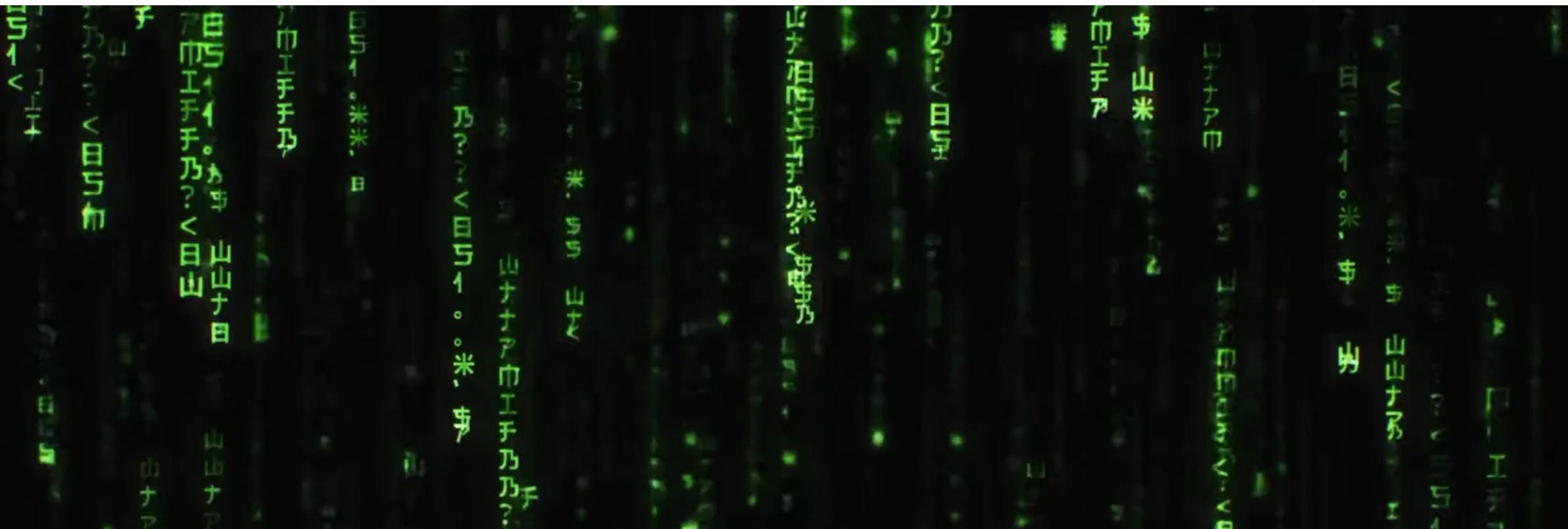
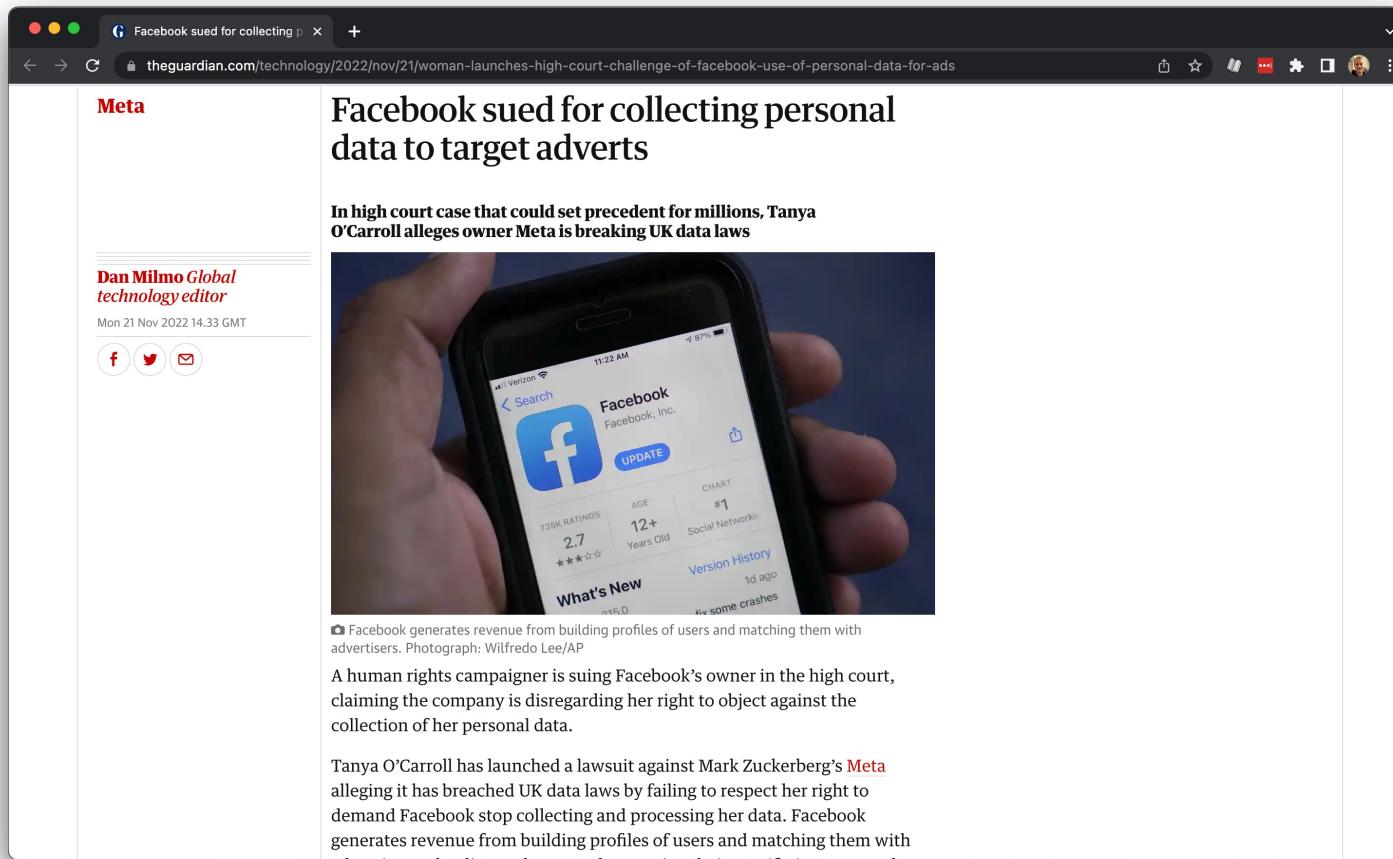


