Big data is not about size: when data transform scholarship

Jean-Christophe Plantin (London School of Economics and Political Science), Carl Lagoze (University of Michigan), Paul N. Edwards (University of Michigan), Christian Sandvig (University of Michigan)

Final draft version, awaiting minor corrections. Forthcoming in Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards, Christian Sandvig, "'Big data' is not about size: when data transform scholarship," in *Towards an Ecology of Data: Political and Scientific Issues of Digital Data,* eds. Clément Mabi and Jean-Christophe Plantin (Paris: Presses de l'École des Mines)

Abstract

"Big data" discussions typically focus on scale, i.e. the problems and potentials inherent in very large collections. Here, we argue that the most important consequences of "big data" for scholarship stem not from the increasing size of datasets, but instead from a loss of control over the sources of data. The breakdown of the "control zone" due to the uncertain provenance of data has implications for data integrity, and can be disruptive to scholarship in multiple ways. A retrospective look at the introduction of larger datasets in weather forecasting and epidemiology shows that more data can at times be counter-productive, or destabilize already existing methods. Based on these examples, we look at two implications of "big data" for scholarship: when the presence of large datasets transforms the traditional disciplinary structure of sciences, as well as the infrastructure for scholarly communication.

Introduction

"Very large" amounts of data in science are far from a new phenomenon. Previous eras each saw their own "data deluge." For example, the expansion of travel following the discovery of the New World brought naturalists an unprecedented number of specimens, observations, and measurements, forcing them to create new classification

systems (Strasser, 2012; Burke, 2011). The phrase "big data," however, connotes something much more recent. It first appeared in the scientific literature in 1970, and the use of the term slowly increased before reaching a peak around 2008. It appeared most often in computer science, but other disciplines are now involved, including electrical engineering, telecommunication, mathematics, and business (Halevi, Moed, 2012). In the sciences, the phrase's appearance coincides with the advent of large infrastructures to support data-driven science. For example, the petabytes of data streaming from high-energy physics experiments (studied thoroughly by Knorr-Cetina (1999)) or those that are components of the Sloan Digital Sky Survey are certainly "big" in terms of sheer size. A petabyte is approximately one million billion bytes; by comparison, the text of all 32 million books in the US Library of Congress would, if digitized, occupy just 20 terabytes, or one-fiftieth of a petabyte. These "petascale" datasets stress the capacities of computers, networks, and storage systems, as well as the budgets of the institutions that manage them. Even today's fastest computer networks cannot transfer petascale datasets in a reasonable amount of time; as a result, computation is often moved to the data, rather than the other way round.

Several authors emphasize the scientific opportunities presented by these larger datasets. "Big data" proponents have challenged the necessity of theory, promoting science conceived as pattern recognition (Anderson, 2008). Others specifically emphasize the size of datasets and their disruptive effects on science. Writing for a Microsoft Research collection, prominent scientists have suggested that a "fourth paradigm" (Hey et al., 2009) of "data-intensive scientific discovery" is arising due to the ready availability of fine-grained environmental and social data from cheap, ubiquitous sensors and social media. This new paradigm complements the prior paradigms of experiment, theory, and simulation as sources of scientific knowledge. Similarly, "computational social science" (Lazer et al., 2009) combines large sets of born-digital social data (e.g. from email, or gathered by mobile phones) with social network analysis to study how people interact, move, and communicate. In the humanities, several projects such as the digital library HathiTrust aims to provide access to large collections of digitized texts.

By contrast, more critical writers have highlighted the mythology (boyd, Crawford, 2012), hidden biases, and cult of personality (or "fundamentalism," Crawford, 2013) associated with the hype-ridden discourses surrounding big data. Some scholars emphasize how large datasets can come to drive research questions and methods — the reverse of the usual relationship — and thus to frame intellectual approaches in ways that exclude what might be learned from smaller datasets or from methods less driven by the exigencies of scale (Markham, 2013; Mahrt, Scharkow, 2013). An undue focus on immediate "snapshot" analysis is one result (Arbesman, 2013), but other critics point to the difficulties in sampling large datasets, e.g. from the Twitter API (Gerlitz, Rieder, 2013), the shortcomings of dealing with API requests (Vis, 2013), the ethical considerations around accessing personal datasets (boyd, Crawford, 2012), and the various limitations of publishing research based on large data (Bruns, 2013).

A healthy dialogue between proponents and critics of "big data" research helps to develop a reflexive point of view on these emerging scientific practices. Yet we see drawbacks in both perspectives. First, the phrase "big data" remains remarkably vague. The use of the term across several social worlds, from industry to science to marketing, results in multiple definitions. Similarly, the proliferation of "computational" and "digital" sub-fields, such as "computational social sciences" or "digital humanities," seems to arise from a perceived need to engage with large data sets, yet succeeds mainly in highlighting the absence of clear definitions. Second, both proponents and critics of "big data" research focus mainly on the increase in *size* of datasets available, often neglecting other, equally important socio-technical transformations of scientific practices.

