FIT2086
Modelling for Data Analysis
Semester 2, 2017

Dr. Daniel Schmidt

What is a “model” anyway?

Why should I care?

Public Companies by Capitalization

Rank	Company	Capitalization (US$ million)
1	Apple Inc	749,124
2	Alphabet	628,610
3	Microsoft	528,778
4	Amazon.com	466,471
5	Berkshire Hathaway	418,880
6	Johnson & Johnson	357,310
7	Facebook	357,176
8	Tencent	344,879
9	Exxon Mobil	341,947
10	JPMorgan Chase	323,838

Public Companies by Capitalization

Rank	Company	Capitalization (US$ million)
1	Apple Inc	749,124
2	Alphabet	628,610
3	Microsoft	528,778
4	Amazon.com	466,471
5	Berkshire Hathaway	418,880
6	Johnson & Johnson	357,310
7	Facebook	357,176
8	Tencent	344,879
9	Exxon Mobil	341,947
10	JPMorgan Chase	323,838

Data Science is fun

- Data science lets you take data (numbers, measurements) and *learn* about the process that generated the data

- It lets you make predictions about the future based on the past
  - Will Manchester United beat Real Madrid in the Champions League?

- It lets you quantify empirical evidence of phenomena
  - Do dogs really bite more frequently on the full moon?

Administrivia

- Classes
  - 2 hr lecture 2:00-4:00 Mondays
  - 2 hr lab – as per allocate+

- Outside class
  - Reading, assignments and self-learning.
  - Note, you will be expected to teach yourself R programming

- Text: Ross, S.M. (2014) Introduction to Probability and Statistics for Engineers and Scientists, 5th ed. Academic Press.

Schedule

Week	Topics	Chapters
1	Introduction, modelling, descriptive statistics	1-2
2	Probability and Probability Distributions	3-5
3	Sampling, Parameter Estimation and Bias	6-7
4	Hypothesis testing	8, 10
5	Hypothesis testing	11-12
6	Regression	9
7	Classification	
8	Classification	
9	Model fitting, overfitting, regularization, model evaluation	
10	Unsupervised learning	
11	Simulation, Bootstrap Statistical Methods, Permutation Tests	15
12	Revision	

Staff

- Lecturer
  - Dr. Daniel Schmidt: Daniel.Schmidt@monash.edu
    + Office space: TBA
    + Consultation: Monday 12:00 – 13:00 (tentative)

- Tutor
  - Lachlan O’Neill

Studios

- You must prepare beforehand
  - You will be using R, but we will not be teaching you R programming…
- The basic idea behind the studios is:
  - to get some hands-on experience analysing data
  - to use computational techniques to understand concepts

Assessment

Assessment task	Value	Due date
Assignment 1	10%	Friday of Week 3.
Assignment 2	20%	Friday of Week 7.
Assignment 3	20%	Friday of Week 11.
Examination	50%	To be advised

Marks and Hurdles

- To pass FIT2086 you must obtain:
  - 40% or more in the exam, and
    + 40% or more in the assignments, and
  - an overall unit mark of 50% or more.

- If get less than 40% for either the exam or the assignments, and the total mark is:
  - equal to or greater than 50%, then a mark of 49-N will be recorded.
  - less than 50%. then the actual mark will be recorded.

What this unit is about?

- Technical overview of Data Science
- Exposure to a variety of models and methods for doing data science
- *Some* hands-on experience with data analysis and R programming
- Gaining an understanding of the relation between data and probabilistic models

- NOT learning in depth each model, method introduced
- NOT becoming an R expert

- Realistic goals for students:
  - Familiarization with basics of a few tools
  - Learning advantages and disadvantages of main techniques and models
  - Practice data analysis
  - Exposure to fundamental ideas behind data analytic tools

Assignments

- In R/R Studio
  - Download and install R and R Studio on your own machines as soon as possible. (https://www.rstudio.com/), then familiarize yourself with it

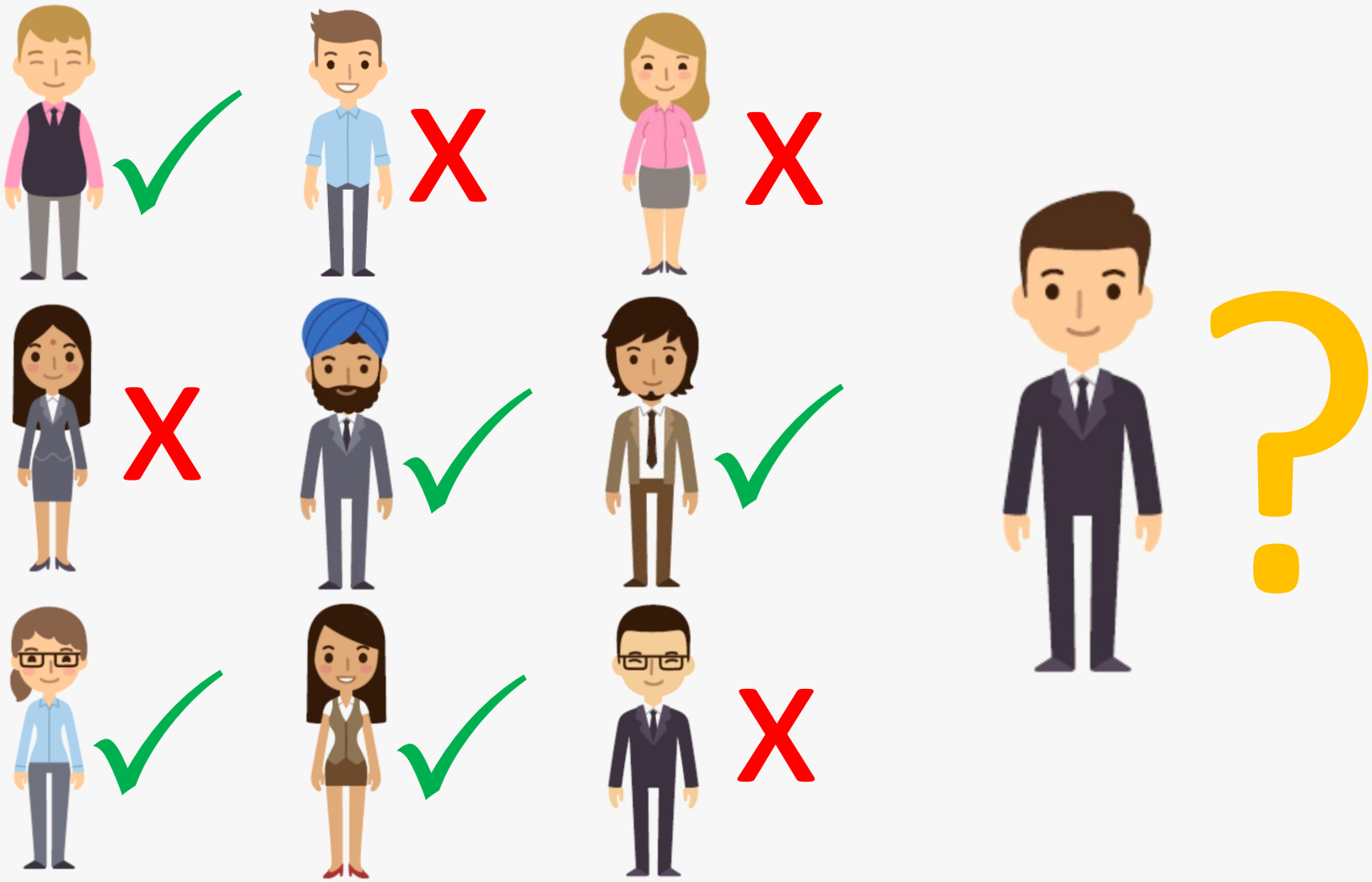
- Each assignment can be expected to flow like this:
  1. Start with a given R or other template file (if one is given) or (if nothing is given) then start from nothing. Hints will be given about the algorithm and output syntax required.
  2. Find (or write) R functions or code that solves some statistical problem. Make sure you comment properly.
  3. Test and debug your code. If you don't know how, **learn** something about testing. Most of these assignments are not expected to be too hard; bad marks will probably at least partly come from a failure to test.
  4. Submit.

