



Bundesministerium  
für Arbeit und Soziales

FORSCHUNGSBERICHT 628

# Untersuchung internationaler Dateninnovationen in der Arbeitsmarkt- und Sozialpolitik

November 2023

ISSN 0174-4992



**Finanziert von der  
Europäischen Union**

NextGenerationEU

# Untersuchung internationaler Dateninnovationen in der Arbeitsmarkt- und Sozialpolitik

**Institut für Angewandte Wirtschaftsforschung e.V.**  
Schaffhausenstraße 73  
72072 Tübingen



In Kooperation mit

**RWI – Leibniz-Institut für Wirtschaftsforschung**  
Hohenzollernstraße 1–3  
45128 Essen



## **Autorinnen und Autoren:**

Ronald Bachmann  
Bernhard Boockmann  
Philipp Kugler  
Paula Roth  
Christina Vonnahme  
Jan Simon Wiemann

## **unter Mitarbeit von:**

Rachel Kühn  
Dilara Öcal  
Marc Schmidt  
Anastasia Suckau  
Jurek Tiedemann  
Diem-Ly Tran  
Lena Walser

Oktober 2023

Erstellt im Auftrag des Bundesministeriums für Arbeit und Soziales.

Die Durchführung der Untersuchungen sowie die Schlussfolgerungen aus den Untersuchungen sind von den Auftragnehmern in eigener wissenschaftlicher Verantwortung vorgenommen worden. Das Bundesministerium für Arbeit und Soziales übernimmt insbesondere keine Gewähr für die Richtigkeit, Genauigkeit und Vollständigkeit der Untersuchungen.

Finanziert im Rahmen des BMAS-Datenlabors durch Mittel der Europäischen Union („NextGenerationEU“)

## Kurzbeschreibung

Das Ziel dieser Untersuchung ist die Identifikation und Beschreibung internationaler Beispiele für eine innovative Nutzung von Daten in der Arbeitsmarkt- und Sozialpolitik. Dabei soll auch die mögliche Übertragbarkeit in den deutschen Kontext berücksichtigt werden. Die dargestellten Innovationen umfassen das gesamte Themenspektrum des Bundesministeriums für Arbeit und Soziales (BMAS), d.h. von der Arbeitsmarkt- und Beschäftigungspolitik einschließlich beruflicher Weiterbildung über Sozialversicherung und Alterssicherung bis zu sozialer Teilhabe. Die für jede Innovation verfassten Steckbriefe basieren auf Erkenntnissen aus Expert\*innengesprächen, einem internationalem Forschungsnetzwerk und einer Literaturrecherche bei Behörden, Forschungseinrichtungen und Datenproduzierenden. Die Bandbreite der relevanten Innovationen reicht von Verfahren im Bereich der Datengenerierung über Prozesse der Datenaufbereitung und -verknüpfung zu Methoden und Technologien im Bereich des Machine Learning und Natural Language Processing.

## Abstract

This study aims to identify and describe international best practices of an innovative use of data in labour market and social policy. In particular, data innovations that are transferable to the German context are of interest. Hence, the conditions of a transfer to Germany are also described. The innovations discussed cover the entire range of topics of the Federal Ministry of Labour and Social Affairs, i.e. labour market and employment policy, including vocational training, social insurance and old-age provision, as well as social participation. The reports written for each of the data innovations are based on findings from expert interviews, an international research network and a literature review covering public offices, research institutions and data producers. The spectrum of relevant innovations ranges from procedures in the area of data generation to processes of data preparation and linking to methods and technologies in the area of Machine Learning as well as Natural Language Processing.

# Inhalt

<b>Kurzbeschreibung</b>	<b>5</b>
<b>Abstract</b>	<b>5</b>
<b>Inhalt</b>	<b>6</b>
<b>Tabellenverzeichnis</b>	<b>8</b>
<b>Abbildungsverzeichnis</b>	<b>9</b>
<b>Abkürzungsverzeichnis</b>	<b>10</b>
<b>Zusammenfassung</b>	<b>13</b>
<b>1. Zielsetzung des Vorhabens</b>	<b>16</b>
<b>2. Ausgangslage und Studiendesign</b>	<b>17</b>
2.1 Ausgangslage in Deutschland	17
2.2 Studiendesign	19
<b>3. Überblick über die Steckbriefe</b>	<b>25</b>
3.1 Datengenerierung und Datenverknüpfung	25
3.2 Methoden	28
3.3 Datenbereitstellung	32
<b>4. Steckbriefe zur Datengenerierung und Datenverknüpfung</b>	<b>34</b>
4.1 Hub- and-Spokes-Architektur und homomorphe Verschlüsselung zur Erstellung von Statistiken (Republik Korea)	34
4.2 Multi-Agency Data Integration Project (MADIP) (Australien)	41
4.3 Proaktive Sozialleistungen durch die Verwirklichung des Once-Only-Prinzips (Estland)	48
4.4 Opportunity Insights: Eine Datenbasis zum sozialen Aufstieg (USA)	54
4.5 Forschung zu Erwerbsminderungsrenten: Verknüpfung von Versicherungs- und Gesundheitsdaten (Niederlande)	65

4.6	Verknüpfung von Stellenanzeigen und Bewerbungen aus einem Online-Stellenportal (Uruguay)	71
4.7	Synthetische Datengeneratoren (Vereinigtes Königreich)	77
<b>5.</b>	<b>Steckbriefe zu innovativen Methoden</b>	<b>82</b>
5.1	Statistische Vorhersagemodelle zur Unterstützung der Beratung und Vermittlung von Arbeitslosen (Portugal)	82
5.2	Machine-Learning-Methoden in Frühwarnsystemen für Leistungsbezug (Australien)	89
5.3	Die Vorhersage von Renteneintrittsentscheidungen durch Machine Learning (mehrere Länder)	95
5.4	Machine-Learning-Verfahren für die Mindestlohnevaluation (USA)	100
5.5	Intelligent Rehabilitation Recommendation System (Republik Korea)	106
5.6	Unterstützung von Betriebsprüfungen zur Aufdeckung von nicht gesetzeskonformem Verhalten (Belgien)	110
5.7	Automatisierte Kategorisierung von Unfällen am Arbeitsplatz (USA)	113
5.8	Sandbox: Ein Reallabor für die Nutzung von Machine-Learning-Methoden (Norwegen)	117
5.9	Randomisierte Kontrollstudie zu Bewerbungsaktivitäten von Arbeitssuchenden (Dänemark)	124
<b>6.</b>	<b>Steckbriefe zu innovativen Datenbereitstellungen</b>	<b>131</b>
6.1	Social Policy Simulation Database and Model (Kanada)	131
6.2	Jobnet und Joblog – Portal für Arbeitssuchende mit Verknüpfungsmöglichkeiten (Dänemark)	142
6.3	The Data-Driven Skills Taxonomy (Vereinigtes Königreich)	148
6.4	JEDI – Jobs and Education Data Infrastructure (Australien)	156
6.5	Job- und Informationsportal auf der Basis von Open Government Data (Spanien)	163
6.6	Nowcast von Beschäftigung und Daten-Dashboard (Australien)	167
<b>7.</b>	<b>Zusammenfassung und Schlussfolgerungen</b>	<b>175</b>
	<b>Literaturverzeichnis</b>	<b>179</b>

# Tabellenverzeichnis

Tabelle 2-1: Dateninnovationen nach Datentyp	21
Tabelle 2-2: Dateninnovationen nach Art der Datenverarbeitung und -bereitstellung	22
Tabelle 2-3: Dateninnovationen nach Ursprungsland	23
Tabelle 3-1: Dateninnovationen mit Fokus Datengenerierung und -verknüpfung	26
Tabelle 3-2: Dateninnovationen mit Fokus Methoden	29
Tabelle 3-3: Dateninnovationen mit Fokus Datenbereitstellung	33
Tabelle 4-1: Datenquellen für Forschungs- und Datenprojekte von Opportunity Insights	56
Tabelle 6-1: Übersicht der Datengrundlagen des SPSPD/M	132
Tabelle 6-2: Verwendete Datenreihen	168

# Abbildungsverzeichnis

Abbildung 2-1: Ablauf des Screeningprozesses in der ersten Projektphase	20
Abbildung 4-1: Hub-and-Spokes-Architektur zur Erstellung von Rentenstatistiken	35
Abbildung 4-2: Homomorphe Verschlüsselung mehrerer Datensätze	37
Abbildung 4-3: The Opportunity Atlas	59
Abbildung 4-4: Economic Tracker	60
Abbildung 5-1: Bedienoberfläche für die Vermittlungsfachkräfte	86
Abbildung 5-2: Genauigkeit und Trefferquote der alternativen Verfahren	102
Abbildung 5-3: Online-Tool zur Information über die individuelle Erfüllung von Pflichten	126
Abbildung 6-1: Auswahl und Variation der Parameter des SPSM	136
Abbildung 6-2: Ergebnisdarstellung einer Simulation	137
Abbildung 6-3: Jobnet-Portal	144
Abbildung 6-4: Überblick über die Methodologie	149
Abbildung 6-5: Anwendung der Kompetenz-Taxonomie - Berufe	151
Abbildung 6-6: Anwendung der Kompetenz-Taxonomie - Wert der Kompetenzen	152
Abbildung 6-7: Anwendung der Kompetenz-Taxonomie - Arbeitsnachfrage und Einkommen	153
Abbildung 6-8: Berechnung der Ähnlichkeit von Berufen	158
Abbildung 6-9: Job Switch Portal, beispielhafte Anwendung	159
Abbildung 6-10: NERO-Dashboard: Überblick Beruf Buchhalter*in	171
Abbildung 6-11: NERO-Dashboard: Zeitliche Entwicklung Beruf Buchhalter*in	171
Abbildung 6-12: NERO-Dashboard: Regionale Beschäftigungsentwicklung nach Berufen	172

# Abkürzungsverzeichnis

3mm-Modell	3-more-months Modell (ebenso: 6mm-, 9mm- und 12mm-Modell)
ABS	Australian Bureau of Statistics
ADEST	Automatisierte Datenübernahme aus Stellenangeboten mittels maschinellen Lernens
AdL	Alterssicherung der Landwirte
ADPP	Automated Data Processing Pipeline
ANZSCO	Australian and New Zealand Standard Classification of Occupations
ASC	Australian Skills Classification
ASiD	Alterssicherung in Deutschland
AU	Arbeitsunfähigkeit
AUC	Area under the ROC curve
BA	Bundesagentur für Arbeit
BAMF	Bundesamt für Migration und Flüchtlinge
BAuA	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin
BASiD	Biografiedaten ausgewählter Sozialversicherungsträger in Deutschland
BAV	Betriebliche Altersversorgung
BDSG	Bundesdatenschutzgesetz
BERT	Bidirectional Encoder Representations from Transformers
BG ETEM	Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse
BLS	Bureau of Labor Statistics
BMAS	Bundesministerium für Arbeit und Soziales
BMFSFJ	Bundesministerium für Familie, Senioren, Frauen und Jugend
BMG	Bundesministerium für Gesundheit
BSV	Berufsständische Versorgung
BV	Beamtenversorgung
CAD	Canadian Dollar
CAS	Classificaties voor Arbo en SV
CBS	Centraal Bureau voor de Statistiek
COMTAX	Commodity Tax Model
COMWEL	Korea Workers' Compensation and Welfare Service
CPS	Current Population Survey
CPS-ORG	CPS Outgoing Rotation Group
DEK	Datenethikkommission
DGUV	Deutsche Gesetzliche Unfallversicherung
DIGIPEG	Digitale Peer-Gruppen-Beratung
DIW	Deutsches Institut für Wirtschaftsforschung
DOMINO	Data On Multiple Individual Occurrences
DOMINO CAD	DOMINO Centrelink Administrative Data
DP	Discussion Paper
DRV	Deutsche Rentenversicherung
DSGVO	Datenschutzgrundverordnung
DSRV	Datenstelle der Rentenversicherung
DSS	Department of Social Services in Australien
DVG	Digitale-Versorgung-Gesetz
EESSI	Electronic Exchange of Social Security Information
Efa	"Einer für Alle"-Prinzip
ELFE	Einfache Leistungen für Eltern

EMSIM	RWI-Einkommensteuer-Mikrosimulationsmodells
ESCO	European Skills, Competences, Qualifications and Occupations
ESCoE	Economic Statistics Centre of Excellence
ESS	Employment Service of Slovenia
EU	Europäische Union
EU-LFS	EU Labor Force Survey
EU-SILC	EU Statistics on Income and Living Conditions
EVS	Einkommens- und Verbrauchsstichprobe
FAST	Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik
FDZ	Forschungsdatenzentrum
FITKO	Föderale IT-Kooperation
FKS	Finanzkontrolle Schwarzarbeit
GAN	Generative Adversarial Network
GBDT	Gradient Boosted Decision Tree
GETTSIM	German Taxes and Transfer Simulator
GPU	Graphics Processing Unit
GRV	Gesetzliche Rentenversicherung
HE	Homomorphic Encryption
IAB	Institut für Arbeitsmarkt- und Berufsforschung
IAB-MSM	IAB-Mikrosimulationsmodell
IAW	Institut für Angewandte Wirtschaftsforschung
IDNr	Steuerliche Identifikationsnummer
IEB	Integrierte Erwerbsbiografien
IEFP	Instituto de Emprego e Formação Profissional
IfSG	Infektionsschutzgesetz
ifo-MSM	ifo Microsimulation Model
IRRS	Intelligent Rehabilitation Recommendation System
ISSN	International Standard Serial Number
IVA	Inkomensvoorziening Volledig Arbeidsongeschikten
IZA	IZA - Institute of Labor Economics
JEDI	Jobs and Education Data Infrastructure
KI	Künstliche Intelligenz
KldB2010	Klassifikation der Berufe 2010
LASSO	Least absolute shrinkage and selection operator
LWR	Laufende Wirtschaftsrechnung
MADIP	Multi-Agency Data Integration Project
MAGS NRW	Ministerium für Arbeit, Gesundheit und Soziales des Landes Nordrhein-Westfalen
MCD	Medicare Consumer Directory
NBER	National Bureau of Economic Research
NERO	Nowcast of Employment by Region and Occupation
NFDI	Nationale Forschungsdateninfrastruktur
NLP	Natural Language Processing
NOW	Nationale Online-Weiterbildungsplattform
NRO	Nichtregierungsorganisation
NRW	Nordrhein-Westfalen
NSC	National Skills Commission
o. D.	Ohne Datumsangabe
OECD	Organisation for Economic Cooperation and Development
OGD	Open Government Data

OJV	Online Job Vacancies
ONS	Office for National Statistics
ONSS	Office Nationale de Sécurité Sociale
ORG	Outgoing Rotation Group
OZG	Onlinezugangsgesetz
OZGÄndG	OZG-Änderungsgesetz
O*NET	Occupational Information Network
PBO	Parliamentary Budget Office
PIT	Personal Income Tax
PREDOC	Pathways to Research and Doctoral Careers
RCT	Randomized Controlled Trial
RatSWD	Rat für Sozial- und Wirtschaftsdaten
RPA	Robotic Process Automation
ROC	Receiver Operating Characteristic Curve
RTZN	Versichertenrentenzugang
RWI	RWI – Leibniz-Institut für Wirtschaftsforschung
SDG	Synthetische Datengeneratoren
SDG-VO	Single-Digital-Gateway-Verordnung
SFTP	Secure File Transfer Protocol
SGB	Sozialgesetzbuch
SHARE	Survey on Health, Ageing, and Retirement in Europe
SHS	Survey of Household Spending
SIAB	Stichprobe der Integrierten Arbeitsmarktbiografien des IAB
SLID	Survey of Labour and Income Dynamics
SOEP	Sozio-oekonomisches Panel
SOEP-RV	Verknüpfung des SOEP mit administrativen Daten der Rentenversicherung
SPF	Service Public Fédéral
SPSD/M	Social Policy Simulation Database and Model
STAR	Styrelsen for Arbejdsmarked og Rekruttering
StatCan	Statistics Canada
StGB	Strafgesetzbuch
SV-Nummer	Sozialversicherungsnummer
SVR	Support Vector Regression
TPU	Tensor Processing Unit
TrEffeR	Treatment Effects and Prediction
USD	US-Dollar
UV	Unfallversicherung
UWV	Uitvoeringsinstituut Werknemersverzekeringen
VAE	Variational Autoencoder
vbw	Vereinigung der Bayerischen Wirtschaft e. V.
VE	Verdiensterhebung
VerBIS	Vermittlungs-, Beratungs- und Informationssystem
VSKT	Versicherungskontenstichprobe
WIA	Wet Werk en Inkomen naar Arbeidsvermogen
WIT	Waterford Institute of Technology
WP	Working Paper
ZÖD	Zusatzversorgung im öffentlichen Dienst

# Zusammenfassung

Datengetriebene Anwendungen und Prozesse spielen eine immer größere Rolle in allen Lebens- und Wirtschaftsbereichen. Dies gilt nicht zuletzt für den Arbeitsmarkt. Das Spektrum reicht dabei von Onlineportalen im Internet, beispielsweise zur Stellensuche, bis zu neuen Verfahren der Arbeitsmarktstatistik. Diese Entwicklung basiert unter anderem auf Fortschritten im Bereich der Künstlichen Intelligenz, insbesondere Machine Learning, und der Verfügbarkeit von Big Data. Durch die Datenstrategie der Bundesregierung soll die innovative und verantwortungsvolle Datenbereitstellung und Datennutzung in der Wirtschaft, der Wissenschaft, der Zivilgesellschaft und der Verwaltung gefördert werden. Daten sollen dabei helfen, dass Entscheidungen auf einer besseren Grundlage getroffen werden und dass Regulierungen, Fördermaßnahmen und Dienstleistungen auf die Bedürfnisse der Bürger\*innen, der Wirtschaft und Wissenschaft eingehen.

Ziel der vorliegenden Studie ist es, Beispiele aus anderen Ländern für eine innovative Nutzung von Daten und Methoden in der Arbeitsmarkt- und Sozialpolitik darzustellen. Angesichts der Breite des Themenfeldes deckt die Auswahl der Beispiele nicht alle Dateninnovationen ab. Sie kann aber einen guten Eindruck davon vermitteln, welche unterschiedlichen Innovationen international bereits eingeführt wurden. Neben der Beschreibung einer Innovation wird auch auf den aktuellen Stand ähnlicher Anwendungen in Deutschland und relevante Rahmenbedingungen eingegangen, um eine erste Einschätzung zur Übertragbarkeit der identifizierten internationalen Dateninnovationen zu vermitteln. Es werden Verfahren und Dienstleistungen aus dem Bereich der Verwaltung und aus der Forschung einbezogen. Die Trennung ist dabei oft nicht scharf, denn Innovationen in der Forschung und in der Verwaltung bedingen sich gegenseitig.

Am Anfang der Untersuchung stand eine Recherchephase (Screening-Phase), in der zunächst rund 90 Dateninnovationen identifiziert wurden. Die Recherche berücksichtigte sowohl wissenschaftliche Publikationen in renommierten Fachzeitschriften oder Working Paper-Reihen als auch Innovationen, die von administrativer Seite entwickelt wurden. Die recherchierten Dateninnovationen wurden hinsichtlich ihrer Relevanz für Deutschland und weiteren Qualitätsmerkmalen bewertet. Daraus wurden 22 besonders relevante Dateninnovationen für eine vertiefte Darstellung ausgewählt. Die Ergebnisse werden in Steckbriefen für jede Dateninnovation dargestellt. Ein Steckbrief enthält zunächst eine Einordnung in den jeweiligen Kontext und fasst das Innovationspotenzial für einen Transfer nach Deutschland zusammen. Anschließend folgt eine technische Beschreibung des zentralen Innovationselements und relevanter Erkenntnisse, welche die beteiligten Akteure im Ausland bei der Entwicklung der Innovation gemacht haben. Abschließend wird in den Steckbriefen die Übertragbarkeit nach Deutschland diskutiert.

Die identifizierten und beschriebenen Dateninnovationen aus dem Ausland lassen sich in drei Kategorien gliedern: Dateninnovationen mit Fokus auf a) der Datenbasis, insbesondere im Hinblick auf die Generierung und Verknüpfung von Daten b) der Methode der Datennutzung und c) der Bereitstellung der Daten für bestimmte Zielgruppen.

Bei den Steckbriefen zur Datenbasis dominiert die Verknüpfung von unterschiedlichen Daten, insbesondere solchen Daten, die im Rahmen administrativer Prozesse erhoben werden. Die Verknüpfung verbindet diese Daten in der Regel auf der Ebene einzelner Personen. Dabei kommen mitunter komplexe Verfahren zum Einsatz, um den Datenschutz zu gewährleisten. Dies erweitert das Merkmalsspektrum der zur Verfügung stehenden Daten und damit die Nutzungsmöglichkeiten erheblich.

Im Bereich der Methoden nutzen viele der Innovationen Verfahren des Machine Learning (maschinelles Lernen). Diese Methodik ist z. B. für verbesserte Vorhersagen geeignet, beispielsweise

für die Vorhersage des Zeitpunkts des individuellen Renteneintritts oder für die Vorhersage des Risikos der Langzeitarbeitslosigkeit. Die Anwendung von Machine Learning resultiert meist in deutlich besseren Vorhersagen als die Nutzung konventioneller statistischer Verfahren. Neben dem Einsatz der Machine Learning zu Vorhersagezwecken gibt es eine Reihe von Anwendungen zur automatisierten Bildung von Kategorien und Clustern.

Im Bereich der Bereitstellung von Daten ist das Ziel der unterschiedlichen Verfahren, die Anwendung der Daten zu verbreitern und dabei die handelnden Personen zu einer selbstständigen Nutzung der Daten zu ermächtigen, beispielsweise durch die Schaffung von Datenportalen. Die Bandbreite reicht hier von einem Mikrosimulationsmodell, das für die eine breite Fachöffentlichkeit zugänglich ist, über die Bereitstellung von aktuellen Beschäftigungsdaten auf regionaler und beruflicher Ebene bis hin zur Bereitstellung eines kleinräumigen Prognosemodells zur Arbeitsmarktentwicklung.

Aus der Analyse der einzelnen Dateninnovationen lassen sich einige allgemeine Schlussfolgerungen ziehen. Die wichtigsten sind die folgenden:

- Die Verknüpfung verschiedener Datenquellen bietet häufig einen hohen Mehrwert für konkrete Prozesse oder Fragestellungen in der Verwaltung, aber genauso für Forschungsfragen. Durch die Verknüpfung gelingt es, das in einzelnen administrativen Datensätzen oft enge, an den Verwaltungszwecken orientierte Merkmalspektrum zu erweitern. Technisch verwandt mit der Verknüpfung ist der Austausch von Daten zwischen Verwaltungen im Zuge der Verwirklichung des Once-Only-Prinzips.
- Im Bereich der Daten ermöglichen Big Data neue und – insbesondere in Verknüpfung mit Machine-Learning-Methoden – zeitnahe Analysen von Arbeitsmarktentwicklungen. Beispielsweise kann mithilfe von Informationen aus Stellenanzeigen aus Online-Portalen (Online Job Vacancies, OJV) der Bedarf an bestimmten Fähigkeiten und Kenntnissen ermittelt werden. Die so gewonnenen Informationen können einzelnen Personen bei der Entscheidungsfindung, z. B. im Hinblick auf die Berufswahl, helfen und unterstützen Vermittlungsfachkräfte und andere in der Verwaltung Tätige.
- Machine Learning kann besonders für Vorhersagen auf der individuellen Ebene genutzt werden. So lassen sich beispielsweise Prognosen zum individuellen Erwerbszustand (z. B. Arbeitslosigkeit und Bezug von Lohnersatzleistungen, Renteneintritt, dauerhafte Erwerbsunfähigkeit) treffen oder die Wahrscheinlichkeit von Compliance-Verstößen z. B. beim Arbeitsschutz oder beim Mindestlohn berechnen. Viele der betrachteten Verfahren finden sich noch in einem Pilotstadium, sodass hier weiteres Anwendungspotenzial besteht. Dabei gilt es, die Chancen mit den Risiken – zum Beispiel möglichen Verzerrungen – abzuwägen.
- Verfahren des Machine Learning können auch zur Kategorisierung und Bündelung, z. B. von großen Datenmengen aus Stellenportalen, genutzt werden. Zusammen mit der automatischen Verarbeitung von Textdaten ergeben sich verbesserte Informationsgrundlagen, unter anderem im Bereich der Arbeitsvermittlung und der Unfallversicherung.
- Daten- und Informationsportale erleichtern es unterschiedlichen Zielgruppen, auf die Ergebnisse von Datenanalysen und Datenverarbeitungen zurückzugreifen.
- Alle betrachteten Dateninnovationen haben Transferpotenzial für die Nutzung in Deutschland oder können zumindest Impulse für ähnliche Innovationen in Deutschland geben, allerdings in unterschiedlichem Ausmaß und auf unterschiedliche Weise. In manchen Fällen ist ein direkter Transfer einer internationalen Dateninnovation denkbar. Dies ist insbesondere dann der Fall, wenn es in Deutschland bereits ähnliche Verfahren oder Vorstufen zu den Dateninnovationen gibt, sodass eine Implementierung in Deutschland auf bereits verwirklichten Verfahren aufsetzen kann. In vielen Fällen ist ein direkter Transfer

jedoch nicht möglich, weil kritische Voraussetzungen nicht gegeben sind. Hier kommt vielfach eine modifizierte Umsetzung in Betracht oder der Transfer einer allgemein verwendbaren Methodik von einem Anwendungsfeld in ein anderes.

- Zwar bestehen allgemein günstige Bedingungen für den Transfer von Dateninnovationen nach Deutschland, insbesondere der vorhandene Bestand an Daten und die Erfahrung mit Daten und Methoden in Verwaltungen und Forschungseinrichtungen. Es gibt jedoch auch eine Reihe von Herausforderungen für den Transfer. Für die Anwendung vieler Dateninnovationen fehlt derzeit noch ein klarer Rechtsrahmen. Die Zusammenführung von Daten wird teilweise durch den föderalen Staatsaufbau erschwert. Schließlich ist die Einrichtung von datenbasierten Verwaltungsverfahren in Deutschland erfahrungsgemäß sehr aufwändig.
- Für die genannten Hemmnisse stehen gleichzeitig vielfach Lösungsmöglichkeiten bereit. Reallabore oder Sandboxes könnten Datenschutzkonformität prüfen und nutzen dabei Aufsichtsbehörden wie auch den Nutzenden. Teilweise lässt sich der Datenschutz selbst durch innovative Datenstrukturen und Verschlüsselungsmethoden gewährleisten.

Bei einem Transfer der Dateninnovationen oder einzelner ihrer Elemente sollten die vorliegenden Erfahrungen aus anderen Ländern genau berücksichtigt werden. Hierfür ist im Einzelfall allerdings eine tiefere Analyse erforderlich, als sie im Rahmen dieses Berichts geleistet werden kann.

# 1. Zielsetzung des Vorhabens

Durch die Datenstrategie der Bundesregierung soll die innovative und verantwortungsvolle Datenbereitstellung und Datennutzung in der Wirtschaft, der Wissenschaft, der Zivilgesellschaft und der Verwaltung gefördert werden. In Politik und Verwaltung sollen Daten dabei helfen, Entscheidungen auf eine solidere Grundlage zu stellen und Regulierungen, Fördermaßnahmen und Dienstleistungen zu schaffen, die besser auf die Bedürfnisse der Bürger\*innen, der Wirtschaft und Wissenschaft eingehen. Eine verbesserte Datennutzung in der Verwaltung soll zu einer effizienteren und bürger\*innenfreundlicheren Verwaltungspraxis führen.

Der zunehmende Bedarf an den Potenzialen bisher nicht erschlossener Daten stößt auf verbesserte Möglichkeiten der Gewinnung und Nutzung von Daten. Hierzu tragen innovative Verfahren und Produkte auf der Basis neuer Methoden in der Forschung bei, insbesondere im Bereich des Machine Learning und der Verknüpfung von Datensätzen, aber auch die stärkere Etablierung von Datenwissenschaften an den Hochschulen bei, beispielsweise durch die Aufnahme von Studiengängen zum Thema „*Data science*“ in die Curricula der Hochschulen.

Innovative Daten und Datennutzungen sind im Bereich der Arbeitsmarkt- und Sozialpolitik in Deutschland bereits vorhanden. So werden Daten aus administrativen Prozessen, oft in Verknüpfung mit Befragungsdaten, zur Analyse der Zielgenauigkeit, Effektivität und Wirtschaftlichkeit von Maßnahmen des Gesetzgebers oder der Verwaltung genutzt. In einigen Bereichen, sei es in der Nutzung von Open Government Data, in der Bereitstellung von Daten aus unterschiedlichen administrativen Prozessen oder in der Nutzung von Big Data durch die Forschung, sind andere Länder allerdings schon weiter fortgeschritten als Deutschland. Dies zeigt sich unter anderem an den Rankings, die internationale Organisationen wie die OECD veröffentlichen und die beispielsweise die Datenverfügbarkeit, den Datenzugang und das Teilen von Daten erfassen (OURdata Index; OECD, 2020a).

Das Ziel der vorliegenden Studie ist es, internationale Beispiele für die innovative Nutzung von Daten in der Arbeitsmarkt- und Sozialpolitik sowie den Sozialverwaltungen zu identifizieren und näher zu beschreiben, die auch für den deutschen Kontext mögliche Anwendungsfälle darstellen könnten. Angesichts der Breite des Themenfeldes kann keine Vollständigkeit geboten werden, es soll jedoch ein Eindruck vermittelt werden, welche unterschiedlichen Innovationen international bereits eingeführt wurden. Dabei geht es zum einen um die Nutzung von Dateninnovationen durch die Verwaltung und die Verbesserung der Transparenz politischer Entscheidungen durch eine bessere Information der Bürger\*innen. Zum anderen geht es um die verbesserte Datenversorgung für die Forschung, die so in die Lage versetzt wird, Zielgenauigkeit und Wirkung von Maßnahmen und Leistungen besser zu ermitteln.

Die im Rahmen der Studie identifizierten Beispiele haben alle einen praktischen Anwendungsbezug zu Themen des Bundesministeriums für Arbeit und Soziales (BMAS). Sie betreffen sowohl die Arbeit des Ministeriums als auch die der Behörden in seinem Geschäftsbereich, z. B. die Bundesagentur für Arbeit (BA), die Deutsche Rentenversicherung (DRV) oder der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA). Adressat\*innen für die Anwendung der Beispiele in Deutschland sind jedoch auch die Bürger\*innen. Die zugrundeliegende Recherche deckt das gesamte Themenspektrum des BMAS ab, von der Arbeitsmarkt- und Beschäftigungspolitik einschließlich beruflicher Weiterbildung über den Arbeitsschutz, das Arbeitsrecht, die Sozialversicherung und Alterssicherung bis zur sozialen Teilhabe und Inklusion. Für die Bearbeitung des Auftrags nutzte das Projektteam umfangreiche Erfahrungen in der Erhebung und wissenschaftlichen Nutzung von Daten, in der empirischen Arbeitsmarkt- und Sozialpolitikforschung sowie bestehende Kontakte zu Stakeholdern und Wissenschaftler\*innen im In- und Ausland.

## 2. Ausgangslage und Studiendesign

### 2.1 Ausgangslage in Deutschland

Zur derzeitigen Verwirklichung von Dateninnovationen in Deutschland im Bereich der Arbeitsmarkt- und Sozialpolitik gibt es bislang keinen systematischen Überblick. Einen solchen Überblick zu schaffen, würde eine erhebliche Herausforderung darstellen. Denn im Bereich der Sozialversicherung, in den Statistischen Ämtern, in Forschungseinrichtungen und in vielen weiteren Institutionen werden eine große Vielzahl an Daten erhoben, verarbeitet und zu Informationszwecken wieder zur Verfügung gestellt. Neben dieser institutionellen Vielgestaltigkeit besteht die Herausforderung auch darin, dass sich die Arbeitsmarkt- und Sozialpolitik in viele Themenfelder ausdifferenziert. Im Folgenden können daher nur einige Anwendungen und Pionierprojekte für innovative Datennutzungen und digitale Verfahren genannt werden.

Im Bereich der Sozialversicherung soll beispielsweise das EESSI-Verfahren (Electronic Exchange of Social Security Information) den papierbasierten Austausch in der europäischen Sozialverwaltung durch elektronische Prozesse ersetzen. Anknüpfend an die im EESSI-Verfahren entwickelten Prozesse wurde in einer Machbarkeitsstudie (Becker et al., 2022) erkundet, welche weiteren Digitalisierungspotenziale es im Bereich der sozialen Sicherung gibt und welche IT-technischen Anforderungen hierfür bestehen. Es wurden acht mögliche Maßnahmenbereiche künftiger Digitalisierung im Bereich der Sozialversicherungen identifiziert. Sie reichen von der Automatisierung von Prozessen durch die Robotic Process Automation (RPA)-Technologie über IT-basierte Anomalie-Erkennung, Fraud-Detection und Missbrauchsprävention bis hin zu bürokratiearmen Verfahren des „One-Stop-Shop“ und „No-Stop-Government“. Das letztgenannte Verfahren bedeutet, dass die Verwaltung bei der Gewährung von Leistungen nicht auf Antrag tätig wird, sondern proaktiv an die Berechtigten herantritt.

Im Bereich der Arbeitsvermittlung und Arbeitsmarktpolitik wird in einem Projekt des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) die Akzeptanz von KI-Anwendungen in der öffentlichen Arbeitsvermittlung erforscht.<sup>1</sup> Es wird untersucht, inwieweit mit KI ausgestattete Tools zur Unterstützung der Vermittlungs- und Beratungsarbeit in Agenturen für Arbeit auf Akzeptanz bei den Vermittlungsfachkräften stoßen. Das Projekt kann auf vorhergehende Erfahrungen im Bereich der BA zurückgreifen. Schon in den 2000er-Jahren wurde mit dem Projekt „TrEffeR“ der BA den Agenturen für Arbeit ein Tool angeboten, das sie dabei unterstützen sollte, Arbeitslose in die Maßnahmen zu vermitteln, die sie am besten auf der Suche nach einer Beschäftigung unterstützen (Stephan et al., 2006). Hierzu wurden Ergebnisse einer regional differenzierten retrospektiven Wirkungsanalyse verwendet. Die individuelle Beratung und Vermittlung kann durch digitale Kommunikationskanäle ergänzt werden. Im Modellprojekt Digitale Peer-Gruppen-Beratung (DIGIPEG) unterstützen sich Arbeitsuchende gegenseitig in einem Online-Chat, der von der Arbeitsagentur angeboten und von zwei Vermittlungsfachkräften moderiert wird (Bähr et al., 2022).

Ein wichtiges Anwendungsfeld digitaler Technologien bieten auch die Daten von Stellenanzeigen. Auf dieser Basis können z. B. die nachgefragten Kompetenzen näher analysiert werden. Im Modellprojekt „Kompetenz-Kompass“ des IAB wurde ein Verfahren entwickelt, das aktuelle fachliche und überfachliche Kompetenzanforderungen aus veröffentlichten Stellenangeboten der BA-Jobbörse identifizierte (Stops et al., 2021). Diese Kompetenzanforderungen lassen sich nach Berufen, Regionen und Anforderungsniveaus auswerten und über die Zeit beobachten. Damit kann z. B. die Bedeutung

---

<sup>1</sup> <https://iab.de/projekt/?id=12498231>

von fachlichen und überfachlichen Anforderungen ermittelt werden. Auf dieser Basis kann auch die zunehmende Verbreitung von technologie-spezifischen Anforderungen untersucht werden, beispielsweise Kompetenzen für die Wasserstofftechnologie (Grimm et al., 2021).

Im Projekt ADEST (Automatisierte Datenübernahme aus Stellenangeboten mittels maschinellen Lernens), das von der BA seit Juli 2022 für operative Prozesse entwickelt wird, wird eine Nutzung von Textverarbeitung verwendet, um Informationen aus Stellenanzeigen zu entnehmen. Die BA verknüpft Stellenanzeigen von Unternehmen mit Profilen von Arbeitsuchenden, um Letzteren gezielt bei der Arbeitssuche zu helfen. Für die Verknüpfung der Angebots- und Nachfrageseite werden konkrete Informationen zu den Stellenanzeigen benötigt, wie beispielsweise die für den Job benötigten Qualifikationen. Die der BA vorliegenden Stellenanzeigen von Unternehmen sind allerdings häufig in Fließtext verfasst, sodass Beschäftigte der BA die notwendigen Informationen aufwändig extrahieren müssen. ADEST übernimmt inzwischen diese Informationsextraktion und erstellt eine Matrix für jede Stellenanzeige, welche die notwendigen Informationen enthält und in die Infrastruktur zur Verknüpfung eingespielt werden kann. Dadurch werden Beschäftigte der BA von administrativen Tätigkeiten entlastet und können vermehrt in der persönlichen Beratung eingesetzt werden. Aktuell befindet sich das Projekt in der Pilotphase. Bis Ende des ersten Quartals 2024 soll ADEST flächendeckend eingesetzt werden (Projekt ADEST, 2022). Darüber hinaus wird von der BA aktuell evaluiert, ob für die internen Qualitätssicherungen auf Unternehmenswebseiten zugegriffen und Informationen extrahiert werden können.

Im Bereich der Arbeitsmarktstatistik veröffentlicht das Statistische Bundesamt in der Rubrik "EXDAT – Experimentelle Daten" innovative Projektergebnisse auf der Basis neuer Datenquellen und Methoden.<sup>2</sup> Im Bereich Arbeit und Soziales wurden in diesem Kontext Pilotdaten zu georeferenzierten SGB II-Hilfequoten bereitgestellt. Dazu wurden Ergebnisse aus der Grundsicherungsstatistik der BA und zum anderen die anonymisierten Daten des Zensus 2011 herangezogen. Das Statistische Bundesamt befasst sich ferner damit, aus Anzeigen von Stellenportalen Indikatoren zur Entwicklung der Arbeitskräftenachfrage auf Online-Stellenmärkten zu entwickeln. Ein weiterer innovativer Ansatz im Bereich der Arbeitsmarktstatistik ist ein Nowcast-Verfahren für die Anzahl der sozialversicherungspflichtig Beschäftigten. Informationen darüber sind bislang nur mit einer Datenaktualität von sechs Monaten zum jeweiligen Zeitpunkt erhältlich. Mithilfe von Informationen zu Beschäftigten mit Versicherungspflicht in der gesetzlichen Krankenversicherung lässt sich jedoch eine präzise zeitnahe Vorhersage errechnen (Hutter, 2020).

Gegenstand von Dateninnovationen ist schließlich die digitale Arbeitswelt selbst. In der Denkfabrik Digitale Arbeitsgesellschaft im BMAS werden Ansätze dazu gebündelt. Hier werden auch spezifische Projekte durchgeführt, beispielsweise das Projekt „KITQAR – KI-Test- und Trainingsdatenqualität in der digitalen Arbeitsgesellschaft“.<sup>3</sup> Hierin sollen Anforderungen an die Datenqualität in Verbindung mit KI-Anwendungen untersucht werden. In einem zweiten Schritt geht es um die Umsetzung und Anwendung von Standards zur Datenqualität. Hierzu werden Fallstudien in unterschiedlichen Einsatzgebieten der KI durchgeführt. In einer weiteren Studie wurden die Chancen und Risiken datengetriebener Produkt- und Preisdifferenzierung untersucht (Bertschek et al., 2021). Dabei geht es insbesondere um die Bereiche Online-Plattformen, Kreditvergabe und Versicherungen. Die Studie zeigt auch die sich daraus möglicherweise ergebenden sozialpolitischen Handlungsbedarfe auf.

---

<sup>2</sup> [https://www.destatis.de/DE/Service/EXDAT/\\_inhalt.html](https://www.destatis.de/DE/Service/EXDAT/_inhalt.html)

<sup>3</sup> <https://www.kitqar.de/de>

Eine neuere Dateninnovation ist die Nationale Online-Weiterbildungsplattform (NOW), die seit 2022 im Rahmen der Nationalen Weiterbildungsstrategie von der BA entwickelt wird.<sup>4</sup> Ziel ist es, künftig ein zentrales und niedrigschwelliges Online-Eingangsportale zum Thema berufliche Weiterbildung anzubieten. Die Plattform soll Informationen über Weiterbildungsangebote bündeln und das komplexe System der beruflichen Weiterbildung digital an einem Ort zugänglich machen. Im Bereich der Rentenversicherung steht seit Juni 2023 die Digitale Rentenübersicht ([www.rentenuebersicht.de](http://www.rentenuebersicht.de)) als Portallösung zur Verfügung, mit der die Bürger\*innen eine Übersicht über ihre persönlichen Altersvorsorgeansprüche aus gesetzlicher, betrieblicher und privater Alterssicherung online abrufen können. Dadurch erhalten sie einen Gesamtüberblick über ihre Altersvorsorge-Situation. Das Portal befindet sich zunächst in der Evaluationsphase, um das Angebot nach und nach zu verbessern und weiterzuentwickeln. Für die Nutzer\*innen sind aktuell noch nicht alle Altersvorsorgeansprüche sichtbar. Die Zahl der am Portal angebotenen Anbieter von Altersvorsorge-Produkten soll stetig zunehmen. So soll das Angebot sukzessive erweitert und attraktiver werden.

Diese Beispiele zeigen, dass derzeit in Deutschland eine Vielzahl neuer Datennutzungen und digitaler Verfahren in vielen Bereichen der Arbeitsmarkt- und Sozialpolitik entwickelt werden. Sie sind vielfach Anknüpfungspunkte für die Verwirklichung der in den Kapiteln 4 bis 6 detailliert dargestellten internationalen Beispiele.

## 2.2 Studiendesign

Um internationale Dateninnovationen mit hohem Transferpotenzial für Deutschland zu identifizieren und zu dokumentieren, sah das methodische Untersuchungskonzept des Projekts mehrere Phasen vor. Die Bearbeitung teilte sich in drei Arbeitspakete:

1. In der Screening-Phase wurden Dateninnovationen recherchiert. Dabei wurden 89 potenziell innovative internationale Datenprodukte identifiziert.
2. Die in der Screening-Phase gewonnene Informationsbasis wurde genutzt, um die Dateninnovationen hinsichtlich ihrer Relevanz für die weitere Betrachtung zu bewerten und auszuwählen.
3. Die ausgewählten Dateninnovationen wurden durch Desk Research und Expert\*innengespräche in Form von 22 Steckbriefen dokumentiert.

### 2.2.1 Screening

Das Ziel der Screening-Phase war eine systematische Zusammenstellung und Beschreibung von potenziell interessanten Dateninnovationen aus anderen Ländern. Hierbei wurden nur solche Dateninnovationen erfasst, die potenziell eine thematische Verbindung zu den Themenbereichen des BMAS aufweisen (Arbeitsmarkt- und Beschäftigungspolitik einschließlich beruflicher Weiterbildung, Arbeitsschutz, Arbeitsrecht, Sozialversicherung und Alterssicherung sowie sozialer Teilhabe und Inklusion). Bei den Dateninnovationen ging es sowohl um Anwendungen aus dem Bereich der öffentlichen Verwaltung als auch um Ergebnisse von Forschungsvorhaben.

Bei der Recherche lag der Schwerpunkt auf den sogenannten Industrieländern. Besonders genau wurden diejenigen Länder betrachtet, die hinsichtlich Dateninnovationen führend sind – beispielsweise Australien im Hinblick auf administrative Daten und ihre Verknüpfung, die USA wegen der Verbreitung von Big-Data-Nutzungen, Estland in Bezug auf die datengestützte Digitalisierung

---

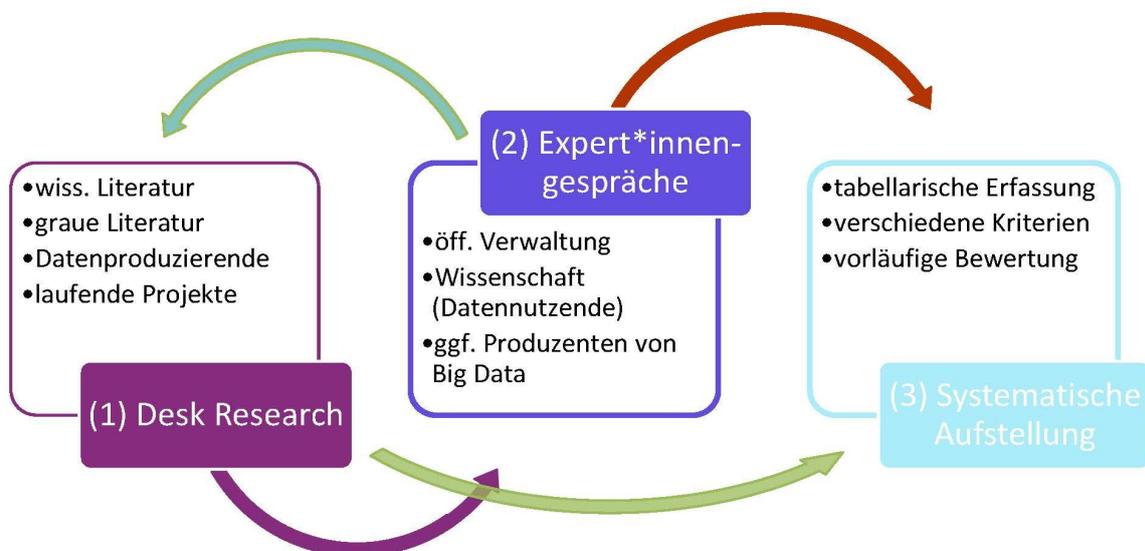
<sup>4</sup> <https://www.bmas.de/DE/Arbeit/Aus-und-Weiterbildung/Weiterbildungsrepublik/Nationale-Online-Weiterbildungsplattform/nationale-online-weiterbildungsplattform.html>

von Verwaltungsprozessen und besonders innovative asiatische Länder wie Südkorea, das den OECD Digital Government Index anführt (OECD, 2020b).

Zur Erfassung der Dateninnovationen wurde ein mehrschrittiges Verfahren durchgeführt (siehe Abbildung 2-1). In einem ersten Schritt (Desk Research) erfolgte eine Identifikation relevanter Beispiele auf Basis einer Sichtung von wissenschaftlicher und grauer Literatur sowie von ergänzenden Internetrecherchen bei möglichen Datenproduzierenden (z. B. Statistische Ämter, Arbeitsverwaltung, zentrale Datenagenturen). Hierbei wurde mit einer Liste von klar spezifizierten Suchbegriffen vorgegangen, die im Laufe der Internetrecherche kontinuierlich aktualisiert wurde. Die Liste und ihre Bearbeitung wurden fortlaufend dokumentiert.

Im Anschluss an das Desk Research wurden im zweiten Schritt Expert\*innengespräche geführt. Die Auswahl erfolgte dynamisch anhand der Ergebnisse des ersten Schritts sowie über bereits bekannte Expert\*innen im Bereich Dateninnovationen aus den Netzwerken des IAW und RWI. Hinweise auf interessante internationale Dateninnovationen aus diesen Gesprächen wurden aufgegriffen und erneut im Desk Research genauer untersucht. Details dazu werden in den Abschnitten 2.2 und 2.3 dargestellt.

**Abbildung 2-1: Ablauf des Screeningprozesses in der ersten Projektphase**



Quelle: Eigene Darstellung

Der konkrete Ablauf der Screeningphase unterschied sich zwischen Dateninnovationen aus der Verwaltung sowie aus der Forschung. Dateninnovationen in der Verwaltung wurden durch eine Internetrecherche sowie durch eine systematische Bestandsaufnahme von Publikationen internationaler Organisationen ermittelt. Ferner wurden sie durch gezielte Recherchen bei Akteur\*innen in den einzelnen Ländern ergänzt. So wurden nach Projektbeginn Expert\*innen aus den statistischen Ämtern und der Wissenschaft angeschrieben und nach deren Kenntnis möglicher Dateninnovationen gefragt.

Bei der Recherche von Dateninnovation aus der Forschung wurde zudem die Tatsache genutzt, dass Forschungsprojekte im Bereich der Arbeitsmarkt- und Sozialforschung, die innovative Datenquellen verwenden und ein wissenschaftlich hohes Niveau aufweisen, häufig in einer von zwei Diskussionspapiererien veröffentlicht werden, unabhängig davon aus welchem Land die Daten stammen oder wo sich die Forschungseinrichtung der beteiligten Wissenschaftler\*innen befindet. Bei

diesen Diskussionspapiererien handelt es sich um die „IZA Discussion Paper Series“ (IZA DP, ISSN: 2365-9793)<sup>5</sup> sowie die „NBER Working Papers“ (NBER WP, ISSN 0898-2937)<sup>6</sup>. Daher wurden, ergänzend zu den genannten Internetrecherchen, alle Veröffentlichungen der IZA DP und der NBER WP im Zeitraum Januar 2017 bis August 2022 systematisch auf die Verwendung innovativer Datenprodukte geprüft. Die so identifizierten Dateninnovationen wurden in Tabellenform erfasst und ausgewertet.

Tabelle 2-1 und Tabelle 2-2 geben einen summarischen Überblick über die identifizierten Dateninnovationen der Screening-Phase. Hinsichtlich des Datentyps beziehen sich viele der erfassten Dateninnovationen auf die Datenquellen Administrative Daten (45) und Big Data (23) hin (siehe Tabelle 2-1). Open Data, Registerdaten, Online-Befragungen, Text Scraping und Befragungsdaten weisen jeweils drei bis sechs Innovationen auf. Nochmals sei auf die Überschneidung der Kategorien hingewiesen. So könnten beispielsweise Online-Befragungen und Social-Media-Befragungen auch in die Kategorie Befragungsdaten aufgenommen werden. Da sich dadurch aber nichts Bedeutendes an der relativen Verteilung verschieben würde und die Dateninnovationen gerade im digitalen Bereich hervorzuheben sind, wird hier die genauere Unterscheidung bevorzugt.

**Tabelle 2-1: Dateninnovationen nach Datentyp**

Datentyp/-generierung	Anzahl
Administrative Daten	45
Big Data	23
Open Data	6
Registerdaten	4
Online-Befragung	4
Text Scraping	3
Befragungsdaten	3
Social-Media-Befragung	1
Geschäftsdaten von Unternehmen	1
<b>Gesamtergebnis</b>	<b>89</b>

Quelle: Eigene Darstellung.

<sup>5</sup> IZA: Das IZA – Institut zur Zukunft der Arbeit (<https://www.iza.org/das-iza>) ist eine unabhängige internationale, in Deutschland angesiedelte Forschungsorganisation und ein weltweit führendes Netzwerk für Arbeitsmarktökonomie mit über 1.800 Mitgliedern aus fast 60 Ländern. Durch eine einzigartige Kombination aus lokal und global vernetzter Forschung stellt es neueste wissenschaftliche Erkenntnisse für Forschungs- und Beratungstätigkeiten bereit, u.a. durch die IZA DP (<https://www.iza.org/publications/dp>) sowie weitere referierte Publikationsreihen. Pro Jahr werden ca. 900 IZA DP veröffentlicht und durch thematisch sortierte Newsletter verbreitet.

<sup>6</sup> NBER: Das National Bureau of Economic Research (<https://www.nber.org/about-nber>) ist eine private, gemeinnützige US-amerikanische Organisation, die Forschung zu verschiedenen Wirtschaftsthemen fördert. Durch die Veröffentlichung von mehr als 1.200 Arbeitspapieren ([https://www.nber.org/papers?page=1&perPage=50&sortBy=public\\_date](https://www.nber.org/papers?page=1&perPage=50&sortBy=public_date)) und die Ausrichtung von mehr als 120 wissenschaftlichen Konferenzen pro Jahr verbreitet die Organisation Forschungsergebnisse an akademische Forschende, öffentliche und private Entscheidungsträger\*innen und die breite Öffentlichkeit.

Hinsichtlich der Methoden und Art der Datenverarbeitung liegt der Schwerpunkt der recherchierten Dateninnovationen auf 40 Beispielen für Datenverknüpfungen (siehe Tabelle 2-2). Mit 17 Beispielen entfallen auf die Kategorie Datenportal die zweitmeisten, aber dennoch deutlich weniger Dateninnovationen. In den Bereichen Maschinelles Lernen und Künstliche Intelligenz (KI) konnten insgesamt zwanzig Dateninnovationen identifiziert werden. Zwar lässt sich Maschinelles Lernen unter KI subsumieren. Da Maschinelles Lernen aber von besonderem Interesse ist, wurde entsprechende Beispiele – sofern aus den Beschreibungen ersichtlich – separat erfasst. In allen weiteren Kategorien, also Datenlabore<sup>7</sup>, Datenvisualisierung, Simulation und Prognose, konnten jeweils ein bis fünf innovative Beispiele gefunden werden. Viele der Beispiele enthalten mehrere der genannten Datenverarbeitungen und Methoden und wurden dem Verfahren zugeordnet, das sie am intensivsten nutzen oder zu dem sie den größten innovativen Beitrag leisten.

**Tabelle 2-2: Dateninnovationen nach Art der Datenverarbeitung und -bereitstellung**

Art der Datenverarbeitung	Anzahl
Datenverknüpfung	40
Datenportal	18
Maschinelles Lernen	14
Künstliche Intelligenz (KI)	6
Datenlabor	5
Datenvisualisierung	4
Simulation	1
Prognose	1
<b>Gesamtergebnis</b>	<b>89</b>

Quelle: Eigene Darstellung.

Viele Dateninnovationen entstammen englischsprachigen Ländern, besonders den USA, die 16 Beispiele zur Gesamtauswahl beitragen, aber auch Kanada, Australien und dem Vereinigten Königreich (siehe Tabelle 2-3). Auch mehrere EU-Mitgliedstaaten sind mit vier oder mehr Beispielen vertreten. Auf die nicht einzeln ausgewiesenen weiteren Ursprungsländer entfallen 22 Beispiele, davon 12 Beispiele aus EU-Mitgliedstaaten sowie jeweils zwei Beispiele auf die Nicht-EU-Länder Norwegen und die Schweiz. Vergleichsweise selten sind außereuropäische nicht-englischsprachige Länder wie Südkorea (vier Beispiele) oder Brasilien, Mexiko und Uruguay (jeweils ein Beispiel) vertreten. Acht Studien betreffen mehr als ein Land.

<sup>7</sup> Darunter fallen Pilotanwendungen im Bereich der Datennutzung, die sich schwer den anderen Kategorien zuordnen lassen. Hierunter fallen zum Beispiel die Entwicklung digitaler Identitäten für den Zugang zu Online-Diensten, neue technische Verfahren der Datenspeicherung und -dokumentation sowie eine Datenbank, die aus Bioproben gewonnene Daten verknüpft.

**Tabelle 2-3: Dateninnovationen nach Ursprungsland**

Land	Anzahl
USA	16
Australien	7
Kanada	7
Österreich	6
Vereinigtes Königreich	6
Niederlande	5
Dänemark	4
Portugal	4
Südkorea	4
Mehrere Länder	8
Andere Länder	22
<b>Gesamtergebnis</b>	<b>89</b>

Quelle: Eigene Darstellung.

### 2.2.2 Auswahl von vertieft zu untersuchenden Dateninnovationen

In der Auswahlphase wurden die im Screening identifizierten Dateninnovationen anhand eines Kriterienkatalogs verglichen und mithilfe einer Gewichtungsmatrix in eine Rangfolge gebracht, um die für den deutschen Kontext interessantesten Anwendungsfälle im Rahmen von Steckbriefen detailliert zu untersuchen. Auf Basis dieser „Longlist“ wurden in einem Auswahl-Workshop gemeinsam mit dem BMAS im Dezember 2022 die finale Auswahl der Beispiele guter Praxis getroffen, die in der folgenden Phase in 22 Steckbriefen vertieft dargestellt wurden.

Im Zuge der Recherche wurden keine Dateninnovationen gefunden, deren primäres Ziel die Herstellung von Geschlechtergerechtigkeit ist. Dennoch haben einige der ausgewählten Beispiele eine Geschlechterdimension, beispielsweise wenn sie Leistungen an Familien betreffen, die wiederum das Gleichstellungsziel beeinflussen könnten (siehe Steckbrief 4.3), oder weil sie sich auf die Analyse von Geschlechtergerechtigkeit anwenden lassen (siehe Steckbrief 6.2). Im Kontext der Verwendung von Machine Learning ist das Risiko gegeben, dass Tendenzen zur Diskriminierung auch nach dem Geschlecht verstärkt werden (siehe Steckbrief 5.2). In solchen Fällen sind entsprechende Risikoanalysen und Gegenmaßnahmen erforderlich.

### 2.2.3 Dokumentation der ausgewählten Beispiele

In der dritten Projektphase ging es darum, die ausgewählten Beispiele guter Praxis in größerer Tiefe zu erschließen und die Bedingungen eines möglichen Transfers der Dateninnovation nach Deutschland herauszuarbeiten.

Die Bedingungen für einen Transfer der guten Praxis sind hierbei zentral, denn ohne Transfermöglichkeit würde der Analyse ein wesentlicher Handlungsbezug fehlen. Die Transferierbarkeit entscheidet sich insbesondere danach, welche spezifischen rechtlichen, institutionellen oder akteur\*innenbezogenen Voraussetzungen kritisch für den Erfolg der Dateninnovation sind und ob diese Voraussetzungen in Deutschland bestehen oder geschaffen werden können (Feenstra & Bunzeck, 2011). Die Identifikation der kritischen Voraussetzungen wurde anhand der Recherche und anhand von Bewertungen der mit der Dateninnovation vertrauten Expert\*innen vorgenommen. Das Vorliegen der kritischen Voraussetzungen in Deutschland wurde – soweit im Rahmen der Studie realisierbar – durch eine Analyse der gesetzlichen, sozialpartnerschaftlichen und anderen Rahmenbedingungen sowie durch deutsche Länderexpert\*innen geprüft. Auch wenn sich gute Praxis in einigen Fällen nicht vollständig übertragen lässt, kann es sein, dass Elemente der dargestellten Innovationen im deutschen Kontext umgesetzt oder auf andere Anwendungsgebiete übertragen werden könnten.

Zur Gewinnung der Informationen wurde zum einen eine vertiefte Dokumentenanalyse und Auswertung der vorhandenen Literatur vorgenommen. Zudem wurden insgesamt 37 Gespräche mit Expert\*innen im In- und Ausland durchgeführt.<sup>8</sup> Darunter wurden 16 Gespräche mit Expert\*innen aus den Forschungs- bzw. Projektteams im Ausland geführt, die direkt zu den jeweiligen Dateninnovationen befragt werden konnten. Weitere 21 Gespräche wurden mit überwiegend deutschen Expert\*innen geführt, um die Beispiele zu diskutieren und ihr Transferpotenzial zu beurteilen. Jeweils etwa die Hälfte dieser Gesprächspersonen waren in der Verwaltung oder in wissenschaftlichen Einrichtungen beschäftigt. Die Ergebnisse dieser Schritte resultierten in der Ausarbeitung der insgesamt 22 Steckbriefe, die in den folgenden Kapiteln dargestellt werden.

---

<sup>8</sup> Darunter waren mehrere Gespräche mit mehr als einem/einer Expert\*in.

### 3. Überblick über die Steckbriefe

Die ausgewählten Dateninnovationen werden im Folgenden drei Schwerpunkten zugeordnet: (1) Datengenerierung und Datenverknüpfung, (2) Methoden der Datenverarbeitung und (3) Datenbereitstellung (z. B. auf Portalen). Die Dateninnovationen kombinieren häufig Vorgehensweisen, die in mehrere dieser Bereiche fallen, und werden hier dem Bereich zugeordnet, durch den die Innovation am besten charakterisiert wird.

Tabelle 3-1, Tabelle 3-2 und Tabelle 3-3 in den folgenden Abschnitten führen die Dateninnovationen auf, die in den Kapiteln 4 bis 6 in Steckbriefform beschrieben werden. Nach der Kapitelnummer und dem Kurztitel der jeweiligen Innovation wird in der dritten Spalte kurz der Inhalt der Innovation, also je nach Tabelle die Art der Daten, die Methodik oder die Art der Bereitstellung, aufgeführt. Die vierte Spalte beschreibt die wesentliche Zielsetzung, die durch die Dateninnovation erreicht werden soll. Dies betrifft beispielsweise die effizientere Gestaltung von Verwaltungsvorgängen, die bessere Zugänglichkeit bestimmter Informationen oder erweiterte Möglichkeiten zur Gewinnung von Informationen im Rahmen der amtlichen Statistik und durch die wissenschaftliche Forschung.

#### 3.1 Datengenerierung und Datenverknüpfung

Die Mehrzahl der ausgewählten Innovationen mit dem Schwerpunkt Datengenerierung und Datenverknüpfung befassen sich mit der Verknüpfung von Daten (siehe Tabelle 3-1), insbesondere von administrativen Daten. Bei der Verknüpfung administrativer Daten geht es darum, die zu Verwaltungszwecken angelegten Daten der Sozialversicherung, der Finanzverwaltung, des Meldewesens und anderer öffentlicher Verwaltungen auf individueller Ebene zu verbinden und dadurch die ausschließliche Nutzung der Daten in voneinander getrennten „Datensilos“ zu beenden. Dies erweitert die Nutzungsmöglichkeiten administrativer Daten erheblich. Diese Daten sind generell auf Merkmale beschränkt, die für die jeweiligen Verwaltungsvorgänge unerlässlich sind, und weisen daher ein geringeres Merkmalspektrum auf. Die Verknüpfung kann das Problem des geringen Merkmalsumfangs mildern oder lösen.

Die Verknüpfung kann zum einen für die einzelnen Verwaltungen vorteilhaft sein, da sie auf Datenbestände anderer Stellen zurückgreifen können und dadurch Arbeitsaufwand sparen. Zum anderen können auch Vorteile für Bürger\*innen entstehen, da sie bestimmte Angaben nur bei einer Stelle machen müssen, was als Once-Only-Prinzip bezeichnet wird. Andere Länder, insbesondere Estland, aus dem das in Steckbrief 4.3 betrachtete Verfahren stammt, sind bei der Anwendung dieses Prinzips deutlich weiter fortgeschritten als Deutschland. Schließlich ergeben sich für die Forschung große Potenziale aus der Nutzung verknüpfter administrativer Daten, denn darin werden die relevanten Sachverhalte meist präziser erfasst als in Datensätzen, die durch wissenschaftliche Befragungen angelegt werden. Zudem müssen sie nicht kostspielig erhoben werden.

**Tabelle 3-1: Dateninnovationen mit Fokus Datengenerierung und -verknüpfung**

Steckbrief	Bezeichnung	Art der Daten oder Datennutzung	Ziel der Innovation
4.1	Hub-and-Spokes-Architektur und homomorphe Verschlüsselung zur Erstellung von Statistiken (Republik Korea)	Verknüpfung administrativer Daten	Zusammenführung von Daten zur Erstellung von Rentenstatistiken
4.2	Multi-Agency Data Integration Project (Australien)	Verknüpfung administrativer Daten	Bereitstellung einer Datenbankstruktur, die die Verknüpfung einer großen Anzahl von Datensätzen zum Zweck von statistischen Analysen ermöglicht
4.3	Proaktive Sozialleistungen durch die Verwirklichung des Once-Only-Prinzips (Estland)	Datenaustausch zwischen einzelnen Behörden und Arbeitgeber*innen	Vereinfachter Zugang zu Familienleistungen
4.4	Opportunity Insights: Eine Datenbasis zum sozialen Aufstieg (USA)	Verknüpfung verschiedenster Datenquellen auf regionaler Ebene	Verbesserte und zeitnahe Studien insbesondere zu Aufstiegschancen auf dem Arbeitsmarkt, Datenbereitstellung
4.5	Forschung zu Erwerbsminderungsrenten: Verknüpfung von Versicherungs- und Gesundheitsdaten (Niederlande)	Verknüpfung von Daten verschiedener Datenträger	Verbesserte Analyse der Effekte der Erwerbsunfähigkeitsversicherung
4.6	Verknüpfung von Stellenanzeigen und Bewerbungen aus einem Online-Stellenportal (Uruguay)	Stellenportal-Daten (Stellenausschreibungen und Lebensläufe)	Verbesserte Analyse von nachgefragten und angebotenen Fähigkeiten
4.7	Synthetische Datengeneratoren (Vereinigtes Königreich)	Umwandlung von Individualdaten in synthetische Daten durch Machine Learning	Sicherstellung des Datenschutzes, verbesserte Möglichkeiten der Datennutzung und -verknüpfung

Die Verknüpfung von Daten ist jedoch nicht auf administrative Daten beschränkt. Verbinden lassen sich auch Daten der amtlichen Statistik und des Zensus, Umfragedaten und Datenbestände privater Unternehmen und Dienstleister, insbesondere Big Data, die z. B. aus den sozialen Netzwerken gewonnen werden. Dabei gilt es die entsprechenden Datenschutzbestimmungen zu beachten, um insbesondere keine persönlichen Informationen preiszugeben und die Identifizierung einzelner Personen oder Betriebe in den Daten zu verhindern. Das Projekt „Opportunity Insights“ (Steckbrief 4.4) steht hier für eine Verknüpfung einer Vielzahl von Datensätzen aus unterschiedlichen Quellen auf der kleinräumigen regionalen Ebene.

Als zentrale Anforderungen müssen bei einer Datenverknüpfung zum einen die korrekte Zusammenführung von Informationen, zum anderen der Schutz vertraulicher Informationen gewährleistet sein. Die betrachteten Dateninnovationen weisen ein weites Spektrum unterschiedlicher Methoden der Verknüpfung auf. Die Steckbriefe 4.1, 4.2 und 4.3 zeigen hierbei unterschiedliche Verknüpfungsarchitekturen sowie unterschiedliche Lösungen für die Vereinheitlichung und Verschlüsselung von Daten im Vorfeld der eigentlichen Verknüpfung.

Neben der Verknüpfung von Daten geht es bei den Dateninnovationen auch um die Generierung und Aufbereitung von Daten, die eine Verarbeitung in der Verwaltung und Forschung erst möglich machen. Im Bereich der Arbeitsmarktpolitik kommt hierbei Daten von Online-Stellenportalen (Online Job Vacancies, OJV), wie sie unter anderem im Steckbrief 4.6 dargestellt werden, eine besondere Bedeutung zu. Hierbei werden Textdaten – aus Stellenanzeigen und Bewerbungen – in quantitativ verwertbare Daten umgewandelt. Verwendet wird dabei das Verfahren des Natural Language Processing (NLP), also die Erkennung und Verarbeitung von menschlicher Sprache u. a. durch Machine-Learning-Verfahren. Die Besonderheit des im Steckbrief 4.6 dargestellten Beispiels des Online-Stellenportals BuscoJobs besteht darin, dass die Informationen aus Stellenanzeigen mit den Informationen von Bewerber\*innen verbunden werden können. Die Daten aus Stellenanzeigen sind auch die Grundlage für weitere Steckbriefe, insbesondere die im Steckbrief 6.3 dargestellte Datennutzung aus dem Vereinigten Königreich.

Für die Übertragung nach Deutschland relevante Aspekte sind insbesondere die technische Verknüpfbarkeit und der Datenschutz. Hinsichtlich des Datenschutzes gilt in Deutschland, in den EU-Mitgliedstaaten sowie in den Nicht-EU-Staaten des Europäischen Wirtschaftsraumes (EWR) die Datenschutz-Grundverordnung (DSGVO). Die einheitliche europäische Rechtsgrundlage sollte eine Übertragung innerhalb ihres Geltungsbereichs erleichtern. Dabei ist allerdings zu berücksichtigen, dass die DSGVO Öffnungsklauseln enthält, durch die der nationale Gesetzgeber eigene Regelungen erlassen kann. Diese Regelungen sind einerseits in den Datenschutzgesetzen des Bundes und der Länder enthalten, andererseits in bundes- und landesrechtlichen Spezialvorschriften, beispielsweise zum Schutz von Sozialdaten in den Sozialgesetzbüchern im Sozialgesetzbuch X – Sozialverwaltungsverfahren und Sozialdatenschutz – (SGB X), (vgl. RatSWD, 2020, S. 15).

Bei der Datenverknüpfung sind die entsprechenden datenschutzrechtlichen Vorschriften zu beachten (vgl. z. B. Lettieri, 2021). Grundsätzlich ist eine Verarbeitung personenbezogener Daten nur zulässig, wenn die Voraussetzungen einer Erlaubnisnorm erfüllt sind, d. h. dass entweder eine wirksame Einwilligung der betroffenen Person oder ein gesetzlicher Erlaubnistatbestand vorliegt. Im letztgenannten Fall kann eine Verknüpfung unter anderem durch § 27 des Bundesdatenschutzgesetzes (BDSG) und § 75 Absatz 4 Satz 1 SGB X auch ohne eine Einwilligung der betroffenen Personen gedeckt sein. In der Vergangenheit wurden einzelne administrative Datensätze zum Zweck der Forschung entsprechend verknüpft, beispielsweise zum BASiD-Datensatz von DRV und IAB (Deutsche Rentenversicherung, 2022). Diese Abwägung treffen die Aufsichtsbehörden für jede einzelne Verknüpfung, was Unsicherheit und zusätzlichen Aufwand erzeugt. Hierzu hatte der Rat für Sozial- und Wirtschaftsdaten im Juni 2022 ein Positionspapier mit Eckpunkten für ein

Forschungsdatengesetz veröffentlicht, das auch die Zulässigkeit der Verknüpfung von Daten regelt (RatSWD, 2022, S. 4).

Für die Verknüpfung von administrativen Daten mit Befragungsdaten gibt es hierzulande ebenfalls einzelne Beispiele, etwa die IAB-SOEP-BAMF-Befragung von Geflüchteten (Bundesamt für Migration und Flüchtlinge, o. D.). Bei dieser Verknüpfungsart ist jedoch eine individuelle Einwilligung der befragten Personen erforderlich, wodurch ein Teil der durch die Befragungsdaten ohnehin kleineren Stichprobe aufgrund fehlender Einwilligung verloren geht.

Nach dem Koalitionsvertrag im Bund soll künftig ein Dateninstitut unter anderem die Datenverfügbarkeit und -standardisierung vorantreiben (Bundesministerium für Wirtschaft und Klimaschutz, Bundesministerium des Innern und für Heimat, 2023). Ob im Rahmen eines solchen Dateninstituts auch die rechtlichen Grundlagen für die Verknüpfung von Daten z. B. nach dem Modell der Person Linkage Spine und ihre Nutzung für Forschungswecke geschaffen wird, lässt sich derzeit nicht absehen. Zumindest sollen die Datenschutzbehörden in die Tätigkeiten eingebunden werden, um das datenschutzkonforme Datenteilen zu gewährleisten.

Um einen hohen Datenschutz auch bei verknüpften Datensätzen zu erreichen, werden in den betrachteten Beispielen Verfahren der Datenvergrößerung angewendet, die es faktisch verhindern, Personen in den Daten zu identifizieren. Ein besonders weitreichender Schutz personenbezogener Daten kann durch die Erzeugung synthetischer Daten gewährleistet werden, deren Beobachtungseinheiten keine realen Personen mehr sind. Dabei wird der statistische Informationsgehalt des Ursprungsdatensatzes möglichst weitgehend gewahrt (Steckbrief 4.7). Synthetische Daten können möglicherweise auch eher über Portallösungen durch die Öffentlichkeit genutzt werden, wie Steckbrief 6.1 zeigt.

Eine weitere, für Anwendungen in der Arbeitsmarkt- und Sozialpolitik derzeit noch wenig genutzte Datenquelle sind die von der Verwaltung bereitgestellten Open Government Data (siehe Steckbrief 6.6). Diese Daten dürfen von vornherein keinen Personenbezug enthalten. Datenschutzfragen stellen sich hier vor allem auf der Seite der Verwaltungen, die Open Data bereitstellen, und weniger auf der Seite der Nutzung.

## 3.2 Methoden

Bei den Dateninnovationen mit Schwerpunkt Methoden (siehe Tabelle 3-2) steht die Methodik des Machine Learning im Vordergrund. Obwohl Machine Learning in der Informatik allgegenwärtig ist und in anderen Bereichen routinemäßig angewendet wird, ist es erst vor relativ kurzer Zeit in die wirtschafts- und sozialwissenschaftliche Forschung vorgedrungen. Grundsätzlich zu unterscheiden sind Verfahren des Supervised Machine Learning (überwachtes maschinelles Lernen) und des Unsupervised Machine Learning.

**Tabelle 3-2: Dateninnovationen mit Fokus Methoden**

<b>Steckbrief</b>	<b>Bezeichnung</b>	<b>Wichtigste Methode</b>	<b>Ziel der Innovation</b>
5.1	Statistische Vorhersagemodelle zur Unterstützung der Beratung und Vermittlung von Arbeitslosen (Portugal)	Machine Learning	Vorhersage eines Risiko-Scores der Langzeitarbeitslosigkeit, um Arbeitsuchende gezielt zu betreuen
5.2	Machine-Learning-Methoden in Frühwarnsystemen für Leistungsbezug (Australien)	Machine Learning	Vorhersage der Bezugsdauer von Leistungen der Mindestsicherung
5.3	Vorhersage von Renteneintritten durch Machine Learning (International)	Machine Learning	Vorhersage von Renteneintritten zur verbesserten Planung im Rentensystem
5.4	Machine Learning in der Mindestlohnevaluation (USA)	Machine Learning	Nutzung von Machine Learning-Methoden zur Evaluation von Mindestlohneffekten
5.5	Intelligent Rehabilitation Recommendation System (Republik Korea)	Machine Learning	Vorhersage der optimalen Strategie zur Wiedereingliederung in den Arbeitsmarkt
5.6	Unterstützung von Betriebsprüfungen zur Aufdeckung von nicht gesetzeskonformem Verhalten (Belgien)	Machine Learning	Vorhersage der Wahrscheinlichkeit von nicht gesetzeskonformem Verhalten am Arbeitsmarkt, Unterstützung der Prüfung und Ermittlung
5.7	Automatisierte Kategorisierung von Unfällen am Arbeitsplatz (USA)	Machine Learning	Kodierung von Arbeitsunfällen zur Erstellung administrativer Statistiken
5.8	Sandbox: Ein Reallabor für die Nutzung von Machine-Learning-Methoden (Norwegen)	Künstliche Intelligenz	Datenschutzkonforme Gestaltung von KI-Projekten, dadurch verbesserte Nutzungsmöglichkeiten
5.9	Randomisierte Kontrollstudie zu Bewerbungsaktivitäten von Arbeitsuchenden (Dänemark)	Randomisierte Kontrollstudie	Evaluation, wie erhöhte Transparenz zur Anspruchsberechnung Transferleistungen beeinflusst

Supervised Machine Learning ist unter anderem dann eine nützliche Methodik, wenn es darum geht, Vorhersagen über Variablen zu treffen, vor allem wenn eine Vielzahl von Variablen und Modellspezifikationen getestet werden soll (zum Folgenden Athey & Imbens, 2019). Hierbei wird ein Modell mithilfe von Trainingsdaten so lange trainiert und mithilfe eines Testdatensatzes getestet, bis es auf dieser Datengrundlage die bestmögliche Vorhersage erzielt. Das erlernte Modell wird dann auf den eigentlichen Analysedatensatz angewendet. Die Unterteilung in einen Trainings- und einen Testdatensatz wird vorgenommen, um zu vermeiden, dass das Modell sich zu sehr an die Trainingsdaten anpasst (Overfitting oder Überlernen). Zur Berechnung der Vorhersage werden unterschiedliche Algorithmen verwendet, z. B. Decision Trees (Entscheidungsbaumverfahren), Random Forests, Gradient Boosting Decision Trees, und Elastic Net Regularisation.<sup>9</sup>

Die Anwendung dieser Verfahren ist vor allem geeignet, aus einer Vielzahl von möglichen Prädiktoren diejenigen auszuwählen, die die beste Vorhersage für den jeweiligen individuellen Fall ergeben. Eine umfangreiche, insbesondere eine durch Verknüpfung gewonnene Datenbasis ist die Voraussetzung dafür, dass Machine Learning seine Vorteile bei der Auswahl der Prädiktoren und ihrer Verwendung in unterschiedlichen Modellspezifikationen ausspielen kann.

Die in den beschriebenen Beispielen enthaltenen, auf Supervised Machine Learning gestützten Vorhersagen beziehen sich auf Langzeitarbeitslosigkeit (Steckbrief 5.1) bzw. den Leistungsbezug (Steckbrief 5.2), Eintritte in den vorzeitigen Rentenbezug (5.3), die Betroffenheit vom Mindestlohn (5.4), gesundheitliche Verbesserungen nach Rehabilitationen (Steckbrief 5.5) oder nicht gesetzeskonformes Verhalten am Arbeitsmarkt wie z. B. die Hinterziehung von Sozialversicherungsbeiträgen (5.6). Die verbesserten Vorhersagen sollen dabei helfen, die Zielgenauigkeit des jeweiligen Instrumenteneinsatzes zu verbessern. Hierfür stehen das portugiesische Beispiel aus dem Steckbrief 5.1, in dem das Vorhersagenmodell bzw. seine Ergebnisse von den Fachkräften der Arbeitsvermittlung als Tool bei der Einteilung der Arbeitslosen in Profilgruppen verwendet werden, oder die maschinell erzeugten Rehabilitationsempfehlungen aus dem südkoreanischen Beispiel in Steckbrief 5.5. In anderen Fällen findet die Erprobung solcher Ansätze im Bereich der Forschung statt (Steckbriefe 5.2 und 5.3). Der Steckbrief 5.5 zeigt das Einsatzpotenzial für die Arbeitsmarktforschung am Beispiel der Beschäftigungseffekte von Mindestlöhnen – ein Gegenstand, der für die Arbeitsmarktpolitik in Deutschland von großer Bedeutung ist und eine große Anzahl an Studien hervorgebracht hat. Die möglichen Anwendungsfelder dieser Innovation reichen jedoch weit über die Mindestlohnforschung hinaus und betreffen alle Bereiche der Arbeitsmarkt- und Sozialpolitik, in denen Evidenz über kausale Beziehungen benötigt wird und in denen mit einem quasi-experimentellen Forschungsdesign gearbeitet werden kann.

Verfahren des Unsupervised Machine Learning werden unter anderem eingesetzt, um Informationen zu bündeln. Ein Beispiel hierfür ist die Klassifikation von Unfallanzeigen in den USA (Steckbrief 5.7). Dabei werden die Texte der Unfallanzeigen mithilfe von Machine-Learning-Methoden in Kodierungen umgewandelt. Dafür sind zunächst Trainings- und Testdaten notwendig, auf deren Basis der Algorithmus lernt, einzelne Wörter und Formulierungen in Kodierungen umzuwandeln. Anschließend werden die Kodierungen klassifiziert, sodass sich eine Systematik ergibt. Grundsätzlich ist dieses Verfahren auf alle Prozesse anwendbar, bei denen zu Verwaltungszwecken Informationen

---

<sup>9</sup> Random Forests sind die logische Weiterentwicklung von Decision Trees und basieren auf einer Mehrzahl von Decision Trees, die auf wiederholten zufälligen Teilstichproben der Daten aufgebaut werden. Das Ziel ist es, die Varianz der einzelnen Bäume zu reduzieren, indem ein Durchschnitt über die verschiedenen Bäume gezogen wird. Die Gradient Boosting Decision Trees Methode teilt dieses Ziel und kann als sequenzielle Ansammlung von mehreren Entscheidungsbäumen angesehen werden, wobei jeder zusätzliche Entscheidungsbaum den Fehlerterm des vorherigen Baumes reduzieren soll. Elastic Net Regularization hingegen ist eine Machine-Learning-Methode zur Reduzierung der Überanpassung bei linearen Regressionsmodellen. Diese Methode „bestraft“ schlechte Schätzvariablen.

aus Texteingaben gewonnen werden sollen. Dieser Ansatz liegt auch der „Skills Taxonomy“ aus dem Vereinigten Königreich zugrunde (Steckbrief 6.3).

Die Nutzung von Machine-Learning-Methoden stellt zugleich neue Herausforderungen. Erstens kann die Gefahr von Diskriminierungen zunehmen. Diese werden beispielsweise dadurch hervorgerufen, dass durch Diskriminierung verursachte Arbeitsergebnisse in Stellenempfehlungen für Arbeitsuchende eingehen. Verbesserte Vorhersagen über den künftigen individuellen Bezug von Lohnersatzleistungen oder künftige individuelle Arbeitslosigkeit durch den Einsatz von Machine-Learning-Verfahren können dazu führen, dass dem sogenannten „Cream Skimming“ Vorschub geleistet wird. Das bedeutet, dass Vermittlungsfachkräfte Anreize haben, vor allem Personen mit niedrigen Risiken einer längeren Arbeitslosigkeit oder eines längeren Leistungsbezugs zu unterstützen, weil sie so schnellere Erfolge bei der Verkürzung von Leistungsbezug vorweisen können. Steckbrief 5.2 enthält Hinweise darauf, dass diese Gefahr real ist, und das Potenzial dafür besteht vermutlich auch in anderen Beispielen (z. B. Steckbrief 5.1). Deshalb ist es zum einen wichtig, Maßnahmen zu ergreifen, um Verzerrungen und Diskriminierungen zu verhindern. Die Selbstverpflichtenden Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung (BMAS, 2022, S. 51f.) nennen eine Reihe von Vorgehensweisen, wie diesem bekannten Problem des Machine Learning begegnet werden kann. Bezüglich der hier vorgestellten Dateninnovationen enthalten die Steckbriefe 5.1 und 5.8 Tools und Verfahren, um Diskriminierungen zu erkennen und zu vermeiden. Zum anderen sollten Fehlanreize wie „Cream Skimming“, die sich durch neue Verfahren verstärken, an der Ursache bekämpft werden, beispielsweise dadurch, dass die Leistung von Arbeitsvermittler\*innen nicht an der Anzahl der Vermittlungen gemessen wird, sondern Anreize für eine Berücksichtigung auch betreuungsintensiver Fälle gesetzt werden (siehe z. B. Courty et al., 2011).

