Edinburgh Napier
UNIVERSITY

Data Cleansing

Dr Pierre Le Bras

About Me

2011 Started to study Software Engineering

2013 Moved to Edinburgh to study Computer Science

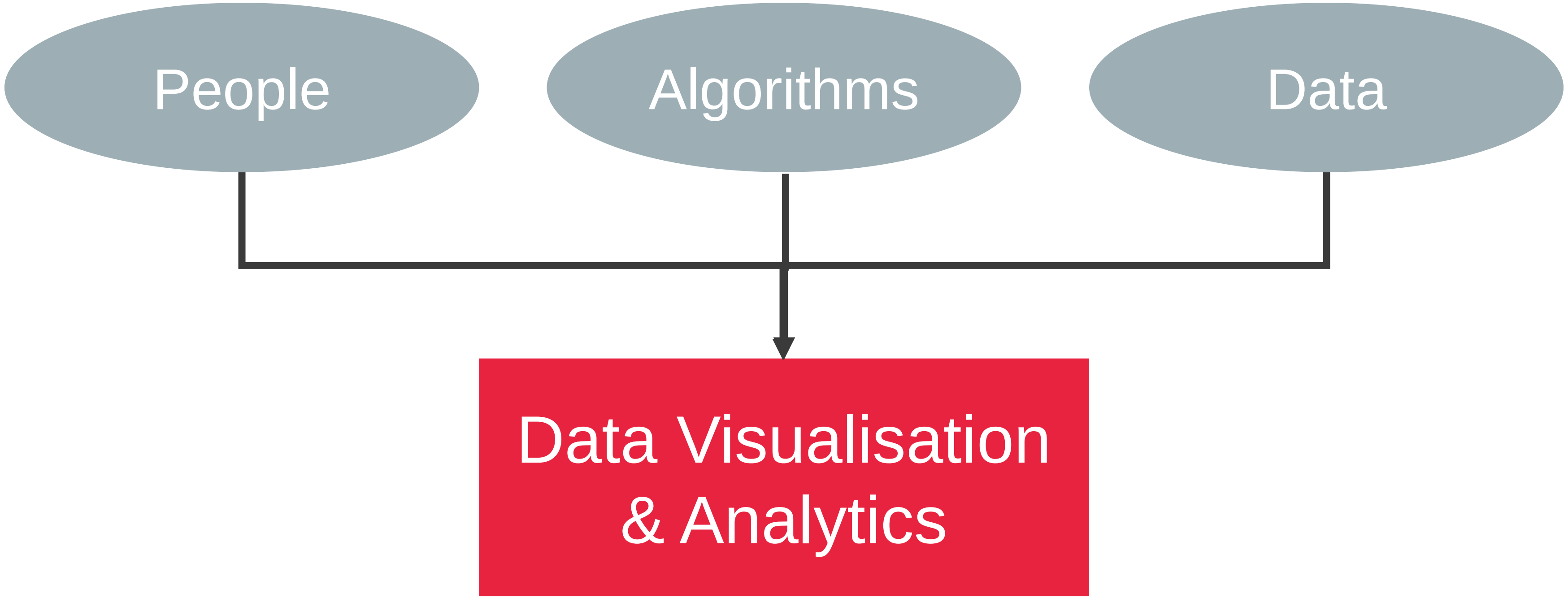
2015 Started my PhD in Data Visualisation and Analytics

