# Data Science on the Ground: Hype, Criticism and Everyday Work

Daniel Carter
School of Information, University of Texas at Austin
1616 Guadalupe Suite #5.202
Austin, Texas
78701-1213
carter.daniel.w@gmail.com
Phone: 281-910-4492
Fax: 512-471-3971

Dan Sholler
School of Information, University of Texas at Austin
1616 Guadalupe Suite #5.202
Austin, Texas
78701-1213
sholler.daniel@gmail.com
Phone: 512-865-7176
Fax: 512-471-3971

Corresponding author: Daniel Carter

## Abstract

Modern organizations often employ data scientists to improve business processes using diverse sets of data. Researchers and practitioners have both touted the benefits and warned of the drawbacks associated with data science and big data approaches, but few studies investigate how data science is carried out “on the ground.” In this paper, we first review the hype and criticisms surrounding data science and big data approaches. We then present the findings of semi-structured interviews with 18 data analysts from various industries and organizational roles. Using qualitative coding techniques, we evaluated these interviews in light of the hype and criticisms surrounding data science in the popular discourse. We found that although the data analysts we interviewed were sensitive to both the allure and the potential pitfalls of data science, their motivations and evaluations of their work were more nuanced. We conclude by reflecting on the relationship between data analysts' work and the discourses around data science and big data, suggesting how future research can better account for the everyday practices of this profession.

## Introduction

Data science and big data have both received significant hype in recent years. For example, *Harvard Business Review* declared data scientist the "sexiest job of the 21st century" (Davenport & Patil, 2012) and others tout the potential business impacts of big data (e.g., Chen, Chiang, & Storey, 2012; Manyika et al., 2011). However, this hype has also been accompanied by criticisms, which call into question the effectiveness (e.g., Ross, Beath, & Quaadgras, 2013) and the assumptions and implications (e.g., boyd & Crawford, 2012) of these approaches.

Data science is an emerging profession that leverages programming and statistical skills to solve business problems (Harris & Merotra, 2014). A related concept, big data refers to the storage, processing and analysis of large amounts of data in order to support existing processes and generate new insights (see, for example, Hashem et al., 2015; LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2010). While data science and big data are distinct concepts, data science approaches often rely on the large data sets to which big data refers, and big data approaches similarly rely on the kind of analysis that data scientists perform. Further, data science and big data are difficult to separate in commentaries found in popular media, with both proponents and critics tending to focus on the amount of data and the potential impact of the analysis that is performed with it. Because we are interested in the arguments that surround these approaches as well as the everyday work practices and values of the analysts working in this area, we use the phrase "data science and big data approaches" as a way to acknowledge these connections.

While proponents and critics of data science and big data approaches have largely focused on potentials and limitations at a very high level, few studies focus on the individuals doing the analysis work that these approaches entail. Workers with job titles like "data analyst," "data scientist" and even "hacker" perform the everyday tasks that comprise data science and big data approaches; they are, in a sense, the boots on the ground, gathering and formatting data, constructing models and tools and communicating their work to members of their organization and clients. As Ribes and Jackson (2013) point out, behind the imagined image of data-driven science are individual actors: "a squeamish student taking spit samples," or "a frustrated field technician recalibrating a vandalized weather monitoring station for the third time that month" (p. 152). While much data used by data science and big data approaches is automatically-collected by by sensors networks and other devices that record behavioral traces, the concepts of data science and big data can still overshadow the individual analysts who clean these data sets, experiment with methods, and make political choices about the presentation of results.

This study connects the high level hype and criticism around data science and big data approaches with qualitative descriptions of data analysts' values and experiences. We do not intend to advocate for data analysts or to support critics or proponents of these approaches. Instead, we use these existing arguments to structure our analysis of a kind of work that is increasingly popular in modern organizations yet not well understood. We believe that a balanced understanding is necessary if researchers wish to build theory through the study of this complex work.

In building a nuanced perspective of data science work, we seek to answer the following question: How do claims made in popular and academic discourses around data science and big data approaches compare with the everyday work practices, motivations, and evaluations of data analysts?

In the next section, we review the broad arguments of both proponents and critics of data science and big data approaches. We then describe our interview-based approach for comparing hype and criticisms with work being done "on the ground." Next, we present the results of our qualitative analysis of semi-structured interviews with 18 data professionals, focusing on the personal and business motivations for data science work and the criteria upon which data science work is evaluated. We then compare themes from the popular discourse around data science and big data approaches with the themes developed through our analysis and conclude by reflecting on the relationship between data analysts' work and the discourses around data science and big data, suggesting how future research can better account for the everyday practices of this profession.

