Data Analytics in Cyber Security
(CT115-3-M)(Version E)

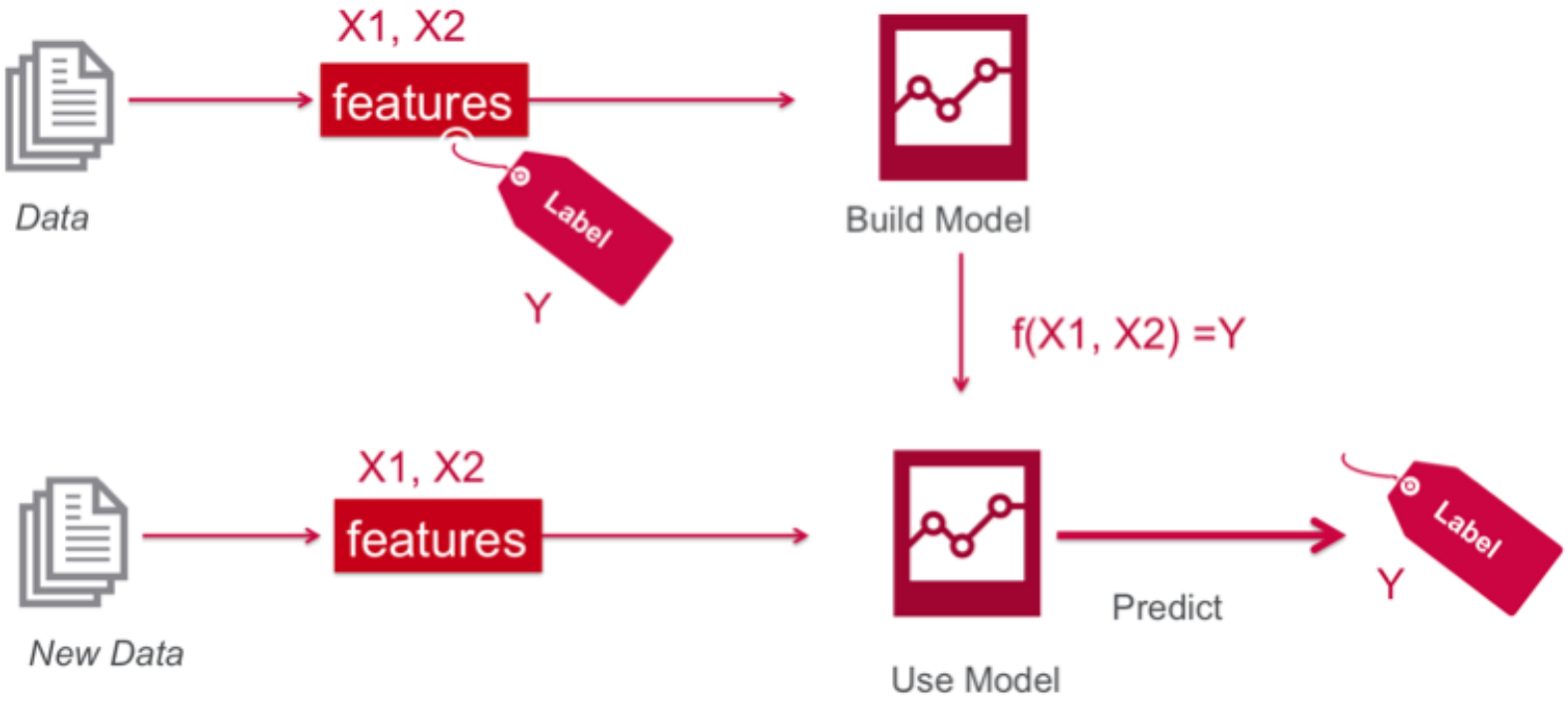
Model Logic

TOPIC LEARNING OUTCOMES

At the end of this topic, you should be able to:

1. To understand concepts in statistics
2. To understand concepts in statistical modelling
3. To understand concepts in predictive modelling
4. To understand concepts in scientific methods

