Prometheus, Vol. 26, No. 4, December 2008

Data Wealth, Data Poverty, Science and Cyberinfrastructure[1]

STEVE SAWYER

Abstract *Changes in access to data are leading to rapid 'data wealth' in some scientific fields, even as others remain 'data-poor'. Furthermore, the current attention towards developing computer-based infrastructures and digital access to common data sets—the basics of scientific 'cyberinfrastructures'—are too-focused on fields of study characterized by data wealth. To better understand the implications of this twin pursuit of data wealth and cyberinfrastructure, I articulate how data-poor scholarly fields differ from data-rich fields. I then suggest four actions that scholars in data-poor fields can take to improve their work's value to science and society in lieu of being data-rich and propose three design considerations for cyberinfrastructures that can better support data-poor scholarly endeavors.*

Keywords: computerization; cyberinfrastructure; data; e-science; science

Through this essay I argue that science is increasingly demarcated by data wealth and this is diverting attention away from the many valuable contributions of relatively data-poor sciences. My thesis is that this heightened attention comes at the risk of mistaking volumes of data with insight. I argue that both forms of science provide value and must co-exist. To make my case as to why, I first outline my position on data wealth and data poverty. Then, I provide a short review of how science is changing due to the pressures of globalization and computerization. Following these, I develop in more detail what it means to be data-rich and data-poor, using examples from several scholarly fields for illustration. I conclude this essay by identifying actions that scholars in data-poor fields can take, individually and collectively, to both leverage data poverty and minimize its negative effects.

Data Wealth and Data Poverty

The rapidly increasing availability of data, due primarily to new forms of digital sensing, collection and representation, is magnifying ongoing changes to the conduct and expectations of science.[2] The result is that some scientific fields are

Prometheus ISSN 0810-9028 print/ISSN 1470-1030 online © 2008 Taylor & Francis
http://www.informaworld.com/journals
DOI: 10.1080/08109020802459348

becoming 'data-rich' while others remain 'data-poor'. Recently and worryingly, this pursuit of data wealth is becoming tightly coupled with the development of cyberinfrastructure.

By data wealth and data poverty, I mean here the volume of data available to the scholarly community. Thus, data wealth and poverty are collective assets that allow scholars to gain access to and use data as a communal resource.[3] Clearly, there is variance among individual scholars on the amount of data they have collected or can use. So, even in a scientific field of relative data poverty, there are scholars—and groups of scholars—that have larger data sets than do their peers. Wealth and poverty are also relative terms because there is no absolute measure of totals. Thus, wealth to scholars in one discipline may be seen as poverty to scholars in another. For example, economists may see 15 years of industry-level data on the roles of information and communication technology in society as relative poverty (compared to the more than 100 years of data on industry-level data on the roles of manufacturing technologies). In contrast, scholars of information systems would see this 15 years of data as wealth relative to the typically smaller data sets that they use. Framing data wealth or poverty as field-level and not at an individual (or laboratory/research group) level, means these are both relativistic (as noted above) and archetypical terms. As archetypes, these terms are used to frame perspectives and serve as orienting concepts. Reality, of course, is not so neat.

My interest in the effects of data wealth and poverty arise in part from the larger issue regarding the cumulative availability of data to scholars. New data collection instruments, and particularly those that are digital, are contributing to rapid changes to data availability in some fields, and this, in turn is helping to drive changes in research methods, findings, concepts, relations with other scholarly fields, and expectations. For example, in a 2007 conference that focused attention on the potential, needs for and issues with developing cyberinfrastructure, representatives of fields such as astrophysics, biology and ecology noted that, primarily as a result of new data collection instruments such as space telescopes, imaging systems, and sensor networks, scientists in these areas are increasingly unable to classify and index, much less analyze, the data they are collecting![4] Some are even arguing that data are now driving science in ways that have never been seen or even imagined and that the focus of research should be driven by, not drive, data collection.[5] This emerging data wealth is impressive, possibly field-changing, and neither pervasive nor well-understood.

A poverty of data has always been a burden; and a burden shared by scholars in most fields of study. The recent and greater access to substantially more data in some fields is exacerbating these burdens for scholars whose fields remain data-poor. As I explain below, data poverty hinders the building of a cumulative knowledge base. And, this poverty often limits scholars' ability to productively engage in larger discourses regarding phenomena. A paucity of data makes it difficult for scholars to discriminate among existing theories as there is insufficient support to effectively dismiss or modify misguided or nascent theoretically forays. Limited amounts of data available to a community of scholars lead to data hoarding; multiple—and often incompatible—forms of data being collected; and to difficulties with sharing and building large-scale data sets. Much of my scholarship is in the area of information systems, a field of relative data poverty[6] that exhibits all of these characteristics.

These concerns stand in contrast to scientific conduct in data-rich fields. In data-rich fields, the theoretical choices are more limited, possibly because large

amounts of data allow for extensive theoretical testing and the pruning of weaker theories. In data-rich fields, scholarly attention turns from data hoarding towards data sharing and tool building—for both data collection and its analysis. Data hoarding may happen in data-rich fields—particularly if novel data—but, hoarding in data-rich fields will likely have less of an effect (for most) because of the availability of other data.

Science is Shifting

Two sets of forces are helping us to rethink the goals, if not the very nature (or 'doing'), of science: globalization and computerization. A full treatment of these forces and the changes they are engendering lies beyond the scope of this essay and these are discussed with great insight by a number of contemporary scholars.[7] My simple summary begins with globalization, where science is seen increasingly as both a source of innovation and as demanding ever higher levels of intellectual and capital resources.[8] Governments see the pursuit of science as a means to foster economic benefit. Corporations pursue science as a means to generate new products and better use their resources. Seen this way, globalization is in part the result of a market-driven, economic model of life that permeates much of the developed west and frames much of the rhetoric for the developing east and global south. Science has more stakeholders, and fewer of them are other scientists.

