REPORT TO THE PRESIDENT

BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE

Executive Office of the President

President's Council of Advisors on Science and Technology

May 2014

REPORT TO THE PRESIDENT

BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE

Executive Office of the President

President's Council of Advisors on Science and Technology

May 2014

About the President's Council of Advisors on Science and Technology

The President's Council of Advisors on Science and Technology (PCAST) is an advisory group of the Nation's leading scientists and engineers, appointed by the President to augment the science and technology advice available to him from inside the White House and from cabinet departments and other Federal agencies. PCAST is consulted about, and often makes policy recommendations concerning, the full range of issues where understandings from the domains of science, technology, and innovation bear potentially on the policy choices before the President.

For more information about PCAST, see www.whitehouse.gov/ostp/pcast

The President's Council of Advisors on Science and Technology

Co-Chairs

John P. Holdren
Assistant to the President for Science and Technology
Director, Office of Science and Technology Policy

Eric S. Lander
President
Broad Institute of Harvard and MIT

Vice Chairs

William Press
Raymer Professor in Computer Science and Integrative Biology
University of Texas at Austin

Maxine Savitz
Vice President
National Academy of Engineering

Members

Rosina Bierbaum
Dean, School of Natural Resources and Environment
University of Michigan

Christine Cassel
President and CEO
National Quality Forum

Christopher Chyba
Professor, Astrophysical Sciences and International Affairs
Director, Program on Science and Global Security
Princeton University

S. James Gates, Jr.
John S. Toll Professor of Physics
Director, Center for String and Particle Theory
University of Maryland, College Park

Mark Gorenberg
Managing Member
Zetta Venture Partners

Susan L. Graham
Pehong Chen Distinguished Professor Emerita in Electrical Engineering and Computer Science
University of California, Berkeley

Shirley Ann Jackson
President
Rensselaer Polytechnic Institute

Richard C. Levin (*through mid-April 2014*)
President Emeritus
Frederick William Beinecke Professor of Economics
Yale University

Michael McQuade
Senior Vice President for Science and Technology
United Technologies Corporation

Chad Mirkin
George B. Rathmann Professor of Chemistry
Director, International Institute for Nanotechnology
Northwestern University

Mario Molina
Distinguished Professor, Chemistry and Biochemistry
University of California, San Diego
Professor, Center for Atmospheric Sciences at the Scripps Institution of Oceanography

Craig Mundie
Senior Advisor to the CEO
Microsoft Corporation

Ed Penhoet
Director, Alta Partners
Professor Emeritus, Biochemistry and Public Health
University of California, Berkeley

Barbara Schaal
Mary-Dell Chilton Distinguished Professor of Biology
Washington University, St. Louis

Eric Schmidt
Executive Chairman
Google, Inc.

Daniel Schrag
Sturgis Hooper Professor of Geology
Professor, Environmental Science and Engineering
Director, Harvard University Center for Environment
Harvard University

Staff

Marjory S. Blumenthal
Executive Director

Ashley Predith
Assistant Executive Director

Knatokie Ford
AAAS Science & Technology Policy Fellow

PCAST Big Data and Privacy Working Group

Working Group Co-Chairs

Susan L. Graham
Pehong Chen Distinguished Professor Emerita in Electrical Engineering and Computer Science
University of California, Berkeley

William Press
Raymer Professor in Computer Science and Integrative Biology
University of Texas at Austin

Working Group Members

S. James Gates, Jr.
John S. Toll Professor of Physics
Director, Center for String and Particle Theory
University of Maryland, College Park

Mark Gorenberg
Managing Member
Zetta Venture Partners

John P. Holdren
Assistant to the President for Science and Technology
Director, Office of Science and Technology Policy

Eric S. Lander
President
Broad Institute of Harvard and MIT

Craig Mundie
Senior Advisor to the CEO
Microsoft Corporation

Maxine Savitz
Vice President
National Academy of Engineering

Eric Schmidt
Executive Chairman
Google, Inc.

Working Group Staff

Marjory S. Blumenthal
Executive Director
President's Council of Advisors on Science and Technology

Michael Johnson
Assistant Director
National Security and International Affairs

EXECUTIVE OFFICE OF THE PRESIDENT
PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY
WASHINGTON, D.C. 20502

President Barack Obama
The White House
Washington, DC 20502

Dear Mr. President,

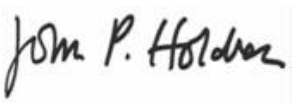
We are pleased to send you this report, *Big Data and Privacy: A Technological Perspective*, prepared for you by the President's Council of Advisors on Science and Technology (PCAST). It was developed to complement and inform the analysis of big-data implications for policy led by your Counselor, John Podesta, in response to your requests of January 17, 2014. PCAST examined the nature of current technologies for managing and analyzing big data and for preserving privacy, it considered how those technologies are evolving, and it explained what the technological capabilities and trends imply for the design and enforcement of public policy intended to protect privacy in big-data contexts.

Big data drives big benefits, from innovative businesses to new ways to treat diseases. The challenges to privacy arise because technologies collect so much data (e.g., from sensors in everything from phones to parking lots) and analyze them so efficiently (e.g., through data mining and other kinds of analytics) that it is possible to learn far more than most people had anticipated or can anticipate given continuing progress. These challenges are compounded by limitations on traditional technologies used to protect privacy (such as de-identification). PCAST concludes that technology alone cannot protect privacy, and policy intended to protect privacy needs to reflect what is (and is not) technologically feasible.

In light of the continuing proliferation of ways to collect and use information about people, PCAST recommends that policy focus primarily on whether specific *uses* of information about people affect privacy adversely. It also recommends that policy focus on outcomes, on the "what" rather than the "how," to avoid becoming obsolete as technology advances. The policy framework should accelerate the development and commercialization of technologies that can help to contain adverse impacts on privacy, including research into new technological options. By using technology more effectively, the Nation can lead internationally in making the most of big data's benefits while limiting the concerns it poses for privacy. Finally, PCAST calls for efforts to assure that there is enough talent available with the expertise needed to develop and use big data in a privacy-sensitive way.

PCAST is grateful for the opportunity to serve you and the country in this way and hope that you and others who read this report find our analysis useful.

Best regards,



John P. Holdren
Co-chair, PCAST

Eric S. Lander
Co-chair, PCAST

Table of Contents

Executive Summary

The ubiquity of computing and electronic communication technologies has led to the exponential growth of data from both digital and analog sources. New capabilities to gather, analyze, disseminate, and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected.

After providing an overview of this report and its origins, Chapter 1 describes the changing nature of privacy as computing technology has advanced and big data has come to the fore. The term privacy encompasses not only the famous "right to be left alone," or keeping one's personal matters and relationships secret, but also the ability to share information selectively but not publicly. Anonymity overlaps with privacy, but the two are not identical. Likewise, the ability to make intimate personal decisions without government interference is considered to be a privacy right, as is protection from discrimination on the basis of certain personal characteristics (such as race, gender, or genome). Privacy is not just about secrets.

Conflicts between privacy and new technology have occurred throughout American history. Concern with the rise of mass media such as newspapers in the 19th century led to legal protections against the harms or adverse consequences of "intrusion upon seclusion," public disclosure of private facts, and unauthorized use of name or likeness in commerce. Wire and radio communications led to 20th century laws against wiretapping and the interception of private communications – laws that, PCAST notes, have not always kept pace with the technological realities of today's digital communications.

Past conflicts between privacy and new technology have generally related to what is now termed "small data," the collection and use of data sets by private- and public-sector organizations where the data are disseminated in their original form or analyzed by conventional statistical methods. Today's concerns about big data reflect both the substantial increases in the amount of data being collected and associated changes, both actual and potential, in how they are used.

Big data is big in two different senses. It is big in the quantity and variety of data that are available to be processed. And, it is big in the scale of analysis (termed "analytics") that can be applied to those data, ultimately to make inferences and draw conclusions. By data mining and other kinds of analytics, non-obvious and sometimes private information can be derived from data that, at the time of their collection, seemed to raise no, or only manageable, privacy issues. Such new information, used appropriately, may often bring benefits to individuals and society – Chapter 2 of this report gives many such examples, and additional examples are scattered throughout the rest of the text. Even in principle, however, one can never know what information may later be extracted from any particular collection of big data, both because that information may result only from the combination of seemingly unrelated data sets, and because the algorithm for revealing the new information may not even have been invented at the time of collection.

The same data and analytics that provide benefits to individuals and society if used appropriately can also create potential harms – threats to individual privacy according to privacy norms both widely

shared and personal. For example, large-scale analysis of research on disease, together with health data from electronic medical records and genomic information, might lead to better and timelier treatment for individuals but also to inappropriate disqualification for insurance or jobs. GPS tracking of individuals might lead to better community-based public transportation facilities, but also to inappropriate use of the whereabouts of individuals. A list of the kinds of adverse consequences or harms from which individuals should be protected is proposed in Section 1.4. PCAST believes strongly that the positive benefits of big-data technology are (or can be) greater than any new harms.

Chapter 3 of the report describes the many new ways in which personal data are acquired, both from original sources, and through subsequent processing. Today, although they may not be aware of it, individuals constantly emit into the environment information whose use or misuse may be a source of privacy concerns. Physically, these information emanations are of two types, which can be called "born digital" and "born analog."

When information is "born digital," it is created, by us or by a computer surrogate, specifically for use by a computer or data processing system. When data are born digital, privacy concerns can arise from over-collection. Over-collection occurs when a program's design intentionally, and sometimes clandestinely, collects information unrelated to its stated purpose. Over-collection can, in principle, be recognized at the time of collection.

When information is "born analog," it arises from the characteristics of the physical world. Such information becomes accessible electronically when it impinges on a sensor such as a camera, microphone, or other engineered device. When data are born analog, they are likely to contain more information than the minimum necessary for their immediate purpose, and for valid reasons. One reason is for robustness of the desired "signal" in the presence of variable "noise." Another is technological convergence, the increasing use of standardized components (e.g., cell-phone cameras) in new products (e.g., home alarm systems capable of responding to gesture).

Data fusion occurs when data from different sources are brought into contact and new facts emerge (see Section 3.2.2). Individually, each data source may have a specific, limited purpose. Their combination, however, may uncover new meanings. In particular, data fusion can result in the identification of individual people, the creation of profiles of an individual, and the tracking of an individual's activities. More broadly, data analytics discovers patterns and correlations in large corpuses of data, using increasingly powerful statistical algorithms. If those data include personal data, the inferences flowing from data analytics may then be mapped back to inferences, both certain and uncertain, about individuals.

Because of data fusion, privacy concerns may not necessarily be recognizable in born-digital data when they are collected. Because of signal-processing robustness and standardization, the same is true of born-analog data – even data from a single source (e.g., a single security camera). Born-digital and born-analog data can both be combined with data fusion, and new kinds of data can be generated from data analytics. The beneficial uses of near-ubiquitous data collection are large, and they fuel an increasingly important set of economic activities. Taken together, these considerations suggest that a policy focus on limiting data collection will not be a broadly applicable or scalable strategy – nor one

likely to achieve the right balance between beneficial results and unintended negative consequences (such as inhibiting economic growth).

If collection cannot, in most cases, be limited practically, then what? Chapter 4 discusses in detail a number of technologies that have been used in the past for privacy protection, and others that may, to a greater or lesser extent, serve as technology building blocks for future policies.

Some technology building blocks (for example, cybersecurity standards, technologies related to encryption, and formal systems of auditable access control) are already being utilized and need to be encouraged in the marketplace. On the other hand, some techniques for privacy protection that have seemed encouraging in the past are useful as supplementary ways to reduce privacy risk, but do not now seem sufficiently robust to be a dependable basis for privacy protection where big data is concerned. For a variety of reasons, PCAST judges anonymization, data deletion, and distinguishing data from metadata (defined below) to be in this category. The framework of notice and consent is also becoming unworkable as a useful foundation for policy.

Anonymization is increasingly easily defeated by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially. While anonymization may remain somewhat useful as an added safeguard in some situations, approaches that deem it, by itself, a sufficient safeguard need updating.

While it is good business practice that data of all kinds should be deleted when they are no longer of value, economic or social value often can be obtained from applying big data techniques to masses of data that were otherwise considered to be worthless. Similarly, archival data may also be important to future historians, or for later longitudinal analysis by academic researchers and others. As described above, many sources of data contain latent information about individuals, information that can be known only if the holder expends analytic resources, or that may become knowable only in the future with the development of new data-mining algorithms. In such cases it is practically impossible for the data holder even to surface "all the data about an individual," much less delete it on any specified schedule or in response to an individual's request. Today, given the distributed and redundant nature of data storage, it is not even clear that data, even small data, *can* be destroyed with any high degree of assurance.

As data sets become more complex, so do the attached metadata. Metadata are ancillary data that describe properties of the data such as the time the data were created, the device on which they were created, or the destination of a message. Included in the data or metadata may be identifying information of many kinds. It cannot today generally be asserted that metadata raise fewer privacy concerns than data.

Notice and consent is the practice of requiring individuals to give positive consent to the personal data collection practices of each individual app, program, or web service. Only in some fantasy world do users actually read these notices and understand their implications before clicking to indicate their consent.

The conceptual problem with notice and consent is that it fundamentally places the burden of privacy protection on the individual. Notice and consent creates a non-level playing field in the implicit privacy negotiation between provider and user. The provider offers a complex, take-it-or-leave-it set of terms, while the user, in practice, can allocate only a few seconds to evaluating the offer. This is a kind of market failure.

PCAST believes that the responsibility for using personal data in accordance with the user's preferences should rest with the provider rather than with the user. As a practical matter, in the private sector, third parties chosen by the consumer (e.g., consumer-protection organizations, or large app stores) could intermediate: A consumer might choose one of several "privacy protection profiles" offered by the intermediary, which in turn would vet apps against these profiles. By vetting apps, the intermediaries would create a marketplace for the negotiation of community standards for privacy. The Federal government could encourage the development of standards for electronic interfaces between the intermediaries and the app developers and vendors.

After data are collected, data analytics come into play and may generate an increasing fraction of privacy issues. Analysis, per se, does not directly touch the individual (it is neither collection nor, without additional action, use) and may have no external visibility. By contrast, it is the *use* of a product of analysis, whether in commerce, by government, by the press, or by individuals, that can cause adverse consequences to individuals.

More broadly, PCAST believes that it is the use of data (including born-digital or born-analog data and the products of data fusion and analysis) that is the locus where consequences are produced. This locus is the technically most feasible place to protect privacy. Technologies are emerging, both in the research community and in the commercial world, to describe privacy policies, to record the origins (provenance) of data, their access, and their further use by programs, including analytics, and to determine whether those uses conform to privacy policies. Some approaches are already in practical use.

Given the statistical nature of data analytics, there is uncertainty that discovered properties of groups apply to a particular individual in the group. Making incorrect conclusions about individuals may have adverse consequences for them and may affect members of certain groups disproportionately (e.g., the poor, the elderly, or minorities). Among the technical mechanisms that can be incorporated in a use-based approach are methods for imposing standards for data accuracy and integrity and policies for incorporating useable interfaces that allow an individual to correct the record with voluntary additional information.

PCAST's charge for this study did not ask it to recommend specific privacy policies, but rather to make a relative assessment of the technical feasibilities of different broad policy approaches. Chapter 5, accordingly, discusses the implications of current and emerging technologies for government policies for privacy protection. The use of technical measures for enforcing privacy can be stimulated by reputational pressure, but such measures are most effective when there are regulations and laws with civil or criminal penalties. Rules and regulations provide both deterrence of harmful actions and incentives to deploy privacy-protecting technologies. Privacy protection cannot be achieved by technical measures alone.

This discussion leads to five recommendations.

Recommendation 1. Policy attention should focus more on the actual uses of big data and less on its collection and analysis. By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals. In the context of big data, these events ("uses") are almost always actions of a computer program or app interacting either with the raw data or with the fruits of analysis of those data. In this formulation, it is not the data themselves that cause the harm, nor the program itself (absent any data), but the confluence of the two. These "use" events (in commerce, by government, or by individuals) embody the necessary specificity to be the subject of regulation. By contrast, PCAST judges that policies focused on the regulation of data collection, storage, retention, a priori limitations on applications, and analysis (absent identifiable actual uses of the data or products of analysis) are unlikely to yield effective strategies for improving privacy. Such policies would be unlikely to be scalable over time, or to be enforceable by other than severe and economically damaging measures.

Recommendation 2. Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.

To avoid falling behind the technology, it is essential that policy concerning privacy protection should address the purpose (the "what") rather than prescribing the mechanism (the "how").

Recommendation 3. With coordination and encouragement from OSTP,[1] the NITRD agencies[2] should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.

Some of the technology for controlling uses already exists. However, research (and funding for it) is needed in the technologies that help to protect privacy, in the social mechanisms that influence privacy-preserving behavior, and in the legal options that are robust to changes in technology and create appropriate balance among economic opportunity, national priorities, and privacy protection.

Recommendation 4. OSTP, together with the appropriate educational institutions and professional societies, should encourage increased education and training opportunities concerning privacy protection, including career paths for professionals.

Programs that provide education leading to privacy expertise (akin to what is being done for security expertise) are essential and need encouragement. One might envision careers for digital-privacy experts both on the software development side and on the technical management side.

[1] The White House Office of Science and Technology Policy

[2] NITRD refers to the Networking and Information Technology Research and Development program, whose participating Federal agencies support unclassified research in advanced information technologies such as computing, networking, and software and include both research- and mission-focused agencies such as NSF, NIH, NIST, DARPA, NOAA, DOE's Office of Science, and the D0D military-service laboratories (see http://www.nitrd.gov/SUBCOMMITTEE/nitrd_agencies/index.aspx).

Recommendation 5. The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today. It can exhibit leadership both by its convening power (for instance, by promoting the creation and adoption of standards) and also by its own procurement practices (such as its own use of privacy-preserving cloud services).

PCAST is not aware of more effective innovation or strategies being developed abroad; rather, some countries seem inclined to pursue what PCAST believes to be blind alleys. This circumstance offers an opportunity for U.S. technical leadership in privacy in the international arena, an opportunity that should be taken.

1. Introduction

In a widely noted speech on January 17, 2014, President Barack Obama charged his Counselor, John Podesta, with leading a comprehensive review of big data and privacy, one that would "reach out to privacy experts, technologists, and business leaders and look at how the challenges inherent in big data are being confronted by both the public and private sectors; whether we can forge international norms on how to manage this data; and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security."[3] The President and Counselor Podesta asked the President's Council of Advisors on Science and Technology (PCAST) to assist with the technology dimensions of the review.

For this task PCAST's statement of work reads, in part,

> PCAST will study the technological aspects of the intersection of big data with individual privacy, in relation to both the current state and possible future states of the relevant technological capabilities and associated privacy concerns.
>
> Relevant big data include data and metadata collected, or potentially collectable, from or about individuals by entities that include the government, the private sector, and other individuals. It includes both proprietary and open data, and also data about individuals collected incidentally or accidentally in the course of other activities (e.g., environmental monitoring or the "Internet of Things").

