Computationally Intensive Theory Construction *

Nicholas Berente, University of Notre Dame
Aron Lindberg, Stevens Institute of Technology
Shaila Miranda, University of Arkansas
Hani Safadi, University of Georgia
Stefan Seidel, University of Cologne

Introduction

Computationally Intensive Theory Construction (CITC) is a research genre focused on generating theoretical insights by combining computational techniques with human interpretation. CITC involves intensive interaction between patterns identified using computational techniques and established theory in an effort to move theoretical knowledge forward (Berente et al 2019; Lindberg 2020; Miranda et al 2022a). Computational analysis extends human abilities to generate novel insights and to account for more of the complexity of the socio-technical phenomena that we study by allowing us to explore a larger variety of potentially relevant aspects of the phenomena.

CITC is enabled by increasing levels of computational power, which allows for the execution of more sophisticated methods to identify patterns. These methods include network, sequence, and text analysis, as well as machine learning, deep learning, large language models, etc., often in conjunction with each other and with more traditional qualitative, survey-based, experimental, or econometric approaches. But it is important to recognize that CITC is not a brute-force approach to generating theory and does not entirely automate the discovery of theory. Rather, the patterns identified through computational techniques provide the foundation for human sensemaking and interpretation. It is the analyst who decides what elements of a phenomenon are relevant, and how the discovered patterns can move a theoretical discourse forward. It is an intensely iterative process where computational capabilities and human interpretation are combined.

The field of information systems is particularly well-suited to CITC because it is a socio-technical field where many researchers are comfortable with computational techniques. The CITC approach is rooted in a handful of foundational information systems articles (Howison et al 2011; Gaskin et al 2014; Miranda et al 2015; Lindberg et al 2016; Vaast et al 2017; see Berente et al 2019). It is inspired by the grounded theory method (Glaser and Strauss 1967) with its

* A subsequent version of this paper was published in the *Routledge Companion to Management Information Systems,* suggested reference:

Berente, N., Lindberg, A., Miranda, S., Safadi, H., Seidel, S. (2024) “Computationally Intensive Theory Construction,” in (Galliers, Stein, Baiyere Eds) *Routledge Companion to Management Information Systems,* 2nd Edition, Routledge.

focus on developing novel insight grounded in empirical data, and the broader field of computational social science with its focus on leveraging computational capabilities to collect and analyze data at unprecedented depth and breadth (Lazer et al. 2009). Now CITC has become established, with a vibrant and growing body of work drawing on a variety of computational techniques (e.g., Pentland et al 2021; Miranda et al 2022b; Lindberg et al 2022; Gal et al 2022; Bachura et al 2022; Lindberg et al 2024; Pienta et al 2024).

While there is extraordinary promise to the CITC approach, there are also risks and challenges. Any new paradigm requires the establishment of new norms for conducting research and deciding what is legitimate practice within that paradigm. The CITC genre is fairly new and it is still in formation. Researchers, reviewers, and editors new to the approach may not be familiar with its goals and practices. They may, for instance, struggle to understand the difference between CITC and other computational, inductive, and abductive approaches to research. Further, there is no single template for writing CITC papers, so researchers may struggle to articulate their findings in a convincing way. In this chapter, we look to help those interested in CITC to think about some of these challenges.

The CITC Genre

CITC, at its core, involves the identification of patterns (Miranda et al 2022a). Researchers use computational techniques to form patterns and draw inferences from those patterns to construct novel theoretical insights. We use the term “construct” to emphasize that theory is not discovered, but rather is developed through an intentional, creative act. Our emphasis on novelty implies novelty to a particular scholarly community where the theoretical insight moves the cumulative tradition of that community forward.

Indeed, there are often a variety of equally valid ways to interpret a given phenomenon, depending on the lens one brings to bear upon the situation. In information systems research, for example, when studying online platforms involving both human actors and machine actors (“bots”), one might focus on bot activity to gain insight into how bots disseminate information (Salge et al 2022), how bots impact human-to-human interaction (Safadi et al 2024), or how bots can help to coordinate work (Hukal et al 2019). The researcher’s goals, the perspectives they draw on, and the methods they use will shape the patterns they can identify and the inferences that they make. Conclusions from any one perspective are not somehow better than conclusions from any others. The world does not determine any particular interpretation of itself (Quine 1951); rather, humans construct their interpretations with reference to the goals, methods, and theoretical perspectives that they use.

CITC offers a pragmatist approach to scholarship that is ontologically agnostic. It is based on an epistemology that is skeptical of absolute knowledge claims, using the consequences of action and other forms of inquiry to validate knowledge claims (Dewey, 1938). Such inquiry is conducted through abductive iteration (Lindberg 2020), which includes the use of induction from data as well as intensive interaction with the scholarly literature to identify potentially fruitful

conceptual and theoretical framings (Berente et al 2019). Abduction is a method of inquiry (Locke, Golden-Biddle, & Feldman, 2008) by which indeterminate situations are iteratively made more tractable by aligning empirical patterns with theoretical explanations (Dewey, 1938). Pragmatism invites researchers to draw on a multitude of analysis methods to identify varied patterns in the empirical world. Each empirical phenomenon can be viewed through a variety of lenses and there can be numerous and overlapping explanations of the same phenomenon. Thus, theorizing always unfolds in relation to some existing discourse. Existing cumulative traditions of scholarship provide the lenses for scholarly contributions. Therefore, researchers need to be familiar with the theoretical discourse they engage with and need to understand its vocabulary, or "lexicon." Pragmatists seek to make a contribution to knowledge by enabling more effective action in the world. Drawing on pragmatism, we next discuss four key central elements of CITC: (1) pattern, (2) method, (3) lexicon, and (4) contribution (Miranda et al 2022a see Figure 1):

