Machine Learning - 1

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Learning Objectives

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

- Explain variables, attributes and traits.
- Explain the basics of knowledge discovery.
- Explain the mechanics of performing prediction.
- Explain the four natures of predictions; TP/FP/TN/FN.
- Explain the evaluation of classifiers and curse of dimensionality.
- Describe the differences between cross-validation and independent-validation.

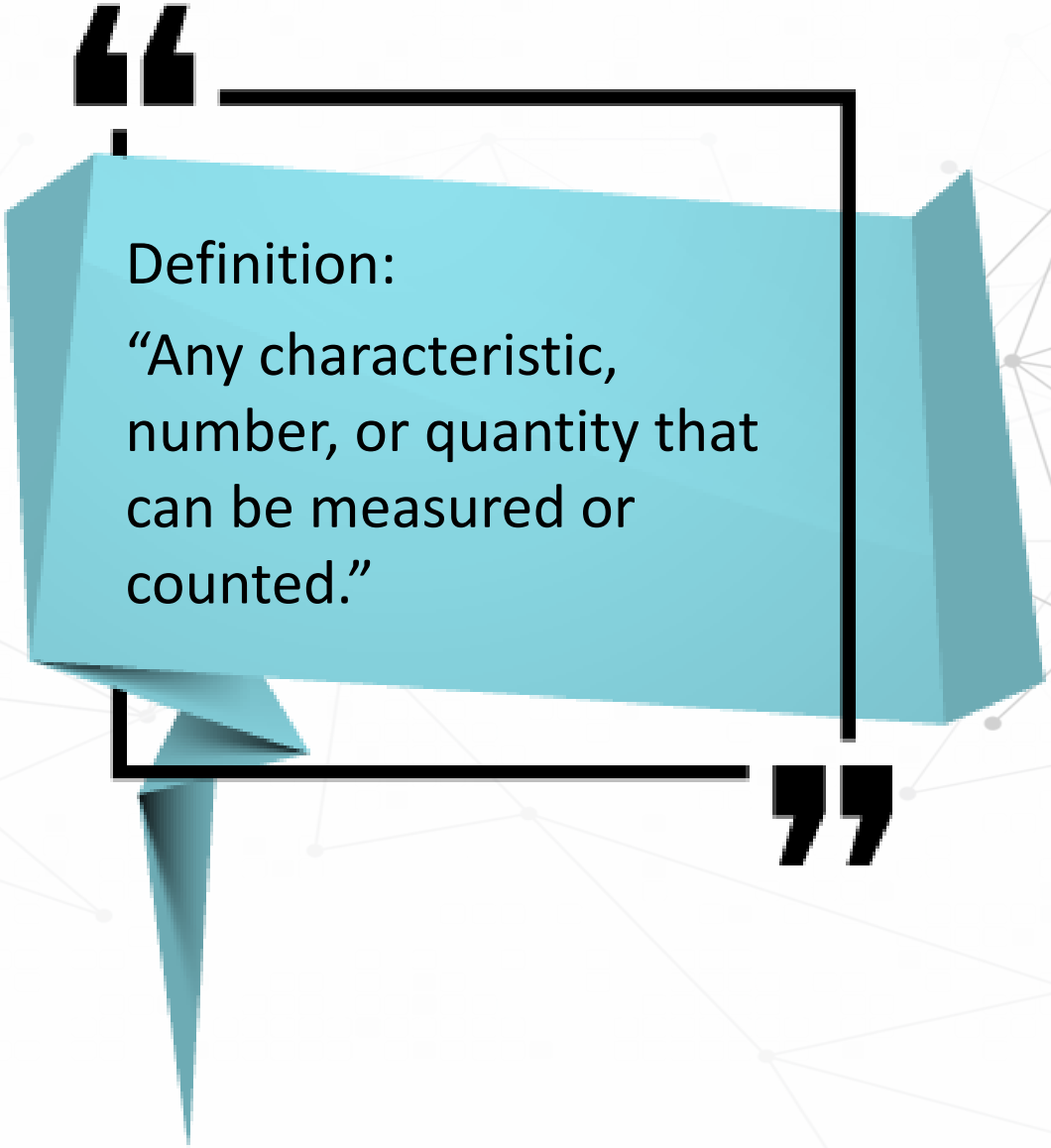
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Mechanics of Knowledge Discovery

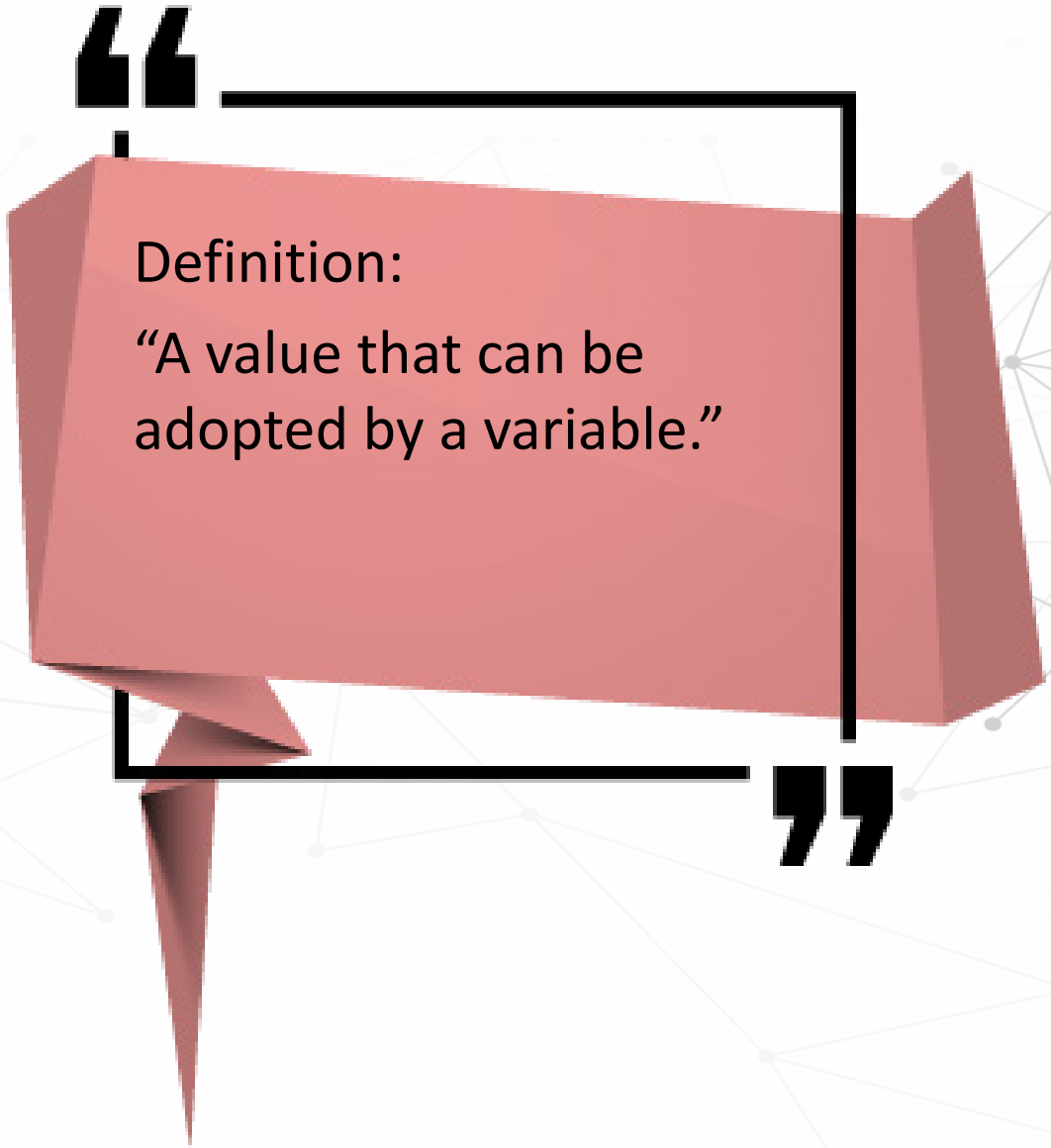
BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

