*Original Research Article*

BIG DATA & SOCIETY

Big Data & Society
January–June 2017: 1–13

DOI: 10.1177/2053951717700925
journals.sagepub.com/home/bds
SAGE

# Algorithmic rationality: Epistemology and efficiency in the data sciences

**Ian Lowrie**

**Abstract**

Recently, philosophers and social scientists have turned their attention to the epistemological shifts provoked in established sciences by their incorporation of big data techniques. There has been less focus on the forms of epistemology proper to the investigation of algorithms themselves, understood as scientific objects in their own right. This article, based upon 12 months of ethnographic fieldwork with Russian data scientists, addresses this lack through an investigation of the specific forms of epistemic attention paid to algorithms by data scientists. On the one hand, algorithms are unlike other mathematical objects in that they are not subject to disputation through deductive proof. On the other hand, unlike concrete things in the world such as particles or organisms, algorithms cannot be installed as the objects of experimental systems directly. They can only be evaluated in their functioning as components of extended computational assemblages; on their own, they are inert. As a consequence, the epistemological coding proper to this evaluation does not turn on truth and falsehood but rather on the efficiency of a given algorithmic assemblage. This article suggests that understanding the forms of algorithmic rationality employed in such inquiry is crucial for charting the place of data science within the contemporary academy and knowledge economy more generally.

**Keywords**

Epistemology, data science, algorithms, rationality, Russia, anthropology

## Introduction

Contemporary data are big. In our moment, data collection is ubiquitous, and their storage and analysis are challenges facing everyone from multinational industrial conglomerates and governments to lone social scientists and would-be digital entrepreneurs. The collection of data has been a central feature of governance since at least the development of statistics in the late 18th century (Foucault, 2009), and their manipulation has become an increasingly digital affair since the emergence of operations research and computing following the Second World War (Rose, 1991). However, the past two decades have seen an explosive growth in the sheer amount of data produced, tied to a radical metastasis of modalities of their collection and manipulation. During this same period, new forms of intellectual and pragmatic attentiveness to data have emerged alongside the quantitative growth, and political economic salience, of data (Halpern, 2015). While each discipline must tarry with data in its own way, there has also been the consolidation of a specific set of techniques and theories proper to the velocity, volume, and variety of contemporary data. In this article, I show how the collective intellectual practice of one group of data scientists is producing a coherent form of algorithmic rationality, irreducible to a congeries of practices or scraps of theory borrowed from mathematics or computer science. This rationality is inextricable from the emergence of big data infrastructures and deeply imbricated with our lately computational modernity (Bratton, 2015; Kitchin, 2014). Studying how such forms of expertise emerge, circulate, and interact with

Rice University, USA

**Corresponding author:**
Ian Lowrie, Department of Anthropology, Rice University, 6100 Main Street, MS-150, Houston, TX 77005, USA.
Email: lowrie.ian@gmail.com

other forms is a crucial component of sociology of contemporary knowledge.

This paper is based upon 12 months of ethnographic fieldwork with Russian data scientists, centered on the 2014–2015 academic year. This research focused on a new department of computer science at the Higher School of Economics, founded with substantial intellectual and material support from Yandex, a web infrastructure firm often referred to by my informants as "the Russian Google." During my time in Russia, I used loosely quota-based snowball sampling to identify and collect interviews from around 80 data scientists, predominantly from Yandex and the Higher School, but also other universities, institutes, and corporations with active research programs in data science. These interviews were divided more or less evenly between academic and industrial interlocutors, as well as across cohorts. Following IRB-approved protocol, the semistructured interviews were one to two hours in length, digitally recorded (unless the interviewee objected), translated and transcribed by the author. Interviews were supplemented by participant observation of classroom instruction, scientific seminars, business meetings, and industry events, as well as discourse analysis of scientific articles, publicity documents, and internal memoranda. The goal of this granular, quantitative observation was to produce an intimate picture of the intellectual milieu, career trajectories, and daily work practices of Russian data scientists.

Anthropologists are generally quite committed to the radical particularity of their field sites; the anthropology of expertise, specifically, has mostly focused on understanding the emergence of local knowledge (e.g., Downey, 2005; Murphy, 2015). Certainly, I chose to study data science in Russia because of its specific place within the local ecology of knowledge. Given the stagnation of Russia's extraction economy, elites from business, education, and government are working frantically to build the institutional and infrastructural supports for a home-grown Russian knowledge economy. This has been a difficult project, in part because of the historical privileging of the theoretical over the applied, and a relatively high barrier between academy and industry, in virtually all sectors of the Russian science system. In this context, one venture capitalist explained to me that pushing the development [*razvitie*] of data science was a clever way of transforming the "human resources" of excellence in fundamental mathematics into the "human capital" of trained computing professionals. Much of my fieldwork focused on the specifically local institutional, pedagogical, and professional transformations tied to this development project.

When I asked my informants what was *intellectually* unique about Russian data science, however, I was often met with blank stares; when I brought this up in conversation, one data scientist friend helpfully told me that this was likely because the question was both "rude" and "a little stupid." The academics and industrial researchers with whom I worked considered themselves full-fledged members of a robust, international scientific community (albeit somewhat self-consciously marginal ones). Although I have come to more or less share their belief that the operational core of data science is not subject to a great deal of geographical variation, it is important to be attentive to the institutional and cultural fields within which this core must inevitably become concrete. In expressing their view of themselves as deploying the epistemological and practical operations of data *science*, understood as a transnationally coherent form of inquiry, my informants also articulated a fierce and particular stance toward that science itself, and a local vision of what it means to be a scientist. This essay, then, should be read not as an exhaustive exploration of the epistemological dynamics of data science as a ready-made whole, but rather as a sympathetic investigation of their intellectual project, of their attempts to build a coherent and epistemologically sophisticated investigation of algorithms. That said, I think that the following analysis of local forms of algorithmic rationality offers a range of laterally portable analytics (Howe and Boyer, 2015) capable of shedding light on the social and intellectual processes structuring data scientific work in other milieux.

