Legal and ethical issues in curating “big new” data

Libby Bishop
GESIS and UK Data Service

CESSDA Expert Workshop
Bergen
12-13 September 2017

UK Data Service

Law & ethics

"Social science researchers using New Forms of Data confront difficult but important ethical questions concerning **consent, privacy risks, and safeguards from harm**. "

OECD(2016)
http://dx.doi.org/10.1787/5jln7vnpxs32-en

What is different about “big data”?

Big data are (usually) not generated **by** researchers or **for** research

- No ethics review before collection
- Protections during collection not done or infeasible, e.g., consent & anonymisation
- Secondary purpose often different from (not compatible with?) original purpose
- Sensitive data or vulnerable users increase risks of harm
- Basic definition of data (personal) undermined by linkage

Bigness is not the (legal/ethical) problem…
…wildness is

Comparison of frameworks for assessing legal and ethical issues in curating new forms of data
DRAFT – not for circulation

Source	Mannheimer and Hull, Dryad and Mont State Uni (STEP),	Townsend and Wallace (Aberdeen workshop SM)	Williams, Burnap, Sloan (Social Data Sci Lab)	Thomson (Digital Preservation Coalition)	Weller and Kinder Kurlanda (GESIS)	Bishop, Big Data Ethics (UKDA)	OECD, Privacy questions
Focus	Open data; Social media	Ethics for social media research	Ethics of publishing Twitter data	Preserving social media	Sharing social media; tech, legal, ethics	Curating big data-legal and ethical	New and novel data
Risk of harm	Sensitivity – how sensitive is data	Risk of harm -sensitivity of data -vulnerable groups	Must check, and note linkage b/t sensitive and other data	Much content sensitive, users not aware of onward usage			Is data personal or sensitive
Transpar-ency and Reprodu-cibility	Transparency			Provenance can not be authenticated	Reproducibility as ethical duty Documentation and metadata	Research integrity, inequality, and digital divide	Curation needed for transpy'y and repro-ducibility
Privacy	Expectations of privacy (includes consent and anonymisation)	Can users expect to be observed by strangers --Infd. consent --Anonymisation	User expectations (w data)	Public private ambiguity	What are expecations of content generators	Privacy --Consent --De-identification	What are users expectations
Legal	Legal (T&C of platforms, etc.)	Legal- comply with platform T&C, also funders, etc.	Compliance required	Primary use is monetising data; research is "different use";Copyright	T&C of platform		Requires legal complicance

Synthesis of seven frameworks –“common moral intuition”

- Legal compliance
  - Data protections laws
  - Policies – funders, journals, codes of conduct
- Ethical duty to minimise harm
  - Are data personal, sensitive, both, commercial sensitivity too…
- Reproducibility (ethical and technical requirement)
  - Metadata for discoverability
  - Documentation – comprehensive
  - Necessary for “critical sociology of data pipeline” (Halford&Savage)
- What are users’ expectations of others seeing the data
  - Was consent obtained? Opt out, opt in?
  - Are data de-identified? Is it it possible to do so?
- Other considerations
  - Other grounds for processing
  - “Public benefit”
  - Restricting access

Protecting confidentiality: the ‘5 Safes’

- Safe data - treat the data to protect respondent confidentiality
- Safe people - educate researchers to use data safely
- Safe projects - research projects for ‘public good’
- Safe settings - SecureLab system for sensitive data
- Safe outputs - SecureLab projects outputs screened

[5 Safes Video]

UK Data Service - example

- Smart Meter data
- Twitter User Ids in Reshare
- …

Questions

ukdataservice.ac.uk/help/

Follow us at:

ukdataservice@jiscmail.ac.uk

twitter.com/ukdataservice

facebook.com/ukdataservice

New and novel data (“Big Data”)

