The Domestication of Data: Why Embracing Digital Data Means Embracing Bigger Questions

DAWN NAFUS
Intel Corporation

The EPIC community has been wrestling with ways to integrate quantitative and qualitative methods in light of the increasing role that digital data plays in business practices. Some focus on methodological issues (digital data as method), while others point to the consumer value in data products (data as thing in the world). This paper argues that "digital data as method" and "digital data as thing in the world" are becoming increasingly intertwined. We are not merely witnessing ethnographers' haulting embrace of digital data, but a wider process of the domestication of data, in which we, alongside the people we study, are participants. The domestication of data involves everyday situations in which ordinary people develop their own sense-making methods—methods remarkably similar to ethnographic knowledge production. In this way, the domestication process tightens the connection between data as thing in the world and data as method. I argue that seeing the interconnection gives us the conceptual resources necessary to open up new areas where ethnographers can gain both intellectual and practical footholds in data-rich environments.

DATA AS METHOD/DATA AS THING IN THE WORLD

Sensor data, click data, and other forms of automated time series data are all now well established elements of contemporary digital culture. Because they are largely numerical, they also reinvigorate longstanding debates about the uses of quantitative versus qualitative research, and the nature of mixed methods. Tools change what is thinkable and knowable. New tools renew questions of epistemology, ontology, and methodology. They shape methods discussions and, in turn, are shaped by them. The distribution of those tools also change the situation. Data collecting technologies may once have been exclusively the domain of scientific and social scientific enquiry, but they are no longer. Particularly through the widespread use of sensors, data have now become everyday facts of life. Neither good, nor bad, nor neutral, data serve as an infrastructure of everyday living, the substrate of the most banal business decisions, and forms of evidence for answering questions of many kinds. While professional researchers debate what sensor-generated data now means to them, data's extension into everyday life opens up questions about who gets to be a knowledge producer. As ethnographers find new ways to engage with contemporary digital data, I hope to suggest that the popularization of data matters to our ongoing methodological debates, and to the choices now available about how we participate in a data-rich social world. I will argue that these are more interrelated than it might at first appear, and that seeing the interconnection might open up new areas where ethnographers can gain both intellectual and practical footholds in data-rich environments.

Within the EPIC community, as well as ethnographically-minded academic communities, there are two lines of discussion about ethnography and "big data" that stand out.[1] The first one lays out a set of methodological concerns and approaches, or what I will call "data as method." Patel (2014) and Curran (2014), for instance, remind us of the reasons why

2016 Ethnographic Praxis in Industry Conference Proceedings, p. 384–399, ISSN 1559-8918, https://www.epicpeople.org

qualitative work and quantitative data are not inherently opposed. Wang (2013) and Boyd and Crawford (2011) take an additive approach by noting how ethnography can add richness to big data. It is no secret that, in the private sector, data analytics is often seen as a substitute for ethnographic work , and ethnographers' businesses have been disrupted by cheaper and worse data. Dig closer, and one finds possibility for collaborative, multi-method approaches. There are ample reasons why data scientists, who often lack domain expertise in social behavior, might want ethnographers as collaborators. In turn, ethnographers have begun to forge their own ways into these sorts of datasets, particularly through temporality (Ladner 2013, Nafus 2016). Anthropology and sociology have good approaches for understanding temporality as forms of social and cultural organization, and electronically collected datasets are particularly good at recording various cadences that fall out of ordinary memory. In this way, some new possibilities for ethnographically researching temporality have opened up.

Other work at EPIC approaches data not in terms of method, but as "things in the world." Margolis (2013), for example, sees direct consumer value in data beyond its aggregation and professional analysis. Roberts (2013) addresses ways that consumer-facing data have value but also create asymmetries between device providers and consumers. Material culture and STS approaches to data similarly seek to understand what data is doing in the social world, and how it mediates social lives. They see in data a kind of performativity, enacting rather than merely quantifying in a neutral way. Price data, for example, do not just measure value but also actively participate in the relations between exchangers, and in doing so directly shape the terms on which future exchanges take place (Munesia 2007). Carbon emissions data become political material that shape the world they also describe (Knox 2015). These scholars also see in data a materiality that is inseparable from data's meanings (Day, Lury and Wakeford 2014, Pryke 2010). For example, market traders' price visualizations shape how those traders act in a market. The numbers are not just abstractions but have color and shape and a screen size, all of which are part of what it means to act in a market. Like clay or paint, data are media through which people shape the world around them.

Latour (2002, 2010) laid vital conceptual ground for thinking about how data as "thing in the world" might be connected to methods. He revisited long-forgotten debates between Emile Durkheim and Gabriele Tarde in order to suggest that a deeper transformation is taking place in what quantitative measurement is about. He argues (2010) that the shift from survey-based work to working with digital traces represents a profound change in how we come to comprehend social patterning. Durkheim saw social structuring as something that takes place outside the individual, and inspired entire fields of inquiry to look for what was "higher" than the individual.   Surveys became the instrument of choice for weeding out individual circumstances in order to identify the social structure that lies beyond any particular person. Tarde, on the other hand, believed structure to be nothing more solid than a perpetually emergent flow of discrete interactions between individual persons. In the Tardean view, there is no structure "out there" to be surveyed and known through what later scholars would call God tricks—a view from everywhere and nowhere at the same time, responsible to no human being in particular. There are instead only transactions between people that create paths that in turn potentiate the terms of future transactions. Here society is not cause, but highly provisional consequence, subject to constant renegotiation (Latour 2002: 10).[2]

Tarde lost a series of debates with Durkheim over this in part because he lacked the technology to do it. What would become Durkheim's technologies—surveys—were more or less ready to hand. "Transactional data"—data that is created through the interaction between people and the technologies they make and use—became more widely available much later, through digitization. Indeed, Tarde himself speculated that one day there would be a "gloriometer" that would both measure reputation, and enable people to follow and reflect on those metrics. Transactional data like our modern-day gloriometers (i.e., social media) are *a priori* embedded in a social relation of some kind. Ontologically, they are fundamentally different from survey data, more comparable to the archival objects of historians or visual anthropologists. They are a part of the relationships that ethnography was designed to uncover, and therefore cannot be fairly treated as mere survey data writ large. Instead, transactional data calls into question sociology's "fictive distinction between micro-interactions and macro-structures" (Venturini and Latour 2010:4).