Modelling

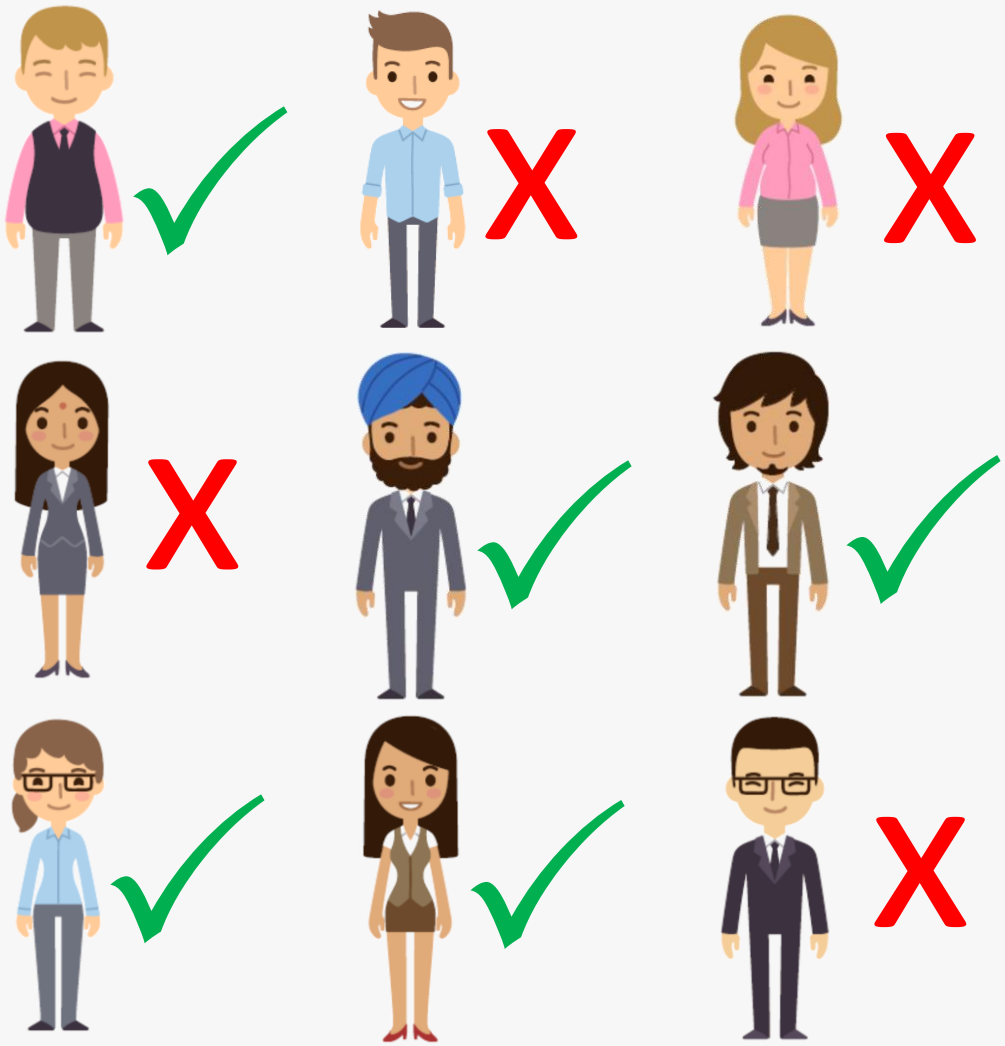
Models

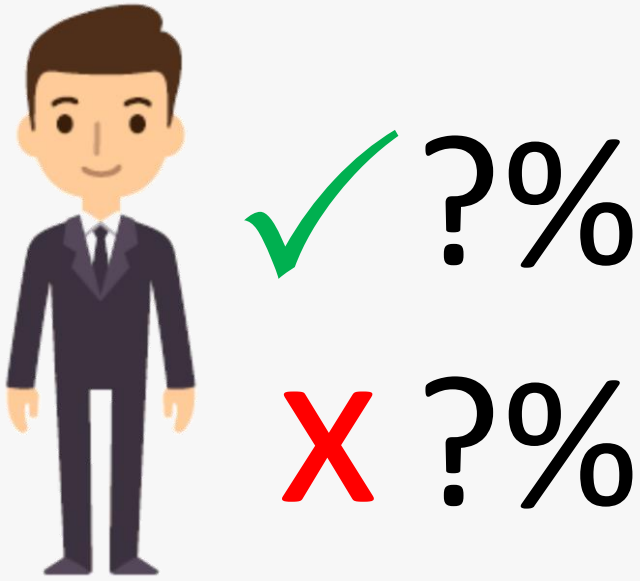
- A model is a thing that represents something else
- Data Science models are mathematical representations
- Models are neither correct or incorrect, they are more or less useful for different purposes
  - one model aircraft might accurately represent the relative dimensions of the wings and body while another might more accurately represent the aerodynamics

Classifiers



Probabilistic classifiers





Scoring (regression)


100

76

84

88

92

82

?

45

95

40

Map anything to anything

knowledge is power

ความรู้คือพลัง

hard work is necessary for success

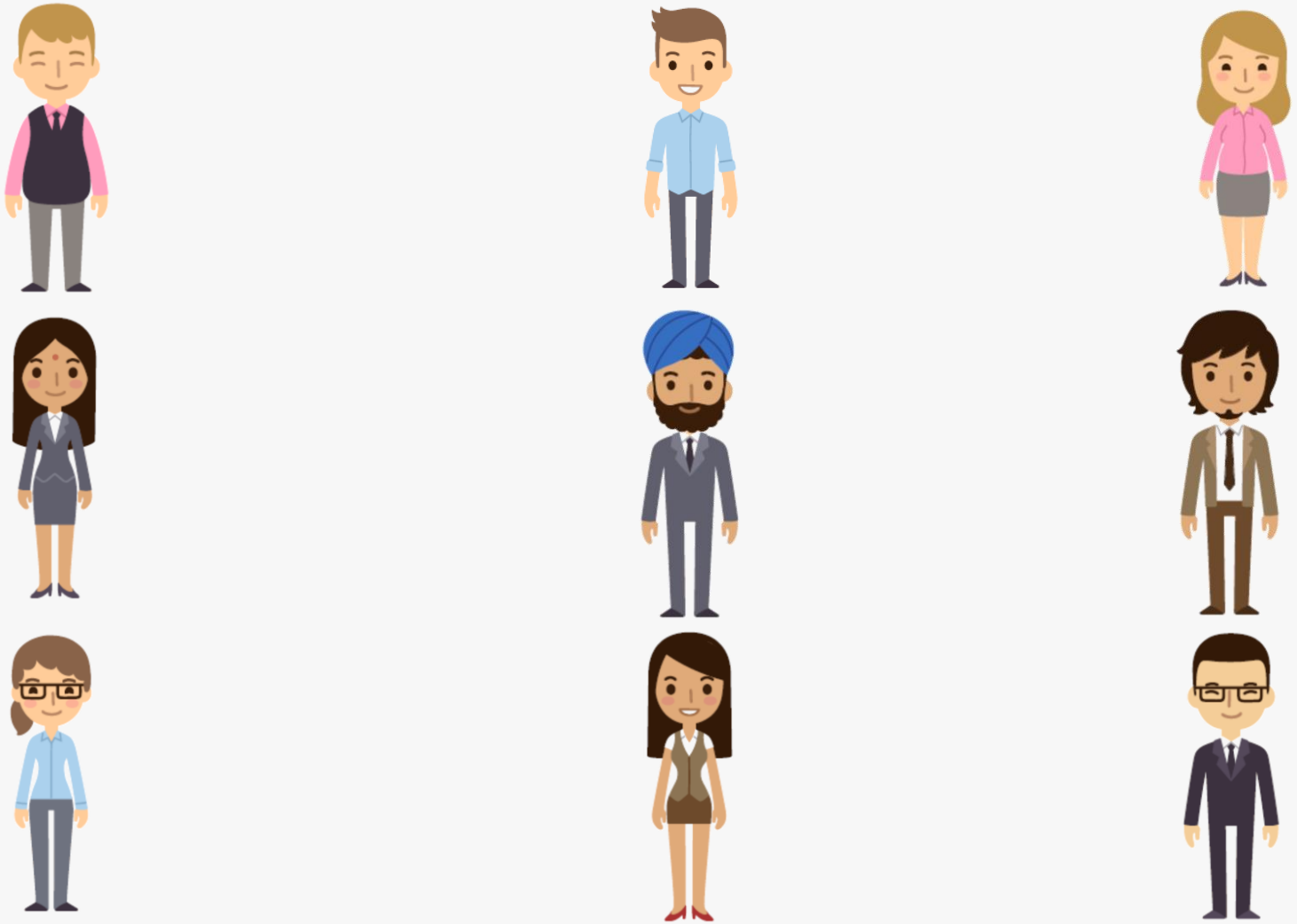
การทำงานอย่างหนักเป็นสิ่งที่จำเป็นสำหรับความสำเร็จ

knowledge is necessary for success

?

Forecasting

Segmenting / clustering



Segmenting / clustering









Segmenting / clustering

Segmenting / clustering

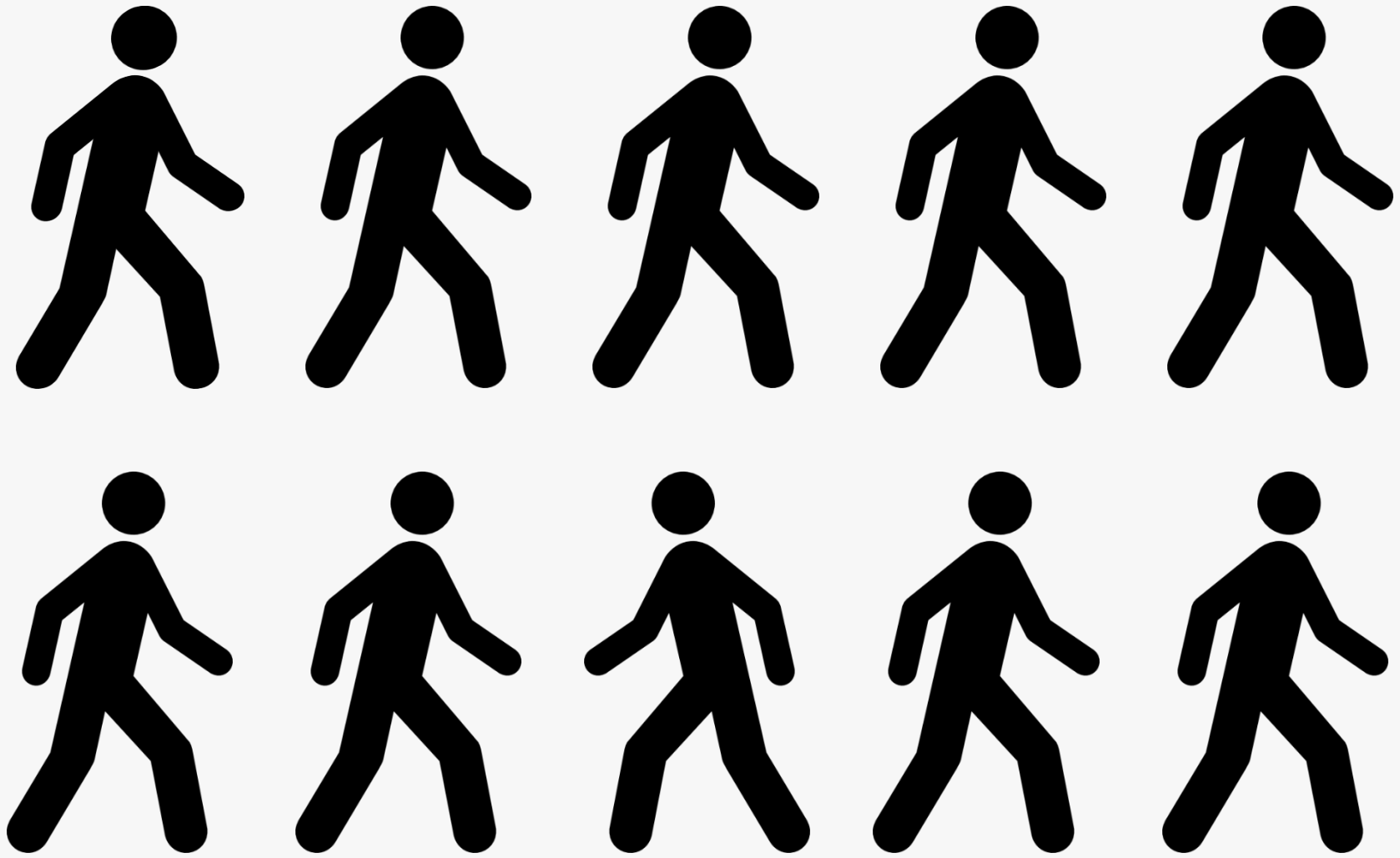




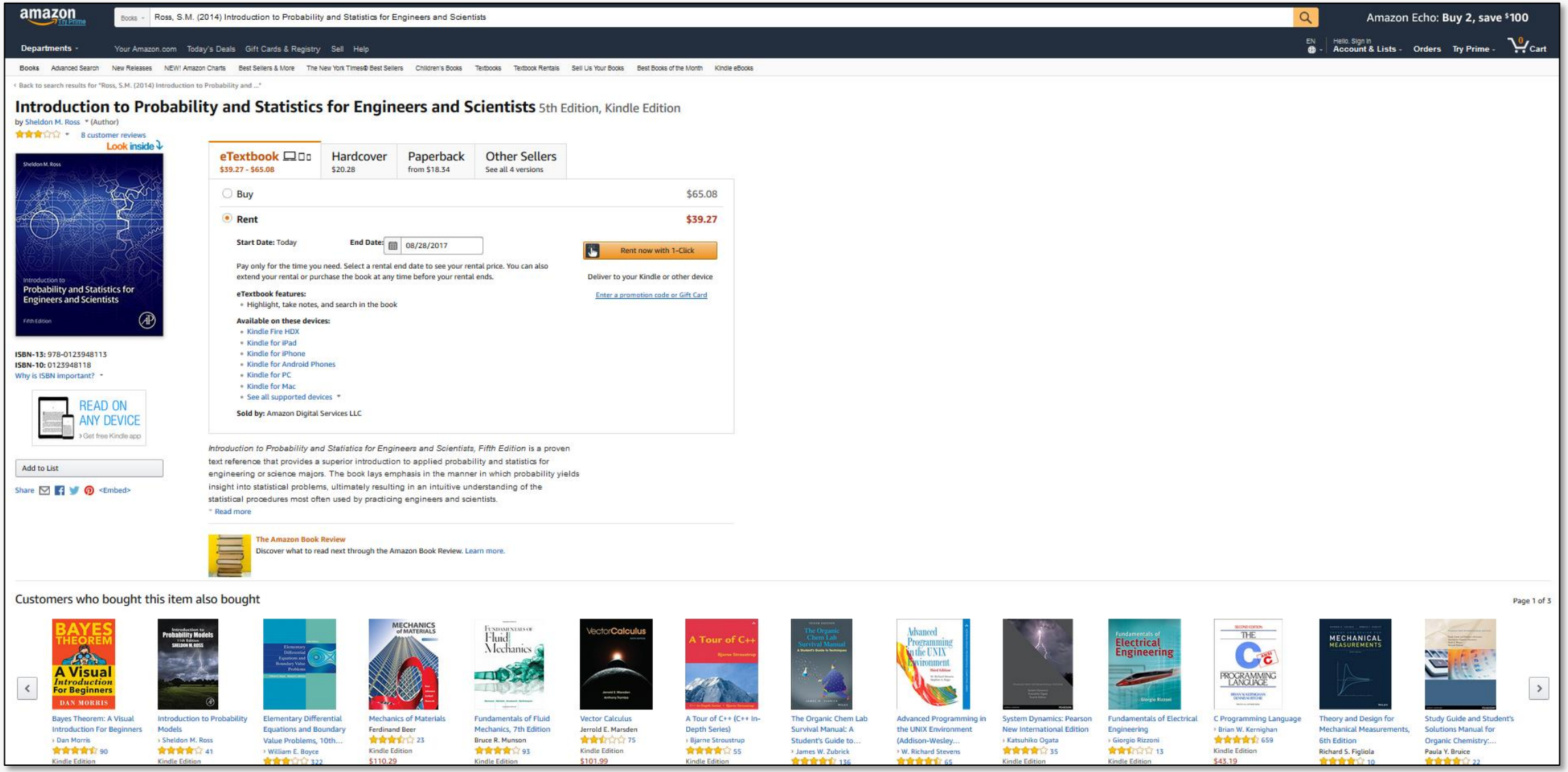




Anomalies detection

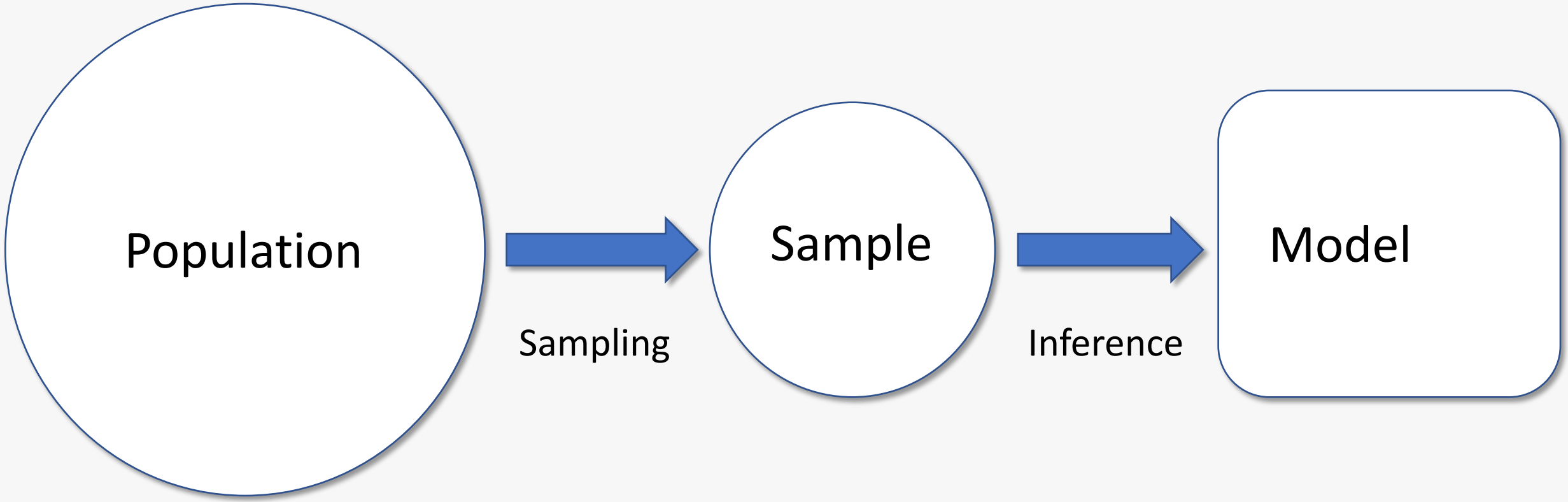


Associating / recommending



Data/samples

From data to models



Some important terms

- Population:
  - A large collection of objects or items with measurable attributes
- Sample:
  - A finite number of recordings of attributes of items from a population
- Model:
  - A mathematical or algorithmic description of the population learned/inferred from the sample

Basic types of data

- *Categorical-Nominal* – discrete number of values, no inherent ordering
  - e.g., Country of birth, sex

- *Categorical-Ordinal* – discrete number of states, but with an ordering
  - e.g., Education status, State of disease progression

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

- *Numeric-Continuous* - Numeric, not enumerable
  - e.g., Weight, Height, Distance from CBD

- *Quantitative* vs *Qualitative*
  - Generally, categorical data is qualitative, numeric data is quantitative

Measurement Scales

Incremental Progress	Uses	Mathematical Operators	Central Tendency	Examples
Nominal	Classification, Membership	=, !=	Mode	Countries
Ordinal	Comparison, Level	>, <	Median	Small, medium, large
Interval	Difference, Affinity	+, -	Mean, Deviation	Dates
Ratio	Magnitude, Amount	*, /	Geometric Mean, Coeff. of Variation	Age

Based on https://en.wikipedia.org/wiki/Level_of_measurement

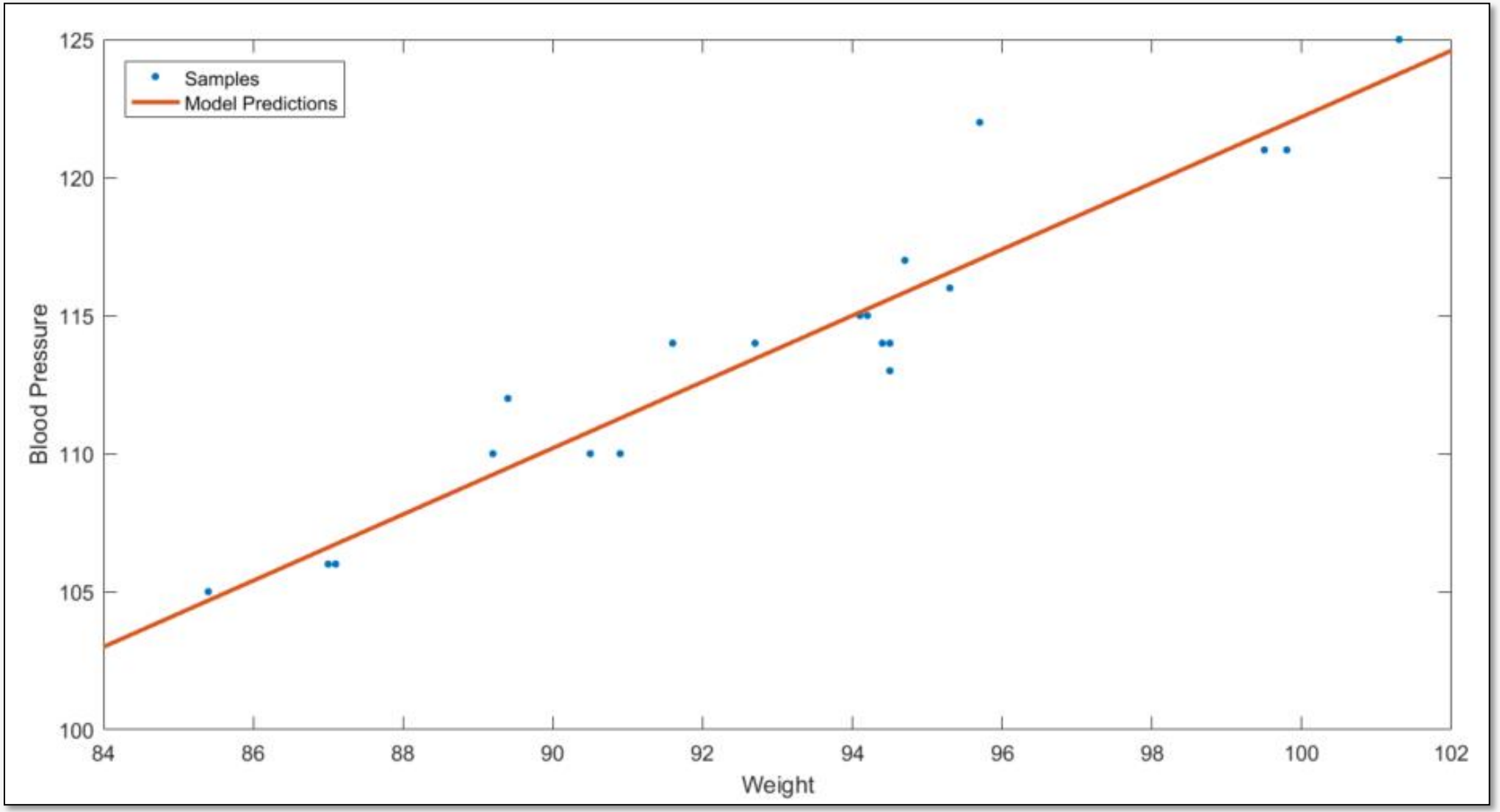
Why do we need formal methods for Data Science?

A Simple Example (1)

Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10.0	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87.0	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	56	95.7	2.09	7.0	75	99

- Knowing weight, can we build a model of blood pressure?

A Simple Example (2)



A Simple Example (3)

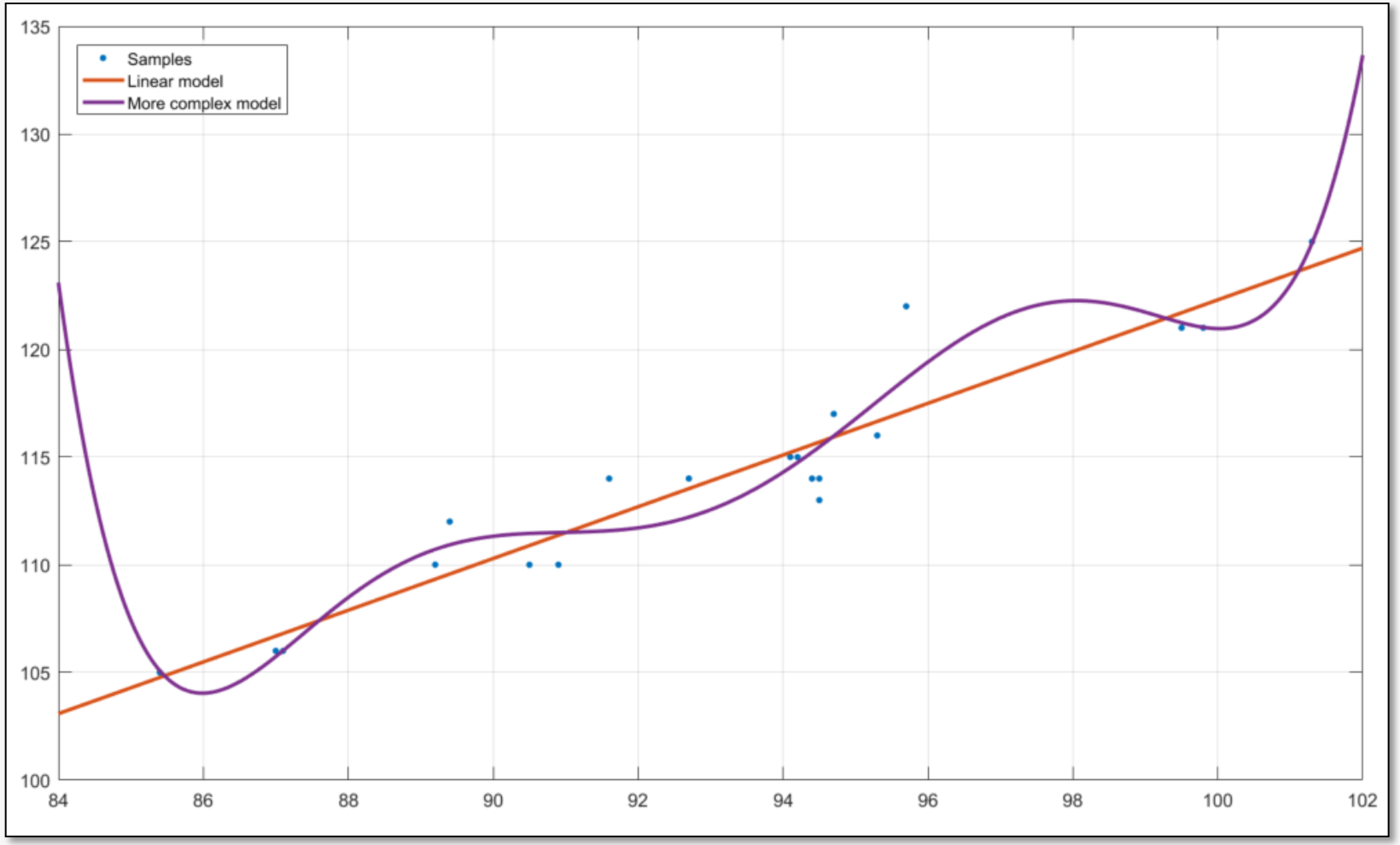
- Formally, our model is:

$$\text{bp} = 1.2 \times \text{weight} + 2.2 + \text{error}$$

- This model relates a person's blood pressure to their weight
  - The relationship is linear
  - The coefficients were *learned* directly from the data

- The "error" term accounts for the discrepancy between the model predictions and the measured data points
  - We handle this error by treating it as a *random, stochastic* quantity

A more complex model



- Fits the *sample* a little better – but is the improvement real?

Data Science techniques let us …

- Find the coefficients of our straight line in an objective fashion
  - "Parameter estimation"
- Answer the question as to which of the two models we looked is the better description of the *population*
  - More complex models fit the sample better, but is it warranted?
- Use many variables simultaneously to find complex relationships
  - Not really possible "by hand"

Descriptive statistics

Descriptive statistics – a refresher

- Summarise aspects of the data
- Usually loses information but gains ease of comprehension
- Contrasts with inferential statistics

- But what is a “statistic?”

Descriptive statistics – a refresher

- Summarise aspects of the data
- Usually loses information but gains ease of comprehension
- Contrasts with inferential statistics

- But what is a “statistic?”
  - Let **X** denote a data sample
  - Then a statistic is any function S(**X**)

- Some functions are more useful than others
  - Describe properties of the data

Measures of centrality

- Let **x** = (x_1, …, x_n) be a sample of *n* data points
- Then the most common measure of centrality, or averageness, is the arithmetic *mean*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The *mode* is the most frequently occurring value in the sample **x**
  - Not particularly useful for continuous data

- Another common measure is the *median,* med(**x**)
  - The value such that 50% of the sample has values less than med(**x**)
  - Easily found by sorting samples and finding the middle sample

Mean vs Median

- The mean uses *all* the values of the sample
  - Any change to any sample changes the mean
  - The mean can be changed as much as desired by changing just one sample by a large enough amount

- The median uses at most two of the values of sample
  - Is very resistant to changes to the samples not in the middle

- Example:
  - **x** = (1,2,3,4,5): $\bar{x}$ = 3, med(**x**) = 3
  - **x** = (1,2,3,4,50): $\bar{x}$ = 12, med(**x**) = 3

- Why would we want to use the mean over the median?

Percentiles

- More general, we can define percentiles
  - The *p*-th percentile is the value, Q(**x**,p) such that *p%* of the values of the sample are lower than Q(**x**,p)

- The median is simply the 50^{th} percentile

- Other important percentiles are the 1^{st} and 3^{rd} *quartiles*
  - i.e., the 25^{th} and 75^{th} percentiles

Measures of spread (1)

- Measures of centrality tell us about the average value of a sample

- Measures of *spread* tell us how much the samples differ, on average, from the average or typical value

- The most straightforward is the range

$$\mathrm{rng}(\mathbf{x}) = \sup\{\mathbf{x}\} - \inf\{\mathbf{x}\}$$

where:

- sup{**x**} denotes the supremum (maximum value)
- inf{**x**} denotes the infimum (minimum value)

Measures of spread (2)

- The most common measure of spread used is the sample standard deviation

$$s(\mathbf{x}) = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\right]^{1/2}$$

- Is the arithmetic mean of the squared deviations from the mean
  - Has the same unit as the data

- Like the mean, is sensitive to changes in the sample

- Often the sample variance

$$v(\mathbf{x}) = s^2(\mathbf{x})$$

is preferred to the standard deviation, as it can be easier to work with

Graphical representations of data

Graphical representations - refresher

- It is often useful to visualise data
  - Can quickly reveal patterns
  - However, going beyond two dimensions is problematic

- For categorical data:
  - Frequency tables
  - Bar graphs
  - Pie charts

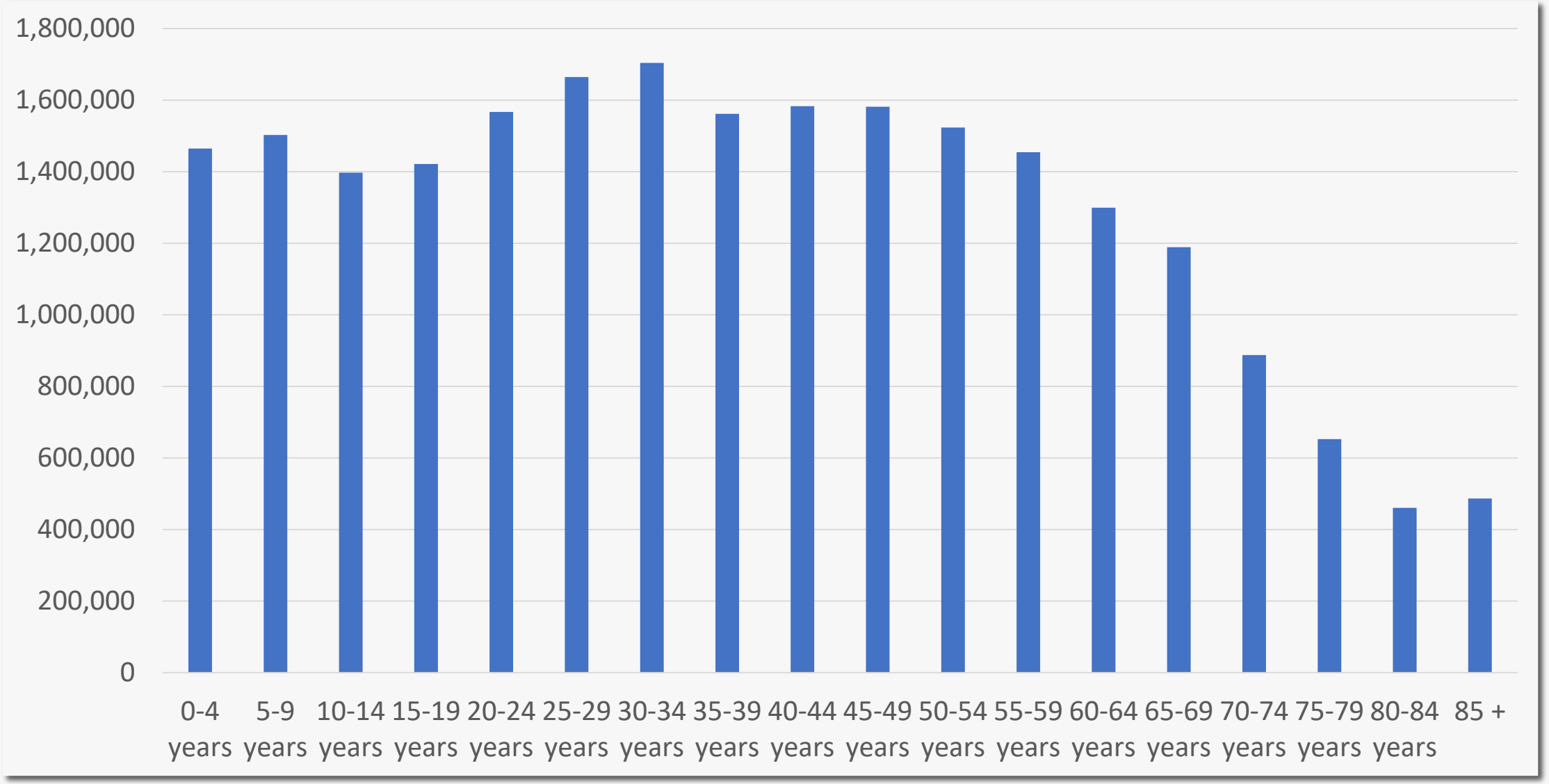
- For numeric data:
  - Histograms
  - Box-and-whisker plots

