

# ETC5512: Wild Caught Data

Week 2

## Open data sources

Lecturer: *Didier Nibbering*

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu



# **Examples of open data?**

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

# **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>

# Why open data useful?

-  Benefits to governments, businesses and individuals.
-  Improves services, grows economies and protects our planet.
-  Restrictions will limit its potential.

# What makes data open?

## Limitations

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

## Cost

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

## Reuse

- ➊ 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.

# Why license open data?

- 📊 Tells anyone that they can access, use and share data.
- 📊 Unless you have a licence, data may be 'publicly available', but users will not have permission to access, use and share it under copyright or database laws.

# 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>

# Open data Machine Readability

## Documents

- ➊ static and frozen in their format

## Data

- ➋ dynamic and can be open to further processing

<https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data>

# Open data Machine Readability



# Open data Machine Readability



# Metadata: data about data

Information necessary to use the data

 Source

 Structure

 Underlying methodology

 Topical

 Geographic and/or temporal coverage

 License

 When it was last updated

 How it is maintained

# Metadata: data about data

 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/>

# Metadata: example

 <https://open.canada.ca/data/en/dataset/4a1b260a-7ac4-4985-80a0-603bfe4aec11>

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

# Open data quality

Legal requirements:

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

# Open data quality

Practical requirements:

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

# Open data quality

Technical requirements:

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

# Five star open data scheme

-  1 star - An open license
-  2 stars - Re-usable format
-  3 stars - Open format
-  4 stars - use URIs
-  5 stars - Link data

<https://5stardata.info/en/>

# Cleaning data

Common pitfalls with data.

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

# Why do we need open data?

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

# Open data sources

 <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/>

# **Open data**

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

# Flavours of open data

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

# 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



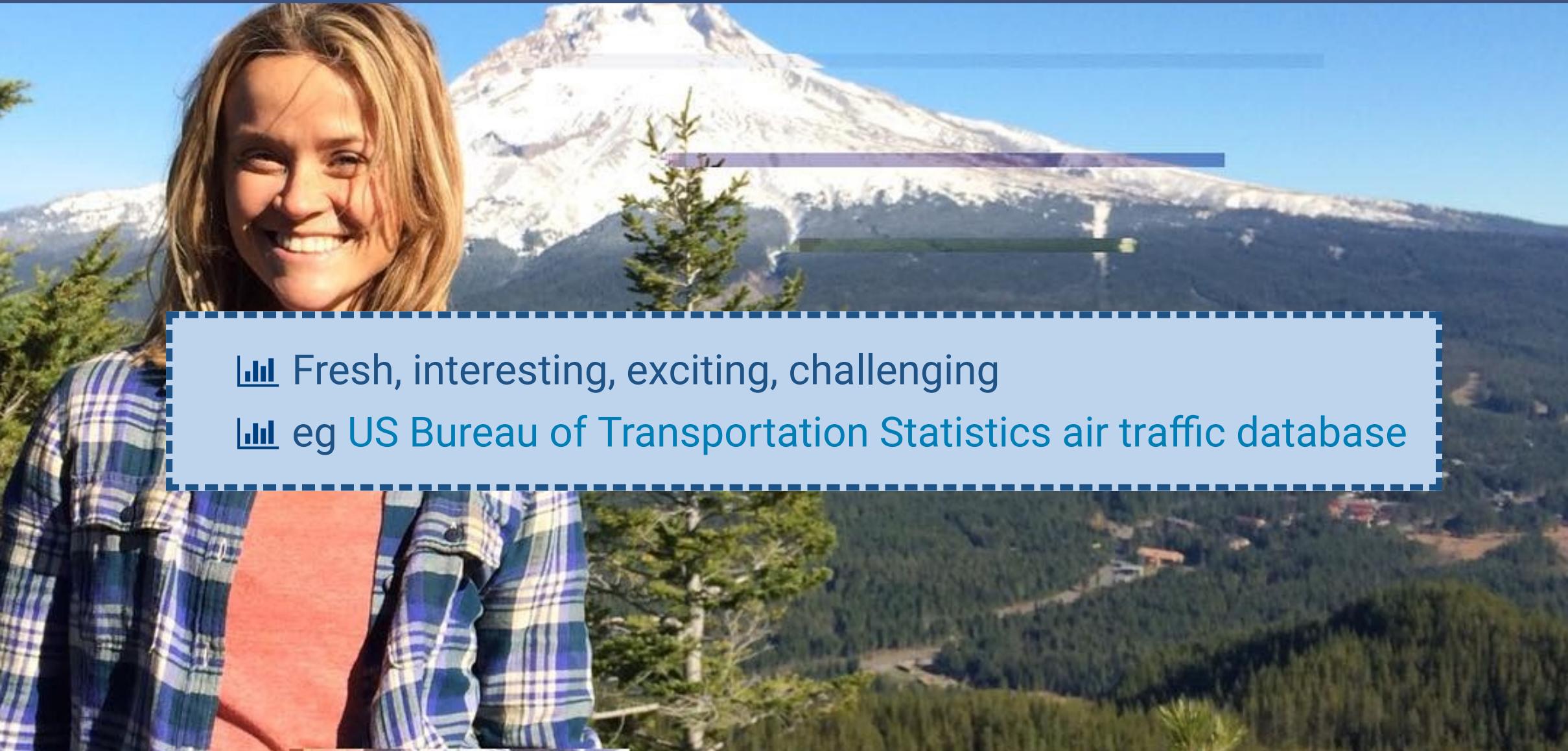
# Orphans

- 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.

# Wild



- Fresh, interesting, exciting, challenging
- eg US Bureau of Transportation Statistics air traffic database

# Fresh and local



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





# That's it!



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

Lecturer: Didier Nibbering

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu