Big Data and the Paradox of Diversity

Bernhard Rieder

Abstract

This paper develops a critique of Big Data and associated analytical techniques by focusing not on errors – skewed or imperfect datasets, false positives, underrepresentation, and so forth – but on data mining that works. *After a quick framing of these practices as* interested *readings of reality, I address the question of how data analytics and, in particular, machine learning reveal and operate on the structured and unequal character of contemporary societies, installing "economic morality" (Allen 2012) as the central guiding principle. Rather than critiquing the methods behind Big Data, I inquire into the way these methods make the many differences in decentred, non-traditional societies knowable and, as a consequence, ready for profitable distinction and decision-making. The objective, in short, is to add to our understanding of the "profound ideological role at the intersection of sociality, research, and commerce" (van Dijck 2014: 201) the collection and analysis of large quantities of multifarious data have come to play. Such an understanding needs to embed Big Data in a larger, more fundamental critique of the societal context it operates in.*

Keywords: Big data; analytical techniques; digital methods; data analysis; data mining.

Introduction

The emergence of Big Data, both as an imminent potentiality and an actual practice, has fuelled considerable discussion of possible social ramifications, ranging from the loss of privacy to the detrimental consequences of statistical bias. The latter aspect, in particular, highlights the implications of statistical models and decision rules learned from datasets that can be skewed or deficient in various ways. Sweeney (2013), for example, shows how racial bias manifests in ad delivery and Barocas/Selbst (2016) list various ways in which discrimination can creep into data mining to produce "disparate impact". An edited volume on *Discrimination and Privacy in the Information Society* (Custers et al. 2013) constitutes the so far most comprehensive effort to bring together computer scientists and legal scholars to investigate issues raised by data mining and consider possible solutions.

DOI 10.14361/dcs-2016-0204

These works highlight the subtle intricacies involved in decision-making[1] that relies on data analysis techniques, which are not only conceptually difficult to apprehend due to their reliance on advanced mathematics, but also practically daunting, in the sense that decision models are not directly designed by a human agent, but derived from the iterative parsing of large quantities of complex and often problematic data. A growing body of work in the humanities and social sciences (cf. Kitchin 2016) has begun to highlight how algorithms intervene in various domains, although the specific operating principles that underpin algorithmic decisions remain elusive and hard to integrate into a vocabulary of agency designed to capture human conduct and reasoning.

At the same time, we witness not only the integration of algorithmic filtering, classification, and recommendation into the fabric of our digital environments, but also the intensification of a long-standing trend in commerce and government to turn to *empirical*, data-driven procedures that heavily rely on measurement and counting to make – and justify – decisions (cf. Porter 1995). In many areas of organizational life, we can see the development and application of approaches that set out to make Hume's (1739) famous consternation with the jump from an *is*, a description, to an *ought*, a prescription, somewhat less distressful. Cost-benefit analysis, evidence-based practice, data-driven government, impact analysis, and many other approaches promise to reconcile the tension between the increasingly accepted moral imperatives of impartiality, fairness, transparency, and accountability on the one side, and, on the other, the factual necessity to make decisions that will impact some people differently than others and thus participate in picking winners and losers. The turn to the empirical – through data collection and analysis – as arbiter is certainly not uncontested as the continuing disputes over global warming or vaccination indicate, but to place *ideology* over *evidence* has become a sure-fire way to get shunned from the community of the rational and reasonable – only the ignorant und unenlightened tune out the facts. At the same time, the "new spirit of capitalism" (Boltanski/Chiapello 1999) may have replaced the rigidities of Taylorist scientific management with principles such as flexibility, enthusiasm, and collaboration, but the associated embrace of uncertainty and probabilistic reasoning has led to *more* empiricism, not less.

This paper approaches Big Data practices as empiricism *on steroids* and develops a critique built around the idea that data analysis is more often than not applied to produce actionable forms of knowledge that are used "as tools for assessment, action, and decision" (Desrosières 2001: 344) instead of disinterested description. This critique deviates from the more common repudiations of Big Data's claims to objectivity (cf. Kitchin 2014) by situating data analysis and associ-

1 The term decision-making is often used in psychology, cognitive science, and behavioral economics to denote the selection between alternative beliefs or courses of action. My use of the term follows this general definition without necessarily subscribing to the larger theoretical frameworks it is usually embedded in.

ated practices in an epistemological paradigm that is informed by the purposes and ideas of the business world rather than modelled upon the predicates of scientific inquiry. Ad targeting techniques may not be able to *know* a user in any meaningful sense of the word, but they are clearly capable of producing higher click-through rates. Assessing Big Data from the angle of its effectiveness in delivering advantageous outcomes rather than its capacity to yield descriptive truth moves the space of potential issues from (broken) epistemic promises to everyday practices of social sorting and, consequently, questions of social justice (Lyon 2003).

These issues are certainly not unfamiliar, but, as I want to argue, Big Data raises them anew and with a twist. However, this paper also departs from broader, more comprehensive critiques (cf. Ekbia et al. 2014) by largely leaving aside the many issues stemming from deficiencies or inaccuracies in data analytics – skewed or imperfect datasets, false positives, underrepresentation, and so forth – to focus on data mining that *works*. Not that these issues are not important, on the contrary. But there is also need for more theoretical inquiries into the epistemic character of Big Data, into the ways of knowing it aspires to, and into its relationship with the dominant normative horizon of contemporary western societies. The objective, in short, is to add to our understanding of the "profound ideological role at the intersection of sociality, research, and commerce" (van Dijck 2014: 201) the collection and analysis of large quantities of multifarious data have come to play. After a quick framing of these practices as *interested* readings of reality, I will therefore focus on the question of how data analytics and, in particular, machine learning reveal and operate on the structured and unequal character of contemporary societies, highlighting "economic morality [as a] guiding logic that conditions and directs our daily lives" (Allen 2012: 19).

Big Data

To begin the argument, it is necessary to outline the technological context, space of application, and normative background I will be referring to throughout this text. There are numerous definitions of Big Data and associated techniques, but for the purpose of this paper, four elements are crucial.

First, we witness the steadily increasing production and availability of very large datasets that often comprise transactional data (logged behaviour) or other forms of non-traditional data such as social interactions, cultural tastes, or sensor readings.

Second, algorithmic techniques for data analysis, many of them probabilistic and capable of *learning*[2], have become widely available. Code libraries for various programming languages, easy-to-use analytics software, and integrated data

2 It should be noted that machine learning implements a specific and partial theory of learning that boils down to forms of statistical induction.

infrastructures offer sophisticated methods to mine vast amounts of stored data at high speeds and from diverse perspectives in the context of a quickly growing set of applications.

Third, the rampant computerization[3] of all aspects of contemporary life means that ever more practices are unfolding in online environments that allow for data collection as well as for the automation of decision-making and the performative implementation of its results. Differential pricing on the web provides an elucidating example: a user's location, software environment, browsing behaviour, and other elements can be situated against a horizon of millions of other users and their shopping behaviour; this knowledge can then be used to set the sales price of an item to the highest level the user has been estimated to support. The result of this calculation, made in the fraction of a second, can then be directly integrated in the interface served to that user, showing an individualized[4] price for an item. This instant *applicability* of data analysis is a crucial step beyond traditional uses of statistics because it integrates and automates the sequence of collecting data, making decisions, and applying results, thereby relegating human discretion to the design and control stages. As a consequence, the scope of data-driven techniques has been continuously extended from bureaucratic management into areas such as information ordering, real-time credit assessment, product pricing, or cultural recommendation.

Fourth, and more broadly, the relentless drift in economic and social organization towards market forms makes techniques that can adapt to and control complex and dynamic situations increasingly attractive. Especially in settings where largely autonomous actors cooperate and compete in shared informational infrastructures, such as online environments, fast, yet informed decisions are rewarded. Paraphrasing Andrejevic (2013), one could argue that algorithms are seen as the go-to solution whenever there is "too much": too many people and things, too much information, and, of course, too little time. This seems to apply to more and more areas of contemporary life, from business and government to the various online platforms that heavily rely on filtering, recommendation, and aggregation.

In addition to these four elements, a broader aspect I have already mentioned needs to be emphasised, namely that "[j]udgment and discretion, normally the prerogatives of elites, are discredited" (Porter 1995: 97), particularly in domains where (social) trust is eroding. Choosing a particular course of action based on

3 While the term has fallen out of fashion, it is highly useful to shift the focus to the computer – and not just its digital code – as the fundamental technological component of our "information societies".

