MONASH University

FIT1043 Introduction to Data Science

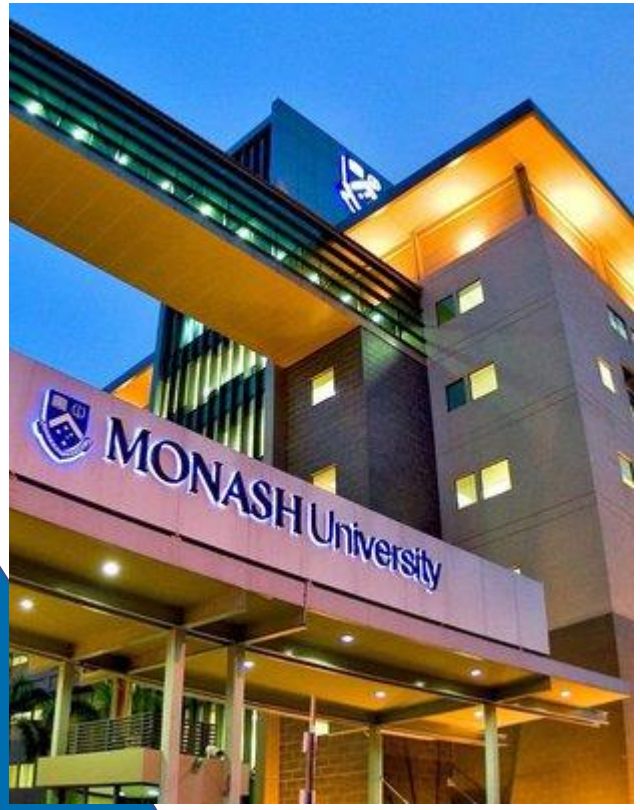
Week 3

Dr. Sicily Ting Fung Fung

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Week 2 Coverage
Overview of Data Science
Python for Data Science

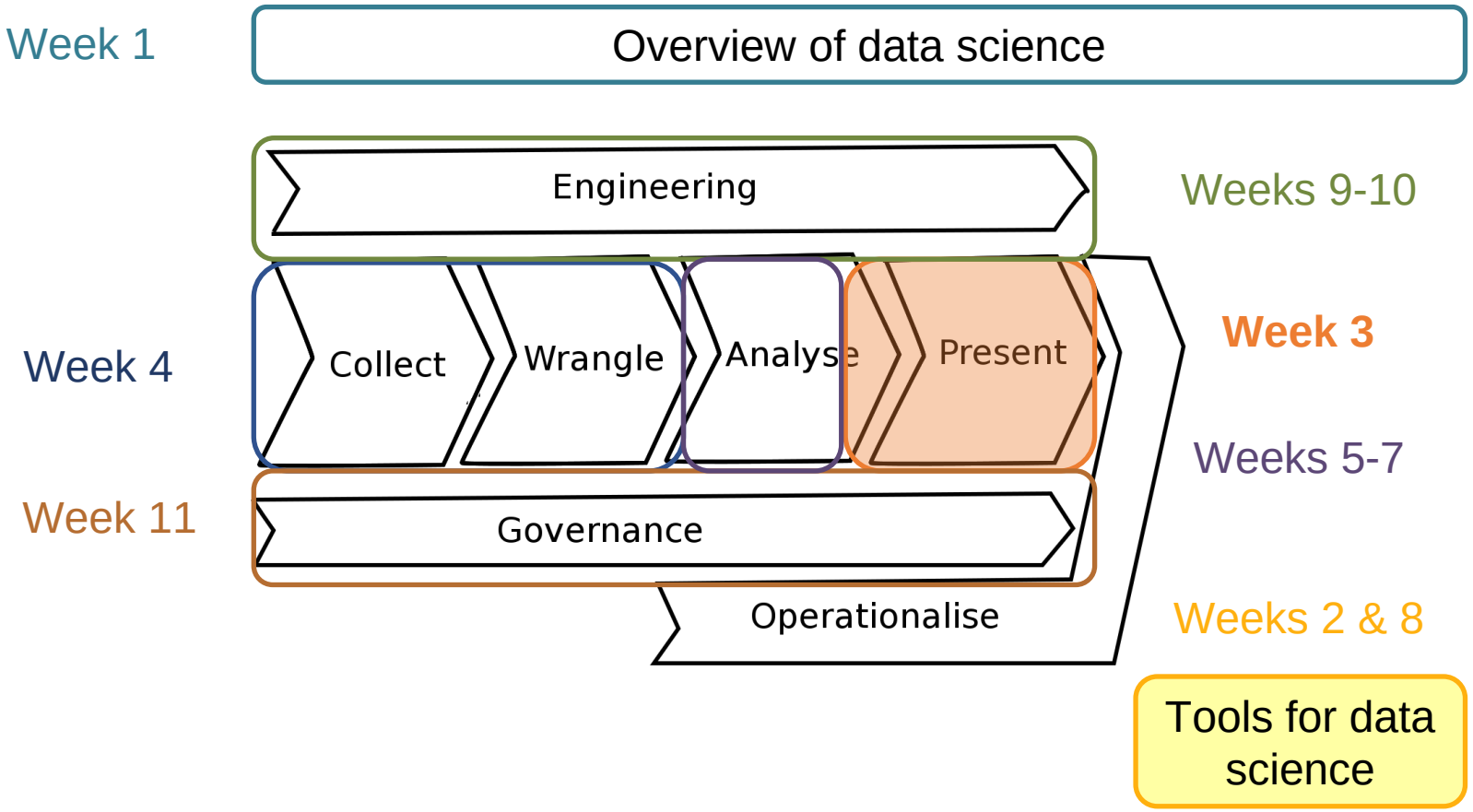


MONASH University

Week	Activities	Assignments
1	Overview of data science	
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture (tentative)	Assignment 3