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

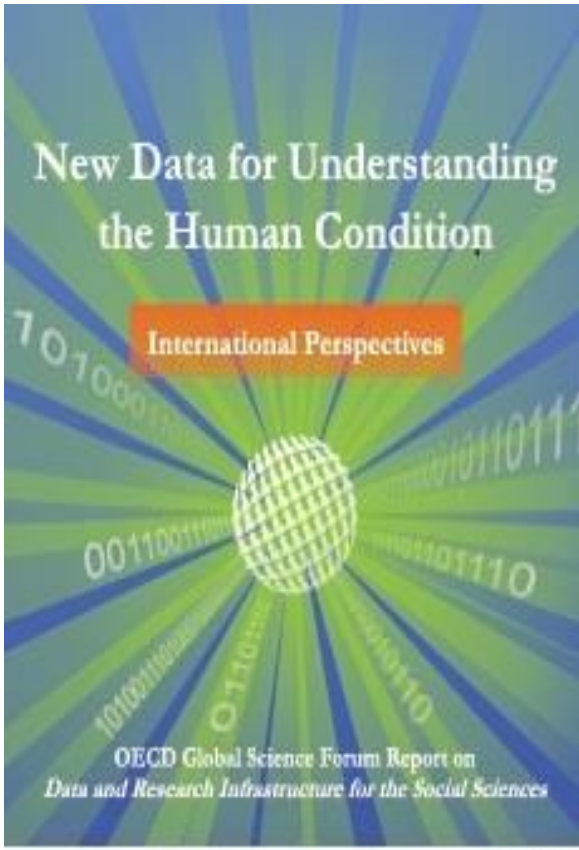
B: Official registrations or licensing requirements.

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

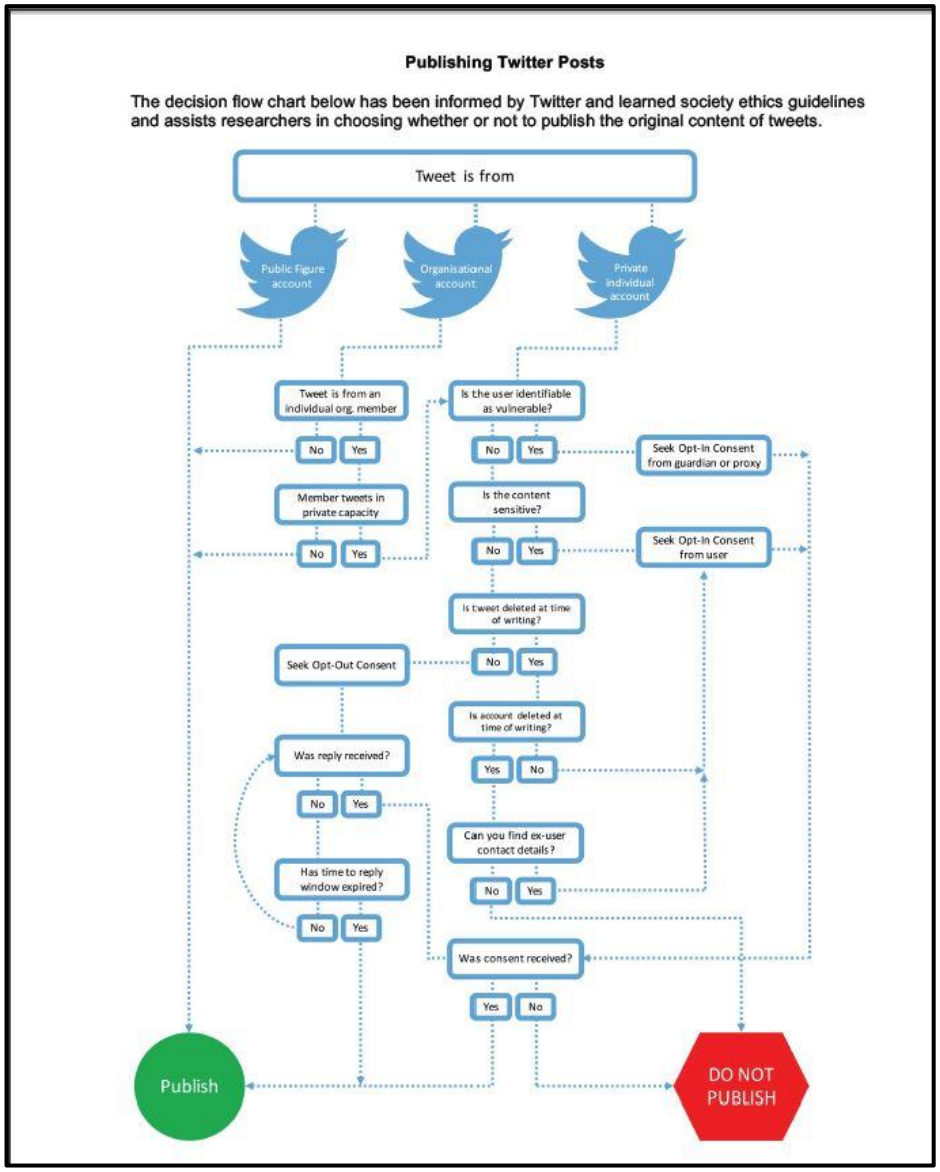
F: **Image** data, particularly aerial and satellite images