In this article, we argue that *uncertainty about the provenance* of data, rather than large scale, best characterizes the real targets of many "big data" discussions. We demonstrate that this uncertainty results from the loss of what Atkinson calls the "control zone" (1996). A retrospective look at the introduction of larger datasets in weather forecast and epidemiology will show that more data can be counter-productive and destabilize already existing research methods. Based on these historical examples, we look at two implications of "big data" for scholarship: transformations of traditional disciplinary structures, and changes in the infrastructure of scholarly communication.

1. The fracturing of the control zone: an alternative view on "big data"

We propose that "big data" fractures the provenance chain that traditionally formed the basis for determining data integrity, and transitively research integrity. Traditionally, provenance chains consisted of physically containable data (e.g., written by hand or stored on magnetic disks or tapes), shared by handing them off physically to colleagues. The physicality of the data, in their containers, and the direct transfer of responsibility to colleagues with stakes in and knowledge of the data and their meaning, provided an unassailable evidence chain. These highly-controlled, regimented procedures were described by Atkinson with the notion of "control zone."

In a seminal 1996 article, the late Ross Atkinson, then Associate University Librarian at Cornell University, described how the notion of the control zone lies at the foundation of the Library. Within this framework, the functioning of the library depends on a clearly defined boundary separating what lies within the library from what is outside. Inside this boundary, within the control zone, the library can lay claim to those resources that have been selected as part of the Collection, and assert curation, or stewardship, of those resources to ensure their consistent availability over the long term. The boundary of the traditional library was easy to define. It was the building that contained and protected the selected physical resources over which the library asserted control and curation responsibility. Correspondingly, from patrons' point of view, the boundary marked what could be called a "trust zone", an area in which they could presume that the integrity guarantees of the library existed. The transition from the physically contained library (bricks and mortar) to the networked digital library has fractured this formerly well-defined control zone (Lagoze, 2010).

This fracturing of the library's control zone offers a useful metaphor for thinking about "big data" in the sciences. Consider social science research. For the past 70 years, many of the most arresting new findings in the social sciences have been derived from expensively produced, but also widely shared datasets. These included survey research and government statistics such as the Roper Poll, the General Social Survey, the American National Election Studies, and the Current Population Survey. In the US, after

the military needs of World War II opened social science to large-scale federal research funding for the first time (Converse, 1987: 242), the government funded an extensive data infrastructure comprised of highly curated, metadata-rich social science archives such as the ICPSR (Inter-University Consortium for Political and Social Research). Like academic libraries, these institutions established control zones permitting data quality and provenance to be preserved, and sometimes enhanced, while making them widely available to the social science community through cooperative inter-institutional arrangements, abroad as well as in the USA.

Today, these archives still play a major role in quantitative social science research. However, the emergence and maturation of ubiquitous networked computing and the ever-growing data cloud has introduced a spectacular quantity and variety of new data sources as well, labeled by some as a "social science data revolution" (King, 2011a, 2011b). These include massive social media data sources such as Facebook, Twitter, and other online communities, which when combined with more traditional data sources provide the opportunity for studies at heretofore unimaginable scales and complexities.

Another example of fracturing the control zone exists in observational science, for example, astronomy, meteorology, and field ecology. In each of these areas there is a growing interest in *crowd-sourced citizen science*, which engages numerous volunteers as participants in large-scale scientific endeavors. Our particular experience is with the eBird project, which originated at the Cornell Laboratory of Ornithology. For over a decade, this highly successful citizen science project has collected observations from volunteer participants worldwide. By nature, citizen science must contend with the problem of highly variable observer expertise and experience. How can we trust data collected or aggregated by individuals who lack traditional scientific credentials such as academic degree, publication record, or institutional affiliation? For this reason, crowd-sourced citizen science makes the established scientific community uneasy (Sauer et al., 1994), especially in fields where people's lives are at stake, such as medicine (Raven, 2012).

These examples illustrate how the traditional control zone for scientific data is breaking down. The reasons for this breakdown are not difficult to discern. For the researcher, an enticing array of data is now available from non-traditional sources, such as social media

platforms. Data mashups, often mixing traditional and non-traditional sources, are becoming increasingly common, sometimes with clear and substantial benefits (Plantin, 2015). Funders, the public, and scientists themselves are demanding better access to data, including fully "open data," in part as a hedge against fraudulent claims based on cherrypicked, illegitimately manipulated, or nonexistent data (Boulton et al., 2012; National Research Council, 2012). This demand grates against the operating principles of some existing data archives, whose organizational *raison d'être* depends on their ability to guarantee data quality and provenance — i.e., on maintaining the control zone. As a result, the traditional criteria for assessing data integrity are being challenged.

While proponents of "big data" present its introduction as fundamentally additive—just more fuel for the fire of research—the arrival of new data sources has, in the past, frequently destabilized disciplines and research practices. In the following section, we present two examples to think through the implications of new data.