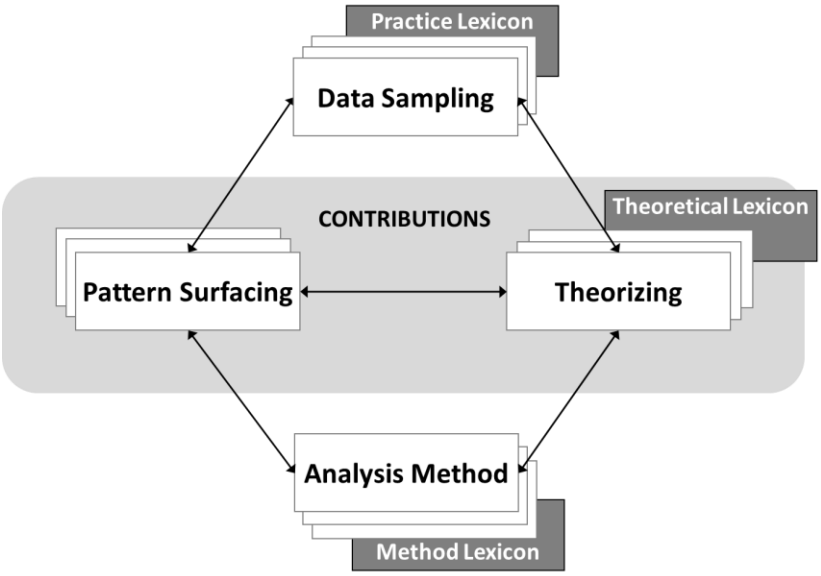

Figure 1: Elements of CITC from Miranda et al 2022a

Pattern: Patterns are descriptions of regularities identified in data that can be expressed in terms of associations among concepts (Berente et al 2019). That is, in surfacing patterns, researchers choose the concepts that matter to them and use analytical techniques to identify associations among these concepts. Patterns are inferences that depend on the data sample, computational techniques, and the operationalization of concepts used. Corroborating patterns–or aspects of patterns–with prevailing understandings, alternative measures, and multiple techniques can help establish confidence in the patterns identified (Miranda et al 2022a). Unlike traditional methods that primarily test predefined hypotheses, CITC uses computational methods to explore data by identifying salient patterns which provide input for human sensemaking and interpretation. Patterns are developed inductively from data and highlight salient relationships between concepts, and theoretical insight is generated as the researcher

constructs a plausible explanation for the occurrence of those patterns (Lindberg 2020) and relates them to a theoretical discourse (Miranda, et al. 2022a).

Method: Computational methods play a pivotal role in CITC, particularly in the identification of patterns from large datasets and for the exploration and validation of theoretical insights derived from those patterns. Computational approaches used in CITC include machine learning, network analysis, sequence mining, and text mining (e.g., Miranda, et al. 2022b, Vaast et al. 2017). These methods enable researchers to uncover novel patterns that might not be evident without the computational capability to process large volumes of data efficiently (Lazer et al. 2020). Computational methods not only facilitate the discovery of new patterns that provide input for theorizing, but also rigorously test these patterns for precision and boundary conditions. CITC emphasizes that whatever patterns are identified, these should not only be statistically significant (if statistical significance is applicable to the study), but also theoretically meaningful, thus contributing to the theoretical discourse (Miranda, et al. 2022a). This requires the use of and engagement with that discourse's lexicon.

Lexicon: Researchers do not develop new concepts out of whole cloth every time they do research. Instead, they draw on the language of a given cumulative tradition of research - the theoretical lexicon. The concepts used to understand the elements identified in patterns indicate the cumulative tradition to which the researchers seek to contribute. If academic research is a conversation (Huff 1999), then the concepts and their theoretical relationships that make up the patterns signal the conversants with whom the researchers seek to engage. Those concepts are theoretically loaded, in that they embody pre-theoretic understandings of that community (Habermas 1984). The conversants' lexicon helps to identify the "general archetypal" abstraction of a problem (Rai 2017). The theoretical lexicon must be intentionally and clearly reconciled with the methodological lexicon (the language used in the methods to identify associations and operationalize the concepts) along with the lexicon of practice (the language used in the world itself). While these may align, typically they do not, and therefore it is up to the researchers to actively reconcile them.

Contribution: The core concern of CITC is to generate theoretical insight that can make a contribution to scholarly discourse. This can involve either generating new theory or identifying patterns with theoretical implications (Miranda et al 2022a). Generating new theory requires the researcher to propose new or alternative explanations for a given phenomenon, whereas patterns with theoretical implications might simply point to a new phenomenon or identify concepts that represent salient elements of a phenomenon that have not been identified or explored before. Either way, researchers must appeal to one or more cumulative traditions and show how this work extends the understanding of those traditions. Because any particular insight is always limited by the data sample, and the approach is intended to generate novel perspectives, validation of any particular inferences is not always possible through strictly empirical methods (Miranda et al 2022a). Indeed, theoretical insights are always relative to existing theoretical understanding, and by drawing on existing knowledge, researchers can help isolate what is new about their particular contribution (Rai 2017). There is no shortcut around a

fundamental understanding of existing, abstracted knowledge (i.e. theory) for a particular body of conversants in justifying a contribution from CITC.

