sciendo

Swiss Journal of Sociology, 49 (3), 2023, 499–517

Claiming Universal Epistemic Authority – Relational Boundary Work and the Academic Institutionalization of Data Science

Bianca Prietl* and Stefanie Raible**

Abstract: This article studies the rise of academic data science in Germany, Austria and Switzerland. By focusing on the boundary work that accompanies this development, we try to understand current transformations in knowledge production within digital academia and beyond. Drawing on qualitative interviews with data science scholars, we identify five lines of demarcation in claiming universal epistemic authority. This boundary work is characterized by multiple tensions and varies depending upon context and counterpart, making it inherently relational.

Keywords: Data science, academic institutionalization, discursive boundary work, epistemic authority

Universelle epistemische Autorität – Relationale Grenzziehungen und akademische Institutionalisierung von Data Science

Zusammenfassung: Dieser Artikel untersucht die akademische Institutionalisierung von Data Science in Deutschland, Österreich und der Schweiz unter Fokussierung auf die damit verbundenen Grenzziehungen (boundary work). Auf Basis qualitativer Interviews mit Data Science-Professor:innen rekonstruieren wir fünf Demarkationslinien, mit Hilfe derer universelle epistemische Autorität beansprucht wird, und zeigen, wie diese Grenzziehungsarbeit von multiplen Spannungen durchzogen ist, kontextabhängig variiert, und so als inhärent relational zu verstehen ist.

Schlüsselwörter: Data Science, akademische Institutionalisierung, diskursive Grenzziehungen, epistemische Autorität

Revendiquer une autorité épistémique universelle – Le travail relationnel de délimitation et l'institutionnalisation académique de la data science

Résumé : Cet article étudie l'essor de la data science académique en Allemagne, en Autriche et en Suisse en se concentrant sur le travail de délimitation (boundary work) qui accompagne ce développement. En nous appuyant sur des entretiens avec des chercheur·e·s en data science, nous identifions cinq lignes de démarcation dans la revendication d'une autorité épistémique universelle. Ce travail de délimitation est caractérisé par de multiples tensions et varie fortement en fonction du contexte et de la contrepartie, ce qui le rend intrinsèquement relationnel.

Mots-clés : Data science, institutionnalisation universitaire, travail sur les frontières discursives, autorité épistémique

* University of Basel, Center for Gender Studies, CH-4051 Basel, Bianca.prietl@unibas.ch.

** Johannes Kepler University Linz, Institute of Sociology, A-4040 Linz, Stefanie.raible@jku.at.

Seismo

DOI 10.2478/sjs-2023-0024

1 Introduction

In recent years, data science has cropped up on the academic landscape with a flurry of newly created chairs, research centers and study programs, all indicative of the increasing academic institutionalization of data science (Lowrie 2017; Saner 2019; Ribes 2019; Ribes et al. 2019; Prietl and Raible *forthcoming*; Slota et al. 2020; Saner 2022). These developments also point to the ongoing professionalization of data science (Dorschel and Brandt 2021) and of algorithmic modes of knowledge production more generally, which are diffusing into more and more areas of society, academic and non-academic (Kitchin 2014; Houben and Prietl 2018; Beer 2019; Bonde Thylstrup et al. 2019; Beaulieu and Leonelli 2022), changing the modes of knowledge production and challenging existing structures of epistemic authority (Bartlett et al. 2018; Prietl 2019a; Kitchin 2022; Jarke et al. *forthcoming*). This article studies the rise of *academic* data science in Germany, Austria and Switzerland by focusing on the complex bundle of boundary work that accompanies this development, in order to gain a better understanding of the current transformations of knowledge production in digital academia and beyond.

Data science has been applied in non-academic contexts for quite some time. It has given rise to the so-called data analytics industry (Beer 2019) and data scientists as a new branch of tech professionals (Dorschel 2021). But data science is only just taking root in academia. Our own empirical research sheds some light on the structural implementation of data science at universities and universities of applied sciences in the so-called DACH region[1]: With a total of 92 study programs in Spring 2021 and 80 newly appointed chairs in data science (out of 146 openings for data science chairs advertised between 2015 and 2021), we find ample evidence for an academic institutionalization of data science in the three countries studied. We can further depict a strong temporal dynamic with a rapid acceleration in the number of open positions in the years observed (advertised chairs in 2015: n = 8; 2016: n = 17; 2017: n = 28; 2018: n = 26; 2019: n = 35; 2020: n = 32). The chairs and degree programs in data science are for the most part situated in university departments related to STEM (science, technology, engineering, mathematics), especially in the area of computer science (e. g. 75 out of 146 advertised chairs and 61 out of 92 study programs). This organizational affiliation is also reflected on a content level, with the majority of data science professors having a background in computer science, and data science study programs focusing on computer science skills and competencies.[2] Where data science is implemented with a domain-specific focus (such as "business analytics and data science"[3] or "bio data science"[4]), there

1 DACH region includes the German-speaking countries Germany, Austria and Switzerland.
2 This might also explain why the majority of data science chairs in the German-speaking countries we studied are currently headed by men.
3 Title of a data science chair at University of Graz (Austria).
4 Title of a master's degree program at University of Applied Sciences Wiener Neustadt (Austria).

is a strong penchant for data science being institutionalized in alliance with either economic (e. g. 11 out of 26 domain-specific degree programs) or bio and life sciences (e. g. 6 out of 26 domain-specific degree programs), while there is hardly any structural affiliation with social sciences (e. g. 2 out of 54 advertised chairs with a domain-specific focus). Albeit this penchant for certain domains, we also find a palpable claim to universality in our interviews with data science scholars and even more so in brochures for data science degree programs. Here, data science is time and again presented as providing a *toolkit* of cutting-edge algorithmic methods for analyzing (big) data sets. Those tools shall allow it to produce "better" answers to a broad variety of questions stemming from heterogeneous disciplines and areas of interest (such as biology, history or industrial businesses), which are referred to as "domains". Domains thereby designate other academic disciplines or non-academic fields such as industrial organizations or areas of political activity where data science methods are applied.