Frequency Table

Australian population
by age
2016 Census

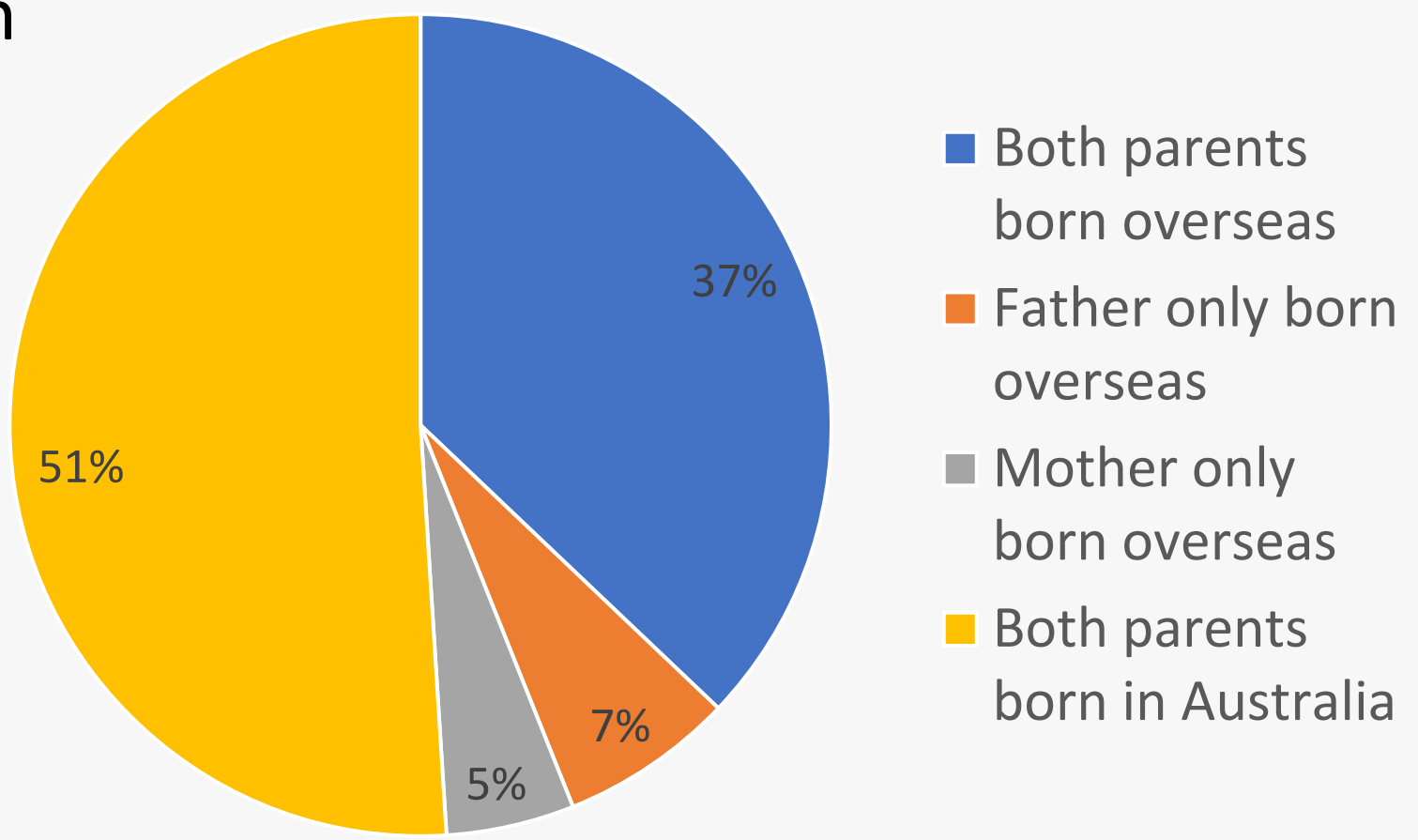
Age	Number
0-4 years	1,464,779
5-9 years	1,502,646
10-14 years	1,397,183
15-19 years	1,421,595
20-24 years	1,566,793
25-29 years	1,664,602
30-34 years	1,703,847
35-39 years	1,561,679
40-44 years	1,583,257
45-49 years	1,581,455
50-54 years	1,523,551
55-59 years	1,454,332
60-64 years	1,299,397
65-69 years	1,188,999
70-74 years	887,716
75-79 years	652,657
80-84 years	460,549
85 years and over	486,842

Bar Graph

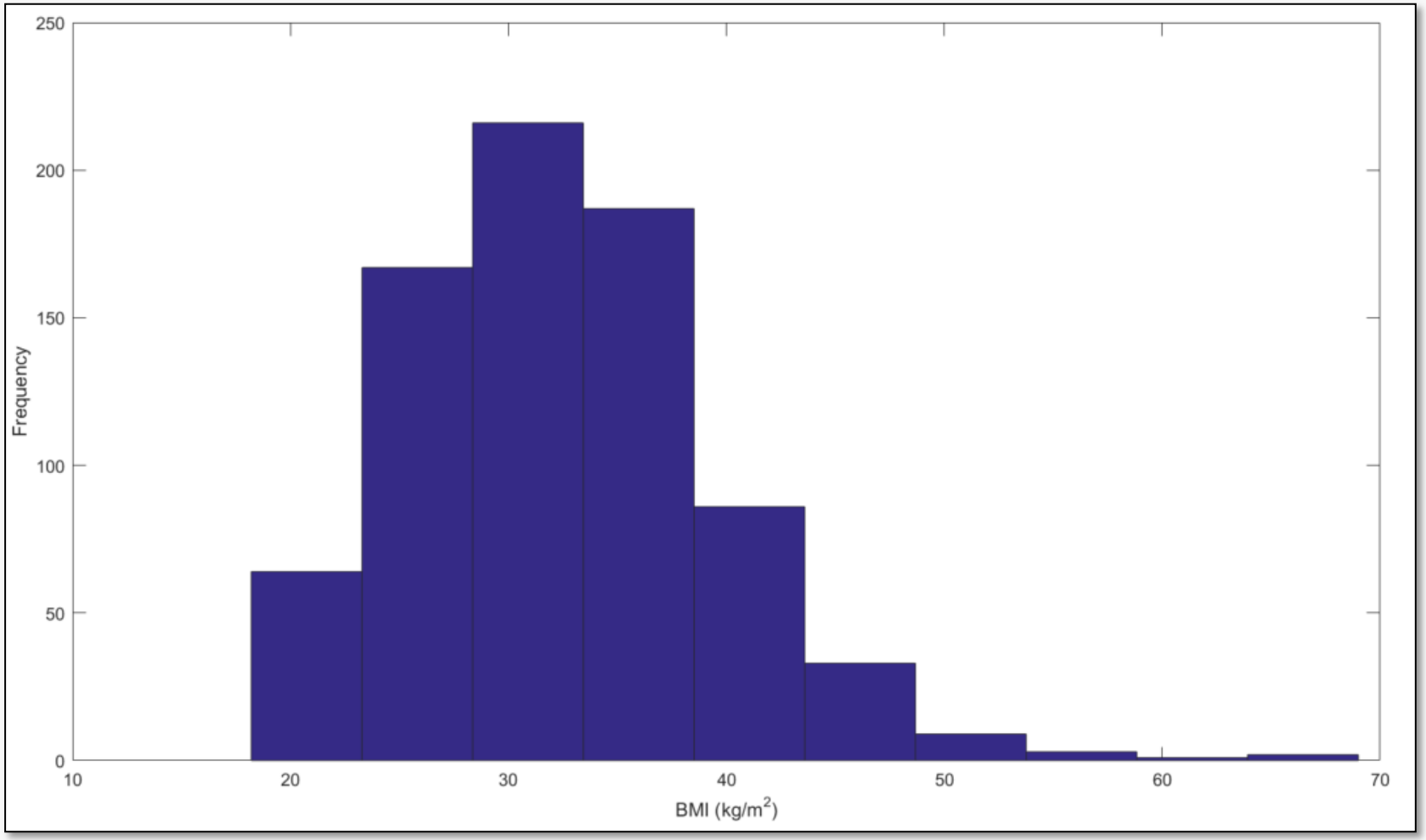

1,800,000
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
0-4 years
5-9 years
10-14 years
15-19 years
20-24 years
25-29 years
30-34 years
35-39 years
40-44 years
45-49 years
50-54 years
55-59 years
60-64 years
65-69 years
70-74 years
75-79 years
80-84 years
85 +

