Kapitel 3 – »Data Science« als soziales Phänomen: Genese und multiple Perspektiven

In diesem Kapitel diskutiere ich wichtige Forschungsarbeiten, die das Phänomen »Data Science« aus unterschiedlichen Perspektiven untersuchen. Die Auswahl der existierenden Literatur schliesst dabei an das in der Einleitung dargelegte zentrale Erkenntnisinteresse, die Fragestellungen und die theoretisch-analytische Perspektive an. Die Diskussion der Studien dient dazu, den Blick auf den Untersuchungsgegenstand zu schärfen und in der weiteren Forschungslandschaft zu verorten. In einem ersten Teil widme ich mich Arbeiten, die durch disziplinäre und theoretische Perspektiven die Etablierung der Datenwissenschaften interpretieren (Kap. 3.1). Dabei skizziere ich die Genese und Etablierung des Begriffs »Data Science« an der Schnittstelle von Wissenschaft, Technologieindustrie sowie Forschungs- und Wissenschaftspolitik. Danach untersuche ich Positionsbezüge innerhalb der Wissenschaft, die den Gegenstand einerseits in einer disziplinären Differenzierungs-, andererseits in einer universalen Entdifferenzierungsperspektive betrachten. Schliesslich adressiere ich Arbeiten, die die Entstehung der Datenwissenschaften primär als Professionalisierung deuten.

Parallel zu den begriffshistorischen, disziplinären und sozialtheoretischen Erklärungs- und Deutungsversuchen beschäftigt sich ein weiteres Spektrum der Forschungsliteratur intensiv mit der Frage nach den relevanten Kompetenzen, Qualifikationen oder Tätigkeitsprofilen. Kompetenzen stellen – wie ich empirisch noch zeigen werde – einen zentralen Begriff des Untersuchungsfeldes dar. Einer konstruktivistischen Perspektive folgend verstehe ich unter Kompetenzen »sozial zugeschriebene Qualitäten, die sich über vielgestaltige Kommunikationen und Interaktionen manifestieren bzw. als sich manifestierend dem Subjekt attestiert werden« (Kurtz 2010: 8). Die Suche nach den ›richtigen‹ Kompetenzen stellt ein zentrales Erkenntnisinteresse für Akteur*innen verschiedener Felder dar, die in die Etablierung der Datenwissenschaften involviert sind. Obwohl viele Arbeiten eher instrumentelle Ziele verfolgen, indem sie darauf ausgerichtet sind, Unternehmen und anderen Akteur*innen Wissen über die Praxis von Data Scientists zu vermitteln und somit deren Rekrutierung zu erleichtern, liefern solche Kompetenzkonstruktionen auch instruktive Rückschlüsse über die transversale Konstruktion der Datenwissenschaften als soziales Phänomen. Entsprechend sind sie von Interesse für die vorliegende Arbeit. Ich fasse deshalb im zweiten Teil zusammen, wie Akteur*innen durch Kompetenzkonstruktionen und Tätigkeitsprofile als feld- und organisationsspezifische Perspektivierungen zur Ge-

nese der Datenwissenschaften im Arbeitsmarkt (Kap. 3.2), in der Bildungs- und Forschungspolitik (Kap. 3.3) sowie im akademischen Feld (Kap. 3.4) beitragen. Mit der Anordnung nehme ich eine Strukturierung vor, die ich in den beiden empirischen Teilen der Arbeit wieder aufgreifen werde. Abschliessend rekapituliere ich in einem Zwischenstand die wichtigsten Punkte des dargestellten Forschungsstandes (Kap. 3.5).

3.1 Begriffliche, disziplinäre und theoretische Perspektiven auf die Datenwissenschaften

3.1.1 »Data Science« als Label: Eine kurze Begriffsgeschichte

> »The absence of clear boundaries defining data science, and the many people co-opting the term for their own, is a good thing for the burgeoning function. It creates more interest in data science, both to support organizational decision-making, as well as to attract more talent into the field of data science« (EMC 2011: 4).

Der Begriff »Data Science« ist keineswegs neu, sondern findet seit den 1960er-Jahren Verwendung in bestimmten akademischen Disziplinen (Chatfield et al. 2014: 3). Dem dänischen Informatiker Peter Naur wird die Verwendung der Begriffe »Data Science« bzw. »Datalogy« als Substitute für Computerwissenschaften zugeschrieben (Irizarry 2020; Donoho 2017: 763). Ebenfalls aus jener Zeit stammen einige hinsichtlich der Emergenz der Datenwissenschaften als Wissensfeld zentrale Texte: Der Artikel »The Future of Data Analysis« des Statistikers und Bell-Labs-Ingenieurs John Tukey (1962) sowie das spätere Lehrbuch »Exploratory Data Analysis« gelten in der Statistik und den Computerwissenschaften als wichtige Beiträge für jene methodisch-technische Expertise, die in den 1990er-Jahren als »Data Mining« Bedeutung erlangt (Dasu & Johnson 2003; Fisher 2001; Han et al. 2012).

Im Laufe der 1990er-Jahre erhält das Label »Data Science« breitere Akzeptanz an der Schnittstelle von Computerwissenschaften, Statistik und weiteren Disziplinen: Verschiedene Repräsentanten im wissenschaftlichen Feld äussern sich dazu (Hayashi 1998; Wu 1997) und erste Konferenzen unter dem Label finden statt (Hayashi et al. 1998; Ueno 2017). Parallel dazu werden wissenschaftliche Zeitschriften gegründet. Als eigentliches Manifest in der jüngeren Geschichte kann ein Artikel des Statistikers William S. Cleveland in der International Statistical Review 2001 gelten. Während Friedman (2001) in derselben Ausgabe dazu aufruft, das Feld der »Data Analysis« nicht den Computerwissenschaften alleine zu überlassen, legt Cleveland (2001: 21f.) mit direktem Bezug auf Tukey einen »Action Plan« für »Data Science« als eigenständige wissenschaftliche Disziplin inklusive eines Curriculums mit eigenen Schwerpunkten und Gewichtungen vor. Obwohl zu jener Zeit in den USA erste Studienprogramme unter dem Label lanciert werden, stiess der Weckruf Clevelands zumindest im Feld der Statistik lediglich auf geringen Nachhall (Kane 2014).

Auch aus wissenschaftspolitischer Sicht häufen sich in jenen Jahren Studien und Bezugnahmen auf das Label »Data Scientist« zur Beschreibung jener Professionen, die innerhalb wissenschaftlicher Forschungsgruppen Aufgaben des Managements, der Kuratierung und Aufbereitung von Daten besorgen (NSB 2005; Swan & Brown 2008; IWGDD 2009). Das US-amerikanische National Science Board (NSB) definierte »Data

Scientists« als jene »information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection« (NSB 2005: 27). Neben der Durchführung wissenschaftlicher Analysen werden die Beratung und Ausbildung in statistischen Analysemethoden, die Visualisierung und Exploration digitaler Daten sowie die Entwicklung neuer Datenbanktechnologien genannt. Die Definition des NSB zeugt von einem breiten, inklusiven Rollenverständnis, das disziplinäre und professionelle Grenzen überschreitet.

Eine breitere Anerkennung über das wissenschaftliche Feld hinaus setzt jedoch erst gegen Ende der 2000er-Jahre ein,[1] als Repräsentanten grosser Social-Media-Plattformen sich selbst als »Data Scientists« zu bezeichnen beginnen bzw. in Anspruch nehmen, die neue Berufsrolle überhaupt erst erschaffen zu haben:

> »At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn't quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization in a clear and concise fashion. To capture the skill set required to perform this multitude of tasks, *we created the role* of ›Data Scientist‹« (Hammerbacher 2009: 84; eigene Hervorhebung).

In der Folge wird die Bezeichnung medial weit verbreitet, was sich in unzähligen Selbst- und Fremdzuschreibungen äussert (Varian 2009; Conway 2010; Davenport & Patil 2012; Voulgaris 2014). Zur selben Zeit häufen sich Bedarfsanalysen von Beratungsunternehmen wie McKinsey und anderen ökonomischen Akteur*innen, die im Einklang mit dem »Hype« um »Big Data« (Carter & Sholler 2016; Donoho 2017; Elish & boyd 2018) eine stetig steigende Nachfrage nach Datenwissenschaftler*innen für die kommenden Jahre prognostizieren, womit das gegenwärtige Angebot der Universitäten nicht mithalten könne (Manyika et al. 2011). Um den wachsenden Bedarf decken zu können, entstanden in den Folgejahren Studiengänge in Datenwissenschaften an zahlreichen Universitäten und Hochschulen rund um den Globus. Trotz der Bemühungen von Wissenschaftler*innen und wissenschaftspolitischen Institutionen waren es Akteur*innen in der Industrie, die das Wissensgebiet über die Professionsbezeichnung »Data Scientist« popularisierten und personifizierten (Davenport & Patil 2012) und diesem somit zum Durchbruch verhalfen (Gehl 2015).

Ein Vergleich verschiedener Definitionen zeigt, dass die meisten Texte Elemente aus Statistik, Mathematik, Informatik und (Software) Engineering, bisweilen auch Ökonomie, Natur- und Sozialwissenschaften umfassen (Chatfield et al. 2014). Insbesondere Definitionen von industriellen Akteur*innen legen den Schwerpunkt auf die

1 Ein Vergleich der Begriffe »Big Data«, »Social Media« und »Data Science« im Web of Science (wissenschaftliche Artikel) sowie bei Google Trends (Google-Suchbegriffe) indiziert, dass die drei Begriffe in den Jahren 2006–2008 wissenschaftlich breit auftauchen und teilweise erst einige Jahre später in Suchanfragen auf Google Prominenz erlangen (»Social Media« 2007/8; »Big Data« 2011/12; »Data Science« ab 2014) (vgl. auch Kane 2014).

Verknüpfung der computerwissenschaftlichen Komponente mit der Lösung konkreter Business-Probleme (Provost & Fawcett 2013). Sowohl wissenschaftliche Definitionen[2] als auch professionelle Selbstbeschreibungen[3] verweisen auf ein grundsätzlich weites Verständnis, unter dem verschiedene wissenschaftliche Disziplinen gefasst werden können. Nur wenige wissenschaftliche Disziplinen arbeiten nicht auf die eine oder andere Weise mit Daten oder computergestützten Auswertungsinstrumenten. Solche inklusiven Definitionen eröffneten vielfältige Möglichkeiten für verschiedene wissenschaftliche Disziplinen, sich fortan unter dem Label »Data Science« zu präsentieren. Als »umbrella term« (Giabbanelli & Mago 2016: 1970; Irizarry 2020) erlaubt es das Label anderen Akteur*innen und Disziplinen, sich affirmativ (oder auch distanzierend) darauf zu beziehen. »Umbrella terms« operieren als Mediatoren zwischen wissenschaftlicher Forschung, Wissenschaftspolitik und industriellen Anwendungsfeldern (Rip & Voss 2013): Durch einheitliche, inklusive Bezeichnungen werden entstehende Wissensfelder gefestigt und mit ökonomischen sowie wissenschaftspolitischen Ansprüchen verknüpft, was deren Identität sowohl nach innen wie nach aussen prägt. Etabliert sich dann ein Wissensfeld, werden die transversalen Verknüpfungen durch Forschungsfinanzierung und Infrastrukturen stabilisiert und erhalten einen dauerhaften Charakter (ebd.: 11). Eine solche konsensbefördernde »Politik von Buzzwords« (Bensaude-Vincent 2014) ist auch im Falle der Datenwissenschaften an der Schnittstelle von Technikwissenschaften, Industrie und Wissenschaftspolitik zu beobachten.

