Note: This is an old, workshop version of a short paper that later turned into a *CSCW* publication. Please cite the final, CSCW version and not this working paper.

Here is the link to the final version: Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects.

Paper Title

Trust in Data Science

From Experimentation to Implementation

Abstract Title:

The Stakes are High

But, do we know what they look like?

Samir Passi
Cornell University

Introduction

The term big data refers not only to specific *types of data* (e.g., large-scale, distributed, granular), but also to *ways of working* with data (e.g., machine-learning, statistics, algorithms). Technical work through computational algorithms and statistical techniques, however, remains at the center of data science research imagination and representation. In data science, and in many other forms of research (including our own!), we often present research setup, process, and results in a dry and straightforward manner. *We had a question, we collected this data, we did this analysis, and here is what we found.* Open and effective conversations about the contingent and discretionary aspects of research and practice – data science or otherwise – tend to escape formal descriptions of method sections and grant applications, reserved instead for water cooler and hallway conversations by which workarounds, 'tricks of the trade', and 'good enough' solutions are shared. The result is an excessively 'neat' picture that fails to communicate the real practices and contingencies by which data science work proceeds.

A pertinent implication of such 'neat' representations of the practice and results of data science is the challenge in *trusting* data science results or applications. What was the business value driving the data science project? What assumptions were made in the analysis, and by whom? What forms of data were analyzed, and which forms of data were left behind? What models were used, and what were the

underlying assumptions of those models? Such questions, though integral to the evaluation of data science applications, are often invisible outside the immediate context of the data science practice. Burrell (2016) notes that the apparent "opacity" in machine learning driven data science applications is because of three main reasons: (a) the algorithms are often considered trade-secrets, (b) analysis of the underlying code and mechanisms requires specialized forms of expertise, and (c) the very nature of specific algorithmic approaches such as deep learning make them harder to understand.

Issues of trust, opacity, and legitimacy are exacerbated owing to the fact that automation – as operationalized with and through algorithms – is often understood as an absence of bias (Bozdag 2013; Dourish 2016; Gillespie 2014; Naik and Bhide 2014). Researchers in this area remind us that it is often difficult to detect algorithmic bias – especially implicit latent bias – particularly given that the development histories of algorithms aren't available for analysis, camouflaging bias that may arise not only from pre-existing values and imperatives, but also from the emergent contexts of algorithmic re-use as opposed to their contexts of development (Friedman and Nissenbaum 1996; Hildebrandt 2011). Mechanistic interpretations of data science applications – coupled with the 'self-evident' nature of floating point numbers, facilitates forms of "deresponsibilisation of human actors" – a "tendency to hide behind the computer." (Mittelstadt et al. 2016) As a result, "non-expert" data subjects may lose trust not only in algorithms, but also in the enterprise of large-scale data science more broadly (Cohen et al. 2014; Rubel and Jones 2014).

Within critical data studies, discussions on trust are found embedded in a dichotomy between corporations and people – *corporations have exclusive data on people, and they use these data in specific ways to alter our agency, experience, practices, and lives*. Algorithmically produced numbers aren't just "supplements" to the knowledge process, but come with their own "rhetoric of factuality, imbued with an ethos of neutrality, and presented with an aura of certainty." (Rieder and Simon 2016) Understanding how trust is built and operationalized in such systems is then as a product of the management – specifically, error and reliability management – and transparency of such systems (e.g.: Symons and Alvarado 2016). However, there are caveats. Transparency, for instance, isn't as straightforward as granting access to the data science process. "What is *clearly transparent* for one community may be totally *opaque* to another." (Mayernik 2017) Moreover, as these algorithmic systems grow and develop, the "complex decision-making structure" required to manage these systems "can quickly exceed the human and organizational resources available for oversight." (Mittelstadt et al. 2016)

Such discussions on trust resonate strongly with the vast scholarship on the relationship between trust, science and technology in Science & Technology Studies [STS]. Within the STS

discourse, notions of trust and objectivity are often considered "inseparable" (Daston and Galison 1992, 2007; Porter 1995). In their seminal work - *Leviathan and the Air Pump* - Shapin and Schaffer (1985) show how the trustworthiness of the scientific method was built up through by ascribing to it a form of mechanical objectivity detached from personal beliefs and biases. Detailed descriptions of scientific experimentation allowed for forms of "virtual witnessing," and third-person narratives provided rules and instruction to allow anyone, anywhere to replicate the same experiment. Mechanically produced, quantitative results were then equated with true, unbiased, and factual representations of reality.

