

ETC5512: Wild Caught Data

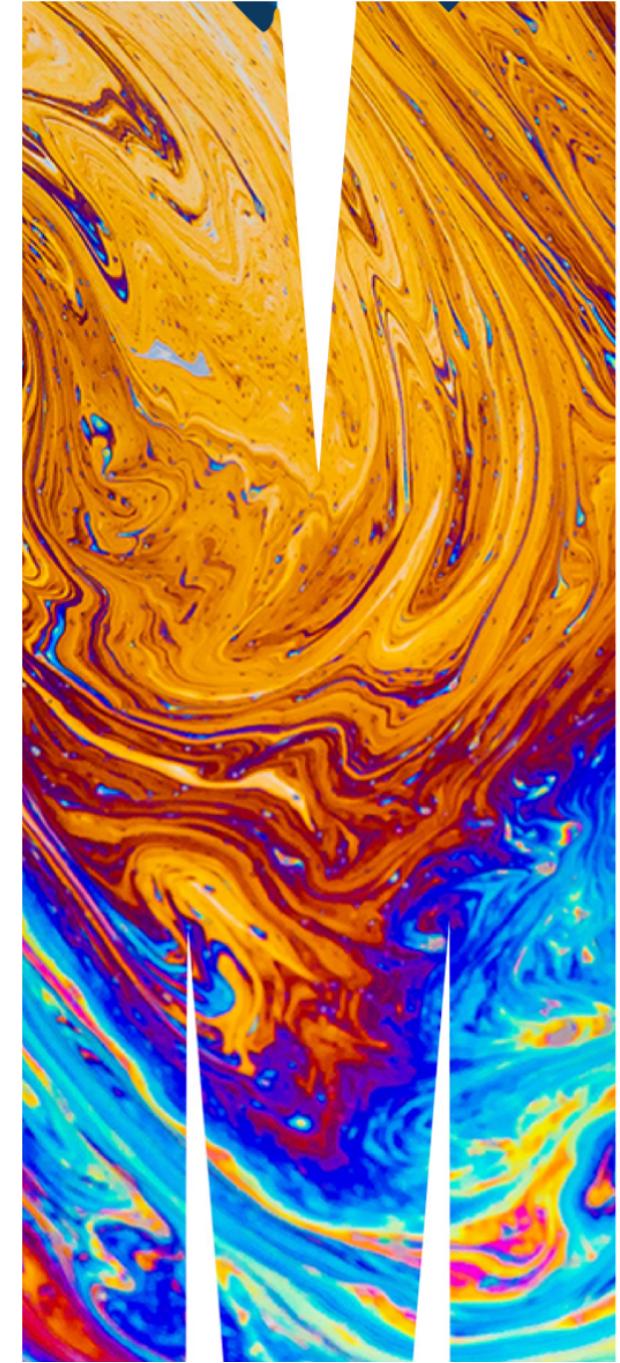
Open data: definitions, sources and examples

Lecturer: *Di Cook*

Department of Econometrics and Business Statistics

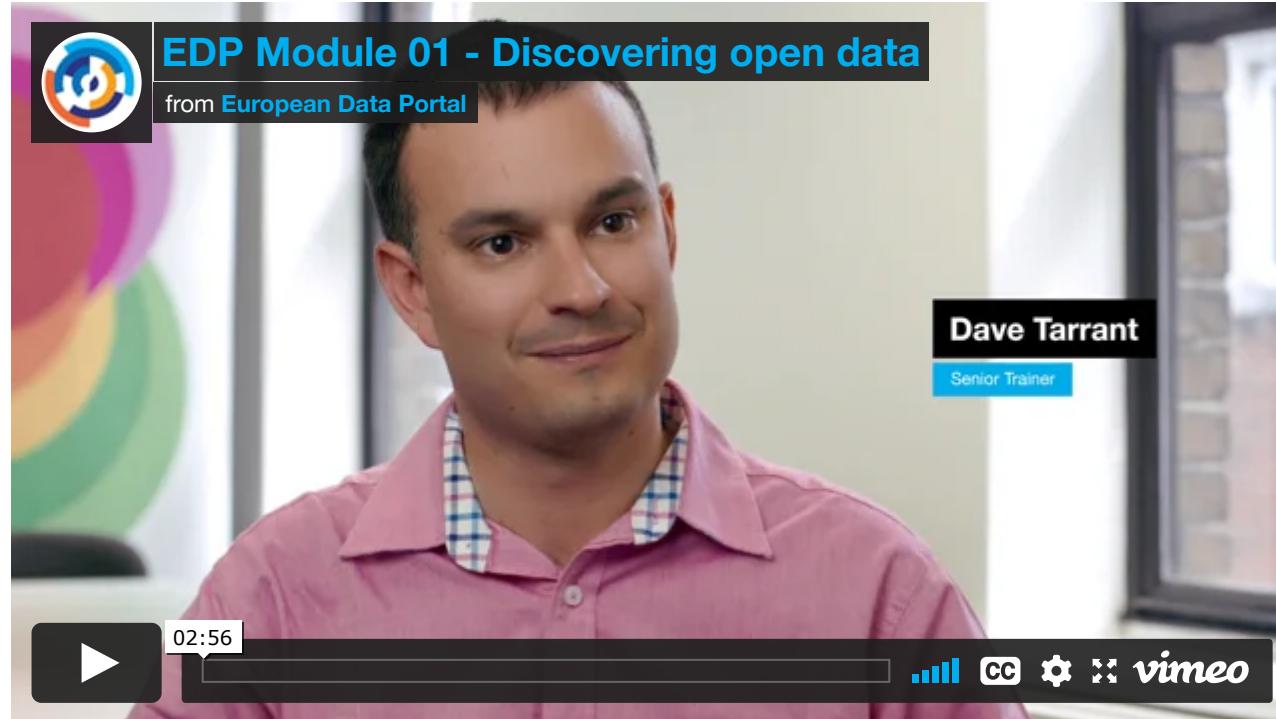
✉ ETC5512.Clayton-x@monash.edu

CALENDAR
Week 2



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.



What makes data open?

Limitations

-  No limitations that prevent particular uses.
-  Anyone free to use, modify, combine and share, even commercially.

Cost

-  Free to use does not mean that it must be free to access.
-  Cost to creating, maintaining and publishing usable data.
-  Live data and big data can incur ongoing costs.

Reuse

-  Free to use, reuse and redistribute it - even commercially.

Definition open data

Open data can be freely used, modified, and shared by anyone for any purpose

There are two dimensions of data openness:

-  The data must be legally open, which means they must be placed in the public domain or under liberal terms of use with minimal restrictions.
-  The data must be technically open, which means they must be published in electronic formats that are machine readable and non-proprietary, so that anyone can access and use the data using common, freely available software tools. Data must also be publicly available and accessible on a public server, without password or firewall restrictions.

<http://opendefinition.org/>

Try the quizzes [here](#)

Why do we need open data?

-  Help make governments more transparent.
-  Open data allowed citizens in Canada to save the government billions in fraudulent charitable donations
-  Building new business opportunities
 -  Transport for London has released open data that developers have used to build over 800 transport apps.
-  Protecting the planet
 -  Open data about weather can provide an early warning system for environmental disasters
 -  Open data is also helping consumers to understand their personal impacts on the environment

Open data from large organisations

 <http://dataportals.org/search>

 <http://data.un.org/>

 <https://datacatalog.worldbank.org/>

Open data Australia:

 <http://www.opendata500.com/au/>

 <https://opendataimpactmap.org/eap>

Government data

 <http://www.data.gov.au/>

 <https://www.data.vic.gov.au/>

 <https://data.melbourne.vic.gov.au/>

Why license open data?

-  Tells anyone that they can access, use and share data.
-  Without a licence, users may find themselves in a legal grey area. Data may be 'publicly available', but users may not have permission to access, use and share it under general copyright or database laws.
-  An open data licence is an explicit permission to use the data for both commercial and non-commercial purposes.
-  Open data publishers should provide easy access to the licence for all datasets that are available to access, use and share.

Open data licenses

-  Standard re-usable license: consistent and broadly recognized terms of use
-  Creative Commons, particularly CC-By and CC0 <https://creativecommons.org/>
-  Open Database License <https://opendatacommons.org/licenses/odbl/>
-  Bespoke licenses: governments and international organizations developed
 -  UK Open Government License <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>
 -  The World Bank Terms of Use <https://data.worldbank.org/summary-terms-of-use>

Try the quizzes [here](#)

Metadata: data about data

Information necessary to use the data appropriately:

-  Source
-  Structure
-  Underlying methodology
-  Topical
-  Geographic and/or temporal coverage
-  License
-  When it was last updated
-  How it is maintained

 Dublin Core Metadata Initiative (DCMI) provides a framework and core vocabulary of metadata terms.

