MONASH University

{xmap}: Unified Tools for Data Harmonisation

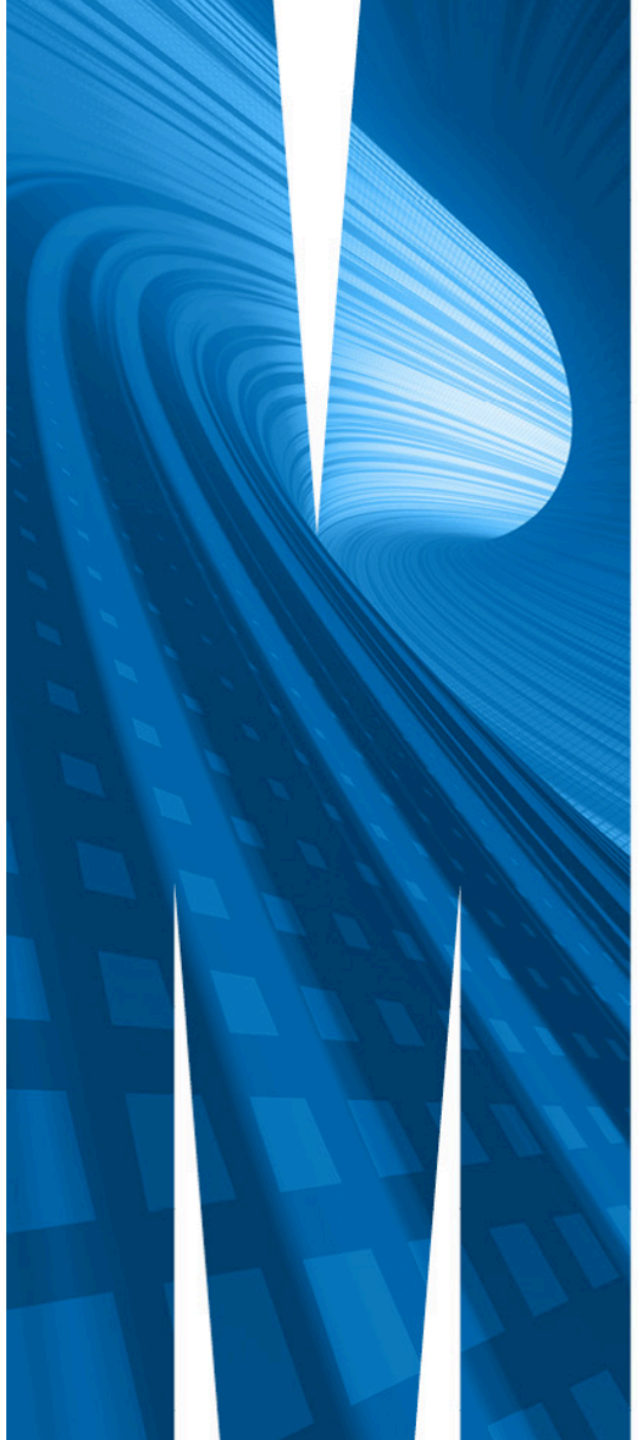
Presented at UseR! 2024, 9 July 2024

Cynthia A. Huang

cynthia.huang@monash.edu

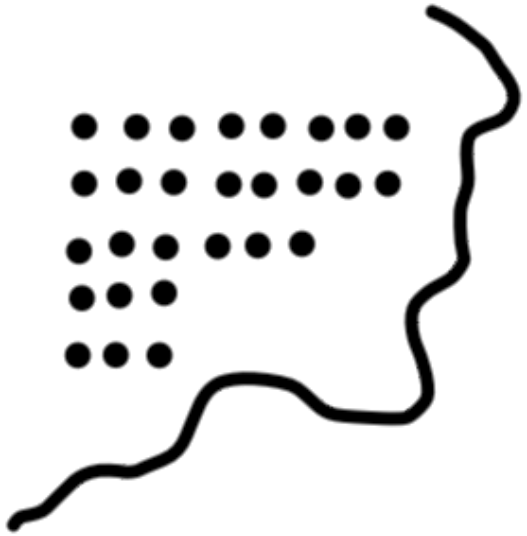
Department of Econometrics and Business Statistics

