The mirages of big data

Ben Li
Department of Information Processing
Science, University of Oulu
PL 3000, 90014 OULUN YLIOPISTO.
Oulu, Finland
banji.li@oulu.fi

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.
ACM 978-1-4503-1051-2/12/02.

Abstract

Organizations have employed topically, spatially, and temporally large data sets for centuries, suggesting that big data is not a fundamentally new problem to humanity. This paper proposes to define the big data problem in social terms of knowledge and collaboration rather than in technological terms of data and silicon.

Author Keywords

Big data; social science; information infrastructure

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human factors, cyberinfrastructure, standards

Introduction

Despite recent (re)cognition of the big data problem in information and communication technology (ICT)-intensive research, big data has been a persistently (re)solved problem. The Tabulating Machine Company of Washington, D.C. was founded to mobilize social information from the U.S. census in the 1880s. Since the 12th century CE, Common Law and Canon Law have evolved to monitor and manage millions of relationships among individuals, organizations, and assets, across spatial and temporal scales. A millennium before the Common Era, the Chinese commandery system centrally mobilized many data items and data sets to manage agricultural production, land use, and resource distribution [4]. Yet researchers continue to discover that: “The era of Big Data has begun” [1]. This paper argues that since big data is a largely solved problem, efforts should focus on social dimensions of new kinds of (scientific) collaboration we seek to build.

Maslow's digital hammer

The “big data” literature reveals some major themes: access and “digital divides”; automation of data collection, processing, and storage; epistemology of computer-mediated social networks; cloud issues; and personal autonomy and privacy. They define outstanding problems in terms of ICTs, not social collaboration. Solutions depend on how such massively multivariate problems are defined [12]. Defining insightful interdisciplinary research as an ICT or data problem—e.g., [6]—tempt us to solve the creative collaboration problem using algorithmic tools. But Shannon warns of diminishing gains from increasing data size alone. The following observations suggest that the social dimensions of big data are not well operationalized by the big data concept.

“Open data” as a big data problem

“Open data”, “open government”, “e-government”, “e-democracy”, “transparency”, etc. all broadly assume that read and write access to larger quantities and varieties of factual and procedural information yields better policy and decision-making processes and outcomes. This approach perpetually yields “apps” that compel stakeholders in the same room to allegedly work together by individually engaging into screens to guess which artificially constrained public policy options presented by programmers is least bad. Long-term accountability benefits potentially arise from analysing patterns of decision- and policy-making. Supplying unmanageable piles of data to individual members of the public is supposed to solve complex long-term knowledge problems that defy coordinated staff and capital resources of governments and media empires. Global citizens’ uprisings in 2011 about poor short- and long- term decision-making and accountability occurred in regions mobilizing minimal through extensive data about citizens’ preferences and activities. How can we reconcile such events with assumptions about big data’s potentials?

Infrastructures for big data

Some public big data infrastructures are well understood in social science. Censuses are reliable sources of information for public officials, private individuals, and other stakeholders. We assume censuses will occur, and only question their absence [2]. Every major liberal democracy also manifests organizations that collect data almost continuously about the current performance of government, including: media, private research firms, political parties, lobby groups, non-governmental organizations, etc. These data consistently cover decades of issues, policies, events, and feedback across diverse factors and settings. Google, Facebook, political dynasties, and journalists appear to similarly mobilize such data (for good and bad) using commodity equipment and skills.

The mere presence of transparent process, data, long-term well-understood information systems, or other infrastructures, does not convincingly result in improved collaboration around such social science big data. Nor does successful collaboration appear to depend on particularly unique information systems.

Interoperability and standards

Enacted laws and pending bills have been public data for as long as the parliamentary system has existed, and are globally based on a small number of standards: Common Law, Civil Law, Sharia Law, and Confucian principles. Laws represent some of the best-understood, most widely adopted, and freely scalable interoperable systems of processes and data humanity has developed. Yet, if laws encoded logics and values well, we would not need packs of lawyers or extensive legal systems to resolve ambiguities. Computers have broadened access to questionable crowd-sourced legal advice, drawn from pages indexed by Google, rather than help us develop better laws.

