ETC5512: Wild Caught Data

Week 12

The proper care and feeding of wild data

Lecturer: *Dianne Cook*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Image source: https://flickr.com/photos/34534185@N00/6081362690, via https://commons.wikime

Time has come to wrap up this unit

Suppose you are the data curator. What should you know.

- Organising data into spreadsheets for analysis
- Rules for caring and feeding your data
- Realistic guide to making data available

Open data is...

a raw material for the digital age but,

unlike coal, timber or diamonds,

it can be used by anyone and everyone at the same time.

https://www.europeandataportal.eu/elearning/en/module1/#/id/co-01

Example in the news

> *Today, three of the authors have retracted "Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis" Read the Retraction notice and statement from The Lancet https://t.co/pPNCJ3nO8n pic.twitter.com/pBoFBj6EXr*
>
> — *The Lancet (@TheLancet) June 4, 2020*


THE LANCET
Volume 395 · Number 10238 · Pages 1709-1758 · May 30-June 5, 2020
www.thelancet.com
The NEW ENGLAND JOURNAL of MEDICINE

An article in Lancet, one of the oldest and best known journals that publishes general medical articles, "found Covid-19 patients who received the malaria drug, hydroxychloroquine, were dying at higher rates and experiencing more heart-related complications than other virus patients". Within days, the World Health Organization had halted its support for trials of hydroxychloroquine. Australian infectious disease researchers began questioning the published results very quickly.

- An important point to note is *The data relied upon by researchers to draw their conclusions in the Lancet is not readily available in Australian clinical databases, leading many to ask where it came from*.
- This is not the norm for research articles today, where most journals require the data and software to be made available so that others can verify the results.
- The numbers for the Australian cases did not match the data that researchers here knew. So they made some phone calls.

> *Once I realised the data in That #LancetGate study was probably fabricated I couldn't do anything else and had to write a blog post about it. Not only is Surgisphere far too small to have software in 671 hospitals, their claimed awards are dodgy: https://t.co/Ro8vEvpZqc*
> *— Peter Ellis (@ellis2013nz) May 30, 2020*

> *Investigation from me in Melbourne and Stephanie Kirchgaessner in the US: Governments and WHO changed Covid-19 policy based on suspect data from tiny US company named Surgisphere: https://t.co/LtyG5UnldX*
> *— Melissa Davey (@MelissaLDavey) June 3, 2020*

The first to the National Notifiable Diseases Surveillance System, who confirmed that they were not the source of the data. Next to health departments in NSW and Victoria, who also confirmed that they did not provide the data. And then to the hospitals themselves, which provoked this response

Dr Allen Cheng, an epidemiologist and infectious disease doctor with Alfred Health in Melbourne, said the Australian hospitals involved in the study should be named. He said he had never heard of Surgisphere, and no one from his hospital, The Alfred, had provided Surgisphere with data. "Usually to submit to a database like Surgisphere you need ethics approval, and someone from the hospital will be involved in that process to get it to a database," he said. He said the dataset should be made public, or at least open to an independent statistical reviewer. If they got this wrong, what else could be wrong?" Cheng said.

> *New piece on the #Surgisphere saga from me: Unreliable data: how doubt snowballed over Covid-19 drug research that swept the world #opendata #openscience #hydroxychloroquine https://t.co/cI4VfcXeZy*
> *— Melissa Davey (@MelissaLDavey) June 4, 2020*

> *Retracted studies may have damaged public trust in science, top researchers fear https://t.co/hNsEM1hYnx*
> *— Melissa Davey (@MelissaLDavey) June 6, 2020*

Success story of open data

- Data related to the COVID-19 pandemic has been collated by many organisations across the globe and made freely available.

- These numbers led to suspicions about the article's claims.

Coronavirus COVID-19 daily update

Print Share

This Chief Health Officer update is intended to provide clinicians and the Victorian public with information about the number of confirmed cases of COVID-19 in Victoria as well as relevant public health response activities in Victoria. Chief Health Officer Alerts will continue to be issued when there are changes to the public health advice related to COVID-19.

