Science AAAS

More Is Less: Signal Processing and the Data Deluge

Richard G. Baraniuk

Science **331**, 717 (2011);
DOI: 10.1126/science.1197448

need to be established if a data-sharing network is to succeed, particularly when it comes to the ethical and privacy issues surrounding patient data (*23*, *24*).

Shifting Attitudes

Widely dispersed researchers in resource-limited countries may have few opportunities to travel to courses or attend meetings, but they can meet online and share experiences, guide each other, and access resources. Learning and knowledge sharing online could play a vital role in adjusting the imbalance in research capacity. However, this medium for learning needs to become accepted, and senior research staff need to encourage and enable their colleagues to take up the numerous free and open-access learning opportunities that are increasingly available online (*13*).

Undoubtedly integration and knowledge sharing can be vastly improved to make the most use of gathered data, but many organizations in global health exist to address a single disease or work in a specific sector. There is a real need for mechanisms allowing research organizations, governments, and universities to collaborate outside their usual remits and locations to maximize the impact of data and available resources.

Governance and ethical issues are also a major concern, because if mistakes are made trust will be quickly lost and enthusiasm for opening access could be stifled. A particular anxiety resulting from disparities between wealthy and resource-limited nations is the removal of data and loss of ownership. Ownership and governance arrangements need to be made transparently for fair access and maintenance of security, and whenever possible the technology should be transferred rather than the data. These issues therefore need to be tackled openly and comprehensively early in the formation of data-sharing collaborations. Groups would be advised to seek advice and obtain example policy documents (such as agreements and terms of reference) from other successful data-sharing groups.

A striking range of data sets spanning a wide range of healthcare issues, including infectious and noncommunicable diseases, are accumulating with use of new technology and online collaboration. All this stands to make real changes in the lives of people affected by diseases of poverty. While scientists are rapidly adapting and taking up these approaches, funding agencies and regulators also need to adapt to ensure that all interested communities are able to take maximum advantage of the digital environment to drive improvements in global health.

References and Notes

1. P. Mwaba, M. Bates, C. Green, N. Kapata, A. Zumla, *Lancet* **375**, 1874 (2010).
2. E. Wenger, W. Snyder, *Harv. Bus. Rev.* **2000**, 139 (Jan.-Feb. 2000).
3. A. de-Graft Aikins *et al.*, *Global. Health* **6**, 5 (2010).
4. P. Kowal *et al.*, *Glob. Health Action* **3** (suppl. 2), 10.3402/gha.v3i0.5302 (2010).
5. OpenXData, www.openxdata.org.
6. EAIDSNet, www.eac.int.
7. H. F. Wertheim *et al.*, *PLoS Med.* **7**, e1000231 (2010).
8. The South East Asia Infectious Disease Clinical Research Network, www.seaicrn.org.
9. S. I. Hay, R. W. Snow, *PLoS Med.* **3**, e473 (2006).
10. The Malaria Genomic Epidemiology Network, *Nature* **456**, 732 (2008).
11. M. Parker *et al.*, *PLoS Med.* **6**, e1000143 (2009).
12. G. W. Fegan, T. A. Lang, *PLoS Med.* **5**, e6 (2008).
13. T. A. Lang *et al.*, *PLoS Negl. Trop. Dis.* **4**, e619 (2010).
14. A. M. Dondorp *et al.*, *Lancet* **376**, 1647 (2010).
15. P. J. Guerin, S. J. Bates, C. H. Sibley, *Curr. Opin. Infect. Dis.* **22**, 593 (2009).
16. M. Pirmohamed, K. N. Atuah, A. N. Dodoo, P. Winstanley, *Br. Med. J.* **335**, 462 (2007).
17. A. S. Kanter *et al.*, *Int. J. Med. Inf.* **78**, 802 (2009).
18. D-Tree International, www.d-tree.org/.
19. S. F. Noormohammad *et al.*, *Int. J. Med. Inf.* **79**, 204 (2010).
20. B. A. Fischer, M. J. Zigmond, *Sci. Eng. Ethics* **16**, 783 (2010).
21. E. Pisani, C. AbouZahr, *Bull. W. H. O.* **88**, 462 (2010).
22. J. Whitworth, *Bull. W. H. O.* **88**, 467 (2010).
23. B. Malin, D. Karp, R. H. Scheuermann, *J. Investig. Med.* **58**, 11 (2010).
24. R. Horton, *Lancet* **355**, 2231 (2000).

