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…

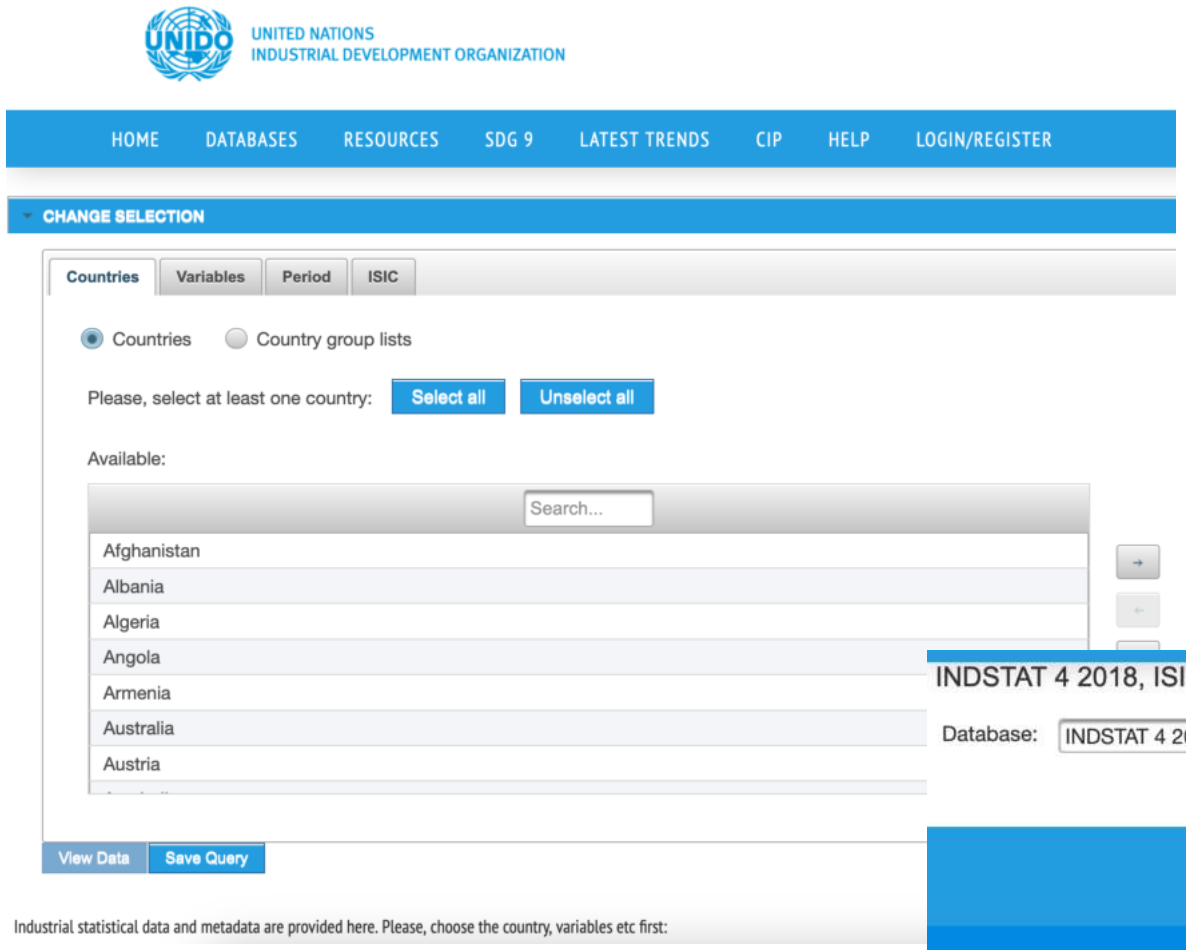
November 12th, 2019

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, trouble-shoot and so on. Perhaps 5-10 hours. (but fine if more than this .. there are always problems...!).
Have you any availability at the moment?

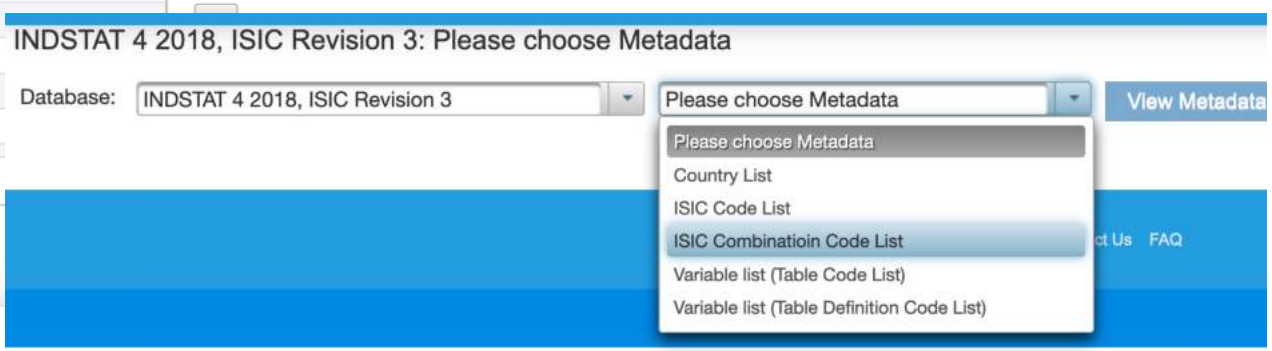
10:49 AM **Cynthia** Hi @sangus & @Laura Puzzello,
I do have availability at the moment so happy to help out. Do you need me to come into campus for briefing?

Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

Collecting data…



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



Collect → Import & Prepare → Harmonise → Validate → ERROR! → Fix → Reboot → Happily Research After…?

