Big Data, the Crowd, and Me

Cathy Marshall
Microsoft Research, Silicon Valley

Like many of you, I've been swept up by the romance of Big Data. But instead of stepping back and taking the long view, in this account I plan to get enmired in the details and tell you a story about my own Big Data dalliance. Furthermore, in an effort to explore some fresh territory, I'm not going to take a curatorial perspective (see, for example, Borgman, Wallis, and Mayernik, 2010, or Tibbo et al., 2009), nor address the challenges of data storage and processing (e.g. Yu et al., 2008) or data sharing (e.g. Reichman, Jones, and Schildhauer, 2011). Instead, I'm going to look outside of the pool of light, in among the digital dust bunnies, to see if I can tell a slightly different, more personal story about *what someone like me* (a qualitative researcher) might need to know to be part of a Big Data effort (bearing in mind boyd and Crawford's (2011) thoughtful scholarly reflections on the implications of manipulating and analyzing Big Data)[1]

Big Data is surely the Gold Rush of the Information Age. Researchers across many disciplines have been seduced: currently, mining a carefully-wrought chunk of the Twitter feed can mean a high-visibility (and relatively straightforward) publication in CS or Information Science; an n-gram analysis of Google Books can form the spine of a Digital Humanities study. Every researcher I've talked to who does one of these analyses realizes its limitations, but like me, they have been seduced by Big Data's availability and held in its thrall.[2]

Big Data, Schmig Data

According to *New York Times* technology journalist Steve Lohr, "Big data refers to the rising flood of digital data from many sources, including the Web, biological and industrial sensors, video, e-mail and social network communications." (Lohr, 2012) It's a broad definition to be sure: there are no particular limits as to size, media type, genre, source, or the degree of curatorial attention they receive. It seems that almost anything that's hard to handle in conventional ways (by, say, pulling semi-structured bits into a spreadsheet and going to town with plots and pivot tables) counts as Big Data. Indeed that's the principle distinguishing factor if you consult Wikipedia, which itself counts as Big Data. It seems that all we know about Big Data is that it's too big to look at closely.

So people do sensible things to give themselves a handle on what they have: they use simple visualization tools (how does this look as a scatterplot? How does some aspect of it change over time? How does it look superimposed on a map using lat/lon pairs?); they talk to people *about* the data rather than looking indiscriminately at the data itself; they look at the data's metadata, hoping the description of data (possibly from its source) will tell them what's inside; or perhaps they break off tractable chunks

[1] My account is more personal, and as such, more anecdotal. I'll tie the two together where it makes sense.

[2] Like the California Gold Rush of the late 1840s and early 1850s—or the California real estate gold rush of the 1970s—the allure of Big Data is considerable and it doesn't take much to get started. "However," as one Wikipedia author admonishes, "many returned home [from the Gold Rush] with little more than they had started with." This is not true, of course, of California real estate.

of the data, sampled randomly or deliberately culled, and hope they are not missing data that represents an inconvenient truth.

Big Data, the Crowd, and Me

Like many of my peers, I've been working on an analysis of portions of the Twitter feed. In particular, my colleagues and I are wondering how people identify tweets that are interesting enough to take notice (and perhaps to favorite or save) or important enough to retweet or attend to for more than a fraction of a second. It's seemingly so simple. Millions of people read—skim, interpret, glance at, retweet, respond to—billions of tweets every day.[3] They don't have a real information need (they aren't searching for the diagnosis of disease symptoms or for a book about Alfred Hitchcock's phobias); perhaps they're just looking for social interaction at the digital water cooler or a serendipitous new bit of knowledge or celebrity gossip (Java et al., 2007). And they know what they're looking for when they see it.

Of course, what makes a tweet interesting is a question that is so laughably vague and subjective that you'd think I'd know better[4]. But the subjectivity of the question hasn't stopped any of us from taking a running leap into the haystack of tweets (see, for example, Alonso et al., 2010; Andre et al., 2012; Duan et al., 2010). And the vagueness of the question is part of what makes it intriguing. No information need has been identified. There is no specific context. Just millions of readers and writers creating and consuming billions of tweets. It's just one of many modern information phenomena that have upended the assumptions we've brought to the table.

To do this research, I'm collaborating with an AI researcher (a peer of mine at Microsoft Research, Silicon Valley) and a colleague in the Social Search portion of Microsoft's Bing product group, a senior technical lead who knows his way around crowdsourcing at scale. What do I bring to the problem? In moments of insecurity, I'd say 'not much', but my story has to do with how my experience doing qualitative field research fits into the group, and more generally, what we needed to do and to know to approach this specific instance of Big Data.

