CC BY

CIVICA RESEARCH

Re-using and repurposing other peoples data

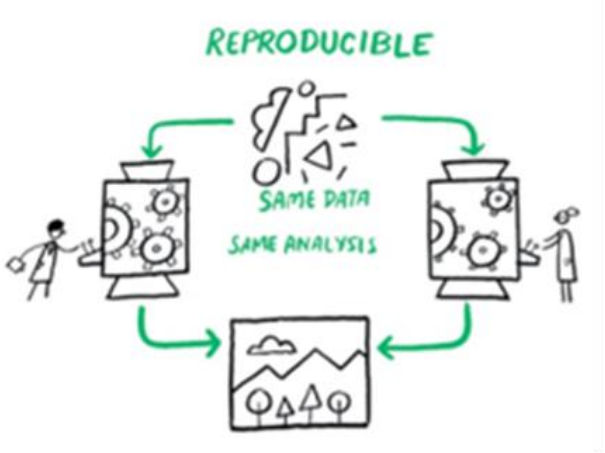
- part of the ***Open Social Science Handbook*** project

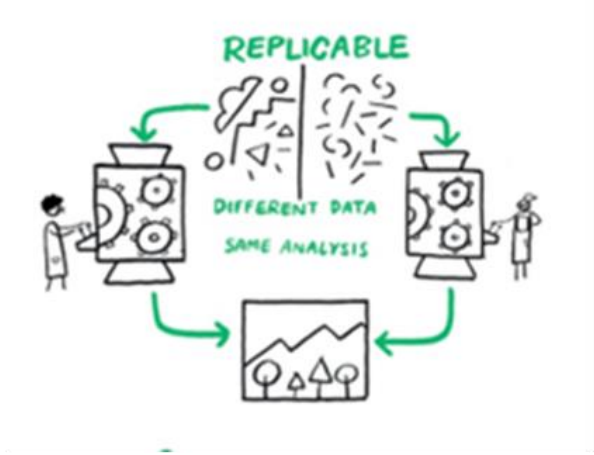
Timothy Monteath (UCL/LSE) and Patrick Dunleavy (LSE)

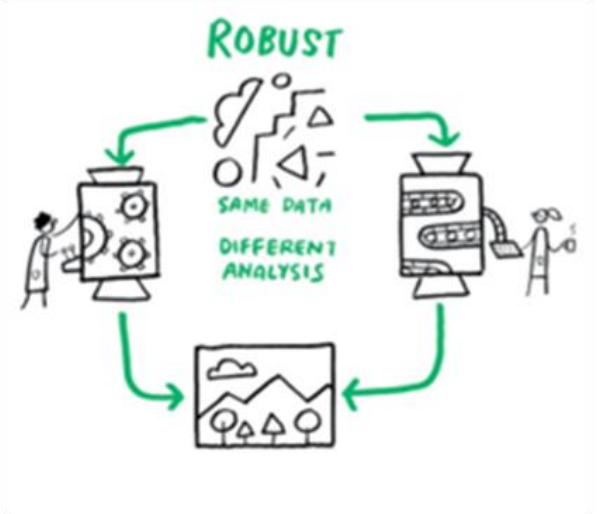
10 November 2022

Bocconi University
Central European University
European University Institute
Hertie School
National University of Political Studies and Public Administration
Sciences Po
Stockholm School of Economics
The London School of Economics and Political Science

Finding and Re-using Data are key for open social science







Source: Turing Way project

Illustration by Scriberia. DOI: 10.5281/zenodo.3332807 Used under a CC-BY 4.0 licence. The component pictograms are slightly re-spaced here.

Commonly used reactive data sources

CIVICA RESEARCH

- Reactive measures depend on how the research subjects respond to questions or experimental stimuli. The unit of analysis is the individual person or organization responding
- *Unique survey dataset*, collected by an individual researcher or team and either self-deposited (but not in an archive) or supplied directly in an article annex or via email
- Similar but *deposited at a formal research archive*
- Single team, unique in-person *experimental dataset*
- Over time, *repeated single-country dataset series*, chiefly common question wordings & demographics, consultation on questions
- *Cross-national, concerted dataset series*, including common common question wordings, set by committee

Questions to ask about reactive data sources

CIVICA RESEARCH

- **RELEVANCE** to your research question – centrality for your work, importance for your conclusions, level of your dependence once started
- **MEANINGS** of variables to be used – critically check how closely a detailed question wording, or a precise demographic code, or combinations of them, identify your key phenomena. If they are proxy variables, how good or imperfect a proxy are they?
- **LIMITATIONS** of the survey/experiment in terms of who respondents are, how many are involved, statistical power of analysis
- **PROVENANCE** – how was the survey or experimental data created, using what methods and level of care in research?
- **REPUTATION** – how is the original source person or organization regarded in academia? Are they seen as biased by other 'tribes'