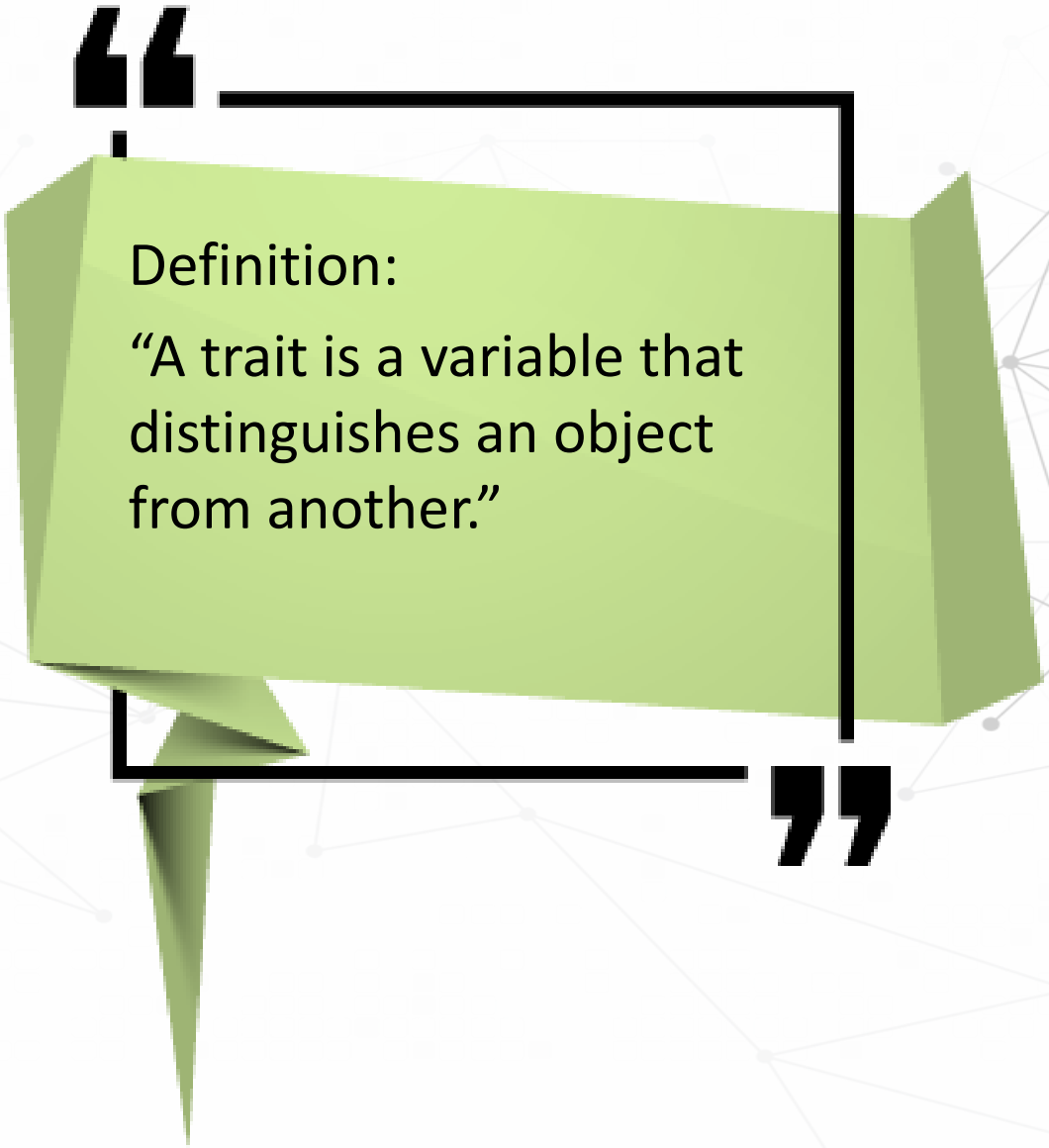
Variables/ Features



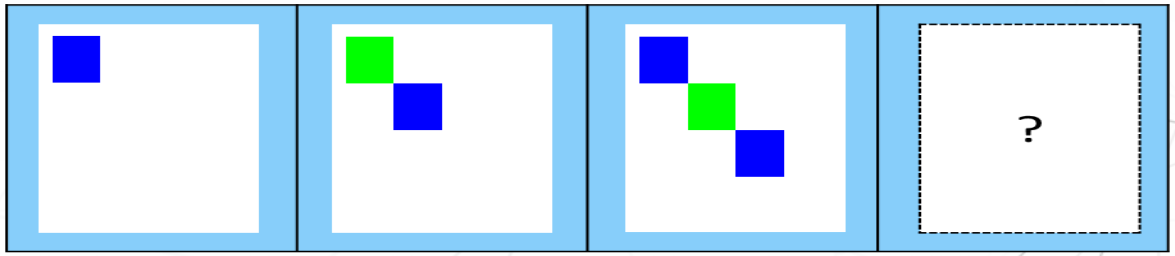
Attributes

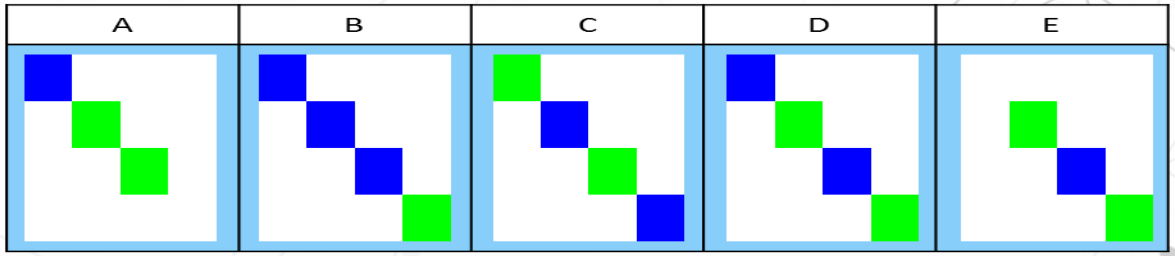


Traits


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

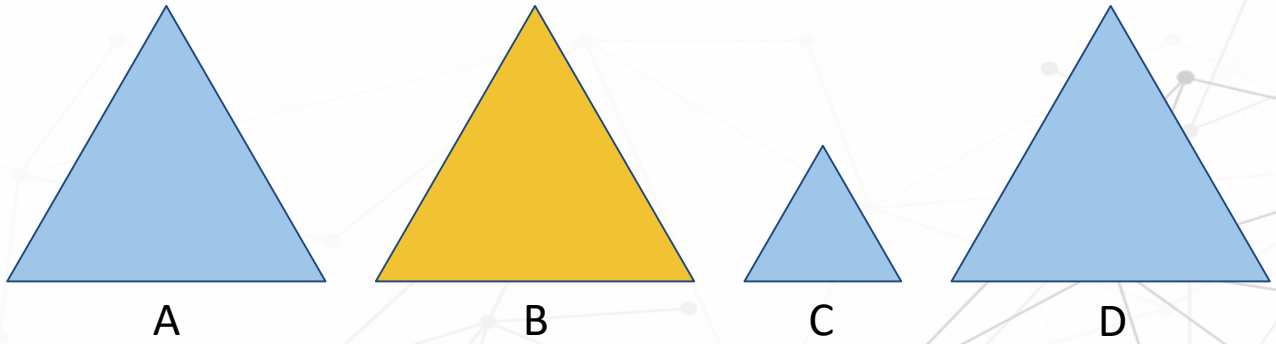
Generating Variable, Attributes, Traits





Variables	Attributes	Traits? (in this context)
Shapes	Square	No
Colours	Blue, green	Yes
Number of shapes	1, 2, 3	Yes
Position of next shape	Diagonally below/above	No

Which is the odd one out?



Which one is the odd one out?

Is shape a trait?

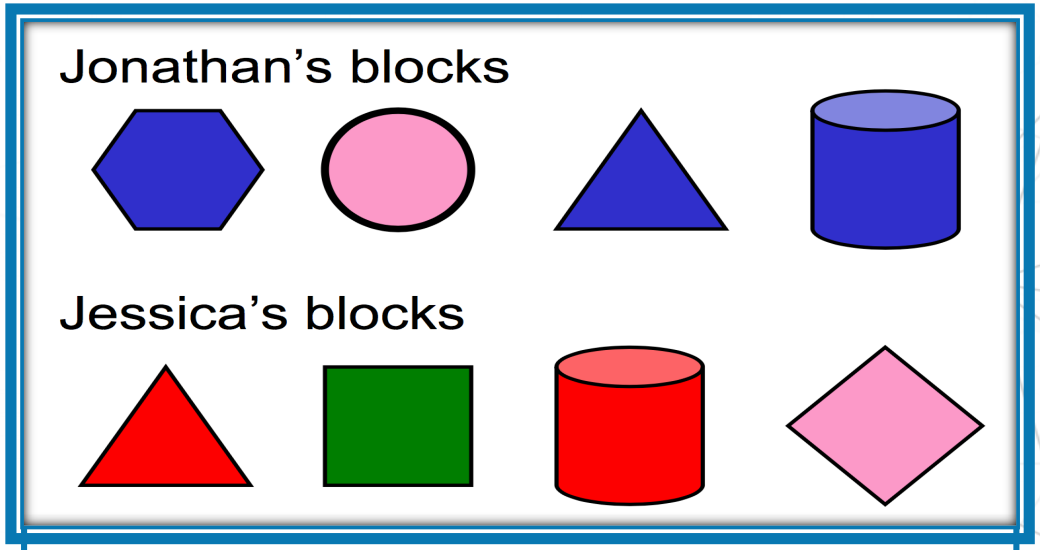
Which is the odd one out?

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.

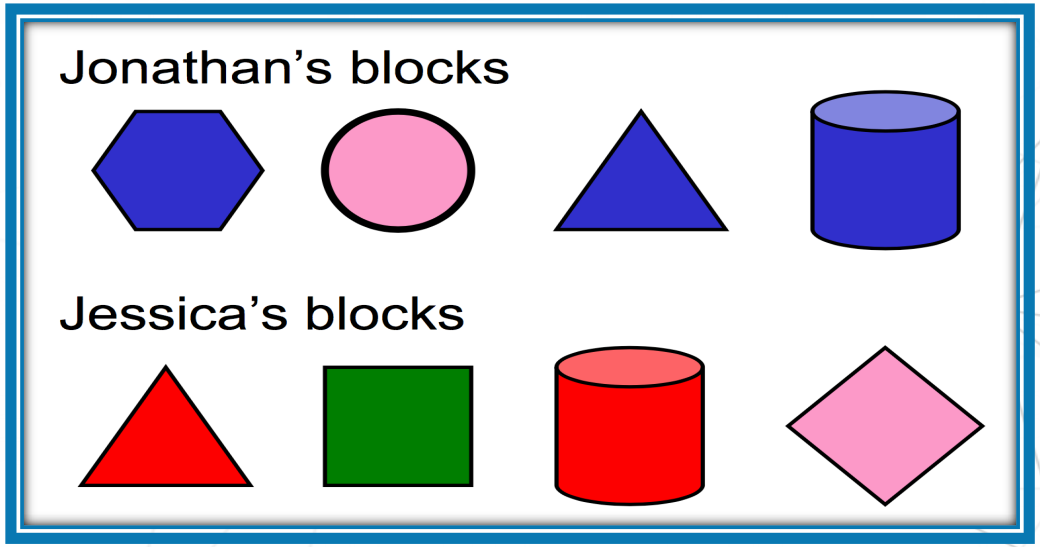
Whose block is this?



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

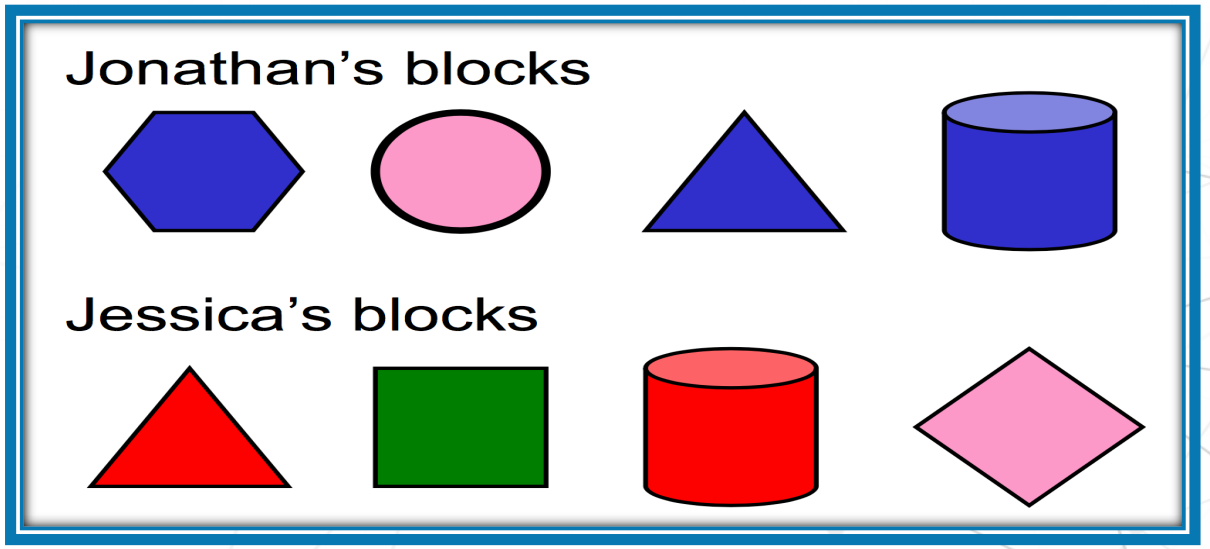
Whose block is this?

Whose block is this?



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

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

Data Gathering

- Shoulder-length Hair
- Black Eyes
- Gold-rimmed Spectacles
- Red Hairband

Variable Generation

- Hairstyle
- Eye Colour
- Facial Accessories
- Hair Accessories

How did you learn to recognise your mother?

Variable Integration

- That lady with shoulder-length hair and gold-rimmed spectacles in my mother!

