COMP4030

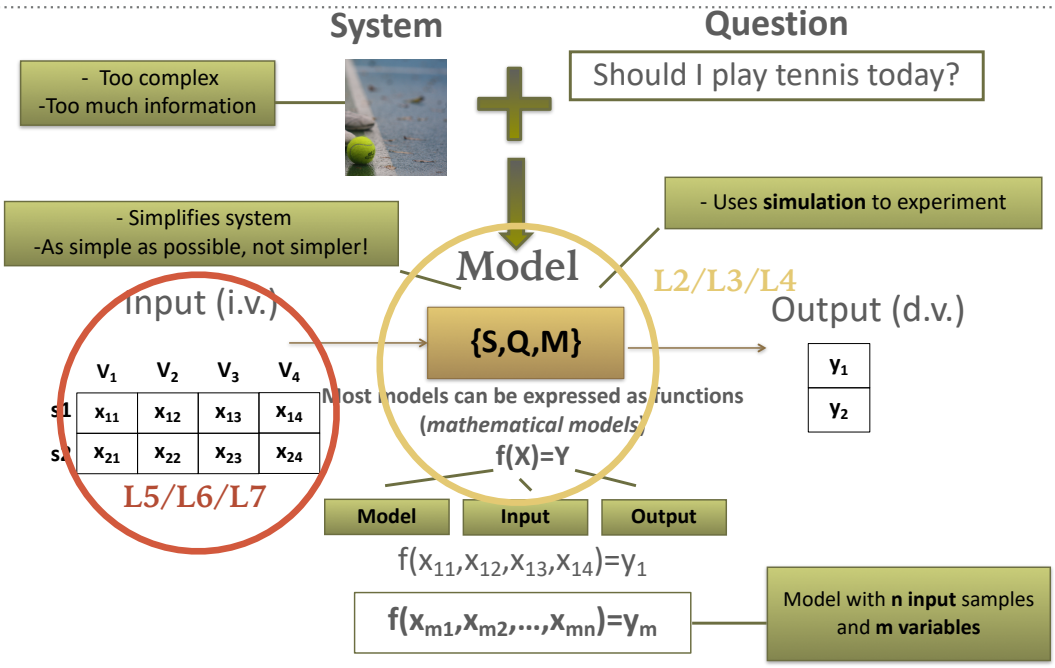
DATA MODELLING AND ANALYSIS

Lecture 5: Data Analysis

LECTURE OUTLINE

1. Summary
2. Lecture Outcomes
3. Modelling Data
4. Data vs knowledge
5. Describing data
    1. Types of data
    2. Statistics for: location, shape, dispersion, association
6. Modelling data
    1. Data overload
    2. Main challenges
    3. Data modelling pipeline

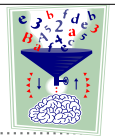
DMA: THE BIG PICTURE



LECTURE OUTCOMES

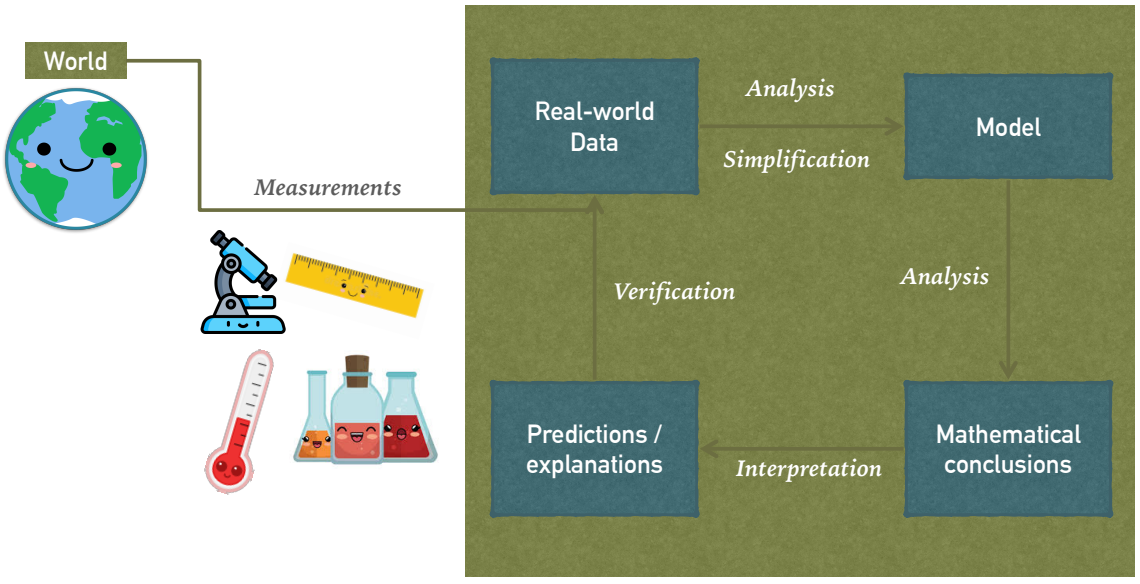
- At the end of this lecture, you should be able to answer these questions:
  - What is modelling data?
  - How is it done?
  - What is the difference between **data** and **knowledge**?
  - What are the main issues found when modelling data?
  - What are the types of data?
  - Which statistics are used to describe data?