Much of the world collaborates through English—a highly interoperable yet exceedingly loose practical standard on- and off- line. Political stakeholders argue for years without stating definitions. Biologists debate ontologies in long-term ecological research. Social scientists disagree about the core principles of their work. Yet they produce solid science from large and diverse data despite gross violations of standards.

Data-intensive collaboration

Big data implies big people: diverse stakeholders need not agree on any particular standard version or view of data as long as each stakeholder's work can draw from and add to the whole. ICTs could help to make some contributions (more) discoverable, but only if ICTs encode the kinds of interpretative flexibility abhorred by some "hard" science practitioners. Computerized or not, censuses and catalogues enable many kinds of stakeholders to finish relatively raw data for their own needs, on demand, indefinitely. The space program developed and assembled physical and knowledge systems from personal and technological components having contracts about requirements and capabilities. The Internet was a public project designed to enable any socially trusted individual to design and connect any data system at any technical, social, and cultural level(s). These problems concerning big data were successfully solved in the 1970s, but not necessarily as scaled up little data problems. The demand is never for big data or any scheme to manage it, but for data that is available to be profitably used.

So let us instead consider the big data problem in terms of developing useful abstractions from data that may be shared and operationalized to generate new kinds of value.

What do we need?

Industry and academia have hundreds of ways to store, describe, and interchange very large data sets. Many pages have been written about successful big data cyber-infrastructures at focused international scientific collaborations such as LHC and others in high-energy physics, SETI and others in astronomy, and GenBank and others in bioinformatics. But such focused social and technical infrastructures may not scale to heterogeneous questions. Diversity of large data sets and problems are even greater outside formal science in efforts such as organ donor registries, peer-to-peer networks, distributed computing, and everything else in the long tail.

Successes in social science suggest considering big data not as piles of servers and databases to be federated, but as a new kind of purposeful sustainable science design that treats each research effort itself as data. (This new design may upset expertise and social structures based on assumptions about non-big data or its technology.) Big data is interesting because it considers emergent phenomena that do not manifest in small and medium scales, and therefore requires new (to each community) knowledge about new tools and how to connect them. Since most uses of statistical methods are improper [3] and go undetected in review, the expectation to handle large quantities of data from different contexts could appear to researchers as a new problem or threat. By definition there can be no quality control mechanisms to evaluate work that is perceived to be completely new.

For work with interlinked data and knowledge ("collaborative insight") to become the rule, rather than the exception, our social and ICT knowledge discovery and recognition systems must re-orient from data toward connections. "Connections" do not primarily consist of long-standing patterns of social networks, communities, and infrastructures *now with 99% more Internet*. Nor does it mean algorithmically adding descriptive meta-data to existing data hoping that someone else will do the hard work of considering specific potential re-uses. Connections must reflect innovative knowledge relationships embodied in people who use ICTs as only one way to formalize some of what occurs in science. Counter-intuitively, that would require us to put less emphasis on publishing data by making that activity as routine as checking e-mail, and more emphasis on value-adding social knowledge activities that defy specific content molds.

Just as researchers assumed an infrastructure of data-gathering instruments in order to focus on data use, so must researchers overcome the urge to build new information storage systems. That would liberate minds to consider and collaborate on systems of knowledge without being locked into any particular system of data.

Current scholarly infrastructures and interfaces do not encourage broad or long-term thinking about data. Handfuls of keywords to enable discoverability by non-human search engines cannot capture meanings of walls of static selfishly produced text with dynamic meanings. That there are few visible interdisciplinary efforts and many great difficulties with meta-data systems, suggests that not only is the "data deluge" not happening, but that it would be a welcome problem for everyone except current structurally protected scholarly elites.