Zum zweiten ist bei Innovationen im Bereich der Verwaltung die Akzeptanz der dort beschäftigten Fachkräfte, die Machine-Learning-Methoden verwenden, erforderlich. Nur wenn die Verfahren transparent sind und den Fachkräften ausreichende Handlungsspielräume bleiben, werden solche Tools produktiv in die beruflichen Tätigkeiten einbezogen. Akzeptanz muss aber auch bei der Zielgruppe bestehen. Der Einsatz von Machine-Learning-Verfahren zur besseren Unterstützung von Arbeitslosen oder Leistungsempfänger\*innen sollte möglichst auch im Einvernehmen mit der leistungsberechtigten Person erfolgen. Nach § 15 SGB II erfolgt etwa die Potenzialanalyse beispielsweise „zusammen mit der erwerbsfähigen leistungsberechtigten Person“. Willigt die Person ein, könnten die Ergebnisse der automatisierten Analyse Bestandteil der gemeinsam erarbeiteten Potenzialanalyse werden.

Fehlerhafte Ergebnisse von maschinellen Verfahren können am Ende dazu führen, dass ihre Potenziale nicht ausgeschöpft werden, wie Steckbrief 5.2 beschreibt.

Drittens ist die Prüfung von Anwendungen des Machine Learning unter Datenschutzaspekten wichtig. Steckbrief 5.8 stellt hierfür ein Verfahren vor. Hierbei können Unternehmen, die mit sensiblen Daten arbeiten und für ein Produkt oder eine Dienstleistung Machine-Learning-Methoden auf diese Daten anwenden möchten, gemeinsam mit der Datenschutzbehörde Prototypen für die Datennutzung in einem Versuchsdesign entwickeln.

Auch wenn Anwendungen des Machine Learning die methodischen Dateninnovationen bei Weitem dominieren, sind sie nicht die einzige Methodik, die neue Möglichkeiten der Steuerung im Bereich der öffentlichen Verwaltung eröffnet. Steckbrief 5.9 zeigt, dass experimentelle Forschungsdesigns genutzt werden können, um die Wirkung von Informationen und Anreizen zu untersuchen. Die Schlussfolgerungen hieraus können insbesondere für eine verbesserte Beratung und Vermittlung von Arbeitslosen genutzt werden und so in die Arbeitsverwaltung hineinwirken.

### 3.3 Datenbereitstellung

Ziel der unterschiedlichen Verfahren zur Bereitstellung der Daten ist es, die Anwendung der Daten zu verbreitern und dabei die Handelnden zu einer selbstständigen Nutzung der Daten zu ermächtigen (siehe Tabelle 3-3). Dies gilt zum Beispiel für die Anwendung komplexer Verfahren der Mikrosimulation, die bisher nur Expert\*innen möglich war; mit einer Portallösung können Simulationen von Politikänderungen auch von nicht fachlich ausgebildeten Personen vorgenommen werden (Steckbrief 6.1). Einige der hier vorgestellten Lösungen richten sich an Fachkräfte in der Leistungserbringung, beispielsweise Arbeitsvermittler\*innen (Steckbrief 6.2). Die Bereitstellung von Informationen ist aber nicht nur an die in der Planung und Umsetzung von Arbeitsmarkt- und Sozialpolitik tätigen Personen gerichtet, sondern besonders auch an die Endnutzer\*innen, z. B. Stellensuchende (Steckbriefe 6.3, 6.4 und 6.5). Es ist nicht auszuschließen, dass durch die Bereitstellung von Daten und Nutzungsmöglichkeiten auch missbräuchliche Verwendungen erleichtert werden. So könnten die Ergebnisse eines öffentlich verfügbaren Simulationsmodells durch die Wahl spezifische Annahmen bewusst zugunsten bestimmter Interessengruppen verzerrt werden. Solche Risiken müssen bei der Erprobung und Einführung einer Innovation im Einzelfall geprüft werden.

Je komplexer die Datenbasis, desto umfangreicher sind die Tools, die für die Bereitstellung der Daten genutzt werden. In den vorgestellten Dateninnovationen wird teilweise auf verknüpfte Daten zurückgegriffen, z. B. im Fall von Steckbrief 6.1. Vielfach sind die Methoden der Datenanalyse, die vor der Bereitstellung aufbereiteter Ergebnisse angewendet werden, selbst innovativ, beispielsweise indem sie auf Machine-Learning-Algorithmen zurückgreifen (Steckbriefe 6.3, 6.4 und 6.6).

Die Beispiele zeigen somit auch die Notwendigkeit eines ganzheitlichen Ansatzes. Bei der Genese von Daten müssen die Verfahren einer späteren Bereitstellung mitberücksichtigt werden. Umgekehrt erfordert die Bereitstellung von Daten möglicherweise neuartige Verknüpfungen und Aufbereitungen von Daten sowie spezifische Analysemethoden.

**Tabelle 3-3: Dateninnovationen mit Fokus Datenbereitstellung**

<b>Steckbrief</b>	<b>Bezeichnung</b>	<b>Art der Bereitstellung</b>	<b>Ziel der Innovation</b>
6.1	Social Policy Simulation Database and Model (Kanada)	Portallösung für ein Mikrosimulationsmodell	Anonymisierung und Verknüpfung von Daten, Nutzung von Simulationen durch Stakeholder
6.2	Jobnet und Joblog – Portal für Arbeitsuchende mit Verknüpfungsmöglichkeiten	Online-Portal für Verwaltung von Arbeitsuchenden	Effiziente Verwaltung von Unterstützungsleistungen und Beratung für Arbeitsuchende
6.3	The Data-Driven Skills Taxonomy (Vereinigtes Königreich)	Frontend mit Visualisierungstool für Arbeitsmarktinformationen, Backend mit Algorithmen zur Identifizierung von Qualifikationen in Stellenanzeigen	Erhöhte Transparenz der nachgefragten Qualifikation auf dem Arbeitsmarkt
6.4	JEDI – Jobs and Education Data Infrastructure (Australien)	Datenbankstruktur, Datenbereitstellung	Bereitstellung von Daten zu neuen Berufen und beruflichen Fähigkeiten
6.5	Job- und Informationsportal auf der Basis von Open Government Data (Spanien)	Einfaches Portal für die Bereitstellung von Open Government Data	Verbindung von Stellenportal und Leistungen für bestimmte Zielgruppen
6.6	Nowcast von Beschäftigung und Daten-Dashboard (Australien)	Webbasiertes Dashboard	Bereitstellung von aktuellen Beschäftigungsdaten auf regionaler und beruflicher Ebene

## 4. Steckbriefe zur Datengenerierung und Datenverknüpfung

### 4.1 Hub- and-Spokes-Architektur und homomorphe Verschlüsselung zur Erstellung von Statistiken (Republik Korea)

#### 4.1.1 Einführung

Bei der Erstellung von Statistiken im Bereich der Arbeitsmarkt- und Sozialpolitik tritt des Öfteren die Schwierigkeit auf, dass die Informationen, die in eine Statistik eingehen sollen, bei verschiedenen Trägern anfallen und dort gespeichert werden. Wenn diese Informationen dann nicht zur Weiterverarbeitung an einen zentralen Ort weitergegeben werden, können gegebenenfalls nur repräsentative Statistiken für Subgruppen, also z. B. pro Versicherungsträger, erstellt werden, nicht jedoch Statistiken, die für die Gesamtbevölkerung repräsentativ sind. Diese Situation lag in Südkorea bis vor Kurzem bei der Erstellung von Rentenstatistiken vor. Ursache hierfür ist, dass es eine Reihe verschiedener Rentenarten gibt, die durch acht öffentliche oder private Träger verwaltet werden. Aufgrund der sehr strengen Datenschutzvorschriften wurden die bei den Trägern anfallenden Daten in der Regel nicht weitergegeben. Daher bestanden in Südkorea bislang keine umfassenden Rentenstatistiken.

Um die Erstellung entsprechender Statistiken möglich zu machen, implementiert Statistics Korea, die südkoreanische Statistikbehörde, im Zeitraum von 2021 bis 2024 in Zusammenarbeit mit dem Ministerium für Wissenschaft und Informations- und Kommunikationstechnologien ein nationales Forschungs- und Entwicklungsprojekt.<sup>10</sup> Dieses ist mit einem Gesamtbudget von 20 Milliarden koreanischen Won (knapp 15 Mio. Euro) ausgestattet. Im Rahmen dieses Projekts wurde in der ersten Phase (2021-2022) ein statistisches Analysesystem entwickelt, in der zweiten Phase (2023-2024) sollen Technologien zum Datenschutz weiterentwickelt werden.

Der innovative Kern des Projekts ist die „Hub-and-spokes“-Architektur („Nabe und Speichen“) eines institutionellen Netzwerks, das die verschiedenen Datenträger miteinander verknüpft und in Austausch bringt. Der Datenschutz wird hierbei durch spezielle Technologien zur Verschlüsselung und Anonymisierung sichergestellt. Dabei spielt eine spezielle Verschlüsselungsmethode, die homomorphe Verschlüsselung, eine entscheidende Rolle, da sie die Ausführungen von Rechenoperationen auf verschlüsselten Daten, z. B. zur Erstellung von Statistiken, ermöglicht (Details siehe Abschnitt 4.1.2.2). Ein Nachteil der Methode ist der hohe Rechenaufwand, der impliziert, dass nur relativ einfache Berechnungen mit vertretbarem Aufwand durchzuführen sind. Außerdem ist Hardware mit höchster Leistungsfähigkeit erforderlich.

Diese Innovation kann hilfreich sein, wenn Statistiken für einen Bereich erstellt werden sollen, in dem die zugrundeliegenden Daten bei verschiedenen Trägern anfallen, die diese Daten nicht untereinander austauschen können. Beispiele in Deutschland könnten im Bereich der Rente, der Krankenversicherungen und bei anderen Sozialleistungen liegen.

---

<sup>10</sup> Die folgenden Ausführungen basieren auf Ryu (2022) sowie einem Expertengespräch.

## 4.1.2 Technische Beschreibung

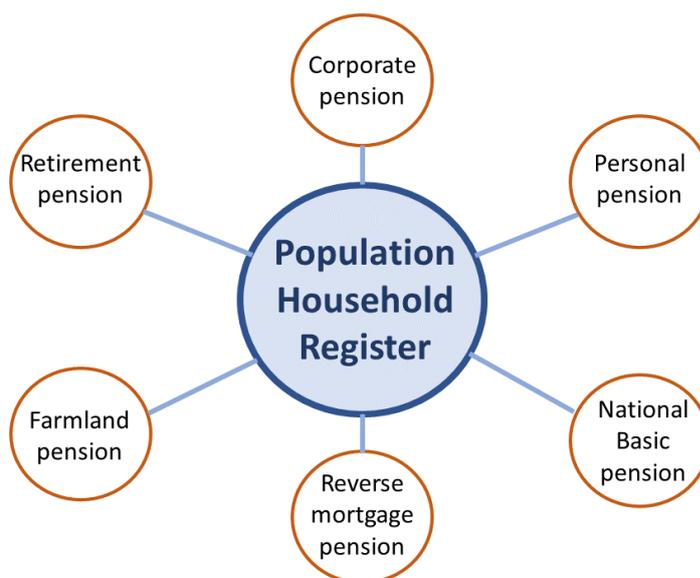
### 4.1.2.1 Datenquellen und Hub-and-Spokes-Architektur

Das südkoreanische Rentensystem ist stark fragmentiert, es existieren acht verschiedene Rentenarten. Die Daten zu Rentenansprüchen und -auszahlungen liegen bei den für die jeweiligen Renten zuständigen Institutionen vor:

- Korea Housing Finance Corporation (Behörde für Immobilienfinanzierung): Immobilienrente
- Ministry of Health and Welfare (Ministerium für Gesundheit und Wohlfahrt): Grundrente
- Nationale Pension Service (Nationale Rentenbehörde): nationale Rente
- Government Employees Pension Service (Rentenbehörde für Staatsbedienstete): betriebliche Altersvorsorge
- National Tax Service (Steuerbehörde) / Financial Services Commission: persönliche Rente
- Ministry of Employment and Labor (Ministerium für Beschäftigung und Arbeit): Ruhestandsrente
- Ministry of Agriculture, Food, Rural Affairs (Ministerium für Landwirtschaft, Nahrungsmittel und ländliche Angelegenheiten): Agrarlandrente.

Um umfassende Statistiken über die Leistungen dieser Trägerinstitutionen erstellen zu können, sollen deren Daten mit denen von Statistics Korea zusammengeführt werden. Dies ermöglicht dann beispielsweise die Auswertung nach demografischen Merkmalen, die im Haushaltregister von Statistics Korea vorliegen. Zum Datenaustausch dient eine Hub- and Spokes-Architektur. Statistics Korea fungiert als Nabe, hier laufen die Daten zusammen. Die Trägerinstitutionen sind die Speichen, die ihre Daten zur Nabe übermitteln. Die Architektur dieses Informationssystems wird in Abbildung 4-1 dargestellt.

**Abbildung 4-1: Hub-and-Spokes-Architektur zur Erstellung von Rentenstatistiken**



Quelle: Ryu (2022).

Um die Weitergabe von Daten unter Gewährleistung des Datenschutzes sicherzustellen, kommen verschiedene Technologien zum Einsatz, die im nächsten Abschnitt beschrieben werden.

#### 4.1.2.2 Homomorphe Verschlüsselung, Differential Privacy und synthetische Daten

Unter Homomorphismus (altgriechisch für „gleiche Gestalt“) versteht man mathematische Operationen, die bei ihrer Anwendung die ursprüngliche mathematische (bzw. statistische) Struktur erhalten. Homomorphe Verschlüsselung (oder Homomorphic Encryption, HE) von Daten ist dementsprechend eine Technologie zur Umwandlung von Daten in Chiffretext<sup>11</sup>, der analysiert und bearbeitet werden kann, sodass Auswertungen auf den umgewandelten Daten das gleiche Ergebnis liefern wie Auswertungen auf den nicht verschlüsselten Daten. Im Gegensatz zu herkömmlichen Verschlüsselungsverfahren ermöglicht die HE somit die Durchführung relativ komplexer mathematischer Operationen an verschlüsselten Daten, ohne die Verschlüsselung im Vorhinein aufzulösen (Yi et al., 2014).

Grundsätzlich gibt es fünf Phasen in der HE: Setup, Schlüsselerzeugung, Verschlüsselung, Datenanalyse, Entschlüsselung (Dai, 2019). Zur homomorphen Verschlüsselung ist die Einrichtung eines Schlüsselverwaltungssystems notwendig. Jede datenanbietende Stelle erzeugt einen privaten Sicherheitsschlüssel, der nur ihr selbst bekannt ist. Eine dritte Partei (im vorliegenden Beispiel Statistics Korea) koordiniert die Erzeugung eines öffentlichen Schlüssels, in den die privaten Schlüssel eingehen. Die privaten Schlüssel sind aber nicht aus dem öffentlichen Schlüssel extrahierbar und die dritte Partei hat keinen Zugriff auf die privaten Schlüssel.

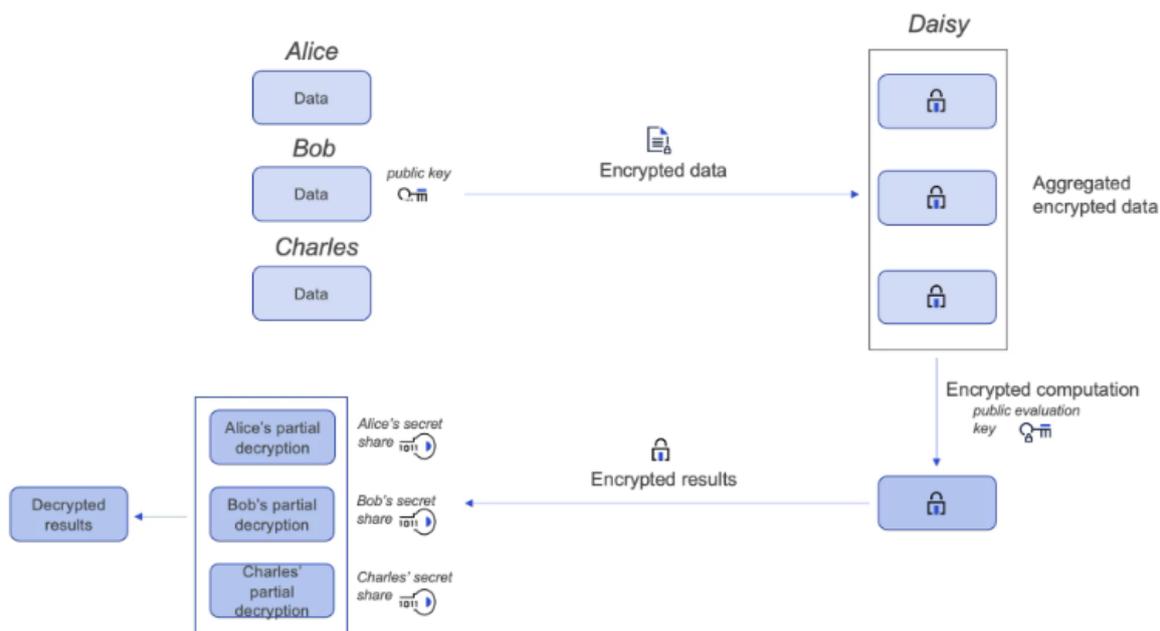
Im nächsten Schritt, der Verschlüsselung, wird der öffentliche Schlüssel von den jeweiligen Datenanbietenden verwendet, um ihre Datensätze zu verschlüsseln. Im Prinzip können die verschlüsselten Datensätze von allen interessierten Parteien zusammengeführt und analysiert werden, seien es die einzelnen Datenanbietenden, Statistics Korea, Forschende oder sogar private Einrichtungen. Im vorliegenden Beispiel ist Statistics Korea in der Lage, die verschiedenen Datensätze für die Berechnung amtlicher Statistiken zusammenzuführen und amtliche Statistiken zu erstellen.

Die verschlüsselten Datensätze werden an Statistics Korea übergeben, daraufhin führt Statistics Korea die verschiedenen Datensätze zusammen. Der Zugang zum gebündelten, verschlüsselten Datensatz zur Durchführung von Berechnungen ist mithilfe des zuvor generierten öffentlichen Schlüssels möglich. Die Analyse erfolgt auf dem verschlüsselten Datensatz, auch die Ergebnisse werden in verschlüsselter Form generiert.

Die Entschlüsselung der Analyseergebnisse, z. B. für die Rentenstatistiken, erfolgt durch die Nutzung aller privaten Sicherheitsschlüssel (secret shares) der verschiedenen Datenanbietenden. Weder die Nutzung eines einzelnen privaten Sicherheitsschlüssels noch des öffentlichen Sicherheitsschlüssels reicht für die Entschlüsselung der Ergebnisse aus. Vielmehr müssen alle Datenanbietenden in gegenseitigem Einverständnis ihre privaten Sicherheitsschlüssel benutzen, damit das Ergebnis entschlüsselt wird. Über die Erteilung des Einverständnisses hinaus müssen die Datenanbietenden jedoch nicht aktiv werden und z. B. einzelne Ergebnisse austauschen. Die zentrale Einsicht in die Ergebnisse mindert dabei nicht die Vertraulichkeit der einzelnen Datensätze (Ryu, 2022).

---

<sup>11</sup> Nach Anwendung eines Verschlüsselungsalgorithmus wird aus Klartext Chiffretext (oder auch Ciphertext/Geheimtext) erzeugt, welcher verschlüsselt ist und somit erst nach der Entschlüsselung mit einem Schlüssel in Klartext umgewandelt und gelesen werden kann. Der Begriff Text ist im weiteren Sinne als Oberbegriff für sämtliche Dateninhalte zu verstehen (Meinel & Sack, 2014).

**Abbildung 4-2: Homomorphe Verschlüsselung mehrerer Datensätze**

Quelle: Erabelli (2022).

Um mathematische Operationen an verschlüsselten Daten durchführen zu können, muss eine Beziehung zwischen Klartext und Chiffretext bestehen. HE kann entweder additiv oder multiplikativ sein, die Verschlüsselung erfolgt entweder etwas, teilweise oder vollständig homomorph. Der Unterschied zwischen diesen HE-Kategorisierungen liegt in der Art und Weise, mit der mathematische Operationen auf dem verschlüsselten Datensatz bzw. den verschlüsselten Datensätzen wie oft durchgeführt werden können (Abbildung 4-2). Im Allgemeinen gilt, dass die Einschränkungen der etwas und teilweise homomorphen Verschlüsselungen geringer sind als bei vollständigen HE, aber auch ihre Datensicherheit schwächer ist (Munjal & Bhatia, 2022). Daher wird von Statistics Korea die vollständige homomorphe Verschlüsselung eingesetzt. Hierbei wurde die CKKS-Ver- und Entschlüsselung gewählt, die bei reellen Zahlen die effizienteste Berechnung ermöglicht.

Ohne den richtigen Entschlüsselungsschlüssel kann nicht auf die Originaldaten zugegriffen werden. Das bedeutet, dass sensible Daten gesendet und analysiert werden können, während sie weiterhin verschlüsselt bleiben. Ein wichtiges Beispiel für die Verwendung von HE ist, dass ein Dateneigentümer Daten zur Verarbeitung in die Cloud senden möchte, aber einem Dienstanbieter seine Daten nicht anvertraut. Der Dateneigentümer ist bei einer HE der Einzige, der die Ergebnisse entschlüsseln kann, da nur er über den geheimen Schlüssel verfügt.

Ein Nachteil von homomorpher Verschlüsselung ist, dass der Rechenaufwand unter Verwendung verschlüsselter Daten deutlich höher ist als bei Berechnungen auf nichtverschlüsselten Daten. Daher sinkt die Geschwindigkeit, mit der Ergebnisse berechnet werden können, um den Faktor 1.000 bis 1.000.000. Folglich sind keine hochkomplexen Berechnungen möglich, zudem erfordern Berechnungen Hardware mit sehr hohen Leistungsparametern.

Eine zusätzlich zur homomorphen Verschlüsselung eingesetzte Methode zur Erhöhung der Anonymität der Analyse ist Differential Privacy. Hierzu wird Daten ein statistisches Rauschen, also ein Fehlerterm, hinzugefügt. Dieser verhindert, dass individuelle Informationen identifiziert werden können, soll aber gleichzeitig das Gesamtergebnis der Auswertung eines Datensatzes nicht verzerren. Grundsätzlich gilt bei der Verwendung der Differential-Privacy-Methode, dass zur Vermeidung einer

solchen Identifizierung umso mehr Rauschen hinzugefügt werden muss, je weniger Personen in einem Datensatz vorhanden sind. Im vorliegenden Fall wendet Statistics Korea Differential Privacy auf ausgewählte Statistiken an, um das Risiko der Deanonymisierung zu minimieren.

Da der Rechenaufwand bei verschlüsselten Daten deutlich höher ist als bei unverschlüsselten Daten, bietet es sich in der Praxis an, die Berechnungen auf den verschlüsselten Daten möglichst selten durchzuführen. Insbesondere ist es von Vorteil, die Entwicklung von Auswertungsprogrammen, z. B. zur Erzeugung von Statistiken, auf nichtverschlüsselten Daten vorzunehmen. Zu diesem Zweck werden synthetische Daten erzeugt, deren Struktur dieselbe ist wie bei den unverschlüsselten Daten, die sich aber nicht auf die ursprünglichen Beobachtungseinheiten beziehen oder extrem verfremdet wurden, sodass eine Identifikation von Originalbeobachtungen nicht möglich ist (siehe dazu näher Steckbrief 4.7). Die synthetischen Daten werden genutzt, um die Auswertungsprogramme zu erstellen und deren Lauffähigkeit zu testen. Ist die Lauffähigkeit sichergestellt und erzeugen die Programme die gewünschten Statistiken, können sie auf die verschlüsselten Daten angewendet werden.

#### 4.1.3 Anwendung und Beispiele

Die beschriebene Technik auf Basis der Hub-and-Spokes-Architektur, der homomorphen Verschlüsselung und der Differential Privacy kommt in Südkorea in mehreren Bereichen zum Einsatz. Hier ist zunächst das bereits beschriebene Beispiel zur Erzeugung von Rentenstatistiken zu nennen. In diesem Zusammenhang kann z. B. auf Grundlage der in Abschnitt 4.1.2.1 genannten Datenquellen das Einkommen und Vermögen älterer Personen berechnet werden, wobei Rentenzahlungen und Immobilienbesitz in die Berechnungen mit einfließen.<sup>12</sup>

Weitere laufende oder geplante Projekte in Südkorea auf Basis des Hub- and Spokes-Ansatzes jenseits der Rentenversicherung sind die Folgenden (Ryu, 2022):

- In einem laufenden Projekt werden umfassende Statistiken zur Verschuldung von Haushalten erzeugt. Hierzu werden Daten eines privaten Kreditbüros mit den Daten des Unternehmensregisters, des Wohneigentumsregisters und des Bevölkerungs- und Haushaltsregisters von Statistics Korea zusammengeführt.
- In einem lokalen Projekt in der Provinz Kyung-Gi wird der Frage nachgegangen, welche Faktoren den Erfolg von kleinen Unternehmen beeinflussen. Hierzu werden die Daten des Unternehmensregisters von Statistics Korea mit den auf Provinzebene vorliegenden zusätzlichen Informationen zu kleinen Unternehmen verknüpft.
- Ein angedachtes Projekt im medizinischen Bereich soll höchst sensible Daten zu Krankheiten und der medizinischen Geschichte von Familienmitgliedern miteinander verknüpfen. Hierdurch könnten wichtige Erkenntnisse zur Bedeutung genetischer Faktoren gewonnen werden.

#### 4.1.4 Übertragbarkeit auf Deutschland

Kern der vorgestellten Dateninnovation ist die Hub-and-Spokes-Architektur zwischen verschiedenen Institutionen im Zusammenspiel mit der homomorphen Verschlüsselung zur Berechnung von Statistiken. Diese Methodik ist sinnvoll einsetzbar, wenn Statistiken für einen Bereich erstellt werden sollen, in dem die notwendigen Daten bei verschiedenen Trägern anfallen und diese Daten (in der Regel aufgrund von Datenschutzbestimmungen, siehe Abschnitt 3.1) nicht einfach ausgetauscht werden können. Zusätzlich ist eine technische Zusammenführung auf individueller Ebene notwendig,

---

<sup>12</sup> Der Immobilienbesitz stammt aus den Daten der Korea Housing Finance Corporation (Behörde für Immobilienfinanzierung).

sie es durch eine personenspezifische Identifikationsnummer oder durch eine ausreichende Anzahl an personenspezifischen Merkmalen, die ein Record Linkage ermöglichen.

Eine Anwendung der Dateninnovation ist in Deutschland in verschiedenen Bereichen möglich. Der erste Bereich ist wie in Korea die Alterssicherung. Hier existiert auch in Deutschland ein relativ fragmentiertes System aus gesetzlicher Rentenversicherung (GRV), Beamtenversorgung (BV), betrieblichen Altersversorgung (BAV), Zusatzversorgung im öffentlichen Dienst (ZÖD), Alterssicherung der Landwirte (AdL) und berufsständischer Versorgung (BSV). Ferner sind für die Sicherung des Einkommens im Alter die Grundsicherung im Alter sowie private Renten (inklusive Riester-Rente) und Renten aus Lebensversicherungen zu berücksichtigen.

Eine Zusammenführung dieser verschiedenen Rentenarten auf der individuellen Ebene könnte dazu beitragen, einen umfassenderen Überblick über die Versorgungssituation und die Einkommensverhältnisse von Rentenbeziehenden zu erlangen. Die statistische Basis zur Analyse solcher Themen bilden in Deutschland momentan Befragungsdaten. Durch ein Zusammenspiel von Verwaltungsdaten könnte eine weitaus genauere Informationsbasis geschaffen werden. In den Berichten zur Alterssicherung in Deutschland (ASiD) wird zwar für einen Teil der genannten Renten angegeben, wie häufig mehr als eine Rente bezogen wird. Die vollständige Erfassung von Einkommen in Befragungen ist jedoch schwierig, oft unvollständig und aufgrund der Kosten auf eine relativ kleine Stichprobe begrenzt. Somit sind differenzierte Analysen zur Einkommenszusammensetzung bisher allenfalls eingeschränkt möglich. Damit könnte in Deutschland neben den Befragungsdaten eine weitere wichtige Datengrundlage zur Messung der Renteneinkommen und damit eine Informationsbasis zu sozialpolitischen Themen wie der Altersarmut genutzt werden.

Ein weiterer sozialpolitischer Bereich mit einer großen Vielfalt an Angeboten und Trägern sind Integrationsleistungen für Zugewanderte, insbesondere Geflüchtete. Leistungen wie z. B. Sprachkurse, Beratungen zur beruflichen Anerkennung, Kompetenzfeststellungen, Maßnahmen des SGB III und SGB II und einzelne Förderprogramme werden nicht in einer einheitlichen statistischen Datenbasis erfasst. Auch für operative Zwecke gibt es keine Datenbasis, aus der ersichtlich würde, welche Person welche Leistungen insgesamt bekommen hat; die bestehenden Informationssysteme der unterschiedlichen Träger erfassen nur den jeweils eigenen Bereich (siehe Boockmann et al. 2017, S. 134ff).

Die Übertragbarkeit erscheint hinsichtlich der Datenschutzbestimmungen grundsätzlich gegeben, da Südkorea ähnlich wie Deutschland ein sehr strenges Datenschutzgesetz hat (siehe hierzu auch Abschnitt 3.1). In Südkorea wurde 2011 ein umfassendes Datenschutzgesetz, der Personal Information Protection Act, verabschiedet, das unter anderem eine Einwilligungspflicht für die Verarbeitung personenbezogener Daten sowie das Recht auf Auskunft, Berichtigung und Löschung von Daten vorsieht. Auch die Erhebung, Nutzung und Bereitstellung personenbezogener Daten ist in Südkorea streng reglementiert (Korea Legislation Research Institute, 2020).

Abgesehen von juristischen Rahmenbedingungen sind auch die institutionellen Rahmenbedingungen zu beachten. So ist eine aktive Zusammenarbeit zwischen den an der Datenzusammenführung beteiligten Institutionen notwendig, insbesondere weil für die Entschlüsselung von Ergebnissen das Einverständnis aller beteiligten Partner erforderlich ist. Hier ergibt sich je nach Anzahl und Art der beteiligten Partner ein relativ hoher Abstimmungsbedarf, der den Willen zur Zusammenarbeit sowie ggf. die Bereitstellung von Personal und Ressourcen (insbesondere Hardware) erfordert. Die Übertragbarkeit dieser Dateninnovation auf Deutschland hängt daher auch entscheidend davon ab, welche Statistik erzeugt werden soll, welche Partner daran zu beteiligen wären, welches institutionelle Umfeld im Bereich dieser Statistik vorherrscht und welche rechtlichen Grundlagen unter Berücksichtigung des föderalen Systems bereits bestehen, die eine Übertragung ermöglichen.

#### 4.1.5 Fazit

Die in Südkorea entwickelte Hub- and Spokes-Architektur zum Informationsaustausch zwischen verschiedenen Datenhaltern nutzt die homomorphe Verschlüsselung von Daten. Mithilfe dieser Verschlüsselungsmethode können Berechnungen auf verschlüsselten Daten durchgeführt werden, was den in Südkorea geltenden hohen Datenschutzerfordernungen Rechnung trägt. Das vorgestellte Anwendungsbeispiel dient der Erstellung von Rentenstatistiken, die in Südkorea durch die stark fragmentierte Trägerlandschaft erschwert wird. Der Ansatz könnte auf Deutschland übertragbar sein. Eingeschränkt wird die Anwendbarkeit dieser Methode allerdings dadurch, dass die Durchführung von Berechnungen auf verschlüsselten Daten deutlich aufwändiger ist als auf nicht verschlüsselten Daten. Zudem erfordert die Anwendung der Hub-and-Spokes-Architektur eine aktive Zusammenarbeit der notwendigen Datenlieferanten, was institutionelle und organisatorische Herausforderungen mit sich bringen kann.

#### 4.1.6 Quellen zum Steckbrief

- Boockmann, B., Hirschfeld, H., Klee, G., Knirsch, C., Lücke, M., Puhe, H., Scheu, T., Späth, J. & Stöhr, T. (2017). *Wie kann eine gesamtwirtschaftlich erfolgreiche Integration der Flüchtlinge gelingen?* [Bericht]. Abschlussbericht des IAW an das Bundesministerium für Wirtschaft und Energie. Abgerufen am 30. März 2023 von [https://www.iaw.edu/team-research-staff/tobias-scheu.html?file=files/dokumente/Abschlussbericht\\_Final\\_2017\\_06\\_01.pdf](https://www.iaw.edu/team-research-staff/tobias-scheu.html?file=files/dokumente/Abschlussbericht_Final_2017_06_01.pdf)
- Dai, W. (2019). *Intro to Homomorphic Encryption* [Seminar session]. Private AI Bootcamp, Microsoft Research Lab, Redmond, Washington, United States. Abgerufen am 30. März 2023 von [https://github.com/WeiDaiWD/Private-AI-Bootcamp-Materials/blob/master/1\\_Wei\\_Intro\\_to\\_HE.pdf](https://github.com/WeiDaiWD/Private-AI-Bootcamp-Materials/blob/master/1_Wei_Intro_to_HE.pdf)
- Erabelli, S. (2022). *How to Use Homomorphic Encryption in the Real World*. Dualitytech. Veröffentlicht am 8. August 2022. Abgerufen am 30. März 2023 von <https://dualitytech.com/how-to-use-homomorphic-encryption-in-the-real-world/>
- Korea Legislation Research Institute (2020). *Personal Information Protection Act*. Veröffentlicht am 4 Februar 2020. Abgerufen am 30. März 2023 von [https://elaw.klri.re.kr/eng\\_service/lawTotalSearch.do](https://elaw.klri.re.kr/eng_service/lawTotalSearch.do)
- Meinel, C. & Sack, H. (2014). Grundbegriffe der Kryptografie. In: *Sicherheit und Vertrauen im Internet*. [https://doi.org/10.1007/978-3-658-04834-1\\_4](https://doi.org/10.1007/978-3-658-04834-1_4)
- Munjal, K. & Bhatia, R. (2022). *A systematic review of homomorphic encryption and its contributions in healthcare industry*. Complex & Intelligent Systems. <https://doi.org/10.1007/s40747-022-00756-z>
- Ryu, K. (2022). *Hub and Spokes Model for Governmental Statistics: Data Linkage and Data Security* [Presentation, 4. Februar 2022]. Private AI: Machine Learning on Encrypted Data, Online. Abgerufen am 30. März 2023 von <https://www.imsi.institute/videos/hub-and-spokes-model-for-governmental-statistics/>
- Yi, X., Paulet, R. & Bertino, E. (2014). Homomorphic Encryption. In: *Homomorphic Encryption and Applications*. [https://doi.org/10.1007/978-3-319-12229-8\\_2](https://doi.org/10.1007/978-3-319-12229-8_2)

## 4.2 Multi-Agency Data Integration Project (MADIP) (Australien)

### 4.2.1 Einführung

Von staatlichen Stellen erhobene administrative Daten bilden eine wichtige Grundlage für das Verwaltungshandeln und die Forschung und können somit informiertes Handeln von Bürger\*innen und staatlichen Akteur\*innen erleichtern. Administrative Daten umfassen z. B. Steuerdaten, Zensusdaten und Krankenversicherungsdaten. Sie erfassen ihre Gegenstände im Vergleich zu Befragungsdaten sehr präzise, insbesondere dann, wenn an die Angaben Rechtsfolgen geknüpft sind und Falschangaben zu rechtlichen Sanktionen führen können. Ein entscheidender Schwachpunkt ist jedoch, dass das Merkmalspektrum von einzelnen administrativen Datensätzen eher klein ist. Daher können Verknüpfungen den Nutzen der Daten oft erheblich steigern. Solche Verknüpfungen sind jedoch in der Regel aufwändig, sodass sie unterbleiben und damit das volle Potenzial administrativer Daten nicht ausgeschöpft wird.

Das Australian Bureau of Statistics (ABS) hat darum 2015 ein umfangreiches Projekt zur Verknüpfung verschiedenster administrativer Datensätze begonnen. Zusätzlich werden auch einzelne auf Befragungen beruhende Datensätze verknüpft. Das Multi-Agency Data Integration Project (MADIP) basiert auf der Methode der Person Linkage Spine (Australian Bureau of Statistics [ABS], o. D.-b). Die Grundidee dieses Datenprojekts ist es, eine Spine (Wirbelsäule) zu erstellen, die mit vielen Datensätzen in Verbindung steht. Werden für ein Forschungsprojekt oder eine Datenauswertung zu Verwaltungszwecken (z. B. im Rahmen von Once-Only) Verknüpfungen zweier oder mehrerer dieser Datensätze benötigt, können diese mit minimalem Aufwand durchgeführt werden. Dies stellt eine Verbesserung zu Verknüpfungsmethoden dar, die Verknüpfungen für neue Projekte immer wieder einzeln erstellen. Die Spine-Methode ist somit eine effiziente Verknüpfungsmethode. Trotz der umfangreichen Datenbasis gewährleistet sie einen hohen Datenschutz, da im Spine-Datensatz selbst über die Verknüpfungsindikatoren hinaus keine Daten enthalten sind, die bei Analysen zum Einsatz kommen. Das Vorgehen ähnelt der Verknüpfung von Informationen über die sogenannte X-Road in Estland (vgl. Steckbrief 4.3). Dort werden allerdings keine vollständigen Datensätze (bzw. Stichproben und Variablen, die für das jeweilige Projekt benötigt werden) zur Verfügung gestellt, sondern nur Einzelinformationen zu Verwaltungszwecken transportiert.

Die Spine des ABS soll die Gesamtbevölkerung Australiens zwischen Januar 2006 und November 2021 erfassen und für diese die Verknüpfung vieler Datensätze ermöglichen. Die Spine wird aus drei Datensätzen gebildet. Jeder dieser Datensätze deckt für sich genommen bereits einen Großteil der australischen Bevölkerung ab. Durch die Kombination der drei umfangreichen Datensätze wird die Wahrscheinlichkeit, dass in Australien ansässige Personen im kombinierten Datensatz nicht enthalten sind, auf ein Minimum reduziert. Selbst wenn sie in einem der Datensätze fehlen, ist die Wahrscheinlichkeit sehr groß, dass sie in mindestens einem der anderen Datensätze vorkommen.

Mithilfe der Spine können die drei Basisdatensätze selbst, darüber hinaus aber auch viele weitere administrative Datensätze sowie einige Datensätze aus Befragungen, paarweise verknüpft werden. Für einige Projekte werden mehrere paarweise Verknüpfungen durchgeführt, was die Kombination einer großen Anzahl an Einzeldatensätzen ermöglicht. Diese Einzeldatensätze können über die gängigen Indikatoren aus den Sozialversicherungen und/oder individuelle Merkmale wie Name, Geschlecht, Adresse und Geburtsdatum leicht mit der Spine verbunden werden. Wenn Forschende oder öffentliche Stellen dann eine Verknüpfung zwischen zwei oder mehreren Datensätzen herstellen wollen, die nicht für die Erstellung der Spine verwendet wurden, müssen die Einzeldatensätze nicht neu kombiniert werden, sondern werden über die Spine verknüpft. MADIP umfasst bereits ein umfangreiches Angebot zur Verknüpfung von bereitstehenden Datensätzen, das weiter ausgebaut wird und zu Forschungszwecken oder von Behörden genutzt werden kann. Dabei sind die

Anwendungsbereiche aufgrund der Vielfalt der verknüpften Daten breit gefächert. Verschiedenste Forschungsprojekte zu den Themen Soziale Sicherung und Arbeitsmarkt nutzen die MADIP-Daten ebenso wie Projekte zu Gesundheits- und Umweltthemen.

Die Idee, verschiedene administrative Datensätze bzw. administrative mit Befragungsdaten zu verknüpfen, ist nicht neu und wird in Deutschland bereits an einigen Stellen umgesetzt. Die Besonderheit von MADIP ist jedoch zum einen die große Zahl der verknüpfbaren Datensätze und zum anderen die Art der Verknüpfung, also der Spine-Ansatz, der einen hohen Datenschutz und gleichzeitig schnell umsetzbare neue Kombinationen von Daten ermöglicht.

## 4.2.2 Technische Beschreibung

### 4.2.2.1 Datenquellen

Das Projekt wird durch eine Kooperation des ABS mit den folgenden Stellen möglich: Australian Taxation Office, Department of Education, Department of Health and Aged Care, Department of Social Services, Services Australia und Department of Home Affairs. Gemeinsam bestimmen die Behörden die Rahmenbedingungen der Nutzung von MADIP und treffen strategische Entscheidungen zur Weiterentwicklung des Datenprodukts.

Datengrundlage für die Erstellung der Spine sind die folgenden drei Kerndatensätze: das Medicare Consumer Directory (MCD) von Services Australia, die DOMINO CAD vom Department of Social Services und die Steuerdaten der Personal Income Tax (PIT) vom Australian Taxation Office (eine Beschreibung der Datensätze folgt in Abschnitt 4.2.2.2).

Insgesamt verknüpft MADIP derzeit mehr als 50 Datensätze von elf Bundesbehörden, vier der sechs Bundesstaaten<sup>13</sup> und einer nicht-öffentlichen Institution, der Australian and New Zealand Intensive Care Society. Die Datensätze können entweder fortlaufend oder einmalig mit der Spine verknüpft werden (ABS, o. D.-a). Fortlaufend verknüpfte Datensätze werden nach Möglichkeit regelmäßig aktualisiert und sind in vorbereiteten Datenmodulen zur Nutzung enthalten (siehe Abschnitt 4.2.3.1). Auf besondere Anfrage hin wurden aber auch schon viele Datensätze für die Dauer einzelner Forschungsprojekte mit der Spine verknüpft. Diese einmaligen Verknüpfungen werden nach Ablauf des entsprechenden Projektes wieder gelöst. Hierbei handelt es sich um eine große Anzahl an Datensätzen zu unterschiedlichen Themen, die sich über unterschiedliche Zeiträume erstrecken. Sowohl für einmalig wie auch fortlaufend verknüpfte Daten muss die Verknüpfung explizit gesetzlich erlaubt sein. Für jeden verwendeten Datensatz werden die jeweiligen relevanten Normen von ABS veröffentlicht (ABS, o. D.-a).

### 4.2.2.2 Verknüpfung – Person Linkage Spine

Die zentrale Innovation des MADIP ist die „Person Linkage Spine“. Basierend auf der Idee, dass ein zentraler Punkt, an den alle Datensätze angeknüpft werden können, die Verknüpfungen vereinfacht und damit effizienter gestaltet, bildet sie die Grundlage für alle Datenverknüpfungen. Außerdem trägt die Spine dazu bei, sicherzustellen, dass die die Datenschutzauflagen vollständig eingehalten werden und dass die bereitgestellten Daten pseudonymisiert sind (siehe 4.2.2.3). Die Spine ist darauf ausgelegt, möglichst alle im Untersuchungszeitraum in Australien wohnhaften Personen zu erfassen. Aktuell wird der Zeitraum Januar 2006 bis November 2021 abgedeckt (ABS, o. D.-b).

---

<sup>13</sup> Zwei der sechs Bundesstaaten steuern keine eigenen Datensätze zum Projekt bei, sind jedoch in den auf gesamtstaatlicher Ebene bereitgestellten Datensätzen enthalten.

Die Spine entsteht durch eine Verknüpfung von Identifikationsmerkmalen (z. B. Alter, Geschlecht, Adresse) in den drei Kerndatensätzen, d.h. es findet ein Record Linkage über diese persönlichen Merkmale statt. Diese werden in sogenannten Clustern zusammengefasst, wobei jedes Cluster einer Person entspricht. Dabei ist es möglich, dass ein Cluster Identifikationsmerkmale aus nur einem oder zwei der Kerndatensätze enthält. In der Spine sind also alle Personen enthalten, die in mindestens einem der drei Kerndatensätze enthalten sind. Die drei genutzten Kerndatensätze sind die folgenden:

1. Medicare Consumer Directory (MCD) von Services Australia, eine Datenbank der allgemeinen australischen Gesundheitsversorgung (ANAO, o. D.)
2. DOMINO Centrelink Administrative Data (DOMINO CAD, Data On Multiple INdividual Occurrences) vom Department of Social Services (DSS). Hierbei handelt es sich bereits um einen zusammengeführten Datensatz, der auf Daten der Sozialversicherung und weiteren Daten aus Programmen, die vom Department of Social Services durchgeführt werden, beruht (DSS, o. D.)
3. Steuerdaten der Personal Income Tax (PIT) vom Australian Taxation Office (Australian Treasury, o. D.).

Die drei Kerndatensätze enthalten jeweils Daten über den gesamten Beobachtungszeitraum, erfassen also auch z. B. Namens- oder Adressänderungen und enthalten insofern besonders valide Identifikationsmerkmale (Frazer, 2021).

Die Spine unterläuft einem ständigen Wartungsprozess durch ABS und der Beobachtungszeitraum wird regelmäßig ausgeweitet, indem neue Identifikationsmerkmale oder Veränderungen in bestehenden Elementen ergänzt werden.

#### 4.2.2.3 Datenschutz

Um mit den zum Teil sensiblen Daten sicher umzugehen, hat das ABS umfangreiche Datenschutzkonzepte, sowohl bei der internen Bearbeitung und Verknüpfung der Daten als auch bei der Herausgabe von Datensätzen, geschaffen (ABS, o. D.-c). Dadurch wird sichergestellt, dass alle gesetzlichen Regelungen zum Datenschutz eingehalten werden. Insbesondere wird in jedem Schritt der Datenerfassung und -bearbeitung der Datenschutz bewertet. Außerdem wird strikt getrennt zwischen sogenannten Linkage-Informationen, die die Identifikationsmerkmale enthalten, und Informationen, die relevant für inhaltliche Analysen sind. Diese Trennung erfolgt unter anderem durch eine Rollentrennung. So haben Mitarbeitende des ABS immer nur Zugang zu den für ihren Aufgabenbereich relevanten Informationen. Die Einhaltung der Datenschutzregelungen bei den verschiedenen Schritten der Datenverknüpfung und -bereitstellung wird zunächst durch ABS selbst geprüft bzw. gewährleistet. Zusätzlich vergibt ABS jedoch auch Aufträge an externe Dienstleistende, die die Einhaltung der Regelungen ebenfalls prüfen (ABS, o. D.-e).

#### 4.2.3 Anwendung

##### 4.2.3.1 Nutzungsvoraussetzungen

Um die verknüpften Daten für die Forschung nutzbar zu machen, wurde das „MADIP Modular Product“ entwickelt. Dabei handelt es sich um vorbereitete Datensätze, die sogenannten Module. Sie sind aktuell für die Jahre 2011 bis 2019 verfügbar. Die Module sind in verschiedene Themen wie Einkommensteuer, Sozialversicherung oder Zensus unterteilt, enthalten die für den jeweiligen Themenkomplex relevanten Variablen und können nach Bedarf kombiniert werden.

Die Nutzung dieser Datenmodule ist für akademische oder von der Regierung durchgeführte Forschung auf Antrag und nach gründlicher Prüfung durch das ABS möglich. Insbesondere wird im

Rahmen des Antragsverfahrens geprüft, ob ein vorgeschlagenes Forschungsprojekt dem öffentlichen Interesse dient und ob alle Datenschutzvoraussetzungen erfüllt werden können. Die Erfüllung der Datenschutzvoraussetzungen wird vorab auch bereits bei der Erstellung der Einzeldatensätze sichergestellt. Dies wird bei den verknüpften Datensätzen wiederholt. Zudem werden zudem stets nur diejenigen Merkmale bereitgestellt, die für die Durchführung des jeweiligen Projekts benötigt werden.

Verwaltet wird der Datenzugriff über das DataLab Portal des ABS. Über diesen Online-Service werden neben den „MADIP Modular Product“-Daten auch andere Mikrodatensätze zur Verfügung gestellt. Der Zugriff erfolgt nach Antragstellung und Bewilligung über einen virtuellen Zugang, auf dem die Daten eingesehen und analysiert werden können. An die Forschenden, die auf die Daten zugreifen, wird eine Reihe von Anforderungen gestellt. So müssen sie über den Datenschutz belehrt und darauf verpflichtet worden sein, und in der Regel werden nur Forschende von australischen Institutionen zugelassen. Der örtliche Zugang ist außerdem nur in Australien erlaubt. Für den Zugang können jährliche Gebühren anfallen (ABS, o. D.-d).

#### 4.2.3.2 Anwendungsbeispiele

Durch die vielfältigen Angebote an Datenmodulen und Kombinationsmöglichkeiten hat MADIP zu einer großen Zahl an Forschungsprojekten beigetragen. Über 150 Projekte benutzen die verschiedenen angebotenen Daten für Projekte zu diversen Themen, insbesondere in den Bereichen Arbeitsmarkt, Bildung und Gesundheit. Ein Großteil der Projekte wird direkt von Regierungsbehörden durchgeführt. So macht sich zum Beispiel die Australian Treasury die Verknüpfung von Daten zu Visumsanträgen zunutze, um ein Modell über die finanziellen Auswirkungen von Migration zu erstellen (Varela et al., 2021). Das Department of Education veröffentlichte eine Studie, die auf der Basis von MADIP den Einfluss von Bildung unter anderem auf Einkommen und Arbeitslosigkeit untersucht (Department of Education, 2020).

Auch an Hochschulen und Universitäten werden die Daten regelmäßig verwendet, beispielsweise um Forschungsfragen über Personengruppen zu beantworten, die in nicht-administrativen Datenquellen wie Befragungsdaten typischerweise fehlen oder unterrepräsentiert sind. Dazu gehören u. a. Personen am oberen Ende der Einkommensverteilung. Forscher an der Australian National University nutzten die MADIP-Daten in mehreren Studien, um die Einkommensverteilung in Australien einschließlich dieser Personengruppe zu beschreiben (Biddle et al. 2019a, 2019b). Unter Verwendung von verknüpften Steuer- und Zensusdaten, die neben Personen- auch Haushaltsidentifikationsnummern enthalten, konnten die Autoren die Einkommensverteilungen sowohl mittels Personen- als auch Haushaltseinkommen berechnen. Durch die Kombination von Einkommensdaten aus mehreren Quellen weisen die MADIP-Daten zudem weniger Lücken auf als einzelne Datensätze zur Einkommensverteilung (Biddle et al., 2021).

#### 4.2.4 Übertragbarkeit auf Deutschland

Daten einiger Sozialversicherungsträger werden in Deutschland bereits vielfältig für die Forschung genutzt, beispielsweise zur Erstellung der Integrierten Erwerbsbiografien des IAB, die als Stichproben zur breiten wissenschaftlichen Nutzung zur Verfügung gestellt werden (IAB-FDZ, o. D.). Auch Steuerdaten können bereits genutzt werden, beispielsweise über das Taxpayer-Panel, das auf Grundlage der Steuererklärungen der veranlagten Steuerpflichtigen erstellt wird (Statistisches Bundesamt, o. D.-a). Da diese Datensätze aber selten das gesamte Merkmalspektrum umfassen, das für wissenschaftliche Fragestellungen relevant ist, könnte die Verknüpfung solcher Datenbestände über die Erstellung einer Spine mit einem hohen Mehrwert verbunden sein.

Der grundlegende Gedanke der Person Linkage Spine ist es, eine Struktur zu schaffen, die die Verknüpfung verschiedener Datensätze auf individueller Ebene ermöglicht. Dies ist in Deutschland grundsätzlich denkbar, allerdings sind dabei die Datenverfügbarkeit und die Datenschutzvorgaben zu berücksichtigen. Hinsichtlich der Datenverfügbarkeit gibt es nur wenige Datensätze, die nutzbar und umfassend genug sind, um als Grundlage für eine Spine zu dienen. In Deutschland stellt die Sozialversicherungsnummer (SV-Nummer) einen Identifikator dar, der in weiten Teilen der staatlichen Sozialversicherung verwendet wird. Im Bereich der Krankenversicherung werden jedoch unterschiedliche Versicherungsnummern vergeben. Daten der Finanzverwaltung können nicht über die SV-Nummer verknüpft werden, weil im Bereich der Besteuerung die Steuernummer bzw. die steuerliche Identifikationsnummer (IDNr) verwendet werden. Da diese seit 2008 für jede in Deutschland geborene Person bzw. für Zugewanderte anschließend an die Meldung bei der Meldebehörde erstellt wird, umfasst die IDNr mehr Personen als z. B. die SV-Nummer, die für Beamte, Selbstständige bzw. Personen ohne sozialversicherungspflichtiges Beschäftigungsverhältnis nicht vorliegt.

Ein weiteres Hindernis besteht darin, dass die Datenhaltung in Deutschland aufgrund des föderalen Staataufbaus und aufgrund der unterschiedlichen fachlichen Zuständigkeiten vielfach dezentral organisiert ist. Zugriff haben nur diejenigen Behörden, bei denen die Daten gehalten werden, und die Register unterschiedlicher Verwaltungsbereiche sind nicht miteinander vernetzt. Die Folge ist, dass die Bürger\*innen dieselben Angaben bei unterschiedlichen Stellen machen müssen und die Daten daher verschieden strukturiert und unterschiedlich aktuell sind. Das Registermodernisierungsgesetz (RegMoG) soll gewährleisten, dass alle Behörden die Daten, die sie für ihre Aufgabenerfüllung benötigen, erhalten. Im Rahmen der Registermodernisierung werden im Registerzensus bis zum Jahr 2031 vorhandene Register erweitert und neue Register aufgebaut. Hierdurch soll es möglich werden, bisher nicht verknüpfbare Daten zu verbinden. So werden die Daten der Melderegister mit sogenannten administrativen „Lebenszeichen“ aus weiteren Verwaltungsdaten abgeglichen. Ferner wird eine Verbindung mit Bildungs- und Arbeitsmarktdaten angestrebt (Statistisches Bundesamt, o. D.-b).

Neben den Problemen einer technischen Verknüpfung müsste auch der Datenschutz bei der Realisierung einer Person Linkage Spine in Deutschland umfassend gewahrt bleiben, wie in Kapitel 3 beschrieben wurde. Während die Möglichkeit zur breiteren Verknüpfung von (administrativen) Datensätzen auch in Deutschland auf verschiedenen Ebenen vorbereitet wird, erscheint eine sehr umfassende Verknüpfung, wie sie in Australien im Rahmen von MADIP vorliegt, aufgrund des Datenschutzes und erforderlicher Prüfverfahren als eine große Herausforderung. Als Vision für eine mögliche Dateninfrastruktur, die die Nutzung von verknüpften Datensätzen in Deutschland effizient gestaltet, kann MADIP allerdings hilfreich sein, sofern eine datenschutzkonforme Ausgestaltung erfolgen kann.

#### 4.2.5 Fazit

Das Multi-Agency Data Integration Project (MADIP) aus Australien stellt eine innovative Struktur der Datenverknüpfung dar, die zum Ziel hat, die Grundlage für die Verknüpfung möglichst vieler Datensätze zu schaffen und somit die Bevölkerung in Australien so weit wie möglich zu erfassen. Das Projekt basiert auf dem Konzept der Person Linkage Spine, d. h. es wird eine Spine (Wirbelsäule) erstellt, in der Personenidentifikatoren enthalten sind. Diese lassen sich mit verschiedenen Datensätzen verknüpfen, was zu einer hohen Effizienz bei der Durchführung der Datenverknüpfung führt. Mittlerweile können mehr als 50 Datensätze mit der Spine verknüpft werden, über 150 Forschungsprojekte nutzen bzw. nutzten bereits diese Dateninfrastruktur. Somit bietet die Spine enorme Forschungsmöglichkeiten in den verschiedensten thematischen Bereichen. Bei einer Übertragung auf Deutschland würden sich mehrere Herausforderungen stellen. Aufgrund der

unterschiedlichen Zuständigkeiten insbesondere auch im föderalen Kontext und der datenschutzrechtlichen Voraussetzungen könnte sich für einen Transfer allenfalls auf längere Sicht eine erfolgversprechende Perspektive ergeben.

#### 4.2.6 Quellen zum Steckbrief

Australian Bureau of Statistics (ABS). (o. D.-a). *MADIP data and legislation*. Abgerufen am 3. April 2023 von <https://www.abs.gov.au/about/data-services/data-integration/integrated-data/multi-agency-data-integration-project-madip/madip-data-and-legislation>

Australian Bureau of Statistics (ABS). (o. D.-b). *Person linkage spine*. Abgerufen am 6. April 2023 von <https://www.abs.gov.au/about/data-services/data-integration/person-linkage-spine>

Australian Bureau of Statistics (ABS). (o. D.-c). *Keeping integrated data safe*. Abgerufen am 6. April 2023 von <https://www.abs.gov.au/about/data-services/data-integration/keeping-integrated-data-safe>

Australian Bureau of Statistics (ABS). (o. D.-d). DataLab. Abgerufen am 6. April 2023 von <https://www.abs.gov.au/statistics/microdata-tablebuilder/datalab>

Australian Bureau of Statistics (ABS). (o. D.-e). MADIP Privacy Impact Assessment. Abgerufen am 10. August 2023 von <https://www.abs.gov.au/about/data-services/data-integration/integrated-data/multi-agency-data-integration-project-madip/madip-privacy-impact-assessment>

Australian National Audit Office (ANAO) (o. D.). Integrity of Medicare customer data. Abgerufen am 6.4.2023 von <https://www.anao.gov.au/work/performance-audit/integrity-medicare-customer-data>

Australian Treasury (o. D.). Tax White Paper. At a Glance. Abgerufen am 6.4.2023 von <https://treasury.gov.au/review/tax-white-paper/at-a-glance>

Bundesamt für Migration und Flüchtlinge (o. D.), IAB-BAMF-SOEP-Befragung von Geflüchteten. Abgerufen am 6. April 2023 von <https://www.bamf.de/SharedDocs/ProjekteReportagen/DE/Forschung/Integration/iab-bamf-soep-befragung-gefluechtete.html?nn=283560>

Biddle, N., Breunig, R., & Markham, F. (2019a). Describing the top of the income distribution in Australia. *CSRM Working Paper 4/2019*.

Biddle, N., Breunig, R., Markham, F., & Wokker, C. (2019b). Introducing the longitudinal MADIP and its role in understanding income dynamics in Australia. *CSRM & SRC Methods Paper 1/2019*.

Biddle, N., & Marasinghe, D. (2021). Using Census, Social Security and Tax Data from the Multi-Agency Data Integration Project (MADIP) to Impute the Complete Australian Income Distribution. *Tax and Transfer Policy Institute Working Paper 8*.

Department of Education. (2020). *Benefits of educational attainment*. Abgerufen am 23. März 2023 von <https://www.education.gov.au/integrated-data-research/benefits-educational-attainment>

Department of Social Services (DSS) (o. D.). DSS Metadata. Abgerufen am 6.4.2023 von: <https://www.dss.gov.au/dss-metadata>.

Frazer, B. (2021). IPDLN 2020 Conference: On-Demand Video on „Person Spine Linkage Methodology And Maintenance“ [Video]. YouTube. Abgerufen am 6. April 2023 von <https://www.youtube.com/watch?v=lKC2QNA7Iao>

IAB-FDZ, Forschungsdatenzentrum des IAB (o. D.). Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB). Abgerufen am 6.4.2023 von <https://fdz.iab.de/unsere-datenprodukte/personen-und-haushaltsdaten/siab/>

Statistisches Bundesamt (o. D.-a), Taxpayer-Panel. Abgerufen am 6.4.2023 von <https://www.forschungsdatenzentrum.de/de/steuern/tpp>

Statistisches Bundesamt. (o. D.-b). *Wie funktioniert der Registerzensus?*. Statistisches Bundesamt. Abgerufen am 13. April 2023 von <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Registerzensus/Methoden/teaser-erlaeuterungen.html>

Varela, P., Husek, N., Williams, T., Maher, R., & Kennedy, D. (2021). The lifetime fiscal impact of the Australian permanent migration program. *Treasury Paper*, Dezember 2021.

## 4.3 Proaktive Sozialleistungen durch die Verwirklichung des Once-Only-Prinzips (Estland)

### 4.3.1 Einführung

Seit fast drei Jahrzehnten verfolgt Estland unter der Bezeichnung e-Estonia eine umfassende Digitalstrategie. Ein wichtiger Bereich ist die Digitalisierung der Verwaltungsleistungen, die 1999 mit der digitalen Steuererklärung begann. Im Jahr 2018 beschloss die estnische Regierung einen Plan zur proaktiven Bereitstellung von weiteren Verwaltungsleistungen. Proaktiv bedeutet, dass eine Leistung nicht auf Antrag bereitgestellt wird, sondern automatisch nach dem Eintreten eines bestimmten Ereignisses. Voraussetzung dafür ist das Vorliegen von Daten, die für diesen Zweck genutzt werden dürfen. Proaktive Verwaltungsleistungen basieren auf dem Once-Only-Prinzip, wonach Daten und Informationen, die in einer Verwaltung über Bürger\*innen oder Unternehmen bereits vorliegen, nicht erneut von einer anderen Verwaltung erfragt werden sollen. Stattdessen soll ein Informationsfluss zwischen den Verwaltungsstellen stattfinden. Die erste, seit dem Jahr 2020 proaktiv angebotene digitale Verwaltungsleistung betrifft die Familienleistungen. Dabei werden nach der Geburt eines Kindes die der Familie zustehenden Leistungen, wie beispielsweise das Kindergeld, automatisch ausgezahlt.

In Deutschland sind für den Bezug von Familienleistungen ebenso wie bei anderen Sozialleistungen Anträge und oft auch Behördengänge notwendig. Beispielsweise muss das Neugeborene in den ersten drei Monaten nach der Geburt von den Eltern beim Standesamt und der Krankenkasse angemeldet werden. Um Kinder- und Elterngeld sowie Mutterschaftsleistungen zu beziehen, müssen die Eltern diese Leistungen beantragen. Bei unverheirateten Eltern sind weitere Schritte wie die Vaterschaftsanerkennung notwendig (Bundesministerium für Familie, Senioren, Frauen und Jugend, o. D.). Wenn diese Prozesse wie in Estland automatisiert würden, ergäbe sich für Eltern eine erhebliche Zeitersparnis. In Deutschland gibt es mit dem Onlinezugangsgesetz (OZG) eine rechtliche Norm, deren Ziel die Verbesserung des digitalen Zugangs zu Verwaltungsleistungen ist. In diesem Rahmen wurde in Bremen das Projekt „Einfache Leistungen für Eltern“ (ELFE) entwickelt, das die Beantragung von Familienleistungen vereinfacht und das Once-Only-Prinzip umsetzt. Aufgrund der Nähe zur estnischen Dateninnovation und dem fortgeschrittenen Entwicklungsstand wird für die Übertragbarkeit nach Deutschland besonders auf ELFE Bezug genommen. Im Bereich der Arbeitsmarkt- und Sozialpolitik gibt es ebenfalls Anwendungen. So macht z. B. die Sozialplattform (MAGS NRW, o. D.) als digitale Anlaufstelle unterschiedliche Sozialleistungen zugänglich.

### 4.3.2 Technische Beschreibung

#### 4.3.2.1 Infrastruktur

Im Rahmen der estnischen e-Government-Strategie wurde die sogenannte X-Road-Infrastruktur<sup>14</sup> aufgebaut (e-Estonia, o. D. -a), die für die Umsetzung des Once-Only-Prinzips von zentraler Bedeutung ist. In der X-Road wird zwischen den Registern unterschiedlicher staatlicher Stellen eine zentrale Plattform geschaffen. An die X-Road können dann die verschiedenen Verwaltungsleistungen über sogenannte Pull-Anfragen angeschlossen werden. Eine Pull-Anfrage bedeutet beispielsweise, dass der Träger des Kindergelds über die X-Road anfragt und anschließend Rückmeldung bekommt, ob eine Person ein Kind hat. Die Pull-Anfragen werden in der Regel monatlich gestellt. Die X-Road ist somit kein zentrales Register, sondern eine Austauschplattform, über die Informationen zwischen den Institutionen übermittelt werden. Das Single-Source-of-Truth-Prinzip, wonach die korrekte

<sup>14</sup> Auch als "National Once Only Technical System" bezeichnet.

Information an genau einer Stelle gespeichert wird, wird durch eine klar geregelte Zuordnung von Informationen zu den einzelnen Verwaltungen verwirklicht.

#### 4.3.2.2 Prozess

Um Familienleistungen wie das Kindergeld zu erhalten, müssen Eltern in Estland das Kind nach der Geburt in einem Onlineportal registrieren. Anschließend wird über die Informationen aus der X-Road die Höhe der auszahlenden Leistungen berechnet. Dabei wird der Erwerbsstatus sowie das gegebenenfalls erzielte Arbeitseinkommen in den 12 Monaten vor Beginn der Schwangerschaft herangezogen. Durch das Once-Only-Prinzip und die Verknüpfung der erforderlichen Daten, die der Verwaltung bereits vorliegen, ist dafür keine weitere Kommunikation zwischen Verwaltung und Eltern notwendig.<sup>15</sup> Die berechneten Leistungen werden den Eltern inklusive der Informationen zur Berechnungsgrundlage zur Bestätigung vorgelegt. Nach der Bestätigung werden die Familienleistungen aktiviert. Ferner werden den Eltern der Zeitpunkt der Zahlung und eventuelle weitere erforderliche Schritte mitgeteilt. Jedes weitere Ereignis, das die Familienzusammensetzung oder die Leistungsansprüche verändert, wird automatisch bei der Berechnung der Familienleistungen berücksichtigt und die Leistungen werden entsprechend angepasst.

Bei der erstmaligen Implementierung der proaktiven Familienleistungen wurden die Anfragen zunächst individuell und manuell begutachtet, um potenzielle Fehlerquellen durch Anpassungen zu minimieren und Vertrauen in das System aufzubauen – sowohl auf Seiten der Beschäftigten in der Verwaltung als auch der Bürger\*innen. Inzwischen werden etwa neun von zehn Aktivierungen von Familienleistungen nach einer Geburt automatisiert durchgeführt. Die weiterhin manuellen Tätigkeiten dienen der Kontrolle des laufenden Systems und komplexen Einzelfällen, die das System nicht automatisch beurteilen kann. Für die verbleibenden automatisierten Fälle bedeutet dies konkret, dass diese Fälle in den Arbeitsprozessen der Verwaltungsbeschäftigten keine Rolle mehr spielen. Tatsächlich haben die Beschäftigten in der Regel keinen Zugriff auf den Datenaustausch und die Leistungsbereitstellung. Als Nebeneffekt konnte dadurch dem Fachkräftemangel in der Verwaltung entgegengesteuert werden. Im Zuge der Implementierung der Innovation müssen zwischen der Verwaltung und den Bürger\*innen zuverlässige digitale Kommunikationskanäle aufgebaut werden.

Eine weitere Erfahrung der estnischen Innovation betrifft die Nichtinanspruchnahme staatlicher Unterstützungszahlungen. Das proaktive Element führt dazu, dass auch solche Bürger\*innen Leistungen erhalten, die sie andernfalls nicht in Anspruch genommen hätten, beispielsweise weil sie nicht wussten, dass ihnen die Leistung zusteht. Es zeigte sich, dass die Familienleistungen infolge der automatischen Gewährung an eine größere Zahl von Familien ausgezahlt wurden als zuvor. Insofern hat sich die Fairness des Leistungssystems erhöht.

Eine mögliche künftige Verbesserung der estnischen Anwendung betrifft die Kommunikation über die X-Road und die Art und Weise des automatisierten Informationsaustauschs. Zusätzlich zu den etablierten Pull-Anfragen wäre es für bestimmte Prozesse hilfreich, Push-Auskünfte zu ermöglichen. Beispielsweise ist für einzelne Leistungen eine Information zum Familienstand notwendig. Sobald eine Person verheiratet ist und dies eine Rolle für deren Leistungen spielt, wäre es effizient, wenn die einzelnen Personenstandsregister vom Register des Standesamtes per Push-Auskunft eine Meldung über die Änderungen bekämen. Pull-Anfragen führen bei solchen Änderungen innerhalb der X-Road zu Verzögerungen, da sie in der Regel nur einmal im Monat versendet werden.

---

<sup>15</sup> Die Kompatibilität des Verfahrens mit europäischem Recht, insbesondere den Grundsätzen für die Verarbeitung personenbezogener Daten nach Art. 5 der DSGVO sowie Art. 14 Abs. 3 Satz 1 der SDG-VO, ist nicht abschließend geklärt, vgl. Mikiver/Tupay (2023).

Mittlerweile werden weitere Verwaltungsleistungen proaktiv angeboten oder dieses Angebot ist in Planung, unter anderem die Beantragung von Altersrente (e-Estonia, o. D. -b). Estland profitiert von Rahmenbedingungen, die begünstigend auf die Entwicklung digitaler Verwaltungsleistungen wirken. Es gibt ein universelles elektronisches Identitätssystem und elektronische Signaturen. Die Zuständigkeiten der Informationsbereitstellung sind gesetzlich geregelt. Definitorische Datenstandards, beispielsweise zum Einkommensbegriff, sind etabliert. Zudem gibt es von Seiten der Gesellschaft ein hohes Vertrauen in die digitale Verwaltung, und die Digitalstrategie erfährt von der Politik nachhaltige Unterstützung.

#### 4.3.3 Ähnliche Anwendungen in anderen Ländern

In Österreich wird das Kindergeld (Familienbeihilfe) seit 2015 ebenfalls automatisiert ab der Geburt eines Kindes an die Eltern ausgezahlt (Österreichische Bundesregierung, o. D.). Die Eltern müssen dafür keine Anträge einreichen und werden von der Verwaltung nur kontaktiert, wenn einzelne Daten wie z. B. die Bankverbindung nicht vorliegen. Die erforderlichen Informationen für den automatischen Ablauf der Auszahlung der Familienbeihilfe in Österreich stammen aus dem zentralen Personenstandsregister. Weitere Informationen, beispielsweise zum Einkommen, sind in diesem Fall nicht erforderlich, da die Familienbeihilfe nur vom Alter des Kindes und der Kinderzahl in der Familie abhängt.

In Neuseeland ist die Beantragung des Kindergeldes direkt mit der Anmeldung des Kindes verknüpft. Bei der erstmaligen Anmeldung des Kindes werden die Eltern im dazugehörigen Onlineformular gefragt, ob sie Kindergeld beantragen möchten. Entscheiden sie sich für diese Option, wird das Kindergeld automatisch ausgezahlt, wenn die Eltern anspruchsberechtigt sind (Inland Revenue, o. D.).

In der EU wird das Once-Only-Prinzip durch die Single-Digital-Gateway-Verordnung (SDG-VO) von 2018 implementiert (Verordnung (EU) 2018/1724). Im Rahmen dieser Verordnung wird das Portal Your Europe aufgebaut, über das bis Ende 2023 verschiedene europäische Leistungen organisiert werden sollen. Beispielsweise sollen über das Portal die Registrierung eines Autos oder die Beantragung einer Rente über die Grenzen der EU-Mitgliedsstaaten hinweg möglich sein (Europäische Kommission, o. D.).

#### 4.3.4 Übertragbarkeit auf Deutschland

In Deutschland ist die Einführung möglichst oder nahezu vollständig automatisch gewährter und ausgezahlter Verwaltungsleistungen und nutzungsfreundlicher Prozesse ein gesetzliches Ziel. Das im Jahr 2017 in Kraft getretene OZG verpflichtet Bund und Länder, bis zum Ende des Jahres 2022 ihre Verwaltungsleistungen auch digital anzubieten<sup>16</sup> und die dafür notwendige Infrastruktur zu einem bundesweiten Portalverbund zu verknüpfen. Im Portalverbund ist für jede Person ein Nutzungskonto vorhanden. Über das Nutzungskonto können Bürger\*innen „barriere- und medienbruchfreien Zugang zu elektronischen Verwaltungsleistungen [der] Verwaltungsträger erhalten“ (§ 3, Abs. 1 OZG, Einfügung der Autor\*innen).

Um eine nutzerfreundliche und effiziente Umsetzung zu gewährleisten, wurden die OZG-Leistungen in 14 übergeordnete Themenfelder unterteilt (zum Beispiel "Familie & Kind" und "Unternehmensführung & -entwicklung") und darin wiederum in Lebens- und Geschäftslagen

---

<sup>16</sup> Der Entwurf zum Gesetz zur Änderung des Onlinezugangsgesetzes sowie weiterer Vorschriften zur Digitalisierung der Verwaltung (OZGÄndG) vom 24.05.2023 sieht die Streichung der OZG-Umsetzungsfrist und die Einführung eines begleitenden Monitorings der Regelungen des OZG (§ 12 OZG) vor.

gebündelt. Die Bündelung in Themenfelder legt die Grundlage für das arbeitsteilige Vorgehen von Bund und Ländern nach dem "Einer für Alle"-Prinzip (EfA). Im Digitalisierungsprogramm Föderal übernimmt ein Tandem aus Bundesressort und Land die Federführung für ein bestimmtes Themenfeld (z. B. BMAS und NRW für das Themenfeld „Arbeit & Ruhestand“, BMFSFJ und Bremen für das Themenfeld „Familie & Kind“) und in der Regel auch für die darin befindlichen Umsetzungsprojekte. Dort erarbeiten sie gemeinsam digitale EfA-Lösungen für die zugehörigen OZG-Leistungen mit dem Ziel, eine deutschlandweite flächendeckende Verfügbarkeit zu erreichen.

Der Stand der Entwicklung der einzelnen Leistungen im Rahmen der OZG-Umsetzung wird anhand eines Reifegradmodells erfasst. Dabei werden die Reifegrade null bis vier unterschieden. Während man bei Reifegrad null noch keine Informationen online findet, ist bei Reifegrad drei die komplette Abwicklung des Verwaltungsprozesses digital möglich. Ist Stufe vier erreicht, müssen von den Antragstellenden zudem keinerlei Nachweise mehr erbracht werden, die der Verwaltung bereits vorliegen („Once-Only-Prinzip“).

Im Rahmen der OZG-Umsetzung seit 2017 wurde in Bremen der Online-Dienst „Einfache Leistungen für Eltern“ (ELFE) entwickelt und implementiert. ELFE hat das Ziel, Familienleistungen über ein Online-Portal ohne Behördengänge oder Datenabfragen bei den Eltern bereitzustellen. Konkret geht es um die Bestimmung des Namens sowie der Beantragung von Eltern- und Kindergeld. ELFE besitzt als eine von sehr wenigen Diensten den Reifegrad vier (vollständig digital plus Datenabruf nach dem Once-Only-Prinzip), weshalb sie im Folgenden genauer beschrieben wird.

Hinter ELFE stehen zwei zentrale Systeme (Freie Hansestadt Bremen, o. D.). Zunächst steht über die Webseite ein Interface zur Verfügung, über das von den Eltern notwendige Angaben, beispielsweise zu Einkommensverlusten durch eine schwangerschaftsbedingte Erkrankung von der Krankenkasse oder der Nachweis einer Frühgeburt, gemacht werden müssen, die im Rahmen des Once-Only-Prinzips anderweitig nicht zur Verfügung stehen. Das zweite Element ist die ELFE -Middleware, welche die Daten verarbeitet und dabei zum Beispiel mit der Krankenkasse oder der Datenstelle der Rentenversicherung (DSRV) kommuniziert. Dadurch werden alle notwendigen Angaben gesammelt und standardisiert, damit sie den Fachverfahren in der Verwaltung zugespielt werden können.

Derzeit sind die ELFE-Leistungen nur für bestimmte Personengruppen verfügbar. Beispielsweise können Einkommensdaten für Selbständige noch nicht automatisiert abgefragt werden, da die dafür notwendige Schnittstelle zwischen Finanzverwaltung und ELFE noch nicht eingerichtet ist. Die betreffenden Personengruppen können demnach die Leistungen nicht über ELFE beantragen. Aktuell wird geprüft, wie die bisher nicht berücksichtigten Personengruppen in das System einbezogen werden können.

Die Erfahrungen des Projektteams in Bremen zeigen, dass verschiedene Faktoren für eine erfolgreiche Entwicklung einer digitalen Verwaltungsleistung wichtig sind (Freie Hansestadt Bremen, o. D.). Zum einen sind die Standardisierung von Rechtsbegriffen und digitalfähige Rechtsbegriffe von hoher Relevanz, damit im Rahmen des OZG verschiedene Leistungen nicht zu Widersprüchen führen und die technische Implementierung effizient umgesetzt werden kann. Beispielsweise müssen für viele Sozialleistungen die Einkommens- und Vermögensbegriffe eindeutig definiert sein, wenn die Leistungen automatisiert berechnet und aktiviert werden. Für diese Aspekte könnten in Analogie zu den Digitalisierungslaboren im Rahmen des OZG sogenannte Rechtslabore hilfreich sein, wo bereits früh und das Gesetzgebungsverfahren unterstützend standardisierte und digitalfähige Rechtsbegriffe entwickelt werden. Ein begünstigender Faktor sind die Skaleneffekte einzelner Elemente der Digitalleistungen. Im Rahmen der Einmalzahlung für Studierende in Deutschland im Frühjahr 2023 nach dem Studierenden-Energiepreispauschalengesetz wurde die dafür notwendige Applikation AusweisApp2 verstärkt heruntergeladen. Da die App auch für die Leistungen von ELFE notwendig ist,

sinkt die Hemmschwelle bzw. der Aufwand für die Nutzung von ELFE-Leistungen durch diesen kurzfristigen Schub zusätzlich. Ähnliches gilt für weitere digitale Leistungen in Deutschland.

Die Realisierung des Innovationspotenzials von ELFE kommt bisher hauptsächlich den Bürger\*innen zugute. Die Fachverfahren innerhalb der Verwaltung profitieren bisher beispielsweise durch die direkt digital vorliegenden Daten und den damit wegfallenden Prozessen, beispielsweise dem Abtippen der Anträge.

Im Prinzip ist ELFE, insbesondere durch die Verwirklichung des Once-Only-Prinzips, der Dateninnovation in Estland ähnlich. Unterschiede bestehen insbesondere aufgrund der Einbettung des estnischen Verfahrens in die umfassende digitale Infrastruktur der öffentlichen Verwaltung sowie aufgrund der unterschiedlichen Abdeckung der Bevölkerung.

In Hamburg gibt es seit 2018 die digitale Verwaltungsleistung „Kinderleicht zum Kindergeld“, die ähnlich funktioniert wie ELFE (Senatskanzlei Hamburg, o. D.). Das Programm wird unabhängig vom EfA-Prinzip entwickelt.

Im Bereich der Arbeitsmarkt- und Sozialpolitik könnte das Once-Only-Prinzip Potenziale für eine Entlastung der Bürger\*innen, Unternehmen und der Verwaltung eröffnen. Im Themenfeld Arbeit und Ruhestand werden auf der für das OZG eingerichteten Webseite vier Umsetzungsprojekte genannt (Bundesministerium des Innern und für Heimat, o. D. -b): die gebündelte Umsetzung von Sozialleistungen über die „Sozialplattform“ (MAGS NRW, o. D.), die Expressdigitalisierung von Leistungen nach § 56 des Infektionsschutzgesetzes (IfSG), die Zulassung für reglementierte Berufe sowie der elektronische Heilberufausweis. Über die „Sozialplattform“ soll künftig eine Vielzahl einzelner Leistungen angeboten werden, vom Bürgergeld über Leistungen aus dem Bildungs- und Teilhabepaket bis zur Schuldner- und Suchtberatung. Vielfach haben diese Leistungen – beispielsweise das Bürgergeld – bereits den Reifegrad drei erreicht.

#### 4.3.5 Fazit

Durch die proaktiven Familienleistungen in Estland erhalten Familien eines Kindes die ihnen zustehenden Leistungen nach der Anmeldung des Kindes automatisch. Dabei wird das Once-Only-Prinzip genutzt, wobei die für die Familienleistung notwendigen Daten zwischen den beteiligten Behörden und Sozialträgern über die sogenannte X-Road ausgetauscht werden und somit nicht bei den Familien abgefragt werden müssen. Durch die Weiterentwicklung der X-Road, insbesondere von Pull- zu Push-Anfragen, könnte das System nochmals an Effizienz gewinnen. Die Entwicklung der proaktiven Leistungen in Estland wurde durch die bestehende digitale Infrastruktur der öffentlichen Verwaltung begünstigt.

In Deutschland werden Verwaltungsleistungen im Rahmen der Umsetzung des OZG digital zugänglich gemacht. Das Land Bremen hat die Verwaltungsleistung ELFE entwickelt, um Familienleistungen in einem einfachen Verfahren und ohne unnötige Datenabfragen bereitzustellen. Auch hier wird das Once-Only-Prinzip angewendet. Eine bundesweite Implementierung von ELFE hätte ein großes Entlastungspotenzial für Eltern. Die Vereinfachung von Antragsverfahren unter Anwendung des Once-Only-Prinzips könnte auch auf weitere Verwaltungsleistungen übertragen werden, beispielsweise solche im Themenfeld Arbeit und Ruhestand. Voraussetzungen hierfür sind unter anderem digitalfähige und standardisierte Rechtsbegriffe und institutionenübergreifende Datenübertragungskanäle.