As has been noted for the data analytics industry (e. g. Beer 2019), our study also demonstrates a clear expansionist tendency in academic data science's claim to epistemic authority, as the discipline increasingly asserts its relevance for academia and society as a whole. Taking into account the growing general demand for algorithmic modes of knowledge production, along with the specific call for data science by science policy as well as industry actors (Saner 2019), better understanding the epistemological claims made in the name of data science emerges as a timely and topical undertaking.

To grasp these developments, this paper explores the multiple forms of boundary work performed in staking the territory of data science, as encountered in qualitative interviews with data science professors (see section 2). It reconstructs the central lines of discursively enacted demarcation to construct data science on a symbolic level as an academic endeavour in its own right. These boundaries serve to distinguish it from other, established, disciplines within academia as well as from industrial data analytics and everyday notions of data science, especially public hypes around big data (see section 3). Finally, we reflect upon these empirical findings to better understand the implications of the academic institutionalization of data science for established structures and modes of knowledge production (see section 4).

2 Analytical Perspectives

Adopting the conflict-theoretical concept of *boundary work* as proposed for studying the emergence of professions (Abbott 1988; 1995a; 1995b) as well as the demarcation of science (Gieryn 1983; 1994; 1999) in a discourse- and practice-theoretical approach (Paulitz 2012), this article employs a relational perspective in order to explore the *discursive practices of distinction* done by data science professors. By exploring these practices of boundary work, we aim to analyse the various discursive strategies for claiming epistemic authority and legitimacy for data science as a new, stand-alone academic endeavour. We analyse how data science is symbolically constituted within a system of established disciplines and in relation to established structures and modes of knowledge production (for the use of the concept of boundary work for the study of professions, science and knowledge, see Lamont and Molnár 2002, 177–181).

Following both Andrew Abbott and Thomas F. Gieryn we take an anti-essentialist view of data science. That means we forgo any assumptions of a core set of characteristics that make data science a profession, a scientific discipline or even something called "data science" itself. Instead, we apply a processual perspective on *how* data science is constituted, especially by means of discursively drawn (professional) boundaries. As Abbott has pointed out, professions regularly compete with each another to secure the "more or less exclusive right to dominate a particular area of work" (1995a, 551), in other words, to be solely or primarily responsible for solving a particular problem in and for society. Because it is rare for a single profession to hold a monopoly in this regard, the professional system remains fluid and in constant negotiation (Abbott 1988, 69–79), with demarcations as well as divisions of labour and professional cooperation patterns that change over time (Abbott 1995b, 872). In order for professions to solve the problems to which they lay claim, they need to develop a professional body of knowledge that enables and legitimizes their inquiry (Abbott 1988, 52). In most cases such a body consists of a rather abstract, formal knowledge system, whose administrators are to be found in the academic field, which is also why "the ability of a profession to sustain its jurisdictions lies partly in the power and prestige of its academic knowledge" (Abbott 1988, 53–54). The academic institutionalization of data science, hence, marks an important milestone in the professionalization of data science as a scholarly discipline – and as a new mode of knowledge production.

Whereas Abbott studies professional "turf wars" primarily on the structural level of actors, organizations, labour divisions and resource distribution, Gieryn proposes, while building on Abbott's earlier work, the concept of "boundary work" to focus on the *symbolic* struggle for *epistemic* authority, thereby highlighting the importance of cultural classifications and representations. Gieryn's own work centers around the question of how science becomes perceived as the sole producer of truth within an "intellectual ecosystem" (1983, 783), especially in contrast to other knowledge-producing fields such as art, religion or politics. Adopting the metaphor of cartography, he understands science as a specific territory on a cultural map that serves as a guide for members of society, especially those who decide upon the distribution of resources in the intellectual ecosystem, and shows them "where" verified knowledge is produced. This scientific terrain, however, is neither fixed in and of itself, nor is it stable over time; rather, it emerges as the result of its demarcation, as an effect of the boundary-drawing work of competing actors and/or organizations (Gieryn 1999). Put differently, science neither exists as an entity *a priori*, nor does it have any fixed characteristics in an essentialist sense. On the contrary, science represents a historically as well as locally specific phenomenon, for "[t]he boundaries of science are ambiguous, flexible, historically changing, contextually variable, internally inconsistent, and sometimes contested" (Gieryn 1983, 792). Gieryn consequently draws attention to the rhetorical processes in which scientific practices and actors are attributed certain properties and are distinguished from others in order to identify them as scientific.

Loosely referring to Foucault and Bourdieu, Gieryn (1994, 417) further stresses the connection between processes of boundary work and questions of power. This becomes clearest when he describes boundary work as a means in the struggle for "credibility, prestige, power, and material resources" (Gieryn 1994, 405), which is achieved through "social interest in claiming, expanding, protecting, monopolizing, usurping, denying, or restricting the cognitive authority of science" (Gieryn 1994, 405). For him, this power struggle takes place primarily at the symbolic level of cultural classifications and, thus, in interest-driven rhetorical negotiations, having nonetheless very material consequences. In order to succeed, scientists – whom Gieryn identifies as the prime actors in these "rhetorical games" (1994, 406) – have to draw on established cultural norms and classifications, and strive to connect new negotiations with previous negotiation outcomes.