supervised by Rob J Hyndman, Sarah Goodwin and Simon Angus



Overview

Motivation

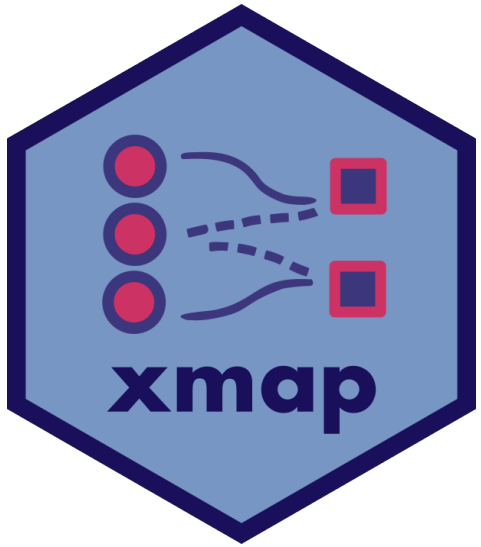


Official Statistics+

Conceptual Solution

Crossmaps Framework

Implementation



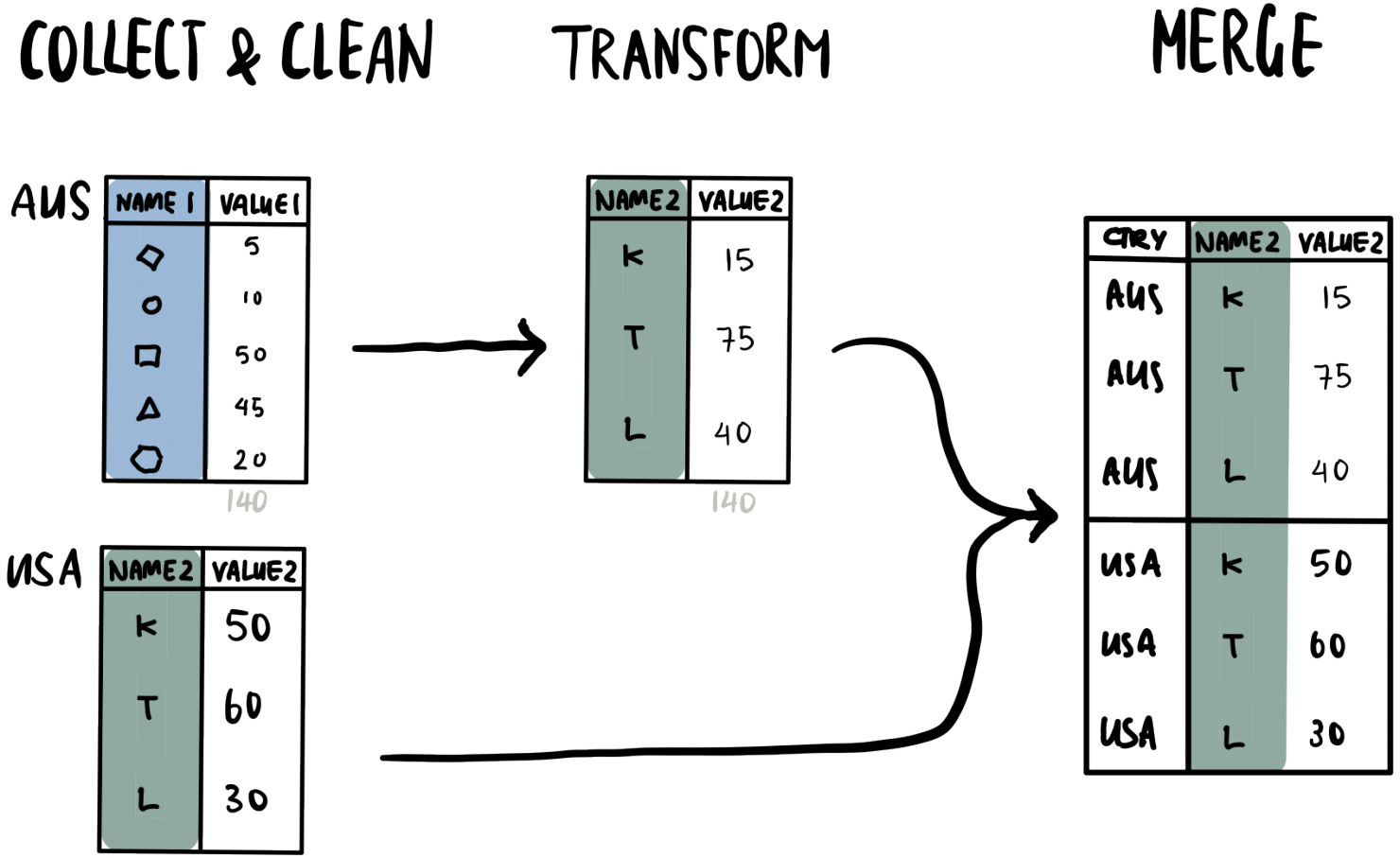
`xmap` Toolkit and Workflows

Motivation

Ex-Post Harmonisation of Survey Statistics

Harmonisation of Aggregate Statistics

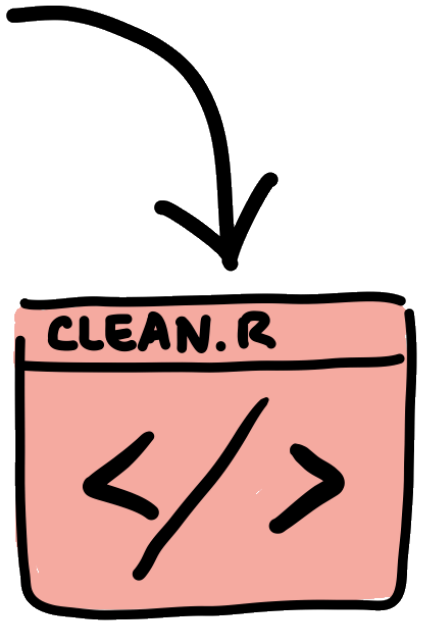
Combining **semantically related** data collected under **distinct survey instruments** into a single analysis dataset.



Current Workflow

AUS

NAME1	VALUE1
◇	5
○	10
□	50
△	45
⬡	20



AUS'