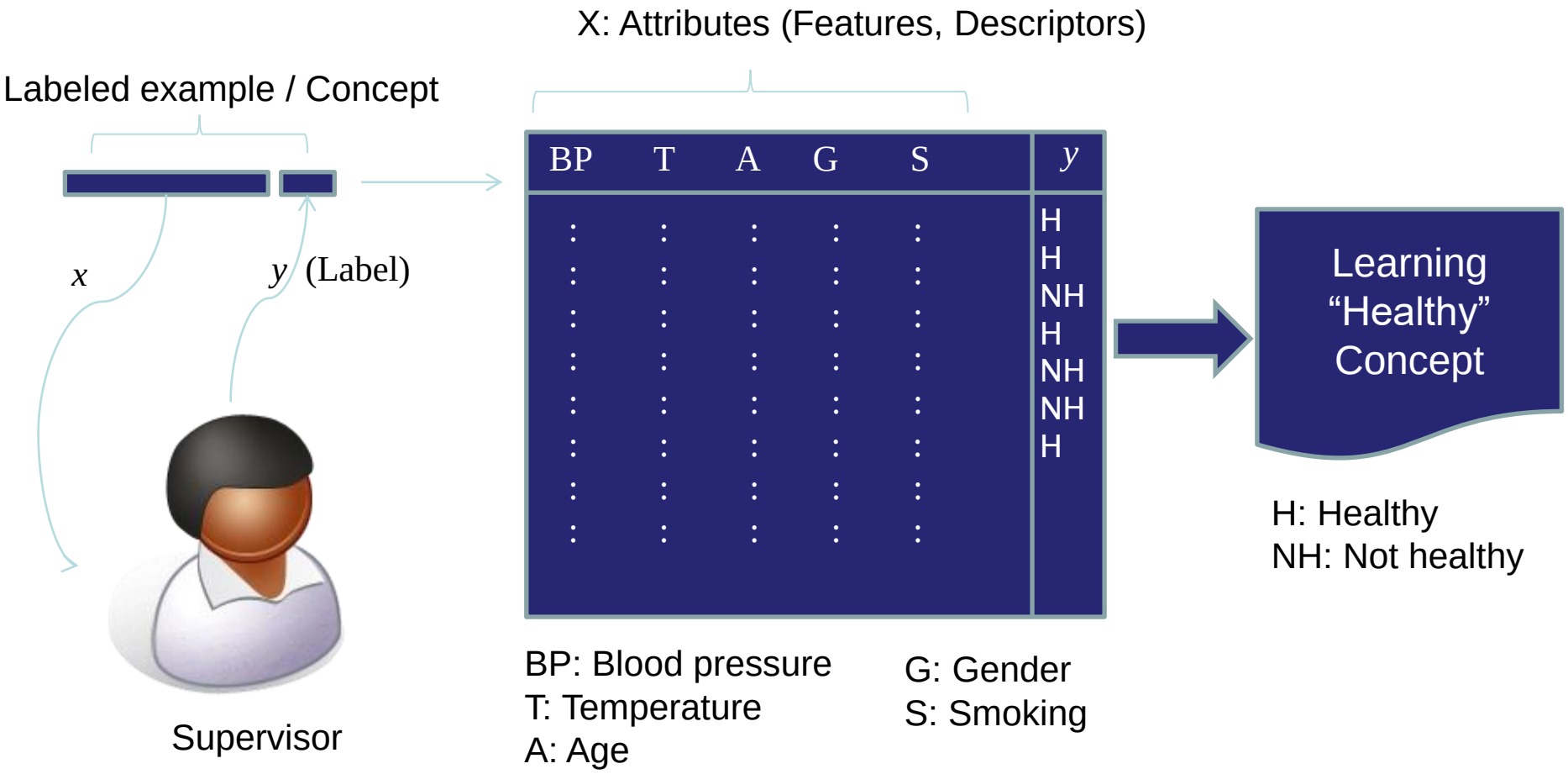
Supervised Learning



That's nice – but where do the labels come from?

Supervised Learning

The labels come from *human intelligence*



Contents & Structure

- Statistics
- Predictive Modeling
- Scientific Method

Why Do We Need Statistics?

"Impossible things usually don't happen."

- Sam Treiman, Princeton University

Statistics helps us quantify "usually."

What is a *Statistic*?

- "A quantity that is computed from a sample [of data]."
  Merriam-Webster

→ a fact or piece of data obtained from a study of a large quantity of numerical data.

Basics

- Independent Events:
  - One event does not affect the other
  - Knowing probability of one event does not change estimate of another
- Covariance:
  - Degree that two variables vary with each other
  - Two independent variables have Covariance of Zero
- Correlation:
  - Normalized Covariance (between –1 and 1)
  - Represents degree of linear relationship
  - Positive: when one gets bigger, the other gets bigger
  - Negative: when one gets bigger, the other gets smaller

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable x and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

