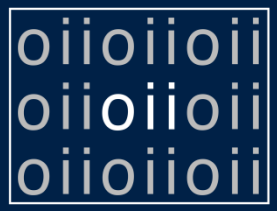


Oxford Internet Institute
University of Oxford

Big Data, Big Science and Beyond

Ralph Schroeder

Collaborators:
Eric T. Meyer, Linnet Taylor, Josh Cowls, Greg Taylor, Monica Bulger

Science 2.0, Hamburg, 27.3.2014

Overview

- Projects
- Questions
- Issues
- Definition
- How knowledge advances
- Examples
- Big Data Issues in Research and Beyond
- Policy Implications
- Conclusion

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

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

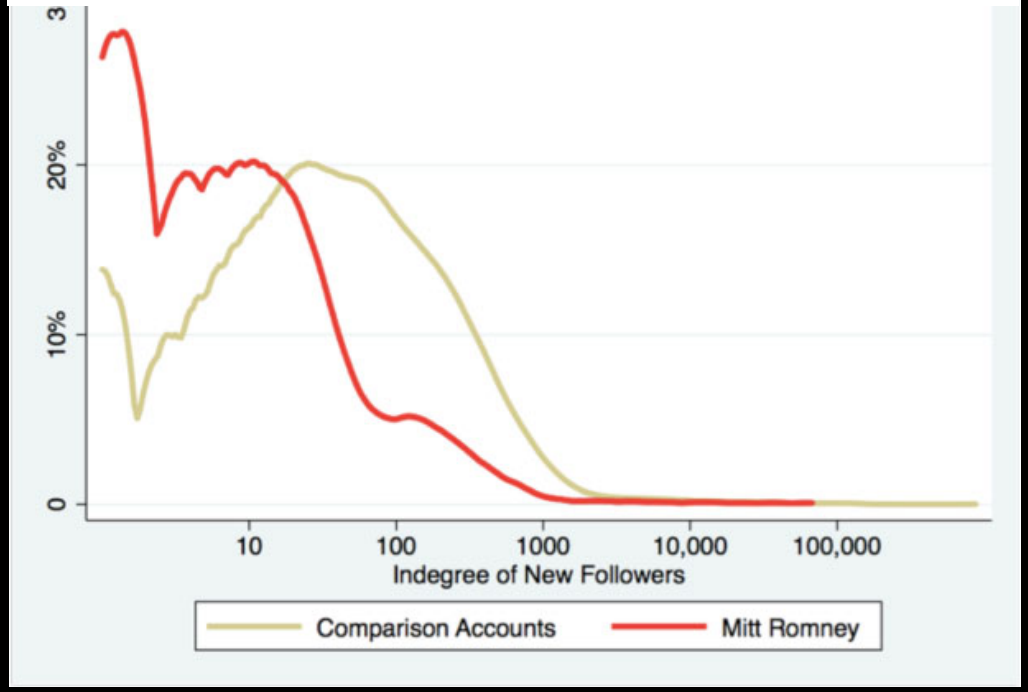
Twitter-bots

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).

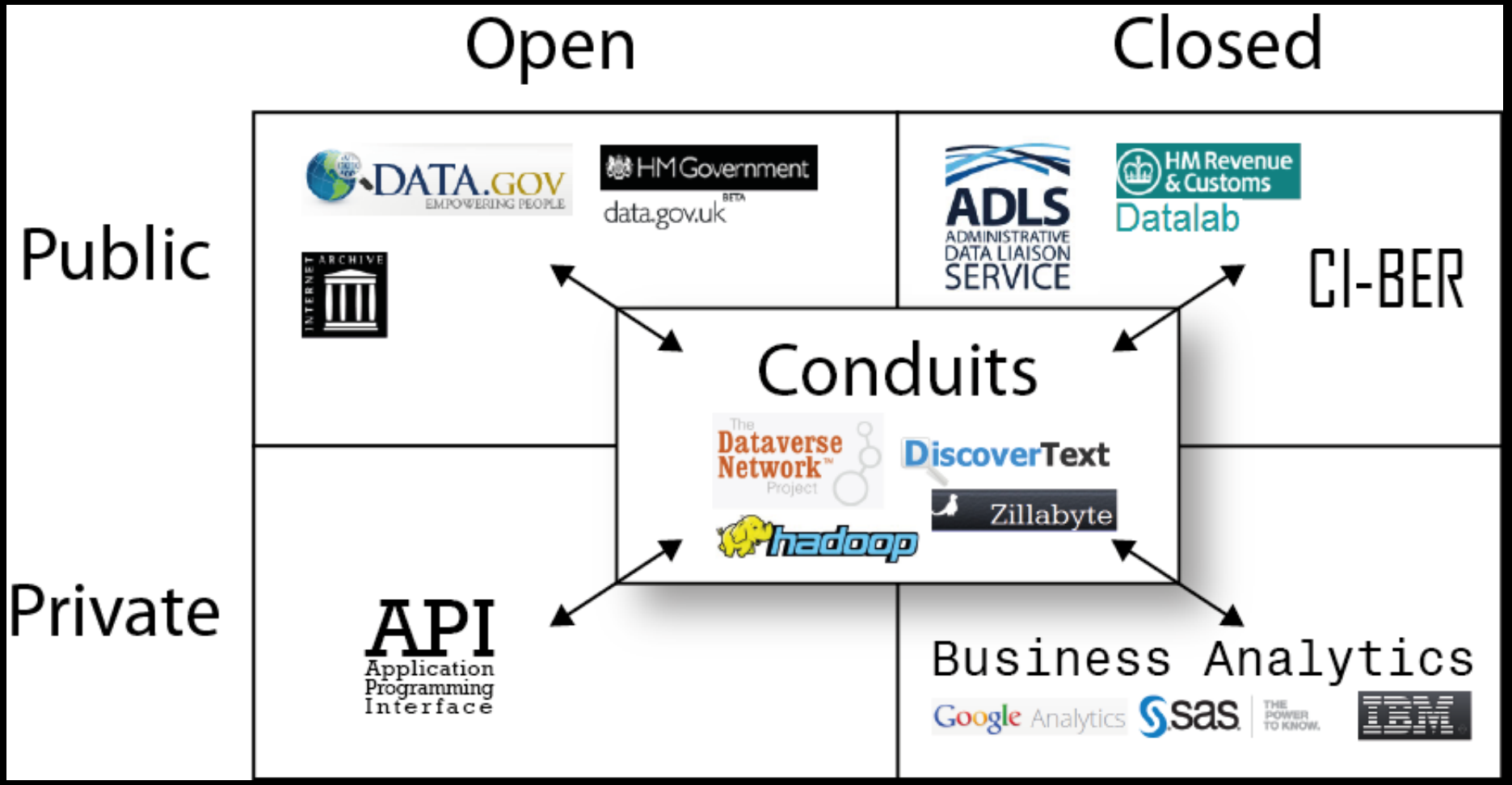
Accessing and Using Big Data to Advance Social Science Knowledge

- Funded by Sloan Foundation
- Data sources
  - 100+ interviews, mainly with social scientists
  - Reports, workshops
  - Publications, conferences
  - No representative sample, but some patterns of disciplinary and skills background and career trajectory