2. How "big data" destabilizes knowledge production

"Uncontrolled" data will inevitably find a place in the research process. Just as libraries cannot return to the era of control over physical resources within bricks and mortar institutions (Lagoze, 2010), it would be unrealistic for any science to deny the reality and potential benefits of a sociotechnical knowledge infrastructure that mixes the formal with the informal. At the same time, in many cases adding data from uncontrolled and potentially unreliable sources may jeopardize historically successful modes of knowledge production. Examples from weather forecasting and epidemiology will illustrate some of these risks.

2.1. The case of weather forecasting

In the history of weather forecasting, the arrival of new data sources, from radiosondes (weather balloons) to satellite radiometers, initially created confusion and disrupted existing knowledge processes. At the same time, meteorologists eagerly anticipated them, and they ultimately proved of enormous value. By the turn of the 20th century, telegraph networks were already used routinely for regional data exchange, especially in Europe. Around 1900, meteorologists called for a *réseau mondial* (worldwide network) that would permit the construction of quasi-global, near-real-time weather

maps. Although the telegraph-based *réseau mondial* never materialized, by the 1920s worldwide meteorological data exchange was in fact possible, using a combination of telegraph, teletype, shortwave radio, and several other media. Yet most forecasters never tried to acquire most of these data, and actually *discarded* much of what they did receive. The reason: pre-computer forecasting techniques simply could not use it within the short time (hours) available for creating a useful forecast. Even climatologists, who did not face the time pressure of forecasting, could not (before computers) make use of much of the available data directly. Instead, they developed a system of distributed calculation. Weather stations were asked to pre-compute such figures as monthly average temperatures and report only those, rather than provide all the raw data to central collectors, for whom the calculating burden would have been overwhelming.

Computer forecast models first became available in the mid-1950s. The pioneers of this all-important technique faced a different problem. Weather models divide the atmosphere into three-dimensional grids and compute transformations of mass, energy, and momentum among grid boxes on a short time-step (every few minutes). Every grid point must be supplied with a value at each time step; they cannot simply be zeroed out. Yet most instrument observations of weather are taken every few hours (not minutes), and very few weather stations or other instruments are located exactly at the gridpoints used by the models. So forecasters developed techniques for interpolating gridpoint values, in time and in space, from observations. In other words, they went from a pre-computer situation in which the large amounts of available data were never used, to a post-computer situation in which most data used were actually generated by calculations (interpolation), rather than measured directly.

This problem later led to a far more complicated technique known as “four-dimensional (4-D) data assimilation.” 4-D data assimilation systems ingest observations as they arrive, using them to correct or steer an essentially model-based forecast process in which the previous forecast is used as a “first guess” for the current time period. Forecasters like to say that weather models “carry information” from data-rich areas, where there are dense observing networks, to data-poor areas which lack them. When a weather system moves from a data-rich to a data-poor area, the forecast made while that system was still in the data-rich area becomes the “first guess” for its development

in the data-poor area — thus transferring the information acquired in the data-rich area forward in time to its later location.

A surprising conclusion over five decades of experience with this process is that more "real" data (i.e. observations) are not necessarily helpful. First, uneven global coverage is more of a problem than is insufficient data volume. Second, observations inevitably contain errors due to instrument bias, weather station siting, local topography, and dozens of other factors. Third, since the error characteristics of observations are not perfectly known, the best forecast centers now generate dozens of different data sets from the observations, in order to simulate the likely range of possibility of the true state of the atmosphere. Then they run forecasts on *each* of these data streams. The idea is that since we can't know exactly what the errors in the observational data actually are, the statistical properties of a few dozen forecasts run on a few dozen variations on those data are most likely to approximate what will really happen.

One of the most striking ways forecasters have pictured this process is to describe 4-D data assimilation as "a unique and independent observing system" (Bengtsson, Shukla, 1988, p. 1134) that can generate *better,* more detailed images of actual planetary weather than can instruments alone. In other words, simulation models (albeit guided by real observations) give you better data than do your instruments. Or, to say it even more provocatively, simulated data — appropriately constrained — are better than real data.

To take another example, when satellite photographs of weather systems first became available in the 1960s, many meteorologists were elated, and expected a revolution. As it happened, though, interpreting the photographs proved much more difficult than most anticipated. Taken from a great distance, at strange angles, the photographs showed weather systems clearly but were hard to relate to existing standard measurements, such as temperature, pressure, and wind speed. The same thing happened when radar first entered meteorologists' repertoire; these data promised revolution, but it took well over a decade to work out how to use radar in daily forecasting. In their first 15-20 years, both satellite photographs and radar found their major uses as imagery for television meteorologists — much more symbol than

substance. Certainly they were not yet used as direct inputs to weather forecasts (Courain, 1991).

