Workshop Discussion Notes: Data Supply Chains

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

This document was produced based on notes that were taken during the Data Supply Chains workshop as part of "The Social, Cultural, and Ethical Dimensions of 'Big Data'". Not all attendees were involved in every part of the conversation, nor does this document necessarily reflect the views and beliefs of individual attendees. All workshop participants were given workshop materials prior to the event to spark discussion. This primer can be found at: http://www.datasociety.net/pubs/2014-0317/DataSupplyChainsPrimer.pdf

Overview

Throughout the group's discussion, several themes emerged. Workshop members spent a lot of time parsing the notion of a "data supply chain," and connecting the metaphor to the history of traditional supply chains while sorting out the various actors involved in the process. Some participants debated the use of "creepiness" as a metric when it comes to data supply chains. Some group members discussed creepiness as being contingent upon context and perception. Another theme revolved around individuals' access and awareness of different nodes in the data supply chains. Can individuals access their data, and are they able to integrate data across various platforms? Participants also debated the problems associated with the visibility of the data subject across supply chains and data retention, as group members wondered about data labor practices and the overlap between public and private data. While individuals may know that their data is used in one context or data environment, what happens when it changes hands? Data may persist for longer than originally intended and may end up in new hands or may be used for different purposes, which may lead to a heightened creep factor.

Some group participants also addressed the question of, how visible and accountable are data brokers and third-party platforms in the data supply chain? Certain actors' visibility may lead to increased responsibility, but those who are invisible may escape the attention of regulators. How do we interface data supply chains with the people who are affected by their existence and use? Some participants also raised the matter of trust when it comes to data supply chains. What happens when public and private data sources are blurred or passed back and forth? Should there be contexts and data types with strict normative rules about their collection, sharing or commodification, i.e. when it comes to DNA information?

Workshop participants found several major sites for further exploration. Some group members discussed the need for balance between visibility and accountability. Because there are new forms of data labor, some actors like data brokers are invisible and thus not subject to regulation or held accountable. Who does the work associated with data supply chains and what is the product? How does information move around? Some group members also discussed other possible metaphors for thinking about data analytics aside from data supply chains, suggesting ecosystem as an alternative. Finally, the group visually mapped the data supply chain in order to better understand the relationships between various actors. Different actors may have multiple and shifting roles inside the chain. How do we make sense of these different sets of relationships?

Themes and Discussion Topics

Understanding data supply chains

A good portion of the discussion in the working group was focused on better understanding data supply chains. Members raised many questions and points with respect to the actors and stakeholders of data supply chains. Some members asked about oversight, wondering who is responsible and accountable for the supply chain and accountable to data subjects. Data moves across the supply chain and is handled by different entities, but many of these entities do not interact directly with data subjects at all. Some group members discussed different ways that policy might be informed by a stakeholder process, as a way to allocate responsibility and facilitate mechanisms for due process in data supply chains and with respect to “big data” in general. Some participants noted that data brokers or less visible actors may not be known stakeholders in the data supply chain, thus allowing them to circumvent legal and ethical responsibility. Workshop participants also discussed the possible unintended social consequences of data supply chains. Organizations within the data supply chain and those who are external to it may be subject to different consequences. Some participants offered that it is also possible that adding new information to existing data sources will have unforeseen downstream consequences.

Group members also addressed the role of data brokers, as some participants cautioned that data brokers are generally misunderstood or overlooked actors. How visible and accountable are they in the data supply chain? Some group members discussed the possibility of coming up with a framework that would allow companies to ethically, legally, and morally do the right thing in data supply chains. How can we ensure that any framing of data supply chains also entails a reasonable understanding of the technology that underlies their workings? Can consent mechanisms be useful and empowering given the flow of data in a data supply chain and its consequences for the data subject? How can we factor context into data supply chains? How can we bring together the objective description of the data supply chain and yet integrate the subjective experiences of people with the data supply chain at particular points?

Creepiness

Some participants raised the issue of "creepiness" in an effort to locate what was at stake. What is it about information moving across data supply chains that sometimes makes people uncomfortable? Surveys and other studies show that consumers, citizens, and data subjects find many practices associated with the big data phenomenon creepy. Creepiness is a term that often gets used to express the discontent that data subjects have with the data collected, processed and shared in supply chains. Specifically, data subjects may find it creepy when they notice that their perception of the data supply chain is different from the actual dimensions and workings of the data supply chain.

