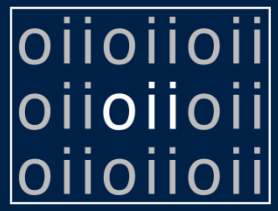


Oxford Internet Institute
University of Oxford

Big Data and the Uses and Disadvantages of Scientificity for Social Research

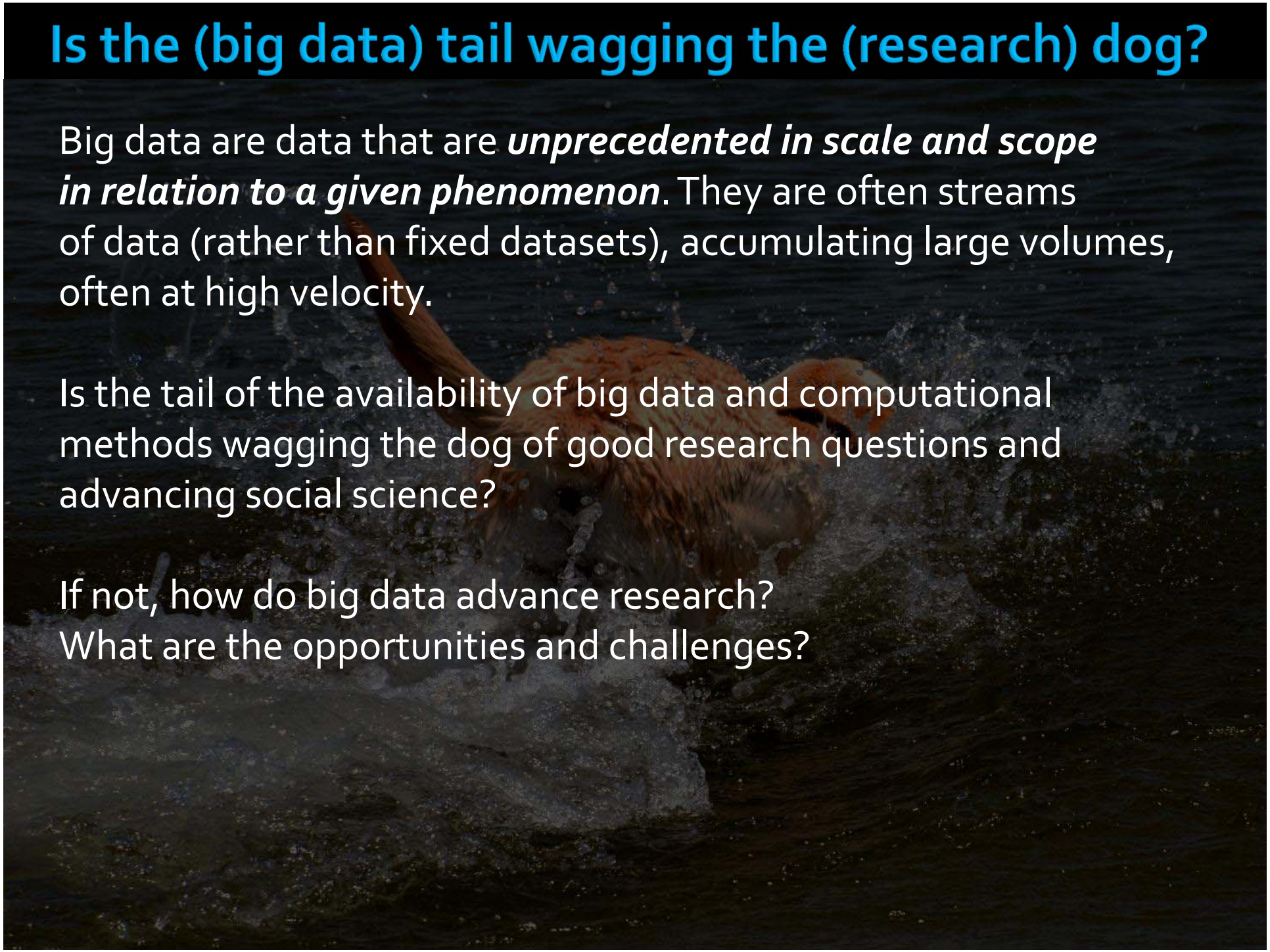
Ralph Schroeder, Professor
Eric Meyer, Research Fellow
Linnet Taylor, Researcher

SPRU, May 24, 2013

Is the (big data) tail wagging the (research) dog?

Source: Leonard John Matthews, CC-BY-SA (http://www.flickr.com/photos/mythoto/3033590171)

Is the (big data) tail wagging the (research) dog?



Big data are data that are ***unprecedented in scale and scope in relation to a given phenomenon***. They are often streams of data (rather than fixed datasets), accumulating large volumes, often at high velocity.

Is the tail of the availability of big data and computational methods wagging the dog of good research questions and advancing social science?

If not, how do big data advance research?
What are the opportunities and challenges?