Variable Selection

- Shoulder-length Hair (Hairstyle)
- Gold-rimmed Spectacles (Facial Accessories)

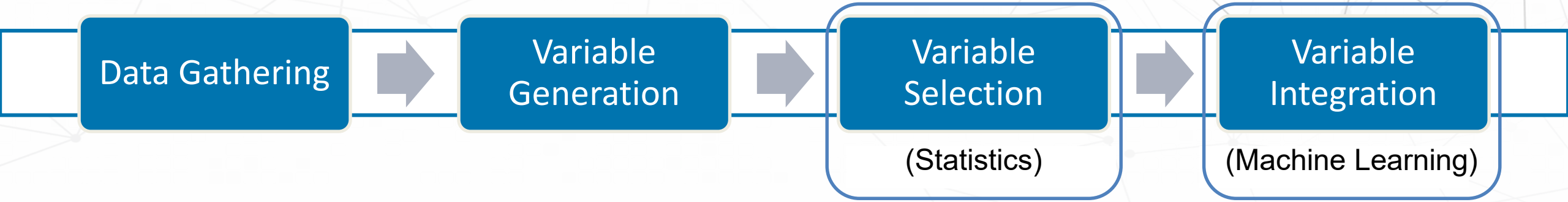
Variables, Attributes, Traits

- 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

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Prediction and Statistics

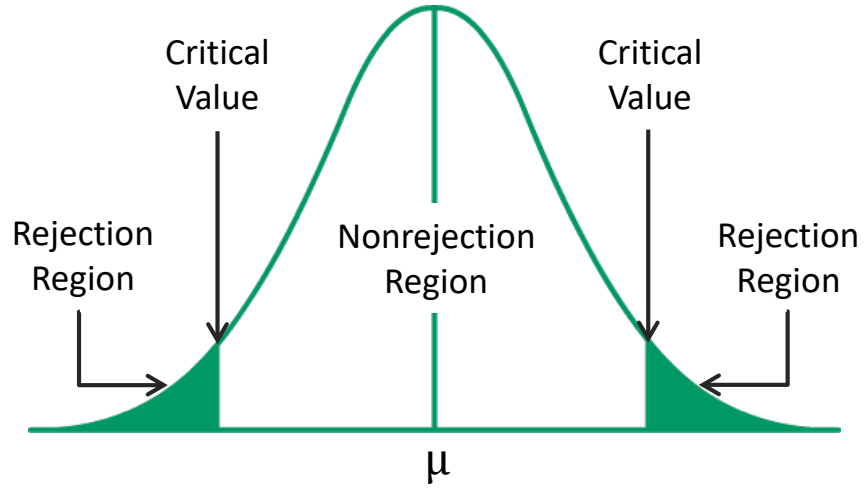
- Prediction is estimating unknown data based on extrapolation of extracted data during knowledge discovery.
- Predictions are grounded on statistics:
  - Distinguishing factor (between H0 and H1): traits → describes significantly different attributes.
  - If a data's attribute is significantly different 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 → person likely to be diabetic



SUGAR LEVEL	FASTING PLASMA GLUCOSE
DIABETES	≥ 7.0 mmol/L (≥ 126 mg/dL)
PRE-DIABETES	6.1 – 6.9 mmol/L (110 – 125 mg/dL)
NORMAL	< 6.1 mmol/L (< 110 mg/dL)

Source: https://www.healthxchange.sg/diabetes/pre-diabetes/pre-diabetes-take-charge-of-your-health
http://www.onekobo.com/Articles/Statistics/03-Hypotheses/Stats3%20-%2010%20-%20Rejection%20Region.htm

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

BS3033 Data Science for Biologists

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

		Predictiion	
		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

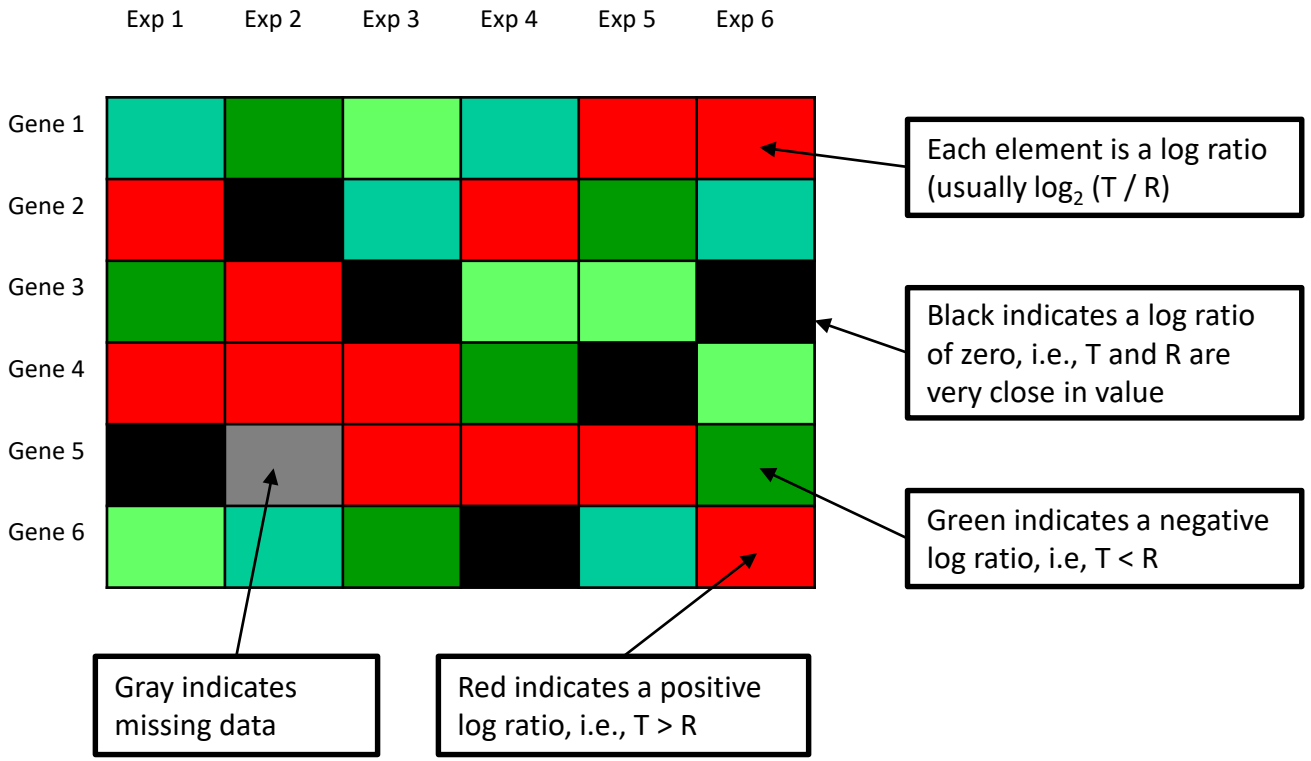
BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Testing many Hypotheses

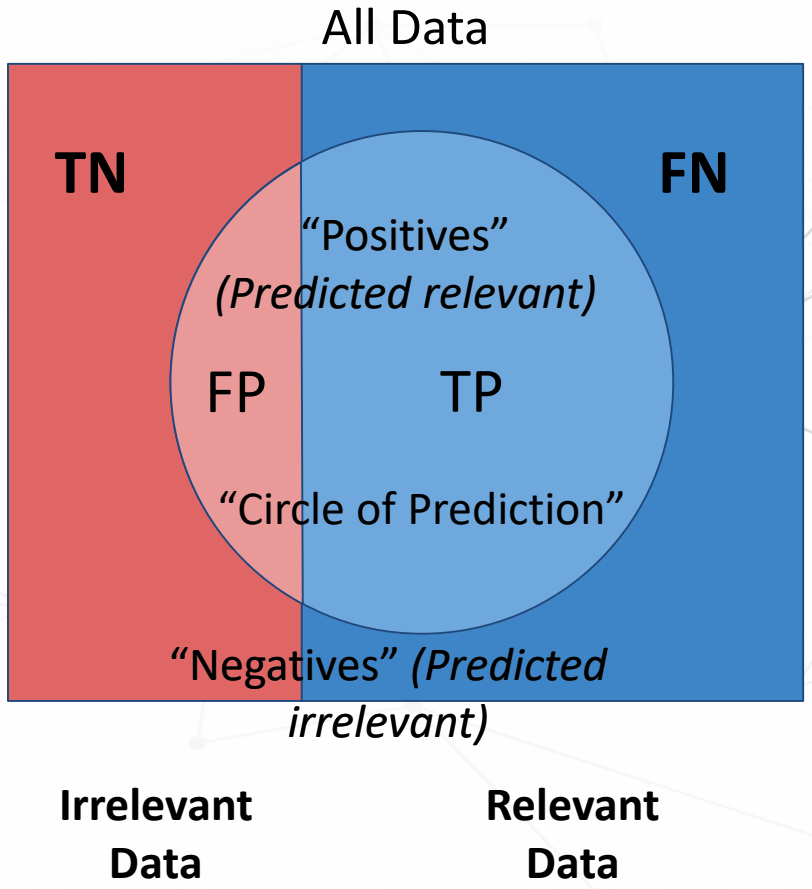
- You may test many hypotheses at once.
- For example, in a microarray experiment, you are effectively performing thousands of hypothesis testing at once (per gene).
- Prediction on each gene falls into 1 of 4 outcomes.

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

BS3033 Data Science for Biologists

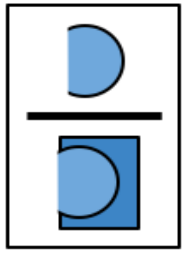
Dr Wilson Goh
School of Biological Sciences

Metrics for performance evaluation

Sensitivity/ Recall

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

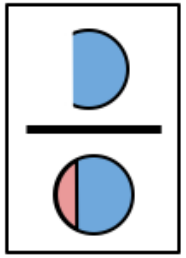
$$R = \frac{TP}{(TP + FN)}$$



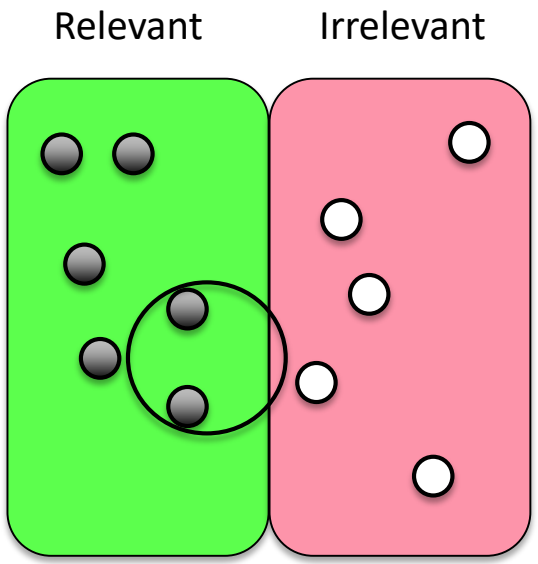
Precision/ Positive Predictive Value (PPV)

What proportion of predicted values are true positive?

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

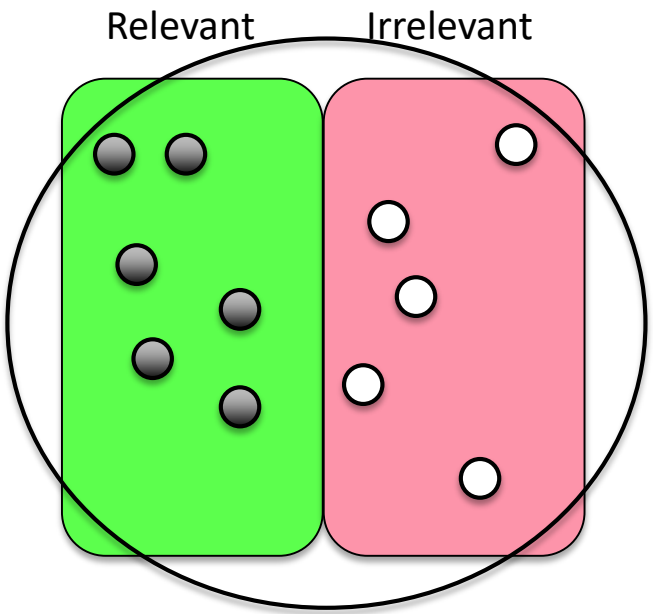


Precision and Recall Work Against Each Other



Stringent p-value Cutoff

Precision ⇅↑
Recall ↓

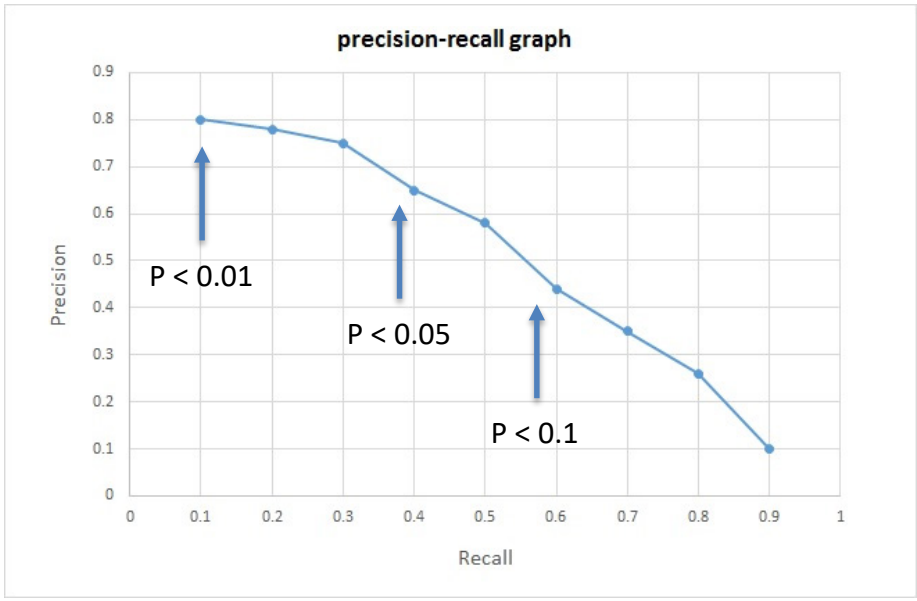


Loose p-value Cutoff

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.