Eine begriffshistorische Perspektive auf den Gegenstand macht also deutlich, dass sowohl Akteur*innen in der Wissenschaft, in der Wissenschaftspolitik und der IT-Industrie in die Genese der Datenwissenschaften involviert sind. Die Durchsetzung des breiten, inklusiven Labels »Data Science« markiert dabei ein integrierendes Moment zwischen den unterschiedlichen Akteur*innen der genannten gesellschaftlichen Sphären.

3.1.2 Die Datenwissenschaften zwischen disziplinärer Differenzierung und Entdifferenzierung

Eine zweite Perspektive interpretiert das Phänomen – vor allem in der Wissenschaft selbst – in einer disziplinären Lesart. Zahlreich sind die Bezüge, die »Data Science« als neue Disziplin rahmen (Cao 2017; NASEM 2017; Song & Zhu 2017). Die disziplinäre Perspektive wird besonders virulent in der Auseinandersetzung über die Ursprünge, die verschiedene ›Kerndisziplinen‹ der Datenwissenschaften für sich in Anspruch nehmen. Womöglich die grösste epistemologische, aber auch politische Herausforderung stellt das als »Data Science« bezeichnete Wissensfeld für die Statistik dar, indem dieses gewissermassen Anspruch auf eine traditionelle Domäne des Feldes, nämlich die Entwicklung von Methoden zur Analyse von (grossen) Datensätzen, erhebt (Grommé et al. 2018). In Reaktion darauf kam es zu einer Flut an Publikationen, die sich mit

2 »Data science is the discipline of drawing conclusions from data using computation« (Adhikari & DeNero 2016).

3 »›Data Science‹ means the scientific study of the creation, validation and transformation of data to create meaning«, vgl. Data Science Association (2020): Code of Conduct. Online: www.datascience-assn.org/code-of-conduct.html (Zugriff: 03.02.2022).

dem Verhältnis von Statistik und Datenwissenschaften beschäftigen.[4] Einige Kommentator*innen argumentieren, dass »Data Science« lediglich eine Art Re-Branding von Statistik sei (Donoho 2017; Yu 2014; Wu 1997). So hielt der Statistiker Jeff Wu bereits 1997 einen Vortrag mit dem Titel »Statistics = Data Science?«, in dem er Statistik als eine Trilogie von Datenerhebung, Datenmodellierung und Analyse sowie *decision making* beschreibt. In seiner Schlussfolgerung fordert er – wie Cleveland (2001) – dazu auf, fortan die Begriffe »Data Science« anstelle von Statistik bzw. »Data Scientists« für Statistiker*innen zu verwenden (Wu 1997).

Ähnlich argumentiert David Donoho, wonach der »Hype« um Big Data, die erforderlichen Kompetenzen und die vielen neuen Jobs für Data Scientists die disziplinären Grundlagen der Statistik, auf denen die Datenwissenschaften aufbauten, verdecken würden (Donoho 2017: 745–49). Er bezieht sich dabei auf den Statistiker Leo Breiman (2001), der die Auffassung zweier unterschiedlicher Statistik-Kulturen (*Data Modeling Culture* vs. *Algorithmic Modeling Culture*) vertrat.[5] Unter »Generative Modeling«, das der »Data Modeling Culture« entspreche, kann nach Breiman ein Grossteil der mathematischen Statistik, Ökonometrie und quantitativen Sozialforschung subsumiert werden. In dieser epistemischen Tradition wird für einen gegebenen Datensatz ein stochastisches Modell entwickelt, zu welchem die vorhandenen Daten ›passen‹. Auf dieser Grundlage werden dann Schlussfolgerungen über die zugrundeliegende Struktur des Modells gezogen (Inferenzstatistik). »Predictive Modeling«, das Breiman als Teil einer »Algorithmic Modeling Culture« sieht, umfasst hingegen Felder wie angewandte bzw. industrielle Statistik oder Computerwissenschaften und verschreibt sich primär der Entwicklung von Algorithmen. Das Modell testet die Vorhersagekraft (*prediction*) für einen gegebenen Datensatz, ohne Annahmen über dessen Entstehung zu machen (Donoho 2017: 751; Hofmann & VanderPlas 2017: 776).[6]

Trotz dieser divergierenden Positionsbezüge wird den Datenwissenschaften von vielen Autor*innen das Potenzial zugeschrieben, die unterschiedlichen Methoden, Datenverständnisse und Denkweisen zu integrieren und disziplinäre Grenzziehungen zu überschreiten.[7] Konträr dazu wird in einer konflikttheoretischen Perspektive die Machtverschiebung über soziale und kulturelle Fragestellungen von den Geistes- und Sozialwissenschaften zu den Computerwissenschaften und Engineering beklagt (Roberge & Seyfert 2016: 10). Manche warnen gar vor einer Kolonisierung sozialwissenschaftlicher Fragestellungen durch ingenieur- bzw. computerwissenschaftliche Disziplinen (Dagiral & Parasie 2017; McFarland et al. 2016). Barocas und boyd (2017: 25)

4 Vgl. die Beiträge im *Journal of Computational and Graphical Statistics 26(4), 2017*, oder das Special Issue on Statistics and the Undergraduate Curriculum in *The American Statistician 69, 2015*.

5 Die Diagnose zweier »Statistik-Kulturen« weist darauf hin, dass es nicht nur um bestimmte Modelle, Messinstrumente oder Analysetools geht, sondern auch um divergierende Denktraditionen und Weltverständnisse im Sinne epistemischer Kulturen (Knorr-Cetina 2002).

6 DiMaggio (2015) und Manovich (2015: 22) entwickeln ähnliche Argumente zum Verhältnis von sozialwissenschaftlicher (Inferenz-)Statistik und Computer- bzw. Datenwissenschaften.

7 Auch sozialwissenschaftliche Autoren argumentieren, dass bestehende Grenzziehungen zwischen wissenschaftlichen Disziplinen und Denktraditionen aufgelöst (DiMaggio 2015: 1) und fruchtbare Kollaborationen zwischen diesen entstehen würden (Salganik 2017). Kauermann (2018: 88) bezeichnet Data Scientists als »Zwitter«, die beide von Breiman skizzierten statistischen Kulturen beherrschen und somit verknüpfen könnten.

verweisen auf die Unterschiedlichkeit von Computer- und Datenwissenschaften auf der einen Seite sowie Geistes- und Sozialwissenschaften auf der anderen Seite bezüglich der Diskussion von ethischen Fragestellungen und der gesellschaftlichen Implikationen, die sich durch zeitgenössische Verwendungsweisen und Methoden in »Data Science« oder »Machine Learning« ergeben würden (ähnlich Wallach 2018).

Gegenüber dieser Perspektive disziplinärer Differenzierung sind jene Beobachter*innen zahlreich, die in »Data Science« weniger die Entstehung einer neuen wissenschaftlichen Disziplin als vielmehr eine grundlegende Neuordnung des Verhältnisses von Wissenschaft, Technologieentwicklung und Ökonomie diagnostizieren: In Verknüpfung mit Big Data als eigentlicher »Revolution« (Mayer-Schoenberger & Cukier 2013; McAfee & Brynjolfsson 2012) entstehe damit eine fundamental neue Wahrnehmung von Gesellschaft, eine neue »Kultur« (Barlow 2013).[8] Hey et al. (2009) erkennen in den veränderten Bedingungen wissenschaftlicher Erkenntnisproduktion mit Bezug auf den Informatiker und Turing-Preisträger Jim Gray ein »viertes Paradigma«: Es bezeichnet in Anlehnung an den Wissenschaftshistoriker Thomas S. Kuhn (1996) einen epochalen Wandel der Wissensproduktion, die auf Experimenten (experimentelle Wissenschaft), Modellen und Generalisierungen (theoretische Wissenschaft) oder Simulationen (computergestützte Wissenschaft) basiert, hin zu einer »explorativen Wissenschaft« (Kitchin 2014: 3), die sich massgeblich auf grosse Datenmengen, immense Rechenkapazitäten und algorithmische Verfahren zu deren Auswertung abstützt (Gray 2009: xviii; Kitchin 2014).[9] Das Verständnis der Datenwissenschaften als neues »Forschungsparadigma« hat zudem Eingang in die Wissenschaftspolitik gefunden (ETH-Rat 2016a: 2).

Ebenfalls im Sinn dieser Entdifferenzierungsperspektive präsentiert Donoho (2017: 758f.) »Data Science« als eine Art ›Meta-Wissenschaft‹, die Sekundäranalysen für jegliche Formen wissenschaftlicher Daten und Forschungsprobleme erlaube. Gemäss Ribes (2019: 516; Ribes et al. 2019) positionieren sich die Datenwissenschaften als »Universalwissenschaft«, die nicht nur für die Wissenschaften selbst, sondern auch für die Wirtschaft, den Staat und andere soziale Felder fundamental sei. Daten spielten in jeder wissenschaftlichen Disziplin eine zentrale Rolle, mehr noch als komplexe Algorithmen und hohe Rechenkapazitäten. Durch diese Positionierung, die durch Akteur*innen in Industrie und Wissenschaftspolitik gestützt werde, würden die Datenwissenschaften gewissermassen von Beginn weg die Gefahr überwinden, zwischen formaler Mathematik und angewandtem Engineering auf den Status einer »Hilfswissenschaft« reduziert zu werden, wie es den Computerwissenschaften über Jahrzehnte hinweg drohte (Ribes et al. 2019: 296).

8 In ihrer ›populärwissenschaftlichen Variante‹ taucht das Argument in Form globaler Heilsversprechungen und technikutopischer Rhetoriken auf: Jede denkbare Herausforderung (wie die Rettung der Menschheit vor dem Klimawandel, globaler Armut, Hunger etc.) gilt als technisch lösbar. Morozov (2013) spricht in diesem Zusammenhang von einer »solutionistischen« Perspektive, die nicht nur geographische, kulturelle und ökonomische Kontexte ausblendet, sondern auch politische und soziale Machtverhältnisse negiert.

9 Ein Paradigma besteht laut Kuhn (1996) in einer akzeptierten Art, die Welt zu untersuchen und das Wissen einer bestimmten Disziplin zu einem gegebenen Zeitpunkt zusammenzufassen. Periodisch erscheint eine neue Art zu denken, die bestehende Theorien und Ansätze herausfordert, wie beispielsweise die Darwin'sche Evolutionstheorie. Es kommt zum Paradigmenwechsel, wenn die dominante wissenschaftliche Vorgehensweise nicht mehr in der Lage ist, ein bestimmtes Phänomen oder Schlüsselfragen zu beantworten und zu erklären.

Die kurze Rekapitulation einiger wichtiger Positionen der disziplinären Differenzierungsperspektive einerseits und universaler Entdifferenzierung andererseits macht deutlich, dass unter Beobachter*innen kaum Konsens über die wissenschaftliche bzw. gesellschaftliche Bedeutung der Datenwissenschaften auszumachen ist. Dennoch können die intensiv geführten Debatten als Beleg dafür gelesen werden, dass sich Akteur*innen in unterschiedlichen Disziplinen und Feldern in vielfältiger Weise mit dem Gegenstand auseinandersetzen. »Data Science« als Wissensgebiet prägt und strukturiert demnach über seinen »Buzzword«-Charakter hinaus die Interessen und Aufmerksamkeiten heterogener Akteur*innen in multiplen sozialen Feldern.

