“Raw Data” Is an Oxymoron

Edited by Lisa Gitelman

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Introduction

Lisa Gitelman and Virginia Jackson

"Raw data" is both an oxymoron and a bad idea.
—Geoffrey C. Bowker, *Memory Practices in the Sciences*

Data are everywhere and piling up in dizzying amounts. Not too long ago storage and transmission media helped people grapple with kilobytes and megabytes, but today's databases and data backbones daily handle not just terabytes but petabytes of information, where *peta-* is a prefix which denotes the unfathomable quantity of a quadrillion, or a thousand trillion. Data are units or morsels of information that in aggregate form the bedrock of modern policy decisions by government and nongovernmental authorities. Data underlie the protocols of public health and medical practice, and data undergird the investment strategies and derivative instruments of finance capital. Data inform what we know about the universe, and they help indicate what is happening to the earth's climate. "Our data isn't just telling us what's going on in the world," IBM advertises; "it's actually telling us where the world is going." The more data the better, by these lights, as long as we can process the accumulating mass. Statisticians are on track to be the next sexy profession in the digital economy, reports the front page of the *New York Times*. "Math majors, rejoice," the newspaper urges in another instance, because businesses are going to need an army of mathematicians as they grapple with increasing mountains of data.[1]

What about the rest of us? What are we to data and data to us? As consumers we tend to celebrate our ability to handle data in association with sophisticated technology. My iPad has 64 gig! My phone is 4G! We don't always know what this means and typically don't know how these devices actually function, but they are "friendly" to users in part according to the ways they empower us to store, manipulate, and transmit data.

Yet if data are somehow subject to us, we are also subject to data, because Google collects so much information on users' interests and behaviors, for instance, and the U.S. National Security Agency mines fiber-optic transmissions for clues about terrorists. Not too long ago it was easier to understand the ways that data was collected about us, first through the institutions and practices of governmentality—the census, the department of motor vehicles, voter registration—and then through the institutions and practices of consumer culture, such as the surveys which told us who we were, the polls which predicted who we'd elect, and the ratings which measured how our attention was being directed. But today things seem different—in degree if not always in kind—now that every click, every move has the potential to count for something, for someone somewhere somehow. Is data about you *yours*, or should it be, now that data collection has become an always-everywhere proposition? Try to spend a day "off the grid" and you'd better leave your credit and debit cards, transit pass, school or work ID, passport, and cell phone at home—basically, anything with a barcode, magnetic strip, RFID, or GPS receiver.[2]

In short, if World War II helped to usher in the era of so-called Big Science, the new millennium has arrived as the era of Big Data.[3] For this reason, we think a book like *"Raw Data" Is an Oxymoron* is particularly timely. Its title may sound like an argument or a thesis, but we want it to work instead as a friendly reminder and a prompt. Despite the ubiquity of the phrase *raw data*—over seventeen million hits on Google as of this writing—we think a few moments of reflection will be enough to see its self-contradiction, to see, as Bowker suggests, that data are always already "cooked" and never entirely "raw." It is unlikely that anyone could disagree, but the truism no more keeps us from valuing data than a similar acknowledgment keeps up from buying jumbo shrimp. The analogy may sound silly, but not as silly as it first appears: just as the economy of shrimp and shrimping has shifted radically in the decades since the birth of industrial aquaculture in the 1970s, so the economy of data has an accelerated recent history. The essays in this volume do not present one argument about that economy, but they do begin to supply a little heretofore-unwritten history for the seismic shift in the contemporary conception and use—the sheer existence—of so much data.

However self-contradicting it may be, the phrase *raw data*—like *jumbo shrimp*—has understandable appeal. At first glance data are apparently before the fact: they are the starting point for what we know, who we are, and how we communicate. This shared sense of starting with data often leads to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself. If we're

not careful, in other words, our zeal for more and more data can become a faith in their neutrality and autonomy, their objectivity. Think of the ways people talk and write about data. Data are familiarly "collected," "entered," "compiled," "stored," "processed," "mined," and "interpreted." Less obvious are the ways in which the final term in this sequence—interpretation—haunts its predecessors. At a certain level the collection and management of data may be said to presuppose interpretation. "Data [do] not just exist," Lev Manovich explains, they have to be "generated."[4] Data need to be imagined *as* data to exist and function as such, and the imagination of data entails an interpretive base.

Here another analogy may be helpful. Like *events* imagined and enunciated against the continuity of time, *data* are imagined and enunciated against the seamlessness of phenomena. We call them up out of an otherwise undifferentiated blur. If events garner a kind of immanence by dint of their collected enunciation, as Hayden White has suggested, so data garner immanence in the circumstances of their imagination.[5] Events produce and are produced by a sense of history, while data produce and are produced by the operations of knowledge production more broadly. Every discipline and disciplinary institution has its own norms and standards for the imagination of data, just as every field has its accepted methodologies and its evolved structures of practice. Together the essays that comprise *"Raw Data" Is an Oxymoron* pursue the imagination of data. They ask how different disciplines have imagined their objects and how different data sets harbor the interpretive structures of their own imagining. What are the histories of data within and across disciplines? How are data variously "cooked" within the varied circumstances of their collection, storage, and transmission? What sorts of conflicts have occurred about the kinds of phenomena that can effectively—can ethically—be "reduced" to data?

Treating data as a matter of disciplines—rather than of computers, for instance—may seem curious at first. The subject of data is bound to alienate students and scholars in disciplines within the humanities particularly. Few literary critics want to think of the poems or novels they read as "data," and for good reason. The skepticism within literary studies about Franco Moretti's "distant reading" approach, which in part reduces literary objects to graphs, maps, and other data visualizations, testifies to the resistance the notion of literature as data might provoke. Similarly, many historians would not like to reduce their subjects to abstract objects useful in the production of knowledge about the past. Their reluctance was evidenced by the hostile reception accorded to cliometrics in the 1960s and it persists today. In some sense, data are precisely *not* the domain of humanistic inquiry. Yet we propose that students and scholars in the humanities do worry about data, broadly speaking, to the extent that they worry about how their

objects of study have been assumed as well as discerned. Don't all questions presuppose or delimit their answers to some degree? Recent work in historical epistemology has challenged the status of the research object, or as Michel Foucault would have it, has raised questions about the boundaries of the archive, about the form, appearance, and regularity of the statements and practices available to us in knowing what we know.[6] When we put our own critical perspectives into historical perspective, we quickly find that there is no stance detached from history, which is to say that there is no persistently objective view.

The conditions of evolving, possessing, and assessing knowledge turn out to be remarkably available to cultural and historical change. The field of science studies has pursued this observation in the greatest detail, and *"Raw Data" Is an Oxymoron* is inspired by science studies while directed beyond it to a broader audience. Evolved over the same decades as other "studies"—like area studies, ethnic studies, cultural and media studies—science studies takes as its object the work of scientists and engineers.[7] The field has helped to confound simplistic dichotomies like theory/practice and science/society in a rich, diverse body of work that, among other things, has explored the situated, material conditions of knowledge production. Looking at the ways scientific knowledge is produced—rather than innocently "discovered," for instance—resembles our project of looking into data or, better, looking *under* data to consider their root assumptions.[8] Inquiries such as these may be seen as contributions toward a critique of objectivity. The point of such a critique—we must quickly emphasize—is not that objectivity is *bad* or that objectivity is mythical. Any such claim must depend, as Lorraine Daston and Peter Galison note, on first achieving a careful understanding of "what objectivity *is*."[9] The point is not how to judge whether objectivity is possible—thumbs up or thumbs down—but how to describe objectivity in the first place. Objectivity is situated and historically specific; it comes from somewhere and is the result of ongoing changes to the conditions of inquiry, conditions that are at once material, social, and ethical.

