Six Provocations for Big Data

Danah Boyd

Microsoft Research

Kate Crawford

University of New South Wales

"Technology is neither good nor bad; nor is it neutral…"
Melvin Kranzberg (1986, p. 545))

"Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."
Geoffrey Bowker (2005, p. 183-184)

Will large-scale analysis of DNA help cure diseases? Or will it usher in a new wave of medical inequality?

Will data analytics help make people's access to information more efficient and effective?Or will it be used to track protesters in the streets of major cities?

Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means?

"Big Data is notable not because of its size, but because of its relationally to other data."

"Big Data is notable not because of its size, but because of its relationally to other data."

The United Kingdom National DNA Database
5,950,612 individuals

10 million Facebook users
personal information

"Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself."

"we are using Big Data here because of its popular salience and because it is the phenomenon around Big Data that we want to address."

Big Data "is the kind of data that encourages the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions."

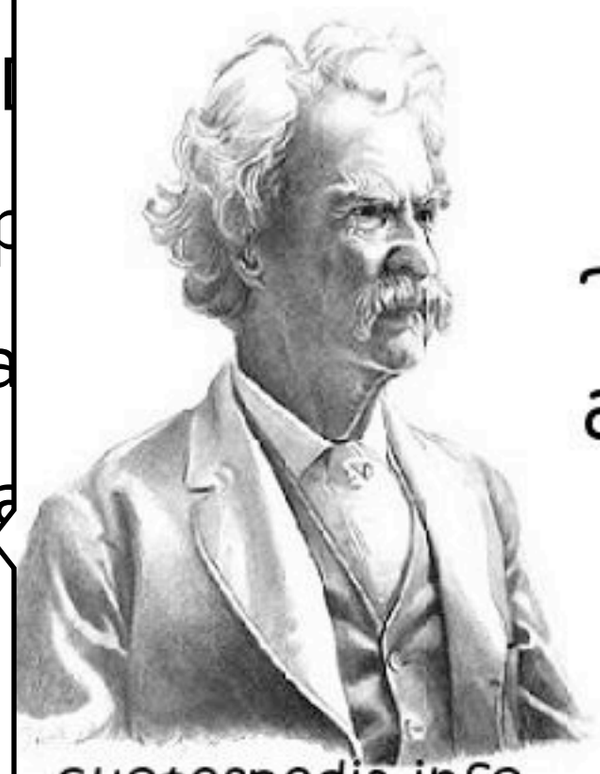
Apophenia /æpɵˈfiːniə/ is the human tendency to perceive meaningful patterns within random data.

Big I ce of

apop imply

beca s that

radia



Apophenia /æpɵˈfiːniə/ is the human tendency to perceive meaningful patterns within random data.

I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DON'T HAVE DATA.

"Data is increasingly digital air: the oxygen we breathe and the carbon dioxide that we exhale. It can be a source of both sustenance and pollution."

Internet ⟷ Big Data

"Yet, it is imperative that we begin asking critical questions about what all this data means, who gets access to it, how it is deployed, and to what ends."

“With Big Data come big responsibilities.”

1. Automating Research Changes the Definition of Knowledge.

"*computational turn* in thought and research"

"Big Data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community."

“Change the instruments, and you will change the entire social theory that goes with them,’ Latour reminds us (2009, p. 9)”

"This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."

?

"Instead of philosophy – which Kant saw as the rational basis for all institutions – 'computationality might then be understood as an onto-theology, creating a new ontological "epoch" as a new historical constellation of intelligibility' (Berry 2011, p. 12)."

"Big Data is about exactly right now"

"what can science learn from Google?', but to ask how Google and the other harvesters of Big Data might change the *meaning* of learning, and what new possibilities and new limitations may come with these systems of knowing."

2. Claims to Objectivity and Accuracy are Misleading

"Sociology has been obsessed by the goal of becoming a quantitative science."
Latour(2010:116)

"every discipline and disciplinary institution has its own norms and standards for the imagination of data."

Data cleaning

"A dataset may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a dataset, we need to know where data is coming from; it is similarly important to know and account for the weaknesses in that data. Furthermore, researchers must be able to account for the biases in their interpretation of the data"

"Spectacular errors can emerge when researchers try to build social science findings into technological systems."

"Big Data is at its most effective when researchers take account of the complex methodological processes that underlie the analysis of social data."

are big data and whole data the same?

"Unfortunately, some who are embracing Big Data presume the core methodological issues in the social sciences are no longer relevant. There is a problematic underlying ethos that bigger is better, that quantity necessarily means quality."

Some accounts are ‘bots’ that produce automated content without involving a person.

accounts and users are equivalent

the notion of an ‘active’ account

others participate as ‘listeners’

make claims about people and users

Is the data representative of all tweets?

Although a handful of companies and startups have access to the firehose, very few researchers have this level of access. Most either have access to a 'gardenhose' (roughly 10% of public tweets), a 'spritzer' (roughly 1% of public tweets), or have used 'white-listed' accounts where they could use the APIs to get access to different subsets of content from the public stream. It is not clear what tweets are included in these different data streams or sampling them represents.

"two people are physically co-present – which may be made visible to cell towers or captured through photographs – does not mean that they know one another."

“Without taking into account the sample of a dataset, the size of the dataset is meaningless.”

“two people are physically co-present – which may be made visible to cell towers or captured through photographs – does not mean that they know one another.”

"Big Data introduces two new popular types of social networks derived from data traces: 'articulated networks' and 'behavioral networks.'"