Stopping Rules

Undertaking a CITC project and crafting a manuscript can be daunting to authors. Guidelines for authors and reviewers are still emerging. Absent extensive research within this tradition and codification of success criteria, reviewers are uncertain about how to judge a work and authors are often held to standards appropriate to other genres, but not necessarily to CITC. Researchers can always collect more data, use different methods, or generate additional theoretical statements. When is the right time to stop? Miranda et al. (2022a) offered "stopping rules" in an attempt to alleviate some of this uncertainty and provide some guidance about scoping appropriate CITC contributions. Next we consider this guidance with respect to three questions that researchers bring to their CITC analysis: Is more data always better? When can I stop adding methods? Is my theoretical development adequate?

Is More Data Always Better? The big data era offers researchers an embarrassment of riches, with a multitude of ways in which our actions, words, and representations can be captured, codified, and quantified. Understandably, authors with access to interesting datasets want to explore them thoroughly. Often, the expectation is that casting a wide net attenuates the risk of not being able to develop a phenomenon-based contribution. Authors may also believe that using as many data elements that describe a phenomenon in their analysis allows them to tell a richer and more comprehensive story about the phenomenon. Many intuitively believe that leveraging more data gives them more statistical power to provide a more rigorous model of a phenomenon, while also addressing the generalizability challenges being levied at abductive work (e.g., Bamberger 2019). However, this is not always the case.

It is important to keep in mind that the first obligation of CITC research is to generate novel or surprising insight. While computational techniques allow the researcher to attend to the entirety of large datasets with a variety of elements (such as features and variables), investigating too broad a dataset and too many elements can actually get in the way of generating theoretical insight. While computational approaches expand the researchers' cognitive processing capabilities, they can also have an adverse effect when too many patterns using too many elements are generated at an arbitrary level of analysis. The researcher must still make sense of the patterns that surfaced. The researcher is often more likely to be able to understand patterns that emerge from a subset of elements and determine what is relevant to the conversant discourse. Undertaking the exploratory data analysis that should be foundational to any CITC effort is often more tractable with a focused dataset and a limited set of elements. Subsets of data may enable researchers to surface deeper insights. One reason for this is that researchers may already have unique vantage points on a subset of the data. For example, someone studying Yelp reviews, who focuses on her local market will have a deeper understanding of the data for that market. She therefore may be better able to identify and explain the anomalies that are prized in abductive work. Further, the multitude of factors that may moderate a causal model in a large dataset may mask the underlying causal mechanisms.

In short, sometimes a well bounded and focused study can help to construct theory better than casting a wide net.

Of course, focusing on particular subsets of data to generate theory may cause some to question the generalizability of such work (e.g., Bamberger 2019). But such an objection is misguided - it is predicated on the assumption that universal and context-free knowledge is somehow achievable and desired from a particular study. But this perspective has been debunked in a variety of ways (Lee and Baskerville 2003; Tsang and Williams 2012; Lincoln and Guba 2000). Findings from large samples are not necessarily more inherently generalizable. Generalization is a human act, where the researcher makes a claim that a particular finding applies to a theory, to another context, or to another time (Tsang & Williams 2012). No particular generalization is perfectly defensible and universal; all generalizations require human judgment. Rather than seek universal generalizability, CITC researchers are often better served deeply understanding a limited subset of data and contributing to a cumulative knowledge tradition. Behfar and Okhuysen (2018: 325) suggested that the abductive work that the CITC researcher should "develop a narrow explanation that is appropriate for local observations." Then it is up to the researcher to situate their observations in the appropriate cumulative theoretical tradition in a plausible and convincing manner.

When Can I Stop Adding Methods? Scholars have often advocated methodological pluralism in phenomenon-driven, data-driven theorizing (Miranda et al. 2022a; Lamont and Swidler 2014; Lindberg 2020). Different methods can help to triangulate, corroborate, and further explore patterns in data and the associated explanations. On the other hand, sometimes a single method can identify an interesting pattern that poses an interesting contribution to a discourse.

Consider, for example, two uses of dimensionalization and category surfacing via topic modeling. In a relatively early use of topic modeling in social science, DiMaggio et al. (2013) applied it to explore the progressive politicization of funding for the arts between 1986 and 1997. The authors undertook a range of analytic moves in addition to topic modeling, including ordinary least squares (OLS) regressions. In another example, Croidieu and Kim's (2018) used topic modeling to investigate the legitimation of the lay expertise of ham radio operators. Here, the authors used topic modeling to automate their discovery of first-order concepts, and they conceptually constructed a plausible theoretical argument from the arrangement of these topics over time. Both examples drew on topic modeling, but the former added regressions to support claims about patterns, whereas the second one did not. Both, of course, required human interpretation and a qualitative interaction with computational results (as topic modeling inevitably requires). Is one with regression somehow superior to one without? Clearly this is not the case. They are both excellent studies, but their goals are different. DiMaggio et al (2013) were looking to draw a causal argument and therefore regression made sense, whereas Croidieu and Kim (2018) were characterizing mechanisms and could do so with topic modeling as the only computational technique.

Therefore, just as more data is not necessarily better, researchers should not add methods simply for the sake of adding methods. A method must contribute to the argument that researchers are looking to make. Too much methodological pluralism can derail a project,

increasing its scope and complexity beyond what is digestible within a single manuscript and compromising the coherence of the work. However, often there is some combination of methods that makes sense. Particularly since a particular dataset may be limited, and in such cases triangulating with different methods can help (e.g., Miranda et al 2022; Pienta et al 2024). Some research combines singular computational techniques with manual qualitative analysis. This approach can enhance the "thin" analysis of a large digital trace dataset using computational tools with in-depth qualitative inquiry (Lindberg et al 2022; Lindberg et al 2024). Deciding on what data to use goes hand-in-hand with deciding on the methods and the goals of the research, and these interrelated choices determine a study's scope and boundaries. Being intentional and clear about these decisions can not only help analysts to make their work manageable but can also help to preemptively manage reviewer expectations.