Business Value versus Academic Value

Strategic Knowledge

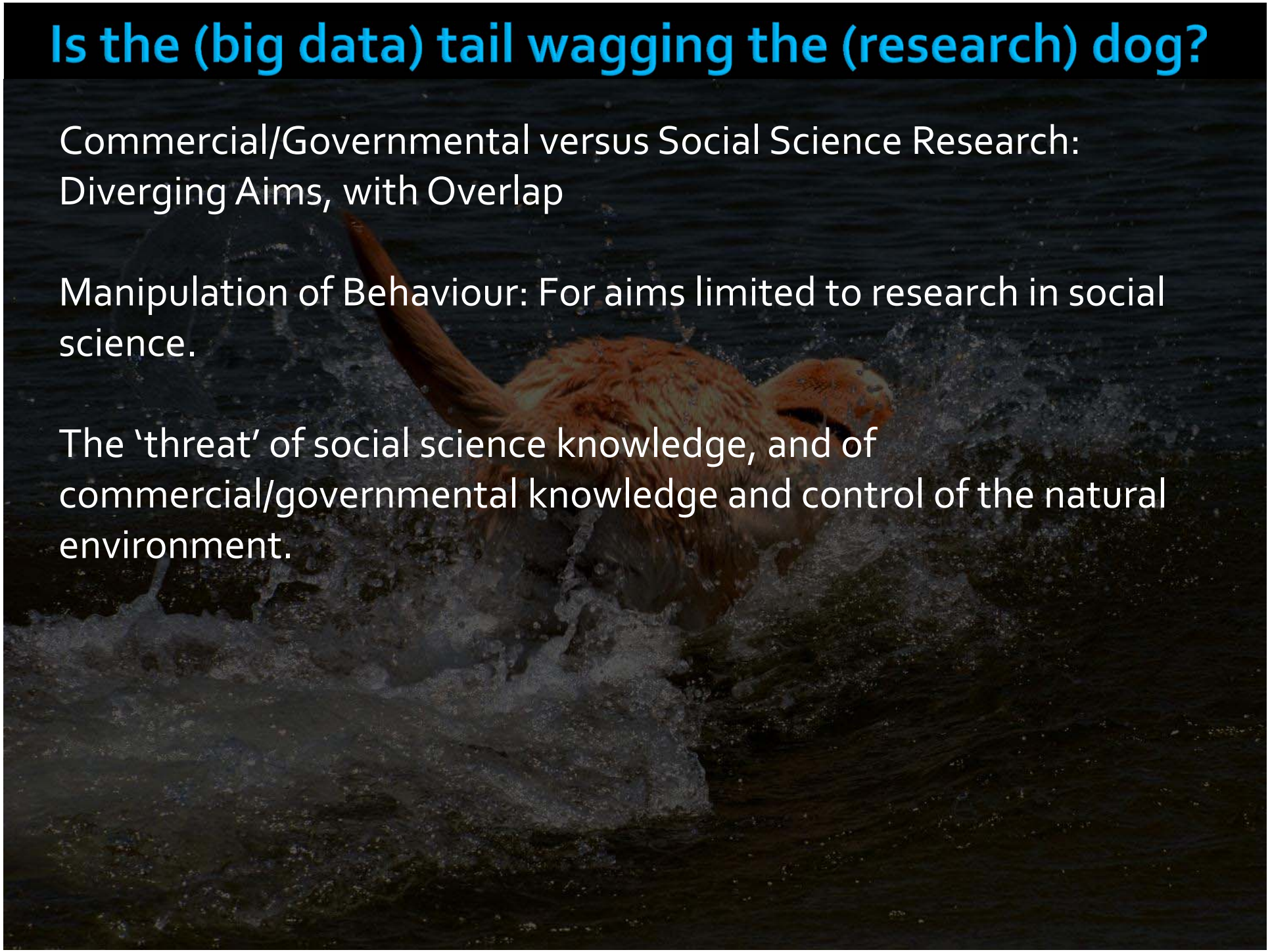
- Generally time-limited (with exceptions)
- Value comes from knowing what your competitors don't
- Often has high monetary value if it can be exploited

Business Value versus Academic Value

Durable Knowledge

- Less time-limited (with exceptions)
- Value comes from adding to the world's knowledge (the global brain is cumulative/scientific)
- Rarely has direct monetary value, but has value in terms of creating the possibility both of future knowledge and of future exploitation and commercial uses

Is the (big data) tail wagging the (research) dog?



Commercial/Governmental versus Social Science Research: Diverging Aims, with Overlap

Manipulation of Behaviour: For aims limited to research in social science.

The 'threat' of social science knowledge, and of commercial/governmental knowledge and control of the natural environment.

Big Data Analytics

- Access to data
- Cost of analytical tools
- Skills to use the tools
- Why should anyone share?
- How different skills and disciplines work together
- Starting with questions, or starting with data?
- Prediction?
  - A/B and other experiments
- Gaps?
- Futures