Weekly Quiz From Week 2-11
- The quiz will open for 48 hours and you are allowed to have 2 attempts

MONASH University



Week 3 Outline

Introduction to Python for Data Science

- Advanced Aggregation in Python

Data Visualisation

- Introduction and Basic Visualisation Plots
- Descriptive Statistics

Learning Outcomes

Week 3

By the end of this week you should be able to:

- Comprehend more sophisticated group-by operations and graphing in Python
- Comprehend the power/**importance of data visualisation**
- Differentiate between **approaches for data visualisation**, and explain where each approach is appropriate to be used
- Explain/differentiate different concepts in **descriptive statistics**

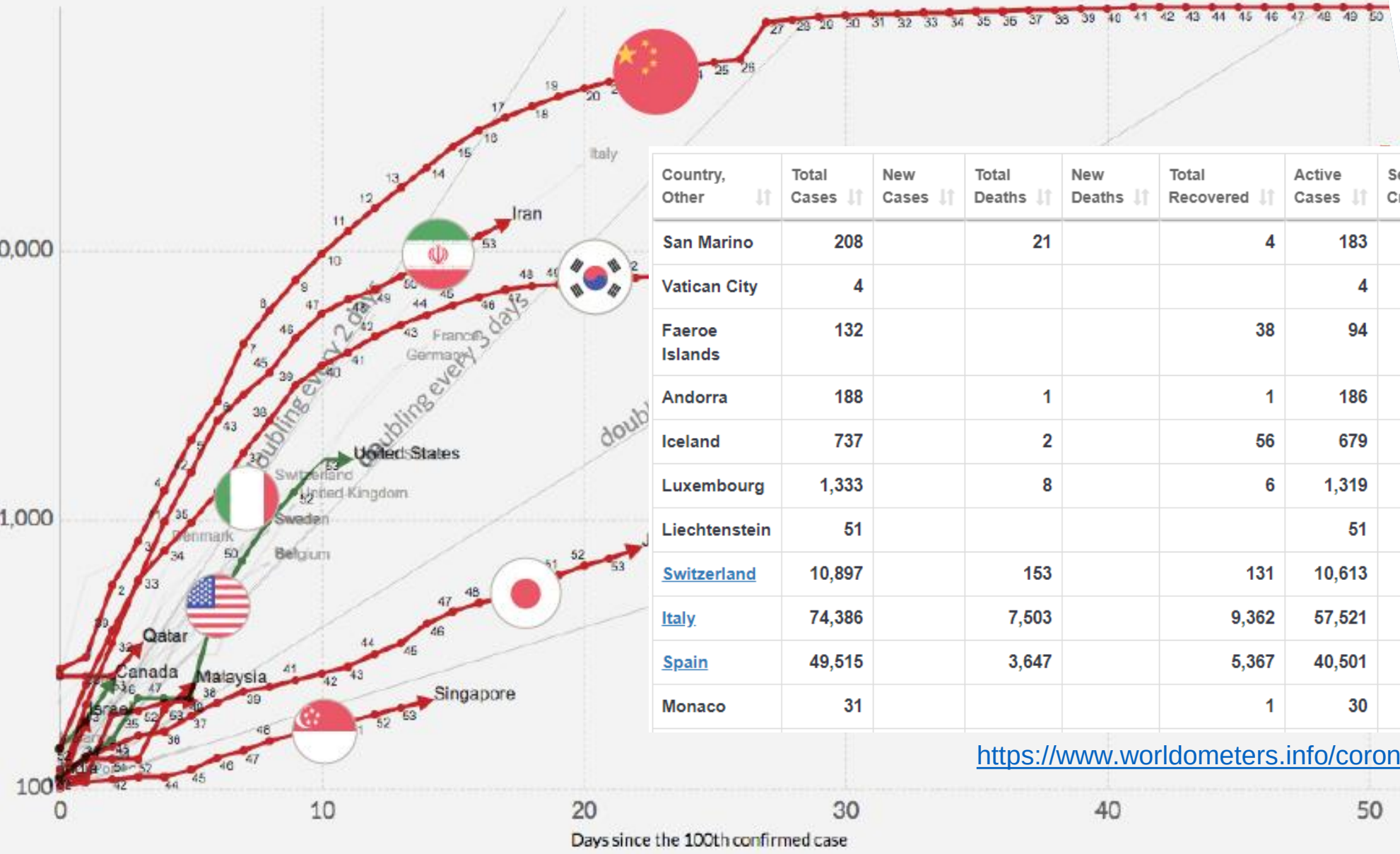
MONASH University

Turning Powerful Statistics into Art

MONASH University

https://www.visualcapitalist.com/infection-trajectory-flattening-the-covid19-curve/