Although not in the direct crosshairs of his focus, Gieryn does point out early on that the concept of boundary work can also be applied to study the negotiation of boundaries and the associated processes of constitution of territories *within* academia, for example in creating (sub)disciplines (1983, 79; for such an application of the concept of boundary work, see e.g. Paulitz et al. 2015). Tanja Paulitz (2012) takes up this notion in her studies of how engineering became constituted as a gendered discipline. Drawing on Foucault's reasoning on the power/knowledge nexus and Bourdieu's field theory, she extends Gieryn's concept of boundary work from a discourse- and practice-theoretical perspective, thus reframing rhetorical negotiations as *discursive practices of distinction* by actors socialized and competing within the academic field. While adhering to the incorporated norms and rules of the academic field, they also fight over the shape and form of these rules in order to position themselves favourably, especially with regards to the realm they claim epistemic authority for.

Following the analytical perspectives outlined above, we endeavour to grasp how data science scholars discursively claim epistemic authority for a certain set of problems and draw the line between their expertise and that of other – already established – actors in academia and beyond, and, in the process, contribute to the professionalization and academic institutionalization of data science. We understand boundary-making as constitutive for the academic field and its disciplines, and we grasp our object of inquiry as both processual in character and structurally entangled with myriad power relations (see also Prietl and Ziegler 2016).

3 Empirical Approach

Empirically, we draw on own empirical research[5], especially 19 in-depth semi-structured qualitative interviews with data science professors in Austria, Germany and Switzerland. Applying the strategy of theoretical sampling (Strauss and Corbin 1996), we collected data covering the categories of gender (with an over-representation of women, 5 out of 19 interviewees), university type (research universities or universities of applied science), (technoscientific and geographic) metropoles or peripheries, generalist or domain-specific data scientists (including dominant domains such as economics as well as "niche" domains like social science). With the exception of two junior professors (equivalent to assistant professors without tenure track), our interviewees held permanent positions. The chair-based system in the DACH region grants them a high level of job security, basic financial and personal resources as well as freedom of research. However, since the turn of the century higher education governance has introduced entrepreneurial elements, especially performance- and project-based funding, pressuring universities as well as scholars to compete for third-party funding and students (Houben 2022, 323–332).

In terms of content, the interviews centered around the interviewees' own characterization of academic data science[6] for which they can be seen as representatives due to their position as data science professor. The interviews lasted two hours on average. All interviews were conducted online via Zoom (for a methodological reflection, see Raible et al. 2023), transcribed verbatim,[7] and analysed with the help of MAXQDA, applying open and selective coding strategies (Strauss and Corbin 1996). Coding was guided by our research interest of better understanding the positioning, discursive constitution, and legitimization of data science as an academic endeavour. For the purpose of this paper, we focused our analysis on the boundaries made relevant by the interviewees when presenting their understanding of data science and doing data science.

As we did not encounter any systematic country-specific differences, we do not distinguish between the three countries in the following results.

5 For the financial support of the research project "The Politics of Data Science" we would like to thank the Dr. Hans Messer Foundation.

6 Our interview guideline contained questions about the interviewees' professional and disciplinary biography, their understanding of data science, doing research and teaching in data science, co-operation and academic networks, their stance on critiques towards data science and perceptions of the future of their discipline.

7 All interviews were conducted and transcribed in German; the quotations cited below were translated by the authors.

4 Empirical Results

Looking at how our interviewees presented themselves and data science in the interviews, several boundaries stand out that we interpret as *relational* as their content and form vary depending upon context and counterpart: First and foremost are the boundaries drawn between methods-driven and applied data science (4.1) as well as between data science and its constitutive disciplines, computer science, mathematics and statistics (4.2.), but also so-called domains (4.3.). Furthermore, we witnessed boundaries drawn between academic data science in contrast to industrial data analytics (4.4.) as well as distinctions drawn between data science and everyday notions surrounding data analytics, particularly the high-publicized promises of big data (4.5.).

4.1 Ambiguous Hierarchies Within Data Science: Methods-Driven Versus Applied Data Science

> *I would describe data science as trying to draw insights from what we hope is a large [laughs] data set and then interpreting them somehow. That's the main idea of what you might call applied data science. Then there's the more methods-driven approach that is mainly focused on developing new methods. I try to strike a healthy balance.* (IV_DE_13-2, Pos. 19)

At first glance, this interview quote from a data science professor describes data science as aiming to distil insights from large data sets. As noted above, data science is frequently presented as an analytical toolkit that offers a new approach to knowledge production in multiple domains by analysing (big) data sets (Slota et al. 2020; Saner 2022). At a second glance, however, a subtle tension becomes apparent between "applied" vs. "methodological" data science. This differentiation between data science that "solely" focuses on the application of data science methods, on the one hand, and a data science that is concerned with advancing those same methods, on the other hand, surfaces in many of our interviews and constitutes an *ambiguous hierarchy* within the emerging discipline of data science.

Especially those interviewees who positioned themselves primarily as applied data scientists were strongly invested in boundary work vis-à-vis their more methods-oriented colleagues. Elaborating further, the interviewee quoted above explained:

> *Because, apart from purely methodological data science, if you say you are developing data science methods, then maybe you can do that* at home in your little office. *But as soon as you want to do something with those methods, you need to actively enter the respective domain. And at that point, at the latest, you cannot avoid working together with the relevant experts and in an interdisciplinary context.* (IV_DE_13-2, Pos. 37; our emphasis)

Here, refining data science methods is associated with isolated, non-communicative work that is separate from relevant domains and experts. This description of methods-driven data science evokes the cliché of the computer nerd: technically accomplished but socially incompetent (cf. Turkle 1986). It also reveals deprecatory notions of doing science in the ivory tower. Applied data science, in contrast, appears to bring both – supposedly mutually exclusive – skill sets together, casting off undesirable notions of scientific work while retaining the claim of epistemic authority.

