IDENTITY AND PRIVACY

Unique in the shopping mall: On the reidentifiability of credit card metadata

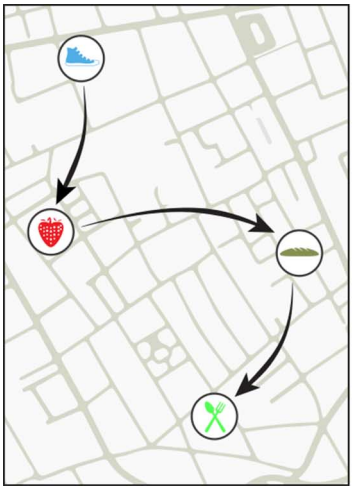
Yves-Alexandre de Montjoye,[1]* Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex "Sandy" Pentland[1]

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that women are more reidentifiable than men in credit card metadata.

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Ubiquitous technologies create personal metadata on a very large scale. Our smartphones, browsers, cars, or credit cards generate information about where we are, whom we call, or how much we spend. Scientists have compared this recent availability of large-scale behavioral data sets to the invention of the microscope (*1*). New fields such as computational social science (*2*–*4*) rely on metadata to address crucial questions such as fighting malaria, studying the spread of information, or monitoring poverty (*5*–*7*). The same metadata data sets are also used by organizations and governments. For example, Netflix uses viewing patterns to recommend movies, whereas Google uses location data to provide real-time traffic information, allowing drivers to reduce fuel consumption and time spent traveling (*8*).

The transformational potential of metadata data sets is, however, conditional on their wide availability. In science, it is essential for the data to be available and shareable. Sharing data allows

[1]Media Lab, Massachusetts Institute of Technology (MIT), 20 Amherst Street, Cambridge, MA 02139, USA. [2]Department of Computer Science, Aarhus University, Aabogade 34, Aarhus, 8200, Denmark. [3]School of Communication and Information, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901, USA.
*Corresponding author. E-mail: yvesalexandre@demontjoye.com



shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5
	4c7af72a	09/23	$12.29	$5 – $16
	89c0829c	09/24	$3.66	$2 – $5
	7abc1a23	09/24	$35.81	$16 – $49

Fig. 1. Financial traces in a simply anonymized data set such as the one we use for this work. Arrows represent the temporal sequence of transactions for user 7abc1a23 and the prices are grouped in bins of increasing size (*29*).

scientists to build on previous work, replicate results, or propose alternative hypotheses and models. Several publishers and funding agencies now require experimental data to be publicly available (*9*–*11*). Governments and businesses are similarly realizing the benefits of open data (*12*). For example, Boston's transportation authority makes the real-time position of all public rail vehicles available through a public interface (*13*), whereas Orange Group and its subsidiaries make large samples of mobile phone data from Côte d'Ivoire and Senegal available to selected researchers through their Data for Development challenges (*14*, *15*).

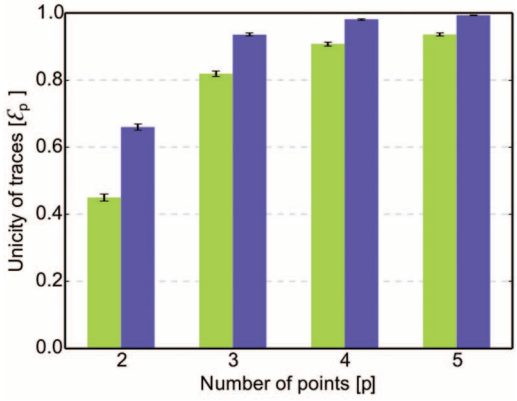
These metadata are generated by our use of technology and, hence, may reveal a lot about an individual (*16*, *17*). Making these data sets broadly available, therefore, requires solid quantitative guarantees on the risk of reidentification. A data set's lack of names, home addresses, phone numbers, or other obvious identifiers [such as required, for instance, under the U.S. personally identifiable information (PII) "specific-types" approach (*18*)], does not make it anonymous nor safe to release to the public and to third parties. The privacy of such simply anonymized data sets has been compromised before (*19*–*22*).

Unicity quantifies the intrinsic reidentification risk of a data set (*19*). It was recently used to show that individuals in a simply anonymized mobile phone data set are reidentifiable from only four pieces of outside information. Outside information could be a tweet that positions a user at an approximate time for a mobility data set or a publicly available movie review for the Netflix data set (*20*). Unicity quantifies how much outside information one would need, on average, to reidentify a specific and known user in a simply anonymized data set. The higher a data set's unicity is, the more reidentifiable it is. It consequently also quantifies the ease with which a simply anonymized data set could be merged with another.

Financial data that include noncash and digital payments contain rich metadata on individuals' behavior. About 60% of payments in the United States are made using credit cards (*23*), and mobile payments are estimated to soon top $1 billion in the United States (*24*). A recent survey shows that financial and credit card data sets are considered the most sensitive personal data worldwide (*25*). Among Americans, 87% consider credit card data as moderately or extremely private, whereas only 68% consider health and genetic information private, and 62% consider location data private. At the same time, financial data sets have been used extensively for credit scoring (*26*), fraud detection (*27*), and understanding the predictability of shopping patterns (*28*). Financial metadata have great potential, but they are also personal and highly sensitive. There are obvious benefits to having metadata data sets broadly available, but this first requires a solid understanding of their privacy.

To provide a quantitative assessment of the likelihood of identification from financial data, we used a data set D of 3 months of credit card transactions for 1.1 million users in 10,000 shops in an Organisation for Economic Co-operation and Development country (Fig. 1). The data set was simply anonymized, which means that it did not contain any names, account numbers, or obvious identifiers. Each transaction was time-stamped with a resolution of 1 day and associated with one shop. Shops are distributed throughout the country, and the number of shops in a district scales with population density ($r^2 = 0.51$, $P < 0.001$) (fig. S1).

Fig. 2. The unicity ε of the credit card data set given *p* points. The green bars represent unicity when spatiotemporal tuples are known. This shows that four spatiotemporal points taken at random ($p = 4$) are enough to uniquely characterize 90% of individuals. The blue bars represent unicity when using spatial-temporal-price triples ($a = 0.50$) and show that adding the approximate price of a transaction significantly increases the likelihood of reidentification. Error bars denote the 95% confidence interval on the mean.



We quantified the risk of reidentification of D by means of unicity ε (*19*). Unicity is the risk of reidentification knowing p pieces of outside information about a user (*29*). We evaluate ε_p of D as the percentage of its users who are reidentified with p randomly selected points from their financial trace. For each user, we extracted the subset $S(I_p)$ of traces that match the p known points (I_p). A user was considered reidentified in this correlation attack if $|S(I_p)| = 1$.

For example, let's say that we are searching for Scott in a simply anonymized credit card data set (Fig. 1). We know two points about Scott: he went to the bakery on 23 September and to the restaurant on 24 September. Searching through the data set reveals that there is one and only one person in the entire data set who went to these two places on these two days. $|S(I_p)|$ is thus equal to 1, Scott is reidentified, and we now know all of his other transactions, such as the fact that he went shopping for shoes and groceries on 23 September, and how much he spent.

Figure 2 shows that the unicity of financial traces is high ($\varepsilon_4 > 0.9$, green bars). This means that knowing four random spatiotemporal points or tuples is enough to uniquely reidentify 90% of the individuals and to uncover all of their records. Simply anonymized large-scale financial metadata can be easily reidentified via spatiotemporal information.