3.1.3 Die Genese der Datenwissenschaften in einer Professionalisierungsperspektive

Ein dritter Strang der Literatur diskutiert die Genese der Datenwissenschaften in einer Professionalisierungsperspektive (Avnoon 2021; Carter & Sholler 2016; Demchenko et al. 2016; Dorschel & Brandt 2021; H. D. Harris et al. 2013; Walker 2015).[10] Die Arbeiten referieren mehrheitlich auf das Professionsverständnis bei Abbott (1988), wonach Professionen ein interdependentes System bilden, das die Kontrolle über Wissen, Kompetenzen und Arbeitsinhalte ausübt. Kennzeichnend ist, dass Professionen generalistisches, abstraktes Wissen auf konkrete Anwendungsfälle applizieren (ebd.: 8).[11]

Die Entwicklung der Datenwissenschaften in den letzten zwei Jahrzehnten entspricht durchaus jener Professionalisierung mit eigenen Konferenzen, Zeitschriften, Lehrbüchern, Fachgesellschaften oder Grundsätzen der Selbstregulierung (Walker 2015), wie sie auch bei anderen, etablierten Professionen zu beobachten ist. Brandt (2016) zeigt in seiner Dissertation auf, wie Data Scientists allgemeine Wissensbestände in Form von Methoden und Technologien auf konkrete – wissenschaftliche, ökonomische, politische – Fragestellungen anwenden, und stellt dies in Beziehung zu etablierten Professionen wie Jurisprudenz oder Systembiologie. Er schlägt den Begriff einer »Denkgemeinschaft« (*thought community*) (ebd.: 3) vor: Daten-Nerds würden formale Ideen mit informellen Interpretationen kombinieren. Daraus resultiere eine Form von Geheimwissen bzw. Expertise (*arcane knowledge*). Durch diese Form der Improvisation gelingt es ihnen, inhaltliche Probleme durch Darstellungen miteinander zu verknüpfen und so bürokratische Kontrolle in den jeweiligen Arbeitskontexten zu unterbinden. Dorschel und Brandt (2021) sprechen von »Professionalisierung mittels Ambiguität«: Sie identifizieren »eine *soziale Logik der Ambiguität*«, die die Konstruktion von Data Scientists in Wissenschaft und Wirtschaft strukturiere und auf einer »Grammatik aus Differenz und basalem Konsens« aufbaue (Dorschel & Brandt 2021: 21; Hervorhebung im Original). Die Rahmungen als »multipolare Akademikerinnen«

10 Gleichzeitig wird aber bisweilen auch bestritten, dass es sich bei Data Scientists überhaupt um eine eigenständige Profession handle, sondern im Rahmen des »Hypes« um Big Data das Label Science für bereits existierende Praktiken der Datenerhebung, -auswertung und -analyse in Feldern wie Business Intelligence angeeignet und Letztere dadurch aufgewertet worden seien (Watson 2014).

11 Professionen befinden sich in einer Art konstanten Auseinandersetzung mit anderen Professionen um jurisdiktische Monopole, d. h. die Legitimität der jeweiligen Expertise auf einem bestimmten Gebiet. Der Begriff der Jurisdiktion bezeichnet die Verknüpfung zwischen einer Profession und ihren Arbeitsinhalten (Abbott 1988: 20).

(Wissenschaft) bzw. »Schnittstellenprofession« (Wirtschaft) einerseits sowie als »Grenzgängerin«, »die sich von Nerds abgrenzt, aber für Weltverbesserung einsetzt« (ebd.), andererseits markieren Mehrdeutigkeit als strategisches Mittel (Eisenberg 1984; Leitch & Davenport 2007) in der Genese und Etablierung der Profession.

Grommé et al. (2018) konzipieren das Verhältnis zwischen traditioneller staatlicher Statistik und den neuen Datenwissenschaften als einen professionellen Konflikt um die Produktion legitimen Wissens über den Staat. Die Genese eines neuen Gegenstandes (Big Data) als epistemisches Objekt intensivierte die Auseinandersetzungen über die legitimen Formen der Expertise. Grommé et al. argumentieren feldanalytisch, dass Data Scientists eine neue Fraktion und somit Differenzierung des Statistikfeldes darstellen, die die bestehende Hierarchie herausfordern und zu einer Rekonzeption der Rollenverständnisse und epistemischen Praktiken nationalstaatlicher Statistikinstitute bzw. Statistiker*innen geführt haben.

Die vielfältigen Definitions- und Bestimmungsversuche in akademischen Publikationen genauso wie in Selbstbeschreibungen (Bowne-Anderson 2018; Strachnyi 2017; Voulgaris 2014) indizieren demnach ein professionelles Feld, dessen Grenzen offen sind, was Anschlussmöglichkeiten an verschiedene wissenschaftliche Disziplinen und Wissensgebiete ermöglicht. Zu den verbindenden Elementen, die Kommunikation über disziplinäre Grenzen hinweg ermöglichen, zählen neben gemeinsamen Karrierewegen und epistemischen Objekten vor allem methodische Herangehensweisen: Fragestellungen aus wissenschaftlichen, ökonomischen oder politischen Zusammenhängen werden als Datenprobleme gerahmt, die mit einem ähnlichen Set an Methoden, Instrumenten und Tools bearbeitet werden können (Schutt & O'Neil 2013: 44). Eine zeitgenössische Sozialform hierfür sind Diskussionen, wie sie an »Data Science« Hackdays oder Meetups[12] geführt werden, in denen Problemstellungen feld- und themenübergreifend vor allem in Bezug auf die verwendeten Methoden diskutiert werden. Brandt (2016) hat dies ethnographisch im Hinblick auf Rollenbeschreibungen von Data Scientists bei der Genese der Profession untersucht. Dahl (2020) untersucht die performative Aushandlung beruflicher Identitäten von Data Scientists auf digitalen Plattformen anhand von Podcasts. In Anlehnung an Goffmans Konzept der Selbstdarstellung identifiziert Dahl drei Typen eines Data-Scientist-Selbst (*Coach*, *Business* und *Academic*), die auf die interne Differenzierung der Berufsgruppe sowie unterschiedliche Betätigungsfelder hinweisen.

Neue Studienangebote und die stark steigende Nachfrage transformieren sowohl die Verfügbarkeit von Data Scientists als auch deren Ausbildungsniveau. Die Lancierung neuer Studiengänge löst dabei etablierte Rekrutierungswege ab: So verfügt die »zweite Generation« von Data Scientists häufiger über Masterabschlüsse in »Data Science«, während die »erste Generation« breiter zusammengesetzt war und Professionelle technik-, natur- und sozialwissenschaftlicher Disziplinen umfasste. Befragungen von Data Scientists zeigen, dass die zunehmende Verfügbarkeit von Absolvent*innen erste Effekte auf das Ausbildungsniveau und die Saläre der US-amerikanischen »Data Science«-Community zeitigt (Burtch 2016, 2018).

Somit stehen professionssoziologische Diagnosen eines »Upgradings« (Dröge 2019: 23) bzw. »Upskillings« datenwissenschaftlicher Praxis, wonach Data Scientists legiti-

12 Dabei handelt es sich um informelle Treffen zu bestimmten Themen mit starkem Networking-Charakter.

me Expertise für das Feld der Datenarbeit beanspruchen können, einem »Downgrading« (ebd.: 24) bzw. »Deskilling« entgegen: Gehl (2015) vermutet, dass der ökonomische »Wert« der »zweiten Generation« von Data Scientists rasch sinken werde: Sobald die Hochschulen genügend Absolvent*innen in hoher Zahl produzierten, seien Data Scientists nicht mehr rar, würden billiger und entsprechend bald auch nicht mehr so »sexy« (ebd.: 422). Strategisches Ziel insbesondere von Technologieunternehmen sei es, das Wissen von Data Scientists zu entpersonalisieren und in Applikationen oder Algorithmen einzubauen, damit es unabhängig von der jeweiligen Person zur Verfügung steht. Solche Anwendungen könnten dann auch von Nicht-Expert*innen benutzt werden und seien entsprechend günstiger (ebd.: 423).

Ob »Buzzword«, »Umbrella Term«, Disziplin, Feld, Paradigma oder Profession: So vielfältig die begrifflichen, disziplinären und theoretischen Perspektiven auf die Emergenz der Datenwissenschaften sind, so konträr wirken die resultierenden Interpretationen. Was aus diesen divergierenden Deutungsangeboten übrig bleibt, scheint mir die Erkenntnis zu sein, dass die Datenwissenschaften ein multidimensionales Phänomen darstellen, das synchron in verschiedenen sozialen Sphären in je eigenen Ausprägungen konzeptualisiert, verhandelt, praktiziert und implementiert wird. Dabei bleiben das Label »Data Science«, die damit assoziierten Wissensbestände, Praktiken, Tools und Methoden sowie die daraus resultierenden Auswirkungen auf die Wissensproduktion die zentralen Bezugspunkte der unterschiedlichen Deutungen. Insofern sind die Interpretationen weniger als konkurrierend zu lesen, sondern vielmehr als Ausprägungen jener strukturellen Eigenschaft des Gegenstandes, nämlich selbst vielstimmig und mehrdeutig zu sein.

3.2 Konstruktionen der Datenwissenschaften im Arbeitsmarkt

> »Despite the recent sensational declaration of a data scientist as ›the sexiest job of the 21st century‹, however, there is a lack of published rigorous studies of what a data scientist is, and what job skills this hottest job title may require« (Chatfield et al. 2014: 1).

3.2.1 Qualifikations- und Kompetenzanforderungen in Stellenausschreibungen

Arbeiten, die die Entstehung der Datenwissenschaften im Arbeitsmarkt untersuchen, wenden im Wesentlichen zwei unterschiedliche Strategien an, um die relevanten Kompetenzen, Qualifikationen oder Tätigkeitsprofile von Data Scientists zu eruieren. Eine erste Strategie in der Forschungsliteratur besteht darin, die neuen Qualifikations- und Kompetenzanforderungen von Data Scientists durch Analysen von Stellenanzeigen empirisch herzuleiten (Dadzie et al. 2018; Debortoli et al. 2014; Djumalieva et al. 2018; Gardiner et al. 2018; Wowczko 2015). Im Fokus sind entstehende sowie sich verändernde Berufsfelder, vor allem im Bereich der Informations- und Kommunikationstechnologien sowie der digitalen Transformation von Wissensarbeit (dazu Boes et al. 2018; Dröge & Glauser 2019):[13] Bildungs- und Qualifikationsanforderungen

13 Die veränderten Anforderungen im schweizerischen Arbeitsmarkt (Sacchi et al. 2005; Salvisberg 2010) und spezifisch in Bezug auf die Digitalisierung (SECO 2017a; Sheldon 2020) sind seit längerem Gegenstand der sozialwissenschaftlichen Arbeitsmarktforschung.

sowie die nachgefragten Tools in diesen Berufsfeldern entwickeln sich derart dynamisch, dass sie kaum mehr zur Festlegung von Tätigkeitsprofilen verwendet werden können. Berufsbezeichnungen sind in diesem Sinne »Container« (Wowczko 2015: 39) für spezifische Kompetenzprofile, die sich kontinuierlich verändern. Die Analyse von digital verfügbaren Stelleninseraten bietet insofern eine einfache, gegenüber anderen Erhebungsmethoden (wie Befragungen) kostengünstige Möglichkeit, um die kurzfristige, stark volatile Nachfrage nach bestimmten Kompetenzen innerhalb sich transformierender Berufsfelder zu eruieren (Boselli et al. 2018). Im Folgenden fasse ich einige wichtige Ergebnisse dieser Forschungsliteratur im Hinblick darauf zusammen, wie Stellenanzeigen und Kompetenzdefinitionen die Datenwissenschaften beschreiben und konstruieren.