4 A recent report by the White House summarizes: "Broadly speaking, big data seems likely to produce a shift from third-degree price discrimination based on broad demographic categories towards personalized pricing and individually targeted marketing campaigns." (Executive Office of the President of the United States 2015: 19)

the analysis of empirical data is by no means a new phenomenon in business and government. But even in mature, impersonal bureaucracies, these processes are riddled with moments of human discretion in the sense that managers and administrators interpret results and have a level of leeway when it comes to deciding how to proceed. This residual element of volition in moving from *is* to *ought* has been coming under scrutiny in areas where individual authority is perceived or portrayed as inadequate, inefficient, partial, paternalistic, corrupt, or illegitimate. In these areas, fully formalized, automated decisions have become more and more attractive as effective and supposedly neutral or even democratic procedures, in particular if they implement an empirical component that can be presented as "carrying" the actual decision. Responsibility can then be shifted to the data themselves. While Porter describes a process spanning the last two hundred years, the ambiguous connection between democratization and quantification clearly echoes through the rhetoric of Big Data with renewed vigour and connects tightly to the intensifying "legitimation crisis" (Habermas 1973) traditional institutions and modes of authority have been experiencing.

Taken together, these elements explain why Big Data is perceived as technically and practically feasible, economically appealing, and socially – and even ethically – desirable.

Data Analysis and Accounting Realism

The turn to the empirical relies heavily on the proliferation of instances of data collection, but the ways these data are being made to signify is particularly relevant for understanding how normativity comes into play. To this end, the algorithms involved in analysis and automated decision making – which generally manifests as some kind of *ordering* (e. g. a ranked list, a categorization, etc.) – need to be distinguished into two morphological lines.

In the first case, the decision model is explicitly designed. The (in)famous impact factor for scientific journals, for example, is the average number of citations papers in a journal received in the two previous years. *Somebody*, in this case Eugene Garfield, decided that this would be a good formula to capture "impact" and a sufficient number of people agreed, turning the metric into a widely accepted means for ranking scientific publications. Outcomes can be – and are – presented as distributed decisions where every researcher gets her "vote", simply by citing others, but the calculative procedure has an identifiable author and, more importantly, a clear and stable content that can be scrutinized and criticized.

In the second case, however, the decision model is derived through statistical learning. A spam filter, for example, requires specimen of *spam* and *ham* emails from its users. Parsing through these examples, it will generate a decision model where each word becomes a probabilistic indicator for the two categories. If the word "Viagra" always appears in emails marked as spam and never as ham, it will

become a strong indicator for "spamminess". All words are taken into account and if the combined score exceeds a certain level, the email is flagged as spam. This is, in a nutshell, how machine learning works and it removes the decision model a step further since it becomes *adaptive* (the classifier changes with shifting email content) and potentially *personalized* (my filter classifies differently from yours because my spam is your ham). There is still a calculative procedure, but it no longer contains a clear normative proposition like the Impact Factor; rather, it orchestrates how the empirical examples – the mails marked as spam or ham – are turned into a decision model through human feedback. Both the content and the author of the decision model become substantially vague (cf. Burrell 2016).

The second group of algorithms informs what I have called "interested readings of reality" (Rieder 2016), assessments that are not just applied in operational settings, but fully permeated by operational goals in terms of their epistemological makeup. We are currently witnessing the proliferation of a particular use of statistical techniques, which, in Desrosières' (2001) terms, does not subscribe to "metrological realism" predicated on a correspondence theory of truth, but to "accounting realism", an epistemological stance that assesses truth – or rather validity – in relation to an operational objective, for example profit maximization. Machine learning techniques fit the requirements of accounting realism almost perfectly since they are *inductive* in the sense that they do not test or apply a hypothesis (e. g. what spam looks like), but generate it from an *interested* appraisal of past experience. Most people are not keen on describing or theorizing spam, they simply want it gone. One could argue that the trained statistical classifier containing the probability values for all parsed words represents a "theory" of spam, but this theory will vary between users and, most importantly, is derived from feedback rather than explicitly laid out.

This has profound consequences for how decisions come to be made and how judgement is operationalized. Rather than formulating a theory of what makes a "good" employee, which may be seen as tainted by common biases, a manager may turn to machine learning for counsel on hiring by submitting a set of well-structured CVs of excellent current or past employees to the computer. The learning technique will then derive ("learn") a statistical model consisting of correlations between the CV data and the performance assessment as target variable – the component carrying what I call *interest*[5]. In the case of spam, it is the binary value spam/ham that is trained for, while in a hiring process it may be the number of sales or some other stand-in for "good performance". The statistical

5 It is important to note that this need not be monetary value. The manager may very well decide that she is looking for the funniest hire possible and select the employee CVs used for training according to their capacity to entertain the office. The classifier will then correlate the submitted data with that assessment and produce a decision model. But, as Desrosières (2001: 342) argues, accounting realism generally relies on *money* as general equivalent.

model can then be used to classify incoming job candidates according to their predicted performance. This requires relatively little discretion or judgment from the manager other than the trust in the method itself, since using job performance as target variable would hardly seem controversial. If the CVs used to train the model happen to indicate that employees with higher educational attainment generate higher value for the company, the model will reflect that. Just like the spam filter, every variable present in the CVs will be correlated with the desired outcome.

