Big Data, Machine Learning, and the Social Sciences | by Hanna Wallach

hannawallach.medium.com/big-data-machine-learning-and-the-social-sciences-927a8e20460d

December 23, 2014

Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency

This essay is a (near) transcript of a talk I recently gave at a NIPS 2014 workshop on "Fairness, Accountability, and Transparency in Machine Learning," organized by Solon Barocas and Moritz Hardt.

Introduction

I want to start by giving you some context for this talk by telling you a little bit about me and my background: I'm a machine learning researcher by training. That said, I wouldn't describe my research over the past few years as being traditional machine learning. Instead, most of my recent research has been in the emerging field of computational social science — that is, the development and use of computational and statistical techniques to study social processes. This shift in research direction has given me an opportunity to start thinking outside the algorithmic boxes typically embraced by the machine learning community and instead focus on the opportunities, challenges, and implications involved developing and using machine learning methods to analyze real-world data about society.

Outside of my research life, I've also spent the past 15 years contributing to the free and open source software community and working to promote and support women in computing. These activities, and their focus on transparency, openness, fairness, and inclusion, have influenced my perspective on the issues surrounding today's workshop, and I'm excited to be speaking here because it's given me an opportunity to tie together some of my thoughts on these ideas and their relationship to machine learning.

We have a really interesting group of people here today, with very diverse backgrounds, and I'm looking forward to seeing what comes out of their interactions. As a result, this talk will be structured around four talking points — intended to prompt discussion — that lie at the heart of fairness, accountability, and transparency in machine learning:

- Data
- Questions
- Models
- Findings

Data

Since the primary impetus for this workshop was President Obama's ninety-day review of big data, I want to start by talking about a couple of definitional issues surrounding big data and its common uses.

> What is big data?

For something so spectacularly pervasive and trendy, I've heard this question surprisingly often. Initially, I thought this was because no one knew (or at least agreed upon) the answer. For example, NSF and NIH — two of the largest funders of academic research on big data — gave the following definition in their 2012 joint program solicitation:

'[B]ig data' […] refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available[.]

Meanwhile, other organizations refer to Gartner, Inc.'s "3Vs" definition:

'Big data' is high volume, high velocity and high variety information assets that demand cost-effective, innovative forms of information processing[.]

In this definition, volume refers to the amount of data in question, velocity refers to the speed with which that data can be obtained and/or processed, while variety refers to the range of different data types and sources.

Nowadays, I think the reason I hear this question so often is that some of the most salient properties of big data, as it's commonly construed, make people very uncomfortable. Yet, these vague, catch-all definitions don't highlight these disquieting properties. For example, there's little in these definitions — especially Gartner's — that distinguishes this "new" big data from the massive data sets arising in physics, for example. So why, then, are people so worried about big data but not about particle physics?

I think there are two reasons. The first is nicely captured in this alternative definition of big data from a recent article by Michelle Chen:

'Big data' [is] the amassing of huge amounts of statistical information on social and economic trends and human behavior.

In other words, unlike the data sets arising in physics, the data sets that typically fall under the big data umbrella are about *people* — their attributes, their preferences, their actions, and their interactions. That is to say, these are *social* data sets that document people's behaviors in their everyday lives.

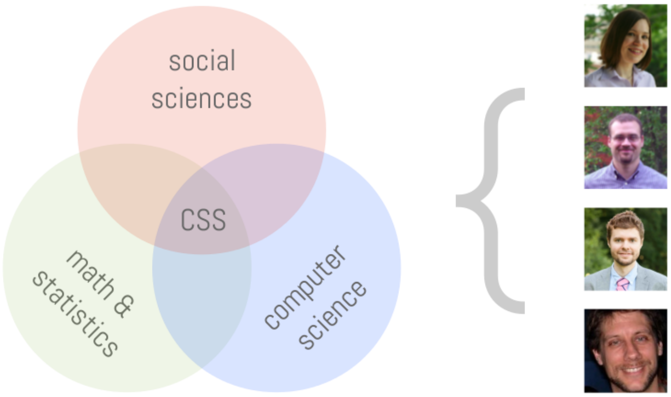
The second reason is highlighted by this quote from Michael Jordan, a machine learning researcher, taken from a recent talk of his:

The issue is not just size—we've always had big data sets — the issue is granularity.

In other words, not only do these data sets document social phenomena, they do so at the granularity of *individual* people and their activities.