MODELLING DATA

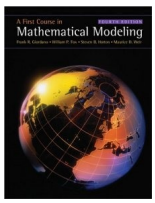


- This module: **Data** Modelling and Analysis
  - Why are we talking about data?
  - Data: motivation behind modelling and analysis
- **Analysis**: organise and understand
- **Modelling**: describe and predict
- **Optimisation**: build & improve models
- Mathematical Models allow for description and prediction of data:
  - Description
    - e.g.: Summarisation, Clustering, Visualisation
  - Prediction (generalise to new cases)
    - e.g.: Classification, Regression, Forecasting

MODELLING DATA: DATA IN RELATION TO THE WORLD



Adapted from:



MODELLING DATA: DATA OVERLOAD

- Information/data overload
- Data is being generated at an ever increasing rate
- Rough estimate:
  - Data volume in the world doubles every 20 months
- What does the data mean?
- What can we do with the data?
- Our ability to analyse data (e.g. through computational tools) is developing slower than our ability to generate and store data

MODELLING DATA: DATA OVERLOAD

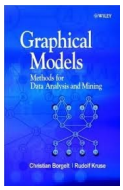
- Data overload comes in many forms
- Many sources of data with many applications, e.g:
  - Images → classification of satellite imagery
  - Sounds → music recognition, speech recognition
  - Text → automated translation
  - DNA → gene identification
  - Demographics → marketing decision support
  - Commercial transactions → inventory optimization
  - Weather → weather forecasting
  - etc.

MODELLING DATA: DATA VS KNOWLEDGE

- For data to be useful, we need to extract **meaning**
  - Data → knowledge
- Differences between data and knowledge

Data	Knowledge
Refers to instances (e.g. objects and events)	Refers to classes (e.g. sets of objects and events)
Describes individual properties	Describes general patterns, principles, etc.
Usually in huge amounts	Aim: to extract a minimal set of statements
Usually easy to obtain	Usually hard to obtain
Can't be used directly for generating predictions	Can be used directly for generating predictions

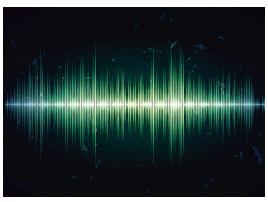
Adapted from

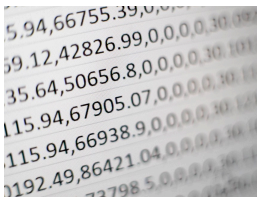


BUT…WHAT IS DATA?

- When we talk about data, what are we talking about?
- Factual information used as a basis for reasoning, discussion, or calculation
  - Includes measurements or statistics
- Information in digital form that can be transmitted or processed

NOT THIS





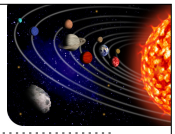
BUT THIS!

