Introduction

We live in an age of Big Data. In the science of learning, for example, the advent of massively open online courses (MOOCs), educational games and simulations, computer-based tests, and other computational tools for learning mean that the data available is growing exponentially—and not only is the volume of data too large for any human to reasonably process, but the type of data collected may be impossible for a human to interpret simply by reading it, as with clickstream data. A critical question is thus how to use Big Data to study deep learning: that is, to understand how people adopt the ways of thinking and acting that people in some culture of learning use to frame, investigate, and solve complex problems. Fortunately, the technological advances that create Big Data also give us the tools to analyze it. The tools of data science make it possible to analyze larger and larger volumes of data.

If we want to understand and enhance deep learning, however, we need to use these sophisticated technological tools and approaches to understand not just the low-level psychological processes of learning, but how people make meaning of what they are learning. We cannot ask merely “What works?” in isolated laboratory contexts under carefully controlled conditions. Rather, we have to understand how learning interventions work—and that they work—in the real world of classrooms and after school centers, of parents and teachers and students, of kids in maker spaces and collaborating online and learning from automated pedagogical agents.

This has led researchers who are interested in assessing deep learning to use both qualitative and quantitative methods—meaning not just that we are agnostic to method, but that we actively work to combine them in our research. There is a growing literature on mixed methods, but the advent of Big Data changes the balance between empirical and theoretical work, tempting us to think that more data can substitute for deeper understanding.

In what follows, I argue that to assess deep learning, we have to look at what it means to use Big Data in a theoretically robust way. This means more than merely “mixing” methods, making informed choices about which variables to include in data mining, or triangulating correlational studies with case studies. Rather, we need to understand how to think about empirical approaches in a way that uses the power of large-scale analyses to answer how and why questions: to use Big Data for Thick Description of Deep Learning.

The Centrality of Culture

Data, by itself, is meaningless.

If I told you that some person—call her “Sara”—bought $10.87 of gas in Mississippi, what sense would you make of it? You would have no way of knowing if she was visiting Mississippi or living there. Or maybe someone stole her credit card, and she was not there at all. If I also told you Sara bought gas in Louisiana earlier that afternoon and in Texas that morning, you might reasonably think that she is driving from Texas to Mississippi. And if I told you Sara was 19 years old and this was all happening on a Saturday in March when the University of Texas was starting its spring break, you might guess that Sara is a college sophomore driving to Florida for vacation. You could not be sure, of course, but if you
accumulated more and more data consistent with your guess, you would be more and more confident about who Sara was and what she was doing.

Researchers looking at this data qualitatively would describe the idea that Sara is driving to Florida for spring break as an inference. Those with a quantitative orientation we would call it a hypothesis. But either way, the point is to understand what is going on. Without an explanation for what we think is happening, the data itself is meaningless. When we create an inference or test a hypothesis, we transform the data into information: part of a story about something that is happening in the world, or that has happened, or that we think is likely to happen (Devlin, 1995).

This is why a situated perspective is so critical. Human beings traffic in symbols: in action, in talk, in writing, and in making things that mean something to ourselves and to others. The things people say and make and do are interpreted by others who share their culture. Culture is how people understand the meaning of things—and not just the meaning of things themselves, but the web of meanings that connect things to each other, things to people, and therefore people to each other.

Culture is what makes data into information by adding meaning. If researchers want to understand how deep learning happens and why particular approaches to deep learning do or do not work, we need a method for analyzing culture at scale to make sensible analyses of Big Data.

The Alchemy of Research Methods

For more than a decade there has been a movement in social science research toward mixed qualitative and quantitative methods, including foundational work by Art Graesser (see, e.g., Magliano & Graesser, 1991). In such studies, sometimes work is done in parallel, and then researchers triangulate findings from one analysis with another. For example, studies where researchers collect survey data, which they analyze quantitatively, and data from focus groups, which they analyze qualitatively. In other studies, researchers use the results from one analysis to inform another. A study might use qualitative analysis to identify groups to survey. One handbook lists more than 70 different types of mixed method research design (Creswell & Plano Clark, 2010, p. 56 ff.).

In chemistry, a mixture is combination of two elements or compounds in which neither substance is changed. If you add meat to a soup, for example, you do not fundamentally change the meat or the broth. In contrast, a vodka martini is a solution. Once shaken (or stirred, if you prefer), there is no way to separate the vermouth from the vodka.