08/06/2020

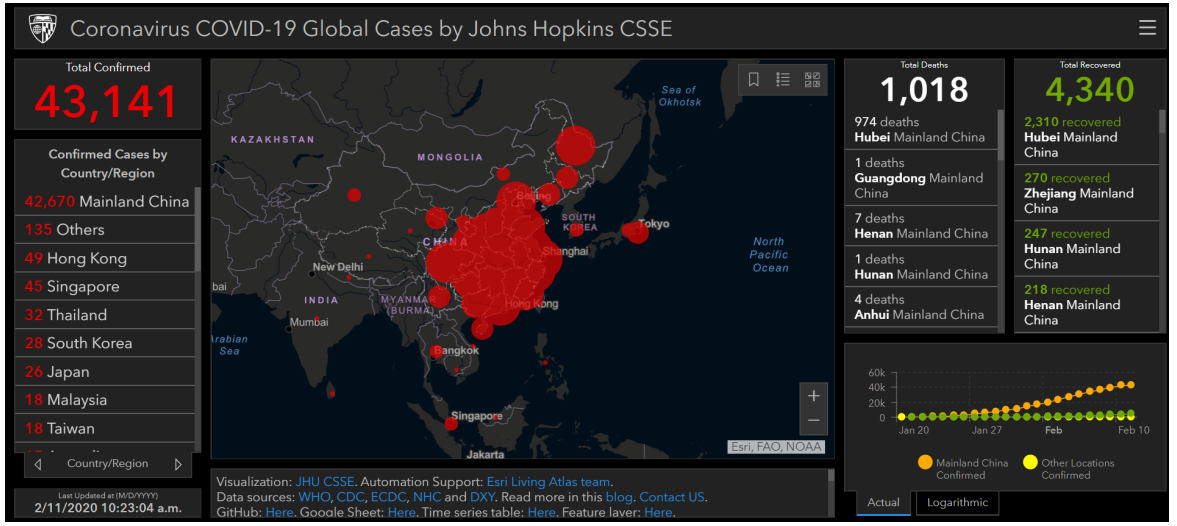
What's new?

Developments in the outbreak

- As of 8 June 2020, the total number of coronavirus (COVID-19) cases in Victoria is 1,687 which is an increase of two since yesterday.
- 177 cases have been acquired in Victoria where the source of infection is unknown, which is the same as yesterday.
- Seven people are in hospital, including two people in intensive care. 19 people have died. 159 people have recovered.
- Of the total 1,687 cases, there have been 1,403 in metropolitan Melbourne and 236 in regional Victoria. A number of cases remain under investigation.
- There have been 188 confirmed cases in healthcare workers, no change since yesterday's report.

Johns Hopkins COVID19

- COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
- Jan 23 (?) start of data collection
- I used this data for my own flexdashboard, started in mid-March, but it didn't have detailed data for Australia.
- Nick Evershed and group at Guardian
- Monash team



Vast number of people and organisations collating data, often (others) cross-checking numbers between sites.

Difficulties

- Changing formats!

> *... collated by Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) ... we will nevertheless scrape data from the relevant wikipedia pages, because it tends to be more detailed and better referenced than the equivalent JHU data ...*
> *Tim Churches blog Mar 1*

- Changing links! (The link to the GBR data from assignment 2 has changed)
- So many links on the website - which data to use?

Spreadsheets

Human consumption

	A	B	C	D	E	F	G	H	I
1	Qualifications by Year Level and Gender								
2					National				
3					Year 11		Year 12		Year 13
4	Qualificat	Gender							
5									
6	National Certificate of Educational Achievement								
7	NCEA (Level 1)								
8		Male			5,929		6,427		5,170
9		Female			0		60		38
10	NCEA (Level 2)								
11		Male			194		5,395		5,027
12		Female			0		58		38
13	NCEA (Level 3)								
14		Male			2		128		3,276
15		Female			0		0		36
16									