"Should someone be included as a part of a large aggregate of data? What if someone's 'public' blog post is taken out of context and analyzed in a way that the author never imagined? What does it mean for someone to be spotlighted or to be analyzed without knowing it? Who is responsible for making certain that individuals and communities are not hurt by the research process? What does consent look like?"

5. Just Because it is Accessible Doesn’t Make it Ethical

5. Just Because it is Accessible Doesn’t Make it Ethical

- Privacy
- Accuracy
- Property
- Access

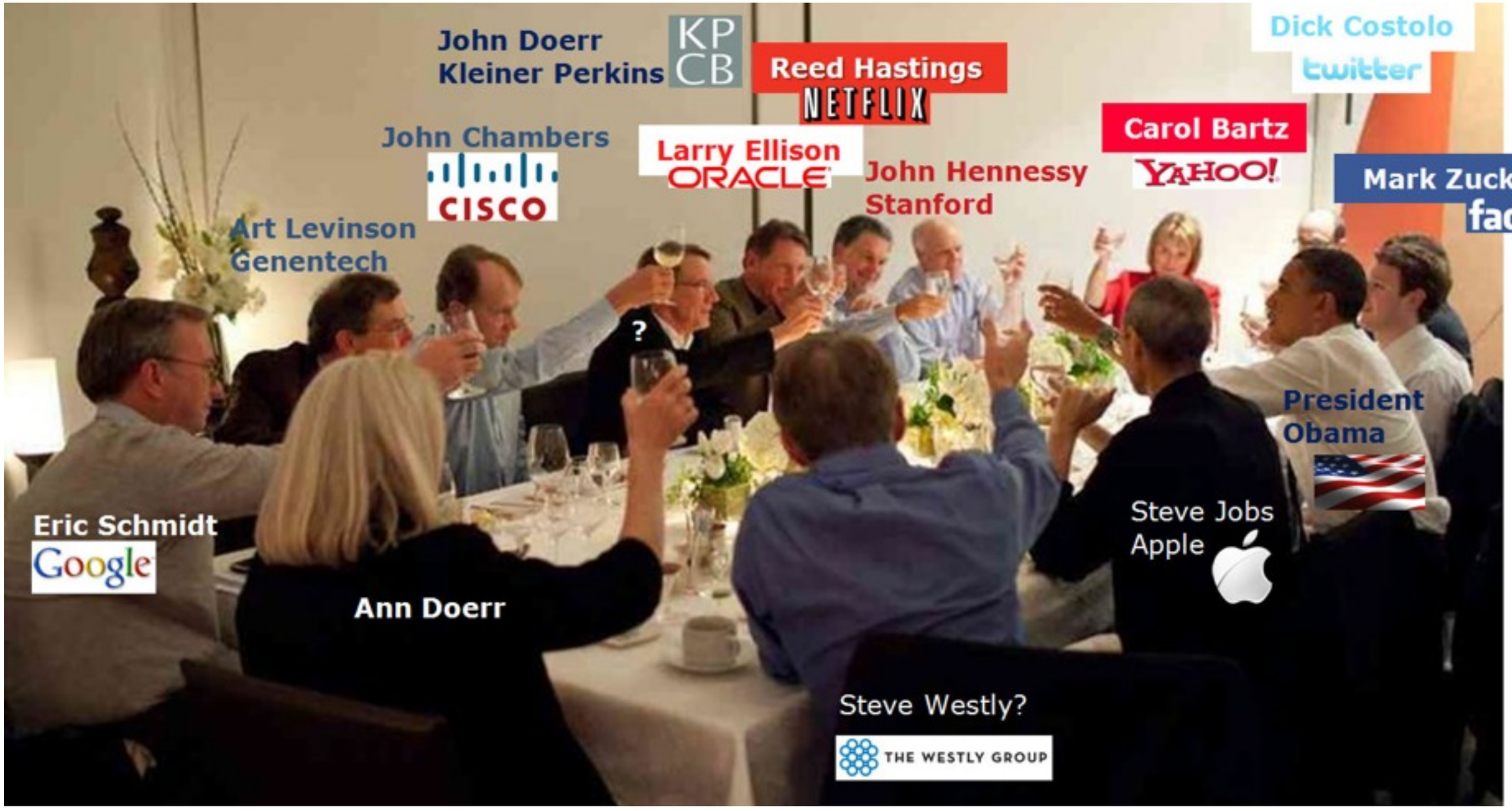
privacy, being in public, being public, accountability

Is it “unethical for researchers to justify their actions as ethical simply because the data is accessible”?

6. Limited Access to Big Data Creates New Digital Divides

But who gets access? For what purposes? In what contexts? And with what constraints?

"the market, the law, social norms, and architecture – or, in the case of technology"


John Doerr
Kleiner Perkins
KP CB
Reed Hastings
NETFLIX
Dick Costolo
twitter
John Chambers
CISCO
Larry Ellison
ORACLE
Carol Bartz
YAHOO!
John Hennessy
Stanford
Mark Zuckerberg
facebook
Art Levinson
Genentech
?
President Obama
Eric Schmidt
Google
Steve Jobs
Apple
Ann Doerr
Steve Westly?
THE WESTLY GROUP

"only social media companies have access to really large social data - especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not."

"There are complex questions about what kinds of research skills are valued in the future and how those skills are taught."

Manovich writes of three classes of people in the realm of Big Data:

- those who create data (both consciously and by leaving digital footprints)
- those who have the means to collect it, and
- those who have expertise to analyze it

- those who add noise to it
- those who rethink it

Thank you