#### 4.3.6 Quellen des Steckbriefs

Bundesministerium des Innern und für Heimat. (o. D. -a). *Digitalisierungslabore*. Bundesministerium des Innern und für Heimat. Abgerufen am 11. April 2023 von <https://www.onlinezugangsgesetz.de/Webs/OZG/DE/grundlagen/agile-methoden/digitalisierungslabore/digitalisierungslabore-node.html>

Bundesministerium des Innern und für Heimat. (o. D. -b). *Themenfeld Arbeit & Ruhestand: Mit großen Schritten zu einem bürgerfreundlichen digitalen Zugang zu Sozialleistungen*. Bundesministerium des Innern und für Heimat. Abgerufen am 14. Juni 2023 von [https://www.onlinezugangsgesetz.de/Webs/OZG/DE/themen/digitalisierungsprogramm-foederal/themenfelder/arbeit\\_und\\_ruhestand/arbeit-und-ruhestand-node.html](https://www.onlinezugangsgesetz.de/Webs/OZG/DE/themen/digitalisierungsprogramm-foederal/themenfelder/arbeit_und_ruhestand/arbeit-und-ruhestand-node.html)

Bundesministerium für Familie, Senioren, Frauen und Jugend. (o. D.). *Checklisten rund um die Geburt*. Bundesministerium für Familie, Senioren, Frauen und Jugend. Abgerufen am 31. März 2023 von <https://familienportal.de/familienportal/lebenslagen/schwangerschaft-geburt/checklisten>

Europäische Kommission. (o. D.). *The single digital gateway and Your Europe*. Europäische Kommission. Abgerufen am 12. April 2023 von [https://single-market-economy.ec.europa.eu/single-market/single-digital-gateway\\_de](https://single-market-economy.ec.europa.eu/single-market/single-digital-gateway_de)

e-Estonia. (o. D. - a). *We have built a digital society and we can show you how*. e-Estonia. Abgerufen am 12. April 2023 von <https://e-estonia.com/>

e-Estonia (o. D. - b). Factsheet proactive government. Abgerufen am 12. April 2023 von <https://e-estonia.com/>

Freie Hansestadt Bremen. (o. D.). *ELFE – One Vision*. Abgerufen am 11. April 2023 von [https://netzwerk-rechtsetzung-buerokratieabbau.de/veranstaltung/Digitalisierung\\_von\\_Verwaltungsleistungen.html](https://netzwerk-rechtsetzung-buerokratieabbau.de/veranstaltung/Digitalisierung_von_Verwaltungsleistungen.html)

Inland Revenue. (o. D.). *Apply for Best Start*. New Zealand Government. Abgerufen am 12. April 2023 von <https://www.ird.govt.nz/working-for-families/applying/best-start>

Mikiver, M., & Tupay, P.K. (2023). Has the GDPR killed e-government? The “once-only” principle vs the principle of purpose limitation. *International Data Privacy Law*, 13 (3), 194-206.

Ministerium für Arbeit, Gesundheit und Soziales des Landes Nordrhein-Westfalen (MAGS NRW) (o. D.). *Sozialleistungen einfach erklärt und beantragt*. Abgerufen am 17. August 2023 von <https://sozialplattform.de/>

Österreichische Bundesregierung. (o. D.). *Familienbeihilfe – Beantragung*. Oesterreich.gv. Abgerufen am 12. April 2023 von [https://www.oesterreich.gv.at/themen/steuern\\_und\\_finanzen/sonstige\\_beihilfen\\_und\\_foerderungen/4/1/Seite.450233.html#Verfahrensablauf](https://www.oesterreich.gv.at/themen/steuern_und_finanzen/sonstige_beihilfen_und_foerderungen/4/1/Seite.450233.html#Verfahrensablauf)

Senatskanzlei Hamburg. (o. D.). *In Hamburg kommen Eltern „Kinderleicht zum Kindergeld“*. Hamburg. Abgerufen am 11. April 2023 von <https://www.hamburg.de/kinderleicht-zum-kindergeld/>

## 4.4 Opportunity Insights: Eine Datenbasis zum sozialen Aufstieg (USA)

### 4.4.1 Einführung

Bei vielen Themen treten Big Data mehr und mehr an die Stelle von konventionellen Datengrundlagen. Diese Daten entstehen häufig im privatwirtschaftlichen Bereich. Vielfach haben sie eine hohe Frequenz, manchmal sind sie sogar tagesgenau. Aufgrund ihres Umfangs, ihrer Detailgenauigkeit und ihrer Aktualität bieten Big Data ein hohes Analysepotenzial. Zudem besteht häufig die Möglichkeit, diese Art von Daten mit öffentlich zugänglichen Daten zu verknüpfen. Allerdings müssen die Daten häufig erst in eine für Analysen nutzbare Form gebracht werden, zudem müssen datenschutzrechtliche Vorgaben beachtet werden.

Ein Datenprojekt, das Big Data aus einer Vielzahl von Quellen bündelt und verknüpft, ist „Opportunity Insights“. Hierbei handelt es sich um eine stiftungsfinanzierte Einrichtung an der Harvard University in den USA, die 2018 von den renommierten Ökonomen Raj Chetty, Nathaniel Hendren und John Friedman als gemeinnützige Organisation gegründet wurde. Die Motivation zur Gründung dieses Projekts war die Feststellung, dass Forschung und Öffentlichkeit oft Daten zur Beantwortung drängender Fragen in Themenbereichen wie soziale Ungleichheit, Geschlechterungleichheit und Bildung nicht zur Verfügung stehen, solche Daten jedoch häufig bereits vorliegen und daher prinzipiell verfügbar gemacht werden können. Hier setzt das Projekt „Opportunity Insights“ an, indem öffentliche Daten gebündelt und Kooperationen mit privaten Unternehmen eingegangen werden (Opportunity Insights Team, 2023b). Die genutzten Datensätze werden nach vorgegebenen, öffentlich nachvollziehbaren Routinen automatisch aktualisiert und liegen damit nahezu in Echtzeit vor. Sofern es die Datenschutzaufgaben erfordern, werden die Daten zudem automatisch verfremdet, d. h. soweit verändert, dass keine Rückschlüsse auf Personen oder Unternehmen möglich sind.

Grundlage für die Nutzung der Daten ist die Zusammenführung von großen Datensätzen (Big Data). Hierbei erfolgt die Datenzusammenführung in der Regel auf regionaler Ebene, für bestimmte Forschungsfragen auch auf Ebene von Highschools und Colleges. Das Datenprojekt basiert generell auf langfristigen Kooperationen mit verschiedenen Firmen und Institutionen. Aufgrund der hohen Frequenz der Daten bedarf es einer regelmäßigen Aktualisierung der Datensätze. Das geschieht mithilfe eines eigens erstellten Algorithmus über eine sogenannte „Automatisierte Datenverarbeitungspipeline“ (Chetty et al., 2022a). Nach der Gewinnung der Daten werden verschiedene Datenanpassungen vorgenommen. Die Daten werden anschließend auf einer zentralen Plattform einer breiten Öffentlichkeit durch eine interaktive Visualisierung zugänglich gemacht. Der gesamte Prozess inklusive der Veröffentlichung läuft mithilfe eines Versionskontrollsystems automatisch ab und verlangt nur in speziellen Fällen manuelle Eingriffe.

Zusätzlich zur Bereitstellung von Daten führt das Team von Opportunity Insights selbst wissenschaftliche Analysen durch. Ziel ist es, ein besseres Verständnis der sozialen Mobilität in den USA unter Berücksichtigung lokaler ökonomischer Gegebenheiten zu erlangen. Dafür arbeitet die Initiative mit lokalen Interessengruppen, politischen Entscheidungsträger\*innen und Privatunternehmen zusammen. So sollen ihre Erkenntnisse in die Forschung einfließen und die Forschungsergebnisse in politische Initiativen umgesetzt werden. Die größten Partner in den bisher veröffentlichten Studien sind die Statistikbehörde United States Census Bureau und privatwirtschaftlich agierende Zahlungs- und Personalvermittlungsunternehmen wie beispielsweise Affinity Solutions oder Lightcast.

Ein Beispiel für ein innovatives Forschungsprojekt ist eine Untersuchung, welchen Einfluss die Stärke von zwischenmenschlichen Beziehungen auf das individuelle Einkommen und die Gesundheit hat.

Diese Analyse wurde durch die Zusammenarbeit mit Facebook ermöglicht. Chetty et al. (2022c) nutzten hierbei den Zugriff auf 21 Milliarden Facebook-Freundschaften. Dabei wird die Stärke zwischenmenschlicher Beziehungen („social capital“) in drei Kategorien gemessen: (i) Kohäsion: das Ausmaß der Fragmentierung zwischenmenschlicher Verbindungen in soziale Gruppen; (ii) ökonomische Verbundenheit: das Ausmaß, mit dem Personen aus niedrigen und aus hohen Einkommensschichten miteinander in Kontakt stehen; (iii) bürgerschaftliches Engagement: das Ausmaß von ehrenamtlicher Tätigkeit und Mitgliedschaft in Gemeinschaftsorganisationen bzw. gemeinnützigen Vereinen. Die berechneten Indikatoren zu diesen drei Kategorien werden auf Ebene von Postleitzahlbezirken und für alle Highschools und Colleges in den USA zur Verfügung gestellt.

Ein weiteres Ziel von Opportunity Insights ist es, die wissenschaftliche Ausbildung von Sozialwissenschaftler\*innen und Praktiker\*innen in Bezug auf Datenverarbeitung und -analyse zu verbessern. Dies erfolgt u. a. durch die Bereitstellung von öffentlich verfügbaren Vorlesungen und Anleitungen zu Big Data sowie den von Opportunity Insights veröffentlichten interaktiven Datenvisualisierungen. Zudem verweist das Team von Opportunity Insights auf die Zusammenarbeit mit den Bildungsplattformen PREDOC und edX, die ebenfalls Lernmodule zum Themenbereich ökonomische Chancengerechtigkeit und verwandten sozialen Themen anbieten.

Die Finanzierung der Einrichtung erfolgte vor allem durch die privaten Stiftungen Bill and Melinda Gates Foundation und Chan Zuckerberg Initiative. Förderungsübersichten dieser beiden Stiftungen zeigen, dass die Fördersummen zum Start der Einrichtung bei jeweils 15 Mio. USD lagen (Gates Foundation, 2023; Chan Zuckerberg, 2022). Die Finanzierung durch die Bill and Melinda Gates Foundation wurde dabei für eine Dauer von fünf Jahren initiiert. Weitere Stiftungen wie z. B. die Overdeck Family Foundation trugen ebenfalls zur Finanzierung bei (Eroh, 2021).

In Deutschland wird zwar die gebündelte Bereitstellung von Daten an verschiedensten Institutionen bereits erfolgreich praktiziert. Jedoch geht Opportunity Insights deutlich darüber hinaus, insbesondere durch die große Zahl an Kooperationen mit Privatunternehmen, die transparente automatische Aktualisierung und Anonymisierung sowie die übersichtlich gestaltete Visualisierung. Das Projekt kann somit Vorbildcharakter für den Ausbau der öffentlich zugänglichen Daten im Bereich der Arbeitsmarkt- und Sozialpolitik in Deutschland haben. Die Übertragbarkeit auf Deutschland wird jedoch durch finanzielle Restriktionen sowie strengere Datenschutzauflagen begrenzt.

#### 4.4.2 Technische Beschreibung

##### 4.4.2.1 Datensammlung und -aufbereitung (private und öffentliche Daten, Kooperation mit Unternehmen)

Grundsätzlich beruhen die auf der Webseite von Opportunity Insights zur Verfügung gestellten Datensätze und die dazugehörigen Visualisierungen auf zwei Datenquellen: anonymisierten Datensätzen von Privatunternehmen wie z. B. Facebook und öffentlich zugänglichen Daten. Diese Daten werden mithilfe von eindeutigen Identifikatoren, insbesondere der Region, in der ein Unternehmen tätig bzw. eine Person ansässig ist, zusammengeführt. Zudem gibt es für spezifische Forschungsfragen Datenzusammenführungen und entsprechende Auswertungen auf Ebene von Highschools und Colleges.

Zu jedem verwendeten Datensatz gibt es eine ausführliche Dokumentation in Form von PDF-Dateien oder GitHub-Archiven (Opportunity Insights Team, 2023c). Diese Dokumentation umfasst unter anderem die jeweils enthaltenen Variablen, den Erhebungszeitraum, die Erhebungsfrequenz sowie Informationen zur räumlichen Abdeckung. Vor ihrer Zusammenführung werden die einzelnen Datensätze vom Projektteam auf etwaige Mängel untersucht und die Mängel behoben, beispielsweise

werden statistische Ausreißer ausgeschlossen (Opportunity Insights Team, 2023d). Im zeitlichen Verlauf der Datenverarbeitung werden Aktualisierungen der Methodik vorgenommen, um politischen oder gesellschaftlichen Veränderungen oder der Veränderung von Datensatzstrukturen Rechnung zu tragen. Die Methodik-Aktualisierungen betreffen z. B. die Änderung bestimmter, für die Analyse notwendiger Schwellenwerte aufgrund einer Mindestloohnerhöhung oder die Verschiebung von Einkommensquantilen, die sich aus allgemeinen Lohn- und Inflationssteigerungen in der breiten Bevölkerung ergeben.

Tabelle 4-1 zeigt eine Auswahl von fünf Forschungsvorhaben von Opportunity Insights, für die jeweils mehrere Datensätze von unterschiedlichen Datenlieferanten zusammengeführt wurden, z. B. über geografische Indikatoren oder Bildungseinrichtungen. Pro Studie bzw. Themenfeld werden auf der Homepage mehrere Datensätze zum Download in unterschiedlichen Formaten angeboten, die verschiedene Variablen bzw. Aggregationsebenen umfassen.

**Tabelle 4-1: Datenquellen für Forschungs- und Datenprojekte von Opportunity Insights**

Studie	Zeitliche Abdeckung	Anzahl der Datensätze zum Download	Zusammenführung über	Kooperationen mit Datenprovidern/ Datenquellen	Dateiformat
Social Capital II: Determinants of Economic Connectedness (Chetty et al., 2022b)	2018	5	Geografische Indikatoren, Bildungseinrichtungen (z. B. Schule, College)	Facebook, National Center for Education Statistics, Civil Rights Data Collection, Integrated Postsecondary Education Data System, College Mobility Report Cards (Tax Data)	CSV, PDF
The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data (Chetty et al., 2022a)	2019-2023	35	Geografische Indikatoren	Affinity Solutions Inc, Coinout, Womply, Paychex, Intuit, Earnin, Lightcast, Zearn, Department of Labor and State Government Agencies, New York Times, Johns Hopkins, CDC, U.S. Department of Health and Human Services, Google	CSV
The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility (Chetty et al., 2020)	1978-2015	12	Geografische Indikatoren	Census Bureau	DTA, EXCEL, PDF
Who Becomes an Inventor in America? The Importance of Exposure to Innovation (Chetty et al., 2019b)	1996-2012	7	Individuen und geografische Indikatoren	United States Patent and Trademark Office, Social Security Administration, New York City school district, Patent-Datensatz (Strumsky, 2014), Erfinderdaten (Jones, 2010), Steuerdaten (Chetty et al., 2014a)	DTA, EXCEL, PDF

Studie	Zeitliche Abdeckung	Anzahl der Datensätze zum Download	Zusammenführung über	Kooperationen mit Datenprovidern/ Datenquellen	Dateiformat
Childhood Environment and Gender Gaps in Adulthood (Chetty et al., 2016)	1980-2012	3	Geografische Indikatoren	Einkommensdatensatz (Chetty et al., 2014b)	DTA, EXCEL, PDF

Quelle: Eigene Darstellung, basierend auf Opportunity Insights (2023a).

Die beschriebenen Datenlieferanten umfassen neben mehreren US-amerikanischen Behörden auch verschiedene Unternehmen. Hierbei handelt es sich beispielsweise um Facebook (Daten zur Identifizierung von Freundschaften), Affinity Solutions Inc. und Coinout (Informationen zu Konsumausgaben), Womply (Informationen zu den Umsätzen von Kleinunternehmen), Lightcast (Online-Stellenanzeigen), Zearn (Bildungsdaten), die Abrechnungsdienstleistungsunternehmen Paychex, Intuit und Earnin (Informationen zu Anstellungen und Einkommen), Zillow, Redfin und Apartments.com (Immobilien Daten), New York Times (Daten zu aktuellen COVID-19-Fällen) und Google (öffentlich verfügbare Informationen zum Aufenthalt von Personen in den USA während der Corona-Pandemie).

#### 4.4.2.2 Datenschutz

Damit die Datenschutzrichtlinien eingehalten werden, verwenden Chetty et al. (2019a) eine automatisierte Methodik. Insbesondere die kleinräumigen Daten bringen ein hohes Risiko mit sich, sensible Daten unwillentlich öffentlich zu machen. Der Ansatz von Opportunity Insights zur Verhinderung von Verletzungen des Datenschutzes ist ein Algorithmus, der die Daten durch das Hinzufügen von statistischem Rauschen gezielt zufällig verändert.<sup>17</sup> Hierzu wird die betreffende Variable geschätzt, in einem nächsten Schritt die lokale Sensitivität berechnet und die maximal beobachtete Sensitivität hergeleitet. Die lokale Sensitivität beschreibt, wie sensibel ein statistisches Modell auf lokale Unterschiede oder Änderungen reagiert. Eine hohe Sensitivität ist für bessere ökonomische Ergebnisse wünschenswert, daraus folgt jedoch eine erhöhte Wahrscheinlichkeit sensible Daten preiszugeben, sodass man auf Personen schließen kann. Auf die Bestimmung der Sensitivität folgt die Bestimmung eines sogenannten Privatsphäre-Parameters, welcher den Datenschutzrichtlinien des jeweiligen Daten-Providers entspricht. Dieser Privatsphäre-Parameter setzt sich aus den von dem jeweiligen Datenlieferanten angeforderten Datenschutz-Richtlinien zusammen und wird dementsprechend individuell erstellt. Anschließend wird ein zufälliges Rauschen, proportional zu der maximalen beobachteten Sensitivität und zum Privatsphäre-Parameter, auf die geschätzte Variable gelegt. Für diesen Prozess wird bei Aufnahme eines neuen Datensatzes in das Projekt Opportunity Insights einmalig manuell ein Code erstellt, der bei späteren Aktualisierungen des Datensatzes automatisch aktualisiert und angewendet wird.

<sup>17</sup> Siehe auch die Methode „Differential privacy“ in Steckbrief 4.1.

#### 4.4.2.3 Automatische Datenaktualisierung

Das Team um Opportunity Insights erhält in regelmäßigen, mit den jeweiligen Unternehmen abgestimmten Abständen ein Datensatz-Update. Diese Datenaktualisierung wird von Opportunity Insights mithilfe einer sogenannten Automated Data Processing Pipeline (ADPP) vorgenommen, sodass manuelle Aktualisierungen bzw. Änderungen auf ein Minimum reduziert werden und die Daten umgehend auf der Homepage bereitgestellt werden. Je nach Datenstruktur erfolgt diese Aktualisierung z. B. auf täglicher oder wöchentlicher Basis. Die ADPP sieht eine dreistufige Vorgehensweise vor (Chetty et al. 2022a).

Im ersten Schritt werden die Daten, die nicht frei zugänglich sind, in regelmäßigen Abständen über einen Cloud-Server von den Datenanbietenden bereitgestellt. Um dabei den Sicherheitsanforderungen der verschiedenen privaten Datenanbietenden gerecht zu werden, werden entsprechende Skripte unter Verwendung von sicheren Dateiübertragungsdiensten, wie z. B. SFTP-Servern, erstellt. Der zweite Schritt beschreibt die Automatisierung der Datenverarbeitung. Für jeden Datensatz gibt es eine eigene Reihe von Programmen, die nach der Bereitstellung der aktualisierten Daten automatisch ausgeführt werden. Diese Abfolge von Programm-Ausführungen wurde von Opportunity Insights in der Statistik-Software Stata automatisiert (Picard, 2013). In diesem Schritt der Datenverarbeitung entstehen zwei Dokumente. Zum einen werden publikationsfähige CSV-Dateien erstellt, zum anderen werden Qualitätskontrollberichte ausgegeben, die Auskunft über Datenveränderungen geben. Mithilfe dieser Berichte sollen den Datenanalytist\*innen potenzielle, manuell notwendige Überarbeitungsstellen schnell erkennbar gemacht werden. Dieser Schritt erfolgt anhand eines Git-Pull-Requests, der die anderen an der Datenanalyse beteiligten Personen über die erfolgten Veränderungen informiert. Diese Veränderungen müssen dann manuell bestätigt werden, um den dritten Schritt der ADPP einzuleiten. Dieser besteht in der Veröffentlichung des Datensatzes. Auch dieser Schritt erfolgt über ein eingangs erstelltes Skript, welches die Daten in zwei Verzeichnissen, eines zur Onlinevisualisierung und eines zum Datendownload, automatisch hochlädt. Dabei wird jeweilig das unter dem Download stehende Aktualisierungsdatum und das Datum der nächsten zu erwartenden Aktualisierung automatisch festgehalten.

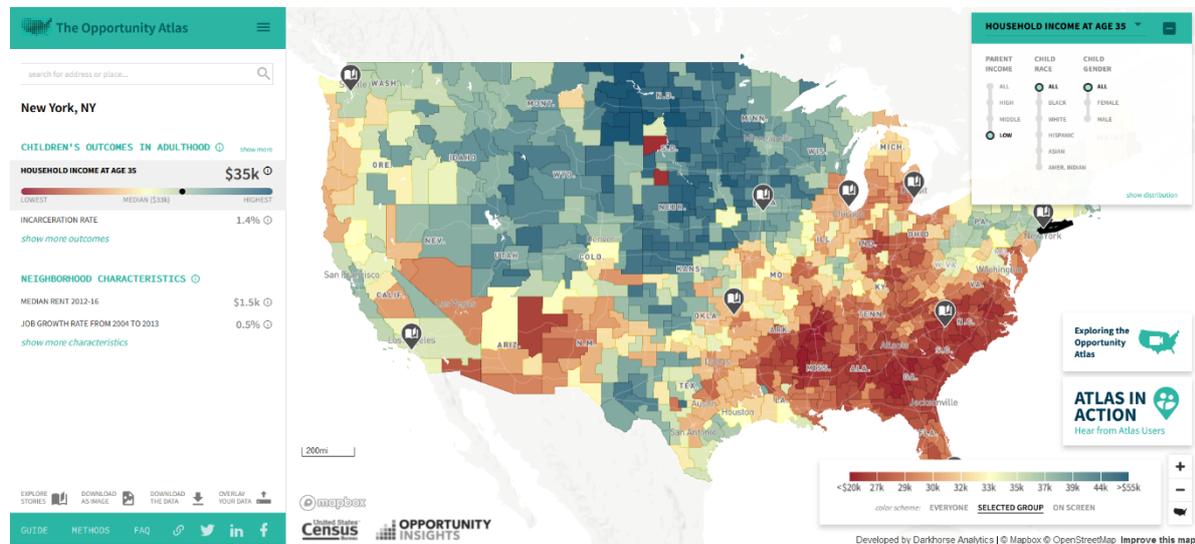
#### 4.4.2.4 Interaktive Visualisierung und Veröffentlichung

Eine der wichtigsten Innovationen der Datenprojekte von Opportunity Insights ist die frei verfügbare und interaktive Visualisierung und Veröffentlichung eines Teils der Daten, die sich an der regelmäßigen Aktualisierung der Datensätze orientiert (Abbildung 4-3). Opportunity Insights hat zu vier seiner Studien eigenständige Homepages entwickelt, auf denen ausgewählte Daten interaktiv visualisiert werden. Dazu gehört der Opportunity Atlas (Chetty et al., 2018), der Economic Tracker (Chetty et al., 2023), der Social Capital Atlas (Chetty et al., 2022c) und die Migration Patterns (Hendren et al., 2022). Zusätzlich gibt es in Zusammenarbeit mit der New York Times eine auf deren Webseite veröffentlichte Visualisierung einer weiteren Studie zu wirtschaftlichen Aufstiegschancen von Studierenden an verschiedenen Universitäten in den USA (New York Times Company, 2017).

Der Opportunity Atlas ist beispielsweise eine Darstellung von verschiedenen Merkmalen (z. B. Einkommen im Alter von 35, Inhaftierungsrate, Anzahl der Kinder) auf Ebene der Counties (was etwa der deutschen Kreisebene entspricht). Mithilfe der Visualisierung lässt sich die regionale Verteilung dieser Merkmale anzeigen. Dabei kann die Betrachtung auf bestimmte Teilpopulationen gemäß der ethnischen Herkunft, des Geschlechts und der Einkommensgruppe der Eltern begrenzt werden. Eine Legende hebt dabei die regionalen Unterschiede hervor. Eine kleinteiligere Betrachtung der Regionen bis auf die Ebene der Census Tracts (Zählbezirke) ist, soweit die Daten es zulassen, durch Scrollen möglich. Die jeweiligen Grafiken lassen sich zudem als Bild oder Datensatz herunterladen. Zusätzlich ist es möglich, einzelne Regionen oder Städte zur besseren Illustration hervorzuheben.

Bei einigen Visualisierungen erfolgt bereits eine regelmäßige Aktualisierung über die ADPP. Dies ist beim Opportunity Atlas und beim Economic Tracker der Fall. Zukünftig sollen noch weitere Visualisierungen mithilfe der ADPP automatisiert erfolgen.

### Abbildung 4-3: The Opportunity Atlas



Quelle: <https://www.opportunityatlas.org/> (Chetty et al., 2018). Visualisierung der durchschnittlichen jährlichen Haushaltseinkommen für die Jahre 2014–2015 nach Gebieten, in denen die jeweiligen Personen in ihrer Kindheit aufgewachsen sind.

## 4.4.3 Anwendung und Beispiele

### 4.4.3.1 Politikmaßnahmen / Politikinitiativen

Die zusammengestellten Datensätze des Teams von Opportunity Insights werden unter anderem von Medien, Politik und Zivilgesellschaft genutzt, teilweise in Zusammenarbeit mit Wissenschaftler\*innen. So haben sich beispielsweise verschiedene Initiativen aus Opportunity Insights entwickelt, die nach der Veröffentlichung von Forschungsergebnissen gegründet wurden. Das Projekt „Creating Moves to Opportunity“ hat z. B. zum Ziel, der Chancengleichheit zwischen unterschiedlichen Wohnorten entgegenzuwirken, also den Unterschied im Potenzial, im Laufe des Lebens einen höheren Lebensstandard zu erreichen. Die CLIMB- („Collegiate Leaders in Increasing MoBility“) Initiative, der unter anderem eine große Anzahl an Hochschulen als Partner angehören, soll die wirtschaftlichen Aufstiegschancen in den USA verbessern und Mobilitätsbarrieren verringern. Eine wichtige Zielgröße ist die Chance, dass Studienabsolvent\*innen aus Familien mit niedrigen Einkommen in eine höhere Einkommensgruppe gelangen. Als weitere, lokale Initiative wurde in der Stadt Charlotte im Bundesstaat North Carolina die Charlotte Opportunity Initiative gegründet, um die wirtschaftlichen Chancen von Kindern in dieser Stadt zu verbessern. Der Anlass waren Forschungsergebnisse von Opportunity Insights, denen zufolge Kinder aus Haushalten mit niedrigem Einkommen in dieser Region besonders benachteiligt sind, da die Region im Vergleich zu den anderen US-Regionen äußerst niedrige Aufstiegschancen aufwies.

Die Veröffentlichungen und Datenvisualisierungen von Opportunity Insights erhalten durch die Berichterstattung in den Medien regelmäßig große Aufmerksamkeit. So zitieren beispielsweise Tages- und Wochenzeitungen wie die Washington Post, die New York Times und das Wall Street Journal die Ergebnisse der Projekte und Studien von Opportunity Insights (Adamy, 2020; McArdle, 2022). Besonderes Augenmerk auf die Arbeit des Teams um Opportunity Insights fiel dabei auf die

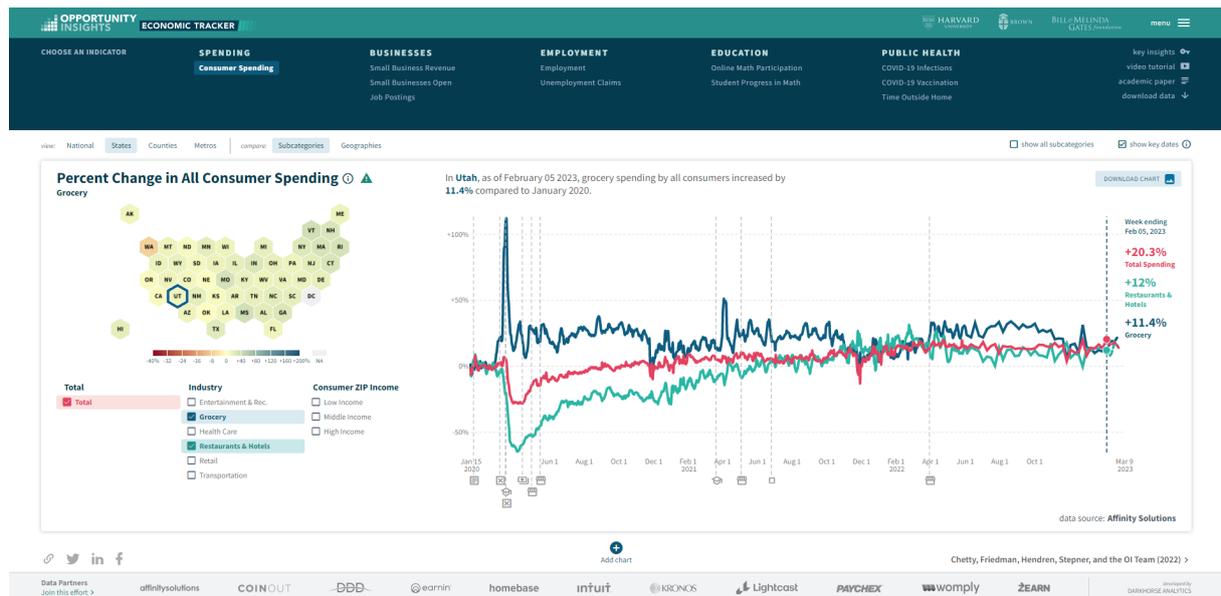
Studie zu den ökonomischen Folgen der Corona-Pandemie und der Wirksamkeit staatlicher Hilfsmaßnahmen (Chetty et al., 2022a).

#### 4.4.3.2 Angegliederte Projekte

Die größten Datenprodukte von Opportunity Insights haben sich zu eigenen Projekten entwickelt. Diese Projekte werden innerhalb von Opportunity Insights größtenteils als eigenständige Arbeitsgruppen gesehen oder sind aus Opportunity Insights ausgegliedert (z. B. Policy Impacts).

Ein aus den Analysen von Opportunity Insights entwickeltes Projekt ist der „Economic Tracker“ (Abbildung 4-4), dessen Ziel die zeitnahe Veröffentlichung von Daten zur wirtschaftlichen Lage ist. Das Projekt bündelt eine große Menge an privaten Unternehmen- und öffentlichen Verwaltungsdaten und veröffentlicht diese mit entsprechenden Visualisierungen auf der Webseite von Opportunity Insights. Zu den Daten gehören u. a. Unternehmensdaten von Abrechnungsunternehmen und Finanzdienstleistern. Die Daten sind Teil der ADPP und werden auf wöchentlicher Basis aktualisiert. Durch die hohe Frequenz und den großen Umfang dieser Daten sind die Ergebnisse aktueller und genauer als andere Indikatoren. Zudem sind die Daten regional kleinräumig gegliedert, anders als bei herkömmlichen Datenquellen zu diesem Themenbereich. Ähnlich wie beim Opportunity Atlas erfolgt die Visualisierung interaktiv auf einer separaten Homepage. Der Economic Tracker umfasst u. a. Informationen zu Konsumausgaben, Geschäftsgründungen und -schließungen, Beschäftigung und COVID-19 Infektionen sowie Impfungen. Alle Informationen lassen sich dann in einem weiteren Schritt in kleinere Betrachtungsobjekte teilen, z. B. nach Region oder Wirtschaftszweig.

Abbildung 4-4: Economic Tracker



Quelle: Chetty et al. (2023). Visualisierung der prozentualen Konsumveränderung für den Bundesstaat Utah für die Industriezweige Lebensmittel, Restaurants und Hotels.

Ergänzend zu den genannten Initiativen haben sich weitere Projekte entwickelt, durch die Daten und Forschungsergebnisse der Öffentlichkeit zugänglich gemacht werden. Das Projekt „Migration Patterns“ befasst sich mit der Migration von jungen Erwachsenen und daraus folgenden Erkenntnissen für Politikmaßnahmen. Dieses Projekt ist eine Zusammenarbeit von Opportunity Insights, Policy Impacts und dem United States Census Bureau. Policy Impacts ist ein Forschungs-

und Politikberatungsinstitut, das 2021 durch Opportunity Insights gegründet wurde. Der Fokus dieser Organisation ist die Verbesserung der politischen Entscheidungen.

#### 4.4.4 Übertragbarkeit auf Deutschland

Opportunity Insights zeichnet sich durch eine Vielzahl von Datenquellen, auch von privaten Unternehmen, die Verknüpfung dieser Datenquellen und eine hohe Aktualität aus. Eine Zusammenarbeit zwischen Forschungsinstituten und Privatunternehmen (beispielsweise für GPS- oder Zahlungsdaten) erscheint aufgrund der vielfältigen Unternehmenslandschaft in Deutschland grundsätzlich möglich. Dies lässt sich auch an bestehenden Kooperationen zwischen Forschenden und privaten Unternehmen in Deutschland ablesen. So kooperiert beispielsweise das RWI – Leibniz-Institut für Wirtschaftsforschung mit Immoscout24, um Immobilienpreise für die Forschung verfügbar zu machen (RWI, 2023). Auf einer übergeordneten Ebene verfolgt die Nationale Forschungsdateninfrastruktur (NFDI) in der Sektion „Industry Engagement“ das Ziel, eine Austauschplattform mit wirtschaftlichen Partnern aufzubauen (NFDI, 2023). Unter der Rubrik EXSTAT veröffentlicht das Statistische Bundesamt „experimentelle“ Projekte zu innovativen Daten und Methoden, beispielsweise zur Generierung von Statistiken zu wöchentlichen Veränderungen von Online-Geschäften, neuen Kreditverträgen und neuen Hypotheken, die auf Daten von Informationsdienstleistern beruhen (Statistisches Bundesamt, o. D.).

Eine Verknüpfung, automatische Verarbeitung und Visualisierung von Daten einschließlich Big Data wie im Projekt Opportunity Insights erscheint daher auch in Deutschland grundsätzlich möglich. Die vielfältigen Erfahrungen aus den USA haben ein hohes Transferpotenzial. Dabei lässt sich die Zahl der an den vorhandenen deutschen Projekten beteiligten Institutionen und Unternehmen noch ausbauen. In Deutschland bisher recht wenig genutzte Datenquellen sind Unternehmen, die sich mit Gehaltsabrechnungen befassen, oder (private) Bildungsträger.

Bei der Übertragung auf Deutschland wären jedoch einige Herausforderungen zu bewältigen. Erstens werden viele verschiedene Datenquellen miteinander verknüpft, was die Zustimmung vieler beteiligter Partner voraussetzt. Dies geht mit einem hohen Organisationsaufwand einher, möglicherweise sind manche Datenquellen in Deutschland nicht so einfach erschließbar wie in den USA. Zweitens hat der Datenschutz in Deutschland eine höhere Bedeutung als in den USA (von Lewinski, 2017). Dies erfordert eine komplexere und genauere Betrachtung der jeweiligen Datensätze, der gesetzlichen Bestimmungen und der Vertragsbedingungen der Datenlieferanten. Erfolgt die Verknüpfung von Datensätzen ausschließlich auf regionaler (im Gegensatz zur individuellen) Ebene, erscheint eine Übertragung auf Deutschland durchaus machbar, sofern die Daten so vergrößert werden, dass keine Rückschlüsse auf einzelne Personen möglich sind. Drittens stellt sich die Frage, wer ein Projekt von der Größe von Opportunity Insights in Deutschland finanzieren könnte.

Mögliche vergleichbare Projekte in Deutschland könnten von technischen Erfahrungen aus dem Opportunity-Insights-Projekt profitieren, wobei natürlich im europäischen Kontext Anpassungen, z. B. bezüglich Cloud-Anbieter und Tools für den jeweiligen Anwendungsfall geprüft werden sollten. So beschreibt Opportunity Insights beim Umgang mit Big Data und bei der Implementierung der ADPP mehrere Probleme. Aufgrund der Größe der genutzten Datensätze ist das Speichern, das Aktualisieren und die Versionskontrolle teilweise schwierig. Zur Lösung dieser Probleme speichert das Team von Opportunity Insights Datensätze über den Objektspeicherdienst Amazon S3 in einem sogenannten Data Lake. Die Daten werden dabei in Objektspeichern gelagert, um Lagerungskosten gering- und die Datenflexibilität hochzuhalten. Objektspeicher sind eine Art von Datenbank, die Daten als Objekte speichert und organisiert. Sie werden oft in der Cloud genutzt, um große

Datenmengen zu speichern. Hinzu kommt, dass Cloud-Objektspeicher durch ihre Skalierbarkeit und Flexibilität eine effektive und kosteneffiziente Verwaltung der Daten ermöglicht.

#### 4.4.5 Fazit

Das Projekt „Opportunity Insights“ ist eine interessante Dateninnovation, die sich aus einer großen Zahl von Quellen speist und umfangreiche und detaillierte Analysen der wirtschaftlichen und sozialen Situation der Bevölkerung in den USA ermöglicht. Im Zentrum stehen dabei die ökonomischen Ungleichheiten und die unterschiedlichen Aufstiegschancen. Die Innovation besteht aus der Erschließung einer großen Anzahl von Datenquellen, ihrer Verknüpfung, der technischen Verarbeitung der Daten, ihrer fortlaufenden Aktualisierung, der Bereitstellung der Daten und Ergebnisse in anonymisierter Form sowie ihrer Visualisierung auf der Website von Opportunity Insights.

Die Daten ermöglichen neue wissenschaftliche und politikrelevante Erkenntnisse. Sie werden in der Forschung weithin genutzt und dienen als Informationsquelle für die Politik, die Zivilgesellschaft und die Öffentlichkeit. Ermöglicht wird das Projekt durch die institutionelle Verankerung an einer der führenden Universitäten der USA, die beteiligten Forschenden, die Bereitschaft zur Mitwirkung bei den verschiedenen Datenlieferanten sowie die auskömmliche Finanzierung, unter anderem aus den Mitteln verschiedener Stiftungen. Eine Übertragung auf Deutschland erscheint daher wünschenswert, um bspw. die ökonomische Mobilität auf regionaler Ebene und auf Grundlage einer umfassenden Datenbasis untersuchen zu können. Die Umsetzung der Dateninnovation scheint in Deutschland grundsätzlich möglich, wobei der organisatorische Aufwand eine Hürde darstellen könnte und der Datenschutz vollumfänglich zu gewähren ist.

#### 4.4.6 Quellen zum Steckbrief

Adamy, Janet. (2020). Job Postings Have Fallen Steadily During Pandemic, New Tracker Shows. *The Wall Street Journal*. Veröffentlicht am 7. Mai 2020. Abgerufen am 15.06.2023 von <https://www.wsj.com/livecoverage/coronavirus-2020-05-07/card/izjgLwld6yfsMTMOBiRG>

Chan Zuckerberg. (2022). *Science, Education, Community, and Justice Grants / CZI*. Chan Zuckerberg Initiative. Abgerufen am 14. März 2023 von <https://chanzuckerberg.com/grants-ventures/grants/>

Chetty, R., Friedman, J.N. & Rockoff, J.E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104 (9), 2593– 2632. <https://doi.org/10.1257/aer.104.9.2593>

Chetty, R., Hendren, N., Kline, P. & Saez, E. (2014b). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, 129(4), 1553–1623. <https://doi.org/10.1093/qje/qju022>

Chetty, R., Hendren, N., Lin, F., Majerovitz, J. & Scuderi, B. (2016). Childhood Environment and Gender Gaps in Adulthood. *American Economic Review Papers and Proceedings*, 106(5), 282–288. <https://doi.org/10.1257/aer.p20161073>

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. & Porter, S. R. (2018). *The Opportunity Atlas*. Opportunity Atlas. Abgerufen am 09. März 2023 von <https://www.opportunityatlas.org/>

Chetty, R., & Friedman, J. N. (2019a). *A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Samples* (NBER Working Paper No. 25626.). NBER Working Papers. <https://doi.org/10.1257/pandp.20191109>

- Chetty, R., Bell, A., Jaravel, X., Petkova, N. & van Reenen, J. (2019b). Who Becomes an Inventor in America? The Importance of Exposure to Innovation. *The Quarterly Journal of Economics*, 134(2), 647-713. <https://doi.org/10.1093/qje/qjy028>
- Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. & Porter, S. R. (2020). *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility* (NBER Working Paper No. 25147). NBER Working Papers. <https://doi.org/10.3386/w25147>
- Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. & Opportunity Insights Team (2022a). *The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data* (NBER Working Paper No. 27431). NBER Working Papers. <https://doi.org/10.3386/w27431>
- Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenberg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barberá, P., Bhole, M. & Wernerfelt, N. (2022b). Social Capital II: Determinants of Economic Connectedness. *Nature*, 608(7921), 122-134. <https://doi.org/10.1038/s41586-022-04997-3>
- Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenberg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barberá, P., Bhole, M. & Wernerfelt, N. (2022c). *The Social Capital Atlas*. Social Capital. Abgerufen am 16. März 2023 von <https://socialcapital.org/>
- Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. & Opportunity Insights Team. (2023). *Economic Tracker*. Track the Recovery. Veröffentlicht am 19. Februar 2023. Abgerufen am 09. März 2023 von <https://www.tracktherecovery.org/>
- Eroh, L. (2021). *Q4 2021 Funding Announcement*. Overdeck Family Foundation. Veröffentlicht am 16. November 2021. Abgerufen am 14. März 2023 von <https://overdeck.org/news-and-resources/article/q4-2021-funding-announcement/>
- Gates Foundation. (2023). *Committed Grants / Bill & Melinda Gates Foundation*. Bill & Melinda Gates Foundation. Veröffentlicht am 11. Februar 2023. Abgerufen am 14. März 2023 von <https://www.gatesfoundation.org/about/committed-grants>
- Hendren, N., Sprung-Keyser, B. & Porter, S. R. (2022). *The Radius of Economic Opportunity: Evidence from Migration and Local Labor Markets* (CES Working Paper 22-27). CES Working Papers. <https://www2.census.gov/ces/wp/2022/CES-WP-22-27.pdf>
- Jones, B. (2010). Age and Great Invention. *The Review of Economics and Statistics*, 92(1), 1-14. <https://doi.org/10.2139/ssrn.608701>
- McArdle, M. (2022). What are friends for? Climbing the income ladder, for one. *The Washington Post*. Veröffentlicht am 7. August 2022. Abgerufen am 15.06.2023 von <https://www.washingtonpost.com/opinions/2022/08/07/harvard-chetty-research-facebook-friends-income/>
- New York Times Company. (2017). College Mobility – Economic diversity and student outcomes at ... University. *The New York Times*. Veröffentlicht am 18. Januar 2017. Abgerufen am 16. März 2023 von <https://www.nytimes.com/interactive/projects/college-mobility/city-college-of-new-york>
- NFDI. (2023). *Über uns – Sektion Industry Engagement*. Nationale Forschungsdateninfrastruktur. Abgerufen am 08. März 2023 von <https://www.nfdi.de/sektion-industry-engagement/>

Opportunity Insights Team. (2023a). *Opportunity Insights / Policy Solution to the American Dream*. Opportunity Insights. Veröffentlicht am 18. Januar 2023. Abgerufen am 8. März 2023 von <https://opportunityinsights.org/>

Opportunity Insights Team (2023b). Opportunity Insights Economic Tracker Data Dictionary. Opportunity Insights. Abgerufen am 08. März 2023 von <https://opportunityinsights.org/paper/tracker/>

Opportunity Insights Team (2023c). Opportunity Insights Economic Tracker Data Documentation. Opportunity Insights. Abgerufen am 08. März 2023 von <https://opportunityinsights.org/paper/tracker/>

Opportunity Insights Team (2023d). Opportunity Insights Economic Tracker Data Revisions. Opportunity Insights. Abgerufen am 08. März 2023 von <https://opportunityinsights.org/paper/tracker/>

Picard, R. (2013). PROJECT: Stata module providing a set of tools to build and manage a Stata project. Statistical Software Components S457685. Boston College Department of Economics.

RWI. (2023). *Regional Real Estate Price Index for Germany - 2008-12/2022*. RWI – Leibniz-Institut für Wirtschaftsforschung. Abgerufen am 28. März 2023 von <https://fdz.rwi-essen.de/doi-detail/id-107807immoredxv10>

Statistisches Bundesamt. (o. D.), *EXSTAT – Experimentelle Statistiken*. Abgerufen am 26. Mai 2023 von [https://www.destatis.de/DE/Service/EXDAT/\\_inhalt.html](https://www.destatis.de/DE/Service/EXDAT/_inhalt.html).

Strumsky, D. (2014). Strumsky Patent Database (verfügbar auf Anfrage).

Von Lewinski, K. (2017). Was Europa und die USA in Sachen Datenschutz unterscheidet. Bundeszentrale für politische Bildung. Abgerufen am 14.8.2023 von <https://www.bpb.de/themen/daten/democracy/254254/was-europa-und-die-usa-in-sachen-datenschutz-unterscheidet>

## 4.5 Forschung zu Erwerbsminderungsrenten: Verknüpfung von Versicherungs- und Gesundheitsdaten (Niederlande)

### 4.5.1 Einführung

In Deutschland gab es im Jahr 2021 rund 1,8 Mio. Personen, die aus gesundheitlichen Gründen einer Erwerbstätigkeit nur in eingeschränktem Ausmaß oder gar nicht nachgehen konnten und daher Anspruch auf eine gesetzliche Erwerbsminderungsleistung hatten (Deutsche Rentenversicherung, 2022). Die Anzahl der hiervon betroffenen Personen ist seit 2013 um 5,3 Prozent angestiegen und könnte aufgrund der fortschreitenden Alterung der Bevölkerung und des höheren Renteneintrittsalters noch weiter ansteigen. Daher kommt einer Analyse der Gründe für und Auswirkungen von Erwerbsminderungsleistungen eine große Bedeutung zu.

In den Niederlanden haben Erwerbsminderungsleistungen ebenfalls eine hohe zahlenmäßige Bedeutung. Da die Datenverfügbarkeit zum Thema Erwerbsminderung dort in mehrerer Hinsicht deutlich besser ist als in Deutschland, gibt es innovative wissenschaftliche Studien in diesem Themenbereich, die Erkenntnisse über wichtige Wirkungsmechanismen liefern. So untersuchen Koning et al. (2022a), ob und inwiefern Erwerbsminderungsleistungen die Rückkehr in den Arbeitsmarkt beeinflussen. Die Autoren nutzen für ihre Studie unter anderem Daten der Erwerbsminderungsversicherung WIA<sup>18</sup> in den Niederlanden, die staatlich organisiert ist (Uitvoeringsinstituut Werknemersverzekeringen, 2014). Insgesamt werden drei verschiedene administrative Datensätze miteinander verknüpft: (1) individuelle Antragsdaten für die Erwerbsminderungsversicherung, (2) monatliche Einkommensdaten und (3) Daten zur Behandlung psychischer Erkrankungen. Die Datenverknüpfung wird dadurch ermöglicht, dass die verwendeten Datensätze alle eine personenspezifische Identifikationsnummer enthalten, die ursprünglich auf der Sozialversicherungsnummer beruht. Der resultierende Datensatz umfasst alle Personen, die seit 2006 in den Niederlanden einen Antrag auf die Leistungen der WIA gestellt haben. Innovativ an diesen Daten und der darauf basierenden Studie ist somit die Verknüpfung verschiedener administrativer Datensätze und die große Zahl der erfassten Personen.

In ihrer Studie untersuchen die Autoren den Effekt von Erwerbsminderungsleistungen auf die Erwerbsbeteiligung von Personen mit psychischen Erkrankungen. Ihre Ergebnisse zeigen, dass die Genesung von einer psychischen Erkrankung zu einem deutlichen Anstieg der Wahrscheinlichkeit einer Beschäftigung führt, dass dieser Anstieg jedoch durch die Möglichkeit, Erwerbsminderungsleistungen zu beziehen, deutlich abgeschwächt wird.

Die WIA-Leistungen unterteilen sich in Leistungen für Personen, die trotz Einschränkung ihrer Erwerbsfähigkeit zumindest noch teilweise arbeiten können (Werkhervatting Gedeeltelijk Arbeidsgeschikten, WGA), und in Leistungen für Personen, die aufgrund vollständiger Erwerbsunfähigkeit nicht mehr am Arbeitsmarkt teilnehmen können (Inkomensvoorziening Volledig Arbeidsongeschikten, IVA). Dementsprechend ist die WGA vergleichbar mit der deutschen Versicherungsleistung bei teilweiser Erwerbsminderung, die IVA mit der vollen Erwerbsminderung.

Grundvoraussetzung für die Inanspruchnahme einer WIA-Leistung ist die aufgrund einer Krankheit entstandene Unfähigkeit, in dem Umfang arbeiten zu können, der das vorherige Einkommensniveau sichert. Außerdem muss die Dauer der Erkrankung mindestens 104 Wochen ohne Unterbrechung betragen, erst danach besteht ein Anrecht auf eine WIA-Leistung in Höhe von maximal 70 Prozent

---

<sup>18</sup> WIA ist die Abkürzung für „wet Werk en Inkomens naar Arbeidsvermogen“ (Gesetz über Arbeit und Einkommen nach Arbeitsfähigkeit), in dem die unterschiedlichen Abstufungen der Erwerbsminderungsrente geregelt sind. Die Abkürzung wird auch als Bezeichnung für die Leistungen der Erwerbsminderungsversicherung selbst verwendet.

der Differenz zwischen dem letzten Bruttogehalt und dem aktuellen Bruttogehalt. Für WGA-Leistungen gilt zusätzlich die Bedingung, dass die betroffene Person voraussichtlich wieder in den Arbeitsmarkt eintreten und dort bis zu 65 Prozent ihres vorherigen Erwerbseinkommens verdienen kann. Für IVA-Leistungen gilt hingegen die zusätzliche Bedingung, dass die betroffene Person bei Eintritt in den Arbeitsmarkt nach 104 Wochen Krankheit nur maximal 20 Prozent des vorherigen Erwerbseinkommens beziehen wird und eine Änderung mit ziemlicher Sicherheit nicht absehbar ist. Die Prüfung des Grads der Erwerbsminderung wird von der Versicherungsagentur für Arbeitnehmer\*innen (Het Uitvoeringsinstituut Werknemersverzekeringen, UWV) vorgenommen. Selbstständig Beschäftigte sind von den WIA-Leistungen ausgeschlossen.

In Deutschland haben Personen Anspruch auf die gesetzliche Erwerbsminderungsrente, die aufgrund einer Krankheit täglich weniger als sechs Stunden arbeiten können. Bei einem verbliebenen Leistungsvermögen von drei bis unter sechs Stunden täglich spricht man von einer teilweisen Erwerbsminderung. Die Rente wegen voller Erwerbsminderung wird gezahlt, wenn die versicherte Person aufgrund einer Krankheit täglich weniger als drei Stunden auf dem allgemeinen Arbeitsmarkt tätig sein kann. Die Feststellung erfolgt in beiden Fällen auf der Grundlage ärztlicher Unterlagen und Gutachten (Deutsche Rentenversicherung, o. D.).

Zwischen den Niederlanden und Deutschland gibt es beim Thema Erwerbsminderung also Gemeinsamkeiten. In beiden Ländern spielen Erwerbsminderungsleistungen eine wichtige Rolle, die institutionelle Ausgestaltung ist zudem vergleichbar. Die Schaffung einer ähnlichen Datenverknüpfung für Analysezwecke in Deutschland würde im Bereich der Rentenversicherung wichtige Erkenntnisse zu den Wirkungen von Erwerbsminderungsrenten ermöglichen. Der Datenschutz ist hierbei ein besonders wichtiges Thema, da es sich größtenteils um äußerst sensible Daten auf Personenebene handelt, die aus unterschiedlichen Quellen miteinander verknüpft werden müssen.<sup>19</sup> Dies wird in Abschnitt 4.5.4 detailliert erläutert.

## 4.5.2 Technische Beschreibung

### 4.5.2.1 Datenquellen

Für die Studie von Koning et al. (2022a) wurden Personendaten aus drei verschiedenen niederländischen Quellen zusammengeführt, welche alle administrativen Ursprungs sind. Sie stammen entweder von der UWV<sup>20</sup> oder dem Centraal Bureau voor de Statistiek (CBS, Niederländische Behörde für Statistik).

Die erste Quelle umfasst Antragsdaten der Erwerbsminderungsversicherung für den Zeitraum von 2006 - 2017 und beinhaltet alle in dieser Zeit gestellten Anträge. Dieser Datensatz umfasst Informationen zum vorherigen Stundenlohn und dem vorherigen Arbeitsumfang in Stunden, Informationen zur Erwerbsminderung inklusive deren Schwere und zur potenziellen zukünftigen Arbeitsleistung und dem dazugehörigen Arbeitsentgelt, das von der UWV ermittelt wird. Ein weiterer wichtiger Aspekt dieser Datenquelle ist die Klassifizierung von medizinischen Diagnosen. Dies geschieht über UWV-Kategorien, die von der UWV festgelegt werden, und über die detaillierteren CAS-Codes (Classificaties voor Arbo en SV), die von Ärzt\*innen verwendet werden, um den Wiedereingliederungsprozess von über längere Zeit erwerbsunfähigen Versicherten zu begleiten. Die CAS-Codes bestehen aus einem Diagnose-Buchstaben, der den primären Grund für eine

---

<sup>19</sup> Siehe Abschnitt 3.1.

<sup>20</sup> Die UWV verwaltet die Sozialversicherungen von abhängig Beschäftigten in den Niederlanden. Dazu zählen u. a. die Arbeitslosenversicherung, Krankengeld, Mutterschutz und die Leistungen für Arbeitseinschränkungen.

Erwerbsminderung nennt, und drei Ziffern, die die Diagnose spezifiziert bzw. Komorbiditäten identifiziert, z. B. P630 für „Psychische Krankheit, Angststörung, Panikstörung“ (Uitvoeringsinstituut Werknemersverzekeringen, 2002). In der Analyse werden sowohl die UWV-Kategorien als auch die CAS-Codes eingesetzt. Letztere sind detaillierter, liegen aber für weniger Personen vor (Koning et al., 2022a).

Die zweite Quelle enthält Informationen zum Erwerbseinkommen, die vom niederländischen Statistikamt zur Verfügung gestellt werden. Dabei werden zwei verschiedene Datensätze herangezogen. Der eine beinhaltet Informationen über alle Arbeitsverträge für die Jahre 2006 - 2018, der andere enthält monatliche Indikatoren für die verschiedenen Einkommensarten für die Jahre 1999 - 2016. Zu diesen Indikatoren gehören u. a. Beschäftigungsart oder Einkünfte über Sozialleistungen. Personen eines Haushalts werden über einen Haushalts-Identifikator miteinander verknüpft, um entsprechende Sozialhilfeansprüche zu ermitteln. Da z. B. die Sozialhilfe abhängig vom Einkommen des/der Lebenspartner\*in ist, ist die Information zum Haushaltseinkommen zur Berechnung des Sozialhilfeanspruchs für das UWV unabdingbar.

Die dritte und letzte Datenquelle besteht aus Gesundheitsdaten. Diese setzen sich zusammen aus gesetzlichen Krankenversicherungsdaten für die Jahre 2009 - 2017 und Informationen zu Behandlungen psychischer Erkrankungen für die Jahre 2011 - 2016. Die obligatorische Grundkrankenversicherung deckt dabei den Großteil der gesamten Gesundheitsversorgung ab und ist für alle Einwohner\*innen der Niederlande gesetzlich vorgeschrieben. Somit enthält auch diese Datenquelle alle für die Analyse relevanten Personen.

#### 4.5.2.2 Verknüpfung und Datenschutz

Die verwendeten Datensätze können mithilfe einer anonymisierten Identifikationsnummer, die ursprünglich auf der Sozialversicherungsnummer beruht, auf individueller Ebene verknüpft werden. Die anonymisierte Identifikationsnummer wird durch einen 1:1-Match mit der Sozialversicherungsnummer erzeugt. Um die Datensicherheit zu gewährleisten, werden alle Daten auf den Servern des CBS gespeichert und Forschende können nur über eine sichere Verbindung (VPN) durch Fernverarbeitung auf die Daten zugreifen. Die Analysen finden also auf den Rechnern des CBS statt. Jeder Export von Ergebnissen wird durch das CBS geprüft, bevor die Forschenden sie an einen Ort außerhalb der sicheren Umgebung exportieren können. Die Ausgaberegeln stellen sicher, dass die Ergebnisse niemals mit einzelnen Arbeitnehmer\*innen oder Unternehmen in Verbindung gebracht werden können. Alle Ergebnisse müssen darüber hinaus auf mindestens zehn Beobachtungen (Personen oder Unternehmen) beruhen.

Auf Grundlage der drei verknüpften Datenquellen wurde eine 0,75- bis 1-Prozent-Stichprobe von Personen erstellt, die Erwerbsminderungsleistungen in den Niederlanden beantragt haben. Für diese Personen sind Informationen aus dem Antragsprozess, zu ihrer Berufs- und Einkommensgeschichte sowie zu ihrer (psychischen) Gesundheit verfügbar. Nach mehrmaliger Stichprobenreduzierung während des Verknüpfungsprozesses enthält diese Datensatzverknüpfung noch 5.003 Anträge. Die Reduzierungen wurden vorgenommen, um die unterschiedlichen Perioden der Datensätze zu vereinheitlichen und etwaige Analyseverzerrungen, z. B. durch zusätzliche körperliche Erkrankungen, zu minimieren.

#### 4.5.3 Anwendung

In ihrer Studie untersuchen Koning et al. (2022a) den Effekt von Erwerbsminderungsleistungen auf die Erwerbsbeteiligung von Personen mit psychischen Krankheiten. Sie nutzen dabei die personenspezifische medizinische Diagnose und den daraus berechneten Prozentwert der Erwerbsminderung. Auf dieser Grundlage vergleichen sie Personen, die sich knapp unterhalb der

Anspruchsgrenze für eine Erwerbsminderungsleistung befinden und diese infolgedessen nicht erhalten, mit Personen, die knapp oberhalb der Anspruchsgrenze liegen und die Leistung folglich erhalten. Dieser Vergleich ermöglicht die Schätzung des kausalen Effekts der Berechtigung zum Bezug von Erwerbsminderungsleistungen auf das Arbeitsangebot nach dem Zeitpunkt der Genesung. Die Ergebnisse zeigen, dass die Genesung von einer psychischen Erkrankung zu einem deutlichen Anstieg der Wahrscheinlichkeit einer Beschäftigung führt, dass dieser Anstieg durch die Erwerbsminderungsleistung aber deutlich abgeschwächt wird. Die Autoren argumentieren daher, dass Fehlanreize, die zu einer geringeren Wiederaufnahme von Beschäftigung oder einem verminderten Erwerbsumfang führen, bei der Gestaltung von Erwerbsminderungsleistungen beachtet werden sollten.

Eine weitere Studie derselben Autoren verwendet den gleichen Ansatz der Datenverknüpfung (Koning et al, 2022b). Ziel dieser Studie ist die Beantwortung der Frage, warum befristet Beschäftigte in den Niederlanden eine schlechtere Gesundheit und ein um 30 Prozent höheres Risiko für den Bezug von Erwerbsminderungsleistungen haben als unbefristet Beschäftigte. Auch hier werden die bereits beschriebenen Datensätze der UWV und des CBS genutzt. Laut der Studie ist die schlechtere Gesundheit der befristet Beschäftigten weder dadurch zu erklären, dass Personen mit schlechter Gesundheit systematisch häufiger befristete Beschäftigungsverhältnisse eingehen als Personen mit guter Gesundheit, noch durch einen negativen Effekt von befristeten Verträgen per se auf die Gesundheit. Wichtig sind hingegen zwei andere Faktoren: Erstens kümmern sich Arbeitgeber\*innen weniger um die Gesundheit ihrer befristet Beschäftigten, z. B. indem sie entsprechende gesundheitliche Maßnahmen einleiten, als um die Gesundheit der unbefristet Beschäftigten. Zweitens gehen geringere Beschäftigungschancen in bestimmten Teilarbeitsmärkten bei befristet Beschäftigten mit einer schlechteren Gesundheit einher.

#### 4.5.4 Übertragbarkeit auf Deutschland

Die beschriebene Dateninnovation beruht auf der Verknüpfung von Datenquellen zur Erwerbsminderungsrente, zu Arbeitsmarktergebnissen und zu medizinischen Diagnosen. Für die Übertragbarkeit auf Deutschland ist die Zulässigkeit des Vorgehens im Hinblick auf den Datenschutz zu prüfen. Die Verknüpfung von personenbezogenen Daten aus unterschiedlichen Quellen erfordert eine sorgfältige Abwägung zwischen dem Schutz der Privatsphäre und der Notwendigkeit, aussagekräftige Daten zu generieren. Datenschutzrechtliche Bestimmungen stellen dabei hohe Anforderungen an die Datenverarbeitung (siehe Abschnitt 3.1). Diese sichern eine mit dem Grundrecht auf informationelle Selbstbestimmung konforme Zusammenführung von Daten. Im Bereich der Gesundheitsinformationen gelten besonders hohe Anforderungen an den Datenschutz (z. B. Privat-, insbesondere Patientengeheimnis nach § 203 StGB).

Zusätzlich zu den Anforderungen des Datenschutzes ist auch die praktische Umsetzung eine Herausforderung, da im vorliegenden Beispiel Daten aus zwei verschiedenen Systemen (Gesundheitssystem, Rentenversicherung) verknüpft werden. In den Niederlanden wird die Datenverknüpfung durch die Existenz einer staatlichen Pflichtversicherung und eine in den Datensätzen enthaltene anonymisierte personenspezifische Identifikationsnummer, die ursprünglich auf der Sozialversicherungsnummer beruht, technisch ermöglicht.

Ein weiteres Problem ist die unterschiedliche Qualität und Struktur der Daten, die von verschiedenen Stellen erfasst werden. Oftmals sind die Daten nicht standardisiert und in verschiedenen Formaten gespeichert, was eine Verknüpfung erschwert. Es ist zu erwarten, dass verschiedene Träger jeweils unterschiedliche (und unterschiedlich viele) Informationen in unterschiedlicher Frequenz sammeln. Dies trifft beispielsweise für Leistungen zur Teilhabe zu, die zwischen der Bundesagentur für Arbeit und der gesetzlichen Renten- sowie Krankenversicherung aufgeteilt sind. Bei der Übertragbarkeit auf

Deutschland besteht eine zusätzliche Herausforderung darin, dass in vielen deutschen Datensätzen nicht identifiziert werden kann, ob Personen dem gleichen Haushalt angehören – dies ist z. B. bei den Daten des IAB ein Problem. Diese Information ist aber notwendig, weil das Haushaltseinkommen für viele Arten von Transferleistungen eine Rolle spielt.

Die empirische Strategie zur Identifikation des kausalen Effekts von Erwerbsminderungsleistungen kann überdies nicht direkt von den Niederlanden auf Deutschland übertragen werden. Grund hierfür ist, dass es in Deutschland, im Gegensatz zu den Niederlanden, kein Punkteverfahren gibt, mit dem Erwerbsunfähigkeit festgestellt wird. Vielmehr wird in Deutschland die Entscheidung auf Basis der Aussagen medizinischer Gutachten gefällt, die nicht standardisiert sind. Daher müssten alternative Verfahren zur Identifikation eines kausalen Effekts, wie z. B. die Nutzung exogener regionaler Variation oder der Einsatz von Instrumentvariablenansätzen, geprüft werden.

Insgesamt sind also rechtliche, technische, organisatorische und forschungspraktische Herausforderungen zu bewältigen, um Daten für die Analyse der Erwerbsminderungsleistungen mit anderen administrativen Daten in Deutschland zu verknüpfen. In diesem Zusammenhang ist die Arbeit des FDZ Gesundheit in Deutschland von großem Interesse. Dieses erschließt die Abrechnungsdaten der gesetzlich Krankenversicherten in Deutschland bereits und plant eine Erweiterung des Datenbestands mit Daten aus dem Bereich der sonstigen Leistungserbringer des Gesundheitswesens (FDZ Gesundheit, o. D.).

#### 4.5.5 Fazit

Die vorliegende Dateninnovation aus den Niederlanden besteht in der Verknüpfung verschiedener Datensätze, insbesondere zu Erwerbsminderungsleistungen, die durch die Nutzung einer persönlichen ID unter Wahrung des Datenschutzes möglich gemacht wird. Hierdurch entstehen neue Analysemöglichkeiten, die zu neuen Erkenntnissen im Bereich Erwerbsminderungsleistungen, Beschäftigung und Gesundheit führen.

Entsprechende Untersuchungen wären auch für den deutschen Arbeitsmarkt von großem Interesse, da die Verknüpfung von Daten aus verschiedenen Quellen ein wichtiger Schritt sein könnte, um ein umfassendes Bild von der Erwerbsminderung in Deutschland erhalten und die sozialpolitischen Maßnahmen entsprechend ausrichten zu können. Eine Übertragung dieser Dateninnovation würde im Bereich der Rentenversicherung wichtige Erkenntnisse zu den Wirkungen von Erwerbsminderungsrenten ermöglichen. Hierbei sind jedoch technische und rechtliche Fragen zu klären, die in der Art der Daten und ihrer Vorhaltung (u.a. in einer fragmentierten Landschaft von Sozialversicherungsträgern) sowie der geltenden Datenschutzbestimmungen (siehe Abschnitt 3.1) begründet sind. Das niederländische Beispiel könnte aber in mancher Hinsicht Lösungsmöglichkeiten aufzeigen und als Vorbild dienen.

#### 4.5.6 Quellen zum Steckbrief

Deutsche Rentenversicherung. (o. D.). *Erwerbsminderungsrenten*. Deutsche Rentenversicherung. Abgerufen am 16. März 2023 von [https://www.deutsche-rentenversicherung.de/DRV/DE/Rente/Allgemeine-Informationen/Rentenarten-und-Leistungen/Erwerbsminderungsrente/erwerbsminderungsrente\\_node.html](https://www.deutsche-rentenversicherung.de/DRV/DE/Rente/Allgemeine-Informationen/Rentenarten-und-Leistungen/Erwerbsminderungsrente/erwerbsminderungsrente_node.html)

Deutsche Rentenversicherung. (2022). *Erwerbsminderungsrenten im Zeitablauf 2022*. Statistik der Deutschen Rentenversicherung. [https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/statistikpublikationen/erwerbsminderungsrenten\\_zeitablauf.html](https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/statistikpublikationen/erwerbsminderungsrenten_zeitablauf.html)

FDZ Gesundheit, (o. D.). *Forschungsdatenzentrum Gesundheit*. Abgerufen am 11. August 2023 von <https://www.forschungsdatenzentrum-gesundheit.de>.

Koning, P., Muller, P. & Prudon, R. (2022a). Do disability benefits hinder work resumption after recovery?. *Journal of Health Economics*, 82(102593), 1-23.

<https://doi.org/10.1016/j.jhealeco.2022.102593>

Koning, P., Muller, P. & Prudon, R. (2022b). *Why Do Temporary Workers Have Higher Disability Insurance Risks Than Permanent Workers?*[Tinbergen Institute Discussion Paper 2022-024/V].

Tinbergen Institute Discussion Papers. <https://dx.doi.org/10.2139/ssrn.4067608>

Uitvoeringsinstituut Werknemersverzekeringen (2002). CAS – Classificaties voor Arbo en SV [CAS - Klassifizierungen für Arbeitsschutz und Sicherheit].

[https://www.steungroep.nl/images/her\\_keuring\\_WIA\\_of\\_WAO/Wetten\\_en\\_regels\\_bij\\_her\\_keuring/CAS\\_Classificaties\\_voor\\_Arbo\\_en\\_SV\\_UWV\\_2002.pdf](https://www.steungroep.nl/images/her_keuring_WIA_of_WAO/Wetten_en_regels_bij_her_keuring/CAS_Classificaties_voor_Arbo_en_SV_UWV_2002.pdf)

Uitvoeringsinstituut Werknemersverzekeringen (2014). *Ik ben ziek (WIA-uitkering) | UWV / Particulieren* [Ich bin krank (WIA-Leistung) | UWV | Einzelpersonen]. UWV. Veröffentlicht am 28. Juli 2014. Abgerufen am 16. März 2023 von [https://www.uwv.nl/particulieren/ziek/ziek-wia-](https://www.uwv.nl/particulieren/ziek/ziek-wia-uitkering/index.aspx)

[uitkering/index.aspx](https://www.uwv.nl/particulieren/ziek/ziek-wia-uitkering/index.aspx)

## 4.6 Verknüpfung von Stellenanzeigen und Bewerbungen aus einem Online-Stellenportal (Uruguay)

### 4.6.1 Einführung

Online-Stellenportale erleichtern das Matching auf dem Arbeitsmarkt und sind mittlerweile sowohl für Arbeitsuchende als auch für Unternehmen von großer Bedeutung. Auch für Politik und Verwaltungshandeln sind Daten aus Online-Stellenportalen informativ, da sie beispielsweise Aussagen zu Suchprozessen von Unternehmen und Personen sowie dem Matching auf dem Arbeitsmarkt ermöglichen. In der Forschung werden Textdaten aus Online-Stellenportalen seit wenigen Jahren unter Verwendung von Methoden des Natural Language Processing (NLP) ausgewertet, um die Nachfrage nach bestimmten Fähigkeiten zu beschreiben. Die Angebotsseite des Arbeitsmarkts wurde bisher allerdings kaum erforscht, da Daten mit detaillierten Informationen zu Arbeitsuchenden kaum verfügbar waren. Ein Großteil der vorliegenden Studien, die Online-Stellenanzeigen verwenden, bezieht sich außerdem auf die USA und auf Fähigkeiten zu ausgewählten Themenbereichen. Die im Folgenden dargestellte Studie von Bennett et al. (2022) trägt zur Schließung dieser Forschungslücken bei, indem Daten des privaten Stellenportals BuscoJobs aus Uruguay, die auf Stellenanzeigen von Unternehmen und Lebensläufen, die von Beschäftigten eingestellt werden, basieren, aufbereitet und ausgewertet werden.

Die Dateninnovation ist durch vier Elemente gekennzeichnet. Erstens ermöglichen die in den Daten enthaltenen Profile der auf dem Portal registrierten Personen die Analyse des Angebots von Fähigkeiten. Zweitens fällt auch die Analyse der Nachfrageseite umfassender aus als in der bisherigen Literatur. Drittens zeigt die Klassifizierung der nachgefragten bzw. angebotenen Fähigkeiten für ein Land außerhalb der USA auf, inwiefern die Ergänzung von US-amerikanischen durch länderspezifische Quellen wichtig ist. Viertens sind die verwendeten Daten insofern innovativ, als sie Informationen darüber enthalten, welche Personen sich auf welche ausgeschriebenen Stellen beworben haben. Dieser letztgenannte Aspekt wurde jedoch in der vorliegenden Studie noch nicht systematisch ausgewertet.

Auch für Politik und Verwaltungshandeln sind Daten aus Online-Stellenportalen informativ, weil dadurch stark nachgefragte Qualifikationen identifiziert und gezielt passende Weiterbildungen gefördert werden können. Weiteres Potenzial für die Forschung besteht im Hinblick auf die Analyse der Lohnentwicklung. Lohnänderungen können durch die veränderliche Nachfrage nach Fähigkeiten erklärt werden, wie Deming & Kahn (2018) in einem einflussreichen Forschungspapier für die USA zeigen. Deming & Noray (2020) kommen außerdem zu dem Schluss, dass das Lohnwachstum in Berufen, die sich schnell verändern, geringer ist als in anderen Berufen, da erlerntes Wissen schnell veraltet. Diese Ergebnisse sind allerdings nicht uneingeschränkt auf Arbeitsmärkte in anderen Ländern übertragbar, da diese beispielsweise eine andere Branchen- und Qualifikationsstruktur aufweisen. Zur länderspezifischen Analyse solcher Entwicklungen ist daher die Erstellung einer länderspezifischen Taxonomie von Fähigkeiten erforderlich, wie sie Bennett et al. (2022) für Uruguay vorlegen.

### 4.6.2 Technische Beschreibung

#### 4.6.2.1 Daten des Online-Stellenportals BuscoJobs

Grundlage für die Studie von Bennett et al. (2022) ist die Erschließung und Beschreibung des vom privatwirtschaftlichen uruguayischen Stellenportal BuscoJobs bereitgestellten Datensatzes. Hierbei handelt es sich um das größte Portal dieser Art in Uruguay, das sowohl Stellenanzeigen von Unternehmen als auch Profile von Stellensuchenden umfasst und damit sowohl die Nachfrage- als

auch die Angebotsseite des Arbeitsmarktes abdeckt. Die Daten stehen für die Jahre 2010 bis 2020 zur Verfügung.

Die Vakanzdaten der Unternehmen in BuscoJobs decken nach Angaben von Bennett et al. (2022) ca. 50 bis 60 Prozent der Online-Stellenanzeigen in Uruguay ab und enthalten im Gegensatz zu z. B. von Webseiten gescrapten Anzeigedaten keine Duplikate, was die Repräsentativität der Daten verbessert. Die verwendeten Inhalte der Stellenanzeigen umfassen neben dem Namen und dem Standort des Unternehmens Freitextbeschreibungen zur konkreten Stelle mit Anforderungen an die Bewerber\*innen wie Qualifikationen, Arbeitserfahrungen und Fähigkeiten. Außerdem werden typischerweise die in der ausgeschriebenen Stelle zu erbringenden Tätigkeiten beschrieben und Angaben zum möglichen Gehalt gemacht. Anhand einer Verknüpfung mit administrativen Datenquellen mittels eindeutiger Identifikationsnummern konnten dem Datensatz außerdem zuverlässige 4-stellige Wirtschaftszweiginformationen zugefügt werden. Zusätzlich werden die Stellenanzeigen anhand der Textinformationen mittels Machine-Learning-Verfahren in eine Berufsklassifikation auf der 2-Steller-Ebene eingeordnet (zu Verfahren der Gewinnung von standardisierten Informationen aus Fließtexten vergleiche Steckbrief 5.7). Die Struktur dieses Teils des Datensatzes ähnelt der Datenstruktur der Stellenanzeigendaten aus der BA-Jobbörse, die von Stops et al. (2021) im Rahmen der Machbarkeitsstudie Kompetenz-Kompass beschrieben werden.

Die uruguayischen Daten enthalten zusätzlich jedoch individuelle Profile mit selbst verfassten Lebensläufen einschließlich früherer und aktuell ausgeübter Beschäftigungen, vorhandenen Qualifikationen und Fähigkeiten der Stellensuchenden. Neben allgemeinen Angaben zu individuellen Fähigkeiten ist auch eine Angabe von bei bestimmten Beschäftigungsverhältnissen erlernten Fähigkeiten möglich. Außerdem können die Stellensuchenden einen Soft-Skills-Test auf dem Portal absolvieren und die Ergebnisse auf ihrer Profilseite anzeigen lassen. Äquivalent zur Vorgehensweise bei den Stellenanzeigen ordnen Bennett et al. (2022) auch den Beschäftigungsverhältnissen der Profilseiten mittels Machine-Learning-Verfahren Berufscodes auf der 2-Steller-Ebene zu.

Die Stellenanzeigen und Profile auf BuscoJobs sind auf zwei Arten miteinander verknüpft. Zum einen können sich Unternehmen und arbeitssuchende Personen gegenseitig finden und kontaktieren, indem sie nach bestimmten Merkmalen von Stellen bzw. Profilen suchen. Zum anderen können sich die registrierten Personen auch direkt über das Portal auf dort inserierte Vakanzen bewerben. Insgesamt beinhaltet der Datensatz Informationen zu ca. 667.000 Einzelpersonen, von denen etwa 388.000 Angaben zu ungefähr 1,2 Mio. aktuellen oder früheren Beschäftigungsverhältnissen machen. Dem stehen ca. 87.000 Stellenanzeigen von mehr als 6.000 Unternehmen gegenüber.

Ein Vergleich des Analysedatensatzes mit einer repräsentativen Haushaltsbefragung zeigt, dass der Datensatz des Stellenportals bis auf den Landwirtschafts- und den Militärssektor alle Wirtschaftszweige umfasst. Die auf anderen Stellenportalen vorhandenen Selektionseffekte, beispielsweise im Hinblick auf das Bildungsniveau (Napierała et al., 2022), fallen relativ gering aus. Einwohner\*innen der Hauptstadt Montevideo und junge Menschen sind stärker überrepräsentiert, die Verzerrung bleibt in den Daten jedoch über die Zeit stabil.

Insgesamt erscheint der Datensatz aufgrund der Größe, der Detailinformationen zu Stellenanzeigen und individuellen Profilen sowie der Möglichkeit zu deren Verknüpfung und zuletzt aufgrund der nicht zu starken Selektivität gut geeignet, um Fähigkeitsprofile auf der Nachfrage- und der Angebotsseite des Arbeitsmarktes zu erstellen.

#### 4.6.2.2 Klassifikation von Fähigkeiten

Der Hauptbeitrag des Papiers von Bennett et al. (2022) ist die Erstellung einer besonders umfassenden Fähigkeiten-Taxonomie für Uruguay und die Anwendung der Taxonomie auf die

Angebotsseite des Arbeitsmarkts. Dabei wird zudem erläutert, welche Bedeutung US-amerikanische und landesspezifische Quellen für die Klassifizierung haben. Die Analyse ermöglicht die Darstellung der resultierenden Anforderungs- und Fähigkeitsprofile von Vakanzen bzw. Personen. Die Autor\*innen argumentieren, dass die jeweiligen Arbeitsmarktstrukturen, aber auch die Sprache des jeweils betrachteten Landes eine länderspezifische Anpassung der Fähigkeiten-Klassifikation erforderlich machen. Die von Bennett et al. (2022) erstellte Fähigkeiten-Taxonomie ist, basierend auf früheren wissenschaftlichen Studien, in drei Oberkategorien gegliedert: kognitive, sozio-emotionale und manuelle Fähigkeiten. Diese Oberkategorien werden in insgesamt 14 Unterkategorien aufgeschlüsselt. Für jede der Unterkategorien wird anschließend eine Liste von eindeutigen Schlüsselwörtern erstellt, d.h. es gibt keine Schlüsselwörter, die in mehreren Unterkategorien auftauchen. Insgesamt umfassen die 14 Unterkategorien 152 Schlüsselbegriffe, die in weiten Teilen auf Wortlisten bisheriger US-amerikanischer und europäischer Literatur beruhen. Zusätzlich werden Informationen aus dem kürzlich initiierten Projekt O\*NET Uruguay ergänzt, das entsprechend dem US-amerikanischen Pendant Anforderungen verschiedener Berufe beschreibt.<sup>21</sup> Werden zusätzlich Synonyme der aufgeführten Schlüsselwörter ergänzt, umfasst die Taxonomie 800 Begriffe.

Wie bei NLP-Projekten üblich, werden nach der Übersetzung der Schlüsselbegriffe (hier: ins Spanische) diese Begriffe sowie die analysierten Textdaten in mehreren Schritten für die anschließende Analyse vorbereitet. Bei der „Normalisierung“ werden Zeichen wie Großbuchstaben und Akzente durch normale Kleinbuchstaben ersetzt sowie weitere Sonder- und Satzzeichen, sogenannte Stopp-Worte, wie Artikel oder Präpositionen, und weitere häufige Wörter ohne Bedeutung für die Fähigkeiten, wie Zeit- und Ortsangaben, entfernt. Mit dem Verfahren des sogenannten Stemming werden die verbleibenden Wörter dann auf den jeweiligen Wortstamm reduziert. Bei der Tokenisierung wird anschließend berücksichtigt, dass bestimmte Wortkombinationen besondere Bedeutungen haben, beispielsweise „Vorhersage von Daten“. Die Wörter werden dabei sowohl einzeln als auch als Kombination aufgenommen. Der Großteil der Datenbereinigungsschritte wird dabei mit Machine-Learning-Methoden des Natural Language Processing (NLP) durchgeführt. An einzelnen Stellen werden jedoch manuelle Überprüfungen bzw. Ergänzungen durchgeführt.

Im entscheidenden Schritt der Analyse werden aus diesen vorbereiteten Textdaten mithilfe der Schlüsselwörter die in den Stellenanzeigen und individuellen Profilseiten beschriebenen Fähigkeiten extrahiert. Für jede Fähigkeits-Unterkategorie wird eine binäre Variable erstellt, die den Wert Eins annimmt, sofern mindestens ein Schlüsselwort aus der entsprechenden Unterkategorie im untersuchten Text enthalten ist. Die resultierenden Anforderungsprofile der Stellen bzw. Fähigkeitsprofile der Personen zeigen somit kategorisiert auf, welche Fähigkeiten von den Unternehmen nachgefragt werden bzw. bei den Arbeitskräften vorhanden sind. Neben einfachen Indikatorvariablen für jede der 14 Unterkategorien von Fähigkeiten werden zusätzlich auch Variablen erstellt, die die Häufigkeit von unterschiedlichen Fähigkeiten (bzw. passender Schlüsselwörter) innerhalb eines Textes pro Unterkategorie abbilden. Diese Häufigkeiten werden als Proxy-Größen für die Intensität der nachgefragten bzw. angebotenen Fähigkeiten interpretiert. Außerdem gibt eine Gruppe zusätzlicher Variablen an, wie hoch der Anteil von Fähigkeiten innerhalb einer Unterkategorie ist, die leicht automatisiert werden können und daher in Zukunft voraussichtlich weniger stark nachgefragt werden. Diese Fähigkeiten werden entsprechend der beschriebenen Einteilung für die drei Oberkategorien anhand von Schlüsselbegriffen aus der Literatur klassifiziert.