Computer consumption

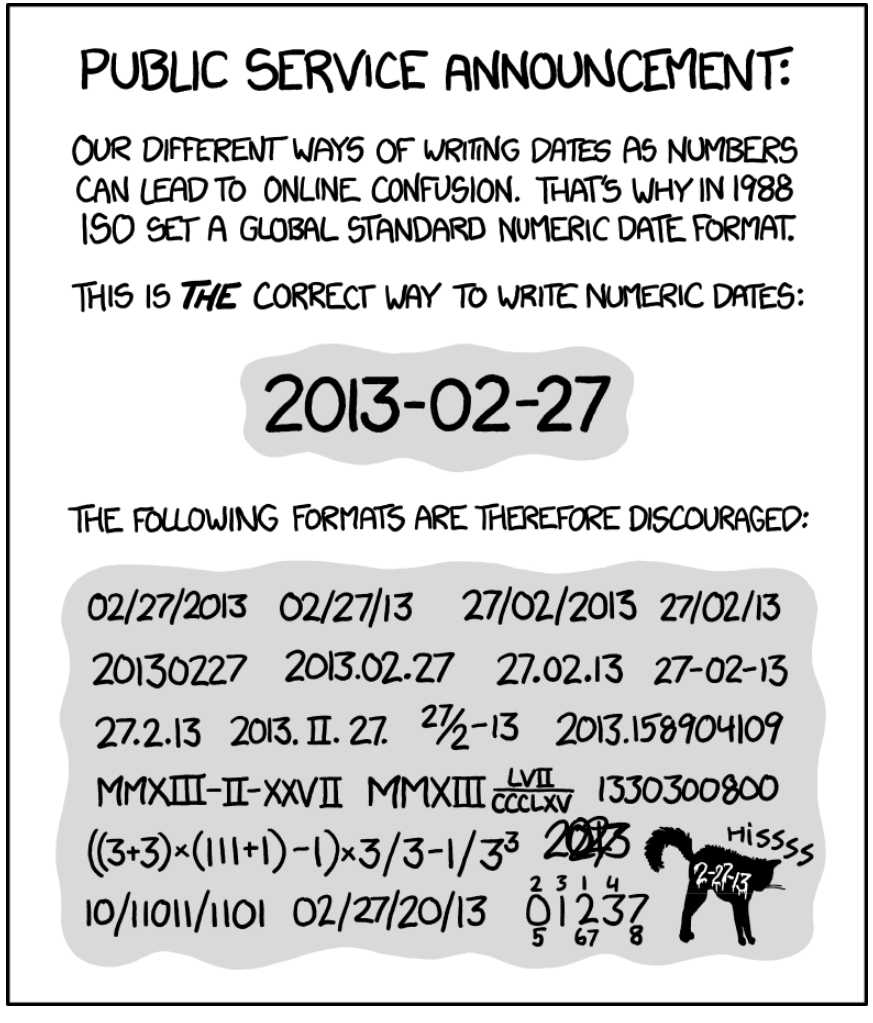
Gender	Qualification	Year	Value
Male	NCEA (Level 1)	Year 11	5929
Female	NCEA (Level 1)	Year 11	0
Male	NCEA (Level 2)	Year 11	194
Female	NCEA (Level 2)	Year 11	0
Male	NCEA (Level 3)	Year 11	2
Female	NCEA (Level 3)	Year 11	0
Male	NCEA (Level 1)	Year 12	6427
Female	NCEA (Level 1)	Year 12	60
Male	NCEA (Level 2)	Year 12	5395
Female	NCEA (Level 2)	Year 12	58
Male	NCEA (Level 3)	Year 12	128
Female	NCEA (Level 3)	Year 12	0
Male	NCEA (Level 1)	Year 13	5170
Female	NCEA (Level 1)	Year 13	38
Male	NCEA (Level 2)	Year 13	5027
Female	NCEA (Level 2)	Year 13	38
Male	NCEA (Level 3)	Year 13	3276
Female	NCEA (Level 3)	Year 13	36

Source: Murrell (2013) Data Intended for Human Consumption

Spreadsheets for computer consumption

- write dates like YYYY-MM-DD,
- do not leave any cells empty,
- put just one thing in a cell,
- organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row),
- create a data dictionary,
- do not include calculations in the raw data files,
- do not use font color or highlighting as data,
- choose good names for things,
- make backups,
- use data validation to avoid data entry errors, and
- save the data in plain text files.

