Journal of Privacy and Confidentiality (2010) **2**, Number 1, pp. 19–50

Thinking About Privacy: Chapter 1 of "Engaging Privacy and Information Technology in a Digital Age"

Committee on Privacy in the Information Age, National Research Council
James Waldo*, Herbert S. Lin†, and Lynette I. Millett‡, Editors

Just as recent centuries saw transitions from the agricultural to the industrial to the information age and associated societal and technological changes, the early 21st century will continue to pose dynamic challenges in many aspects of society. Most importantly from the standpoint of this report, advances in information technology are proceeding apace. In this rapidly changing technological context, individuals, institutions, and governments will be forced to reexamine core values, beliefs, laws, and social structures if their understandings of autonomy, privacy, justice, community, and democracy are to continue to have meaning. A central concept throughout U.S. history has been the notion of privacy and the creation of appropriate borders between the individual and the state. In the latter 19th century, as industrial urban society saw the rise of large bureaucratic organizations, notions of privacy were extended to the borders between private organizations and the individual. This report focuses on privacy and its intersections with information technology and associated social and technology trends.

1 Introduction

One of the most discussed and worried-about aspects of today's information age is the subject of privacy. Based on a number of other efforts directed toward analyzing trends and impacts of information technology (including the evolution of the Internet, a variety of information security issues, and public-private tensions regarding uses of information and information technology), the National Research Council saw a need for a comprehensive assessment of privacy challenges and opportunities and thus established the Committee on Privacy in the Information Age.

The committee's charge had four basic elements:

- To survey and analyze potential areas of concern—privacy risks to personal information associated with new technologies and their interaction with non-technology-based risks, the incidence of actual problems relative to the potential, trends in technology and practice that will influence impacts on privacy, and so on;
- To evaluate the technical and sociological context for those areas as well as new

*Sun Microsystems, Mountain View, CA, mailto:jim.waldo@east.sun.com
†Computer Science and Telecommunications Board, National Research Council of the National Academies, Washington, DC, mailto:hlin@nas.edu
‡Computer Science and Telecommunications Board, National Research Council of the National Academies, Washington, DC, mailto:ldc@cdc.gov

collection devices and methodologies—why personal information is at risk given its storage, communication, combination with other information, and various uses; trends in the voluntary and involuntary (and knowing and unknowing) sharing of that information;

- To assess what is and is not new about threats to the privacy of personal information today, taking into account the history of the use of information technology over several decades and developments in government and private sector practices; and

- To examine the tradeoffs (e.g., between more personalized marketing and more monitoring of personal buying patterns) involved in the collection and use of personal information, including the incidence of benefits and costs,[1] and to examine alternative approaches to collection and use of personal information.

Further, in an attempt to paint a big picture that would at least sketch the contours of the full set of interactions and tradeoffs, the charge called for these analyses to take into account changes in technology; business, government, and other organizational demand for and supply of personal information; and the increasing capabilities for individuals to collect and use, as well as disseminate, personal information. Within this big picture, and motivated by changes in the national security environment since the September 11, 2001, attacks on the World Trade Center and the Pentagon, the committee addressed issues related to law enforcement and national security somewhat more comprehensively than it did other areas in which privacy matters arise.

To what end does the committee offer this consideration of privacy in the 21st century? Most broadly, to raise awareness of the spider web of connectedness among the actions we take, the policies we pass, the expectations we change, the "flip side" of impacts policies have on privacy. There should not be unintended consequences to privacy created by policies we write or change to address the continuing shifts in our society. We may decide to tolerate erosion on one side of a continuum—privacy and security sometimes pose a conflict, for example. We may decide it makes sense to allow security personnel to open our bags, to carry a "trusted traveler" card, to accept "profiling" of people for additional examination. But we should not be surprised by the erosion of our own and other people's privacy by this shift in the continuum. Policies may create a new and desirable equilibrium, but they should not create unforeseen consequences.

The goals here are not to evaluate "good" and "bad," whether in changes in the continuums privacy moves on, policies, technologies, and laws. Rather, the committee hopes that this report will contribute to a recalibration of the many issues that play a part in privacy and will contribute to the analysis of issues involving privacy. The degree of privacy traded for security or public health, for example, should be a result of thoughtful decisions following public discussion in which all parties can participate.

[1] Throughout this report, the term "benefits and costs" should be construed broadly, and in particular should not be limited simply to economic benefits and costs.

Only then will the policies that emerge from the pressures at work during the early years of the 21st century be understood in their impacts on privacy.

To be clear, the committee does not claim that this report presents comprehensive solutions to the many privacy challenges confronting society today. Nor does it provide a thorough and settled definition of privacy. Debate will continue on this complicated and value-laden topic for the foreseeable future. This report does provide ways to think about privacy, its relationship to other values, and related tradeoffs. It emphasizes the need to understand context when evaluating the privacy impact of a given situation, piece of legislation, or technology. And it provides an in-depth look at ongoing information technology trends as related to privacy concerns.

2 What is Privacy?

The committee began by trying to understand what privacy is, and it quickly found that privacy is an ill-defined but apparently well-understood concept. It is ill-defined in the sense that people use the term to mean many different things. Any review of the literature on privacy will reveal that privacy is a complicated concept that is difficult to define at a theoretical level under any single, logically consistent "umbrella" theory, even if there are tenuous threads connecting the diverse meanings. Specifying the concept in a way that meets with universal consensus is a difficult if not impossible task, as the committee found in doing its work.

At the same time, the term "privacy" is apparently well understood in the sense that most people using the term believe that others share their particular definition. Nonetheless, privacy resists a clear, concise definition because it is experienced in a variety of social contexts. For example, a question may be an offensive privacy violation in one context and a welcome intimacy in another.

The committee believes that in everyday usage, the term "privacy" generally includes reference to the types of information available about an individual, whether they are primary or derived from analysis. These types of information include behavioral, financial, medical, biometric, consumer, and biographical. Privacy interests also attach to the gathering, control, protection, and use of information about individuals. Informational dimensions of privacy thus constitute a definitional center of gravity for the term that is used in this report, even while recognizing that the term may in any given instance entail other dimensions as well—other dimensions that are recognized explicitly in the discussion.[2]

The multidimensional nature of privacy is explicated further in Chapter 2, and a theme that becomes apparent is the situational and contextual nature of privacy—that is, it depends on a number of specific factors that often do not cleanly and clearly overlap, rather than being identified by a sweeping universal calculus or definition.

[2]The term "private" can have both descriptive and normative meanings. To describe information as "private information" might mean "information that is not accessible to others," or it could mean "information that should not be accessible to others." Generally the context will specify the meaning, but these two different meanings are noteworthy.

Moreover, privacy in any given situation may be in tension with other values or desires of the individual, subgroups, and society at large. Privacy, like most other values in modern democratic societies, is not an absolute but rather must be interpreted and weighed alongside other socially important values and goals. How this balancing (which need not mean equivalent weighing) is to be achieved is often the center of the controversy around privacy, because different people and groups balance in different ways these values that are in tension.

A further complication is that participants in the balancing debate often confuse the needs of privacy with other values that might be tied to privacy but that are, in fact, distinct from it. For example, concerns over whether an individual's HIV status should be private may in fact reflect, in part, a concern about his or her ability to obtain health insurance.

In short, as with most interesting and contentious social topics, where privacy is concerned there are both costs and benefits, and these vary by the group, context, and time period in question, as well as by the means used to measure them. Sometimes, tradeoffs are inevitable (Box 1.1 provides some illustrative examples). Advocates for various positions who argue vigorously for a given policy thus run the risk of casting their arguments in unduly broad terms. Though rhetorical excesses are often a staple of advocacy, in truth the factors driving the information age rarely create simple problems with simple solutions.

Perhaps the best known of the general tradeoffs in the privacy debate is that which contrasts privacy with considerations of law enforcement and national security. At this writing, there is considerable debate over the Bush administration's use of warrantless wiretapping in its counterterrorism efforts against al-Qaeda. Furthermore, the USA PATRIOT Act, passed in the immediate wake of the September 11, 2001, attacks on the World Trade Center and the Pentagon and extended and amended in early 2006, changed a number of privacy-related laws in order to facilitate certain law enforcement and national security goals. (Chapter 9 contains an extensive discussion of these issues.)

But the law enforcement/national security versus privacy debate is hardly the only example of such tradeoffs that are being made. Box 1.1 provides some illustrations. Privacy concerns interact with the delivery of health care and the information needed to contribute to public health as well as the information needed to discover and understand risk factors that any individual may have. Privacy concerns interact with the ability to do long- and short-term sociological studies. Techniques that are believed to increase productivity and profitability may come at a cost to the privacy interests of many consumers and workers. Privacy concerns also are reflected in the debates about new forms of intellectual property.

