Big Data Is the Answer . . . But What Is the Question?

by Bruno J. Strasser and Paul N. Edwards§*

ABSTRACT

Rethinking histories of data requires not only better answers to existing questions, but also better questions. We suggest eight such questions here. What counts as data? How are objects related to data? What are digital data? What makes data measurable, and what does quantification do to data? What counts as an "information age"? Why do we keep data, and how do we decide which data to lose or forget? Who owns data, and who uses them? Finally, how does "Big Data" transform the geography of science? Each question is a provocation to reconsider the meanings and uses of "data" not only in the past but in the present as well.

Unquestionably the most vocal exponent of epistemic hierarchy, the American writer, performer, and all-around wild man Frank Zappa frequently declaimed on stage that "information is not knowledge, knowledge is not wisdom . . . and music is the best."[1] Zappa was just one among many to adopt the "DIKW" (data, information, knowledge, wisdom) hierarchy. Long before rock 'n' roll, T. S. Eliot may have been DIKW's first exponent: "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" he asked, in *The Rock* (1934).[2] But whereas early authors focused on the relationships between information, knowledge, and wisdom (and music), since the late 1980s "data" has taken center stage, especially after philosopher of science and systems researcher Russell L. Ackoff published his short article "From Data to Wisdom."[3]

The contributions in this volume offer a unique opportunity not only to rethink the meanings of "data," today and in the past, but also to consider why data has become

* Section of Biology, University of Geneva, CH-1211 Geneva 4, Switzerland, and Section of the History of Medicine, Yale University School of Medicine, New Haven, CT 06510; bruno.strasser@unige.ch.

§ Center for International Security and Cooperation, Stanford University, C-226 Encina Hall, Stanford, CA 94309, and School of Information, University of Michigan, 105 South State Street, Ann Arbor, MI 48109; pedwards@stanford.edu.

We would like to thank all the participants at the two workshops on "Historicizing Big Data" at the Max Planck Institute for the History of Science, especially the organizers, Elena Aronova, Christine von Oertzen, and David Sepkoski. We also thank Jérôme Baudry, Dana Mahr, and two anonymous reviewers for helpful comments and suggestions.

[1] Frank Zappa, "Packard Goose," *Joe's Garage: Act III*, 1978.

[2] Jennifer Rowley, "The Wisdom Hierarchy: Representations of the DIKW Hierarchy," *J. Inform. Sci.* 33 (2007): 163–80.

[3] Russell L. Ackoff, "From Data to Wisdom," *J. Appl. Syst. Analysis* 16 (1989): 3–9.

so central to contemporary discussions about the production of knowledge. A genuine rethinking implies not only better answers to existing questions, but also better questions (including those proposed in this volume's introduction). Here we suggest eight further ways to reframe the discussion about data, especially "Big Data," in its historical contexts.

1. WHAT COUNTS AS DATA?

What are data? This obviously fundamental question recurs in virtually all discussions that attempt to historicize Big Data (including many of those in this volume).[4] One answer—also recurrent—is that the category of data describes something basic, foundational, elemental, or *Ur-* (the convenient German prefix). Labeling something "data" usually serves the rhetorical purpose of qualifying some trace or sample as objective, precognitive, unanalyzed, a sure foundation for knowledge claims. The traction of the expression "raw data" comes precisely from this claim that data are unaltered, uncooked, unprocessed by human subjectivity, and thus beyond doubt.[5] Labeling something "data" carries epistemic weight. Dropping the phrase "I have data" in a conversation often seems sufficient to end debate.

More interesting than searching for an ontological conception of data, as if a datum were an atom of knowledge (now that even atoms are no longer atoms), is to ask how *what counts as data* has changed over time. In crystallography, for example, x-rays diffracted by a crystal produce an image containing dark spots, whose intensities are used to calculate "structure factors," which in turn are used to determine the coordinates of each atom composing the crystal. But where are the data? Crystallographers were first content to publish atomic coordinates as the "data" supporting a proposed structure, before they were asked to provide more foundational data, the "structure factors," and eventually yet more foundational data, the original diffraction images.[6] To take another example, since the 1980s climate scientists have referred not only to recorded instrument readings, but also to simulation model outputs as "data," full stop. Today the volume of data from simulation models vastly exceeds that of the entire historical instrument record. Furthermore, data from simulation models may be created to represent either the real past, including the distant past of paleoclimates, or an experimental past (an Earth with different continental positions, different atmospheric composition, etc.), or possible planetary futures.[7]

Thus "data" is not a natural kind, but a property attributed at a given moment in history to a set of signs, traces, indices ("inscriptions," if one insists on Latourian language). As philosopher Sabina Leonelli writes, data are "fungible objects defined by their portability and prospective usefulness as evidence."[8] To attach the label "data" to something is to place that thing specifically in the long chain of transformations that

[4] Rob Kitchin, *The Data Revolution* (London, 2014).

[5] On the critique of "raw data," see Lisa Gitelman, ed., *Raw Data Is an Oxymoron* (Cambridge, Mass., 2013). On the rhetorical power of "data," see Daniel Rosenberg's chapter in that volume, "Data before the Fact," 15–40.

[6] Bruno J. Strasser, "The 'Data Deluge': Turning Private Data into Public Archives," in *Science in the Archives: Pasts, Presents, Futures*, ed. Lorraine Daston (Chicago, 2017), 185–202.

[7] Paul N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (Cambridge, Mass., 2010).

[8] Sabina Leonelli, "What Counts as Scientific Data? A Relational Framework," *Phil. Sci.* 82 (2015): 810–21.

moves from nature to knowledge;[9] this act of categorization marks a particular moment in time when someone thought some inscription or object could serve to ground a knowledge claim. At any such moment, "data" is the closest thing to nature that is no longer "natural"—or, to put it differently, data is the first transformation of nature in the production chain that culminates in knowledge. Thus seeing "data" almost as an adjective, as a *relational* property (like being the youngest child in a family), makes apparent why what counts as data changes over time: as the frontier between nature and knowledge evolves, so do the data that inhabit this moving frontier.

2. HOW ARE OBJECTS RELATED TO DATA?

Less often discussed are questions about the materiality of data. A revealing opening move can be to ask, *What are data made of?* From texts and numbers carved in stone to digital bits stored as electromagnetic charges on silicon chips, all data without exception have a material aspect. The rapidly rising energy consumption of "cloud" servers around the world (currently equal to the output of more than fifty average-sized nuclear power plants; i.e., twelve percent of the world's nuclear electricity) serves as a useful reminder that data can exist only within a material infrastructure.[10]

Thinking about data as objects can focus attention on "data friction": the forms of resistance data offer to circulation and to transformation, and the means developed to smooth their path.[11] A key goal of the 1957–58 International Geophysical Year (IGY) was to share the billions of data records, such as meteorological observations and seismographs, collected around the world. In principle, the IGY's World Data Centers were supposed to provide IGY data on request—but how? By the time the IGY took place, many scientists already saw computers as the future of data processing, but computer technology was young, and few standards for computer-readable records had been established. Rather than decree such a standard themselves, IGY planners chose to distribute them on printed "microcards" or microfilm, instead of on punch cards or magnetic tapes, the two major computer media of the day, as Elena Aronova (this volume) mentions in her discussion of IGY data regimes.[12] This fateful choice forced anyone wishing not just to examine, but also to use IGY data to first transcribe the microcard or microfilm records onto computer-readable media. This created a substantial barrier to data reuse. Computer-readable data are, of course, hardly immune to data friction. Everyday examples abound, while large-scale friction in computerized data can be glimpsed in such cases as the periodic "migration" (to new tapes) of the National Oceanic and Atmospheric Administration's vast libraries of data tapes. Without the migration, data would be lost to the rapid technological obsolescence of tape readers and the physical decay of the underlying media.[13] In these examples, we see

[9] Bruno Latour, *Pandora's Hope: Essays on the Reality of Science Studies* (Boston, 1999), chap. 2.

[10] Jonathan G. Koomey estimated the worldwide server electricity consumption at 271.8 terawatt hours (TWh) in 2010. World nuclear power plant electricity production was 2,364 TWh in 2014. Koomey, *Growth in Data Center Electricity Use 2005 to 2010* (El Dorado Hills, Calif., 2011).

[11] The concept of data friction is developed at length in Edwards, *A Vast Machine* (cit. n. 7).

[12] Elena Aronova, "Geophysical Datascapes of the Cold War: Politics and Practices of the World Data Centers in the 1950s and 1960s," in this volume.

[13] M. Halem, F. Shaffer, N. Palm, E. Salmon, S. Raghavan, and L. Kempster, "Technology Assessment of High Capacity Data Storage Systems: Can We Avoid a Data Survivability Crisis?," in *Government Information Technology Issues 1999: A View to the Future* (Washington, D.C., 1999).

that data-as-objects—even electronic ones—have mass and momentum. Their material characteristics strongly shape the cost and physical possibilities of storage, retrieval, and use, as most of the papers in this volume demonstrate in one way or another. Even today, in the age of cloud storage, when scientists need to move the largest datasets to another physical location, they put them on disk drives and ship them off by mail. The data are then delivered to their destination the old-fashioned way—by a mail carrier on foot.

What happens if, instead of asking after data-as-objects, we shift to the relational view discussed above, asking, *How are objects related to data?* We say, without thinking, that we "collect" or "assemble" data, as if they were shells on the beach or a drawer full of random Lego pieces. Such locutions almost certainly descend from practices of collecting objects for scientific analysis. Historically, collections of objects in a single place created unique opportunities for classification, comparison, and analysis. Indeed, it would be difficult to overstate the centrality to science of "collecting" and "collections." They lie at the core of modern knowledge institutions, from the all-important library to museums to laboratories.[14]

Collected objects such as tissue samples, plant specimens, rocks, and molecules *become* data by being brought into a collection, that is, into relationships with other objects and with a knowledge institution that considers them to represent nature, to be a "second nature." Their material characteristics matter (so to speak) to their lives as data. Mirjam Brusius's chapter in this volume details the complexity of transforming fragments found at an archeological site into reconstructed artifacts, a process that involved extensive detours into mapping, drawing, and cataloging, thereby arousing tensions over the status of these derivative forms of data.[15] To take a related but different example, specimens have carefully defined locations in a natural history museum; the museum itself exists to solve the problem of locating, comparing, storing, and preserving them—not a trivial issue given that (for example) many biological species are known only from a single specimen.[16] In such museums, specimens are physically moved around in order to compare them visually or examine them with special instruments. Transporting specimens between physically distant museums may sometimes be necessary, but in many places and times it has also been perilous and expensive. As a result, practices such as scientific and medical illustration, taking plaster casts of specimens, scale model making, and taxidermy developed to obviate this problem: a simulacrum, model, or representation of a specimen could be moved without risking the destruction or loss of the original. An ever-increasing effort to transform objects-as-data into more portable data-as-objects, including physical simulacra, images, texts, numbers, and digital bits, has been fundamental to the history of science.[17] Crucially, these data-as-objects need not necessarily dwell in just one place; reproducible representations (including electronic ones) can be here *and* be there, allowing the simultaneous and coordinated production of knowledge in many places at once.[18]

[14] Bruno J. Strasser, "Collecting Nature: Practices, Styles, and Narratives," *Osiris* 27 (2012): 303–40.

[15] Mirjam Brusius, "The Field in the Museum: Puzzling Out Babylon in Berlin," in this volume.

[16] Geoffrey C. Bowker, "Biodiversity Datadiversity," *Soc. Stud. Sci.* 30 (2000): 643–83; Lorraine Daston, "Type Specimens and Scientific Memory," *Crit. Inq.* 31 (2004): 153–82.

[17] See, e.g., Martin Rudwick, "George Cuvier's Paper Museum of Fossil Bones," *Arch. Natur. Hist.* 27 (2000): 51–68.

[18] David Weinberger, *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room* (New York, 2011).

Data-as-objects can also, of course, be collected, arranged, compared, and analyzed. Historically, this has been the role of librarians and archivists, for whom the materiality, weight, and volume of information have always been foreground concerns.[19] The ease and low cost of copying and distributing digital information has far-reaching consequences for assembling and accessing collections—or, we might say, metacollections—of data, and thus to participation in the production of knowledge. So the question of data's material infrastructures can be reformulated, at least in part, as *What do collections do to (and for) data?* How do they allow data to circulate, or limit their circulation? What kinds of processing do they require? How do they create relationships among individual data items? Keeping their Aristotelian duality in mind—data are both form and matter—can reveal both similarities and differences among collections and data practices across time. Where data (and in-*form*-ation) are conceptualized primarily as abstract forms (texts, numbers, sequences, etc.), historians should ask after their material basis, what we have called "data-as-objects." Where data are conceptualized primarily as material objects (specimens, bones, rocks, etc.), historians can query the structures and practices that transform these objects into data, what we have called "objects-as-data."

3. WHAT ARE DIGITAL DATA?

Today we refer to almost anything computers can process as "digital," but this is a relatively recent convention. Through the 1990s, the terms "electronic" or "computerized" were more common, and even in the early 2000s, "electronic" was a viable alternative to "digital" (e.g., "e-science"). Today's designation of computable electronic formats as "digital" proves problematic for historians of data, because it implies a temporal divide between pre- and postcomputer data. That divide is very real for some kinds of data but not for others: hence our question, *What are* "*digital*" *data?*

The root meaning of "digital" refers not to electronic but to numerical (or more loosely, symbolic) representations based on a finite set of discrete elements. In this sense, all scientific data originally recorded as numbers or text, whether in lab notebooks, printed books and articles, or other media, were already digital. And indeed, preelectronic digital data storage and processing technologies, especially punch cards, were robust and widely used. Many of the first computerized databases were simply electronic versions of data collections originally stored on cards or microfilm.[20] Long before World War II, libraries at weather centers around the world already held tens of millions of punch cards containing weather data. By 1960, the data library at the U.S. National Weather Records Center contained over 400 million cards, so many that administrators feared the building might collapse under their weight. Transcribed onto magnetic tapes, these punch card libraries became the first electronic databases for global climatology. Similarly, the National Library of Medicine created MEDLARDS in 1964 as the largest publicly accessible electronic database. MEDLARDS was essentially a digitized version of the Index Medicus, a printed bibliographic index to medical literature whose advent dated to 1879 (MEDLARDS eventually became today's enormous PubMed). Finally, the *Atlas of Protein Sequences and Structures*,