Conclusions

When the data deluge does occur outside the social sciences, collaborative insight challenges will not be solvable as just one problem. Already, ecologists routinely argue that every piece of collected data should be coveted since it may never be collected again [7], despite lack of clear future uses. Others include: tracking epidemics by re-using data in public tweets [5]; asking the public to socially sample the same instant world-wide [9]; and providing the public with new instruments to collect immediately useful scientific and public policy data [10]. The implicit standard(s) model(s) employed appear to work well enough for both citizens and scientists.

Interdisciplinary knowledge infrastructure gaps may be less easily identified by individual experts, while the interdisciplinary nature of diverse collaborations would require more diverse attention as the work is produced. We can only know that collaborative insight has succeeded when it meets some design, practical, or theoretical ideal. As of yet, we have few good theories about how macro-scale human systems (should) operate more generally, and therefore lack ways to gauge how we perform with respect to the specific case of macro-scale human collaborative insight.

Successful collaborative insight practices will not necessarily meet hard science's hypo-deductive pure ideals. Instead they more closely resemble grounded (re-)explorations that enable the social sciences and humanities to provide context-aware insights in an inherently ambiguous and ever-changing world. Unlike [8] who criticized Tai for rediscovering, from first-year calculus, the trapezoid rule for finding the area under a curve [11], we should celebrate useful connections and (re-)cognitions among diverse knowledge stakeholders. Before designing new ICT infrastructures we must first socially envision the kinds of collaboration, science, scholarship, or other practices that exceed our current practices and capabilities.

The idea of collaborative insight enabled by big data systems lets us consider how our vast technological and communication infrastructures enable both expert and non-expert consumers of science to contemplate and undertake new research to generate immediately useful and reusable knowledge.

Acknowledgements

This work was funded by the Academy of Finland. It was made possible by informative discussions with the staff, students, and visitors at the University of Oulu's Oulanka Research Station, Helsinki University's Kilpisjärvi Biological Station, and the Taiwan Forestry Research Institute.

References

[1] Boyd, D., & Crawford, K. Six Provocations for Big Data. A Decade in Internet Time. Proc Symposium on the Dynamics of the Internet and Society, 2011.

[2] Green, D. A., & Milligan, K. The Importance of the Long Form Census to Canada. Canadian Public Policy, 36, 3 (2010), 383-388.

[3] Ioannidis, J. P. Why Most Published Research Findings Are False. PLoS Medicine, 2, 8 (2005), 696-721.

[4] Kiser, E., & Cai, Y. War and Bureaucratization in Qin China: Exploring an Anomalous Case. American Sociological Review, 68, 4 (2003), 511-539.

[5] Lampos, V., Bie, T. D., & Cristianini, N. Flu Detector - Tracking Epidemics on Twitter. In J. L. Balcázar, F. Bonchi, A. Gionis, & M. Sebag (Ed.), European Conference Machine

Learning and Knowledge Discovery in Databases Proceedings, Part III, 2010 (pp. 599-602).

[6] Manovich, L. Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), Debates in the Digital Humanities. The University of Minnesota Press, 2012.

[7] Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. Nongeospatial Metadata for the Ecological Sciences. Ecological Applications, 7, 1 (1997), 330-342.

[8] Monaco, J., & Anderson, R. Tai's Formula Is the Trapezoidal Rule. Diabetes Care, October (1994) 1224-1225.

[9] One Day On Earth. http://www.onedayonearth.org/

[10] Parvaz, D. Crowdsourcing Japan's radiation levels, Al Jazeera, April 28, 2011, http://www.aljazeera.com/news/asia/2011/04/201142317359479927.html

[11] Tai, M. M. A Mathematical Model for the Determination of Total Area Under Glucose Tolerance and Other Metabolic Curves. Diabetes Care, 17, 2 (1994), 152-154.

[12] Whelton, M., & Ballard, G. Wicked Problems in Project Definition. Proc International Group for Lean Construction, 2002.

About the author

The author has held various stakeholder roles in data and knowledge management as a data manager in a national social science research network, as a research and development manager in an international ICT firm, as a scholar studying research management and governance in large open source projects and organizations, and as a policy analyst for a regional government. The author’s current research interests connect e-democracy, international governance, knowledge management, and information infrastructures.