Big Data's Big Unintended Consequences

Marcus R. Wigan, *Oxford Systematics, Swinburne University, and the University of Melbourne*
Roger Clarke, *Xamax Consultancy, University of New South Wales, and Australian National University*

Businesses and governments exploit big data without regard for issues of legality, data quality, disparate data meanings, and process quality. This often results in poor decisions, with individuals bearing the greatest risk. The threats harbored by big data extend far beyond the individual, however, and call for new legal structures, business processes, and concepts such as a Private Data Commons.

Big data has been coming for years.

Dataveillance, introduced in 1988, offers a more economical method for monitoring individuals than physical and electronic surveillance.[1] Early techniques included front-end verification and data matching. Profiling, an important development in this area, involves inferring from existing data holdings a set of characteristics for a particular category of person, with the intention of singling out other individuals who closely fit that set of characteristics.

Following the development and application of neural networks and other rule-generation tools, a larger-scale process emerged. The search for a new term to excite customers and achieve sales led to the adoption of "data mining." This term framed the data as raw material, and the process as the exploitation of that resource to extract subtle, complex, or multidimensional relationships.

"Big data," an expression that's been in the formal literature since the 1990s, typically refers not only to specific, large datasets, but also to data collections that consolidate many datasets from multiple sources, and even to the techniques used to manage and analyze the data. Its original use appears to have been in the physical sciences, where economics has dictated that computational analysis and experimentation complement and even supplant costly, messy physical laboratories. A vast amount of data is generated by applications of big data techniques in such undertakings as the Search for Extra-Terrestrial Intelligence (SETI), genome projects, CERN's Large Hadron Collider, and the Square Kilometer Telescope Array.

Big data techniques subsequently found application in other disciplines, and gave rise to the field of computational social science. Large quantities of health and social welfare data already exist. New sources of big data include locational data arising from traffic management, and from the tracking of personal devices such as smartphones. This article focuses not on data about physical phenomena, but on data that relates to individuals who are identifiable, or to categories of individuals.

Corporations see big data as a tool for commercial advantage, particularly in consumer marketing.[2] Much of the populist management literature is expressed in vague terms, but some authors deal with specific cases.[3] More recently, the big data idea has been grasped as a mantra by government agencies, with the expectation of attacking

 Published by the IEEE Computer Society

waste and fraud, and by law enforcement and national security agencies promising more frequent and earlier detection of terrorists.

Businesses and governments exploit big data, often pressing the limits of legality, data quality, disparate data meanings, and process quality. This can result in poor decisions, with individuals bearing the greatest risk. The potential negative outcomes enabled by big data extend far beyond the individual, into social, economic, and political realms. We need new balances to handle the resulting power shifts. A suitable framework for the coherent treatment of these side effects can be derived from recent responses to environmental losses and the concept of the commons, by a proposed analogous concept of the private data commons.

THE POLITICAL ECONOMY OF BIG DATA

Some important and commonly overlooked presumptions underlie the wave of big data enthusiasm.

In some cases, a big data collection can arise from a single coherent and consistent data acquisition process. In other cases, however, quantities of data are acquired from multiple sources and combined. The legality of the collection activity, the disclosure, the consolidation, and the mining of the consolidated database might be resolved, asserted, or merely assumed.

The quality of the original data varies, with inherent accuracy, precision, and timeliness problems. When data is repurposed, disclosed, or expropriated, the widely varying quality levels of data in the individual databases result in yet lower quality levels in the overall collection.

The meaning of each item in a database is frequently far from clear. Nonetheless, drawing together these items into a single database implicitly assumes that data items from different databases with apparent similarities are sufficiently compatible that equivalence can be imputed.

Legality, data quality, and semantic coherence appear to be of little concern to those responsible for national security applications. These organizations, by their nature, tend to assume that the risk of unjustified but potentially serious impacts on individuals is of little consequence when compared with the (claimed) potential to avert (what are asserted to be) sufficiently probable major calamities. The same justifications do not apply to social control applications in areas such as tax and welfare fraud, or to commercial uses of large-scale data assemblies, but organizations in both the private and public sectors have exploited the gray edges between national security intelligence and other applications to achieve a default presumption that the ends in these cases justify the means adopted to achieve them.

