NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Machine Learning – 2

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:

- Describe machine learning.
- Describe the major classes of ML methods.
- Describe how rule-based decision trees are constructed.
- Describe how KNN works.
- Describe how hierarchical clustering works.
- Describe overfitting.
- Describe the various considerations for model building.

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

What is Machine Learning?

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

What is Machine Learning (ML)?

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."

--- Arthur Samuel (1959)

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

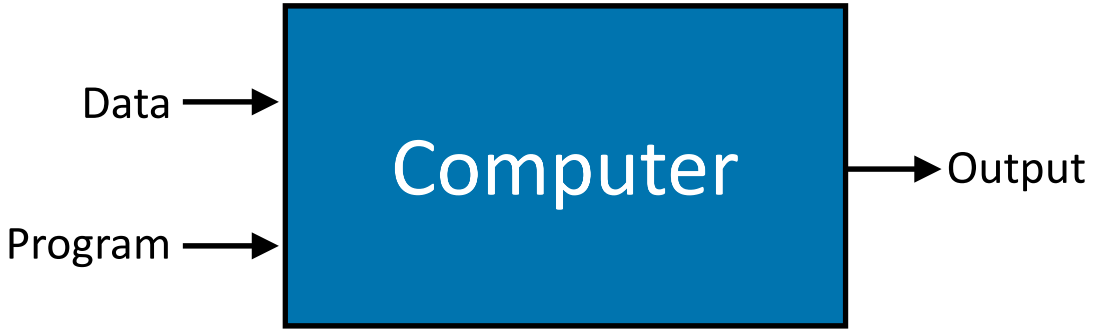
-- Tom Mitchell (1997)

ML solves complex problems that cannot be solved by numerical means alone.

What is Machine Learning (ML)?

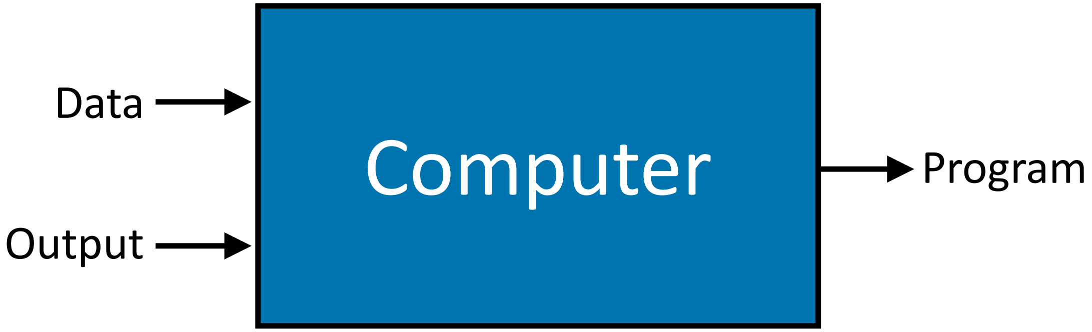
AKA Human-guided analysis

Traditional Programming



AKA Data/Machine-guided analysis

Machine Learning



The output is the sample class or the variable we want to predict.

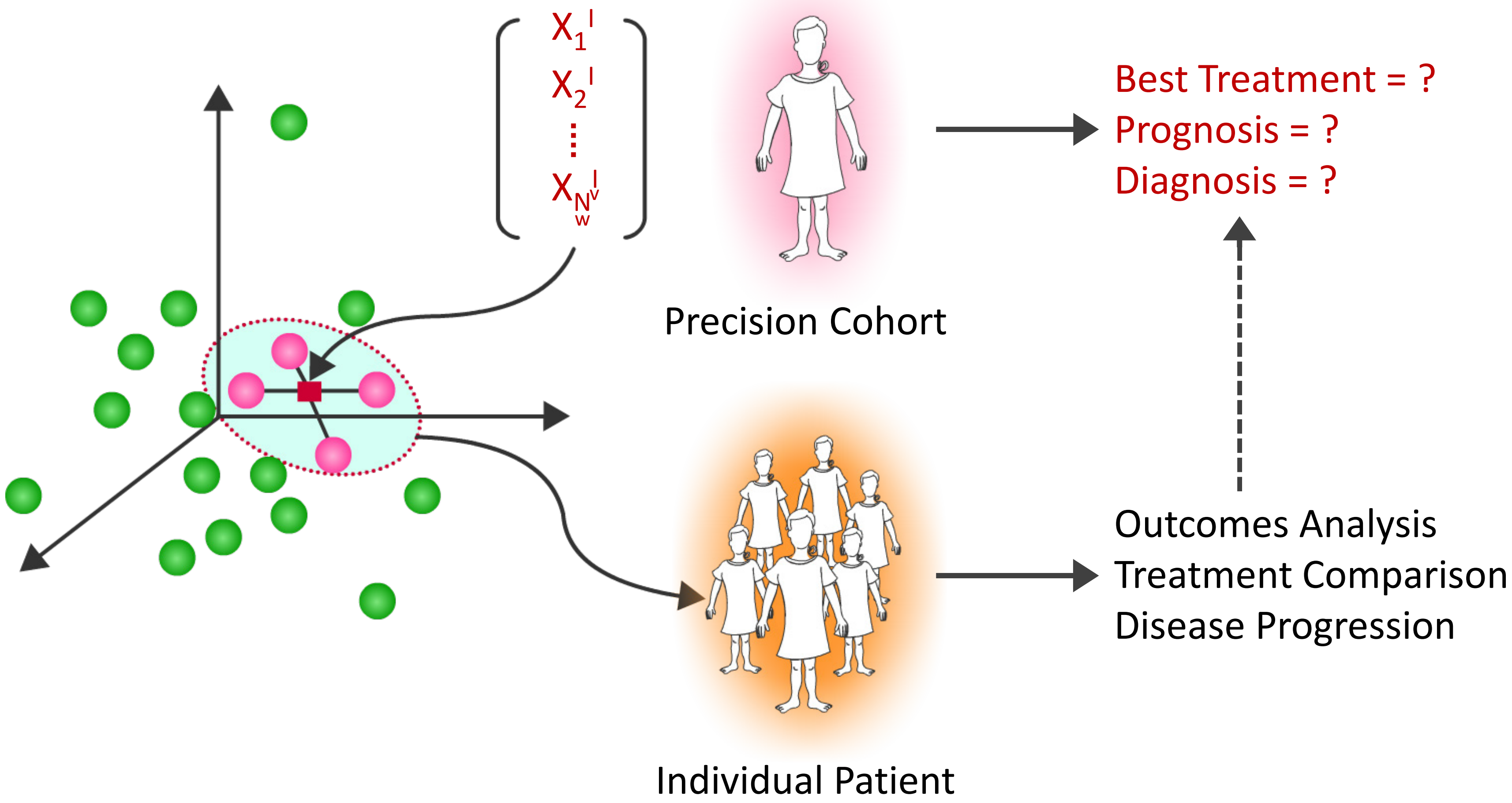
Pedro Domingos, CSE446

Suitable Problems for ML

- The highly complex nature of many real-world problems, though, often means that inventing specialised algorithms that will solve them perfectly every time is impractical, if not impossible.

- Examples of machine learning problems include, “Will this patient die from this cancer?”, “What is the market value of this house?”.

Will this patient die from this cancer?