25. The author received no specific funding for this work and has no conflicts of interest to declare.

10.1126/science.1199349

PERSPECTIVE

More Is Less: Signal Processing and the Data Deluge

Richard G. Baraniuk

The data deluge is changing the operating environment of many sensing systems from data-poor to data-rich—so data-rich that we are in jeopardy of being overwhelmed. Managing and exploiting the data deluge require a reinvention of sensor system design and signal processing theory. The potential pay-offs are huge, as the resulting sensor systems will enable radically new information technologies and powerful new tools for scientific discovery.

Until recently, the scientist's problem was a "sensor bottleneck." Sensor systems produced scarce data, complicating subsequent information extraction and interpretation. In response to the resulting challenge of "doing more with less," signal-processing researchers have spent the last several decades creating powerful new theory and technology for digital data acquisition (digital cameras, medical scanners), digital signal processing (machine vision; speech, audio, image, and video compression), and digital communication (high-speed modems, Wi-Fi) that have both enabled and accelerated the information age.

These hardware advances have fueled an even faster exponential explosion of sensor data produced by a rapidly growing number of sensors of rapidly growing resolution. Digital camera sensors have dropped in cost to nearly $1/megapixel; this has enabled billions of people to acquire and share high-resolution images and videos. Millions of security and surveillance cameras, including unmanned drone aircraft prowling the skies, have joined high-resolution telescopes, digital radio receivers, and many other types of sensors in the environment. As a result, a sensor data deluge is beginning to swamp many of today's critical sensing systems.

In just a few years, the sensor data deluge has shifted the bottleneck of many data acquisition systems from the sensor back to the processing, communication, or storage subsystems (Fig. 1). To see why, consider the exponentially growing gap between global sensing and data storage capabilities. A recent report (*1*) found that the amount of data generated worldwide (which is now dominated by sensor data) is growing by 58% per year; in 2010 the world generated 1250 billion gigabytes of data—more bits than all of the stars in the universe. In contrast, the total amount of world data storage (in hard drives, memory chips, and tape) is growing 31% slower, at only 40% per year. A milestone was reached in 2007, when the world produced more data than could fit in all of the world's storage; in 2011 we already produce over twice as much data as can be stored. This expanding gap between sensor data production and available data storage means that sensor systems will increasingly face a deluge of data that will be unavailable later for further analysis. Similar exponentially expanding gaps exist between sensor data production and both computational power and communication rates.

The danger is that more sensor data can lead to less efficient sensor systems. Consider two brief illustrations. The first is the Defense Advanced Research Projects Agency (DARPA) Autonomous Real-Time Ground Ubiquitous

Department of Electrical and Computer Engineering, Rice University, Houston, TX 77251–1892, USA. E-mail: richb@rice.edu

Surveillance Imaging System (ARGUS-IS) developed for military reconnaissance and real-time monitoring that features a 1.8-gigapixel digital camera constructed from hundreds of cell phone camera chips (*2*). Each camera image covers up to 160 km^2 (almost the size of greater Los Angeles) with a 30-cm ground resolution. When acquiring video at 15 frames per second, the camera produces raw data at a rate of 770 gigabits per second (Gbps). In stark contrast, the wireless communications link to the ground station (where the data are to be exploited by signal processing algorithms) has a maximum rate of just 274 megabits per second (Mbps). Even using today's state-of-the-art video compression algorithms, the camera sensor produces hundreds of times more image and video data than can ever be communicated off the platform. Moreover, moving the ground station's signal processing hardware up to the sensing platform is out of the question, because it occupies several large racks of computers.