It is perhaps not coincidence that the statistical approaches designed to find patterns in digital traces have more in common epistemologically with ethnography than survey-based sociology, even if the computational methods are harder for us to get our heads around. Bayesian approaches (the everyday workhorses of big data computation) tolerate the absence of a strong hypothesis when survey-wielding Gaussians get twitchy. Bayesians accept 'found' data not optimized for a particular question, which survey researchers often reject as poor research design. There is, of course, a much richer epistemic diversity in data science than I am portraying here. Some machine vision specialists, for example, work from general first principles and build algorithms accordingly, while others take the mess of real-world images as a given, and try to reduce that mess into a pattern that seems workable for the current, contingent situation at hand (Suzanne Thomas, pers. comm.). The differences in ways of knowing are far more complex that whether one is positivist or not. The newfound richness, made possible by a widening diversity of computational methods, means that it is now more possible for ethnographers to find the sorts of kinships and alliances they need to work in a data-rich world than were available even ten years ago.

The differences are also proliferating in disciplines closer to home. Building on the Tardean turn and other intellectual currents, there has been growing interest in social liveliness in sociology (Lury 2012, Marres and Weltevrede 2013, Ruppert, Law and Savage 2013). Here, researchers seek better approaches to the ongoingness of social life. This work asks, if data are things in the world—if they are not mere signifiers or indicators but actually act in the world—are new methods required to maintain the liveliness that sociological evidence is also a part of? For these researchers, data as thing-in-the-world and data-as-method are no longer separate—there is a social life of methods (Law, Ruppert, Savage 2013). Data and devices become possible to work with in new ways as a result.

This work is an important example of how it is possible to ask methodological questions about numerical data without concluding that one must choose between aping the methods of positivist social science, and relegating oneself to the role of friendly addendum—the storyteller that situates the numbers that scale. These scholars are working directly with digital data in ways that further their intellectual commitments, even if in a context that privileges resources for researching all things digital at the expense of other kinds of work.[3] Whether these sociologists can open up a broader conversation about what constitutes quantitative methods, of course, an unanswered political question. Nevertheless, they have created certain "facts on the ground" that *de facto* enrich the methodological diversity now available.

Indeed, if we glance over to the discussions happening in the digital humanities about the use of digital data, we find an analogous set of debates about what it means to be a humanities scholar with new methodological options. There is a good deal of concern about how funding for digital humanities projects is so often predicated on scholars' willingness to render humanities scholarship into a positivistic exercise in computer science. Like our colleagues in the humanities and academic sociology, we also face the *realpolitik* of working conditions. Even if ethnographers do forge ways of working collaboratively with data scientists, we occupy the epistemological minority position. Genuine collaboration that truly respects epistemological difference is a wonderful thing, but it is also hard to come by. Wider social conditions and cultural norms can make meaningful collaboration easier or harder to establish. We could ask, then, what are those wider conditions exactly, and what kind of a say do we have in them?

A SPECIAL CASE OF DATA AS THING

One important part of our overall working conditions is the widespread preference for talking about "data" in very general ways. Much of the technology market (and some data scientists, and some scholarship) holds the assumption that data can be disembedded from its particulars entirely, and stockpiled. While most ethnographers would describe this assumption as a kind of magical thinking, our technical systems make it a social fact: I can download my steps count file as a .csv, and ship it to you, and you can fuse it with whatever you would like to attempt. Data's set-forming ability leads to the urge to build up ever larger datasets, and the urge to minimize the incommensurabilities between different kinds of data--incommensurabilities that become more apparent when one is tasked with having to actually work with it. In this talk, the labor it takes to move data around is largely invisible. It all just becomes "the data," writ large, like a kind of greased pig of truth. Here, "the data" is spoken of as if it were already encoded into one single database, awaiting querying. This reifying discourse is itself a (real) thing in the world. Young tech workers can be spotted in t-shirts proclaiming "data is the new bacon," as if it came vacuum-sealed packets, ready to be sprinkled on everything in sight. When we get a greased pig instead of bacon, it is no wonder that the sight of actually existing data so often seems deflating!

This reified way of talking about data is so widespread as to be unavoidable-- a significant part of how discussions in various disciplines unfold. Pointing out its faults only gets us so far. I suspect Tarde would find it more fruitful to pay greater attention to data's various concrete manifestations, and its everyday life as a medium like clay, paint, or language. The datasets I have in mind—sensor data largely—can be thought about in abstract ways, but when they matter, the where and the how matters most. We can see this when we look at sensor-generated datasets in a spreadsheet and then through a visualization tool, or in an app. We can readily see how some things change, while others stay the same. They have the same begin date and end date, for example, and an indication of what they refer to—steps counted in one minute, say. They can be re-calculated and re-visualized, but they cannot bend to any mathematical or visual will without losing meaning. It is only in its concrete forms where questions about what signifies what to whom can really emerge—questions that ethnographers are quite used to puzzling through.