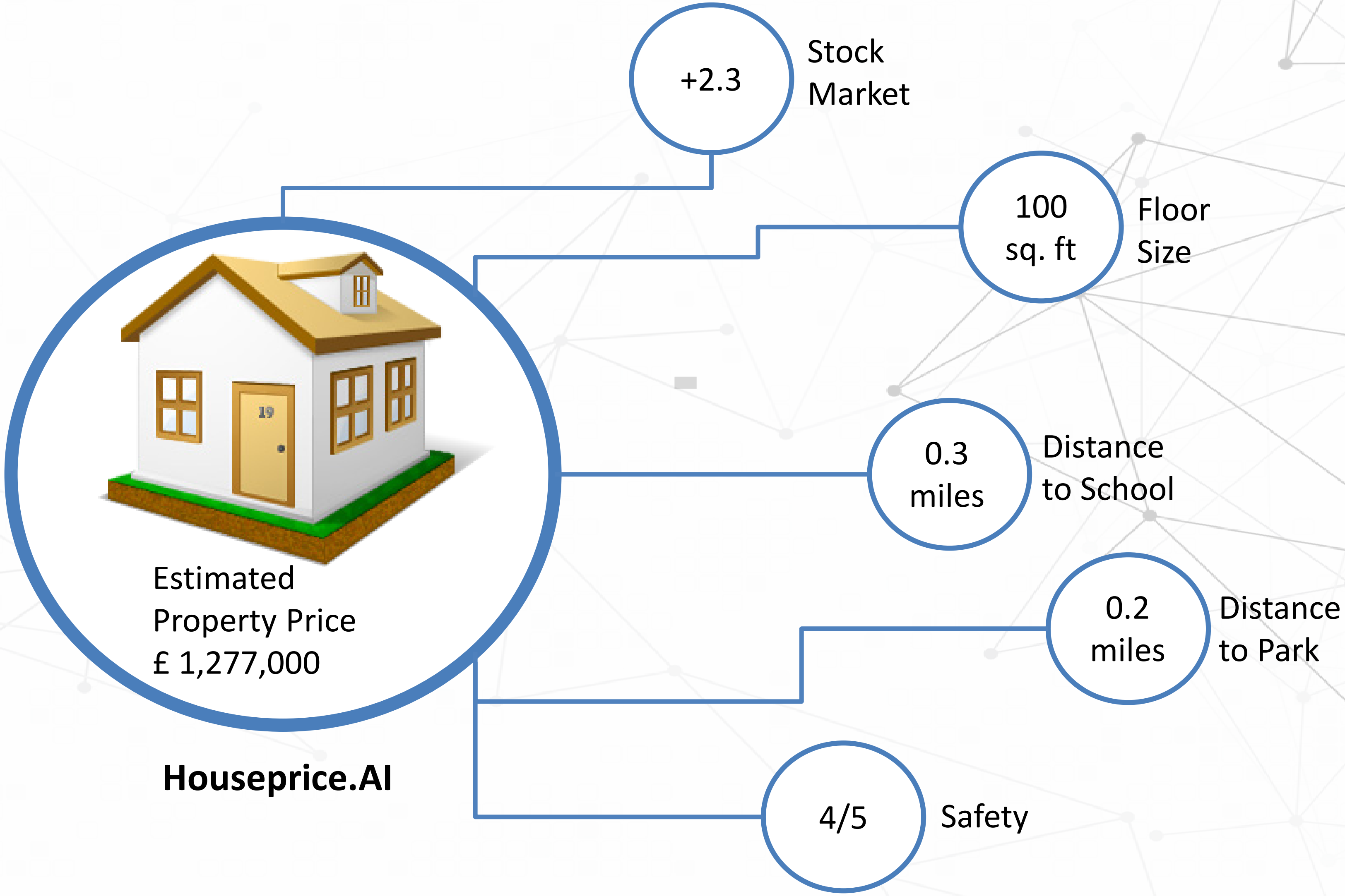
Precision Cohort

Individual Patient

Best Treatment = ?
Prognosis = ?
Diagnosis = ?

Outcomes Analysis
Treatment Comparison
Disease Progression

What is the market value of this house?



Houseprice.AI

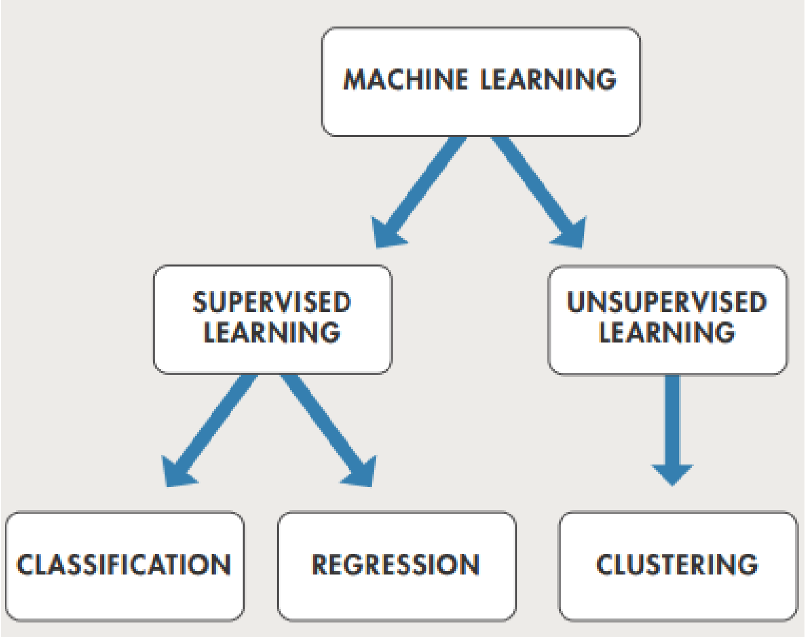
NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Overview of Machine Learning

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Supervised and Unsupervised ML



Supervised machine learning:

- **Classification machine learning systems:** guess the class (e.g. survive or die).
- **Regression:** guess the value Y when $X_1..X_n$ is observed.

Unsupervised machine learning: The program is given data and must find patterns and relationships therein **without** explicitly using class information (output).

- **Clustering**: Group together samples that are more similar to one another (then check for corroboration with output/class).

Supervised Learning (Classification)

- **Learn from past experience, and use the learned knowledge to classify new data.**
- **Knowledge learned by intelligent algorithms.**
- **Examples:**
  - Clinical diagnosis for patients
  - Cell type classification

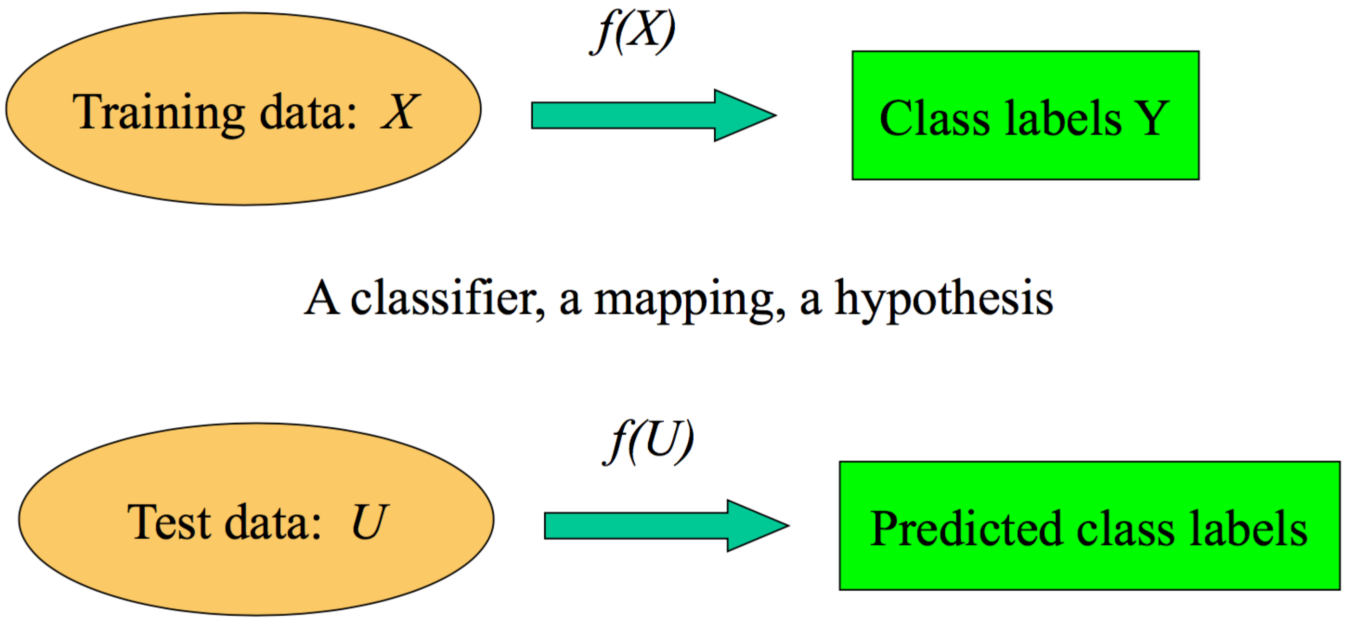
- **Classification involves > 1 class of data. E.g.,** Normal vs disease cells for a diagnosis problem.
- **Training data is a set of instances (samples, points, etc.) with known class labels.**
- **Test data is a set of instances whose class labels are to be predicted.**

Some Notation

- Training data:
  $\{<x_1, y_1>, <x_2, y_2>, \ldots, <x_m, y_m>\}$
  - where x_j are n-dimensional vectors and y_j are from a discrete space Y. E.g., Y = {normal, disease}.

- Test data:
  $\{<u_1, ?>, <u_2, ?>, \ldots, <u_k, ?> \}$
  - Where u_k is an n-dimensional vector and ? are the classes to be predicted.