Most used non-reactive data sources

CIVICA RESEARCH

- Non-reactive measures track the behaviour of research subjects' objective or other objective indices, without depending on their responses or involvement (may be 'unobtrusive measures')
- *Unique dataset of statistics/codings*, collected by an individual researcher or team and self-deposited (not in an archive) or supplied directly in an article annex or via email
- Similar dataset *deposited at a formal research archive*
- *Social media data* downloaded from the APIs of key firms
- *Administrative data* produced by firms or agencies
- *Data from an official statistics organization*

Questions to ask about non-reactive data sources

CIVICA RESEARCH

- **RELEVANCE** to your research question – centrality for your work, importance for your conclusions, level of your dependence once started
- **MEANINGS** of variables to be used – critically check how closely the detailed measurement of a behaviour, or accompanying contextual or demographic data, identify your key phenomena. If they are proxy variables, how good or imperfect a proxy are they?
- **LIMITATIONS** of the dataset in terms of the size and coverage of the data involved, time/context shaping ecology factors, breadth of unit or country coverage, statistical power of analysis
- **PROVENANCE** – how was the data created, using what methods and with what coding accuracy?
- **REPUTATION** – how is the original source person or organization regarded in academia? Are they seen as biased by other 'tribes'

Finding Data

CIVICA RESEARCH

Speak to librarians

Ask around amongst researchers

Ask on Twitter or other social media

Developing a research question

What can you do with the dataset?

AND

How can this fit with my research interests and trajectory?

Finding Data 1 – As an explorer

CIVICA RESEARCH

- See what data is available, keep an open mind
- Start by looking for datasets rather than with a specific research question
- Search widely !

- Where possible download the datasets and explore them before deciding on your research question

- How could datasets speak to other research in your field?
- Think about how datasets could be combined with other data

Finding Data 2 – As a Reverse engineer

CIVICA RESEARCH

- Picture your perfect dataset:
  - What does it contain?
  - How was the data been collected?
  - How would you use it?
- The perfect dataset will never exist!
- Search for things that approximate or are similar to your perfect dataset
- How can you modify, combine or add to existing datasets to get them as close as is (reasonably possible) to your perfect dataset
- What compromises can you make? What compromises would undermine your research?

Exploring a dataset

CIVICA RESEARCH

- How big is the dataset?
  - Is it human sized?
  - Does it need to be parsed programmatically?
- How well is it documented?
  - Has it been used by other researchers?
  - Is the data collection process documented?
  - Are the variables and what they mean explained?
- What's missing?
  - How does the dataset deal with missing values?
  - Do missing values matter to you?
  - Can they be corrected?