Privacy concerns also interact with sociological and policy research. In order to conduct these kinds of research, substantial amounts of personal information are often necessary. However, in general, these data never have to be associated with specific individuals. This situation contrasts sharply with the societal needs described above: law enforcement authorities are interested in apprehending a specific individual guilty of criminal wrongdoing, national security authorities are interested in identifying a parti-

Box 1.1
Some Illustrative Tradeoffs in Privacy

- Government or privately controlled cameras monitoring the movement of ordinary citizens in public places for the stated purpose of increasing public safety.
- Government collection of data on peoples' political activities for the stated purpose of increasing public safety or homeland security.
- Collection by a retailer of personal information about purchases for the stated purpose of future marketing of products to specific individuals.
- Collection by a bank of personal financial information about an individual for the stated purpose of evaluating his or her creditworthiness for a loan.
- Aggregation by insurers of medical data obtained through third parties for the stated purpose of deciding on rates or availability of health insurance for an individual.
- Provision of information to law enforcement agencies about library patrons (including who they are and what they read or saw in the library) for the stated purpose of increasing public safety or homeland security, and a prohibition of discussing or acknowledging that this has been done.
- Availability of public government records (including criminal records, family court proceedings, real estate transactions, and so on, and formerly only available in paper format) on the World Wide Web for the stated purpose of increasing the openness of government.
- Geographic tracking of cell-phone locations at all times for the stated purpose of enabling emergency location.

Note also that privacy concerns are often grounded in information that may be used for purposes other than a stated purpose. Indeed, in each of the examples given above, another possibleand less benignpurpose might easily be envisioned and thus might change entirely one's framing of a privacy issue.

-cular terrorist, or a business wants to identify a specific customer who will buy a product. For these reasons, protected data collections such as those found in social science data archives and census public-use files serve the interests of groups and communities with less controversy; when controversy does exist, it usually relates to whether the data contained in these files and archives are sufficiently anonymized, or to specific nonstatistical uses of these data.

Tradeoffs are also not limited to the value of information to an organization versus the value of privacy to an individual—they also arise in the same situation of an individual alone. For example, an individual might regard his or her personal information as a commodity that can be traded freely in exchange for some other good or service of value—and thus he or she might well be willing to provide personal information on shopping habits at a chain drugstore or supermarket in exchange for a 2 percent discount on all purchases. Furthermore, even if the tradeoffs do appear to pit value to an organization against value to an individual, some would argue that there is benefit to the individual as well (albeit not specific benefit to him or her) if the organization can be construed as "all or most of society." This point is discussed in greater detail in Section 4 of Chapter 6.

Not only are these tradeoffs complex, difficult, and sometimes seemingly intractable, but they are also often not made explicit in the discussions that take place around the policies that, when they are enacted, quietly embody the value tradeoffs. Clarifications on these points would not necessarily relieve the underlying tensions, but they would likely help illuminate the contours of the debate. A major purpose of this report is to contribute to that illumination.

3 An Illustrative Case

In early 2005, a firm known as ChoicePoint announced that "a crime committed against ChoicePoint...MAY have resulted in [consumer] name[s], address[es], and Social Security number[s] being viewed by businesses that are not allowed access to such information."[3] Specifically, ChoicePoint reported that "several individuals, posing as legitimate business customers, recently committed fraud by claiming to have a lawful purpose for accessing information about individuals, when in fact, they did not." ChoicePoint explained its business as verifying for its business customers information supplied by individuals as part of a business transaction, often as part of an application for insurance, a job, or a home loan. ChoicePoint notified approximately 143,000 individuals that their personal information might have been compromised. In early 2006, the U.S. Federal Trade Commission (FTC) announced that ChoicePoint would pay $15 million in fines and other penalties for lax security standards in verifying the credentials of its business customers. Furthermore, the FTC noted that "this breach occurred because ChoicePoint failed to implement reasonable and appropriate procedures for approving new customers and for monitoring existing ones."[4] It also said that more than 800 cases of identity theft arose from this breach in security.

For purposes of this study, the truth or falsity of the FTC's allegations about ChoicePoint's security practices per se is not relevant. But what is relevant is that the personal information of more than 143,000 individuals was released to parties that did

[3] "Choicepoint's Letter to Consumers Whose Information Was Compromised," *CSO Magazine*, available at `http://www.csoonline.com/read/050105/choicepoint_letter.html`.

[4] Federal Trade Commission, "ChoicePoint Settles Data Security Breach Charges; to Pay $10 Million in Civil Penalties, $5 Million for Consumer Redress," available at `http://www.ftc.gov/opa/2006/01/choicepoint.htm`.

not have a lawful purpose in receiving that information, and that a number of cases of identity theft arose from this release.

Several questions immediately come to mind:

1. How is ChoicePoint able to aggregate such voluminous information? The data that ChoicePoint collects on individuals includes criminal histories, Social Security numbers, and employment histories.

2. Why do ChoicePoint and other similar firms collect such voluminous data on individuals?

3. What was the harm suffered by the individuals whose identities were not stolen? Eight hundred individual cases of identity theft were attributed to the breach, a number corresponding to about $\frac{1}{2}$ of 1 percent of the 143,000 individuals involved.

4. To what extent were individuals notified by ChoicePoint surprised by the existence of such aggregations of personal data?

Question 1 points to the availability of great quantities of personal information on a large scale to organizations that have no direct involvement in the creation of the data. ChoicePoint is not the primary collector of such information; it is an aggregator of it. It also points to the fact that information collected for one purpose (e.g., a job application with a certain employer) can be "repurposed" and used for an entirely different purpose (e.g., verification of job history in connection with a background investigation).

Question 2 points to a demand on the part of private businesses and government agencies for personal information about its employees and customers. Indeed, such information is so important to these businesses and government agencies that they are willing to pay to check and verify the accuracy of information provided by employees and customers. (Note also that by insisting that employees and customers provide personal information, these businesses and agencies often add to the personal information that is available to data aggregators.)

Question 3 focuses attention on the value of privacy and the nature of the harm that can accrue to individuals when their privacy is breached even if they have not been the victims of identity theft. In this case, the answer is that these individuals suffer the same harm that Damocles experienced when he was partying and feasting under the sword. No physical harm came to Damocles, yet the cost to his sense of well-being was high indeed. A person whose privacy has been breached is likely to be concerned about the negative consequences that might flow from the breach, and those kinds of psychological concerns constitute a type of actual though intangible harm entirely apart from the other kinds of tangible harm that the law typically recognizes. A second kind of intangible harm experienced by Damocles might have been his reluctance to engage in dancing and making loud noises that might have caused the thread holding the sword to break—a so-called chilling effect on his activities and behaviors. In short—harm need

not be tangible to be real or actual.[5]

Question 4 alerts us to issues involving the commodification of personal information and its being treated as a kind of marketable property to be used as those who come to possess it choose. Question 4 also calls attention to several collateral issues surrounding privacy. In this case, the issue is the role of notification in privacy, and whether notification that personal information is being collected about an individual in any sense ameliorates any breach of privacy that might be associated with that collection. Given legal requirements to notify individuals after privacy violations have been documented, are such violations thus less likely?

Questions and issues such as these recur frequently in this report, although in no sense do these examples exhaust the kinds of questions and issues that arise. Privacy provides a useful filter through which to think about individual and societal benefits and costs.

4 The Dynamics of Privacy

Privacy is part of a social context that is subject to a range of factors. While a relationship between privacy and society has always existed, the factors (or pressures) affecting privacy in the information age are varied and deeply interconnected. These factors, individually and collectively, are changing and expanding in scale with unprecedented speed in terms of our ability to understand and contend with their implications to our world, in general, and our privacy, in particular. Some of these factors include the volume, magnitude, complexity, and persistence of information; the expanding number of ways to collect information; the number of people affected by information; and the geographic spread and reach of information technology.

4.1 The Information Age

What is meant by the term "information age," and what are the factors so profoundly affecting the dynamics of privacy? With respect to the information age, a great deal has been written about the fact that almost no part of our lives is untouched by computing and information technology. These technologies underlie new ways of collecting and handling information that in turn have ramifications throughout society, as they mediate much private and public communication, interaction, and transactions. They are central components of contemporary infrastructures involving (but certainly not restricted to) commerce, banking and finance, utilities, communications, national defense, education, and entertainment.

This brief characterization of the information age highlights the three major factors, indeed drivers, of the vast changes affecting current notions, perceptions, and expecta-

[5]Nor is "harm" a concept that is relevant only to individuals. As Section 3 in Chapter 2 addresses in greater detail, certain kinds of harm may relate to groups or to society as a whole. Group or societal harms may be related to individually suffered harm, but are conceptually separate notions.

