Big Data & Inductive Theory Development:  
Towards Computational Grounded Theory? 

Completed Research Paper

Nicholas Berente  
University of Georgia  
berente@uga.edu

Stefan Seidel  
University of Liechtenstein  
stefan.seidel@uni.li

Abstract

It has been argued that the unprecedented availability of trace data may revolutionize the social sciences. Still, methodological knowledge is scarce as to how this abundance of data can be used to develop novel and important theory. In this essay, we inquire into how the lessons learned from Grounded Theory Method (GTM) can be used to build theory from big data. To do so, we review GTM in light of three key concepts in social analysis: the continuum of induction, the continuum of generalization, and the level of lexicon and theory. Using Habermas's concept of rational reconstruction we articulate a broader "grounded paradigm" that emphasizes the notion of emergence and provides a pragmatic epistemological foundation for different types of grounded analysis. On this basis, we propose a model that describes the process of theorizing from big data.

Keywords

Grounded theory, big data, theory development, computational grounded theory.

Introduction

“When we wake up in the morning, we check our e-mail, make a quick phone call, walk outside (our movements captured by a high definition video camera), get on the bus (swiping our RFID mass transit cards) or drive (using a transponder to zip through the tolls). We arrive at the airport, making sure to purchase a sandwich with a credit card before boarding the plane, and check our BlackBerries shortly before takeoff. Or we visit the doctor or the car mechanic, generating digital records of what our medical or automotive problems are. We post blog entries confiding to the world our thoughts and feelings, or maintain personal social network profiles revealing our friendships and our tastes. Each of these transactions leaves digital breadcrumbs which, when pulled together, offer increasingly comprehensive pictures of both individuals and groups, with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago" (Lazer et al., 2009, p. 721).

As this quotation illustrates, “big data,” comprised of abundant and ever-increasing trace data, offers boundless opportunities for a computational social science (Lazer et al., 2009). It is precisely this unprecedented level of trace data that Latour (2010) indicates might revolutionize the social sciences. Through direct computational attention to trace data, researchers can potentially skip some pre-established latent constructs (what Latour refers to as “imaginary”—see Latour 2005) and generate richer and more accurate understandings of social life. According to Latour, researchers now have the opportunity to directly access large amounts of empirical phenomena and develop insights closer to the source. Big data requires novel visualizations and pattern identification (Latour, 2010; Lazer et al., 2009), but there is a dearth of guidance for how, broadly, to attend to such a wealth of data. Big data offers an opportunity to generate novel theory, but there are limited methodological options for researchers who wish to do so. To unleash the power of trace data, scientists might benefit from methodological guidance from an approach that focuses on inductive generation of novel theory. Grounded Theory Method (GTM) may have something to offer computational social science.

GTM is intended to enable novel theorizing from large amounts of data (Legewie and Schervier-Legewie, 2004). Although GTM’s adherents typically work with qualitative data, GTM was originally formulated to
Research Methods

accommodate both quantitative and qualitative data (Glaser, 2008; Glaser and Strauss, 1967). However, the method is extremely labor-intensive and researchers cannot be expected to pour over gigabytes of trace data using coding strategies commonly associated with qualitative data—it would simply be too manually intensive. Similarly, big data typically does not come in readily comparable numerical form, which would lend itself to the sort of ordering that was recommended for quantitative data (Glaser, 2008; see also Glaser and Strauss, 1967). The lessons of GTM thus do not necessarily transfer directly to the computational analysis of big data. However, perhaps there is some value in using GTM in a more fundamental sense—since this is the only widespread existing method for generating theory from large sets of unstructured data.

In this essay, we inquire into how the lessons from GTM apply to computational analysis of big data in an effort to generate theory. To do so we first review GTM and then unpack some of the theoretical underpinnings of the method by drawing upon three key concepts in social analysis:

1. **The Continuum of Induction**: The central tension of GTM involves the role of existing theory in induction—can researchers really derive theory from a “blank slate” approach to empirical data? Or are their interpretations sensitized by existing theory? This tension is evident in the very foundations of social theorizing. For example, in Merton’s (1957) critique of empirically-grounded theorizing, he distinguishes between two forms of “post factum” analysis: one that uses existing theory to explain empirical contexts, and one that uses empirical contexts to derive new theory.