Working with a dataset

- Descriptive
- Causal
- Predictive
- Methodological

Data Cleaning

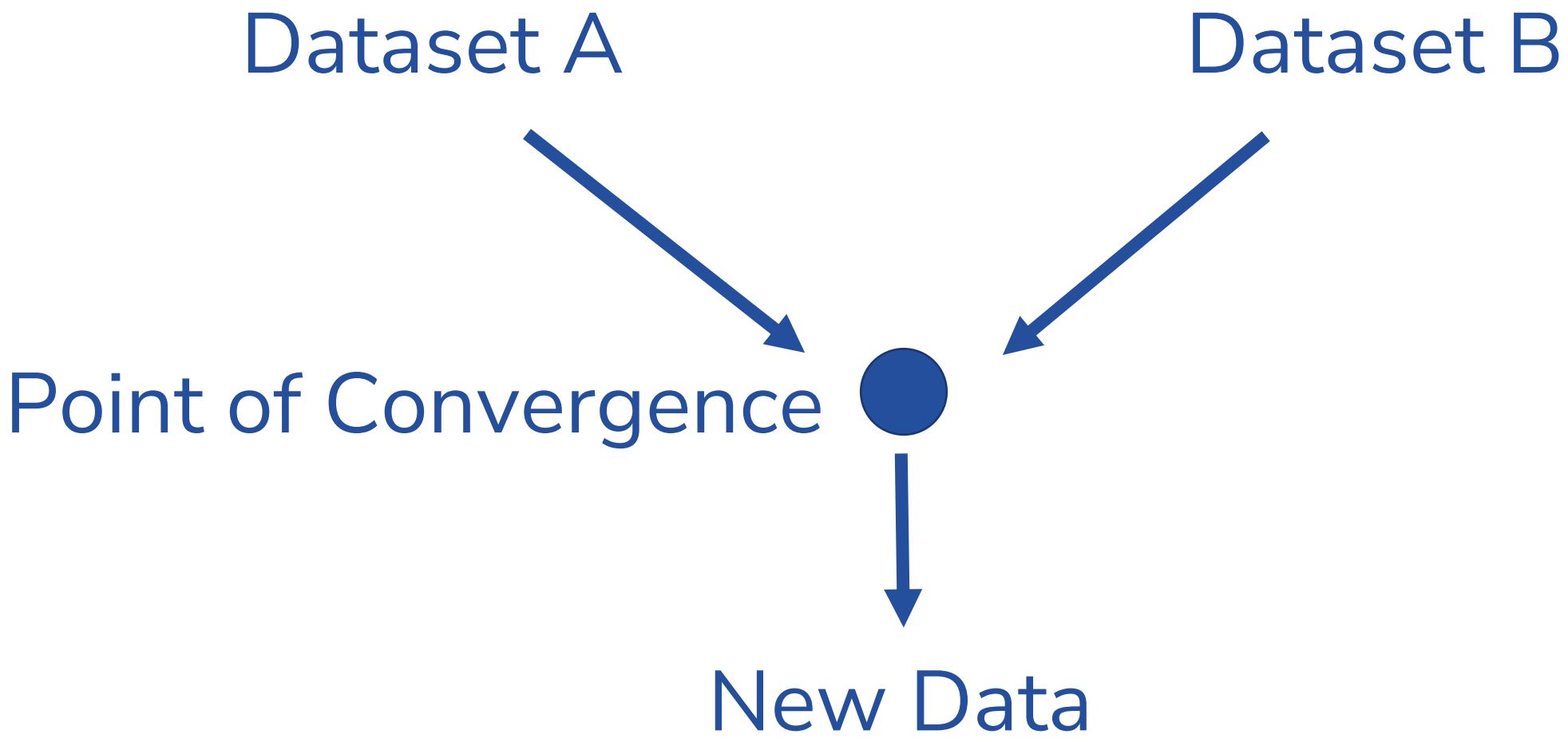
CIVICA RESEARCH

- This will take up most of your time!
- As a rubric you will spend:
  - 80% of your time getting and cleaning the data
  - 15% analyzing the data
  - 5% turning this into a research output

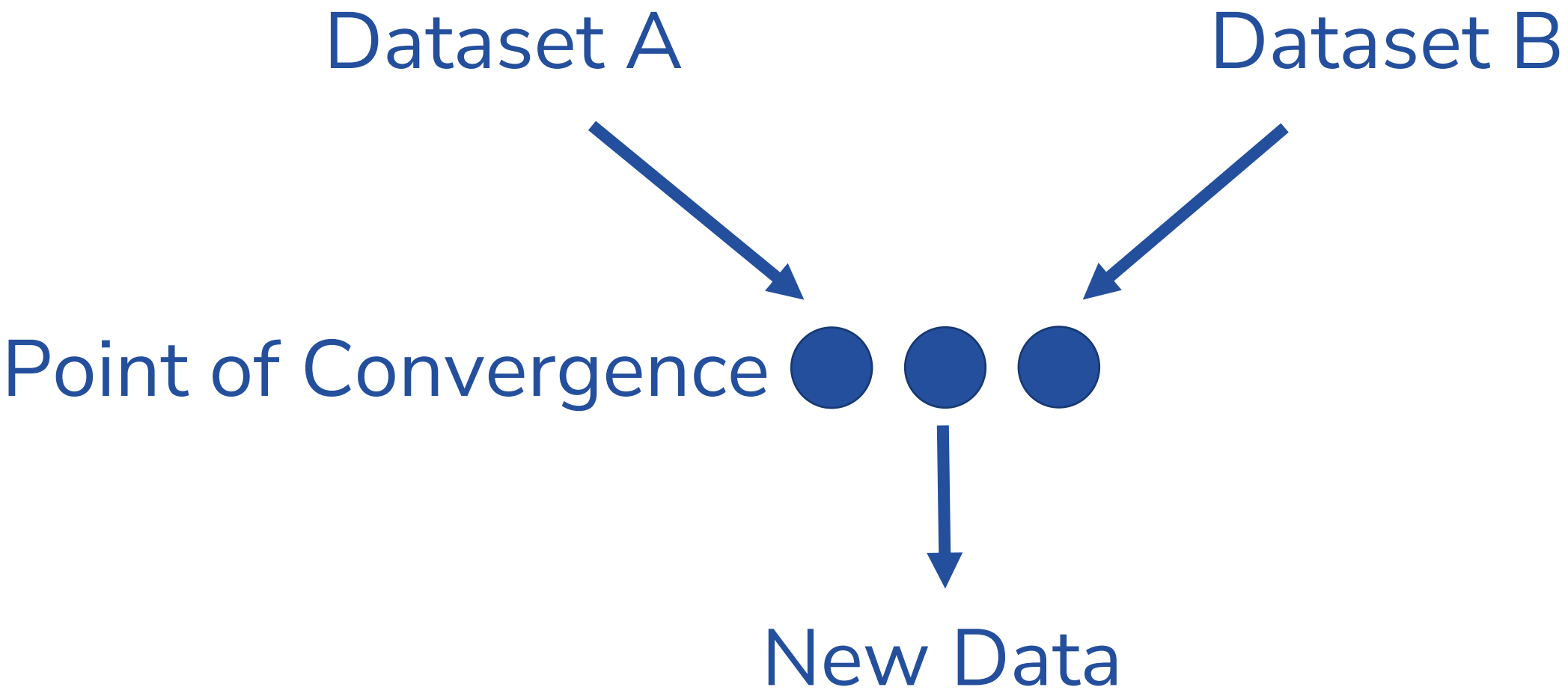
- Data cleaning is the most important part of the work