---

<sup>21</sup> Da das Projekt O\*NET Uruguay noch nicht vollständig entwickelt und bisher unveröffentlicht ist, sind darüber nur wenige Informationen vorhanden. Das durchführende Ministerio de Trabajo y Seguridad Social (Ministerium für Arbeit und Soziale Sicherung) veröffentlichte im Mai 2022 lediglich einen Beitrag, der die Ziele des Projekts und den aktuellen Stand beschreibt (Ministerio de Trabajo y Seguridad Social, 2022).

Insgesamt kann der Großteil der Stellenausschreibungen bzw. Beschäftigungsverhältnisse aus den Lebensläufen einer oder mehrerer Fähigkeits-Unterkategorien zugeordnet werden. Ohne Berücksichtigung von Synonymen der Schlüsselbegriffe liegen die jeweiligen Anteile bei 86 bzw. 53 Prozent. Unter Verwendungen von Synonymen steigen diese Anteile auf 94 bzw. 64 Prozent an. Werden die Fähigkeiten einer Person über all ihre beschriebenen Beschäftigungsverhältnisse aggregiert, beträgt der Anteil der Personen, denen mindestens eine Fähigkeits-Unterkategorie zugeordnet werden kann, 75 Prozent (in der Variante inkl. Synonyme).

#### 4.6.3 Anwendung

Die Anwendung der beschriebenen Methode auf die Daten des uruguayischen Stellenportals ist für den deutschen Kontext insbesondere im Hinblick auf allgemeine Erkenntnisse aus der Analyse von Stellenanzeigen außerhalb der USA relevant. Eine Erkenntnis der Studie ist, dass die Einbeziehung von Schlüsselwörtern aus verschiedenen Quellen wichtig ist für die Erstellung der Taxonomie und die anschließende Kategorisierung der beruflichen Anforderungen und Fähigkeiten. 61 Prozent der Fähigkeiten in den Personenprofilen bzw. 72 Prozent in den Vakanzdaten wurden anhand von Schlüsselwörtern aus der Literatur zu Online-Stellenanzeigen identifiziert. Weitere, sukzessive hinzugezogene Literaturquellen sowie O\*NET Uruguay wurden mit Anteilen von 9 bis 31 Prozent ebenfalls häufig genutzt.

Die Bedeutung länderspezifischer Quellen wurde durch eine Vergleichsanalyse unterstrichen, bei der die Autor\*innen Informationen aus der US-amerikanischen Datenbank O\*NET auf Berufsebene verwendeten. Die Ergebnisse dieser Analyse unterscheiden sich stark von den mittels der BuscoJobs-Daten für Uruguay gewonnenen Erkenntnissen. Dies zeigt, dass Daten, die sich auf die USA beziehen, nicht zur Klassifikation von Berufen in anderen Ländern verwendet werden sollten. Die Erfahrung der Autor\*innen mit der Textanalyse durch NLP-Verfahren zeigt außerdem, dass eine Arbeit in der jeweiligen Sprache zu präziseren Ergebnissen führt als beispielsweise Daten privater Anbieter von Online-Stellenanzeigen wie Lightcast, deren Taxonomien auf Übersetzungen aus dem Englischen beruhen.

Im Vergleich zu in Deutschland erstellten Taxonomien wie dem IAB-Kompetenz-Kompass (s. u.) verwenden Bennett et. al (2022) außerdem eine breitere Kategorisierung der Fähigkeiten. Dies vereinfacht den Vergleich der Profile auf der Angebots- und der Nachfrageseite des Arbeitsmarktes, welcher – laut den Autor\*innen – durch eine begrenzte Zahl systematisch aggregierter Kategorien erleichtert wird. Der Nachteil dieser breiteren Kategorisierung ist jedoch, dass sich dieses möglicherweise negativ auf die Qualität der Matches zwischen Arbeitsuchenden und Unternehmen auswirkt.

#### 4.6.4 Übertragbarkeit auf Deutschland

Grundsätzlich stehen auch in Deutschland bereits Datensätze zur Verfügung, die die Angebotsseite des Arbeitsmarktes, also die Bewerbungen, analysieren. Beispielsweise verwendeten Bauer et al. (2020) Daten des Online-Karrierenetzwerks LinkedIn, um das Bewerbungsverhalten von Stellensuchenden in den ersten Monaten des Corona-Lockdowns im Jahr 2020 zu untersuchen. Darüber hinaus gibt es weitere Stellenportale mit Informationen zu Lebensläufen, die zur Analyse der Angebotsseite genutzt werden könnten. Auch Daten aus Beratungen der Agenturen für Arbeit oder der Jobcenter erscheinen für Analysen der Fähigkeiten auf der Angebotsseite grundsätzlich möglich. Allerdings werden dort Daten aus Bewerbungen, die Arbeitsuchende versenden, bisher nicht systematisch erfasst bzw. aufbereitet.

Eine Taxonomie von Fähigkeiten in Bezug auf die Nachfrageseite, also die Stellenangebote, wurde in Deutschland mit dem IAB-Kompetenz-Kompass bereits entwickelt (Stops et al., 2021). Somit sind

auch für Deutschland detaillierte Analysen zur Personalsuche der Unternehmen möglich und erfolgen bereits. Allerdings stellt die Repräsentativität der Daten – wie bei vielen Online-Stellenportalen – eine Herausforderung dar (Rengers, 2018). Zudem existiert kein aufbereiteter Analysedatensatz, der Angebots- und Nachfrageseite miteinander verknüpft, wie dies bei BuscoJobs möglich ist. Durch eine Übertragung des Vorgehens von BuscoJobs einschließlich der Informationen zu Bewerbungen wären auch für Deutschland tiefergehende Analysen zum Such- und Bewerbungsverhalten von Unternehmen bzw. Stellensuchenden und dem Matching auf dem Arbeitsmarkt möglich. Die Übertragung dieses innovativen Aspekts scheint allerdings schwierig, da aufgrund von Datenschutzbestimmungen die Informationen zum Bewerbungsverhalten auf individueller Ebene ohne explizite Zustimmung der Betroffenen nicht genutzt werden kann (vgl. Abschnitt 3.1).

#### 4.6.5 Fazit

Die Verwendung der Daten von Stellenportalen bietet erhebliche Analysemöglichkeiten. In der wissenschaftlichen Literatur werden Informationen aus Stellenausschreibungen, also der Nachfrageseite des Arbeitsmarktes, schon seit einigen Jahren ausgewertet. Die Analysen beziehen sich jedoch auf einige wenige Länder. Die vorgestellte Dateninnovation aus Uruguay ermöglicht es zusätzlich, die Angebotsseite sowie das Zusammenspiel von Angebots- und Nachfrageseite näher zu untersuchen. Sie beruht auf einer Klassifizierung von Fähigkeiten, die aus den Lebensläufen hervorgehen, die Stellensuchende auf dem Stellenportal hinterlegen. Zudem soll zukünftig genauer analysiert werden, auf welche Stellen sich diese Personen bewerben und wie erfolgreich sie dabei sind.

In Deutschland wird die Nachfrageseite bereits eingehend durch den IAB-Kompetenz-Kompass untersucht. Die Daten für eine Analyse der Angebotsseite, d.h. zum Bewerbungsverhalten Arbeitssuchender, stehen in Daten privater Unternehmen wie LinkedIn oder Stepstone ebenfalls grundsätzlich zur Verfügung, sodass eine Übertragung der Dateninnovation auf Deutschland generell möglich erscheint. Entsprechende Daten öffentlicher Behörden werden bislang nicht umfassend erhoben, da der Datenschutz der Nutzung der Daten von Arbeitssuchenden Grenzen setzt.

#### 4.6.6 Quellen zum Steckbrief

Bauer, A., Keveloh, K., Mamertino, M. & Weber, E. (2020). Competing for jobs: How COVID-19 changes search behaviour in the labour market (*IAB Discussion Paper 33/2020*). *IAB Discussion Paper*. Abgerufen am 14. März 2023 von <https://doku.iab.de/discussionpapers/2020/dp3320.pdf>

Bennett, F., Escudero, V., Liepmann, H., & Podjanin, A. (2022). Using Online Vacancy and Job Applicants' Data to Study Skills Dynamics (*ILO Working Paper 75*). *ILO Working Paper*. Abgerufen am 14. März 2023 von [https://www.ilo.org/global/publications/working-papers/WCMS\\_853821/lang--en/index.htm](https://www.ilo.org/global/publications/working-papers/WCMS_853821/lang--en/index.htm)

Deming, D. & Kahn, L. B. (2018). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. *Journal of Labor Economics*, 36(S1), 337-369. <https://doi.org/10.1086/694106>

Deming, D. J. & Noray, K. (2020). Earnings Dynamics, Changing Job Skills, and STEM Careers. *The Quarterly Journal of Economics*, 135(4), 1965–2005. <https://doi.org/10.1093/qje/qjaa021>

Ministerio de Trabajo y Seguridad Social. (2022). *O\*NET Uruguay: El Banco Mundial y el MTSS desarrollaron herramienta para mejorar políticas públicas [O\*NET Uruguay: Die Weltbank und die MTSS entwickelten ein Instrument zur Verbesserung der öffentlichen Politik]*. Gub. Abgerufen am 14.

März 2023 von <https://www.gub.uy/ministerio-trabajo-seguridad-social/comunicacion/noticias/onet-uruguay>

Napierała, J., Kvetan, V. & Branka, J. (2022). *Assessing the representativeness of online job advertisements* (Working Paper No 17 / December 2022). Cedefop Working paper series. <https://doi.org/10.2801/807500>

Rengers, M. (2018). Internetgestützte Erfassung offener Stellen. *WISTA – Wirtschaft und Statistik*, 5(18), 11–33.

Stops, M., Bächmann, A.-C., Glassner, R., Janser, M., Matthes, B., Metzger, L.-J., Müller, C., Seitz, J. (2021). Extracting Skill Requirements from Job Ads – the “Machbarkeitsstudie Kompetenz-Kompass”. IAB-Forschungsbericht 07/2021, Nürnberg.

## 4.7 Synthetische Datengeneratoren (Vereinigtes Königreich)

### 4.7.1 Einführung

Viele Institutionen aus dem Bereich der amtlichen Statistik, der öffentlichen Verwaltung, der Unternehmen und der Forschungseinrichtungen erheben Datensätze mit personenbezogenen Informationen oder anderweitig geschützten Inhalten, bereiten diese auf und speichern sie. Aufgrund von datenschutzrechtlichen oder anderen gesetzlichen Auflagen können diese Datensätze meistens nicht zu Forschungszwecken genutzt werden. Selbst wenn eine Nutzung für die Forschung zulässig ist, ist sie oft nur unter strikten Auflagen möglich, beispielsweise, dass nur vergrößerte Informationen verwendet werden oder dass nur an bestimmten Orten wie den Forschungsdatenzentren mit den Daten gearbeitet wird. Mithilfe von synthetischen Datengeneratoren (SDG) kann die Nutzung von geschützten Daten vereinfacht werden. Durch die SDG werden synthetische Daten generiert, welche den Informationsgehalt der echten Daten weitestmöglich wahren, aber die Privatsphäre der in den Daten enthaltenen Personen im Vergleich zum Ursprungsdatensatz besser schützen.

Am Alan Turing Institute in London beschäftigt sich eine Gruppe von Wissenschaftler\*innen mit synthetischen Daten und organisiert einen Erfahrungsaustausch zur Bereitstellung und Nutzung über unterschiedliche Anwendungsbereiche hinweg. Die Generierung synthetischer Daten ist schon seit den 1990er Jahren Gegenstand der statistischen Forschung.<sup>22</sup> Es wurden unterschiedliche Vorgehensweisen entwickelt, mit denen synthetische Daten erzeugt werden können. Durch die Entwicklung immer komplexerer Algorithmen, beispielsweise im Rahmen von Machine-Learning-Methoden, gewann die Generierung von synthetischen Daten in den letzten Jahren vor allem in der Informatik an Bedeutung. Das Projekt „Synthetic data and privacy preservation“ (Szpruch et al., o. D.) des Alan Turing Institute erforscht die Potenziale synthetischer Daten, indem es konventionelle Vorgehensweisen mit diesen neuen Ansätzen verbindet.

### 4.7.2 Technische Beschreibung

Modelle zur Synthetisierung von Daten können unterschiedliche Verfahren nutzen (Jordon et al., 2022). Üblicherweise werden für die Synthetisierung agentenbasierte oder ökonometrische Modelle (Bonabeau, 2002) verwendet. Neuere Herangehensweisen basieren auf Deep Generative Models. Dabei handelt es sich um Machine-Learning-Methoden, die mithilfe von neuronalen Netzwerken hochdimensionale Wahrscheinlichkeitsverteilungen schätzen. Zwei der am häufigsten verwendeten Algorithmen sind Generative Adversarial Networks (GANs, Goodfellow et al., 2020) und der Variational Autoencoder (VAE, Kingma & Welling, 2013).

Das Projekt des Alan Turing Institute kombiniert konventionelle Modelle zur Generierung synthetischer Daten mit solchen Deep Generative Models. Damit kann der Datengenerierungsprozess der echten Daten (zum Beispiel Querschnitts- oder Zeitreihendaten) genau geschätzt werden. Unabhängig von der Funktion und dem Zweck umfasst ein System zur Generierung synthetischer Daten die Aufbereitung der Daten, die Entwicklung von Methoden zur Synthetisierung und einen Feedback-Mechanismus, dessen Aufgabe die Überwachung des gesamten Prozesses ist und der beispielsweise anzeigt, ob die Verteilung der synthetischen Daten die Verteilung der echten Daten widerspiegelt und ob die Vertraulichkeit personenbezogener Daten sichergestellt werden kann.

Viele Methoden zur Generierung synthetischer Daten können durch Verfahren der Optimierung und der Wahl von Modellparametern so verschlüsselt werden, dass der Rückschluss von den erzeugten

---

<sup>22</sup> Für eine Übersicht verschiedener aktueller Fragestellungen im Bereich von SDGs, siehe Jordon et al. (2022).

synthetischen Daten auf die echten Daten nicht mehr möglich ist. Da keine direkte Verknüpfung zwischen den Beobachtungen in den synthetischen und den echten Daten möglich ist, wird die Aufdeckung geschützter Informationen verhindert.

Dennoch muss bei der Erzeugung und Bereitstellung synthetischer Daten mit großer Vorsicht vorgegangen werden, da Rückschlüsse auf die Originaldaten nicht automatisch auszuschließen sind (Jordon et al., 2022). Zum einen besteht die Gefahr, dass einzelne Personen mit ihren Merkmalen trotz des Wegfalls der direkten Zuordnung zwischen synthetischen und echten Daten direkt identifiziert werden können. Dieses Risiko ist umso eher gegeben, je mehr Informationen die Daten enthalten. Zum anderen könnte aus den synthetischen Daten die Information extrahiert werden, ob eine Person bzw. Institution in den ursprünglichen Daten vorhanden ist, auch wenn die Merkmale der Personen nicht direkt identifizierbar sind (Risiko eines sogenannten Membership Attack). Dies ist unter dem Aspekt des Datenschutzes relevant, da das Vorhandensein einer Person in sensiblen Daten bereits Informationen über die Person enthält. So lässt die Tatsache, dass eine Person in den Arbeitssuchendendaten der BA enthalten ist, bereits Rückschlüsse auf die Person zu.

Synthetische Daten werden bisher hauptsächlich dazu genutzt, um Vorhersagemodelle zu schätzen beziehungsweise deskriptive Zusammenhänge zu analysieren. Da die Identifikation von kausalen Zusammenhängen auf Annahmen über den datengenerierenden Prozess beruht, ist weitere Forschung erforderlich, um festzustellen, inwieweit synthetische Daten für diesen Zweck verwendet werden können (Jordon et al., 2022).

#### 4.7.3 Anwendung

Die vom Alan Turing Institute entwickelten Methoden könnten künftig die Datenversorgung für eine Vielzahl von wissenschaftlichen Disziplinen und Fragestellungen verbessern. Sie ermöglichen Forschenden einen einfacheren Zugang zu Daten und geben gleichzeitig den für die Daten Verantwortlichen und den in den Daten enthaltenen Personen und Unternehmen Sicherheit in Bezug auf den Datenschutz. Des Weiteren bieten synthetische Daten die Möglichkeit, Informationen aus unterschiedlichen Datensätzen zu verknüpfen und systematische Fehler in den Originaldaten wie beispielsweise Messfehler zu beheben.

Die Kooperation des Alan Turing Institute mit dem Office for National Statistics (ONS) lässt für die Zukunft eine Reihe von Projekten mit Bezug zum öffentlichen Sektor erwarten. Der Schwerpunkt des Instituts liegt derzeit noch eher auf anderen Themenfeldern als der Arbeitsmarkt- und Sozialpolitik. Eine konkrete Anwendung ist die Simulation von Daten im Gebiet der Finanzwirtschaft, beispielsweise zur Analyse der Bereitschaft von Bankkunden, Produkte wie Hypotheken oder ein Darlehen zu erwerben.

#### 4.7.4 Übertragbarkeit auf Deutschland

Die Synthetisierung von Daten ist in Deutschland schon seit vielen Jahren ein Thema, mit dem sich Expert\*innen beschäftigen. In ihrem Gutachten empfiehlt die Datenethikkommission der Bundesregierung, Forschung im Bereich synthetischer Daten zu fördern und identifiziert unter anderem großen Forschungsbedarf zu der Frage, inwieweit und in welchen Kontexten synthetische Daten die Verarbeitung echter Daten ersetzen können und wie eng die synthetischen Daten an die Eigenschaften von echten Daten angelehnt sein sollen (DEK, 2019). Im Folgenden werden einige wichtige in Deutschland aktuell durchgeführte oder bereits abgeschlossene Projekte zu synthetischen Daten dargestellt, die im Rahmen von Expertengesprächen diskutiert wurden.

Das Statistische Bundesamt erprobt die Synthetisierung von Daten in mehreren Projekten. Einige haben das Ziel, Forschenden bisher noch nicht verfügbare Daten zugänglich zu machen.

Beispielsweise könnten durch die Synthetisierung den Forschenden georeferenzierte Daten auf einer Ebene von bis zu 100x100m-Rastern zugänglich gemacht werden. Ferner beschäftigt sich das Statistische Bundesamt mit der grundsätzlichen Frage, inwiefern die Synthetisierung dabei hilft, ein frei verfügbares Datenangebot zu schaffen. Ein solches Angebot ist zwar in Form von Public Use Files oft schon vorhanden, aber das Analysepotenzial der Public Use Files wird dadurch begrenzt, dass bestimmte Variablen wie beispielsweise Regionalinformationen – bedingt durch die Anonymisierung – nur in klassierter und stark vergrößerter Form verfügbar sind.

Die Synthetisierung von Daten wird am Statistischen Bundesamt auch im Rahmen der Veröffentlichung von aggregierten Daten erprobt. Um die Anonymität der Beobachtungen zu gewährleisten, wendet das Statistische Bundesamt bisher primäre und sekundäre Zellsperren an. Mithilfe von synthetischen Daten können Zellsperren vermieden werden. In diesem Zusammenhang wendet das Statistische Bundesamt auch aktuell schon synthetische Daten an: Die Ergebnisse des Zensus 2011 und des Zensus 2022 werden unter Verwendung synthetisierender Methoden in aggregierter Form publiziert.

Am Forschungsdatenzentrum des Bundesinstituts für Arzneimittel und Medizinprodukte wird zurzeit geprüft, inwieweit Abrechnungsdaten des Bundes der Krankenkassen synthetisiert werden können (Schneider, 2022). Dies soll Forschenden unter Einhaltung der Datenschutzvorschriften einen Zugang zu hochsensiblen Gesundheitsdaten gewähren. Hierbei sollen ähnlich wie bei den Verfahren, die das Alan Turing Institute entwickelt, Machine-Learning-Algorithmen zum Einsatz kommen.

Im Jahr 2021 entwickelte die Bundesregierung auf Anregung des Wissenschaftlichen Beirats beim Bundesministerium der Finanzen Pläne zur Gründung des Instituts für empirische Steuerforschung (IfeS). Das Institut soll unter anderem durch die Erstellung synthetischer Datensätze die Dateninfrastruktur für die Wissenschaft im Bereich Steuern verbessern (Deutscher Bundestag, 2021, Dörrenberg & Peichl, 2021).<sup>23</sup>

Mehrheitlich stehen diese Projekte noch am Anfang und untersuchen die Grundlagen der Generierung synthetischer Daten. Die Synthetisierung von Daten wird von Expert\*innen als sehr zeitaufwändig eingeschätzt. Dies zeigt auch die Erfahrung am Forschungsdatenzentrum des IAB. Dort wurde in einem Testversuch die erste Welle des IAB-Betriebspanels, der größten Panelbefragung von deutschen Betrieben, synthetisiert (Drechsler, 2011). Für die Synthetisierung war ein hoher Arbeitsaufwand erforderlich. Da die Daten über das IAB-Forschungsdatenzentrum bereits für die Forschung zugänglich waren, war die Nachfrage zudem gering. Aus diesen Gründen wurde das Projekt eingestellt.

Der hohe Zeitaufwand der Synthetisierung resultiert daraus, dass zwei Anforderungen erfüllt werden müssen. Einerseits muss gewährleistet werden, dass keine Rückschlüsse auf einzelne Merkmalsträger mehr möglich sind. Eine solche Anonymisierung ist in manchen Fällen, beispielsweise in Unternehmensbefragungen im Fall von großen Unternehmen, nicht durchführbar. Andererseits müssen der Aufbau und die Logik der ursprünglichen Daten erhalten werden. Durch diese komplexen Anforderungen sind routinemäßige Synthetisierungen einzelner Datensätze nicht möglich, sondern es muss bei jedem Datensatz unterschiedlich vorgegangen werden. Zusätzlich betonen Expert\*innen, dass die IT-Infrastruktur bisher nicht den Anforderungen der für die Synthetisierung angewandten rechenintensiven Algorithmen genügt.

Synthetische Daten haben insbesondere in Bereichen einen Mehrwert, in denen mit sensiblen Merkmalen gearbeitet wird, die aus Datenschutzgründen aktuell nicht für Forschende zugänglich

---

<sup>23</sup> Da das IfeS zum Zeitpunkt der Verfassung dieses Berichts noch nicht gegründet wurde, liegen inhaltliche Ergebnisse bisher nicht vor.

sind. Dies könnte insbesondere im Bereich der Gesundheits- und Steuerdaten zu wichtigen Verbesserungen der Informationsbasis führen. Synthetische Daten können aber auch in Bereichen nützlich sein, in denen bereits jetzt ein guter Zugang zu Forschungsdaten besteht. Gerade in der Arbeitsmarkt- und Sozialpolitik sind schon viele Datensätze, wie die Daten der BA, der Rentenversicherung oder des Statistischen Bundesamts, für Forschende zugänglich. Diese Daten könnten durch eine Synthetisierung einem breiteren Nutzer\*innenkreis zugänglich gemacht werden. Außerdem könnten durch die Synthetisierung und Verknüpfung von Daten neue Datenprodukte entstehen. Vorstellbar wäre die Verbindung von Daten wie z. B. den Integrierten Erwerbsbiografien (IEB) des IAB mit kleinräumigen Informationen oder die Kombination von Daten unterschiedlicher Institutionen wie dem IAB, der Bundesbank oder dem Statistischen Bundesamt.

Wegweisend hierfür könnten die Erfahrungen sein, die derzeit im Projekt MikroSim gewonnen werden, das von der Deutschen Forschungsgemeinschaft (DFG) gefördert wird. Die Generierung synthetischer Datensätze als Grundlage für Mikrosimulationsmodelle ist notwendig, weil die einzelnen Datensätze jeweils für sich für die Simulation unzureichend sind. Im Projekt MikroSim werden Informationen aus unterschiedlichen Datenquellen – Daten des Zensus, weitere Daten der amtlichen Statistik, Befragungsdaten aus der empirischen Sozialforschung – verknüpft, um auf dieser Basis ein kleinräumiges Simulationsmodell der Bevölkerung der Bundesrepublik Deutschland auf der Ebene von Personen und Haushalten zu entwickeln (MIKROSIM, o. D.). Mithilfe des Simulationsmodells können die Konsequenzen verschiedener sozialpolitischer und demografischer Entwicklungsszenarien auf kommunaler und regionaler Ebene abgeschätzt werden. Hierzu wird eine synthetische Bevölkerungsdatenbank aus bedingten Verteilungen erstellt, die auf Basis unterschiedlicher Erhebungen geschätzt werden. Der Datenschutz spielt in diesem Projekt eine große Rolle, ist jedoch in diesem Zusammenhang einfacher einzuhalten als für die Synthetisierung bestimmter Datensätze. Der Datensatz wird rekursiv generiert und enthält stochastische Elemente. Somit gibt es in diesem Zusammenhang keine echten Beobachtungen und Analysen werden auf einer Pseudo-Grundgesamtheit durchgeführt. Seltene Ereignisse oder Ausreißer können allerdings nicht abgebildet werden.

#### 4.7.5 Fazit

Synthetische Datengeneratoren bzw. synthetische Daten sind ein Forschungsfeld, das sich aktuell in einem anfänglichen Stadium befindet, aber hinsichtlich des Nutzens für die Forschung vielversprechende Potenziale aufweist. Das Forschungsprojekt des Alan Turing Institute trägt zur Grundlagenforschung in diesem Gebiet bei.

Die Synthetisierung von Daten ist vor allem sinnvoll, wenn dadurch der Zugang zu den Daten wesentlich vereinfacht und das Nutzungspotenzial, beispielsweise durch neue Datenprodukte, erheblich erweitert werden kann. In Deutschland ist der Bedarf an zusätzlichen Forschungsdaten, z. B. im Bereich der Steuer- und Gesundheitsdaten, hoch. Zudem ist die Erstellung von synthetischen Daten sinnvoll, wenn zur Beantwortung gesellschaftsrelevanter Fragen viele Informationen benötigt werden, die in keinem einzelnen Datensatz enthalten sind. Synthetische Daten können hier zur Verknüpfung bereits bestehender Datensätze genutzt werden.

Es gibt in Deutschland bereits Erfahrungen aus einer Reihe von Projekten, in denen die Anwendung von synthetischen Daten geprüft wird. Sie zeigen einerseits, dass die Synthetisierung von Daten mit einem hohen zeitlichen Aufwand und hohen notwendigen Rechenleistungen verbunden ist. Andererseits sind die derzeit durchgeführten Projekte ein Beleg dafür, dass die zur Entwicklung synthetischer Datengeneratoren notwendigen Methoden und Kompetenzen vorhanden sind.

#### 4.7.6 Quellen des Steckbriefs

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99 (3), 7280-7287.

Datenethikkommission (DEK) (2019). Gutachten der Datenethikkommission. Bundesministerium des Inneren und der Heimat.

Deutscher Bundestag (2021). Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Stephan Brandner, Dr. Bruno Hollnagel, Kay Gottschalk, weiterer Abgeordneter und der Fraktion der AfD, Drucksache 19/31668

Drechsler, J. (2011). *Methodenreport: Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels*, FDZ-Methodenreport 01/2011, Forschungsdatenzentrum der Bundesagentur für Arbeit am Institut für Arbeitsmarkt- und Berufsforschung.

Dörrenberg, P. & Peichl, A. (2021). Was gute Steuerpolitik braucht. Süddeutsche Zeitung. Abgerufen am 16. Mai 2023 von <https://www.sueddeutsche.de/wirtschaft/steuern-bundestagswahl-folgen-1.5408922>.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N. & Weller, A. (2022). Synthetic Data – what, why and how? *arXiv*. <https://doi.org/10.48550/arXiv.2205.03257>

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv*. <https://doi.org/10.48550/arXiv.1312.6114>

MikroSim. (o. D.). MikroSim – Sektorenübergreifendes kleinräumiges Mikrosimulationsmodell. MikroSim. Abgerufen am 13. April 2023 von <https://mikrosim.uni-trier.de/>

Schneider, K. (2022). Forschung meets Datenschutz: Mit Künstlicher Intelligenz synthetische Gesundheitsdaten analysieren. Bundesinstitut für Arzneimittel und Medizinprodukte. Veröffentlicht am 10. März 2022. Abgerufen am 15.06.2023 von <https://www.bfarm.de/DE/Aktuelles/Blog/docs/2022-03-10-forschungsdatenzentrum.html#:~:text=Das%20Forschungsdatenzentrum%20Gesundheit%20am%20BfArM,dann%20mittels%20K%C3%BCnstlicher%20Intelligenz%20untersuchen.>

Szpruch, L., Cohen, S., Daniel, O., Houssiau, F. & Mole, C. (o. D.). Synthetic data and privacy preservation. The Alan Turing Institute. Abgerufen am 3. April 2023 von <https://www.turing.ac.uk/research/research-projects/synthetic-data-and-privacy-preservation>

## 5. Steckbriefe zu innovativen Methoden

### 5.1 Statistische Vorhersagemodelle zur Unterstützung der Beratung und Vermittlung von Arbeitslosen (Portugal)

#### 5.1.1 Einführung

Mit Hilfe der vorgestellten Dateninnovation können Vermittlungsfachkräfte in der öffentlichen Arbeitsverwaltung Portugals (IEFP)<sup>24</sup> über eine nach ihren Bedarfen gestaltete Schnittstelle auf die Ergebnisse eines statistischen Wahrscheinlichkeitsmodells zugreifen. Dadurch werden ihnen Informationen über den zu erwartenden Verbleib einer Person in der Arbeitslosigkeit zur Verfügung gestellt. Gemäß der geschätzten Wahrscheinlichkeit der Langzeitarbeitslosigkeit werden die Arbeitslosen in Risikogruppen eingeteilt. Das Modell ist damit ein Tool für das sogenannte „Profiling“. Das Tool soll dabei helfen, die Zielgenauigkeit der Beratung und Vermittlung zu erhöhen und den Eintritt von Arbeitslosen in Langzeitarbeitslosigkeit zu vermeiden. Um dies zu erreichen, erhalten Arbeitslose, deren Risiko der Langzeitarbeitslosigkeit als hoch eingeschätzt wird, zusätzliche Leistungen der Arbeitsförderung.

Das Projekt wird von der Nova School of Business and Economics an der Universidad de Nova de Lisboa in Partnerschaft mit dem IEFP durchgeführt. Statistische Modelle zur Schätzung der Wahrscheinlichkeit von Langzeitarbeitslosigkeit wurden am IEFP schon seit dem Jahr 2011 verwendet. Auch in einer Reihe weiterer Länder gibt es Erfahrungen mit dem statistischen Profiling von Arbeitslosen (Desiere et al., 2019). Das Innovationspotenzial der Kooperation zwischen IEFP und der Nova School of Business and Economics liegt neben der innovativen Methodik des Wahrscheinlichkeitsmodells vor allem in der Schnittstelle für die Vermittlungsfachkräfte in den regionalen Arbeitsvermittlungen. Über diese Schnittstelle können die Vermittlungsfachkräfte die relevanten Risikofaktoren bestimmen und die Betreuung der oder des Arbeitslosen individuell und effizient gestalten. In Deutschland steht den Vermittlungsfachkräften der BA und der Jobcenter ein solches Instrument derzeit nicht zur Verfügung.

#### 5.1.2 Technische Beschreibung

##### 5.1.2.1 Datenquellen

Die für das Vorhersagemodell in Portugal verwendeten Daten umfassen den Zeitraum zwischen 2012 und 2019. Insgesamt sind 3,25 Mio. Personen in den Daten enthalten, die in diesem Zeitraum zu mindestens einem Zeitpunkt arbeitslos waren. Zu jeder Person sind etwa 160 Merkmale verfügbar. Dazu gehören Daten des IEFP auf der Ebene der einzelnen Person sowie aggregierte Daten auf Ebene der regionalen Dienststellen des IEFP, Daten des portugiesischen Zensus zu regionalen Merkmalen und makroökonomische Kennziffern.

Das Vorhersagemodell löste ein bereits bestehendes, seit dem Jahr 2011 am IEFP verwendetes Modell ab. Der Mehrwert der Neuentwicklung liegt in der detaillierten Beschreibung der Arbeitsmarkthistorie, insbesondere der Berücksichtigung der individuellen Dauer der Arbeitslosigkeit zum Zeitpunkt der Wahrscheinlichkeitsvorhersage, und der sich daraus ergebenden verbesserten

---

<sup>24</sup> Das Instituto de Emprego e Formação Profissional [Institut für Beschäftigung und Berufsbildung] (IEFP) untersteht dem portugiesischen Ministerium für Wirtschaft und Beschäftigung.

Vorhersagequalität. Zudem wurden das Schätzverfahren mithilfe von Machine-Learning-Methoden verbessert und die Schnittstelle zur Arbeitsverwaltung neugestaltet.

### 5.1.2.2 Entwicklung des Vorhersagemodells

Das Vorhersagemodell nutzt Methoden des Machine Learning, um innerhalb der ersten zwölf Monate der Arbeitslosigkeit das Risiko zu berechnen, dass die betrachtete Person länger als zwölf Monate in Arbeitslosigkeit verbleibt und damit – ähnlich wie in Deutschland gemäß der Definition des § 18 SGB III – langzeitarbeitslos wird. Je nach dem Zeitpunkt im Verlauf der Arbeitslosigkeit wird ein unterschiedliches Schätzmodell verwendet. So wird für eine Person, die bereits zwischen neun und zwölf Monaten arbeitslos ist, das sogenannte 3mm-Modell geschätzt („3 more months“), und entsprechend ein 6mm-, 9mm- und 12mm-Modell. Die Modelle verwenden jeweils dieselben Merkmale.

Die Modelle gehören zur Kategorie der überwachten Algorithmen des Machine Learning (siehe Kapitel 3). Es werden vier verschiedene Algorithmen getestet: ein Decision Tree (Entscheidungsbaumverfahren), eine skalierte logistische Regression, ein Random Forest und der XGBoost-Algorithmus.<sup>25</sup> Das Modell wird auf Daten über einen Zeitraum von zwei Jahren innerhalb des Zeitraums von 2012 bis 2019 trainiert. Anschließend wird das Modell mithilfe von Daten für ein Jahr getestet. Dabei liegt zwischen den Trainings- und Testdaten eine Zeitspanne in Höhe des jeweiligen Modellhorizonts, sodass jeweils ersichtlich wird, ob eine Person tatsächlich zum Beispiel sechs weitere Monate arbeitslos ist (Bell et al., 2020). Aufgrund der vier Basismodelle und der verschiedenen Substichproben der verfügbaren Daten sowie von unterschiedlichen im Algorithmus verwendeten Parametern wurden insgesamt 4.000 verschiedene Modellvarianten geschätzt, die im Anschluss anhand verschiedener Kriterien evaluiert wurden und aus denen das finale Modell ausgewählt wurde.<sup>26</sup>

### 5.1.2.3 Evaluation des Vorhersagemodells

Als zu optimierendes Kriterium wird der Anteil der Arbeitslosen, die durch das Modell korrekterweise als Personen mit hohem Risiko für Langzeitarbeitslosigkeit identifiziert wurden, im Verhältnis zu allen als Hochrisiko-Personen eingestuften Personen betrachtet (diese Metrik wird *precision@k* genannt).<sup>27</sup> Durch dieses Kriterium wird der Anteil der fälschlich als Hochrisiko-Personen Eingestuften minimiert. Wenn die Ergebnisse des Profiling für die Zuordnung zu einer intensiven Beratung und Vermittlung verwendet werden, dient das Kriterium dazu, die Vergabe von arbeitsmarktpolitischen Maßnahmen an Personen zu vermeiden, die diese nicht benötigen, und ist somit ressourcenschonend (Bell et al., 2020).

Aus allen geschätzten Modellen wurde in einem mehrstufigen Auswahlverfahren das beste Modell für jeden Zeithorizont und jedes Verfahren ausgewählt.<sup>28</sup> In die Auswahl ging erstens die Präzision der Vorhersage gemäß dem erwähnten *precision@k*-Maß ein. Zweitens wurde das Kriterium der

---

<sup>25</sup> Details zu den Algorithmen sind in Breiman (1984, 2001), Chen & Guestrin (2016) und Menard (2002) enthalten.

<sup>26</sup> Dabei wurde das Tool „Triage“ verwendet, das vom Data Science and Public Policy Lab an der University of Chicago erstellt wurde. Dieses Tool findet insbesondere bei Datenprojekten Anwendung, die Machine Learning im Kontext von Politikberatung und -analyse verwenden. Triage verfügt über Funktionen zur Datenaufbereitung, Generierung weiterer Merkmale, sowie zum Trainieren und Testen der Modelle (Bell et al., 2020).

<sup>27</sup> Dafür wurden in Absprache mit dem IEFP die vorhergesagten Wahrscheinlichkeiten mithilfe von Schwellenwerten in drei Wahrscheinlichkeitskategorien (geringe, mittlere und hohe Wahrscheinlichkeit der Langzeitarbeitslosigkeit) übersetzt.

<sup>28</sup> Hier wurde erneut auf das Tool Triage zurückgegriffen. Für Details siehe Bell et al. (2020), ab Abschnitt 4.2.

Fairness herangezogen. Das bedeutet, dass Diskriminierungen bestimmter Gruppen vermieden werden sollen. Zur Aufdeckung von Diskriminierungen wird das open source-Tool „Aequitas“ verwendet. Dieses Tool wurde an der University of Chicago entwickelt, um Vorhersagen und Risikobewertungen, die auf Machine Learning beruhen, auf Verzerrungen und Diskriminierungen zu prüfen. Drittens wurden solche Algorithmen bevorzugt, die weniger komplex sind und deren Ergebnisse von den Vermittlungsfachkräften anhand einer begrenzten Anzahl von Merkmalen gut nachvollzogen werden können. Schlussendlich wurde der XGBoost-Algorithmus als optimal ausgewählt.

Die Beteiligten der Nova School of Business and Economics haben dem IEFP die erforderlichen Kenntnisse über Strukturen und Prozesse vermittelt, sodass das IEFP in Zukunft Aktualisierungen der Daten und Schätzparameter selbstständig durchführen kann. Allerdings besteht weiterhin Kontakt zwischen den Projektpartner\*innen, da Fehler in den Algorithmen weiterhin auftreten können und behoben werden müssen.

### 5.1.3 Anwendung

Die individuelle Vorhersage des Verbleibs in Arbeitslosigkeit durch den Machine-Learning-Algorithmus soll zur Verbesserung der Betreuung der Arbeitslosen beitragen. Arbeitslose, die durch das Modell in die mittlere und hohe Risikogruppe für den Eintritt in Langzeitarbeitslosigkeit gruppiert werden, erhalten von den Vermittlungsfachkräften gemäß ihrem „persönlichen Beschäftigungsplan“<sup>29</sup> verstärkte Maßnahmen und werden in ein Fallmanagement aufgenommen. Dabei können die Vermittlungsfachkräfte die Zuordnung durch das Modell gemäß ihrer eigenen Risikoeinschätzung korrigieren.

Für die Anwendung in der Praxis ist es zentral, dass das Modell bestmöglich in den Alltag der Vermittlungsfachkräfte eingebettet wird. Aufgabe des Projektteams war daher nicht nur die Bereitstellung eines Modells, das die Arbeit der Vermittlungsfachkräfte unterstützt, sondern auch die Entwicklung einer Benutzeroberfläche.

Eine Vorläuferversion des Modells hatten die Vermittlungsfachkräfte nach den Ergebnissen einer Studie nicht als nützlich bewertet, obwohl ein unterstützendes System grundsätzlich positiv gesehen wurde (Bell et al., 2020). Die Vermittlungsfachkräfte arbeiteten nicht wie vorgesehen mit dem Vorläufermodell, beispielsweise änderten sie den geschätzten Risikofaktor nicht gemäß ihrer eigenen Einschätzung, was ursprünglich angedacht war. Es stellte sich heraus, dass der Grund dafür die zeitintensive Begründung der Korrektur war, die für eine Anpassung des Risikofaktors zwingend erforderlich war.

Auf der Basis dieser Vorerfahrungen wurden im Rahmen eines kontrollierten Feldexperiments (Randomised Controlled Trial, RCT) unterschiedliche Modelle in zwölf regionalen Dienststellen des IEFP erprobt. Kriterien für die Auswahl waren die Akzeptanz des Modells und das Vertrauen der Vermittlungsfachkräfte in die Modellergebnisse, die Vornahme von manuellen Korrekturen und die Nutzeneinschätzung.

Abbildung 5-1 zeigt die Bedienoberfläche, die den Vermittlungsfachkräften in einer der getesteten Varianten bereitgestellt wurde. Im oberen Teil der Maske sind Identifikationsmerkmale enthalten, die aus der Datenbank des IEFP entnommen wurden. Die Menükarten im unteren Bereich sind mit

<sup>29</sup> Eine gemeinsame Verpflichtung, die zwischen der Dienststelle der Arbeitsverwaltung (Centro de Empleo) und der arbeitslosen Person abgeschlossen wird und die in Übereinstimmung mit dem Profil und den spezifischen Umständen der Person sowie dem Arbeitsmarkt die Schritte für die Integration in den Arbeitsmarkt festlegt, ähnlich wie die Eingliederungsvereinbarung nach dem deutschen SGB II und SGB III.

„Angaben zur Beschäftigung“, „Beruf“, „Voraussetzungen“, „Kenntnisse“, „Profiling“, „andere Daten“, „Ehepartner“, „Mobilität“ und „weitere Berufsklassifikationen“ überschrieben.

Darunter werden die ermittelten Risikobewertungen des Vorgängermodells und des verbesserten Modells angezeigt. In Abbildung 5-1 zeigt sich beispielsweise eine unterschiedliche Einschätzung der Modelle. Darunter werden zur Nachvollziehbarkeit der Einstufung die sechs wichtigsten risikoh erhöhenden Faktoren dargestellt. Die Grafik zeigt den Zeitverlauf der Risikoeinstufung. Im Rahmen der dargestellten Bedienoberfläche werden die Vermittlungsfachkräfte außerdem gefragt, inwieweit die Risikoeinschätzung des Modells nützlich für ihre eigene Einschätzung der Arbeitsuchenden ist. Im unteren Teil der abgebildeten Menüseite sollen die Vermittlungsfachkräfte ihre eigene Einschätzung eintragen und erläutern, aufgrund welcher Faktoren sie ihre eigene Risikoeinstufung abgegeben haben. Hier besteht also die Möglichkeit, eine andere Einstufung als die des Modells zu wählen.

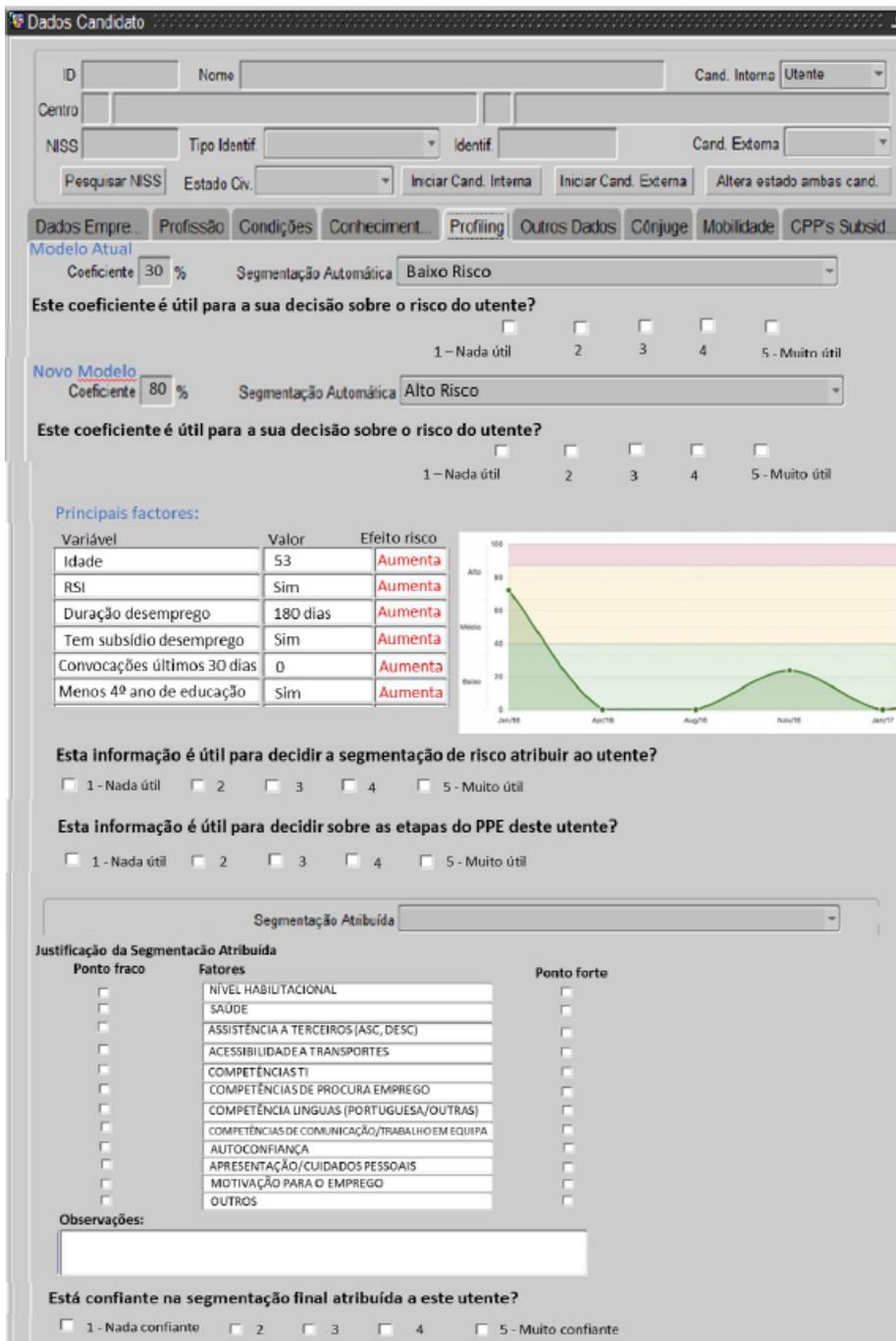
Im RCT wurde festgehalten, wie häufig der vorhergesagte Risikofaktor des Modells manuell geändert wurde. Die Auswertungen des RCT ergaben, dass eine Vereinfachung der Korrektur des vorhergesagten Risikofaktors, beispielsweise ohne notwendige Begründung, zu einer deutlich häufigeren manuellen Änderung durch die Vermittlungsfachkräfte führt. Allerdings wurde in einem Expertengespräch auch festgestellt, dass die Vermittlungsfachkräfte häufig Vorhersageeinschätzungen korrigierten, die sich im Anschluss als korrekt herausstellten, wenn beispielsweise eine arbeitsuchende Person, wie vom Modell eingeschätzt, tatsächlich langzeitarbeitslos wurde, während die Vermittlungsfachkraft diese Einschätzung geändert hatte.

In einem Expertengespräch wurde herausgestellt, dass eine enge Absprache und gegenseitige Zusammenarbeit zwischen den Projektpartner\*innen und den Vermittlungsfachkräften zur Implementierung der Bedienoberfläche notwendig ist, um Akzeptanz für das Verfahren zu schaffen. Um Befürchtungen entgegenzuwirken, dass die Innovation Vermittlungsfachkräfte überflüssig machen könnte, wurde in der Kommunikation mit diesen herausgestellt, dass die Innovation sie unterstützen, jedoch nicht ersetzen soll. Inzwischen wurde das Programm auf alle regionalen Dienststellen des IEFP ausgeweitet und wird dort von den Vermittlungsfachkräften in den Beratungsgesprächen verwendet.

Aus Sicht der Projektbeteiligten sind Strategien zur Eindämmung von Diskriminierungen im Kontext der Unterstützung von Arbeitslosen sehr wichtig. In der Vorversion des Modells wurde beispielsweise festgestellt, dass der Algorithmus präzisere Ergebnisse für jüngere Personen erzielt hatte. Daraufhin wurden Maßnahmen ergriffen, die dieser Tatsache Rechnung trugen und der Algorithmus erzielte ähnlich genaue Vorhersagen über alle Altersgruppen hinweg. Die Vermeidung von solchen Verzerrungen erhöhte die Akzeptanz des Verfahrens erheblich.

Während die Risikoeinschätzung den Prozess der Beratung und Vermittlung beeinflusst, erfolgt die Vergabe einzelner arbeitsmarktpolitischer Maßnahmen nach wie vor ausschließlich anhand der Einschätzung der Vermittlungsfachkräfte. Künftig ist auch eine Empfehlung zur Vergabe einzelner arbeitsmarktpolitischer Maßnahmen mithilfe eines Algorithmus geplant. Dieser Aspekt deutet an, dass die Anwendungsmöglichkeiten einer solchen Dateninnovationen sehr flexibel erweitert werden können.

Abbildung 5-1: Bedienoberfläche für die Vermittlungsfachkräfte



Quelle: Bell et al. (2020), S. 56.

### 5.1.4 Übertragbarkeit auf Deutschland

In Deutschland werden statistische Modelle zur Schätzung des Risikos für Langzeitarbeitslosigkeit derzeit ausschließlich in der Forschung verwendet (Kunaschk & Lang, 2022). Eine Nutzung von Modellen in den Prozessen der Beratung und Vermittlung von Arbeitslosen und Arbeitssuchenden, ähnlich wie im Projekt des IEF, findet derzeit nicht statt. Im Bereich der BA wird in der bewerberorientierten Vermittlung die weiterentwickelte Beratungskonzeption verwendet. Danach gliedert sich die Integrationsarbeit in vier Phasen. In der ersten Phase, der Situationsanalyse, findet

eine Standortbestimmung statt, aber – anders als im früheren 4-Phasen-Modell – kein Profiling im Sinne einer Eingliederung in „Profillagen“. Die Situationsanalyse beruht auf den Beratungsgesprächen zwischen der Vermittlungsfachkraft und der beratenen Person.

Mit dem Projekt „TrEffeR“ der BA wurde vor Längerem eine datengesteuerte Unterstützung des Vermittlungsprozesses entwickelt (Stephan et al., 2006). Kernstück war eine regional differenzierte quantitative Wirkungsanalyse der arbeitsmarktpolitischen Maßnahmen auf der Basis von Prozessdaten aus den Datenbeständen der BA. Die Ergebnisse der Wirkungsanalyse sollten den Vermittlungsfachkräften Prognosen der Effekte von einzelnen Maßnahmen im individuellen Fall zur Verfügung stellen (sogenanntes Targeting). Diese Anwendung hat sich nach Auskünften von Expert\*innen nicht in der Praxis etabliert, da sie eine wenig intuitive Bedienoberfläche besaß und die Vermittlungsfachkräfte nicht ausreichend in das System eingeführt wurden.

Allerdings hat sich auch gezeigt, dass statistische Modelle als Tools zur Unterstützung von Vermittlungsfachkräften nicht grundsätzlich abgelehnt werden. Unter der Voraussetzung, dass das Modell eine gute Prognosegüte aufweist und zusammen mit der Bedienoberfläche im Dialog mit den Vermittlungsfachkräften entwickelt wird, könnte eine datengesteuerte Unterstützung für die Vermittlungsfachkräfte auch in Deutschland hilfreich sein. Insbesondere könnte dabei das Ziel der schnellstmöglichen Integration in Arbeit in den Vordergrund gestellt werden, also insbesondere die Frage nach dem besten Instrument der aktiven Arbeitsmarktpolitik für die jeweilige Person – ähnlich, wie es auch in Portugal geplant ist.

Bezüglich des Datenschutzes existieren keine grundsätzlichen Hürden, denn die Vermittlungsfachkräfte haben über das System VerBIS bereits jetzt Zugriff auf die Daten der BA, die auch in das Modell eingehen würden. Allerdings müsste vor einer Einführung eines Tools eine Datenschutzprüfung hinsichtlich des Merkmalsumfangs, des Differenzierungsgrades, des Gültigkeitszeitraums und weiterer Eigenschaften der verwendeten Daten sowie hinsichtlich der Voraussetzungen für automatisierte Einzelfallentscheidungen und übrigen Zulässigkeitsvoraussetzungen für die Datenverarbeitung vorgenommen werden.

Nach dem Vorschlag der Europäischen Kommission (EU-Kommission) für eine EU-Verordnung für Künstliche Intelligenz (KOM(2021)206) (KI-Verordnung) würde eine solche Dateninnovation in Deutschland als Hochrisiko-KI-System eingestuft werden. Daraus ergeben sich u.a. konkrete Anforderungen an das Risikomanagement, die verwendeten Daten, die Daten-Governance, die technische Dokumentation und die Beaufsichtigung durch Menschen. Die Einstufung einer solchen Dateninnovation als Hochrisiko-KI-System lässt sich insbesondere mit der Größe der betroffene Personengruppe, der Relevanz einer funktionierenden Vermittlung für das Wohlergehen der davon betroffenen Personen begründen (Bundesministerium für Arbeit- und Soziales, 2022) sowie mit den möglichen negativen Folgen einer fehlerhaften Einstufung durch das System für die betroffene Person. Die Anforderungen der KI-VO sind bei einem solchen Hochrisiko-KI-System von Beginn an verbindlich.

Die „Selbstverpflichtenden Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung“ (Bundesministerium für Arbeit- und Soziales, 2022) können dabei schon heute als Orientierung bei der verantwortungsvollen Einführung von KI-Anwendungen herangezogen werden.

### 5.1.5 Fazit

Die hier vorgestellte Innovation aus Portugal unterstützt die Vermittlungsfachkräfte in den lokalen Dienststellen der Arbeitsverwaltung bei der Einschätzung des Risikos der Langzeitarbeitslosigkeit. Diese Unterstützung erfolgt über eine Schnittstelle, die neben der Wahrscheinlichkeit der Langzeitarbeitslosigkeit auch die ausschlaggebenden Faktoren und die Entwicklung des Risikos im

Zeitverlauf darstellt und den Vermittlungsfachkräften die Möglichkeit bietet, die Risikoeinschätzung des Modells zu ändern. Das Portal wurde in enger Zusammenarbeit mit den Vermittlungsfachkräften vor Ort entwickelt und getestet. Der zugrunde liegende Algorithmus zur Schätzung der Wahrscheinlichkeit ist eine Anwendung von Machine Learning. Durch den implementierten Algorithmus kann die drohende Langzeitarbeitslosigkeit besser prognostiziert werden als mit konventionellen Methoden. Zugleich wird das Risiko von Diskriminierungen durch den Algorithmus berücksichtigt. Die Anwendung ergibt ein Profiling in Risikogruppen, das den weiteren Prozess der Beratung und Vermittlung mitbestimmt. Die Zuweisung zu bestimmten arbeitsmarktpolitischen Maßnahmen im Sinne eines Targeting ist in der Planung, aber derzeit noch nicht implementiert.

Die portugiesische Arbeitsverwaltung (IEFP) ist offen für Innovationen und an datengestützten Tools für die Vermittlungsarbeit interessiert. Dies war eine wichtige Vorbedingung für die enge Zusammenarbeit zwischen der Nova School of Business and Economics und dem IEFP. Insgesamt zeichnet sich die Implementierung der Innovation durch eine enge Kommunikation zwischen den Entwickler\*innen, dem IEFP und den Vermittlungsfachkräften aus.

### 5.1.6 Quellen des Steckbriefs

Bell, A., Soares, C., Martínez de Rituerto de Troya, Í., Mestrinho, L., Zejnilovic, L., Ghani, R. & Lavado, S. (2020). *Predicting long-term unemployment: an applied approach for intelligent unemployment management - Technical Report*. NovaSBE, Data Science Knowledge Center.

Breiman, L. (1984). *Classification and Regression Trees* [1. Version]. Routledge.  
<https://doi.org/10.1201/9781315139470>

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.  
<https://doi.org/10.1023/A:1010933404324>

Bundesministerium für Arbeit- und Soziales (2022). Selbstverpflichtende Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung. Herausgeber: Denkfabrik Digitale Arbeitsgesellschaft im Bundesministerium für Arbeit und Soziales.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.  
<https://doi.org/10.1145/2939672.2939785>

Desiere, S., Langenbucher, K., & Struyven, L. (2019). *Statistical profiling in public employment services: An international comparison* (OECD SEM Working Paper No. 224). OECD Social, Employment and Migration Working Papers. <https://doi.org/10.1787/b5e5f16e-en>

Kunaschk, M. & Lang, J. (2022). Can Algorithms Reliably Predict Long-Term Unemployment in Times of Crisis? – Evidence from the COVID-19 Pandemic (IAB-Discussion Paper 8|2022). IAB-Discussion Papers. Abgerufen am 1. August 2023 von <https://doku.iab.de/discussionpapers/2022/dp0822.pdf>

Menard, S. (2002). *Applied logistic regression analysis* [2. Version]. SAGE.  
<https://doi.org/10.4135/9781412983433>

Stephan, G., Rässler, S. & Schewe, T. (2006). Das TrEffeR-Projekt der Bundesagentur für Arbeit: Die Wirkung von Maßnahmen aktiver Arbeitsmarktpolitik. *Zeitschrift für ArbeitsmarktForschung*, 39 (3/4), 447-465.

## 5.2 Machine-Learning-Methoden in Frühwarnsystemen für Leistungsbezug (Australien)

### 5.2.1 Einführung

Das im Folgenden beschriebene Forschungspapier von Sansone & Zhu (2023) untersucht, welchen Mehrwert der Einsatz von Machine-Learning-Methoden für die Prognose des Bezugs von Mindestsicherungsleistungen im Vergleich zu konventionellen ökonomischen Methoden bzw. zu Risikoanalysen hat, die bislang in Australien in Beratungsstellen für Empfänger\*innen von Sozialleistungen verwendet wurden. Hierzu verwenden die Autor\*innen umfangreiche australische Verwaltungsdaten. Die auf Basis dieser Daten erstellten Ergebnisse zeigen, dass Machine-Learning-Methoden eine teils deutlich höhere Vorhersagekraft haben als konventionelle Ansätze. Dies ist laut den Autor\*innen insbesondere darauf zurückzuführen, dass die Machine-Learning-Methoden eine größere Zahl an Determinanten und insbesondere auch zahlreiche Interaktionen zwischen den Determinanten eines langen Bezugs von Sozialleistungen untersuchen können. Somit kann die komplexe, multifaktorielle Entstehung entsprechender Abhängigkeiten besser abgebildet und es können effizientere Frühwarnsysteme entwickelt werden.

Frühwarnsysteme, die Machine-Learning-Methoden nutzen, haben daher das Potenzial, solche Personen besonders gut zu identifizieren, die ohne Interventionen dauerhaft auf Sozialleistungen angewiesen sein könnten. Wenn diese Personen dann besonders frühe, umfassende oder besonders gut zugeschnittene Unterstützungsangebote erhalten, könnte ein Teil der langen Bezugsdauern verhindert werden. Sofern die Daten, die in die Machine-Learning-Modelle einfließen, wie in Australien bereits vorliegen, sind die Kosten einer Erweiterung der Frühwarnmodelle überschaubar. Durch die Unterstützung der datenbasierten Risikoanalysen könnten Vermittlungsfachkräfte außerdem einen größeren Teil ihrer Ressourcen für die zielgenaue Beratung aufwenden und auch damit zur Effizienz des Sozialleistungssystems beitragen. Datenbasierte Risikoanalysen zum Leistungsbezug auf individueller Ebene könnten grundsätzlich auch verstärkt in Deutschland angewendet werden.

### 5.2.2 Technische Beschreibung

#### 5.2.2.1 Datenquellen

Der Vergleich der Vorhersagekraft von Machine-Learning-Modellen mit konventionellen Methoden zur Vorhersage des individuellen Risikos für den Bezug von Sozialleistungen wird von Sansone & Zhu (2023) mit administrativen Daten aus Australien durchgeführt.

Im australischen System zur sozialen Sicherung werden alle Leistungen von der Behörde Centrelink verwaltet und ausgezahlt. In den von Centrelink bereitgestellten Daten sind für die Jahre 2000 bis 2019 über 32 Mio. Personen enthalten, die eine Leistung von der Behörde erhalten haben. Neben Mindestsicherungsleistungen werden auch einkommensunabhängige Leistungen, z. B. Elterngeld oder Pflegezuschläge, von Centrelink verwaltet.<sup>30</sup> Basierend auf den Daten von Centrelink wird vom Department of Social Services (DSS) der Datensatz DOMINO (Data On Multiple INdividual Occurrences) erstellt, der für Forschung und Behörden verfügbar ist und alle bei Centrelink hinterlegten Informationen bündelt. In diesem Datensatz werden neben den Leistungsarten, -zeiträumen und -höhen auch ausführliche demografische Informationen wie

<sup>30</sup> Eine Übersicht der von Centrelink verwalteten Leistungen ist dargestellt unter Services Australia (2023). Die Vergleichbarkeit dieser australischen mit den entsprechenden deutschen Sozialleistungen ist für die Frage der Übertragbarkeit der im Folgenden dargestellten Methode jedoch nicht unmittelbar relevant.

Haushaltsstruktur, Familienstand, Bildung, Wohnort und Wohnsituation, sowie Informationen zu Beschäftigungsverhältnissen, Einkommen und Vermögen der Leistungsbeziehenden erfasst. Im DOMINO-Datensatz liegen somit individuelle Informationen zum Leistungsbezug und zu demografischen Merkmalen für den Zeitraum des Leistungserhalts besonders detailliert vor.

Da der DOMINO-Datensatz alle Leistungen des Sozialsystems abdeckt, also beispielsweise auch Unterstützungsleistungen für Studierende oder Elterngeld, enthält er einen Großteil der Bevölkerung aus verschiedensten Gruppen und ermöglicht damit die Beantwortung unterschiedlichster Forschungsfragen. Für die Gesamtbevölkerung ist der Datensatz jedoch nicht repräsentativ, da Personen ohne Leistungsbezug nicht enthalten sind. Umgekehrt sind z. B. Personen mit geringem und mittlerem Einkommen und Frauen überrepräsentiert. Daher sollten diese Merkmale bei den Analysen, die auf diesen Daten beruhen, z. B. als Kontrollvariablen berücksichtigt oder die Ergebnisse entsprechend gewichtet werden.

Die von Sansone & Zhu (2023) für die Prognose der Leistungsbezugsdauer verwendete Stichprobe des DOMINO-Datensatzes umfasst grundsätzlich alle Personen, die am 1. Januar 2014 zwischen 15 und 66 Jahre alt waren und im Kalenderjahr 2014 mindestens eine Leistung von Centrelink erhalten haben. Hieraus wird eine zufällige 1-Prozent-Stichprobe gezogen, da die Anwendung der Machine-Learning-Verfahren mit dem Gesamtdatensatz die verfügbaren Rechenkapazitäten überschritten hätte. Der Analysedatensatz umfasst etwa 50.000 Personen, deren Leistungsbezug zwischen 2014 und maximal 2018 tagesgenau beobachtet wird. Hieraus berechnen Sansone & Zhu (2023) als erklärende Variable den Anteil der Tage, an denen eine Person im Zeitraum 2015 bis 2018 Mindestsicherungsleistungen erhalten hat. Hierbei handelt es sich nur um diejenigen 24 Sozialleistungen, die nach einer Bedürftigkeitsprüfung durch Centrelink ausgezahlt werden.<sup>31</sup>

Als erklärende Variablen werden die bereits aufgeführten soziodemografischen Variablen verwendet und zusätzliche Variablen berechnet, die die Dynamiken des Bezugs verschiedener Sozialleistungen und auch die Dynamiken von Beschäftigungsverhältnissen abdecken. Diese erfassen z. B. Anfang und Ende sowie die Höhe von Leistungsbezügen und die Anzahl und den Umfang von Beschäftigungsverhältnissen im Analysezeitraum. Insgesamt enthält der Datensatz 1.800 mögliche Kovariate (ohne Interaktionsterme). Zur Vorhersage der Dauer des Leistungsbezugs werden Variablen aus dem Basisjahr 2014 verwendet, die entsprechend der Stichprobenziehung für alle Personen im Analysedatensatz vorliegen.

### 5.2.2.2 Beschreibung der Methodik

Der Hauptbeitrag des Papiers besteht im Vergleich der Vorhersagekraft verschiedener Modelle in Bezug auf die Bezugsdauern von Mindestsicherungsleistungen. Die erste Modellgruppe umfasst konventionelle ökonomische Methoden, die zweite Modellgruppe neuere Machine-Learning-Methoden.

In der ersten Modellgruppe mit konventionellen Methoden wird der Bezug von Leistungen der Mindestsicherung zunächst mit einfachen Kleinst-Quadrat-Regressionen geschätzt, die zunächst nur verschiedene einzelne soziodemografische Kontrollvariablen sowie Indikatoren für vergangenen Leistungsbezug beinhalten. Anschließend werden die Kontrollvariablen in einem Modell gemeinsam verwendet und durch Ortsindikatoren ergänzt. Die Vorhersagen werden zudem mit einem weiteren

---

<sup>31</sup> Konkret umfassen diese Leistungen: Age Pension, Austudy, Bereavement Allowance, Carer Payment, Disability Support Pension, Exceptional Circumstances Payment, Farm Family Restart, Mature Age Allowance, Mature Age Partner Allowance, Newstart Mature Age Allowance, Newstart Allowance, Parenting Payment Partnered, Parenting Payment Single, Partner Allowance, Sickness Allowance, Special Benefit, Widow Allowance, Wife Pension Age, Wife Pension DSP, Widow B Pension, Youth Allowance (Apprentice), Youth Allowance (Other), Youth Allowance (Student), Youth Training Allowance.

Schätzmodell verglichen, das lediglich Indikatoren für die drei Zielgruppen (junge pflegende Angehörige, Studierende, die besondere Unterstützungsleistungen erhalten, minderjährige Eltern) enthält. Diese drei Gruppen stehen bisher im Fokus der öffentlichen Beratungsstellen, entsprechend bildet das Schätzmodell mit diesen Indikatoren das bisher verwendete „Frühwarnmodell“ ab. Dieses Modell zum Zielgruppenfokus kann die Bezugsdauer von Sozialleistungen jedoch deutlich schlechter vorhersagen als die einfachen ökonomischen Modelle, die mindestens das Alter oder weitere der beschriebenen Kontrollvariablen beinhalten.

In der zweiten Modellgruppe werden verschiedene Verfahren des Supervised und Unsupervised Machine Learning angewendet. Zunächst werden verschiedene Modelle mit Verfahren des Supervised Machine Learning berechnet: Least Absolute Shrinkage and Selection Operator (LASSO), Boosting und Support Vector Regression (SVR). Hierzu wird die gleiche Aufteilung der Stichprobe wie für die konventionellen Modelle verwendet: 80 Prozent der Daten dienen als Trainingsdaten, der verbleibende Teil als Testdaten. Mit den Trainingsdaten werden die Modelle berechnet, mit den Testdaten werden sie evaluiert und verglichen.

Die Ergebnisse aller drei Modelle der zweiten Modellgruppe haben eine deutlich höhere Vorhersagekraft als die Vergleichsmodelle der ersten Gruppe, die mit konventionellen Regressionen berechnet worden waren. Die Vorhersagekraft wird anhand des Bestimmtheitsmaßes ( $R^2$ ) verglichen. Die Vorhersagegüte aller drei einzelnen Machine-Learning-Modelle lag mit ca. 75 Prozent 12 Prozentpunkte oberhalb derer des besten konventionellen Modells.

Die Vorhersagegüte konnte zudem durch eine entsprechend der Empfehlung von Mullainathan & Spiess (2017) gewichtete Kombination der drei Machine-Learning-Modelle um weitere 2 Prozentpunkte gesteigert werden.

Die Ergebnisse der Machine-Learning-Modelle zeigen zudem, dass auch Determinanten für die Bezugsdauer entscheidend sind, die standardmäßig nicht in den australischen Frühwarnsystemen für Sozialleistungsbezug verwendet werden. Dazu gehören nach Einschätzung von Sansone & Zhu (2023) Indikatoren für Einkommensfluktuation, Umzüge sowie das Überschreiten von Bedürftigkeitsgrenzen. Andere Indikatoren, z. B. für indigene Bevölkerungsgruppen oder im Ausland geborene Personen, tragen in den Modellen dagegen nicht entscheidend zur Erklärung der Zielgröße bei, weshalb ihre Aufnahme in das Frühwarnsystem nicht zu einer Verbesserung der Vorhersagegüte führen würde.

Die Ergebnisse zeigen auch, dass die Ergebnisse von der verwendeten Zielvariable für die Prognose abhängen. Wird als abhängige Variable nicht die Bezugsdauer von Mindestsicherungsleistungen allgemein, sondern von spezifischen Bürgergeld-ähnlichen Leistungen, betrachtet, sind die Vorteile des Einsatzes von Machine-Learning-Methoden deutlich kleiner als für den allgemeinen Fall.

Zusätzlich zu Modellen des Supervised Machine Learning, die Personen mit einem hohen Risiko für eine lange Bezugsdauer von Sozialleistungen identifizieren, setzen Sansone & Zhu (2023) auch Methoden des Unsupervised Machine Learning ein. Diese Methoden sind besonders gut geeignet, um die Personen mit hohen Risiken in Cluster einzuteilen, die bestimmte Risikogruppen mit ähnlichen Kovariaten beschreiben. Die Risikogruppen werden also nicht vorgegeben, sondern die Machine-Learning-Methoden erzeugen eine Gruppeneinteilung. Dies sind im vorliegenden Fall z. B. sehr junge und sehr alte Personen, Beschäftigte mit mehreren gleichzeitigen Jobs sowie Unverheiratete. Die Identifikation solcher Gruppen ist aus Sicht der Autor\*innen gut geeignet, um gebündelte Unterstützungsmaßnahmen zu entwickeln. Für den australischen Datensatz liefern die Modelle die besten Ergebnisse, wenn nur wenige (fünf) Cluster gebildet werden.

Die zentrale Aussage des Papiers von Sansone & Zhu (2023) ist, dass die Bezugsdauer von Mindestsicherungsleistungen mit den in staatlichen Sozialversicherungsbehörden vorliegenden Daten und relativ einfachen Machine-Learning-Methoden gut vorhergesagt werden kann. Somit ist eine zielgenauere Unterstützung von Bedürftigen möglich, die die langfristige Abhängigkeit von solchen Sozialleistungen reduzieren kann. Selbst ohne den Einsatz von Machine-Learning-Methoden können die Ergebnisse dieser Modelle nach Einschätzung der Autor\*innen verwendet werden, um Variablen zu identifizieren, die auch die Vorhersagekraft konventioneller Modelle verbessern können.

Beim Vergleich und der Kombination verschiedener Machine-Learning-Methoden zeigte sich außerdem, dass eine Kombination der verschiedenen Methoden des Supervised Machine Learning die Erklärungskraft noch einmal erhöhen kann.

### 5.2.3 Anwendung

Da das beschriebene Forschungspapier erst kürzlich veröffentlicht wurde, gibt es noch keine direkten Anwendungen in der Sozialverwaltung. Die Autor\*innen der Studie (Sansone & Zhu, 2023) stehen jedoch mit den australischen Behörden im Austausch dazu, unter welchen Bedingungen der Ansatz zukünftig in die Beratungsprozesse implementiert werden kann.

Die Anwendung in Australien könnte dadurch erschwert werden, dass es dort in den vergangenen Jahren eine kontroverse öffentliche Debatte zum Einsatz von Machine-Learning-Methoden gab. Ausgangspunkt waren Rückzahlungsforderungen für Schulden von Personen und Haushalten gegenüber dem Staat (Hegarty, 2022; ABC News, 2022). Diese Forderungen wurden automatisiert erzeugt und basierten teilweise auf Machine-Learning-Methoden; in der Debatte wurde die Methode automatisch berechneter Rückzahlungsforderungen daher als „Robodebt scheme“ bezeichnet. Die erstellten Bescheide waren teilweise fehlerhaft und hatten für einige Personen dramatische Folgen (Nikidehaghani et al., 2023; Lewis, 2020). Der Prozess wird aktuell von einer öffentlichen Kommission aufgearbeitet (Royal Commission, o. D.). Dieses Beispiel illustriert die allgemeine Schlussfolgerung, dass – wie bereits in Kapitel 3 dargestellt – eine zu starke Automatisierung von Prozessen in der öffentlichen Verwaltung Risiken birgt.

### 5.2.4 Übertragbarkeit auf Deutschland

Generell ist die Übertragbarkeit der von Sansone & Zhu (2023) vorgeschlagenen Methode der Entwicklung eines Frühwarnsystems für den Bezug von Mindestsicherungsleistungen auf individueller Ebene auch für Deutschland möglich. Zwar sind die in der BA vorliegenden Daten weniger umfangreich als die im australischen DOMINO-Datensatz vorhandenen, wo Daten zu verschiedensten Bezügen von staatlichen Transfers gebündelt werden, doch auch in der BA sind Daten zu zahlreichen individuellen Merkmalen mit einer hohen Datenqualität verfügbar. Durch die größere Bevölkerungszahl in Deutschland im Vergleich zu Australien stehen zudem genug Beobachtungen zur Verfügung, um selbst mit einer relativ kleinen Stichprobe (1 Prozent im beschriebenen Forschungspapier) valide Aussagen treffen zu können. Der konkrete Nutzen neuer Risikomodelle ist jedoch auch immer kontextabhängig, wie die Robustheitsanalysen von Sansone & Zhu (2023) zeigen. Dort erwiesen sich die vorgeschlagenen Machine-Learning-Methoden für den Bezug von bestimmten Leistungen der Mindestsicherung als weniger hilfreich als für den Bezug sämtlicher Mindestsicherungsleistungen zusammengenommen.

Auch wenn die Kosten für die Implementierung eines datenbasierten Frühwarnsystems zum langfristigen Bezug von Mindestsicherungsleistungen überschaubar wären, müssen weitere Gesichtspunkte in eine Kosten-Nutzen-Abwägung aufgenommen werden. In diesem Kontext sollten neben einer vollumfänglichen Gewährung des Datenschutzes insbesondere datenethische Argumente berücksichtigt werden. Im konkreten Fall der Mindestsicherungen kann es durch die

Unterstützung durch automatisierte Risikoanalysen zu Cream-Skimming-Effekte kommen, die auch von Sansone & Zhu (2023) beschrieben werden. Das bedeutet, dass Vermittlungsfachkräfte Anreize haben, vor allem Personen mit niedrigen Risiken zu unterstützen, weil sie so schnellere Erfolge bei der Verkürzung von Leistungsbezug vorweisen können. In diesem Fall könnte eine verbesserte Risikoanalyse den unerwünschten Effekt haben, dass das Cream Skimming noch effektiver vorgenommen wird. Zudem sind negative Auswirkungen im Sinne sich selbst erfüllender Prophezeiungen möglich. Personen, denen ein hohes Risiko für einen permanenten Bezug von Mindestsicherungsleistungen prognostiziert wird, sowie deren Vermittlungsfachkräfte werden durch diese Prognose möglicherweise davon abgehalten, Investitionen zur Vermeidung einer Realisation dieses Risikos zu tätigen. Auch sind Effekte eines perpetuierenden Systems denkbar, wenn ein auf vergangenen Daten basierendes Modell z. B. für Frauen geringere Beschäftigungswahrscheinlichkeiten prognostiziert, weil ihre Erwerbsbeteiligung in der Vergangenheit niedriger lag als die von Männern. Es muss daher genau abgewogen werden, ob die Implementierung eines automatisierten Frühwarnsystems die Entscheidungen von Vermittlungsfachkräften in die gewünschte Richtung beeinflusst oder unerwünschte Nebeneffekte hervorruft. Falls Risikoanalysen auf individueller Ebene auch in Deutschland verstärkt eingesetzt werden, sollten aktive Maßnahmen getroffen werden, um diese möglichen unerwünschten Nebeneffekte zu minimieren.

### 5.2.5 Fazit

Die vorgestellte Studie von Sansone & Zhu (2023) hat zum Ziel, Risikofaktoren bzw. bestimmte Personengruppen zu identifizieren, bei denen die Gefahr des permanenten Bezugs von Leistungen der Mindestsicherung besonders groß ist. Hierzu wurden auf der Grundlage von Daten einer australischen Behörde und mithilfe von Machine-Learning-Methoden entsprechende Analysen durchgeführt. Die Studie zeigt, dass Machine-Learning-Methoden einen Beitrag leisten können, solche Berechnungen zu verbessern. Vor allem, wenn Daten mit einem umfangreichen Merkmalspektrum eingesetzt werden, liefern Machine-Learning-Methoden bessere Vorhersageergebnisse als konventionelle Methoden. Die Ergebnisse der Studie zeigen das Potenzial dafür auf, wie gefährdete Personen besser identifiziert und anschließend zugeschnittene Unterstützungsangebote implementiert werden können, um die Abhängigkeit von Sozialleistungen auf individueller Ebene zu reduzieren.

Auch in Deutschland könnten entsprechende Methoden und Risikoanalysen auf individueller Ebene eingesetzt werden, da zumindest ein Teil der in den DOMINO-Daten enthaltenen Informationen auch bei der BA in Deutschland vorliegt. Ob die Erstellung entsprechender Analysen tatsächlich sinnvoll ist, müsste im Rahmen einer Abwägung mit den betroffenen Behörden diskutiert werden. Bisherige Kosten-Nutzen-Abwägungen könnten durch die Einbeziehung neuer ökonomischer Methoden ggf. anders ausfallen als entsprechende in der Vergangenheit erstellte Analysen.

### 5.2.6 Quellen zum Steckbrief

ABC News. (2022). Here's how we got to this point. *ABC News*. Veröffentlicht am 26. August 2022. <https://www.abc.net.au/news/2022-08-26/robodebt-royal-commission-explained/101374912>

Hegarty, N. (2022). Robodebt royal commission to investigate unlawful debt collection from hundreds of thousands of Australians. *ABC News*. Veröffentlicht am 25. August 2022. Abgerufen am 15.06.2023 von <https://www.abc.net.au/news/2022-08-25/robodebt-royal-commission-to-probe-unlawful-debt-collection/101357198>

Lewis, P. (2020). *I, Robodebt - Submission to the Australian Human Rights Commission discussion paper*. The Australia Institute. <https://policycommons.net/artifacts/2038854/i-robodebt/2791297/>. CID: 20.500.12592/fg1jv1.

Mullainathan, S. & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87–106. <https://doi.org/10.1257/jep.31.2.87>

Nikidehaghani, M., Andrew, J. & Cortese, C. (2023). Algorithmic accountability: robodebt and the making of welfare cheats. *Accounting, Auditing & Accountability Journal*, 36(2), 677–711. <https://doi.org/10.1108/AAAJ-02-2022-5666>

Royal Commission (o. D.). *Royal Commission into the Robodebt scheme*. Royal Commission Into the Robodebt Scheme. Abgerufen am 11. April 2023 von <https://robodebt.royalcommission.gov.au/>

Sansone, D. & Zhu, A. (2023). Using Machine Learning to Create an Early Warning System for Welfare Recipients [Vorzeitige Online-Veröffentlichung]. *Oxford Bulletin of Economics and Statistics*. <https://doi.org/10.1111/obes.12550>

Services Australia. (2023). *A guide to Australian Government payments*. Services Australia. Abgerufen am 11. April 2023 von <https://www.servicesaustralia.gov.au/guide-to-australian-government-payments?context=1>

## 5.3 Die Vorhersage von Renteneintrittsentscheidungen durch Machine Learning (mehrere Länder)

### 5.3.1 Einführung

Die Determinanten des individuellen Zeitpunkts des Renteneintritts (insbesondere vor dem regulären Rentenzugangsalter) sind vor dem Hintergrund des demografischen Wandels von entscheidender Bedeutung für viele arbeitsmarkt- und rentenpolitische Fragestellungen. Volkswirtschaftliche Ansätze stellen hierbei oft das mikroökonomische Entscheidungskalkül der Versicherten in den Vordergrund, z. B. durch die Berechnung des „Optionswertes“ eines aufgeschobenen Rentenbeginns (Stock & Wise, 1990). Dieser Optionswert wird dadurch beeinflusst, dass ein späterer Renteneintritt die Möglichkeit höherer Rentenzahlungen schafft. In der Soziologie werden weitere Gesichtspunkte wie z. B. der Einfluss von Paarbeziehungen untersucht (Eismann et al., 2019), während die psychologische Forschung das Verhalten in Modellen wie dem Annäherungs-Vermeidungs-Konflikt abbildet, der entsteht, wenn eine Entscheidung sowohl positive als auch negative Folgen hat (Feldman & Beehr, 2011).

Die bessere Verfügbarkeit von Längsschnittdaten – sei es aus Befragungen oder aus Verwaltungsprozessen – hat neue Potenziale zur empirischen Prüfung dieser Modelle eröffnet. Der Einsatz von Machine Learning-Methoden erweitert zudem die Möglichkeiten, neben theoriebasierten Ansätzen auch datengetriebene Verfahren zur Prognose des individuellen Renteneintritts zu nutzen. Ein aktueller Forschungsbeitrag von Garibay et al. (2022) geht der Frage nach, ob die jüngsten Fortschritte im Bereich Machine Learning zur aktuellen Forschung über Rentenzugangsentscheidungen beitragen können. Der Artikel vergleicht die Prognosen zum Rentenzugang auf der Basis von Machine Learning mit konventionellen statistischen Modellierungen von Renteneintrittsentscheidungen.

Die im Beitrag von Garibay et al. (2022) auf Basis des SHARE-Datensatzes verwendete Methodik hat Potenziale für die Vorhersage der Eintritte in den Rentenbezug durch die gesetzliche Rentenversicherung in Deutschland – jedenfalls bei den Rentenarten, bei denen individuelle Entscheidungen eine wichtige Rolle spielen. Die besseren Prognosen könnten die Planung und Steuerung im Bereich der gesetzlichen Rentenversicherung unterstützen. z. B.