[19] Michael K. Buckland, "Information as Thing," *J. Amer. Soc. Inform. Sci.* 42 (1991): 351–60.
[20] Markus Krajewski, *Paper Machines: About Cards and Catalogs, 1548–1929* (Cambridge, Mass., 2011); Lisa Gitelman, *Paper Knowledge* (Durham, N.C., 2014).

launched in 1965 as the first electronic database of protein sequence data, was a computerized version of its author's collection of printed sequences. Thus, although databases seem synonymous with electronic, computerized collections, many of them were in fact born long before the computer age (as David Sepkoski reminds us in this volume for the case of paleontology).[21] In other words, many early electronic databases began as straightforward transcriptions of older collections. Any data that already existed in the form of numbers, or indeed of discrete symbols of any kind (such as texts), had to be changed into an electronic form for the computer to use it. In these cases, the transition from paper, microfilm, and so on, to magnetic tape, disk, and so on, was merely a change of substrate: they became electronic, but they were already born digital in the most direct sense of the term. Of course, inscriptions often have many properties other than the symbols they contain, and these sometimes matter. Paul Duguid recounts the story of a researcher he once encountered in an archive, not just reading old documents but also sniffing them. By detecting the smell of vinegar, used in the nineteenth century as a disinfectant against cholera, he could date and sometimes place an epidemic.[22]

The question, *What are digital data?* can also help us reflect on the more significant digital data revolution, namely, analog-to-digital conversion. This data transition was far more difficult and took much longer than the mere transcription of numbers and text, which were (as we have been saying) in an important sense already digital. Scientists, programmers, and engineers had to figure out how to represent continuous signals such as waveforms, images, video, and sound in discrete (digital) forms, as well as how to analyze and manipulate such analog data with a digital computer. This required a revolution in numerical methods and discrete mathematics, which began in the 1940s and was substantially complete by the late 1960s, though refinements continue to this day.[23] The transition from analog to digital data and computing brought innovations in analog-digital conversion. In the 1960s and 1970s, for example, the isolines of hand-drawn synoptic maps used in weather forecasting were digitized for computer processing, requiring a human operator to manually follow the map's grid and enter (via a keyboard) the value of the nearest isoline, using such devices as the Bendix Datagrid Digitizer (figs. 1, 2). A similar system was used to digitize early Landsat photographs.[24]

In the same period, digital sensors designed to produce numerical outputs directly began to replace analog instruments that required human interpretation, for example, reading off numerical values from the scales marked on mercury-column thermometers, or estimating the intensities of x-ray diffraction spots from a crystallographic image. Recorded on magnetic tapes, these newly digitized data could be more readily stored, reproduced, and transported, and of course—the ultimate payoff—much more

[21] David Sepkoski, "The Database before the Computer?," in this volume.

[22] John Seely Brown and Paul Duguid, *The Social Life of Information* (Boston, 2000).

[23] William Aspray, *John Von Neumann and the Origins of Modern Computing* (Cambridge, Mass., 1990); Paul N. Edwards, *The Closed World: Computers and the Politics of Discourse in Cold War America* (Cambridge, Mass., 1996), chaps. 2, 3; Michael R. Williams, *A History of Computing Technology* (Englewood Cliffs, N.J., 1985); Joseph A November, *Biomedical Computing: Digitizing Life in the United States* (Baltimore, 2012), chap. 2.

[24] R. H. Rogers, C. L. Wilson, L. E. Reed, N. J. Shah, R. Akeley, T. G. Mara, and V. Elliott Smith, "Environmental Monitoring from Spacecraft Data," Laboratory for Applications of Remote Sensing Symposia, Purdue University, Paper 76, 1975, http://docs.lib.purdue.edu/lars_symp/76 (accessed 20 April 2017).

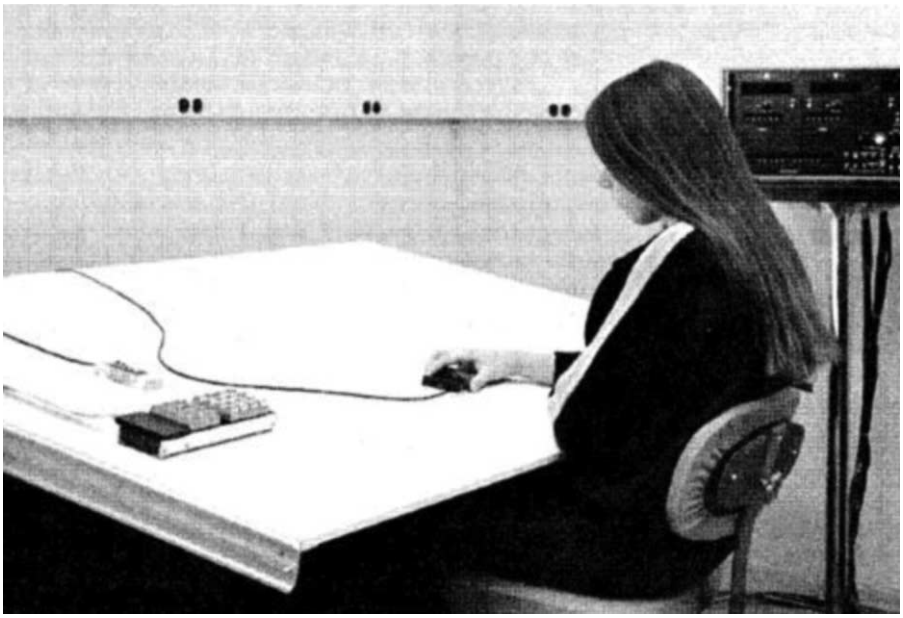

Figure 1. *A Bendix Datagrid Digitizer in use at the U.S. National Center for Atmospheric Research (NCAR) in the 1970s. The operator moved a crosshair-style cursor across analog maps placed on the table's gridded surface. Centering the cursor on a gridpoint, she then used a keyboard (see fig. 2) to enter the value of the isoline nearest to the gridpoint. Image courtesy of NCAR.*

readily manipulated by computers. This "digital convergence" has, of course, continued into the Internet era, when virtually all media use digital formats. Data are increasingly "born digital" (and electronic), a fact well known to all photographers who mourn the loss of analog film. So while the growth of digital data started long before the advent of computers and the Internet, its most dramatic expansion began with the full-scale conversion of previously analog techniques and instruments to "digital" (numerical, discrete) ones starting in the 1940s.

