Editor: Cecilia Mascolo • *cecilia.mascolo@cl.cam.ac.uk*

Big Mobile Data Mining: Good or Evil?

Mirco Musolesi • *University of Birmingham, UK*

Increasingly, data are collected through sensors embedded in smartphones as well as through the cellular infrastructure. This information is extremely valuable for marketing applications, but also has an incredible potential to benefit society as a whole. Because mobile data are highly personal, however, privacy concerns are increasingly at the center of the public debate.

I still remember the first time, as a child, I entered the Biblioteca of Archiginnasio — one of the oldest libraries in my hometown of Bologna — and looked with a mixture of awe and curiosity at the hundreds and hundreds of dusty books sitting on the wooden shelves. The Biblioteca represented — and, indeed, represents — the accumulation of knowledge over the centuries, a sort of "big data" repository. This type of massive library always reminds me of the infinite library of Babel (described in the book *Fictions*[1]), a place where information is present somewhere but difficult to find and use.

In the Biblioteca, knowledge is printed on beautiful book pages, especially the medieval, miniated manuscripts. This knowledge is wonderful and profound, but static, in that it represents the past, not the present (and certainly not the future). This is the difference between the old concept of big data and that now emerging: today, we're used to information that isn't static, but is rather dynamic and continuously changing and updated over time. This trend is particularly visible in big data that are acquired in real-time, particularly those generated by personal mobile devices. In fact, for the first time in history, mining data from smartphones and the cellular infrastructure lets us access detailed knowledge about people, things, and events that are happening right now, potentially everywhere in the world. However, this trend has sparked a considerable debate among practitioners, researchers, and the public: is big mobile data mining "good" or "evil"?

The Big Mobile Data Trend

We've witnessed several fast-moving revolutions in the technological landscape in the past few years, particularly related to the emergence of powerful, always-connected, and extremely popular portable devices. Smartphones and tablets let us receive information through multiple channels while generating massive amounts of information about us. Data collected from the sensors embedded in smartphones — especially GPS receivers — provide an incredible wealth of information that service providers and applications can collect, store, and analyze in real time.

Computing is truly personal now, not only because we access information through mobile devices, but also because the information itself is usually highly personalized and relevant to our location and context. Typical examples are location-based services, such as Foursquare, which provide suggestions about restaurants and shops close to the area where users have checked in, and considers their previous mobility history. Other examples include search engines that are increasingly context- and location-aware. Moreover, users generate information themselves using mobile devices. For example, in June 2013, Facebook had, on average, 819 million monthly active mobile users. In other words, we should be talking not about the big data revolution but — at least as far as consumer applications are concerned — about the "big mobile data" phenomenon.

However, another trend will be even more central in the years to come: big mobile data are increasingly used not just for analyzing the past or understanding the present, but also for predicting the future. A new paradigm is slowly emerging that we might define as *anticipatory*

 1089-7801/14/$31.00 © 2014 IEEE

mobile computing.[2] Examples include companies that are developing innovative mining applications for real-time marketing or for supporting strategic decisions in retailing, such as the Telefonica Smart Steps project (http://dynamicinsights.telefonica.com/488/smart-steps). GoogleNow already provides users with relevant information related to the specific place a person is currently located and his or her habits, such as the current travel time to a location identified as the user's workplace. Soon, we will witness the emergence of new services (some start-ups are already working in this space) that give us information relevant to our future activities, such as possible events of interest on the weekend displayed on a Friday afternoon, or the expected traffic next Sunday in a certain motorway connecting two of our favorite leisure locations (GoogleTomorrow?).

Mining Big Mobile Data for Good

The information that smartphones generate directly (that is, from sensors embedded on devices) or that is indirectly collected by mobile operators (through cell registration patterns and call records) provides an immense opportunity for society at large.

