2014

Governing, Exchanging, Securing: Big Data and the Production of Digital Knowledge

Bernard E. Harcourt
Columbia Law School, bharcourt@law.columbia.edu

COLUMBIA LAW SCHOOL

PUBLIC LAW & LEGAL THEORY WORKING PAPER GROUP

PAPER NUMBER 14-390

GOVERNING, EXCHANGING, SECURING:

BIG DATA AND THE PRODUCTION OF DIGITAL KNOWLEDGE

BERNARD E. HARCOURT

MAY 23, 2014

GOVERNING, EXCHANGING, SECURING:

BIG DATA AND THE PRODUCTION OF DIGITAL KNOWLEDGE

BERNARD E. HARCOURT
UNIVERSITY OF CHICAGO/EHESS

Paper presented at a daylong conference organized by
Professor Pierre-Michel Menger, Collège de France
« Big data, *entreprises et sciences sociales - Usages et partages des données numériques de masse* »
Collège de France, Paris, June 2, 2014

DRAFT: MAY 23, 2014

GOVERNING, EXCHANGING, SECURING:

BIG DATA AND THE PRODUCTION OF A DIGITAL KNOWLEDGE

BERNARD E. HARCOURT*

Let me begin with three anecdotes from the arts of governing, of exchanging, of securing. The first arose in mid-2009 and involved a trove of data—a large database consisting of half a million telephone numbers of cell phone users in Cuba—or, more precisely, a trove of data *and* a few entrepreneurial officials and subcontractors of the United States Agency for International Development (USAID), a United States government agency that delivers billions of dollars of aid and humanitarian assistance to needy countries.[1] The cell phone numbers had surreptitiously been given by an employee at Cubacel, the Cuban state-owned cell phone provider, to a Cuban engineer who was living at the time in Spain; that Cuban engineer then gave the database, "free of charge" according to the documents reviewed by the Associated Press, to officials at USAID and a for-profit Washington DC company, Creative Associates International (which had received millions of dollars in contracted business from USAID).

A manager with Creative Associates then got, well, creative. With her brother in Nicaragua, she hatched the idea to begin sending bulk text messages from different countries to the Cuban cell phone users as a way to circumvent the strict Cuban state control of the Internet and begin a form of social media on the model of Twitter. The idea would be to try to surreptitiously set up, from scratch, an entire platform for Cubans, a basic "Cuban Twitter," with the long-term goal of fomenting political dissent. As the Associated Press reports, after carefully reviewing over 1,000 pages of documents:

* Julius Kreeger Professor of Law and Political Science, University of Chicago; Stephen and Barbara Friedman Visiting Professor of Law, Columbia Law School (academic year 2013-2014); *Directeur d'études*, École des hautes études en sciences sociales. I would like to thank Daniel Henry for exceptional research assistance.

> Documents show the U.S. government planned to build a subscriber base through "non-controversial content": news messages on soccer, music and hurricane updates. Later when the network reached a critical mass of subscribers, perhaps hundreds of thousands, operators would introduce political content aimed at inspiring Cubans to organize "smart mobs" — mass gatherings called at a moment's notice that might trigger a Cuban Spring, or, as one USAID document put it, "renegotiate the balance of power between the state and society."[2]

They gave it the name "ZunZuneo" which means, in Cuban slang, a hummingbird's tweet. At one point, in March 2011, there were about 40,000 subscribers. None of them had any idea that the social media was created, sustained, and fed by USAID operatives. None of them realized that their involvement was being profiled by USAID contractors to determine their political allegiance. None of them realized that the messaging was intended to politically motivate them. USAID set up "a byzantine system of front companies using a Cayman Islands bank account, and recruit[ed] executives who would not be told of the company's ties to the U.S. government," according to the AP investigation.[3] They had a British company set up a corporation in Spain to run ZunZuneo. They also set up a companion website to the cell phone texting service so that the cell phone users could subscribe, offer feedback, and send texts themselves for free. The documents reveal discussion of how to make this all seem legitimate: "Mock ad banners will give it the appearance of a commercial enterprise," a document reads.[4] Most importantly, as AP reports: "'there will be absolutely no mention of United States government involvement,' according to a 2010 memo from Mobile Accord, one of the project's contractors. 'This is absolutely crucial for the long-term success of the service and to ensure the success of the Mission.'"[5] Accordingly, the ZunZuneo team hired a Havana-born satirical artist, Alen Lauzan Falcon, to write twitter-like messages that had a Cuban flavor.

One USAID contractor, Paula Cambronero, was tasked with categorizing the Cuban subscribers as either "pro-revolution," "apolitical," or "anti-revolution" based

on their answers to prompts.[6] According to the AP, "Cambronero collected a sample of more than 700 responses and analyzed them according to two variables. The first was the level of interest in the messages received, and the second was the political nature of the response. She wrote in her report that 68 percent of responses showed mild interest in the texts."[7] Along another dimension, "USAID divided Cuban society into five segments depending on loyalty to the government. On one side sat the 'democratic movement,' called 'still (largely) irrelevant,' and at the other end were the 'hard-core system supporters,' dubbed 'Talibanes' in a derogatory comparison to Afghan and Pakistani extremists."[8] The key concern of the project, according to the AP, was to "move more people toward the democratic activist camp without detection."[9] "USAID documents say their strategic objective in Cuba was to 'push it out of a stalemate through tactical and temporary initiatives, and get the transition process going again toward democratic change.'"[10]

ZunZuneo shut down in mid-2012 when the USAID money dried up. Senator Patrick Leahy of Vermont, chairman of the Senate Appropriations Committee's State Department and foreign operations subcommittee, is now holding hearings on this USAID program.

The second anecdote concerns the arts of exchange. In 2004, Google launched Gmail—a platform most of you know well. What made the idea of Gmail so attractive to consumers was that Google provided its users with a large amount of free cloud space: 1 gigabyte of free storage space for each user account.[11] Even before the product went on-line, many were clamoring—and some, in fact, paying—to get priority access to Gmail. In exchange of the free storage space, though, users had to sign off on Google having their own bit of freedom: free access to all of the user's e-mails and attachments—all their *content*—as well as free access to the e-mails of non-subscribing

incoming traffic, that is, the e-mails of any non-subscriber communicating with a Gmail user.

Within the Gmail platform, Google would automatically scan all the e-mails and content in order to target advertising to its users. This is something that most of us know and have come to accept, to live with—as the cost of doing free business. But as an investigative reporter, Yasha Levine, has revealed and emphasizes, Google's commercial surveillance went far beyond that:

> Google was not simply scanning people's emails for advertising keywords, but had developed underlying technology to compile sophisticated dossiers of everyone who came through its email system. All communication was subject to deep linguistic analysis; conversations were parsed for keywords, meaning and even tone; individuals were matched to real identities using contact information stored in a user's Gmail address book; attached documents were scraped for intel — that info was then cross-referenced with previous email interactions and combined with stuff gleamed from other Google services, as well as third-party sources…[12]

Based on a close analysis of two patents that Google filed prior to launching its Gmail service, Levine reports that the company uses a wide range of technologies to construct profiles of its users—to produce *digital knowledge*. These include analyzing the concepts and topics that are discussed in users' emails, as well as in their email attachments; analyzing the content of the websites that different users have visited; tracking the demographic information about users, including their income, sex, race, and marital status; linking this to their geographic information; inferring their psychological and "psychographic" information, such as their personality type, values, lifestyle interests and attitudes; dissecting the previous Internet searches that users have made; collecting information about any documents that the user has viewed and edited; analyzing their Internet browsing activity; and studying their previous purchases.[13]

The goal, naturally, is to produce profiles of each user so as to target advertising and help facilitate consumption—and thereby to reap a substantial profit for Google.

Or, in more coded terms, to "make the on-line experience more personal and more pleasurable for the user/consumer." This has now been enhanced with the other platforms that Google has introduced, from Google Calendar to Google Docs to Google Hang-outs, etc. As Levine writes, "Google isn't a traditional Internet service company. It isn't even an advertising company. Google is a whole new type of beast: a global advertising-intelligence company that tries to funnel as much user activity in the real and online world through its services in order to track, analyze and profile us: it tracks as much of our daily lives as possible — who we are, what we do, what we like, where we go, who we talk to, what we think about, what we're interested in — all those things are seized, packaged, commodified and sold on the market ..."[14]

This type of corporate commercial surveillance has become so significant, in fact, that it has come to dwarf—and, as we will see, it feeds—the type of surveillance being conducted by the NSA, the GCHQ, and the Direction Générale de la Sécurité Extérieure (DGSE). As Levine suggests, Google's Gmail and other services represent "a massive surveillance operation that intercepts and analyzes terabytes of global Internet traffic every day, and then uses that data to build and update complex psychological profiles on hundreds of millions of people all over the world — all of it in real time."[15]

The third and final anecdote concerns the arts of securing and securitizing. The British intelligence initiative was code-named "Optic Nerve."[16] It is not known today whether the Government Communications Headquarters (GCHQ), the British signals intelligence agency, is still doing it, although there is evidence they were still operating the program in 2012. What is known is that, during a period of six months in 2008, the British signals intelligence division surreptitiously intercepted screenshots of the webcam video communications of about 1.8 million internet users using video chat platforms like those provided by Yahoo Messenger.