This is a tall order, especially on the ambitious timescale requested by the President. The literature and public discussion of big data and privacy are vast, with new ideas and insights generated daily from a variety of constituencies: technologists in industry and academia, privacy and consumer advocates, legal scholars, and journalists (among others). Independently of PCAST, but informing this report, the Podesta study sponsored three public workshops at universities across the country. Limiting this report's charge to technological, not policy, aspects of the problem narrows PCAST's mandate somewhat, but this is a subject where technology and policy are difficult to separate. In any case, it is the nature of the subject that this report must be regarded as based on a momentary snapshot of the technology, although we believe the key conclusions and recommendations have lasting value.

1.1 Context and outline of this report

The ubiquity of computing and electronic communication technologies has led to the exponential growth of online data, from both digital and analog sources. New technological capabilities to create, analyze, and disseminate vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromised or protected.

This report discusses present and future technologies concerning this so-called "big data" as it relates to privacy concerns. It is not a complete summary of the technology concerning big data, nor a complete summary of the ways in which technology affects privacy, but focuses on the ways in which big-data and privacy interact. As an example, if Leslie confides a secret to Chris and Chris broadcasts that secret by email or texting, that might be a

[3] "Remarks by the President on Review of Signals Intelligence," January 17, 2014. http://www.whitehouse.gov/the-press-office/2014/01/17/remarks-president-review-signals-intelligence

privacy-infringing use of information technology, but it is not a big-data issue. As another example, if oceanographic data are collected in large quantities by remote sensing, that is big data, but not, in the first instance, a privacy concern. Some data are more privacy-sensitive than others, for example, personal medical data, as distinct from personal data publicly shared by the same individual. Different technologies and policies will apply to different classes of data.

The notions of big data and the notions of individual privacy used in this report are intentionally broad and inclusive. Business consultants Gartner, Inc. define big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making,"[4] while computer scientists reviewing multiple definitions offer the more technical, "a term describing the storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to, NoSQL, MapReduce, and machine learning."[5] (See Sections 3.2.1 and 3.3.1 for discussion of these technical terms.) In a privacy context, the term "big data" typically means data about one or a group of individuals, or that might be analyzed to make inferences about individuals. It might include data or metadata collected by government, by the private sector, or by individuals. The data and metadata might be proprietary or open, they might be collected intentionally or incidentally or accidentally. They might be text, audio, video, sensor-based, or some combination. They might be data collected directly from some source, or data derived by some process of analysis. They might be saved for a long period of time, or they might be analyzed and discarded as they are streamed. In this report, PCAST usually does not distinguish between "data" and "information."

The term "privacy" encompasses not only avoiding observation, or keeping one's personal matters and relationships secret, but also the ability to share information selectively but not publicly. Anonymity overlaps with privacy, but the two are not identical. Voting is recognized as private, but not anonymous, while authorship of a political tract may be anonymous, but it is not private. Likewise, the ability to make intimate personal decisions without government interference is considered to be a privacy right, as is protection from discrimination on the basis of certain personal characteristics (such as an individual's race, gender, or genome). So, privacy is not just about secrets.

The promise of big-data collection and analysis is that the derived data can be used for purposes that benefit both individuals and society. Threats to privacy stem from the deliberate or inadvertent disclosure of collected or derived individual data, the misuse of the data, and the fact that derived data may be inaccurate or false. The technologies that address the confluence of these issues are the subject of this report.[6]

The remainder of this introductory chapter gives further context in the form of a summary of how the legal concept of privacy developed historically in the United States. Interestingly, and relevant to this report, privacy rights and the development of new technologies have long been intertwined. Today's issues are no exception.

Chapter 2 of this report is devoted to scenarios and examples, some from today, but most anticipating a near tomorrow. Yogi Berra's much-quoted remark – "It's tough to make predictions, especially about the future" – is

[4] Gartner, Inc., "IT Glossary." https://www.gartner.com/it-glossary/big-data/

[5] Barker, Adam and Jonathan Stuart Ward, "Undefined By Data: A Survey of Big Data Definitions," arXiv:1309.5821. http://arxiv.org/abs/1309.5821

[6] PCAST acknowledges gratefully the assistance of several contributors at the National Science Foundation, who helped to identify and distill key insights from the technical literature and research community, as well as other technical experts in academia and industry that it consulted during this project. See Appendix A.

germane. But it is equally true for this subject that policies based on out-of-date examples and scenarios are doomed to failure. Big-data technologies are advancing so rapidly that predictions about the future, however imperfect, must guide today's policy development.

Chapter 3 examines the technology dimensions of the two great pillars of big data: collection and analysis. In a certain sense big data is exactly the confluence of these two: big collection meets big analysis (often termed "analytics"). The technical infrastructure of large-scale networking and computing that enables "big" is also discussed.

Chapter 4 looks at technologies and strategies for the protection of privacy. Although technology may be part of the problem, it must also be part of the solution. Many current and foreseeable technologies can enhance privacy, and there are many additional promising avenues of research.

Chapter 5, drawing on the previous chapters, contains PCAST's perspectives and conclusions. While it is not within this report's charge to recommend specific policies, it is clear that certain kinds of policies are technically more feasible and less likely to be rendered irrelevant or unworkable by new technologies than others. These approaches are highlighted, along with comments on the technical deficiencies of some other approaches. This chapter also contains PCAST's recommendations in areas that lie within our charge, that is, other than policy.

1.2 Technology has long driven the meaning of privacy

The conflict between privacy and new technology is not new, except perhaps now in its greater scope, degree of intimacy, and pervasiveness. For more than two centuries, values and expectations relating to privacy have been continually reinterpreted and rearticulated in light of the impact of new technologies.

The nationwide postal system advocated by Benjamin Franklin and established in 1775 was a new technology designed to promote interstate commerce. But mail was routinely and opportunistically opened in transit until Congress made this action illegal in 1782. While the Constitution's Fourth Amendment codified the heightened privacy protection afforded to people in their homes or on their persons (previously principles of British common law), it took another century of technological challenges to expand the concept of privacy rights into more abstract spaces, including the electronic. The invention of the telegraph and, later, telephone created new tensions that were slow to be resolved. A bill to protect the privacy of telegrams, introduced in Congress in 1880, was never passed.[7]

It was not telecommunications, however, but the invention of the portable, consumer-operable camera (soon known as the Kodak) that gave impetus to Warren and Brandeis's 1890 article "The Right to Privacy,"[8] then a controversial title, but now viewed as the foundational document for modern privacy law. In the article, Warren and Brandeis gave voice to the concern that "[i]nstantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that 'what is whispered in the closet shall be proclaimed from the house-tops,'" further noting that "[f]or years there has been a feeling that the law must afford some remedy for the unauthorized circulation of portraits of private persons…"[9]

[7] Seipp, David J., *The Right to Privacy in American History*, Harvard University, Program on Information Resources Policy, Cambridge, MA, 1978.

[8] Warren, Samuel D. and Louis D. Brandeis, "The Right to Privacy." *Harvard Law Review* 4:5, 193, December 15, 1890.

[9] Id. at 195.

Warren and Brandeis sought to articulate the right of privacy between individuals (whose foundation lies in civil tort law). Today, many states recognize a number of privacy-related harms as causes for civil or criminal legal action (further discussed in Section 1.4).[10]

From Warren and Brandeis' "right to privacy," it took another 75 years for the Supreme Court to find, in *Griswold v. Connecticut*[11] (1965), a right to privacy in the "penumbras" and "emanations" of other constitutional protections (as Justice William O. Douglas put it, writing for the majority).[12] With a broad perspective, scholars today recognize a number of different legal meanings for "privacy." Five of these seem particularly relevant to this PCAST report:

(1) The individual's right to keep secrets or seek seclusion (the famous "right to be left alone" of Brandeis' 1928 dissenting opinion in *Olmstead v. United States*).[13]
(2) The right to anonymous expression, especially (but not only) in political speech (as in *McIntyre v. Ohio Elections Commission*[14])
(3) The ability to control access by others to personal information after it leaves one's exclusive possession (for example, as articulated in the FTC's Fair Information Practice Principles).[15]
(4) The barring of some kinds of negative consequences from the use of an individual's personal information (for example, job discrimination on the basis of personal DNA, forbidden in 2008 by the Genetic Information Nondiscrimination Act[16]).
(5) The right of the individual to make intimate decisions without government interference, as in the domains of health, reproduction, and sexuality (as in *Griswold*).

These are asserted, not absolute, rights. All are supported, but also circumscribed, by both statute and case law. With the exception of number 5 on the list (a right of "decisional privacy" as distinct from "informational privacy"), all are applicable in varying degrees both to citizen-government interactions and to citizen-citizen interactions. Collisions between new technologies and privacy rights have occurred in all five. A patchwork of state and federal laws have addressed concerns in many sectors, but to date there has not been comprehensive legislation to handle these issues. Collisions between new technologies and privacy rights should be expected to continue to occur.

---

[10] Digital Media Law Project, "Publishing *Personal* and Private Information." http://www.dmlp.org/legal-guide/publishing-personal-and-private-information

[11] *Griswold v. Connecticut*, 381 U.S. 479 (1965).

[12] Id. at 483-84.

[13] *Olmstead v. United States*, 277 U.S. 438 (1928).

[14] *McIntyre v. Ohio Elections Commission*, 514 U.S. 334, 340-41 (1995). The decision reads in part, "Protections for anonymous speech are vital to democratic discourse. Allowing dissenters to shield their identities frees them to express critical minority views . . . Anonymity is a shield from the tyranny of the majority. . . . It thus exemplifies the purpose behind the Bill of Rights and of the First Amendment in particular: to protect unpopular individuals from retaliation . . . at the hand of an intolerant society."

[15] Federal Trade Commission, "Privacy Online: Fair Information Practices in the Electronic Marketplace," May 2000.

[16] Genetic Information Nondiscrimination Act of 2008, PL 110–233, May 21, 2008, 122 Stat 881.

1.3 What is different today?

New collisions between technologies and privacy have become evident, as new technological capabilities have emerged at a rapid pace. It is no longer clear that the five privacy concerns raised above, or their current legal interpretations, are sufficient in the court of public opinion.

Much of the public's concern is with the harm done by the use of personal data, both in isolation or in combination. Controlling access to personal data after they leave one's exclusive possession has been seen historically as a means of controlling potential harm. But today, personal data may never be, or have been, within one's possession – for instance they may be acquired passively from external sources such as public cameras and sensors, or without one's knowledge from public electronic disclosures by others using social media. In addition, personal data may be derived from powerful data analyses (see Section 3.2) whose use and output is unknown to the individual. Those analyses sometimes yield valid conclusions that the individual would not want disclosed. Worse yet, the analyses can produce false positives or false negatives -- information that is a consequence of the analysis but is not true or correct. Furthermore, to a much greater extent than before, the same personal data have both beneficial and harmful uses, depending on the purposes for which and the contexts in which they are used. Information supplied by the individual might be used only to derive other information such as identity or a correlation, after which it is not needed. The derived data, which were never under the individual's control, might then be used either for good or ill.

In the current discourse, some assert that the issues concerning privacy protection are collective as well as individual, particularly in the domain of civil rights – for example, identification of certain individuals at a gathering using facial recognition from videos, and the inference that other individuals at the same gathering, also identified from videos, have similar opinions or behaviors.

Current circumstances also raise issues of how the right to privacy extends to the public square, or to quasi-private gatherings such as parties or classrooms. If the observers in these venues are not just people, but also both visible and invisible recording devices with enormous fidelity and easy paths to electronic promulgation and analysis, does that change the rules?

Also rapidly changing are the distinctions between government and the private sector as potential threats to individual privacy. Government is not just a "giant corporation." It has a monopoly in the use of force; it has no direct competitors who seek market advantage over it and may thus motivate it to correct missteps. Governments have checks and balances, which can contribute to self-imposed limits on what they may do with people's information. Companies decide how they will use such information in the context of such factors as competitive advantages and risks, government regulation, and perceived threats and consequences of lawsuits. It is thus appropriate that there are different sets of constraints on the public and private sectors. But government has a set of authorities – particularly in the areas of law enforcement and national security – that place it in a uniquely powerful position, and therefore the restraints placed on its collection and use of data deserve special attention. Indeed, the need for such attention is heightened because of the increasingly blurry line between public and private data.

While these differences are real, big data is to some extent a leveler of the differences between government and companies. Both governments and companies have potential access to the same sources of data and the same analytic tools. Current rules may allow government to purchase or otherwise obtain data from the private

sector that, in some cases, it could not legally collect itself,[17] or to outsource to the private sector analyses it could not itself legally perform.[18] The possibility of government exercising, without proper safeguards, its own monopoly powers and also having unfettered access to the private information marketplace is unsettling.

What kinds of actions should be forbidden both to government (Federal, state, and local, and including law enforcement) and to the private sector? What kinds should be forbidden to one but not the other? It is unclear whether current legal frameworks are sufficiently robust for today's challenges.

1.4 Values, harms, and rights

As was seen in Sections 1.2 and 1.3, new privacy rights usually do not come into being as academic abstractions. Rather, they arise when technology encroaches on widely shared values. Where there is consensus on values, there can also be consensus on what kinds of harms to individuals may be an affront to those values. Not all such harms may be preventable or remediable by government actions, but, conversely, it is unlikely that government actions will be welcome or effective if they are not grounded to some degree in values that are widely shared.

In the realm of privacy, Warren and Brandeis in 1890[19] (see Section 1.2) began a dialogue about privacy that led to the evolution of the right in academia and the courts, later crystalized by William Prosser as four distinct harms that had come to earn legal protection.[20] A direct result is that, today, many states recognize as causes for legal action the four harms that Prosser enumerated,[21] and which have become (though varying from state to state[22]) privacy "rights." The harms are:

- Intrusion upon seclusion. A person who intentionally intrudes, physically or otherwise (now including electronically), upon the solitude or seclusion of another person or her private affairs or concerns, can be subject to liability for the invasion of her privacy, but only if the intrusion would be highly offensive to a reasonable person.
- Public disclosure of private facts. Similarly, a person can be sued for publishing private facts about another person, even if those facts are true. Private facts are those about someone's personal life that have not previously been made public, that are not of legitimate public concern, and that would be offensive to a reasonable person.

---

[17] One Hundred Tenth Congress, "Privacy: The use of commercial information resellers by federal agencies," *Hearing before the Subcommittee on Information Policy, Census, and National Archives of the Committee on Oversight and Government Reform*, House of Representatives, March 11, 2008.

[18] For example, Experian provides much of Healthcare.gov's identity verification component using consumer credit information not available to the government. See *Consumer Reports*, "Having trouble proving your identity to HealthCare.gov? Here's how the process works," December 18, 2013. http://www.consumerreports.org/cro/news/2013/12/how-to-prove-your-identity-on-healthcare-gov/index.htm?loginMethod=auto

[19] Warren, Samuel D. and Louis D. Brandeis, "The Right to Privacy." *Harvard Law Review* 4:5, 193, December 15, 1890.

[20] Prosser, William L., "Privacy," *California Law Review* 48:383, 389, 1960.

[21] Id.

[22] (1) Digital Media Law Project, "Publishing Personal and Private Information." http://www.dmlp.org/legal-guide/publishing-personal-and-private-information. (2) Id., "Elements of an Intrusion Claim." http://www.dmlp.org/legal-guide/elements-intrusion-claim

- "False light" or publicity. Closely related to defamation, this harm results when false facts are widely published about an individual. In some states, false light includes untrue implications, not just untrue facts as such.
- Misappropriation of name or likeness. Individuals have a "right of publicity" to control the use of their name or likeness in commercial settings.

It seems likely that most Americans today continue to share the values implicit in these harms, even if the legal language (by now refined in thousands of court decisions) strikes one as archaic and quaint. However, new technological insults to privacy, actual or prospective, and a century's evolution of social values (for example, today's greater recognition of the rights of minorities, and of rights associated with gender), may require a longer list than sufficed in 1960.

Although PCAST's engagement with this subject is centered on technology, not law, any report on the subject of privacy, including PCAST's, should be grounded in the values of its day. As a starting point for discussion, albeit only a snapshot of the views of one set of technologically minded Americans, PCAST offers some possible augmentations to the established list of harms, each of which suggests a possible underlying right in the age of big data.

PCAST also believes strongly that the positive benefits of technology are (or can be) greater than any new harms. Almost every new harm is related to or "adjacent to" beneficial uses of the same technology.[23] To emphasize this point, for each suggested new harm, we describe a related beneficial use.

- **Invasion of private communications.** Digital communications technologies make social networking possible across the boundaries of geography, and enable social and political participation on previously unimaginable scales. An individual's right to private communication, secured for written mail and wireline telephone in part by the isolation of their delivery infrastructure, may need reaffirmation in the digital era, however, where all kinds of "bits" share the same pipelines, and the barriers to interception are often much lower. (In this context, we discuss the use and limitations of encryption in Section 4.2.)
- **Invasion of privacy in a person's virtual home.** The Fourth Amendment gives special protection against government intrusion into the home, for example the protection of private records within the home; tort law offers protection against similar non-government intrusion. The new "virtual home" includes the Internet, cloud storage, and other services. Personal data in the cloud can be accessible and organized. Photographs and records in the cloud can be shared with family and friends, and can be passed down to future generations. The underlying social value, the "home as one's castle," should logically extend to one's "castle in the cloud," but this protection has not been preserved in the new virtual home. (We discuss this subject further in Section 2.3.)
- **Public disclosure of inferred private facts.** Powerful data analytics may infer personal facts from seemingly harmless input data. Sometimes the inferences are beneficial. At its best, targeted advertising directs consumers to products that they actually want or need. Inferences about people's health can lead to better and timelier treatments and longer lives. But before the advent of big data, it could be assumed that there was a clear distinction between public and private information: either a fact was "out there" (and could be pointed to), or it was not. Today, analytics may discover facts that

[23] One perspective informed by new technologies and technology-medicated communication suggests that privacy is about the "continual management of boundaries between different spheres of action and degrees of disclosure within those spheres," with privacy and one's public face being balanced in different ways at different times. See: Leysia Palen and Paul Dourish, "Unpacking 'Privacy' for a Networked World," *Proceedings of CHI 2003,* Association for Computing Machinery, April 5-10, 2003.

are no less private than yesterday's purely private sphere of life. Examples include inferring sexual preference from purchasing patterns, or early Alzheimer's disease from key-click streams. In the latter case, the private fact may not even be known to the individual in question. (Section 3.2 discusses the technology behind the data analytics that makes such inferences possible.) The public disclosure of such information (and possibly also some non-public commercial uses) seems offensive to widely shared values.