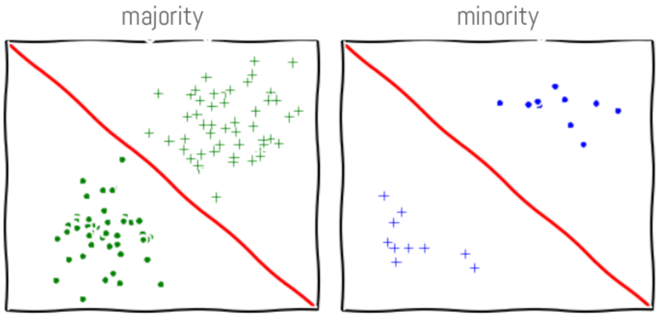
So why, then, does *granular, social data* make people uncomfortable? Well, ultimately—and at the risk of stating the obvious—it's because data of this sort brings up issues regarding ethics, privacy, bias, fairness, and inclusion. In turn, these *issues* make people uncomfortable because, at least as the popular narrative goes, these are *new* issues that fall outside the expertise of those those aggregating and analyzing big data. But the thing is, these issues aren't actually new. Sure, they may be new to computer scientists and software engineers, but they're not new to social scientists.

This is why I think the world of big data and those working in it — ranging from the machine learning researchers developing new analysis tools all the way up to the end-users and decision-makers in government and industry — can learn something from computational social science.



One of the things I love most about working in computational social science is that, by definition, it's an inherently interdisciplinary field, in which collaboration is necessary to make ground-breaking progress: social scientists provide vital context and insight regarding pertinent research questions, data sources, acquisition methods, and interpretations, while statisticians and computer scientists contribute expertise in developing mathematical models and computational tools. Granted, interdisciplinary collaboration is hard — for example, you have to establish a common language for communication — and it takes time — often more time than within-discipline collaborations — but, ultimately, it produces results that no discipline could have produced single-handedly. In fact, there's research in social psychology and sociology indicating that rapid, breakthrough innovations are best facilitated by bringing together people with diverse backgrounds. When working in diverse groups, people behave differently — they are more likely to express different viewpoints and to emphasize diverse characteristics. In contrast, when working in homogeneous groups, people tend to overemphasize the qualities they share.

So, if technology companies and government organizations — the biggest players in the big data game — are going to take issues like bias, fairness, and inclusion seriously, they need to hire social scientists — the people with the best training in thinking about important societal issues. Moreover, it's important that this hiring is done not just in a token, "hire one social scientist for every hundred computer scientists" kind of way, but in a serious, "creating interdisciplinary teams" kind of kind of way.



Thanks to Moritz Hardt for the picture!

While preparing for my talk, I read an article by Moritz Hardt, entitled "How Big Data is Unfair." In this article, Moritz notes that even in supposedly large data sets, there is always proportionally less data available about minorities. Moreover, statistical patterns that hold for the majority may be invalid for a given minority group. He gives, as an example, the task of classifying user names as "real" or "fake." In one culture — comprising the majority of the training data — real names might be short and common, while in another they might be long and unique. As a result, the classic machine learning objective of "good performance on average," may actually be detrimental to those in the minority group.

As another, more contrived, example, consider the following randomized algorithm for fairly distributing birthday cake between six people: roll a fair, six-sided die and give the entire cake to the winning person. On average, everyone gets 1/6 of the cake, so, on average, the performance is fair; however, the probability that someone is unhappy is always one.

Within computer science, there's a lot of enthusiasm about big data at the moment. This is unsurprising and completely reasonable: we finally have computing infrastructure that supports the storage and analysis of massive data sets. This makes for some really exciting computational opportunities. But when it comes to addressing bias, fairness, and inclusion, perhaps we need to focus our attention on the *granular* nature of big data, or the fact that there may be many interesting data sets, nested within these larger collections, for which average-case statistical patterns may not hold.

As a machine learning researcher, working in the social sciences, I think this is an exciting shift in perspective, and one that raises very different, but extremely compelling, computational challenges: For example, how can we build models for drawing accurate conclusions about heterogeneous data? Should we develop aggregate models, perhaps inspired by ensemble methods, that account separately for smaller, nested subsets of this data? Can we do so without violating ethics or privacy concerns? And, if so, how can we build accurate models for these diverse, nested subsets?

I don’t have any concrete answers to these questions. However, when thinking about the last one — building accurate and reliable models for diverse, nested data sets — perhaps we should again be drawing upon the experience of social scientists, who already have significant expertise in this area. It’s often the case that social scientists want to answer questions for which massive, homogeneous data sets just don’t exist: for instance, there are political scientists interested in the relationship between gender and US Supreme Court opinions — i.e., the opinions of just nine individuals, of which even fewer are female. Other social scientists have multiple data sets with very different emphases, which, when combined, reveal a much more complete picture of some underlying social process. For example, consider a publication and a patent describing the same piece of work. Although each is separately useful, together they form a richer view of technological innovation. Finally, many social scientists work with “artisanal data” — i.e., data sets that have been carefully selected, hand-labeled, or curated in some way by a small group of experts for some specific purpose.

With that, I’m going to move onto my second talking point — questions — but before I do so, I want to briefly summarize my take-home messages regarding data. First, we need to start explicitly stating that big data, at least as the term is commonly used, is actually *granular, social* data. Second, in order to responsibly articulate and address issues relating to bias, fairness, and inclusion, we need to stop thinking of big data sets as being homogeneous, and instead shift our focus to the many diverse data sets nested within these larger collections. Finally, addressing both of these points necessitates collaboration with social scientists.

Questions

With a sufficiently large data set, it can be relatively easy to pick out a coarse-grained signal from noise with even the simplest mathematical methods. Discovering finer-grained signals, however, such as those often associated with data about minorities, can be much harder. Luciano Floridi, a philosopher, addresses this point in a recent paper of his. He says,

The real, epistemological problem with big data is small patterns. […] [But] small patterns may be significant only if properly aggregated. So what we need is a better understanding of which data are worth preserving.

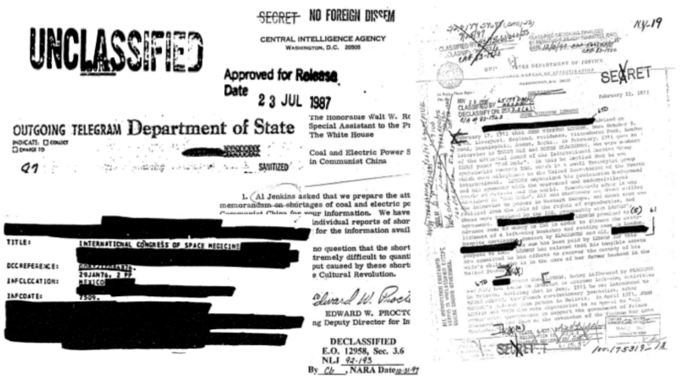
In other words, fine-grained patterns may not be readily visible using existing computational techniques. He continues, *And this is a matter of grasping which questions are or will be interesting. […] [T]he game will be won by those who 'know how to ask and answer questions.'*

Again, this underscores the need for social scientists, who are trained to ask and answer important questions about society; however, it also highlights one of my biggest concerns about some of the big data research and development coming out of the computer science community.

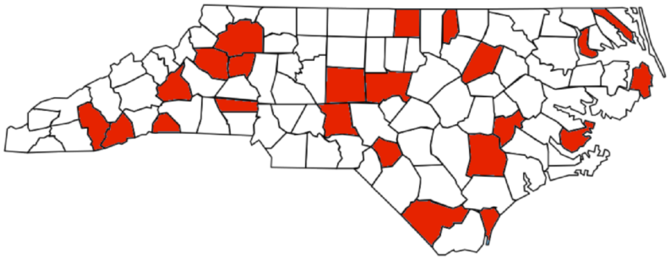
One of the most common driving forces behind big data research is the convenient availability of data sets. In other words, it's common for researchers or developers to encounter a new data set and then structure their workflow — be it designing mathematical models or developing computational tools — around figuring out some way to use it. A related, but different, approach, is to create a new model or tool — a hammer — in order to solve some abstract problem and then search for data sets — nails — with which to showcase it. While conducive to fast-paced work, these "data-first" and "method-first" approaches can amplify issues relating to bias, fairness, and inclusion. First, it's easy to immediately zone in on the coarse-grained patterns typically evidenced by the majority and then concentrate on analyzing those patterns instead of the fine-grained, harder-to-see patterns associated with minority groups. Second, it's extremely common to focus on only those data sets that are readily available, such as Twitter. The problem with these "convenience" data sets is that they typically reflect only a particular, privileged segment of society — e.g., white, Americans or young people with smart phones. Consequently, many of the methods developed to analyze these data sets prioritize accurately modeling that majority over other segments of the population.

