

Silver Linings:
*How a Data Harmonisation Mishap
Turned into a PhD Thesis*

Cynthia Huang

PhD Candidate at Monash University, Australia

Once upon a time...

I was a SoDa Labs RA..

November 12th, 2019 ▾



9:47 AM **sangus** Hi @Cynthia!

@Laura Puzzello and I are looking into an **Alternative data approach to trade, knowledge and prosperity.**

As part of this, we have a discrete job that we need to get done in the next four weeks or so (for funding reasons), which would be to automate a scraping exercise to get **output data at the country-industry-year level.**

@Laura Puzzello has the details on specifics but I think you would be well able to progress this, if you had the time.

I would guess it would take around half a day to a day of work to set up, troubleshoot and so on. Perhaps **5-10 hours.** (but fine if more than this .. there are always problems...!).

Have you any availability at the moment?

10:49 AM **Cynthia** Hi @sangus & @Laura Puzzello,

I do have availability at the moment so happy to help out. Do you need me to come into campus for briefing?

Collecting data..



HOME DATABASES RESOURCES SDG 9 LATEST TRENDS CIP HELP LOGIN/REGISTER

CHANGE SELECTION

Countries Variables Period ISIC

Countries Country group lists

Please, select at least one country: [Select all](#) [Unselect all](#)

Available:

Search...

- Afghanistan
- Albania
- Algeria
- Angola
- Armenia
- Australia
- Austria

[View Data](#) [Save Query](#)

Industrial statistical data and metadata are provided here. Please, choose the country, variables etc first:

Section D	Manufacturing
Division 15	Manufacture of food products and beverages
151	Production, processing and preservation of meat, fish, fruit, vegetables, oils and fats
1511	Production, processing and preserving of meat and meat products
1512	Processing and preserving of fish and fish products
1513	Processing and preserving of fruit and vegetables
1514	Manufacture of vegetable and animal oils and fats
152	1520 Manufacture of dairy products
153	Manufacture of grain mill products, starches and starch products, and prepared animal feeds
	1531 Manufacture of grain mill products
	1532 Manufacture of starches and starch products
	1533 Manufacture of prepared animal feeds
154	Manufacture of other food products
	1541 Manufacture of bakery products
	1542 Manufacture of sugar
	1543 Manufacture of cocoa, chocolate and sugar confectionery
	1544 Manufacture of macaroni, noodles, couscous and similar farinaceous products
	1549 Manufacture of other food products n.e.c.
155	Manufacture of beverages
	1551 Distilling, rectifying and blending of spirits; ethyl alcohol production from fermented materials
	1552 Manufacture of wines
	1553 Manufacture of malt liquors and malt
	1554 Manufacture of soft drinks; production of mineral waters

INDSTAT 4 2018, ISIC Revision 3: Please choose Metadata

Database:

Please choose Metadata

- Please choose Metadata
- Country List
- ISIC Code List
- ISIC Combination Code List
- Variable list (Table Code List)
- Variable list (Table Definition Code List)

[View Metadata](#)



Importing data..

14-Output.csv

```
> nested_indstat
# A tibble: 1,587 × 4
# Rowwise: country, year
  country year isic_rev data
  <chr>   <dbl> <dbl> <list<tibble[,7]>>
1 004     2002     3 [152 × 7]
2 004     2003     3 [152 × 7]
3 004     2004     3 [152 × 7]
4 004     2005     3 [152 × 7]
5 004     2006     3 [152 × 7]
6 004     2007     3 [152 × 7]
7 004     2008     3 [152 × 7]
8 004     2009     3 [152 × 7]
9 004     2010     3 [152 × 7]
10 004     2011     3 [152 × 7]
# i 1,577 more rows
# i Use `print(n = ...)` to see more rows
```

```
> nested_indstat$data[[1]]
# A tibble: 152 × 7
  ctable isic isiccomb value utable source unit
  <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>
1 14 151 151 NA 14 0 $
2 14 1511 1511
3 14 1512 1512
4 14 1513 1513
5 14 1514 1514
6 14 1520 1520
7 14 153 153
8 14 1531 1531
9 14 1532 1532
10 14 1533 1533
# i 142 more rows
# i Use `print(n = ...)` to see more rows
```

```
> nested_indstat$data[[35]]
# A tibble: 152 × 7
  ctable isic isiccomb value utable source unit
  <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>
1 14 151 151 13879638 14 1 $
2 14 1511 1511 5998235 14 1 $
3 14 1512 1512 1836669 14 1 $
4 14 1513 1513 2580636 14 1 $
5 14 1514 1514 3464097 14 1 $
6 14 1520 1520 137331692 14 1 $
7 14 153 153 36291649 14 1 $
8 14 1531 1531 12252210 14 1 $
9 14 1532 1532 0 14 1 $
10 14 1533 1533 24039440 14 1 $
# i 142 more rows
# i Use `print(n = ...)` to see more rows
```