There has lately been a flurry of writing in the humanistic sciences about algorithms and data. Much of it has focused on technical transformations in the analysis and storage of data. This literature has produced critical insights into the design processes behind technological infrastructures subtending big data handling (Kockelman, 2013), and how these infrastructures interface with broad structural changes in governance (Rouvroy and Berns, 2013), consumer capitalism (Carah, 2017), and scientific inquiry (Leonelli, 2012). There is a complementary strand of inquiry that focuses instead on how big data technologies face users, both professional (Leonelli, 2014) and lay (Bucher, 2016). However, there has been little attention paid to the expertise of those who are building the epistemological frameworks within which these technologies are being developed and implemented. In focusing on this expertise, my work is most closely in dialog with those working to understand the epistemologies produced by and within big data architectures (Kitchin, 2014; Leonelli, 2012, 2014). This latter group, though, has been focused on the ways that the techniques of data science restructure existing fields of inquiry. My own project, by contrast, looks toward those knowledge practices that take algorithms, the mathematically

formalizable procedures for operating upon data, as their object of inquiry. This article, then, is not about everyone who calls themselves a ''data scientist'' (many of whom my informants felt to be simply technologists or analysts, concerned only with building machines or using their outputs). Nor is it about the epistemological status of the predictions, classifications, or cluster analyses produced by algorithmic assemblages. Rather, it focuses on the epistemological dynamics of an inquiry into algorithms themselves and the computational assemblages with which they are imbricated, as they operate among a specific group of scientists.

It should be said that, although they were committed to producing what they consistently called ''scientific'' knowledge about algorithmic assemblages, the primary business of most of my informants remained the identification of structures and tendencies in data, and the provision of actionable interpretations of such structures and tendencies. While the elegance of a particular algorithmic approach to a problem *tout court* might attract academic notice, even the most scholastic of practitioners are interested in technological questions of feasibility and expense, at least in terms of both the practical difficulties posed by its working out in code and the computational time required. Thus, the expertise of most data scientists I worked with bridges the conceptual and the pragmatic: to function as a competent data scientist requires familiarity with the mathematical operations in question, their implementations in code, and the hardware architectures underlying such implementations.

Given the constant pressure to minimize the expenditure of processing time, the rationalization of computational and storage architectures, and the system-internal standardization of data handling protocols is as much a component of an elegant solution to a problem as the algorithmic sophistication of the actual analysis. This focus on expenditure and output might lead us to conclude that we are dealing here with technology, which Lyotard calls ''a game pertaining not to the true... but to efficiency,'' rather than science proper, which tends to remain a cultural system or game dominated by the criteria of truth (1984: 44; see also Luhmann, 1996).[1] Certainly, we must concede that data science is dominated by the putatively technical ''criterion of efficiency,'' that it is governed by the quest to find the ''best possible input/output equation'' (Lyotard, 1984: 46). As with technology, we shall see that in data science a given '''move' is 'good' when it does better and/or expends less energy than another,'' rather than when it produces some ''truth'' about an algorithm (1984: 44).

My informants, however, are firmly committed to a vision of themselves as scientists, and of their work as science rather than engineering. Although they like it when things ''work,'' when their projects ''give people something they can use,'' they most often told me they were interested primarily in doing ''good science,'' in publishing ''serious papers,'' or in making ''discoveries.'' Certainly, this commitment to a ''science'' of algorithms is in part disciplinary boundary maintenance, an effort to establish the intellectual basis for the institutionalization of a new research field, separate from the existing and locally well-established domains of mathematics, software engineering, or operations research. Their insistence that they were focused on new knowledge about algorithms, rather than on building new machines or solving problems, is also conditioned by a Russian intellectual milieu that has long valued the abstraction of theoretical science and basic research above applied science and engineering. It is also part of constructing and enacting their ''socially recognised capacity to speak and act legitimately... in scientific matters,'' which in the contemporary global science system has consistently meant being able to convincingly claim that one is producing truth rather than merely building or using machines (Bourdieu, 1975: 19).

Although their striving to purify data science of the applied and the technological is undoubtedly structured by these institutional and cultural forces, I saw no reason to doubt the sincerity of their intellectual commitment to a profound understanding of the objective characteristics of algorithms, rather than their enlistment in the application of technological ''force'' (Lyotard, 1984: 46). However, in trying to separate their science from technology, they cannot abandon the criterion of efficiency altogether. This is in large part because, unlike the objects taken by many other sciences, algorithms can only be approached obliquely, through an evaluation of their functioning within extended computational assemblages. Further, the technical components of these assemblages themselves often emerge as objects of epistemic attention in their own right, as their function is implicated in the evaluation of the assemblage as a whole. It is also, however, due to the inevitable imbrication of these technical components with not only the mathematical abstraction of algorithms, but with real-world data and processes that are not fully compatible with the true/false code of the science system. In short, my argument is that my informants have repurposed the technological criterion of efficiency as an epistemological standard in the service of a new form of scientific inquiry.

In order to understand the forms of algorithmic rationality and epistemology at the core of this inquiry, then, this article begins with a brief look at the concrete knowledge practices that comprise these data scientists' professional toolkits. With this background established, it moves on to a description of the epistemology

that dominates their investigation of algorithms, before turning to a discussion of data itself—the material upon which these algorithms operate.

## Knowledge practices

With few exceptions, the core operations of data-scientific inquiry are borrowed. This should not be particularly surprising. Data science, situated as it is at the intersections of statistics, computer science, and mathematics, rather predictably sources many of its techniques from these disciplines. It is interesting to note that while most of these techniques and modes of expression have obtained a high degree of semantic abstraction and practical flexibility within data science, many were in fact developed in highly specific, applied domains of inquiry.

In taking up these techniques, however, it modifies them, combines them, breaks them down, and recombines them in ways that render them strange when viewed from the vantage point of their home disciplines. In both pedagogical and developmental contexts, traditional statistical modeling techniques percolate together with a range of other mathematical approaches to characterization and prediction, drawn from fields such as graph theory, complex systems analysis, and mathematical logic. The modes of inquiry dominant within data science, however, are drawn from a class of algorithmic approaches to classification and prediction known as machine learning. In machine learning, to quote a definition that was widely paraphrased by my informants: ''a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E'' (Mitchell, 1997: 2). To put it another way, machine learning aims to build programs that develop their own analytic or descriptive approaches to a body of data, rather than employing ready-made solutions such as rule-based deduction or the regressions of more traditional statistics. They do so through repeated trials, following each of which error is identified and fed back into the system, adjusting the approach for each subsequent trial. (These trials are *costly*: learning through iteration takes place in computational time that always threatens to spiral out of control; this is another crucial reason that data science cannot eschew the criterion of efficiency.)