Unlike some other programs that only capture metadata, Optic Nerve was able to access the content of the video communications—the actual video images streaming on the chat. Apparently, it "automatically downloaded the content of video communications—taking a screenshot from the video feed every five minutes."[17] According to a secret report revealed by Edward Snowden, British intelligence aspired to capture more images at a faster clip and hoped to get the full webcam videos at some point, at least for surveillance targets, intending to "identify targets using automatic facial recognition software as they stared into their computer's webcams."[18] GCHQ was assisted, in these efforts, by the NSA: "Webcam information was fed into NSA's XKeyscore search tool, and NSA research was used to build the tool which identified Yahoo's webcam traffic."[19]

Apparently, the operation has netted a trove of X-rated images. An intelligence document regarding the program stated that "It would appear that a surprising number of people use webcam conversations to show intimate parts of their body to the other person;" according to one informal analysis, somewhere around seven percent of the recorded images contained "undesirable nudity."[20] As the Associated Press noted, "The collection of nude photographs also raises questions about potential for blackmail. America's National Security Agency has already acknowledged that some analysts have been caught trawling databases for inappropriate material on partners or love interests. Other leaked documents have revealed how U.S. and British intelligence discussed leaking embarrassing material online to blacken the reputations of their targets."[21]

Other documents leaked by Snowden reveal that the NSA has been investigating the possibility of using game console video communication as a way to intercept data: "the NSA were exploring the video capabilities of game consoles for surveillance purposes," the Guardian reported. "Microsoft, the maker of Xbox, faced a privacy backlash last year when details emerged that the camera bundled with its new console, the Xbox One, would be always-on by default."[22]

***

Each one of these initiatives in governing, exchanging, and securing generate a trove of additional data about individuals that can be captured, recorded, interlinked, connected to other data, mined, studied, and analyzed. Each one can produce a trove of data for purposes of social science research—and we will be hearing a lot about that today. But equally, each one can produce a trove of data to identify individuals, nudge them politically, manipulate them slightly, encourage and stimulate their consumption and disclosures, watch and surveil them, detect, predict, and punish. The data allow governments, transnational corporations and everyday businesses, employers, salesmen, advertisers, the police and parole workers to track individuals' physical movements, follow their internet browsing, know what they read, what they like, what they wear, whom they communicate with, where and on what they spend their money.

A new *digital knowledge* has emerged that has begun to challenge the boundaries between governing, exchange, and security—in other words, to challenge the conventional lines between politics, economics, and policing. A *digital knowledge* that ambiguates the lines between commerce and surveillance, between governing and exchanging, between democracy and the police state. This new *digital knowledge* produces and reproduces consuming subjects who wittingly or unwittingly allow themselves to be watched, tracked, linked and predicted in a blurred amalgam of commercial and government projects. Linking back and forth from consumer data to government information to social media, these new webs of information become available to anyone who can purchase the information.

How is it that governmental, commercial and security interests have converged, coincided, and also diverged in ways, in this production of Big Data? Which sectors have stimulated the production and mining of all this information? How have the

various projects aligned or contradicted each other? How is it, for example, that every new digital technology seems to make security easier?

In the remainder of this paper, I will explore these questions along two dimensions. First, I will lay out the historical development and growth in the production of the digital realm. I will offer some categories to understand the mass of data that surrounds us today, and lay some foundation for the notion of a *digital knowledge*. Second, I will investigate the new political economy of data that has emerged as a way to suggest some of the larger forces that are at play in our new digital age.

I. A Historico-Sociological View of Big Data

I will begin, then, by tracing in broad strokes the historical trajectory of the development of these large troves of data, locate which sectors have promoted the accumulation, and propose various ways of categorizing the mass of information we have collected today.

A. The Historical Trajectory of Big Data

"Big Data"[23] consists of a wide variety of things, including mobile communications, social media, e-mails, videos, chats, vital statistics, government census data, genomics, satellites, and sensors.[24] (Sensors include hospital and home sensors in the health care industry used to monitor key biochemical markers, in real time flows of information that are analyzed on reception.[25]) The IDC, a premier IT research and consulting firm, provides a basic measure of the extent of "big data" in 2013.[26] The total amount of information stored in what IDC refers to as the "Digital Universe" measures to roughly 4.4 zettabytes (a zettabyte is equivalent to $1.18059162 \times 10^{21}$ bytes). By 2020, this number is expected to increase to 44 zettabytes.[27]

The most scientific, rigorous measurement of the world's technological capacity to store, communicate, and compute data was collected and presented in a research

article in *Science* magazine in April 2011—it is the source of most of the descriptions today about the quantity of data and the folkloric image of stacking CD-ROM's of data to the moon.[28] The authors—Martin Hilbert and Priscila López—tried to measure and compare the technological capacity using analog versus digital technologies over the period from 1986 to 2007, including the following technologies[29]:

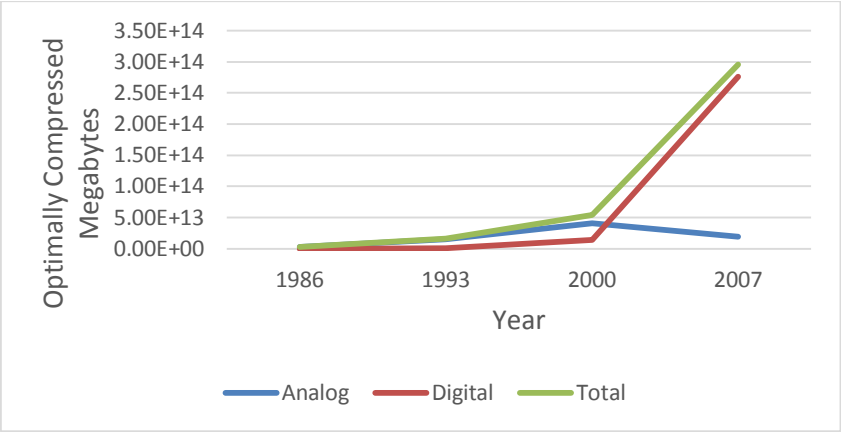
	Analog	Digital
Storage	Video analog; photo print; audio cassette; photo negative; cine movie film; vinyl LP; TV episodes film; X Rays; TV movie film; newsprint; other paper and print; books	PC hard-disk; DVD and Blu-Ray; digital tape; server and mainframe hard-disk; CDs and minidisks; other hard-disks (portable); portable media player; memory cards; mobile phones and PDA; videogames; floppy disks; digital camera and camcorders internal; chipcards
Communication	Broadcasting: TV-terrestrial; TV-cable; TV-satellite; radio; newspapers; paper advertisements	Broadcasting: TV-terrestial; TV-cable; TV-satellite; radio; personal navigation GPS
	Telecom: Fixed (voice) phone; mobile (voice) phone; paper postal letters	Telecom: Fixed (voice) phone; internet; mobile (data) phone; mobile (voice) phone
Computation		General purpose: Personal computers; videogame consoles; servers and mainframe; supercomputers; pocket calculators; mobile phones/PDA Digital Signal Processors: CD, DVD and PVR, cameras and camcorders, modems and setup boxes, GPS, portable media, printer and fax, radio, fixed phone, mobile phone; microcontrollers; graphic processing

Hilbert and López rigorously measure and compare the technological capacity to perform these three tasks, showing the increase over the period from 1986 to 2007.

Here, first, is the storage capacity from a historical perspective. Notice that the overall capacity was originally growing steadily because of analog technological

developments, at least through 2000. But after that, the growth was exponential in the digital realm—as illustrated by the graph below Table 1.

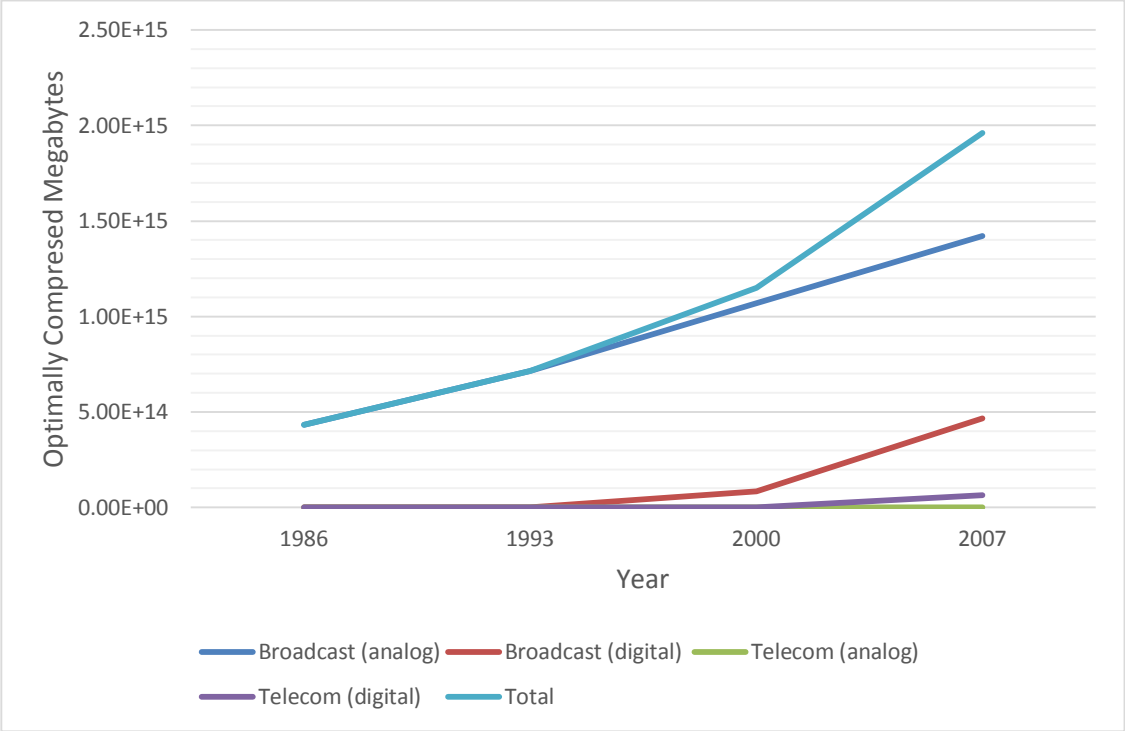
Table 1. World's technological installed capacity to store information, in optimally compressed Megabytes (MB) per year, for 1986, 1993, 2000 and 2007.[30]				
	1986	1993	2000	2007
Analog	2.62E+12	1.52E+13	4.08E+13	1.89E+13
Digital	2.08E+10	5.33E+11	1.37E+13	2.76E+14
Total	*2.64E+12*	*1.58E+13*	*5.45E+13*	*2.95E+14*