Pie Chart

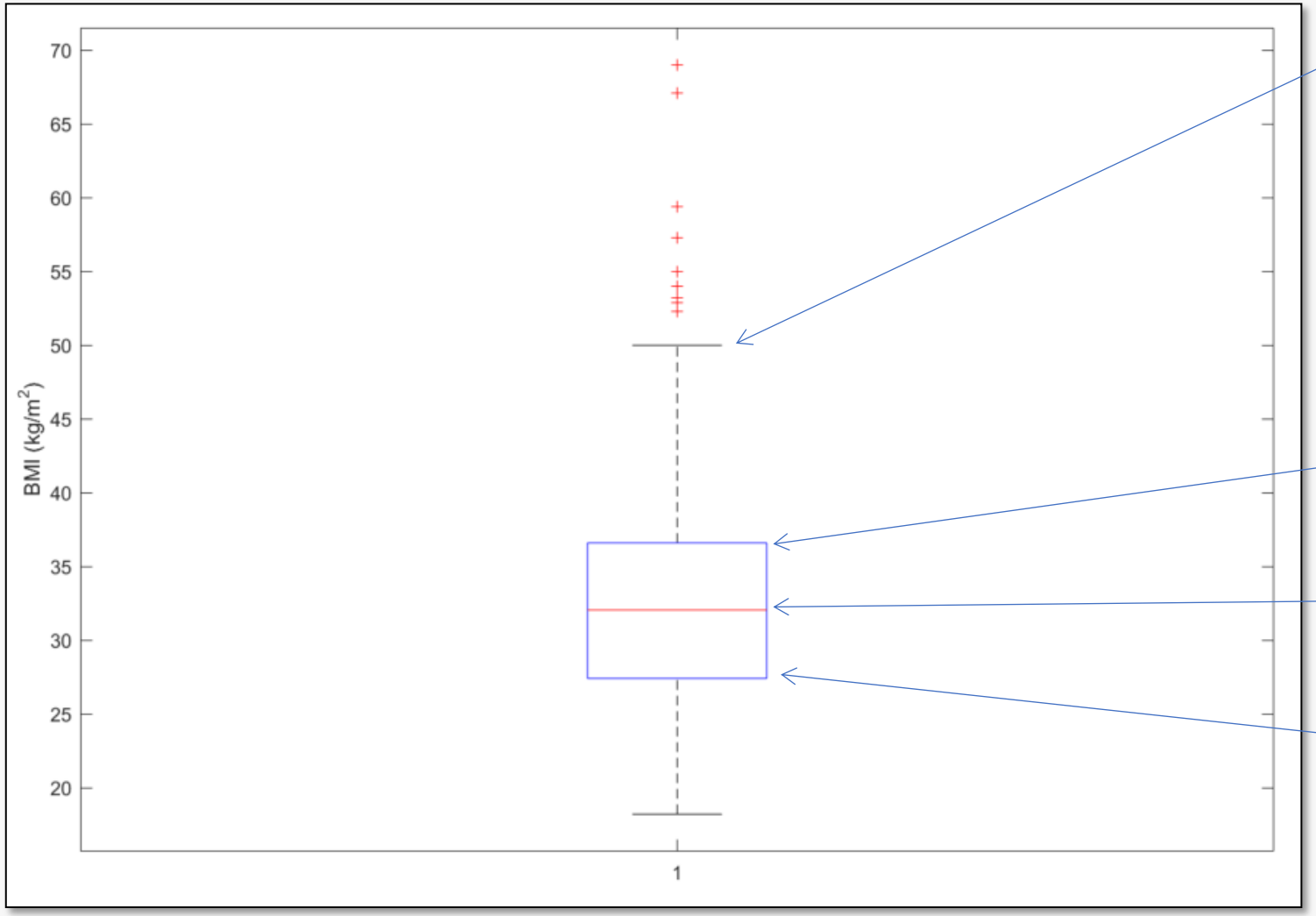
Australian population
by parental country
of birth
2016 Census



Histogram



Box-and-whiskers plot



Scatter plots and correlation coefficients

Correlation coefficient

- Let **x** = (x_1,...,x_n) and **y** = (y_1,...,y_n) be two numeric variables

- Pearson correlation measures association between **x** and **y**

$$R(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n\,s(\mathbf{x})s(\mathbf{y})}$$

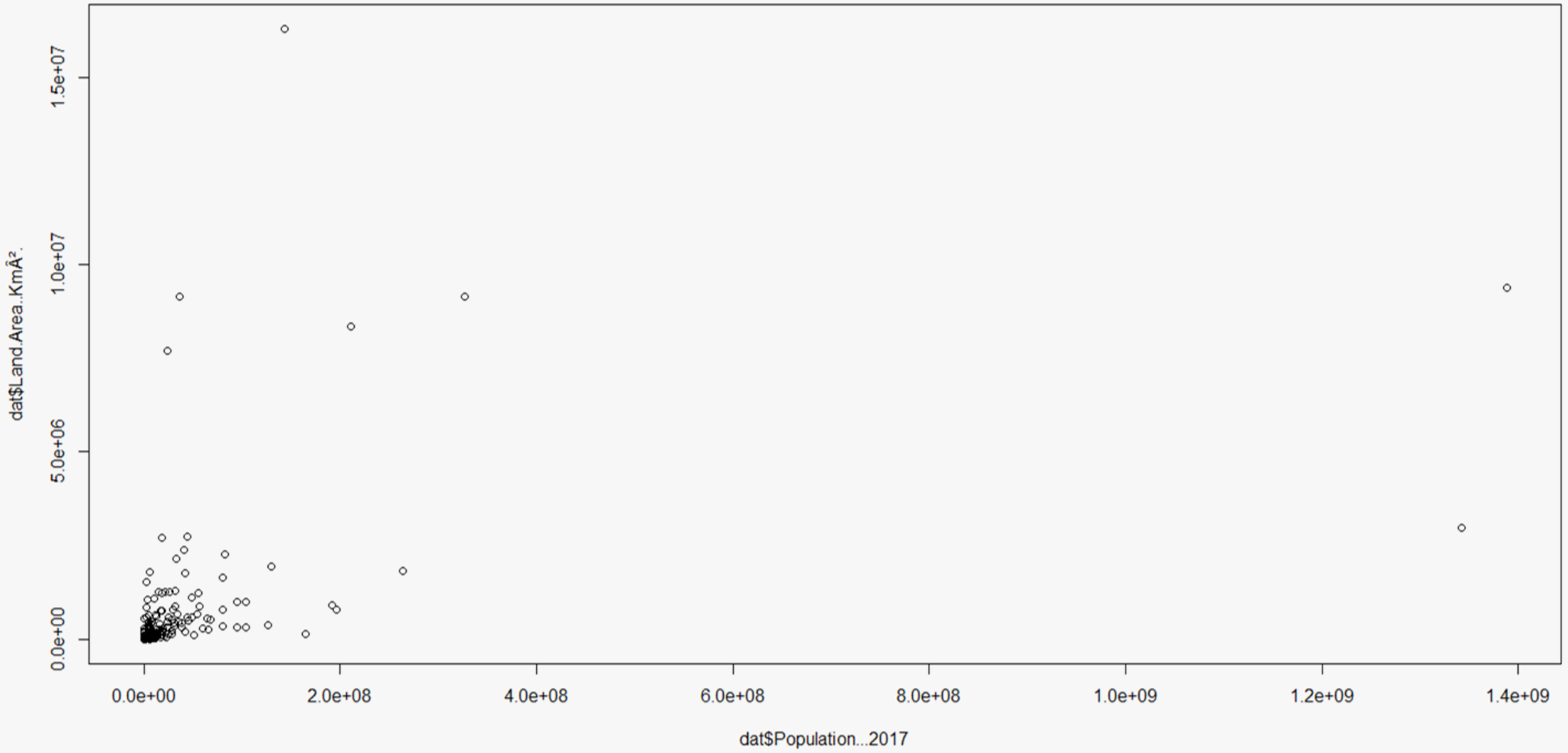
- Pearson correlation measures *linear* association
  - Always lies between -1 and 1, with zero meaning no association
  - Variables can be highly associated but be uncorrelated

- Remember correlation does not imply causation!