## Background

Descriptions of data scientists often emphasize the diverse skills needed to fill these positions. Hempel (2012), for example, calls for "one part mathematician, one part product-development guru, and one part detective." Data scientists are described as "a hybrid of data hacker, analyst, communicator, and trusted advisor" (Davenport & Patil, 2012, (p. 73) and as "better programmers than most statisticians and better statisticians than most programmers" (Harris 2014). Other descriptions focus less on specific skill sets and instead emphasize data scientists' need for certain personality characteristics such as curiosity (Davenport & Patil, 2012, p. 73), skepticism (Ferguson, 2013a) and creativity (Ferguson, 2013b). These broad descriptions have led to a perception of data scientists as rare and mysterious. One CEO, quoted in the *Sloane Business Review*, responds to these descriptions: "They're unicorns; you can't find them. Or there are a very limited number of people that fit the criteria" (Ferguson, 2013b).

Indeed, data scientists are in high demand. According to a report by the McKinsey Global Institute (2011), the U.S. will face a shortage of between 140,000 and 190,000 data analysts and 1.5 million executives and staff who understand data analysis by 2018. Current research shows the demand to be already felt in some sectors (Davenport & Patil, 2012), and a survey of IT firms reports that the most significant challenge to big data projects is finding analytics talent (Kaskade, 2013).

Kandel et al. (2012) provides one of the few empirical investigations aimed at understanding the processes and technologies involved in data science from the analyst's perspective. Drawing on interviews with 35 analysts, the authors analyze the processes (e.g., locating data, verifying assumptions and generating reports) and tools used (e.g., SQL, Java and Excel) and describe three categories of analysts: Hackers, or those who are proficient in many tools but limited in the statistical models they employ due to the larger data sets with which they work; scripters, or those who work within a software package to perform sophisticated analysis of data that was delivered in an expected format; and application users, or those who primarily work within a spreadsheet-like application such as Excel.

As the majority of existing research on data scientists is oriented to business and management audiences, the focus has often been on informing hiring decisions and organizational positioning (e.g., Harris & Merotra, 2014). These types of analyses are useful for integrating data science and big data approaches into existing organizational structures but are less fruitful for understanding the actual experiences of data professionals and the implications of their work. Moreover, research from this perspective risks blindly accepting that data science and big data approaches are inherently good or necessary for business processes. To provide a more nuanced perspective, we review a growing body of criticisms below. Although these criticisms sometimes come from popular media and are often not the product of empirical research, they represent an acknowledged perspective that can be compared with the practices and motivations of those doing data analysis work.

## Criticisms of Data Science and Big Data Approaches

Along with the hype surrounding data science and big data, some have questioned the effectiveness of these approaches as well as their underlying assumptions and implications. Richards and King (2013) argue that proponents of data science and big data have not considered the approaches' drawbacks such as threats to privacy and the potential to privilege governments and large corporations over individuals. Marcus and Davis (2014) highlight additional limitations such as the inability to explain which correlations found in data are meaningful.

We focus here on three broad criticisms consistently raised in relation to data science and big data approaches: 1) that proponents of these approaches see no need for theory or for more traditional methods of data collection and analysis, 2) that proponents of these approaches see them as objective and disinterested when this is not the case and 3) that these approaches introduce issues around privilege and access to data.

### *Theory and Traditional Methods*

Critics have claimed that proponents of data science and big data approaches ignore the need for theory or more traditional methods. As Mayer-Schönberger and Cukier (2013) put it, what matters is if a statistical model can predict an outcome—not why that outcome occurred (p. 4). Anderson (2008) represents the object of this critique well:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. ... With enough data, the numbers speak for themselves.

From this perspective, generalizable models of how the world works are not needed because large amounts of data can be analyzed in relation to specific contexts and questions. This perspective also implies that domain knowledge and specialization become irrelevant in the face of large data sets and analysis.

However, critics of this perspective point out that data science and big data approaches are sometimes less effective than traditional approaches (Lazer, Kennedy, King, & Vespignani,

2014) and that they still rely on domain experts to contextualize results and draw attention to what is not explained by data (Graham, 2012). boyd and Crawford (2012) draw out the implications of relying on large-scale data analysis at the expense of all other methods, describing a new "system of knowledge" that replaces those that came before and that contains its own, largely unexamined constraints and limitations (p. 665). Berry (2011) makes a similar point, describing these approaches as depoliticized ways of knowing that bypass traditional mechanisms (such as theory and philosophy) developed to ensure rational thought (p. 8). While the hype around data science and big data approaches suggests that they supplant all previous methods and theories, these critics suggest that doing so reduces the effectiveness of the analysis and also ignores larger epistemological implications.

### *Objectivity and Disinterest*

Critics of data science and big data approaches have also argued that these approaches claim objectivity and disinterest when, in fact, they are subjective and rely on interpretation. For example, Jurgenson (2014) argues that proponents of these approaches tend to downplay the expertise of the researcher, instead foregrounding the data itself. Jurgenson specifically cites OkCupid co-founder and prominent data scientist Christian Rudder's insistence on his "mediocre statistical skills," noting that this makes truth "no longer a matter of analytical approach and instead one of sheer access to data."

