Data Supply Chains

by Tamara Kneese, Alex Rosenblat, and danah boyd

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

Brief Description

As data moves between actors and organizations, what emerges is a data supply chain. Unlike manufacturing supply chains, transferred data is often duplicated in the process, challenging the essence of ownership. What does ethical data labor look like? How are the various stakeholders held accountable for being good data guardians? What does clean data transfer look like? What kinds of best practices can business and government put into place? What upstream rights to data providers have over downstream commercialization of their data?

Detailed Topic Description:

As individuals go about their everyday lives, they accumulate an inordinate amount of data. Many transactions, communications, and interactions are traceable, allowing a patterned profile of individual and group activity to emerge. Instances of communication traces are connected to large corporations, such as telecommunications' providers or banks, which provide the scaffolding and infrastructure for those communications to take place. Not only do individuals communicate via text, email, blogs, website commenting systems, and social media platforms, but they also engage in self-tracking behaviors, such as using fitness applications to monitor their daily routines. In addition to these forms of data, telephone companies have records of people's locations, insurance companies know who received what medical service, and financial companies have records of purchasing patterns. Sometimes, the data collected is connected to personally identifiable information - including names, addresses, phone numbers - and sometimes it is linked to less seemingly obvious identifiers, such as IP addresses, zip codes, or gender.

Marketers cobble together these bits of information to create a profile, but even innocuous seeming information like a zip code can be added to a person's birth date and gender in order to pinpoint an individual. According to Latanya Sweeney, a professor at Harvard, up to 87% of Americans are potentially identifiable from their zipcode, birthday, and gender.

Corporate and governmental sources of data are often intertwined. Commercial enterprises sell data to government agencies, just as public records are used to target individuals with marketing campaigns. Palantir, a private security company that collates and mines datasets for various government agencies, is able to cross-reference massive amounts of data from a variety of places. Commercial databases, such as those belonging to Facebook, are often accessible to government agencies, which can then combine this information with surveillance video cameras outside of stores, credit card transactions, emails, flight records, phone calls, online search information, and a host of other data in order to form highly detailed profiles on individuals. In this way, slippage occurs between data collected for commercial purposes, datasets for localized security systems, such as those monitoring fraud or network security, and for national security reasons. As government agencies and commercial entities share data back and forth, how does the context of information change? What sorts of profiles are both commercial companies and government bureaus able to use to identify individuals as well as certain social groups? What problems could arise as a result of this relationship between federal and commercial interests, particularly when it comes to individual privacy?

Furthermore, even supposedly anonymous data has imperfections. Re-identification of subjects is often easy. Other issues include the inability of users to opt out from such tracking methods and discrimination from the framing and use of data. Data anonymization also gives databases lives of their own, disconnecting information from the individuals theoretically represented by it. Anonymization breaks down connections between individuals, communities, and populations and the datasets that are then used to "act" upon them. This disconnection in itself can be a problem and requires rethinking big data governance in a way that can involve these parties regardless of anonymization. New methods, such as differential privacy, have the ability to automate privacy protection in order to protect individuals' privacy in databases, but there are limits to where this can be deployed, and what kinds of data can be queried.

While data is often given to an actor for one purpose or in one context, data is often repurposed by the original actor, or by a third-party actor. Anyone who has signed up for a credit card, for example, knows that a wave of new credit card offers is sure to follow. In another example, new parents can find themselves inundated with adverts for diapers, education information, vaccine updates, and other paraphernalia as a function of public birth records, which are accessed by a variety of organizations for multiple purposes. Online, browser cookies placed by one website are consumed by another website upon visitation. The internet gives the appearance of anonymity, so many consumers are not perturbed by cookies and other online tracking methods, or necessarily aware of them.

Yet, even for savvier users, some slippages aren't nearly as well known. Many brick-and-mortar stores have begun implementing services that scan the room for cell phones, grabbing public information, including phone browser histories and digital cookies, in order to better target shoppers with location-based discount information and other incentives. While clicking on an online advert doesn't directly deliver your information to the

advertiser, data still transfers. For example, Facebook adverts are only shown to you because you fit into pre-selected demographics that the advertiser was targeting; thus, when you click on that advert, the advertiser knows that you're in those categories.

Data slippages can be quite beneficial, enabling you to get discounts at stores and learn about new services. Data aggregators can also address a variety of societal problems - in addition to the marketing challenges that businesses face - by pulling together disparate sources of information. There may be cases in which individual data subjects want to integrate their aggregated data and associated services. How far does an individual get when she wants to create her own data supply chain? A data subject may want to use data from one service, e.g., Fitbit data, 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.

Yet, public discomfort with these practices is also very real, especially when people feel as though they have no control over how information flows. Creepiness is a term that often gets used to express the discontent data subjects have with data supply chains. Specifically, individuals may find it creepy when they notice that their perception of the data supply chain is different from its actual workings. The creepiness factor arises when data traces people leave behind are linked with something else. For example, it may seem creepy when companies can infer that a person is pregnant, has HIV, or has not been taking her medication. Creepiness has a lot to do with the context. It is possible to come up with examples or framings that have social value, such as tracking suicidal individuals with wearables. However, if the framing changes even slightly, people may get uncomfortable as information is presented or used in unexpected ways. How do we deal with the contextual aspects of creepiness? With time the creepiness factor may decrease: months after the revelations, many people may shrug off knowledge about intrusive tracking or government surveillance. Is creepiness factor a reliable "metric" for the existence of ethical, social, cultural and political problems?

The sharing of commercial data is lucrative for companies who profit by selling said data, and also for the companies who buy it and use it for marketing purposes. Public data records are more complex, where the collection of this data is rarely for commercial purposes (e.g., birth records, tax filings, etc.) and, yet, there are very real commercial implications. Likewise, commercial data is often sold to government agencies. While individuals may believe their information is going to their favorite store or perhaps their employer, it is also going to the federal government. Commercial databases may contain faulty information, so what are the implications when this misinformation ends up in the hands of government employees?

When it comes to data about people - public and private records, actively shared and passively collected - who actually "owns" this data? What rights do those data guardians have to transfer, aggregate, interpret the data? When data is transferred, who holds the new guardians accountable? Who is responsible for the data when it is hacked or leaked? If it is

common practice for companies to buy and sell data, regularly sharing and reproducing it, then what protocols are in place to ensure privacy? If this privacy is compromised, as was the case with Target's credit card breach, what are the potential consequences and who is ultimately responsible for fixing the problem? Big databases are also attractive targets of "attacks". How can data be secured? What happens when brokers or guardians spill their data? In light of such risks, is it safer to move towards decentralized architectures and organizations rather than centralized repositories?

Data supply chains can also break down, jeopardizing people's autonomy and privacy. Data guardians may have a difficult time ensuring the quality of the data they store, especially as data is passed back and forth. If a database becomes polluted or compromised, individuals may not know which guardian is responsible. Is the data broker who originally collected the information at fault, or is it a data guardian further down the chain? Who is responsible for ensuring data quality, addressing user complaints and making up for damages? Data also relies on infrastructure and servers, which can go down, as well as on particular corporate entities providing service. As individuals learn to rely on their data, having it become suddenly unavailable can be just as detrimental as having it widely disseminated or repurposed.

Case Study 1: Fitness Tracking

Companies like Fitbit, Nike, and Garmin allow users to track various components of their overall health, including calories burned, weight, sleep quality, and sexual activity. Users are then able to submit this information to websites or applications in order interpret it. Health tracking application users reap great benefit from the information they collect about themselves. Whether an individual is attempting to lose weight or discern potential food allergies, these sorts of applications are useful tools.

This data is reproducible and portable, so advertisers may use this information to target individuals with particular forms of marketing. In addition to being inundated with adverts, users of these programs may also find that their self-tracking is being used to determine their insurance rates or credit scores. While companies like Nike and Fitbit suggest in their privacy policies that they do not currently sell or share users' data, users often consent to allowing other third parties to access this data in order to enhance the activity, not realizing the ways in which it then transfers.

The companies themselves are also more likely to sell information in aggregate so that it cannot be immediately traced back to particular individuals. But, as evidenced by the ability of zip codes, birthdates, and gender to pinpoint individuals, even the most seemingly general information can be linked back to particular users. Fitbit users record intimate details of their lives, including sexual activity, to calculate how many calories they have burned throughout the day. Unbeknownst to them, this personal information is visible in Google search results if users don't change Fitbit's default privacy settings.

Data also transfers when startups sell to bigger companies or when companies go bankrupt. Users may receive legal notices notifying them of pending changes, but most individuals have little understanding of the implications or the ways in which their data flows. While fitness tracking users may want their data to be used by their doctor or, along with other patients' aggregated data, for medical research purposes, they are less likely to want their data to appear in Google search results or to increase their insurance rates. Yet, it is often hard to make sense of the kinds of data transfers that regularly take place here. Further, if individual or aggregate data are incorrect or misleading, individuals may incur damages to their health, access to institutions and medical care. Users may be left with the heavy burden of discovering where things went wrong and finding a responsible party to solve their problems with.

Case Study 2: Public Records

Some kinds of data collected about individuals, including vital records, census statistics, and tax information, are not intended for commercial use. When combined with other information, however, this data can have commercial implications. In one recent case, the father of a teenaged girl killed in an auto accident was horrified to see that an advertisement from OfficeMax included the line "daughter killed in car crash" on the envelope addressed to him. OfficeMax spokespeople blamed a third-party mailing list provider for the faux pas, but the information was most likely originally gleaned from public death records or insurance information. While such records are not intended for use by advertisers, they may be combined with other information to target individuals with specific marketing campaigns. This particular instance exposes what is in fact a common practice. What are the ethical implications for using public records in such a way? Is it better for individuals to be aware of these practices and, if they are made aware, what complications could arise? These are public records, but how should personal and sensitive be integrated into advertising campaigns? For instance, would the relatives of a recently deceased person wish to be targeted with advertisements from the funeral industry? Targeted adverts can backfire when they are disrespectful of a user's sense of propriety or privacy, especially when the advert is nearly a real-time response to a negative situation.

Another place where public records have raised concerns is regarding the birth of a new child. Because birth records are public, many companies flood such parents with baby-related offers. The commercial interest is so intense that companies go to great measures to try to assess the likelihood of a pending birth before the public records are submitted in order to catch new parents before the flood of competing marketing campaigns arrived. Some have predicted birth based on marriage records (or even online marriage announcements). In a now infamous case, Target tried to stay ahead of this curve by using customer data to determine when people were about to have a baby. Cocoa butter, large purses, and calcium pills were some of the items generally purchased by expecting mothers, so Target knew to bombard them with baby centered adverts.

Although some parents are pleased to get useful information about relevant local programs and health material, others are upset by the targeted marketing, particularly when it begins during pregnancy. In order to make their marketing plan less apparent, Target started mixing baby-related adverts in with general adverts so that women were unaware of Target's specific knowledge. In this way, Target was able to utilize their data on customers without always making it visible to its customers. Is this process of obfuscating marketing practices more ethical than blatantly targeting consumers? If customers receive more relevant adverts and offers, is this practice doing any harm and is harm the right criteria for assessing the existence of a wrongdoing? If public records are collected on a mandatory basis, what is a meaningful form of consent for the reuse of this data in other contexts? What happens if public offices start relying on statistical data about citizens? Finally, as data slips between public records and algorithmic inference, who is responsible for the chain of data being used?

Case Study 3: In-Store Tracking

Thanks to recent startups like Nomi and SocialSign, numerous retailers, including many national chains, have implemented sensors to track shoppers' cell phones and in-store movements. Brick-and-mortar stores hope to gain information about their potential customers in order to target them with specific offers and advertisements while also learning how they engage with products in the store. Stores can then target shoppers in specific locations as they shop, alerting them to relevant deals when they approach display cases.

Brick-and-mortar store owners argue that online retailers have an advantage because online shoppers automatically leave digital crumbs, allowing e-commerce websites to learn about potential customers' browsing histories and personal information profiles. As shoppers browse physical stores, however, they may not like the idea of being stalked. Employees at a physical store can link an actual embodied person with their browsing habits, both online and off. Cameras may note how long consumers pause at displays, as well as their relative age and gender presentation, while sensors may determine how many people walk past a store versus how many enter. In addition to these forms of surveillance, some stores have also implemented services that use shoppers' cell phones in order to learn more information about how they shop. RetailNext, for instance, uses its WiFi network to show where a shopper is in the store while also registering each mobile phone as a unique user, allowing retailers to know when someone is a repeat customer. Realeyes targets customers with online adverts based on their facial expressions. Nomi not only tracks in-store customers through their WiFi service, but also matches each phone to an individual. The retailer then gets a holistic view of a shopper, putting together online browsing information with physical information, including the person's shopping history at the store.

It's clear how reproducing the Amazon shopping experience is beneficial to retailers, but it's less evident how these practices impact consumers. If customers know they are being tracked, both through their in-store movements and online habits, they may be perturbed. Aside from the privacy issue or the creepiness factor, however, what can retailers infer about

individuals from this wealth of data? Should information be transferred across contexts without users' knowledge? Can this information be used against shoppers? In one case, Facebook displayed adverts encouraging a young man to come out despite the fact that he had never divulged his sexuality. What happens if such targeted adverts appear on a mobile phone and are then captured by store cameras, allowing employees to discern an individual shopper's sexual orientation? Consumer profiles are extremely valuable, so how can individuals be certain this information won't be sold, stolen, or disseminated? As consumer profiles are increasingly held by third-parties, what kind of influence may these third-parties have on retailers, their surveillance practices, and their product offerings? What happens when the tracking practices we take for granted online are combined with physical surveillance? How does associating a mobile phone's history with an embodied person change the stakes?

Questions to Consider

- What are the major social, cultural, and ethical tensions that emerge because of the flow of data? What needs to be better understood to address these?
- What conflicting values and tradeoffs are at stake? How do we understand relevant actors, stakeholders, and "camps"?
- In what ways does the metaphor of a data supply chain work or not work to capture social, cultural and ethical issues? Can it be used to rethink the shared responsibility of the involved actors? What kinds of environmental, labor, and social issues should be considered?
- What rights, if any, do individuals have to data collected by and about them? What expectations can they have with respect to the collection, processing, distribution, aggregation and deletion of this data?
- How are data supply chains different in different domains (e.g., criminal justice vs. healthcare vs. marketing)? Should we think about different types differently?
- What are salient case studies that highlight the issues surrounding data supply chains? How can we understand the benefits as well as the concerns?
- Who should be responsible for thinking about accountability across a data supply chain, be it for personal data or (anonymized) data aggregates? What is the role of the government? Of data providers? Of tools that allow people to manipulate their own data? Of educational institutions? Of media?
- Beyond legal requirements, are corporations ethically bound to tell consumers when - and exactly how - they are going to share, sell, or otherwise enable the transfer of data? What are the protocols in place to secure data transfers between companies? How is this data protected from being hacked or leaked? In the event of a major privacy breach, who is ultimately responsible and how should companies be punished for failing to keep data safe?
- Who should serve as a data caretaker? What is the role of the government in

supporting, regulating, protecting data caretakers? What kinds of industry self-regulation should be put into place?

- Should there exist limits on how public records data can be used and, vice versa, on how commercially collected data is used by public institutions? Should individuals have the right to opt out of public records data?