Developing a Research Agenda for Human-Centered Data Science

Cecilia Aragon
University of Washington
Seattle, WA 98195, USA
aragon@uw.edu

Joseph Bayer
University of Michigan
Ann Arbor, MI 48109, USA
joebayer@umich.edu

Andy Echenique
University of California, SD
San Diego, CA 92122, USA
aechenique@eng.ucsd.edu

Yun Huang
Syracuse University
Syracuse, NY 13210, USA
yhuang@syr.edu

Clayton Hutto
Georgia Institute of Technology
Atlanta, GA 30332, USA
Clayton.Hutto@gtri.gatech.edu

Jinyoung Kim
University of Maryland
College of Park, MD 20740, USA
jkim0204@umd.edu

Gina Neff
University of Washington
Seattle, WA 98195, USA
gneff@uw.edu

Wanli Xing
University of Missouri
Columbia, MO 65211, USA
wxdg5@mail.missouri.edu

Abstract

The study and analysis of large and complex data sets offer a wealth of insights in a variety of applications. Computational approaches provide researchers access to broad assemblages of data, but the insights extracted may lack the rich detail that qualitative approaches have brought to the understanding of sociotechnical phenomena. How do we preserve the richness associated with traditional qualitative methods while utilizing the power of large data sets? How do we uncover social nuances or consider ethics and values in data use?

These and other questions are explored by *human-centered data science*, an emerging field at the intersection of human-computer interaction (HCI), computer-supported cooperative work (CSCW), human computation, and the statistical and computational techniques of data science. This workshop, the first of its kind at CSCW, seeks to bring together researchers interested in human-centered approaches to data science to collaborate, define a research agenda, and form a community.

An early version of this workshop was developed and piloted at the 2015 CSST Summer Institute where it was voted best workshop of the conference.

Author Keywords
Data Science, Human-Centered Design, Human-Centered Data Science, Research Methods, Social Media

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

Overview

Large-scale data analysis has opened up opportunities in a wide variety of social, scientific, and technological areas and led to a variety of innovations [1, 2]. Yet the increasing focus on purely statistical or computational approaches may fail to capture social nuances, affective relationships, or ethical, value-driven, and other human-centered concerns.

Small scale, qualitative approaches to data collection and analysis offer researchers the opportunity to obtain very rich, deep insights about very specific phenomena —often in a very bounded or limited context [3]. Such studies often face challenges related to generalization, extension, verification, and validation. On the other hand, large scale, quantitative approaches to data collection and analysis offer researchers access to broad assemblages of data, but the insights gleaned are often much more shallow—lacking the rich detail associated with deep study [4].

But what happens as qualitative data sets grow ever larger? With the ease of collecting qualitative data such as social media text and multimedia photos and videos, such data sets are becoming an increasing challenge to analyze with the same level of detail and depth. How do we preserve the richness associated with traditional qualitative techniques in data-driven research? How can we be sure not to lose the compelling and inspiring stories of individuals in the sea of aggregated data at scale?

There are clear advantages of each perspective—one can choose methods and techniques which facilitate deep, but narrow analysis, or one can be broad, but shallow [5, 6, 7]. Various techniques allow researchers to track down traces of human behaviors, but affective elements and social context might not be well represented. Human interpretation of data is, in the end, still necessary [8, 9, 10, 11].

Human-centered data science includes opportunities for researchers of both qualitative and quantitative traditions. Researchers have addressed this trend and attempted to integrate quantitative research methods into a qualitative research workflow [9, 12, 13]. Digital or virtual ethnography [12] has gained widespread adoption as qualitative researchers adapt traditional ethnographic methods to online spaces. Data science tools that integrate seamlessly into the sociotechnical ecosystem of the domain they are designed for have demonstrated the greatest success. Human-centered design is particularly effective in the development of software for the analysis of large data sets.

Among the many unanswered questions surrounding human-centered data science include issues of sampling, selection, and privacy. What are the ethical questions raised by the necessity to process vast data sets? How should we treat crowdworkers? Who owns personal medical data, the company whose machines and software collect it, or the individual who generates it? Can design be effectively crowdsourced? What are

the policies we need to develop to protect human rights in this new age of “big data”?

The questions are legion and we are only beginning to explore the territory of potential answers [5, 14, 15, 16, 17].

We welcome researchers interested in exploring how data-driven and qualitative research can be integrated to address complex questions in a diverse range of areas, including but not limited to social computing, urban, health, or crisis informatics, scientific, business, policy, technical, and other fields. Researchers and practitioners working with large data sets and/or qualitative data sets looking to expand their methodological toolbox are invited to participate and share their experiences while learning from the broader community.