Process



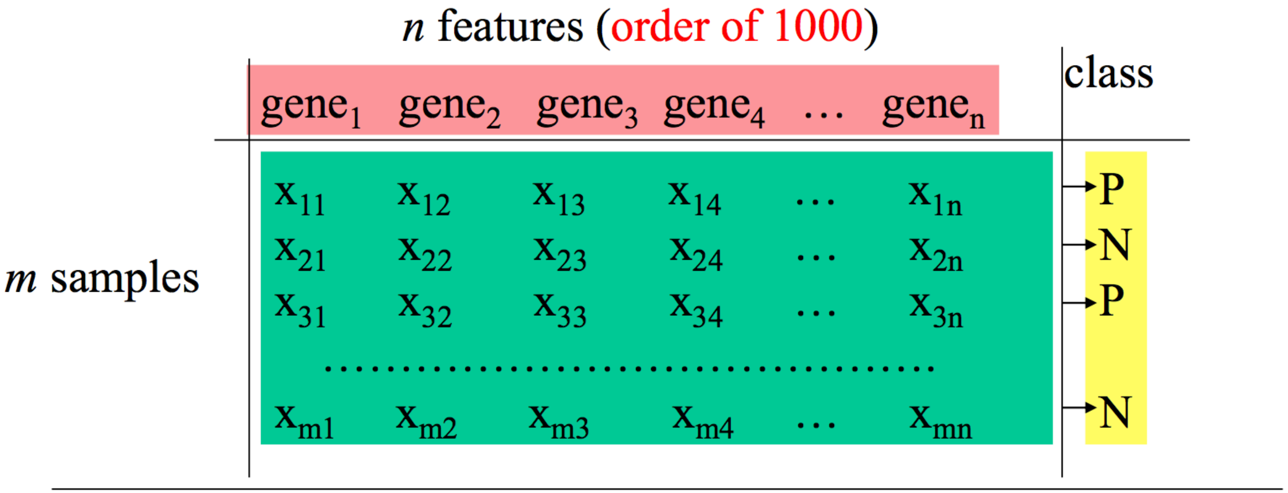
Training data: *X* $\xrightarrow{f(X)}$ Class labels Y

A classifier, a mapping, a hypothesis

Test data: *U* $\xrightarrow{f(U)}$ Predicted class labels

Relational Data Representation (X and Y)

X is $\text{gene}_1 \ldots \text{gene}_n$



n features (order of 1000)

	gene_1	gene_2	gene_3	gene_4	…	gene_n	class
	x_{11}	x_{12}	x_{13}	x_{14}	…	x_{1n}	→ P
m samples	x_{21}	x_{22}	x_{23}	x_{24}	…	x_{2n}	→ N
	x_{31}	x_{32}	x_{33}	x_{34}	…	x_{3n}	→ P
	………………………………						
	x_{m1}	x_{m2}	x_{m3}	x_{m4}	…	x_{mn}	→ N

Class = Y

Which sources of big biological data are amendable to this? Genomics, Transcriptomics, RT-PCR, Proteomics or combinations of these.

Variables/ Features

- **Categorical features (Nominal/ Ordinal)**
  - Colour = {red, blue, green}

- **Continuous or numerical features (Interval/ Ratio)**
  - Gene Expression
  - Age
  - Blood Pressure

Data Example

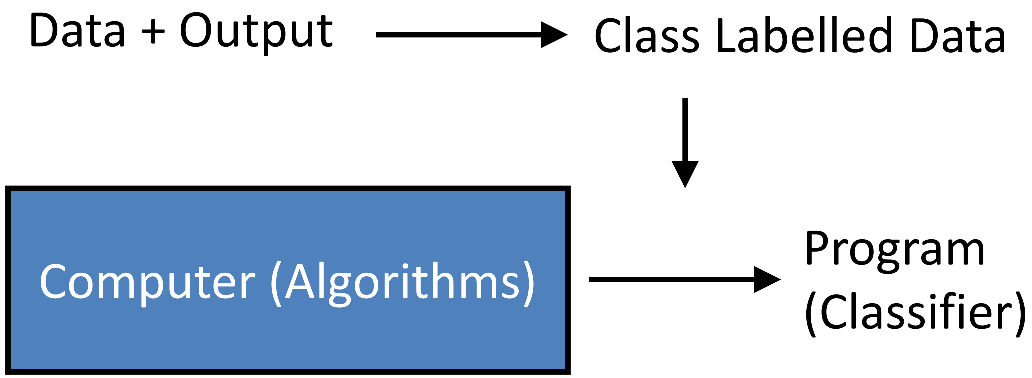
Each column is a variable

Each row is a Sample

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Categorical Continuous Categorical

Supervised Learning (Global View)



Data + Output

Class Labelled Data

Computer (Algorithms)

Program (Classifier)

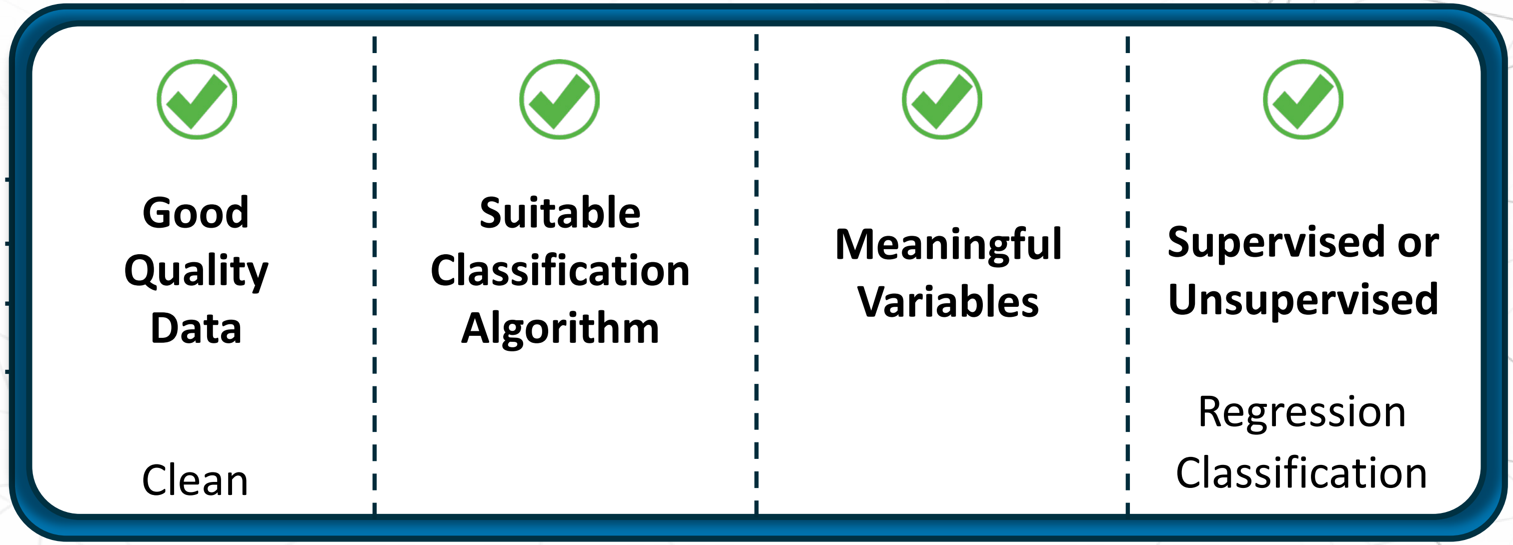
How do you know if your predictions are good?

- **Many measures:**
  - Accuracy, error rate, false positive rate, false negative rate, sensitivity, specificity, precision.
- **K-fold cross validation:**
  - Given a dataset, divide it into k even parts, k-1 of them are used for training, and the rest one part treated as test data.
- **Independent validation (Performance on independent blind test data):**
  - Blind test data properly represent real world.

Requirements of a Good Classifier

- High accuracy, sensitivity, specificity and precision (Is this truly possible?).
- High comprehensibility.

What determines good performance?



✓ **Good Quality Data**

Clean

✓ **Suitable Classification Algorithm**

✓ **Meaningful Variables**

✓ **Supervised or Unsupervised**

Regression

Classification

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Decision Trees

BS3033 Data Science for Biologists

Dr Wilson Goh

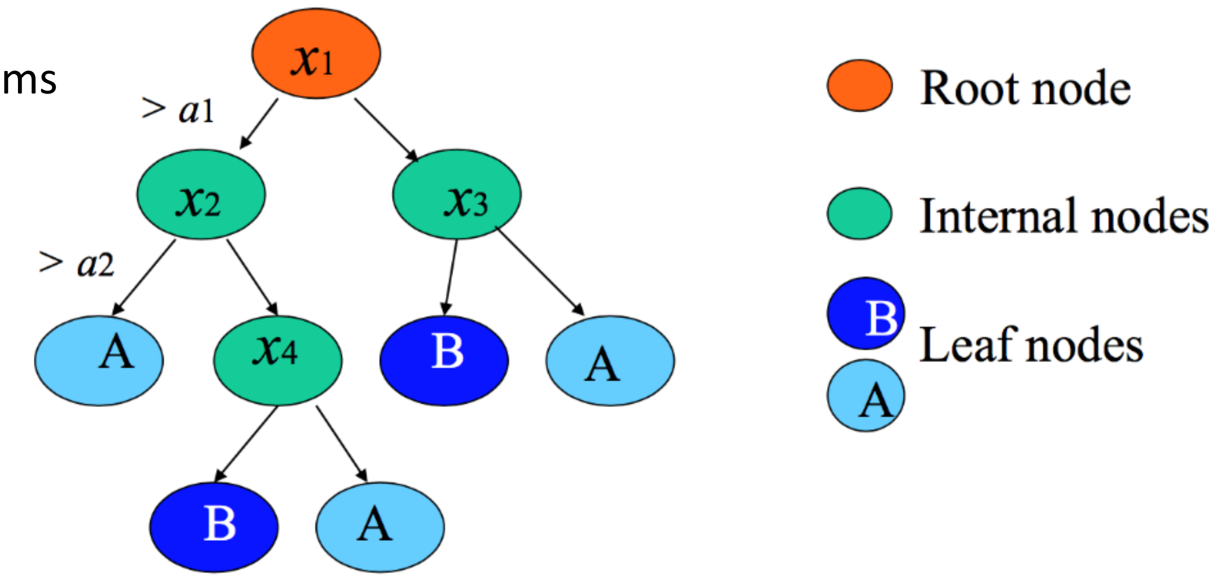
School of Biological Sciences

Decision Trees

- A group of rule-based methods useful for classification.
- Systematic selection/ ordering of a small number of features used for the decision making.
- This increases comprehensibility of the knowledge patterns (tells us which variables are the most important).

