Example "Big Data" Research Controversies

DRAFT VERSION

by danah boyd and Jacob Metcalf / November 10, 2014
Produced for Council for Big Data, Ethics, and Society[1]

This document is meant to provide some example controversies to spark a conversation about how to learn from existing controversies and the battles that rage around them. This list is by no means exhaustive and the categories that we're proposing may not be the right way of thinking about these issues. Read this document purely as fodder for conversation and to help trigger other ways of thinking about existing concerns and the cases that prompted them.

Some of the general questions to keep in the back of your mind include:

- What are the boundaries of research? What ethical concerns are unique to researchers and, in the realm of "big data," how do research practices interact with practices in other sectors (government, business, philanthropic, etc.)?
- Although there are some controversies that are purely about research, many of the major ethical issues emerge when research practices have impact beyond academia. What responsibility do scholars have for thinking about how their data and techniques might get used (or abused)?

Blurring Research and Practice, Science and Entertainment

Frustrated with their limited access to data, some researchers have begun creating companies or products in order to incentivize the public to contribute data to a research effort. These take on different flavors and have different levels of transparency. SETI@home is probably one of the most famous examples of a collective action science experiment where everyday people are invited to run software that turns people's personal computers into a supercomputer. Luis von Ahn, a computer scientist at CMU, is known for building platforms that incentivize large numbers of people to contribute actions and data to advance research puzzles; he sold reCAPTCHA to Google and started Duolingo as an independent company while continuing to operate as a professor. The former asked people to do reading tasks to help computer vision efforts while the latter invites people to learn a language while providing researchers with data about how language is learned. This kind of dual-purpose work is often celebrated by academia as beneficial to both researchers and the public, but it also raises questions about how people

[1] Funding for this Council was provided by the National Science Foundation (#IIS-1413864).

understand their contributions. Example:

- **23andme and "personal" genomics:** 23andme was created by Stanford bioinformatics PhD students who were frustrated by their inability to get enough research data to do their analysis. With venture money, they started a company to invite people to share genetics data for research by offering them a public service in the form of personalized information and simple social and genetic networking features (e.g., "find lost relatives"). They also gave people a way to make sense of the wide range of scientific literature on genetic markers by connecting published journal articles to personal markers in a person's sequence. Because of the cost of genetic testing when they started and their personal networks, their early adopters were primarily science and technology geeks who understand the limitations of the research, their genetic markers, and basic probabilities; they contributed their data because they wanted to advance science and were curious to learn what was found. As the service expanded and testing became cheaper, their customers began to use the site as a predictive tool, devoid of any understanding of what information they were getting or the scientific inquiry underpinning the service itself. Controversially, the FDA shut down key parts of the company for providing unregulated medical equipment, prompting 23andme to claim that they were an entertainment company, not a scientific one. This service raises a series of ethical questions, ranging from the role of consumers in the scientific process to the ethics of medicine vs. entertainment.

Immutability and Blurring of Records

Longitudinal datasets are very valuable for researchers and researchers have long used government datasets to do analysis, from census data to public health data. Although there have been plenty of controversies around these datasets, the processes put in place are both rigorous and, in the case of the census, protected by law. Increasingly, the "big data" datasets that researchers are working with and the analysis that they're doing significantly affect the populations that they're studying, blurring the lines between research and practice. Government datasets are fed into industry tools; industry datasets are sold to data brokers and fed back into tools for government usage. Social media datasets are given to the Library of Congress. Researcher datasets are fed into both proprietary and government algorithms. The level of controversy surrounding this practice depends significantly on the domain – criminal justice, human rights, corporate marketing, etc. The same tools and data that can be used to empower are also feared as potentially harmful in other contexts. Data that was collected for one purpose is made persistent and is often immutable, which can be particularly harmful when it gets fed into another system with directed consequences. How do we account for ethical considerations when the context surround data collection and usage changes? What is the responsibility of researchers in light of their efforts to design and build tools to collect and analyze data? Example:

- **inBloom and Student Data:** Education researchers have long clamored for access to more data in order to develop new ways to improve education and learning. For a long time, this got tethered together with testing, as researchers sought tests alongside governments and other interested parties. Increasingly, the possibility of scaffolding data

analytics tools into the classroom to enable personalized learning has prompted both excitement and horror among researchers. These conflicting attitudes boiled over and mixed with corporate interests, policy advocacy, parental concern, and government actions surrounding a controversial philanthropically backed company called inBloom, which had brokered deals with various governments to collect and store student data. The issues that were raised ranged from concerns over monied interests to anger over the inability to monitor and correct third party datasets about students that may be used beyond their intended goals.

