



MONASH  
University

MONASH  
INFORMATION  
TECHNOLOGY

**TURN ON RECORDING!**





**FIT5147**

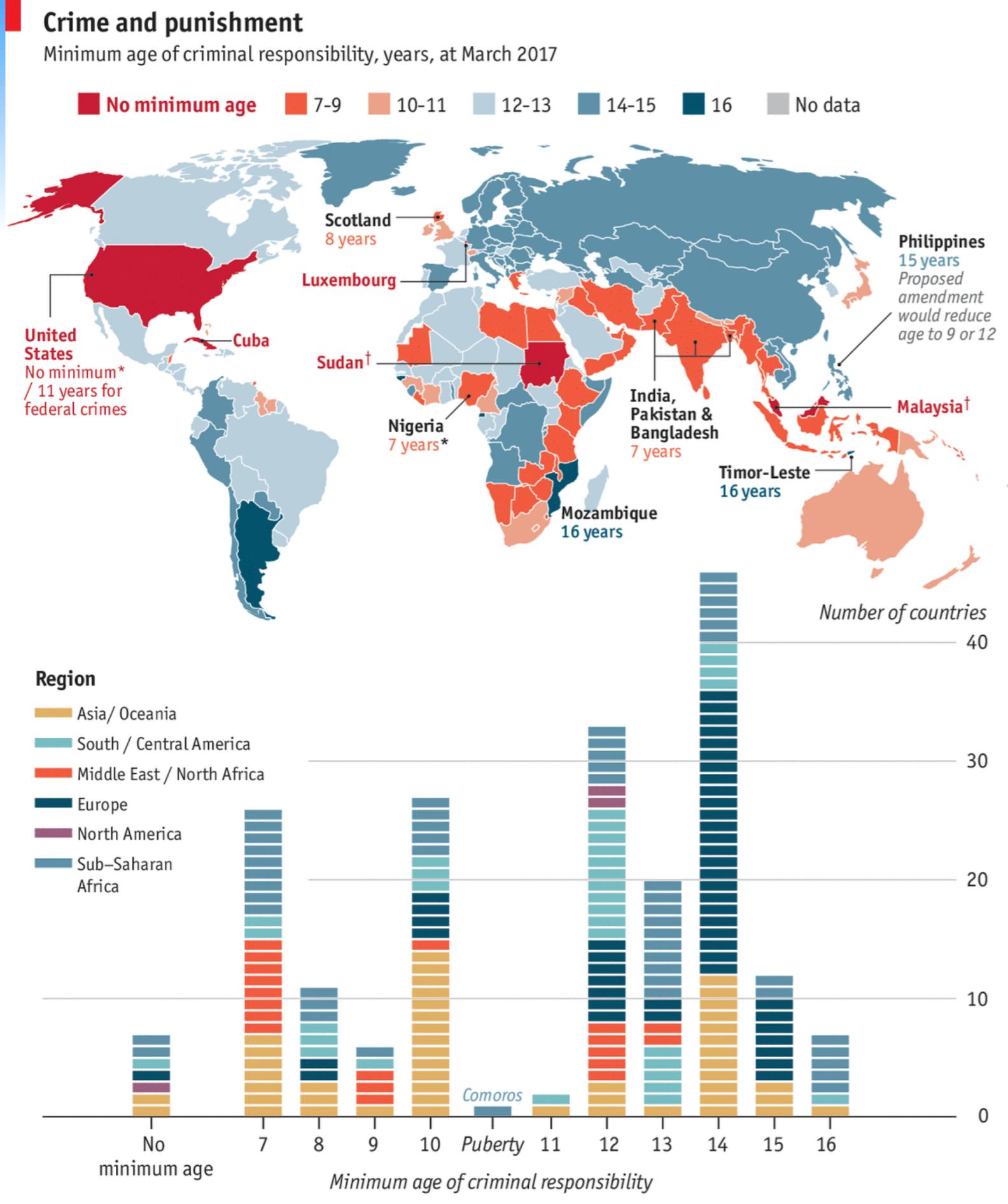
# Data Exploration & Visualisation

## Meet-up 4

# Meet-up 4 Agenda

- Visualisations to Share
- Module 4: Visualising Network and Textual Data and Analysis
- Tasks for the week
- Week 4 Design Critique

# Visualisation to Share



<https://www.economist.com/blogs/graphicdetail/2017/03/daily-chart-7>

Figure 5. Renewable Power Capacities in World, BRICS, EU-28 and Top 6 Countries, 2016