Structure of Decision Trees

Every path from root to a leaf forms a **decision rule**.



- If $x_1 > a_1$ & $x_2 > a_2$, then it's class A.
- Easy interpretation, but accuracy may be unattractive.

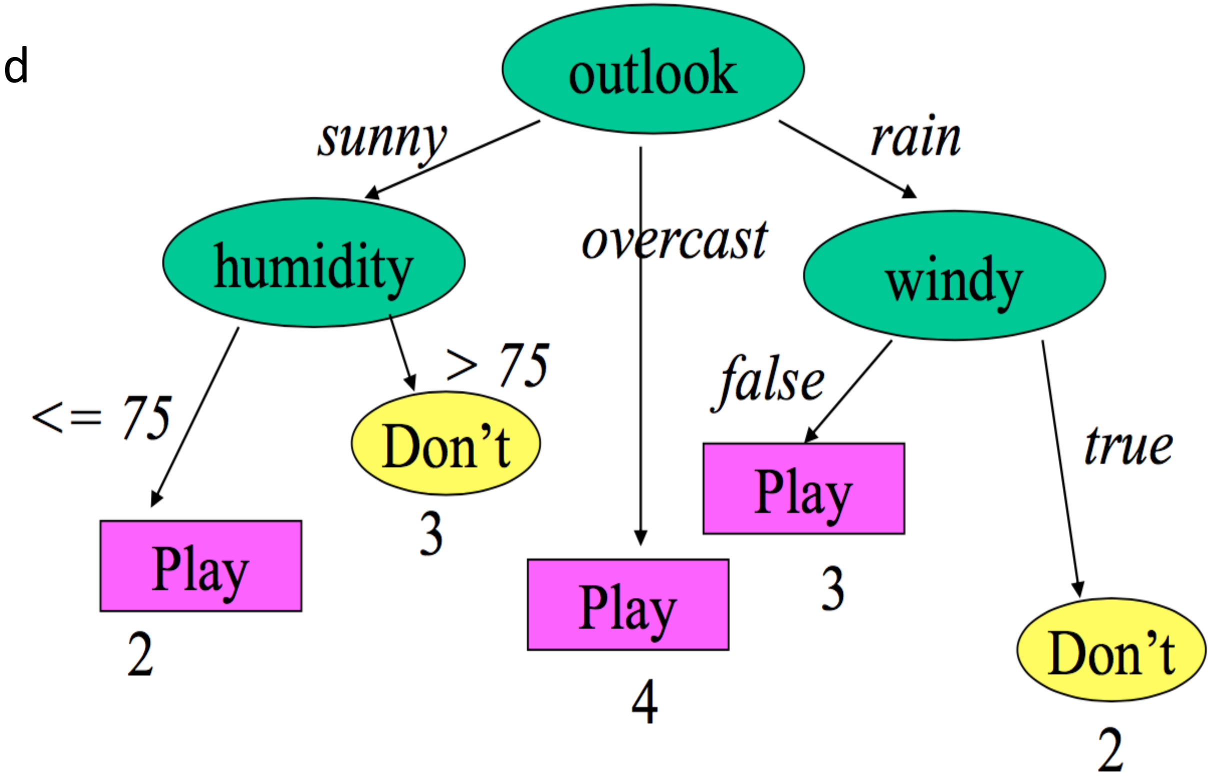
Decision Tree Example

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

A total of 14 outcomes:
9 Play
5 Don’t Play

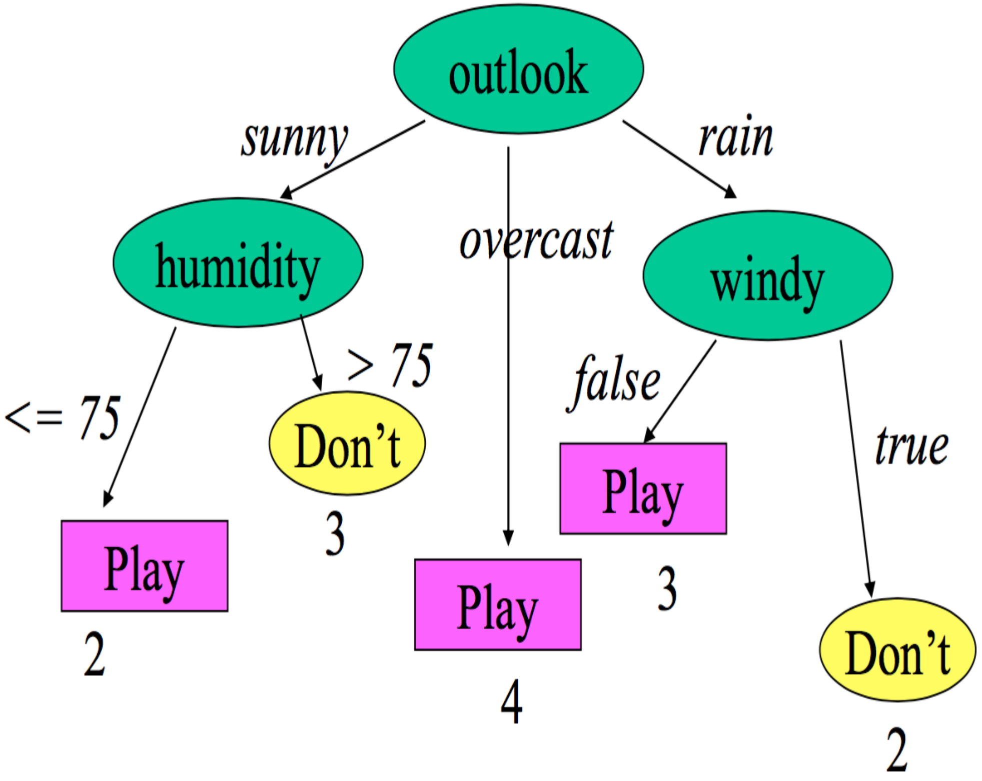
Decision Tree Example

Construction of a tree is equivalent to determination of root node of the tree and root nodes of its sub-trees.



Decision Tree Example

Outlook	Temp	Humidity	Windy	Class	Predicted	Verdict
Sunny	75	70	True	Play	Play	TP
Sunny	80	90	True	Don’t	Don’t	TN
Sunny	85	85	False	Don’t	Don’t	TN
Sunny	72	95	True	Don’t	Don’t	TN
Sunny	69	70	False	Play	Play	.
Overcast	72	90	True	Play	Play	.
Overcast	83	78	False	Play	Play	.
Overcast	64	65	True	Play	Play	.
Overcast	81	75	False	Play	Play	.
Rain	71	80	True	Don’t	Don’t	.
Rain	65	70	True	Don’t	Don’t	.
Rain	75	80	False	Play	Play	.
Rain	68	80	False	Play	Play	.
Rain	70	96	False	Play	Play	.



Most Discriminatory Variable

- Every variable can be used to partition the training data e.g., “Play and Don’t Play”.

- If the partitions contain at least 1 pure class of training instances, then this variable is most certainly discriminatory.

Partitions

- Categorical feature:
  - Number of partitions of the training data is equal to the number of values of this feature e.g. Number of partitions {Play, Don’t Play} = 2.
- Numerical feature:
  - Two partitions based on some threshold e.g. A > 100 (splits into values which are greater than 100 or otherwise).