The very idea of objectivity as the abnegation, neutrality, or irrelevance of the observing self turns out to be of relatively recent vintage. Joanna Picciotto has recently suggested that "the question raised by objectivity is how innocence, traditionally understood to be a state of ignorance, ever came to be associated with epistemological privilege."[10] As a moment in which we can see the emergence of a modern privileging of objectivity, Picciotto nominates "the seventeenth century's conversion of the original subject of innocence, Adam, into a specifically intellectual exemplar. Used to justify

experimental science, an emergent public sphere, and the concept of intellectual labor itself," Adam became emblematic of "a new ideal of estranged and productive observation."[11] This means that Milton's *Paradise Lost* and *Paradise Regain'd* may be as important to the development of experimental science as the invention of the microscope.

The innocent observer has had a long, diverse career. Looking at scientific atlases, not Milton poems, Daston and Galison discern the arrival of a version of objectivity that is mechanical: characterized by the observer's restraint and distinguishable from other versions in which the skill and discernment of the observing self counts for something, such as cases in which knowledgeable observers idealize multiple, idiosyncratic specimens into a single type, or in which practiced diagnosticians exert trained judgment in order to make sense of blurry scans. Mechanical objectivity emerged as a dominant ideal in the sciences only in the middle of the nineteenth century, and it is perhaps simplest to describe it contextually with reference to the development of photography during those same years. When Louis Daguerre, Henry Fox Talbot, and others developed and then popularized the first photographic processes, observers were struck by the apparent displacement of human agency in the production of lifelike images. Fox Talbot's lavish account of his calotype process captures this displacement in its title, *The Pencil of Nature*. No artist necessary. Light itself is enough. Photography is objective.

David Ribes and Steven Jackson (chapter 8) direct attention toward some of the difficulties that mechanical objectivity presents today in scientific practice, when biologists rely upon data collected by remote sensors. But mechanical objectivity was something of a conundrum even in Fox Talbot's day. From the very first, the mechanical objectivity of photography was framed by a counter discourse in which photographers were praised for their ability to capture "inner" or "higher" truths on film. The pencil of nature is not enough. Artists are necessary. Photography is subjective. This isn't a question of *either/or* as much as a matter of *and yes*: mechanical objectivity is an "epistemic virtue" among other competing virtues.[12] The presumptive objectivity of the photographic image, like the presumptive rawness of data, seems necessary somehow—resilient in common parlance, utile in commonsense—but it is not sufficient to the epistemic conditions that attend the uses and potential uses of photography. At the very least the photographic image is always framed, selected out of the profilmic experience in which the photographer stands, points, shoots. Data too need to be understood as framed and framing, understood, that is, according to the uses to which they are and can be put. Indeed, the seemingly indispensable misperception that data are ever

raw seems to be one way in which data are forever contextualized—that is, framed—according to a mythology of their own supposed decontextualization.

Thus the history of objectivity turns out to be inescapably the history of subjectivity, of the self,[13] and something of the same thing must hold for the concept of data. Data require our participation. Data need us. Yet for all of the suggestive parallels, the history of objectivity is not the history of data. Where did the modern concept of data come from? The first two chapters in this volume tackle this question in different ways. In "Data before the Fact" (chapter 1), Daniel Rosenberg plumbs the derivation and use of *datum* (the singular form) and *data*, offering an intellectual history of the concept that stretches back to the Enlightenment, before the virtue of mechanical objectivity had fully taken shape. Rosenberg is aided in his study—if also provoked—by a new set of tools that offer ways to find and visualize patterns within the digitized corpus of Western printed thought. He gives us the data on data, as it were. Travis D. Williams heads even further back in time, to the Renaissance, in order to consider the history behind one of the strongest epistemic conditions shaping the contemporary data imaginary: the self-evidence of numbers and arithmetic fact as such. Previous scholars have rendered the history of math as or relating to a pre-history of capitalism, and Williams's "Procrustean Marxism and Subjective Rigor" (chapter 2) seeks an additional path, giving an account of English math books with their hilariously prosaic story problems. Like Rosenberg's self-conscious use of present tools in rendering the past, Williams is at pains to take early modern math on its own terms while also considering just what such an endeavor means, since the terms of math are supposed to be universal in time and space. Two plus two equals four, always and everywhere, and "Numbers never lie."

No two chapters could exhaust the multiple origins of data as a concept; Rosenberg and Williams only open the question in different ways. The association of data with diagrams and graphs, in the first instance, and with numbers and mathematical functions, in the second, leads us to the general precept that *data are abstract*. While this quality can make it hard to think or write about data in general—that is, in the abstract—it follows from their abstraction that data ironically require material expression. The retention and manipulation of abstractions require stuff, material things. Just as Cambridge University could become a training ground for mathematical physics only after the introduction of written exams at the end of the eighteenth century (paper and pencil are the things of things where modern abstractions are concerned), so the contemporary era of Big Data has been enabled by the widespread availability of electronic storage media, specifically mainframe computers, servers and server farms, and storage

area networks.[14] Both the scale and ontology of electronic storage pose an interesting challenge across the humanities, where lately there has been a renewed interest in *things*.[15] Indeed, as Wendy Hui Kyong Chun has observed, this current scholarly interest in things or "thing theory" needs to be seen against the context of digital media within which things "always seem to be disappearing" in such crucial ways.[16] What sort of things are electronic data, after all?

As we suggested earlier, one productive way to think about data is to ask how different disciplines conceive their objects, or, better, how disciplines and their objects are mutually conceived. The second pair of chapters in this volume takes that tack. In "From Measuring Desire to Quantifying Expectations" (chapter 3), Kevin R. Brine and Mary Poovey address the discipline of economics, and in "Where Is That Moon, Anyway?" (chapter 4), Matthew Stanley considers astronomy. Brine and Poovey follow the work of Irving Fisher, the twentieth-century economist who created the scaffolding for today's financial modeling by linking capital to the concept of present value, which calculates value by taking into account expectations about future yields or benefits. Although the data he used needed to be "scrubbed" to be usable, models like those that Fisher created continue to be influential because they claim a basis that is situated as the objective source of information it can never actually be. As Rosenberg's history helps us understand, this fundamental contradiction may actually be intrinsic to the concept of data, since "the semantic function of data is specifically *rhetorical*." Data by definition are "that which is given prior to argument," given in order to provide a rhetorical basis. (Facts are facts—that is, they are true by dint of being factual—but data can be good or bad, better or worse, incomplete and insufficient.) Yet precisely because data stand as a given, they can be taken to construct a model sufficient unto itself: given certain data, certain conclusions may be proven or argued to follow. Given other data, one would come to different arguments and conclusions.

