Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination

Michael Muller[a], Shion Guha[b], Eric P.S. Baumer[c], David Mimno[c], and N. Sadat Shami[d]
[a]IBM Research, [b]Marquette University, [c]Cornell University, [d]IBM
[a]One Rogers Street
Cambridge MA USA 02142
{michael_muller, sadat}@us.ibm.com, shion.guha@marquette.edu,
{ericpsb,dm655}@cornell.edu

ABSTRACT
Grounded Theory Method (GTM) and Machine Learning (ML) are often considered to be quite different. In this note, we explore unexpected convergences between these methods. We propose new research directions that can further clarify the relationships between these methods, and that can use those relationships to strengthen our ability to describe our phenomena and develop stronger hybrid theories.

Keywords
Grounded theory; Axial coding; Coding families; Machine learning; Supervised learning; Unsupervised learning.

1. INTRODUCTION
Researchers in CHI and (E)CSCW often think of qualitative analysis and machine learning as very different domains (e.g., [1, 9, 10, 21]), even though some researchers may use one of these methods as a tool to understand or summarize work done via the other method (e.g., [2, 3, 33, 58]). However, there are multiple convergences between these two research traditions (e.g., [6, 45]). These convergences can lead to interesting hypotheses and potential methods (e.g., [6]). In this note, we continue a conversation begun by [5] about two specific convergences, which we hope will open a discussion of additional convergences and new ways of conducting mixed-methods research.

This note proceeds as follows. In the Background section, we provide brief introductions to machine learning (ML) and grounded theory method (GTM), with a focus on setting up our points of convergence. In the Convergences section, we explicate similarities between these two distinct research traditions. We also consider some differences between the two approaches (Divergences). In the Discussion section, we propose three ways to transform the convergences into new research hypotheses and mixed-methods inquiries.

2. Background
Researchers are increasingly interested in combining quantitative “big data” methods with qualitative “small data” methods [10, 23]. Traditional qualitative scholarship often focuses on in-depth analysis of specific, local phenomena, with the intention of generalizing to other sites and other people. Traditional quantitative scholarship often focuses on generalizable results from more numerous data points, which tend to have less context associated with each observation.

GROUP '16, November 13-16, 2016, Sanibel Island, FL, USA
© 2016 ACM. ISBN 978-1-4503-4276-6/16/11…$15.00
DOI: http://dx.doi.org/10.1145/2957276.2957280

2.1 Combining Big Data and Small Data
In one example, in an influential series of studies, De Choudhury and colleagues studied linguistic signs of affect [13], and more specifically depression [14], in social media posts – a characteristic quantitative, “big data” study method. Their goal was not to provide detailed analyses of postings or posters, but rather to understand large-scale phenomena. In a second example, Menking and Erickson used grounded theory method to understand aspects of women’s experiences in with participation in Wikipedia [40]. Their goal was not to provide a statistical summary of such phenomena, or a large-scale analysis of quantitative patterns among participation data, but rather to develop a theory that described the experiences of individual women. These are two out of many examples that seem to argue that the quantitative domain works differently from the qualitative domain. But is that true? What if we could enjoy the virtues of both ways of inquiring?

Computational social science and digital humanities are beginning to use large-scale computational analysis in conjunction with close analytic reading. Blevins used topic modeling as a way to structure a detailed inquiry into the extensive and idiosyncratic diaries of Martha Ballard, an 18th-19th-century midwife [8]. Jockers used a topic model to study themes in 3000 novels, but only after a laborious, iterative process of document and vocabulary curation [34]. Rhody used similar topic-based methods to identify figurative language in poetic texts [49]. Only with contextual human analysis were the identified word-clusters recognized as the surface form of a repeated metaphor

Within the CHI and CSCW disciplines, Thom-Santelli et al. used “heuristics from grounded theory” in alternation with large-scale data analysis, for iterative guidance on a series of interview samples [56]. In a very recent demo, Arazy et al. similarly proposed using machine-learning to reveal patterns in log data from peer production; they suggested that those patterns could then be studied using grounded theory method [2].