Importing data…

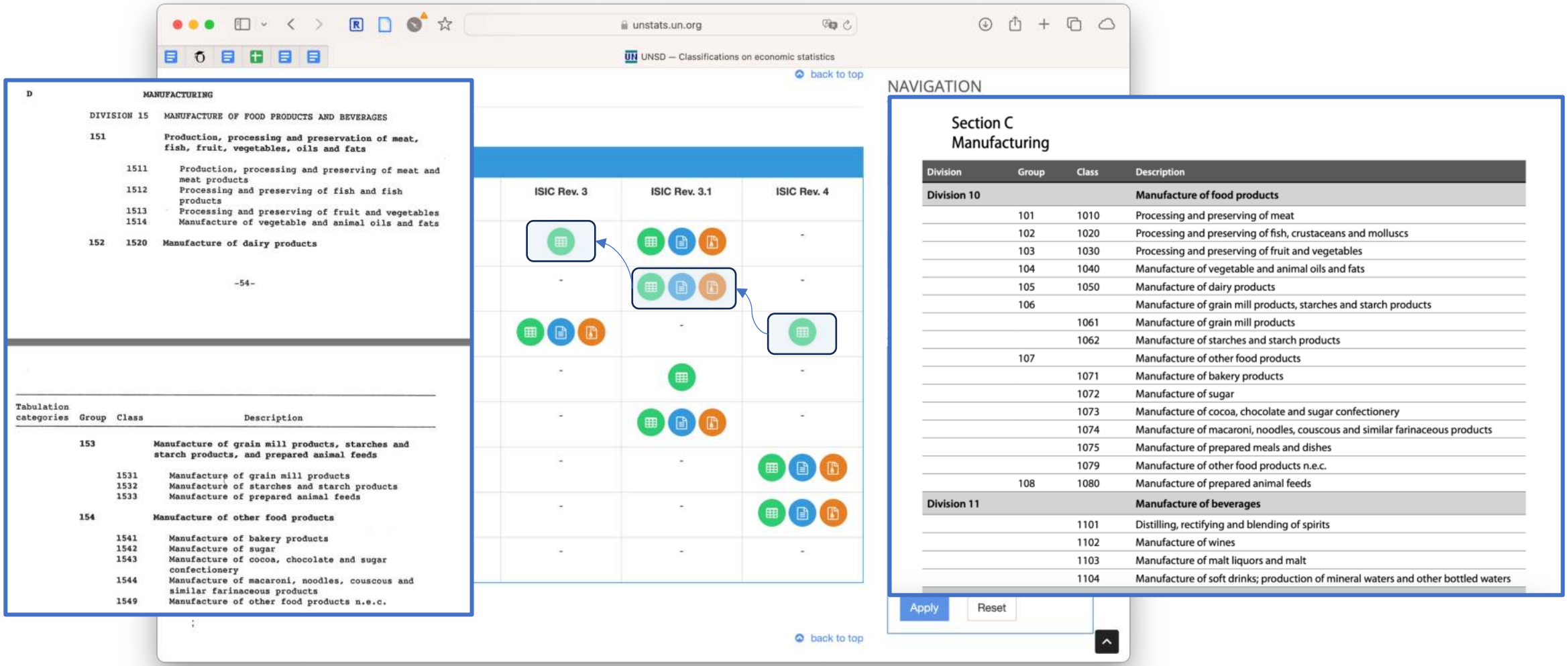
```
│   │   ├── 14-Output.csv
```

```
> 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  15
 3     14 1512  15
 4     14 1513  15
 5     14 1514  15
 6     14 1520  15
 7     14 153   15
 8     14 1531  15
 9     14 1532  15
10     14 1533  15
# ℹ 142 more rows
# ℹ Use `print(n =
```

```
> 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!



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

-54-

Tabulation categories	Group	Class	Description
	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.

