identification card causes a furious libertarian reaction from parties not usually outspoken in defense of individual freedom. Under the REAL ID act of 2005, uniform federal standards are being implemented for state-issued drivers' licenses. Although it passed through Congress without debate, the law is opposed by at least 18 states. Resistance pushed back the implementation timetable first to 2009, and then, in early 2008, to 2011. Yet even fully implemented, REAL ID would fall far short of the true national ID preferred by those charged with fighting crime and preventing terrorism.

As the national ID card debate continues in the U.S., the FBI is making it irrelevant by exploiting emerging technologies. There would be no need for anyone to carry an ID card if the government had enough biometric data on Americans—that is, detailed records of their fingerprints, irises, voices, walking gaits, facial features, scars, and the shape of their earlobes. Gather a combination of measurements on individuals walking in public places, consult the databases, connect the dots, and—bingo!—their names pop up on the computer screen. No need for them to carry ID cards; the combination of biometric data would pin them down perfectly.

As the national ID card debate continues in the U.S., the FBI is making it irrelevant by exploiting emerging technologies.

Well, only imperfectly at this point, but the technology is improving. And the data is already being gathered and deposited in the data vault of the FBI's Criminal Justice Information Services database in Clarksburg, West Virginia. The database already holds some 55 million sets of fingerprints, and the FBI processes 100,000 requests for matches every day. Any of 900,000 federal, state, and local law enforcement officers can send a set of prints and ask the FBI to identify it. If a match comes up, the individual's criminal history is there in the database too.

But fingerprint data is hard to gather; mostly it is obtained when people are arrested. The goal of the project is to get identifying information on nearly everyone, and to get it without bothering people too much. For example, a simple notice at airport security could advise travelers that, as they pass through airport security, a detailed "snapshot" will be taken as they enter the secure area. The traveler would then know what is happening, and could have refused (and stayed home). As an electronic identification researcher puts it, "That's the key. You've chosen it. You have chosen to say, 'Yeah, I want this place to recognize me.'" No REAL ID controversies, goes the theory; all the data being gathered would, in some sense at least, be offered voluntarily.

Friendly Cooperation Between Big Siblings

In fact, there are two Big Brothers, who often work together. And we are, by and large, glad they are watching, if we are aware of it at all. Only occasionally are we alarmed about their partnership.

The first Big Brother is Orwell's—the government. And the other Big Brother is the industry about which most of us know very little: the business of aggregating, consolidating, analyzing, and reporting on the billions of individual transactions, financial and otherwise, that take place electronically every day. Of course, the commercial data aggregation companies are not in the spying business; none of their data reaches them illicitly. But they do know a lot about us, and what they know can be extremely valuable, both to businesses and to the government.

The new threat to privacy is that computers can extract significant information from billions of apparently uninteresting pieces of data, in the way that mining technology has made it economically feasible to extract precious metals from low-grade ore. Computers can correlate databases on a massive level, linking governmental data sources together with private and commercial ones, creating comprehensive digital dossiers on millions of people. With their massive data storage and processing power, they can make connections in the data, like the clever connections the MIT students made with the Chicago homicide data, but using brute force rather than ingenuity. And the computers can discern even very faint traces in the data—traces that may help track payments to terrorists, set our insurance rates, or simply help us be sure that our new babysitter is not a sex offender.

And so we turn to the story of the government and the aggregators.

Acxiom is the country's biggest customer data company. Its business is to aggregate transaction data from all those swipes of cards in card readers all over the world—in 2004, this amounted to more than a billion transactions a day. The company uses its massive data about financial activity to support the credit card industry, banks, insurers, and other consumers of information about how people spend money. Unsurprisingly, after the War on Terror began, the Pentagon also got interested in Acxiom's data and the ways they gather and analyze it. Tracking how money gets to terrorists might help find the terrorists and prevent some of their attacks.

ChoicePoint is the other major U.S. data aggregator. ChoicePoint has more than 100,000 clients, which call on it for help in screening employment candidates, for example, or determining whether individuals are good insurance risks.

Acxiom and ChoicePoint are different from older data analysis operations, simply because of the scale of their operations. Quantitative differences have

qualitative effects, as we said in Chapter 1; what has changed is not the technology, but rather the existence of rich data sources. Thirty years ago, credit cards had no magnetic stripes. Charging a purchase was a mechanical operation; the raised numerals on the card made an impression through carbon paper so you could have a receipt, while the top copy went to the company that issued the card. Today, if you charge something using your CapitalOne card, the bits go instantly not only to CapitalOne, but to Acxiom or other aggregators. The ability to search through huge commercial data sources—including not just credit card transaction data, but phone call records, travel tickets, and banking transactions, for example—is another illustration that more of the same can create something new.

Privacy laws do exist, of course. For a bank, or a data aggregator, to post your financial data on its web site would be illegal. Yet privacy is still developing as an area of the law, and it is connected to commercial and government interests in uncertain and surprising ways.

A critical development in privacy law was precipitated by the presidency of Richard Nixon. In what is generally agreed to be an egregious abuse of presidential power, Nixon used his authority as president to gather information on those who opposed him—in the words of his White House Counsel at the time, to "use the available federal machinery to screw our political enemies." Among the tactics Nixon used was to have the Internal Revenue Service audit the tax returns of individuals on an "enemies list," which included congressmen, journalists, and major contributors to Democratic causes. Outrageous as it was to use the IRS for this purpose, it was not illegal, so Congress moved to ban it in the future.

The Privacy Act of 1974 established broad guidelines for when and how the Federal Government can assemble dossiers on citizens it is not investigating for crimes. The government has to give public notice about what information it wants to collect and why, and it has to use it only for those reasons.

The Privacy Act limits what the government can do to gather information about individuals and what it can do with records it holds. Specifically, it states, "No agency shall disclose any record which is contained in a system of records by any means of communication to any person, or to another agency, except pursuant to a written request by, or with the prior written consent of, the individual to whom the record pertains, unless ...." If the government releases information inappropriately, even to another government agency, the affected citizen can sue for damages in civil court. The protections provided by the Privacy Act are sweeping, although not as sweeping as they may seem. Not every government office is in an "agency"; the courts are not, for example. The Act requires agencies to give public notice of the uses to which they will put the information, but the notice can be buried in the

Federal Register where the public probably won't see it unless news media happen to report it. Then there is the "unless" clause, which includes significant exclusions. For example, the law does not apply to disclosures for statistical, archival, or historical purposes, civil or criminal law enforcement activities, Congressional investigations, or valid Freedom of Information Act requests.

In spite of its exclusions, government practices changed significantly because of this law. Then, a quarter century later, came 9/11. *Law enforcement should have seen it all coming*, was the constant refrain as investigations revealed how many unconnected dots were in the hands of different government agencies. *It all could have been prevented if the investigative fiefdoms had been talking to each other. They should have been able to connect the dots.* But they could not—in part because the Privacy Act restricted inter-agency data transfers. A response was badly needed. The Department of Homeland Security was created to ease some of the interagency communication problems, but that government reorganization was only a start.

In January 2002, just a few months after the World Trade Center attack, the Defense Advanced Research Projects Agency (DARPA) established the Information Awareness Office (IAO) with a mission to:

> imagine, develop, apply, integrate, demonstrate, and transition information technologies, components and prototype, closed-loop, information systems that will counter asymmetric threats by achieving total information awareness useful for preemption; national security warning; and national security decision making. The most serious asymmetric threat facing the United States is terrorism, a threat characterized by collections of people loosely organized in shadowy networks that are difficult to identify and define. IAO plans to develop technology that will allow understanding of the intent of these networks, their plans, and potentially define opportunities for disrupting or eliminating the threats. To effectively and efficiently carry this out, we must promote sharing, collaborating, and reasoning to convert nebulous data to knowledge and actionable options.

Vice Admiral John Poindexter directed the effort that came to be known as "Total Information Awareness" (TIA). The growth of enormous private data repositories provided a convenient way to avoid many of the prohibitions of the Privacy Act. The Department of Defense can't get data from the Internal Revenue Service, because of the 1974 Privacy Act. *But they can both buy it from private data aggregators!* In a May 2002 email to Adm. Poindexter, Lt. Col Doug Dyer discussed negotiations with Acxiom.

> Acxiom's Jennifer Barrett is a lawyer and chief privacy officer. She's testified before Congress and offered to provide help. One of the key suggestions she made is that people will object to Big Brother, wide-coverage databases, but they don't object to use of relevant data for specific purposes that we can all agree on. Rather than getting all the data for any purpose, we should start with the goal, tracking terrorists to avoid attacks, and then identify the data needed (although we can't define all of this, we can say that our templates and models of terrorists are good places to start). Already, this guidance has shaped my thinking.
>
> Ultimately, the U.S. may need huge databases of commercial transactions that cover the world or certain areas outside the U.S. This information provides economic utility, and thus provides two reasons why foreign countries would be interested. Acxiom could build this mega-scale database.

The *New York Times* broke the story in October 2002. As Poindexter had explained in speeches, the government had to "break down the stovepipes" separating agencies, and get more sophisticated about how to create a big picture out of a million details, no one of which might be meaningful in itself. The *Times* story set off a sequence of reactions from the Electronic Privacy Information Center and civil libertarians. Congress defunded the office in 2003. Yet that was not the end of the idea.

The key to TIA was data mining, looking for connections across disparate data repositories, finding patterns, or "signatures," that might identify terrorists or other undesirables. The General Accountability Office report on Data Mining (GAO-04-548) reported on their survey of 128 federal departments. They described 199 separate data mining efforts, of which 122 used personal information.

Although IAO and TIA went away, Project ADVISE at the Department of Homeland Security continued with large-scale profiling system development. Eventually, Congress demanded that the privacy issues concerning this program be reviewed as well. In his June 2007 report (OIG-07-56), Richard Skinner, the DHS Inspector General, stated that "program managers did not address privacy impacts before implementing three pilot initiatives," and a few weeks later, the project was shut down. But ADVISE was only one of twelve data-mining projects going on in DHS at the time.

Similar privacy concerns led to the cancellation of the Pentagon's TALON database project. That project sought to compile a database of reports of

suspected threats to defense facilities as part of a larger program of domestic counterintelligence.

The Transportation Security Administration (TSA) is responsible for airline passenger screening. One proposed system, CAPPS II, which was ultimately terminated over privacy concerns, sought to bring together disparate data sources to determine whether a particular individual might pose a transportation threat. Color-coded assessment tags would determine whether you could board quickly, be subject to further screening, or denied access to air travel.

The government creates projects, the media and civil liberties groups raise serious privacy concerns, the projects are cancelled, and new ones arise to take their place. The cycle seems to be endless. In spite of Americans' traditional suspicions about government surveillance of their private lives, the cycle seems to be almost an inevitable consequence of Americans' concerns about their security, and the responsibility that government officials feel to use the best available technologies to protect the nation. Corporate databases often contain the best information on the people about whom the government is curious.

Technology Change and Lifestyle Change

New technologies enable new kinds of social interactions. There were no suburban shopping malls before private automobiles became cheap and widely used. Thirty years ago, many people getting off an airplane reached for cigarettes; today, they reach for cell phones. As Heraclitus is reported to have said 2,500 years ago, "all is flux"—everything keeps changing. The reach-for-your-cell phone gesture may not last much longer, since airlines are starting to provide onboard cell phone coverage.

The more people use a new technology, the more useful it becomes. (This is called a "network effect"; see Chapter 4, "Needles in the Haystack.") When one of us got the email address `lewis@harvard` as a second-year graduate student, it was a vainglorious joke; all the people he knew who had email addresses were students in the same office with him. Email culture could not develop until a lot of people had email, but there wasn't much point in having email if no one else did.

Technology changes and social changes reinforce each other. Another way of looking at the technological reasons for our privacy loss is to recognize that the social institutions enabled by the technology are now more important than the practical uses for which the technology was originally conceived. Once a lifestyle change catches on, we don't even think about what it depends on.

Credit Card Culture

The usefulness of the data aggregated by Acxiom and its kindred data aggregation services rises as the number of people in their databases goes up, and as larger parts of their lives leave traces in those databases. When credit cards were mostly short-term loans taken out for large purchases, the credit card data was mostly useful for determining your creditworthiness. It is still useful for that, but now that many people buy virtually everything with credit cards, from new cars to fast-food hamburgers, the credit card transaction database can be mined for a detailed image of our lifestyles. The information is there, for example, to determine if you usually eat dinner out, how much traveling you do, and how much liquor you tend to consume. Credit card companies do in fact analyze this sort of information, and we are glad they do. If you don't seem to have been outside Montana in your entire life and you turn up buying a diamond bracelet in Rio de Janeiro, the credit card company's computer notices the deviation from the norm, and someone may call to be sure it is really you.

The credit card culture is an economic problem for many Americans, who accept more credit card offers than they need, and accumulate more debt than they should. But it is hard to imagine the end of the little plastic cards, unless even smaller RFID tags replace them. Many people carry almost no cash today, and with every easy swipe, a few more bits go into the databases.

Email Culture

Email is culturally in between telephoning and writing a letter. It is quick, like telephoning (and instant messaging is even quicker). It is permanent, like a letter. And like a letter, it waits for the recipient to read it. Email has, to a great extent, replaced both of the other media for person-to-person communication, because it has advantages of both. But it has the problems that other communication methods have, and some new ones of its own.

Phone calls are not intended to last forever, or to be copied and redistributed to dozens of other people, or to turn up in court cases. When we use email as though it were a telephone, we tend to forget about what else might happen to it, other than the telephone-style use, that the recipient will read it and throw it away. Even Bill Gates probably wishes that he had written his corporate emails in a less telephonic voice. After testifying in an antitrust lawsuit that he had not contemplated cutting a deal to divide the web browser market with a competitor, the government produced a candid email he had sent, seeming to contradict his denial: "We could even pay them money as part of the deal, buying a piece of them or something."

Email is as public as postcards, unless it is encrypted, which it usually is not.

Email is bits, traveling within an ISP and through the Internet, using email software that may keep copies, filter it for spam, or submit it to any other form of inspection the ISP may choose. If your email service provider is Google, the point of the inspection is to attach some appropriate advertising. If you are working within a financial services corporation, your emails are probably logged—even the ones to your grandmother—because the company has to be able to go back and do a thorough audit if something inappropriate happens.

Email is as public as postcards, unless it is encrypted, which it usually is not. Employers typically reserve the right to read what is sent through company email. Check the policy of your own employer; it may be hard to find, and it may not say what you expect. Here is Harvard's policy, for example:

> Employees must have no expectation or right of privacy in anything they create, store, send, or receive on Harvard's computers, networks, or telecommunications systems. .... Electronic files, e-mail, data files, images, software, and voice mail may be accessed at any time by management or by other authorized personnel for any business purpose. Access may be requested and arranged through the system(s) user, however, this is not required.

Employers have good reason to retain such sweeping rights; they have to be able to investigate wrongdoing for which the employer would be liable. As a result, such policies are often less important than the good judgment and ethics of those who administer them. Happily, Harvard's are generally good. But as a general principle, the more people who have the authority to snoop, the more likely it is that someone will succumb to the temptation.

Commercial email sites can retain copies of messages even after they have been deleted. And yet, there is very broad acceptance of public, free, email services such as Google's Gmail, Yahoo! Mail, or Microsoft's Hotmail. The technology is readily available to make email private: whether you use encryption tools, or secure email services such as Hushmail, a free, web-based email service that incorporates PGP-based encryption (see Chapter 5). The usage of these services, though, is an insignificant fraction of their unencrypted counterparts. Google gives us free, reliable email service and we, in return, give up some space on our computer screen for ads. Convenience and cost trump privacy. By and large, users don't worry that Google, or its competitors, have all their mail. It's a bit like letting the post office keep a copy of every letter you send, but we are so used to it, we don't even think about it.

Web Culture

When we send an email, we think at least a *little* bit about the impression we are making, because we are sending it to a human being. We may well say things we would not say face-to-face, and live to regret that. Because we can't see anyone's eyes or hear anyone's voice, we are more likely to overreact and be hurtful, angry, or just too smart for our own good. But because email is directed, we don't send email thinking that no one else will ever read what we say.

The Web is different. Its social sites inherit their communication culture not from the letter or telephone call, but from the wall in the public square, littered with broadsides and scribbled notes, some of them signed and some not. Type a comment on a blog, or post a photo on a photo album, and your action can be as anonymous as you wish it to be—you do not know to whom your message is going. YouTube has millions of personal videos. Photo-archiving sites are the shoeboxes and photo albums of the twenty-first century. Online backup now provides easy access to permanent storage for the contents of our personal computers. We entrust commercial entities with much of our most private information, without apparent concern. The generation that has grown up with the Web has embraced social networking in all its varied forms: MySpace, YouTube, LiveJournal, Facebook, Xanga, Classmates.com, Flickr, dozens more, and blogs of every shape and size. More than being taken, personal privacy has been given away quite freely, because everyone else is doing it—the surrender of privacy is more than a way to social connectedness, it is a social institution in its own right. There are 70 million bloggers sharing everything from mindless blather to intimate personal details. Sites like `www.loopt.com` let you find your friends, while `twitter.com` lets you tell the entire world where you are and what you are doing. The Web is a confused, disorganized, chaotic realm, rich in both gold and garbage.

The "old" web, "Web 1.0," as we now refer to it, was just an information resource. You asked to see something, and you got to see it. Part of the disinhibition that happens on the new "Web 2.0" social networking sites is due to the fact that they still allow the movie-screen illusion—that we are "just looking," or if we are contributing, we are not leaving footprints or fingerprints if we use pseudonyms. (See Chapter 4 for more on Web 1.0 and Web 2.0.)

But of course, that is not really the way the Web ever worked. It is important to remember that even Web 1.0 was never anonymous, and even "just looking" leaves fingerprints.

In July 2006, a *New York Times* reporter called Thelma Arnold of Lilburn, Georgia. Thelma wasn't expecting the call. She wasn't famous, nor was she involved in anything particularly noteworthy. She enjoyed her hobbies, helped her friends, and from time to time looked up things on the Web—stuff about her dogs, and her friends' ailments.

Then AOL, the search engine she used, decided to release some "anonymous" query data. Thelma, like most Internet users, may not have known that AOL had kept every single topic that she, and every other one of their users, had asked about. But it did. In a moment of unenlightened generosity, AOL released for research use a small sample: about 20 million queries from 658,000 different users. That is actually not a lot of data by today's standards. For example, in July 2007, there were about 5.6 billion search engine queries, of which roughly 340 million were AOL queries. So, 20 million queries comprise only a couple of days' worth of search queries. In an effort to protect their clients' privacy, AOL "de-identified" the queries. AOL never mentioned anyone by name; they used random numbers instead. Thelma was 4417149. AOL mistakenly presumed that removing a single piece of personal identification would make it hard to figure out who the users were. It turned out that for some of the users, it wasn't hard at all.

It didn't take much effort to match Thelma with her queries. She had searched for "landscapers in Lilburn, GA" and several people with the last name "Arnold," leading to the obvious question of whether there were any Arnolds in Lilburn. Many of Thelma's queries were not particularly useful for identifying her, but were revealing nonetheless: "dry mouth," "thyroid," "dogs that urinate on everything," and "swing sets."

Thelma was not the only person to be identified. User 22690686 (Terri) likes astrology, and the Edison National Bank, Primerica, and Budweiser. 5779844 (Lawanna) was interested in credit reports, and schools. From what he searched for, user 356693 seems to have been an aide to Chris Shays, Congressman from Connecticut.

One of the privacy challenges that we confront as we rummage through the rubble of the digital explosion is that information exists without context. Was Thelma Arnold suffering from a wide range of ailments? One might readily conclude that from her searches. The fact is that she often tried to help her friends by understanding their medical problems.

Or consider AOL user 17556639, whose search history was released along with Thelma Arnold's. He searched for the following:

how to kill your wife 23 Mar, 22:09
wife killer 23 Mar, 22:11
poop 23 Mar, 22:12
dead people 23 Mar, 22:13
pictures of dead people 23 Mar, 22:15
killed people 23 Mar, 22:16
dead pictures 23 Mar, 22:17
murder photo 23 Mar, 22:20
steak and cheese 23 Mar, 22:22
photo of death 23 Mar, 22:30
death 23 Mar, 22:33
dead people photos 23 Mar, 22:33
photo of dead people 23 Mar, 22:35
www.murderdpeople.com 23 Mar, 22:37
decapitated photos 23 Mar, 22:39
car crashes3 23 Mar, 22:40
car crash photo 23 Mar, 22:41

Is this AOL user a potential criminal? Should AOL have called the police? Is 17556639 about to kill his wife? Is he (or she) a researcher with a spelling problem and an interest in Philly cheese steak? Is reporting him to the police doing a public service, or is it an invasion of privacy?

There is no way to tell just from these queries if this user was contemplating some heinous act or doing research for a novel that involves some grisly scenes. When information is incomplete and decontextualized, it is hard to judge meaning and intent.

In this particular case, we happen to know the answer. The user, Jason from New Jersey, was just fooling around, trying to see if Big Brother was watching. He wasn't planning to kill his wife at all. Inference from incomplete data has the problem of false positives—thinking you have something that you don't, because there are other patterns that fit the same data.

Information without context often leads to erroneous conclusions. Because our digital trails are so often retrieved outside the context within which they were created, they sometimes suggest incorrect interpretations. Data interpretation comes with balanced social responsibilities, to protect society when there is evidence of criminal behavior or intent, and also to protect the individual when such evidence is too limited to be reliable. Of course, for every example of misleading and ambiguous data, someone will want to solve the problems it creates by collecting more data, rather than less.

Beyond Privacy

There is nothing new under the sun, and the struggles to define and enforce privacy are no exception. Yet history shows that our concept of privacy has evolved, and the law has evolved with it. With the digital explosion, we have arrived at a moment where further evolution will have to take place rather quickly.

Leave Me Alone

More than a century ago, two lawyers raised the alarm about the impact technology and the media were having on personal privacy:

> Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the house-tops."

This statement is from the seminal law review article on privacy, published in 1890 by Boston attorney Samuel Warren and his law partner, Louis Brandeis, later to be a justice of the U.S. Supreme Court. Warren and Brandeis went on, "Gossip is no longer the resource of the idle and of the vicious, but has become a trade, which is pursued with industry as well as effrontery. To satisfy a prurient taste the details of sexual relations are spread broadcast in the columns of the daily papers. To occupy the indolent, column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle." New technologies made this garbage easy to produce, and then "the supply creates the demand."

And those candid photographs and gossip columns were not merely tasteless; they were bad. Sounding like modern critics of mindless reality TV, Warren and Brandeis raged that society was going to hell in a handbasket because of all that stuff that was being spread about.

> Even gossip apparently harmless, when widely and persistently circulated, is potent for evil. It both belittles and perverts. It belittles by inverting the relative importance of things, thus dwarfing the thoughts and aspirations of a people. When personal gossip attains the dignity of print, and crowds the space available for matters of

> real interest to the community, what wonder that the ignorant and thoughtless mistake its relative importance. Easy of comprehension, appealing to that weak side of human nature which is never wholly cast down by the misfortunes and frailties of our neighbors, no one can be surprised that it usurps the place of interest in brains capable of other things. Triviality destroys at once robustness of thought and delicacy of feeling. No enthusiasm can flourish, no generous impulse can survive under its blighting influence.

The problem they perceived was that it was hard to say just why such invasions of privacy should be unlawful. In individual cases, you could say something sensible, but the individual legal decisions were not part of a general regime. The courts had certainly applied legal sanctions for defamation—publishing malicious gossip that was false—but then what about malicious gossip that was true? Other courts had imposed penalties for publishing an individual's private letters—but on the basis of property law, just as though the individual's horse had been stolen rather than the words in his letters. That did not seem to be the right analogy either. No, they concluded, such rationales didn't get to the nub. When something private is published about you, something has been taken from you, you are a victim of theft—but the thing stolen from you is part of your identity as a person. In fact, privacy was a right, they said, a "general right of the individual to be let alone." That right had long been in the background of court decisions, but the new technologies had brought this matter to a head. In articulating this new right, Warren and Brandeis were, they asserted, grounding it in the principle of "inviolate personhood," the sanctity of individual identity.

Privacy and Freedom

The Warren-Brandeis articulation of privacy as a right to be left alone was influential, but it was never really satisfactory. Throughout the twentieth century, there were simply too many good reasons for *not* leaving people alone, and too many ways in which people *preferred* not to be left alone. And in the U.S., First Amendment rights stood in the way of privacy rights. As a general rule, the government simply cannot stop me from saying *anything*. In particular, it usually cannot stop me from saying what I want about your private affairs. Yet the Warren-Brandeis definition worked well enough for a long time, because, as Robert Fano put it, "The pace of technological progress was for a long time sufficiently slow as to enable society to learn pragmatically how to exploit new technology and prevent its abuse, with society maintaining its equilibrium most of the time." By the late 1950s, the emerging

electronic technologies, both computers and communication, had destroyed that balance. Society could no longer adjust pragmatically, because surveillance technologies were developing too quickly.

The result was a landmark study of privacy by the Association of the Bar of the City of New York, which culminated in the publication, in 1967, of a book by Alan Westin, entitled *Privacy and Freedom*. (Fano was reviewing Westin's book when he painted the picture of social disequilibrium caused by rapid technological change.) Westin proposed a crucial shift of focus.

Brandeis and Warren had seen a loss of privacy as a form of personal injury, which might be so severe as to cause "mental pain and distress, far greater than could be inflicted by mere bodily injury." Individuals had to take responsibility for protecting themselves. "Each man is responsible for his own acts and omissions only." But the law had to provide the weapons with which to resist invasions of privacy.

Westin recognized that the Brandeis-Warren formulation was too absolute, in the face of the speech rights of other individuals and society's legitimate data-gathering practices. Protection might come not from protective shields, but from control over the uses to which personal information could be put. "Privacy," wrote Westin, "is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."

> ... what is needed is a structured and rational weighing process, with definite criteria that public and private authorities can apply in comparing the claim for disclosure or surveillance through new devices with the claim to privacy. The following are suggested as the basic steps of such a process: measuring the seriousness of the need to conduct surveillance; deciding whether there are alternative methods to meet the need; deciding what degree of reliability will be required of the surveillance instrument; determining whether true consent to surveillance has been given; and measuring the capacity for limitation and control of the surveillance if it is allowed.

So even if there were a legitimate reason why the government, or some other party, might know something about you, your right to privacy might limit what the knowing party could do with that information.

This more nuanced understanding of privacy emerged from the important social roles that privacy plays. Privacy is not, as Warren and Brandeis had it, the right to be isolated from society—privacy is a right that makes society work. Fano mentioned three social roles of privacy. First, "the right to maintain the privacy of one's personality can be regarded as part of the right of

self-preservation"—the right to keep your adolescent misjudgments and personal conflicts to yourself, as long as they are of no lasting significance to your ultimate position in society. Second, privacy is the way society allows deviations from prevailing social norms, given that no one set of social norms is universally and permanently satisfactory—and indeed, given that social progress requires social experimentation. And third, privacy is essential to the development of independent thought—it enables some decoupling of the individual from society, so that thoughts can be shared in limited circles and rehearsed before public exposure.

Privacy is the way society allows deviations from prevailing social norms, given that social progress requires social experimentation.

Privacy and Freedom, and the rooms full of disk drives that sprouted in government and corporate buildings in the 1960s, set off a round of soul-searching about the operational significance of privacy rights. What, in practice, should those holding a big data bank think about when collecting the data, handling it, and giving it to others?

Fair Information Practice Principles

In 1973, the Department of Health, Education, and Welfare issued "Fair Information Practice Principles" (FIPP), as follows:

- **Openness.** There must be no personal data record-keeping systems whose very existence is secret.
- **Disclosure.** There must be a way for a person to find out what information about the person is in a record and how it is used.
- **Secondary use.** There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
- **Correction.** There must be a way for a person to correct or amend a record of identifiable information about the person.
- **Security.** Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for its intended use and must take precautions to prevent misuses of the data.

These principles were proposed for U.S. medical data, but were never adopted. Nevertheless, they have been the foundation for many corporate privacy policies. Variations on these principles have been codified in international trade agreements by the Organization of Economic Cooperation and Development (OECD) in 1980, and within the European Union (EU) in 1995. In the United States, echoes of these principles can be found in some state laws, but federal laws generally treat privacy on a case by case or "sectorial" basis. The 1974 Privacy Act applies to interagency data transfers within the federal government, but places no limitations on data handling in the private sector. The Fair Credit Reporting Act applies only to consumer credit data, but does not apply to medical data. The Video Privacy Act applies only to videotape rentals, but not to "On Demand" movie downloads, which did not exist when the Act was passed! Finally, few federal or state laws apply to the huge data banks in the file cabinets and computer systems of cities and towns. American government is decentralized, and authority over government data is decentralized as well.

The U.S. is not lacking in privacy laws. But privacy has been legislated inconsistently and confusingly, and in terms dependent on technological contingencies. There is no national consensus on what should be protected, and how protections should be enforced. Without a more deeply informed collective judgment on the benefits and costs of privacy, the current legislative hodgepodge may well get worse in the United States.

U.S. Privacy Laws

The Council of Better Business Bureaus has compiled a "Review of Federal and State Privacy Laws":

`www.bbbonline.org/UnderstandingPrivacy/library/fed_statePrivLaws.pdf`

The state of Texas has also compiled a succinct summary of major privacy laws:

`www.oag.state.tx.us/notice/privacy_table.htm`.

The discrepancy between American and European data privacy standards threatened U.S. involvement in international trade, because an EU directive would prohibit data transfers to nations, such as the U.S., that do not meet the European "adequacy" standard for privacy protection. Although the U.S. sectorial approach continues to fall short of European requirements, in 2000 the European Commission created a "safe harbor" for American businesses with multinational operations. This allowed individual corporations to establish their practices are adequate with respect to seven principles, covering notice, choice, onward transfer, access, security, data integrity, and enforcement.

It is, unfortunately, too easy to debate whether the European omnibus approach is more principled than the U.S. piecemeal approach, when the real question is whether either approach accomplishes what we want it to achieve. The Privacy Act of 1974 assured us that obscure statements would be buried deep in the Federal Register, providing the required official notice about massive governmental data collection plans—better than nothing, but providing "openness" only in a narrow and technical sense. Most large corporations doing business with the public have privacy notices, and virtually no one reads them. Only 0.3% of Yahoo! users read its privacy notice in 2002, for example. In the midst of massive negative publicity that year when Yahoo! changed its privacy policy to allow advertising messages, the number of users who accessed the privacy policy rose only to 1%. None of the many U.S. privacy laws prevented the warrantless wiretapping program instituted by the Bush administration, nor the cooperation with it by major U.S. telecommunications companies.

Indeed, cooperation between the federal government and private industry seems more essential than ever for gathering information about drug trafficking and international terrorism, because of yet another technological development. Twenty years ago, most long-distance telephone calls spent at least part of their time in the air, traveling by radio waves between microwave antenna towers or between the ground and a communication satellite. Government eavesdroppers could simply listen in (see the discussion of Echelon in Chapter 5). Now many phone calls travel through fiber optic cables instead, and the government is seeking the capacity to tap this privately owned infrastructure.

High privacy standards have a cost. They can limit the public usefulness of data. Public alarm about the release of personal medical information has led to major legislative remedies. The Health Information Portability and Accountability Act (HIPAA) was intended both to encourage the use of electronic data interchange for health information, and to impose severe penalties for the disclosure of "Protected Health Information," a very broad category including not just medical histories but, for example, medical payments. The bill mandates the removal of anything that could be used to re-connect medical records to their source. HIPAA is fraught with problems in an environment of ubiquitous data and powerful computing. Connecting the dots by assembling disparate data sources makes it extremely difficult to achieve the level of anonymity that HIPAA sought to guarantee. But help is available, for a price, from a whole new industry of HIPAA-compliance advisors. If you search for HIPAA online, you will likely see advertisements for services that will help you protect your data, and also keep you out of jail.

Ever Read Those "I Agree" Documents?

Companies can do almost anything they want with your information, as long as you agree. It seems hard to argue with that principle, but the deck can be stacked against the consumer who is "agreeing" to the company's terms. Sears Holding Corporation (SHC), the parent of Sears, Roebuck and Kmart, gave consumers an opportunity to join "My Sears Holding Community," which the company describes as "something new, something different ... a dynamic and highly interactive online community ... where your voice is heard and your opinion matters." When you went online to sign up, the terms appeared in a window on the screen.

The scroll box held only 10 lines of text, and the agreement was 54 boxfuls long. Deep in the terms was a detail: You were allowing Sears to install software on your PC that "monitors all of the Internet behavior that occurs on the computer ..., including ... filling a shopping basket, completing an application form, or checking your ... personal financial or health information." So your computer might send your credit history and AIDS test results to SHC, and you said it was fine!

At the same time as HIPAA and other privacy laws have safeguarded our personal information, they are making medical research costly and sometimes impossible to conduct. It is likely that classic studies such as the Framingham Heart Study, on which much public policy about heart disease was founded, could not be repeated in today's environment of strengthened privacy rules. Dr. Roberta Ness, president of the American College of Epidemiology, reported that "there is a perception that HIPAA may even be having a negative effect on public health surveillance practices."

The European reliance on the Fair Information Practice Principles is often no more useful, in practice, than the American approach. Travel through London, and you will see many signs saying "Warning: CCTV in use" to meet the "Openness" requirement about the surveillance cameras. That kind of notice throughout the city hardly empowers the individual. After all, even Big Brother satisfied the FIPP Openness standard, with the ubiquitous notices that he was watching! And the "Secondary Use" requirement, that European citizens should be asked permission before data collected for one purpose is used for another, is regularly ignored in some countries, although compliance practices are a major administrative burden on European businesses and may cause European businesses at least to pause and think before "repurposing" data they have gathered. Sociologist Amitai Etzioni repeatedly asks European

audiences if they have *ever* been asked for permission to re-use data collected about them, and has gotten only a single positive response—and that was from a gentleman who had been asked by a U.S. company.

The five FIPP principles, and the spirit of transparency and personal control that lay behind them, have doubtless led to better privacy practices. But they have been overwhelmed by the digital explosion, along with the insecurity of the world and all the social and cultural changes that have occurred in daily life. Fred H. Cate, a privacy scholar at the Indiana University, characterizes the FIPP principles as almost a complete bust:

> Modern privacy law is often expensive, bureaucratic, burdensome, and offers surprisingly little protection for privacy. It has substituted individual control of information, which it in fact rarely achieves, for privacy protection. In a world rapidly becoming more global through information technologies, multinational commerce, and rapid travel, data protection laws have grown more fractured and protectionist. Those laws have become unmoored from their principled basis, and the principles on which they are based have become so varied and procedural, that our continued intonation of the FIPPS mantra no longer obscures the fact that this emperor indeed has few if any clothes left.

Privacy as a Right to Control Information

It is time to admit that we don't even really know what we want. The bits are everywhere; there is simply no locking them down, and no one really wants to do that anymore. The meaning of privacy has changed, and we do not have a good way of describing it. It is not the right to be left alone, because not even the most extreme measures will disconnect our digital selves from the rest of the world. It is not the right to keep our private information to ourselves, because the billions of atomic factoids don't any more lend themselves into binary classification, private or public.

The bits are everywhere; there is simply no locking them down, and no one really wants to do that anymore.

Reade Seligmann would probably value his privacy more than most Americans alive today. On Monday, April 17, 2006, Seligmann was indicted in connection with allegations that a 27-year-old performer had been raped at a party at a Duke fraternity house. He and several of his lacrosse teammates instantly became poster children for everything that is wrong with

American society—an example of national over-exposure that would leave even Warren and Brandeis breathless if they were around to observe it. Seligmann denied the charges, and at first it looked like a typical he-said, she-said scenario, which could be judged only on credibility and presumptions about social stereotypes.

But during the evening of that fraternity party, Seligmann had left a trail of digital detritus. His data trail indicated that he could not have been at the party long enough, or at the right time, to have committed the alleged rape. Time-stamped photos from the party showed that the alleged victim of his rape was dancing at 12:02 AM. At 12:24 AM, he used his ATM card at a bank, and the bank's computers kept records of the event. Seligmann used his cell phone at 12:25 AM, and the phone company tracked every call he made, just as your phone company keeps a record of every call you make and receive. Seligmann used his prox card to get into his dormitory room at 12:46 AM, and the university's computer kept track of his comings and goings, just as other computers keep track of every card swipe or RFID wave you and I make in our daily lives. Even during the ordinary movements of a college student going to a fraternity party, every step along the way was captured in digital detail. If Seligmann had gone to the extraordinary lengths necessary to avoid leaving digital fingerprints—not using a modern camera, a cell phone, or a bank, and living off campus to avoid electronic locks—his defense would have lacked important exculpatory evidence.

Which would we prefer—the new world with digital fingerprints everywhere and the constant awareness that we are being tracked, or the old world with few digital footprints and a stronger sense of security from prying eyes? And what is the point of even asking the question, when the world cannot be restored to its old information lock-down?

In a world that has moved beyond the old notion of privacy as a wall around the individual, we could instead regulate those who would inappropriately *use* information about us. If I post a YouTube video of myself dancing in the nude, I should expect to suffer some personal consequences. Ultimately, as Warren and Brandeis said, individuals have to take responsibility for their actions. But society has drawn lines in the past around which facts are relevant to certain decisions, and which are not. Perhaps, the border of privacy having become so porous, the border of relevancy could be stronger. As Daniel Weitzner explains:

> New privacy laws should emphasize usage restrictions to guard against unfair discrimination based on personal information, even if it's publicly available. For instance, a prospective employer might be able to find a video of a job applicant entering an AIDS clinic or a

> mosque. Although the individual might have already made such facts public, new privacy protections would preclude the employer from making a hiring decision based on that information and attach real penalties for such abuse.

In the same vein, it is not intrinsically wrong that voting lists and political contributions are a matter of public record. Arguably, they are essential to the good functioning of the American democracy. Denying someone a promotion because of his or her political inclinations *would be* wrong, at least for most jobs. Perhaps a nuanced classification of the ways in which others are allowed to use information about us would relieve some of our legitimate fears about the effects of the digital explosion.

In *The Transparent Society*, David Brin wrote:

> Transparency is not about eliminating privacy. It's about giving us the power to hold accountable those who would *violate* it. Privacy implies serenity at home and the right to be let alone. It may be irksome how much other people know about me, but I have no right to police their minds. On the other hand I care very deeply about what others *do* to me and to those I love. We all have a right to some place where we can feel safe.

Despite the very best efforts, and the most sophisticated technologies, we cannot control the spread of our private information. And we often want information to be made public to serve our own, or society's purposes.

Yet there can still be principles of accountability for the *misuse* of information. Some ongoing research is outlining a possible new web technology, which would help ensure that information is used appropriately even if it is known. Perhaps automated classification and reasoning tools, developed to help connect the dots in networked information systems, can be retargeted to limit inappropriate use of networked information. A continuing border war is likely to be waged, however, along an existing free speech front: the line separating my right to tell the truth about you from your right not to have that information used against you. In the realm of privacy, the digital explosion has left matters deeply unsettled.

Always On

In *1984,* the pervasive, intrusive technology could be turned off:

> As O'Brien passed the telescreen a thought seemed to strike him. He stopped, turned aside and pressed a switch on the wall. There was a sharp snap. The voice had stopped.
>
> Julia uttered a tiny sound, a sort of squeak of surprise. Even in the midst of his panic, Winston was too much taken aback to be able to hold his tongue.
>
> "You can turn it off!" he said.
>
> "Yes," said O'Brien, "we can turn it off. We have that privilege. ...Yes, everything is turned off. We are alone."

Sometimes we can still turn it off today, and should. But mostly we don't want to. We don't want to be alone; we want to be connected. We find it convenient to leave it on, to leave our footprints and fingerprints everywhere, so we will be recognized when we come back. We don't want to have to keep retyping our name and address when we return to a web site. We like it when the restaurant remembers our name, perhaps because our phone number showed up on caller ID and was linked to our record in their database. We appreciate buying grapes for $1.95/lb instead of $3.49, just by letting the store know that we bought them. We may want to leave it on for ourselves because we know it is on for criminals. Being watched reminds us that they are watched as well. Being watched also means we are being watched over.

And perhaps we don't care that so much is known about us because that is the way human society used to be—kinship groups and small settlements, where knowing everything about everyone else was a matter of survival. Having it on all the time may resonate with inborn preferences we acquired millennia ago, before urban life made anonymity possible. Still today, privacy means something very different in a small rural town than it does on the Upper East Side of Manhattan.

We cannot know what the cost will be of having it on all the time. Just as troubling as the threat of authoritarian measures to restrict personal liberty is the threat of voluntary conformity. As Fano astutely observed, privacy allows limited social experimentation—the deviations from social norms that are much riskier to the individual in the glare of public exposure, but which can be, and often have been in the past, the leading edges of progressive social changes. With it always on, we may prefer not to try anything unconventional, and stagnate socially by collective inaction.

For the most part, it is too late, realistically, ever to turn it off. We may once have had the privilege of turning it off, but we have that privilege no more. We have to solve our privacy problems another way.

The digital explosion is shattering old assumptions about who knows what. Bits move quickly, cheaply, and in multiple perfect copies. Information that used to be public in principle—for example, records in a courthouse, the price you paid for your house, or stories in a small-town newspaper—is now available to everyone in the world. Information that used to be private and available to almost no one—medical records and personal snapshots, for example—can become equally widespread through carelessness or malice. The norms and business practices and laws of society have not caught up to the change.

The oldest durable communication medium is the written document. Paper documents have largely given way to electronic analogs, from which paper copies are produced. But are electronic documents really like paper documents? Yes and no, and misunderstanding the document metaphor can be costly. That is the story to which we now turn.

CHAPTER 3

Ghosts in the Machine

Secrets and Surprises of Electronic Documents

What You See Is Not What the Computer Knows

On March 4, 2005, Italian journalist Giuliana Sgrena was released from captivity in Baghdad, where she had been held hostage for a month. As the car conveying her to safety approached a checkpoint, it was struck with gunfire from American soldiers. The shots wounded Sgrena and her driver and killed an Italian intelligence agent, Nicola Calipari, who had helped engineer her release.

A fierce dispute ensued about why U.S soldiers had rained gunfire on a car carrying citizens of one of its Iraq war allies. The Americans claimed that the car was speeding and did not slow when warned. The Italians denied both claims. The issue caused diplomatic tension between the U.S. and Italy and was a significant political problem for the Italian prime minister.

The U.S. produced a 42-page report on the incident, exonerating the U.S. soldiers. The report enraged Italian officials. The Italians quickly released their own report, which differed from the U.S. report in crucial details.

Because the U.S. report included sensitive military information, it was heavily redacted before being shared outside military circles (see Figure 3.1). In another time, passages would have been blacked out with a felt marker, and the document would have been photocopied and given to reporters. But in the information age, the document was redacted and distributed electronically, not physically. The redacted report was posted on a web site the allies used to provide war information to the media. In an instant, it was visible to any of the world's hundreds of millions of Internet users.

(U) [redacted] has Direct Liaison Authorized (DIRLAUTH) to coordinate directly with [redacted] for security along Route Irish. This is the same level of coordination previously authorized by [redacted] Division to [redacted]. When executing DIRLAUTH, [redacted] directly coordinates an action with units internal or external to its command and keeps the [redacted] commander informed. The [redacted] TOC passes all coordination efforts through the [redacted] Brigade TOC to [redacted] JOC. (Annex 58C).

Source: http://www.corriere.it/Media/Documenti/Classified.pdf, extract from page 10.

FIGURE 3.1 Section from page 10 of redacted U.S. report on the death of Italian journalist Nicola Calipari. Information that might have been useful to the enemy was blacked out.

One of those Internet users was an Italian blogger, who scrutinized the U.S. report and quickly recovered the redacted text using ordinary office software. The blogger posted the full text of the report (see Figure 3.2) on his own web site. The unredacted text disclosed positions of troops and equipment, rules of engagement, procedures followed by allied troops, and other information of interest to the enemy. The revelations were both dangerous to U.S. soldiers and acutely embarrassing to the U.S. government, at a moment when tempers were high among Italian and U.S. officials. In the middle of the most high-tech war in history, how could this fiasco have happened?

(U) 1-76 FA has Direct Liaison Authorized (DIRLAUTH) to coordinate directly with 1-69 IN for security along Route Irish. This is the same level of coordination previously authorized by 1st Cavalry Division to 2-82 FA. When executing DIRLAUTH, 1-76 FA directly coordinates an action with units internal or external to its command and keeps the 31D commander informed. The 1-76 FA TOC passes all coordination efforts through the 4th Brigade TOC to 31D JOC. (Annex 58C).

Source: http://www.corriere.it/Media/Documenti/Unclassified.doc.

FIGURE 3.2 The text of Figure 3.1 with the redaction bars electronically removed.

Paper documents and electronic documents are useful in many of the same ways. Both can be inspected, copied, and stored. But they are not equally useful for all purposes. Electronic documents are easier to change, but paper documents are easier to read in the bathtub. In fact, the metaphor of a series of bits as a "document" can be taken only so far. When stretched beyond its breaking point, the "document" metaphor can produce surprising and damaging results—as happened with the Calipari report.

Office workers love "WYSIWYG" interfaces—"What You See Is What You Get." They edit the electronic document on the screen, and when they print it, it looks just the same. They are deceived into thinking that what is in the

computer is a sort of miniaturized duplicate of the image on the screen, instead of computer codes that produce the picture on the screen. In fact, the WYSIWYG metaphor is imperfect, and therefore risky. The report on the death of Nicola Calipari illustrates what can go wrong when users accept such a metaphor too literally. What the authors of the document saw was dramatically different from what they got.

The report had been prepared using software that creates PDF files. Such software often includes a "Highlighter Tool," meant to mimic the felt markers that leave a pale mark on ordinary paper, through which the underlying text is visible (see Figure 3.3). The software interface shows the tool's icon as a marker writing a yellow stripe, but the user can change the color of the stripe. Probably someone tried to turn the Highlighter Tool into a redaction tool by changing its color to black, unaware that what was visible on the screen was not the same as the contents of the electronic document.

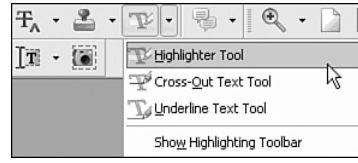
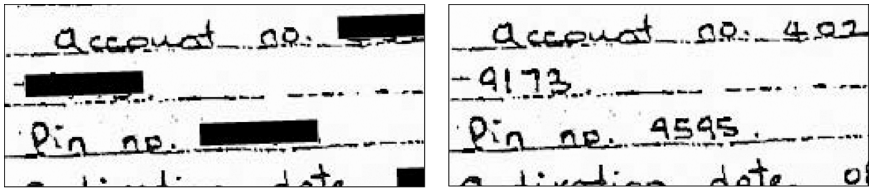


FIGURE 3.3 Adobe Acrobat Highlighter Tool, just above the middle. On the screen, the "highlighter" is writing yellow ink, but with a menu command, it can be changed to any other color.

The Italian blogger guessed that the black bars were nothing more than overlays created using the Highlighter Tool, and that the ghostly traces of the invisible words were still part of the electronic document that was posted on the web. With that realization, he easily undid the black "highlighting" to reveal the text beneath.

Just as disturbing as this mistake is the fact that two major newspapers had quite publicly made the same mistake only a few years before. On April 16, 2000, the *New York Times* had detailed a secret CIA history of attempts by the U.S. to overthrow Iran's government in 1953. The newspaper reproduced sections of the CIA report, with black redaction bars to obscure the names of CIA operatives within Iran. The article was posted on the Web in mid-June, 2000, accompanied by PDFs of several pages of the CIA report. John Young, who administers a web site devoted to publishing government-restricted documents, removed the redaction bars and revealed the names of CIA agents. A controversy ensued about the ethics and legality of the disclosure, but the names are still available on the Web as of this writing.

The *Washington Post* made exactly the same mistake in 2002, when it published an article about a demand letter left by the Washington snipers, John Allen Muhammad and John Lee Malvo. As posted on the *Post*'s web site, certain information was redacted in a way that was easily reversed by an inquisitive reader of the online edition of the paper (see Figure 3.4). The paper fixed the problem quickly after its discovery, but not quickly enough to prevent copies from being saved.



Source: Washington Post web site, transferred to web.bham.ac.uk/forensic/news/02/sniper2.html. Actual images taken from slide 29 of http://www.ccc.de/congress/2004/fahrplan/files/316-hidden-data-slides.pdf.

FIGURE 3.4 Letter from the Washington snipers. On the left, the redacted letter as posted on the *Washington Post* web site. On the right, the letter with the redaction bars electronically removed.

What might have been done in these cases, instead of posting the PDF with the redacted text hidden but discoverable? The Adobe Acrobat software has a security feature, which uses encryption (discussed in Chapter 5, "Secret Bits") to make it impossible for documents to be altered by unauthorized persons, while still enabling anyone to view them. Probably those who created these documents did not know about this feature, or about commercially available software called Redax, which government agencies use to redact text from documents created by Adobe Acrobat.

A clumsier, but effective, option would be to scan the printed page, complete with its redaction bars. The resulting file would record only a series of black and white dots, losing all the underlying typographical structure—font names and margins, for example. Whatever letters had once been "hidden" under the redaction bars could certainly not be recovered, yet this solution has an important disadvantage.

One of the merits of formatted text documents such as PDFs is that they can be "read" by a computer. They can be searched, and the text they contain can be copied. With the document reduced to a mass of black and white dots, it could no longer be manipulated as text.

A more important capability would be lost as well. The report would be unusable by programs that vocalize documents for visually impaired readers. A blind reader could "read" the U.S. report on the Calipari incident, because software is available that "speaks" the contents of PDF documents. A blind reader would find a scanned version of the same document useless.

Tracking Changes—and Forgetting That They Are Remembered

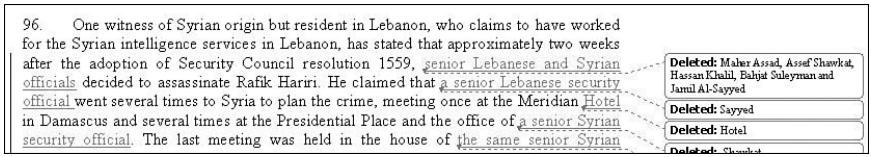
In October, 2005, UN prosecutor Detlev Mehlis released to the media a report on the assassination of former Lebanese Prime Minister Rafik Hariri. Syria had been suspected of engineering the killing, but Syrian President Bashar al-Assad denied any involvement. The report was not final, Mehlis said, but there was "evidence of both Lebanese and Syrian involvement." Deleted, and yet uncovered by the reporters who were given the document, was an incendiary claim: that Assad's brother Maher, commander of the Republican Guard, was personally involved in the assassination.

Microsoft Word offers a "Track Changes" option. If enabled, every change made to the document is logged as part of the document itself—but ordinarily not shown. The document bears its entire creation history: who made each change, when, and what it was. Those editing the document can also add comments—which would not appear in the final document, but may help editors explain their thinking to their colleagues as the document moves around electronically within an office.

Of course, information about strategic planning is not meant for outsiders to see, and in the case of legal documents, can have catastrophic consequences if revealed. It is a simple matter to remove these notes about the document's history—but someone has to remember to do it! The UN prosecutor neglected to remove the change history from his Microsoft Word document, and a reporter discovered the deleted text (see Figure 3.5). (Of course, in Middle Eastern affairs, one cannot be too suspicious. Some thought that Mehlis had intentionally left the text in the document, as a warning to the Syrians that he knew more than he was yet prepared to acknowledge.)

A particularly negligent example of document editing involved SCO Corporation, which claimed that several corporations violated its intellectual property rights. In early 2004, SCO filed suit in a Michigan court against Daimler Chrysler, claiming Daimler had violated terms of its Unix software agreement with SCO. But the electronic version of its complaint carried its modification history with it, revealing a great deal of information about SCO's litigation planning. In particular, when the change history was revealed, it

turned out that until exactly 11:10 a.m. on February 18, 2004, SCO had instead planned to sue a different company, Bank of America, in federal rather than state court, for copyright infringement rather than breach of contract



Source: Section of UN report, posted on Washington Post web site, `www.washingtonpost.com/wp-srv/world/syria/mehlis.report.doc`.

FIGURE 3.5 Section from the UN report on the assassination of Rafik Hariri. An earlier draft stated that Maher Assad and others were suspected of involvement in the killing, but in the document as it was released, their names were replaced with the phrase "senior Lebanese and Syrian officials."

Saved Information About a Document

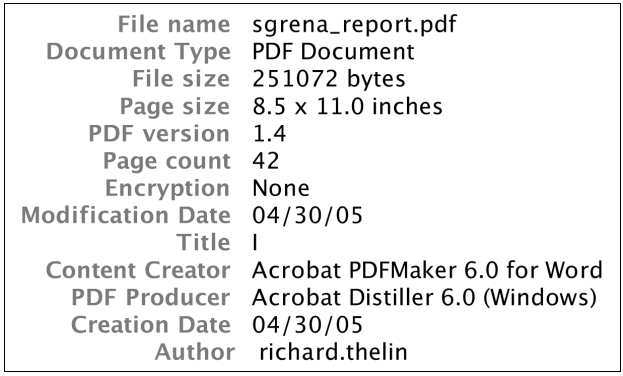
FORGING METADATA

Metadata can help prove or refute claims. Suppose Sam emails his teacher a homework paper after the due date, with a plea that the work had been completed by the deadline, but was undeliverable due to a network failure. If Sam is a cheater, he could be exposed if he doesn't realize that the "last modified" date is part of the document. However, if Sam *is* aware of this, he could "stamp" the document with the right time by re-setting the computer's clock before saving the file. The name in which the computer is registered and other metadata are also forgeable, and therefore are of limited use as evidence in court cases.

An electronic document (for example, one produced by text-processing software) often includes information that is *about* the document—so-called *metadata*. The most obvious example is the name of the file itself. File names carry few risks. For example, when we send someone a file as an email attachment, we realize that the recipient is going to see the name of the file as well as its contents.

But the file is often tagged with much more information than just its name. The metadata generally includes the name associated with the owner of the computer, and the dates the file was created and last modified—often useful information, since the recipient can tell whether she is receiving an older or newer version than the version she already

has. Some word processors include version information as well, a record of who changed what, when, and why. But the unaware can be trapped even by such innocent information, since it tends not to be visible unless the recipient asks to see it. In Figure 3.6, the metadata reveals the name of the military officer who created the redacted report on the death of Nicola Calipari.



File name	sgrena_report.pdf
Document Type	PDF Document
File size	251072 bytes
Page size	8.5 x 11.0 inches
PDF version	1.4
Page count	42
Encryption	None
Modification Date	04/30/05
Title	I
Content Creator	Acrobat PDFMaker 6.0 for Word
PDF Producer	Acrobat Distiller 6.0 (Windows)
Creation Date	04/30/05
Author	richard.thelin

FIGURE 3.6 Part of the metadata of the Calipari report, as revealed by the "Properties" command of Adobe Acrobat Reader. The data shows that Richard Thelin was the author, and that he altered the file less than two minutes after creating it. Thelin was a Lieutenant Colonel in the U.S. Marine Corps at the time of the incident.

Authorship information leaked in this way can have real consequences. In 2003, the British government of Tony Blair released documentation of its case for joining the U.S. war effort in Iraq. The document had many problems—large parts of it turned out to have been plagiarized from a 13-year-old PhD thesis. Equally embarrassing was that the electronic fingerprints of four civil servants who created it were left on the document when it was released electronically on the No. 10 Downing Street web site. According to the *Evening Standard of London*, "All worked in propaganda units controlled by Alastair Campbell, Tony Blair's director of strategy and communications," although the report had supposedly been the work of the Foreign Office. The case of the "dodgy dossier" caused an uproar in Parliament.

You don't have to be a businessperson or government official to be victimized by documents bearing fingerprints. When you send someone a document as an attachment to an email, very likely the document's metadata shows who actually created it, and when. If you received it from someone else

and then altered it, that may show as well. If you put the text of the document into the body of your email instead, the metadata won't be included; the message will be just the text you see on the screen. Be sure of what you are sending before you send it!

Can the Leaks Be Stopped?

Even in the most professional organizations, and certainly in ordinary households, knowledge about technological dangers and risks does not spread instantaneously to everyone who should know it. The Calipari report was published five years after the *New York Times* had been embarrassed. How can users of modern information technology—today, almost all literate people—stay abreast of knowledge about when and how to protect their information?

It is not easy to prevent the leakage of sensitive information that is hidden in documents but forgotten by their creators, or that is captured as metadata. In principle, offices should have a check-out protocol so that documents are cleansed before release. But in a networked world, where email is a critical utility, how can offices enforce document release protocols without rendering simple tasks cumbersome? A rather harsh measure is to prohibit use of software that retains such information; that was the solution adopted by the British government in the aftermath of the "dodgy dossier" scandal. But the useful features of the software are then lost at the same time. A protocol can be established for converting "rich" document formats such as that of Microsoft Word to formats that retain less information, such as Adobe PDF. But it turns out that measures used to eradicate personally identifiable information from documents don't achieve as thorough a cleansing as is commonly assumed.

At a minimum, office workers need education. Their software has great capabilities they may find useful, but many of those useful features have risks as well. And we all just need to think about what we are doing with our documents. We all too mindlessly re-type keystrokes we have typed a hundred times in the past, not pausing to think that the hundred and first situation may be different in some critical way!

Representation, Reality, and Illusion

René Magritte, in his famous painting of a pipe, said "This isn't a pipe" (see Figure 3.7). Of course it isn't; it's a painting of a pipe. The image is made out

of paint, and Magritte was making a metaphysical joke. The painting is entitled "The treachery of images," and the statement that the image isn't the reality is part of the image itself.

Los Angeles County Museum of Art. Purchased with funds provided by the Mr. and Mrs. William Preston Harrison Collection. Photograph © 2007 Museum Associates/LACMA.

FIGURE 3.7 Painting by Magritte. The legend says "This isn't a pipe." Indeed, it's only smudges of paint that make you think of a pipe, just as an electronic document is only bits representing a document.

When you take a photograph, you capture inside the camera something from which an image can be produced. In a digital camera, the bits in an electronic memory are altered according to some pattern. The image, we say, is "represented" in the camera's memory. But if you took out the memory and looked at it, you couldn't see the image. Even if you printed the pattern of 0s and 1s stored in the memory, the image wouldn't appear. You'd have to know *how* the bits represent the image in order to get at the image itself. In the world of digital photography, the format of the bits has been standardized, so that photographs taken on a variety of cameras can be displayed on a variety of computers and printed on a variety of printers.

The general process of digital photography is shown in Figure 3.8. Some external reality—a scene viewed through a camera lens, for example—is turned into a string of bits. The bits somehow capture useful information

about reality, but there is nothing "natural" about the way reality is captured. The representation is a sort of ghost of the original, not identical to the original and actually quite unlike it, but containing enough of the soul of the original to be useful later on. The representation follows rules. The rules are arbitrary conventions and the product of human invention, but they have been widely accepted so photographs can be exchanged.

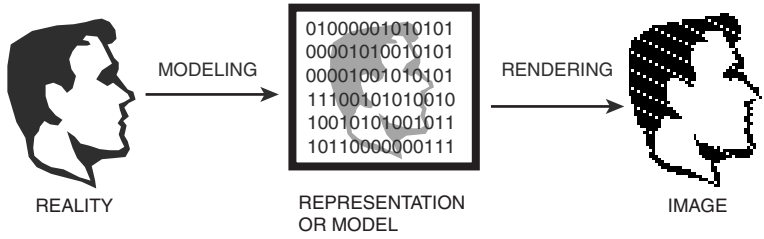


FIGURE 3.8 Reproducing an image electronically is a two-stage process. First, the scene is translated into bits, creating a digital model. Then the model is rendered as a visible image. The model can be stored indefinitely, communicated from one place to another, or computationally analyzed and enhanced to produce a different model before it is rendered. The same basic structure applies to the reproduction of video and audio.

The representation of the photograph in bits is called a *model* and the process of capturing it is called *modeling*. The model is turned into an image by *rendering* the model; this is what happens when you transfer the bits representing a digital photograph to a computer screen or printer. Rendering brings the ghost back to life. The image resembles, to the human eye, the original reality—provided that the model is good enough. Typically, a model that is not good enough—has too few bits, for example—cannot produce an image that convincingly resembles the reality it was meant to capture.

Modeling always omits information. Magritte's painting doesn't smell like a pipe; it has a different patina than a pipe; and you can't turn it around to see what the other side of the pipe looks like. Whether the omitted information is irrelevant or essential can't be judged without knowing how the model is going to be used. Whoever creates the model and renders it has the power to shape the experience of the viewer.

The process of modeling followed by rendering applies to many situations other than digital photography. For example, the same transformations happen when music is captured on a CD or as an MP3. The rendering process produces audible music from a digital representation, via stereo speakers or a

headset. CDs and MP3s use quite distinct modeling methods, with CDs generally capturing music more accurately, using a larger number of bits.

Knowing that digital representations don't resemble the things they represent explains the difference between the terms "analog" and "digital." An analog telephone uses a continuously varying electric signal to represent a continuously varying sound—the voltage of the telephone signal is an "analog" of the sound it resembles—in the same way that Magritte applied paint smoothly to canvas to mimic the shape of the pipe. The shift from analog to digital technologies, in telephones, televisions, cameras, X-ray machines, and many other devices, at first seems to lose the immediacy and simplicity of the old devices. But the enormous processing power of modern computers makes the digital representation far more flexible and useful.

Indeed, the same general processes are at work in situations where *there is no "reality" because the images are of things that have never existed.* Examples are video games, animated films, and virtual walk-throughs of unbuilt architecture. In these cases, the first step of Figure 3.8 is truncated. The "model" is created not by capturing reality in an approximate way, but by pure synthesis: as the strokes of an artist's electronic pen, or the output of computer-aided design software.

The severing of the immediate connection between representation and reality in the digital world has created opportunities, dangers, and puzzles. One of the earliest triumphs of "digital signal processing," the science of doing computations on the digital representations of reality, was to remove the scratches and noise from old recordings of the great singer Enrico Caruso. No amount of analog electronics could have cleaned up the old records and restored the clarity to Caruso's voice.

And yet the growth of digital "editing" has its dark side as well. Photo-editing software such as Photoshop can be used to alter photographic evidence presented to courts of law.

Can We Be Sure a Photo Is Unretouched?

Cryptographic methods (discussed in Chapter 5) can establish that a digital photograph has not been altered. A special camera gets a digital key from the "image verification system," attaches a "digital signature" (see Chapter 5) to the image and uploads the image and the signature to the verification system. The system processes the received image with the same key and verifies that the same signature results. The system is secure because it is impossible, with any reasonable amount of computation, to produce another image that would yield the same signature with this key.

The movie *Toy Story* and its descendants are unlikely to put human actors out of work in the near future, but how should society think about synthetic child pornography? "Kiddie porn" is absolutely illegal, unlike other forms of pornography, because of the harm done to the children who are abused to produce it. But what about pornographic images of children who do not exist and never have—who are simply the creation of a skilled graphic synthesizer? Congress outlawed such virtual kiddie porn in 1996, in a law that prohibited any image that "is, or appears to be, of a minor engaging in sexually explicit conduct." The Supreme Court overturned the law on First Amendment grounds. Prohibiting images that "appear to" depict children is going too far, the court ruled—such synthetic pictures, no matter how abhorrent, are constitutionally protected free speech.

In the world of exploded assumptions about reality and artifice, laws that combat society's problems may also compromise rights of free expression.

In this instance at least, reality matters, not what images appear to show. Chapter 7, "You Can't Say That on the Internet," discusses other cases in which society is struggling to control social evils that are facilitated by information technology. In the world of exploded assumptions about reality and artifice, laws that combat society's problems may also compromise rights of free expression.

What Is the Right Representation?

DIGITAL CAMERAS AND MEGAPIXELS

Megapixels—millions of pixels—are a standard figure of merit for digital cameras. If a camera captures too few pixels, it can't take good photographs. But no one should think that more pixels invariably yield a better image. If a digital camera has a low-quality lens, more pixels will simply produce a more precise representation of a blurry picture!

Figure 3.9 is a page from the Book of Kells, one of the masterpieces of medieval manuscript illumination, produced around A.D. 800 in an Irish monastery. The page contains a few words of Latin, portrayed in an astoundingly complex interwoven lacework of human and animal figures, whorls, and crosshatching. The book is hundreds of pages long, and in the entire work no two of the letters or decorative ornaments are drawn the same way. The elaborately ornate graphic shows just 21 letters (see Figure 3.10).

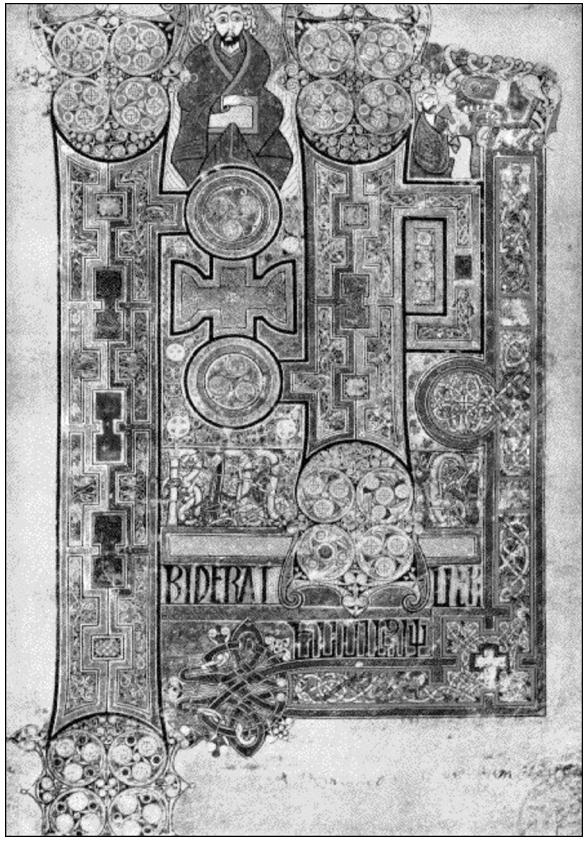


FIGURE 3.9 Opening page of the Gospel of St. John from the Book of Kells.

IN PRINCIPIO ERAT VERBUM

FIGURE 3.10 The words of the beginning of the gospel of St. John. In the book of Kells, the easiest word to spot is ERAT, just to the left of center about a quarter of the way up the page.

Do these two illustrations contain the same information? The answer depends on what information is meant to be recorded. If the only important thing were the Latin prose, then either representation might be equally good, though Figure 3.10 is easier to read. But the words themselves are far from

the only important thing in the Book of Kells. It is one of the great works of Western art and craftsmanship.

A graphic image such as Figure 3.9 is represented as a rectangular grid of many rows and columns, by recording the color at each position in the grid (see Figure 3.11). To produce such a representation, the page itself is scanned, one narrow row after the next, and each row is divided horizontally into tiny square "picture elements" or *pixels*. An image representation based on a division into pixels is called a *raster* or *bitmap representation*. The representation corresponds to the structure of a computer screen (or a digital TV screen), which is also divided into a grid of individual pixels—how many pixels, and how small they are, affect the quality and price of the display.

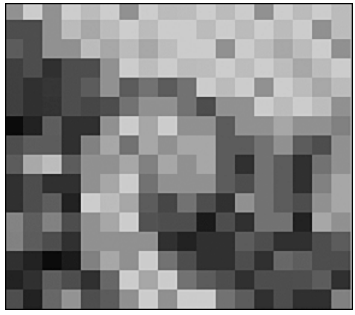


FIGURE 3.11 A detail enlarged from the upper-right corner of the opening page of John from the Book of Kells.

What would be the computer representation of the mere Latin text, Figure 3.10? The standard code for the Roman alphabet, called ASCII for the American Standard Code for Information Interchange, assigns a different 8-bit code to each letter or symbol. ASCII uses one byte (8 bits) per character. For example, A = `01000001`, a = `01100001`, $ = `00100100`, and 7 = `00110111`.

The equation 7 = `00110111` means that the bit pattern used to represent the symbol "7" in a string of text is `00110111`. The space character has its own code, `00100000`. Figure 3.12 shows the ASCII representation of the characters "IN PRINCIPIO ERAT VERBUM," a string of 24 bytes or 192 bits. We've separated the long string of bits into bytes to improve readability ever so slightly! But inside the computer, it would just be one bit after the next.

```
01001001 01001110 00100000 01010000
01010010 01001001 01001110 01000011
01001001 01010000 01001001 01001111
00100000 01000101 01010010 01000001
01010100 00100000 01010110 01000101
01010010 01000010 01010101 01001101
```

FIGURE 3.12 ASCII bit string for the characters of "IN PRINCIPIO ERAT VERBUM."

So `01001001` represents the letter I. But not always! Bit strings are used to represent many things other than characters. For example, the same bit string `01001001`, if interpreted as the representation of a whole number in binary notation, represents 73. A computer cannot simply look at a bit string `01001001` and know whether it is supposed to represent the letter I or the number 73 or data of some other type, a color perhaps. A computer can interpret a bit string only if it knows the conventions that were used to create the document—the intended interpretation of the bits that make up the file.

The meaning of a bit string is a matter of convention. Such conventions are arbitrary at first. The code for the letter I could have been `11000101` or pretty much anything else. Once conventions have become accepted through a social process of agreement and economic incentive, they became nearly as inflexible as if they were physical laws. Today, millions of computers assume

FILENAME EXTENSIONS

The three letters after the dot at the end of a filename indicate how the contents are to be interpreted. Some examples are as follows:

Extension	File Type
.doc	Microsoft Word document
.odt	OpenDocument text document
.ppt	Microsoft PowerPoint document
.ods	OpenDocument Spreadsheet
.pdf	Adobe Portable Document Format
.exe	Executable program
.gif	Graphics Interchange Format (uses 256-color palette)
.jpg	JPEG graphic file (Joint Photographic Experts Group)
.mpg	MPEG movie file (Moving Picture Experts Group)

that 01001001, if interpreted as a character, represents the letter I, and the universal acceptance of such conventions is what makes worldwide information flows possible.

The document format is the key to turning the representation into a viewable document. If a program misinterprets a document as being in a different format from the one in which it was created, only nonsense will be rendered. Computers not equipped with software matching the program that created a document generally refuse to open it.

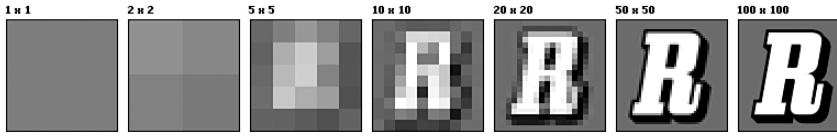
Which representation is "better," a raster image or ASCII? The answer depends on the use to which the document is to be put. For representation of freeform shapes in a great variety of shades and hues, a raster representation is unbeatable, provided the pixels are small enough and there are enough of them. But it is hard even for a trained human to find the individual letters within Figure 3.9, and it would be virtually impossible for a computer program. On the other hand, a document format based on ASCII codes for characters, such as the PDF format, can easily be searched for text strings.

The PDF format includes more than simply the ASCII codes for the text. PDF files include information about typefaces, the colors of the text and of the background, and the size and exact positions of the letters. Software that produces PDFs is used to typeset elegant documents such as this one. In other words, PDF is actually a *page description language* and describes visible features that are typographically meaningful. But for complicated pictures, a graphical format such as JPG must be used. A mixed document, such as these pages, includes graphics within PDF files.

Reducing Data, Sometimes Without Losing Information

Let's take another look at the page from the Book of Kells, Figure 3.9, and the enlargement of a small detail of that image, Figure 3.11. The computer file from which Figure 3.9 was printed is 463 pixels wide and 651 pixels tall, for a total of about 300,000 individual pixels. The pages of the Book of Kells measure about 10 by 13 inches, so the raster image has only about 50 pixels per inch of the original work. That is too few to capture the rich detail of the original—Figure 3.11 actually shows one of the animal heads in the top-right corner of the page. A great deal of detail was lost when the original page was scanned and turned into pixels. The technical term for the problem is *undersampling*. The scanning device "samples" the color value of the original document at discrete points to create the representation of the document, and in this case, the samples are too far apart to preserve detail that is visible to the naked eye in the original.

The answer to undersampling is to increase the resolution of the scan—the number of samples per inch. Figure 3.13 shows how the quality of an image improves with the resolution. In each image, each pixel is colored with the "average" color of part of the original.



Credit as in Wikipedia, en.wikipedia.org/wiki/Image:Resolution_illustration.png.

FIGURE 3.13 A shape shown at various resolutions, from 1×1 to 100×100 pixels. A square block consisting of many pixels of a single shade can be represented much more compactly than by repeating the code for that shade as many times as there are pixels.

But, of course, a price is paid for increased resolution. The more pixels in the representation of an image, the more memory is needed to hold the representation. Double the resolution, and the memory needed goes up by a factor of four, since the resolution doubles both vertically and horizontally.

Standard software uses a variety of representational techniques to represent raster graphics more concisely. Compression techniques are of two kinds: "lossless" and "lossy." A *lossless* representation is one that allows exactly the same image to be rendered. A *lossy* representation allows an approximation to the same image to be rendered—an image that is different from the original in ways the human eye may or may not be able to discern.

One method used for lossless image compression takes advantage of the fact that in most images, the color doesn't change from pixel to pixel—the image has *spatial coherence*, to use the official term. Looking at the middle and rightmost images in Figure 3.13, for example, makes clear that in the 100×100 resolution image, the 100 pixels in a

AUDIO COMPRESSION

MP3 is a lossy compression method for audio. It uses a variety of tricks to create small data files. For example, human ears are not far enough apart to hear low-frequency sounds stereophonically, so MP3s may record low frequencies in mono and play the same sound to both speakers, while recording and playing the higher frequencies in stereo! MP3s are "good enough" for many purposes, but a trained and sensitive ear can detect the loss of sound quality.

10 × 10 square in the top-left corner are all the same color; there is no need to repeat a 24-bit color value 100 times in the representation of the image.

Accordingly, graphic representations have ways of saying "all pixels in this block have the same color value." Doing so can reduce the number of bits significantly.

Depending on how an image will be used, a lossy compression method might be acceptable. What flashes on your TV is gone before you have time to scrutinize the individual pixels. But in some cases, only lossless compression is satisfactory. If you have the famous Zapruder film of the Kennedy assassination and want to preserve it in a digital archive, you want to use a lossless compression method once you have digitized it at a suitably fine resolution. But if you are just shipping off the image to a low-quality printer such as those used to print newspapers, lossy compression might be fine.

Technological Birth and Death

The digital revolution was possible because the capacity of memory chips increased, relentlessly following Moore's Law. Eventually, it became possible to store digitized images and sounds at such high resolution that their quality was higher than analog representations. Moreover, the price became low enough that the storage chips could be included in consumer goods. But more than electrical engineering is involved. At more than a megabyte per image, digital cameras and HD televisions would still be exotic rarities. A *megabyte* is about a million bytes, and that is just too much data per image. The revolution also required better algorithms—better computational methods, not just better hardware—and fast, cheap processing chips to carry out those algorithms.

For example, digital video compression utilizes *temporal coherence* as well as spatial coherence. Any portion of the image is unlikely to change much in color from frame to frame, so large parts of a picture typically do not have to be retransmitted to the home when the frame changes after a thirtieth of a second. At least, that is true in principle. If a woman in a TV image walks across a fixed landscape, only her image, and a bit of landscape that newly appears from behind her once she passes it, needs be transmitted—*if* it is computationally feasible to compare the second frame to the first before it is transmitted and determine exactly where it differs from its predecessor. To keep up with the video speed, there is only a thirtieth of a second to do that computation. And a complementary computation has to be carried out at the other end—the previously transmitted frame must be modified to reflect the newly transmitted information about what part of it should change one frame time later.

Digital movies could not have happened without an extraordinary increase in speed and drop in price in computing power. Decompression algorithms are built into desktop photo printers and cable TV boxes, cast in silicon in chips more powerful than the fastest computers of only a few years ago. Such compact representations can be sent quickly through cables and as satellite signals. The computing power in the cable boxes and television sets is today powerful enough to reconstruct the image from the representation of what has changed. Processing is power.

By contrast, part of the reason the compact disk is dying as a medium for distributing music is that it doesn't hold enough data. At the time the CD format was adopted as a standard, decompression circuitry for CD players would have been too costly for use in homes and automobiles, so music could not be recorded in compressed form. The magic of Apple's iPod is not just the huge capacity and tiny physical size of its disk—it is the power of the processing chip that renders the stored model as music.

The birth of new technologies presage the death of old technologies. Digital cameras killed the silver halide film industry; analog television sets will soon be gone; phonograph records gave way to cassette tapes, which in turn gave way to compact disks, which are themselves now dying in favor of digital music players with their highly compressed data formats.

The periods of transition between technologies, when one emerges and threatens another that is already in wide use, are often marked by the exercise of power, not always progressively. Businesses that dominate old technologies are sometimes innovators, but often their past successes make them slow to change. At their worst, they may throw up roadblocks to progress in an attempt to hold their ground in the marketplace. Those roadblocks may include efforts to scare the public about potential disruptions to familiar practices, or about the dollar costs of progress.

Data formats, the mere conventions used to intercommunicate information, can be remarkably contentious, when a change threatens the business of an incumbent party, as the Commonwealth of Massachusetts learned when it tried to change its document formats. The tale of Massachusetts and OpenDocument illustrates how hard change can be in the digital world, although it sometimes seems to change on an almost daily basis.

Data Formats as Public Property

No one owns the Internet, and everyone owns the Internet. No government controls the whole system, and in the U.S., the federal government controls only the computers of government agencies. If you download a web page to

your home computer, it will reach you through the cooperation of several, perhaps dozens, of private companies between the web server and you.

UPLOADING AND DOWNLOADING

Historically, we thought of the Internet as consisting of powerful corporate "server" machines located "above" our little home computers. So when we retrieved material from a server, we were said to be "downloading," and when we transferred material from our machine to a server, we were "uploading." Many personal machines are now so powerful that the "up" and "down" metaphors are no longer descriptive, but the language is still with us. See the Appendix, and also the explanation of "peer-to-peer" in Chapter 6, "Balance Toppled."

This flexible and constantly changing configuration of computers and communication links developed because the Internet is in its essence not hardware, but protocols—the conventions that computers use for sending bits to each other (see the Appendix). The most basic Internet Protocol is known as IP. The Internet was a success because IP and the designs for the other protocols became public standards, available for anyone to use. Anyone could build on top of IP. Any proposed higher-level protocol could be adopted as a public standard if it met the approval of the networking community. The most important protocol exploiting IP is known as TCP. TCP is used by email and web software to ship messages reliably between computers, and the pair of protocols is known as TCP/IP. The Internet might not have developed that way had proprietary networking protocols taken hold in the early days of networking.

It was not always thus. Twenty to thirty years ago, all the major computer companies—IBM, DEC, Novell, and Apple—had their own networking protocols. The machines of different companies did not intercommunicate easily, and each company hoped that the rest of the world would adopt its protocols as standards. TCP/IP emerged as a standard because agencies of the U.S. government insisted on its use in research that it sponsored—the Defense Department for the ARPANET, and the National Science Foundation for NSFnet. TCP/IP was embedded in the Berkeley Unix operating system, which was developed under federal grants and came to be widely used in universities. Small companies quickly moved to use TCP/IP for their new products. The big companies moved to adopt it more slowly. The Internet, with all of its profusion of services and manufacturers, could not have come into existence had one of the incumbent manufacturers won the argument—and they failed even though their networking products were technologically superior to the early TCP/IP implementations.

File formats stand at a similar fork in the road today. There is increasing concern about the risks of commercial products evolving into standards. Society will be better served, goes the argument, if documents are stored in formats hammered out by standards organizations, rather than disseminated as part of commercial software packages. But consensus around one *de facto* commercial standard, the `.doc` format of Microsoft Word, is already well advanced.

Word's `.doc` format is proprietary, developed by Microsoft and owned by Microsoft. Its details are now public, but Microsoft can change them at any time, without consultation. Indeed, it does so regularly, in order to enhance the capabilities of its software—and new releases create incompatibilities with legacy documents. Some documents created with Word 2007 can't be opened in Word 2003 without a software add-on, so even all-Microsoft offices risk document incompatibilities if they don't adjust to Microsoft's format changes. Microsoft does not exclude competitors from adopting its format as their own document standard—but competitors would run great risks in building on a format they do not control.

In a large organization, the cost of licensing Microsoft Office products for thousands of machines can run into the millions of dollars. In an effort to create competition and to save money, in 2004 the European Union advanced the use of an "OpenDocument Format" for exchange of documents among EU businesses and governments. Using ODF, multiple companies could enter the market, all able to read documents produced using each other's software.

In September, 2005, the Commonwealth of Massachusetts decided to follow the EU initiative. Massachusetts announced that effective 15 months later, all the state's documents would have to be stored in OpenDocument Format. About 50,000 state-owned computers would be affected. State officials estimated the cost savings at about $45 million. But Eric Kriss, the state's secretary of administration and finance, said that more than software cost was at stake. Public documents were public property; access should never require the cooperation of a single private corporation.

Microsoft did not accept the state's decision without an argument. The company rallied advocates for the disabled to its side, claiming that no available OpenDocument software had the accessibility features Microsoft offered. Microsoft, which already had state contracts that extended beyond the switchover date, also argued that adopting the ODF standard would be unfair to Microsoft and costly to Massachusetts. "Were this proposal to be adopted, the significant costs incurred by the Commonwealth, its citizens, and the private sector would be matched only by the levels of confusion and incompatibility that would result...." Kriss replied, "The question is whether a sovereign state has the obligation to ensure that its public documents remain forever free

OpenDocument, Open Source, Free

These three distinct concepts all aim, at least in part, to slow the development of software monopolies. OpenDocument (opendocument.xml.org) is an open standard for file formats. Several major computer corporations have backed the effort, and have promised not to raise intellectual property issues that would inhibit the development of software meeting the standards. Open source (opensource.org) is a software development methodology emphasizing shared effort and peer review to improve quality. The site openoffice.org provides a full suite of open source office productivity tools, available without charge. Free software—"Free as in freedom, not free beer" (www.fsf.org, www.gnu.org)—"is a matter of the users' freedom to run, copy, distribute, study, change, and improve the software."

and unencumbered by patent, license, or other technical impediments. We say, yes, this is an imperative. Microsoft says they disagree and want the world to use their proprietary formats." The rhetoric quieted down, but the pressure increased. The stakes were high for Microsoft, since where Massachusetts went, other states might follow.

Three months later, neither Kriss nor Quinn was working for the state. Kriss returned to private industry as he had planned to do before joining the state government. The *Boston Globe* published an investigation of Quinn's travel expenses, but the state found him blameless. Tired of the mudslinging, under attack for his decision about open standards, and lacking Kriss's support, on December 24, Quinn announced his resignation. Quinn suspected "Microsoft money and its lobbyist machine" of being behind the *Globe* investigation and the legislature's resistance to his open standard initiative.

The deadline for Massachusetts to move to OpenDocuments has passed, and as of the fall of 2007, the state's web site still says the switchover will occur in the future. In the intervening months, the state explains, it became possible for Microsoft software to read and write OpenDocument formats, so the shift to OpenDocument would not eliminate Microsoft from the office software competition. Nonetheless, other software companies would not be allowed to compete for the state's office software business until "accessibility characteristics of the applications meet or exceed those of the currently deployed office suite"—i.e., Microsoft's. For the time being, Microsoft has the upper hand, despite the state's effort to wrest from private hands the formats of its public documents.

Which bits mean what in a document format is a multi-billion dollar business. As in any big business decisions, money and politics count, reason becomes entangled with rhetoric, and the public is only one of the stakeholders with an interest in the outcome.

