



Simple Ways of Evaluating Machine Learning Algorithms

BS0004 Introduction to Data Science

Dr Wilson Goh

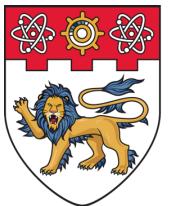
School of Biological Sciences

Learning Objectives

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

- Explain the basics of knowledge discovery.
- Explain the mechanics of performing prediction.
- Explain the four natures of predictions; TP/FP/TN/FN.
- Describe the differences between cross-validation and independent-validation.





NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

The Mechanics of Knowledge Discovery

BS0004 Introduction to Data Science

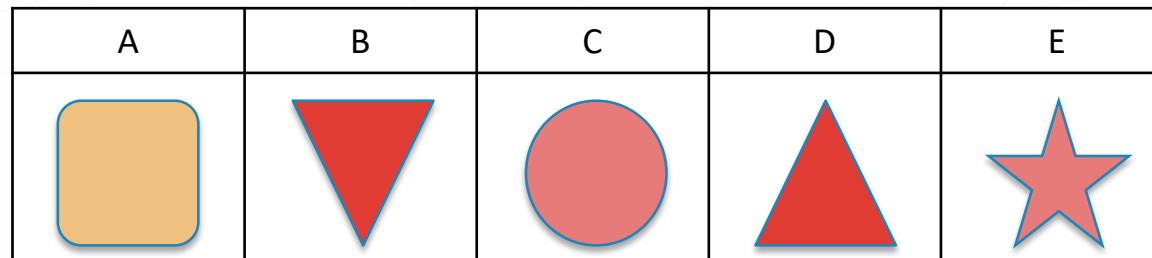
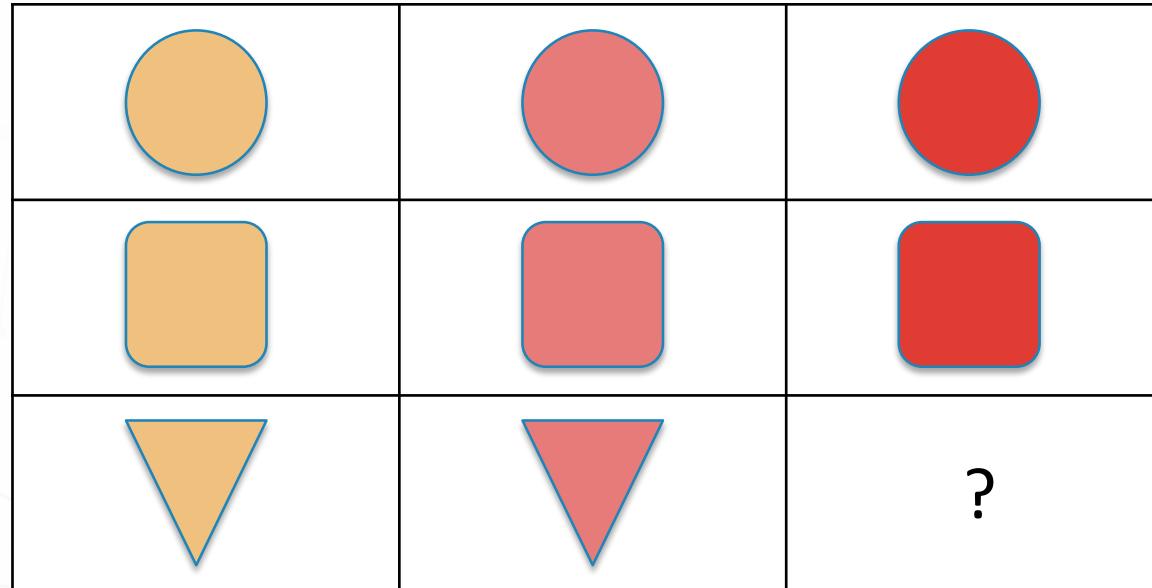
Dr Wilson Goh
School of Biological Sciences



Coming Up: IQ Test!



Which of these figures continues the sequence?

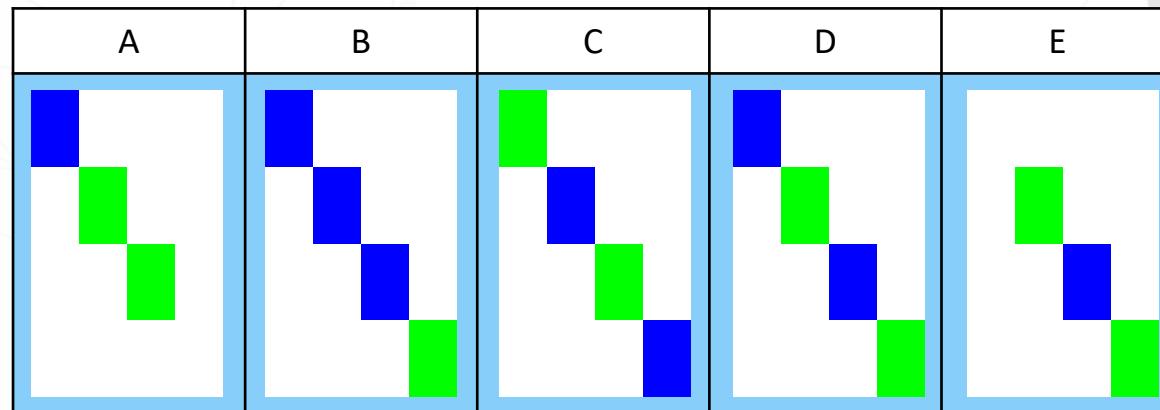
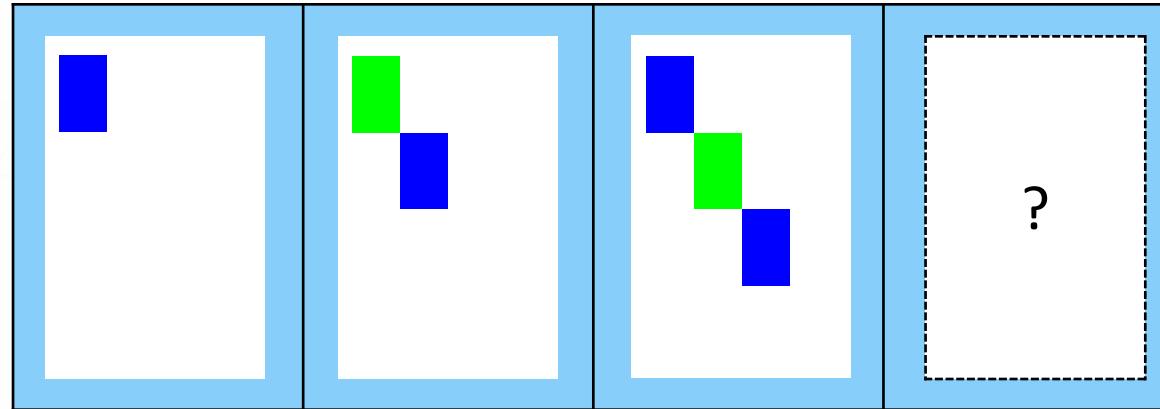


Which of these figures continues the sequence?