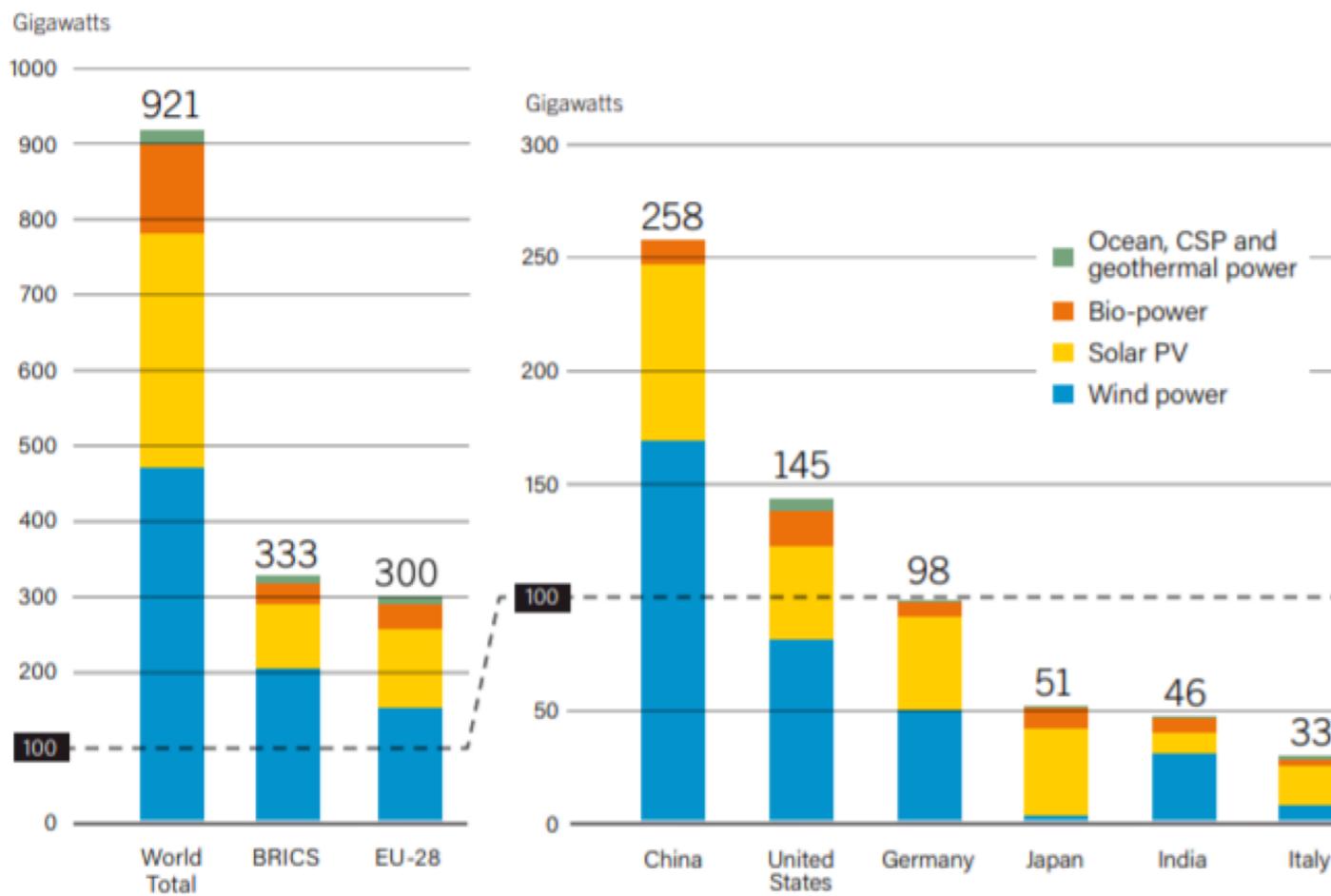
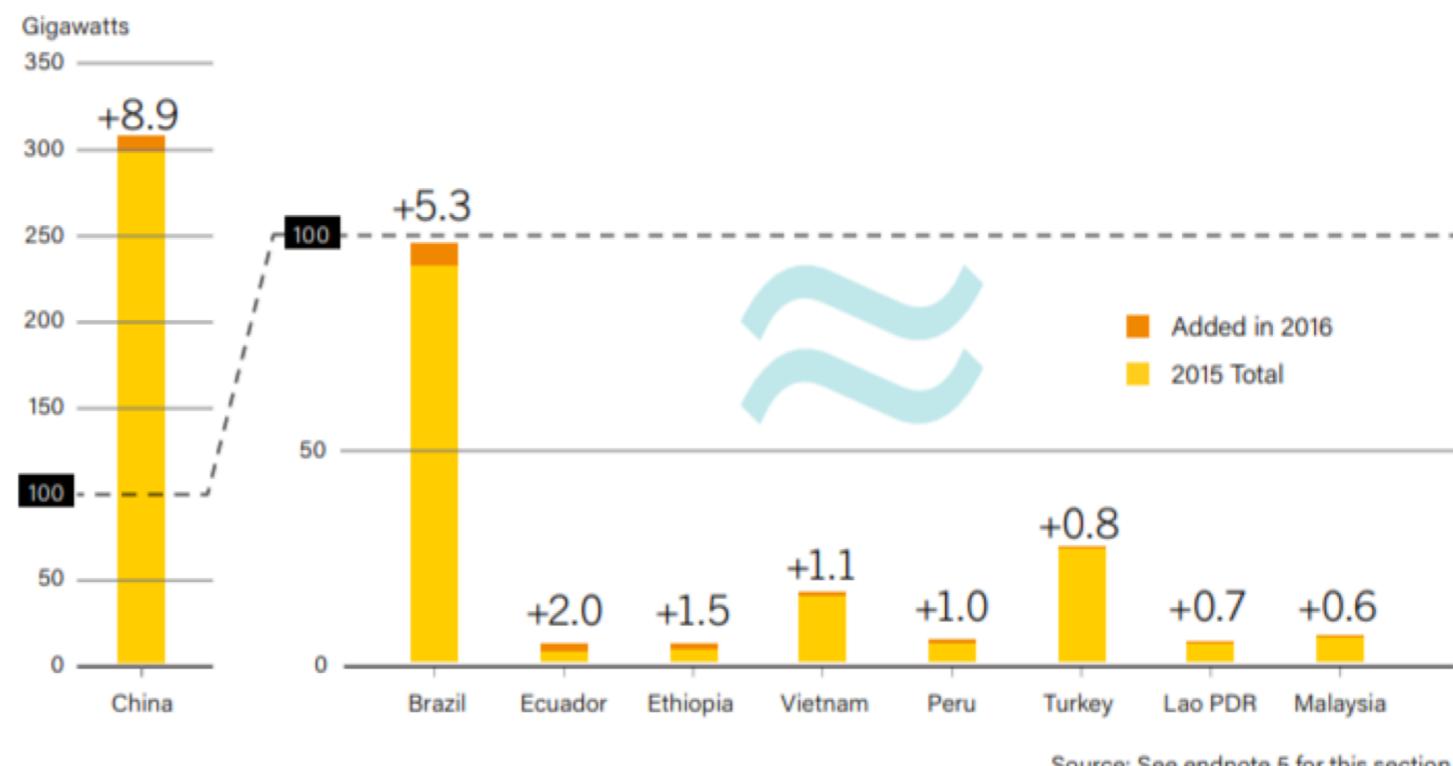


Figure 14. Hydropower Capacity and Additions, Top 9 Countries for Capacity Added, 2016



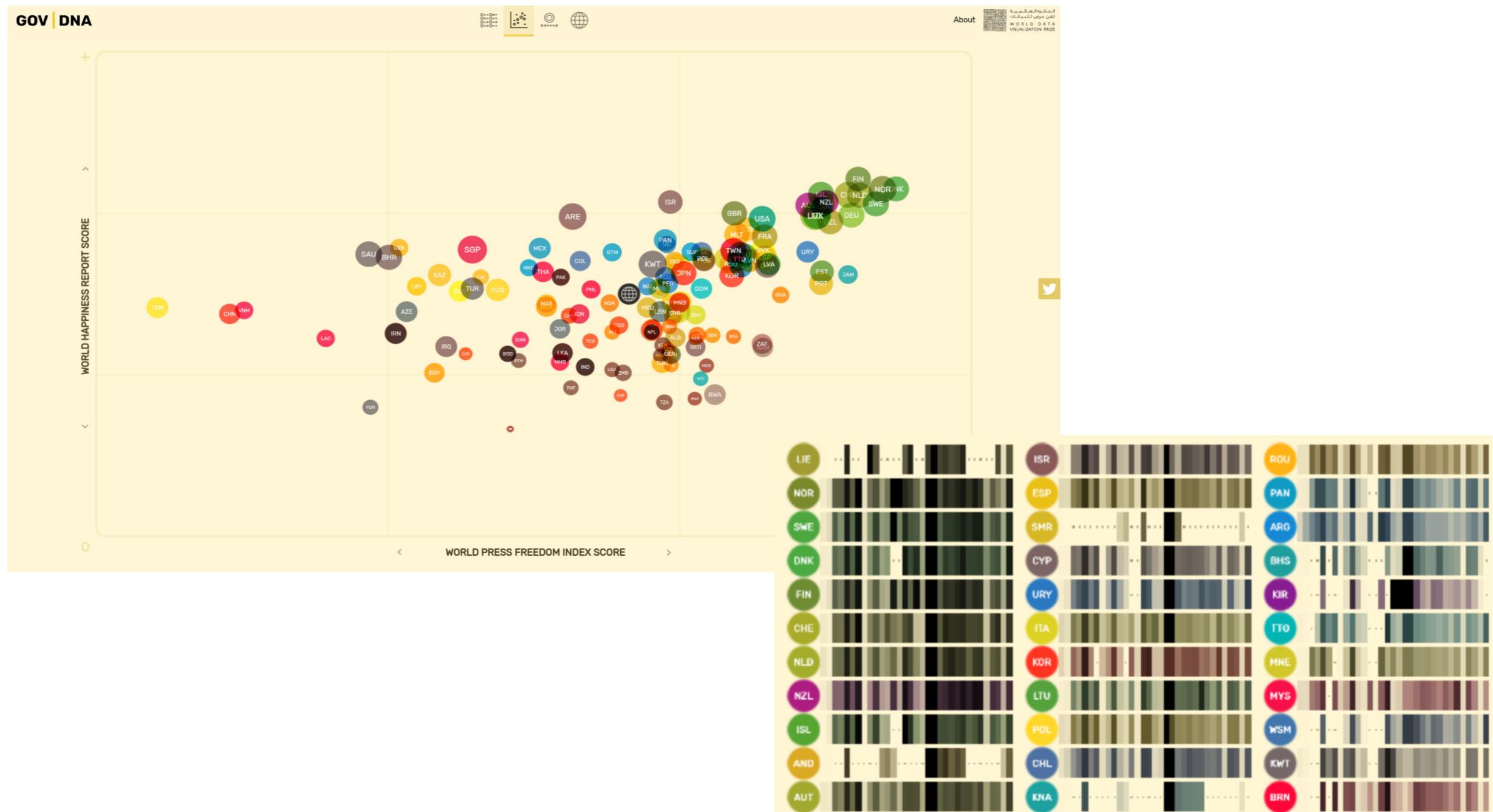
# Visualisation to Share

Source: See endnote 5 for this section.