- **Tracking, stalking, and violations of locational privacy.** Today's technologies easily determine an individual's current or prior location. Useful location-based services include navigation, suggesting better commuter routes, finding nearby friends, avoiding natural hazards, and advertising the availability of nearby goods and services. Sighting an individual in a public place can hardly be a private fact. When big data allows such sightings, or other kinds of passive or active data collection, to be assembled into the continuous locational track of an individual's private life, however, many Americans (including Supreme Court Justice Sotomayor, for example[24]) perceive a potential affront to a widely accepted "reasonable expectation of privacy."
- **Harm arising from false conclusions about individuals, based on personal profiles from big-data analytics.** The power of big data, and therefore its benefit, is often correlational. In many cases the "harms" from statistical errors are small, for example the incorrect inference of a movie preference; or the suggestion that a health issue be discussed with a physician, following from analyses that may, on average, be beneficial, even when a particular instance turns out to be a false alarm. Even when predictions are statistically valid, moreover, they may be untrue about particular individuals – and mistaken conclusions may cause harm. Society may not be willing to excuse harms caused by the uncertainties inherent in statistically valid algorithms. These harms may unfairly burden particular classes of individuals, for example, racial minorities or the elderly.
- **Foreclosure of individual autonomy or self-determination.** Data analyses about large populations can discover special cases that apply to individuals within that population. For example, by identifying differences in "learning styles," big data may make it possible to personalize education in ways that recognize every individual's potential and optimize that individual's achievement. But the projection of population factors onto individuals can be misused. It is widely accepted that individuals should be able to make their own choices and pursue opportunities that are not necessarily typical, and that no one should be denied the chance to achieve more than some statistical expectation of themselves. It would offend our values if a child's choices in video games were later used for educational tracking (for example, college admissions). Similarly offensive would be a future, akin to Philip K. Dick's science fiction short story adapted by Steven Spielberg in the film *Minority Report,* where "pre-crime" is statistically identified and punished.[25]
- **Loss of anonymity and private association.** Anonymity is not acceptable as an enabler of committing fraud, or bullying, or cyber-stalking, or improper interactions with children. Apart from wrongful behavior, however, the individual's right to choose to be anonymous is a long held American value (as, for example, the anonymous authorship of the Federalist papers). Using data to (re-) identify an individual who wishes to be anonymous (except in the case of legitimate governmental functions, such as law enforcement) is regarded as a harm. Similarly, individuals have a right of private association with groups or other individuals, and the identification of such associations may be a harm.

[24] "I would ask whether people reasonably expect that their movements will be recorded and aggregated in a manner that enables the Government to ascertain, more or less at will, their political and religious beliefs, sexual habits, and so on." *United States v. Jones* (10-1259), Sotomayor concurrence at http://www.supremecourt.gov/opinions/11pdf/10-1259.pdf.

[25] Dick, Phillip K., "The Minority Report," first published in *Fantastic Universe* (1956) and reprinted in *Selected Stories of Philip K. Dick*, New York: Pantheon, 2002.

While in no sense is the above list intended to be complete, it does have a few intentional omissions. For example, individuals may want big data to be used "fairly," in the sense of treating people equally, but (apart from the small number of protected classes already defined by law) it seems impossible to turn this into a right that is specific enough to be meaningful. Likewise, individuals may want the ability to know what others know about them; but that is surely not a right from the pre-digital age; and, in the current era of statistical analysis, it is not so easy to define what "know" means. This important issue is discussed in Section 3.1.2, and again taken up in chapter 5, where the attempt is to focus on actual harms done by the *use* of information, not by a concept as technically ambiguous as whether information is *known*.

2. Examples and Scenarios

This chapter seeks to make Chapter 1's introductory discussion more concrete by sketching some examples and scenarios. While some of these applications of technology are in use today, others comprise PCAST's technological prognostications about the near future, up to perhaps 10 years from today. Taken together the examples and scenarios are intended to illustrate both the enormous benefits that big data can provide and also the privacy challenges that may accompany these benefits.

In the following three sections, it will be useful to develop some scenarios more completely than others, moving from very brief examples of things happening today to more fully developed scenarios set in the future.

2.1 Things happening today or very soon

Here are some relevant examples:

- Pioneered more than a decade ago, devices mounted on utility poles are able to sense the radio stations being listened to by passing drivers, with the results sold to advertisers.[26]
- In 2011, automatic license-plate readers were in use by three quarters of local police departments surveyed. Within 5 years, 25% of departments expect to have them installed on all patrol cars, alerting police when a vehicle associated with an outstanding warrant is in view.[27] Meanwhile, civilian uses of license-plate readers are emerging, leveraging cloud platforms and promising multiple ways of using the information collected.[28]
- Experts at the Massachusetts Institute of Technology and the Cambridge Police Department have used a machine-learning algorithm to identify which burglaries likely were committed by the same offender, thus aiding police investigators.[29]
- Differential pricing (offering different prices to different customers for essentially the same goods) has become familiar in domains such as airline tickets and college costs. Big data may increase the power and prevalence of this practice and may also decrease even further its transparency.[30]

[26] ElBoghdady, Dina, "Advertisers Tune In to New Radio Gauge," *The Washington Post*, October 25, 2004. http://www.washingtonpost.com/wp-dyn/articles/A60013-2004Oct24.html

[27] American Civil Liberties Union, "You Are Being Tracked: How License Plate Readers Are Being Used To Record Americans' Movements," July, 2013. https://www.aclu.org/files/assets/071613-aclu-alprreport-opt-v05.pdf

[28] Hardy, Quentin, "How Urban Anonymity Disappears When All Data Is Tracked," *The New York Times*, April 19, 2014.

[29] Rudin, Cynthia, "Predictive policing: Using Machine Learning to Detect Patterns of Crime," *Wired*, August 22, 2013. http://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/.

[30] (1) Schiller, Benjamin, "First Degree Price Discrimination Using Big Data," Jan. 30. 2014, Brandeis University. http://benjaminshiller.com/images/First_Degree_PD_Using_Big_Data_Jan_27,_2014.pdf and http://www.forbes.com/sites/modeledbehavior/2013/09/01/will-big-data-bring-more-price-discrimination/ (2) Fisher, William W. "When Should We Permit Differential Pricing of Information?" *UCLA Law Review* 55:1, 2007.

- The UK firm FeatureSpace offers machine-learning algorithms to the gaming industry that may detect early signs of gambling addiction or other aberrant behavior among online players.[31]
- Retailers like CVS and AutoZone analyze their customers' shopping patterns to improve the layout of their stores and stock the products their customers want in a particular location.[32] By tracking cell phones, RetailNext offers bricks-and-mortar retailers the chance to recognize returning customers, just as cookies allow them to be recognized by on-line merchants.[33] Similar WiFi tracking technology could detect how many people are in a closed room (and in some cases their identities).
- The retailer Target inferred that a teenage customer was pregnant and, by mailing her coupons intended to be useful, unintentionally disclosed this fact to her father.[34]
- The author of an anonymous book, magazine article, or web posting is frequently "outed" by informal crowd sourcing, fueled by the natural curiosity of many unrelated individuals.[35]
- Social media and public sources of records make it easy for anyone to infer the network of friends and associates of most people who are active on the web, and many who are not.[36]
- Marist College in Poughkeepsie, New York, uses predictive modeling to identify college students who are at risk of dropping out, allowing it to target additional support to those in need.[37]
- The Durkheim Project, funded by the U.S. Department of Defense, analyzes social-media behavior to detect early signs of suicidal thoughts among veterans.[38]
- LendUp, a California-based startup, sought to use nontraditional data sources such as social media to provide credit to underserved individuals. Because of the challenges in ensuring accuracy and fairness, however, they have been unable to proceed.[39,40]

---

[31] Burn-Murdoch, John, "UK technology firm uses machine learning to combat gambling addiction," *The Guardian*, August 1, 2013. http://www.theguardian.com/news/datablog/2013/aug/01/uk-firm-uses-machine-learning-fight-gambling-addiction

[32] Clifford, Stephanie, "Using Data to Stage-Manage Paths to the Prescription Counter," *The New York Times*, June 19, 2013. http://bits.blogs.nytimes.com/2013/06/19/using-data-to-stage-manage-paths-to-the-prescription-counter/

[33] Clifford, Stephanie, "Attention, Shoppers: Store Is Tracking Your Cell," *The New York Times*, July 14, 2013.

[34] Duhigg, Charles, "How Companies Learn Your Secrets," *The New York Times Magazine*, February 12, 2012. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0

[35] Volokh, Eugene, "Outing Anonymous Bloggers," June 8, 2009. http://www.volokh.com/2009/06/08/outing-anonymous-bloggers/; A. Narayanan et al., "On the Feasibility of Internet-Scale Author Identification," IEEE Symposium on Security and Privacy, May 2012. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6234420

[36] Facebook's "The Graph API" (at https://developers.facebook.com/docs/graph-api/) describes how to write computer programs that can access the Facebook friends' data.

[37] One of four big-data applications honored by the trade journal, *Computerworld*, in 2013. King, Julia, "UN tackles socio-economic crises with big data," *Computerworld*, June 3, 2013. http://www.computerworld.com/s/article/print/9239643/UN_tackles_socio_economic_crises_with_big_data

[38] Ungerleider, Neal, "This May Be The Most Vital Use Of "Big Data" We've Ever Seen," *Fast Company*, July 12, 2013. http://www.fastcolabs.com/3014191/this-may-be-the-most-vital-use-of-big-data-weve-ever-seen.

[39] Center for Data Innovations, *100 Data Innovations*, Information Technology and Innovation Foundation, Washington, DC, January 2014. http://www2.datainnovation.org/2014-100-data-innovations.pdf

[40] Waters, Richard, "Data open doors to financial innovation," *Financial Times*, December 13, 2013. http://www.ft.com/intl/cms/s/2/3c59d58a-43fb-11e2-844c-00144feabdc0.html

- Insight into the spread of hospital-acquired infections has been gained through the use of large amounts of patient data together with personal information about uninfected patients and clinical staff.[41]
- Individuals' heart rates can be inferred from the subtle changes in their facial coloration that occur with each beat, enabling inferences about their health and emotional state.[42]

2.2 Scenarios of the near future in healthcare and education

Here are a few examples of the kinds of scenarios that can readily be constructed.

2.2.1 Healthcare: personalized medicine

Not all patients who have a particular disease are alike, nor do they respond identically to treatment. Researchers will soon be able to draw on millions of health records (including analog data such as scans in addition to digital data), vast amounts of genomic information, extensive data on successful and unsuccessful clinical trials, hospital records, and so forth. In some cases they will be able to discern that among the diverse manifestations of the disease, a subset of the patients have a collection of traits that together form a variant that responds to a particular treatment regime.

Since the result of the analysis could lead to better outcomes for particular patients, it is desirable to identify those individuals in the cohort, contact them, treat their disease in a novel way, and use their experiences in advancing the research. Their data may have been gathered only anonymously, however, or it may have been de-identified.

Solutions may be provided by specific new technologies for the protection of database privacy. These may create a protected query mechanism so individuals can find out whether they are in the cohort, or provide an alert mechanism based on the cohort characteristics so that, when a medical professional sees a patient in the cohort, a notice is generated.

2.2.2 Healthcare: detection of symptoms by mobile devices

Many baby boomers wonder how they might detect Alzheimer's disease in themselves. What would be better to observe their behavior than the mobile device that connects them to a personal assistant in the cloud (e.g., Siri or OK Google), helps them navigate, reminds them what words mean, remembers to do things, recalls conversations, measures gait, and otherwise is in a position to detect gradual declines on traditional and novel medical indicators that might be imperceptible even to their spouses?

At the same time, any leak of such information would be a damaging betrayal of trust. What are individuals' protections against such risks? Can the inferred information about individuals' health be sold, without additional consent, to third parties (e.g., pharmaceutical companies)? What if this is a stated condition of use of

[41] (1) Wiens, Jenna, John Guttag, and Eric Horvitz, "A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to Enhance Hospital-Specific Predictions," *Journal of the American Medical Informatics* Association, January 2014. (2) Weitzner, Daniel J., et al., "Consumer Privacy Bill of Rights and Big Data: Response to White House Office of Science and Technology Policy Request for Information," April 4, 2014.

[42] Frazer, Bryant, "MIT Computer Program Reveals Invisible Motion in Video," *The New York Times* video, February 27, 2013. https://www.youtube.com/watch?v=3rWycBEHn3s

the app? Should information go to individuals' personal physicians with their initial consent but not a subsequent confirmation?

2.2.3 Education

Drawing on millions of logs of online courses, including both massive open on-line courses (MOOCs) and smaller classes, it will soon be possible to create and maintain longitudinal data about the abilities and learning styles of millions of students. This will include not just broad aggregate information like grades, but fine-grained profiles of how individual students respond to multiple new kinds of teaching techniques, how much help they need to master concepts at various levels of abstraction, what their attention span is in various contexts, and so forth. A MOOC platform can record how long a student watches a particular video; how often a segment is repeated, sped up, or skipped; how well a student does on a quiz; how many times he or she misses a particular problem; and how the student balances watching content to reading a text. As the ability to present different material to different students materializes in the platforms, the possibility of blind, randomized A/B testing enables the gold standard of experimental science to be implemented at large scale in these environments.[43]

Similar data are also becoming available for residential classes, as learning-management systems (such as Canvas, Blackboard, or Desire2Learn) expand their roles to support innovative pedagogy. In many courses one can now get moment-by-moment tracking of the student's engagement with the course materials and correlate that engagement with the desired learning outcomes.

With this information, it will be possible not only to greatly improve education, but also to discover what skills, taught to which individuals at which points in childhood, lead to better adult performance in certain tasks, or to adult personal and economic success. While these data could revolutionize educational research, the privacy issues are complex.[44]

There are many privacy challenges in this vision of the future of education. Knowledge of early performance can create implicit biases[45] that color later instruction and counseling. There is great potential for misuse, ostensibly for the social good, in the massive ability to direct students into high- or low-potential tracks. Parents and others have access to sensitive information about children, but mechanisms rarely exist to change those permissions when the child reaches majority.

2.3 Challenges to the home's special status

The home has special significance as a sanctuary of individual privacy. The Fourth Amendment's list, "persons, houses, papers, and effects," puts only the physical body in the rhetorically more prominent position; and a house is often the physical container for the other three, a boundary inside of which enhanced privacy rights apply.

---

[43] For an overview of MOOCs and associated analytics opportunities, see PCAST's December 2013 letter to the President. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_edit_dec-2013.pdf

[44] There is also uncertainty about how to interpret applicable laws, such as the Family Educational Rights and Privacy Act (FERPA). Recent Federal guidance is intended to help clarify the situation. See: U.S. Department of Education, "Protecting Student Privacy While Using Online Educational Services: Requirements and Best Practices," February 2014. http://ptac.ed.gov/sites/default/files/Student%20Privacy%20and%20Online%20Educational%20Services%20%28February%202014%29.pdf

[45] Cukier, Kenneth, and Viktor Mayer-Schoenberger, "How Big Data Will Haunt You Forever," *Quartz*, March 11, 2014. http://qz.com/185252/how-big-data-will-haunt-you-forever-your-high-school-transcript/

Existing interpretations of the Fourth Amendment are inadequate for the present world, however. We, along with the "papers and effects" contemplated by the Fourth Amendment, live increasingly in cyberspace, where the physical boundary of the home has little relevance. In 1980, a family's financial records were paper documents, located perhaps in a desk drawer inside the house. By 2000, they were migrating to the hard drive of the home computer – but still within the house. By 2020, it is likely that most such records will be in the cloud, not just outside the house, but likely replicated in multiple legal jurisdictions – because cloud storage typically uses location diversity to achieve reliability. The picture is the same if one substitutes for financial records something like "political books we purchase," or "love letters that we receive," or "erotic videos that we watch." Absent different policy, legislative, and judicial approaches, the physical sanctity of the home's papers and effects is rapidly becoming an empty legal vessel.

The home is also the central locus of Brandeis' "right to be left alone." This right is also increasingly fragile, however. Increasingly, people bring sensors into their homes whose immediate purpose is to provide convenience, safety, and security. Smoke and carbon monoxide alarms are common, and often required by safety codes.[46] Radon detectors are usual in some parts of the country. Integrated air monitors that can detect and identify many different kinds of pollutants and allergens are readily foreseeable. Refrigerators may soon be able to "sniff" for gases released from spoiled food, or, as another possible path, may be able to "read" food expiration dates from radio-frequency identification (RFID) tags in the food's packaging. Rather than today's annoying cacophony of beeps, tomorrow's sensors (as some already do today) will interface to a family through integrated apps on mobile devices or display screens. The data will have been processed and interpreted. Most likely that processing will occur in the cloud. So, to deliver services the consumer wants, much data will need to have left the home.

Environmental sensors that enable new food and air safety may also be able to detect and characterize tobacco or marijuana smoke. Health care or health insurance providers may want assurance that self-declared non-smokers are telling the truth. Might they, as a condition of lower premiums, require the homeowner's consent for tapping into the environmental monitors' data? If the monitor detects heroin smoking, is an insurance company obligated to report this to the police? Can the insurer cancel the homeowner's property insurance?

To some, it seems farfetched that the typical home will foreseeably acquire cameras and microphones in every room, but that appears to be a likely trend. What can your cell phone (already equipped with front and back cameras) hear or see when it is on the nightstand next to your bed? Tablets, laptops, and many desktop computers have cameras and microphones. Motion detector technology for home intrusion alarms will likely move from ultrasound and infrared to imaging cameras – with the benefit of fewer false alarms and the ability to distinguish pets from people. Facial-recognition technology will allow further security and convenience. For the safety of the elderly, cameras and microphones will be able to detect falls or collapses, or calls for help, and be networked to summon aid.

People naturally communicate by voice and gesture. It is inevitable that people will communicate with their electronic servants in both such modes (necessitating that they have access to cameras and microphones).

[46] Nest, acquired by Google, attracted attention early for its design and its use of big data to adapt to consumer behavior. See: Aoki, Kenji, "Nest Gives the Lowly Smoke Detector a Brain," *Wired*, October, 2013. http://www.wired.com/2013/10/nest-smoke-detector/all/

Companies such as PrimeSense, an Israeli firm recently bought by Apple,[47] are developing sophisticated computer-vision software for gesture reading, already a key feature in the consumer computer game console market (e.g., Microsoft Kinect). Consumer televisions are already among the first "appliances" to respond to gesture; already, devices such as the Nest smoke detector respond to gestures.[48] The consumer who taps his temple to signal a spoken command to Google Glass[49] may want to use the same gesture for the television, or for that matter for the thermostat or light switch, in any room at home. This implies omnipresent audio and video collection within the home.

All of these audio, video, and sensor data will be generated within the supposed sanctuary of the home. But they are no more likely to stay in the home than the "papers and effects" already discussed. Electronic devices in the home already invisibly communicate to the outside world via multiple separate infrastructures: The cable industry's hardwired connection to the home provides multiple types of two-way communication, including broadband Internet. Wireline phone is still used by some home-intrusion alarms and satellite TV receivers, and as the physical layer for DSL broadband subscribers. Some home devices use the cell-phone wireless infrastructure. Many others piggyback on the home Wi-Fi network that is increasingly a necessity of modern life. Today's smart home-entertainment system knows what a person records on a DVR, what she actually watches, and when she watches it. Like personal financial records in 2000, this information today is in part localized inside the home, on the hard drive inside the DVR. As with financial information today, however, it is on track to move into the cloud. Today, Netflix or Amazon can offer entertainment suggestions based on customers' past key-click streams and viewing history on their platforms. Tomorrow, even better suggestions may be enabled by interpreting their minute-by-minute facial expressions as seen by the gesture-reading camera in the television.