### 5.3.2 Technische Beschreibung

#### 5.3.2.1 Verwendung von Machine Learning zu Vorhersagezwecken

Die Anwendung nutzt Algorithmen des Supervised Machine Learning (siehe Abschnitt 3.2), um individuelle Renteneintrittsentscheidungen vorherzusagen. Es wird geprüft, inwieweit die Vorhersagekraft dieser Algorithmen höher ist als bei konventionellen, nicht-maschinellen statistischen Methoden.

Als Datengrundlage werden Längsschnittdaten aus sechs der sieben Wellen des Survey on Health, Ageing, and Retirement in Europe (SHARE) verwendet. Der SHARE-Datensatz umfasst eine große Anzahl von Variablen, die potenziell einen Einfluss auf den Rentenübergang haben: Land, Alter, Geschlecht, Gesundheit, Bildung, Vorhandensein von Enkelkindern und finanzielle Verpflichtungen (Miet- oder Hypothekenzahlungen). Machine Learning wird in der Analyse dafür genutzt, um aus der Vielzahl dieser Variablen diejenigen zu identifizieren, die für den Renteneintritt maßgebliche Prädiktoren sind.

### 5.3.2.2 Ergebnisse

In der Studie wird eine Reihe von Modellen verglichen, die den Einfluss der genannten Variablen auf den Übergang in die Rente schätzen. Hierbei werden sowohl konventionelle Übergangsratenmodelle mit oder ohne zeitabhängige Kovariate als auch sogenannte Random Forests (siehe Abschnitt 3.2) eingesetzt.

Das gewählte Kriterium für die Güte der Vorhersage ist das Kriterium der „Area under the ROC curve“ (AUC-Kriterium). Dieses bezieht sich auf die „receiver operating characteristic curve“ (ROC), die das Verhältnis zwischen richtig positiven und falsch positiven Werten angibt. Der Wertebereich des AUC-Kriteriums reicht von null (für 100 Prozent falsche Vorhersagen) bis eins (für 100 Prozent richtige Vorhersagen). Das Kriterium wird im vorgestellten Projekt über unterschiedliche Vorhersagezeiträume berechnet. Über einen kurzen und mittleren Prognosehorizont führt die Anwendung von unterschiedlichen Machine-Learning-Algorithmen stets zu höheren Werten für das AUC-Kriterium als konventionelle Methoden, d. h. die bei der Modellierung von Übergangsentscheidungen üblicherweise verwendeten konventionellen statistischen Methoden liefern eine geringere Vorhersagegenauigkeit.

Die Einflussfaktoren auf die Rentenzugangsentscheidung, die sich dabei als besonders vorhersagestark erweisen, unterscheiden sich allerdings zwischen den unterschiedlichen Schätzmethode und stimmen nicht notwendigerweise mit den Faktoren überein, denen man aufgrund von anderen Studien einen hohen kausalen Erklärungsgehalt zumessen würde. Insgesamt ziehen die Autor\*innen den Schluss, dass auf konventionelle (nicht-maschinelle) Modelle nicht verzichtet werden kann, wenn es darum geht, kausale Aussagen zu treffen. Für die reine Prognose der Entscheidungen im Einzelfall haben allerdings die Machine-Learning-Verfahren Vorteile. Diese sind insbesondere vorzuziehen, wenn sich das Rentenzugangsverhalten in einem konstanten institutionellen Rahmen bewegt, während die durch Rentenreformen hervorgerufenen Verhaltensänderungen kausale Analysen erfordern.

Die Ergebnisse zum Mehrwert der Machine-Learning-Methoden sind nicht notwendigerweise verallgemeinerbar, sondern sind von der Anwendung und den verwendeten Daten abhängig. In einer ähnlichen Studie für mexikanische Daten finden Rocha Salazar und Boado Penas (2019) keinen klaren Unterschied in der Vorhersagekraft zwischen Machine-Learning- und konventionellen Verfahren.

Beiden Studien ist gemeinsam, dass die Anzahl der verwendeten erklärenden Variablen in den Daten relativ gering ist. Würden mehr Variablen einbezogen, könnte die systematische Auswahl der verwendeten Prädiktoren einen Vorteil gegenüber einem konventionellen Vorgehen haben, das oft zu einer Verwendung von zu vielen Prädiktoren (Overfitting) neigt, um zu vermeiden, dass relevante Faktoren nicht berücksichtigt werden. Dadurch reduziert sich beim konventionellen Vorgehen jedoch die Güte der Vorhersage (Athey und Imbens, 2019).

### 5.3.3 Übertragbarkeit auf Deutschland

Im Bereich der gesetzlichen Rentenversicherung besteht in Deutschland grundsätzlich die Möglichkeit, Machine-Learning-Verfahren zur Vorhersage des Renteneintritts einzusetzen. Eine Vorhersage des durchschnittlichen Rentenzugangszeitpunktes wäre für die Vorhersage künftiger Einnahmen und Ausgaben der gesetzlichen Rentenversicherung relevant. Es könnte darüber hinaus sinnvoll sein, auch individuelle Vorhersagen zu berechnen. Damit könnte z. B. beurteilt werden, wie hoch das Risiko einzelner Personen ist, künftig in eine Erwerbsminderungsrente zu wechseln. Auf dieser Basis könnten Reha-Bedarfe genauer identifiziert werden. Solche Prognosen könnten die Berater\*innen der DRV in ihrer Beratungstätigkeit unterstützen. Bei Entscheidungen über die Gewährung von Leistungen können sie nach derzeitigem Recht nicht verwendet werden. Eine

Rechtsänderung, die dies ermöglichen würde, müsste sich auf eine breite gesellschaftliche Debatte stützen.

Die Ergebnisse von Studien wie derjenigen von Garibay et al. (2022) sind bisher nicht durch die gesetzliche Rentenversicherung aufgenommen worden. Im Folgenden sollen die Bedingungen für einen Transfer der bisher nur im akademischen Bereich realisierten Anwendung von Machine Learning auf die Frage des Renteneintrittszeitpunkts in die Praxis der Alterssicherung in Deutschland erörtert werden. Dabei wird insbesondere eine Nutzung im Bereich der gesetzlichen Rentenversicherung thematisiert. Für einen Praxistransfer sind eine Reihe von Kriterien zu berücksichtigen. Erstens muss die Nutzung zu einer verbesserten Vorhersage von Rentenzugangsentscheidungen führen, was voraussetzt, dass seitens der Versicherten überhaupt ein Entscheidungsspielraum existiert. Diese Frage ist bei unterschiedlichen Rentenarten unterschiedlich zu beantworten. Zweitens setzt eine Anwendung voraus, dass entsprechende Daten vorhanden sind. Drittens müssen diese Daten nutzbar sein, was insbesondere eine Frage des Datenschutzes ist. Viertens müssen Infrastrukturen (insbesondere bei der Deutschen Rentenversicherung) vorhanden sein, die die Implementierung der Verfahren gewährleisten können. Ein fünftes Kriterium ist der rechtliche Rahmen bei der Verwendung der Ergebnisse, z. B. in Bezug auf die Gewährung von Leistungen in Abhängigkeit von Prognosen.

#### 5.3.3.1 Entscheidungsspielräume der Versicherten beim Renteneintritt

Neben der regulären Altersrente gibt es unter bestimmten Voraussetzungen die Möglichkeit, zu einem früheren als dem regulären Renteneintrittsalter, Leistungen der gesetzlichen Rentenversicherung zu beziehen. So können Personen ab 63 Jahren die Altersrente für langjährig Versicherte in Anspruch nehmen, die nach frühestens 35 Jahren an anrechenbaren Zeiten in der Rentenversicherung gezahlt werden kann. Hierbei müssen sie allerdings einen Abschlag von bis zu 14,4 Prozent der monatlichen Rentenleistungen in Kauf nehmen. Damit besteht ein Abwägungsproblem zwischen dem frühen Renteneintritt und den Rentenabzügen. Für die Entscheidung ist eine Vielzahl individueller Faktoren maßgeblich, beispielsweise die Höhe der erreichten Rentenanwartschaft, die gesundheitliche Situation, die Art der Tätigkeit (arbeitsbedingte Belastungen, Nutzung von bestimmten Arbeitsmitteln usw.) und weitere Sachverhalte.

Entscheidungsspielräume für die Versicherten gibt es auch bei anderen Rentenarten, beispielsweise der Altersrente für schwerbehinderte Menschen und der Erwerbsminderungsrente für Personen, die aus gesundheitlichen Gründen nicht mehr arbeitsfähig sind und das reguläre Rentenzugangsalter noch nicht erreicht haben. Über die Rentenberechtigung wird individuell auf der Basis medizinischer Gutachten entschieden. Auch wenn der Grad der Behinderung bzw. Erwerbsminderung nicht durch die Versicherten beeinflusst werden, gibt es individuelle Entscheidungsspielräume insofern, als die Versicherten bei der Feststellung der Erwerbsfähigkeit bzw. des Grades der Behinderung mitwirken (und zumeist den ersten Schritt dazu tun). Ferner stellt sich bei der Altersrente für schwerbehinderte Menschen ein ähnliches Abwägungsproblem zwischen früherem Renteneintritt und Rentenabzügen wie bei der Altersrente für langjährig Versicherte.

Insgesamt ist das Renteneintrittsalter in der gesetzlichen Rente nicht allein durch die gesetzlichen Regeln determiniert, sondern es bestehen erhebliche individuell bedingte Entscheidungsspielräume, die eine Prognose des Renteneintritts durch Machine-Learning-Methoden sinnvoll erscheinen lassen.

#### 5.3.3.2 Verfügbare Daten

In der Anwendung von Garibay et al. (2022) wird mit dem SHARE-Datensatz ein Paneldatensatz aus Befragungsdaten verwendet. Wollte man den Machine-Learning-Ansatz auf Fragestellungen der gesetzlichen Rentenversicherung übertragen, wären dagegen Daten aus dem Bereich der Deutschen

Rentenversicherung besser geeignet, weil nur mit diesen Daten die Berechtigung zum Bezug der im letzten Abschnitt genannten Rentenarten hinreichend genau nachgebildet werden kann.

Die von der DRV zur Verfügung gestellten Daten (insbesondere Versicherungskontenstichprobe, VSKT, und Versichertenrentenzugang, RTZN) enthalten eine Vielzahl von Informationen, die sich auf den Rentenzugang auswirken könnte (z. B. Rentenpunkte, Qualifikation und Tätigkeiten durch den Tätigkeitsschlüssel, zurückliegende Zeiträume der Erwerbsbiografie, Behinderung, Arbeitsunfähigkeitstage).

Für Forschungs-, allerdings nicht für Verwaltungszwecke besteht durch die seit einiger Zeit zur Verfügung stehenden Verknüpfungen zwischen den Daten der DRV und Befragungsdaten (Sozio-oekonomisches Panel (SOEP) und SHARE-Daten) die Möglichkeit, das Merkmalspektrum wesentlich zu erweitern. Dies ermöglicht die Einbeziehung familiärer Umstände oder des Vermögens, was insbesondere bei der Entscheidung über den Eintritt in die Rente für langjährig Versicherte wichtig sein dürfte. Die Daten stehen allerdings nur für eine Stichprobe von Haushalten bzw. Versicherten zur Verfügung, was die Nutzungsmöglichkeiten für die Analyse von Teilgruppen begrenzt.

Für Personen, die an einer medizinischen und beruflichen Rehabilitation teilgenommen haben, liegt ein erweitertes Merkmalspektrum insbesondere im gesundheitlichen Bereich vor. Neben den bewilligten Maßnahmen sind auch medizinische Diagnosen Gegenstand der Daten. Die Nutzung dieser Informationen dürfte die Vorhersage des Renteneintrittszeitpunktes bei gesundheitlich eingeschränkten Menschen deutlich verbessern. Der Grundsatz „Prävention vor Rehabilitation vor Rente“ besagt, dass der Eintritt in den vorzeitigen Rentenbezug durch eine Rehabilitation möglichst abgewendet werden soll. Dazu ist es erforderlich, das Risiko eines Renteneintritts möglichst präzise abzuschätzen. Das Ziel sollte es dabei sein, solche Reha-Maßnahmen zu identifizieren, die eine Fortdauer der Beschäftigung möglichst bis zur Regelaltersgrenze am besten fördern. Eine ähnliche Zielsetzung verfolgt Steckbrief 5.5.

Gesundheitliche Determinanten des Rentenzugangs sind jedoch auch für andere Versicherte als für Personen, die an einer medizinischen und beruflichen Rehabilitation teilgenommen haben, relevant. Hier bietet es sich an, gesundheitliche Daten, insbesondere Daten der gesetzlichen Krankenversicherung, mit den Daten der Deutschen Rentenversicherung zu verknüpfen und gemeinsam für die Vorhersage des individuellen Renteneintritts zu nutzen.

#### 5.3.3.3 Nutzung der Daten, Datenschutz

Grundsätzlich sind für die Nutzung der im vorigen Abschnitt genannten Daten die Bestimmungen des Zweiten Abschnitts des SGB X sowie insbesondere des SGB VI abschließend, soweit nicht die DSGVO unmittelbar gilt (§ 35 Absatz 2 SGB I). Gemäß dem § 75 SGB X ist die Übermittlung von Sozialdaten für die Forschung und Planung zulässig, soweit sie „für ein bestimmtes Vorhaben der wissenschaftlichen Forschung im Sozialleistungsbereich [...] oder der Planung im Sozialleistungsbereich durch eine öffentliche Stelle im Rahmen ihrer Aufgaben“ erforderlich ist. Die Übermittlung der Sozialdaten und die weitere Verarbeitung bedürfen der vorherigen Genehmigung der obersten Bundes- oder Landesbehörde, die für den Bereich, aus dem die Daten herrühren, zuständig ist.

Die in Abschnitt 5.3.3.2 genannten Daten werden bereits vielfach innerhalb der DRV und in der Forschung verwendet.

### 5.3.3.4 Infrastrukturen zur Umsetzung

Für eine Nutzung durch die DRV könnten die dafür vorhandenen Infrastrukturen genutzt werden. Gegebenenfalls kommt auch die Hinzunahme externer Expertise in Betracht. Mit dem Forschungsnetzwerk Alterssicherung und dem Forschungsdatenzentrum der Rentenversicherung verfügt die DRV über bewährte Kooperationen mit der Wissenschaft.

### 5.3.4 Fazit

Machine Learning hat als Methode für die Vorhersage von Renteneintritten insbesondere dann einen Vorteil gegenüber konventionellen statistischen Verfahren, wenn die Daten eine Vielzahl von Merkmalen enthalten. Die Nutzung von Machine-Learning-Methoden kann die Grundlage für eine verbesserte Beratung und darüber hinaus die Gewährung von Leistungen im Bereich der Rentenversicherung sein, sofern die rechtlichen Voraussetzungen dafür geklärt bzw. geschaffen werden. Im Bereich der DRV sind geeignete Daten vorhanden, die eine aussagekräftige Prognose des vorzeitigen Renteneintritts in Rentenarten wie z. B. die Altersrente für langjährig Versicherte ermöglichen würden. Besonders detaillierte Daten gibt es für die Versicherten, die an einer Reha-Maßnahme teilgenommen haben. Lohnend wären daher beispielsweise Analysen zur Frage, wie sich der Rentenübergang dieser Personengruppe gestaltet, welche individuellen Risikofaktoren insbesondere für den Übergang in eine Erwerbsminderungsrente bestehen und wie den entsprechenden Risiken begegnet werden kann.

### 5.3.5 Quellen des Steckbriefs

Athey, S. & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685-725. <https://doi.org/10.1146/annurev-economics-080217-053433>

Eismann, M., Henkens, K. & Kalmijn, M. (2019). Origins and Mechanisms of Social Influences in Couples: The Case of Retirement Decisions. *European Sociological Review*, 35 (6), 790-806. <https://doi.org/10.1093/esr/jcz037>

Feldman, D. C., & Beehr, T. A. (2011). A three-phase model of retirement decision making. *American Psychologist*, 66 (3), 193-203. <https://doi.org/10.1037/a0022153>

Garibay, M. G., Srakar, A., Bartolj, T. & Sambt, J. (2022). Does Machine Learning Offer Added Value Vis-à-Vis Traditional Statistics? An Exploratory Study on Retirement Decisions Using Data from the Survey of Health, Ageing, and Retirement in Europe (SHARE). *Mathematics*, 10, 152. <https://doi.org/10.3390/math10010152>

Rocha Salazar, J. D. J. & Boado Penas, M. (2019). Scoring and prediction of early retirement using machine learning techniques: application to private pension plans. *Anales del Instituto de Actuarios Españoles*, 25 (4), 119-145.

Stock, J. & Wise, D. (1990). Pensions, the Option Value of Work, and Retirement. *Econometrica*, 58 (5), 1151-1180.

## 5.4 Machine-Learning-Verfahren für die Mindestlohnevaluation (USA)

### 5.4.1 Einführung

Die Auswirkungen des Mindestlohns auf den Arbeitsmarkt sind seit Jahrzehnten Gegenstand intensiver wissenschaftlicher Forschung (Card & Krueger, 1994, Neumark & Wascher, 2010; Dustmann et al., 2022). In der Regel wurde der Effekt auf Personen untersucht, die direkt durch die Einführung oder Erhöhung des Mindestlohns betroffen waren, weil ihr Stundenlohn nach der Mindestlohnreform angepasst werden musste. Eine der wenigen Ausnahmen hiervon ist die Studie von Card & Krueger (1995), die vom Mindestlohn *potenziell* betroffene Personen in den Blick nimmt, beispielsweise Personen, die vor einer Mindestlohnreform keiner Beschäftigung nachgingen. Dieser Ansatz wurde durch Cengiz et al. (2022) aufgenommen, wodurch potenziell betroffene Personen, zusätzlich zu tatsächlich betroffenen Personen, in der Evaluation von Mindestlohnreformen betrachtet werden können. Die dabei verwendeten Machine-Learning-Methoden führen zu einer objektiveren Vorgehensweise bei der Wahl des Schätzmodells und der Modellspezifikation und erhöhen die Validität der Schätzergebnisse.

Für Deutschland ist diese Innovation aus zwei Gründen von großem Interesse. Erstens ist es auch für die deutsche Mindestlohnforschung eine Herausforderung, die potenziell oder auch direkt vom Mindestlohn betroffenen Personen zu identifizieren, zumal bislang kein ausreichend großer Paneldatensatz verfügbar ist, der die Berechnung von Stundenlöhnen ermöglicht. Zweitens ist die eingesetzte Machine-Learning-Methodik von generellem Interesse, wenn es um die Evaluation von Politikmaßnahmen geht, bei der die Betroffenheit eine Rolle spielt. Neben dem Mindestlohn ist dies beispielsweise in der aktiven Arbeitsmarktpolitik, bei Maßnahmen zur Qualifizierung oder auch im Bereich Mindestsicherungsleistungen, in dem die Anspruchsberechtigung ebenfalls häufig schwierig zu ermitteln ist, der Fall.

### 5.4.2 Technische Beschreibung

#### 5.4.2.1 Datenquellen

Die Daten für die Studie von Cengiz et al. (2022) stammen im Wesentlichen aus zwei Datenquellen. Die erste ist der Current Population Survey (CPS) des U.S. Census Bureau und des U.S. Bureau of Labor Statistics, wobei sowohl die monatliche CPS-Befragung von rund 60.000 US-Haushalten als auch die Outgoing Rotation Group (CPS-ORG) verwendet werden. Die CPS-ORG-Daten enthalten neben demografischen Informationen und persönlichen Merkmalen auch Daten zum Stundenlohn, dem wöchentlichen Erwerbseinkommen und der wöchentlichen Arbeitszeit (US Census Bureau, 2021; Bureau of Labor Statistics, 2014). In dieser Studie wurden Daten für die Jahre von 1979 bis 2019 verwendet.

Der zweite Datensatz enthält Informationen zur Entwicklung von Mindestlöhnen auf regionaler Ebene enthält und wurde von Vaghul & Zipperer (2016) zusammengestellt. Darin sind Informationen zu 172 wichtigen Mindestlohn-Maßnahmen in den USA enthalten, insbesondere zur Einführung und zu Änderungen von Mindestlöhnen, sodass die Höhe des Mindestlohns in sämtlichen Staaten und Städten in den USA seit 1974 nachverfolgt werden kann. Die genutzten Informationen stammen aus verschiedenen Quellen. Zu diesen Quellen gehören u. a. die (bundes-) staatlichen Gesetzgebungen sowie Beschlüsse und Veröffentlichungen auf den Internetseiten der Arbeitsverwaltung.

Die Analyse nutzt Daten von insgesamt ca. 470.000 Personen, die im Zeitraum von 1979 bis 2019 potenziell von Änderungen des Mindestlohns betroffen gewesen sein könnten. Diese Personen müssen sich im Bereich einer künftigen, in den kommenden drei Jahren eingeführten Änderung des jeweiligen regionalen Mindestlohns befinden. Um sicherzustellen, dass sich die Effekte vorheriger

Mindestlohn-Maßnahmen stabilisiert haben und somit keine verzögerten Effekte auf die derzeitige Arbeitsmarktsituation der betrachteten Personen ausüben, werden nur Beobachtungen von Personen verwendet, die in den letzten fünf Jahren nicht von einer Änderung des regionalen Mindestlohns betroffen waren. Dies bedeutet beispielsweise, dass Personen zu Beginn des Jahres 2010 nur dann in den Datensatz aufgenommen werden, wenn im Zeitraum von Anfang 2010 bis Ende 2012 eine Änderung des Mindestlohns in ihrer Arbeitsregion erfolgte, der Mindestlohn auf regionaler Ebene aber seit Anfang 2005 stabil war.

#### 5.4.2.2 Machine-Learning-Ansatz

In einem ersten Schritt werden durch verschiedene ökonometrische Verfahren die Personen identifiziert, die potenziell von einer Mindestloohnerhöhung betroffen sind und die deshalb in eine Wirkungsanalyse einbezogen werden sollten. Im zweiten Schritt werden die hierfür verwendeten Methoden miteinander verglichen und die leistungsstärkste ausgewählt.

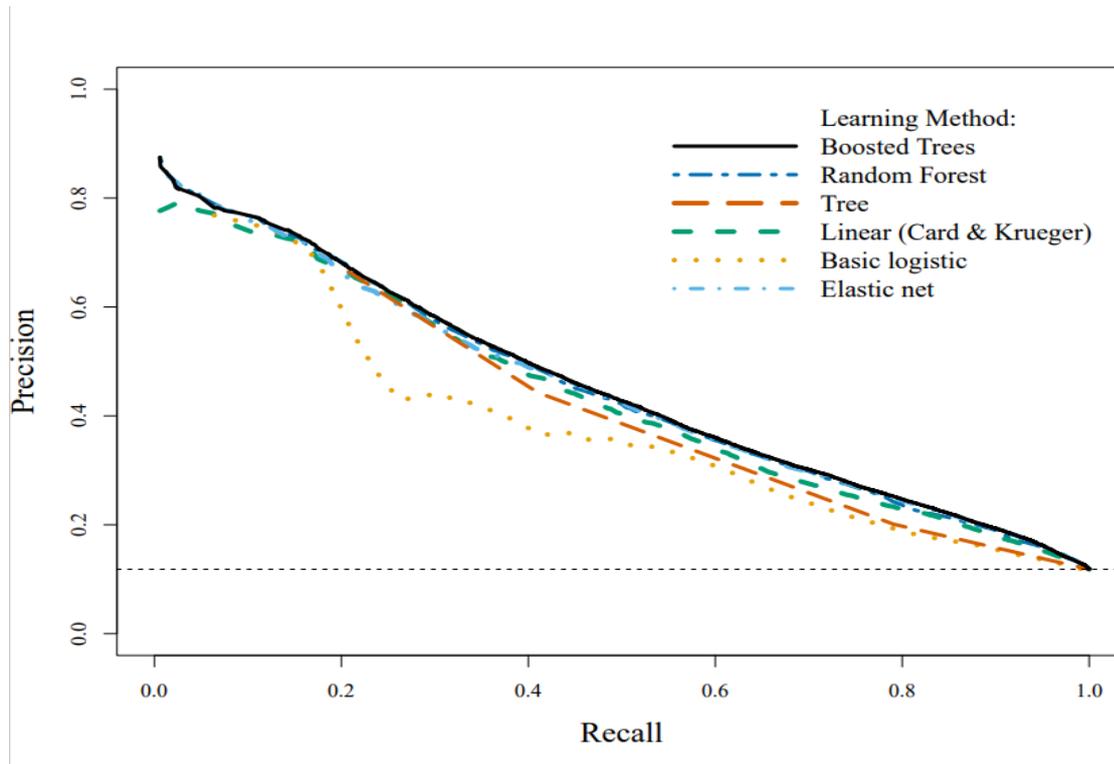
Der erste Schritt der Analyse besteht darin, die Stichprobe auf diejenigen Personen einzuengen, die vom Mindestlohn betroffen gewesen sein könnten. Denn die Arbeitsmarktergebnisse von Personen, die Löhne weit oberhalb des Mindestlohns erhalten, werden durch den Mindestlohn nicht beeinflusst. Diese Personen sollten also aus den Daten ausgeschlossen werden. Dazu wird ein Vorhersagemodell für die Wahrscheinlichkeit entwickelt, mit der eine Person mit bestimmten sozio-demografischen Merkmalen vom Mindestlohn betroffen sein könnte. Konkret wird die Wahrscheinlichkeit geschätzt, höchstens 125 Prozent des Mindestlohns zu verdienen. In der Mindestlohnforschung werden die betroffenen Personen meist so abgegrenzt, dass ihr beobachteter Lohn mit dem Mindestlohn verglichen wird. Die Vorhersage von Cengiz et al. (2022) basiert dagegen auf Merkmalen, die auch für Personen verfügbar sind, die zum Zeitpunkt der Beobachtung nicht auf dem Arbeitsmarkt tätig waren, die aber von einer Änderung des Mindestlohns betroffen wären, wenn sie künftig eine Beschäftigung aufnehmen wollten. Der Vorteil hiervon ist, dass auch die Wirkung des Mindestlohns auf die Entscheidung zur Partizipation am Arbeitsmarkt ermittelt werden kann.

Für die Schätzung der Wahrscheinlichkeit, vom Mindestlohn betroffen zu sein, verwendet die Studie von Cengiz et al. (2022) neben dem linearen Wahrscheinlichkeitsmodell von Card & Krueger (1995) verschiedene Verfahren des Supervised Machine Learning: Decision Trees, Random Forests, Gradient Boosting Decision Trees und Elastic Net Regularization (siehe dazu Kapitel 3). Der Vergleich der Machine-Learning-Methoden hinsichtlich ihrer Leistungsfähigkeit erfolgt im zweiten Schritt. Cengiz et al. (2022) vergleichen die Ergebnisse der genannten Methoden anhand ihrer Genauigkeit (precision) und ihrer Trefferquote (recall). Die Genauigkeit gibt den Anteil der Personen, die später tatsächlich den Mindestlohn erhalten, an allen als Betroffene klassifizierten Personen an. Die Trefferquote gibt hingegen an, wie hoch der Anteil der als Betroffene klassifizierten Personen an den Mindestlohnbeziehenden ist.

Beim Vergleich der Methoden stellt sich heraus, dass die Gradient-Boosted-Decision-Trees-Methode die beste Leistung erzielt, da sie die höchste Genauigkeit für fast jede Trefferquote aufweist. Die Unterschiede zwischen den verschiedenen Methoden fallen allerdings nicht sonderlich groß aus (siehe Abbildung 5-2). Dennoch argumentieren Cengiz et al. (2022), dass die eingesetzte Machine-Learning-Methode einen grundlegenden Vorteil besitzen, da sie eine objektive Vorgehensweise bei der Wahl des Modells und der jeweiligen Modellspezifikation (d.h. bei der Wahl der Modellvariablen) erlaubt. Der Einsatz von Machine-Learning-Methoden erlaubt es somit, ein Schätzmodell anhand objektiver Kriterien aus den zugrundeliegenden Daten zu entwickeln. Card & Krueger (1995) wählten hingegen eine von subjektiven Einschätzungen geleitete Vorgehensweise. Hierzu gehörten Einschätzungen über die verwendeten Prädiktoren und besonders über die funktionale Form, die komplizierte Mehrfachinteraktionen umfasst. Die Ergebnisse zur Genauigkeit und Trefferquote der

Vorhersage zeigen zwar, dass die auf diesen Einschätzungen beruhende Vorhersage kaum zu verbessern ist. Dies kann aber auch auf der besonderen Erfahrung der Autoren oder schlicht auf Glück beruhen, also auf Faktoren, die in der Forschung nicht immer gegeben sind. Im Vergleich dazu ist Machine Learning besser anhand objektiver Kriterien nachvollziehbar.

**Abbildung 5-2: Genauigkeit und Trefferquote der alternativen Verfahren**



Quelle: Cengiz et al. (2022). Die Abbildung zeigt die Evaluation der verschiedenen Machine-Learning-Methoden auf Basis der jeweiligen Genauigkeiten und Trefferquoten. Es ist ersichtlich, dass die "Gradient Boosted Decision Trees"-Methode die beste Leistungsfähigkeit hat. Zusätzlich aber auch, dass die Unterschiede nicht von signifikanter Größe sind.

### 5.4.3 Anwendung

Die leistungsstärkste Machine-Learning-Methode wird auf den gesamten Datensatz angewendet, um die Auswirkungen von Mindestloohnerhöhungen auf die betroffenen Personen zu schätzen. Wie beschrieben werden dadurch aus dem Gesamtdatensatz diejenigen Personen identifiziert, die durch den Mindestlohn betroffen sein könnten. Für diese Personen wird die eigentliche Schätzung durchgeführt, die durch ein konventionelles Regressionsmodell vorgenommen wird. Die Daten werden dazu auf der Ebene des Bundesstaats und des Quartals aggregiert. Die abhängigen Variablen sind die jeweiligen Löhne sowie die Wahrscheinlichkeit der Beschäftigung, der Arbeitslosigkeit und der Partizipation am Arbeitsmarkt. Diese Ergebnisse werden über einen Zeitraum von fünf Jahren nach der jeweiligen Mindestlohnänderung betrachtet. Die geschätzten Effekte zeigen signifikante Wirkungen der Mindestlohnanpassung auf die gezahlten Löhne, während die Effekte auf Beschäftigung, Arbeitslosigkeit und Arbeitsmarktpartizipation nicht statistisch signifikant sind.

Für die Anwendung von Machine-Learning-Methoden im Allgemeinen sind ausreichend große und repräsentative Datensätze eine wichtige Voraussetzung. Dies gilt besonders für die hier dargestellte Anwendung auf den Mindestlohn. Entscheidend ist hier eine ausreichend große Anzahl von

Beobachtungen für Personen oder Haushalte, die von Mindestlohnänderungen betroffen sein können. Daraus werden die Aggregatdaten berechnet, auf deren Basis die Schätzung der Wirkungen vorgenommen wird. Der verwendete CPS umfasst 60.000 Haushalte und erfüllt die Voraussetzungen. Die Daten des CPS sind auf der Webseite des United States Census Bureau (US Census Bureau, 2021) verfügbar, die Daten des CPS-ORG auf der Webseite der US-amerikanischen Denkfabrik CEPR (Center for Economic and Policy Research, 2020).

Neben großen Datensätzen ist eine große Rechenleistung erforderlich. Diese wird benötigt, um komplexe Algorithmen zu trainieren und zu evaluieren sowie letztlich die Machine-Learning-Methoden auf die Analyse anzuwenden. Es muss also gewährleistet werden, dass die notwendige Recheninfrastruktur bereitgestellt wird. Dies kann durch den Einsatz von Cloud-basierten Plattformen oder spezialisierten Hardware-Systemen wie graphics processing units (GPUs) oder tensor processing units (TPUs) erreicht werden.

Eine weitere Voraussetzung für die Studie ist eine ausreichende Variation in der Änderung der Mindestlöhne. In den USA gibt es seit dem Fair Labor Standards Act von 1938 einen bundesweiten Mindestlohn, aber die meisten Bundesstaaten erlassen darüber hinaus eigene Vorschriften zum Mindestlohn. Dies ermöglicht den Vergleich zwischen Bundesstaaten, der letztlich der von Cengiz et al. (2022) durchgeführten Regression zugrunde liegt. Zu den vielfältigen Änderungen der Mindestlöhne müssen auch Daten zur Verfügung stehen. Für die USA steht der regelmäßig aktualisierte Datensatz von Vaghul & Zipperer (2022) zu Mindestlöhnen im Zeitverlauf auf GitHub zum Download bereit.

#### 5.4.4 Übertragbarkeit auf Deutschland

Auch in Deutschland sind ähnliche Herausforderungen zu bewältigen wie in anderen Ländern, wenn es darum geht, die Betroffenheit durch den Mindestlohn zu ermitteln. Daher ist die Anwendung des Vorhersagemodells in Verbindung mit Machine Learning zur Wirkungsanalyse von Mindestlöhnen, wie in Cengiz et al. (2022) verwendet, eine vielversprechende Methode für die deutsche Mindestlohnforschung. Hierbei ist insbesondere die Identifikation von Personen, die zu einem bestimmten Zeitpunkt kein Erwerbseinkommen erzielen, aber aufgrund ihrer Merkmale, einschließlich ihrer Erwerbsbiografie, potenziell vom Mindestlohn betroffen sind, von Interesse. Entsprechende Analysen sind in Deutschland bisher nicht vorhanden. Potenziell nutzbare Datensätze sind diejenigen, die in einer Reihe von Evaluationsstudien bereits zum Einsatz kamen (z. B. Bachmann et al., 2020), insbesondere die Integrierten Erwerbsbiografien (IEB) des IAB und das Sozio-oekonomische Panel (SOEP).<sup>32</sup>

Zudem besteht in Deutschland die Herausforderung, dass die IEB aufgrund fehlender Angaben zur genauen Arbeitszeit keine Berechnung des Stundenlohns erlauben. Diese Situation kann potenziell verbessert werden, indem mithilfe von Machine-Learning-Verfahren die direkte Mindestlohn Betroffenheit in den Daten der Verdienststrukturerhebung ermittelt und dann auf die IEB übertragen wird (Dumpert & Beck, 2023).

Neben der Mindestlohnforschung könnte die vorgestellte methodologische Innovation allgemein in der Evaluationsforschung in Bereichen, die für das BMAS relevant sind, zum Einsatz kommen. Sie ließe sich als vorbereitenden Schritt in der Verwendung von Regressions- oder Differenz-von-

---

<sup>32</sup> Ein Nachteil des SOEP – insbesondere im Vergleich zu den Integrierten Erwerbsbiografien des IAB – sind seine relativ geringen Fallzahlen, die dazu führen, dass Subgruppenanalysen nur eingeschränkt durchgeführt werden können. Zudem handelt es sich um einen Befragungsdatensatz, der mehr Messfehler als andere Datensätze enthält. Dies erschwert die Identifikation der potenziellen Betroffenheit (Bachmann et al., 2020).

Differenzen-Verfahren einsetzen, indem diejenigen Haushalte oder Personen identifiziert werden, die von der zu evaluierenden Maßnahme potenziell betroffen sind. Dazu würde – entsprechend der vorgestellten Dateninnovation – die Bildung eines Vorhersagemodells vorgenommen. Somit können durch die Betrachtung der potenziell Betroffenen mehr Personengruppen einbezogen werden als bei einer Betrachtung der tatsächlich Betroffenen.

Mögliche Anwendungsbeispiele sind Maßnahmen der aktiven Arbeitsmarktpolitik oder der Qualifizierung, beispielsweise die Förderung von Weiterbildung für Beschäftigte. Ferner könnten auch Maßnahmen zur Steigerung der Inanspruchnahme von Sozialleistungen nach diesem Ansatz evaluiert werden. In jedem Einzelfall ist es wichtig, den Umfang, die Tiefe, Qualität und Verfügbarkeit der Daten zu prüfen, da sich diese je nach Kontext unterscheiden.

#### 5.4.5 Fazit

Die vorgestellte methodische Innovation ermöglicht eine bessere Evaluation von wirtschaftspolitischen Maßnahmen. Insbesondere ist die Identifikation von Personen, die potenziell von einer Maßnahme betroffen sind, besser möglich, wodurch diese Personengruppe – zusätzlich zu den tatsächlich vom Mindestlohn betroffenen Personen – in Evaluationsstudien aufgenommen werden können. Prinzipiell ist dies zwar auch mit traditionellen ökonometrischen Methoden machbar. Allerdings erhöht sich durch die Verwendung von Machine Learning die Effizienz der Analyse, insbesondere wenn sehr große Datensätze verwendet werden. Zudem bieten Machine-Learning-Methoden bessere Möglichkeiten, die Eignung verschiedener Modellspezifikationen und erklärender Variablen zu testen.

Die möglichen Anwendungsfelder der Innovation sind breit. Wie in der vorgestellten Analyse für die USA ließen sich auch in Deutschland die Effekte von Mindestloohnerhöhungen analysieren. Hierbei könnten dann vermehrt potenziell vom Mindestlohn betroffene Personen in den Blick genommen werden, d. h. Personen, die zum Zeitpunkt der Mindestloohnerhöhung kein Erwerbseinkommen erzielen, beispielsweise weil sie arbeitslos sind. Prinzipiell ließe sich die Methodik aber auf eine Reihe von Politikmaßnahmen anwenden, bei denen es von Interesse ist, potenziell Betroffene zu identifizieren und die Wirkungen der Politikmaßnahme auf diese Personen zu evaluieren.

#### 5.4.6 Quellen zum Steckbrief

Bachmann, R., Bonin, H., Boockmann, B., Demir, G., Felder, R., Isphording, I., Kalweit, R., Laub, N., Vonnahme, C. & Zimpelmann, C. (2020). *Auswirkungen des gesetzlichen Mindestlohns auf Löhne und Arbeitszeiten*. Studie im Auftrag der Mindestlohnkommission. IAW, IZA, RWI.

Bureau of Labor Statistics (2014). *Redesign of the Sample for the Current Population Survey* [Technische Dokumentation]. [https://www.bls.gov/cps/sample\\_redesign\\_2014.pdf](https://www.bls.gov/cps/sample_redesign_2014.pdf)

Card, D., Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4), 772-793.

Card, D. & Krueger, A. B. (1995). Myth and measurement: The new economics of the minimum wage. *Princeton University Press*. <https://doi.org/10.2307/j.ctv7h0s52>

Cengiz, D., Dube, A., Lindner, A. S. & Zentler-Munro, D. (2022). Seeing beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes. *Journal of Labor Economics*, 40(S1), 203-247. <https://doi.org/10.1086/718497>

Center for Economic and Policy Research (2020). *CPS ORG Uniform Extracts* [Version 2.5]. CEPRdata. Abgerufen am 28. März 2023 von <https://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-data/>

Dumpert, F., & Beck, M. (2023). Verbesserung der Datengrundlage der Mindestlohnforschung mittels maschineller Lernverfahren. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 17, 5-34. <https://doi.org/10.1007/s11943-023-00318-w>

Dustmann, C., Lindner, A., Schönberg, U., Umkehrer, M. & vom Berge, P. (2022). Reallocation Effects of the Minimum Wage. *Quarterly Journal of Economics*, 137(1), 267-328. <https://doi.org/10.1093/qje/qjab028>

Neumark, D. & Wascher, L. (2010). Minimum Wages and Employment. *Foundations and Trends® in Microeconomics*, 3(1-2), 1-182. <http://dx.doi.org/10.1561/07000000015>

US Census Bureau (2021). *Methodology*. Census.gov. Abgerufen am 21. März 2023 von <https://www.census.gov/programs-surveys/cps/technical-documentation/methodology.html>

Vaghul, K. & Zipperer, B. (2016). *Historical state and sub-state minimum wage data* [Working paper 090716]. Washington Center for Equitable Growth Working Papers. <http://cdn.equitablegrowth.org/wp-content/uploads/2016/09/02153029/090716-WP-Historical-min-wage-data.pdf>

Vaghul, K. & Zipperer, B. (2022). *State and sub-state historical minimum wage data*. GitHub. Abgerufen am 27. März 2023 von <https://github.com/benzipperer/historicalminwage>

## 5.5 Intelligent Rehabilitation Recommendation System (Republik Korea)

### 5.5.1 Einführung

Beim Intelligent Rehabilitation Recommendation System (IRRS) handelt es sich um ein in Südkorea implementiertes System, welches mithilfe von Machine-Learning-Methoden Personen identifiziert, die aufgrund eines Arbeitsunfalls eine Rehabilitation brauchen, da sie ohne Rehabilitation eine niedrige Rückkehrwahrscheinlichkeit auf den Arbeitsmarkt hätten. Die ausgewählten Personen werden anschließend durch Rehabilitationsexpert\*innen intensiv betreut. Für die Rehabilitationsplanentwicklung werden ebenfalls Empfehlungen des IRRS genutzt, die auf vergangenen erfolgreichen Maßnahmen mit jeweils ähnlichen Verletzungsmustern und Personenmerkmalen basieren. Durch das IRRS wird die Effektivität von Rehabilitationsmaßnahmen erhöht und die betroffenen Rehabilitand\*innen werden besser bei der Überwindung von Unfallfolgen und der Reintegration auf den Arbeitsmarkt unterstützt. Bereits seit 2011 entwickelt der Korea Workers' Compensation and Welfare Service (COMWEL) individuelle Rehabilitationspläne. Auf Grundlage der seitdem gewonnenen Erfahrungen und Daten zu erfolgreichen Rehabilitationsfällen wurde das IRRS im Jahr 2020 implementiert.

In Deutschland gibt es mit den Weller-Tabellen ein ähnliches datenbasiertes System im Bereich der Rehabilitationsentwicklung, jedoch dient dieses System vorwiegend der Abschätzung der Arbeitsunfähigkeitsdauer der Rehabilitanden. Die Rehabilitationspläne werden in Deutschland von Expert\*innen bzw. Ärzt\*innen entwickelt, die keine mit dem südkoreanischen System vergleichbare datenbasierte Unterstützung erhalten. Aufbauend auf den Daten der Weller-Tabellen hätte ein Transfer der Innovation nach Deutschland daher Potenzial zur Verbesserung der Versorgung der durch Arbeitsunfälle verletzten Personen. Auch im Bereich der gesetzlichen Rentenversicherung besteht Transferpotenzial.

### 5.5.2 Technische Beschreibung

Seit 2011 sammelt COMWEL Daten zu Personen, die Rehabilitationsmaßnahmen erhalten haben. Neben dem Geschlecht, dem Alter, der Verletzung und der Diagnose gehören dazu auch die Erwerbsbiografien und vergangene Leistungen der Arbeitslosenversicherung. Seit 2020 werden diese Daten genutzt, um mithilfe eines algorithmischen Vorhersagemodells einen Vulnerabilitätsindex zu erstellen, der die Wahrscheinlichkeit einer Rückkehr auf den Arbeitsmarkt der Rehabilitanden widerspiegelt (Korea Workers' Compensation & Welfare Service, 2021). Dieses Element des IRRS dient der Identifikation jener Fälle, die voraussichtlich besondere Betreuung durch Expert\*innen des COMWEL benötigen. Der Vulnerabilitätsindex wird mithilfe der Methoden des Rule-Based Filtering und Case-Based Reasoning geschätzt (Korea Workers' Compensation & Welfare Service, 2021). Diese Methoden nutzen Machine-Learning-Verfahren und erkennen Ähnlichkeiten in den Merkmalen zwischen Arbeitsunfällen und Rehabilitationen und den aktuellen zu bewertenden Fällen. Anschließend werden für die Fälle mit niedriger Rückkehrwahrscheinlichkeit mithilfe der gleichen Machine-Learning-Methoden individuelle und sequenzielle Rehabilitationspläne bis zur Wiedereingliederung auf den Arbeitsmarkt erstellt. Auch diese basieren auf einem Vergleich bereits betreuter Rehabilitanden, deren Rehabilitationsplan und dem jeweiligen Erfolg der Maßnahmen.

Die Maßnahme richtet sich an Personen mit besonders stark ausgeprägtem Rehabilitationsbedarf nach einem Arbeitsunfall. Im Jahr 2020 wurden durch das IRRS insgesamt ca. 2.600 Personen mit hohem Vulnerabilitätsindex identifiziert, welche insgesamt ca. 9.200 Maßnahmen im Rahmen ihrer Rehabilitationen zur Wiedereingliederung auf dem Arbeitsmarkt erhielten. Personen, die vom IRRS nicht als besonders förderbedürftig identifiziert werden, werden im traditionellen Rehabilitationssystem Südkoreas betreut.

Die im Rehabilitationssystem von COMWEL verfügbaren Maßnahmen lassen sich in drei Kategorien unterteilen: a) medizinische („medical“), b) soziale („social“) und c) berufliche („vocational“) Rehabilitationen. Darüber hinaus gibt es fünf Phasen der Rehabilitationen: a) akut, b) Genesung, c) Stabilisierung, d) vollständige Genesung und e) Rückkehr an den Arbeitsplatz. Innerhalb dieser Kategorien und Phasen gibt es dann im Einzelfall spezifische Maßnahmen, die durch den Algorithmus empfohlen werden. Beispielsweise gibt es innerhalb der medizinischen Rehabilitation die Maßnahmen Reha-Sport und Prävention von Komplikationen. Je nach Erfolg vergangener Maßnahmen bei ähnlichen oder vergleichbaren Diagnosen empfiehlt das System auf Basis von Rule-Based-Filtering- und Case-Based-Reasoning-Algorithmen nur eine oder auch beide Maßnahmen. In Einzelfällen lässt sich zeigen, dass bei identischem Verletzungsbild die Dauer der Arbeitsunfähigkeit bzw. die Zeit bis zur Wiedereingliederung auf den Arbeitsmarkt durch das IRRS kürzer ausfällt als vor Implementation des IRRS (Korea Workers' Compensation & Welfare Service, 2020). Im Mittelwert befinden sich die Rehabilitand\*innen jedoch länger in Behandlung, wenn sie Unterstützung aus dem IRRS erhalten (162 Tage ohne und 167 Tage mit IRRS Unterstützung in 2022). Dies ist jedoch zumindest teilweise auf die deutlich schwerwiegenden Diagnosen der im IRRS behandelten Personen zurückzuführen. Deutlich positiver fällt der Vergleich bei Betrachtung der Rückkehrwahrscheinlichkeit auf den Arbeitsmarkt aus. Trotz höherem Rehabilitationsbedarf kehren etwa 74 % der durch das IRRS unterstützten und 69 % der traditionell betreuten Rehabilitand\*innen auf den Arbeitsmarkt zurück.

### 5.5.3 Übertragbarkeit auf Deutschland

In Deutschland werden nach einem Arbeitsunfall die Rehabilitationsmaßnahmen in der Regel von medizinischem und therapeutischem Fachpersonal sowie den Unfallversicherungsträgern selbst entwickelt und begleitet. Besonders schwere Fälle werden in der Regel durch das Fachpersonal bisher ohne datenbasierte Unterstützung identifiziert und, ähnlich wie im Rahmen der hier beschriebenen Dateninnovation, intensiver betreut als Rehabilitanden mit leichteren Verletzungen. Für die Rehabilitationsplanentwicklung gibt es in Deutschland bereits einige Orientierungshilfen, wie beispielsweise das Fachbuch „Standards für Heilverfahren und Rehabilitation“ (Froese, 2015).<sup>33</sup>

Eine datenbasierte Hilfestellung, die vor und während der medizinischen Betreuung der verletzten Personen herangezogen wird, sind die sogenannten Weller-Tabellen. Diese Tabellen enthalten eine Übersicht von vergangenen Diagnosen und typischen Behandlungsplänen sowie den daraus resultierenden Arbeitsunfähigkeitsdauern (fsa, o. D.). Die Tabelle wurde 1999 eingeführt und basiert auf anonymisierten Daten von vier Unfallkassen in Deutschland. Das System wird alle zwei Jahre aktualisiert (Bischoff et al., 2021) und kann mit dem Schlüsselkatalog (Schlüsselhandbuch zur Unfallanzeige) der gesetzlichen Unfallversicherung (DGUV) verbunden werden (siehe Steckbrief 5.7). Es liegen circa 800.000 Datenpunkte vor, auf Basis derer die voraussichtliche Arbeitsunfähigkeitsdauer ermittelt wird. Derzeit werden die Weller-Tabellen nicht zur Entwicklung der tatsächlichen Rehabilitationsmaßnahmen herangezogen, laut Einschätzung verschiedener Expert\*innen wäre dies allerdings vorstellbar.

Einzelne algorithmische Verfahren werden in den Berufsgenossenschaften in Deutschland bereits entwickelt. In der Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse (BG ETEM) gibt es mit RehaPlus eine ähnliche datengestützte Innovation zur Verbesserung des Rehabilitationsangebots (Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse, 2022). In der Berufsgenossenschaft der Bauwirtschaft (BG BAU) wird derzeit eine vom BMAS geförderte KI-Anwendung entwickelt, die

---

<sup>33</sup> Für eine Implementation in ein algorithmisches Vorhersagemodell, das sich auf möglichst alle Arbeitsunfälle bezieht, eignet es sich aufgrund des Bezugs auf nur 25 Diagnosen allerdings nicht.

Aufsichtspersonen der BG BAU bei der Auswahl von sogenannten Unfalltendenzbetrieben unterstützt, bei denen also ein Arbeitsunfall wahrscheinlicher ist als bei anderen und deshalb am dringendsten kontrolliert und beraten werden müssen (Denkfabrik Digitale Arbeitsgesellschaft, 2023).

Ein Transfer der südkoreanischen Innovation muss im Lichte des geltenden datenschutzrechtlichen Rechtsrahmens sowie der Trägerlandschaft in Deutschland bewertet werden. In Südkorea übernimmt der COMWEL als zentrale Institution sowohl die Entwicklung und Implementation der Innovation als auch die Durchführung der Rehabilitationen. In Deutschland gibt es demgegenüber eine Vielzahl von Unfallversicherungsträgern, die jeweils ihre eigenen Prozesse und Daten verwalten. Durch eine trägerübergreifende Entwicklung der Innovation könnten Skaleneffekte erzielt werden, wofür allerdings zunächst Fragen der Datenzusammenführung zwischen den einzelnen Unfallversicherungsträgern geklärt werden müssten. Die Verwendung der Weller-Tabellen zeigt jedoch, dass Gesundheitsdaten trägerübergreifend genutzt werden können. Es wäre daher interessant zu prüfen, inwieweit der Dachverband der Unfallversicherungen in Deutschland, die DGUV, die Entwicklung einer solchen Innovation tragen könnte.

Insgesamt nimmt die Entwicklung der sogenannten elektronischen Heilverfahrenssteuerung seit Inkrafttreten des Digitalen-Versorgung-Gesetzes (DVG) im Jahr 2019 in Deutschland zu (Bischoff et al., 2021). In diesem Rahmen sind Entwicklungen geplant, die einen Transfer erleichtern oder die Vorhersagequalität verbessern könnten. So könnten im Rahmen eines Transfers Befragungsdaten der geplanten trägerübergreifenden Qualitätsanalyse eingebunden werden, die neben den durchgeführten Rehabilitationsmaßnahmen den Gesundheitszustand der Rehabilitanden erheben soll. Diese Qualitätsanalyse dient der Leistungskontrolle der Unfallversicherungen und den durchführenden Rehabilitationseinrichtungen. Sie enthält somit potenziell nützliche Daten für die Evaluierung der durchgeführten Rehabilitationen. Zudem ist bei der DGUV geplant, die einzelnen Rehabilitationsmaßnahmen elektronisch zu erfassen. Bisher stehen den Unfallversicherungsträgern nur unpräzise Informationen über die tatsächlich durchgeführten Rehabilitationsmaßnahmen zur Verfügung, die sich aus den übermittelten Kostenabrechnungen ableiten lassen. Beispielsweise ist nur erkennbar, ob eine Rehabilitation stationär oder ambulant durchgeführt wurde, aber nicht, welche konkreten Maßnahmen durchgeführt wurden. Diese Entwicklungen im Bereich der elektronischen Heilverfahrenssteuerung können einen günstigen Rahmen für einen Transfer der Innovation nach Deutschland bieten.

Das Innovationspotenzial bei der Einleitung und Durchführung von Rehabilitationen könnte über die gesetzliche Unfallversicherung hinaus auch auf die gesetzliche Rentenversicherung übertragen werden, die noch weitaus mehr Rehabilitationen erbringt als die Unfallversicherung. Die gesetzliche Rentenversicherung ist, anders als die Unfallversicherung, nicht bei Unfällen oder Berufskrankheiten zuständig. In der gesetzlichen Rentenversicherung werden stattdessen u.a. medizinische Rehabilitationen durchgeführt. Dies hätte entsprechende Konsequenzen für die Anwendung eines Algorithmus, beispielsweise durch anderweitige Datengrundlagen, da der Unfallbericht entfällt. Im Rahmen des DVG müsste entsprechend geprüft werden, welche Daten bei einem Transfer auf die gesetzliche Rentenversicherung genutzt werden könnten (siehe dazu auch Steckbrief 5.3).

#### 5.5.4 Fazit

Das IRRS nutzt ein Vorhersagemodell auf Basis von Machine-Learning-Methoden, um für Personen nach einem Arbeitsunfall einen optimalen Rehabilitationsplan zu entwickeln und den Wiedereinstieg in den Arbeitsmarkt schnellstmöglich zu erreichen. Für das System werden biografische Daten der Rehabilitanden und Daten zu bisher erfolgten Rehabilitationen und deren Erfolg genutzt. In Deutschland steigt die Nutzung der elektronischen Heilverfahrenssteuerung seit Inkrafttreten des

DVG. Mit dem System der Weller-Tabellen liegt bereits eine nützliche und datenbasierte Hilfestellung zur Abschätzung der Arbeitsunfähigkeitsdauer vor. Ein auf der Basis dieser Daten entwickeltes Vorhersagemodell zur Verbesserung der Rehabilitationsversorgung hätte für Deutschland ein hohes Innovationspotenzial.

### 5.5.5 Quellen des Steckbriefs

Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse (2022). Künstliche Intelligenz – Die Zukunft ist schon da. Herausgeber: Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse.

Bischoff, H. J., Kehl, P. & Schreyer, L. (2021). *Weller-System: Perspektiven der elektronischen Heilverfahrenssteuerung*. DGUV Forum, Ausgabe 9/2021.

Denkfabrik Digitale Arbeitsgesellschaft (2023). Mit Künstlicher Intelligenz die Verwaltung modernisieren: Für mehr Sicherheit in der Bauwirtschaft. Abgerufen am 13. September 2023 von <https://www.denkfabrik-bmas.de/projekte/ki-in-der-verwaltung/mit-kuenstlicher-intelligenz-die-verwaltung-modernisieren-fuer-mehr-sicherheit-in-der-bauwirtschaft>

Froese, E. (2015). *Standards für Heilverfahren und Rehabilitation* [4. Auflage]. Gentner Verlag.

fsa. (o. D.). *Steuerung des Heilverfahrens: Weller-Tabelle*. Forschungsgesellschaft für angewandte Systemsicherheit und Arbeitsmedizin. Abgerufen am 11. April 2023 von <https://www.fsa.de/produkte/weller-tabelle/>

Korea Workers' Compensation & Welfare Service (2021). *Big data-based services to enhance customized rehabilitation: Intelligent Rehabilitation Recommendation System (IRRS) (Good practices in social security)*. International Social Security Association, ISSA.

## 5.6 Unterstützung von Betriebsprüfungen zur Aufdeckung von nicht gesetzeskonformem Verhalten (Belgien)

### 5.6.1 Einführung

Mining Watch ist ein belgisches Analyseinstrument zur Verbesserung der Aufdeckung nicht gesetzeskonformen Verhaltens von Betrieben auf dem Arbeitsmarkt, beispielsweise nicht angemeldeter Erwerbstätigkeit, Missbrauch von Teilzeitregelungen oder Scheinselbstständigkeit. Basierend auf betriebspezifischen Geschäfts- und Regionaldaten erstellt ein Machine-Learning-Algorithmus einen individuellen Risikofaktor, der die Wahrscheinlichkeit ungesetzlichen Verhaltens widerspiegelt. Dieser Risikofaktor wird den Fahndungskräften der belgischen Aufsichtsbehörde zur Verfügung gestellt. Die Innovation wird vom Office Nationale de Sécurité Sociale (ONSS, Landesamt für Soziale Sicherheit) in Belgien betrieben.

Auf Basis des Risikofaktors können die Fahndungskräfte Betriebe gezielt für Prüfungen auswählen. Dadurch kann eine effizientere Nutzung der Ressourcen der Aufsichtsbehörde erzielt werden. Zudem soll mithilfe des Machine-Learning-Ansatzes der Tatsache entgegengewirkt werden, dass der Aufwand zur Aufdeckung nicht gesetzeskonformen Verhaltens zunimmt. Etwa 70 % der Vor-Ort-Prüfungen durch den Inspektionsservice des ONSS werden aufgrund offizieller Anordnungen durchgeführt. Die verbleibenden 30 % der Prüfungen basieren auf Einzelentscheidungen der Fahndungskräfte. Für einen Teil der Inspektionen auf Basis von Einzelentscheidungen werden mittlerweile die Analysen und Vorhersagen von Mining Watch genutzt (Williams, 2021; Europäische Kommission, o. D.). Nach Einschätzung der Beteiligten ist die Aufdeckungsquote durch den Einsatz von Mining Watch stark gestiegen (Williams, 2021).

In Deutschland werden derzeit im Bereich der Prüfung und Ermittlung von Betrieben, die sich im Kontext von Beschäftigung nicht an gesetzliche Vorgaben halten, keine datenbasierten Anwendungen eingesetzt. Die Verwendung eines ähnlichen Verfahrens wie Mining Watch würde daher eine Neuerung darstellen. Die Umsetzung könnte sich an dem hier vorgestellten Beispiel eines datengestützten Analyseverfahrens orientieren.

### 5.6.2 Technische Beschreibung

Das Ziel des im Rahmen von Mining Watch eingesetzten Algorithmus ist die Vorhersage eines Risikofaktors, der die Wahrscheinlichkeit von nicht gemeldeter Beschäftigung und anderen Formen von Betrug für das einzelne Unternehmen widerspiegelt. Dieser vorhergesagte Risikofaktor kann dann von den Fahndungskräften der Aufsichtsbehörde bei der Planung ihrer Prüfungen von Unternehmen herangezogen werden. Der Risikofaktor wird mithilfe von Machine-Learning-Algorithmen anhand von häufig in Verbindung mit Wirtschaftsbetrug stehenden Unternehmensmerkmale berechnet.

In der ersten Version der Innovation („Mining Watch 1“) wurden zunächst Personal- und Geschäftsdaten<sup>34</sup>, die der durchführenden Institution ONSS zur Verfügung standen, für die Vorhersage genutzt. In den Daten waren beispielsweise Informationen zur Personalfluktuation und zu Neueinstellungen, zu Umsatz- und Steuerdaten sowie Angaben zu gewerblichen Partnern und Kunden enthalten. Diese Daten lagen bereits vor Projektstart bei der SPF Sécurité Sociale (Föderaler Öffentlicher Dienst Soziale Sicherheit) vor. Ihre Inhalte sind aus Gründen der Geheimhaltung nicht öffentlich verfügbar. Bei der Nutzung von Mining Watch 1 konnten die Inspektor\*innen gezielt eine

---

<sup>34</sup> Aus Geheimhaltungsgründen dürfen diese Daten, insbesondere die verwendeten Merkmale, nicht genannt werden.

Branche, Betrugsart und Region auswählen und erhielten innerhalb dieser Filterung das Ranking der Unternehmen nach dem geschätzten Risikofaktor. Durch den Einsatz der Innovation ist die Aufdeckungsquote der Inspektor\*innen von 16 % auf 45 % gestiegen (Williams, 2021).

Die Aussagekraft des Risikofaktors im Rahmen von Mining Watch 1 war durch die eingeschränkte Weitergabe von sensiblen Unternehmensinformationen begrenzt, denn der behördenübergreifende Datenaustausch ging aufgrund der geltenden Datenschutzbestimmungen mit Informationsverlusten einher. Zudem bezog sich Mining Watch 1 nur auf die Sektoren Bau-, Reinigungs-, Hotel- und Gaststättengewerbe.

Die Weiterentwicklung von Mining Watch 1 zu Mining Watch 2 greift diese beiden Aspekte auf und erhöht die Nützlichkeit des Risikofaktors für die Inspektor\*innen. Zum einen liegen nun genauere Unternehmensinformationen vor. Zum anderen soll Mining Watch in naher Zukunft alle Sektoren umfassen. Ein weiteres neues Element von Mining Watch 2 ist die Darstellung der Unternehmensmerkmale, die für die Größe des Risikofaktors maßgeblich sind, in einem Dashboard. Inspektor\*innen können daher nachvollziehen, aufgrund welcher Sachverhalte ein Unternehmen beispielsweise einen hohen Risikofaktor erhält. Aktuell befindet sich Mining Watch 2 in der Entwicklung und wird nur von einer Testgruppe verwendet.

Für die Inspektor\*innen ist die Nutzung von Mining Watch nicht an dessen Empfehlungen gebunden, sondern dient als Unterstützung. Die finale Entscheidung, welches Unternehmen geprüft wird, obliegt immer den Inspektor\*innen. Dabei wurde in der Vergangenheit ca. 10 % der Einzelentscheidungen auf Basis der Empfehlungen von Mining Watch 1 getroffen. Informationen zur Nutzung von Mining Watch 2 sind noch nicht verfügbar. Zwischen dem Team, das für die technische Entwicklung des Algorithmus verantwortlich ist, und den Inspektor\*innen besteht ein ständiger Austausch. Dies betrifft zum einen die Weitergabe von neuen Informationen hinsichtlich der Relevanz spezifischer Unternehmensmerkmale durch den Mining Watch-Algorithmus an die Inspektor\*innen. Diese Informationen können dann bei der Prüfung von Betrieben vor Ort genutzt werden. Zum anderen sind die Erfahrungen der Inspektor\*innen eine wertvolle Grundlage für Weiterentwicklungen der Algorithmen, insbesondere was die Berücksichtigung verschiedener Unternehmensmerkmale betrifft.

### 5.6.3 Übertragbarkeit auf Deutschland

In Deutschland prüfen und kontrollieren verschiedene Institutionen die Einhaltung der sozialversicherungsrechtlichen Meldepflichten und den Missbrauch von Leistungen nach den SGB II und III.<sup>35</sup> Zu nennen ist hier beispielsweise die Finanzkontrolle Schwarzarbeit (FKS). Derzeit werden bei der FKS keine datengestützten Verfahren zur Unterstützung bei nicht hinweisbezogenen Prüfungen eingesetzt, die mit dem belgischen Verfahren Mining Watch vergleichbar sind. Es ist jedoch für die Zukunft angedacht, behördenübergreifend Daten zu Analysezielen hinsichtlich potenzieller Risikomerkmale zusammenzuführen und auszuwerten. Bestehende datenschutzrechtliche Vorschriften müssen dabei umfassend beachtet werden. Risiken wie Fehleinschätzungen, Diskriminierungen und sonstige Verzerrungen müssen mitigiert werden (siehe Abschnitt 3.2). In Anlehnung an die Praxis in Belgien sollten daher die Empfehlungen eines Algorithmus nicht bindend sein. Das Verfahren sollte zunächst anhand einer Minderheit der Fälle erprobt werden.

---

<sup>35</sup> [https://www.zoll.de/DE/Fachthemen/Arbeit/Bekaempfung-der-Schwarzarbeit-und-illegalen-Beschaeftigung/Aufgaben-und-Befugnisse/aufgaben-und-befugnisse\\_node.html](https://www.zoll.de/DE/Fachthemen/Arbeit/Bekaempfung-der-Schwarzarbeit-und-illegalen-Beschaeftigung/Aufgaben-und-Befugnisse/aufgaben-und-befugnisse_node.html), zuletzt abgerufen: 05.04.2023.

#### 5.6.4 Fazit

Mit Mining Watch hat die belgische Aufsichtsbehörde zur Kontrolle von Schwarzarbeit und Wirtschaftsbetrug ein hilfreiches Tool, um Inspektor\*innen bei der Entscheidung, welche Unternehmen zu kontrollieren sind, zu unterstützen. Die von Mining Watch erstellten Empfehlungen basieren auf der Auswertung von Unternehmensmerkmalen mithilfe von Machine-Learning-Methoden. Das Tool wird von den Inspektor\*innen intensiv genutzt und hat die Aufdeckungsquote von nicht gesetzeskonformen Verhalten von Betrieben erhöht. Für Deutschland sind vergleichbare Auswertungen denkbar.

#### 5.6.5 Quellen des Steckbriefs

Europäische Kommission (o. D.). *Undeclared Work: MiningWatch, Belgium* [Bericht]. Europäische Plattform. Abgerufen am 20. März 2023 von

<https://ec.europa.eu/social/BlobServlet?docId=18372&langId=en>

Williams, C. (2021). Miningwatch Data Analytics Tool, Belgien. Technische Dokumentation.

[https://www.researchgate.net/publication/351820008\\_MININGWATCH\\_DATA\\_ANALYTICS\\_TOOL\\_BELGIUM](https://www.researchgate.net/publication/351820008_MININGWATCH_DATA_ANALYTICS_TOOL_BELGIUM)

## 5.7 Automatisierte Kategorisierung von Unfällen am Arbeitsplatz (USA)

### 5.7.1 Einführung

Das Bureau of Labor Statistics (BLS) in den USA führt Statistiken über Arbeitsunfälle und deren Hergang, um Arbeitsunfällen besser entgegenwirken zu können. Die Statistiken basieren auf Daten, die aus Unfallberichten von Betrieben gewonnen werden. Vor der Implementierung des im Folgenden beschriebenen Verfahrens wurden diese von den Betrieben übersendeten Textdokumente manuell kodiert und katalogisiert, bevor sie zu statistischen Zwecken genutzt wurden.<sup>36</sup> Die manuellen Zwischenschritte erwiesen sich dabei jedoch als sehr fehleranfällig. Deshalb automatisiert das BLS inzwischen einen Großteil der für die Codierung notwendigen Schritte. Measure (2017) zeigt, dass der Algorithmus für den Arbeitsschritt des Kodierens 39 Prozent weniger Fehler macht als eine manuelle Auswertung. Dadurch werden zwei Ziele erreicht: Zum einen sinkt durch die Automatisierung der Personaleinsatz,<sup>37</sup> zum anderen werden die Ergebnisse präziser und die auf dieser Basis entwickelten Präventionsmaßnahmen sind wirkungsvoller. Das Projekt wurde 2014 im BLS implementiert und seitdem stetig erweitert. Inzwischen werden ca. 90 Prozent aller Berichte zu Arbeitsunfällen automatisch kodiert (Measure, o. D.).

Der Kern der Innovation ist das Einlesen und Verarbeiten von Textdokumenten mithilfe von Machine-Learning-Methoden. Grundsätzlich ist die Innovation nicht ausschließlich für das Erfassen von Arbeitsunfällen anwendbar, sondern in allen Prozessen, die Informationen zu bestimmten, vorher festgelegten Kriterien aus einem Text gewinnen sollen. Insofern ist eine Übertragbarkeit weit über den konkreten Fall der Kategorisierung von Arbeitsunfällen denkbar.

### 5.7.2 Technische Beschreibung und Anwendung

In den USA erstatten Betriebe nach einem Arbeitsunfall über dessen Hergang und Folgen Bericht. Diese Berichte sind die Basis für die Bildung von mehreren übergeordneten Merkmalskategorien, die später in den Statistiken des BLS abgebildet werden. Zu den Merkmalskategorien gehören die Berufsklassifikation der verletzten Person, die Verletzungsart, das betroffene Körperteil, der Unfallhergang und die beteiligten Personen.

Zunächst müssen die in Textform vorliegenden Unfallberichte kodiert werden. Seit dem Jahr 2014 werden die Texte der Unfallanzeigen mithilfe von Machine-Learning-Methoden in Kodierungen umgewandelt. Dieses Verfahren wird als Natural Language Processing (NLP) bezeichnet und ist eine weit verbreitete Methode des Machine Learning. Mithilfe des NLP werden die Informationen so kodiert, dass jede relevante Information der richtigen Merkmalskategorie zugeordnet wird (z. B. „Zerrung“ der Kategorie „Verletzungsart“). Die Machine-Learning Verfahren werden detailliert in Measure (2017) beschrieben.

Zur Erkennung der relevanten Textelemente und der daraus resultierenden Klassifikation sind zunächst Trainings- und Testdaten notwendig, auf deren Basis der Algorithmus lernt, einzelne Wörter und Formulierungen in Kodierungen umzuwandeln. Die bei der erstmaligen Kodierung verwendeten Trainings- und Testdaten bestanden aus über einer Million Berichten zu Arbeitsunfällen, die dem BLS aus der Vergangenheit vorlagen. Da die Trainings- und Testdaten allerdings fehlerhaft waren, wurden diese zur Vorbereitung für das Trainieren und Testen des NLP-

---

<sup>36</sup> Die Statistiken werden unter der Bezeichnung Survey of Occupational Injuries and Illnesses (SOII) veröffentlicht.

<sup>37</sup> Nach Angaben in Measure (o. D.) werden innerhalb des BLS 25.000 Arbeitsstunden für das reine Kodieren benötigt. Dabei sind Arbeitsstunden zur Korrektur von Fehlern nicht inbegriffen.

Algorithmus zunächst von je drei Beschäftigten des BLS kodiert und anschließend bei nicht-identischen Kodierungen durch Expert\*innen des BLS überarbeitet (Measure, 2017).

Die Daten werden jedes Jahr um die neu dazugekommenen Arbeitsunfälle ergänzt. Die Kodierungen des Algorithmus werden jährlich durch eine manuelle stichprobenartige Kontrolle überprüft, um systematische Fehler zu identifizieren. Zudem sind manuelle Eingriffe erforderlich, wenn neue Arten von Arbeitsunfällen entstehen. Denn diese sind im Trainings- und Testdatensatz zunächst nicht enthalten und benötigen dementsprechend so lange manuelle Kodierungen, bis diese einen ausreichenden Umfang zur Einbindung in das NLP haben (Measure, 2017). Beispielsweise musste im Zuge der Covid-19-Pandemie das Wort „Covid“ gelernt werden, da eine Covid-19-Erkrankung in den USA teilweise als Arbeitsunfall gilt. Entsprechend mussten zu Beginn der Pandemie die dem Algorithmus unbekannt Fälle gesammelt und manuell kodiert werden. In der Regel werden bei solchen manuellen Eingriffen zunächst Verbote in den Algorithmus eingebaut, um ein falsches Kodieren zu unterbinden. Ähnliche Begriffe oder Kodierungen, die potenziell aus dem Wort „Covid“ resultieren könnten, werden nicht automatisch kodiert, sondern müssen in einer manuellen Schleife bearbeitet werden.

In allen sechs Merkmalskategorien sind die Kodierungen der Berichte über Arbeitsunfälle durch den NLP-Algorithmus präziser als die manuell erstellten Kodierungen, die vor Implementierung der Standard im BLS waren. Die routinemäßigen manuellen Kodierungen, wie sie vor der Innovation stattfanden, hatten im Vergleich zu den dreifach durchgeführten Kodierungen der Expert\*innen, die den Benchmark für den Algorithmus darstellen, je nach Kategorie eine Präzision von 52 Prozent bis 84 Prozent, während der Algorithmus mit 70 Prozent bzw. 92 Prozent deutlich zuverlässiger kodiert.

### 5.7.3 Übertragbarkeit auf Deutschland

Ähnlich wie in den USA müssen Arbeitsunfälle in Deutschland durch die Arbeitgeber beim Träger der jeweiligen gesetzlichen Unfallversicherung (UV) gemeldet werden.<sup>38</sup> Das dafür zu übermittelnde Formular enthält Merkmale zum Betrieb, zum Unfall sowie zur betroffenen Person und muss vom Arbeitgeber ausgefüllt werden. Zentral ist dabei ein Freitextelement, in dem der genaue Unfallhergang beschrieben wird. Aus diesem werden später die Informationen extrahiert, die von der Deutschen Gesetzlichen Unfallversicherung (DGUV) für Statistiken zu Arbeitsunfällen aufbereitet werden. Ein zweiter Weg der Meldung eines Arbeitsunfalls führt über Durchgangsärzt\*innen<sup>39</sup>, die ebenfalls Details zur Person und zum Unfallhergang an den betreffenden Träger der Unfallversicherung melden. Auch hier wird der Hergang des Arbeitsunfalls in Form eines Freitextes beschrieben.<sup>40</sup> Diese Art der Meldung findet nur dann statt, wenn die Meldung durch den Arbeitgeber nicht vorgenommen wurde.

Ist die Meldung beim jeweiligen Träger der Unfallversicherung eingegangen, läuft das weitere Vorgehen nach trägerspezifischen Standards ab. In der Regel wird eine Stichprobe der Meldungen, die normalerweise zwischen sechs Prozent und zehn Prozent liegt, nach einem Schlüsselkatalog (Schlüsselhandbuch zur Unfallanzeige) kodiert. Der Schlüsselkatalog ist dabei im Rahmen der

---

<sup>38</sup> In Deutschland gibt es circa 30 Unfallversicherungsträger, die teilweise branchenspezifisch organisiert sind.

<sup>39</sup> Durchgangsärzt\*innen sind die erste Anlaufstelle nach einem Arbeitsunfall und entscheiden über das weitere Verfahren zur Heilung (DGUV, o.D.-a)

<sup>40</sup> Nach Auskunft eines Experten sind die Beschreibungen der Durchgangsärzt\*innen deutlich standardisierter als die Meldungen der Arbeitgeber\*innen. Dies ist insofern relevant, als das die Informationsextraktion durch einen Algorithmus von der Komplexität des Textes abhängt. Präzisere und kürzere Texte sind dabei weniger fehleranfällig.

Europäischen Statistik über Arbeitsunfälle standardisiert.<sup>41</sup> Es werden beispielsweise Merkmale wie der Ort, die Tätigkeit beim Unfall, mögliche Abweichungen vom normalen Ablauf der Tätigkeit sowie die Art und Schwere der Verletzungen in vierstellige Kodierungen übersetzt. Die Kodierung wird von Beschäftigten der Unfallversicherung durchgeführt und ist trotz der Verwendung nur einer Stichprobe aller Meldungen sehr zeitintensiv.

Das in Deutschland bestehende System der Meldung von Arbeitsunfällen bietet ähnliche Innovationspotenziale wie das System in den USA. Unmittelbare Vorteile durch einen geringeren Ressourceneinsatz hätten jeweils die einzelnen UV-Träger. Allerdings sind drei Gegebenheiten des in Deutschland angewendeten Systems der Unfallmeldungen zu berücksichtigen, die Hemmnisse für eine Automatisierung darstellen könnten. Erstens wird der Schlüsselkatalog, wie in den USA, laufend um neue Merkmale erweitert. Die Innovation aus den USA zeigt, dass dadurch weiterhin manuelle Eingriffe notwendig sind und die Zeitersparnis eines automatisierten Kodierungssystems reduziert wird. Zudem ist der Freitext in Deutschland deutlich weniger strukturiert als in den USA, wo die einzelnen Elemente des Unfalls im Detail in eine Kodierung übersetzt werden müssen (Measure, 2017). In Deutschland könnten jedoch die Meldungen der Durchgangsärzt\*innen herangezogen werden, die durch die Routine und einheitliche Fachtermini stärker strukturiert und damit einfacher zu verarbeiten sind. Drittens müsste aufgrund der verschiedenen trägerspezifischen Prozesse und Datenspeicherungen geprüft werden, inwieweit der Umfang der jeweils vorliegenden Stichproben der Unfallmeldungen als Grundlage der Trainings- und Testdaten ausreicht.

Nach Einschätzung von Expert\*innen sind die für die Innovation notwendigen Machine-Learning-Methoden ausreichend erforscht und könnten mit angemessenem Aufwand im hier beschriebenen Anwendungsfall implementiert werden. Dabei sollte von Beginn an darauf geachtet werden, die Trainings- und Testdaten sorgfältig zu pflegen, um die Wahrscheinlichkeit von systematischen Fehlern über die Zeit zu reduzieren.

#### 5.7.4 Fazit

Zur Erstellung von Statistiken zu Arbeitsunfällen nutzt das BLS Machine-Learning-Methoden und wendet diese auf Unfallmeldungen an, um daraus die erforderlichen Informationen zu extrahieren und zu kodieren. Neben der schnelleren Verarbeitung der gemeldeten Unfälle führt die automatisierte im Vergleich zur manuellen Kodierung zu genaueren Ergebnissen. Es sind aber weiterhin manuelle Prozesse nötig, um neue Unfallarten korrekt zu erfassen und die Qualität der Kodierungen zu gewährleisten.

Da in Deutschland die Prozesse innerhalb der Unfallversicherung ähnlich zu denen in den USA sind und die Machine-Learning-Methoden inzwischen eine hohe Qualität haben, hat die beschriebene Dateninnovation aus den USA ein hohes Transferpotenzial. Zentral ist dabei allerdings zum einen die Datenqualität der vorhandenen Kodierungen, die in Deutschland bisher nicht systematisch erfasst wird. Zum anderen wird eine ausreichend große und einheitliche Datengrundlage benötigt, was in Deutschland aufgrund des komplexen Systems der Unfallversicherungsträger problematisch sein könnte.

Grundsätzlich kann diese Innovation auch in andere Prozesse integriert werden, bei denen eine Informationsextraktion aus Text notwendig ist. Diese dürfte für viele Geschäftsprozesse in Behörden der Arbeitsmarkt- und Sozialverwaltung der Fall sein. Mit ADEST entwickelt beispielsweise die BA derzeit ein innovatives und nützliches Tool (siehe Abschnitt 2.1).

---

<sup>41</sup> Weitere Informationen sind in Eurostat (2013) enthalten.

### 5.7.5 Quellen des Steckbriefs

DGUV. (o. D.-a). Durchgangsarztverfahren. Deutsche Gesetzliche Unfallversicherung. Abgerufen am 5. April 2023 von [https://www.dguv.de/landesverbaende/de/med\\_reha/d-arzt-verfahren/index.jsp](https://www.dguv.de/landesverbaende/de/med_reha/d-arzt-verfahren/index.jsp)

Eurostat (2013). Europäische Statistik über Arbeitsunfälle (ESAW) [Methodik-Zusammenfassung]. Eurostat Methodologies & Working Papers. <https://doi.org/10.2785/40882>

Measure, A. (2017). *Deep neural networks for worker injury autocoding* [WP-Entwurf vom 18. September 2017]. U.S. Bureau of Labor Statistics, Working Paper. <https://doi.org/10.13140/RG.2.2.30587.41765>

Measure, A. (o. D.). *Automated Coding of Injury and Illness Data*. U.S. Bureau of Labor Statistics. Abgerufen am 23. März 2023 von <https://www.bls.gov/iif/automated-coding.htm>

## 5.8 Sandbox: Ein Reallabor für die Nutzung von Machine-Learning-Methoden (Norwegen)

### 5.8.1 Einführung

In Norwegen können sich Unternehmen, die mit sensiblen Daten arbeiten und für ein Produkt oder eine Dienstleistung Machine-Learning-Methoden nutzen, seit dem Jahr 2020 auf die Sandbox (Sandkasten, norwegisch: Sandkasse, in Deutschland auch als Reallabor bezeichnet) bewerben, die von der Datenschutzbehörde Datatilsynet entwickelt wurde und dort als Pilotprojekt angegliedert ist. In der Sandbox werden in engem Austausch mit Datatilsynet Konflikte mit dem Datenschutz untersucht, die sich für Unternehmen bei der Anwendung von Machine-Learning-Methoden auf sensible Daten ergeben. Zugelassene Unternehmen werden bei der Entwicklung ihrer Anwendung von einem Expert\*innenteam der Sandbox begleitet, das laufend den Datenschutz, ethische Gesichtspunkte und die Sicherheit aus der Perspektive der Gesetzgebung und Gesellschaft bewertet und prüft. Seit dem Projektstart wurden 14 Unternehmen mit ihren Anwendungen in der Sandbox betreut.

Die Zusammenarbeit zwischen Unternehmen und Sandbox-Team hat für beide Seiten Vorteile. Auf der einen Seite können Unternehmen Datenschutzkompetenzen erwerben. Darüber hinaus können sie in Fällen, in denen ihre in der Entwicklung befindlichen Produkte und Dienstleistungen nicht mit dem Datenschutz übereinstimmen, konkrete Lösungen erarbeiten. Auf der anderen Seite kann die Datenschutzbehörde in der Sandbox aktuelle Entwicklungen im Bereich der Anwendung von Machine Learning begleiten. Dadurch können potenzielle Datenschutzverletzungen oder sonstige Risiken frühzeitig erkannt werden. Aufgrund dieser Erfahrungen kann das gesetzliche Regelwerk permanent weiterentwickelt werden, sodass der Datenschutz und gesellschaftliche Interessen berücksichtigt und zugleich die Innovationskraft der Unternehmen gefördert werden. Neben den Vorteilen für die beteiligten Unternehmen und die Datenschutzbehörde gibt es auch Vorteile für die Öffentlichkeit, deren Vertrauen in Anwendungen, die mit Machine-Learning-Methoden arbeiten, gestärkt wird.

Für Deutschland liegt das Innovationspotenzial von Reallaboren im Bereich der Arbeitsmarkt- und Sozialpolitik insbesondere in der Förderung neuer Anwendungen, die sensible Daten nutzen, welche bisher nur eingeschränkt oder gar nicht genutzt werden konnten. Zudem kann die Kommunikation der Erkenntnisse zu einem gestiegenen Vertrauen der Bevölkerung in Datenanwendungen beitragen, die Machine-Learning-Methoden anwenden.