Data Example

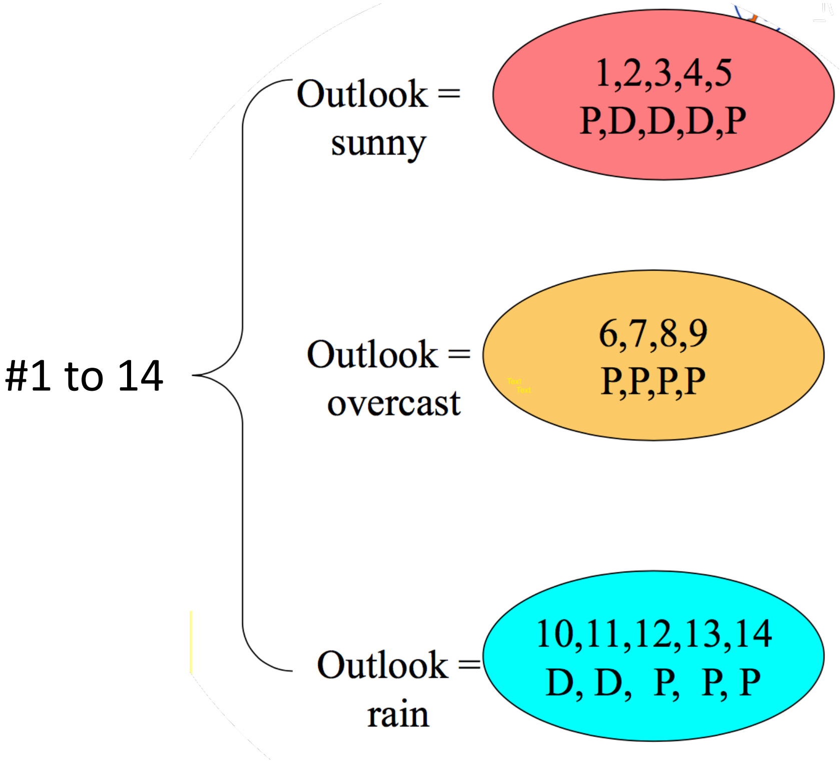
Each column is a variable

Each row is a Sample

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Categorical Continuous Categorical

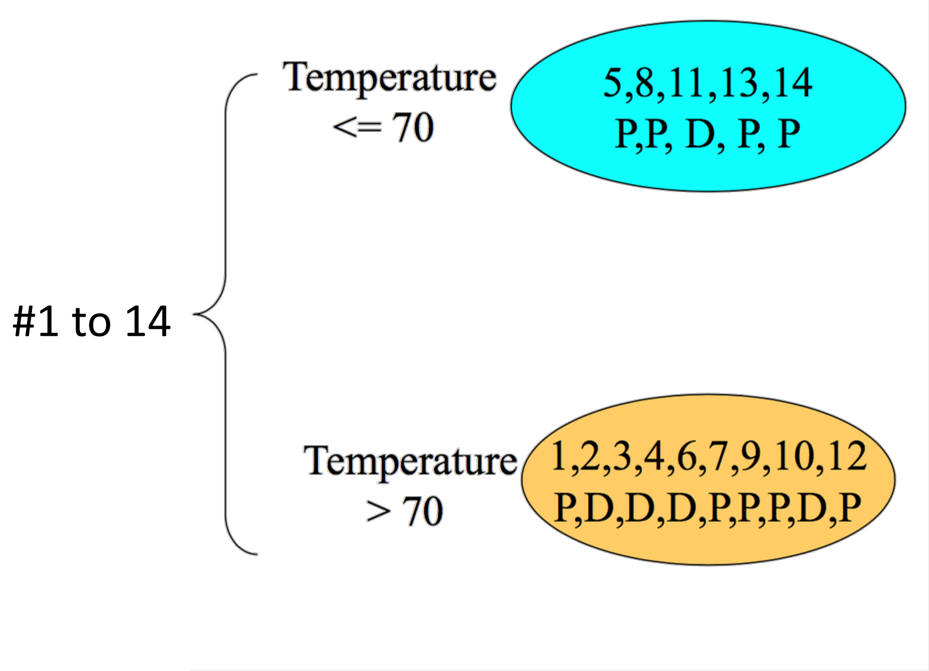
Partitioning Variables



A categorical feature is partitioned based on its number of possible values.

Partitioning Variables

A numerical feature is generally partitioned by choosing a “cutting point”.



#1 to 14

Temperature <= 70

5,8,11,13,14
P,P, D, P, P

Temperature > 70

1,2,3,4,6,7,9,10,12
P,D,D,D,P,P,P,D,P

Decision Tree Construction

1. Select the “best” feature as root node of the whole tree.
2. Partition dataset into subsets using this feature so that the subsets are as “pure” as possible.
3. After partition by this feature, select the best feature (with respect to the subset of training data) as root node of this sub-tree.
4. Recursively, until the partitions become pure or almost pure.

Let’s Construct a Decision Tree

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Gini Coefficient

- Gini Index or coefficient can be used as an approximation of the power of a variable.
  - Split is completely pure, Gini index = 0
  - Split is impure, max Gini index = 1 – 1/k (where k = number of class levels)

$$Gini = \sum_{i \neq j} p(i)p(j)$$

i and j are levels of the target variable

- The sum of the joint probabilities of all impure combinations.
- Minimum value of Gini Index will be 0 when all observations belong to one class label.

Gini Coefficient

Suppose we have class label with 2 levels -> Normal (N) and Cancer (C). There are 4 possible permutations.

1	2	3	4
Normal	Cancer	Cancer	Normal
Normal	Cancer	Normal	Cancer

P(Class=N).P(Class=N) + P(Class=C).P(Class=C) + P(Class=C).P(Class=N) + P(Class=N).P(Class=C) = 1

P(Class=N).P(Class=C) + P(Class=C).P(Class=N) = 1 - P(Class=N).P(Class=N) - P(Class=C).P(Class=C)

P(Class=N).P(Class=C) + P(Class=C).P(Class=N) = 1 – P^2(Class=N) - P^2(Class=C)

Maximum value of Gini Index = 1 – (P^2(Class=N) + P^2(Class=C))

Maximum value of Gini Index = $1-\sum_{t=0}^{t=k} P_t^2$

Where t is the class, and k are attributes of class (N and C).

Gini Coefficient

1	2	3	4
Normal	Cancer	Cancer	Normal
Normal	Cancer	Normal	Cancer

- Max Gini Index value = $1 - (1/2)^2 - (1/2)^2 = 1 - 2*(1/2)^2 = 1 - 2*(1/4) = 1-0.5 = 0.5$
- Similarly for Nominal variable with k level, the maximum value Gini Index is= 1 - 1/k.
- Since the play data has 2 levels (play and don't play), its max Gini index is also 0.5.
- However, knowing the min and max Gini coefficients don't tell us what is the quality of a split given a variable.

Gini Coefficient of a Split

GINI (s,t) = GINI (t) – P_L GINI (t_L) – P_R GINI (t_R)

where

s: split

t: node

GINI (t): Gini Index of input node t

P_L: Proportion of observation in Left Node after split, s

GINI (t_L): Gini of Left Node after split, s

P_R: Proportion of observation in Right Node after split, s

GINI (t_R): Gini of Right Node after split, s

Gini Coefficient of Outlook

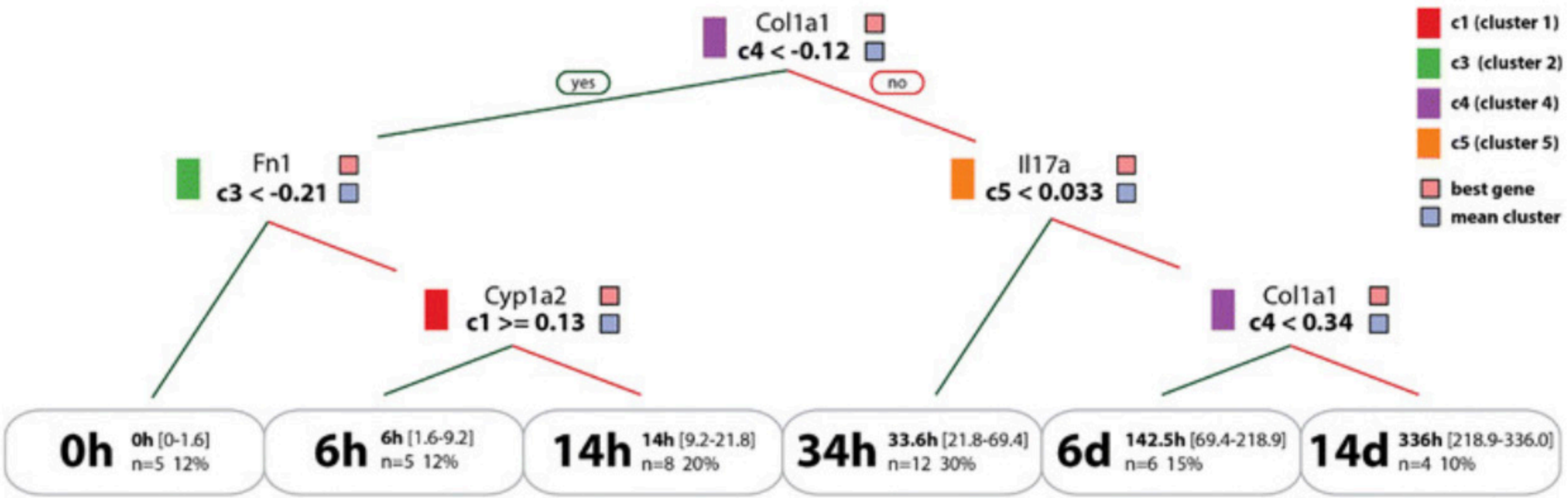
- GINI (t) ≠ 0.5
- GINI (t) = $1 - (5/14)^2 - (9/14)^2 = 0.46$ [the distribution between classes is not equal!]
- Gini (Sunny) = $1 - (2/5)^2 - (3/5)^2 = 0.48$
- Gini (Overcast) = $1 - (4/4)^2 - (0/4)^2 = 0$
- Gini (Rain) = $1 - (3/5)^2 - (2/5)^2 = 0.48$
- Gini (Outlook) = $0.46 - (5/14 * 0.48 + 4/14 * 0 + 5/14 * 0.48) = 0.46 - 0.34 = 0.12$

#note that Gini (overcast) is a pure sub-cluster

#Try doing Gini (Windy) and Gini (Humidity <= 75) yourself

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Decision Tree in Action



- When considering high-throughput data with thousands of variables, the split rules are often not so clear. In this case.
- A “representative best gene” is shown at the top but these are by no means exhaustive (there can be many equivalent best genes at each level) nor does the selection of best genes necessarily mean anything biologically beyond prediction value.