# Visualisation to Share

World Data Visualisation Prize - [GOV|DNA](#)

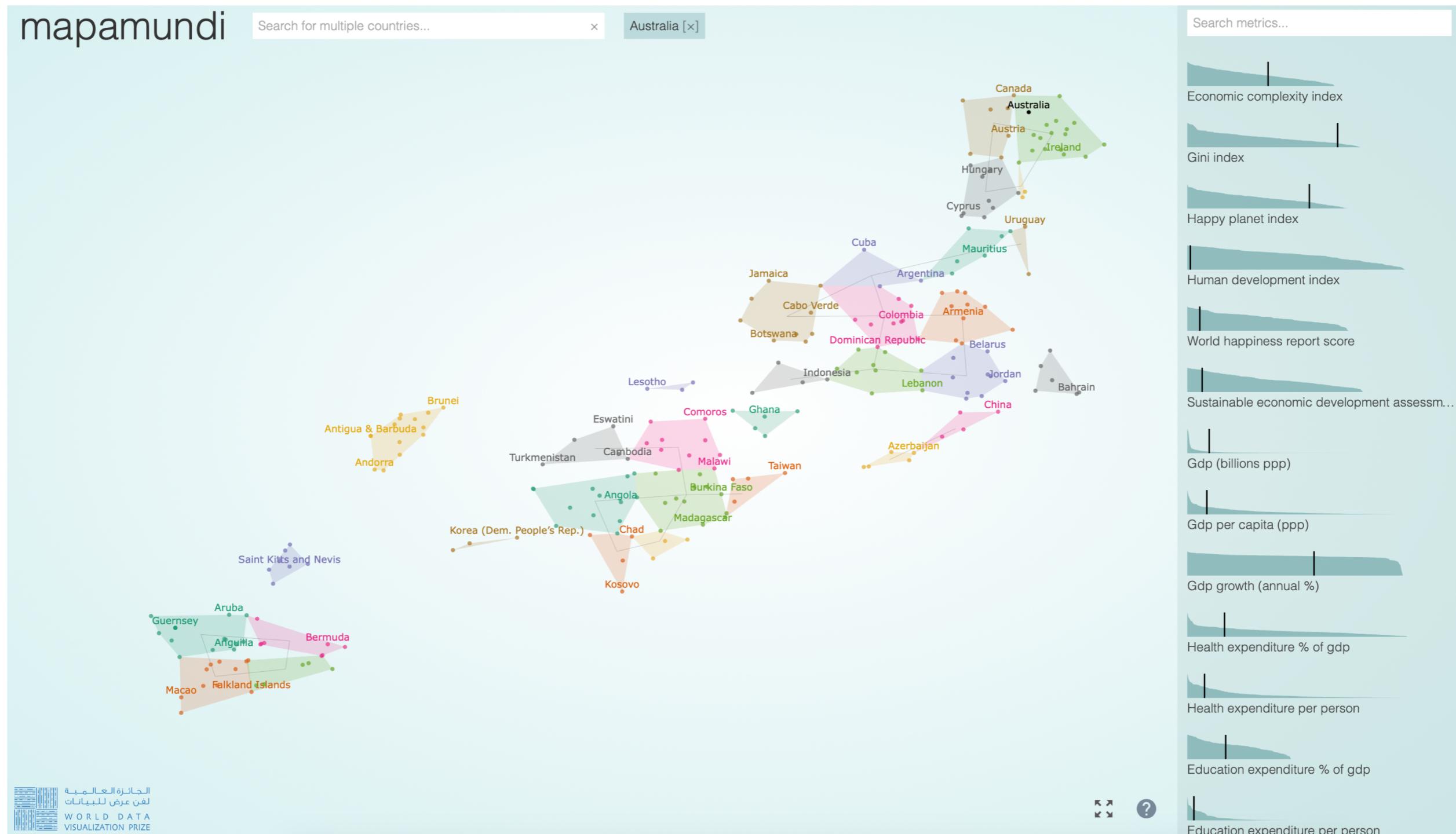
<https://informationisbeautiful.net/2019/winners-of-the-world-data-visualization-prize/>



# Visualisation to Share

World Data Visualisation Prize - [Mapamundi](#)