In approaches that mix quantitative and qualitative methods, one kind of analysis confirms or refutes or extends the other. Ideally, each analysis informs the other in an iterative process. But instead of a pragmatic a mixture of research methods, it is possible to create a research solution, where the power of statistics to examine Big Data and the power of ethnographic methods to provide understanding of deep learning are inseparably linked. Before we explore such an approach, however, it is important to understand the goals and assumptions of quantitative and qualitative methods.

While there are many techniques for quantitative analysis of data, statistical tools depend on the idea of sampling. In a quantitative analysis, data (the sample) is drawn in an unbiased way from some larger population, and the question statistics answer is whether some characteristic of the sample reflects a characteristic of that larger population. In the context of learning research, the sample is often some collection of individual students and the population is all students who are “similar to those in the study,” which explains why quantitative researchers are concerned about recruitment, self-selection bias, and other things that could limit the kind of students who are “similar” to those in the sample. It also explains
why quantitative researchers are concerned about collecting and recording data systematically: statistical tools depend on having data that is consistent, well-structured, and ideally complete.

Quantitative analyses justify claims that observations about a sample generalize to a population by distinguishing between “true effects” or relationships in the population and the normal random variations among individuals. In larger samples, the impact of systematic effects increases because systematic effects follow a pattern (hence the term systematic); the effects of random errors cancel out. In quantitative analyses, larger samples provide more power because they make it possible to justify claims about more effects and relationships in the data.

This kind of non-contextualized generalization is not the goal of qualitative research, which rejects the notion that there are “true effects” that can be identified in subjects or interventions. Instead, qualitative methods take as a premise that observations are produced through the contingent interactions among participants, contexts, and researchers. This means that specifying a priori the kind of structure for data collection that quantitative methods typically require is epistemologically problematic.

There are many approaches to qualitative research, but the overarching goal of qualitative inquiry is to provide some form of what Geertz (1973b) popularized as Thick Description: an explanation of how and why events unfolded in a particular place and time. Some theorists argue that descriptions of causal mechanism are naively realist (e.g., Tashakkori & Teddlie, 1998), but any qualitative analysis has to assert some claim to being more a portrait of the experience of participants than a reflection of the biases of the researchers.

Qualitative researchers use a number of techniques to account for bias—not to eliminate it, that being impossible anyway, but to understand and account for its effect on how the data is interpreted. The data analysis tradition of Grounded Theory provides a useful way of thinking about this problem through the concept of theoretical saturation: An analysis is theoretically saturated when researchers have collected and analyzed enough data such that additional observations confirm existing hypotheses rather than lead to new insights (Glaser & Strauss, 1967). In qualitative analysis, power comes from collecting a great deal of information about a small number of subjects, so that researchers can examine a large corpus of observations to find stable interpretations about what people are doing and why.

Both quantitative and qualitative techniques gain increasing power from collecting more data; however, the concepts of analytic power in qualitative and quantitative research are contradictory. Qualitative analysis tries to provide a thick description of people and their actions in some specific context. Quantitative analysis tries to provide general claims about more subtle differences in observed data. Qualitative analysis needs a large amount of data about individual participants. Quantitative analysis needs data about a large number of individuals. In the context of the finite resources available in any study, this creates an irresolvable conflict.

Or anyway it did, until the advent of Big Data.

**Toward a Research Solution**

Data from MOOCs, educational games, simulations, and other computer-based learning environments are in many cases rich enough to make possible Thick Descriptions of deep learning for large numbers of students. They thus provide an opportunity for learning scientists to combine the tools of meaning-based, qualitative inquiry for understanding deep learning in context with quantitative methods for understanding deep learning at scale. But to accomplish this, researchers studying deep learning need to confront the logistical and epistemological challenge of integrating two different analytic frameworks that have, until
now, remained distinct. We need to develop methods for using significance tests to support Thick Description of the cultures of learning, and thus also the means to conduct situated analyses of Big Data.

There is not room in a single chapter to discuss all the components of recombining qualitative and quantitative methods in this way. Rather, in what follows I will attempt to outline the central conceptual underpinnings of one such approach, called \textit{quantitative ethnography}, with a goal of using this one particular approach to examine the considerations that are critical in any attempt to integrate qualitative and quantitative methods—and thus are central to developing an approach to the assessment of deep learning using Big Data.

Quantitative ethnography begins with the premise that any culture of learning is characterized by a \textit{Discourse} (with a Big D), which Gee (1999) defines as a particular way of “talking, listening, writing, reading, acting, interacting, believing, valuing, and feeling (and using various objects, symbols, images, tools, and technologies)” (p. 25). A Big-D Discourse is a pattern of communication within some community: it is how some group of people makes meaning of the world. Deep learning requires learning more than just basic facts and skills. It means understanding how to talk, think, and act (and value, and feel, and make decisions) like experts in some field. That is, deep learning means learning some Big-D Discourse: learning how to make meaning of the world in a particular way.