Decision Trees

Advantages

- Single coverage of training data (elegance).
- Divide-and-conquer splitting strategy (simple).
- Rules are obvious (understandable).

Disadvantages

- Fragmentation problem => Locally reliable but globally insignificant rules.
- Miss many globally significant rules.
- Mislead system.

Some Examples of Use of Decision Trees in Biological data

- In prostate and bladder cancers (Adam et al. Proteomics, 2001).
- In serum samples to detect breast cancer (Zhang et al. Clinical Chemistry, 2002).
- In serum samples to detect ovarian cancer (Petricoin et al. Lancet; Li & Rao, PAKDD 2004).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

K-Nearest Neighbours

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

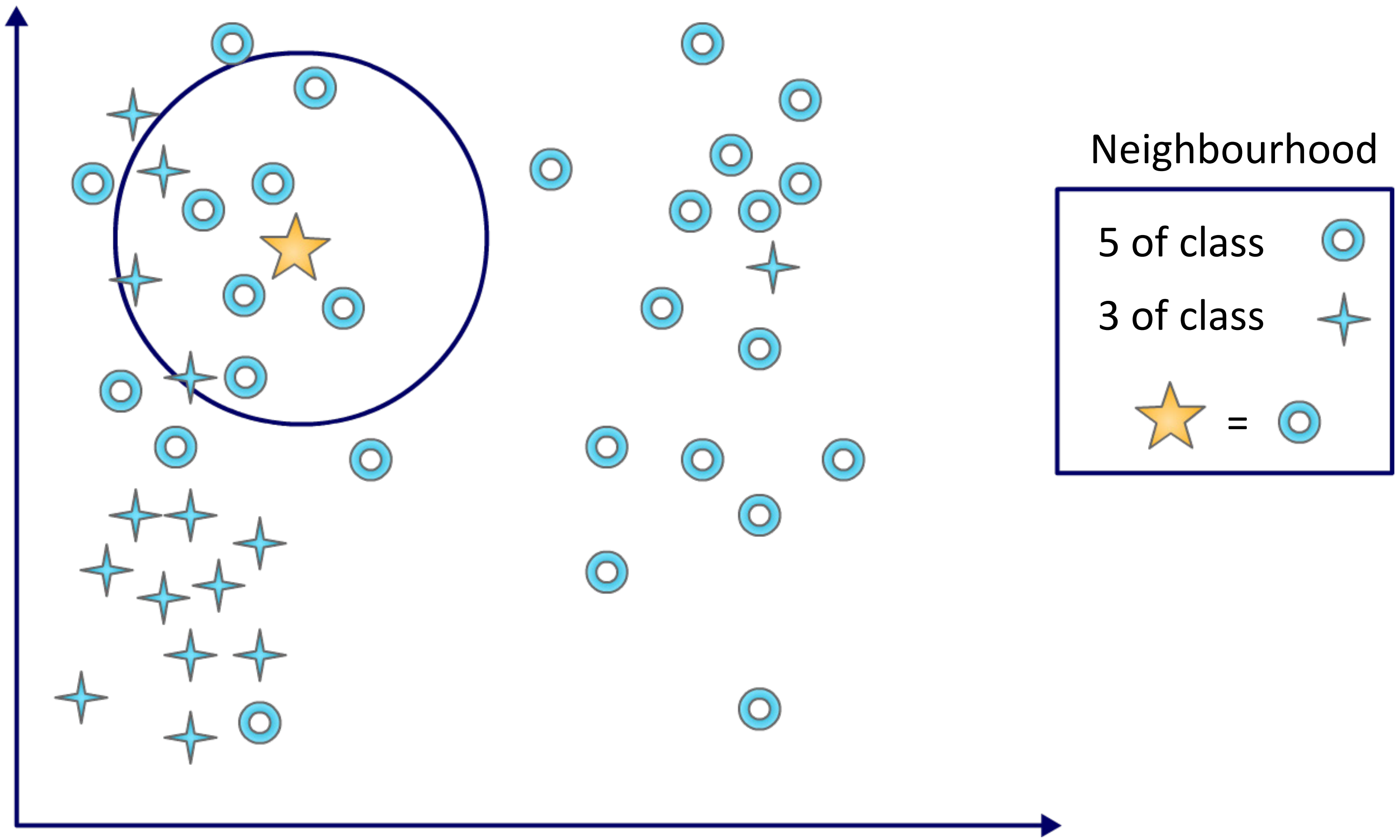
K-Nearest Neighbours (kNN)

Given a new case:

- Find k “nearest” neighbours, i.e., k most similar points in the training data set.
- Assign new case to the same class to which most of these neighbours belong .
- A common “distance” measure between samples x and y is $\sqrt{\sum_f (x[f] - y[f])^2}$.
- Where f ranges over variables of the samples.

Illustration of kNN (k=8)

What should the class of ★ be?



Neighbourhood

5 of class ◎

3 of class ✦

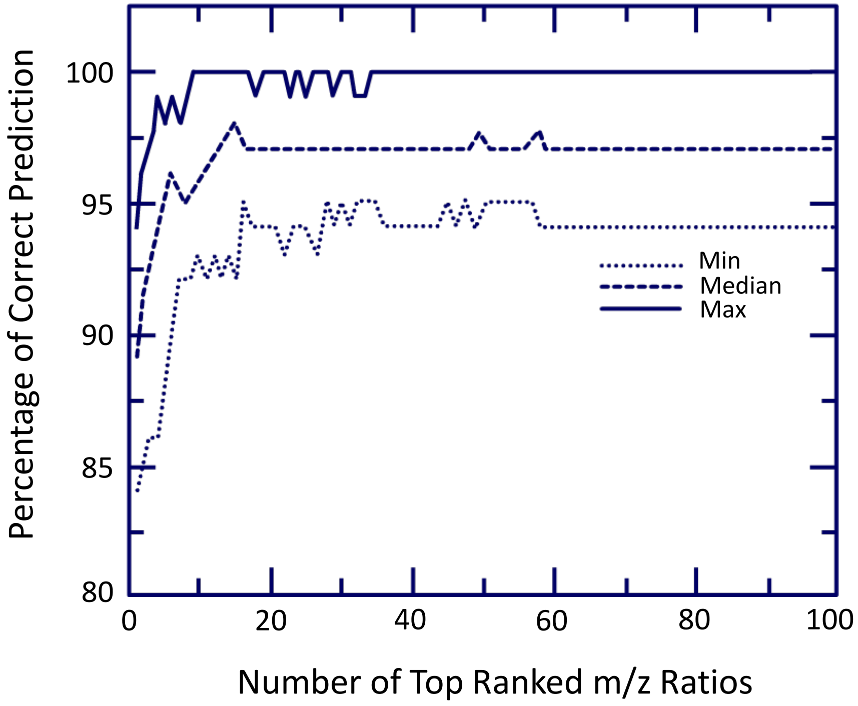
★ = ◎

Some Issues

- Simple to implement.
- Must compare new case against all training cases.
  - May be slow during prediction.
- No need to train.
- But need to design distance measure properly.
  - May need expert for this.
- Can’t explain prediction outcome.
  - Can’t provide a model of the data.

Example Use of kNN

- Li et al, *Bioinformatics* 20:1638-1640, 2004.
  - ○ Use kNN to diagnose ovarian cancers using proteomic spectra.
  - ○ Data set is from Petricoin et al., *Lancet* 359:572-577, 2002.



Minimum, median and maximum of percentages of correct prediction as a function of the number of top-ranked *m/z* ratios on 50 independent partitions into learning and validation sets.

Random Forest and Other Supervised Methods

BS3033 Data Science for Biologists

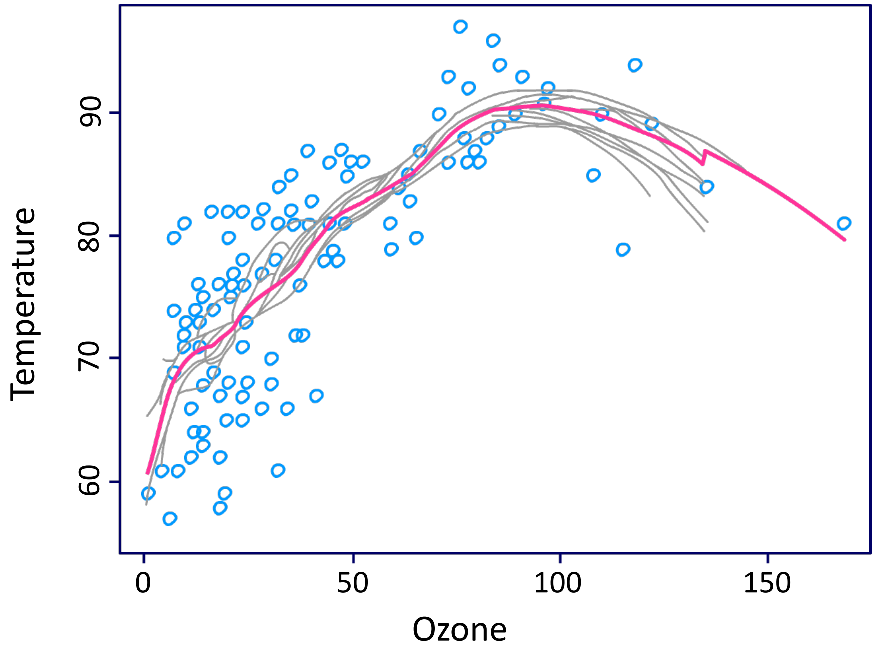
Dr Wilson Goh
School of Biological Sciences

Ensemble ML Algorithms

- The random forest belongs to a class of ML methods called “Ensemble”.
- Brute-force: Instead of using just one classifier, use hundreds/ thousands of classifiers at once.