unstats.un.org

UNSD — Classifications on economic statistics

back to top

ISIC Rev. 3	ISIC Rev. 3.1	ISIC Rev. 4
		-
-		-
	-	
-		-
-		-
-	-	
-	-	
-	-	-

back to top

NAVIGATION

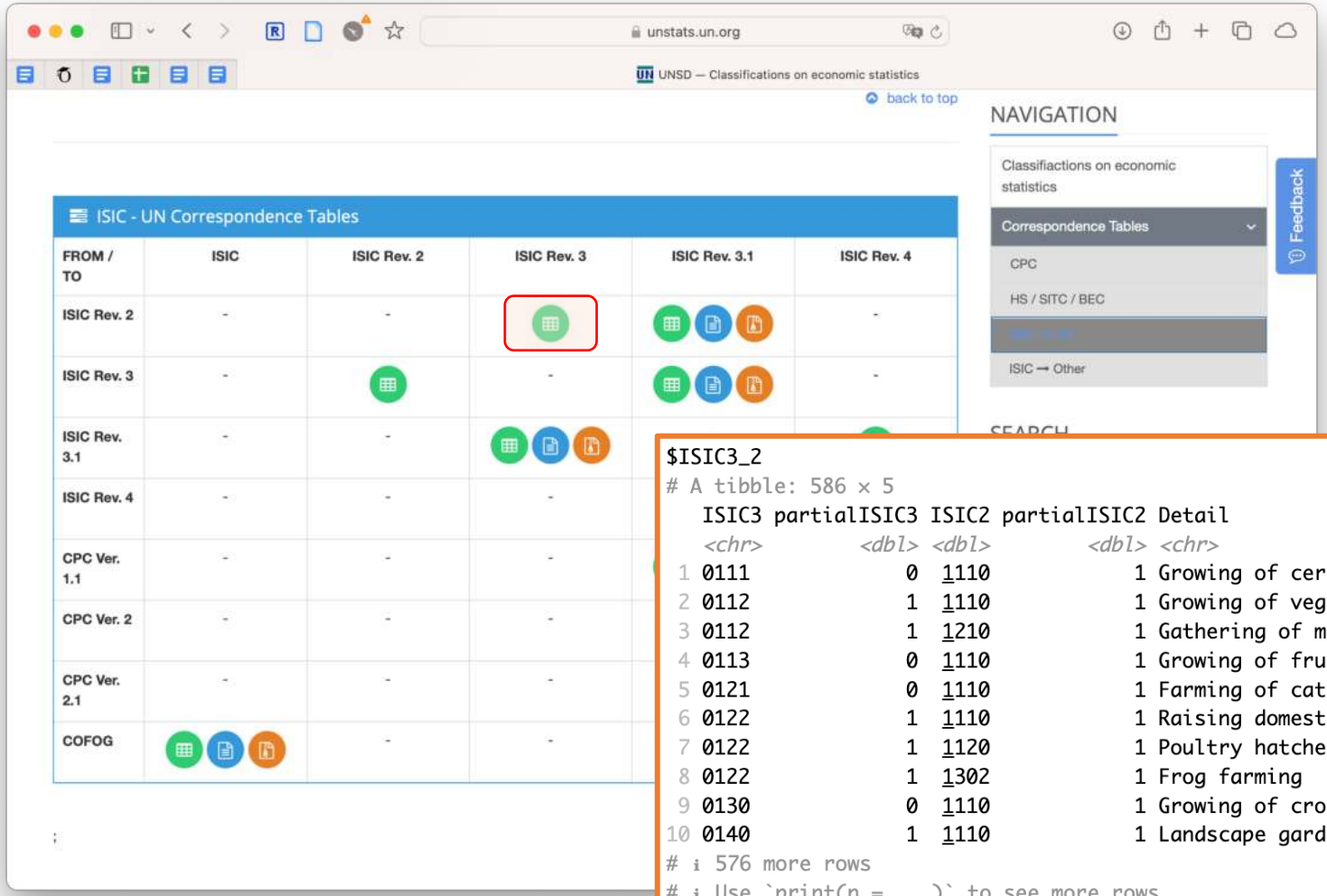
Section C
Manufacturing

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

Apply Reset

Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

Ready to harmonise…



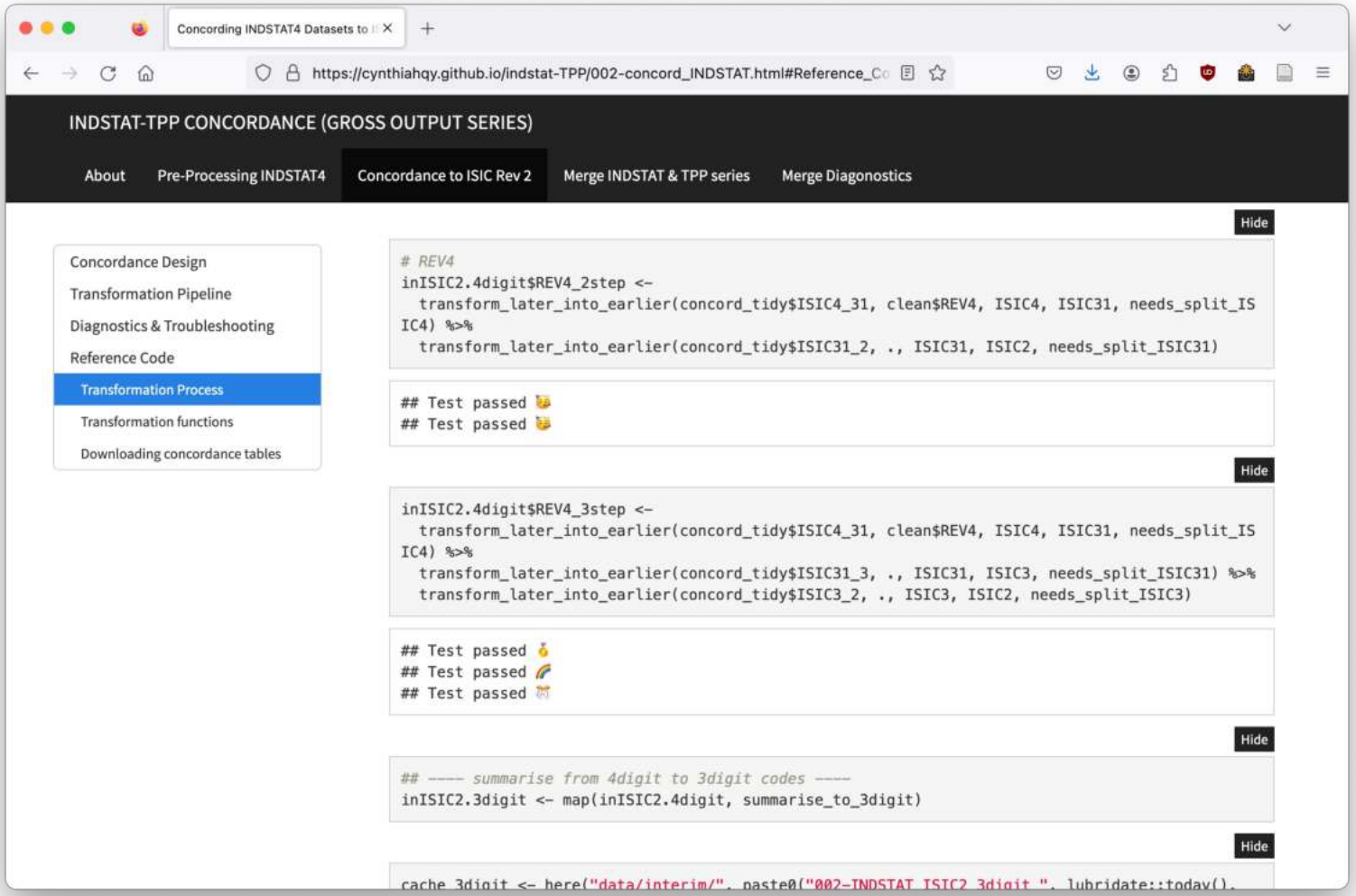

```
data/
├── external/
│   ├── TPP_data/
│   │   ├── TPP_rev1_76-90.csv
│   │   └── TPP_rev1_91-04.csv
│   └── concordances/
│       ├── ISIC3-ISIC2.txt
│       ├── ISIC_3.1_3_note.htm
│       ├── ISIC_Rev_2-ISIC_Rev_3_1_correspondence.txt
│       ├── ISIC_Rev_31-ISIC_Rev_3_correspondence.txt
│       ├── ISIC4_ISIC31.txt
│       └── isicTables_2020-05-15.Rds
└── raw/
    └── datasets/
        ├── REV_3/
        │   └── 14-Output.csv
        └── REV_4/
            └── 14-Output.csv
```

```
$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
```

Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

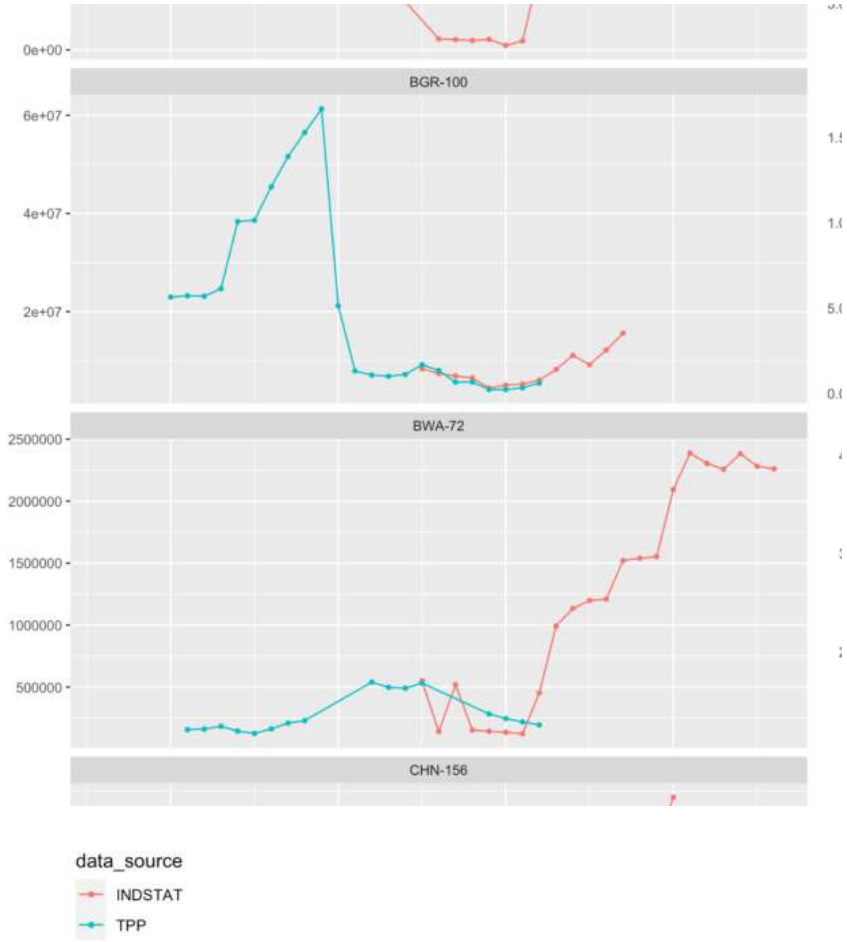
Running scripts…

```
.