- Without clean, useable and reliable data there is no research!

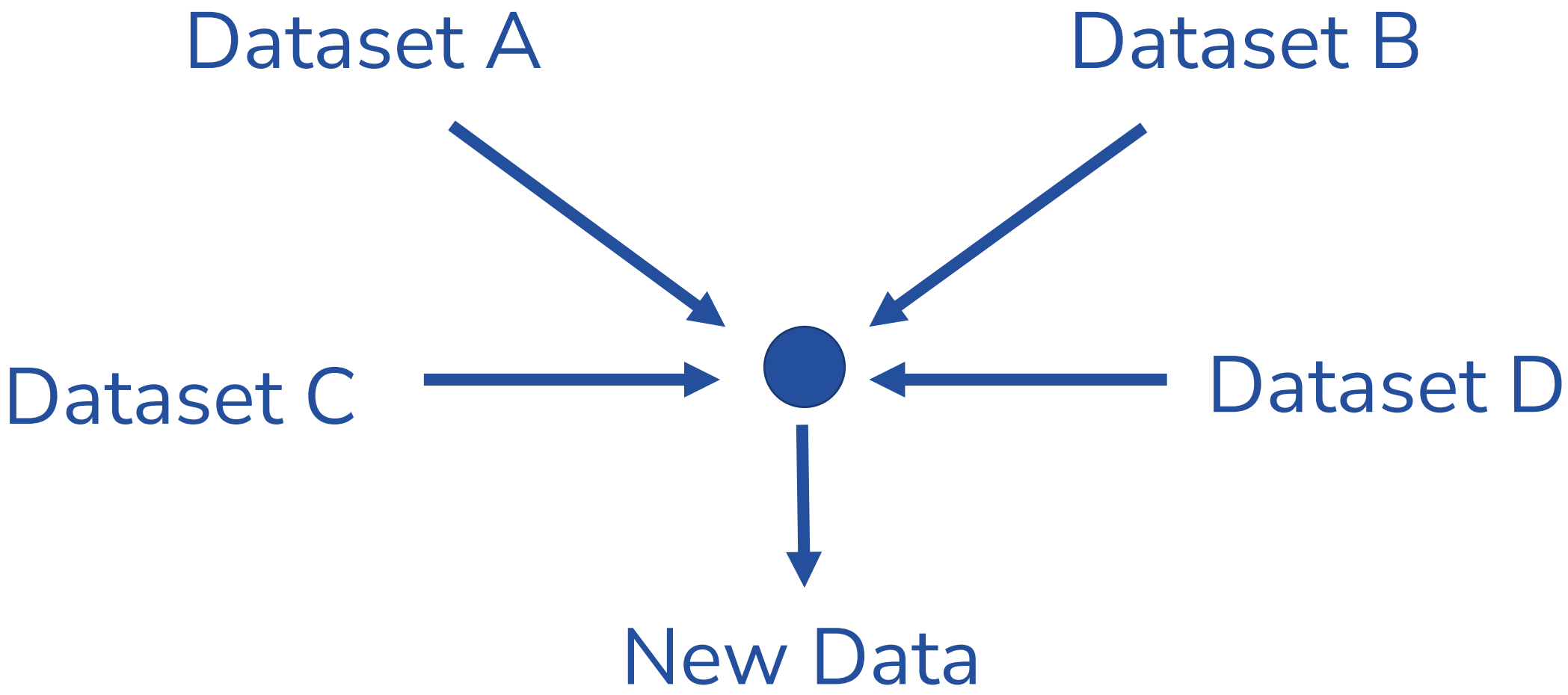
Data Merging



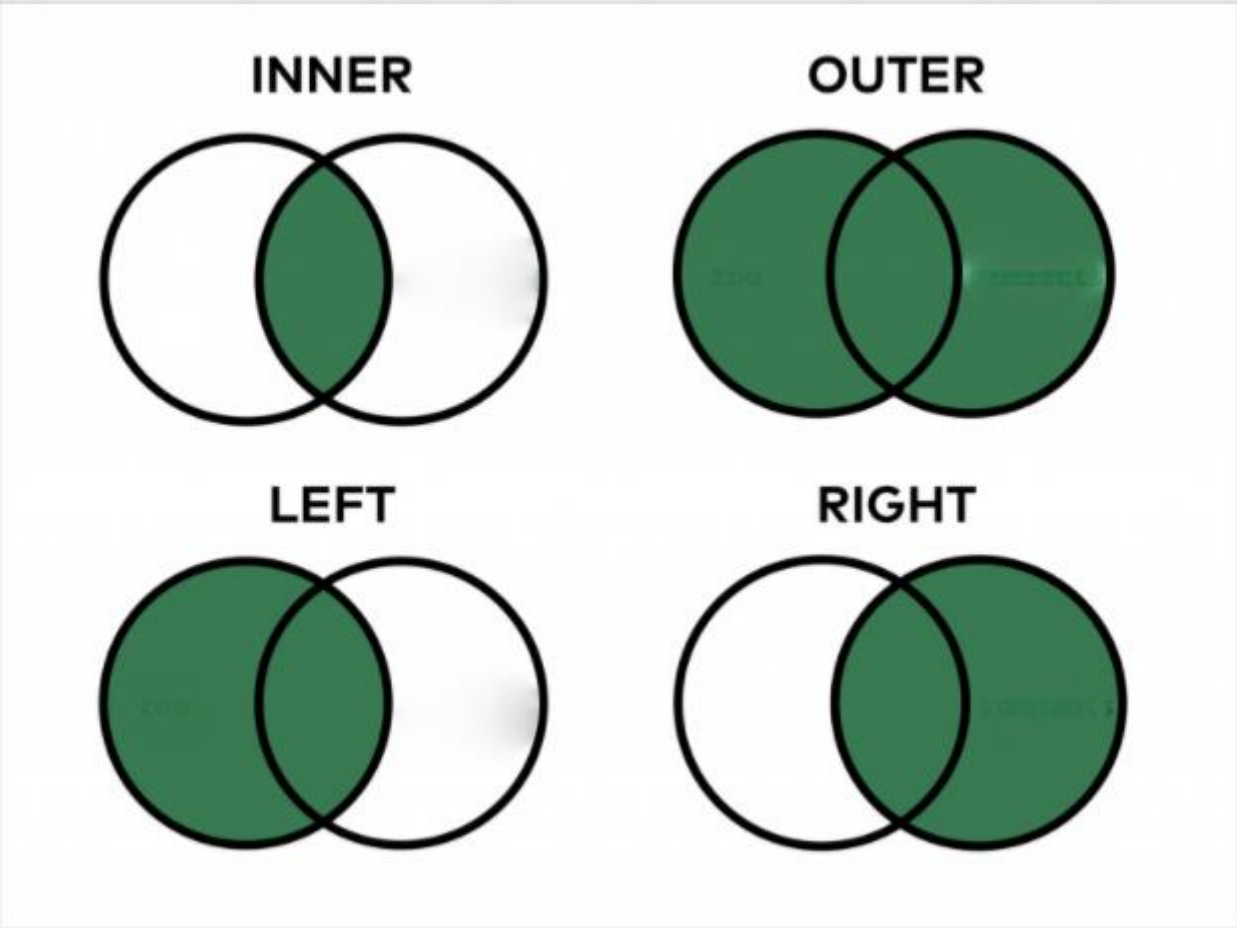
Data Merging



Data Merging



Merging – Types of merging



Merging Data – Multiple Iterations

CIVICA RESEARCH

- Sometimes there can be multiple possible different ways to merge a dataset
- You do not have to pick one, you can try all of them