Data, Politics and Society

W7 – Safe Research



In the news

A screenshot of a news article from theguardian.com. The title is "Facebook sued for collecting personal data to target adverts". Below the title is a subtitle: "In high court case that could set precedent for millions, Tanya O'Carroll alleges owner Meta is breaking UK data laws". The main image shows a hand holding a smartphone displaying the Facebook app store page. The text below the image states: "Facebook generates revenue from building profiles of users and matching them with advertisers. Photograph: Wilfredo Lee/AP". The article continues: "A human rights campaigner is suing Facebook's owner in the high court, claiming the company is disregarding her right to object against the collection of her personal data. Tanya O'Carroll has launched a lawsuit against Mark Zuckerberg's Meta alleging it has breached UK data laws by failing to respect her right to demand Facebook stop collecting and processing her data. Facebook generates revenue from building profiles of users and matching them with".

Where we at?

W6

W7



Regulations and governance

W8

W9



Crowdsourcing, VGI, and Geographic Citizen Science

W10



Critical Data Studies

Today*

- Importance of data access for social science
- Data disclosure risk
- How can we conduct research safely?
- Slightly meme heavy

* Slides available later today.

Why is access important?

Bender *et al.* 2020:

- Validating the data generating process
- Replication
- Building knowledge infrastructure

How to organise this safely?



Approaches to giving access

Bender *et al.* 2020:

- Statistical disclosure techniques
- Research Data Centers
- A combination of both?

Quiz



Approaches to giving access

Bender *et al.* 2020:

- Statistical disclosure techniques
- Research Data Centers
- A combination of both?