Digital convergence seems to render all data equivalent, and this is certainly the rhetoric of our times. Yet numbers and symbols, no matter what their medium of storage and treatment, remain interestingly distinct from born-analog information. Here we offer just one example. About one megabyte of computer storage can contain the entire text of a 500-page book, where "text" means characters, numbers, and other discrete symbols only. Yet just one page-size (analog) photograph, scanned at reasonably high resolution (1,700 × 1,000 pixels) as a JPEG digital file, is also one megabyte (or more) in size—even though the JPEG format's "lossy" compression algorithms delete irrelevant and redundant information to conserve storage space. Thus, all the text contained in the U.S. Library of Congress's 32 million books and other printed items could be represented in about 20 terabytes (at this writing, four matchbox-sized disk drives). Yet when printed pages are treated as *analog* data—for example, scanned as grayscale images to capture illustrations, marginalia, and other potentially important features—the resulting files are 100–1,000 times larger. Digitizing the approximately 110 million nonprint (analog) items in the Library of Congress's physical collections, such as pho-

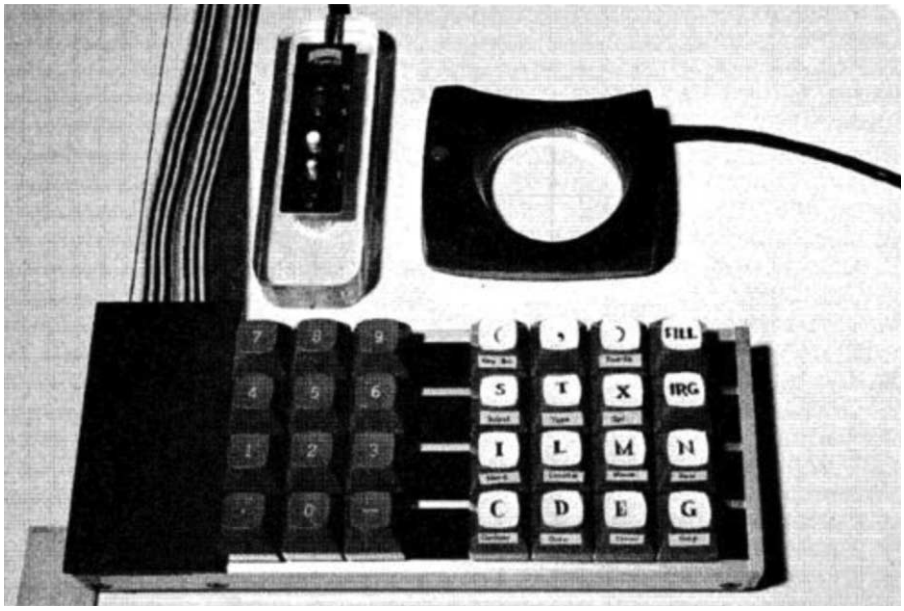

Figure 2. *A Bendix Datagrid keyboard and cursor with crosshairs. Image courtesy of NCAR.*

tographs, sound recordings, and films, to an archival standard of quality would require hundreds of petabytes, perhaps even an exabyte or two—tens to hundreds of thousands of times more storage than all the books in the Library of Congress.[25] Analog data proves remarkably resistant to its digital reduction.

The digital revolution has affected most of the arenas that previously produced and used analog data: photographs, sounds, drawings, and countless scientific recordings, such as electrocardiograms, bubble chamber traces, or electrophoresis gels. In some cases, such as photography, digital sensors have almost completely replaced their analog counterparts, but others are still digitized after the fact from analog media. Once digitized, these data enjoy a greatly increased potential for circulation, calculation, and comparison. At the same time, they have lost the trace of their physical and intimate contact with nature, upon which rested belief in their authenticity.[26] The potential for fraud, in science and elsewhere, is increased by the lack of a unique analog original that could vouch for the reality of the phenomenon being represented.[27] Some new

[25] One exabyte = 1,000 petabytes = 1,000,000 terabytes = 1,000,000,000 gigabytes. The calculation in this paragraph is highly approximate but probably correct to an order of magnitude. For its basis, see the following blog posts: Matt Raymond, "How Big Is the Library of Congress?," *Library of Congress Blog*, 11 February 2009, http://blogs.loc.gov/loc/2009/02/how-big-is-the-library-of-congress/ (accessed 20 April 2017); Michael Lesk, "How Much Information Is There in the World?," 1997, http://www.lesk.com/mlesk/ksg97/ksg.html (accessed 20 April 2017); and Nicholas Taylor, "Transferring 'Libraries of Congress' of Data," *The Signal* (blog), 11 July 2011, http://blogs.loc.gov/digitalpreservation/2011/07/transferring-libraries-of-congress-of-data/ (accessed 20 April 2017).

[26] Roland Barthes, *La chambre claire: Note sur la photographie* (Paris, 1980).

[27] Thus returning to the essential problem of trust in human testimony, inescapable before photography and the age of "mechanical reproduction"; see Lorraine Daston and Peter Galison, *Objectivity* (New York, 2007); and Steven Shapin, *A Social History of Truth: Civility and Science in Seventeenth-Century England* (Chicago, 1995). On fraud and trust in images, see Nick Hopwood, *Haeckel's Embryos: Images, Evolution, and Fraud* (Chicago, 2015).

digital data have also lost the remarkable "depth" that allowed the photographer in Michelangelo Antonioni's 1966 movie, *Blow-Up*, to enlarge his negative repeatedly and thereby to discover a dead body in the background of a romantic park scene. In the digital age of big but "thin" data, his enlargement might only have revealed meaningless gray pixels.

4. WHAT MAKES DATA MEASURABLE? WHAT DOES QUANTIFICATION DO TO DATA?

Our own foregoing discussion is an instance of a widespread practice: the quantification of data, treating all data as having a measurable and comparable size. Today these quantities are usually expressed in bytes, or units of digital storage capable of encoding one character, whether a numeral, letter, or some other symbol. A byte is made up of bits (short for binary digits; i.e., the famous zeros and ones digital computers work with). In an Ur-document of information theory, "A Mathematical Theory of Communication" (1948), Claude Shannon introduced bits as a universal measure of information content.[28] Like money, bits rapidly became a kind of currency used to compare computer storage and processor capability.

The byte superseded the bit as a standard measure when the American Standard Code for Information Interchange (ASCII) text encoding system gained widespread currency in the 1960s. Although 6- and 7-bit bytes were once used, today standard measures such as kilobytes (Kb) and megabytes (Mb) always reference 8-bit bytes. Each 8-bit byte can represent a maximum of 256 characters. This is enough to handle the alphanumeric systems used by many Western languages, but the 8-bit byte cannot accommodate the thousands of characters in logographic languages such as Chinese, or even the many diacritical marks used in such Western languages as Swedish and Czech. Well into the 1970s, this problem stymied the introduction of native-language computing in Asia and later created a technical bottleneck in the early spread of the Internet.[29] Further, as a universal unit for quantifying data, the byte responds poorly to human experience. As discussed above, a 500-page book and a single scanned photograph require the same number of bytes of computer memory, yet from a human point of view, the book usually contains far more information. The conundrums presented by this seeming paradox were evident even at the dawn of information theory. Warren Weaver's exposition in *Scientific American* (1949) noted that Shannon's clarity came at the expense of separating the relatively easy "technical problem" (how information travels from transmitter to receiver; how transmission errors can be prevented or corrected) from two counterparts, the "semantic problem" (how "the information" conveys meaning) and the "effectiveness problem" (how a message affects the recipient's behavior).[30]

The quantification of data almost invariably serves to support a narrative about an "explosion" of data, and it rarely extends much beyond an exclamation of wonder. Yet

[28] Claude Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.* 27 (1948): 379–423, 623–656.