Chatfield et al. definieren sechs Kompetenzbereiche von Data Scientists als Schnittmenge von zwei Dutzend akademischen und industriellen Definitionen: »(1) Entrepreneurship and business domain knowledge, (2) Computer scientist, (3) Effective Communication skills, (4) Create valuable and actionable insights, (5) Inquisitive and curious, and (6) Statistics and modeling« (Chatfield et al. 2014: 7). Die Verknüpfung computerwissenschaftlicher und statistischer Fertigkeiten mit unternehmerischem Wissen sowie den dazugehörigen Softskills fördert sehr ähnliche Ergebnisse zutage wie die Analyse der Kompetenzanforderungen der untersuchten Curricula. Schumann et al. (2016) konstruieren Data Scientists als »Allrounder«, die vielfältige Fachkompetenzen, soziale Fähigkeiten und Selbstkompetenzen wie beispielsweise Kreativität besitzen. Je nach Anforderungen bezüglich Tätigkeit bzw. der Rolle innerhalb der stellenausschreibenden Organisation empfehlen sie die Bildung von unterschiedlichen »Kompetenzprofile[n] (Data-Scientist-Typen)«, wobei keine der drei Kompetenzdimensionen (Fach-, Sozial-, Selbstkompetenz) vernachlässigt werden soll (ebd.: 465).

Dadzie et al. (2018) eruieren durch die Analyse zehntausender Stellenanzeigen aus diversen europäischen Ländern Datenbanken, Statistik, Tools (vor allem NoSQL, Hadoop) sowie Erfahrungen bezüglich Datenvisualisierung (vor allem Tableau, D3.js) als die am häufigsten genannten Kompetenzanforderungen. Debortoli et al. (2014) zeigen, dass bei Inseraten mit dem Suchbegriff »Big Data« zwischen technischen und Business-Kompetenzen differenziert werden kann, wobei erstere in rund 70 % aller Anzeigen nachgefragt werden. Die am häufigsten genannten technischen Kompetenzen sind »NoSQL Databases«, »Software Engineering« und »Programming«, während »Machine Learning« und »Quantitative Analysis« weniger bedeutend sind (Debortoli et al. 2014). Gardiner et al. (2018) identifizieren in der Analyse von Stellenanzeigen mit »Big Data« im Titel ebenfalls starke technische Orientierungen, insbesondere im Management von Daten sowie bei analytischen Informationssystemen, und folgern daraus, dass die Anforderungen an Fertigkeiten und Wissen von Datenwissenschaftler*innen hauptsächlich aus der Informatik stammen (ebd.: 9). Eine Industriestudie von IBM und Burning Glass Technologies erschliesst durch Data Mining von 130 Millionen Stellenanzeigen 300 analytische Kompetenzen für datenbezogene Professionen, die anschliessend aufgrund von Ähnlichkeiten in den Kompetenzprofilen in sechs Gruppen eingeteilt werden (Markow et al. 2017). Für jeden Beruf wird sodann ein analytischer Wert berechnet, wobei Data Scientists die höchsten Werte erzielen würden:

> »Data Scientists are the most analytical roles in the market. They require proficiency with a large range of specialized analytical skills and tools, such as Machine Learning, Apache Hadoop, and Data Mining, in addition to generalized [Data Science and Analytics] skills like SQL, R, and Data Analysis« (Markow et al. 2017: 9).

Das Beispiel zeigt, dass viele der Kompetenzen, über welche Datenwissenschaftler*innen verfügen sollten, trotz solcher Kategorisierungen unscharf sind: So bezeichnet »Machine Learning« ein Kompendium statistischer Modelle und Algorithmen, mit deren Hilfe Computersysteme Aufgaben ohne spezifische Instruktionen ausführen können, sondern sich auf die Analyse von Mustern und Inferenzen aus bestehenden Daten stützen. In der thematischen Anwendung und Komplexität reichen diese von linearen Regressionsmodellen bis zu neuronalen Netzwerken, die divergierende Anforderungen an mathematische Wissensbestände, technisch-methodische sowie analytisch-interpretative Fähigkeiten stellen. »Machine Learning« kann folglich in verschiedenen Anwendungskontexten Unterschiedliches bedeuten (Engemann & Sudmann 2017). Der Umstand, dass notwendige methodisch-technische Kenntnisse oft summarisch (beispielsweise »Erfahrung in Machine Learning«, »Statistische Analysen in R«) repräsentiert werden (Ismail & Abidin 2016), indiziert demnach eine disziplinen- bzw. feldübergreifende Offenheit, die eine Vielzahl von Praktiker*innen ansprechen und potenziell inkludieren soll.[14] Solche Listen mehrdeutiger Begriffe stellen insofern strategische Positionierungen der rekrutierenden Organisationen dar.

Data Scientists werden in vielen der zitierten Studien als multidisziplinär orientierte Praktiker*innen imaginiert, die als sogenannte »data unicorns« (Hermida & Young 2017) über eine ganze Reihe verschiedener Kompetenzen verfügen sollen. In einer kultursoziologischen Perspektive werden Data Scientists sodann als ›Allesfresser‹ (*omnivores*) bezeichnet, die ihre professionelle Identität durch einen generalistischen, inklusiven Zugang zum Kompetenzerwerb konstruieren und aufrechterhalten würden (Avnoon 2021). Die grundlegende Distinktion gegenüber anderen Berufsgruppen erfolge gerade nicht durch Spezialisierung, sondern durch eine generalistische Haltung, die Expertise für unterschiedlichste Gegenstände beansprucht (ebd.: 345).

Da solche vielseitig ausgebildeten Generalist*innen jedoch auf dem Arbeitsmarkt nicht verfügbar waren, fand allmählich ein Übergang zur Wahrnehmung als Teamplayer statt und rückte die Zusammensetzung von »Data Science Teams« in den Vordergrund (Davenport 2020; H. D. Harris et al. 2013; Kim et al. 2018; Patil 2011; Schumann et al. 2016). Diese werden zudem als Lösung für den andauernden »Fachkräftemangel« in den Datenwissenschaften präsentiert (J. G. Harris et al. 2013). Aufgrund der anhaltenden Unsicherheit bezüglich der verfügbaren Kompetenzprofile streben vor allem Akteur*innen in der Geschäftswelt formale Klassifikationsschemata sowie Zertifizierungen der nachgewiesenen Kompetenzen von Data Scientists an (Davenport 2020), um diese gegenüber anderen Praktiker*innen, die ebenfalls Daten verarbeiten, abgrenzen zu können.

Unabhängig von der Frage, wie sich ein Profil von Data Scientists zusammensetzt und welche Kompetenzen dazu vonnöten sind, zeigt die Rekapitulation dieses ersten

14 Umgekehrt vermuten manche Autor*innen eine ›Aufpolierung‹ der Stellenanzeigen durch Unternehmen, weil konkrete Definitionen für das entstehende Berufsfeld noch weitgehend fehlen würden (Debortoli et al. 2014).

Teils von Arbeitsmarktstudien zu Datenwissenschaften, dass unterschiedliche Kombinationen von Skills charakteristisch sind. Qualifikations- und Kompetenzanforderungen werden durch offene Aufzählungen von zugeschriebenen Fähigkeiten, Bildungstiteln, Methoden oder Tools formuliert. In dieser Offenheit manifestieren sich einerseits Unsicherheiten über die zentralen Begrifflichkeiten des noch jungen Feldes. Auf der anderen Seite kennzeichnen die Listen bestimmte Kategorien, die sich bereits in einer frühen Phase etabliert haben. Die empirische Analyse der Stellenanzeigen wird zeigen, inwiefern und welche Begriffe sich als flexibel genug erweisen, um die multiplen Bedeutungen und Anforderungen an datenwissenschaftliche Praktiker*innen zu vereinen und zu repräsentieren.

3.2.2 Arbeitspraktiken von Data Scientists

Neben der Analyse von Stellenanzeigen besteht eine zweite Strategie der Forschungsliteratur darin, praktizierende Data Scientists und verwandte Professionen nach den erforderlichen Kompetenzen, Ausbildungs- und Karrierewegen sowie den epistemischen Praktiken im Arbeitsalltag zu befragen (EMC 2011; Feldman et al. 2017; H. D. Harris et al. 2013; Ismail & Abidin 2016; Kandel et al. 2012; Kim et al. 2018; Roberts & Roberts 2013; Swan & Brown 2008). Aus den Erkenntnissen werden oft Forderungen an die Strukturen und Inhalte von Ausbildungsprogrammen abgeleitet.

Eine Studie des Hard- und Softwareherstellers EMC (2011) untersucht Datenpraktiken, Tools und Ausbildungswege der »data science community« sowie die Frage, wie die Organisationen mit datengetriebener Problemlösung umgehen. Die Analyse von Surveydaten von knapp 500 Data Scientists und Business-Intelligence (BI)-Professionellen zeigt, dass sich Data Scientists von den BI-Professionellen insbesondere dadurch unterscheiden, dass sie öfter Programmierkenntnisse aufweisen und in der Lage sind, Experimente mit grossen Datenmengen durchzuführen.[15] Bezüglich der Ausbildung verfügen sie häufiger über Master- und PhD-Abschlüsse als die Vergleichsgruppe, bei der der Bachelor der häufigste Abschluss ist. Als zentral für die Praxis von Data Scientists werden zudem Interdisziplinarität und Arbeitsteilung mit anderen Professionellen in Unternehmen betrachtet (ebd.: 4).

Kandel et al. (2012) untersuchen, wie die Arbeitssituation und die verfügbaren Tools die Praktiken dreier Gruppen von Analyst*innen (»Business Analysts«, »Data Analysts« und »Data Scientists«) beeinflussen. Mittels qualitativer Interviews mit Praktiker*innen aus verschiedenen ökonomischen Feldern identifizieren sie drei »Archetypen«: »Hackers« sind erfahrene Programmierer*innen, verwenden die meisten Tools und operieren meist unabhängig von der IT. Sie führen die Schritte der Datenerhebung und -analyse selbstständig durch, erschliessen Datenquellen ausserhalb der Organisation und verknüpfen sie mit internen Daten. »Scripters« sind erfahren in Programmier- und Statistiksoftware wie R, erheben aber die Daten nicht selbst. Die Gruppe verwendet viel Arbeitszeit für ihre Modelle, die sie meist innerhalb einer Programmierumgebung erarbeitet. Die »Application Users« schliesslich verwenden hauptsächlich Excel-Tabellen und Software wie SAS oder SPSS. Sie benötigen Hilfe

15 Zu ähnlichen Schlussfolgerungen gelangen Debortoli et al. (2014: 325), die vergleichend Stellenanzeigen von BI- und Big-Data-Professionellen analysieren.

bei der Datenaufbereitung, arbeiten meist mit kleineren Datenmengen und selten ausserhalb der vertrauten Software (Kandel et al. 2012: 2918f.).

Roberts und Roberts (2013) unterscheiden anhand von Surveydaten (N = 300) vier Gruppen von Analytics- und Data-Science-Professionellen, die sie unter anderem aufgrund der aufgewendeten Arbeitszeit für verschiedene Analysetätigkeiten sowie zugeschriebenen Fähigkeiten zuordnen. Daraus leiten sie detaillierte Beschreibungen ab, wie die vier Gruppen von Analytics- und Data-Science-Professionellen rekrutiert, gemanagt und in Organisationen gehalten werden können (ebd.: 8ff.).