Quiz

Go to www.menti.com and use the code 5483 7523



Statistical disclosure control

- Addressing residual risk of re-identification in results for publication
- Precautionary, but: consistent with good research practices and balancing utility and risk

Statistical disclosure control

Variable	Description
id	random ID number
male	male dummy
age	age
ethnicity	census ethnicity category
diabetic	diabetes diagnosed
lcovid	long covid
education	highest education
soceco	socio-economic group
income	annual income in £
incomeqrt	income quartile
imputed	imputed value dummy

Statistical disclosure control

Variable	Description
id	random ID number
male	male dummy
age	age
ethnicity	census ethnicity category
diabetic	diabetes diagnosed
lcovid	long covid
education	highest education
soceco	socio-economic group
income	annual income in £
incomeqrt	income quartile
imputed	imputed value dummy

Statistical disclosure control

Variable	Description
id	random ID number
male	male dummy
age	age
ethnicity	census ethnicity category
diabetic	diabetes diagnosed
lcovid	long covid
education	highest education
soceco	socio-economic group
income	annual income in £
incomeqrt	income quartile
imputed	imputed value dummy

Statistical disclosure control

	Has long covid?		
	No	Yes	Total
Gender			
Female	85	6	91
Male	58	1	59
Total	143	7	150

Statistical disclosure control

	Has long covid?		
	No	Yes	Total
Gender			
Female	85	6	91
Male	58	1	59
Total	143	7	150

Statistical disclosure control

		Has long covid?		
		No	Yes	Total
Diabetes	No			
	No	114	2	116
Yes	Yes	29	5	34
Total	Total	143	7	150

Statistical disclosure control

		Has long covid?		Total
		No	Yes	
Diabetes	No	114	2	116
	Yes	29	5	34
Total	143	7		150

Statistical disclosure control

- It is not only unique observations that matter
- Average salary of the highlighted cell in the first and second table: £30,000
 - (1) male, with long covid (count 1): £30,000
 - (2) no diabetes, with long covid (count 2): each can calculate salary of the other
 - (3) counts above 3: no certainty on income of others
- So: statistical disclosure control is about the risk of disclosure

Statistical disclosure control

	At least one value imputed?		
	No	Yes	Total
Gender			
Female	89	2	91
Male	58	1	59
Total	147	3	150

Statistical disclosure control

	Income Quartile				Total
	1	2	3	4	
Education					
PG Degree	1	1	8	18	28
UG Degree	2	6	14	17	39
College	8	18	16	3	45
School	13	9	0	0	22
None	13	3	0	0	16
Total	37	37	38	38	150

Statistical disclosure control

	Income Quartile				Total
	1	2	3	4	
Education					
PG Degree	1	1	8	18	28
UG Degree	2	6	14	17	39
College	8	18	16	3	45
School	13	9	0	0	22
None	13	3	0	0	16
Total	37	37	38	38	150

Statistical disclosure control

	Income Quartile					Total
	1	2	3	4		
Education						
PG Degree	1	1	8	18	28	
UG Degree	2	6	14	17	39	
College	8	18	16	3	45	
School	13	9	0	0	22	
None	13	3	0	0	16	
Total	37	37	38	38	150	

Statistical disclosure control

	Age				Total
	16-17	18-19	20-23	24-29	
Education					
UG Degree	0	0	51	64	115
College	0	25	33	57	115
School	15	18	19	41	93
None	8	7	12	17	44
Total	23	50	115	179	367

Statistical disclosure control

	Income Quartile				Total
	1	2	3	4	
Education					
PG Degree	1	1	8	18	28
UG Degree	2	6	14	17	39
College	8	18	16	3	45
School	13	9	0	0	22
None	13	3	0	0	16
Total	37	37	38	38	150

Statistical disclosure control

	Income Quartile				
	1	2	3	4	Total
Education					
PG Degree	< 3	< 3	8	18	26
UG Degree	< 3		14	17	37
College	8	18	16	3	45
School	13	9	< 3	< 3	22
None	13	3	< 3	< 3	16
Total	34	36	38	38	146

Statistical disclosure control

	Income Quartile				Total
	1	2	3	4	
Education					
PG Degree	0	0	10	20	30
UG Degree	0	5	15	15	35
College	10	20	15	5	50
School	15	10	0	0	25
None	15	5	0	0	20
Total	40	40	40	40	150

Statistical disclosure control

	Income Quartile					Total
	1	2	3	4		
Education						
Postgrad / Degree	3	7	22	35	57	
College	8	18	16	3	45	
School	13	9	0	0	22	
None	13	3	0	0	16	
Total	37	37	38	38	150	

Which is best?

- Depends on the output, not all approaches will work all the time.
- Depends on the message you want to present.