Our general method—as preparation to training classifiers that would eventually be used to identify interesting tweets—went something like this:

(1) *Sample the Twitter data*. This meant grabbing a relatively small chunk of the public English-language Twitter feed. This limited sample is used two ways: first it is winnowed further into a set of tweets that's tractable on a scale suitable for human computation. These are the tweets to be judged by an internal crowdsourcing workforce, one that specializes in relevance judgment, to form a labeled training set. The remaining large sample can then act as a test set for the trained classifiers.

[3] According to the official Twitter blog, as of March 2012, 140 million active users produce about 340 million tweets per day. We can assume there are many more readers than writers, that people have multiple accounts, and that people who use Twitter may not represent the population at large.

[4] After all, most well-regarded papers marry important solvable problems with social good.

(2) *Label the tweets*. This involves picking an existing labeling scheme or designing a new one, and developing a way to present the tweet to a worker and collect the label and any other information deemed necessary to assess the label's potential veracity (for example, the worker's level of Twitter experience). This crowdsourced work is monitored, keeping an eye out for fraud and assessing what seems to be steady progress toward completion (Alonso, 2012). If the human computation task isn't moving along, it must be debugged and redesigned.

(3) *Analyze the data*. This means a couple of things: First, ensure data quality by looking at the labels the workers produce. Then decide how many workers need to evaluate each tweet, and what constitutes sufficient consensus (i.e. do 2 out of 3 judges need to agree? 3 out of 5? 4 out of 5? 4 out of 7? As the numbers go up, so does the cost). This initial analysis will determine whether the task was interpreted correctly and that the work is meaningful in addition to being high-quality (in other words, even if the work was done correctly, the results may not be helpful). Statistics may then help identify patterns in the data.

(4) *Add a secondary data source*. Bring a secondary data source into the picture to help interpret the first one. In this case, we had access to query data, since one of us is associated with Bing.[5] This supports the interestingness model we will use to train classifiers.

(5) *Reflect on the results*. In other words, evaluate the results in a way that is convincing to the research community. We can anticipate criticisms because, after all, it would've been more straightforward if there had been a clear-cut information need (e.g. emergency workers who need to locate hurricane victims (Hughes and Palen, 2009); London residents who wanted know about the truth of rumors about the unrest (Lewis, 2011); or perhaps a DJ who wants to spin records suitable to match the apparent moods of millions (Poblete et al., 2011)). As always, reading related work is nerve-wracking after you've finished an initial round of data gathering and analysis—projects inevitably shift subtly as you're working, and a project whose closest relative was far away when you started might be too close for comfort later on. Big Data has the potential to fuel new kinds of inquiry (e.g. the emerging field of Climate Science[6]). It also has the possibility of telling us what we already know.

Putting together this list was straightforward. Accomplishing the 5 items on it wasn't. I'll pull back the curtain and tell my side of the story.[7]

I came to the problem skeptical. In the abstract, I believed that Twitter users could tell you which tweets they found interesting or important in their own feeds. But I wasn't sure that they could articulate the

[5] As boyd and Crawford (2011) point out, this association is significant, and can be an element of what distinguishes who is on each side of the digital divide.

[6] In 2010, at the American Geophysical Union's Fall Meeting, I noticed that an entire half-aisle of the poster session (perhaps 20-30 posters) was devoted to Climate Science. Of course, the program says there were 11,517 posters in all, but at least there were a few of them.

[7] The reader should be cautioned that my side of the story may bear little resemblance to the version(s) told by my two colleagues, even though we've worked together closely for the last six months. To our credit, no-one has blamed anyone for anything.

criteria they used; nor was I certain that they could label tweets that weren't in their own feed. Would they be able to pick out tweets of general interest? Would they have noticed Keith Urbahn's tweet[8] as it scrolled by, especially if they weren't interested in global events?

Thus before we started, we asked the crowd what they looked for as they read their Twitter feed. This seemed to me like asking for trouble; I'd say we should ask a specific question like, "What's the last tweet you retweeted or favorited?" But the answers they gave us seemed reasonable. They didn't seem to be purely aspirational; nor were they telling us what we apparently wanted to hear. They admitted to being on the lookout for celebrity gossip, for humor and inspiration, for photos, for tidbits to which they could attach their own names and turn into memes.

Step 1 was another story. The first sample file, just a few days' worth of tweets, was far too large to open in a text editor and too unwieldy to use the Unix-derived strategy of looking at the first lines of the file. It was even difficult to move the flat file around. Finally we broke off a much smaller chunk *of the sample* so I could take a look at it. Although I'm sure most of the people doing these analyses operate on faith when they start, I can't bear to begin without rolling around in the data.