r_{xy} = correlation coefficient between x and y

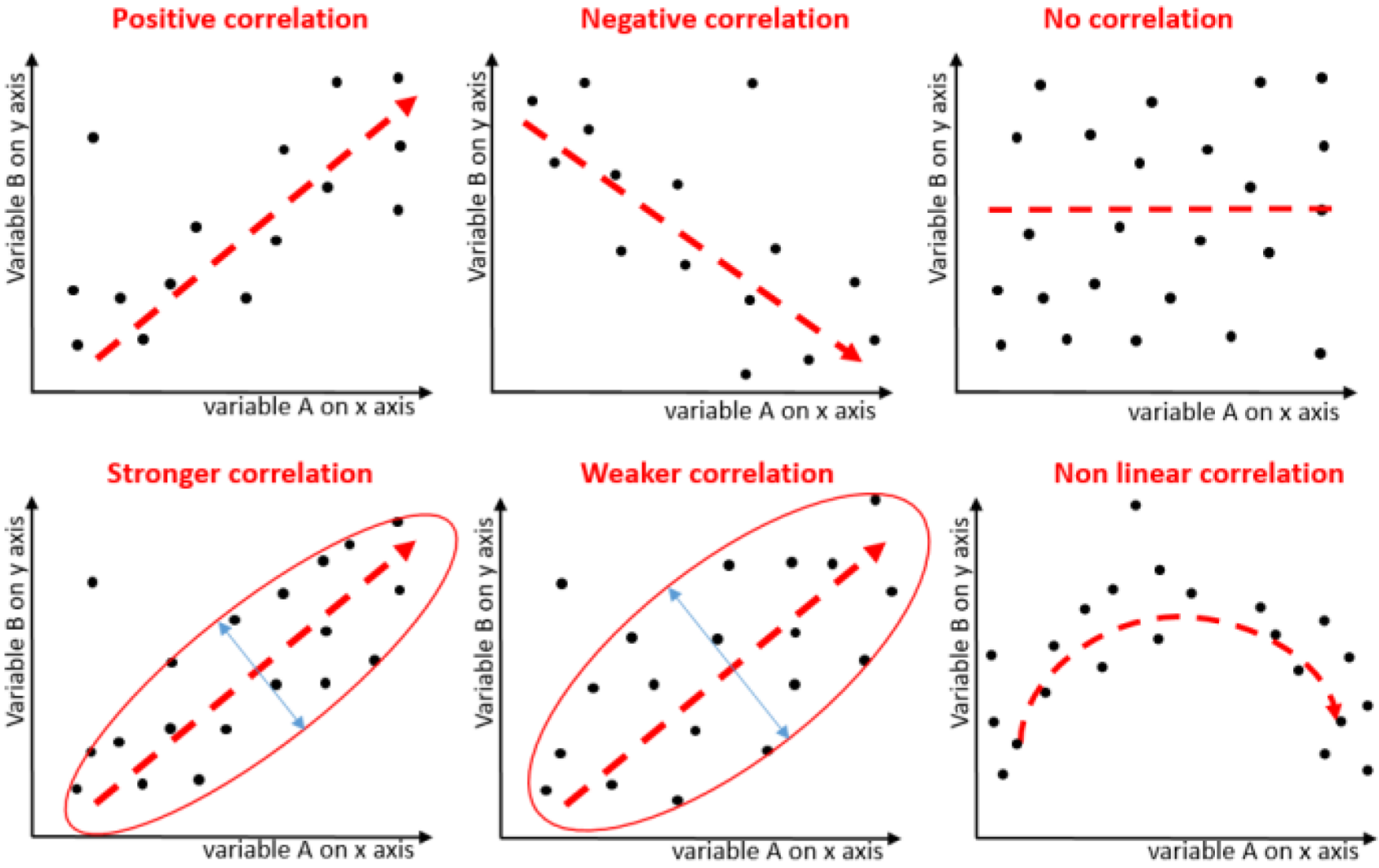
x_i = the values of x within a sample

y_i = the values of y within a sample

$\bar{x}$ = the average of the values of x within a sample

$\bar{y}$ = the average of the values of y within a sample

Correlation



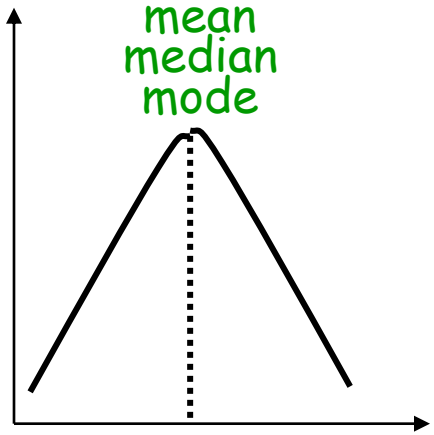
Basics

Indices of central tendency

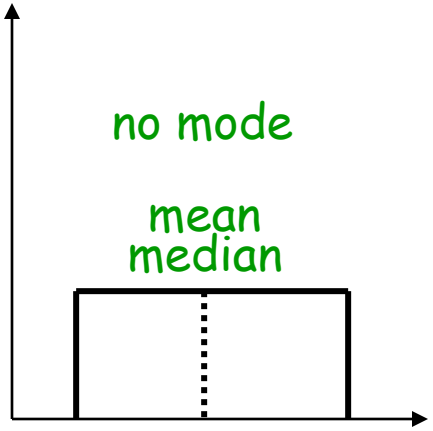
Summarize Data by a Single Number

- Three most popular: mean, median, mode
- Mean – sum all observations, divide by number of observations
- Median – midpoint value when sorted
- Mode – most frequent value observed

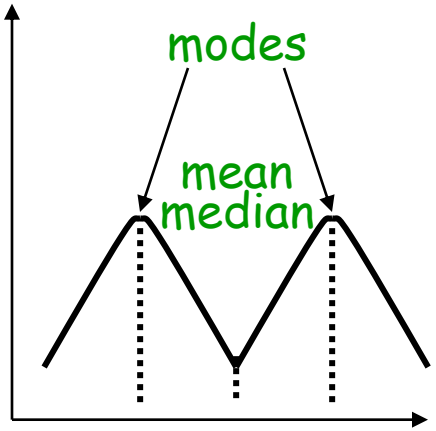
Relationship Between Mean, Median, Mode



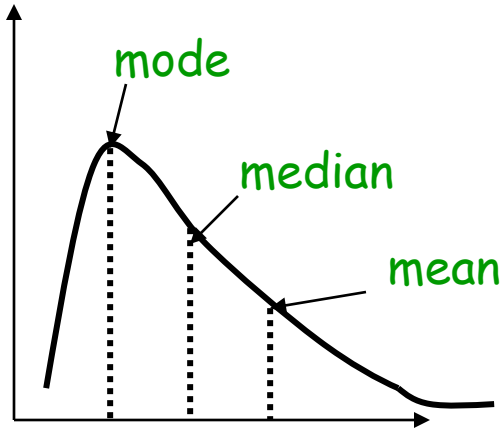
"Gaussian" or "normal"



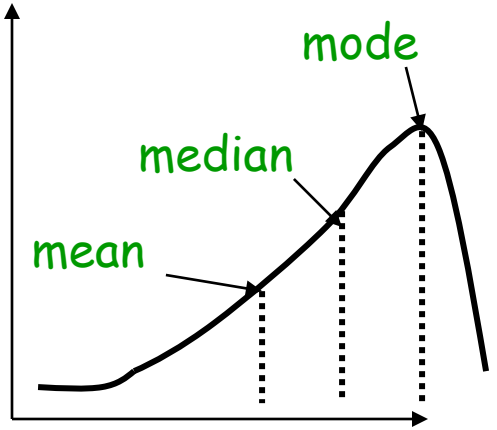
Possible in theory



"bimodal"



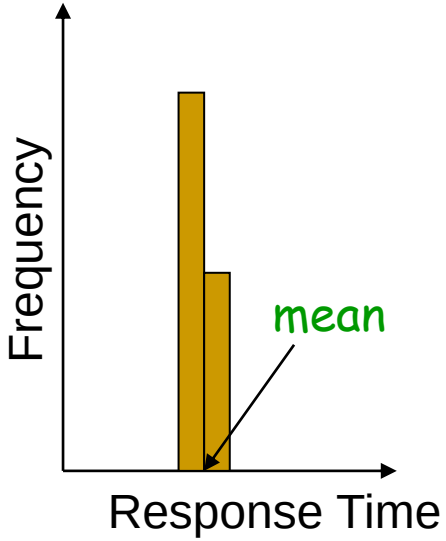
"skewed"

