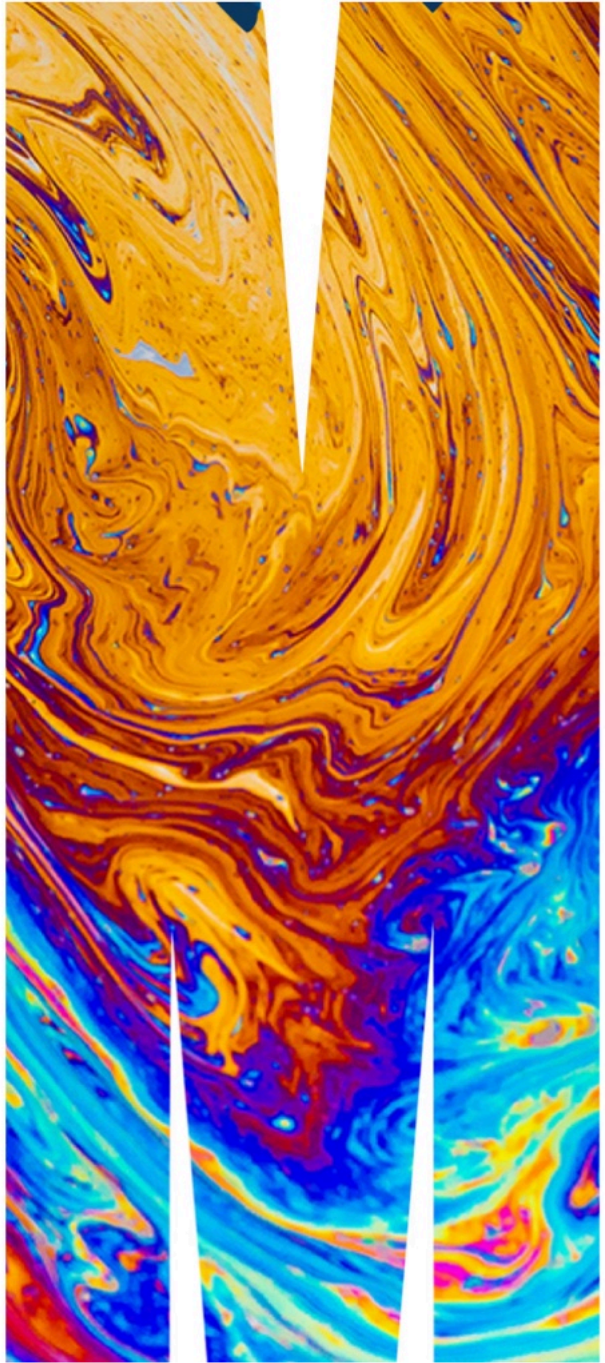
ETC5512: Wild Caught Data

The proper care and feeding of wild data



Lecturer: *Kate Saunders*

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu

📅 Week 12

Time has come to wrap up this unit

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

Today you will learn:

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

We will also discuss SETUs and Assignment 4 today.

Back in week 1 ...