Today, machine learning is a diverse congeries of algorithmic approaches, software implementations of such approaches, and hardware configurations designed to handle such implementations. There is a wide range of algorithmic approaches to core machine learning tasks, such as clusterization, classification, and prediction. These rest on quite diverse logical or mathematical foundations. Some of my informants at the Higher School, for example, are committed to advancing a machine learning approach based on a rather obscure branch of mathematical logic called Formal Concept Analysis (e.g., Poelmans et al., 2013). What they all share, however, is the commitment to the conceptual understanding of machine learning outlined by Mitchell. While some data scientists are partisans of one or another approach to such learning, most are flexible and willing to employ those techniques that they feel most appropriate to the task at hand.

This is in part because there is another, orthogonal sense in which data science might be said to be a congeries of practices. While we might think of ''datafication'' (Van Dijck, 2014) as a process of distilling the complexity of the world into a homogenous digital medium, the reality is that data-scientific inquiry not only comprises a motley of mathematical approaches but it also handles variegated forms of data. The presentation of homogeneity, I want to suggest, might ultimately be understood not as a property of data, but as an effect of the work of commensuration (Espeland and Stevens, 1998) done by data scientists to bring different forms of data together in constructing functioning algorithmic assemblages.

The formal differences between text and image are stark. Performing feature identification on a series of images, for example, requires an altogether different conceptual and computational apparatus than performing network analysis on a series of interlinked, textual objects. Although the final operations may be substantively the same (i.e. we might want to classify a certain type of object or predict how individual objects might change over time), my informants were clear that the procedures for managing the coupling between data and algorithm were wildly different, requiring a fair bit of knowledge about the specificities of extracting ''usable information'' from a given medium. This difference persisted, although at the conceptual rather than technical level, when moving between objects that, to me, initially seemed to be in the same class. In a series of talks on network analysis, for example, participants would take off on flights of theoretical fancy, discussing the finer mathematical points of clusterization, only to be grounded by the injunction to ''speak concretely [*govori konkretno*]'' about the real things and processes underlying the network: specific assertions, drawn from the graph-theoretical observation of social networks made up of people interacting with each other, simply rang absurd when applied to the world of hyperlinked Wikipedia pages, and vice versa.

I'd like to suggest that these different forms are not *only* the result of the ''second-order, distinctively human activity'' of categorization or organization

(Larkin, 2015); rather, some distinctions, such as between images and text, seem to inhere in the formal qualities of the data themselves. At the same time, I agree with Larkin that these formal qualities cannot be reduced to the characteristics of "technical devices" (2015), but rather emerge through a series of interpretive choices about how to apply such devices and structure information in ways that make it amenable to analysis. In short, data are produced by specific, human methods of capturing the world, and this technological process of capture carves the world up into various, more or less incommensurate forms of data. This is a stronger claim than the commonplace, acceded to by most data scientists I spoke with, that there is no such thing as "raw" data (Gitelman, 2013); form pushes us to think further about the internal ontological and conceptual differences inherent in the process of making the word into data. As a consequence of this incommensurability as much as of its intellectual history, the field of data-scientific inquiry is also organized into clusters of purpose-built algorithms for different areas of inquiry, clustered together under names such as "computer vision," "natural language processing," "social network analysis," and so on. The core epistemological principles and algorithmic procedures animating such algorithms may be laterally quite similar, but the specific features of their implementation in extended computational assemblages can be quite distinct.

Thus, data science is structured by two cross-cutting distinctions: between specific analytic approaches and between specific domains of problems. Many of the data scientists I spoke with tended to specialize according to *either* analytic approach or problem domain, leaving them more or less free to generalize along the other axis. That is to say, some might specialize in natural language processing, but graze freely from various fields of algorithmic approach. Others might be obsessively committed to a certain class of algorithms, eager to test their applicability to a huge range of concrete forms and sets of data. Data scientists' competence, however, tends to extend somewhat in either direction, with perhaps a preference for specialization according to type of data. Of course, we should not be taken to overstate the radicality of such specialization. Indeed, the business of data science, in many cases, is finding ways to smooth the frictions (Nafus, 2014) that build up as inquiry moves across different forms of data. My informants devoted a great deal of time and effort to building higher level computational abstractions capable of bringing together, for example, social network data with purchasing data, or dashcam video with natural language analysis of driver speech. Indeed, they routinely pointed to this ability to abstract over forms of data as a crucial and distinguishing component of their expertise; this work of commensuration is one of the key factors they identified as separating data science from earlier forms of computational analysis organized around a specific field, such as computer vision, or mathematical approach, such as Bayesian statistics.

In addition to these forms of scientific expertise, the data scientists I met were generally more or less competent technicians and builders of infrastructure, as well. They considered their professional competence to extend through the entire data pipeline, from collection through analysis to visualization. The importance of material infrastructures for contemporary social life has been well documented by anthropologists (Star, 1999; see Larkin, 2013 for a review) and media theorists (Kittler, 1999; Parks and Starosielski, 2015), and there are a growing number of ethnographies exploring how data infrastructures subtend and facilitate the deployment of expert and scientific knowledge (e.g., Leonelli, 2014). However, there has been less focus on the emergence of infrastructural arrangements as the objects of scientific attention, as what Rheinberger calls epistemic objects, in their own right.

Certainly, those who considered themselves data scientists spent less time tinkering with infrastructure than those who called themselves developers or engineers. This is in large part because well-developed, functional computing architectures will tend to "recede" from view, to function as merely the "technical, preparative subroutines" of data-scientific inquiry (Rheinberger, 1997: 21). However, the infrastructural arrangements underlying data scientific research also sometimes emerged as objects of intense scientific scrutiny in their own right. Most often, this happened when something broke down. One researcher described his ongoing contributions to Hadoop, an open-source information retrieval and storage framework, as the "selfish" result of "falling down the rabbit hole" when one of his algorithms failed to function properly with an off-the-shelf implementation of Hadoop. He just wanted to build something that would let him "get back to business." However, the emergence of the computing architecture as an object of epistemic concern does not only occur through the failure of infrastructural components to function as ready-to-hand tools. Another of my informants, for example, set herself a dissertation project focused primarily on developing big data architecture. She still considered herself to be a data scientist, rather than engineer, and her advisor agreed. She justified this self-identification through explicit reference to the criterion of efficiency: if data science is, after all, about learning how to create more efficient algorithms, then part of the scientific project is a "rational" approach to investigating the "relative efficiency" of various "technical things,

like storage,'' not just ''the mathematics'' behind data science.