Conversion needed!

The screenshot shows the UNstats website interface. On the left, a list of manufacturing categories is shown under 'MANUFACTURING'. In the center, a table compares ISIC Rev. 3, ISIC Rev. 3.1, and ISIC Rev. 4. On the right, a 'NAVIGATION' panel shows 'Section C Manufacturing' with a list of divisions and classes.

ISIC Rev. 3	ISIC Rev. 3.1	ISIC Rev. 4
[Grid Icon]	[Grid Icon] [Document Icon] [Folder Icon]	-
-	[Grid Icon] [Document Icon] [Folder Icon]	[Grid Icon]
[Grid Icon] [Document Icon] [Folder Icon]	-	[Grid Icon]
-	[Grid Icon]	-
-	[Grid Icon] [Document Icon] [Folder Icon]	-
-	-	[Grid Icon] [Document Icon] [Folder Icon]
[Grid Icon] [Document Icon] [Folder Icon]	-	[Grid Icon] [Document Icon] [Folder Icon]
-	-	-

Section C Manufacturing

Division	Group	Class	Description
Division 10			
Manufacture of food products			
101		1010	Processing and preserving of meat
102		1020	Processing and preserving of fish, crustaceans and molluscs
103		1030	Processing and preserving of fruit and vegetables
104		1040	Manufacture of vegetable and animal oils and fats
105		1050	Manufacture of dairy products
106			Manufacture of grain mill products, starches and starch products
		1061	Manufacture of grain mill products
		1062	Manufacture of starches and starch products
107			Manufacture of other food products
		1071	Manufacture of bakery products
		1072	Manufacture of sugar
		1073	Manufacture of cocoa, chocolate and sugar confectionery
		1074	Manufacture of macaroni, noodles, couscous and similar farinaceous products
		1075	Manufacture of prepared meals and dishes
		1079	Manufacture of other food products n.e.c.
108		1080	Manufacture of prepared animal feeds
Division 11			
Manufacture of beverages			
		1101	Distilling, rectifying and blending of spirits
		1102	Manufacture of wines
		1103	Manufacture of malt liquors and malt
		1104	Manufacture of soft drinks; production of mineral waters and other bottled waters

Ready to harmonise...

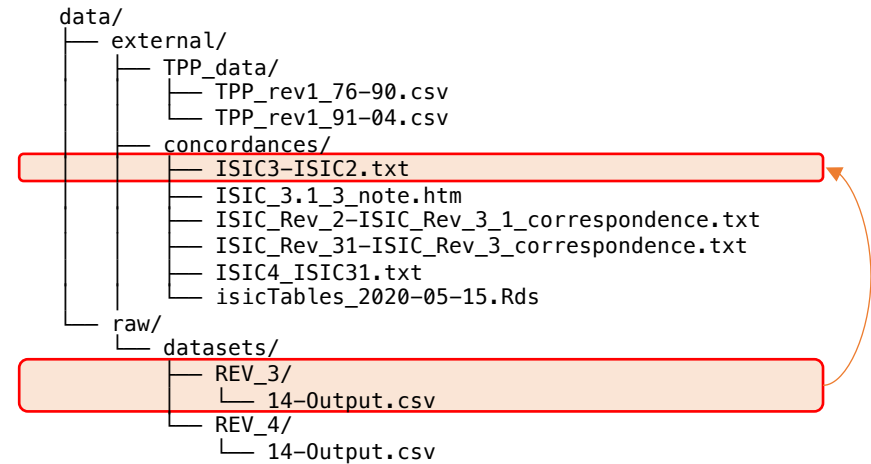
UNSD - Classifications on economic statistics