As boyd and Crawford (2012) argue, there is a tendency to claim computational work as based in fact and not interpretation—however, "as soon as a researcher seeks to understand what [a model] means, the process of interpretation has begun" (p. 669). Drucker (2010) similarly points out that the "data" of data science and big data are not objective facts but instead are "capta," pieces of information gathered in response to researchers' goals and assumptions. This criticism of the created nature of data has been repeated by others (e.g., Boellstorff, 2013; Gitelman & Jackson, 2013), including Bowker (2005), who quips, "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care" (p. 183-84).

Additionally, the technologies data analysts use to accomplish their work may add another layer of interpretation. Features of technologies can act as affordances or constraints on particular actions (Leonardi & Barley, 2008), influencing the way data is collected, manipulated, and analyzed. Further, these technologies are embedded in organizational and social contexts (Orlikowski & Scott, 2008). Researchers interested in work practices have noted that, when conducting and communicating work using technology, workers' choices are influenced by social factors such as peer use of technology and expectations about audience (Barley, Leonardi, and Bailey, 2012).

### *Privileged Access to Data*

Related to the above discussions of objectivity and disinterest are criticisms that data science and big data approaches rely on privileged access to data. For example, Osberg (2014), in a review of Rudder's book *Dataclysm*, points out that although Rudder downplays his individual technical skills, he relies on personal relationships with tech company executives to gain access to the proprietary data required for his analysis.

This privileged access to data leads boyd and Crawford (2012) to claim that big data approaches create “a new kind of digital divide: the Big Data rich and the Big Data poor” (p. 674). While Mayer-Schönberger and Cukier (2013) point to an analyst who created a model to predict price change in airfares using data gathered from airlines’ public websites (p. 3), suggesting that data science and big data approaches democratize research, boyd and Crawford’s argument is that who you are matters and that the position of the individual researcher cannot be ignored. And, as Manovich (2012) points out, researchers are better positioned if they work for a large social media corporation—while public APIs offer limited access to data, full data sets are available only to those working within companies or who have the resources to purchase data from them. Further, these datasets are collected for very specific purposes and thus are not always suitable for making claims about real-world phenomena, such as social behavior (Ruths & Pfeffer, 2014). As a result, these critics claim that data science and big data approaches place limits on who can conduct this research and what questions can be asked.

## Methods

Based on this review of popular media and academic literature around data science and big data approaches, we constructed an interview protocol intended to elicit the everyday work practices of data analysts, as well as the motivations of their work and the criteria by which it is evaluated. Sample questions are included below:

- What was the last piece of analysis that you played a part in?
- When you are finished with a task, how do evaluate whether you were successful or not?
- What was the final product of the analysis? Was it presented to others?

### Data Collection

In order to define our ideal participant profile, we employed Kandel et al.’s (2012) definition of a data analyst as “anyone whose primary job function includes working with data to answer questions that inform business decisions.” In line with this definition, we sought individuals from a variety of industries with varying levels of experience extracting, analyzing, and using data in an organizational setting. We recruited participants through professional networks on LinkedIn and Meetup.com, as well as through snowball sampling. This method allowed us to locate individuals who self-identified as data analysts, either through their job title of through their participation in groups related to data analysis and data science.

Following Yin’s (2013) technique, we performed 17 semi-structured, 20- to 50-minute interviews with 18 individuals. One interview was conducted with two analysts from the same organization; all others were conducted individually. Interviews were conducted in person, via webconferencing technologies, or via phone conferencing. Each interview was recorded and transcribed, along with any notes or supplementary materials. We then loaded the files into Atlas.ti qualitative coding software for analysis.

### Participant Profile

The participants we interviewed held a variety of job titles, worked in organizations of different sizes and types, and had various levels of experience working in data analysis. Participants self-reported job title, highest degree, degree field, and experience. We determined organization size by reviewing the organizations' LinkedIn profiles and categorizing each as small (under 200 employees), medium (200-1,000 employees), and large (over 1,000 employees). Over half of our participants worked in small start-up organizations, which served as both an opportunity (discussed below) and a limitation (discussed in the conclusion) for this study. An overview of the participant profile is shown in Table 1.

Insert Table 1 here

### *Job Titles and Roles*

The data analysts we interviewed held a variety of job titles. Half of our participants held the title "data analyst" or "data scientist." We included participants with other job titles because, although not formally "analysts," these participants performed work that aligned with Kandel et al.'s (2012) definition. These workers held titles that reflected specialized duties or expertise (e.g., GIS analyst) or that designated formal authority in the organization (e.g., Head of Research and Development), allowing us to examine work at various levels of the organization.