THE DOMESTICATION OF DATA

There other conditions, beyond the prevalence of fantasy thinking, that are worth our attention. I would argue that in EPIC's current discussion of digital data, we are not merely witnessing ethnographers' halting embrace of working with such data, but a wider process of the domestication of data, in which we, alongside the people we study, are participants. By "domestication of data" I mean to evoke the processes of consumption and adoption that have taken place with other kinds of technologies, like computers and mobile phones. These were designed for narrow types of use, and yet meanings and practices quickly proliferated once people adapted them for the richness of everyday life (Silverstone and Haddon 1996). Few of these were anticipated or anticipatable by their designers. Text messaging famously was an afterthought to the mobile phone, and with use, entire genres of communication have been developed. Early personal computers were going to or business-sify our homes, and instead have become platforms for myriad other activities. In each case, users of new technologies push device capabilities, and a market ecosystem began to pay attention and recalibrate their wares accordingly. When adoption scales, what once was "a device" becomes myriad possibilities created by a combination of consumers, prosumers, artists, open source developers, companies and other institutional actors. Consumption is an active process that changes both the consumer and consumed.

An analogous domestication process is well underway with respect to data. Data is rarely sold to consumers as such, but devices and apps where the consumer is an audience for the data are now commonplace. This is relatively new, as data's long social history has largely been an institutional one. Long before electronic systems, data creation was an important technique of early European state-making, enabling large-scale taxation and conscription (Scott 1998, Desrosières 2002). Similar measurement practices were then adopted by the bourgeoisie in an attempt to legitimate their businesses as activity comparable to scientific practice (Poovey 1998). As more of Western social life became caught up in formal institution-making, and later audit culture (Strathern 2000), the tropes of measuring that were commonplace in institutions became facts of everyday life. Test scores, measures of height and weight, land ownership in so many meters or acres, are all largely taken for granted today as frames for how the world works (not just how institutions work). While quantification as a personal practice goes back at least to Benjamin Franklin, data as something that many people consume is relatively new. And yet here we are. Some obsess over the number of social media followers they have, while others occupy themselves with daily step count, and others still watch closely air quality or miles per gallon on their cars. That is, not only is data a "thing in the world" in the way that prices or test scores are, it is also now a *consumer* thing.

Consumption, of course, is never solely a market activity, even while it is also at key moments squarely a market activity (Slater 2002). Consumers live in much bigger social and cultural worlds than markets, even in the most neoliberal of societies. They bring their own non-market frames to data and devices. Social media, for example, is consumed for individualistic pleasure, and for articulating cultural and class identities of various kinds, but it is also used for the coordination of protest. That is, social media is sometimes a consumer good to be managed inside the moral economy of the household, and sometimes the means of (social, non-market) production. Some adaptations and uses of data by consumers will feed the cycle of design evolution, but many will fall outside the scope of what a for-profit firm can optimize for. In these ways, then, we can recognize that when we ethnographers are working on this or that consumer good, we are shaping a much wider set of social circumstances beyond the particular markets are clients are in, even when those clients are in no position whatsoever to recognize it as such.

The data part of data products, similarly, is sometimes a market commodity and sometimes a means of social reproduction. When sensor data goes beyond consumption, and becomes a means of production, one important thing that is produced is knowledge. Sensor data opens up spaces in which everyday notions of evidence and research unfold. In consumers' hands, sensor data mobilizes everyday notions of health, biology, environmental sciences, and the like. People start asking research-like questions about whether the measurement is true, whether it is relevant, and whether more data is needed, largely because sensor data tends to be semiotically rather vague (Nafus 2014). It tends to offer up partial, messy answers to half-formulated questions.

Many good examples of using data as a means of non-professional knowledge production can be found in the Quantified Self community (QS), a community I study, participate in, and design for. QS members are people who get together in person to discuss what they have learned from the data they collect about themselves through various technologies new and old. QS is an environment where people do not merely ask "how can I take more steps?" but also "why 10,000?" or "Is that appropriate for me?" Often, the key to making meaning from data is outside the dataset itself. Therefore, members of QS often emphasize the importance of context, because that is largely where the value in sensor data lies. One has to know the individual context well enough to know what aspect of the data is or is not relevant. It takes work to puzzle through these matters, and many people put real intellectual effort into making meaning from data (Kragh-Furbo et al 2016).

One example of using activity tracker data "off-script" in order to get meaning from it comes from Jacqueline Wheelwright (2015), who spoke quite movingly at the 2015 QS Conference about how she worked with her activity tracker data to draw important conclusions about her autoimmune disease. Most activity trackers emphasize a daily step count by default, but Wheelwright reworked daily counts of steps into a total aggregation across a month, which better revealed a long-term pattern. She then carefully annotated those monthly time bins with a timeline of flare-ups, which led her to the conclusion that taking 10,000 steps per day was triggering her autoimmune disease. Avoiding 10,000 steps a day was the only way to get her autoimmune disease under control.

This sort of adaptation is what studies of consumption teach us to expect: she took a product's data apart and reassembled it anew, using the data science skills she had to make it speak to her circumstances (an autoimmune disease). We can even speculate on how designers working in this space might respond to ethnographies that explain and document this sort of activity. Some designers might respond by elaborating a "steps for autoimmune tracking" app, while others might reject the finding entirely as a kind of exception, and choose to double down on culturally loaded assumptions about "healthiness" as only ever more exercise. Other designers still might respond by developing a data analysis tool that make it easier to spot patterns in their steps data, which is what my colleagues and I did (Nafus et al 2016). Different designers will respond by emphasizing different valences of the underlying data (Fiore-Silfvast and Neff 20013), thus building up the ecosystem over time.