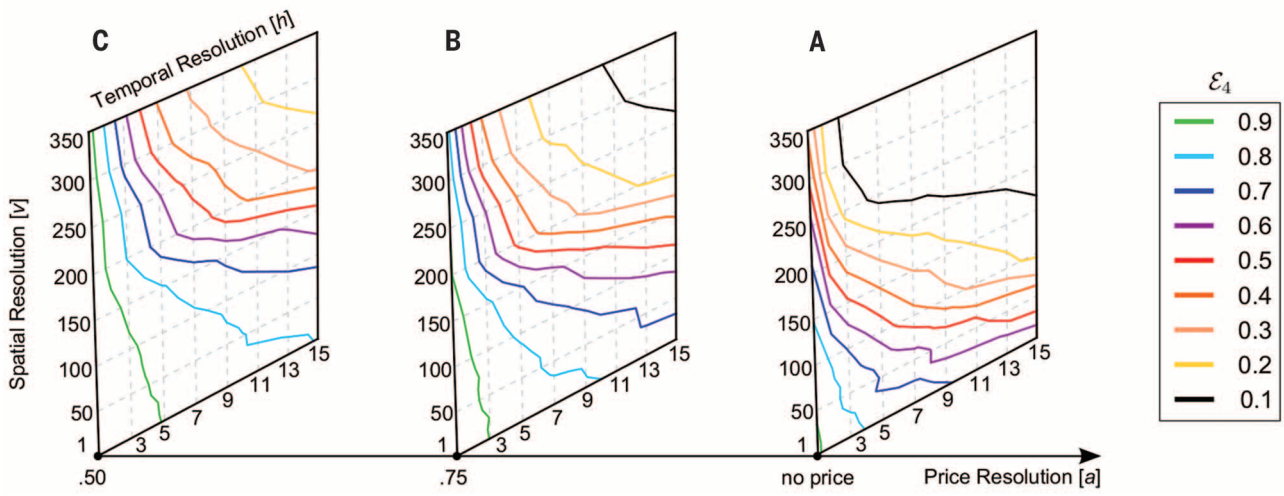


Fig. 3. Unicity (ε_4) when we lower the resolution of the data set on any or all of the three dimensions; with four spatiotemporal tuples [(A), no price] and with four spatiotemporal-price triples [(B), *a* = 0.75; (C), *a* = 0.50]. Although unicity decreases with the resolution of the data, the decrease is easily overcome by collecting a few more points. Even at very low resolution (*h* = 15 days, *v* = 350 shops, price *a* = 0.50), we have more than an 80% chance of reidentifying an individual with 10 points ($\varepsilon_{10} > 0.8$) (table S1).

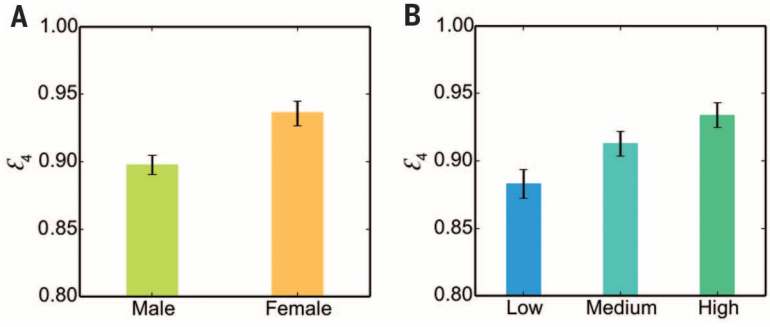


Fig. 4. Unicity for different categories of users (*v* = 1, *h* = 1). (**A**) It is significantly easier to reidentify women (ε_4 = 0.93) than men (ε_4 = 0.89). (**B**) The higher a person's income is, the easier he or she is to reidentify. High-income people (ε_4 = 0.93) are significantly easier to reidentify than medium-income people (ε_4 = 0.91), and medium-income people are themselves significantly easier to reidentify than low-income people (ε_4 = 0.88). Significance levels were tested with a one-tailed *t* test (*P* < 0.05). Error bars denote the 95% confidence interval on the mean.

Furthermore, financial traces contain one additional column that can be used to reidentify an individual: the price of a transaction. A piece of outside information, a spatiotemporal tuple can become a triple: space, time, and the approximate price of the transaction. The data set contains the exact price of each transaction, but we assume that we only observe an approximation of this price with a precision *a* we call price resolution. Prices are approximated by bins whose size is increasing; that is, the size of a bin containing low prices is smaller than the size of a bin containing high prices. The size of a bin is a function of the price resolution *a* and of the median price *m* of the bin (*29*). Although knowing the location of my local coffee shop and the approximate time I was there this morning helps to reidentify me, Fig. 2 (blue bars) shows that also knowing the approximate price of my coffee significantly increases the chances of reidentifying me. In fact, adding the approximate price of the transaction increases, on average, the unicity of the data set by 22% (fig. S2, when *a* =0.50, $\langle\Delta\varepsilon\rangle$ = 0.22).