Answer: B

- **Variables:** Shapes, colour
- **Traits:** Shapes, colour

Which of these figures continues the sequence?

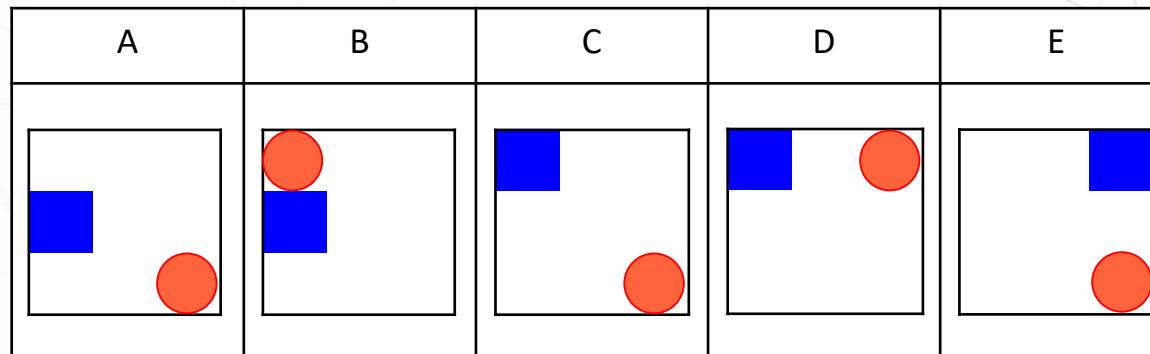
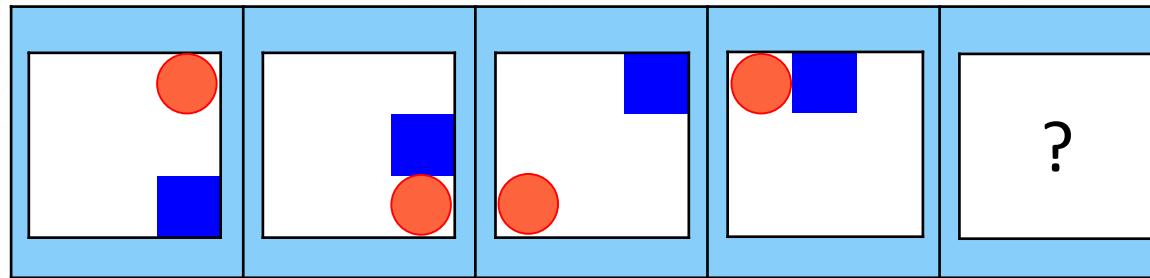


Which of these figures continues the sequence?

Answer: C

- **Variable:** Shapes, colours, sequence of colour, position
- **Trait:** Colours, sequence of colour, position

Which of these figures continues the sequence?



Which of these figures continues the sequence?

Answer: D

- **Variable:** Shapes, colours, number of shapes, relative positions of shapes, shape “movement”
- **Trait:** shape “movement”

Variables/ Features

“

Definition:

“Any characteristic,
number, or quantity that
can be measured or
counted.”

”

Attributes

“

Definition:

“A value that can be adopted by a variable.”

”

Traits

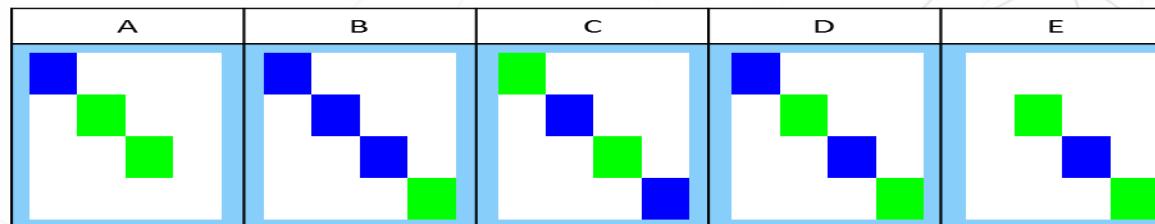
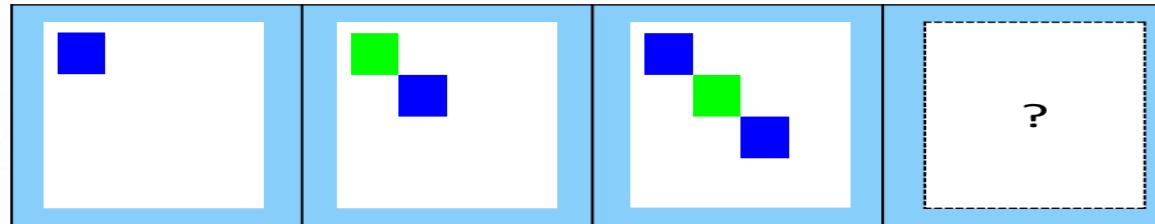
“

Definition:

“A trait is a variable that distinguishes an object from another.”

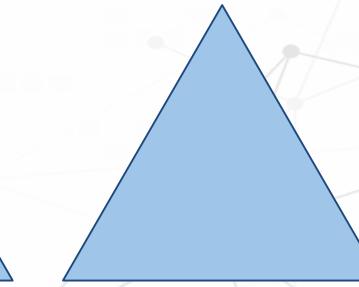
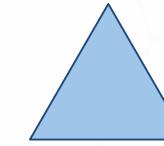
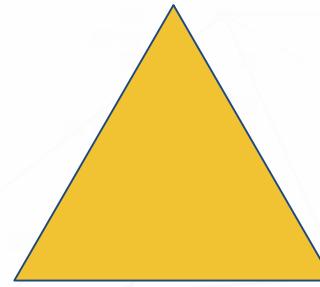
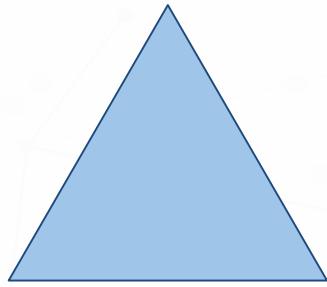
”

Generating Variable, Attributes, Traits



Variables	Attributes	Traits? (in this context)
-----------	------------	---------------------------

Which is the odd one out?



Which one is the odd one out?

Is shape a trait?

Which is the odd one out?

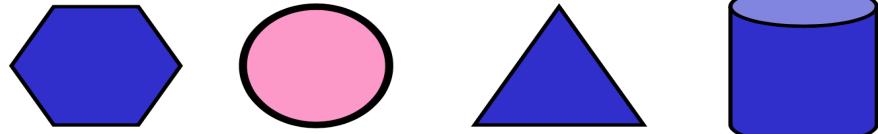
Shape is indeed a trait. But it is not the only one. Colour is also a trait.

This makes the yellow triangle and the small triangle both equally different, if both traits have the same level of importance attached to them.