These collections of data are benign, in the sense that they are necessary for products and services that consumers will knowingly demand. Their challenges to privacy arise both from the fact that their analog sensors necessarily collect more information than is minimally necessary for their function (see Section 3.1.2), and also because their data practically cry out for secondary uses ranging from innovative new products to marketing bonanzas to criminal exploits. As in many other kinds of big data, there is ambiguity as to data ownership, data rights, and allowed data use. Computer-vision software is likely already able to read the brand labels on products in its field of view – this is a much easier technology than facial recognition. If the camera in your television knows what brand of beer you are drinking while watching a football game, and knows whether you opened the bottle before or after the beer ad, who (if anyone) is allowed to sell this information to the beer company, or to its competitors? Is the camera allowed to read brand names when the television set is supposedly off? Can it watch for magazines or political leaflets? If the RFID tag sensor in your refrigerator usefully detects out-of-date food, can it also report your brand choices to vendors? Is this creepy and strange, or a consumer financial benefit when every supermarket can offer you relevant coupons?[50] Or (the dilemma of

[47] Reuters, "Apple acquires Israeli 3D chip developer PrimeSense," November 25, 2013. http://www.reuters.com/article/2013/11/25/us-primesense-offer-apple-idUSBRE9AO04C20131125

[48] Id.

[49] Google, "Glass gestures." https://support.google.com/glass/answer/3064184?hl=en

[50] Tene, Omer, and Jules Polonetsky, "A Theory of Creepy: Technology, Privacy and Shifting Social Norms," *Yale Journal of Law and Technology* 16:59, 2013, pp. 59-100.

differential pricing[51]) is it any different if the data are used to offer *others* a better deal while *you* pay full price because your brand loyalty is known to be strong?

About one-third of Americans rent, rather than own, their residences. This number may increase with time as a result of long-term effects of the 2007 financial crisis, as well as aging of the U.S. population. Today and foreseeably, renters are less affluent, on average, than homeowners. The law demarcates a fine line between the property rights of landlords and the privacy rights of tenants. Landlords have the right to enter their property under various conditions, generally including where the tenant has violated health or safety codes, or to make repairs. As more data are collected within the home, the rights of tenant and landlord may need new adjustment. If environmental monitors are fixtures of the landlord's property, does she have an unconditional right to their data? Can she sell those data? If the lease so provides, can she evict the tenant if the monitor repeatedly detects cigarette smoke, or a camera sensor is able to distinguish a prohibited pet?

If a third party offers facial recognition services for landlords (no doubt with all kinds of cryptographic safeguards!), can the landlord use these data to enforce lease provisions against subletting or additional residents? Can she require such monitoring as a condition of the lease? What if the landlord's cameras are outside the doors, but keep track of everyone who enters or leaves her property? How is this different from the case of a security camera across the street that is owned by the local police?

2.4 Tradeoffs among privacy, security, and convenience

Notions of privacy change generationally. One sees today marked differences between the younger generation of "digital natives" and their parents or grandparents. In turn, the children of today's digital natives will likely have still different attitudes about the flow of their personal information. Raised in a world with digital assistants who know everything about them, and (one may hope) with wise policies in force to govern use of the data, future generations may see little threat in scenarios that individuals today would find threatening, if not Orwellian. PCAST's final scenario, perhaps at the outer limit of its ability to prognosticate, is constructed to illustrate this point.

Taylor Rodriguez prepares for a short business trip. She packed a bag the night before and put it outside the front door of her home for pickup. No worries that it will be stolen: The camera on the streetlight was watching it; and, in any case, almost every item in it has a tiny RFID tag. Any would-be thief would be tracked and arrested within minutes. Nor is there any need to give explicit instructions to the delivery company, because the cloud knows Taylor's itinerary and plans; the bag is picked up overnight and will be in Taylor's destination hotel room by the time of her arrival.

Taylor finishes breakfast and steps out the front door. Knowing the schedule, the cloud has provided a self-driving car, waiting at the curb. At the airport, Taylor walks directly to the gate – no need to go through any security. Nor are there any formalities at the gate: A twenty-minute "open door" interval is provided for passengers to stroll onto the plane and take their seats (which each sees individually highlighted in his or her wearable optical device). There are no boarding passes and no organized lines. Why bother, when Taylor's identity (as for everyone else who enters the airport) has been tracked and is known absolutely? When her known information emanations (phone, RFID tags in clothes, facial recognition, gait, emotional state) are known to the cloud, vetted, and essentially unforgeable? When, in the unlikely event that Taylor has become deranged and dangerous, many detectable signs would already have been tracked, detected, and acted on?

[51] See references at footnote 30.

Indeed, everything that Taylor carries has been screened far more effectively than any rushed airport search today. Friendly cameras in every LED lighting fixture in Taylor's house have watched her dress and pack, as they do every day. Normally these data would be used only by Taylor's personal digital assistants, perhaps to offer reminders or fashion advice. As a condition of using the airport transit system, however, Taylor has authorized the use of the data for ensuring airport security and public safety.

Taylor's world seems creepy to us. Taylor has accepted a different balance among the public goods of convenience, privacy, and security than would most people today. Taylor acts in the unconscious belief (whether justified or not, depending on the nature and effectiveness of policies in force) that the cloud and its robotic servants are trustworthy in matters of personal privacy. In such a world, major improvements in the convenience and security of everyday life become possible.

3. Collection, Analytics, and Supporting Infrastructure

Big data is big in two different senses. It is big in the quantity and variety of data that are available to be processed. And, it is big in the scale of analysis ("analytics") that can be applied to those data, ultimately to make inferences. Both kinds of "big" depend on the existence of a massive and widely available computational infrastructure, one that is increasingly being provided by cloud services. This chapter expands on these basic concepts.

3.1 Electronic sources of personal data

Since early in the computer age, public and private entities have been assembling digital information about people. Databases of personal information were created during the days of "batch processing."[52] Indeed, early descriptions of database technology often talk about personnel records used for payroll applications. As computing power increased, more and more business applications moved to digital form. There now are digital telephone-call records, credit-card transaction records, bank-account records, email repositories, and so on. As interactive computing has advanced, individuals have entered more and more data about themselves, both for self-identification to an online service and for productivity tools such as financial-management systems.

These digital data are normally accompanied by "metadata" or ancillary data that explain the layout and meaning of the data they describe. Databases have schemas and email has headers,[53] as do network packets.[54] As data sets become more complex, so do the attached metadata. Included in the data or metadata may be identifying information such as account numbers, login names, and passwords. There is no reason to believe that metadata raise fewer privacy concerns than the data they describe.

In recent times, the kinds of electronic data available about people have increased substantially, in part because of the emergence of social media and in part because of the growth in mobile devices, surveillance devices, and a diversity of networked sensors. Today, although they may not be aware of it, individuals constantly emit into the environment information whose use or misuse may be a source of privacy concerns. Physically, these information emanations are of two types, which can be called "born digital" or "born analog."

3.1.1 "Born digital" data

When information is "born digital," it is created, by us or by a computer surrogate, specifically for digital use – that is, for use by a computer or data-processing system. Examples of data that are born digital include:

- email and text messaging
- input via mouse-clicks, taps, swipes, or keystrokes on a phone, tablet, computer, or video game; that is, data that people intentionally enter into a device

[52] Such databases endure and form the basis of continuing concern among privacy advocates.

[53] Schemas are formal definitions of the configuration of a database: its tables, relations, and indices. Headers are the sometimes-invisible prefaces to email messages that contain information about the sending and destination addresses and sometimes the routing of the path between them.

[54] In the Internet and similar networks, information is broken up into chunks called packets, which may travel independently and depend on metadata to be reassembled properly at the destination of the transmission.

- GPS location data
- metadata associated with phone calls: the numbers dialed from or to, the time and duration of calls
- data associated with most commercial transactions: credit-card swipes, bar-code reads, reads of RFID tags (as used for anti-theft and inventory control)
- data associated with portal access (key card or ID badge reads) and toll-road access (remote reads of RFID tags)
- metadata that our mobile devices use to stay connected to the network, including device location and status
- increasingly, data from cars, televisions, appliances: the "Internet of Things"

Consumer-tracking data provide an example of born-digital data that has become economically important. It is generally possible for companies to aggregate large amounts of data and then use those data for marketing, advertising, or many other activities. The traditional mechanism has been to use cookies, small data files that a browser can leave on a user's computer (pioneered by Netscape two decades ago). The technique is to leave a cookie when a user first visits a site and then be able to correlate that visit with a subsequent event. This information is very valuable to retailers and forms the basis of many of the advertising businesses of the last decade. There has been a variety of proposals to regulate such tracking,[55] and many countries require opt-in permission before this tracking is done. Cookies involve relatively simple pieces of information that proponents represent as unlikely to be abused. Although not always aware of the process, people accept such tracking in return for a free or subsidized service.[56] At the same time, cookie-free alternatives are sometimes available.[57] Even without cookies, so-called "fingerprinting" techniques can often identify a user's computer or mobile device uniquely by the information that it exposes publicly, such as the size of its screen, its installed fonts, and other features.[58] Most technologists believe that applications will move away from cookies, that cookies are too simple an idea, and that there are better analytics coming and better approaches being invented. The economic incentives for consumer tracking will remain, however, and big data will allow for more precise responses.

Tracking is also the enabling technology of some more nefarious uses. Unfortunately, many social networking apps begin by taking a person's contact list and spamming all the recipients with advertising for the app. This technique is often abused, especially by small start-ups who may assess the value gained by reaching new customers as being greater than the value lost to their reputation for honoring privacy.

---

[55] Federal Trade Commission, "FTC Staff Revises Online Behavioral Advertising Principles," Press Release, February 12, 2009. http://www.ftc.gov/news-events/press-releases/2009/02/ftc-staff-revises-online-behavioral-advertising-principles

[56] (1) Cf. *The Wall Street Journal*'s "What they know" series (http://online.wsj.com/public/page/what-they-know-digital-privacy.html). (2) Turow, Joseph, *The Daily You: How the Advertising Industry is Defining your Identity and Your Worth*, Yale University Press, 2012. http://yalepress.yale.edu/book.asp?isbn=9780300165012

[57] DuckDuckGo is a non-tracking search engine that, while perhaps yielding fewer results than leading search engines, is used by those looking for less tracking. See: https://duckduckgo.com/

[58] (1) Tanner, Adam, "The Web Cookie Is Dying. Here's The Creepier Technology That Comes Next," *Forbes*, June 17, 2013. http://www.forbes.com/sites/adamtanner/2013/06/17/the-web-cookie-is-dying-heres-the-creepier-technology-that-comes-next/ (2) Acar, G. et al., "FPDetective: Dusting the Web for Fingerprinters," 2013. http://www.cosic.esat.kuleuven.be/publications/article-2334.pdf

All information that is born digital shares certain characteristics. It is created in identifiable units for particular purposes. These units are in most cases "data packets" of one or another standard type. Since they are created by intent, the information that they contain is usually limited, for reasons of efficiency and good engineering design, to support the immediate purpose for which they are collected.

When data are born digital, privacy concerns can arise in two different modes, one obvious ("over-collection"), the other more recent and subtle ("data fusion"). Over-collection occurs when an engineering design intentionally, and sometimes clandestinely, collects information unrelated to its stated purpose. While your smartphone could easily photograph and transmit to a third party your facial expression as you type every keystroke of a text message, or could capture all keystrokes, thereby recording text that you had deleted, these would be inefficient and unreasonable software design choices for the default text-messaging app. In that context they would be instances of over-collection.

A recent example of over-collection was the *Brightest Flashlight Free* phone app, downloaded by more than 50 million users, which passed back to its vendor its location every time the flashlight was used. Not only is location information unnecessary for the illumination function of a flashlight, but it also discloses personal information that the user might wish to keep private. The Federal Trade Commission issued a complaint because the fine print on the notice-and-consent screen (see Section 4.3) had neglected to disclose that location information, whose collection was disclosed, would be sold to third parties, such as advertisers.[59,60] One sees in this example the limitations of the notice-and-consent framework: A more detailed initial fine-print disclosure by *Brightest Flashlight Free*, which almost no one would have actually read, would likely have forestalled any FTC action without much affecting the number of downloads.

In contrast to over-collection, data fusion occurs when data from different sources are brought into contact and new, often unexpected, phenomena emerge (see Section 3.1). Individually, each data source may have been designed for a specific, limited purpose. But when multiple sources are processed by techniques of modern statistical data mining, pattern recognition, and the combining of records from diverse sources by virtue of common identifying data, new meanings can be found. In particular, data fusion frequently results in the identification of individual people (that is, the association of events with unique personal identities), the creation of data-rich profiles of an individual, and the tracking of an individual's activities over days, months, or years.

By definition, the privacy challenges from data fusion do not lie in the individual data streams, each of whose collection, real-time processing, and retention may be wholly necessary and appropriate for its overt, immediate purpose. Rather, the privacy challenges are emergent properties of our increasing ability to bring into analytical juxtaposition large, diverse data sets and to process them with new kinds of mathematical algorithms.

---

[59] Federal Trade Commission, "Android Flashlight App Developer Settles FTC Charges It Deceived Consumers," *Press Release*, December 5, 2013. http://www.ftc.gov/news-events/press-releases/2013/12/android-flashlight-app-developer-settles-ftc-charges-it-deceived

[60] (1) FTC File No. 132-3087 Decision and order. http://www.ftc.gov/system/files/documents/cases/140409goldenshoresdo.pdf (2) "FTC Approves Final Order Settling Charges Against Flashlight App Creator." http://www.ftc.gov/news-events/press-releases/2014/04/ftc-approves-final-order-settling-charges-against-flashlight-app

3.1.2 Data from sensors

Turn now to the second broad class of information emanations. One can say that information is "born analog" when it arises from the characteristics of the physical world. Such information does not become accessible electronically until it impinges on a "sensor," an engineered device that observes physical effects and converts them to digital form. The most common sensors are cameras, including video, which sense visible electromagnetic radiation; and microphones, which sense sound and vibration. There are many other kinds of sensors, however. Today, cell phones routinely contain not only cameras, microphones, and radios but also analog sensors for magnetic fields (3-D compass) and motion (acceleration). Other kinds of sensors include those for thermal infrared (IR) radiation; air quality, including the identification of chemical pollutants; barometric pressure (and altitude); low-level gamma radiation; and many other phenomena.

Examples of born-analog data providing personal information and in use today include:

- the voice and/or video content of a phone call – born analog but immediately converted to digital by the phone's microphone and camera
- personal health data such as heartbeat, respiration, and gait, as sensed by special-purpose devices (Fitbit has been a leading provider[61]) or cell-phone apps
- cameras/sensors in televisions and video games that interpret gestures by the user
- video from security surveillance cameras, mobile phones, or overhead drones
- imaging infrared video that can see in what people perceive as total darkness (and also see evanescent traces of past events, so-called heat scars)
- microphone networks in cities, used to detect and locate gunshots and for public safety
- cameras/microphones in classrooms and other meeting rooms
- ultrasonic motion detectors
- medical imaging, CT, and MRI scans, ultrasonic imaging
- opportunistically collected chemical or biological samples, notably trace DNA (today requiring slow, off-line analysis, but foreseeably more nimble)
- synthetic aperture radar (SAR), which can image through clouds and, under some conditions, see inside of non-metallic structures
- unintended radiofrequency emissions from electrical and electronic devices

When data are born analog, they are likely to contain more information than the minimum necessary for their immediate purpose, for several valid reasons. One is that the desired information ("signal") must be sensed in the presence of unwanted extraneous information ("noise"). The technologies typically work by sensing the environment ("signal plus noise") with high precision, so that mathematical techniques can then be applied that will separate the two even in the worst anticipated case when the signal is smallest or the noise is largest.

Another reason is technological convergence. For example, as the cameras in cell phones become smaller and cheaper, the use of identical components in other products becomes a favored design choice, even when full images are not needed. Where a big-screen television today has separate sensors for its IR remote control, room brightness, and motion detection (a feature that turns off the picture when no one is in the room), plus a true video camera in the add-on game console, tomorrow's model may integrate all of these functions in a single, cheap, high-resolution, IR-sensitive camera, a few millimeters in size.

[61] See: http://www.fitbit.com/

In addition to the information available from digital and analog sources consciously intended to provide information about people, inadvertent disclosure abounds from the emerging "Internet of Things," an amalgamation of sensors whose primary purpose is enhanced by "smart" network-connected computational capabilities. Examples include "smart" thermostats that detect human presence and adjust air temperatures accordingly, "smart" automobile-ignition systems, and locking systems that are biometrically triggered.

The privacy challenges of born-analog data are somewhat different from those of born-digital data. Where over-collection (as was defined above) is an irrational design choice for the principled digital designer – and therefore an identifiable red flag for privacy issues – over-collection in the analog domain can be a robust and economical design choice. A consequence is that born-analog data will often contain information that was not originally expected. Unexpected information could in many cases lead to unanticipated beneficial products and services, but it could also give opportunities for unanticipated misuse.

As a concrete example, one might consider three key parameters of video imaging: resolution (how many pixels in the image), contrast ratio (how well can the image see into dark regions), and photometric precision (how accurate is the image in brightness and color). All three parameters have improved by orders of magnitude and are likely to keep improving. Today, with special cameras, one can image a cityscape from a high rooftop and see clearly into every facing house and apartment window within several miles.[62] Or, already mentioned, the ability exists to sense remotely the pulse of an individual, giving information on health status and emotional state.[63]

It is foreseeable, perhaps inevitable, that these capabilities will be present in every cell phone and security-surveillance camera, or every wearable computer device. (Imagine the process of negotiating the price for a car, or negotiating an international trade agreement, when every participant's Google Glass (or security camera or TV camera) is able to monitor and interpret the autonomic physiological state of every other participant, in real time.) It is unforeseeable what other unexpected information also lies in signals from the same sensors.

Once they enter the digital world, born-analog data can be fused and mined along with born-digital data. For example, facial-recognition algorithms, which might be error-prone in isolation, may yield nearly perfect identity tracking when they can be combined with born-digital data from cell phones (including unintended emanations), point-of-sale transactions, RFID tags, and so forth; and also with other born-analog data such as vehicle tracking (e.g., from overhead drones) and automated license-plate reading. Biometric data can provide identity information that enhances the profile of an individual even more, and data on behavior (as from social networks) are being used to analyze attitudes or emotions ("sentiment analysis," for individuals or groups[64]). In short, more and more information can be captured and put in a quantified format so it can be tabulated and analyzed.[65]

---

[62] Koonin, Steven E., Gregory Dobler and Jonathan S. Wurtele, "Urban Physics," *American Physical Society News*, March, 2014. http://www.aps.org/publications/apsnews/201403/urban.cfm

[63] Durand, Fredo, et al., "MIT Computer Program Reveals Invisible Motion in Video," *The New York Times,* video, February 27, 2013. https://www.youtube.com/watch?v=3rWycBEHn3s

[64] Feldman, Ronen, "Techniques and Applications for Sentiment Analysis," *Communications of the ACM*, 56:4, pp. 82-89.

[65] Mayer-Schönberger, Viktor and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Boston, NY: Houghton Mifflin Harcourt, 2013.

3.2 Big data analytics

Analytics is what makes big data come alive. Without analytics, big datasets could be stored, and they could be retrieved, wholly or selectively. But what comes out would be exactly what went in. Analytics, comprising a number of different computational technologies, is what fuels the big-data revolution.[66] Analytics is what creates the new value in big datasets, vastly more than the sum of the values of the parts.[67]

3.2.1 Data mining

Data-mining, sometimes loosely equated to analytics but actually only a subset of it, refers to a computational process that discovers patterns in large data sets. It is a convergence of many fields of academic research in both applied mathematics and computer science, including statistics, databases, artificial intelligence, and machine learning. Like other technologies, advances in data mining have a research and development stage, in which new algorithms and computer programs are developed, and they have subsequent phases of commercialization and application.

Data mining algorithms can be trained to find patterns either by supervised learning, so-called because the algorithm is seeded with manually curated examples of the pattern to be recognized, or by unsupervised learning, where the algorithm tries to find related pieces of data without prior seeding. A recent success of unsupervised-learning algorithms was a program that, searching millions of images on the web, figured out on its own that "cat" was a much-posted category.[68]

The desired output of data mining can take several forms, each with its own specialized algorithms.[69]

- Classification algorithms attempt to assign objects or events to known categories. For example, a hospital might want to classify discharged patients as high, medium, or low risk for readmission.
- Clustering algorithms group objects or events into categories by similarity, as in the "cat" example above.
- Regression algorithms (also called numerical prediction algorithms) try to predict numerical quantities. For example, a bank may want to predict, from the details in a loan application, the probability of a default.
- Association techniques try to find relationships between items in their data set. Amazon's suggested products and Netflix's suggested movies are examples.
- Anomaly-detection algorithms look for untypical examples within a data set, for example, detecting fraudulent transactions on a credit-card account.
- Summarization techniques attempt to find and present salient features in data. Examples include both simple statistical summaries (e.g., average student test scores by school and teacher), and higher-level analysis (e.g., a list of key facts about an individual as gleaned from all web postings that mention her).

---

[66] National Research Council, *Frontiers in Massive Data* Analysis, National Academies Press, 2013.

[67] (1) Thill, Brent and Nicole Hayashi, *Big Data = Big Disruption: One of the Most Transformative IT Trends Over the Next Decade*, UBS Securities LLC, October 2013. (2) McKinsey Global Institute, Center for Government, and Business Technology Office, *Open data: Unlocking innovation and performance with liquid information*, McKinsey & Company, October 2013.

[68] Le, Q.V. et al., "Building High-level Features Using Large Scale Unsupervised Learning," http://static.googleusercontent.com/media/research.google.com/en/us/archive/unsupervised_icml2012.pdf

[69] Bramer, M., "Principles of Data Mining," *Springer*, 2013.

Data mining is sometimes confused with machine learning, the latter a broad subfield of computer science in academic and industrial research.[70] Data mining makes use of machine learning, as well as other disciplines, while machine learning has applications to fields other than data mining, for example, robotics.

There are limitations, both practical and theoretical, to what data mining can accomplish, as well as limits to how accurate it can be. It may reveal patterns and relationships, but it usually cannot tell the user the value or significance of these patterns. For example, supervised learning based on the characteristics of known terrorists might find similar persons, but they might or might not be terrorists; and it would miss different classes of terrorists who don't fit the profile.

Data mining can identify relationships between behaviors and/or variables, but these relationships do not always indicate causality. If people who live under high-voltage power lines have higher morbidity, it might mean that power lines are a hazard to public health; or it might mean that people who live under power lines tend to be poor and have inadequate access to health care. The policy implications are quite different. While so-called confounding variables (in this example, income) can be corrected for when they are known and understood, there is no sure way to know whether all of them have been identified. Imputing true causality in big data is a research field in its infancy.[71]

Many data analyses yield correlations that might or might not reflect causation. Some data analyses develop imperfect information, either because of limitations of the algorithms, or by the use of biased sampling. Indiscriminate use of these analyses may cause discrimination against individuals or a lack of fairness because of incorrect association with a particular group.[72] In using data analyses, particular care must be taken to protect the privacy of children and other protected groups.

Real-world data are incomplete and noisy. These data-quality issues lower the performance of data-mining algorithms and obscure outputs. When economics allow, careful screening and preparation of the input data can improve the quality of results, but this data preparation is often labor intensive and expensive. Users, especially in the commercial sector, must trade off cost and accuracy, sometimes with negative consequences for the individual represented in the data. Additionally, real-world data can contain extreme events or outliers. Outliers may be real events that, by chance, are overrepresented in the data; or they may be the result of data-entry or data-transmission errors. In both cases they can skew the model and degrade performance. The study of outliers is an important research area of statistics.

3.2.2 Data fusion and information integration

Data fusion is the merging of multiple heterogeneous datasets into one homogeneous representation so that they can be better processed for data mining and management. Data fusion is used in a number of technical domains such as sensor networks, video/image processing, robotics and intelligent systems, and elsewhere.

---

[70] Mitchell, Tom M., "The Discipline of Machine Learning," Technical Report CMU-ML-06-108, Carnegie Mellon University, July 2006.

[71] DARPA, for example, has a project involving machine learning and other technologies to build medical causal models from analysis of cancer literature, leveraging the greater capacity of a computer than a person to process information from a large number of sources. See description at http://www.darpa.mil/Our_Work/I2O/Programs/Big_Mechanism.aspx

[72] "Data mining breaks the basic intuition that identity is the greatest source of potential harm because it substitutes inference for identifying information as a bridge to get at additional facts." Barocas, Solon and Helen Nissenbaum, "Big Data's End Run Around Anonymity and Consent," Chapter II, in Lane, Julia, et al., *Privacy, Big Data, and the Public Good,* Cambridge University Press, 2014.

Data integration is differentiated from data fusion in that integration more broadly combines data sets and retains the larger set of information. In data fusion, there is usually a reduction or replacement technique. Data fusion is facilitated by data interoperability, the ability for two systems to communicate and exchange data.

Data fusion and data integration are key techniques for business intelligence. Retailers are integrating their online, in-store, and catalog sales databases to create more complete pictures of their customers. Williams-Sonoma, for example, has integrated customer databases with information on 60 million households. Variables including household income, housing values, and number of children are tracked. It is claimed that targeted emails based on this information yield ten to 18 times the response rate of emails that are not targeted.[73] This is a simple illustration of how more information can lead to better inferences. Techniques that can help to preserve privacy are emerging.[74]

There is a great amount of interest today in multi-sensor data fusion.[75] The biggest technical challenges being tackled today, generally through development of new and better algorithms, relate to data precision/resolution, outliers and spurious data, conflicting data, modality (both heterogeneous and homogeneous data) and dimensionality, data correlation, data alignment, association within data, centralized vs. decentralized processing, operational timing, and the ability to handle dynamic vs. static phenomena. Privacy concerns may arise from sensor fidelity and precision as well as correlation of data from multiple sensors. A single sensor's output might not be sensitive, but the combination from two or more may raise privacy concerns.

3.2.3 Image and speech recognition

Image- and speech-recognition technologies are able to extract information, in some limited cases approaching human understanding, from massive corpuses of still images, videos, and recorded or broadcast speech.

Urban-scene extraction can be accomplished using a variety of data sources from photos and videos to ground based LiDAR (a remote-sensing technique using lasers).[76] In the government sector, city models are becoming vital for urban planning and visualization. They are equally important for a broad range of academic disciplines including history, archeology, geography, and computer-graphics research. Digital city models are also central to popular consumer mapping and visualization applications such as Google Earth and Bing Maps, as well as GPS-enabled navigation systems.[77] Scene extraction is an example of the inadvertent capture of personal information and can be used for data fusion that reveals personal information.

Facial-recognition technologies are beginning to be practical in commercial and law-enforcement applications.[78] They are able to acquire, normalize, and recognize moving faces in dynamic scenes. Real-time video surveillance with single-camera systems (and some with multi-camera systems, which can both recognize objects and analyze activity) has a wide variety of applications in both public and private environments, such as homeland

[73] Manyika, J. et al., "Big Data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, 2011.

[74] Navarro-Arriba, G. and V. Torra, "Information fusion in data privacy: A survey," *Information Fusion,* 13:4, 2012, pp. 235-244.

[75] Khaleghi, B. et al., "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion,* 14:1, 2013, pp. 28-44.

[76] Lam, J., et al., "Urban scene extraction from mobile ground based lidar data," *Proceedings of 3DPVT*, 2010.

[77] Agarwal, S., et al., "Building Rome in a day," *Communications of the ACM,* 54:10, 2011, pp. 105-112.

[78] Workshop on Frontiers in Image and Video Analysis, National Science Foundation, Federal Bureau of Investigation, Defense Advanced Research Projects Agency, and University of Maryland Institute for Advanced Computer Studies, January 28-29, 2014. http://www.umiacs.umd.edu/conferences/fiva/

security, crime prevention, traffic control, accident prediction and detection, and monitoring patients, the elderly, and children at home.[79] Depending on the application, use of video surveillance is at varying levels of deployment.[80]

Additional capabilities of image recognition include

- Video summarization and scene-change detection (that is, picking the small number of images that summarize a period of time)
- Precise geolocation in imagery from satellites or drones
- Image-based biometrics
- Human-in-the-loop surveillance systems
- Re-identification of persons and vehicles, that is, tracking the same person or vehicle as it moves from sensor to sensor
- Human-activity recognition of various kinds
- Semantic summarization (that is, converting pictures into text summaries)

Although systems are expected to become able to track objects across camera views and detect unusual activities in a large area by combining information from multiple sources, re-identification of objects remains hard to do (a challenge for inter-camera tracking), as is video surveillance in crowded environments.

Although the data they use are often captured in public areas, scene-extraction technologies like Google Street View have triggered privacy concerns. Photos captured for use in Street View may contain sensitive information about people who are unaware they are being observed and photographed.[81]

Social-media data can be used as an input source for scene extraction techniques. When these data are posted, however, users are unlikely to know that their data would be used in these aggregated ways and that their social media information (although public) might appear synthesized in new forms.[82]

Automated speech recognition has existed since at least the 1950s,[83] but recent developments over the last 10 years have allowed for novel new capabilities. Spoken text (e.g., news broadcasters reading part of a document) can today be recognized with accuracy higher than 95 percent using state-of-the-art techniques. Spontaneous speech is much harder to recognize accurately. In recent years there has been a dramatic increase in the corpuses of spontaneous speech data available to researchers, which has allowed for improved accuracy.

---

[79] For example, Newark Airport recently installed a system of 171 LED lights (from Sensity [http://www.sensity.com/]) that contain special chips to connect to sensors and cameras over a wireless system. These systems allow for advanced automatic lighting to improve security in places like parking garages, and in doing so capture a large range of information.

[80] This was discussed at the workshop cited in footnote 78.

[81] Such concerns are likely to grow as commercial satellite imagery systems such as Skybox (http://skybox.com/) provide the basis for more services.

[82] Billitteri, Thomas J., et al. "Social Media Explosion: Do social networking sites threaten privacy rights?" *CQ Researcher*, January 25, 2013, 23:84-104.

[83] Juang, B.H. and Lawrence R. Rabiner, "Automated Speech Recognition – A Brief History of the Technology Development," October 8, 2004. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

Over the next few years speech-recognition interfaces will be in many more places. For example, multiple companies are exploring speech recognition to control televisions and cars, to find a show on TV, or to schedule a DVR recording. Researchers at Nuance say they are actively planning how speech technology would have to be designed to be available on wearable computers.[84] Google has already implemented some of this basic functionality in its Google Glass product, and Microsoft's Xbox One system already integrates machine vision and multi-microphone audio input for controlling system functions.

3.2.4 Social-network analysis

Social-network analysis refers to the extraction of information from a variety of interconnecting units under the assumption that their relationships are important and that the units do not behave autonomously.[85] Social networks often emerge in an online context. The most obvious examples are dedicated online social media platforms, such as Facebook, LinkedIn and Twitter, which provide new access to social interaction by allowing users to connect directly with each other over the Internet to communicate and share information. Offline human social networks may also leave analyzable digital traces, such as in phone-call metadata records that record which phones have exchanged calls or texts, and for how long. Analysis of social networks is increasingly enabled by the rising collection of digital data that links people together, especially when it is correlated to other data or metadata about the individual.[86] Tools for such analysis are being developed and made available,[87] motivated in part by the growing amount of social network content accessible through open application-programming interfaces to online social-media platforms. This sort of analysis is an active arena for research.

Social-network analysis complements analysis of conventional databases, and some of the techniques used (e.g., clustering in association networks) can be used in either context. Social-network analysis can be more powerful because of the easy association of diverse kinds of information (i.e., considerable data fusion is possible). It lends itself to visualization of the results, which aids in interpreting the results of the analysis. It can be used to learn about people through their association with others, in a context of people's tendency to associate with others who are have some similarities to themselves.[88]

Social-network analysis is yielding results that may surprise people. In particular, unique identification of an individual is easier than from database analysis alone. Moreover, it is achieved through more diverse kinds of

---

[84] "Where Speech Recognition is Going," *Technology Review*, May 29, 2012. http://www.kurzweilai.net/where-speech-recognition-is-going

[85] Wasserman, S. "Social network analysis: Methods and applications," *Cambridge University Press*, 8, 1994.

[86] See, for example: (1) Backstrom, Lars, et al., "Inferring Social Ties from Geographic Coincidences," *Proceedings of the National Academy of Sciences*, 2010. (2) Backsrom, Lars, et al., "Wherefore Art Though R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," *International World Wide Web Conference* 2007, Alberta, Canada, May 12, 2007.

[87] A variety of tools exist for managing, analyzing, visualizing and manipulating network (graph) datasets, such as Allegrograph, GraphVis, R, visone and Wolfram Alpha. Some, such as Cytoscape, Gephi and Netviz are open source.

[88] (1) Geetoor, L. and E. Zheleva, "Preserving the privacy of sensitive relationships in graph data," *Privacy, security, and trust in KDD*, 153-171, 2008. (2) Mislove, A., et al., "An analysis of social-based network Sybil defenses," *ACM SIGCOMM Computer Communication Review*, 2011. (3) Backstrom, Lars, et al., "Find Me If You Can: Improving Geographic Prediction with Social and Spatial Proximity," *Proceedings of the 19th international conference on World Wide Web,* 2010. (4) Backstrom, L. and J. Kleinberg, "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook," *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW), 2014.

data than many people may understand, contributing to the erosion of anonymity.[89] The structure of an individual's network is unique and itself serves as an identifier; co-occurrence in time and space is a significant means of identification; and, as discussed elsewhere in this report, different kinds of data can be combined to foster identification.[90]

Social-network analysis is used in criminal forensic investigations to understand the links, means, and motives of those who may have committed crimes. In particular, social-network analysis has been used to better understand covert terrorist networks, whose dynamics may be different from those of overt networks.[91]

In the realm of commerce, it is well-understood that what a person's friends like or buy can influence what he or she might buy. For example, in 2010, it was reported that having one iPhone-owning friend makes a person three times more likely to own an iPhone than otherwise. A person with two iPhone-owning friends was five times more likely to have one.[92] Such correlations emerge in social-network analysis and can be used to help predict product trends, tailor marketing campaigns towards products an individual may be more likely to want, and target customers (said to have higher "network value") with a central role (and a large amount of influence) in a social network.[93]

Because disease is commonly spread via direct contact between individuals (humans or animals), understanding social networks through whatever proxies are available can suggest possible direct contacts and thereby assist in monitoring and stemming the outbreak of disease.

A recent study by researchers at Facebook analyzed the relationship between geographic location of individual users and that of their friends. From this analysis, they were able to create an algorithm to predict the location of an individual user based upon the locations of a small number of friends in their network, with higher accuracy than simply looking at the user's IP address.[94]

There are many commercial "social listening" services, such as Radian6/Salesforce Cloud, Collective Intellect, Lithium, and others, that mine data from social-networking feeds for use in business intelligence.[95] Coupled

---

[89] (1) Narayanan, A. and V. Shmatikov, "De-anonymizing social networks," *30th IEEE Symposium on Security and Privacy*, 173-187, 2009. (2) Crandall, David J., et al., "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences,* 107:52, 2010. (3) Backstrom, L, C. Dwork and J. Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," *Proceedings of the 16th Intl. World Wide Web Conference*, 2007. (4) Saramäki, Jari, et al., "Persistence of social signatures in human communication," *Proceedings of the National Academy of Sciences,* 111.3:942-947, 2014.

[90] Fienberg, S.E., "Is the Privacy of Network Data an Oxymoron?" *Journal of Privacy and Confidentiality,* 4:2, 2013.

[91] Krebs, V.E., "Mapping networks of terrorist cells," *Connections,* 24.3:43-52, 2002.

[92] Sundsøy, P. R., et al., "Product adoption networks and their growth in a large mobile phone network," *Advances in Social Networks Analysis and Mining (ASONAM)*, 2010.

[93] Hodgson, Bob, "A Vital New Marketing Metric: The Network Value of a Customer," *Predictive Marketing: Optimize Your ROI With Analytics.* http://predictive-marketing.com/index.php/a-vital-new-marketing-metric-the-network-value-of-a-customer/

[94] Backstrom, Lars et al, "Find me if you can: improving geographical prediction with social and spatial proximity," *Proceedings of the 19th international conference on World Wide Web*, 2010.

[95] "Top 20 social media monitoring vendors for business," *Socialmedia.biz*, http://socialmedia.biz/2011/01/12/top-20-social-media-monitoring-vendors-for-business/

with social-network analysis, this information can be used to evaluate changing influences and the spread of trends between individuals and communities to inform marketing strategies.

3.3 The infrastructure behind big data

Big-data analytics requires not just algorithms and data, but also physical platforms where the data are stored and analyzed. The related security services used for personal data (see Sections 4.1 and 4.2) are also an essential component of the infrastructure. Once available only to large organizations, this class of infrastructure is now available through "the cloud" to small businesses and to individuals. To the extent that the software infrastructure is widely shared, privacy-preserving infrastructure services can also be more readily used.

3.3.1 Data centers

One way to think about big-data platforms is in physical units of "data centers." In recent years, data centers have become almost standard commodities. A typical data center is a large, warehouse-like building on a concrete slab the size of a few football fields. It is located with good access to cheap electric power and to a fiber-optic, Internet-backbone connection, usually in a rural or isolated area. The typical center consumes 20-40 megawatts of power (the equivalent of a city with 20,000-40,000 residents) and today houses some tens of thousands of servers and hard-disk drives, totaling some tens of petabytes.[96] Worldwide, there are roughly 6000 data centers of this scale, about half in the United States.[97]

Data centers are the physical locus of big data in all its forms. Large data collections are often replicated in multiple data centers to improve both performance and robustness. There is a growing marketplace in selling data-center services.

Specialized software technology allows the data in multiple data centers (and spread across tens of thousands of processors and hard-disk drives) to cooperate in performing the tasks of data analytics, thereby providing both scaling and better performance. For example, MapReduce (originally a proprietary technology of Google, but now a term used generically) is a programming model for parallel operations across a practically unlimited number of processors; Hadoop is a popular open-source programming platform and program library based on the same ideas; NoSQL (the name derived from "*not* Structured Query Language") is a set of database technologies that relaxes many of the restrictions of traditional, "relational" databases and allows for better scalability across the many processors in one or more data centers. Contemporary research is aimed at the next generation beyond Hadoop. One path is represented by Accumulo, initiated by the National Security Agency and transitioned to the open-source Apache community.[98] Another is the Berkeley Data Analytics Stack, an open-source platform that outperforms Hadoop by a factor of 100 for memory-intensive data analytics and is being used by such companies as Foursquare, Conviva, Klout, Quantifind, Yahoo, and Amazon Web Services.[99] Sometimes termed "NoHadoop" (to parallel the movement from SQL to NoSQL), technologies that fit this trend include Google's Dremel, MPI (typically used in supercomputing), Pregel (for graphs), and Cloudscale (for real-time analytics).