As an alternative, I would advocate prioritizing vital social questions over data availability — an approach more common in the social sciences. Moreover, if we're prioritizing social questions, perhaps we should take this as an opportunity to prioritize those questions explicitly related to minorities and bias, fairness, and inclusion. Of course, putting questions first — especially questions about minorities, for whom there may not be much available data — means that we'll need to go beyond standard convenience data sets and general-purpose "hammer" methods. Instead we'll need to think hard about how best to instrument data aggregation and curation mechanisms that, when combined with precise, targeted models and tools, are capable of elucidating fine-grained, hard-to-see patterns.

As a concrete example, my political science collaborator Bruce Desmarais and I are currently studying the role of gender in local government organizations. This project is exciting for two reasons: First, there is very little data-driven work on government at the local level. Second, although there has been work in organizational science suggesting that women tend to occupy disadvantaged positions in organizational communication networks, this work mostly consists of individual- and firm-level case studies, rather than large-scale analysis of real-world data.

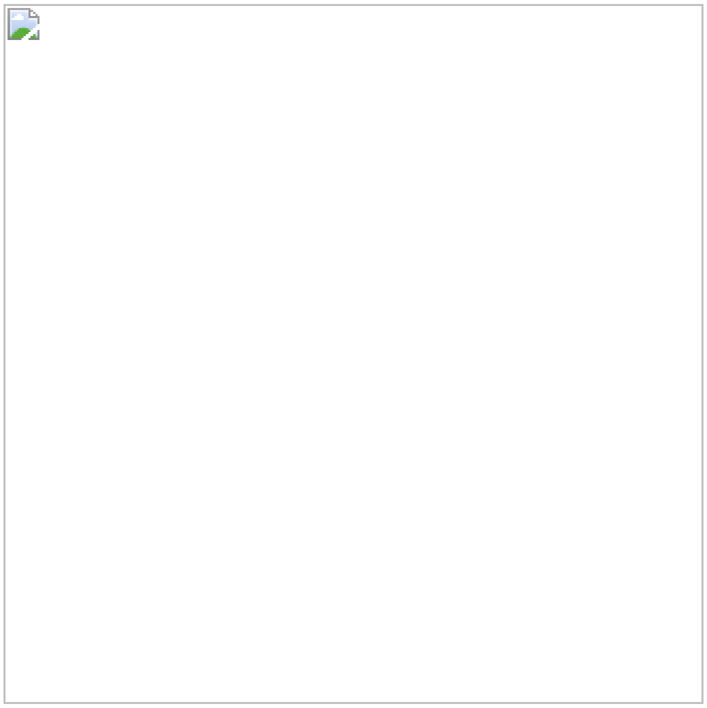

UNCLASSIFIED
~~SECRET~~ NO FOREIGN DISSEM
CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505
Approved for Release
Date 23 JUL 1987
OUTGOING TELEGRAM Department of State
SANITIZED
[redacted]
TITLE: INTERNATIONAL CONGRESS OF SPACE MEDICINE
DCCREFERENCE:
29JAN76, 2 PP
INFCLCCATION: MEXICO
INFCDATE: 7509.
[redacted]
The Honorable Walt W. Ro
Special Assistant to the Pr
The White House
Coal and Electric Power S
in Communist China
DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb, NARA Date 10-31-97
NY-19
UNITED STATES DEPARTMENT OF JUSTICE
SECRET
February 22, 1972
[redacted]
SECRET
100-175319-18

At first glance, this might seem like a hard area to study using a data-driven approach — these are not the types of question readily answered using Twitter or Facebook data. However, over the past few years, many "open government" data sets have been made available to the public with the stated goal of transparency. These data sets are instances of what I call "push" transparency — that is, government organizations proactively facilitated their distribution. Unfortunately, for our research questions, even these data sets (or the ones that we could find) are insufficient. But, as it turns out, there are other transparency mechanisms for addressing social questions, especially those relating to government — what I call "pull" transparency mechanisms. These mechanisms can be used as an opportunity to move beyond convenience data sets and even to request data that explicitly relates to bias and fairness. For example, most US states have sunshine laws that mimic the federal Freedom of Information Act. These laws require local government organizations to archive textual records — including, in many states, email — and disclose them to the public upon request. As a result, it's possible to obtain all kinds of local government data via public records requests, including data on bias, fairness, and inclusion. Of course, in order to do this, you have to know about these laws, how to issue a public records request, and so on and so on — all of which is arguably more difficult than pulling in data from the Twitter firehose, but may ultimately help address bigger societal issues.