Disciplines operate according to shared norms, and data scrubbing is an accepted and unexceptional necessity in economics and finance. Disciplines also operate by dint of "data friction"—Paul Edwards's term—friction consisting of worries, questions, and contests that assert or affirm what should count as data, or which data are good and which less reliable, or how big data sets need to be.[17] Stanley's chapter offers a fascinating example of data friction in the field of astronomy. In efforts to derive a particular lunar constant—called the secular acceleration—astronomers have repeatedly engaged in research that on its face seems a lot less like astronomy than it does textual analysis, history, and psychology: poring over the works of classical authors to evaluate their

accounts of solar eclipse. The apparent intrusion of psychology into astronomy, or history into climate science, or bibliography into botany—to mention additional examples recently documented—serves as a reminder of just how diverse and dynamic disciplines are.[18] Disciplines aren't just separate subjects you pick out of a course catalogue. They involve infrastructures comprised of "people, artifacts, and institutions that generate, share, and maintain specific knowledge" in complex and interconnected ways.[19] The bodies of knowledge made and maintained by the professions can be more or less specific than those of academic disciplines, but they involve related infrastructures and a similarly evolved and evolving "trust in numbers."[20]

Data aren't only or always numerical, of course, but they do always exist in number in the sense that data are particulate or "corpuscular, like sand or succotash." Something like information, that is, data exist in little bits.[21] This leads us to a second general precept, that *data are aggregative*. They pile up. They are collected in assortments of individual, homologous data *entries* and are accumulated into larger or smaller data *sets*. This aggregative quality of data helps to lend them their potential power, their rhetorical weight. (More is better, isn't it?) Indeed, data are so aggregative that English usage increasingly makes many into one. The word *data* has become what is called a mass noun, so it can take a singular verb. Sentences that include the phrase "data is . . ." are now roughly four times as common (on the web, at least, and according to Google) as those including "data are . . ." despite countless grammarians out there who will insist that *data* is a plural. So far in this introduction we have been assiduous in using the word *data* with plural verbs, and some readers may already have sensed the strain. Data's odd suspension between the singular and the plural reminds us of what aggregation means. If a central philosophical paradox of the Enlightenment was the relation between the particular and the universal, then the imagination of data marks a way of thinking in which those principles of logic are either deferred or held at bay. The singular *datum* is not the particular in relation to any universal (the elected individual in representative democracy, for example) and the plural *data* is not universal, not generalizable from the singular; it is an aggregation. The power within aggregation is relational, based on potential connections: network, not hierarchy.

To be sure, data also depend upon hierarchy. Part of what distinguishes data from the more general category, information, is their discreetness. Each datum is individual, separate and separable, while still alike in kind to others in its set. It follows that the imagination of data is in some measure always an act of classification, of lumping and splitting, nesting and ranking, though the underlying principles at work can be hard

to recover. Once in place, classification schemes are notoriously difficult to discern and analyze, since "Good, usable systems disappear almost by definition. The easier they are to use, the harder they are to see."[22] This is the provocation animating an important book by Bowker and Susan Leigh Star entitled *Sorting Things Out*. Working with a group of examples—such as classifying causes of death; classifying the labor of healthcare workers; and classifying race in apartheid-era South Africa—Bowker and Star illuminate the ways that classifications function, for good and ill, to underpin the social order. When phenomena are variously reduced to data, they are divided and classified, processes that work to obscure—or *as if* to obscure—ambiguity, conflict, and contradiction.

Today the ubiquitous structures of data aggregation are computational forms called relational databases. Described and developed since 1970, relational databases organize data into separate tables ("relational variables") in such a way that new data and new kinds of data can be added or subtracted without making the earlier arrangement obsolete. Data are effectively made independent of their organization, and users who perform logical operations on the data are thus "protected" from having to know how the data have been organized.[23] The technical and mathematical details are not important here, but imagine sorting a giant stack of paperwork into separate bins. Establishing which and how many bins are appropriate would be your first important task, but it is likely that as you proceed to sort your papers, you will begin to have a nagging sense that different bins are needed, or that some bins should be combined, or that some papers impossibly belong in multiple bins. You may even wind up with an extra bin or two marked "miscellaneous" or "special problems." It is just this sort of tangle that database architecture seeks to obviate while making relational variables (bins) and their data (papers) available to a multiplicity of desirable logical operations, like queries.

The third pair of chapters in this volume, "facts and FACTS" by Ellen Gruber Garvey (chapter 5) and "Paper as Passion" by Markus Krajewski (chapter 6), takes our paperwork metaphor at face value. Each imagines a different prehistory of the database by considering a specific trove of paper. Garvey describes a giant mass of clippings taken from Southern newspapers to document the horrors of slavery in the antebellum United States, while Krajewski describes the enormous file amassed in the twentieth century by the German sociologist and prolific theorist Niklas Luhmann. Two examples could hardly exhaust the possible prehistories of databases—papery and not—which reach at least as far back as early modern note-taking practices and the accompanying sense of what can anachronistically be called "information overload" that together led to giant

compendia with elaborate finding aids.[24] Yet Garvey's example comes from that important moment when the concept of information—close relative of data—finally emerged in something like its present form, as the alienable, abstract contents of an *inform*ative press,[25] while Krajewski's example comes from the equally important moment of systems theory and cybernetics in the second half of the twentieth century.

Garvey's trick, or rather, the trick of the Grimké sisters she writes about, is to fix on an instance where information collected in one locale can take on wholly different meanings in another, as advertisements for runaway slaves become data in the argument against slavery. This is fully remaking the power of the press in the user-dimension, where users may differ in locale if also in their gender, race, and politics. Krajewski by contrast addresses a single user, Niklas Luhmann, who is famous in some quarters for working from his own huge and all-encompassing card index. Author of more than forty books—not a few of them considered "difficult"—Luhmann developed his systems theory, Krajewski suggests, because of, out of, and in collaboration with his card index, a sort of paper machine—a system—for remembering and for generating thought. Papery databases are only metaphorically databases, of course, yet the example of Luhmann's card index helps to clarify the extraordinary generative power that data aggregation can possess while also raising the question of the human or—one must wonder—the posthuman, the human-plus-machine/machine-plus-human hybrids that living with computers make increasingly integral to our understanding.

The final pair of chapters, "Dataveillance and Countervailance" by Rita Raley (chapter 7) and "Data Bite Man" by David Ribes and Steven J. Jackson (chapter 8), pursues the question of data in the present day. Readers will be challenged to think in some detail about the kinds of data being collected about them today, and they will be challenged to consider the difficulties that scientists and policy makers confront when they try to make data useful today and also reusable potentially by others in the future. What are the logics and the ethics of "dataveillance," now that we appear to be moving so rapidly from an era of expanding data resources into an era in which we have become the resource for data collection that vampirically feeds off of our identities, our "likes," and our everyday habits? If while using the Internet we click on a book or a pair of shoes at Amazon.com, or in a box to sign a petition to stop a Congressional bill, or on a link to a porn website, or on a Google Books page or on an online map to find directions, are we making a choice or are we giving Amazon and the federal government and the pornographers (and the security agencies trolling them) and their advertisers ways to guide our choices, calculate our votes, or put us in jail? Both, Raley answers, and

suggests that activist projects that exploit dataveillance—that do not opt out but instead "insist on a near-total inhabitation of the forcible frame"—might stand the best chance of at least offering an immanent critique of the predicament that we have created and now must find a way to inhabit.

Ribes and Jackson address the predicament experienced by today's scientists, who must not only collect and analyze data but also make sure their data remain useable over the life of a research program and beyond, available to readers of resulting publications as well as for potential research in the future. A recent survey confirms that researchers across the sciences are dealing with vast quantities of data (a fifth report generating data sets of 100 gigabytes or more) while at the same time lacking the resources to preserve that data sensibly (four fifths acknowledge insufficient funding for data curation).[26] Ribes and Jackson show the surprising complexities in something as apparently simple as collecting water samples from streams, while they challenge readers to think of scientists and their data as evolved and evolving symbionts, mutually dependent species adapted amid systems ecological and epistemic.