Box 1.2
Large-scale Factors Affecting Privacy

Technological Change	*Societal Trends*	*Discontinuities in Circumstance*
• Ubiquity	• Globalization	• Catastrophic attacks in 2001 on the World Trade Center/Pentagon
• Connectivity	• Mobility	• Watergate scandal in 1972-1973
• Data collection	• Virtuality	• Church Committee Hearings of 1976 (also known as the Hearings of the United States Senate Select Committee to Study Governmental Operations with Respect to Intelligence Activities)
• Storage	• Urbanization	• Attack on Pearl Harbor in 1941
• Computational power	• Constant accessibility	• Invention in 1995 of the World Wide Web[1]
• Commoditization of hardware	• Litigiousness	• National and international health threats (SARS and avian flu)
• Software usability	• Demographic/Aging	
• Encryption	• New ways of living and communicating	
• Privacy-relevant biotechnology	• Increases in social networking	
• Extensions of human senses	• Increased societal interdependance	
• Portability of data and communications devices	• Increase in electronic communication literacy	
• Persistence of information	• Increase in expectations for information availability	
• Affordability of data and communications	• Linked monetary systems	
• Advances in sensor technology	• Linked production systems	

[1]The World Wide Web is a product of technology trends, but it was also the primary driving force underlying the explosion of easy-to-use Internet application that ultimately made enormous amounts of information—personal and otherwise—publicly accessible.

-tions of privacy: *technological change*, *societal shifts*, and *discontinuities in circumstances* (Box 1.2).

Technological change (Column 1 in Box 1.2) refers to major differences in the technological environment of today as compared to that existing many decades ago (differences that have a major influence on today's social and legal regime governing privacy). Column 2 in Box 1.2 identifies a number of trends that set a large-scale social and cultural context for discussions of privacy. Societal shifts refer to evolutionary trends in soci-

ety writ large. Discontinuities in circumstances (Column 3 in Box 1.2) are events and emergent concerns that transformed the national debate about privacy in a very short time (and thus did not allow for gradual adjustment to a new set of circumstances).

Society is thus experiencing the effects of changes in these factors. For example:

- Changes in technology have enhanced access to information and images previously available to the public but then much more difficult to access. New technologies that extend the senses have made new kinds of data available as a result of covert "soft surveillance." The fact that such surveillance permits the collection of personal information without the consent or knowledge of the subject offers temptations for misuse.

- Changes in business models, which are increasingly based on the notion of greater customization of services and products, a process that in turn requires large amounts of personal information so that the appropriate customization can be employed.

- Changes in expectations of security following the terrorist attacks of 2001 have reduced people's expectations of the privacy rights of foreign nationals and U.S. citizens in this country, as did the attack on Pearl Harbor in 1941. Similarly, the post-Watergate revelations of government abuse of records containing personal information increased peoples expectations of the privacy rights to which they were entitled.

Subsequent chapters characterize these rapid changes in some detail. For the purposes of this introduction to thinking about privacy, it is sufficient to note that each of these changes is having significant impacts on society. However, in combination, these changes are key drivers of the information society and underlie fundamental changes in how we, as individuals and as a society, grapple with privacy, business activities, social interaction, and information. These systemic and profound changes in turn have a most direct influence on the dynamics of privacy—and indeed privacy's salience as a topic of importance to this committee and to citizens generally.

4.2 Information Transformed and the Role of Technology

Technological advancements, coupled with changes in other areas, combine to make the privacy challenge particularly vexing. Technological change is, of course, not new. The printing press has been described as a precursor to the World Wide Web; e-mail and cell phone text messaging have revolutionized interpersonal and group correspondence. Affordability and advances in sensor technologies have broadened the volume and scope of information that can be practically acquired. The privacy debate in the United States itself has part of its roots in the technological changes involving the press and technology for photographyWarren and Brandeis, in their landmark 1890 *Harvard Law Review* paper,[6] were responding to, as they put it, "recent inventions and business

methods."

The business method at issue was the popular press, and the most striking of the recent inventions—the technology—was the unposed photograph. Suddenly, it was easy to take spontaneous and often uninvited photos of people—which Warren and Brandeis denounced as "invad[ing] the sacred precincts of private and domestic life"—and to show the results to a large, literate, curious, and gossipy audience.[7]

What makes information special is that it is reproducible. In digital form, information can be copied an infinite number of times without losing fidelity. Digitized information is also easy to distribute at low cost. Today, in the information age, the sheer quantity of information; the ability to collect unobtrusively, aggregate, and analyze it; the ability to store it cheaply; the ubiquity of interconnectedness; and the magnitude and speed of all aspects of the way we think about, use, characterize, manipulate, and represent information are fundamentally and continuously changing. Consider concepts of:

- *Information search.* Within half a generation we have moved from dusty card catalogues and file drawers full of rolls of microfiche to warehouses of servers connected to the worldwide Internet that allow, among many other things, much of the Internet to be searched for keywords at the click of a button.

- *Information production.* In just the world of publishing alone, we have moved from mimeographs and hand distribution for the truly dedicated amateur to parents creating, modifying, and publishing entire photo and video albums of their children in ways that are accessible almost instantly around the globe. Blogging enables many of us to publish nearly anything we want on the Internet.

- *Information manipulation.* The ways in which information can be manipulated have expanded—both in terms of capability and also in terms of who has access to the tools that allow such manipulation. Photoediting software and sound-editing technologies are now bundled with many common personal computers. What might have taken hours to correct or modify in the days of the professional darkroom or recording studio can now be trivially accomplished by anyone with a PC.

- *Information storage.* In many cases, records containing information are no longer thrown away. It has become less expensive to keep the data on larger, cheaper storage devices than to cull the information accurately so as to remove data. As a result data that has outlived its original use is retained and becomes subject to future unanticipated uses.

[6]Samuel D. Warren and Louis D. Brandeis, "The Right to Privacy," *Harvard Law Review IV* (December 15, No. 5):195, 1890, available at `http://www.lawrence.edu/fac/boardmaw/Privacy_brand_warr2.html`.

[7]George Radwanski, Address to the Privacy Lecture Series, Toronto, Ontario, March 26, 2001, available at `http://www.privcom.gc.ca/speech/02_05_a_010326_2_e.asp`.

- *Information acquisition.* It is easier today than ever before to acquire many kinds of information about individuals. Sensors such as video surveillance cameras and radio-frequency ID tags are rapidly dropping in cost and are increasingly ubiquitous in the environment. Cell phones are capable of localizing to an accuracy of 100 meters the real-time whereabouts of the individuals carrying them. Electronic fare cards for public transportation often identify entry and exit points, along with the time of day.
- *Information analysis.* Sophisticated algorithms are increasingly capable of finding patterns buried in large quantities of data. Basic statistics of data can be generated on board sensor platforms, or even the sensors themselves, before even being transmitted to a central point of analysis. And the ingenuity of users knows few bounds, as such users find new ways of using information already collected for new purposes.

Trends in information technology have made it easier and cheaper by orders of magnitude to gather, retain, and analyze information. Other trends have also enabled access to new kinds of information that historically would have been next to impossible to gather about another individual. For example, certain kinds of data acquisition devices are already widely deployed (e.g., video cameras). The cost of such devices is dropping, which will enable even more ubiquitous deployment. And it will be increasingly easy to collect information from them as they are deployed not as standalone devices but in networks. Such devices have many socially beneficial applications, ranging from health care monitoring to monitoring of weather and geophysical variables to traffic control. But even if the data from these systems are not intended to monitor human interaction and behavior, they can often be repurposed to do exactly that. Moreover, information about human behavior can be inferred from seemingly innocuous data (such as heat sources in buildings or the way a person walks).

Still another effect of new information technologies is the erosion of privacy protection once provided through obscurity or the passage of time; e.g., youthful indiscretions can now become impossible to outlive as an adult. For much of the past, the effects of data collection were not a major issue, perhaps because the relevant data were inaccessible for practical purposes or individual pieces of data were stored in different locations so that patterns contained within the potentially aggregated data were difficult to find. Often either the sheer volume of input would overwhelm the method of analysis or the patterns would be lost in a sea of data. It is not quite the case that data were inaccessible, but they were contained in the form of, for example, public records stored in filing cabinets in county clerks' basements, and were in practice expensive and difficult to access.

Today and increasingly in the future, electronic storage of information is less expensive and potentially more persistent than paper storage.[8] Also, information systems have moved from isolated systems to clustered systems of users and machines to what now is becoming a mesh of interconnected information and analysis systems, which can share information and work collectively, leading to a much greater ease of data aggregation. Once data is aggregated, new and more powerful techniques and technologies

for analyzing information (generically known as data mining) will make it much easier to extract and identify personally identifiable patterns that were previously protected by the vast amounts of data "noise" around them. Furthermore, as the interrelationships between systems become more closely identified, the issues of ownership, control, prerogative, and privacy also become more difficult to discern or manage.

Similarly worrisome to many is the emergence of biometric identification, the use of information technologies to measure and record biological or physiological characteristics of the human body for identification purposes. Such characteristics can include DNA sequences, gait, retinal patterns, fingerprints, and so on. The primary significance of biometrics in a privacy context is that certain markers are selected for large-scale use because they are believed to be more or less invariant over an individual's lifetime. (Whether this is in fact the case in any given instance can be a subject of great debate.)

The comments above should not be taken to mean that the advance of technology has only negative effects on privacy. As the discussion in Chapter 3 indicates, some advances in technology can promote or enhance privacy. For example, technology enables the maintenance of audit trails that can keep track of who accesses what data. The possibility of accessing sensitive data improperly on an anonymous basis often presents a strong temptation for doing so, and the keeping of audit logs can often deter such activity. However, such privacy-protecting technologies must be deployed in order to enhance privacy, and because they generally have no operational or business value other than protecting privacy, it is often the case that such protective technologies are not deployed.

4.3 Societal Shifts and Changes in Institutional Practice

Focusing solely on technological advancements provides an incomplete view of how values, understandings, and expectations shift over time. Important consideration must be given to societal institutions—the organizations and the activities and practices that make use of the technological systems described above—and to the transformation of social institutions through their routine use.

Modern society is characterized, in part, by a multitude of demands for personal information not just from families and one's immediate community but also from governments and institutions. Whether these demands are the result of new technologies searching for problems to solve at lower cost, or whatever they serve to stimulate the growth of new technologies, is open to question—as with most such questions, the most likely answer is "some of both."[9] But what is clear today is that making personal infor-

[8] Whether paper or electronic storage is in fact more persistent in the long run (measured in decades) is not known with certainty. Whereas paper is a very simple and enduring medium, today's high-capacity CD-ROMs and DVDs may be largely unusable in 10 years. The problem of electronic media obsolescence as it affects access to stored information can be addressed by periodically rewriting the information onto new media, but such rewriting presents logistical challenges that can be daunting for individuals and organizations alike. (On the difficulties faced by organizations, see National Research Council, *LC21 : A Digital Strategy for the Library of Congress*, National Academy Press, Washington, D.C., 2000.)

mation available to institutions and organizations is absolutely essential for individual participation in everyday life.

Consider, for example, the information demands involved in:

- Licensing practices, of which the driver's license is the most ubiquitous example. To obtain a driver's license, an individual must provide personal information (e.g., name, address, and so on) as well as proof of driving ability. But over time, a driver's license comes to contain a driver's history of moving violations and accidents as well. Furthermore, a driver's license is a de facto ID standard for many purposes, ranging from admission to facilities and air travel to check cashing. Though automated systems are not in place today to collect driver's license numbers in all of these applications, they could be—and the volume of personal information about spending habits, locations, travel, and so on that might be assembled through such systems is rather large. For other licensing applications, such as licenses needed for various professions, other kinds of information may be needed, such as various histories of education, records of previous practice, and customer complaints and/or disciplinary actions taken. To receive an amateur radio operator's license, a person regardless of age—must allow his or her name and address to be posted on the Internet.

- Many benefits in society are conferred by law only to particular classes of people (e.g., veterans, the unemployed, those with low income, homeowners). Establishing eligibility and verifying claims require individual information. In response to concerns about fraud, administering government agencies are asking for more information and have increasingly turned to computer matching involving diverse databases. In contrast, such agencies rarely do computer matching to identify potential clients who are not utilizing benefits to which they are entitled.

- Many private sector institutions make distinctions between categories of people. For example, the granting and the terms of credit to individuals both depend heavily on many of the details of their financial history (e.g., records of payment, length of time at particular addresses, employment record, income). Admissions to many institutions of higher education depend on a detailed history and record of an individual's curricular and extracurricular activities.

- Many institutions require personal information as a condition of providing service at all. In some cases, the need for personal information is intrinsic to the service itself—health care services for an individual are perforce information-intensive, and given societal pressures to deliver more effective health care at lower cost,

[9] As one example, a string of technological innovations that shaped, and were shaped by, the development of the modern bureaucracy between 1890 and 1939 is described in James Beniger, *The Control Revolution: Technological and Economic Origins of the Information Society*, Harvard University Press, Cambridge, Mass., 1986. The duality of causation is reflected quite well in the example of the use of Herman Hollerith's punch card system to increase the efficiency of the 1890 census. While Hollerith's machines cannot be blamed, there is little doubt that they were an integral part of the transformation of the national governments data gathering, processing, and distribution activities.

are likely to become more so in the future. In other cases, the need for personal information is externally motivated—for example, as a matter of regulation for the purpose of inhibiting money laundering, banks are legally required to collect and file information from customers that is not intrinsically connected to the provision of financial services.

- Employers are demanding more information about employees as they seek to validate employment credentials, to better match a person to a job, and to avoid liability suits. Would-be employees submit extensive application forms documenting previous work histories and education; once hired, they are often subject to drug tests and location checks to help ensure that they are continuing to observe the conditions of their employment. On the job, intensive work monitoring has increased, particularly as individuals work with computers and or work in areas subject to video surveillance. This may go beyond monitoring of work products per se, to the monitoring of behavior unrelated to work and sometimes behavior off-duty.

- Retailers of goods and services galore are seeking to provide more personalized products and targeted attention to their customers. From customization of goods and services to individual needs targeted at marketing specific products to selected audiences likely to buy them, detailed personal information about the preferences and habits and buying histories of customers is an enabler for personalization.

- Members of the public demand information as well. Through the legislative process, previously private information such as physician malpractice histories, sexual offender status, and political contributions are now public—and more importantly, easily available—for all to see.

- Individuals demand more information from each other in many contexts. For example, it is common that individuals—especially young people—using social networking sites post large amounts of personal information. No one forces these people to do so, and yet the social context of the sites' use provides a strong impetus for doing so.

The examples above illustrate current information demands. They also suggest how our attitudes toward privacy are context dependent. It is difficult to hold a broad view, absent consideration of what kind of information is sought, who seeks it, and how it is to be collected, protected, and used. There are, for example, some things one might not mind the government knowing that one would object to an employer knowing (and vice versa). And there are other things that one would not object to either of them knowing, but would not want passed on to aunts and uncles, just as there are things that one would like to keep within the family. Determining what should be left to the realm of ethics and common courtesy, what should be incentivized or discouraged, and what should be formalized into a code of law is yet another balancing question that comes up when contemplating privacy.

A further complicating factor is the changing nature of expectations about the revelation and concealment of personal information. Social and cultural trends over the

last century (perhaps accelerated during the 1960s) have softened traditional beliefs that opposed the easy revelation of certain kinds of personal information. Although many individuals do seek a certain measure of privacy in their lives (e.g., they purchase homes with privacy-protecting features such as enclosed porches or obscuring bushes), there has been a lessening or outright ending of reticence in mass culture as seen in the popularity of reality television shows and talk show confessionals. In addition, an emphasis in some parts of society on sharing and building trust and community through openness in communication and discussion may conflict with privacy notions regarding what is (or should be) kept as "personal information" and what should be revealed.

Finally, in some cases personal information is used to determine a category into which a given individual might fall, and what is of interest to another party is the category rather than the person.[10] The availability of personal information enables the assignment of an individual to one or more categories, such as those who share a characteristic such as age, race, or genetic marker. For example, the popularity of geo-demographic targeting for the marketing of goods and services at the neighborhood level reflects a determination that there is quite a bit of predictive utility in the differences between 100 types of communities definable at the ZIP + 4 level of precision.[11] Political parties use personal information to determine how to target their voter turnout efforts towards those most likely to vote for their candidates.[12]

Undertaken in the context of selling different products based on a zip code's socioeconomic status indicators, such a practice may be benign. Nevertheless, it is important to consider the implications of less benign applications, such as political campaigns run on a similar basis, in which different messages are targeted to different geographical areas, or redlining— the practice of denying service (or increasing the cost of service) to people in selected geographic areas—which may serve as a proxy for race, ethnicity, or income. Such issues reflect the potential for an information-based ability for discrimination of many different kinds—against individuals and against groupsin the name of increasing efficiency. (Note that this notion of discrimination is not necessarily confined to discrimination against categories of people protected by law.)

4.4 Discontinuities in Circumstance and Current Events

Current events can be important factors in shaping attitudes toward privacy.

4.4.1 National Security and Law Enforcement

The events of September 11, 2001, have led to a renewed emphasis on homeland security and how best to achieve it in the United States. The primary focus of homeland

[10]Geoffrey C. Bowker and Susan L. Star, *Sorting Things Out: Classification and Its Consequences*, MIT Press, Cambridge, Mass., 1999.

[11]See the discussion of geo-demographic clustering and commercial services offered by Claritas Corporation in Mark Monmonier, *Spying with Maps*, University of Chicago Press, 2002.

[12]See, for example, Chris Cillizza and Jim VandeHei, "In Ohio, a Battle of Databases," Washington Post, September 26, 2006, p. A-1.

security is now prevention of deliberate catastrophically harmful incidents rather than prosecution of those responsible for such acts. Prevention and disruption of a terrorist act are much more difficult than is prosecution of those responsible after the act, primarily because investigative activities can be focused much more precisely, working backward from the event.

This new focus has resulted in a number of privacy-relevant changes in the policy environment. One of the most important changes has been to elide the traditional separation of law enforcement and national security intelligence gathering. But this change poses numerous challenges, the most important of which is the final disposition of "law enforcement" information versus "national security" information. Law enforcement officials operate in a prosecutorial role, which means that "law enforcement" information must be usable in open court, along with information about its origins and provenance. "National security intelligence" information is often tied to the sources and methods used to gather it, most of which must remain secret if they are to be productive sources in the future. This means, for example, that it is not generally feasible to allow individuals about whom information has been collected to challenge the accuracy of such information, or even to notify these individuals about the fact of such collection.

A second change has been a greater willingness to focus information-gathering efforts on the continental United States. Although they were foreign citizens, the September 11 hijackers operated from U.S. soil and used U.S. airliners flying from U.S airports to strike U.S. targets. Thus, attention has been focused on identifying other possible terrorist cells operating in the United States by detecting their operational "signatures" through domestically focused information gathering and analysis. While the concerns of law enforcement and national security officials regarding the possibility of U.S.-based terrorist operations cannot be discounted, the mere fact of including information about U.S. persons within the scope of counterterrorist operations inevitably raises privacy concerns as well.

These issues are addressed at greater length in Chapter 9.

4.4.2 Disease and Pandemic Outbreak

In recent years, concerns about pandemic disease outbreaks have also advanced to the top of the public policy agenda. By definition, pandemics result from the emergence of a new disease (or a variant of an old one) that is both infectious in humans and highly contagious. Pandemics have occurred throughout human history, but the cost and time required to travel great distances have diminished now to the point where long-distance travel is within the reach of a large part of the world's population. Along with increased cultural exchange and commerce, this rapid and accessible travel, especially by airplane, has increased the chances for rapid spread of communicable diseases across local and national borders. A person may become infected with a disease and fly to a foreign country before even realizing that he or she is sick—an especially relevant point when the symptoms of the disease in question take a long time to appear.

As this report is written, world scientists are monitoring two diseases in particular,

SARS and the avian (bird) flu. In both cases, the public health response calls for rapid detection of a medical anomaly and, if possible, identification of the location and direction of spread of the disease so that, for example, quarantine and inoculation zones can be established to stem the spread of disease.

The options available from a public health standpoint to prevent pandemic outbreaks originating from outside national borders are limited. The volume of air traffic is so large that it cannot be shut down or even seriously attenuated without enormous economic consequences. Thus, the only other option is to monitor closely for the outbreak of disease in other nations and to seek to prevent those who are disease carriers from crossing one's own national borders.

Although individuals seeking to enter the United States have fewer and more limited privacy protections than they would if they were already present in-country, monitoring and obtaining information on the health of individuals have implications for privacy. Monitoring for the outbreak of disease can entail the acquisition of a great deal of personal information so that public health officials can track a disease as it spreads. But even more (potentially) invasive is the idea of obtaining information from travelers (who may be either foreign nationals or one's own citizens returning from abroad) in order to differentiate them into disease carriers and nondisease carriers. Thus, in the pursuit of public health, nations have sometimes required individuals seeking to enter to undergo tests for HIV, fill out detailed medical questionnaires, take medical examinations at the border, and undergo (sometimes covert) thermal scans that detect the presence of fever.

5 Important Concepts and Ideas Related to Privacy

Debates over privacy often make use of specialized concepts whose intuitive meaning is not necessarily clear on the face of it. Moreover, these debates are often conducted without much cognizance of the topic's long history in the public policy sphere. This section addresses key concepts, and Section 6 addresses important historical lessons.

5.1 Personal Information, Sensitive Information, and Personally Identifiable Information

Personal information can be regarded as the set of all data that is associated with a specific individual, e.g., date of birth, gender, address, name of first pet, favorite chocolate, high school of graduation, geographical location at 3:14 p.m. on March 30, 2005, and on and on and on. The specific value of any given element in that set (e.g., a date of birth January 2, 1957) can almost always be associated with more than one individual (many people were born on January 2, 1957).

Personal information thus has meaning only through the ways in which it associates or differentiates an individual from others. The value of any given data element (call that data element D_1) divides the set of all human beings in the universe into two subsets—a set S_1 comprising those with whom the value of D_1 can validly be associated,

and the complement of that set. Multiple data (D_2, D_3, D_4...) result in sets S_2, S_3, S_4. Combining the values of personal data elements D_1, D_2, D_3, D_4 means taking the intersection of S_1, S_2, S_3, S_4 (call the intersection S_1, and the number of people in S_1 the bin size. In general, S_1 has more than one person in it (i.e., the bin size is more than one). In the case when the bin size is one, S_1 has exactly one person in it, and the data values associated with S_1 can be said to uniquely specify a specific individual.

Several points are worth noting here:

- Privacy is perforce a relative concept. In a specific context, I may feel that my "privacy" is adequately protected if I can be identified within a bin size of 1,000; you may feel that your privacy is adequately protected only if you can be identified within a bin size of 10,000.

- Certain combinations of data elements can be particularly—and surprisingly—effective in reducing bin size. For example, 87 percent of the U.S. population can be uniquely specified by knowledge of his or her 5-digit ZIP code of residence, gender, and date of birth.[13] None of these individual pieces of information are individually identifying, but most of the general public would be surprised by their collective power in identifying individuals.[14]

- A person's identity (whether defined by the individual in question or others labeling him or her) is defined by some subset of personal information. By convention and for legal purposes, that subset generally includes the name of the person in question. But people often operate with multiple identities (or may have identities imposed upon them) ones identity as a parent, as an employee, as a Social Security recipient, as a member of America Online with several screen names, and so on. Reconciling multiple identities is, in essence, the process of taking the union of all of these subsets, although efforts to link multiple identities through a common identifier are often controversial. (Also, knowing a person's name will not necessarily permit access to that person if his or her location (whether in real space or in cyberspace) is unknown.)

- In general, it is the values of data elements and combinations thereof that specify unique individuals, not the data elements themselves. In some cases, "unique identifiers"—if genuinely unique—could be said to specify unique individuals. For example, ruling out the case of identical twins, an individual's complete genomic sequence (the specific sequence of all 3 billion DNA base pairs) could specify a unique individual. Barring errors and fraud, the Social Security number was

[13] Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population*, LIDAP-WP4, Laboratory for International Data Privacy, Carnegie Mellon University, Pittsburgh, Pa., 2000.

[14] Date of birth is an especially powerful tool for reducing bin size. Knowing the day of the year splits the population into 366 groups. Knowing the year of birth splits the population into an additional 90 to 100 years, depending on one's estimate of the age of the oldest individuals. Thus, date of birth alone splits a population into some 32,940 to 36,600 bins. A 5-digit ZIP code splits the population into 100,000 bins. These attributes taken together constitute approximately 3.5×10^9 bins, a number that is about 10 times the population of the United States. Thus, Sweeney's empirical result is not entirely surprising.

originally intended to be a unique identifier. But in general, no one data element specifies a unique individual.

- Unique identifiers require special protection from a privacy perspective. Because it is a data element (and not a specific value) that can be used to uniquely specify an individual, a unique identifier for a person can greatly facilitate the linkage of other information about that person and hence the collection of large amounts of information under that one identifier. When such unique identifiers fall into criminal hands, and especially when it is impossible to revoke an identifier and obtain a new one, impersonation, identity theft, and even location tracking become much easier to accomplish.

- The value of any given data element may or may not be permanently associated with a given individual. An individual's date of birth does not change, but an individual's weight does. Matters of historical fact, if recorded correctly and accurately, do not change and thus are permanent, although their meaning is subject to interpretation and those interpretations may change—e.g., what to make of an individual who undergoes a sex change operation. Names and addresses do change with some frequency, although one may be able to make some general sociodemographic inferences with knowledge of such changes over time. An individual's DNA sequence does not change throughout his or her lifetime, but the longevity and stability of many other biometric indicators have not been definitively established.

Individuals vary considerably in their privacy demands or expectations for different kinds of data and for the same individual data element in different situations. That is, in one situation, an individual may regard a particular data element as highly private (one that might require a large bin size) and in a different situation regard the same data element as not at all private (i.e., he would be perfectly fine with a bin size of one). Relevant situational factors may include:

- *The specific value of the data element and whether or not it stigmatizes or disadvantages.* For example, an HIV-positive individual may require a bin size of one million to feel that his HIV status is private; an HIV-negative individual may feel entirely comfortable with a bin size of one (i.e., being identified with certainty as being HIV-negative).

- *The stated purpose for which any given data element is requested.* The closer the fit between the goals of the supplier and the requester of information and between the information requested and the goal, the more likely it is to be provided. In most doctor-patient contexts, the patient is only too glad to offer information. If a newspaper's Web site asks a visitor her income, she may refuse to provide it, whereas she would willingly supply that same information in filling out an online application for a mortgage. Note also that if there is an incentive or reward for supplying personal information, many consumers sell that information more cheaply than their statements about the value of their personal information would lead one to expect.

- *The accessibility of the given data element.* Data that are public and hard to access (e.g., paper records, such as property taxes or divorce proceedings, that are kept in the physical facilities of many jurisdictions) are very different from data that are public and very easy to access (e.g., the same public information posted online). The ease or difficulty of finding a particular type of data element may also contribute to accessibility.

- *The transience of the given data element.* For example, when information is stored in paper form, it may be discarded eventually because it is expensive to store. There may be different privacy implications if the data element is available only for an instant (e.g., a conversation being heard in real time), for one hour, one year, one decade, or one century.[15]

The above discussion also illuminates the distinction between three categories of information—personal information, sensitive information, and personally identifiable information.[16]

- Personal information is the set of all information that is associated with a specific person X. Personal information is thus defined in a technical or objective sense.

- Sensitive information is the set of personal information that some party believes should be kept private. If the party is the person associated with that information (call that person X), the set is defined by personal preferences of X, and X's definition of private (which may be highly context dependent and linked to particular cultural standards regarding the revelation or withholding of information). Note that context may reflect a temporal aspect as well. In some circumstances, one might regard a certain item of personal information as less sensitive if it referred to his or her information "state" in the past rather than in the present. (For example, I may regard my physical location now as being a more sensitive item of information than my physical location 3 weeks ago.) The converse may be true as well. The party defining sensitive information may also be a party other than person X. This other party may take into account the interests and preferences of person X, but may also be taking other factors into consideration. For example, person X may prefer that her criminal record be kept private, but most criminal records are regarded legally as public information. Who defines what information should count as "sensitive" is often a controversial matter.

[15]Privacy is not necessarily a monotonically decreasing function of the holding period. Personal information held for 100 years after the death of the person involved is arguably nonsensitive as far as that person is concerned, although it may be highly sensitive to grandchildren if it contains genetic or severely stigmatizing information. The converse may be true as well.

[16]These and some additional distinctions are discussed in Gary T. Marx, "Varieties of Personal Information as Influences on Attitudes Towards Surveillance," in R. Ericson and K. Haggerty, eds., *The New Politics of Surveillance and Visibility*, University of Toronto Press, 2006; and "Identity and Anonymity: Some Conceptual Distinctions and Issues for Research," in J. Caplan and J.T. Torpey, *Documenting Individual Identity: The Development of the State Since the French Revolution*, Princeton University Press, Princeton, N.J., 2000.

- Personally identifiable information (PII) refers to any information that identifies or can be used to identify, contact, or locate the person to whom such information pertains. This includes information that is used in a way that is personally identifiable, including linking it with identifiable information from other sources, or from which other personally identifiable information can easily be derived, including, but not limited to, name, address, phone number, fax number, e-mail address, financial profiles, Social Security number, and credit card information.[17] Although PII is also said to not include information collected anonymously, the discussion above suggests both that the ability to make an identification may depend on the specific values of the PII in question and on the ability to aggregate data in ways that reduce significantly or even eliminate the anonymity originally promised or implied. Thus, information that previously was not PII may at a later date become PII as new techniques are developed or as other non-PII information becomes available.

5.2 False Positives, False Negatives, and Data Quality

In many societies, alleged criminals are tried by jury. In any given trial, the jury finds a defendant either innocent or guilty (apart from jury deadlocks). If a defendant found guilty did not in fact commit the crime for which he or she is being tried, the result is a "false positive." If a defendant found innocent did in fact commit the crime for which he or she is being tried, the result is a "false negative."

False positives and false negatives arise in any kind of classification exercise.[18] For example, a credit-card-issuing bank examines personal information of potential clients and classifies them as good credit risks (likely to pay their bills) and bad credit risks (unlikely to pay their bills). Some individuals identified as good credit risks will, in fact, not pay their bills—these are the false positives. Some individuals identified as bad credit risks would, in fact, pay their bills—these are the false negatives. These errors can arise either from the problems in the data or from the classification mechanism. For example, if the credit card company has information on two John Smith's mixed together, it is easy to see how a classification of John Smith might be erroneous. However, even if the data are entirely accurate, mistaken classifications are still possible, even though they would be less likely than in the case of conflated data.

Or, an intelligence analyst examines financial transactions and phone records of a set of individuals, searching for possible indications of terrorist planning. He classifies them as "unlikely to be involved in terrorist activity" and "likely to be involved in terrorist activity," and sends only those in the latter category up the chain of command for further investigation. A false positive is someone in the latter category who, upon further investigation, has no terrorist connection at all. A false negative is someone in

[17] This definition is a commonly used one, although the precise wording may vary depending on the user in question.

[18] An extensive treatment of false positives and false negatives (and the tradeoffs thereby implied) can be found in National Research Council, *The Polygraph and Lie Detection*, The National Academies Press, Washington, D.C., 2003.

the former category who should have received further investigation but did not.

Two important points arise in this discussion.

- For a given database and given analytical approach, false positives and false negatives are in some sense complementary. More precisely, for a given database, one can drive the rate of false positives to zero, or the rate of false negatives to zero, but not simultaneously. For example, it is easy to identify all individuals who are bad credit risks—just deny everyone credit. This approach catches all of the bad credit risks—but also results in a huge number of false negatives. Decreases in the false positive rate are inevitably accompanied by increases in the false negative rate, and vice versa, though not necessarily in the same proportion. However, if the quality of the data is improved, or if the classification technique is improved, it is possible to reduce both the false positive rate and the false negative rate.
- Identifying false negatives in any given instance may be problematic. In the case of credit card issuers, the bank will probably not issue cards to the bad credit risks. Thus, it may never learn that these individuals are in fact creditworthy—and these individuals may forevermore be saddled with another declination of credit on their records without being given the chance to prove their creditworthiness. In the case of the terrorist investigation, it is essentially impossible to know if a person is a false negative until he or she commits the terrorist act.

False positives and false negatives are important in a discussion of privacy because they are the language in which the tradeoffs described in Section 2 are often cast. Banks obtain personal information on individuals for the purpose of evaluating their creditworthiness. All of these individuals surrender some financial privacy, but some do not receive the benefit of obtaining credit, and some of those not receiving credit are deserving of credit. A law enforcement official may obtain personal information on individuals searching evidence of criminal activity. All of these individuals surrender some privacy, and those who have not been involved in criminal activity have had their privacy violated despite the lack of such involvement.

Data quality is the property of data that allows them to be used effectively, economically, and rapidly to inform and evaluate decisions.[19] Typically, data should be correct, current, complete, and relevant. Data quality is intimately related to false positives and false negatives, in that it is intuitively obvious that using data of poor quality is likely to result in larger numbers of false positives and false negatives than would be the case if the data were of high quality.

[19]Alan F. Karr, Ashish P. Sanil, and David L. Banks, "Data Quality: A Statistical Perspective," *Statistical Methodology* 3:137-173, 2006; Thomas C. Redman, "Data: An Unfolding Quality Disaster," *DM Review Magazine*, August 2004, available at `http://www.dmreview.com/article_sub.cfm?articleId=1007211`; Wayne W. Eckerson, "Data Warehousing Special Report: Data Quality and the Bottom Line," May 1, 2002, available at `http://www.adtmag.com/article.aspx?id=6321&page=`; Y. Wand and R. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM* 39(11):86-95, November 1996; and R. Wang, H. Kon, and S. Madnick, "Data Quality Requirements Analysis and Modelling," Ninth International Conference of Data Engineering, Vienna, Austria, 1993.

Data quality is a multidimensional concept. Measurement error and survey uncertainty contribute (negatively) to data quality, as do issues related to measurement bias. But in the context of using large-scale data sets assembled by multiple independent parties using different definitions and processes, many other issues come to the fore as well.

It is helpful to distinguish between issues related to data quality in a single database and data quality associated with a collection of databases. Data quality issues for a single database include (but are not limited to) missing data fields; inconsistent data fields in a given record, such as recording a pregnancy for a 9-year-old boy; data incorrectly entered into the database, such as that which might result from a typographical error; measurement error; sampling error and uncertainty; timeliness (or lack thereof); coverage or comprehensiveness (or lack thereof); improperly duplicated records; data conversion errors, as might occur when a database of vendor X is converted to a comparable database using technology from vendor Y; use of inconsistent definitions over time; and definitions that become irrelevant over time.

Data quality issues for multiple databases include all of those issues for a single database, and also syntactic inconsistencies (one database records phone numbers in the form 202-555-1212 and another in the form 2025551212); semantic inconsistencies (weight measured in pounds vs. weight measured in kilograms); different provenance for different databases; inconsistent data fields for records contained in different databases on a given data subject; and lack of universal identifiers to specify data subjects.