Readers concerned with metrological truth may protest that correlation does not mean causation. They may interject that there is no "raw" data (cf. Gitelman 2013; van Dijck 2014), that the data used in machine learning is itself *produced* in various ways – skewed, incomplete, and generated through interfaces that more often than not impose discrete choices on fluid matters. They may, more fundamentally, take issue with the very idea that often highly decontextualized data can be used to produce adequate knowledge about complex social situations that require a "situated and embodied" (Haraway 1988) perspective. These objections would clearly have merit – and yet miss the point. The epistemological problem of accounting realism is not to *describe* the sociological makeup of society, but to *decide* whether a given job candidate should be hired or not. The goal is not truthful description, but good – i.e. profitable – decision-making in situations where too much information meets too little time. What matters most, here, is that contemporary forms of mechanical reasoning propose methods that seemingly circumvent normative commitment by turning to the empirical, reading it through the lens of operational goals. Judgement, understood as the evaluation of evidence to make a decision, becomes a product of statistical analysis and thus acquires an aura of objectivity, rationality, and – most importantly – a legitimacy that derives from its empirical underpinnings. The manager in our CV example could easily claim that her decision was as objective as they come. Not only on account of the computational technique used, but also because productivity in monetary terms is the very value or interest seen as requiring no further justification. So where is the problem? To answer this question, we first need to make a detour to comment on the makeup of contemporary societies, which constitute the material data analysis processes.

Structured Societies and Big Data

Data proliferate in contemporary societies because an ever growing number of things we do are, in one way or another, touched by computerization. Mediation through interfaces, databases, and algorithms may well involve a loss of immediacy or some other element of "artificialization", but this can be said of all aspects of culture. For all intents and purposes, the technical environments we inhabit are, indeed, our *real*, and the data these environments produce so effortlessly reflect

part of it. There would be many caveats to add at this point, but for the sake of the argument I am trying to develop, I propose that we consider the possibility that the masses of data are not a hallucinatory fever dream, but a somewhat spotty and skewed window on societies that are, in part, organized through the same technical structures that produce these data in the first place. Their analysis therefore reveals our societies, at least particular aspects from particular vantage points. When reflecting on potential ramifications of Big Data and the machine learning techniques I just described, we need to think about what it means to know these societies through the lens of accounting realism.

Modernity, and in particular the period since the Second World War, is characterized by processes of *individualization* and *diversification* of situations and styles of living (Beck 1986: 122). The emergence of consumer capitalism has shifted the focus from production to consumption and produces an ever more fine-grained variety of products and experiences in virtually all areas of human existence, from food to cultural goods and vacations. Societies adopting liberal democracy have seen many traditional social segmentations and taboos erode, continuously extending individuals' capacities to live lives that differ substantially from those lived by both previous generations and the next door neighbours. According to Giddens, ours are decentred, non-traditional societies "where social bonds have effectively to be *made,* rather than inherited from the past" (1994: 107) and where "choice has become obligatory" (1994: 76). One may rightfully wonder whether there is any "real" difference between the many breakfast cereals available in every supermarket, but my objective, here, is not to adjudicate whether these variations in patterns of consumption, in socio-economic status, in geographical anchoring, in political and social values, in sexual preferences, in cultural identities and tastes, and so forth are meaningful or not. The argument I want to put forward is threefold: first, we live in societies characterized by high degrees of diversity in terms of lived lives; second, these lives are constantly logged and surveyed in various ways, leading to enormous amounts of data that reflect (some of) their diverse character; third, these lived lives are *patterned* and not random. The last point requires additional elaboration.

The social sciences have spent the last two-hundred years trying to understand how individuals and society relate, how variation and commonality entwine to produce complex and dynamic arrangements that stabilize through various continuities and institutions. The most common term used to address stability in society is that of *structure,* whether it is understood descriptively to denote non-randomness or analytically to refer to actual social forces. The notion of social structure is partially tied to instances of group membership, both externally attributed or used by actors to demarcate themselves. Categories along the lines of estate, class, caste, ethnicity, race, nationality, profession, and so forth are the result of historically produced (socioeconomic) classification and stratification that resulted in more or less consistent groups that shared characteristics and social standing, which, in turn, differentiated them from other groups. These segmen-

tations have – at least in part – lost their "binding force" (Giddens 1994: 63) and structuring capacity, as well as their utility as descriptive concepts. Traditional arrangements have been disrupted and the new ones are more complex, dynamic, and opaque.

