Ethical issues in accessing and using “big data”

Libby Bishop
Research Data Management Team
UK Data Service – University of Essex

Big Data and Analytics Summer School
BD014 – Secure Access Protocols for Big Data
24 August 2015

UK Data Service

Overview

1. what is at stake?
2. clarify key concepts
3. facebook case – discuss in groups
4. legal and ethics overview
5. current practices that protect privacy
6. BD challenges to current practices
7. emergent practices – good enough? Too soon to tell…

1. What is at stake?

search jobs dating more UK edition

theguardian
Winner of the Pulitzer prize

ort football opinion culture economy lifestyle fashion environment tech travel browse all section

Care.data is in chaos. It breaks my heart

Ben Goldacre

Medical data has huge power to do good, but it presents risks too. When leaked, it cannot be unleaked. When lost, public trust cannot be easily regained

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

"Our findings indicate there is a 'data trust deficit' whereby trust in institutions to use data appropriately is lower than trust in them in general." Royal Statistical Society 2014 *Trust in Data*

	High trust (8-10)	Low trust (0-4)
Your GP surgery	41%	15%
The NHS	36%	17%
The police	28%	26%
Academic researchers and universities	25%	22%
The Office for National Statistics (ONS)	23%	24%
Charities	15%	35%
Banks	14%	45%
Your local authority	14%	35%
Online retailers, for example, Amazon, Asos and play.com	13%	38%
The British government	13%	46%
Supermarkets	10%	42%
Insurance companies	7%	55%
Telecommunications companies, such as mobile phone	6%	54%

…how do we design systems that make use of our data collectively to benefit society as a whole, while at the same time protecting people individually?…This is it: this is the fundamental issue of the information age.”

Bruce Schneier 2015 *Data and Goliath*

2. Clarify concepts: New and novel data (“Big Data”)

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

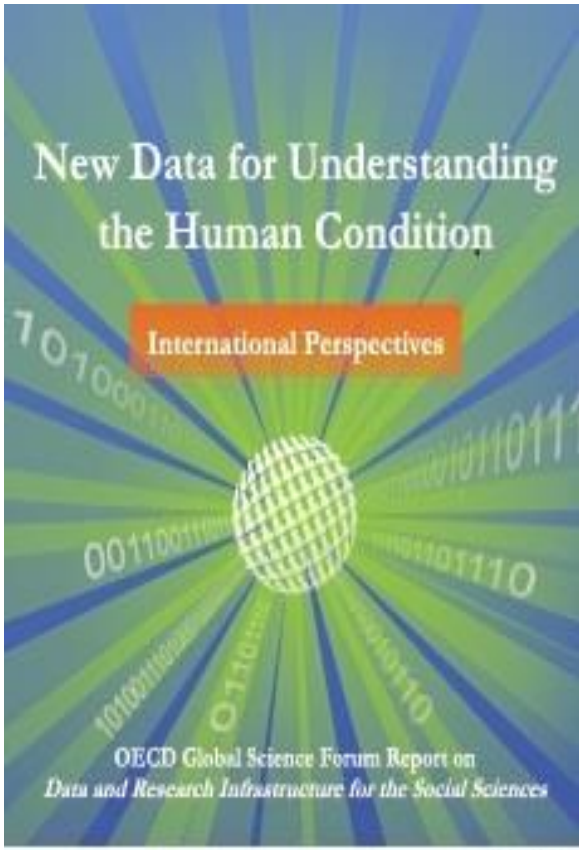
B: Official registrations or licensing requirements.