At the same time, however, some interviewees voiced their admiration for data scientists who work on new algorithms, thus, marking the distinction between "applied" and methods-driven data science a hierarchical one. One professor at a university of applied sciences described her position as follows:

> *I wouldn't – well, I'm definitely not – among the top researchers [smiles], the ones refining or tinkering with new algorithms, but rather in practical applications and communicating results.* (IV_DE_08, Pos. 88)

Besides, once again, associating practical application with communication, this scholar also links research excellence with the idea of improving the very toolkit that constitutes data science. This association of excellence and prestige with methods-driven data science can also be observed in other interviews. A data science professor who works in the domain of engineering makes a deeply personal argument out of this distinction:

> *I especially feel that this project goes beyond merely applying these approaches. When I do research, it's very unsatisfying to just pull something out of the drawer [clears throat] that I've used in another context, for example, and I now apply it to the specific problem at hand.* (IV_DE_05, Pos. 68)

With not being satisfied to "just" apply ready-to-use methods, working on improving the analytical methods of data science is here again positioned as superior to their mere application – also when it comes to one's own self-image as a data scientist.

Considering the hierarchical notion underpinning the distinction between applied and methodological data science, the boundary work done by scholars on the application side of the data science spectrum seems to be stemming from a symbolically subordinate position and with a rather (self-)defensive and justifying goal. Conversely, methods-driven data science emerges as symbolically prestigious. This distinction between methods-driven and applied data science resembles the well-documented theory versus practice-boundary in engineering that served as a flexible means to distinguish academic engineering from non-academic "tinkering", but also to distinguish between theoretical versus applied areas of engineering (for theory versus practice distinctions in science and engineering, see Paulitz 2012; Paulitz et al. 2015). Discursive constructions of data science therefore seem to build upon past traditions of the symbolic construction of engineering. At the same time the symbolic hierarchy remains ambiguous as the subordinated pol, i.e. practical application of data science, is of value in itself, not least in the context of the entrepreneurial university that calls for practical relevance of science.

4.2 One for All: Uni-Dimensional Disciplines Versus Integrative Data Science

Data science is not only structurally implemented at the intersection of mathematics, statistics and computer science (see with regards to curricula construction Saner 2019), but also described by our interviewees as the perfect synergy of these established disciplines. In their depictions, data science overcomes their respective one-sidedness by integrating the strong suits of its "parent disciplines", as well as additional interdisciplinary expertise, including social skills and domain knowledge. In drawing these distinctions, data science professors once again rely on the theory versus practice-boundary. This time, however, the distinction is manifested by drawing lines between science and the "real world", or between the technical and the social. Furthermore, while mathematics, statistics and computer science are associated with more theoretical (that is, science and technical) territory, data science is presented as integrating both – supposedly mutually exclusive – ends of the spectrum.

A smaller circle of data science professors, mainly those with a disciplinary background in mathematics or statistics, underscored the scientific and academic nature of data science – especially in contrast to "mere" computer science. Depicting data science degree courses as a breeding ground for future data scientists who would then embark on a career in academia, one interviewee argued for shaping data science in such a – scientific – way, suggesting that developing the discipline more toward computer science would not produce the talent pool that the discipline needs:

> *Because we train the young talents who go on to get their PhDs, and because I come from the field of statistics, I also have an interest in shaping data science to be more than just another word for computer science. My hope is that we can shape it in this way.* (IV_DE_06, Pos. 105)

Other interviewees, by contrast, stress the what has been called "real-world orientation" (Saner 2022) of data science. The distinction here is between data science and neighbouring disciplines, such as mathematics, computer science or statistics, which are said to be disinterested in the application of their knowledge and expertise. One data science professor specialized in biomedicine described potential students who would not be a good fit for the master's degree program in which he teaches and which is highly focused on data science methods:

> *Yes, let's say a pure computer scientist or mathematician who is not really enthusiastic about a specific application domain, I would actually advise them against it, because even at [a university, our anonymization], when you choose this degree program, you have to choose a specialization. And you*

> *should also want to do it, because otherwise why would you do data science?* (IV_DE_13-2, Pos. 83)

Whereas the "pure" mathematician or computer scientist is viewed as oblivious to the world, data science is defined precisely by openness to worldly topics. Hence, with computer science, mathematics and statistics being depicted as either too close or too distant to science and its methodological rigor, data science is presented as being both open to the "real" world yet scientific in nature.

The idea of worldly oblivion is often linked to a lack of social and communicative skills, when for instance describing "pure" mathematicians and computer scientists as less socially competent than data scientists. When asked to give an assessment about what data science entailed for her, another interviewee describes these skills as a feature of data science in contrast to computer science, the discipline that she herself studied:

> *And, of course, the basics that you bring with you from mathematics and computer science are important. But the social skills are particularly important for data science. I don't think they are that important for pure computer scientists. But for data science in particular. […] Because you have to communicate a lot. Right from the start. That means you sit down with the domain experts and you have to understand first, what the problems are that you actually want to solve with the approach?* (IV_DE_04, Pos. 41–43)

When compared to mathematics, computer science or statistics, data science is associated repeatedly with practical applications, a real-world perspective and social skills. The message being: data science can move beyond a single focus on scientific progress or technical knowledge by *also* succeeding in the real world and in the social realm. In contrast to its "parental disciplines" that are depicted as uni-dimensional, data science is portrayed as integrative, uniting technical and social skills, such as data science expertise and communication skills. By aligning the theory–practice boundary with the technical–social boundary, data science appears to "have it all". This boundary work also serves the goal of attacking the epistemic authority of those disciplines that have so far held jurisdiction over high-level quantitative data analysis.