2. **The Continuum of Generalization**: In further conceptualizing the concept of induction, we distinguish between substantive and formal theory (Glaser and Strauss, 1967; Urquhart, Lehmann, and Myers, 2010). Substantive theorizing involves highly contextualized and specific accounts of phenomena, whereas formal theory is intended to apply more broadly across contexts.

3. **Level of Theory and Lexicon**: To support theorizing in GTM, theoretical codes are used to generate mid-level concepts and categories from empirical data (Glaser, 1978). Concepts are developed by researchers whose interpretations are sensitized by prior experience and familiarity with existing theory. This approach is consistent with Habermas’s “rational reconstruction” method for social scientific analysis, and his identification of the importance of lexicons as critical enablers of scientific knowledge in a community (Habermas 1984; 2003).

In this paper, we briefly review GTM in light of these three key concepts. We conclude by articulating a broader “grounded paradigm” in information systems research that emphasizes the notion of emergence—emergence of (1) study design, (2) sense-making artifacts (i.e., categories), and (3) theory. This view of the grounded theory paradigm is not intended to supplant GTM for qualitative analysis. Rather, it is intended to leverage the insights associated with GTM more widely, and to provide a pragmatic epistemological basis all sorts of grounded analysis. This might inform GTM research, but is intended to make the key ideas portable for opportunities such as computational analysis of big data.

**Overview: GTM & Computational Analysis of Big Data**

In the middle part of the 20th century, Barney Glaser and Anselm Strauss sought to bring together the rigorous empirical approach of the Columbia school of sociology with the creative, but often less rigorous, approach of the Chicago school (Glaser and Strauss, 1967). This was motivated largely by the desire to inspire new and innovative efforts at theorizing rather than application of existing theories (what Strauss referred to as “working” these existing theories, see Legewie and Schervier-Legewie, 2004) in a way that was legitimate and acceptable in light of the prevailing, positivistic social science of the time. Thus the original goal of the grounded theory effort might be characterized as twofold: (1) to encourage novel theorizing; and (2) to develop an empirically-driven methodology to enable this novel theorizing.

The result was GTM, which has been one of the strongest catalysts for widespread acceptance of qualitative research across a variety of social science disciplines (Bryant and Charmaz, 2007). Grounded theory seeks to develop theoretical concepts and relationships while being informed by intense analysis of empirical data (Glaser and Strauss, 1967; Strauss and Corbin, 1990). Since their seminal work on the

---

1 Of course, in the literature GTM has almost exclusively been used with qualitative data.
discovery of new theory through this rigorous, empirically-grounded methodology, information systems (IS) researchers have adopted the perspective with varying degrees of faithfulness to the method (Matavire and Brown, 2011; Urquhart et al., 2010). Sometimes GTM is adopted quite rigorously, and other times it is invoked as a sort of catch-all method for qualitative or interpretive research that does not necessarily attend to the fundamental tenets of the method (Urquhart and Fernández, 2013). This has led to recent calls to take the method more seriously in the IS field (Urquhart and Fernández, 2013; Urquhart et al., 2010) as well as in other fields (e.g., Suddaby, 2006).

In recent years, this attention to GTM as a rigorous and increasingly well-understood method has been incorporated into the review process of the IS field, and there is an established discourse around reflective and appropriate application of GTM (e.g., Matavire and Brown, 2011; Urquhart et al., 2010). These recent trends indicate a maturation of the methodology in the IS field. With this maturation, however, it is important to avoid the dogmatism that sometimes accompanies the establishment of discourses within the IS field. Dogmatism may close down the discourse precisely when it should be evolving and staying current (Ciborra, 1998). In primarily emphasizing the second goal of the grounded theory movement—the establishment of a valid methodology—we must be careful not to undermine the first goal—novel forms of theorizing. A move toward stringent application of GTM as a method, however positive for the discourse concerning that particular approach, may in some ways undermine the spirit of creativity that led to the method in the first place.