2018 Worked as a Research Associate

2021 Lecturer in Software Engineering and Data Science

Edinburgh Napier
UNIVERSITY

Data Visualisation

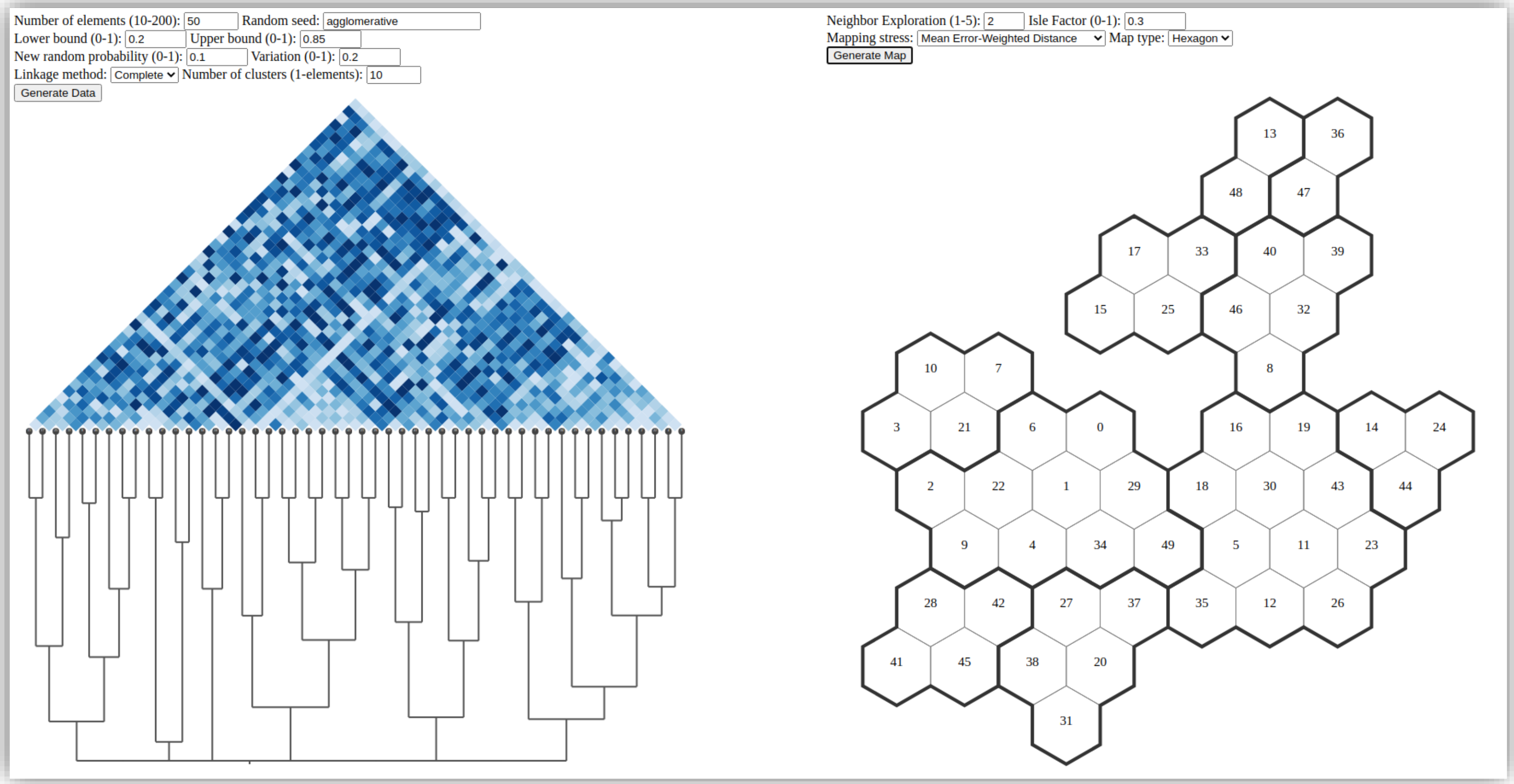

People

Algorithms

Data

Data Visualisation & Analytics

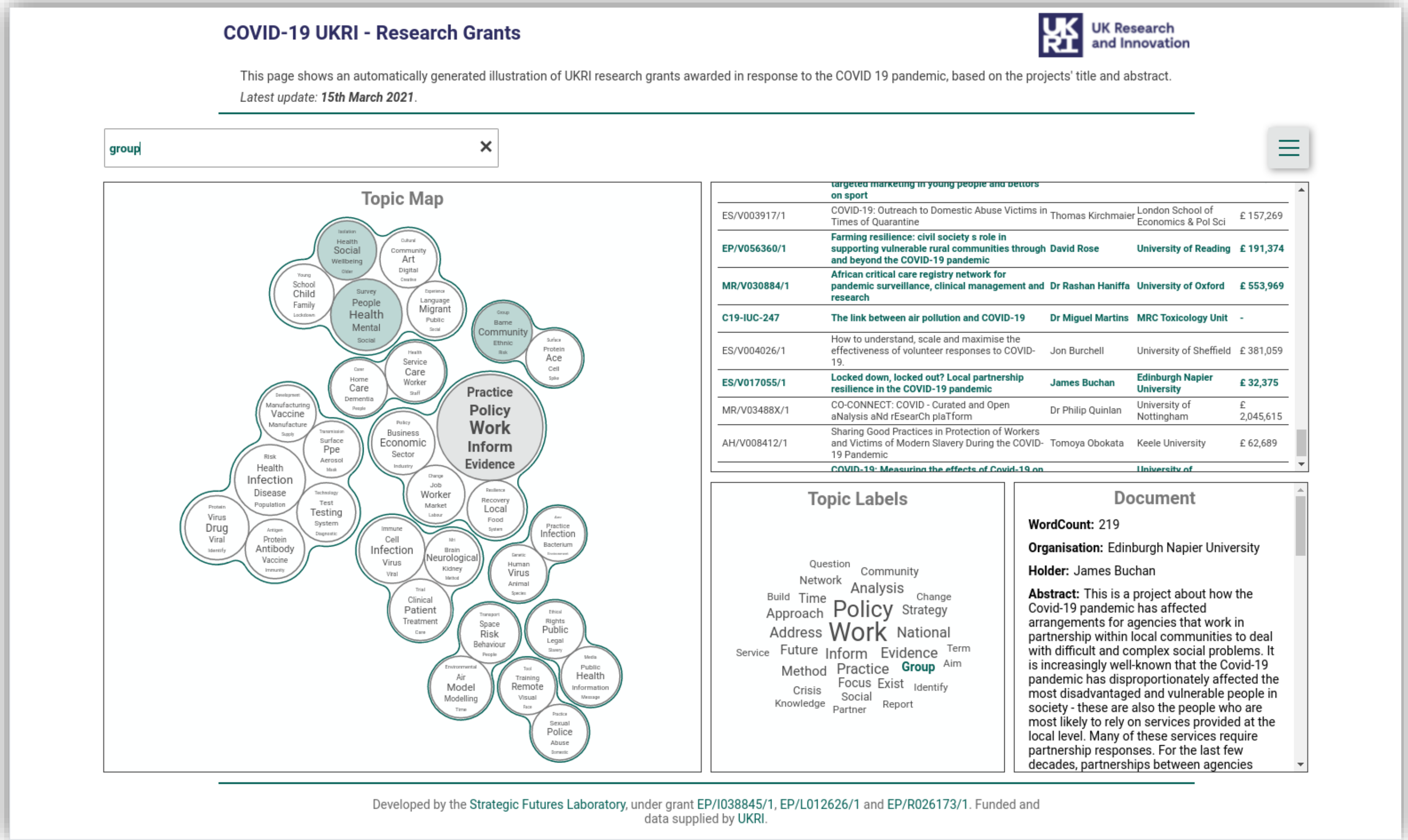
Data Visualisation



Number of elements (10-200): 50 Random seed: agglomerative
Lower bound (0-1): 0.2 Upper bound (0-1): 0.85
New random probability (0-1): 0.1 Variation (0-1): 0.2
Linkage method: Complete Number of clusters (1-elements): 10