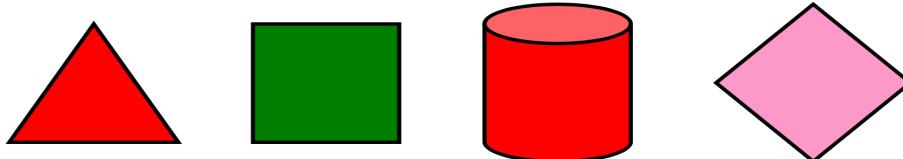
Note: Some of you might be tempted to point out that being small or being yellow creates “more difference”. However, this is a personal bias.

Whose block is this?

Jonathan's blocks



Jessica's blocks



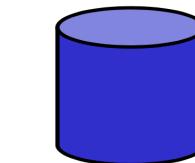
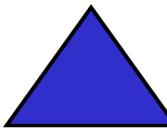
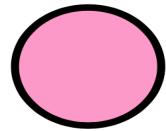
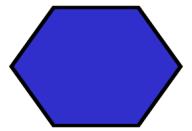
Lets start by writing some rules on each person's blocks. What do you observe?



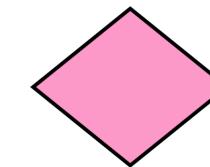
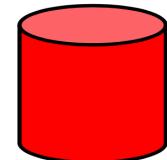
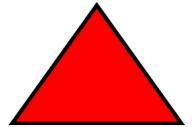
Whose block is this?

Whose block is this?

Jonathan's blocks

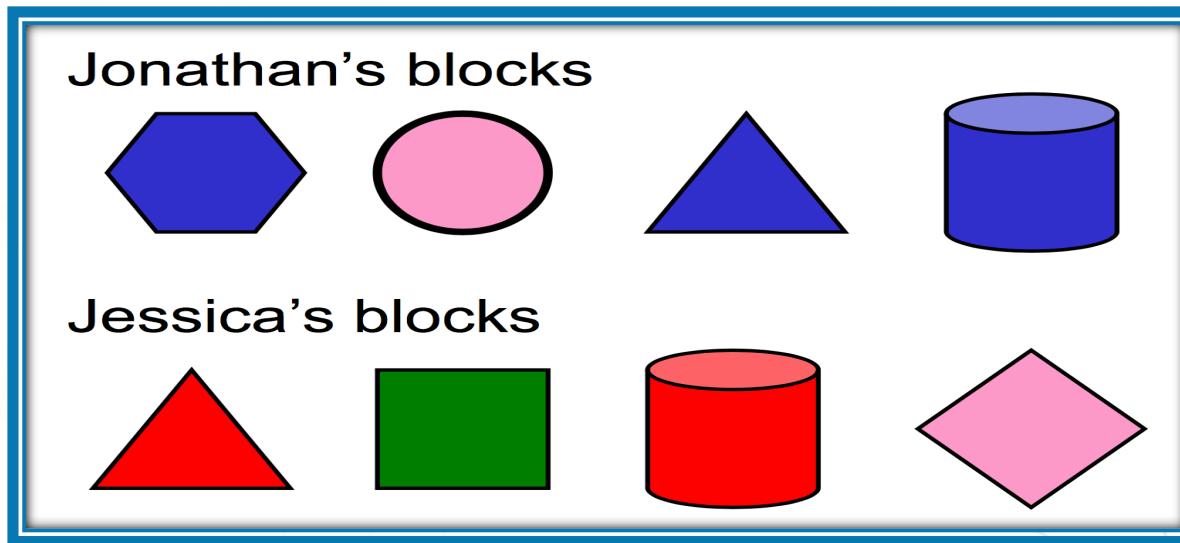


Jessica's blocks



Variables	Attributes	Traits?

Whose block is this?



Variables	Attributes	Traits?
Shape	Hexagonal, Circle..	Possible (The closest shape to square is found only in Jessica)
Colour	Blue, Pink, Red, Green	Possible (blue is only found in Jonathan)
3D	3D, Non-3D	No
Size	1 size only	No

Question



How did you learn who
your mother is?

Knowledge Discovery

“A term describing the finding of useful knowledge from data.”

A natural process:

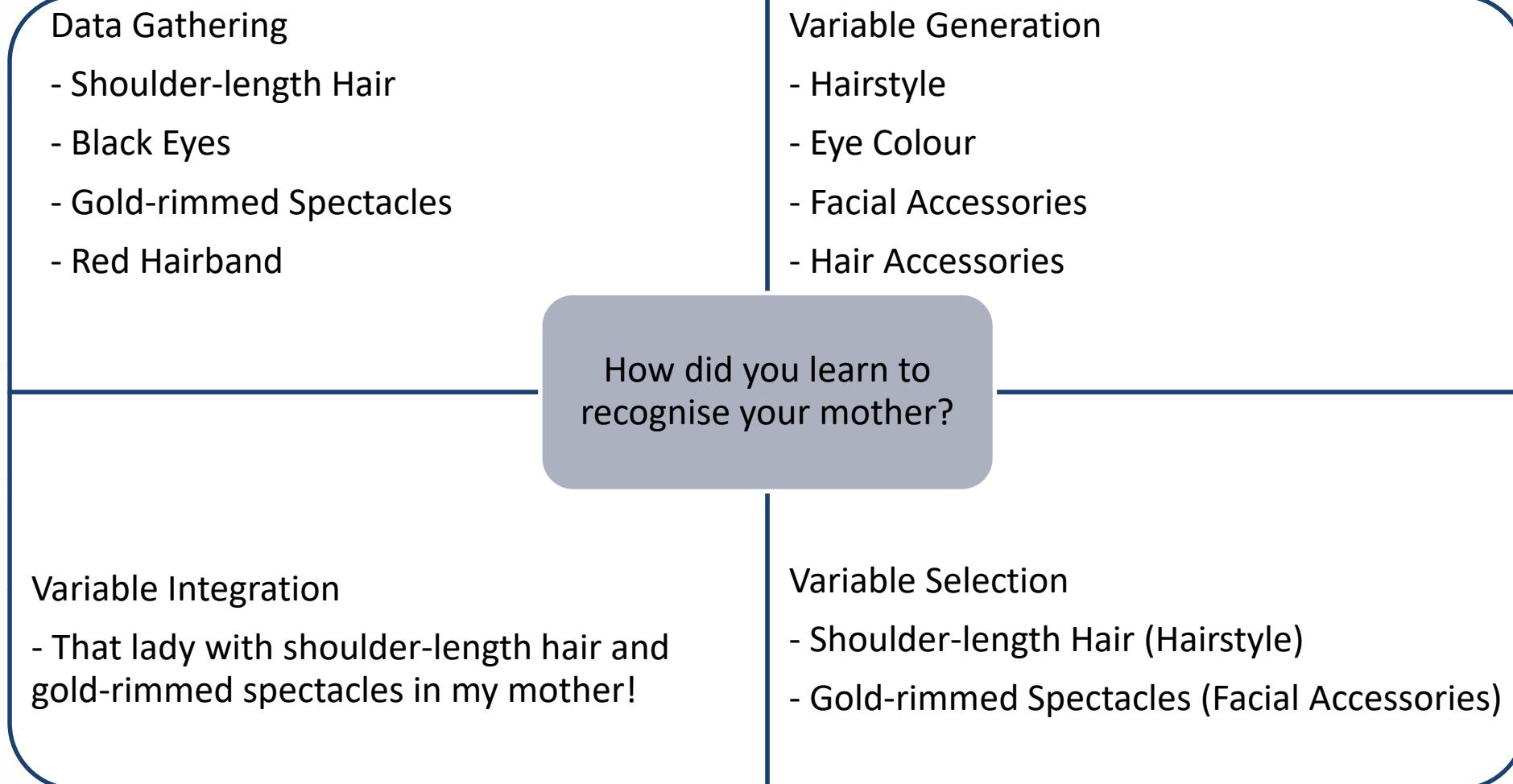
Data Gathering → Variable Generation → Variable Selection → Variable Integration

(Statistics)

(Machine Learning)

We actually do it all the time.