In our case, Bruce and I (and our students) ended up issuing public records requests to county governments in North Carolina, asking for county managers' email spanning a period of a few months. We're now using this data to investigate whether women occupy disadvantaged positions in local government communication networks and, if so, the extent to which this varies with the topic of communication.

As another example of question-driven work that moves beyond readily available convenience data, I also want to briefly mention some really interesting research being done by Nick Diakopoulos and others in the computational journalism community. Fueled by concerns that businesses and governments are amplifying their already considerable power by their use of opaque and often proprietary, automated, decision-making algorithms, increasing numbers of investigative journalists are using reverse engineering techniques to obtain data that they can then use to understand algorithmic decision-making. This line of work — known as "algorithmic accountability reporting" — serves as another example of choosing to focus on hard questions about bias, fairness, and inclusion — in this case, the nature and extent of biased algorithmic power structures — rather than the convenient availability of digital data sets.



To conclude this section, I want to summarize my take-home messages about questions. First, rather than engaging in data-first or model-first research and development, we should prioritize questions — especially questions about bias and fairness. Second, although taking a question-driven approach can be more challenging, in that standard convenience data sets may not be appropriate, collecting data that pertains to bias, fairness, and inclusion will ultimately better serve everyone's needs.

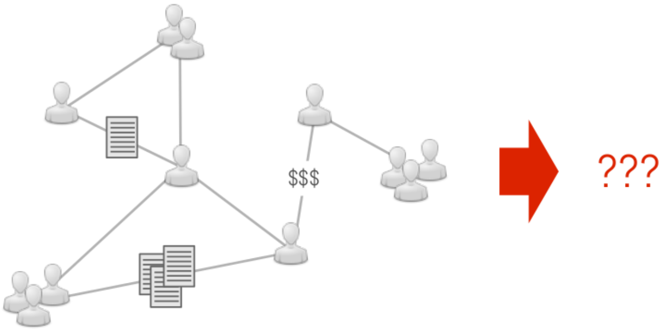
Models

As I was writing this talk, I noticed that many of the models discussed in the literature about fairness and transparency in machine learning are *predictive* models. I found this emphasis on predictive models particularly interesting because of a disconnect that I kept encountering when I first started working in computational social science. Specifically, I kept overhearing conversations between computer scientists and social scientists that involved sentences like, "I don't get it — why is that a research question?" And I couldn't work out why. But then I found this quote by Gary King and Daniel Hopkins (both political scientists) that I think really captures the essence of this disconnect:

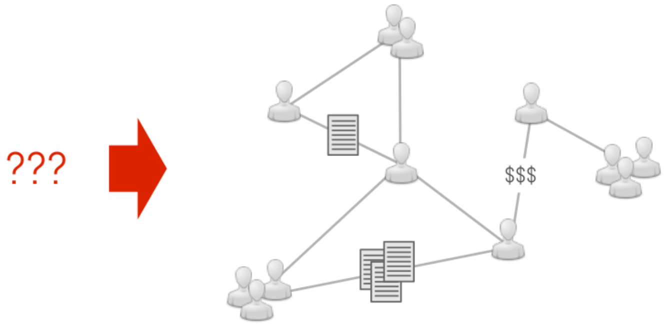
> *[C]omputer scientists may be interested in finding the needle in the haystack — such as […] the right web page to display from a search — but social scientists are more commonly interested in characterizing the haystack.*

After reading this quote, I did some thinking about the different types of modeling tasks typically undertaken by computer scientists and social scientists and — roughly speaking — I think they fall into three categories.

The first is prediction. Prediction tasks are all about using observed data to make predictions about missing information or future, yet-to-be-observed data. These are “finding the needle” tasks — in general, it is computer scientists and decision-makers who are most interested in them. Commonly used predictive models include logistic regression, conditional random fields, naive Bayes, Gaussian processes, and support vector machines.