Here, we suggest a deeper convergence. In the next two subsections, we describe two analytic trends in machine learning, and we show that they are closely related to two well-established approaches in grounded theory method.

2.2 Machine Learning: Current Approaches
Machine learning centers around statistical models and algorithms. The models consist of two types of variables: those that are observed in data (*features*) and those that are not observed. The algorithms find values for unobserved variables

that best fit the values of the observed variables. For instance, given a corpus of emails, for each message we can observe a set of words and a label of spam or not-spam. We can then use an algorithm to infer *latent* variables that correspond to the "spamminess" of each word. Most ML algorithms function by iteratively refining a model until it fits the data within some threshold of acceptability.

2.2.1 Types of Machine Learning Methods

We often divide machine learning into two categories: supervised and unsupervised learning [37]. In supervised models, the latent variables define a relationship between two types of observed variables, inputs and outputs. Regression, classification, and confirmatory factor analysis are common examples. In unsupervised models, the latent variables suggest a (hopefully) simpler description of the observed variables. Clustering, principal components analysis, topic modeling, and exploratory factor analysis are common examples.

There is a strong bias towards standard, pre-existing data that is accompanied by well-known quantitative evaluation metrics. This approach obviates the difficult human labor involved in creating such data sets [34]. When new labeled data sets are collected, they are increasingly built using crowdsourced annotation platforms such as Amazon Mechanical Turk [51]. These crowdsourced, human annotated datasets are often referred to as gold standards and have become widely used in HCI/CSCW [13, 50].

There is growing recognition of the important role to be played in ML by in-depth human analysis and interpretation. Wagstaff argues that machine learning research has become dangerously disconnected from the real world, and recommends that domain experts be involved [59]. Some scholars [60] have attempted to answer this call by incorporating focus groups and feedback from doctors to build a system that predicts fetal hypoxia and makes judicious medical interventions. Sen et al. [50] criticize the focus on labeled datasets, saying that the notion of a one -size-fits-all gold standard is risky and inconsistent with existing social science theory. They show that labeling with different cultural groups on the same dataset often creates markedly different labeled datasets – a concept already understood in social sciences [39].

2.3 Grounded Theory Method

Grounded theory method provides a set of rigorous methods for constructing theory directly from data, in a way that "grounds" that theory in data. The founders of grounded theory, Barney Glaser and Anselm Strauss, published four major works [27, 28, 29, 30] before they fell into disagreement and developed major bodies of work and method that contrasted with one another (e.g., [17, 26, 54, 55], and especially [24]). This history of their disagreement has been well-described in the research literature [11, 35, 43], and is beyond the scope of this note. Here, we focus on the potential relationship of each respective author's work to trends in machine learning.

One of several crucial aspects of the intellectual discipline of grounded theory is a rigorous way of coding data [15, 48, 52]. Over several revisions of their account of methods, Corbin and Strauss provided a series of data-driven operations to develop a high-quality description of a domain [17, 54, 55]; grounded theory researchers sometimes refer to these volumes as "the cookbook." Charmaz integrated aspects of the Straussian and Glaserian approaches into a disciplined set of practices that she refers to as "Grounded Theory Method" (GTM) [15]. There is general agreement [11, 35, 43, 44, 52] on the following overview of the method: The researcher begins with *open coding*, in which each description is specific to particular points in the data. Over time, the researcher formalizes and organizes these codes through *axial coding*, which begins to establish relationships among the open codes. Through the use of *constant comparison* of data with data, and of data with emerging theory, the researcher constructs *dimension*s, which are collections of open and axial codes that, collectively, describe major differences across the range of the data. The emergent theory (codes, categories, dimensions) suggests new questions, and the researcher uses that theory to select new sites to study, or new people to interview, to test the emergent theory at its weakest points. This process is generally termed *theoretical sampling*, and is a procedural manifestation of the underlying concept of *abductive logic* [15, 48, 57].

