The Human is the Data Science

Sebastian Benthall
School of Information
University of California Berkeley
sb@ischool.berkeley.edu

Abstract

The increasing size of qualitative data sets prompts an interest in bridging between data-driven research and qualitative methods. Human-centered data science can be both sensitive to context and objective if it conceives of traditionally qualitative methods in the same formal languages used to define data-driven methods. We consider how to operationalize the "richness" of qualitative methods in terms of computational statistics and outline the challenge of representing composite perspective as examples of human-centered data science.

Author Keywords

CSCW '16, February 27 - March 02, 2016, San Francisco, CA, USA.

Data science; ethnography; computational psychology.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

To compute is human

"Human-centered data science" is notionally the bridge between broad, shallow data-driven research and deep, narrow qualitative research [2]. The former depends on statistics and computation to derive generalizations or summaries of data. The latter is prototypically a combination of human activities: rich observation [19] and "thick" description [13].

The increasing size of qualitative data sets prompts the questions of what human-centered data science can be. The metaphorical terms– "broad", "shallow", "deep", "narrow", "thick"– used to frame these questions are terms of space and scale. This language signals a confrontation between qualitative research traditions and the fact and theory of data quantity.

The foundational theory of computer science is the mathematical definition of the algorithm and the tools of assessing the complexity in time and space of specific tasks as a function of the size of the input data [23]. Statistics provides a rigorous basis for generalizing from quantities of data by assigning measures of probability to well-defined hypotheses [12]. In so far as computers are employed to perform tasks in service of qualitative research, these tasks will necessarily be defined in terms of computation. As far as research methods can be rigorously shown to provide generalizable results, they will have to at least approximate statistics.

There is ample evidence that suggests qualitative methods do precisely this. Computational psychology has shown how human perception and higher-level cognition implicitly perform statistical operations [24]. The psychological paradigm of rational analysis [1] explains human cognition in terms of the computational tasks it has evolved to perform. The Bayesian modeling approach emphasizes the question of how humans are able to successfully generalize from their experience. It empirically shows the deep correspondence between statistics and thought. [15, 20, 26]

Qualitative researchers are, in effect, psychological instruments, observing, analyzing, and articulating data. As such, they are bound by the same mathematical limits of data processing and generalization that have been used to design computer algorithms. Fortunately for human-centered data science, this means that qualitative and quantitative methods can in principle be described in a shared formal language. This is good news for mixed-methods research because science depends on formulation of methods. [5]

How much richness?

As an example of how we can formally bridge between qualitative and computational methods, let's ask: Can we address the question of 'richness' of qualitative data in a quantitative way? Consider the following sketch of an answer.

Statistical data analysis often involves data sets with many instances and few features. Examples include a satellite image that assigns a few variables of color to every spatially distinct pixel, a time series data set that gives a stock price at every minute, or a census of many citizens and their demographics.

In contrast, the data collected in a qualitative study have many features and few instances. This is because the emphasis on rich data collection and thick writing encourages representations that track connections between phenomena as opposed to sampling in ways that provide statistical independence. An ethnographer's notes from a field site are a good example of this.

Today there are many data sets that have both many features and many instances. Large corpuses of documents, for example, are rich in features and also broad in number. There are also large data sets about connected objects, such as social network data. These may be treated quantitatively by making simplifying assumptions about their contents, for example by assuming that nodes are only affected by neighbors on the social graph. They may also be studied by a qualitative researcher. In general, with these large digital data sets, the data itself is neither strictly quantitative nor strictly qualitative. It is always both. If there is a distinction to be made between the qualitative and quantitative, it is in methods not data.

One way in which these methods differ is in the representation of the data in the research product. Computational methods have been described as "shallow" [2]. One thing this can mean is that computational models are frequently used to dramatically reduce the number of features in the final representation. For example, principle component analysis (PCA) takes a high-dimensional data set and transforms it into a lower-dimensional space. Linear regression with L1-regularization will fit a model with a lower number of relevant features than the original data set.

In contrast, qualitative methods will turn a rich data set like ethnographic field notes into "thick" writing that summarizes the data's rich interconnectivity, perhaps also with reference to themes from other thick literature. The expressive potential for qualitative methods is much wider than for traditional statistical measures. In the language of statistical model fitting, qualitative methods have high *variance.* Intuitively this means that they are very sensitive to the particulars of the data. This is often exactly what qualitative researchers want. The downside is that high variance methods have high statistical variance error and so are generally not good for providing generalizable results [14]. Many qualitative researchers recognize this limitation to their methods and claim that their results are not intended to be generalizable. A formal definition of qualitative work would serve to provide guidelines for how to present the data in an unbiased way, which is not an easy task [21]. It would also help prevent unintentional bias, an important part of ethical social scientific research [9].

Measuring context

Qualitative research is important not just for its richness, but also for its contextual sensitivity. This has historically been a missing component of many technical fields, including computational cognitive science, whose research subjects are typically isolated in lab conditions. This has led some qualitative researchers to argue that prioritizing technical language endangers the legitimacy of situated understanding [6,18] and risks being unethical with respect to values, such as privacy, that depend on context [22]. Human-centered data science is an opportunity to resolve these tensions.

The framework of distributed cognition is a way of reconciling the importance of context with a computational view of learning [17]. Meanwhile, cognitive scientists are recognizing how their earlier work was shaped by the isolated conditions of the psychology lab [16]. Many of the important questions raised by the 'human' element of human-centered data science, such as ethical questions around privacy and fairness, have already been addressed in a formal way [10, 11]. At the same time, the fact of data quantity challenges notions of context and demands objectivity that may be unfamiliar to some qualitative researchers. Whereas human-centered design aims to address situated users [4], data science often aims to provide programmatic tools that can be shared between contexts and used on “the cloud”. Formalization of qualitative methods in technical terms is necessary for integration with data science. Insensitivity to context is not. Rather, human-centered data science can aspire to be objectively sensitive to context in novel ways. An example of this is the problem of representing the combined perspectives of many interviewed research subjects.

Composite Perspective

In an ongoing research project the author is involved in, many research subjects have been interviewed about a historical event, the failure of a legislative bill that was widely thought to be likely to pass. The subjects are contextually dispersed: some are political insiders, others are grassroots activists. Some opposed the bill, others were in favor. The research team conducting the interviews is interested in a factual account of what happened. But we have discovered that while there are some topics of broad agreement, other narratives are consistently associated with a particular context.

In general we can think of any inference, whatever the method by which it was derived, as a kind of summary or compression of the original data with reference to prior information and a representational language [8, 24]. This extends to inferences involved in the interpretation of human language. Coding of qualitative data, such as interview data, is operationally akin to feature learning, the process in machine learning of developing intermediary representations of data that are informative for further analysis. [3] With this understanding we have begun experimenting with using qualitative codings of interviews as an input to downstream computational data processing and visualization. We aim to develop a rigorous method for representing composite perspectives that reflects both the shared views and particular situated understandings of the participants involved.

In summary, the tension between data-driven and qualitative methods is based on a false dichotomy. Understanding the deep connection between qualitative methods and formally defined data-driven research techniques opens new opportunities for mixed-methods research that is both scalable and contextually sensitive. In particular, we see potential in methods that explore the problem of representing composite perspective as a promising frontier for human-centered data science.

Acknowledgements

I gratefully acknowledge the financial support of the Berkeley Institute for Data Science, and the helpful comments of David Bamman, Coye Cheshire, Paul Duguid, Tom Griffiths, and Brian Keegan.

References

1. John Anderson. 1991. Is human cognition adaptive? *Behavioral and Brain Sciences* 14: 471–517.
2. Cecilia Aragon, Joseph Bayer, Andy Echenique, Yun Huang, Clayton Hutto, Jinyoung Kim, Gina Neff, Wanli Xing. 2015. Developing a Research Agenda for Human-Centered Data Science.
3. Yoshua Bengio and Aaron C. Courville and Pascal Vincent. 2012. Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives. *CoRR* abs/1206.5538 http://arxiv.org/abs/1206.5538 Wed, 10 Oct 2012 21:28:53 +0200
4. Jeanette Blomberg, Mark Burrell, Greg Guest. 2002. An ethnographic approach to design. *The human-computer interaction handbook*. p. 964-986. L. Erlbaum Associates Inc. Hillsdate, NJ, USA. 2003.
5. Pierre Bourdieu. 2004. *Science of Science and Reflexivity* Translated by Richard Nice. The University of Chicago Press.
6. danah boyd, and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly, phenomenon. *Information, Communication & Society*, 15(5), 662-679.
7. Nick Chater, Mike Oaksford. 1999. Ten years of the rational analysis of cognition*. Trends in cognitive sciences* 3 (2): 57–65. 10.1016/s1364-6613(98)01273-x.
8. Nick Chater, Paul Vitanyi. 2003. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, Volume 7, Issue 1, January 2003, Pages 19-22, ISSN 1364-6613, http://dx.doi.org/10.1016/S1364-6613(02)00005-0.
9. Garret Christensen and Courtney Soderberg, Manual of Best Practices in Transparent Social Science Research. 2015. GitHub repository, https://github.com/garretchristensen/BestPracticesManual. Retrieved on December 10, 2015.
10. Cynthia Dwork. 2011. Differential privacy. In *Encyclopedia of Cryptography and Security* (pp. 338-340). Springer US.
11. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science*

Conference (ITCS '12). ACM, New York, NY, USA, 214-226. http://dx.doi.org/10.1145/2090236.2090255

12. Fisher, Ronald A. 1971/1935. *The Design of Experiments* (9th ed.). Macmillan.
13. Clifford Geertz. 1973. Thick Description: Toward an Interpretive Theory of Culture. In *The Interpretation of Cultures: Selected Essays.* New York: Basic Books. 3-30.
14. Stuart Geman, Elie Bienenstock, Rene Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural computation*. MIT Press.
15. Tom Griffiths, Charles Kemp, and Josh Tenenbaum. 2008. Bayesian models of cognition. In Ron Sun (ed.), *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.
16. Tom Griffiths. 2014. Manifesto for a new (computational) cognitive revolution. *Cognition*, http://dx.doi.org/10.1016/j.cognition.2014.11.026
17. Edwin Hutchins. 1996. Learning to navigate. In *Understanding Practice: Perspectives on Activity and Context*. Seth Chaiklin, Jean Lave (Eds). Cambridge University Press.
18. Jean Lave. 1996. "The practice of learning." Seth Chaiklin, Jean Lave. 1996. Understanding Practice: Perspectives on Activity and Context. Cambridge University Press.
19. John Lofland, David Snow, Leon Anderson, Lyn H. Lofland. 2006. *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Wadsworth.
20. David Marr. 1982. *Vision*. San Francisco, CA: W. H. Freeman.
21. Mark Monmonier. 1996. *How to lie with maps*. The University of Chicago Press.
22. Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford University Press.
23. Michael Sipser. 1996. *Introduction to the Theory of Computation*. Course Technology.
24. Ming Li and Paul Vitanyi. 2008. *An introduction to Kolmogorov Complexity and its Applications (3rd Edition)*. Springer.
25. Ron Sun. 2008. Introduction to Computational Cognitive Modeling. In *The Cambridge Handbook of Computational Psychology*, edited by Ron Sun. Cambridge University Press.
26. Josh Tenenbaum, Charles Kemp, Tom Griffiths, and Noah Goodman. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331 (6022), 1279-1285