Workshop Objectives

This workshop has the following goals:

- Build and connect an international community of researchers in human-centered data science.
- Define and develop the terms and techniques of human-centered data science.
- Identify, encourage, and facilitate opportunities for research cooperation, especially multi-institute, and interdisciplinary collaboration.
- Develop a research agenda for human-centered data science.
- Produce a co-authored paper or edited journal issue on the research agenda for human-centered data science.

Workshop Topics and Themes

This workshop provides a venue for attendees to discuss a variety of topics in human-centered data science. Topics and themes of interest include, but are not limited to:

- **Deep ethnographic methods**: How do we preserve the richness of traditional qualitative techniques in data science?
- **Scaling up qualitative data analysis**: How do we deal with ever growing qualitative datasets?
- **Quantitative and behavioral methods**: How are quantitative and behavioral methods related to data mining, machine learning, and qualitative methods?
- **Connecting across levels of analysis**: How can we integrate the analysis of personal data with large-scale data?
- **Ethics and values of data use**: What ethical questions should we raise in using large-scale online data?
- **Privacy of data use:** How can we preserve anonymity and privacy within data ecosystems that can easily expose users?
- **Human-centered algorithm design**: How do we design machine learning algorithms tailored for human use and understanding?
- **Understanding community data**: How can we integrate knowledge gained about communities from their aggregate social data as well as their personal experiences?
- **Health and well-being at micro and macro scales**: What understandings can be exposed or occluded by aggregate or granular perspectives on health and well-being?

Activities and Format

This one-day workshop will be structured to facilitate deep discussions of current issues in human-centered data science. The workshop will be limited to no more than 25 participants to facilitate in-depth discussion.

An early version of this workshop was developed and piloted at the 2015 CSST Summer Institute (http://www.sociotech.net/summer-institutes/2015-summer-research-institute-colorado-springs-colorado/), where participants voted it the most effective and interesting proposal of the conference. We were awarded a small amount of funding which we intend to use to invite a keynote speaker, rent a projector, and procure supplies for the CSCW workshop.

Pre-workshop activities

We will produce a workshop website and solicit position papers 2-4 pages in length. Papers will be peer-reviewed. Authors of selected papers will be notified of their selection and asked to read and comment on the other papers before the conference.

Workshop activities

We will begin with a 20-minute keynote on the methodological challenges related to combining quantitative and qualitative approaches to research on very large data sets. This framing session will be followed by small group activities. We will avoid the "series of 30-minute talks" approach, which is less effective in fostering collaboration and building community.

We will encourage the group to have lunch together to continue informal conversation and generate potential research collaborations.

Activities will include:

Lighting talks: Strict 3-minute madness-style presentations where individuals present key points from their position paper.

Research speed dating: Form two lines and take 90 seconds per person to describe to a partner your research interests in human-centered data science. We have used this technique with success in past workshops to jump-start research collaborations.

Case studies: Several data-driven research questions are given to each group. Small groups of 3-4 discuss how the project might be improved through a human-centered data science approach.

Small group note cards: Participants gather in small groups to write out common or key issues, problems, hurdles, or challenges related to human-centered data science. The whole group will organize these into thematic clusters, and then begin to identify common narratives.

We will conclude the workshop with a discussion of publication plans and the key issue of how to maintain momentum going forward.

Post-workshop activities

A group mailing list for further communication and collaboration will be generated. We will produce a summary paper or poster from the workshop discussion. We will then initiate further discussion among the group as to whether we should try to publish an article or journal special issue on the research agenda for human-centered data science.

Participation

Participants will be recruited from the CSCW community, CHI, AoIR and the extended research networks of the organizers. We will also create a public website outlining the purpose and goals of the workshop and will advertise it widely. We hope to attract a balanced mix of participants from academia, industry, and the public sector to attain a comprehensive view of human-centered data science.

Interested individuals should submit a 2-4 page position paper in CSCW extended abstracts format that addresses the workshop theme and one or more of the highlighted topics. Submissions will go through a peer-reviewed process under the workshop program committee, and will be accepted based on relevance and potential to contribute to the workshop discussion and goals.

Workshop Organizers

The organizers of this workshop comprise a diverse set of researchers and practitioners from academic and non-academic settings with substantial and varied combined experience in data science and human-centered design as well as workshop organization. We have included students, postdocs, data and research scientists, and junior and senior faculty, and look forward to the cross-pollination of a variety of perspectives in the structure and content of the workshop.

Cecilia Aragon is an associate professor in the Department of Human Centered Design & Engineering and a Senior Data Science Fellow at the eScience Institute at the University of Washington, where she directs the Human-Centered Data Science Lab. Her research interests include visual analytics, data-intensive scientific collaborations, and human-centered data science.

Gina Neff is an associate professor in the Departments of Communication and Sociology and is a Senior Data Science Fellow at the eScience Institute at the University of Washington. She studies how multidisciplinary teams collaborate with visualizations and data.

