MONASH University

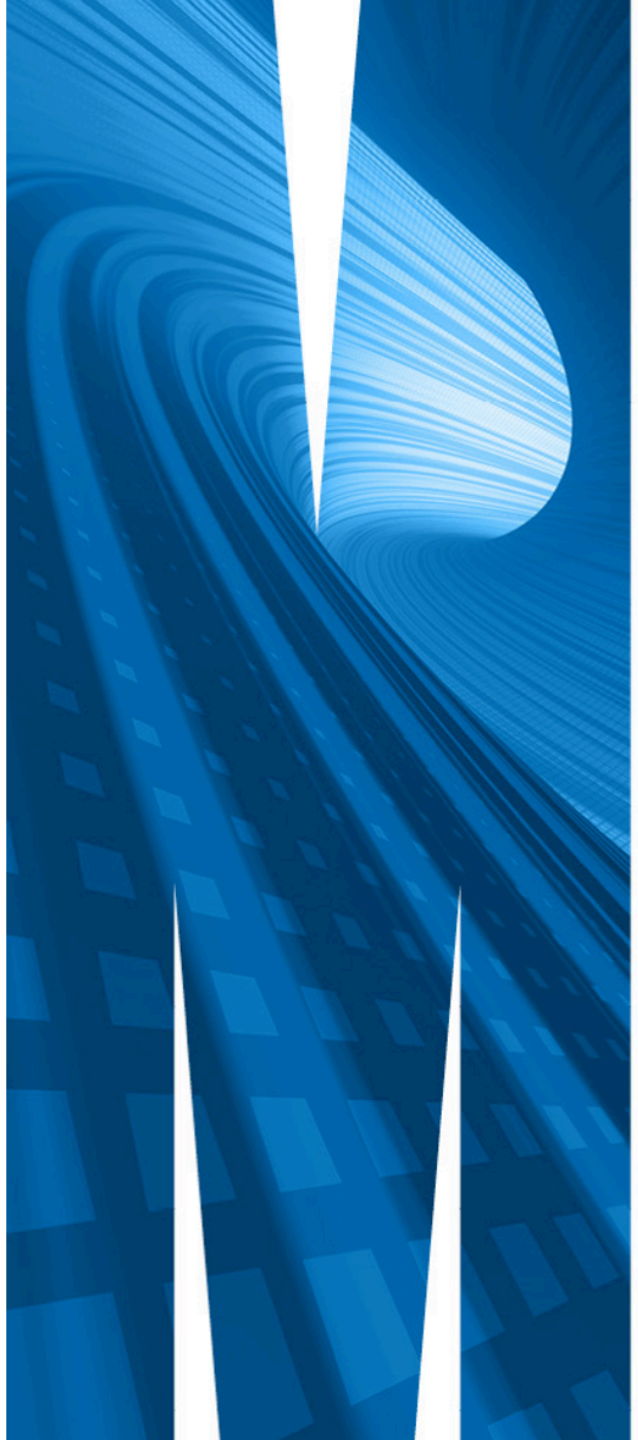
Ex-Post Harmonisation and “Statistical” Data Provenance

SoDa Lab Meeting, LMU Munich, 16 Jun 2024

Cynthia A. Huang

cynthia.huang@monash.edu

Department of Econometrics and Business Statistics, Monash Business School



Introduction

About Me!

- 👩‍🎓 Third Year PhD Candidate in Econometrics & Business Statistics, supervised by, and affiliated with:
  - Prof. Rob Hyndman, EBS
    - NUMBATs, Non-Uniform Monash *Business Analytics* Team
  - A.Prof Simon Angus, Economics
    - SoDa Laboratories, *alternative data for social science insights*
  - Dr. Sarah Goodwin, Human Centered Computing
    - Embodied Visualisation group, *immersive technologies for data analysis*

About Me!

- 💱 Previously:
  - Economics at the University of Melbourne
  - Tutoring undergraduate economics
  - Assisting with data collection & curation for empirical economists
- 👩 Outside of Research:
  - 🧗‍♀️ Climbing, 🧘‍♀️ Yoga, 👩‍🍳 Foodie
  - 🎙️ Regular host on The Random Sample podcast

About Me!