Similar examples of the intellectual work that people do with data can be found in the emerging class of domestic Internet of Things (IoT) devices. In a study of early adopters of home energy monitors, we found that these users either elaborated their sensing capabilities after an initial foray, by adding additional sensors or infrastructure, or else they abandoned the project altogether upon seeing that total kilowatts consumed did not actually help them measure "energy efficiency," which they came to see as much bigger than kilowatts consumed (Nafus and Beckwith 2016). While we saw less interest in the meaning-making part than in Quantified Self (and more interest in hardware setup), the presence of sensors did prompt people to come to a point of view about what constituted "efficiency" and what an appropriate measurement of it would be, if not the one on offer. As with self-tracking, we can only expect that such views will shape the terms of subsequent adoption of these sorts of devices.

Finally, environmental data is also, albeit more slowly, becoming a matter of personal, scientific, and community-based inquiry. Devices such as the Speck, or Netatmo, monitor different types of air quality. The Speck is designed so as to enable individuals to monitor their own home, and also to contribute to a publicly available GIS of air quality conditions. While there is not yet available research on how people use such devices in their homes, there are precedents in environmental activism where "ordinary" people take samples of the air and have them analyzed in a lab in order to make claims about pollution in their communities. This research shows that people who live in areas strongly affected by pollution can and do develop notions about what the data is telling them (Ottinger 2010). In areas with large oil refineries, for example, Ottinger reports that it is not difficult to find people with fairly well defined beliefs about the inadequacy of year-long or twenty four hour averages as a way of processing pollutant data. Organized through civic groups, their data collection protocols are designed precisely to make the point that averaging over long periods of time is inadequate. While this use of data is not a consumer use *per se*, it supports the broader conceptual point that as data collection tools move out of the lab and into people's hands, those people develop sensibilities about what the data means, and whether it is credible or useful.

In each of these examples we have a situation where expert knowledges are not necessarily the most important factor in drawing conclusions about what data means, even if they are part of the cultural milieu that people draw upon to make meaning. In the self-tracking example, there is no validated protocol developed by clinical research to connect steps aggregated monthly with autoimmune symptoms. In effect, self-trackers are doing their own N of 1 experiments. Indeed, Kragh-Furbo (2016) has found that some users of direct-to-consumer genomics data develop highly elaborate data wrangling skills in order to make sense of direct-to-consumer genomics data in the context of chronic conditions. Similarly, Marres (2009) found that people with green homes often framed their consumption as a kind of experiment in what is possible—i.e., the point was the production of knowledge about ecologically sound houses, and sharing that knowledge with others, not individual optimization *per se*. The publicness of it, as opposed to using specific scientific protocols, was what made it a "real experiment."

It is also notable that where we see these sensibilities develop most quickly, it is in community. Self-trackers or environmental justice organizations are not working alone—they have social organizations that support and facilitate learning and sense-making. Marres's smart home experimenters have an audience, and are not just quietly going about their individual business. Just as Tarde would have it, the data they work with is not merely "out there" recording some abstract social structure or objective phenomenon. Data records the air people breathe and embody, and breathing problematic air demands an account of why, which in turn includes social relations. Recording and seeing those things further entangles people in their social worlds, and becomes the fodder for exchanges about what is knowable and known.

WHO GETS TO KNOW

We can also see in each domain different levels of interest in gate-keeping from expert communities of practice. In the air quality example, citizens' claims about proper data processing are hugely contested by other actors with whom civic groups have conflicts (Ottinger 2010). These contests play out in decidedly "mixed methods" territory. We cannot say that "lay" people only have stories. Now they have numbers, too, and they do not hesitate to use them alongside stories, which also have their own power. The production of "legitimate" numbers—numbers produced through one of many possible scientific vetting processes--sometimes do mobilize action, but when they do, it is not because there is an uncontested notion of universal legitimacy. Which vetting process is used matters, and in a political contest, who pays for the science has a remarkable way of inflecting it. There are times too, when the scientific legitimacy of numbers is utterly besides the point (Ottinger, pers. comm.).

Responses to self-tracking from the medical world have been complex and varied, ranging from ignoring it as irrelevant and non-clinical, to complaining about it as a form of self-diagnosis, to acceptance of self-tracking data as a matter of working with patients. I have also seen the very rare medical professional embracing self-tracking as its own form of medical research. Indeed, the language that one might use to describe this activity is deeply loaded. If I had described Wheelwright's work as "medical self-experimentation," surely it would be immediately thought unwise, dangerous even. Yet she had effectively done a post-hoc A/B test to discover the impact of activity levels on autoimmune disease. In our own response to the language used here, we can see just how powerful medical gatekeepers are.

In the health domain, fights over data access take place through notions of who is properly expert to responsibly use it. Medical device firms often prohibit patients from accessing their own data on the grounds that patients are too inexpert to make sense of it, and thus deliberately thwart the domestication of medical data. In turn, patient groups like #wearenotwaiting mobilize for data access, noting that the are indeed experts in their own bodies, and that matters as much as a knowledge of the underlying biology. Type 1 diabetics, for example, learn the how their bodies respond to various factors, and develop fine-tuned practices of eating, taking insulin, etc., which can take a good deal of time to learn and refine. The #wearenotwaiting group gathers together to work out how to reprogram their insulin pumps to respond to data about glucose levels in the way that they would. However, they also encounter difficulties from device manufacturers that make it difficult to access glucose data in ways that would enable their insulin pumps to use it programmatically, as opposed to having a person look at the glucose data on a screen and reset the insulin pump accordingly. At its core, this fight over access is a fight about who gets to know.