A last and even more striking episode involves the use of data from satellite radiometers, which measure the energy Earth radiates into space. These instruments, first flown around 1979, generate very large amounts of data continuously (unlike many other weather data sources, such as surface stations, which take readings on a periodic basis). The radiation they measure comes from the entirety of a huge column of air, typically about 70 km wide and thousands of kilometers deep. Because weather forecast models required values only at regular grid points, and because they already stretched the limits of computer power, in the 1980s and 1990s forecasters converted these continuous, volumetric satellite measurements into periodic point measurements — in effect treating the satellites as if they were radiosondes. This massive data reduction went on for two decades before computer power became sufficient to incorporate satellite data directly (Edwards, 2010, ch. 15).

The example of weather forecasting shows that the quest for more data sometimes leads in strange and possibly counter-productive directions. Weather forecasters sought and eventually integrated many new data sources, especially satellites, leading to vast data volumes anyone would describe as "big data." Yet the actual practice of forecasting still incorporates only some of these observational data, while discarding most of it. It also shows how the very meaning of "data" has shifted, over time, toward a definition that accepts processed, simulated, and/or analyzed data as ultimately more useful *and more accurate* than anything instruments alone can provide.

2.2. The case of epidemiology

As we have noted, many conceptions of "big data" do not clearly define what they mean by "big." Researchers in the social sciences, for instance, seem to use the phrase "big data" when they really mean "web data" or "social media data," and the datasets they ultimately collect may not be large by any standard. "Bigness," when discussed at all, is imagined only in terms of procedure—as a problem of "performance requirements," for example (Magoulas, Lorica, 2009, p. 2). This contrasts with other fields, where increasing scale has proven epistemologically significant.

Our example here is epidemiology. In an initially controversial movement termed "evidence-based medicine" (Sackett et al., 1996), doctors of the 1980s sought to join forces with epidemiologists to better integrate systematic findings from associative population studies into the clinical practice of medicine. These "observational" datasets—meaning data from studies that did not involve random assignment to treatment groups and a control group—were large by the standards of the day. They held out the promise of solving large numbers of clinical puzzles at one swoop. The biggest concern at the time wasn't that these methods wouldn't work, but that they were "a dangerous innovation, perpetrated by the arrogant" (71) that would render medical practice impersonal by replacing the subjective judgements of doctors with cold statistics.

But the cold statistics did not work as expected. Recall that pundits such as Anderson (2008) imagine that pattern-finding algorithms will now quasi-automatically generate new truths from large datasets. In medicine of the 1990s, however, more data led to more falsehoods. Epidemiologists produced a plethora of new medical associations such as "Vitamin E lowers the risk of heart attack" and "a low-fat diet prevents cancer in women" (examples from Young, 2009). These new findings received widespread press coverage, only to be refuted when tested via expensive, randomized, controlled clinical trials. A 1997 editorial cartoon proclaimed the *New England Journal of Medicine* to be the "New England Journal of Panic-Inducing Gobbledygook" and depicted the research process as a series of spinning roulette wheels (Borgman, 1997).

One cause of this state of affairs was that statistical techniques then in wide use were calibrated, sometimes implicitly, to smaller sample sizes. Researchers who were used to straining to detect any effect at all were suddenly able to easily detect small statistical associations of questionable clinical value, and they did so, and these were published. Trained to worry habitually about false negatives (that is, the danger of missing genuine, but small effects), researchers had little experience in worrying about false *positives*, which proliferated (Ziliak, McCloskey, 2007). The situation where a sensational new finding appears and then is quickly proven false was termed "the Proteus Effect"

(Ioannidis, 2005, p.698) as researchers learned that the truth, like Proteus and the sea, is mutable.

More and larger epidemiological research studies also revealed the fact that all associations are not created equal. After detecting all of the "conspicuous" relationships between risk factors and disease, such as the fact that smoking can increase the risk of lung cancer by 3000% (Taubes, Mann, 1995: 164), researchers were left to sift the data for the remaining, inconspicuous associations that proved subtle and weak by comparison. As these remaining factors might influence disease by 300%, 30%, or perhaps 3%, they proved very difficult to isolate. In the words of one researcher in the mid 1990s, "we're pushing the edge of what can be done with epidemiology" (ibid., p. 167). The first victories with "big data" promised future successes that proved impossible to deliver, and this imperiled the whole evidence-based approach.

At a more fundamental level, one new truth that emerged is that many of the older truths in epidemiology were also wrong. New data destabilized old knowledge, leading to recent assessments that the bulk of *all* published findings in biomedicine are incorrect (Ioannidis, 2005). Just as computer power allowed larger datasets to be processed with existing methods, it also enabled the creation of new methods, invigorating a 200-year-old philosophical debate about the use of Bayesian inference in place of Frequentist inference (currently the dominant technique in applied statistics). Advances in computers and networks allowed researchers to use more data, but analyzing these data highlighted some of the flaws in Frequentist methods. Advances in computers and networks also gave researchers the computational power to make alternative, Bayesian models tractable.