In this paper, we look to both preserve the maturation of GTM as a method, but also to carve out a space for alternative empirically-grounded inductive approaches that do not necessarily follow the prescriptions of GTM, nor the traditionally qualitative approach usually associated with GTM. Particularly in the context of the “big data” revolution and computational social science, the unprecedented access to different forms of data can drive alternative inductive approaches. Latour (2010), for example, argued that the current explosion of digital trace and other computational data offers an unprecedented opportunity to explore empirical phenomena without the baggage of traditional qualitative and quantitative approaches—they offer the potential for strong, empirically-grounded inductive theorizing, yet cannot really be incorporated into the solidifying discourse around GTM.

**GTM in a Nutshell**

Over the years, GTM has evolved into a contested “family” of methodologies, rather than one, very specific method (Bryant and Charmaz, 2007). This family of methods is replete with variants and rich in reflective discourse. There are disagreements on coding procedures (e.g., Kelle, 2007a), the role of existing research (e.g., Jones and Noble, 2007), epistemological foundations (e.g., Charmaz, 2000), and a host of other divisions. At a very broad level, one can distinguish between ‘classical’ grounded theory that is much in line with the original version of GTM as proposed by Glaser and Strauss (Glaser and Strauss, 1967), and ‘evolved’ grounded theory as, for instance, proposed by Strauss (1987), Strauss and Corbin (1990, 1998), or Charmaz (2000; 2006) in the form of ‘constructivist’ grounded theory. From a unifying perspective, however, the method can be thought to involve a number of key elements. Strauss alludes to these key elements when he describes the motivation behind writing the book (Glaser and Strauss, 1967) and founding the method:

“\textit{The important thing about that book, which is still not understood by most people, is that... it really had three different purposes. One was, we were trying to legitimate qualitative research in a period when it wasn’t yet legitimated... The second reason we did it is that we wanted to attack people like Blau, Parsons, and Merton, because their theories were being taken over by students and younger sociologists... they weren’t challenging the theories, they were just working them... they were dotting i’s and crossing t’s of these theories. So we attacked so-called “received” theories that were defective.... The third reason we wrote the book is that we wanted to put forward the idea of doing theory that was grounded.}” (Strauss 1994 in Legewie and Schervier-Legewie, 2004).

Thus, GTM’s current manifestations flow from these original purposes. First, GTM involves studying qualitative data (Strauss and Corbin, 1998; Urquhart, 2013)—in fact, some have argued that GTM was one of the key methodologies for establishing the legitimacy of qualitative research in a number of fields (Bryant and Charmaz, 2007). Second, GTM is primarily concerned with the generation of theory—preferably novel and interesting theory—rather than just “working” pre-existing views (compare also
Third, this theory generation is a result of intense empirical analysis. GTM is “grounded” in data—typically lots of data. Beyond these three original motivations, there are a number of practices that are often explicitly associated with GTM, including theoretical sampling (Glaser and Strauss, 1967; Hood, 2007; Morse, 2007), memo writing (Lempert, 2007; Strauss and Corbin, 1998), and the development and use of emergent concepts (Glaser, 1978; Kelle, 2007a; Strauss and Corbin, 1998). However, the three key areas—qualitative research, new theory development, and empirically grounded induction—are each always present in GTM efforts. This is interesting, particularly since each of these fundamental tenets of GTM have been deemed problematic. Next, we briefly describe challenges to each of these areas: (1) although it is typically a qualitative method, GTM can also accommodate quantitative analysis; (2) although it is aimed at new theory, GTM necessarily allows for existing theory; and (3) although GTM involves a rigorous procedure for induction, so do a variety of other methods:

- **Qualitative vs. Quantitative:** Although GTM is credited with helping legitimize qualitative research, its founders never intended for this process to be exclusively qualitative. In their pioneering text, Glaser & Strauss (1967) spell out a procedure for theoretically ordering quantitative data in “elaboration tables” to enable the generation of new theory. Although qualitative data was most definitely the focus, Glaser & Strauss (1967) went through great lengths to describe rigorous alternatives to the “verification” oriented research that is rooted in quantitative data. They suggested that rich, quantitative data can also be explored inductively to identify patterns and generate (rather than test) hypotheses. Glaser continues to promote this same methodology as a way to mine the realms of unused and unacceptable survey data that have been generated over the years for further insight (Glaser, 2008), and others have begun defining different roles for quantitative analysis in the pursuit of grounded theory (Urquhart, 2013).