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



Forms of data with new research potential

Broad category of data	Detailed categories	Examples
Category A: Government transactions	Individual tax records	Income tax; tax credits
	Corporate tax records	Corporation tax; sales; tax, value added tax
	Property tax records	Tax on sales of property; tax on value of property
	Social security payments	State pensions; hardship payments; unemployment benefits; child benefits
	Import/export records	Border control records; import/export licensing records
Category B: Government and other registration records	Housing and land use registers	Registers of ownership
	Educational registers	School inspections; pupil results
	Criminal justice registers	Police records; court records
	Social security registers	Registers of eligible persons
	Electoral registers	Voter registration records
	Employment registers	Employer census records: registers of persons joining/leaving employment
	Population registers	Births; marriages; civil unions; deaths; immigration/emigration records; census records
	Health system registers	Personal medical records; hospital records
	Vehicle/driver registers	Driver licence registers; vehicle licence registers
	Membership registers	Political parties; charities; clubs

Category C: Commercial transactions	Store cards	Supermarket loyalty cards
	Customer accounts	Utilities; financial institutions; mobile phone usage
	Other customer records	Product purchases; service agreements
Category D: Internet usage	Search terms	Google©; Bing©; Yahoo© search activity
	Website interactions	Visit statistics; user generated content
	Downloads	Music; films; TV
	Social networks	Facebook©; Twitter©; LinkedIn©
	Blogs; news sites	Reddit
Category E: Tracking data	CCTV images	Security/safety camera recordings
	Traffic sensors	Vehicle tracking records; vehicle movement records
	Mobile phone locations: GPS data	
Category F: Satellite and aerial imagery	Visible light spectrum	Google Earth©
	Night-time visible radiation	Landsat
	Infrared; radar mapping	

Clarify concepts: What are ethics?

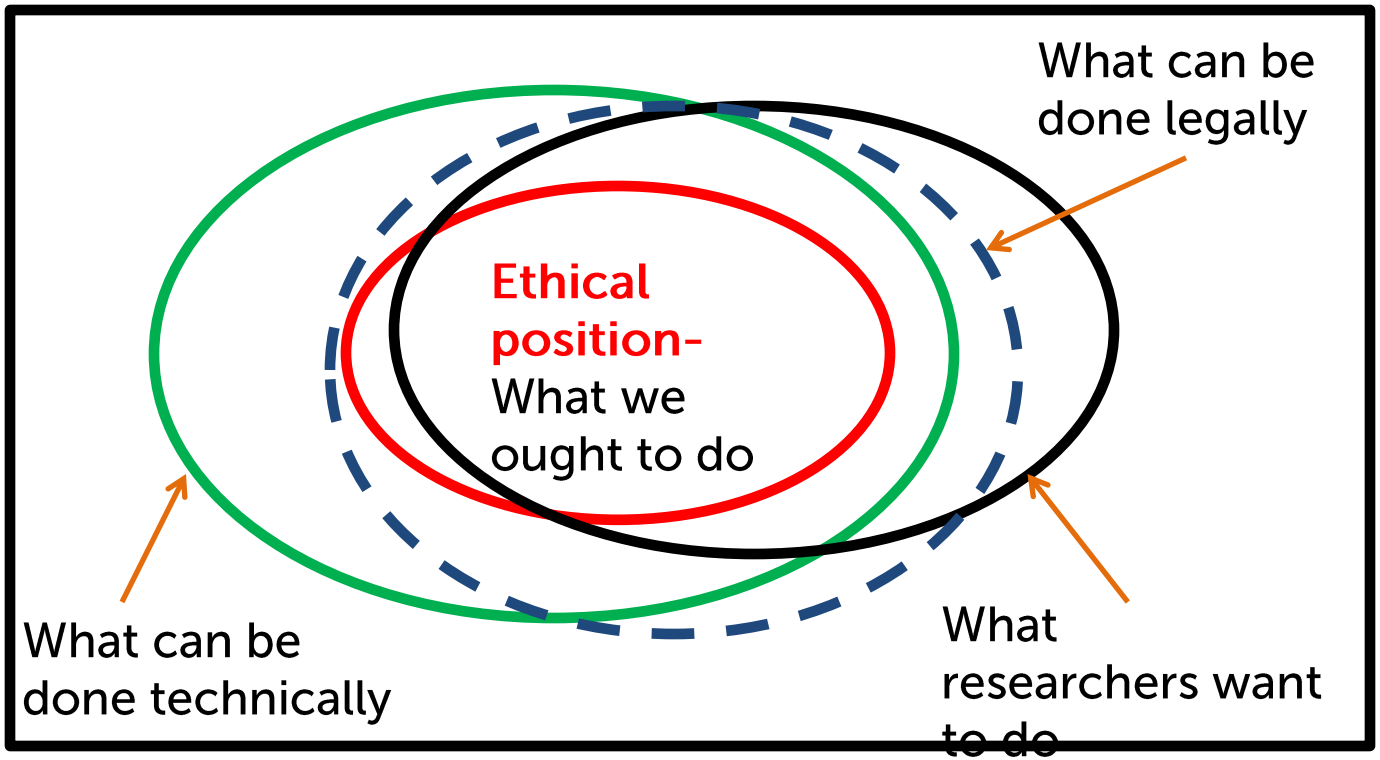
- …standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues. …Such standards are adequate standards of ethics because they are supported by consistent and well-founded reasons.
- Ethics also means the continuous effort of studying our own moral beliefs and moral conduct, and striving to ensure that we, and the institutions we help to shape, live up to standards that are reasonable and solidly-based.
- http://www.scu.edu/ethics/practicing/decision/whatisethics.html