Big Data

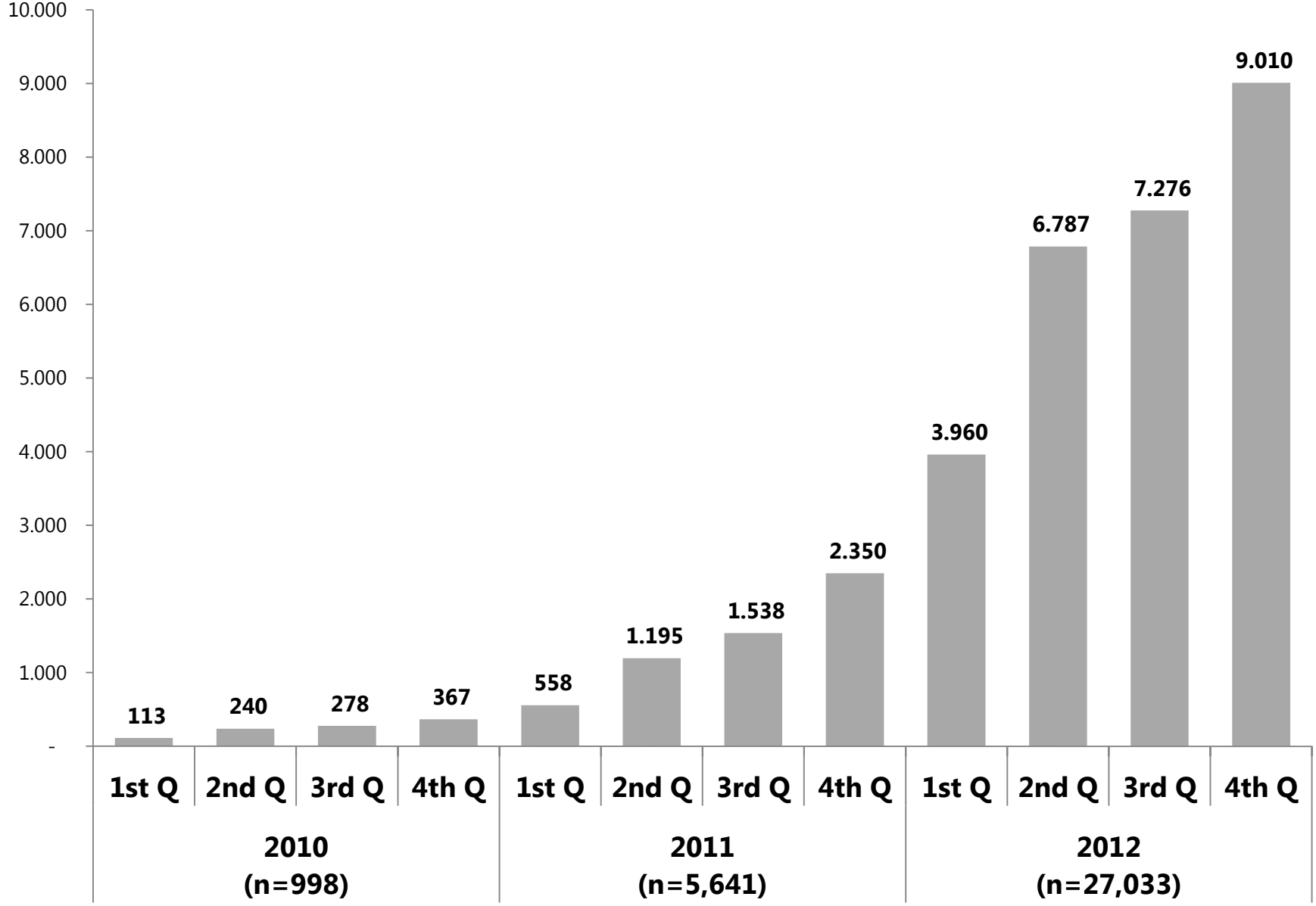
ALFRED P. SLOAN FOUNDATION

Accessing and Using Big Data to Advance Social Science Knowledge



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

Number of News Articles on Big Data

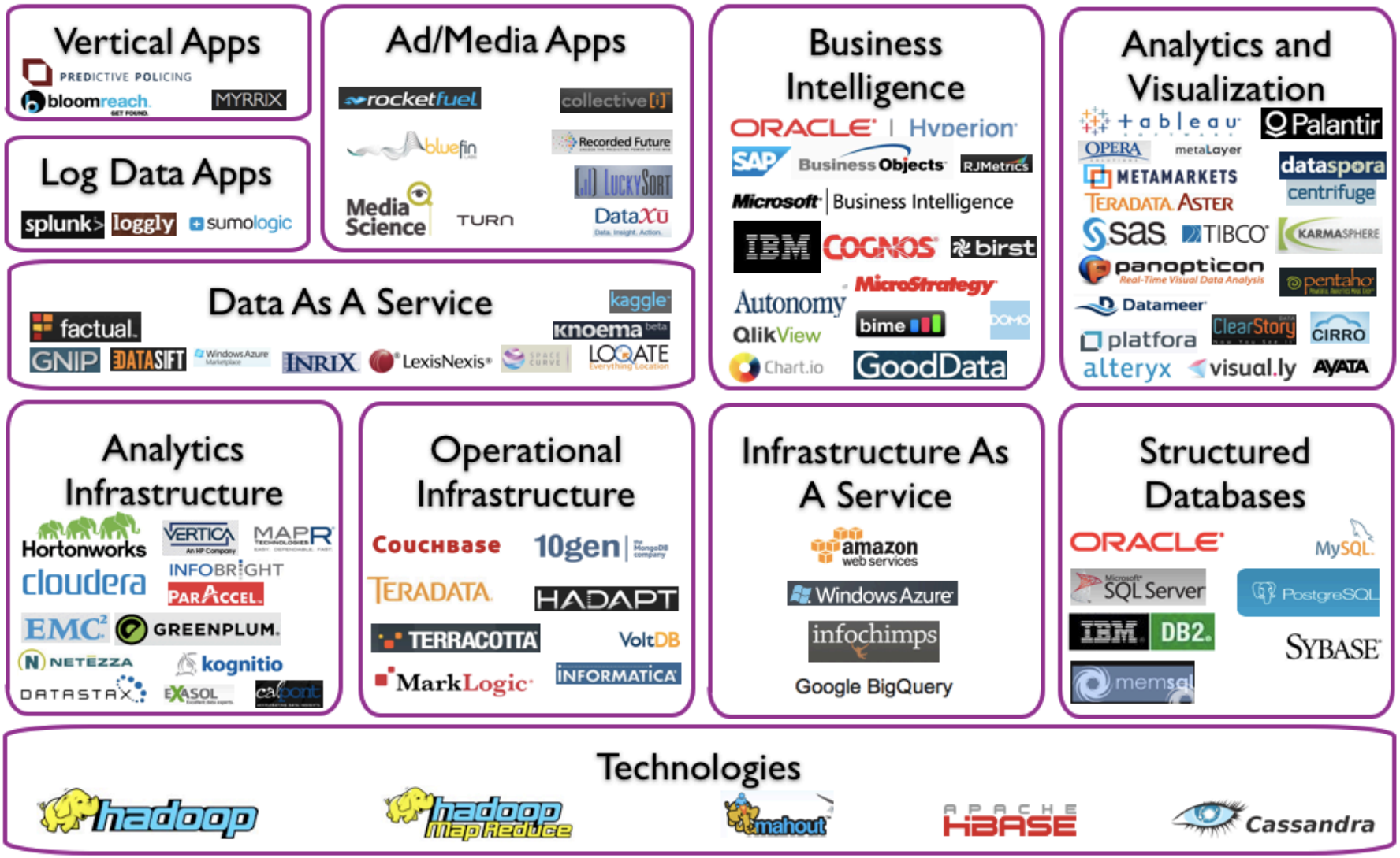


Source: Nexis data compiled by Meyer & Schroeder

Data-driven economic models: challenges and opportunities of big data

- Funded by Research Councils UK (RCUK), New Economic Models in the Digital Economy (NEMODE) network
- Data Sources:
  - 25+ interviews
  - Case studies
  - Issues include how models relate to national contexts (ie. privacy laws in Germany), where skills are located (plus gaps), use of public/private data, standardization

Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

Source: http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/

Big data in the commercial world

- Commercial uses are: ‘in house’, ‘outsourced own data’, ‘data analysis as a consultancy service’
- Careers in data analysis entail as a baseline computer science/statistical expertise, plus different domains of ‘sorting people’ and being able to ‘manipulate’ them (ie. predict their behaviour)

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

322 comments, 169 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into

Target has got you in its aim

Source: Hill, K. (Feb 16, 2012). *Forbes.com*. Available at: http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Based on Duhigg, C. (Feb 16, 2012). "How Companies Learn Your Secrets." *New York Times Magazine.*

Definition

- ‘Big data’
  - the advance of knowledge via a leap in the scale and scope in relation to a given object or phenomenon

‘Data’

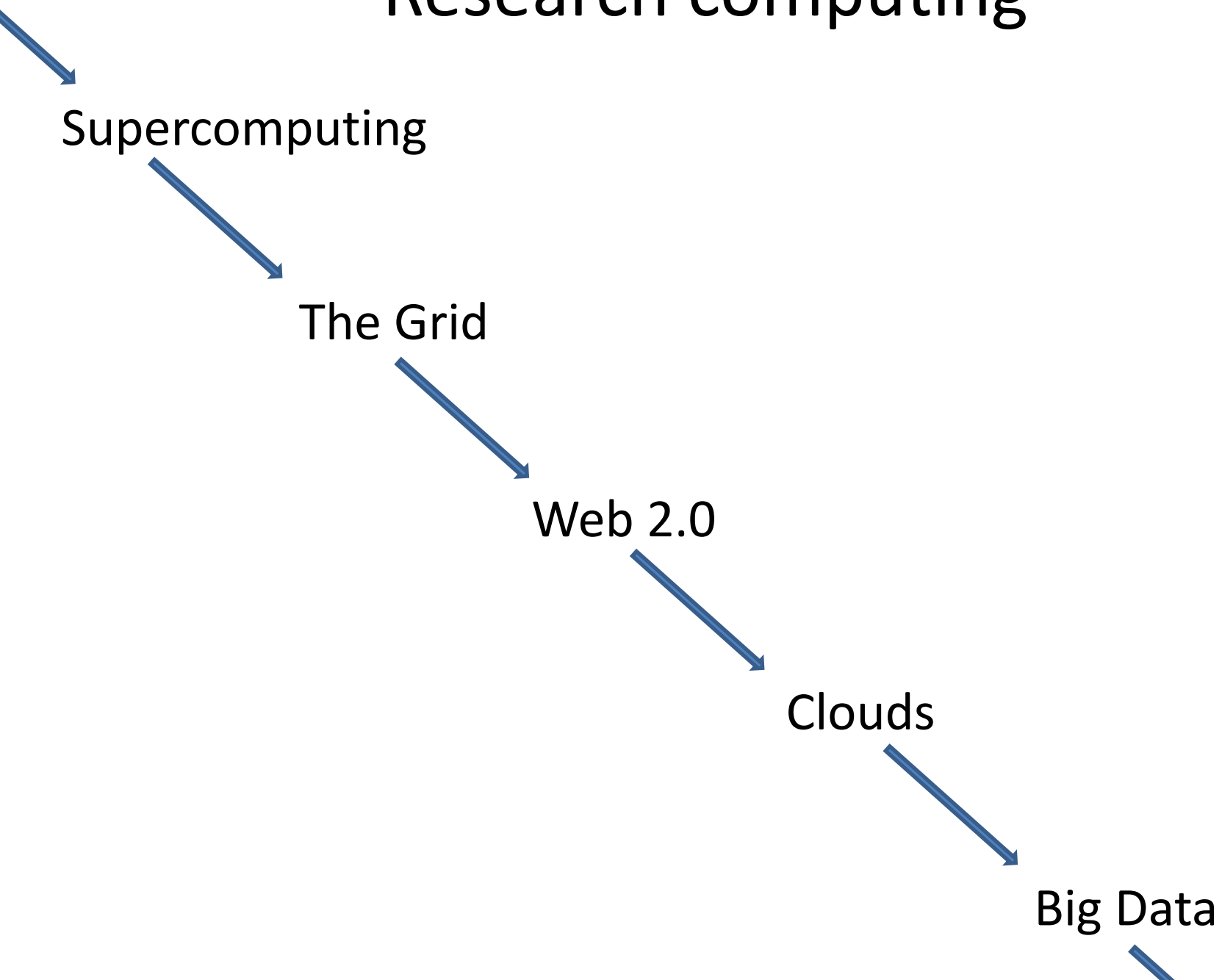
- Belongs to the object
- ‘taking…before interpreting’ (Ian Hacking)
  - the view that ‘all data are of their nature interpreted’ is misleading: ‘data are made, but as a good first approximation, the making and taking come before interpreting’
- The most atomizable useful unit of analysis

Computational Manipulability?

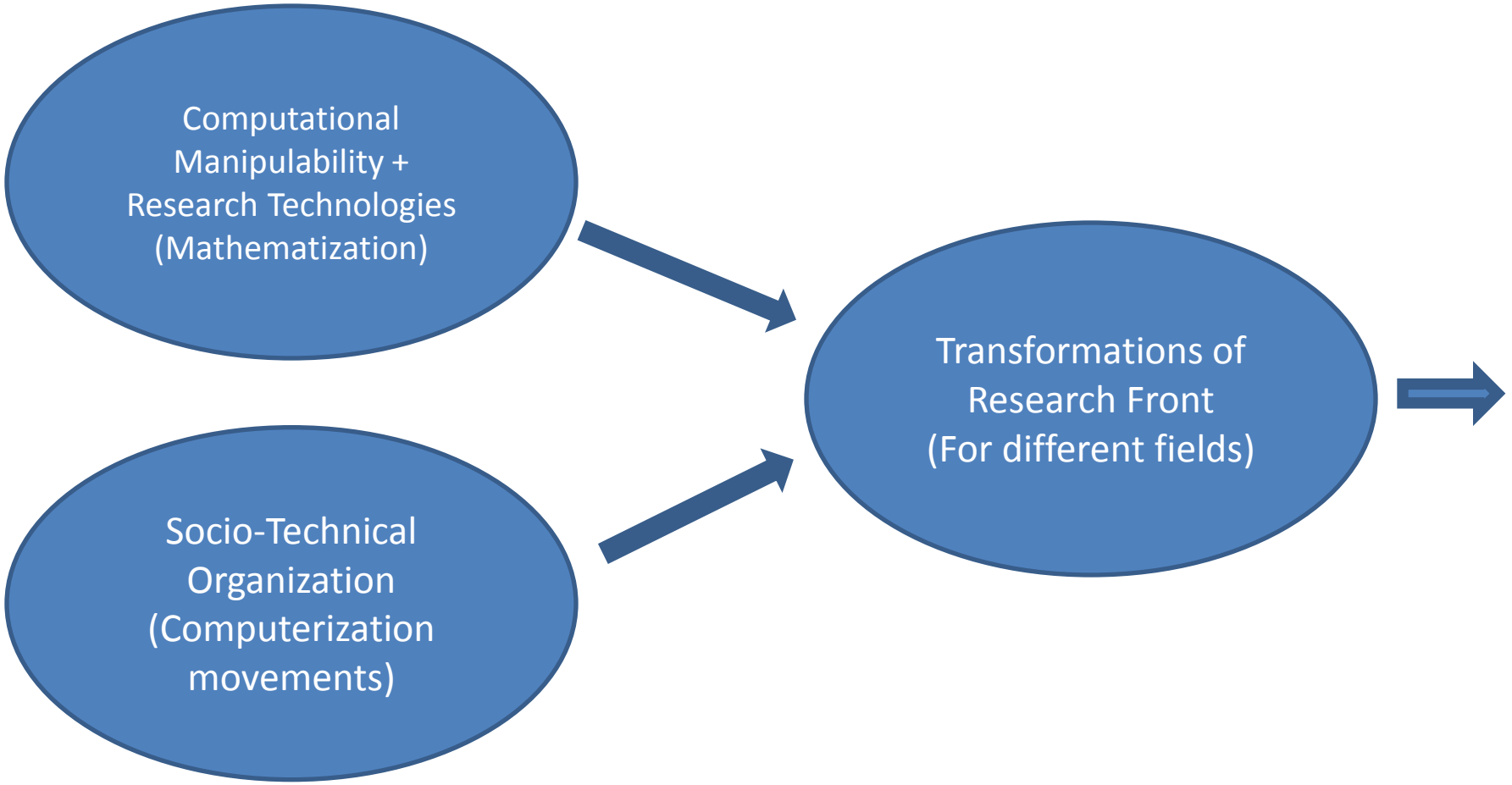
- ‘the distinctiveness of the network of mathematical practitioners is that they focus their attention on the pure, contentless form of human communicative operations: on the gestures of marking items as equivalent and of ordering them in series, and on the higher-order operations which reflexively investigate the combinations of such operations’