Mobile big data can be used for "good" — for example, to improve transportation in developing countries,[3] devise strategies for epidemic containment,[4] or study social response during major disasters such as earthquakes.[5] Rather interestingly, given the fact that companies and other governmental and nongovernmental organizations own most of these data, a successful way to make them available for researchers and the public in general has been through challenges (such as the Orange Data for Development challenge; www.d4d.orange.com) or via hackathon events, which usually attract vast numbers of programmers and researchers. Other applications include analyzing big data to understand the behavior and emotional states of group of individuals[6] — for example, to tackle mental health problems such as depression by devising effective behavior intervention strategies at group or population levels.[7]

Another potential application for large-scale datasets is building analytical and predictive applications for law and order enforcement. You can already read titles in newspapers about the development of *Minority Report*-like systems.[8] Such headlines probably aren't true and shouldn't be considered worrying. Rather, most of these studies involve trying to identify crime hotspots in cities, analyze these hotspots' characteristics (for example, via census information), identify any emerging geographic patterns (that thefts happen on quiet roads close to major ones so perpetrators have easy access to fast escape routes, for instance) and — possibly — predict the evolution of criminal patterns. Such predictions could be extremely important for concentrating the efforts of law enforcement personnel in specific areas, especially today, when many countries' public resources are scarce.

In other words, big mobile data offer considerable potential not only for commercial applications but also for projects with high societal impact. Because this field is quite recent, however, researchers, practitioners, and society as a whole have yet to address many potential concerns. Let's examine some key, mostly nontechnical (or socio-technical) issues that are emerging in this area.

Privacy and Our Mobile Future

Because this article is, in a sense, about the future of computing, I can't avoid referring to one of the first people who thought about the digital age we're living in; he wrote probably the best article about the future possibilities computing has opened up. I am talking of Vannevar Bush's "As We May Think," published in the *Atlantic Monthly* in July 1945.[9] In this masterpiece, he also wrote about the upcoming deluge of scientific (and nonscientific) data, and computing technologies' potential, in the following terms:

> Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. If the aggregate time spent in writing scholarly works and in reading them could be evaluated, the ratio between these amounts of time might well be startling. [...] Mendel's concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it; and this sort of catastrophe is undoubtedly being repeated all about us, as truly significant attainments become lost in the mass of the inconsequential. [...]The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships. [...] But there are signs of a change as new and powerful instrumentalities come into use. Photocells capable of seeing things in a physical sense, advanced photography which can record what is seen or even what is not, thermionic tubes capable of controlling potent forces under the guidance of less power than a mosquito uses to vibrate his wings, cathode ray tubes rendering visible an occurrence so brief that by comparison a microsecond is a long time, relay combinations which will carry out involved sequences of movements more reliably than any human operator and thousands of times as fast — there are plenty of mechanical aids with which to effect a transformation in scientific records.

What Vannevar Bush likely couldn't imagine was that one day, most of this information would be

generated by our personal mobile devices, and that data wouldn't come in the form of recordings, but rather as streams of information from people all over the world in real time. He also probably wasn't expecting the merging of the physical world (in terms of physical and social sensing) and large-scale data mining. But what role do researchers and academics play in this new era of big mobile data? First, we must do good research as usual. But not only that: researchers and academics should educate society about what is possible (to do and know) with such data's availability.

Privacy has been a fundamental problem since ubiquitous computing emerged,[10] but this issue is now at the center of the public discussion because mobile devices are part of our lives. People are realizing that information about them is collected from their phones and stored in company and government databases. I believe that the fears related to big data — as seen as a way to control society — derive from the fact that most of these technologies look mysterious to the public. Indeed, large-data mining has recently received a fairly negative connotation.

I'm personally not against the collection of mobile data by companies and other organizations. As far as companies are concerned, consumers must be informed about what type of information companies hold about them, and about what can be extracted from these data. With respect to governmental organizations, I understand the importance of not revealing all the existing procedures, but we as a society must strengthen the checks that are in place (at least in democratic societies) to avoid any form of abuse. This can happen only if politicians and decision makers are also educated about the potential applications of today's large-scale data mining technologies.

New issues are also emerging with respect to the availability of large-scale predictive tools. In fact, concerns have arisen related not only to the privacy of past data or current information (such as my current location), but also to the inferences that can be made using available data. Even more concerning is that predictions and inferences might be incorrect, leading to erroneous deductions. So, should we worry about the privacy of our mobile future? Who owns our future in terms of big data prediction? Who controls the inferred information from big mobile data? Is it possible to devise systems that can exploit large-scale mobile data mining technologies and simultaneously protect our privacy?