Some group members discussed the importance of context, as something may seem creepy and then become normalized or vice versa. With time, the creepiness factor may decrease: months after the revelations, many people may shrug off knowledge about intrusive tracking or government surveillance. Is the creepiness factor then a reliable metric for the existence of ethical, social, cultural and political problems? The creepiness factor arises when data traces people leave behind are linked with something else. Real inferences are thus a source of the creepiness factor, some participants explained. It is creepy when companies can infer that a person is pregnant, has HIV, or has not been taking their medication, from 'data'. On the other hand, a pro-active approach to inferred associations could result in something like at-risk veterans being asked to carry wearables because they are more likely to be depressed or suicidal. Would such tracking be ethical if it could be shown to potentially save their lives? This and other examples or framings, such that an issue that could be creepy actually becomes productive, may have social value. If you change those framings even slightly, however, people may become uncomfortable. How do we deal with the contextual aspects of creepiness?

Creepiness may be a side effect of surprising the user with unexpected data collection and use. Companies and governments heavily engage in perception management in order not to surprise the user. How can we go beyond thinking about privacy as an illusion that we maintain through perception management? In further discussion of just what creepiness is made of, some participants offered that surveys may help express how people feel about data practices, like finding certain data practices creepy. However, some group members argued that they are not good at capturing people's everyday practices and the role that data play in these.

Access and integration

The question of access to data as it moves down the data supply chain also emerged as an issue throughout the discussion. Members of the group debated who has access to the data in a data supply chain and who should have access. Which other actors, other than the data collecting and sharing entities, should be guaranteed access to a data supply chain? The actors in question included, but were not limited to, the data subject herself, journalists, scientists, as well as members of citizen science initiatives. When thinking about these actors, some participants emphasized the increasing ambiguity between professional and amateur initiatives, as well as the erosion of the distinction between public and private actors as a result of the fact that more and more data is held by private entities. Specifically, the following questions were raised with respect to guaranteeing and opening access to data supply chains: How can citizen science and other distributed knowledge making practices be sustained, as data supply chains get instantiated? What are some mechanisms to counter privileged access to data that may be architected into data supply chains? How can we ensure that data subjects have access to and control of their own data? Is it always desirable and ethical that data subjects have access to this data? Participants also discussed the possibility of open data, debating who should have access and what kinds of information should be accessible. FOIA, Freedom of Information Act, requests are one way of getting access to data. Yet, such requests are limited to public entities: private entities are not subject to FOIAs. Even further, sometimes organizations that do make use of FOIAs to access government data may not release the data to the commons. What are some appropriate licensing schemes that would ensure that FOIA and open data become part of the data commons?

Participants also addressed the question of how data subjects can integrate their own data. A data subject may want to use data from one service, e.g. FitBit, with another service. For this to work, the individual needs to have access to her own data repository in an interoperable fashion. Depending on the sensitivity of the data and the purpose of its collection, the reliability, accuracy, security and ethical use of this data may gain greater eminence for the data subject. If companies start offering discounts or premiums in return for medical data, will data subjects be incentivized to share sensitive health data? Should the sensitivity of the data that is collected be an element of whether or not that data can be collected and aggregated with other data? The institution that the data subject has a relationship with may be the one that introduces a third-party service, e.g., a university may distribute FitBit or educational products to its students. As a result, this institution may now be sharing data with the third party or combining data from multiple sources. Will this combined data be accessible to the data subjects? Will they have a say over and knowledge of where else this data flows? What if the recipients of such data are mandated to report to the government? If so, what difference would it make if the data was used to report assessment-driven decision making for educational policy versus for the commercialization of the school system? Is it possible to prohibit the sale of certain data in certain contexts, even if data subjects want to sell this information? Institutions may compel data subjects to participate in the data collection and processing, so individuals may consent due to lack of choice. Should it be possible for people to say they want to share their data but they do not want to let certain things happen with it? Some group members argued that individuals are no better than committees at making good decisions. Sometimes people need to see in practice what happens with their data before they can make better-informed decisions. How could this kind of transparency be arranged? How can we capture and address the unforeseen negative consequences associated with allowing data subjects to use or sell their own data?

Data on the move

Group members also discussed the issues of visibility and data retention in relation to data supply chains. Some actors in the data supply chain are more public than others. Alternatively, some people may be more visible or vulnerable to data supply chains. As time goes by, actors may be introduced or taken out of the supply chain and the data supply chain may change shape, purpose, and effect. Those actors that are most visible in the data supply chain are more likely to be held accountable for their actions than those parties that are less visible or invisible. If invisibility leads to a lack of accountability, what are some measures that we can take? Which actors within the supply chain are visible? Are the data visible? Are the transformed data visible in the downstream data supply chain? Is the entire data supply chain visible? What data subjects are visible or invisible? The logic of efficiency and concerns about privacy may be in tension. People may have time constraints or may not keep up with the speed and massiveness of data supply chains and hence may be unable to react to privacy issues. Can transparency and visibility within the data supply chain empower the individual data subject or communities of data subjects?