The unicity ε of the data set naturally decreases with its resolution. Coarsening the data along any or all of the three dimensions makes reidentification harder. We artificially lower the spatial resolution of our data by aggregating shops in clusters of increasing size *v* based on their spatial proximity (*29*). This means that we do not know the exact shop in which the transaction happened, but only that it happened in this geographical area. We also artificially lower the temporal resolution of the data by increasing the time window *h* of a transaction from 1 day to up to 15 days. Finally, we increase the size of the bins for price *a* from 50 to 75%. In practice, this means that the bin in which a $15.13 transaction falls into will go from $5 to $16 (*a* = 0.50) to $5 to $34 (*a* = 0.75) (table S2).

Figure 3 shows that coarsening the data is not enough to protect the privacy of individuals in financial metadata data sets. Although unicity decreases with the resolution of the data, it only decreases slowly along the spatial (*v*), temporal (*h*), and price (*a*) axes. Furthermore, this decrease is easily overcome by collecting a few more points (table S1). For instance, at a very low resolution of *h* = 15 days, *v* = 350 shops, and an approximate price *a* = 0.50, we have less than a 15% chance of reidentifying an individual knowing four points ($\varepsilon_4 < 0.15$). However, if we know 10 points, we now have more than an 80% chance of reidentifying this person ($\varepsilon_{10} > 0.8$). This means that even noisy and/or coarse financial data sets along all of the dimensions provide little anonymity.

We also studied the effects of gender and income on the likelihood of reidentification. Figure 4A shows that women are easier to reidentify than men, whereas Fig. 4B shows that the higher somebody's income is, the easier it is to reidentify him or her. In fact, in a generalized linear model (GLM), the odds of women being reidentified are 1.214 times greater than for men. Similarly, the odds of high-income people (and, respectively, medium-income people) to be reidentified are 1.746 times (and 1.172 times) greater than for low-income people (*29*). Although a full causal analysis or investigation of the determinants of reidentification of individuals is beyond the scope of this paper, we investigate a couple of variables through which gender or income could influence unicity. A linear discriminant analysis shows that the entropy of shops, how one shares his or her time between the shops he or she visits, is the most discriminative factor for both gender and income (*29*).

Our estimation of unicity picks the points at random from an individual's financial trace. These points thus follow the financial trace's nonuniform distributions (Fig. 5A and fig. S3A). We are thus more likely to pick a point where most of the points are concentrated, which makes them less useful on average. However, even in this case, seven points were enough to reidentify all of the traces considered (fig. S4). More sophisticated reidentification strategies could collect points that would maximize the decrease in unicity.

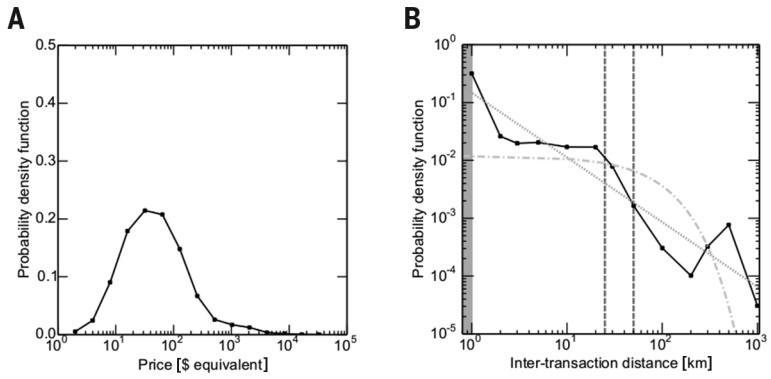


Fig. 5. Distributions of the financial records. (**A**) Probability density function of the price of a transaction in dollars equivalent. (**B**) Probability density function of spatial distance between two consecutive transactions of the same user. The best fit of a power law (dotted line) and an exponential distribution (dot-dashed line) are given as a reference. The dashed lines are the diameter of the first and second largest cities in the country. Thirty percent of the successive transactions of a user are less than 1 km apart (the shaded area), followed by, an order of magnitude lower, a plateau between 2 and 20 km, roughly the radius of the two largest cities in the country. This shows that financial metadata are different from mobility data: The likelihood of short travel distance is very high and then plateaus, and the overall distribution does not follow a power-law or exponential distribution.