Neighbor Exploration (1-5): 2 Isle Factor (0-1): 0.3
Mapping stress: Mean Error-Weighted Distance Map type: Hexagon

Data Visualisation



COVID-19 UKRI - Research Grants

UK Research and Innovation

This page shows an automatically generated illustration of UKRI research grants awarded in response to the COVID 19 pandemic, based on the projects' title and abstract.
Latest update: ***15th March 2021***.

group

Topic Map

Isolation Health Social Wellbeing Older
Cultural Community Art Digital Creative
Young School Child Family Lockdown
Survey People Health Mental Social
Experience Language Migrant Public Social
Group Bame Community Ethnic Risk
Surface Protein Ace Cell Spike
Carer Home Care Dementia People
Health Service Care Worker Staff
Practice Policy Work Inform Evidence
Development Manufacturing Vaccine Manufacture Supply
Transmission Surface Ppe Aerosol Mask
Policy Business Economic Sector Industry
Risk Health Infection Disease Population
Technology Test Testing System Diagnostic
Change Job Worker Market Labour
Resilience Recovery Local Food System
Protein Virus Drug Viral Identify
Antigen Protein Antibody Vaccine Immunity
Immune Cell Infection Virus Viral
Mri Brain Neurological Kidney Method
Practice Infection Bacterium
Genetic Human Virus Animal Species
Trial Clinical Patient Treatment Care
Transport Space Risk Behaviour People
Ethical Rights Public Legal Slavery
Environmental Air Model Modelling Time
Tool Training Remote Visual Face
Media Public Health Information Message
Practice Sexual Police Abuse Domestic

	targeted marketing in young people and bettors on sport			
ES/V003917/1	COVID-19: Outreach to Domestic Abuse Victims in Times of Quarantine	Thomas Kirchmaier	London School of Economics & Pol Sci	£ 157,269
EP/V056360/1	**Farming resilience: civil society s role in supporting vulnerable rural communities through and beyond the COVID-19 pandemic**	**David Rose**	**University of Reading**	**£ 191,374**
MR/V030884/1	**African critical care registry network for pandemic surveillance, clinical management and research**	**Dr Rashan Haniffa**	**University of Oxford**	**£ 553,969**
C19-IUC-247	**The link between air pollution and COVID-19**	**Dr Miguel Martins**	**MRC Toxicology Unit**	**-**
ES/V004026/1	How to understand, scale and maximise the effectiveness of volunteer responses to COVID-19.	Jon Burchell	University of Sheffield	£ 381,059
ES/V017055/1	**Locked down, locked out? Local partnership resilience in the COVID-19 pandemic**	**James Buchan**	**Edinburgh Napier University**	**£ 32,375**
MR/V03488X/1	CO-CONNECT: COVID - Curated and Open aNalysis aNd rEsearCh plaTform	Dr Philip Quinlan	University of Nottingham	£ 2,045,615
AH/V008412/1	Sharing Good Practices in Protection of Workers and Victims of Modern Slavery During the COVID-19 Pandemic	Tomoya Obokata	Keele University	£ 62,689
	COVID-19: Measuring the effects of Covid-19 on		**University of**	

Topic Labels

Document

WordCount: 219

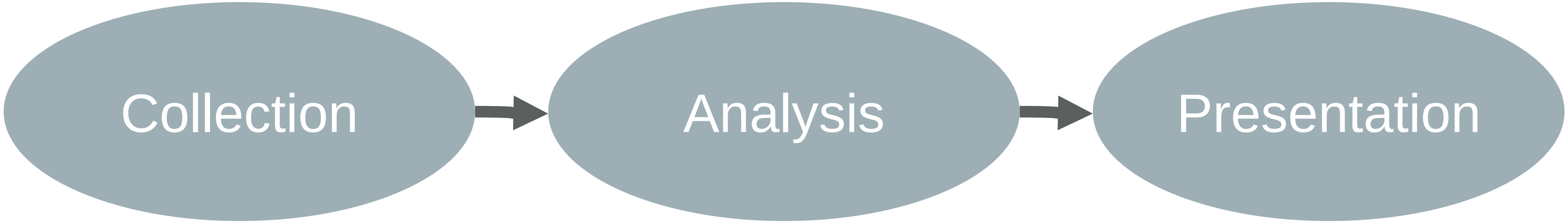
Organisation: Edinburgh Napier University

Holder: James Buchan

Abstract: This is a project about how the Covid-19 pandemic has affected arrangements for agencies that work in partnership within local communities to deal with difficult and complex social problems. It is increasingly well-known that the Covid-19 pandemic has disproportionately affected the most disadvantaged and vulnerable people in society - these are also the people who are most likely to rely on services provided at the local level. Many of these services require partnership responses. For the last few decades, partnerships between agencies

Developed by the Strategic Futures Laboratory, under grant EP/I038845/1, EP/L012626/1 and EP/R026173/1. Funded and data supplied by UKRI.