Statistical disclosure control

	Socio-economic group		
	X1	X2	Total
Age			
50-54	21	11	32
55-59	25	11	36
60-64	28	12	40
65+	31	11	42
Total	105	45	150

	Socio-economic group		
	X1	X2	Total
Age			
50-54	21	11	32
55-59	25	11	36
60-64	28	12	40
65+	31	11	42
Total	105	45	150
	Socio-economic group non-diabetics		
	X1	X2	Total
Age			
50-54	17	7	32
55-59	19	9	36
60-64	23	8	40
65+	23	10	42
Total	82	34	150

	Socio-economic group		
	X1	X2	Total
Age			
50-54	21	11	32
55-59	25	11	36
60-64	28	12	40
65+	31	11	42
Total	105	45	150
	Socio-economic group non-diabetics		
	X1	X2	Total
Age			
50-54	17	7	32
55-59	19	9	36
60-64	23	8	40
65+	23	10	42
Total	82	34	150

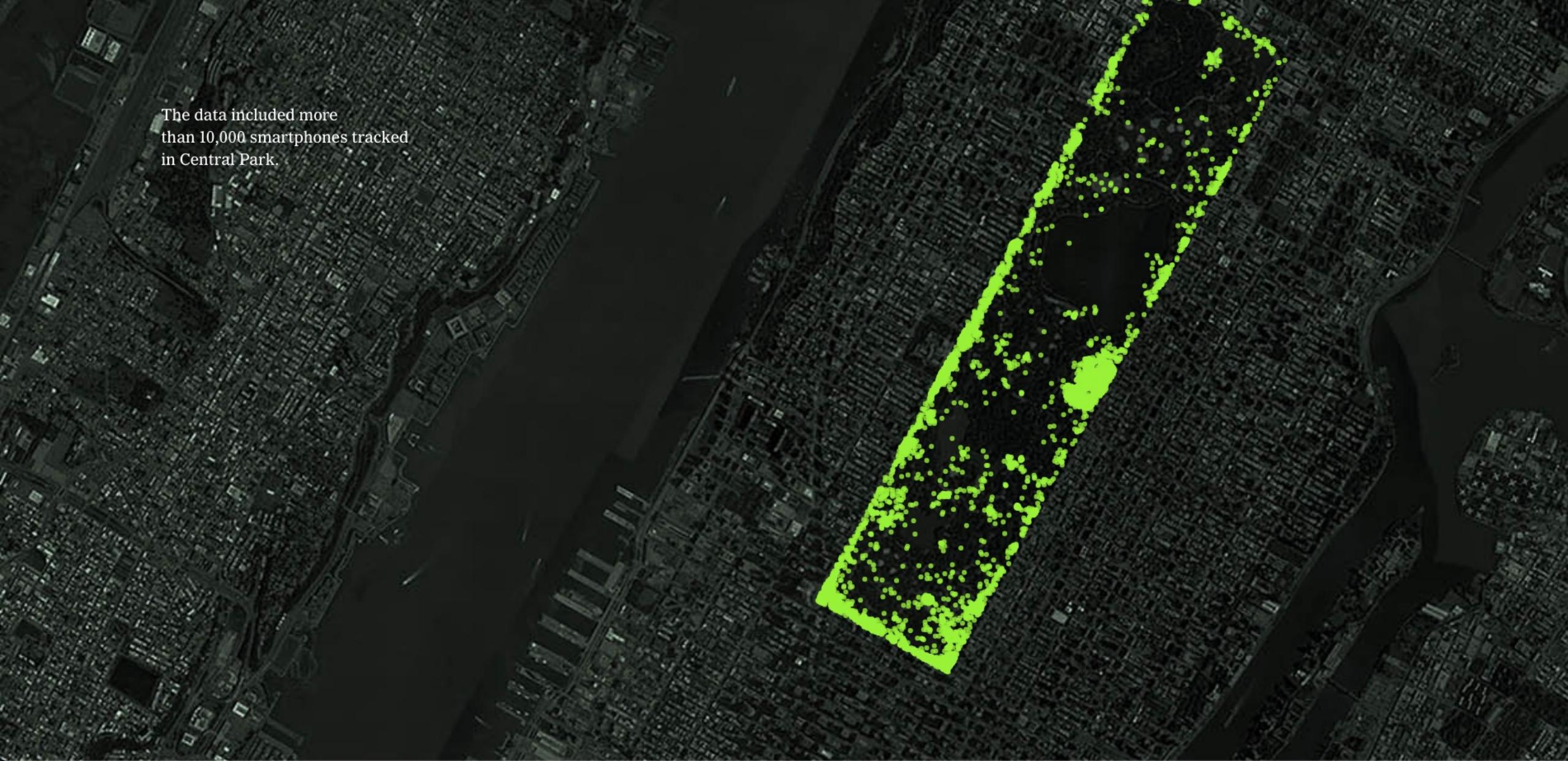
Beyond tables

Same rules apply to:

- Linear regression coefficients
- Scatter plot of regression residuals
- Box plots
- Minimum, maximum, median
- Ranks

And of course: maps!