Next, Hilbert and López measure the growth in the capacity to broadcast and telecommunicate information:

Table 2. World's technological effective capacity to broadcast and telecommunicate information, in optimally compressed Megabytes (MB) per year, for 1986, 1993, 2000 and 2007.[31]				
	1986	1993	2000	2007
Broadcast	**4.32E+14**	**7.15E+14**	**1.15E+15**	**1.89E+15**
Analog	4.32E+14	7.15E+14	1.07E+15	1.42E+15
Digital	-	-	8.37E+13	4.68E+14
Telecom	**2.81E+11**	**4.71E+11**	**2.24E+12**	**6.54E+13**
Analog	2.25E+11	1.48E+11	5.15E+10	3.63E+10
Digital	5.57E+10	3.23E+11	2.19E+12	6.53E+13
Total	*4.32E+14*	*7.16E+14*	*1.15E+15*	*1.96E+15*

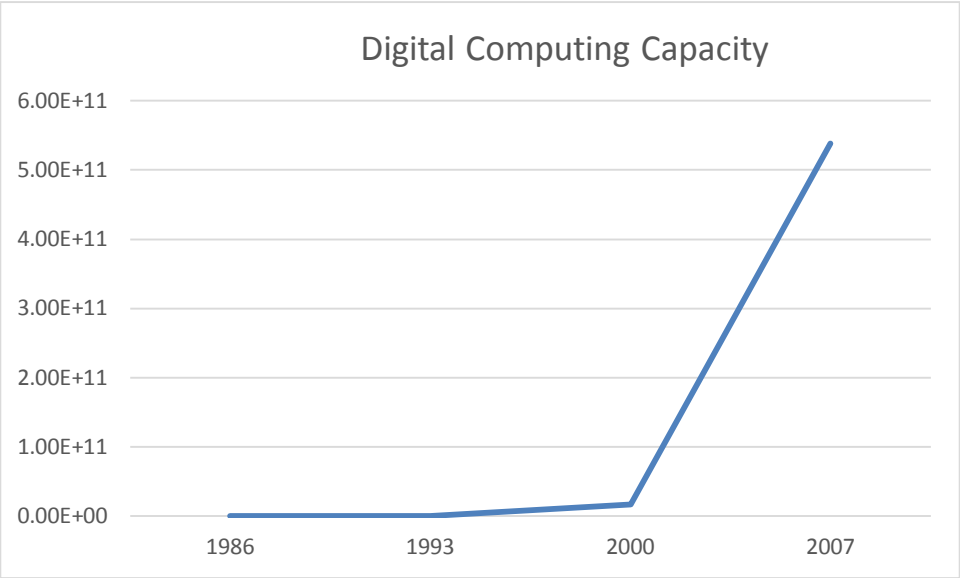
Notice here that most of the capacity remains analog, and that analog capacity has been increasing steadily, but that the augmented shift since 2000 has been in the digital area, as represented in this graph visualizing table 2:



Computation has always been digital, so there is no point of comparison here, simply an exponential increase since 2000 in technological capacity, as evidenced by the graph following Table 3:

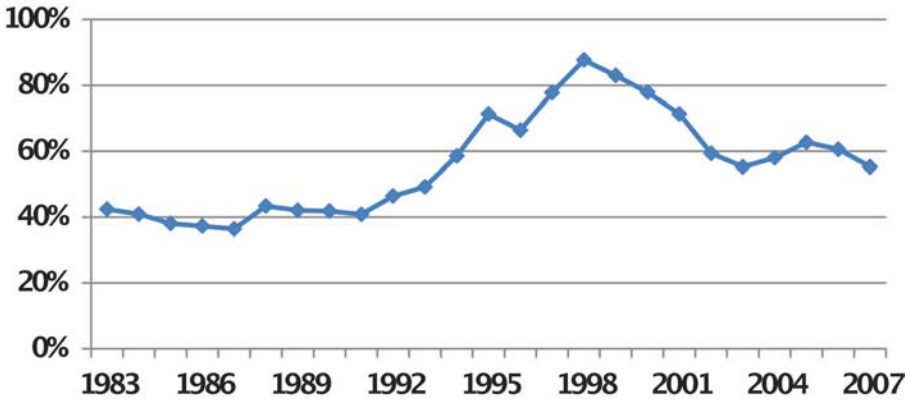
Table 3. World's technological effective capacity to compute information on general-purpose computers (gross usage), in MIPS (million instructions per second), for 1986, 1993, 2000 and 2007.[32]

	1986	1993	2000	2007
Total	1.73E+07	4.02E+08	1.66E+10	5.38E+11



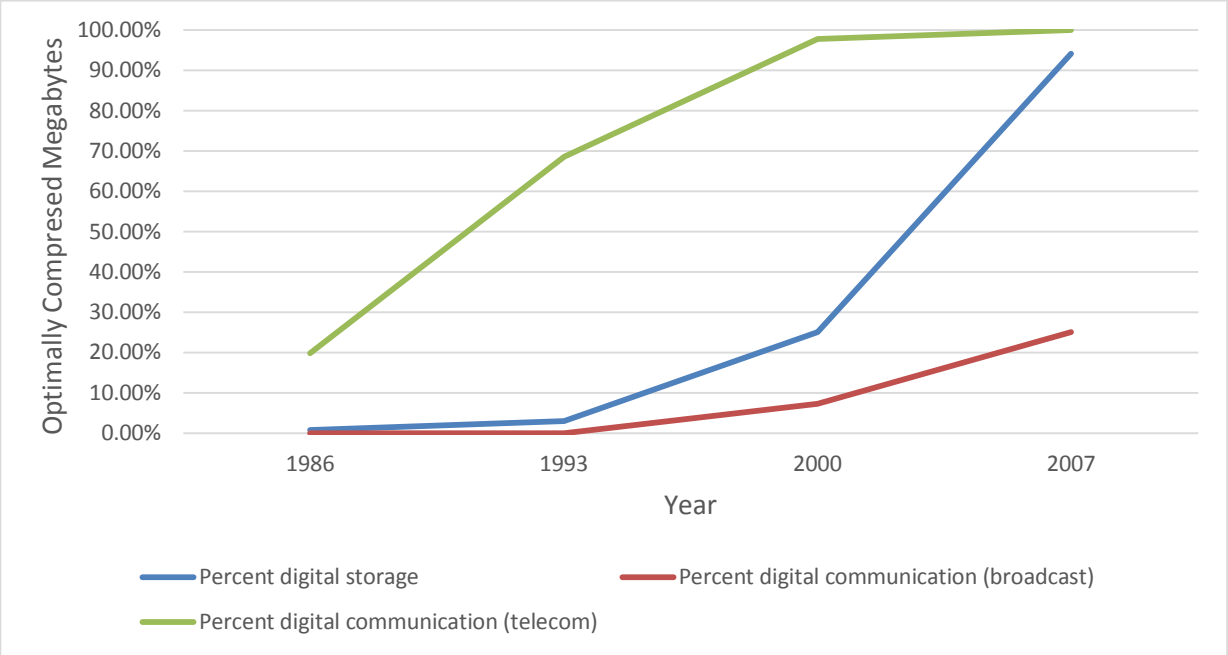
Hilbert and López's study reveals an absolutely remarkable increase in the annual percentage rate of growth in computing power. This is represented in their own Figure 6, reproduced below:

"Fig. 6. Annual growth of installed general-purpose computational capacity as percentage of all previous computations since 1977 (year t / S[1977, year t – 1])."[33]



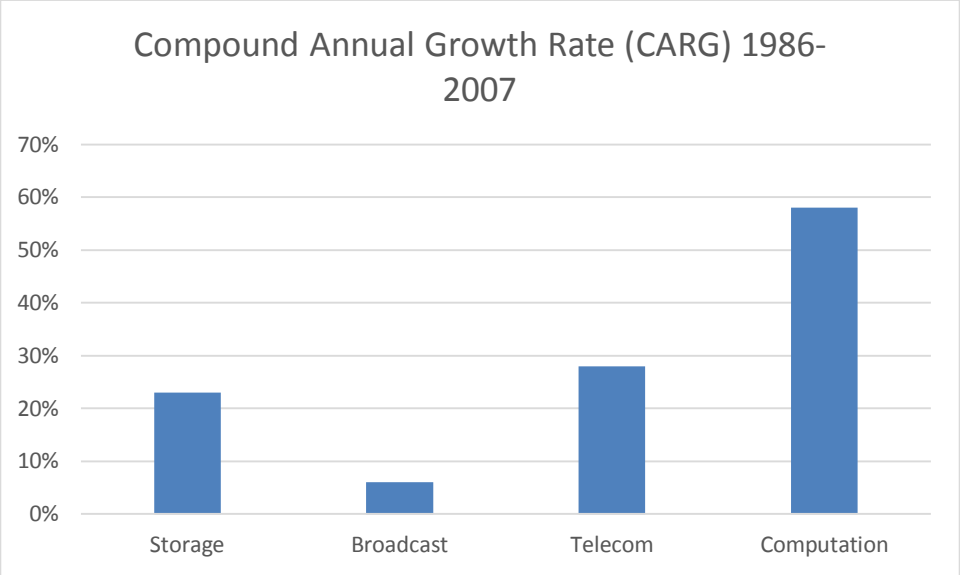
As their data demonstrate, the percentage of the technological capacity being taken over by digital media is increasing at remarkable rates, as evidenced in Table 4 and the following graph:

Table 4: Evolution of the world's capacity to store and communicate information, CARG and percentage in digital format.[34]						
		1986	1993	2000	2007	CARG 1986-2007
Percent digital storage		.8%	3%	25%	94%	23%
Percent digital communication	Broadcast	0%	0%	7.3%	25%	6%
	Telecom	19.8%	68.5%	97.7%	99.9%	28%



If you pull the data on annual growth from the last column of Table 4, you get the following in terms of the compound annual rate of growth in the three major areas of technological capacity:

Table 5: Evolution of the world's technological capacity in terms of Compound Annual Growth Rate (CARG) of digital format.[35]		
		Compound Annual Growth Rate (CARG) 1986-2007
Digital storage		23%
Digital communication	Broadcast	6%
	Telecom	28%
Computation		58%



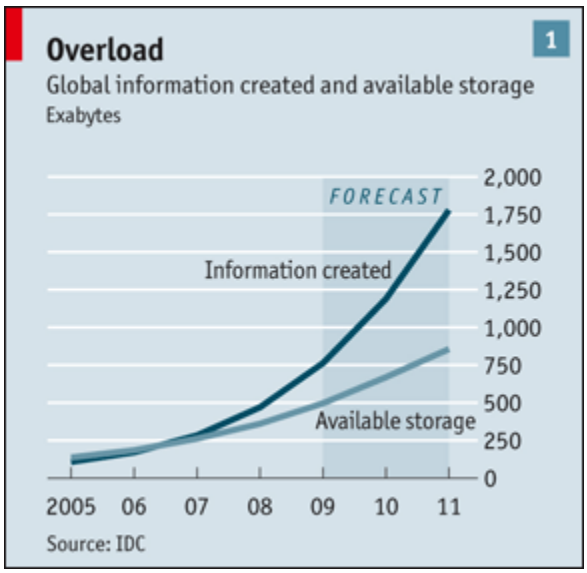
The IDC measured the prevalence of "digital things" within the total universe of "things" that manage the physical world—trying to map onto the physical world the old analog stuff like cars and turbines alongside the new digital devices. It defines the "Internet of Things" as consisting of "adding computerization, software, and intelligence to things as varied as cars, toys, airplanes, dishwashers, turbines, and dog collars." In 2013, "connected 'things' were 7% of the total" number of connectable things in the world; however, "by 2020, that number will grow to 15%." The Internet of

Things is thus "exploding" as "analog functions managing the physical world migrate to digital functions."[36]

Furthermore, the Internet of Things is a primary driver of big data accumulation. This is especially the case for mobile devices—for purposes of this analysis, not just mobile phones, but also "devices such as RFID tags, GPS devices, smart cards, cars, toys, and even dog collars")—which in 2014 "generate 18% of the digital universe," and will constitute 27% in 2020.[37] The IDC adds that internet of things—which triggers "the migration of analog functions monitoring and managing the physical world to digital functions involving communications and software telemetry"—has a profound effect on our surroundings:

> All earlier eras involved the computerization of enterprises or people, of which there are a finite number on the planet. This era involves the computerization, adding software and intelligence, to things – things as varied as cars and toys, airplanes and dishwashers, turbines and dog collars."[38]

We know, then, that big data has grown exponentially over recent time. The Federal Big Data Commission of the TechAmerica Foundation estimates that, in 2011, "1.8 zettbytesabytes of information were created globally, and that amount is expected to double every year."[39] This is based on the IDC's Digitial Universe Study, and is consistent with the *Economists*' numbers. According to IDC, the exponential growth is expected to reach 35 Zettabytes in 2020.[40] As *The Economist* noted several years ago, and is reflected in the above data, the growth of data production far surpasses the capability to store it. This graph is from a detailed report issued on February 27, 2010[41]—which, in our digital age, feels like pre-history:



McKinsey underscores this as well: the rate of data growth is far outstripping storage capacity. In 2007, for the first time, data creation exceeded worldwide storage capacity. The worldwide data total is expected to increase at a rate of 40% per year, with both data storage and data computation capacity also continuing to rise—as noted by Hilbert and Lopez, storage having increased at an annual rate of 23% between 1986 and 2007, while computation capacity grew at a rate of 58% per year during the same period.[42] In 2011, McKinsey reported some amazing statistics about what it called this "growing torrent" of data:

- "5 billion mobile phones in use in 2010
- 30 billion pieces of content shared on Facebook every month
- 235 terabytes data collected by the US Library of Congress by April 2011
- 15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress."[43]

Harvard Business Review reports that, "As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so. More data cross the internet every second than were stored in the entire internet just 20 years ago."

Walmart alone, in fact, "collects more than 2.5 petabytes of data every hour from its customer transactions."[44]

B. The Composition of Big Data

Big data is sliced and diced in a number of different ways—by form, source, type, purpose, etc. Here are a few useful ways to describe the emerging torrent of data.

(a) *Sources of data*

According to McKinsey, in 2009, the three top sectors relative to the quantity of information stored by sector were (1) discrete manufacturing (966 petabytes), (2) government (848 petabytes), and (3) communications and media (715 petabytes).[45]

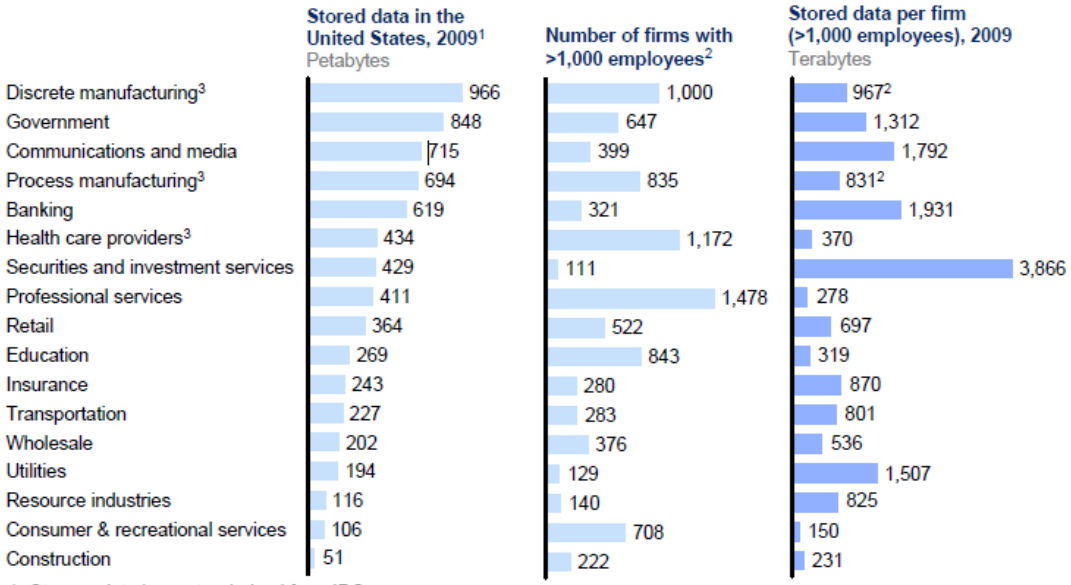
The IDC reports in 2013 that roughly two-thirds of the information in the digital universe, or about 2.9zb, is "generated by consumers," the rest, or about 1.5zb, is generated by enterprises. Business touches, though, about 85% or 2.3zb, of the data, in contrast to about 0.6zb or 15% that is not touched by enterprises.[46]

Only a portion, then, of data is generated or collected by governments. Despite this, it is fair to say that, since 2000 in the United States, "the amount of information the federal government captures has increased exponentially."[47] To give a sense of this, in 2009 alone, "the U.S. Government produced 848 petabytes of data and U.S. healthcare data alone reached 150 exabytes. Five exabytes (10^18 gigabytes) of data would contain all words ever spoken by human beings on earth. At this rate, Big Data for U.S. healthcare will soon reach zetabyte (10^21 gigabytes) scale and soon yottabytes (10^24 gigabytes)."[48]

The other sectors of the economy that store most of the country's data can be visualized on the following chart, prepared by McKinsey[49]:

Exhibit 7

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

(b) *Structured versus unstructured*

The Economist reported in 2010 that "Only 5% of the information that is created is 'structured,' meaning it comes in a standard format of words or numbers that can be read by computers. The rest are things like photos and phone calls which are less easily retrievable and usable. But this is changing as content on the web is increasingly 'tagged,' and facial-recognition and voice-recognition software can identify people and words in digital files."[50]

About three years later, TechAmerica Foundation reports that about 15% of today's data is structured, consisting of rows and columns of data; about 85% consists of unstructured or information that is humanly generated.[51] This means a ton of