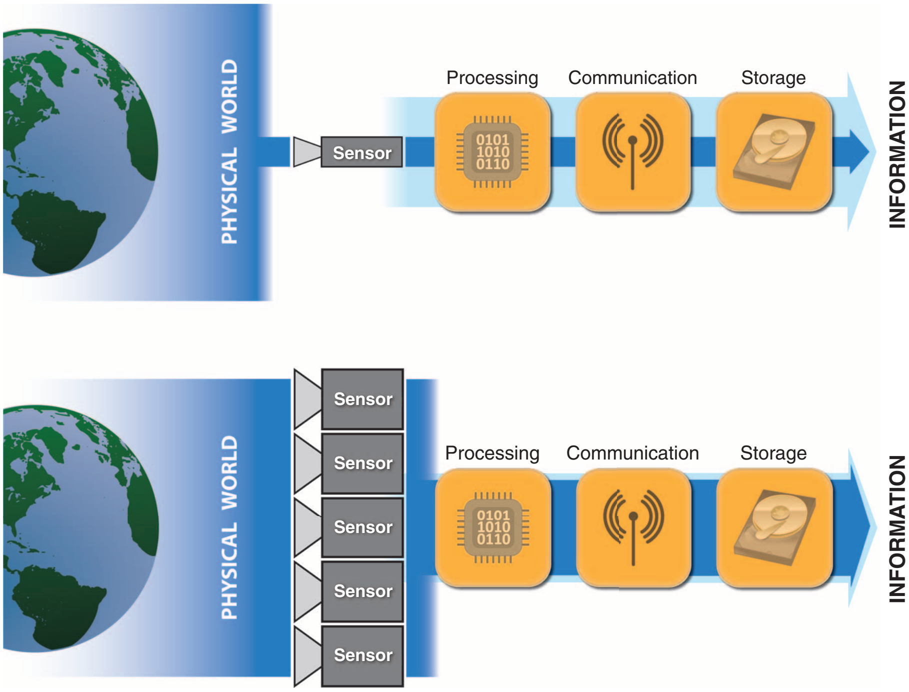

Fig. 1. Dealing with the sensor data deluge. In a conventional sensing system (top), the sensor is the performance bottleneck. In a data deluge–era sensing system (bottom), the number and resolution of the sensors grow to the point that the performance bottleneck moves to the sensor data processing, communication, or storage subsystem.

The second example is the Compact Muon Solenoid (CMS) detector of the Large Hadron Collider at CERN, which will produce raw measurement data at a rate of 320 terabits per second (Tbps), far beyond the capabilities of either processing or storage systems today (*3*). As a stopgap measure, custom hardware carefully triages the raw data stream to a rate of 800 Gbps by selecting only the potentially "interesting" events, which are then fed into a real-time computing farm to further process and compress for storage. All other events are lost in the acquisition process.

Given the growing gap between the amount of data we produce and the amount of data we can process, communicate, and store, systems like ARGUS-IS and the CMS will become more the norm than the exception over time. Successfully navigating the data deluge calls for fundamental advances in the theory and practice of sensor design; signal processing algorithms; wideband communication systems; and compression, triage, and storage techniques.

A recent Frontiers of Engineering event examined some of the encouraging preliminary results in these directions (*4*). One promising direction is the design of new kinds of data acquisition systems that replace conventional sensors with compressive sensors that combine sensing, compression, and data processing in one operation. The key enabler is the recognition that the amount of information in many interesting signals is much smaller than the amount of raw data produced by a conventional sensor. More technically, many interesting signals inhabit an extremely low-dimensional subset of the high-dimensional raw sensor data space. Rather than first acquiring a massive amount of raw data and then boiling it down into information via signal processing algorithms, compressive sensors attempt to acquire the information directly.

Such low-dimensional signal structure may manifest itself in a number of different ways. In a sparse signal model, N raw data samples can be transformed to a domain where only K (much less than N) representation coefficients are nonzero (*5*, *6*). Sparse models lie at the heart of popular compression and processing algorithms such as JPEG. In a manifold signal model, the raw data can be parameterized (nonlinearly, in general) using just K parameters (*7*). Such a model is natural for imaging problems involving a known object and K unknown camera parameters. Recent research on compressive sensing has led to two results that in combination promise to temper the data deluge. First, signals from both sparse and manifold models can be acquired without information loss using just on the order of $K\log N$ compressive measurements rather than N raw-data measurements (*5*, *6*, *8*). Second, a range of different signal processing algorithms can extract the salient signal characteristics directly from the low-rate compressive measurements (*9*). The sensing protocols that achieve this low measurement rate are inherently random and distinct from the classical Shannon-Nyquist sampling theory that dominates digital sensing theory and practice.

In another promising direction, researchers are turning the data deluge to their advantage by replacing conventional signal processing algorithms based on mathematical models with new algorithms that mine the deluge. One striking example is a tool that fuses a large collection of unorganized images of a scene (say, photos of Notre Dame cathedral from the photo-sharing Web site Flickr) and automatically computes each photo's viewpoint and a three-dimensional model of the scene (*10*).

In the long run, without radical superexponential advances in computer processing, communication, and storage capabilities, the data deluge is here to stay. The next generation of sensor designs and signal processing theory will have to harness the deluge in order to do more, rather than less, with its bounty. The broader implications for science and engineering are appreciable. Can scientific conclusions be trusted when the raw experimental data are lost and the data triage or compression algorithm might be suspect? Can we resist the temptation to equate correlation with causation when mining massive data sets for scientific conclusions? Can we develop the new low-complexity mathematical models and the new practical sensing protocols that are needed to effectively extract information from the bulk of the deluge? Clearly, these are exciting times for sensor system design.

References

1. J. Gantz, D. Reinsel, "The Digital Universe Decade—Are You Ready?" IDC White Paper, May 2010; http://idcdocserv.com/925.
2. DARPA ARGUS-IS program, www.darpa.mil/i2o/programs/argus/argus.asp.

3. The CMS Collaboration, *J. Instrumentation* **3**, S08004 (2008).
4. U.S. National Academy of Engineering and Royal Academy of Engineering, Frontiers of Engineering, EU-US Symposium, Cambridge, UK, 31 August to 3 September 2010; www.raeng.org.uk/international/activities/frontiers_engineering_symposium.htm.
5. E. J. Candès, J. Romberg, T. Tao, *IEEE Trans. Inf. Theory* **52**, 489 (2006).
6. D. L. Donoho, *IEEE Trans. Inf. Theory* **52**, 1289 (2006).
7. J. B. Tenenbaum, V. de Silva, J. C. Langford, *Science* **290**, 2319 (2000).
8. R. G. Baraniuk, M. B. Wakin, *Found. Comput. Math.* **9**, 51 (2009).
9. S. Muthukrishnan, *Found. Trends Theor. Comput. Sci.* **1** (issue 2), 117 (2005).
10. N. Snavely, S. M. Seitz, R. Szeliski, *ACM Trans. Graph.* **25**, 835 (2006).

10.1126/science.1197448

PERSPECTIVE

Ensuring the Data-Rich Future of the Social Sciences

Gary King

Massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress. I address these changes and challenges and suggest what can be done.

Fifteen years ago, *Science* published predictions from each of 60 scientists about the future of their fields (*1*). The physical and natural scientists wrote about a succession of breathtaking discoveries to be made, inventions to be constructed, problems to be solved, and policies and engineering changes that might become possible. In sharp contrast, the (smaller number of) social scientists did not mention a single problem they thought might be addressed, much less solved, or any inventions or discoveries on the horizon. Instead, they wrote about social science scholarship—how we once studied *this*, and in the future we're going to be studying *that.*

Fortunately, the editor's accompanying warning was more prescient: "history would suggest that scientists tend to underestimate the future" (*2*).

