Workshop Primer: Inequalities and Asymmetries

The Social, Cultural & Ethical Dimensions of "Big Data"
March 17, 2014 - New York, NY
http://www.datasociety.net/initiatives/2014-0317/

Brief Description

The availability of data is not evenly distributed. Some organizations, agencies, and sectors are better equipped to gather, use, and analyze data than others. If data is transformative, what are the consequences of defense and security agencies having greater capacity to leverage data than, say, education or social services? Financial wherewithal, technical capacity, and political determinants all affect where data is employed. As data and analytics emerge, who benefits and who doesn't, both at the individual level and the institutional level? What about the asymmetries between those who provide the data and those who collect it? How does uneven data access affect broader issues of inequality? In what ways does data magnify or combat asymmetries in power?

Detailed Topic Description:

Thus far, the "big data" phenomenon has primarily benefited the financial, technology, advertising, and defense sectors. The organizations that have statistical expertise, technical resources, and access to data tend to be those that have tremendous public or private resources available to them. While other sectors - like healthcare, transportation, and education - are beginning to recognize the potential of data mining, they have not yet implemented the kinds of systems that are standard fare in more advanced sectors.

Data analytics has the potential to transform many areas, but significant issues arise when mining techniques are unevenly distributed. The potential of large-scale data mining in various sectors is notable, but it also raises significant questions, particularly when technologies implemented to help the public-at-large could also be used to target individuals. The recent exposure of unexpected ways in which the public is tracked and targeted by advertisers and law enforcement has raised concerns about the potential for unfair, coercive, or inappropriate collection and use of information in other social domains.

For example, Siemens and the U.S. Department of Transportation initiated a program in Texas in 2011 that uses cell phone signals to regulate traffic. Using their navigation systems or smartphones, drivers were able to determine the fastest route. Thanks to the signals put out by drivers' devices, traffic lights could also be adjusted according to traffic flows. Citizens benefit when transportation is more seamless and commutes are reduced. Even so, this information may be put to use far beyond its original purpose. Should the data be accessible to automobile insurance companies and law enforcement? Who will be targeted as a result of "more perfect" implementations of regulation through technology?

Education is another sector in which the potential transformative value of data analytics must be understood alongside the potential abuses. While data-driven instructional technology is often heralded as a means of empowering students, it may also have unintended adverse effects. Projects underway often presume the former. Ed-tech advocates often promote their programs as permitting personalization that will increase student engagement, individualize lessons according to skill level, provide teachers with detailed assessment, and supply pedagogical and curriculum feedback for educators and researchers. For example, a collaboration between IBM and Mobile (Alabama) County Public Schools is designed to identify trouble spots automatically rather than relying on student self-reporting. Researchers can then compile this data from thousands of schools to assess which particular lessons need more work. Individual students who are too shy to articulate or too confused to identify their difficulties may benefit from increasingly tailored instruction and curriculum developers can pinpoint lessons in need of revision. IBM claims that this sort of tracking enables researchers to predict which students will complete math problems at satisfactory levels, and permits teachers to provide early intervention to at-risk students. The promises and potentials of transforming learning are driving foundation investment in student data and educational interventions.

Yet, there are also potential downsides to integrating data analytic techniques into the education sectors. Tracking - or placing students in sets of classes according to perceived ability - has a long history in American educational systems. In theory, this was designed to benefit those who needed additional help. In practice, it became a mechanism of segregating youth. Studies have shown that early childhood tests are low predictors of potential, while tracking has mental health and socialization implications. The use of data analytics in schools could uncritically reinforce existing tracking procedures. Conversely, a constant stream of assessment and feedback may provide a more accurate picture of student progress than a one-time test. It's not clear what will be most beneficial for individual students and the educational system overall.

Additionally, parents may not have the same access to technology that teachers do and may be unable to view the recommendations that educators make. A disadvantaged population may not have means of understanding programs' terms and conditions, meaning that parents may not fully understand what they are agreeing to in having their children tracked at school. How can this process be clarified so that both the educated and uneducated consumer can make informed decisions?

Analyzing and predicting student success based on prior performance also raises the issue of aptitude versus passion. In many countries, aptitude tests affect the opportunities students have to pursue particular career paths, military roles, and educational openings. To what degree should students' paths be shaped by their abilities versus their desires?

There is rising concern that the "big data" phenomenon has the potential to amplify inequalities rather than solve them. Data tracking is expensive, requiring massive amounts of infrastructure as well as human labor. While security and defense systems may have the requisite funding and technology, this is not necessarily the case when it comes to education, healthcare, transportation, and social services. Even if data is collected, it still needs to be interpreted. This also requires the right tools and personnel, meaning that gathered information may sit unused. Researchers need to ensure that they have the tools and resources to account for every variable. Numbers can also obscure other social, cultural, and economic factors. Who decides how these numbers are used and what they mean? How can the financial sector and local communities implement these tracking methods and also provide the infrastructure and training necessary to correctly interpret and use the data?

While data tracking may benefit researchers, corporations, or government agencies, it is unclear what impact it will have on individuals. If individuals are unaware of how they are being tracked or how the data is being used, they may be blissfully clueless or increasingly fearful of existing institutions. Increased transparency may lessen the gap or it may become disempowering if people feel as though they can't use this knowledge. For example, researchers monitored social networks in Chicago and used them to compile a list of the 20 individuals most likely to kill or be killed in the area. Such measures may help protect against further violence, but they also increased the surveillance of already marginal populations. The subjects being monitored didn't have access to how those data points were created, reinforcing the power differential between those being tracked and those gathering and using the data. People in vulnerable positions are often compelled to share data by law enforcement, employers, and institutions and almost never get insight into what happens to that data. One example is the gang databases maintained in most major us cities. Individuals may be

put on the list for a variety of reasons - search entry terms or gang symbols or dress or social networks. Even if you aren't a gang member or have reformed, your data persists in the database. The Rampart list from an LA police program, which was disbanded because of concerns regarding racism, still exists even after it was proven to factor into wrongful convictions. Durable information with racial, gender and class assumptions may restrict marginalized groups' access to upward mobility.

Furthermore, the potential disconnect between individuals who are being monitored and those who are collecting and interpreting data may unintentionally widen existing inequalities, even when the goals are to address societal inequality. The technologies used to gather and safeguard information about individuals and groups have values embedded in them. Technology can also make assumptions about people and may reinforce existing social inequalities. Because of the assumptions built into technology a small group of engineers can have an enormous impact. This is complicated by the fact that certain groups are underrepresented in the engineer community and will not be represented by new technologies. What can researchers, educators, and government officials do to ensure that typically invisible populations are represented by data analytics? How can designers and developers create programs that will benefit marginalized groups or communities as well as affluent ones? What technological infrastructures are needed to implement the use of data analytics in sectors like education, healthcare, and transportation? How can researchers tie abstract data to culturally and geographically specific elements? If aggregated data leaves gaps when it comes to certain communities, how can researchers attempt to fill those in?

Data is being collected from a wide variety of places and there may be asymmetries between different datasets and data brokers. For instance, which organizations and researchers have the ability to combine user-generated data from applications like Fitbit and genomic information? Informally trained data workers may shape how data is collected and interpreted, having potentially far-reaching consequences. For instance, the EPA encourages citizen scientists to monitor air quality. Do these new opportunities help correct power imbalances between individuals and government agencies?

Finally, it is important to address public wariness regarding the widespread application of big data as a tool of power, for both ethical and pragmatic purposes. Generally speaking, we lack a vocabulary for discussing inequality and power differentials with regard to data analytics; power, not just privacy, is an ethical issue. Vulnerable members of the public and civil rights organizations may opt-out or actively challenge emergent data practices if appropriate safeguards are not incorporated from their inception.

Case Study 1: Reproducing Civic Inequalities

The city of Boston implemented Street Bump in order to flag potholes and to expedite the repair process. Bumps are registered when drivers with smartphones placed on their dashboards drive through the city and hit potholes. As media scholar Kate Crawford pointed out, however, the "digital divide" exists even within major metropolitan centers in the US, meaning that big data can leave glaring gaps. The reports were mostly from areas with high concentrations of smartphones, meaning that wealthier locations were more likely to receive attention than poorer areas. Similarly, elderly people are less likely to have smartphones and thus were unable to contribute to the pothole map. While theoretically this measure should have benefitted all Bostonians, marginalized groups were left out because of their lack of access to requisite technologies.

Not only did Street Bump leave out information from poorer areas of the city, the fact that wealthier areas received more attention could actually exacerbate existing inequalities. "So if you think about how this might be used to fix roads, we might see a future where the wealthy areas with young people get more attention and resources, unlike the areas with older citizens, who might get fewer resources," notes Crawford, "So if you're off the map, this could have some really material consequences for social inequity." Despite having the best of intentions, Street Bump's originators may have inadvertently contributed to widening the gaps between the rich and the poor or the young and the old. At least, as Crawford notes, "Boston's Office of New Urban Mechanics is aware of this problem, and works with a range of academics to take into account issues of equitable access and digital divides."

Relying on smartphone use, or other expensive devices or working knowledge of such tools, means that certain groups will be left out. How might both city officials and application designers work together with communities to ensure that marginalized groups are not left out? While technologies like smartphones are a boon to researchers, what happens when assumptions are made about their ubiquity and ease of use?