Processing Data

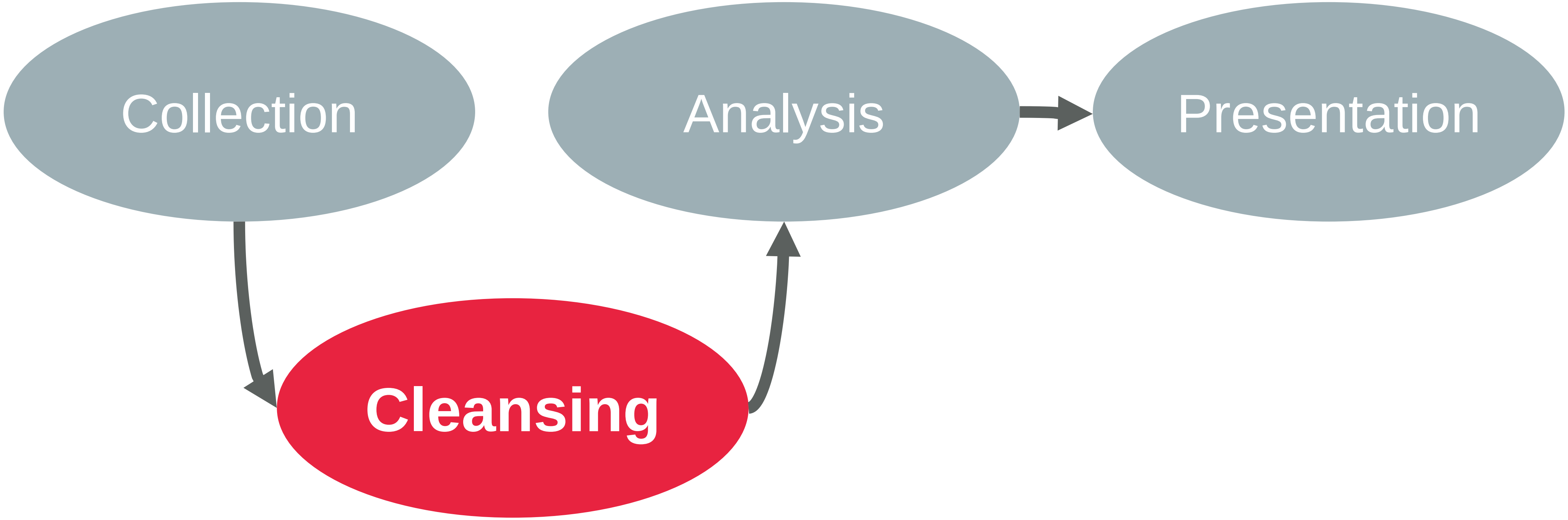


Collection

Analysis

Presentation

Processing Data



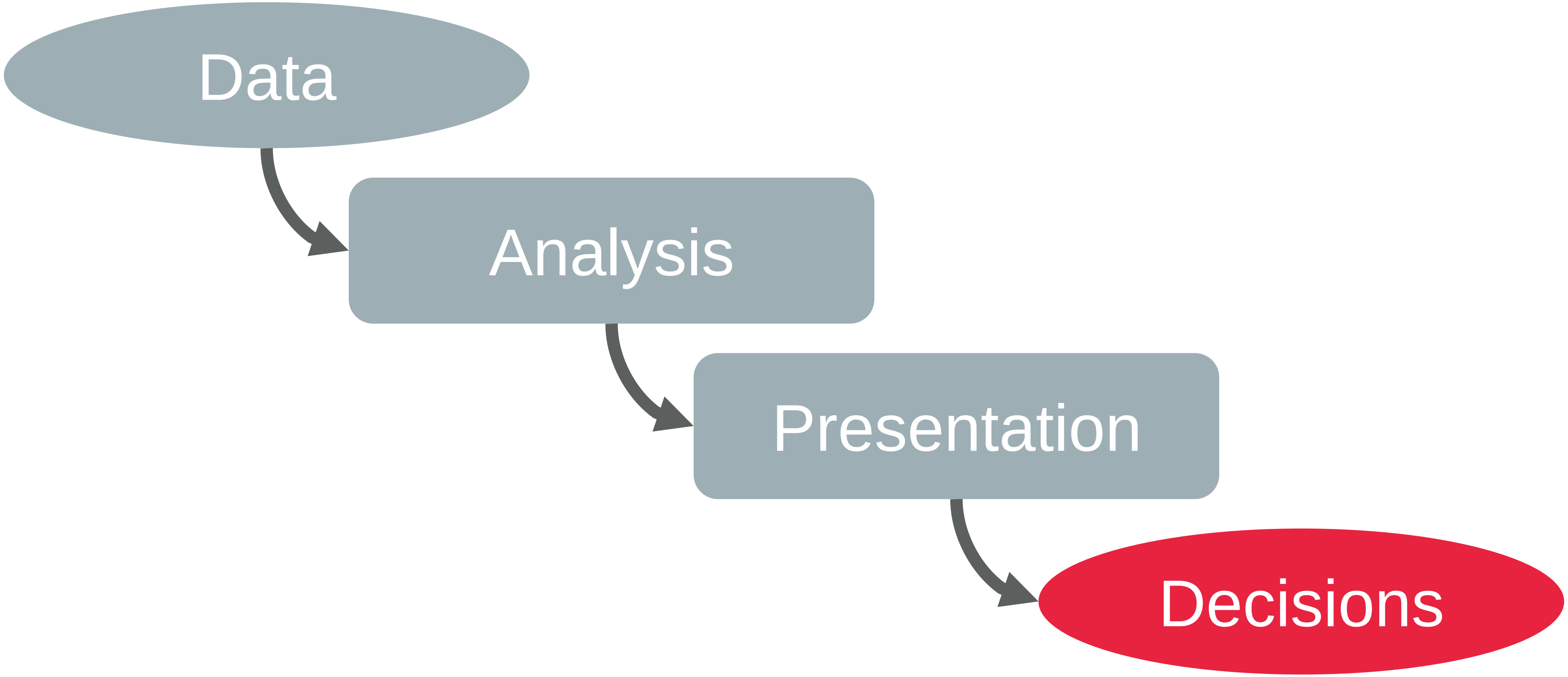
Why do you want Clean Data?