What can I say about the public English-language Twitter feed? The many papers I'd read over the past few years did not prepare me for what I saw. Nor did my own Twitter feed or the feeds of the people I follow. Not even my occasional searches of the public feed helped me get my head around what I saw when I scanned through the random tweets.

On the upside, the tweets reminded me of the importance of considering how one recruits participants, that the people we know and encounter in our own everyday lives—our samples of convenience—might be very different than most real users.

That the tweets were so incoherent and dopey[9] was both disheartening and exhilarating, even though this was exactly the problem we were anticipating. I've watched colleagues work with scientific datasets, with telemetry, with sensor data. I've interviewed CIA image analysts as they pursued open-ended search tasks, looking through endless imagery to discover something new in the landscape.[10] I've interviewed a researcher to learn how he programmatically sifted through search indices for spam (as we watched text stream by on the screen, he cautioned me that some of it might be 'adult content'[11]). In fact, I've always loved to watch search voyeurs, those displays like the one in Google's main lobby that shows peoples' aggregated queries whiz by.[12]

[8] If you don't recognize his name, you will after I tell you that he's the much retweeted guy who is responsible for, "So I'm told by a reputable person they have killed Osama Bin Laden. Hot damn."

[9] There's no other way to say this. What are we to make of @OscvrBoy's tweet "thirsty bitches are so annoying bruh. ⍰"?

[10] The Cuban Missile Crisis began with just such a revelation—something new was being built. President Kennedy, upon seeing the imagery did not know what he was looking at (perhaps a football field, he speculated), but an image analyst knew what he was seeing.

[11] I've always thought 'adult' was an odd euphemism for porn.

[12] I suspect they filter this display; the query feed always looks remarkably G-rated and upbeat.

But the Twitter feed's sheer banality and size overwhelmed me. What if we gave workers 10,000 ordinary tweets, and not a single one was interesting or important? Yet, if we screened the tweets to start with, wouldn't we be doing exactly what we were trying to avoid (filtering a dataset so we could find exactly what we were looking for)? And certainly, if you filter Big Data (like many current research projects do), you'd want to keep very close track of the relationship between what you have and what you're leaving behind.

Once I had a conversation with a TSA screener as my carry-on luggage crept forward on the conveyer belt through the x-ray tunnel. He was watching a window washer clean fingerprints (gross, greasy, kids' fingerprints) off of a restaurant window; the window was facing onto the concourse where he stood all day. "Looks like fun," he said to me with a sigh. Looks like fun? Cleaning fingerprints off the window? But relative to looking at the ghostly outlines of stuff inside bags, it probably *was* fun. "They have magazines just for window washers," I told him. "One of them is called *American Window Cleaner*." He looked decidedly jealous.

How would he ever notice anything interesting (interesting in a gun or bomb way) in that steady stream of shoes, keys, cell phones, laptops, carry-on luggage, and Ziploc baggies full of toiletries?

But that's Big Data for you.

And that brings me to the second step in our method, crowdsourcing the tweets to obtain what my colleagues were calling a "gold set," a set of tweets labeled through expertise and consensus.

In practice, we discovered that the greatest flaw in the labeling task was the task itself: it was boring; it was fatiguing; and it was frustrating. At first I thought it would be relatively fun and easy. But take a look at three tweets our judges assessed as interesting. In fact, all 5 of them agreed *these tweets are interesting*:

- ***Recent Advances in Ultrasound Diagnosis: 3rd: International Symposium Proceedings (International congress series): http://t.co/9Bqd266l***
- ***Stoned officer calls 911 thinking he's dead... http://t.co/OUBvsEMw***
- ***This is NUTS! Been using this app for Twitter, getting 100s of followers a day! Check it out: http://t.co/QjDsUw6a***

Do these tweets stand out from the others? The judges say they're interesting. Are they perhaps spam? Interesting spam? It's hard to tell. We might say the first one looks legitimate, but the link leads the reader to a volume for sale in Amazon; the symposium proceedings are from a medical meeting that occurred in 1981. The volume looks distinctly unpromising. Yet five judges agreed that this is EXACTLY what we're looking for. Probably the judges are just worn down. After all, at least the words in the tweet are spelled correctly, and the partnership between Amazon and Twitter (the program that is the source of this tweet) is legitimate by some measure. In fact, a surprisingly large number of tweets that the judges labeled for us were exactly of this form, items in Amazon, everything from plastic screws to laptop batteries to bumpers for a 1980s-era Dodge.[13]

The second tweet in the list refers to a goofy animated video made from a recording of a 911 call in 2007. It weighs in on YouTube at under a million views. Does this count as viral? Perhaps. Does it count as humor? At least to its intended audience it does. Is it timely? Probably not, but it's supposed to be funny, not breaking news. Four out of five judges thought it was interesting. According to our survey, Twitter readers are indeed reading their feed with an eye toward being entertained.