4.3 The Great Integrator: Isolated Domains Versus Transversal Data Science

Data science is not only defined by drawing boundaries to established disciplines such as mathematics, statistics and computer science. It is also characterized by its relation to what our interviewees and the literature refer to as "domains" (Ribes et al. 2019). As stated before, domains may designate other academic disciplines or non-academic areas where data science methods are applied. Domains supply the lines of inquiry to be pursued using data science methods, along with the necessary data to do so. Thus, domains are not *part* of data science, but *constitutive* for data science. When it comes to these domains, data science is portrayed as an advanced way of producing domain-immanent knowledge. Most of our interviewees presented their role as data scientists in relation to domains and domain experts with confidence, but also modesty, with the exception of one data science professor who rather jokingly, yet tellingly, described his experience that "using very primitive means" (IV_DE_02, Pos. 94) was often enough to make a big impression. All, however, argued that data science enables researchers to pursue a whole new set of questions that previously had not been possible to investigate. When asked about the societal benefits of data science, one interviewee spoke about deploying data science methods to other disciplines:

> *And in this respect, there are simply new possibilities to conduct research. And also to evaluate data. It just brings new energy and, of course, also new possibilities to do research in new ways.* (IV_AT_02, Pos. 2)

For many interviewees, the specific thing that data science brings to the table was opening up research for new *types* of data and unprecedented *quantities* of data that both had not been possible to analyse before.

Another strength of data science that is repeatedly put forward is its genuinely *integrative* character. While our interviewees described other disciplines and domains as uni-dimensional since they are often restricted by a specific research focus or specific theoretical or empirical approaches, they saw data science as having no such limitations because – by its very nature – it combines different disciplinary perspectives and approaches. That includes the integration of domain-specific expertise. A data science professor explained this point as follows:

> *Because I don't want to contradict my other colleagues, on the contrary, every discipline or every domain has its strengths and its expertise. But sometimes I feel like they have blinders on: like, OK, this is my focus, this is what I can do. And everything else almost becomes secondary. And as a data scientist you have to somehow manage a balancing act between: this is important here, and this aspect from a different area is important as well. You really need to juggle a bit between all the domains that are somehow connected. And you're also doing a bit of bridge-building between all those areas.* (IV_DE_11, Pos. 2)

Here, domains are presented as restricted by "blinders" and overly focused on certain approaches. Going a step further, that also makes them in need of data science's support in bringing different perspectives together by way of "bridge-building". Data science is again positioned as overcoming the weaknesses of established uni-dimensional disciplines as they investigate questions within domains. This boundary work done in relation to so-called domains furthermore presents data science as a means of revitalizing these research areas. By opening up new possibilities for cooperation, data science is positioned as a potentially beneficial research partner. The emerging discipline is symbolically constructed as bridging the gaps between different areas of research, bringing together diverse fields and actors, and thus as a quintessentially *transversal* endeavour (see also Saner 2022), albeit one with a clear symbolic asymmetry: data science is the active partner, the one required for a domain to rise to the task, while the partner fields remain relatively passive and in need of data science.

4.4 A Matter of Interest: Interest-Driven Industrial Data Analytics Versus Value-Free Data Science

Data science professors are well aware that what is now becoming institutionalized as "data science" has long been practiced by (self-taught) data scientists in a range of industrial fields. In presenting and positioning data science as an *academic* endeavour, they also distance themselves from industrial data analytics on the one hand and (in)famous "big tech" companies on the other hand.

When asked how she came to work in data science, one professor who collaborates in several joint projects in the medical field outlined her understanding of data science:

> *And what I understand by that is not simply something like data analysis, not simply data analytics, which is often done in businesses, just analysing or describing data or exploiting data, but really the science behind it.* (IV_DE_06, Pos. 11)

By pointing out the academic and scientific character of her work, the interviewee distinguishes her perception of data science versus "mere" data analytics. Whereas she presents the latter as highly outcome-driven, to the extent of "exploiting" data, she finds the edges of the former in its scientific character. Subtly, a boundary is drawn between non-academic and academic data science based on the distinction of interest-driven vs. interest-free, with the reference to "exploitation" evoking notions of (economic) interests as a driving force behind data analytics. Depicting academic data science as "really" all about the science symbolically links it with established notions of impartial, value-free science, similar to pursuing *l'art pour l'art*.
This idea of doing data science as a goal in and of itself, independent from economic or business interests, was also brought up by another interviewee who linked this argument even more strongly to prevalent scientific ideals of free research (and speech):

> *We also saw at Google recently how they just fired a researcher, […] because she wrote something that Google didn't like. And I wouldn't see that in academia, these restrictions. I don't have a company behind me where I have to be careful to toe the company line or anything.* (IV_AT_01, Pos. 2)

At the same time, Google and other “big tech” companies, often referred to as the “usual suspects” (IV_AT_01, Pos. 2) and therefore in no need of an introduction, are also portrayed as Goliath-like competitors with whom academic data science simply cannot compete. Due to the economic, technical and human resources available to them, it is an accepted view that these companies are able to do a level of data science that is out of reach for any university data scientist. Our interviewees noted the perceived superiority of big tech companies with mixed feelings, admiring some achievements and possibilities, while also remaining sceptical about potential threats, especially due to the monopolistic positions of these enterprises. One data science professor reflected on their influence, not only on the institutionalization of data science as a discipline, but also on society as a whole:

> *They were able to anticipate very well what will be in high demand, what – maybe they can also steer it, what people want, that's always such a question. But solutions are often offered for really urgent problems or [problems that] are made urgent – I don't know, it is difficult to judge, I think, but that is of course an issue that these are the solutions that people then use.* (IV_DE_05, Pos. 52)

As the argument develops from “offering” solutions to urgent problems, to “making” these problems urgent in the first place, to finally diffusing solutions to potentially fabricated challenges, this quote clearly has a critical undertone: doing data analytics in non-academic contexts is once again linked to (economic) interests and academic data science, on the other hand, is positioned as interest-free.

In terms of available resources, industrial data analytics, especially the opportunities in “big tech” companies, are presented as superior to academic data science. But the latter is seen as adhering to scientific values, especially those of non-partisan and value-free research, independent from economic interests or company concerns. This distinction between academic data science and industrial data analytics aligns closely with the science–non-science boundary of central interest to Gieryn (1999), situating it along the axis of value-free versus interest-driven data analysis.

4.5 A Question of Honour: Public Hype Versus Serious Data Science

Although all interviewees – little to our surprise – pointed out the advantages of data science, claiming its rightful place in the academic field, they were also critical towards common public perceptions and ideas about data analytics. Many explicitly distanced data science from public (and sometimes, academic) discourses surrounding big data, AI and data science, which they described as “hype” and even hyperbole, both with potentially perilous consequences.

One repeated distinction we found was the one between data science and big data. An interviewee described his path into data science as paralleling the development from earlier big data hypes towards the supposedly much more (self)reflexive and, thus, serious endeavour of data science:

> *I think data science somehow came out of big data. [...] Early in 2010 and in the 2010s and so on, there was a lot of hype: big data. [And we all thought] there's just so much data now and we don't need to look after it or anything, it's just all there. All we need to do is analyse and then we'll get what we're looking for. But that was really all hype in the end. It didn't turn out that way. Instead, it took a different turn [we found out that] the data has to be looked at very well. The distortions, the biases have to be found. And it's not possible to find something in the data that's not there. So basically, all the things we've known since the dawn of statistics. It became clear that it all applies to big data as well. And that's when we started moving toward data science.* (IV_AT_01, Pos. 2)

In this historical narrative, data science is not only presented as self-reflexive and self-critical – in other words, aware of biases and reflecting on its method and data choices – but, having "overcome" big data, also as capable of advancement and self-improvement.

Another interviewee drew comparisons between past (and present) promises made in the name of "multimedia" and artificial intelligence. Linking these hypes with economic interests, he hoped for data science to "normalize" and become a solid part of (computer) science:

> *So, whoever did multimedia got money. [...] That's exactly how I see it. […] And that's also how I see AI. So, AI is a farce, data science is a part of AI. It's a part of computer science, of course. Yes, it has a certain validity. But there is a certain belief in hype. […] My hope is that it won't end up in the same bucket we put multimedia in today [...] My fear is that it might end up like that. But my hope is that it will be normalized and data science will just be a normal part of working in different domains. And a solid part of computer science.* (IV_DE_14, Pos. 35 and Pos. 153)

Looking at the boundary work done to distinguish data science from hype, data science is positioned as a serious and robust scientific enterprise in contrast to dubious, exaggerated and untrustworthy promises. Some of our interviewees also brought up the problematic role of some data science scholars who fuel unrealistic promises that eventually lead to disappointment but also endanger the epistemic authority of academic data science:

> *What I think is more likely to happen is that we start to over-promise, in the sense that it's all very simple: you press a button and then a perfect decision, your perfect decision model, your perfect prognosis comes out, which is then also fully understandable and totally explainable and free of error and*

> *anyone can do it. You do a weekend course on, well, on Coursera and then everything works. But the thing is, we are very, very, very, very far away from that. […] Because they* [scholars appropriating the label data science; the authors] *are travelling under the wrong flag and with the wrong label. And I'm really afraid of that. That they'll say, this is it and we're publishing it. Because, well, it sounds very nice, fancy and all that. So somehow, it's all new and there's a lot of hype behind it, but then maybe it's no longer true in detail.* (IV_DE_03, Pos. 88)

As the above quote shows, unrealistic promises are problematic not only because they eventually lead to disappointment, but also because they undermine the epistemic authority of data scientists. They suggest that everyone could ultimately do data analytics in the blink of an eye, an implication that belittles the expertise and skills necessary for serious data analytics. Again, boundary work involves contrasting public perceptions of data science against academic data science along the line between unrealistic and serious or untrustworthy versus reliable.

The boundary work observed above, undertaken to distance data science from commonly held public notions surrounding data analytics or data science, can also be interpreted as proactive engagement with critical voices and public concerns, especially with questions of bias that have gained considerable public attention in recent years and even lead to regulatory attempts aimed at responsible companies and technologies (Andrews et al. 2017; Prietl 2021).

5 Discussion

This paper set out to study the boundary work accompanying the academic institutionalization of data science in order to understand current developments in the structures and modes of knowledge production in society.

As we have shown, there are several lines of demarcation discursively drawn by data science professors to construct academic data science on a symbolic level. In claiming a place for data science in academia, data science is distinguished from industrial data analytics on the one hand and popular notions of (big) data analysis on the other. Compared to industrial data analytics, which is presented as driven by – mainly economic – interests, academic data science is associated with long established ideals and norms of doing science, especially those of value-free and interest-free research for the sole sake of advancing science and knowledge. In critically distancing academic data science from what is often referred to as "hypes" surrounding big data and AI, it is furthermore positioned as a serious, trustworthy and reliable scientific endeavour, one that is even capable of self-critique and self-improvement. *Within* academia, data science is distinguished from its "parent disciplines" – mathematics, statistics and computer science – as well as from so-called domains that are each depicted as being uni-dimensional and limited in scope. In contrast data science is characterized as "having it all", being "real-world oriented" as well as scientific, technical as well as social, generalist as well as domain-specific. The boundary work done to distinguish data science from mathematics, statistics and computer science primarily serves the goal of positioning data science *in place of* these established disciplines, thus attacking their epistemic authority. By comparison, the boundary work done in distinction to so-called domains emphasizes the transversal openness of data science and the promise of bringing new energy into the research agenda of domains, thus, positioning data science as a beneficial partner for joint (research) projects. Last but not least, referring to the theory–practice boundary or pure–applied science boundary, hierarchical lines are being drawn within data science itself, positioning methods-driven data science as symbolically superior to applied data science.