Harlan D. Harris et al. (2013) untersuchen, weshalb für unterschiedliche berufliche Tätigkeiten die gemeinsame Bezeichnung »Data Scientists« verwendet wird. Sie argumentieren, dass professionelle Bezeichnungen aus Statistik, Software- und Datenbank-Engineering sowie Natur- und Sozialwissenschaften wie im »Fleischwolf« zusammengeführt und durch relativ unklare Profile, Fähigkeiten und Kompetenzen abgelöst worden sind (ebd.: 3). Dies habe einerseits zusammen mit dem »Hype« um »Data Science« zu nicht einlösbaren Erwartungen geführt, andererseits zu einem unnötigen Aufwand bei der Rekrutierung von Mitarbeitenden.

Die Crowdsourcing-Plattform CrowdFlower (2015) (heute Figure Eight Inc.) befragte 153 Data Scientists, um deren Einbettung in Organisationen sowie den Beitrag zu deren Operationsweise zu verstehen. Interessant sind die Ergebnisse der zeitlichen Aufwendungen für die unterschiedlichen Tätigkeiten: Die zeitintensivste Tätigkeit besteht in der Aufbereitung und Bereinigung von sogenannten »messy data«, d. h. das Vorhandensein von Daten in schmutzigem und unordentlichem Zustand (Mützel et al. 2018: 115). Daraus resultieren sehr eingeschränkte zeitliche Kapazitäten für die beliebtesten Tätigkeiten, nämlich die »predictive analysis« sowie das »mining data for patterns« (CrowdFlower 2015: 7f.).

Die Feststellung, dass die ungeliebte Tätigkeit der Datenerhebung und -aufbereitung rund 80 % der Arbeitszeit einnimmt, wird in vielen Selbstbeschreibungen der *Community* geteilt (Schutt & O'Neil 2013; Wickham 2014; Wilson 2017). Wie Mützel et al. (2018) zeigen, bleiben solche Praktiken in der Diskussion oftmals auf der »Hinterbühne« verborgen, obwohl die darin implizierten Entscheidungen (der Datenerhebung, -bereinigung etc.) spätere Analyseergebnisse und somit organisationale Entscheidungen vorstrukturieren.

Carter und Sholler (2016) diagnostizieren anhand von semistrukturierten Interviews mit achtzehn Datenanalyst*innen in unterschiedlichen ökonomischen Feldern eine Diskrepanz zwischen der medial konstruierten Wahrnehmung von »Data Science« und den Alltagspraktiken von Data Scientists »on the ground«. Dies betrifft sowohl die Rolle von Theorie und traditionellen statistischen Methoden, die Bedeutung von Objektivität sowie den Zugang zu – öffentlichen oder proprietären – Datenquellen. Dabei zeigt sich, dass die Datenanalyst*innen ein reflexiveres und differenzierteres Verständnis der Herausforderungen von Datenpraktiken in zeitgenössischen Organisationen entwickeln, als dies die medialen Verlockungen und Verheissungen nahelegen (ebd.: 2317). Obwohl sie meist aus der Wissenschaft in industrielle Felder gewechselt sind mit der Motivation, rasch und unbürokratisch Zugang zu grossen Datenbeständen zu erhalten und diese bearbeiten zu können, werden neue Limitationen (insbesondere im Umgang und der Kommunikation mit Kund*innen) deutlich, die die Handlungsspielräume der Interviewten in verschiedener Hinsicht einschränken.

Feldman et al. (2017) untersuchen am Beispiel von Data Analytics die Personalrekrutierung auf Freelancing-Plattformen. Durch qualitative Interviews eruieren die Autor*innen, welche Tätigkeiten festangestellte Data Scientists an Freelancer*innen auslagern (würden) und welche Kompetenzen dazu notwendig sind.[16] Neben Programmierfähigkeiten sowie statistischen und mathematischen Skills erachten die Interviewten vor allem Fachexpertise (*domain knowledge*), Interdisziplinarität und Kommunikation als wichtig, da mit unterschiedlichen Professionen und Funktionen kollaboriert wird. Auch die Freelancer*innen werden vor allem für jene Tätigkeiten eingesetzt und hoch bewertet, die supplementär zu den Aufgaben der Data Scientists selbst sind, nämlich Dokumentation, Visualisierung oder Datenbank-Engineering (ebd.: 13f.).

Viele existierende Studien fokussieren primär Arbeitspraktiken und Kompetenzerfordernisse in industriellen Anwendungsfeldern der Datenwissenschaften. Allerdings setzte sich eine frühe empirische Arbeit, die im Auftrag des britischen *Joint Information Systems Committee* entstand (Swan & Brown 2008), mit den Berufsrollen und der Karriereentwicklung von Datenprofessionen in der Wissenschaft auseinander. Auch sollte der Bedarf entsprechender Kompetenzen und Fähigkeiten für die Forschungsgemeinschaft evaluiert werden. Mittels Interviews und Fokusgruppen mit verschiedenen Berufsgruppen, die im wissenschaftlichen Feld mit grossen Datensätzen arbeiten, entwickeln die Autorinnen verschiedene Professionsbezeichnungen und unterscheiden zwischen »data creators or data authors«, »data scientists«, »data managers« sowie »data librarians« (ebd.: 8). Bei den Data Scientists sind oft keine klaren Karrierewege erkennbar; zudem ziehen sie informelle berufliche Weiterbildung gegenüber formellen Ausbildungsprogrammen vor. Die Interviewten sind auch der Ansicht, dass »computational skills« alleine nicht ausreichten, um in interdisziplinär zusammengesetzten Forschungsgruppen mitarbeiten zu können; vielmehr sei dazu auch eine spezifische Fachexpertise (*domain knowledge*), z. B. in den Naturwissenschaften, notwendig (ebd.: 15).

Ein Jahrzehnt später untersuchen Geiger et al. (2018) Karrierewege und Zukunftsaussichten von Data Scientists an renommierten US-amerikanischen Forschungsuniversitäten. Sie zeigen, dass sich die Karrierewege und Zukunftsvorstellungen von akademischen Data Scientists teilweise stark von den angestammten disziplinären Werdegängen unterscheiden, indem etwa gewisse Praktiken (wie der Unterhalt von Forschungsinfrastruktur oder das Schreiben statistischer Pakete) hinzugekommen sind, die zwar von den Forschenden geschätzt werden, die allerdings von Berufungsgremien kaum honoriert würden (ebd.: 27ff.).[17]

Die Ergebnisse des zweiten Teils der diskutierten Forschungsliteratur zum Arbeitsmarkt der Datenwissenschaften deuten intensive Suchprozesse und Bestrebungen nach einer Taxonomie im Bereich der Datenprofessionen an, die Aufschluss, Überblick und Eingrenzung bieten soll über das breite Feld datenbezogener Praktiken, Berufsrollen und dazugehöriger Kompetenzen in zeitgenössischen Organisationen. Die

16 Bei den Freelancer*innen handelt sich oftmals um junge, hochqualifizierte Studierende mit noch wenig Berufserfahrung.

17 Metzler et al. (2016) legen eine ähnliche Analyse zu den Arbeitspraktiken und Herausforderungen von Sozialwissenschaftler*innen (Computational Social Scientists) vor, die mit grossen Datenmengen arbeiten.

Vielfalt der identifizierten Bezeichnungen indiziert, dass trotz des neuen Labels »Data Science« bzw. der entstehenden Profession der »Data Scientists« kaum Klarheit, geschweige denn Konsens darüber besteht, wie sich eine solche Gruppe zusammensetzt, was sie charakterisiert und was sie von anderen abgrenzt. Die empirische Analyse von Stellenanzeigen und Curricula wird zeigen, ob und inwiefern solche Taxonomien die Such- und Rekrutierungsstrategien von Organisationen im schweizerischen Arbeitsmarkt bzw. in der Hochschulbildung prägen.

3.3 Konstruktionen der Datenwissenschaften in der Bildungs- und Forschungspolitik

Bereits in den 2000er-Jahren beschäftigten sich forschungs- und wissenschaftspolitische Akteur*innen vermehrt mit der Transformation der Wissensproduktion infolge der exponentiellen Zunahme digital verfügbarer Daten (NSB 2005; IWGDD 2009; Swan & Brown 2008). Kennzeichnend für die Studien ist ein tendenziell ›pessimistischer‹ Grundton (Swan & Brown 2008: 18f.; IWGDD 2009: 18; Brown 2009), da Probleme bei der Anerkennung, mangelnde Ausbildungsmöglichkeiten sowie Karriereoptionen für Data Scientists innerhalb von wissenschaftlichen Forschungsgruppen diagnostiziert werden. Diese eher negativen Einschätzungen weichen innert weniger Jahre fast komplett einer (sehr) positiven Beurteilung: Fortan dominiert die Zuschreibung eines transformativen Potenzials und die Betonung der grossen Nachfrage nach ausgebildeten Fachkräften die Bewertung der Datenwissenschaften in der Wissenschafts- und Forschungspolitik (National Research Council 2014; ETH-Rat 2014; Getoor et al. 2016; Horvitz & Mitchell 2010; Labrinidis & Jagadish 2012). Damit einher geht nicht nur eine Perspektiven-, sondern auch eine Deutungsverschiebung, indem sich die dominanten Positionen im Diskurs zu »Data Science« vom wissenschaftlichen und wissenschaftspolitischen Feld hin zum technologischen und ökonomischen Feld – und dabei insbesondere den grossen Internetfirmen – verschieben.

Fortan richtet sich die wissenschaftspolitische Aufmerksamkeit auf die zukünftigen Potenziale der Datenwissenschaften zur Lösung zentraler wissenschaftlicher, politischer oder ökonomischer Herausforderungen. Zugleich bleibt ein negativer Gegenhorizont in den Publikationen bestehen, indem sie unter Verweis auf einen McKinsey-Bericht von 2011 einen gravierenden »Fachkräftemangel« diagnostizieren (Manyika et al. 2011) – alleine in den USA würden bis ins Jahr 2018 zwischen 140'000 und 190'000 Data Scientists fehlen.[18] Das fehlende Angebot an datenwissenschaftlichen Praktiker*innen würde demnach die Erkenntnisgewinne und Handlungsmöglichkeiten in verschiedenen sozialen Feldern gefährden. Diese Rahmung der Datenwissenschaften erwies sich als äusserst einflussreich und schrieb sich sowohl in die Aktivitäten ökonomischer Akteur*innen als auch in bildungs- und forschungspolitische Massnahmen ein (Saner 2019).