- **Existing Theory:** Glaser and Strauss originally thought to develop theories without respect to existing theories in order to generate new insight. This view, however, has been criticized because it is both unrealistic (people always have pre-existing theories in their minds) and leading to trivial or non-original results (if analyzing phenomena without existing literature, researchers may reconstruct existing explanations) (Urquhart and Fernández, 2013). Similarly, Kelle (2007a, 2007b) has argued that the understanding of GTM as a purely inductive method does not hold by either the Straussian or Glaserian schools of thought. Reichertz (2007) observes that since observation is always conditioned by implied theory, GTM is more abductive in the sense of Peirce (1992). Proponents of constructivist grounded theory have argued that grounded theories are constructed by researchers based on their interactions with the field (Bryant, 2002; Charmaz, 2000; Charmaz, 2006), much in line with the pragmatist origins of the method (Strübing, 2007).

- **Empirically-grounded Induction:** In a (relatively) recent article, Jane Hood (2007) distinguishes between GTM and what she describes as the “Generalized Inductive Qualitative Model”—which is shorthand for other forms of case-based qualitative research with inductive goals. According to Hood (2007), it is not the qualitative aspect, nor the theory generation, nor the inductive empirical basis that mark grounded theory. Instead, it is the emergent nature of the theory hand-in-hand with research as it is being conducted. Sampling should progress based on findings as data are continuously compared (constant comparison) in a theoretically-driven way to enhance more insight (theoretical sampling) and should only end when no new findings are emerging (saturation)\(^2\). In this respect, the concept of theoretical sensitivity (Glaser, 1978) has been discussed intensively, and Glaser (1978; 2005) has introduced a total of 41 coding families that are intended to help the researcher identify relationships in the data, thereby sensitizing her to allow for emergence while avoiding to use preconceived ideas and conceptualizations. Hood (2007) argues that it is the emergence, not solely the induction or qualitative methods that mark GTM. Grounded research grows organically from within, not through a pre-structured plan. If a researcher goes to her data with a purposeful sample based on pre-existing concepts and categories, then the design of the research does not emerge, but is pre-specified. In GTM research, the “design, like the concepts, must be allowed to emerge during the research process” (Strauss and Corbin, 1998, p. 33).

\(^2\) Glaser similarly identified theoretical coding, theoretical sampling, and constant comparison as key elements for research to be called grounded theory (Legewie & Schervier-Legewie 2004).
On the basis of this brief analysis, we find that it is neither qualitative data, nor completely novel theory, nor the idea of induction that mark grounded theory. The key element appears to be the idea of emergence. The design, the concepts, the sample, and the eventual theory, are all expected to emerge through the creative, adaptive, yet rigorous processes of the researchers with the data and the sample.

**Abstracting GTM**

For the purpose of guiding computational analysis of big data, next we draw on three areas that are fundamental to GTM and relate the method to common streams of social science: (1) the role of existing theory in generating new theory (induction); (2) the level of abstraction of theory (generalization); and (3) the importance of theoretical codes as concepts to theory generation (rational reconstruction / lexicons).

**The Continuum of Induction: Analysis to Explain & Analysis to Derive**

It is a bit ironic that Glaser was a student of Robert Merton, one of the staunchest and most high profile advocates for positivistic, functionalist social science. According to Merton (1957), findings from the interpretation of existing data have the illusory advantage of plausibility, but offer very little of value in terms of “compelling evidence value” (p. 93). Essentially, Merton argued that while theory used to explain existing data may fit that data, there is little reason (in his mind) to accept one theory instead of another that might also conceivably fit that data. There is theoretically any number of possible explanations for a phenomenon, and interpretive conclusions cannot make claim to providing the “right” lens for the data. Of course, a veritable army of interpretive researchers have debunked this view over the years—from a variety of pragmatist and interactionist perspectives (e.g., Strauss) or phenomenological and ethnomethodological perspectives (e.g., Guba and Lincoln, 1994; Van Maanen, 1979).

However, Merton did make an interesting distinction in the relationship of theory to interpreted data. He describes the difference between using theory to explain vs. deriving theory from the data. When deriving theory, the researcher proceeds from observations via empirical generalizations to theory and thus reasons inductively (Handfield and Melnyk, 1998). That is, concepts and relationships are derived from the data. In explaining data with theory, on the other hand, consequences are deduced from a hypothesis, and then compared with empirical observations. That is, existing concepts and relationships are used to make sense of data. These two modes of reasoning are captured by the continuum of induction (Figure 1).

![Figure 1. The Continuum of Induction](image)

In practice, the distinction between “derive” and “explain” is blurry, and both types of reasoning are typically applied in conjunction (Handfield and Melnyk, 1998). This is also the case in grounded theory studies, where the researcher starts with initial slices of data, but then compares emergent concepts with incoming data (Glaser and Strauss, 1967; Strauss and Corbin, 1998).

**The Continuum of Generalization: Substantive & Formal Theorizing**

Glaser & Strauss (1967) distinguished between substantive and formal theorizing. Substantive theorizing involves highly contextual relationships between variables that are situated in a particular context. Formal theorizing generates broader, more general theory. The difference is thus in the level of abstraction (Kearney 2007). At this, one can proceed from substantive to (more) formal theory, as indicated in Glaser (1967) and elaborated on in Urquhart et al. (2010). Urquhart et al. (2010) provide a nuanced of levels of generalization in theorizing by distinguishing between bounded, substantive, and formal concepts.
Bounded concepts are intensely contextual, substantive concepts address a “middle range” between the contextual and the general, and formal concepts are intended for broad generalizations (see Figure 2).

![Figure 2. The Continuum of Generalization](image)

While the two dimensions have been discussed in seminal works on grounded theory method, they are not only relevant to grounded theory studies, but indeed to any study that focuses on emergence. In the next section, we will use Habermas’s rational reconstruction to explain how theories in a grounded paradigm are developed using theoretical coding—which requires the elicitation of concepts from empirical data.

**Rational Reconstruction, the Level of Theory, and the Level of Lexicon**

Habermas (1984; 2003) conceived of social science in terms of a “rational reconstruction” of empirical phenomena within a particular community of researchers. In characterizing social science in such a way, Habermas looked to cut a line between the objectivist paradigm of scientific method and the subjectivist paradigm associated with interpretive positions (Pedersen 2008). The rational reconstructive perspective enabled Habermas to approach science with an appreciation for rigor, evidence, and the methodological advantages of the scientific method, without the naïve positivism that often accompanies it. Similarly, this view allowed him to appreciate individual interpretations without falling into the relativism of “anything goes.”

Rational reconstruction is essentially an epistemological position. Through a rationally reconstructive view, communities of social scientists who share a “lifeworld” necessarily also share some level of an intersubjective understanding of the world that they study (Habermas 1984). Fundamental to this intersubjective understanding is a lexicon of shared vocabulary involving key concepts that reflect the assumptions, history, and institutional context of the community.

The lexicon is the scaffolding upon which the scientific community is constructed. To Habermas, any knowledge is always with respect to a community, and made sense of (i.e., reconstructed) with respect to the lexicon of that community. This lexicon involves “pre-theoretic” elements that enable its members to communicate with each other and compare findings—generalizations are always made with respect to a particular lexicon (Pedersen 2008).

The lexicon reflects the theoretical lens through which scientists make sense of empirical data. For example, the languages of structuration theory (agency, structure, modes of legitimation, domination, signification); actor-network theory (translation, material agency, enrollment, intermediaries, mediators, etc.); social network analysis (nodes, links, centrality, distance, homophily, etc.) are all lexicons. Specific theoretical positions are inherent in the particular terminologies in their contexts of use. Any computational analysis of big data requires a theoretical lexicon. “Map reduce” (a process by which
computational researchers might define a structure around unstructured data) essentially requires a theoretical lexicon—whether explicit or implicit. Applying theoretical concepts and relationships provided by a lexicon thus allows the analyst to move beyond mere description.

Habermas’s concept of rational reconstruction can be used to describe how social scientists draw upon lexicons to analyze empirical phenomena in relation to the above two continua of induction and generalization, and this leads to four major modes of grounded inquiry. From a broad perspective, in deriving theory, lexicons are generated or extended in either a substantive or formal setting, based on the level of generalization. The result is a new or adapted lexicon. Using a lexicon to explain phenomena either substantially (contextually) or more generally may result in new knowledge for that community, but does so strictly within the bounds of that existing lexical framework. A framework of these four modes of the relationship between theoretical lexica and induction is summarized in Table 1. Next, we provide examples of each mode, using structuration theory and social network analysis as examples of lexicons commonly used in information systems.

| Table 1. Continua of Induction, Generalization, and the Level of Lexicon |
|-----------------------------|-----------------------------|
| **Derive**                  | **Explain**                 |
| **Formal Theory**           |                             |
| **Generate Lexicon:**        | **Use Lexicon:**            |
| Develop or modify           | Broadly apply general theories; use lexicon to explain. |
| “general theories”          | Example: Apply structuration theory to IS field (Jones and Karsten, 2008) |
| Example: Extend Social Network Analysis (SNA) to include multidimensional networks (Contractor, Monge, and Leonardi, 2011) | |
| **Substantive Theory**      |                             |
| **Generate Lexicon:**        | **Use Lexicon:**            |
| Develop local or mid-range theories | Extend general theory in domain |
| Example: Extend structuration theory to the area of IT (Orlikowski, 1992) | Example: Apply SNA to open source software (Grewal, Lilien, and Mallapragada, 2006) |

**Generate Lexicon for Formal Area of Inquiry:** In deriving formal theory based on empirical data, the lexicon must often be adapted to accommodate new modes of inquiry or forms of data. Thus, while generating new theory, the researchers would need to extend the original lexicon with new concepts. This may involve drawing on another lexicon. For example, Contractor et al. (2011) extended SNA using the sociomaterial lens, thus resulting in a new lexicon to explore “multidimensional networks.”

**Generate Lexicon for Substantive Area of Inquiry:** In this case, an adaptation of an existing lexicon occurs in two ways. First, it is applied to a context. Second, the lexicon is extended in a non-trivial way. In deriving substantive theory through a primarily inductive process, one generates a lexicon that serves as the scaffolding that allows analyzing a substantive area of inquiry. An example is Orlikowski’s (1992) seminal extension of structuration theory to accommodate IT artifacts in organizational contexts.

**Use Lexicon in Formal Area of Inquiry:** In using a formal theory in a broad domain, a general lexicon is applied very generally. This does not extend the lexicon, but applies it and makes general claims. An example in information systems research is Jones & Karsten (2008), who reported on a review of structuration theory in the information systems literature. The general lexicon provided by structuration theory (i.e., concepts and relationships) was applied to an entire field.

**Use Lexicon in Substantive Area of Inquiry:** In applying an existing lexicon in a specific domain, a more general model is used in order to explain the practices in a more narrow (i.e., substantive) field. For instance, Grewal et al. (2006) used well-established SNA concepts such as embeddedness and related them to success in the context of open source software development. Typical theory-testing studies would also fall into this category, as a general lexicon is applied to explain observations made in a sample of a population.
The four modes as described above are idealized cases, and in research practice we often use aspects of different modes.

**Discussion and Conclusion**

“...discovery cannot be stopped, but breaks through both verifications and conceptual schemas to provide us with very interesting and important theory.” —Glaser & Strauss 1967, p.185

Glaser and Strauss led a revolution of sorts in social analysis. Through a program of intense attention to empirical data, they legitimized a way to generate novel theory that can revitalize a stale discourse. Some argue that organizational and IS literature may be stagnating (M Davison, 2010) or not reaching its potential (Grover, 2013). Now, particularly given the opportunity that the data explosion provides, it is not the time to close down methods for theory generation that are grounded in empirical data. At the same time, it is important to capitalize on the maturity of GTM, and to encourage further methodological attention in this regard. By proposing a broader grounded theory paradigm, of which GTM is a methodology, we look to accomplish this.

The framework we proposed leaves room for different modes of grounded theory development, and suggests that GTM as described in seminal works (e.g., Charmaz, 2006; Glaser and Strauss, 1967; Strauss and Corbin, 1990) is but one way to do so. The framework has some important implications for IS research. While GTM has been alleged to be used as a toolbox for coding data in IS research, instead of developing theory (Matavire and Brown, 2011; Urquhart et al., 2010), we contend that the opposite should be the case. Not only should we use GTM in order to build theory, but as a discipline we should be open to rigorous approaches to developing theory that can be creative and adaptive, where the common denominator is that of emergence. At the same time, more recent developments in computational social science (Lazer et al., 2009) promise an unprecedented access to different forms of data, thereby calling for alternative inductive approaches that move beyond the manual consideration of qualitative data only. After all, the big data available to IS scholars is rife with opportunity to rethink phenomena in fundamentally different ways, driven by intensive empiricism (Latour, 2010). If one were to compare this opportunity in social science to physics, “it is as if every physicist had a supercollider dropped into his or her backyard” (Davis p. 696).

The IS field is particularly well-positioned to lead this revolution in social research. First, as a discipline, we investigate those phenomena that have made the “big data revolution” possible- the phenomena that make computational social science (Lazer et al., 2009) possible in the first place. Second, our discipline is devoted to investigating complex socio-technical settings that require us to make sense of large amounts of data that pertain to the interaction of ‘the social’ and ‘the technical’ (Orlikowski, 2007). Third, there is a very real need to develop novel and accurate theory grounded in large amounts of data instead of “working” existing theories, as we are challenged to further develop our intellectual core (Webster and Watson, 2002).

Figure 3 illustrates the process of building theory from big data, thereby highlighting the importance of a lexicon in this process.
Traditional qualitative GTM (see left side of the diagram) starts with a sample and generates concepts through constant comparison. In this process, concepts are derived from the data and are constantly compared to incoming data. Identified concepts may thus extend an existing lexicon and provide the building blocks to develop theory. At the same time, an existent lexicon may be applied as the analyst uses theoretical codes (e.g., Glaser’s earlier mentioned coding families) in order to make sense of the data.

Computational analysis of big data looks for patterns in the data, but these patterns only make sense when filtered through a lexicon such as social network analysis or sequence analysis (see the right side of the diagram). The coding strategies used in GTM are too manually intensive for large amounts of data. Identified patterns can both extend the lexicon and build the foundation for novel theory. Together we see a general model of theory development for empirical data.

In this paper, we took first step towards a grounded paradigm and highlighted the importance of a lexicon in this process. We have argued that we, as a discipline, are challenged to develop a better understanding of how to develop novel theories. While grounded theory is a promising approach, we can observe that it has not yet been made use of to its fullest potential. Specifically, GTM is typically used as a method of qualitative inquiry, and there has been debate of some of its fundamental premises, most notably the role of prior theory. In this paper, we have argued that we need to focus on the all-important metaphor of emergence, and understand how we can (a) use qualitative, quantitative, and computational data and (b) consider prior theory in our attempts of building and justifying theory. Based on three key concepts in sociological research (level of induction, level of generalization, and lexicon), we have proposed an analytical framework that proposes four major modes of research that either generate or explain lexicons. The framework is intended to provide a basis to inform theory building efforts in the IS discipline.

References