---

[96] A petabyte is 10^{15} bytes. One petabyte could store the individual genomes of the entire U.S. population. The human brain has been estimated to have a capacity of 2.5 petabytes.

[97] McLellan, Charles, "The 21st Century Data Center: An Overview," *ZDNet*, April 2, 2013. http://www.zdnet.com/the-21st-century-data-center-an-overview-7000012996/

[98] See: http://accumulo.apache.org/

[99] See: https://amplab.cs.berkeley.edu/software/

3.3.2 The cloud

The "cloud" is not just the world inventory of data centers (although much of the public may think of it as such). Rather, one way of understanding the cloud is as a set of platforms and services *made possible* by the physical commoditization of data centers. When one says that data are "in the cloud," one refers not just to the physical hard-disk drives that exist (somewhere!) with the data, but also to the complex infrastructure of application programs, middleware, networking protocols, and (not least) business models that allow that data to be ingested, accessed, and utilized, all with costs that are competitively allocated. The commercial entities that, in aggregate, provision the cloud exist in an ecosystem that has many hierarchical levels and many different coexisting models of value added. There may be several handoffs of responsibility between the end user and the physical data center.

Today's cloud providers offer some security benefits (and through that, privacy benefits) as compared to yesterday's conventional corporate data centers or small-business computers.[100] These services may include better physical protection and monitoring, as well as centralized support staffing, training, and oversight. Cloud services also pose new challenges for security, a subject of current research. Both benefits and risks come from the centralization of resources: More data are held by a given entity (albeit distributed across multiple servers or sites), and a cloud provider can perform better than separately held data centers by applying high standards to recruiting and managing people and systems.

Usage of the cloud and individual interactions with it (whether witting or not) are expected to increase dramatically in coming years. The rise of both mobile apps,[101] reinforcing the use of cell phones and tablets as platforms, and broadly distributed sensors is associated with the growing use of cloud systems for storing, processing, and otherwise acting on information contributed by dispersed devices. Although progress in the mobile environment improves the usability of mobile cloud applications, it may be detrimental to privacy to the extent that it more effectively hides information exchange from the user. As more core mobile functionality is transitioned to the cloud, larger amounts of information will be exchanged, and users may be surprised by the nature of the information that no longer remains localized to their cell phone. For example, cloud-based screen rendering (or "virtualized screens") for cell phones would mean that the images shown on a cell-phone screen will actually be calculated on the cloud and transmitted to the mobile device. This means all the images on the screen of the mobile device can be accessed and manipulated from the cloud.

Cloud architectures are also being used increasingly to support big-data analytics, both by large enterprises (e.g., Google, Amazon, eBay) and by small entities or individuals who make ad hoc or routine use of public cloud platforms (e.g., Amazon Web Services, Google Cloud Platform, Microsoft Azure) in lieu of acquiring their own infrastructure. Social-media services such as Facebook and Twitter are deployed and analyzed by their providers using cloud systems. These uses represent a kind of democratization of analytics, with the potential to facilitate new businesses and more. Prospects for the future include exploration of options for federating or

[100] Cloud Security Alliance, "Big Data Working Group: Comment on Big Data and the Future of Privacy," March 2014. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Comment_on_Big_Data_Future_of_Privacy.pdf

[101] Qi, H. and A. Gani, "Research on mobile cloud computing: Review, trend and perspectives," *Digital Information and Communication Technology and it's Applications (DICTAP), 2012 Second International Conference on*, 2012.

interconnecting cloud applications and for reducing some of the heterogeneity in application-programming interfaces for cloud applications.[102]

[102] Jeffery, K. et al., "A vision for better cloud applications," *Proceedings of the 2013 International Workshop on Multi-Cloud Applications and Federated Clouds*, Prague, Czech Republic, MODAClouds, ACM Digital Library, April 22-23, 2013.

4. Technologies and Strategies for Privacy Protection

Data come into existence, are collected, and are possibly processed immediately (including adding "metadata"), possibly communicated, possibly stored (locally, remotely, or both), possibly copied, possibly analyzed, possibly communicated to users, possibly archived, possibly discarded. Technology at any of these stages can affect privacy positively or negatively.

This chapter focuses on the positive and assesses some of the key technologies that can be used in service of the protection of privacy. It seeks to clarify the important distinctions between privacy and (cyber-)security, as well as the vital, but yet limited, role that encryption technology can play. Some older techniques, such as anonymization, while valuable in the past, are seen as having only limited future potential. Newer technologies, some entering the marketplace and some requiring further research, are summarized.

4.1 The relationship between cybersecurity and privacy

Cybersecurity is a discipline, or set of technologies, that seeks to enforce policies relating to several different aspects of computer use and electronic communication.[103] A typical list of such aspects would be

- identity and authentication: Are you who you say you are?
- authorization: What are you allowed to do?
- availability: Can attackers interfere with authorized functions?
- confidentiality: Can data or communications be (passively) copied by someone not authorized to do so?
- integrity: Can data or communications be (actively) changed or manipulated by someone not authorized?
- non-repudiation, auditability: Can actions (payments may provide the best example) later be shown to have occurred?

Good cybersecurity enforces policies that are precise and unambiguous. Indeed, such clarity of policy, expressible in mathematical terms, is a necessary prerequisite for the Holy Grail of cybersecurity," provably secure" systems. At present, provable security exists only in very limited domains, for example, for certain functions on some kinds of computer chips. It is a goal of cybersecurity research to extend the scope of provably secure systems to larger and larger domains. Meanwhile, practical cybersecurity draws on the emerging principles of such research, but it is guided even more by practical lessons learned from known failures of cybersecurity. The realistic goal is that the practice of cybersecurity should be continuously improving so as to be, in most places and at most of the time, ahead of the evolving threat.

Poor cybersecurity is clearly a threat to privacy. Privacy can be breached by failure to enforce confidentiality of data, by failure of identity and authentication processes, or by more complex scenarios such as those compromising availability.

[103] PCAST has addressed issues in cybersecurity, both in reviewing the NITRD programs and directly in a 2013 report, *Immediate Opportunities for Strengthening the Nation's Cybersecurity*. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_cybersecurity_nov-2013.pdf

Security and privacy share a focus on malice. The security of data can be compromised by inadvertence or accident, but it can also be compromised because some party acted knowingly to achieve the compromise – in the language of security, committed an attack. Substituting the words "breach" or "invasion" for "compromise" or "attack," the same concepts apply to privacy.

Even if there were perfect cybersecurity, however, privacy would remain at risk. Violations of privacy are possible even when there is no failure in computer security. If an authorized individual chooses to misuse (e.g., disclose) data, what is violated is privacy policy, not security policy. Or, as we have discussed (see Section 3.1.1), privacy may be violated by the fusion of data – even if performed by authorized individuals on secure computer systems.[104]

Privacy is different from security in other respects. For one thing, it is harder to codify privacy policies precisely. Arguably this is because the presuppositions and preferences of human beings have greater diversity than the useful scope of assertions about computer security. Indeed, how to codify human privacy preferences is an important, nascent area of research.[105]

When people provide assurance (at some level) that a computer system is secure, they are saying something about applications that are not yet invented: They are asserting that technological design features already in the machine today will prevent such application programs from violating pertinent security policies in that machine, even tomorrow.[106] Assurances about privacy are much more precarious. Since not-yet-invented applications will have access to not-yet-imagined new sources of data, as well as to not-yet-discovered powerful algorithms, it much harder to provide, today, technological safeguards against a new route to violation of privacy tomorrow. Security deals with tomorrow's threats against today's platforms. That is hard enough. But privacy deals with tomorrow's threats against *tomorrow's* platforms, since those "platforms" comprise not just hardware and software, but also new kinds of data and new algorithms.

Computer scientists often work from the basis of a formal policy for security, just as engineers aim to describe something explicitly so that they can design specific ways to deal with it by purely technical means. As more computer scientists begin to think about privacy, there is increasing attention to formal articulation of privacy policy.[107] To caricature, you have to know what you are doing to know whether what you are doing is doing the right thing.[108] Research addressing the challenges of aligning regulations and policies with software

---

[104] There are also choices in the design and implementation of security mechanisms that affect privacy. In particular, authentication or the attempt to demonstrate identity at some level can be done with varying degrees of disclosure. See, for example: Computer Science and Telecommunications Board, *Who Goes There: Authentication Through the Lens of Privacy*, National Academies Press, 2003.

[105] Such research can inform efforts to automate the checking of compliance with policies and/or associated auditing.

[106] This future-proofing remains hard to achieve; PCAST's cybersecurity report advocated approaches that would be more durable than the kinds of check-lists that are easily rendered obsolete. See: http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_cybersecurity_nov-2013.pdf

[107] See, for example: (1) Breaux, Travis D., and Ashwini Rao, "Formal Analysis of Privacy Requirements Specifications for Multi-Tier Applications," *21st IEEE Requirements Engineering Conference* (RE 2013), Rio de Janeiro, Brazil, July 2013. http://www.cs.cmu.edu/~agrao/paper/Analysis_of_Privacy_Requirements_Facebook_Google_Zynga.pdf (2) Feigenbaum, Joan, et al., "Towards a Formal Model of Accountability," *New Security Paradigms Workshop 2011*, Marin County, CA, September 12-15, 2011. http://www.nspw.org/papers/2011/nspw2011-feigenbaum.pdf

[108] Landwehr, Carl, "Engineered Controls for Dealing with Big Data," Chapter 10, in Lane, Julia, et al., *Privacy, Big Data, and the Public Good,* Cambridge University Press, 2014.

specifications includes formal languages to express policies and system requirements; tools to reason about conflicts, inconsistencies, and ambiguities within and among policies and software specifications; methods to enable requirements engineers, business analysts, and software developers to analyze and refine policy into measurable system specifications that can be monitored over time; formalizing and enforcing privacy through auditing and accountability systems; privacy compliance in big-data systems; and formalizing and enforcing purpose restrictions.

4.2 Cryptography and encryption

Cryptography comprises a set of algorithms and system-design principles, some well-developed and others nascent, for protecting data. Cryptography is a field of knowledge whose products are encryption technology. With well-designed protocols, encryption technology is an inhibitor to compromising privacy, but it is not a "silver bullet."[109]

4.2.1 Well established encryption technology

Using cryptography, readable data of any kind, termed plaintext, are transformed into what are, for all intents and purposes, incomprehensible strings of provably random bits, so-called cryptotext. Cryptotext requires no security protection of any kind. It can be stored in the cloud or sent anywhere that is convenient. It can be sent promiscuously to both the NSA and Russian FSB. If they have only cryptotext – and if it was properly generated in a precise mathematical sense – it is useless to them. They can neither read the data nor compute with it. What is needed to decrypt, to turn cryptotext back into the original plaintext, is a "key," which is in practice a string of bits that is supposed to be known to (or computable by) only authorized users. Only with the key can encrypted data be used, i.e., their value read.

In the context of protecting privacy, it is primarily not the cryptography that is of concern.[110] Rather, compromises of data will occur in one of two main ways:

- Data can be stolen, or mistakenly shared, before they have been encrypted or after they have been decrypted. Many attacks on supposedly encrypted data are actually attacks on machines that contain – however briefly – unencrypted plaintext. For example, in Target's 2013 breach of one hundred million debit card number and personal-identification numbers (PINs), the PINs were present in unencrypted form only ephemerally. They were stolen nonetheless.[111]
- Keys must be authorized, generated, distributed, and used. At every stage of a key's life, it is potentially open to compromise or misuse that can ultimately compromise the data that the key was intended to protect. No system based on encryption is secure, of course, if persons with access to private keys can be coerced into sharing them.

[109] The use of this term in computing originated with what is now viewed as a classic article: Brooks, Fred P., "No silver bullet – Essence and Accidents of Software Engineering", *IEEE Computer* 20:4, April 1987, pp. 10-19.

[110] Attacks that compromise the hardware or software that does the encrypting (for example, the promulgation of intentionally weak cryptography standards) can be considered to be a variant of attacks that reveal plaintext.

[111] "Krebs on Security, collected posts on Target data breach," 2014. http://krebsonsecurity.com/tag/target-data-breach/

Until the 1970s, keys were distributed physically, on paper or computer media, protected by registered mail, armed guards, or anything in between. The invention of "public-key cryptography"[112] changed everything. Public-key cryptography, as the name implies, allows individuals to broadcast publicly their personal key. But this public key is only an encryption key, useful for turning plaintext into cryptotext that is meaningless to others. Its corresponding "private key," used to transform cryptotext to plaintext, is still kept secret by the recipient. Public-key cryptography thus turns the problem of key distribution into a problem of identity determination. Alice's messages (encrypted data transmissions) to Bob are completely protected by Bob's public key – but only if Alice is certain that it is really *Bob's* public key that she is using, and not the public key of someone merely masquerading as Bob.

Luckily, public-key cryptography also provides some techniques for helping to establish identity, namely the electronic "signing" of messages to document their authenticity. Electronic signatures, in turn, enable messages of the form "I, a person of authority known as X, certify that the following is really the public key of subordinate person Y. (Signed) X." Messages like this are termed certificates. Certificates can be cascaded, with A certifying the identity of B, who certifies C, and so on. Certificates essentially transform the identity problem from one of validating the identity of millions of possible Y's to validating the identity of much smaller number of top-level certificate authorities (CAs). Yet it is a matter of concern that more than 100 top-level CAs are widely recognized (e.g., accepted by most all web browsers), because there may be several intermediate steps in the hierarchy of certificates from a CA to a user, and at every step a private key must be protected by some signer on some computer. The compromise of this private key potentially compromises the privacy of all users lower down the chain – because forged certificates of identity can now be created. Such exploits have been seen. For example, the 2011 apparent theft of a Dutch CA's private key compromised the privacy of potentially all government records in the Netherlands.[113,114]

Many major companies have recently introduced or strengthened their use of encryption to transmit data.[115] Some are now using "(perfect) forward secrecy," a variant of public-key cryptography that ensures that the compromise of an individual's private key can compromise only messages that he receives subsequently, while the confidentiality of past conversations is maintained, even if their cryptotext was previously recorded by the same eavesdropper now in possession of the purloined private key.[116]

4.2.2 Encryption frontiers

The technologies thus far mentioned enable the protection of data both in storage and in transit, allowing those data to be fully decrypted by users who either (i) have the right key already (as might be the case for persons

[112] Public-key encryption originated through the secret work of British mathematicians at the U.K.'s Government Communications Headquarters (GCHQ), an organization roughly analogous to the NSA, and received broader attention through the independent work by researchers including Whitfield Diffie and Martin Hellman in the United States.

[113] Fisher, Dennis, "Final Report on DigiNotar Hack Shows Total Compromise of CA Servers," *ThreatPost*, October 31, 2012. http://threatpost.com/final-report-diginotar-hack-shows-total-compromise-ca-servers-103112/77170.

[114] It is not publicly known whether or not the earlier 2010 compromise of servers belonging to VeriSign, a much larger CA, led to compromises of certificates or signing authorities. Bradley, Tony, "VeriSign Hacked: What We Don't Know Might Hurt Us," *PC World*, February 2, 2012. http://www.pcworld.com/article/249242/verisign_hacked_what_we_dont_know_might_hurt_us.html

[115] A sample report-card: https://www.eff.org/deeplinks/2013/11/encrypt-web-report-whos-doing-what#crypto-chart

[116] Diffie, Whitfield, et al., "Authentication and Authenticated Key Exchanges" *Designs, Codes and Cryptography* 2:2, June 1992, pp.107-125.

storing data for their own later use), or (ii) are authorized by the data owner and have identities certified by a CA that is itself trusted by the data owner. A frontier of cryptography research, with some inventions now starting to make it into practice, is how to create different kinds of keys, ones which give only limited access of various kinds, or which allow messages to be sent to classes of individuals without knowing in advance exactly who they may be.

For example, "identity-based encryption" and "attribute-based encryption" are ways of sending a message, or protecting a file of data, for the exclusive use of "a person named Ramona Q. Doe who was born on May 23, 1980," or for "anyone with the job title ombudsman, ombudsperson, or consumer advocate." These techniques require a trusted third party (essentially a certificate authority), but the messages themselves do not need to pass through the hands of that third party. These tools are in early stages of adoption.

"Zero-knowledge" systems allow encrypted data to be queried for certain higher-level abstractions without revealing the low-level data. For example, a website operator could verify that a user is over age 21 without learning the user's actual birthdate. What is remarkable is that this can be done in a way that proves mathematically that the user is not lying about his age: The operator learns with mathematical certainty that a certificate (signed by some CA of course!) attests to the user's birthdate, without ever actually seeing that certificate. Zero-knowledge systems are just beginning to be commercialized in simple cases. They are not foreseeably extendable to complex and unstructured situations, such as what might be needed for the research mining of health-record data from non-consenting patients.

In some simpler domains, for example location privacy, practical cryptographic protection is closer to reality. The typical case might be that a group of friends want to know when they are close to one another, but without sharing their actual locations with any third party. Applications like this are, of course, much simpler if there is a trusted third party, as is *de facto* the case for most such commercial applications today.

Homomorphic encryption is a research area that goes beyond the mere querying of encrypted databases to actual computations (e.g., the collection of statistics) using encrypted data without ever decrypting it. These techniques are far from being practical, and they are unlikely to provide policy options on the timescale relevant to this report.

In secure multi-party computation, which is related to homomorphic encryption and is of particular interest in the financial sector, computation may be done on distributed data stores that are encrypted. Although individual data are kept private using "collusion-robust" encryption algorithms, data can be used to calculate general statistics. Parties that each know some private data use a protocol that generates useful results based on both information they know and information they do not know, without revealing to them data they do not already know.

Differential privacy, a comparatively new development related to but different from encryption, aims to maximize the accuracy of database queries or computations while minimizing the identifiability of individuals with records in the database, typically via obfuscation of query results (for example, by the addition of spurious information or "noise").[117] As with other obfuscation approaches, there is a tradeoff between data anonymity

[117] (1) Dwork, Cynthia, "Differential Privacy," 33rd International Colloquium on Automata, Languages and Programming, 2006. (2) Dwork, Cynthia, "A Firm Foundation for Private Data Analysis," *Communications of the ACM*, 54.1, 2011.

and the accuracy and utility of the query outputs. These ideas are far from practical application, except insofar as they may enable the risks of allowing any queries at all to be better assessed.

4.3 Notice and consent

Notice and consent is, today, the most widely used strategy for protecting consumer privacy. When the user downloads a new app to his or her mobile device, or when he or she creates an account for a web service, a notice is displayed, to which the user must positively indicate consent before using the app or service. In some fantasy world, users actually read these notices, understand their legal implications (consulting their attorneys if necessary), negotiate with other providers of similar services to get better privacy treatment, and only then click to indicate their consent. Reality is different.[118]

Notice and consent fundamentally places the burden of privacy protection on the individual – exactly the opposite of what is usually meant by a "right." Worse yet, if it is hidden in such a notice that the provider has the right to share personal data, the user normally does not get any notice from the next company, much less the opportunity to consent, even though use of the data may be different. Furthermore, if the provider changes its privacy notice for the worse, the user is typically not notified in a useful way.