The story of Proteus in epidemiology did not lead to a backlash against evidence, and larger sample size is still seen as a positive. Yet the "big data" hopes from early evidence-based medicine have been sharply curtailed. Some medical statisticians now advocate that the discovery of an apparently important new finding in a research study should be taken first and foremost as evidence of an error, as any associational study should properly be expected to find nothing of value (Ioannidis, 2008a).

The earlier crises in epidemiology eventually led from cries for "more evidence" to a wide-ranging conversation about the institutions of science and scientific publishing, which were implicated as drivers of false results. Medical statistician John Ioannidis recently characterized epidemiology as potentially supporting two kinds of researchers, the "aggressive discoverer" and the "thoughtful replicator" (2008b, 646). The "aggressive discoverer," he argued, is currently rewarded by the system of scientific publication, yet he produces findings that are extremely unlikely to be true. In the Ioannidis call to action, in addition to more data, successful epidemiology demands Bayesian statistical methods, a new system of scientific publication that rewards replication and tracks null findings, and ultimately a revaluation of the role of data itself. In Ioannidis's terms, the "aggressive discoverer" in medicine conceptualizes datasets as "private goldmines not to be shared" while the successful practice of research demands that databases be public (Ioannidis, 2008b, 646) so that suspect findings can be checked.

As noted previously, in meteorology new data led to an increased reliance on elaborate modes of conditioning and assimilating data, and the field moved away from the analysis of needlessly raw and needlessly large datasets. In the context of epidemiology, new data led to a proliferation of contradictory findings. Although the datasets (presumably) had a known provenance, the discipline is now arguing that the researchers themselves are not trustworthy and that data need to be widely shared in order to allow all new claims to be extensively checked.

3. "Big data" as transformative for scholarship

Our argument thus far has been that there is more to "big data" than just its "bigness". In particular, we see the term as describing data sources and practices that are disruptive on the level of knowledge infrastructures and sociotechnical systems, rather than existing as only a scaling problem that requires a technical solution (e.g., more bandwidth, larger disk drives, parallel computing techniques, etc.). In the first section we related "big data" to breakdowns in the traditional control zone, where the provenance of data sources and collections is no longer guaranteed by established knowledge institutions, generating questions of trust and integrity that remain unresolved. In the second section we saw that new data sources and practices can destabilize disciplines in unpredictable ways. In this section, we explore ways in which

"big data" are disrupting established scholarly communities, as well as the scholarly infrastructure for dissemination and publication.

3.1. Unexpected scientific collaborations

Universities, funding agencies, publication venues, and learned societies all assume, and support, institutionally-defined "disciplines" as basic organizational units. Disciplines can be characterized as path dependencies, in the sense that they represent the continuing imprint of historical choices and accidents. As administrative units and long-lasting professional organizations, they shape not only the nature of research, but also the reward systems — especially promotion and tenure decisions— that drive scholarly careers. Yet close examination immediately reveals that most disciplines encompass a wide variety of methodologies, epistemologies, publication practices, and other norms. This raises the question of whether disciplines are really the most significant levels or structures in academic research communities.

Research on this question has called out a number of other levels and structures, among them invisible colleges (Lievrouw, 1989; Crane, 1969; de Solla Price and Beaver, 1966; Wagner, 2009), communities of practice (Wenger, 1998; Lave and Wenger, 1991), grant-funded projects (Cummings, Kiesler, 2005; Cummings, Pletcher, 2011), and team science[1], which better capture the characteristics of active research communities and work groups. These differ along a variety of dimensions: size, from the solitary bench chemist to the 1000+ person teams of high-energy physics; methodology, including experimentation, simulation, observation, interpretation, and clinical intervention; primary publishing mode, including journals, archival conference proceedings, and monographs; geographical proximity; intellectual and social diversity; and others.

We argue that "big data" simultaneously emerge from and help to generate and strengthen inter-, trans-, and post-disciplinary structures of scientific work, because of the manner in which generating and using big data conflicts with the cultural norms of disciplines. We see this disruption firstly on the scale of a single unit of scholarly practice, and secondly in contexts that bring together multiple scholarly cultures.

[1] See the extensive resource list at www.scienceofteamscience.org/scits-a-team-science-resources.

As an example of disruption within a single field, consider the Human Genome Project. Conceived in the mid-1980s, the Human Genome Project has successfully mapped the entire human genome and made the results available through a fully open, shared, network-accessible database. The resulting data arguably represent one of the great achievements of a cooperative scientific enterprise. Creating and opening this primary data source to the scientific community had profound effects on fields such as microbiology (Glasner, 1996) (Hilgartner, 1998). Laboratory-scale work groups, each producing and guarding data for its own use, became dysfunctional. Instead, progress now depended on a broader collaborative framework and more substantial cooperation. This was manifested through data sharing in online, readily accessible data repositories, which required releasing data from the control zone of the traditional laboratory into a wider, messier arena. Thus, in this well-known example, "big data" transformed scientific practice.