The cynical among us might recognize the third tweet as spam; a promise to automatically increase one's followers usually falls into this category. Yet the majority of judges (three out of five) thought it was interesting. And here's what puzzled me the most: Items in Amazon are one thing, but out-and-out spam is another. Should I pretend that this is a good label? This tweet-label pair will become part of what my collaborators are calling the "Gold Set."

And with this, I lost some of my confidence in the crowd's wisdom. Even in aggregate, the crowd seemed misguided, like it was just milling around. "Aw, I haven't seen anything interesting in a while. This one's gotta be interesting," they seemed to be saying.

Although my machine learning collaborator seemed untroubled by the apparent non-wisdom of the crowd, I began to feel some angst. I started to pick apart the judges (was Worker 7475 working a bit too quickly? Did Worker 11101 assign his labels in a pattern?), the judgments (could the judges really pick out an incipient meme?), and the information we were giving them (the endless feed of meaningless tweets). Different elements of the labeling could easily be going haywire:

> *The judges*. Were they working too fast? Perhaps they were missing key semantic aspects of the tweet and being fooled by its form. What kind of speed bumps would keep them from working so quickly?
>
> Or perhaps the judges were becoming fatigued. Maybe they needed to judge more interesting tweets. One surely couldn't look at 10,000 boring and poorly written tweets without losing one's mind.
>
> And there was nothing to say that every judge was familiar with Twitter. What if they weren't? A #FF (follow Friday) wasn't recognized for what it was (a standard Twitter convention), and this made me suspicious that we were relying on an expertise—the ability to quickly scan a Twitter feed—that was not uniformly held by our crowd workforce.
>
> *The tweets*. Were we giving the judges sufficient exogenous information to judge the tweets? We know that a tweet will be judged differently if it flies from Kim Kardashian's keyboard than if it's from a profile called ***@Ishy_Wishy99***. In the first labeling task, we presented the tweets as they appear in most Twitter clients (a profile picture and name, plus the tweet itself). Perhaps it

[13] Here's an example: ***Sony Vaio AR Series Laptop Battery (Replacement): 6-Cell Sony Vaio AR Series 11.1V 4800mAh LiIon Laptop Battery.... http://t.co/RM2fWgae***.

> would help to give the judges the profile's number of followers in addition to its name and photo. After all, these weren't the profiles the judges normally followed.
>
> And what about all that spam? We know that between 1 and 14% of tweets are spam.[14] Would the judges know it when they saw it? Would it dishearten them the way it's disheartening me?
>
> *The labels*. Maybe it was the labels. At times, we gave the judges multiple categories. Then we cut back to an interesting/not interesting judgment. More nuanced labels (without creating such fine categorical distinctions that the task would become unbearably cognitively taxing) might help. What if there was a *ProbablySpam* label? Would that help the judges to recognize spam?

The punchline is, in my efforts to fix the task, I not only spoiled the training data, I potentially alienated the crowd workforce and possibly ticked off my colleagues. I'm still not sure how to fix things, but I'm starting to know what I don't know.

First, I'll confess what I did. As a qualitative researcher I thought, let's find out something about the judges. I began asking the judges to tell us how often they used Twitter. Perhaps we could correlate work quality with Twitter familiarity. Then we upped the number of possible judges from fewer than ten to more than a hundred. Perhaps that would ameliorate the fatigue problem. We asked them to give us rationale for their labeling decisions. Perhaps asking the judges to reflect would improve the quality of their labels. We also expanded the label set—if we added a secondary interest category (*LimitedInterest*), it would allow judges to make a more nuanced distinction; and if we added a spam category (*ProbablySpam*), the judges would realize that some of what they were seeing was spam.

What a mess!

Suddenly the judgment task was cluttered with incomplete responses. Out of 534 responses, only 14 were reasonably complete.[15] This is something no-one says very often about human computation: the humans are, well, HUMAN, and you, the requestor, can violate their trust.[16] You can bore them. You can irritate them. You can frustrate them by asking them to do something unpleasant or impossible. I'm afraid we may have done just that.

And before we did that, we did something else that both embarrasses and puzzles me. We filtered the data. I had thought, what if there were more interesting tweets for the judges to label? Some workers evaluated more than 6000 tweets and found fewer than 150 interesting ones. No wonder they were fatigued.