[29] Jeffrey Shapard, "Islands in the (Data) Stream: Language, Character Codes, and Electronic Isolation in Japan," in *Global Networks: Computers and International Communication*, ed. Linda Harasim (Cambridge, Mass., 1993). For computing in Asia, see James W. Cortada, *The Digital Flood: The Diffusion of Information Technology across the US, Europe, and Asia* (New York, 2012); and Basile Zimmermann, *Waves and Forms: Electronic Music Devices and Computer Encodings in China* (Boston, 2015), chap. 12.

[30] Warren Weaver, "The Mathematics of Communication," *Sci. Amer.* 181 (1949): 11–5.

to a historian or an anthropologist, the idea of a single unit of measure for everything that has ever counted as data—whether object, inscription, specimen, or sample; analog or electronic; and so on—seems fraught with puzzles, if not patently absurd.[31] The fact that many kinds of scientific data, but also so many aspects of our informational lives—from family pictures to favorite music, to epistolary relations—have come to be quantified, and quantified using the *same* metric, constitutes a historically significant turning point deserving of scholarly attention. So when thinking about data, instead of marveling at its size, one might ask such questions as, *How did data become a measurable quantity? What does quantification do to data? More broadly, how does data measurement fit into the history of the relationship between quantification, trust, and objectivity?*[32] Elena Aronova offers one answer in the context of the IGY, where national contributions to data collections were quantified in service of the Cold War competition among the scientific powers (including the invention of a "data gap" echoing the "missile gap").[33] To take another example, in the history of taxonomy, counting species has seemed like an obvious thing to do, but as Staffan Müller-Wille has convincingly shown, many ways of doing taxonomy simply described species without quantifying their numbers.[34] The very impulse to measure data—as well as the highly abstract, and misleading, currency of bits and bytes—can thus be questioned, and their history traced.

5. WHAT KIND OF INFORMATION AGE IS THIS, ANYWAY?

Discourses about the unprecedented quantity of data—Terabytes! Petabytes! Exabytes!—serve to justify the diffuse feeling (most likely shared by anyone with an email account) that we are suffering from an "information overload" resulting from a "data deluge," a unique feature of our "information society." A more useful perspective might stem from cultural historian Robert Darnton's remark that "every age was an age of information, each in its own way."[35] Claims of a period-specific information overload have a long history. In his remarkable *Avatars of the Word*, classicist James O'Donnell showed that contemporaries of every major transition to a new technology of writing—from papyrus scrolls to the codex, to the printing press, to the computer age—expressed exhaustion and despair when confronting the increasing quantity of written information.[36] In the Renaissance, botanists complained about the vast number of new species to be described.[37] Darnton and other scholars of the Enlightenment similarly point to the challenges of collecting, storing, and making sense of an increasing

[31] Bruno J. Strasser, "Data-Driven Sciences: From Wonder Cabinets to Electronic Databases," *Stud. Hist. Phil. Biol. Biomed. Sci.* 43 (2011): 85–7.

[32] Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, N.J., 1995)

[33] Aronova, "Geophysical Datascapes" (cit. n. 12).

[34] Staffan Müller-Wille, "Names and Numbers: 'Data' in Classical Natural History, 1758–1859," in this volume; see also Staffan Müller-Wille and Isabelle Charmantier, "Natural History and Information Overload: the Case of Linnaeus," *Stud. Hist. Phil. Biol. Biomed. Sci.* 43 (2011): 4–15.

[35] Robert Darnton, "An Early Information Society: News and the Media in Eighteenth-Century Paris," *Amer. Hist. Rev.* 105 (2000): 1–35, on 1.

[36] James J. O'Donnell, *Avatars of the Word: From Papyrus to Cyberspace* (Cambridge, Mass., 1998).

[37] Brian W. Ogilvie, *The Science of Describing: Natural History in Renaissance Europe* (Chicago, 2006); Daniel Rosenberg, "Early Modern Information Overload," *J. Hist. Ideas* 64 (2003): 1–9.

quantity of knowledge—a principal justification of the encyclopedia movements of the eighteenth and nineteenth centuries. A question for today's Big Data could thus be, *What kind of information age is this?*

If we focus solely on the twentieth century, complaints of an information overload or a data deluge abound. In 1934, for example, the Belgian documentalist Paul Otlet highlighted the rapidly increasing number of written documents: "Their enormous mass, accumulated in the past, grows daily, hourly, by new additions in disconcerting, sometimes frightening numbers. . . . Of them, as of rain falling from the sky, one can say that they could set off a flood and a deluge, or trickle away as beneficial irrigation" (fig. 3).[38] The abundance of information called for new ways of organizing knowledge, the topic of Otlet's illuminating essay, as discussed by Markus Krajewski in this volume.[39] Much like Otlet's heroic Mundaneum, which undertook to store and catalog all the world's knowledge as well as to provide it on demand to researchers, both Vannevar Bush's imaginary Memex (1945), a microfilm-based, proto-hypertext retrieval system, and computer science pioneer J. C. R. Licklider's vision for *Libraries of the Future* (1965) promised to tame the rising tide through finer-grained classification, division of knowledge into atomic units, and associative linking, search, and retrieval.[40] Social commentator Alvin Toffler's influential *Future Shock*, published in 1970, popularized the notion of information overload and explored its psychological consequences.[41] Since then, such claims have become too numerous to count.

Understandably, a number of commentators have attempted to explain today's perception of an information overload by pointing to our time's unique features, such as the ability to produce, store, and share digital data with devices embedded in networked infrastructures. By asking *How is Big Data produced?* they have traced the history of the technologies responsible for the data deluge, from high-energy particle detectors to massive relational databases, and produced stories that fit into standard narratives of technological progress.[42] This often technologically deterministic perspective helps with the "how," yet it does not explain *why* data were deemed sufficiently valuable to store.

6. WHY DO WE KEEP DATA?

Even if one should remain critical about narratives of an unprecedented Big Data era, one must concede that data volumes *are* growing rapidly, while means of producing digital data are proliferating (from high-throughput DNA sequencers to high-definition smartphone cameras). Yet one should not assume that this increase in the amount of data, by itself, leads to Big Data or an information overload. To historicize twenty-first-century Big Data and its specific kind of information overload, it may be more meaningful to examine the relationships among the amount of data considered signif-

[38] Paul Otlet, *Traité de Documentation: Le Livre sur le Livre* (Brussels, 1934), 3; translation by Paul N. Edwards.