Within this framing, the nature, goals, and roles of science relative to society are shifting from the pursuit of knowledge to the pursuit of economic, social and cultural advantage.[9] That is, the 'doing' of science is increasingly seen as a means of direct economic benefit or social value, and not as a means for the betterment of all. Science, in this globalized economic world, is a renewable source of fuel for the powerful engines of production.

Globalization is, of course, bound up in the discourses of computerization. Relative to science, computerization efforts to date have centered primarily on developing software applications and devices to gather, analyze, and represent data, findings and collections of materials. These are possible in large part because in doing this, the very materials of science—data, its analysis, and presentation/representation—are increasingly gathered, stored and presented in digital forms.

This digitization of data involves the uses of computer-based sensors and digital representations of phenomena in a form consumable by computers. This leads towards computerization of data collection instruments and data analyses and—broadly—the production of knowledge as shaped phenomena reflecting recent technological advances that build on and shape a series of trends in science that have been evolving over decades, if not centuries.[10] Computerization is both deeply intertwined with digitization and distinct. Digitization is the raw material and product of computerization. Moreover, digitization and computerization are at the heart of discussions in the scientific community of the growing need to invest in what is called cyberinfrastructure in the United States and E-Science in Europe (and possibly the rest of the world).[11]

Cyberinfrastructure

The major elements of cyberinfrastructure are depicted in Figure 1 as a 'layer-cake' model.[12] The premise behind this model of cyberinfrastructure is of a digital infrastructure comprised of a set of distinct elements that, when combined (stacked)

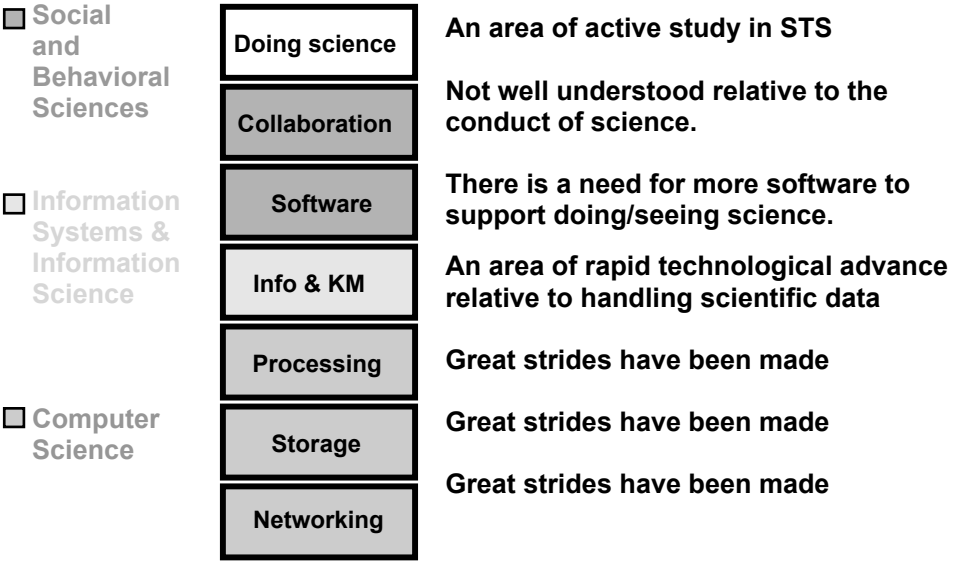

Figure 1. Cyberinfrastructure: layer-cake perspective.

together, provide a means to do science in new ways with greater value to scientists and, more broadly, for society.[13] The elements include sophisticated and large-bandwidth telecommunications networks, powerful and distributed computer processing and large-scale, distributed, secure data storage. This is the province of computer science and engineering and has been the primary focus of contemporary cyberinfrastructure funding and attention. Not surprisingly, these are also the elements of cyberinfrastructure seen by many as the most well-developed.

Information scientists have become increasingly involved in the issues of information retrieval and knowledge management relative to cyberinfrastructure in particular and large-scale distributed knowledge bases more generally.[14] These elements of cyberinfrastructure include the data structures, search, retrieval, manipulation and representation of data, meta-data and use. More profoundly, access to and the uses of data are intimately tied to the research processes and doing of science—those aspects of cyberinfrastructure which are both higher-up in the layer-cake model and have not been as actively studied or well-understood. Thus, the areas of data access and knowledge management are now seen as areas of critical importance to leveraging the technological infrastructure to the best advantage for scientists. In this area of active attention are the software tools which can gather, store, use, visualize and analyze data.[15] As detailed in the United States' (US) National Science Foundation (NSF) vision for cyberinfrastructure, this is the area of dire short-term need.[16]

In the layer-cake model of cyberinfrastructure, the elements of collaboration and virtual organizing, the ways of doing science (processes) and the practices of scientists have been relatively under-attended to date. While these elements are seen as critical, few resources have been devoted to studying them.[17] Luckily, a growing number of scholars in Science and Technology Studies (STS) have, however, been very actively engaged in the issues of the conduct of science, the roles of data and information (and its form and structure) and the increasingly digital nature of

science: all directly relative to cyberinfrastructure.[18] These STS scholars have begun to both provide empirical details and increasingly complex theorizing on the ways in which institutional structures, policy guidance and behavioral issues will influence cyberinfrastructure use, as has previously been found in lesser-scale efforts relative to digital libraries and studies of 'collaboratories'.[19]

Contemporary Examples of Cyberinfrastructure

Three examples of cyberinfrastructures being developed to handle rapid data-wealth are: the Physicist's Open Science Grid, the National Center for Microscopy and Imaging Research (NCMIR) and Long Term Ecological Research (LTER) networks. The Physicist's Open Science Grid (see www.opensciencegrid.org/) has three goals. The first goal is to provide support and infrastructure to engage large-scale and shared experiments. The second goal is to store the results of these experiments in a large shared-data repository. The intention is to move towards hundreds of petabytes[20] of data to be collected and made available for shared use. The third goal is to provide the research education, training and support to allow physicists to pursue this form of science.