- 📊 Research Interests
  - 🧉 Statistically sound, well-documented and low-friction adaptation of “alternative” data for research purposes.
  - 🖇️ Data provenance models that capture both **statistical** decisions, and **computational** implementation details.
- 👩‍🎓 Thesis: **Unified Statistical Principles and Computational Tools for Data Harmonisation and Provenance**
  - Conceptual framework for redistributing numeric mass between categories in related statisical classifications
  - Software implementation in R

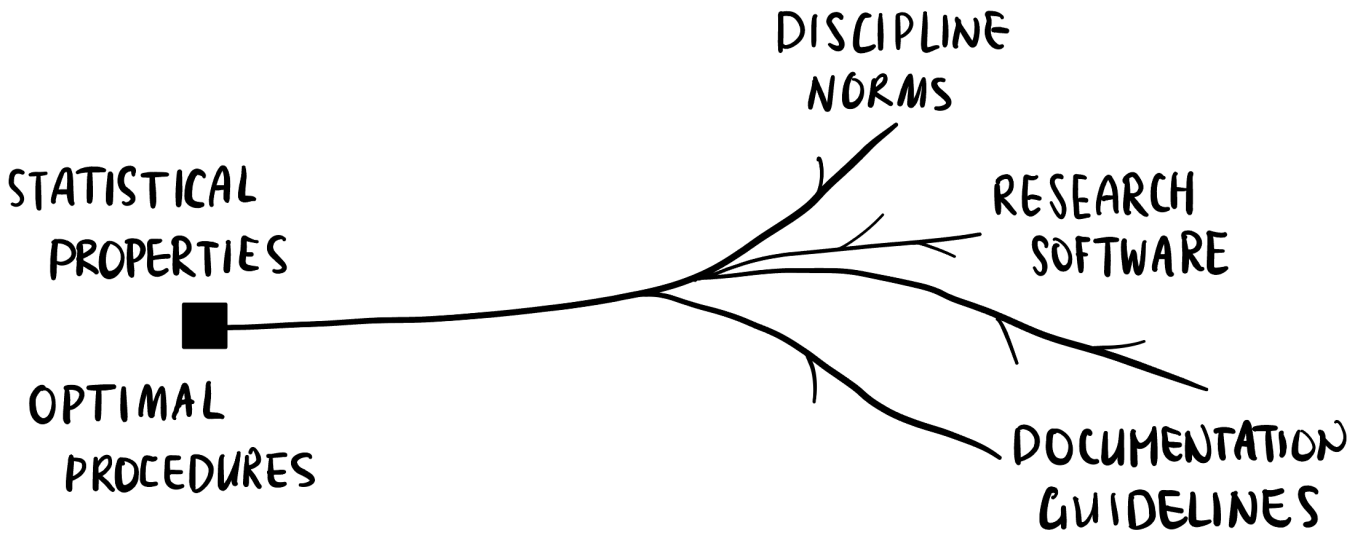
About Me!

- 📋 Collaborative work:
  - Review of Data Provenance approaches across CS and Statistics
  - Adapting web-scraped retail product & price data for public health research
  - Human in the Loop verification for data extraction from spreadsheets using Generative AI
- 💡 Reproducible and reusable research and teaching tools:
  - learnr.academy built on Quarto & webR
  - Reproducibility for Economists also with Quarto

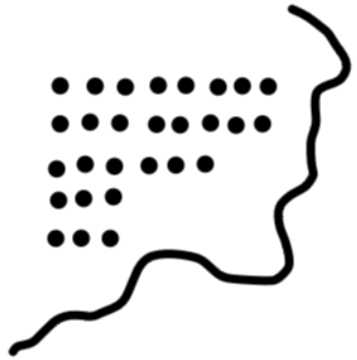
Thesis Background & Motivation

Harmonising and Integrating Data

- Opportunities to combine existing data for analysis abound,
- Existing literature exists on a spectrum from conceptual to applied,
- with keywords such as **data** *preprocessing*, *cleaning*, *fusion*, *integration*, *harmonisation* etc.

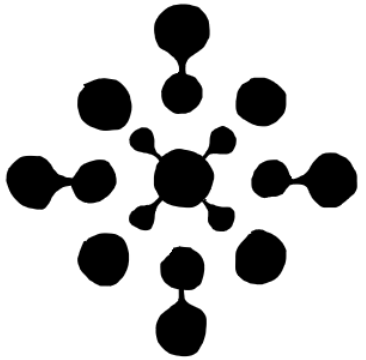


Aspects of Ex-Post Harmonisation



Defining or selecting mappings between classifications or taxonomies,

Implementing and validating mappings on given data,



Documenting and analysing the implemented mapping.