Edinburgh Napier UNIVERSITY

Why Investing in Data?

Data is used by organisations or individuals seeking **informed decision making**

From Data to Decisions



Data

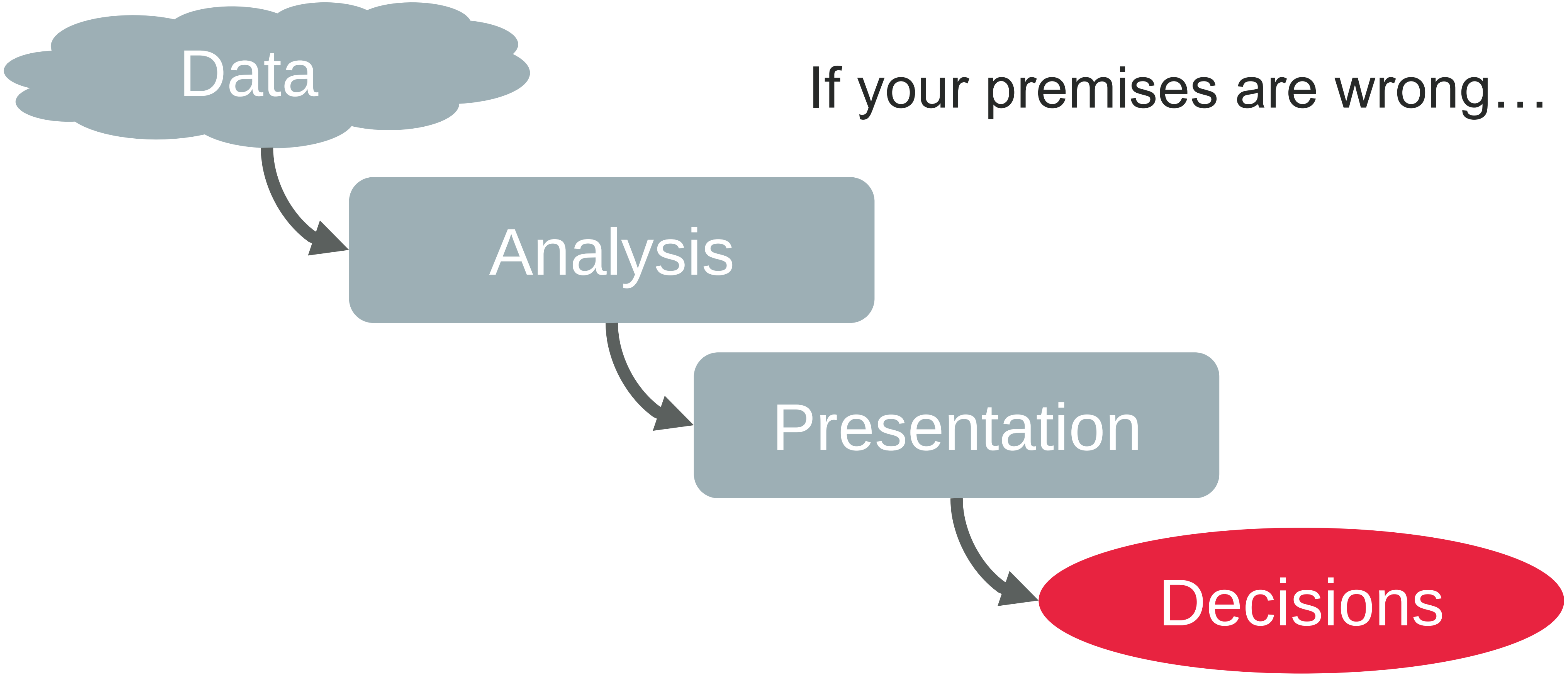
Analysis

Presentation

Decisions

Edinburgh Napier
UNIVERSITY

From Data to Decisions


Data

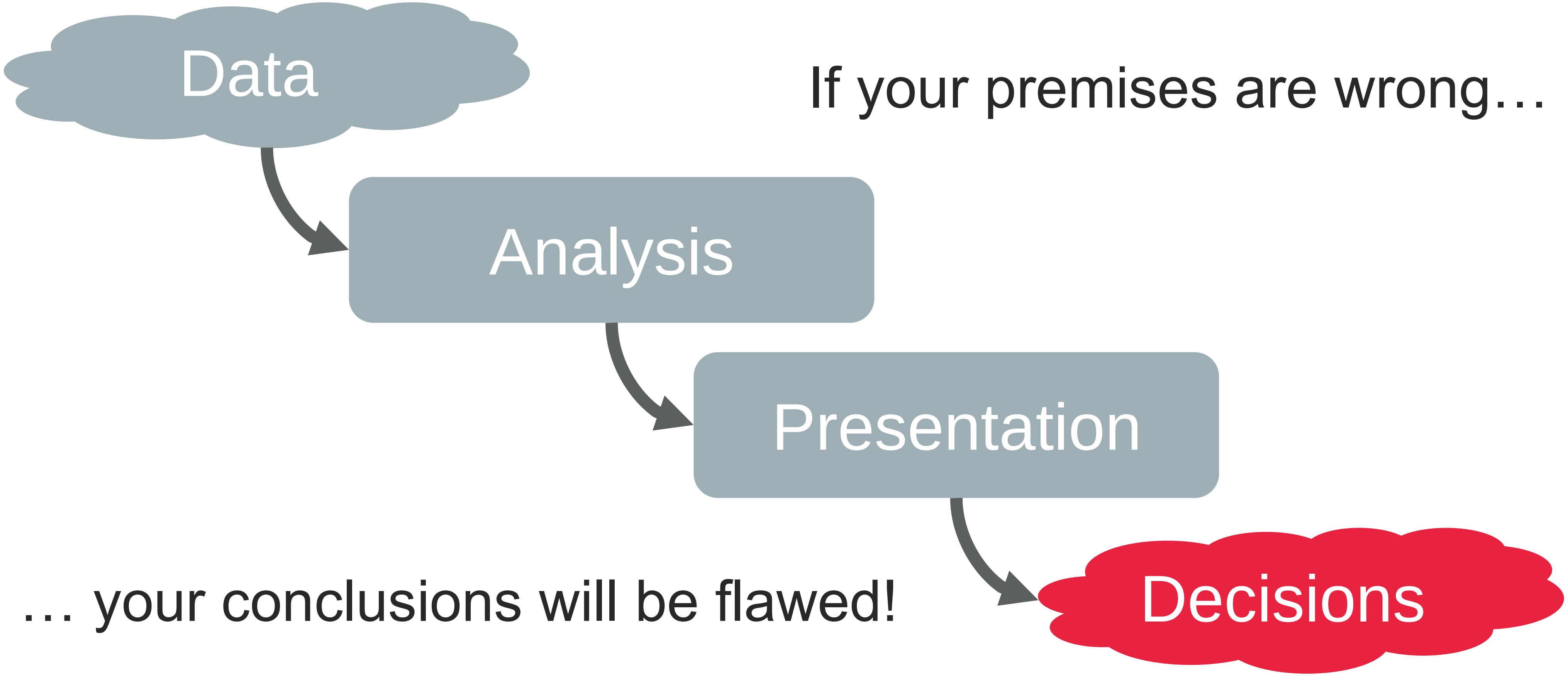
Analysis

Presentation

Decisions

If your premises are wrong…

From Data to Decisions



If your premises are wrong…

… your conclusions will be flawed!

Impacts of Dirty Data

- Bad data lead to inaccurate results and false insights
- ... which lead to the wrong decisions being taken
- ... which can have serious consequences

- For example:
  - Wrong census data and public funds: clickondetroit.com/news/michigan/2021/10/15/census-analysis-shows-black-americans-may-have-been-significantly-undercounted/

Where does it get dirty?

Human Error

- At an individual level
  - Incorrect or missed entry
  - Mistake when copying data
- At an organisational level
  - Poor data management policies
  - Poor communication between services
  - Poor documentation

System Failures

- Hardware and/or software malfunction
  - Data file corruption
  - Loss of data

Collection Methods

- Collecting and merging data from multiple sources
  - With likely different formats

- Unforeseen exceptions during automatic collections