As a useful policy tool, notice and consent is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data. It is simply too complicated for the individual to make fine-grained choices for every new situation or app. Nevertheless, since notice and consent is so deeply rooted in current practice, some exploration of how its usefulness might be extended seems warranted.

One way to view the problem with notice and consent is that it creates a non-level playing field in the implicit privacy negotiation between provider and user. The provider offers a complex take-it-or-leave-it set of terms, backed by a lot of legal firepower, while the user, in practice, allocates only a few seconds of mental effort to evaluating the offer, since acceptance is needed to complete the transaction that was the user's purpose, and since the terms are typically difficult to comprehend quickly. This is a kind of market failure. In other contexts, market failures like this can be mitigated by the intervention of third parties who are able to represent significant numbers of users and negotiate on their behalf. Section 4.5.1 below suggests how such intervention might be accomplished.

4.4 Other strategies and techniques

4.4.1 Anonymization or de-identification

Long used in health-care research and other research areas involving human subjects, anonymization (also termed de-identification) applies when the data, standing alone and without an association to a specific person, do not violate privacy norms. For example, you may not mind if your medical record is used in research as long as you are identified only as Patient X and your actual name and patient identifier are stripped from that record.

Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data. In

[118] Gindin, Susan E., "Nobody Reads Your Privacy Policy or Online Contract: Lessons Learned and Questions Raised by the FTC's Action against Sears," *Northwestern Journal of Technology and Intellectual Property* 1:8, 2009-2010.

general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially.[119]

One compelling example comes from Sweeney, Abu, and Winn.[120] They showed in a recent paper that, by fusing public, Personal Genome Project profiles containing zip code, birthdate, and gender with public voter rolls, and mining for names hidden in attached documents, 84-97 percent of the profiles for which names were provided were correctly identified.

Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods. PCAST does not see it as being a useful basis for policy. Unfortunately, anonymization is already rooted in the law, sometimes giving a false expectation of privacy where data lacking certain identifiers are deemed not to be personally identifiable information and therefore not covered by such laws as the Family Educational Rights and Privacy Act (FERPA).

4.4.2 Deletion and non-retention

It is an evident good business practice that data of all kinds should be deleted when they are no longer of value. Indeed, well-run companies often mandate the destruction of some kinds of records (both paper and electronic) after specified periods of time, often because they see little benefit in keeping the records as well as potential cost in producing them. For example, employee emails, which may be subject to legal process by (e.g.) divorce lawyers, are often seen as having negative retention value.

Counter to this practice is the new observation that big data is frequently able to find economic or social value in masses of data that were otherwise considered to be worthless. As the physical cost of retention continues to decrease exponentially with time (especially in the cloud), there will be a tendency in both government and the private sector to hold more data for longer – with obvious privacy implications. Archival data may also be important to future historians, or for later longitudinal analysis by academic researchers.

Only policy interventions will counter this trend. Government can mandate retention policies for itself. To affect the private sector, government may mandate policies where it has regulatory authorities (as for consumer protection, for example). But it can also encourage the development of stricter liability standards for companies whose data, including archived data, cause harm to individuals. A rational response by the private sector would then be to hold fewer data or to protect their use.

The above holds true for privacy-sensitive data about individuals that are held overtly – that is, the holder knows that he has the data and to whom they relate. As was discussed in Section 3.1.2, however, sources of data increasingly contain latent information about individuals, information that becomes known only if the holder expends analytic resources (beyond what may be economically feasible), or that may become knowable only in the future with the development of new data-mining algorithms. In such cases it is practically impossible for the data holder even to surface "all the data about an individual," much less delete those data on any specified schedule.

---

[119] De-identification can also be seen as a spectrum, rather than a single approach. See: "Response to Request for Information Filed by U.S. Public Policy Council of the Association for Computing Machinery," March 2014.

[120] Sweeney, et al., "Identifying Participants in the Personal Genome Project by Name," *Harvard University Data Privacy Lab*. White Paper 1021-1, April 24, 2013. http://dataprivacylab.org/projects/pgp/

The concepts of ephemerality (keeping data only on-the-fly or for a brief period), and transparency (enabling the individual to know what data about him or her are held) are closely related, and with the same practical limitations. While data that are only streamed, and not archived, may have lower risk of future use, there is no guarantee that a violator will play by the supposed rules, as in Target's loss of 100 million debit card PINs, each present only ephemerally (see Section 4.2.1).

Today, given the distributed and redundant nature of data storage, it is not even clear that data *can* be destroyed with any useful degree of assurance. Although research on data destruction is ongoing, it is a fundamental fact that at the moment that data are displayed (in "analog") to a user's eyeballs or ears, they can also be copied ("re-digitized") without any technical protections. The same holds if data are ever made available in unencrypted form to a rogue computer program, one designed to circumvent technical safeguards. Some misinformed public discussion notwithstanding, there is no such thing as automatically self-deleting data, other than in a fully controlled and rule-abiding environment.

As a current example, SnapChat provides the service of delivering ephemeral snapshots (images), visible for only a few seconds, to a designated recipient's mobile device. SnapChat promises to delete past-date snaps from their servers, but it is only a promise. And, they are careful *not* to promise that the intended recipient may not contrive to make an uncontrolled and non-expiring copy. Indeed, the success of SnapChat incentivizes the development of just such copying applications.[121]

From a policymaking perspective, the only viable assumption today, and for the foreseeable future, is that data, once created, are permanent. While their *use* may be regulated, their continued *existence* is best considered conservatively as unalterable fact.

4.5 Robust technologies going forward

4.5.1 A Successor to Notice and Consent

The purpose of notice and consent is that the user assents to the collection and use of personal data for a stated purpose that is acceptable to that individual. Given the large number of programs and Internet-available devices, both visible and not, that collect and use personal data, this framework is increasingly unworkable and ineffective. PCAST believes that the responsibility for using personal data in accordance with the user's preferences should rest with the provider, possibly assisted by a mutually accepted intermediary, rather than with the user.

How might that be accomplished? Individuals might be encouraged to associate themselves with one of a standard set of privacy preference profiles (that is, settings or choices) voluntarily offered by third parties. For example, Jane might choose to associate with a profile offered by the American Civil Liberties Union that gives particular weight to individual rights, while John might associate with one offered by *Consumer Reports* that gives weight to economic value for the consumer. Large app stores (such as Apple App Store, Google Play, Microsoft Store) for whom reputational value is important, or large commercial sectors such as finance, might choose to offer competing privacy-preference profiles.

[121] See, for example: Ryan Whitwam, "Snap Save for iPhone Defeats the Purpose of Snapchat, Saves Everything Forever," *PC Magazine*, August 12, 2013. http://appscout.pcmag.com/apple-ios-iphone-ipad-ipod/314653-snap-save-for-iphone-defeats-the-purpose-of-snapchat-saves-everything-forever

In the first instance, an organization offering profiles would vet new apps as acceptable or not acceptable within each of their profiles. Basically, they would do the close reading of the provider's notice that the user should, but does not, do. This is not as onerous as it may sound: While there are millions of apps, the most popular downloads are relatively few and are concentrated in a relatively small number of portals. The "long tail" of apps with few customers each might initially be left as "unrated."

Simply by vetting apps, the third-party organizations would automatically create a marketplace for the negotiation of community standards for privacy. To attract market share, providers (especially smaller ones) could seek to qualify their offerings in as many privacy-preference profiles, offered by as many different third parties, as they deem feasible. The Federal government (e.g., through the National Institute of Standards and Technology) could encourage the development of standard, machine-readable interfaces for the communication of privacy implications and settings between providers and assessors.

Although human professionals could do the vetting today using policies expressed in natural language, it would be desirable in the future to automate that process. To do that, it would be necessary to have formalisms to specify privacy policies and tools to analyze software to determine conformance to those policies. But that is only part of the challenge. A greater challenge is to make sure the policy language is sufficiently expressive, the policies are sufficiently rich, and conformance tests are sufficiently powerful. Those requirements lead to a consideration of context and use.

4.5.2 Context and Use

The previous discussion, particularly that of Sections 3.1 and 3.2, illustrates PCAST's belief that a focus on the collection, storage, and retention of electronic personal data will not provide a technologically robust foundation on which to base future policy. Among the many authors that have touched on these issues, Kagan and Abelson explain why access control does not suffice to protect privacy.[122] Mundie gives a cogent and more complete explanation of this issue and advocates that privacy protection is better served by controlling the use of personal data, broadly construed, including metadata and data derived from analytics than by controlling collection.[123] In a complementary vein, Nissenbaum explains that both the context of usage and the prevailing social norms contribute to acceptable use.[124]

To implement in a meaningful way the application of privacy policies to the use of personal data for a particular purpose (i.e., in context), those policies need to be associated both with data and with the code that operates on the data. For example, it must be possible to ensure that only apps with particular properties can be applied to certain data. The policies might be expressed in what computer scientists call natural language (plain English or the equivalent) and the association done by the user, or the policies might be stated formally and their association and enforcement done automatically. In either case, there must also be policies associated with the outputs of the computation, since they are data as well. The privacy policies of the output data must be computed from the policies associated with the inputs, the policies associated with the code, and the intended use of the outputs (i.e., the context). These privacy properties are a kind of metadata. To achieve a reasonable level of reliability, their implementation must be tamper-proof and "sticky" when data are copied.

---

[122] Abelson, Hal and Lalana Kagal, "Access Control is an Inadequate Framework for Privacy Protection," *W3C Workshop on Privacy for Advanced Web APIs 12/13*, July 2010, London. http://www.w3.org/2010/api-privacy-ws/papers.html

[123] Mundie, Craig, "Privacy Pragmatism: Focus on Data Use, Not Data Collection," *Foreign Affairs*, March/April, 2014.

[124] Nissenbaum, H., "Privacy in Context: Technology, Policy, and the Integrity of Social Life," *Stanford Law Books*, 2009.

There has been considerable research in areas that would contribute to such a capability, some of which is beginning to be commercialized. There is a history of using metadata ("tags" or "attributes") in database systems to control use. While the formalization of privacy policies and their synthesis is a research topic,[125] manual interpretation of such policies and the human determination of usage tags can be found in recent products. Identity management systems (to authenticate users and their roles, i.e., their context) are also evident both in research[126] and in practice.[127]

Commercial privacy systems for implementing use control exist today under the name of Trusted Data Format (TDF) implementations, developed principally for the United States intelligence community.[128] TDF operates at the file level. The systems are primarily being implemented on a custom basis by large consulting firms, often assembled from open-source software components. Customers today are primarily government agencies, such as Federal intelligence agencies or local-government criminal intelligence units, or large commercial companies in vertically integrated industries like financial services and pharmaceutical companies looking to improve their accountability and auditing capabilities. Consulting services that have expertise in building such systems include, for example, Booz Allen, Ernst & Young, IBM, Northrop Grumman, and Lockheed; product-based companies like Palantir and new startups pioneering internal usage auditing, policy analytics, and policy reasoning engines have such expertise, as well. With sufficient market demand, more widespread market penetration could happen in the next five years. Market penetration would be further accelerated if the leading cloud-platform providers like Amazon, Google, and Microsoft implemented usage-controlled system technologies in their offerings. Wider-scale use through the government would help motivate the creation of off-the-shelf standard software.

4.5.3 Enforcement and deterrence

Privacy policies and the control of use in context are only effective to the extent that they are realized and enforced. Technical measures that increase the probability that a violator is caught can be effective only when there are regulations and laws with civil or criminal penalties to deter the violators. Then there is both deterrence of harmful actions and incentive to deploy privacy-protecting technologies.

It is today straightforward technically to associate metadata with data, with varying degrees of granularity ranging from an individual datum, to a record, to an entire collection. These metadata can record a wealth of auditable information, for example, provenance, detailed access and use policies, authorizations, logs of actual access and use, and destruction dates. Extending such metadata to derived or shared data (secondary use) together with privacy-aware logging can facilitate auditing. Although the state of the art is still somewhat ad hoc, and auditing is often not automated, so-called accountable systems are beginning to be deployed (Section

---

[125] See references at footnote 107 and also: (1) Weitzner, D.J., et al., "Information Accountability," *Communications of the ACM,* June 2008, pp. 82-87. (2) Tschantz, Michael Carl, Anupam Datta, and Jeannette M. Wing, "Formalizing and Enforcing Purpose Restrictions in Privacy Policies." http://www.andrew.cmu.edu/user/danupam/TschantzDattaWing12.pdf

[126] For example, at Carnegie Mellon University, Lorrie Cranor directs the CyLab Usable Privacy and Security Laboratory (http://cups.cs.cmu.edu/). Also, see *2nd International Workshop on Accountability: Science, Technology and Policy*, MIT Computer Science and Artificial Intelligence Laboratory, January 29-30, 2014. http://dig.csail.mit.edu/2014/AccountableSystems2014/

[127] Oracle's eXtensible Access Control Markup Language (XACML**)** has been used to implement attribute-based access controls for identity management systems. (Personal communication, Mark Gorenberg and Peter Guerra of Booz Allen)

[128] Office of the Director of National Intelligence, "IC CIO Enterprise Integration & Architecture: Trusted Data Format." http://www.dni.gov/index.php/about/organization/chief-information-officer/trusted-data-format

4.5.2). The ability to detect violations of privacy policies, particularly if the auditing is automated and continuous, can be used both to deter privacy violations and to ensure that violators are punished.

In the next five years, with regulation or market-driven encouragement, the large cloud-based infrastructure systems (e.g., Google, Amazon, Microsoft, Rackspace) could, as one example, incorporate the data-provenance and usage-compliance aspects of accountable systems into their cloud application-programming interfaces (APIs) and additionally provide APIs for policy awareness. These capabilities could then readily be included in open-source-based systems like Open Stack (associated with Rackspace)[129] and other provider platforms. Applications intended to run on such cloud-based systems could be built with privacy concepts "baked into them," even when they are developed by small enterprises or individual developers.

4.5.4 Operationalizing the Consumer Privacy Bill of Rights

In February 2012, the Administration issued a report setting forth a Consumer Privacy Bill of Rights (CPBR). The CPBR addresses commercial (not public sector) uses of personal data and is a strong statement of American privacy values.

For purposes of this discussion, the principles embodied in CPBR can be divided into two categories. First, there are obligations for data holders, analyzers, or commercial users. These are passive from the consumer's standpoint – the obligations should be met whether or not the consumer knows, cares, or acts. Second, and different, there are consumer empowerments, things that the consumer should be empowered to initiate actively. It is useful here to rearrange the CPBR's principles by category.

In the category of obligations are these elements:

- Respect for Context: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- Focused Collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
- Security: Consumers have a right to secure and responsible handling of personal data.
- Accountability: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.

In the category of consumer empowerments are these elements:

- Individual Control: Consumers have a right to exercise control over what personal data companies collect from them and how they use it.
- Transparency: Consumers have a right to easily understandable and accessible information about privacy and security practices.
- Access and Accuracy: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data are inaccurate.

PCAST endorses as sound the principles underlying CPBR. Because of the rapidly changing technologies associated with big data, however, effective operationalization of CPBR is at risk. Up to now, debate over how to operationalize CPBR has focused on the collection, storage, and retention of data, with an emphasis on the

[129] See: http://www.openstack.org/

"small-data" contexts that motivated CPBR development. But, as discussed at multiple places in this report (e.g., Sections 3.1.2, 4.4 and 4.5.2), PCAST believes that such a focus will not provide a technologically robust foundation on which to base future policy that also applies to big data. Further, the increasing complexity of applications and uses of data undermines even a simple concept like "notice and consent."

PCAST believes that the principles of CPBR can readily be adapted to a more robust regime based on recognizing and controlling harmful uses of the data. Some specific suggestions follow.

Turn first to the rights classified above as obligations on the data holder.

The principle of Respect for Context needs augmentation. As this report has repeatedly discussed, there are instances in which personal data are not provided by the customer. Such data may emerge as a product of analysis well after the data were collected and after they may have passed through several hands. While the intent of the right is appropriate, namely that data be used for legitimate purposes that do not produce certain adverse consequences or harms to individuals, the CPBR's articulation in which "consumers provide the data" is too limited. This right needs to state in some way that data about an individual – however acquired – not be used so as to cause certain adverse consequences or harms to that individual. (See Section 1.4 for a possible list of adverse consequences and harms that might be subject to some regulation.)

As initially conceived, the right to Focused Collection was to be achieved by techniques like de-identification and data deletion. As discussed in Section 4.4.1, however, de-identification (anonymization) is not a robust technology for big data in the face of data fusion; in some instances, there may be compelling reasons to retain data for beneficial purposes. This right should be about use rather than collection. It should emphasize utilizing best practices to prevent inappropriate use of data during the data's whole life cycle, rather than depending on de-identification. It should not depend on a company's being able itself to recognize "all" the data about a consumer that it holds, which is increasingly technically infeasible.

The principles underlying CPBR's Security and Accountability remain valid in a use-based regime. They need to be applied throughout the value chain that includes data collection, analysis, and use.

Turn next to the rights here classified as consumer empowerments.

Where consumer empowerments have become practically impossible for the consumer to exercise meaningfully, they need to be recast as obligations of the commercial entity that actually uses the data or products of data analysis. This applies to the CPBR's principles of Individual Control and of Transparency.

Section 4.3 explained how the non-obvious nature of big data's products of analysis make it all but impossible for an individual to make fine-grained privacy choices for every new situation or app. For the principle of Individual Control to have meaning, PCAST believes that the burden should no longer fall on the consumer to manage privacy for each company with which the consumer interacts by a framework like "notice and consent." Rather, each company should take responsibility for conforming its uses of personal data to a personal privacy profile designated by the consumer and made available to that company (including from a third party designated by the consumer). Section 4.5.1 proposed a mechanism for this change in responsibility.

Transparency (in the sense of disclosure of privacy practices) suffers from many of the same problems. Today, the consumer receives an unhelpful blizzard of privacy-policy notifications, many of which say, in essence, "we

providers can do anything we want."[130] As with Individual Control, the burden of conforming to a consumer's stated personal-privacy profile should fall on the company, with notification to the consumers by a company if their profile precludes that company's accepting their business. Since companies do not like to lose business, a positive market dynamic for competing privacy practices would thus be created.

For the right of Access and Accuracy to be meaningful, personal data must include the fruits of data analytics, not just collection. However, as this report has already explained (Section 4.4.2), it is not always possible for a company to "know what it knows" about a consumer, since that information may be unrecognized in the data; or it may become identifiable only in the future, when data sets are combined using new algorithms. When, however, the personal character of data is apparent to a company by virtue of its use of the data, its obligation to provide means for the correction of errors should be triggered. Consumers should have an expectation that companies will validate and correct data stemming from analysis and, since not all errors will be corrected, will also take steps to minimize the risk of adverse consequences to consumers from the use of inaccurate data. Again, the primary burden must fall on the commercial user of big data and not on the consumer.

[130] Lawyers may encourage companies to use over-inclusive language to cover the unpredictable evolution of possibilities described elsewhere in this report, even in the absence of specific plans to use specific capabilities.

5. PCAST Perspectives and Conclusions

Breaches of privacy can cause harm to individuals and groups. It is a role of government to prevent such harm where possible, and to facilitate means of redress when the harm occurs. Technical enhancements of privacy can be effective only when accompanied by regulations or laws because, unless some penalties are enforced, there is no end to the escalation of the measures-countermeasures "game" between violators and protectors. Rules and regulations provide both deterrence of harmful actions and incentives to deploy privacy-protecting software technologies.