Broman and Woo (2018) Data Organization in Spreadsheets https://doi.org/10.1080/00031305.2017



- Microsoft Excel's treatment of dates can cause problems in data
- It stores them internally as a number, with different conventions on Windows and Macs
- Excel also has a tendency to turn other things into dates.

The cells in your spreadsheet should each contain one piece of data. Do not put more than one thing in a cell.

You might have a column with "plate position" as "plate-well", it would be better to separate this into "plate" and "well" columns

- Remember, airlines data, time zone on one column, departure time in another. This is partly technical because multiple time zones can't be stored in a single column.
- Also, the data is distributed as Year, Month, Day columns, which is safer across systems

Create a data dictionary

Remember, the PISA data. Extensive data dictionary for each year distributed, giving variable names, and also explanation of levels in categorical variables.

Codebook_Stu06_Dec07.pdf (page 4 of 105)

```
ST05Q01 (12) Mother ISCO code Q5a
  Format: A4        Columns: 32-35
          1000      LEGISLATORS, SENIOR OFFICIALS & MANAGERS
          1100      LEGISLATORS & SENIOR OFFICIALS
          1110      LEGISLATORS [incl. Member of Parliament, Member of Local Council]
          1120      SENIOR GOVERNMENT OFFICIALS [incl. Minister, Ambassador]
          1130      SENIOR LOCAL GOVERNMENT OFFICIALS
          1140      SENIOR OFFICIALS SPECIAL-INTEREST ORGANISATIONS
          1141      Senior officials political-party organisations
          1142      Senior officials economic-interest organisations
          1143      Senior officials special-interest organisations
          1200      CORPORATE MANAGERS [LARGE ENTERPRISES]
          1210      [LARGE ENTERPRISES] DIRECTORS & CHIEF EXECUTIVES
          1220      [LARGE ENTERPRISE OPERATION] DEPARTMENT MANAGERS
          1221      Production dep. managers agriculture & fishing
          1222      Production dep. managers manufacturing [incl. Factory Manager]
          1223      Production dep. managers construction
          1224      Production dep. managers wholesale & retail trade
          1225      Production dep. managers restaurants & hotels
          1226      Production dep. managers transp., storage & communic. ]
          1227      Production dep. managers business services [incl. Bank Manager]
          1228      Production dep. managers personal care, cleaning etc
          1229      Production dep. managers nec [incl. Dean,School Principal]
          1230      [LARGE ENTERPRISES] OTHER DEPARTMENT MANAGERS
          1231      Finance & admin. department managers [incl. Company Secretary]
          1232      Personnel & industrial relations department managers
          1233      Sales & marketing department managers
          1234      Advertising & public relations department managers
          1235      Supply & distribution department managers
          1236      Computing services department managers
          1237      Research & development department managers
          1239      Other department managers nec
          1240      OFFICE MANAGERS [incl. Clerical Supervisor]
          1250      MILITARY OFFICERS
          1251      Higher military officers [Captain and above]
          1252      Lower grade commissioned officers [incl. Army Lieutenant]
          1300      [SMALL ENTERPRISE] GENERAL MANAGERS
          1310      [SMALL ENTERPRISE] GENERAL MANAGERS [incl. Businessman, Trader]
          1311      [Small enterprise] General managers agr., forestry & fishing
          1312      [Small enterprise] General managers manufacturing
          1313      [Small enterprise] General managers constr. [incl. Contractor]
          1314      [Small enterprise] General managers wholesale & retail trade
          1315      [Small enterprise] General managers restaurants & hotels
          1316      [Small enterprise] General managers transp., storage & comm.
          1317      [Small enterprise] General managers business services
```