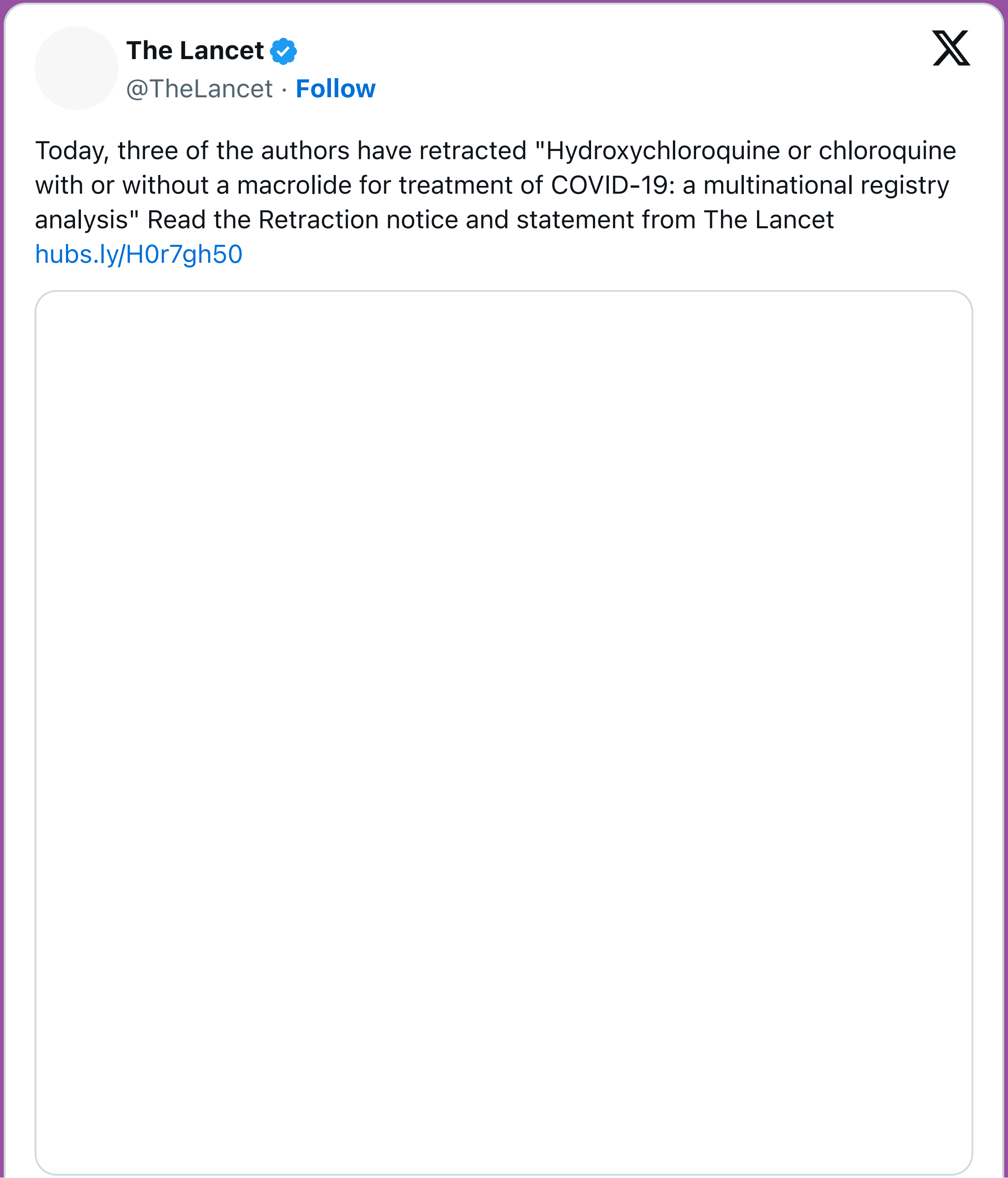
We learnt **OPEN DATA** is a raw material for the digital age but,

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

Let's remind ourselves with an example why open data is important!



The Lancet
@TheLancet · Follow

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 hubs.ly/H0r7gh50

THE LANCET
Volume 395 · Number 10238 · Page[illegible]-1738 · May 30-June 5, 2020
www.thelancet.com
The NEW ENGLAND JOURNAL of MEDICINE

Results published in The Lancet

-An article in The Lancet, "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".

-The Lancet is *one of the oldest and best known journals* that publishes general medical articles,

-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.

Something fishy

-The data the researchers used to draw their conclusions in the Lancet article is not readily available in Australian clinical databases

-This lead many to ask where the data came from?

-The numbers for the Australian cases did not match the data that researchers here knew.

-Most journals require the data and software to be made available so that others can verify the results. This is becoming more and more the standard.

So the Australian infectious disease specialists made some phone calls ...

Hello, can I ask you about your data?

The first call was 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 prompted 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.