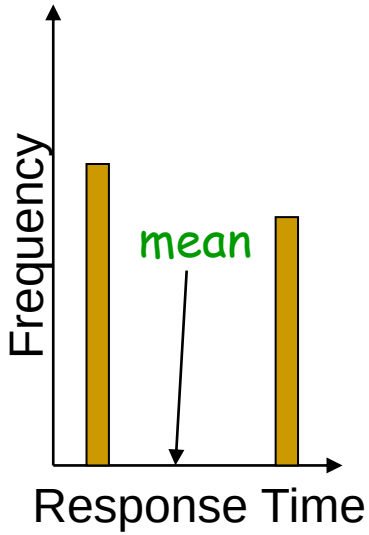


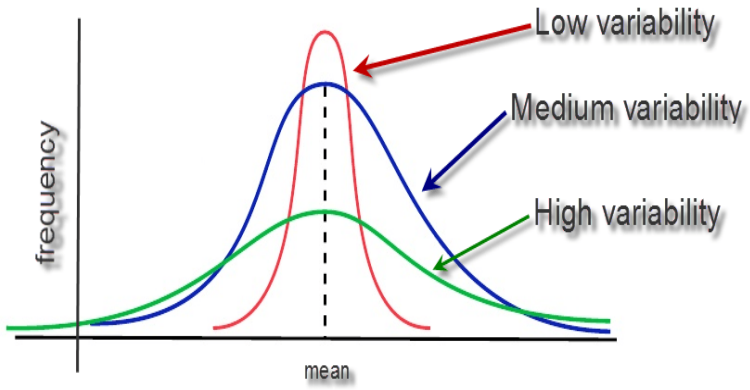
"skewed"

Summarizing Variability

- Summarizing by a single number is rarely enough → need statement about *variability*
  - If two systems have same mean, tend to prefer one with less variability







Variance and Standard Deviation

- *Variance* = square of the distance between x and the mean $\sigma^2 = (x - \bar{x})^2$
  - variance is often denoted σ^2
  - Also called *degrees of freedom*

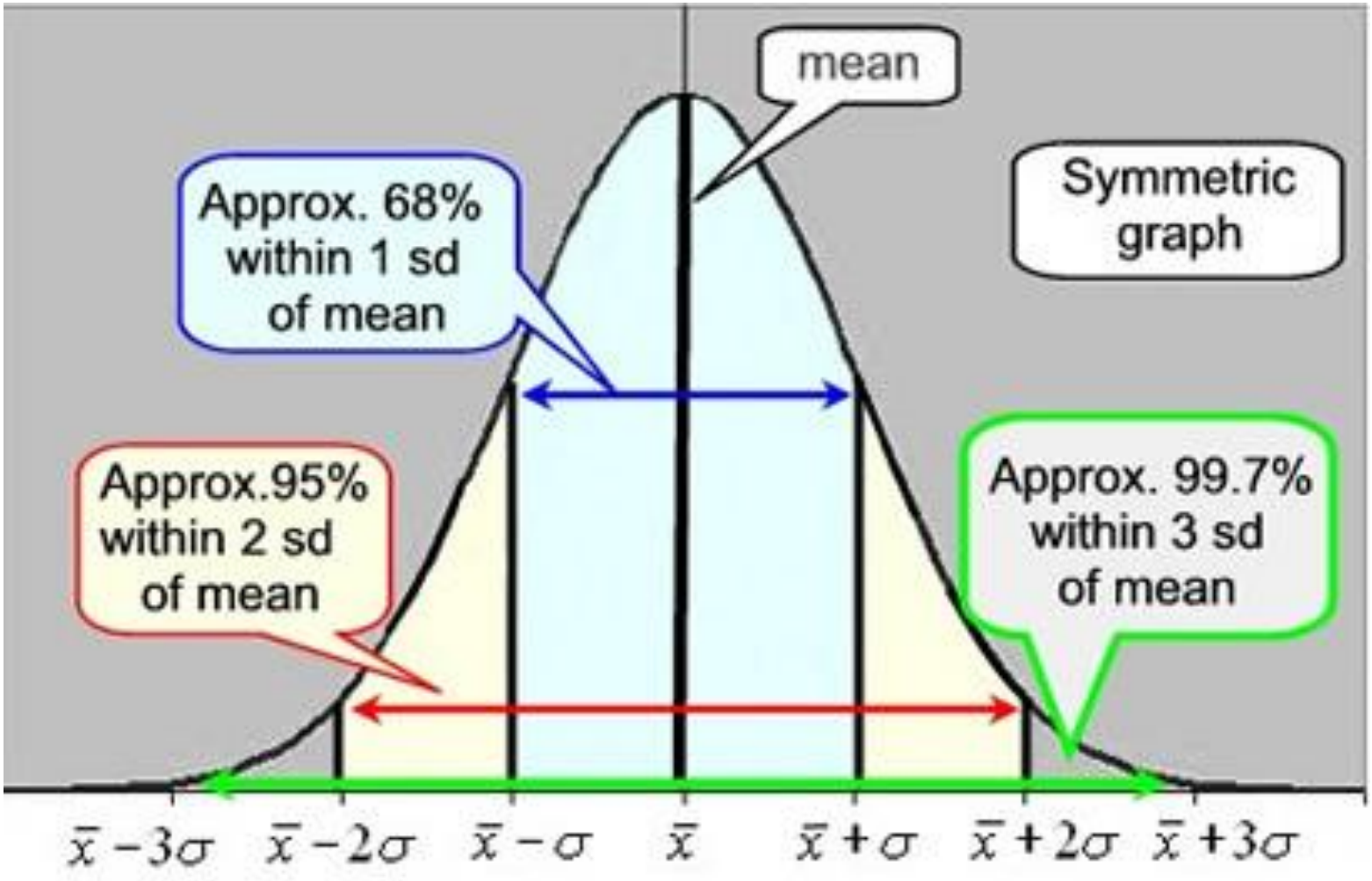
$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