As the codes become more abstract, the researcher can begin to relate the data-grounded theory (i.e., based on the codes) with formal, published theory [15, 17, 28, 55]. The researcher chooses certain of these axial codes (categories and dimensions) as the focus of her/his report, and engages in further *selective coding* around these chosen attributes, deepening the analysis further [15, 17, 24-29, 35, 48, 52, 54, 55].

GTM thus provides a set of agreed core practices (three levels of coding, from open to axial to selective) and core *ways of thinking* about data (constant comparison, theoretical sampling) which are derived from philosophical bases in pragmatism and abduction [11, 15, 16, 43, 44, 48, 54, 55, 57].[1]

2.3.1 Types of Grounded Theory Method Approaches

Among the points of disagreement between Glaser and Strauss (e.g., [24]) was the method and formality of coding – i.e., the influence (or constraint) of prior work on developing new theory from new data [35]. While Strauss made a series of differing proposals about core concepts (i.e., codes) in grounded theory (e.g., [54, 55]), many grounded theory researchers who derive their work from Strauss, tend to ignore these divergent advices, focusing instead on the core notion of the gradual development of open and axial codes into categories and dimensions of analysis, as summarized above (e.g., [15]; see also multiple chapters in [11]). Thus, one of Strauss's (later, Corbin's and Strauss's) contributions to GTM is a rigorous yet relatively unencumbered approach to coding. Through a rigorous process, codes emerge [17, 55] or are constructed [15] de novo, without constraints based on previous coding schemes or reference volumes.

By contrast, Glaser developed a rigorous set of coding families [26]. These families grew incrementally over time, from an initial set of 18 coding families [26] to 40 coding families [16, 25]. Thus, one of Glaser's contributions to grounded theory is a rich set of tested coding families that can serve as a bridge between initial open codes (nascent theory) and more formal theory, and between past and present in GTM.

3. Convergences

Despite the different world-views, there are attributes in common between ML and GTM. Each process makes major claims to be grounded in the data. Each process begins with, and returns to, the data. Each process develops interim components of theory (GTM dimensions or ML features) that describe major differences across

[1] For comparative purposes, we note that a conventional statistical analysis may at first appear to be a simple, executable method, but in fact involves many decisions by the researcher regarding sample size and provenance, data quality issues and possible transformations of variables, selection of appropriate control variables and conditions, moderating factors, and covariates.

the range of the data. Each process is iterative. Neither process is complete when the data are analyzed; rather each process requires interpretation and theory-building after the data are complete. This is to say, while each process is strongly guided by methods, practices, and protocols, each process crucially requires human interpretation and human judgment at each stop of theory-building. We describe two similarities of method in depth here.

3.1 Pattern Discovery / Unsupervised ML / Straussian GTM

Straussian GTM is often recommended as a disciplined way of entering a new field of study, unconstrained by prior theory [11, 15, 17, 28, 44, 52, 54, 55]. Through a series of *constant comparisons* of data with data, and of data with theory, the researcher develops an initial theory, iteratively tests it with new data, and gradually develops a stronger and more interesting theory with better fit to the data. To do this, the researcher must make a series of decisions about which categories or dimensions of the data are important. S/he can then collect more data to confirm or disconfirm those candidate dimensions, and to establish new hypotheses for further qualitative inquiries (new samples or sites for the next iteration of GTM analysis) [15, 17, 28, 43, 44, 54, 55].

Unsupervised learning approaches in ML, such as clustering and topic-modeling, have been characterized in similar ways [22, 32]. Through a series of data-driven clustering and grouping explorations, the researcher develops an initial sense of the possible underlying patterns in the data, iteratively tests clustering parameters, and gradually finds a stronger and more interesting set of parameters with better fit to the data. To do this, the researcher must make a series of decisions about which parameters are important. S/he can then refine those parameters and experiment with higher-order interactions that involve the subset of parameters, thus confirming or disconfirming those candidate parameters.

While these two descriptions are not perfectly matched, they exhibit strong similarities. Both methods are exploratory; both methods are grounded in the data; both methods involve iterative and testable intermediate versions, and both methods require human discernment to refine their models. The researcher's problem is, in part, to find the best data that correspond to the hypothesized structure of the current knowledge.