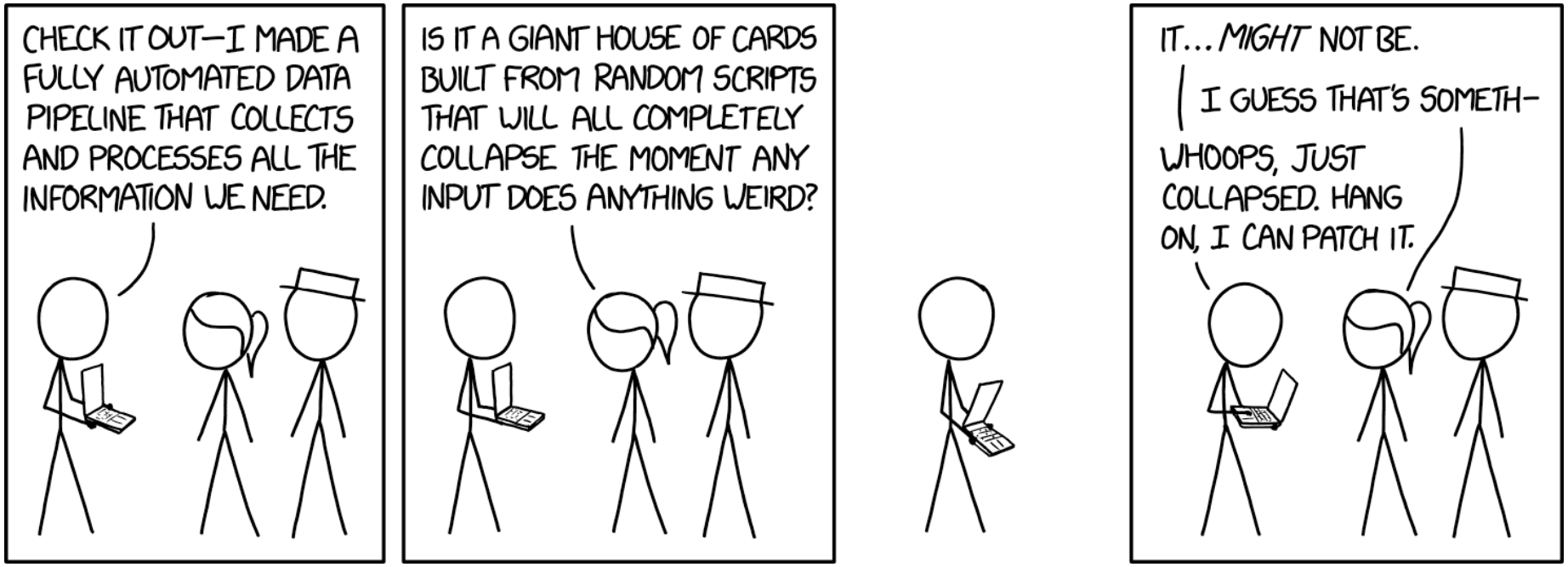

Beware your spreadsheets don't bite your data!

You can validate the integrity of your csv file with

http://csvlint.io

Goodman et al (2014) Ten Simple Rules for the Care and Feeding of Scientific Data

🤔 As we look at these rules, think about what this implies for business and government data.



Care and feeding

1. Love Your Data, and Help Others Love It, Too
2. Share Your Data Online, with a Permanent Identifier
3. Conduct Science with a Particular Level of Reuse in Mind
4. Publish Workflow as Context
5. Link Your Data to Your Publications as Often as Possible
6. Publish Your Code (Even the Small Bits)
7. State How You Want to Get Credit
8. Foster and Use Data Repositories
9. Reward Colleagues Who Share Their Data Properly
10. Be a Booster for Data Science

Love Your Data, and Help Others Love It, Too

What are some ways to show your love?

What data have we seen that isn't loved?

- Nurture:
  - feed,
  - hug, check on it
  - dress it nicely
  - give it a name
- Show it off:
  - tell someone about it
  - demonstrate how it can be used

Share Your Data Online, with a Permanent Identifier

- Give is a name: digital object identifier (DOI)
- Adequate documentation and metadata
- Employing good curation practices

Common resources:

- Zenodo
- FigShare
- Dataverse
- Dryad

Conduct Science with a Particular Level of Reuse in Mind

Replace "science" with "data science", "data analysis", "analytics", "business intelligence".