Trust in numbers, however, as Porter (1995: 90) puts it, is but one form of "technologies of trust" in which "mechanical objectivity serves as an alternative to personal trust." People, institutions, and politics has as much to do with trust in science and technology as practices of standardization, objectivity, and mechanization (e.g.: Porter 1995; Shapin 1994, 1995a, 1995b) - "trust within science functions in a way similar to that in which it functions in any area of life involving human skills" (Pinch 1986: 207). Moreover, scientific and technological artifacts aren't confined to laboratories - they form an integral part of every sphere of modern life. As these artifacts travel further from their immediate communities and contexts of production, "distance lends enchantment" Collins (1985) and they take on forms of certainty that are often unwarranted (e.g.: MacKenzie 1993).

Studying how trust operates in the interaction between corporate data science applications and people has been pivotal in understanding the role and implications of data science for social, cultural, economic, and policy practices. In a way, such discussions of trust focus on a specific relation: between *people and organizations* i.e. how can and do people trust corporate data science systems. In my own research, I have found that there is another important relation between trust and corporations: *how can and do organizations trust their own data science practices?* Is organizational trust in a corporation's own algorithms and models simply an operationalization of numerical metrics and cross-validation scores? How does the understanding of underlying data and its messiness impact such discussions of trust? Is there a singular culture of trust in a data science organization, or are there multiple frameworks of trust within which data science algorithms, models, and systems are conceptualized, developed, and evaluated? In this paper, I begin to scratch the surface of such questions concerning organizational trust. In the section – *Methods and Empirics* – I describe my research methodology and field-site. In the section – *Negotiating Trust: An Empirical Vignette* – I describe one specific example of organization trust based on my fieldwork. Finally, in the section – *Discussion* – I describe preliminary and open questions coming out of the vignette.

Methods and Empirics

In my research, I study forms of data science work ethnographically within two contexts: (1) *academic data science* (one-year of ethnographic research conducted in two graduate-level machine learning courses in an American university, and participant-observation in a series of three digital humanities workshop organized at the same university), and (2) *corporate data science* (six-months of ethnographic research conducted at an American technology firm working as a data scientist including approximately 50 interviews with data scientists, project/product managers, business personnel, and company executives). For this workshop, I draw on my study of corporate data science practices.

I worked as a data scientist at DeepNetwork Inc.[1] from June 2017 until November 2017. Founded in the nineties, DeepNetwork is a mid-size (2500-5000 employees) ecommerce and new media corporation that owns 100+ online companies in domains such as health, legal, and automotive. The company has a core data science team based on the US West Coast that works with multiple businesses across different domains and states on a variety of data science projects. During my time at the company, the size of the team ranged from seven to ten members including myself. The team is formed by and under the direct supervision of Justin – DeepNetwork's Chief Technology Officer (CTO). The team is headed by Martin – DeepNetwork's Director of Data Science. My approach in this research was to study corporate data science practices not simply as an observer of data science teams, but as an active participant shaping decisions while working with data scientists and business teams. As I analyze my fieldwork notes, interviews, and ethnographic data through grounded-theory method, this workshop paper is a first step towards outlining an insight coming out of my fieldwork.

