Sociology in the Era of Big Data: The Ascent of Forensic Social Science

Daniel A. McFarland[1] · **Kevin Lewis**[2] · **Amir Goldberg**[1]

Abstract The rise of big data—data that are not only large and massively multivariate but concern a dizzying array of phenomena—represents a watershed moment for the social sciences. These data have created demand for new methods that reduce/simplify the dimensionality of data, identify novel patterns and relations, and predict outcomes, from computational ethnography and computational linguistics to network science, machine learning, and in situ experiments. Such developments have led scholars to begin new lines of social inquiry. Company engineers, computer scientists, and social scientists have all converged on big data, creating the possibility of a vibrant "trading zone" for collaboration. However, strong differences in research frameworks help explain why big data may not be an egalitarian trading zone across fields, but rather—at least in the short term—a moment when engineering colonizes sociology more than vice versa. In the long term, however, we suggest there may be the possibility of a constructive synthesis across paradigms in what we term 'forensic social science.'

Keywords Big data · Computational social science · Sociology of science · Forensic social science

This paper has benefited from conversations with Jure Leskovec and Kylie Swall who helped us formulate some of this material. This paper also benefitted from the feedback of Ron Breiger and John Mohr at the annual conference of the American Sociological Association in 2014. The work presented here is in part supported by Stanford's Center for Computational Social Science and NSF # 0835614.

✉ Daniel A. McFarland
mcfarland@stanford.edu

[1] Stanford University, Stanford, CA 94305, USA

[2] University of California, San Diego, CA, USA

Published online: 17 September 2015

Introduction

Science in general, and the social sciences in particular, are facing a watershed moment where data and methods are dramatically expanding. This dramatic expansion stems from novel, engineering-led, technological means of data collection and analysis. In biomedicine, scholars refer to this as the "big data" revolution (Cukier and Mayer-Schoenberge 2013; Hilbert and López 2011; Lohr 2012); in the social sciences, they refer to it by the same name or as "computational social science" (Lazer et al. 2009). As a consequence of this movement, we are witnessing the collision of distinct cultures of inquiry: for the first time, the field of engineering, the disciplines of social science, and the industry for social media are all focusing on similar types of data (e.g., digital information on social transactions) and similar types of questions (e.g., what promotes certain types of social behaviors?).

As a result of these shared data and questions, there has arisen a potential "convergence" of scientific perspectives, methods, and technologies (National Research Council 2014). This convergence means that big data has the potential to become a "trading zone" where researchers from entirely different paradigms, despite differences in language and culture, collaborate with each other to exchange tools, information, and knowledge (Galison 1997; Collins et al. 2007). In addition, this shared focus will likely produce important *theoretical* changes for the social sciences—and for sociology in particular.

But what might these changes be? Will we witness an era of intellectual recombination, an instance of paradigm formation (from pre-paradigm to paradigm; see Kuhn 1996), or a moment of colonization where the field of sociology and its traditions are subverted to other fields like computer science—including their ends and organizing structures? To understand this moment in the history of social science and to foretell the impending theory-change, we begin by briefly reflecting on the previous empirical watershed and then turn our attention to the current situation. The ensuing article is therefore divided into multiple sections concerning various facets of this transformation: how the last transformation is similar; how the current moment is redefining data and methods; how the current era is both altering our approach to old questions and posing new questions altogether; how all of these changes are occurring at the intersection of somewhat incommensurable fields (engineering, social science, and social media industry); how this intersection will have likely winners and losers who embed their perspectives in one another; and how the future will likely mean (in the short term) the colonization of sociology (and social science more generally) by computer science perspectives and practices.

This transformation will likely find a mixed reception: on the one hand, it will draw sociology into Pasteur's quadrant of actionable knowledge and afford sociologists more employable skills; but on the other, it could diminish theory to a secondary role by encouraging piecemeal (de-unified) explanations of social life. We argue that an approach we term forensic social science—by which we mean a middle ground that is both inductive and theory-oriented—might help mitigate the colonization of social science by an atheoretical scientific program (see also Goldberg In press).

The Last "Watershed" Moment

The last empirical watershed moment has been well documented by others (Camic and Xie 1994; Converse 1987). The anti-disciplinary movement of the 70's aside (Menand 2010), the trend in the social sciences has been toward increasing quantification: statistical methods from science, technology, engineering, and mathematics (hereafter "STEM") fields (Hacking 2006; Porter 1995) were imported to psychology, economics, and sociology upon the advent and general use of survey research in the 1930's and 40's (Converse 1987; Platt 1996). Statistical modeling became more prominent as computing technology advanced, making more and more sophisticated calculations feasible.

As a result of this convergence, the social sciences, and sociology in particular, moved away from ethnographic community studies and adopted a methodological individualistic perspective (Porter and Ross 2003). This perspective was evident in the nature of surveys asking for individual responses and viewpoints. It was also evident in statistical procedures that relied upon assumptions of independent observations (Agresti and Finlay 2009). With the advent of ordinary least squares (OLS) regression, social science journals became replete with regression tables, OLS equations, path models, and a variable-centric viewpoint (Abbott 1988). With survey research and the accompanying statistical models came assumptions about social actors, their interrelation (or lack thereof), and even conceptions of time (as panels). These assumptions were never believed to be an accurate conception of social life, nor a warrant for developing methodological individualism. Rather, they were acknowledged after the fact as necessary byproducts of performing regression-based research using survey data. Nevertheless, this style of research compiled further and further evidence and became increasingly central to mainstream sociological inquiry.[1]

The advent of survey research and statistical modeling also institutionalized hypothesis-testing as the predominant scientific paradigm in the social sciences. Generations of researchers were consequently trained to ask research questions in terms of null-hypotheses that they refuted with statistical evidence. This paradigmatic shift implied that theory precedes data collection and research aims to find statistical support for a preconceived set of hypotheses. Moreover, the limited availability of data and costs associated with data collection reinforced a methodological orientation that required inductive hypothesizing and relied on statistical sampling. Though there were attempts to challenge hypothesis-testing as the prevailing social scientific methodology (e.g., Glaser and Strauss's formulation of Grounded Theory in 1967), the continuous development of statistical methods and data collection capacities only entrenched its unassailable position throughout the twentieth century.

In stating this, we do not intend to judge the past transformation. Rather, our intent is merely to note that the transformation occurred and that there were clear shifts in data, method, and theory. The shifts were not endogenous to sociology as a discipline, either; they were the result of converging fields spanning statistics, polling/opinion research, and social science more generally. They were also the result of business and industrial demands, such as contracts from the military to study the American soldier (Stouffer

[1] This approach also had an elective affinity with certain social scientific theories over others: for instance, rational choice theory (Coleman 1994a) was arguably more readily translated into the data and methods of the time, than say the more abstract theories that preceded it, such as structural functionalism.

1949), government funding to perform a census (Anderson 1988), Good Society efforts to transform society (Coleman 1986), and private industry pouring resources into marketing research. All these efforts shared a focus on similar types of data and similar types of questions and called for the complementary use of diverse expertise; a network of institutional supports appears to have then catalyzed this research focus and intellectual development. In short, survey research became a "trading zone" across multiple social domains (Galison 1997; Collins et al. 2007)—one that ultimately transformed the social sciences.

The Current "Watershed" Moment

Sea changes in research paradigms typically arise from fundamental changes in the nature of research content—and along with it, the convergence of various intellectual partnerships and the networks used to support them. As a consequence of big data, the research content of social science is going through such a change. Big data are a new style of *data*. With that comes an assortment of new analytic techniques and methods that render these data into novel *information* about social phenomena. Inevitably following this is a change in theories that take the available information and render it into *knowledge*, or narratives explaining how social phenomena occur. In this manner, the new watershed in social science research is already happening as a gestalt switch across different thought styles, or complexes of data-information-knowledge (Fleck 1979). With it will come a shift in the thought community, or the networks of partnerships and resources of which this community is composed. In this section, we explain the current changes in data and information, and then examine in ensuing sections the implications for sociological theory.