Process of Knowledge Discovery



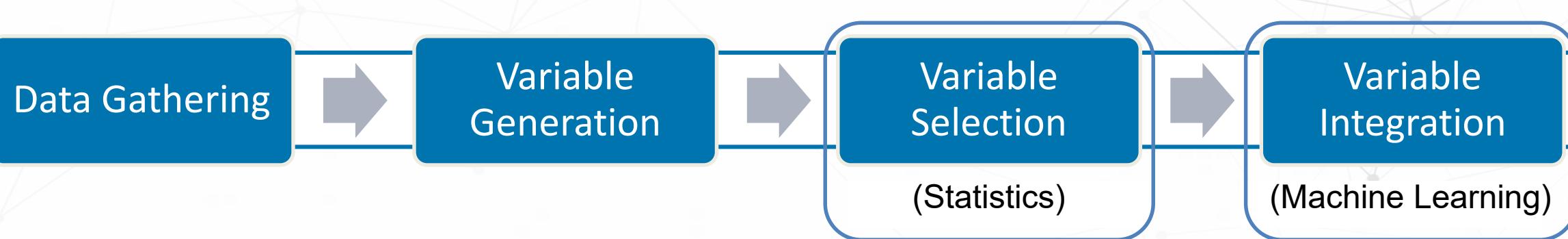
Variables, Attributes, Traits

- By now, you would have commented on the **colour**, **size**, **shape**, and so on, to make some decision in each game.
- Each of these descriptors tell us something about the entity and we can use these to make some decision or create some learning rules.
- A **variable** is any entity that can take on different values (e.g. gender). A variable is also called a **feature**.
- An **attribute** is a specific value on a variable. For instance, the variable *gender* has two attributes: *male* and *female*.
- A **trait** is a distinguishing quality (variable).

What is Knowledge Discovery?

We have just created rules based on observation for the purpose of answering specific questions.

Knowledge discovery is similar. It is the process of extracting useful information from data (observations)

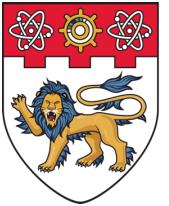


Purpose of Knowledge Discovery

In Biology:

Diagnosis: Through data analytics and hypothesis testing, computers will be able to pinpoint the genes that are responsible for causing different diseases.

Prognosis: With the help of computers and new biomedical knowledge, we can predict who has a higher chance of having certain diseases with higher accuracy and more certainty.



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Prediction and Statistics

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Prediction and Statistics

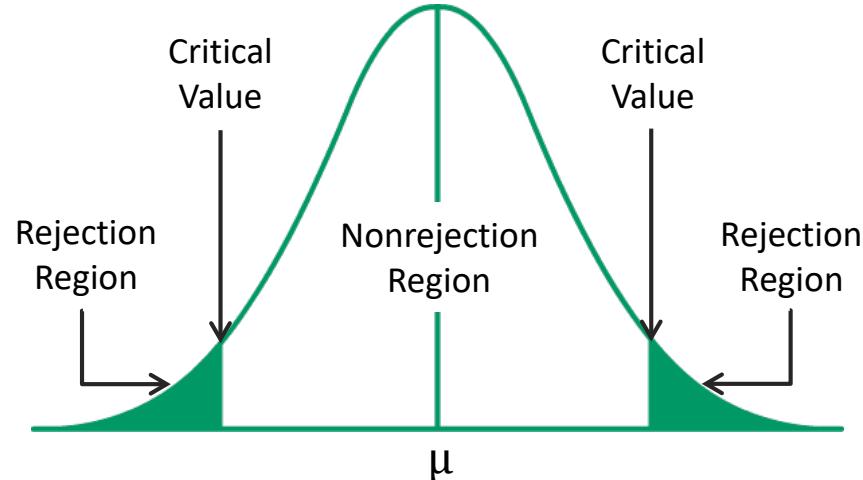
- Prediction is estimating unknown data based on extrapolation of extracted data during knowledge discovery.
- Predictions are grounded on statistics:
 - Distinguishing factor (between H₀ and H₁): traits → describes significantly different attributes.
 - If a data's attribute is significantly different from others (i.e. small p-value), then the data is predicted to be relevant (i.e. H₀ is rejected).

Source: Finlay, Steven (2014). *Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods* (1st ed.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 1137379278.

Prediction and Statistics

Diagnosis of Type 2 Diabetes:

- Trait: blood glucose level
- Normal blood glucose level determined from previous data (knowledge discovery)
- H_0 : non-diabetic, H_1 : diabetic
- If blood glucose level significantly higher (small p-value), H_0 rejected \rightarrow person likely to be diabetic



SUGAR LEVEL	FASTING PLASMA GLUCOSE
DIABETES	$\geq 7.0 \text{ mmol/L}$ ($\geq 126 \text{ mg/dL}$)
PRE-DIABETES	$6.1 - 6.9 \text{ mmol/L}$ ($110 - 125 \text{ mg/dL}$)
NORMAL	$< 6.1 \text{ mmol/L}$ ($< 110 \text{ mg/dL}$)

Prediction and Statistics

Predictions are not perfect:

- Sometimes, predictions does not equate the actual condition.
- Example, wrongly claiming that the person has the disease (when he does not have it).

4 ways to describe the nature of predictions against the actual data:

- True Positive (TP)
- False Positive (FP)
- True Negative (TN)
- False Negative (FN)

Does She Like Me?

H_0 : I do not think she likes me.

H_1 : I think she likes me.

		Expectations	
		I think she does	Nah, don't think she does
Reality	She does	True Positive	False Negative
	She doesn't	False Positive	True Negative

Gypsy's Prediction

The gypsy says... “soon you will meet the girl of your dreams...”

Reality... “you will die single”

That is a false positive.

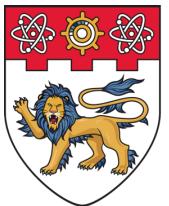
4 possible outcomes (relating prediction to reality) → True positive (TP), False positive (FP), True Negative (TN) and False Negative (FN).

	Predicted as Positive	Predicted as Negative
Positive	TP	FN
Negative	FP	TN

Gypsy's Prediction

- But the gypsy is actually not randomly guessing. Although she has no real powers, she looked at your clothes, the car you drove, your age, your enthusiasm and perhaps, also your innate charm.
- She knows that you are young, relatively good looking, looking for love actively.
- Given what she has seen in the past, she rated your chance as “not bad”, which led to her prediction.
- This is exactly what machine learning does. It looks at variables, and decide which are traits. It then bases the decision on past knowledge, and makes a prediction.