- ‘mathematical rapid-discovery science…the lineage of techniques for manipulating formal symbols representing classes of communicative operations’

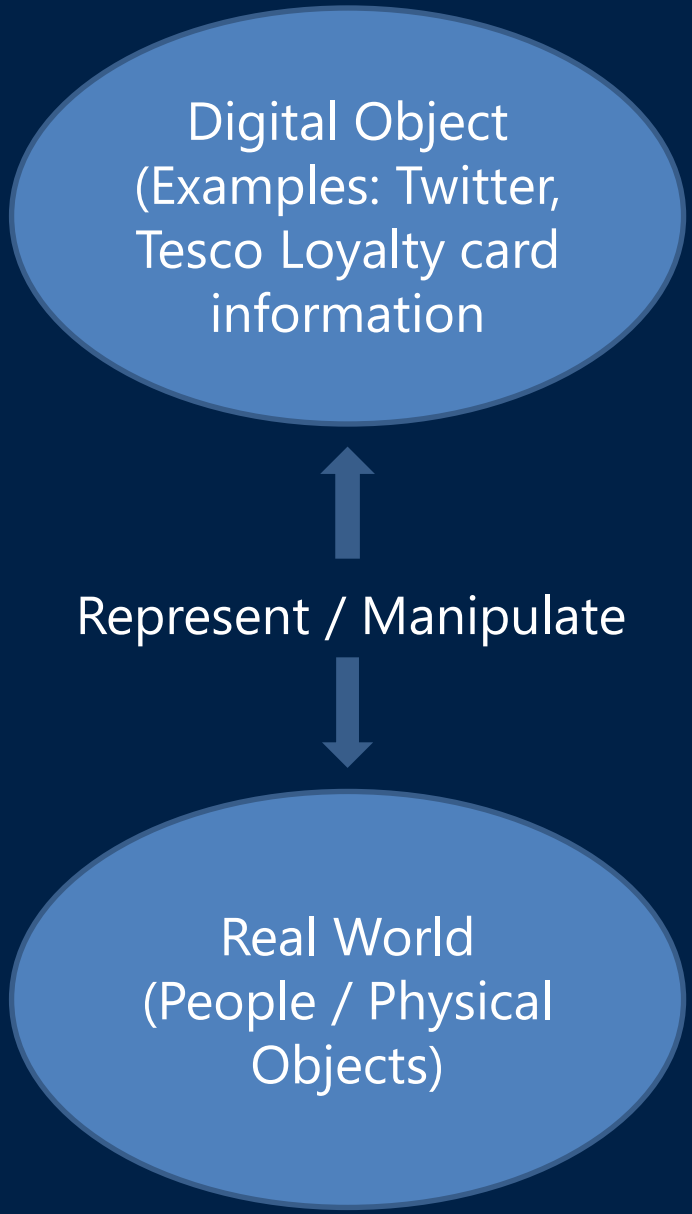
Research computing



Digital transformations of research



Digital Objects and their Referents



Uses and Limits

- Big data research uses (academic, commercial, government) are limited to the exploitation of suitable objects
- The knowledge produced is aimed at ‘sorting people’ and advancing ‘representing and intervening’ (but without ‘manipulating’, except where this is warranted by practical economic and political objectives)

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

Example 1:
Search engine behaviour

Waller's analysis of Australian Google Users

Key findings:

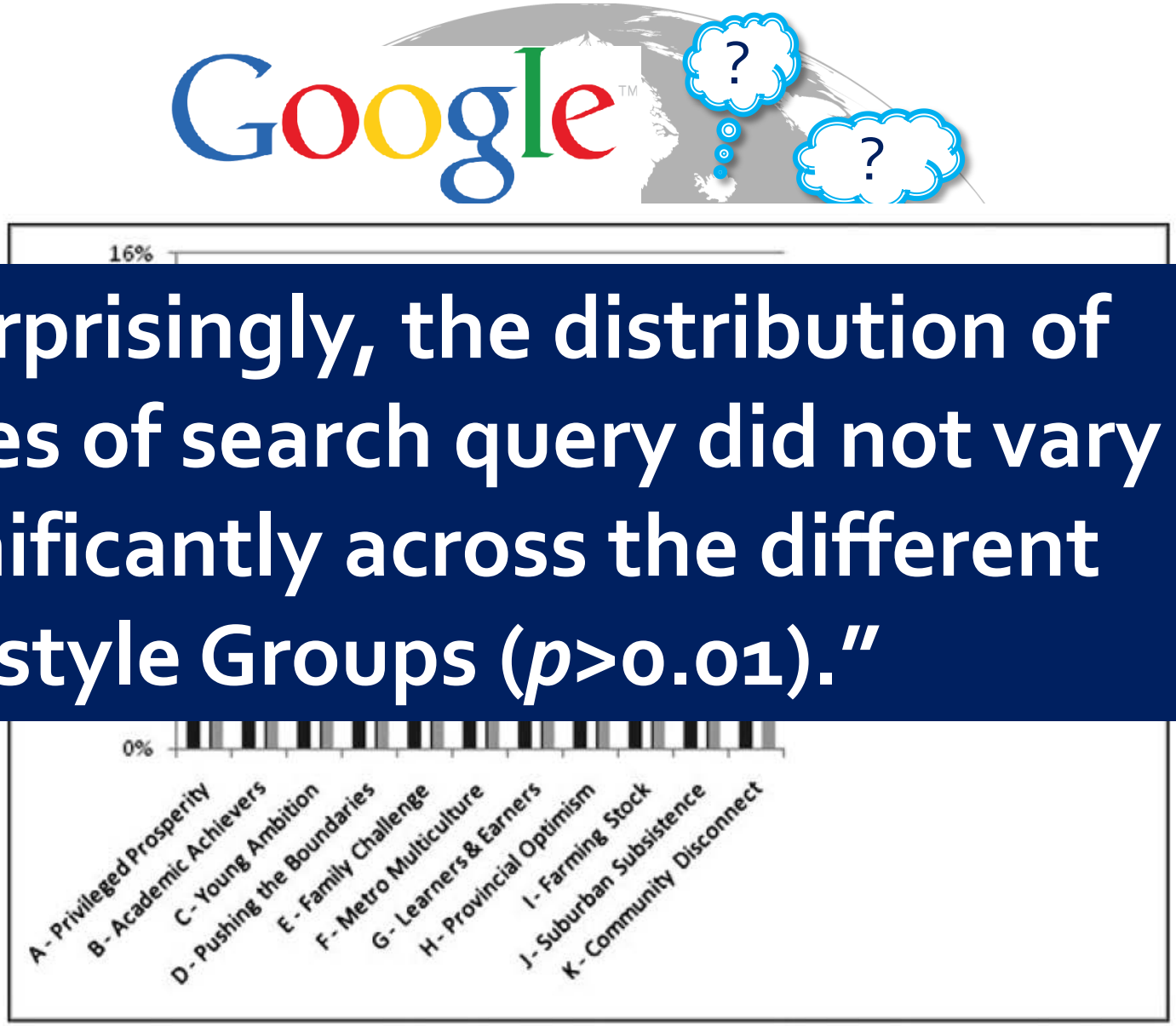
- Mainly leisure
- > 2% contemporary issues
- No perceptible 'class' differences

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.

Example 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

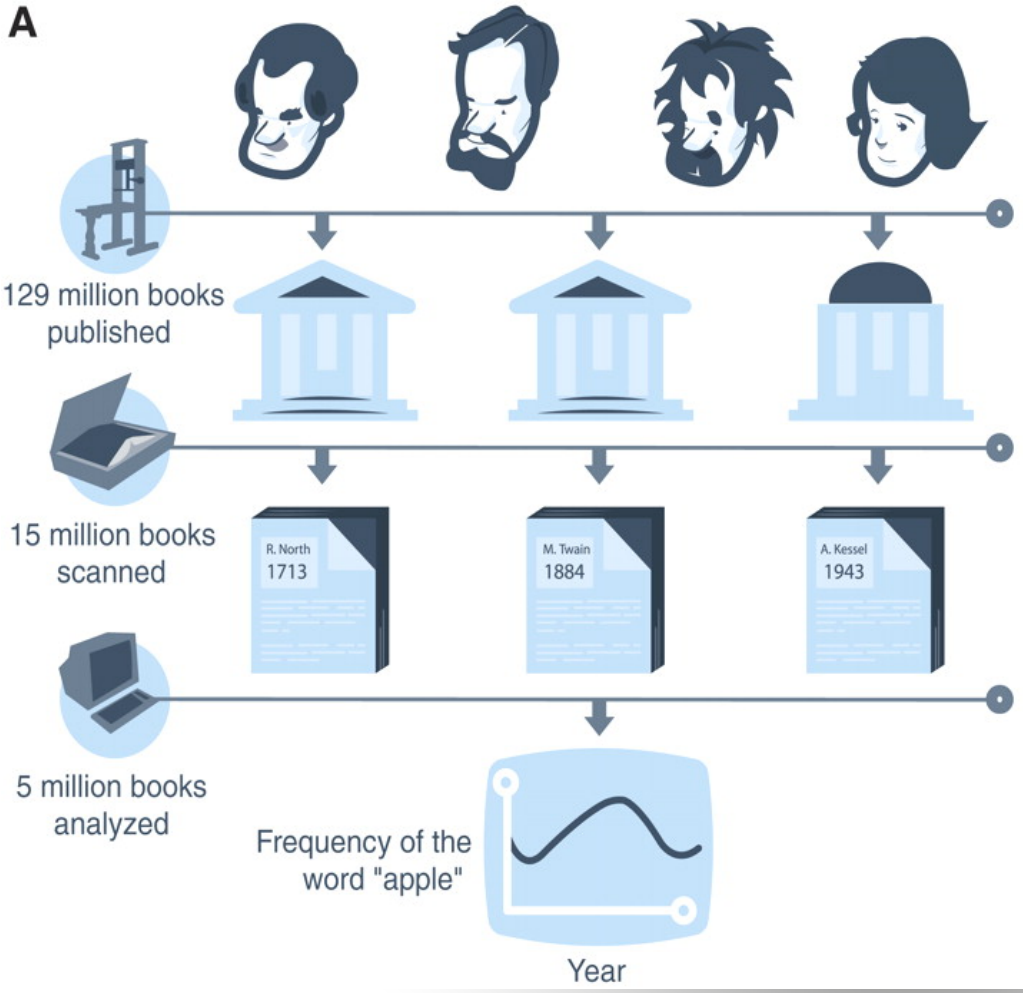
Novel advance:

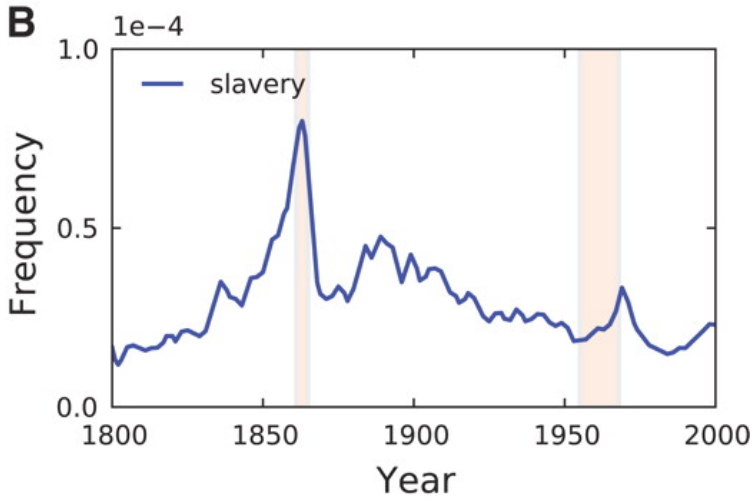
- Replicability, extension to other areas, systematic analysis of cultural materials

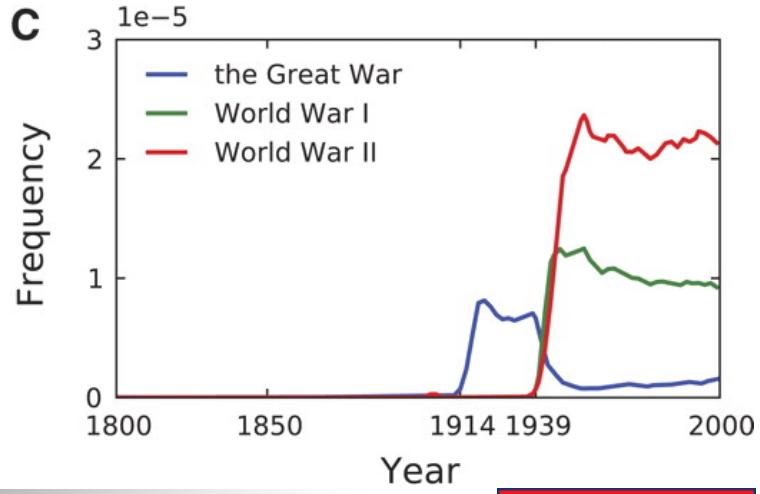
Challenge:

- Data quality

Google books







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

Science
AAAS

Example 3:
Social network or news?

Kwak et al.'s analysis of Twitter

Key findings:

- 1.47 billion social relations
- 2/3 of users are not followers or not followed by any of their followings
- Celebrities, politicians and news are among top 20 being followed

Novel advance:

- Volume of relations and topics

Challenge:

- News or social network needs to be contextualized in media ecology
- Securing access to commercial data