New Data

The current empirical and methodological watershed has been coined many things, and most prominently "big data." In many regards, this label is incomplete and does not adequately capture the full array of changes that are transforming the nature of data. Perhaps the most fundamental is a shift in data collection toward digital records of every kind. This is due to changes in industry and technology that have rendered digital records ubiquitous. Today, a growing number of organizations retain digital data on millions of persons and their moments, and much of social life is mediated by technology that retains a digital record of every action. We live in an "age of engineering" where we rely upon technology for a staggering variety of tasks (Brown and Duguid 2002).[2] The types of data that are collected and the range of contexts from which they are sourced are remarkable. For brevity's sake, only a few can

[2] However, we are not necessarily living an age of science. By this we mean that we are black-boxing information and knowledge in tools and treatments that bring about desired outcomes without ever really understanding why or how they do so. While in the past, scientific facts were black-boxed due to their complexity (Latour 1988), we now find ourselves in a time when we often seek solutions without any concern or desire for explanation at all. Not everyone wants this; but rather, the prevailing pressures of industry, engineering, and practical concerns of life in a technology-mediated age demand this. Scientists will still seek explanations—but their voices may remain an (increasingly small) numerical minority.

be noted here: robotic medical devices that retain data on every movement; insurance and hospital records of patient health and treatment; credit card records of every transaction; online social media retaining every click, communication, and timestamp; internet protocol (IP) addresses registered on every cell tower; pixel data on videos and images; WAV files for voices; all sorts of individual judgments on products (stars) and other individuals (reputation points)—the list goes on and on. All of these data are catalogued (or "scraped") and amassed, collectively constituting not just one, but countless "fire-hoses" of potentially continuous information.

Suffice it to say, the term "big data" only begins to capture the richness, variety, dynamism, and massively multivariate nature of the data now being collected. The new digital data often concern relational events like transactions, and of many different kinds. They also concern expressions of meaning like texts that arguably convey multiple dimensions of information about the actor (see Bail 2014). The data are not (only) information expressed by a particular person placed in a particular (narrowly delimited) relation with a researcher. Most of the data concern actual social behaviors and information about persons going about activities in their daily lives; they are "digital footprints" of human activity and interaction (Golder and Macy 2014). In most instances, the data are not mediated by a survey, but by a technology, device, or interface; these devices are created not for the artificial use of research but for purposes individuals or social institutions naturally select.[3] In other words, these data do not necessitate a research-driven hypothesis in order to be generated.

Big data, and the computational advances they entail, therefore present a paradigm shifting opportunity because they remove, or at least significantly attenuate, two significant limitations that severely constrain traditional statistical modeling-based sociological analysis. First, they provide access to data about basic social behaviors that have always been practiced, but hitherto have been rarely documented. Second, they obviate the need to draw representative samples in contexts in which the complete universe in question—such as all transactions between buyers and sellers on an electronic marketplace such as eBay—is documented.[4] These enormous, unstructured corpora of data require immense computational horsepower and prowess to be

[3] This feature of big data, in particular, is a curse as much as a blessing. Several years ago, Lazer et al. questioned whether "computational social science could become the exclusive domain of private companies and government agencies" (2009:721). Equally problematic as questions of ownership and access, however, are related concerns about data quality and interpretation: in eliminating the participation of the academic researcher, we also eliminate the guiding force that orients data collection towards the pursuit of knowledge rather than the maximization of profit. The data that are collected in industry are not always the data that are most useful for science (as we elaborate in great detail below); worse, too seldom acknowledged is the basic observation that technologies constrain as much as they enable—and so any given dataset may tell us less about human agency and more about interfaces and algorithms that subtly influence user behavior (cf. Lewis 2015).

[4] An additional dimension to these new types of data relates to behaviors that are made possible by digital intermediation and hitherto did not exist. For example, in pre-internet times, people were simply technically unable to share photos on the scale and frequency they do today. These types of technologically enabled social transactions are a specific category of behaviors, some of which may (or may not) significantly affect social dynamics and structures. Though the same technological advances that make big data possible enable these new categories of data, these advances are, in principle, no different from previous technological and ideational transformations that catalyzed social change (such as the invention of the printing press or the emergence of the formal organization). In that respect, data generated on digitally-mediated platforms that represent new categories of social action are no different from other phenomena of sociological interest.

meaningfully tamed and made intelligible. The advantage in this disarray lies, however, in the fact that researchers can now inductively build theory from the ground up, rather than presuppose it and collect the data that will either support or refute it. We discuss the implications of this potential below.

Even so, larger-broader-richer data does not mean we now have data that are perfectly accurate, perfectly generalizable, or directly reflective of social life, to say the least; ethnomethodologists have provided a long line of critique in this regard. This new type of data is “found data” that is often prone to error and bias (See McFarland and H.R. McFarland in press). For example, each “big” dataset tends to be a single “dive” into one electronic platform and one dimension of social activity (e.g., Facebook and friendships; Twitter and affective expressions; the Web of Science and research articles). In addition, these datasets frequently consist of biased selections of individuals, including persons who have access to technological devices (e.g. smart phones and the internet), persons who use these technologies more (e.g. extroverts), and persons who generate the particular types of records being accumulated (e.g. academics who write articles, not books). As scholarship moves forward, there needs to be a discussion about the social and temporal boundaries of these new data and whether they reflect nominal or realist notions of a phenomenon (i.e., boundaries attributable to a technology’s data collection or boundaries actually perceived by social actors; see Bender-deMoll and McFarland 2006; Laumann et al. 1983) or one or more dimensions of a social system. Moreover, we need to think critically about what sort of data represents the greatest potential for scientific progress. A potential way forward would be to match data points generated by the same actors across multiple online platforms and contexts—though naturally, such an endeavor represents significant computational, legal, and ethical challenges.[5]

One question worth asking is whether the scale, breadth, and depth of this information will offset concerns about sampling bias and missing data and still afford valid inferences? For example, González-Bailón et al. (2014) consider the sampling bias introduced by collecting data through publicly available APIs. Based on samples of Twitter activity surrounding the same political protests, they find that the structure of sampled networks is significantly influenced by both the API and the number of hashtags used to retrieve messages. Meanwhile, Wang et al. (2012) identify a variety of different types of systematic missing data in large-scale corpora, such as different persons with the same names, the same person having different names, and persons underrepresented in the data. How much missing data and error undermines our results? Do large samples correct for this—and under what circumstances? Answers to these questions are needed if we are to believe the stories big data tell us.

On the other hand, one could argue that the *kind* of data that are collected and the nature of insights gleaned from their analysis could be well worth the tradeoff in generalizability. In other words, detailed, moment-to-moment behavioral data—even if they are observed for a small and/or non-representative population—are surely preferable for answering *some* types of research questions than are survey data from

[5] For our part, we believe the linking of multiple corpora for an entire domain will bring the greatest advances to the social sciences. With rich, multifaceted data for an entire social system—of, say, politics, a market, or academe—we can ask and answer a variety of social science questions with less concern of confounding, missing data, and selection bias (see Coleman 1994b).

even a perfect random sample of respondents (for an early discussion of this tradeoff in the networks literature, see Rogers 1987). For certain applications, surveys may offer the best available measurement of some theoretical construct we are interested in (cf. Vaisey 2009); for others, they offer an immensely impoverished representation of social reality. Ironically, then, the new massively multivariate data may share more in common with the deep ethnographic research of the Chicago school than the large-scale quantitative work of the last watershed; we elaborate on this affinity below.

In summary, it is difficult to deny the advantages of using these new types of data. The size of the new datasets may offset some of their error and systematic biases in ways that remain to be fully explored; but the richness and breadth of the new data undoubtedly offer more information with greater nuance than we have had in the prior generation of research. Further, these new types of data will likely reveal new ways to ask old questions the prior "paradigm" of surveys/methodological individualism left unanswered and open up a multitude of entirely new questions about social life that the prior research framework is unable to identify or address.

New Techniques

It is well known that computational power and storage have grown over the last century, making quantitative analyses possible for more and more data. But today's datasets have become so large, so longitudinal, and so multivariate, that some old statistical techniques are impossible to apply (known as "NP complete") and no amount of computing power could enable the algorithm to work; meanwhile, other, more recent developments in quantitative methods (especially those that employ simulation, search, and/or combinatorial techniques) are prohibitively inefficient and potentially untrustworthy at massive scales (insofar as they rely on assumptions that are severely violated). On the other hand, precisely because these new data require new ways of handling them so they can be rendered into *information*, the expansion of big data has spawned a number of new, exciting, and rapidly-developing computational techniques as well as galvanized traditional methods.

