Linked Digital Trace and Survey Data for Secondary Research

Potential and Constraints

Dr. Conor Gaughan
Postdoctoral Research Associate
Cathie Marsh Institute for Social Research (CMI)
University of Manchester

conor.gaughan@manchester.ac.uk

As part of the UKRI ‘smart data’ accelerator project **DIGISURVOR**

Contents

- Introduction to DIGISURVOR
- **Session 1**: Linking Survey and Digital Trace Data (30 minutes)
  - What is data linkage?
  - How is data linkage used in survey research?
  - What are the strengths and challenges of survey-to-digital data linkage?
  - What challenges does survey-to-digital data linkage pose for open research?
  - Q&A
- **Session 2**: Processing Linked Data for Open Research (30 minutes)
  - How does data linkage increase the risk to disclosure?
  - How do we assess these risks?
  - How might we reduce this risk in relation to linked survey to social media data?
  - How might we reduce this risk in relation to linked survey to web browsing data?
  - Q&A
- Mentimeter (topics for future webinars)
- Conclusion

DIGISURVOR Team

Linking **DIGI**tal and **SURV**ey data for **O**pen **R**esearch

Smart Data Research UK

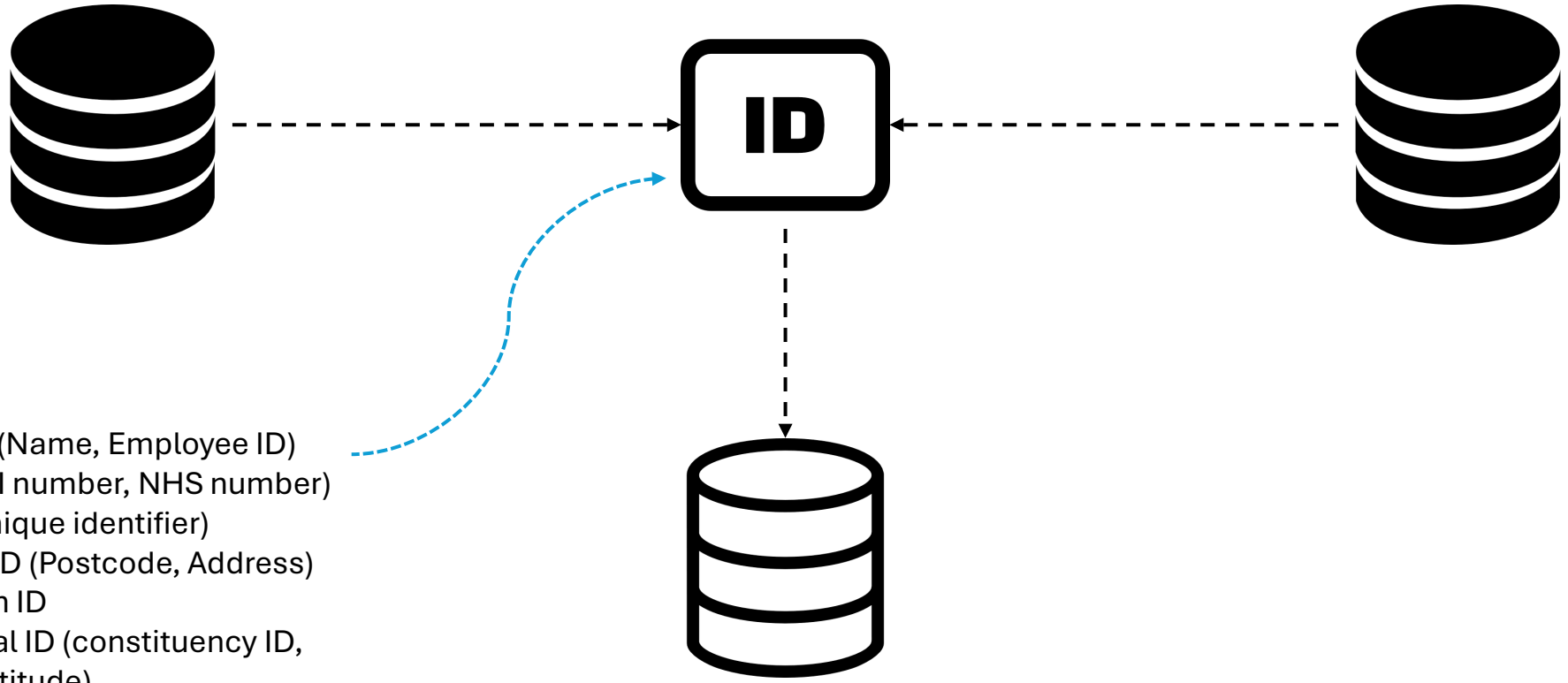
UK Research and Innovation

Session 1

Linking Survey and Digital Trace Data

What is Data Linkage?

"Data linkage is the activity of bringing together separate data sources by identifying and matching the same entity in each and then bringing those different sources of information together into a single dataset." – **Understanding Society (UKLS)**



- Personal ID (Name, Employee ID)
- Admin ID (NI number, NHS number)
- Study ID (Unique identifier)
- Household ID (Postcode, Address)
- Organisation ID