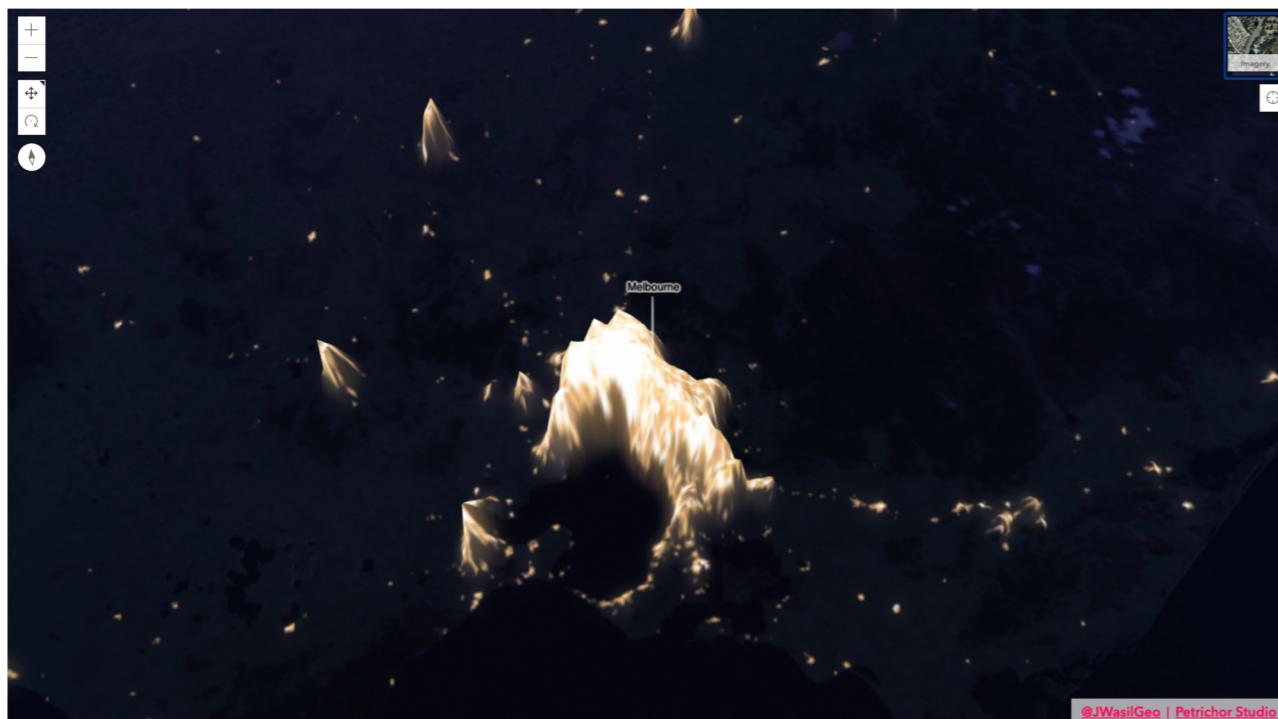
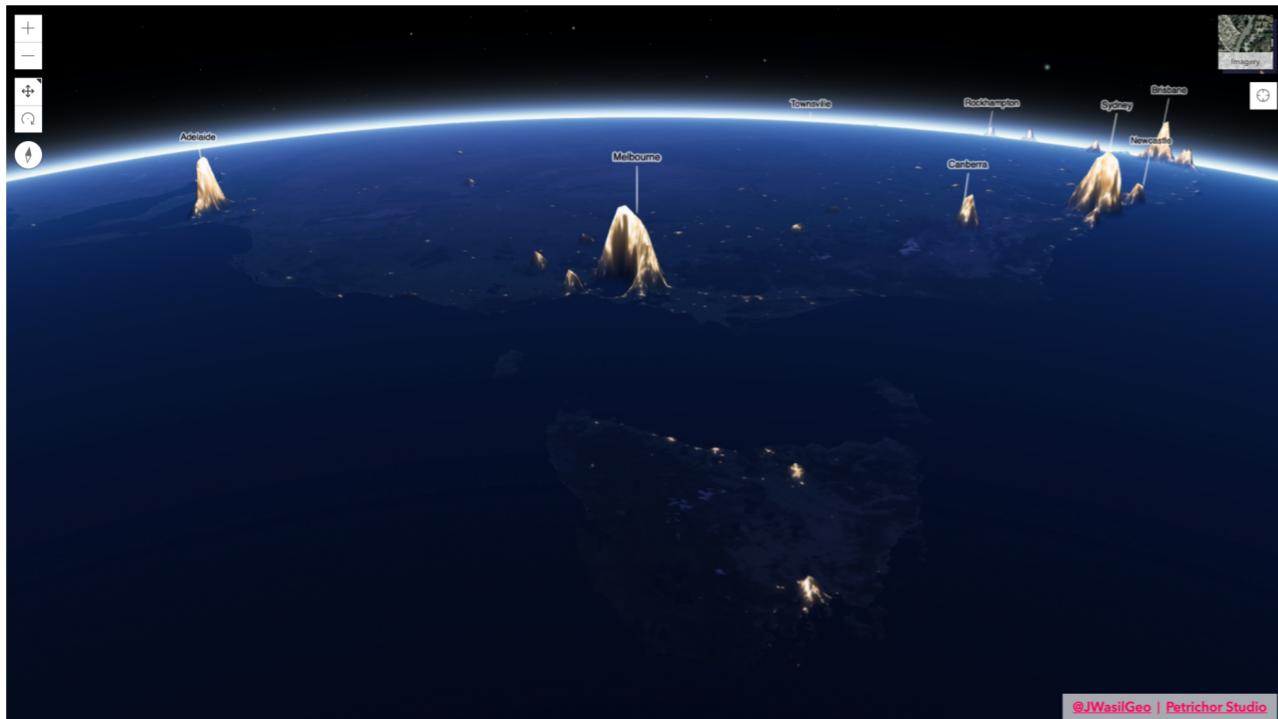
<https://informationisbeautiful.net/2019/winners-of-the-world-data-visualization-prize/>



# Visualisation to Share

Earth at Night, Mountains of Light

<https://petrichor.studio/2019/02/14/earth-at-night-mountains-of-light/>

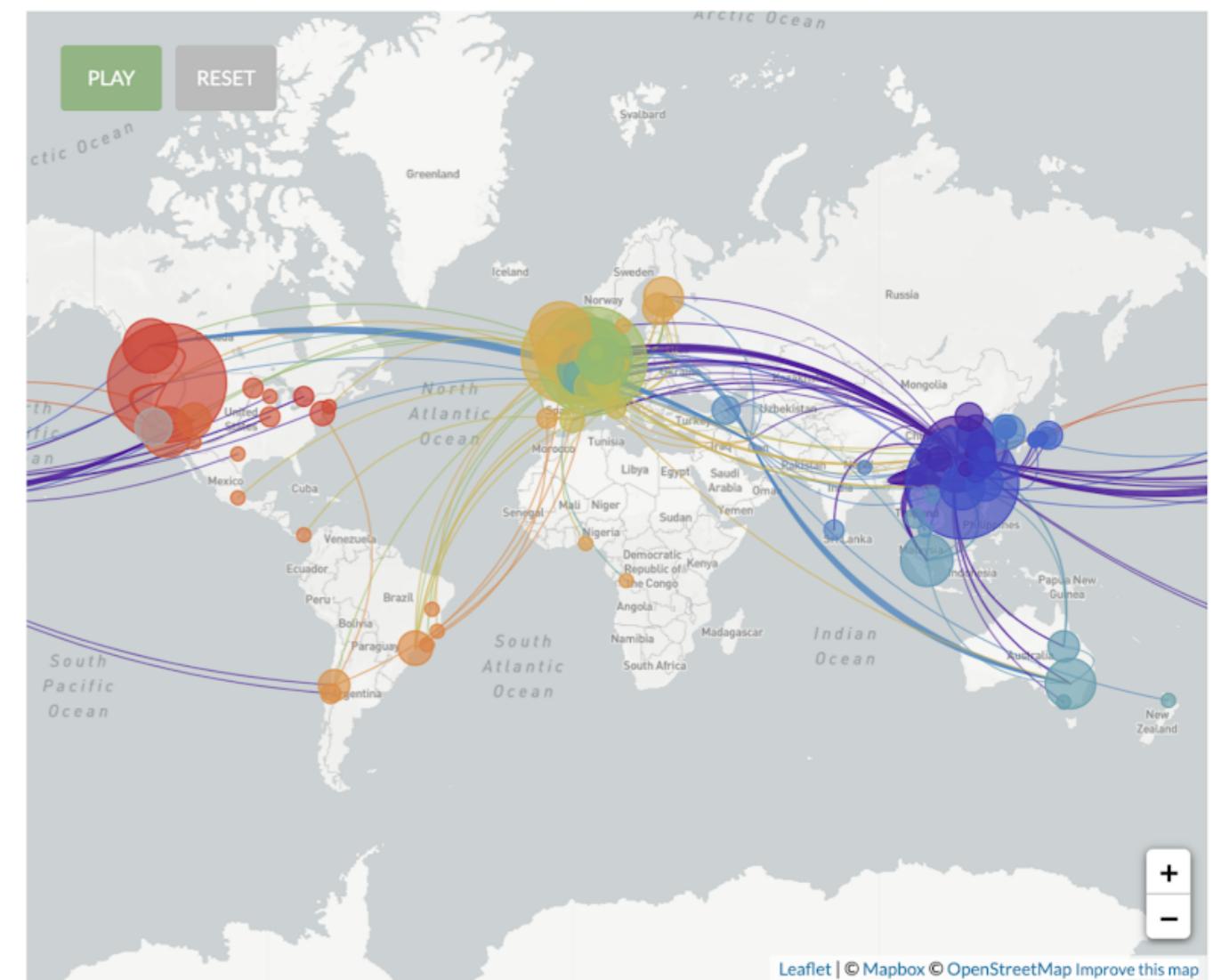
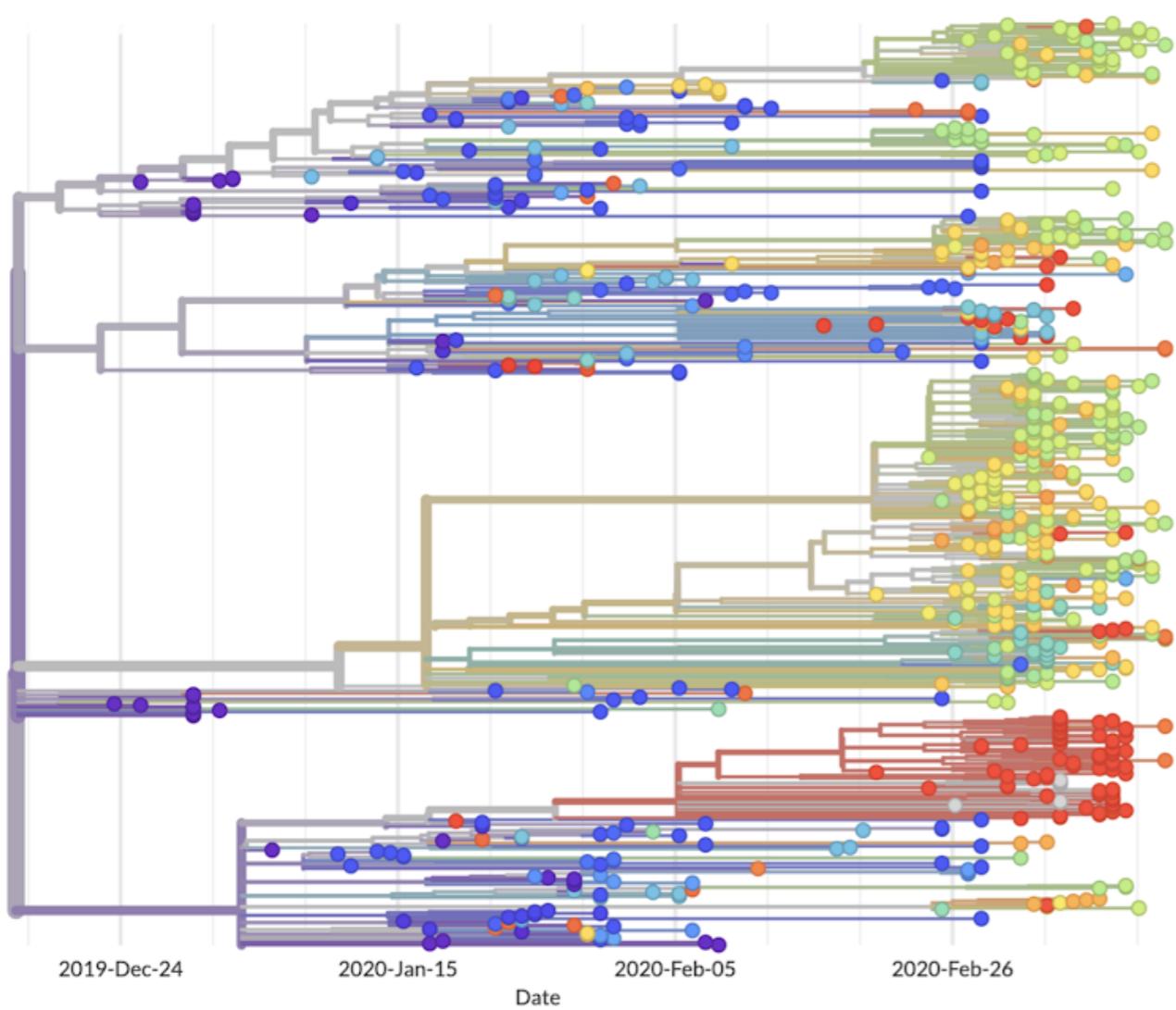


