Practical ethics for big data research: An introduction

The webinar will begin at 3pm

- You now have a menu in the top right corner of your screen.
- The red button with a white arrow allows you to expand and contract the webinar menu, in which you can write questions/comments.
- We will answer your questions at the end.
- If we don’t get to a question, we will reply later by email.
- You will be on mute throughout – we need to do this in order to ensure a high quality recording.

Practical ethics for big data research: An introduction

Libby Bishop
UK Data Service
UK Data Archive, University of Essex

Webinar
27 October 2016

UK Data Service

"When lost, public trust cannot be easily regained."

Trust

My recommendations centre on **trust**. Building public trust for the use of health and care data means giving people confidence that their private information is kept secure and used in their interests.

Dame Fiona Caldicott, National Data Guardian

Principles of research ethics

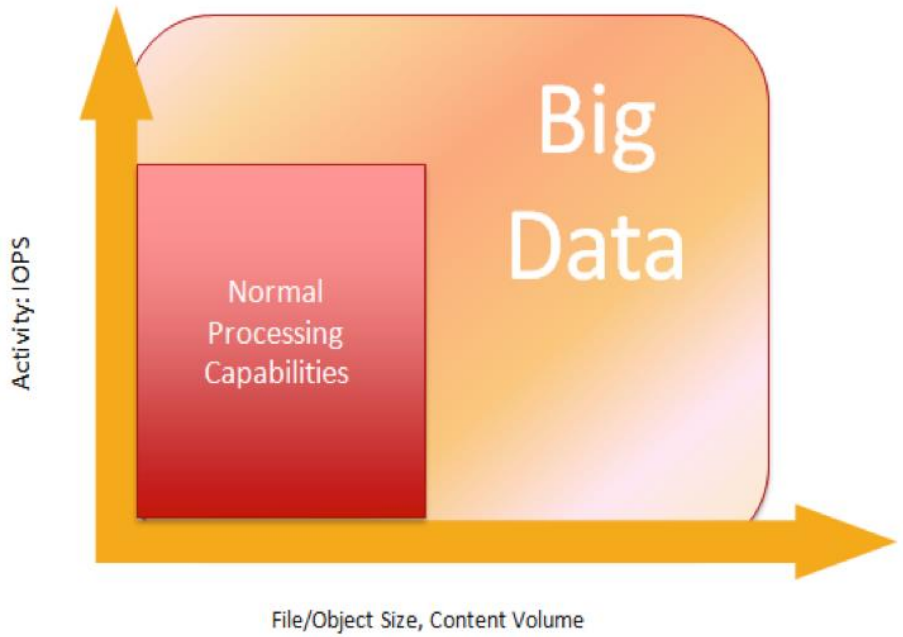
- **Respect** for persons – for autonomy as human beings, not mere means, includes privacy
- **Beneficence** – doing good; maximise benefits, and minimise harm
- **Justice** – fair distribution, of risks and benefits

Belmont Report, 1978; Common Rule, 1981 (US)

Universal Declaration of Human Rights (1948)
(right to privacy; legally binding Bill of HR in 1976)

A working definition of Big Data

Data sets that exceed the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach



New and novel data (“Big Data”)

A: Transactions of **government**, e.g., tax data

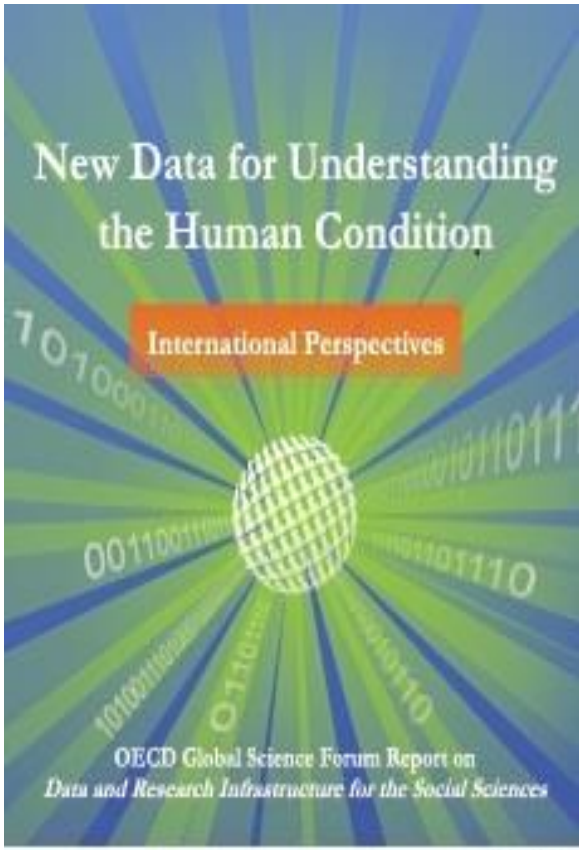
B: Official registrations or licensing requirements.

C: **Commercial** transactions made by individuals and organisations

D: **Internet data** from search and social networking activities

E: Tracking data, monitoring **movement** of individuals or objects

F: **Image** data, particularly aerial and satellite images



What is different about “big data”?

- Big data are (usually) not collected by researchers
  - No formal ethics review

- Big data were (usually) not generated for research
  - Protections used at time of collection not possible or not feasible (e.g., consent and anonymisation)

Bigness is not a problem–wildness is

Case 1 - Twitter and cyberhate speech

- The research question: does a terrorist act trigger an increase in cyberhate language in social media? (Williams and Burnap, 2015)*
- The data: N=427,330 tweets over 15 days on Lee Rigby murder in Woolwich by two British-born men.
- Collected via streaming API
  - Trending keywords, e.g., Woolwich

*Source: Williams, M. L. and Burnap, P. 2015. Cyberhate on social media in the aftermath of Woolwich: British Journal of Criminology 56(2), pp. 211-238. (10.1093/bjc/azv059)

COSMOS
Collaborative Online Social Media Observatory

Publishing aggregated data and protecting privacy

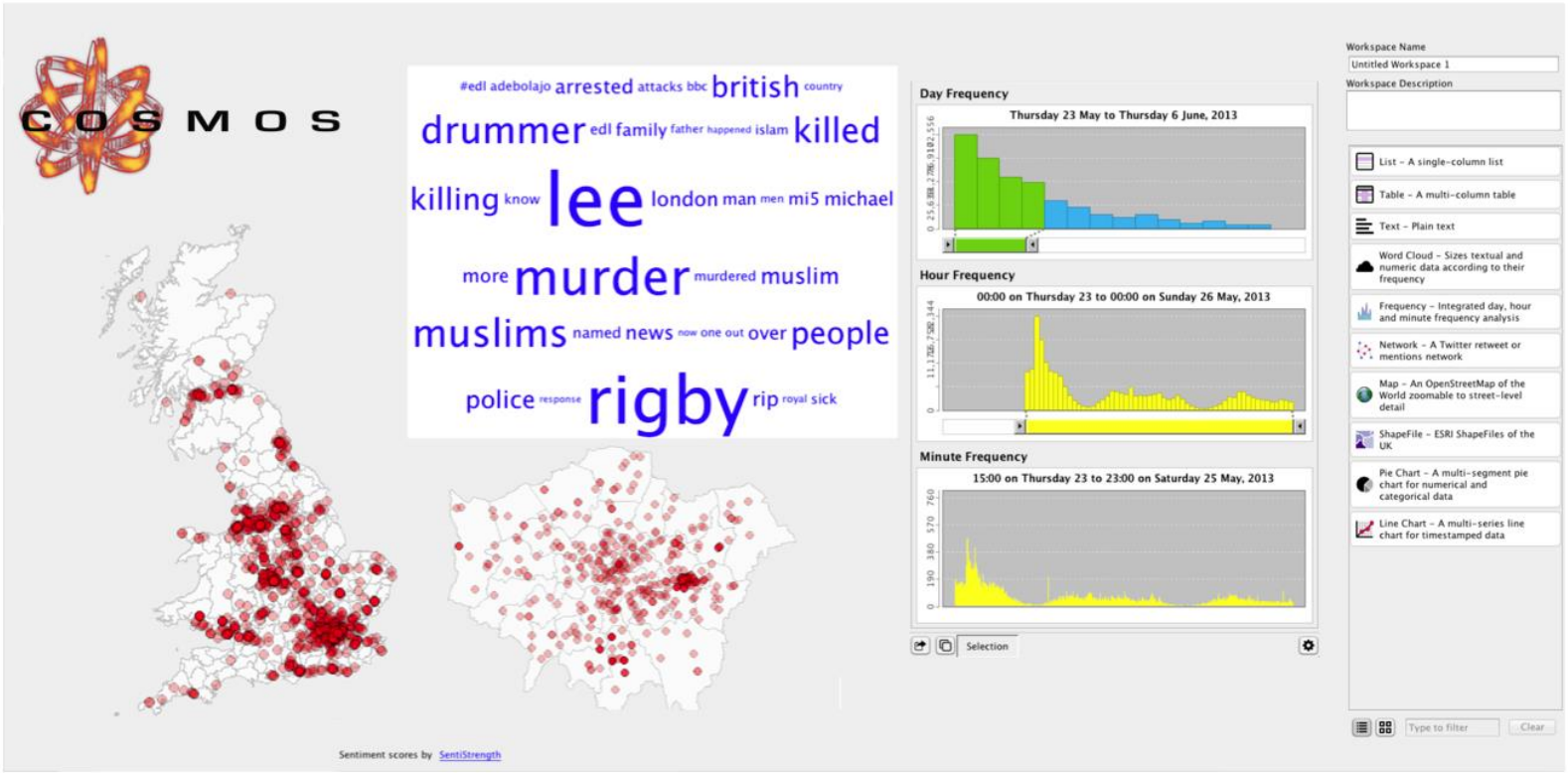
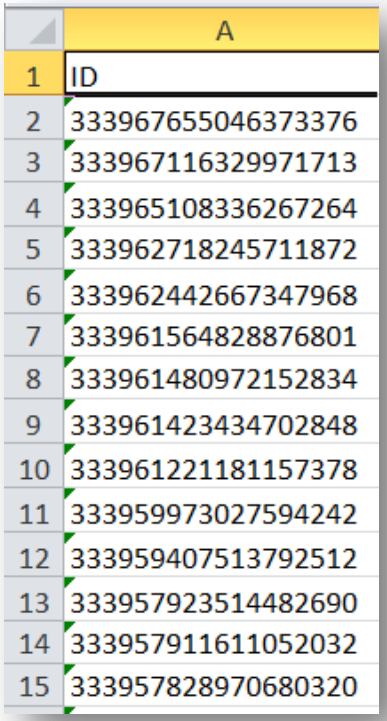


FIG 2: Woolwich Inventory Stage – COSMOS visualisation of twitter traffic during the first four days of collection (left to right: Geo-located tweets in UK & London, WordCloud of tweet content and Frequency of Tweets in the 15 day collection window)

Source: Williams, M. L. and Burnap, P. 2015.
Cyberhate on social media in the aftermath of Woolwich:
British Journal of Criminology 56(2), pp. 211-238. (10.1093/bjc/azv059)

Publishing Twitter research – still some issues



	A
1	ID
2	333967655046373376
3	333967116329971713
4	333965108336267264
5	333962718245711872
6	333962442667347968
7	333961564828876801
8	333961480972152834
9	333961423434702848
10	333961221181157378
11	333959973027594242
12	333959407513792512
13	333957923514482690
14	333957911611052032
15	333957828970680320

- Original and “re-hydrated” tweet datasets almost certainly will not match
- The ways they don’t match cannot be known with certainty
- Availability of tweets depends on Twitter’s discretion
- Does not meet highest standards for transparency, replication

Weller, K. and Kinder-Kurlanda, K. **A Manifesto for Data Sharing in Social Media Research**
http://dl.acm.org/citation.cfm?id=2908172&CFID=686568339&CFTOKEN=73278098

Data Availability: The data is subject to Twitter terms and conditions which prevent us from redistributing any data collected from the 1% API beyond the research team. However, we are able to supply two limited datasets as Supporting Information files: For age: user ID and derived age in years. For occupation: user ID, derived occupation, SOC2010 code and NS-SEC group. Whilst we cannot provide the text from which we have derived our proxy demographics, according to the T&Cs of Twitter we are able to supply user IDs so that others can retrieve the same data as us from the Twitter API.
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115545

Sloan, L. et al. 2015 Who Tweets?

Poll

- Do you think it is acceptable to publish the content of public tweets without anonymising them?

Publishing tweet content – more challenges

- Constraints from Twitter's Terms and Conditions
  - Tweets may not be altered, even for anonymisation
  - Re-publishers subject to requests for deletion
    - Not possible if paper with quotes has been published or data archived
- What about publishing unanonymised tweets?
  - Tweeters consent to limited 3rd party reuse in T&C
  - "Already public" … remember OK Cupid?

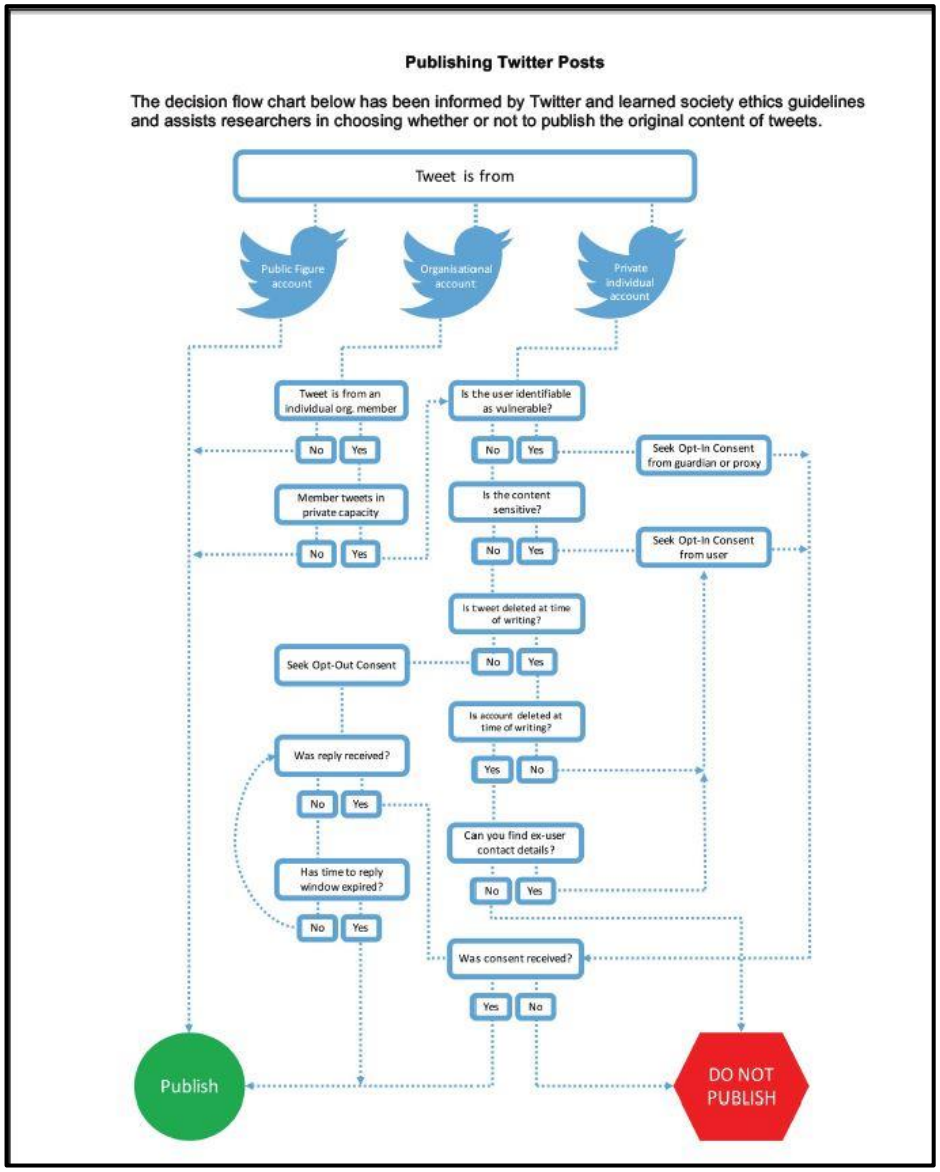
Publishing tweets – using principles

- COSMOS/SDSL research showed users objected to unconsented reuse, even research, if identifiable
  - Respect = going beyond minimal legal obligations

- For small scale project, consent was feasible
  - Contacted tweeters directly by tweet to request consent w/o anonymity
  - Full information was provided via a website

- Researchers have duty of beneficence to participants
  - Riskiness of content (hate speech) mandated a precautionary approach

Publishing Twitter research – solutions!



http://socialdatalab.net/wp-content/uploads/2016/08/EthicsSM-SRA-Workshop.pdf

UK Data Service

Case 2 - Genomes, anonymity, linkage

"Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially."

(US ExOffofPres p.xi in Tech Watch Review)

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Gymrek, et al. Science, 18 January, v339, 2013.

Debate - effectiveness of anonymisation

No silver bullet: De-identification still doesn't work

Arvind Narayanan
arvindn@cs.princeton.edu

Edward W. Felten
felten@cs.princeton.edu

July 9

Dispelling the Myths Surrounding De-identification:

Anonymization Remains a Strong Tool for Protecting Privacy

nn Cavoukian, Ph.D.
tion & Privacy Commissioner, Ontario, Canada

Khaled El Emam, Ph.D.
Canada Research Chair in Electronic Health Information, CHEO Research Institute and University of Ottawa

June 2011

Big Data and Innovation, Setting the Record Straight: De-identification *Does* Work

33 Bits of Entropy

The End of Anonymous Data and what to do about it

HOME | ABOUT 33 BITS | SITEMAP | ARVIND NARAYANAN

One more re-identification demonstration, and then I'm out

Ann Cavoukian, Ph.D.
Information and Privacy Commissioner
Ontario, Canada

Daniel Castro
Senior Analyst, Information Technology and Innovation foundation

June 16, 2014

Why de-identification is a key solution for sharing data responsibly

Khaled El Emam (University of Ottawa, CHEO Research Institute & Privacy Analytics Inc.)

Luk Arbuckle (CHEO Research Institute, Privacy Analytics Inc.)

Anonymisation – where does it stand?

- Anonymisation as magic pill no longer defensible
  - Linkage
  - Risks increase with dimensionality of data

- BUT, anonymisation remains a vital tool – as part of “risk mitigation strategy”
  - ICO and others defend it as effective (ICO Big Data par. 42).

Case 3–Facebook's contagion research

- Facebook altered feeds of c.700,000 users
- Reducing positive inputs resulted in users making fewer positive posts, and more negative ones (same when negative posts reduced)
- Significant public and research outcry
  - Manipulation
  - Consent – not adequately informed
- "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." This statement has since been confirmed by Cornell University. (Editorial Expression of Concern)

Kramer, Adam D., Jamie E. Guillory, and Jeffrey T. Hancock"Experimental evidence of massive- scale emotional contagion through social networks," 24, June 17, 2014, of Proc Natl Acad Sci USA (111:8788–8790)

Poll

- Would you feel comfortable using data from Facebook if that data had been approved for research by Facebook's Institutional Review Board?

Facebook – summary

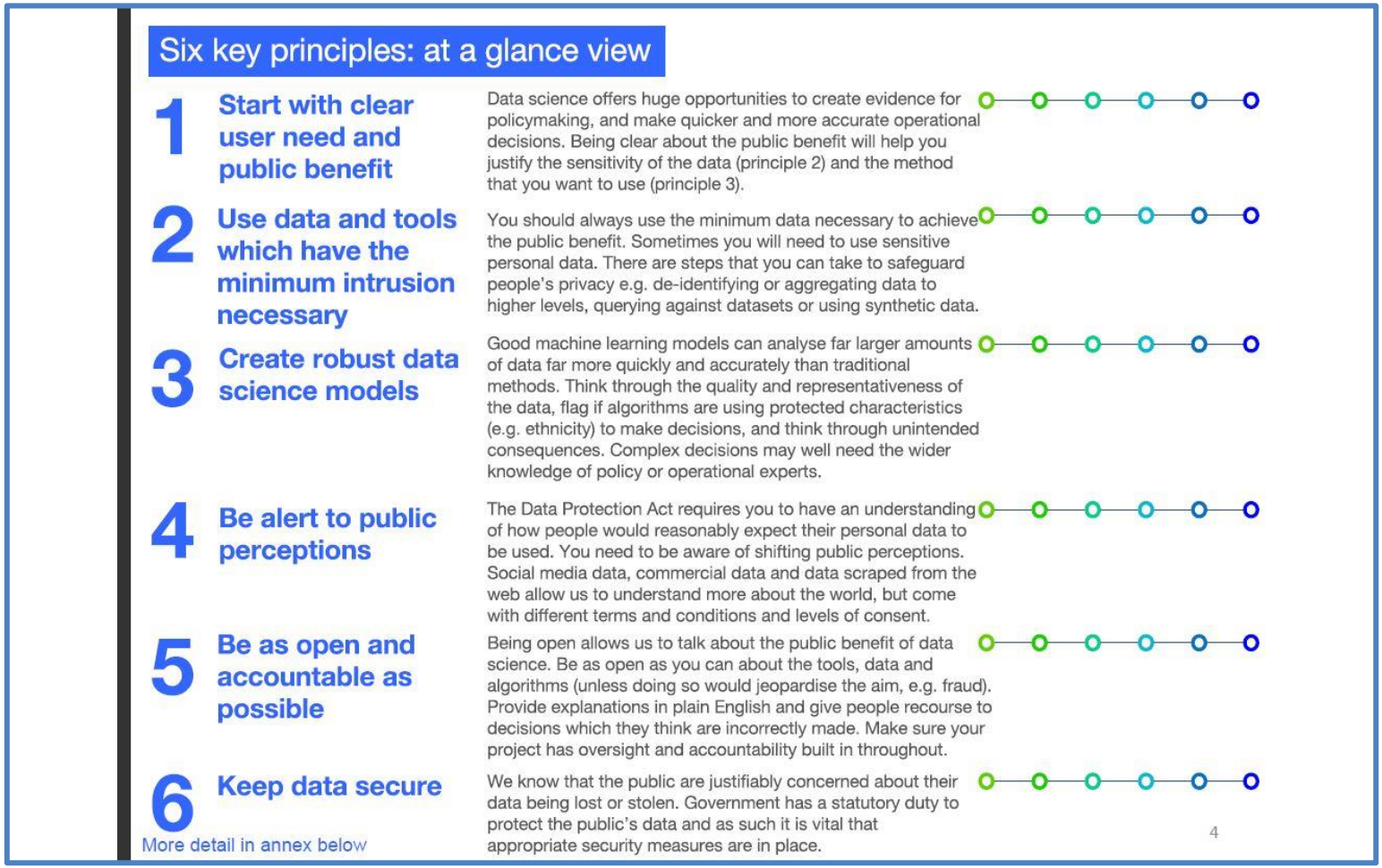
- Questions without easy answers
  - What counts as research? New knowledge? Public good?
  - Is pre-existing data always safe to use?
  - Does it matter what type of entity created the data?

- Details about Facebook's IRB
  - Member names and any info about proceedings not made public
  - http://www.wsj.com/articles/facebook-offers-details-how-it-handles-research-1465930152

- Key point – **provenance –** sources…
  - Where did your data come from?
  - How do you know?
  - How did data subjects think it was going to be used?

Jackman, M., & Kanerva, L. (2016). "Evolving the IRB: Building Robust Review for Industry Research". *Washington and Lee Law Review Online*,*72*(3), 442.

Tools for good practice – 1
UK Cabinet Office-Data Science Ethical Framework



UK Data Service

Tools for good practice - 2

- Specific to problems of personal/sensitive data covered by Data Protection Act

Personal data	Does your big data project need to use personal data at all? If you are using personal data, can it be anonymised? If you are processing personal data you have to comply with the Data Protection Act.
Privacy impact assessments	Carry out a privacy impact assessment to understand how the processing will affect the people concerned. Are you using personal data to identify general trends or to make decisions that affect individuals?
Repurposing data	If you are repurposing data, consider whether the new purpose is incompatible with the original purpose, in data protection terms, and whether you need to get consent. If you are buying in personal data from elsewhere, you need to practice due diligence and ensure that you have a data protection condition for your processing.
Data minimisation	Big data analytics is not an excuse for stockpiling data or keeping it longer than you need for your business purposes, just in case it might be useful. Long term uses must be articulated or justifiable, even if all the detail of the future use is not known.
Transparency	Be as transparent and open as possible about what you are doing. Explain the purposes, implications and benefits of the analytics. Think of innovative and effective ways to convey this to the people concerned.
Subject access	People have a right to see the data you are processing about them. Design systems that make it easy for you to collate this information. Think about enabling people to access their data on line in a re-usable format.

ICO: Big Data and Data Protection
https://ico.org.uk/media/1541/big-data-and-data-protection.pdf

Consent – good practice

- First best is still consent – also a legal requirement for personal or sensitive data
- But there are exceptions for consent – ex. Administrative Data Service
- Seek ethical review, e.g., for govt research https://www.statisticsauthority.gov.uk/national-statistician/national-statisticians-data-ethics-advisory-committee/
- Protect identities, consider expectations, and duty to protect from harm remains.

- Disability networks study (Trevisan and Reilly) http://dx.doi.org/10.1080/1369118X.2014.889188

Anonymisation – good practice

- Useful guidance
  - Office of National Statistics
  - UKAN
  - ICO Anon Code of Practice
  - Anonymisation Decision making Framework http://tinyurl.com/ADF-TALK

- Consider risk of linkage when publishing and sharing
- Not absolute, reframe as element of risk management
- And this will be even more true under the European General Data Protection Regulation – 2018.

Anonymisation – tools and resources

Existing and emerging tools:

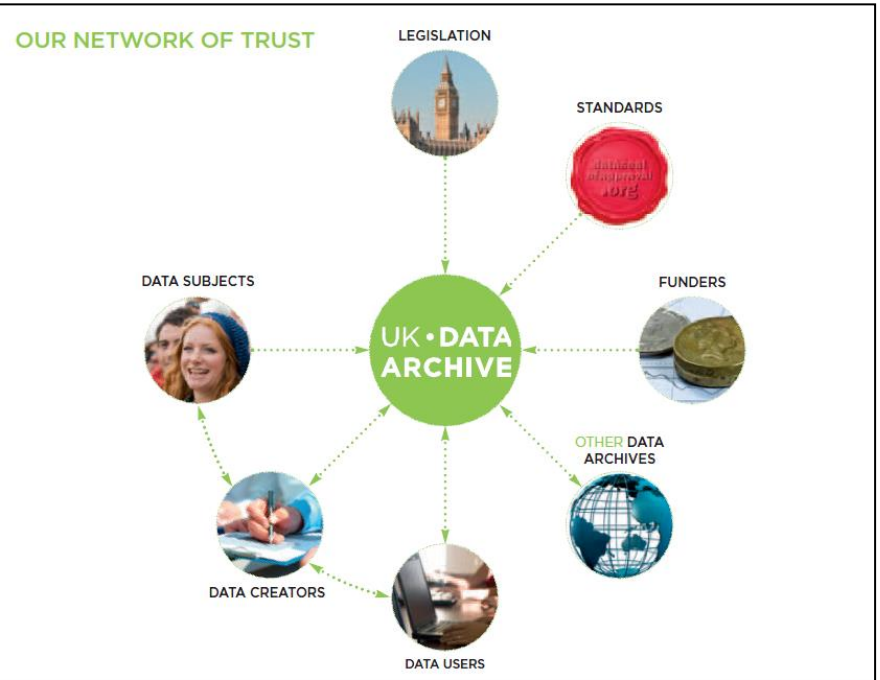
- Statistical disclosure control software e.g., Mu-argus, ARX
- Tools for qualitative data
  - http://data-archive.ac.uk/curate/standards-tools/tools

Published resources:

- UKAN Anonymisation Decision-making Framework http://ukanon.net/ukan-resources/ukan-decision-making-framework/
- ONS *Disclosure control guidance for microdata produced from social surveys* http://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata
- *UCD Research Data Management Guide* http://libguides.ucd.ie/data/ethics

UK Data Service – all about trust

- Three tiers of data access
- Five Safes
- Secure Lab
- Administrative Data Service
- https://www.ukdataservice.ac.uk/



Future ideas?

Ethics training for BD social research

Case studies for **social** data researchers, like these for data scientists http://bdes.datasociety.net/output/

UK Data Service

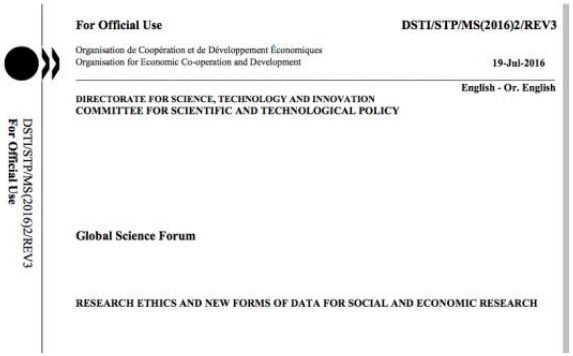
Just a few of those deserving thanks…

- OECD Global Science Forum-Expert Group-Research ethics and new forms of data for social and economic research, http://www.oecd.org/sti/sci-tech/oecdglobalscienceforum.htm
- Sara Day Thomson, Preserving Social Media, DPC Tech Watch Report 16-01, Feb 2016. (Also transaction data 16-02) http://dpconline.org/publications/technology-watch-reports
- Web Science Institute, Southampton, http://www.southampton.ac.uk/wsi/research/index.page?
- GESIS, http://www.gesis.org/en/home/
- Data & Society's Ethics in "Big Data" Research

gesis Leibniz Institute for the Social Sciences

SOCIAL DATA SCIENCE LAB.