The second type of disruption occurs when "big data" research requires collaboration between scholars previously separated by established and historical disciplinary, field, or institutional barriers. Data-driven multidisciplinary collaborations can create affinities between historically separate epistemologies and methods, providing the context for radically new ways of thinking and doing things. As Kertcher (2010) has said, "[c]ombining knowledge from different domains is the essence of innovation, as it offers individuals and organizations a potent recipe to break away from cemented, path dependent cognitive molds."

At the same time, "forced marriage" collaborations around large data sets often confront profound cultural differences. For example, the study of a "big data" project involving ecologists and computer scientists showed how both had different levels of tolerance towards uncertainty and relations to power — the former bring highly intolerant of uncertainty and more comfortable with hierarchies, the latter being highly tolerant of uncertainty and comfortable with loose organizational structures (Finholt, Birnholtz, 2006). Yet another potential source of friction (Edwards et al., 2011) is differing attitudes towards openness and sharing (Kervin et al., 2012) within as well as between fields, often influenced more by particular research group norms than by the subject of study (Vertesi et al., 2011). Finally, a substantial source of friction sometimes conflicting, often delay-inducing temporal patterns of team members and work groups, stemming

from a wide range of factors including research objects, time zone differences, other work commitments, personal habits, bottlenecks in peer review, and so on (Jackson et al., 2011). For example, a field ecologist's work rhythm may be determined by natural events (e.g., the annual migration of a particular bird species), while that of a climate modeler may be shaped by available supercomputer time. In a closely collaborative project, these differing temporal patterns, along with others, will generate a specific "collaborative rhythm," and may also create friction and conflict.

3.2. The alternative dissemination of scholarship

Besides disrupting research communities and their work patterns, big data has disruptive effects on how scholarly work is recorded and disseminated. Since the origins of modern science in the 17th century, virtually all infrastructures for scholarly communication have embodied a "scholarly value chain" characterized by the following functions:

- *Registration* establishes the precedence of claims and findings.
- *Certification* by other scholars validates claims.
- *Awareness* mechanisms keep researchers abreast of new work.
- *Archiving* preserves the scholarly record over time.
- *Rewarding* creates incentives that increase the quality and quantity of scholarly contributions (Roosendaal, Geurts, 1997).

For many decades, the registration function has been fulfilled principally by published articles. These were packaged in journals, books, and archival conference proceedings which (along with the related citation practices) served as awareness mechanisms, as well as providing archivable records. Peer review has been the principal certification mechanism. These mechanisms, highly (though not entirely) dependent on print technology, have all been severely challenged by Internet- and web-based publication and dissemination. In the print tradition, data were rarely published in raw form. Instead, publication presented researchers' synthesis and analysis of data, for example

in graphs or tables. Raw data remained effectively the intellectual property of their producers. Though peer reviewers occasionally questioned data analysis or asked to see the raw data from which some result was derived, instead they almost always simply assumed the integrity of data analysis. Laboratories and research groups thus functioned as control zones, within which data were produced, managed, archived, and eventually lost. In principle, they guaranteed data integrity — though some scientists famously took advantage of this principle to publish shaky or fraudulent claims.

By making it possible to circulate raw data nearly as easily as analysis and synthesis, electronic media place new demands on the scholarly communication system (Borgman, 2011; Pepe et al., 2009; Wallis et al., 2011; Wynholds et al., 2012). One example is the disruption of publication and citation practices. The publication and citation systems for articles, book, and conference papers is very well established. Today, however, datasets are increasingly recognized as important, publishable scholarly work products. What "publication" means with respect to data remains very poorly defined, encompassing everything from barely documented ftp sites, to obligatory posting of datasets along with the journal articles built from them, to formal stand-alone publication. Data citation schemes remain largely experimental, with many competing versions (Lawrence et al., 2011; Parsons et al., 2010).

Data publication, as an emerging norm, is also challenging peer review, the principal certification mechanism of the traditional scholarly communication system. But mechanisms for peer review of data still remain highly problematic (Parsons et al., 2010). For example, the technological requirements for article review are simple: display text, images, and graphs on screen, or print them on paper. In contrast, evaluating data (whose forms range from simple spreadsheets to petabytes of binary information) may require elaborate technical scaffolding, access to software, and computational resources. Despite these complications, there is general consensus in the scholarly, publishing, and funding communities that in order to re-establish the currently broken scholarly value chain, data must be integrated into the full cycle of scholarly communication.

4. Conclusion

In this article, we aimed to define "big data" through associated transformations in the nature and level of control over the data that underlie research, rather than as a simple reflection of scale or scope. We focused on how data resources and data publication stress traditional knowledge infrastructures, especially the disciplines, the role of methods, scientific collaborations, and the publication infrastructure. If large and potentially exhaustive datasets are often presented as disruptive, we showed that the transformation of the control zone can also destabilize modes of knowledge production: the case studies of weather forecasting and epidemiology showed how the availability of larger datasets could be counter-effective or destabilizing. But this fracturing of the control zone can cause other disruptions, such as upsetting the traditional separations between scientific disciplines and disturbing the peer-reviewed journal as benchmark for scholarly dissemination.