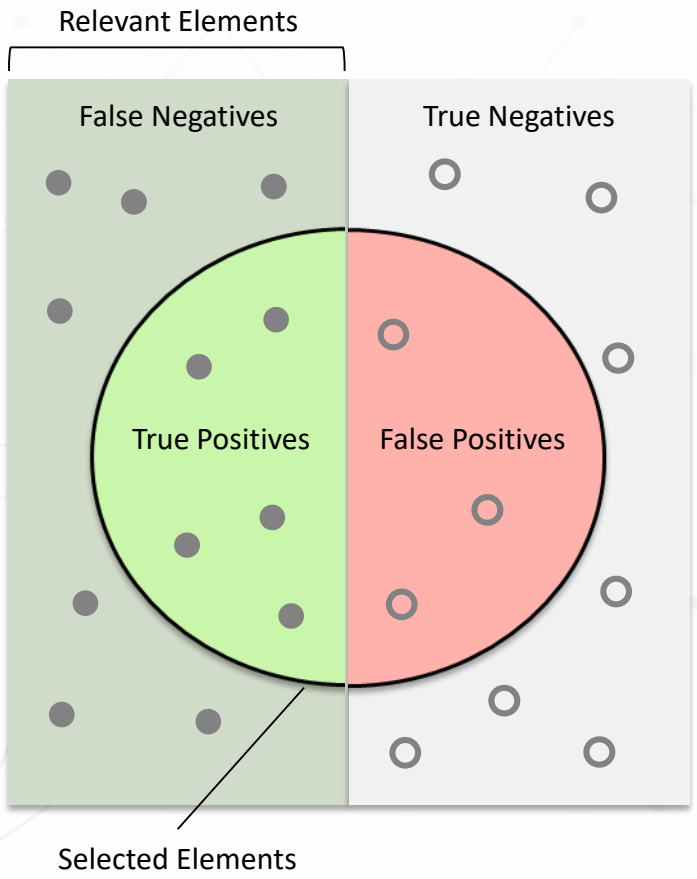
$$F = 2\left(\frac{precision\ X\ recall}{precision + recall}\right)$$

- 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



$$Accuracy = \frac{No.\,of\ correct\ predictions}{No.\,of\ predictions}$$

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Case Study 1: Limitations of the Accuracy Scoring Metric

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Case Study 1: Accuracy

- Balanced scenario (Half are positive, half are negative).

Classifier	TP	TN	FP	FN	Accuracy
A	25	25	25	25	50%
B	50	25	25	0	75%
C	25	50	0	25	75%
D	37	37	13	13	74%

- B,C,D are all better than A.
- B is good at finding TPs, but it does so at the expense of also making a lot of false predictions (FPs).
- C is more conservative but makes less false predictions. But ends up with a lot of false negatives.
- D is intermediate of B and C.

- Unbalanced scenario (50 are positive, 150 are negative).

Classifier	TP	TN	FP	FN	Accuracy
A	25	75	75	25	50%
B	0	150	0	50	75%
C	50	0	150	0	25%
D	30	100	50	20	65%

- Which do you think should be the best method?
- Should it be B?
- What do you think is B's prediction strategy?
- **High accuracy is meaningless if population is unbalanced.**

Case Study 1: Precision, Recall and F-score

Classifier	TP	TN	FP	FN	Accuracy
A	25	75	75	25	50%
B	0	150	0	50	75%
C	50	0	150	0	25%
D	30	100	50	20	65%

- What are the recall and precision of A, B, C and D?
- Is B still better than A, C, D?

Case Study 1: Precision, Recall and F-score

Classifier	TP	TN	FP	FN	Accuracy
A	25	75	75	25	50%
B	0	150	0	50	75%
C	50	0	150	0	25%
D	30	100	50	20	65%

Precision	Recall	F-score
0.25	0.50	0.33
0	0	NA
0.25	1	0.4
.38	0.60	0.47

Case Study 1: Precision, Recall and F-score

- B is doing well because it is not making any predictions (if you don't make any predictions, you don't make any mistakes). There are 50 that are correct. B picks up none of them.

- But because there are so many that are negatives, by not making any predictions, B is effectively ignoring all these negatives, effectively bolstering its accuracy score (it is because of the TNs).

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Case Study 2: Decision Making Process of a Machine Learning Classifier

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Decision Making Process of a Machine Learning Classifier

The machine learning algorithm must determine a reasonable cutoff for deciding its prediction decision.

This can involve:

- Computing scores for both supporting and non-supporting evidences
- Deriving a metric for considering both evidences
- Establish a decision rule
- Establish a threshold if the decision rule is quantitative

Decision Making Process of a Machine Learning Classifier

Given a test sample S:

- Compute scores p(S), n(S)
- Predict S as negative if p(S) / n(S) < t
- Predict S as positive if p(S) / n(S) > t

t is the decision threshold of the classifier. Changing t affects the recall and precision, and hence accuracy, of the classifier.

The decision threshold, t can be a pre-determined value, a test-statistic or a p-value (in this case, it is a pre-determined value, a golden ratio).

Decision Making of a Classifier?

S	P (S)	N (S)	Actual Class	Predicted Class @t = 3	Predicted Class @ t = 2
2	0.961252	0.038748	P	P	P
3	0.435302	0.564698	N	N	N
6	0.691596	0.308404	P	N	P
7	0.180885	0.819115	N	N	N
8	0.814909	0.185091	P	P	P
10	0.887220	0.112780	P	P	P
			Accuracy	5/6	6/6
			Recall	3/4	4/4
			Precision	3/3	4/4

Recall that…

- Predict S as negative if p(S) / n(S) < t
- Predict S as positive if p(S) / n(S) ≥ t

Decision Making of a Classifier?

In this case, it is better to just take t=2.

You get better recall while preserving precision.

You also have better accuracy.

It should be clear by now that cutoffs can be moved to optimise these metrics.

When optimisation is possible, you should not limit yourself to just one cutoff → it is plain silly!

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

What Makes a Good Prediction?

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

When to use which one?

In many cases, accuracy is an obvious measure. But it conveys the right intuition only when the positive and negative populations are roughly equal in size.

Recall and precision together form a better measure. But what do you do when A has better recall than B and B has better precision than A?

F-score

Take the harmonic mean of recall and precision.

$$F = \frac{2 * recall * precision}{recall + precision}$$ (wrt positives)

Classifier	TP	TN	FP	FN	Accuracy	F-measure
A	25	75	75	25	50%	33%
B	0	150	0	50	75%	Undefined
C	50	0	150	0	25%	40%
D	30	100	50	20	65%	46%

Does not accord with intuition: C predicts everything as +ve, but still rated better than A.

Not a perfect solution. Not necessarily better than accuracy always.

What Makes a Good Prediction?

An ideal prediction should be able to:

1. Include ALL the positive (relevant) data in the prediction:
   - 'Circle of prediction' lies fully in the 'relevant data' region
   - High true positive rate
   - **High sensitivity/recall**

2. Exclude ALL the negative (irrelevant) data from the prediction:
   - 'Circle of prediction' lies fully out of the 'irrelevant data' region
   - Low false positive rate
   - **High specificity**

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ROC Curves

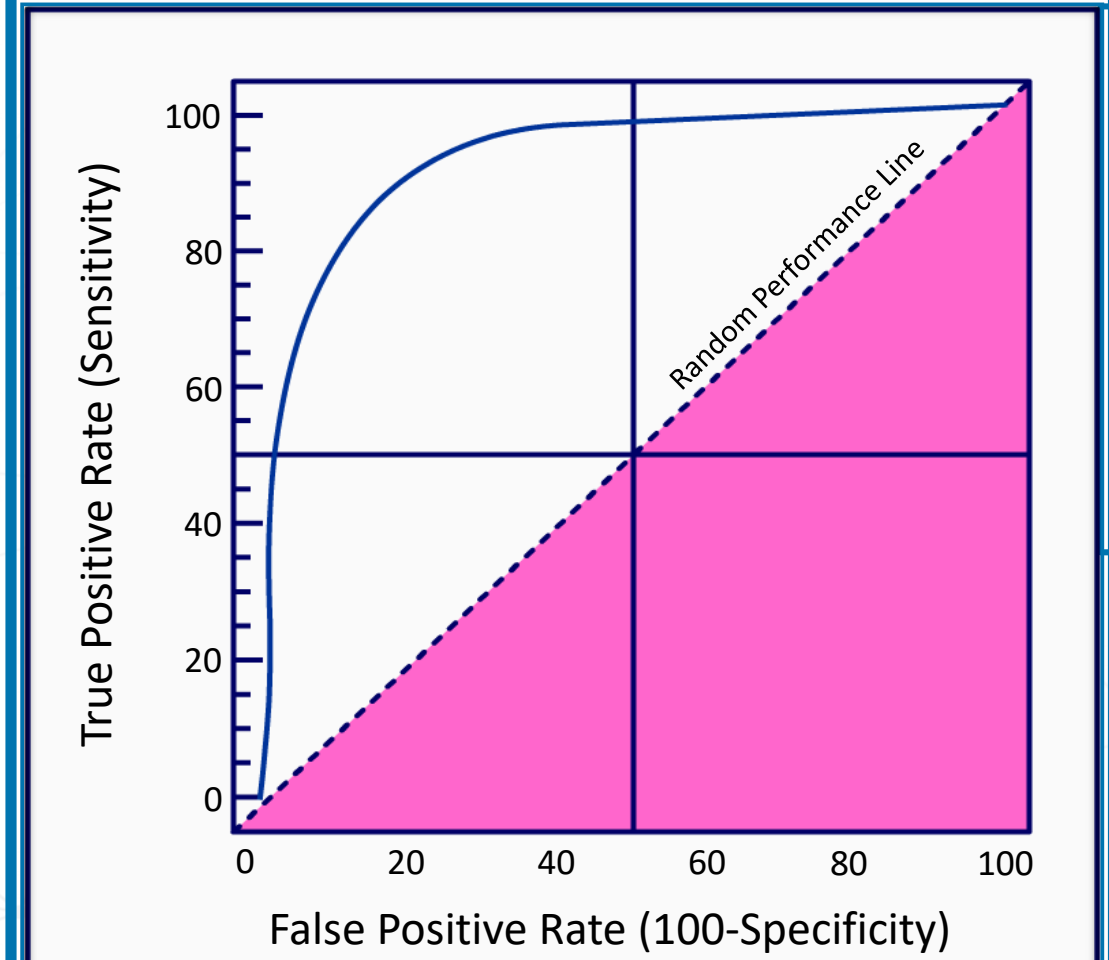
BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Threshold-free Evaluation of Performance

- The accuracy, F-score, precision and recall are all dependent on some arbitrary cutoff (e.g. at p-value below 0.05).
- What if we want to look at overall performance?

Receiver Operator Characteristic (ROC) Curve



Perfect situation!
(100% sensitivity, 100% specificity)

The closer the curve is to the upper left corner, the higher the overall accuracy of the test!

Interpreting ROC/ AUC Curves

The **ROC curve** is created by plotting the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)** at various threshold (e.g. p-value) settings.

The **TPR** is also known as **sensitivity or recall** in machine learning.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

The **FPR** is also known as the fall-out OR **(1 − specificity)**.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Retrieved From: "Detector Performance Analysis Using ROC Curves - MATLAB & Simulink Example". www.mathworks.com. Retrieved 11 August 2016.