NAME2	VALUE2
K	15
T	75
L	40

BESPOKE

```
IF ◇ . THEN K = ... .

ASSERT_THAT(
  SUM(VALUE1) ==
  Sum(VALUE2) ,

  nrow(AUS) >
  nrow(AUS'
)
```

Current Workflow

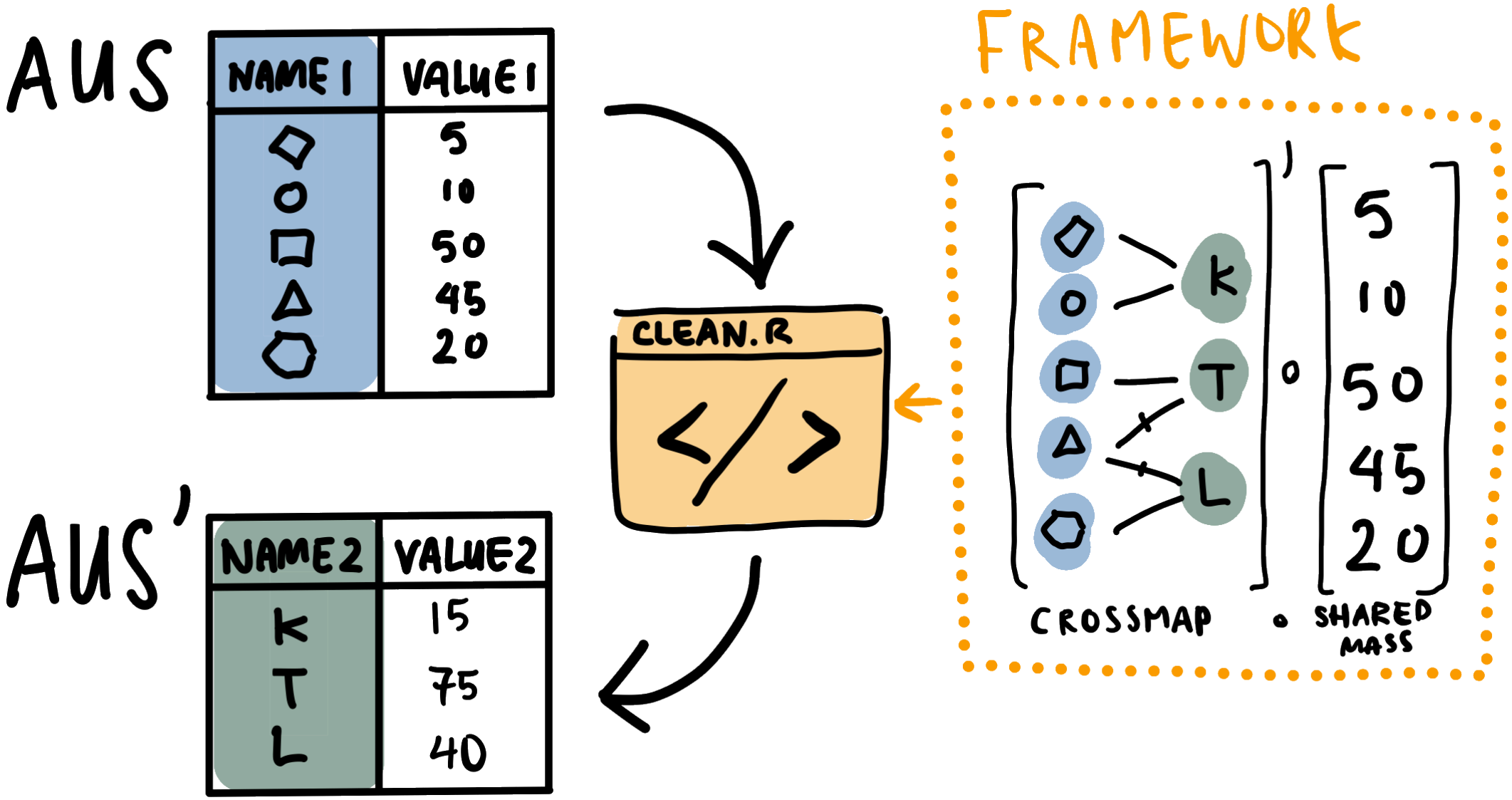
schott_algorithm_28.do [800+ lines]

```
/*

HS-SIC-NAICS- Concordance Project

This program

   1. reads in the hs-sic and hs-naics concordances from the monthly trade cd files and
         from Census and uses two mechanical matches to fill in naics matches prior to 2000
         matches after 2001.
  2. Although step 1 succeeds in generating many matches, there are some that remain unmatch
```