Visualization vs Pure Data



Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Tot Deaths/ 1M pop
San Marino	208		21		4	183	12	6,130	619
Vatican City	4					4		4,994	
Faeroe Islands	132				38	94	2	2,701	
Andorra	188		1		1	186	6	2,433	13
Iceland	737		2		56	679	11	2,160	6
Luxembourg	1,333		8		6	1,319	3	2,129	13
Liechtenstein	51					51		1,338	
Switzerland	10,897		153		131	10,613	141	1,259	18
Italy	74,386		7,503		9,362	57,521	3,489	1,230	124
Spain	49,515		3,647		5,367	40,501	3,166	1,059	78
Monaco	31				1	30		790	

https://www.worldometers.info/coronavirus/

MONASH University

Data Visualisation

From Introduction to Probability and Statistics for Engineers and Scientists, by S. M. Ross

From the previous slide, what do you think that was the purpose of the visualisation?

*“ … data visualisation is useful as a preliminary form of data analysis to get a "**feel**" for the data …”*

Data Visualisation

It is often useful to visualise data

- Can sometimes quickly reveal patterns
- However, going beyond two dimensions is problematic

Piece of paper or monitor now only limits us to 2 dimensional data (but with clever tricks, we can visualize up to 6 or 7 dimensions on a flat screen. We aren't going to cover that in this course but you get to see 3 dimensional data in your lab (by using colour).

Basic Types of Data

Numeric-Discrete

- Numeric, but the values are **enumerable**
- E.g., Number of live births, Age (in whole years)

Numeric-Continuous

- Numeric, **not enumerable** (i.e., real numbers)
- E.g., Weight, Height, Distance from CBD

Basic Types of Data

Categorical-Nominal

- Discrete numbers of values, **no inherent ordering**
- E.g., country, state, gender

Categorical-Ordinal

- Discrete number of states, but with an **ordering**
- E.g., Education status, State of disease progression

Data Visualisation

Visualise data to quickly reveal patterns

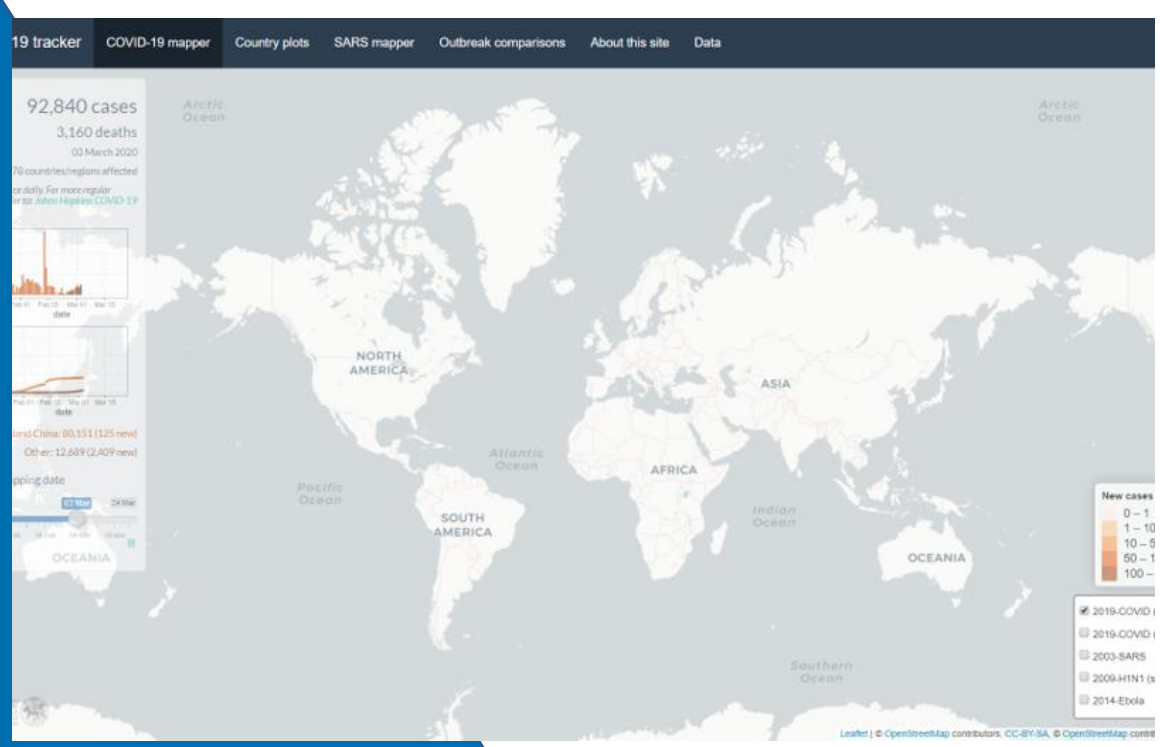
For numeric data (continuous and discrete), we can use:

- Histograms
- *Box plots (will revisit this in Descriptive Statistics)*
- Motion charts

For categorical data, standard visualisations include:

- Frequency tables
- Bar graphs
- Pie charts

Visualizing Categorial and Numeric Data



MONASH University

Frequency Tables

(Not a Graph … obviously)

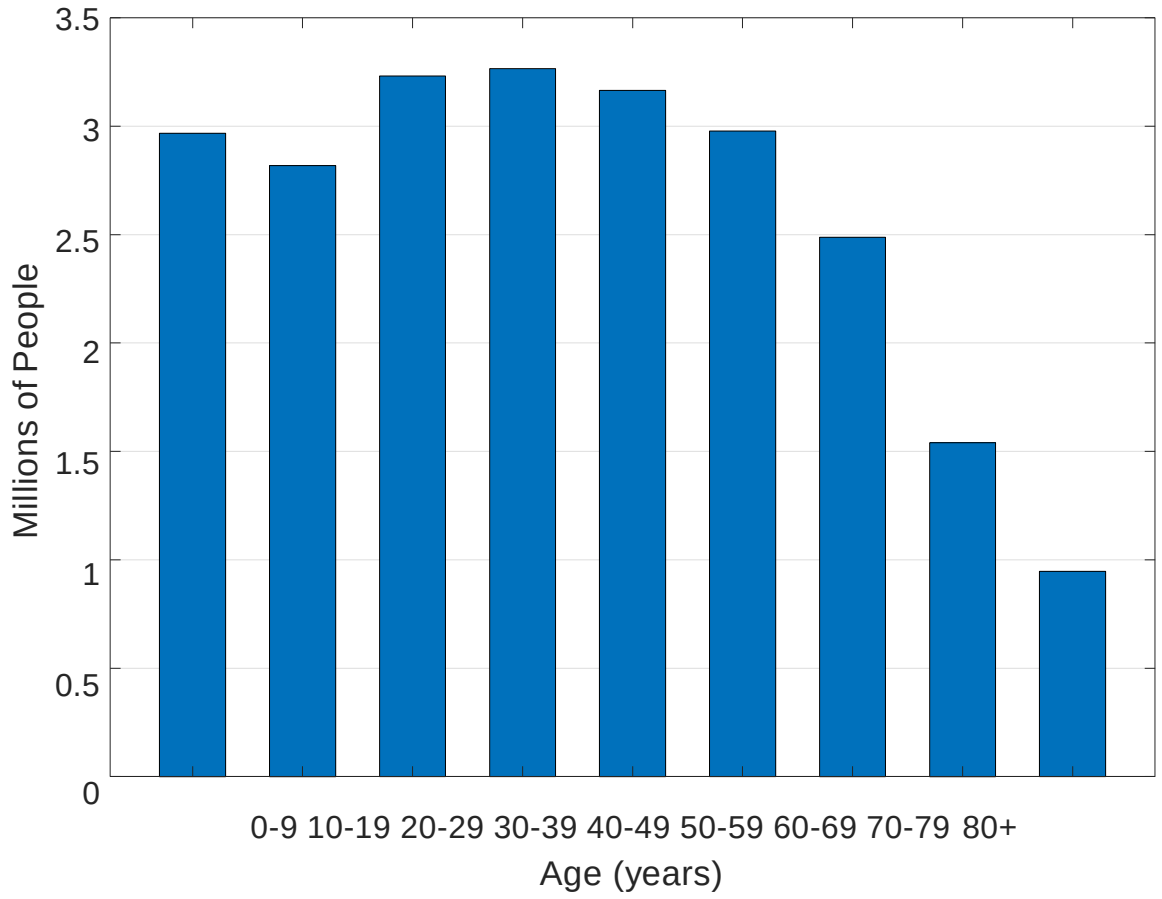
Age (years)	Number of People
0-9	2,967,425
10-19	2,818,778
20-29	3,231,395
30-39	3,265,526
40-49	3,164,712
50-59	2,977,883
60-69	2,488,396
70-79	1,540,373
80+	947,411

Australian Population by Age (2016 Census)

Frequency table is a **chart** that summarizes **values** and their **frequency**. It's a useful way to **organize** data if you have a list of numbers that represent the frequency of a certain outcome in a sample. Putting it into this form helps make it simpler to understand and further analyse.

