Reflections on Our Human-Centred *Qualitative* Data Science Journey

ROBERT P GAUTHIER and JAMES R WALLACE, University of Waterloo, Canada

OUR RESEARCH EXPERIENCE

We are HCI researchers from the University of Waterloo's School of Public Health Sciences. Many of our colleagues are social science researchers, and are tackling important issues like vaccine hesitancy, mental health, and addiction. These colleagues are very excited about the potential benefits of data science in applied settings, but they are also limited by the technical and programming skills required to engage with current data science practices. Put simply, there is a gap between those with domain expertise and the means to enact positive changes in our health care systems, and those with the technical skills required to currently perform data science.

Towards these ends, over the past 4 years we have explored how we can make computational techniques more accessible to our colleagues to support and enhance their *qualitative* research, and specifically reflexive thematic analysis [2]. We are optimistic that the HCI community — a hub for multi-, trans-, and inter-disciplinary technology research — is a critical venue for human-centred data science to evolve. Yet, we simultaneously have experienced difficulty in engaging with the community and have concerns about how the research process has unfolded. To situate these concerns, we first summarize our own research experiences, before discussing interrogations and provocations for the workshop.

Addiction Recovery on Reddit

We first set out to develop holistic understandings of addiction recovery from communities' discussions on Reddit — we are presenting this work this year at CHI [7]. To do so, we performed a topic-guided thematic analysis: we used Latent Dirichlet Allocation (LDA) topic models [1] as a purposive sampling tool for reflexive thematic analysis [2]. We then (manually) performed inductive coding to develop and report on themes from that analysis.

Over the course of this study we identified two benefits of using topic modelling for thematic analysis. First, the act of performing the modelling helped us build data familiarity, and thus informed our reflexive thematic analysis. Second, the model enabled us to identify threads with reoccurring interesting phenomena and then to inductively identify codes and themes in a reasonable time. However, we also noted that our computational tasks took much of our time and energy relative to the qualitative tasks, which were the primary contributor to our analysis. Most of our time was spent on generic programming tasks, such as data object transfers between libraries and loading data into training scripts, that impact model performance/optimization but did not help us develop our understanding of the data. These issues are a barrier to wider adoption of computational techniques in fields like public health.

The Computational Thematic Analysis Toolkit

We then decided to explore how to enable (non-technical) qualitative researchers to use computational tools to perform thematic analyses, and developed the Computational Thematic Analysis Toolkit — published this year at GROUP [8]. Our toolkit provides a flexible, iterative, and visual interface to common data science and thematic analysis workflows (e.g., [3]). We also strove to encourage ethical research practices, as identified by the SIG CHI community (e.g., [4, 6, 9]), and support transparency to enable researchers to reflect on, improve, and report their research process and choices [11].

We have released the software as open source under the MIT licence. We hope that it serves to spark discussion about what tools should look like for the future of human-centred data science to be inclusive, ethical, and transparent. Since publishing the toolkit, we have experienced a significant degree of interest in the project from external groups, some of whom are now using the toolkit for their own analyses. For example, our toolkit is currently being used by health researchers to examine vaccine hesitancy on Canadian news website posts, and by HCI researchers to understand how people are discussing use of technology on Reddit. We hope to iterate and refine the toolkit using this feedback, and feedback from the broader HCI community over the coming months.

The toolkit's latest releases and living source code are available on github at:
https://github.com/rpgauthier/ComputationalThematicAnalysisToolkit

PROVOCATIONS AND CALLS TO ACTION

In reflecting on this work, we have encountered some challenges that we feel offer opportunities to interrogate how human-centred data science is approached by the SIG CHI community. We have observed that epistemological differences often make dissemination more difficult. We have also encountered challenges in navigating the transparency and ethical expectations created by the highly contextual nature of human-centred data science. We elaborate on these as potential interrogations below.

Technical Focus

There is a common belief that data science requires a programming and math expertise, particularly when conceptualized within SIG CHI, a special interest group of the Association for Computing Machinery. In short, Computer Scientists like to create new technology. This belief centres post-positivist thinking and encourages researchers to focus on quantitative and programming-based research, such as optimization on quantitative metrics. It also de-centres interpretivist research and research that focuses on applications of technology and its impact. While we understand the historical and cultural reasons behind these beliefs, we also believe that they are obstacles to the development of human-centred data science.

We have found that academic structures within the ACM make de-centring technology, and centring the human, incredibly difficult. For instance, the burden of proof is much higher when publishing qualitative, interpretive work than it is for quantitative or artifact-driven research [11]. We have to admit more than a little frustration in publishing (or failing to) our qualitative work at CHI. But this isn't just us. Voices from around the CHI community have acknowledged biases in the treatment of qualitative research [11], highlighting issues like sample size fallacy [5] and the over-use of IRR [10].

And so we feel that it's important ask, is CHI the right place for human-centred data science to evolve? It *should be*. CHI is one of the few places where multi-, inter-, and trans-disciplinary work is embraced, and it welcomes researchers from a variety of fields. But our experience also tells us that there is much room for improvement; by recognizing our own biases, by including a larger group of voices in the conversation, and by considering the constraints that our peer review system places on us.

Transparency and Ethics

Transparency and ethics are important considerations for any research. However, to be human-centred data science, there is an implicit expectation that researchers behave ethically, use data appropriately [6], and avoid harming the people who created, used, and analyzed the data [4]. However, these two considerations can end up in conflict when interpreted through the diverse and sometimes contradictory community contexts. For instance, how can researchers be transparent with sources of quotations when that data could out people in stigmatized situations?

We appreciate the availability of transparency and ethics guidelines for both researchers and the community (e.g.,[4, 6, 9]). However, we also worry that such guidelines can become problematic. They can become standardized rules that researchers *have to* satisfy regardless of context. They can become checklists that are applied without thought. And they can create expectations for authors to disclose information even when inappropriate. When research communities begin to use guidelines as wrote rules, they begin to downplay the necessity of weaving transparency and ethics thinking into the research process itself.

Instead, we believe that we need to make space for discussions about ethics and transparency in our publications and peer review process. Research tools need to aid researchers with built-in support for ethics and transparency, and support them in making appropriate decisions and disclosures based on the ethics and context of their research. There should be more room for discussion of these choices in our papers and within the reviewing process. And we need to provide some shelter from potential negative consequences of these disclosures — authors typically expose themselves and their research to additional criticism (in an already highly critical setting) by making their research more transparent. Detailing the ethical and transparency decisions made in research can enable the research community to ask more detailed questions, to scrutinize, and even potentially be more skeptical; but these discussions will ultimately improve our human-centred data science practices.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[3] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, Pranee Liamputtong (Ed.). Springer, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[4] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4, 3 (2002), 217–231. https://doi.org/10.1023/A:1021316409277

[5] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[6] Casey Fiesler, Nathan Beard, and Brian C. Keegan. 2020. No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 187–196.

[7] Robert P. Gauthier, Mary Jean Costello, and James R. Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, 17. https://doi.org/10.1145/3491102.3502076

[8] Robert P. Gauthier and James R. Wallace. 2022. The Computational Thematic Analysis Toolkit. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 25 (jan 2022), 15 pages. https://doi.org/10.1145/3492844

[9] Annette Markham. 2012. Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts. *Information, Communication & Society* 15, 3 (2012), 334–353. https://doi.org/10.1080/1369118X.2011.641993

[10] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov 2019), 72:1–72:23. https://doi.org/10.1145/3359174

[11] Poorna Talkad Sukumar, Ignacio Avellino, Christian Remy, Michael A. DeVito, Tawanna R. Dillahunt, Joanna McGrenere, and Max L. Wilson. 2020. Transparency in Qualitative Research: Increasing Fairness in the CHI Review Process. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381066