Once the legal, data quality, and semantic issues have been resolved (or more commonly assumed away), it is possible to use a wide array of available algorithms—or invent new ones—to draw inferences from the amassed data. In scientific fields, those inferences are commonly generalizations. Managerial applications, on the other hand, lean toward using analyses of big data not for generalization but for particularization. The payoff is the discovery of individuals of interest and the customization of activities targeted at specific individuals or categories of individuals.

When generalizing, the statistical inference methods used might justify assuming away data-quality issues, and even ignoring incompatibilities between data items acquired from different sources, at different times, for different purposes. On the other hand, when dealing with particular cases and categories, not confronting these problems undermines decision-making quality and inevitably creates the problems generated by the resulting type two errors.

Individuals often bear the consequences of an organization's low-quality decision making.

Moreover, individuals often bear the consequences of an organization's low-quality decision making. For example, an applicant might be denied a loan or access to a government benefit, or be singled out for attention at a border crossing. Service denial has been increasingly apparent in many contexts, including government licensing, financial services, transport, and even health. In some cases, the user might not even know about the decision, or about the basis on which it was made. Even when individuals are aware of the problem, they often lack the expertise and institutional power to force corrective actions.

In this regard, the combination of sources to a level that permits identification of an individual creates a responsibility within an organization to comply with any relevant privacy legislation in any administrative regions from which the datasets have been drawn. Almost all the big data exploiters in the commercial domain have neglected this simple point.

The techniques applied to big data are a complex mix of pattern matching, Bayesian inference, and other automated deductive algorithms. Consequently, the resulting inferences are difficult to explain in a manner the general public can understand, and many inferences have no straightforward, coherent, logical, or easily explained justification. Before such inferences are used to make decisions and take significant actions, they must be tested empirically.

On the other hand, testing costs money, incurs delays, and can undermine business models. It also lacks the appeal of the apparent immediate emergence of useful information from a huge mass of often disparate data. So organizations tend to assume the truth-value of the

inferences rather than demonstrate it, and judge the outcomes against criteria dreamt up by proponents of the technology or its application. These frameworks rarely regard analytical integrity as having any great significance. An appearance of success can justify the use of data mining, whether or not the outcomes prove effective against an appropriate external measuring stick.

The exploitation of these intensive data collections gives rise to concerns about the legal and logical justification for the activities, quality controls over data management, applicability of the analytical techniques used, and lack of external standards for evaluating results.

BIG DATA CONTEXTS

To show how these issues have steadily emerged, we describe some long-standing instances of big data, and then move on to more recent and still-emergent forms. In some cases, the example involves a relatively coherent dataset, whereas others involve a mélange of sources. All, however, can be integrated with other sources to generate bigger data collections.

Since the turn of the century, and particularly since about 2005, consumers have volunteered volumes of personal data.

Consolidation of government data holdings

Mechanisms to facilitate dataveillance include database consolidation and organizational mergers.[1] The scale of the data involved is challenging, but smaller countries such as Denmark, Finland, and Malaysia have achieved considerable concentration, supported by the imposition of a comprehensive national identification scheme.

Key government agencies in Australia have spent the last quarter of a century working toward the same kind of consolidation. Since 1997, all of the approximately 100 social welfare programs have been funneled through a single operator, Centrelink. In 2011, Centrelink merged with the national health insurance and pharmaceutical benefits schemes operator into the Department of Human Services (DHS). Recently, the federal government moved to bring all Australian health databases within reach of the Department of Health, using an identifier managed by DHS. Agencies in Australia have thereby made a complete mockery of data protection laws, done everything possible to override the public's strongly expressed opposition to a national identification scheme, and enabled cross-agency data consolidation, warehousing, and mining.

In various countries, interactions of an individual with government have increasingly been consolidated onto a single identifier in an attempt to deny the legality of multiple identities, and to destroy the protection that data silos and identity silos once provided.[4,5]

The bureaucratic desire is for a singular identity per person that is outside the individual's control. Some (mostly small) governments have funded schemes that go some way toward achieving this goal: others have tried and failed. Currently, several governments are endeavoring to develop partnerships with financial services institutions to leverage the identity management and authentication schemes that have been imposed on those corporations by counterterrorism. One mechanism has been "know your customer" legislation. Despite Microsoft Passport's failure, governments are also considering whether supranational corporations such as Facebook and Google, with their extensive coverage and "real names" policies, might provide a basis for a less expensive, more publicly acceptable, "good enough" identity-management framework. Furthermore, Facebook's recent alliances with major global commercial data brokers give governments a convenient arm's-length distancing from this commercial conflation.