Not found

> 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

Not found

> 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

Not found

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

Not found

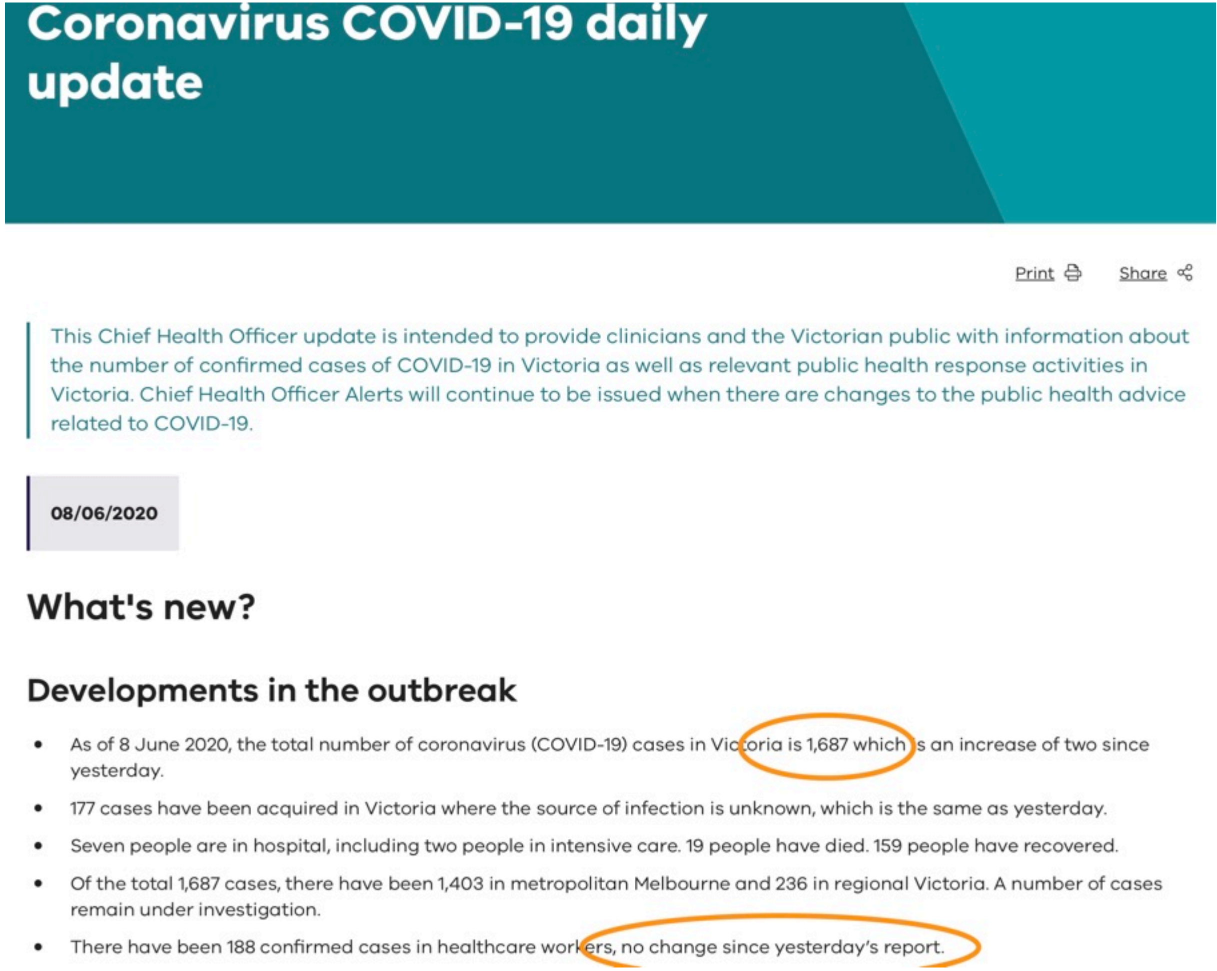
Retracted studies may have damaged public trust in science, top researchers fear https://t.co/hNsEM1hYnx
— Melissa Davey (@MelissaLDavey) June 6, 2020

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

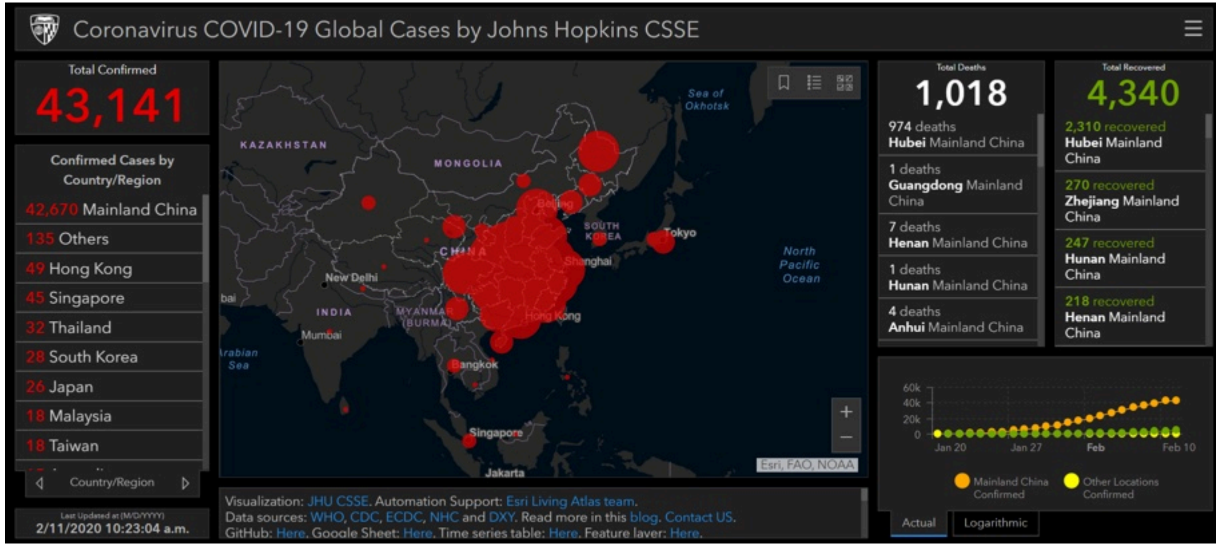
What's new?