### 5.8.2 Technische Beschreibung

#### 5.8.2.1 Bewerbungsprozess

Die Sandbox begleitet jährlich etwa drei bis fünf Unternehmen. Die Unternehmen werden zunächst in einem Bewerbungsprozess bewertet und für die Sandbox zugelassen. Dabei gibt es verschiedene Kriterien, anhand derer Datatilsynet die Bewerbungen evaluiert. Zunächst muss die Anwendung einen Bezug zu Machine-Learning-Methoden haben. Dabei kann die Methode auch nur ein Teil eines Produkts sein, wie beispielsweise in Robotern, die mithilfe von Machine Learning gesteuert werden. Zudem ist Datatilsynet an den Implikationen einer solchen Anwendung für den Datenschutz interessiert, weswegen ein Bezug zu sensiblen Daten, wie beispielsweise Gesundheitsdaten, innerhalb der Anwendung gegeben sein sollte. Ein weiteres Kriterium ist der Mehrwert der Prüfung für die Gesellschaft. Ein solcher Mehrwert ergibt sich zum Beispiel bei Anwendungen im Gesundheits- oder Sozialbereich. Insbesondere sollen begleitete Anwendungen Erfahrungswerte für andere

Unternehmen schaffen. So wird im Endbericht eines abgeschlossenen Projekts auf fördernde Faktoren oder Hürden verwiesen, die sich auch in ähnlichen Anwendungen zeigen könnten.

Die beteiligten Unternehmen sind größtenteils in einer frühen Entwicklungsphase ihrer Anwendung und kommen vorwiegend aus dem privaten Sektor. Die Ansprache potenzieller Unternehmen fand insbesondere in der ersten Bewerbungsrunde der Sandbox über Pressearbeit statt. Inzwischen hat sich das Projekt in Norwegen etabliert und es gibt jedes Jahr ca. 25 Unternehmen, die sich für die Sandbox bewerben.

### 5.8.2.2 Prozesse in der Sandbox

Die Begleitung eines Unternehmens dauert in der Regel drei bis sechs Monate. Während dieser Zeit wird das Unternehmen von Expert\*innen von Datatilsynet aus verschiedenen Fachrichtungen begleitet. Datatilsynet bezieht dabei drei verschiedene Gruppen von Expert\*innen ein. Die erste Gruppe besteht aus Jurist\*innen aus der norwegischen Datenschutzbehörde, die unterschiedliche Spezialisierungen haben und anhand ihrer Kenntnisse und Erfahrungen in ein passendes Projekt in der Sandbox eingebunden werden. Die zweite Gruppe sind Expert\*innen aus dem Bereich Technologie und Machine Learning. Die dritte Gruppe besteht aus externen Expert\*innen, die bei konkretem Bedarf hinzugezogen werden. Dazu gehören z. B. Expert\*innen aus der Diskriminierungs- oder der Finanzbehörde Norwegens. Zu Beginn eines Projekts wird geprüft, welche Fachrichtungen besonders relevant sind, sodass für jedes Projekt ein individuelles Expert\*innenteam zur Verfügung steht.

Die Zusammenarbeit zwischen den Expert\*innen der Sandbox und den Unternehmen findet während der Projektphase in vier bis fünf Workshops statt. Dabei unterscheiden sich Teilnehmendenzahl und Dauer je nach Bedarf. In den Workshops werden rechtliche oder technische Einzelheiten besprochen sowie der projektbezogene Endbericht verfasst. Die Diskussionen folgen keiner festgelegten Vorlage, sondern hängen davon ab, in welchen Bereichen ein Unternehmen Hilfe wünscht. Die bisher begleiteten Unternehmen bringen vorwiegend Anliegen in die Sandbox, die in die Bereiche Transparenz der Algorithmen und Ergebnisse, Diskriminierungsverzerrungen der Modelle und Datenverarbeitungsprozesse fallen, und erhalten in diesen Bereichen juristische und technische Beratung. Über diese inhaltliche Zusammenarbeit hinaus wird von Datatilsynet keine technische oder finanzielle Unterstützung gegeben. Im Gegensatz zu ähnlichen Konzepten einer Sandbox in anderen Ländern bzw. Themenbereichen, wie beispielsweise Sandboxes im Finanzsektor, finden in der hier beschriebenen Sandbox die üblichen Vorschriften und Gesetze Anwendung. In anderen Sandboxes bzw. Reallaboren ist es demgegenüber möglich, unter bestimmten Bedingungen den Rechtsrahmen flexibel auszulegen.<sup>42</sup>

Die Betreuung in der Sandbox ergibt keine rechtsverbindlichen Auskünfte. Es ist theoretisch möglich, dass die Datenschutzbehörde Datatilsynet bei einer Anwendung, die zuvor die Sandbox durchlaufen hat, im Nachgang Verstöße gegen bestehende Datenschutzvorschriften feststellt. Bisher ist dieser Fall jedoch nicht eingetreten.

### 5.8.2.3 Anwendungsbeispiel

Die Unternehmen und deren Anwendungen umfassen viele verschiedene Themen und Zielgruppen. Das Unternehmen Simplifai hat beispielsweise einen Algorithmus entwickelt, der Beschäftigten im öffentlichen Sektor hilft, Emails korrekt zu organisieren. Das Ziel ist die Vermeidung von Informationsverlust aufgrund falscher Archivierung von Emails. Dabei ergeben sich allerdings

---

<sup>42</sup> Für exemplarische Anwendungen in Deutschland, siehe Bundesministerium für Wirtschaft und Energie (2019).

Datenschutzfragen. Insbesondere stehen hier rechtliche Bedenken aufgrund der Verarbeitung der potenziell sensiblen Inhalte der Emails im Vordergrund, wenn es also um die Verarbeitung personenbezogener Daten wie Krankheitsmeldungen geht. Sollte der Algorithmus nicht auf diese personenbezogenen Daten zugreifen dürfen, wäre der Mehrwert der Anwendung deutlich geringer. Im Verlauf der Sandbox bewerteten Jurist\*innen, ob der Algorithmus nach den Bestimmungen der europäischen Datenschutzgrundverordnung (DSGVO) auf die personenbezogenen Daten für die Zwecke der Weiterverarbeitung zugreifen darf. Die DSGVO gilt seit 2018 auch in Norwegen. Des Weiteren wurden Fragen der Verantwortlichkeit der Gewährleistung des Datenschutzes in der Sandbox diskutiert und geprüft.

In einem projektbezogenen Abschlussbericht werden die wesentlichen Ergebnisse der Sandbox dargestellt und Quellen zusammengefasst, die im jeweiligen Rechtsbereich als relevant betrachtet werden. Beispielsweise enthält der Abschlussbericht des Unternehmens Simplifai Erläuterungen zur Privacy-by-Design-Architektur, die im Projekt eingearbeitet wurden. Die Abschlussberichte sind im Internet frei zugänglich verfügbar.

Für die Zukunft sind Anpassungen der Sandbox geplant. Die Förderphase, in der sich die Sandbox aktuell befindet, läuft im Jahr 2023 aus. Da großes Interesse am Fortbestehen der Sandbox besteht, soll sie nun direkt in die norwegische Datenschutzbehörde eingegliedert werden. Im Zuge dessen soll sich allerdings der Fokus von rein auf den Datenschutz bezogenen Themen im Kontext von Machine Learning zu einem breiteren Konzept im Bereich der Digitalisierung und Innovation verschieben, um für ein breiteres Spektrum von Unternehmen relevant zu werden.

Insgesamt sollen die Projektphasen deutlich flexibler gestaltet werden. So wird von Datatilsynet geplant, verschiedene Formate der Begleitung anzubieten. Dabei sind auch Formate von beispielsweise nur einem Tag denkbar.

#### 5.8.2.4 Ergebnisse

Inhaltlich haben die einzelnen Projekte konkrete Folgewirkungen erzielt. Beispielsweise stellten die Beteiligten während eines Projekts in der Sandbox fest, dass es für die betrachtete Anwendung gesetzlich erlaubt ist, Machine-Learning-Methoden anzuwenden, dass die dafür benötigten Daten jedoch nach geltendem Gesetz nicht zum Training der Modelle verwendet werden dürfen. Infolgedessen wurde eine Gesetzesänderung geplant, die demnächst implementiert werden soll. Die Ergebnisse jedes Projekts werden dabei im projektbezogenen Endbericht detailliert beschrieben und so für die Öffentlichkeit zugänglich gemacht. Im Endbericht werden rechtliche, technische und ethische Aspekte der konkreten Anwendung und dessen Merkmale beleuchtet, die für Unternehmen mit ähnlichen Anwendungen ebenfalls relevant sind. Konkrete Rückschlüsse auf einzelne Elemente des Geschäftsmodells werden dabei nicht im Endbericht veröffentlicht, um die Geschäftsgrundlage der teilnehmenden Unternehmen nicht zu beeinträchtigen. Datatilsynet hat dabei die Rückmeldung bekommen, dass der Endbericht insbesondere für andere Institutionen und Unternehmen ein wertvolles Produkt ist.

Viele Unternehmen nutzen die Sandbox in einem frühen Stadium der Entwicklung von Produkten und Dienstleistungen. Dies stellte sich auch als notwendig heraus, da die Anwendungen beim Projektstart noch flexibler sind. Bei einigen Projekten wurden unüberwindliche Hindernisse festgestellt, die nur durch eine grundsätzliche Änderung des Geschäftsmodells umgangen werden konnten. Die Teilnahme an der Sandbox in einem frühen Stadium führte auch dazu, dass die Kooperationsprojekte zwischen Unternehmen und der Sandbox nicht abgebrochen wurden. Falls größere Hindernisse festgestellt wurden, waren die Unternehmen in der Lage, ihre Anwendung oder das Geschäftsmodell so zu ändern, dass am Ende des Projekts eine Lösung gefunden wurde.

Für Datatilsynet stellte es sich, besonders direkt nach der Einführung der Sandbox, als schwierig dar, junge Unternehmen zu erreichen, da diese das Angebot der Sandbox nicht kannten. Erst im Lauf der Zeit erhöhte sich die Reichweite der Sandbox und junge Unternehmen wurden rechtzeitig auf die Sandbox aufmerksam. Daher sind die Kommunikation und Öffentlichkeitsarbeit besonders in der Zielgruppe der jungen Unternehmen für die Sandbox eine wichtige Aufgabe.

Die in der Sandbox zwischen den Unternehmen und Datatilsynet diskutierten Fragen bezogen sich größtenteils auf grundlegende Probleme im jeweiligen Anwendungsgebiet und weniger auf technische Einzelfragen. Nach Einschätzung von Expert\*innen war der Erkenntnisgewinn hinsichtlich des Datenschutzes im Kontext von Machine Learning für Datatilsynet deutlich höher als erwartet.

### 5.8.3 Ähnliche Anwendungen in anderen Ländern

Testräume wie die norwegische Sandbox werden im Englischen unter dem Begriff „Regulatory Sandboxes“ zusammengefasst und im deutschen Sprachgebrauch meist als „Reallabore“ bezeichnet.<sup>43</sup> Dazu gehören die Sandboxes im Vereinigten Königreich (ico, o. D.) und Spanien (Europäische Kommission, 2022). In Frankreich gibt es eine Sandbox im Bereich Energie und Innovation (cre, 2022). Des Weiteren gibt es viele Sandboxes im Finanzbereich, die auch Ausnahmen von den im Finanzwesen geltenden Vorschriften anbieten, um innovative Produkte zu testen (World Bank, 2020).

Hervorzuheben ist die spanische Sandbox, die im Rahmen eines EU-Pilotprojekts entwickelt wurde. Dort sollen insbesondere die nach EU-Recht geltenden Vorschriften auf ihre Praxistauglichkeit getestet werden und Erkenntnisse für Verbesserungsansätze liefern (Europäische Kommission, 2022). Damit soll die spanische Sandbox Grundlage für eine EU-weite Reihe von nationalen Sandboxes werden, die in Zukunft aufgebaut werden sollen. Dazu hat der Rat der EU im Jahr 2020 eine Schlussfolgerung angenommen, die den Mehrwert eines experimentellen Rahmens, wie es eine Sandbox verspricht, anerkennt. Die EU-Kommission soll daher sogenannte Experimentierklauseln, also im Einzelfall eine mögliche Abweichung von geltenden Regelungen, diskutieren und evaluieren (Europäischer Rat, 2020). Wie bereits oben beschrieben gibt es in der norwegischen Sandbox bisher keine solche Experimentierklauseln.

Des Weiteren werden im Entwurf zum Artificial Intelligence Act des Europäischen Parlaments (IMCO & LIBE, 2022) Anforderungen und Ziele eines gesetzlichen Rahmens für Sandboxes im Bereich der KI formuliert. Für Details sei auf Madiaga & Van De Pol (2022) und die dortigen Quellen verwiesen (siehe auch: Europäischer Datenschutzausschuss, 2021).<sup>44</sup> Aufgrund dieser Entwicklungen ist damit zu rechnen, dass es auch in Deutschland in absehbarer Zeit einen Rechtsrahmen für Sandboxes und damit gute Voraussetzungen für einen Transfer nach Deutschland gibt.

### 5.8.4 Übertragbarkeit auf Deutschland

In Deutschland gibt es noch keine datenschutzfokussierte Sandboxes oder Reallabore im Bereich der Arbeitsmarkt- und Sozialpolitik.<sup>45</sup> Das Interesse an einer solchen Sandbox könnte speziell unter Start-Ups und Ausgründungen von Forschungsprojekten mit datenorientiertem Geschäftsmodell groß sein, denn diese zeigten bereits in der Vergangenheit ein erhöhtes Interesse an beratenden Stellen für ihre in der Entwicklung befindlichen Produkte und Anwendungen. Bezüglich des

---

<sup>43</sup> Neben den an den Behörden angegliederten Sandboxes gibt es auch kommerzielle Sandboxes.

<sup>44</sup> Siehe Legislative Observatory (o.D.) für aktuelle Entwicklungen des Artificial Intelligence Acts.

<sup>45</sup> Reallabore werden jedoch in anderen Themenbereichen genutzt, beispielsweise im Energiesektor (<https://www.energieforschung.de/spotlights/reallabore>, zuletzt abgerufen: 01.06.2023).

Innovationspotenzials einer Sandbox in Deutschland im Bereich der Arbeitsmarkt- und Sozialpolitik wäre es insbesondere interessant, etwaige Anwendungsfälle in Deutschland unter Berücksichtigung der Vorgaben des Sozialgesetzbuches, der Datenschutz-Grundverordnung (DSGVO) und weiterer jeweils einschlägiger rechtlicher Vorgaben zu eruieren.

Ein Transfer der norwegischen Sandbox nach Deutschland wird insbesondere durch das föderale System der Aufsichtsbehörden erschwert. Auf Bundesebene ist der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit für öffentliche Stellen des Bundes und Unternehmen, die Telekommunikations- oder Postdienstleistungen erbringen, zuständig. Sonstige privatwirtschaftliche Unternehmen und öffentliche Stellen der Länder liegen dagegen im Wesentlichen im Zuständigkeitsbereich der jeweiligen Landesdatenschutzbehörden.<sup>46</sup> Dezentrale Sandboxes hätten dabei den Nachteil, dass die Prozesse und Kommunikation zwischen den dezentralen Sandboxes eng abgestimmt werden müssen, um ein koordiniertes Vorgehen und eine einheitliche Beratungs- und Regelungspraxis zu schaffen. Zu prüfen wäre, ob durch einen Zusammenschluss von Datenschutzaufsichtsbehörden eine zentrale Anbindung von Sandboxes möglich wäre. Damit könnten sowohl öffentliche als auch private Einrichtungen angesprochen und die datenschutzrechtlichen Rahmenbedingungen einer etwaigen Sandbox koordiniert werden.

Allerdings müssten bei einer Entwicklung durch beziehungsweise Angliederung an die Datenschutzbehörden mögliche Interessenkonflikte berücksichtigt werden.<sup>47</sup> Bereits heute ist die Beratungstätigkeit im Rahmen der jeweiligen gesetzlichen Zuständigkeiten eine Aufgabe der Datenschutzbehörden (Art. 57 DSGVO). Bei einer engen Einbindung der Datenschutzbehörden sollte allerdings berücksichtigt werden, dass es derzeit keine normativen Anweisungen oder Anleitungen dazu gibt, wie es sich in Einklang bringen lässt, dass die Datenschutzbehörden einerseits Aufsichts- und Abhilfebefugnisse haben und andererseits im Rahmen von Reallaboren detaillierte Anleitungen erteilen sollen. Wichtig für die Ausgestaltung der Sandbox ist daher, dass die Erfüllung der datenschutzrechtlichen Anforderungen in jedem Fall der verantwortlichen Stelle, hier also den „Nutzenden“ der Sandbox obliegt.

### 5.8.5 Fazit

Die Sandbox der norwegischen Datenschutzbehörde Datatilsynet bietet Unternehmen die Möglichkeit, Unterstützung von einem Expertenteam für ihre Produkte bzw. Anwendungen im Bereich sensibler Daten und Machine Learning zu erhalten. Dabei zeigt sich, dass eine frühe Nutzung dieser Dienstleistung noch in der Entwicklung der Geschäftsidee erfolgsversprechend ist. Bislang wurden in der Sandbox sehr unterschiedliche Projekte betreut und die dabei gewonnenen Erkenntnisse in der Öffentlichkeit publik gemacht, sodass ein Mehrwert auch für externe Interessierte geschaffen wurde.

Die Idee der Sandbox ist grundsätzlich nicht neu und wird besonders häufig im Finanzsektor angeboten. In Bezug auf den Datenschutz und Machine-Learning-Methoden haben unter anderem das Vereinigte Königreich und Spanien ebenfalls solche Sandboxes. Die spanische Sandbox sammelt aktuell Erfahrungen für eine EU-weite Entwicklung nationaler Sandboxes, welche im Rahmen des Artificial Intelligence Act in europäisches Recht eingebettet werden sollen. Im Rahmen dieser Initiative wäre es auch für Deutschland möglich, eine Sandbox aufzubauen. Dadurch erhalten die

---

<sup>46</sup> Schnittmengen gibt es allerdings zum Beispiel bei technischen Anwendungen von Unternehmen der Privatwirtschaft, deren Anwendungen von öffentlichen Institutionen bzw. Einrichtungen genutzt werden. Beispielhaft sind hier Cloud-Dienste zu nennen.

<sup>47</sup> Diese Interessenskonflikte bestehen auch in der norwegischen Sandbox, wobei aktuell kein Fall bekannt ist, bei dem Unternehmen nach dem Prozess in der Sandbox in Konflikt mit der norwegischen Datenschutzaufsichtsbehörde kamen.

Unternehmen Kenntnisse und Handlungssicherheit, was ihre Innovationsfähigkeit erhöht. Zwar besteht nicht über alle Unternehmen hinweg Interesse an einer Sandbox, einzelne Unternehmen haben allerdings schon in der Vergangenheit Interesse an einer Begleitung durch ein Expertenteam bekundet. Deren allgemeine Erkenntnisse im Rahmen einer Sandbox könnten zudem für alle Unternehmen einen Mehrwert schaffen, deren Geschäftsmodell auf Anwendung von Machine Learning basiert. Durch eine Sandbox wird darüber hinaus in der Gesellschaft das Vertrauen in die Anwendung von Machine-Learning-Methoden gestärkt. Nicht zu unterschätzen sind ebenso die Vorteile des internen Kompetenzaufbaus der Datenschutzaufsichtsbehörden in Deutschland.

### 5.8.6 Quellen des Steckbriefs

Bundesministerium für Wirtschaft und Energie (2019). *Freiräume für Innovationen*. Abgerufen am 12. Juni 2023 von [https://www.bmwk.de/Redaktion/DE/Publikationen/Digitale-Welt/handbuch-fuer-reallabore.pdf?\\_\\_blob=publicationFile&v=1](https://www.bmwk.de/Redaktion/DE/Publikationen/Digitale-Welt/handbuch-fuer-reallabore.pdf?__blob=publicationFile&v=1).

cre. (2022). *Regulatory Sandbox*. Commission de Régulation de L'Énergie. Veröffentlicht am 29. November 2022. Abgerufen am 15. Juni 2023 von <https://www.cre.fr/en/Energetic-transition-and-technologic-innovation/regulatory-sandbox>

Europäische Kommission. (2022). *Launch Event for the Spanish Regulatory Sandbox on Artificial Intelligence*. Europäische Kommission. Veröffentlicht am 27. Juni 2022. Abgerufen am 15. Juni 2023 von <https://digital-strategy.ec.europa.eu/en/events/launch-event-spanish-regulatory-sandbox-artificial-intelligence>

Europäischer Rat. (2020). *Regulatorische „Sandkästen“ und Experimentierklauseln als Instrumente für eine bessere Rechtsetzung: Rat nimmt Schlussfolgerungen an*. Rat der Europäischen Union. Veröffentlicht am 16. November 2020. Abgerufen am 15. Juni 2023 von <https://www.consilium.europa.eu/de/press/press-releases/2020/11/16/regulatory-sandboxes-and-experimentation-clauses-as-tools-for-better-regulation-council-adopts-conclusions/>

Europäische Datenschutzausschuss (2021). Gemeinsame Stellungnahme 5/2021 zum Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union 18. Juni 2021, Rdnr. 61 ff. ([https://edpb.europa.eu/system/files/2021-10/edpb-edps\\_joint\\_opinion\\_ai\\_regulation\\_de.pdf](https://edpb.europa.eu/system/files/2021-10/edpb-edps_joint_opinion_ai_regulation_de.pdf), zuletzt abgerufen: 31.07.2018).

ico. (o. D.). *Regulatory Sandbox*. Information Commissioner's Office. Abgerufen am 13. April 2023 von <https://ico.org.uk/sandbox>

IMCO & LIBE (2022). *DRAFT REPORT – on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts* [Report No. COM2021/0206 – C9-0146/2021 – 2021/0106(COD)]. Europäisches Parlament. [https://www.europarl.europa.eu/doceo/document/CJ40-PR-731563\\_EN.pdf](https://www.europarl.europa.eu/doceo/document/CJ40-PR-731563_EN.pdf)

Legislative Observatory. (o. D.). *Artificial Intelligence Acts*. Europäisches Parlament. Abgerufen am 6. April 2023 von [https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2021/0106\(COD\)&l=en](https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2021/0106(COD)&l=en)

Madiega, T. & Van De Pol, A. L. (2022). *Briefing des Europäischen Parlamentarischen Forschungsservices: Artificial intelligence act and regulatory sandboxes* [Briefing PE 733.544].

[https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS\\_BRI\(2022\)733544\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf)

World Bank. (2020, 1. November). *Key Data from Regulatory Sandboxes across the Globe*. The World Bank. Veröffentlicht am 1. November 2020. Abgerufen am 15. Juni 2023 von <https://www.worldbank.org/en/topic/fintech/brief/key-data-from-regulatory-sandboxes-across-the-globe>

## 5.9 Randomisierte Kontrollstudie zu Bewerbungsaktivitäten von Arbeitssuchenden (Dänemark)

### 5.9.1 Einführung

Randomisierte Kontrollstudien (auch als Feldexperimente bezeichnet) stellen nach wie vor den Goldstandard für Evaluationsstudien dar, da sie kausale Effekte transparent und valide identifizieren können. Bei solchen Studien werden die Studienteilnehmenden nach dem Zufallsprinzip in eine Treatment- und eine Kontrollgruppe eingeteilt. Die Intervention, die evaluiert werden soll, wird anschließend nur für die Treatmentgruppe durchgeführt. Durch den Vergleich der Zielvariable für die Treatment- und die Kontrollgruppe können Veränderungen der Zielvariable, die nicht ursächlich auf die Intervention zurückzuführen sind, herausgerechnet und der Effekt der Intervention isoliert werden. Im Themenfeld der Arbeitsmarkt- und Sozialpolitik werden randomisierte Kontrollstudien allerdings weniger häufig eingesetzt als beispielsweise in der Entwicklungs- oder Verhaltensökonomie. Dies kann darauf zurückgeführt werden, dass es in der Arbeitsmarktpolitik oft um Interventionen in der Gesamtbevölkerung geht, in deren Rahmen Kontrollstudien schwieriger durchzuführen sind, da ethische Leitlinien und gesetzliche Normen in vielen Ländern nicht erlauben, solche Interventionen nur für einen Teil der Gesamtbevölkerung durchzuführen (Bodry et al., 2023, S. 4). Stattdessen werden, auch in Deutschland, häufiger quasi-experimentelle Studien, die „natürliche Experimente“ ausnutzen, durchgeführt. Solche Methoden beruhen jedoch teilweise auf stärkeren Annahmen.

Das im Folgenden beschriebene Forschungspapier zeigt beispielhaft, dass sich randomisierte Kontrollstudien grundsätzlich auch für die Evaluation von arbeitsmarktpolitischen Maßnahmen eignen – insbesondere, wenn hierfür umfangreiche administrative Daten zur Verfügung stehen. Das Beispiel soll die generellen Möglichkeiten zum Erkenntnisgewinn für solche Studien in Deutschland darstellen, wenn auch die Datengrundlage, strengere rechtliche Auflagen sowie abweichende institutionelle Ausgestaltungen der Einkommensersatzleistungen eine direkte Übertragung nicht ermöglichen. Zwar liegen auch für Deutschland im Themenfeld der Arbeitsmarkt- und Sozialpolitik bereits einige Studien vor, die auf randomisierten Kontrollstudien beruhen. Allerdings sind solche Studien im genannten Themenfeld in Deutschland weniger verbreitet als in anderen Ländern (Bodry et al., 2023).

Das Feldexperiment von Cairo & Mahlstedt (2021) wurde im Rahmen des Jobnet-Portals (Jobnet, o. D.) in Dänemark (siehe Steckbrief 6.2) durchgeführt. Die Intervention bestand darin, zufällig ausgewählten Arbeitssuchenden im Online-Portal individuell zugeschnittene Informationen über das persönliche Risiko von Leistungsanktionen anzuzeigen, die aus der Verletzung von Pflichten der Arbeitssuchenden resultieren. Im Vergleich zu in ähnlichen Studien durchgeführten Interventionen wurde in der beschriebenen Studie von häufigeren Aktualisierungen dieser Informationen Gebrauch gemacht, sodass die Teilnehmenden zeitnah über Änderungen ihres individuellen Sanktionsrisikos informiert wurden. Die Studienergebnisse zeigen, dass die Bereitstellung dieser Informationen teilweise die Beschäftigungswahrscheinlichkeit erhöhte, teilweise jedoch auch Leistungskürzungen häufiger in Kauf genommen wurden.

### 5.9.2 Technische Beschreibung

Die im Forschungspapier von Cairo & Mahlstedt (2021) beschriebene Intervention bezieht sich auf Informationen zu Sanktionsrisiken im Rahmen von Unterstützungszahlungen für Arbeitssuchende, die dem deutschen Bürgergeld bzw. der Sozialhilfe entsprechen (dänisch: Kontanthjælp, im Folgenden

als Bürgergeld bezeichnet), d. h. die Zahlungen sind an eine Bedürftigkeitsprüfung und an Mitwirkungspflichten gebunden. Die Auszahlungshöhe hängt u. a. vom Alter, von der Anzahl der Kinder, die im gleichen Haushalt leben, sowie vom Einkommen des/der Ehepartner\*in ab. Die Größenordnung der monatlichen Unterstützungszahlungen lag im Jahr 2020 bei circa 1.500 bis 2.000 Euro.

Um das Bürgergeld im Rahmen der sozialen Grundsicherung zu beziehen, ist es in Dänemark verpflichtend, eine bestimmte Mindestanzahl an Stunden pro Jahr zu arbeiten. Andernfalls können Arbeitsuchende dauerhaft ihre Ansprüche auf das Bürgergeld verlieren. In den ersten zwölf Monaten des Leistungsbezugs gibt es keine entsprechenden Pflichten. Ab dem 13. Monat des Leistungsbezugs bestehen die Ansprüche nur fort, sofern die Arbeitsuchenden innerhalb der vorangegangenen zwölf Monate mindestens 225 Stunden in nicht-subventionierten Beschäftigungsverhältnissen aktiv waren, was einer Tätigkeit von fünf Wochenstunden oder einer sechswöchigen Vollzeitbeschäftigung entspricht.

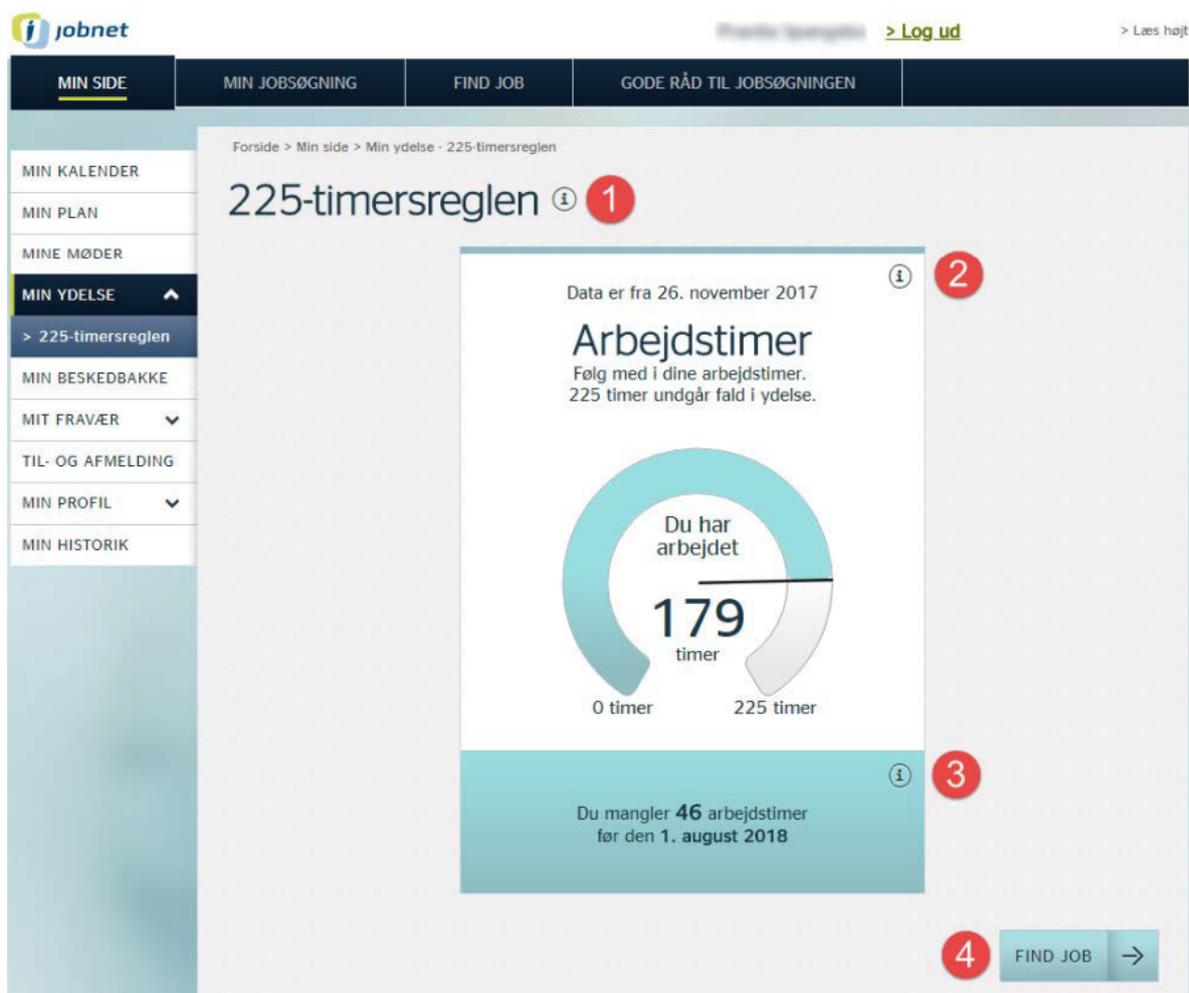
Wenn die Pflicht nicht erfüllt wird, können die Auszahlungen dauerhaft um etwa vier bis sieben Prozent gekürzt werden. Dies betrifft im Durchschnitt etwa 16,5 Prozent der Leistungsbeziehenden. Ausnahmen von Kürzungen können z. B. bei Vorliegen von gesundheitlichen Gründen gestattet werden. Zusätzlich zu den Kürzungen bei einer Verletzung der Pflicht wird der Zählerstand der bis dahin gesammelten Arbeitsstunden für in den vergangenen zwölf Monaten geleistete Arbeitsstunden auf null gesetzt und die Arbeitsstunden müssen neu aufgebaut werden.

Der Sanktionsmechanismus der Leistungskürzungen im Falle einer Verletzung der Pflichten hat zwei Ziele. Zum einen stellt er einen Anreiz für Arbeitsuchende dar, aktiv nach einer neuen Beschäftigung zu suchen und eine neue Beschäftigung anzunehmen. Zum anderen wird eine Selbstselektion herbeigeführt, wenn arbeitsuchend gemeldete Personen keine großen Suchambitionen haben und daher in Kauf nehmen, auf Unterstützungszahlungen zu verzichten.

Für ein funktionierendes Anreizsystem ist allerdings zentral, dass die Betroffenen den Mechanismus, d. h. die zu erfüllenden Pflichten und die Sanktionen im Falle der Nichterfüllung, tatsächlich kennen. Dies war in Dänemark vor dem Feldexperiment für den Großteil der Betroffenen nicht der Fall. Hier setzt das von Cairo & Mahlstedt (2021) durchgeführte Experiment an. Im Rahmen eines Feldexperiments wurden einem Teil der Arbeitsuchenden im Portal Jobnet mithilfe eines Online-Tools in Echtzeit Informationen dazu bereitgestellt, welche Pflichten für den Bezug von Bürgergeld generell bestehen und wie sich diese für den individuellen Fall darstellen. Konkret beinhaltete das Online-Tool tagesaktuelle individuelle Informationen darüber, wie viele Arbeitsstunden eine Person in den zurückliegenden zwölf Monaten bereits geleistet hatte und bis zu welchem Datum sie ggf. noch weitere Stunden leisten musste. Dieses Online-Tool stand jedoch nur Arbeitsuchenden in der Treatmentgruppe zur Verfügung. Diese Gruppe erhielt über Jobnet für einen Zeitraum von sechs Monaten monatliche Benachrichtigungen mit einem Link zu einer Übersicht zu den individuell bereits erbrachten bzw. noch zu erbringenden Arbeitsstunden (siehe Abbildung 5-3).

Die Treatmentgruppe wurde mit zwei verschiedenen Kontrollgruppen verglichen. Die erste Kontrollgruppe erhielt ebenfalls Benachrichtigungen über die allgemeinen Pflichten beim Leistungsbezug, allerdings weder Zugang noch Link zum Online-Tool und damit keine individuellen Informationen zur Erfüllung der Pflichten. Die zweite Kontrollgruppe erhielt keinerlei gesonderte Benachrichtigungen, sondern nur einen Standardbrief mit allgemeinen Informationen zu den Pflichten von der zuständigen Kommunalverwaltung, den alle Leistungsbeziehenden sechs Monate nach der ersten Auszahlung erhalten.

Abbildung 5-3: Online-Tool zur Information über die individuelle Erfüllung von Pflichten



Quelle: Cairo & Mahlstedt (2021). Anmerkungen: Element (1) enthält Informationen zur allgemeinen Pflicht, mindestens 225 Arbeitsstunden innerhalb der zurückliegenden zwölf Monate nachzuweisen. Element (2) zeigt relativ dazu die bis zum aktuellen Tag bereits geleisteten Stunden an. Element (3) enthält das individuelle Datum und einen Countdown-Zähler, bis wann die bis zum Ziel fehlenden Stunden geleistet werden müssen. Element (4) enthält einen Link zur Stellenbörse von Jobnet.

Die Treatment- und Kontrollgruppen wurden per Zufallsprinzip aus der Grundgesamtheit aller Leistungsbeziehenden zum Stichtag 15. August 2018 gebildet, die bereits ein Standardschreiben von der für sie zuständigen Kommunalverwaltung mit allgemeinen Informationen zu den Pflichten beim Leistungsbezug erhalten hatten und daher seit mindestens sechs Monaten Bürgergeld bezogen. Insgesamt enthielten die Gruppen knapp 50.000 Personen.

Der empirische Ansatz zur Berechnung der Auswirkungen der Intervention auf verschiedene abhängige Variablen besteht – wie bei Feldexperimenten üblich – in einer einfachen Kleinst-Quadrat-Regression mit einem Indikator für die Treatmentgruppe und soziodemografischen Kontrollvariablen, die zum Zeitpunkt vor der Intervention gemessen werden. Zusätzlich enthält die Schätzgleichung Indikatoren für den Wohnort der Leistungsbeziehenden in allen 98 dänischen Kommunen. Die Schätzung misst den Treatmenteffekt der Verfügbarkeit des Online-Tools – unabhängig davon, ob die Leistungsbeziehenden es tatsächlich nutzen (sogenannter Intention-to-Treat-Effekt). Um auch den Effekt der Intervention nur für diejenigen zu ermitteln, die tatsächlich

vom Online-Tool Gebrauch machten, nehmen Cairo & Mahlstedt (2021) zusätzliche Berechnungen vor.

Die Jobnet-Daten wurden mit weiteren administrativen Datenquellen verknüpft, um die Auswirkung der Intervention auf mehrere abhängige Variablen untersuchen zu können und dabei für wichtige erklärende Variablen zu kontrollieren. Neben den Daten von Jobnet bzw. der dänischen Arbeitsverwaltung mit Informationen zu Leistungsauszahlungen (aus der Datenbank DREAM, Styrelsen for Arbejdsmarked og Rekruttering, o. D.), Sanktionen bzw. Ausnahmen davon, gesammelten Arbeitsstunden und Klick-Zahlen im Jobnet-Portal stehen den Autor\*innen nach der Datenverknüpfung auch Informationen zu Beschäftigungsverhältnissen und Einkommen (aus den E-Income-Daten, Danish Customs and Tax Administration, o. D.) zur Verfügung.

Die Ergebnisse des Papiers zeigen, dass die Intervention – also die Möglichkeit der Nutzung des Online-Tools – die Wahrscheinlichkeit einer Leistungskürzung im Vergleich zur Kontrollgruppe, die keinerlei Benachrichtigungen erhielt, um fünf Prozent reduziert. Die Intervention führte außerdem zu positiven Beschäftigungs- bzw. Lohneffekten für die Treatmentgruppe im Vergleich zur ersten Kontrollgruppe ohne individuelle Informationen über die Erfüllung der Pflichten. Diese Effekte wurden bis zu einem Jahr nach Start des Zugangs zum Online-Tool gemessen. Innerhalb eines Jahres hatte die Treatmentgruppe sechs Prozent höhere Löhne, sieben Prozent mehr akkumulierte Arbeitsstunden und eine um drei Prozent erhöhte Wahrscheinlichkeit, eine neue Beschäftigung gefunden zu haben. Der Anstieg der Beschäftigung war insbesondere dadurch getrieben, dass mehr Vollzeitstellen angetreten wurden. Positive Beschäftigungs- bzw. Einkommenseffekte lagen allerdings nur für die Personen in der Treatmentgruppe vor, die ohnehin ein relativ niedriges Sanktionsrisiko hatten. Arbeitsuchende, die wenig Arbeitsstunden nachweisen konnten und somit schon vor der Intervention ein hohes Sanktionsrisiko hatten, verließen das Unterstützungssystem durch die Intervention um sieben Prozent häufiger. Gleichzeitig stiegen die Bezüge von Erwerbsminderungsleistungen sowie Leistungen zur Unterstützung von Ausbildung bzw. Studium.

Cairo & Mahlstedt (2021) weisen darauf hin, dass die Verwendung eines Online-Tools, wie es in der evaluierten Intervention zur Verfügung gestellt wurde, die persönlichen Beratungen in Jobcentern nicht ersetzen soll bzw. kann. In Heterogenitätsanalysen arbeiten sie heraus, dass besonders positive Effekte erzielt wurden, wenn parallel zum Online-Tool eine intensive Beratung durch Fallmanager\*innen stattfand.

### 5.9.3 Weitere Anwendungen

Im Folgenden werden wissenschaftliche Studien dargestellt, die weitere Einsatzmöglichkeiten für Feldexperimente im Arbeitsmarktkontext aufzeigen. Ein US-amerikanisches Beispiel ist die Studie von Flory et al. (2015), die ein groß angelegtes Feldexperiment in 16 Großstädten durchführten um zu analysieren, inwiefern eine leistungsabhängige Bezahlung das Bewerbungsverhalten von Frauen vs. Männern auf dem Arbeitsmarkt beeinflusst. Hierzu wurden in zwei Teilstudien insgesamt ca. 9.000 Personen verschiedene Arten von fiktiven Stellenanzeigen über eine Online-Stellenbörse zugespielt. Die Anzeigen unterschieden sich u.a. in Bezug auf die Art der Bezahlung, die in unterschiedlichen Abstufungen von der Leistung abhing und darin, ob typisch männliche Konnotationen enthalten waren. Die Ergebnisse zeigten, dass sich Männer – wie zuvor schon in anderen Laborexperimenten belegt – eher auf Ausschreibungen mit leistungsbezogener Vergütung bewarben. Zusätzlich ergab die Studie, dass die Bewerbungslücke durch eine verstärkte Hervorhebung von Teamaufgaben und dem Verzicht auf die Verwendung männlicher Konnotationen reduziert werden konnte. In Städten mit einem hohen Lohnniveau war die Lücke allerdings größer – Frauen nahmen eine leistungsabhängige Bezahlung also weniger gerne hin, wenn es gut bezahlte Alternativangebote gab. Die Größe der

Stichprobe ermöglichte außerdem Heterogenitätsanalysen u.a. nach Alter. Diese Analysen zeigten, dass die Bewerbungslücken bei jüngeren Personen größer waren als bei älteren.

Ein groß angelegtes Feldexperiment auf dem deutschen Arbeitsmarkt wurde von Altmann et al. (2018) durchgeführt. Ziel der Studie war es zu untersuchen, wie die Bereitstellung von Informationen die Beschäftigungsaussichten und die Arbeitsmarktergebnisse von Arbeitssuchenden beeinflusst. Personen, die der Treatmentgruppe des Experiments zugewiesen wurden, erhielten eine Broschüre, die sie über Strategien der Arbeitssuche und die Folgen von Arbeitslosigkeit informierte und sie motivierte, aktiv nach einer neuen Beschäftigung zu suchen. Die kausalen Effekte der Vergabe der Broschüre wurden untersucht, indem die Arbeitsmarktergebnisse von Arbeitssuchenden in der Treatment- bzw. Kontrollgruppe anhand von Daten der BA aus Verwaltungsprozessen (Integrierte Erwerbsbiografien des IAB, abgekürzt IEB) verglichen wurden. Die Ergebnisse zeigen tendenziell positive Effekte, die sich aber auf Arbeitssuchende konzentrieren, bei denen das Risiko besteht, für längere Zeit arbeitslos zu sein. Für die Gesamtstichprobe sind die Effekte im Durchschnitt positiv, aber meist nicht statistisch signifikant. Gleichzeitig können ausgeprägte und statistisch signifikante Effekte für Personen, die ein erhöhtes Risiko für Langzeitarbeitslosigkeit aufweisen, nachgewiesen werden. Für diese Gruppe erhöht die Broschüre die Beschäftigung und das Einkommen im Jahr nach der Intervention um etwa vier Prozent. Die Autoren argumentieren, dass die gezielte Bereitstellung von Informationen in Anbetracht der geringen Kosten ein äußerst effizientes arbeitsmarktpolitisches Instrument sein kann.

Ein Feldexperiment, das die administrativen Abläufe bei der Betreuung von Arbeitssuchenden verändert und dadurch einen kausalen Effekt identifiziert, wurde von van den Berg et al. (2021) durchgeführt. Hierbei wurden die deutschen Eingliederungsvereinbarungen im Bereich des SGB III untersucht.<sup>48</sup> Dabei handelt es sich um verpflichtende Vereinbarungen zwischen der Agentur für Arbeit und den Arbeitssuchenden, die von diesen und dem/r Sachbearbeiter\*in gemeinsam unterzeichnet werden. Die Eingliederungsvereinbarungen legen Rechte und Pflichten fest, werden aber im Allgemeinen als Instrumente zur Kontrolle des Suchverhaltens wahrgenommen. Bei der randomisierten kontrollierten Studie mit Tausenden von neu arbeitslos gewordenen Arbeitnehmer\*innen wurde der Zeitpunkt der Eingliederungsvereinbarung sowie der zeitliche Abstand, mit dem dieser Zeitpunkt vor dem Treffen angekündigt wird, für jede einzelne Person randomisiert. Die Arbeitsmarktergebnisse wurden, wie bei Altmann et al. (2018), mithilfe der IEB analysiert. Die Ergebnisse zeigen, dass Eingliederungsvereinbarungen zu Beginn der Arbeitslosigkeit im Durchschnitt einen kleinen positiven Effekt auf die Aufnahme einer Beschäftigung innerhalb eines Jahres haben. Dieses Ergebnis ist auf Personen mit schlechten Beschäftigungsaussichten zurückzuführen. Bei diesen Personen erhöht sich die Wahrscheinlichkeit einer Wiederbeschäftigung innerhalb eines Jahres von 45 Prozent auf 53 Prozent, wenn sie einer frühen Eingliederungsvereinbarung zugewiesen werden.

#### 5.9.4 Übertragbarkeit auf Deutschland

Wie im vorherigen Abschnitt anhand ausgewählter Beispiele dargestellt, werden auch in Deutschland im Themenfeld der Arbeitsmarkt- und Sozialpolitik randomisierte Kontrollstudien durchgeführt. Hierbei bestehen durch die Rechtsordnung Grenzen (Body et al. 2023, S. 4). Als übergeordnete Norm impliziert das Verfassungsgebot der Gleichbehandlung nach Art. 3 Abs. 1 GG, dass Bürger\*innen nicht zufällig von ihnen zustehenden Leistungen ausgeschlossen werden können. Das macht Feldexperimente im Bereich gesetzlich definierter Ansprüche, beispielsweise im Bereich der Sozialversicherung, unmöglich. Randomisierte Kontrollstudien sind nur in Bereichen zulässig, in

---

<sup>48</sup> Ein ähnliches Experiment für den Bereich des SGB II ist in der Studie von Bernhard et al. (2022) enthalten.

denen die Rechtsordnung der Verwaltung entsprechende Entscheidungsspielräume für den Einzelfall gibt, wie es in den beiden im vorherigen Abschnitt beschriebenen Beispielen der Fall ist.

Zum anderen beschränkt die staatliche Aufgabenverteilung zwischen den föderalen Ebenen die Möglichkeit von Feldexperimenten. Kleine Interventionseffekte können nur in großen Datensätzen mit einer hohen Beobachtungszahl mit ausreichender Präzision geschätzt werden. Dies erfordert, dass eine große Zahl an Personen am jeweiligen Feldexperiment teilnimmt. In der Arbeitsmarktpolitik in Deutschland wird dies jedoch aufgrund der eher dezentralen Umsetzung der Arbeitsmarktpolitik in den Agenturen für Arbeit und Jobcentern und den großen lokalen Ermessensspielräumen (Boockmann et al., 2013) erschwert. Doch zeigt die Studie von van den Berg et al. (2021), dass die nötigen Fallzahlen durchaus erreicht werden können.

Wie die Studie von Bernhard et al. (2022) zeigt, sind randomisierte Feldexperimente in Deutschland auch im Bereich der Grundsicherung möglich, zum Beispiel bei den Mitwirkungspflichten – ähnlich wie in der hier vorgestellten Studie, wenn auch nicht unter Verwendung der technischen Möglichkeiten eines digitalen Portals.

### 5.9.5 Fazit

Die vorgestellte Studie aus Dänemark zeigt den Nutzen von randomisierten Feldexperimenten, die es ermöglichen, die Effekte von Politikmaßnahmen kausal zu evaluieren und somit die Wirksamkeit der Arbeitsmarkt- und Sozialpolitik zu verbessern. Im spezifischen Beispiel wurden Arbeitsuchenden über ein Online-Portal individuell zugeschnittene und regelmäßig aktualisierte Informationen über das persönliche Risiko von Leistungssanktionen bei Nichterfüllung von Pflichten angezeigt. Die Bereitstellung dieser Informationen erhöhte insbesondere bei einer Teilgruppe mit niedrigen Risiken für Sanktionen die Beschäftigungswahrscheinlichkeit. Die grundsätzliche Idee des dänischen Feldexperiments, Arbeitsuchenden aktualisierte und personalisierte Informationen zur Verfügung zu stellen, erscheint auf den deutschen Kontext übertragbar. Hierbei ist eine Abwägung zwischen Nutzen und Kosten zu treffen, insbesondere unter dem Gesichtspunkt der Verwaltungspraktikabilität.

Existierende Studien aus Deutschland zeigen, dass Feldexperiment mit einer ähnlichen Zielsetzung grundsätzlich auch in der deutschen Arbeitsmarktpolitik durchgeführt werden können. Gleichzeitig erschweren gesetzliche Vorgaben, die stärker dezentral umgesetzte Arbeitsmarktpolitik sowie fehlende digitale Verfahren die Umsetzung groß angelegter Feldexperimente.

### 5.9.6 Quellen zum Steckbrief

Altmann, S., Falk, A., Jäger, S. & Zimmermann, F. (2018). Learning about job search: A field experiment with job seekers in Germany. *Journal of Public Economics*, 164, 33-49.  
<https://doi.org/10.1016/j.jpubeco.2018.05.003>

Bernhard, S., Stephan, G., Uhlendorff, A., & van den Berg, G. (2022). Verträge zwischen Arbeitslosen und ihrem Jobcenter: Die Wirkung von Eingliederungsvereinbarungen im Rechtskreis SGB II. *IAB-Forschungsbericht* 16/2022. Abgerufen am 15. September 2023 von  
<https://doku.iab.de/forschungsbericht/2022/fb1622.pdf>.

Bodry, Y., Boockmann, B., Kugler, P., von Zitzewitz, V. (2023). Experimentelle Public-Policy Evaluation von Sozialpolitik. DIFIS-Studie 2023/7. *Deutsches Institut für Interdisziplinäre Sozialpolitikforschung*.

Boockmann, B., Osiander, C., Stops, M., Verbeek, H. (2013). Effekte von Vermittlerhandeln und Vermittlerstrategien im SGB II und SGB III (Pilotstudie). *IAB-Forschungsbericht* 07 (2013). Abgerufen am 11. April 2023 von <http://doku.iab.de/forschungsbericht/2013/fb0713.pdf>

Cairo, S. & Mahlstedt, R. (2021). *Transparency of the Welfare System and Labor Market Outcomes of Unemployed Workers* (IZA DP No. 14940). IZA Discussion Paper Series. <https://docs.iza.org/dp14940.pdf>

Danish Customs and Tax Administration. (o. D.). *Reporting pay in E-income*. SKAT, Ministry of Taxation. Abgerufen am 14. April 2023 von <https://skat.dk/data.aspx?oid=2279580#:~:text=E%2Dincome%20is%20a%20register,public%20benefits%2C%20grants%2C%20etc.>

Flory, J.A., Leibbrandt, A. & List J.A. (2015). Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions, *The Review of Economic Studies*, 82(1), 122–155. <https://doi.org/10.1093/restud/rdu030>

Jobnet. (o. D.). *Jobnet-Portal*. Jobnet. Abgerufen am 11. April 2023 von <https://job.jobnet.dk/CV/frontpage>

Styrelsen for Arbejdsmarked og Rekruttering (o. D.). DREAM vejledning version 47 [DREAM-Leitfaden Version 47]. Styrelsen for Arbejdsmarked og Rekruttering. Abgerufen am 11. April 2023 von <https://www.dst.dk/Site/Dst/SingleFiles/GetArchiveFile.aspx?fi=36881106234&fo=0&ext=forskning>

Van den Berg, G. J., Hofmann, B., Stephan, G. & Uhlendorff, A. (2021). *Mandatory Integration Agreements for Unemployed Job Seekers: A Randomized Controlled Field Experiment in Germany* (IZA DP No. 14026). IZA Discussion Paper Series. <https://docs.iza.org/dp14026.pdf>

## 6. Steckbriefe zu innovativen Datenbereitstellungen

### 6.1 Social Policy Simulation Database and Model (Kanada)

#### 6.1.1 Einführung

Die Social Policy Simulation Database and Model (SPSD/M) ist ein vom kanadischen Statistikamt (StatCan) entwickeltes Mikrosimulationsmodell für das Steuer- und Transfersystem in Kanada. Basis der Simulationen bilden vier Datensätze, welche von StatCan exklusiv für dieses Tool verknüpft werden.<sup>49</sup> Die Daten bilden die Bevölkerung, Haushalte und ökonomischen Lebensbedingungen ab und werden mit einer Vielzahl von einzelnen Parametern des Steuer- und Transfersystems Kanadas verbunden. Das Simulationsmodell erlaubt die Variation verschiedener Parameter der Steuer- und Transfersysteme, um hypothetische Situationen wie beispielsweise eine höhere Einkommenssteuer oder ein höheres Kindergeld abzubilden und deren Auswirkungen auf die finanzielle Situation der Bevölkerung, Haushalte und Personen zu ermitteln.

Ziel des SPSPD/M ist die Simulation sozialpolitisch relevanten Kenngrößen, wie beispielsweise die Verteilung von Haushaltseinkommen, unter hypothetischen Politikmaßnahmen. Es liefert damit eine statistische Grundlage für den politischen Diskussions- und Entscheidungsprozess auf verschiedenen Ebenen. Angesichts des komplexen Systems staatlicher Transfers und unterschiedlicher Steuer- und Abgabensätze in den kanadischen Provinzen stellt die Analyse von Änderungen bei Steuern und Transfers – ähnlich wie in Deutschland – eine Herausforderung dar.

Kern der Innovation ist neben der Verknüpfung verschiedener mikroökonomischer Datenquellen ein barrierearmer Zugriff auf das Simulationsmodell für Akteur\*innen aus Politik, Forschung und der Gesellschaft. Der Zugriff ist kostenlos und über den eigenen Computer möglich (auch im Ausland). Hierfür wird von StatCan ein Programm zur Verfügung gestellt, über dessen Bedienoberfläche die Simulationsparameter bereitgestellt werden. Um den Einstieg und den Umgang mit den Funktionen des Modells zu erleichtern, bietet StatCan Schulungen an. Dieser öffentliche Zugriff ist im Vergleich zu anderen Mikrosimulationsmodellen, u. a. in Deutschland, ein Alleinstellungsmerkmal.

#### 6.1.2 Technische Beschreibung

##### 6.1.2.1 Datenquellen

Die Daten für das Modell setzen sich aus vier Quellen zusammen. Zwei basieren auf Befragungen der kanadischen Bevölkerung bzw. Haushalte (Survey of Labour and Income Dynamics [SLID] und Survey of Household Spending [SHS]); zwei Datensätze basieren auf Daten aus administrativen Prozessen (Personal Income Tax Returns [T1 Family File] und EI Claimant History Data). Tabelle 6-1 stellt die Datensätze schematisch dar.

---

<sup>49</sup> *SPSD* bezieht sich auf die Datenbasis (Social Policy Simulation Data), während sich *SPSM* auf das Simulationsmodell bezieht (Social Policy Simulation Model). Der Titel *SPSD/M* (Social Policy Simulation Data and Model) beschreibt die Innovation als Ganzes.

**Tabelle 6-1: Übersicht der Datengrundlagen des SPSPD/M**

<u>Survey - Daten</u>		
	<b>The Survey of Labour and Income Dynamics (SLID)</b>	<b>Survey of Household Spending (SHS)</b>
Quelle	StatCan	StatCan
Teilnahme	freiwillig	freiwillig
Einheiten	Haushalte	Haushalte
Fallzahl	~17.000	~15.000
Inhalte der Daten (Auszug)	<ul style="list-style-type: none"> <li>○ Merkmale des Haushalts</li> <li>○ Regionale Angaben</li> <li>○ Aktivität auf dem Arbeitsmarkt</li> <li>○ Unterstützungszahlungen</li> <li>○ Ausgaben für Kinderbetreuung</li> <li>○ Einkommen</li> </ul>	<ul style="list-style-type: none"> <li>○ Merkmale des Haushalts</li> <li>○ Regionale Details</li> <li>○ Wohnsituation/Eigentum</li> <li>○ regelmäßige Ausgaben (tagesgenau)</li> <li>○ Ausgaben für Gesundheit und Pflege</li> <li>○ Ausbildung und Freizeitgestaltung</li> </ul>
<u>Administrative Daten</u>		
	<b>Personal Income Tax Returns (T1 Family File, auch bekannt als <i>Green Book</i>)</b>	<b>EI Claimant History Data</b>
Quelle	StatCan	Arbeitsministerium <sup>50</sup>
Einheiten	Personen	Personen
Fallzahl	>400.000 Einkommenssteuererklärungen	zehn Prozent der Grundgesamtheit, ca. 200.000 Personen
Inhalte der Daten (Auszug)	<ul style="list-style-type: none"> <li>○ Einkommen aus Transfers</li> <li>○ Erwerbseinkommen</li> <li>○ Einkommen aus Rentenversicherung/ Pension</li> <li>○ Familienstand</li> </ul>	biografische und individuelle Ansprüche aus der Arbeitslosenversicherung

Quellen: Expert\*innengespräch, Décarie et al. (2012), Statistics Canada (o. D.-a), Statistics Canada (o. D.-b), Statistics Canada (2020). Eigene Darstellung. Erläuterung: Die Stichprobengröße bezieht sich jeweils auf die für die Verknüpfung relevante Größe des Datensatzes. Im Ursprung sind die einzelnen Datensätze teilweise größer.

<sup>50</sup> Das Ministerium besaß je nach Kabinett wechselnde Bezeichnungen (u. a. Human Resources Development Canada, Employment and Social Development Canada, Employment Workforce Development and Labour).

Der Survey of Labour and Income Dynamics (SLID) ist eine von StatCan selbst durchgeführte Haushaltsbefragung. Kernelement ist das Einkommen unter Berücksichtigung verschiedener Quellen, wie beispielsweise das Erwerbseinkommen sowie staatliche Transfers im Rahmen des Kindergeldes. Des Weiteren sind im SLID Haushaltsmerkmale vorhanden (Statistics Canada, o. D.-b). Der Survey of Household Spending (SHS) bildet die zentralen Ausgabenposten eines Haushalts ab. Dies ist für die Effekte auf die Haushaltsausgaben bzw. -budgets unter hypothetischen Steuer- bzw. Transferleistungen relevant. Ferner dient der SHS als Referenz für die Gewichtung des finalen Datensatzes. Sowohl der SLID als auch der SHS sind freiwillige Befragungen.

Administrative Informationen kommen aus den Personal Income Tax Returns (auch Green Book genannt). Dessen Merkmale dienen hauptsächlich zur Korrektur verzerrter Einkommensinformationen im SLID, vor allem von Personen mit extrem hohem Einkommen. Informationen über Transfers aus der Arbeitslosenversicherung kommen aus individuellen Ansprüchen der Arbeitslosenversicherung (Employment Insurance Claimant History Data) und dienen ebenfalls der Erfassung des verfügbaren Einkommens von Personen und Haushalten.

Die einzelnen Datensätze werden mithilfe von Matching- und Imputationsverfahren unter geltenden Datenschutzregeln verknüpft (siehe Abschnitt 4.2.2.2).<sup>51</sup> Die Beobachtungen werden nach der Zusammenführung der vier Datenquellen um Hochrechnungsfaktoren ergänzt, um ein repräsentatives Bild der kanadischen Bevölkerung zu erhalten.<sup>52</sup> Die Datenbasis wird unter der Bezeichnung Social Policy Simulation Database (SPSD) geführt.

Neben der SPSP wird für die Simulationsberechnungen das Commodity Tax Model (COMTAX) benötigt. COMTAX bildet die Steuersysteme auf Bundes- und Provinzebene ab und differenziert Konsumgüter und Dienstleistungen mit unterschiedlichen Umsatzsteuersätzen. Dies ist notwendig, um die Effekte einer Änderung dieser Steuersätze präzise zu berechnen, da das Steuersystem in Kanada komplex ist (Statistics Canada, o. D.-a).

### 6.1.2.2 Verknüpfung

Die Verknüpfung der beschriebenen Datensätze hat laut StatCan den Anspruch, eine nachvollziehbare und plausible Datenstruktur zu erzeugen, die alle relevanten Dimensionen und Merkmale der Original-Datensätze nachbildet, während gleichzeitig die Datenschutzvorgaben eingehalten werden (Statistics Canada, o. D.-a). Es werden verschiedene Verfahren angewendet, um die oben beschriebenen Datenquellen aufzubereiten und zur SPSP zu verknüpfen (Statistics Canada, o. D.-b).<sup>53</sup> Der durch die Verknüpfung entstandene Datensatz enthält ca. 1 Mio. „synthetische“ Bürger\*innen und ca. 300.000 „synthetische“ Haushalte. Dies bedeutet, dass diese Einheiten plausible Bürger\*innen und Haushalte darstellen, jedoch nicht tatsächlichen Beobachtungen in den Daten entsprechen. Der verknüpfte Datensatz bildet die Grundlage für das Simulations-Modell SPSM.

---

<sup>51</sup> Da im finalen Datensatz keine Verknüpfung von individuellen Informationen aus den vier zugrundeliegenden Datensätzen vorgenommen wird, sind die datenschutzrechtlichen Anforderungen für die Bereitstellung und Verwendung im Rahmen des Simulationsmodells gewährleistet.

<sup>52</sup> Einschränkungen in den Daten (Survey-Non-Response, Verzerrungen an den Einkommensrändern) werden berücksichtigt. Die Region Yukon, die Nunavut- und die Nordwest-Territorien sind nicht in den Daten berücksichtigt. Ebenfalls sind Personen, die in Reservaten wohnen, oder Angehörige der Streitkräfte, die in Kasernen wohnen, nicht in den Daten enthalten (Statistics Canada, o.D.-a).

<sup>53</sup> Diese sind in diesem Kontext gängig und werden deshalb nicht näher beschrieben. Für Details wird auf Statistics Canada (o.D.-b) verwiesen.

### 6.1.2.3 Simulation

Insgesamt stehen interessierten Personen und Einrichtungen ca. 2.000 Parameter zur Verfügung, die sie für eine Simulation verändern können (Statistics Canada, o. D.-a). Dazu gehören einzelne Steuern, zum Beispiel Einkommenssteuersätze oder Steuern auf Konsumgüter wie Spirituosen, aber auch Kindergeld und Kindersteuervergünstigungen.

Die Simulationen basieren auf eigens von StatCan entwickelten Algorithmen, die je nach Fragestellung (und Veränderung der dafür relevanten Parameter) auf Daten der individuellen oder der Haushaltsebene angewendet werden (Statistics Canada, o. D.-a).<sup>54</sup> Anschließend werden die interessierenden Zielgrößen auf die zu betrachtende Gesamtpopulation hochgerechnet und erlauben dadurch eine makroökonomische Betrachtung der Effekte. Die Outputs des Modells umfassen beispielsweise die hypothetischen Mittelwerte von Haushaltseinkommen oder das aggregierte Steueraufkommen unter Berücksichtigung der geänderten Parameter. Mikroökonomische Verhaltensanpassungen, beispielsweise eine Reaktion des Arbeitsangebots bei Änderung der Einkommenssteuer, werden im SPSM nicht berücksichtigt.

Seit der Entwicklung und Implementation des SPSD/M in den 1980er Jahren wurden im Laufe der Jahre die Inhalte des Modells (z. B. die abzubildenden Leistungen) auf dem aktuellen Stand gehalten, die technische Art der Programmierung wurde jedoch beibehalten. Aktuell benötigt die Programmierung des Simulationsprogramms daher eine Aktualisierung, da der Code und dessen Architektur weiterhin auf den Bedingungen basiert, die sich aus der Rechenleistung eines Computers aus dem Jahr 2000 (letzte Softwareaktualisierung) ergibt. Die damit verbundenen Kosten der Entwicklung und Implementierungen der Software sind nun einmalig sehr hoch. Eine stetige Verbesserung der Programmierung der Simulationsalgorithmen im Laufe der Jahre wäre nach Auskunft einer Expertin von StatCan daher wünschenswert gewesen.

Des Weiteren ist der Merkmalsumfang des SHS in den vergangenen Jahren gesunken, sodass es zunehmend schwierig wird, die Ausgabenseite der Haushalte detailliert abzubilden. Hier steht StatCan nun vor der Aufgabe, die Daten bzw. Datenqualität wiederherzustellen. Eine Aufrechterhaltung der Befragungsdatenqualität durch eine stetige Kommunikation zwischen der Organisation des SHS und den Anforderungen für die Simulation hätte dieser Herausforderung entgegengewirkt.

Von Seiten der Nutzenden bekommt StatCan zunehmend Anfragen bezüglich eines dynamischen Modells, das Verhaltensänderungen bei beispielsweise steigenden Preisen oder Löhnen abbilden kann. Grundsätzlich hat StatCan darauf verzichtet, diese Elastizitäten abzubilden. Denn für die Abbildung solcher Dynamiken im Programmcode müssen Annahmen getroffen werden und Bedingungen erfüllt sein, die je nach Anwendungsfall abzuwägen sind. Allerdings mussten infolge der Covid-19-Pandemie staatliche Förderprogramme implementiert werden, deren Inanspruchnahme durch dramatisch geänderte Zugangsbedingungen statisch nicht abzubilden war. Daher wurden im Zuge der Covid-19-Pandemie vereinzelt dynamische Elemente in die Programmierung aufgenommen. Ferner gibt es in Kanada eine sogenannte Fuel Charge, die kohlendioxidintensive Güter preislich unattraktiver machen soll und über Konsumänderungen wirkt. Bisher hat dies im Simulationsmodell keine Effekte, da Haushalte und Personen die betroffenen Güter trotzdem in gleichem Umfang konsumieren, jedoch zu höheren Preisen. Politische Entscheidungsträger\*innen sind jedoch an einer Abschätzung von Verhaltensanpassungen infolge der Fuel Charge sehr

---

<sup>54</sup> Das Mikrosimulationsmodell des SPSD/M basiert auf Programmcodes, die ausschließlich von StatCan entwickelt wurden.

interessiert. Diese Präferenzen der Nutzenden und die potenziellen Limitationen eines statischen Modells hätte StatCan gerne früher antizipiert.

### 6.1.3 Anwendung

Grundsätzlich wird das SPSPD/M von verschiedenen Ministerien, regionalen Verwaltungen, wissenschaftlichen Einrichtungen und privatwirtschaftlichen Unternehmen genutzt (Statistics Canada, o. D.-a).<sup>55</sup> Nach Auskunft von StatCan wird es beispielsweise mindestens monatlich vom Parliamentary Budget Office (PBO, o. D.) verwendet.

Interessierte haben auf die zugrundeliegenden Daten und Algorithmen keinen Zugriff, sondern können Simulationen über eine Bedienoberfläche erstellen, die die Berechnungen im Hintergrund durchführt. Die Bedienoberfläche ist so gestaltet, dass eine Nutzung auch durch Personen möglich ist, die keine Expertise im Bereich der Datenanalyse haben.

Nutzende können je nach Anspruch und Expertise zwischen zwei Modi der Anwendung des SPSPD/M wählen:

- **Black-Box Mode:** In diesem Modus ist die Bedienung des SPSPD/M besonders benutzungsfreundlich und übersichtlich ausgerichtet. In diesem Modus können die Parameter der Simulationsalgorithmen ohne große technische Expertise verändert werden. Dieser Modus wird von Nutzenden am häufigsten verwendet.
- **Glass-Box Mode:** Hier können Programmierer\*innen mit Kenntnissen in C++ die bestehenden Algorithmen verändern und neue entwickeln. Dadurch lassen sich weitaus komplexere und spezifischere Fragestellungen beantworten als im *Black-Box Mode*.

StatCan formuliert beispielhaft folgende Fragestellungen, die mit SPSPD/M beantwortet werden können (Statistics Canada, o. D.-a):

- Welche Erhöhung des Kindergeldes kann durch eine Erweiterung des Kindergeld-Bundesprogramms in Höhe von 500 Mio. kanadischen Dollar (CAD) erreicht werden?
- Welche Provinz hat für eine Person mit zwei Kindern, einem Arbeitseinkommen von 45.000 CAD und weiteren 15.000 CAD Kapitaleinkommen die vorteilhafteste Steuerstruktur?
- Welches Finanzierungsvolumen wäre erforderlich, um alle Familien und Personen mit niedrigem Einkommen auf die von StatCan für 2014 festgelegten Grenzwerte für niedrige Einkommen anzuheben?

Nach der Durchführung einer gewünschten Simulation lassen sich Output-Tabellen erstellen, die anschließend für die Ergebnispräsentation oder weitere Analysen verwendet werden können (Statistics Canada, o. D.-a).

Eine beispielhafte Simulation und die vorangehende Interaktion mit dem Programm werden im Folgenden dargestellt:<sup>56</sup> Zunächst kann zwischen dem Black-Box- und dem Glass-Box-Modus gewählt werden. Zudem kann das zu betrachtende Jahr und ein Referenzjahr bezüglich des Preisniveaus ausgewählt werden. Anschließend öffnet sich das Drop-Down-Menü, in welchem auf

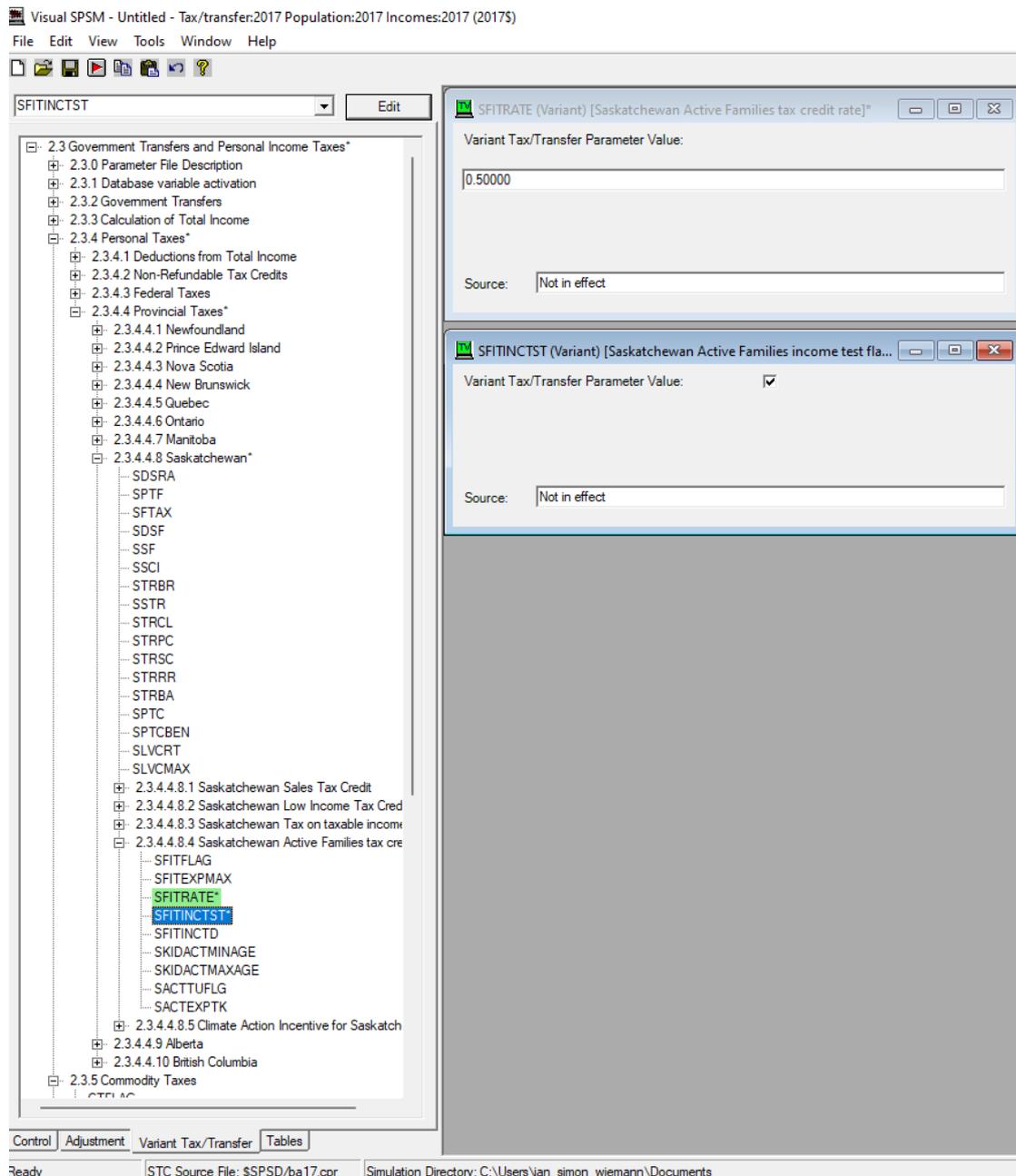
---

<sup>55</sup> Das kanadische Finanzministerium verwendet allerdings ein eigenes Mikrosimulations-Tool aufgrund der Verwendung sensibler Daten. Dies läuft unter dem Titel „Canada’s Federal Corporate Income Tax: A Microsimulation Model“ und wurde 2000 initiiert (Décarie et al., 2012)

<sup>56</sup> Hier wird das Visual SPSPD/M, v29 verwendet. Wie der Name andeutet, besteht die Interaktion mit dem Programm aus einem visuell dargestellten Drop-Down Menu und sich öffnenden Fenstern bei der Variation von Parametern. Im Classic SPSPD/M v29 besteht die Interaktion aus der Eingabe von Befehlen wie bei der Programmierung.

der linken Seite alle Parameter in einem Baum-Schema dargestellt werden und auf der rechten Seite die Variation eines Parameters nach dem Auswählen durchgeführt werden kann (Abbildung 6-1). Beispielhaft wird in der Abbildung die Active-Families-Tax-Credit-Rate in der Region Saskatchewan auf 0,5 Prozent gesetzt.

**Abbildung 6-1: Auswahl und Variation der Parameter des SPSM**



Quelle: StatCan, SPSP/M version 29.0, Visual SPSM version 28.1.1.0. Copyright: 2016 Minister of Innovation, Science and Economic Development. Anmerkung aufgrund rechtlicher Vorschriften: "This analysis is based on Statistics Canada's Social Policy Simulation Database and Model. The assumptions and calculations underlying the simulation results were prepared by Jan Simon Wiemann and the responsibility for the use and interpretation of these data is entirely that of the author."

Nach dem Speichern dieser angepassten Parameter wird die Simulation über das kleine rote Dreieck im linken oberen Fenster gestartet. Sobald die Simulation fertig ist, öffnet sich ein Fenster mit zwei Tabellen zu Eckdaten der Simulation (Abbildung 6-2). Zum einen werden aggregierte Daten für

Kanada dargestellt, zum anderen werden die Effekte auf die verschiedenen Regionen dargestellt. Die Tabellen enthalten Informationen beispielsweise zur verwendeten Grundgesamtheit, verschiedene Einkommensinformationen und die Zahl der von der Maßnahme profitierenden und benachteiligten Personen.

Abbildung 6-2: Ergebnisdarstellung einer Simulation

The image displays two screenshots of the SPSP/M simulation software interface. Both screenshots show the 'SPSP/M (Database 29.0)' window with the date 'Tue Mar 21 11:50:46 2023' and the driver version '29.0 : 97-26, File: sim\_3\_B.mpr'. The variant description is 'No variant results' and the sample size is '1.0000 AGENAME=Standard'.

The left screenshot, titled 'Table 0', shows 'Results for Census Families'. The table lists various variables (x1,000,000) and their total values:

Variable (x1,000,000)	TOTAL
Family Units (x1000)	16996.8
Persons (x1000)	36234.7
CIS Survey Records	49179.0
SPSD Records	525113.0
Income (Base)	1077681.8
Income (Variant)	1077681.8
Change	0.0
Number of Gainers (x1000)	0.0
Number of Losers (x1000)	0.0
No Change (x1000)	16996.8
Gainer's Gain	0.0
Loser's Loss	0.0
Total Income	1482543.9
Market Income	1272555.7
Wages and Salaries	912432.3
Self-Employment Income	55148.8
Investment Income	150009.8
Other Income	146963.8
Transfer Income	209988.3
Total Tax	404862.5
Net Transfers	-194874.2
Disposable Income	1190636.7
Consumable Income	1077681.8
Federal Taxes	236300.6
Federal Income Tax	148999.7

The right screenshot, titled 'Table 1', shows 'Results for Census Families by Province'. The table lists the same variables as Table 0, but with values for each province: NFLD, PEI, NS, NB, QUE, ONT, MAN, and SASK.

Variable (x1,000,000)	NFLD	PEI	NS	NB	QUE	ONT	MAN	SASK
Family Units (x1000)	249.1	70.1	463.1	365.3	4129.5	6409.4	568.0	507.9
Persons (x1000)	524.1	150.8	929.7	745.2	8426.1	14122.6	1262.9	1093.4
CIS Survey Records	2066.0	1439.0	2894.0	2777.0	8896.0	11858.0	4721.0	3887.0
SPSD Records	8837.0	5395.0	14426.0	8575.0	107627.0	189022.0	17458.0	12077.0
Income (Base)	13574.9	3713.5	23643.2	18717.9	223753.9	427325.0	34106.4	32864.6
Income (Variant)	13574.9	3713.5	23643.2	18717.9	223753.9	427325.0	34106.4	32864.6
Change	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Number of Gainers (x1000)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Number of Losers (x1000)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
No Change (x1000)	249.1	70.1	463.1	365.3	4129.5	6409.4	568.0	507.9
Gainer's Gain	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Loser's Loss	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Income	19915.1	5255.0	33825.5	26386.4	315145.1	587619.3	47648.6	44337.5
Market Income	15902.9	4215.3	27711.5	21236.9	258983.0	508078.7	40895.5	38549.5
Wages and Salaries	12486.3	3072.2	19691.0	15913.5	185737.0	363546.1	29843.0	27116.2
Self-Employment Income	658.6	281.2	1046.7	711.5	11399.9	22056.6	1925.0	1960.6
Investment Income	814.6	370.7	2747.7	1502.9	26384.1	62563.1	4394.8	5575.5
Other Income	1943.4	491.2	4226.2	3108.9	35462.1	59128.0	4736.7	3897.1
Transfer Income	4012.2	1039.8	6114.1	5149.6	56162.1	79540.6	6749.1	5788.0
Total Tax	6340.2	1541.6	10182.4	7668.5	91391.3	160294.4	13542.2	11472.9
Net Transfers	-2327.9	-501.8	-4068.3	-2519.8	-35229.2	-80753.8	-6793.1	-5884.9
Disposable Income	15785.8	4259.2	27112.4	21479.4	251739.0	472924.2	37889.1	36020.1
Consumable Income	13574.9	3713.5	23643.2	18717.9	223753.9	427325.0	34106.4	32864.6
Federal Taxes	3250.2	809.4	5258.0	4075.1	42760.9	96487.6	7615.4	7221.5

Quelle: StatCan, SPSP/M version 29.0, Visual SPSP/M version 28.1.1.0. Copyright: 2016 Minister of Innovation, Science and Economic Development. Anmerkung aufgrund rechtlicher Vorschriften: "This analysis is based on Statistics Canada's Social Policy Simulation Database and Model. The assumptions and calculations underlying the simulation results were prepared by Jan Simon Wiemann and the responsibility for the use and interpretation of these data is entirely that of the author."

### 6.1.3.1 Unterstützende Produkte von StatCan zur Nutzung des SPSP/M

Um das SPSP/M bestmöglich zu nutzen, werden Interessierten drei Arten von Hilfestellungen angeboten:

- **Handbuch:** Diese Dokumentation besteht aus einem Einführungs-Handbuch (*Introduction Manual*) in das SPSP/M und gibt Interessierten einen ersten Einblick in die Funktionsweise und Bedienung des Tools. Ferner wird hier eine Anleitung zur Nutzung der Online Hilfe (siehe unten) angeboten. Das Benutzerhandbuch (*Users' Manual*) erklärt das COMTAX Modell, den Black-Box sowie Glass-Box Mode. Das Referenz-Handbuch (*Reference Manual*) beinhaltet eine Beschreibung der Implementierung der Steuer- und Transfersysteme in das SPSP/M, sowie der Merkmale und Parameter des Modells.
- **Trainingskurse:** Diese werden für Black-Box und Glass-Box-Nutzende getrennt angeboten. Das Training für den Black-Box-Modus beläuft sich derzeit auf ca. 15 Stunden, verteilt auf vier Tage mit bis zu 15 Personen je Kurs. Vorwiegend wird dort das Verständnis der Komponenten des SPSP/M, die Verwendung von Schlüsselfunktionen (Steuerung von Szenarien, Erstellung neuer Variablen und benutzerdefinierter Tabellen) geschult. Ferner werden praktische Erfahrungen durch das Durchführen von Fallstudien gesammelt. Das Training für den Glass-Box Modus ist auf bereits mit dem Black-Box Modus vertraute Anwendende zugeschnitten. Dieses Training dauert ca. 22 Stunden, verteilt auf acht Tage, wobei jeder Kurs aus bis zu acht Personen besteht. Neben komplexeren Fallstudien werden

auch Modifizierungen bestehender SPSM-Algorithmen in C++ geschult (Expertengespräch mit StatCan, 20. März 2023).

Neben diesen Trainingskursen bei StatCan gibt es bei Bedarf die Möglichkeit für Einzeltrainings sowie Besuche vor Ort (der nutzenden Einrichtung) durch Beschäftigte bzw. Expert\*innen von StatCan. Diese Sondertrainings müssen im Einzelfall angefragt werden.