Combinations of methods and data must be in the service of contributing to the discourse and established validity criteria apply (Miranda et al. 2022a) and must serve the goals of the project. In each CITC research process scholars thus need to consider the discourse that they are part of, both within the wider literature, as well as within a particular review process, and calibrate their methodological approach so that it helps to communicate their findings and attendant insights in an effective manner to other participants within the scholarly discourse.

Is My Theoretical Development Adequate? Since the objective of CITC work is to generate novel or surprising insight, expectations of a theoretical contribution can be daunting. Stopping rules with regards to a theoretical contribution, such as originality and importance, are subjective and cannot ensure that such a contribution is sufficient. The question of whether a contribution has been made is negotiated between authors and editors.

Of course, one can generate novel causal theoretical contributions using CITC. Causal inferences about regularities among variables are typically thought to be the most useful type of theoretical contribution (Maxwell 2012; Donmoyer 2012). Traditionally, the experimental method has been the holy grail for establishing causal inference because of the ability to establish precedence and control. Researchers have devised ways to approximate experimental approaches in studying empirical phenomena using econometric techniques. These techniques have created sophisticated, empirical means for establishing causal inferences through a variety of practices, including instrumental variables, Heckman models, difference-in-difference models, and propensity score matching. This is indeed a powerful set of techniques that do a phenomenal job in isolating probabilistic causal influence among variables. Yet this focus on a particular view of causality is not without problems. Even researchers within the most causally-focused traditions of information systems research have bemoaned that "the explosion of attention to causality … [has] seen several unwelcome consequences" (Mithas et al 2022, p.iii). These unwelcome consequences include increasingly narrow research questions which enable causally robust, yet largely unsurprising findings, along with an intense focus on ruling out competing explanations and tunnel-vision about specific methods for doing so (Mithas et al. 2022). It is important for CITC researchers to understand that causal research is indeed important, but it is not the only way to conduct useful scholarship.

There are (at least) two alternative positions to the prevailing causal approach rooted in the experimentalist paradigm – process and configural explanations. These explanations challenge the determinism and temporal ordering presumed by causal inference - that cause is indeed unidirectional from an exogenous to an endogenous variable. Indeed, most empirical phenomena instead arise from complex interactions among variables that combine and mutually shape the results (Lincoln and Guba 1985: 38). Process explanations are often more appropriate in research that seeks to characterize this complex interaction. A second challenge to the experimentalist paradigm involves the goal of generalizability–that posited causal mechanisms will appear with regularity across empirical contexts (e.g., Donmoyer 2012). Instead, many argue that causation is "fundamentally local rather than general" (Maxwell 2012), and that attention to broad statistical regularities will not uncover local particulars. The focus on broad, general regularities often uncover less interesting, obvious, and unimportant findings - instead the focus on subcategories - the rare, exceptional, and particular will lead to more useful findings (Starbuck 1993). Given the focus of CITC projects on emerging technology-related phenomena, our studies often represent edge cases, rather than those that are commonplace at the time.

Certainly, answering *how* and *why* questions (Bertilsson 2015) that are of interest to some community of inquiry and conveying findings using some propositional or process grammar (Cornelissen 2023) is a well-understood way of making a theoretical contribution that most would agree on. However, rich descriptions or insightful conceptualizations of phenomena can also put us on the path to a theoretical contribution. Rich descriptions provide other researchers with the vicarious experience of a phenomenon (Lincoln and Guba 2000). For example, Gill (1995) theorized information technology as an impediment to organizational learning from Harvard Business School cases on the Mrs. Fields Cookies and Batterymarch companies. Insightful conceptualizations of a phenomenon that challenge the taken-for-granted can likewise be generative. For example, the blogosphere polarization description by Adamic and Glance (2005) motivated subsequent inquiry into such dynamics in digital communication. Thus, answers to *what* questions that describe the phenomenon (Bertilsson 2015), along with configural and contingency explanations offer an alternative to the identification of general relations.

In the end, it is clear that there are many paths to a contribution, including causal, process, and configural arguments. But whichever approach requires an attitude of humility on the part of the researcher and invokes an attitude of mindful engagement on consumers of the research. A CITC project concludes with theoretical arguments that invite subsequent research to test, replicate, and extend.

What CITC Is Not

Because CITC looks to explore patterns in computational data, naive critics of CITC may dismiss the technique as simply p-hacking, cherry-picking, or HARKing (hypothesizing after results are known) and, therefore a questionable research practice (Andrade 2021; Bamberger

2019). After all, the patterns identified are necessarily a function of the data sample, and every sample has limitations (Tsang & Williams 2012). Thus researchers may simply be capitalizing on chance and noise and thus may be identifying patterns that are fleeting and not replicable.

CITC is markedly different from these practices. First, p-hacking is a vestige of small-N quantitative research, where small changes to sample size or model specification may push a particular coefficient over the threshold for statistical significance. This practice is increasingly irrelevant, since large digital trace data tends to make a larger share of coefficients statistically significant. Hence, we increasingly need to rely on analyses of effect sizes rather than statistical significance to assess whether the coefficients of a statistical model are substantively significant (Mertens and Recker 2020). Second, HARKing is problematic when it is done secretly in the frontend of a paper, since this misleads the reader about the actual research process. When HARKing is done transparently in the discussion section, this represents a more honest description of how many research processes unfold, and therefore a desirable activity (Hollenbeck & Wright 2017). This marks the difference between the hypothetico-deductive approach that requires up-front theory development, and an abductive approach consistent with CITC's pragmatist underpinnings. A key aspect of pragmatist abductive inquiry is fallibility - the view that every theoretical insight is tentative and subject to refutation and extension over time. When propositions are developed in the discussion section, this provides an opportunity for scholars to explicitly and transparently lay out the rigor of their abductive, iterative research process, thus indicating how the developed propositions are grounded in a particular dataset and a particular theoretical framing (Hollenbeck & Wright, 2017). Of course, such propositions can then be subjected to rigorous testing across contexts and time in a hypothetico-deductive way. CITC provides the empirical foundation for developing those propositional statements.