In the first assessment task, the judges seemed to like tweets with links in them. What if they were tagging a training set in which every tweet had a link? And perhaps we should discard the tweets coming from profiles with fewer than 250 followers. 250 was an arbitrary number; it wasn't informed by what I

[14] Depends on when you look and who you ask.

[15] We iterated five more times, and gradually got better response rates. We may have done irreparable damage to our reputation among the workers however.

[16] Far more emphasis is placed on the judges' competence and their ethics (are they willing to spam?).

know now (in our datasets, spammers sometimes had over 10,000 followers). Furthermore research by Yardi, Romero, Schoenebeck, and boyd (2010) puts the number of followers that a spammer has at an average of 1230 (median 225), while a legitimate user has an average of 536 followers (median 111)[17]. So the number of followers may be a rather poor indicator of the profile holder's intentions. Of course we should discard any tweet whose first character was "@", since it signified a conversation—these were by definition unimportant[18].

And this is how the trouble started. What's more, I suspect this is relatively common practice when it comes to Big Data: it's like sculpting. You keep throwing away stuff that seems like it shouldn't be there, and when you're finished, you have just what you want. I see this when I read my peers' work. They've thrown away data that looks irrelevant (data without the right topical hashtags or data without the desired keywords or data outside the desired geographic region).

There's just so much data that we can all afford to throw quite a bit of it away. *We can throw away data until we find what we're looking for*.

At some point, my machine learning colleague began complaining about datasets we'd been referring to as D2 and D3. The correlations were terrible, he said. And the data was bizarre. All of the tweets had links, and the negative correlation he'd found between what he called "@ mentions" no longer held.

"Oh," I said. "I wonder why THAT happened."

A Research Background for Approaching Big Data

To me, this story illustrates so much about the coming Big Data work, and what a qualitative person needs to know to get along (and what a qualitative person can contribute, if she is careful and doesn't start throwing away data prematurely). As I said at the beginning, I'm not considering the curatorial skills that are mandatory, nor the basic data storage and processing that is necessary to deal with a to-scale dataset. In the first case, I'd hope I'd know the curatorial skills from other parts of my education; in the second case, I'd hope I'd be partnering with a computer scientist (one with data management skills, and one with machine learning chops) to fill in these gaps. What I will focus on instead are the Big Data boundary objects (Star, 2010)[19].

Human computation. At its best, human computation is compelling. Tasks that are ambiguous or difficult can be performed by people instead of computer programs. But programming a human system is nothing like programming a parallel computer. Computers don't get bored, frustrated, nor do they generate inconsistent results. Although much has been made about eliminating spam workers and bad results (Jakobsson, 2009), little emphasis has been put on the requestor, the nature of the tasks the

[17] Albeit again, the universality of these numbers depends crucially on how they filtered the data.

[18] Unless, of course, you're listening in on a conversation between, say, Justin Bieber and Lady Gaga. Here too, I find our assumptions troubling, although they were validated by our initial rounds of labeling assignments.

[19] I'm sure you are worried about my liberal interpretation of the term "boundary objects." I am too. You already know I'm a worrier. But this will not stop me from pressing forward. They're data boundary objects in the sense that these are the points at which data passes between people playing different roles who read the data differently.

requestor designs, and the quality of the tasks (with the notable exception of Alonso, 2012). Learning how to use human computation in its many variations (for example, to do OCR via CAPTCHAs (von Ahn, 2008) or to answer questions about social norms via scenarios (Marshall and Shipman, 2011)) seems important to dealing with the numerous tasks that are required to effectively use Big Data.

In twenty years, human computation will change. Perhaps the workers will organize, unionize: THE UNIVERSAL BROTHERHOOD OF RELEVANCE ASSESSORS 358. Or perhaps they'll be exploited to an even greater degree.[20] But the layer of communications infrastructure between requestor and workers will surely change. Already there are crowd aggregators like Crowdflower, and on the other side there are communications forums for the workers (c.f. turkernation.com and mturkforum). Furthermore, the relationship between requestor and worker has not escaped notice (Silberman, Irani, and Ross, 2010).

Statistics. Here I'm not talking about using Pearson Coefficients. I can look up the formulae or call the functions and plug in the numbers. A good statistics course can show you how to establish the significance of your results. Instead I'm talking about statistics as a meaningful translation between data writ small and data writ large. When there's a spike on a graph, a data scientist needs to be able to know how to ask questions of the data to see what the spike means and whether it represents a meaningful trend or an anomaly in the data.