Existing Conceptual Contributions

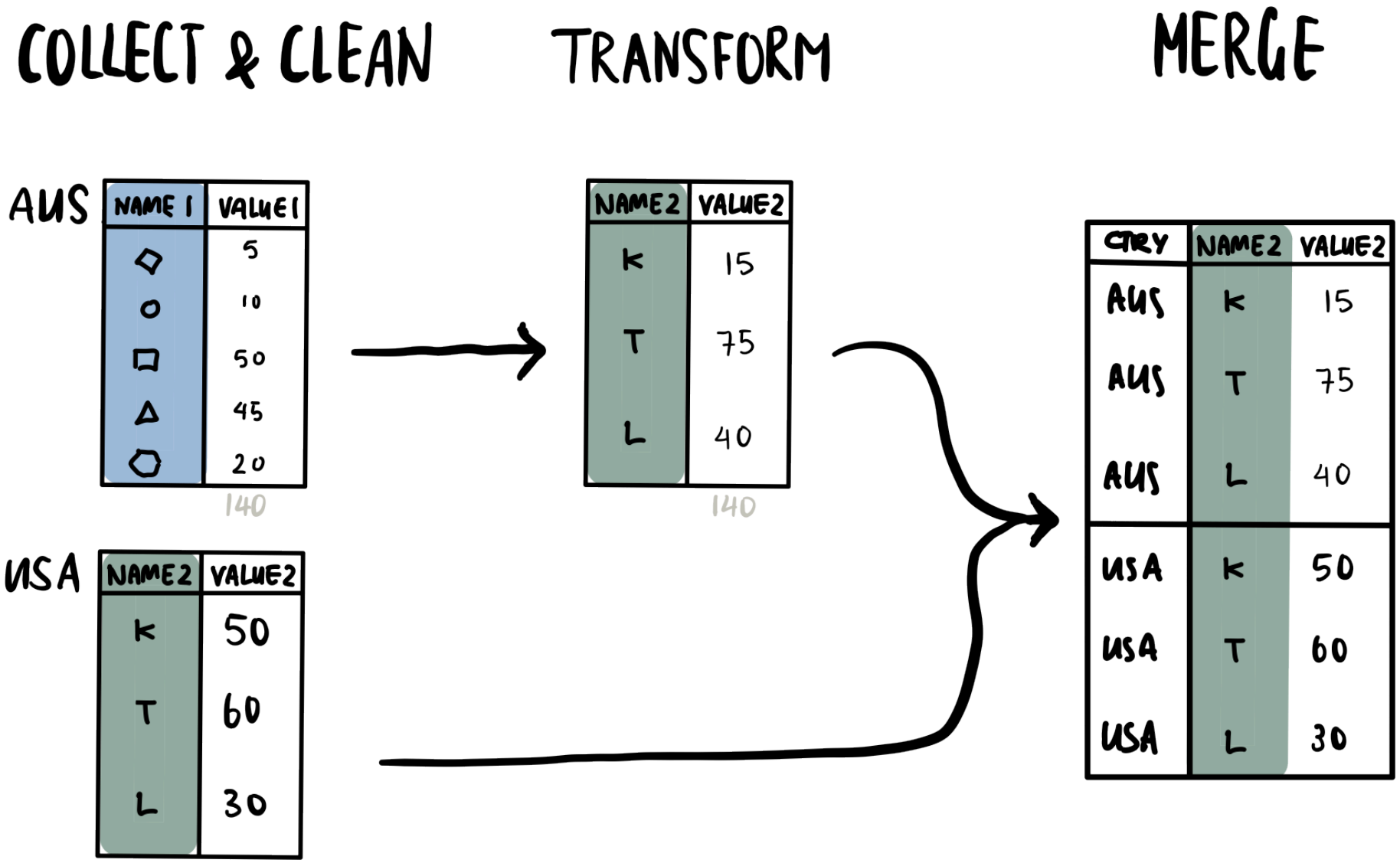
- Approaches and insights are scattered across “data science” fields:
  - Multi-resolution, multi-phase and multi-source inference (Meng 2014)
  - Data wrangling workflows and domain specific languages (Kandel et al. 2011; Wickham 2014)
  - Guidelines for data quality and provenance for trustworthy research (Gebru et al. 2021; Peng 2011; Wilkinson et al. 2016)
  - Frameworks for considering alternative preprocessing decisions(Steegen et al. 2016; Blocker and Meng 2013)

