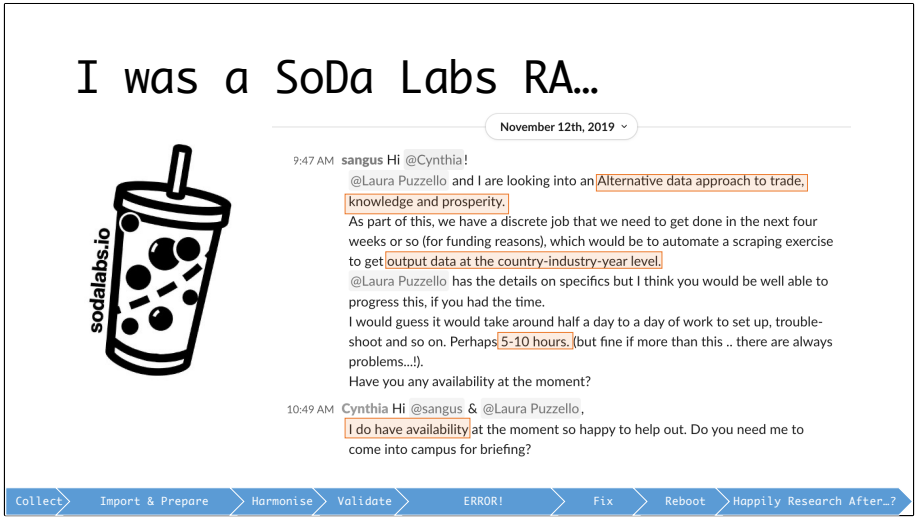
Silver Linings:
How a Data Harmonisation Mishap Turned into a PhD Thesis

Cynthia Huang

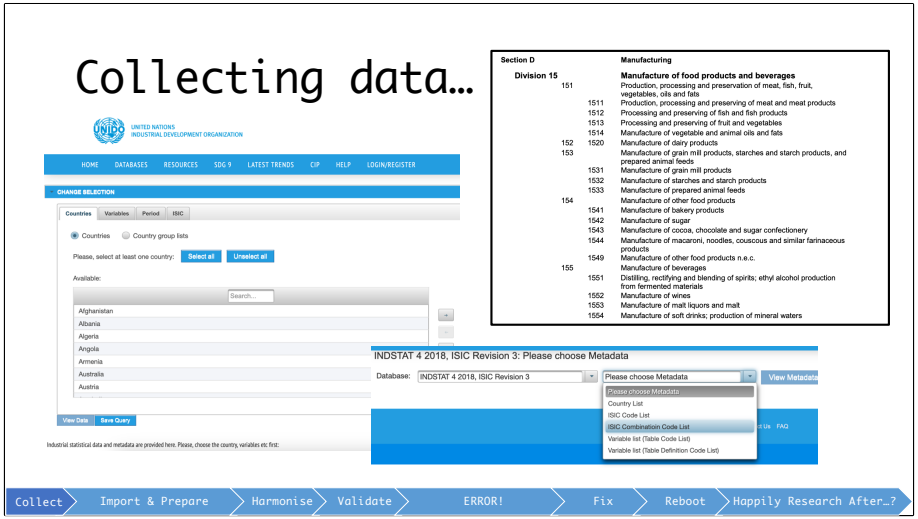
PhD Candidate at Monash University, Australia

Once upon a time…

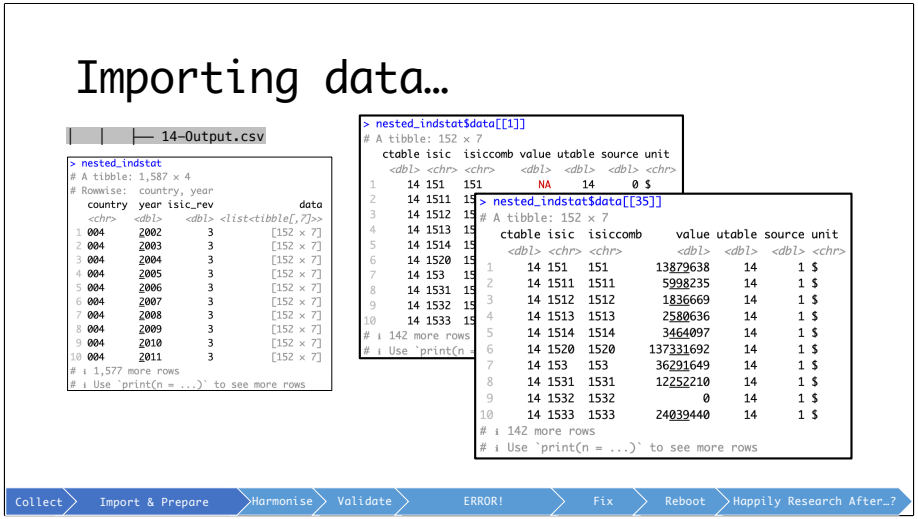
My story starts just like any other data story…