As researchers who want to understand and enhance deep learning, then, we need to be able to assess the extent to which students can enact a Big-D Discourse. But we cannot see Big-D Discourse directly. Rather, our data is what Gee calls \textit{discourse} (with a small d): the things people actually “say,” where people express themselves through “bodies, clothes, non-linguistic symbols, objects, tools, technologies, times and places” as well as “ways of acting [and] interacting” (Gee, 1999, p. 25). Small-d discourse is the overt manifestation of communication: what we can observe about how people interact.

To understand a culture of learning, researchers therefore need to find a way to go from small-d discourse to Big-D Discourse—to take specific things that some people said and did and infer their meanings. And a key part of how qualitative researchers make sense of culture is through the process of \textit{coding}.

A Code (with a Big C, because it is part of a Big-D Discourse) describes the culturally relevant meaning of some event. Goodwin (1994) argues that learning in any field means developing \textit{professional vision}: “socially organized ways of seeing and understanding events that are answerable to the distinctive interests of a particular social group” (p. 606). In other words, enculturation requires learning the Codes of some culture. For example, he writes about how archaeologists are trained:

The medium that archaeologists work in is dirt. Students are given a form that contains an elaborate set of categories for describing the color, consistency, and texture of whatever dirt they encounter…. Moreover some of the categories are supported by additional tools of inscription, such as a Munsell Color chart, used by archaeologists all over the world as a standard for color descriptions (p. 608).

Archaeologists do this because they are interested in reading a record of the past from the evidence of the present. For example, when a post decays in the ground, it leaves a different color and consistency of dirt behind. Finding precise distinctions between dirt in different locations thus provides evidence about human settlement in the long past. Archaeologists learn the scientifically (that is, culturally) appropriate meanings—the different Codes—for \textit{DIRT: COLOR, CONSISTENCY, and TEXTURE}.

Different researchers, working in different settings, with different kinds of data, or using different tools of qualitative analysis, use different techniques to identify Codes in a culture of learning. But the
goal of qualitative analysis is almost always to develop Codes that are *grounded*, in the sense that they are (a) based on things that matter to participants in the domain and (b) come from data itself rather than from some existing theory or theories held by the researchers (Charmaz, 2006; Remenyi, 2014). Once a researcher identifies a set of culturally relevant Codes in the data, he or she then needs to specify how to identify those Codes in the data systematically. Each Big-C Code in the Big-D Discourse needs a small-c code that describes what in the small-d discourse counts as evidence for the Big-C Code. This is, of course, just a more technical way of saying that qualitative researchers develop codebooks that define their Codes and show how to apply those Codes to their data.

Critically, though, while understanding a Discourse requires making sense of the Codes, it is not enough to just identify Codes in the data. Qualitative researchers create Thick Descriptions by understanding how Codes are systematically related to one another. The meaning of the MUNSELL COLOR CHART for Goodwin’s archaeologists is tied up in the fact that it is used to analyze DIRT, and that the reason it is used to analyze DIRT is that differences in DIRT indicate things like EVIDENCE OF HUMAN SETTLEMENT. Codes are thus symbols, in the sense that, as Deacon (1998) explains, “the pairing between a symbol (like a word) and some object or event is… some complex function of the relationship that the symbol has to other symbols” (p. 83). Cultures—what Geertz (1973a) calls “organized systems of significant symbols”—are those composed of symbols that interact to form a web of meanings. Thus, researchers can understand a culture by analyzing how Codes in a Discourse are systematically related to one another.

### An Example of Quantitative Ethnography

Quardokus Fisher et al. (2016) studied Industrially-Situated Virtual Laboratory (ISVL) projects. In these projects, students work in teams with a simulation of a manufacturing process, and each team is guided by a coach: a more experienced engineer whose job is to help “enculturate students to the expectations of industrial project work.” ISVL projects let students solve authentic engineering tasks through an iterative process of experimentation, analysis, and reflection. The researchers were interested in how coaches guide students’ work and the effects of different coaching styles on deep learning. Just as Goodwin was interested in understanding how, in the culture of archaeologists, the Munsell Color Charts, relate to evidence of human settlement, and the color, texture, and consistency of dirt, so Quardokus Fisher and her colleagues wanted to understand how coaches used particular mentoring techniques, such as questioning and directive dialogue, to guide students to think about the relationships among important engineering content, such as the rates of kinetic reactions, experimental design, and the choice of input parameters to an experiment.

