Reassembling Our Digital Selves

Deborah Estrin & Ari Juels

Abstract: Digital applications and tools that capture and analyze consumer behaviors are proliferating at a bewildering rate. Analysis of data from large numbers of consumers is transforming advertising, generating new revenue streams for mobile apps, and leading to new discoveries in health care. In this paper, we consider a complementary perspective: the utility of these implicitly generated data streams to the consumer.

Our premise is that people can unlock immense personal value by reassembling their digital traces, or small data, into a coherent and actionable view of well-being, social connections, and productivity. The utility of reassembling the self arises in diverse contexts, from wellness to content-recommendation systems. Without design attention to the unique characteristics of small data, however, the image that these data provide to individual users will be, at best, like a cubist portrait: a fragmented picture of the self.

Management of small data presents fundamental design questions regarding the "who, what, and where" of access rights and responsibilities. The blend of competing and cooperating entities handling small data breaks down distinctions such as that between shared and private, and renders questions like *whose data are they?* hard to answer. Conceptual boundaries blur further as data increase in sensitivity and become "activated," such as when personal apps process and fuse longitudinal data streams to drive context-rich personalization algorithms on the consumer's behalf.

We explore this confusing landscape by drawing attention to three critical design objectives: programmatic access to the digital traces that make up small data, activation of small data for personal applications, and creating privacy and accountability measures for the apps and services consuming small data. We point out the limitations of existing perspectives on both data ownership and control, and on privacy mechanisms, such as sanitization and en-

DEBORAH ESTRIN, a Fellow of the American Academy since 2007, is Professor of Computer Science at Cornell Tech and Professor of Healthcare Policy and Research at Weill Cornell Medical College.

ARI JUELS is Professor at the Jacobs-Technion Cornell Institute at Cornell Tech.

*(*See endnotes for complete contributor biographies.)*

doi:10.1162/DAED_a_00364

cryption. Rather than attempting to provide answers, we pose key questions that should inform new system designs.

The term *big data* expresses the potential of extracting meaning and value using data sets covering large numbers of people, or a large *n*. Big data's humbler counterpart, *small data*, promises to be equally transformative, allowing individual users to harness the data they generate through their own use of online and mobile services; or, in other terms, when $n = 1$.[1]

The explosion of data about individuals is no secret. Personal data sources include: continuous activity and location data sourced from mobile devices and wearables; URL and click data from online searches; text and voice data from social and personal communications (emails and posts, texts and tweets); photographs both taken and viewed; entertainment preferences and consumption; daily purchases made online and offline; personal records from digital education files and medical systems; transportation preferences and patterns; and emerging sources such as genomic data, implanted medical devices, and wearables.

In this paper, we discuss the promise and problems associated with small data. We use the term *small data* to refer to the digital traces produced by an individual in the course of her daily activities, which she can use not to understand general trends across a population, but to understand herself. All of the benefits of small data require a reassembly of the self, a partial to comprehensive drawing together of diverse small data sources pertaining to the individual. While there arise many technical challenges related to data standards, storage, and computation, we focus on the issues in greatest need of architectural attention: *access*, *activation*, *privacy*, and *accountability*.

When should a person have programmatic access to the digital traces he generates, along with the capability to activate these data through applications and services of his choosing? Several examples highlight how small data, as a complement to big data, promises powerful new insights and opportunities:

1) *Custom small-data analytics*. Consider a mobile health application that guides a patient through preparation and recovery from hip surgery. Someday, such an app could analyze her daily walking patterns, provide predictive analytics on her recovery time, and engage her in the physical therapy regimen that best matches her unique medical situation. Such approaches are expanding beyond their origins in the quantified-self movement into broad-based health management practices.[2] The perspective of small data, rather than big data, will provide not only global insights that lead to new therapies, but personalization of these therapies to the patient, time, and place.

2) *Rich user-modeling to facilitate social services*. The quantified self has a natural counterpart in the quantified student. For example, a teacher or tutor could gain great insight from a synthesis of individual students' detailed analytics, as captured by patterns in their online consumption of lectures and readings, or in their online input during homework exercises and examinations. Similar functionality could enrich relationships between mentors and mentees, coaches and clients, and provide crucial support to those whose job it is to safeguard the well-being of teens in the foster system.