- Geographical ID (constituency ID, longitude/latitude)

Data Linkage: Research Use Cases

Healthcare

How does **socioeconomic background** affect the risk of **chronic diseases** such as Type 2 diabetes?

Deprivation Index (IMD) <-> NHS Health Records

Crime

How do childhood **educational outcomes** affect the likelihood of **future offending**?

Attainment Records (DfE) <-> Criminal Records (MoJ)

Education

Does participation in **higher education** affect long-term **employment outcomes**?

University Enrolment (UCAS) <-> Employment Records (HMRC)

Economic and Social Policy

How does investment in **local community programmes** affect **mental health**?

Local authority data <-> NHS Health Records

Politics

How does exposure to **online party campaigning** affect **voting behaviour**?

Survey Data (BES) <-> Social Media Data

Data Linkage in Survey Research

Four Eras of Survey Research (Groves et al. 2011; Burke-Garcia et al. 2021; Hill et al., 2021)

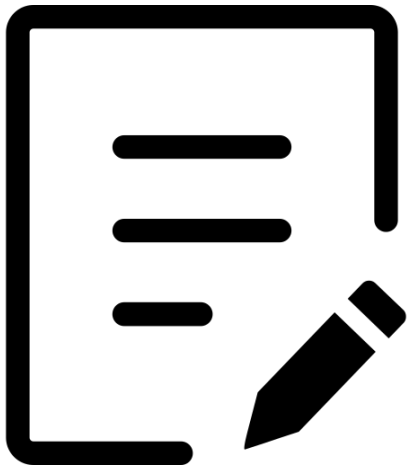
- Probability sampling
- Structured questionnaires
- Standardised interviews
- Statistical inference

- Large national surveys
- Telephone surveying
- Longitudinal studies
- Government funded

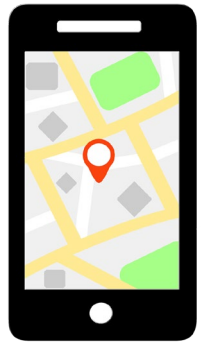
- Online surveying
- Faster data processing
- Declining response rates
- Increasing cost

- Mixed method surveys
- Online panels
- Big data
- External linkage

Linked Survey to Digital Trace Data



Social Media



Mobile/App Usage

Wearable Trackers

Web History

Online Transactions

Streaming History

All images free for use under the Pixabay Content License

SURVEY DATA

- Representative Sampling
- Tailored Questions
- Anonymity
- Self-Reported Information