One approach to this problem would have been to try to identify these relationships directly: to create a Code for each of the different connections in the cultural web of meanings. So, for example, the researchers could have defined Codes for DIRECTIVE DIALOG TO ELICIT INFORMATION ABOUT RATES OF KINETIC REACTIONS, or QUESTIONS ABOUT THE RELATIONSHIP BETWEEN EXPERIMENTAL DESIGN AND INPUT PARAMETERS. However, it is often easier to identify evidence for Codes when they are broken into smaller pieces, such as a Code for GUIDING, another for KINETICS, another for EXPERIMENTAL DESIGN, and so on. This is because it is typically more efficient to develop primary codes (simple codes that are easy to identify directly in the data) and then use computation tools to model their connections.

Quardokus Fisher and her colleagues used epistemic network analysis (ENA) to model the connections between Codes in their data. ENA is a network analysis technique designed to model qualitative data by looking at how Codes are systematically related to one another in discourse (see, e.g., Shaffer & Ruis, 2017). Without going into any of the underlying mathematics of ENA here, an ENA model shows the Codes of a Discourse as nodes in a network graph. The pattern of association between Codes in the data is
represented by the lines connecting the Codes, where the strength of the connection between any two Codes is indicated by the thickness of the line connecting them.

The researchers analyzed 27 coaching sessions, 14 by one coach and 13 by another, and used that data to look for similarities and differences between the two coaches’ approaches to mentoring engineering students. The resulting network graphs looked something like Figure 1.

![Figure 1: Models showing the discourse patterns of two different coaches in an Industrially-Situated Virtual Laboratory project, adapted from Quardokus Fisher et al. (2016).](image)

The network graphs show the mean strength of connection between Codes for the coaching sessions of the two different coaches (the left network and right network in Figure 1). The model confirmed and extended what the researchers saw in their data. Both the Left and Right Coach focused their guidance heavily on understanding the rate of kinetic reactions, which was central to the experiments students were conducting. Both guided students to think about experimental design and input parameters in an experimental setting. But the Left Coach integrated these topics: The left network shows that averaged across coaching sessions for 14 different teams, the Left Coach facilitated connections between INPUT PARAMETERS, EXPERIMENTAL DESIGN, and KINETICS, as indicated by the more robust network of connections in the left network. Facilitation from the Right Coach, in contrast, was less integrated, and specifically did not connect INPUT VARIABLES to either EXPERIMENTAL DESIGN or KINETICS. The researchers explain that the Left Coach systematically “preferred to use ‘Input Parameters’ as an access point for discussion of the project,” which the Right Coach did not.

However, ENA models such as these do more than just visualize patterns in discourse. The two network graphs both represent the mean discourse network across multiple coaching sessions. That is, they show the means of two samples of discourse: one from the hypothetical population of all coaching sessions by the Left Coach, and one from the hypothetical population of all coaching sessions by the Right Coach. The researchers were thus able to use inferential statistics to warrant that the difference between the two samples was statistically significant.

Notice, though, that this statistical claim was not about coaches in general. Quardokus Fisher and her colleagues did not show anything about the mentoring of all coaches or even some particular type of coaches. Their claim was about these two particular coaches. A statistical analysis showed that the
researchers had enough data to conclude that the Left Coach and the Right Coach had different mentoring styles. In other words, the researchers were able to provide a statistical warrant for theoretical saturation of their analysis. Critically, however, they also developed a model that could be used to investigate the mentoring of coaches more widely in the future. Although this model involves only two coaches and a small number of observations about each of them, the process with which this model was created enabled the researchers to make an inference about key differences in approaches to coaching—and the same process can be used to test that inference with a much larger population of coaches without the need to examine all the data by hand.

Yet those larger investigations are predicated on statistical warrants for theoretical saturation, which in turn require codes that are both valid and reliable. That is, codes need to accurately and consistently reflect the Codes of interest in the Discourse. While it might seem obvious that coding should accurately reflect something that is meaningful in the discourse, many Big Data approaches to automated classification elide this point. For example, in topic modeling, an area of research in which Art Graesser has made considerable contributions, algorithms identify clusters of related words in a dataset, and then researchers identify what each collection—each topic—means (see, e.g., Moldovan, Rus, & Graesser, 2011). But a topic model does not “understand” the text, so the topics need to be validated by comparing actual data that is coded by human raters and the topics identified by the machine.