3) *Service and product personalization*. Rich user-modeling is equally relevant to service personalization, recommendation systems, and advertising. Popular online platforms like Amazon and Netflix and sharing-economy services like Uber and Airbnb, are largely informed by a very narrow set of data available to them, either directly or through third-party acquisition. Imagine the immersive recommendation systems

that could be built by drawing on users' *full* suites of data, from online retail and service transactions to location and mobile communication data.

This direction could continue to be pursued strictly as a big data play to sell more products and services to targeted customers, such that utility is measured in terms of sales figures. However, we have already seen signs of customer pushback against the perceived "creepiness" of platforms mining personal data to boost sales. If individuals can, instead, demonstrably benefit from personalization on their behalf – in other words, if *utility is instead shown in terms of small data benefiting the individual* – then "getting it right" can advance whole industries beyond contention with consumers.

(4) *Enriching the arc of individual-to-community knowledge*. Individuals share data with communities to accumulate shared knowledge and a collection of experiences. Small data streams, contributed by individual users could, for instance, amplify the great success of manual data entry for sites such as PatientsLikeMe and Inspire, which help patients and caregivers understand and navigate the choices and challenges of specific medical conditions.[3] The small data perspective also points to a path for this collective knowledge to return to the individual in the form of moment-to-moment guidance. Knowledge and predictions about matters from food allergies to triggers of seizures can be mapped continuously onto an individual's small data from a bank of collective experience.

These potential benefits are uncontroversial. But controversy arises and design focus is most needed when we consider an individual's access to her own small data. In order to realize the benefits inherent in the above examples, a consumer needs to have access to her own digital traces, and also needs to be able to activate them, such as by unlocking them in one piece of software and making them available for use in another. It might seem self-evident that this combination of access and activation of one's own data is an imperative, even a universal right. But it is not.

Do you have an irrevocable right to your own physiological data? It is hard to imagine an answer other than yes. But most fitness tracking devices and mobile apps do not give users direct access to raw data on their physiological measures, such as number of steps taken, skin temperature, body weight, speed of food intake, and heart activity. Instead, users must upload the data to a device- or software-maker's service for analysis and display. Users often cannot download or export this raw data because the makers of fitness devices and apps frequently rely on business models that exploit control of their users' data and outline terms of use that claim broad rights to user-generated data.[4]

Tensions around the rights of the individual to her physiological data are not new. But in the past these concerns primarily affected the small segment of the population with implantable medical devices, such as pacemakers and insulin pumps.[5] Now, these issues of physiological data use and ownership impact every user of a mobile phone, smartwatch, or fitness-device.

One complication is the fact that the physiological data recorded by apps are created not just by an individual, but in collaboration with an app. The mobile app that records your footstep is, in fact, a collaboration between your body, which produces the motion, and the accelerometers in your mobile device, which detect it. Their outputs are then translated by the app (or a cloud service) into human-consumable data, like pedometer readings. Moreover, the model that translates the data most likely benefits from data from other users, further complicating the issue!

Joint creation is a widespread feature of small data. If a user interacts with a service provider's content – such as when buying a

video online or posting a comment on the provider's site – ownership and control of the resulting data, transaction, or text can be a complicated matter. Activities like taking group photos, videoconferencing, and gathering nutrition data on shared meals all result in the joint creation of small data.

As the amount, variety, and multiplicity of stakeholders in small data balloon, the questions of rights and control become increasingly complicated.

If people are going to have ready access to sensitive information about themselves, what platforms and methods will support the privacy and accountability needed for apps and services fueled by these data? Small data can furnish powerful insights, for good and ill. So it is essential that the technical community start to develop mechanisms and build products that allow users to access and activate their small data, while protecting them from abuses in digital and commercial ecosystems far too complex for them to reason through, let alone manage.

Today's exploration of privacy violations foreshadows tomorrow's challenges in small data protection. Consider this example: *On July 8 at 11:20 a.m., Olivia hailed a taxi on Varick Street in the West Village in Manhattan. An eleven-minute ride brought her to the Bowery Hotel. She paid $6.50 for the ride. She did not tip.* In the future, small data elements gathered in a narrative like this will be generated by a constellation of devices carried by the user, and by her supporting services. A data-rich payment ecosystem based on NFC (near field communication)-enabled devices will create a record of the payment and harvest ride details automatically. These data will then feed into personal applications such as automated diaries, personal expense reports, and time-management aids.