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
- COVID Live
- NYTimes
- Monash team



Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE

Total Confirmed
43,141

Confirmed Cases by Country/Region

42,670	Mainland China
135	Others
49	Hong Kong
45	Singapore
32	Thailand
28	South Korea
26	Japan
18	Malaysia
18	Taiwan

Country/Region

Last Updated at (M/D/YYYY)
2/11/2020 10:23:04 a.m.

Total Deaths
1,018

974 deaths Hubei Mainland China
1 deaths Guangdong Mainland China
7 deaths Henan Mainland China
1 deaths Hunan Mainland China
4 deaths Anhui Mainland China

Total Recovered
4,340

2,310 recovered Hubei Mainland China
270 recovered Zhejiang Mainland China
247 recovered Hunan Mainland China
218 recovered Henan Mainland China

Visualization: JHU CSSE. Automation Support: Esri Living Atlas team.
Data sources: WHO, CDC, ECDC, NHC and DXY. Read more in this blog. Contact US.
GitHub: Here. Google Sheet: Here. Time series table: Here. Feature layer: Here.

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

Difficulties

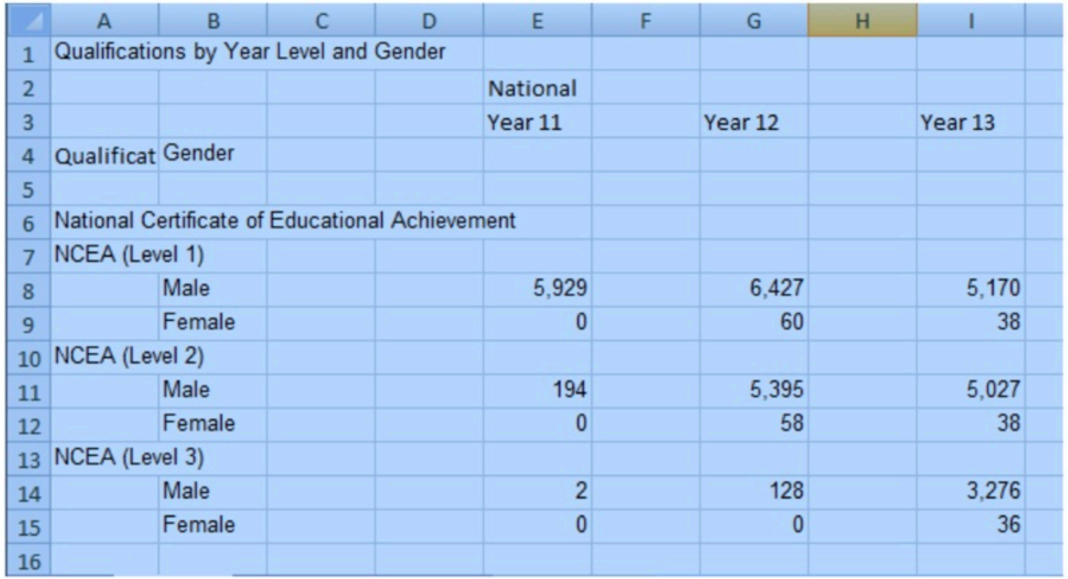
- Changing formats!

... 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!
- 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