- Sampling Biases
- Measurement Biases
- One-Dimensional

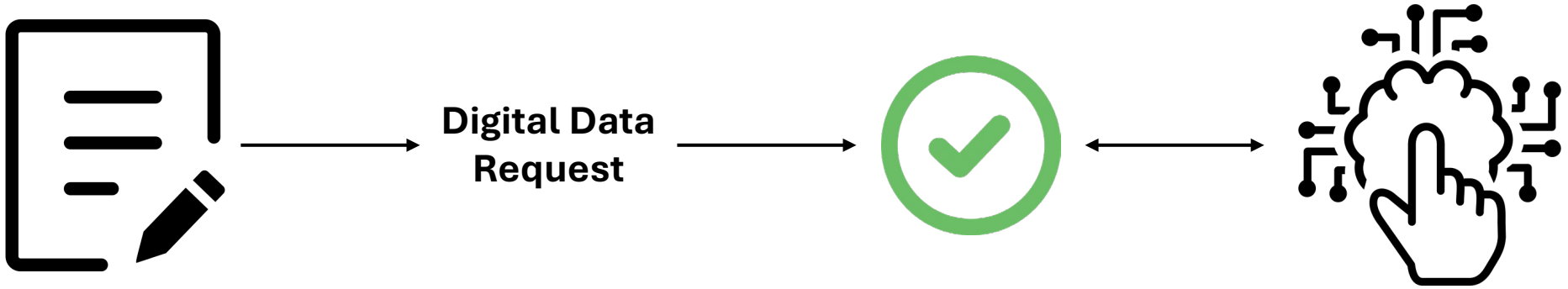
DIGITAL TRACE DATA

- Non-Representative
- Unstructured
- Highly Sensitive

- Non-Reactive
- Real-Time Data
- Unlock New Insights
- Publicly Accessible (Sometimes….)

Enhancing Data Through Linkage

Survey with Digital Trace Data (plugging the “say-do” gap)

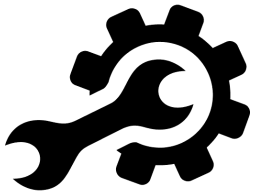


Digital Trace with Survey Data (giving digital data a human face)

Challenges to Digital Data Linkage

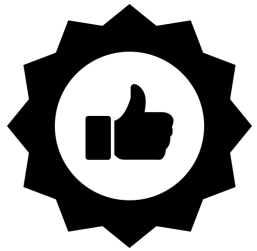
Ethical

- Ensuring informed consent to linkage
- Legal restrictions to digital data processing
- Increased risk of re-identification



Methodological

- Lack of common identifiers for linkage
- Can require advanced technical skills
- Digital data can often be large and messy



Data Quality

- Coverage bias: uneven usage of digital technologies
- Selection bias: certain groups more likely to consent
- Digital data can be inconsistent and incomplete

Linked Datasets for Open Research (FAIR)

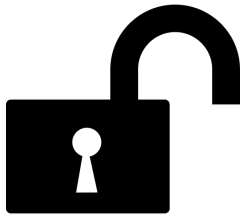
*"**Open research** embodies good research practices by opening up participation in, and access to, the research lifecycle. It covers a wide range of practices and principles related to how research is carried out and enables research to take advantage of technical advances."*

"By improving access to research outputs according to best practices that enable research to be findable, accessible, reuseable and interoperable (FAIR) researchers have more opportunity to:

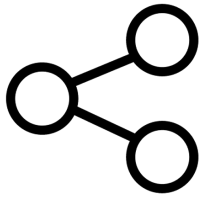
- *engage, replicate and accelerate knowledge discoveries*
- *benefit society and the economy"*

UK Research and Innovation (UKRI)

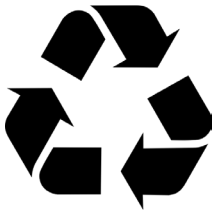
Findable



Accessible



Interoperable



Reusable

UKDS Access Levels

UK Data Service

Tier 1: Open

Users can access open data collections without registration.

Tier 2: Safeguarded

Safeguarded data collection can be downloaded by registering and accepting our End User Licence Agreement.

Tier 3: Controlled

Data that are very detailed, sensitive or confidential are accessed through the UKDS SecureLab, the UK Data Service Trusted Research Environment. After registering, experienced researchers can apply to access controlled data.

Session 1: Q&A

Session 2

Processing Linked Data for Open Research

Safely Working with Survey Data

Participants are fully informed

Data protection requirements are met

Personal data remains confidential

Data is stored securely

"Data should be as open as possible, as closed as necessary." – **FAIR data**

Three Areas of Guidance on Effective Anonymisation:

(**Information Commissioner's Office (ICO)** Anonymisation, pseudonymisation and privacy enhancing technologies guidance)

1) **Identifiability:** Can an individual be directly identified in a dataset, or indirectly "singled out" from other participants?
2) **Linkability:** Can an individual be identified when the data is linked with other external data sources?
3) **Inference:** Is there any potential to reasonably infer, guess or predict personal data based on the data provided?