- Before the merge assign a unique identifier to each row
- Then try all combinations of the merge
- Once complete use the unique identifier to keep only one of the matches

Merging Data – Multiple Iterations

For example:

Dataset 1 – Merge on	Dataset 2 – Merge om
Address 1	Full Address
Address 2	Full Address
Address 3	Full Address
Address 1 + 2	Full Address
Address 1 + 3	Full Address
Address 2 + 3	Full Address
Address 1 + 2 + 3	Full Address

Fuzzy Matching

- Fuzzy matching is a way to match data points that are similar but not identical
- Can be useful for words/names/places with different spellings or mistakes in data input

```
In [28]: from fuzzywuzzy import fuzz

In [29]: fuzz.ratio("Timothy Monteath","Timothy W Monteath")
Out[29]: 94

In [30]: fuzz.ratio("Monteath Timothy","Timothy Monteath")
Out[30]: 50

In [31]: fuzz.token_sort_ratio("Timothy Monteath", "Monteath, Timothy")
Out[31]: 100
```

Regular Expressions

- Extremely powerful!
- Allow you to match and modify any form of text based based on complex expressions

REGULAR EXPRESSION

```
r" .\ [0-9]{1,2}: [A-Z]
```

TEST STRING

```
Testing String. 1 not formated correctly
Testing_String. 2: Formated correctly
Test.String. 33333: too long
test0string. 44: Formated correctly
test string. 00: formated correctly
```

- After a full stop match between one and two digits – no more - followed by a colon a space and an upper case character
- Many sites to help you build regex's for example: https://regex101.com