Indeed, each study is merely a brick in the edifice of our knowledge (Tiwana & Kim 2019), and building a strong cumulative knowledge base requires an accumulation of abstracted knowledge in a particular research tradition over time. Theory is this accumulation of abstracted knowledge–the concepts and the associations that have been established among a community of researchers. Therefore, any CITC contribution requires understanding this cumulative tradition, its concepts, and associations, and making a generative contribution using data. It is not blind data reporting but an iterative process whereby researchers explore their operationalizations of concepts, how they establish associations, and triangulate and corroborate across methods–all the while engaging substantively with the research of conversants (Berente et al 2019). To establish the validity of inferences, one can adopt common quantitative and qualitative techniques, but one can also use a deep understanding of a theoretical tradition to argue for why insights are justified, even in situations where the data may not be perfect (Miranda et al 2022a). In the end, CITC is like detective work (Miranda et al 2022a) that involves a trail of evidence that uses some sort of computational technique, along with argumentation, to justify contributions.

CITC researchers use computational techniques, sometimes a variety of them, combined with creative human investigation to explore data in order to generate theoretical insights. But is this just another name for the grounded theory method? Mixed methods research? Data science?

Computational social science? Computational design? If there are a variety of already established, related methodological approaches, why do we need a new concept? To answer this question, we will first relate the concept to traditional approaches and then to the more recent data and computationally intensive approaches to highlight how we see CITC as a *genre* for conducting research using a variety of data source and combining computational methods with human interpretation to generate theoretical insight of various forms..

First, it is important to note that CITC, although recently articulated, is not a new concept. Computational techniques such as sequence analysis (Abbott 1995) and social network analysis (Wasserman & Faust 1994) have been around for a long time. Scholars using these techniques have been generating theoretical insights for decades following a similar approach to the one we articulate here. In information systems research, the trailblazing work in CITC was consistent with this work, drawing on sequence analysis (e.g., Lindberg et al 2016) and social network analysis (e.g., Vaast et al 2017) to make theoretical contributions. Further, exploratory approaches to theory development exist in many fields, particularly in the grounded theory methodology in social science and computational theory discovery approaches in the sciences (Berente et al 2019). Rather than an altogether new approach, CITC is a set of formulations (i.e. Berente et al 2019; Lindberg 2020; Miranda et al 2022a) to help researchers think through the process and to offer reviewers and editors an updated framework to reference. After all, not all researchers are familiar with the grounded theory methodology–and those who are familiar tend to be qualitative researchers. One can consider CITC a general variant of grounded theory methodology or an approach that is consistent with the spirit of a broadly conceived grounded theory (Walsh et al 2015) that expressly and intentionally adapts computational techniques as part of the method. At the same time, this approach avoids the baggage that grounded theory reviewers require, and with which non-devotees are not familiar. So, although CITC is consistent with grounded theory methodology, it is not reducible to grounded theory.

Similarly, CITC is not necessarily mixed methods research. CITC may involve mixed methods, but it does not require them. One might draw on a single method. Further, in their typical formulation (i.e., Tashakkori & Teddlie 1998), mixed methods generally involve combining quantitative and qualitative approaches. Although CITC can involve computational and qualitative techniques (Linberg 2020), this is not the same as traditional quantitative and qualitative mixed methods. Computational approaches tend to be quantitative, but not necessarily quantitative in a traditional sense. Next-generation methods for pattern recognition increasingly blur the line between quantitative and qualitative methods, such as machine learning and deep learning for content analysis, or the use of large language models to interrogate textual datasets. So CITC can indeed include mixed methods in the general sense, but not necessarily so, and it mixes methods in much different ways than the traditional distinction between quantitative and qualitative methods.

Further, CITC is neither synonymous with data science nor computational social science, but one might consider CITC to be a subset of both. Data science is sometimes described as the “fourth paradigm for scientific discovery” (Abbasi et al 2023), whereby analysts use computational techniques to gain insights. While, of course, this fits a description of CITC, it is

important to note that data science is a much broader term. Data science refers to a wide variety of goals and approaches and is often not concerned with theoretical insight. For example, some approaches to data science might look for predictive power over specific instances rather than theory–they are atheoretical. Sutton and Staw (1995) referred to this as abstracted empiricism. Data science without the goal of generating theory is not CITC, much as conceptual description of qualitative data is not grounded theory. Computational social science (Laser et al 2009) can also be quite consistent with CITC; both involve generating insights using computational approaches. However, computational social science is the broader term and could also include hypothesis testing research, as well as research on methods and a-theoretical work. Further, CITC often involves some qualitative analysis, which is not typical in prevailing computational social science.

Finally, there is the information systems research tradition of computational design (Abbasi et al 2023), which is often computationally intensive and generates artifacts and can be theoretically informed, but is not CITC unless those artifacts generate data used to extend existing theoretical perspectives in exploratory ways.