Neuroscientists are pursuing their cyberinfrastructure differently than the physicists. The National Center for Magnetic Image Resonance (NCMIR, see http://ncmir.ucsd.edu/) is both more multidisciplinary and more focused on a particular type (and source) of data—namely, brain scans. The NCMIR staff are pursuing the infrastructure to support millions of brain scans, each of which may take nearly one petabyte of storage. Since there are approximately seven billion brains currently available (with more to come), the scale of storage, access and indexing issues are both impressive and known.

A third example is from ecology and in particular, oceanic science where entities known as long term ecological research (LTER; see http://www.lternet.edu/) are being created. These exist to support large and very diverse data sets that are, in turn, being assembled and curated for a very broad group of scientists. Currently, the LTER data volume is moving from terabytes towards petabytes, with increasing demands on access, analysis and uses.

Possibilities

Beyond these three examples are others. We could also point to astrophysics (and in particular cosmology), various subfields of biology and other scientific communities who are also pursuing cyberinfrastructure projects. The list of such projects is impressive and growing in both number and sophistication. The basis of science is seemingly, and increasingly, a globalized, digital enterprise enabled by computers.

While the examples above draw from the natural sciences, cyberinfrastructures for various social sciences and the humanities are also emerging. For example, scholars in information science who focus on information retrieval are benefiting greatly from the rapid growth of the Internet and uses of search engines such as Yahoo and Google. These systems are capturing and cataloging billions of searches. The results of these searches provide information-retrieval scholars with extensive data sets that were unthinkable even 10 years ago. A second example is the explosion of online social engagement via email, mobile phones, and more recently, social networking sites like MySpace and LinkedIn that provide an unprecedented

amount of data on social engagement, communication and interaction. These data are increasingly being used by sociologists and communications scholars. A third example can be seen in economics, where scholars are benefitting from the extensive data collection activities of the many governments and non-governmental organizations such as the World Bank, United Nations, and Organization for Economic Cooperation and Development. And economists are also increasingly drawing on new sources of digital data (from Internet traffic to online economic activity) to bolster these existing large-scale data sets.

Beyond the recent and rapid expansion of digital data, information scientists, sociologists, and economists have a second commonality. Scholars in each of these areas have experience with large-scale, common data sets. Information scientists doing retrieval relied for years on library catalogs and the federally-supported text retrieval conference (TREC) competitions (and TREC data sets see http://trec.nist.gov/). The similarity of library catalogs and the common use of TREC data sets have led information-retrieval scholars to grow comfortable with using common data sets and shared data. Likewise, some sociologists (primarily the demographers) and economists have grown accustomed to working with common data sets like the general social survey, the Census, and the material available from US federal government sources such as the Bureau of Economic Analysis, National Institutes of Justice, and others.[21]

Data-Rich Fields

More broadly, these areas of scholarship all share characteristics that represent what I call data-rich fields. By data-rich field, I mean an area of scholarship where much data is available, and it is seen as a common asset. And, as I outline below, beyond sheer volume, data-rich fields can be identified by three additional common characteristics: access to data sets is shared; method choices are limited; and only a small number of theoretical camps have been validated.

Pooling and Sharing Data Resources Expected

Scholars in data-rich fields typically pool their data. It may be that individual scholars work with a data set before providing more open access. Or, it may be that most—if not all—data are shared, with individuals drawing from this as their interests and research work dictate. In either situation, common access to data is the issue. This common access to data leads to scholars developing a shared understanding of the data and of the issues with its collection (such as the level of data quality or issues with this) and curation (such as how to best use the data). In some cases, data collection becomes a specialized (and perhaps field-engaging) activity. In many large-scale physics experiments, for example, specialists focus on developing and using instruments to collect data. For example, in polar research, data collection is a field-centering activity, with data collection being done collectively by specialists. In the social sciences, the US Census and the Bureau of Economic Analysis employ specialists to gather and develop the data sets for others' uses.

Form(s) Drive Methods

In data-rich fields, the form of data often dictates the analysis approach taken. For example, the collections of search-term strings that search engine companies

make available lead to statistical analyses of search and reduce the likelihood that more detailed contextual data will be collected. In this case, the volume of search-term data is likely to reduce the chances of someone doing field research regarding who is the searcher sitting with and interacting as they search, and to what effect. As a second example, the uses of radio-telegraphy for gathering data for cosmologists leads to developing specific methods tuned to the mass of data being gathered.

The cumulative nature of work that builds on common data and common methods, such as is seen in economics, reduces the likelihood of alternative data sources/forms and other methods being common, or perhaps even accepted. This reduced variation means that, over time, it may be that the choice of analysis approach becomes one of the key differentiators in discerning acceptable or unacceptable scientific activities. Said differently, minute variations in the nature and uses of a common method may become much more divisive when these are the analytic differentiators.

Data-Rich Fields have Few(er) Theoretical Camps

In data-rich fields, the number of theoretical choices is limited because there is enough data to adequately test competitive theories. Simply, data pressure forces out un-supportable theoretical positions.[22] That said, existing theories become even more difficult to reconcile because these rival positions have sufficient empirical support to withstand criticism. Moreover, theories that survive in data-rich environments are likely to develop through the twin pressures of repeated tests and clearly delineated boundary conditions and assumptions. In this way, theory choices become defining—it is hard to move among different theoretical camps. I further note that data wealth and theoretical camps are related in that theories help to shape which data are collected while data help to shape which theories provide greater insight. Issues with the mutually-constituted relationships among data and theory deserve more than the brief in-passing acknowledgment received in this essay.