In der Folge widmeten sich zahlreiche Förderprogramme als auch universitäre Initiativen der Etablierung, Vernetzung und Priorisierung von disziplinen- und hoch-

18 Vorhersagen für die EU gingen sogar von 500'000 fehlenden Data Scientists aus, vgl. Offerman, Adrian (2016): »500'000 data scientists needed in European open research data«. Online: https://joinup.ec.europa.eu/collection/open-government/news/500000-data-scientists-need (Zugriff: 03.02.2022).

schulübergreifenden Gefässen der Datenwissenschaften in Forschung und Lehre. Die US-amerikanische *National Science Foundation* (NSF) lancierte verschiedene Instrumente zur Förderung der Datenwissenschaften: Dazu zählen unter anderem geographische Innovationssysteme, sogenannte Big Data Hubs (NSF 2015), oder das Schwerpunktprogramm »Harnessing the Data Revolution« (NSF 2019), das als langfristiger, strategischer Forschungsbereich an der Schnittstelle von Wissenschaft und Industrie etabliert wurde.[19] In der Schweiz wird unter Verweis auf das Primat ›liberaler Innovationspolitik‹ (Merz & Sormani 2016: 9) in der Regel keine Förderung bestimmter Technologien vorgenommen. Im Sinne einer kompetitiven Förderpolitik engagierten sich trotzdem eine Reihe wissenschaftspolitischer Akteur*innen in der Förderung der Datenwissenschaften: Der Schweizerische Nationalfonds (SNF) schrieb 2015 das Nationale Forschungsprogramm (NFP) Big Data (SNF 2015) aus. Parallel dazu lancierte der bundesstaatlich geförderte ETH-Bereich im Jahr 2016 die »Initiative for Data Science in Switzerland« (ETH-Rat 2016a), wozu die Gründung des Swiss Data Science Centers (SDSC) sowie neue Masterstudiengänge in Data Science an den ETH Lausanne und Zürich gehören (vgl. ausführlich zu den Massnahmen Kap. 6.4.2). Die bildungs- und forschungspolitischen Massnahmen erläutert und konkretisiert der Bericht »Herausforderungen der Digitalisierung für Bildung und Forschung in der Schweiz« (SBFI 2017). Die Datenwissenschaften werden darin als eine neue »Basiswissenschaft« präsentiert, »auf welcher andere Wissenschaften und Anwendungen aufbauen können« (ebd.: 70). Die bildungs- und forschungspolitischen Akteur*innen erhoffen sich zudem nicht nur technologisch-wissenschaftliche Innovationen, sondern auch ein gänzlich neues Verhältnis zwischen Hochschulen, Forschung und Industrie.

Des Weiteren beteiligen sich auch zahlreiche nichtstaatliche Akteur*innen an der Forschungsförderung im Bereich Datenwissenschaften: Die Gordon und Betty Moore Foundation sowie die Alfred P. Sloan Foundation schufen mit den »Moore-Sloan Data Science Environments« eine gemeinsame disziplinen- und organisationsübergreifende Struktur zur Förderung von Forschung und Ausbildung in Datenwissenschaften an drei renommierten US-amerikanischen Forschungsuniversitäten (University of California Berkeley, New York University und University of Washington-Seattle) (Moore-Sloan Data Science Environments 2018). Damit förderten die beiden Stiftungen nicht nur Forschung, Lehre und Beratung in datenintensiver Wissenschaft an den beteiligten Institutionen, sondern schufen eigentlich Rollenmodelle, die durch Veröffentlichung und Diffusion der erarbeiteten Inhalte und Prozesse an zahlreichen anderen Universitäten und Hochschulen aufgenommen und implementiert wurden (Katz 2019; Geiger et al. 2019). Ein weiteres Beispiel bildet die feldübergreifende Initiative »Social Science One« (King & Persily 2020): Damit lancierten Akteur*innen im akademischen Feld in Kollaboration mit Facebook, dem *Social Science Research Council* sowie mehreren gemeinnützigen Stiftungen ein neues Förder- und Begutachtungsmodell, das Daten von Facebook-Nutzer*innen unter Berücksichtigung von Datenschutzrichtlinien für sozialwissenschaftliche Forschung zugänglich macht. Die geförderten Projekte werden durch Stiftungsgelder finanziert, um die Unabhängigkeit gegenüber den Unternehmensinteressen von Facebook zu wahren.

Ein zweites Aktionsfeld ist die Förderung und Weiterentwicklung von Lehrprogrammen und Studiengängen im Bereich Datenwissenschaften. Die US-amerikani-

19 Für eine Übersicht zu staatlichen Förderaktivitäten in den Datenwissenschaften vgl. SBFI 2017: 94ff.

schen *National Academies of Sciences, Engineering and Medicine* (NASEM) beauftragten mehrere Arbeitsgruppen, die sich mit der Implementierung der Datenwissenschaften auf den unterschiedlichen Stufen der Hochschulbildung beschäftigten (NASEM 2017, 2018). Parallel dazu führten mehrere Divisionen der NASEM gemeinsam eine Serie von Roundtables durch (Kloefkorn et al. 2020), die Vertreter*innen von Universitäten, Fördereinrichtungen, Stiftungen und der Industrie zusammenbrachten, um Ausbildungs- und Praxismodelle sowie die Bedürfnisse der »Data Science«-Community und von Unternehmen als Arbeitgeber*innen zu diskutieren (NASEM 2020).

Auch in der europäischen Forschungspolitik wurden zahlreiche Initiativen lanciert, um Studiengänge in Datenwissenschaften zu fördern und die Eigenschaften des entstehenden Feldes schärfer zu konturieren. Das im Rahmen von Horizon 2020 geförderte Projekt EDISON[20] setzt sich zum Ziel, durch eine koordinierte Zusammenarbeit von Forschungspolitik, Wissenschaft und Industrie die Datenwissenschaften als Profession zu etablieren, die Bedürfnisse des Arbeitsmarktes mit den verfügbaren Skills in Übereinstimmung zu bringen und die Curricula an Hochschulen anzupassen. Damit sollen die Zahl ausgebildeter Data Scientists in Europa signifikant erhöht und deren Kompetenzen sowie die Ausbildungsqualität verbessert werden (Demchenko et al. 2016). Ähnliche Ziele verfolgt die ebenfalls durch Horizon 2020 geförderte *European Data Science Academy* (Mikroyannidis et al. 2018).

Sowohl die wissenschaftspolitischen Fördermassnahmen als auch die Vorbildfunktion etablierter Forschungs- und Ausbildungszentren unterstützten die Diffusion der Datenwissenschaften als Wissensfeld und führten zur Etablierung unzähliger Studienprogramme an Universitäten und Hochschulen weltweit (Anderson et al. 2014; Asamoah et al. 2015; Buckingham Shum et al. 2013; Giabbanelli & Mago 2016). Dies trug einerseits zur Professionalisierung bestehender Methodenausbildungen hinsichtlich der gelehrten Kompetenzen bei und andererseits zu einer klareren Grenzziehung gegenüber Nachbardisziplinen wie Computerwissenschaften, Software Engineering, Mathematik oder Statistik.

Der Literaturüberblick macht deutlich, dass Akteur*innen der Bildungs- und Forschungspolitik die Diagnosen eines Mangels an Data Scientists sowie damit einhergehende Forderungen – sowohl aus dem Feld der Ökonomie als auch der Wissenschaft – nach einem verstärkten Aufbau von Kompetenzen in den Datenwissenschaften aufgenommen und in ihre Förderpolitiken integriert haben. Sie schreiben den Datenwissenschaften hohe Potenziale zu und artikulieren kollektive Visionen und Entwicklungsverläufe des neuen Wissensgebiets. In der qualitativen Inhaltsanalyse von bildungs- und forschungspolitischen Strategiedokumenten wird zu überprüfen sein, welche Zukunftsszenarien der Datenwissenschaften entworfen werden und inwiefern sie die Entwicklung des entstehenden Wissensfeldes strukturieren. Ferner sind die begleitenden Massnahmen und Investitionen von Interesse, durch welche den Datenwissenschaften gegenüber anderen Wissensgebieten prioritäre Förderung zukommt.

20 Das Akronym steht für »Education for Data Intensive Science to Open New science frontiers«. Für eine Projektbeschreibung vgl. online: https://edison-project.eu/edison/edison-project/ (Zugriff: 03.02.2022).

3.4 Konstruktionen der Datenwissenschaften im akademischen Feld

3.4.1 Curricula als Forschungsgegenstand

Curricula bilden ein zentrales Forschungsgebiet der Bildungs- und Hochschulforschung (Pinar et al. 1995; Brüsemeister 2008). Neue bzw. veränderte Inhalte und Lehrformate in Curricula gehen oft mit intensiven Debatten über die Positionierung, Rolle und Ziele der verantwortlichen Bildungsinstitutionen einher. Insofern artikulieren sich in einer konflikttheoretischen Perspektive in den Auseinandersetzungen um Curricula auch Konzeptionen sozialer Ordnung und damit gesellschaftliche Wertvorstellungen (Bernstein 1975; Binder 2002).

Im Feld der Hochschulbildung sind Curricula ein zentraler Mechanismus, mit der sich wissenschaftliche Disziplinen reproduzieren (Holley 2009). Sie repräsentieren für die disziplinäre Gemeinschaft zentrale Werte und Normen und schaffen damit die Bedingungen für Forschungs- und Lehrpraktiken. Curricula sind viel mehr als nur die Summe der disziplinären Wissensbestände; sie umfassen auch soziale Prozesse und Relationen zwischen verschiedenen Akteur*innen wie Dozierenden, Studierenden, den Universitäten und Hochschulen als Organisationen sowie verschiedenen externen Entitäten wie professionellen Fachgesellschaften, politischen Behörden, Akkreditierungsagenturen und Akteur*innen im Feld der Ökonomie; Letztere motivieren insbesondere die Nachfrage nach Absolvent*innen mit bestimmten Kompetenzprofilen. Entsprechend existiert eine Vielzahl wissenschaftsinterner und -externer Faktoren, die die Planung, Etablierung oder Anpassung von Curricula rahmen und beeinflussen (Knight et al. 2013; Lattuca & Stark 2009).

Die Herausbildung eines neuen Wissensfeldes wie der Datenwissenschaften bedingt insofern neben der Festlegung von Vermittlungsformaten, Lehrmethoden und Wissensinhalten auch die Positionierung gegenüber verwandten disziplinären Feldern, die über etablierte Curricula verfügen. Darüber hinaus formulieren sie stets auch Konzeptionen über die Rolle und Bedeutung des Wissensgebiets in der Gesellschaft: Indem bestimmte Protagonisten die Datenwissenschaften als unverzichtbar für die Bewältigung globaler Menschheitsprobleme deuten, setzen sie einen Definitionsanspruch weit über das Feld der Wissenschaft hinaus. Manche Curricula nehmen diese Positionierung auf und vermitteln sie an Studierende und prägen somit nicht nur deren Verständnis auf die Datenwissenschaften, sondern auch ihr Weltbild an sich. Dies zeigen insbesondere auch Debatten über die sozialen und epistemologischen Bedingungen der Wissensproduktion in den Datenwissenschaften (Bates et al. 2020; Kross et al. 2020).

Curricula bilden insofern einen interessanten Forschungsgegenstand, um einerseits zeitgenössische Vorstellungen über die gesellschaftliche Verortung eines Wissensfeldes und andererseits die Austauschbeziehungen zwischen verschiedenen Akteur*innen im Feld der Hochschulbildung selbst und jenen anderer Felder zu untersuchen.

3.4.2 Charakteristika und Herausforderungen interdisziplinärer Curricula

In den letzten Jahrzehnten ist ein signifikanter Anstieg an interdisziplinären Curricula auf unterschiedlichen Stufen des tertiären Bildungswesens zu beobachten (Brint et al. 2009; Holley 2009; Knight et al. 2013). Interdisziplinäre Curricula erheben den Anspruch, den limitierten Blick einer einzelnen Disziplin auf einen Gegenstand oder ein Forschungsgebiet zu überwinden und stattdessen multiple Perspektiven darauf zu eröffnen. Beispiele aus jüngerer Zeit sind Neuro-, Klima- oder Umweltwissenschaften, aber auch die diversen Ausprägungen der Cultural Studies in den Geistes- und Sozialwissenschaften (Brint et al. 2009). Insbesondere bei den Studiengängen mit natur- und technikwissenschaftlicher Ausrichtung handelt es sich oft um technologieintensive, forschungsorientierte Wissensgebiete, die meist auf Masterstufe angeboten werden, jedoch keine eigene departementale Entsprechung haben und sich eng an ausseruniversitäre Forschungsbereiche, insbesondere in der Industrie, anlehnen (Abbott 2005: 265; Brint et al. 2009). Daraus resultiert eine zunehmende Entkopplung von interdisziplinären Forschungsgebieten und organisationalen Einheiten des akademischen Feldes (wie Instituten und Fakultäten). Dies lässt sich empirisch auch am Beispiel der Datenwissenschaften beobachten (vgl. Kap. 8.5).

