Human-Centered Data Science: a Primer

Alex H. Poole

LEADING

June 23, 2021

Agenda

- Big Data
- Two cultures
- Need for HCDS
- HCDS defined
- HCDS foci
    - Data ethics
    - Design
    - Translation
    - Privacy
    - Empathy
    - Data feminism
    - The opportunity for IS

Big Data revisited

- Data
  - Any type of information systematically collected, organized, and analyzed
  - Reduction of life world experiences
- “Data-driven decisions tend to be better decisions” (McAfee & Brynjolfsson, 2012, 68).
- Datafication
  - “Each of us is now a walking data generator” (McAfee & Brynjolfsson, 2012, 63)
- “Every single dataset is likely to have some intrinsic, hidden, not yet unearthed value” (Mayer-Schönberger & Kukier, 2013, 15)
- Prospect of displacing human subjectivity
  - What people do, not what they say they do (Tanweer et al., 2016)
- Social media, financial transactions, and public transportation use (Kogan et al., 2020)
  - Big data failures: 1) Google flu; 2) 2016 US presidential election

Quantitative and qualitative methods

- Clash between qualitative and qualitative research traditions (Benthall, 2016)
  - Specificity versus aggregation
- Methodological separation
  - Social network, collective action, peer-production, and crowdsourcing research (Maddock et al., 2016)
- Statistics, e.g. machine learning
  - Generalize from large amounts of data
    - Assign measures of probability to well-defined hypotheses
  - Statistical data analysis often involves data sets with many instances and few features, e.g. U.S. Census
- Qualitative researchers as instruments, e.g. ethnography
  - Data: many features and few instances.
    - Contextual sensitivity/situated understanding
      - Discern connections among phenomena
    - But not generalizable

The need for a Human-Centered Data Science (HCDS)

- Human-centered approaches novel for data science
    - Data-driven analytics supersede human judgments (Anderson & Parker, 2019)
- Statistical/computational approaches
    - Miss social nuances, affective relationships, or ethical, value-driven, and other human-centered concerns (Aragon et al., 2016)
    - “technological redlining” (Noble, 2018) and dataveillance
        - Mask and deepen social inequality
    - Fairness, responsibility, human rights (Floridi & Taddeo, 2016)
- Need methods
    - Automated power of computational techniques
    - Situated and nuanced social activity (Kogan et al., 2020)
- Embrace both quantitative and qualitative approaches
    - E.g.: machine learning algorithms augment ethnographic work
        - Identify suitable individuals for interviews from a massive dataset of digital logs
        - High-level aggregate trends within a population for further investigation in the field

HCDS definition

- Human-centered: “used to describe computers, technology, systems, etc. that are designed to work in ways that people can easily understand and learn” (Cambridge English Dictionary)
- “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” (Aragon et al., 2016, 529)
  - “understand, theorize, and codify the complex interactions between human-centered and machine-based approaches to dealing with vast human-generated data sets” (Kogan et al., 2020, 152)
  - Nuance, richness, value-drive, relational, complex, context-dependent

Data ethics

- Macroethics
    - Overarching, holistic, and inclusive framework (Floridi & Taddeo, 2016)
    - Privacy, anonymity, transparency, trust, responsibility
- Foci: interactions among hardware, software, and data
    - Data
        - Generation, recording, curation, processing, dissemination, sharing and use
    - Algorithms
        - Artificial intelligence, artificial agents, machine learning and robots
    - Practices
        - Responsible innovation, programming, hacking, professional codes
- Goal: formulate and implement morally good solutions
    - “social preferability” is sine qua non (Floridi & Taddeo, 2016, 2)

Design

- Data science workers intuitively and actively shape their data (Muller, Lange, et al., 2019)
- Diverse stakeholders in data science
  - System developers, analysts, data subjects
    - Understand moral obligations and choices
    - Impact data system design and use (Shilton, 2016)
    - Design datasets that preserve multiplicities of experience (Young, 2016)
    - Shape experiences that improve lives (Girardin & Lathia, 2017)
- Values in infrastructure
  - Implications: free speech and censorship, content privacy and security, network neutrality
    - Imagining users and use cases
    - Interdisciplinary conversation
    - Self-testing prototypes

Translation

- Translational work in both directions (data science and domain) essential (Bica, 2019)
- IDR work and communicating design values
  - Likely using same terms (“data,” “model,” “segment,” and “trend”)
  - But referring to very different things! (Girardin & Lathia, 2017)
- “overly-nuanced questions or categories, for instance differentiating between a risk and a threat, do not reflect the ways that people actually tweet, and using ML techniques to understand or categorize such themes will not be fruitful” (Bica, 2019)