# Visualisation to Share

## Coronavirus: The Hammer and the Dance

<https://medium.com/@tomasgueyo/coronavirus-the-hammer-and-the-dance-be9337092b56>

Chart 6: Mutations in the Coronavirus



Source: Nextstrain, based on open source information gathered through GISAID

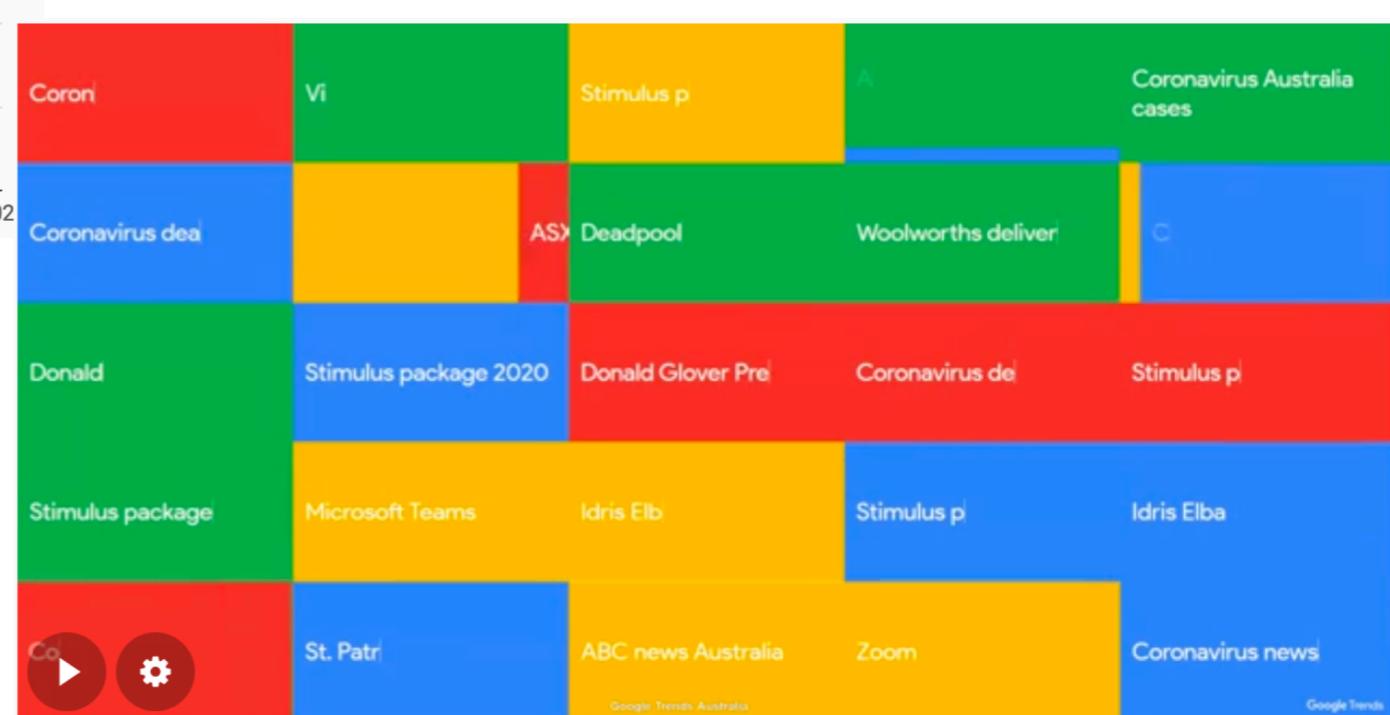
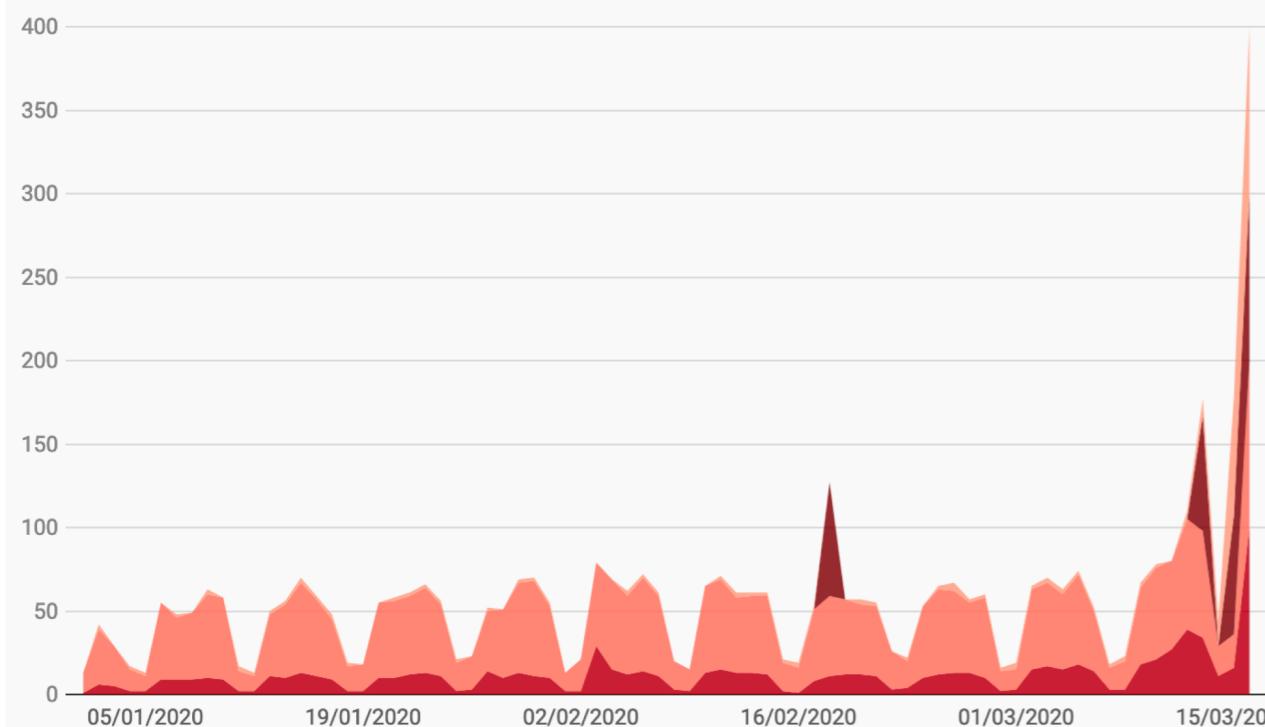
# Visualisation to Share