5.3 Privacy and Anonymity

Privacy is an umbrella concept within which anonymity is located. A vandal may break a window, but his or her identity may not be directly known. Someone may send an unsigned or pseudonymous e-mail, or make a charitable contribution. Anonymity may involve a protected right, as in the delivery of political messages. Or it may simply be an empirical condition generated by stealth or circumstance. Unsigned graffiti illustrates the former and "faceless" individuals in a crowd the latter.

The distinction between privacy and anonymity is clearly seen in an information technology context. Privacy corresponds to being able to send an encrypted e-mail to another recipient. Anonymity corresponds to being able to send the contents of the e-mail in plain, easily readable form but without any information that enables a reader of the message to identify the person who wrote it. Privacy is important when the contents of a message are at issue, whereas anonymity is important when the identity of the author of a message is at issue. Depending on the context, privacy expectations (and actualities apart from the rules) may extend to content or the identity of the sender or to both.

The relationship between privacy and anonymity can be made more formal. If personal information about an individual is denoted by the set P, the individual has privacy to the extent that he or she can keep the value of any element in the set private. Consider then another set Q, a subset of P, which consists of all elements that could

be used—individually or in combination—to identify the individual. The anonymity of the individual thus depends on keeping Q private.

For example, one might define a number of different sets: the set of all people with black hair, the set of all people who work for the National Academies, the set of all people who type above a certain rate, and so on. Knowledge that an individual is in any one of these sets does not identify that individual uniquely—he or she is thus "anonymous" in the usual meaning of the term. But knowledge that an individual is in all of these sets—that is, considering the intersection of all of these sets—might well result in the ability to identify the individual uniquely (and hence in the loss of anonymity).[20]

Note also that anonymity is often tied to the identification of an individual rather than the specification of that individual. A person may be specified by his or her complete genomic sequence, but in the absence of databases that tie that sequence to a specific identity the person is still anonymous. A fingerprint may be found on a gun used in a murder, but the fingerprint does not directly identify the shooter unless the fingerprint is on file in some law enforcement databank. In short, the specification of a unique individual is not necessarily the same thing as identifying that individual.[21]

An additional consideration is that "identification" usually means unique identification —using any of these sets would result in a bin size of one. In other words, in the usual discussion of anonymity, an anonymous person is someone whose identity cannot be definitively ascertained. However, for some purposes, a bin size of three would be insufficient to protect his or her identity—if a stool pigeon for an organized crime syndicate were kept "anonymous" within a bin size of three, it is easy to imagine that the syndicate would be perfectly willing and able to execute three murders rather than one. Here again is a situational factor that contributes to the relative nature of such concepts.

The anonymity dimension of privacy is central to the problem of protecting data collected for statistical purposes. For example, many agencies of the federal government collect information about the state of the nation—from the national economy to household use of Medicare—in order to evaluate existing programs and to develop new ones. That information is often derived from data collected by statistical agencies or others under a pledge of confidentiality. A most critical data source is microdata, which includes personal information about individuals, households, and businesses, and a central concern of the federal statistical agencies is that the responses provided by

[20]More precisely, Q is the set of all subsets of P that could be used to identify the individual. Imagine that elements P_2, P_4, P_{17} of P could be used together to identify the individual, as could elements P_2, P_3, P_{14} taken together, and elements P_3, P_7, P_{14}. Then anonymity would require that these three sets be kept private, that is $\{P_2, P_4, P_{17}\}$, $\{P_2, P_3, P_{14}\}$, and $\{P_3, P_7, P_{14}\}$. In practice, this might well imply keeping private the union of all these sets $\{P_2, P_3, P_4, P_7, P_{14}, P_{17}\}$.

[21]It is worth noting that despite the common "intuitively obvious" usage of the term "identity," identity is fundamentally a social construct and thus has meaning only in context. I may know a person who sends me e-mail only by his or her e-mail address, but the identity "JohnL7534@yahoo.com" may be entirely sufficient for our relationship—and it may not matter if his first name is really John, whether his last name begins with L, or even whether this person is male or female. In this sense, specification might be regarded as a decontextualized identification.

information providers will be less candid if their confidentiality cannot be guaranteed.[22] (This issue is addressed at greater length in Section 8 of Chapter 6.)

This issue also arises explicitly, although in a somewhat different form, in contemplating the significance of an organization's privacy—that is, information about an organization with whom a number of individuals may be associated. Information about an organization can reveal information about individuals, although it may not be uniquely associated with an individual. For example, if a survey of employers shows that a company pays a large amount in employee health care benefits to medical care providers that specialize in treating AIDS, then it can be inferred that some employees of that company have AIDS. This fact may have significance for all of the employees—those with AIDS may face a greater likelihood of having their status revealed, and those without AIDS may face higher health care premiums in the future if their past employment history becomes known.

5.4 Fair Information Practices

Fair information practices are standards of practice required to ensure that entities that collect and use personal information provide adequate privacy protection for that information. These practices include notice to and awareness of individuals with personal information that such information is being collected, providing individuals with choices about how their personal information may be used, enabling individuals to review the data collected about them in a timely and inexpensive way and to contest that data's accuracy and completeness, taking steps to ensure that the personal information of individuals is accurate and secure, and providing individuals with mechanisms for redress if these principles are violated.

Fair information practices were first articulated in a comprehensive manner in the U.S. Department of Health, Education, and Welfare's 1973 report *Records, Computers and the Rights of Citizens*.[23] This report was the first to introduce the Code of Fair Information Practices (Box 1.3), which has proven influential in subsequent years in shaping the information practices of numerous private and governmental institutions and is still well accepted as the gold standard for privacy protection.[24]

---

[22]See, for example, National Research Council, *Expanding Access to Research Data: Reconciling Risks and Opportunities*, The National Academies Press, Washington, D.C., 2005; National Research Council, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, National Academy Press, Washington, D.C., 1993.

[23]U.S. Department of Health, Education, and Welfare, *Records, Computers and the Rights of Citizens*, Report of the Secretary's Advisory Committee on Automated Personal Data Systems, MIT Press, Cambridge, Mass., 1973.

[24]Fair information principles are a staple of the privacy literature. See, for example, the extended discussion of these principles in D. Solove, M. Rotenberg, and P. Schwartz, *Information Privacy Law*, Aspen Publishers, 2006; Alan Westin, "Social and Political Dimensions of Privacy," *Journal of Social Issues* 59(2):431-453, 2003; Helen Nissenbaum, "Privacy as Contextual Integrity," *Washington Law Review* 79:101-139, 2004; and an extended discussion and critique in Roger Clarke, "Beyond the OECD Guidelines: Privacy Protection for the 21st Century," available at `http://www.anu.edu.au/people/Roger.Clarke/DV/PP21C.html`.

Box 1.3
Codes of Fair Information Practice

Fair information practices are standards of practice required to ensure that entities that collect and use personal information provide adequate privacy protection for that information. As enunciated by the U.S. Federal Trade Commission (other formulations of fair information practices exist),[1] the five principles of fair information practice include:

- *Notice and awareness.* Secret record systems should not exist. Individuals whose personal information is collected should be given notice of a collector's information practices before any personal information is collected and should be told that personal information is being collected about them. Without notice, an individual cannot make an informed decision as to whether and to what extent to disclose personal information. Notice should be given about the identity of the party collecting the data, how the data will be used and the potential recipients of the data, the nature of the data collected and the means by which it is collected, whether the individual may decline to provide the requested data and the consequences of a refusal to provide the requested information, and the steps taken by the collector to ensure the confidentiality, integrity, and quality of the data.
- *Choice and consent.* Individuals should be able to choose how personal information collected from them may be used, and in particular how it can be used in ways that go beyond those necessary to complete a transaction at hand. Such secondary uses can be internal to the collector's organization, or can result in the transfer of the information to third parties. Note that genuinely informed consent is a sine qua non for observation of this principle. Individuals who provide personal information under duress or threat of penalty have not provided informed consentand individuals who provide personal information as a requirement for receiving necessary or desirable services from monopoly providers of services have not, either.
- *Access and participation.* Individuals should be able to review in a timely and inexpensive way the data collected about them, and to similarly contest that data's accuracy and completeness. Thus, means should be available to correct errors, or at the very least, to append notes of explanation or challenges that would accompany subsequent distributions of this information.
- *Integrity and security.* The personal information of individuals must be accurate and secure. To assure data integrity, collectors must take reasonable steps, such as using only reputable sources of data and cross-referencing data against multiple sources, providing consumer access to data, and destroying untimely data or converting it to anonymous form. To provide security, collectors must take both procedural and technical measures to protect against loss and the unauthorized access, destruction, use, or disclosure of the data.
- *Enforcement and redress.* Enforcement mechanisms must exist to ensure that the fair information principles are observed in practice, and individuals must have redress mechanisms available to them if these principles are violated.

[1]See http://www.ftc.gov/reports/privacy3/fairinfo.htm.