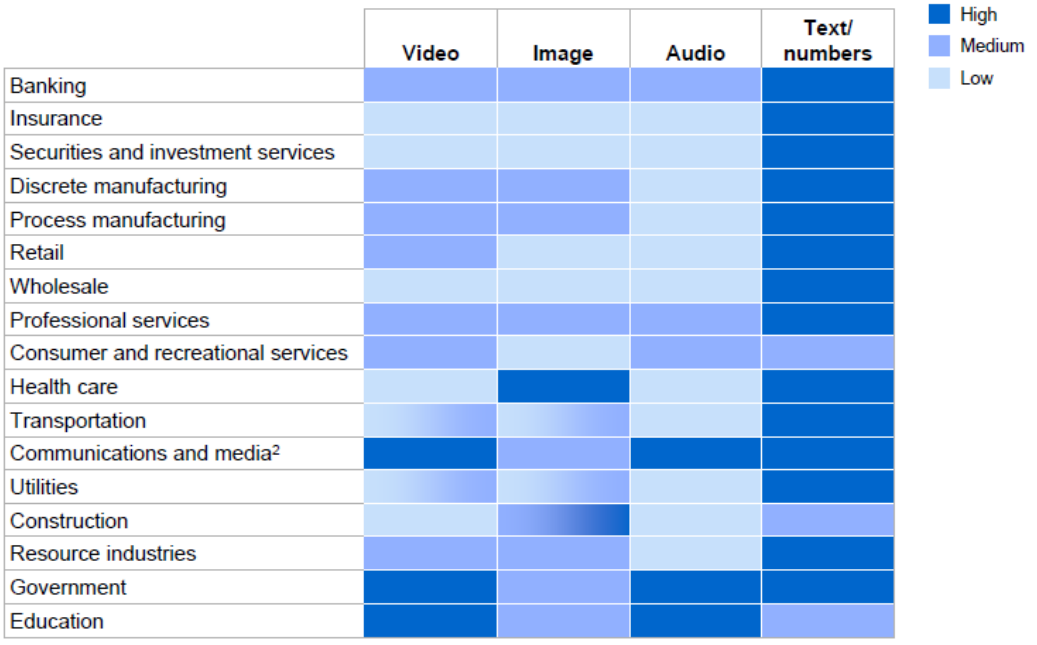
unstructured data that is difficult to process. As the *MIT Sloan Management Review* reported in 2012:

> There is no question that organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analyzed through traditional means. Among its burgeoning sources are the clickstream data from the Web, social media content (tweets, blogs, Facebook wall postings, etc.) and video data from retail and other settings and from video entertainment. But big data also encompasses everything from call center voice data to genomic and proteomic data from biological research and medicine. Everyday, Google alone processes about 24 petabytes (or 24,000 terabytes) of data. Yet very little of the information is formatted in the traditional rows and columns of conventional databases.[52]

However, the structured or unstructured nature of the data—whether audio, visual, or text—varies by the different sectors of the economy, as demonstrated by McKinsey's table of the different sectors[53]:

Exhibit 8

The type of data generated and stored varies by sector[1]



	Video	Image	Audio	Text/ numbers
Banking				
Insurance				
Securities and investment services				
Discrete manufacturing				
Process manufacturing				
Retail				
Wholesale				
Professional services				
Consumer and recreational services				
Health care				
Transportation				
Communications and media[2]				
Utilities				
Construction				
Resource industries				
Government				
Education				

1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

(c) *Types of Data I: smart, identity, and people*

Higinio Maycotte, in an article in *Wired*, suggests that there are three particularly important and prevalent types of big data that should be explained: "smart data," "identity data," and "people data."[54]

"Smart data" is the term for big data that has been "siloed, segmented and then visualized" for a particular need.[55] It has been made legible from its original state as a large collection of binary numbers, and no longer requires an expert to be analyzed.

"Identity data" is possibly the most important type of big data going forward. It is "the force behind predictive modeling and machine learning," and it seeks to "tell the story of who you are in the digital age, including what you like, what you buy, your lifestyle choices and at what time or intervals all of this occurs."[56] It includes social media, purchases, behavior analytics, and more. As Maycotte notes, when Target's customer data was hacked last year, "it was the loss of identity data...that became the biggest issue;" in this case, the identity data stolen was "Credit card numbers associated with names and physical addresses, as well as email addresses."[57]

Lastly, "people data" is created by "aggregating social data over time." This is collected by looking at the "data exhaust" of a large number of users: "who your audience likes and follows on social media, what links they click, how long they stay on the site that they clicked over to and how many converted versus bounced."[58]

(d) *Types of Data II: social, machine, and transactional.*

This first tripartite division overlaps, in part, with another way of dissecting digital data into "social data," "machine data," and "transactional data."[59]

Social data here refers largely to the "people data" described previously by Maycotte. It is generated mostly by consumer behavior on the Internet. It includes social

media data, which is massive: "with 230 million tweets posted on Twitter per day, 2.7 billion Likes and comments added to Facebook every day, and 60 hours of video uploaded to YouTube every minute."[60]

Machine data "consists of information generated from industrial equipment, real-time data from sensors that track parts and monitor machinery (often also called the Internet of Things), and even web logs that track user behavior online." An example of this is the Large Hadron Collider (LHC) machine at the CERN, the largest research center on particle physics on the globe, that generates approximately 40 terabytes of data per second when experiments are ongoing.

Lastly, transactional data is generated by aggregated, recorded transactions from a company's daily business, including the items sold, "product IDs, prices, payment information, manufacturer and distributor data."[61] Think here of Amazon.com or Domino's Pizza—the latter of which serves about a million customers per day—that produce huge quantities of big data on a daily basis.

(e) *Types of Data III: individuals, public sector, private sector*

The World Economic Forum tends to think of data as generated from three different sectors. The first are individuals, who provide a particular type of data: "Crowdsourced" information, social data from mobile phones, and other "data exhaust." The WEF emphasizes, in this regard, that "The data emanating from mobile phones holds particular promise, in part because for many low-income people it is their only form of interactive technology, but it is also easier to link mobile-generated data to individuals."[62] And there is tons of this data being produced: "online or mobile financial transactions, social media traffic, and GPS coordinates now generate over 2.5 quintillion bytes of so-called 'big data' every day."[63]

The second involves the public and development sectors. This includes, especially, governmental census data, health and vital statistics and indicators, tax and expenditure data, and other "facility data." The third sector is the private sector, and it includes transaction data, purchases, spending, consumption and use information.[64]

These then are some of the ways of categorizing the torrent of data that are emerging in this digital age—and that can help us understand this new digital realm.

II. A Political Economy of Data

With the exponential growth of technological capacity, there has emerged an entire political economy of data—an industry replete with data brokers and full-fledged markets. In 2012 alone, the data brokerage industry reached $156 billion in revenue, which, as Senator John D. Rockefeller IV noted, is "twice the size of the entire intelligence budget of the United States government—all generated by the effort to detail and sell information about our private lives."[65] There are, today, more than 4,000 data broker companies, some of them large publicly-traded corporations and household names like Lexis-Nexis or Experian, and others much smaller and less well known.[66]

A. The Costly Market in Data

Senator Rockefeller held hearings in the U.S. Senate Committee on Commerce, Science, and Transportation on Wednesday, December 18, 2013, examining this data brokerage industry—and throwing some light on their sometimes appalling practices. The hearings disclosed that, for instance, one data broker in Lake Forest, Illinois, Medbase200, offered to sell to pharmaceutical companies a list of "rape sufferers" at a cost of $79 for 1,000 names.[67] Medbase200 marketed this list on its website as follows:

> These rape sufferers are family members who have reported, or have been identified as individuals affected by specific illnesses, conditions or ailments relating to rape. Medbase200 is the owner of this list. Select from families affected by over 500 different ailments, and/or who are consumers of over 200 different

> Rx medications. Lists can be further selected on the basis of lifestyle, ethnicity, geo, gender, and much more. Inquire today for more information.

Medbase200 took the "rape sufferers" database off its website after these revelations, as well as removing "lists of domestic violence victims, HIV/AIDS patients and 'peer pressure sufferers' that it had been offering for sale."[68] But the number and variety of other lists it offers for sale is simply staggering. Here are the lists for sale starting with an A. The size of each database and the price information (in dollars per 1,000 pieces of information) are in the right column:

Allergy/Immunology Nurses	53423 Total Universe @ 59/M
AARP Members Mailing List	20435556 Total Universe @ 79/M
Abscess Sufferers	> (Inquire) Total Universe @ 79.00/M
Abuse Sufferers	> (Inquire) Total Universe @ 79.00/M
Acetaminophen Users	21092445 Total Universe @ 79/M
Achondroplasia Sufferers	> (Inquire) Total Universe @ 79.00/M
Acid Reflux Disease (GERD) Sufferers	> (Inquire) Total Universe @ 79.00/M
Acid Reflux Disease (GERD) Sufferers at Home	5456709 Total Universe @ 79/M
Acne Sufferers	> (Inquire) Total Universe @ 79.00/M
Addiction Sufferers	> (Inquire) Total Universe @ 79.00/M
Addiction/Substance Abuse (Drug Abuse) Nurses	38009 Total Universe @ 59/M
Addison's Disease Sufferers	> (Inquire) Total Universe @ 79.00/M
Adenoma Sufferers	> (Inquire) Total Universe @ 79.00/M
Adolescent Medicine Nurses	20198 Total Universe @ 59/M
Adult Medicine/Adult Care Nurses	98996 Total Universe @ 59/M
Advanced Practice Nurses	92231 Total Universe @ 59/M
Aestheticians at Home	116545 Total Universe @ 59/M
Agoraphobia Sufferers	> (Inquire) Total Universe @ 79.00/M
Aids And Hiv Infection Sufferers	> (Inquire) Total Universe @ 79.00/M
AIDS/HIV Nurses	300893 Total Universe @ 59/M
Ailments, Diseases & Conditions - Hispanic Sufferers	17234554 Total Universe @ 79/M
Ailments, Diseases & Conditions - Sufferers	227453121 Total Universe @ 79/M
Ailments, Diseases & Conditions - Sufferers (Vol)	227453121 Total Universe @ 39.5/M
Ailments, Diseases & Conditions - Sufferers via Email	173209889 Total Universe @ 129/M
Albinism Sufferers	> (Inquire) Total Universe @ 79.00/M
Alcoholic Hepatitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Alcoholism Sufferers	> (Inquire) Total Universe @ 79.00/M
Allergies Sufferers	> (Inquire) Total Universe @ 79.00/M
Allergy / Immunology Nurses	57886 Total Universe @ 59/M
Allergy Sufferers at Home	25698121 Total Universe @ 79/M
Alli Users	1985452 Total Universe @ 79/M
Alopecia (Thinning Hair/Hair Loss) Sufferers	> (Inquire) Total Universe @ 79.00/M
Altitude Sickness Sufferers	> (Inquire) Total Universe @ 79.00/M

Alzheimer's Disease Sufferers	> (Inquire) Total Universe @ 79.00/M
Amblyopia Sufferers	> (Inquire) Total Universe @ 79.00/M
Ambulatory Care Nurses	72234 Total Universe @ 59/M
Amebiasis Sufferers	> (Inquire) Total Universe @ 79.00/M
Amnesia Sufferers	> (Inquire) Total Universe @ 79.00/M
Amyotrophic Lateral Sclerosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Anemia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anesthesiology Nurses	172339 Total Universe @ 59/M
Aneurdu Sufferers	> (Inquire) Total Universe @ 79.00/M
Aneurysm Sufferers	> (Inquire) Total Universe @ 79.00/M
Angina Sufferers	> (Inquire) Total Universe @ 79.00/M
Animal Bites Sufferers	> (Inquire) Total Universe @ 79.00/M
Anorexia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anosmia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anotia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anthrax Sufferers	> (Inquire) Total Universe @ 79.00/M
Antisocial Personality Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
Anxiety And Anxiety Disorders Sufferers	> (Inquire) Total Universe @ 79.00/M
Anxiety Disorders Sufferers	> (Inquire) Total Universe @ 79.00/M
Anxiety Sufferers (GAD) Sufferers at Home	3983434 Total Universe @ 79/M
Appendicitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Apraxia Sufferers	> (Inquire) Total Universe @ 79.00/M
Argyria Sufferers	> (Inquire) Total Universe @ 79.00/M
Arthritis Nurses	180371 Total Universe @ 59/M
Arthritis Sufferers	> (Inquire) Total Universe @ 79.00/M
Arthritis, Infectious Sufferers	> (Inquire) Total Universe @ 79.00/M
Ascariasis Sufferers	> (Inquire) Total Universe @ 79.00/M
Aseptic Meningitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Asperger Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
Asthenia Sufferers	> (Inquire) Total Universe @ 79.00/M
Asthma Sufferers	> (Inquire) Total Universe @ 79.00/M
Astigmatism Sufferers	> (Inquire) Total Universe @ 79.00/M
Atherosclerosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Athetosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Athlete's Foot Sufferers	> (Inquire) Total Universe @ 79.00/M
Atrophy Sufferers	> (Inquire) Total Universe @ 79.00/M
Attention Deficit Hyperactivity Disorder (ADHD) Sufferers	> (Inquire) Total Universe @ 79.00/M
Attention Sufferers	> (Inquire) Total Universe @ 79.00/M
Autism Sufferers	> (Inquire) Total Universe @ 79.00/M
Autism Sufferers at Home	2983342 Total Universe @ 79/M
Avandia Users	6898545 Total Universe @ 79/M

Here are the datasets for sale that start with a B:

Babesiosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Back Pain Sufferers	> (Inquire) Total Universe @ 79.00/M
Bacterial Infections Sufferers	> (Inquire) Total Universe @ 79.00/M
Bacterial Meningitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Bedsores (Pressure Sores) Sufferers	> (Inquire) Total Universe @ 79.00/M

Bedwetting (Enuresis) Sufferers	> (Inquire) Total Universe @ 79.00/M
Bell's Palsy Sufferers	> (Inquire) Total Universe @ 79.00/M
Bends Sufferers	> (Inquire) Total Universe @ 79.00/M
Beriberi Sufferers	> (Inquire) Total Universe @ 79.00/M
Binge Eating Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
Bioterrorism Sufferers	> (Inquire) Total Universe @ 79.00/M
Bipolar Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
BiPolar Disorder Sufferers at Home	1985233 Total Universe @ 79/M
Birth Defects And Brain Development Sufferers	> (Inquire) Total Universe @ 79.00/M
Bites And Stings Sufferers	> (Inquire) Total Universe @ 79.00/M
Bladder Cancer Sufferers	> (Inquire) Total Universe @ 79.00/M
Blindness Sufferers	> (Inquire) Total Universe @ 79.00/M
Body Dysmorphic Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
Body Image Sufferers	> (Inquire) Total Universe @ 79.00/M
Bone Densitometry Nurses	27045 Total Universe @ 59/M
Botulism Sufferers	> (Inquire) Total Universe @ 79.00/M
Brain Injuries Sufferers	> (Inquire) Total Universe @ 79.00/M
Brain Tumor Sufferers	> (Inquire) Total Universe @ 79.00/M
Breast Cancer Sufferers	> (Inquire) Total Universe @ 79.00/M
Broken Bones And Fractures Sufferers	> (Inquire) Total Universe @ 79.00/M
Bronchiolitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Bronchitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Bronchitis, Infectious Sufferers	> (Inquire) Total Universe @ 79.00/M
Brucellosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Bulimia Sufferers	> (Inquire) Total Universe @ 79.00/M
Bullying Sufferers	> (Inquire) Total Universe @ 79.00/M
Bunions Sufferers	> (Inquire) Total Universe @ 79.00/M
Bunions Sufferers	> (Inquire) Total Universe @ 79.00/M
Burn Care Nurses	45814 Total Universe @ 59/M
Burns Sufferers	> (Inquire) Total Universe @ 79.00/M

As *The New York Times* recently reported, InfoUSA, one of the largest data brokers in the country, "advertised lists of 'Elderly Opportunity Seekers,' 3.3 million older people 'looking for ways to make money,' and 'Suffering Seniors,' 4.7 million people with cancer or Alzheimer's disease. 'Oldies but Goodies' contained 500,000 gamblers over 55 years old, for 8.5 cents apiece. One list said: 'These people are gullible. They want to believe that their luck can change.'"[69] As you can imagine, these types of lists are often sold to people who then prey on those listed.

One data broker is called the Acxiom Corporation and referred to, by another *New York Times* reporter, as "the quiet giant of a multibillion-dollar industry known as database marketing."[70] "Few consumers have ever heard of Acxiom," *The Times* reports.

"But analysts say it has amassed the world's largest commercial database on consumers — and that it wants to know much, much more. Its servers process more than 50 trillion data 'transactions' a year. Company executives have said its database contains information about 500 million active consumers worldwide, with about 1,500 data points per person. That includes a majority of adults in the United States."[71] Here's a flavor of what that data broker knows:

> It peers deeper into American life than the F.B.I. or the I.R.S., or those prying digital eyes at Facebook and Google. If you are an American adult, the odds are that it knows things like your age, race, sex, weight, height, marital status, education level, politics, buying habits, household health worries, vacation dreams — and on and on.[72]

Consumer data has become a commodity, its trade a highly competitive market. The going rate for information about individual consumers varies depending on its character, competition between data providers, and the "sheer ubiquity of details about hundreds of millions of consumers."[73] In general, however, "The more intimate the information, the more valuable it is." For instance, Steel writes that "Basic age, gender and location information sells for as little as $0.0005 per person, or $0.50 per thousand people, according to price details seen by the *Financial Times*." But information about individuals "believed to be 'influential' within their social networks sells for $0.00075, or $0.75 per thousand people." Data about "income details and shopping histories" sell for $.001—so, a dollar per thousand.[74] Certain developments in a consumer's life also increase the price of their information:

> Certain milestones in a person's life prompt major changes in buying patterns, whether that's becoming a new parent, moving homes, getting engaged, buying a car, or going through a divorce. Marketers are willing to pay more to reach consumers at those major life events. Knowing that a woman is expecting a baby and is in her second trimester of pregnancy, for instance, sends the price to tag for that information about her to $0.11.[75]