From Big Data to Big (Hi-res) Picture

Marketing → Tailoring

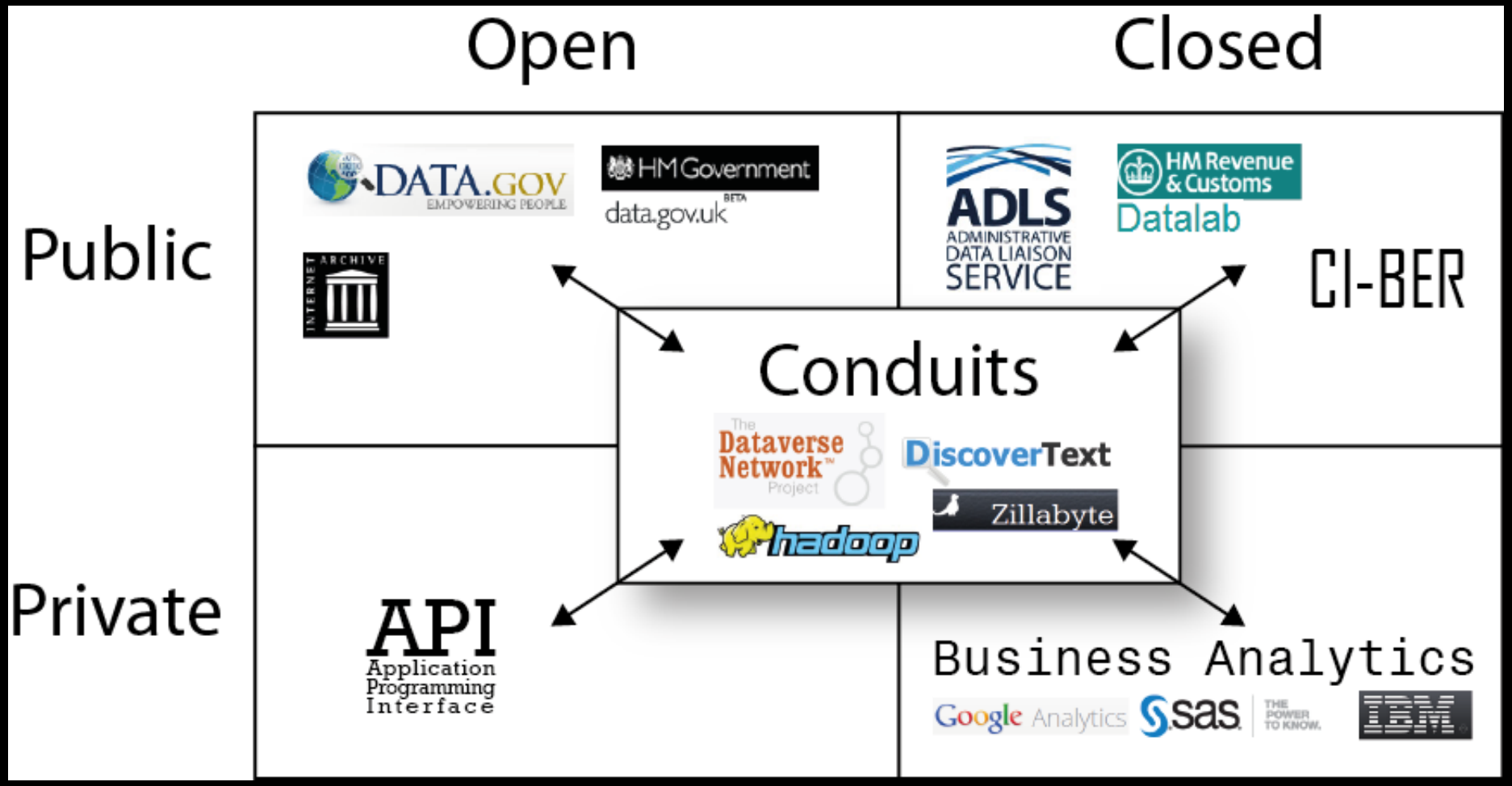
Forecasting → Prediction

Complex Trends → Linking datasets plus modelling

Big Data

Alfred P. Sloan Foundation

Accessing and Using Big Data to Advance Social Science Knowledge



See http://www.oii.ox.ac.uk/research/projects/?id=98

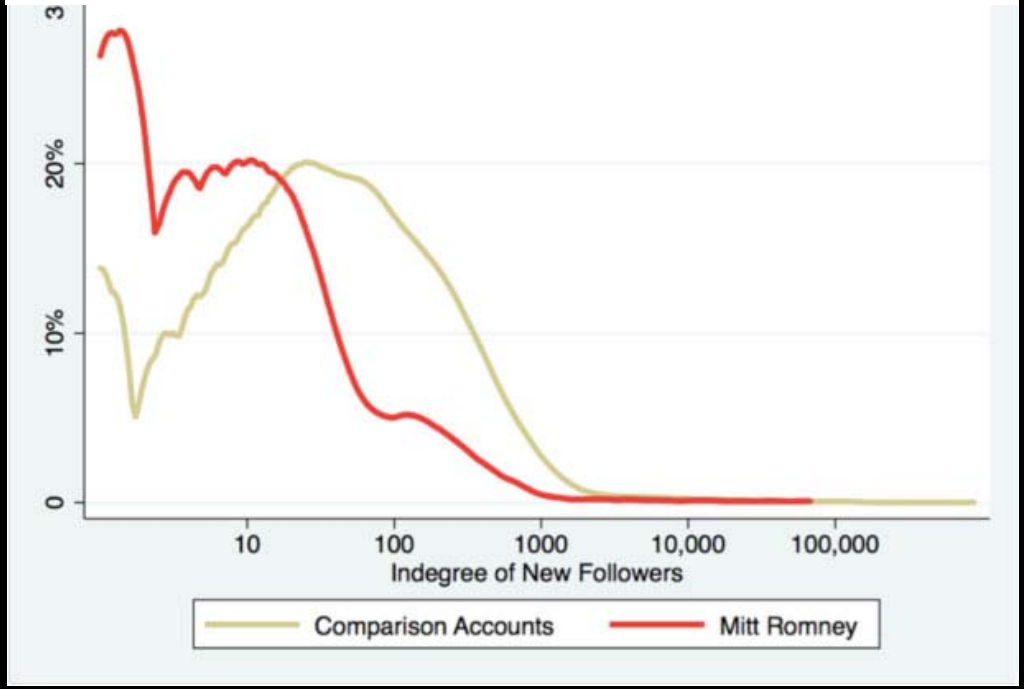
Twitter-bots

the Atlantic

OII master's students Alexander Furnas and Devin Gaffney saw a large spike in then-US presidential candidate Mitt Romney's Twitter followers, and decided to look at the new followers:

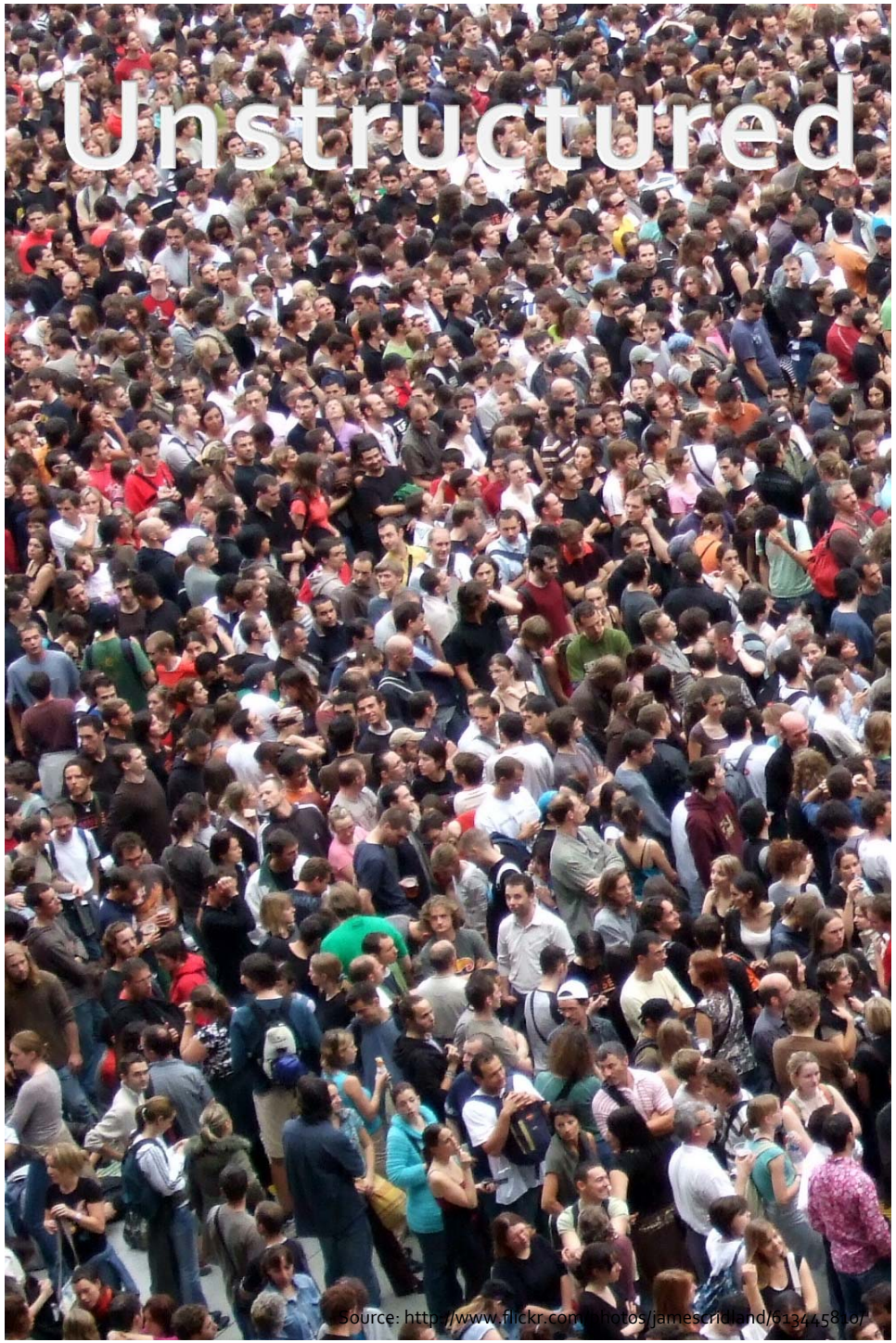
Statistical Probability That Mitt Romney's New Twitter Followers Are Just Normal Users: 0%

JUL 31 2012, 11:37 AM ET 132



Furnas, A. and Gaffney, D. (2012). 'Statistical Probability That Mitt Romney's New Twitter Followers Are Just Normal Users: 0%'. *The Atlantic*, July 31, http://www.theatlantic.com/technology/archive/2012/07/statistical-probability-that-mitt-romneys-new-twitter-followers-are-just-normal-users-0/260539/ (accessed August 31, 2012).

Source: http://www.flickr.com/photos/nakedcharlton/597075830/

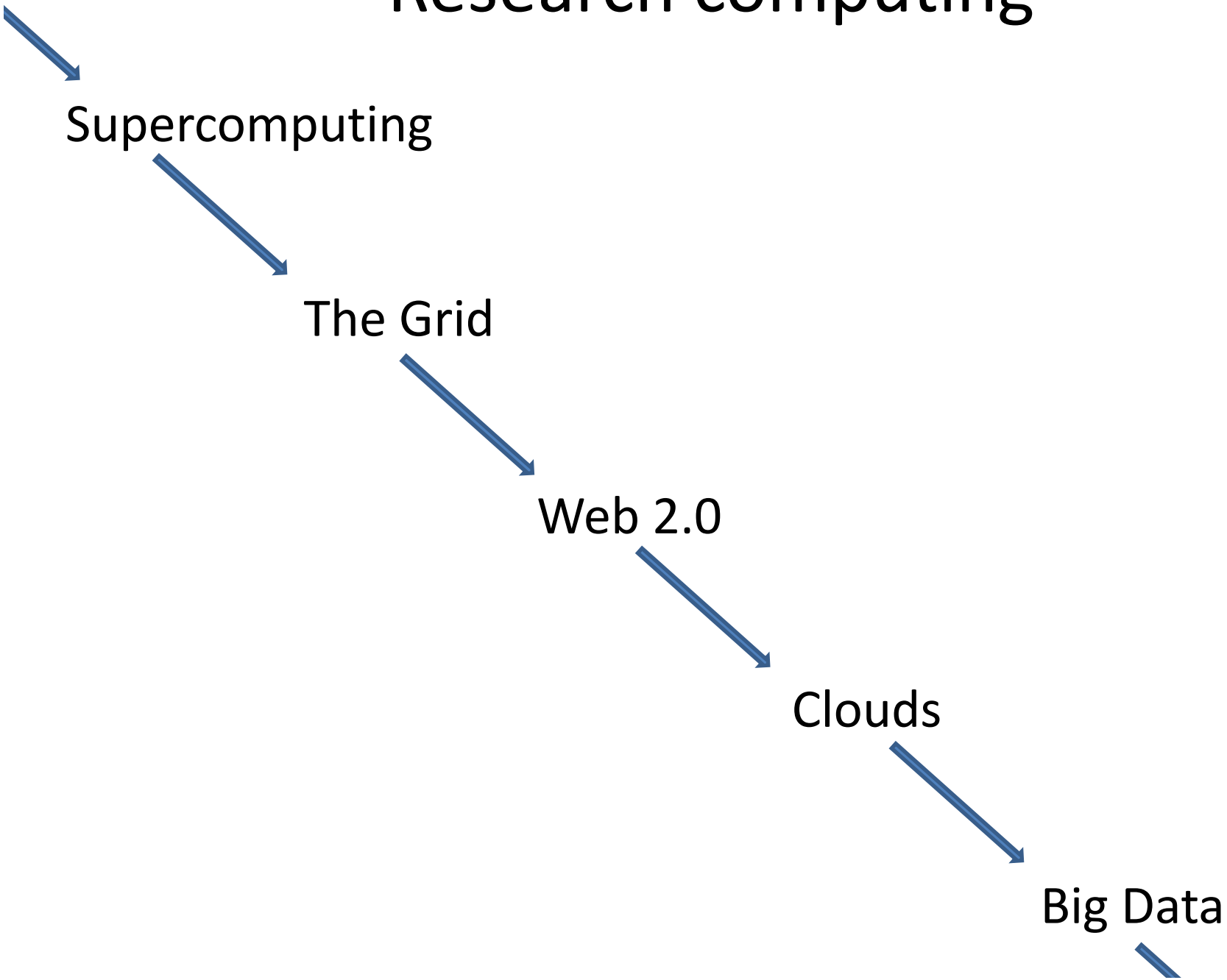


Source: http://www.flickr.com/photos/jamescridland/6134458107/

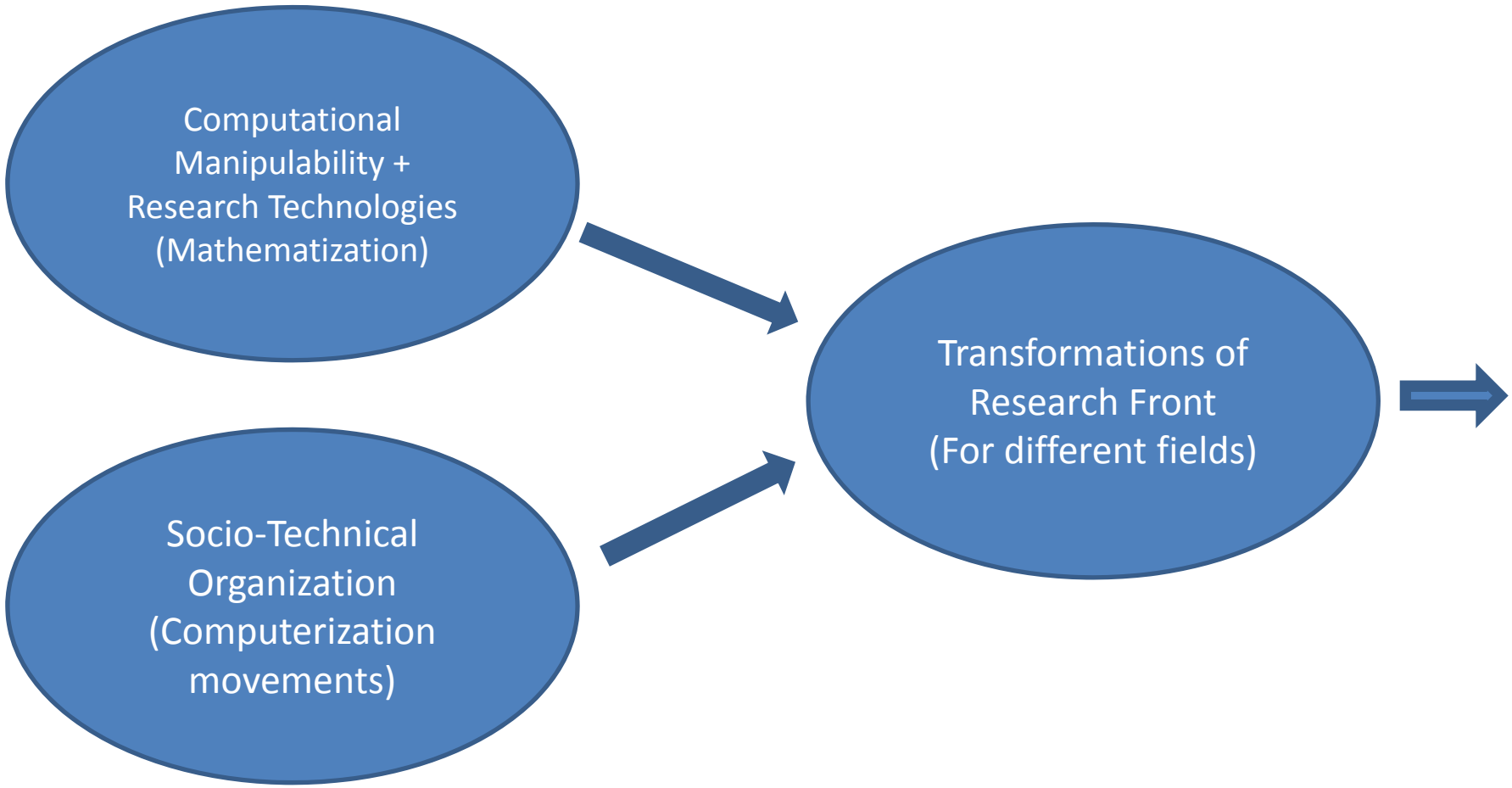
Computational Manipulability?