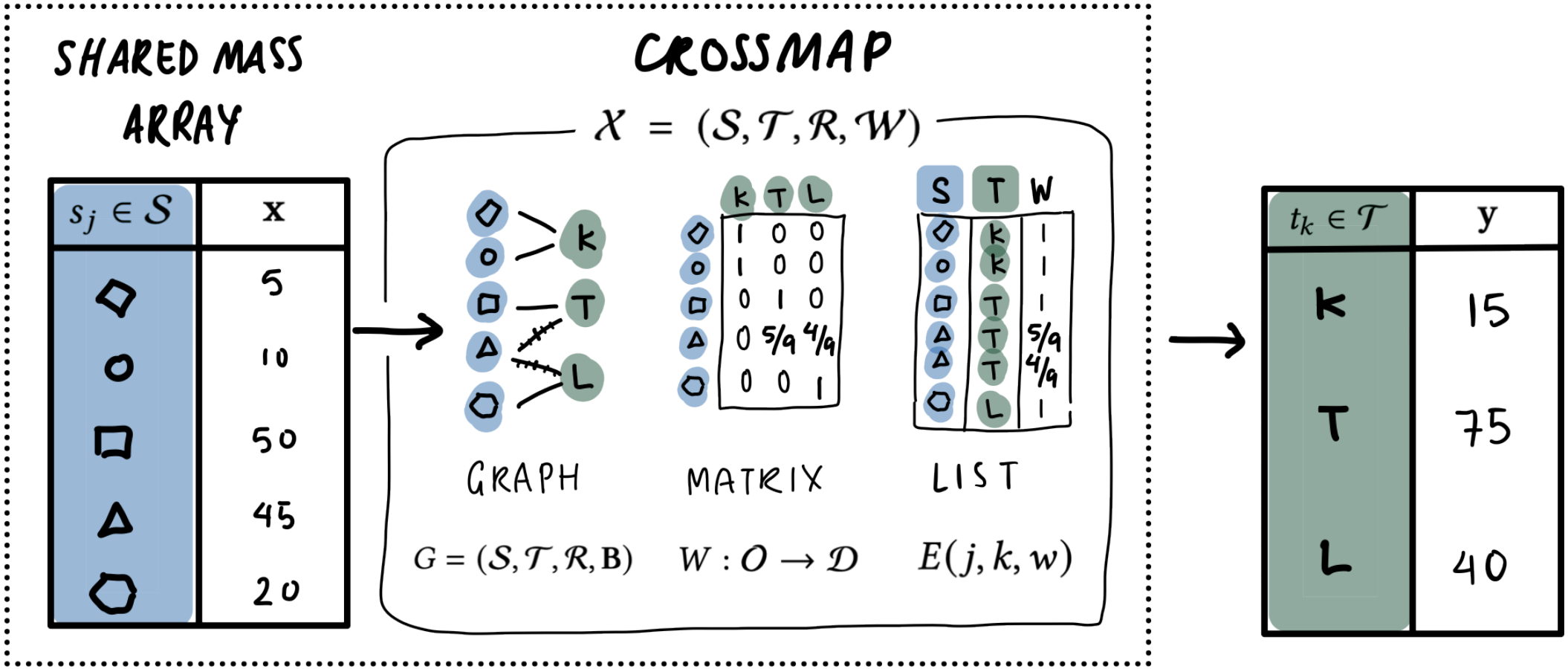
Conceptual Solution

Crossmaps Framework

Proposed Solution

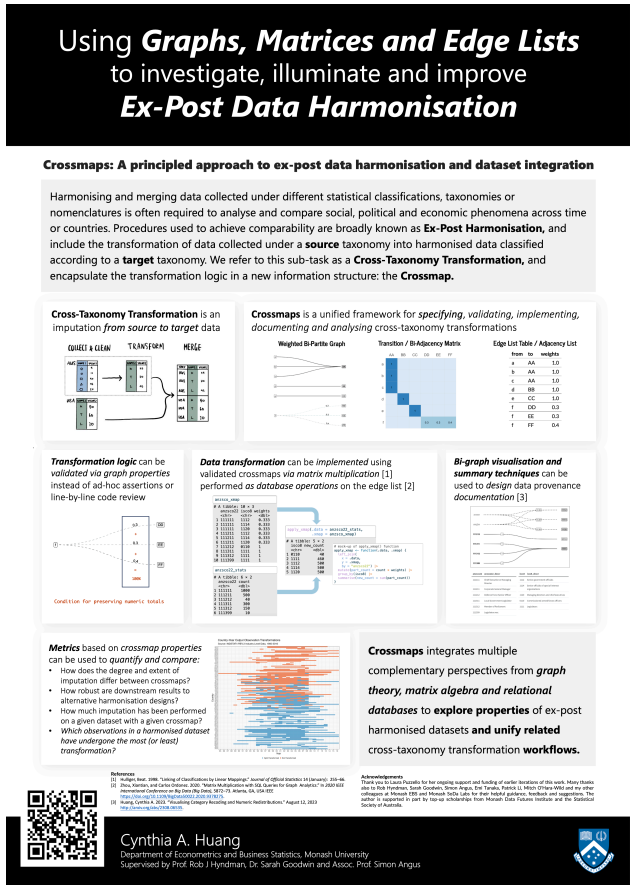


Crossmaps Framework

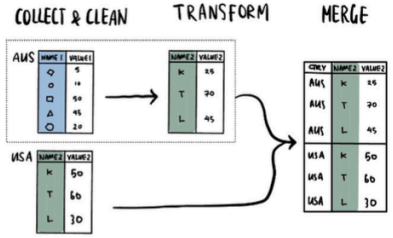


For details see preprint: A Unified Statistical And Computational Framework For Ex-Post Harmonisation Of Aggregate Statistics <arxiv.org/abs/2406.14163>