Consumer profile databases

Consumer-profiling companies have long gathered data, often surreptitiously and often in breach of public expectations and even the laws of countries with strong data protection statutes, which includes almost all of Europe. The US Federal Trade Commission (FTC) announced at the end of 2012 that it will investigate the operations of the shadowy nine "data brokers": Acxiom, CoreLogic, Datalogix, EBureau, ID Analytics, Intelius, Peekyou, Rapleaf, and Recorded Future.

Loyalty cards

Loyalty cards give consumer-marketing corporations access to data trails generated at points of sale far beyond their own cash registers and Web commerce sites.

This has fed into customer relationship management (CRM) systems, which Ngai and colleagues argue are the most significant initial big data application in the commercial sector.[6] Corporations can combine the data derived from these sources with that from the micromonitoring of individual shoppers' movements and actions on retailers' premises and websites. Building on that data, they can manipulate consumer behavior not only through targeted and timed advertising and promotions, but also through dynamic pricing, wherein the price offered does not necessarily benefit the buyer.

Social media

Since the turn of the century, and particularly since about 2005, consumers have volunteered volumes of personal data through social media services. Google has amassed vast quantities of data about users of its search

facilities, and progressively of other services. The company's acquisition, retention, and exploitation of all Gmail traffic have allowed it to build archives of its users and their correspondents' communications.

Since about 2004, users of social networking services and other social media have gifted to a range of corporations, but most substantially Facebook, a huge amount of content that is variously factual, inaccurate, salacious, malicious, and sheer fantasy. Users understood that they were paying for the services by accepting advertisements in their browser windows, but very few appreciated how extensive the accumulation, use, and disclosure of their data was to become.

Issues arise concerning users' data, including informed consent for use and disclosure, retention (even after the account is closed), access, and the adequacy of the consideration provided. Much of the data, however, is also about the users' colleagues, friends, and others they come into contact with. Social media providers are gathering and exploiting vast amounts of personal data, without quality controls and without the consent of the individuals to whom that data relates. Individuals who volunteer such data have moral responsibility for their actions, but little or no legal responsibility. In some social media systems, such as Facebook, biometric and other connection linkages can be set up without the knowledge or permission of the individuals affected, typically by tagging people in photographs. Service providers, meanwhile, can use obscurity, data havens, jurisdictional arbitrage, and market power to escape data protection laws.

As any new market structure matures, consolidation occurs. The transaction-based content, trails, and social networks generated by social media corporations complements the decades of behind-the-scenes work by consumer profiling corporations. Mergers of old and new databases are inevitable—and the United States puts few legal constraints on corporate exploitation of and trafficking in personal data. Such mergers will likely occur as cash-rich Internet companies take over key profiling companies in the same way they have taken over key players in parallel markets. Just as Microsoft saw advantage in acquiring Skype, Acxiom is a natural target for Google.

Analysts have documented examples of new kinds of inferences that can be drawn from this vast volume of data, along the lines of "your social media service knows you're pregnant before your father does." Such inferences arise from the application of predictive analytics developed in loyalty contexts,[7] but become much more threatening when they move beyond a specific consumer-supplier relationship.

To marketers, this social media data is a treasure trove, both on its own and even more so when linked to other sources. To individuals, it's a morass of hidden knowledge whose exposure will have some seriously negative consequences. Some harmful inferences will arise from what careful analysis could reveal as false matches. In other cases, ambiguities will provide fertile ground for speculation, innuendo, and the exercise of preexisting biases for, and particularly against, racial, ethnic, religious, and socioeconomic stereotypes.

Sensor data

The flows of data generated by various kinds of sensors are rapidly becoming an avalanche. RFID tags have extended beyond the industry value chain not only in packaging, but also in consumer items, notably clothing. RFID has also been applied to public transport ticketing and road toll payment mechanisms. The use of RFID tags in books was a particularly chilling development, as it extends surveillance far beyond mere consumption behavior toward social and political choices, attitudes, and even values.

RFID product tags are not inherently associated with individuals, but they can become so in various ways. The rich trail associated with a commonly carried item, such as a purse or wallet, can render superfluous a name and address or a company ID code. Meanwhile, many applications of RFID in transport include user identification in their design, in some cases by requiring the person's identity as a condition of discounted purchase, and in others by ensuring that payment is made at least once by inherently identified means such as credit and debit cards. RFID tags in clothing let manufacturers, wholesalers, and retailers track clothing and the individuals carrying it within the store. These trails can be associated with the individual through loyalty cards or in-store video. Elsewhere, intelligent transport systems (ITS) developments have given rise to car and in-car monitoring. This generates trails that are closely associated with individuals and are available to various organizations.