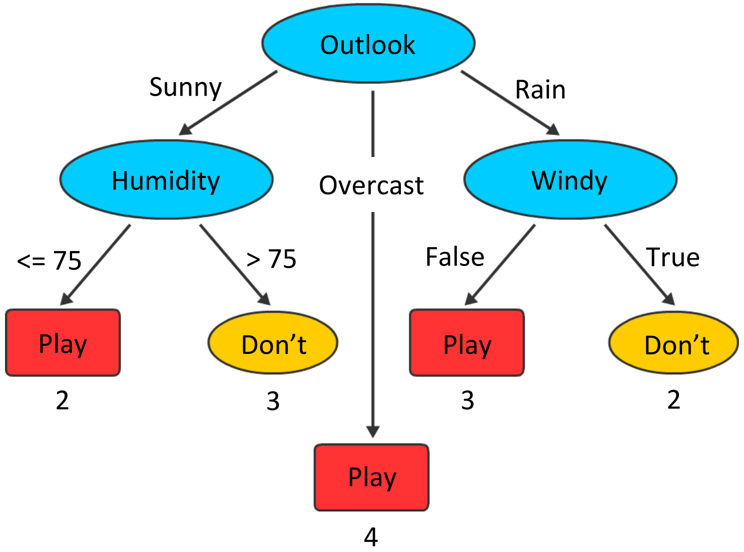
The “wisdom” of Crowds

- The main principle behind Ensemble methods is that a group of “weak learners” can come together to form a “strong learner”.
- Each classifier (grey), individually, is a “weak learner,” while all the classifiers taken together (red line) are a “strong learner”.
- Data points are in blue.

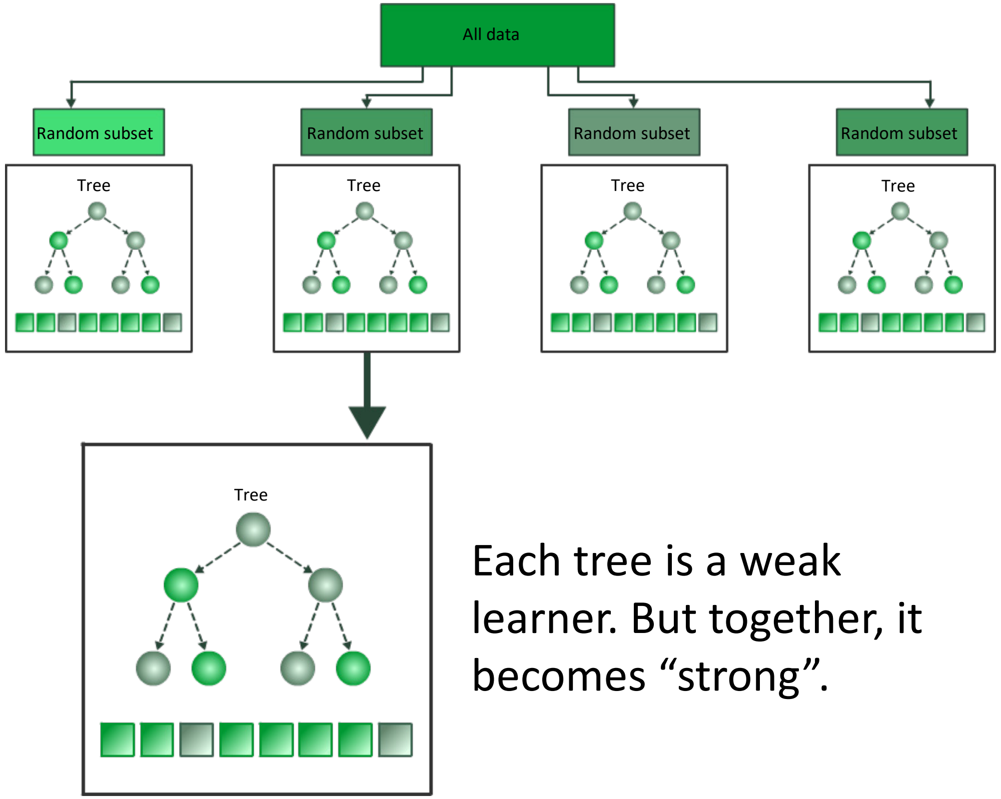


How does the RF work?

A Single Tree



A "Forest"



Each tree is a weak learner. But together, it becomes "strong".

How does RF work?

How does RF work?

Define a value of m, define the number of trees n.

1

For each tree, take a random subset of samples.
At each node in tree:

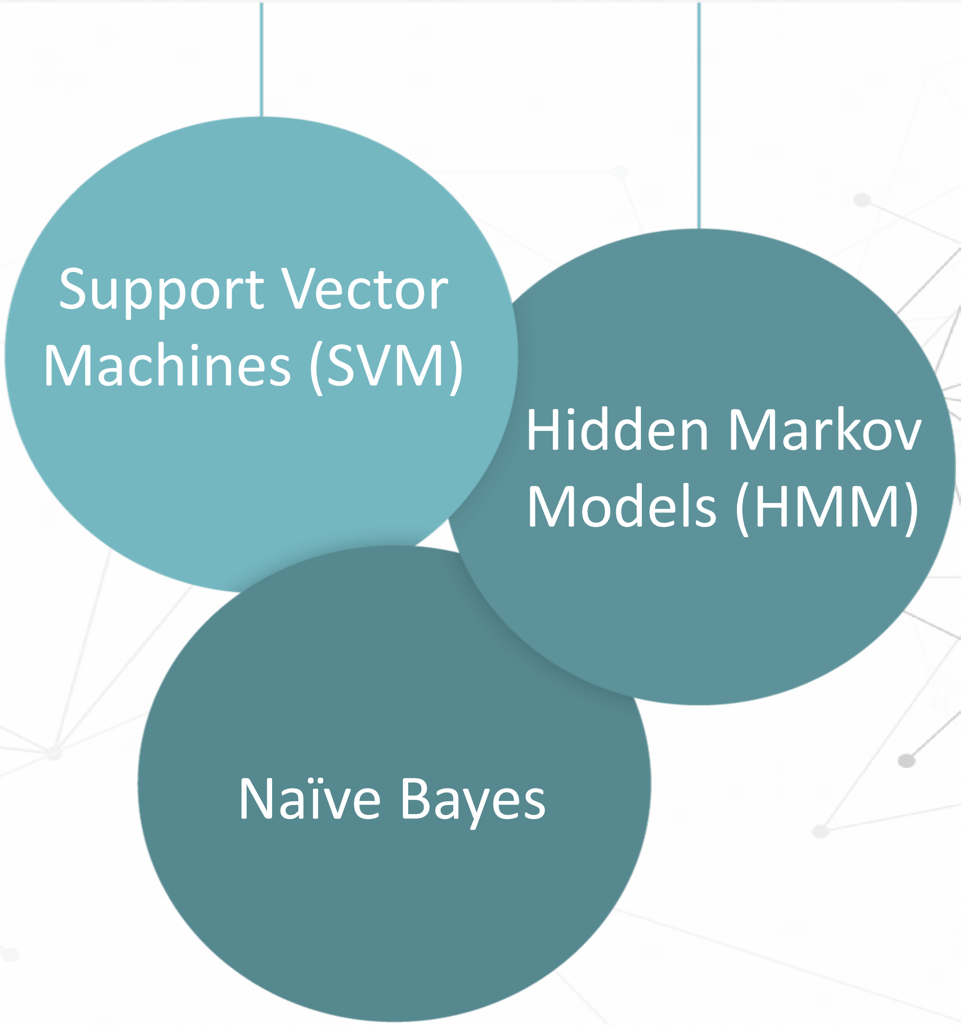
- *Select m* predictor variables randomly.
- The predictor variable that provides the best split, is used to do a binary split on that node.
- At the next node, choose another *m* variables at random from all predictor variables and do the same.

2

Evaluate aggregate performance over n trees (majority voting).

3

Other Examples of Supervised ML Approaches



Support Vector Machines (SVM)

Hidden Markov Models (HMM)

Naïve Bayes

Not covered in lectures. Just for general knowledge. You will encounter some of these in the tutorial.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Unsupervised Learning

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Unsupervised Machine Learning

Unsupervised learning typically is tasked with finding relationships within data.