- Online-Hilfe: Hier wird neben einer Beschreibung aller Parameter und Merkmale die Unterstützung bei deren Anpassung im Dialogfenster des Anwendungstools angeboten. Diese Hilfe richtet sich vorwiegend an Nutzende während der Arbeit mit dem SPSD/M.

### 6.1.3.2 Nutzungsvoraussetzungen

Um das SPSD/M nutzen zu können, muss eine Lizenzvereinbarung mit StatCan unterschrieben werden. Anschließend werden die Dokumentationen (siehe oben), die SPSD und die Software für die Simulationen elektronisch übermittelt, sodass das SPSD/M lokal verwendet werden kann. Nutzende benötigen einen PC mit Windows-Betriebssystem, auf dem Windows 7 oder höher installiert ist. Für Nutzende des Glass-Box-Modus ist außerdem Microsofts Visual C++.

### 6.1.3.3 Angegliederte Projekte und Forschung

Henessy et al. (2015) erweitern das SPSD/M um individuelle Gesundheitsdaten. Die Autor\*innen evaluieren damit alternative Finanzierungsstrategien des kanadischen Gesundheitswesens. Dabei werden die Daten des SPSD/M mit zwei weiteren Erhebungen von StatCan und einer externen administrativen Datenquelle des Canadian Institute for Health Information verknüpft.

Interessante Forschungsergebnisse, die auf dem SPSD/M basieren, finden sich beispielsweise bei Barker (2018). Hier werden höhere Rentenansprüche und ein höheres Renteneintrittsalter simuliert und deren Effekte auf Altersarmut, insbesondere bei Frauen, berechnet. Tarrow (2012) vergleicht die Einkommensverteilung von kanadischen Haushalten unter der Annahme eines einmaligen Einkommenstransfers mit der tatsächlich beobachteten Verteilung und nutzt dafür das SPSD/M. Krzepkowski & Mintz (2013) simulieren die Effekte eines dem Ehegattensplitting in Deutschland ähnlichen Steuersystem im Vergleich zu der (damals) geltenden Regel eines individuellen Steuersystems in Kanada mithilfe des SPSD/M.

### 6.1.4 Übertragbarkeit auf Deutschland

In Deutschland werden an verschiedenen Forschungsinstituten eine Reihe von Mikrosimulationsmodellen verwendet. Hier sind unter anderem die folgenden Institutionen zu nennen, die ein solches Modell entwickelt haben: a) das Tax-Benefit Microsimulation Model des DIW (STSM, ursprünglich am ZEW entwickelt, siehe Steiner, Wrohlich, Haan und Geyer, 2012), b) das ifo Microsimulation Model (ifo-MSM, Blömer & Peichl, 2020), c) das RWI-Einkommensteuer-Mikrosimulationsmodells (EMSIM, Bechara et al., 2015), d) das MikroSim an der Universität Trier und der Universität Duisburg-Essen in Zusammenarbeit mit dem Statistischen Bundesamt (MIKROSIM, o. D.).

Eine neuere Entwicklung ist das Open-Source-Projekt GETTSIM. Als besonderes Merkmal ist hier zum einen die Open-Source-Programmierung unter Verwendung der Programmiersprache Python zu nennen. Dadurch kann jede Person mit Python-Kenntnissen GETTSIM lokal auf dem eigenen Rechner verwenden und Simulationen durchführen. Zum anderen beinhaltet GETTSIM im Allgemeinen keinen Datensatz. Nutzende müssen somit einen Datensatz zur Verfügung haben, auf den GETTSIM zugreift. Dabei müssen einige Mindestanforderungen an die im Datensatz verfügbaren

Merkmale erfüllt sein, damit die Simulationen sinnvoll durchgeführt werden können. Die Vorteile von GETTSIM liegen in der präzisen Abbildung des Steuer- und Transfersystems in Deutschland. Zudem lässt sich das Simulationstool aufgrund der Programmierung in Python gut in bestehende Workflows einbetten.<sup>57</sup> In der Vergangenheit wurde eine aus Python herausgelöste Bedienoberfläche programmiert, jedoch war hier der Erhaltungsaufwand zu hoch, sodass dieses Merkmal inzwischen nicht mehr bereitgestellt bzw. weiterentwickelt wird. Daher ist die aktuelle Bedienoberfläche eher an Personen mit Programmierkenntnissen in Python gerichtet.

Insgesamt stehen in Deutschland also eine Reihe von Mikrosimulationsmodellen zur Verfügung mit denen Änderungen bei Steuern, Sozialabgaben und Transfers analysiert werden können. Diese unterscheiden sich zum SPSP/M hauptsächlich in der (fehlenden) Verfügbarkeit für die breite Öffentlichkeit und einer (fehlenden) barrierearmen Bedienoberfläche. Eine solche Bedienoberfläche ließe sich durchaus in eines der existierenden Modelle integrieren. Allerdings würde bei der Programmierung und dem laufenden Betrieb einer Bedienoberfläche erheblicher Aufwand entstehen.

Als Datenbasis käme zunächst das SOEP in Betracht, da dieses viele relevante Merkmale umfasst, die für eine Simulation notwendig sind. Eine Verknüpfung mit administrativen Daten würde darüber hinaus die Genauigkeit der Erfassung des Steuer- und Transfersystems deutlich verbessern. Das SPSP/M wird, wie beschrieben, durch die Verknüpfung von Befragungsdaten und Daten aus administrativen Quellen hinsichtlich der Merkmale und ihrer Erfassung deutlich umfangreicher und präziser. Auch dieses innovative Element ließe sich nach Deutschland übertragen, beispielsweise durch die Nutzung bereits verknüpfter Daten wie z. B. dem SOEP-RV. In diesem Datensatz werden die Befragungsdaten aus dem SOEP auf individueller Ebene mit Informationen aus den Daten der Deutschen Rentenversicherung verknüpft. Dadurch können zusätzliche Merkmale aus dem Bereich der Rentenversicherung wie z. B. die erworbenen Entgeltpunkte oder der Bezug bestimmter Rentenarten genutzt werden, die im SOEP nicht zur Verfügung stehen. Alternativ wäre auch zu prüfen, ob die Biografiedaten ausgewählter Sozialversicherungsträger in Deutschland (BASiD, Hochfellner et al., 2012), gegebenenfalls in Verbindung mit weiteren Daten, für Simulationszwecke genutzt werden könnten. Die Daten beinhalten verknüpfte Daten aus der gesetzlichen Rentenversicherung und Prozessdaten der BA. Durch letztere würden in den verknüpften Daten Informationen zu den beruflichen Tätigkeiten der Personen und den Betriebsstätten erfasst werden, sodass spezifischere Analysen, beispielsweise zu einzelnen Wirtschaftsabschnitten oder Berufen, möglich würden.

Die Nutzung verknüpfter Daten würde in Deutschland einen hohen Mehrwert in der bereits existierenden Mikrosimulationslandschaft schaffen. Sollte das Simulationsmodell für die breite Öffentlichkeit zur Verfügung gestellt werden, entstehen allerdings datenschutzrechtliche Fragen, die noch ungeklärt sind. Zudem müssen die darunterliegenden Annahmen und Parameter stets bei der Interpretation der Ergebnisse berücksichtigt und im Einzelfall diskutiert werden.

### 6.1.5 Fazit

Das SPSP/M ist ein von StatCan bereitgestelltes Mikrosimulationsmodell. Sein Ziel ist die Simulation sozialpolitisch relevanter Ergebnisvariablen, wie beispielsweise die Verteilung der Haushaltseinkommen, unter hypothetischen Politikmaßnahmen. Die Innovation besteht in erster Linie darin, dass Interessierte mithilfe einer benutzungsfreundlichen Oberfläche Simulationen selbst individuell durchführen können. Die kanadische Anwendung ist aber auch deshalb innovativ, weil das Simulationsmodell auf verknüpften Daten aus administrativen Prozessen und Befragungen beruht.

---

<sup>57</sup> Beispielsweise kann und wird bereits über die Statistiksoftware STATA auf GETTSIM zugegriffen.

Durch eine Übertragung des kanadischen Beispiels auf Deutschland würden neue Nutzungsmöglichkeiten für die Politiksimulation eröffnet. Die so gewonnenen Informationen könnten die öffentliche Debatte über sozialpolitische und fiskalische Maßnahmen bereichern und ihre Transparenz erhöhen. Ein Mehrwert über bestehende deutsche Mikrosimulationsmodelle hinaus würde durch die Nutzung von verknüpften Daten, beispielsweise der Verknüpfung des SOEP mit Daten der Deutschen Rentenversicherung, entstehen.

Ein Hindernis können dabei Beschränkungen der Datennutzung aufgrund der geltenden Datenschutzbestimmungen darstellen. Insbesondere für verknüpfte Daten wäre zu prüfen, unter welchen Bedingungen eine der SPSD/M ähnliche Nutzung möglich ist. Für die Entwicklung und den fortlaufenden Betrieb einer Bedienoberfläche ist zudem ein hoher Ressourcenaufwand zu erwarten. Zudem müssten mögliche Risiken einer missbräuchlichen Verwendung bedacht werden (siehe Abschnitt 3.3)

### 6.1.6 Quellen des Steckbriefs

Barker, A. (2018). *Increasing inclusiveness for women, youth and seniors in Canada* (OECD ED Working Papers No. 1519). OECD Economics Department Working Papers. <https://dx.doi.org/10.1787/83cb8b8d-en>

Bechara, P., Kasten, T. & Schaffner, S. (2015). *Dokumentation des RWI-Einkommensteuer-Mikrosimulationsmodells (EMSIM)* (RWI Materialien, No. 86). RWI Materialien. Abgerufen am 20. März 2023 von <https://nbn-resolving.de/urn:nbn:de:101:1-201607146488>

Blömer, M., & Peichl, A. (2020). *The ifo tax and transfer behavioral microsimulation model* (ifo Working Paper No. 335). ifo Working Papers.

Décarie, Y., Boissonneault, M., Légaré, J. (2012). *An Inventory of Canadian Microsimulation Models* (SEDAP Research Paper No. 298). Program for Research on Social and Economic Dimensions of an Aging Population Research Papers. Abgerufen am 20. März 2023 von [https://www.researchgate.net/publication/238597596\\_An\\_Inventory\\_of\\_Canadian\\_Microsimulation\\_Models](https://www.researchgate.net/publication/238597596_An_Inventory_of_Canadian_Microsimulation_Models)

Henessy, D., Sanmartin, C., Eftekhary, S., Plager, L., Jones, J., Onate, K., McEvoy, N., Hicks, C. & Deber, R. (2015). Creating a synthetic database for use in microsimulation models to investigate alternative health care financing strategies in Canada. *International Journal of Microsimulation*, 8 (3), 41-74. <https://doi.org/10.34196/ijm.00121>

Hochfellner, D., Müller, D. & Wurdack, A. (2012): Biographical Data of Social Insurance Agencies in Germany – Improving the Content of Administrative Data. *Schmollers Jahrbuch 132 (2012)*, 443 – 451.

Krzepkowski, M. & Mintz, J. (2013). *No More Second-Class Taxpayers: How Income Splitting Can Bring Fairness to Canada's Single-Income Families* (SPP Research Paper No. 6–15). School of Public Policy Research Papers. <https://ssrn.com/abstract=2262525>

MIKROSIM. (o. D.). *MikroSim – Sektorenübergreifendes kleinräumiges Mikrosimulationsmodell*. Universität Trier. Abgerufen am 12. April 2023 von <https://mikrosim.uni-trier.de/>

PBO. (o. D.). Parliamentary Budget Officer | Home. Office of the Parliamentary Budget Officer. Abgerufen am 20. März 2023 von <https://www.pbo-dpb.ca/en/>

Steiner, V., Wrohlich, K., Haan, P. & Geyer, J. (2012). *Documentation of the Tax-Benefit Microsimulation Model STSM*.

Statistics Canada (o. D.-a). *SPSD/M, Product Overview*. Produktübersicht. Abgerufen am 13. Februar 2023 von <https://www.statcan.gc.ca/eng/microsimulation/spsdm/spsdm/spsdm-overview.pdf>

Statistics Canada (o. D.-b). *Database Creation Guide* [Version 20.0]. Datensatz. **Fehler! Linkreferenz ungültig.** Abgerufen am 13. Februar 2023 von [https://datalib.usask.ca/dli/social/spsdm/spsdm-v20.0/doc\\_en/database%20creation%20guide.pdf](https://datalib.usask.ca/dli/social/spsdm/spsdm-v20.0/doc_en/database%20creation%20guide.pdf)

Statistics Canada (2020). *T1 Family File, Final Estimates, 2018*. Technisches Dokument. Abgerufen am 13. Februar 2023 von <https://www150.statcan.gc.ca/n1/pub/72-212-x/72-212-x2020001-eng.pdf>

Tarroux, B. (2012). Are equalization payments making Canadians better off? A two-dimensional dominance answer. *The Journal of Economic Inequality*, 10, 19–44.

## 6.2 Jobnet und Joblog – Portal für Arbeitsuchende mit Verknüpfungsmöglichkeiten (Dänemark)

### 6.2.1 Einführung

Das staatliche dänische Portal Jobnet bündelt zahlreiche Funktionen der Beratung und Vermittlung von Arbeitsuchenden. Neben einer umfassenden Stellenbörse enthält es Funktionen zur Kommunikation zwischen der Arbeitsverwaltung und den Arbeitsuchenden. In einem „Joblog“ genannten Bereich des Portals können Suchaktivitäten und Bewerbungen erfasst werden, was Beratungsgespräche und die Kontrolle von Erfüllungen von Auflagen für den Bezug von Transferleistungen erleichtert.

Jobnet und Joblog bieten zusammengenommen die Möglichkeit einer hohen Effizienz der Verwaltung bei der Beratung von Arbeitsuchenden und der Gewährung von Leistungen. Dadurch, dass die Daten dieser Prozesse digital in einem einzigen Portal gebündelt sind, können Kosten gespart werden und die relevanten Informationen für Arbeitsuchende und Vermittlungsfachkräfte liegen jederzeit leicht einsehbar vor. Durch die strukturierte Erfassung von Stellenanzeigen und individuellen Bewerbungsaktivitäten werden die Arbeitsuchenden umfassend unterstützt. Zugleich kann die Erfüllung von Auflagen für den Leistungsbezug ebenfalls systematisch überprüft werden. Dänische Arbeitsuchende müssen auf Joblog u. a. ihren Lebenslauf sowie vollständige Bewerbungsunterlagen für mindestens zwei Bewerbungen pro Monat hochladen und außerdem nachweisen, dass sie zwei Bewerbungen pro Woche verfassen. Bei Nichterfüllung erfolgen Vorladungen zu „Verfügbarkeitsinterviews“ und unter Umständen die Kürzungen von Leistungen.

Für die Forschung bietet die Verknüpfung von Stellenanzeigen und Bewerbungen auf die inserierten Stellen eine außergewöhnlich gute Möglichkeit zur Beantwortung vielfältiger Forschungsfragen zum Suchverhalten auf dem Arbeitsmarkt – insbesondere, da die Jobnet-Daten über eine universelle Identifikationsnummer mit vielen weiteren administrativen Daten verknüpft und für die gesamte dänische Bevölkerung ausgewertet werden können.

Im Folgenden wird zunächst die zugrundeliegende digitale Ausrichtung der dänischen Arbeitsverwaltung skizziert und anschließend das Portal Jobnet und insb. die Funktionen von Joblog vorgestellt. Schließlich werden laufende oder kürzlich veröffentlichte Forschungsprojekte, die mit Daten von Jobnet bzw. Joblog erstellt wurden, dargestellt. Steckbrief 5.9 präsentiert ein weiteres innovatives Projekt, das auf der Datenbasis von Jobnet und Joblog beruht und illustriert, wie die Daten in Kombination mit Feldexperiment-Methoden genutzt werden.

### 6.2.2 Technische Beschreibung

#### 6.2.2.1 Digitale Arbeitsverwaltung

Dänemark ist ein Vorreiterland in Europa in Bezug auf Digitalisierung. Im Jahr 2022 belegte es nach Finnland Platz zwei des EU Digital Economy and Society Index (Digitalisierungsranking der EU) (Europäische Kommission, o. D.). Auch die Digitalisierung öffentlicher Dienstleistungen ist überdurchschnittlich weit fortgeschritten, wie auch der deutsche Verband der Informations- und Telekommunikationsbranche unterstreicht (Bitkom, o. D.).

Ein zentrales Merkmal der Verwaltungsstrukturen in Dänemark ist der Einsatz einer universell einsetzbaren Identifikationsnummer, der „MitID“ (deutsch: „Meine ID“, bis 2022 „NemID“, deutsch: „Einfache ID“), die von 85 Prozent der dänischen Bevölkerung genutzt wird. Hiermit sind u. a. die Nutzung verschiedener digitaler Portale für diverse Verwaltungsangelegenheiten möglich. Die

Webseiten borger.dk (KL & Uddannelses- og Forskningsministeriet, o. D.) für Bürger\*innen und virk.dk für Unternehmen (Erhvervsstyrelsen, o. D.) erlauben etwa Meldungen von Wohnsitzwechseln, Schulanmeldungen, Registrierungen neuer Unternehmen usw. Somit sind gedruckte Anträge bzw. Antworten von Behörden fast nicht mehr nötig. Insgesamt sind mehr als 100 Dienstleistungsgebiete innerhalb des öffentlichen Sektors nahezu vollständig digitalisiert (Dänisches Außenministerium, o. D.-a).

Die erfolgreiche Digitalisierung vieler Verwaltungsprozesse erstreckt sich auch auf den Bereich der dänischen Arbeitsverwaltung STAR (Styrelsen for Arbejdsmarked og Rekruttering), die für die Umsetzung der dänischen Arbeitsmarktpolitik zuständig ist. Diese folgt dem Konzept der Flexicurity, also der Verbindung von Sicherheit, die beispielsweise durch hohe Einkommensersatzleistungen gewährleistet wird, mit der Forderung nach Flexibilität und damit einer hohen Eigeninitiative der Arbeitslosen bei der Suche nach einer neuen Beschäftigung. Dazu werden entsprechende Anreize gesetzt (Österreichisches Bundesministerium für Arbeit und Wirtschaft, 2022; Bogedan, 2005; Danish Agency for Labour Market and Recruitment, 2022).

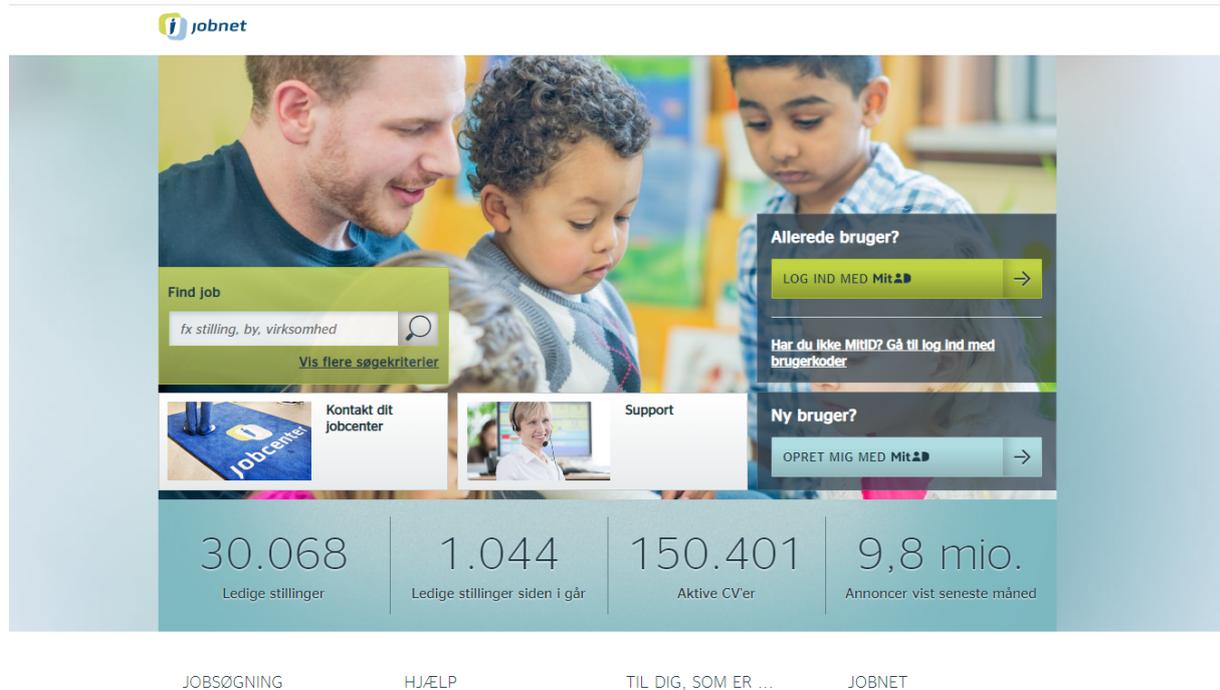
Die Arbeitsverwaltung STAR ist insbesondere auch für die Verwaltung der sozialen Mindestsicherung (dänisch: Kontanthjælp, im Folgenden als Bürgergeld bezeichnet) zuständig. Diese Leistungen werden nach einer Bedürftigkeitsprüfung an Arbeitsuchende gezahlt, die keine Ansprüche auf Versicherungsleistungen der Arbeitslosenversicherung (A-kasser) haben, da sie schon mehr als zwei Jahre arbeitslos sind. Da die Arbeitslosenversicherung in Dänemark nicht verpflichtend ist und von Beschäftigten individuell bei einem von mehreren privat organisierten Versicherungsunternehmen abgeschlossen werden muss, gibt es auch Erwerbslose, die direkt nach einem Beschäftigungsverlust auf Bürgergeld angewiesen sind.

Für den Bezug sowohl von Leistungen der Arbeitslosenkassen als auch von Bürgergeld gibt es unterschiedliche Auflagen. Beziehende von Versicherungsleistungen müssen dem Arbeitsmarkt zur Verfügung stehen und u. a. Angaben zum Nachweis von Suchaktivitäten machen. Ähnlich wie in Deutschland durch die Eingliederungsvereinbarungen (ab 1. Juli 2023: Kooperationspläne) werden auch in Dänemark individuelle Pflichten vereinbart, vor allem während des Bezugs von Versicherungsleistungen. Diese Pflichten sind relativ einheitlich geregelt. Die meisten Arbeitsuchenden müssen durchschnittlich 1,5 bis zwei Bewerbungen pro Woche verfassen und pro Monat mindestens zwei vollständige Bewerbungen oder Lebensläufe im Portal Joblog hochladen. Für den Bezug von Bürgergeld gelten andere Verpflichtungen. Hier müssen ab dem 13. Monat der Bezugsdauer trotz Arbeitslosigkeit mindestens 225 Arbeitsstunden z. B. in einer temporären Beschäftigung nachgewiesen werden (siehe Steckbrief 5.9). Werden die Pflichten für den Bezug von Versicherungsleistungen oder Bürgergeld verletzt, können Leistungen gekürzt bzw. gestrichen werden (CA Karrierepartner & A-Kasse, o. D.; United Federation of Danish Workers, o. D.; A-kasser, 2023; Dänisches Außenministerium, o. D.-b; Braun, 2003).

#### 6.2.2.2 Jobnet und Joblog

Das Portal Jobnet (Abbildung 6-3) bündelt verschiedene Dienstleistungen und Verwaltungsprozesse der Arbeitsverwaltung STAR bzw. der Arbeitslosenkassen. Zum einen enthält das Portal eine umfassende Stellenbörse mit diversen Such- und Filtermöglichkeiten. Zum anderen werden im Joblog seit 2015 Aktivitäten zur Stellensuche von Arbeitsuchenden eingetragen. Insbesondere können auf Joblog Informationen zu verfassten Bewerbungen gespeichert werden. Diese Angaben dienen auch zum Nachweis der Erfüllung der Auflagen für den Bezug von Bürgergeld bzw. Leistungen der Arbeitslosenversicherung. Zudem können Arbeitsuchende weitere persönliche Einträge wie Kalendereinträge und Notizen hinzufügen.

Abbildung 6-3: Jobnet-Portal



Quelle: Jobnet (o. D.).

Die Nutzung von Joblog und eine wöchentliche Aktualisierung der eigenen Aktivitäten ist für Empfänger\*innen von Bürgergeld und Leistungen der Arbeitslosenkassen verpflichtend und die Joblog-Inhalte dienen als Ausgangspunkt für Beratungsgespräche von Arbeitsuchenden mit Berater\*innen der Arbeitsverwaltung. Die Berater\*innen haben ebenfalls Zugriff auf die in den individuellen Joblog-Profilen gespeicherten Daten. Damit bietet Joblog sowohl den Berater\*innen als auch den Arbeitsuchenden selbst einen guten Überblick über den individuellen Suchfortschritt (Styrelsen for Arbejdsmarked og Rekruttering, 2022).

### 6.2.3 Anwendung

Die umfangreichen Daten, die in Jobnet und Joblog zur Verfügung stehen, ermöglichen die Beantwortung vieler Forschungsfragen rund um das Matching auf dem Arbeitsmarkt und insbesondere das Suchverhalten von Arbeitslosen. Dabei nutzen die Autor\*innen vieler Studien die Möglichkeit, die Jobnet-Daten bzw. die Joblog-Einträge mit weiteren dänischen administrativen Datensätzen verknüpfen zu können (z. B. Informationen zu öffentlichen Transferleistungen, Styrelsen for Arbejdsmarked og Rekruttering, o. D.).

Ein aktuelles Diskussionspapier von Fluchtman et al. (2021) untersucht unter Verwendung der Daten von Joblog, inwiefern sich die Stellen, auf die sich Frauen bzw. Männer bewerben, unterscheiden. Da in den Joblog-Daten eine Identifikationsnummer enthalten ist, die für viele verschiedene administrative Vorgänge verwendet wird, können die Bewerbungen mit Unternehmensdaten bzw. Beschäftigungsverhältnissen, die sich aus erfolgreichen Bewerbungen ergeben, verknüpft werden. Auf dieser Grundlage können zahlreiche Merkmale der Stellen, für die Bewerbungen versendet werden, ermittelt bzw. berechnet werden. U. a. verwenden die Autor\*innen die Unternehmen- und Beschäftigteninformationen für Vorhersagemodelle, um zu berechnen, wie hoch der erwartete Lohn einer Stelle unabhängig von der Person ist, die tatsächlich eingestellt wird. Die Analyse dieser Zielgröße ist der Hauptfokus des Forschungspapiers. Anhand der in den Daten

beobachteten Bewerbungen kann untersucht werden, inwiefern die in vielen bisherigen Forschungsarbeiten beschriebenen geschlechterspezifischen Lohndifferenziale auch von einem unterschiedlichen Bewerbungsverhalten von Frauen und Männern abhängen.

Die in der Studie von Fluchtmann et al. (2021) verwendete Stichprobe besteht aus Arbeitslosen, die mindestens acht Wochen lang Leistungen einer Arbeitslosenkasse beziehen, mindestens vier Bewerbungen versenden und innerhalb eines Jahres eine neue Beschäftigung finden. Diese Stichprobe umfasst über 100.000 Personen und knapp drei Mio. Bewerbungen. Die Ergebnisse der Studie zeigen, dass ein Großteil der Unterschiede in der Entlohnung von Frauen und Männern, der nicht durch Merkmale der Beschäftigten bzw. Unternehmen selbst erklärt werden kann, auf ein unterschiedliches Bewerbungsverhalten zurückzuführen ist. Dies kann einerseits eine Reaktion auf Erwartungen hinsichtlich des Einstellungsverhaltens sein: Frauen bewerben sich deshalb nicht auf Stellen in typischen Männerberufen, weil sie glauben, dass Männer bei der Einstellung bevorzugt werden. Andererseits spiegeln sich in den Bewerbungen auf unterschiedliche Stellen auch geschlechtsspezifische Rollenbilder wider. So bewerben sich Frauen häufiger auf Teilzeitstellen sowie bei Unternehmen, die näher am eigenen Wohnort liegen und familienfreundlicher sind. Dies lässt eine traditionelle Rollenverteilung vermuten, bei denen Frauen die Kinderbetreuung übernehmen und Männer die Hauptverdiener sind.

Ein weiteres Forschungspapier (Maibom et al., 2023) verwendet den gleichen Datensatz und eine ähnliche Stichprobe, die aus Leistungsbeziehenden der Arbeitslosenversicherung besteht. Hier untersuchen die Autor\*innen das Suchverhalten von Arbeitslosen – unabhängig vom Geschlecht – über den Verlauf der Arbeitslosigkeit. So kann u. a. gezeigt werden, dass sich die Lohnerwartungen von Bewerber\*innen sehr stark unterscheiden, jedoch über die Zeit hinweg leicht abnehmen. Die Lohnerwartungen reagieren somit dynamisch auf Erfolge bzw. Misserfolge im Bewerbungsprozess. Darüber hinaus liegen weitere Forschungspapiere vor, die die Daten im Rahmen von Feldexperimenten nutzen (Steckbrief 5.9).

#### 6.2.4 Übertragbarkeit auf Deutschland

Die beschriebene Dateninnovation zeichnet sich dadurch aus, dass umfassende Informationen über Stellenangebote (Arbeitsnachfrage) und Arbeitslose und deren Bewerbungen (Arbeitsangebot) zusammengeführt und in der Verwaltung genutzt werden können, was die Beratung und die Kontrolle der Erfüllung von Auflagen zum Leistungsbezug erleichtert. Zudem stehen die Daten der Forschung zur Verfügung. Grundsätzlich erscheint eine Übertragbarkeit auf Deutschland möglich. So enthalten Jobbörse-Daten<sup>58</sup> der BA die notwendigen Informationen auf Seiten der Arbeitsnachfrage, in den Jobcentern werden grundsätzlich auch Informationen zum Arbeitsangebot, also auch zu Bewerbungen, gesammelt. Diese müssten für eine umfassende Auswertung jedoch standardisierter erfasst und für die Analyse verfügbar gemacht und aufbereitet werden. Zudem erfordern Datenschutzbestimmungen die explizite Zustimmung der Betroffenen, sofern deren Informationen zum Bewerbungsverhalten auf individueller Ebene genutzt werden sollen. (siehe Steckbrief 4.2).

#### 6.2.5 Fazit

Das dänische Portal Jobnet mit den Funktionen von Joblog sammelt und bündelt umfassende Informationen zur Arbeitsnachfrage von Unternehmen, zum Arbeitsangebot und zu den Suchbemühungen von Arbeitslosen. Diese Informationen können für verschiedene Zwecke genutzt werden. So können zum einen Leistungen rund um die Beratung und Vermittlung von Arbeitslosen

---

<sup>58</sup> Die amtliche Stellenbörse (Jobbörse) der BA wurde kürzlich in das Portal [arbeitsagentur.de](https://arbeitsagentur.de) integriert.

effizienter vergeben werden. Zum anderen können die Daten für Forschungszwecke genutzt und mit weiteren Daten verknüpft werden. Die bisherigen Studien auf der Basis dieser Daten weisen auf ein großes Analysepotenzial hin.

### 6.2.6 Quellen zum Steckbrief

A-kasser. (2023a, 3. April). *Arbeitslosengeld*. A-kasser. Abgerufen am 11. April 2023 von <https://www.a-kasser.dk/voraussetzungen/index.html>

Bitkom. (o. D.). *Dänemark / Verwaltungsleistungen so einfach wie Onlineshopping*. Bitkom. Abgerufen am 11. April 2023 von <https://www.bitkom.org/Themen/Politik-Recht/Oeffentliche-Verwaltung/Daenemark-Verwaltungsleistungen-so-einfach-wie-Onlineshopping>

Bogedan, C. (2005). *Mit Sicherheit besser? Aktivierung und Flexicurity in Dänemark* (ZeS-Arbeitspapier, 6/2005). ZES-Arbeitspapiere. [https://www.ssoar.info/ssoar/bitstream/handle/document/10930/ssoar-2005-bogedan-mit\\_sicherheit\\_besser.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2005-bogedan-mit\\_sicherheit\\_besser.pdf](https://www.ssoar.info/ssoar/bitstream/handle/document/10930/ssoar-2005-bogedan-mit_sicherheit_besser.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2005-bogedan-mit_sicherheit_besser.pdf)

Braun, T. (2003). *Ein neues Modell für Flexicurity – der dänische Arbeitsmarkt* (WSI Mitteilungen 2/2003). WSI Mitteilungen. [https://www.boeckler.de/data/wismit\\_2003\\_02\\_braun.pdf](https://www.boeckler.de/data/wismit_2003_02_braun.pdf)

CA Karrierepartner & A-Kasse. (o. D.). *Registering your job search activities in your job log*. Unemployment Insurance. Abgerufen am 11. April 2023 von <https://www.ca.dk/en/the-benefit-rules/registering-your-job-search-activities-in-your-job-log/>

Dänisches Außenministerium. (o. D.-a). *Digitalisierung in Dänemark*. Tyskland.um. Abgerufen am 11. April 2023 von <https://tyskland.um.dk/de/aussenwirtschaftsrat/digitalisierung>

Dänisches Außenministerium. (o. D.-b). *Soziale Leistungen*. Tyskland.um. Abgerufen am 11. April 2023 von <https://tyskland.um.dk/de/uber-danemark/leben-und-arbeiten/soziale-leistungen>

Danish Agency for Labour Market and Recruitment. (2022, 5. Juli). *Flexicurity*. Danish Agency for Labour Market and Recruitment. Abgerufen am 11. April 2023 von <https://www.star.dk/en/about-the-danish-agency-for-labour-market-and-recruitment/flexicurity/>

Erhvervsstyrelsen. (o. D.). Virk.dk. Virk.dk. Abgerufen am 14. April 2023 von <https://virk.dk/vejledning/virk-om-virk/>

Europäische Kommission. (o. D.). *The Digital Economy and Society Index (DESI)*. Digital Strategy – European Commission. Abgerufen am 11. April 2023 von <https://digital-strategy.ec.europa.eu/en/policies/desi>

Fluchtmann, J., Glenny, A., Harmon, N. & Maibom, J. (2021). *The Gender Application Gap: Do Men and Women Apply for the Same Jobs?* (IZA DP No. 14906). IZA Discussion Paper Series. <https://docs.iza.org/dp14906.pdf>

Jobnet. (o. D.). *Jobnet-Portal*. Jobnet. Abgerufen am 11. April 2023 von <https://job.jobnet.dk/CV/frontpage>

KL & Uddannelses- og Forskningsministeriet. (o. D.). Borger.dk. Abgerufen am 14. April 2023 von <https://www.borger.dk/>

Maibom, J., Glenny, A., Harmon, N. & Fluchtmann, J. (2023). Unemployed Job Search Across People and Over Time: Evidence from Applied-For Jobs. Demnächst in *Journal of Labor Economics*.  
<https://doi.org/10.1086/725165>

Österreichisches Bundesministerium für Arbeit und Wirtschaft. (2022, 2. März). *Das dänische Arbeitslosenversicherungsmodell setzt auf effiziente Vermittlung bei hoher Einkommenssicherung*. Bundesministerium Arbeit und Wirtschaft. Abgerufen am 11. April 2023 von <https://www.bmaw.gv.at/Presse/News/Kocher-in-Kopenhagen.html>

Styrelsen for Arbejdsmarked og Rekruttering (o. D.). DREAM vejledning version 47 [DREAM-Leitfaden Version 47]. Styrelsen for Arbejdsmarked og Rekruttering.  
<https://www.dst.dk/Site/Dst/SingleFiles/GetArchiveFile.aspx?fi=36881106234&fo=0&ext=forskning>

Styrelsen for Arbejdsmarked og Rekruttering. (2022, 5. Oktober). *Joblog – borgerens værktøj til en struktureret jobsøgning* [Joblog – das Bürgertool für eine strukturierte Stellensuche]. Styrelsen for Arbejdsmarked og Rekruttering. Abgerufen am 11. April 2023 von <https://star.dk/it/borger-it/find-job-og-joblog/joblog-borgerens-vaerktoej-til-en-struktureret-jobsoegning/>

United Federation of Danish Workers. (o. D.). *Du skal registrere din jobsøgning i joblog* [Sie müssen Ihre Jobsuche im Joblog registrieren]. 3F. Abgerufen am 11. April 2023 von <https://www.3f.dk/find-svar/jobsoegning/joblog>

## 6.3 The Data-Driven Skills Taxonomy (Vereinigtes Königreich)

### 6.3.1 Einführung

Das Ziel dieser Innovation ist es, Informationen über die derzeit nachgefragten Kompetenzen und deren Wert auf dem britischen Arbeitsmarkt bereitzustellen. Insbesondere im Lichte der Digitalisierung und weiterer globaler Trends entsteht durch solche Informationen ein Mehrwert für Bürger\*innen sowie politische Akteure (siehe auch Stops et al., 2021). Des Weiteren kann im Rahmen dieser Innovation über die Webseite (Djumaliewa & Sleeman, o. D.) in Erfahrung gebracht werden, welche Kompetenzen für die Ausübung eines bestimmten Berufs benötigt werden und welche Berufe ähnliche Kompetenzprofile haben. Zudem lässt sich mithilfe der Innovation darstellen, inwieweit sich die Nachfrage nach Kompetenzen im Vereinigten Königreich im Zeitverlauf ändert. Darüber hinaus wird das zu erwartende Einkommen aus der Ausübung der unterschiedlichen Tätigkeiten, die einer Kompetenz zugeordnet sind, abgebildet. Das Projekt wird von Nesta<sup>59</sup>, dem Economic Statistics Centre of Excellence (ESCoE) und dem Office for National Statistics (ONS) im Vereinigten Königreich durchgeführt.

Die hier beschriebene Taxonomie basiert auf der Auswertung von Daten in Form von öffentlich verfügbaren Stellenanzeigen mithilfe von Machine-Learning-Methoden. Die Anzahl der identifizierten Kompetenzen in dieser datenbasierten Taxonomie hängt stark von der Ausgestaltung der Parameter der genutzten Algorithmen ab. Aufgrund dieser Volatilität ist eine datenbasierte Kompetenz-Taxonomie, wie die hier beschriebene Taxonomie aus dem Vereinigten Königreich, als Ergänzung und als Orientierung für etwaige Aktualisierungen bestehender Taxonomien zu verstehen (Gallagher et al., 2022).

Das Potenzial einer solchen Innovation für Deutschland liegt zum einen in der Möglichkeit, die Nachfrage nach Kompetenzen auf dem Arbeitsmarkt besser und aktueller darzustellen. So lässt sich eine Aktualisierung der Kompetenzen leicht anhand aktueller Stellenanzeigen durchführen. Schnelle Änderungen der nachgefragten Kompetenzen, wie beispielsweise durch die Covid-19-Pandemie, lassen sich somit gut nachzeichnen. Gleiches gilt für die Erfassung neu entstehender Kompetenzen oder Verschiebungen in der Relevanz einzelner Kompetenzen im Zeitverlauf. Zum anderen präsentiert die Innovation aus dem Vereinigten Königreich ihre Ergebnisse interaktiv und zugänglich auf einem Webportal. Eine solche umfangreiche Visualisierung wird in Deutschland in einer ähnlichen Innovation, dem IAB-Kompetenz-Kompass, bisher nicht angeboten.

### 6.3.2 Technische Beschreibung

#### 6.3.2.1 Aufbereitung der Stellenanzeigen und Erstellung der Taxonomie

Die zentrale Komponente der Innovation ist die Extraktion der auf dem Arbeitsmarkt geforderten Kompetenzen aus aktuellen öffentlichen Stellenanzeigen. Die Erstellung der Kompetenz-Taxonomie<sup>60</sup> fand durch die Projektpartner in Zusammenarbeit mit TextKernel statt, einem Unternehmen, das Machine-Learning-Verfahren für Prozesse im Personalmanagement entwickelt

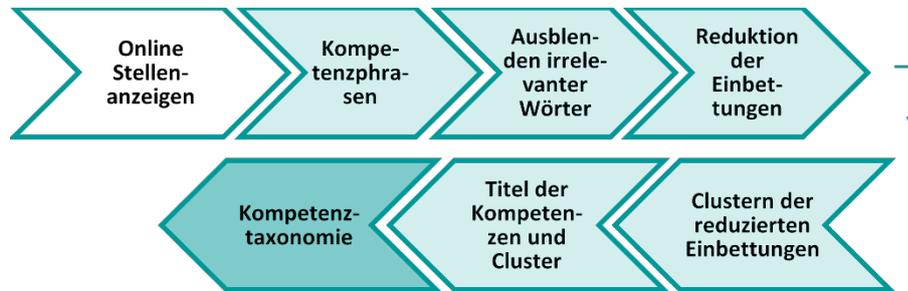
---

<sup>59</sup> Nesta ist eine seit 1998 bestehende britische Innovationsagentur, die ihre Arbeit an sozialen Belangen orientiert. Das Ziel ist die Entwicklung von Innovationen mit gesellschaftlichem Mehrwert.

<sup>60</sup> Die derzeit online verfügbare Taxonomie basiert auf der von Djumaliewa & Sleeman (2018) entwickelten Methodik. Einen weiteren Ansatz, der auf den Originaltexten der Stellenanzeigen beruht, bieten Gallagher et al. (2022). In ihrem Forschungspapier wird eine alternative Methode zur Identifikation von Kompetenzen in den Anzeigen und Einteilung der Gruppen erläutert. Um den neuesten Stand der Forschung darzustellen, wird im Folgenden insbesondere auf diese neue Vorgehensweise Bezug genommen. Diese Ansätze sind in der im Internet abrufbaren Taxonomie bisher jedoch noch nicht implementiert.

(Gallagher et al., 2022). Grundlage für die Erstellung der Taxonomie sind die Inhalte der Stellenanzeigen. Im Detail haben die Projektpartner Zugriff auf die Volltexte der Anzeigen sowie Informationen zu Veröffentlichungsdatum der Anzeigen, des Berufstitels und der Region im Vereinigten Königreich. Die Basis für die aktuelle Taxonomie bildet eine Stichprobe von über 62 Mio. Online-Stellenanzeigen, die über den Zeitraum von 2015 bis 2021 gesammelt wurden. Für die Analyse wurden jedoch nur fünf Millionen Anzeigen genutzt, da diese Stichprobe ausreichend groß war und Rechenzeit gespart werden konnte. Die einzelnen Schritte, von den Stellenanzeigen bis zur Erstellung der Taxonomie, sind in Abbildung 6-4 dargestellt.

**Abbildung 6-4: Überblick über die Methodologie**



Quelle: Eigene Darstellung in Anlehnung an Gallagher et al. (2022).

Im Folgenden werden die Schritte kurz dargestellt, allerdings nicht näher beschrieben, da die dabei verwendeten Methoden kein Alleinstellungsmerkmal der Innovation sind. In Deutschland finden ähnliche Methoden bei der Erstellung des IAB-Kompetenz-Kompass Anwendung (Stops et al., 2021).

Um die Taxonomie zu entwickeln, werden zunächst die unterschiedlichen Kompetenzen aus den Stellenanzeigen extrahiert. Dafür wird ein Klassifizierungssystem aufgebaut, das Sätze erkennt, in denen Kompetenzen enthalten sind. Mithilfe dieses Systems konnten ca. 20 Millionen dieser Sätze identifiziert werden. Aus diesen wurden irrelevante Wortarten wie Eigennamen und Interpunktionszeichen ausgeblendet. Nach der erfolgten Reduktion der Kompetenzphrasen wurden in einem dritten Schritt konkrete Kompetenzen identifiziert. Dies erfolgte durch die Bildung von Clustern, wobei ein Cluster je eine Kompetenz repräsentieren soll. Nach der Anwendung des Clusteralgorithmus auf die Stichprobe ergaben sich 6.685 verschiedene Gruppen, die je eine Kompetenz darstellen. Durch das beschriebene Vorgehen können auch Kompetenzen identifiziert werden, die vergleichsweise selten in den Stellenanzeigen vorkamen (200 oder weniger Nennungen).

Im letzten Schritt wurden die Kompetenzen automatisch benannt und in eine grobe, übergeordnete und anschließend eine feine, detaillierte Berufsgruppe eingebettet. Beispielsweise wird der Beruf eines/einer Buchhalter\*in zunächst der Betriebswirtschaft und dort dem Rechnungswesen zugeordnet. Die am häufigsten nachgefragten Kompetenzen sind Lohn- und Steuerbuchhaltung („payroll and tax accounting“) sowie Rechnungswesen und Finanzverwaltung („accounting and financial management“). Für detaillierte Ausführungen, die die Methodik und Algorithmen näher beschreiben, wird auf Djumalieva & Sleeman (2018), Djumalieva et al. (2018) sowie McInnes et al. (2018) und Ester et al. (1996) verwiesen.

### 6.3.2.2 Unterschied zu katalogbasierten Kompetenz-Taxonomien

Die hier beschriebene Kompetenz-Taxonomie aus dem Vereinigten Königreich gehört zu den datenbasierten Taxonomien, denn sie wird hauptsächlich aus den Inhalten der vorliegenden Stellenanzeigen generiert, die in diesem Sinne die Datenbasis oder Grundgesamtheit darstellen.

Damit unterscheidet sie sich von anderen Taxonomien wie BERUFENET<sup>61</sup> der BA, der ESCO<sup>62</sup> oder dem O\*NET-Portal<sup>63</sup>. Diese Taxonomien stützen sich auf die Einschätzung und Bewertung durch Expert\*innen der Arbeitsmarkt- oder Bildungsforschung. Sie können als katalogbasierte Taxonomien bezeichnet werden, da sie durch eine qualitative Auswertung verschiedener Quellen, wie beispielsweise Ausbildungsverträge oder Nennung in Tarifabschlüssen (Stops et al., 2021), erstellt werden.

Die datenbasierte Herangehensweise bietet den Vorteil, dass die Begriffe so erfasst werden, wie sie auf dem Arbeitsmarkt verwendet werden. Somit können sprachliche Verschiebungen abgebildet werden und Arbeitsuchende können ihre Kompetenzen beispielsweise in Bewerbungsschreiben präzise benennen. Des Weiteren werden neuartige Kompetenzen schnell erkannt und die Taxonomie kann aufgrund der automatisierten Verarbeitung der Stellenanzeige einfach aktualisiert werden. Auch hier ist der Einsatz von Expert\*innen notwendig, insbesondere für die Qualitätskontrolle der Erkennung der Sätze in Stellenanzeigen, in denen Kompetenzen enthalten sind. Zudem müssen die Cluster der herausgearbeiteten Kompetenzen manuell anhand der jeweiligen Begriffe in den Clustern benannt werden.

Das Ziel bei der Identifizierung von Kompetenzen in den Stellenausschreibungen sind robuste Kompetenzgruppen. Dies erfordert eine ex-ante Festlegung der Granularität. In der beschriebenen Version der Taxonomie werden circa 6.700 Kompetenzen durch den Algorithmus identifiziert. Die durch Expert\*innen erstellte ESCO-Taxonomie umfasst etwas weniger als 14.000 Hard und Soft Skills. Die Auswahl der Parameter, z. B. die Zahl der Dimensionen, auf die die Worteinbettungen reduziert werden, oder die Kriterien zur Beschneidung von Sätzen und zum Ausblenden irrelevanter Begriffe, müssen demnach festgelegt werden, denn sie beeinflussen die Anzahl der identifizierten Kompetenzen maßgeblich (Gallagher et al., 2022). Andernfalls kann nach Erstellung der Taxonomie die Anzahl an Kompetenzen willkürlich erscheinen und damit gegenüber katalogbasierten Taxonomien für analytische bzw. wissenschaftliche Zwecke weniger nutzbar erscheinen. Obwohl die Anzahl der Kompetenzen der datenbasierten Taxonomie im Vergleich zur katalogbasierten Taxonomie bereits jetzt gering ausfällt, erwägen die Autor\*innen, in künftigen Anpassungen des Algorithmus einzelne Kompetenzcluster zusammenzulegen, um die Robustheit der Kompetenzen zu erhöhen. Alternativ kann durch eine Variation der Parameter der Methoden zur Identifizierung der Kompetenzen eine höhere Anzahl an Kompetenzen erreicht werden, die dann allerdings womöglich weniger robust sind.

### 6.3.3 Anwendung

Die Taxonomie kann interaktiv über eine öffentlich zugängliche Webseite aufgerufen werden. In Abbildung 6-5 wird das Vorgehen am Beispiel einer Lehrkraft für das Fach Englisch dargestellt. Wie bereits beschrieben, gibt es in einem ersten Schritt eine zweistufige Zuordnung zu Berufsgruppen, gefolgt von den relevantesten geforderten Kompetenzen und einer Angabe zum Medianeinkommen. Im dargestellten Beispiel erfordert die Tätigkeit von Englischlehrkräften die Kompetenz „teaching“. Angezeigt werden im Bereich „teaching“ Abgrenzungen, z. B. nach der inhaltlichen Ausrichtung der Tätigkeit. Würde man demnach nach den erforderlichen Kompetenzen einer Chemielehrkraft suchen,

---

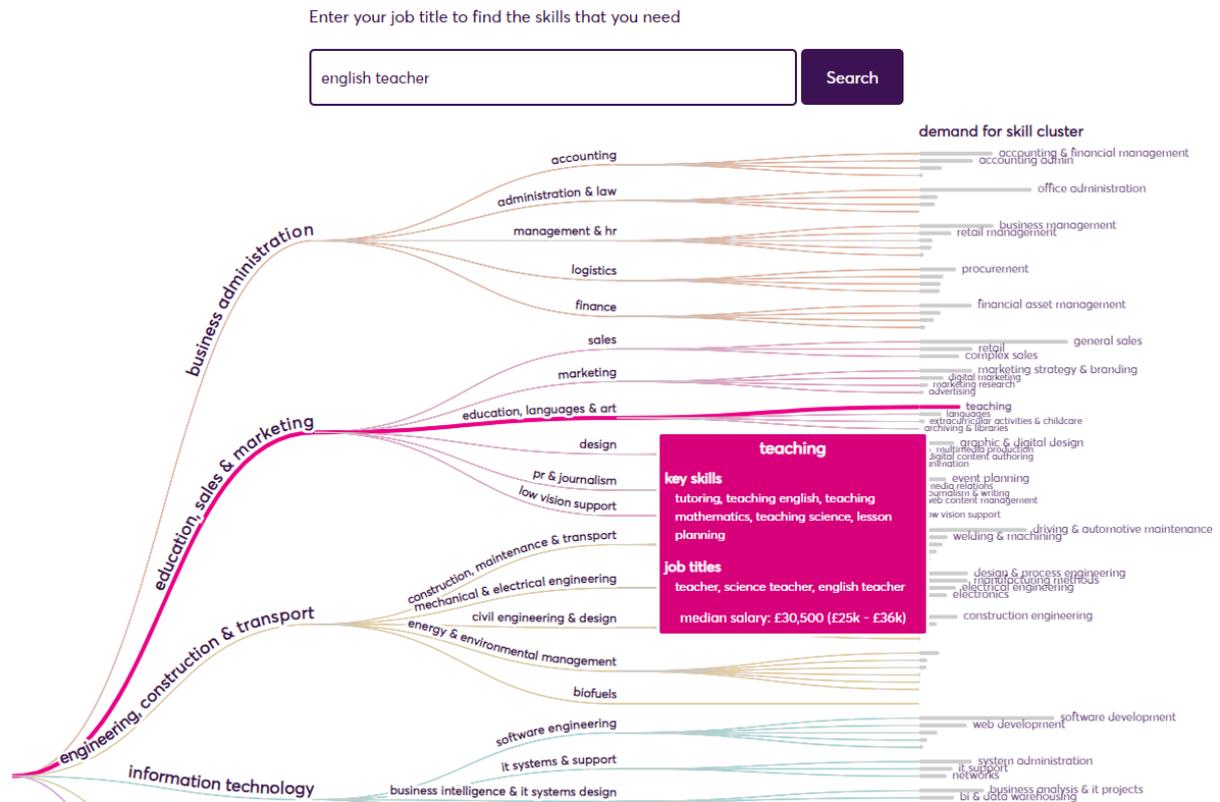
<sup>61</sup> Diese Taxonomie beschreibt umfangreich konkrete Berufe, beispielsweise hinsichtlich ihrer Tätigkeiten oder Berufsperspektiven (Bundesagentur für Arbeit, o.D.)

<sup>62</sup> Diese auf Europa bezogene Taxonomie wurde 2017 veröffentlicht und umfasst etwas weniger als 3.000 Berufsbeschreibungen sowie circa 13.500 Kompetenzen.

<sup>63</sup> Dieses Klassifizierungssystem identifiziert circa 1.100 Berufe, die anhand nötiger Qualifikationen, Fähigkeiten, Wissen und Arbeitstätigkeiten näher beschrieben werden.

würde man zunächst ebenfalls in den Bereich „teaching“ geleitet. Darüber hinaus können sich Nutzende über die Möglichkeiten eines Berufswechsels informieren, indem sie Tätigkeiten mit ähnlichen Kompetenzprofilen identifizieren.

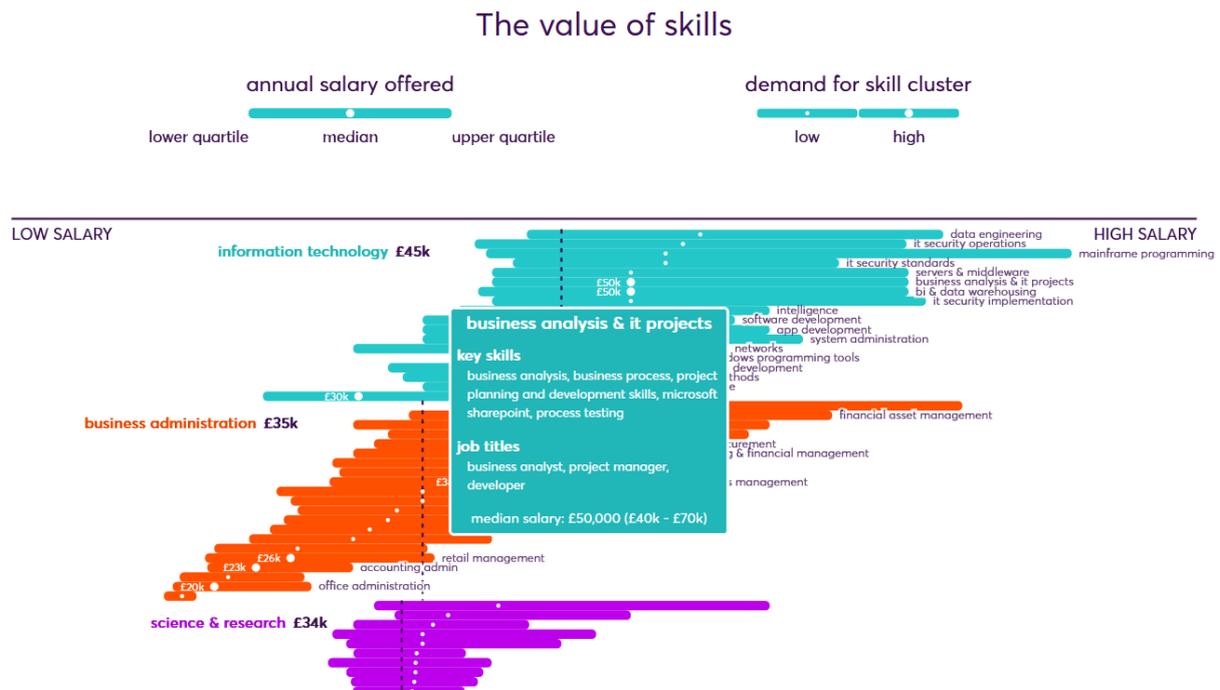
Abbildung 6-5: Anwendung der Kompetenz-Taxonomie - Berufe



Quelle: Djumalieva & Sleeman (o. D.). Zuletzt abgerufen: 22.03.2023.

Des Weiteren stellt die Webseite aggregierte Informationen zur Entlohnung verschiedener Kompetenzen und deren Nachfrage auf dem Arbeitsmarkt dar. Dabei wird das untere und obere Quartil des Einkommens sowie das Medianeinkommen dargestellt (Abbildung 6-6). Die Höhe der Nachfrage nach einer Kompetenz wird durch die Stärke der Punkte des Medianeinkommens visualisiert.

Abbildung 6-6: Anwendung der Kompetenz-Taxonomie - Wert der Kompetenzen



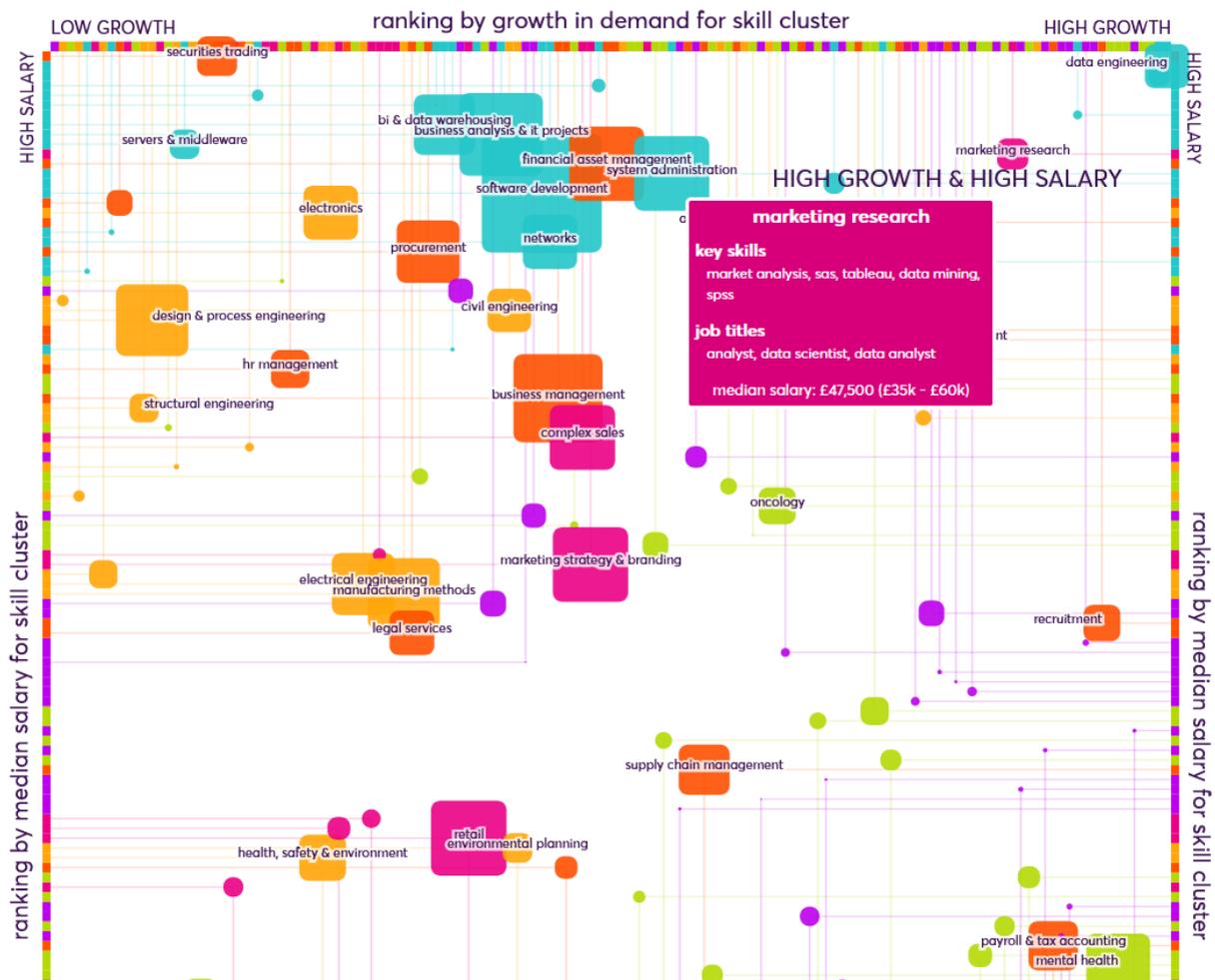
Quelle: Djumalieva & Sleeman (o. D.). Zuletzt abgerufen: 22.03.2023.

Die Kompetenz-Taxonomie wird auch in der Forschung verwendet. Gallagher et al. (2022) nutzen sie, um die räumliche Verteilung von Kompetenzen im Vereinigten Königreich zu messen und den Einfluss der Covid-19-Pandemie auf die Nachfrage nach Kompetenzen zu evaluieren. Beispielsweise zeigen sie, dass das größte Wachstum ab März 2020 jene Kompetenzen verzeichneten, die besonders für die Arbeit in der medizinischen Versorgung benötigt werden.

Als weiterer Baustein kann, wie bereits erwähnt, auf der Webseite die Entwicklung der Nachfrage nach einer Kompetenz und des damit verbundenen Einkommens abgerufen werden (Abbildung 6-7). Dabei zeigt sich beispielsweise, dass die Kompetenz Data Engineering sowohl ein hohes Einkommen als auch eine wachsende Nachfrage auf dem Arbeitsmarkt verzeichnet.<sup>64</sup> Die Verknüpfung mit Einkommensdaten ist ein besonders innovatives Element der Taxonomie.

<sup>64</sup> Betrachtet werden die Zeiträume 2012-2014 und 2015-2017.

Abbildung 6-7: Anwendung der Kompetenz-Taxonomie - Arbeitsnachfrage und Einkommen



Quelle: Djumalieva & Sleeman (o. D.). Zuletzt abgerufen: 22.03.2023.

### 6.3.4 Übertragbarkeit auf Deutschland

In Deutschland wurde eine explizite datenbasierte Taxonomie von Kompetenzen für die Analyse von Kompetenzanforderungen in der Gesundheits- und Pflegebranche in Bayern verfolgt (vbw, 2020). Eine Darstellung und Zugänglichkeit über ein frei zugängliches Portal wie bei der Taxonomie im Vereinigten Königreich ist hier jedoch nicht erfolgt. Zudem beschränkt sich diese Taxonomie auf Stellenanzeigen in der Gesundheits- und Pflegebranche (mit insgesamt 50.000 Stellenanzeigen in Bayern).

Mit dem Kompetenz-Kompass des IAB ist in Deutschland hingegen bereits eine weitgehende Verwirklichung einer katalogbasierten Taxonomie vorhanden, die auch Elemente einer datenbasierten Taxonomie aufweist (Stops et al., 2021). Dieses Pilotprojekt hat zunächst das Ziel, aktuell nachgefragte fachliche und überfachliche Kompetenzen des deutschen Arbeitsmarktes abzubilden. Exemplarisch werden dafür die Branchen Maschinenbau, Informationsdienstleistungen sowie das Gesundheits- und Sozialwesen betrachtet. Der hierbei genutzte Kompetenz-Katalog wird durch die Analyse der Stellenanzeigen um neue, bei der Betrachtung der Stellenanzeigen gefundene Kompetenzen, erweitert. Diese Form der Katalog-Aktualisierung durch datenbasierte Analysen kommt häufig bei katalogbasierten Taxonomien zum Einsatz (Flake et al., 2020). Die Stellenanzeigen, die zur Erstellung des IAB-Kompetenz-Kompass genutzt werden, werden, anders als bei der vorgestellten Dateninnovation aus dem Vereinigten Königreich, nicht mithilfe von frei zugänglichen

Stellenanzeigen und Text-Mining Methoden gesammelt, sondern aus einer zentralen Quelle entnommen, der Jobbörse der BA. Eine weitere Taxonomie, die Stellenanzeigen mithilfe von Machine-Learning-Methoden analysiert und ebenfalls auf einem Kompetenz-Katalog aufbaut, wurde in Baden-Württemberg zur Bestimmung der beruflichen Kompetenzen in der Automobil- und Zulieferindustrie erstellt (Flake et al., 2020).

Zusammenfassend zeigt sich somit, dass es in Deutschland bereits sowohl katalogbasierte als auch datenbasierte Kompetenz-Taxonomien gibt. Diese haben bisher jedoch einen deutlich geringeren Umfang und beschränken sich teilweise auf einzelne Branchen. Zudem gibt es aktuell kein Portal, auf dem Interessierte interaktiv auf die Erkenntnisse der Taxonomien, insbesondere im Zeitverlauf, zugreifen können. Die Lage ändert sich jedoch laufend. So ist für den Kompetenz-Kompass des IAB geplant, die Betrachtung auf den gesamten Arbeitsmarkt auszuweiten, auch unter Berücksichtigung von dynamischen Entwicklungen der geforderten Kompetenzen über die Zeit. Beispielsweise sollen stärker nachgefragte Kompetenzen (sogenannte emerging skills) gezielt betrachtet werden. Ferner hat der Kompetenz-Kompass das mittelfristige Ziel, auch für die breite Öffentlichkeit leicht zugänglich zu sein (Stops et al., 2021). Dafür sind allerdings noch Validierungsschritte hinsichtlich der Repräsentativität der Kompetenzen sowie Partner für den Aufbau eines webbasierten Portals erforderlich.

Ein Vergleich mit der von Expert\*innen entwickelten ESCO-Taxonomie könnte helfen, die Verzerrung in der datenbasierten Taxonomie aus dem Vereinigten Königreich durch nicht-erfasste, nicht online ausgeschriebene Stellenanzeigen besser einschätzen zu können. Dieser Schritt ist für den IAB-Kompetenz-Kompass ebenfalls angedacht.

Seitens der auswertbaren öffentlichen Stellenanzeigen, die für die Erstellung einer Taxonomie genutzt werden können, bestehen insbesondere bei privaten Anbietern teilweise rechtliche Einschränkungen für die Nutzung und Weiterverarbeitung. Zudem ist es im Gegensatz zur Praxis im Vereinigten Königreich nicht die Regel, dass Informationen zum zu erwartenden Einkommen in den Stellenanzeigen verfügbar sind. Daher müssten hier weitere Quellen genutzt werden. Hier könnte die Infrastruktur der BA bzw. des IAB genutzt werden, um Stellenanzeigen aus der Stellenbörse sowie Einkommensinformationen aus den administrativen Daten der BA zu verknüpfen.

Die notwendigen Methoden und Machine-Learning-Algorithmen zur Erstellung einer datenbasierten Kompetenz-Taxonomie sind in der Forschung geläufig und auch im Kontext der Erstellung der hier beschriebenen Innovation ausgearbeitet. Zum einen sind die von Nesta und der ESCO angewandten Algorithmen öffentlich zugänglich. Zum anderen werden verschiedene technische Ansätze in dem Forschungspapier von Djumalieva et al. (2018) diskutiert. Darüber hinaus existieren bereits am IAB und weiteren Forschungsinstituten in Deutschland die notwendigen Fachkenntnisse, um auf Basis von öffentlichen Stellenanzeigen aktuell auf dem Arbeitsmarkt gefragte Kompetenzen zu identifizieren (Flake et al., 2020; Stops et al, 2021; vbw, 2020).

### 6.3.5 Fazit

Die Data-Driven Skills Taxonomy aus dem Vereinigten Königreich extrahiert aus öffentlichen Stellenanzeigen mithilfe von Machine-Learning-Methoden aktuell nachgefragte Kompetenzen auf dem Arbeitsmarkt. Inhaltlich lassen sich die identifizierten Kompetenzen auf Basis der Analyse der Stellenanzeigen mit verschiedenen Merkmalen verknüpfen, wie den dazu passenden Berufen oder Einkommensinformationen. Dieses Spektrum an Informationen wird in einem interaktiven Portal auf einer Webseite dargestellt. Über dieses Portal ist auch der Verlauf der Nachfrage einzelner Kompetenzen über die Zeit nachvollziehbar. Für Arbeitsuchende und politische Akteure ist die Innovation daher ein hilfreiches Tool. Eine solche datenbasierte Taxonomie zeichnet sich im Vergleich zu einer katalogbasierten Taxonomie insbesondere durch eine höhere Aktualität aus, da

neu auftretende Kompetenzen zeitnah erfasst werden können. Nichtsdestotrotz bedarf es auch bei einer datenbasierten Taxonomie, ähnlich wie bei einer katalogbasierten Taxonomie, manueller Arbeiten, insbesondere im Bereich der Qualitätssicherung.

In Deutschland werden nachgefragte Kompetenzen auf dem Arbeitsmarkt bereits von verschiedenen Akteuren erfasst und publiziert. Der IAB-Kompetenz-Kompass und das Portal BERUFENET weisen einige Ähnlichkeiten mit der Innovation aus dem Vereinigten Königreich auf. Beispielsweise ist mit BERUFENET ein Vergleich verschiedener Berufe anhand der geforderten Kompetenzen möglich. In diesem Zusammenhang ist die vorgestellte Dateninnovation vor allem hinsichtlich ihres interaktiven Webportals von Interesse, das die Erkenntnisse der Analysen für interessierte Akteure gut zugänglich darstellt und auch die zeitliche Dimension sowie Einkommensinformationen beinhaltet.

### 6.3.6 Quellen des Steckbriefs

Bundesagentur für Arbeit. (o. D.). BERUFENET. Bundesagentur für Arbeit. Abgerufen am 13. April 2023 von <https://web.arbeitsagentur.de/berufenet/>

Djumaliewa, J. & Sleeman, C. (o. D.). Making sense of skills: A UK skills taxonomy. Nesta. Abgerufen am 13. April 2023 von <https://data-viz.nesta.org.uk/skills-taxonomy/index.html>

Djumaliewa, J. & Sleeman, C. (2018). *An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts* [ESCoE DP 2018-13]. Economic Statistics Centre of Excellence Discussion Papers. Abgerufen am 13. April 2023 von <http://escoe-Webseite.s3.amazonaws.com/wp-content/uploads/2018/08/29113127/ESCoE-DP-2018-13.pdf>

Djumaliewa, J., Lima, A. & Sleeman, C. (2018). *Classifying Occupations According to Their Skill Requirements in Job Advertisements* [ESCoE DP 2018-04]. Economic Statistics Centre of Excellence Discussion Papers. Abgerufen am 13. April 2023 von <https://escoe-website.s3.amazonaws.com/wp-content/uploads/2020/07/13154228/ESCoE-DP-2018-04U.pdf>

Ester, M., Kriegel, H. P., Sander, J. & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Knowledge Discovery and Data Mining, Vol. 96(34)*, 226-231.

Flake, R., Goecke, H., Hickmann, H., Mertens, A. & Seyda, S. (2020). Expertise eines methodischen Ansatzes zur Identifizierung von beruflichen Übergangspfaden in der Automobil- und Zulieferindustrie in Baden-Württemberg. Gutachten im Auftrag der Bertelsmann Stiftung.

Gallagher, E., Kerle, I., Sleeman, C. & Richardson, G. (2022). A New Approach to Building a Skills Taxonomy [ESCoE TR-16]. Economic Statistics Centre of Excellence Technical Reports.

McInnes, L., Healy, J., Saul, N. & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software, 3(29)*, 861. <https://doi.org/10.21105/joss.00861>

Stops, M., Bächmann, A.-C., Glassner, R., Janser, M., Matthes, B., Metzger, L.-J., Müller, C. & Seitz, J. (2021). Extracting Skill Requirements from Job Ads – the “Machbarkeitsstudie Kompetenz-Kompass”. IAB-Forschungsbericht 07/2021, Nürnberg.

Vereinigung der Bayerischen Wirtschaft e. V. (vbw) (2020). *Zukunft der Pflege in Bayern – eine Big Data Analyse der Herausforderungen und Chancen*. Studie von WifOR Institute GmbH. Abgerufen am 13. April 2023 von <https://www.vbw-bayern.de/Redaktion/Frei-zugaengliche-Medien/Abteilungen-GS/Sozialpolitik/2020/Downloads/Studie-Pflegelandschaft-Juni-2020.pdf>

## 6.4 JEDI – Jobs and Education Data Infrastructure (Australien)

### 6.4.1 Einführung

Das Projekt „Jobs and Education Data Infrastructure“ (JEDI) wurde im Jahr 2020 von der australischen National Skills Commission (NSC) initiiert. Ziel des Projekts ist es, detaillierte Informationen zur Arbeitsnachfrage und zu erforderlichen Kompetenzen bereitzustellen (National Skills Commission, 2023). Dazu wurde eine Dateninfrastruktur entwickelt, die Berufen die jeweils erforderlichen Fähigkeiten zuordnet und neue Berufe sowie neu erforderliche Fähigkeiten identifiziert. Die Rohdaten sowie die methodischen Werkzeuge (Programmierung etc.) des Projekts selbst sind nicht für die Öffentlichkeit zugänglich, die erzeugten Informationen finden aber in vielen Unterstützungsangeboten für Arbeitsuchende Anwendung. So wird die Datenbasis z. B. in Job- und Beratungsportalen verwendet, die Arbeitsuchenden Vorschläge machen, in welchen Jobs sie ihre Fähigkeiten ideal einsetzen können oder welche Fortbildungsmaßnahme besonders gut für sie geeignet ist. Begonnen wurde das Projekt mit der Beschreibung von 600 Berufen. Inzwischen sind für mehr als 1.000 Berufe insgesamt über 2.000 Anforderungen beschrieben, die allgemeine Kompetenzen, spezifische ausgeübte Tätigkeiten sowie die Fähigkeit zum Umgang mit bestimmten Hilfsmitteln umfassen.

Diese australische Dateninnovation weist Überschneidungen mit entsprechenden Produkten und Angeboten in Deutschland auf, insbesondere der Klassifikation der Berufe (KldB) 2010, der Datenbank BERUFENET und dem in der Entwicklung befindlichen Kompetenz-Kompass des IAB (vgl. Stops et al., 2021). Im Vergleich dazu ist JEDI derzeit umfassender, identifiziert systematisch neu erforderliche Fähigkeiten und Werkzeuge und wird systematisch in sich anschließende Datenprodukte eingespeist.<sup>65</sup> Die Übertragbarkeit auf Deutschland erscheint technisch machbar. Allerdings ist der erforderliche personelle Aufwand bei der Umsetzung eines solchen Projekts relativ hoch.

### 6.4.2 Technische Beschreibung

Für JEDI wurde durch die NSC eine eigene Australian Skills Classification (ASC) geschaffen, die die Anforderungsprofile von Berufen ermittelt und Berufe entsprechend klassifiziert. Ziel der Klassifikation ist es, eine einheitliche Taxonomie zu entwickeln, die Entwicklungen bei Anforderungsprofilen von Berufen und die Übertragbarkeit von Fähigkeiten zwischen Berufen abbildet.

Als Grundlage für die Klassifikation wurden eine Datenbank mit Berufsbeschreibungen aus den USA, O\*NET (siehe auch Steckbrief 4.7), und Vorgängerversionen der australischen Klassifikation genutzt. Diese Datenbasis wurde insbesondere durch Daten aus Stellenanzeigen (mithilfe von Machine-Learning-Methoden), aber auch aus Firmenumfragen und Weiterbildungsberichten, überprüft und ergänzt – vor allem, um neu entstehende Berufe zu identifizieren und in die Klassifikation aufzunehmen. Begonnen wurde mit 600 Berufen auf 4- und 6-Stellerebene der Berufsklassifizierung Australian and New Zealand Standard Classification of Occupations (ANZSCO). Die ANZSCO enthält seit dem Jahr 2022 364 Berufe auf der 4-Steller- und 1.417 Berufe auf der 6-Steller-Ebene (Australian Bureau of Statistics, 2022). Seit der ersten Entwicklung der ASC wird diese kontinuierlich weiterentwickelt und durch weitere Berufe ergänzt, wobei auch vorher nicht klassifizierte Berufe auftauchen, die nach einem neuen System als sogenannte „nowhere else classified“ (nec) Berufe aufgeführt werden. Die entstehenden Anforderungsprofile sind in drei Kategorien eingeteilt: Core

---

<sup>65</sup> Kapitel 6.3 beschreibt eine vergleichbare Dateninnovation, die Textdaten aus Stellenanzeigen nutzt.

Competencies (Kernkompetenzen), Specialist Tasks (spezialisierte Tätigkeiten) und Technology Tools (technologiebasierte Arbeitsmittel) (Jobs and Skills Australia, 2022).

In der ersten Kategorie sind als Core Competencies zehn Kompetenzen beschrieben, die in unterschiedlichem Ausmaß für die überwiegende Mehrheit an Berufen benötigt werden. Zu den Core Competencies gehören IT-Kompetenzen, Eigeninitiative, Lernen, mündliche Kommunikation, Planen und Organisieren, Problemlösung, Rechnen und Lesen. Auf einer Zehn-Punkte-Skala wird bewertet, wie wichtig die jeweilige Kompetenz für einen Beruf ist. Da diese Kompetenzen in fast allen Berufen erforderlich sind, sind sie besonders einfach übertragbar. Die zweite Kategorie, Specialist Tasks, umfasst berufsspezifische Tätigkeiten, die jeweils in weniger Berufen ausgeübt werden als Core Competencies. Somit sind Specialist Tasks auch weniger leicht übertragbar. Sie verändern sich außerdem schneller, zum Beispiel durch technologischen Wandel, und sind darum besonders relevant für Veränderungen der Anforderungsprofile. Beispielsweise wurden kürzlich die Specialist Tasks „Entwicklung von Inhalten für soziale Medien“ sowie „Anwendung von Techniken maschinellen Lernens auf Daten“ zur ASC hinzugefügt (NSC, 2022). Weitere Beispiele sind „Unterrichtsmaterialien entwickeln“, „Computerprogramme installieren“, „Landkarten erstellen“, „Essen und Trinken servieren“, „Schminke auftragen“, „Medizinische Notfälle behandeln“, „Proben und Vorstellungen planen“ und „Pakete zur Auslieferung verpacken“.

Um die Übertragbarkeit der Specialist Tasks und der damit verbundenen Fähigkeiten abzubilden, wurden sogenannte Skills Cluster entwickelt. Ein Cluster beinhaltet jeweils ähnliche, komplementäre Tätigkeiten und Fähigkeiten. Die Übertragbarkeit von Fähigkeiten zwischen den Berufen innerhalb eines Clusters ist somit relativ hoch. Beispiele für Skills Cluster innerhalb der Oberkategorie der Skills-Cluster-Familie „Gesundheit und Pflege“ sind „Medizinische Informationen und Prozeduren erklären“, „Krankheiten diagnostizieren und Behandlungen verschreiben“ und „Brillen und Sehhilfen anpassen“.

In der dritten Kategorie Technology Tools werden alle technologiebasierten Arbeitsmittel wie Soft- und Hardware abgebildet, die für den jeweiligen Beruf benötigt werden. Beispiele hierfür sind Instant Messenger, Social-Media-Plattformen, Tele- oder Videokonferenzprogramme, Navigationssysteme, Programmiersprachen, Buchhaltungsprogramme, Datenanalyse-Software, wissenschaftliche Instrumente und medizinische Geräte. Die Übertragbarkeit der zugehörigen Fähigkeiten unterscheidet sich damit stark zwischen den jeweiligen Arbeitsmitteln.

In der aktuellsten Ausgabe von November 2022 sind in der ASC 1.347 Berufe enthalten, die nach insgesamt 10 Kernkompetenzen, 2.136 Specialist Tasks und 70 Technology Tools unterschieden werden können. Dabei werden jedem Beruf im Schnitt 17,8 Specialist Tasks und 4,1 Technology Tools zugeordnet. Für viele Berufe ist zusätzlich angegeben, wieviel Prozent der Arbeitszeit einer bestimmten Specialist Task gewidmet wird. Außerdem wird dargestellt, welche Specialist Tasks oder Technology Tools innerhalb eines Berufs wichtiger werden („trending“) oder neu hinzukommen („emerging“) (Jobs and Skills Australia, 2022).

### 6.4.3 Anwendung

Die entstandenen Daten können für viele Projekte und Maßnahmen als Informationsquelle dienen. So hat die australische Regierung zum Beispiel mithilfe des JEDI-Projekts besonders gefragte Fähigkeiten identifiziert, um gezielt Weiterbildungskurse zu diesen Fähigkeiten zu fördern.<sup>66</sup> Im August 2020 veröffentlichte die NSC eine Liste mit den 25 am stärksten nachgefragten neuen Berufen, sogenannten Emerging Occupations. Die Liste wurde mithilfe von JEDI erstellt und soll dazu

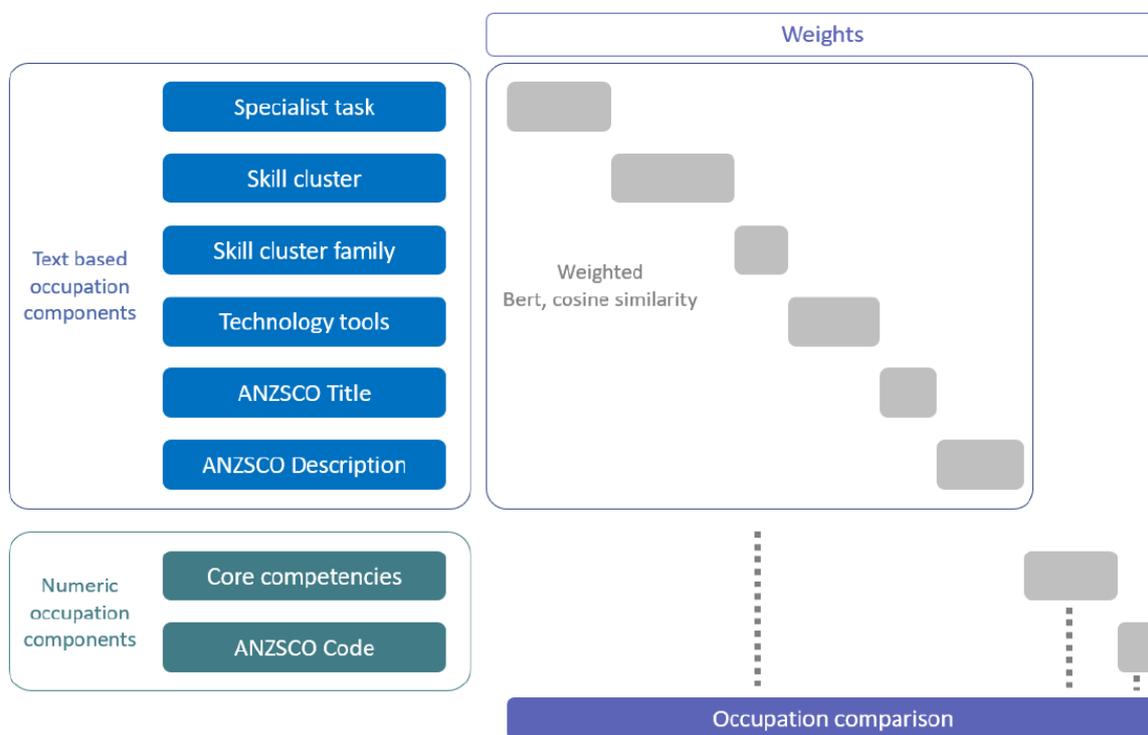
<sup>66</sup> Die Förderung wurde im Rahmen des JobTrainer Fund durchgeführt (Department of Employment and Workplace Relations, o.D.).

dienen, Veränderungen auf dem Arbeitsmarkt zu erfassen und notwendige Anpassungen von Aus- und Weiterbildung bzw. beruflichen Neuorientierungen zu fördern (National Skills Commission, 2020).