Constructing Theory from Data through Patterns

Describing patterns necessarily requires the use of conceptual labels to capture multiple instances of what is observed as well as to describe co-occurrences between those instances. Before researchers can represent data in terms of a pattern, they need to choose a general structure to make sense of the data. Data does not come ready-made for representing patterns—the data needs to be structured according to a particular language, or lexicon, for representation. Following Habermas (1984), we refer to this as the “pre-theoretic” lexicon for identifying the pattern (Figure 2).

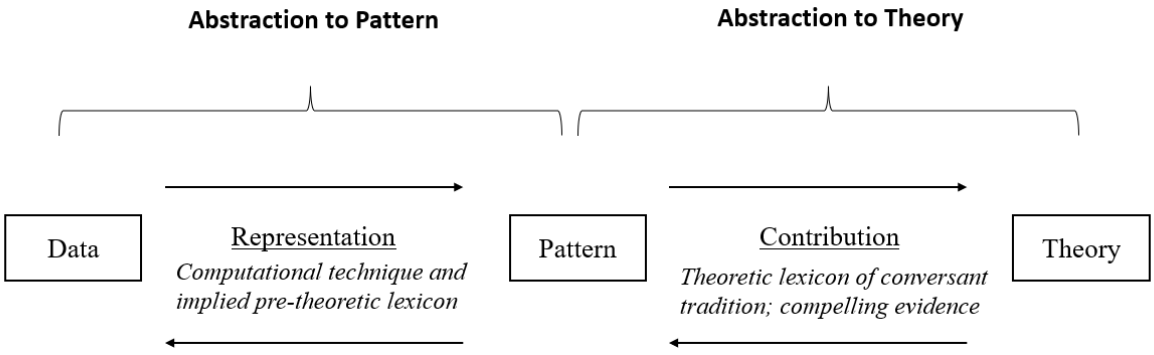

Figure 2: Theory Development as a Process of Abstraction

The process of identifying regularities is a process of abstracting from data to pattern. The same data can lead to different abstractions–different patterns–depending on the computational approach used. Through abstraction, the analyst removes elements of the phenomenon and maps one representation, like the raw dataset, on another, such as the pattern (Giunchiglia and

Walsh, 1992). Different computational approaches imply specific pre-theoretical lexicons that shape the subsequent patterns (see Table 1). For example, researchers use a variety of computational approaches to study open-source software projects. Some research projects focus on developers as nodes and their joint activity as evidence of relations, or edges in a network—and therefore they identify network patterns (Singh et al 2011; Maruping et al 2019; Tang et al 2020). Other research projects look at essentially the same sort of open-source data but emphasize sequences of activities among developers and essentially take the research in a much different direction (e.g., Lindberg et al 2016; Bradley et al 2020). Instead of nodes and edges, sequence analysis requires temporally ordered events that are distinguished by distance and variation among sequences. Instead of network concepts, sequence analysis draws on constructs such as sequential variety, task complexity, order and activity variation, coherence, and inertia. We call this process *abstraction to pattern*.

Table 1: Patterns Identified with a Variety of Computational Methods

Focus	**Computational method family**	**Specific computational method**	**Method lexicon**
Social structure	Social network analysis	Graph modeling	Nodes, edges, paths, dyads, triads, brokerage,
Routines, practices	Sequences analysis	Hidden Markov models	Elements, sequences, events, states, transitions
Discourse	Text analysis	Topic modeling	Corpora, documents, words, bag of words (word association)
Social systems	Computational modeling	Agent-based modeling	Agents, interactions, systems, emergence, and complexity

Theory construction flows from the pattern as researchers identify some regularity in a set of empirical observations (e.g., a certain change in an input variable leading to a certain change in an output variable, or an activity regularly following another activity). But it further requires that researchers relate this pattern to existing, expected theoretical explanations. That is, relate the pattern to a cumulative tradition of knowledge to see if and how it is consistent and different from the expectations implied by that cumulative tradition. We refer to situating patterns in theoretical discourse as *abstraction to theory*.

The goal of CITC is to construct theoretical statements, which are statements of association among concepts. In order to construct a theoretical statement from a pattern identified, one must translate it to the theoretical lexicon of the conversant scholarly community. That is, abstract patterns are translated to abstract theoretical statements that are related to some theoretical discourse through the concepts that constitute those statements. Each community

has its own lexicon through which they linguistically reconstruct phenomena to make sense of those phenomena (Habermas 1984). CITC researchers must frame their contributions using the terms for concepts that the conversants understand. To do so, they need to be aware of how those conversants currently describe, explain, and make predictions about similar phenomena. This construction of explanations involves translating representations into the language of a community of conversants is what we refer to as *abstraction to theory*. This requires that CITC researchers explain how their theoretical statements relate to the existing cumulative tradition in the discourse.

An anomaly is an unexpected pattern relative to existing knowledge. Anomalies are prized in CITC research, since they trigger researchers to employ abductive reasoning to explain them and relate them to regularities. Anomalies provide the seed for further exploration to elaborate on the patterns identified (Sætre and Van De Ven 2021).

Discussion

At the dawn of the 21st century, researchers began to appreciate how so much of human activity had become digitally enabled. Organizational processes, customer journeys, social networks, mobile activity–virtually everything now leaves digital traces. Many recognized that this provides a wonderful opportunity to gain new insights about human, organizational, and social phenomena across fields (Lazer et al. 2009; Latour 2010; Davis 2010; DiMaggio 2015).