Though these hypothetical personal applications do not yet exist in the market, this type of small data is generated every minute and has already contributed to documented failures of data protection and control. The taxi ride cited above is real: the actress Olivia Munn traveled from Varick Street to the Bowery Hotel in 2013. In 2014, an enterprising researcher chose to mine public data, and published findings of public interest. The researcher had first noticed that publicly posted photos of celebrities entering and exiting New York City taxis often show legible taxi medallion numbers.[6] Though the government of New York City does make data on individual taxi rides publicly available, it takes care to conceal medallion numbers to protect riders and drivers. Unfortunately, in one large data set, the city implemented this protection ineffectively – through misuse of a hash function – making it possible to associate ride information with specific medallion numbers.

Ridesharing services like Uber are yet another way that these types of data are being generated. Thanks to its use of user-generated location data for pickups, Uber has transformed the use of small data in urban transportation. Like many other shared-service providers, the company is blurring the boundary between customer data, which is used to generate sales, and small data, or personal information, which is used to benefit the user. One could imagine Uber consuming additional user-generated data, such as its users' personal calendars, in order to provide more convenient – and powerful – services. A dark facet of Uber's convenience is the "God view," a (once secret) viewing mode available to Uber employees to track any user. Uber has purportedly used the God view to harass journalists who have written critically about the company.[7] In 2012, Uber infamously published a blog post that tracked what the company called "rides of glory": rides whose timing seemed to indicate passengers had engaged in one-night stands.[8] Given that Uber is generating at least a portion of this personal data, the question arises: should individual users

have the ability to delete personal data stored with such services, or should they learn about how their data are used in order to hold service providers like Uber accountable for abuses?

While these particular privacy violations may not be of great concern to the general public, they illustrate the principle that personal data do not always originate directly with the user. More and more, personal data can be sourced from many different places and can emerge unpredictably. When personal information is turned into small data and made available for individual benefit, it comes burdened with complex provenance; thus, consumers will struggle to control small data they perceive as "theirs." Moreover, as small-data use transforms life-altering, positive realms, such as health care and education, the hazards and conflicting interests involved in data creation could bring additional serious issues to the fore, including data entanglement and data integrity.

There are important limitations to existing designs and models for privacy and control. Several existing approaches to data protection, such as *sanitization*, *cryptography*, and *ownership assignment*, do not address the perspective of small data used by and for the individual. Sanitization is, very broadly speaking, the practice of redacting, aggregating, or adding noise to a collection of data to prepare it for safe release in a privacy-sensitive context. This is the approach that the New York City government took to prevent its taxi-ride data from being used to identify customers; it replaced medallion numbers with cryptographically constructed pseudonyms. As that example shows, data sanitization can be a fragile process. One mistake or unanticipated correlation can lead to unwanted data disclosures.

Another problem with sanitization is the trade-off between privacy and utility. Generally, with an increase in utility comes a decrease in privacy. This tension was strikingly demonstrated by a data set from sixteen MOOCs (massive open online courses) run by MITX and HarvardX on the edX platform.[9] To comply with a federal statute known as the Family Educational Rights and Privacy Act (FERPA), scientists "deidentified" the data set, using a privacy measure called "k-anonymity." Subsequently, these data sets were widely studied by researchers. However, the scientists who produced the data set also discovered that the sanitized data differed in marked ways from the original data set. For instance, in the deidentified data set, the percentage of certified students, or those who successfully completed courses, dropped by nearly one half from the true data set. In this case, protecting privacy could have the drawback of invalidating studies meant to improve instruction quality for students.

In the case of small data, the privacy-utility trade-off is particularly problematic, though not unique to it. There are many big-data analyses, such as medical studies, that can be done more or less safely using sanitized data.[10] Sanitization, however, often does not scale down to the protection of small data: it is not possible to hide an individual's data within a crowd's when the utility of the data stems from its successful integration with other data pertaining to that individual. This problem is illustrated by a study of personalized medicine in which researchers examined estimates of stable dosages for warfarin, an anticoagulant medication, that were made using patients' genetic markers.[11] Researchers demonstrated that in the standard model for such recommendations, a patient's estimated stable dose of warfarin leaks information about his genetic markers. Sanitizing the dose data – in other words, preventing leakage of genetic information by using standard privacy-protecting tools within the model – does not work.[12] The model consumes a tiny amount of infor-

mation (only two genetic markers), and the information is only sourced from one individual. Further, the cost of strong sanitization could be fatal. Degrading the fidelity of the model could result in inaccurately estimated stable warfarin dosages, which could very likely cause patient deaths.