Data-Poor Fields

Data-poor fields can also be defined by three characteristics in addition to a shortage of data. In data-poor fields, data are a (if not the) prized possession. In these fields, the types of data available often dictate the methods taken. And, third, in data-poor fields there are many theoretical camps.

Data is a Prized Possession

In data-poor fields, hoarding data is common, if not expected. Given the limited amount of data that people get access to, and the difficulties in getting it, scholars tend to be loath to share. Sharing weakens any one scholar's ability to contribute by reducing his or her control over a prized asset. Because of this, there is much less focus on developing common data sets. This leads to where this is typically little (or no) common understanding of data sets, even to the point of contention on the specifics and the murkiness of data being reported. Data suspicion is more typical than data commonality in data-poor fields, meaning that much of the scholarly

discourse is focused on justifying data relative to validity, generalizability and perhaps even accuracy.

In data-poor fields, gathering data becomes a focal activity of scholarly work, and it often consumes a substantial portion of one's research time and funding. That is, because data are rare, or hard to get, or both, most scholars in data-poor fields spend a significant amount of their time (and resources) engaged in gathering their own data. Thus, it remains a private activity even though there is shared interest for many about how best to collect data, and even though substantial—though diffuse—resources are put towards data collection.[23] Essentially this is subsistence data collection: each scholar gathering enough to survive, with some luckier than others in gathering data, making them data-rich relative to their peers. This suggests that the getting of data, and the quality of the data collected, is often a differentiating characteristic of scholars and, perhaps, of institutions.

In data-poor fields, some institutions will be valued for their location relative to hard-to-get data or the ability of the institution to provide access to hard-to-reach data sources. In some ways, this is similar to variations among institutions relative to their library holdings (often the empirical source for scholars in the humanities) or special instruments for collection and analysis that data-rich fields require (such as access to telescopes, microscopes or other unique resources). More pointedly, having or not having access to data is likely to be a career-changing event in data-poor fields.

Because of this hardscrabble existence relative to data, scholars take data in whatever forms they can find. This leads to a state in which most scholars have data compiled in many forms, structured in many ways, and often idiosyncratic to the collector or collection instrument (not to the phenomena). Such diffuse and heterogeneous data collection leads to complex analytic approaches and higher levels of innovation relative to combining and drawing insights from these data. Limited and heterogeneous data sets can also lead to boutique (or one-off) analyses and great confusion both to the meaning and nature of data.

Access to Data Drives Methods

In data-poor fields, the forms of data or the means which data can be collected often drive both the research design and data collection approach. Or, more subtly, scholars make choices of methods based in part on the types of data that they think they can get. It may even be that access to a data set drives people's interests or activities. This activity might be a partial explanation for the incredible interest across many scholarly communities about free/libre and open source software (F/LOSS): there are publicly accessible data archives.

Many Theoretical Camps

Data-poor fields typically are host to many theoretical camps as well as multiple theories. There are at least two reasons for this multiplicity. First, the type of work being done, and the forms and volumes of data being used, lead to much theory development, complicating theory elaboration and theory testing. Second, because there is more exploration and less accumulation, theories in data-poor fields are relatively easy to generate and much more difficult to validate (or deter). Given the relatively large number of theoretical choices and relatively low level of

completeness of any one theory, theory choice(s) are not defining. It is possible to move from theory one to theory two without violating epistemological frames.

Rich Data or Wealth?

The archetypical concepts of data wealth and data poverty I develop here provide a means to characterize perspectives on doing science. Certainly, archetypical representations of scholarly fields such as data-rich or data-poor gloss over the important and nuanced variations among fields, scholars, research approaches and issues with data, instrumentation, sharing, and other mechanisms that shape the doing of science.[24] And, there are other ways to consider the nature and roles of data. For example, the concept of rich data as discussed in ethnography is that detailed data in multiple formats collected *in situ* and over time, used to describe for others a nuanced exploration of a phenomena and its locale, does not easily map to the volumetric depiction of data wealth or poverty. Detailed data may provide rich insights into a situated phenomenon, but these data do not easily aggregate and are often difficult for others to use. In this vernacular, the data may be rich but the field is not wealthy.

So What? Accelerating Differences!

Data-poor fields have always existed, but I argue that data paucity matters more now because of the interest in many scientific fields and those who are primary sources of funding (for instance, the US's National Science Foundation and National Institutes of Health) towards doing science through cyberinfrastructure. The combination of increasing disparities in access to data with increasing enthusiasm to build infrastructures to support data-rich fields leads to where data-rich fields are seeing substantial increases in their access to resources (funding, attention) and greater influence relative to political power. These represent the basis of what social scientists call political economy (the political economy of data) that threatens to further differentiate the contributions of various scientific endeavors by data volume beyond the insights provided, forms in which the insights are made, or the methods used.

If data wealth becomes one of the primary factors for establishing both legitimacy of insight and a basis for receiving additional resources, it becomes even easier for the lay public, political leadership, and perhaps funding organizations, to allow data volume to overtake insight and contribution as key measures of scientific value. I call this the 'freakonomics effect'.[25] One of the basic premises of freakonomics is that economists are more excited about large data sets than about pursuing questions of real value: questions that require special (and often small and difficult-to-get) data sets. In a series of chapter-length study write-ups, the freakonomics authors demonstrate how small data sets can provide great economic insights.