There is much more in the essays collected here than this introduction has mentioned or could encapsulate, and we hope that readers will consider as they read what the ideas are that emerge across the essays as well as what gaps there are among them. One omission, certainly, which this Introduction accentuates with its brief attention to English usage and the history of concepts, is any account of non-Western contexts or intercultural conjunctions that might illuminate and complicate data past and present. How have non-Western cultures arrived at data and allied concepts like information and objectivity? How have non-Western cultures been subject to data, in the project of colonialism, for example, or otherwise? Indeed, how are data putatively raw—and not—in non-Anglophone contexts? Do other languages deploy the food metaphor that English does? Do their speakers semantically align supposedly raw data with supposedly raw text (that is, ASCII) and supposedly raw footage (unedited film or video) the way that English speakers do? How do different languages differently resolve the dilemma of singular and plural? No collection of essays could exhaust the subject of data, of course, and that is one reason we earlier called our title a prompt rather than an argument. The authors collected in *"Raw Data" Is an Oxymoron* all hope to open the question of data, to model some of the ways of thinking about data that seem both interesting and productive, as well as to encourage further discussion. The ethics surrounding the collection and use of today's "Big Data" are a particularly pressing concern.[27]

As an additional gesture toward further discussion, we include a brief section of color images, most of them selected and described by additional contributors. The images in this color insert extend the types of data considered in this volume—some in challenging ways—while some of them also broach the important subject of representation and, more specifically, data visualization, which is not always addressed directly in the chapters that follow but which haunts them nonetheless. As the neologism "dataveillance" suggests, data provide ways to survey the world (the noun *surveillance* is related to *survey*), yet it is important to remember that surveying the world with data at some level means having data visibly before one's eyes, looking *through* the data if not always self-consciously looking *at* the data. There is then a third and final precept closely related to the other two. Not only are data abstract and aggregative, but also *data are mobilized graphically*. That is, in order to be used as part of an explanation or as a basis for argument, data typically require graphical representation and often involve a cascade of representations.[28] Any interface is a data visualization of sorts—think of how many screens you encounter every day—and so are spreadsheets, charts, diagrams, and other graphical forms. Data visualization amplifies the rhetorical function of data, since different visualizations are differently effective, well or poorly designed, and all data sets can be multiply visualized and thereby differently persuasive.

More than a few contemporary visual artists make obvious the rhetoric of data visualization: Jenny Holzer's LED feeds of poems in the place of stock quotes or headlines and "truisms" in the place of public information, for instance, confront spectators with variations on the data frames they face every day. Like the digital network, the database is an already rich and still emerging conceptual field for artwork, while a varied and variously evocative "database aesthetics" demonstrates—as we hope the chapters in this collection make clear—that recognizing the power of data visualization is an important part of living with data.[29]

Notes

1. Geoffrey C. Bowker, *Memory Practices in the Sciences* (Cambridge, MA: MIT Press, 2005), 184. This is an IBM advertising campaign from 2009 to 2010. *New York Times*, August 5, 2009, and May 13, 2011.

2. For more on data obfuscation generally, see Finn Brunton and Helen Nissenbaum, "Vernacular Resistance to Data Collection and Analysis: A Political Theory of Obfuscation," *First Monday* 16, no. 5 (May 2, 2011). The question of whether data about you is yours came before the U.S. courts in the form of a question about privacy: whether the police need a warrant to attach a

GPS device to your car and then monitor your movements. According to *United States v. Jones* (2012), they do.

3. On the bigness of data, see, for instance, Lev Manovich, "Trending: The Promises and the Challenges of Big Social Data," http://lab.softwarestudies.com/2011/04/new-article-by-lev-manovich-trending.html (accessed June 20, 2011). For an example linking big science and big data, see Peter Galison, *Image and Logic: A Material Culture of Microphysics* (Chicago: University of Chicago Press, 1997).

4. Lev Manovich, *The Language of New Media* (Cambridge, MA: MIT Press, 2001), 224.

5. See Hayden White, *Metahistory: The Historical Imagination in Nineteenth-Century Europe* (Baltimore, MD: Johns Hopkins University Press, 1975).

6. Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2005); and Michel Foucault, *The Archaeology of Knowledge & The Discourse on Language* (New York: Vintage, 1982), part 3 (French eds. 1969, 1971).

7. See Mario Biagioli, "Postdisciplinary Liaisons: Science Studies and the Humanities," *Critical Inquiry* 35 (Summer 2009): 816–833; and Mario Biagioli, ed., *The Science Studies Reader* (New York: Routledge, 1999).

8. Looking under is a gesture of "infrastructural inversion" within the sociology of knowledge; see Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge, MA: MIT Press, 1999), 34–36.

9. Lorraine Daston and Peter Galison, *Objectivity* (New York: Zone Books, 2007), 51. On critique itself, see Bruno Latour, "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern," *Critical Inquiry* 30 (Winter 2004): 225–248.

10. Joanna Picciotto, *Labors of Innocence in Early Modern England* (Cambridge, MA: Harvard University Press, 2010), 1.

11. Ibid., 2–3.

12. Daston and Galison, *Objectivity*, 27.

13. Ibid., 37.

14. Andrew Warwick, *Masters of Theory: Cambridge and the Rise of Mathematical Physics* (Chicago: University of Chicago Press, 2003), chap. 3.

15. For instance, Bill Brown, "Thing Theory," *Critical Inquiry* 28 (Autumn 2001): 1–22; Lorraine Daston, ed., *Things That Talk: Object Lessons from Art and Science* (New York: Zone Books, 2004); Lorraine Daston, ed., *Biographies of Scientific Objects* (Chicago: University of Chicago Press, 2000); Hans-Jörg Rheinberger, *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube* (Stanford, CA: Stanford University Press, 1997).

16. Wendy Hui Kyong Chun, *Programmed Visions: Software and Memory* (Cambridge, MA: MIT Press, 2011), 11. "Thing Theory" is Bill Brown's title (see note 15).

17. Paul N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (Cambridge, MA: MIT Press, 2010), xiv.

18. On climate science as a form of history, see Edwards, *A Vast Machine*, xvii; on botany and bibliography, see Lorraine Daston, "Type Specimens and Scientific Memory," *Critical Inquiry* 31 (Autumn 2004): 153–182, esp. 175.

19. See Edwards, *A Vast Machine*, 17.

20. The phrase comes from a title by Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, NJ: Princeton University Press, 1995), which we recommend (along with works already cited) for readers who wish more prolonged exposure to the kinds of questions introduced here.

21. Geoffrey Nunberg, "Farewell to the Information Age," in *The Future of the Book*, ed. Geoffrey Nunberg (Berkeley: University of California Press, 1996), 117.

22. Bowker and Star, *Sorting Things Out*, 33.

23. E. F. Codd, "A Relational Model for Large Shared Data Banks," *Communications of the ACM* 13, no. 6 (June 1970): 377–387. Alan Liu led us to Codd; see his *Local Transcendence: Essays on Postmodern Historicism and the Database* (Chicago: University of Chicago Press, 2008), 239–262.

24. See Ann M. Blair, *Too Much to Know: Managing Scholarly Information Before the Modern Age* (New Haven, CT: Yale University Press, 2010); and Daniel Rosenberg, "Early Modern Information Overload," *Journal of the History of Ideas* 64 (January 2003): 1–9.

25. Nunberg, *Farewell*, 110–111.

26. See "Challenges and Opportunities," *Science* 331 (February 11, 2011): 692–693.

27. See danah boyd and Kate Crawford, "Six Provocations for Big Data," paper presented at "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society," Oxford Internet Institute, September 21, 2011; and Jay Stanley, "Eight Problems with 'Big Data,'" *ACLU.org*, April 25, 2012.

28. On mobilization and cascades, see Bruno Latour, "Drawing Things Together," *Representation in Scientific Practice*, ed. Michael Lynch and Steve Woolgar (Cambridge, MA: MIT Press, 1990), 19–68; on the effectiveness of visualizations, see, for instance, Edward Tufte, *The Visual Display of Quantitative Information*, 2nd ed. (Cheshire, CT: Graphics Press, 2001).

29. For an overview, see, for instance, Victoria Vesna, ed., *Database Aesthetics: Art in the Age of Information Overflow* (Minneapolis: University of Minnesota Press, 2007).

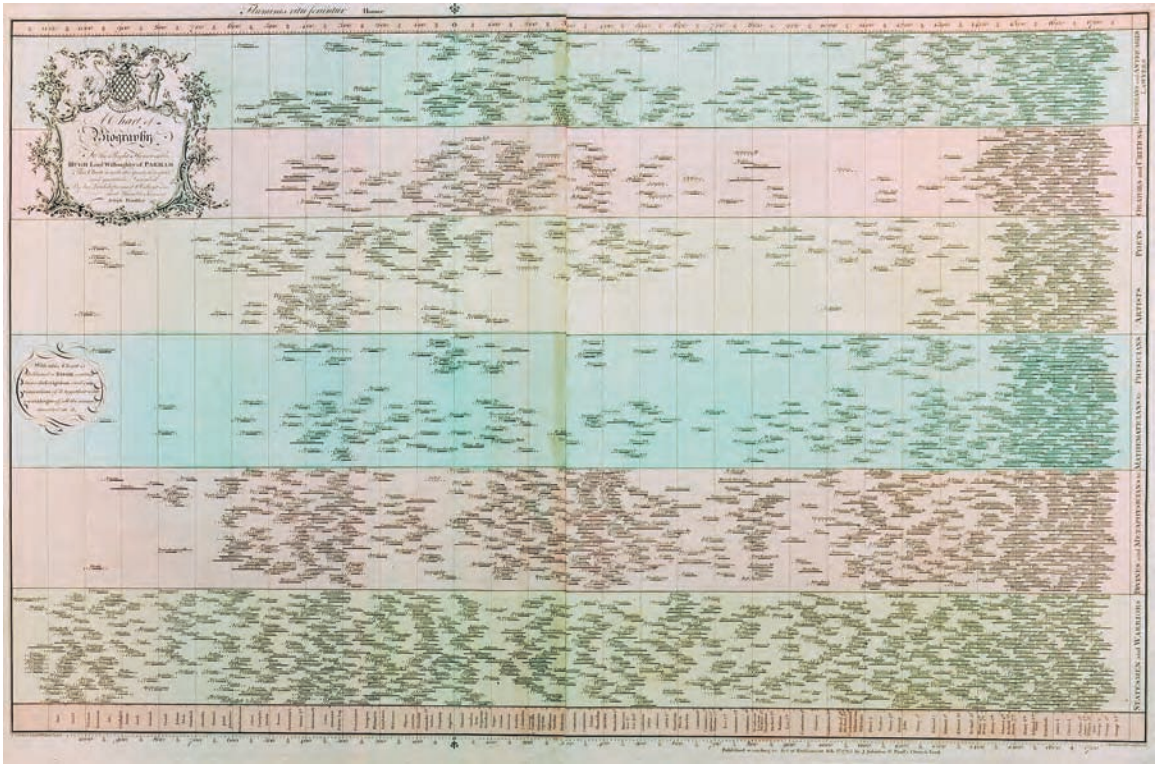
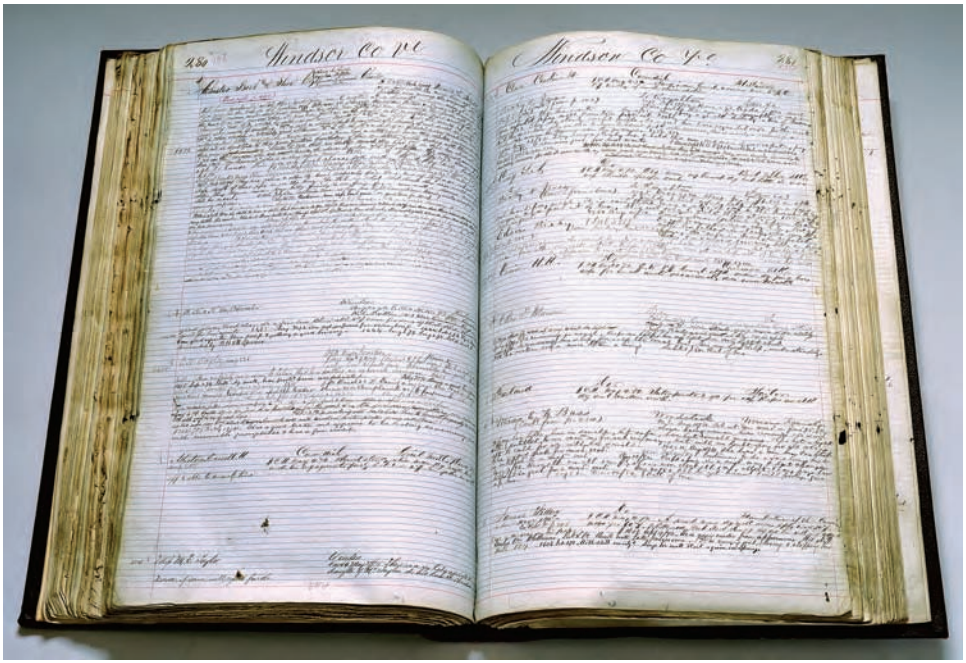


Chart of Biography (1765) The notion that human affairs may be studied through quantitative mechanisms was significantly advanced both in practice and theory during the seventeenth and eighteenth centuries. Theorists and philosophers from William Petty to Jeremy Bentham promoted the power of quantitative research into social and psychological phenomena, while the application of quantitative methods spread by imitation among many domains of research. Joseph Priestley's 1765 *Chart of Biography* is representative of both trends. Priestley is best remembered for his work in chemistry, including his experimental isolation of oxygen in 1772, but he made the *Chart* during his employment as a teacher at the dissenting academy in Warrington, where his duties included teaching history and politics as well as natural philosophy. In his *Chart*, Priestley applied a scientist's intuition to the problem of visualizing historical data. The *Chart of Biography* solved several problems at once. It served as an index to the dates of birth and death of famous historical figures, it clarified the relative position of these lives, and it made visible patterns of achievement in history. Priestley noted that his division of figures into categories of achievement, such as "Statesmen and Warriors" and "Mathematicians and Physicians," revealed that "the world hath never wanted competitors for empire and power, and least of all in those periods in which the sciences and the arts have been the most neglected." Priestley understood his timeline as a heuristic tool capable of making certain phenomena visible while necessarily obscuring others. Few of Priestley's immediate successors shared his reflexivity about the formal and interpretive character of the timeline. (Image courtesy of the Library Company of Philadelphia)