There is little motivation for sanitization when data are consumed by the individual who produced them, as is sometimes the case for small data. But given how many opportunities now exist for sharing small data, it would be natural to appeal to sanitization as a privacy-preserving tool.

Another technical approach to enforcing data confidentiality is the use of cryptography, particularly encryption. Take the example of medical data, also known as protected health information (PHI), which is a particularly sensitive form of small data. The federal Health Insurance Portability and Accountability Act (HIPAA) promotes encryption of such data. Organizations that properly encrypt data and store keys can, in the case of a breach, claim safe harbor status and bypass breach notifications.

When properly deployed today, encryption is very robust: a standard algorithm, such as the Advanced Encryption Standard (AES), cannot be broken even by a powerful adversary. At first glance, properly implemented encryption seems like a cure-all for confidentiality issues.

But encryption, like sanitization, acts at odds with utility. Encrypted data cannot be computed on. (Theoretical and application-specific approaches to computing on encrypted data exist, but have limited utility in practice.) A system must have access to data in order to process it, and thus, if presented with encrypted data, must be able to decrypt it. Further, if a system has access to the encrypted data, then so, too, does an attacker that breaches the system or steals credentials, such as passwords, from a person with access. While encryption is an alluring technical approach to protecting privacy, it is not a magical, cure-all solution.

Given the limitations of technical measures in the protection of privacy, a call has arisen to appeal to economic protections, and perhaps even stimulate open markets for personal data. This approach, which we here refer to as ownership assignment, is exemplified by computer scientist Alex Pentland's "Reality Mining of Mobile Communications: Toward a New Deal on Data," which urges that users should "own their own data."[13] Old English Common Law encapsulates this idea in three general rights for tangible property: users should control the possession, use, and destruction, or dispersion, of their data. The "New Deal on Data" goes a step further: users should also be able to treat data handlers like banks, withdrawing their data if desired and, as with Swiss banks, storing it anonymously.

This deal, which is grounded in a common sense physical model, is enticing. But data are distinctly different from land or money. Data management is far more complicated, and it defies physical transactional models. An acre or a dollar cannot be arbitrarily replicated by anyone who sees it. Nor can it be mathematically transformed into a new object.

Data, on the other hand, are infinitely malleable. They can arise in unexpected places and be combined and transmogrified in an unimaginable number of ways.

To understand the complexities of data ownership, we might ask: who owns the data you generate when you purchase electronic toys from Amazon or food from FreshDirect? Who owns the information produced by your viewings of movies on Netflix, or videos on YouTube? Who owns the data generated by your Android phone, purchased from Cyanogen, and connected to the T-Mobile network, to say nothing of the "physiological" data generated by third-party software on your Fitbit or Apple Watch?

In a previous issue of *Dædalus* on "Protecting the Internet as a Public Commons," legal scholar Helen Nissenbaum articulated relevant alternatives to property rights through her suggestion that we understand privacy as contextual integrity (the idea that privacy is a function of social norms and the environment in which disclosure occurs). She argues that instead of focusing on ownership assignment, we focus on the right to access.[14] Our essay is an argument for that right, and further, for the embodiment of that right in the data and services markets and architectures that we are investing in as leaders of organizations, designers of products, executors of regulations, and consumers of services.

But even with this formulation of property rights, complications arise. Many small data settings invoke *involuntary hazard*, in which the handling of small data by one person can affect the privacy or rights of another without his or her knowledge or involvement. This can occur either from joint data creation or from interactions between individuals on a given platform. Emails, blog posts, and group photos all implicate people captured or referenced in these media with or without their consent, just as a Facebook "gift" creates a record of the sender and the (potentially unwitting) recipient. Many more forms of involuntary hazard will arise as cameras and sensors proliferate, as small data are increasingly aggregated in the cloud, and as analysis and correlation of small data streams yield new insights. Innocent bystanders in photographs, for example, could also be implicated by data sharing.