**Coronavirus Google searches show how the world is coping with the COVID-19 pandemic** by Alan Weedon (ABC News)

<https://www.abc.net.au/news/2020-03-22/google-searches-in-the-time-of-coronavirus/12064150/>

## Trending UK Google search topics

Microsoft Teams Slack nhs.co.uk/coronavirus Waitrose home delivery



# Data analyses and visualisation

- screencast by David Robinson (TidyTuesday)

<https://www.youtube.com/channel/UCeiiqmVK07qhY-wvg3IZiZQ>

Watch the Tidy Tuesday screencast:  
Analyzing data on women in the workplace

<https://www.youtube.com/watch?v=fv9SQ4IFNr4>

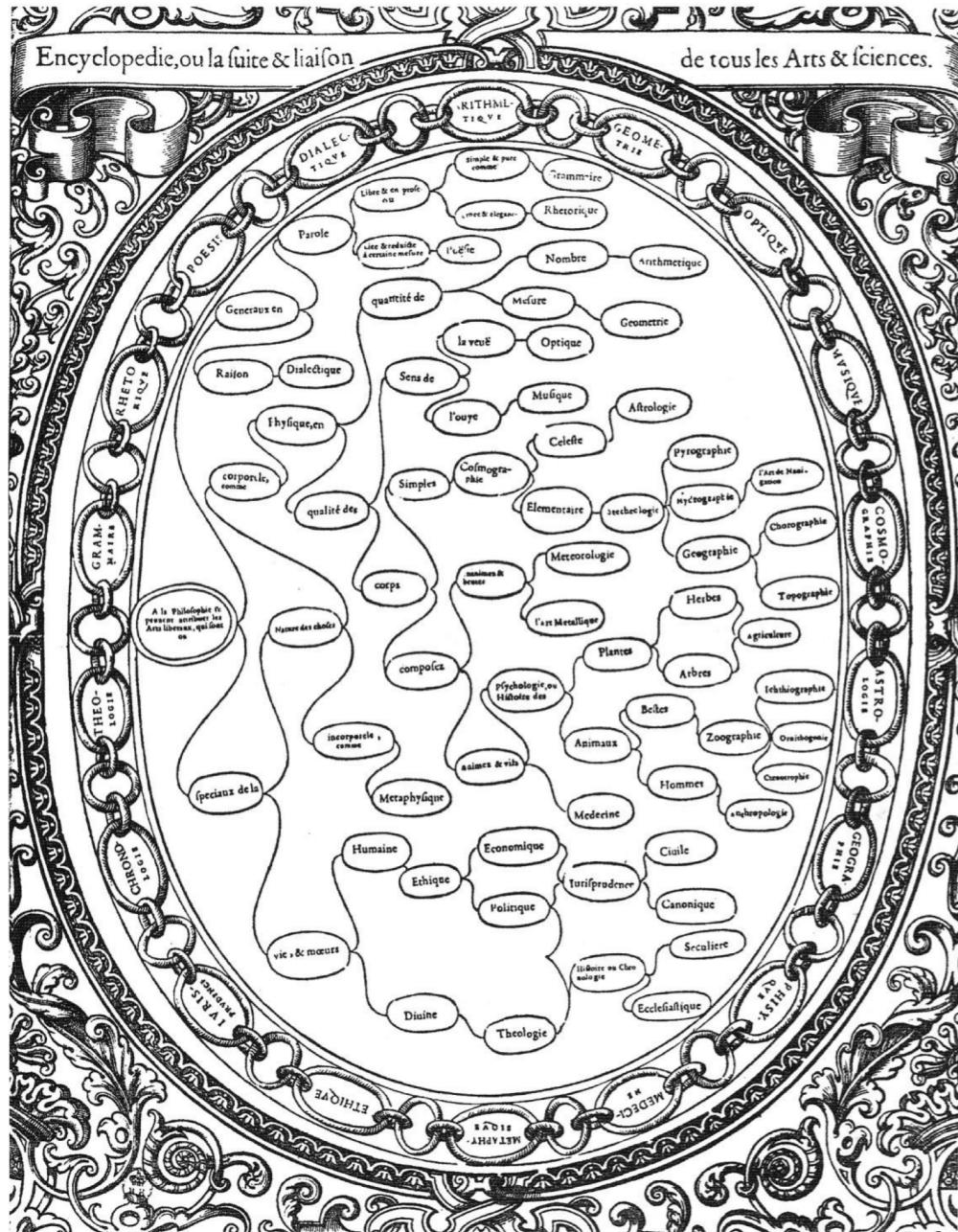


Week	Material	Activities	Assignment
1	Visual Analytics; History of Data Vis; Tools for Data Exploration & Visualisation	Intro to common data exploration tools; Five design sheet visualisation design methodology	
2	Visualisation of tabular data; analysis of trends & patterns in tabular data	Advanced and interactive graphics with R;	Quiz 1
3	Exploring spatial data with data maps	Tools for creating data maps;	Project: Project Proposal
4	Network data analysis & visualization; Textual data analysis & visualisation	Tools for network and text analysis; Project progress interview	DataVis Design Exercise
5	Human visual system; Visual communication	Design of effective visualisations; Introduction to D3; Project progress interview;	Quiz 2
6	Interactive data visualisation; the future of data visualisation	Design of visual analytics tools; Selected design groups presentations	Project: Final product and report

# Main Kinds of Data Sets

- **Tabular data:** Data organised in tables, a row for each data item and a column for each of its attributes.
- **Spatial data:** Data which is naturally organised and understood in terms of its spatial location or extent.
- **Network data:** Nodes in the network are data items and links between the nodes are relations between. For instance a social network.
- **Sequential data:** Data organised in a sequence. Text is an example of this.