ISIC - UN Correspondence Tables

FROM / TO	ISIC	ISIC Rev. 2	ISIC Rev. 3	ISIC Rev. 3.1	ISIC Rev. 4
ISIC Rev. 2	-	-			-
ISIC Rev. 3	-		-		-
ISIC Rev. 3.1	-	-		-	-
ISIC Rev. 4	-	-	-		-
CPC Ver. 1.1	-	-	-	-	-
CPC Ver. 2	-	-	-	-	-
CPC Ver. 2.1	-	-	-	-	-
COFOG		-	-	-	-

```

$ISIC3_2
# A tibble: 586 × 5
  ISIC3 partialISIC3 ISIC2 partialISIC2 Detail
  <chr>          <dbl> <dbl>          <dbl> <chr>
1 0111            0  1110            1 Growing of cereals and other crops n.e.c
2 0112            1  1110            1 Growing of vegetables, horticultural specialities, nursery products
3 0112            1  1210            1 Gathering of mushrooms, truffles
4 0113            0  1110            1 Growing of fruit, nuts, beverage and spice crops
5 0121            0  1110            1 Farming of cattle, sheep, goats, horses, asses, mules and hinnies; dairy farming
6 0122            1  1110            1 Raising domesticated or wild animals n.e.c. (e.g. swine, poultry, rabbits)
7 0122            1  1120            1 Poultry hatchery, silkworm raising, on a fee or contract basis
8 0122            1  1302            1 Frog farming
9 0130            0  1110            1 Growing of crops combined with farming of animals (mixed farming)
10 0140           1  1110            1 Landscape gardening
# i 576 more rows
# i Use `print(n = ...)` to see more rows
  
```



Running scripts..

