Critical approaches to B/big D/data

18th February 2015

Objectives

- Find out what you're all doing here
- Consider the critical questions for Big Data research set out by boyd and Crawford (2012)
- Discuss good and bad examples of Big Data research in the wild!

Why are you here?

Seriously, why?

Context: a provocation

The Web makes possible radical kinds of democracy, but also preserves the hegemony. In many respects, it has shifted over time toward the latter.

- In the old days, people disseminated information, sold things, met likeminded others
- Remember *Altavista*, *Lycos*, *excite*?
- As late as *Myspace*, people could customise their pages quite a lot (making it hard to mine data!)

Web 2.0 came, and brought a shift toward standardisation within social network pages (ergo, *Bigger Data*)

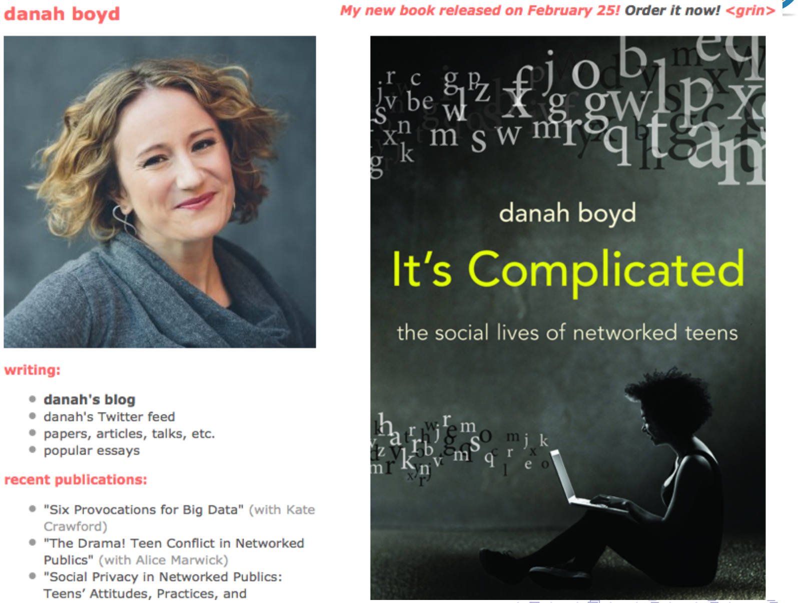
- Mergers, buy-outs
- Start-ups **aiming** to be bought out
- Fewer popular sites now, personal devices, fewer accounts

So, what's converged?

Just in terms of content—I imagine we could make a list for hardware too.

- Shopping
- Gaming
- Academia
- News
- SNSs
- Film and television
- Language learning
- Searching
- Dictionaries, encyclopædiæ
- Programming
- ?

danah.org



danah boyd

My new book released on February 25! ***Order it now!*** *<grin>*

writing:

- **danah's blog**
- danah's Twitter feed
- papers, articles, talks, etc.
- popular essays

recent publications:

- "Six Provocations for Big Data" (with Kate Crawford)
- "The Drama! Teen Conflict in Networked Publics" (with Alice Marwick)
- "Social Privacy in Networked Publics: Teens' Attitudes, Practices, and

Kate Crawford

I'm an academic researcher who works on issues of data, ethics and power. I'm a Principal Researcher at Microsoft Research, a Visiting Professor at the MIT Center for Civic Media, a Senior Fellow at NYU's Information Law Institute, and an Associate Professor at the University of New South Wales. I'm based in New York City.

You can email me at kate [at] katecrawford [dot] net. Or try @katecrawford on Twitter. If you're so inclined, here's my PGP Key. (from katecrawford.net)

Big Data

“We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of …

1. **Technology:** maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
2. **Analysis**: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
3. **Mythology**: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy” (p. 3)

Issue 1

"Big Data Changes the Definition of Knowledge"

- "Big Data not only refers to very large data sets and the tools and procedures used to manipulate and analyze them, but also to a computational turn in thought and research" (p. 6)
- "It is a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality" (p. 6)
- Archiving of Big Data is often not so great, forcing us to shift our understanding of what can be studied toward the present (think RSS feeds)
- *Do numbers speak for themselves?*

Issue 2

“Claims to Objectivity and Accuracy are Misleading”

- We can use Big Data to cherry pick or to mask/hide problems within the dataset.
- “**Data needs to be imagined as data in the first instance, and this process of the imagination of data entails an interpretative base**” (p. 10).
- “In the case of social media data, there is a ‘data cleaning’ process: making decisions about what attributes and variables will be counted, and which will be ignored. This process is inherently subjective” (p. 10).

Issue 3

“Bigger Data are Not Always Better Data”

- What can and can’t be in Twitter/Facebook data?
- “Twitter data has serious methodological challenges that are rarely addressed by those who embrace it. When researchers approach a dataset, they need to understand—and publicly account for—not only the limits of the dataset, but also the limits of which questions they can ask of a dataset and what interpretations are appropriate” (p. 14)
- Google N-grams: impressive because of its size, but there is:
  - no possibility of seeing meaning in context
  - a common conflation of word and concept

Issue 4

"Taken Out of Context, Big Data Loses its Meaning"

- An emerging obsession with mapping/charting/graphing social network data
  - Do these visuals *do* anything?
- Visualisation is really great, especially for journalism etc., but it often doesn't tell us much
- We often conflate digital measures of intimacy with closeness (*if they speak a lot on Facebook, they're great friends*)
- Do you think we will shift toward a culture where online/offline intimacy are more-or-less the same?

Issue 5

“Just Because it is Accessible Doesn’t Make it Ethical”

- The National Statement doesn’t cover Big Data
- We tend to just do what suits us and justify it…
- How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Issue 6

“Limited Access to Big Data Creates New Digital Divides”

The old digital divide was about people’s access to technologies. The new digital divide is about institutional access to the data produced thereon.

- Data has become potentially very cheap to collect (i.e. pay a computer science student for a few hours)
- Huge private datasets academics aren’t allowed to use and can’t replicate
- “those with money—or those inside the company—can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access” (p. 22)
- . . . though I suspect this has always been the case . . .

Access, ownership, control

Big Data is very much at the centre of the online economy. Hackers actually steal email addresses!

- "*If you don't pay for it, you're the product*" (?)
  - ...eh, sort of. It's complicated:
  - Sometimes you pay for more advanced features or more content (Busuu, NYT)
  - Sometimes you're exposed to ads
  - Sometimes your data is sold
  - On SNSs, user-generated content is the drawcard for new customers
- How do you even go about learning about these issues?
- Where is the issue actually being discussed?!

Adam Kramer

Q: Why did you join Facebook?

A: Facebook data constitutes the largest field study in the history of the world. Being able to ask—and answer—questions about the world in general is very, very exciting to me. At Facebook, my research is also immediately useful: When I discover something, we can use this to make improvements to the product. In an academic position, I would have to have a paper accepted, wait for publication, and then hope someone with the means to usefully implement my work takes notice. At Facebook, I just message someone on the right team and my research has an impact within weeks if not days.

(from facebook.com interview)

Our experiences

What kind of encounters have we all had with Big Data?

References I

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.