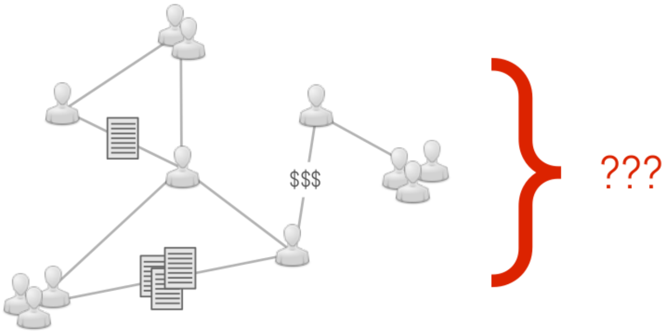


The second is explanation. Here the focus is on “why” questions —in other words finding plausible or probable explanations for observed data. These explanations can then be compared with established social theories or previous findings. Explanation tasks are therefore “characterizing the haystack” tasks and, in general, it is social scientists who are most interested in them. I’m not going to say much about explanatory models today — in part because they’re somewhat outside my area of expertise.



Finally, there’s exploration. Exploration is all about uncovering patterns in observed data — usually patterns that we don’t already know about. Equivalently, exploration is concerned with answering the question, “what do these data tell me that I don’t already know?” Tasks that fall into this category are also “characterizing the haystack” tasks, although — contrary to the quote by King and Hopkins — both computer scientists *and* social scientists perform exploratory analyses. Interestingly, though, neither group tends to explicitly acknowledge exploration as a “first order” modeling task. Commonly used exploratory models include latent Dirichlet allocation, factor analysis, and stochastic block models.

When thinking about the ways in which models for granular, social data can exhibit bias or unfairness, it's important to think not only about predictive models, but also about models for explanation and exploration. In part, this is because such models are extremely well suited to answering questions about granular, social data, but it's also because explanatory and exploratory analyses tend to be an important precursor to predictive modeling. The first step, when building or deploying a predictive model, is to decide which features of the observed data to focus on — that is, to choose appropriate input and output representations. To do this, we typically rely upon existing knowledge — often from previous explanatory or exploratory analyses, performed by ourselves or by others. Of course, the models used to perform these previous analyses — and any bias or unfairness in them — will necessarily influence the resultant findings and hence the representations we choose to use in our predictive models.

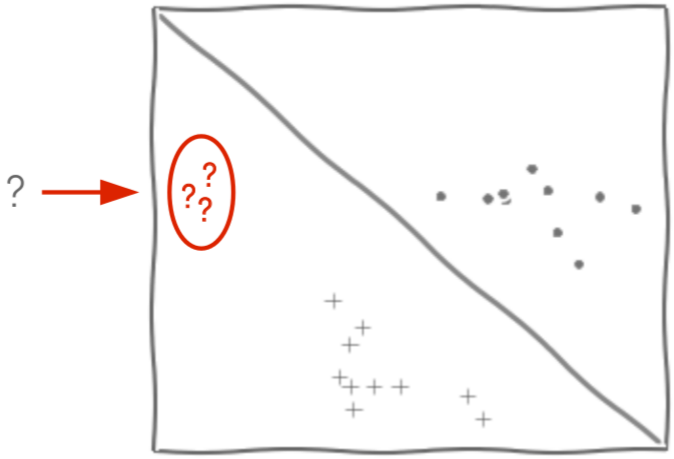


When reading about predictive models in the literature on fairness and transparency in machine learning, one of the things that really stood out to me was a repeated discussion of the dangers of unnoticed errors. As Moritz notes, when classifying user names as "real" or "fake," there's a big difference between a model that's 95% accurate because of noise and one that's 95% accurate because it nails classifying white, American names but achieves only 50% accuracy when classifying names from other cultures. This example clearly highlights the importance of careful error analysis.

I want to take this opportunity to note that error analysis needn't be a tedious exercise in "good practices" or "covering your tail." It can also be an opportunity to identify unexpected research directions. Personally, it is at this point — where I've run my experiments, seen where my model fails, *and then gone on to figure out why* — that some of my most innovative machine learning research has taken place. Moreover, if errors do relate to bias, fairness, and inclusion, fixing them will yield not only performance gains, but the kinds of performance gains that are actually important to society.