Information systems research was no exception–"big data" provided a unique opportunity for our sociotechnical field to make new and interesting contributions to knowledge (Howison et al 2011; Agarwal & Dhar 2014; Abbasi et al 2016). However, the information systems field (along with adjacent fields of operations and management) responded to the big data opportunity with essentially one main approach to inference from big data: unidirectional exogenous causation using econometric methods. This stands to reason since large datasets lend themselves to econometric analysis (Varian 2014) because the data is constructed in the world and thus can lead to insight close to the phenomenon in question (Schumpeter 1933). Further, econometric techniques offer sophisticated, empirical ways of isolating effects and establishing the validity of probabilistic inferences about unidirectional, exogenous causation. If the only patterns that research is interested in are unidirectional, exogenous, and causal, and we had perfect datasets to deal with counterfactuals in every study, then there would be no problem.

But there is a problem. Many phenomena are not so readily isolated and involve mutual constitution or simultaneity. If only one type of causal inference is admissible for research, researchers will seek datasets that meet the ever more stringent requirements of econometric analysis, and they will not be studying some interesting phenomena with less perfect datasets. This is commonly referred to as the "streetlight effect" (Rai 2017). Rather than exploring research questions programmatically for strong theoretical inferences (Tiwana & Kim 2019), a single approach to analysis and inference limits the datasets that can be studied (Abbasi et al. 2023). Dysfunctions from a preponderance of econometric analyses are beginning to be

evident. For example, the restaurant rating platform, Yelp, indicated for a time whether businesses were “black owned” or not. The dataset that the platform made available fit the econometric model of a field experiment perfectly. As a result, many research teams conducted studies on this data - Abbasi and colleagues (2023) identified 108 research teams in total. Of course, this is an interesting dataset that led to some interesting observations, but one might be correct in pointing out that perhaps not quite so many of our most talented researchers should be working on the exact same dataset and the same problem. This example illustrates the danger of an over-reliance on a singular approach to inference which admits ever more stringent requirements for data. It encourages the streetlight effect.

CITC offers a way out. CITC is a general approach that relies on deep engagement with both data and theory. By focusing on the creative, generative practice of theory construction and making strong justifications for the steps one makes, the data does not necessarily have to be perfect. Of course, the researcher needs to be rigorous in their method and their inferences (Miranda et al 2022a), but it is important to keep in mind the creative act of theory construction. To construct new theory, one arguably does not need data at all. Theory can be constructed with deep attention to existing knowledge and extending it through rigorous thinking and appeals to logic. Certainly, adding data to this process adds considerable value. CITC offers a way to thoughtfully do this through using computational approaches as an integral part of the theory construction process.

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. Journal of the association for information systems, 17(2), 3.

Abbasi, A., Chiang, R. H., & Xu, J. (2023). Data Science for Social Good. Journal of the Association for Information Systems, 24(6), 1439.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. Annual review of sociology, 21(1), 93-113.

Adamic, L. A., & Glance, N. (2005, August). The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (pp. 36-43).

Agarwal R, Dhar V (2014) Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. Inform. Systems Res. 25(3):443–448

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. The Journal of clinical psychiatry, 82(1), 25941.

Bachura, E., Valecha, R., Chen, R., & Rao, H. R. (2022). The OPM data breach: An investigation of shared emotional reactions on Twitter. MIS Quarterly, 46(2).

Bertilsson, M. 2015. "On Why's, How's, and What's—Why What's Matter," in Action, Belief and Inquiry—Pragmatist Perspectives on Science, Society and Religion, U. Zackariasson (ed.). Helsinki: Nordic Pragmatism Network, pp. 209-229.

Cornelissen, J. P. 2023. "The Problem with Propositions: Theoretical Triangulation to Better Explain Phenomena in Management Research," Academy of Management Review (forthcoming).

Croidieu, G., and Kim, P. H. 2018. "Labor of Love: Amateurs and Lay-Expertise Legitimation in the Early Us Radio Field," Administrative Science Quarterly (63:1), pp. 1-42.

Davis GF (2010) Do theories of organizations progress? Organ. Res. Methods 12(2):690–709.

Dewey, J. (1938). Logic: The theory of inquiry. Henry Holt.

DiMaggio PJ (1995) Comments on "What Theory is Not." Admin. Sci. Quart. 40(3):391–397.

DiMaggio, P., Nag, M., and Blei, D. (2013) "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Us Government Arts Funding," Poetics (41:6), pp. 570-606.

Folger, R., & Turillo, C. J. (1999). Theorizing as the Thickness of Thin Abstraction. The Academy of Management Review, 24(4), 742.

Gal, U., Berente, N., & Chasin, F. (2022). Technology Lifecycles and Digital Technologies: Patterns of Discourse across Levels of Materiality. Journal of the Association for Information Systems, 23(5), 1102-1149.

Gaskin, J., Berente, N., Lyytinen, K., & Yoo, Y. (2014). Toward generalizable sociomaterial inquiry. Mis Quarterly, 38(3), 849-A12.

Gill, T. G. 1995. "High-Tech Hidebound: Case Studies of Information Technologies That Inhibited Organizational Learning," Accounting, Management and Information Technologies (5:1), pp. 41-60.

Giunchiglia, F., & Walsh, T. (1992). A theory of abstraction. Artificial Intelligence, 57(2-3), 323-389.

Habermas, J. (1984) The Theory of Communicative Action. Boston: Beacon.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. Journal of Management, 43(1), 5-18.

Howison J, Wiggins A, Crowston K (2011) Validity issues in the use of social network analysis with digital trace data. J. Assoc. Inform. Systems 12(12):767–797.

Huff, A. S. (1999). Writing for scholarly publication. Sage.

Hukal, P., Berente, N., Germonprez, M., & Schecter, A. (2019). Bots coordinating work in open source software projects. *Computer*, *52*(9), 52-60.