Kinship gives rise to a particularly striking example in the small data handling of involuntary hazard. Parenthood – the ultimate act of joint creation – creates shared genetic material among kin. This genetic data provide a long-term window into a person's health prospects and behavioral characteristics. Given the sensitivity of such data, the U.S. Federal Genetic Information Nondiscrimination Act (GINA) of 2008 prohibits the use of genetic information by health insurers and employers. As a result of direct-to-consumer genetic testing, however, some people choose to post their genetic data online in repositories, such as OpenSNP, to help catalyze medical discoveries.[15] This choice impacts their kin and descendants, potentially for many decades. As shown in a study by security and privacy researcher Mathias Humbert and colleagues, genetic data enable strong inferences about the predisposition of people related to carriers of Alzheimer's disease toward developing it themselves.[16]

Of all the problems with privacy and accountability mechanisms described here, the most fundamental challenge is, perhaps, psychological in nature. As with health risks associated with exposure to toxins in the air and water, individuals' welfare in terms of privacy is typically degraded more by cumulative exposure than by acute events. Galvanizing people to address gradual threats is a significant and major challenge, without a simple solution.

Given the challenges we have enumerated, key design decisions made today will determine whether we can foster an equitable future society that, while flooded with small data, respects both the value of individuality and personal rights and preferences. Such a society could empower individuals to improve their well-being, social connections, and productivity through intentional use of their small data, while largely avoiding the harmful side-effects of data sharing, such as loss of privacy and vulnerability to predatory businesses. Amid an explosion in the generation, collection, and analysis of small data, however, as well as a resulting erosion of existing models of rights and control, how can we articulate and navigate the decisions needed to realize this vision?

We believe that it is both critical to take a step back from existing models of small-data use, confidentiality, and control, and to frame and reflect on three foundational questions.

What are the practically realizable roles and rights of the individual in the management of small data? Granting ownership and control to individuals over their small data alone will not enable meaningful stewardship. History has shown that many individuals do not have the time or interest to administer fine-grained policies for data access and use. (Facebook privacy settings continue to baffle the majority of users.)[17] It is increasingly impractical for people even to be aware of what data they have produced and where it is stored. As small data become ubiquitous, confidentiality will become increasingly difficult to protect, and leaks may be inevitable. What practical remedies are there?

We suspect that any workable remedy will foremost recognize that individuals' rights should not end with disclosure, and should instead extend to data use. Thus, policies such as HIPAA, which emphasize confidentiality as a means of restricting data flow, will need to be supplemented by rights protections that encompass disclosed data and create fair use and accountability. Consider, again, the example of GINA: if people publish their genetic data, their kin should remain protected.

What fundamental bounds and possibilities exist in data privacy and accountability for small data? There will always be trade-offs between utility and confidentiality. As described above, encryption and sanitization can achieve confidentiality, but often at the expense of the data's usefulness. While emerging cryptographic technologies (such as secure multiparty computation and fully homomorphic encryption) have basic limitations and probably will not alter the landscape for many years to come, they delineate possibilities.[18] They show, for example, that it is possible to mathematically simulate a "trusted third party" that discloses only preagreed-upon results of computation over data without ever revealing the underlying data. Trusted hardware such as the pending Intel SGX technology holds similar potential, but with much more practical, medium-term promise.[19]

Access in such a trusted third-party model can be time-bounded – granted for past data, present data, and/or future data – and limited according to any desired criterion or algorithm. Such a permissions-based model is applicable to both streaming and static data, and is especially useful for "activated" linked-data that is long-lived, streaming, and distributed and used in various ways to drive models and algorithms. This model can create a high degree of accountability by constraining and recording the flow of data.

A simulated trusted third party can offer richer options than releasing sanitized data to researchers. For example, the MOOC data in the HarvardX and MITx study could be made available to researchers not as a sanitized dataset, but as an interface (an API, or application program interface) to a system that manages the raw data.

Nonetheless, public or semipublic release of data will always exist in society; thus, we ought to understand privacy as contextual integrity, a function of social norms and the environment in which disclosure occurs.[20] This concept points toward a future in which semantics automatically govern the flow of data. An intelligent system could discover and determine when and how to process and release data on behalf of consumers, and when and how to fabricate plausible white lies on their behalf (a key social norm). This would be a boon for data-enriched social engagement that would also help restore control to users.