Although future work is needed, it seems likely that most large-scale metadata data sets—for example, browsing history, financial records, and transportation and mobility data—will have a high unicity. Despite technological and behavioral differences (Fig. 5B and fig. S3), we showed credit card records to be as reidentifiable as mobile phone data and their unicity to be robust to coarsening or noise. Like credit card and mobile phone metadata, Web browsing or transportation data sets are generated as side effects of human interaction with technology, are subjected to the same idiosyncrasies of human behavior, and are also sparse and high-dimensional (for example, in the number of Web sites one can visit or the number of possible entry-exit combinations of metro stations). This means that these data can probably be relatively easily reidentified if released in a simply anonymized form and that they can probably not be anonymized by simply coarsening of the data.

Our results render the concept of PII, on which the applicability of U.S. and European Union (EU) privacy laws depend, inadequate for metadata data sets (*18*). On the one hand, the U.S. specific-types approach—for which the lack of names, home addresses, phone numbers, or other listed PII is enough to not be subject to privacy laws—is obviously not sufficient to protect the privacy of individuals in high-unicity metadata data sets. On the other hand, open-ended definitions expanding privacy laws to "any information concerning an identified or identifiable person" (*30*) in the EU proposed data regulation or "[when the] re-identification to a particular person is not possible" (*31*) for Deutsche Telekom are probably impossible to prove and could very strongly limit any sharing of the data (*32*).

From a technical perspective, our results emphasize the need to move, when possible, to more advanced and probably interactive individual (*33*) or group (*34*) privacy-conscientious technologies, as well as the need for more research in computational privacy. From a policy perspective, our findings highlight the need to reform our data protection mechanisms beyond PII and anonymity and toward a more quantitative assessment of the likelihood of reidentification. Finding the right balance between privacy and utility is absolutely crucial to realizing the great potential of metadata.