The freakonomics effect can also be seen in the information-retrieval research community. Here the enormous wealth of data regarding search-term entries into information retrieval sites such as Yahoo! and Google is making it very difficult for more sustained research on user behavior (demonstrated most clearly by those scholars pursuing information search in context research[26]) to be done. In this community, scholars were data-poor up through the early 1990s. Thus, the particular approach to doing research and the types of data did not differentiate their

scholarship's value. However, as the use of search engines on the web exploded, so, too, did data for the search-term scholars. Neither the search term nor search-in-context scholars are rewarded for gathering data; but, the search-term scholars have suddenly become data-rich, and this translates to a higher likelihood for career success relative to garnering resources or producing publications.

More to the point, an increasing number of the social sciences are moving, sometimes with dizzying rapidity, from data poverty towards data wealth. Each area is proceeding differently—depending upon the local institutional logics of the field's expectations and norms and by the forms and shape of the field's data that are increasingly prevalent. These rapid changes relative to data wealth have some commonality, and the one I note here is that these changes benefit some scholars over others. For example, and as noted, the field of economics has, over the past 40 years, shifted to be primarily focused on building and testing models using large (and often federal) secondary data sets. More recently, economists have even begun advocating for synthetic data sets to increase the amount of data they can use.[27] Likewise, marketing scholars used the development of the Profit Impact of Market Strategy (PIMS) database to support the development of an area of marketing expertise and legitimacy relative to other scholarly areas in business schools.[28] Likewise, software engineering scholars leveraged the NASA software engineering laboratory (SEL) data sets in the 1980s to legitimize their claims that they should be a sub-discipline of computer science and to raise their relative value to other scientific disciplines regarding how to build software.[29] Of course, these efforts have their limitations. For example, the NASA SEL data set does not provide much insight into the development of commercial software products or free/libre and open sources software (F/LOSS) approaches.[30] The PIMS data set has not been universally accepted by marketing scholars, and it does not have much data relevant to online sales and marketing.

More broadly, as data sets become large, they begin to change more slowly. This makes them less likely to accommodate new phenomena or to incorporate new methodological or analytical needs. And, these larger data sets draw more stakeholders: policy-makers, business interests, curators, and the public all become engaged in large-scale data issues. Meyer[31] argues that these large data sets become, essentially, institutions that have structure and governance issues which both constrain and enable their growth and uses. Large data sets become 'path dependent' in that their current form and structure shape what will be done in the future. Clearly this is beneficial for building cumulative insight. It is also, as noted, likely to constrain what choices are made about adding new types and categories of data. And, a singular pursuit of large data sets may be a barrier to gathering new data or focusing attention at new phenomena that arise.

The current interest in large data sets is particularly problematic for scholars in fields where the focus is towards emerging—and, thus, under-examined and difficult to detail—phenomena. Given the nature of the phenomena, it will be difficult to engage in sustained data collection. And, the data being collected are likely to be of multiple forms, making the construction of larger data sets complex. For fields such as information systems, new media/Internet studies, and human/computer interaction, it may be that developing large common data sets is both prohibitively difficult and would forestall the field's ability to respond to emerging phenomena. That is, the nature of the phenomena is such that it will be difficult to develop large data sets, or that large data sets obscure the nuances of initial introduction and subsequent innovation. Since the intention of studies in these fields is

typically to explore the nature and effects of emerging technologies and phenomena relative to current empirical and conceptual understanding, data wealth is less important than insight.

The issues with building large data sets in fields focused on emerging phenomena are also characteristics of data poverty of a form that is structural (the nature of the phenomena) not behavioral (due to the actions of the scholars). If so, then sustained attention to leveraging the burgeoning collections of case studies and phenomenological studies—which are typical approaches in fields focusing on emerging phenomena—into a larger data set would be the sign of a field interested in creating data wealth—and also a goal for a cyberinfrastructure for data-poor fields.

There are many positives for increasing the amount of and access to data. Across many scholarly fields—and particularly the social sciences—there are emerging opportunities regarding data. For example, the explosion of digital data collection supporting the uses of radio-frequency identification tags (RFID) and increasing digital traces of communication and online activities regarding both electronic commerce and collaboration (such as virtual team work) portend opportunities for scholars to both get access to large amounts of relevant data and to consider pooling and other aspects of data wealth. Increasingly, online traces of social activity (such as MySpace and Facebook) and the growth of wireless internetworking blur the boundaries of family, work and commuting. More broadly, online organizing and social activities are opening up new forms of study and social interaction that are both of direct interest to information systems scholars and of the form to encourage data sharing activities.

What to Do?

The premise of this essay is that data-poor fields of scholarship stand the risk of being accidentally marginalized relative to both their legitimacy to claims of insight and their access to resources to move their research forward because of the current and perhaps implicit bias towards data-rich fields and the current focus to what cyberinfrastructure means and who it is intended to help. Here I raise four actions that data-poor scholars can take—some individually, some collectively—to both better pursue data wealth and to leverage the skills honed from decades of relative data poverty: (1) better connecting micro-studies to macro-data (emphasizing the macro/micro value); (2) focusing more on theory elaboration and less on theory testing or theory borrowing; (3) drawing on expertise in analyzing mixed data sets to do more multi-method research; and (4) focusing on data pooling. These are not the only four. They do reflect, however, the issues with data poverty and the opportunities to leverage skills honed by many data-poor scholars to date.

(1) Emphasizing the Macro/Micro Value of Data-Poor Research

Much of the work done in data-poor fields focuses on micro-level studies at the expense of more macro-level studies. Often, these larger trends are mentioned in passing without a detailed connection being made. This first suggestion is that scholars in data-poor fields should be more aggressive in connecting the macro to the micro by drawing on secondary data and by more explicitly locating the micro-level study both conceptually and empirically. Connecting the micro data with macro data in ways helps to create a place for data-poor scholarship in the debate

as providing insight and countering un-contested (or un-testable) claims based on other forms of data.