But what about…..???

- What if users are behaving “as if” space were private?
  - “Context collapse” – “know” it is public, but behave as if private
  - AoIR: privacy must consider “expectations &consensus of users”
- Should users’ expectations be considered?
  - Common (moral) sense to take them into account
  - DPA considers “reasonable expectations” of data subjects
- What % of Tweeters expect to be asked for consent?
  - About 80% of (500+) respondents expected to be asked for consent to publish their posts in academic outputs
- Does the subject matter make any difference?
  - Cyberhate, misogyny, suicidal tendencies, crime…
- When linking with data derived from algorithms?
  - Analytics can predict gender, age, sexual orientation
  - These are sensitive data under data protection law

Publishing Twitter posts – practical tool



Williams, M. L., Burnap, P. & Sloan, L. (2017) 'Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation', *Sociology*.

Gray, D. Talking About Women: Misogyny on Twitter, Master of Science Dissertation, Cardiff University, September 2015.

UK Data Service

Consent – good practice

- First best is consent – almost always a legal requirement for personal or sensitive data
- But there are exceptions, e.g., administrative data, but requires “public interest” and more
- Seek ethical review, e.g., for ONS data https://www.statisticsauthority.gov.uk/national-statistician/national-statisticians-data-ethics-advisory-committee/
- Protect identities, consider expectations, protect from harm

Consent – usually legally required

always ethically preferred

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."

President's Council of Advisors on Science and Technology. Report to the President: Big Data and Privacy: A Technological Perspective, at 38-39 (May 2014).

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.

De-identification – where do we stand?

- De-identification not binary or magic
  - Risks increase with dimensionality, linkage, esp. linkage across genres

- Emerging trends
  - Identifiability is a spectrum
  - Risk can be mitigated, not eliminated
  - De-identification remains a vital tool – as part of “risk mitigation strategy” (ICO Big Data https://ico.org.uk/media/1541/big-data-and-data-protection.pdf)

Rubinstein, I. Identifiability: policy and practical solutions for anonymisation and pseudonymisation, Brussels Privacy Forum. https://fpf.org/wp-content/uploads/2016/11/Rubinstein_framing-paper.pdf

Ex. Using de-identification for disclosure control

- ONS Wealth and Assets Survey
- Longitudinal; Wave 1 30K households; Wave 2 20K
- Remove households of size 10 and above
- Top code Individual Age at 80
- Give special consideration to the Wealth variable
  - all variables relating to wealth and finance top-coded
  - compromise-variables of lesser research importance removed to reduce the risk of identification (geography)
  - data reviewed wave by wave, due to dynamic nature

Apply optimal SDC techniques that reduce disclosure risks with minimal information loss, and preserve data utility

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
- United Nations Economic Commission for Europe: Managing statistical confidentiality and microdata access