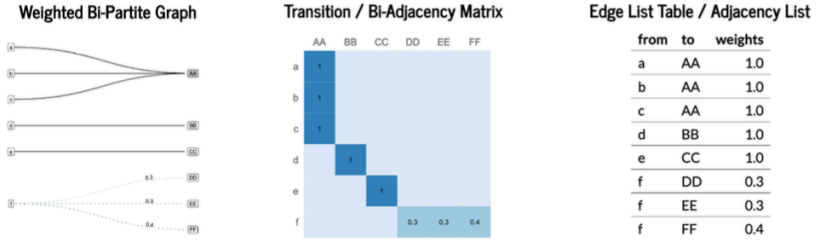
Unified and Principled Workflows



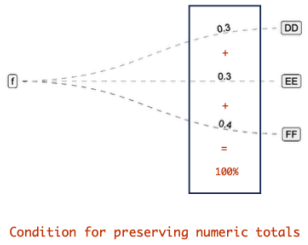
Cross-Taxonomy Transformation is an imputation *from source to target* data



Crossmaps is a unified framework for *specifying*, *validating, implementing, documenting and analysing* cross-taxonomy transformations

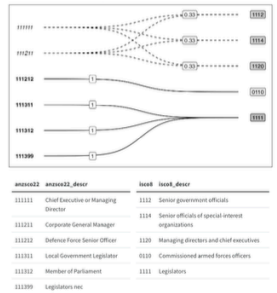


Transformation logic can be *validated via graph properties* instead of ad-hoc assertions or line-by-line code review



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

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



For details see poster on cynthiahqy.com

Implementation in R

Data Structures and Functions for using the Crossmaps Approach

Software Overview

Core Features

Specify and Validate Mappings

- `{dplyr}` verbs
- `as_xmap_tbl()`
- `diagnose_as_xmap_tbl()`

Match and Apply Transformations

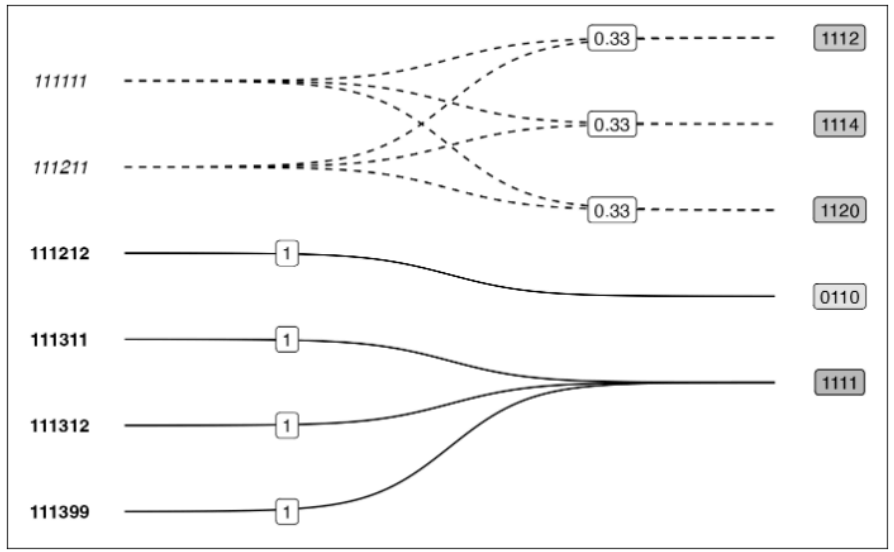
- `apply_xmap()`
- `diagnose_apply_xmap()`

Auxillary Features [WIP]

- extract logic from existing scripts
- graph and matrix classes
- symbolic fractional weights
- manage collections of crossmaps
- analysis of crossmap properties
- visualise and edit crossmap graphs
- …

Example: Occupation Code Mappings

ANZSCO22 to ISCO8



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

Stylised Occupation Level Counts

anzsco22	count
111111	1000
111211	500
111212	40
111311	300
111312	150
111399	10

Hypothetically aggregated from **2000** individual responses

Creating a Valid Crossmap

Start with *crosswalk*, or *lookup table*:

```
# A tibble: 10 × 5
   anzsco22 anzsco22_descr                    isco8 partial isco8_descr
   <chr>    <chr>                             <chr> <chr>   <chr>
 1 111111   Chief Executive or Managing Director 1112  p       Senior governmen…
 2 111111   Chief Executive or Managing Director 1114  p       Senior officials…
 3 111111   Chief Executive or Managing Director 1120  p       Managing directo…
 4 111211   Corporate General Manager          1112  p       Senior governmen…
 5 111211   Corporate General Manager          1114  p       Senior officials…
 6 111211   Corporate General Manager          1120  p       Managing directo…
 7 111212   Defence Force Senior Officer       0110  p       Commissioned arm…
 8 111311   Local Government Legislator        1111  p       Legislators
 9 111312   Member of Parliament               1111  p       Legislators
10 111399   Legislators nec                    1111  p       Legislators
```

Creating a Valid Crossmap

Add some naive (equal) distribution weights:

```r
library(dplyr)
library(xmap)

crosswalk <-
  xmap::demo$anzsco22_isco8_crosswalk |>
  select(anzsco22, isco8)

(
links <- crosswalk |>
  group_by(anzsco22) |>
  mutate(equal = 1 / n_distinct(isco8)) |>
  ungroup()
)
```

```
# A tibble: 10 × 3
   anzsco22 isco8 equal
   <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
```

Creating a Valid Crossmap

… and coerce to a crossmap:

```
(
occp_xmap <- links |>
  xmap::as_xmap_tbl(
    from = anzsco22,
    to = isco8,
    weight_by = equal
  )
)
```

```
# A crossmap tibble: 10 × 3
# with unique keys:  [6] anzsco22 -> [5] isco8
   .from$anzsco22 .to$isco8 .weight_by$equal
   <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
```

! ERROR: Invalid Weights

What if we try to naively use unit weights?

```
1  crosswalk |>
2    mutate(ones = 1) |>
3    xmap::as_xmap_tbl(
4      from = anzsco22,
5      to = isco8,
6      weight_by = ones
7    )
```

```
Error in `xmap_tbl()`:
! Invalid `.weight_by` found for some links
✖ The total outgoing `.weight_by` for some
`.from` nodes are not near enough to
  1
ℹ Modify `.weight_by` or adjust `tol` and try
again.
ℹ Use `diagnose_xmap_tbl() for more
information.
```

```
1  crosswalk |>
2    mutate(ones = 1) |>
3    xmap::diagnose_as_xmap_tbl(
4      anzsco22, isco8,
5      weight_by = ones
6    )
```

```
$bad_dups
NULL

$miss_weight_by
NULL

$bad_froms
# A tibble: 2 × 2
  .from$anzsco22 .sum.weight_by
  <chr>                   <dbl>
1 111111                      3
2 111211                      3
```

Applying a Valid Crossmap to Conformable Data

Stylised counts from before…

```
(occp_stats <-
  xmap::demo$anzsco22_stats |>
  mutate(ref = 100)
)
```

```
# A tibble: 6 × 3
  anzsco22 count   ref
  <chr>    <dbl> <dbl>
1 111111    1000   100
2 111211     500   100
3 111212      40   100
4 111311     300   100
5 111312     150   100
6 111399      10   100
```

Transformed, with redistribution and aggregation:

```
occp_stats |>
  xmap::apply_xmap(
    .xmap = occp_xmap,
    values_from = c(count, ref),
    keys_from = anzsco22
  )
```

```
# A tibble: 5 × 3
  isco8 count   ref
  <chr> <dbl> <dbl>
1 0110     40 100
2 1111    460 300
3 1112    500  66.7
4 1114    500  66.7
5 1120    500  66.7
```

Verifying Crossmaps and Transformations

Helper functions for keeping inside crossmap guardrails:

`diagnose_as_xmap_tbl()`:

- weights from a given `.from` key sum to one, to preserve total mass before and after
- no missing values in the links
- no duplicated links

`diagnose_apply_xmap()`:

- all key-value pairs in `.data` have matching instructions in `.xmap`
- no missing values in `.data`

Thanks for listening!

Try the package:

- Install from GitHub: cynthiahqy/xmap
- …soon to be on CRAN

Ask me about related work and applications:

- 🇦🇹 ***UseR!, Salzburg (Jul 8-11)***
- 🇺🇸 JSM, Portland (Aug 3-9)
- 🇺🇸 posit::conf(2024), Seattle (Aug 12-14)
- 🇨🇦 UBC, Vancouver, (Jul-Nov)
- 🇦🇺 Monash University, Melbourne (Nov onwards)

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