├── 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
```

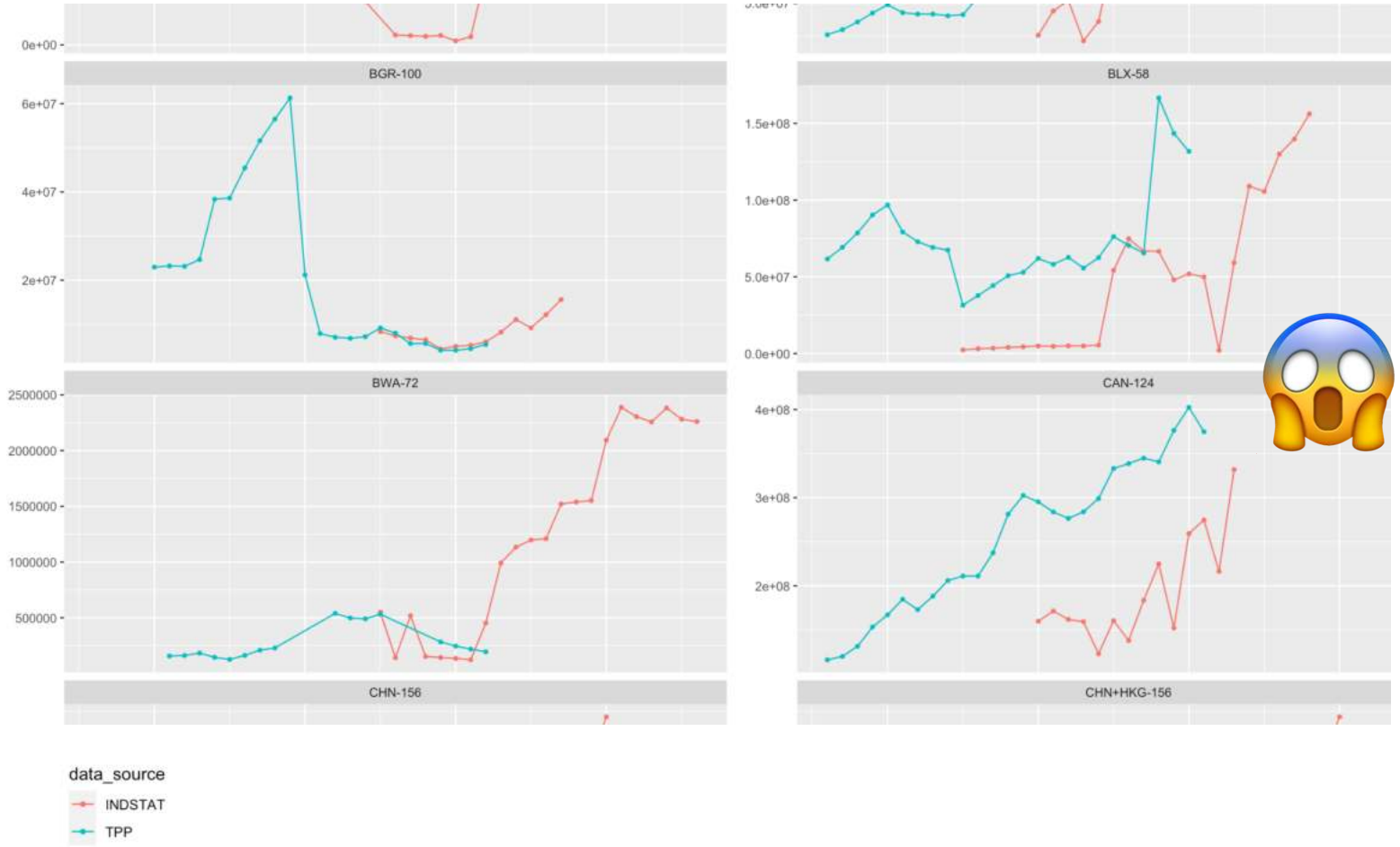


Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

Validating data quality…



Issue found…


BGR-100
BLX-58
BWA-72
CAN-124
CHN-156
CHN+HKG-156
0e+00
6e+07
4e+07
2e+07
1.5e+08
1.0e+08
5.0e+07
0.0e+00
2500000
2000000
1500000
1000000
500000
4e+08
3e+08
2e+08
data_source
INDSTAT
TPP

Error located…

Cynthia Huang <cynthia.huang@monash.edu>
to Laura

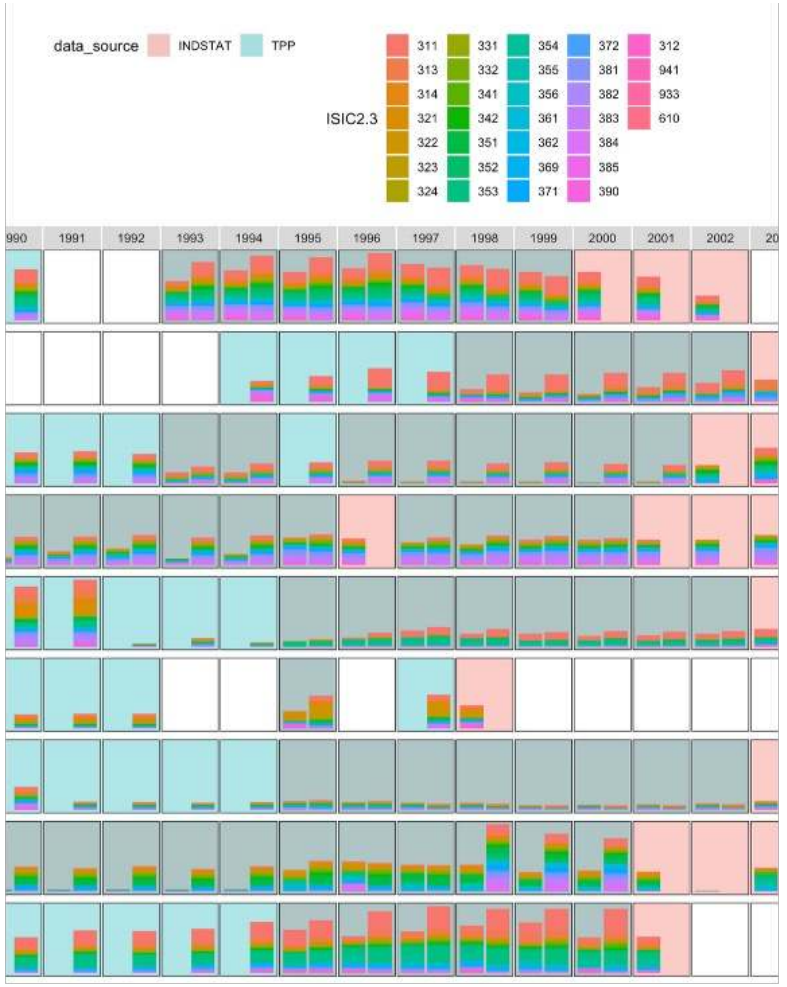
Hi Laura,

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.