Wanli Xing is a PhD candidate in Information Science and Learning Technologies at University of Missouri. He has an interdisciplinary background in mathematics, statistics, learning sciences and computer science. His research interests are at the nexus of data science, human-centered computing, and learning analytics.

Jinyoung Kim is a PhD candidate in College of Information Studies at the University of Maryland, College Park. Her research interests lie in the areas of information behavior and women's life transitions.

Andy Echenique is a joint research fellow at the University of California, San Diego and San Diego Supercomputer Center. His research interests include cyberinfrastructure and the use of high performance computing tools for international collaboration.

Yun Huang is an Assistant Professor at the School of Information Studies in Syracuse University. She directs the Social Computing Systems Lab, SALT lab (http://salt.ischool.syr.edu). Her research interests include crowdsourcing systems, human-computer interaction, mobile applications, and accessible technologies.

Clayton Hutto is a Research Scientist working in the Human Systems Integration (HSI) Division at the Georgia Tech Research Institute (GTRI). He has a B.S. in Human Factors, a M.S. in Human Computer Interaction (HCI), and is currently finishing a doctoral degree in Human Centered Computing (HCC) from the Georgia Institute of Technology.

Joseph Bayer is a PhD Candidate at the University of Michigan. His research combines two perspectives on social connectedness in order to understand everyday social interaction, emotional experience, and personal well-being: (1) how social cognitive processes relate to mobile and social media use, and (2) how social network characteristics are linked to social cognition.

References

1. Julia Gluesing, Kenneth Riopelle, and James Danowski. 2014. Mixing Ethnography and Information Technology Data Mining to Visualize Innovation Networks in Global Networked Organizations. *Mixed Methods Social Networks Research: Design and Applications*, 36, 203.
2. Markus Luczak-Roesch, Ramine Tinati, Kieron O'Hara, and Nigel Shadbolt. (February, 2015). Socio-technical computation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (pp. 139-142). ACM.
3. Kai Zheng, David Hanauer, Nadir Weibel, and Zia Agha. 2015. Computational Ethnography: Automated and Unobtrusive Means for Collecting Data In Situ for Human–Computer Interaction Evaluation Studies. In *Cognitive Informatics for Biomedicine* (pp. 111-140). Springer International Publishing.
4. Tera Marie Green, Richard Arias-Hernandez, R., and Brian Fisher. 2014. Individual Differences and Translational Science in the Design of Human-Centered Visualizations. In *Handbook of Human Centric Visualization* (pp. 93-113). Springer New York.
5. 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.
6. Michael Brooks, John Robinson, Megan Torkildson, and Cecilia Aragon, 2014. Collaborative Visual Analysis of Sentiment in Twitter Events. In *Cooperative Design, Visualization, and Engineering* (pp. 1-8). Springer International Publishing.
7. David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science,* 343(6176), 1203-1205.
8. David Brooks. (2013, February 18). What Data Can't Do. *The New York Times*. Retrieved from http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html
9. Wanli Xing, Rui Guo, Eva Petakovic, and Sean Goggins. 2015. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, 47, 168-181.
10. Z-Q Liu, and Sadaaki Miyamoto. (Eds.). 2012. Soft computing and human-centered machines. Springer Science & Business Media.
11. Wanli Xing, Bob Wadholm, Eva Petakovic, and Sean Goggins, S. 2015. Group Learning Assessment: Developing a Theory-Informed Analytics. Journal of Educational Technology & Society, 18(2), 110-128.
12. Dhiraj Murthy. (2011). Emergent Digital Ethnographic Methods for Social Research. In S.

Hesse-Biber (Ed.), *Handbook of Emergent Technologies in Social Research* (pp. 158–179). New York: Oxford University Press.

13. Ashok Goel and Michael Helms. 2014. Theories, Models, Programs, and Tools of Design: Views from Artificial Intelligence, Cognitive Science, and Human-Centered Computing. In *An Anthology of Theories and Models of Design* (pp. 417-432). Springer London.
14. France Bélanger, and Robert Crossler. 2011. Privacy in the digital age: a review of information privacy research in information systems. *MIS quarterly*, 35(4), 1017-1042.
15. Sangita Ganesh, and Rani Malhotra. 2014. Designing to scale: A human centered approach to designing applications for the Internet of Things. In *Advanced Computing and Communications (ADCOM), 2014 20th Annual International Conference on* (pp. 26-28). IEEE.
16. Yang Wang, Yun Huang, and Claudia Louis. (September, 2013). Towards A Framework for Privacy-Aware Mobile Crowdsourcing. In Social Computing (SocialCom), 2013 International Conference on (pp. 454-459). IEEE.
17. Yang Wang, Huichuan Xia, and Yun Huang. 2016. Examining American and Chinese Internet Users' Contextual Privacy Preferences of Behavioral Advertising. To appear in Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2016)