- ISIC_concordance.Rproj
- README.md
- code
 - fncs_cleaning.R
 - fncs_concordance.R
 - fncs_matchTPP.R
 - fncs_probing.R
 - get_concordance_tables.R
- data
 - archived
 - external
 - final
 - interim
 - raw
- notebooks
 - 001-clean_INDSTAT.Rmd
 - 002-concord_INDSTAT.Rmd
 - 003-match_TPP_INDSTAT.Rmd
 - 004-compare-matched.Rmd
- reference
 - ISIC manuals
 - TPP database
 - country codes
 - unido-metadata

```
INDSTAT-TPP CONCORDANCE (GROSS OUTPUT SERIES)
About Pre-Processing INDSTAT4 Concordance to ISIC Rev 2 Merge INDSTAT & TPP series Merge Diagonistics

Concordance Design
Transformation Pipeline
Diagnostics & Troubleshooting
Reference Code
Transformation Process
Transformation functions
Downloading concordance tables

# REV4
inISIC2.4digit$REV4_2step <-
  transform_later_into_earlier(concord_tidy$ISIC4_31, clean$REV4, ISIC4, ISIC31, needs_split_IS
IC4) %>%
  transform_later_into_earlier(concord_tidy$ISIC31_2, ., ISIC31, ISIC2, needs_split_ISIC31)

## Test passed 🎉
## Test passed 🎉

inISIC2.4digit$REV4_3step <-
  transform_later_into_earlier(concord_tidy$ISIC4_31, clean$REV4, ISIC4, ISIC31, needs_split_IS
IC4) %>%
  transform_later_into_earlier(concord_tidy$ISIC31_3, ., ISIC31, ISIC3, needs_split_ISIC31) %>%
  transform_later_into_earlier(concord_tidy$ISIC3_2, ., ISIC3, ISIC2, needs_split_ISIC3)

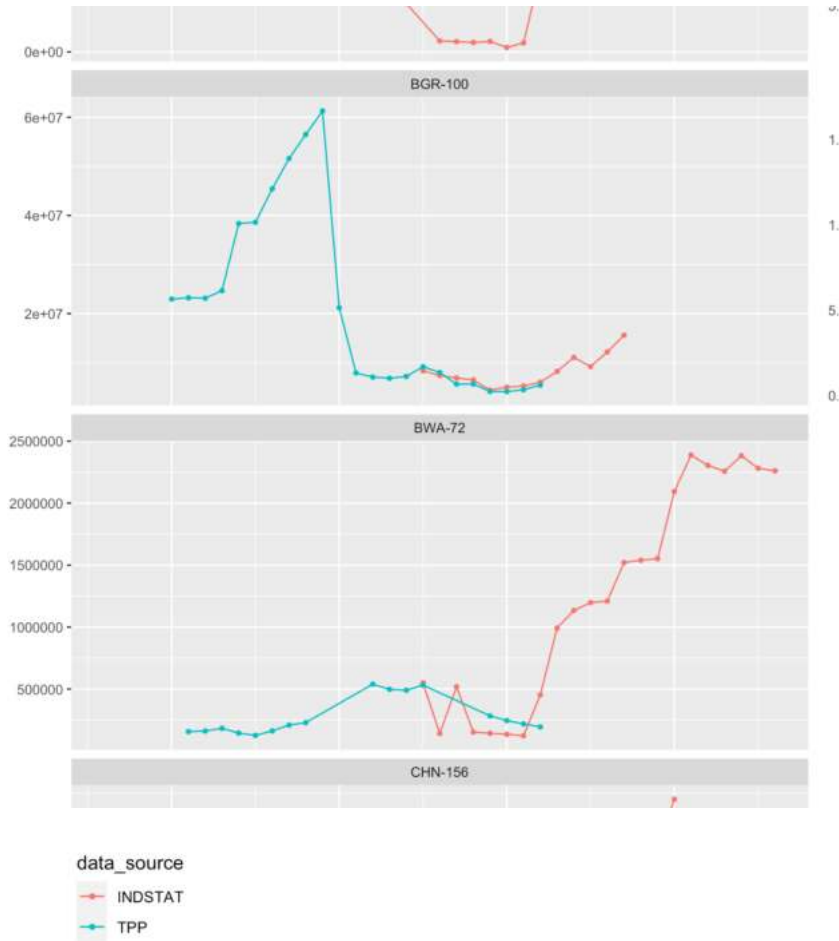
## Test passed 🎉
## Test passed 🎉
## Test passed 🎉

## ---- summarise from 4digit to 3digit codes ----
inISIC2.3digit <- map(inISIC2.4digit, summarise_to_3digit)

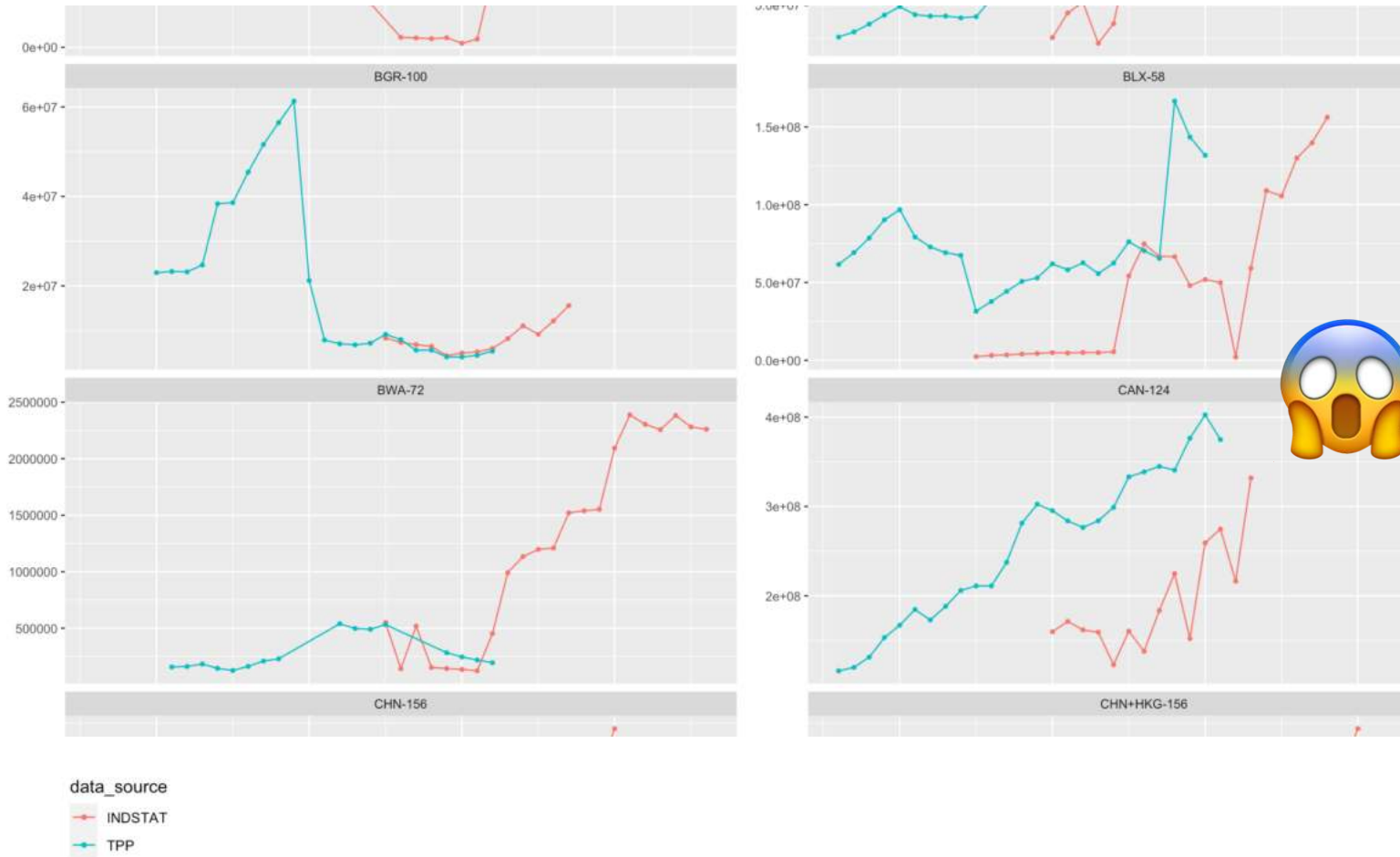
cache_3digit <- here("data/interim/", paste0("002-INDSTAT_ISIC2_3digit_", lubridate::today().
```