Mergers of old and new databases are inevitable—and the United States puts few legal constraints on corporate exploitation of and trafficking in personal data.

These issues were debated when RFID-based smart passports were introduced, but the international bodies promoting and agreeing to their use only conditionally acknowledged these issues because the international agreements were restricted to border crossing, the domain of the international forums involved in negotiating passport agreements. The threats involved have penetrated far enough into public consciousness that wallets providing shielding of RFID chips are now readily procurable.

Some forms of visual surveillance also give rise to data that can be associated with one or more individuals. Crash cameras in cars, for example, could be imposed as a

condition of purchase, insurance, or rental. Like so many other data trails, the data can be used for more than originally intended (accident investigation), and with or without informed, freely given, and granular consent. In the UK and some other countries, automated number plate recognition (ANPR) has exceeded its nominal purpose of traffic management to provide vast mass transport surveillance databases.

Devices that use cellular and Wi-Fi networks are locatable not merely within a cell, but within a small area within that cell, by various means. The device's disclosure of its cell location is intrinsic to network operation; but networks can deliver much more precise positional data, extraneous to network operations and intended to add value—in some cases for the individual, but in all cases for other parties. Devices and apps, meanwhile, are designed to be promiscuous with location data, mostly without effective consent.

Devices and apps are designed to be promiscuous with location data.

Smartphones, tablets, and other mobile devices can not only be located with considerable precision—with or without the user's knowledge and meaningful consent—but also accurately tracked, in real time.[8] This has implications not only for individuals' ability to exercise self-determination, but also for their physical safety.

In less than a decade, the explosion in smartphone usage has resulted in almost the entire population in many countries having been recruited as unpaid, high-volume suppliers of highly detailed data about their locations and activities. This data is highly personal even before it's combined with loyalty card data, marketers' various sources of consumer data, and the locations and activities of other people.

Smart meters

In many respects, the Internet of Things is still emerging, although some elements, such as energy consumption monitoring, have arrived.

Smart meters essentially give energy providers sensitive, time-based data about consumers, their activities, and their presence or absence from their premises. In accordance with the "warm frog" principle, whereby a frog is placed in a pot and the temperature is increased incrementally until finally the frog is cooked, monitoring has thus far been infrequent, and providers have in general not yet exploited their capacity to act based on the data, or to sell it. However, most designs support highly intensive monitoring and let providers directly control and discontinue power supply to the home and even to individual devices. This results in a mix of detailed usage data and control over power access, creating a new form of natural monopoly that is attractive to investors, especially as the power distributors assert that they, and not their customers, own all the smart meter data, however detailed in profile and exact usage levels of devices (and potentially real time in coverage).[9]

Aerial surveillance

Satellite imagery delivers volumes of raw material for big data operators. At higher resolutions, satellites disclose a substantial amount of personal data. For example, local government agencies can and do use satellite imagery to find unregistered backyard swimming pools.

Aerial surveillance from lower altitudes used to be sufficiently expensive to restrict its application to activities with high economic value or a military purpose. Costs have dropped significantly in the last decade. Drones have migrated beyond military contexts, and unmanned aerial vehicles (UAVs) have been democratized. Carrying high-resolution video, and controlled by smartphones, UAVs are now sufficiently inexpensive that individuals, businesses, and government agencies can deploy them for unobtrusive data collection.

Aircraft licensing and movement regulators have not yet resolved important operational aspects of drones, but meanwhile do not appear to be interfering in their use. Parliaments and regulatory agencies almost everywhere have failed their responsibility to impose reasonable privacy constraints on long-standing, fixed closed-circuit TV and open-circuit TV. As a result, the new, drone-borne mobile CCTV and OCTV cameras are operating largely free of regulation.