  - E.g., web crawlers and weird HTML structures

What is Clean Data?

Dirty Data can be…

- Duplicated
- Out of date
- Inconsistent
- Incorrect
- Invalid
- Unformatted
- Non-uniform
- Inaccurate
- Missing
- Incomplete

5 Rules to Data Quality

- Validity
- Accuracy
- Completeness
- Consistency
- Uniformity

Validity

- How much the data conforms to a set of rules
- Plays at the individual entries and fields level

	Field 1	Field 2	Field 3	Field 4	…
Entry 1	✓	✓	✓	✓	✓
Entry 2	✓	✓	✓	✓	✓
Entry 3	✓	✓	✓	✓	✓
…	✓	✓	✓	✓	✓

Validity

- Each field has correct types (String, Number, Date, Boolean, …)
- Numbers and dates must fall within the right ranges
- Discrete values must be consistent
    - E.g. a female gender value must be “Female” or “F” across all entries but not both
- Discrete values must be from a finite set
    - E.g. UK regions can only be “England”, “Wales”, “Scotland” and “Northern Ireland”

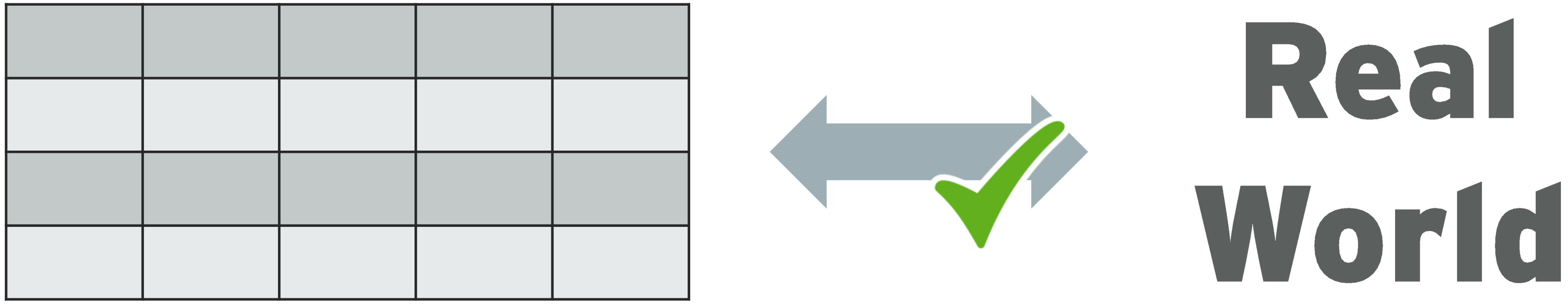
Validity

- Text values must have the right pattern
  - E.g. phone numbers must have a “XXXX-XXX-XXXX” pattern
- Each entry should be unique
  - A field or combination of fields must be defined as unique key
- Mandatory fields are not empty *
- Related fields are consistent *
  - E.g. a birth date and age fields should be coherent (age = current date – birth date)