# Understanding Network & Textual Data



Family trees and tables have existed for a long time but not other ways to understand networks

# This module

- Understanding networks
  - Understanding text
  - Tools for understanding networks and text

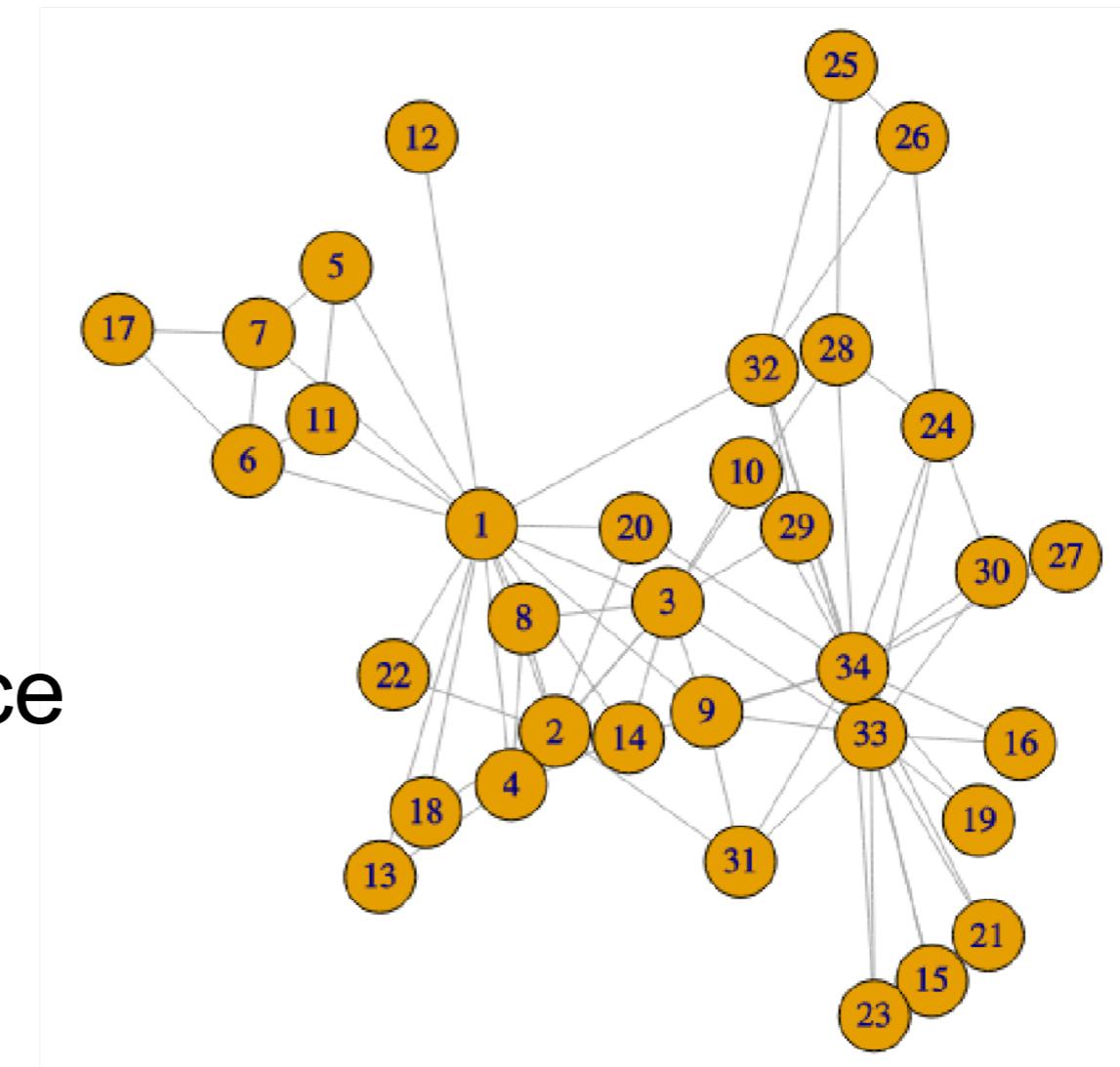
# Networks

Networks encode relations (edges) between items (nodes).

Both edges and nodes can have attributes.

## Key concepts

- Neighbour
- Directed
- Shortest path
- Graph theoretic distance between two nodes
- Connected
- Weight



# Ways to Visualise Relational Data

- Node-link diagram
- Orthogonal node-link diagram
- Radial network diagram
- Layered network diagram
- Adjacency matrix
- Chord diagram
- Hive plot
- Arc diagram
- Indented list
- Layered tree
- Dendrogram
- Cladogram
- Tree map
- Icicle plot
- Sunburst plot
- Argument map
- Sankey diagram
- Flow diagram
- MatrixWave
- Venn diagram
- Euler diagram
- Linear diagram
- Spider diagram
- Power graph

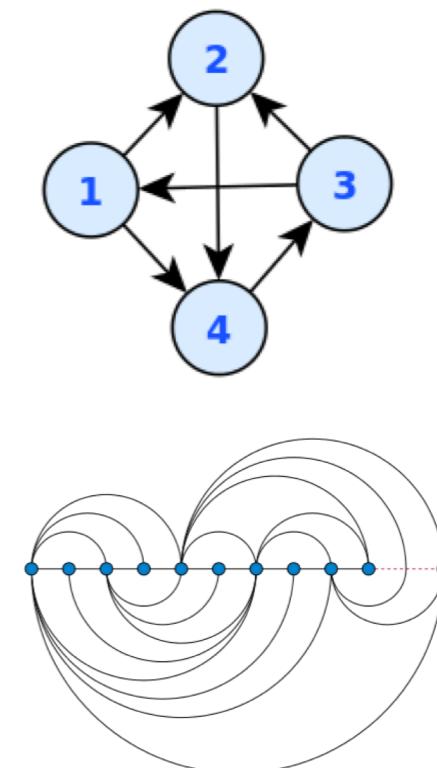
# Group activity: Visualising Relational Data

- Later in the meet-up, we will break up into groups. Each group will have 3 of those visualisations to research, providing information on:
  - Description
  - What is it good for
  - Example image
  - Update these details on your group's version of the Wall Chart of Network Visualisation, found in

<https://drive.google.com/open?id=1EIEEnHoVVZhzLvls1rkNfLvZJRjKnhOm>

- If you can't directly edit the file, be sure to email it to Michael by Monday 30 March.

- Node-link diagram -- Sample
- Description:** consist of nodes (representing entities) and links (representing relations).
- Good for:** visualizing network data, entities and the relation between entities relations



From:

[https://en.wikipedia.org/wiki/Graph\\_drawing](https://en.wikipedia.org/wiki/Graph_drawing)