[39] Markus Krajewski, "Tell Data from Meta: Tracing the Origins of Big Data, Bibiliometrics, and the OPAC," in this volume.

[40] Vannevar Bush, "As We May Think," *Atlantic Monthly* 176 (1945): 101–8; J. C. R. Licklider, *Libraries of the Future* (Cambridge, Mass., 1965).

[41] Alvin Toffler, *Future Shock* (New York, 1970).

[42] For example, Viktor Mayer-Schönberger and Keneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston, 2013).

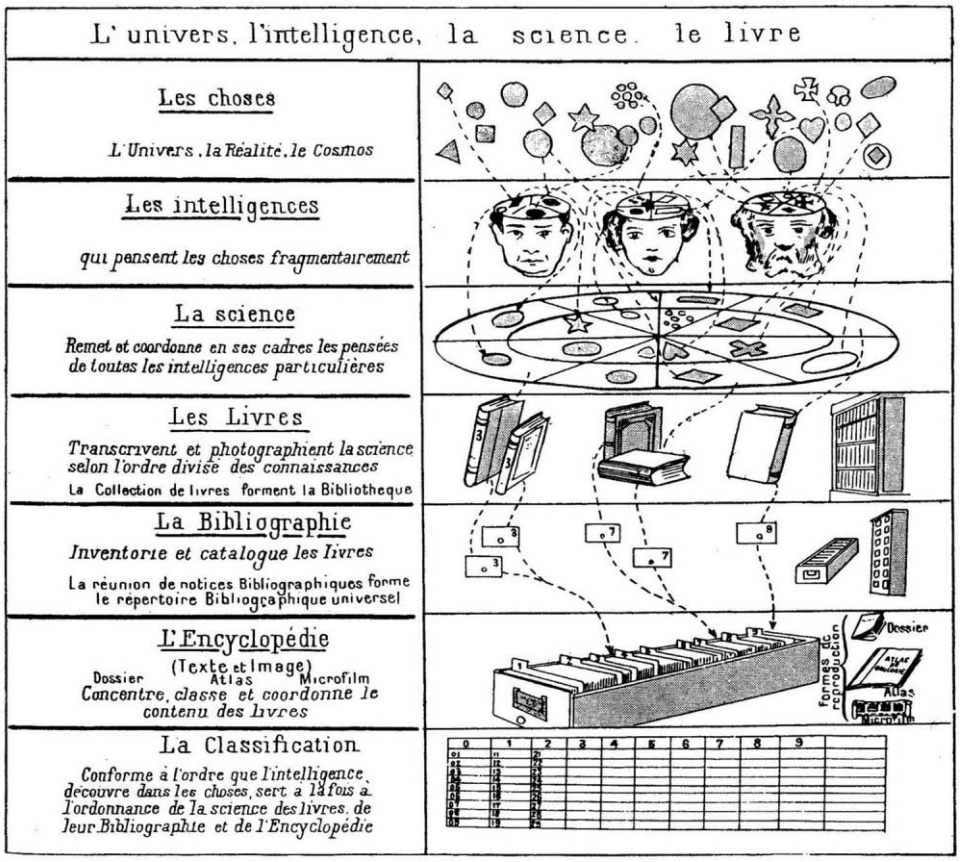

Figure 3. *Paul Otlet's view of the chain of transformations from the world to the classified card indexes, in his* Traité de Documentation *(cit. n. 38), 34.*

icant, the technologies available to handle it, and the perceived benefits of mastering that information. Ann Blair's eloquent *Too Much to Know: Managing Scholarly Information before the Modern Age* epitomizes this approach.[43] From this perspective, key questions would be not only *How big is Big Data?* or *How much data are we producing?* but also *Why do we keep data?* In other words, what has made every shred of scientific information seem so valuable that it should be stored electronically, backed up multiple times, and often subjected to a DMP ("data management plan") for making it available to everyone, thus putting all scientists in the business of data librarianship? Does this value stem from the belief that mining data can produce scientifically or economically useful insights ("data is the new oil")?[44] Or perhaps from an anxiety about the fragility of digital media, which might lead to the instantaneous erasure of a decade's worth of family photographs or laboratory data? Or simply from the pack-rat impulse to keep everything, and thus avoid the challenging (and annoying) task of

[43] Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven, Conn., 2010).

[44] Perry Rotella, "Is Data the New Oil?," *Forbes.com*, 2 April 2012, http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/ (accessed 22 July 2015).

choosing what to discard? These questions redirect attention from the production of data to the uses (and non-uses) of data.

In the life sciences, it is the persistent belief, already common in the eighteenth century, that comparative perspectives can bring unique insights into the relationship between biological structures and functions that has made data about every single species worthy of being stored (think of Georges Cuvier's comparative anatomy). This perceived opportunity to derive knowledge (or value) from data goes further in explaining why we store so much of it than the simple fact that we can produce data in large amounts. After all, the vast majority of data produced is not preserved. Utopias of "ubiquitous computing" or a "quantified self," relying on clunky Google glasses, health-monitoring wristbands, wearable computers, and other devices, will never succeed in capturing more than a very small fraction of the data constantly streaming in through our senses—and will likely one day seem as naive as earlier attempts to create the databases of dreams described by Rebecca Lemov (this volume) or the cloud atlases of the nineteenth and early twentieth centuries.[45] Thus, questions of selection and valuation are perhaps more important than the simple fact of production, and these operations depend on imagined future uses. Thus, one important reason for the emergence of Big Data is the belief that data we store now can (and will?) be (re)used for the production of knowledge or value later on.

In their ongoing quest to reduce the cost and improve the quality of science, funders such as the U.S. National Science Foundation and the National Institutes of Health have come to the not unreasonable view that broad data sharing might strengthen science and save money. With wide access to other researchers' data, some scientists might avoid having to repeat costly observations or experiments, while others might discover new, unanticipated ways to analyze it. Yet the vast majority of data preserved is still used only once (if ever), and so far—with important exceptions—even most published data are never reused by anyone other than the original producer.[46]

Revelations about fraud and error in scientific publishing have generated another, essentially moral rationale for sharing data: replicability and accountability. If data and models were habitually published alongside the scientific articles that make use of them, other researchers could check the claimed results directly, or so the argument goes. This rationale plays into, and perhaps stems from, a larger ideology of "transparency" in public affairs, often pictured as an obvious, quasi-obligatory way to buttress trust and the moral value of openness in democratic societies. Yet publishing data is rarely as simple or as costless as proponents of data publication assume, while evidence for the putative benefits of reuse and replication remains minimal. To put it differently, in science (and in our everyday life more generally), data hoarding seems to reflect more a growing distrust and insecurity about people, institutions, and memory than some transcendental necessity to preserve human knowledge.

[45] Rebecca Lemov, "Anthropology's Most Documented Man, Ca. 1947: A Prefiguration of Big Data from the Big Social Science Era," in this volume; Lorraine Daston, "On Scientific Observation," *Isis* 99 (2008): 97–110.

[46] Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, Mass., 2015); Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman, "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLOS ONE* 8 (2013): e67332, https://doi.org/10.1371/journal.pone.0067332 (accessed 20 April 2017).