 <https://www.dublincore.org/>

 Governments develop metadata models to provide further uniformity to government-wide Open Data initiatives.

 <https://project-open-data.cio.gov/v1.1/schema/>

 Australian government metadata standards

 [National Archives of Australia, Australian Institute of Health and Welfare](#)

Examples from Canadian government

 Resettled refugees

 Canada emergency wage subsidy
(CEWS)

 Title: what data contains and where it comes from.

 Description: details to quickly understand whether data is relevant to you

 publisher: dataset originated, who is responsible for maintaining, credibility

 license:

 contact information: questions or incomplete metadata

 frequency: interval data is updated. check for updates? data out of date?

 date modified: relevant for your work?

 spatial coverage: geographic area data is relevant

 temporal coverage:

Machine Readable



'machine readable' is not synonymous with 'digitally accessible'

- 📊 Historical efforts have focused on
 - 📊 pushing static information about government programs and services to the web,
 - 📊 where the intended use is a human who can read, print, and take actions based on reading.
 - 📊 It's a narrow vision of the expected users and uses of the information.
- 📊 Machine readable formats expand field of vision to new users and new uses and require technologies like XML and JSON
 - 📊 😢 PDF is not suitably machine readable
 - 📊 😊 CSV (or XLSX, XLS) is common, and universally accessible, but should be structured for analysis not for reading
 - 📊 😷 XML, JSON is verbose, can contain metadata, but needs special readers

Five star open data scheme

The web site 5 ★ Open Data at <https://5stardata.info/en/> reports a rating system for deploying open data.

-  ★ - [An open license](#): make your stuff available on the Web (whatever format) under an open license
-  ★★ - [Re-usable format](#): make it available as structured data (e.g., Excel proprietary instead of image scan of a table)
-  ★★★ - [Open format](#): make it available in a non-proprietary open format (e.g., CSV instead of Excel)
-  ★★★★ - [use \(Uniform Resource Identifiers \(URLs\)\)](#) to denote things, so that others can link to it, and also give context to the values
-  ★★★★★ - [Link data to definitions and context for various aspects](#)

FAIR principles for scientific data

Findable

Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Publishing data

Research data is increasingly seen as part of the corpus of scholarly publications. Publishers, funders and governments support researchers to publish their data outputs by various policies, guidelines and mandates.

-  Obtaining a Digital Object Identifier system (DOI) provides a persistent identifier, and can be used for data.
-  Many open data sets provide information on how to cite them, when used in other forms of publication.

Open data quality

Legal requirements:

-  Protect sensitive information like personal data
-  Preserve the rights of data owners
-  Promote correct use of the data

Practical requirements:

-  Link to the data from their website
-  Update the data regularly if it changes
-  Commit to continue to make the data available