Negotiating Trust: An Empirical Vignette

When I joined DeepNetwork, the data science team was involved in multiple projects across different industries. Three of these projects dealt with business churn prediction i.e. the problem of detecting currently active paying customers who are likely to cancel their subscriptions or services in the near future. In this paper, I want to focus on aspects of the DeltaTribe churn prediction project. Owned by DeepNetwork, DeltaTribe is a multi-million-dollar online marketing and customer relation and retention management company with over 50,000 clients across the United States. While DeltaTribe's headquarters are in a different city, its churn prediction project is handled by the data science team

[1] All company and personnel names in this ethnographic vignette have been changed to pseudonyms for research participant anonymity as per Cornell University's Institutional Review Board (IRB) for Human Participants' guidelines for the approved research protocol #1705007175.

based in DeepNetwork headquarters. When I joined the company, the project had already been going on for two months, and the data science team had already decided on two specific machine learning models that they were going to use for this project.[2] There were two data scientists working on this project – Max and David. Both, Max and David, had tried out a few machine learning models, and then settled on two of them. When asked how these two models were chosen, David mentioned that the selection of the two models wasn't done only by the data scientists, but collaboratively with Daniel – the Project Manager for the team – who "chose them based on how well they performed." [3] The models were different not only in their algorithmic approach, but also differed in the datasets on which they worked.

> **David:** Max's model [XGBoost] works on customer subscription data… like user features… whether the user has enabled email, voice services, what are the different settings and features. This data is collected every day. My model [Random Forests] works on customer trending data. This consists of actions… what has the user been doing on the platform. User actions. This is collected every month.[4]

Moreover, the datasets used by Max and David were provided directly by David who scraped company datasets, preparing two different sets of datasets:

> **Daniel:** In my personal opinion, and this is different from what Martin believes in, people who build the models are better at building training sets. But… the caveat is… this is also what Martin says… people who build the model are more likely to over fit because of their biases or because they know… they know how their algorithm works. This is why Martin wants me – as an independent person – to create the datasets used by the machine learning team.[5]

Daniel was, in addition to Max and David, also involved in evaluating the model results. Daniel's evaluation of the results, thus, played a key role in how he "prepared" the data sets:

> **Daniel:** Guess what we found when we started looking at the results… at the output. Remember that customer churn data is collected over the last 2 years, every month. So, a customer who has been with us for those two years will have 24 records. We found that the model was treating the number of records as the strongest indicator to predict – in the past – who was a churning customer. A lot of the customers it correctly predicted had… say only 10 records. Of course, if a person has only 10 records, and the 10th record is before the last month of data that we have, they have churned! [So…] If a customer starts with us 4 months ago, and we use the model to predict, it would predict that customer

[2] *Model 1:* XGBoost, *Model 2:* Random Forests.
[3] David, Personal Communication, Fieldwork Notes, June 8, 2017.
[4] David, Personal Communication, Fieldwork Notes, June 8, 2017.
[5] Daniel, Personal Communication, Fieldwork Notes, June 16, 2017.

> as 'likely to churn' just because it has only 4 records, but that's just because it is a new customer, and has nothing to do with the confidence of whether they will churn or not.
>
> **Samir:** And you figured out that the algorithm was doing that?
>
> **Daniel:** Yea, as I was looking at the results. I told you that I have to interpret what the model is doing based on the results I see. So, when I saw what customers the model was predicting as likely to churn, I noticed that most people the algorithm was identifying had less than 10 records. I told this to the guys, and said – hey, why is this happening, and that's when we realized that the model was prioritizing past number of records.
>
> **Samir:** So, what did you do?
>
> **Daniel:** I corrected that by balancing out the training set.
>
> **Samir:** What does that mean – balancing them out?
>
> **Daniel:** Creating balanced sets takes time. Once I realized that the number of past records was overriding model features, I started to balance training set to include people with different start dates. For instance, if I include someone who has been with us for only 3 months and is still a current customer, I have to include an older customer too in order to have a balance of new people and old people. When we do that, our accuracy of predicting churn customers goes down drastically… currently it is at 30%.[6]

While Daniel may have tried to create "independent" training sets and "balance" them out, Martin – DeepNetwork's Director of Data Science – wasn't satisfied with the model accuracies:[7]