Gypsy Predicts	Reality	TP/FP/ TN/FN
You will meet someone	You did	TP
You will not meet someone	You did not	TN
You will not meet someone	You did	FN
You will meet someone	You did not	FP



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Relating Prediction to Hypothesis-driven Statistics

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

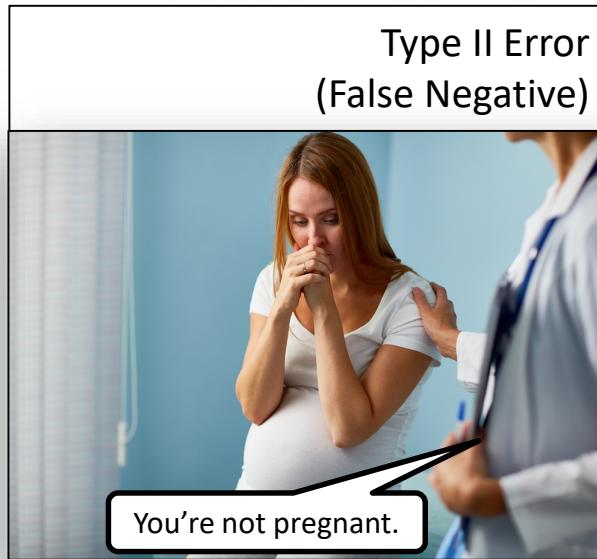


How to Remember?

Type I Error
(False Positive)



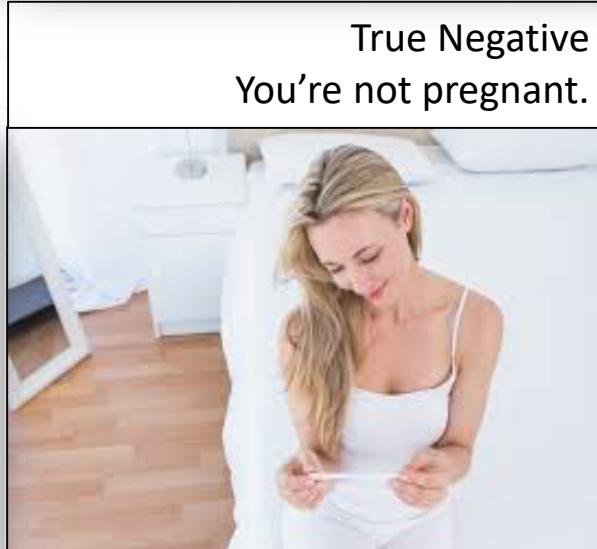
Type II Error
(False Negative)



True Positive
You're pregnant.



True Negative
You're not pregnant.

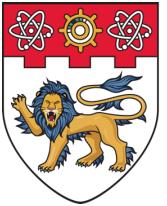


Do you recall type I and II statistical errors?

Type I: Reject the null when the null is true.
Type II: Fail to reject the null when the null is not true.

Confusion Matrix

		Prediction	
		Relevant (Supports H_1)	Irrelevant (Supports H_0)
Reality	Relevant (Supports H_1)	True Positive	False Negative
	Irrelevant (Supports H_0)	False Positive	True Negative



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Multiple Testing in Predictions

BS0004 Introduction to Data Science

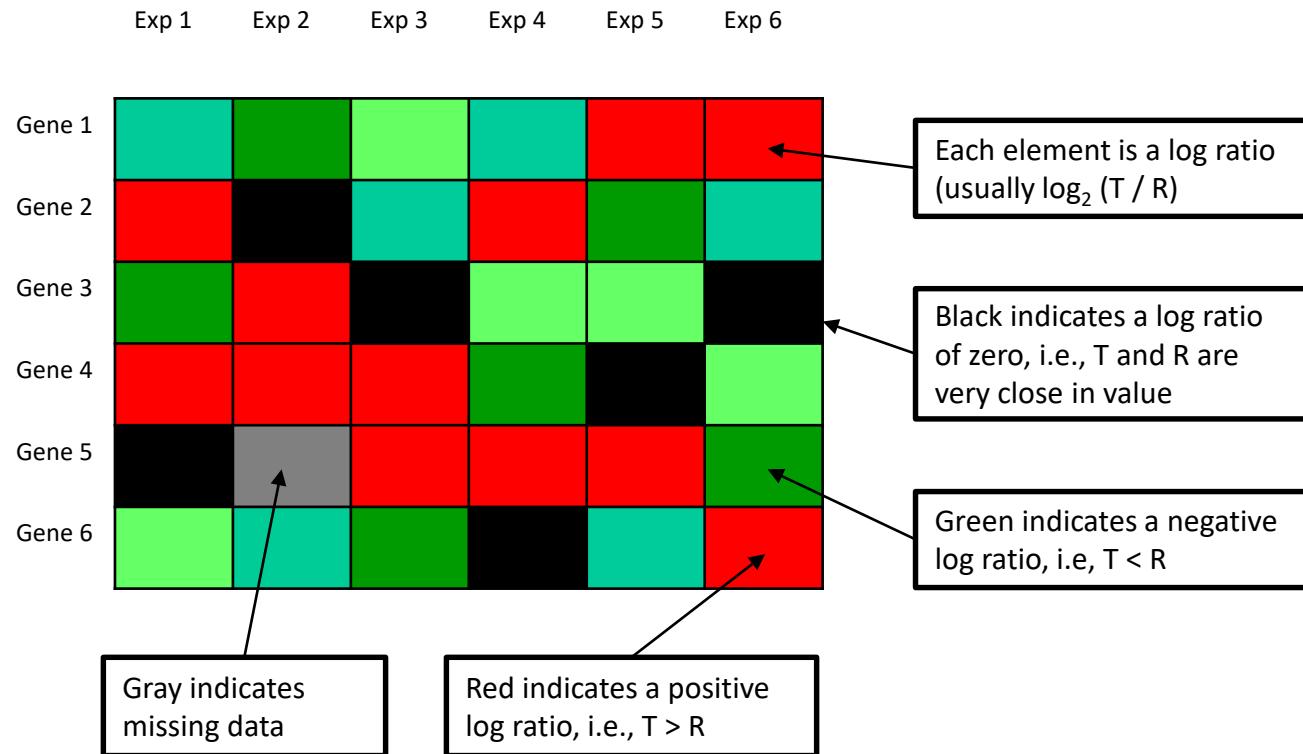
Dr Wilson Goh
School of Biological Sciences