Validating data quality..



Issue found..



Error Located...



Cynthia Huang <cynthia.huang@monash.edu>

to Laura ▾

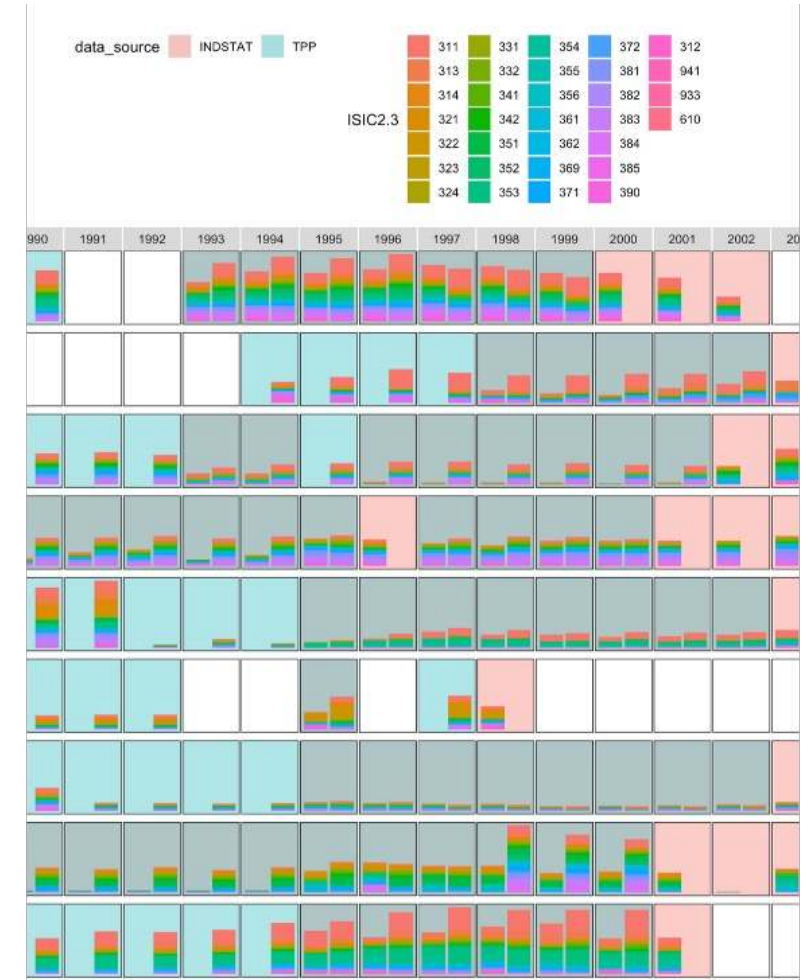
Hi Laura,

Many apologies, but it turns out I lost a bunch of rows when cleaning the INDSTAT data because I didn't correctly define the case where 4-digit totals exactly matched the corresponding 3-digit values.

The missing data is about 31,000 rows (163,782 - 132,191). This might explain the difference with the TPP data.

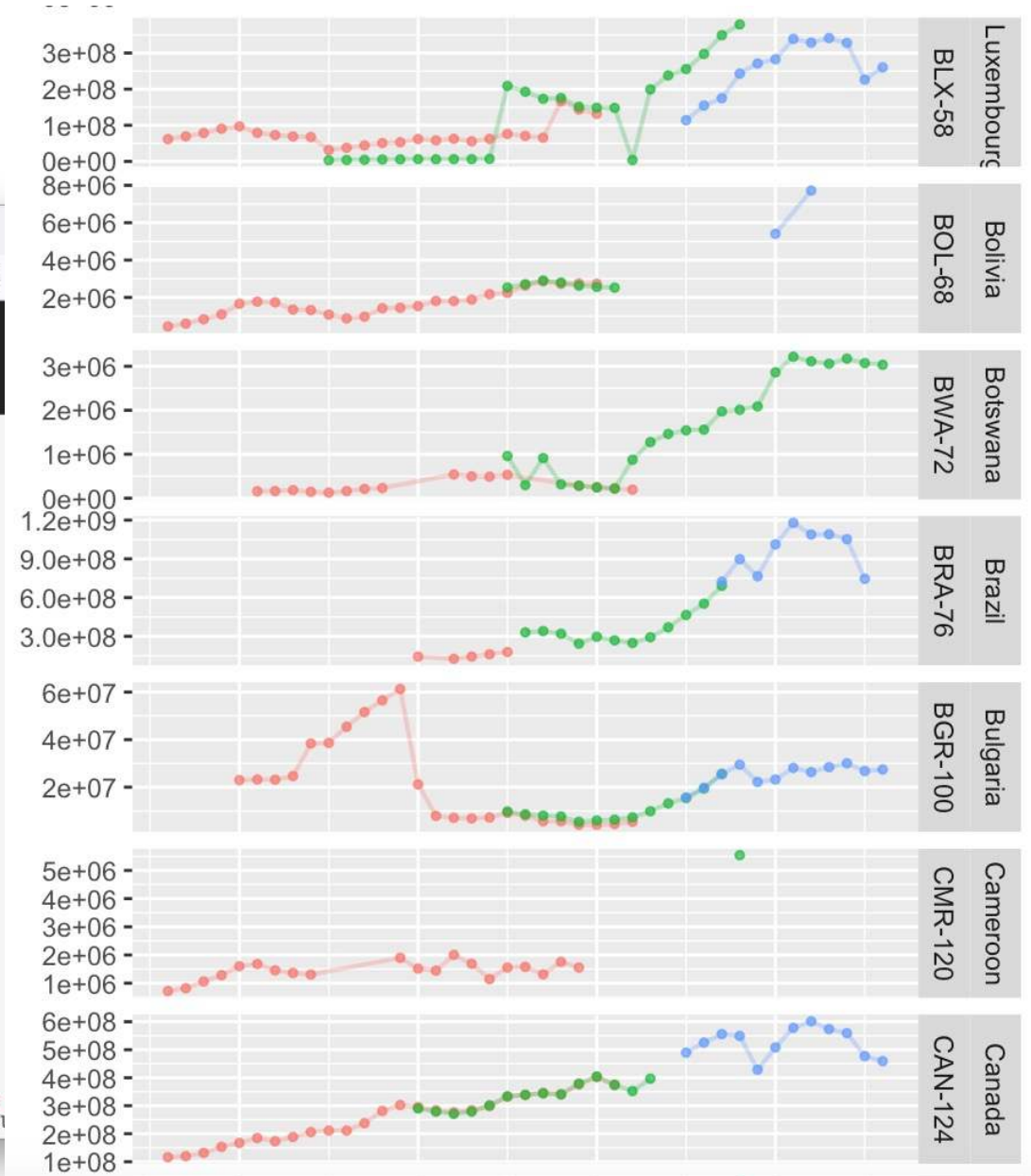
I should be able to send through updated data (remerged with TPP) tomorrow or Wednesday. I picked up the mistake when reworking the code for use with Revision 4, so I'm also not far from finishing up with that.

Hopefully you haven't spent too much time looking at the data yet. I'm very sorry about this.



...and fixed..

```
by = c('country', 'year', 'isic.3')
}
# wrapper for processing & tidying cases
process_3_cases <- function(comparison_df, four_df){
  # define cases
  larger <- split_cases_by_which.larger(comparison_df)
  # apply processing to each case
  processed <- list()
  processed$value3 <-
    add_avgDiff_from_value3(larger$value3, four_df)
  processed$equal <-
    keep_value4_only(larger$equal, four_df) %>%
    drop_na(value)
  processed$total4 <-
    keep_value4_only(larger$total4, four_df) %>%
    drop_na(value)
  # verify that all cases have been processed
  test_that("All cases have been processed",
    {
      n.cases <- length(larger)
      n.fixed <- length(processed)
      expect_equal(n.cases, n.fixed)
    })
  test_that("Extracted expected number of value.4 based on n_obs.4 in comparison",
    expected_n_obs <- comparison_df %>% filter(which.larger != "value3") %>% pull
```



~~Lessons~~

Questions for next time?

Is there a better way to harmonise data?

Why don't we try different mappings?

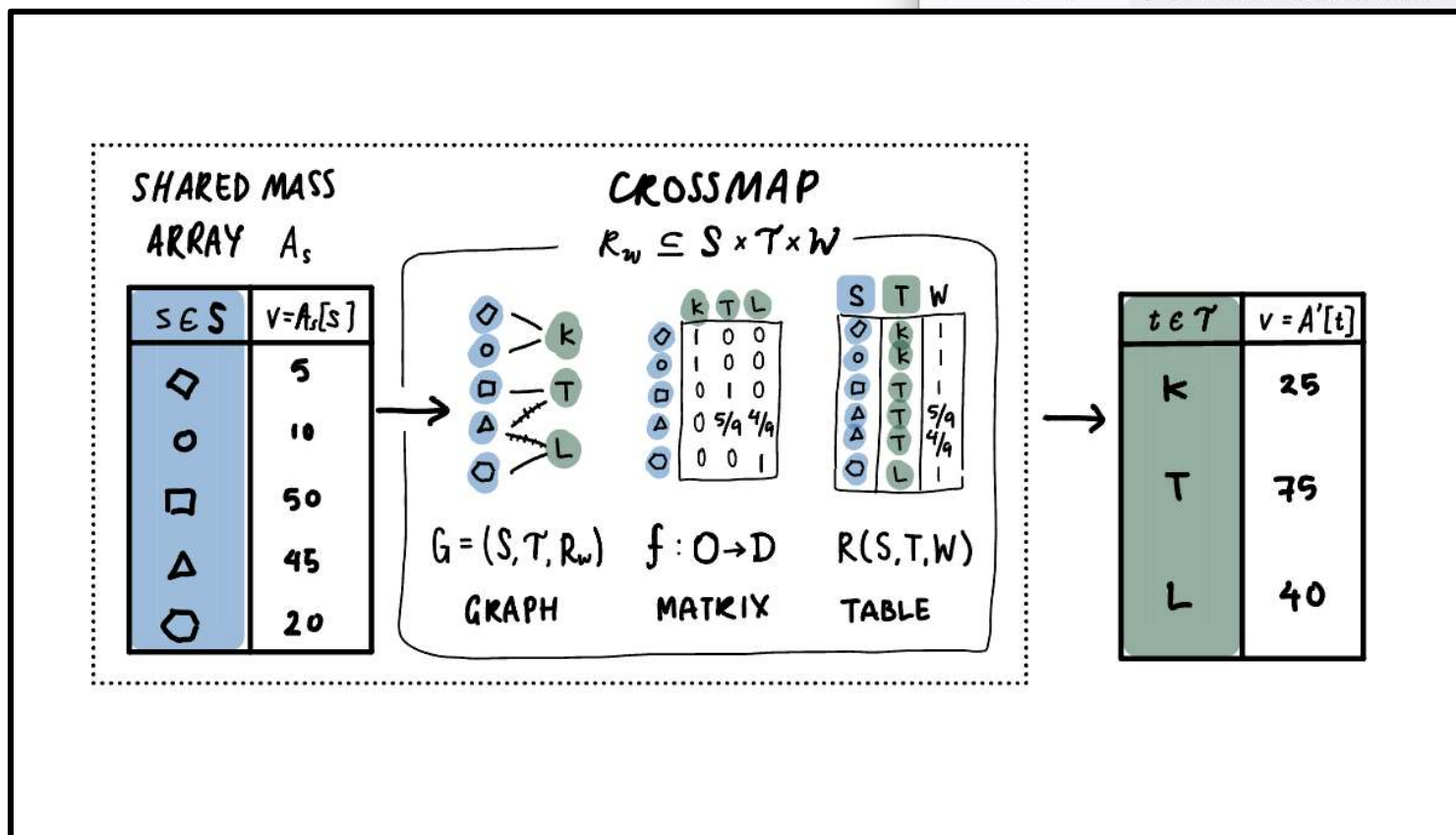
What's actually happening to data through multiple concordance steps?

What if we didn't have ground truth to compare with?

Answers... in my thesis?

Using **Graphs, Matrices and Edge Lists** to investigate, illuminate and improve **Ex-Post Data Harmonisation**

Crossmaps: A principled approach to ex-post data harmonisation and dataset integration



Cynthia Huang - Research

https://www.cynthiahq.com/research

SELECT & CLEAN

NAME1	VALUE1
Q	5
O	10
□	50
△	45
○	20

TRANSFORM

NAME2	VALUE2
K	25
T	70
L	45

MERGE

COPY	NAME2	VALUE
AU5	K	25
AU5	T	70
AU5	L	45
USA	K	50
USA	T	60
USA	L	30

Advances in Ex-Post Harmonisation using Graph Representations of Cross-Taxonomy Transformations

TALK

Initially presented at IDWSDS 2023 for the PhD contest on Oct 10, 2023, and then at NUMBATS group seminar on Oct 12, 2023.

ons, taxonomies or economic phenomena across time

Ex-Post Harmonisation, and armonised data classified

ny Transformation, and map.

precifying, validating, implementing, my transformations

gency Matrix

Edge List Table / Adjacency List

from	to	weights
A	AA	1.0
B	AA	1.0
C	AA	1.0
D	BB	1.0
E	CC	1.0
F	DD	0.3
G	EE	0.3
H	FF	0.4

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

integrates multiple tary perspectives from graph rix algebra and relational o explore properties of ex-post datasets and unify related my transformation workflows.

For ongoing support and handling of earlier iterations of this work, Maria Ramon (London, Simon Angus, Eui Tanaka, Paris Li, Muel CHen, Woe and my other Harman India Labs for their helpful guidance, feedback and suggestions. The y Harup scholarships from Harman Data Future Institute and the Statistical

Thanks for listening!

Talk to me: @cynthiahqy / cynthia.huang@monash.edu

Learn more: cynthiahqy.com/research