Case Study 2: Metadata and Social Networks

In the 1979 case *Smith v. Maryland*, the Supreme Court declared that metadata about communications is not subject to the same protections as the content of those communications. The original decision was based on the particular workings of the rotary telephone. Pen registers, or electronic devices that record all phone calls made from a particular number, were not deemed to constitute a search according to the Fourth Amendment, and could thus be installed without a warrant. As technology has

changed, this precedent has been expanded to include mobile phones and internet-based communications, which allows for the accumulation of much more extensive kinds of metadata than landline phone records alone. Revelations associated with documented evidence provided by Edward Snowden indicate that the NSA is regularly using metadata to analyze information about who communicates with whom. Privacy advocates and computer scientists, both of whom recognize just how much information can be discerned by metadata alone, are outraged by this revelation. MIT's Media Lab released Immersion, a tool that shows users just how revealing metadata can be, in part because of its searchability. A recent Stanford study shows that phone record metadata can be used to identify information about individuals including gun ownership and religious affiliation, as well as sexual, financial, political, professional, and social associations. Indeed, part of the power of social network analysis is that the graph of social relationships can be hugely informative for anything from targeted advertising to criminal interrogations.

Public health researchers have long found that social network analysis is valuable for understanding populations and deploying interventions for everything from smoking cessation to sexual reproductive health education. Given the role of personal networks in socio-economic status, poverty researchers have also turned to social network analysis to guide action-oriented projects. For instance, researchers have found that social networks strongly influence hiring practices. In Chicago, social networks, more than structural factors like race and poverty, were found to determine how likely someone was to be the victim of gun violence. Groups that work towards developing social services do not have the same level of sophisticated network analysis tools - let alone the data - as intelligence agencies. What does it mean that government agencies are more likely to collect and use personal network data for law enforcement, defense, and intelligence than to address structural inequalities within society or implement social interventions? If citizens are accustomed to having metadata used against them, how can researchers and officials use the valuable information produced by social network analysis to minimize social inequalities or solve community issues without causing alarm? What is needed for metadata analysis to be implemented outside of the defense, advertising, and security sectors?

Case Study 3: Health Data Analysis

Public health researchers have long found that social network analysis is valuable for understanding populations and deploying interventions for everything from smoking cessation to sexual reproductive health education. The tracking of individuals and of populations can provide information about the spread of disease and the

likelihood of illness. While government agencies have long worked to do this kind of analysis, corporations are increasingly having equally (if not more) reliable data. Some findings point to Google Flu Trends being potentially as accurate as the CDC.

Sometimes, however, the use of data tracking can lead to incorrect assumptions or faulty information. For instance, Google Flu Trends overestimated peak flu levels in 2013. Monitoring flu-related search terms does not take into account the impact of news stories or other media on such searchers. Just because one Googles "achy, fever, flu" does not mean that one is actually sick with flu. What is the societal implication when Google - who does not specialize in verifying the accuracy of its health data - is more widely recognized by the public than the CDC? How should the government respond to private analysis of public phenomena?

As different organizations start to amass data about people's health and societal disease tracking, who is responsible for piecing it together both on an individual level and for society as a whole? Should Google's data be turned over to researchers for verification? What are the implications for data being released into the public when the public may not be qualified to interpret what they are given? Should individual physicians trust tracking information provided by patients using unregulated systems?

Microbiologists at Harvard's School of Public Health have petabytes of raw data that could be used to prevent TB outbreaks. Health researchers have the ability to use data analytics to solve major crises all over the world, focusing on epidemics in the Global South. The information could save many lives, offering new means of diagnosis, treatment, and even the possibility of a vaccine. Unfortunately, analyzing the data was a difficult task, requiring lots of labor. The head of one microbiology lab at HSPH, Sarah Fortune, decided to crowdsource some of this labor, enlisting volunteers to measure and label the distance between cells, a task too complex for computer algorithms alone. One thousand volunteers agreed to the project, although none of them had scientific backgrounds. While crowdsourcing is a creative way of handling the enormity of such data, what are the risks of having non-professionals engage in this labor, especially as unpaid participants? Who is verifying their measurements to ensure their accuracy and how are they being trained? These tactics have the ability to reduce structural inequalities, but there are also risks of the data being misinterpreted or mishandled by volunteers.

As the ability to collect, use, and interpret data is open to more people and organizations, who is assessing the emergent inequalities? What kinds of data asymmetries exist? Who benefits and who loses?

Questions to Consider

- What are the major social, cultural, and ethical tensions that emerge when thinking about data-related inequalities and asymmetries? What needs to be better understood to address what's happening?
- What conflicting values and tradeoffs are at stake? How do we understand relevant actors, stakeholders, and "camps"?
- How do inequalities play out differently in different domains (e.g., social services vs. health care vs. marketing vs. intelligence)? In particular, what aspects of power are at play?
- Should data aggregation be treated different in different domains? What is the role of transparency?
- How are societal values implicated? What does it mean that intelligence and marketing have greater access to data and analysis than other sectors?
- What are additional salient case studies that highlight what's at stake, where lines need to be drawn, and how we should be thinking about empowering vulnerable populations?
- Who should be responsible for addressing data and analysis divides? What is the role of the government? Of corporations? Of data providers? Of technologies and tools? Of educational institutions? Of media institutions? Of civil rights organizations?
- What structures should be put into place to make certain that divisions are being addressed? What can be done to empower vulnerable populations before they are further marginalized?
- Should data aggregation be treated different in different domains?