Interdisziplinäre Curricula entstehen in Räumen zwischen etablierten disziplinären Strukturen, in denen die existierenden institutionellen Einheiten kaum Kontrolle ausüben können (Lindvig et al. 2019). Dies schafft Opportunitäten für Lehrende und Lernende, interdisziplinäre Aktivitäten zu kreieren, die sich nicht in die existierenden disziplinären Strukturen einfügen müssen – und die sich, sofern sich eine entsprechende Nachfrage einstellen sollte, auch in hohe Profite feldspezifischen Kapitals transformieren lassen. Eine schwache Positionierung interdisziplinärer Aktivitäten kann insofern auch strategische Absicht sein, Räume zu schaffen bzw. offenzuhalten, damit gerade keine Institutionalisierung erfolgt (Eyal & Pok 2015).

Im organisationalen Alltag führen interdisziplinäre Curricula allerdings zu diversen Herausforderungen und Konflikten (Holley 2009: 242): Zum einen sind interdisziplinäre Curricula Teil von Universitäten, die grossmehrheitlich disziplinär, in Instituten und Fakultäten, organisiert sind. Verschiedene Disziplinen können demnach Expertise auf einem solchen Gebiet beanspruchen. Dabei droht gewissermassen die Übertragung existierender disziplinärer Grenzziehungen und Konflikte auf interdisziplinäre Gegenstände und Gefässe. Zum anderen sind die Lehrenden eines interdisziplinären Curriculums in der Regel in einem disziplinären Feld trainiert, d. h., es fehlen gemeinsame Werte und Normen, die eine spezifische Fachkultur mit entsprechend sozialisierten Praktiker*innen begründen.

Neben der institutionellen Verankerung, Grösse und Struktur der lokalen organisationalen Einheiten (Brint et al. 2009; Merz & Sormani 2016; Small 1999) beeinflussen auch historische Besonderheiten, beispielsweise bei der Lancierung des Programms, die Ausgestaltung und das Ausmass an Interdisziplinarität massgeblich (Holley 2009; Lindvig et al. 2019). Augsburg und Henry (2009) charakterisieren interdisziplinäre Curricula als ein Kontinuum zwischen ›starken‹ und ›schwachen‹ Programmen: Erstere sind hochgradig strukturiert und weisen einen hohen Anteil verpflichtender Studienleistungen im interdisziplinären Bereich eines Programms auf. Dies geht oft mit einer höheren Anzahl Lehrender einher, deren Denomination im Kernbereich verortet ist. Schwache Programme hingegen bieten mehr Wahlmöglichkeiten ausserhalb des

definierten Kernbereichs und übertragen somit die Integration der involvierten Wissensbereiche und Forschungsperspektiven auf die Studierenden selbst.

3.4.3 Curricula in den Datenwissenschaften

Zeitgenössische Curricula in den Datenwissenschaften sind meist nach dem »Baukastenprinzip« organisiert und gemäss aktuellen berufs- und wirtschaftspädagogischen Gesichtspunkten erarbeitet werden (Schumann et al. 2016). Gemeinhin sind Curricula in Datenwissenschaften derart strukturiert, dass ein interdisziplinärer »Kern« in den Computerwissenschaften, Statistik, Mathematik und Engineering verortet wird, der sich auf Methoden, Technologien und andere Formalisierungen abstützt. Bisweilen werden zusätzliche Disziplinen wie Ökonomie, Semantik, Linguistik, Business, Design oder Visualisierung zum erweiterten Kern gezählt (Chatfield et al. 2014). Auch kommunikative und soziale Kompetenzen werden in empirischen Studien wiederkehrend als bedeutend identifiziert (Feldman et al. 2017; Ismail & Abidin 2016; Schumann et al. 2016).

Der nachfolgende Forschungsstand zu Curricula in den Datenwissenschaften umfasst äusserst heterogene Perspektiven unterschiedlicher Akteur*innen und ist in drei Teilen strukturiert. Erstens werden in einer disziplinenorientierten Perspektive – wie bereits gezeigt – normative Anforderungen an solche Programme formuliert (Cleveland 2001; Donoho 2017; Song & Zhu 2017; De Veaux et al. 2017). So versuchen primär hochschulpolitische Akteur*innen, datenwissenschaftliche Kompetenzprofile für unterschiedliche Bildungsstufen zu identifizieren und zu definieren (BHEF 2016; ETH-Rat 2016a; NASEM 2018). Parallel dazu skizzieren Praktiker*innen aufgrund eigener Forschungs- und Lehrerfahrungen sowie bestehender Ausbildungsgänge normativ die Anforderungen an solche Curricula (Cleveland 2001; Donoho 2017; Gupta et al. 2015; Kane 2014; De Veaux et al. 2017). In den USA haben sich verschiedene wissenschaftliche Institutionen intensiv mit den Herausforderungen der epistemologischen Transformation von Wissenschaft und der Rolle der Datenwissenschaften auseinandergesetzt (Berman et al. 2016; Kloefkorn et al. 2020; NASEM 2017, 2018; NASEM & The Royal Society 2018). Eine Expert*innenkommission der NASEM fasst zehn verschiedene Kompetenzbereiche als »Data Acumen«[21] (was mit »Datenscharfsinn« oder »Datenverstand« übersetzt werden kann) für die Konzeption datenwissenschaftlicher Studiengänge auf *Undergraduate*-Stufe zusammen (NASEM 2018: 2–8). Das bereits erwähnte Projekt EDISON entwickelte ein Kompetenzprofil sowie ein Modell-Curriculum für die Datenwissenschaften. Damit wurde somit eine Art »Landkarte« der unterschiedlichen Berufsprofile der entstehenden Data Economy formuliert, die zu einer europaweiten Vereinheitlichung und Zusammenführung von universitärem Angebot und industrieller Nachfrage nach datenwissenschaftlichen Kompetenzen beitragen sollten (Demchenko et al. 2016).[22]

---

21 Es handelt sich um die folgenden zehn Kompetenzbereiche: »Mathematical foundations, Computational foundations, Statistical foundations, Data management and curation, Data description and visualization, Data modeling and assessment, Workflow and reproducibility, Communication and teamwork, Domain-specific considerations, and Ethical problem solving« (NASEM 2018: 2–8ff.).

22 EDISON entwirft die folgenden Kompetenzen als Bestandteil einer »Data Science Literacy«: »Statistical techniques […], Computational thinking and programming with data […], Programming languages

Unter solche normativen Anforderungen fallen auch Bemühungen, Curricula in den Datenwissenschaften zu ›demokratisieren‹ (Cornelissen 2018; Kross et al. 2020), d. h. ihre soziale Verortung und die gesellschaftlichen Implikationen datenwissenschaftlicher Praxis aktiv zum Thema zu machen (Green 2018). Dabei sind insbesondere zwei Themen virulent: Einerseits ist das Teilen von Inhalten, Daten, Methoden, Algorithmen etc. unter Losungen wie Open Science und Open Data von Beginn weg in die Praxis des neuen Wissensgebiets eingeschrieben. Dies wird besonders durch Kurse zur Reproduzierbarkeit datenwissenschaftlicher Forschung, beispielsweise als Teil des erwähnten »Data Acumen« (NASEM 2018: 2–8), gefördert. Andererseits sollen Gender, Race, Diversity und weitere Dimensionen von *Social Justice* in den Curricula und Praktiken der Datenwissenschaften Berücksichtigung finden (Berman et al. 2016, 2018; Berman & Bourne 2015; Duranton et al. 2020; Geiger et al. 2019; Rawlings-Goss 2018).

Zweitens schildern Fallstudien ihre Erfahrungen und *Best Practices* einzelner Studiengänge (Anderson et al. 2014; Borne et al. 2009; Kreuter et al. 2018; McNamara et al. 2017; Stockinger et al. 2016) oder Veranstaltungen (Asamoah et al. 2015; Kross et al. 2020; Schuff 2018). Eine der ersten Publikationen stammt von Borne et al. (2009), die die Etablierung des Undergraduate-Programms »Data and Computational Sciences« an der George Mason University beschreiben. Das Programm vereint datengetriebene und rechenintensive Praktiken mit Anwendungsfeldern der Naturwissenschaften. Die Studiengangbezeichnung ist Ausdruck davon, dass zum Zeitpunkt der Etablierung multiple Bezeichnungen für die epistemische Transformation der Wissensproduktion koexistierten. Entsprechend fokussiert das Programm stärker auf das Management und die Analyse wissenschaftlicher Daten (ebd.: 78f.), als dies spätere Angebote unter dem Label »Data Science« tun.

Der erste Studiengang, der explizit »Data Science« im Namen trägt, ist ein Programm am College of Charleston in den USA, das seit Mitte der 2000er-Jahre existiert (Anderson et al. 2014). Neben dem Umstand, dass es sich um ein Provinzcollege handelt, ist auch die Verortung auf Undergraduate-Stufe aussergewöhnlich. Solche wurden an den meisten Universitäten erst nach Erfahrungen mit (professionellen) Masterprogrammen implementiert. Inhaltlich fällt auf, dass von den fünfzehn Kursen, die den »Kern« des Studienprogramms bilden, lediglich deren drei in »Data Science« sind, gegenüber sieben Kursen in Informatik und acht in Mathematik (inkl. Statistik) (ebd.: 147). Dennoch skizziert das Programm eine curriculare Struktur, die später von diversen Hochschulen in ähnlicher Form aufgenommen wurde und sich auch in den Empfehlungen der National Academies (2018) wiederfindet.

Stockinger et al. (2016) schildern ihre Erfahrungen anhand eines Weiterbildungsprogramms in Datenwissenschaften (Diploma of Advanced Studies in Data Science) an der Zürcher Hochschule für angewandte Wissenschaften (ZHAW) in Winterthur. Das Curriculum legt den Schwerpunkt auf die Verknüpfung von theoretischen Wissensbeständen in Datenanalyse und Information Engineering mit praktischen Umsetzungen (Data Science Applications), um damit Berufstätige zu adressieren. Die Ergebnisse eines Surveys unter Absolvent*innen deuten die Heterogenität der beruflichen und feldspezifischen Verortung der Teilnehmenden an (ebd.: 73f.). Die Befrag-

and tools for data analysis […], Data visualization languages and tools […], Data Management« (Demchenko et al. 2016: 30).

ten erwarten positive Effekte der Weiterbildung auf ihre berufliche Stellung, was sich etwa darin manifestiert, dass sie »mehr Verantwortung bei analytischen, quantitativen und technischen Aufgaben« erwarten. Auch verwenden sie fortan häufiger die Bezeichnung »Data Scientist« anstelle anderer Stellentitel (ebd.: 77f.).

Kross et al. (2020) beschreiben die Einführung des MOOCS (Massive Open Online Course) »Data Science Specialization« der John Hopkins University. Die Ubiquität grosser Datenmengen und die postulierte »Demokratisierung« des Datenzugangs und der Verfügbarkeit entsprächen allerdings nicht der Verfügbarkeit der akademischen Ausbildungsmöglichkeiten. Aufgrund der grossen Nachfrage unter Studierenden habe das neue Programm, das auf existierenden Onlineangeboten aus dem Bereich Biostatistik basierte, zu einer Angleichung anderer Studienangebote (online und offline) an die curricularen Inhalte des MOOC geführt.