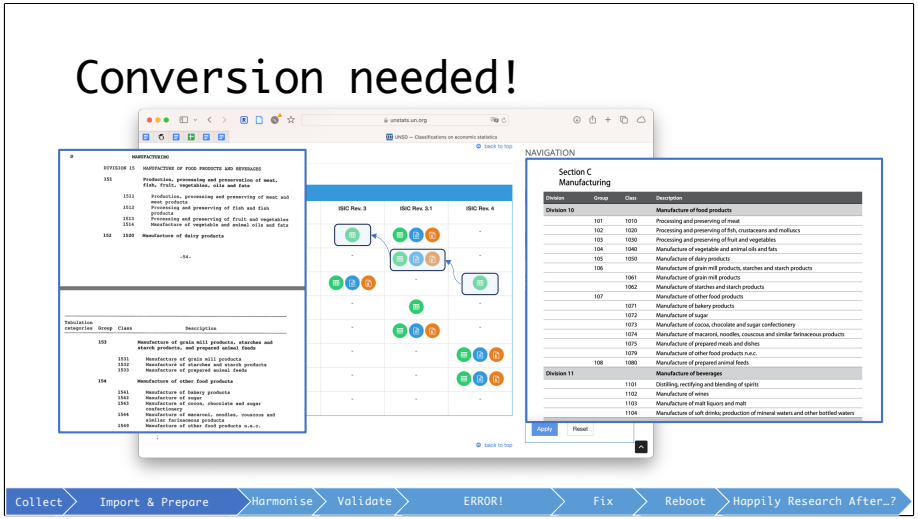
I was an economist at SoDa Labs Monash University, and I received this slack message, asking me if I could help prepare some data for an alternative data approach to researching Trade, Knowedge and Prosperity… It was only 5-10hours of work (famous last words), so I said yes of course I could help out, and thus my data mishap story began like any good data story…



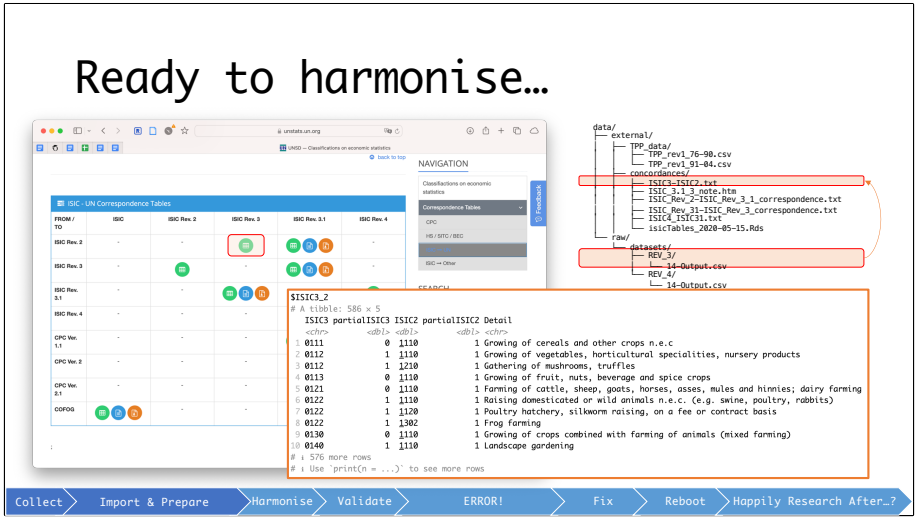
With the pain of collecting data… but that’s not the point of this story, so let’s just snap our fingers and skip straight to having