One way of preventing or mitigating errors is to explicitly represent and maintain uncertainty. For any predictive model, some predictions will be less confident than others. Moreover, it is likely that those less confident predictions will be about minority groups. To illustrate this point, I want to return to Moritz's observation that even for large data sets, there is always proportionally less data available about minorities, and that this data may even be substantively different to that of the majority. Error bars or some other representation of uncertainty can therefore help with error analysis by indicating when even correct predictions may be predicated on weak information. They can also be used by decision-makers in order to identify those decisions that may require additional human input.

For example consider this classifier with a decision boundary that does a good job of separating the data with which it was trained, represented by pluses and dots. How should this model classify these points out here? Granted, their perpendicular distance from the decision boundary is relatively large, suggesting that they may be pluses, but they're also in a very different region of space to that of the training data. As a result, it seems prudent to make less certain predictions about these points and — importantly — to report, rather than ignore, that uncertainty.



Many popular machine learning methods do not represent uncertainty when making predictions. Furthermore, even when using methods that *do* represent uncertainty, this uncertainty is often discarded. Sometimes these choices are justified by the argument that ultimately a concrete decision has to be made, so representing and maintaining uncertainty is an unnecessary waste of storage or computation. While this may be true for some tasks, I would argue that when building and using predictive models for social data—where the resultant decisions can affect real-world people — representing, maintaining, and reporting uncertainty along with any decisions should be standard practice. A lot of my own research has focused on Bayesian methods for exactly this reason.

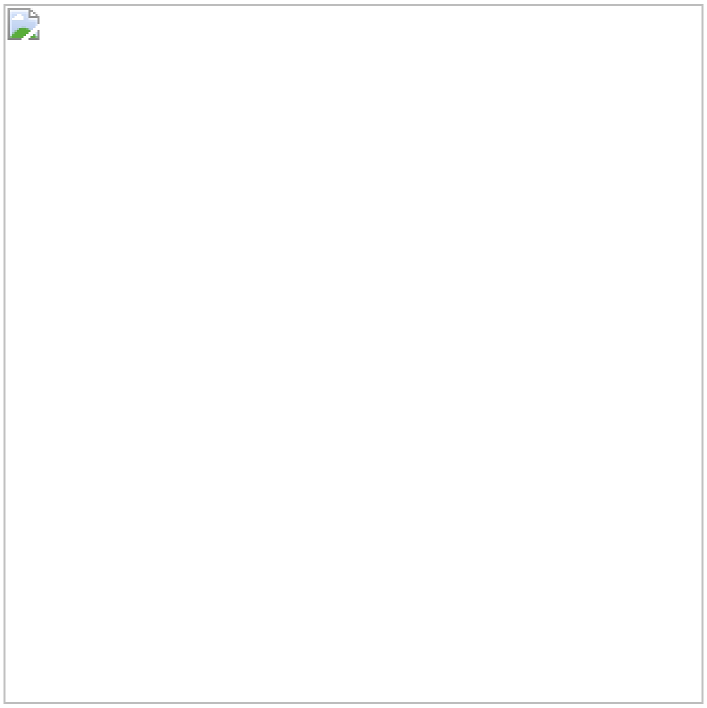
Model validation and error analysis are much trickier for exploratory models. Unlike prediction tasks, most exploration tasks don't typically have a single right answer, so it can be hard to know whether or not an exploratory model is doing a good job. However, rather than ignoring the problem entirely, I would advocate taking a multifaceted approach to model validation by identifying several different qualitative and quantitative tasks, each intended to validate a particular aspect or property of the model in question. For example, a good model for exploring network data should perform reasonably at link prediction, even though link prediction is not the ultimate modeling goal. Meanwhile, a good statistical topic model should exhibit coherent topics according to either an automated coherence metric or a domain expert. Both types of models should pass various posterior predictive checks. Taken together, these validation tasks can then be thought of as effectively validating the model as a whole. Poor performance on particular tasks can reveal information about the situations in which the model should and shouldn't be used.

I want to conclude this section by reiterating my take-home messages about models. First, when thinking about the ways in which models can exhibit bias or unfairness, we need to consider models for explanation and exploration, as well as models for prediction. Second, if we want to achieve fairness, we need to perform rigorous error analysis and model validation. Finally, we should use modeling frameworks that allow us to represent, maintain, and report uncertainty in decisions or other findings.