> **Martin:** We have been on this for a long time… on multiple kinds of DeltaTribe data going back to two years… a lot. We want to predict who will churn. The disappointment is… we have a high accuracy in predicting who will stay on, but [laughs] we are not good at predicting who will leave. 30%… I am not going to go to somebody with a machine learning model that is only 30% accurate. That makes no sense.[8]

The number "30%" wasn't the only number associated with the two models. There were, in fact, two sets of accuracies associated with the models as explained by David and Max:

> **David:** Churn, in general, is low… say around 10%. This is true for us also since a lot of our customers aren't going to churn, they are old customers. Now, if only 10% are going to churn, and we predict that no one is going to churn, we will have accuracy of 90%… just by making the same prediction for everyone.[9]
>
> **Max:** Getting a high accuracy… historical accuracy… is easy. We take past data, hold some out for validation, and train the model on remaining data. We get high accuracies… 95%… sometimes 98%.

[6] Daniel, Personal Communication, Fieldwork Notes, June 16, 2017.

[7] In simple terms, the accuracy of a machine learning model is a numerical score indicating what percentage of the model's output is correct.

[8] Martin, Personal Communication, Fieldwork Notes, May 30, 2017. All quotes, unless explicitly put between double quotation marks (""), are paraphrased versions taken from ethnographic field notes.

[9] David, Personal Communication, Fieldwork Notes, June 8, 2017.

> Not useful. We need future testing… hold out the data from recent months, and then predict… as if looking into the future, that is where we get only 30%... at best.[10]

Max's articulation of the difference between "historical" and "future" accuracy was important for the data science team. The aim was to predict churn in the future, and a model that was only 30% accurate wasn't going to cut it for anyone. The interim solution that the data science team had come up with was to not focus not on the accuracy scores, but on the small set of active customers with the highest churn probabilities as calculated by the model:

> **Martin:** What we have done is to create an ordered list of 450 customers with a probability of their churning. In doing that we give our business team a list of clients that are likely to churn. Think of it as a list that… if you are on it, you are going to die. We provide the 450 with the highest probability of dying, and tell our business team that maybe you should reach out to these folks as they may very well leave us. By doing this we deemphasize the 30%. Instead, we say – 'OK, how can a low 30% score still be made useful for business purposes?' An ordered list is one such way.[11]

Daniel, however, had an additional reason for "deemphasizing" the 30% score through the ordered list. For Daniel, the score was "problematic to begin with" given the orientation of the business team to such numbers:

> **Daniel:** When we say to the business team that we have a 90% accuracy, they think that 'oh so you are saying with a 90% confidence that these people are going to churn.' No, that's not what we are saying! They think in black and white – that they are only 2 answers: likely to churn or not likely to churn. But, in total we have 4 answers: (a) people who we say will churn, and who churn [in machine learning terminology[12], these are called *true positives*], (b) people who we say will churn, but don't [*false positives*], (c) people who we say won't churn, and don't [*true negatives*], and (d) people who we say won't churn, but churn [*false negative*s]. So, when we say something, it is in relation to four categories, not just two.
>
> **Samir:** Why do you think the business team thinks like that?
>
> **Daniel:** I think that the binary labels are maybe restrictive for the business team. We have found that explaining in terms of 'confidence.' is easier. It's like saying – OK, here are a few people at higher risk than others, deal with them first. […] We don't say that here are 10 people who are going to die, and others won't. We say that here are 10 people at a higher risk, deal with them first. […] And you know what, I mean… the overall accuracy might not even be valuable. Even if we are wrong 7 out of 10 times, we are still giving extra information on at-high-risk customers…[13]

[10] Max, Personal Communication, Fieldwork Notes, June 8, 2017.

[11] Martin, Personal Communication, Fieldwork Notes, May 30, 2017.

[12] One class of machine learning models deal with the problem of binary classification i.e. predicting which of two classes does a given data sample belong to. For the churn project, these two classes are 'likely to churn' and 'not likely to churn.' The class in question, here the 'likely to churn' class, is considered the positive class. In this setup, the classification model's output can be categorized into four parts as explained by Daniel.