DATA OWNERSHIP, CONTROL, AND RIGHTS

Analyses of big data economics often refer to "data ownership." However, data is not real estate, and hence property law does not apply. Nor is it a tangible object to which the law of chattels applies. Under specific circumstances, data can be considered intellectual property. Patent, copyright, trademark, and similar laws attempt to encourage innovation by letting corporations not merely recover costs but make (often very substantial) profits by exercising their monopoly powers and restricting their competitors' activities. However, these rights are not applicable to the kinds of data we focus on here. Ownership might be a relevant concept in specific contexts, but as a general analytical tool, current notions of property in data have little value.[9]

In the personal data arena, data possession and data control are more common and effective notions. These notions recognize that often multiple parties have, or have access to, copies of particular data, multiple parties have an interest in it, and multiple parties might have some form of right to it. The ability to transfer or link to data, with or

without a price, is a relevant form of "ownership" in terms of the capacity to monetize its possession.

Big data aggregators typically assume they have rights to the data, or at least to the data collection as a whole. They claim at least the rights to possess it or have access to it, to analyze it, and to exploit results arising from their analyses. They might claim the right to disclose parts of the data, share or rent access to it, or sell copies of some or all of it. Other organizations might claim conflicting rights. In the cases of asset liquidations, company failures, and takeovers, big data assemblies represent a valuable asset whose value the seller will naturally maximize. Any existing privacy protections are unlikely to survive the asset's sale or a bankruptcy, as any data held is often regarded as an asset for cost recovery in a liquidation.

When data directly or indirectly identifies an individual, that individual can claim rights to it. In many countries, human rights instruments, such as statutes or even constitutions, support these claims. It is a poor reflection on the rule of law in these countries when highly uncertain claims of rights by government agencies and corporations receive greater protection than the much clearer claims of individuals.

Tensions among interests in personal data have always existed. A useful test case is the public health interest in, for example, reports of highly contagious diseases such as bubonic plague, which most people agree outweigh an individual's interest in suppressing the data. The public health interest has been generalized far beyond the public health issue of contagious diseases, however. Cancer registries contain rich collections of sensitive socioeconomic and health data, on the partly reasonable and partly spurious basis that rich datasets are essential to cancer research. The same justifications are being used to override the interests of individuals in their genetic data—with little public debate and few mitigating measures.

Big data proponents are keen to develop vast warehouses of personal data. They prefer to do so unhampered by public debate, let alone government policy, soft regulation, or legal constraints. In expropriating the data for corporate benefit, the state, or the "common good," they are implicitly withdrawing Western civilization from the centuries-old dominance of individualism back to a time when governing bodies and elites fostered a sense of collectivism as a convenient means of achieving hegemony over an uneducated and largely powerless population. In some regions (for example, East Asia) groups as diverse as religious fundamentalists, environmentalists, medical researchers, and consumer marketing corporations are working to subjugate individual rights in favor of corporate and state rights.

Storage capacities have grown exponentially and dropped in cost. In addition, organizations that are distant from the individuals they deal with depend on detailed data holdings about them. Consequently, organizations increasingly tend to retain all data indefinitely. This tendency clashes substantially with some important social needs.

Personal data is potentially sensitive. In the past, few people would know of an individual's youthful indiscretions. Today, however, such indiscretions are increasingly recorded and broadcast over space and time. Criminal justice systems are designed to not broadcast information about minor offenses, and in many countries criminal records systems actively omit old offenses when performing criminal records checks. The system thus favors forgiveness and rehabilitation, rather than permanently labeling people as criminals.

Indiscriminate data retention conflicts with such constructive social processes, and big data's expropriation of datasets greatly exacerbates the problem. A further concern arises from the inherent insecurity of data, particularly when it exists in many copies, and when it has been consolidated into "honey pots" of potential value to many organizations.

Big data aggregators typically assume they have rights to the data, or at least to the data collection as a whole.

Calls are emerging for teaching technology to forget, and for giving individuals the legal right to enforce deletion of data (the "right to be forgotten" in Europe). In countries where disadvantaged socioeconomic groups include indigenous peoples, this right is critical. Policy-making bodies and governments are not recognizing the social costs resulting from the growing scale of data and its detailed intensity generally, and in the case of really "big" data in particular. The current data breach epidemic reflects both carelessness and active measures to gain unauthorized access to data collections.

Organizations that apply big data methods currently do not need to consider the economic costs to individuals or the broader social costs, as these costs are externalities for them, and will remain so unless a revised legal framework changes this. In the same way that coal-fired electricity generators and other highly polluting industrial activities are being forced to confront and mitigate their negative impacts, big data operators must also be denied a free ride in this transactional space.