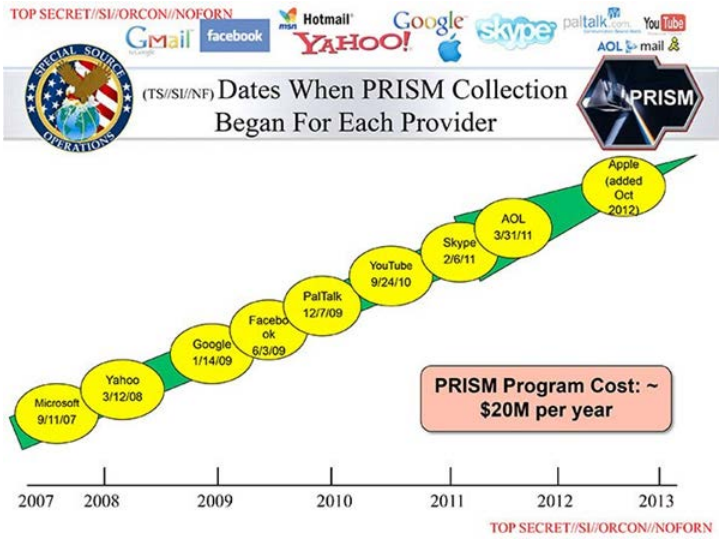
Medical information is worth much more: "For $0.26 per person, LeadsPlease.com sells the names and mailing addresses of people suffering from ailments such as cancer, diabetes and clinical depression. The information includes specific medications including cancer treatment drug Methotrexate and Paxil, the antidepressant, according to price details viewed by the [*Financial Times*]."[76]

Consumer data is used to predict future buying behavior, to classify existing behavior into "predetermined categories"; to associate certain behaviors with others, as when, to borrow an example from a recent article in the *Atlantic*, Amazon makes a recommendation for martini glasses based on the recent purchase of a cocktail shaker; and to form "clusters" of information based on consumer behaviors, as when a group of consumers are separated into their specific hobbies and interests.[77] Consumer surveillance, as Lyons notes, is "The most rapidly growing sphere of surveillance…, outstripping the surveillance capacities of most nation-states. And even within nation-states, administrative surveillance is guided as much by the canons of consumption as those of citizenship, classically construed."[78]

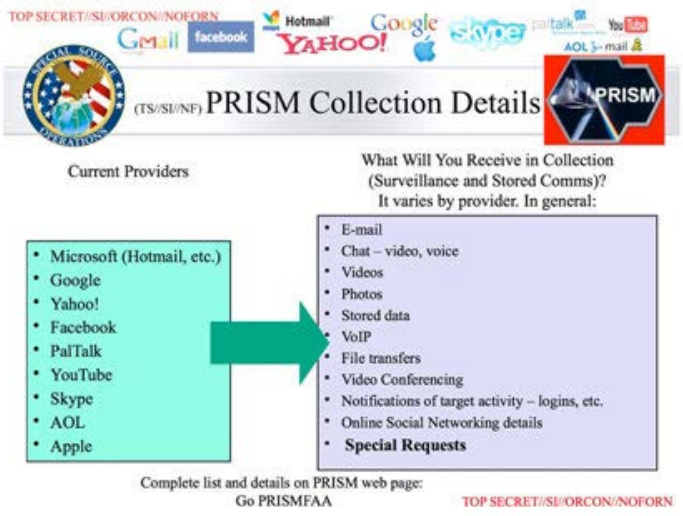
B. The Free Market of Surveillance

There is little that is free about this market for data—except, perhaps, government access to it. One thing that seems pretty costless, or virtually so, is for the government to access the troves of data. This is what, in part, is deconstructing the link between the state and exchange.

The PRISM program is a good example of that. Launched in 2007, the PRISM program allows the NSA to access data from Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, Apple, and more—at a total program cost of only about $20 million per year.[79] This is the Powerpoint slide used by the NSA to describe the history of the PRISM program, as leaked by Edward Snowden and published by the Guardian:



The PRISM program gives the government access to individuals' e-mails, photos, videos, attachments, VoIP, etc. At practically no cost to the government, it provides almost complete access to their digital lives. And although the NSA denies that is has immediate access to the data, the NSA documents leaked by Snowden state that the agency "claims 'collection directly from the servers' of major US service providers."[80] The types of data available are wide-ranging:



As the Guardian reports, "the Prism program allows the intelligence services direct access to the companies' servers. The NSA document notes the operations have 'assistance of communications providers in the US.'"[81] The Guardian explains, based on an NSA document dating to April 2013, that:

> The Prism program allows the NSA, the world's largest surveillance organisation, to obtain targeted communications without having to request them from the service providers and without having to obtain individual court orders.
>
> With this program, the NSA is able to reach directly into the servers of the participating companies and obtain both stored communications as well as perform real-time collection on targeted users.[82]

According to the Guardian, there is no warrant requirement or the need for individual authorization under FISA for any of these collection and analysis activities so long as the analyst searching the communications had "reasonable suspicion that one of the parties was outside the country at the time of the records were collected by the NSA."[83]

The program is apparently leading to exponential rates of growth of search queries. "The document highlights the number of obtained communications increased in 2012 by 248% for Skype – leading the notes to remark there was 'exponential growth in Skype reporting; looks like the word is getting out about our capability against Skype'" the Guardian reports. "There was also a 131% increase in requests for Facebook data, and 63% for Google."[84]

Conclusion

The age of spectacle in the Ancient world was evidence of the cost of publicity. To render something public was expensive, and so the Ancients needed to amass in a common space to watch, to share, to partake together in the act of seeing. The age of surveillance in the nineteenth century, on the other hand, offered evidence of the cost of

security. To surveil and correct individuals was expensive. The invention of new institutions of confinement was driven, in part, by the search for more efficient means of seeing.[85]

We have entered an era, now, of costless publicity. Today, it costs nothing to disseminate information. On the contrary, the information itself is valuable. Most individuals give it away for free. In many cases, we are forced to—we have become our own administrators and spend our time doing all our own "administration."[86] We enter our own personal data on airplane reservation websites, on Zappos.com, on Amazon. And it is precisely this natural resource that has given rise to a new political economy of data—a political economy of publicness.

What follows when our modes of communication themselves and our everyday acts of existence—when our every word spoken, each note taken, every letter written, every photograph and video, even our heartbeat, cholesterol levels, facial features, bank withdrawals, subway swipes, etc.—can be recorded, stored, linked, mined and analyzed by machines with computation powers that so far exceed the human brain? The answer, it seems, is a thriving market in data that makes possible forms of monitoring and surveillance that we never would have imagined—except, perhaps, in Orwellian dystopias.

This new and emerging political economy of data has been made possible not only by technological innovation—though the technology has certainly been at the forefront of all this—but also by a new form of power that has taught us to willingly give our information when asked, to identify ourselves, to reveal our deepest secrets, to comply with requests—and ironically, in a world of private property, to never feel entitled to express private property ownership over our own identity and all this personal information. Here, for instance, is a list of "good" security questions for websites to ask and collect, in order to render user ID and passwords more secure:

What was your childhood nickname?
In what city did you meet your spouse/significant other?
What is the name of your favorite childhood friend?
What street did you live on in third grade?
What is your oldest sibling's birthday month and year? (e.g., January 1900)
What is the middle name of your oldest child?
What is your oldest sibling's middle name?
What school did you attend for sixth grade?
What was your childhood phone number including area code?
What is your oldest cousin's first and last name?
What was the name of your first stuffed animal?
In what city or town did your mother and father meet?
Where were you when you had your first kiss?
What is the first name of the boy or girl that you first kissed?
What was the last name of your third grade teacher?
In what city does your nearest sibling live?
What is your oldest brother's birthday month and year? (e.g., January 1900)
What is your maternal grandmother's maiden name?
In what city or town was your first job?
What is the name of the place your wedding reception was held?
What is the name of a college you applied to but didn't attend?
Where were you when you first heard about 9/11?[87]

If you know my answers to these questions, I suspect, you know more of my secrets than my own partner of 25 years. You know me through and through. Perhaps you know me even better than I can remember myself. The consequences are far-reaching, especially insofar as they break down the conventional boundaries that separate governing from commerce from security.

NOTES

[1] Desmond Butler, Jack Gillum and Alberto Arce, "US secretly created 'Cuban Twitter' to stir unrest," Associated Press, April 4, 2014, available at http://bigstory.ap.org/article/us-secretly-created-cuban-twitter-stir-unrest.
[2] *Id*.
[3] *Id*.
[4] *Id*.
[5] *Id*.
[6] Desmond Butler and Alberto Arce, "US Contractors Profiled 'Cuban Twitter' Responses," Associated Press, April 30, 2014, available at http://abcnews.go.com/International/wireStory/usaid-contractors-profiled-cuban-twitter-users-23532155.
[7] Desmond Butler and Alberto Arce, "US Contractors Profiled 'Cuban Twitter' Responses."
[8] Desmond Butler, Jack Gillum and Alberto Arce, "US secretly created 'Cuban Twitter' to stir unrest."
[9] Desmond Butler, Jack Gillum and Alberto Arce, "US secretly created 'Cuban Twitter' to stir unrest."
[10] Desmond Butler, Jack Gillum and Alberto Arce, "US secretly created 'Cuban Twitter' to stir unrest."
[11] Yasha Levine, "The Psychological Dark Side of Gmail: Google is using its popular Gmail service to build profiles on the hundreds of millions of people who use it," *AlterNet.org*, December 31, 2013 (first appeared in *PandoDaily*), available at http://www.alternet.org/media/google-using-gmail-build-psychological-profiles-hundreds-millions-people.
[12] Yasha Levine, "The Psychological Dark Side of Gmail."
[13] Yasha Levine, "The Psychological Dark Side of Gmail."
[14] Yasha Levine, "The Psychological Dark Side of Gmail."
[15] Yasha Levine, "The Psychological Dark Side of Gmail."
[16] Spencer Ackerman and James Ball, "Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ," *The Guardian*, Thursday 27 February 2014, available at http://www.theguardian.com/world/2014/feb/27/gchq-nsa-webcam-images-internet-yahoo; The Associated Press, "British Spies Intercept Webcam Pictures, Report Says," Feb. 27, 2014, available at *The New York Times* at http://www.nytimes.com/aponline/2014/02/27/world/europe/ap-nsa-surveillance-naked-pictures.html?partner=rss&emc=rss.
[17] Associated Press, "British Spies Intercept Webcam Pictures, Report Says," Feb. 27, 2014, *2.
[18] Associated Press, "British Spies Intercept Webcam Pictures, Report Says," Feb. 27, 2014, *2; *see also* Spencer Ackerman and James Ball , "Optic Nerve," *2.
[19] Spencer Ackerman and James Ball, "Optic Nerve," *2.
[20] Spencer Ackerman and James Ball, "Optic Nerve," *3; Associated Press, "British Spies Intercept Webcam Pictures, Report Says," Feb. 27, 2014, *2.
[21] Associated Press, "British Spies Intercept Webcam Pictures, Report Says," Feb. 27, 2014, *2-3.
[22] Spencer Ackerman and James Ball, "Optic Nerve," *5.
[23] The term is defined in different ways. TechAmerica Foundation's Federal Big Data Commission defines the term as one that "describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." "Demystifying Big Data," at p. 10, available at http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf (2013). McKinsey and Company defines big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." *See* James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity." McKinsey and Company, June, 2011, at p. 1 (available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, accessed May 16, 2014). Another commentator defines it, broadly, as "an umbrella term. It encompasses everything from digital data to health data (including your DNA and genome) to the data collected from years and years of paperwork issued and filed by the government." Higinio Maycotte, "The Evolution of Big Data, and Where We're Headed," *Wired*, March 26, 2014.
[24] TechAmerica Foundation's Federal Big Data Commission, "Demystifying Big Data," at p. 9 and 11.
[25] TechAmerica Foundation's Federal Big Data Commission, "Demystifying Big Data," at p. 9 and 11.

[26] "The Digital Universe of Opportunities." *IDC*, April, 2014. Presentation available at http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf . White paper available at http://idcdocserv.com/1678. Accessed May 15, 2014.
[27] IDC 2014 Presentation at slide 2.
[28] Martin Hilbert and Priscila López, "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* Vol. 332, 60-65 (1 April 2011). Lib.uchicago.edu. Accessed May 15, 2014.
[29] *Id*. at p. 60.
[30] *Id*. at pp. 6-7, Table S A-1.
[31] *Id*. at p. 7, Table S A-2.
[32] *Id*. at p. 8, Table S A-4.
[33] *Id*. at p. 63.
[34] *Id*.
[35] *Id*.
[36] IDC 2014 Presentation at slide 5.
[37] IDC 2014 Presentation at slide 8.
[38] IDC 2014 White Paper at p. 3.
[39] TechAmerica Foundation's Federal Big Data Commission, "Demystifying Big Data," at p. 9, available at http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf.
[40] TechAmerica Foundation's Federal Big Data Commission, "Demystifying Big Data," at p. 11.
[41] *The Economist*, report issued on February 27, 2010.
[42] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey and Company, June, 2011, at p. 16 and 17 (available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, accessed May 16, 2014).
[43] "Big data: The next frontier for innovation, competition, and productivity," McKinsey and Company," preface.
[44] Andrew McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution." *Harvard Business Review*, October, 2012, at p. 62. http://hbr.org/2012/10/big-data-the-management-revolution/ar . Accessed May 15, 2014.
[45] "Big data: The next frontier for innovation, competition, and productivity," McKinsey and Company, p. 19, exhibit 7.
[46] IDC 2014 Presentation at slide 4.
[47] TechAmerica Foundation's Federal Big Data Commission, "Demystifying Big Data," at p. 9.
[48] "Demystifying Big Data," at p. 9.
[49] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey and Company, June, 2011, at p. 19, Exhibit 7.
[50] "Data, data everywhere: A special report on information management." *The Economist*, February 27, 2010. http://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf. Accessed on May 11, 2012.
[51] TechAmerica Foundation's Federal Big Data Commission, "Demystifying Big Data," at p. 11.
[52] Thomas H. Davenport, Paul Barth, and Randy Bean, "How 'Big Data' Is Different," *MITSloan Management Review*, Vol 54, No. 1, Fall 2012: p. 22.
[53] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey and Company, June, 2011, at p. 20, Exhibit 8.
[54] Higinio Maycotte, "The Evolution of Big Data, and Where We're Headed," *Wired*, March 26, 2014. http://www.wired.com/2014/03/evolution-big-data-headed/. Accessed May 16, 2014.
[55] *Id*.
[56] *Id*.
[57] *Id*.
[58] *Id*.

[59] "Big Data FAQs—A Primer." Arcplan Business Intelligence Blog, March 23, 2012. http://www.arcplan.com/en/blog/2012/03/big-data-faqs-a-primer/#more . Accessed May 17, 2014.
[60] "Big Data FAQs—A Primer."
[61] *Id*.
[62] "Big Data, Big Impact: New Possibilities for International Development," The World Economic Forum, 2012, at p. 2. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf . Accessed May 15, 2014.
[63] "Big Data, Big Impact: New Possibilities for International Development," The World Economic Forum, 2012, at p. 2.
[64] *Id*., infographic at p. 5.
[65] Statement of Senator John D. (Jay) Rockefeller IV at the Senate Hearings on Dec 18 2013, "What Information Do Data Brokers Have on Consumers, and How Do They Use It?", U.S. Senate Committee on Commerce, Science, and Transportation (statement available at http://www.commerce.senate.gov/public/index.cfm?p=Hearings&ContentRecord_id=a5c3a62c-68a6-4735-9d18-916bdbbadf01&Statement_id=a47c081a-d653-4272-8d12-d6edc1e04dc6&ContentType_id=14f995b9-dfa5-407a-9d35-56cc7152a7ed&Group_id=b06c39af-e033-4cba-9221-de668ca1978a&MonthDisplay=12&YearDisplay=2013; hearings available at http://www.commerce.senate.gov/public/index.cfm?p=Hearings&ContentRecord_id=a5c3a62c-68a6-4735-9d18-916bdbbadf01&ContentType_id=14f995b9-dfa5-407a-9d35-56cc7152a7ed&Group_id=b06c39af-e033-4cba-9221-de668ca1978a); *see also* Yasha Levine, "You'll Be Shocked at What 'Surveillance Valley' Knows About You: The most intimate details about a person's life, packaged and sold over and over again to anyone willing to pay," *AlterNet.org*, January 8, 2014 (first published at *Pando Daily*), available at http://www.alternet.org/media/what-surveillance-valley-knows-about-you?page=0%2C0.
[66] Yasha Levine, "You'll Be Shocked at What 'Surveillance Valley' Knows About You."
[67] Elizabeth Dwoskin, "Data Broker Removes Rape-Victims List After Journal Inquiry," *Wall Street Journal Digits*, Dec 19, 2013, available at http://blogs.wsj.com/digits/2013/12/19/data-broker-removes-rape-victims-list-after-journal-inquiry/.
[68] Elizabeth Dwoskin, "Data Broker Removes Rape-Victims List After Journal Inquiry."
[69] Charles Duhigg, "Bilking the Elderly, With a Corporate Assist," *The New York Times*, May 20, 2007 available here: http://www.nytimes.com/2007/05/20/business/20tele.html. Accessed February 9, 2014.
[70] Natasha Singer, "Mapping, and Sharing, the Consumer Genome," *The New York Times*, June 16, 2012, available at http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html. Accessed February 9, 2014.
[71] Natasha Singer, "Mapping, and Sharing, the Consumer Genome."
[72] Natasha Singer, "Mapping, and Sharing, the Consumer Genome."
[73] Emily Steel, "Financial worth of data comes in at under a penny a piece." FT.com, June 12, 2013, http://search.proquest.com/docview/1367072051?accountid=14657. Accessed February 9, 2014.
[74] Emily Steel, "Companies in Scramble for Consumer Data." See also Emily Steel, Callum Locke, Emily Cadman and Ben Freese, "How much is your personal data worth?" FT.com, June 12, 2013, http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html#axzz32g6ooslw. Accessed February 9, 2014.
[75] Emily Steel, "Financial worth of data comes in at under a penny a piece."
[76] Emily Steel, "Companies in Scramble for Consumer Data."
[77] Alexander Furnas, "Everything You Wanted to Know About Data Mining but Were Afraid to Ask." *Theatlantic.com*, April 3, 2012, http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388/. Accessed February 16, 2014.
[78] David Lyon, "Surveillance Technology and Surveillance Society," in *Modernity and Technology*, T. Misa, P. Brey, and A. Feenberg, eds. Cambridge, MA: MIT Press, p. 172.
[79] Glenn Greenwald and Ewen MacAskill, "NSA Prism program taps in to user data of Apple, Google and others," *The Guardian*, June 6, 2013, http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. Accessed March 11, 2014.
[80] *Id*.

[81] *Id*.
[82] *Id*.
[83] *Id*.
[84] *Id*.
[85] Foucault develops in detail the movement from spectacle to surveillance in M. Foucault, *La Société punitive* (Gallimard 2013), at pp. 24-25.
[86] Elizabeth Emens at Columbia Law School is developing a marvelous project on this question.
[87] Taken from http://goodsecurityquestions.com/examples.htm.