The Privacy Project



The data included more than 10,000 smartphones tracked in Central Park.

The Privacy Project



Here is one smartphone, isolated
from the crowd.

The Privacy Project



Here are all pings from
that smartphone over the period
covered by the data.

The Privacy Project



Connecting those pings reveals a diary of the person's life.

Five safes

Safe projects: Is this an appropriate use of the data?

Safe people: How trustworthy are the researchers?

Safe setting: Does the environment prevent misuse?

Safe data: Is the level of detail appropriate?

Safe outputs: Is there any confidentiality risk from publication?

Attitudes toward data sharing

Imagine you are the Data Provider. How would your impression of researchers affect the way you make data available?

- If you believe researchers will try to look after the data, but could make mistakes then you can train them?
- If you do not trust researchers then perhaps you will only make Public Use Files available, hugely restrictive in the detail available.

Default of most data services is that users can be trusted but will need some training on the specifics.

Data protection

Research Data Centers typically host data in different tiers:

- Open data
- Safeguarded data
- Controlled / Secure data
- Source data

Data protection

- Open data: data which are freely available to all for any purpose. Open data are often accessed via basic registration and download.
- Safeguarded data: data to which access is restricted due to license conditions, but where data are not considered 'personally-identifiable' or otherwise sensitive. Access is typically available via a remote service with registration and project approval requirements.
- Controlled data: data which need to be held under the most secure conditions with stringent access restrictions. Access is available via secure services, with registration and project approval requirements.

Data protection

- License agreements: these stipulate whether the data can be made available via the Open, Safeguarded or Secure services. It also sets out the conditions of use, for example it may limit the use to research for academic purposes only or look to accommodate the commercial interests of the data provider.
- Research Approvals Groups: conducting reviews for project proposals.
- Safe Results: statistical disclosure control, typically by one or two independent approved researchers.

The screenshot shows a web browser window for the UK Data Service. The URL in the address bar is ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/. The page title is "Apply to access controlled data in SecureLab". The header includes the UK Data Service logo, a search bar, and a "Login" button. A purple navigation bar at the top has links for "Find data", "Deposit data", "Learning hub", "Training and events", "About", "News", "Impact", "Help", and "Contact". Below the header, a breadcrumb trail shows "Home > Find data > Access conditions > Apply to access controlled data in SecureLab". The main content area features a large image of a leaf's vascular pattern. A prominent message reads: "COVID-19 update: To enable continued research access during the pandemic, please follow the normal SecureLab application process. Following project approval, you may apply for temporary home-working access to specific datasets through an additional short form and agreement." Below this, a link says "Information on how to apply is available on our [Covid-19 SecureLab home-working page](#)". A dropdown menu titled "SecureLab application requirements" is open. At the bottom, a section titled "Which secure data application process?" contains a note about pathway variations.

COVID-19 update: To enable continued research access during the pandemic, please follow the normal SecureLab application process. Following project approval, you may apply for temporary home-working access to specific datasets through an additional short form and agreement.

Information on how to apply is available on our [Covid-19 SecureLab home-working page](#).

SecureLab application requirements

Which secure data application process?

Different application pathways set by data providers and legislative requirements mean there are slight variations in application processes. Please check carefully which pathway you need to follow.

The screenshot shows a web browser window for the UK Data Service. The URL in the address bar is beta.ukdataservice.ac.uk/databatalogue/studies/study?id=7481. The page title is "UK Data Service > Study". The main content is about the "Integrated Census Microdata (I-CeM), 1851-1911" study (Study number 7481). The "Details" tab is selected, showing the following information:

Title:	Integrated Census Microdata (I-CeM), 1851-1911
Alternative title:	I-CeM
Study number (SN):	7481
Access:	These data are safeguarded
Persistent identifier (DOI):	10.5255/UKDA-SN-7481-2
Data creator(s):	Schurer, K., University of Essex, Department of History Higgs, E., University of Essex, Department of History

Below the details, there is a section for "Sponsors and contributors" which is currently empty.