Assessing the Risk of Identifiability

The "**Motivated Intruder Test**":

(**Information Commissioner's Office (ICO)** Anonymisation, pseudonymisation and privacy enhancing technologies guidance)

How likely is it that a motivated intruder would be able to successfully identify an individual in your anonymised data?

- The test accounts for identifiability risk in the context of the realistic cost of identification by other motivated actors in human, economic, temporal and technological terms

- A motivated intruder would be an individual who is:

✔ Reasonably competent

✔ Access to appropriate resources

✔ Uses investigative techniques

✘ Specialised knowledge

✘ Access to specialist equipment

✘ Resort to criminal activity

Statistical Disclosure Control (SDC)

- SDC is a set of methods and techniques used to reduce and test disclosure risks

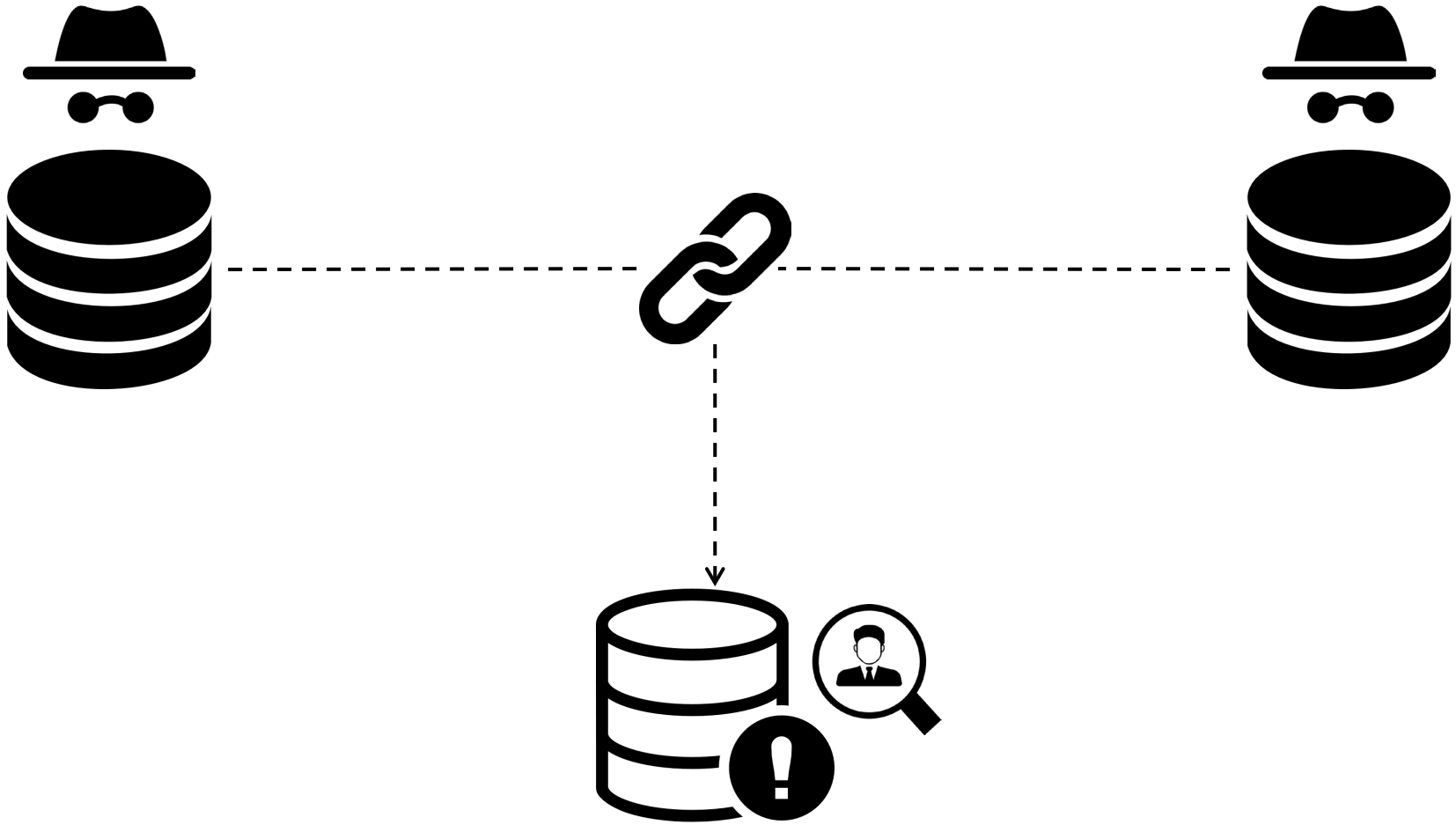
- Even if one anonymises *direct* identifiers, other variables may be what we call *quasi-identifiers*