[13] Daniel, Personal Communication, Fieldwork Notes, June 8, 2017.

This ordered list was presented to Charles and Parth – two Senior Business Operations Analysts from DeltaTribe – on June 9, 2017. Since Charles and Parth weren't in the same city, they had dialed in to the meeting. The ordered list presented to them contained a list of unique customer IDs along with their model-produced 'likely to churn' probabilities. This list was produced through Max's model. Charles and Parth said that the list is "useful," and that they would run a pilot test with this list to see if it made any difference to DeltaTribe's churn rate. Running the pilot test required the machine learning results to be integrated into DeltaTribe's existing customer relations portal so that customer agents could use the associated 'likely to churn' probabilities with those accounts. Daniel sent the final version of the list to Charles and Parth on June 21, 2017. For the next two months, things were quiet on this project and the data science team got small updates on the status of the integration.

It was in the first week of August that we had another meeting with Charles and Parth. We learned that while the technical integration of data science results into the customer relations portal had passed quality assurance, the results were never completely incorporated in the portal. The reason was that DeltaTribe's technology development team had been dealing with other issues and challenges that had come up, and so the data science project was never made a priority. Thus, the pilot test never took off. However, things had settled down now, and DeltaTribe wanted a fresh set of results given that the previous set of results were a couple of months old now. Martin asked Max to retrain the model, and produce results for the most recent dataset.

There was another point of discussion in the meeting. Martin had learned from his colleagues that there was a data analyst working at DeltaTribe who had also been trying to do machine learning driven churn prediction for DeltaTribe. His name was Hector. Given that Hector had essentially been trying to do the same thing that Max and David had already done, Charles and Parth asked Martin if the data science team could have a meeting with Hector. They informed us that Hector – in his capacity as a data analyst – "knows a lot about churn and churn analysis." Thus, they felt that the meeting could be beneficial to both: the data science team (i.e., we could gain from Hector's business knowledge about churn), and to Hector (i.e., he could learn more about machine learning from us).

We had a meeting with Hector on August 16, 2017. Since Hector wasn't in the same town, he dialed into the meeting. While Hector admitted that he "wasn't a crazy good data scientist," he did inform us that his algorithmic approach for churn prediction was very similar to Max's approach. The difference was that Hector had been using a different dataset for the task – one that overlapped with Max and David's datasets, but also had additional data from other company sources. At this point, Hector was – as Max pointed out – "still learning… [and] didn't know much about prediction… or

the problems with it." Max's remark stemmed from the fact that when asked about future testing, Hector not only indicated that he hadn't done any future testing, but also recapitulated his model's historical computational scores (the scores that the data science team considered not useful). The meeting ended with a decision to share Max's code with Hector, and Hector's code with Max. Things became quiet on this project for some more time.

Fast forward two and a half months, Martin informed the data science team during the daily meeting that Hector had continued working on churn prediction and he had, in fact, presented his results to DeltaTribe's executives and business heads – something that even DeepNetwork's core data science team hadn't been able to do. Martin organized a meeting with Charles and Parth to discuss the same. We met again with Charles and Parth on November 2, 2017. Martin began the meeting by setting out an agenda:

> **Martin:** It isn't worthwhile to continue with two separate models. Either we consolidate our efforts to refine one model, or… we decide which model has potential and drop the other one. If Hector's model is better, then he can take the lead on the project, and the data science team would act only in a consulting capacity.[14]

It was in this meeting that Charles told Martin that "everyone at DeltaTribe likes Hector's model better." Martin asked why.

> **Parth:** Hector's model is segmented by AOP [Area of Practice].[15] From churn analysis we know that churn is different across different AOPs, and so it makes sense that there be different models to capture churn across different AOPs.
>
> **Martin:** OK, but our next step was also going to be splitting models by AOPs. We just never reached that point.
>
> **Charles:** Yea, but… Hector has been part of the DeltaTribe team as a data analyst for a long time… He understands business in a more in-depth manner.
>
> **Martin:** That is correct. He is embedded in the business in a way that we aren't.[16]

Martin pressed on. He told Charles and Parth that the models used by Hector and Max were in fact very similar, and asked how the comparison between the two models was done.