Hiding Information in Images

The surprises in text documents are mostly things of which the authors were ignorant or unaware. Image documents provide unlimited opportunities for hiding things intentionally—hiding secrets from casual human observers, and obscuring open messages destined for human recipients so anti-spam software won't filter them out.

The Spam Wars

Many of us are used to receiving email pleas such as this one: *I am Miss Faatin Rahman the only child/daughter of late mrs helen rahman Address: Rue 142 Marcory Abidjan Cote d'ivoire west africa, I am 20 years old girl. I lost my parent, and I have an inheritance from my late mother, My parents were very wealthy farmers and cocoa merchant when they were alive, After the death of my father, long ago, my mother was controling his business untill she was poisoned by her business associates which she suffered and died, ... I am crying and seeking for your kind assistance in the following ways: To provide a safe bank account into where the money will be transferred for investment....*

If you get such a request, don't respond to it! Money will flow out of, not into, your bank account. Most people know not to comply. But mass emails are so cheap that getting one person out of a million to respond is enough to make the spammer financially successful.

"Spam filters" are programs that intercept email on its way into the in-box and delete messages like these before we read them. This kind of spam follows such a standard style that it is easy to spot automatically, with minimal risk that any real correspondence with banks or African friends will be filtered out by mistake.

But the spam artists have fought back. Many of us have received emails like the one in Figure 3.14. Why can't the spam filter catch things like this?

Word-processing software includes the name and size of the font in conjunction with the coded characters themselves, as well as other information, such as the color of the letters and the color of the background. Because the underlying text is represented as ASCII codes, however, it remains relatively easy to locate individual letters or substrings, to add or delete text, and to perform other such common text-processing operations. When a user positions a cursor over the letter on the screen, the program can figure out the location within the file of the character over which the cursor is positioned. Computer software can, in turn, render the character codes as images of characters.

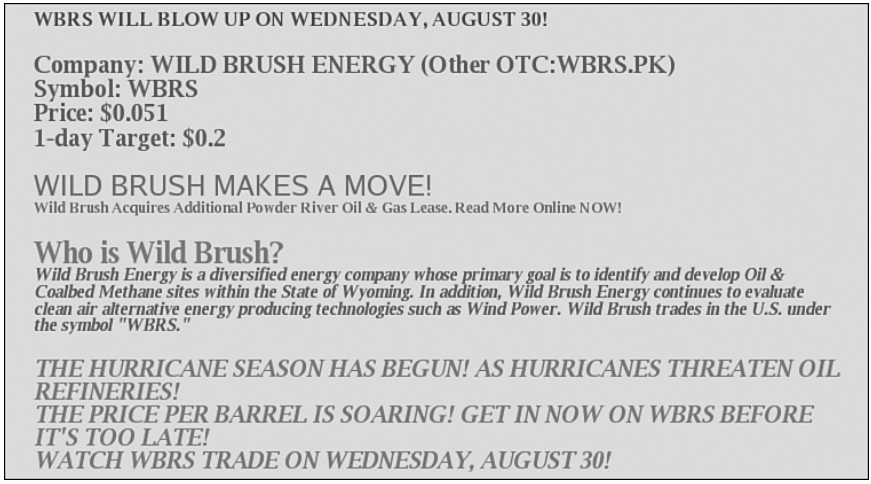


FIGURE 3.14 Graphic spam received by one of the authors. Although it looks like text, the computer "sees" it as just an image, like a photograph. Because it doesn't realize that the pixels are forming letters, its spam filters cannot identify it as spam.

But just because a computer screen shows a recognizable letter of the alphabet, this does not mean that the underlying representation is by means of standard character codes. A digitized photograph of text may well look identical to an image rendered from a word-processing document—that is, the two utterly different representations may give rise to exactly the same image.

And that is one reason why, in the battle between spam producers and makers of spam filters, the spam producers currently have the upper hand. The spam of Figure 3.14 was produced in graphical form, even though what is represented is just text. As the underlying representation is pixels and not ASCII, spam like this makes it through all the filters we know about!

The problem of converting raster graphics to ASCII text is called *character recognition*. The term *optical character recognition*, or OCR, is used when the original document is a printed piece of paper. The raster graphic representation is the result of scanning the document, and then some character recognition algorithm is used to convert the image into a sequence of character codes. If the original document is printed in a standard typeface and is relatively free of smudges and smears, contemporary OCR software is quite accurate, and is now incorporated into commercially available scanners commonly packaged as multipurpose devices that also print, photocopy, and fax. Because OCR algorithms are now reasonably effective and widely available, the next generation of spam filters will likely classify emails such as Figure 3.14 as spam.

OCR and spam are merely an illustration of a larger point. Representation determines what can be done with data. In principle, many representations may be equivalent. But in practice, the secrecy of formatting information and the computation required to convert one format to another may limit the usefulness of the data itself.

Hiding Information in Plain Sight

During World War I, the German Embassy in Washington, DC sent a message to Berlin that began thus: "PRESIDENT'S EMBARGO RULING SHOULD HAVE IMMEDIATE NOTICE." U.S. intelligence was reading all the German telegrams, and this one might have seemed innocuous enough. But the first letters of the words spelled out "PERSHING," the name of a U.S. Navy vessel. The entire telegram had nothing to do with embargoes. It was about U.S. ship movements, and the initial letters read in full, "PERSHING SAILS FROM N.Y. JUNE 1."

Steganography is the art of sending secret messages in imperceptible ways. Steganography is different from *cryptography*, which is the art of sending messages that are indecipherable. In a cryptographic communication, it is assumed that if Alice sends a message to Bob, an adversary may well intercept the message and recognize that it holds a secret. The objective is to make the message unreadable, except to Bob, if it falls into the hands of such an eavesdropper or enemy. In the world of electronic communication, sending an encrypted message is likely to arouse suspicion of electronic monitoring software. By contrast, in a steganographic message from Alice to Bob, the communication itself arouses no suspicion. It may even be posted on a web site and seem entirely innocent. Yet hidden in plain sight, in a way known only to Alice and Bob, is a coded message.

Steganography has been in use for a long time. The *Steganographia* of Johannes Trithemius (1462–1516) is an occult text that includes long conjurations of spirits. The first letters of the words of these mystic incantations encode other hidden messages, and the book was influential for a century after it was written. Computers have created enormous opportunities for steganographic communications. As a very simple example, consider an ordinary word-processing document—a simple love letter, for example. Print it out or view it on the screen, and it seems to be about Alice's sweet nothings to Bob, and nothing more. But perhaps Alice included a paragraph at the end *in which she changed the font color to white*. The software renders the white text on the white background, which looks exactly like the white background.

But Bob, if he knows what to look for, can make it visible—for example, by printing on black paper (just as the text could be recovered from the electronically redacted Calipari report).

If an adversary has any reason to think a trick like this might be in use, the adversary can inspect Alice's electronic letter using software that looks for messages hidden using just this technique. But there are many places to look for steganographic messages, and many ways to hide the information.

Since each Roman letter has an eight-bit ASCII code, a text can be hidden within another as long as there is an agreed-upon method for encoding 0s and 1s. For example, what letter is hidden in this sentence?

```
Steganographic algorithms hide messages inside photos, text, and
other data.
```

The answer is "I," the letter whose ASCII character code is `01001001`. In the first eight words of the sentence, words beginning with consonants encode `0` bits and words beginning with vowels encode `1`s (see Figure 3.15).

```
Steganographic algorithms hide messages inside photos, text, and other data.
0              1          0    0        1      0       0     1
```

FIGURE 3.15 A steganographic encoding of text within text. Initial consonants encode `0`, vowels encode `1`, and the first eight words encode the 8-bit ASCII code for the letter "I."

A steganographic method that would seem to be all but undetectable involves varying ever so slightly the color values of individual pixels within a photograph. Red, green, and blue components of a color determine the color itself. A color is represented internally as one byte each for red, green, and blue. Each 8-bit string represents a numerical value between 0 and 255. Changing the rightmost bit from a `1` to a `0` (for example, changing `00110011` to `00110010`), changes the numerical value by subtracting one—in this case, changing the color value from 51 to 50. That results in a change in color so insignificant that it would not be noticed, certainly not as a change in a single pixel. But the rightmost bits of the color values of pixels in the graphics files representing photographs can then carry quite large amounts of information, without raising any suspicions. The recipient decodes the message not by rendering the bits as visible images, but by inspecting the bits themselves, and picking out the significant `0`s and `1`s.

Who uses steganography today, if anyone? It is very hard to know. *USA Today* reported that terrorists were communicating using steganography in early 2001. A number of software tools are freely available that make steganography easy. Steganographic detectors—what are properly known as steganalysis tools—have also been developed, but their usefulness as yet seems to be limited. Both steganography and steganalysis software is freely available on the World Wide Web (see, for example, `www.cotse.com/tools/stega.htm` and `www.outguess.org/detection.php`).

The use of steganography to transmit secret messages is today easy, cheap, and all but undetectable. A foreign agent who wanted to communicate with parties abroad might well encode a bit string in the tonal values of an MP3 or the color values of pixels in a pornographic image on a web page. So much music and pornography flows between the U.S. and foreign countries that the uploads and downloads would arouse no suspicion!

The Scary Secrets of Old Disks

By now, you may be tempted to delete all the files on your disk drive and throw it away, rather than run the risk that the files contain unknown secrets. That isn't the solution: Even deleted files hold secrets!

A few years ago, two MIT researchers bought 158 used disk drives, mostly from eBay, and recovered what data they could. Most of those who put the disks up for sale had made some effort to scrub the data. They had dragged files into the desktop trash can. Some had gone so far as to use the Microsoft Windows FORMAT command, which warns that it will destroy all data on the disk.

Yet only 12 of the 158 disk drives had truly been sanitized. Using several methods well within the technical capabilities of today's teenagers, the researchers were able to recover user data from most of the others. From 42 of the disks, they retrieved what appeared to be credit card numbers. One of the drives seemed to have come from an Illinois automatic teller machine and contained 2,868 bank account numbers and account balances. Such data from single business computers would be a treasure trove for criminals. But most of the drives from home computers also contained information that the owners would consider extremely sensitive: love letters, pornography, complaints about a child's cancer therapy, and grievances about pay disputes, for example. Many of the disks contained enough data to identify the primary user of the computer, so that the sensitive information could be tied back to an individual whom the researchers could contact.

Cloud Computing

One way to avoid having problems with deleted disk files and expensive document-processing software is not to keep your files on your disks in the first place! In "cloud computing," the documents stay on the disks of a central service provider and are accessed through a web browser. "Google Docs" is one such service, which boasts very low software costs, but other major software companies are rumored to be exploring the market for cloud computing. If Google holds your documents, they are accessible from anywhere the Internet reaches, and you never have to worry about losing them—Google's backup procedures are better than yours could ever be. But there are potential disadvantages. Google's lawyers would decide whether to resist subpoenas. Federal investigators could inspect bits passing through the U.S., even on a trip between other countries.

The users of the computers had for the most part done what they thought they were supposed to do—they deleted their files or formatted their disks. They probably knew not to release toxic chemicals by dumping their old machines in a landfilll, but they did not realize that by dumping them on eBay, they might be releasing personal information into the digital environment. Anyone in the world could have bought the old disks for a few dollars, and all the data they contained. What is going on here, and is there anything to do about it?

Disks are divided into blocks, which are like the pages of a book—each has an identifying address, like a page number, and is able to hold a few hundred bytes of data, about the same amount as a page of text in a book. If a document is larger than one disk block, however, the document is typically not stored in consecutive disk blocks. Instead, each block includes a piece of the document, and the address of the block where the document is continued. So the entire document may be physically scattered about the disk, although logically it is held together as a chain of references of one block to another. Logically, the structure is that of a magazine, where articles do not necessarily occupy contiguous pages. Part of an article may end with "Continued on page 152," and the part of the article on page 152 may indicate the page on which it is continued from there, and so on.

Because the files on a disk begin at random places on disk, an *index* records which files begin where on the disk. The index is itself another disk file, but one whose location on the disk can be found quickly. A disk index is very much like the index of a book—which always appears at the end, so readers know where to look for it. Having found the index, they can quickly find the page number of any item listed in the index and flip to that page.