Given that so many different kinds of people do have something to say about the data that refers to them, the "big data divide" (Andrejevic 2014) might not be so big after all. The "big data divide" points to a state of power asymmetries when large, profitable companies that offer little to no transparency to users about how they parse data, exert control over the largest datasets and therefore how that data acts in the world. While these power imbalances are quite real, my examples suggest that we do not have a completely sharp line between the data haves and haves-not. It is also notable that there are so many lead user groups who have found themselves in the position of demanding more transparency and data access because they have *uses* for the data in question. That is, the fights for transparency take place not because some believe that in general more transparency is better, but because particular kinds of people have compelling domestic uses for data and therefore need access. Cheaper data collection tools create the conditions of possibility for non-professionals to begin to formulate their own questions with data, and mount challenges to those gatekeepers who only offer it partially.

THE STAKES FOR EPISTEMIC DIVERSITY

These interactions between "lay" or "citizen-driven" knowledge and expert-driven knowledge are not new. Ottinger (2016) usefully suggests that we should not be fooled into believing that all citizen science projects are citizen-driven. She distinguishes between social movement-based citizen science and scientific authority-based citizen science. The latter uses unpaid volunteers to conduct low-level data collection on behalf of scientists. We see this modality at work in consumer health data products, where medical researchers call for people to donate that data to medical research, but without asking questions about the agenda of research or formulating views on its worth. Indeed, technical systems built to facilitate research on donated data largely exclude citizen scientists who have questions, as opposed to just data to give. This situation is under constant re-negotiation, and some projects do provide results back to data donators, while others consult patient-experts on research design and direction. I have not yet encountered an example of a medical research project that makes granular data available to uncredentialed researches. Even writing those words makes the notion sound absurd: how would a non-expert know about proper data handling, or have the mathematical chops necessary to do it? The instant, visceral response comes from the culturally-ingrained view of the uncredentialed citizen scientist as only ever lacking (Irwin and Wynne 1996): what comes to mind is the figure of the hacker, a bull in the digital china shop, not someone also suffering from the disease who knows all too well what to look for. This view of citizens as people who only ever lack is likely to strengthen when big datasets are at stake, given the near magical powers society invests in those who can wrangle them.

According to Ottinger, social movement-based citizen science is not inherently opposed to scientific methods, but it is much more skeptical of scientists' claims to be able to find universally applicable knowledge. It is more transparent about the values and agendas that motivate data collection, and embedded in concrete action such as public deliberation about what should be a concerning level of a particular pollutant—a discussion that requires moral reasoning as well as evidence. People working in this way are positioned see things that desk-based data science cannot. Popular epidemiology, for instance, has instigated important discussions about what to do when statistically significant concentrations of disease cannot be mathematically produced, because harms are occurring in small communities that the lack large populations that make it possible to eliminate random coincidence as a possible cause in the usual way. The absence of large numbers does not meant the absence of harm being done.

Contributions to "formal" knowledge making from "below" are not rare, even if less visible. Chronic fatigue syndrome did not become a diagnostic category because a medical researcher saw fit to investigate it—it was a disease that patients had to fight to get by summoning evidence of its existence as a disease category (Dumit 2006). Early AIDs research was spurred not merely by a disease constituency lobbying medical experts to invest in more research, but because advocates developed an alternative basis of expertise that enabled them to challenge how science was supposed to be done (Epstein 1996). Activists successfully argued that "elegant science"—cleanly parse-able double-blind clinical trials--was effectively killing people by delaying access to experimental treatment. They did not black box the science, but developed views about appropriate methods, and advocated for them. No small amount of self-experimentation was done in the process. Similarly, air quality problems do not simply become "known"; citizens who suffer the consequences often have do the research themselves in the absence of others taking an interest (Corburn 2005). These contributions from below are not instances of people simply playing nice with experts. They are born out of social contestation, and are made because they matter to the people making them. They happen when ordinary people can see what scientists cannot. This is not to say that scientific authority-driven citizen science are an inherently misguided, but that approach is more likely to be socially productive when values, worldviews, and interests between citizens and professionals are more closely aligned. The ability for the public to scrutinize scientific agendas when those things are not aligned also matters.

It is important to recognize that while "citizen science" might be powerful to our collective knowledge-making, it might not be a frame with which people who domesticate data see themselves. #Wearenotwaiting undeniably does research and development, but not to straightforwardly produce generalizable scientific knowledge. Self-trackers rarely think of themselves as researchers, even though what they find can in fact make a contribution to public knowledge about health, the environment, or other matters. While doing research is an everyday act ("I was researching cars today…"), the presence of data raises the cultural stakes, as traditionally only "researchers" research with data. *De facto*, however, people are doing it all the time.

Far from lacking skill, people who domesticate data have at their disposal the methodological advantage of situated knowledge. The observational powers of a diabetic to understand her own glucose levels are quite high; in a sense she has more "data" than a lab does because she can see what else is going (stress, lack of sleep and so forth). No big data company is ever going to create an inference engine to detect Wheelwright's autoimmune disease fluctuations, or infer that that is the cause (or result) of steps going up or down. It is only she who is in a position to put those two pieces of the puzzle together, and recognize what to look for. The resigned techno-optimist might argue that systems will eventually get better at putting the two together, but this is a big assumption to make. It relies on the assumption that fatigue can be successfully quantified, that symptoms will be consistently tracked, that the technical systems to put the two together are sufficiently integrated, and the person who has visibility into the data across that system also thinks that autoimmune triggers are worthwhile to look for. They might think to look for autoimmune indicators, but they don't know how to look in the way that Wheelwright knows how to look. She will always run faster than large sociotechnical systems, because she has more flexible resources at her disposal.