- Three key SDC assessments:
  - k-anonymity (prevents singling out)
  - l-diversity (prevents linkage attacks)
  - t-closeness (prevents inference)

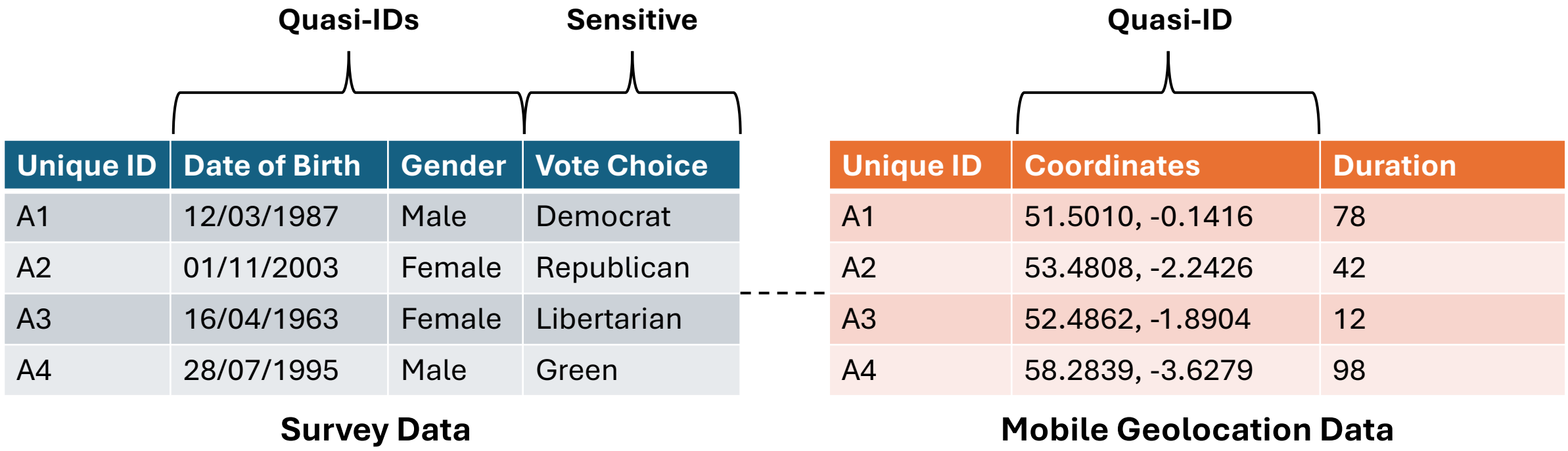
- SDC techniques include binning, aggregation, suppression, randomisation, deletion, and perturbation

- However, the more we **protect** the data, the more **utility** we destroy…

The Disclosure Risks of Linkage

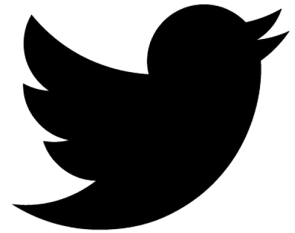


Quasi-Identifiers



DIGISURVOR Linked Datasets (1)

- Part of a previous ERC-funded research project on digital campaigning during elections (DiCED)
- Two surveys taken during US Presidential election campaigns in 2020 and 2024
- Fielded to c. 5,000 - 6,000 respondents
- Survey contained questions relating to online behaviour, media consumption, political attitudes and behaviours, and sociodemographics