Following David Beer's (2022) argument that *tensions* are constitutive of algorithmic thinking, we can see that the boundary work done to constitute data science is also quite charged. Whereas when distinguishing data science from mathematics, statistics or computer science, it is viewed as an applied, worldly and socially competent discipline, these ascribed characteristics change once the focus is turned onto itself. Within data science, an emphasis on methods and the scientificness of academic data science are used to establish a hierarchy between methods-driven and application-oriented data science. Robert Dorschel (2021) argued that data scientists are constructed as *hybrid professionals*: they integrate supposedly mutually exclusive characteristics such as being technically skilled and socially capable, or exploiting data while also caring about privacy and ethics. Our analysis goes a step further in demonstrating that the boundary work done to symbolically construct academic data science is more than just inherently tense and hybrid; it is flexible and above all *relational*, yet in no way arbitrary. We observed a pronounced flexibility of boundary work in terms of content. Data science was linked at one moment to theory and "pure" science, and at the other to the seemingly contrasting elements of practice and applied research. These variations, however, are by no means random; they become understandable once they are set in relation to the respective context and subject of distinction: The specific form, content and references that boundary work draws upon vary depending on what (or whom) data science is being related to, and whether that relation is one of distinction or connection (for a similar finding with respect to the symbolic construction of engineering in renewable energies, see also Prietl 2019b, 108–109).

Content-wise, data science's constant balancing act around the theory–practice boundary furthermore resembles the symbolic construction of engineering as a discipline (Paulitz 2012; Paulitz and Prietl 2013). Thus, it seems that data science not only builds on the tradition of engineering when it comes to the structuring and organization of its curricula (Saner 2022), but also in how it is symbolically constituted.

What are the implications of these findings on a profession in becoming for established structures and modes of knowledge production in society? What social demarcations (Lamont and Molnár 2002) could result from the symbolic boundaries depicted? The professionalization and academic institutionalization of data science might affect the (academic) "system of professions and disciplines" and how "turf" is (re)divided among different actors in several ways: the epistemological claims made in the name of data science primarily attack established disciplines, and destabilize their epistemic authority as the guardians of high-level quantitative data analysis. Said disciplines include statistics, mathematics, and computer science, but also quantitative social science. While at first glance offering possible collaborations, data science also challenges the former "sole" epistemic authority of experts in other disciplines – now also symbolically reduced to domain experts in contrast to the seemingly limitless mandate of data scientists. More generally, the widespread demand for data science methods in academic as well as non-academic "domains", might delegitimize other – especially non-quantifying – modes of doing research and knowing in these areas. At the same time, scholars working with data science methods or collaborating with data scientists might see their standing rise. As Gieryn has already pointed out with regards to the distinction of science versus non-science, these symbolic struggles need to be understood as having serious material consequences. In short, actors compete for epistemic authority for good reason: epistemic authority is *the* key asset when competing for (research) funding and talent.

Looking ahead to future avenues of research, studying reactions to the rise of data science in different domains where data science methods are now applied could be one interesting angle. Interdisciplinary contexts in particular could offer an insightful setting to research the "turf wars" between data scientists and representatives of other – established – disciplines as they negotiate epistemic authority and disciplinary boundaries. Considering the importance of cultural representations for claiming epistemic authority and how performative promises drive (technological) research (Borup et al. 2006) and technology implementation in organizations (Raible 2022), further research on the role of *expectations in technology development* might also be revealing in the context of data science, especially discovering how data scientists balance tensions surrounding narratives of hopes and promises, on the one hand, and (self-)critical assessments of hypes and containment of unrealistic expectations, on the other.

6 References

Abbott, Andrew.1988. *The System of Professions: An Essay on the Division of Expert Labor*. Chicago: University of Chicago Press.

Abbott, Andrew. 1995a. Boundaries of Social Work or Social Work of Boundaries? *Social Service Review* 69(4): 545–562.

Abbott, Andrew. 1995b. Things of Boundaries. *Social Research* 62(4): 857–882.

Andrews, Leighton, Bile Benbouzid, Jeremy Brice, Lee A. Bygrave, David Demortain, Alex Griffiths, Martin Lodge, Andrea Mennicken, and Karen Yeung. 2017. Algorithmic Regulation. *LSE Discussion Paper* 85. London School of Economics and Political Science.

Bartlett, Andrew, Jamie Lewis, Luis Reyes-Galindo, and Neil Stephens. 2018. The Locus of Legitimate Interpretation in Big Data Sciences: Lessons for Computational Social Science from -omic Biology and High-Energy Physics. *Big Data & Society* 5(1): 1–15.

Beaulieu, Anne, and Sabina Leonelli. 2022. *Data and Society. A Critical Introduction*. Thousand Oaks: Sage.

Beer, David. 2019. *The Data Gaze: Capitalism, Power and Perception*. Thousand Oaks: Sage.

Beer, David. 2022. *The Tensions of Algorithmic Thinking. Automation, Intelligence and the Politics of Knowing*. Bristol: Bristol University Press.

Bonde Thylstrup, Nanna, Mikkel Flyverbom, and Rasmus Helles. 2019. Datafied Knowledge Production: Introduction to the special theme. *Big Data & Society* 6(2): 1–5.

Borup, Mads, Nik Brown, Kornelia Konrad, and Harro van Lente. 2006. The Sociology of Expectations in Science and Technology. *Technology Analysis & Strategic Management* 18(3–4): 285–298.

Dorschel, Robert. 2021. Discovering Needs for Digital Capitalism. The Hybrid Profession of Data Science. *Big Data & Society* 8(2): 1–13.

Dorschel, Robert, and Philipp Brandt. 2021. Professionalisierung mittels Ambiguität: Die diskursive Konstruktion von Data Scientists in Wirtschaft und Wissenschaft. *Zeitschrift für Soziologie* 50(3–4): 193–210.

Gieryn, Thomas F. 1983. Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review* 48(6): 781–795.

Gieryn, Thomas F. 1994. Boundaries of Science. Pp. 393–443 in *Handbook of Science, Technology and Society*, edited by Sheila Jasanoff, Gerald Markle, James Peterson, and Trevor Pinch. Thousand Oaks: Sage.

Gieryn, Thomas F. 1999. *Cultural Boundaries of Science: Credibility on the Line.* Chicago: University of Chicago Press.

Houben, Daniel. 2022. *Die verborgenen Mechanismen der Governance*. Wiesbaden: Springer.

Houben, Daniel, and Bianca Prietl (eds.). 2018. *Datengesellschaft. Einsichten in die Datafizierung des Sozialen.* Bielefeld: transcript

Kitchin, Rob. 2014. Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society* 1(1): 1–12.

Kitchin, Rob. 2022. *The Data Revolution. A Critical Analysis of Big Data, Open Data & Infrastructure*. 2nd edition. Thousand Oaks: Sage.

Jarke, Juliane, Bianca Prietl, Simon Egbert, Yana Boeva, Hendrik Heuer, and Maike Arnold (eds.). forthcoming. *Algorithmic Regimes. Methods, Interactions, Politics*. Amsterdam: Amsterdam University Press.

Lamont, Michèle, and Virág Molnár. 2002. The Study of Boundaries Across the Social Sciences. In: *Annual Review of Sociology* 28: 167–195.

Lowrie, Ian. 2017. Algorithmic Rationality: Epistemology and Efficiency in the Data Sciences. *Big Data & Society* 4(1): 1–13.

Paulitz, Tanja. 2012. "Hegemoniale Männlichkeit" als narrative Distinktionspraxis im Wissenschaftsspiel. *Österreichische Zeitschrift für Soziologie* 37(1): 45–64.

Paulitz, Tanja, and Bianca Prietl. 2013. Spielarten von Männlichkeit in den «Weltbildern» technikwissenschaftlicher Fachgebiete. *Informatik-Spektrum: Archive of Applied Mechanics* 36: 300–308.

Paulitz, Tanja, Susanne Kink, and Bianca Prietl 2015. Fachliche Distinktion und Geschlechterunterscheidung in Technik- und Naturwissenschaften. Grundlagen- und anwendungsorientierte Wissenskulturen im Vergleich. Pp. 207–225 in *Akademische Wissenskulturen und soziale Praxis. Geschlechterforschung zu natur-, technik- und geisteswissenschaftlichen Fächern,* edited by Tanja Paulitz, Barbara Hey, Susanne Kink, and Bianca Prietl. Münster: Westfälisches Dampfboot.

Prietl, Bianca. 2019a. Die Versprechen von Big Data im Spiegel feministischer Rationalitätskritik. *GENDER* 3: 11–25.

Prietl, Bianca. 2019b. Changierende Subjektpositionen und implizite Vergeschlechtlichungen. Zur diskursiven (Re-)Produktion hierarchischer Geschlechterverhältnisse am Beispiel der Ingenieurarbeit im Bereich erneuerbarer Energien. *Soziale Welt* 70(1): 93–118.

Prietl, Bianca. 2021. Warum Ethikstandards nicht alles sind. Zu den herrschaftskonservierenden Effekten aktueller Digitalisierungskritik. *Behemoth, Special Issue Future(s) of Critique:* 19–30.

Prietl, Bianca, and Armin Ziegler. 2016. Machtvolle Grenzen als konstitutive Momente des Sozialen. Grenzziehungen als Analysekonzept für eine Soziologiegeschichte. Pp. 1–16 in *Handbuch Geschichte der deutschsprachigen Soziologie. Band 2: Forschungsdesign, Theorien und Methoden*, edited by Stephan Moebius and Andrea Ploder. Wiesbaden: Springer.

Prietl, Bianca, and Stefanie Raible. *Forthcoming*. The Politics of Data Science. Institutionalising Algorithmic Regimes of Knowledge Production. In *Algorithmic Regimes: Methods, Interactions, Politics*, edited by Juliane Jarke, Bianca Prietl, Simon Egbert, Hendrik Heuer, and Maike Arnold. Amsterdam: Amsterdam University Press.

Raible, Stefanie. 2022. Organisationen als Treiber und Getriebene von Digitalisierung – Zur Dualität von Digitalisierungsnarrativen. *Arbeits- und Industriesoziologische Studien* 15(2): 62-75.

Raible, Stefanie, Werner, René, and Laube, Stefan. 2023. Begegnungen im Digitalen. Qualitative Sozialforschung als intersituative und polymediale Praxis. *Forum Qualitative Sozialforschung* 24(2): Art. 11.

Ribes, David. 2019. STS, Meet Data Science, Once Again. *Science, Technology, & Human Values* 44(3): 514–539.

Ribes, David, Andrew S. Hoffman, Steven C. Slota, and Geoffrey C. Bowker, 2019. The Logic of Domains. *Social Studies of Science* 49(3): 281–309.

Saner, Philippe. 2019. Envisioning Higher Education: How Imaging the Future Shapes the Implementation of a New Field in Higher Education. *Swiss Journal of Sociology* 45: 359–381.

Saner, Philippe. 2022. *Datenwissenschaften und Gesellschaft. Zur Genese eines transversalen Wissensfeldes.* Bielefeld: transcript.

Slota, Steven C., Andrew S. Hoffman, David Ribes, and Geoffrey C. Bowker. 2020. Prospecting (in) the Data Sciences. *Big Data & Society* 7(1): 1–12.

Strauss, Anselm, and Juliet Corbin. 1996. *Grounded Theory: Grundlagen Qualitativer Sozialforschung.* Weinheim: Beltz.

Turkle, Sherry. 1986. Computational Reticence: Why Women Fear the Intimate Machine. Pp. 41–61 in *Technology and Women's Voices*, edited by Cheris Kramarae. New York: Pergamon Press.