Area Under (ROC) Curve

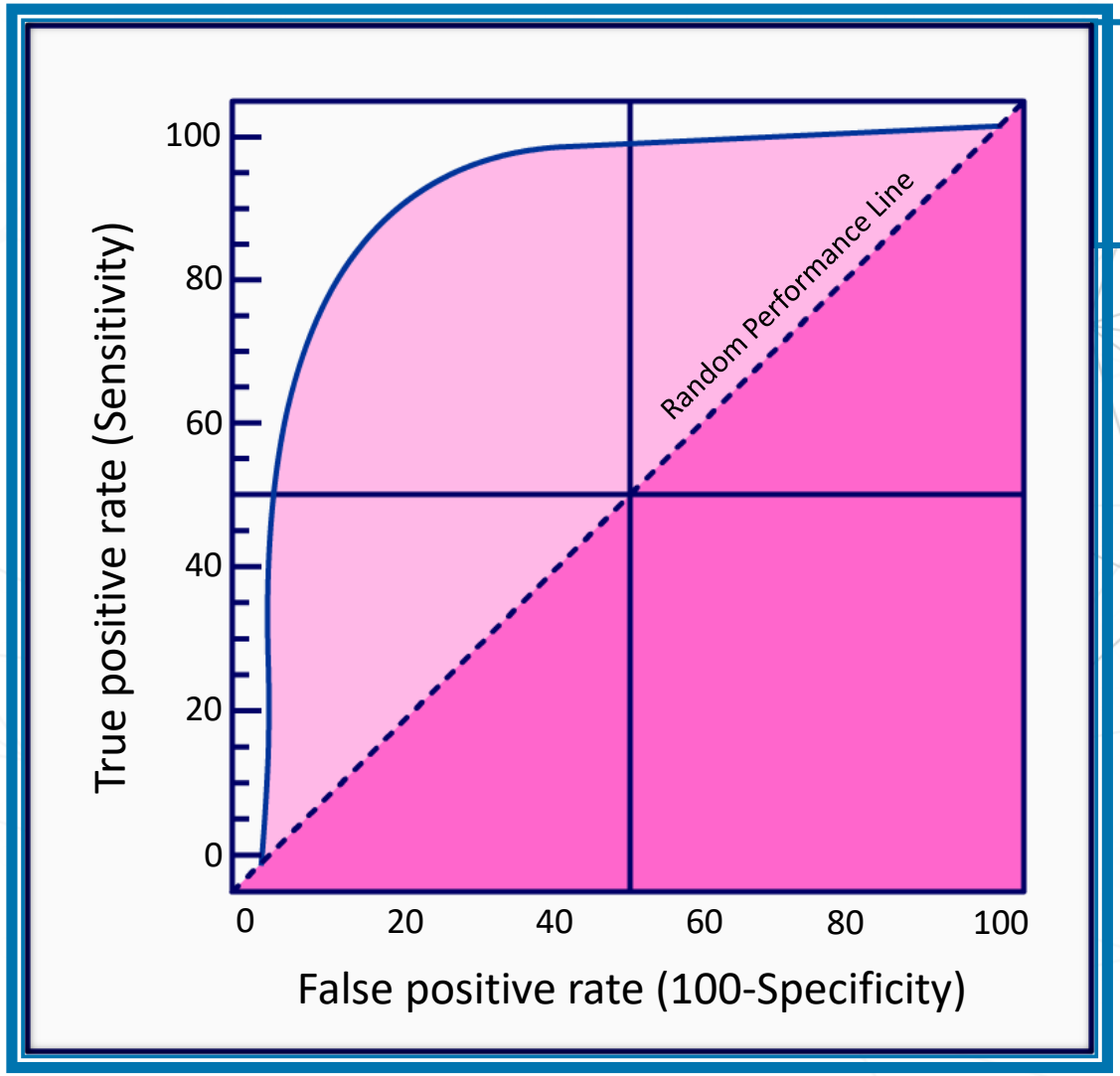
A measure of how well a parameter can distinguish between two diagnostic groups.

An ROC curve that reaches the “perfect situation” point (top-left) will have a greater area under the curve.

Therefore, the AUC/AUROC value can describe how good the variable is at predicting the data accurately:

- Capture most of the relevant data
- Exclude most of the irrelevant data

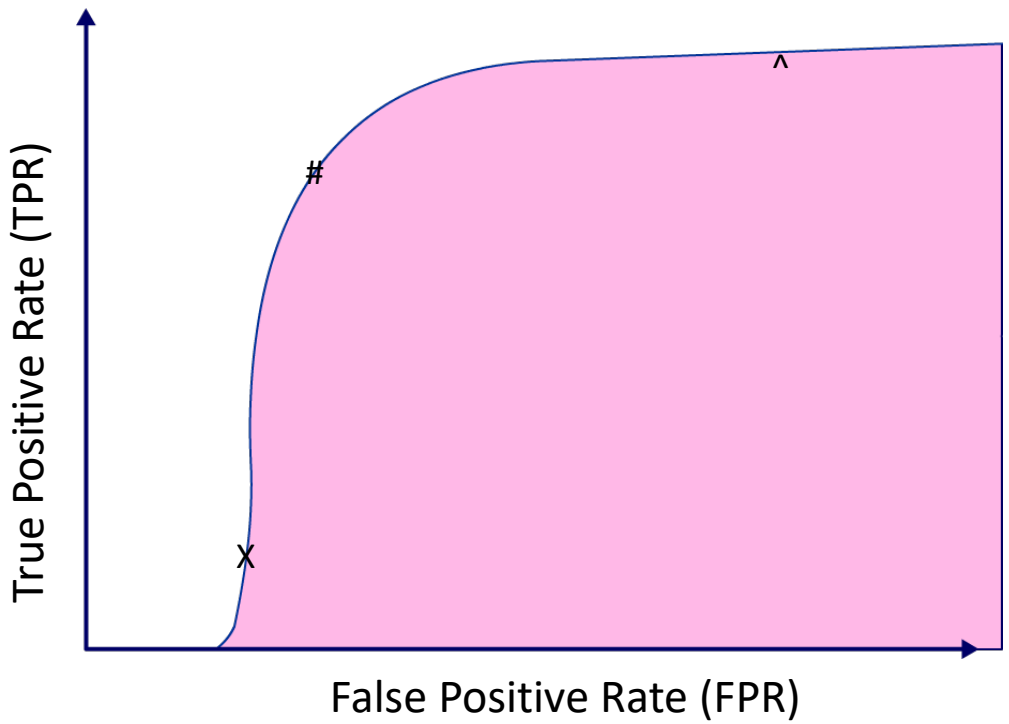
ROC Curve



Perfect situation!
(100% sensitivity,
100% specificity)

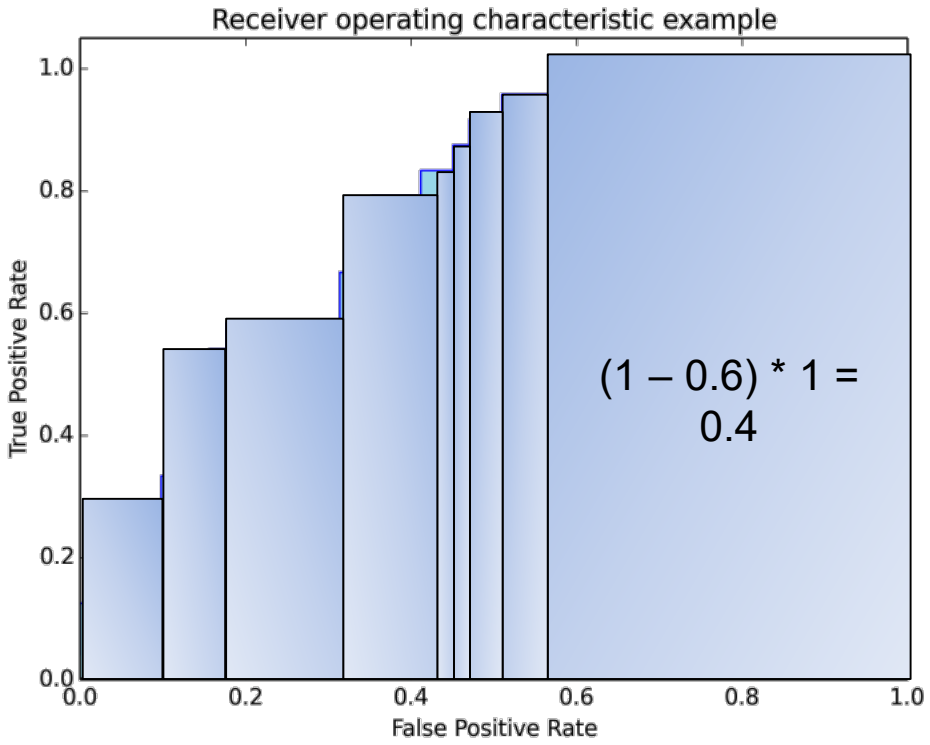
Plotting the ROC

p-value	TPR	FPR
0.01	X	X
0.1	#	#
1.0	^	^



Area Under Curve (AUC)

More rightfully called Area Under the ROC curve (AUROC).



- The blue area corresponds to the AUROC. The dashed line in the diagonal is expected performance due to random chance (so we have to be better than chance).

- Total area = 1 x 1 = 1
- Half area under the diagonal = ½ = 0.5
- One simple way to get the AUROC is to simply calculate the area using simple length x breadth. But of course one may use calculus.

A Question from your Peers

- Precision and Recall (Sensitivity) seems to be mutually exclusive. If we want to increase the sensitivity we need to reduce the amount of false negatives.
- However, by reducing the amount of false negatives we risk increasing the amount of false positives in the process and thus compromising on the precision of our results.
- Thus there seems to be a trade-off between precision and sensitivity. This dilemma is encapsulated in the ROC Curve.

Wil's Thoughts

- The ROC curve does not specifically look at the trade-off between precision and recall. There is such a thing called the precision-recall curve, which is used for this purpose (you've actually seen this in the earlier slides).

- The ROC looks at FPR vs TPR. Think of the FPR as the potential of making a type I error (mistakenly calling a negative a positive). The TPR relates to the proportion of calling out all the correct answers (it is directly related to power → or the tendency to not make Type II error). If you think of it this way, it is in fact finding the sweet spot between Type I and Type II error.

What do you think?

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

False Discovery Rates

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

False Discovery Rate

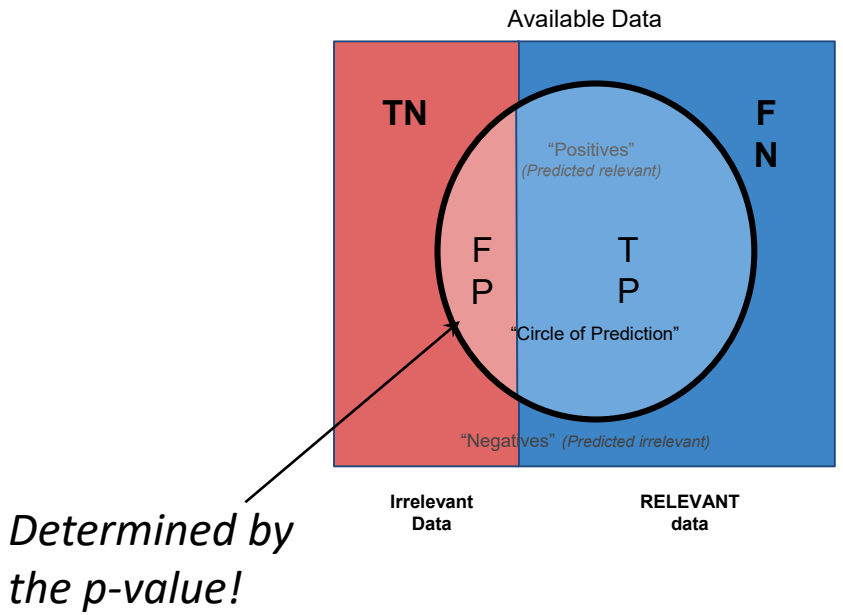
The **False Discovery Rate (FDR)** is a measure of the amount of False Positives among the total predictions made.