More attention to making explicit the macro/micro links demands that data-poor scholars learn to better use secondary data such as is available from federal and other sources (perhaps even partnering with for-profit research companies to leverage the strengths of both the company and the scholar). In doing this, data-poor scholars should focus more on developing comparisons and contrasts because these can illuminate alternative insights and the deeper understanding that micro-studies can provide. For example, many ICT-related phenomena demand micro-level studies because of the relative novelty and recentness of the technology and its constituent phenomena. This action plays to the strengths of much current work in my scholarly field of information systems.[32] That is, detailed studies on implementation of new systems provide the type of rich insights into problems with the take up and uses of computing in organizations that larger-scale data do not.

(2) Focusing on Theory Elaboration

One of the characteristics of data-poor fields is that the contemporary theories are incomplete. This suggests that rather than doing more theory generation, the focus should be towards refining and extending these indeterminate theories—what Vaughan[33] calls theory elaboration. For example, again using my knowledge of the field of information systems, Orlikowski and Iacono[34] have argued that scholars should be more aggressive about directly conceptualizing ICT. Similarly, scholars in STS have been very focused on encouraging work in this field to drive towards theory elaboration.[35]

(3) Leveraging Mixed-Data Expertise

Because most data-poor scholars live a data-constrained existence, they have paid great attention and worked to develop the skills needed to take advantage of mixed data forms. One area of great potential is to move from the small-scale efforts that characterize much of the research in this area to larger-scale attempts. More broadly, data-poor scholars can also benefit themselves and their claims to insights by building their work on linking multiple data sets together into a larger corpus.[36]

(4) Pooling Data

Despite the paucity of data available to most individual scholars in data-poor fields, and despite the difficulties with gaining access to data about relevant phenomena, there exists ever more data due to the twin forces of computerization and digitization. So, each scholar has some data (some more than others), and one possible step is to begin pooling this together. Certainly, there are many issues with respect to pooling data such as ownership, institutional review, compatibility, source distortion and other forms of bias, etc. Many of these issues are not unique to data-poor fields, and they are being discussed and dealt with in other areas of science.[37] Simply, while pooling can be problematic, the benefits to individual scholars, the research community, and possibly society, warrant more attention to pursuing this action.

Data-poor scholars can choose to pursue individually several of these actions such as better developing the macro/micro linkages in their research, being more directed towards doing theory elaboration, and to leveraging mixed-data set expertise. Further, scholars, collectively, in data-poor fields can build institutional capacity in ways such as support for pooling data and encouraging mixed-data set work. And, one approach does not constrain another as all need to happen. Data-poor scholars may not desire data wealth, and wealth may not always be possible. However, most of these scholars can certainly do better than simply sustaining data poverty.

Designing Cyberinfrastructure for Data-Poor Fields

The premise here is that scholarly fields will always differ relative to their access to and uses of data. This means that scholarly practices, tools and uses of resources will also differ. The current focus of, and rhetoric about, cyberinfrastructure is too focused on large data sets and attendant computational tools, with much less attention towards data-poor fields. To counterbalance this, here I articulate three design considerations for cyberinfrastructures to support data-poor fields of study: (1) focusing on supporting multiple and often small-scale collaborations; (2) attending to data curation, and in particular support for mixed-form data sets; and (3) designing for flexible, if not simple, architectures.

Supporting Collaboration

Evidence shows that social scientists tend to work in relatively small groups (relative to those in natural and physical sciences) and to pay greater homage to sole-authored work.[38] I argue that data poverty provides additional incentives to work alone or in small groups. This suggests that any useful cyberinfrastructure for data-poor science must be designed to allow for small groups to collaborate and for solo-scholars to benefit. Moreover, these collaborations should be seen as short-term and likely with restricted access for non-participants. These collaborative spaces must allow for both private and public data, and accommodate the research team's desires to share, or not, all or some of their data.[39]

Curation for Heterogeneous Data

Data-poor fields are likely to have many types and forms of data. Moreover, there is likely little collective competence in data-poor fields with managing data sets, preserving data (or its instrumentation) or doing any of the data curation and meta-data tagging involved in running data repositories. Thus, cyberinfrastructures for data-poor sciences will need to develop the technological infrastructure and explicit practices to support and manage distributed, heterogeneous and small data sets. I see this as something much like special collections in contemporary libraries, but all digital and mostly distributed. A cyberinfrastructure for data-poor fields that provides a means to store, preserve and annotate small data sets provides something that no single scholar can hope to develop but most (if not all) scholars can value. Moreover, if particular software-based tools to engage these data sets are also developed, this serves as a means to support data pooling in ways not yet possible in most data-poor fields.

Designing Flexible Infrastructures

The third design consideration for data-poor cyberinfrastructure is more holistic and reflects two assumptions about the current working practices and work infrastructures of data-poor scientists. The first assumption is that most data-poor scholars rely on relatively simple computing infrastructures—personal computers, commercial or open-source software, small applications developed for their own use: essentially a commodified computing infrastructure. The second assumption is that these commodified infrastructures often rely on other resources (like the university's course-management system; local data back-up, security and loss-recovery processes/systems; ad-hoc web hosting/presence; and local technology support). In socio-technical terms, these two assumptions capture elements of domesticated technologies[40]—ones that are commonplace and that fade into the daily activities.

The domestication assumption leads me to argue that cyberinfrastructures for data-poor fields should be designed to allow for inter-operability with commercial course management systems such as Blackboard. These cyberinfrastructures should demand minimal technological infrastructures/rely on common computer systems, and should be designed in modules to allow scholars to select and draw on these resources in ad-hoc and idiosyncratic ways. Further, data-poor cyberinfrastructures should be designed with little embedded workflow and, instead, to maximize local and flexible uses.

Beyond my simple depiction of three design considerations for cyberinfrastructure for data-poor fields, what are needed are at least two additional efforts. First, there must be specific attention to developing a deeper understanding and improving data-poor research processes. Why? Because, data-poor research is done differently and has different needs than does data-rich scholarship. This suggests setting aside resources specifically to advance those scholars who work in data-poor environments (such as those exploring emergent phenomena) and yet have the potential to provide substantial contributions to science and society. Second, more work is needed that focuses on understanding the incentives (and disincentives) for using cyberinfrastructures. For instance, we need studies which contrast data-poor and data-rich sciences relative to the higher-levels elements of the cyberinfrastructure layer-cake. One-size-fits-all conceptions of cyberinfrastructure are clearly too limited and the massive investments being made for developing these demand greater scholarly accountability. These cyberinfrastructures are after-all, an emergent and data-poor phenomena.

Notes and References

1. Earlier versions of this essay were presented to the Information Systems Department at Case Western Reserve University's Weatherhead School of Management in March 2007; the 2007 Annual Meeting of the Society for the Social Studies of Science in October 2007; and as part of a panel at the International Conference of Information Systems in December 2007. Comments from conference program members and members of these audiences have helped improve this essay. This essay has benefited from discussions with Sandeep Purao, Madhu Reddy, John Jordan, Michel Avital, Kalle Lyytinen, Noriko Hara, Wayne Lutters, Eric Meyer, Carsten Osterlund, Howard Rosenbaum and Ben Light. Comments from the *Prometheus* reviewers have further improved the current version of this essay.
2. See E. Meyer, 'Moving from small science to big science: social and organizational impediments to large scale data sharing', paper presented at the *Third International Conference on*

e-Social Science, Ann Arbor, MI, 7–9 October 2007 and available online at: http://ess.si.umich.edu/papers/paper218.pdf; M. Nedeva and R. Boden, 'Changing science: the advent of neoliberalism', *Prometheus*, 24, 3, 2006, pp. 269–81; R. Schroeder, *Rethinking Science, Technology and Social Change*, Stanford University Press, Stanford, CA, 2007; T. Hey and A. Trefethen, 'The data deluge: an e-science perspective', in G. Fox and T. Hey (eds), *Grid Computing—Making the Global Infrastructure a Reality*, Wiley, London, 2003, pp. 804–24.

3. The vibrant debates on what data are, how to measure, and how to use data are beyond the scope of this essay. For more on this see the work of L. Floridi, 'What is the philosophy of information?', *Metaphilosophy*, 1&&2, 2002, pp. 123–45; I. Cornelius, 'Theorizing information for information science', in B. Cronin (ed.), *Annual Review of Information Science and Technology*, 36, Information Today, Medford, NJ, 2002, pp. 393–425.
4. See www.cyberinfrastructure.us. Moreover, the rapid growth in data collection in these fields stems in large part from using ICT to collect data in more sustained and larger-scale ways.
5. See C. Anderson, 'The end of theory: the data deluge makes the scientific method obsolete', *Wired*, 23 June 2008 and available online at: www.http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.
6. For more on this field, see www.aisnet.org and www.isworld.org. A rough estimate is that there are about 5,000 people, with about half being research-active, worldwide members of this research community.
7. The forum for this debate, however, is found in discussions of what is known in the United States as Cyberinfrastructure (which is called E-Science in Europe and perhaps the rest of the world).
8. See C. Burton, *Places of Inquiry: Research and Advanced Education in Modern Universities*, University of California Press, Berkeley, 1995.
9. Nedeva and Boden, *op. cit.*; H. Rose and S. Rose, *Science and Society*, Neguin Books, New York, 1971.
10. See S. Jackson, P. Edwards, G. Bowker and C. Knobel, 'Understanding infrastructure: history, heuristics, and cyberinfrastructure policy', *First Monday*, 12 June 2007 and available online at: http://firstmonday.org/issues/issue12_6/jackson/index.html.
11. D. Atkins *et al.*, *Revolutionizing Science and Engineering through Cyberinfrastructure*, Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Directorate for Computer and Information Science and Engineering, National Science Foundation, Arlington, VA, July 2003.
12. Much of the section draws from the presentation by Stuart Feldman (of IBM) at the 2007 NSF/OED Workshop on Cyberinfrastructure. For more on this workshop, please visit: http://www.cyberinfrastructure.us/.
13. Layer-cake models of technology assemblages are notorious over-simplifications of a much more complex and complicated set of inter-relationships, dependencies and paths. That said, there is utility in this simplification, particularly since this essay is focusing on the issues with the layers and not their relationships.
14. Current examples of this include: S. Adolphs *et al.*, 'Integrating cyberinfrastructure into existing e-social science research', *Proceedings of the 2007 E-Social Science Conference*, 10–13 October 2007, Ann Arbor, MI and available online at: http://ess.si.umich.edu/; W. Dutton, 'Reconfiguring access to information and expertise in the social sciences: the social shaping and implications of cyberinfrastructure', *Proceedings of the 2007 E-Social Science Conference*, 10–13 October 2007, Ann Arbor, MI and available online at: http://ess.si.umich.edu/; J. Ure, R. Procter and Y. Lin, 'Aligning technical and human infrastructures in the semantic web: a socio-technical perspective', *Proceedings of the 2007 E-Social Science Conference*, 10–13 October 2007, Ann Arbor, MI and available online at: http://ess.si.umich.edu/.
15. See W. Turner, G. Bowker, L. Gasser and M. Zacklad, 'Information infrastructures for distributed collective practices', *Computer Supported Cooperative Work*, 15, 2–3, 2006, pp. 93–110 for an overview.
16. See NSF/OECD, *Social and Economic Factors Shaping the Future of the Internet*, Workshop Proceedings, 31 January 2007, NSF, Washington, DC; NSF, *Cyberinfrastructure Vision for 21st*

Century Discovery, Version 5, January 2006, NSF, Washington, DC; T. Hey and A. Trefethen, 'Cyberinfrastructure for e-science', *Science,* 308, 5723, 2005, pp. 817–21. Additional technical reports from the UK's E-Science projects can be found at: http://www.nesc.ac.uk/technical_papers/uk.html.