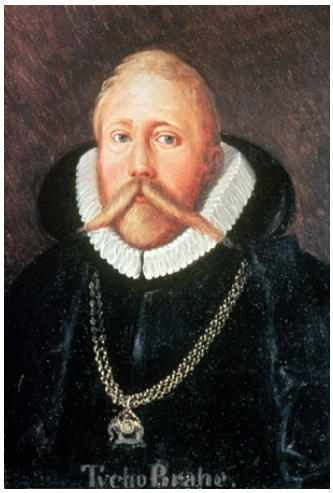
DESCRIBING DATA: DATA REPRESENTATION

- Different formats
- Common representation: tables
  - Columns: attributes, features
    - Measurable characteristics
    - Last column: Output, label, class
  - rows (tuples): data objects, instances
    - case, point, record, sample, entity
  - Input(s) / output(s)

	Outlook	Temp.	Humid.	Windy	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

MODELLING DATA: DATA VS KNOWLEDGE – AN EXAMPLE





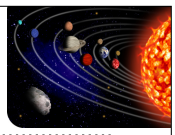
Tycho Brahe

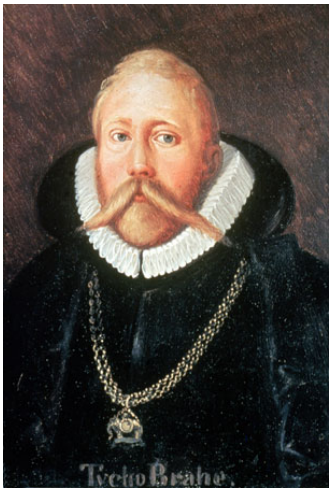
(1546 – 1601)

Johannes Kepler

(1571-1630)

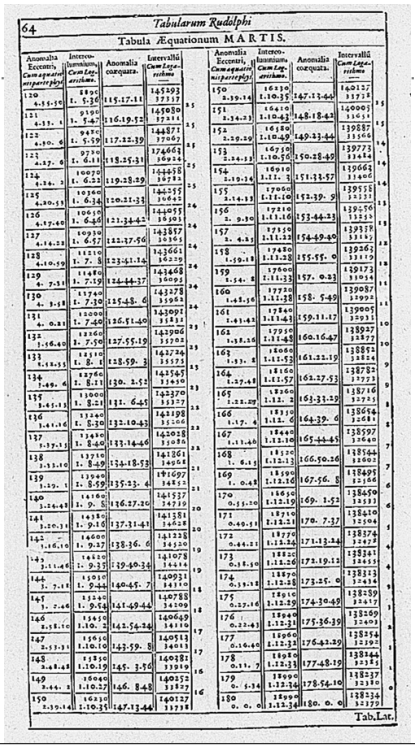
MODELLING DATA: DATA VS KNOWLEDGE





Tycho Brahe

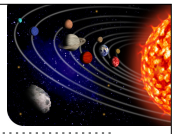
(1546 – 1601)


64 Tabularum Rudolphi

Tabula Æquationum MARTIS.

Tab. Lat.

MODELLING DATA: DATA VS KNOWLEDGE



KEPLER'S LAWS OF PLANETARY MOTION

1. *The orbit of a planet is an ellipse with the Sun at one of the two foci.*
2. *A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.*
3. *The square of the orbital period of a planet is proportional to the cube of the semi-major axis of its orbit.*

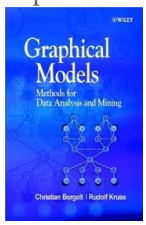
Johannes Kepler

(1571-1630)