Existing Applied Contributions

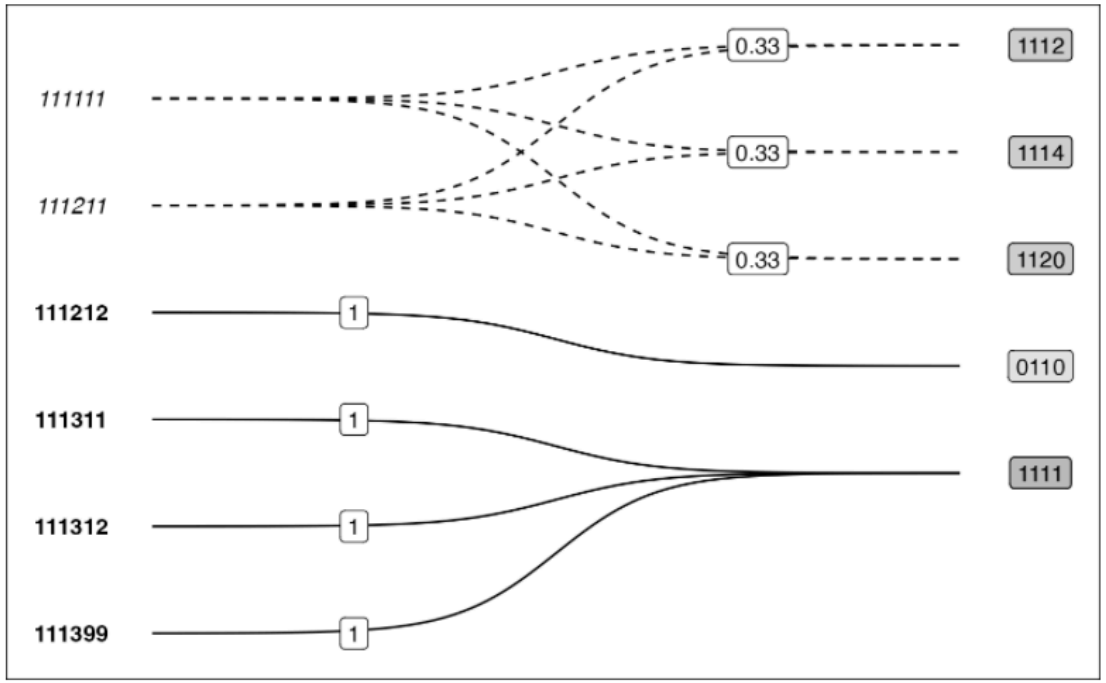
- Many attempts at domain-specific and general purpose tools exist:
  - Survey Data Harmonisation (Kołczyńska 2022)
  - Data validation tools (van der Loo and de Jonge 2021)
  - Data provenance tracking and documentation (Lucchesi et al. 2022)
  - Interactive tools for *Multiverse Analysis* (Sarma et al. 2018; Liu et al. 2021)

Ex-Post Harmonisation of Aggregate Statistics

Stylised Example



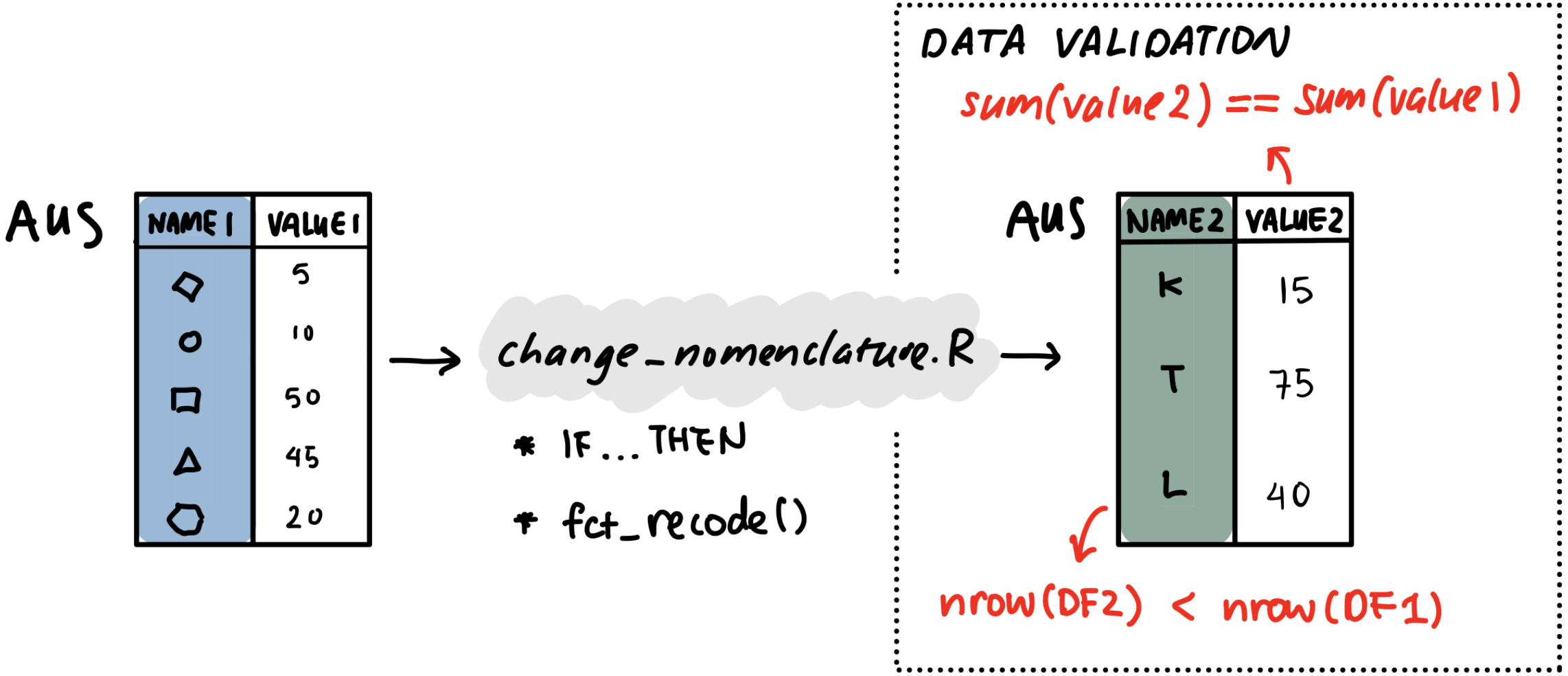
Example: ANZSCO22 and ISCO8 Occupation Codes