- keep careful track of versions of data and code
- to be fully reproducible, then *provenance* information is a must
  - working pipeline analysis code,
  - a platform to run it on, and
  - verifiable versions of the data.
- what types of re-use do you think others might make of your work?

Reward Colleagues Who Share Their Data Properly

- Build promotion and award systems that count data and code-sharing activities.
- Consider this activity an important part of your own data science work.
- Clear guidelines for credit

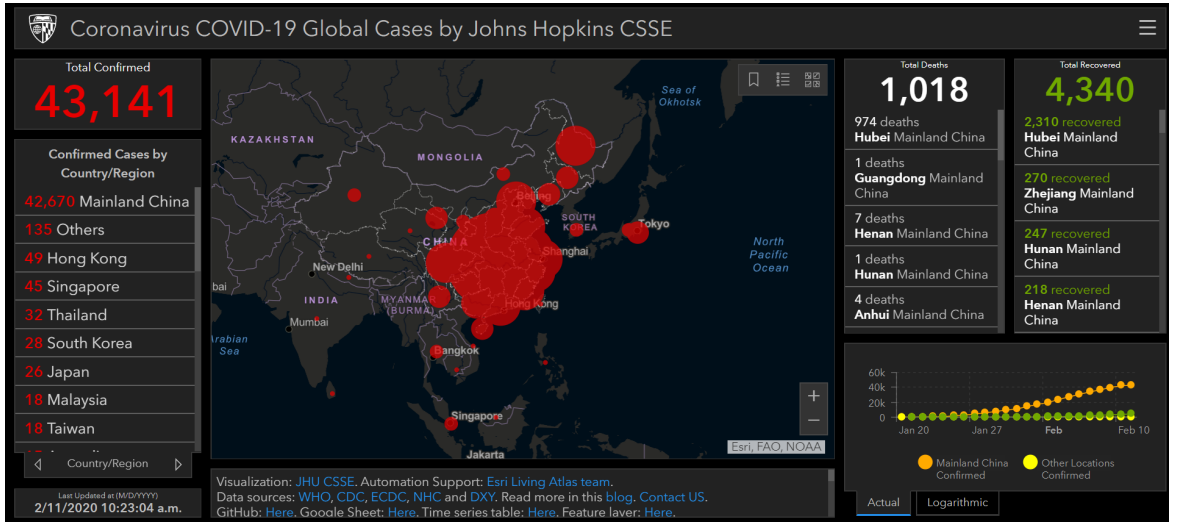
Lorraine and her husband Rick celebrating the Society's conferral of the Serventy Conservation Award in March.

Source: https://www.aws.org.au/serventy/

Johns Hopkins COVID19

What's really nice 😄

- Github page
- Compiled data from various sources, sources listed
- Update time stamp
- Versioning
- Issues for two way conversations with users



Macroeconomic data

Survey of Professional Forecasters (Assignment 1)

- 🐾 Need to know what you are looking for, many links, and several clicks deep ❌
- 🐾 Regularly updated, time stamp ✅
- 🐾 Web interface ✅
- 🐾 API for other software, like ALFRED package, to extract subsets ✅
- 🐾 `csv` file is nicely rectangular ✅