Schliesslich entwerfen Kreuter et al. (2018) ein Online-Mastercurriculum für Professionelle der Sozial- und Marktforschung, das Surveyforschung mit Methoden und Technologien der Datenwissenschaften verbinden soll. Zu den fünf Schwerpunkten des Programms zählen die Neudefinition von Forschungsfragen, Datenerhebung, Datenaufbereitung und -speicherung, Datenanalyse sowie die Produktion von Output und Zugänglichkeit der Daten (ebd.: 2). Das Programm beabsichtigt so eine datenwissenschaftliche Ergänzung existierender Curricula.

Drittens existieren schliesslich einige Studien, die die Strukturen und Inhalte von Curricula in den Datenwissenschaften, primär von Universitäten in den USA, empirisch untersuchen (Aasheim et al. 2014, 2015; Bukhari 2020; Ortiz-Repiso et al. 2018; Tang & Sae-Lim 2016). Aasheim et al. (2014) untersuchen die Implementierung von Business-Analytics-, Data-Analytics- und Data-Science-Programmen auf Undergraduate-Stufe. Zum damaligen Zeitpunkt boten lediglich 21 Universitäten im Sample Majors in den genannten Feldern an, davon lediglich fünf in Datenwissenschaften. Ein Vergleich der curricularen Strukturen und Inhalte der Studiengänge verweist auf signifikant ähnliche Verteilungen von Kursen und *Credit Hours* in den Bereichen Programmieren, Statistik, Mathematik, Data Analytics/Modeling sowie Data Mining. Differenzen existieren insbesondere in den Bereichen Datenbanken, Big Data und Visualisierung (ebd.: 17). Die Programme in Business und Data Analytics werden primär von Business Schools angeboten, während diejenigen in »Data Science« mehrheitlich in computerwissenschaftlichen Departements und Schools verortet sind.

Tang und Sae-Lim (2016) untersuchen Strukturen, Inhalte und Beschreibungen von dreissig Data-Science-Studiengängen an US-amerikanischen iSchools (vgl. auch Ortiz-Repiso et al. 2018). Trotz einer gewissen Heterogenität der Studiengänge hinsichtlich der involvierten Disziplinen zeigt die Inhaltsanalyse der Materialien, dass die Studiengänge mathematische Kursinhalte insbesondere in den Kernbestandteilen der Curricula gegenüber informationswissenschaftlichen sowie Visualisierungskompetenzen vernachlässigen würden.

Schliesslich analysiert auch Bukhari (2020) Strukturen und Inhalte von dreissig Studiengängen der Datenwissenschaften an US-amerikanischen Universitäten. Es zeigen sich signifikante Überschneidungen im curricularen Aufbau, in der Anzahl erforderlicher *Credit Hours* sowie der Bedeutung, die Praktika und Capstone-Kursen zukommt. Bezüglich der institutionellen Verortung in den Hochschulen sind Management und Business Schools führend, während Schools of Engineering, Sciences sowie Computer Science nachgelagert sind (ebd.: 6). Es dominieren Kursinhalte in Daten-

analyse und -management, (Business) Analytics, Statistik und Mathematik gegenüber solchen in Informatik, Programmierung, künstlicher Intelligenz und Machine Learning sowie Datenschutz und Ethik (ebd.: 8).

Für das deutsche Hochschulfeld geben Lübcke und Wannemacher (2018) sowie Heidrich et al. (2018) einen Überblick über existierende Studiengänge, Kursangebote und deren institutionelle Verortung. Die ab 2014 entwickelten Studiengänge sind überwiegend generalistisch angelegt und weisen nur selten eine fachliche Konzentration auf. Mehrheitlich handelt es sich um Masterstudiengänge an staatlichen Universitäten und Fachhochschulen, die auf vorangehenden Informatik- oder Mathematikstudiengängen auf Bachelorstufe aufbauen. Viele Studienangebote sind eher forschungsorientiert, was aufgrund der steigenden Nachfrage im Arbeitsmarkt als Nachteil interpretiert wird (Lübcke & Wannemacher 2018: 3). Bezüglich der in den Studiengängen und weiteren Ausbildungsformaten vermittelten Data Literacy orten Heidrich et al. (2018) eine hohe Diversität und unklare Definitionen der verwendeten Begrifflichkeiten. Obwohl in vielen Implementierungen ähnliche Ansätze verfolgt würden, resultiere durch die Vielfalt der Begriffe eine gewisse Ambiguität. Deshalb brauche es standardisierte Kompetenz-Frameworks für »Data Literacy« (wie EDISON, vgl. oben) auf den unterschiedlichen Stufen des Bildungssystems.

Die Diskussion existierender Studien indiziert eine Breite disziplinärer und institutioneller Implementierungen und Positionierungen von Curricula in den Datenwissenschaften im Feld der Hochschulbildung. Trotz der identifizierten Diversität, die im Kontext der Herausbildung des Wissensfeldes zu betrachten ist – der Grossteil der Studiengänge existiert seit weniger als fünf Jahren –, zeigen sich jedoch bereits Prozesse von Angleichung bzw. mimetischer Isomorphie (DiMaggio & Powell 1983: 151f.), die in strukturellen und inhaltlichen Ähnlichkeiten der Curricula resultieren. Die Forschung zu interdisziplinären Studiengängen verweist auf Herausforderungen in der Etablierung und Implementierung solcher Curricula. Hierbei ist insbesondere auch der Einfluss von curricularen Empfehlungen und anderer normativer Anforderungen an solche Programme zu prüfen. Die Etablierung von akademischen Studiengängen in Datenwissenschaften an Schweizer Universitäten und Hochschulen stellt insofern einen geeigneten empirischen Fall dar, um solche Prozesse von Diffusion und Kanonisierung, aber auch organisationaler Aushandlung und Koordination sowie die Herausforderungen, die die Genese interdisziplinärer Curricula markieren, exemplarisch zu untersuchen.

3.5 Zwischenstand

Der Forschungsstand hat die Bedeutung multipler Begriffsverständnisse, disziplinärer und theoretischer Perspektiven, von Qualifikations- und Kompetenzzuschreibungen sowie kollektiven Visionen und Entwicklungsszenarien bei der Konstruktion der Datenwissenschaften als entstehendes transversales Wissensfeld aufgezeigt. Nachfolgend fasse ich einige zentrale Erkenntnisse zusammen, um die Analyse der empirischen Untersuchungsfälle im bestehenden Forschungs- und Erkenntnisstand zu verorten.

Bei den Datenwissenschaften handelt es sich um einen empirischen Gegenstand, dessen Genese in verschiedenen sozialen Sphären synchron verortet werden kann. Im

Zentrum steht weniger die Frage, welche Disziplin oder welches soziale Feld das Primat bei der Genese der Datenwissenschaften für sich beanspruchen kann, sondern der Umstand, dass Akteur*innen aus verschiedenen sozialen Feldern darin involviert sind. Insofern sind die Datenwissenschaften ein transversales Phänomen: Als epistemisches Feld basieren sie auf einer Reihe von technologischen und wissenschaftlichen Innovationen (vor allem in Methoden, algorithmischen Systemen und leistungsfähiger Hardware), die in andere soziale Felder diffundiert sind. Sie stützen sich auf technologische Infrastrukturen, die wiederum durch Kollaborationen und Wissenstransfer von industriellen Akteur*innen und wissenschaftlicher Forschung entstanden sind. Trotz der wissenschaftlichen Konnotation des Begriffs »Data Science« lässt sich das entstehende Wissensgebiet nicht auf eine feldspezifische Handlungslogik, beispielsweise der Wissenschaft oder der Industrie, zurückführen, sondern muss als Mischform, als Hybrid betrachtet werden.

Die Auseinandersetzung mit verschiedenen Definitionen legt nahe, dass Interdisziplinarität und Multiperspektivität konstitutive Merkmale der Datenwissenschaften darstellen: Als Kernbereich werden gemeinhin Statistik, Mathematik und Informatik genannt, die sich je nach Ausrichtung mit weiteren Wissensfeldern bzw. Disziplinen (Engineering, Wirtschaftswissenschaften, Linguistik etc.) verknüpfen. Anwendungsfelder (*domains*) liegen in der Wissenschaft, in der Medizin, in ökonomischen Feldern, in der staatlichen Verwaltung, im Non-Profit-Bereich und in anderen sozialen Feldern. Analog zu anderen technischen Wissensformationen zirkulieren nicht nur Methoden, Tools und Wissensbestände zwischen den involvierten Feldern, sondern auch individuelle Akteur*innen können zwischen technischen Hochschulen, Forschungslabors und der Tech-Industrie, aber auch zwischen der staatlichen Verwaltung, internationalen Organisationen oder Think Tanks hin- und herwechseln (»*revolving doors*«) (Safavi et al. 2018; Beckert et al. 2008). Somit profitieren auch Organisationen als kollektive Akteur*innen davon, denn es eröffnen sich vielfältige Möglichkeiten zur Kooperation: Während Unternehmen ihre humanen und methodischen Ressourcen erweitern können, bietet sich Universitäten und Hochschulen die Gelegenheit, Innovationsfähigkeit und Praxisnähe unter Beweis zu stellen. Auch bildungs- und forschungspolitische Institutionen tätigen hohe Investitionen in die Datenwissenschaften. Schliesslich profitiert auch die Verwaltung von der Expertise, indem sie zur Grundlage staatlicher Politiken oder direkt in staatliche Strukturen inkorporiert wird (Bundesrat 2020).

Über die Grenzen, d. h. was noch dazugezählt werden kann und was nicht, herrscht tendenziell Unklarheit. Sie sind Gegenstand von Aushandlungsprozessen, die unterschiedliche Formen annehmen. In der Verhandlung der Grenzen der Datenwissenschaften, was noch dazugehört und was nicht, wird die fundierende Rolle von Kompetenzen und Begriffen erkennbar: So beschäftigt sich eine Vielzahl unterschiedlicher Akteur*innen mit der Frage nach den ›richtigen‹ Kompetenzen und Tätigkeitsprofilen. Charakteristisch für solche Diskussionen ist, dass umfangreiche Listen unterschiedlicher Methoden, Tools, Bildungsqualifikationen und individueller Eigenschaften angeführt und miteinander verglichen werden. Solche Aufzählungen verweisen einerseits auf manifeste Unsicherheiten über die relevanten Begrifflichkeiten, andererseits aber auch darauf, dass sich bestimmte Kategorien in den Datenwissenschaften etabliert haben, die in Stellenausschreibungen, hochschulpolitischen Strategiedokumenten oder Curricula Verwendung finden.

Schliesslich entwerfen Akteur*innen aus Wissenschaft, Ökonomie und Politik durch multiple Bedeutungszuschreibungen divergierende Perspektiven auf die Datenwissenschaften als einen Raum zwischen Feldern. Bedeutsam sind insbesondere Zuschreibungen von zukünftigen Potenzialen und Entwicklungsverläufen, durch die kollektive Visionen der gesellschaftlichen Nutzung des Wissensgebiets artikuliert werden (Jasanoff 2015; Mische 2014). So strukturiert etwa die Identifikation einer grossen »Nachfrage« nach Data Scientists in der Wirtschaft die Aktivitäten eines zu schaffenden »Angebots« und so die Erwartungen anderer Akteur*innen, namentlich von Universitäten sowie in der Bildungs- und Forschungspolitik.