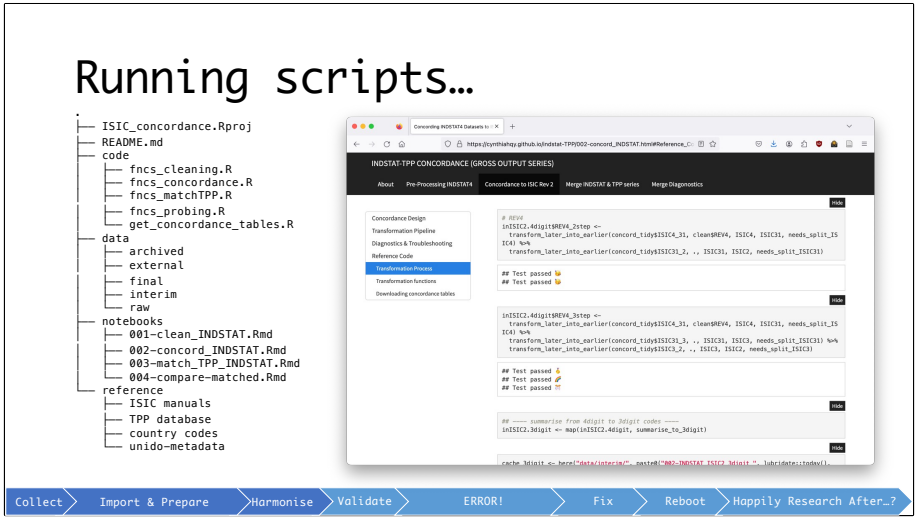
country, year output statistics, broken down by industry…



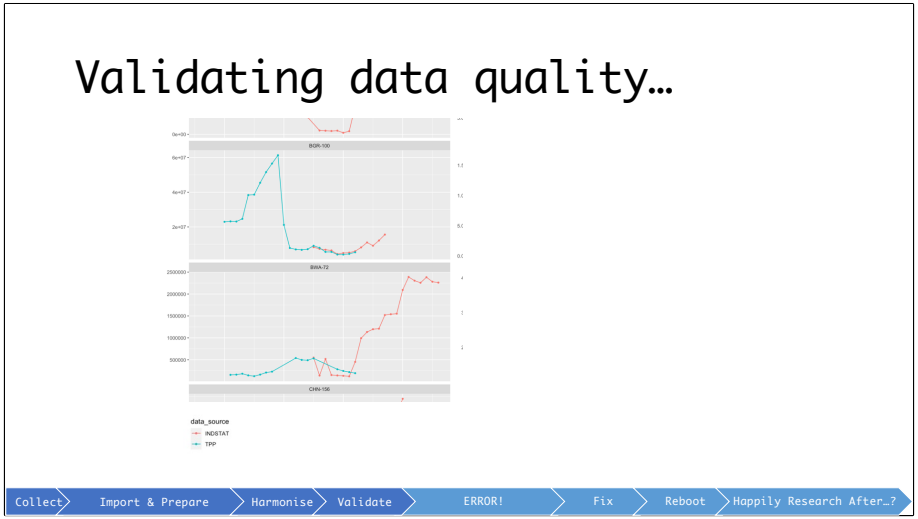
Unfortunately, the data couldn't be used straight away because the industry classification codes changed over the years and some countries adopt the revisions faster than others, while others decide to use variations on the international standard. The upshot is that if you want to analyse trends over time, you need to resolve these classification changes. Luckily for us, UN stats kindly provides selected correspondence tables between revisions…



… so in theory I had everything I needed – the raw data, logic for mapping between ISIC revisions… all I needed to do was *just* code it up.

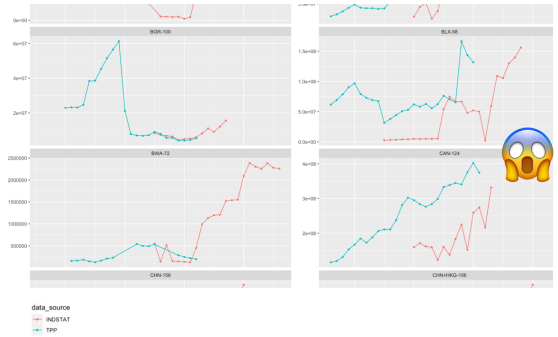


So that's what I did, I took the correspondence tables and turned them into R functions and scripts. Not only did I do that, but I also meticulously documented what I did, kept my project folder neat and tidy, and even tested things like row numbers and totals as I went along!