From everything already said, it should be obvious that new sources of big data are abundant; that they will continue to grow; and that they can bring enormous economic and social benefits. Similarly, and of comparable importance, new algorithms, software, and hardware technologies will continue to increase the power of data analytics in unexpected ways. Given these new capabilities of data aggregation and processing, there is inevitably new potential for both the unintentional leaking of both bulk and fine-grained data about individuals, and for new systematic attacks on privacy by those so minded.

Cameras, sensors, and other observational or mobile technologies raise new privacy concerns. Individuals often do not knowingly consent to providing data. These devices naturally pull in data unrelated to their primary purpose. Their data collection is often invisible. Analysis technology (such as facial, scene, speech, and voice recognition technology) is improving rapidly. Mobile devices provide location information that might not be otherwise volunteered. The combination of data from those sources can yield privacy-threatening information unbeknownst to the affected individuals.

It is also true, however, that privacy-sensitive data cannot always be reliably recognized when they are first collected, because the privacy-sensitive elements may be only latent in the data, made visible only by analytics (including those not yet invented), or by fusion with other data sources (including those not yet known). Suppressing the collection of privacy-sensitive data would thus be increasingly difficult, and it would also be increasingly counterproductive, frustrating the development of big data's socially important and economic benefits.

Nor would it be desirable to suppress the combining of multiple sources and kinds of data: Much of the power of big data stems from this kind of data fusion. That said, it remains a matter of concern that considerable amounts of personal data may be derived from data fusion. In other words, such data can be obtained or inferred without intentional personal disclosure.

It is an unavoidable fact that particular collections of big data and particular kinds of analysis will often have both beneficial and privacy-inappropriate uses. The appropriate use of both the data and the analyses are highly contextual.

Any specific harm or adverse consequence is the result of data, or their analytical product, passing through the control of three distinguishable classes of actor in the value chain:

First, there are *data collectors*, who control the interfaces to individuals or to the environment. Data collectors may collect data from clearly private realms (e.g., a health questionnaire or wearable sensor), from ambiguous situations (e.g., cell-phone pictures or Google Glass videos taken at a party or cameras and microphones placed

in a classroom for remote broadcast), or – increasing in both quantity and quality – data from the "public square," where privacy-sensitive data may be latent and initially unrecognizable.

Second, there are *data analyzers*. This is where the "big" in big data becomes important. Analyzers may aggregate data from many sources, and they may share data with other analyzers. Analyzers, as distinct from collectors, create uses ("products of analysis") by bringing together algorithms and data sets in a large-scale computational environment. Importantly, analyzers are the locus where individuals may be profiled by data fusion or statistical inference.

Third, there are *users of the analyzed data* – business, government, or individual. Users will generally have a commercial relationship with analyzers; they will be purchasers or licensees (etc.) of the analyzer's products of analysis. It is the user who creates desirable economic and social outcomes. But, it is also the user who is the locus of producing actual adverse consequences or harms, when such occur.

5.1 Technical feasibility of policy interventions

Policy, as created by new legislation or within existing regulatory authorities, can, in principle, intervene at various stages in the value chain described above. Not all such interventions are equally feasible from a technical perspective, or equally desirable if the societal and economic benefits of big data are to be realized.

As indicated in Chapter 4, basing policy on the control of collection is unlikely to succeed, except in very limited circumstances where there is an explicitly private context (e.g., measurement or disclosure of health data) and the possibility of *meaningful* explicit or implicit notice and consent (e.g., by privacy preference profiles, see Sections 4.3 and 4.5.1), which does not exist today.

There is little technical likelihood that "a right to forget" or similar limits on retention could be meaningfully defined or enforced (see Section 4.4.2). Increasingly, it will not be technically possible to surface "all" of the data about an individual. Policy based on protection by anonymization is futile, because the feasibility of re-identification increases rapidly with the amount of additional data (see Section 4.4.1). There is little, and decreasing, meaningful distinction between data and metadata. The capabilities of data fusion, data mining, and re-identification render metadata not much less problematic than data (see Section 3.1).

Even if direct controls on collection are in most cases infeasible, however, attention to collection practices may help to reduce risk in some circumstances. Such best practices as tracking provenance, auditing access and use, and continuous monitoring and control (see Sections 4.5.2 and 4.5.3) could be driven by partnerships between government and industry (the carrot) and also by clarifying tort law and defining what might constitute negligence (the stick).

Turn next to data analyzers. One the one hand, it may be difficult to regulate them, because their actions do not directly touch the individual (it is neither collection nor use) and may have no external visibility. Mere inference about an individual, absent its publication or use, may not be a feasible target of regulation. On the other hand, an increasing fraction of privacy issues will surface only with the application of data analytics. Many privacy challenges will arise from the analysis of data collected unintentionally that was not, at the time of collection, targeted at any particular individual or even group of individuals. This is because combining data from many sources will become more and more powerful.

It might be feasible to introduce regulation at the "moment of particularization" of data about an individual, or when this is done for some minimum number of individuals concurrently. To be effective such regulation would

need to be accompanied by requirements for tracking provenance, auditing access and use, and using security measures (e.g., robust encryption infrastructure) at all stages of the evolution of data, and for providing transparency, and/or notification, at the moment of particularization.

Big data's "products of analysis" are created by computer programs that bring together algorithms and data so as to produce something of value. It might be feasible to recognize such programs, or their products, in a legal sense and to regulate their commerce. For example, they might not be allowed to be used in commerce (sold, leased, licensed, and so on) unless they are consistent with individuals' privacy elections or other expressions of community values (see Sections 4.3 and 4.5.1). Requirements might be imposed on conformity to appropriate standards of provenance, auditability, accuracy, and so on, in the data they use and produce; or that they meaningfully identify who (licensor vs. licensee) is responsible for correcting errors and liable for various types of harm or adverse consequence caused by the product.

It is not, however, the mere development of a product of analysis that can cause adverse consequences. Those occur only with its actual use, whether in commerce, by government, by the press, or by individuals. This seems the most technically feasible place to apply regulation going forward, focusing at the locus where harm can be produced, not far upstream from where it may barely (if at all) be identifiable.

When products of analysis produce imperfect information that may misclassify individuals in ways that produce adverse consequences, one might require that they meet standards for data accuracy and integrity; that there are useable interfaces that allow an individual to correct the record with voluntary additional information; and that there exist streamlined options for redress, including financial redress, when adverse consequences reach a certain level.

Some harms may affect groups (e.g., the poor or minorities) rather than identifiable individuals. Mechanisms for redress in such cases need to be developed.

There is a need to clarify standards for liability in case of adverse consequences from privacy violations. Currently there is a patchwork of out-of-date state laws and legal precedents. One could encourage the drafting of technologically savvy model legislation on cyber-torts for consideration by the states.

Finally, government may be forbidden from certain classes of uses, despite their being available in the private sector.

5.2 Recommendations

PCAST's charge for this study does not ask it to make recommendations on privacy policies, but rather to make a relative assessment of the technical feasibility of different broad policy approaches. PCAST's overall conclusions about that question are embodied in the first two of our recommendations:

Recommendation 1. Policy attention should focus more on the actual uses of big data and less on its collection and analysis.

By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals. In the context of big data, these events ("uses") are almost always actions of a computer program or app interacting either with the raw data or with the fruits of analysis of those data. In this formulation, it is not the data themselves that cause the harm, nor the program itself (absent any data), but the confluence of the two. These "use events" (in commerce, by government, or by individuals)

embody the necessary specificity to be the subject of regulation. Since the purpose of bringing program and data together is to accomplish some identifiable desired task, use events also capture some notion of intent, in a way that data collection by itself or program development by itself may not. The policy question of what kinds of adverse consequences or harms rise to the level of needing regulation is outside of PCAST's charge, but an illustrative set that seem grounded in common American values was provided in Section 1.4.

PCAST judges that alternative big-data policies that focus on the regulation of data collection, storage, retention, a priori limitations on applications, and analysis (absent identifiable actual uses of big data or its products of analysis) are unlikely to yield effective strategies for improving privacy. Such policies are unlikely to be scalable over time as it becomes increasingly difficult to ascertain, about any particular data set, what personal information may be latent in it – or in its possible fusion with every other possible data set, present or future. The related issue is that policies limiting collection and retention are increasingly unlikely to be enforceable by other than severe and economically damaging measures. While there are certain definable classes of data so repugnant to society that their mere possession is criminalized,[131] the information in big data that may raise privacy concerns is increasingly inseparable from a vast volume of the data of ordinary commerce, or government function, or collection in the public square. This dual-use character of information, too, argues for the regulation of use rather than collection.

Recommendation 2. Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.

To avoid falling behind the technology, it is essential that policy concerning privacy protection should address the purpose (the "what") rather than the mechanism (the "how"). For example, regulating disclosure of health information by regulating the use of anonymization fails to capture the power of data fusion; regulating the protection of information about minors by controlling inspection of student records held by schools fails to anticipate the student information capturing by online learning technologies. Regulating control of the inappropriate disclosure of health information or student performance, no matter how the data are acquired is more robust.

PCAST further responds to its charge with the following recommendations, intended to advance the agenda of strong privacy values and the technological tools needed to support them:

Recommendation 3. With coordination and encouragement from OSTP, the NITRD agencies[132] should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies.

Some of the technology for controlling uses already exists. Research (and funding for it) is needed, however, in the technologies that help to protect privacy, in the social mechanisms that influence privacy-preserving

[131] Child pornography is the most universally recognized example.

[132] NITRD refers to the Networking and Information Technology Research and Development program, whose participating Federal agencies support unclassified research in in advanced information technologies such as computing, networking, and software and include both research- and mission-focused agencies such as NSF, NIH, NIST, DARPA, NOAA, DOE's Office of Science, and the D0D military service laboratories (see http://www.nitrd.gov/SUBCOMMITTEE/nitrd_agencies/index.aspx). There is research coordination between NITRD and Federal agencies conducting or supporting corresponding classified research.

behavior, and in the legal options that are robust to changes in technology and create appropriate balance among economic opportunity, other national priorities, and privacy protection.

Following up on recommendations from PCAST for increased privacy-related research,[133] a 2013-2014 internal government review of privacy-focused research across Federal agencies supporting research on information technologies suggests that about $80 million supports either research with an explicit focus on enhancing privacy or research that addresses privacy protection ancillary to some other goal (typically cybersecurity).[134] The funded research addresses such topics as an individual's control over his or her information, transparency, access and accuracy, and accountability. It is typically of a general nature, except for research focusing on the health domain or (relatively new) consumer energy usage. The broadest and most varied support for privacy research, in the form of grants to individuals and centers, comes from the National Science Foundation (NSF), engaging social science as well as computer science and engineering.[135,136]

Research into privacy as an extension or complement to security is supported by a variety of Department of Defense agencies (Air Force Research Laboratory, the Army's Telemedicine and Advanced Technology Research Center, Defense Advanced Research Projects Agency, National Security Agency, and Office of Naval Research) and the Intelligence Advanced Research Projects Activity (IARPA) within the Intelligence Community. IARPA, for example, has hosted the Security and Privacy Assurance Research[137] program, which has explored a variety of encryption techniques. Research at the National Institute for Standards and Technology (NIST) focuses on the development of cryptography and biometric technology to enhance privacy as well as support for federal standards and programs for identity management.[138]

Looking to the future, continued investment is needed not only in privacy topics ancillary to security, but also in automating privacy protection for the broadest aspects of use of data from all sources. Relevant topics include cryptography, privacy-preserving data mining (including analysis of streaming as well as stored) data,[139] formalization of privacy policies, tools for automating conformance of software to personal privacy policy and to legal policy, methods for auditing use in context and identifying violations of policy, and research on enhancing people's ability to make sense of the results of various big-data analyses. Development of technologies that support both quality analytics and privacy preservation on distributed data, such as secure multiparty computation, will become even more important, given the expectation that people will draw increasingly from

---

[133] *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology* (http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd2013.pdf [2012] and http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf [2010]).

[134] Federal Networking and Information Technology Research and Development Program, "Report on Privacy Research Within NITRD [Networking and Information Technology Research and Development], National Coordination Office for NITRD, April 23, 2014. http://www.nitrd.gov/Pubs/Report_on_Privacy_Research_within_NITRD.pdf

[135] The Secure and Trustworthy Cyberspace program is the largest funder of relevant research. See: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504709

[136] In December 2013, the NSF directorates supporting computer and social science joined in soliciting proposals for privacy-related research. http://www.nsf.gov/pubs/2014/nsf14021/nsf14021.jsp.

[137] See: http://www.iarpa.gov/index.php/research-programs/spar

[138] NIST is responsible for advancing the National Strategy for Trusted Identities in Cyberspace (NSTIC), which is intended to facilitate secure transactions within and across public and private sectors. See: http://www.nist.gov/nstic/

[139] Pike, W.A. et al., "PNNL [Pacific Northwest National Laboratory] Response to OSTP Big Data RFI," March 2014.

data stored in multiple locations. The creation of tools that analyze the panoply of National, state, regional, and international rules and regulations for inconsistencies and differences will be helpful for the definition of new rules and regulations, as well as for those software developers that need to customize their services for different markets.

Recommendation 4. OSTP, together with the appropriate educational institutions and professional societies, should encourage increased education and training opportunities concerning privacy protection, including professional career paths.

Programs that provide education leading to privacy expertise (akin to what is being done for security expertise) are essential and need encouragement. One might envision careers for digital-privacy experts both on the software development side and on the technical management side. Employment opportunities should exist not only in industry (and government at all levels), where jobs focused on privacy (including but not limited to Chief Privacy Officers) have been growing, but also for consumer and citizen advocacy and support, perhaps offering "annual privacy checkups" for individuals. Just as education and training about cybersecurity has advanced over the past 20 years within the technical community, there is now opportunity to educate and train students about privacy implications and privacy enhancements, beyond the present small niche area occupied by this focus within computer science programs.[140] Privacy is also an important component of ethics education for technology professionals.

Recommendation 5. The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today. This country can exhibit leadership both by its convening power (for instance, by promoting the creation and adoption of standards) and also by its own procurement practices (such as its own use of privacy-preserving cloud services).

Section 4.5.2 described a set of privacy-enhancing best practices that already exist today in U.S. markets. PCAST is not aware of any more effective innovation or strategies being developed abroad; rather, some countries seem inclined to pursue what PCAST believes to be blind alleys. This circumstance offers an opportunity for U.S. technical leadership in privacy in the international arena, an opportunity that should be seized. Public policy can help to nurture the budding commercial potential of privacy-enhancing technologies, both through U.S. government procurement and through the larger policy framework that motivates private-sector technology engagement.

As it does for security, cloud computing offers positive new opportunities for privacy. By requiring privacy-enhancing services from cloud-service providers contracting with the U. S. government, the government should encourage those providers to make available sophisticated privacy enhancing technologies to small businesses and their customers, beyond what the small business might be able to do on its own.[141]

---

[140] A basis can be found in the newest version of the curriculum guidance of the Association for Computing Machinery (http://www.acm.org/education/CS2013-final-report.pdf). Given all of the pressures on curriculum, progress—as with cybersecurity—may hinge on growth in privacy-related research, business opportunities, and occupations.

[141] A beginning can be found in the Federal Government's FedRAMP program for certifying cloud services. Initiated to address Federal agency security concerns, FedRAMP already builds in attention to privacy in the form of a required Privacy Threshold Analysis and in some situations a Privacy Impact Analysis. The office of the U.S. Chief Information Officer

5.4 Final Remarks

Privacy is an important human value. The advance of technology both threatens personal privacy and provides opportunities to enhance its protection. The challenge for the U.S. Government and the larger community, both within this country and globally, is to understand what the nature of privacy is in the modern world and to find those technological, educational, and policy avenues that will preserve and protect it.

provides guidance on Federal uses of information technology that addresses privacy along with security (see http://cloud.cio.gov/). It provides specific guidance on the cloud and FedRAMP (http://cloud.cio.gov/fedramp), including privacy protection (http://cloud.cio.gov/document/privacy-threshold-analysis-and-privacy-impact-assessment).

Appendix A. Additional Experts Providing Input

Yochai Benkler
Harvard

Eleanor Birrell
Cornell University

Courtney Bowman
Palantir

Christopher Clifton
Purdue University

James Costa
Sandia National Laboratory

Lorrie Faith Cranor
Carnegie Mellon University

Deborah Estrin
Cornell NYC

William W. (Terry) Fisher
Harvard Law School

Stephanie Forrest
University of New Mexico

Dan Geer
In-Q-Tel

Deborah K. Gracio
Pacific Northwest National Laboratory

Eric Grosse
Google

Peter Guerra
Booz Allen

Michael Jordan
University of California, Berkeley

Philip Kegelmeyer
Sandia National Laboratory

Angelos Keromytis
Columbia University

Thomas Kalil
OSTP

Jon Kleinberg
Cornell University

Julia Lane
American Institutes for Research

Carl Landwehr
George Washington University

David Moon
Ernst & Young

Keith Marzullo
National Science Foundation

Martha Minow
Harvard Law School

Tom Mitchell
Carnegie Mellon University

Deirdre Mulligan
University of California, Berkeley

Leonard Napolitano
Sandia National Laboratory

Charles Nelson
OSTP

Chris Oehmen
Pacific Northwest National Laboratory

Alex "Sandy" Pentland
Massachusetts Institute of Technology

Rene Peralta
National Institute of Standards and Technology

Anthony Philippakis
Genome Bridge

Timothy Polk
OSTP

Fred B. Schneider
Cornell University

Greg Shipley
In-Q-Tel

Lauren Smith
OSTP

Francis Sullivan
Institute for Defense Analysis

Thomas Vagoun
NITRD National Coordination Office

Konrad Vesey
Intelligence Advanced Research Activity

James Waldo
Harvard

Peter Weinberger
Google, Inc.

Daniel J. Weitzner
Massachusetts Institute of Technology

Nicole Wong
OSTP

Jonathan Zittrain
Harvard Law School

Special Acknowledgment

PCAST is especially grateful for the rapid and comprehensive assistance provided by an ad hoc group of staff at the National Science Foundation (NSF), Computer and Information Science and Engineering Directorate. This team was led by Fen Zhao and Emily Grumbling, who were enlisted by Suzanne Iacono. Drs. Zhao and Grumbling worked tirelessly to review the technical literature, elicit perspectives and feedback from a range of NSF colleagues, and iterate on descriptions of numerous technologies relevant to big data and privacy and how those technologies were evolving.

NSF Technology Team Leaders

Fen Zhao, AAAS Fellow, CISE
Emily Grumbling, AAAS Fellow, Office of Cyberinfrastructure

Additional NSF Contributors

Robert Chadduck, Program Director
Almadena Y. Chtchelkanova, Program Director
David Corman, Program Director
James Donlon, Program Director
Jeremy Epstein, Program Director
Joseph B. Lyles, Program Director
Dmitry Maslov, Program Director
Mimi McClure, Associate Program Director
Anita Nikolich, Expert
Amy Walton, Program Director
Ralph Wachter, Program Director

President's Council of Advisors on Science and Technology (PCAST)

www.whitehouse.gov/ostp/pcast