Tools for good practice

- 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

Conducting Privacy Impact Assessments: Code of Practice

Summary – ways to use and share big data

To share data—even big data—ethically and legally

- Seek informed consent for data sharing and long-term preservation
- Protect identities when promised
- Regulate access where needed (all or part of data) e.g. by group, use, time period

Access conditions

Open

- available for download/online access under open licence without any registration

Safeguarded

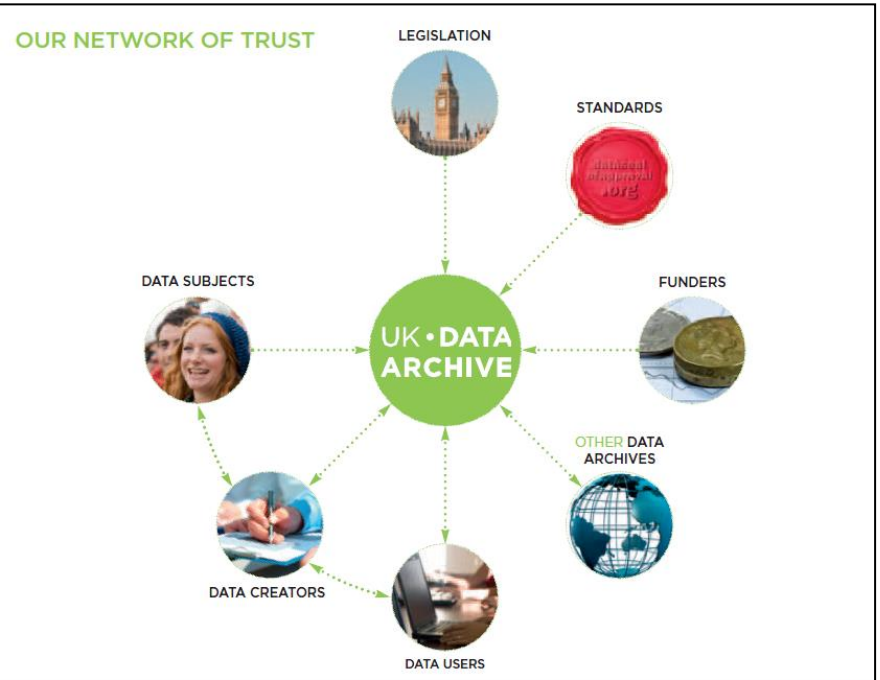
- available for download/online access to logged-in users who have registered and agreed to an End User Licence

Controlled

- available for remote or safe room access registered users whose research proposal has been approved by an access committee and who have received specialist training