I find this view compelling. In other disciplines, researchers are sometimes disinclined toward infrastructure building precisely because it is viewed as low-status, technical work, ancillary to their primary scholarly aims (Leonelli, 2014). Given the high degree of integration between computational architectures and software implementations of algorithmic approaches in data science, however, it seems impossible to extricate the ''scientific'' investigation of such approaches from the ''engineering'' challenges posed by the computational architecture, or even to clearly separate the algorithm from its assembly with data and computational architectures. As technical objects become objects of epistemic concern, while the algorithms themselves recede from view, they confirm Star and Ruhleder's assertion that infrastructure is ''fundamentally and always a relation, never a thing'' (1994: 253). For my informants, this refocusing, this inversion of the relationship between technical and epistemic objects did not fundamentally complicate their vision of themselves as scientists, intellectually and practically dedicated to producing new knowledge about the properties and function of algorithms; sometimes doing science means doing a bit of engineering (cf. Rheinberger, 1997: 31–32).

## Epistemological standards

The *sine qua non* of mathematical reasoning, historically, has been the work of proof. Ian Hacking argues that people have been drawn to mathematics in large part because ''they have *experienced* mathematics and found it passing strange'' (italics original, 2014: 84). This experience, mostly, has been the experience of proof: proof that smacks us with the inescapability of its conclusions, that strikes as an intellectual *coup de foudre*, rendering the unknown as the obvious. What makes this kind of proof so exciting and so different from other forms of rationality is that it seems to bypass the mucky business of empirical inquiry, to produce new knowledge directly from the mind itself.

However, although it makes use of axioms and certain procedures derived from the logical, deductive traditions of mathematics, data science operates according to a different set of epistemological standards and produces different truth experiences. The work of the ideal-typical proof-building mathematician or theoretical physicist crescendos in a series of Aha!-moments, of moments when her cognition snaps into clarity and rigor, when new knowledge has been produced out of raw thought. The data scientist, however, must content herself with the more pragmatic, somewhat exasperated ''Finally!'' when her system begins to produce useful outputs. (Anyone who has ever tinkered, though, will know that this somewhat paler form of cognition can be tied to its own, equally-intense affective investments.)

In data analysis, the goal is not proof of anything. The algorithms employed either function well or they do not. Whether we are engaged in classification, clusterization, time series forecasting, or the visualization of networks, the goal is decidedly not the demonstration of logical deduction from axioms. Neither, however, are we searching for the elegant congruence of mathematical models to physical structures or dynamic processes that is the hallmark of a great deal of applied mathematics (cf. Steiner, 1998). While such forms of reasoning can and sometimes do figure in discussions about the development of new algorithmic approaches to classes of problems, they ultimately play second fiddle to an altogether more pragmatic logic of feasibility, practicality, and computational time.

One way of thinking about the intellectual procedures of mathematics is as the socially mediated interpretation of structures (Wagner, 2010). These interpretations are made and contested in mathematical discussions, which are dominated by logical criteria appropriate to the work of proof and refutation (Lakatos, 1976). Data science, by contrast, might be considered to be the collective development of algorithmic approaches to data handling. Rather than through logical proof and refutation, these algorithms are developed by their assembly within computational architectures and the social evaluation of their inputs and outputs with respect to some real-world set of tasks. Thus, this form of inquiry cannot function except as a by-product of the primary implementation of such assemblages in the real business of increasing the efficiency of certain types of these tasks. This is why efficiency has emerged as the operative logic of evaluation shared by my interlocutors working in both academic and industrial data science. Efficiency as a term is here chosen in part for its capacity and abstraction; what counts as efficiency depends largely upon the concrete problem space in which the data scientists is working at any given moment. As such, unlike proof and refutation in mathematics, the discussions over efficiency that I witnessed frequently included stakeholders in the specific domains operated upon by algorithmic assemblages.[2]

This is perhaps most obvious when the problems are those of business: data science applications succeed when they lead to an increased rate in client retention, return business, or total cost of items in a customer's basket. One informant suggested that, in such cases,

> the scientific considerations of optimization *are* immediately *also* business concerns, right? It's an area where it's hard to separate out whether you are doing

> something in the name of doing good science, or in the name of doing good business, because they come together.

In most cases, an efficient application of industrial data science is one in which the cost of the application is outweighed by the increased revenue in which it results. This need not be by a huge margin. In sufficiently large operations, such as mobile telecom firms, marginal improvements such as an increase in monthly client retention of 0.5–1% through targeted communications can translate into millions of dollars in annual revenue. In many academic or scientific contexts, we can draw some relatively straightforward parallels to this pecuniary logic. For example, we might measure efficiency by the application of similarly external metrics: a decrease in false positives for cancer diagnoses, improvements in the rate at which the clusterization of social networks bring out relevant features, or an increase in the rate of meaningful results when topic sorting natural language texts.

Ultimately, however, the logic of efficiency in both cases is much more immanent to the experimental system than either of these lists of examples might indicate. It turns on the *cost* of implementing a given algorithm in any particular software and computational substrate. Computational time costs money, but it is also a finite resource to be managed by scientists whose access to quality server time is often institutionally constrained. With this in mind, this focus on efficiency might again seem to be a contingent result of the current situation of data science within the contemporary knowledge economy. We might object that, were data science relieved of its requirement to be useful, granted unlimited access to computational resources, and allowed to pursue its own ends as a "pure" science, we might begin to see an altogether different, more respectably "scientific" set of epistemological standards for the evaluation of algorithms to emerge. This, however, would be to somewhat miss the point. Efficiency is the epistemological code through which data science produces knowledge about algorithms. This is because algorithms of data science always, constitutively, address themselves (Peirce, 1935 [1902]) to a world of practical tasks.

The personal and professional interests of academic data scientists may tend toward the purely theoretical, in that their object of inquiry may be the algorithmic process itself, rather than the data being manipulated at any given time. Indeed, as scientists, the consistency of the algorithmic approach, its elegance and feasibility, is generally what is at issue in any given iteration of data scientists' inquiries. However, the investigation of such algorithmic epistemic objects requires their working out in and through concrete engagements with real data. Engaging real data means engaging delimited problem spaces, and producing practical, useful results through such engagements is the only feasible way to produce insight into the operations of the algorithmic propositions that ultimately form the epistemic core of such inquiry. Unlike the epistemic objects discussed by Rheinberger (1997), algorithmic processes cannot be installed as the *objects* of an experimental system. We cannot substantivize and produce knowledge about an algorithm as we can the protein synthesis pathways of a given model organism. They must be approached obliquely, through their application as technical components of systems of inquiry. The evaluation of their functioning *as* technical objects is the primary source of knowledge about algorithms *qua* epistemic objects. One cannot "directly" investigate an algorithm. Such an inquiry is a process that is essentially a by-product (to borrow from Elster (1981)).

This may seem quite abstract, so let us look at a concrete example. Viktor[3] is a Higher School professor who works in computer vision, which he describes as "basically information extraction from images." This is his central problem space: the development of algorithms that efficiently and reliably identifies and operates on the content of video and still images. Even more so than most academic data scientists that I spoke with, he is deeply invested in the practical outcomes of his theoretical research, one of which is road mapping:

> For example, you have a camera on a car, and the goal is to map all objects, all types of objects that are observable from that car, to the map. Starting from simple road infrastructure like traffic signs, and finishing to the trees, poles, traffic poles, road markings – everything that can influence the map.

Viktor is *invested* in the application of computer vision to a real world of problems. He thinks that mapping software is both useful and interesting and has hopes for a future with fewer car crashes and deaths on the road. However, as a theoretical data scientist, his ultimate goal is not this application; his output is, rather, improvements to the class of algorithms which might potentially come to be incorporated in public or commercial mapping software. This class of algorithms faces unique problems, because unlike in facial recognition, his other area of inquiry,

> The objects seem simple, but… the requirements of precision and reliability are much higher. For example, if you have a traffic maintenance service, they have to check that all traffic signs are in place and are clear and visible from the road… Every traffic sign should be accounted for because if you don't… you open the possibility for some traffic accidents, and the service

> will be held accountable. So, the precision should be much higher. And there are a lot of different looking traffic signs, so the number of classes are much larger than for example human face. And so the problem still exists. For some subtasks, for example if you want to detect speed signs only, then it's ok... But if you have two hundred types of objects, then it's still not solved. Probably next, maybe, five, maximum ten years, it will be solved. But right now, it's still not ready.

His goal, ultimately, is not to build a market-ready image recognition machine, tailored to the specific demands of road mapping. Rather, it is to "get the algorithms [he is] working with ready" to tackle the problem of mapping hundreds of classes of objects in real, living data. It is the class of algorithms that is the object of his epistemic attention. However, he cannot abstract the algorithms lying at the core of such applications *from* those applications. He cannot remove them from play and tinker with them on the sidelines—at least not without immediately sending them back out to the field. Without their ramification in extended, practical assemblages of code, data, and computational architecture, they would be inert objects. Unlike in some other branches of mathematics, where proofs and formulae are themselves open for discussion, here there is no internal metric by which to evaluate improvements to the algorithms at hand. Instead, the efficiency of a given system, measured through accurate identifications made in real-world data, is what Viktor uses to differentiate and evaluate various algorithmic approaches to computer vision. Precision and efficiency, here, are standards immanent both to algorithms and the practical tasks to which they address themselves.

The companies who feed him the grist for his scientific mill in the form of real data sets are aware that Viktor's primary goal is the investigation of new algorithmic approaches, not the development of industrial applications. They do not look to him either to do the work of classifying the objects in any specific data set, or to once and for all solve the problems facing their own, industrial algorithmists. There are no "formal requirements" attached to the data they send him: "they know that currently there is no perfect solution" and cannot expect immediate return on investment. This does not mean that they are disinterested; rather, they know that algorithmic science needs many people working through many successive iterations on real data sets in order to progress, and that "if they provide the data, then it's probable that the result will be sooner rather than later." More fundamentally, as stakeholders in the domain of road mapping, they are substantively involved in constituting the problem space in which Viktor is working, and consequently in discussions over, ultimately, what constitutes an efficient algorithmic approach within that space. Even Victor, though who more than many of the academics I spoke to is invested in the eventual, practical applications of his research—views practical trials as "testbeds" for theoretical improvements, rather than as ends in themselves. As he put it: he's a scientist.

In short, algorithms are the object of Viktor's inquiry. They are also, essentially, processes and tools. While we may be interested in knowing about them in the abstract, about how they work, and articulating a scientific approach to their development and implementation, we can only observe them in their functioning. While we can represent them in mathematical language, and further elaborate them in a specific software environment, without their enactment in a computational substrate and concrete articulation with a specific set of data and its attendant problem space, we cannot learn anything about them. They remain inert.

However, I want to suggest (more speculatively) that this is more robustly essential to the nature of algorithms than the preceding might imply. It seems to me that for data scientists, the *test* of an algorithm, of its elegance and conceptual unity, is not only found in its syntactical parsimony but in its practical economy. Ultimately, an algorithm is a crystallization and representation of cognitive processes that might otherwise be performed by humans.[4] Of course, few algorithms aim to directly model human forms of cognition, although this latter provides a rich metaphorical repertoire for commentary on their behavior. However, in confronting data, they all attempt to replace human cognition with cognition in a different computational substrate. There is nothing magical about algorithmic information processing: it simulates a particular form of searching through data. These algorithms are queries. As such, it is neither scientifically interesting nor aesthetically satisfying if we come up with an algorithm that searches everything, that brute forces problems by checking every possible outcome, pattern, or configuration. Rather, we're enamored by parsimonious approaches that let us search only try out a few organizational schemes or pathways, while still producing reliably valuable insights. In short, algorithmic economy is an evaluative standard that, like much in data science, straddles "materiality, mathematics, and metaphysics" (Kockelman, 2013: 49). However, it is a standard immanent to the fact of being an algorithm, rather than one imposed post facto by the practical constraints of data science as a form of inquiry; this is the case to the exact extent (and no further) that "function" is immanent to the fact of being an organ rather than being a framework imposed by the physiologist or anatomist.