Example - Investigating the super-rich

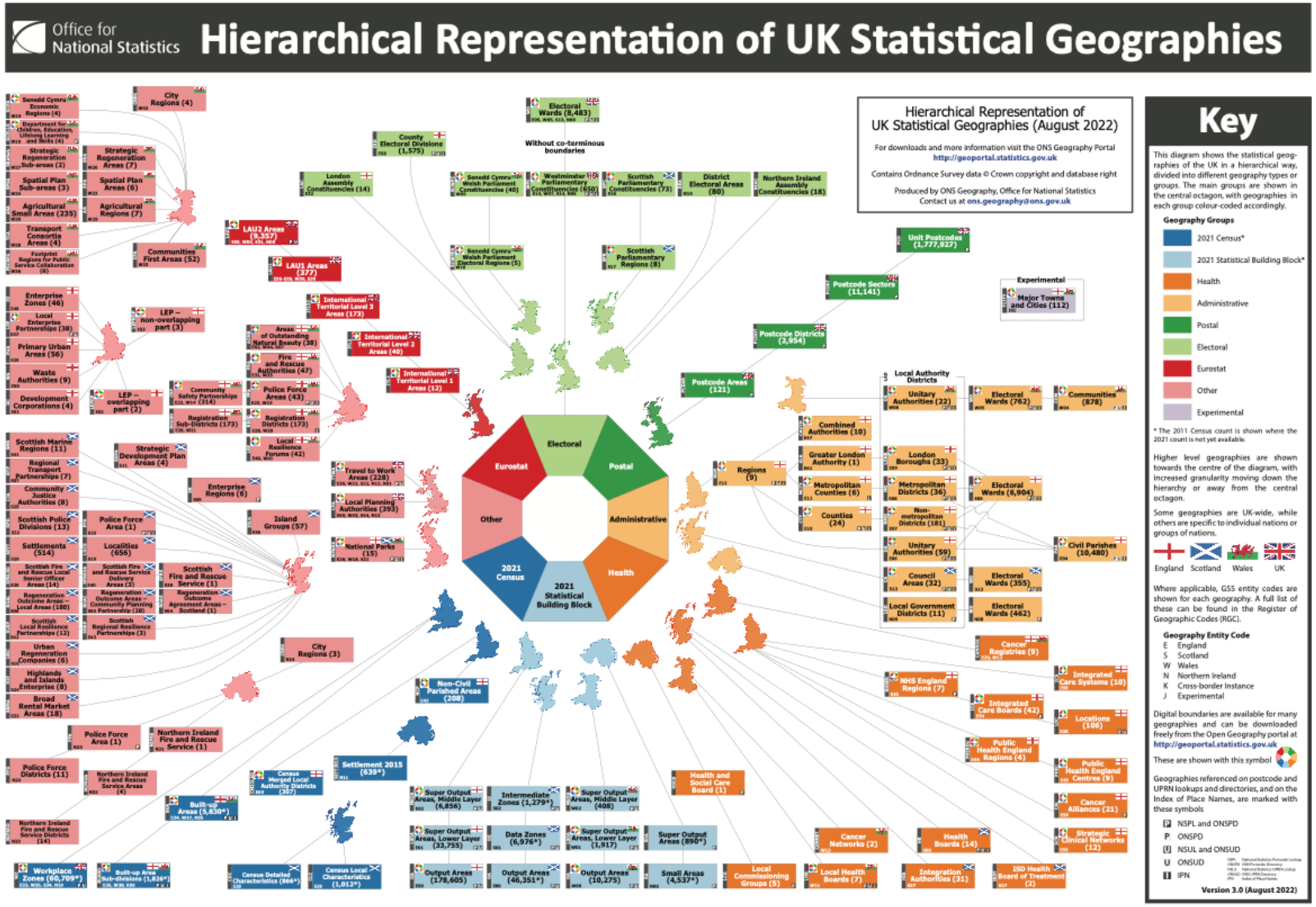
CIVICA RESEARCH

- Sunday Times Rich List of 1000 wealthiest individuals in the UK
- Charity Commission data on all trustees of UK charities
  - 100,000 plus names, no standardization
- Word count analysis of names fields to spot common non-name patterns i.e. Mr, Miss, Doctor, Rev etc
- Regular expression to match common non-name patterns but not catch names that may contain same matters, i.e. Missthorpe
- Fuzzy matching to names on the rich list, calibrate metric by testing
- Resulting in 6,000 matches that were then hand checked for accuracy

Upscaling

- Grouping together variables to align them between datasets by aggregating them
- Works well for temporal and spatial variables
- For spatial variables think about the many different scales and areas you can work within to make the data fit together