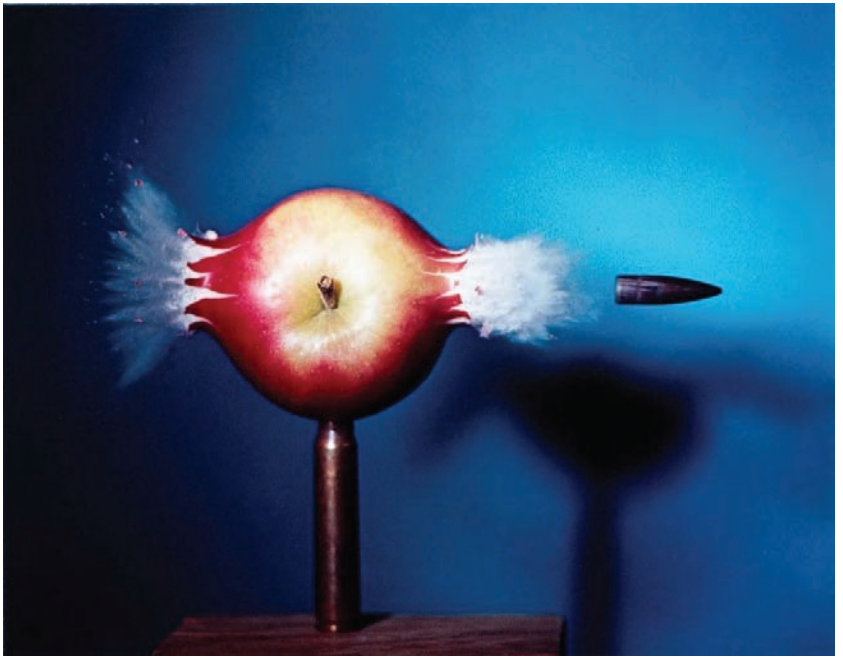
—Daniel Rosenberg



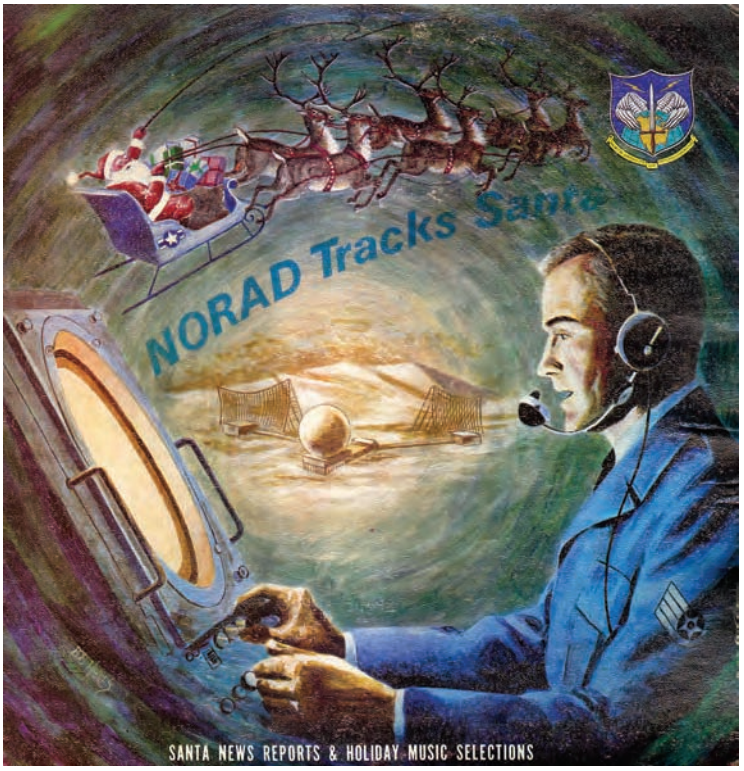
Moral Data Beginning in the 1840s, commercial reporting agencies of Lewis Tappan and R. G. Dun—which together later became Dun and Bradstreet—created networks to gather and evaluate the credit of retailers throughout the United States. Detailing personal events, finances, past performance, and character traits, the companies' anonymous agents compiled histories designed to help predict the behavior of the actors in commercial markets increasingly defined by impersonality and geographic distance. Should individual merchants be trusted to perform obligations they assumed? In claiming to offer credit information that was objective and accurate, reports solicited by the agencies considered matters of property affecting the capacity to pay (bankruptcies, divorces, assets and liabilities), typical of credit reporting in our own day. They also assessed moral capacities—about honesty and the reliability of intentions, about habits of temperance, frugality, and work, in more qualitative, if not impressionist, ways, all of which counted as evidence of success and failure in nineteenth-century America. As clerks inscribed thousands of reports into large folio registers, and made them available to subscribers seeking to hedge the risks they assumed by extending credit to customers, nineteenth-century credit agencies assembled a massive, leather-bound bank of moral data that documented the evolving language in which Americans understood and diagnosed human nature in business. That language was suffused with what Scott Sandage has termed the "folklore of American capitalism," drawing on secular and religious traditions of moral cultivation, concerned with the virtues and vices, as well as Victorian regard for public opinion, for visible signs of respectability, propriety that conveyed one's "character" to strangers. Applying new technologies and systems to the storage, organization, and transmission of local knowledge, the credit agencies developed more efficient tools of information management, and in the twentieth century replaced narrative representations of financial identity with quantitative modes of fortune-telling. (*Image:* Vermont, Vol. 25, R. G. Dun & Co. Credit Report Volumes, Baker Library Historical Collections, Harvard Business School)
—Thomas Augst

"Ascertaining Capacity of Cranial Cavity by Means of Water" (1884) War Department, Surgeon General's Office, United States Medical Museum, Washington, D.C. Here is a quick lesson from the 1880s on how to measure the internal capacity of a human skull. The image captures a scene from the years when an old race science was giving way to a new physical anthropology, when the capacity of skulls might measure racial difference or offer clues to human history. Craniologists had tried beans, buckshot, and sand but worried when no two men measured alike. At the end of the century, water looked promising. Water sometimes seeped into porous bones or dripped through putty plugged into eye sockets, but sure knowledge of hydrostatics and hydraulics made water a good bet as a means to gauge a skull's capacity. You needed a collection of skulls, a beaker, a scale, and a metronome. A thin rubber lining kept water from settling into squamous sutures and leaking out through sinuses. And best of all, as one skull measurer remembered, water guaranteed objectivity, protecting scientists from the temptation to use "muscular exertion" to press a few more beans into a head. (Photography Collection, Miriam and Ira D. Wallach Division of Art, Prints, and Photographs, The New York Public Library, Astor, Lenox, and Tilden Foundations)

—Ann Fabian

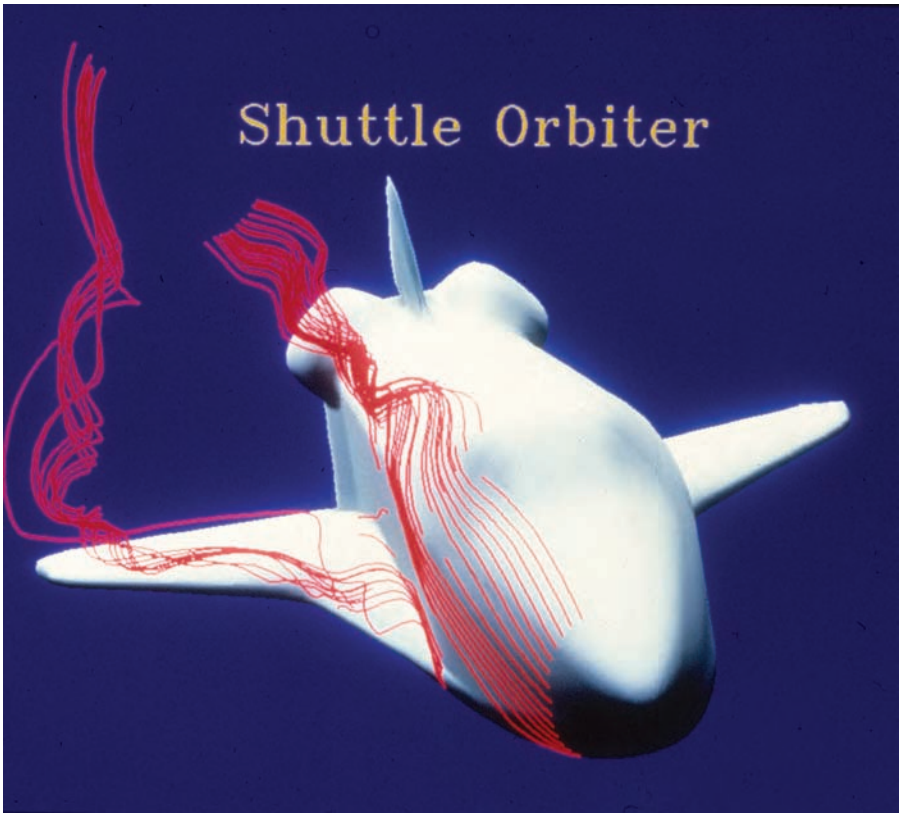


A Bullet through an Apple (1939) How should we interpret photographs of time intervals so small that even a flying bullet could appear perfectly immobile in midair? James R. Killian, a young science writer who would later become president of MIT and one of Eisenhower's must trusted advisors, started his illustrious career by writing about the famous strobe photographs of his colleague and friend Harold E. Edgerton. These photographs, he argued, were raw representations of the natural world. They were a "unique and literal transcription" of nature—a "scientific record" written in a "universal language for all to appreciate." Killian described Edgerton's method as a technique to "contract and expand not only space but time." His strobe was an instrument for "manipulating time as the microscope or telescope manipulates space." From Aristotle to Einstein, most scientists and philosophers felt justified in treating time as space. Although radical thinkers from Hegel to Bergson fought against this space-time conception, an orthodox interpretation of high-speed photographs as *expanding time* coalesced by mid-century. The "instantaneity" of each photograph guaranteed that these images could be studied as temporal and spatial data—easily transformed into mathematical (x, y, z, t) coordinates. But at least one anonymous observer remained skeptical, publishing a humorous critique in *The Electrical Journal* (1931). Remarking on the strobe's alleged ability to stretch time, he titled his commentary: "If Money Could Be Stretched Like That." (Copyright Harold & Esther Edgerton Foundation, 2011, courtesy of Palm Press, Inc.) —Jimena Canales

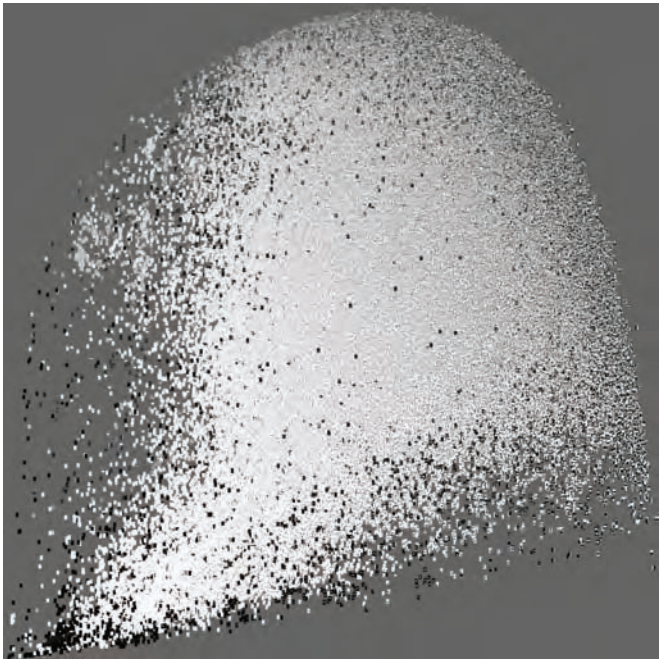


NORAD Santa Tracker (1964) This image, the cover of an album titled NORAD TRACKS SANTA CLAUS, shows a military analyst peering into a radar screen at a North American Aerospace Defense Command facility, feeding data into one of the enormous off-site computers that comprised the SAGE (Semi Automatic Ground Environment) bomber detection system. It is a Cold War image with a twist, however: instead of performing customary surveillance operations, the analyst is tracking the location of Santa's sleigh on Christmas Eve. NORAD released this compilation of radio spots and assorted Christmas tunes to commemorate the ten-year anniversary of a public relations effort that began due to happenstance, but came to serve as the humorous face of an organization whose primary purpose was to monitor U.S. and Canadian airspace in anticipation of nuclear attack. In 1955, a misprint in a Sears advertisement meant that children intending to dial Sears' own Santa hot line instead reached the phone line reserved by NORAD for communication of an impending Soviet missile strikes. Playing along, the organization began to field calls from children interested in Santa's whereabouts, and eventually to issue brief broadcast "updates" of Santa's location, claiming to use NORAD's "satellites, high-powered radars, and jetfighters" to track Santa's journey. The radio broadcasts continued until 1997, when the Santa Tracker moved to the Internet. In 2007, Google partnered with NORAD on the endeavor, creating 2D Google maps and 3D Google Earth images based on NORAD's tracking data. In 2011 the Santa Tracking program drew on over 1,000 U.S. and Canadian military volunteers to field over 100,000 phone calls and emails; the Apple/Android Santa-Tracker smartphone app was downloaded 1.4 million times, while the NORAD Santa Tracker Web site received 2.2 million hits. (*Image:* Bob Haynes)
—Lisa Lynch

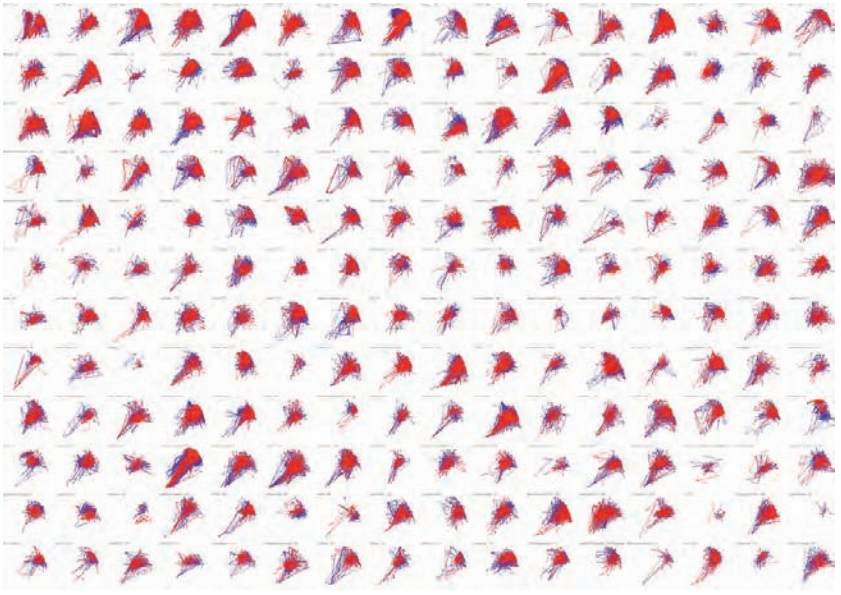
Rumsfeld, Ford, and Cheney (1974) Missing minutes of secret audio recording and other intrigues and malfeasance revealed during the Watergate scandal in the United States led to pressure for greater openness. In 1974 Congress passed a toothsome amendment to the 1966 Freedom of Information Act, but it was vetoed by President Gerald Ford. Ford vetoed the bill at the urging of his chief of staff, Donald Rumsfeld, and his deputy, Richard Cheney, who consulted with a government lawyer, Antonin Scalia ("Veto Battle 30 Years Ago," National Security Archive, nsarchive.org, November 23, 2004). Congress handily overrode Ford's veto. Responding to additional public concern about computer databases, Congress also passed the Privacy Act of 1974, which requires federal agencies to inform the public about the systems of records they use at the same time that it establishes rules for the protection of information that makes individuals identifiable. Both gestures by Congress helped to initiate the information regime in which Americans live and which has been structured further by an extended sequence of laws of fluctuating stricture and enforcement (and, in some cases, evasion) that govern privacy and the retention or destruction of records both private and public. (*Image:* Courtesy of Gerald R. Ford Library)
—Lisa Gitelman