## Material domain of operation

So far, I have been arguing that algorithms are the primary objects of data-scientific inquiry. Data science is also, however, quite obviously about data. When I asked data scientists to tell me what data was, however, the answers were surprisingly hollow: ''I can't tell you what data is, because I can't tell you what *isn't data*.'' ''Data is anything capable of being operated upon by an algorithm.'' The most ubiquitous answer, though, was that data was just another word for ''information.'' Information, however, is a notoriously polysemic term even within specific communities of practice. Labeling data as ''information'' doesn't ultimately tell us anything, doesn't do any particularly useful conceptual work – which fact my interlocutors would be the first to admit. Empirically speaking, I think we can here at least give data a synthetic definition both more constrained than and encompassing of the three versions I heard in my fieldwork: as they emerge within data science, data are digital traces of real processes or objects capable of being manipulated within computational environments.[5]

Now, data scientists already use the concept of digital trace to describe at least a particular kind of datum—namely, those that humans produce as they move through computational modernity. These ''fragments of past interactions or activities'' have lately figured quite prominently in the projects of consumer capitalists and the national security state alike (Reigeluth, 2014: 250). Giving a more capacious description of data, however, forces us to expand our conception of the trace: I want to think, with my informants, about data as being composed of traces not just of human activities, but of the world. That is to say, data can be understood as the marks that some section of the world makes when it moves through some recording field. The traces that emerge within these fields, then, are both signs that retain indexical relations of actual spatiotemporal contiguity to the objects for which they stand (Peirce, 1906), and the marks such objects produce within the graphematic space installed at the core of data scientific assemblages (Rheinberger, 1997: 104–112). That is to say, they are both ''real,'' in the sense of having definite relations to actual things in the world, and ''fictive,'' in the sense of being ordered and conditioned by the human-built sociotechnical systems within which they emerge. These two faces of data are complementary, emphasizing their ontological and ontic aspects, respectively.

It should be immediately apparent that, given this definition, everything is always on the cusp of being data. Given the vast array of techniques for shepherding analog material across the digital threshold and formalizing it as data, all that is required is a bit of money and some elbow grease. Indeed, my informants would often ask if I had ever considered applying algorithmic approaches to the analysis of my own data. At first, these questions struck me as rather odd, perhaps not least because of my own insecurities about the rigor of my chosen disciplinary approach to analysis. The first time this happened, I didn't really even know how to respond properly.

''I'm not really sure that I *have* data in the sense that you mean, Sergei,'' I demurred.

''Of course you do,'' he replied cheerily. ''It's just a very small set of messy and perhaps not so useful data, but we can always give it a try.''

He went on to explain that while my field notes were probably a hopeless cause, being too far gone into the wilds of literary text to tell us about anything other than my own idiosyncrasies, my interview data represented a ''small'' but perhaps interesting collection of ''natural language data, just waiting to be cleaned up and played around with.'' To integrate it with one of his ongoing investigations into a new class of natural language processing machines, we would just have to type up the transcripts, tag them with a little metadata, and see what we could see.[6]

Other informants were less optimistic. One initially expressed similar hopes for my corpus of interviews, until he found out that I was only planning on collecting around a hundred or so; ''generally,'' he said: ''it's not so interesting for us to work with any set of texts less than one or two orders of magnitude bigger than that.'' (For comparison, he and his team had a current project working on the set of all mathematics articles published in Russian over the previous decade.)

Beyond its form, then, the size of the data set matters. On the one hand, we can chart the growth of data sets quantitatively. Indeed, at the industry events that I attended it was *de rigueur* to have at least one slide chronicling the explosion in sources, types, and sheer amount of data available to be manipulated over the past 10 or 20 years. Academics and industry people alike would frequently and not-so-subtly brag about the magnitude of data to which they had access in both presentations and informal conversation. However, as data grow, the problem space within which they may be approached does *not* always expand linearly, but might instead undergo a series of rapid state changes, mutations, which signal genuine, qualitative shifts in their status as objects of inquiry.

That is to say, simple statistical approaches scale quantitatively with the size of the data set upon which they are operating. The reliability of the descriptions and inferences made by such approaches simply increases, often linearly, as new data points are added to the set. For many of the algorithmic approaches employed by data scientists, however, at certain levels

of scale new operations and approaches become epistemologically and practically feasible. Once certain thresholds of bigness are passed, for example, machine learning algorithms shift from producing nearly random clusterings or classifications to reliable, repeatable identifications. New graph-theoretical approaches to network analysis become available as the amount of data about relationships between members of the network increases.

Of course, the techniques that emerge as certain thresholds of quantity are passed *could* be applied to smaller sets of data. As one industrial algorithmist put it, there was nothing preventing him from calibrating a neural network using a training set of 20 pictures of his family. However, the results of a machine trained on such an impoverished set would not merely be poor, but profoundly "incoherent." As he explained, knees might be identified as faces, or the same image identified as different people over successive iterations of the machine. Conversely, the machine might "overtrain," becoming excellent at performing categorization on his family, but incapable of extending its analysis to new data. However, even in cases where such incoherencies resolve into coherence at a mathematically linear rate, there are nevertheless hidden thresholds, subject to case-by-case evaluation, at which new methods become appropriate. Part of data scientific expertise is being able to choose the appropriate algorithmic approach for the properties and scale of a given data set, precisely in the absence of firm, logical, generalizable rules.

These transformations speak at least partially to the Hegelian-Marxist dictum that beyond certain thresholds, changes in quantity have the uncanny ability to resolve themselves into changes of state or quality (cf. Bukharin 2013: 79–81). However, the emergence of coherence described here diverges from Hegel's model in that they are emphatically *not* due to the ontological qualities of data. As I've said, data objectively scale linearly and quantitatively: you have more, or fewer. You know more, or less, about each datum. The transformations I observed, instead, are products of the sociotechnical systems in which data are embedded, of the questions being asked of them, the techniques employed to extract answers to those questions, and the forms of rationality evaluating those answers. It is impossible to attribute such state changes in the problem space adjacent to any given algorithmic assemblage to either epistemological or pragmatic criteria alone; indeed, I have been arguing that for data science there is little difference between the two. In practice, as in the anecdote above regarding my own impoverished data sets, there are definite, discriminating, if ad hoc and disputable lines demarcating the lower quantitative bounds of the investigatory envelope for particular classes of solutions.[7] Charting the work of establishing such boundaries remains a critical task for the ethnography of data science.

## Conclusion

I have argued that there are two levels of inquiry operative in my informants' project of building a properly scientific data science: the exploration of concrete domains of applied tasks, on the one hand, and the epistemic inquiry into the nature and functioning of the algorithmic assemblages used in such exploration. My informants consistently stated their primary intellectual commitment to the latter. However, it is ultimately parasitic upon the former. Practically speaking: "algorithms are inert, meaningless machines until paired with databases upon which to function" (Gillespie, 2014: 169).

These databases must come from somewhere. When teaching students new techniques, for example, or testing out some new approach to a well-trod class of problems, the data scientists I met preferred to work with more or less public and well-characterized data sets, whose properties and contours have already been established. Their students learned to work graph-theoretical approaches to network analysis, for example, on classics such as the university karate club described by Zachary (1977). One computer vision researcher I interviewed hired Mechanical Turk workers to clean and prepare training sets of images celebrity faces from Google image search; most researchers I knew, however, disdained such sets, viewing them as at best teaching tools, test beds for exploratory techniques still in their infancy, or as opportunities for controlled comparison with colleagues' approaches to certain classes of problems. What they variously call "real-world," "lively," or simply "interesting" data sets almost universally entered their experimental assemblages from outside. While pedagogical activities and certain forms of algorithmic research do employ well-characterized sets of data analogous, perhaps, to certain well-characterized "model organisms" in laboratory biology, the real work of data science is done in its encounters with *new* data sets. There is a manifold of reasons for this tendency. Perhaps the most sociological is, simply, that data sets come with their own problem spaces, and that any significantly worked-over data set has had its problem spaces relatively well plumbed. However, more epistemologically speaking, novel data sets function as both the occasion for a rigorous test of existing algorithms and their implementations and as a crucial, empirically chaotic ground for the emergence of epistemic innovation. There is

something *lively* about the encounter between algorithms and new data.

It should be no great surprise, then, that data scientists, and their experimental systems, are voracious and omnivorous consumers of data; similarly, they are dedicated producers of structures that facilitate its circulation. Given the forms of inquiry proper to their discipline, it is also not unexpected to find these systems tending toward acting as quilting points for a variegated multitude of social processes. Their operations necessitate certain forms of input, drawing together a wide range material from outside of the experimental system, which must retain its connection to that outside world in order to fuel the analytic work of algorithmic assemblages. That is to say, the nature of this inquiry necessitates a relationship with what my informants frequently referred to as ''the world of applied tasks.'' As a consequence, they found themselves continually forced to develop relationships with business and government, to build the communicative infrastructures (Elyachar, 2010) that would ensure their continued access to sufficiently novel data. Indeed, some of them pursued careers in industry for no other reason that the ease of finding new data sets and attendant problem spaces to explore.

My informants, generally, were more committed to their theoretical work than to the resolution of specific problems. From this view, these communicative infrastructures were part of the ground state for the investigation of the relative intellectual merits of various algorithmic approaches. This perspective, however, is exceptionally misleading: it ignores that the criterion of efficiency dominating this investigation can never be entirely system internal. Rather, it must be coconstructed in dialog with other stakeholders in a specific problem area. What constitutes successful road mapping, cancer diagnosis, or financial prediction is ultimately an empirical question, requiring intellectual engagement with the social worlds of transportation, medicine, or business. For this reason, it seems to me unlikely that my informants will ever totally succeed in purifying their practice of its ''technological'' or ''applied'' character. Most of the dedicated researchers I met during my fieldwork, however, didn't spend a great deal of time fretting about this remainder. For even the most scholastic of them, efficiency is, simply and inescapably, the epistemological criterion best suited to the collective investigation of the function of algorithms, their possible uses, and their modes of assembly.

## Acknowledgement

I would like to thank my informants for their willingness to share and explain their world. Andrei Kozhanov and Leila Ashurova at the Higher School of Economics provided invaluable logistical and collegial support. James Faubion, Dominic Boyer, Marcel LaFlamme, Derek Woods, and the three anonymous reviewers at *Big Data & Society* provided invaluable input on earlier drafts of this article. Of course, all errors and oversights remain are my own.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for the research underlying this article was provided by the Wenner-Gren Foundation for Anthropological Research, American Councils for International Education, and the Rice Social Science Institute. Support during the writing phase was provided by the National Academy of Education and the Spencer Foundation.

## Notes

1. Of course, there are other ways of playing a recognizably scientific game. Gelfert (2016), for example, argues that scientific work with models is dominated by a logic of conceptual and practical usefulness rather than of truth. Further, the distinction between science and technology is itself contested by second-order commentators such as actor-network theorists, who prefer to use the term ''technoscience'' to describe the mixed networks of knowledge and engineering practices they study (e.g., Latour, 1987), and new materialists, who prefer to speak of ''matterings'' produced by technological and conceptual assemblages (e.g., Barad, 2007). However inescapably mixed these networks and matterings may be at the level of quotidian practices, though, it is impossible to deny the continued social reality of the divide between science and technology at the level of communicative systems (Luhmann, 1996), or its importance for building institutions, careers, and concepts (Bourdieu, 1975). My argument here primarily charts the impact of this conceptual distinction on practice; certainly, for my informants, it was a difference that made a difference. I am not committed, here, to defending its ontological or practical inviolability at all resolutions of analysis. That said, I do agree with Rheinberger that, at the very least, that the concept of technoscience ''needs to be handled with caution,'' lest its latent suggestion of ''an identity of science and technology'' occlude the fundamental specificity of the ''research process'' which, after all, is the topic here at hand (1997: 31).
2. When computer science is described as the scientific study of algorithms (as does, influentially, Knuth, 1974), it is generally limited to the system-internal dynamics of their operations upon data, rather than a more comprehensive investigation of their function within these extended assemblages. Data scientific efficiency, then, differs from

the concept operative, and occasionally dominant, in computer science in that the object of evaluation is not only the performance of the computational substrate, but the entire sociotechnical system within which the algorithmic assemblage is embedded.

3. All names in this article are pseudonyms.
4. There is no coherent, technical definition of "algorithm" afforded by the computer science or mathematics literature. It seems to my mind that this is, in part, because of the capacious yet ephemeral nature of the cognitive processes they attempt to capture and fix in formal, logical description. See, for example, the debates in Gurevich (2011) and Moschovakis (2001). For discussions definitional efforts from the secondary literature, see Hill (2015) and Seaver (2013).
5. In so doing, I am somewhat short-circuiting ongoing debates in the second-order sociology and philosophy of science about the definition of data (Leonelli, 2015) and information (Floridi, 2005). My goal is not to deny the importance of such debates for either second- or first-order observers of epistemologies and infrastructures. Rather, I aim to provide a definition that both cleaves closely to the empirical and discursive realities of the community under discussion here, while expanding the conceptual reach of the existing use of "traces" and "tracing" within this second-order literature.
6. Interestingly, although they viewed their algorithm-facing research as simply orthogonal to anthropological investigation, my informants were often willing to consider the goals of my own, qualitative form of analysis as epistemologically consonant with at least the outputs of their algorithmic systems. As one put it, both strove to identify "tendencies and structures" in a "messy" world through selecting specific strands of it for careful dissection and comparison.
7. Of course, new techniques also emerge due to qualitative changes in the data themselves. For example, the introduction of semantic or ontological information about the edges of a graph—of relations between nodes—rather than just information about the nodes of the graph, radically changes the quality of the data about those nodes and produces attendant changes in problem spaces of graph theoretical inquiry.