MONASH University

Bar Charts



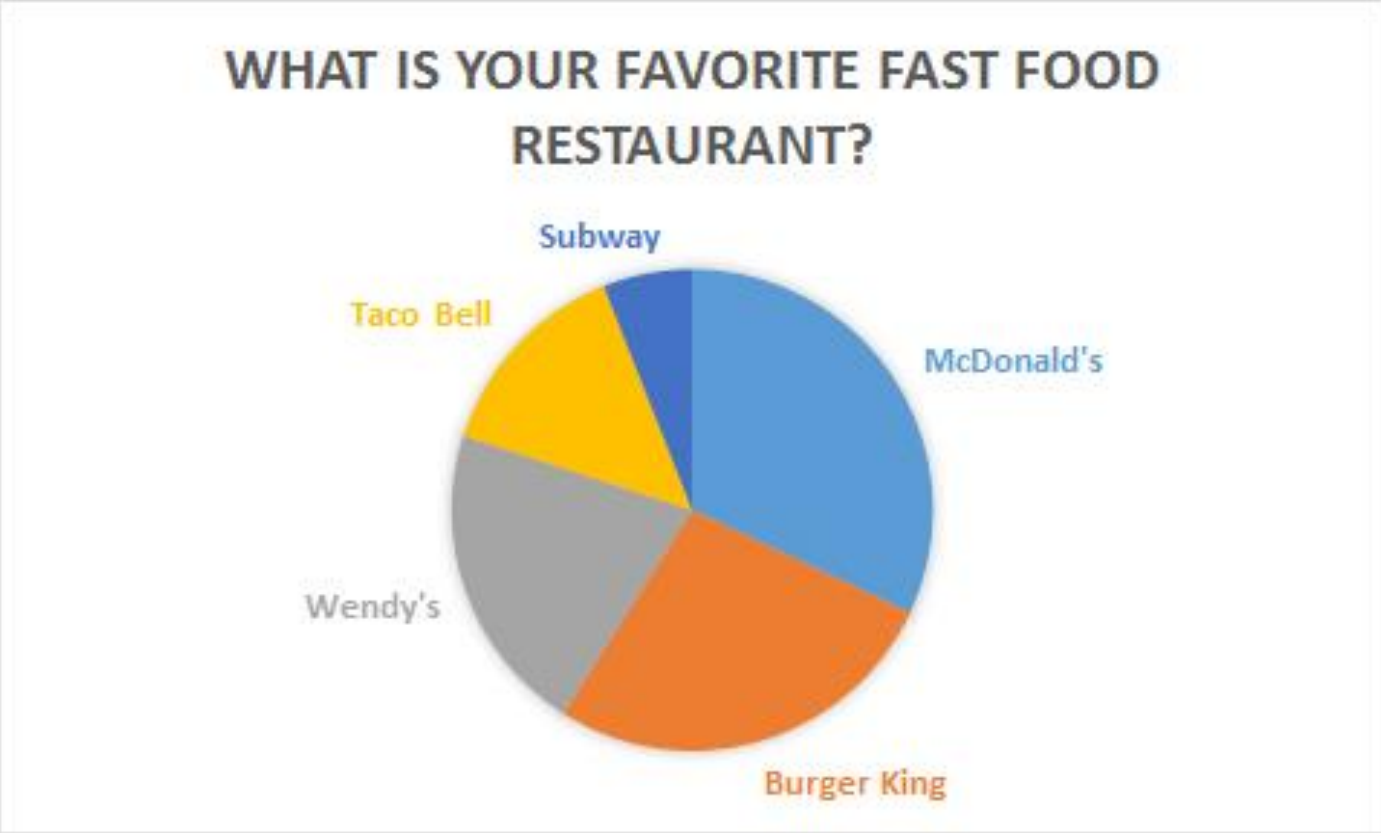
Australian Population by Age (2016 Census)

Compare between different groups, or
Show changes over time

A bar diagram makes it easy to **compare** sets of **data between different groups at a glance**.

MONASH University

Pie Charts



Pie Chart is a type of graph in which a circle is divided into sectors that each represent a proportion of the whole

Note, usually, for 6 or less categories. When there are more categories, it is difficult for the eye to distinguish between the relative sizes of the different sectors and so the chart becomes difficult to interpret.

Image source: varsitytutors.com

MONASH University

Histograms

Group **numeric** data into categories by putting into bins

If $\boldsymbol{y} = (y_1, \ldots, y_n)$ are our data points, we divide them into $\boldsymbol{K}$ equally spaced bins, i.e.,

The number of samples that fall in bin (category) $\boldsymbol{K}$ are

$$v_k = \#\{y_j \in (\min\{y\} + (\boldsymbol{K} - 1)w, \min\{y\} + \boldsymbol{K}w)\}$$