From their origin in 1973, fair information practices "became the dominant U.S. approach to information privacy protection for the next three decades."[25] The five principles not only became the common thread running through various bits of sectoral regulation developed in the United States, but they also were reproduced, with significant extension, in the guidelines developed by the Organisation for Economic Cooperation and Development (OECD). These principles are extended in the context of OECD guidelines that govern "the protection of privacy and transborder flows of personal data" and include eight principles that have come to be understood as "minimum standards...for the protection of privacy and individual liberties."[26] They also include a statement about the degree to which data controllers should be accountable for their actions. This generally means that there are costs associated with the failure of a data manager to enable the realization of these principles.

5.5 Reasonable Expectations of Privacy

A common phrase in discussions of privacy is "reasonable expectation of privacy." The phrase has a long history in case law, first introduced in *Katz v. United States*, 389 U.S. 347 (1967), that reflects the fact that expectations are shaped by tradition, common social practices, technology, law, regulations, the formal and informal policies of organizations that are able to establish their own rules for the spaces that they control, and the physical and social context of any given situation. Expectations of privacy vary depending on many factors, but place and social relationships are among the most important.

Historically, the home has been the locale in which the expectation of privacy has been the most extensive and comprehensive. Yet there are different zones of privacy even within the home, and within the sets of interpersonal relationships that are common to one's home. While customs vary across cultures and individual families, there is a well-distributed sense of the nature of these spatial boundaries within the home. Kitchens and living rooms are common or relatively public spaces within the home, and they are places into which outsiders may be invited on special occasions. Bedrooms and bathrooms tend to be marked off from the more public or accessible spaces within the home because of the more intimate and personal activities that are likely to take place within them.

In U.S. workplaces, individuals have only very limited expectations of privacy. The loss of privacy begins for many with the application, and reaches quite personal levels

[25]Westin, "Social and Political Dimensions of Privacy," 2003, p. 436.
[26]Marc Rotenberg, *The Privacy Law Sourcebook 2001*, Electronic Privacy Information Center, 2001, pp. 270-272.

for those jobs that require drug tests and personality assessments. On the other hand, privacy does not evaporate entirely on the job. Closets may be provided for the storage of personal effects, and depending on the relative permanence of assigned spaces, desk drawers may be treated as personal space. The presence or absence of doors within workspaces affects the ability of workers to control direct observation by others.

Technology also affects reasonable expectations of privacy. Technology can be used to enhance human senses and cognitive capabilities, and these enhancements can affect the ability to collect information at a distance. The result is that space is not the marker it once was for indicating boundaries between private and public interactions. In the case of information technology, the "objects" about which one is private (digital objects such as electronic files or streams of bits as communications) are quite distinct from objects that were originally the focus of privacy concerns (physical, tangible objects made of atoms). Thus, Kerr argues, for example, that the well-established history of Fourth Amendment law governing permissible searches (and also reasonable expectations of privacy) must be rethought in light of the manifest differences between physical and digital objects.[27]

Critical events such as the terrorist attacks of 2001 have dramatically increased the level of personal and records surveillance that travelers encounter. Heightened concern about threats of violence means that searches of personal effects are becoming more common at sporting events, popular tourist sites, and even schools.

Formal and informal policies that define the boundaries between the public and the private also help to shape our expectations of privacy that develop over time. Privacy policies are not only established by legislatures, administrative agencies, and the courts. Individual firms, trade unions, professional associations, and a host of other institutional actors have also developed policies to govern the collection and use of personal information. Individuals also have policies, or norms, that govern the ways in which they will interact with organizations and with other individuals. Indeed, individuals' reciprocal behavior with respect to asking for, and offering, information is conditioned by custom and manners that are no less significant for not being less formal than the written policies.

Cross-cultural differences with respect to expectations of privacy can be noted. For example, compared to Western cultures, a number of Eastern cultures place a far lower value on certain kinds of privacy in the home, and an Asian child often grows up with very different expectations of privacy than might an American child.

Finally, the concept of "reasonable expectations of privacy" has a normative meaning as well as a descriptive meaning. For example, in a world where electronic surveillance technologies make surveillance easy to conduct on a wide scale, one could argue that no one today has a "reasonable expectation" that his or her phone calls will not be tapped. But both statutory law (e.g., Title III in the U.S. Code) and case law (e.g.,

[27] Orin Kerr, "Searches and Seizures in a Digital World," *Harvard Law Review* 119:531, 2005. Kerr's normative reformulation of Fourth Amendment law calls for maintaining "the specific goals of specific doctrinal rules in light of changing facts," although he clearly recognizes that other normative reformulations are possible.

Katz v. United States, 389 U.S. 347 (1967)) stipulate that under most circumstances, an individual does have a reasonable expectation that his or her phone calls will not be tapped.

6 Lessons From History

In the history of the United States, a number of societal shifts have taken place that relate to contemporary visions of privacy (Appendix A). For example, a move from primarily rural to a more urban (or suburban) society resulted in changes to the scale of one's community and increased ones proximity to strangers. In addition, the impact of information technologies is often to compress time and distance in the social sphere, and one result has been an increasingly diminished utility of time and space as markers of the boundaries between private and public space. Associated changes in how trust is developed and sustained have all shaped our understanding and appreciation of the value of privacy and the limits on it in a more impersonal society.

Furthermore, there is an increased appetite on the part of many sectors of society for information collection and analysis and verification. The kinds of interactions individuals have with institutions and with each other have changed as a result. Increased societal needs, increased interdependence, new kinds of risks, ever greater complexity, and an increase in the number of rules one needs to be aware of to move safely and smoothly through society have radically altered the kinds of interactions individuals have with institutions and with each other. Both private organizations and government agencies are increasingly concerned with the ability to document compliance and discover violations. This is a major motivation for collection of information about individuals and about organizations.

As the discussion in Appendix A (on the history of surveillance and privacy in the United States) suggests, a number of lessons can be gleaned from history. The first is that surveillance has been intensifying as society has grown more complex.[28]

The second lesson is that each technological advance in the spheres of sensing, communication, and information processing invites greater surveillance, and often those invitations are accepted. The invention of the telegraph led almost immediately to the invention of wiretapping. The invention of automated fingerprint matching led to the FBI's integrated automated fingerprint identification system. The development of the computer resulted in unprecedented record-keeping power, and the emergence of networking technology has further increased that power. This is not to suggest that technologies make things happen on their own, but they do facilitate the activities and ambitions of those who might use them and who can afford the costs of those new technologies.

The third lesson is that times of crisis or war are often marked by contractions

[28] Living in small towns or tightly knit communities is often associated with lesser degrees of privacy (where "everyone knows everyone else's business"). But lesser privacy in these communities is not generally the result of explicit acts of surveillance or information gathering—rather, it is a by-product of routine day-to-day living.

in the scope of civil liberties. Often, when U.S. government leaders have come to believe that the security or the core interests of the nation were being threatened from without, the government has increased its surveillance of groups within its borders. In case after case whether British Loyalists, or Japanese-Americans, or Arab-Americans, the unequal weight of government surveillance on these groups has been justified on the basis of alleged links between the groups in question and threats to the national interest. Moreover, as the putative threat from these groups has faded with history, actions taken against these groups have generally been regarded with a degree of retrospective shame.

The fourth lesson is that although U.S. conceptions of privacy can be traced historically, the meaning of the concept has been highly varied and vague, and there has never been an agreed-upon meaning. One result is that the legal and regulatory framework surrounding privacy has been a patchwork without a unifying theme or driving principles. This state of affairs in the United States contrasts sharply with those of certain other nations (notably the member states of the European Union) that often take a more comprehensive approach to privacy-related issues. This point is discussed further in Chapter 4.

7 Scope and Map of This Report

This report examines privacy from several perspectives and offers analysis and ways of thinking through privacy questions at the same time that it provides a snapshot of the current state of affairs.

Part I is this chapter (Chapter 1).

Part II includes Chapters 2 through 5, which are primarily expository. Chapters 2 and 3 seek to lay the groundwork for what privacy is and how it affects and is affected by societal and technological complexities. Chapters 4 and 5 address the legal landscape of privacy in the United States and the political forces shaping that landscape throughout recent history.

Part III (Chapters 6 through 9) considers privacy in context, examining privacy issues in different sectors of society. Chapter 6 looks at institutional practice in privacy broadly in several different sectors. Chapter 7 provides a more in-depth look at health and medical privacy. Chapter 8 explores privacy and the U.S. library community and also mentions the issue of intellectual property and privacy (where technology, policy, and privacy intersect strongly). Chapter 9 looks at law enforcement and national security.

Part II can be skipped without loss of continuity if the reader wishes to consider the various case studies first in Part III. However, Parts I and II supply important background information that provides a context for Part III.

Part IV consists of a single and final chapter (Chapter 10) and provides the bulk of the report's look to the future. It examines mechanisms and options for privacy protection and presents the report's findings and recommendations.

Appendix A presents a short history of surveillance and privacy in the United States. Appendix B provides a look at international considerations.