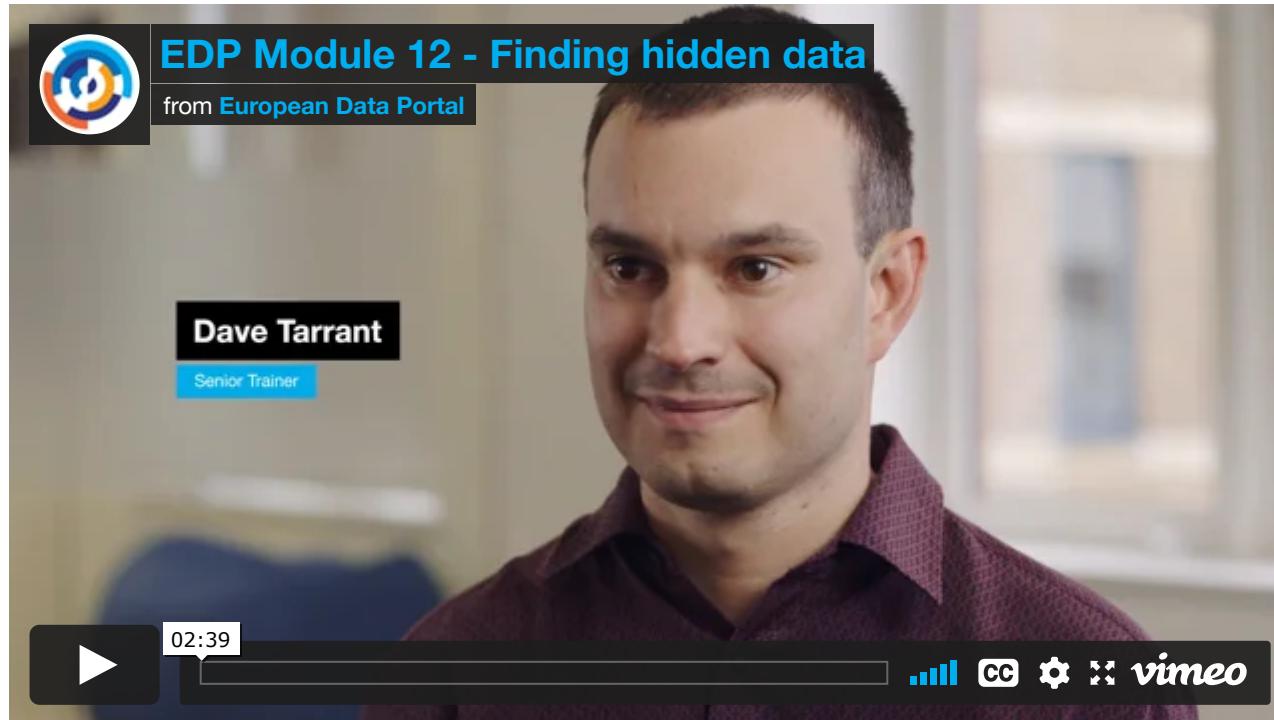
Technical requirements:

-  The format in which the data is published
-  The structure of the data
-  The channels through which the data is available

Common pitfalls with open data

-  Mixed date formats american/european
-  Multiple representations differences in abbreviations, capitalisation, spacing
-  Duplicate records
-  Redundant data
-  Mixed numerical scales
-  Spelling errors
-  Inconsistent naming
-  Missing values

What is hidden data?



Let's look at <https://www.realestate.com.au/buy>.

And do the quizzes [here](#)

Some of my favourite examples of open data

-  Airline traffic in the USA <https://www.bts.gov>
-  Australian Bureau of Statistics <http://stat.data.abs.gov.au>
-  Australian Electoral Commission <https://www.aec.gov.au>
-  National Longitudinal Survey of Youth (NLSY) <https://www.nlsinfo.org/investigator/pages/search?s=NLSY79>
-  Atlas of Living Australia <https://www.ala.org.au>
-  Australian bushfires from satellite hotspot remote sensing
https://www.eorc.jaxa.jp/ptree/registration_top.html (also see resulting analysis at
<https://ebsmonash.shinyapps.io/VICfire/>)
-  John Hopkins Coronavirus tracking <https://coronavirus.jhu.edu/data>
-  OECD Programme for International Student Assessment <http://www.oecd.org/pisa/data/>
-  Melbourne pedestrian counting system <http://www.pedestrian.melbourne.vic.gov.au/>

We'll spend some time here taking a look at these open data examples

Consider the interface

Look for licensing

Explanations of what's in the data

Metadata

Flavours of open data

How to tell if the open data is not so good to consume?

This is Prof Di Cook's taxonomy

Long shelf life, highly processed

- Convenient, but contains unhealthy ingredients, and is a bad habit
- eg iris, mtcars, titanic, handwritten digits
- Found at eg [UCI Machine learning archive](#)