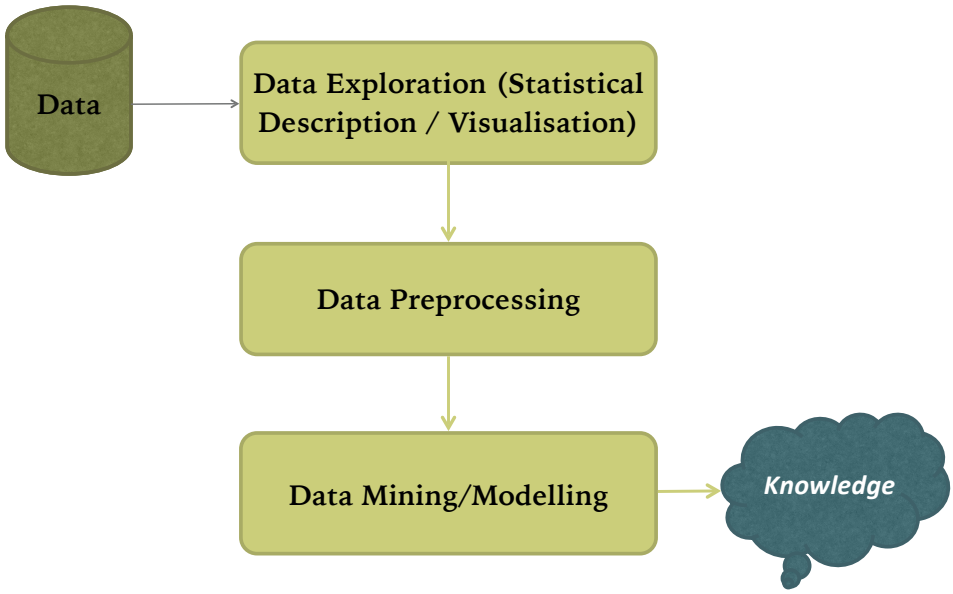
MODELLING DATA: DATA VS KNOWLEDGE – ASSESSMENT

- Not all knowledge is equal: how do we assess it?
- Assessment Criteria
  - Correctness (probability, success in tests)
  - Generality (range of validity, conditions of validity)
  - Usefulness (relevance, predictive power)
  - Comprehensibility (simplicity, clarity, parsimony)
  - Novelty (previously unknown, unexpected)
- Priority
  - Science: correctness, generality, comprehensibility
  - Economy: usefulness, comprehensibility, novelty

Adapted from



MODELLING DATA: DATA MODELLING PIPELINE



DESCRIBING DATA TYPES OF DATA: CATEGORICAL (DISCRETE) DATA

A. Nominal

- Data characterised by names, labels or categories
- There is no order
  - Blood type (A, B, AB, O)
  - Gender
  - Country of origin
  - Attributes with yes/no values
- Data is distinct
- valid statistics: mode

DESCRIBING DATA TYPES OF DATA: CATEGORICAL (DISCRETE) DATA

B. Ordinal

- Data can be arranged in some **order** but differences between values are meaningless.
- It may be linguistic, but also numeric!
  - Product rating: poor, good, excellent.
  - Gymnastic scores.
  - Satisfaction level (very unsatisfied, somewhat unsatisfied, neutral, somewhat satisfied, very satisfied).
- Distinctness and order
- valid statistics: median, percentile

DESCRIBING DATA TYPES OF DATA: NUMERIC (METRIC) DATA

C. Interval

- Data are ordered and differences are meaningful.
- There is no true 0.
- Multiplication and division and are meaningless, so ratios have no meaing
  - Temperature: 50° – 40° = 60° – 50°, but 100° is not 2 * 50°.

- Distinctness, order and addition (subtraction)
- valid statistics: mean, st. dev., correlation, regression

DESCRIBING DATA TYPES OF DATA: NUMERIC (METRIC) DATA

D. Ratio

- data is ordered and all arithmetic operators are applicable; ratios are meaningful
  - e.g. litres of petrol consumed per day per individual
- distinctness, order, addition and multiplication (division)
- valid statistics: all

DESCRIBING DATA: EXERCISE: DATA CATEGORISATION

1. IQ scores? (100 = 'average'; 140 = 'genius')
2. Sport team squad numbers? (1 to 22)
3. Golf score?
4. Age?
5. Date of birth?
6. Political party (Labour, Lib-Dem, Conservative)?
7. Exam grades (A, B, C, D, E, F)?
8. pH (acidity = -log[H^+])?
9. Number of children

DESCRIBING DATA: EXERCISE: DATA CATEGORISATION

10. Income?
11. Eye colour?
12. Degrees in Farenheit?
13. Degrees in Kelvin?
14. Number of sales?
15. Level of spiciness?
16. Time?
17. Constellations
18. Zodiac sign