UNINTENDED CONSEQUENCES

Big data's marketing message and mythology stress the extraction of new generalities that have social and economic value. In the commercial arena, the archetypal (but apparently apocryphal) example is the discovery of hitherto unknown market segments, such as men driving

home from work and stopping at the supermarket to buy "diapers and beer." Each submarket for big data services has spawned its own pseudo-examples of the brave new world that the techniques allegedly lead to. The application of the new generalities discovered through big data affects individuals. Some impacts are to their benefit, while others are to their detriment. New forms of unfair discrimination—some financial, some social—could also arise.[10,11]

However, big data is not just about the extraction of generalities. Many datasets contain explicit identifiers for individuals, such as name and birthdate, or a unique code issued by a government agency or corporation. Even when no formal identifier exists, the richness of the data collection is such that an analyst can draw a reasonably reliable inference, and hence the data is re-identifiable.

Big data aims to improve efficiencies in social control and marketing.

Various studies have shown that very little data is needed to re-identify individuals, even in putatively anonymized datasets.[5] Claims about big data anonymity are highly contestable, even spurious.

A considerable amount of big data effort exploits this identified data, and is about particularization, not generalization. In this case, the impacts on each individual are not just because inferences have been drawn about a category that they statistically fall within. The inferences are being drawn about them in particular, on the basis of a mélange of data from multiple sources that was assembled without their consent through the exercise of market or institutional power, that is at least partially internally incompatible, and that might include data spuriously associated with them.

Big data aims to improve efficiencies in social control and marketing, but many unintended consequences have arisen. Organizations manipulate consumer behavior by inferring individuals' interests from the big data accumulated about them, outside their control. They are denying consumer choice through the inference-based narrowcasting of marketing information. Social control agencies are unjustifiably targeting individuals because they fit an obscure model of infraction, even though they have little understanding of why the individual has been singled out, and hide behind vague security justifications to deny the individual access to their automated accuser.[12]

Decision making comes to be based on difficult-to-comprehend and low-quality data that is nonetheless treated as authoritative. Consequences include unclear accusations, unknown accusers, inversion of the onus of proof, and hence denial of due process. Further extension includes ex-ante discrimination and guilt prediction, and a prevailing climate of suspicion.[1] Franz Kafka could paint the picture, but could not foresee the specifics of the enabling technology.

THE PRIVATE DATA COMMONS

Until the mid-20th century, information about individuals was shared locally. In villages, there were few secrets, but there was also little trafficking in information beyond that village. Urbanization separated the locations of work, play, and sleep, and enabled anonymity within the crowd and multiple identities in different contexts. Information about the individual continued to be shared, but on a more compartmentalized basis than in villages. Individuals' health information was shared with medical professionals. Their financial information was known to the organizations they deposited funds with and borrowed from. Information about their family was shared among the family and with trusted friends. In the city context, the information was not localized as it was in villages, but no one confidante had access to all of it.

The terms that we propose to describe these issues, "private data commons" and "community data commons," convey the key characteristics of these phenomena. The information was private in that it went no further than the people the individual shared it with. Each of these closed communities treated the information as a commons. A person's accountants and solicitors exploited the information within its limited context, but controlled its use to protect the individual's interests. For a few decades, some telephone switchboard operators might have intercepted and passed on interesting snippets, but there was little systematic leakage of community information, and seldom to organizations that took commercial advantage of it.

By the middle of the 20th century, financial services organizations had grown much larger, and operated over much wider geographical areas than they had in the past. Organizations progressively converted information that they had stored in a relatively informal and localized manner into structured data, and shifted its storage to a central location, distant from the individual and community. Computing and then telecommunications accelerated this change. Bigger government accompanied the growth in transfer payments and social welfare programs, adopting structured approaches to data and centralized approaches to its storage. Government agencies expropriated some categories of medical data at any early stage—for example, communicable-disease- and cancer-related information.

During the early 21st century, healthcare data more generally has been migrating from local storage to regional and national collections. Meanwhile, personal computing applications that replicate local control are being replaced by cloud services for which "default is social." People's photo albums have not only been digitized, but also opened to the world; their address books, diaries,

scribbled notes, and sotto voce asides have been converted from private material to public property.

The reduction of nuanced information to structured data and the centralization of previously localized information have furthered the assault on the private data commons. The data can be expropriated for new purposes, disclosed to further parties remote from the individual and their communities, and replicated within new accumulations of data that are not merely remote from the individual, but also unknown to them, and that "know" them not as human beings, but as database entries.