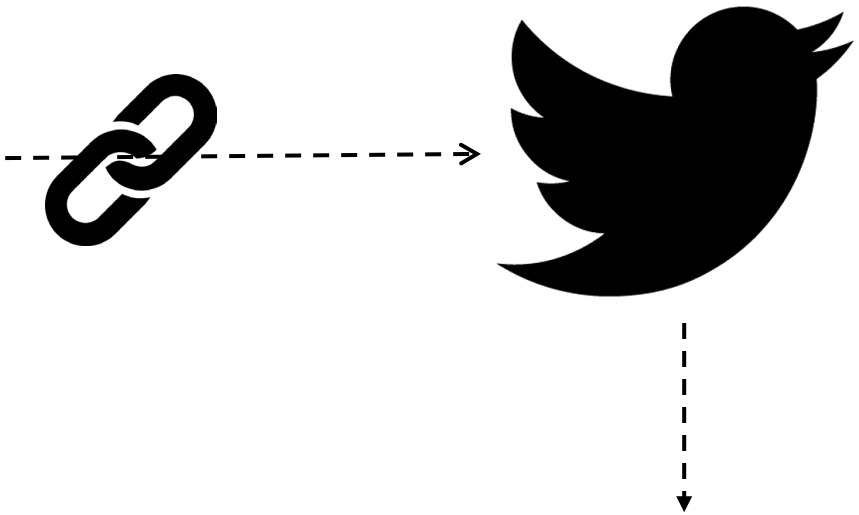
- Participants invited to share their Twitter/X usernames
- Just over one in four participants had an account and agreed to linkage in both surveys respectively
- Their Twitter/X profiles, timelines, follows, and likes were extracted from the Twitter/X platform
- The profiles and timelines of all the accounts that they followed were also extracted

Social media often contains publicly searchable data!

Quasi-IDs | Sensitive

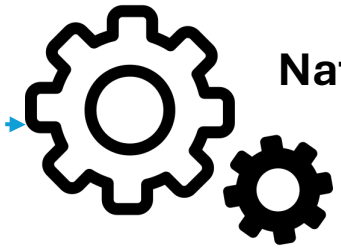
Unique ID	Date of Birth	Gender	Vote Choice
A1	12/03/1987	Male	Democrat
A2	01/11/2003	Female	Republican
A3	16/04/1963	Female	Libertarian
A4	28/07/1995	Male	Green

Survey Data



Unique ID	Name	Username	Location	Tweet
A1	John Smith	@johnsmith2	Dallas, TX	“I can’t believe what’s happened in the news!”
A2	Jane Doe	@jane_doe45	Birmingham, AL	“Just had the worst day at work ever! 😤 ”
A3	Jack Bryan	@jbryan18	Phoenix, AZ	“Nothing going right at the moment 🙄 ”
A4	Wilma Woods	@wwoods	New York, NY	“Anyone watching the tele right now?”

“Just wrapped up a productive day and feeling inspired 🚀✨ Ready to take on what’s next! Sometimes it’s the small wins that matter most—keep going 💪🔥 #Motivation #DailyGrind. Check this out for a boost: https://example.com 🌐”