DATA ANALYSIS: DESCRIBING DATA – LOCATION

A. Mode (nominal)

- Attribute value that is most frequent in the sample
- Not unique
  - more than one value may have the largest frequency
- The most general localisation measure
  - applies to all scales of measurement

DATA ANALYSIS: DESCRIBING DATA – LOCATION

B. Median (Ordinal):

- If the data elements $x_1, x_2, \ldots x_n$ are sorted, the median is:

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if n is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if n is even} \end{cases}$$

C. Arithmetic mean (interval, ratio):

- Only applicable to metric attributes

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

DATA ANALYSIS: DESCRIBING DATA – DISPERSION

D. Range:

A. difference between the maximum and minimum value

E. Inter-quantile range:

- *p*-quantile: the value relative to which the fraction *p* of the data is smaller
- The (1/2)-quantile value is the median
- the *p*-inter-quantile range (IQR) is the difference between the (1-*p*)-quantile value and the *p*-quantile value
  - note: $0 < p < ½$
- It is common to use $p = ¼$, *inter-quartile* range
  - In which case: IQR = Q3 - Q1

DATA ANALYSIS: DESCRIBING DATA – LOCATION

F. Average absolute deviation

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

G. Variance

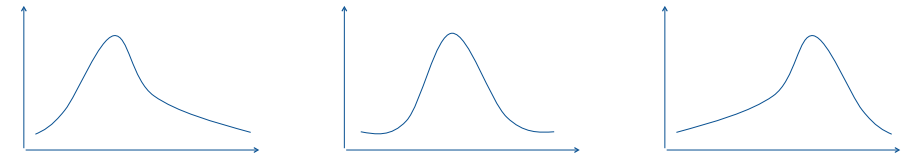
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

H. Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

DATA ANALYSIS: DESCRIBING DATA – SHAPE

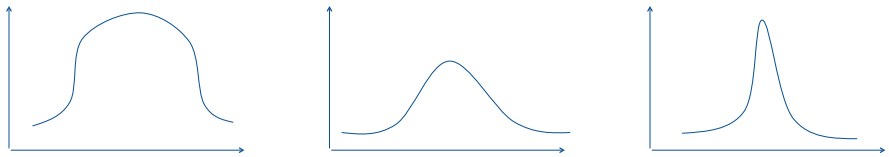
- Skewness:
  - Measures the asymmetry of probability distribution of a real-valued random variable about its mean.
  - It can be positive, negative and even undefined.



Positive (right) skew *Symmetric* *Negative (left) skew*

DATA ANALYSIS: DESCRIBING DATA – SHAPE

- Kurtosis:
  - Measures how peaked a distribution is
  - Measure is usually computed relative to the Gaussian



Platikurtic *Gaussian* *Leptokurtic*

DATA ANALYSIS: DESCRIBING DATA – RELATIONSHIPS

- Measures a statistical relationship between two variables.
- **Pearson coefficient:** measures strength and direction of the linear relationship between two variables.
  - Ranges between $[-1,+1]$

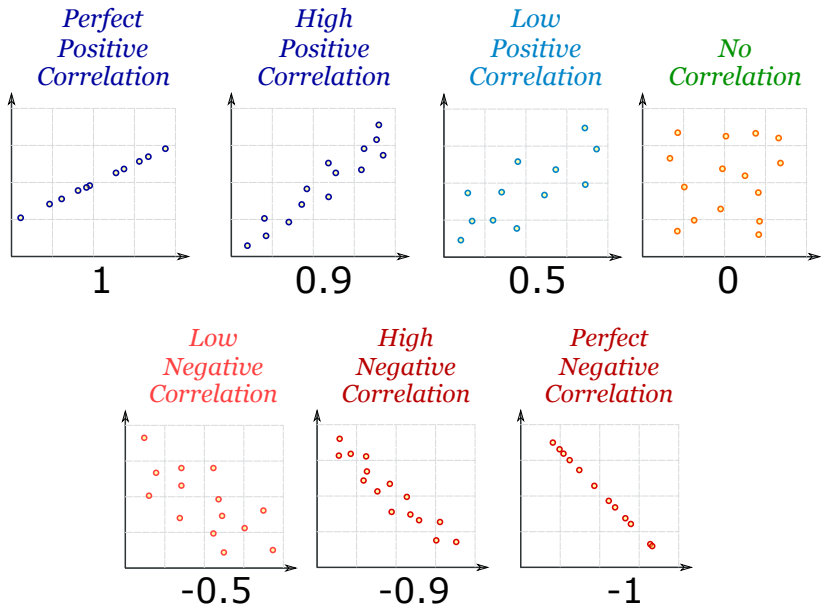
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