- Main problem is units squared
  - changing the units changes the answer squared
- So, use *Standard Deviation* $\sigma = \text{sqrt}(\sigma^2)$
  - Same unit as *mean*, so can compare to *mean*

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

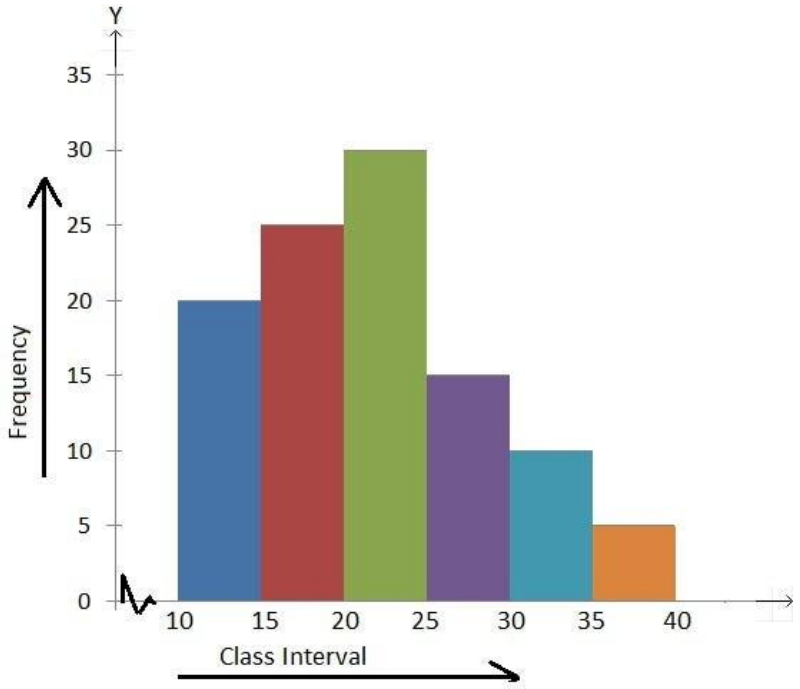
- Ratio of *standard deviation* to *mean*?
  - Called the *Coefficient of Variation* (C.O.V.)
  - C.O.V. = σ / μ
  - Takes units out and shows magnitude

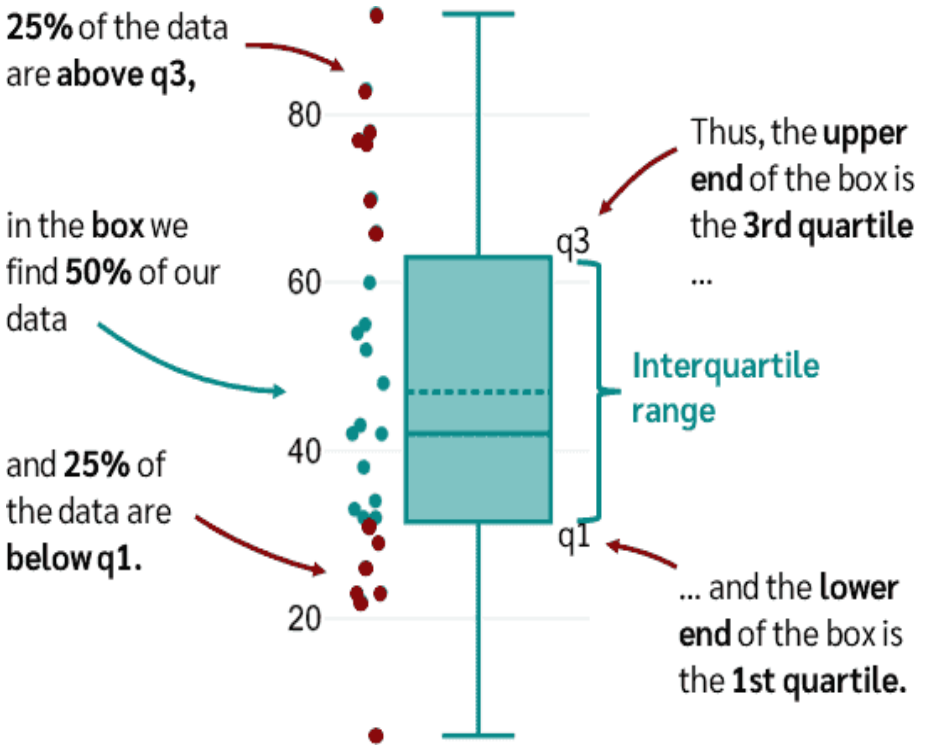
Variance and Standard Deviation



Histogram and Box-Plots

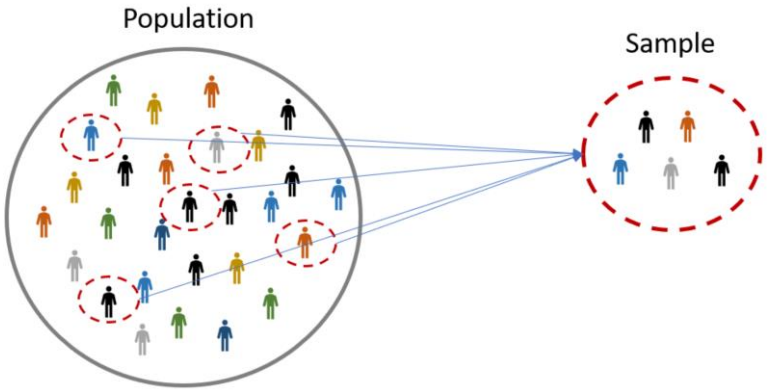
Useful statistical plots that summarise the variability of the data