Testing many Hypotheses

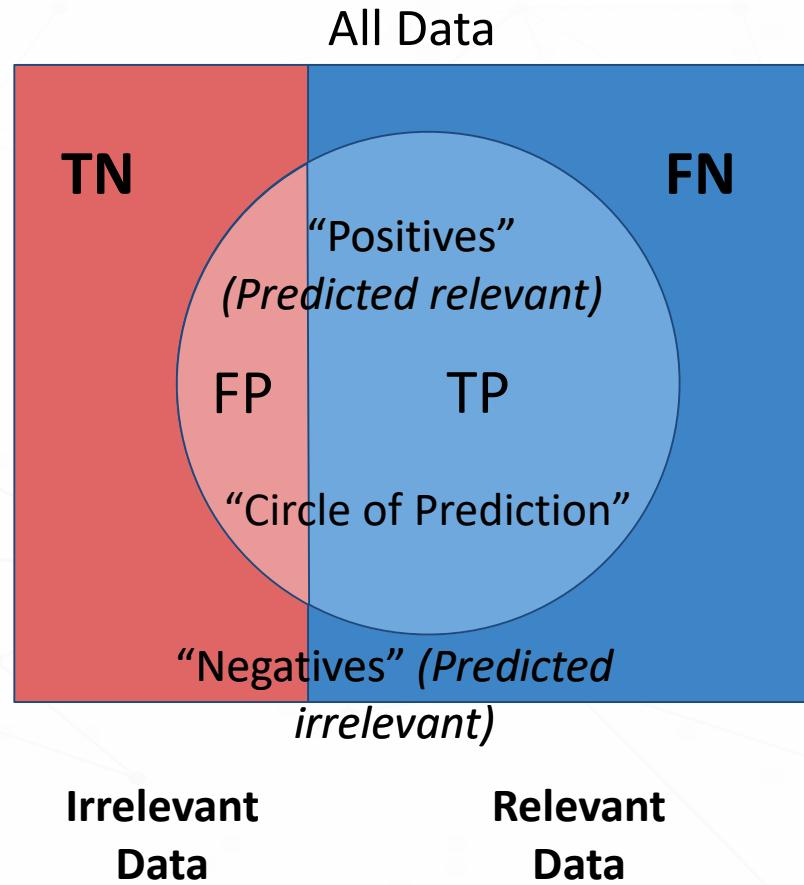
- You may test many hypotheses at once.
- For example, in a microarray experiment, you are effectively performing thousands of hypothesis testing at once (per gene).
- Prediction on each gene falls into 1 of 4 outcomes.

The Expression Matrix is a representation of data from multiple microarray experiments.



T is the gene expression level in the testing sample, R is the gene expression level in the reference sample.

Testing many Hypotheses

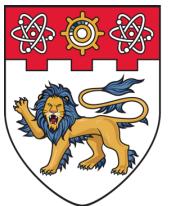


Prediction = Actual
→ **TRUE** (Positive/Negative)

Prediction ≠ Actual
→ **FALSE** (Positive/Negative)

Testing many Hypotheses

- The implication is that if every one of your predictions are a true positive, then you are very much in luck!
- But notice that because now that given n predictions, it can be split down into 4 quadrants, we do need some metrics for evaluating the overall performance of our experiment (i.e., is everything we are testing a true positive? What is the overall rate of mistakes?)



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Prediction Performance Evaluation

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

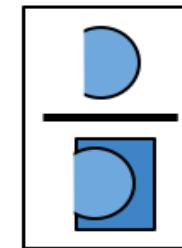


Metrics for performance evaluation

Sensitivity/ Recall

How well it can capture all relevant results in the prediction?

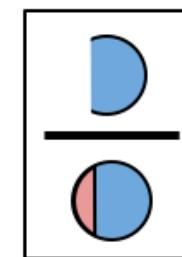
$$R = \frac{TP}{(TP + FN)}$$



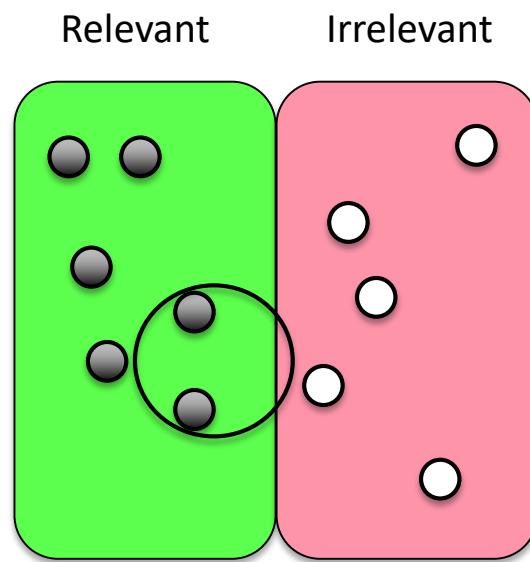
Precision/ Positive Predictive Value (PPV)

What proportion of predicted values are true positive?