- 'the distinctiveness of the network of mathematical practitioners is that they focus their attention on the pure, contentless form of human communicative operations: on the gestures of marking items as equivalent and of ordering them in series, and on the higher-order operations which reflexively investigate the combinations of such operations'
- 'mathematical rapid-discovery science…the lineage of techniques for manipulating formal symbols representing classes of communicative operations'

Research computing



Digital transformations of research



Case 1: Search engine behaviour

Waller's analysis of Australian Google Users

Key findings:

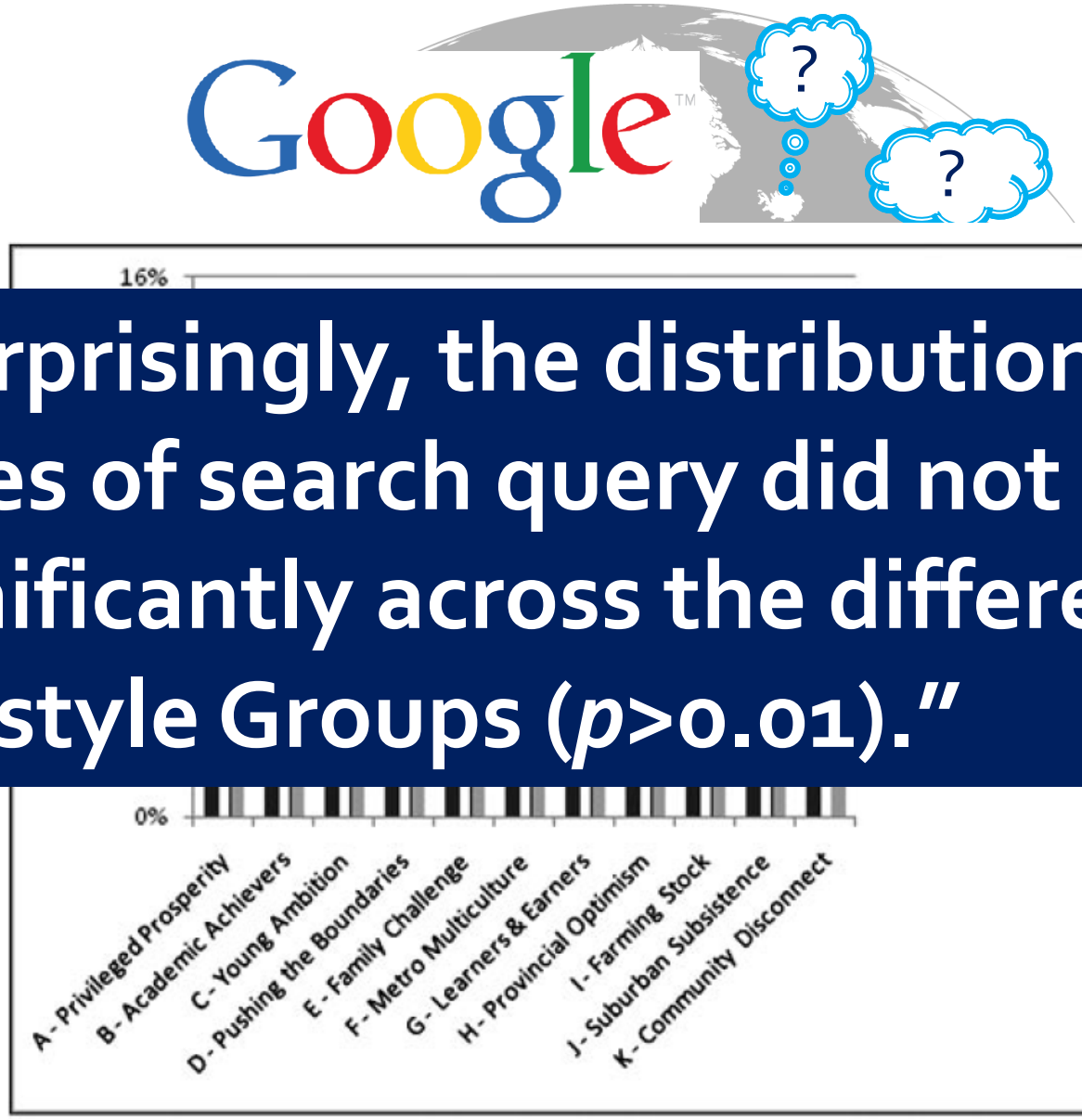
- Mainly leisure
- > 2% contemporary issues
- No perceptible 'class' differences

Novel advance:

- Unprecedented insight into what people search for

Challenge:

- Replicability
- Securing access to commercial data



"Surprisingly, the distribution of types of search query did not vary significantly across the different Lifestyle Groups (p>0.01)."

FIG. 2. Lifestyle profile of visitors to Google, compared to their representation in the Australian population (Data source: Hitwise).

Source: Waller, V. (2011). "Not Just Information:Who Searches for What on the Search Engine Google?" *Journal of the American Society for Information Science & Technology* 62(4): 761-775.