The extent to which codes accurately reflect Codes is typically measured using some interrater reliability (IRR) statistic, which provides a quantitative representation of the level of agreement between coding processes (human or machine). In a typical IRR analysis, two processes code a subset of the data available. If the IRR statistic between the two is over some established threshold, the coding processes are reliable: it consistently identifies Codes in the data. If both coders were human the process is also valid: the interpretation of the Code is sound. If the IRR statistic is below the threshold, then the coding processes are altered—coders discuss the differences and models (mental or computational) are changed—and the process repeats. Stated this way, though, IRR generalizes from a sample (the subset of data coded by both processes) to a population (all of the data). Thus, researchers need to control for potential Type I errors; however, work by Eagan et al. (2017) shows conclusively that this rarely happens.

For this reason, my colleagues and I have developed a statistic, Shaffer’s rho, for estimating the Type I error rate of interrater reliability statistics (Shaffer, 2017). If a measure of interrater reliability is statistically significant, researchers have a justification that their coding processes is theoretically saturated: although they only tested whether two raters agree on codes for some sample, they can say (with some limit to their error rate) that the raters would agree throughout the data. Only then can researchers confidently use automated classifiers to reliably identify meaningful Codes in large datasets.

Using Big Data for Thick Description thus depends on two forms of quantitative support for qualitative analyses:

1. Measures (such as rho) of Type I error in interrater reliability provide justifications that Codes are theoretically saturated; and

2. Inferential statistics applied to models (such as ENA) of how codes are linked to one another in discourse provide justifications for claims about how Codes are systematically related to one another in Discourse.

As Figure 2 shows, in quantitative ethnography, quantitative analyses thus support Thick Description by providing statistical warrants for claims of theoretical saturation.
Lessons for the Assessment of Deep Learning

The point of this brief example is, of course, not to suggest that one particular study is a model for all research on deep learning, nor that one specific set of analytical techniques are the best or only way for statistical tools to support qualitative analyses. And, of course, there are many conceptual and practical considerations that have been described here: things like identifying Codes, organizing and segmenting data, and so on. These considerations are described in much more detail in Shaffer (2017).

Rather, the purpose of this example is to suggest how and why the critical step in using Big Data for Thick Description is not using quantitative and qualitative analyses on the same data, but in putting both forms of analysis within the same conceptual framework. Statistical tools applied to Big Data support Thick Description when we reframe their role from generalizing to some population beyond the data at hand to generalizing within the data. When we reframe statistical analysis in this way, the logics of quantitative and qualitative inquiry become compatible, and we can connect these two different epistemological stances toward research.

That having been said, all formal analyses are simplifications of reality. Like any scientific instrument, the tools of quantitative ethnography make it possible to analyze one kind of phenomenon—in this case, Big Data about peoples’ actions and interactions in a range of social media and immersive learning environments—by combining the techniques of ethnographic and statistical analysis. Research on deep learning has much to gain from such an approach. What we potentially lose is the certainty that a person or group of people have read and interpreted all of the data. We rely on estimations of error instead hermeneutic scrutiny; we trade scale for precision.
But unlike purely statistical or computational techniques, when we analyze Big Data on deep learning with a quantitative ethnographic approach of the kind described here, the models are both grounded and based on a theory of deep learning. They are not just models of what people do, but how of how people make meaning. The models are not just patterns that a researcher happened to find in the data, but warrants that a close examination of discourse has told us something meaningful about a culture of learning.

Of course, more typical approaches to data mining and natural language processing can also model patterns in discourse, and even in some cases predict which students are likely to do well in a class and which are not (see, e.g., Gray, McGuinness, Owende, & Hofmann, 2016). But predicting that a student is likely to do poorly in a class is not particularly useful unless a model can explain why. For example, a data mining approach to measuring whether someone is a good chess player would almost certainly be able to identify expert and novice players: advanced players move their pieces faster than beginners. But one of the worst pieces of advice you could give a beginning chess player is: Just move faster. This is just another way of saying that predictive data mining models are saturated but not grounded.

Sometimes all we need is to make a prediction without really understanding why the prediction is right. And sometimes it is helpful to use data mining to find patterns in data in an exploratory way. But as researchers studying deep learning, one thing we absolutely need to be able to do is construct grounded models at scale. To pretend otherwise may be mathematically rigorous, but in the end is conceptually empty. Instead, I argue, like Geertz (1973b), that those who study deep learning should be skeptical of “claims that structural linguistics, computer engineering, or some other advanced form of thought is going to enable us to understand men [and women] without knowing them” (p. 30).

References


Acknowledgments

This work was funded in part by the National Science Foundation (DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DRL-1661036, DRL-1713110, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000), the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.