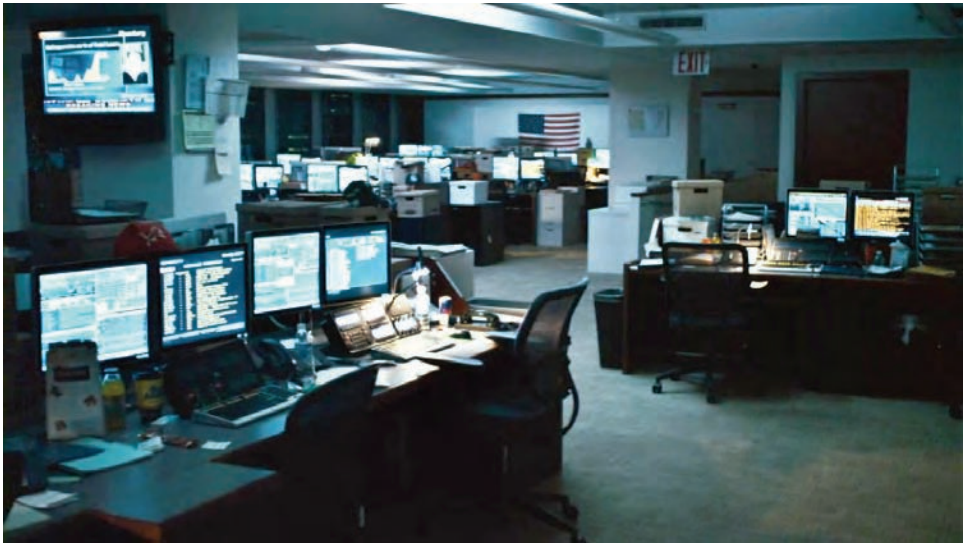
Aerodynamics (ca. 1980) When the U.S. Space Shuttle was under development in the 1970s, its designers faced a number of challenges. Modeling the flow of air over the craft's wings, traditionally done using wind tunnels, was especially difficult, as the Shuttle operated over a wide range of velocities as it returned to Earth from space. Wind tunnels always involved compromises because it was impractical to replicate exactly the conditions of actual flight. Translating data obtained from a model in a tunnel into data about the flying qualities of the actual aircraft was a complex process. Shuttle designers were able to use a new tool that had just become became available: the supercomputer—a digital computer optimized for very fast numerical calculations. Usually associated with the work of computer engineer Seymour Cray, the new computers were offered first from the Control Data Corporation in the late 1960s and later on from Cray Research, a company Seymour Cray founded after leaving Control Data in the early 1970s. Supercomputers created a "virtual" wind tunnel by dividing the region around the aircraft into a grid; assigning numbers corresponding to pressure, temperature, velocity, and so on to each point on that grid; and using equations of aerodynamics to compute those values at the next step in time. To aid in analysis, these numbers were then rendered graphically in false-color, replicating the streams of smoke that were used in traditional tunnels. The image reproduced here is from the NASA Ames Research Center, in Mountain View, California, ca. 1980. It shows a supercomputer-generated image of air flowing over the right side of the Shuttle fuselage and wing. (*Credit:* NASA-Ames Research Center)
—Paul Ceruzzi



Digital Image Analysis: Manga Style Space (2010) Images pose particular challenges for computational analysis. In 2009, we downloaded 883 Manga series containing 1,074,790 unique pages and then used our custom software system installed on a supercomputer at National Department of Energy Research Center (NERSC) to analyze their visual features, turning style into data. This visualization maps all of the pages according to their grayscale measurements, plotting the standard deviation of pixels' grayscale values in a page (x-axis) against the entropy measured over grayscale values in a page (y-axis). The pages in the bottom part of the visualization are the most graphic and have the least amount of detail. The pages in the upper right have lots of detail and texture. The pages with the highest contrast are on the right, while pages with the least contrast are on the left. Among these four extremes, we find every possible graphic variation. This suggests that our basic concept of "style" maybe not appropriate when we consider large cultural data sets. The concept assumes that we can partition a set of works into a number of discrete categories. However, the space of manga graphical variations does not have any distinct clusters, so if we try to divide this space into discrete categories, any such attempt will be arbitrary.



Digital Image Analysis: Manga Style Matrix (2009–2010) We can use the same method to visualize the space of graphical variations in individual Manga series. Some series have been relatively short lived while the longest running began in 1976; the most popular Manga series contains over 10,000 pages to date. This visualization shows 192 different Manga series, each rendered as a separate scatter plot in which pages are represented by points. As in the previous visualization, the position of every point is determined by the corresponding page's visual characteristics, measured by software. (The points in the bottom part of each plot correspond to pages that are more graphic, and contain little detail, while the points in the upper right of a plot correspond to pages with lots of detail and texture.) Page order within each series is represented by color, using a blue-red gradient (pure blue—first page; pure red—last page). This mapping of page order in a series into color creates distinct visual patterns, which indicate whether visual language in a given series changes over the period of its publication. A scatter plot matrix is very useful for working with large cultural data sets. It allows us to quickly see which artifacts in a set stand out from the rest and should be investigated more closely.
—Lev Manovich, Jeremy Douglass, and William Huber



Visualizing the Financial Markets (2011) First sold in 1982, the Bloomberg terminal can now be found on the desks of over 300,000 subscribers around the globe, including at investment banks, hedge funds, government agencies, and even the Vatican. A dedicated portal used to access a vast, proprietary suite of data, tools, and news, the Bloomberg terminal combines real-time quotes from a diverse array of capital and product markets—ranging from the familiar to the esoteric—with a historical repository stretching back decades, all on a single platform. As the devices have become ubiquitous in the financial sector, they have become increasingly essential for anyone hoping to understand, monitor, analyze, and participate in the modern economy. Market participants without them, or away from their desks, find themselves falling entire minutes, seconds, and nanoseconds behind. Like the telegraph and ticker before it, the Bloomberg terminal is credited with market effects. Through the process of aggregating, disseminating, and contextualizing data—modeling markets—the Bloomberg terminal actively shapes the decisions that investors make, thereby confounding the causality between principal and agent as well as human and machine. (Director of photography: Frankie DeMarco; reproduced courtesy of Roadside Attractions/Lionsgate)
—Vikas Mouli