Findings

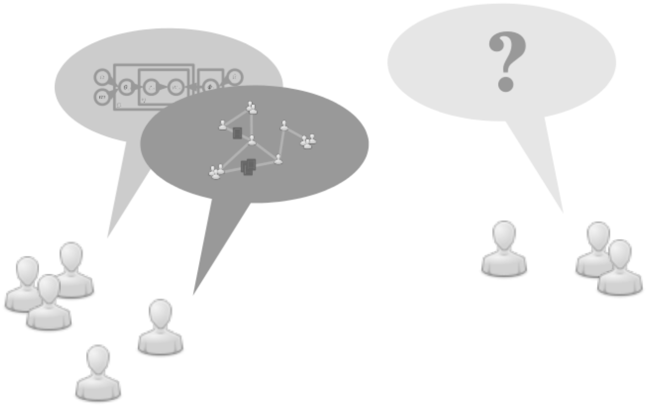
In the final section of my talk, I want to focus briefly on drawing responsible, fair conclusions from granular, social data.

Few computer scientists or engineers would consider developing models or tools for analyzing astronomy data without involving astronomers. So, why, then, are so many methods for analyzing social data developed without the involvement social scientists? I think, in part, it's because, as humans, we have strong intuitions about the social world. Duncan Watts' recent book, "Everything is Obvious," addresses this exact point — that humans are really good using at using intuition, at rationalizing, and at narrativizing. But, intuition is often wrong and narratives are not historical fact. For example, we all possesses attitudes, stereotypes, or other “cognitive shortcuts” that unconsciously influence our understanding, actions, and decisions. Some of these shortcuts may be positive, but others may be negative, especially towards people different to ourselves. Being aware that these “implicit biases” exist, and that everyone possesses them — even scientists — is an important step toward drawing fair and unbiased conclusions.



I would therefore argue that in order to aggregate and analyze granular, social data in a way that responsibly addresses issues of bias and inclusion, we need to make sure we’re working in a fair, accountable, and scientific fashion, rather than relying solely — or even primarily — on intuition. Returning to one of my take-home messages from the first part of my talk, one effective way to do this is to work closely with social scientists, who tend to be better trained at interpreting findings so as to acknowledge, account for, and challenge bias. Their training, as well as their research, can help us to be more aware of human factors that can influence both the process of drawing conclusions, as well as the types of conclusions drawn.

As an example, I want to highlight the phenomenon of confirmation bias. Confirmation bias is the process of seeking or favoring information that confirms one’s preconceived hypotheses, while ignoring or downplaying contradictory evidence. Although confirmation bias can occur with predictive analyses, it tends to be a bigger issue with exploratory analyses, where the goal is to answer the question, “what do these data tell me?” This can be especially problematic when working with large quantities of social data, not only because people often have stronger preconceptions, but because the sizes of and imbalances in these data sets can make it (comparatively) easy to keep searching until one finds some kind of supposed supporting evidence for these preconceptions. Addressing confirmation bias involves being willing to search for both confirmatory and contradictory evidence for hypothesized findings — a time-consuming process. Claude Steele, a social psychologist, has spent years demonstrating that academic performance can be harmed by the cognitive load of stereotype threat — that is, the fear of confirming a negative stereotype. For instance, when taking a math test, simply being reminded that they are female causes women to feel the threat of confirming negative stereotypes about women and math, and their test scores drop accordingly. This finding possesses a very different causal structure to “woman are inherently worse at math than men.” Teasing apart these alternatives required years of careful experimentation on the part of Steele and his colleagues, as well as the willingness to question a controversial, though common, belief about women and their inherent mathematical ability.



Finally, I want to return very briefly to transparency and accountability. Although I’m a firm believer in interdisciplinary teams, it’s likely that many machine learning methods will, at some point, be used by social scientists, policy-makers, or other end-users, without the involvement of those who developed them. If we want people to draw responsible conclusions using our models and tools, then we need people to understand how they work, rather than treating them as infallible “black boxes.” This means not only publishing academic papers and making research code available, but also explaining our models and tools to general audiences and, when doing so, focusing on elucidating implicit assumptions, best practices for selecting and deploying them, and the types of conclusions they can and can’t be used to draw.

With that, I’m going to end, but before I do so, I want to briefly summarize my take-home messages about findings. First, it’s important to be aware of concepts like implicit bias and to be willing to challenge negative stereotypes about minority groups, even when this requires additional effort. Second, it we want others to use our methods fairly and responsibly, and without our hands-on involvement, we need a community-wide increase in commitment to the public understanding of science.