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?

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
Accessing and Using Big Data to Advance Social Science Knowledge

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

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:

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

Supercomputing

The Grid

Web 2.0

Clouds

Big Data
Digital transformations of research

Computational Manipulability + Research Technologies (Mathematization)

Socio-Technical Organization (Computerization movements)

Transformations of Research Front (For different fields)
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 ($p > 0.01$).”

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


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 Logo]