Zusätzlich finden die Daten auch in mehreren Online-Tools Verwendung, die sich an Forschung und Öffentlichkeit, insbesondere an Berufseinsteiger\*innen und Berufswechsler\*innen richten.

Hierzu zählt das Online-Tool „Similar Occupations“, das ebenfalls mithilfe der ASC geschaffen wurde. Es zeigt zu ausgewählten Berufen eine Liste mit bis zu zehn ähnlichen Berufen und den jeweiligen Grad der Übereinstimmung an (Jobs and Skills Australia, 2023). Der Vergleich zwischen Berufen erfolgt mithilfe von Natural Language Processing, d.h. es findet ein automatisierter Vergleich der in der JEDI-Datenbank in Textform enthaltenen Anforderungsprofile statt. Genutzt wird hierbei das von Google zur Verfügung gestellte Tool „Bidirectional Encoder Representations from Transformers“ (BERT). Hierbei gehen die in JEDI auf Berufsebene identifizierten Kompetenzen, Tätigkeiten und Arbeitsmittel ein, zusätzlich die Berufsbezeichnungen und Berufsbeschreibungen aus der ANZSCO. Dabei werden zunächst die Informationen aus den Texten in einen numerischen Vektor überführt, der dann zur Bestimmung der Ähnlichkeit von Berufen herangezogen wird. Die Vorgehensweise wird in Abbildung 6-8 dargestellt.

**Abbildung 6-8: Berechnung der Ähnlichkeit von Berufen**



Quelle: National Skills Commission (2022), Abb. 6.

Auch die Plattform Job Switch, die vom australischen Department of Employment and Workplace Relations betrieben wird, beruht auf JEDI-Daten, ist jedoch umfassender als Similar Occupations und richtet sich explizit an Berufswechsler\*innen. Job Switch verbindet die Informationen zu ähnlichen Berufen und zur Übertragbarkeit von Fähigkeiten zwischen diesen Berufen mit Daten zur Nachfrage nach bestimmten Berufen. Arbeitsuchende Personen können ihren bisher ausgeübten Beruf angeben und bekommen mögliche neue Berufe, die entsprechend der aktuellen Nachfrage auf dem Arbeitsmarkt geordnet werden, angezeigt. Außerdem sind mit diesen Berufen beim Department of

Employment and Workplace Relations gemeldete offene Stellen direkt verknüpft, sodass neben den Informations- auch die Suchkosten für neue Stellen stark reduziert werden. Die verknüpfte Stellensuche kann nach Bundesstaat gefiltert werden. Zu den empfohlenen Berufen wird zudem eine Kurzbeschreibung mit Informationen zu nötigen formalen Qualifikationen gegeben (Workforce Australia, o. D.). Eine beispielhafte Anwendung auf den Beruf „Dieselmotor-Mechaniker\*in“ wird in Abbildung 6-9 dargestellt. Für diesen Beruf werden zehn alternative Berufe vorgeschlagen. Die Aufgaben in diesen Berufen sind dem ursprünglichen Beruf teilweise ähnlich, wie z. B. Motor-Mechaniker\*in, teilweise werden aber auch etwas weiter entfernte Berufe vorgeschlagen, wie z. B. Lastwagenfahrer\*in oder Flugzeugwartungsingenieur\*in (siehe Abbildung 6-9). Die weiteren beruflichen Alternativen im Beispielfall sind Mechaniker\*in für Kleinmotoren (z. B. Kettensägen, Rasenmäher), Motorradmechaniker\*in, Automobilelektriker\*in, Karosserieschlosser\*in, Monteur\*in von KFZ-Teilen und Zubehör, Bau und Wartung von Spezialfahrzeugen sowie Fahrzeuglackierer\*in.

**Abbildung 6-9: Job Switch Portal, beispielhafte Anwendung**

**Job title or occupation you've done before\***

Begin typing a job title and select from the list

**Location**

I would like to search in

Queenland
▼

Find Job Switch options

We found 10 job types that use the same skills as Diesel Motor Mechanic.

Sort by Demand (High to Low) ▼

### Motor Mechanic (General)

📈 In demand

🎯 Very similar to Diesel Motor Mechanic

👤 You may not need much training

Maintains, tests and repairs petrol engines and the mechanical parts of lightweight motor vehicles such as transmissions, suspension, steering and brakes. Registration or licensing may be required.

**Study options**

📌 You don't need a qualification but it might improve your chances!

- | [Certificate III in Automotive Underbody Technology](#)
- | [Certificate III in Light Vehicle Mechanical Technology](#)

View jobs

☆ Interested ?

### Truck Drivers

📈 In demand

🎯 Very similar to Diesel Motor Mechanic

👤 You may not need much training

Truck Drivers drive heavy trucks, removal vans, tankers and tow trucks to transport bulky goods and liquids.

**Study options**

📌 You don't need a qualification but it might improve your chances!

- | [Certificate IV in Driving Operations](#)
- | [Certificate II in Driving Operations](#)

View 2576 jobs

☆ Interested ?

## Aircraft Maintenance Engineer (Mechanical)

Very similar to Diesel Motor Mechanic

You may not need much training

View jobs

Inspects, tests, repairs and installs aircraft hydromechanical and flight system components and aircraft engines, subassemblies and components. Registration or licensing may be required.

☆ Interested ?

### Study options

You don't need a qualification but it might improve your chances!

Diploma of Aeroskills (Mechanical)

Certificate IV in Aeroskills (Mechanical)

Quelle: Workforce Australia (o. D.)

Das Portal „Your Career“ des National Careers Institute verwendet ebenfalls JEDI-Daten und stellt ähnliche Services wie Job Switch bereit. Your Career richtet sich im Gegensatz zu Job Switch jedoch nicht an Berufswechler\*innen, sondern an Berufseinsteiger\*innen. Basierend auf Interessen, Qualifikationen und Arbeitserfahrung können auch hier Informationen über passende Berufe, ausgeschriebene Stellen und nötige Qualifikationen gefunden werden. Zudem werden mögliche Weiterbildungskurse direkt aufgeführt (National Careers Institute, o. D.).

Der Kern der Innovation, die Taxonomie, ist vergleichbar mit der Data-Driven Skills Taxonomy aus dem Vereinigten Königreich (Steckbrief 6.3). Allerdings beruhen die Taxonomien auf unterschiedlichen Datentypen. Zudem ist JEDI im Gegensatz zur Data-Driven Skills Taxonomy nicht öffentlich zugänglich, dafür ist JEDI im Hinblick auf die Einbeziehung des Weiterbildungsangebots breiter angelegt.

### 6.4.4 Übertragbarkeit auf Deutschland

In Deutschland werden bereits einige Datenprodukte bereitgestellt, die JEDI teilweise ähnlich sind. Dies macht JEDI zu einem interessanten Vergleichspunkt für Deutschland und zeigt zudem, dass eine Übertragbarkeit besonders innovativer Elemente von JEDI möglich sein sollte. Zum einen liefern BERUFENET und die damit in Zusammenhang stehenden Arbeiten an der KldB2010 Erkenntnisse über neu entstehende Berufe, zudem werden über BERUFENET Informationen zu Ähnlichkeiten zwischen Berufen hinsichtlich deren Anforderungsprofilen sowie empfehlenswerten Weiterbildungen zur Verfügung gestellt. Mit dem IAB-Kompetenz-Kompass (Stops et al., 2021) befindet sich zudem ein Datenprodukt in Entwicklung, das hinsichtlich der Verwendung von Stellenanzeigen Parallelen aufweist.

Im Vergleich zu diesen deutschen Datenprodukten umfasst JEDI zusätzliche Elemente. Erstens werden neu entstehende Berufe, neu geforderte Kompetenzen und technische Werkzeuge systematisch erfasst und dargestellt. Hierdurch wird die Dynamik des Arbeitsmarktes hinsichtlich neuer Erfordernisse, wie sie beispielsweise durch den digitalen und ökologischen Wandel entstehen, sichtbar gemacht. Dies geht teilweise über die Arbeiten an der KldB2010 hinaus, die ebenfalls neue Berufe systematisch identifiziert, nicht jedoch neu geforderte Kompetenzen und technische Werkzeuge. Zweitens wird die Bedeutung von Kompetenzen in Berufen klar quantifiziert. Drittens wird die Datenbasis von JEDI systematisch in daran anschließende Daten- und Beratungsprodukte eingespeist, z. B. bei der Berechnung der Ähnlichkeit von Berufen und der Plattform Job Switch. BERUFENET liefert ebenfalls Informationen zu Berufen, allerdings erscheint z. B. der Ähnlichkeitsvergleich von Berufen weniger stark ausgeprägt.

Hinsichtlich der Datenverfügbarkeit ist die Situation bei den Stellenanzeigen in Deutschland vergleichbar, wenn nicht sogar besser, da neben kommerziellen Angeboten von Online-Stellenanzeigen durch z. B. Lightcast und TextKernel<sup>67</sup> auch die Stellenanzeigen der BA genutzt werden können. Diese werden bereits im IAB-Kompetenz-Kompass, aber auch in anderen Forschungsprojekten (Burstedde et al., 2020) verwendet. Diese Verwendungsbeispiele zeigen zudem, dass nicht nur die Datenverfügbarkeit gewährleistet ist, sondern dass auch der Datenschutz bei der Verwendung von Stellenanzeigen keine größere Hürde darstellt, insbesondere sofern keine personenbezogenen Daten verarbeitet werden. Auch die zweite benötigte Datenquelle, eine ausführliche textliche Beschreibung von Berufen, steht in Deutschland mit BERUFENET (Bundesagentur für Arbeit, o. D.) zur Verfügung. Diese wird zwar regelmäßig aktualisiert, die Darstellung neu entstehender Berufe erscheint im australischen Beispiel jedoch umfassender, systematischer und transparenter zu erfolgen.

Bei der praktischen Umsetzung sollte beachtet werden, dass ein entsprechendes Vorhaben mit einem erheblichen personellen Aufwand im datenwissenschaftlichen Bereich verbunden ist, was sowohl die Anzahl des notwendigen Personals als auch deren erforderliches Wissen angeht. Eine solche umfangreiche Dateninnovation ist besonders dann sinnvoll, wenn die Nutzung der entstehenden Daten in verschiedenen angeschlossenen Datenprojekten bereits in der Projektgenese mitgedacht wird.

#### 6.4.5 Fazit

Das australische Projekt “Jobs and Education Data Infrastructure” (JEDI) hat zum Ziel, für möglichst viele Berufe die jeweils notwendigen Kompetenzen, spezifische ausgeübte Tätigkeiten sowie die verwendeten technischen Arbeitsmittel zu identifizieren. Dabei werden systematisch neu entstehende Berufe sowie Veränderungen in den Anforderungsprofilen von Berufen erfasst. Die generierten Erkenntnisse fließen zudem in direkt angeschlossene weitere Datenprodukte ein, wie z. B. die Analyse der Ähnlichkeit von Berufen oder die Informationsplattform Job Switch. Insgesamt entstehen somit höchst relevante Erkenntnisse über Transformationsprozesse auf dem Arbeitsmarkt, die vor allem Arbeitnehmer\*innen sowie der Wirtschafts- und Bildungspolitik zugutekommen.

In Deutschland bestehen insbesondere mit dem Kompetenz-Kompass am IAB sowie BERUFENET und der Kldb 2010 Datenprojekte, die Überschneidungen mit JEDI aufweisen. Im Vergleich dazu ist JEDI umfassender, stellt systematischer neue Berufe und neu erforderliche Fähigkeiten und Werkzeuge dar, und wird systematischer in sich anschließende Datenprodukte eingespeist.

#### 6.4.6 Quellen zum Steckbrief

Australian Bureau of Statistics. (2022). *ANZSCO - Structure*. Australian Bureau of Statistics. Abgerufen am 03. April 2023 von <https://www.abs.gov.au/statistics/classifications/anzsco-australian-and-new-zealand-standard-classification-occupations/2022/anzsco%202022%20structure.xlsx>

Büchel, J., Engler, J. F., Mertens, A. (2023), Gesuchte Datenkompetenzen in Deutschland. IW-Trends 2/2023. Institut der deutschen Wirtschaft, Köln.

Bundesagentur für Arbeit. (o. D.). *BERUFENET*. Bundesagentur für Arbeit. Abgerufen am 13. April 2023 von <https://web.arbeitsagentur.de/berufenet/>

---

<sup>67</sup> Die Daten von TextKernel werden z. B. in Büchel et al. (2023) genutzt.

Burstedde, A., Flake, R., Jansen, A., Malin, L., Risius, P., Seyda, S., Schirner, S., & Werner, D. (2020). *Die Messung des Fachkräftemangels* (IW Report 59/2020). Institut der deutschen Wirtschaft. <https://www.iwkoeln.de/studien/alexander-burstedde-regina-flake-anika-jansen-lydia-malin-paula-risius-susanne-seyda-sebastian-schirner-dirk-werner-die-messung-des-fachkraeftemangels.html>

Department of Employment and Workplace Relations. (o. D.). *Job Trainer Fund*. Australian Government Department of Employment and Workplace Relations. Abgerufen am 13. April 2023 von <https://www.dewr.gov.au/skills-reform/jobtrainer-fund>

Jobs and Skills Australia. (2022). *Australian Skills Classification*. Jobs and Skills Australia. Abgerufen am 03. April 2023 von <https://www.jobsandskills.gov.au/data/ASC/release-2022.09/Australian%20Skills%20Classification%20-%20November%202022.xlsx>

Jobs and Skills Australia. (2023). *Australian Skills Classification – Similar Occupations*. Jobs and Skills Australia. Abgerufen am 31. März 2023 von <https://www.jobsandskills.gov.au/australian-skills-classification#similarity>

National Careers Institute. (o. D.). *Your Career*. Australian Government Your Career. Abgerufen am 23. März 2023 von <https://www.yourcareer.gov.au/>

National Skills Commission (2020). *Emerging occupations: How new skills are changing Australian jobs*. [https://www.nationalskillscommission.gov.au/sites/default/files/2021-12/Emerging%20occupations%20How%20new%20skills%20are%20changing%20Australian%20jobs\\_0.pdf](https://www.nationalskillscommission.gov.au/sites/default/files/2021-12/Emerging%20occupations%20How%20new%20skills%20are%20changing%20Australian%20jobs_0.pdf)

National Skills Commission. (2022). *The Australian Skills Classification release 2.1*. Australian Government National Skills Commission. Abgerufen am 31. März 2023 von <https://www.nationalskillscommission.gov.au/data/ASC/release-2022.09/Australian%20Skills%20Classification%202.1%20Report.pdf>

National Skills Commission (2023). *Jobs and Education Data Infrastructure (JEDI)*. Australian Government National Skills Commission. Abgerufen am 31. März 2023, vom <https://www.nationalskillscommission.gov.au/topics/jedi>

Stops, M., Bächmann, A.-C., Glassner, R., Janser, M., Matthes, B., Metzger, L.-J., Müller, C. & Seitz, J. (2021). *Extracting Skill Requirements from Job Ads – the “Machbarkeitsstudie Kompetenz-Kompass”*. IAB-Forschungsbericht 07/2021, Nürnberg.

Workforce Australia. (o. D.). *Job Switch*. Australian Government Workforce Australia. Abgerufen am 23. März 2023 von <https://www.workforceaustralia.gov.au/individuals/coaching/careers/job-switch/search>

## 6.5 Job- und Informationsportal auf der Basis von Open Government Data (Spanien)

### 6.5.1 Einführung

Die spanische Webseite [discapacidad.es](http://discapacidad.es) veröffentlicht tagesaktuell Stellenangebote und Informationen zu finanziellen und personellen Unterstützungsmaßnahmen beim Übergang in eine Beschäftigung. Die dargestellten Stellen- und Unterstützungsangebote werden direkt und automatisiert aus Open Government Data (OGD) der Verwaltung der Region Kastilien und León gezogen. OGD sind Rohdaten, die durch Bürger\*innen nicht ohne weitere Aufbereitung genutzt werden können. Daher sind die zielgruppenspezifische Aufbereitung der Daten und ihre Darstellung auf der Webseite die entscheidenden Elemente der Innovation. Die Zielgruppe des Portals sind Menschen mit Behinderung, aber das Innovationspotenzial lässt sich auch für andere Zielgruppen realisieren.

Die Bereitstellung der Daten ist das Ergebnis einer Initiative der Region Kastilien und León zur Stärkung der politischen Teilhabe und gesellschaftlicher Initiative. Schon im Jahr 2012 wurde ein frei zugängliches Datenportal entwickelt. Dort werden seitdem verschiedene Datensätze maschinenlesbar und kostenfrei zur Verfügung gestellt, beispielsweise zur detaillierten Entwicklung der Anzahl der Betriebe, der Arbeitsuchenden oder der Beschäftigten in verschiedenen Tätigkeiten. Neben Arbeitsmarktstatistiken werden auch Daten zu vielen weiteren Themen, beispielsweise der regionalen Betroffenheit durch die Covid-19-Pandemie oder Waldbrandrisikogebieten bereitgestellt (Junta de Castilla y León, o. D. -a).<sup>68</sup> Im Themenbereich Arbeitsmarkt enthält das Datenportal auch Stellenanzeigen von Unternehmen und Stellenanzeigen des öffentlichen Dienstes sowie einen Datensatz zu Fördermöglichkeiten und Programmen für Menschen mit Behinderung, die auf der Webseite [discapacidad.es](http://discapacidad.es) aufbereitet dargestellt werden. Diese Darstellung wird im Folgenden als Beispiel für die Nutzung von OGD vertieft.

In Deutschland gibt es bereits eine Vielzahl von OGD-Angeboten durch die öffentliche Verwaltung oder öffentliche Unternehmen, die über das Portal [www.govdata.de](http://www.govdata.de) zugänglich sind. Allerdings ist die Nutzung der Daten durch Unternehmen, die Zivilgesellschaft oder einzelne Bürger\*innen im Vergleich zu den Nutzungsmöglichkeiten in Deutschland und der tatsächlichen Nutzung in anderen Ländern noch nicht weit fortgeschritten (Boockmann et al. 2020).

Die Webseite [discapacidad.es](http://discapacidad.es) verdeutlicht, welches Potenzial in granularen OGD-Angeboten steckt. Aufgrund des Bedarfs an zielgruppenspezifischer Information für die Vermittlung und Förderung am Arbeitsmarkt könnte diese Anwendung auch für Deutschland Vorbildcharakter haben. Der Aufbau eines Stellenportals für Menschen mit Behinderung in Deutschland ist allerdings nicht sinnvoll, da hierzulande Stellen nicht spezifisch für Menschen mit Behinderung ausgeschrieben werden.

### 6.5.2 Technische Beschreibung

Auf der Webseite der Verwaltung der Region Kastilien und León können Datensätze zu unterschiedlichen Themen kostenfrei heruntergeladen werden. Die Daten beziehen sich vorwiegend auf wirtschaftliche und geografische Merkmale innerhalb der Region und liegen in maschinenlesbarer Form vor, sodass sie beispielsweise unter geringem Aufwand in Web-Scraping-Methoden eingebunden werden können. Die Dateninitiative der Region Kastilien und León listet über 50

---

<sup>68</sup> Ferner können Bürger\*innen und Unternehmen über ein Anfrageformular neuartige Datensätze anfordern, die bei Genehmigung produziert und herausgegeben werden.

Projekte auf, die allein auf dieser regionalen OGD-Schnittstelle basieren (Junta de Castilla y León, o. D. -b). Unter anderem kann von der Webseite ein frei zugänglicher und tagesaktueller Datensatz zu Stellenangeboten im öffentlichen Dienst und der Privatwirtschaft in der Region abgerufen werden, welcher in die Webseite [discapacidad.es](http://discapacidad.es) integriert wird (Junta de Castilla y León, o. D. -c). Der Datenabruf für [discapacidad.es](http://discapacidad.es) geschieht mithilfe eines speziellen Web-Scraping-Programms namens RSS-Ground und ist aufgrund der Formate der zur Verfügung gestellten Daten technisch sehr einfach zu implementieren. RSS-Ground erlaubt die Verknüpfung und automatische Aktualisierung zwischen einem im Internet frei verfügbarem Ursprungselement und einer individuellen Anwendung. Beispielsweise kann damit auf einem Blog automatisch die neuste Nachricht eines Nachrichtenportals eingebettet werden, wenn das Portal ein RSS-Feed anbietet. Für die Anwendung bei [discapacidad.es](http://discapacidad.es) werden die Stellenangebote der Webseite von Kastilien und León tagesaktuell automatisch eingebunden. Nach eigenen Angaben existieren weltweit über RSSGround circa 5,3 Mio. dieser Verknüpfungen von über 80.000 Nutzenden (RSSGround, o. D.). Die Dienstleistung wird von einem privaten Unternehmen angeboten und ist teilweise kostenpflichtig.

Auf der Webseite [discapacidad.es](http://discapacidad.es) werden zusätzlich zur Verlinkung noch zentrale Informationen wie der Beginn der Tätigkeit, die erforderlichen Qualifikationen, die Region und ein Link zur tatsächlichen Stellenausschreibung automatisch extrahiert und separat dargestellt. Über die Stellenanzeigen hinaus werden auf [discapacidad.es](http://discapacidad.es) auch weitere Informationen zu staatlichen Fördermaßnahmen für Menschen mit Behinderung bereitgestellt, z. B. Beihilfen für die Anschaffung von digitalen Geräten oder Fördermaßnahmen zur Schaffung von Arbeitsplätzen. Diese richten sich sowohl an Menschen mit Behinderung als auch an Arbeitgeber\*innen, die Beschäftigte mit einer Behinderung eingestellt haben oder planen einzustellen. Von den auf [discapacidad.es](http://discapacidad.es) dargestellten Informationen führt ein Weiterleitungslink direkt auf die zugeordnete Ursprungswebseite, welcher bereits im Originaldatensatz der Dateninitiative der Region Kastilien und León enthalten ist. Menschen mit einer Behinderung bzw. Arbeitgeber\*innen können somit das breite Angebot an Fördermöglichkeiten zunächst via [discapacidad.es](http://discapacidad.es) durchsuchen und dann über die Verlinkung Details der sie interessierenden Angebote gezielt nachlesen.

Die Webseite [discapacidad.es](http://discapacidad.es) wurde von einem Experten im Bereich Sozialhilfe für Menschen mit Behinderung entwickelt und wird fortführend von ihm unterhalten und weiterentwickelt. Die Projektdurchführung erfolgt privat und unabhängig von der Verwaltung der Region Kastilien und León. Zu Beginn der Implementierung gab es zwischen der Verwaltung der Region und dem Projektleiter allerdings regelmäßig Austausch, insbesondere um die Daten korrekt in die Webseite einzubinden. Inzwischen findet nur noch selten ein Austausch statt, da die Einbettung der Stellenanzeigen auf [discapacidad.es](http://discapacidad.es) automatisch geschieht und es selten Bedarf für manuelle Korrekturen gibt.

### 6.5.3 Anwendung

Aktuell beziehen sich die verfügbaren Angebote nur auf die Provinzen der Region Kastilien und León. Die nachgelagerten Webseiten mit der Auflistung der Stellenanzeigen in der betreffenden Provinz enthalten jeweils den Titel der Stelle und das Datum der Stellenanzeige. Darunter stehen einige Zeilen aus der Stellenanzeige. Von dort aus gelangt man mit einem Klick zur vollständigen Anzeige. In einem anderen Menü gelangt man zu einer Auflistung von Angeboten und Maßnahmen für Menschen mit Behinderung. Die Webseite ist einfach strukturiert. Hervorzuheben sind das zielgruppenspezifische Datenangebot von Stellenanzeigen und Maßnahmen sowie die einfache Navigationsführung.

Nach Auskunft des Projektleiters wird die Webseite von circa 500 bis 1.000 Personen im Monat aufgerufen. Die Nutzenden sind meist Privatpersonen, aber auch Nichtregierungsorganisation, die

mit Menschen mit Behinderung zusammenarbeiten. Das Projekt [discapacidad.es](http://discapacidad.es) arbeitet nicht gewinnorientiert und wird privat durch persönliche Mittel des Projektleiters finanziert. Dies zeigt, dass die Nutzung von OGD auch mit geringem Ressourceneinsatz möglich ist.

Die Webseite soll künftig um drei Elemente erweitert werden. Zunächst soll die Webseite alle Regionen Spaniens abdecken. Des Weiteren soll eine Hotline eingerichtet werden, welche die Informationen der Webseite auch für Menschen mit starker Sehbeeinträchtigung zugänglich macht. Drittens soll die Zusammenarbeit mit Sozialarbeiter\*innen forciert werden, um sowohl die Reichweite der Webseite als auch die Informationsweitergabe an Multiplikator\*innen zu verbessern.

#### 6.5.4 Übertragbarkeit auf Deutschland

Das zentrale Innovationselement von [discapacidad.es](http://discapacidad.es) ist die automatisierte Verknüpfung und Aufbereitung von zwei von öffentlicher Hand zur Verfügung gestellten Datensätzen. In Deutschland werden OGD auf der Webseite [govdata.de](http://govdata.de) zur Verfügung gestellt. Diese Webseite verfolgt im Sinne verschiedener Bundes- und Landesgesetze das Ziel, den Zugang zu Verwaltungsdaten aus Bund, Ländern und Kommunen sowie zu Daten von Unternehmen der Daseinsvorsorge, Hochschulen, Forschenden, Forschungseinrichtungen und Forschungsfördereinrichtungen (FITKO, o. D.) zu erleichtern. Die Webseite erlaubt den Zugriff auf über 80.000 Datensätze, es sind jedoch nur 24 Anwendungen aufgeführt, die auf diesen Daten basieren.

Dass dem relativ großen Datenangebot nur eine geringe Nutzung durch Bürger\*innen und Zivilgesellschaft gegenübersteht, ist auch das Ergebnis einer Studie im Auftrag der Hans-Böckler-Stiftung (Boockmann et al., 2020). Auch kommerzielle Geschäftsmodelle auf Basis von OGD sind danach nicht vorhanden. Die Nutzung von OGD vollzieht sich fast ausschließlich innerhalb und zwischen Verwaltungen.

Die häufigsten OGD-Angebote kommen aus dem Bereichen Geografie, Geologie und Geobasisdaten. Danach folgen Strukturdaten der Kommunen (z. B. Haushalt, Politik und Wahlen, Verkehr, Umwelt, Infrastruktur). Etwa ein Fünftel der befragten Kommunen stellt OGD aus dem Bereich Soziales zur Verfügung (Boockmann et al., 2020). Stellenanzeigen werden über [govdata.de](http://govdata.de) nur aus der öffentlichen Verwaltung Nordrhein-Westfalens als OGD zur Verfügung gestellt. Für einen direkten Transfer ähnlich des Beispiels [discapacidad.de](http://discapacidad.de) – ggf. mit einem anderen Fokus als Menschen mit Behinderung – ist die Datengrundlage damit noch nicht ausreichend.

Durch eine Ausdehnung des OGD-Angebots könnten sich möglicherweise neue Nutzungen etablieren. Dies allein dürfte jedoch kaum ausreichend sein, um die Anzahl der OGD-Nutzungen – auch im Bereich Arbeit und Soziales – spürbar zu erhöhen. Die Komplexität der Daten, ihre mangelnde Standardisierung und fehlende Verknüpfungsfähigkeit machen Nutzungen derzeit unattraktiv (Boockmann et al., 2022). Zur Nutzung von OGD in vielen Bereichen sind spezielle Fachkenntnisse erforderlich. Die Autoren plädieren daher dafür, intermediäre Organisationen mit der Aufbereitung und Qualitätssicherung von OGD zu beauftragen.

#### 6.5.5 Fazit

Die Webseite [discapacidad.es](http://discapacidad.es) ist ein Stellen- und Informationsportal für Menschen mit Behinderung. Die dargestellten Informationen werden aus einer OGD-Schnittstelle der spanischen Region Kastilien und León gezogen und automatisiert aufbereitet. Aufgrund dieser Automatisierung sind die Kosten für den Unterhalt der Webseite gering. Die Möglichkeit der automatisierten Einbettung von OGD ist ein zentrales und notwendiges Element, um gesellschaftliche Datenprojekte aus OGD zu fördern. In Deutschland gibt es bereits ein breites Angebot an OGD, allerdings wird dieses noch wenig genutzt. Dies ist auf die komplexen und teilweise uneinheitlichen Datenstrukturen und -merkmale

zurückzuführen. Um die Nutzung und damit auch den gesellschaftlichen Mehrwert zu steigern, ist es von Interesse, die Komplexität und Heterogenität der zur Verfügung gestellten Daten zu reduzieren.

#### 6.5.6 Quellen zum Steckbrief

Boockmann, B., Hochmuth, U., Klee, G., Mangold, M., & Scheu, T. (2020). Open Government Data: Ziele, Umsetzung und mögliche künftige Verlaufsformen (No. 442). Study der Hans-Böckler-Stiftung.

Boockmann, B., Hochmuth, U., Mangold, M., & Scheu, T. (2022). Intermediäre Strukturen zur Qualitätssicherung von Open Government Data. *Die Öffentliche Verwaltung*, 75(11), 463-467.

FITKO (o. D.) GOVDATA - Das Datenportal für Deutschland. Abgerufen am 24.05.2023 von <https://www.govdata.de/web/guest/hilfe>

Junta de Castilla y León. (o. D.-a). Datos Abiertos de Castilla y León [Castilla y León Open Data]. Junta de Castilla y León. Abgerufen am 5. April 2023 von <https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html>

Junta de Castilla y León. (o. D.-b). Datos Abiertos de Castilla y León [Castilla y León Open Data]. Junta de Castilla y León. Abgerufen am 24. Mai 2023 von <https://datosabiertos.jcyl.es/web/es/aplicaciones.html>

Junta de Castilla y León. (o. D.-c). Convocatorias de Empleo Público [Aufrufe zur öffentlichen Beschäftigung]. Abgerufen am 24. März 2023 von <https://datosabiertos.jcyl.es/web/jcyl/set/es/sector-publico/convocatoriaspublicas/1284165772128>

RSSGround. (o. D.). *Streamline Your Content Discovery & Sharing*. RSSGround. Abgerufen am 13. April 2023 von <https://www.rssground.com/>

## 6.6 Nowcast von Beschäftigung und Daten-Dashboard (Australien)

### 6.6.1 Einführung

Um schnell und angemessen auf aktuelle Entwicklungen wie die Covid-19-Pandemie reagieren zu können, sind die Arbeitsmarktpolitik und die Akteure am Arbeitsmarkt auf aktuelle und genaue Informationen angewiesen. Echtzeitdaten über die Beschäftigung in einzelnen Regionen und Berufen, die öffentlich einsehbar und verständlich aufbereitet sind, sind darum von hohem Wert. Mit dem Projekt „Nowcast of Employment by Region and Occupation“ (NERO) ist diese Idee in Australien umgesetzt worden.

Die durch NERO erzeugten Daten zur detaillierten Beschäftigungsentwicklung sollen die aktuelle Entwicklung des australischen Arbeitsmarktes detailliert quantifizieren, was eine deutliche Verbesserung gegenüber den bisherigen Datenquellen darstellt. Vor der Entwicklung von NERO standen zwei Arten von Datensätzen zur Verfügung (siehe auch Tabelle 6.2). Einerseits handelt es sich hierbei um den vom Australian Bureau of Statistics durchgeführte Census of Population. Dieser enthält tief gegliederte Informationen, wird aber nur alle fünf Jahre erhoben. Andererseits handelt es sich um verschiedene Datenquellen, die mit geringerer Verzögerung veröffentlicht werden, aber weniger tief gegliedert sind. Diese weisen zudem aufgrund geringer Fallzahlen hohe Schwankungen und starke Messungenauigkeiten auf.

NERO bietet Daten auf der detaillierten regionalen und beruflichen Ebene an und aktualisiert diese monatlich. Dies wird durch die Methode des Nowcastings ermöglicht, das auf einer Vielzahl von Datenquellen sowie Machine-Learning-Methoden basiert. Die Darstellung erfolgt mittels eines Dashboards, das verschiedene grafische Darstellungsweisen bietet. Auch ein Download der Daten ist auf dem Portal möglich. Es werden Ergebnisse zu 355 verschiedene Berufe und 88 Regionen berechnet und veröffentlicht. Von September 2015 an stehen damit über 30.000 monatliche Datenreihen bereit. Insgesamt ermöglichen die Daten also aktuelle und detaillierte Einblicke in den australischen Arbeitsmarkt (Wu et al., 2021).

Kurzfristige Entwicklungen genau nachzuverfolgen, war während der pandemiebedingten Schutzmaßnahmen im Jahr 2020 besonders wichtig und darum einer der Beweggründe für die Entwicklung des Datenprodukts. Für die Umsetzung zuständig war zunächst die australische National Skills Commission (NSC), eine 2020 von der australischen Regierung gegründete Institution, die Behörden, Regierungen und Privatpersonen mit wissenschaftlich gestützten und aktuellen Informationen zur Entwicklung des Arbeitsmarkts, insbesondere in Bezug auf die Nachfrage nach Kompetenzen, versorgen soll. Mittlerweile wurden die Aufgaben der NSC von der neu gegründeten Stelle Jobs and Skills Australia übernommen, auf deren Webseite nun auch das NERO-Dashboard zu finden ist.

Der Innovationsgehalt von NERO besteht zum einen in der Methode des Nowcastings, das auf verschiedenen Datenquellen und Machine Learning beruht, zum anderen in einer sehr aktuellen und nutzerfreundlichen Visualisierung detaillierter Informationen auf regionaler und beruflicher Ebene. Insofern könnte die Dateninnovation beim Nowcasting als interessanter Vergleichspunkt für in Deutschland laufende Projekte (insbesondere bei der Bundesagentur für Arbeit) dienen. Auch die Visualisierung kann Ansatzpunkte für eine Optimierung ähnlicher Portale in Deutschland bieten.

## 6.6.2 Technische Beschreibung

### 6.6.2.1 Datenquellen

Grundlage für die Berechnungen zur Erstellung des Nowcastings ist eine breite Auswahl an administrativen Daten aus dem Zensus, dem Labour Force Survey (Arbeitskräfteerhebung) des Australian Bureau of Statistics (ABS), Visumsdaten, Wirtschaftsindikatoren und internen Daten des Department of Employment and Workplace Relations. Außerdem benutzt Jobs and Skills Australia eigene Daten zu Online-Stellenanzeigen sowie Daten zu Online-Stellenanzeigen vom US-amerikanischen Unternehmen Lightcast. Alle für NERO verwendeten Datenreihen sind in Tabelle 6-2 aufgeführt.

**Tabelle 6-2: Verwendete Datenreihen**

Datenquelle	Datenreihe	Frequenz der Veröffentlichung
ABS – Census	Beschäftigung nach Region und Beruf	alle fünf Jahre
ABS – Labour Force Survey	Beschäftigung landesweit nach Beruf	Quartalsweise
	Beschäftigung nach Region und Beruf	Quartalsweise
	Beschäftigung nach Region gesamt	Monatlich
NSC – Internet Vacancy Index	Online-Stellenanzeigen nach Region und Beruf	Monatlich
Lightcast	Online-Stellenanzeigen nach Region und Beruf	Täglich
Department of Education, Skills and Employment – Jobactive-Programm Daten <sup>a</sup>	Jobactive-Vermittlungen nach Region und Beruf	Zweiwöchentlich
ABS – Weekly payroll jobs	Jobs mit Verdienst nach Region und Industrie	Wöchentlich
ABS – Job Vacancies	Offene Stellen nach Bundesstaat und Beruf	Quartalsweise
Home Affairs	Visumsinhaber*innen nach Bundesstaat und Beruf	Quartalsweise
ABS – National Accounts	Bruttoinlandsprodukt der Bundesstaaten	Jährlich

Quelle: Shamiri et al., 2022. Eigene Darstellung. Erläuterung: a) Jobactive war ein Regierungsprogramm, das unter anderem Arbeitsuchende mit Arbeitgebenden zusammenbrachte und so Arbeitsuchende in Arbeitsstellen vermittelte. Es wurde 2022 durch Workforce Australia ersetzt (Department of Employment and Workplace Relations, 2022).

Wie in Tabelle 6-2 zu sehen ist, werden als Datenquellen für NERO u.a. Datenreihen verwendet, die mit relativ hoher Frequenz auf Ebene der Regionen und Berufe vorhanden sind. Allerdings werden diese Daten außerhalb von NERO nur quartalsweise veröffentlicht, normalerweise zudem mit einer größeren Verzögerung. Außerdem werden die zugrunde liegenden Rohdaten als ungenau, unvollständig und anfällig für hohe Schwankungen bewertet (Shamiri et al. 2022). Aufgrund unterschiedlicher Erhebungsmethoden und teilweise kleiner Fallzahlen weisen die verschiedenen Datenquellen zudem teilweise abweichende Ergebnisse auf. Die NERO-Daten bündeln und glätten die Daten (Abschnitt 6.6.2.2) und stellen sie aktueller sowie mit genauerer zeitlicher Abdeckung zur Verfügung, was eine erhebliche Verbesserung gegenüber den Rohdaten darstellt.

### 6.6.2.2 Technisches Verfahren

Die Berechnung der aktuellen, detaillierten Daten erfolgt mithilfe eines Nowcasting-Verfahrens (Shamiri et al., 2022). Dieser Begriff bezeichnet eine Vorhersage der aktuellen Situation, die aufgrund von unvermeidlichen Verzögerungen noch nicht in der amtlichen Statistik oder sonstigen Messungen wiedergegeben werden kann. Nowcasting wird darum auch als Gegenwartsvorhersage bezeichnet und benutzt Daten von vergangenen Zeitpunkten und, wo vorhanden, aktuelle Indikatoren, um Daten am aktuellen Rand zu berechnen. Im Fall von NERO werden Daten für den gerade abgeschlossenen Monat erzeugt und bereitgestellt.

Umgesetzt wird das NERO-Verfahren mithilfe von Machine-Learning-Methoden. Zuvor müssen die vorhandenen Daten bereinigt und kontrolliert werden. Insbesondere ist beim Erstellen (bzw. „Trainieren“) des Modells wichtig, dass alle Daten einem konkreten und korrekten Zeitpunkt zugeordnet werden, um die Vorhersage nicht durch die Verwendung von Daten, die nach dem Vorhersagezeitpunkt erhoben wurden, zu verfälschen. Grund hierfür ist, dass es problematisch wäre, wenn das Modell beim Training mit historischen Daten eine andere Datengrundlage verwenden würde, als dies am aktuellen Rand möglich ist. Das Modell soll also beispielsweise nicht für das Nowcasting des Monats Juli 2015 Werte vom August 2015 verwenden, da am aktuellen Rand, beispielsweise für den Monat Juli 2023, keine Werte für August 2023 zur Verfügung stehen. Außerdem werden in einigen Datenreihen fehlende Werte imputiert und die Datenreihen insgesamt geglättet, um eventuelle Ausreißer nicht zu sehr ins Gewicht fallen zu lassen.

Wie bei Machine Learning üblich, werden Trainingsdaten und Lernalgorithmen genutzt, um die Vorhersagemodelle zu entwickeln. Es wird also zunächst eine Teilmenge der verfügbaren Daten (die Trainingsdaten) verwendet, um das Nowcasting durchzuführen, wobei die Lernalgorithmen das verwendete Modell bereits optimieren. Dann werden die (vorläufigen) Ergebnisse mit beobachtbaren Daten (den Testdaten) verglichen. Dabei werden insbesondere Daten aus dem Zensus von August 2016 zum Testen der Modelle verwendet, da diese besonders vollständig und umfangreich sind und als qualitativ hochwertig eingeschätzt werden. Der Vergleich der Testdaten für diesen Datenausschnitt mit den aus den zugehörigen Trainingsdaten erstellten Vorhersagen ist daher ein guter Indikator für die Genauigkeit des Modells. Der Vorhersagefehler, also die Differenz zwischen den Nowcasting-Vorhersagen auf Grundlage der Trainingsdaten und den Testdaten, geht abschließend erneut in das Modell ein, um dessen Spezifikation und Parameter weiter zu optimieren.

Das für NERO verwendete Nowcasting-Verfahren basiert auf drei Modellen mit unterschiedlichen Machine-Learning-Methoden: Random-Forest-, Gradient-Boost- und Elastic-Net-Regressionen (siehe Kapitel 3). Das finale NERO-Modell ist eine mittels einer linearen Regression entwickelte Kombination aus den drei Modellen. Die Vorhersagen, die das kombinierte Modell liefert, werden schließlich noch in zweierlei Hinsicht angepasst. Erstens werden die Ergebnisse des Nowcasting geglättet angegeben, wenn einzelne Beobachtungspunkte in detaillierten Datenreihen stark vom

nationalen Durchschnitt abweichen. Dadurch werden möglicherweise falsche oder irreführende Signale vermieden, die entstehen können, weil detaillierte Datenreihen oft nur auf kleinen Beobachtungszahlen beruhen. Zweitens werden aus Datenschutzgründen nur Kombinationen aus Region und Beruf ausgegeben, die mindestens zehn Beobachtungen umfassen. Andernfalls wird für diese Kombinationen ein gerundeter Wert ausgegeben.

Die NERO-Schätzungen werden geglättet, um langfristige Trends in den lokalen Arbeitsmärkten abzubilden. Infolge dieses Glättungsprozesses erfassen diese Schätzungen möglicherweise nicht die kurzfristige Dynamik auf dem Arbeitsmarkt, wie z. B. die beträchtlichen Auswirkungen der Schließungen infolge der Covid-19-Pandemie.

### 6.6.3 Anwendung

Die NERO-Daten werden planmäßig immer am ersten Mittwoch eines Monats für den jeweiligen Vormonat auf dem zum Projekt gehörigen Dashboard veröffentlicht. Dort sind die Daten frei zugänglich und auf verschiedene Arten, entsprechend den betrachteten Variablen, visuell aufbereitet. Auf zwei Themenseiten zu Berufen und Regionen können die Daten durchsucht und aktuelle Werte betrachtet werden. Außerdem stehen die verschiedenen Datenreihen vollständig, also nicht nur für den jeweils visualisierten Zeitraum, sondern seit Beginn der jeweiligen Erhebung, zum Download bereit.<sup>69</sup>

Durch die übersichtliche Gestaltung, den einfachen und öffentlichen Zugang und den großen Umfang der Daten bietet das Dashboard unterschiedlichen Zielgruppen neue Möglichkeiten zur Datennutzung. Insbesondere die Visualisierung hat das Potenzial, auch Personen anzusprechen, die ähnliche Daten vor Beginn des NERO-Projekts nicht genutzt hatten. Dies gelingt auch durch die einfache Bedienung des Dashboards und durch die kleinteilige Darstellung der Daten.

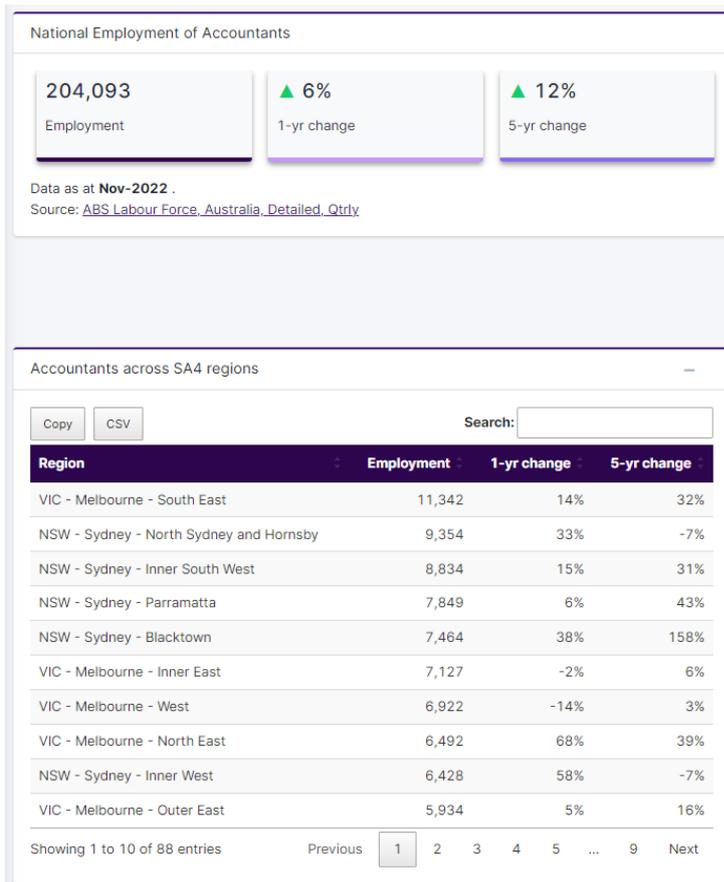
Abbildung 6-10 zeigt beispielhaft einen Ausschnitt der Themenseite Berufe, mit dem ausgewählten Beruf Buchhalter\*innen. Gezeigt wird die Zahl der in ganz Australien Beschäftigten in diesem Beruf sowie Veränderungen im Vergleich zum Vorjahresmonat und dem entsprechenden Monat fünf Jahre zuvor. In einer Tabelle sind außerdem die Beschäftigungszahlen im gewählten Beruf für alle Regionen aufgelistet.

Abbildung 6-11 verdeutlicht, dass auch einzelne Datenreihen, hier Buchhalter\*innen in der Region Melbourne – South East, betrachtet werden können. Zusätzlich wird die 5-Jahres-Veränderung in einer Karte, aufgeschlüsselt nach Regionen, dargestellt.

---

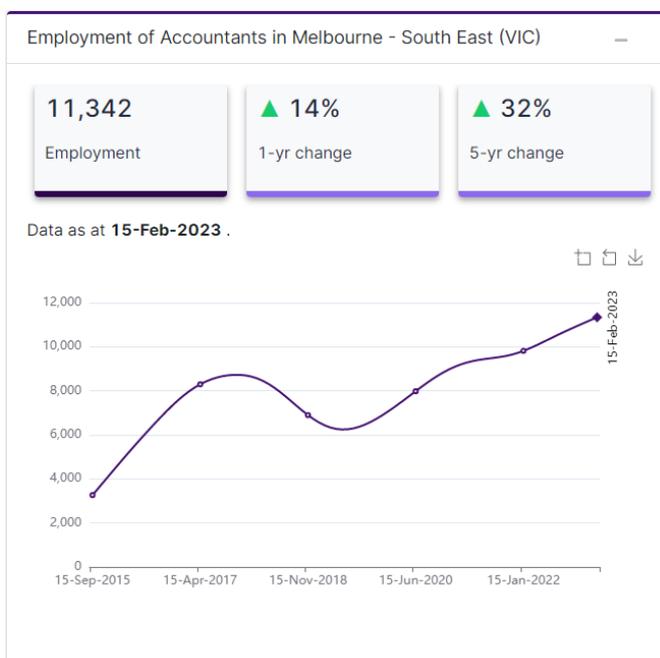
<sup>69</sup> Ex-post Revisionen der vorhergesagten Daten, wenn die tatsächlich beobachteten Daten zur Verfügung stehen, sind angedacht.

Abbildung 6-10: NERO-Dashboard: Überblick Beruf Buchhalter\*in



Quelle: Jobs and Skills Australia (2022). Aufgenommen am 20. März 2023.

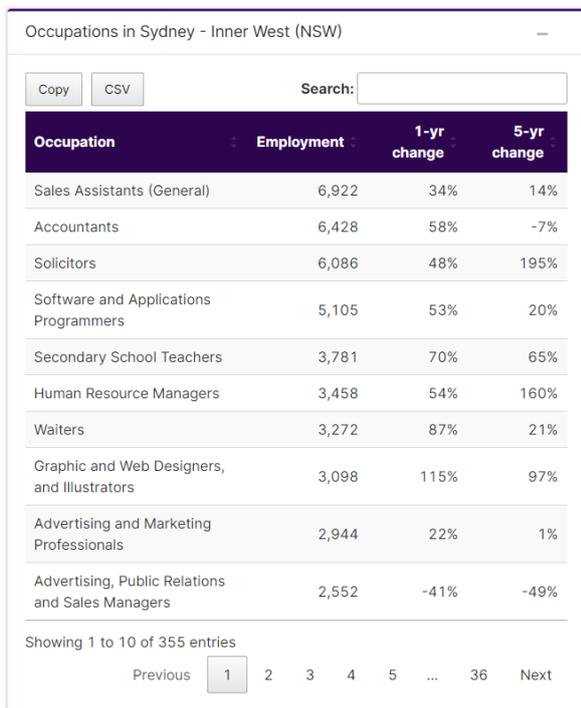
Abbildung 6-11: NERO-Dashboard: Zeitliche Entwicklung Beruf Buchhalter\*in



Quelle: Jobs and Skills Australia (2022). Aufgenommen am 20. März 2023.

Auf der Themenseite Regionen liegt der Fokus auf der Entwicklung von einzelnen Berufen in den Regionen. So wird in der in Abbildung 6-12 dargestellten Tabelle eine Übersicht über die Beschäftigungszahlen in allen Berufen in der Region Sydney – Inner West gegeben. Diese Tabelle lässt auch nach bestimmten Berufen durchsuchen und nach Spalten sortieren.

**Abbildung 6-12: NERO-Dashboard: Regionale Beschäftigungsentwicklung nach Berufen**



Occupations in Sydney - Inner West (NSW)

Copy CSV Search:

Occupation	Employment	1-yr change	5-yr change
Sales Assistants (General)	6,922	34%	14%
Accountants	6,428	58%	-7%
Solicitors	6,086	48%	195%
Software and Applications Programmers	5,105	53%	20%
Secondary School Teachers	3,781	70%	65%
Human Resource Managers	3,458	54%	160%
Waiters	3,272	87%	21%
Graphic and Web Designers, and Illustrators	3,098	115%	97%
Advertising and Marketing Professionals	2,944	22%	1%
Advertising, Public Relations and Sales Managers	2,552	-41%	-49%

Showing 1 to 10 of 355 entries

Previous **1** 2 3 4 5 ... 36 Next

Quelle: Jobs and Skills Australia (2022). Aufgenommen am 20. März 2023.

#### 6.6.4 Übertragbarkeit auf Deutschland

Die Dateninnovation liefert einen interessanten Vergleichspunkt zu in Deutschland laufenden Datenprojekten. Für den Aspekt des Nowcastings ist hierbei relevant, dass die BA in verschiedenen Statistiken Daten zu aktuellen Entwicklungen auf dem deutschen Arbeitsmarkt erhebt. In der Beschäftigungsstatistik werden alle sozialversicherungspflichtig und geringfügig Beschäftigten erfasst. Die Daten werden monatlich veröffentlicht, allerdings auf Grund von Meldefristen und der Erhebungsmethode mit einer Wartezeit von sechs Monaten (Bundesagentur für Arbeit, 2021). Zusätzlich wird monatlich der Bestand an sozialversicherungspflichtig bzw. geringfügig Beschäftigten mit 2 und 3 Monaten Wartezeit ermittelt und auf 6-Monatswerte hochgerechnet (Wirtschaftsabteilungen x Bundesländer) (Bundesagentur für Arbeit, 2022). Im Vergleich hierzu erscheint die für NERO gewählte Vorgehensweise umfassender und transparenter.

Ein internes Nowcasting wird jedoch für die IAB-Stellenerhebung durchgeführt, indem einmal im Quartal der aktuelle Stand der sozialversicherungspflichtig Beschäftigten geschätzt wird. Hierbei wird der von Hutter (2020) entwickelte Ansatz genutzt, der Daten der gesetzlichen Krankenversicherung verwendet, die zeitnah zur Verfügung stehen. Zudem ist dieses Nowcasting-Verfahren auch im Rahmen der zweimal pro Jahr veröffentlichten IAB-Kurzfristprognose im Einsatz (Bauer et al., 2023). Zentral für diese Kurzfristprognosen ist das integrierte Arbeitsmarktmodell (IAB-IAM), ein State-Space-Modell, welches die unterschiedlichen Frequenzen und unterschiedlichen Verfügbarkeiten der Variablen am aktuellen Rand berücksichtigt. Die Nowcasting-Methoden werden im IAB laufend weiterentwickelt. Während der verwendete Nowcasting-Ansatz vergleichbar ist, legt das NERO-

Modell einen stärkeren Schwerpunkt auf eine regionale und berufliche Disaggregation. Diese Möglichkeit besteht grundsätzlich auch für den deutschen Arbeitsmarkt.

Hinsichtlich des Dashboards stehen in Deutschland ähnlich wie beim australischen NERO-Projekt Daten zur sozialversicherungspflichtigen Beschäftigung in einer webbasierten Datenbank öffentlich zur Verfügung. In diesem Online-Tool können verschiedene Variablen betrachtet werden, unter anderem die Beschäftigungszahlen auf Kreisebene und auf Ebene der Berufe (3-Steller nach KldB 2010). Durch die Auswahl von Variablen und gewünschten Kennzahlen können individuelle Tabellen erstellt werden. Für diese Datenbank werden die Daten quartalsweise aktualisiert. Aus Datenschutzgründen werden sie außerdem gerundet. Angegeben ist jeweils der auf das nächste Vielfache von zehn gerundete Wert (Bundesagentur für Arbeit, 2020). Alle erstellten Tabellen stehen auch zum Download bereit (Bundesagentur für Arbeit, o. D.). Im Vergleich hierzu enthält das NERO-Dashboard Daten am aktuellen Rand, die durch Nowcasting gewonnen wurden, zudem sind die Daten visuell aufbereitet. Die Kombination aus Aktualität und visueller Aufbereitung könnte somit als zusätzliche Inspiration für in Deutschland existierende Datenprojekte dienen.

### 6.6.5 Fazit

Das australische Datenprojekt “Nowcast of Employment by Region and Occupation” (NERO) berechnet die aktuelle Beschäftigung auf regionaler und beruflicher Ebene und stellt die resultierenden Daten ansprechend visualisiert und zum Download bereit. Ein entscheidender Punkt bei der Nutzung von Nowcast-Modellen wie NERO ist die Verlässlichkeit der berechneten Daten, d.h. inwiefern die Ergebnisse des Nowcasts letztlich mit der Wirklichkeit übereinstimmen, was insbesondere in Krisenzeiten eine große Herausforderung darstellt. Dies wird im australischen Beispiel deutlich thematisiert, mit dem Ergebnis, dass die Modellergebnisse grundsätzlich valide sind, die Performance während der Corona-Pandemie jedoch zurückgeht (Shamiri et al., 2022).<sup>70</sup>

NERO ist auf zwei Arten innovativ. Erstens wird die Methode des Nowcastings verwendet, wobei eine Vielzahl von Datenquellen und Machine-Learning-Verfahren eingesetzt werden. Hierdurch stehen detaillierte Daten zur aktuellen Beschäftigung zur Verfügung, die für verschiedene Arbeitsmarktakteure von großem Interesse sind. Zweitens werden die Daten auf einem Dashboard zur Verfügung gestellt, das sich durch seine Nutzungsfreundlichkeit auszeichnet. Hinsichtlich beider Aspekte, also bezüglich Nowcast-Methode und Dashboard, erscheint NERO als interessanter Vergleichspunkt für verschiedene in Deutschland laufende Datenprojekte.

### 6.6.6 Quellen zum Steckbrief

Bauer, A., Gartner, H., Hellwagner, T., Hummel, M., Hutter, C., Wanger, S., Weber, E. & Zika, G. (2023), Rekord-Arbeitskräftebedarf in schwierigen Zeiten. IAB-Kurzbericht 5/2023. Abgerufen am 26.5.2023 von <https://doku.iab.de/kurzber/2023/kb2023-05.pdf>.

Bundesagentur für Arbeit. (o. D.). *Datenbank Beschäftigung*. Statistik der Bundesagentur für Arbeit. Abgerufen am 12. April 2023 von <https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Interaktive-Statistiken/Datenbanken/Datenbanken-BST-Nav.html>

<sup>70</sup> Der Weighted Mean Square Error (WAPE) wird für die Random-Forest-Spezifikation mit 13 Prozent angegeben. Die Performance des Modells wird bei größeren Veränderungen wie in der Corona-Pandemie jedoch schlechter.

Bundesagentur für Arbeit (2020). *Kurzinformation Statistische Geheimhaltung in den neuen webbasierten Anwendungen der Statistik der BA*. <https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Interaktive-Statistiken/Eckwerte-Arbeitsmarkt/Generische-Publikationen/Kurzinfo-Statistische-Geheimhaltung-webbasierte-Anwendungen.html? blob=publicationFile>

Bundesagentur für Arbeit (2021). *Kurzinformation Beschäftigungsstatistik*. <https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Definitionen/Generische-Publikationen/Kurzinformation-Beschaeftigungsstatistik.pdf? blob=publicationFile>

Bundesagentur für Arbeit (2022). Statistik der sozialversicherungspflichtigen und geringfügigen Beschäftigung. Abgerufen am 13. September 2023 von <https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Methodik-Qualitaet/Qualitaetsberichte/Generische-Publikationen/Qualitaetsbericht-Statistik-Beschaeftigung.pdf? blob=publicationFile>

Department of Employment and Workplace Relations. (2022). *Jobactive*. Department of Employment and Workplace Relations. Abgerufen am 12. April 2023 von <https://www.dewr.gov.au/jobactive#:~:text=What%20is%20jobactive%3F%20jobactive%20was%20the%20Australian%20Government,Workforce%20Australia%20replaced%20jobactive%20on%204%20July%202022.>

Hutter, C. (2020). A new indicator for nowcasting employment subject to social security contributions in Germany. *Journal for Labour Market Research*, 54(1), 4.

Jobs and Skills Australia. (2022). *NERO dashboard*. Jobs and Skills Australia. Abgerufen am 12. April 2023 von <https://www.jobsandskills.gov.au/work/nero/nero-dashboard>

Shamiri, S., Ngai, L., Lake, P., Shan, Y., McMillan, A., Smith, T. & Sharma, K. (2022). Nowcasting the Australian Labour Market at Disaggregated Levels. *The Australian Economic Review*, 55(3), 389-303. Wu, D., Ngai, L., Shamiri, S., Mcmillan, A., Shan, Y., Smith, T. & Lake, P. (2021). *Finding NERO: A Nowcast of Employment by Region and Occupation | National Skills Commission*. National Skills Commission. Abgerufen am 23. März 2023 von <https://www.nationalskillscommission.gov.au/insights/finding-nero-nowcast-employment-region-and-occupation>

## 7. Zusammenfassung und Schlussfolgerungen

Die Steckbriefe in diesem Bericht stellen beispielhafte Dateninnovationen anderer Länder aus dem Bereich der Arbeitsmarkt- und Sozialpolitik dar. Sie beziehen sich auf drei Bereiche: Verfahren und Produkte der Datengenerierung und -verknüpfung, Methoden der Datennutzung und Bereitstellung von Daten. Viele der beschriebenen Innovationen betreffen mehr als einen dieser Bereiche. Bei den Dateninnovationen geht es sowohl um Anwendungen aus dem Bereich der öffentlichen Verwaltung als auch um Ergebnisse von Forschungsvorhaben. Die Dateninnovationen leisten unmittelbare Beiträge zur Verbesserung von Prozessen und Ergebnissen in der Verwaltung und in der Forschung, darüber hinaus kommen sie aber auch spezifischen Zielgruppen wie Stellensuchenden, Versicherten, Arbeitslosen oder allgemein allen Bürger\*innen zugute.

Die internationalen Dateninnovationen wurden aufgrund ihres Innovationsgrades und ihres Transferpotenzials für Deutschland aus einer größeren Zahl von recherchierten Anwendungen ausgewählt. Es wurden nur solche Dateninnovationen erfasst, die potenziell eine thematische Verbindung zu den Themenbereichen des BMAS haben, also der Arbeitsmarkt- und Beschäftigungspolitik einschließlich der beruflichen Weiterbildung, dem Arbeitsschutz, dem Arbeitsrecht, der Sozialversicherung und Alterssicherung sowie der sozialen Teilhabe und Inklusion.

Die betrachteten Dateninnovationen erleichtern die Datenbereitstellung und damit die Nutzung durch unterschiedliche Stakeholder sowie die Zivilgesellschaft. Dabei zeigen sich bei den ausgewählten Dateninnovationen in den drei Bereichen deutliche Schwerpunkte:

- Im Bereich der Datengenerierung und Datenverknüpfung dominiert die Verknüpfung von unterschiedlichen Daten, insbesondere solchen Daten, die im Rahmen administrativer Prozesse erhoben wurden. Die Verknüpfung verbindet diese Daten in der Regel auf der Ebene einzelner Personen. Dabei kommen mitunter komplexe Verfahren zum Einsatz, um den Datenschutz zu gewährleisten. Dies erweitert das Merkmalspektrum der zur Verfügung stehenden Daten und damit die Nutzungsmöglichkeiten erheblich.
- Im Bereich der Methoden nutzen viele der Innovationen Verfahren des Machine Learning. Diese Methodik ist z. B. für verbesserte Vorhersagen geeignet, beispielsweise für die Vorhersage des Zeitpunkts des individuellen Renteneintritts oder für die Vorhersage des Risikos der Langzeitarbeitslosigkeit. Die Anwendung von Machine Learning resultiert meist in deutlich besseren Vorhersagen als die Nutzung konventioneller statistischer Verfahren. Neben dem Einsatz der Machine Learning zu Vorhersagezwecken gibt es eine Reihe von Anwendungen zur automatisierten Bildung von Kategorien und Clustern.
- Im Bereich der Bereitstellung von Daten ist das Ziel der unterschiedlichen Verfahren, die Anwendung der Daten zu verbreitern und dabei die Akteur\*innen zu einer selbstständigen Nutzung der Daten zu ermächtigen, beispielsweise durch die Schaffung von Datenportalen.

Für diese drei Bereiche ergeben sich Schlussfolgerungen für die Frage, wo aus Dateninnovationen ein besonders hoher Mehrwert zu erwarten ist.

- Die Verknüpfung verschiedener Datenquellen bietet häufig einen hohen Mehrwert für konkrete Prozesse oder Fragestellungen in der Verwaltung, aber genauso für Forschungsfragen. Die hohe Qualität von Daten aus administrativen Prozessen ist hierbei von großem Vorteil, der von vielen der betrachteten Innovationen genutzt wird. Durch die Verknüpfung gelingt es, das in einzelnen administrativen Datensätzen oft enge, an den Verwaltungszwecken orientierte Merkmalspektrum zu erweitern.
- Eng verwandt mit der Verknüpfung ist der Austausch von Daten zwischen Verwaltungen im Zuge der Verwirklichung des Once-Only-Prinzips. Die technischen Voraussetzungen für den

Austausch und die Verknüpfung sind teilweise ähnlich, es geht hier jedoch nicht um die Anlage eines neuen, größeren Datensatzes, sondern um den Transfer von Einzelinformationen im Bedarfsfall.

- Im Bereich der Daten ermöglichen Big Data neue und – insbesondere in Verknüpfung mit Machine-Learning-Methoden – zeitnahe Analysen von Arbeitsmarktentwicklungen. Beispielsweise kann mithilfe von Informationen aus Stellenanzeigen aus Online-Portalen (Online Job Vacancies, OJV) der Bedarf an bestimmten Fähigkeiten und Kenntnissen ermittelt werden. Besonders viele Möglichkeiten bietet dabei die Verbindung von Stellenanzeigen mit Bewerbungen. Weitere Anwendungen von Big Data greifen auf Daten aus sozialen Netzwerken zurück. Die so gewonnenen Informationen können einzelnen Personen bei der Entscheidungsfindung, z. B. im Hinblick auf die Berufswahl, helfen. Sie unterstützen darüber hinaus Vermittlungsfachkräfte und andere in der Verwaltung Tätige und schaffen eine verbesserte Informationsgrundlage für die Arbeitsmarkt- und Sozialpolitik.
- Verfahren des Machine Learning können besonders für Vorhersagen auf der individuellen Ebene genutzt werden. So lassen sich beispielsweise Prognosen zum individuellen Erwerbszustand (z. B. Arbeitslosigkeit und Bezug von Lohnersatzleistungen, Renteneintritt, dauerhafte Erwerbsunfähigkeit) treffen oder die Wahrscheinlichkeit von Compliance-Verstößen z. B. beim Arbeitsschutz oder beim Mindestlohn berechnen. Viele der betrachteten Verfahren finden sich noch in einem Pilotstadium, sodass hier weiteres Anwendungspotenzial besteht. Dabei gilt es, die Chancen mit den Risiken – zum Beispiel möglichen Verzerrungen – abzuwägen.
- Andere Verfahren des Machine Learning werden zur Kategorisierung und Bündelung, z. B. von großen Datenmengen aus Stellenportalen, benutzt. Zusammen mit der automatischen Verarbeitung von Textdaten ergeben sich verbesserte Informationsgrundlagen, unter anderem im Bereich der Arbeitsvermittlung und der Unfallversicherung.
- Das Innovationspotenzial von Machine Learning resultiert in der Regel weniger aus grundlegend neuen Methoden als aus ihrer innovativen Verwendung. So lässt sich beispielsweise die Evaluierung der Arbeitsmarkteffekte von Mindestlöhnen durch Machine-Learning-Methoden verbessern.
- Portallösungen machen es einfacher, auf die Ergebnisse von Daten und Datenverarbeitungen zurückzugreifen. Sie können beispielsweise die Informationen verbessern, die Stellensuchende über für sie geeignete offene Stellen erhalten. Sie erleichtern aber auch den Zugang zu anderweitig nicht verfügbaren Ergebnissen für Stakeholder (Politik, Verwaltung und Interessengruppen).

Durch die Darstellung der Innovationen zieht sich als roter Faden, dass sich Innovationen für die Forschung und für die Verwaltung gegenseitig bedingen. Machine-Learning-Verfahren werden meist durch Forschungseinrichtungen entwickelt und dann in der Verwaltung umgesetzt. Umgekehrt gewinnt die Forschung, wenn sie ursprünglich für die Verwaltung vorgesehene Daten nutzen kann. Die Vorteile, die für beide Seiten entstehen, werden nur dann ausgenutzt, wenn Verwaltung und Forschung gemeinsam an der Entwicklung neuer Verfahren und Produkte beteiligt sind. Deshalb ist es sinnvoll, Möglichkeiten des Dialogs zwischen Forschung und Verwaltungen vorzusehen und einzurichten.

Alle betrachteten Dateninnovationen haben Transferpotenzial für die Nutzung in Deutschland oder können zumindest Impulse für ähnliche Innovationen in Deutschland geben, allerdings in unterschiedlichem Ausmaß und auf unterschiedliche Weise. In manchen Fällen ist ein direkter Transfer einer internationalen Dateninnovation möglich. Dies wird dadurch erleichtert, dass es in Deutschland oft bereits ähnliche Verfahren oder Vorstufen zu den Dateninnovationen gibt, so dass eine Implementierung in Deutschland auf bereits verwirklichten Verfahren aufsetzen kann, oder dass Datensätze bereits angelegt sind, die mit modernen Verfahren verknüpft und analysiert werden

können. In vielen Fällen ist ein direkter Transfer jedoch nicht möglich, weil kritische Voraussetzungen nicht gegeben sind. Hier kommt vielfach eine modifizierte Umsetzung in Betracht oder der Transfer einer allgemein verwendbaren Methodik von einem Anwendungsfeld in ein anderes.

Der Aufwand eines Transfers würde sich zwischen den betrachteten Dateninnovationen erheblich unterscheiden. In einigen Fällen, beispielsweise der Nutzung von Open Government Data, aber auch dem Nowcasting der Arbeitsmarktentwicklung, das in Deutschland teilweise schon implementiert wird, ist der Aufwand vergleichbar gering. Die Schaffung von Verknüpfungsmöglichkeiten zwischen Daten verlangt dagegen nach einer Vielzahl unterschiedlicher Schritte und ist entsprechend aufwändig. Die Entscheidung zur Verwirklichung einer Dateninnovation liegt immer eine Kosten-Nutzen-Abwägung zugrunde. Dabei muss beachtet werden, dass diese Abwägung durch die Einbeziehung neuer Verfahren beeinflusst wird, sodass Erfahrungswerte und in der Vergangenheit erstellte Analysen ihre Gültigkeit verlieren.

Allgemein bestehen günstige Bedingungen für den Transfer nach Deutschland, weil neben dem Vorhandensein von Daten auch die vorhandene Erfahrung mit Daten und Methoden in den Verwaltungen und Forschungseinrichtungen wie z. B. der BA, der DRV und anderen Sozialversicherungen, dem IAB und der BAuA gute Ausgangsbedingungen darstellen. Daneben bestehen jedoch auch Barrieren für den Transfer, die sich aus dem rechtlichen Rahmen, dem föderalen Staatsaufbau, der Aufgabenteilung und Zuständigkeit der Akteure, aber auch der Akzeptanz technischer Innovationen oder anderen Wertvorstellungen ergeben.

Ein Teil der betrachteten Dateninnovationen ließe sich vermutlich auch im Rahmen der in Deutschland geltenden gesetzlichen Regelungen zur Datennutzung und zum Datenschutz und des entsprechenden Verwaltungshandelns umsetzen. Derzeit sind allerdings viele Fragen durch gesetzliche Vorschriften nicht ausreichend klar geregelt. So ist die Verknüpfung von administrativen Datensätzen zwar im Rahmen des Bundesdatenschutzgesetzes (BDSG) prinzipiell möglich, und einzelne administrative Datensätze wurden in Deutschland bereits verknüpft. Dies setzt jedoch einen aufwändigen Genehmigungsprozess voraus, denn die Entscheidung wird von den Aufsichtsbehörden für jede einzelne Verknüpfung getroffen.

Für die Anwendung vieler Dateninnovationen fehlte bislang noch ein klarer Rechtsrahmen. Auf europäischer Ebene bildet der Data Governance Act eine wichtige Grundlage für die Nutzung und Weitergabe von Daten öffentlicher Stellen. Diese Verordnung ist von September 2023 direkt anwendbar. In der Rechtsanwendung fehlen digitalfähige Rechtsbegriffe; häufig verwendete Begriffe oder Definitionen, beispielsweise im Kontext von Machine-Learning-Methoden, sollten daher rechtssicher ausgearbeitet werden. Auch ist die Einrichtung entsprechender Verwaltungsverfahren in Deutschland sehr zeitaufwändig. Dies erschwert in Deutschland die Entwicklung und Implementation von Dateninnovationen.

Machine-Learning-Methoden bieten neben ihrem hohen Innovationspotenzial die Gefahr der Diskriminierung. Dies kann z. B. dadurch hervorgerufen werden, dass durch Diskriminierung verursachte Arbeitsergebnisse von einem Machine-Learning-Algorithmus als anzustrebendes Marktergebnis bewertet werden und daher in Stellenempfehlungen für Arbeitsuchende eingehen. Daher sollte stets auch die ethische Dimension des Einsatzes von Machine-Learning-Methoden beachtet werden. Die Selbstverpflichtenden Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung (BMAS, 2022) nennen eine Reihe von Vorgehensweisen, wie diesem bekannten Problem des Machine Learning begegnet werden kann. Diese sollten auf dem jeweiligen Stand der Technik weiterentwickelt werden.

Für die genannten Hemmnisse stehen vielfach Lösungsmöglichkeiten bereit. Reallabore oder Sandboxes (vgl. Steckbrief 5.8) könnten die Datenschutzkonformität von Dateninnovationen prüfen –

davon könnten sowohl die Aufsichtsbehörden als auch die Innovator\*innen sowie die Zielgruppe der Innovation profitieren. Teilweise lässt sich der Datenschutz selbst durch innovative Datenstrukturen, z. B. das Hub-and-Spokes-Modell, und durch Verschlüsselungstechnologien wie z. B. die homomorphe Verschlüsselung gewährleisten (vgl. Steckbrief 4.1). Auch die Erstellung von synthetischen Daten kann eine Möglichkeit sein, Datennutzungen und Verknüpfungen von geschützten Daten mit dem Datenschutz zu vereinbaren (vgl. Steckbrief 4.7). Solche technischen Möglichkeiten gilt es zu berücksichtigen, wenn über die Genehmigung von Datennutzungen und die Verknüpfung von Daten entschieden wird.

Etliche Dateninnovationen zeigen, dass die Akzeptanz bei den Betroffenen und in der Öffentlichkeit eine wesentliche Voraussetzung für den Erfolg der Innovation ist. Die Zielgruppe, die künftig mit einer Dateninnovation arbeiten soll (z. B. Vermittlungsfachkräfte in der Arbeitsverwaltung oder Arbeitsuchende) sollte in den Entwicklungsprozess eingebunden werden, insbesondere in die Gestaltung von Bedienoberflächen für Portale, die Eröffnung von Möglichkeiten der Änderung von Einstufungen und den Umgang mit Diskriminierungen. Tools und Portale sind entsprechend zu erproben und zu evaluieren.

Bei einem Transfer der Dateninnovationen oder einzelner ihrer Elemente sollten die vorliegenden Erfahrungen aus anderen Ländern genau berücksichtigt werden. Hierfür ist eine tiefere Analyse erforderlich, als sie im Rahmen dieses Berichts geleistet werden konnte. Diese ergibt sich am besten im direkten Dialog zwischen den Entwickler\*innen aus den anderen Ländern und den Beteiligten in Deutschland.

## Literaturverzeichnis

Hinweis: Die in den einzelnen Steckbriefen enthaltenen Literaturangaben finden sich am Ende des jeweiligen Steckbriefes.

- Athey, S. & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685-725. <https://doi.org/10.1146/annurev-economics-080217-053433>
- Bähr, H., Broder, H., Dietz, M., Förster, M. & Klier, M. (2022). Neue Wege bei Beratung und Arbeitsvermittlung durch Digitalisierung: Online-Chat ermöglicht Informationsaustausch und gegenseitige Unterstützung. *IAB-Kurzbericht* 09/2022, Nürnberg.
- Becker, J., Benke, A., Fettke, P., Gutermuth, O., Halsbenning, S., Houy, C., Hoffmeister, B., Müller, R. & Scholta, H. (2022). Machbarkeitsstudie zu den IT-technischen Anforderungen einer weitergehenden Digitalisierung im Bereich Soziale Sicherheit. *BMAS-Forschungsbericht* 600, Berlin.
- Bertschek, I., Bonin, H., Janßen, R. & Wenzel, T. (2021). Einsatz datengetriebener Produkt- und Preisdifferenzierungen und Folgen für den Sozialstaat. Kurzexpertise. *BMAS-Forschungsbericht* 579, Berlin.
- Bundesministerium für Arbeit und Soziales (BMAS) (2022). Selbstverpflichtende Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung. Abgerufen am 16. August 2023 von <https://www.bmas.de/DE/Service/Publikationen/Broschueren/a862-leitlinien-ki-einsatz-behoerdliche-praxis-arbeits-sozialverwaltung.html>
- Bundesministerium des Innern und für Heimat. (2022, 9. Dezember). Gründungskommission stellt erste Ideen für das Dateninstitut auf dem Digital-Gipfel vor. BMI. Abgerufen am 24. Mai 2023 von <https://www.bmi.bund.de/SharedDocs/pressemitteilungen/DE/2022/12/dateninstitut.html>
- Bundesministerium für Wirtschaft und Klimaschutz, Bundesministerium des Innern und für Heimat. (2023). Konzept zum Aufbau des Dateninstituts. Abgerufen am 26. Juli 2023 von <https://www.bmi.bund.de/DE/themen/it-und-digitalpolitik/it-des-bundes/dateninstitut/dateninstitut-node.html>
- Courty, P., Kim, D.H. & Marschke, G. (2011). Curbing cream-skimming: Evidence on enrolment incentives. *Labour Economics*, 18(5), 643-655.
- Deutsche Rentenversicherung. (2022). Forschungsdatenzentrum der Rentenversicherung. Eservice Deutsche Rentenversicherung. Abgerufen am 24. Mai 2023 von [https://www.eservice-drv.de/FdzPortalWeb/dispccontent.do?id=main\\_fdz\\_forschung\\_laengst&chmenu=ispvwNavEntriesByHierarchy41](https://www.eservice-drv.de/FdzPortalWeb/dispccontent.do?id=main_fdz_forschung_laengst&chmenu=ispvwNavEntriesByHierarchy41)
- Digitalstrategie Deutschland. (2023). Die Pläne für ein deutsches Dateninstitut schreiten voran. BMDV. Abgerufen am 24. Mai 2023 von <https://digitalstrategie-deutschland.de/dateninstitut/>
- Feenstra, C. Y., & Bunzeck, I. I. (2011). Copy & paste policies?! An analysis of the transferability of successful local and national policies related to alternative fuels. Energy research centre of the Netherlands. <http://www.ecn.nl/docs/library/report/2010/m10079.pdf>. Last accessed: 23.02.2023.

- Grimm, V., Janser, M. & Stops, M. (2021). Neue Analyse von Online-Stellenanzeigen: Kompetenzen für die Wasserstofftechnologie sind jetzt schon gefragt. *IAB-Kurzbericht* 11/2021, Nürnberg.
- Hutter, C. (2020). A new indicator for nowcasting employment subject to social security contributions in Germany. *Journal for Labour Market Research* 54(1), 4.
- Lettieri, V. (2021). Record Linkage: Datenverknüpfung von (personenbezogenen) Daten zu Forschungszwecken. Datenschutz Notizen. Abgerufen am 26. Juni 2023 von <https://www.datenschutz-notizen.de/record-linkage-datenverknuepfung-von-personenbezogenen-daten-zu-forschungszwecken-0132992/>
- OECD (2020a). *OECD Open, Useful and Re-usable data (OURdata) Index: 2019*. Paris: OECD.
- OECD (2020b). Digital Government Index. 2019 results. *OECD Public Governance Policy Papers*. Paris: OECD.
- Projekt ADEST (2022, Oktober). *ADEST - Automatisierte Datenübernahme aus Stellenangeboten mittels maschinellen Lernens* [Grundlagenpräsentation]. Bundesagentur für Arbeit.
- Rat für Sozial- und Wirtschaftsdaten (RatSWD) (2020). Handreichung Datenschutz. 2. Vollständig überarbeitete Auflage.
- Rat für Sozial- und Wirtschaftsdaten (RatSWD) (2022). Positionspapier des RatSWD: Eckpunkte für ein Forschungsdatengesetz. Empfehlungen des Rates für Sozial- und Wirtschaftsdaten (RatSWD) für die Erarbeitung eines Forschungsdatengesetzes.
- Stephan, G., Rässler, S. & Schewe, T. (2006). Das TrEffeR-Projekt der Bundesagentur für Arbeit: die Wirkung von Maßnahmen aktiver Arbeitsmarktpolitik, *Zeitschrift für ArbeitsmarktForschung - Journal for Labour Market Research*, 39(3/4), 447-465.
- Stops, M., Bächmann, A.-C., Glassner, R., Janser, M., Matthes, B., Metzger, L.-J., Müller, C. & Seitz, J. (2021). Extracting Skill Requirements from Job Ads – the “Machbarkeitsstudie Kompetenz-Kompass”. *IAB-Forschungsbericht* 07/2021, Nürnberg.

Diese Publikation wird im Rahmen der Öffentlichkeitsarbeit des Bundesministeriums für Arbeit und Soziales kostenlos herausgegeben. Sie darf weder von Parteien noch von Wahlbewerbern oder Wahlhelfern während des Wahlkampfes zum Zwecke der Wahlwerbung verwendet werden. Dies gilt für Europa-, Bundestags-, Landtags- und Kommunalwahlen. Missbräuchlich ist insbesondere die Verteilung auf Wahlveranstaltungen, an Informationsständen der Parteien sowie das Einlegen, Aufdrucken oder Aufkleben parteipolitischer Informationen oder Werbemittel. Untersagt ist gleichfalls die Weitergabe an Dritte zum Zwecke der Wahlwerbung. Unabhängig davon, wann, auf welchem Weg und in welcher Anzahl diese Publikation dem Empfänger zugegangen ist, darf sie auch ohne zeitlichen Bezug zu einer bevorstehenden Wahl nicht in einer Weise verwendet werden, die als Parteinahme der Bundesregierung zugunsten einzelner politischer Gruppen verstanden werden könnte. Außerdem ist diese kostenlose Publikation - gleichgültig wann, auf welchem Weg und in welcher Anzahl diese Publikation dem Empfänger zugegangen ist - nicht zum Weiterverkauf bestimmt.

Alle Rechte einschließlich der fotomechanischen Wiedergabe und des auszugsweisen Nachdrucks vorbehalten.