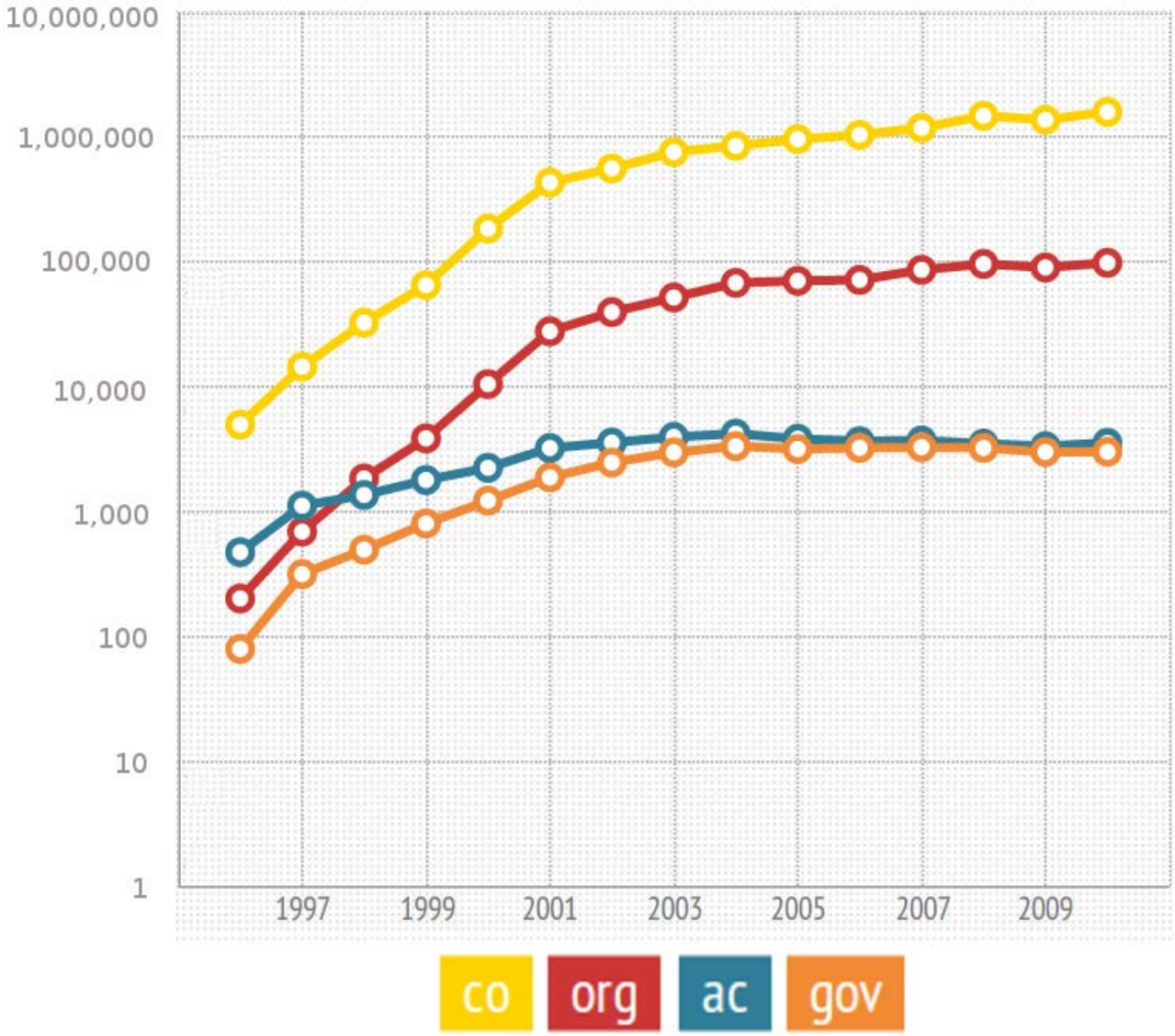
Example 4: The UK Webspace

- Data
  - Internet Archives data of .uk back to 1996
  - Annual crawls of .uk websites since 2013
  - 2.7 billion nodes, 40 TB compressed
- Features
  - Full text search (in progress, IHR)
  - Network analysis (OII)
  - N-gram analysis
- Limitations
  - Page content data access limited