The screenshot shows a web browser window for the UK Data Service. The URL in the address bar is beta.ukdataservice.ac.uk/databatalogue/studies/study?id=7856. The page title is "UK Data Service > Study". The main content area displays a study titled "Integrated Census Microdata (I-CeM) Names and Addresses, 1851-1911: Special Licence Access" (Study number 7856). The study details include:

Title:	Integrated Census Microdata (I-CeM) Names and Addresses, 1851-1911: Special Licence Access
Alternative title:	I-CeM
Study number (SN):	7856
Access:	These data are safeguarded
Persistent identifier (DOI):	10.5255/UKDA-SN-7856-2
Data creator(s):	Schurer, K., University of Essex, Department of History Higgs, E., University of Essex, Department of History

On the left sidebar, there are links for "Studies" and "Series". Below the study details, there is a "Copy study DOI" button. At the bottom of the page, there is a footer with the URL <https://beta.ukdataservice.ac.uk>.

Office for National Statistics

A screenshot of a web browser displaying the Office for National Statistics (ONS) website. The URL in the address bar is ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/accessthedatasecurely. The page title is "Access the data securely - Off". The header includes the ONS logo, language links (English (EN) | [Cymraeg \(CY\)](#)), and navigation links (Release calendar, Methodology, Media, About, Blog). A search bar at the top has the placeholder "Search for a keyword(s) or time series ID" and a green search icon. Below the search bar is a purple banner with the text "census 2021 Data and analysis from Census 2021". The main content area has a grey header with the title "Access the data securely" and the subtitle "How to access our service and data securely.". Underneath is a section titled "In this section" with two columns of links:

1. Accessing the Secure Research Service (SRS)	4. Assured Organisational Connectivity (AOC)
2. Safe Rooms	5. Contact details
3. SafePods Network	

At the bottom left is a call-to-action button labeled "1. Accessing the Secure Research". At the bottom right is a link labeled "Related downloads".

A screenshot of a web browser displaying the 'HMRC Datalab datasets available' page on the GOV.UK website. The page is under the 'Guidance' section and is titled 'HMRC Datalab datasets available'. It provides information about datasets available in the HM Revenue and Customs (HMRC) Datalab. The page includes details from HM Revenue & Customs, published on 14 November 2014, last updated on 17 April 2019, and a link to see all updates. There are also buttons for getting emails about the page and printing it. A note at the bottom states that the team is working to add new datasets. On the right side, there is a 'Related content' section with links to 'Research at HMRC' and 'Compliance Perceptions Survey 2011'.

HMRC Datalab datasets available

Information about the datasets that are currently available in the HM Revenue and Customs (HMRC) Datalab.

From: [HM Revenue & Customs](#)
Published 14 November 2014
Last updated 17 April 2019 — [See all updates](#)

[Get emails about this page](#)

[Print this page](#)

The HMRC Datalab Team is working to keep adding new datasets to those available for use in the HMRC Datalab. This page is updated when these become available.

Related content

[Research at HMRC](#)
[Compliance Perceptions Survey 2011](#)

Consumer Data Research Centre

The screenshot shows a web browser window for the CDRC Data search interface at data.cdrc.ac.uk/search/type/dataset. The page title is "Search | CDRC Data". The header includes the CDRC logo, an "An ESRC Data Investment" badge, a search bar, and navigation links for "CDRC", "Datasets", "Stories", "Tutorials", "Topics", "Geodata Packs", "About Data", "Log in", and "Register".

The main content area shows a breadcrumb navigation "Home » Dataset » Search". On the left, there is a sidebar with filters for "Content Types", "Topics", "Type", and "Controller". The "Content Types" filter is expanded, showing "Dataset" selected. The "Topics" filter is expanded, showing "Population & Mobility (49)", "Retail Futures (22)", "Finance & Economy (13)", "Transport & Movement (9)", and "Digital (6)". The "Type" filter is expanded, showing "Open (43)", "Safeguarded (30)", and "Secure (14)". The "Controller" filter is expanded, showing "University College London (UCL) (57)" and "University of Leeds (16)".