Participants served in a variety of organizational roles—interns, mid-level professionals, and managers—yet all analyzed data to inform business decisions. The most common tasks involved in these roles included data gathering, analysis and presentation of results to peers or managers. Because our participants worked in a variety of industries (see below), tasks within each of these stages varied. For example, some analysts used proprietary data and technologies, while others worked with public data and open source technologies. Some worked with others on a day-to-day basis, while others communicated less frequently. We believe that this balance in our sample facilitated a deeper understanding of the variety of positions data analysts hold and the sorts of work they perform in modern organizations.

The products of our participants' work varied based on industry, organizational role, and a number of other factors. However, at all levels of the organization and across these industries, data analysts in our sample interpreted the results of their work for presentation to others. The output of their work was often a presentation containing visualizations from the analysis they conducted. Sometimes these presentations resulted in the implementation of an algorithm (e.g., a new way to order search results or predict supply shortages). At other times, presentations of analysis resulted in less direct business decisions such as how to market a product.

### *Organizations and Industries*

Ten of our eighteen participants worked in small organizations with fewer than 200 employees. The limitations related to this factor are discussed at the end of this paper. However, we used this characteristic to benefit our study by asking questions about communication and evaluation of data work. We believe that interviewing individuals from smaller organizations allowed for richer and more accurate recall of interactions with peers and made these interactions easier to trace. Furthermore, our sample includes data analysts working in a number of important industries, including finance, education, government, security and energy. We identified commonalities in ideas, experiences, and attitudes of data analysts across these industries, indicating that data analysis work looks similar in many key sectors of the economy.

### *Degree, Degree Field, and Experience*

Little is known about the educational backgrounds of data analysts, a lack that initially served as a motivation of our investigation. We interviewed analysts with bachelor's, master's, and Ph.D.'s in a number of fields. Unsurprisingly, a large portion of our sample held degrees in computer science and statistics, fields that have obvious connections to data analysis. Our participants had varying levels of experience with data analysis work. Some considered their educational background, such as serving as a research assistant in graduate school, as relevant experience. Others believed educational background was only tangentially related to their current work and thus did not report education as years of experience. Overall, our sample represented practitioners who are early in their careers (1-3 years), who have some experience (4-6 years), and who have many years of experience (6+ years). This diversity of experience levels offered insight into the experiences of both newcomers and veterans in the data science community.

## Data Analysis

Upon completion of observations and interviews, we imported our interview transcripts into ATLAS.ti software for review and coding. The analysis process comprised three steps: iterative readings of the interview transcripts, open coding, and selective coding (Strauss & Corbin, 1990). In the first step, we read through our entire corpus of interview transcripts to identify salient points in the data, those quotes and actions that demanded our attention to reach deeper understanding of their meaning. We then discussed the points we identified and generated preliminary code lists to probe the data for themes. We initially identified 54 themes, covering organizational and team characteristics, types of interactions, and forms of results. We then discussed these themes and opportunities to combine codes, refine them, and add a greater level of specificity. After considering these themes in light of our reading of the literature, we revised the coding list and developed a final scheme focusing on motivations and evaluations, presented in Table 2.

Insert Table 2 here

Following these themes, we performed selective coding to identify episodes in which motivations or evaluations were discussed.

# Results

Our analysis yielded several themes with respect to motivations of data analysts. These themes often reflect an awareness of popular hype and criticisms of data science and big data approaches. However, in responding to questions about their motivations and evaluations, the data analysts provided a more nuanced picture of their work practices. Below, we describe the common motivations and evaluation criteria we identified before discussing the consistencies and contradictions with hype and criticisms surrounding data science and big data approaches.

## Motivations

To begin to compare the criticisms of data science and big data approaches with the experiences and values of data analysts, we outline the three most common motivations for performing analysis reported by our respondents: creativity and curiosity, interesting problems, and data availability. Each of these motivations is distinct, but the relationships between them could offer valuable insight into daily work practices involved in data science. In particular, each category of motivations demonstrates a tendency for data analysts' practices to be exploratory and iterative, even when the end goal of the analysis appears straightforward.

### *Creativity and Curiosity*

Nine participants mentioned creativity and curiosity as motivators for performing analysis work. The majority of these participants worked in small organizations. These participants often described data analysis work as ideal for individuals who are always curious to learn new kinds of data, techniques and tools. For example, a participant with over ten years of experience in a range of industries noted:

> I think the most valuable skill is just the ability to kind of keep exploring and keep learning, right? Whether you're exploring the data or learning about the data or, you know, continuing to learn about new methods, techniques, new tools.

This curiosity also translates to a desire for certain kinds of positions among data analysts. One participant with a Ph.D. described leaving a position at an organization where creativity and curiosity weren't encouraged: "They had a problem keeping [data analysts] because, you know, it was kind of lame … it was like, oh, here's the data, turn out the same report."