Indeed. What the social scientists did not foresee in 1995 was the onslaught of new social science data—enormously more informative than ever before—and what this information is now making possible. Today, huge quantities of digital information about people and their various groupings and connections are being produced by the revolution in computer technology, the analog-to-digital transformation of static records and devices into easy-to-access data sources, the competition among governments to share data and run randomized policy experiments, the new technology-enhanced ways that people interact, and the many commercial entities creating and monetizing new forms of data collection (*3*).

Analogous to what it must have been like when they first handed out microscopes to microbiologists, social scientists are getting to the point in many areas at which enough information exists to understand and address major previously intractable problems that affect human society. Want to study crime? Whereas researchers once relied heavily on victimization surveys, huge quantities of real-time geocoded incident reports are now available. What about the influence of citizen opinions? Adding to the venerable random survey of 1000 or so respondents, researchers can now harvest more than 100 million social media posts a day and use new automated text analysis methods to extract relevant information (*4*). At the same time, parts of the biological sciences are effectively becoming social sciences, as genomics, proteomics, metabolomics, and brain imaging produce large numbers of person-level variables, and researchers in these fields join in the hunt for measures of behavioral phenotypes. In parallel, computer scientists and physicists are delving into social science data with their new methods and data-collection schemes.

The potential of the new data is considerable, and the excitement in the field is palpable. The fundamental question is whether researchers can find ways of accessing, analyzing, citing, preserving, and protecting this information. Although information overload has always been an issue for scholars (*5*), today the infrastructural challenges in data sharing, data management, informatics, statistical methodology, and research ethics and policy risk being overwhelmed by the massive increases in informative data. Many social science data sets are so valuable and sensitive that when commercial entities collect them, external researchers are granted almost no access. Even when sensitive data are collected originally by researchers or acquired from corporations, privacy concerns sometimes lead to public policies that require the data be destroyed after the research is completed—a step that obviously makes scientific replication impossible (*6*) and that some think will increase fraudulent publications (*7*).

Indeed, we appear to be in the midst of a massive collision between unprecedented increases in data production and availability about individuals and the privacy rights of human beings worldwide, most of whom are also effectively research subjects (Fig. 1).

Consider how much more informative to researchers, and potentially intrusive to people, the new data can be. Researchers now have the possibility of continuous-time location information from cell phones, Fastlane or EZPass transponders, IP addresses, and video surveillance. We have information about political preferences from person-level voter registration, primary participation, individual campaign contributions, signature campaigns, and ballot images. Commercial information is available from credit card transactions, real estate purchases, wealth indicators, credit checks, product radio-frequency identification (RFIDs), online product searches and purchases, and device fingerprinting. Health information is being collected via electronic medical records, hospital admittances, and new devices for continuous monitoring, passive heart beat measurement, movement indicators, skin conductivity, and temperature. Extensive quantities of information in unstructured textual format are being produced in social media posts, e-mails, product reviews, speeches, government reports, and other Web sources. Satellite imagery is increasing in resolution and scholarly usefulness. Social everything—networking, bookmarking, highlighting, commenting, product reviewing, recommending, and annotating—has been sprouting up everywhere on the Web, often in research-accessible ways. Participation in online games and virtual worlds produces even more detailed data. Commercial entities are scrambling to generate data to improve their business operations through tracking employee behavior, Web site visitors, search patterns, advertising click-throughs, and every manner of cloud services that capture more and more information.

Efforts in the social sciences that make data, code, and information associated with individual published articles available to other scholars have been advancing through software, journal policies, and improved researcher practices for some time (*8*, *9*). However, this movement is at risk of

Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, MA 02138, USA. E-mail: king@harvard.edu