UK Data Service – it comes down to trust

- Five Safes
- Consent
- De-identify
- Regulate data access



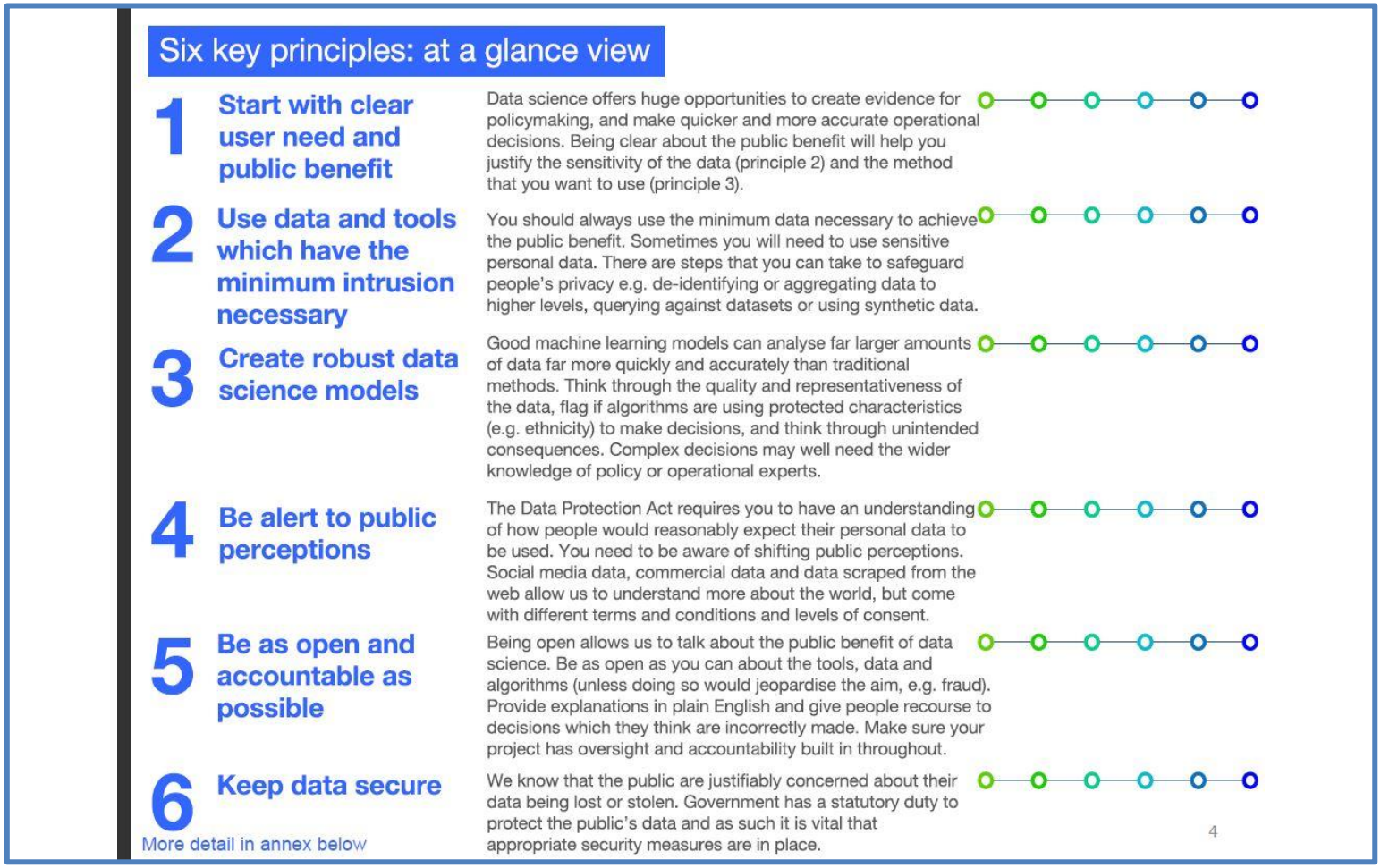
Big data and data sharing: Ethical issues

UK Data Service

https://bigdata.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf

UK Data Service

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



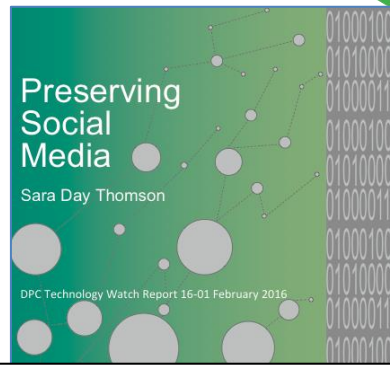
UK Data Service

OECD Privacy Questions for Reflection [partial list]

- What?
  - Is the data personal or sensitive (under Data Protection laws)?
- Who?
  - Do data subjects have resources to object?
  - Might the research discriminate against groups, not just individuals?
  - Do any of the data groupings constitute vulnerable groups?
- Where?
  - In what settings is the information gathered, and what uses are expected in those settings?
- Why?
  - Does the research undermine the values of the place from which the data are gathered?
  - Do the purposes of the research harmonise or conflict with the aims, goals, and values of the sites the research might be affecting?
  - What are data subjects' reasonable expectations concerning the research project's re-contextualisation of their information? (purpose compatibility)

Resources

- OECD (2016), “Research Ethics and New Forms of Data for Social and Economic Research”, OECD Science, Technology and Industry Policy Papers, No. 34, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jln7vnpxs32-en
- 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



SOCIAL DATA SCIENCE LAB.

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

Resources

- 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.

Resources

- 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