REFERENCES AND NOTES

1. S. Higginbotham, "For science, big data is the microscope of the 21st century" (2011); http://gigaom.com/2011/11/08/for-science-big-data-is-the-microscope-of-the-21st-century/.
2. D. Lazer *et al.*, *Science* **323**, 721–723 (2009).
3. J. Giles, *Nature* **488**, 448–450 (2012).
4. D. J. Watts, *Winter Issue of The Bridge on Frontiers of Engineering* **43**, 5–10 (2013).
5. A. Wesolowski *et al.*, *Science* **338**, 267–270 (2012).
6. S. Charaudeau, K. Pakdaman, P.-Y. Boëlle, *PLOS ONE* **9**, e83002 (2014).
7. N. Eagle, M. Macy, R. Claxton, *Science* **328**, 1029–1031 (2010).
8. V. Padmanabhan, R. Ramjee, P. Mohan, U.S. Patent 8,423,255 (2013).
9. G. Boulton, *Nature* **486**, 441 (2012).
10. M. McNutt, *Science* **346**, 679 (2014).
11. T. Bloom, "Data access for the open access literature: PLOS's data policy" (2013); www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy.
12. K. Burns, "In US cities, open data is not just nice to have; it's the norm" *The Guardian*, 21 October 2013; www.theguardian.com/local-government-network/2013/oct/21/open-data-us-san-francisco.
13. Massachusetts Bay Transportation Authority, "Real-time commuter rail data" (2010); www.mbta.com/rider_tools/developers/default.asp?id=21899.
14. Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, V. D. Blondel, D4D-Senegal: The second mobile phone data for development challenge. (2014); http://arxiv.org/abs/1407.4885.
15. V. D. Blondel *et al.*, Data for Development: The D4D challenge on mobile phone data. (2012); http://arxiv.org/abs/1210.0137.
16. P. Mutchler, "MetaPhone: The sensitivity of telephone metadata" (2014); http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata/.
17. Y.-A. de Montjoye, J. Quoidbach, F. Robic, A. Pentland, Predicting personality using novel mobile phone-based metrics. in *Proc. SBP* (Springer, Berlin, Heidelberg, 2013), pp. 48–55.
18. P. M. Schwartz, D. J. Solove, *Calif. Law Rev.* **102**, 877–916 (2014).
19. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, *Sci. Rep.* **3**, 1376 (2013).
20. A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets. in *IEEE Symposium on Security and Privacy*, Oakland, CA, 18 to 22 May 2008 (IEEE, New York, 2008), pp. 111–125.
21. A. C. Solomon, R. Hill, E. Janssen, S. A. Sanders, J. R. Heiman, Uniqueness and how it impacts privacy in health-related social science datasets. in *Proc. IHI* (Association for Computing Machinery, New York, 2012), pp. 523–532.
22. L. Sweeney, *Int. J. Unc. Fuzz. Knowl. Based Syst.* **10**, 557–570 (2002).
23. 2013 Federal Reserve payments study (2013); www.frbservices.org/files/communications/pdf/research/2013_payments_study_summary.pdf.
24. eMarketer, "US mobile payments to top $ 1 billion in 2013" (2013); www.emarketer.com/Article/US-Mobile-Payments-Top-1-Billion-2013/1010035.
25. "The trust advantage: How to win with big data" (2013); www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/.
26. C.-L. Huang, M.-C. Chen, C.-J. Wang, *Expert Syst. Appl.* **33**, 847–856 (2007).
27. S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, *Decis. Support Syst.* **50**, 602–613 (2011).
28. C. Krumme, A. Llorente, M. Cebrian, A. S. Pentland, E. Moro, *Sci. Rep.* **3**, 1645 (2013).
29. Materials and methods are available as supplementary materials on *Science* Online.
30. European Commission, "General data protection regulation" (2012); http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.
31. Deutsche Telekom, "Guiding principle big data" (2014); www.telekom.com/static/-/205808/1/guiding-principles-big-data-si.
32. Y.-A. de Montjoye, J. Kendall, C. Kerry, Enabling Humanitarian Use of Mobile Phone Data. *Brookings Issues in Technology Innovation Series* (Brookings Institution, Washington, DC, 2014), vol. 26.
33. Y.-A. de Montjoye, S.S. Wang, A.S. Pentland, *IEEE Data Eng. Bull.* **35**, 5–8 (2012).
34. C. Dwork, in *Automata, Languages and Programming* (Lecture Notes in Computer Science Series, Springer, Berlin, Heidelberg, 2006), vol. 4052, pp. 1–12.

ACKNOWLEDGMENTS

For contractual and privacy reasons, we unfortunately cannot make the raw data available. Upon request we can, however, make individual-level data of gender, income level, resolution (*h*, *v*, *a*), and unicity (true, false), along with the appropriate documentation, available for replication. This allows the re-creation of Figs. 2 to 4, as well as the GLM model and all of the unicity statistics. A randomly subsampled data set for the four points case can be found at http://web.media.mit.edu/~yva/uniqueintheshoppingmall/ and in the supplementary materials. This work was supported in part by the Geocrowd Initial Training Network funded by the European Commission as an FP7-People Marie Curie Action under grant agreement number 264994, and in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. Y.-A.d.M. was partially supported by the Belgian American Educational Foundation and Wallonie-Bruxelles International. L. R. did part of this work while visiting the MIT Media Lab. We gratefully acknowledge B. Bozkaya and a bank that wishes to remain anonymous for access to the data. Views and conclusions in this document are those of the authors and should not be interpreted as representing the policies, either expressed or implied, of the sponsors.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/347/6221/536/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 and S2
Algorithms S1 and S2
Reference (*35*)
Subsampled Data

20 May 2014; accepted 23 December 2014
10.1126/science.1256297