Participants' motivations to follow their curiosity and be creative in their analysis work were sometimes associated with desires to not be slowed down or inhibited by external standards. One participant who had recently completed a bachelor's degree in statistics, for example, described business as a context in which he could be freer in his analysis than in academia:

> I feel like out here you can take more liberties with the data. ... I can kind of be creative with what I have and not get bogged down in what existing studies or that sort of thing have shown.

### *Interesting Problems*

Related to curiosity and creativity, the desire to solve interesting problems was cited by nine participants as a motivator for performing analysis work. Further, participants associated this motivation with performing analysis that would have a direct impact in the world. One participant with a background in educational psychology described his decision to not work as a research statistician, noting, "I was more interested in programming and solving real world problems."

Often, the problems that participants described as interesting could be related to social issues or causes. One participant, for example, described being motivated to work on solutions to help individuals monitor their energy usage. Another, working in social media, described a desire to work on problems that would make life easier for users.

However, participants also discussed being motivated by problems that were not directly related to social issues—one participant described trying teaching before realizing that he was more interested in technical challenges. Sometimes a problem was described as interesting just because it hadn't been solved well in the past. Other times, a problem was interesting because it gave the analyst the opportunity to work with sophisticated techniques or coworkers. For example, one participant who worked in a past job on payday loans (small, short-term loans often portrayed as cash advances on a salary) described the high levels of experience and sophistication of the individuals he worked with: "I don't miss being in payday, but I miss having access to those kind of clients."

### *Data Availability*

Data availability was the most commonly cited motivator for participants, with eleven participants mentioning it. While data availability was sometimes described as a benefit to the organization, participants more often described it as a personal motivator or as a reason they took one position over another. Desirable jobs were described as having access to large and interesting data sets and as putting few barriers between the analyst and the data. For example, one participant with over ten years of experience discussed his future career trajectory, noting that while he would like to start a department in small company: "There's something very exciting about being in a small, dynamic environment, where you don't have a lot of bureaucracy and you can get to the data you need without having to fill out a bunch of requests."

Participants described data availability in two ways. First, data sets were described as having advantages such as being very large or combining data in a way that allowed new connections to be made. One participant described receiving such a data set:

> We just received this huge Excel workbook that is essentially every customer for a utility's minute-by-minute energy use by end use. … You get this really intimate picture of, wow, this lady makes really bad decisions and her bill is like 40% higher than it should be.

Second, participants described data sets that were neglected but that could be used for interesting analysis work if they were analyzed in sufficient depth:

> We'd find out what sorts of data these companies had laying around and doing nothing with it, you know, just kind of sitting there. … So I got really interested in that, in how this data that's just sitting there can be looked at in a way that helps.

Poor data availability was also described as a reason one position might be less desirable than another. This was particularly evident when participants described why they preferred to work outside of academia. The same participant who described not pursuing a career in research because he wanted to make an impact in the real world also went on to note the advantages of the choice in terms of data availability:

> It turns out that businesses usually have really good data and they have lots of it, and it's already kind of there just waiting to be used. And academics kind of have to collect data, right, you have to do a lot of work to just get something to work with.

## Evaluation Criteria

In addition to motivations, we also assessed how data analysis work is evaluated in everyday practice. When coding episodes related to the evaluation of data analysts' work, we noted both who performed the evaluation and the criteria on which it was made. Because some participants performed work for clients outside their organization and others for clients within their organization, we differentiated only between whether evaluation was performed by members of the data analysis team or by an individual external to the team. In doing so, we gained insight into the similarities and differences between how data analysts and their coworkers or clients evaluated the results of analysis. Interestingly, we found that data analysts primarily evaluate their own work based on whether or not it makes sense based on previous experience and intuition. Others, however, evaluated data analysts' work according to its presentation.

### *Makes Sense*

Eight participants described episodes of evaluation based on whether the product of analysis aligned with intuition or made sense. Unsurprisingly, this criteria was almost always applied by members of the data analysis team and not by others.

Participants described applying this evaluation criteria during the development of models or algorithms as a way to decide if a current approach was worth pursuing. Multiple participants referred to this evaluation as a "sanity check." If the results did not align with that the analyst expected, it was often taken as a sign to try something else. A participant in charge of a research and development team noted:

> It's probably not the most scientific way, but given our depth and breadth in security, everybody on the team has a general idea of, "This is what I would expect to come out of the data," or, "This is what I would expect the end result to look like."

Indeed, domain expertise and relevant theory were often cited as important for this kind of evaluation. One participant with a background in social psychology pointed out that you can find correlations in any data set but what is important is whether those correlations are theoretically sound: "I'm like, do humans actually function this way?"