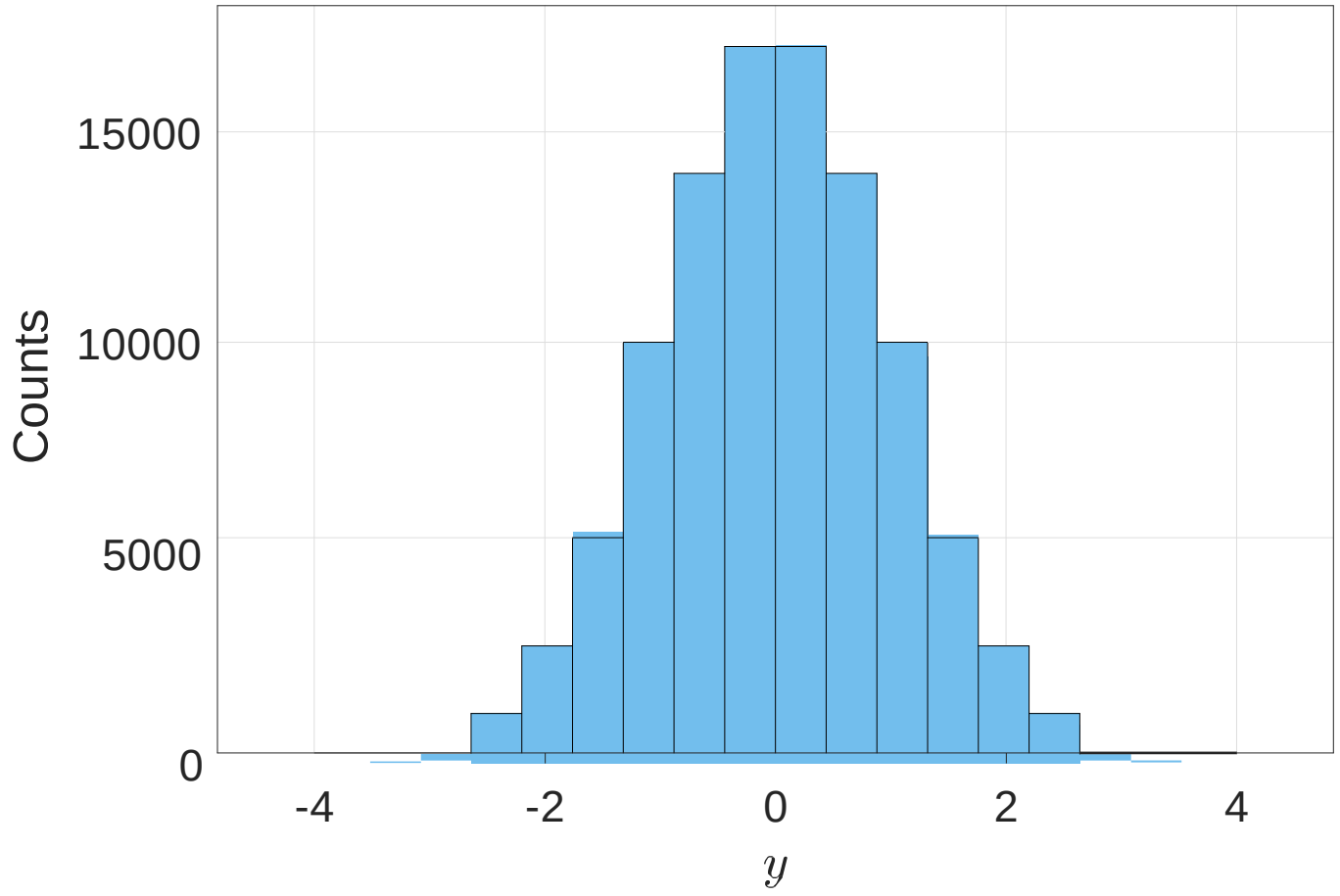
Where $$w = \frac{\max\{\mathbf{y}\} - \min\{\mathbf{y}\}}{K}$$