As an overly simple example, in one dataset we were using, there were a surprising number of tweets that were 99 characters long. String matching showed that they were not identical. We could put the graph in the paper, and note the spike, or we could discover that the spike was the result of a ubiquitous piece of spam, "***GET MORE FOLLOWERS MY BEST FRIENDS? I WILL FOLLOW YOU BACK IF YOU FOLLOW ME - <shortened link>***" and realize that we now have thousands of judgments of that one tweet (albeit with different link shortenings), published by a startling variety of profiles, some with zero followers and computer-generated names (@***fsdfsdf5y5y45h4***) and others with 13K followers and human-sounding names (@***alexjoshthomas***). Thousands of inadvertent judgments of the same tweet are oddly interesting. They can tell you that one judge in the crowd struggled with the tweet's validity (because he or she spent a great deal of time on the judgments of that tweet and sometimes labeled it *TRUE* and other times labeled it *FALSE*). By breaking off a very small chunk of data, we begin to straddle the qualitative and quantitative.

The ability to use statistics to straddle the quantitative and qualitative means that we don't end up with meaningless laws that are neither laws of nature, nor laws of data, but rather accidents of human interaction with technology and statistics.

Data visualization and manipulation. Data visualization has been a topic hailed as promising for almost 20 years. Yet many of the most imaginative visualizations turned out to be unintelligible to the scientists,

[20] I had never really taken a Labor view of human computation. That is, not until a paper I'd written with a colleague was rejected because one reviewer was offended by how we were exploiting the workers. "You're not even paying them minimum wage," the review exhorted. Yet some early research shows that US Mechanical Turk crowd workers participate in these tasks not simply as information piecework, but rather because they find the work somehow entertaining, diverting, or motivating (Ipeirotis, 2010).

analysts, and others the visualizations were supposed to serve.[21] Yet Big Data is well served, especially by the simplest of presentations (time-based or place-based mappings). How do you know what you have? How do you know that the data is okay or that it's what you think it is? How do you discover anomalies in the data and figure out what caused them? Most importantly, how do you establish the relationship between your sample and the rest of the data?

One of the big changes from campaign based data gathering (where a scientist went to the data site and used instruments to collect data) and sensor-based data gathering (where the data is collected and accumulated remotely) is the loss of direct contact with the data, and hence an explanation for bad data values (e.g. bird poop on a sensor or a sensor that only works in partial sunlight).

Furthermore, most data visualization is not interactive in a useful way (i.e. although you can manipulate the presentation, you cannot change the underlying data—for example, to compare different algorithms for cleaning the data).[22] There remains a substantial research agenda, well beyond the beautiful information quilts and information geographies that cause us to ooh and aah and secretly scratch our heads.[23]

Identifying ancillary datasets. One pervasive aspect of Big Data is that no matter how big a dataset is, there are others, and often there's one (or more) that can be brought to bear on question we are asking (provided, of course, that differences in the context of production can be bridged (boyd and Crawford, 2011)). Maybe it's someone else's snowfall records when you're looking at plant respiration and carbon production. Maybe it's Bing social queries when you're looking at a dataset of labeled tweets. The ability to identify ancillary datasets, to interpret them, to know which ones to trust, to understand the ways in which they compromise privacy, and to form partnerships that will give you access to them may seem like an atheoretical skill, but it advances a research agenda in untold ways.

Privacy. When we analyze Big Data from social media—especially when we start to interlock one dataset with another—privacy questions come to the fore. What do we really need to know about privacy? The literature is extensive, so extensive that the last time I looked, I became overwhelmed and decided that anything I could possibly say about privacy (either from the perspective of personal practice, or from the perspective of the data itself) had been said already. And when I make my best effort to read the privacy theory papers, I am overwhelmed by the sophistication of their models. What could these numerous insights into practice or these formal models mean for the information

[21] I realize I'm talking about this topic without being specific. This is deliberate. Some visualizations, e.g. Wordle, are visually appealing, but ultimately a little silly. Others are straightforward, but possibly deceptive. Without deep knowledge about a topic, a corpus, and how to interpret the visualization, Big Data can be viewed deceptively. I also don't want to pick on visualizations that are beautiful, but ultimately unintelligible and meaningless.

[22] I once discarded the Amazon River from a world map databank (large for its time, small now). I surely would not have done so had I been working with a visual representation of the data while I was cleaning it (by algorithmically throwing away line segments with impossible offsets).