* These requirements fit with other data quality principles

Accuracy

- How well the data fits with the truth
- Difficult to achieve, requires an external “gold standard” dataset
- Accurate data can become inaccurate (out-of-date)

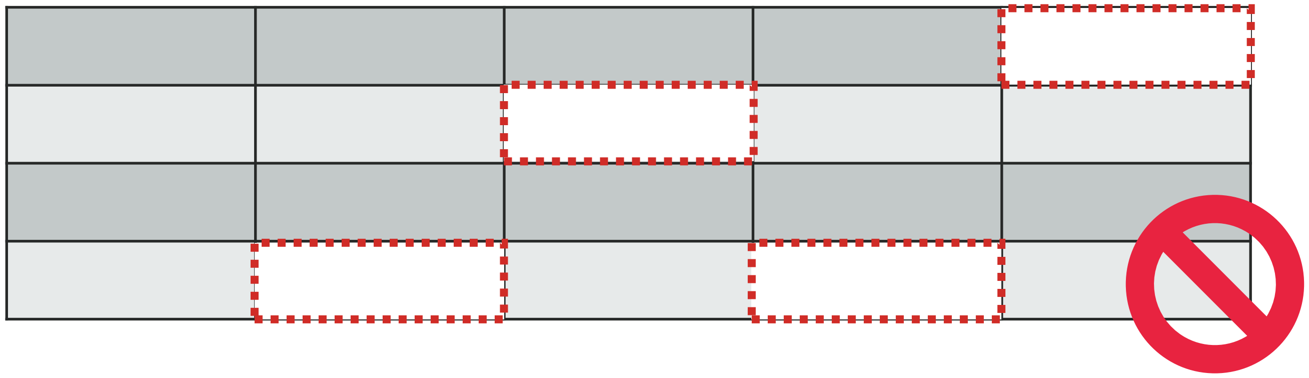


Real World

Completeness

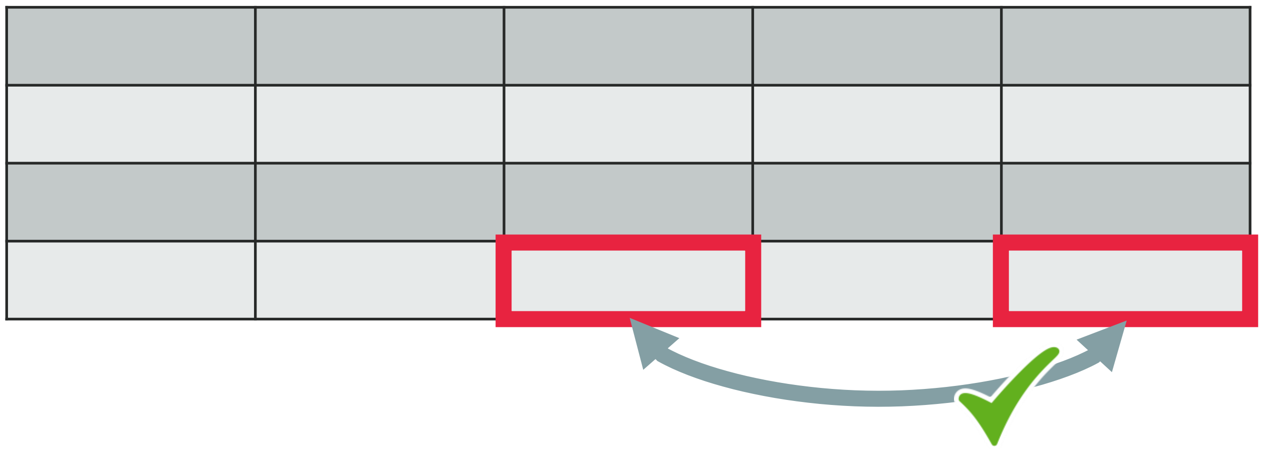
- How little required data is missing

- Difficult to fix, “*Missing*”, “*Unknown*” or “*NA*” value might be needed → **Missing data is still data**