$$PPV = \frac{TP}{(TP + FP)}$$

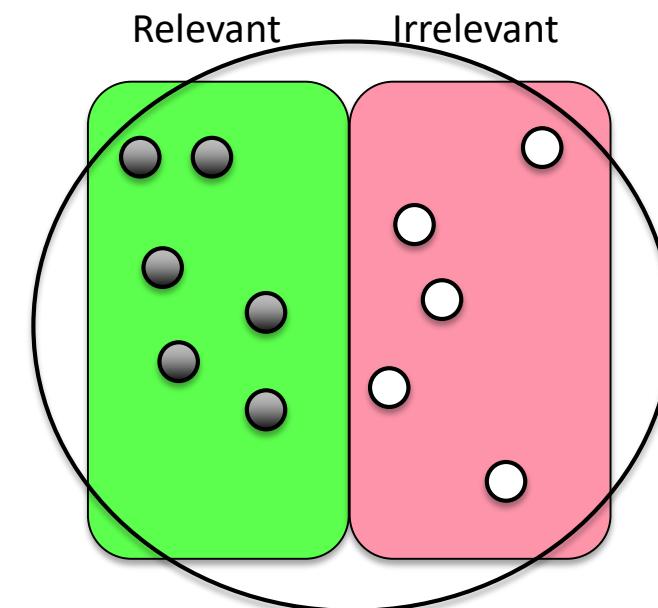


Precision and Recall Work Against Each Other



Stringent p-value Cutoff

Precision ↑
Recall ↓

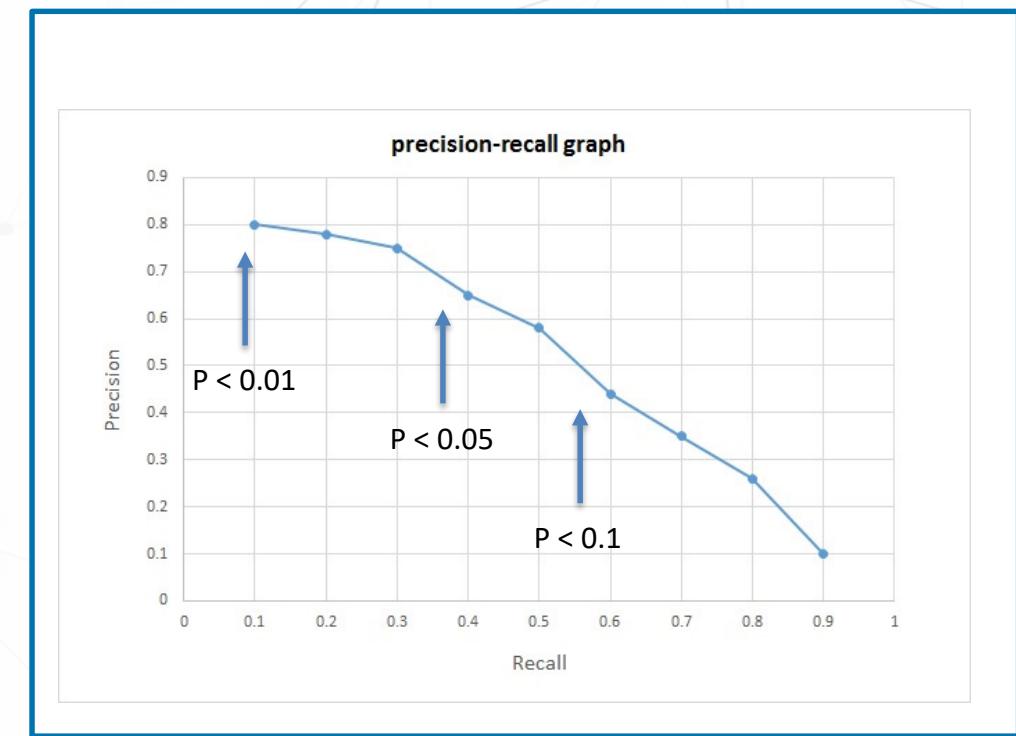


Loose p-value Cutoff

Precision ↓
Recall ↑

Precision and Recall tradeoff

- A predicts better than B if A has better recall and precision than B.
- There is a trade-off between recall and precision.
- In some apps, once you reach satisfactory precision, you optimise for recall.
- In some apps, once you reach satisfactory recall, you optimise for precision.
- The particular tradeoff level between precision and recall is determined by the threshold, t (which can be a score or p-value).



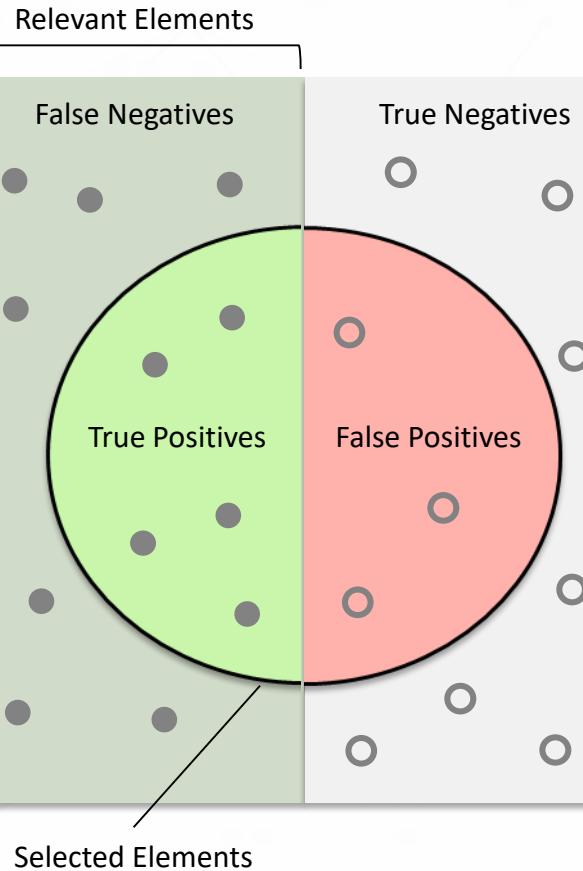
F-Score

- Combines the precision and recall values into a single value.

$$F = 2 \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right)$$

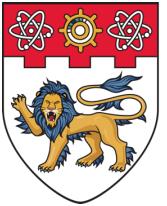
- Gives an idea of the overall ‘quality’ of prediction.
- High F-score:
 - Most of the predictions are true positives (**high precision**).
 - Captures most of the relevant (+) data (**high recall**).
- May oversimplify quality of prediction when used alone.

Accuracy



$$\begin{aligned} \text{Accuracy} &= \frac{\text{No. of correct predictions}}{\text{No. of predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN} \end{aligned}$$

Accuracy seems to be a good measure of overall performance of the predictor. **But is it really?**



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Cross-validation and Independent Validation

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

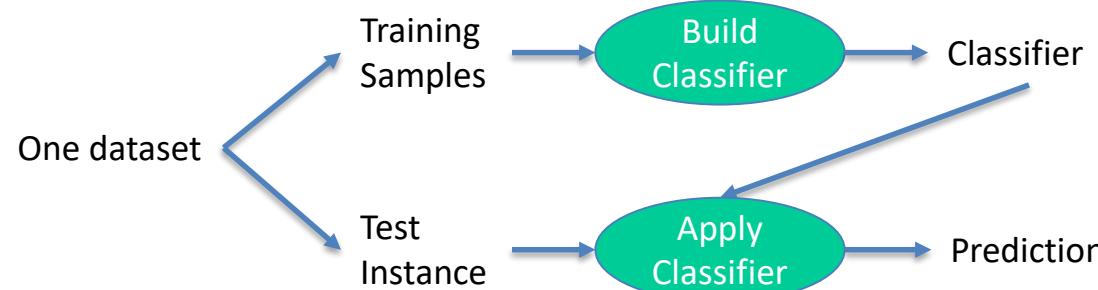


Working with Incomplete Information

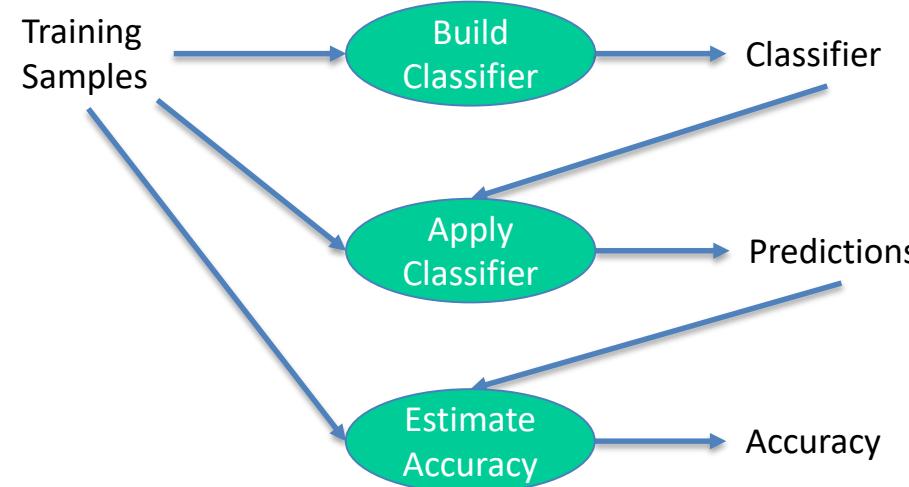
- By now, you might appreciate that in many cases, we do not know which are the relevant or irrelevant variables in real data.
- In which case, we cannot directly calculate precision/recall, etc.
- However, we do know the class labels of our samples (e.g. normal and disease).
- We want to know which variables are relevant. And relevance is evaluated by the ability of each variable to correctly predict the class label.