## References


Barad K (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Bourdieu P (1975) The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information* 16(4): 19–47.

Bratton B (2015) *The Stack: On Software and Sovereignty*. Cambridge: MIT Press.

Bucher T (2016) The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication and Society* 20(1): 30–44.

Bukharin N (2013) *Historical Materialism: A System of Sociology*. London: Routledge.

Carah N (2017) Algorithmic brands: A decade of brand experiments with mobile and social media. *New Media and Society* 19(3): 384–400.

Downey G (2005) *Learning Capoeira: Lessons in Cunning from an Afro-Brazilian Art*. Oxford: Oxford University Press.

Elster J (1981) States that are essentially by-products. *Social Science Information* 20(3): 431–473.

Elyachar J (2010) Phatic labor, infrastructure, and the question of empowerment in Cairo. *American Ethnologist* 37(3): 452–464.

Espeland W and Stevens M (1998) Commensuration as a social process. *Annual Review of Sociology* 24: 313–343.

Floridi L (2005) Is semantic information meaningful data? *Philosophy and Phenomenological Research* 70(2): 351–370.

Foucault M (2009) *Security, Territory, Population: Lectures at the Collège de France 1977–1978*. New York, NY: Macmillan.

Gelfert A (2016) *How to Do Science with Models: A Philosophical Primer*. New York: Springer.

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge: MIT Press, pp. 167–194.

Gitelman L (ed.) (2013) *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press.

Gurevich Y (2011) What is an algorithm. Microsoft Technical Report MSR-TR-2011-116. Redmond, WA: Microsoft Research.

Hacking I (2014) *Why Is There Philosophy of Mathematics at All?* Cambridge: Cambridge University Press.

Halpern O (2015) *Beautiful Data: A History of Vision and Reason Since 1945*. Durham, NC: Duke University Press.

Hill RK (2015) What an algorithm is. *Philosophy and Technology* 29(1): 35–59.

Howe C and Boyer D (2015) Later analytics and portable theory. In: Marcus G, Faubion J and Boyer D (eds) *Theory Can Be More Than It Used to Be: Learning Anthropology's Method in a Time of Transition*. Ithaca, NY: Cornell University Press, pp. 15–38.

Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12.

Kittler FA (1999) *Gramophone, Film, Typewriter*. Redwood City, CA: Stanford University Press.

Knuth D (1974) Computer science and its relation to mathematics. *The American Mathematical Monthly* 81(4): 323–343.

Kockelman P (2013) The anthropology of an equation: Sieves, spam filters, agentive algorithms, and ontologies of transformation. *HAU: Journal of Ethnographic Theory* 3(3): 33–61.

Lakatos I (1976) *Proofs and Refutations*. Cambridge: Cambridge University Press.

Larkin B (2013) The politics and poetics of infrastructure. *Annual Review of Anthropology* 42: 327–343.

Larkin B (2015) Form. Available at: http://culanth.org/fieldsights/718-form (accessed 10 December 2015).

Latour B (1987) *Science in Action*. Cambridge, MA: Harvard University Press.

Leonelli S (ed.) (2012) "Special section: Data-driven research in the biological and biomedical sciences". *Studies in History and Philosophy of Science Part C* 43(1): 1–87.

Leonelli S (ed.) (2014) "What difference does quantity make? On the epistemology of big data in biology". *Big data & Society* 1(1): 1–11.

Leonelli S (ed.) (2015) "What counts as scientific data? A relational framework". *Philosophy of Science* 82(5): 810–821.

Luhmann N (1996) *Social Systems*. Redwood City, CA: Stanford University Press.

Lyotard J-F (1984) *The Postmodern Condition: A Report on Knowledge*. Minneapolis, MN: University of Minnesota Press.

Mitchell TM (1997) *Machine Learning*. Boston, MA: McGraw-Hill.

Moschovakis Y (2001) What is an algorithm?? In: Engquist B and Schmid W (eds) *Mathematics Unlimited—2001 and Beyond* Berlin: Springer, pp. 919–936.

Murphy K (2015) *Swedish Design: An Ethnography*. Ithaca, NY: Cornell University Press.

Nafus D (2014) Stuck data, dead data, and disloyal data: The stops and starts in making numbers into social practices. *Distinktion: Scandinavian Journal of Social Theory* 15(2): 208–222.

Parks L and Starosielski N (eds) (2015) *Signal Traffic: Critical Studies of Media Infrastructures*. Champaign, IL: University of Illinois Press.

Peirce CS (1906) Prolegomena to an apology for pragmaticism. *The Monist* 16(4): 492–546.

Peirce CS (1935 [1902]) Logic as semiotic: The theory of signs. In: *Philosophical Writings*. New York: Dover.

Poelmans J, Ignatov DI, Kuznetsov SO, et al. (2013) Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications* 40(16): 6538–6560.

Reigeluth T (2014) Why data is not enough: Digital traces as control of self and self-control. *Surveillance and Society* 12(2): 243–254.

Rheinberger HJ (1997) *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Redwood City, CA: Stanford University Press.

Rose N (1991) Governing by numbers: Figuring out democracy. *Accounting, Organizations and Society* 16(7): 673–692.

Rouvroy A and Berns T (2013) Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux* 1: 163–196.

Seaver N (2013) *Knowing algorithms. Media in Transition 8*. Cambridge, MA, April, 2013.

Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391.

Star SL and Ruhleder K (1994) Steps toward an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems. In: *Proceedings of the 1994 ACM conference on computer supported cooperative work*, Chapel Hill, North Carolina, USA, 22–26 October 1994, pp.253–264. New York: ACM.

Steiner M (1998) *The Applicability of Mathematics as a Philosophical Problem*. Cambridge, MA: Harvard University Press.

Van Dijck J (2014) Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society* 12(2): 197.

Wagner R (2010) For a thicker semiotic description of mathematics. In: Löwe B and Müller T (eds) *Philosophy of Mathematics: Sociological Aspects and Mathematical Practice*. London: College Publications, pp. 361–384.

Zachary WW (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33(4): 452–473.