BRITISH LIBRARY

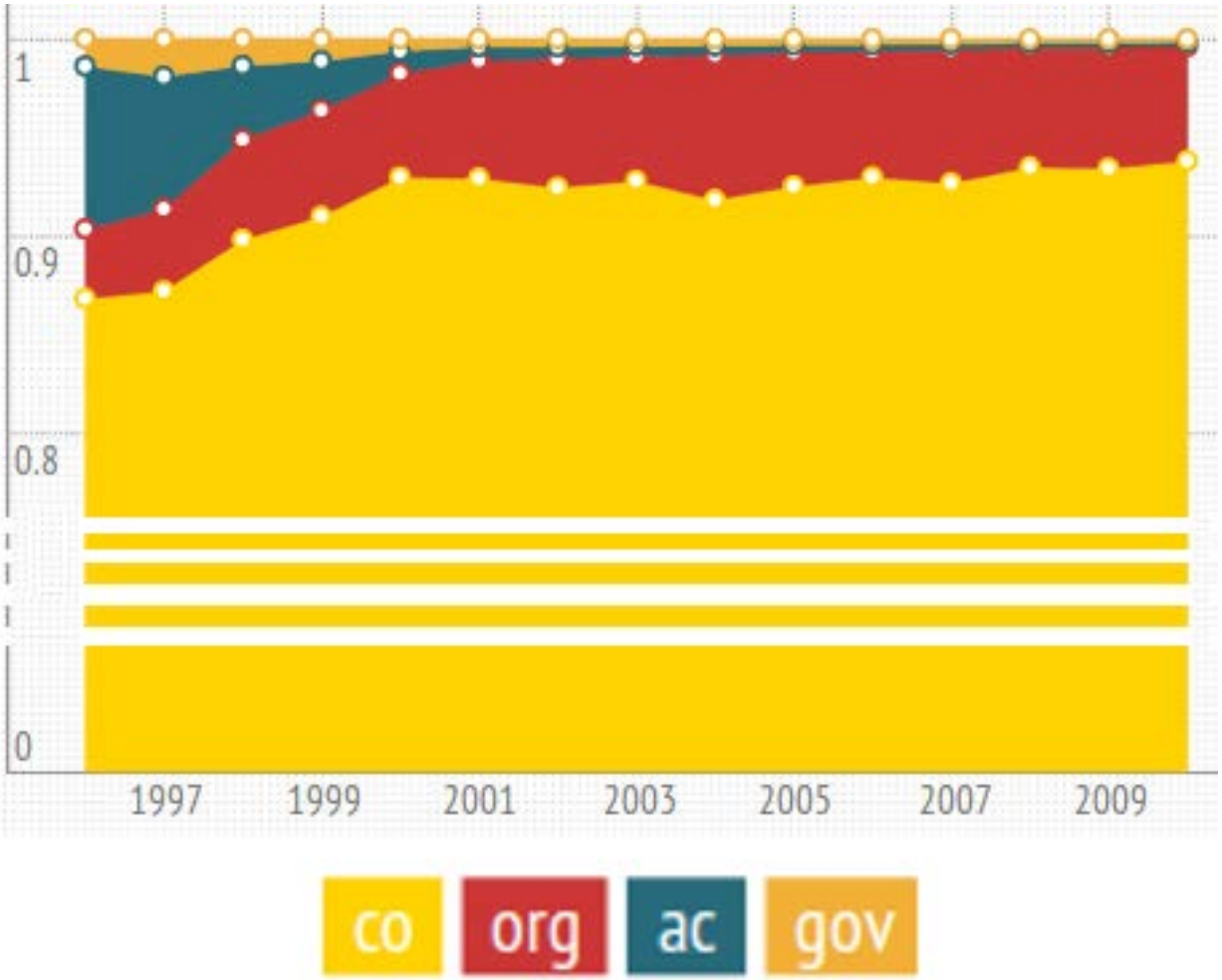
JISC

Growth of subdomains

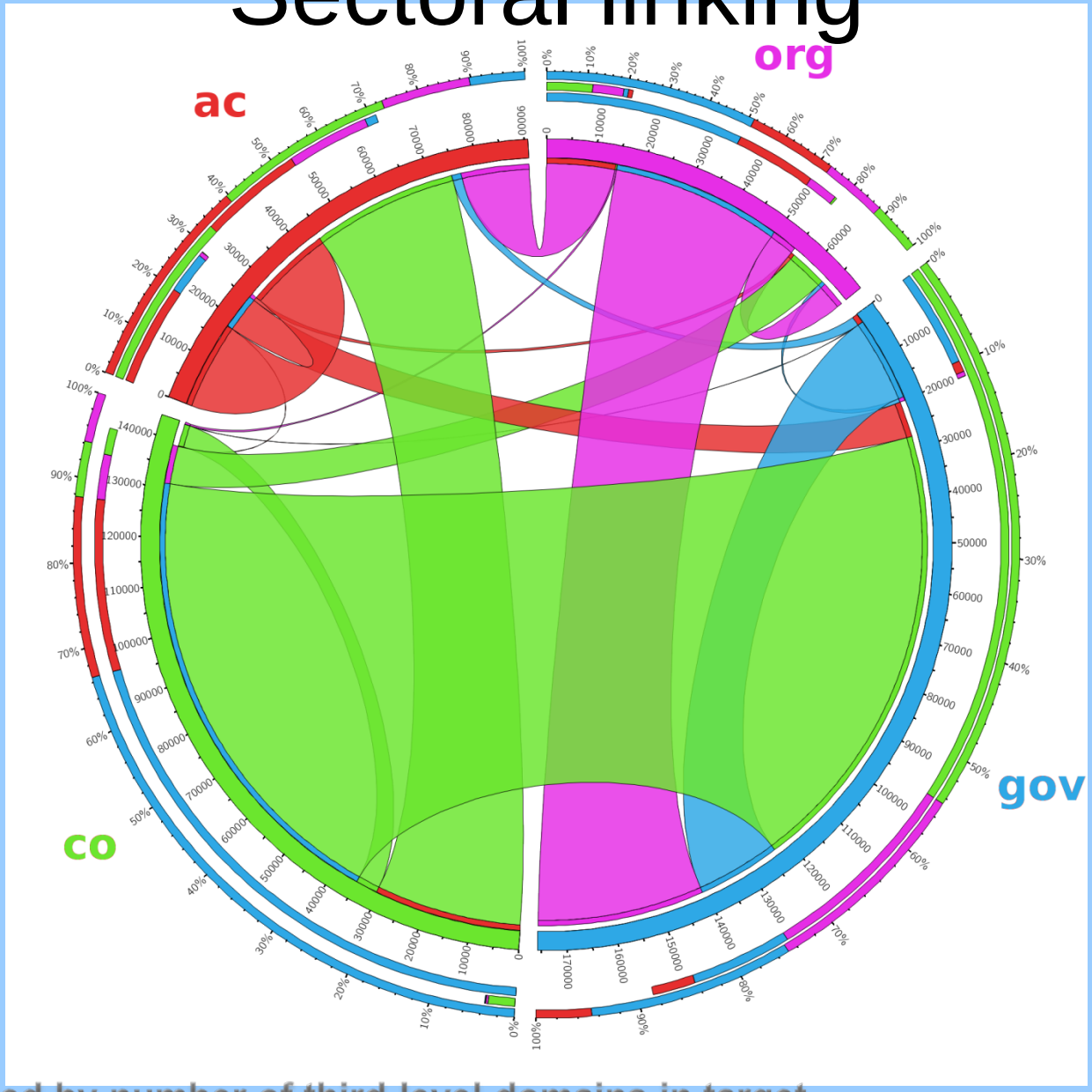


N.B. y-axis on log scale

Relative sector size on the web



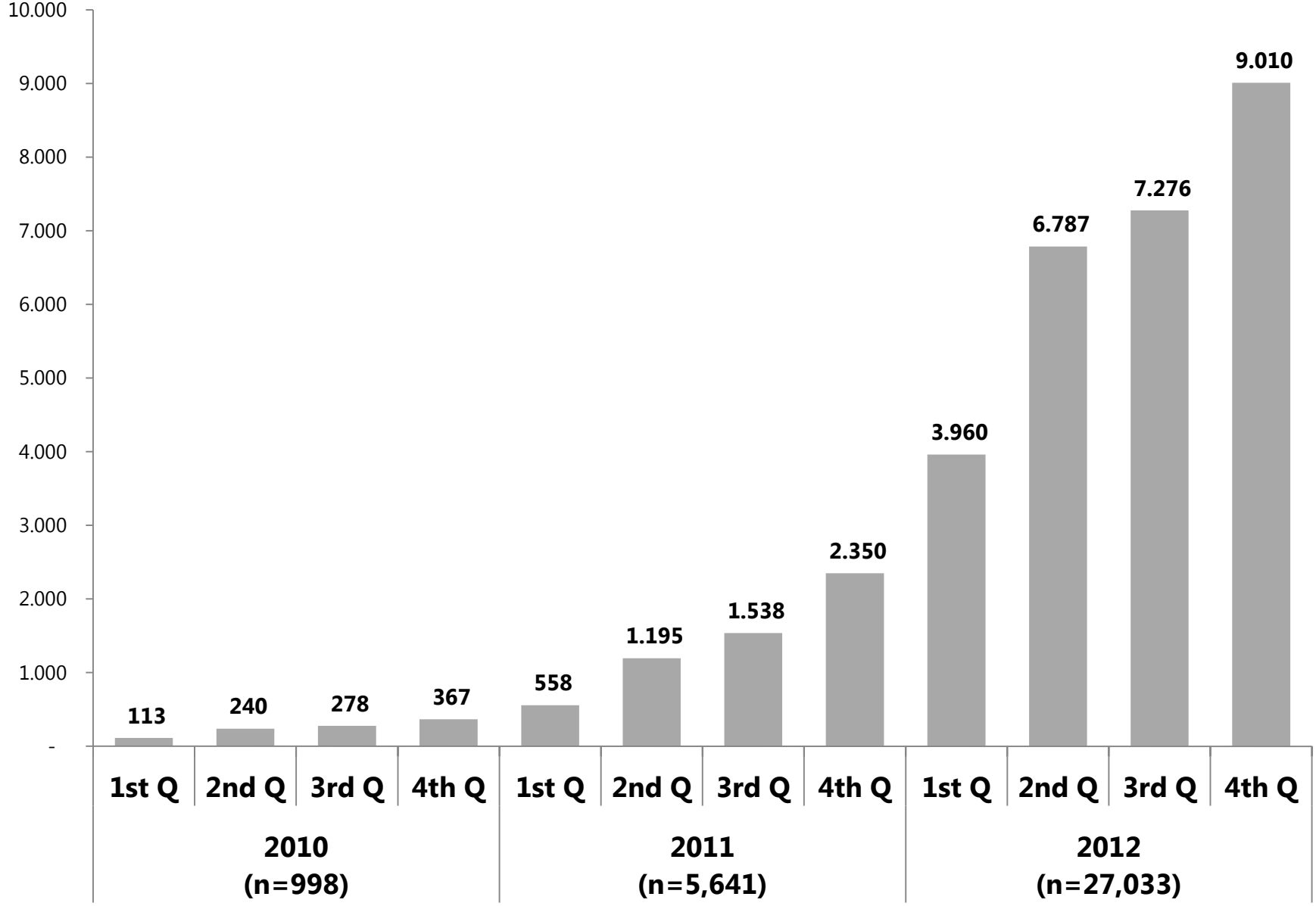
Sectoral linking



2010

Links normalized by number of third-level domains in target second-level domain

Number of News Articles on Big Data

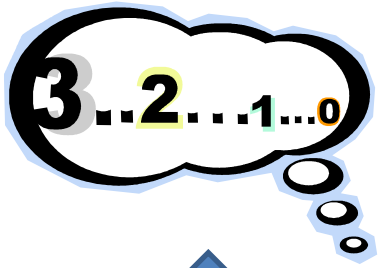


Source: Nexis data compiled by Meyer & Schroeder

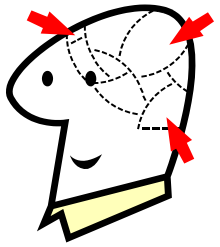
Representing

Manipulating

Limits



Digital Data



(Big) data definition enables pinpointing impacts and threats

- 'Google Plus may not be much of a competitor to Facebook as a social network, but…some analysts…say that Google understands more about people's social activity than Facebook does.'
  - New York Times, 15.2. 2014, p. A1 'The Plus in Google Plus? It's Mostly for Google'.
- Facebook Likes: 'Predicting users' individual attributes and preferences can beused to improve numerous products and services. For instance, digital systems and devices (such as online stores or cars) could be designed to adjust their behavior to best fit each user's inferred profile…online insurance…advertisements might emphasize security when facing emotionally unstable (neurotic) users but stress potential threats when dealing with emotionally stable ones'
  - 'Private traits and attributes are predictable from digital records of human behavior.' Kosinski M, Stillwell D, Graepel T.,Proc Natl Acad Sci 2013 Apr 9;110(15):5802-5.
- More powerful knowledge will enable better services, and more manipulation

Big Data is…Big Science, but only for limited domains

- Objects which 'give off' digital data are limited
- Social objects which 'give off' digital data are increasing (and data they give off is increasing), but how many are there, what phenomena do they lay bare (and which aspects of the phenomena), and how can their relations to the phenomena be theorized to advance scientific knowledge?
  - For science 2.0, how many useful identifiers are there for a given 'publication' that enables tracking its impact?

‘Big data‘ for understanding society

- Real-time transactional data (unlike survey data, traditional staple of social science)
- Outside capability of normal desktop computing environment (‘Too big to handle’)
- Big potential for understanding institutions and individual behaviour

Social Science and Big Data Research

- Dominated by social media
- Issues of 'whole universe'
  - What population, offline and online, does it represent
  - Data quality and replicability
  - How does 'modality' determine findings about implications
- How to embed the research
  - In existing theory (but also advance theory)
  - In existing ecology of media uses in society (including ones that extend existing ones)

Scientificity and Big Data: Pro and Con

- Pro
  - Replicability, extension to new domain
  - 'Total' datasets, 'whole universe'
  - (Often) 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
  - (Often) not known who users are
  - No or little knowledge of how (commercial) data were 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

Ethical and Social Issues in Big Data Research

- Objects with 'total' knowledge (universes)
  - Danger is inferring behaviour not of individuals, but of classes of people
- Asymmetry of knower and the subjects of knowledge is greater than elsewhere
- Based not on individuals' but on aggregate behaviour
  - Hence only utilitarian, not Kantian justification?
- Why does prediction or uncovering laws of behaviour 'grate'?
- Benefits: greater scientific power and more specific details
- Relation to smaller data? 'Creep'
- Solution: ethical = greater researcher and public awareness, regulatory (would apply to academic researchers?) = prevent legal and specific harms

Other positions on Big Data Implications 1

- Mayer-Schoenberger and Cukier, boyd and Crawford argue that not all information can or should be captured
  - No, need to create the legal and ethical social space which protects the individual. The solution does not rely on denying the powerfulness of knowledge, but harnessing it appropriately.
- Mayer-Schoenberger and Cukier solution of 1.more transparent algorithm, 2. Certifiying validity of algorithm 3. Allowing disprovability of prediction (p.176) –
  - Yes, but within social science, solution is to make knowledge more scientific.
- Underlying all these problems is more powerful knowledge
  - This goes against free, untrammelled behaviour
  - Solution: Society becomes more self-aware and shapes knowledge to constrain it
- Crawford, Marwick: big data is product of neoliberal capitalism? No, uses by different societies, and for purposes apart from 'neoliberal capitalist' ones, such as open government data and Wikipedia analysis

Other Positions on Big Data Implications 2

- Savage and Burrows: ask are commercial data outpacing social science?
- Boyd and Crawford: does big data raise epistemological conundrums, and isn't it always already (social) contextual ?
- Mayer-Schoenberger and Cukier: what are the political and commercial harms of wrong knowledge, especially when it changes 'everything'?

… No …

- Knowledge depends on the relation between research technologies and the advance of knowledge
- The threats and opportunities are not contextual, but depend on how more powerful knowledge is used
- Big data contributes to more 'scientific' (i.e. cumulative) social sciences, but within limits, and there are limits to commercial and political uses too

Consumer (and gov't) Big Data

- Consumer data and privacy (ie. Target pregnancy case)
  - Solution: data protection
- Consumer data and prediction and control (ie. click behaviour): affects consumer without transparency, predictive privacy harm
  - Solution: transparency, 'due process' (Crawford and Schultz)
- Consumer data – and government data - and exclusion from benefits thereof (ie. no or little use of digital devices) - if not captured by data, left out
  - Solution: Data antisubordination (Lerman)
  - Solution: government may need more data about us (and counteract the data invisibility of parts of the population)
- Consumer data from digital media (ie. search engines) – manipulate what is found without transparenyc, inappropriate personalization (Pariser)
  - Solution: transparency, consumer protection

Big Data and Policy

- Probabilistic rather than 'causal' commercial and government uses of data (ie. profiling) - only probable, not definite causal behaviour of data emitters established (Mayer-Schoenberger and Cukier)
  - Solution: more accurate knowledge
- Exposure of Data emitter because of identifiers in large-scale and linked data (Netflix, AOL, Google Streetview, National Security Administration), such that anonymization does not work
  - Solution: data protection, better anonymization, opting out, consent
- Social media used in authoritarian regimes for control (Weibo in China)
  - Solution: more commercial independence, more civil society pushback, researcher non-cooperation

Future of Big Data Research

- Difference commercial versus academic world is that knowledge provides competitive advantage as against advancing (high-consensus rapid-discovery) knowledge
- The limits in both cases are the objects (to which the data 'belong'), and that need to have available digitally manipulable data points
- How available these objects are differs
- There are many objects, for non-academics and scientists to humanities scholars (physical, human, cultural), but they are not infinite
- This availability, not skills or other issues, determines the future of big data research

The Outlook for Big Data Research

- There is an overlap between real world research and the world of academic research which is closer than elsewhere
  - because this is the research front in both
  - because they share common objects
- Research in the sciences (outside of social science) also consists of computer science/statistics plus domain expertise, and will provide more powerful knowledge for manipulating the physical world
- Ethical and social issues matter, but there is also 'creep'
- The main (social science) objects are related to digital media, whose limits are their expanding uses

Implications

- For research
  - Develop theoretical frame in which to embed big data (for social 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
- Big Brother?
  - Yes, but also Brave New World of Omniscience, with Social Science as Handmaiden

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).

Project Papers

Schroeder, Ralph (Forthcoming). 'Big Data: Towards a More Scientific Social Science and Humanities' in Mark Graham and William H Dutton (eds.), Society and the Internet: How Networks of Information are Changing our Lives. Forthcoming.

Schroeder, Ralph, & Taylor, Linnet (Forthcoming). 'Is bigger better? The emergence of big data as a tool for international development policy.' GeoJournal.

Meyer, Eric T., Schroeder, Ralph, & Taylor, Linnet (2013, August). 'Big Data in the Study of Twitter, Facebook and Wikipedia: On the Uses and Disadvantages of Scientificity for Social Research.' Paper presented at the proceedings of the Annual Meeting of the American Sociological Association.

Schroeder, Ralph, & Taylor, Linnet. 'Big Data and Wikipedia Research: Social Science Knowledge across Disciplinary Divides'. Submitted to Information, Communication and Society.

Schroeder, Ralph. 'Big Data in the Study of Twitter, Facebook and Wikipedia: On the Uses and Disadvantages of Scientificity for Social Research'. New Media and Society. Status?

Taylor, Linnet. 'No place to hide? The ethics and analytics of tracking mobility using African mobile phone data. Submitted to Population, Space and Place.

Meyer, Eric T., Schroeder, Ralph, & Taylor, Linnet. 'Big Data in the Social Sciences: Towards a New Research Paradigm?' International Journal of Communication. Revised and resubmitted.

Meyer, Eric T., Schroeder, Ralph, & Taylor, Linnet (2013, November). 'The Boundaries of Big Data.' Paper presented at SIG-SI Symposium, ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

Schroeder, Ralph and Cowls, Josh. 'Answering Questions and Questioning Answers in the Era of Big Data.' In preparation.

Taylor, Linnet, Meyer, Eric T., & Schroeder, Ralph. 'Bigger and better, or more of the same? Emerging practices and perspectives on big data analysis in economics". Submitted to Big Data & Society.

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