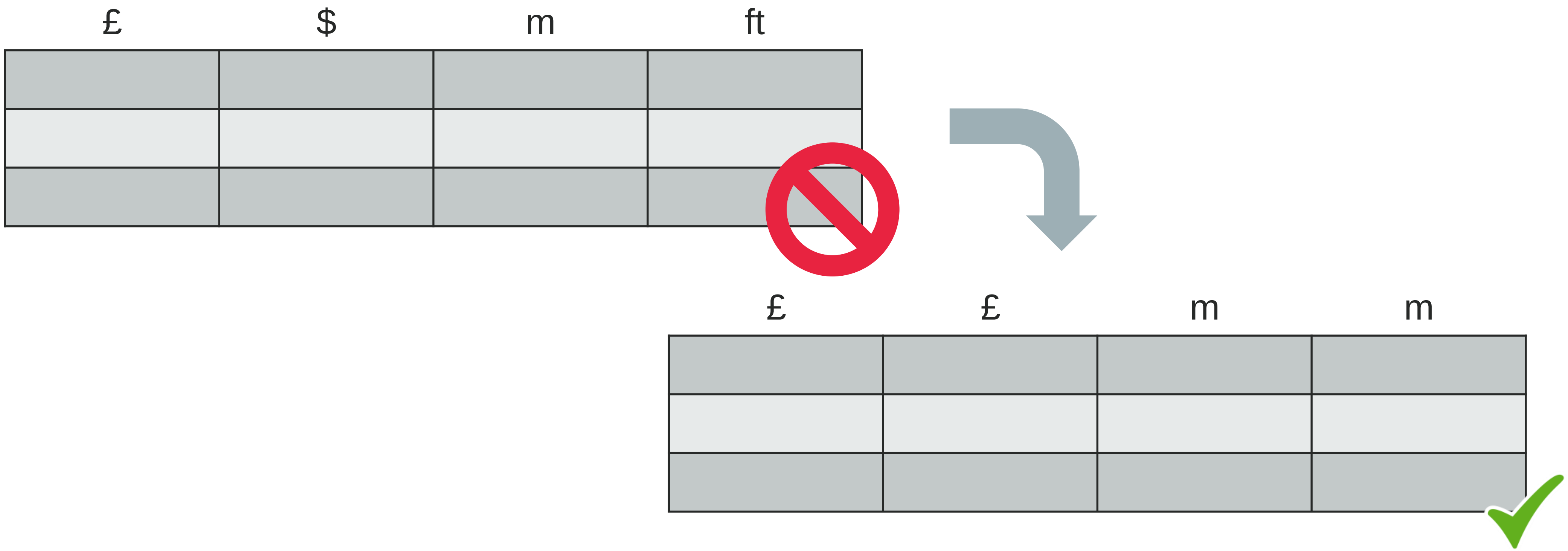
Consistency

- How much the data values agree or are coherent
- To fix: check which is most up-to-date, which source is most reliable, etc.



Uniformity

- How units of measures are similar across fields



£	$	m	ft

£	£	m	m

Other Data Quality Aspects

- Comparability
- Relevance
- Credibility
- Currency (up-to-date)
- Confidentiality

http://www.dama-nl.org/data_quality/

How to Clean Data?

Step 1 - Inspection

- Exploring the data and detecting unexpected, incorrect and/or inconsistent entries or fields.

- Data Profiling:
  - Are values corresponding to their field’s format and pattern specification?
  - Are there missing fields?
  - Are the mean, median, range, sum, etc. coherent?
  - Is the distribution of values coherent?
  - Are there outliers?
  - Is there duplicate entries?

- Visualisation

Step 2 - Cleansing

- Clear irrelevant data
- Handle duplicates
- Handle missing values
- Convert types
- Fix the content
- Fix the format
- Scale and/or normalise numbers

Step 2 - Cleansing

Drop irrelevant data: remove fields or entries which are not needed for our analysis

- *Column-wise*: fields which provide information outside of our task domain, e.g., phone numbers in an analysis of student grades
- *Row-wise*: entries with attributes outside of our task domain, e.g., postgraduate students in an analysis of undergraduate student grades

- **Only delete data if you are sure about it!**
    - You might need it later
    - It might make some interesting correlation analysis

Step 2 - Cleansing

Remove duplicates

- *Column-wise*: a field repeating another field
- *Row-wise*: duplicate entries
  - Identify a field or combination of fields as unique keys
  - Flag and remove duplicates based on the keys

Step 2 - Cleansing

Check and correct missing values

- Drop: if a field or entry is mostly empty, it might be worth removing
- Complete: some values can be filled, using another dataset, doing manual search, etc.
- Infer: some values may be inferred from other observation (e.g., with linear regression)
- Flag: **missing data can be data!**
    - Numerical data can be set to absurd values: *-1* or *0*
    - Categorical data can be set to a new category: *“missing”* or *“NA”*
    - *“missing”*, *“NA”* or *“unknown”* have different meaning

Step 2 - Cleansing

Convert values to their correct types

- Dates should be date objects (Unix timestamp)
- Numerical data should be numbers
- Categorical data can be strings or numbers
- Text data should be strings
- Boolean values (true/false, yes/no) should be Booleans

Step 2 - Cleansing

Fix the content: Correct syntax, mistakes and conversion errors

- Strings: trim whitespaces, fix typos
  - Easy on categorical data
  - Difficult with free text data (tools can help)
- Numbers: fix errors (from your inspection)
  - Can use the same approaches as with missing data (drop / complete / infer / flag)

Step 2 - Cleansing

Fix the format: Standardize the values across your dataset

- Strings:
  - Patterns (e.g., XXXX-XXX-XXXX for phone numbers)
  - Format (e.g., lower case)
- Numbers:
  - Units (e.g., distances, weights, currencies)
  - Format (e.g., zero-padding, decimal digits)
- Date:
  - Format (e.g. YYYY-MM-DD or DD/MM/YY)
- Categories:
  - Values (e.g. *“F”* or *“female”*, *“Sat.”* or *“saturday”*)

Step 2 - Cleansing

Scale and/or normalise number ranges

- Scaling
  - Making your numbers comparable, on the same scale
- Normalisation
  - Making your number distribution normally distributed (for later statistical analysis)

Edinburgh Napier
UNIVERSITY

Step 3 - Verification

- Re-inspect the data to make sure of its correct cleaning
- Correct mistakes (manually) if needed

Step 4 - Report

- Reporting your cleansing is important

- It helps measure data quality, highlight issues and identify common problems

- It can inform future data practices
  - Collection (e.g., survey question often unanswered, web crawler cropping text, etc.)
  - Storage (e.g., mismatch between datasets)

Data Cleansing Tools

Programming

python™

pandas

R Studio®

Software

Excel

TRIFACTA

Data Ladder
Get the most out of your data

WinPure™

Software

OpenRefine

- Free
- Open Source
- Cross Platform
- Design for Cleansing
- Powerful

Next…

- Questions?
- Break
- Live demonstration