Some participants also addressed the issue of trust with respect to data as it changes hands. What are some of the challenges of being a good data guardian or caretaker in a data supply chain? Certain companies may be more trustworthy than others and can use trustworthiness as a branding technique. Some group members discussed the possibility of having penalties for those companies who have not been honest with their data practices. Data use agreements (e.g. terms of use or privacy policies) can function as a barrier to truth or can function as pacifiers to show the public that the data is being handled well. Most of the time people don't have a choice with respect to privacy as they do not have control over data as it moves. Most do not realize there is a secondary market for their data. Another issue has to do with the way that data is actually supplied, which makes a potential quality control feature to protect users tricky to implement. Trackers may piggyback on other services: for instance, tracking companies may hide tracking scripts in advertisements. Data is customarily sold as is and the entity that sells it is not liable for the quality of the data. If a person is dependent on this data to get insurance or a mortgage, faulty data is a huge problem. Who is obligated to correct incorrect data? There is a tension between the volume and diversity of data versus the quality of data. The ability to aggregate data from a number of sources may be prioritized over making sure the data are accurate.

Some members of the group also addressed the issues inherent to public data and personal information merging, and the downstream effects of that merger. Data generated by the public sector may benefit democracy, offered some participants. However, there are many problems with public records. Governments collect a lot of data, engage in secondary use, and are at great risk when this data is exposed. Then there are other issues with how government-collected data is disseminated beyond its initial collection point. For instance, people are told that marketers will not have access to public records, yet this is not always the case. Moreover, those in the public sector are often not mandated to consider the downstream use of this data. What are some issues with sharing public records within government, private use of public records, as well as public access to these records? In addition, we need to consider how governments use private data sets. How can we address different needs -- journalistic, academic, commercial, and governmental -- within a data supply chain, be it public or private data? What public information should be accessible for non-commercial use?

There are different constituencies for different data sets. A lot of data, even public data, is in the hands of private entities who pay for public records. Should we be clearer about distinguishing public and private data? Private data is also moving to government sectors. Should there be rules that apply to these kinds of flows between public and private data? Is it even possible to make a distinction between public, government, private, and non-commercial data?

Further Exploration

Through conversation and debate, three main challenges emerged as major points of concern. For one, participants struggled with the balance between visibility and accountability. Visibility allows individuals to see what is being done with their data, but it is also intended to keep platforms accountable. If users are able to clearly see to what ends their information is being used, this may reinforce the responsibility of data brokers and various platforms.

Some participants also debated and discussed at length usefulness of data supply chains as a metaphor. Is it the best possible one to use, or are there other metaphors that would be just as good if not better? Who determines what metaphors work best and for whose purposes are these metaphors then used? A data ecosystem may be a better metaphor than 'data supply chain' since it emphasizes the interrelationships between the different actors. In which case, we may also be able to make use of environmental metaphors for what is happening with the big data phenomenon. Taking this thread further, it may be more useful to think of the data removed out of context as a form of pollution rather than an actor at the far end of a long chain. By expanding the metaphor to a data ecosystem, it may be easier to think about responsibility and accountability in terms of pinpointing relationships that result in harm.

In the discussion, two themes were regarded as elemental to a data supply chain: 1) Information production: what are the labor and workforce issues? What is the work associated with information production? Who does the work and what is the product? 2) Information exchange: How does information get moved around? The discussion addressed the ways in which data supply chains differ from traditional ones. The revolving role of producers and consumers, or prosumers, of data do not cleanly map to the same actors as in physical supply chains. Further, sometimes the data supply chain itself may be the product. Nevertheless, the metaphor offers a great starting point as can be evidenced from the diversity of issues that were raised with respect to data supply chains

Many of the tensions in discussing data supply chains became more evident after some participants worked on visually mapping the flow of a data supply chain with post-it notes. It is tricky to actually track data as it moves between private and public sectors, across platforms, and through data brokers. Different actors may have multiple and shifting roles inside the chain. How do we make sense of these different sets of relationships? Where, for example, does credit card data fit into this schema? Users may generate data, send it out across various platforms and corporate entities, and this data may end up in the hands of data brokers. Public sector data may be used for commercial purposes and public sector entities may likewise integrate corporate data. The role of data brokers was not completely understood, as much of this work tends to be invisible, so some participants flagged this as something in need of further exploration.