n is the sample size

x_i, y_i are the single samples indexed with i

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of X

DATA ANALYSIS: DESCRIBING DATA – RELATIONSHIPS



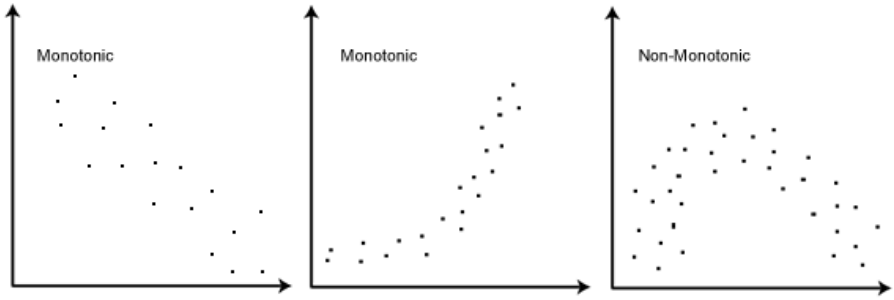
DATA ANALYSIS: DESCRIBING DATA – RELATIONSHIPS

- The Pearson coefficient is not applicable to ordinal data.
- **Spearman coefficient:** non-parametric rank-based correlation coefficient
- It measures monotonic relationship
- Values range between [-1,1]

$$rho = \frac{\sum (x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum (x' - m_{x'})^2 \sum (y' - m_{y'})^2}}$$

- where m *is the mean,* $x' = rank(x)$, and $y' = rank(y)$

DATA ANALYSIS: DESCRIBING DATA – RELATIONSHIPS



MODELLING DATA: MAIN CHALLENGES

- In this lecture, we will cover three challenges [1]:
  - Too much data
  - Too little data
  - Fractured data
- Analysis will help us understand which of these challenges, if any, are present in our problems.
  - Lecture 6 and 7 will delve in how to "fix" them, of possible.

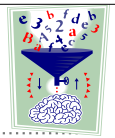
A. Famili, W.M. Shen, R. Weber, E. Simoudis, Data preprocessing and intelligent data analysis, Intell. Data Anal. 1 (1) (1997) 1–28.

MODELLING DATA: MAIN CHALLENGES

- Too much data
  - Corrupt and noisy data
  - Irrelevant data
  - Very large data-sizes
  - Very large number of attributes
  - Different types of data

A. Famili, W.M. Shen, R. Weber, E. Simoudis, Data preprocessing and intelligent data analysis, Intell. Data Anal. 1 (1) (1997) 1–28.

MODELLING DATA: TOO MUCH DATA

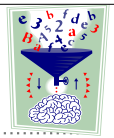


- Some domains (e.g. satellite imagery, medical imagery, telecommunications) have very large:
  - Volumes of data
  - Rates of data production
- Real-time data analysis: very difficult if not impossible (depending on hardware available)
- Example solutions:
  - Sampling
  - Scaling down the resolution of the data (effectively attribute and instance selection of some kind)
  - Dimensionality reduction (for attributes)
  - High-performance computing, etc.