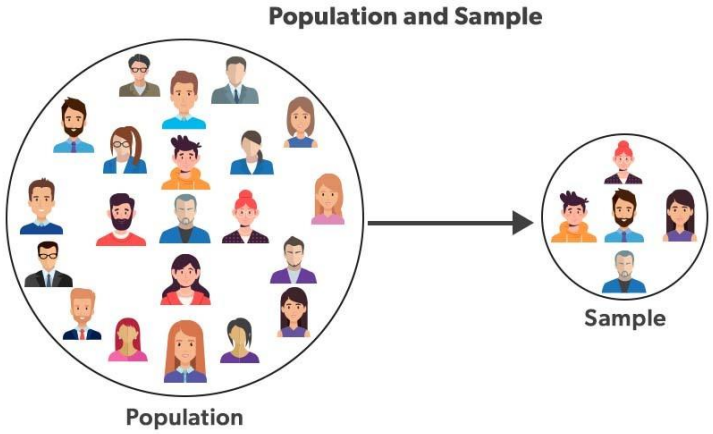




Population versus Sample

- Population refers to the entire group or set of individuals, objects, or events being studied, while a sample is a subset of the population that is used for analysis.
- The word “sample” comes from the same root word as “example”; one sample does not prove a theory, but rather is an example.
- Basically, a definite statement cannot be made about characteristics of all systems. Instead, make probabilistic statement about the range of most systems, i.e., *Confidence intervals*





Summary

- Statistics are tools
  - Help draw conclusions
  - Summarize in a meaningful way in presence of noise

- Indices of central tendency and Indices of central dispersion
  - Summarize data with a few numbers

Statistical Modelling / Machine learning

- **Statistical modelling** is the formalization of relationships between variables in the form of mathematical equations.
- **Machine learning** is an algorithm to optimize a performance criterion using data of particular examples.
- **Machine learning** relies on **Statistical modelling**

Dependent Variable (DV)
e.g.: Salary, University Grade

Constant

b1, b2, b3,....bn
Coefficients/Weights

X1, X2, X3,....Xn
Independent Variable (IV)
e.g.: Experience, Hours of Study

$$Y = b0 + b1*X1 + b2*X2 + b3*X3 + ...bn*Xn$$

ML/Statistical Modelling

- The fewer assumptions in a predictive model, the higher will be the predictive power.
- **Machine Learning** as the name suggest needs minimal human effort.
  - Machine learning works on iterations where computer tries to find out patterns hidden in data.
  - Because machine does this work on comprehensive data and is independent of all the assumptions, predictive power can be very strong for these models.
- **Statistical modelling** are mathematics intensive and based on coefficient estimation.
  - It requires the modeler to understand the relation between variables before putting them in.

Predictive Modeling

- There is a common principle that underlies all supervised machine learning algorithms for predictive modeling.
- Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y)

$$Y = f(X)$$

- This is a general learning task where we would like to make predictions in the future (Y) given new examples of input variables (X)
- We don't know what the function (f) looks like or its form. If we did, we would use it directly and we would not need to learn it from data using machine learning algorithms.

The Perils of Forecasting

Nobel Laureate Physicist Niels Bohr said:

“Prediction is very difficult, especially if it’s about the future.”

Henri Poincare was more positive:

“It is far better to foresee even without certainty than not to foresee at all.”

Anonymous quotes about the perils of forecasting:

“Forecasting is the art of saying what will happen, and then explaining why it didn’t.”

“There are two kinds of forecasts: lucky and wrong.”

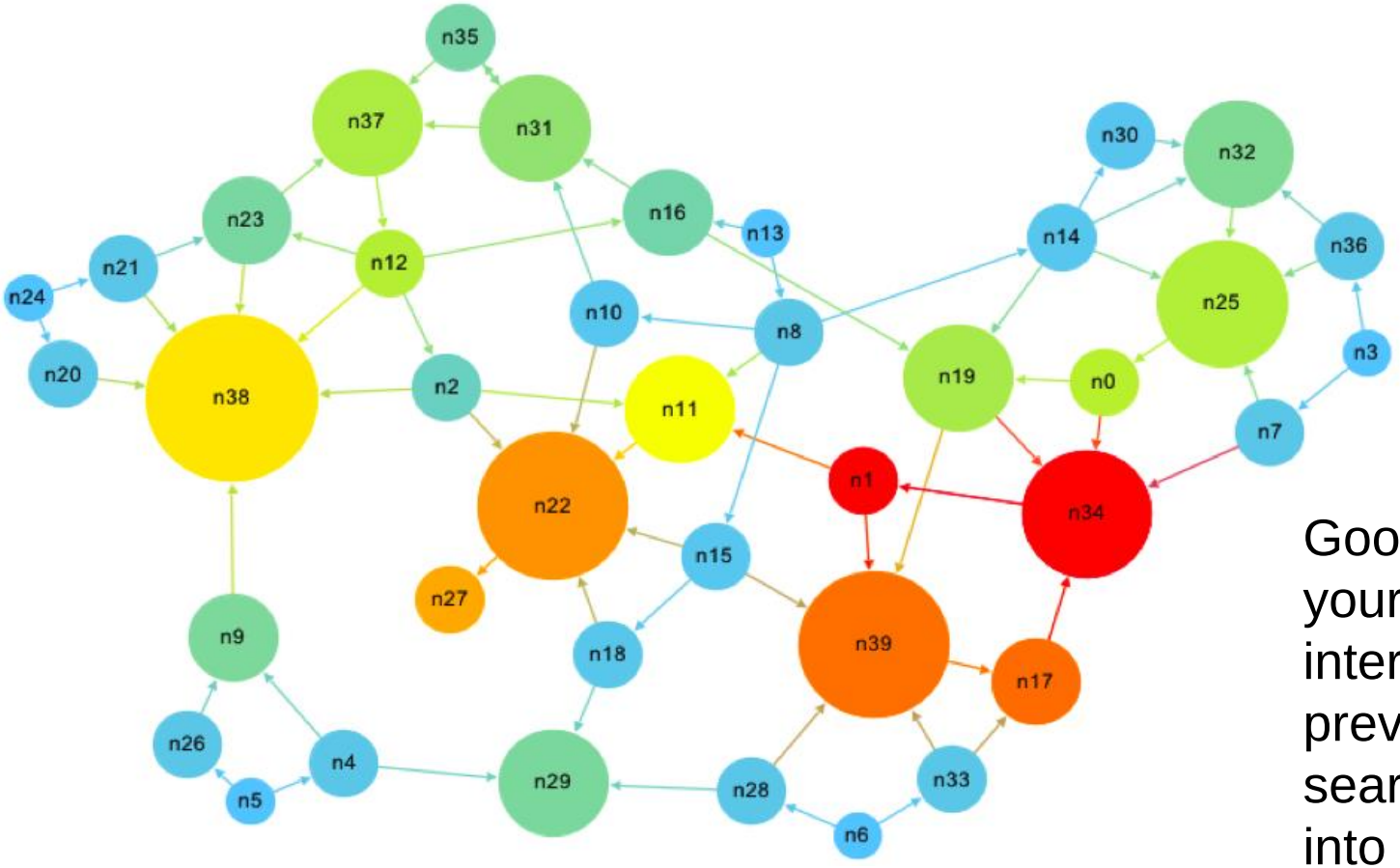
“A good forecaster is not smarter than everyone else; he merely has his ignorance better organized.”