Cross-validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis (from the training data set) will generalise to a test data set.

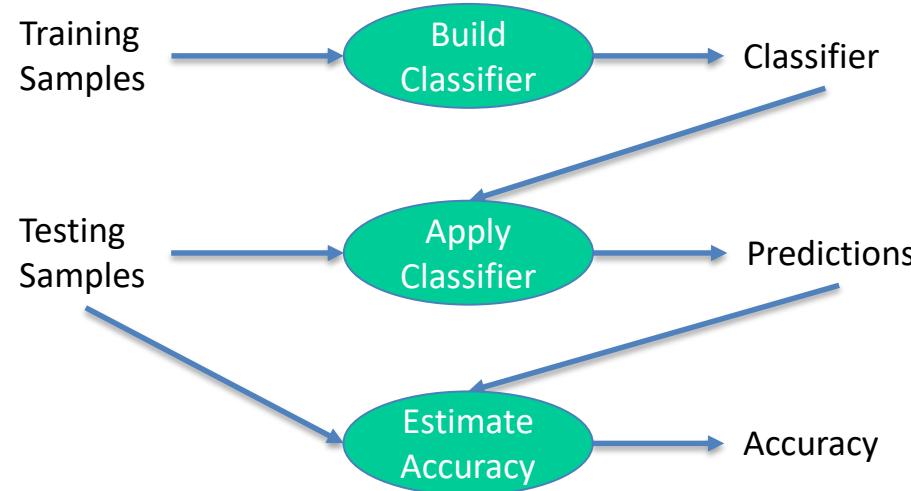


Cross-validation (The Wrong Way)



Why is this way of estimating accuracy wrong?

Cross-validation (The Right Way)



Testing samples are NOT to be used during “Build Classifier” (No cheating!).

How Many Training and Testing Samples?

- No fixed ratio between training and testing samples; but typically 2:1 ratio.
- Proportion of instances of different classes in testing samples should be similar to proportion in the real world, and preferably also to proportion in the training samples.
- What if there are insufficient samples to reserve 1/3 for testing?

Cross-validation

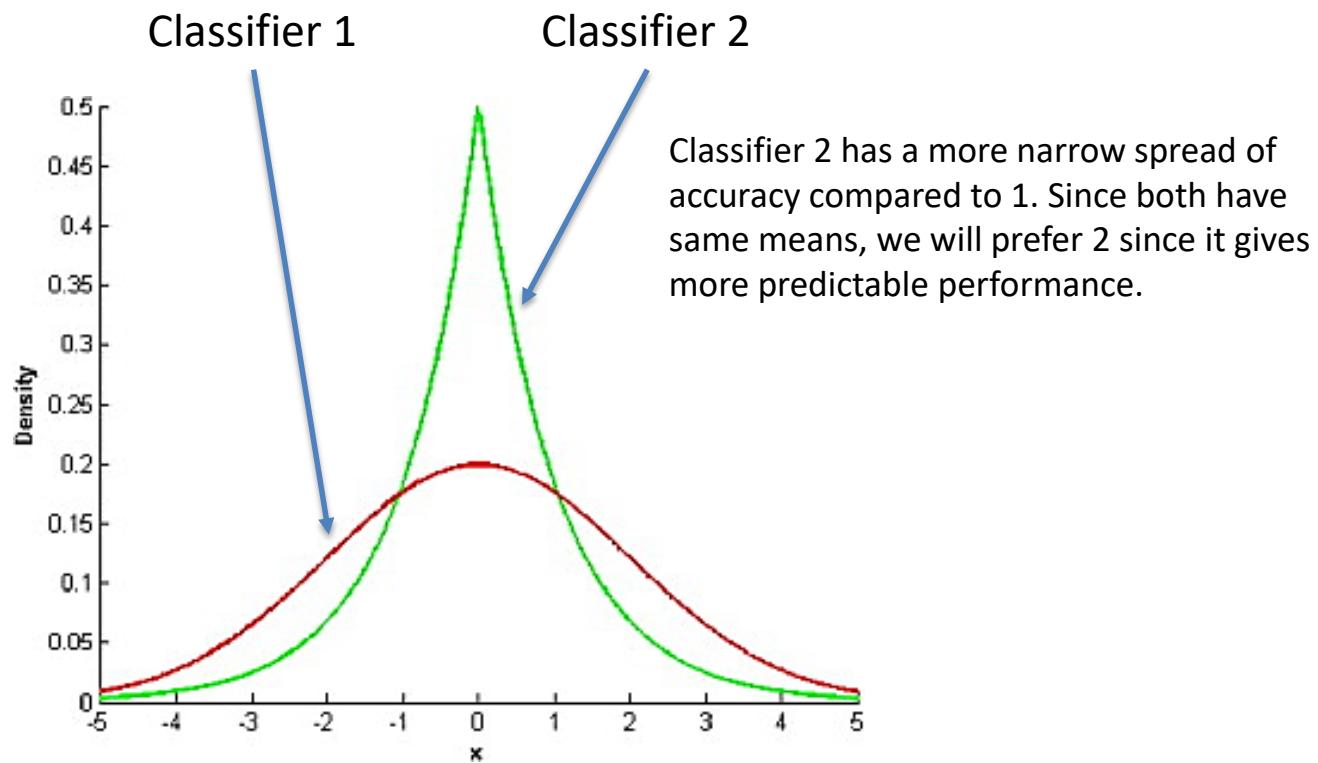
- Divide samples into k roughly equal parts.
- Each part has similar proportion of samples from different classes.
- Use each part to test other parts.
- Total up accuracy.



- The number of k parts also determines the number of times we evaluate accuracy.
- This is also referred to as k or N -fold cross-validation where $N = k$.
- $k = 5$ or 10 are popular choices.

Cross-validation

Z-normalised distribution of prediction accuracies.



So why do cross-validation?

What is the logical basis of cross validation?

Hint: **Central limit theorem**

What/ whose accuracy does it really estimate?

Do the results tell us more about the data or the classifier?

Independent validation

Instead of using cross-validation which splits the same data into many folds.

We train the model using dataset A, and then evaluate it onto dataset B (where B is produced by a completely different group).

B is never considered as part of a cross-validation evaluation.

Cross-validation and independent validation

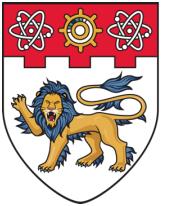
Cross-validation:

Cross-validation methods are often used to obtain estimates of classification accuracy, but both simulations and case studies suggest that, when inappropriate methods are used, bias may ensue (overestimate classifier performance).

Independent validation:

Bias can be bypassed and generalisability can be tested by external (independent) validation.

Usually used together.



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Summary

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



Summary

1. Model evaluation is complicated: the precision-recall, F-Score, ROC curves and FDR are all imperfect. So you have to exercise discretion and discern.
2. Cross-validation is commonly used as an approach for evaluating the machine learning model.
3. Be wary about the curse of dimensionality issues when you have small sample sizes and large variable sizes.