While these epistemic diversities are not new, they play out on terms newly framed by continuously collected datasets. The 1990s are not the same as 2016. We might speculate that the next major citizen-led change in health or environmental knowledge is highly likely to use data from sensors already or newly deployed. Even though there are major uncertainties in how to coordinate citizen-led work organizationally and technologically, access to the means of knowledge production now clearly involves access to sensor data. That is something that we too have a stake in. The epistemic approaches in social movement-based citizen science, self-tracking, and other everyday acts of domestication should be instantly recognizable to ethnographers as matters near and dear to us. We too insist on the importance of contextual knowledge. Our intellectual starting point is that what one sees is always partial, contingent upon the position from where one sees it. We too use the embodied self as the instrument of knowing (Ortner 1995), and demand radical truthfulness about our values and assumptions rather than erasing them through techniques of bias removal--techniques which remove only biases of a particular kind. Therefore, we are not disinterested observers in the question of who gets to formulate views about what data says. Making room for diverse research methods in a data-rich social world—including the methods of people who may not even think of themselves as researchers—also makes room for ourselves. A world in which data gatekeepers prevent “consumers” or “citizens” from getting access to data is a world that dismisses our ways of knowing as only ever substandard science. If we are to be valued as people who can see what desk data scientists are poorly positioned to see, we need access to the means of knowledge production just as much as citizen scientists do.

WORKING THE METHOD/THING CONNECTION

These cases of data domestication suggest that what we applied ethnographers are dealing with when we deal with data is not just the “analytics” performed by a business, or by another researcher, but the forms of analysis that are ready-to-hand in the wider social world. One matters to the other, and the connection can be fairly direct in a business context. The types of data and analysis that businesses choose to take seriously shapes how they embed data into their products. For example, the urge to apply machine learning to self-tracking systems is not an impulse that is coming from consumers themselves, even if appropriate consumer-facing applications can be found. Finding those appropriate uses relies on how consumers choose to domesticate that data—what people make of what the machines are telling them, and what resources are available to do this. Businesses are keenly interested in what people make of their inferences. They should not be terribly concerned if what their customers do is in fact a kind of phenomenology of data, or a science project with it, but they do need to know, and we are fairly well positioned to tell them.

This back-and-forth between businesses and their customers is precisely the nexus in which many ethnographic practitioners reside. Collectively, we work with many different kinds of companies that provide a wide range of social, cultural and technical affordances that could encourage (or discourage) the domestication of data. We can help tip the balance, either in broadening our client’s understanding of what this or that product should do, or by showing ourselves to be capable of working with client’s datasets. We are also perhaps more likely to be in the room when gatekeepers make it their goal to narrow who can know what, or choose to ignore the customers we interview as only ever "fringe exceptions" who "aren't doing it right." Clearly there are businesses that have a direct economic reason to make strong claims about what data means, and benefit from being hostile to alternatives. Individuals steeped in MBA-type thinking might never be prepared to see situated ways of working with data as anything more than the "bad science" that their customers do. They will be the first to decline us the opportunity to get involved developing strategies for parsing big data based on what that "bad science" shows, and they will lose a competitive advantage because of it. Indeed, we can see this cultural politics already on the wearables market, where there is widespread unwillingness to design data from wearable technologies to do much beyond cajoling people into a shame-driven, biomedicalized notion of health (Nafus 2016). This limitation constrains adoption, yet voicing alternative ways of doing data remains difficult because biomedical normativities are held for reasons that are much deeper than the need for profit-making. Shifting this situation requires more than thoughtful, socially-aware presentations to clients, but it also desperately needs those, too. Whether we are working from a position of relative strength or weakness, many of us will find ourselves in situations where we need to consider what our strategies will be for handling the epistemic and ideological differences that arise when digital data is part of the picture.

I would like to conclude by beginning a conversation about the kinds of things we can do when we do have some say in these matters. The first thing to do might be to acknowledge that the concerns of citizen science are almost never going to be the stated topic of applied research. What we tell businesses about what people do with data, shapes how the public can meaningfully participate in knowledge systems. Therefore, when we work on data products, we inadvertently participate in "citizen science" whether that is the stated goal of the project or not. If knowledge production systems are, realistically, more likely to be an afterthought, then we might focus on encouraging businesses to provide low-cost enablers that do not conflict with business goals, but might otherwise get overlooked. In some cases, certain enablers might support business goals, just in the long-term rather than the short-term.

Here are a few examples of some potential enablers. Many of us are in a position to remind our clients that some of their customers will get much more value out of a product if given good access to the raw data that their product creates, and just how cheap it is to build a data export button. In talking with colleagues in the self-tracking ecosystem, I find that often this vital component to the domestication of data is left on the table not because of any particular business interests, but simply because so few people are aware that it is even conceivable that people would want to look for something in the data outside of the app provided, or outside of other services that might connect to that app. Similarly, companies strive for accuracy in their sensors, but accuracy is never 100%. Giving users the ability to add approximations for times when the sensor was left charging, or delete obviously false data will enable users to have more accurate tallies across weeks, months, or years. One self-tracker I spoke with fantasizes about building the "Forgetbit" that would allow him to do such things with his activity tracker. This would overcome the issue some user have that they feel like exercise "doesn't count" if they leave their activity tracker at home. I suspect that the inability for users to edit sensor data from within apps largely comes from an oversight rather than a strong distrust of people "making up" data.