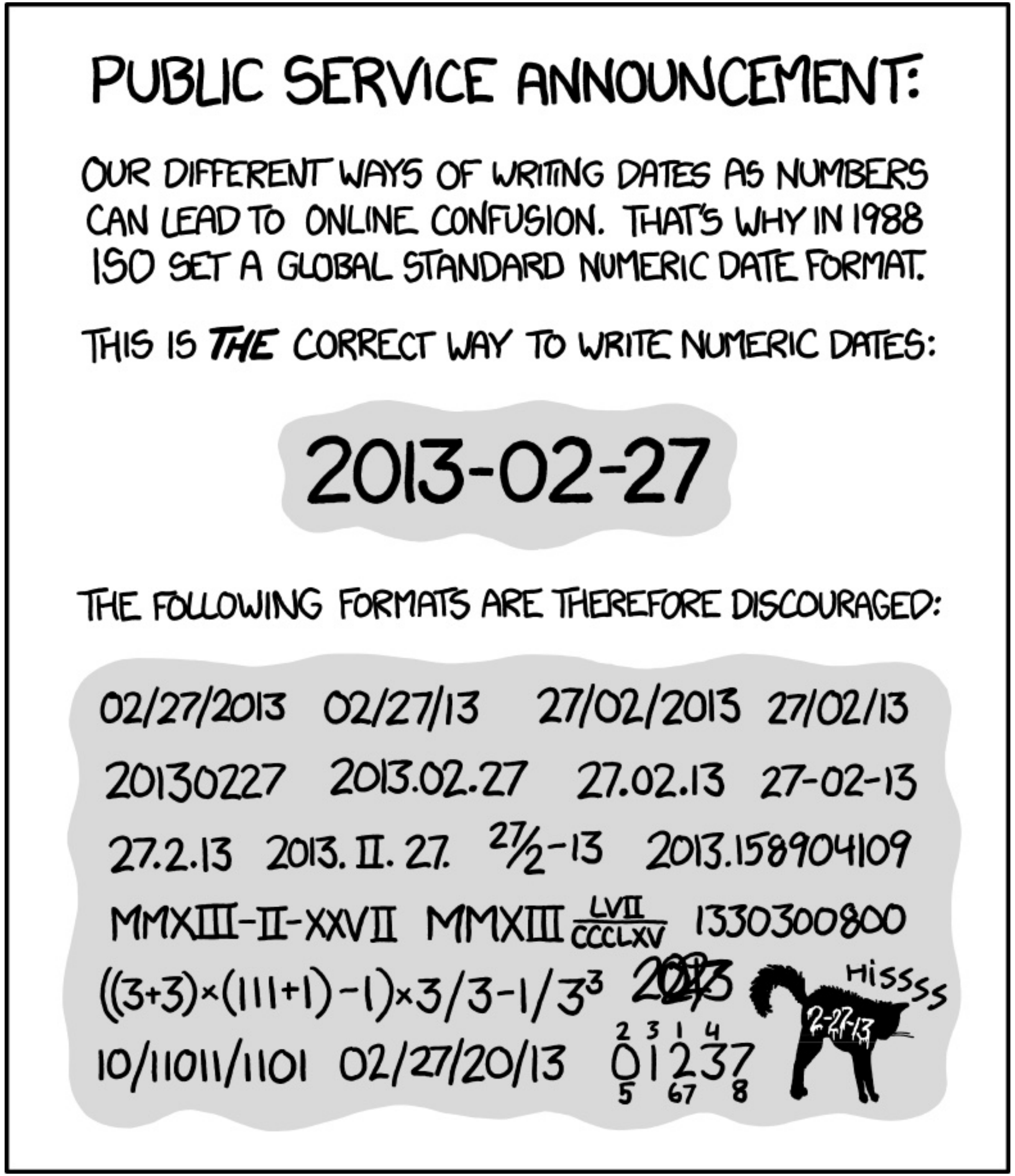
```
Gender    Qualification      Year   Value
  Male  NCEA (Level 1)    Year 11    5929
Female  NCEA (Level 1)    Year 11       0
  Male  NCEA (Level 2)    Year 11     194
Female  NCEA (Level 2)    Year 11       0
  Male  NCEA (Level 3)    Year 11       2
Female  NCEA (Level 3)    Year 11       0
  Male  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
```

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/10.1080/00031305.2017.1375989


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13 01237
2 3 1 4
5 67 8
HISSSSS
2-27-13

- 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.

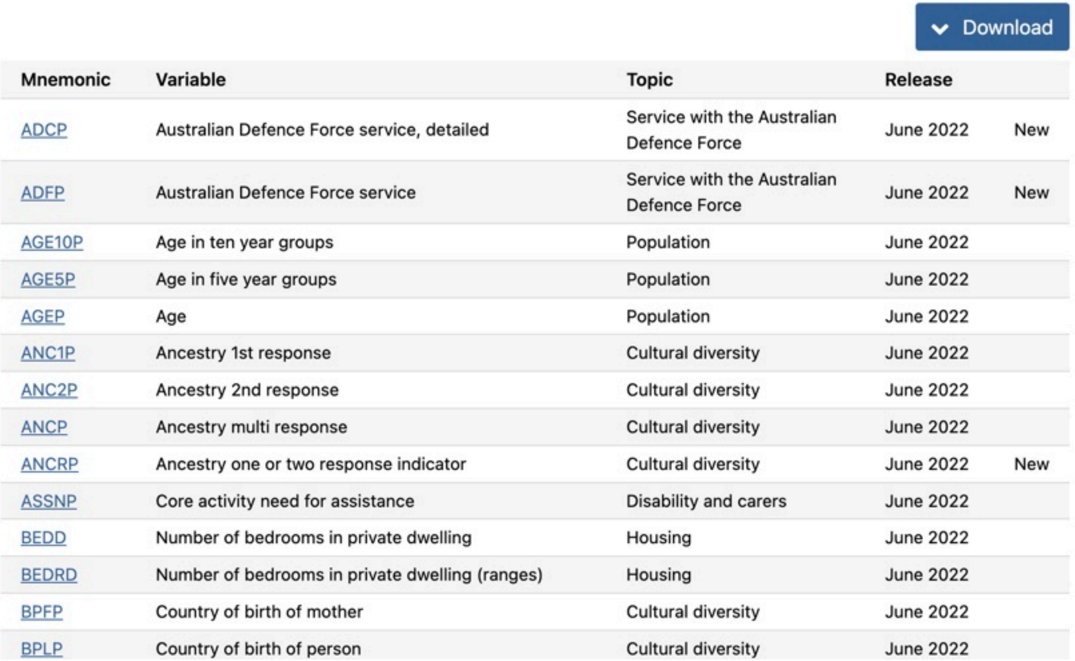
You might have a column with "plate position" as "plate-well", it would be better to separate this into "plate" and "well" columns.

- 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

The census has an extensive data dictionary for each year distributed, giving variable names, and also explanation of levels in categorical variables.

But, these are stored totally separately to where you access the census, making it all a bit more of a hassle than it needs to be.



Mnemonic	Variable	Topic	Release	
ADCP	Australian Defence Force service, detailed	Service with the Australian Defence Force	June 2022	New
ADFP	Australian Defence Force service	Service with the Australian Defence Force	June 2022	New
AGE10P	Age in ten year groups	Population	June 2022	
AGE5P	Age in five year groups	Population	June 2022	
AGEP	Age	Population	June 2022	
ANC1P	Ancestry 1st response	Cultural diversity	June 2022	
ANC2P	Ancestry 2nd response	Cultural diversity	June 2022	
ANCP	Ancestry multi response	Cultural diversity	June 2022	
ANCRP	Ancestry one or two response indicator	Cultural diversity	June 2022	New
ASSNP	Core activity need for assistance	Disability and carers	June 2022	
BEDD	Number of bedrooms in private dwelling	Housing	June 2022	
BEDRD	Number of bedrooms in private dwelling (ranges)	Housing	June 2022	
BPFP	Country of birth of mother	Cultural diversity	June 2022	
BPLP	Country of birth of person	Cultural diversity	June 2022	

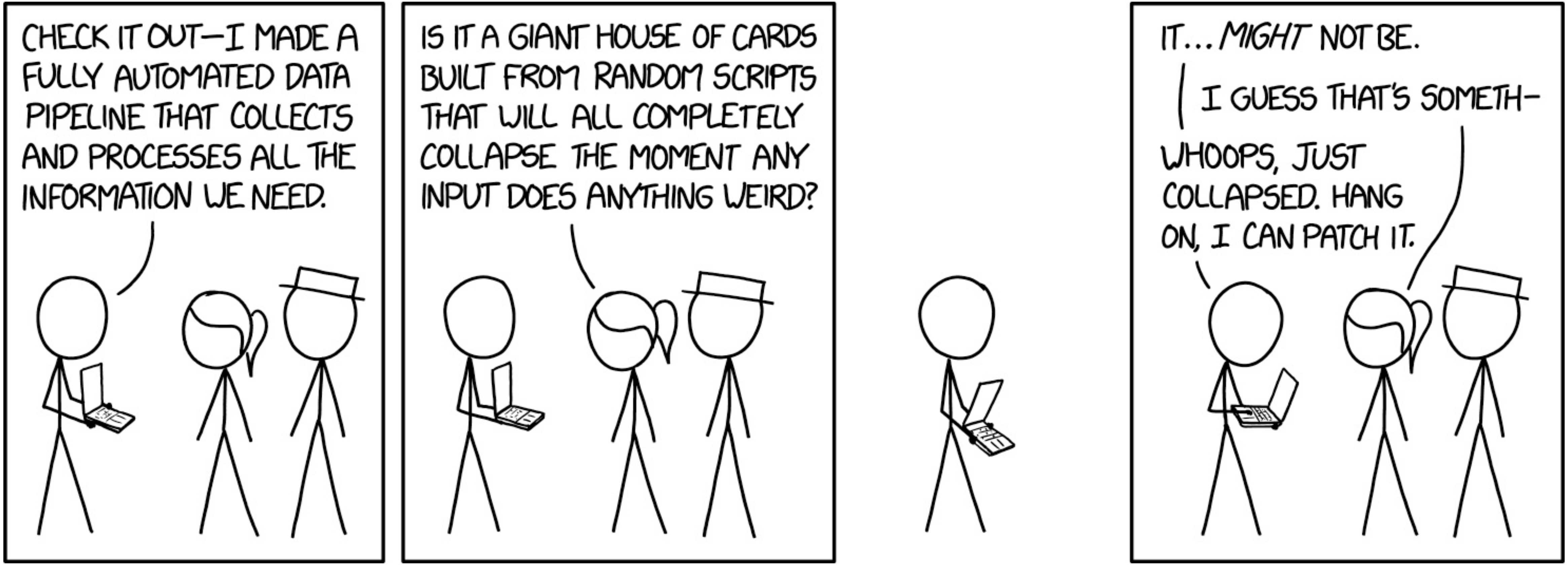
Beware your spreadsheets don't bite your data!