What market demands, government regulations, industry self-regulation (through standardized terms of service), and social norms will shape

the rights of consumers to have access to their small data? To answer this question, we might look at the striking tensions in the commercial handling of health and fitness data today. Recognizing the growing importance of health-related data, Apple has offered HealthKit, an app that serves as a hub for such data drawn from mobile apps. At the same time, the company is treating personal data like a hot potato: Apple does not store or otherwise access its users' health data, leaving this task and liability with app developers. Meanwhile, app developers are ravenously collecting "fitness" data that are likely to function as health data in the future. For example, researchers have shown strong correlations between general health and physical movement throughout the day, and many such apps track physical movement. None of these data are being managed under the aegis of HIPAA.[21]

Users often acquiesce to service providers, such as social networks, health and fitness app developers, and online retailers that take possession of their small data and hold it captive, not sharing it with the users themselves or facilitating its export. Online superpowers like Facebook maintain control over user data and interactions to the extent of being able to influence voter turnout in national U.S. elections.[22] Will our digital selves be reassembled on our behalf by monolithic service providers? Or will an ensemble of entities instead act in concert, according to individual users' tastes and objectives? For more than a decade, mobile phones were controlled by the mobile service provider; since the emergence of smartphones and app stores, control has shifted to the consumer. Which future should we design for? Should ownership of personal data be an inalienable right, rather than one that can be blithely signed away through a terms-of-service agreement?

As Vint Cerf, coinventor of the Internet architecture and basic protocols, has remarked, privacy as we conceive of it today is a historical anomaly, possibly born of the urban revolution.[23] In an age of selfies and social networks, there is every reason to believe that the individual's notions of boundaries and privacy, and of what constitutes personal and public, will continue to shift. Explicit models of the social norms driving policy and practice, and their relationships with market forces and government regulation, must be central to the project of designing the architecture of next-generation small-data systems.

Building methods, tools, and systems with small data in mind is an explicit design choice. By recognizing the role of the individual as beneficiary of her own data-driven applications and services, we are choosing to consider design criteria that are different from those faced by service providers.

If we build systems and market practices that routinely provide people with direct programmatic access to their small data, along with the ability to export and use it, applications and services can offer users the benefit of highly individualized modeling and recommendations that would neither be possible nor acceptable otherwise. And yet, in building such systems, how do we also provide consumers with the safeguards to manage small-data exposure and its consequences in the long term, while still maximizing individual benefit?

We have presented what we believe to be the core challenges raised by small data. We hope that posing these questions is a first step in the direction of secure and beneficial use of small data – by individuals, governments, and enterprises alike.

ENDNOTES

* Contributor Biographies: DEBORAH ESTRIN, a Fellow of the American Academy since 2007, is Professor of Computer Science at Cornell Tech and Professor of Healthcare Policy and Research at Weill Cornell Medical College. She is Founder of the Jacobs Institute Health Tech Hub and Cofounder of the nonprofit Open mHealth. Her recent publications include articles in *Journal of Medical Internet Research*, *Journal of Acquired Immune Deficiency Syndromes*, and *ACM Transactions on Intelligent Systems and Technology*.

ARI JUELS is Professor at the Jacobs-Technion Cornell Institute at Cornell Tech. He has recently published articles in *Journal of Cryptology*, *Communications of the ACM*, and *IEEE Security & Privacy Magazine*.

1 Deborah Estrin, "Small Data, Where *n* = Me," *Communications of the ACM* 47 (4) (2014): 32 – 34.

2 See the collaboration between users and manufacturers of self-tracking tools at http://quantifiedself.com/.

3 See https://www.patientslikeme.com/; and https://corp.inspire.com/patients-caregivers/.

4 For an example of such terms of use, see MyFitnessPal, "Terms of Use," http://www.myfitnesspal.com/account/terms_and_privacy?with_layout=true (accessed January 23, 2015).

5 "Fighting for the Right to Open His Heart Data: Hugo Campos at TEDxCambridge 2011," TEDx Talks, uploaded January 19, 2012, https://www.youtube.com/watch?v=or019-l5M8k.

6 J. K. Trotter, "Public NYC Taxicab Database Lets You See How Celebrities Tip," *Gawker*, October 23, 2014, http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546.