I'm way past 5-10 hours at this point, so I'm ready to send the data off to be analysed and turned into some interesting economics research… I just needed to quickly check the data quality… here I'm checking that the totals before and after recoding the industry codes line up…

Issue found…



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

Error located…

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

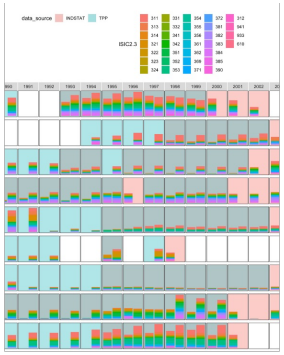
Hi Laura,

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.

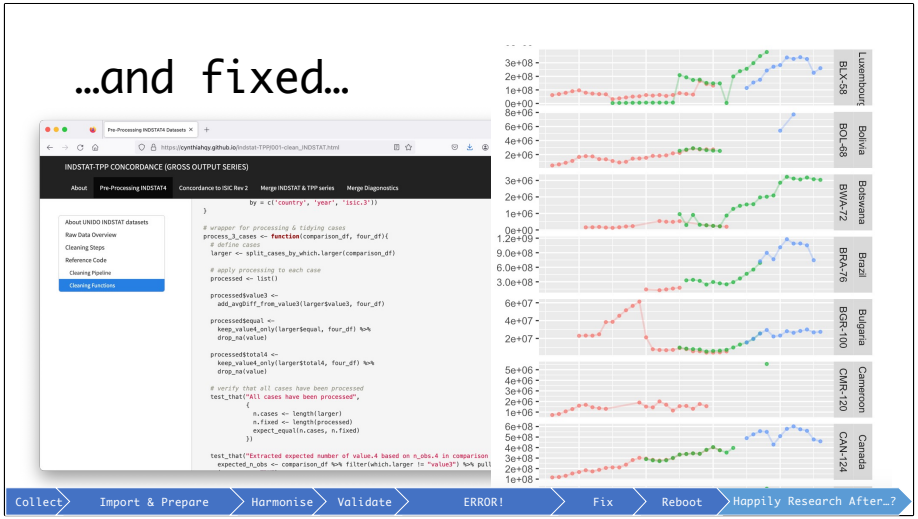
The missing data is about 31,000 rows (163,782 - 132,191). This might explain the difference with the TPP data.

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…?

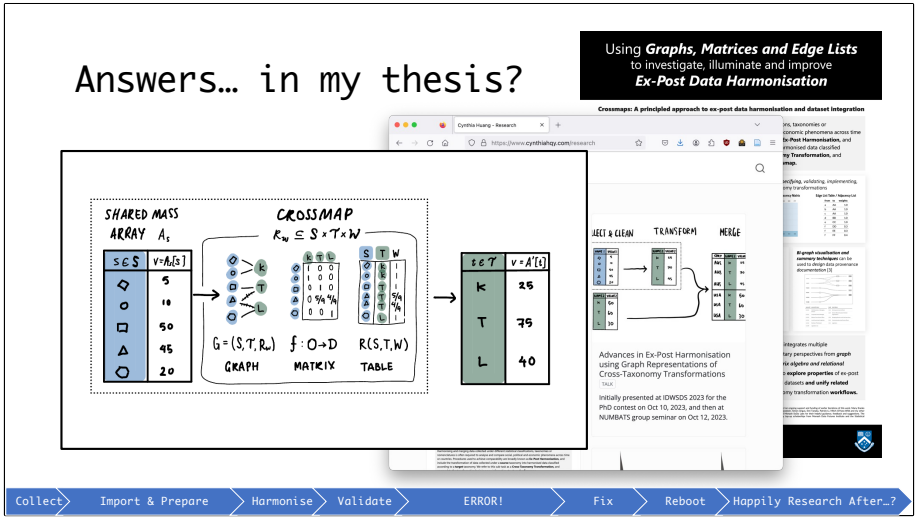


I fixed it… mostly, to this day I still can't tell you exactly what causes the remaining mismatch. But everyone on the team was satisfied that the data was good enough to analyse… so no harm no foul right?

~~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?

Well not quite, this project left me with a lot of questions…



Abstracting, formalizing and picking names for things…

Thanks for listening!

Talk to me: @cynthiahqy / cynthia.huang@monash.edu
Learn more: cynthiahqy.com/research