If "big data" means "big uncertainty" about the provenance of data, a major challenge for scientists willing to use new data sources remains the questions of validity, integrity, and quality. Specifically, how is it possible to assess the quality of datasets coming from outside traditional control zones in science? Inevitably, questions such as these reduce to more fundamental debates in science about positivism, constructionism and the like. What is data quality and validity if "raw data is an oxymoron" (Gitelman, 2013), implying that data validity is measured by level of community agreement rather than its "correctness" as a transcription of some "underlying reality"? A possible amelioration of this big data problem lies in developing new tools and techniques that provide the basis for community-agreed-upon trust of new data sources. One path to achieving this is the development of mechanisms to allow reproducibility and replicability of scholarly work (Jasny et al., 2011), and to acquire trust in data by making them open and reusable (Molloy, 2011), thereby encouraging community quality determination. Another possible approach is retrospective determination of data integrity: recovering traces of origin, provenance, and the like from a digital artifact itself, perhaps drawing on the practice of *digital forensics* (Reith et al., 2002), a technique increasingly popular in the intelligence and legal communities. One example is work that we have done in the context of citizen science to infer observers' expertise, and thus the quality of their contributed data, from trace artifacts embedded in the data. None of these techniques bring us back to the "good old days" when well-defined control zones provided a

physical trust environment for data. Nostalgia will not move us forward and, hopefully, the more nuanced approach here will provide us the advantages of the new networked world without throwing out the fundamental supports that science depends upon.

References

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Atkinson, R. (1996). Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone. *The Library Quarterly*, 66(3).

Bengtsson L., Shukla J. (1988) Integration of Space and in Situ Observations to Study Global Climate Change. *Bulletin of the American Meteorological Society.* 69(10), 1130-43.

Borgman, J. (1997, April 27). "Today's Random Medical News." *The New York Times:* E4.

Borgman, C. L. (2011). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science,* 63(6), 1–40.

Boulton, R., Campbell, P., et al. (2012). *Science as an Open Enterprise: Open Data for Open Science* (London: Royal Society).

boyd, d., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679.

Burke, P. (2011). *A social history of knowledge II,*. Cambridge: Polity Press.

Bruns, A. (2013). Faster than the speed of print: Reconciling "big data" social media analysis and academic scholarship. *First Monday*, 18(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4879

Crawford, K. (2013). The Raw and the Cooked: The Mythologies of Big Data. *DataEDGE 2013* (video). Retrieved from http://www.ischool.berkeley.edu/node/24376

Converse, J. M. (1987). *Survey Research in the United States: Roots and Emergence, 1890-1960*. Berkeley: University of California Press.

Courain, M.E. (1991). "Technology Reconciliation in the Remote-Sensing Era of United States Civilian Weather Forecasting: 1957-1987." PhD dissertation, Rutgers University.

Crane, D. (1969). "Social Structure in a Group of Scientists: A Test of the 'invisible College' Hypothesis." *American Sociological Review*: 335-52.

Cummings, J.N., and Kiesler, S. (2005). "Collaborative Research Across Disciplinary and Organizational Boundaries," *Social Studies of Science* 35, no. 5: 703.

Cummings, J.N. and Pletcher, C. (2011) "Why Project Networks Beat Project Teams," *MIT Sloan Management Review*. Retrieved from: http://sloanreview.mit.edu/article/why-project-networks-beat-project-teams/.

de Solla Price, D.J., and Donald Beaver. (1966). "Collaboration in an Invisible College," *American Psychologist* 21, no. 11: 1011.

Gitelman, Lisa (2013) (eds), *"Raw Data" Is an Oxymoron*, Cambridge, MA: MIT Press.

Edwards, P., Mayernik, M. S., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, 41(5).

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

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Data replication & reproducibility. Again, and again, and again .... Introduction. *Science*, *334*(6060), 1225.

Finholt, T. A., & Birnholtz, J. P. (2006). If We Build it, Will They Come? The Cultural Challenges of Cyberinfrastructure Development. In W. S. Bainbridge & M. Roco (Eds.), *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society* (pp. 89–102). Springer Verlag.

Gerlitz, C., & Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C Journal*, 16(2). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/view/620

Glasner, P. (1996). From community to "collaboratory"? The Human Genome Mapping Project and the changing culture of science. *Science and Public Policy*, 23(2), 109–116.

Halevi, G., Moed, F. (2012). The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature. *Research Trends*, issue 30, september 2012.

Hey, T., Tansley, Stewart, & Tolle, K. (2009). *The fourth paradigm data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research.

Hilgartner, S. (1998). Access to Data and Intellectual Property: Scientific Exchange in Genome Research. Washington, DC: National Academy of Sciences.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine 2*(8): 696-701.