Similarly, the cultural history of twentieth-century "time capsules," those buried boxes and coffee cans in which people consigned artifacts of their times to a future more or less distant, reveals much about what people of one time imagined would be most informative to their counterparts in the future.[47] Today, it seems obvious that many of these boxes in fact reveal very little about their times, while many things we would like to know about the past were never preserved because they were deemed uninteresting at the time the boxes were buried in the ground. The trash heaps of past societies are often more valuable to today's historians, as *Sacred Trash,* the history of the Cairo Genizah storing centuries of miscellaneous notes from Jewish life, so beautifully illustrates.[48] In 2008, Apple Computer marketed its backup drive as the "Time Capsule," which works together with the company's "Time Machine" software to preserve all the data contained in a computer. To future historians, the content of a given Apple Time Capsule—should any survive, along with the software required to read them—might be more revealing than that of its older eponym, but it will still capture only a thin slice from the fabric of our daily lives. Even Big Data will always be too small to contain everything we would like to know, especially about ourselves.

The question of the rise of Big Data can thus be reframed as one about the *reasons* we increasingly stockpile data. As the cultural historian Krzysztof Pomian pointed out, the growth of artifact collections in the sixteenth and seventeenth centuries reflected the tastes, passions, and curiosities of their times.[49] Today is no different. We keep data because we imagine future epistemic uses, or imagine that future people will find such uses, and because of the moral imperatives of sharing, openness, and accountability—creating a data deluge along the way. These issues suggest paying more attention to the actual uses and users of data and to the inevitable question of data ownership.

7. WHO OWNS DATA? WHO USES DATA?

Although the notion of data is often presented as impersonal and anonymous, the issue of *who owns data* matters deeply, either because data are about people, as Joanna Radin and Dan Bouk discuss in this volume, or because data were produced by people.[50] The "data exhaust" left behind by hundreds of millions in the course of daily online activity has become a supremely valuable resource for those—mostly large information technology companies—who know how to use it to target advertising, or even to predict what an individual will need before that individual herself is aware of the need.[51] Certain data about individuals is continually controversial, especially when it is collected without their consent or knowledge; among the numerous exam-

[47] William Jarvis, *Time Capsules: A Cultural History* (Jefferson, N.C., 2002).

[48] Adina Hoffman and Peter Cole, *Sacred Trash: The Lost and Found World of the Cairo Geniza* (New York, 2011).

[49] Krzysztof Pomian, *Collectors and Curiosities: Paris and Venice, 1500–1800* (Cambridge, Mass., 1990).

[50] Joanna Radin, "'Digital Natives': How Medical and Indigenous Histories Matter for Big Data"; Dan Bouk, "The History and Political Economy of Personal Data over the Last Two Centuries in Three Acts," both in this volume.

[51] In one famous case, a man complained to the Target department store chain about baby-related advertising material sent to his teenage daughter. Unbeknownst to him, she was in fact pregnant. Her online behavior had triggered the store's algorithms to send the advertising. Charles Duhigg, "How Companies Learn Your Secrets," *New York Times Magazine*, 16 February 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (accessed 20 April 2017).

ples one could cite is the National Security Agency's surveillance of Americans' telephone conversations as revealed by Edward Snowden.

But the issue of data ownership goes far beyond issues of privacy related to data about human beings.[52] Data are often the product of work processes, such as laboratory experiments or field studies, carried out by human beings employed by institutions. As a result, historically, scientists and even institutions have claimed at least metaphorical, and often literal, ownership of "their" data, whether DNA sequences or astronomical data, as W. Patrick McCray shows in this volume, typically through authorship or patents.[53] In some scientific fields, data sharing and rights to reuse have existed all along, however. Early Modern genealogists producing data did not seem to clash with historians who made use of it, as Markus Friedrich points out.[54] Since the nineteenth century, many governments have collected and shared certain kinds of data very widely: censuses such as those discussed by Christine von Oertzen, economic and labor statistics, geographical survey maps, weather data.[55] Geneticists working on model organisms were famously open about sharing their data.[56] Yet in other fields, data have been hoarded or jealously guarded, whether for profit or simply to retain control. The consequences can sometimes be dramatic, even tragic, as when data about clinical trials gone awry are concealed or surreptitiously altered, or damaging for the scientific enterprise, as when scientific claims cannot be properly evaluated because data were withheld.

The question of data ownership has become more acute as what we might call "data supply chains" become more visible and more differentiated.[57] Most data were once produced by their eventual users and thus did not need to be shared; in that receding world, observation and analysis were part of the same process, with storage and reuse by others an afterthought at best. Across the last century or so, data suppliers, data managers, and data users have become far more differentiated and specialized. As the collection, organization, and curation of data become increasingly professionalized, a divide has appeared between the scientists who produce data, those who manage it, and those who analyze it.[58] Because data managers do not themselves seek to exploit the data they handle, they usually do not conflict with data producers. Instead, it is the professionalization of data analysis that has led to moral tensions with data producers. This professionalization results in increasingly distinct communities, such as "bioinformatics" or "lexomics," as discussed by Judith Kaplan in this volume, each

[52] For an overview of these issues, see Kitchin, *The Data Revolution* (cit. n. 4), chap. 10.

[53] W. Patrick McCray, "The Biggest Data of All: Making and Sharing a Digital Universe," in this volume; Bruno J. Strasser, "The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine," *Isis* 102 (2011): 60–96.

[54] Markus Friedrich, "Genealogy as Archive-Driven Research Enterprise in Early Modern Europe," in this volume.

[55] Christine von Oertzen, "Machineries of Data Power: Manual versus Mechanical Census Compilation in Nineteenth-Century Europe," in this volume.

[56] Robert E. Kohler, *Lords of the Fly:* Drosophila *Genetics and the Experimental Life* (Chicago, 1994).

[57] Richard B. Rood and Paul N. Edwards, "Climate Informatics: Human Experts and the End-to-End System," *Earthzine*, 22 May 2014.

[58] Information schools and other academic units have recently begun to offer degrees in "data science," a broad, unsettled category that includes data management as well as analysis, while job advertisements for "data scientists" are proliferating rapidly in industry. These programs and positions are intriguingly agnostic with respect to the actual content of data, reflecting the commodification of data mentioned earlier.

with its own rules and norms, including about authorship and credit.[59] Bioinformaticians routinely analyze other researchers' data, publish articles about the results of such analysis, and get tenure for their work (although they do not win Nobel prizes—yet), creating tensions or even "long strings of clashes"[60] with those who produce the data in the first place.

Although the social rewards attached to scientific activities have always been highly contingent, expert observations of nature have long held a central place in the very idea of scientific discovery.[61] As nature is being replaced by a "second nature" made of data, and as borrowed data become the primary object of investigation for a growing number of researchers, one wonders whether data analysis will increasingly be valued as *the* key intellectual activity in science—perhaps even eclipsing experimentation and theory.[62] Beyond the naive empiricism of "data-driven science" lies an entire world where the practices of correlation, comparison, and classification of data, as discussed by Hallam Stevens in this volume, are deeply embedded with experimentation and theory, and all play their role in the production of scientific knowledge.[63]

In such a world, published data take on characteristics of commodities, or commons, and the questions *Who owns data?* and *Who uses data?* coalesce, since data increasingly belong to those who use them. Today, there is an increasing expectation that scientific data should be free to use and transform, like software, where "free" should be conceptualized "as in 'free speech,' not as in 'free beer,'" as computer scientist Richard Stallman famously put it.[64] Thus data may now be used by people who have no connection to, or any practical knowledge of, the original conditions of data production, resulting in new epistemic risks.

8. HOW DOES BIG DATA TRANSFORM THE GEOGRAPHY OF SCIENCE?

The increased circulation of knowledge and the ability of computerized data to be in many places at the same time—to be virtually everywhere at the same time—should not obscure the fact that data still have a geography. The point is not simply that databases are located somewhere (the staff in one place, the servers often in another), or that users are located almost exclusively in places with Internet access (which now includes almost the entire world), but that the geography of science reflects the specific epistemic practices of Big Data, past and present,[65] requiring specific kinds of local

[59] Judith Kaplan, "From Lexicostatistics to Lexomics: Basic Vocabulary and the Study of Language Prehistory," in this volume; Strasser, "Collecting Nature" (cit. n. 14).

[60] Eliot Marshall, "Data Sharing—DNA Sequencer Protests Being Scooped with His Own Data," *Science* 295 (2002): 1206.

[61] Lorraine Daston and Elizabeth Lunbeck, *Histories of Scientific Observation* (Chicago, 2011).

[62] Tony Hey, Stewart Tansley, and Kristin Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Redmond, Wash., 2009); Chris Anderson, "The End of Theory," *Wired* 16 (2008), https://www.wired.com/2008/06/pb-theory/ (accessed 20 April 2017). On "second nature," see Strasser, "Collecting Nature" (cit. n. 14).

[63] Hallam Stevens, "A Feeling for the Algorithm: Working Knowledge and Big Data in Biology," in this volume; Kitchin, *The Data Revolution* (cit. n. 4), chap. 8; Sabina Leonelli, "Integrating Data to Acquire New Knowledge: Three Modes of Integration in Plant Science," *Stud. Hist. Phil. Biol. Biomed. Sci.* 44 (2013): 503–14.

[64] Richard Stallman, "Free Software Philosophy," https://www.gnu.org/philosophy/free-sw.html (accessed 13 August 2015)

[65] David N. Livingstone, *Putting Science in Its Place: Geographies of Scientific Knowledge* (Chicago, 2003).

infrastructures. The American Museum of Natural History, then the largest natural history museum in the United States, opened in 1877 on New York City's Upper West Side, while the Rockefeller Institute for Medical Research, a temple of the new experimentalism made iconic by Sinclair Lewis's 1925 novel *Arrowsmith*, was founded in 1901 just a few blocks away, on the Upper East Side. The massive collections of specimens and the specialized laboratory instrumentation—that era's Big Data infrastructures—required both the physical spaces of large buildings and the social spaces of a modern urban environment, making natural history as much a field science as one performed in an urban museum.

The geographical logic of centralized depositories held so long as both data-as-objects and objects-as-data remained physically large, heavy, and difficult or impossible to reproduce. While these characteristics still hold for many collections, such as dinosaur fossils or the ice cores drilled from glaciers around the world, the rise of (electronic, digital) Big Data has brought the value of centralized research infrastructures into question. Yet they are unlikely to disappear anytime soon. Today's most powerful computers still require large, expensive, specialized facilities, and the biggest data of all—such as climate simulations, particle accelerator experiments, or satellite data streams—remain too cumbersome and expensive to move easily. Further, the social advantages of large research institutions, where face-to-face collaboration is easier and serendipitous interaction more likely, will probably never be entirely effaced by distant interactions or anonymous crowdsourcing.[66]

Still, data analysis is now also carried out in a distributed fashion, through crowdsourcing. Individual researchers, located outside the main research institutions, can access and use Big Data. As a result, the social landscape of science is changing. Lay citizens are beginning to contribute to the production of scientific knowledge by analyzing scientific data—often from home. In the citizen-science project Galaxy Zoo, citizens study telescope photographs to classify galaxies; in Eyewire, they map data about neuron networks; in Old Weather, they decipher weather data contained in maritime logbooks. These projects have already resulted in numerous publications, often in high-profile journals, and following every standard of a scientific publication—except that they include collective and distributed authors, such as "the Eyewirer" in a recent publication in *Nature*.[67] Big Data, because it is often too big to be fully exploited by those who have produced it, is increasingly made publicly available, opening the possibility of a new kind of citizen science in which lay people act not merely as observers or sensors, but also as analysts. Citizens have long contributed to the making of Big Data, as observers of ephemeral comets and migrating birds, and have even engaged in their own research projects, as Etienne Benson shows in this volume.[68] Today, citizens continue to generate large quantities of observational data, about their natural environment and their own bodies, contributing to the "data deluge" and to overcoming it at the same time through citizen projects controlled by scientific institutions. More important, perhaps, as Big Data escape the control of academic research

[66] Gary M. Olson and Judith S. Olson, "Distance Matters," *Human-Computer Interaction* 15 (2000): 139–78.

[67] Jinseop S. Kim et al., "Space-Time Wiring Specificity Supports Direction Selectivity in the Retina," *Nature* 509 (2014): 331–36, on 331.

[68] Etienne S. Benson, "A Centrifuge of Calculation: Managing Data and Enthusiasm in Early Twentieth-Century Bird Banding," in this volume.

institutions, corporations, and the state, they are produced, appropriated, and mobilized by other actors, such as Public Lab (best known for its participatory mapping of the Gulf Coast following the Deepwater Horizon oil spill) to weigh in on environmental and other issues. As a result, the geography of knowledge—and power—may begin to shift.

CONCLUSION

What makes the Big Data era seem revolutionary is not only a series of incremental technological changes, but also an erasure of Big Data's past. The contributions to this volume show that concerns regarding the collection, storage, and uses of vast amounts of data are not new, and that computers and the Internet are not necessary ingredients. In doing so, they help us distinguish what *is* new with today's Big Data. More important, these studies show that the current concerns with Big Data can be better understood by looking at past situations where, in widely different contexts, people were confronted with large amounts of data and devised solutions to deal with it. The rich historical literature on natural history collections, for example, offers insights into the moral and political economies of today's Big Data and helps us understand current tensions around property, privacy, and credit.[69] Conversely, current issues—the professionalization of "data science," the "placelessness" of data, and the measurement of data—help us revisit earlier historical studies with new eyes. Finally, questioning the current fascination with Big Data can contribute to understanding more broadly our present condition, with its anxieties about lost memories and its hopes for predictable futures. Although the power of data may never match that of music (or so Frank Zappa might proclaim), it is having tremendous consequences for how we produce knowledge and how we live our lives.

[69] Strasser, "The Experimenter's Museum" (cit. n. 53).