### *Works in Context*

Another frequently mentioned evaluation criteria was whether the product of analysis worked in the context for which it was built. Eight participants described scenarios in which this criteria was applied, both by the data analysis team as well as by others. Interestingly, the majority of participants who discussed this criteria worked in small organizations.

While analysts create models based on historical data or data created specifically for testing purposes, participants repeatedly mentioned the need to evaluate how a model performs in the context for which it was constructed. One analyst described this evaluation process as testing “toy models”; another referred to it as “playing with live ammo.”

Evaluation based on effectiveness in context was often described as leading to further analysis work. For example, one analyst explained that once a model designed to help schools monitor student progress was in place, the analyst team could begin to evaluate whether it was less effective with a certain subset of students and how it might be improved. Because many participants discussed the creation of models that would be incorporated into tools or products that were developed by other parts of the organization, this phase of evaluation was also described as relying on other departments to first implement and then report back on effectiveness. A participant whose job involved creating algorithms that would be a part of larger software products described this interaction:

> So we’ll say, “OK, here’s the data. Here’s what we ran, here’s our ... model. Can you take this and put this in a module that is blessed for going into the product? Furthermore, can you let us know when it maybe hits, or it doesn’t hit, so we can begin to evaluate the correctness of the model and look for model drift and all that kind of stuff in customer-type environments?”

In this example, the participant describes asking a coworker from another department to implement a statistical model in a piece of software that is approved, or “blessed,” to be put in production and used on real data. While data analysts typically test their models on historical data that may not reflect current conditions, coordinating with coworkers in other departments allows them to evaluate whether the model works in the context for which it was built.

Working in context was often described as attaining some measure of accuracy, for example leading to few false positives. However, whether a product of analysis works in context was sometimes also described as a matter of domain expertise, at which point the formal measures of the data analysis team became secondary to the more contextualized knowledge of others. As one participant with over ten years of experience noted, “We want the most accurate model, but that’s not necessarily the best model. The most accurate model might not be the best model that should be used.”

### *Presented Appropriately*

The format in which analysis results were presented was described by six participants as a criteria by which individuals outside the data analysis team evaluated work. While the data analysis team evaluates work based on formal measurements or whether it is intuitive, when delivering work to others within the organization or to clients, participants often described the importance of presenting it in an appropriate way.

Participants consistently described differences between the appropriate presentation format for other data analysts and for others in the organization or clients: “There is layman’s like, consumer-facing stuff, and then there is data scientist stuff. The consumers don’t like to see numbers.”

Other times, participants described less a need to avoid numbers and more a need to show the business impact of their analysis. The manager of a data science team described interacting with other departments in the organization:

> What you have to show them is like, here’s how you can use it. You know, by doing this, you can either reprioritize your call queue, you could segment your messaging better, and look what happens when you do. You’ve actually improved your renewal rate, you’ve increased the average selling price, or you’ve shortened the hold times.

Presentation format was also described as another kind of sanity check that took place within the data analysis team, in addition to intuition. One participant described this process: “Run [the analysis] by the rest of your team and they will say, you know what, I know that’s what the number says, but they are not going to like the way you say it. … It’s not pretty.”

## Discussion

We began this paper by describing two voices in debates over data science and big data approaches: 1) proponents largely coming from business literature and popular media and 2) critics who tend to come from academia and, particularly, the humanities and social sciences. This paper has worked to introduce a third voice to these discussions: that of on-the-ground, practicing data analysts. In this section, we return to the criticisms of data science and big data approaches discussed above and reflect on them in relation to these three voices.

### Theory and Traditional Methods

Insert Table 3 here

While critics have argued that data science and big data approaches ignore theory and, similarly, domain knowledge, our findings indicate that these do play an important role, especially as supports for data analysts’ evaluations of their own work at early stages.

Interestingly, participants’ descriptions of theory and domain knowledge were largely restricted to evaluations within data analyst teams. When presenting analysis work to others in the organization or to clients, analysis was more likely to be evaluated on whether it produced results.

Similarly, while participants did rely on domain knowledge and theory, they viewed traditional research methods as less useful. In discussing their desire to work outside of academia, several participants indicated that they had greater freedom to pursue approaches that might be restricted in other contexts. As Berry (2011) argues, traditional approaches and theories serve as gatekeepers, controlling what passes on to be considered knowledge. In describing their desires to not be "bogged down" by such academic processes, participants indicated, at times, a willingness to pursue interesting results at the expense of the mechanisms for rigor on which other approaches rely.

## Objectivity and Disinterest

Insert Table 4 here

Similarly, our results question the extent to which data analysts see their work as objective or disinterested. In line with criticisms of data science and big data approaches, participants consistently discussed the data they worked with as objective. However, they saw other aspects of their work as involving subjective judgment. This was especially notable in discussions of the presentation of their work. All participants were involved in presenting their work in some way, and many discussed the persuasive aspects of such presentations, pointing to the need to craft messages for specific audiences. In some instances, participants described presenting results that were created specifically as compromises with the desires of a client, and one participant described a position in which analysis work was performed specifically to support existing arguments.