Case 2: Large-scale text analysis

Michel et al. 'culturomic' analysis of 5 Million Digitized Google Books and Heuser & Le-Khac of 2779 19th Century British Novels

Key findings:

- Patterns of key terms
- Industrialization tied to shift from abstract to concrete words

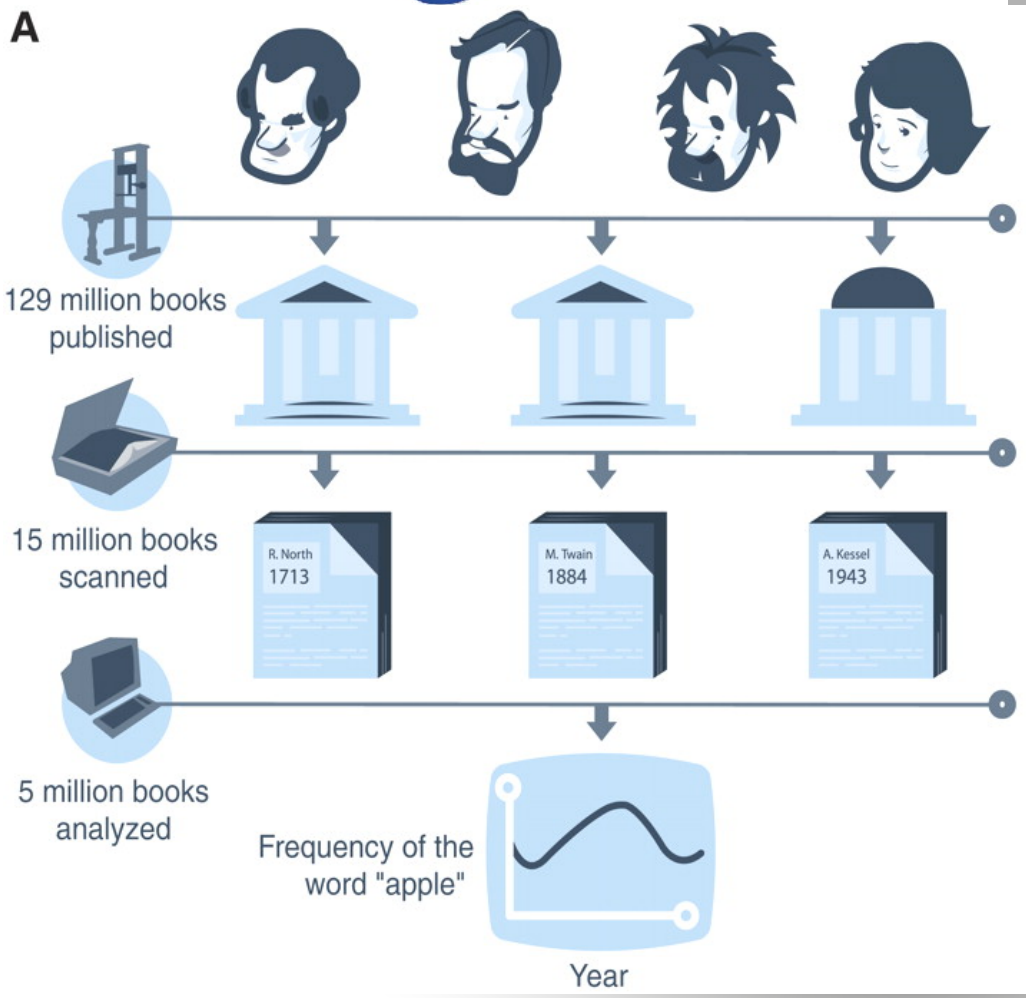
Novel advance:

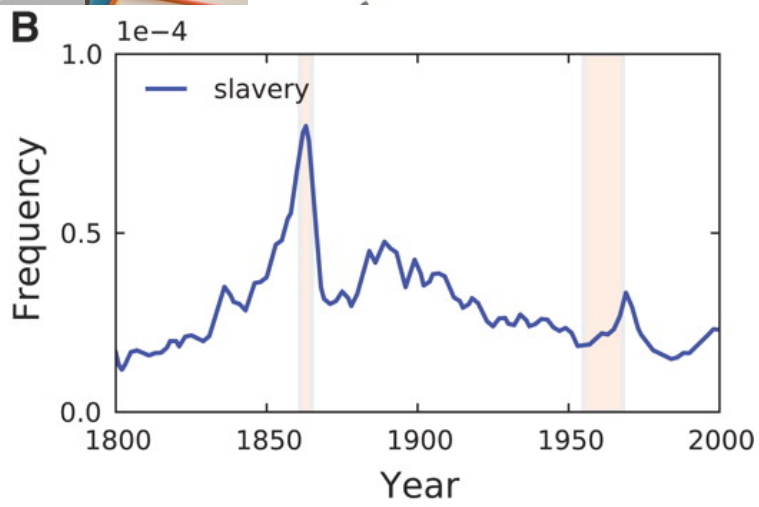
- Replicability, extension to other areas, systematic analysis of cultural materials