The growth and exploitation of big data directly affects democratic governance as well. The transparency of individual behavior to employers, suppliers, and social control agencies chills not only criminal and antisocial behavior, but also creative behavior and economically and technologically innovative activities. Western nations, through the big data epidemic, are risking a stasis as grinding as that experienced in post-WWII East Germany.

As the volumes of data grow, and the Internet of Things takes hold, universal surveillance is graduating from a paranoid delusion to a practicable proposition. The survival of free societies depends on individuals' rights in relation to the data being asserted, and the interests of big data proponents being subjected to tight controls.

A regulatory framework for big data is essential. That framework must be constructed with a clear understanding of the ravages that have been wrought on personal interests by the reduction of information to data, its centralization, and its expropriation. Individuals need protections that recover the benefits to individualism that the private data commons afforded.

References

1. R. Clarke, "Information Technology and Dataveillance," *Comm. ACM*, vol. 31, no. 5, 1988, pp. 498-512.
2. B. Ratner, *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data,* CRC Press, 2003.
3. T. Craig and M.E. Ludlof, *Privacy and Big Data: The Players, Regulators, and Stakeholders*, O'Reilly Media, 2011.
4. R. Clarke, "Human Identification in Information Systems: Management Challenges and Public Policy Issues," *Information Technology & People*, vol. 7, no. 4, 1994, pp. 6-37; www.rogerclarke.com/DV/HumanID.html.
5. M.R. Wigan, "Owning Identity—One or Many—Do We Have a Choice?" *IEEE Technology and Society*, vol. 29, no. 2, 2010, pp. 33-38.
6. E.W.T. Ngai, L. Xiu, and D.C.K. Chau, "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification," *Expert Systems with Applications*, vol. 36, no. 2, 2009, pp. 2592-2602.
7. C. Duhigg, "How Companies Learn Your Secrets," *The New York Times*, 16 Feb. 2012; www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=2&hp&.
8. K. Michael and R. Clarke, "Location and Tracking of Mobile Devices: Überveillance Stalks the Streets," to appear in *Computer Law & Security Rev.*, vol. 29, no. 2; http://works.bepress.com/cgi/viewcontent.cgi?article=1396&context=kmichael.
9. M.R. Wigan, "Smart Meter Technology Tradeoffs," *Proc. IEEE Int'l Symp. Technology and Society in Asia* (ISTAS), IEEE, 2012, pp. 1-8.
10. d. boyd and K. Crawford, "Six Provocations for Big Data," *A Decade in Internet Time: Proc. Symp. Dynamics of the Internet and Society*, Oxford Internet Inst., Sept. 2011; http://ssrn.com/abstract=1926431.
11. A. Croll, "Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It: What the Data Is Must Be Linked to How It Can Be Used," *O'Reilly Radar*, 2012; http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html#more-50087.
12. A. Oboler, K. Welsh, and L. Cruz, "The Danger of Big Data: Social Media as Computational Social Science," *First Monday*, vol. 17, no. 7, 2012; http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3993/3269.

Marcus R. Wigan *is the principal of Oxford Systematics, Melbourne, Australia, and adjunct professor of information and communications technology at Swinburne University in Melbourne. He is also a partner in the GAMUT Faculty of Architecture Building and Planning and a professorial fellow in the Melbourne Sustainable Society Institute, both at the University of Melbourne; and vice chair of Electronic Frontiers Australia. His research interests include transport policy, modeling, data and planning, freight and logistic systems, privacy and surveillance, and issues of mobility and accessibility. Wigan has a doctorate in nuclear physics from Oxford University. He is a Fellow of the Australian Computer Society, the UK Institute of Physics, the UK Institution of Civil Engineers, the Institution of Engineers Australia, the Chartered Institute of Logistics and Transport, and the Australian Institute of Traffic Planning and Management. Contact him at www.mwigan.com.*

Roger Clarke *is the principal of the Xamax Consultancy in Canberra, Australia; visiting professor in cyberspace law and policy at the University of New South Wales; and visiting professor in computer science at the Australian National University. He is also chair of the Australian Privacy Foundation and a member of the advisory committee of Privacy International. His research interests include privacy and dataveillance. Clarke has a PhD in information systems from the Australian National University. He is a Fellow of the Australian Computer Society. Contact him at roger.clarke@xamax.com.au.*