Experimentation and A/B Testing

For many scientists, experimentation is considered a commonly used and acceptable method of inquiry. When applied to data analytics, this often takes the form of a practice known to computer scientists as A/B testing. Because A/B testing is common in software development, it is often not treated with the level of ethical consideration and rigor that is typical in scientific inquiry. These two approaches come into conflict when the data being invoked is related to people. This issue is significant in academia, but it is especially acute when academic inquiry intersects with industry, raising a host of questions both about ethics and procedures. Example:

- **Facebook's Newsfeed/Emotional Contagion Study.** Facebook's 'Newsfeed' is algorithmically produced, thereby determining what users see from their 'friends' based on a variety of input, most of which is not publicly known. The introduction of this function was itself controversial, and the research experiment that underpinned the emotional contagion study (conducted by both academics and corporate data scientists) was dependent on Facebook's ability to manipulate the Newsfeed. This erupted into a large-scale controversy with issues as varied as the role of an IRB, the framing of the study itself, the collaboration between industry actors and academics, the ethics of manipulating users for research, and much more. Commentary.

Networked Consent

Consent is often understood as the gold standard for doing ethical research, but consent is not always viable and not all research can work under consent (e.g., discrimination studies). Academic IRBs have protocols for navigating data collection procedures where consent isn't appropriate or possible. Increasingly, data scientists are turning to "found" data, creating new challenges for thinking about how to achieve consent or how to think ethically about the people behind that data. Although this data is sometimes collected for scientific inquiry outside of academia (e.g., the Facebook emotional contagion study), this data is often collected for a different purpose and then the question is how to use the dataset once it's produced (e.g., AOL's search data set, Enron's data set, Orange Mobile's data set). In some cases, people are purportedly protected through anonymity, but controversies erupt when researchers re-identify people in these datasets, prompting questions about how privacy-protective they are. In other cases, the people are known, raising questions about how much say they should have over information that is in the public. Which researchers should use this data, under what circumstances, and with what protections is not always clear. Example:

- **Henrietta Lacks:** After journalist Rebecca Skloot published *The Immortal Life of Henrietta Lacks*, researchers and the scientific community began to think more deeply about the ethical issues raised by the HeLa cell line. At the time, there was particular concern over genetic studies that were being done with the HeLa line without Lacks' ancestors knowing of the studies until after research about their genetic makeup was published. In response to the conversation that unfolded, the National Institute of Health worked with the family to develop a protocol for handling HeLa genome data that allowed the family to be involved in the research process. Link

Algorithmic Accountability

In "big data" systems, algorithms play a significant role in determining the output of what a system does. Algorithms are typically designed to maximize for a desired end goal (typically defined by the developer), but dependent on the type of algorithm and the complexity of it, figuring out how to hold it accountable and to assess its biases can be tricky at best. This is particularly notable in any system that relies on machine learning techniques where even those designing the systems are often unsure of how the systems are evolving. Furthermore, because it is not possible to test for every possible outcome of a system, choosing which tests to use becomes an important part of the development process. While accountability is challenging for all algorithmic development, this is particularly true in complex systems that involve dozens if not hundreds of programmers, making it impossible for any one person to fully understand the code at play. Such is the case in many long-term collaborative projects in both industry (e.g., a search engine) and science (e.g., NASA R&D projects). It is also important to take into account the various ways in which these systems can be gamed. Example:

- **Twitter's 'Trending' function.** Twitter's researchers designed the "trending topics" feature to help understand what topics were gaining the most traction on the site. Although an internal tool at first, they were given clearance to turn it into a product with the idea that this type of analytics may be of broader interest. This, in turn, prompted users to game the system in the hopes of shaping the trending topics, which forced the designers to alter their algorithm. Because the deployment of the feature suggests that these are the most popular topics on Twitter, there is constant outrage over the manipulation of the system. At a most basic level, few people realize that the goal was to present second order changes (a.k.a. spikes over slow builds) and so there is constant outrage when a topic that slowly gained traction never hits the Trending Topics even if it is one of the most discussed topics on Twitter. While the mechanism is reasonable from a research or feature perspective, the algorithm produces certain cultural effects that prompt others to be critical of the social impact of the company's research activities. Controversies have emerged when political topics like elections or Ebola do not trend or when trends have prompted cross-cultural conflict. Link