Natural Language Processing Pipeline

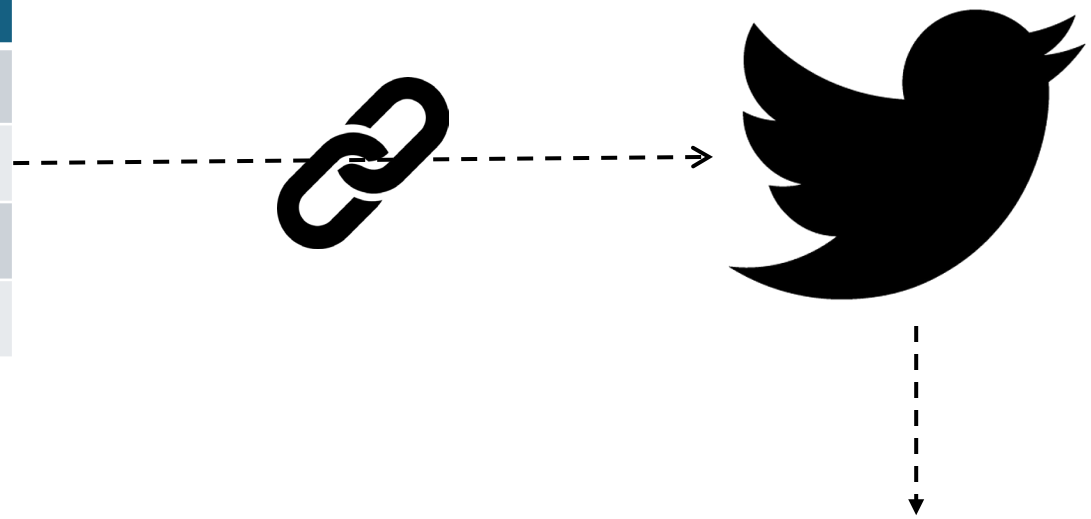
Letters	Emojis	Hashtags	URLs	Sentiment	Topic	Emotion	Sophistication
155	5	2	1	Positive	Health and Lifestyle	Optimistic	Moderate

Quasi-IDs

Sensitive

Unique ID	Date of Birth	Gender	Vote Choice
A1	12/03/1987	Male	Democrat
A2	01/11/2003	Female	Republican
A3	16/04/1963	Female	Libertarian
A4	28/07/1995	Male	Green

Survey Data



Unique ID	Tweet Count	Mean Length	Mean Emojis	Mean Hashtags	% Positive	% Toxic	% Political
A1	40	155	2	2	0.35	0.05	0.25
A2	11,562	63	0	0	0.21	0.12	0.46
A3	2,678	44	1	5	0.57	0.03	0.12
A4	846	236	3	1	0.86	0	0.05

DIGISURVOR Linked Datasets (2)

- Part of a previous research project on information consumption during COVID-19 (Charlemagne Prize Academy Fellowship).
- Survey fielded to 599 UK participants between March and May 2020
- Survey contained questions relating to online behaviour, media consumption, trust, political attitudes and behaviours, and sociodemographics

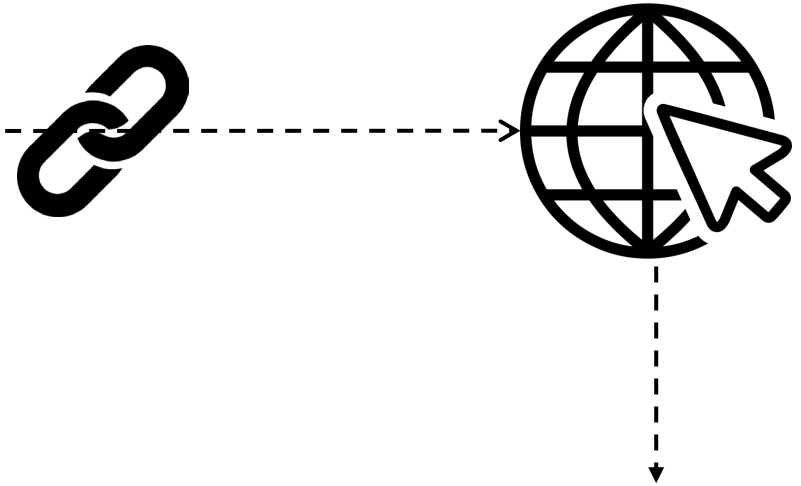
- Web trackers installed on all participants' mobile devices over a period of 9 weeks between March and May 2020
- All URLs visited by participants during the observation period were stored by the tracker
- Web-tracking data includes complete URL links (domain + path), as well as a date-time stamp and duration spent on the webpage

Web browsing history can be extremely granular!