Ioannidis, J. P. A. (2008a). Finding Large Effect Sizes: Good News or Bad News? *The Psychologist 21*(8): 690-691.

Ioannidis, J. P. A. (2008b). Why Most Discovered True Associations are Inflated. *Epidemiology 19*(5): 640-648.

Jackson, S. J., Ribes, D., Buyuktur, A., & Bowker, G. C. (2011). Collaborative rhythm. *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW 2011* (p. 245). New York, USA: ACM Press.

Kertcher, Z. (2010). Gaps and Bridges in Interdisciplinary Knowledge Integration. In M. Anandarajan (Ed.), e-Research Collaboration SE - 4 (pp. 49–64). Springer Berlin Heidelberg.

Kervin, K., Finholt, T., & Hedstrom, M. (2012). Macro and micro pressures in data sharing. In Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012 (pp. 525–532).

Knorr-Cetina, K. (1999). *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

King, G. (2011a). The Social Science Data Revolution. Horizons in Political Science. Cambridge, MA: Harvard University. Retrieved from http://gking.harvard.edu/files/gking/files/evbase-horizonsp.pdf

King, G. (2011b). Ensuring the data-rich future of the social sciences. Science (New York, N.Y.), 331(6018), 719–21.

Lagoze, C. (2010). Lost Identity: The Assimilation of Digital Libraries into the Web. Cornell University, Ithaca. Retrieved from http://carllagoze.files.wordpress.com/2012/06/carllagoze.pdf

Lave, J., and Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation (New York: Cambridge University Press).

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital* Curation, 6(2), 4–37.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … Alstyne, M. V. (2009). Computational Social. *Science*, *323*(5915): 721–723.

Lievrouw, L.A. (1989). "The Invisible College Reconsidered: Bibliometrics and the Development of Scientific Communication Theory," *Communication Research* 16, no. 5: 615-28.

Magoulas, R., Lorica, B. (2009). *Big Data: Technologies and Techniques for Large-Scale Data - Strata*. O'Reilly. Retrieved from http://strata.oreilly.com/2009/03/big-data-technologies-report.html

Mahrt, M., & Scharkow, M. (2013). The Value of Big Data in Digital Media Research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33.

Markham, A. N. (2013). Undermining "data": A critical examination of a core term in scientific inquiry. *First Monday*, *18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4868

Molloy, J. C. 2011). The open knowledge foundation: Open data means better science. *PLoS Biology*, *9*, 4.

National Research Council (2012). *The Future of Scientific Knowledge Discovery in Open Networked Environments: Summary of a Workshop* (Washington DC: National Academies Press).

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data Citation and Peer Review. Eos, Transactions American Geophysical Union, 91(34), 297.

Parsons, M. A., Godoy, O., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37(6), 555–569.

Pepe, A., Mayernik, M., Borgman, C. L., & Van de Sompel, H. (2009). From artifacts to aggregations: Modeling scientific life cycles on the semantic Web. *Journal of the American Society for Information Science and Technology,* 61(3) 567-582

Plantin, J-C (2015). Data Politics in Mapping Plaforms: Participatory Radiation Mapping After the Fukushima Dai-ichi Disaster. *Media, Culture & Society*, 37 (6), 904-921

Raven, K. (2012). 23andMe's face in the crowdsourced health research industry gets bigger. *spoonful of medicine: a blog from Nature Medicine*. Retrieved from

http://blogs.nature.com/spoonful/2012/07/23andmes-face-in-the-crowdsourced-health-research-industry-gets-bigger.html

Reith, M., Carr, C., & Gunsch, G. (2002). An examination of digital forensic models. *International Journal of Digital Evidence*, 1, 1–12.

Roosendaal, H., Geurts, P. (1997). Forces and functions in scientific communication: an analysis of their interplay. In Cooperative Research Information Systems in Physics. oldenberg, Germany.

Sackett, D. L., Rosenberg, W. M., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence-based medicine: What it is and what it isn't. *British Medical Journal 312*: 71-72.

Sauer, J. R., Peterjohn, B. G., & Link, W. A. (1994). Observer Differences in the North American Breeding Bird Survey. The Auk, 111(1), 50–62.

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 85–87.

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10).

Vertesi, J., Dourish, P., & Ca, I. (2011). The Value of Data : Considering the Context of Production in Data Economies. Human Factors, 533–542.

Wallis, J. C. & Borgman, C. L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. American Society for Information Science and Technology, New Orleans, Information Today.

Wagner, C.S. (2009). *The New Invisible College: Science for Development* (Washington, D.C.: Brookings Institution Press).

Wenger, E. (1998) Communities of Practice: Learning, Meaning, and Identity (New York: Cambridge University Press).

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry. In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL '12 (p. 19). New York, New York, USA: ACM Press.

Ziliak, S. T. & McCloskey, D. N. (2007). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

Young, S. S. (2009). "Everything is Dangerous." *American Scientist.* (video). http://www.americanscientist.org/science/pub/everything-is-dangerous-a-controversy