3.2 Prediction-Validation / Supervised ML / Glaserian GT

A second area of similarity occurs in Glaserian GTM. Glaserian GTM explores a new domain, but with guidance from one or more of Glaser's coding families [16, 25, 26]. Part of the researcher's task is to choose the coding family that has the strongest fit to the data, and the best relationship to formal (preceding) theory. Different coding families may be confirmed or disconfirmed by collecting more data, in new samples or at new sites chosen to make a strong test of the current coding family(ies).

Supervised learning in ML (e.g., regression, decision trees, random forests, and many other predictive methods) operates in somewhat similar ways. In supervised learning, there is a concept of *ground truth* or *gold standard*, in which the nature of the categories in the data is known a priori, and for which the researcher tries to find the best predictors [38, 61]. Different predictive parameters may be tested, en masse or individually, and may be confirmed or disconfirmed through further analysis.

We see similarities between GTM coding families and ML ground truth measures: Both are known a priori; both serve as a kind of bridge between established knowledge and new data; both require human discernment to refine their models. The researcher's problem is, in part, to find the best data that correspond to the hypothesized structure of the current data, based on preceding knowledge.

4. Evaluating the Quality of Results

ML and GTM differ in their approaches to evaluating the quality of results.

4.1 Quality and Fit

A frequent concern in ML is to avoid *overfitting*, which is a model that performs well on training data, but fails to generalize to new data. ML establishes quality using statistical comparisons between a model's predictions and actual data. For example, cross-validation tests the robustness of a solution within a data sample [36]. A computed model may also be applied to a new set of data or to held-out data, or different models compared on the same set of data [19]. In each case, statistical tests for the degree of fit determine quality.

In the case of supervised learning, it is also common to test the accuracy of the computed model from one dataset, through a comparison of correctly vs. incorrectly classified data points in a second dataset (i.e., in a confusion matrix) [53]. Thus, the robustness of a ML analysis *depends crucially on a search for unexpected* data (i.e., disconfirmation of the current version of the model), and is *tested by recourse to more data*.

Formal GTM follows a process that iteratively tests both internal coherence of the emergent theory, and the ability of the theory to describe new data. Some GTM researchers make reference to the logic of *abduction* [15, 48, 57]. While the details are beyond the scope of a note, the core steps of abductive logic include: collecting a small amount of data; constructing an initial theory based on those data; determining the weakest point of the theory; and then collecting a second small amount of data designed to break the theory – i.e., to test it at its weakest point; and iterating this process until there is no new learning (the theory is stable)[2]. This abductive process in GTM is broadly described as *theoretical sampling*, because each iteration of the emerging theory is used to select additional data to test it [15, 48, 52, 57]. The comparisons of data with data, and data with theory, are consistent with the principle of constant comparison mentioned earlier. Similarly to issue of overfitting in ML, the robustness of an emergent theory in grounded theory method *depends crucially on a search for unexpected data* (which signal a need to continue developing the emergent theory), and is *tested by recourse to more data* (see [34, 44] for further details on the necessity of unexpected data to guide theoretical sampling in GTM).

Essentially, the difference is analogous to performance vs. description. ML assesses quality by testing how well a learned model performs on new data. GTM assess quality by asking how well a theory describes data from a novel context.

4.2 Iteration and Human Evaluation

Both ML and GTM follow iterative processes. In GTM, human researcher(s) iteratively analyze data while revising their theories

[2] Some ML algorithms, such as perceptron and SVM, focus on those data points where the model fits poorest. These points are tested in a prescribed iterative process, though, rather than strategically selected by the researcher.

about those data. In ML, an algorithm iteratively refines a mathematical model for the data until the model converges and satisfactorily explains the data. Both these iterative processes also occur at multiple, nested levels, drawing attention to two differences.

First, at the lowest level, ML fully automates the iteration. Algorithmic iteration takes place while a model is fitted to data, and requires no human intervention. In grounded theory, iteration through the data is done by the human researchers.