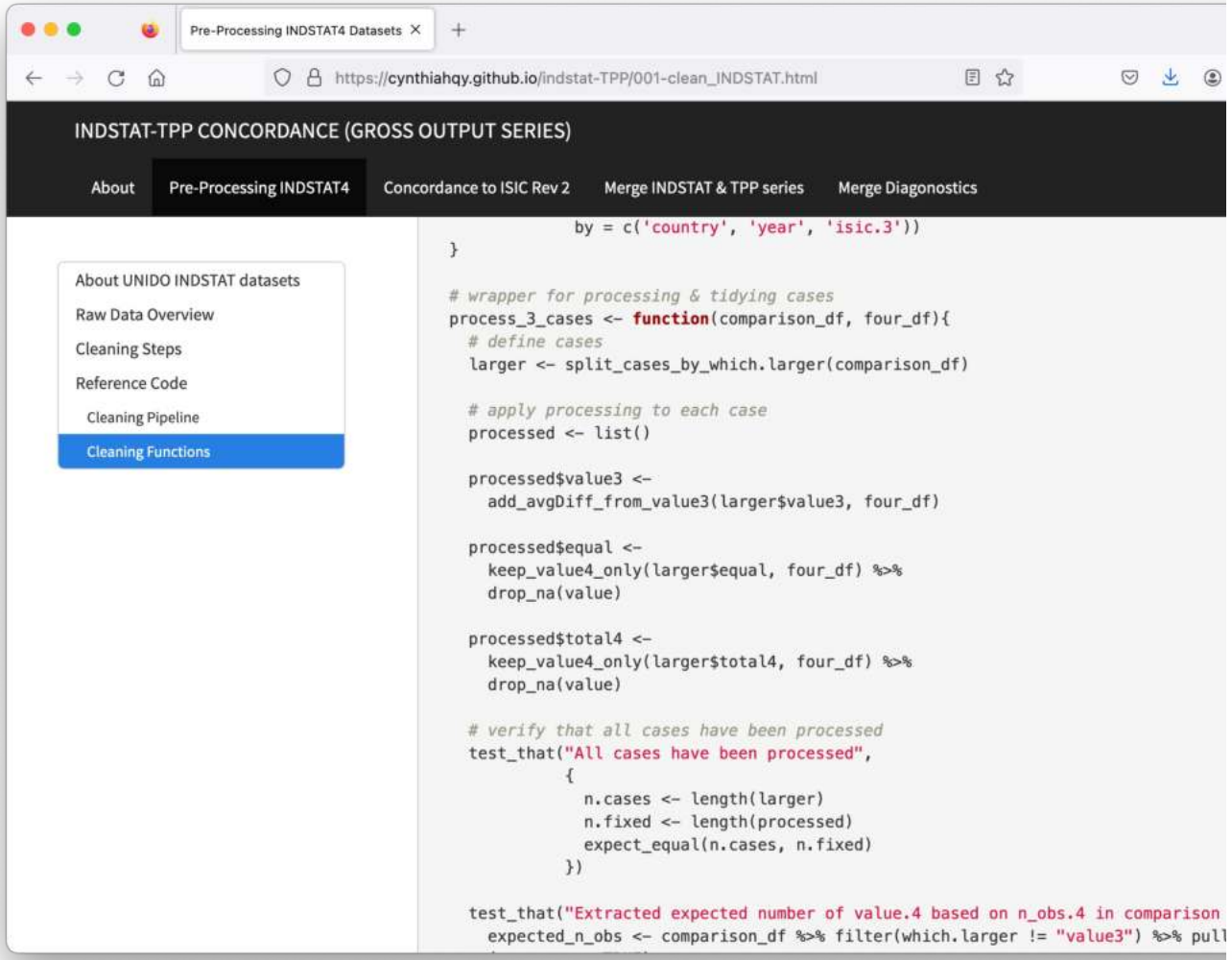
I should be able to send through updated data (remerged with TPP) tomorrow or Wednesday. I picked up the mistake when reworking the code for use with Revision 4, so I'm also not far from finishing up with that.

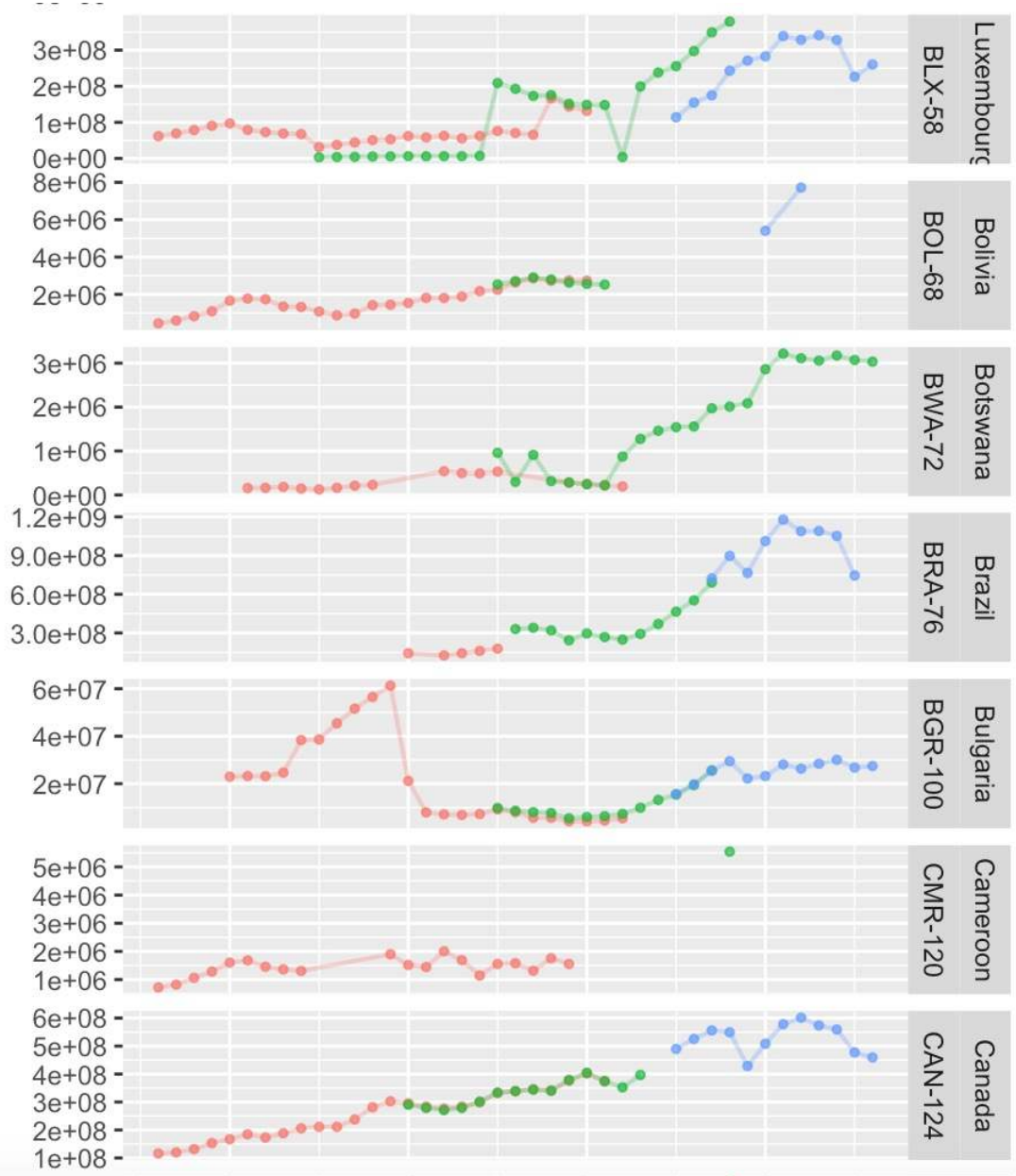