These instances involving power and the analyst's position suggest that disinterest may not always be possible, especially for workers in business settings that involve contact with a client. Analysts might aspire to objectivity but be forced by circumstance to recognize their own positioning and the role of communication in data analysis.

This finding is consistent with prior research on work practices that reveals how practitioners with particular skills, knowledge, or expertise present their work in ways that align with their expectations about their audience. Barley et al. (2012), for example, note that engineers strategically favored either ambiguity or clarity in presentations depending on expectations about the intended audience. When soliciting feedback from others, engineers favored a level of ambiguity that fostered rich, long-term relationships with collaborators. Ambiguity in the objects they created served "to establish common understanding and future direction, to promote compromise, and to avoid potential conflict" (Barley et al., 2012, p. 301). While our study did not provide the kind of observational data that would allow detailed description of the strategies employed by analysts when presenting their work, our results do indicate that such a level of focus could reveal interesting comparisons with the existing literature. When presenting their work, data analysts must navigate a complex array of technologies, organizational structures, and social influences. Studying these aspects of their work could yield fruitful analyses for theorizing about communication in organizations, specifically as it relates to the introduction of data science and big data approaches.

## Privileged Access to Data

Insert Table 5 here

Participants showed a consistent awareness of issues around data availability, expressing the desire to be in positions with access to large and interesting data sets. In this context, boyd and Crawford's (2012) distinction between the data rich and the data poor describes not just the organizations that own data but also the workers that analyze it. For these workers, gaining access to data often means advancing their career and seeking positions at different kinds of organizations.

This perspective usefully highlights a tension in how data analysts are portrayed. Proponents of data science and big data approaches often foreground examples of insights generated from publicly available data—see, for example, Mayer-Schönberger & Cukier's (2013) discussion of predicting airline fares using available website data or *Wired*'s (2013) discussion of predicting Osama bin Laden's location using tweets. However, our results indicate that, in the contexts we examined, this pattern is not the norm, as participants showed consistent interest in working for organizations with access to proprietary data.

Perhaps more interesting are the motivations that participants described alongside data availability. Participants indicated that they want access to large and interesting data sets, the freedom to pursue promising analysis and the opportunity to implement the results of their analysis in ways that impact the world. Further, participants described wanting to work in companies where bureaucracy and excessive procedures don't hinder their ability to access data. Participants expressed a desire to be able to work quickly and creatively, and this was often associated with decisions to work outside the standards and rigor of academia. These motivations suggest a further distinction among the data rich and the data poor—from the perspective of data analysts, it is not just which institution has the data but also which can allow the data to be accessed easily, used creatively and implemented in impactful ways.

## Conclusion

Returning to the three voices with which we began discussion of our results, we see that practicing data analysts, unsurprisingly, fail to align perfectly with either proponents or critics of data science and big data approaches. While our participants were largely unaware of the exact criticisms made by academics in the humanities and social sciences, they did express an awareness of some of the issues raised in that literature—especially in relation to subjectivity, reliance on existing theory and data availability—while remaining consistently unaware of issues around the collection of data that are central to critics' arguments. Further, awareness of issues did not always translate into agreement.

Ultimately, our results indicate a complicated relationship between practicing analysts' work and discourses around data science and big data. At times, the image of data science and big data functioned as a goal for participants who wanted to use data creatively to make changes in the world. At other times participants held a more nuanced view of these approaches than do proponents, understanding the realities of working with proprietary data and within institutional structures. While critics' arguments are still relevant in relation to the work of our participants, we suggest that they seem to respond more to the popular media's hype than to the work practices our participants described. And while the popular media has tended to focus on the work of academics, independent researchers and prominent social media companies, our participants represent a manifestation of data science and big data approaches that has received little attention. At the same time, our participants' work uses data science and big data approaches to guide business decisions in impactful industries such as finance, energy and security. We suggest that further research on the everyday work practices of data analysts could allow critics and proponents of data science and big data approaches to better represent the reality of contemporary businesses.

In relation to these high level discussions of data science and big data, we suggest three ways future work might better account for the everyday work of data analysis:

1. Increased emphasis should be placed on data science and big data approaches at the individual level, rather than at the organizational or industry levels. At higher levels, evaluations of data science and big data approaches too often focus on financial or other outcomes to the exclusion of changes to work practices, interpersonal dynamics, and other individual-level phenomena.
2. Increased attention should be paid to the stages of analysis, as values are not consistent through the analysis process. For example, participants described domain knowledge and theory as important in the early stages of analysis but as far less relevant during later stages when other members of the organization and clients became involved in the evaluation process.
3. Increased attention should be paid to the context of analysis. While many criticisms of data science and big data approaches seem aimed at the academic researcher who has a great degree of control of data collection and the direction of analysis, many participants described working in far more constrained contexts. Focusing on these workers and the business contexts they inhabit suggests that data analysis needs to be considered in relation to the kind of organization in which it is embedded.