Quasi-IDs | Sensitive

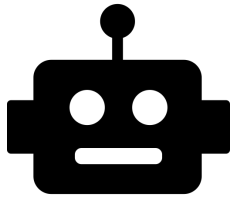
Unique ID	Date of Birth	Gender	Vote Choice
A1	12/03/1987	Male	Conservative
A2	01/11/2003	Female	Labour
A3	16/04/1963	Female	Green
A4	28/07/1995	Male	Reform UK

Survey Data



Unique ID	Date	URL Link	Duration
A1	25/03/2025	https://www.theboltonnews.co.uk/news/	15
A1	26/03/2025	https://www.staffnet.manchester.ac.uk/	22
A1	26/03/2025	https://www.localboltonschool.gov.uk/updates	8
A1	27/03/2025	https://www.northernrailway.co.uk/service-updates	11

URL Link	Domain
https://www.theboltonnews.co.uk/news/	theboltonnews.co.uk
https://www.staffnet.manchester.ac.uk/	staffnet.manchester.ac.uk
https://www.localboltonschool.gov.uk/updates	localboltonschool.gov.uk
https://www.theguardian.co.uk/sport	theguardian.co.uk



OpenAI (GPT), Google (Gemini), Anthropic (Claude)

*Between **76%** - **94%** accuracy depending on model

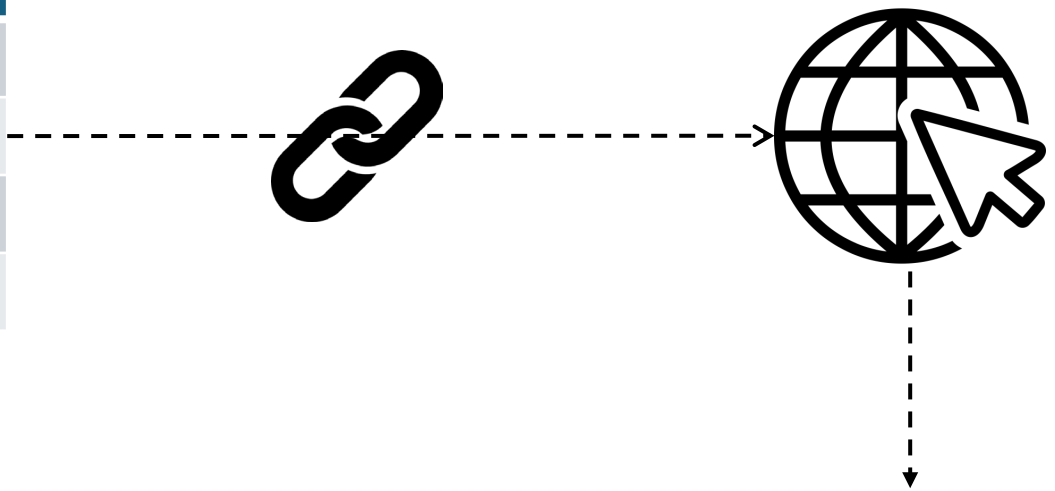
Domain	Domain Classification
theboltonnews.co.uk	Online News Media
staffnet.manchester.ac.uk	Other
localboltonschool.gov.uk	Other
theguardian.co.uk	Online News Media

Quasi-IDs

Sensitive

Unique ID	Date of Birth	Gender	Vote Choice
A1	12/03/1987	Male	Conservative
A2	01/11/2003	Female	Labour
A3	16/04/1963	Female	Green
A4	28/07/1995	Male	Reform UK

Survey Data



Unique ID	Date	Domain Classification	Duration
A1	25/03/2025	Online News Media	15
A1	26/03/2025	Other	22
A1	26/03/2025	Other	8
A1	27/03/2025	Online News Media	11

Conclusions

- Linking survey data to external data can significantly increase the risk of identifiability within a dataset
- The rich, and in some cases publicly accessible, nature of digital trace data makes this an especially risky source of data to link with
- We can enlist several advanced quantitative and computational techniques to help create new fields of interest which can retain utility but reduce the risk of identifiability
- However, we need to remain vigilant about disclosure risks, especially when combining multiple quasi-identifiers together

Final Takeaways

- Linking survey data with other sources of data can help to unlock new avenues for secondary research

- In particular, linkage with digital trace data can give us access to rich new insights into human behaviour and online usage

- However, linkage also poses a number of ethical, technical and methodological challenges which should be accounted for

Thank you!

Dr. Conor Gaughan
Postdoctoral Research Associate
Cathie Marsh Institute for Social Research (CMI)
University of Manchester

conor.gaughan@manchester.ac.uk

As part of the UKRI ‘smart data’ accelerator project **DIGISURVOR**