[14] Martin, In-Meeting, Fieldwork Notes, November 2, 2017.

[15] DeltaTribe offered its services to customers in different industries ranging from health to legal and automotive. The different industries were considered different Areas of Practice [AOPs]. While Max clubbed all data from all AOPs together (and considered AOP as one feature of a given customer), Hector's model separated the data into different datasets – each belonging to one specific AOP.

[16] In-Meeting, Fieldwork Notes, November 2, 2017.

> **Charles:** Hector's model has higher accuracy. This has been the main bone of contention between the two models. His model is AOP-specific and seems to produce better results.
>
> **Martin:** Can you describe what you mean by accuracy, and how are you calculating it?
>
> **Parth:** It makes sense that an automotive client isn't using the portal in the same way a health client is, so AOP is the way to go on this.
>
> **Charles:** When it comes to accuracy… the results… we mostly did spot-checking. During the spot-checking, we found that we couldn't find causes in your identified higher probability customers… but we could for Hector's results. He was heavily involved in our own churn meetings, and was our main point of contact.
>
> **Max:** We chatted with Hector sometime back, and at that time we had been using very similar models. We… I did future testing, but last time I checked he [Hector] didn't do it. Is the spot checking from the future testing results, or historical results?[17]

Charles indicated that while he "wasn't sure of the answer," he felt that the results were from "future testing with accuracy over 90%." Hearing the number "90" made almost everyone from the data science team shift in their seats. A ninety percent future churn prediction accuracy was something that was unheard of at least in the data science team. The meeting concluded with Martin, Charles, and Parth deciding that perhaps it was best to inquire into the "more technical details" with Hector himself. We had an on-call meeting with Hector on November 8. As soon as Hector joined the meeting, Martin asked him about his evaluation strategy:

> **Martin:** Hector, Charles and Parth are unable to give me forms of evaluation that I can trust… try do spot-checking… spot-checking cannot work. I wanted to know from you, what kinds of testing have you done, and what the numbers are.[18]

Hector responded by telling Martin about his accuracy scores – across AOPs, scores ranged from high 80s to low 90s. These were the same numbers that Hector had shared with DeltaTribe business teams and executives. At this point, Max interrupted Hector and asked him about the difference between "historical testing scores and future testing scores." Has Hector performed future testing? In his reply, Hector not only indicated that he had not done future testing (at least the kind that Max had in mind), but also argued that future testing probably wasn't a good metric to begin with:

> **Hector:** If we do future testing… like a controlled study… our predictions would anyways be low. To be honest, it… future testing is kind of unfair… as we are asking people to take action on a list… there is already influence, and we don't know if... they were going to churn or not.

[17] In-Meeting, Fieldwork Notes, November 2, 2017.

[18] Martin, In-Meeting, Fieldwork Notes, November 8, 2017.

> **Martin:** Even if you predict to cancel… people are at different phases of contracts… maybe they want to quit, but they have four months of service already paid for.
> **Hector:** Yea, but folks over here… they wanted percent probabilities… 6 months… even before.[19]

Hector went on to say that after segmenting customers into AOPs, and establishing a time period, he had sliced records up by churn probability ranges, and that "results looked good at higher numbers." Martin, however, wasn't convinced:

> **Martin:** There are ways to evaluate models… we can do spot checking, but that cannot hold. We get very high numbers on historical data as Max has shown up. I am sure you can do that too with your model given that models are very similar anyways. What has been the challenge is to get Charles and Parth and everyone to agree that we need to take model results, and ask our customer agents to follow-up on people to see if we are going anywhere. Without that feedback, I don't know how to make the models better.

Hector agreed that the final deliverable of this project was indeed a list of at-risk customers that the customers representatives could call. But, he also felt that the numbers generated from such a test would be skewed since these customers were now influenced by the agents. However, he agreed with Martin that spot-checking was "not a sure shot way of knowing if things were working well." The meeting ended with no clear answer in sight. In the meantime, Martin had asked Samuel – the new Project Manager who had replaced Daniel back in August – to send an email to Charles and Parth, asking them more details on the "spot-checking evaluation" that they had done. Here is what Charles replied to Samuel:

> **Charles:** The way we were evaluating the results was initially uploading the model you provided and running reports for the accounts that showed a high probability for canceling. Parth and I did a lot of spot checking to see if there was any indicator that would show why they are showing a high probability and for a large number there wasn't anything we could see. Hector's model honestly came out of nowhere and he brought it to my attention during a churn meeting with company executives and based on his analysis it was showing very accurate numbers due to the fact that it was industry specific.[20]

It is at this point that I end this vignette. The project moved through various hoops and loops, and by the time I left it was decided to hand over the project to Hector with Martin's team acting in a consulting capacity to help DeltaTribe in case need be. As I continue to refine and analyze this vignette, there are a few broad (albeit rough) points of interest. First, even within the data science team, there were different approaches to model evaluation. While the data scientists constantly

[19] In-Meeting, Fieldwork Notes, November 8, 2017
[20] Personal Email Communication between Samuel and Charles, November 27, 2017.

reiterated either the need for future testing or the futility of historical accuracy scores, the project manager felt that the models were biased in other ways, needing forms of data balancing. The Director of Data Science, all this while, was focused on on-the-ground use of model results by customer representatives to better understand model performance. On the other hands, the business team felt that the segmentation of the model into specific AOPs was key to "better results." While Hector may not have performed future testing in the way Max had done, his high historical accuracy numbers coupled with Charles and Parth's spot-checking approach (and the fact that they felt Hector had in-depth knowledge of the business) made Hector's model appear better to DeltaTribe. I followed up on these discussions in person with Martin, Samuel, Max, Charles, Parth, and Hector in my one-on-one interviews with them. However, since I am still in the process of transcribing those interviews, I haven't included them in here yet. Given that, much of the details that I found during the interviews further corroborate and nuance the existing narrative.

Discussion

Trust – as seen clearly through the ethnographic vignette – is *subjective*. This isn't a surprise. We already know that negotiations of trust are embedded in a variety of factors including, but not limited to, organizations, power, and professions. However, this vignette problematizes the simplistic dichotomy that has been established in critical data studies: between data science corporations and people. Data science corporations, in fact, comprise multiple groups of professionals with each conceptualizing and enacting trust in different ways. For instance, while data scientists may have their own quantitative metrics, these often require alignment with business and organizational values. Corporate data science projects don't just comprise of a group of homogeneous professionals working in a self-contained epistemological vacuum. These projects require collaboration. As all other forms of collaboration, these then necessitate myriad forms of work in articulating, translating, and negotiating across multiple repertoires of trust. Fortunately, or unfortunately, not everything boils down to a number.

Currently, I am in the process of analyzing such forms of 'trust work' involved in corporate data science projects. My aim is to advance a sociotechnical understanding of how "agreements of trust" (Boltanski and Thévenot 2006) in data science systems are accomplished on the ground. I wish to take the opportunity of this workshop to get ideas and feedback form the workshop participants with regard to how they understand the relation between trust, algorithms, and data science. Give that this is still a work-in-progress, here are two short preliminary remarks to kick-start the discussion:

a) Trust isn't a binary between yes and no. It operates on a spectrum, and in a collaborative organizational setting, groups often lie at different points on the spectrum. Again, this isn't a surprise to anyone in organizational studies. However, in corporate data science projects, it is often unclear whether there is, in fact, just one spectrum at play. Two groups may trust the machine learning application, but for two very different reasons. While data science research on evaluation and performance of such systems may focus extensively on one such spectrum (e.g., computational scoring metrics), experimentation and implementation of corporate data science applications are embedded within multiple repertoires of trust.