Second, at the higher level, human judgments play a role in both processes. In GTM, following principles of theoretical sampling [15, 52], a human researcher may intentionally seek out data that challenges an existing theory. In ML, a human researcher may use preliminary results to tune aspects of the analysis, such as a stopword list, number of categories, number of iterations, optimization calculation, hyperparameter values, etc. [34].

However, the role of human judgments is reported differently in research results. The human level of iteration is integral to GTM, and is both described in canonical texts and often reported in research results. By contrast, in ML, this higher level creates a dialogic iteration between human researcher and computational algorithm, but it is rarely described as such. Instead, ML is often described as a linear process, from data collection and curation through feature extraction to classifier training and testing, rather than an iterative process of trial and error. This difference draws attention to divergent disciplinary norms, divergences to which future work will need to attend.

5. Research Directions

In related work, Baumer et al. [6] proposed a *Conjecture* about the relationship between GTM and the specific ML technique of topic modeling. This conjecture asserts that any theoretical concept identified via grounded theory or similar analyses should occur in part via linguistic regularities that can be detected by algorithms. Put differently, the two different methods should be capable of generating comparable results. Baumer et al. suggest that future research should investigate the conditions under which this conjecture holds.

In this section, we propose several more integrated combinations of methods, in which conclusions would be based on the weaving-together of methods and outcomes. We consider three "configurations of methods." Each of these configurations is from unpublished research. We provide a sketch of how the analyses might proceed. We hope that other researchers will propose more diverse configurations.

5.1 Comparative Sequential Methods

Our first two configurations propose to use the methods sequentially, with one method structuring a series of smaller research projects using the other method. There are two variations in this approach, differing in terms of primacy.

GTM→ML. In an analysis of group discourse, we might discover different types of contributions. Once we have texts grouped in this way, we could then apply ML methods in a quantitative analysis of the linguistic attributes of each type of contribution.

> **Example GTM→ML:** In a study of social media use within an organizations' Intranet, we have found 30 distinct types of postings. GTM could help us to combine these types into higher-order conceptual classifications, such as texts that *originate* a discussion; vs. texts that *respond* to a previously-originated discussion; vs. texts that *curate* or *evaluate* the contributions to a discussion. We could then apply text analysis and ML to understand the large-scale linguistic phenomena within each type of contribution, such as implicit sentiment [31], linguistic style matching [18], or framing [6, 12].

ML→GTM. The inverse of this strategy could also be adopted. We could begin with a ML classification against a ground-truth outcome measure. Once this formal classification was complete, we could then apply GTM to search for emergent properties within each of the ML classes.

> **Example ML→GTM:** Mitra and Gilbert used ML to classify and predict *successful* vs. *unsuccessful* Kickstarter proposals [42]. In an extension of this work, we could apply GTM to study how the words and tuples that they identified combined conceptually into persuasive strategies and implicit appeals that were made in each class of proposals.

5.2 Hybrid Iterative Methods

Our third configuration puts the two sets of methods in a more integrated, iterative process. Following the examples of [8, 49], and especially [2, 56], we propose an expansion of the GTM concept of theoretical sampling to put these two methods into constructive, iterative dialogue. An initial analysis could perform large-scale pattern analysis, similarly to Azary et al.'s analysis of Wikipedia texts [2]. The revealed patterns would then point to interesting cases, which could be studied via GTM, similarly to Thom-Santelli et al.'s approach to using macro-scale data patterns in order to select sites (persons) for open-ended interviews, which were then analyzed via GTM approaches [56].

We propose that researchers could alternate between developing each iterative step in a series of evolving GTM *theories* of a phenomenon, and each iterative step in a successively refined ML *model* of that phenomenon (i.e., one complete run of a topic-modeling algorithm). Each of these small iteration steps is a form of constant comparison, and would require comparison to each type of data. Each method would inform the other method, with a progressive strengthening of the core representation of each method (i.e., a *theory* composed of *dimensions* in GTM; a *model* composed of *features* in ML). Theoretical sampling would become more complex, and potentially more powerful, because each theory or model would suggest new data to sample in the domain of the other theory or model (i.e., from GTM theory to ML data, and from ML model to GTM data).