# Network Centrality

## Node centrality measures

- Degree
- Closeness
- Betweenness
- Status or rank

## Edge centrality

- Betweenness

# Understanding Text

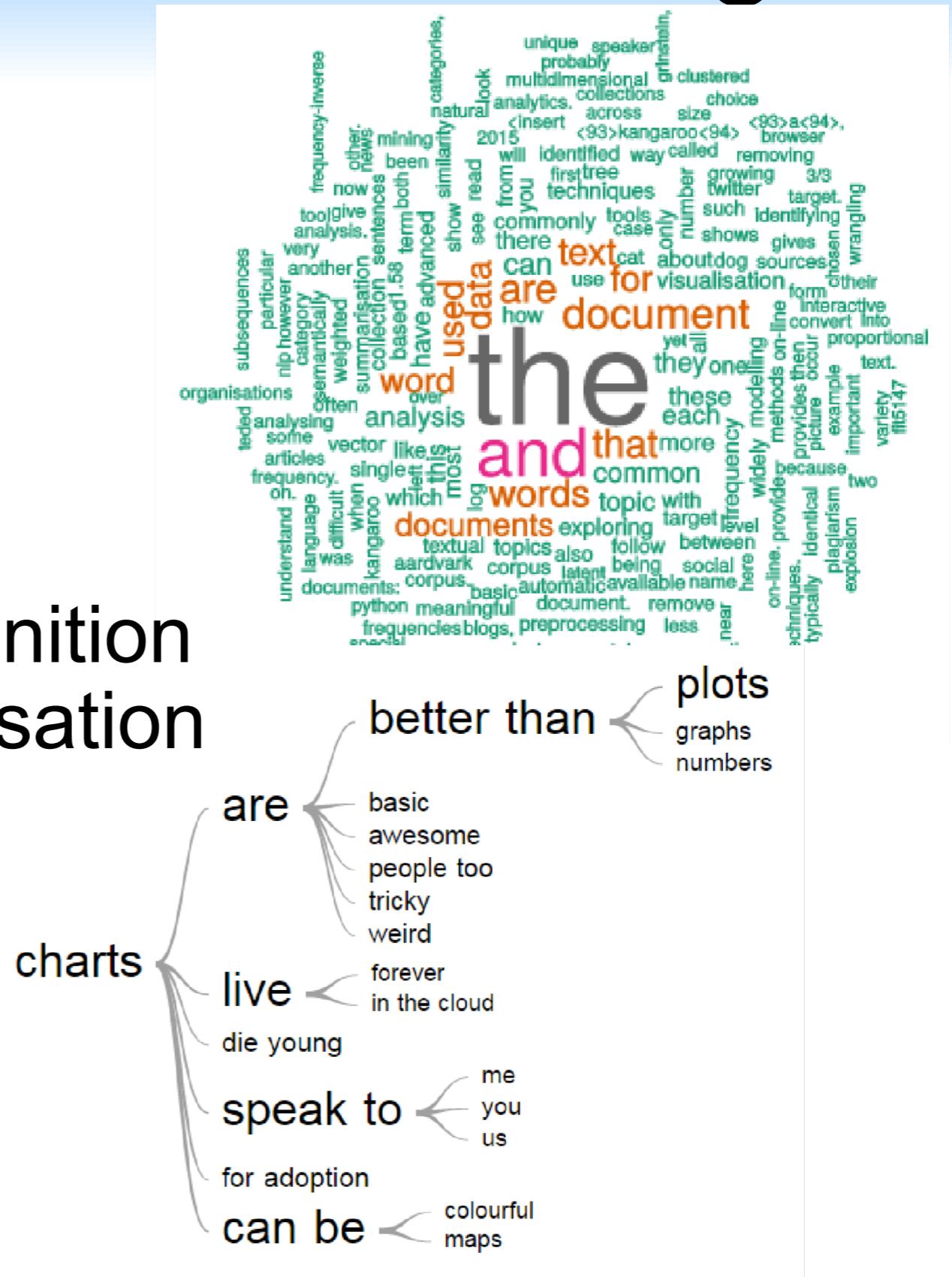
Text is unstructured and hard to understand  
but some automated processing and analysis  
is still possible.

Preprocessing steps:

- Remove punctuation
- Remove numbers
- Normalise to lower (upper) case
- Expand contractions
- Remove “stop words”
- Recognise compound words
- Determine stem words

# Document Understanding

- Word cloud
- Word tree
- Sentiment analysis
- Named entity recognition
- Automatic summarisation



# Understanding Document Collections

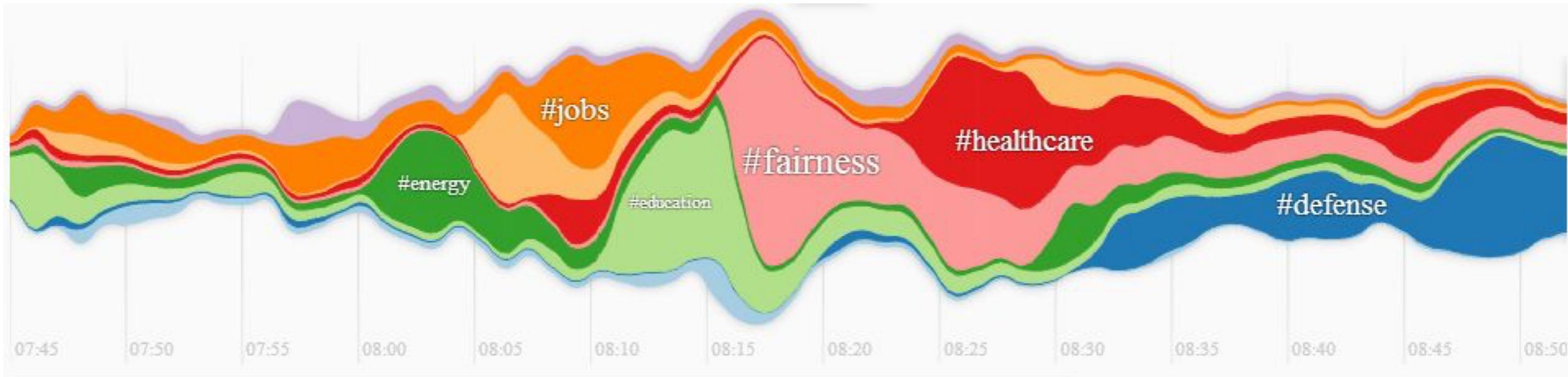
Document similarity: use a vector of words and for each document and word compute the

$$\text{tf-idf}(w) = tf(w) \times \log \left( \frac{N}{df(w)} \right)$$

Gives a measure of similarity between documents in the collection

Use clustering or dimension reduction techniques to understand the collection

# Understanding Document Collections



## Topic modelling

- Identify groups of words that occur together across the collection: these form “topics”
- Identify the topics in each document

# Tools for Network & Text Analysis & Visualisation

- R – igraph, tm
- SAS Visual Analytics
- Python NTLK
- Java Stanford NLP Tools
- Other specialised text mining tools

# Summary

- Visualisation and analytical methods for understanding tabular and spatial data are well-understood—studied for centuries.
- However lots of the most interesting data is network data or textual data.
- Understanding text is particularly difficult because it is unstructured.
- Visualisation and analytics for network and textual data is not as mature, though rapidly evolving.

# Rest of the Week...

- Read eTextbook for Module 4
- Participate in your Meet-up group's activities
- Finish the e-Textbook activities
- **A1 | Project Proposal**
  - Should have been submitted already!
- **A2 | DataVis Design Exercise**
  - Work with your group designing using FDS methodology
  - Due to be submitted Monday Week 5
- **A3 | Data Exploration and Visualisation Project**
  - Stage 4: Complete your FDS visualisation design.
  - Stage 5: Start implementing your final design in R or D3
- **Interviews:**
  - We will inform you if a Zoom interview is needed and how to set it up

# Week 4 Design Critique

You are to come up with a visualisation to help understand a data set collected by agricultural scientists who are comparing the yield rate (tonnes per hectare) for two different strains of wheat across Australia and exploring how soil type, average rainfall and average yearly temperature affects this. Their data set consists of 1000 records. Each record gives the yield rate for one of the strains of wheat, soil type, rainfall and temperature at a different location (latitude, longitude) in Australia. The locations are not uniformly distributed but concentrated around the coast.

- You will be allocated to a group of 3-4 to brainstorm and design a way to visualise the data and explain the reason for your design and any alternatives you considered.
- Note that the data does not need to be shown in a single view. If you like, you can use multiple views, in which case you should describe how they are linked and what interaction links them.

<https://drive.google.com/open?id=1EIEnHoVVZhzLvls1rkNfLvZJRjKnhOm>