Companies could also do more to enable groups of users to make their own pools of data. Today, "social" features in wearables initiate competitions between people, or share the fact that the user had taken a particularly long bike ride, which in some social circles amounts to the same thing. But it could be socially productive (and encourage end user continued use) to let patient or other kinds of communities create their own data sub-groups that would allow individual users to situate their data in a more relevant context. A group of people with a particular disease, or exposed to a particular environmental hazard, might want to know whether they are indeed experiencing more sleep disturbances than "typical" activity tracker users (for instance). This means developing systems that allow groups to better self-form in relation to their data, as companies are poorly positioned to pre-identify what "relevant" groupings actually are. Features like these are more costly to develop, but would allow companies to see what their customers thought their data was actually relevant to.

Finally, because consumer-facing data increasingly is processed into "upleveled" recommendations or insights yielded from artificial intelligence systems, ethnographers could very well play key roles in helping companies meet their obligations to customers to transparently explain why their systems made the inference or recommendation that they did. In 2016, the European Union created a legal "right to explanation" of any adverse action taken by artificially intelligent systems. This legislation is a response to data systems that increasingly make guesses about people's likely behavior in discriminatory ways. What a meaningful explanation actually looks like is a significant unknown, and exactly the sort of problem that applied ethnographers are well positioned to tackle. Forward thinking companies might want to tackle it not just out of legal obligation, but also as a way to help customers get more meaning from their products in the first place.

These are just ideas that come out of my own experience in this space. No doubt you have your own to add to the list. My broader point, however, is that it is worthwhile considering the potential "citizen science" value in every data project, regardless of whether the public good is under explicit consideration in the scope of the brief. It is worth doing so not because the idea of participatory knowledge-making feels nicely moral, but because we have a professional interest in making a social world where diverse inquiry remains thinkable. Pointing out the need for a "Forgetbit" feature, or data export, raises the possibility that customers might also be knowledge producers who require certain resources from the business to proceed. Showing that a wearable device user might have their own phenomenological approach to data, potentially brings businesses one step closer to a deeper acknowledgement that there is more than one approach at all.

Data products will not create baby data scientists, but they will and do prompt people to wrestle with data using their own methods. When we focus our attentions on how people get their hands dirty in data, or when we also attempt this ourselves, we can build methods and approaches that do not just "add context" but build context into how data is cleaned, parsed, and represented. That is, our methods can start to occupy a plane commensurable with data science, and more meaningful conversation can begin with those corners of data science willing to try. We cannot deny that more heartbreak in a social world where we yet again constitute exotica might also ensue. However, we do have some conceptual resources that we can mobilize to encourage the acknowledgement of epistemological diversity—an acknowledgement that might come in a form that never uses those exact words, but makes room for us nevertheless.

Dawn Nafus is a Senior Research Scientist at Intel Corporation, where she conducts anthropological research for new product innovation. *dawn.nafus@intel.com*

NOTES

Special thanks to Suzanne Thomas, who significantly helped the ideas in this piece mature, and for the felicitous phrase "greased pig of truth." Thanks also goes to Nimmi Rangaswamy and Simon Pullman-Jones for their sharp editorial contributions.

1. In this work I focus on digital data, or sensor data, rather than "big" data, because what constitutes bigness is highly contested, and in any case the sheer volume is not necessarily what makes data interesting from a context-sensitive point of view. Nevertheless, the data types that I have in mind quite often become part of big datasets.

2 More theoretically minded readers might wonder if this view amounts to transactionalism, or a return to individual psychologism. Latour (2002) is quite adamant that it is not—that the "individual" actor here is already in fact composed of countless previous social interactions, and does not act as a single atomized unit.

3 There is much more going on in this intellectual turn than a crude response to funding priorities. I do not wish to reduce their work to that, but simply note that it cannot be seen outside of that context.

REFERENCES

Andrejevic, M. (2014). "The big data divide." *International Journal of Communication*, 8, 1673-1689.

Boyd, D. and K. Crawford. (2011). Six Provocations for Big Data.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Corburn, Jason. (2005) *Street science: Community knowledge and environmental health justice*. Cambridge, MA: MIT Press.

Curran, J. (2014) "Big Data or Big Ethnographic Data?" Proceedings of EPIC 2014.
https://www.epicpeople.org/big-data-or-big-ethnographic-data-positioning-big-data-within-the-ethnographic-space/

Day, S., Lury, C., & Wakeford, N. (2014). Number ecologies: numbers and numbering practices. Distinktion: Scandinavian Journal of Social Theory, 15(2), 123-154.

Desrosières, A. (2002). *The politics of large numbers: A history of statistical reasoning*. Harvard University Press.

Dumit, J. (2006). Illnesses you have to fight to get: facts as forces in uncertain, emergent illnesses. *Social science & medicine*, 62(3), 577-590.

Epstein, S. (1996). *Impure science: AIDS, activism, and the politics of knowledge*. Berkeley, CA: University of California Press.