Privacy

- Privacy as contextual integrity
  - Privacy judgments not individual preferences
    - Shared reactions to data uses in specific social contexts (Shilton, 2016)
- Weaknesses in Big Data privacy protection
  - 1) Many anonymized datasets re-identifiable
    - Correlating dataset with publicly available, identified data (a “join”)
  - 2) Notice and consent policies overemphasize individual responsibility (Young, 2016)
    - FTC’s Fair Information & Privacy Practices
      - Subjects may opt-in or opt out
      - But too much responsibility on the user’s ability to decide
- Open data
  - Seeming monolith
  - But should be seen as set of choices with respect to platforms, formats, licensing, and uses

Data empathy

- Empathy
    - “the ability to share in, understand, and identify with the experience of others; to take on another’s perspective in an affective or cognitive sense.
- Data empathy
    - “developing this ability for sharing and understanding different data valences, or the values, intentions, and expectations around data” (Tanweer et al., 2016)

Data feminism

- Data feminism
  - Intersectional feminism
  - Address oppression and power inequities
  - How can we use data to remake the world?
  - Both goal and process (D’Ignazio & Klein, 2020)
- Intersectionality
  - Race, class, sexuality, ability, age, religion, geography
  - Individual identity
  - Societal intersections of privilege and oppression
- “before there are data, there are people—people who offer up their experience to be counted and analyzed, people who perform that counting and analysis, people who visualize the data and promote the findings of any particular project, and people who use the product in the end”
  - Those who go uncounted!
  - Not all problems amenable to data representation or solution

Operationalizing data feminism

7 principles (D’Ignazio and Klein, 2020):

- Examine power
  - Analyze how power operates in the world.
- Challenge power.
  - Challenge unequal power structures and pursue justice
- Elevate emotion and embodiment
  - Value multiple forms of knowledge, e.g. that from people as living, feeling bodies
- Challenge binaries and hierarchies
- Embrace pluralism
  - Synthesize multiple perspectives
  - Prioritize local, Indigenous, and experiential ways of knowing.
- Consider context
  - Data are not neutral or objective
  - Products of unequal social relations
- Make labor visible so it can be recognized

An opportunity for Information Science

- Human-centered
  - Values such as privacy, human rights, and ethics
  - Ask what technology *should* do
  - Consider individuals, organizations, and communities behind production and consumption (Shah et al., 2021)
- Socially-responsible
  - Diversity, equality, and sustainability as well as utility
  - Inclusivity
    - Accessibility to people from all races, genders, ethnicities, and abilities

References

Anderson, T. D., & Parker, N. (2019). Keeping the human in the data scientist: Shaping human-centered data science education. *Proceedings of the Association for Information Science and Technology*, *56*(1), 601–603. https://doi.org/10.1002/pra2.103

Aragon, C., Hutto, C., Echenique, A., Fiore-Gartland, B., Huang, Y., Kim, J., Neff, G., Xing, W., & Bayer, J. (2016). Developing a Research Agenda for Human-Centered Data Science. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion - CSCW '16 Companion*, 529–535. https://doi.org/10.1145/2818052.2855518

Benthall, S. (2016, March 27). *The Human is the Data Science*. CSCW '16, San Francisco, CA.

Bica, M. (2019). Human-Centered Data Science for Collaborative, Interdisciplinary Research. *Proceedings of CHI EA '19: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*.

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Girardin, F., & Lathia, N. (2017). *When User Experience Designers Partner with Data Scientists*. 376–381. https://www.bbvaaifactory.com/publications/ux_datascientists.pdf

Kogan, M., Halfaker, A., Guha, S., Aragon, C., Muller, M., & Geiger, S. (2020). Mapping Out Human-Centered Data Science: Methods, Approaches, and Best Practices. *Companion of the 2020 ACM International Conference on Supporting Group Work*, 151–156. https://doi.org/10.1145/3323994.3369898

Maddock, J., Gergle, D., & Starbird, K. (2016, March 27). *Two is Better Than One: A Mixed Methods Approach to Human-Centered Data Science*. CSCW '16, San Francisco, CA.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 60–66, 68, 128.

Muller, M., Feinberg, M., George, T., Jackson, S. J., John, B. E., Kery, M. B., & Passi, S. (2019). Human-Centered Study of Data Science Work Practices. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3290607.3299018

Noble, S. U. (2018). *Algorithms of Oppression How Search Engines Reinforce Racism*. New York University Press.

Shah, C., Anderson, T., Hagen, L., & Zhang, Y. (2021). An iSchool approach to data science: Human-centered, socially responsible, and context-driven. *Journal of the Association for Information Science and Technology*, asi.24444. https://doi.org/10.1002/asi.24444

Shilton, K. (2016, March 27). *Empirical Ethics: Studying Values in Data Science Practice*. CSCW '16, San Francisco, CA.

Tanweer, A., Fiore-Gartland, B., Neff, G., & Aragon, C. (2016, March 27). *Data Empathy: A Call for Human Subjectivity in Data Science*. CSCW '16, San Francisco, CA.

Young, M. (2016, March 27). *A Human-Centered Approach to Data Privacy: Political Economy, Power, and Collective Data Subjects*. CSCW '16, San Francisco, CA.