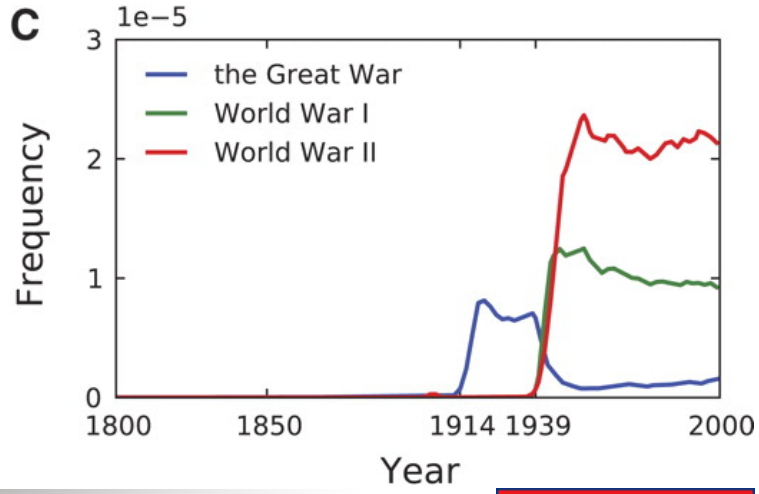
Challenge:

- Data quality

Google books







J Michel et al. Science 2011;331:176-182

Science
AAAS

Platform	Paper	Size of Data in relation to phenomenon investigated	Theoretical question/practical aim	Key findings
Facebook	Backstrom et al. (2012)	69 billion friendship links between 721 million Facebook users	Re-examine Milgram's 'six degrees of separation' online	Four degrees of separation on Facebook
	Ugander et al. (2012)	54 million invitation emails to Facebook users	How does structure of contacts affect invitation acceptance?	Not number of contacts, but number of distinct contexts, matters for acceptance
	Bond et al. (2012)	600000 Facebook users	Facebook experiment about how to mobilize voters	Voters can be mobilized via Facebook friends more than via informational messages
Twitter	Kwak et al. (2010)	1.47 billion directed Twitter relations	Is Twitter a broadcast medium or a social network?	Most use is for information, not as a social network
	Cha et al. (2010)	1.7 billion tweets among 54 million users	Who influences whom?	Top influentials dominate, but some variation by topic
	Bakshy et al. (2011)	1.6 million Twitter users	Who influences whom?	'Ordinary user' influencers can sometimes be more effective than top influencers
Wikipedia	Loubser (2009)	All Wikipedia activity	How is editing organized?	Administrators can impact negatively on participation
	Yasseri, Kertesz (2012)	Editorial activity on Wikipedia, especially reverts	Understanding conflict and collaboration	Types of conflicts can be modelled
	West, Weber and Castillo (2012)	Wikipedia contributions related to Yahoo! browsing	What characterizes Wikipedia contributors' information behaviour compared to Wikipedia readers and non-readers	Wikipedia contributors are more 'information hungry', especially about their topics

Scientificity and Big Data: Pro and Con

- Pro
  - Replicability, extension to new domain
  - ‘Total’ datasets, ‘whole universe’
  - No sampling needed, data for all behaviour and over whole existence
  - Ready made manipulability
  - Powerful relation of data to object
- Con
  - Limited access to object, skills needed for manipulability
  - Not known who users are often
  - Company does not say how data gathered
  - Researcher does not ask what is of interest without ‘givenness’
  - Datasets capture limited dimensions, and about one object
  - Object in isolation, not framed for social change significance

Conclusions

Savage and Burrows?, who ask are commercial data outpacing social science?
Boyd and Crawford?, who ask if big data raise ethical and epistemological conundrums?

… No …

The connection between research technologies and the advance of knowledge

The threats and opportunities represented by unprecedented windows into people's minds and thoughts

Does this lead to more 'scientific' (i.e. cumulative) social sciences and humanities?

Implications

- For research
  - Develop theoretical frame in which to embed big data (for new media), including power/function, relation to traditional media, and role in society
- For research policy
  - Robust base for advancing research, including shared and open databases
- For society
  - Awareness of how research can generate transparency and manipulability

Additional readings and references

Bond, Robert et al. (2012). 'A 61-million-person experiment in social influence and political mobilization', *Nature* 489: 295–298.

Bruns, A. and Liang, Y.E. (2012). 'Tools and methods for capturing Twitter data during natural disasters', *First Monday*, 17 (4 – 2), http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3937/3193

Furnas, A. and Gaffney, D. (2012). 'Statistical Probability That Mitt Romney's New Twitter Followers Are Just Normal Users: 0%'. *The Atlantic*, July 31, http://www.theatlantic.com/technology/archive/2012/07/statistical-probability-that-mitt-romneys-new-twitter-followers-are-just-normal-users-0/260539/ (accessed August 31, 2012).

Giles, J. (2012). 'Making the Links: From E-mails to Social Networks, the Digital Traces left Life in the Modern World are Transforming Social Science', *Nature*, 488: 448-50.

Kwak, H. et al. (2010). 'What is Twitter, a Social Network or a News Media?' *Proceedings of the 19th International World Wide Web (WWW) Conference*, April 26-30, 2010, Raleigh NC.

Manyika, J. et al. (2011). 'Big data: the next frontier for innovation, competition and productivity', McKinsey Global Institute, available at: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation (last accessed August 29, 2012).

Silver, Nate. (2012). *The Signal and the Noise: The Art and Science of Prediction*. London: Allen Lane.

Tancer, B. (2009). *Click: What Millions of People are Doing Online and Why It Matters*. New York: Harper Collins, 2009.

Wu, S. , J.M. Hofman, W.A. Mason, and D.J. Watts, (2011). 'Who says what to whom on twitter', *Proceedings of the 20th international conference on World Wide Web*. (on Duncan Watts webpage, http://research.microsoft.com/en-us/people/duncan/, last accessed August 29, 2012).

Oxford Internet Institute

Ralph Schroeder
ralph.schroeder@oii.ox.ac.uk
http://www.oii.ox.ac.uk/people/?id=26

See http://www.oii.ox.ac.uk/research/projects/?id=98

With support from:

ALFRED P. SLOAN FOUNDATION