MODELLING DATA: MAIN CHALLENGES

- Too little data
- Depending on the problem, a much more serious issue than having too much data
- Issues include:
  - Missing attributes
  - Missing attribute values
  - Small amount of data

MODELLING DATA: TOO LITTLE DATA



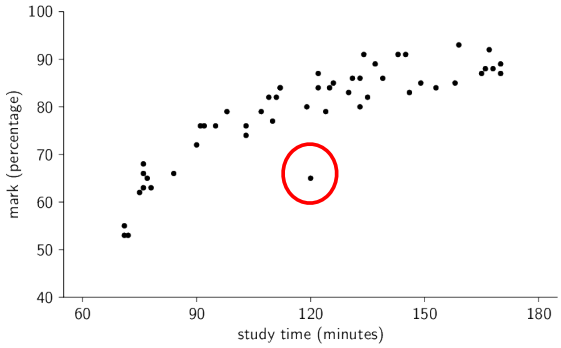
- Even if no missing values, data set may be too small
- Some techniques require large data sets
  - Machine learning models: require a lot of examples to be able to distinguish between classes
- Reliability of predictions might be too low
  - Example are not enough to generalise
- Statistical tests and sample size
  - Confidence of results

MODELLING DATA: MAIN CHALLENGES

- Fractured data
- Fractured data is a very common issue in problems where fusion of data is necessary
- Issues include:
  - Outliers
  - Incompatible data
  - Multiple sources of data
  - Data from multiple levels of granularity

MODELLING DATA: FRACTURED DATA

- Values lies far away from other population values
- Causes:
  - Noise
  - Mixed distributions
  - Natural exceptions
- Problem
  - Distort predictions
- Measures
  - Test abnormality
  - Eliminate (carefully!)



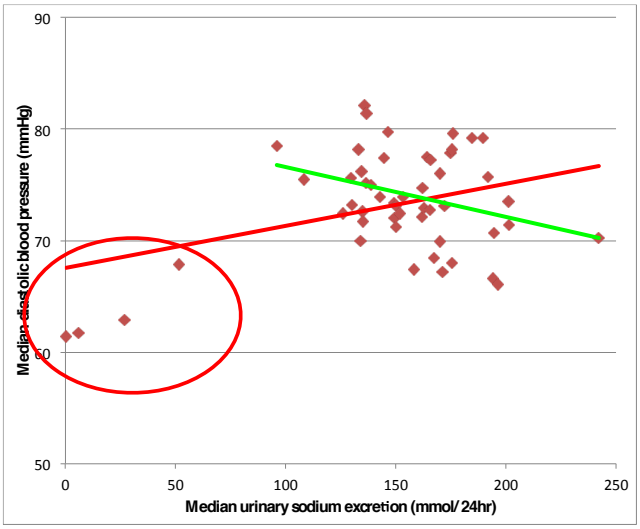
MODELLING DATA: FRACTURED DATA – OUTLIER

- The **identification** and **elimination** of outliers has to be done very carefully.

Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion.

British Medical Journal;
297: 319-328, 1988.

If we exclude these four 'outliers', which are non-industrialised countries with non-salt diets, we get a quite different result!



MODELLING DATA: FRACTURED DATA – MULTIPLE SOURCES

- Data collected by different groups can often lead to data compatibility issues
- Data collected by different teams/projects at different times
- Examples of specific causes of incompatibility:
  - different experimental designs
  - different choices of attributes
  - different hardware

MODELLING DATA: FRACTURED DATA – MULTIPLE SOURCES

- In large enterprises it may be that:
  - data is scattered in different departments and platforms
  - data is acquired and maintained using different software
  - the goal, depth and standard of data collection may vary
- These subtle differences can cause problems when the data from the different sources is combined

MODELLING DATA: FRACTURED DATA – MULTIPLE SOURCES

- In real world applications data can come from different levels of granularity
- Levels of granularity: aerospace example
  - fleet level
    - e.g. all aircraft of a particular type
  - aircraft level
    - e.g. a particular fin-number
  - system level
    - e.g. a particular engine of an aircraft
  - system operation level
    - e.g. operation of the engine for a particular duration or cycle
- The finer the granularity of the data, the more difficult it will be to collect it and to model it.

LECTURE OUTCOMES

- Now, you should be able to answer these questions:
  - What is modelling data?
  - How is it done?
  - What is the difference between data and knowledge?
  - What are the main issues found when modelling data?
  - What are the types of data?
  - Which statistics are used to describe data?

THE END

Questions