One may wonder in how far attempts to think social structure from the bottom up are reactions to these transformations. Simmel's (1908) "social geometry" can already be seen as a way of conceptualizing "societification" *(Vergesellschaftung)* from the individual, who, due to increasing social differentiation, enters into complex relationships with various others and is less and less confined to her primary group. The recent interest in Tarde's monadological understanding of society (Latour et al. 2012), as well as the continued popularity of other "atomistic" currents – from social exchange theory to social network analysis – can be seen as methodological trends or, more fundamentally, as attempts to grapple, conceptually, with decentred societies that are grouping in more flexible, transient, and diverse ways. However, if it has become hard to speak of a working class today, it is not because economic exploitation has disappeared, but because forms of economic exploitation have become too intricate and varied to summarize them easily into a clear-cut sociological concept. The diversity of lived lives does not imply equality and both domination and stratification continue to exist, even if their consequences are increasingly individualized.

But why am I talking about the shape of society and our conceptual means to describe it in a paper on Big Data? Because in a situation characterized by social differentiation on the one side and ambivalent forms of global and local integration on the other, data collection and analysis promise to make the social *legible* again, to reinstall mastery over societies that continuously diversify, creating differentiations that no longer conform to traditional groupings and categorizations. This is the *raison d'être* of computational data analysis. As complexity and opacity grow, the epistemic and commercial value of techniques that promise to produce viable descriptions and decisions grows as well. This promise, however, still hinges on the "structuredness" of society in the sense that elements may be arranged in increasingly complicated ways, yet not devolve into randomness. Forms of coherence, commonality, and stability continue to exist even if they can no longer be reduced to conceptual pivots such as class. As Giddens remarks, individuals' capacity to make decisions in virtually every sphere of life does not guarantee egalitarian pluralism since "it is also a medium of power and of stratification" (1994: 76). And Bourdieu's (1979) assessment that different forms of capital – economic, social, and cultural – are connected in various ways still holds as well, which means that, for example, years of education, level of income, and cultural tastes correlate. Forms of analysis that make it possible to analyse and act upon such multivariate relationships spanning different domains of life proliferate for this very reason. A recent study in attribute prediction makes a good case in point:

> "Facebook Likes can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender." (Kosinski et al. 2013: 5802)

One may rightfully interject that the researchers used contestable concepts, for example concerning gender, but this would, again, apply criticism from the epistemological headspace of metrological realism to procedures that are more deliberately applied in settings where accounting realism dominates. To put it bluntly: in a situation where the task is to distinguish between a seemingly amorphous mass of customers or other entities, the benchmark is not necessarily to get the prediction right in every case, but to make (quick) decisions *that are more accurate than a coin toss*,[6] speculative inferences that produce an advantageous outcome more often than not. And machine learning generally performs much better than that. The above mentioned study, for example, was able to predict gender with an accuracy of 0.93 and sexual orientation with 0.88. In many cases, a level of 0.51 would be enough to justify applying the technique. The targeting of advertisement, for example, does not have to be perfect to make it economically attractive, merely better than purely random placement. Machine learning is a powerful way to produce such better-than-coin-toss performance at very little cost and it has the additional benefit of providing an empiricist narrative that includes a widely acceptable rationale as well as moments of testability and verifiability when effects, for example on click-through rates, can be directly observed.

Publications like *Dataclysm* by Christian Rudder (2010), one of the co-founders of the dating site OkCupid, provide many examples for patterns and correlations between gender, race, class, and cultural tastes that may seem spurious until one considers their considerable commercial potential. Data mining reveals these structures and social fault lines, and it can order individual cases or ad-hoc groups accordingly. The connection with potential social and political ramifications becomes clear when one considers that, first, decisions based on data analysis can have concrete consequences and, second, the existing structures in society, seen through the lens of data, are informing these decisions. A recent paper on data mining's "disparate impact" formulates the issue clearly:

> "Data mining takes the existing state of the world as a given, and ranks candidates according to their predicted attributes in *that* world." (Barocas/Selbst 2016: 731)

That world, *our* world, is ripe with inequalities and the large and small variations between (datafied) individuals are becoming easily detectable and practical to

6 There are, of course, many areas where higher precision is required, but this (slight) exaggeration should again highlight the fundamental difference between metrological and accounting realism in terms of their epistemic requirements.

distinguish and sort. The structures of differentiation are *read* from an interested perspective and the interest is, more often than not, tied to performance targets: lower loan default ratios, more productive employees, longer time on site, higher click-through rates, and so forth. Big Data, then, is a means to know and act on society on the basis of an empiricism that is epistemically biased in a way that the opposition between objective and subjective does not capture: it is, in a sense, a most impartial way to pursue deeply partial objectives. The capacity to make every data point signify in relation to a goal is the starting point of the third part of my argument, which is concerned with its social and political significance.