On the right, there are search and sort controls: a search input field, a "Sort by" dropdown set to "Relevance", and buttons for "Descending", "Apply", and "Reset". Below these are "87 results". Two datasets are listed:

- High Street Retailer - Retail and Consumer Data (2012 - 2017 only)** Secure
This dataset contains information on customer and retail characteristics and transactions from a specific non-grocery retailer chain which has a presence on a number of the main shopping streets in the UK as well as in some other settings (...)
- Airbnb Property Rentals and Reviews (supplied by AirDNA)** Safeguarded
This data profile describes a dataset held by the CDRC which has been supplied by

Consumer Data Research Centre

- The Consumer Data Research Centre was established in 2014 to lead academic engagement between industry and the social sciences and utilise consumer data for academic research purposes.
- Led by leading UCL Academics together with Leeds, Liverpool, Oxford.
- Focus on consumer data – i.e. large-scale human-generated datasets.
- Access through several data licensing agreements with industry partners.
- Funded till at least September 2024.

Data Products

Examples:

- WhenFresh/Zoopla Property Transactions (2014-2019)
- Customer and Ticket Sales data from a regional transport provider
- Analysis-ready products: indices in relation to population (e.g. ethnicity estimates, population churn, residential mobility), retail centre boundaries, geodemographic classifications.

Data Products

The screenshot shows the GBNames web application. At the top, there is a header bar with the title "GBNames" and the URL "apps.cdrc.ac.uk/gbnames/". Below the header is a navigation bar featuring the Consumer Data Research Centre logo, an "An ESRC Data Investment" badge, a search bar with a "Family name" input field and a "Search" button, and a dark blue menu bar with links for "GBNames", "Methodology", and "About".

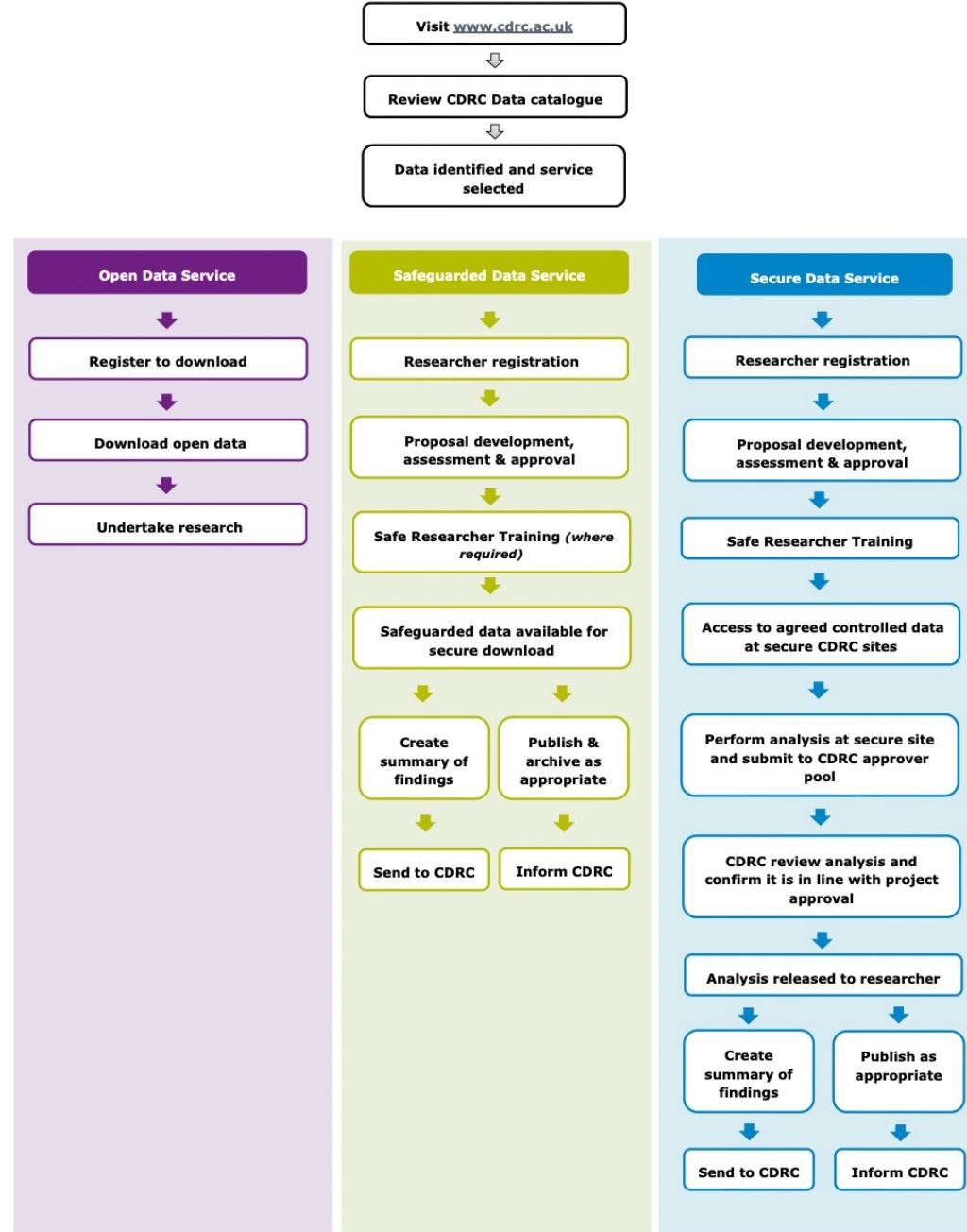
The main content area has a light blue background. It features a large "Welcome to GBNames" heading and a descriptive paragraph about the research conducted at the ESRC Consumer Data Research Centre (CDRC) and the UCL Urban Dynamics Lab. This paragraph discusses the exploration of generational and inter-generational residential movements of family groups across Great Britain using historic censuses and recent consumer registers. It highlights the CDRC's novel consumer datasets and their focus on geographical and social mobility.