The most straightforward and common approach being used can be called "**computational ethnography.**" It is ethnographic in the sense that researchers proceed by a process of induction akin to the Glaser and Strauss (1967), but here scholars can share their data and potentially reproduce each other's results. In many published computer science proceedings, this is the "science in the making" rather than the "science presented" (Latour and Woolgar 1986). The discovery process begins with an initial hunch or intuition for how a social phenomenon arises. Then the researchers look for patterns in the data that build into a narrative they can employ to explain that phenomenon. Given the massively multivariate nature of the data, scholars employing this approach begin by examining trends and testing a variety of cross-tabulations and correlations to see what is going on in their data. When they find an interesting result, they illustrate it in a variety of ways—much like ethnographers do when they inductively arrive at a theory from seeing patterns in their observations—and then reassess and develop their theory with continued observation. Such examples can be found in many of the descriptive studies using big data (Golder and Macy 2011).

One clear example can be observed in a computational history presented on the Association of Computational Linguistics (Anderson et al. 2012). There, the researchers were familiar with the field's development from performing an extensive literature review (Jurafsky and Martin 2009) and from interviewing key participants in the field, but they wanted to illustrate the field's transformations via the available textual and relational data found in the association's many thousands of compiled publications. As such, they used vast amounts of behavioral records and identified a variety of language trends, network trends, and field turning points that affirmed participants' observations and even gave them further nuance. In this manner, a form of computational ethnography was employed to flesh out a rich social intellectual history.

That said, scholars are doing more than probing rich data like ethnographers. A new set of methods has arisen to confront massive, multivariate, interdependent data acquired from a variety of sources, and especially from links (relational ties and events) and text (meanings). The basic questions being asked of these data are questions of simplification and sense-making: how do we reduce the raw data to manageable but still meaningful dimensions—particularly without sacrificing their richness—and what kinds of patterns can we discern? At least four varieties of new techniques are being used to accomplish this. One variety of method has arisen in the field of *computational linguistics* and is used to identify patterns in speech and texts (Jurafsky and Martin 2009; Manning and Schütze 1999). A second variety has arisen in *network science* (a confluence of social network researchers across computer science and social science) and is used to identify patterns in large-scale, dynamic linkages (Brandes et al. 2013; Easley and Kleinberg 2010; Newman 2009). A third variety of research concerns algorithms employing these features (and others) to predict various outcomes. For lack of a better term, many of these algorithms adopt a *machine learning* approach, albeit one that is often augmented with simulation modeling (Alpaydin 2004). A fourth approach stems from human-computer interaction and uses *experimentation* to take advantage of large-scale, real-time manipulations of user experiences on social media platforms so as to identify causal relations (Centola 2010; Dodds et al. 2003; Salganik et al. 2006).

Computational linguistics is a field long in existence that arose from the overlapping pursuits of linguistics, artificial intelligence, and cognitive science when they sought to develop mathematical models for machine translation (automatic language translation). A large amount of government funding and industry support catalyzed this field by briefly drawing in engineers whose statistical methods greatly outperformed the efforts of linguists. This influx of probabilistic models transformed the field and established a line of research that continues today (Anderson et al. 2012). Computational linguistics now affords technologies useful to a variety of internet and web-based industries. The field has not only created techniques for rendering speech into quantifiable information (e.g., pitch, loudness, discourse markers), but also methods for identifying clustered uses of text, sentence recognition, and much more (e.g., with topic modeling or latent Dirichelet allocation, deep neural networks or learning models). For social scientists, this provides a fleet of invaluable methods for linguistic analysis, many of which bear a semblance of interactional and sociolinguistic theories. In particular, topic modeling renders vast reservoirs of text into sub-languages, or bundles of words frequently used together (Blei 2012; McFarland et al. 2013a). This technique, along with distance metrics on how close texts are to each other, is increasingly used to simplify and make sense of large bodies of text.

Network science is a field that has merged together the network analytic efforts of computer scientists, physical scientists, and methodologically oriented social scientists who study social networks (Brandes et al. 2013; Easley and Kleinberg 2010; Newman 2009). In many instances, the computer scientists have replicated prior social network research but on a larger scale (and often without recognition). Moreover, they have shifted the focus toward community detection, simulation, mathematical modeling, and hypothesis testing (as opposed to social scientists' traditional interest in observational research, static networks, structural properties, and small-scale settings; Borgatti et al. 2009; Wasserman and Faust 1994). Many of the network motifs being identified in this interdisciplinary domain are readily applicable to all sorts of new social media and website data. These studies have focused on website linkages (Barabasi 2003), e-mail exchanges (McCallum et al. 2005), and relational data drawn from an assortment of social media, from following ties on Twitter (González-Bailón et al. 2011) to friendships and picture-postings on Facebook (Lewis et al. 2008). Perhaps the most promising line of work attempts to take the mass of links and render it into identifiable clusters of linked nodes (Leskovec et al. 2010; Newman 2001; Newman and Girvan 2004). These techniques touch on fundamental sociological questions relating to the social categories that structure social interaction and have been increasingly used to make sense of large bodies of linkages. Many in industry have also used these features to predict a variety of outcomes important to business.

In internet companies there are large assortments of engineers focused solely on improving prediction. For example, a company may have many records on their website's usage, such as information on clicks, site referrals, posted texts, network positions, as well as time spent on each page, purchasing behavior, and product quality ratings. Most companies ask their engineers to take these data and develop models informing the company when a customer service effort, advertisement, or search result will lead users to purchase more goods, stay on the site longer, or otherwise become an ideal consumer. The engineer typically approaches this problem without any concern for theory and instead applies **machine learning** (Alpaydin 2004; Bishop 2007).[6] This proceeds when the engineer takes, say, half the collected data (the "training set") and identifies a variety of user actions (and their timing) most associated with a desired outcome. In effect, the engineer trains a logit model on an outcome of interest and throws as many features (variables) as possible at it in order to develop highly predictive weights. Then the engineer utilizes these weights to create an algorithm and assesses whether it can accurately predict the desired outcome in the remaining data that were not used for training (the "test set"). When the algorithm reaches certain levels of accuracy, the engineer can use it to determine which users need a "push" so as to proceed in desired directions.

Machine learning is a powerful tool that has assisted companies in many domains with various engineering questions (Talley et al. 2011). In fact, machine learning is the foundation of machine translation: for instance, how Chinese is accurately translated into English. The algorithm proceeds by merely identifying common word sequences across verified translations, and as the "training set" of

[6] A note of caution is in order here: by no concern for theory, we are referring only to theories that relate to *explaining* the social phenomenon in question. There are, of course, multiple statistical assumptions, informed by theory, that are embodied in the data-mining algorithms being employed.

known translations grows, so does the probability of accurately “predicting” word associations in “test sets” of future text. The “theory” is nothing more than probabilities of word associations as identified by many known translations. There is no linguistic theory of how a language or translation works (in fact, theories are notoriously inaccurate at translation). All that is desired is an accurate translation—utility is paramount, while understanding and explanation are superfluous.

When the machine learning approach is combined with theory and scientific research it can lead to surprising results. The atheoretical perspective of machine learning can reveal patterns a theory did not predict or a new way to formulate the theory that perhaps the analyst had overlooked. However, machine learning on its own (and by design) results in little to no understanding if there is no effort to derive a theory or explanation. In sum, the use of machine learning is atheoretical, but it is potentially powerful when used as an agnostic search for potential explanations. In contrast, theory is a somewhat narrow-minded but powerful tool in that it is a focusing device that identifies which constructs are to be selected and formed from the millions of possible variables (or features) and it afford potential explanations for how features interrelate. As such, the iterative combination of atheoretical induction and theory-led deduction can be quite powerful.

In spite of the effort to identify patterns in content and links—and to develop predictive models via a theory-augmented machine learning approach—many still feel that quasi-inferential explorations of big data are too messy and complex. A large number of companies and scholars have instead become concerned with causation and the establishment of **experimental methods** to identify what changes in a social environment result in a desired effect. In many cases, these experiments take place in studies of human-computer interaction, where internet platforms provide ideal conditions for experimental control (Kohavi and Longbotham 2007). Online experiments have the benefit of controlling most conditions and estimating the effect of a single treatment or intervention on large samples of individuals. Consequently, many companies have adopted experiments as the gold standard for evidence before deciding to alter their products or add a new feature. The same is arising for scientists studying social media and big data. Increasingly, academic journals view this type of test as more conclusive than others. That said, many such experiments are run quickly and then the product changes and/or the researchers move on; there is strikingly little effort to perform second studies or to assess the robustness of results. Moreover, the narrow findings of individual experiments are seldom coordinated to produce broader, synthetic contributions to current knowledge.

In sum, the shift toward big data has created a new watershed moment where the nature of data collected and the analytic techniques used to establish information and scientific facts are shifting. With this comes problems, and we have alluded to some (see also Boyd and Crawford 2012). At the same time, we are witnessing the arrival of new lines of inquiry—including opportunities to pursue old questions through new means and opportunities to ask new questions in a world awash in big data. From understanding these shifts in inquiry, we will gain further insight into how theory and knowledge in the social sciences will likely shift in turn.

New Lines of Inquiry

As the availability of these new data and these new methods grows, many old social science research questions will be approached anew from novel angles.[7] In addition, the plethora of data—and living in a world of big data access—has created a whole new set of questions begging to be addressed. These shifts in inquiry correspond with a theoretical shift, as well.

An example of an old social science question concerns the *successful functioning of democracy through dialogue*. Today's researchers have access to all sorts of relevant new data and analytic techniques. Political scientists and pollsters alike can examine debate recordings, transcripts, blogs, opinions, C-Span recordings, legislation, tweets, ongoing polls, and betting on vote outcomes (Grimmer et al. 2014). In addition to records of what is said when, where, how, and by whom, there are sometimes even streaming data on how viewers react to a debate, event, or speech as it unfolds. In short, we now have the means to discern what makes for better or worse outcomes in political dialogues, where "better" or "worse" can be defined in any number of ways (for example, see Backstrom et al. 2013).

Another old social science question is *inequality.* No longer confined to census records and surveys, researchers can observe how people move in their daily lives (Anderson 1988)—but on an unprecedented scale. They have records of when people come into contact with different cell phone towers and where their phones and computers register (at least among those who own cell phones and computers—respectively, 85 % and 78 % of American adults as of November 2012[8]). They know which Wi-Fi devices are consistently utilized. They can track personal movement via global positioning systems on phones. They know a great deal of information regarding where, when, and to precisely what extent people are actually segregated—not just whether their responses and residences indicate it. Additionally, a variety of traditionally collected records from the census, police, and health and county services are now being linked. In this manner, researchers with big data can assess inequality anew in terms of segregation, behavioral patterns, and access to a variety of societal provisions (Bruch and Mare 2012).

A key concern of economic science and business is *value creation and profit acquisition*. With the advent of online markets like eBay—and even "virtual economies," e.g., in massively multiplayer online games (Szell and Thurner 2010)—companies collect data on auctions and naturally occurring experiments where sellers try to sell the same product by a variety of means (Einav et al. 2014). For example, on eBay, people experiment with the same product (like a golf club) to see if it sells better via an auction or as a fixed price. This information is empirically observed and performed by the users. It is neither represented in a math model or simulation nor contaminated by the presence of a researcher. These natural occurrences of different market strategies afford economic insights on profit acquisition and marketing we have never had access to before.

[7] "Novel" often amounts to "more comprehensive" as system-wide, societal-wide, and even planetary-wide data are increasingly available (e.g. Leskovec and Horvitz 2008).

[8] http://www.pewinternet.org/data-trend/mobile/device-ownership/

With new data and new technologies also arrive new research questions and problems. A prominent problem now confronted by researchers in industry and academe is that of *information overload* (Brown and Duguid 2002). With growing bodies of publicly available data and even larger troves of private data, there is pressure to mine it for useful information. But what information is useful? For companies it is what works—or what creates a desired outcome; but for scientists it is what fits—or what answers an important question. In many cases, companies are beginning to link their data with those of public datasets, thereby creating richer information on consumers. Scientists hope to use such linked datasets to address issues of the public good, such as noticing where policies work.

One potential benefit of expanded and linked datasets is that they provide a *more complete view of entire social systems and markets* than heretofore. For example, if insurance companies linked their datasets on policy members, they would quickly gain insight into how policy-holders move and select different health plans and insurance companies. Economists would have a treasure trove of information on a single market. Similarly, the linking of academic corpora—e.g., via Web of Science, dissertations, patents, NSF/NIH grants, or Google Scholar—could afford a more complete vantage on behaviors in the academic knowledge domain. From these social systemic perspectives, we can begin to assess the presence of various micro–macro processes, and relate them more fully to prior social theoretical endeavors (both classical and contemporary). This linkage would be inherently sociological, as it would afford an unprecedented opportunity for understanding how social systems operate as systems. Here we have the opportunity to observe the same actors in different contexts and social arrangements, thus parsing out institutional and socio-structural variation across different domains.

As we increasingly develop datasets on entire social systems, the concern will shift toward *how different social domains interrelate*. For instance, researchers are already asking how purchasing and profit behaviors influence political dialogue (Grimmer et al. 2014). Here, political action committee (PAC) contributions have an effect on politics, so we begin to see how economic and political domains interrelate and collide. In many regards, this focus will bring us back to key social-theoretical concerns about how a society functions from the interrelation of different domains: political, economic, social, and cultural.

With information expansion to the level of social systems combined with exponential accumulation of data on individual behaviors, there follows increasing concern with *data privacy and ethics* (see Golder and Macy 2014). In this new age of social media, digital records, and many trillions of transactional events, new technologies and modes of analysis can identify traces of activity belonging to most of us.[9] While much of these data are public, in combination and linked to one another they allow researchers to make somewhat invasive predictions about our lives and actions. In some instances,

[9] Naturally, conclusions from big data will always require qualification insofar as 1) the "digital divide" persists and 2) usage patterns of a given technology are differentiated even among those who can access it (see Lewis in press). Nonetheless, digital—and especially mobile—communications technologies are diffusing at a staggering rate (e.g. Castells et al. 2007); many population-level datasets are compiled by government or other record-keeping organizations and inclusion is not biased by "self-selection"; and given our unprecedented reliance on technology for communication, information retrieval, and relationship formation and maintenance (Bohn et al. 2014; Rosenfeld and Thomas 2012; Sparrow et al. 2011), the sheer size of available data and the proportion of humanity to whom it pertains is staggering.

these may be embarrassing and potentially damaging. In other instances, the information may be used to accurately foretell our social behaviors in ways that both for-profit and non-profit institutions will want to use to serve their goals. What are the implications of increased information about our lives and analysis of that information for institutional—as well as academic—ends? Questions about privacy and the ethics of big data will only grow stronger over time.

Big Data as Trading Zone

These new questions and problems will be the focus of social scientists as well as engineers (computer scientists) in both academe and industry. Researchers from each of these domains have become concerned with the same types of data, and they will use similar sets of analytic techniques that render these data into usable information. As such, big data have the potential to become a "trading zone" (Collins et al. 2007) across domains and to create a convergence in the intellectual and social networks undergirding them (National Research Council 2014). Should this happen, the new watershed in social science research will come to fruition.

The common view of trading zones is that they equally benefit each participating domain—so computer scientists benefit from engaging with social scientists as much as vice versa. But is this really so? Do these fields share a common research culture, and an overlapping set of research interests, so as to make exchange possible? Or do they have incommensurable views that will make exchange difficult? Can we expect their shared focus on big data to result in equal, reciprocal forms of exchange where theoretical perspectives and research frameworks are exported in either direction? Or is it more likely that some domains will take more of a lead—colonizing those who follow?

In the following sections, we describe how these domains—social science, on the one hand, and engineering/industry (we collapse the two for simplicity's sake), on the other—adopt very different frameworks and cultures of research. In so doing, we present short examples illustrating how the approaches of each domain imply distinct research paradigms.[10] In addition, we describe how the locus of resources, societal demand, cultural legitimacy, and resultant energies will likely privilege the more applied, atheoretical perspective of engineering/industry and lead engineering frameworks to colonize the social sciences—at least in the near term. In the final section, we speculate on the precise nature of the colonization process that will take place.

Prior work identifies STEM fields as having distinct research cultures from the social sciences and humanities (Kagan 2009; Snow 2001). However, in many ways, the field of engineering is distinct from the pursuit of science in general (Stokes 1997). In Table 1 below, we lay out a brief caricature of the differences between a social scientific research approach and that of engineering with respect to big data. In many regards, the

[10] The term "paradigm" may be too strong for the social science disciplines, as they often lack a shared set of standards and questions. In fact, Thomas Kuhn regarded them as pre-paradigmatic (1996). That said, we maintain it is still reasonable to regard social science and engineering as entailing different research frameworks or distinct gestalts of epistemology and research activity.

Table 1 Comparison of Distinct Research Cultures

	Social Science	Engineering/Industry
Goals	• Search for explanation and why something important happens • Develop theory to advance knowledge	• Search for accurate and novel prediction of what happens • Create algorithm/tool to make an accurate prediction
Perspective	• Focus on an explanation • Theory-driven	• Focus on what works/predicts useful outcomes • Applied
Beliefs	• Respondents and data are biased—validate	• Respondents and data have ground truth cases—align to them
Practices	• Few authors (monastic science) • Long journal papers (30–80 pgs.) • Long theory section • Long publication cycles (~2 yrs.) • Long memory/distant citation common (5–50 yrs.) • Low consensus (pre-paradigmatic) • Disparate lineages of work with internal synthesis • Conceptual fusion	• Many authors (team science) • Short conference papers (4–10 pgs.) • No theory section • Short publication cycles (3 mos.) • Short memory/recent citation common (1–5 years max) • High consensus (paradigmatic) • Research fronts with interrelation across projects and little synthesis • Task modularization

approach adopted in industry is consistent with that of engineers in that it is applied, and that some form of utility (often profit) is the ultimate end or problem being solved.

Social science (and sociology in particular) aspires to be a paradigmatic science, and as such, our main goal as social scientists is to explain why something happens. With that comes knowledge and understanding. From this stems a general perspective focused on explanation; the use of theories as possible narratives; and the assessment of these theories in observed phenomena. As we discussed earlier, the advent of statistical modeling privileged scientific hypothesis testing as the ultimate approach toward theory validation. Whereas inductive approaches have been prevalent in certain subfields in sociology—especially those that rely more consistently on ethnographic work—the division of labor within the overall field has been mostly that insights drawn from qualitative work inform theory building, which is consequently validated through hypothesis testing.

A variety of research practices also define the culture of social science (and sociology in particular). Many of these practices are common knowledge and generally well known relative to the practices of STEM fields. For the most part, social science research entails few coauthors and remains more of a monastic science in comparison to STEM fields. Our papers tend to be longer, with extended theory sections and long review and publication cycles that can last several years. Our use of citations also tends to extend further into the past than in STEM fields, and our references are more idiosyncratic and targeted toward small clusters of research (Shi et al. 2010; Shwed and Bearman 2010). In short, the social sciences are generally slower, more theory-focused, and entail less consensus than STEM fields. What we find in social science are usually disparate lineages of work, with internal synthesis and occasional dialogues across them (Levine 1995).

By contrast, engineering has a very different research culture (see Table 1). The goal is to predict and solve applied problems, typically through some sort of technological product. The perspective adopted is atheoretical, and the focus is placed on creating some product that works at solving an applied problem. In many cases, engineers believe there is a "ground truth" or "gold standard" to which they can train models and align their solutions. This can vary from an expert viewpoint, a desired outcome, or other possible anchors. When applied to social settings, such a "social physics" (Pentland 2014) approach underestimates the possibility that these outcomes are in and of themselves socially constructed and therefore endogenous to the process they are used to evaluate (Berger and Luckmann 1966). Such ground truth explorations are therefore, by definition, subject to biases and problems of validity.

The culture of engineering also has very distinct research practices. In contrast with social science, they conduct team science with many collaborators. Their papers tend to be conference proceedings (as opposed to journal articles or books) that are very short (4–10 pages) and written in short time spans. They seldom have a theory section or even much of an introduction. Their review cycles are quick and their citations seldom extend further back than 5 years (Shi et al. 2010). Over time, engineers have also developed increasing consensus in techniques and standards as evidenced by the greater density of their citations (lessened modularity; see Shwed and Bearman 2010). The collaborative nature of engineering research cultures also implies that research teams are required to solve significant coordination challenges. This is often done through task modularization. Such modularization is ideally suited to solve large data processing challenges efficiently and quickly, but at the same time makes ideational fusion that results in new theoretical propositions less likely to emerge. In sum, engineering is a fast-paced, collaborative research enterprise concerned with generating products that work and building a fleet of these approaches. These fleets resemble fronts of research with little historical synthesis.[11]

A few examples will help illustrate these distinct research cultures more clearly; we list them in summary form in Table 2, below. Take the case of language translation mentioned earlier in this paper. This topic was the focus of linguists and computer scientists, and eventually spawned the field of computational linguistics. The main problem of language translation was how to automate the procedure. On the one hand, linguists had theories of language and translation they tried to operationalize into tools (artificial intelligence and math models). While these theories afforded explanations, they did not perform terribly well as automated translation devices and tended to be terribly complicated. On the other hand, engineers approached the problem purely atheoretically. They used machine translation based on machine-learning techniques. The procedure was to take as many translated texts as possible and identify via probabilistic techniques those word associations that were most likely given the localized context (bi-grams and tri-grams) of a word's placement. These models heavily outperformed the theoretically derived linguistic predictions and we see them operationalized via many automated translation devices (e.g., Google Translate).

[11] Distinct kinds of training and opportunities for employment are also important, but brevity requires this to remain a caricature of the two perspectives. We merely intend for the reader to grasp these differences on an intuitive level, so she can see incommensurability as one reason for less reciprocal forms of exchange in the trading zone of big data.

Table 2 Examples of Distinct Approaches to Shared Problems

	Social Science	Engineering/Industry
Example 1: Language translation	• Theory of language predicts translation—poor job, some explanation	• Probabilistic model predicts translation—more accurate job, no explanation
Example 2: Date selection and satisfaction	• Present different theories of compatibility (rating-dating) or making a match	• Take online dating data from company—and identify those items most predictive of satisfactory choice and date
	• Render data into information useful for testing these theories	• Develop algorithm so as to afford users better matches
	• Can mildly predict and give explanation why	• No explanation why, but can predict
Example 3: Murder investigation	• View scene with theory guiding what you see/look for	• View scene as aggregation of tons of features/evidence
	• Test theory by collecting evidence for or against it	• Develop accurate prediction in this case based on features associated with known outcomes in the past

Nonetheless, the models afford little explanation or discussion of *why* languages operate or translate the way they do, and linguists frequently argue that computational linguistics has become nothing more than a subfield of engineering and does not resemble a science that contributes to our body of knowledge. Therein one can see divergent research cultures (Anderson et al. 2012).

Let's consider a second example of mate selection and satisfaction—taking online dating as an (increasingly popular) application (Smith and Duggan 2013). Both social scientists and engineers are confronted with data procured from users of a technology or website. Social scientists will consider different theories of compatibility and mate-selection, such as matching models, rating-dating optimization, and so on. Then they identify data on dating outcomes that they can render into information capable of testing these theories. In many instances, social scientists do a decent job of predicting mate selection and likelihoods of satisfaction, and we can find these results in a variety of journal publications (see McFarland et al. 2013b for review). By contrast, engineers (and especially engineers in industry) are asked to study a key outcome like user choices or date satisfaction because it will help the company retain users and increase its profits. Here, they will take all the company's collected data on time per page, clicks, survey responses on every item, etc., and use them as features predictive of said outcomes in a training set (portion of data used to develop weights); they then use the weightings from the predictive model to identify the accuracy of their predictions for the test set (portion of data held out). If the model performs well, they will implement it as an algorithm that redirects users and pushes them in certain directions (e.g., how potential dates are ordered for viewing on a page) so users have a more productive experience. The approach offers no explanation or understanding for why dates are selected and successful; and the engineers typically land on odd behaviors and peculiar questions in surveys that most would not expect to be associated with such an outcome but which nonetheless do a good job at differentiating tastes.[12]

A third, and admittedly unrealistic albeit illustrative example might afford even greater clarity on these cultural differences—and possibly a means to seeing how a theoretical bridge or synthesis could be possible in the future. Let's take the example of a murder investigation. Social scientists arrive at the scene with theories of murder guiding what they see and then collect evidence to test each of these theories. Engineers approach the context differently. They see the scene as entailing tons upon tons of features or evidence. Using prior murder scenes and evidence from them, they train a model on "known" culprits, identifying those features most associated with murder. They then use those weighted features to predict the likely murderer. Motives, narratives, and explanations are secondary and ad-hoc.

Theoretical Embeddings

In sum, there are very distinct research frameworks and general perspectives at work when it comes to the analysis of big data. How does this play out in the trading zone focused on the same empirical phenomena? Ideally, we would expect big data to afford these fields a shared focus where both can learn from each other. Realistically, these perspectives entail distinct, potentially incommensurable gestalts for the study of social phenomena. This is not reason for hopelessness, however. Just as social psychology has repeatedly shown us that persons can live with great inconsistencies of world-view, so it may be possible that incommensurable views can be jointly adopted within a trading zone focused on the same topics. At worst, we may see multidisciplinary study, or independent efforts on same topic—where there is little collaboration and instead mutual observation (Fig. 1).

We believe the reality of current big data collaborations does entail trading and exchange of knowledge. We already see knowledge sharing in how social scientists and computer scientists use the same data and methods. Both seem to be adopting a focus on links and text, and both seem to be increasingly adopting a perspective that eschews methodological individualism in favor of relationalism or methodological transactionalism (Kirchner and Mohr 2010; McFarland et al. 2011). Where they differ is in how these data and methods are employed: either they are employed toward distinct ends or the scientists' respective expertise resides in different facets of conducting the same study. In many instances, their interdisciplinary collaborations reflect this via their divisions of labor or embeddings of perspectives (see Fig. 2).

In an interdisciplinary collaboration where there is a division of labor, the social scientist will likely write the theory and explanation section and the computer scientist will perform the analysis (see "interdisciplinary joining" in Fig. 2). Over time, the two will come to understand a degree of each other's perspective—but not as extensively as if they fully learned each other's perspective and adopted it as their own.

In many instances, multi-method approaches hide the fact that one viewpoint is favored over the other (layered approaches). In this sense, the interdisciplinary effort

[12] Here is an example of where the study of online dating sites by engineers reveal which odd questions differentiate tastes (http://blog.okcupid.com/index.php/the-best-questions-for-first-dates/).

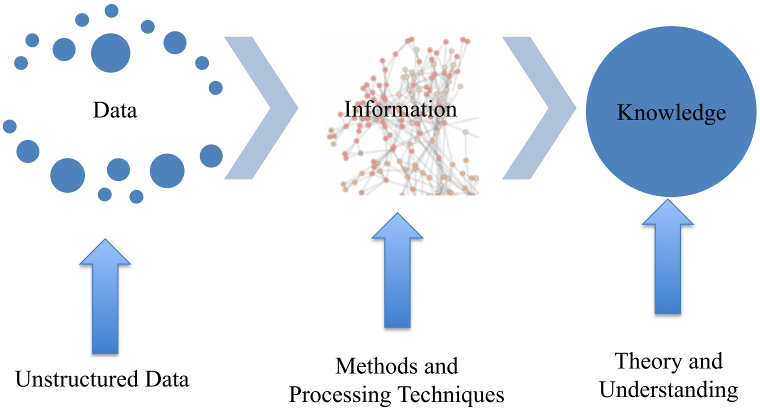
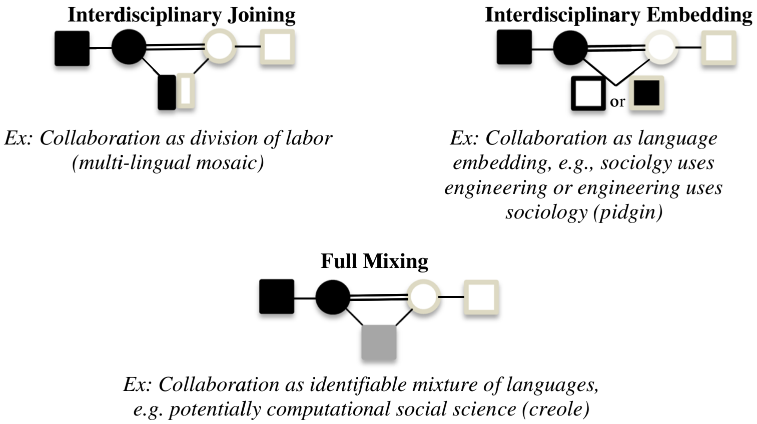


Fig. 1 Schematic of Data, Information, and Knowledge

will embed one perspective in service of another, more primary viewpoint (Diehl and McFarland 2010). This too is arguably already occurring as engineers seek to create solutions to problems and use theory as an ad-hoc explanation, or when social scientists seek to explain phenomena and use computational techniques of prediction to test their theories (see "interdisciplinary embedding" in Fig. 2).

A full hybridization of viewpoints may be difficult to accomplish for established scientists and engineers (see "full mixing" in Fig. 2), but new students and scholars will



Ex: Collaboration as division of labor (multi-lingual mosaic)

Ex: Collaboration as language embedding, e.g., sociolgy uses engineering or engineering uses sociology (pidgin)

Ex: Collaboration as identifiable mixture of languages, e.g. potentially computational social science (creole)

Key	***Symbol***	***Interpretation***
Product	■	Product or paper on subject matter of discipline A (area) with approach of discipline A (rim).
Person	○	Person from discipline B.
Relation (2-mode)	/	A person's connection to a product.
Relation (1-mode)	═	A person's connection to another person via a collaborative product.

Fig. 2 Increasing Mixtures of Collaboration

likely fuse these perspectives in novel ways: instead of speaking a pidgin form of social science infused with engineering terms and methods, they will speak a creole language that could form a new field and discipline.

Is There Hope for Synthesis?

Is there hope for a reciprocal form of exchange and theoretical synthesis across the fields of social science and engineering? In certain regards, our murder example offered a potential route, for it implied **forensic social science** as a potential hybrid ("full mixing") solution of these perspectives (Goldberg In Press). Forensic social science is an approach that merges applied and theory-driven perspectives. In a sense, deductive and inductive approaches are combined as mutually informing. This perspective is similar to Diane Vaughan's approach to qualitative sociological research (2014). According to Vaughan, one should not adopt a purely inductive approach akin to Glaser and Strauss (1967), but rather one should use theory to partly guide deductive explorations of the data while also using induction to discover which theories afford an explanation. In the case of a murder scene, the forensic scientist approaches the setting with an arsenal of tools and understandings of murder scenes. There, theory and evidence co-evolve quickly in a fast-paced iterative fashion.

There is a benefit of such an approach for big data. On one hand, there are an endless number of features one can construct from massively multivariate data. On the other hand, theories often act as blinders; they do not allow for serendipitous findings or unexpected discoveries outside the purview of their lens. When the theories define the data being collected—as is the case with traditional hypothesis testing social science—such exploration is by definition precluded. It makes sense to at least allow for some exploration of alternative theories and explanations. Here, one may find inductive approaches to reveal which sets of features matter, and from there derive which alternative explanations are most salient.

Take, for example, prior work by McFarland, Jurafsky, and Rawlings on speech features using WAV files and transcripts from dating encounters (2013). There, the authors could have generated a seemingly endless number of features about pitch, loudness, and rates of speech depending on what unit of time was considered (e.g., max, min, stdev, range, average, median, stdev of stdev) as well as how words were utilized (e.g., rates of any type of word, counts from sentiment dictionaries, discourse features). The authors could have run machine learning models that employed all of these features to predict an outcome like date selection. However, even with a huge dataset they would have run out of degrees of freedom, identified small cells, and bumped into collinearity issues. In fact, they pursued such an approach in other work (Ranganath et al. 2012); and while the authors unearth a number of specific associations, there is no larger narrative or understanding that helps explain what leads persons to select each other as dates. Rather than throwing a "kitchen sink" of features at the question, it was much more plausible to use theory as a sensitizing guide for how to explore this massively multivariate space. In this manner, one at least accomplishes the assessment of a particular theory as a null hypothesis.

Conversely, consider the matrix factorization algorithms developed by computer scientists to solve the Netflix Prize challenge introduced in 2006 (Koren et al. 2009). The one million dollar prize was eventually granted to a team of researchers who were able to improve the company's movie rating prediction error by ten percent. The algorithm, ultimately too complicated to be effectively implemented by Netflix as a means to improve its recommendation system, comprised a variety of adaptive movie-specific and user-specific effects that were calibrated over historical time and throughout a user's tenure on the website. The main collaborative filtering rationale behind the algorithm is to induce latent factors across the various movies as a function of their likelihood of being sampled and liked (or disliked) by the same people.

This prediction machinery would constitute the first step—rather than the outcome—of a forensic approach to social scientific exploration. Which of the components in this algorithm explain cultural taste, and to what cognitive and social processes do they correspond? Do they relate to the contents of the movies being evaluated; to the identities of the people who consume them; or perhaps to the manner by which people influence one another through their consumption choices? To become a theory building apparatus, not merely a prediction black box, forensic social science tools need to identify patterns in the data and then trace them back to meaningful analytical constructs. The promise of forensic social science is particularly exciting with respect to unstructured textual and audio-visual data, which have so far eluded social science. The age of big data might therefore also represent a paradigmatic shift from structural positivism to scientific constructivism that methodically explores and theorizes about the processes by which meaning (as manifest in text and image) is collectively negotiated through interpersonal interaction.

In sum, the data and methods adopted in big data analysis afford us some route to a trading zone across atheoretical, applied, and inductive procedures on the one hand, and theory-driven, explanation focused, hypothesis-testing deductive procedures on the other. Forensic social science weaves these approaches together—neither offering a purely inductive nor purely deductive perspective. Perhaps the influx of computation and big data will press social science in this direction, bringing back in a level of practical application and speeding up cycles of hypothesis testing and exploration.

Conclusion: The Future for Sociology and big Data

The emergence of big data is a watershed moment for the social sciences akin to that seen in the prior century's statistical and survey turn. We are witnessing a potential shift from methodological individualism and affinitive theories (e.g. rational choice theory) to methodological transactionalism and an openness to a wider range of social theories concerning interaction, attention focus, and meaning usage. Big data represent the emergence of a new class of data gleaned from digital records from a dizzying array of social phenomena. The data are not only big but rich (dynamic and massively multivariate), and they often concern the form and content of communications (links and texts). Old analytic techniques are often inapplicable. As a result, there is demand for new methods that reduce and simplify the dimensionality of data, identify novel patterns and relations (**computational ethnography**, **computational linguistics**, **network science**), predict outcomes (**machine learning**), and implement in situ **experiments** that reveal how we can alter action in desired directions. These new approaches constitute a shift away from standard OLS techniques and lab experiments on psychology 101 students (Anand 2010). The newly available data and methods have also spawned a revisiting of old social science questions (now potentially answerable at the societal level) and enabled the creation of entirely new directions of inquiry (thanks largely to new technologies that both mediate and record a striking array of social experiences).

All of these transformations—a shift in data, methods, and inquiry—are arising because company engineers, as well as recruited computer scientists and social scientists, are all focusing on digital social information collected on large swaths of the human populace. However, the shared focus and concern is at odds with the fact that there are strong differences in research frameworks, and especially those between engineers/industry and social scientists. The nature of these differences helps explain why an easy, egalitarian trading zone may not arise across these fields focusing on big social data. It may also explain why, at least in the short term, we may be more likely to witness engineering colonize sociology and the social science than vice versa.

We are currently experiencing a fevered pitch of technological innovation, where engineering solutions to practical problems carry greater value, both financial and symbolic, than scientific explanation and understanding. This is particularly so for a variety of structural reasons. Engineering questions and solutions are far more applied to companies' concerns for profit. Employment of social scientists may hinge on their ability to adopt a computer science approach and utilize social science merely as an afterthought to help color what was found. Simply put, theory and explanation are not as valuable as discovering what works. Careful sampling techniques and the hypothesis testing confirmatory science that generations of social scientists have been trained to conduct will not keep pace with engineers' brute force application of machine learning to predict a variety of behavioral and consumer outcomes using big data. As more and more of our gadgets are black boxes, we operate with greater faith in what they do without understanding why they work the way they do (Latour 1988). Furthermore, granting agencies afford far more funding for engineering problems and student training than they do for the social sciences. The far greater resource investment and demand for engineering applications with relation to social media and the mining of digital information will likely mean social science is less represented in these problems and less able to define the dialogue surrounding them. For these reasons, we expect theory to decline as it is rendered a more secondary role.

That said, at some point both the reservoirs of online information and the demand for this information will be filled, and then there will be a heightened concern with improving the quality of information and how we make sense of it (e.g., how do we translate information into knowledge?). In this phase of the engineering age, we may see an effort to take stock, to understand and explain why things happen the way they do, and to define and identify what it means for a task or procedure to perform better or worse. There, social science and its theories will come into demand; and in spite of all this, social science will always remain essential to efforts at preserving the public good. In addition, there will grow a need for synthesis of information and narratives that enable us to understand the multitude of findings we have made—tasks for which sociological theory is distinctly suited.

In spite of these structural conditions, we believe the middle ground of a **forensic social science** approach may be most amenable and successful in the long run. We believe this not only because it seems defensible in an age of pluralistic social theory—but also in a pragmatic age where critical comparisons of approaches and efforts at robust explanation will afford the longest lasting contributions to knowledge (Stokes 1997).

References

Abbott, A. (1988). Transcending general linear reality. *Sociological Theory, 6*(2), 169–86.

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River: Prentice Hall.

Alpaydin, E. (2004). *Introduction to machine learning*. Cambridge: MIT Press.

Anand, G. (2010). A weird way of thinking has prevailed worldwide. *New York Times* (August 25, 2010).

Anderson, M. J. (1988). *The American census: a social history*. New York: Yale University Press.

Anderson, A., McFarland, D. A., & Jurafsky, D. (2012). Towards a computational history of the ACL: 1980–2008. *Association of Computational Linguistics, Workshop* (ACL Workshop 2012).

Backstrom, L., Kleinberg, J., Lee, L., & Danescu-Niculescu-Mizil, C. (2013). Characterizing and curating conversation threads: expansion, focus, volume, re-entry. *Proceedings of WSDM*, 2013.

Bail, C. A. (2014). The cultural environment: measuring culture with big data. *Theory and Society, 43*, 465–482.

Barabasi, A. (2003). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. New York: Plume.

Bender-deMoll, S., & McFarland, D. A. (2006). The art and science of dynamic network visualization. *Journal of Social Structure, 7*(2).

Berger, P., & Luckmann, T. (1966). *The social construction of reality: a treatise in the sociology of knowledge*. New York: Anchor.

Bishop, C. (2007). *Pattern recognition and machine learning (information science and statistics)*. Cambridge: Springer.

Blei, D. (2012). Probabilistic topic models. *Review article, Communication of the ACM, 55*(4), 77–84.

Bohn, A., Buchta, C., Hornik, K., & Mair, P. (2014). Making friends and communicating on facebook: implications for the access to social capital. *Social Networks, 37*, 29–41.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*, 892–95.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*, 662–79.

Brandes, U., Robins, G., McCranie, A., & Wasserman, S. (2013). What is network science? *Network Science, 1*, 1–15.

Brown, J. S., & Duguid, P. (2002). *The social life of information*. Harvard Business Review Press

Bruch, E. E., & Mare, R. D. (2012). Methodological issues in the analysis of residential preferences and residential mobility. *Sociological Methodology, 42*, 103–54.

Camic, C., & Xie, Y. (1994). The statistical turn in American social science: Columbia University, 1890 to 1915. *American Sociological Review, 59*(5), 773–805.

Castells, M., Fernández-Ardèvol, M., Qiu, J. L., & Sey, A. (2007). *Mobile communication and society: a global perspective*. Cambridge: MIT Press.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–97.

Coleman, J. S. (1986). Social theory, social research, and a theory of action. *American Journal of Sociology, 91*(6), 1309–35.

Coleman, J. S. (1994a). *Foundations of social theory*. Cambridge: Belknap Press.

Coleman, J. S. (1994b). A vision for sociology. *Society, 30*, 29–34.

Collins, H., Evans, R., & Gorman, M. (2007). Trading zones and interactional expertise. *Studies in History and Philosophy of Science, 38*(4), 657–66.

Converse, J. M. (1987). *Survey research in the United States: roots and emergence 1890–1960*. Berkeley: University of California Press.

Cukier, K., & Mayer-Schoenberge, V. (2013). The rise of big data: how it's changing the way we think about the world. *Foreign Affairs,* 28–41.

Diehl, D., & McFarland, D. A. (2010). Towards a historical sociology of situations. *American Journal of Sociology, 115*(6), 1713–52.

Dodds, P. S., Muhamad, R., & Watts, D. (2003). An experimental study of search in global social networks. *Science, 301*(5634), 827–9.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: reasoning about a highly connected world*. Cambridge: Cambridge University Press.

Einav, L., Levin, J., Popov, I., & Sundaresan, N. (2014). Growth, adoption and use of mobile e-commerce. *American Economic Review: Papers and Proceedings, 104*(5), 489–94.

Fleck, L. (1979). *Genesis and development of a scientific fact*. Chicago: University of Chicago Press.

Galison, P. (1997). *Image and logic: a material culture of microphysics*. Chicago: University of Chicago Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. Chicago: Aldine Pub. Co.

Goldberg, A. (in press). In defense of forensic social science. *Big Data & Society.*

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures. *Science, 333*(6051), 1878–81.

Golder, S. A., & Macy, M. W. (2014). Digital footprints: opportunities and challenges for online social research. *Annual Review of Sociology, 40*, 129–52.

González-Bailón, S., Borge-Holthoeter, J., Rivero, A., & Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific Reports, 1*, 197.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38*, 16–27.

Grimmer, J., Westwood, S. J., & Messing, S. (2014). *The impression of influence: legislator communication, representation, and democratic accountability.* Princeton: Princeton University Press.

Hacking, I. (2006). *The emergence of probability.* Cambridge: Cambridge University Press.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science, 332*(6025), 60–5.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. New York: Prentice Hall.

Kagan, J. (2009). *The three cultures: natural sciences, social sciences, and the humanities in the 21st century.* New York: Cambridge University Press.

Kirchner, C., & Mohr, J. W. (2010). "Meanings and relations: an introduction to the study of language, discourse, and networks.". *Poetics, 38*(6), 555–66.

Kohavi, R., & Longbotham, R. (2007). Online experiments: lessons learned. *Computer, 40*(9), 103–5.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30–37.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.

Latour, B. (1988). *Science in action*. Cambirdge: Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts* (2nd ed.). Princeton: Princeton University Press.

Laumann, E. O., Marsden, P., & Prensky, D. (1983). "The boundary specification problem in network analysis.". In R. S. Burt & M. J. Minor (Eds.), *Applied network analysis: A methodological introduction*. London: Sage Publications.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational social science. *Science, 323*(5915), 721–3.

Leskovec, J., & Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. *International World Wide Web Conference (WWW).*

Leskovec, J., Lang, K., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*. New York: ACM.

Levine, D. N. (1995). *Visions of the sociological tradition*. Chicago: University of Chicago Press.

Lewis, K. (2015). Studying online behavior: comment on Anderson et al. 2014. *Sociological Science, 2*, 20–31.

Lewis, K. (in press). Three fallacies of digital footprints. *Big Data & Society.*

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: a new social network dataset using facebook.com. *Social Networks, 30*(4), 330–42.

Lohr, S. (2012). "The Age of Big Data." *New York Times* (February 11, 2012)

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *IJCAI* (International Joint Conferences on Artificial Intelligence).

McFarland, D.A. and H.R. McFarland. (in press). Big data and the danger of being precisely inaccurate. *Big Data & Society.*

McFarland, D. A., Diehl, D., & Rawlings, C. (2011). "Methodological transactionalism and the sociology of education.". In H. Maureen (Ed.), *Chapter 5 in Frontiers in sociology of education* (pp. 87–109). New York: Springer.

McFarland, D. A., Manning, C. D., Ramage, D., Chuang, J., Heer, J., & Jurafsky, D. (2013a). Differentiating language usage through topic models. *Poetics, 41*(6), 607–25.

McFarland, D. A., Jurafsky, D., & Rawlings, C. (2013b). Making the connection: social bonding in courtship situations. *American Journal of Sociology, 118*(6), 1596–1649.

Menand, L. (2010). *The marketplace of ideas: issues of our time*. New York: W.W. Norton & Company.

National Research Council. (2014). *Convergence: Facilitating transdisciplinary integration of life sciences, physical sciences, engineering and beyond*. National Research Council.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98*, 404–409.

Newman, M. E. J. (2009). *Networks: an introduction*. Oxford: Oxford University Press.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Pentland, A. (2014). *Social physics: How good ideas spread–the lessons from a new science*. New York: Penguin Press.

Platt, J. (1996). *A history of sociological research methods in America, 1920–1960*. Cambridge: Cambridge University Press.

Porter, T. M. (1995). *Trust in numbers*. Princeton: Princeton University Press.

Porter, T. M., & Ross, D. (Eds.). (2003). *The modern social sciences*. New York: Cambridge University Press.

Ranganath, R., Jurafsky, D., & McFarland, D. A. (2012). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language, 27*(1), 89–115.

Rogers, E. M. (1987). Progress, problems and prospects for network research: investigating relationships in the age of electronic communication technologies. *Social Networks, 9*, 285–310.

Rosenfeld, M. J., & Thomas, R. J. (2012). Searching for a mate: the rise of the internet as a social intermediary. *American Sociological Review, 77*(4), 523–47.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*, 854–6.

Shi, X., Leskovec, J., & McFarland, D. A. (2010). Citing for high impact. *Joint Conference on Digital Libraries, (JCDL 2010*).

Shwed, U., & Bearman, P. S. (2010). The temporal structure of scientific consensus formation. *American Sociological Review, 75*(6), 817–40.

Smith, A., & Duggan, M. (2013). *Online dating & relationships*. Washington: Pew Research Center.

Snow, C. P. (2001). *The two cultures*. London: Cambridge University Press.
1959.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: cognitive consequences of having information at our fingertips. *Science, 333*, 776–8.

Stokes, D. E. (1997). *Pasteur's quadrant: basic science and technological innovation*. Washington: Brookings Institution Press.

Stouffer, S. A. (1949). *In The American Soldier, 4 vols Studies in social psychology during World War II.*. Princeton, NJ: Princeton University Press.

Szell, M., & Thurner, S. (2010). Measuring social dynamics in a massive multiplayer online game. *Social Networks, 32*, 313–29.

Talley, E., Newman, D., Herr, B., II, Wallach, H., Burns, G., Leenders, M., & McCallum, A. (2011). A database of national institutes of health (NIH) research using machine learned categories and graphically clustered grant awards. *Nature Methods, 8*, 443–4.

Vaisey, S. (2009). Motivation and justification: a dual-process model of culture in action. *American Journal of Sociology, 114*, 1675–1715.

Vaughan, D. (2014). Analogy, cases, and comparative social organization. In R. Swedberg (Ed.), *Theorizing in social science: the context of discovery* (pp. 61–84). Stanford: Stanford University Press.

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in social network data: a re-classification. *Social Networks, 34*(4), 396–409.

Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge: Cambridge University Press.