FDR = 1 - Precision

FDR can be changed by changing the p-value! Because by increasing the p-value, you are increasing the area predicted as relevant.

So if number of samples in
False Negative > True Negative, FDR will decrease!

(darker blue region) (darker red region)



Case Study: FDR in Proteomics

False Discovery Rate (FDR): **The expected fraction of false positives among the significant test statistics. (FP/FP+TP)**. Compare this against the false positive rate which is FP/(FP+TN).

Score	Type	
7.5	Correct	
7.2	Correct	
6.9	Correct	
6.8	Correct	
6.7	Incorrect	
6.5	Correct	
6.4	Correct	
6.4	Correct	
6.3	Incorrect	
6.1	Correct	
6.0	Incorrect	Threshold
5.9	Correct	
5.7	Incorrect	
…	…	

So how do we look at this? Let's say we have a set of PSM scores and decide to draw the line at 6, i.e., we accept all PSMs with scores > 6.

Let's also assume we have perfect knowledge of correct and wrong matches. We note that 10 PSMs are retained. Of these, 2 are wrong. So the FDR is therefore

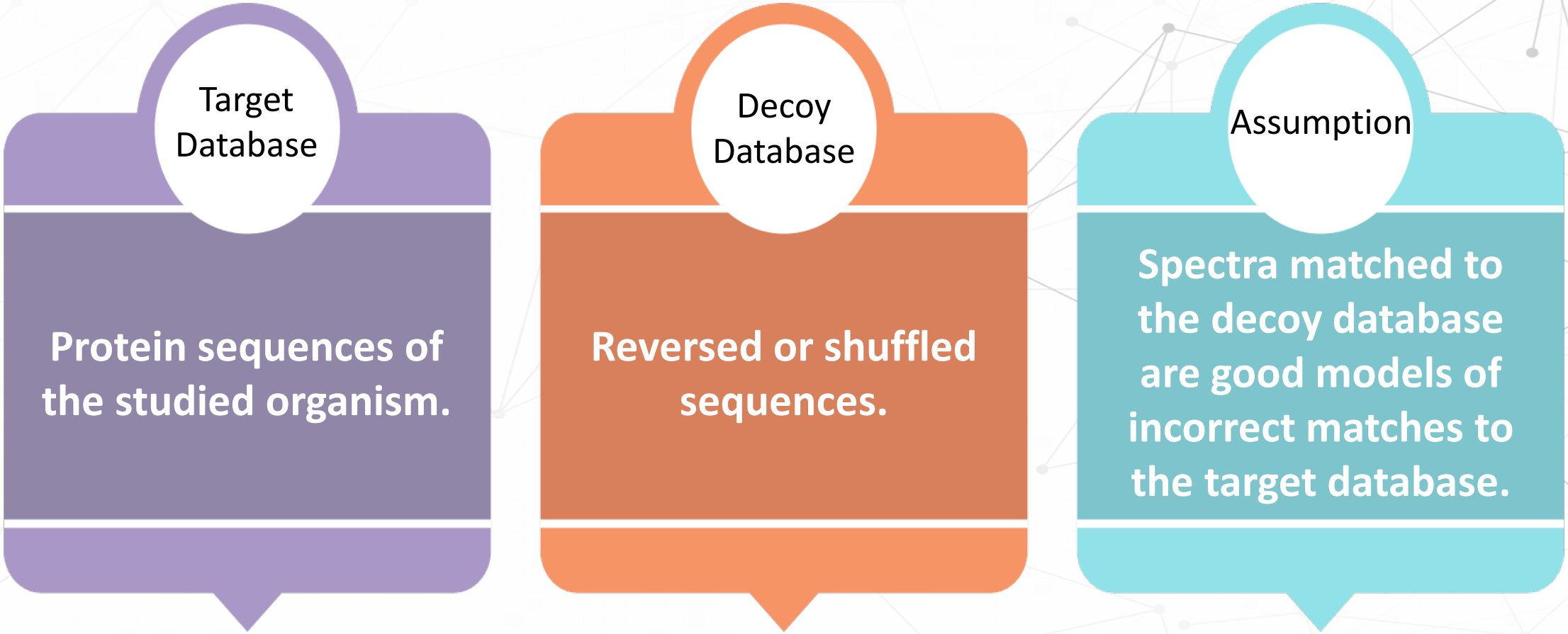
FDR = 2/10 = 20% = 0.2

This seems great. But in reality, we don't know which ones are wrong. This is similar to the null problem in p-value generation. So how do we create something which we know to be wrong or sure?

Source: Benjamini & Hochberg, JSTOR, 1995; Storey & Tibshirani, PNAS, 2003

False Discovery Rate

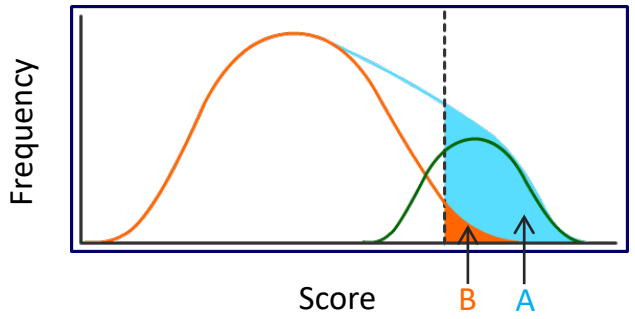
The target-decoy analysis. Estimating FDR: How to purposely create your incorrect PSMs.



In other words, all matches to decoy are false positives.

FDR Estimation Based on Decoy

No Decoy



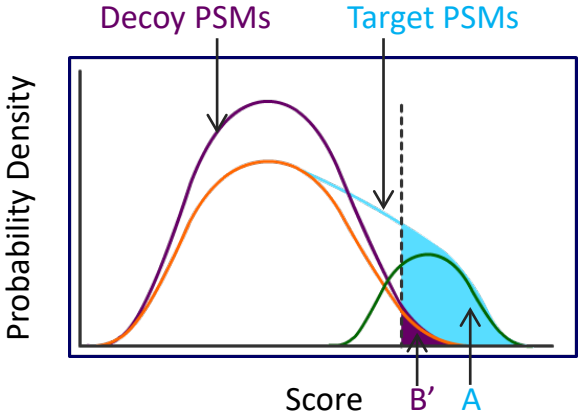
$$FDR = \frac{B}{A}$$

Combined searches: Target and decoy database are searched together.

$$\widehat{FDR} = \{\#decoys\ over\ threshold / \#targets\ over\ threshold\}$$

i.e., $\pi 0$ is 1. Simpler, since estimating $\pi 0$ can be tricky.

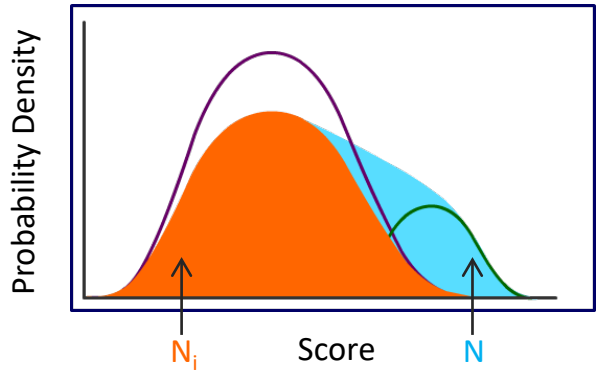
With Decoy



$$FDR = \frac{B}{A} = \frac{\pi_0 B'}{A}$$

π_0 is the fraction of incorrect target PSMs among target PSMs.

Target-decoy Analysis



$$FDR = \frac{B}{A} = \frac{\pi_0 B'}{A}$$

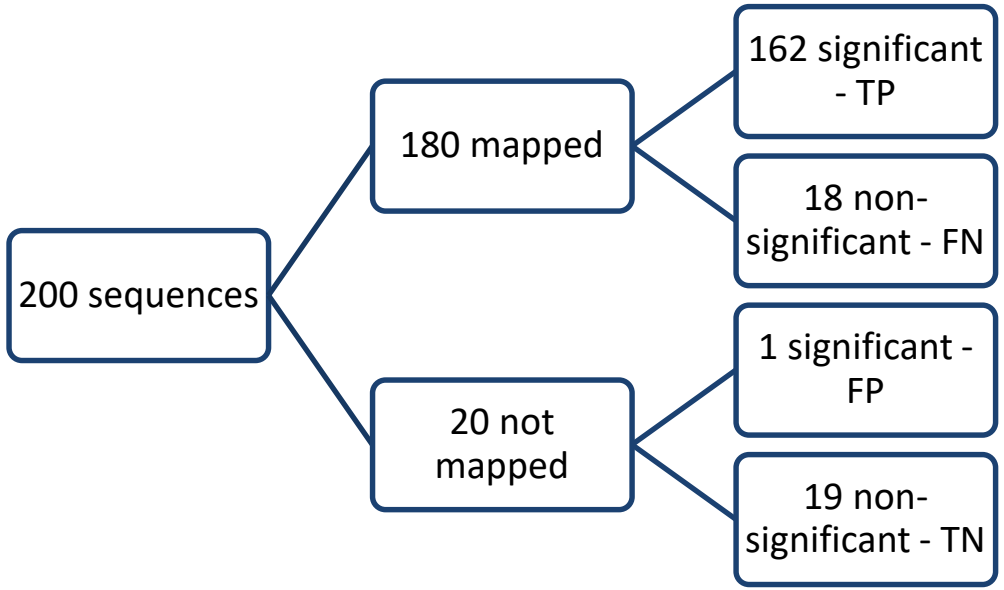
$$\pi_0 = \frac{N_i}{N}$$

π_0 is the fraction of incorrect target PSMs among target PSMs.

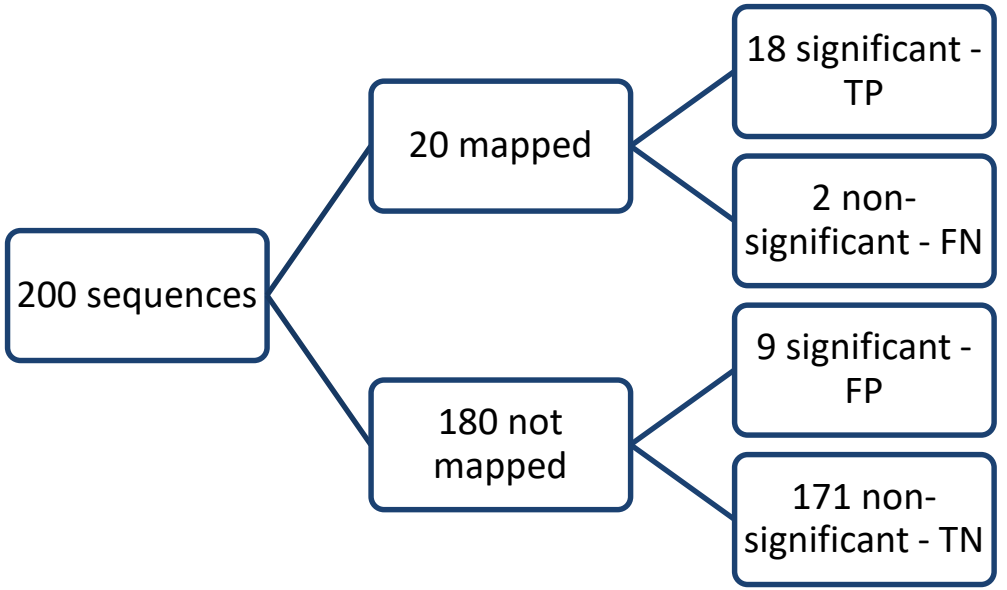
The False Discovery Rate

The FDR relates to the proportion of errors amongst predictions. It is equals to 1 – precision.

False Discovery rate = FP/(FP+TP)



1/163



9/27=1/3

The FDR is sensitive to the proportion of true features in the data.