7 In fact, Emil Michael, Uber's senior vice president of business, had suggested to a private audience that the company might dedicate a portion of its financial resources to private researchers to investigate adversarial journalists' personal lives in retaliation for their negative coverage. See Gail Sullivan, "Uber Exec Proposed Publishing Journalists' Personal Secrets to Fight Bad Press," *The Washington Post*, November 18, 2014, http://www.washingtonpost.com/news/morning-mix/wp/2014/11/18/uber-exec-proposed-publishing-journalists-personal-secrets-to-fight-bad-press/; and Chanelle Bessette, "Does Uber Even Deserve Our Trust?" *Forbes*, November 25, 2014, http://www.forbes.com/sites/chanellebessette/2014/11/25/does-uber-even-deserve-our-trust/.

8 Bessette, "Does Uber Even Deserve Our Trust?"

9 For example, Munn and other celebrities whose rides surfaced in this data-mining exercise were criticized for not tipping their taxi drivers. Some alleged, though, that the taxi drivers themselves intentionally failed to record tips. In other words, Ms. Munn's small data may have been corrupted by a "privacy-conscious" (a euphemism for "tax-evading") taxi driver.

10 Jon P. Daries, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Daniel Thomas Seaton, Andrew Dean Ho, and Isaac Chuang, "Privacy, Anonymity, and Big Data in the Social Sciences," *ACM Queue* 12 (7) (2014), http://queue.acm.org/detail.cfm?id=2661641.

11 Benjamin C.M. Fung, Ke Wang, Rui Chen, and Philip S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys* (CSUR) 42 (4) (2010), doi:10.1145/1749603.1749605.

12 Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," *Proceedings of the 23rd USENIX Security Symposium* (Berkeley: USENIX, 2014), https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew.

13 Cynthia Dwork, "Differential Privacy: A Survey of Results," in *Theory and Applications of Models of Computation*, ed. Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Berlin: Springer Berlin Heidelberg, 2008), 1 – 19.

14 Alex Pentland, "Reality Mining of Mobile Communications: Toward a New Deal on Data," in *The Global Technology Report 2008 – 2009: Mobility in a Networked World*, ed. Soumitra Dutta and Irene Mia (Geneva: World Economic Forum, 2009).

15 Helen Nissenbaum, "A Contextual Approach to Privacy Online," *Dædalus* 140 (4) (Fall 2011): 32 – 48.

16 OpenSNP, https://opensnp.org.

17 Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti, "Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy," *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (2013): 1141 – 1152, doi:10.1145/2508859.2516707.

18 Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove, "Analyzing Facebook Privacy Settings: User Expectations vs. Reality," *Proceedings of the ACM SIGCOMM Conference on Internet Measurement* (2011): 61 – 70, doi:10.1145/2068816.2068823.

19 Craig Gentry, *A Fully Homomorphic Encryption Scheme*, Ph.D. dissertation for Stanford University Department of Computer Science (September 2009), https://crypto.stanford.edu/craig/craig-thesis.pdf; and Marten van Dijk and Ari Juels, "On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing," *HotSec 2010 Proceedings of the 5th USENIX Conference on Hot Topics in Security* (Berkeley: USENIX, 2010), 1 – 8.

20 Ittai Anati, Shay Gueron, Simon P. Johnson, and Vincent R. Scarlata, "Innovative Technology for CPU Based Attestation and Sealing," *Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy* (June 2013).

21 Nissenbaum, "A Contextual Approach to Privacy Online."

22 Robert Ross and K. Ashlee McGuire, "Incidental Physical Activity is Positively Associated with Cardiorespiratory Fitness," *Medicine and Science in Sports and Exercise* 43 (11) (2011): 2189 – 2194.

23 Zoe Corbyn, "Facebook Experiment Boosts U.S. Voter Turnout," *Nature News*, September 12, 2012, http://www.nature.com/news/facebook-experiment-boosts-us-voter-turnout-1.11401.

24 Gregory Ferenstein, "Google's Cerf Says 'Privacy May Be An Anomaly.' Historically, He's Right," TechCrunch, November 20, 2013, http://techcrunch.com/2013/11/20/googles-cerf-says-privacy-may-be-an-anomaly-historically-hes-right/