Unearthing Big Data Logic

In order to set the scenes, we recall some of the details that emerged from a 2012 class action lawsuit against Google regarding its scanning of Gmail messages (Rosenblatt, 2014a). The legal claim against Google included individuals who were not Gmail account holders. These were individuals who had email accounts with different providers but had nonetheless corresponded with Gmail users. As such, Google could not attempt to rely on their own wide-ranging terms of service agreement as a defense in the claim (Johnston, 2014). The plaintiff's action eventually failed because the presiding judge ruled that the various parties did not constitute a "class" for the purposes of a class action against Google (Rosenblatt, 2014b). However, the case is important because Google's defense to that action provides a significant insight into the corporation's consideration of personal data and its use of analytical methods.

Google argued that mainstream media had long discussed its data collection practices and therefore anyone who used its e-mail service had implicitly consented to having their correspondence scanned, sorted and mined since, allegedly, the public had widespread knowledge of Google's data mining practices (Mendoza, 2013). Google's attempted implied consent defense was rejected by Justice Koh who stated

Accepting Google's theory of implied consent — that by merely sending e-mails to or receiving e-mails from a Gmail user, a non-Gmail user has consented to Google's interception of such e-mails for any purposes — would eviscerate the rule against interception (Cain Miller, 2013).

This rejection is important because contentions that the use of digital platforms, applications, and services like Google should be based upon an implicitly accepted exchange of personal information for service are becoming increasingly commonplace. Thus, for example, some commentators have blithely observed that,

[N]orms are changing, with confidentiality giving way to openness. Participating in YouTube…Flickr, and other elements of modern digital society means giving up some privacy, yet millions of people are willing to make that trade-off every day (McCullagh, 2010).

There are some telling conflations in this formulation of user complicity. First, the fact that a trade-off has been made is equated to its being made knowingly, when in fact many users are unaware of the extent and uses of data collection. The conflation is a useful one, insofar as it highlights the corporate belief that the mere use of a technology or application amounts to implicit consent to its information handling practices. Moreover, the assumption is that accepting the stated terms of use (where available) consequently amounts to informed consent.

We contend that Google's implied consent defense is important to wider and embedded considerations of big data. It could be argued that Google was simply defending itself against a legal claim. However, we attempt to demonstrate that Google's defense reveals a deeper disconnect between how users and data collectors think about personal data. We argue that details of this disconnect are important because they provide insights into the nature of big data – big data as embedded meaning – which in turn provides an understanding of the contested meaning of big data and some of its core constituent elements, namely, personal data, data analytics and informed consent. These issues of contested meaning were borne out in the study's findings.

Contested Disconnects

The interview data indicate a relatively low level of knowledge on the part of even regular technology users about data collection and handling practices. In fact, participant understandings often failed to take into account contemporary data handling practices and strategies. Thus, for example, there was a recurring tendency amongst participants to treat the high volume of data collection as providing a form of anonymity – as if the significance of any particular bit of information is drowned in a data deluge {Solove, 2013 #215, 1899}. That is, participants repeatedly stated that they did not worry about the fate of their personal information because there was so much information out there that it was unlikely anyone would notice or care about their data. The analogy here is to a human information processor: the more data that is collected the harder it would be for any human to look through it and pay attention to the individual details of particular users.

This way of thinking about data – by analogy to human scale data processing – is reinforced by Google's familiar and repeated response to privacy concerns related to Gmail: "no human reads your email." (Webb, 2004) That is, user data is in a sense "safe" because it is added to such a large pool that no actual humans could possibly read it all. The data, like people before them, are lost in the crowd. The typical follow-up to this observation on the part of participants is that even if someone were to stumble across their data, it would be deemed utterly mundane and uninteresting. In an apparently humble vein, the message seemed to be: "I'm not that interesting, so why would anyone pay attention?" This self-effacing logic did double-duty, providing a rationale for explaining why participants were not overly concerned about emerging forms of data collection. In effect, "why should we worry about who is collecting the data if it is boring?"

These attempts to deflect concerns about data collection and tracking, of course, illustrate a fundamental misunderstanding of how automated forms of collection and tracking work. Data does not get "lost in the crowd" thanks to technologies that operate on an extra-human scale. Yes, no individual human could keep track of all the information being generated by millions of people online, but of course, humans are not being asked to do this work.

While it is true that big data firms such as Google may not be interested in particular individuals per se, it is certainly interested in collecting their data in order to

aggregate it with information from other users as a means of more effectively managing and manipulating them – at the individual level. In other words, both of the following claims can be true simultaneously: (a) that Google is not interested in particular individuals but that (b) it will nonetheless collect and store detailed information about users in order to more effectively tailor advertising and other forms of content to them, and indeed to the general population of users. In that sense, lack of interest does not mean lack of impact. Indeed, potentially pathological effects result from the very fact that data miners do not care about particular individuals because these are the very individuals about whom the data is being used to make decisions that may have significant impact on their life chances.

A second important disconnect regards the relevance of particular forms of data to various decision-making processes. Thus, one of the examples of data mining that we used in the structured interviews was based on the finding of a company called Evolv. For certain categories of work, the Web browser used by job applicants to fill out an online application served as an important predictor of future job performance (Anonymous, 2013). It was argued that applicants who used browsers that had to be downloaded and installed (e.g. Firefox Chrome) rather than those that tend to come bundled with a device (e.g. IE or Safari), were statistically more likely to "perform better and change jobs less often." (Anonymous, 2013). That is, a piece of data that was nothing more than an artifact of the online application process turned out to have some direct bearing on this process.

The typical response in our focus group interviews to this finding was two-fold. First, participants attempted to posit an underlying explanation for this correlation even though the data miners do not offer such an explanation – the correlation is enough. Participants when asked how they would feel if their own job applications were data mined, then tended to protest that information about what browser an applicant used was irrelevant to the job search process and that only relevant information should enter into the decision-making process. In a sense, these participants (the vast majority of them) were implicitly critiquing the very premise of data-mining: to unearth unintuitable and unanticipatable correlations.

The participants contended that relevant information was information that could be directly intuited and anticipated to bear on a particular decision-making process. In this case, whether to hire a job applicant or not. But of course, the data unearthed by Evolv certainly was relevant, insofar as it predicted job performance with a relatively high degree of accuracy. So too would any other detail be relevant that demonstrated a robust correlation with superior job performance, no matter how unrelated it might seem (Burdon & Harpur, 2014) (Rosenblat, Kneese, & boyd, 2014). The participants in arguing the need for data relevance attempted to assert a standard of relevance before the fact based on conventional historical understandings of what data would be relevant for that particular decision: Job history; educational qualifications; references, and so on. Data that does not seem to have any meaningful connection to the employment process, other than inexplicably

predicting job performance, was for the participants, irrelevant and should not be considered.

The refusal of participants to reformulate their understanding of relevance in light of data-driven revelations helps explain the third disconnect: the need for informed consent and the speculative character of data mining. When participants were asked what they thought would be a fair policy for governing or regulating the collection and use of personal data, respondents typically said that they would like to be told in advance the purpose for which their data was being collected. Specifically, participants stated that they would like confirmation that data collection would be relevant to the purpose at hand. We have just considered the "relevance" issue: participants understand this as referring to a pre-established category whereas data miners see it as an emergent one (e.g. you cannot know whether any particular piece of data is relevant until you run the correlation). But if relevance is, indeed, emergent, then prior, informed consent becomes structurally impossible because no one can know what use any particular dataset will have until it is run through the data mining process. It is impossible to place any limits on the collection of data for a particular purpose if it cannot be known in advance which data will be potentially useful for that purpose.

These study findings suggest a profound misunderstanding on the part of the public regarding the character of emerging forms of data analytics. The findings and the discussion of disconnects is important because it highlights that understandings of big data are themselves contested. These contests do not simply represent conflicting actions and perspectives of data collectors and data subjects. They are instead representative of a much greater contest that aims to establish societal understanding through constructed meaning. In that sense, attempts to create understandings of big data, particularly by data collectors the size and scale of Google, are also representative of attempts to embed meaning. The argument of societal informed consent is not only indicative of how users do act but it is also symptomatic of how users should act. The user passification inherent in notions of implied consent is fundamental to sensorized development that is predicated on the cyclical and ever-expanding logic of big data. It is therefore important to consider a view of big data as embedded infrastructure in a newly emerging sensor society to better understanding attempted justifications for embedding big data meanings.