is the width of the bins

Note: In other words, it is a plot of $(v_1, \ldots, v_K)$ using bar chart

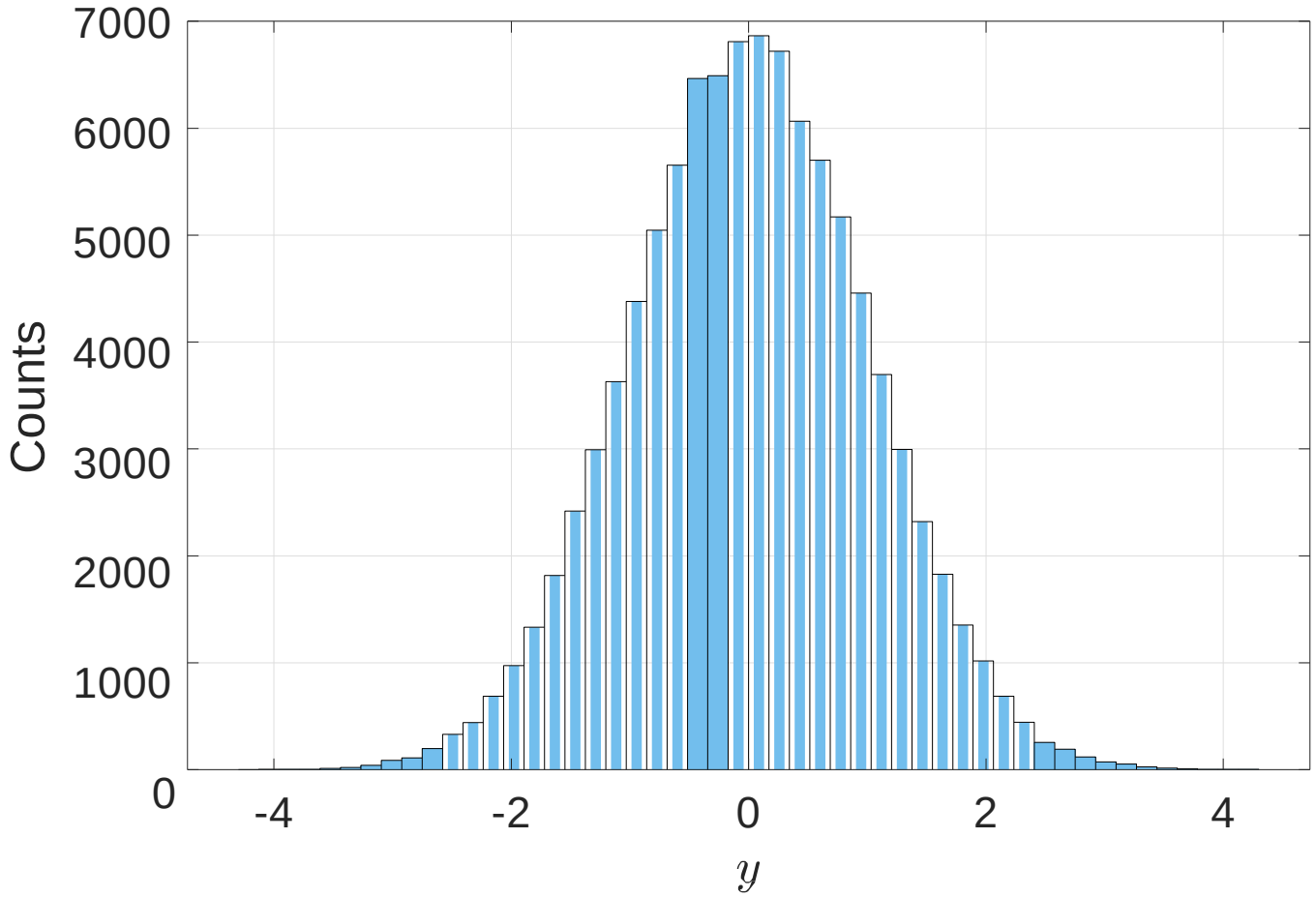
Histogram

K = 20 bins



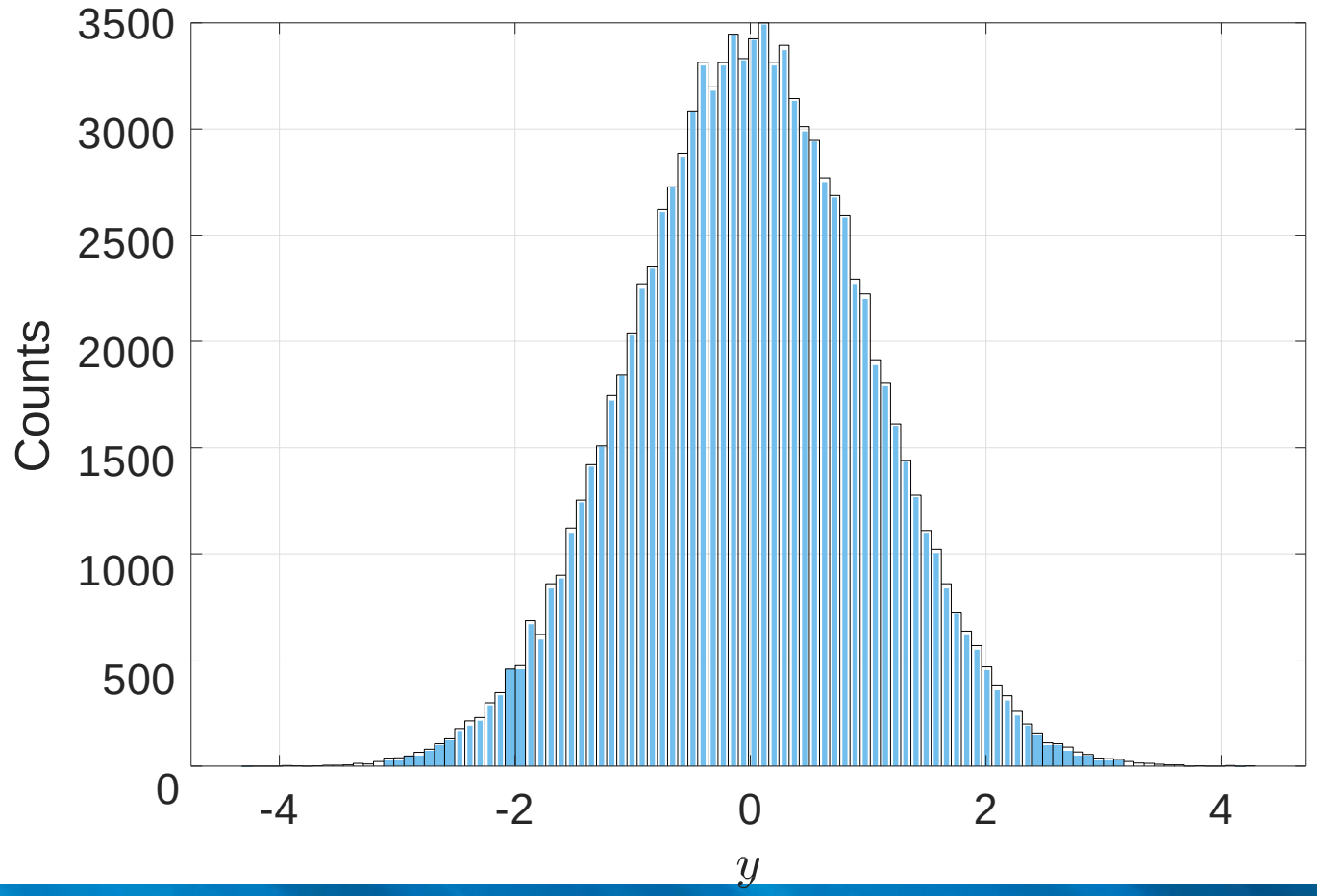
Histogram

K = 50 bins (looks smoother)