17. See D. Rhoten, 'A multi-method analysis of the social and technical conditions for interdisciplinary collaboration', Final Report for award BCS-0129573, National Science Foundation, NSF Press, Arlington, VA, 2003.
18. See P. Edwards, S. Jackson, G. Bowker and C. Knobel, *Understanding Infrastructure: Dynamics, Tensions, and Design*, DeepBlue, Ann Arbor, 2007 and available online at: http://hdl.handle.net/2027.42/49353; Jackson *et al.*, *op. cit.*; C. Mackie, 'Cyberinfrastructure, institutions and sustainability', *First Monday*, 12, 6, 2007, and available at: http://www.firstmonday.org/issues/issue12_6/mackie/index.html; T. Finholt and G. Olson, 'From laboratories to collaboratories: a new organizational form for scientific collaboration', *Psychological Science*, 9, 1, 1997, pp. 28–36; S. Scott and W. Venters, 'The practice of e-science and e-social science: method, theory, and matter', in K. Crowston, S. Sieber and E. Wynn (eds), *Virtuality and Virtualization*, Springer, London, 2007, pp. 267–79; and C. Lee, P. Dourish and G. Mark, 'The human infrastructure of cyberinfrastructure', *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, ACM Press, New York, pp. 483–92.
19. See S. L. Star and K. Ruhleder, 'Steps toward an ecology of infrastructure: design and access for large information spaces', *Information Systems Research*, 7, 1, 1996, pp. 111–34; Finholt and Olson, *op. cit.*
20. A petabyte is 2^{50} or 10^{15}. For more on this see the series of essays in Anderson, *op. cit.*
21. Another example would be the National Institutes of Health and the National Library of Medicine. The library has always been a curator for some data sets and plays a pivotal role in medicine and medical research.
22. Of course, given the issue raised above regarding limited variations in the types of data, it may be argued that the data needed to test rival theories are just not being collected.
23. For example, the information systems research community has developed extensive online resources on methods, theories and literature. See http://www.isworld.org (and note that there are no common data sets).
24. See: K. Knorr-Cetina, 'The disunity of two leading sciences', in P. Galison and D. Stump (eds), *The Disunity of Science, Boundaries, Context, and Power*, Stanford University Press, Stanford, CA, 1994a; and K. Knorr-Cetina, *Epistemic Cultures: How Scientists Make Sense*, Indiana University Press, Bloomington, IN, 1994b.
25. For the recent book by that name: S. Leavit and S. Dubner, *Freakonomics*, Harper Collins, New York, 2005 that uses small-scale empirical work to call into question deeply-held positions in contemporary neo-classical economics.
26. For more on this see the issue of *Information Research* dedicated to key papers from the 6th conference on Information Seeking in Context, available at: http://informationr.net/ir/11-4/infres114.html.
27. For example, see J. Abowd and J. Lane, 'New approaches to confidentiality protection: synthetic data, remote access and research data centers', in J. Domingo-Ferrer and V. Torra (eds), *Privacy in Statistical Databases*, Springer-Verlag, Berlin, 2004, pp. 282–9.
28. See R. Buzzell and B. Gale, *The PIMS Principles: Linking Strategy to Performance*, Free Press, New York, 1987; G. Tellis and P. Golder, 'First to market, first to fail: the real causes of enduring market leadership', *Sloan Management Review*, 37, 2, 1996, pp. 11–9.
29. For more on NASA SEL, see https://www.thedacs.com/databases/sled/sel.php; B. Boehm, *Software Engineering Economics*, Prentice-Hall, New York, 1981.
30. For F/LOSS scholars there is www.sourceforge.org: a publicly-accessible data repository on open source software projects.
31. See Meyer, *op. cit.*
32. See S. Sawyer and H. Huang, 'Conceptualizing information, technology and people: comparing information science and information systems literatures', *Journal of the American Society of Information Science and Technology*, 58, 10, 2007, pp. 1436–47.

33. See D. Vaughan, 'Theory elaboration: the heuristics of case analysis', in C. Ragin and H. Becker (eds), *What is a Case? Exploring the Foundations of Social Inquiry*, Cambridge University Press, Cambridge, MA, 1992, pp. 173–202.
34. See W. Orlikowski, W. and S. Iacono, 'Desperately seeking the "IT" in IT research: a call to theorizing the IT artifact', *Information Systems Research*, 12, 2, 2001, pp. 121–4.
35. See R. Williams and D. Edge, 'The social shaping of technology', *Research Policy*, 25, 1996, pp. 865–99.
36. See National Academies of Science, *Facilitating Interdisciplinary Research*, NAS Press, Washington, DC, 2005.
37. Mackie, *op. cit.*; Turner *et al., op. cit.*
38. See W. Dutton, 'The web of technology and people: challenges for economic and social research', *Prometheus*, 17, 1, 1999, pp. 5–20.
39. There is also strong reason to encourage more collaboration, as Dutton, 1999, *op. cit.* argues.
40. See L. Haddon, 'The contribution of domestication research to in-home computing and media consumption', *The Information Society*, 22, 4, 2006, pp. 3–19.