Below this welcome message is a form titled "What's your family name?" containing a "Family name" input field and a "Search" button. A note below the form states that GBNames records surname searches for statistical purposes and directs users to the Consumer Data Research Centre website for more information on available data products, related maps, and access details.

At the bottom of the main content area is a map of the United Kingdom showing the locations of major cities: Aberdeen, Glasgow, Bradford, Manchester, and Stoke-on-Trent. On the left side of the map are zoom controls (+, -, ×).

www.apps.cdrc.ac.uk/gbnames

Pipeline



How does data get into the data catalogue?

- Confidential and sensitive data will require some form of informed consent.
- Datasets need be cleaned, prepared, clear variables – typically no raw data dumps.
- Documentation needs to be included on how the data set was constructed (e.g. links from variables to questions in the questionnaire, codebook, description of data linkage).
- Access and licensing conditions need to be specified.
- ESRC grant holders are **contractually obliged** to offer their data for deposit with a responsible digital repository within three months of the end of their grant.

How is data secured?

- Physical secure lab at an undisclosed location with specialised access procedures
- Online facilitated research environments ("Trusted Research Environments")
- ISO27001 certified

Data Safe Haven

- UCL's facility for Secure Research
- A technical solution for storing, handling and analysing identifiable data
- 'Walled garden' approach where research stays within a secure environment with carefully controlled access
- Project-based
- Safe-researcher training required
- Output is controlled

Data Safe Haven

What not to do:

- Using data for which you are not licensed
- Using data for anything other than the proposed project
- Linking or matching data without permission
- Handing out usernames and passwords to others
- Attempting to identify individuals, households, or firms
- Copying anything from the screen
- Writing down anything from the screen

Data Safe Haven

What DSH offers:

- Several pre-installed software programmes (python, R, Stata, SAS, SPSS, NVIVO)
- Following a recent refresh: dedicated HPC VMs / queue-based shared HPC facilities
- Dedicated PostgreSQL or MySQL databases
- Local copies of CRAN, pypi and conda

Limitations

- Very technological focus where the data themselves is not questioned.
- Still partly focused on 'traditional' ways of data collection and analysis, not always provision for large datasets.
- A confusing landscape of data providers and research facilities.

Conclusion

- Safe Research using privacy sensitive data predominantly focuses on conducting research in a safe research environment.
- Data Services tend to offer data in a graduated manner (tier system), depending on the level of 'sensitivity' of the data.
- Data typically can only be deposited with metadata, documentation, depends on the creator of the data to what extent attention is paid to issues of data and representation.

Seminar preparation

There is no preparation required for this week's seminar other than carefully reading the articles on the reading list. Use the remainder of your time to work on your coursework assignment.

Seminar preparation



Formative feedback

- Formative feedback on a 300-word proposal / outline.
- Suggestions to include in the proposal / outline: provisional title, aim, proposed structure with main arguments; bullet points are perfectly fine.
- Submit by: Wednesday December 7 at noon (2 weeks) – details TBC.
- Feedback by: Friday December 16.

Formative feedback



Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk