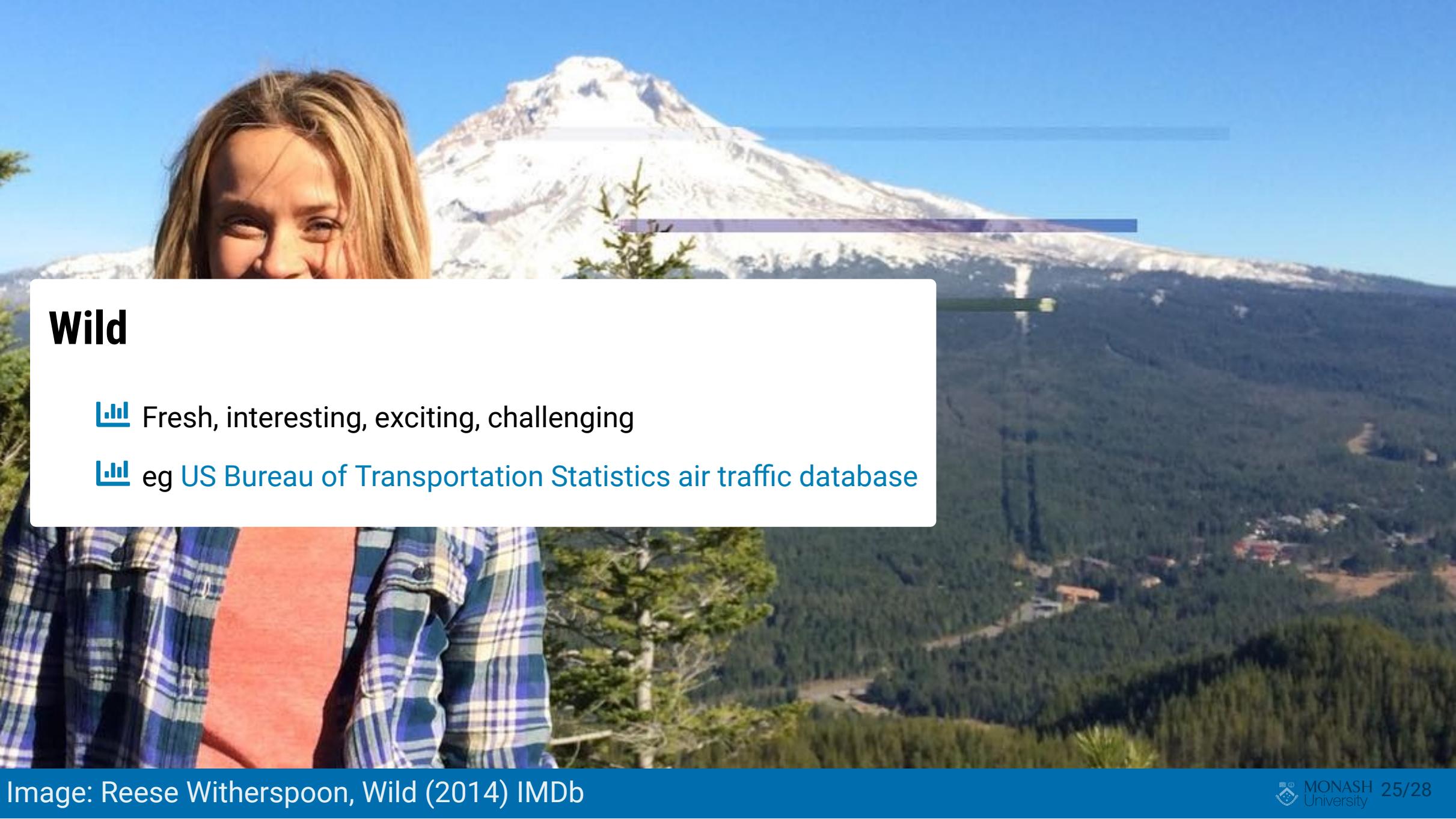
Orphans

- File dumped on an archive
- Stale, could date your results
- Found in places like <https://data.gov.au>



Synthetic

- Used primarily these days for privacy protection
- Correct up to the model used to simulate the data - misses interesting structure in data not captured by model
- Very pretty, very consistent, but it can burn you
- eg [OECD Programme for International Student Assessment](#) A generalised linear model is fitted to the scores, with predictors such as school, gender, ... Model is used to simulate a score for each student.
- eg Also be aware of fraud [Article in the Lancet \(2020\)31180-6/fulltext](#))

A photograph of a woman with blonde hair smiling at the camera. She is wearing a red top and a plaid jacket. In the background is a large, snow-capped mountain under a clear blue sky.

Wild

- 📊 Fresh, interesting, exciting, challenging
- 📊 eg [US Bureau of Transportation Statistics air traffic database](#)



Image: Reese Witherspoon, Wild (2014) IMDb



Fresh and local

- ─ Wild data, collected locally, and impacting our own lives
- ─ eg Melbourne pedestrian counts



Our working definition of wild-caught data will be:



Wild-caught data

The data can be freely used, modified, and shared by anyone for any purpose

AND

The data source is traceable, the data collection is transparent, and the data is updated as new measurements arrive. In case of data processing, the process is clearly described and reproducible.



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Lecturer: *Di Cook*

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu

CALENDAR Week 2