Share Price Prediction

Models try to predict behavior or range of behaviors

Google's "Pagerank" Algorithm

Algorithm to rank the hits so the "most useful" ones come first



Google takes your profile, interests, and previous searches into account

Success Factors

1. Data and domain understanding
   - Generation of data and task
   - Cleaning and representation/transformation
2. Statistical insights
   - Statistical properties
   - Test validity of assumptions
   - Performance measure
3. Modeling and learning approach
   - Choice or development of most suitable algorithm
   - Experiments, Scientific Method (Test, Test, Test)
   - Model Validation

Scientific Methods

- In science, there is a constant interplay between inductive inference (based on observations) and deductive inference (based on theory),
- In science, the goal is to get closer and closer to the 'truth,' which we can only approach but not ascertain with complete certainty.
- In science, like everyday life, most people accept that the usefulness of a theory based on partial knowledge and probabilities is more important than its absolute validity.

Inference

In the words of the American philosopher C.S.Peirce:

- "**Deduction** proves that something **must** be;
  - If A and B then always C
- **Induction** shows that something actually is **operative;**
  - If A and B then most probably C
- **Abduction** merely suggests that something **may** be."
  - If A and B then C ?

Charles Sanders Peirce, "Pragmatism and Abduction," (lecture, Harvard University,Cambridge, MA, May 14, 1903), In The Collected Papers of Charles Sanders Peirce, vol. 5, Pragmatism and Pragmaticism, 180-212, CP 5.186, C. Hartshorne and P. Weiss, eds. (Cambridge, MA: Harvard University Press, 1934)

Deductive Reasoning

- Deduction is the process of deriving the consequences of what is assumed. If the assumptions are true, a valid deduction guarantees the truth of the conclusion.
- In other words, the deductive approach begins with a hypothesis, and we predict what the observations should be if the hypothesis is correct.
- These predictions are tested with experiments, and the outcomes of the experiments support or refute our predictions (hypotheses)
  - We use deductive constructs (association rules) of the form ***if A and B then C*** to help explain observations all the time

Inductive Reasoning

- Where the deductive approach goes from the general (theory) to the specific (observations), the inductive approach goes from the specific to the general.
- We make many observations, discern a pattern, and infer a general explanation
- This is commonly referred to as empirical logic – logic derived from observations about the world.

Inductive Reasoning

- With induction, we observe as much as we can about the world and then define general laws based on a number of observations or experiences of recurring patterns.
- Each observation is a complete and separate instance from the one before, so the premises of inductive reasoning support the conclusion but do not ensure it.

Abductive reasoning

- Abductive reasoning means forming a premise: what we believe to have happened based on our observations.
- The selection criterion is that the hypothesis provides a satisfactory explanation of the observed facts we are interested in explaining
- Abductive inference is not so much a process of inventing hypotheses but rather as one of **adopting a hypothesis**, which is not considered true or verified or confirmed, but **seems to be a worthy candidate for further investigation**

Abductive reasoning

Things to consider:

1. Data must be collected before a hypothesis can be formed
2. Abductive judgements can change, emphasizing or de-emphasizing certain leads.

In this sense, abduction and induction are closely tied: *abduction is the end result of the induction process, and the beginning of the deduction process*