[23] We can go all of the way back to the NoteCards browser for examples of this. Users would compute the hypertext graph, print it, and hang it on their walls. When you asked them what it meant, they'd invariably say, "I don't know. But I like the way it looks. It's inspiring!" See, e.g., http://www8.informatik.uni-erlangen.de/IMMD8/Lectures/HYPERMEDIA/Vorlesung/Design/DD/map3/map3gif/notecards1.gif

professional or researcher who is anonymizing a dataset? Surely I have nothing to say here either. Would I even have known enough to run into Abdur Chowdhury's office shouting "DON'T DO IT!"[24]

Yet personally I feel so exposed on one hand (I'm constantly fearful Facebook is going to inform my whole social graph that I read Dlisted.com), and completely baffled by the bizarre twists of other peoples' understanding of privacy on the other. My collaborators and I have interviewed countless people with surprising privacy beliefs, e.g.:

- if you pay for a service, your data is more secure (from the study reported in Marshall and Tang, 2012);
- if someone puts a picture of your kids on the Internet, a child pornographer will do unspeakable things with it (from the study reported in Marshall, Bly, and Brun-Cottan, 2006); and
- a letter you read at a funeral is substantially more private than your finances (again, from the data we gathered for Marshall, Bly, and Brun-Cottan, 2006).

What's more, I have witnessed people inadvertently compromising their own fiercely guarded privacy by giving out the one small fact (e.g. a birthdate) necessary to weave together IMDb and blockshopper, which will yield far more personal information than one would ever tell one's friends (e.g. the details of a long-ago house purchase or personal tax liability). And just when I feel smug about my examples, I come across something like this (in among the Ricola lozenges and Zagat guides):

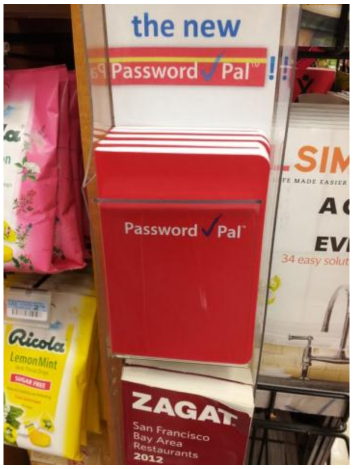

Figure 1. The Password Pal: A blank book speaks volumes

Much to my surprise, writing down passwords on paper is an uncontroversial solution that is endorsed by prominent Windows security experts. [25]

[24] It seems that *Business 2.0* included the release of AOL's search data on a list called "101 Dumbest Moments in Business." (see http://money.cnn.com/galleries/2007/biz2/0701/gallery.101dumbest_2007/57.html) Easy enough for them to say; I'm not so sure most of us would know better.

[25] http://msinfluentials.com/blogs/jesper/archive/2008/02/04/write-down-your-passwords.aspx

Big Data reminds us that privacy problems are far from solved, and that there's an enormous gap between theory and practice.[26] Some of these problems are explored in boyd and Crawford's 2011 Big Data Provocations paper under the rubric of ethics.

Reading the world. If there's anything true of Big Data that isn't true of smaller, more tractable sources, is that you must be able to read the world—the data's world—to understand it (see. Let's go back to the tweet I cited earlier, an example of an entire mystifying genre of tweets:

- ***Recent Advances in Ultrasound Diagnosis: 3rd: International Symposium Proceedings (International congress series): http://t.co/9Bqd266l***

This tweet did not seem to bother anyone but me. The judges thought it might be interesting, and demonstrably labeled it and many of its fellow tweets—all pointing to items for sale in Amazon—as interesting:

- ***Sony Vaio AR Series Laptop Battery (Replacement): 6-Cell Sony Vaio AR Series 11.1V 4800mAh LiIon Laptop Battery.... http://t.co/RM2fWgae***
- ***6 Piece Stacking Rainbow Mug And Stand Set by Collections Etc: 6pc Rainbow Mug Set: Space-saving design! Set of ... http://t.co/qfhS1u10***
- ***Irish Hallowe'en, An: On the Emerald Isle, Halloweíen becomes even trickier, courtesy of three good-for-nothing ... http://t.co/Mwi0MWco***
- ***A/C UV Air Sanitizer 8,000 BTU: A/C UV Air Sanitizer w/Electronic Remote-8,000 BTU http://t.co/0aJfAJ0m***

In fact, a significant proportion of the tweets the judges labeled as interesting are exactly of this form. Is Twitter now a place to run classified ads? Are these squibs spam? Or are they just the result of millions of people acting in accordance with Amazon's Associates program, which gives its members the ability to "Share with Twitter" (aka *Social Advertising*)? I rummaged around my search results (query: *Amazon Twitter*) for quite some time before I found this entry on readwriteweb:

> Last night, Amazon sent out emails to their Amazon Associates members touting the latest addition to the company's affiliate program: a new feature called "Share with Twitter." According to the email, participants can generate "tweetable" links to any Amazon product after first logging into their Associates account. ... After updating Twitter, any person who clicks through on the link and makes a purchase will earn the participant referral fees payable through the Associates program.[27]

Another blog post asked rhetorically if it was spam, hidden advertising, or both. It answered its own question: "It's product placement, Internet-style. Subliminal advertising is rampant on TV (Don Draper in his London Fog coat on *Mad Men*, anyone?), and now it's going to show up in Twitter streams." The blogger ended, however, by saying there's something deceptive about social advertising of this sort.