Clarify concepts:
What are research ethics?

- Research should be done with integrity, quality and **transparency**
- Research participants must normally be **informed fully** about the purpose, methods and **intended possible uses**
- The **confidentiality** of participants' information must be respected
- Research participants must take part **voluntarily**
- **Harm** to research participants must be avoided **in all [?]** instances
- The **independence** of research must be clear, and any conflicts of interest or partiality must be explicit

- e.g. ESRC Framework for Research Ethics, 2010

Ethics-what we *ought* to do



Adapted from: Mandy Chessell, (2014) *Ethics for Big Data and Analytics*
http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG Study Report - Ethics for BD&A.pdf

3. Facebook – discuss case in groups

- Undergraduate student data

Michael Zimmer

“But the data is already public”: on the ethics of research in Facebook, *Ethics and Information Technology* (2010) 12:313–325

DOI 10.1007/s10676-010-9227-5

Published online: 4 June 2010

4. Relevant UK legislation – privacy focus

- Data Protection Act (1998)
- Freedom of Information Act (2000)
- Statistics and Registration Services Act (2007) + bespoke government legislation
- Human Rights Act (1998)
- Environmental Information Regulations (2004)
- Duty of confidentiality

Data Protection Act, 1998

- Personal data:
  - relate to a living individual
  - individual can be identified from those data or from those data and other information
  - include any expression of opinion about the individual
- Only disclose personal data if consent given to do so (and if legally required to do do)
- DPA does not apply to anonymised data

- processed fairly and lawfully
- obtained and processed for specified purpose
- adequate, relevant and not excessive for purpose
- accurate
- not kept longer than necessary
- processed in accordance with the rights of data subjects, e.g. right to be informed about how data will be used, stored, processed, transferred, destroyed; right to access info and data held
- kept secure
- not transferred abroad without adequate protection

Data Protection Act and research

- Exceptions for personal data collected as part of research:
  - can be retained indefinitely (if needed)
  - can be used for other purposes in some circumstances
  - people should still be informed

The Data Protection Act is not intended to, and does not, inhibit ethical research