We built these recommendations from a study of 18 data analysts that has several limitations. First, many participants worked for small organizations in a confined geographic area. Data science may look different in large organizations with high degrees of centralization and long traditions regulating division of labor and use of technology. Additionally, while interviewing data analysts in a variety of industries allowed us to develop a general view of data science work, this limited our ability to make comparisons between groups. Greater attention to analysts and their work from the broader research community can help to overcome these limitations and assess the present and future of data science in modern organizations.

## References

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *WIRED*. Retrieved from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Barley, W. C., Leonardi, P. M., & Bailey, D. E. (2012). Engineering Objects for Collaboration: Strategies of Ambiguity and Clarity at Knowledge Boundaries. *Human Communication Research*, *38*(3), 280–308. http://doi.org/10.1111/j.1468-2958.2012.01430.x

Berry, D. M. (2011). The Computational Turn: Thinking About the Digital Humanities. *Culture Machine*, *12*.

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, *18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4869

Bowker, G. C. (2005). *Memory Practices in the Sciences*. Cambridge: MIT Press.

boyd,  danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job Of the 21st Century. *Harvard Business Review*, *90*(10), 70–76.

Drucker, J. (2010). Graphesis. *paj:The Journal of the Initiative for Digital Humanities, Media, and Culture*, *2*(1), 1–50.

Ferguson, R. B. (2013a, January 9). Bridging the Talent Gap: How to Find the Right Data Scientist. *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu.ezproxy.lib.utexas.edu/article/bridging-the-talent-gap-how-to-find-the-right-data-scientist/

Ferguson, R. B. (2013b, April 26). Predicting the Performance of Analytics Talent. *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu.ezproxy.lib.utexas.edu/article/predicted-to-perform-how-to-hire-analytic-talent/

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *Raw Data Is an Oxymoron* (pp. 1–14). Cambridge: MIT Press.

Graham, M. (2012, March 9). Big data and the end of theory? *The Guardian*. Retrieved from http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory

Harris, J. G., & Merotra, V. (2014, September 16). Getting Value From Your Data Scientists. *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu.ezproxy.lib.utexas.edu/article/getting-value-from-your-data-scientists/

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115. http://doi.org/10.1016/j.is.2014.07.006

Hempel, J. (2012). The Hot Tech Gig of 2022: Data Scientist. *Fortune*, *165*(1), 62–62.

Jurgenson, N. (2014, October 9). View From Nowhere. *The New Inquiry*. Retrieved from http://thenewinquiry.com/essays/view-from-nowhere/

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2917–2926. http://doi.org/10.1109/TVCG.2012.219

Kaskade, J. (2013, January 24). CIOs & Big Data: What Your IT Team Wants You to Know. Retrieved from http://blog.infochimps.com/2013/01/24/cios-big-data/

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2010, December 21). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu.ezproxy.lib.utexas.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. http://doi.org/10.1126/science.1248506

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Marcus, G., & Davis, E. (2014, April 6). Eight (No, Nine!) Problems With Big Data. *The New York Times*. Retrieved from http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

Osberg, M. (2014, September 11). The OKCupid data blog is back, in book form. Retrieved November 6, 2014, from http://www.theverge.com/2014/9/11/6132023/okcupid-data-blog-is-back-in-book-form

Ribes, D., & Jackson, S. J. (2013). Data Bite Man: The Work of Sustaining a Long-Term Study. In L. Gitelman (Ed.), *Raw Data Is an Oxymoron* (pp. 147–166). Cambridge: MIT Press.

Richards, N. M., & King, J. H. (2013). Three Paradoxes of Big Data. *Stanford Law Review Online*, *66*, 41–46.

Ross, J. W., Beath, C. M., & Quaadgras, A. (2013, December). You May Not Need Big Data After All. *Harvard Business Review*. Retrieved from https://hbr.org/2013/12/you-may-not-need-big-data-after-all

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. http://doi.org/10.1126/science.346.6213.1063

Steadman, I. (2013, January 25). Big data, language and the death of the theorist (Wired UK). Retrieved January 28, 2015, from http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory

Strauss, A. L., & Corbin, J. M. (1990). *Basics of qualitative research: grounded theory procedures and techniques*. Thousand Oaks, CA: Sage Publications.

Viaene, S. (2013). Data Scientists Aren't Domain Experts. *IT Professional*, *15*(6), 12–17. http://doi.org/10.1109/MITP.2013.93

Yin, R. K. (2013). *Case Study Research: Design and Methods*. Thousand Oaks, CA: SAGE Publications.