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.


CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... MIGHT NOT BE.
I GUESS THAT'S SOMETH—
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

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 it 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?

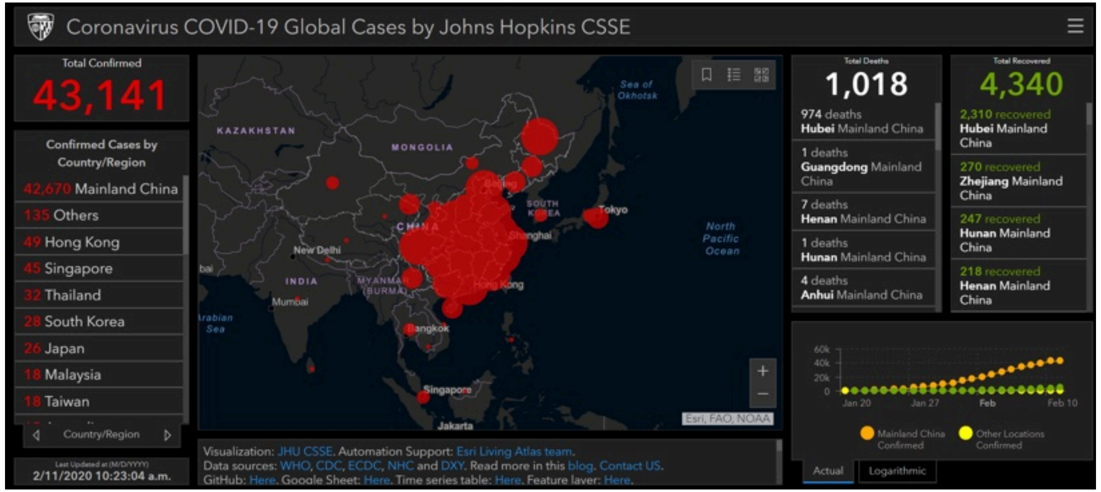
Reward Colleagues Who Share Their Data Properly

- 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

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



Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE

Total Confirmed
43,141

Confirmed Cases by Country/Region

42,670	Mainland China
135	Others
49	Hong Kong
45	Singapore
32	Thailand
28	South Korea
26	Japan
18	Malaysia
18	Taiwan

Last Updated at (M/D/YYYY)
2/11/2020 10:23:04 a.m.

Total Deaths
1,018

974 deaths Hubei Mainland China
1 deaths Guangdong Mainland China
7 deaths Henan Mainland China
1 deaths Hunan Mainland China
4 deaths Anhui Mainland China

Total Recovered
4,340

2,310 recovered Hubei Mainland China
270 recovered Zhejiang Mainland China
247 recovered Hunan Mainland China
218 recovered Henan Mainland China

Visualization: JHU CSSE. Automation Support: Esri Living Atlas team.
Data sources: WHO, CDC, ECDC, NHC and DXY. Read more in this blog. Contact US.
GitHub: Here. Google Sheet: Here. Time series table: Here. Feature layer: Here.

BTS air traffic

Bureau of Transportation Statistics (Assignment 1)

- Many, many different tables. The extent and value of the ontime performance database may not be immediately obvious. Need to know what you are looking for, many links, and several clicks deep ❌
- Sporadically missing chunks ❌
- No API for other software, laborious to download large chunks ❌
- Data provided by airlines is required, regular reporting is incentivised. Regularly updated, time stamp ✅
- Small chunk `csv` file is nicely rectangular ✅

Atlas of Living Australia

CSIRO (Assignment 1)

- Vast amount of data ✅
- Many different ways to access, including API ✅
- Hard to navigate the ways to access and what information is provided ❌
- Data stored is sporadic, on a volunteer basis ❌
- Data identifier (DOI) is provided with each download ✅

ABS Census Data

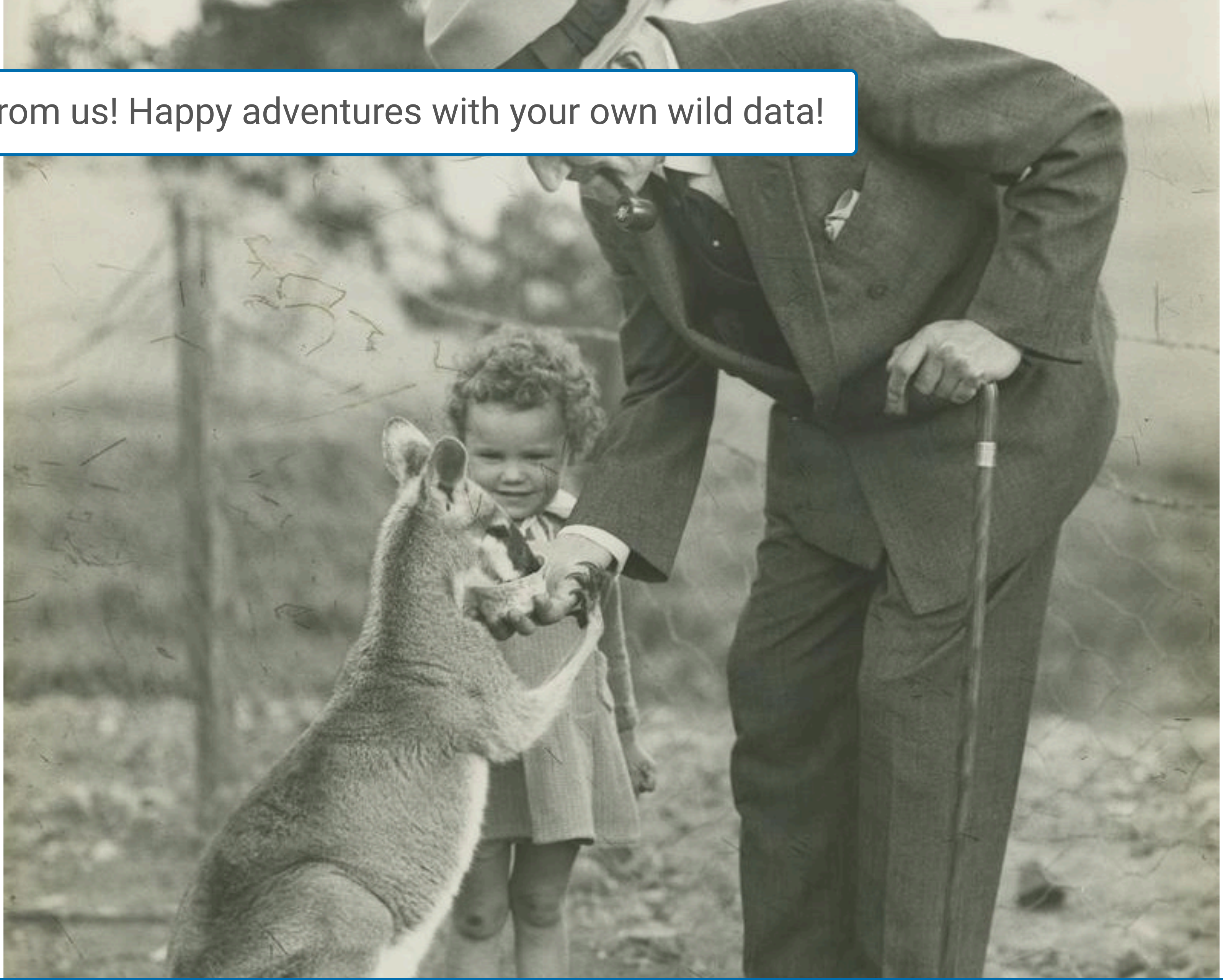
- Updated regularly, for each census ✅
- Data packs, easy to find ✅
- Download has regular file structure ✅
- Finding variable of interest is hard, though ❌
- Spreadsheet with a gazillion tables, and variables are coded into column headers ❌

Summary

Wrapping up this unit you should:

1. Understand the definitions, allowed usage, digital identification and licensing of open data
2. Know about common open data sources, how they are used and effectively search for new sources
3. Explain the differences between data collection methods and the limitations for data analysis
4. Work with the range of different data formats of open data
5. Understand ethical constraints and privacy limits when working with open data
6. Recognise the components of effective curation needed for open data

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



Prof. Di Cook's grandpa feeding little Beverley Purd's pet kangaroo, 1930, State Library of Queensland

Slides developed by Prof. Di Cook. Maintained and updated by Dr. Kate Saunders

MONASH University

Lecturer: *Kate Saunders*

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu

Week 12