Again, I personally believe that mining user data isn't wrong in principle — but, only if it isn't used against individuals in any way. At the same time, the public should be educated about the possibilities recent advances in machine learning, data mining (in particular text mining), mobile sensing, and large-scale data processing open up. The "feats" that are possible thanks to such technologies should be part of the standard body of knowledge for 21st century citizens and should thus be included in school curricula. As scientists, we should popularize these concepts and technologies and explain not only the risks, but also the benefits that society can derive by exploiting this increasing amount of mobile data.

This is important because we should look at the possibilities these emerging technologies open up with confidence. The potential applications are still largely unexplored. We should consider the possibility of collecting, mining, and extracting information from these data as exciting — not only as computer scientists, but also as citizens. We're entering a new and unprecedented era for citizens' science: new wearables will soon be available to the public (such as Google Glass), and for the first time, millions of users worldwide will be collecting a large amount of information (also longitudinal in nature).

We should consider the possibility of collecting and extracting useful information from mobile data and being able to make sense of them for the good of society as a triumph for computer science. After all, it's a testament to this field's power. As computer scientists, we shouldn't be ashamed of what's happening, but rather proud of today's possibilities. As I keep repeating, it is a fantastic time to be a computer scientist. And it is indeed a fantastic time for society to have computer scientists around who can provide new ways of accessing and using information to improve our present and our future. I am a bibliophile and probably a technophile too: for sure, my next mobile device won't be as beautiful as one of those illuminated manuscripts in the Biblioteca dell'Archiginnasio — but it will be as useful as its progenitor.

References

1. J.L. Borges, *Fictions*, Penguin Classics, 2000.
2. V. Pejovic and M. Musolesi, *Anticipatory Mobile Computing: A Survey of the State of the Art and Research Challenges*, tech. report, School of Computer Science, Univ. of Birmingham, 2013; Arxiv 1306.2356.
3. M. Berlingerio et al., "AllBoard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data," *Machine Learning and Knowledge in Databases*, LNCS 8190, Springer, 2013, pp. 663–666.
4. A. Lima, M. De Domenico, and M. Musolesi, "Exploiting Cellular Data for Disease Containment and Information Campaign Strategies in Country-Wide Epidemics," *Proc. 3rd Int'l Conf. Analysis of Mobile Phone Datasets* (NetMob 13), 2013.
5. B. Moumni, V. Frias-Martinez, and E. Frias Martinez, "Characterizing Social

Response to Urban Earthquakes using Cell-Phone Network Data: The 2012 Oaxaca Earthquake," *Proc. Workshop on Pervasive Urban Applications* (PURBA 13), ACM, 2013, pp. 1199–1208.
6. D. Lazer et al., "Computational Social Science," Science, vol. 323, no. 5915, 2009, pp. 721–723.
7. N. Lathia et al., "Smartphones for Large-Scale Behavior Change Intervention," *IEEE Pervasive Computing*, July–September, 2013, pp. 66–73.
8. M.B. Short et al., "A Statistical Model of Criminal Behavior," *Mathematical Models and Methods in Applied Sciences*, vol. 18, supp. 1, 2008, pp. 1249–1267.
9. V. Bush, "As We May Think," *Atlantic Monthly*, vol. 176, no. 1, 1945, pp. 101–108.
10. M. Langheinrich, "Privacy by Design: Principles of Privacy-Aware Ubiquitous Systems," Proc. *UbiComp 2001*, LNCS 2201, Springer, 2001, pp. 273–291.

Mirco Musolesi is a senior lecturer in the School of Computer Science at the University of Birmingham, UK. His research interests lie at the intersection of ubiquitous computing, mobile sensing, large-scale data mining, and network science. Musolesi has a PhD in computer science from University College London. Contact him at m.musolesi@cs.bham.ac.uk; www.cs.bham.ac.uk/~musolesm.