b) Multiple corporate groups need to trust a data science application for it to be implemented as part of a business practice. However, while such forms of trust are essential, the nature of trust within and across these groups isn't that of 'absolute trust' *(i.e., this definitely works, and we should absolutely use it)*, but that of 'practical trust' *(i.e., this seems to work, and is good enough for our purposes)*. There seems to be a pragmatic approach to trust in corporate data science projects, and there is often a relation between a system considered 'trustworthy' and a system considered 'good enough.' The notion of a 'good enough application' is an interesting dimension to investigate not only how different actors approach and understand a machine learning application, but also the mechanisms through which good-enough-ness is ascertained.

Bibliography

Boltanski, Luc, and Laurent Thévenot. 2006. *On Justification: Economies of Worth*. Princeton: Princeton University Press.

Bozdag, Engin. 2013. "Bias in Algorithmic Filtering and Personalization." *Ethics and Information Technology* 15(3): 209–27.

Burrell, Jenna. 2016. "How the Machine 'thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3(1): 1–12.

Cohen, Glenn et al. 2014. "The Legal and Ethical Concerns That Arise from Using Complex Predictive Analytics in Health Care." *Health Affairs* 33(7): 1139–47.

Collins, Harry M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London: Sage.

Daston, Lorraine, and Peter Galison. 1992. "The Image of Objectivity." *Representations* 40: 81–128.

———. 2007. *Objectivity*. Cambridge, Massachusetts: MIT Press.

Dourish, Paul. 2016. "Algorithms and Their Others: Algorithmic Culture in Context." *Big Data & Society* 3(2).

Friedman, Batya, and Helen Nissenbaum. 1996. "Bias in Computer Systems." *ACM Transactions on Information Systems (TOIS)* 14(3): 330–47.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, eds. Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. Cambridge: MIT Press, 167–94.

Hildebrandt, Mireille. 2011. "Who Needs Stories If You Can Get the Data? ISPs in the Era of Big Number Crunching." *Philosophy & Technology* 24(4): 371–90.

MacKenzie, Donald. 1993. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. Cambridge, Massachusetts: MIT Press.

Mayernik, Matthew S. 2017. "Open Data: Accountability and Transparency." *Big Data & Society* 4(2).

Mittelstadt, Brent D. et al. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3(2).

Naik, Gauri, and Sanika S. Bhide. 2014. "Will the Future of Knowledge Work Automation Transform Personalized Medicine?" *Applied & Translational Genomics* 3(3): 50–53.

Pinch, Trevor J. 1986. *Confronting Nature: The Sociology of Solar-Neutrino Detection*. New York: Springer.

Porter, Theodore. 1995. *Trust in Numbers: The Pursuit of Objectivity in Scienceand Public Life*. Princeton: Princeton University Press.

Rieder, Gernot, and Judith Simon. 2016. "Datatrust: Or, the Political Quest for Numerical Evidence and the Epistemologies of Big Data." *Big Data & Society* 3(1): 1–6.

Rubel, Alan, and Kyle M. L. Jones. 2014. "Student Privacy in Learning Analytics: An Information Ethics Perspective." *The Information Society* 32(2): 143–59. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2533704.

Shapin, Steven. 1994. *A Social History of Truth: Civility and Science in Seventeenth Century England*. Chicago: University of Chicago Press.

———. 1995a. "Cordelia's Love: Credibility and the Social Studies of Science." *Perspectives on Science* 3(3): 255–75.

———. 1995b. "Trust, Honesty, and the Authority of Science." In *Society's Choices: Social and Ethical Decision Making in Biomedicine*, eds. Ruth Ellen Bulger, Elizabeth Meyer Bobby, and Harvey Fineberg. Washington D.C.: National Academies Press, 388–408.

Shapin, Steven, and Simon Schaffer. 1985. *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life*. Princeton: Princeton University Press.

Symons, John, and Ramón Alvarado. 2016. "Can We Trust Big Data? Applying Philosophy of Science to Software." *Big Data & Society* 3(2).