Best practice for legal compliance

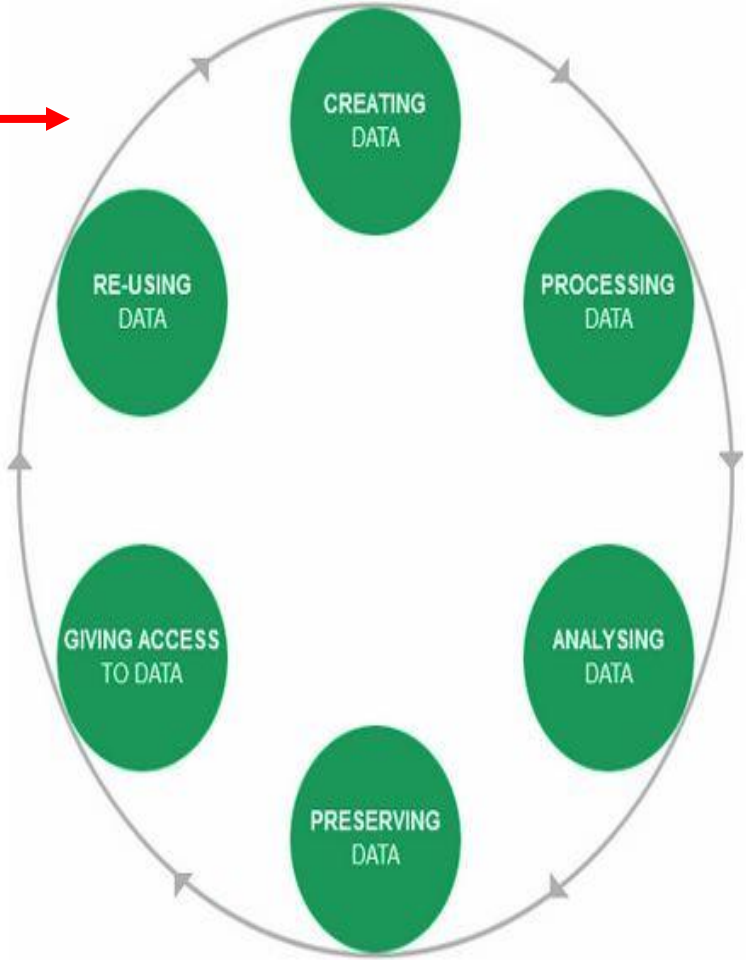
The public benefit of data access should be greater than the harm or distress caused by disclosure

- ✓ Investigate early which laws apply to data
- ✓ Researchers not collect/store personal or sensitive data if not essential to your research
- ✓ Seek advice from local research office experts
- ✓ Plan early in research
- ✓ If need to deal with personal or sensitive data:
  - ✓ inform participants about how their data will be used
- ✓ Not all research data are personal (e.g. anonymised)

5. Current practices to privacy throughout the data lifecycle

Consent

Anonymise/ de-identify



Control access

In practice: wording in consent form / information sheet

Use of the information I provide beyond this project		
I agree for the data I provide to be archived at the UK Data Archive.[2]	☐	☐
I understand that other genuine researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I understand that other genuine researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐

As the ESRC is a publicly funded body, it has developed ways to share data among academic researchers (subject to strict conditions). To this end, we hope you will allow your anonymised transcript to be stored as part of the UK Data Archive (a service provider for the Economic and Social Data Service).

Informed consent – information sheets and forms

- Meet requirements of Data Protection laws:
  - purpose of the research
  - what is involved in participation
  - benefits and risks
  - mechanism of withdrawal
  - usage of data – for primary research and sharing
  - strategies to ensure confidentiality of data, where relevant
- Complete for all purposes: use, publishing, sharing
- Simple and avoiding excessive warnings

Anonymising data

- Plan or apply editing at time of transcription *except: longitudinal studies - (linkages)*
- Consistency within research team and throughout project
- Avoid over-anonymising - removing/aggregating information distort data, make them unusable, unreliable or misleading
  - ex. image data

Controlling access a better option than over-anonymising

In practice: example anonymisation

Ex 1. Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

Date of Interview: 21/02/02

Interview with Lucas Roberts, DEFRA field officer
Date of birth: 2 May 1965
Gender: Male
Occupation: Frontline worker
Location: Plumpton, North Cumbria

Lucas was living at home with his parents, "but I'm hoping to move out soon" so we met at his parents' small neat house. We sat in a very comfortable sitting room with an open fire and Lucas made me coffee and offered shortbread. Although at first Lucas seemed a little nervous, quick to speech and very watchful he seemed to relax as we spoke and to forget abut the tape.

I will just start by asking you to tell me a little bit about yourself and your background.

Well it is an agricultural background. I grew up on the farm where my brother is now. After I left school I did work on the farm but went to college and did exams, did land use recreation, sort of countryside/ environmental management course. So I obviously left agriculture, did the course and came back [to the farm] at weekends.