Data Mining and the Question of Values

A return to recent developments in sociological theory and methodology can help us gain a better understanding of how data mining produces both interested readings of the data it processes and specific "levers on 'reality'"[7] that reach back into society. The question of social grouping is, again, crucial. Musing on the new availability of datasets and analytical tools, Bruno Latour and colleagues have recently argued that the differentiation between micro and macro, between individuals and aggregates is gradually rendered obsolete as it becomes "possible to account for longer lasting features of social order by learning to navigate through overlapping 'monads' instead of alternating between the two levels of individual and aggregate" (2012: 592). But as researchers navigate digital data not by moving "from the particular to the general, but from particular to more particulars" (2012: 599), aggregating and individualizing at will, so do the algorithmic tools used in settings ruled by accounting realism. The notion of the group ceases to be a stable analytical category and becomes a speculative ensemble assembled to inform a decision and to enable a course of action. The *creditworthy*, for example, are not a distinct class of people, but those the decision model deems capable, at this instant, to pay back a loan – and nothing more. Ordered for a different purpose, the groups scatter and reassemble differently. Foucault (1976: 183) still felt the need to distinguish between *anatomo-politics* targeting the individual and *bio-politics* aiming at the population, even if he contended that they are necessarily linked. When considering contemporary data analysis, this distinction melts before our eyes. We witness the emergence of methods that define units and ensembles at will and project the individual as element of ad-hoc aggregates and vice versa. All of this means that fine-grained differentiation between people, things, or situations – a task which used to be difficult and costly – is becoming easy and cheap,

7 Goody argues that writing facilitates information ordering and retrieval through decontextualization and thereby "gives the mind a special kind of lever on 'reality'" (1977: 109).

making it feasible to individualize and aggregate far beyond the granularity of postal codes, income brackets, gender, or skin colour.

Does this mean that data mining will usher in a future without discriminations based on race or gender? If the rhetoric of impartiality and fairness that accompanies data-driven decision-making is an indication, we should not dismiss the possibility outright. But caveats apply. Recent work on data mining (Calders/Verwer 2010; Custers et al. 2013) have put the finger on how even in highly diversified societies seemingly "innocent" variables, such as cultural tastes, correlate strongly with class, gender, race, and so forth, making it easy to dissimulate explicitly discriminatory decisions, even if the race of a person is not directly assessable. Moreover, insufficient or erroneous data may lead to effects of statistical discrimination if certain groups are over- or underrepresented. But there is a set of more complicated issues that point to the core of the normative argument I am trying to make. The notion of "objective racism" (Barocas/Selbst 2016) highlights the troubling fact that race and other "sensitive attributes" correlate with variables that would seem uncontroversial, for example educational achievement as a factor in hiring decisions. The problem, here, is not that data mining can be biased, but that, after centuries of inequality and discrimination, *empirical reality is biased*. This problem has led to proposals that operate on principles similar to affirmative action (Calders/Verwer 2010). But these solutions require the sensitive attribute to be explicitly present in the data in order to correct for it, which may not be feasible in certain contexts, and they indeed raise the question which attributes should be singled out in the first place. It is in this sense that the very makeup of contemporary societies again comes into view.

Even if one could find ways to create somewhat equal starting conditions, a society that attempts to produce *fair* competition by correcting for past discrimination is by no means one that eschews picking winners and losers. Behind the question of how Big Data may disadvantage particular groups sits the broader issue of what it means for a society that fully embraces many forms of competition and discrimination if every data trace can be used to decide "who should be targeted for special treatment, suspicion, eligibility, inclusion, access, and so on" (Lyon 2003: 20). In contemporary western societies, the ideal of *meritocracy* provides a widely shared normative horizon that justifies instances of selection, hierarchisation, and, indeed, disparate impact, even if – or precisely because – it is considered to be at least partially responsible for the breaking up of segmentations based on gender, ethnicity, and race (Kett 2012). Traditionally, meritocracy installs (educational) achievement as selection mechanism and relies on credentials and tests to signal these achievements. Data driven assessments allow for the inclusion of a much wider array of factors and for the extension of the principle to new applications such as the modulation of health insurance payments based on individuals' level of physical activity. This supports the larger trend towards a setting where "[m]eritocracy has shifted from impersonal technology to a situation where the relation between abilities and rewards has been deeply personalised" (Allen

2012: 5). The use of data analysis for hiring decisions is one specific case, but the proliferation of Big Data implies that many other sorting decisions can and will be made on broad assessments of individuals' seemingly superfluous data traces, which can nevertheless become meaningful and actionable indicators when considered as part of the webs of correlation that permeate structured societies. Due to incomplete or faulty data and errors in speculative prediction, but also due to the fact that the ponderation of the factors going into a decision are both fundamentally opaque (Burrell 2016) and potentially dynamic, people may find themselves in conditions of *paranoid meritocracy*, constantly wondering whether their practices and preferences signal their adherence to "economic morality" (Allen 2012) and their genuine desire to contribute and succeed.

One way to think about possible outcomes are generalized versions of the credit score mechanism, which records individuals' financial behaviour and computes a score that is supposed to express their creditworthiness. Explicit attempts in that direction, such as China's "social credit" score[8], can be understood as disciplinary mechanisms, but they also call attention to the question of social values that is made explicit in such endeavours. While scholars rightfully criticize Big Data and associated practices in terms of method, we should not forget that data mining implements deeply value-laden perspectives due to the normativity implied in the target variable. As I have already mentioned, the legitimacy of data-driven decision-making hinges not only on the presumed objectivity of its methods, but on the unquestioned acceptance of productivity, performance, merit, and, in short, of "economic morality [as a] guiding logic that conditions and directs our daily lives" (Allen 2012: 19).

In extremis, Big Data may simply be a means to project our current value systems more pervasively, thoroughly, and effectively into society. This would then mean that a critique of Big Data requires a critique of these values, for example of meritocracy. As Dahrendorf remarks, "nowadays meritocracy seems to be simply another version of the inequality that characterises all societies" and it may even be "a particularly cruel form of inequality, as those who do not succeed cannot argue that they were unlucky or kept down by those in power" (2005). While Kett (2012) notes that the ideal of merit was long perceived as a value in tension with equality, Littler argues that is has since become "an alibi for plutocracy" by "seizing the idea, practice and discourse of greater social equality" (2013: 69).

Seen through this lens, Big Data appears as a means to extend the logic of pervasive and skewed competition, paired with the rhetoric of impartiality, into further spheres of life. The idea that Big Data "works" – in the scope set by accounting realism – may, then, be much more terrifying than its possible failure. What if the real problem is not too little, but too much objectivity? What if the problem is knowing individuals and groups too well rather than not well enough? What if our Facebook Likes are, indeed, indicative of our future job performance? What if

8 "China 'social credit': Beijing sets up huge system", 26 October, 2015 (http://www.bbc.com/news/world-asia-china-34592186).

the saying "ignorance is bliss" holds true for society more generally, in the sense that *not-knowing* creates spaces where a plurality of practices and lives is possible because we cannot mechanically relate them to notions of performance and profit?

Conclusions

In concluding the somewhat experimental argument developed in this paper, I want to posit that the grand challenge of universally available data is not only surveillance understood as permanent policing, but also surveillance understood as permanent appraisal of compatibility with economic morality, the dominant value in contemporary society. It is certainly not new that differentiation, in all its forms, implies economic opportunity. But acting on the networks of difference that characterize our societies used to be costly. In many cases, it no longer is and, as a consequence, the many inequalities that persist in our societies are quickly becoming more consequential. Data mining and associated techniques have begun to read these inequalities from the perspective of operational goals. We would be well advised, however, to scrutinize not only the methods, but also the goals and the values that inform them.

The story of Rolf Buchholz, the current record holder in number of body piercings, makes for an instructive parable for the paradox of diversity. While Buchholz was denied entry into Dubai because airport staff feared he may practice black magic[9], he works apparently without any issues as a computer engineer with Deutsche Telekom in Germany. As long as he complies with the tenets of economic morality, the way he decorates his body is simply irrelevant. This is how contemporary capitalism liberates. But the moment Buchholz' job performance dips, his job is up for grabs. This is how contemporary capitalism disciplines. Even if the principle is not fully implemented, the direction is clear: diversity is welcome, as long as it does not interfere with the bottom line. In practice, however, there is one other thing keeping Buchholz from losing his job, namely Germany's labour laws. These laws are an example for deliberate limits to generalized economic morality and Big Data will increasingly force us to consider such limits.

As interested reading of reality, data mining makes it possible to assess economic utility in profound ways. In order to tame accounting realism, we therefore need to engage these techniques as deep, embedded, and performative forms of judgment, as modes of governing through measuring. Critique should strive to combine two different pathways. The first concerns the problems, limitations, and biases of the method. Here, Big Data's claims need to be critically examined. But the second needs to take these claims at face value and ask how the growing capacity to know society highlights the deep ambiguities in the dominant value system.