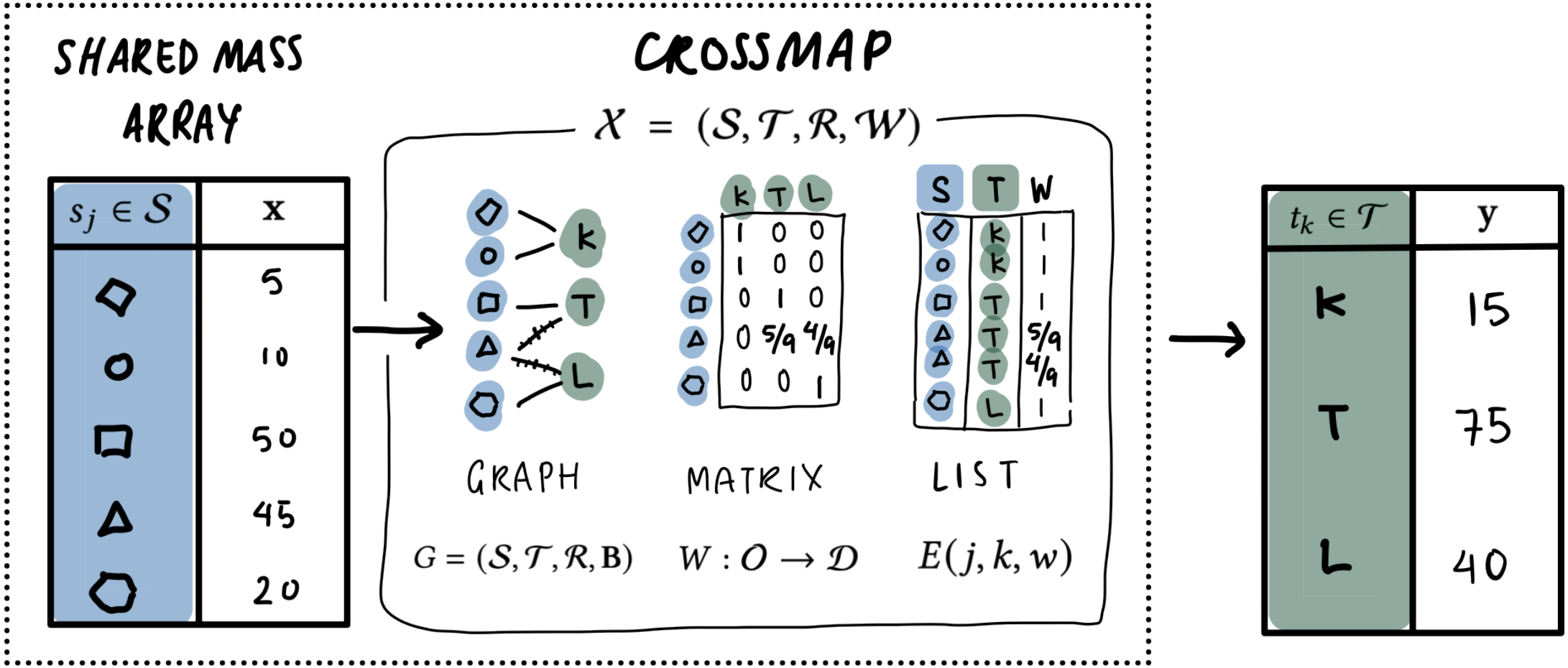
anzsco22	anzsco22_descr
111111	Chief Executive or Managing Director
111211	Corporate General Manager
111212	Defence Force Senior Officer
111311	Local Government Legislator
111312	Member of Parliament
111399	Legislators nec

isco8	isco8_descr
1112	Senior government officials
1114	Senior officials of special-interest organizations
1120	Managing directors and chief executives
0110	Commissioned armed forces officers
1111	Legislators

Current Approach: Input/Output Comparison



Proposed Alternative: Input & Function Capture



Proposed Approach: Task Abstraction

The **crossmap transform** takes *(data input)*:

- numeric values which form a conceptually shared mass and are indexed by a specific set of keys (e.g. occupation codes), a **shared mass array**

and *(function)*:

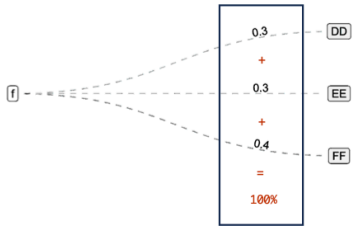
- redistributes the numeric values into a new set of index keys, according to a mapping, the **crossmap**, between the source and target keys

produces *(output)*:

- a counter-factual/imputed **shared mass array** indexed by the target keys

Insights from Equivalent Encodings

Transformation logic can be *validated via graph properties* instead of ad-hoc assertions or line-by-line code review



Condition for preserving numeric totals

Data transformation can be *implemented* using validated crossmaps *via matrix multiplication* [1] performed *as database operations* on the edge list [2]

```
anzsco_xmap
# A tibble: 10 × 3
   anzsco22 isco8 weights
   <chr>    <chr>   <dbl>
 1 111111   1112    0.333
 2 111111   1114    0.333
 3 111111   1120    0.333
 4 111211   1112    0.333
 5 111211   1114    0.333
 6 111211   1120    0.333
 7 111212   0110    1
 8 111311   1111    1
 9 111312   1111    1
10 111399   1111    1
```

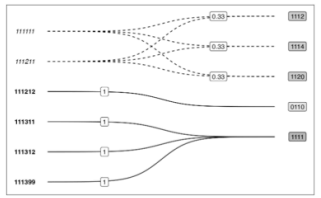
```
anzsco22_stats
# A tibble: 6 × 2
  anzsco22 count
  <chr>    <dbl>
1 111111    1000
2 111211     500
3 111212      40
4 111311     300
5 111312     150
6 111399      10
```

```
apply_xmap(.data = anzsco22_stats,
           .xmap = anzsco_xmap)
```

```
# A tibble: 5 × 2
  isco8 new_count
  <chr>     <dbl>
1 0110         40
2 1111        460
3 1112        500
4 1114        500
5 1120        500
```

```
# mock-up of apply_xmap() function
apply_xmap <- function(.data, .xmap) {
  left_join(
    x = .data,
    y = .xmap,
    by = "anzsco22") |>
  mutate(part_count = count * weights) |>
  group_by(isco8) |>
  summarise(new_count = sum(part_count))
}
```

Bi-graph visualisation and summary techniques can be used to *design* data provenance *documentation* [3]



anzsco22	anzsco22_descr
111111	Chief Executive or Managing Director
111211	Corporate General Manager
111212	Defence Force Senior Officer
111311	Local Government Legislator
111312	Member of Parliament
111399	Legislators nec

isco8	isco8_descr
1112	Senior government officials
1114	Senior officials of special-interest organizations
1120	Managing directors and chief executives
0110	Commissioned armed forces officers
1111	Legislators

Crossmaps can be encoded as:

- **Computational graphs:** multi-partite graph visualisation
- **Linear mappings:** matrix multiplication constraints
- **Edge lists:** rectangular data wrangling tools