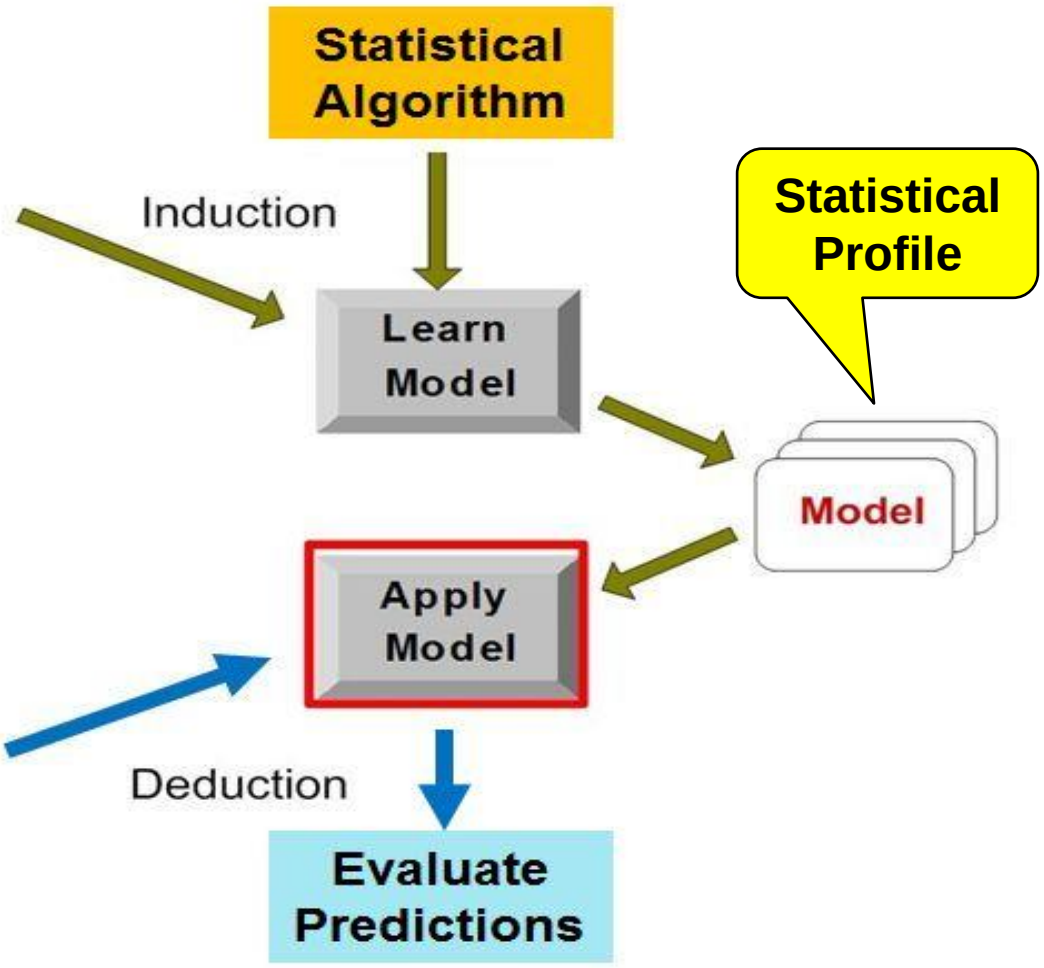
Machine Learning

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

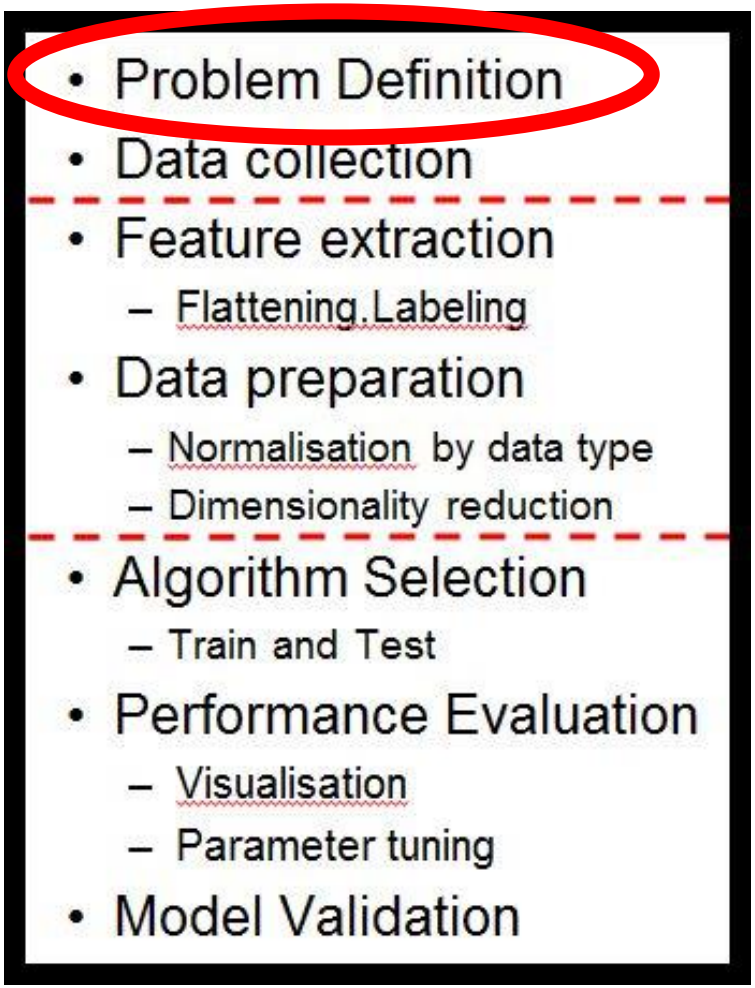
Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Machine Learning Pipeline



Abductive Reasoning:

Problem Definition

- **Goal:** Decide what information is needed
- **Methods:** Typically, committee meetings, brainstorming, and analysis of business objectives
- **Outcome:** A project specification for the Data Scientists.

Further Reading

- **Why AI can't solve unknown problems**

https://bdtechtalks.com/2021/03/29/ai-algorithms-representations-herbert-roitblat/

- **Unsupervised learning can detect unknown adversarial attacks**

https://bdtechtalks.com/2021/08/30/unsupervised-learning-adversarial-attacks-detection/

- **What Google's AI-designed chip tells us about the nature of intelligence**

https://bdtechtalks.com/2021/06/14/google-reinforcement-learning-ai-chip-design/amp/

Review Questions

1. What are the concepts in statistics
2. What is statistical modelling
3. What is predictive modelling
4. What are scientific methods