9 "World's most pierced man Rolf Buchholz barred from Dubai", August 17, 2014 (http://www.bbc.com/news/world-middle-east-28831106).

References

Allen, Ansgar (2012): "Life Without the 'X' Factor: Meritocracy Past and Present." In: Power and Education 4/1, pp. 4–19.

Andrejevic, Mark (2013): Infoglut, New York-Abingdon: Routledge.

Barocas, Solon/Selbst, Andrew D. (2016): "Big Data's Disparate Impact." In: California Law Review 104/3, pp. 671–732.

Beck, Ulrich (1986): Risikogesellschaft, Frankfurt am Main: Suhrkamp.

Boltanski, Luc/Chiapello, Ève (1999): Le nouvel esprit du capitalisme, Paris: Gallimard.

Burrell, Jenna (2016): "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." In: Big Data & Society 3/1.

Bourdieu, Pierre (1979): La distinction, Paris: Editions de Minuit.

Calders, Toon/Verwer, Sicco (2010): "Three naive Bayes approaches for discrimination-free classification." In: Data Mining and Knowledge Discovery 21/2, pp. 277–292.

Custers, Bart/Calders, Toon/Schermer, Bart/Zarsky, Tal (2013): Discrimination and Privacy in the Information Society, Berlin-Heidelberg: Springer.

Dahrendorf, Ralf (2005): "The Rise and Fall of Meritocracy." In: Project Syndicate, April 13 (https://www.project-syndicate.org/commentary/the-rise-and-fall-of-meritocracy).

Ekbia, Hamid et al. (2015): "Big Data, Bigger Dilemmas: A Critical Review." In: Journal of the Association for Information Science and Technology Volume 66/8, pp. 1523–1545.

Executive Office of the President of the United States, "Big Data and Differential Pricing", February 2015 (https://www.whitehouse.gov/sites/default/files/docs/Big_Data_Report_Nonembargo_v2.pdf).

Foucault, Michel (1976): Histoire de la sexualité 1. La volonté de savoir, Paris: Gallimard.

Giddens, Anthony (1994): "Living in a Post-Traditional Society." In: Ulrich Beck/Anthony Giddens/Scott Lash (eds.): Reflexive Modernization Politics. Tradition and Aesthetics in the Modern Social Order, Stanford: Stanford University Press, pp. 56–109.

Gitelman, Lisa (2013): "Raw Data" Is an Oxymoron, Cambridge, MA: MIT Press.

Goody, Jack (1977): The Domestication of the Savage Mind, Cambridge, UK: Cambridge University Press.

Haraway, Donna (1988): "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." In: Feminist Studies 14/3, pp. 575–599.

Hume, David (1739): A Treatise of Human Nature, London: John Noon.

Kett, Joseph F. (2012): Merit: The History of a Founding Ideal, Ithaca: Cornell University Press.

Kitchin, Rob (2014): "Big Data, New Epistemologies and Paradigm Shifts." In: Big Data & Society 1/1.

Kitchin, Rob (2016): "Thinking Critically About and Researching Algorithms." In: Information, Communication & Society 19/1, pp. 14–29.

Kosinski, Michal/Stillwell, David/Graepel, Thore (2013): "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." In: PNAS 110/15, pp. 5802–5805.

Latour, Bruno/Jensen, Pablo/Venturini, Tommaso/Grauwin, Sébastian/Boullier, Dominique (2012): "'The Whole is Always Smaller than its Parts' – a Digital Test of Gabriel Tardes' Monads." In: The British Journal of Sociology 63/4, pp. 590–615.

Littler, Jo (2013): "Meritocracy as Plutocracy: the Marketising of 'Equality' Under Neoliberalism." In: New Formations 80-81, pp. 52–72.

Lyon, David (2003): Surveillance as Social Sorting, London-New York: Routledge.

Porter, Theodore M. (1995): Trust in Numbers, Princeton: Princeton University Press.

Rieder, Bernhard (2016): "Scrutinizing an Algorithmic Technique: The Bayes Classifier as Interested Reading of Reality." Information, Communication & Society, 19/1, pp. 100–117.

Rudder, Christian (2014): Dataclysm, New York: Crown.

Simmel, Georg (1908): Soziologie, Berlin: Duncker & Humblot.

Sweeney, Latanya (2013): "Discrimination in Online Ad Delivery." In: Communications of the Association of Computing Machinery 56/5, pp. 44–54.

van Dijck, José (2014): "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." In: Surveillance & Society 12/2, pp. 197–208.