Scatter plots (1)

- Scatter plots help visualise relationships between two variables
- Correlation coefficients are statistics that quantitatively measure the strength of association between the two variables
- Three-variable scatter plots, like almost all three-dimensional plots, should be avoided!

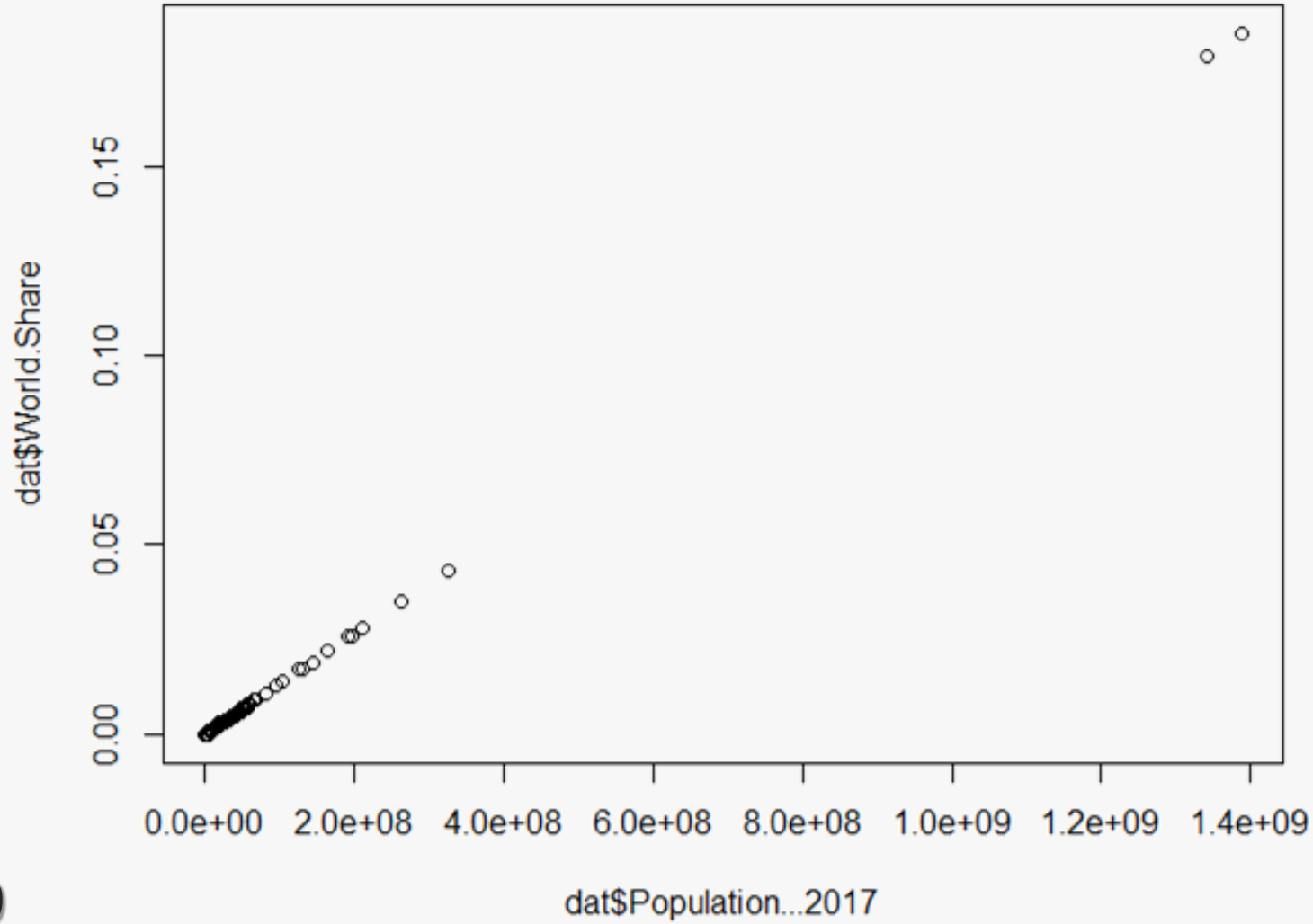
Scatter plots and correlation (1)



- $R = 0.457$

Data from http://www.worldometers.info/world-population/population-by-country/

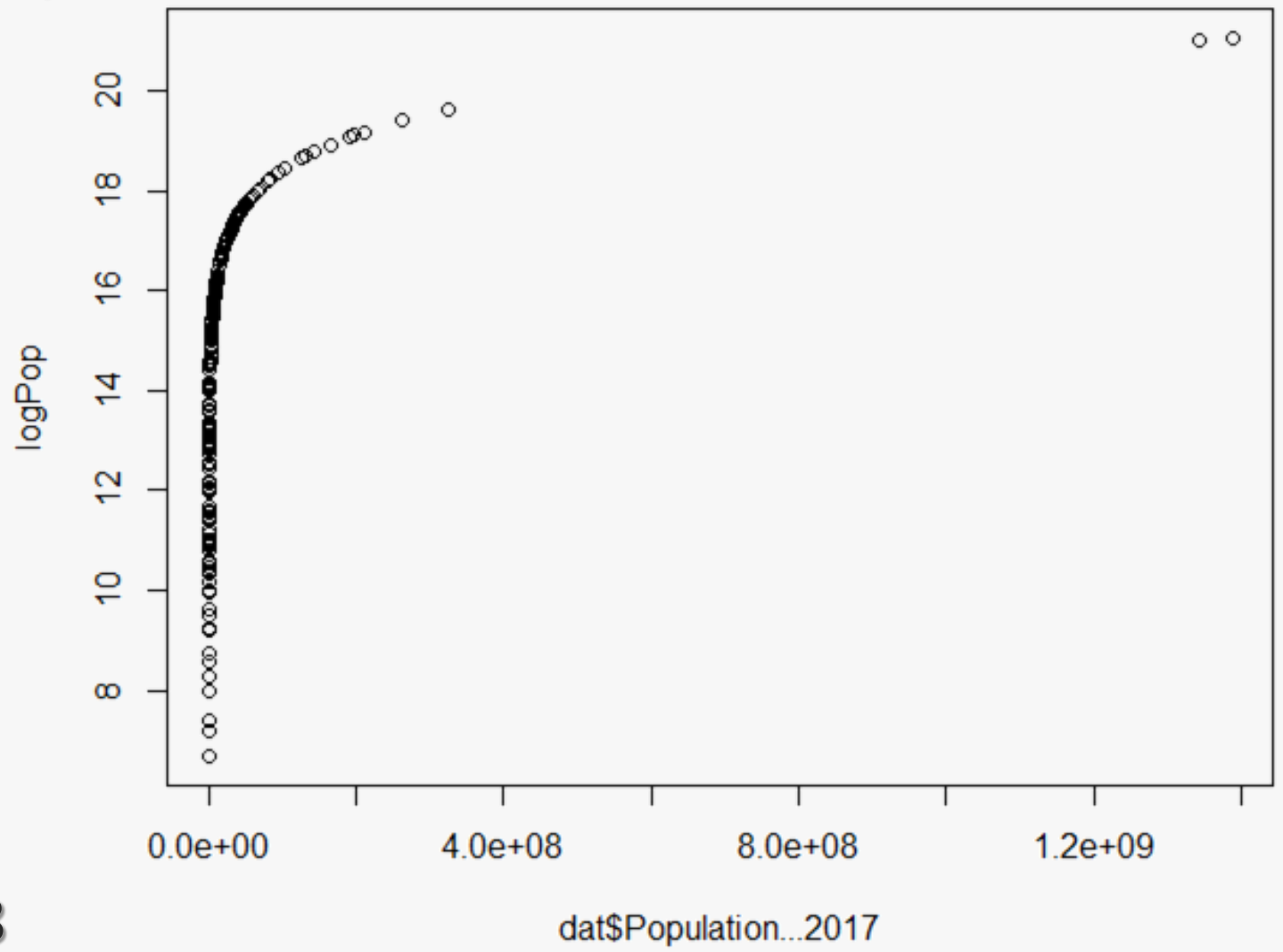
Scatter plots and correlation (2)



- R = 0.999

Data from http://www.worldometers.info/world-population/population-by-country/

Scatter plots and correlation (3)



- R = 0.378

Data from http://www.worldometers.info/world-population/population-by-country/

Terms you need to revise

- Population
- Sample
- Nominal, ordinal, interval and ratio scales
- Discrete vs continuous data
- Quantitative vs qualitative data

Terms you need to learn from Ross Chapter 2

- Histogram
- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of dispersion
  - Standard deviation
  - Variance
  - Range
- Quantiles (percentiles and quartiles)
- Scatter plot
- Correlation coefficient