Hopefully you haven't spent too much time looking at the data yet. I'm very sorry about this.



Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

…and fixed…





Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

~~Lessons~~
Questions for next time?

Is there a better way to harmonise data?

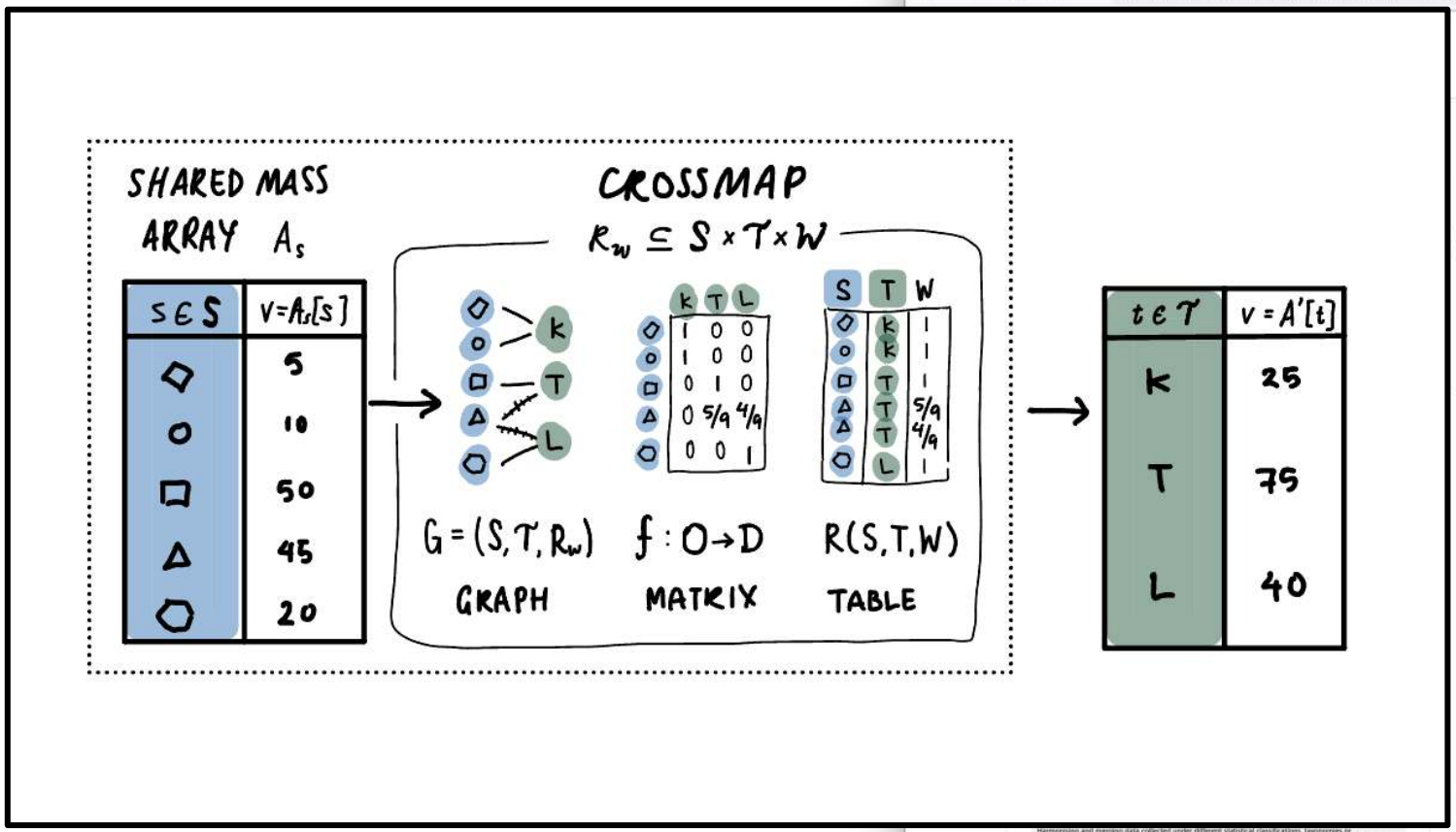
Why don’t we try different mappings?

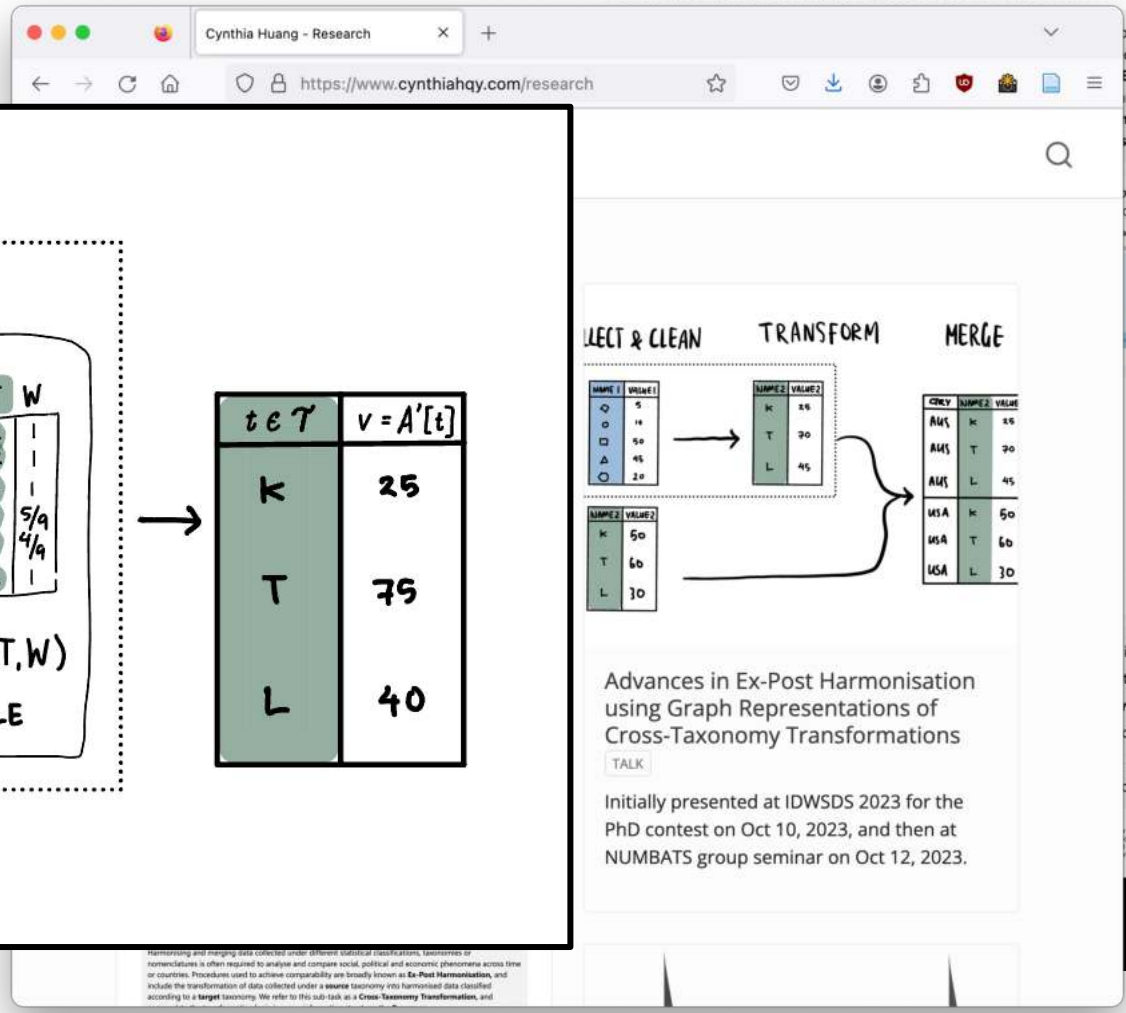
What’s actually happening to data through multiple concordance steps?

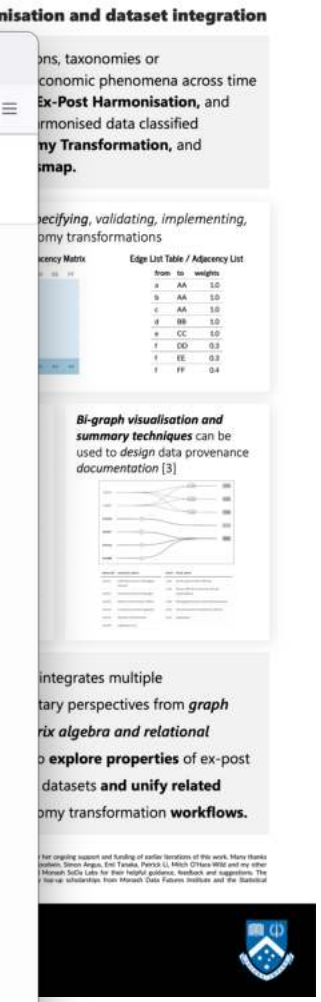
What if we didn’t have ground truth to compare with?

Answers… in my thesis?

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







Collect | Import & Prepare | Harmonise | Validate | ERROR! | Fix | Reboot | Happily Research After…?

Thanks for listening!

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

Learn more: cynthiahqy.com/research