Comment [v1]: Replace: Ken
Comment [v2]: delete
Comment [v3]: delete
Comment [v4]: Replace: Ken
Comment [v5]: Replace: Ken
Comment [v6]: Replace: Ken

Managing access to data at UKDS

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 • Special agreements (depositor permission; approved researcher; embargo for fixed time period
Controlled	• available for remote or safe room (Secure Lab) access to authorised and authenticated users whose research proposal has been and who have received training

In practice: access conditions ReShare

AVAILABLE FILES

Data

− Security_%26_Networks.xlsx

Accessible to:	Registered users only (safeguarded data)
File or bundle content:	Data
File or bundle description:	Security & Networks
File format:	application/octet-stream
License:	UK Data Service End User Licence
File size:	10Kb

+ Sierra_Leone_Security_%26_Networks__Coded.xls

+ Copy_of_Somalia_SC_peace_initiatives.xlsx

Documentation

− Sierra_Leone_methods.doc

Accessible to:	Anyone (open data)
File or bundle content:	Documentation
File or bundle description:	Sierra Leone methods
File format:	application/msword
License:	UK Data Service End User Licence
File size:	64Kb

6. Big data – challenges for consent, de-identification, and access controls

	Social Media	Admin Data	Geo/spatial	Linkage	'Reach-Ability"
Consent	Not "fully informed"; Individual vs. network members	Unconsented; reuse for very different purposes	User does not even know data being collected	Mixed terms of consent	Consent of few may implicate many
De-identi-fication	Too much info lost w/ de-id.	Often personal, sensitive	Highly disclosive		May be "anon", yet reachable
Access controls	Opaque IP; if private, what about replication?	Not for research purposes; commercial use; complex AC is costly		Depends on unambigu-ous categories	

Consent – big data challenges

- indeterminate – unknowable future uses that might prove valuable
- unending – no time limits
- unpredictable – where data will go, who can access, how used

…how does the data controller explain that it is impossible to know in advance what further information might be discoverable? These factors diminish the value of informed consent because they seem to require notice that does not delimit future uses of data and the possible consequences of such uses. As many have now argued, consent under those conditions is not meaningful.

Barocas, S. and Nissenbaum, H. (2014) Big Data's End Run around Anonymity and Consent

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.)

Access summary for UKDS

Access depends on being able to classify:

User type and location	• Higher education/further education • non HE/FE • UK • non UK
Data access conditions	• End User Agreement • Special Conditions • Special Licence (incl. Approved Researcher) • Secure Lab access only
Usage/project characteristics	• Commercial • non-Commercial

7. Current good practices – no one solution

Good practice: ICO, Big Data, and DP

- ICO advice suggests thinking carefully about core principles of DP
- Especially when dealing with new forms, not explicitly covered.
- Fairness, no harm to data provider,

- Recent ICO report: Big data and data protection

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.

Emerging models - consent

- If strictly interpreted, then may require explicit, specific consent; written – exemptions?
- Broad consent – Genome Project, biobanks. Accepted in UK, but not across Europe (e.g., Germany)
- Dynamic consent – maintain contact, seek content for any new use "substantially" different from original
- Alternative legitimate basis (in data protection laws)
  - Administrative Data Research Network
  - Access review committees – medical, Understanding Society
  - Citizen panels
- Proposed EU General Data Protection ***Regulation***

Proliferating guides – making choices

- There IS an app for that…
  http://www.scu.edu/ethics/practicing/decision/
- IBM – Ethics for BD and analytics
- Privacy Impact Assessment (ICO recommends)
- ICO's Data sharing code of practice
- UK Anonymisation Network (UKAN)
- Best practice on survey data from ONS

References

- 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.

- Chessell, M. (2014) Ethics of Big Data and Analytics, http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf.

- 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

Contact

Collections Development and Producer Relations team
UK Data Service
University of Essex
ukdataservice.ac.uk/help/get-in-touch.aspx