Scalable real-time social data analytics for research, policy & practice


For Official Use DSTI/STP/MS(2016)2/REV3

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

19-Jul-2016

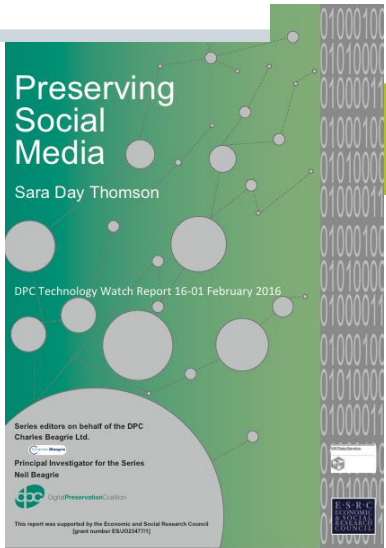
English - Or. English

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY

DSTI/STP/MS(2016)2/REV3
For Official Use

Global Science Forum

RESEARCH ETHICS AND NEW FORMS OF DATA FOR SOCIAL AND ECONOMIC RESEARCH

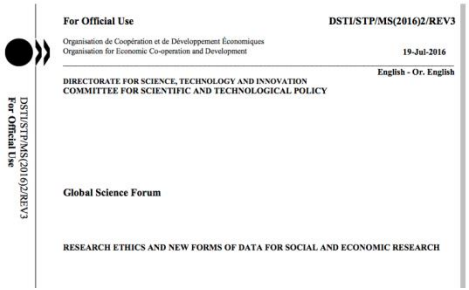


References – a rather random collection

- Barocas, S. and Nissenbaum, H. (2014) Big Data's End Run around Anonymity and Consent, in J. Lane et al. (eds) *Privacy, Big Data and the Public Good*. Cambridge University Press.
- Belmont Report. http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html
- Narayanan, A. and Felton, E. (2014) No silver bullet: de-identification still doesn't work, http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.
- ICO. Big data and data protection.
  https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

References - continued

- Academics Get Personal Over Big Data, IT News Australia, 11 July 2014
  - http://cacm.acm.org/news/176645-academics-get-personal-over-big-data/fulltext
- Nissenbaum's theory [contextual integrity]: see her 2010 slides http://www.practicalethics.ox.ac.uk/__data/assets/pdf_file/0009/21303/Nissenbaum.pdf
- Sara Day Thomson, Preserving Social Media, DPC Tech Watch Report 16-01, Feb 2016. (Also transaction data 16-02) http://dpconline.org/publications/technology-watch-reports
- OECD Global Science Forum-Expert Group-Research ethics and new forms of data for social and economic research, http://www.oecd.org/sti/sci-tech/oecdglobalscienceforum.htm


For Official Use

DSTI/STP/MS(2016)2/REV3

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

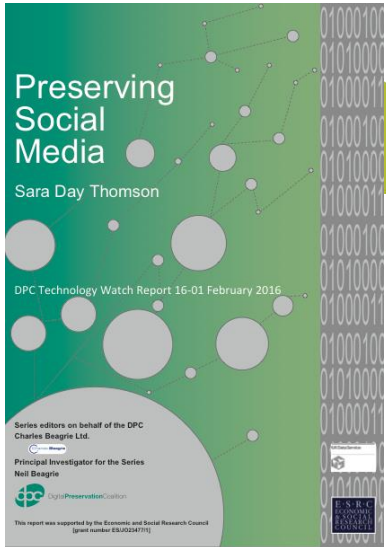
19-Jul-2016

English - Or. English

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY

Global Science Forum

RESEARCH ETHICS AND NEW FORMS OF DATA FOR SOCIAL AND ECONOMIC RESEARCH



More refs, tools, guides, cases, etc.

- Big Data and Society – Data ethics case studies
  - Is it ethical to use hacked public data?
  - Should you violate your employer's rules for public interest?
  - Do you use internet data without consent?
  - http://datasociety.net/blog/2016/04/13/data-ethics-case-studies/
- Markkula Center for Applied Ethics-U of Santa Clara
  - https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/thinking-ethically/
- Williams, M. L. and Burnap, P. 2015. Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology* 56(2), pp. 211-238. (10.1093/bjc/azv059)
- http://the-sra.org.uk/wp-content/uploads/ethics-in-social-media-research-matthew-williams.pdf

Acknowledgements

Many people have been generous with their ideas and materials: Katrin Weller, Susan Halford, Matt Williams, Luke Sloan, Sara Day Thomson, Audrey Guinchard, and many more. Thank you all.