Framework Implications

Domain Problem: *Ex-Post Harmonisation*

Provenance Model: *Crossmaps*

Documenting & Auditing

Interactive Tools

Data Imputation Models

Floating Point Computation

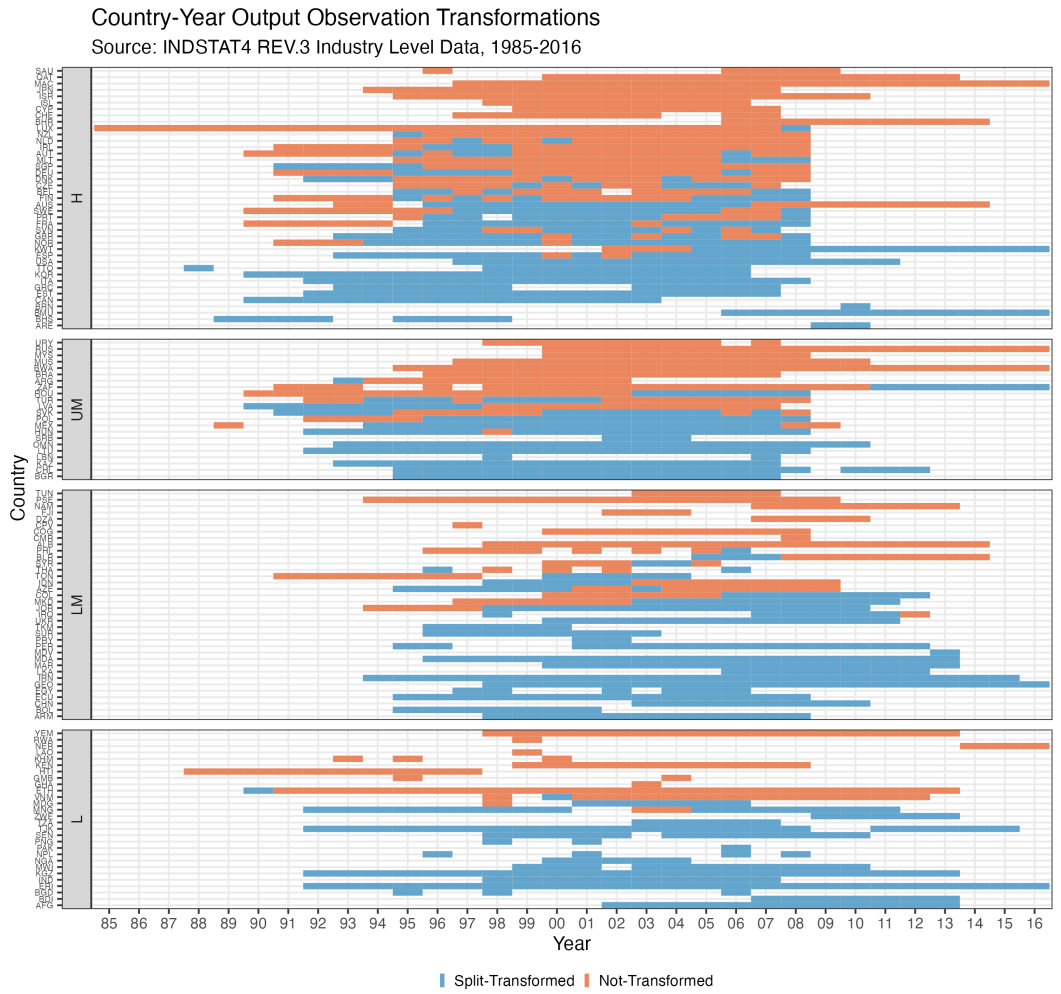
Visual Encoding

Sensitivity and Robustness Analysis

Conceptual and Statistical Implications

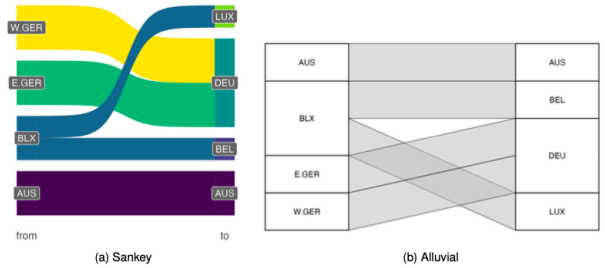
Crossmap (graph) properties could be used to quantify and explore:

- How does the degree and extent of imputation differ between crossmaps?
- How robust are downstream results to alternative harmonisation designs?
- How much imputation has been performed on a given dataset with a given crossmap?
- Which observations in a harmonised dataset have undergone the most (or least) transformation?

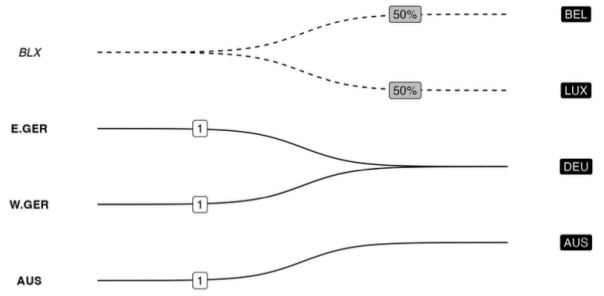


Computational and Design Implications

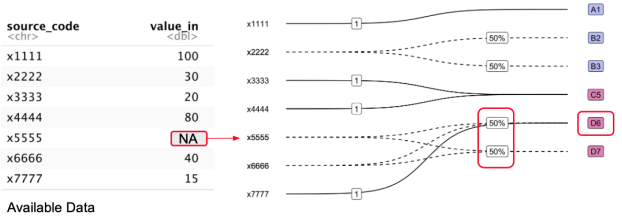
- *data provenance documentation*
  - multi-partite graph layouts
  - graph summaries
- *extracting mapping logic from existing scripts*
  - manipulate data input
  - parse AST into computational graph
- *authoring and auditing interfaces*
  - interactive (multi-table) data merging
  - workflow constraints (missing values etc.)



(a) Sankey (b) Alluvial

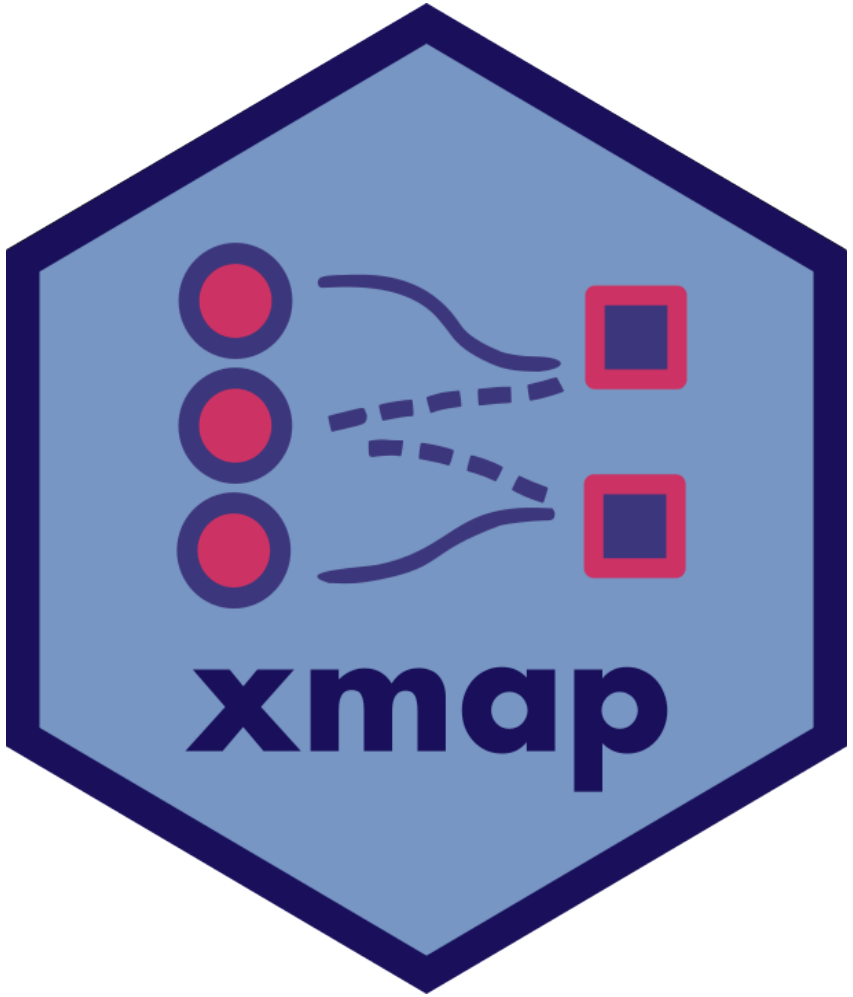


(c) Node-Link



Discussion & Future Work

Current: Software Implementation



WIP:

- Presenting at UseR! (Jul 8-11)
- Will be on CRAN (soon), with accompanying R Journal paper

Package goals:

- implements graph, matrix & table representation in R, with symbolic (fractional) weights
- worked examples in vignettes

Soon: Review of Data Provenance Approaches

- Joint work with PhD Candidate Francis Nguyen, supervised by Prof. Tamara Munzner at the InfoVis group in Dept. Computer Science, University of British Columbia
- Aiming to describe approaches to data provenance across:
  - statistical theory
  - statistical computing
  - database systems
  - data analytics and visualisation

Publication Venues?

🤔 Where to publish & share work on data harmonisation, provenance and quality?

- **Data Science:** ACM/IMS Journal of Data Science*, Harvard Data Science Review, ???
- **CS/HCC:** IEEE VIS*, CHI, ???
- **Statistics & Statistical Programming:** R Journal*, JSS[1], JCGS[2]
- **Applied Venues:** e.g. “Data Reviews” in Australian Economic Review

1. Journal of Statistical Software
2. Journal of Computational and Graphical Statistics

Thanks for Listening!

Connect with me (and other cool Monash folks):

- 🇩🇪 *LMU until Weds, June 26*
- 🇦🇹 UseR!, Salzburg (Jul 8-11)
- 🇺🇸 JSM, Portland (Aug 3-9)
- 🇺🇸 posit::conf(2024), Seattle (Aug 12-14)
- 🇨🇦 UBC, Vancouver, (Jul-Nov)
- 🌏 ???, March 2025 onwards…

Or online: `@cynthiahqy` & `cynthiahqy.com`

References

Blocker, Alexander W., and Xiao-Li Meng. 2013. “The Potential and Perils of Preprocessing: Building New Foundations.” *Bernoulli* 19 (4). https://doi.org/10.3150/13-BEJSP16.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

Kandel, Sean, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. “Wrangler: Interactive Visual Specification of Data Transformation Scripts.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3363–72. Vancouver BC Canada: ACM. https://doi.org/10.1145/1978942.1979444.

Kołczyńska, Marta. 2022. “Combining Multiple Survey Sources: A Reproducible Workflow and Toolbox for Survey Data Harmonization.” *Methodological Innovations* 15 (1): 62–72. https://doi.org/10.1177/2059799121077923.

Liu, Yang, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. “Boba: Authoring and Visualizing Multiverse Analyses.” *IEEE Transactions on Visualization and Computer Graphics* 27 (2): 1753–63. https://doi.org/10.1109/TVCG.2020.3028985.

Lucchesi, Lydia R., Petra M. Kuhnert, Jenny L. Davis, and Lexing Xie. 2022. “Smallset Timelines: A Visual Representation of Data Preprocessing Decisions.” In *2022 ACM Conference on Fairness,*