> **Example Iterative:** In a study of online forums for patients and informal caregivers, we might want to understand how people are evaluating the option of palliative care. One forum for "Palliative Care" contains 5758 postings organized into 177 human-specified topics, authored by 969 contributors.[3] Postings can be long and complex, and therefore resist classification approaches that place each data item into at most one class.
>
> We might use the more flexible ML method of topic modeling, which assumes that each posting may be drawn from one or more latent topics [8]. We could then select the subset of individuals associated with certain topics, and apply GTM to understand the themes and concerns of the patients and caregivers. So far, this step is an example of the **ML→GTM** configuration discussed above. We could then use the emergent dimensions from the GTM analysis (e.g.,

[3] See, e.g., https://community.breastcancer.org/forum/77. Posts were extracted on 25 November 2015.

complexly interleaved themes of fear and hope) to reconfigure the topic-modeling analysis, and then derive a second iteration of topics. This step looks more like the **GTM→ML** configuration discussed above. The revised topic model would then suggest new sets of postings to be analyzed via GTM.

Overall, this process sketch applies the GTM principle of *constant comparison* of data with theory (in ML and in GTM), and of data with data (in the interplay of GTM and ML). The process also applies the GTM principle of *theoretical sampling*, in which the current state of the analysis (i.e., the emergent theory) is used to select the next site or dataset for the next iteration of the analysis.

6. Conclusion

In writing this paper, we wondered if the similarities between GTM and ML arose from a deeper resonance. We note that the Hybrid idea makes sense because GTM and ML pursue similar core research program ("strategies," as it were), even though they use different methods ("tactics," as it were). It may be that these two methods – and, probably, other methods in HCI and CSCW – reflect common goals of inquiry in knowing and describing our phenomena (e.g., [47]). While clearly beyond the scope of the current paper, we hope other researchers will explore to what extent similarities (or divergences) among different methods may contribute to a deeper convergence of research methods.

7. REFERENCES

[1] Chee Siang Ang, Ania Bobrowicz, Diane J. Schiano, and Bonnie Nardi (2013). Data in the wild: Some reflections. *Interactions 20*(2), 39-43.

[2] Ofer Arazy, Henry Brausen, David Turner, Adam Balila, Eleni Stroulia, and Joel Lanir (2015). coDNA: Visualizing peer production processes. *CSCW 2015 Companion*, 5-8.

[3] Ohad Barzilay, Orit Hazzan, and Amiram Yehudai (2009). Characterizing example embedding as a software activity. *Proc. SUITE 2009*, 5-8.

[4] Eric P.S. Baumer, Elisha Elovic, Ying "Crystal" Qin, Francesca Polletta, and Geri K. Gay (2015). Testing and comparing computational approaches for identifying the language of framing in political news. *Proc NAACL 2015*, 1472–1482.

[5] Eric P. S. Baumer, Shion Guha, Emily Quan, David Mimno, and Geri K. Gay (2015). How social media non-use influences the likelihood of reversion: Perceived addiction, boundary negotiation, subjective mood, and social connections. *Social Media + Society 1*(2),.

[6] Eric P.S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri Gay (2016). Comparing grounded theory and topic-modeling: Extreme divergence or unlikely convergence? *JASIST*, in revision.

[7] Christopher M Bishop (2006). *Pattern Recognition and Machine Learning*, Springer, NY, NY, USA.

[8] Cameron Blevins (2010). Topic modeling Martha Ballard's diary. *Cameron Blevins* (blog), 1 April 2010, http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/ .

[9] Jeanette Blomberg and Helena Karasti (2013). Reflections on 25 years of ethnography in CSCW. *CSCW 22*(4-6), 373-423.

[10] danah boyd and Kate Crawford (2012). Critical questions for big data. *Info. Comm. & Soc. 15*(5), 662-679.

[11] Anthony Bryant and Kathy Charmaz (2007). *The Sage handbook of grounded theory*. London, UK, Sage.

[12] Dallas Card,Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith (2015). The media frames corpus: Annotations of frames across issues. *Proc ACL 2015*, 438-444.

[13] Munmun De Choudhury and Scott Counts (2013). Understanding affect in the workplace via social media. *Proc. CSCW 2013*, 303-316.

[14] Munmum De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff (2014). Characterizing and predicting postpartum depression from shared facebook data. *Proc. CSCW 2014*, 626-638

[15] Kathy Charmaz (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. London, UK, Sage.

[16] Ólavur Christiansen (2008). The rationale for the use of classic GT. *Grounded Theory Review 7*(2), http://groundedtheoryreview.com/2008/06/30/1046/

[17] Juliet Corbin and Anselm L. Strauss (2007). *Basics of quail-tative research: Techniques and procedures for developing grounded theory. 3rd edition*. Newbury Park, CA, USA: Sage.

[18] Cristian Danescu-Niculescu-Mizil, Lilian Lee,, Bo Pang, and Jon Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. *Proc WWW 2012*, 699-708.

[19] Janez Demšar. (2006). "Statistical comparisons of classifiers over multiple data sets." *The Journal of Machine Learning Research* 7,1-30.

[20] James Dougherty, Ron Kohavi, and Mehran Sahami (1995). Supervised and unsupervised discretization of continuous features. *Proc. 12th Int. Conf. Machine Learning*, 194-202.

[21] Paul Dourish (2014). Reading and interpretation ethno-graphy. In Judith S. Olson and Wendy A. Kellogg (eds.), *Ways of knowing in HCI*. New York, NY, USA, Springer.

[22] Jennifer G. Dy and Carla E. Brodley. (2004). Feature selection for unsupervised learning. *J. Mach. Learn. Res.* 5 (December 2004), 845-889.

[23] Danyel Fisher, Rob DeLine, Mary Czerwinski, and Steven Drucker (2012). Interactions with big data analytics. *Interactions 19*(3), 50-59.

[24] Barney G. Glaser (1992). *Emergences vs. forcing: Basics of grounded theory analysis*. Mill Valley, CA, USA: Sociology Press.

[25] Barney G. Glaser (2005). *The grounded theory perspective III: Theoretical coding*. Mill Valley, CA, USA: Sociology Press.

[26] Barney G. Glaser (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, CA, USA: Sociology Press.

[27] Barney G. Glaser and Anselm L. Strauss (1965a). *Awareness of dying*. Chicago, Aldine.

[28] Barney G. Glaser and Anselm L. Strauss (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, Aldine.

[29] Barney G. Glaser and Anselm L. Strauss (1965b). Discovery of substantive theory: A basic strategy for qualitative analysis. *Am. Beh. Sci. 8*, 5-12.

[30] Barney G. Glaser and Anselm L. Strauss (1968). *Time for dying*. Chicago, Aldine.

[31] Stephan Greene and Philip Resnik (2009). More than words: Syntactic packaging and implicit sentiment. *Proc HLT 2009*, 503–511.

[32] Isabelle Guyon and André Elisseeff. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3 (March 2003), 1157-1182

[33] Rashina Hoda (2011). *Self-organizing agile teams: A grounded theory*. PhD thesis, Victoria University of Wellington, 2011.

[34] Matthew Jockers. 2013. *Macroanalysis*. University of Illinois Press: Champaign, IL.

[35] Udo Kelle (2005). "Emergence" vs. "forcing" of empirical data? A crucial problem of "grounded theory" reconsidered. *FQS 6*(2), art. 27.

[36] Ron Kohavi. (1995) "A study of cross-validation and bootstrap for accuracy estimation and model selection." *IJCAI 14*(2).

[37] Ron Kohavi, (1998). Glossary of terms**,** "Special issue on applications of machine learning and the knowledge discovery process." http://robotics.stanford.edu/~ronnyk/glossary.htm

[38] Scott Krig (2014). *Computer vision metrics: Survey, taxonomy, and analysis*. Apress, Chapter 7.

[39] Karen S. Kurasaki (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field methods*, *12*(3), 179-194.

[40] Amanda Menking and Ingrid Erickson (2015). The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. *Proc. CHI 2015*, 207-210.

[41] Tom Mitchell, 2006, "The discipline of machine learning". URL: http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf

[42] Tanushree Mitra and Eric Gilbert (2014). The language that gets people to give: Phrases that predict success on Kickstarter. *Proc. CSCW 2014*, 49-61.

[43] Janice M. Morse, Phyllis Noerager Stern, Juliet Corbin, Barbara Bowers, Adele E. Clarke, and Kathy Charmaz (2009). *Developing grounded theory: The second generation*. Walnut Creek, CA, USA, Left Coast Press.

[44] Michael Muller (2014). Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In Judith S. Olson and Wendy A. Kellogg (eds.), *Ways of knowing in HCI*. New York, NY, USA, Springer.

[45] Michael Muller, Shion Guha, Matthew Davis, Werner Geyer, and Sadat Shami (2015). Developing data-driven theories via grounded theory method and machine learning. Presentation at Human Computer Interaction Consortium, Watsonville, CA, USA, June 2015. Slides available at http://www.slideshare.net/traincroft/hcic-muller-guha-davis-geyer-shami-2015-0629

[46] Wanda J. Orlikowski and Jack J. Baroudi. 1991. Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research 2*(1), 1–28. http://doi.org/10.1287/isre.2.1.1

[47] Judith S. Olson and Wendy A. Kellogg (2014) (eds.), *Ways of knowing in HCI*. New York, NY, USA, Springer.

[48] Jo Reichertz (2007). Abduction: The logic of discovery of grounded theory. In Anthony Bryant and Kathy Charmaz (eds.), *The Sage handbook of grounded theory*. Thousand Oaks, CA, USA: Sage.

[49] Lisa Rhody. 2012. Topic modeling and figurative language. *Journal of Digital Humanities 2*, 1.

[50] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao "Ken" Wang, and Brent Hecht. (2015). Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. *Proc. CSCW,* 826-838.

[51] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. *Proc. Conf. Empir. Meth. Nat. Lang. Proc.,* 254-263.

[52] Susan Leigh Star (2007). Living grounded theory. In Antho-ny Bryant and Kathy Charmaz (eds.), *The Sage handbook of grounded theory*. Thousand Oaks, CA, USA: Sage.

[53] Stephen V Stehman. (1997). "Selecting and interpreting measures of thematic classification accuracy." *Remote sensing of Environment* 62.1,77-89.

[54] Anselm L. Strauss (1987). *Qualitative analysis for social scientists*. Cambridge, UK: Cambridge University Press.

[55] Anselm L. Strauss and Juliet Corbin (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA, USA: Sage.

[56] Jennifer Thom-Santelli, Michael Muller, and David Millen (2008). Social tagging roles: Publishers, evangelists, leaders. *Proc CHI 2008*, 1041-1044.

[57] Stefan Timmermans and Iddo Tavory (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Socio. Theor. 30*(3), 167-186.

[58] Christopher Vendome, Mario Linares-Vásquez, Gabriele Bavota, Massimiliano Di Penta, Daniel German, and Denys Poshyvanyk (2015). License usage and changes: A large-scale study of Java projects on GitHub. *Proc. IEEE Int. Conf. Prog. Comp. 2015*, 218-228.

[59] Kiri L. Wagstaff, K. (2012). Machine learning that matters. arXiv preprint arXiv:1206.4656, or http://www.wkiri.com/ research/ papers/wagstaff-MLmatters-12.pdf

[60] Philip A. Warrick, Emily F. Hamilton, Robert E. Kearney, and Doina Precup (2012). A machine learning approach to the detection of fetal hypoxia during labor and delivery. *AI Magazine, 33*(2), 79-90.

[61] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA, 246-253.