Fiore, Silfvast, B., & Neff, G. (2013). "What we talk about when we talk data: Valences and the social performance of multiple metrics in digital health." In *Ethnographic Praxis in Industry Conference Proceedings* (pp. 74-87). https://www.epicpeople.org/what-we-talk-about-when-we-talk-data-valences-and-the-social-performance-of-multiple-metrics-in-digital-health/

Knox, H (2015) 'Thinking Like A Climate'. *Distinktion: Scandinavian Journal of Social Theory*.
http://discovery.ucl.ac.uk/1462480/1/1600910x.2015.1022565.pdf

Kragh-Furbo, M. Mackenzie, A., Mort, M., & Roberts, C. (2016). "Do biosensors biomedicalize?: sites of negotiation in DNA based biosensing data practices". In Nafus, D. (ed). *Quantified: Biosensing Technologies in Everyday Life.* Cambridge, MA: MIT Press.

Ladner, S. (2013). Ethnographic Temporality: Using Time-Based Data in Product Renewal. In *Ethnographic Praxis in Industry Conference Proceedings*. https://www.epicpeople.org/ethnographic-temporality-using-time-based-data-in-product-renewal/

Latour, B. (2002). "Gabriel Tarde and the End of the Social." The social in question: New bearings in history and the social sciences" 117-132.

Latour, B. (2010). "Tarde's Idea of Quantification." In Mattei Candea (ed.) *The Social After Gabriel Tarde: Debates and Assessments*. London: Routledgepp. 145-162.

Law, J., Ruppert, E., & Savage, M. (2011). "The double social life of methods." CRESC Working Paper Series, #95. http://research.gold.ac.uk/7987/1/The Double Social Life of Methods CRESC Working Paper 95.pdf

Lury, C. (2012). "Going live: towards an amphibious sociology." *The Sociological Review*, 60(S1), 184-197.

Margolis, A. (2013). "Five Misconceptions about Personal Data: Why We Need a People-Centered Approach to "Big" Data." *Ethnographic Praxis in Industry Conference Proceedings*. https://www.epicpeople.org/five-misconceptions-about-personal-data-why-we-need-a-people-centred-approach-to-big-data/

Marres, N. (2009). "Testing powers of engagement green living experiments, the ontological turn and the undoability of involvement." *European Journal of social theory*, 12(1), 117-133.

Marres, N., & Weltevrede, E. (2013). "Scraping the social? Issues in live social research." *Journal of Cultural Economy,* 6(3), 313-335.

Muniesa, Fabian. 2007. 'Market Technologies and the Pragmatics of Prices.' *Economy and Society* 36 (3): 377–395.

Nafus, D. (2014). Stuck data, dead data, and disloyal data: the stops and starts in making numbers into social practices. *Distinktion: Scandinavian Journal of Social Theory*, 15(2), 208-222.

Nafus, D. (2016) "Introduction." In D. Nafus (ed.) *Quantified: Biosensing Technologies in Everyday Life.* Cambridge, MA: MIT Press.

Nafus, D. and R. Beckwith (2016). "Number in Craft: Situated Numbering Practices in Do-It-Yourself Sensor Systems." in C. Wilkinson-Weber and A. DeNicola (eds). *Taking Stock: Anthropology, Craft and Artisans in the 21st Century*. London: Bloomsbury Press.

Nafus, D., Denman, P., Durham, L., Florez, O., Nachman, L., Sahay, S., Sharma, S., Savage, E., Strawn, D. & Wouhaybi, R. H. (2016). "As Simple as Possible but No Simpler: Creating Flexibility in Personal Informatics.: In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*(pp. 1445-1452). ACM.

Neff, G. and D. Nafus (2016). *Self-Tracking*. Cambridge, MA: MIT Press.

Ortner, S. (1995). Resistance and the problem of ethnographic refusal. Comparative studies in society and history, 37(01), 173-193.

Ottinger, Forthcoming. "Reconstructing or Reproducing?Scientific Authority and Models of Change in Two Traditions of Citizen Science." In David Tyfield,Rebecca Lave, Samuel Randalls, and Charles Thorpe (eds). *The Routledge Handbook of the Political Economy of Science.* New York: Routledge.

Ottinger, G. (2010). "Constructing empowerment through interpretations of environmental surveillance data." *Surveillance & Society*, 8(2), 221-234.

Patel, N (2014). "Methodological Rebellion: Overcoming the Quantitative-Qualitative Divide" in Denny, R. M. T., & Sunderland, P. L. (Eds.). *Handbook of anthropology in business*. Left Coast Press.

Poovey, M. (1998). *A history of the modern fact: Problems of knowledge in the sciences of wealth and society.* University of Chicago Press.

Pryke, M. (2010). "Money's eyes: the visual preparation of financial markets." *Economy and Society*, 39(4), 427-459.

Roberts, S. (2013) "Big Data, Asymmetries, and Business." Blog post at http://www.ideasbazaar.com/bigdata/.

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society*, 30(4), 22-46.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed.* Yale University Press.

Silverstone, R., & Haddon, L. (1996). "Design and the domestication of ICTs: technical change and everyday life." In Silverstone, R. and Mansell, R (1996) (eds) Communication by Design. The Politics of Information and Communication Technologies,. Oxford: Oxford University Press. 44-74.

Slater, D. (2002). "From calculation to alienation: disentangling economic abstractions.*" Economy and Society,* 31(2), 234-249.

Strathern, M. (2000). *Audit cultures: Anthropological studies in accountability, ethics, and the academy.* Psychology Press.

Venturini, T. and B. Latour (2010). "The social fabric: Digital traces and quali-quantitative methods." *Proceedings of Future En Seine*, pp. 87-101.

Wang, T. (2013). The Conceit of Oracles." EPIC2013 keynote. https://www.epicpeople.org/the-conceit-of-oracles

Wheelwright, J. (2015). "Self-Tracking for Autoimmune Mastery". https://vimeo.com/144678614