Who cares about the FDR?

Why use the FDR when you have p-values?

- You can never be exactly sure that a p-value of 0.01 means exactly that there is an error of 1 in 100. This is the general assumption from a theoretical distribution. It may not hold true in your data.
- Use of the FDR allows you to infer from your data score distributions the corresponding cutoff where you expect to have a set error rate (1 - precision).
- The FDR is also more sensitive than the p-value when considering many variables simultaneously (requires multiple-test correction).
- Why is it that you do not need to do multiple test correction in FDR estimation?

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Cross-validation and Independent Validation

BS3033 Data Science for Biologists

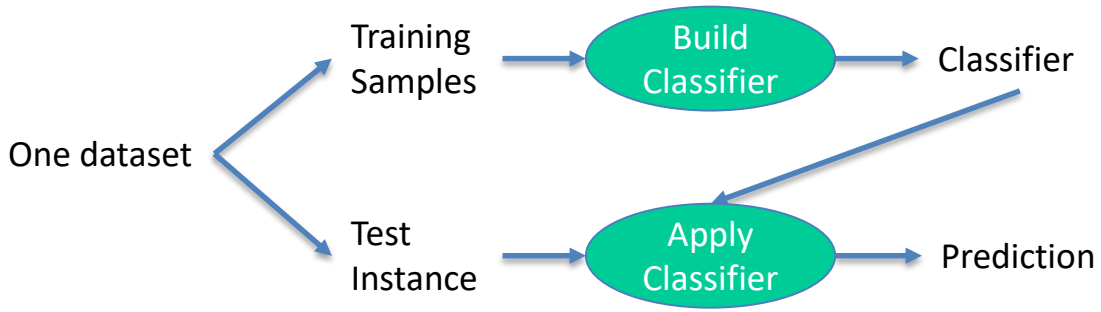
Dr Wilson Goh
School of Biological Sciences

Working with Incomplete Information

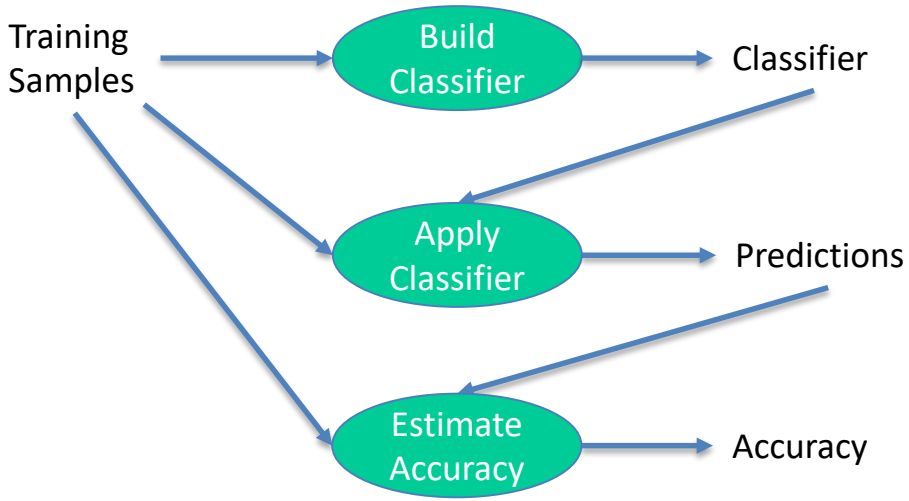
- By now, you might appreciate that in many cases, we do not know which are the relevant or irrelevant variables in real data.
- In which case, we cannot directly calculate precision/recall, etc.
- However, we do know the class labels of our samples (e.g. normal and disease).
- We want to know which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

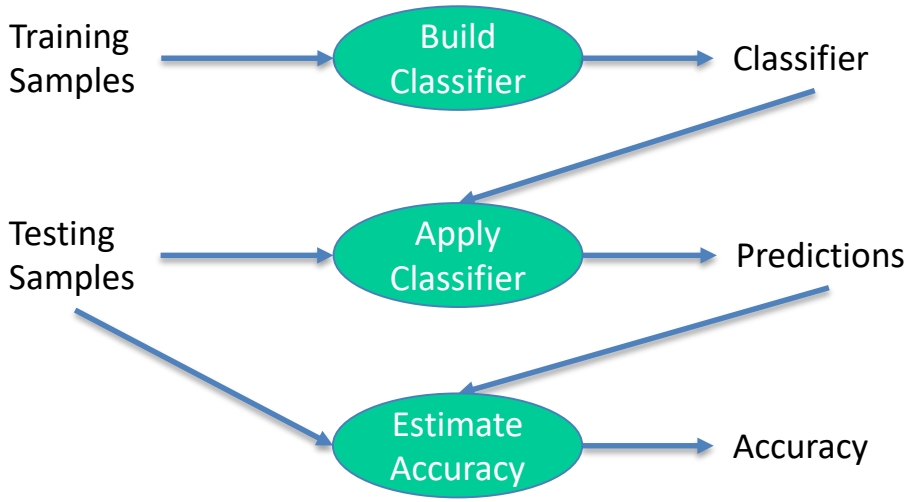


Cross-validation (The Wrong Way)



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?

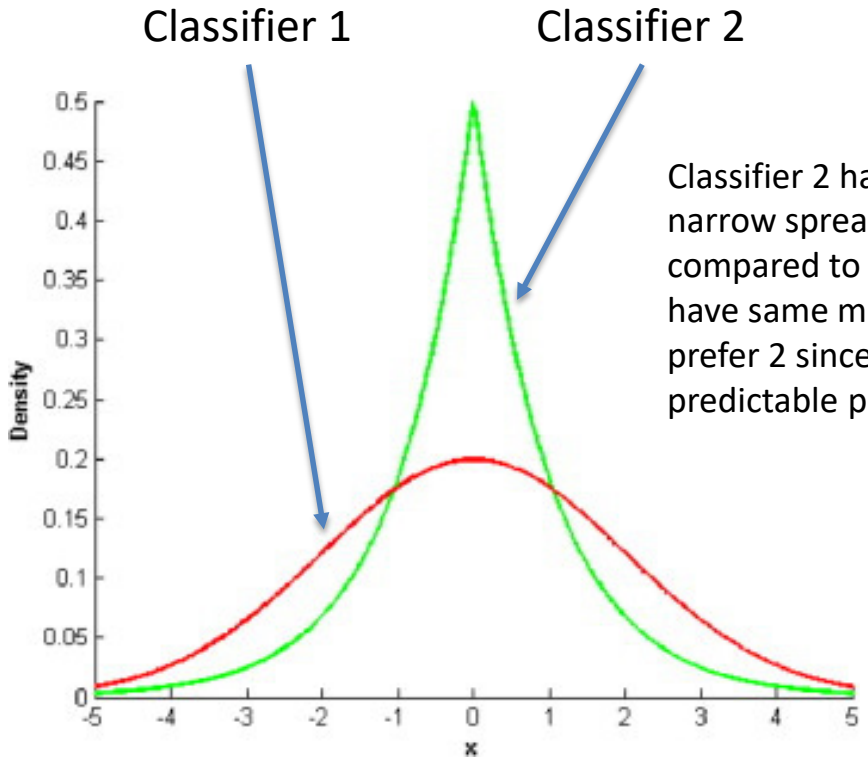
Cross-validation

- 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.



Classifier 2 has a more narrow spread of accuracy compared to 1. Since both have same means, we will prefer 2 since it gives more predictable performance.

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.

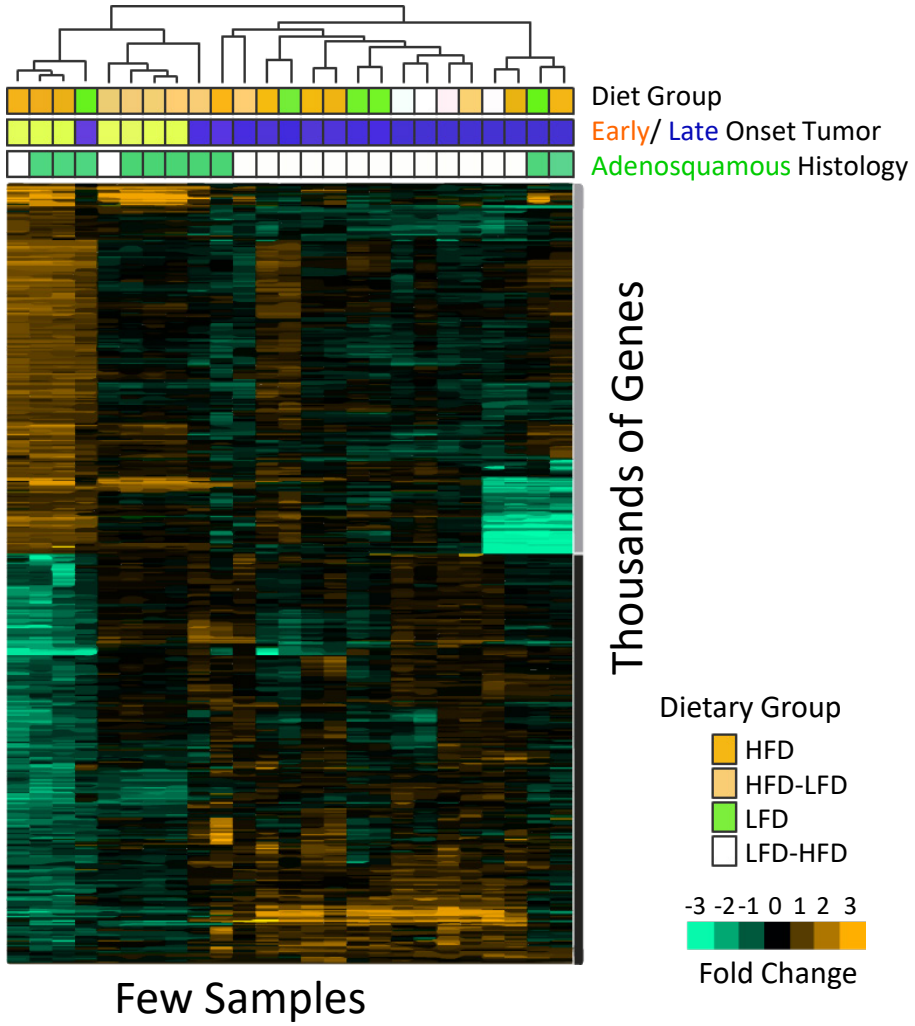
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Curse of Dimensionality

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

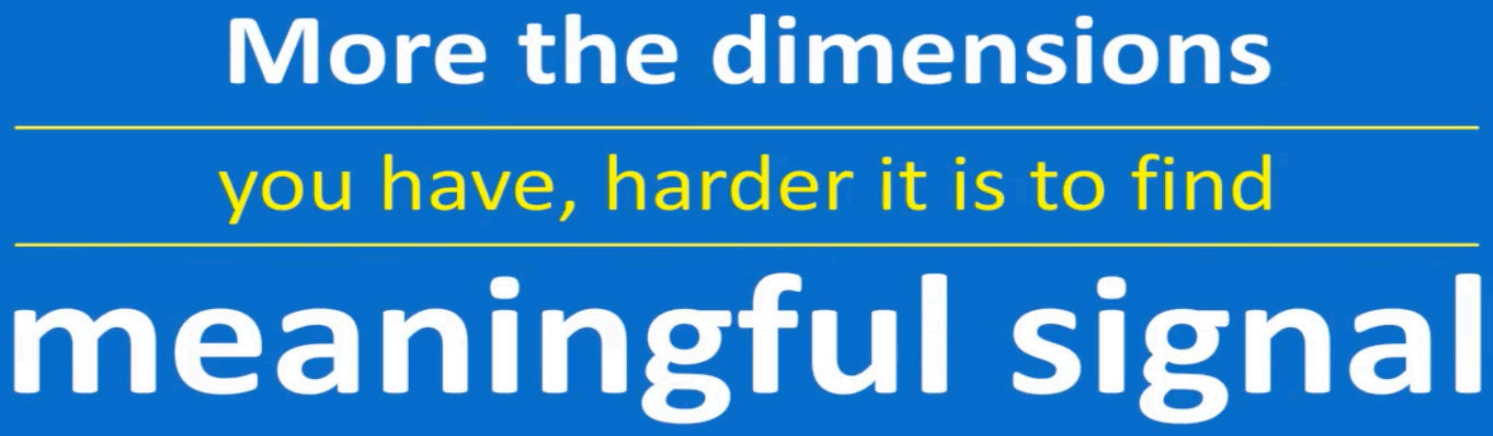
What happens if there are many variables but few samples?