Histogram

K = 100 (starting to look ragged)



Motion Charts

Motivation

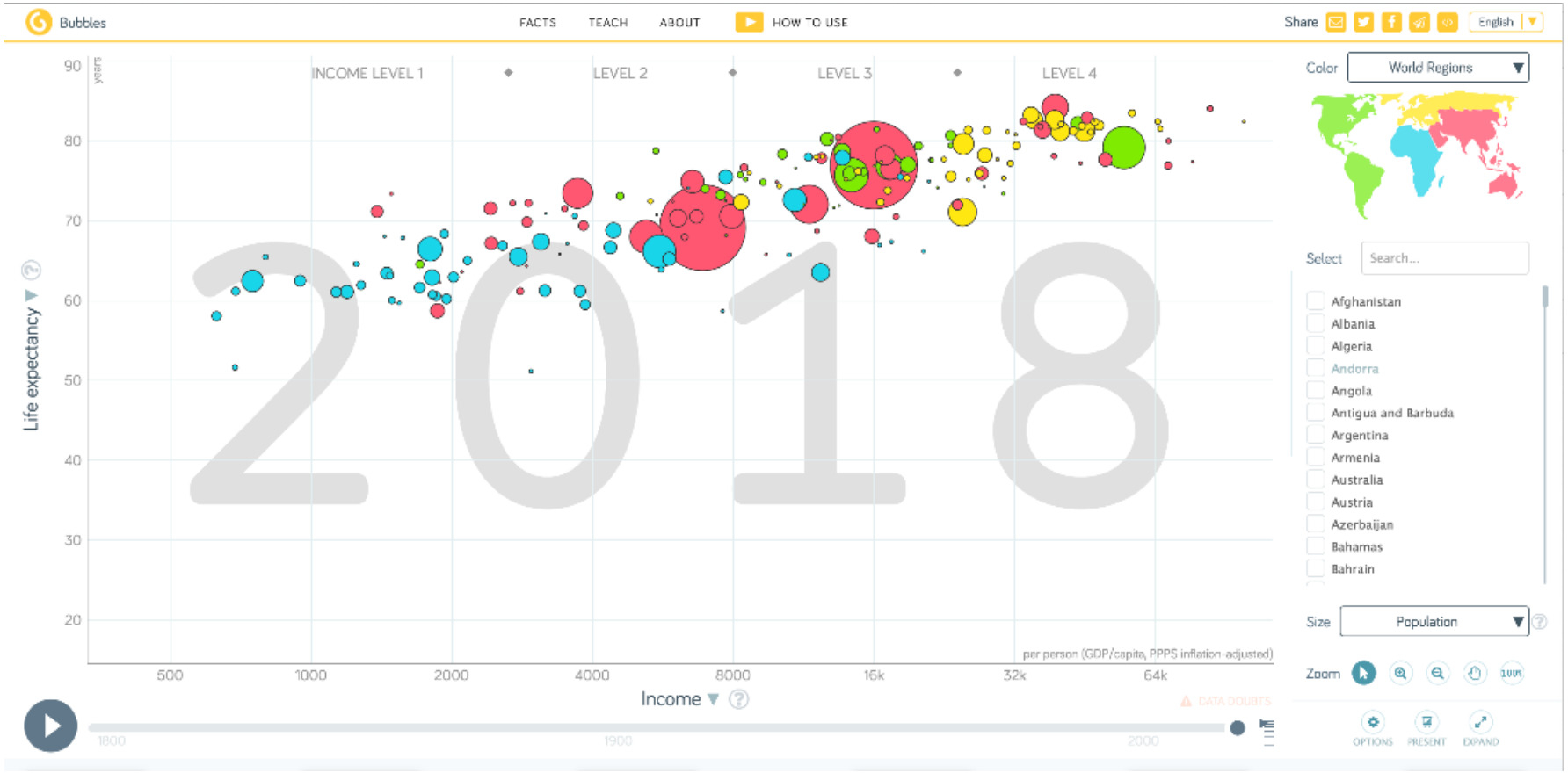
- Motion Charts are interactive multi-dimensional data visualisations
- Originally introduced to the world as GapMinder by Hans Rosling and made famous by his TED talks.

History

- The GapMinder technology was bought by Google and the name of motion charts changed to bubble charts
- But the GapMinder website is now up as a not-for-profit.

Motion Charts

Visualizing data in five dimensions: x-axis, y-axis, size of bubble, colour of bubble, and time.



Motion Charts

Advantages

- Time dimension allows deeper insights and observing trends
- Good for exploratory work
- Appeals to the brain at a more instinctual intuitive level

Disadvantages

- Not suited for static media
- Display can be overwhelming, and controls are complex
- Data scientists who branch into visualization must be aware of the limitations of uses