Upscaling


Office for National Statistics
Hierarchical Representation of UK Statistical Geographies
Hierarchical Representation of UK Statistical Geographies (August 2022)
For downloads and more information visit the ONS Geography Portal
http://geoportal.statistics.gov.uk
Contains Ordnance Survey data © Crown copyright and database right
Produced by ONS Geography, Office for National Statistics
Contact us at ons.geography@ons.gov.uk
Key
This diagram shows the statistical geographies of the UK in a hierarchical way, divided into different geography types or groups. The main groups are shown in the central octagon, with geographies in each group colour-coded accordingly.
Geography Groups
2021 Census*
2021 Statistical Building Block*
Health
Administrative
Postal
Electoral
Eurostat
Other
Experimental
* The 2011 Census count is shown where the 2021 count is not yet available.
Higher level geographies are shown towards the centre of the diagram, with increased granularity moving down the hierarchy or away from the central octagon.
Some geographies are UK-wide, while others are specific to individual nations or groups of nations.
England
Scotland
Wales
UK
Where applicable, GSS entity codes are shown for each geography. A full list of these can be found in the Register of Geographic Codes (RGC).
Geography Entity Code
E England
S Scotland
W Wales
N Northern Ireland
K Cross-border Instance
J Experimental
Digital boundaries are available for many geographies and can be downloaded freely from the Open Geography portal at
http://geoportal.statistics.gov.uk
These are shown with this symbol
Geographies referenced on postcode and UPRN lookups and directories, and on the Index of Place Names, are marked with these symbols
NSPL and ONSPD
ONSPD
NSUL and ONSUD
ONSUD
IPN
Version 3.0 (August 2022)
Electoral
Postal
Administrative
Health
2021 Statistical Building Block
2021 Census
Other
Eurostat
Electoral Wards (8,483)
Without co-terminous boundaries
County Electoral Divisions (1,575)
London Assembly Constituencies (14)
Senedd Cymru Welsh Parliament Constituencies (40)
Westminster Parliamentary Constituencies (650)
Scottish Parliamentary Constituencies (73)
District Electoral Areas (80)
Northern Ireland Assembly Constituencies (18)
Senedd Cymru Welsh Parliament Electoral Regions (5)
Scottish Parliamentary Regions (8)
Unit Postcodes (1,777,927)
Postcode Sectors (11,141)
Postcode Districts (2,954)
Postcode Areas (121)
Experimental
Major Towns and Cities (112)
Local Authority Districts
Unitary Authorities (22)
Electoral Wards (762)
Communities (878)
Combined Authorities (10)
Greater London Authority (1)
London Boroughs (33)
Regions (9)
Metropolitan Counties (6)
Metropolitan Districts (36)
Electoral Wards (6,904)
Counties (24)
Non-metropolitan Districts (181)
Unitary Authorities (59)
Civil Parishes (10,480)
Council Areas (32)
Electoral Wards (355)
Local Government Districts (11)
Electoral Wards (462)
Cancer Registries (9)
NHS England Regions (7)
Integrated Care Systems (10)
Integrated Care Boards (42)
Locations (106)
Public Health England Regions (4)
Public Health England Centres (9)
Cancer Alliances (21)
Strategic Clinical Networks (12)
Health Boards (14)
Cancer Networks (2)
Local Health Boards (7)
Integration Authorities (31)
ISD Health Board of Treatment (2)
Local Commissioning Groups (5)
Small Areas (4,537*)
Output Areas (10,275)
Super Output Areas, Lower Layer (1,917)
Super Output Areas (890*)
Output Areas (46,351*)
Data Zones (6,976*)
Super Output Areas, Lower Layer (33,755)
Output Areas (178,605)
Intermediate Zones (1,279*)
Super Output Areas, Middle Layer (408)
Super Output Areas, Middle Layer (6,856)
Health and Social Care Board (1)
Non-Civil Parished Areas (208)
Settlement 2015 (639*)
Census Merged Local Authority Districts (287)
Built-up Areas (5,830*)
Census Detailed Characteristics (866*)
Census Local Characteristics (1,012*)
Workplace Zones (60,709*)
Built-up Area Sub-divisions (1,836*)
City Regions (3)
Police Force Districts (11)
Northern Ireland Fire and Rescue Service Districts (14)
Northern Ireland Fire and Rescue Service Areas (4)
Police Force Area (1)
Northern Ireland Fire and Rescue Service (1)
Senedd Cymru Economic Regions (4)
Department for Children, Education, Lifelong Learning and Skills (4)
Strategic Regeneration Sub-areas (2)
Strategic Regeneration Areas (7)
Spatial Plan Sub-areas (3)
Spatial Plan Areas (6)
Agricultural Small Areas (235)
Agricultural Regions (7)
Transport Consortia Areas (4)
Footprint Regions for Public Service Collaboration (6)
City Regions (4)
Communities First Areas (52)
LAU2 Areas (9,357)
LAU1 Areas (377)
Enterprise Zones (46)
Local Enterprise Partnerships (38)
LEP – non-overlapping part (3)
Primary Urban Areas (56)
Waste Authorities (9)
Development Corporations (4)
LEP – overlapping part (2)
International Territorial Level 3 Areas (173)
International Territorial Level 2 Areas (40)
International Territorial Level 1 Areas (12)
Areas of Outstanding Natural Beauty (38)
Fire and Rescue Authorities (47)
Police Force Areas (43)
Community Safety Partnerships
Registration Sub-Districts (173)
Registration Districts (173)
Local Resilience Forums (42)
Scottish Marine Regions (11)
Regional Transport Partnerships (7)
Strategic Development Plan Areas (4)
Community Justice Authorities (8)
Enterprise Regions (6)
Travel to Work Areas (228)
Local Planning Authorities (393)
National Parks (15)
Scottish Police Divisions (13)
Police Force Area (1)
Island Groups (57)
Settlements (314)
Localities (656)
Scottish Fire and Rescue Local Senior Officer Areas (14)
Scottish Fire and Rescue Service Delivery Areas (3)
Scottish Fire and Rescue Service (1)
Regeneration Outcome Areas – Local Areas (180)
Regeneration Outcome Areas – Community Planning Partnerships (28)
Regeneration Outcome Agreement Areas – Scotland (1)
Scottish Local Resilience Partnerships (12)
Scottish Regional Resilience Partnerships (3)
Urban Regeneration Companies (6)
Highlands and Islands Enterprise (8)
Broad Rental Market Areas (18)

Downscaling

CIVICA RESEARCH

- Is it appropriate? Is it ethical?

- For example, our dataset may only have a region coded as a geographic variable but other geographic information may be contained within the dataset.
- What different variables can be used to infer or deduce smaller units?
- Can we merge in other data that will allow us to do this?

Infilling

CIVICA RESEARCH

- Can assumptions reasonably be made about the data?
- Can missing data be brought in from other datasources?
  - API's can be a great source of parsing variables for more information
- Can missing data be estimated?
  - Machine Learning techniques
  - Can be unsupervised or semi-supervised

CC BY

CIVICA RESEARCH

Re-using and repurposing other peoples data

- part of the ***Open Social Science Handbook*** project

Timothy Monteath (UCL/LSE) and Patrick Dunleavy (LSE)

10 November 2022

Bocconi University
Central European University
European University Institute
Hertie School
National University of Political Studies and Public Administration
Sciences Po
Stockholm School of Economics
The London School of Economics and Political Science