- An enormous amount of training data is required to ensure that there are several samples with each combination of values (for each variable).
- With a fixed number of training samples, the predictive power reduces as the dimensionality increases.
- This is also known as the Curse of Dimensionality (COD).

Curse of Dimensionality

Watch the video lecture to view the animation.



Curse of Dimensionality

Watch the video lecture to view the animation.

So how does having more samples help?

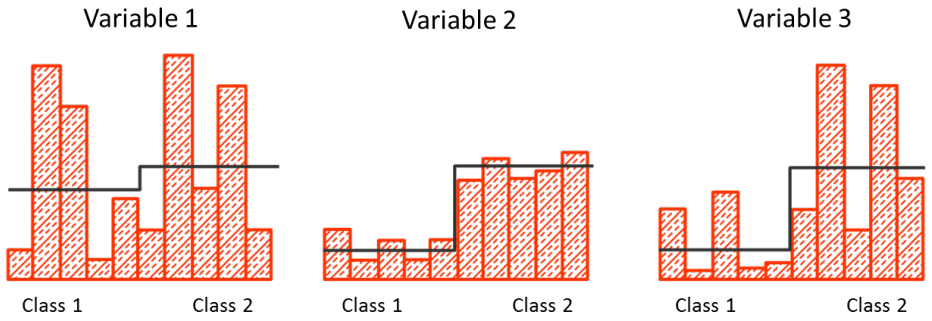
How to overcome COD?

Tackling the curse:

- Given a sample space of p dimensions (variables).
- It is possible that some dimensions are irrelevant.
- Need to find ways to separate those dimensions (aka features) that are relevant (aka signals) from those that are irrelevant (aka noise).

One idea (signal selection):

- Choose a feature w/ low intra-class distance.
- Choose a feature w/ high inter-class distance.



Which one would you choose?

How to overcome COD?

Statistics comes to the rescue:

The t-stats of a signal is defined as:

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

where σ_i^2 is the variance of that signal in class i, μ_i is the mean of that signal in class i, and n_i is the size of class i.

Something to think about:

- How is the t-statistic typically used?
- What are the assumptions required for this way of using the t-statistic?

The process of removing variables that do not pass statistical or test statistic size threshold is known as **Dimensionality Reduction** (DR).

Gene Reduction Improves Classification

- Most learning algorithms look for non-linear combinations of features -- can easily find many spurious combinations given small # of records and large # of genes.
- Classification accuracy improves if we first reduce the number of genes by a linear method, e.g. T-values of mean difference.
- Heuristic: select equal number of genes from each class.
- Then apply a favourite machine learning algorithm.

Iterative Wrapper Approach to Selecting the Best Gene Set

- Test models using 1,2,3, …, 10, 20, 30, 40, ..., 100 top genes with x-validation.
- Heuristic 1: evaluate errors from each class; select # number of genes from each class that minimises error for that class.
- For randomised algorithms, average 10+ Cross-validation runs!
- Select gene set with lowest average error.

Case Study: What if there is actually no real signal?

Statistical feature selection creates false effects.

- Construct artificial dataset with 100 samples, each with 100,000 randomly generated variables and randomly assigned class labels.
- Select 20 variables with the best t-statistic.
- Evaluate accuracy by cross validation using the 20 selected variables.
- The resulting accuracy can be ~90% .
- But the true accuracy should be 50%, as the data were derived randomly (there is actually no real signal).

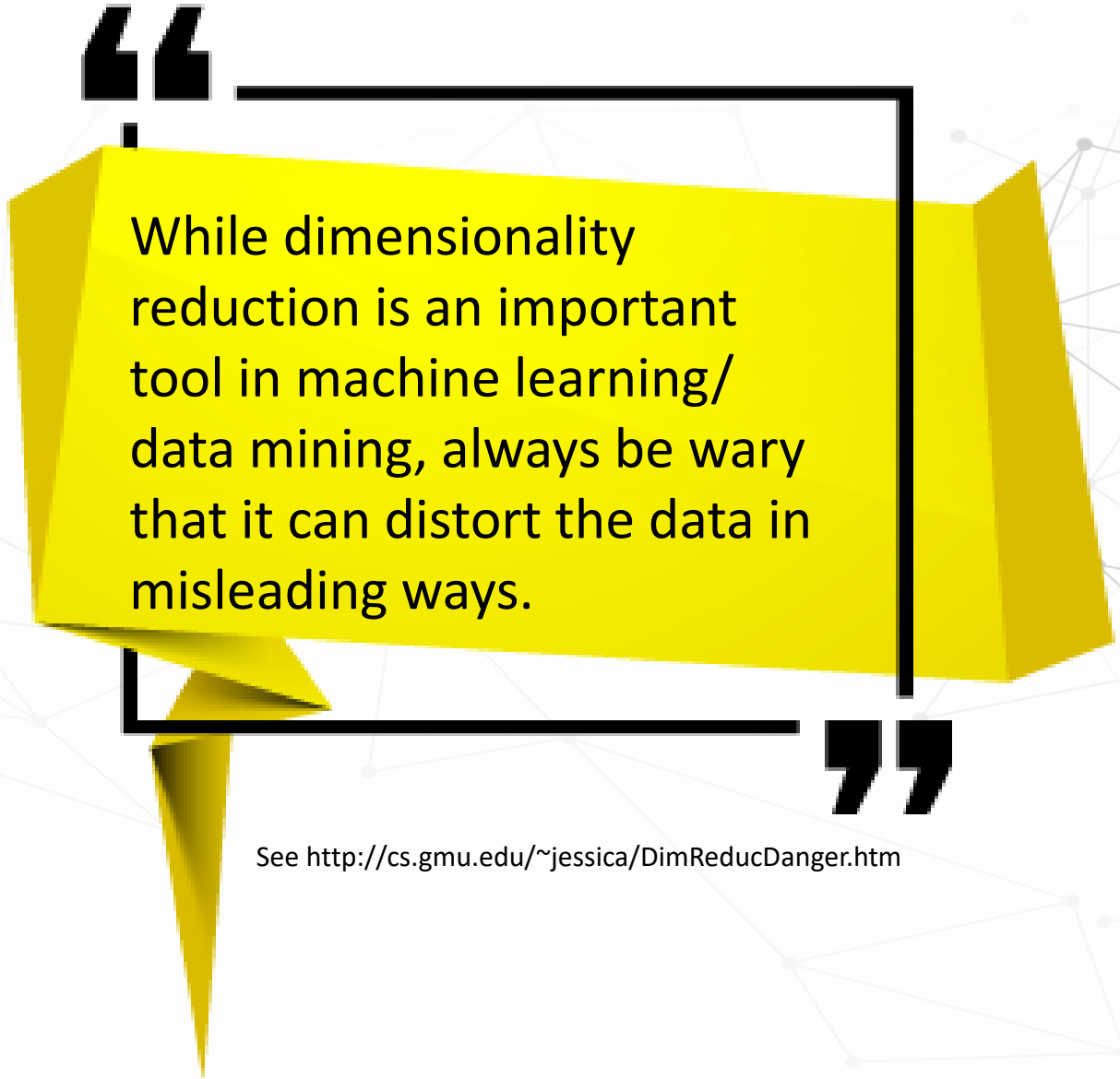
Case study: What went wrong?

The 20 variables were selected from whole dataset.

Information in the held-out testing samples has thus been “leaked” to the training process.

The correct way is to re-select the 20 variables at each fold; better still, use a totally new set of variables for testing (independent validation).

Beware Beware!



See http://cs.gmu.edu/~jessica/DimReducDanger.htm

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

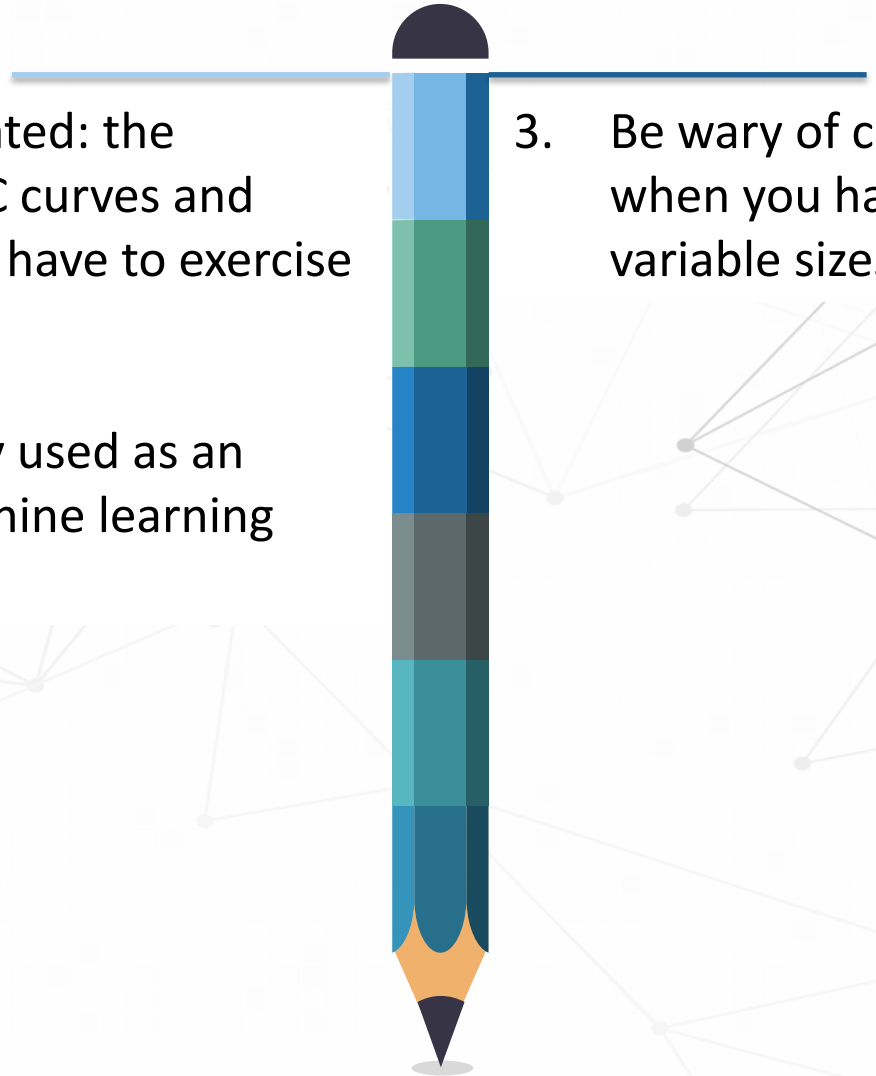
BS3033 Data Science for Biologists

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, we have to exercise discretion and discern.

2. Cross-validation is commonly used as an approach for evaluating machine learning model.



3. Be wary of curse of dimensionality issues when you have small sample sizes and large variable sizes.

References

John A. Swets, Measuring the accuracy of diagnostic systems, Science 240:1285--1293, June 1988

Trevor Hastie et al., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2001. Chapters 1, 7

Lance D. Miller et al., Optimal gene expression analysis by microarrays, Cancer Cell 2:353--361, 2002

David Hand et al., Principles of Data Mining, MIT Press, 2001

Jinyan Li et al., Data mining techniques for the practical bioinformatician, The Practical Bioinformatician, Chapter 3, pages 35—70, WSPC, 2004

Acknowledgements