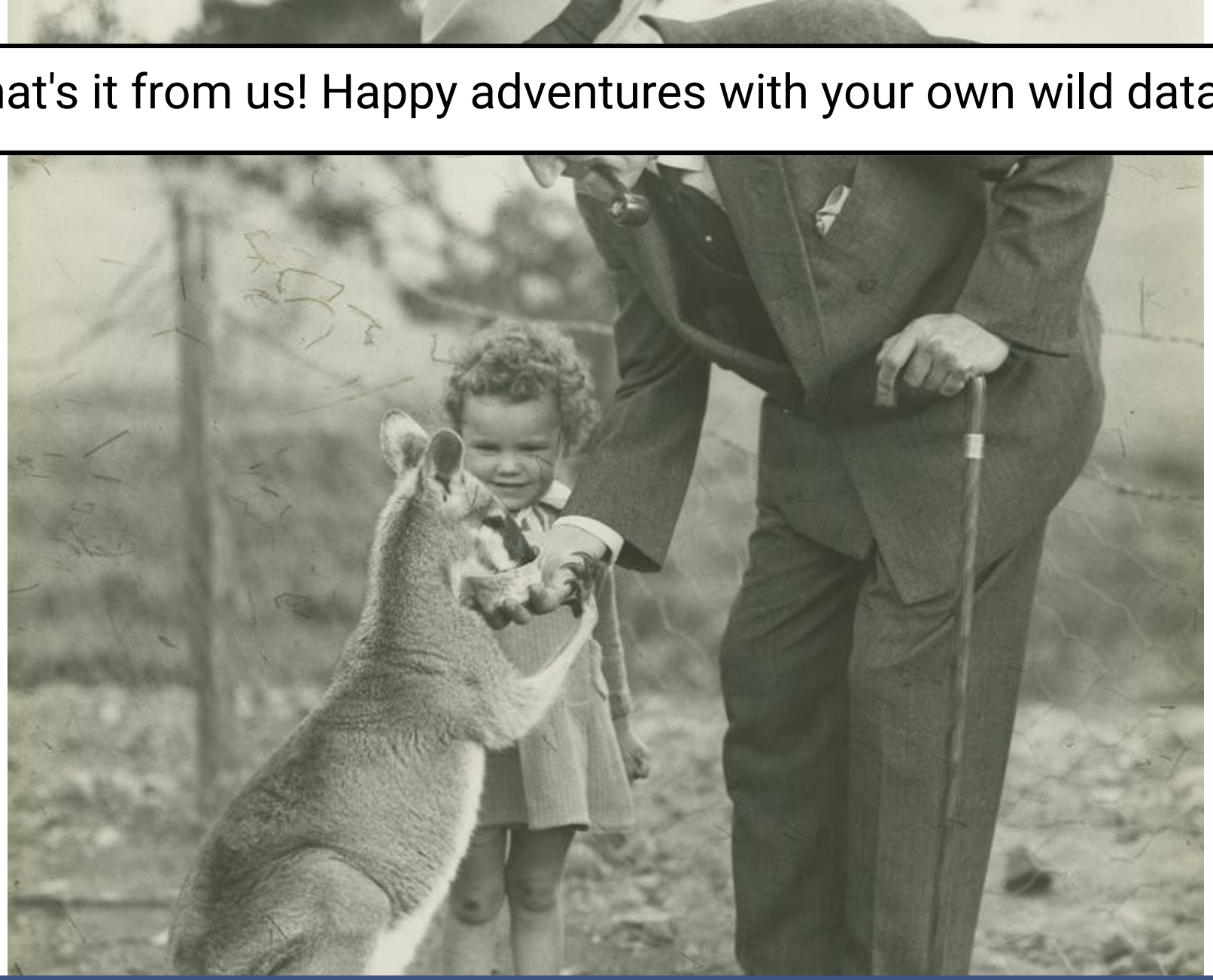
ABS Census Data

- updated regularly, for each census ✅
- data packs, easy to find ✅
- download has regular file structure ✅
- finding variable of interest is hard, though ❌
- spreadsheet with a gazillion tables, and variables are coded into column headers ❌

OECD PISA

- 🐾 nice web interface, now with simple queries and interactive plots ✅
- 🐾 updated regularly ✅
- 🐾 extensive documentation on data collection - very technical ✅
- 🐾 data dictionary, extensive! ✅
- 🐾 data from each available in various formats, with code to read it ✅
- 🐾 format for each year is different, variables collected differ (see learningtower) ❌

That's it from us! Happy adventures with your own wild data!



Grandpa feeding little Beverley Purd's pet kangaroo, 1930, State Library of Queensland