[26] By practice, I don't just mean personal practice. I'm including database administrators, digital curators, researchers, and everyone else, probably even the people who have published the most about privacy.

[27] http://www.readwriteweb.com/archives/amazon_turns_twitter_into_a_marketplace.php

Without deep-ending on this one example, I'm just trying to say that to read Big Data, you have to read the Big World.[28]

References

Alonso, O. (2012) Implementing Crowdsourcing-based Relevance Experimentation: An Industrial Perspective. Information Retrieval Journal (in press).

Alonso, O., Carson, C., Gerster, D., Ji, X. and Nabar, S. U. (2010) Detecting Uninteresting Content in Text Streams. Proceedings of CSE 2010, ACM Press, pp. 39-42.

Andre, P., Bernstein, M. S., and Luther, K. (2012) Who gives a tweet?: evaluating microblog content value. In Proceedings of CSCW12, pp. 471-474, 2012.

Borgman, C., Wallis, J., and Mayernik, M. (2010) Who's got the data? Interdependencies in Science and Technology Collaborations, *Journal of Computer Supported Collaborative Work*.

boyd, d. and Crawford, K. (2011) Six Provocations for Big Data. Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society", delivered on September 21, 2011. SSRN-id1926431.

Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010) An empirical study on learning to rank tweets. In Proceedings of COLING2010, pp. 295-303.

Hughes, A. L. and Palen, L. (2009) Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6 (3/4), pp. 248-260.

Ipeirotis, P. (2010) Analyzing the Amazon Mechanical Turk Marketplace. *ACM XRDS 17*, 2.

Jakobsson, M. (2009) Experimenting on Mechanical Turk: 5 How Tos. *ITWorld*, September 3.

Java, A., Song, X., Finin, T. & Tseng, B. (2007) Why we twitter: Understanding microblogging usage and communities. *Proceedings of SIGKDD'07*, ACM Press.

Lewis, P. (2011) Reading the riots: Investigating England's Summer of Disorder. *The Guardian*, September 5.

Lohr, S. (2012) New U.S. Research Will Aim at Flood of Digital Data, *New York Times*, 29 March.

Marshall, C.C., Bly, S., and Brun-Cottan, F. (2006) The Long Term Fate of Our Personal Digital Belongings: Toward a Service Model for Personal Archives. *Proceedings of Archiving 2006*. Society for Imaging Science and Technology, Springfield, VA, 2006, pp. 25-30.

Marshall, C.C. and Shipman, F.M. (2011) Social media ownership: using Twitter as a window onto current attitudes and beliefs. *Proceedings of CHI'11*, ACM Press, pp. 1081-1090.

Marshall, C.C. and Tang, J. (2012) That Syncing Feeling: Early User Experiences with the Cloud. Proc. of DIS'12, ACM Press.

[28] It's not even our Big World; it's the data's Big World, the world at the time and in the place that gave rise to the data.

Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. (2011) Do all birds tweet the same? Characterizing twitter around the world. In Proceedings of CIKM’11, pp. 1025–1030.

Reichman,O.J., Jones, M.B., and Schildhauer, M.P. 2011. Challenges and Opportunities of Open Data in Ecology. *Science* 331 (6018), pp. 703-705.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of WWW2010*, pp. 851-860.

Silberman, M.S., Irani, L., and Ross, J. (2010) Ethics and tactics of professional crowdwork. *ACM XRDS*, 17 (2), pp. 39-43.

Star, S.L. (2010) This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, and Human Values*, 35 (5), pp. 601-617.

Tibbo, H., Hank, C., Lee, C.A., Clemens, R. (eds.) (2009) *Proceedings of DigCCurr2009: Digital Curation: Practice, Promise, and Prospects.* University of North Carolina SILS.

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008) reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, vol. 321, pp. 1465-1468.

Yardi, S., Romero, D. M., Schoenebeck, G., and boyd. d. (2010) Detecting spam in a twitter network. *First Monday* 15(1).

Yu, Y., Isard, M., Fetterly, D., Budiu, M., Erlingsson, U., Gunda, P.K., and Currey, J. (2008) DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language. In *Proceedings of OSDI'08*. USENIX, pp. 1-14.