Six provocations for Big Data

Danah Boyd, Kate Crawford
Samira Shaikh, Veena Ravishankar

Introduction: Importance of Big Data

- Notable because of relationality to other data
  - patterns derived from pieces of data
  - networked-individual, groups
- Analytic phenomenon in academia and industry
  - seeing patterns in data
  - find out about research methodology, technology used
- Culminates multiple disciplines
- Current decisions will have impact on future

Introduction

- Central questions of computational culture of big data are fundamental to Social and cultural researchers
- Addresses 6 provocations

Automating Research changes the definition of knowledge

- Big data creates a radical shift in how we think about research
- A profound change at the level of ethics and epistemology but not just a matter of scale or depth of data
- Reframes key questions:
  - constitution of knowledge
  - processes of research
  - engagement with information
  - nature and categorization of reality
- ex: correlation with Henry Ford devising automation and assembly lines of manufacturing

Automating Research changes the definition of knowledge

- Do massive numbers speak for themselves? NO.
  - Why people do things, write things is erased by sheer volume of numerical repetitions and large patterns
  - No space for older forms of intellectual craft
- Realize the limitations attached to Big Data
  - No historical context that is predictive and focuses on right now or immediate past
- Automation of tasks requires consideration of inbuilt flaws

Claims to objectivity and accuracy are misleading

- Big data is subjective and what it quantifies is not closer to objective truth
- Reinscribes established divisions in the debates of scientific method
    - Scientific methods develop hypothesis (objective) which are necessarily made by subjects based on subjective observation and choices
    - ex: data cleaning from social media is inherently subjective

Claims to objectivity and accuracy are misleading

- Issue of data errors
  - Understanding of properties and limits of datasets necessary
  - Social scientists carefully collect data and account for biases in it
  - To account for biases in data requires recognizing that one's identity and perspective informs one's analysis
  - ex: Dunbar's work of analyzing gossips in humans lead to wrong results
- Interpretation center of data analysis

Bigger data are not always better data

- Problematic underlying ethos: Quantity means quality, bigger the data the better
- Twitter provides an example in context of statistical analysis
  - Twitter does not represent the whole population
  - Meaning of user, participant and listener needs to be examined
  - no track of multiple accounts, lurkers
- Access to Twitter dataset through APIs varies
  - firehose: all public tweets except ones made private
  - spitzer: 1% of public account
  - gardenhose: 10% of public tweets
  - white-lists: different subsets of data

Bigger data are not always better data

- Limitations with data collection rarely acknowledged
  - Big data not whole data
  - Results of data collection do not reveal biases associated
  - Twitter data has methodological challenges rarely addressed
- Important to recognize value of small data
  - ex: work of Tiffany Veinot on a vault inspector to understand information practices of blue-collar worker
  - reframed definition of informal practices away from white-collar workers and spaces outside offices

Not all data are equivalent

- Context of data matters
- Two datasets modeled similarly does not make them equivalent
  - equating social media analysis with social graphs and social network analysis
  - not interchangeable
  - does not capture social relations
- Social science uses diverse methodological and analytical approaches
  - collection of data through surveys, interviews, observations and experiments
  - developed personal networks-relationship an individual develops and maintains
  - articulated and behavioral network not equivalent to personal network

Not all data are equivalent

- Measurement of tie strength through frequency or public articulation erroneous
- Risk in treating every connection as equivalent to other

Just because it is accessible Doesn't make it ethical

- What constitutes best ethical practice for researchers?
  - ex: Harvard based project released Facebook data publicly which was easily able to de-anonymize identities
- Researchers unaware of the harm
  - educational intervention seeking to discourage people from suicide increased in attempts
- Institutional Review Boards (IRBs)
  - framework for evaluating ethics of particular line of research
  - balances and checks in place to protect the subjects

Just because it is accessible Doesn't make it ethical

- Accountability
  - of research being done and research subjects required
  - multi-directional relationship
  - relation to colleagues, superiors, participants, public
- Researchers are provided with tools to breach privacy
- Subjects unaware of the agents and algorithms collecting data

Limited access to Big Data creates new digital divides

- Gap exists in the level of access people have to data
  - ex: social media companies have full access compared to outsiders
  - considerable unevenness in the system
  - Well-resourced universities buy access to data compared to others
- Gap based on skillset
- Gap based on gender
- Digital divide: Big data poor and Big data rich

Conclusion

- Important to start questioning about the assumptions, values, biases, methodologies associated with Big Data
- Current decisions will have an impact on future

Discussion: Questions to be addressed in class

- Under what circumstance is small data better than big data?

Discussion: Questions to be addressed in class

- Can any measures be taken to minimize (if not avoid) bias and subjectivity while working with big data?

Discussion: Questions to be addressed in class

- How much exactly the negative impacts of the new methods of Big data extraction outweigh the constructive results or is it the other way around??

Discussion: Questions to be addressed in class

- As the big data not representing everyone’s opinion or behavior. How much accurate information we can retrieve using that kind of data?

Discussion: Questions to be addressed in class

- What are the (ethical) limits of collecting user data from the social services and applications?

Discussion: Questions to be addressed in class

- Are privacy laws too strict or too relaxed?

Discussion: Questions to be addressed in class

- Making Big Data “open” will solve the problem where only a limited number of researchers have access to it, or different problems could emerge?

Discussion: Questions to be addressed in class

- How do one define accountability to the research subjects?

Discussion: Questions to be addressed in class

- Is it sometimes necessary to let go privacy of data for a greater good? (eg. surveillance systems that detect terrorist activities)

Thank you!