Hollenbeck, John R., and Patrick M. Wright. “Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data.” *Journal of Management* 43, no. 1 (2017): 5–18. https://doi.org/10.1177/0149206316679487.

Latour B (2010) Tarde’s idea of quantification. Candea M, ed. The Social After Gabriel Tarde: Debates and Assessments (Routledge, New York), 145–162.

Lamont, M., and Swidler, A. 2014. "Methodological Pluralism and the Possibilities and Limits of Interviewing," Qualitative sociology (37), pp. 153-171.

Lazer DMJ, Pentland A, Watts DJ, Aral S, Athey S, Contractor N, Freelon D, et al. (2020) Computational social science: Obstacles and opportunities. Science (80-.). 369(6507):1060–1062.

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, et al. (2009) Life in the network: The coming age of computational social science. *Science* 323(5915):721–723

Lee, A. S., and Baskerville, R. L. 2003. "Generalizing generalizability in information systems research," Information systems research (14:3), pp. 221-243.

Lindberg, A. (2020). Developing theory through integrating human and machine pattern recognition. Journal of the Association for Information Systems, 21(1), 7.

Lindberg, A. (2023) “Analysis Chaining: Conceptual and Empirical Framing of Digital Traces.” In *Handbook of Qualitative Research Methods for Information Systems*, edited by Robert Davison, 360–75, 2023.

Lindberg, A., Berente, N., Gaskin, J., & Lyytinen, K. (2016). Coordinating interdependencies in online communities: A study of an open source software project. Information Systems Research, 27(4), 751-772.

Lindberg, A., Majchrzak, A., & Malhotra, A. (2022). How Information Contributed After an Idea Shapes New High-Quality Ideas in Online Ideation Contests. MIS Quarterly, 46(2).

Lindberg, A., Schecter, A., Berente, N., Lyytinen, K., Hennel, P. (2024) “The Entrainment of Task Allocation and Release Cycles in Open Source Software Development" MIS Quarterly, 48(1).

Locke, K., Golden-Biddle, K., & Feldman, M. S. (2008). Making doubt generative: Rethinking the role of doubt in the research process. Organization Science, 19(6), 907-918.

Mertens, Willem, and Jan Recker. “New Guidelines for Null Hypothesis Significance Testing in Hypothetico-Deductive Is Research.” *Journal of the Association for Information Systems* 21, no. 4 (2020): 1072–1102. https://doi.org/10.17705/1jais.00629.

Miranda SM, Kim I, Summers JD (2015) Jamming with social media: How cognitive structuring of organizing vision facets affects IT innovation diffusion. Mis Quart. 39:591–614

Miranda S, Berente N, Seidel S, Safadi H, Burton-Jones A (2022a) Computationally Intensive Theory Construction: A Primer for Authors and Reviewers. MIS Q. 46(June):iii–xviiii.

Miranda, S. M., Wang, D. D., & Tian, C. A. (2022b). Discursive Fields and the Diversity-Coherence Paradox: An Ecological Perspective on the Blockchain Community Discourse. Management Information Systems Quarterly, 46(3), 1421-1452.

Mithas, S., Xue, L., Huang, N., & Burton-Jones, A. (2022). Editor's comments: Causality meets diversity in information systems research. Management Information Systems Quarterly, 46(3), iii-xviii.

Pentland, B., Vaast, E., & Wolf, J. R. (2021). Theorizing process dynamics with directed graphs: A diachronic analysis of digital trace data. MIS Quarterly, 45(2).

Pienta, D., Somanchi, S., Vishwamitra, N., Berente, N., Thatcher, J., (2024) “Do Crowds Validate False Data? Systematic Distortion and Affective Polarization," MIS Quarterly.

Rai, A. (2017). Avoiding type III errors: formulating IS research problems that matter. MIS QUARTERLY, 41(2), III-VII.

Safadi, H., Lalor, J., Berente, N. (2024) "The Effect of Bots on Human Interaction in Online Communities," MIS Quarterly.

Salge, C. A. D. L., Karahanna, E., & Thatcher, J. B. (2022). Algorithmic Processes of Social Alertness and Social Transmission: How Bots Disseminate Information on Twitter. MIS Quarterly, 46(1).

Schofield, J. W. 2000. "Increasing the Generalizability of Qualitative Research," in Case Study Method, R. Gomm, P. Foster and M. Hammersley (eds.). Thousand Oaks: Sage, pp. 69-97.

Schumpeter, J. (1933). The common sense of econometrics. Econometrica: Journal of the Econometric Society, 5-12.

Starbuck, W. H. (1993). Keeping a butterfly and an elephant in a house of cards: The elements of exceptional success. Journal of Management Studies, 30(6), 885-921.

Sutton, R. I., & Staw, B. M. (1995). What theory is not. Administrative science quarterly, 371-384.

Tiwana, A., & Kim, S. K. (2019). From bricks to an edifice: Cultivating strong inference in information systems research. Information Systems Research, 30(3), 1029-1036.

Tsang, E. W., & Williams, J. N. (2012). Generalization and induction: Misconceptions, clarifications, and a classification of induction. MIS quarterly, 729-748.

Vaast E, Safadi H, Lapointe L, Negoita B (2017) Social media affordances for connective action: An examination of microblogging use during the Gulf of Mexico oil spill. MIS Quart. 41(4):1179–1206.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of economic perspectives, 28(2), 3-28.

Walsh I, Holton JA, Bailyn L, Fernandez W, Levina N, Glaser BG (2015) What grounded theory is: A critically reflective conversation among scholars. Organ. Res. Methods 18(4):581–599.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.

Wiener, N. (1988). The human use of human beings: Cybernetics and society (Issue 320). Da capo press.