No training examples used. The system is given a set data and tasked with finding patterns and correlations therein. A good example is identifying close-knit groups of friends in social network data.

The algorithms used to do this are very different from those used for supervised learning e.g. clustering algorithms such as k-means and hierarchical clustering and dimensionality reduction systems such as principal component analysis.

Things that we can do with Unsupervised Learning

- We can distinguish samples from one another.
- We can group/ cluster the samples, and understand their characteristics (Do they form a true class?).
- Discover interesting structures/ substructures within the data.
- Extract insights for determining next course of action.

Hierarchical Clustering, HCL (Unsupervised)

Hierarchical clustering techniques are an important category of clustering methods. There are two basic approaches for generating a hierarchical clustering:

Agglomerative: Start with the points as individual clusters and, at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity (use similarity).

Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide which cluster to split at each step and how to do the splitting (use distance).

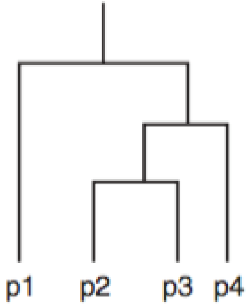
Agglomerative methods are more common.

Hierarchical Clustering, HCL (Unsupervised)

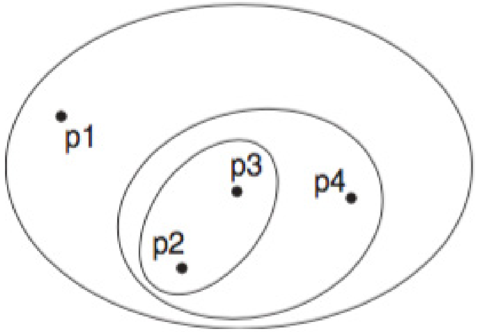
- A hierarchical clustering is often displayed graphically using a tree-like diagram called a dendrogram, which displays both the cluster-sub-cluster relationships and the order in which the clusters were merged (agglomerative view) or split (divisive view).

- For sets of two-dimensional points, a hierarchical clustering can also be graphically represented using a nested cluster diagram.

- Dendrograms are by far the more common graphical representation format.



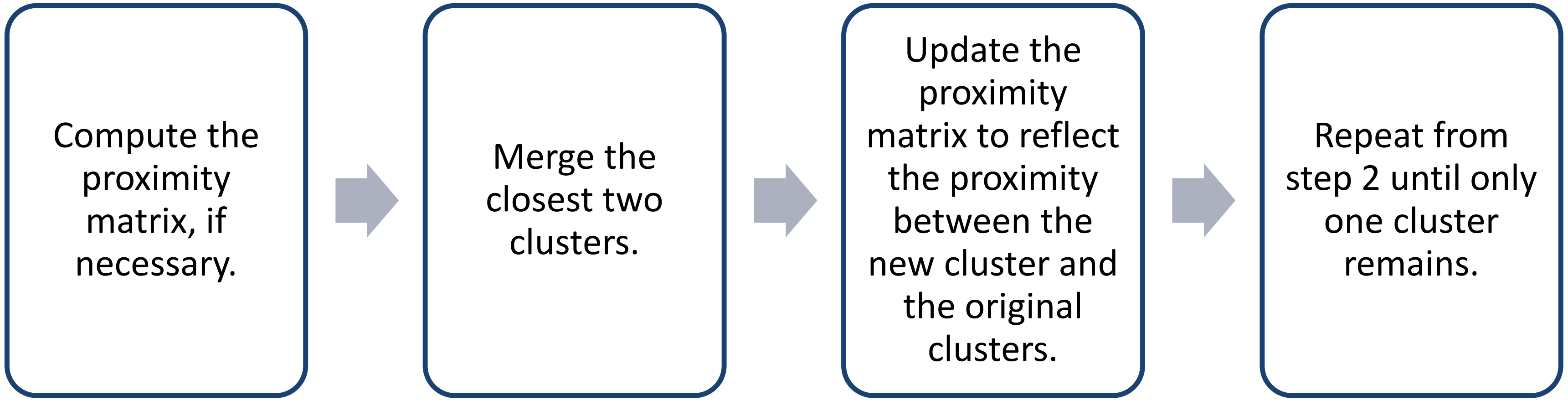
Dendrogram



Nested Cluster Diagram

Hierarchical Clustering, HCL (Unsupervised)

Starting with individual points as clusters, successively merge the two closest clusters until only one cluster remains (Deterministic algorithm).



Compute the proximity matrix, if necessary.

Merge the closest two clusters.

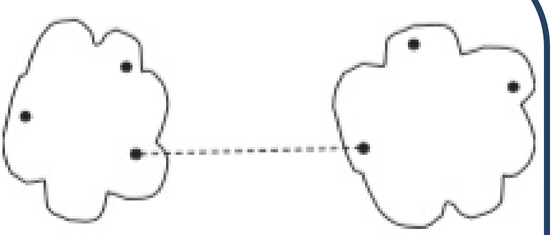
Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.

Repeat from step 2 until only one cluster remains.

Basic Agglomerative Hierarchical Clustering Algorithm

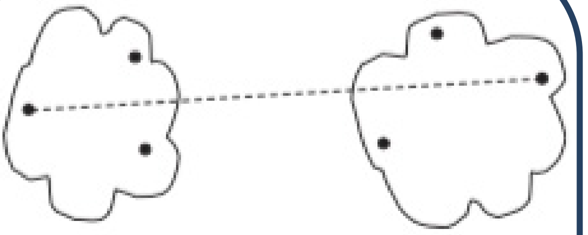
Hierarchical Clustering, HCL (Unsupervised)

Three main computations of the proximity between two clusters (linkage):



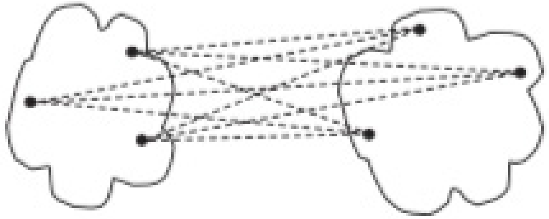
MIN (single link)

MIN defines cluster proximity as the proximity between the closest two points that are in different clusters.



MAX (complete link)

MAX takes the proximity between the farthest two points in different clusters to be the cluster proximity.



Group Average

Group average technique, defines cluster proximity to be the average pairwise proximities (average length of edges) of all pairs of points from different clusters.

Hierarchical Clustering, HCL (Unsupervised)

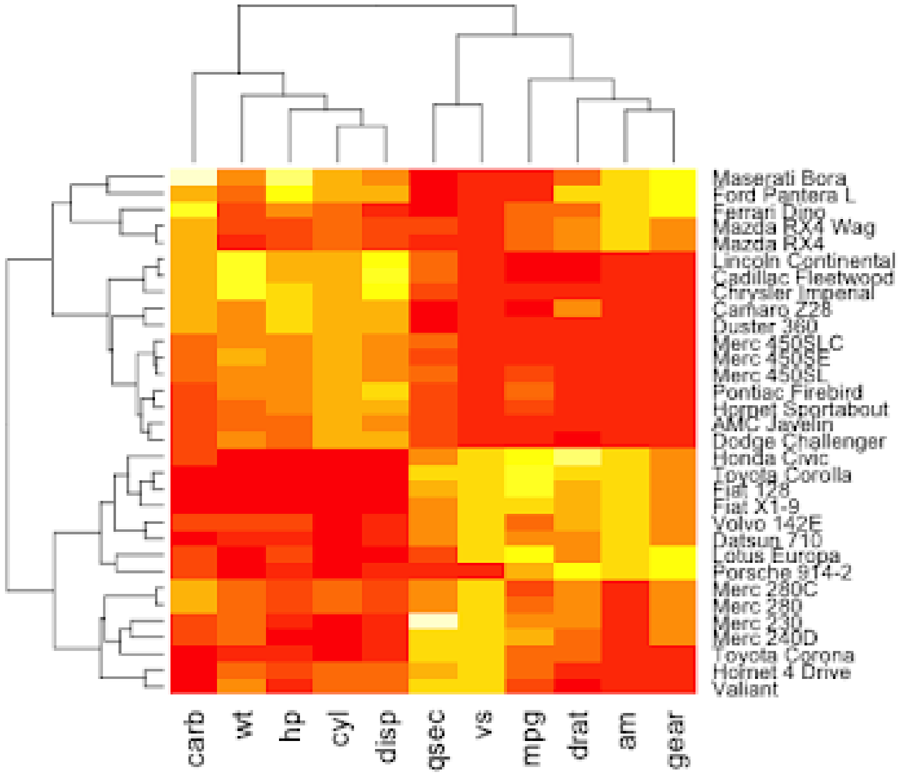
Agglomerative hierarchical clustering algorithms tend to make good local decisions about combining two clusters since they can use information about the pairwise similarity of all points. However, once a decision is made to merge two clusters, it cannot be undone at a later time.

This approach prevents a local optimisation criterion from becoming a global optimisation criterion. (What was optimal given a smaller problem, might not be optimal for the greater problem).

Tend to produce good quality clusters. Deterministic (Does not change if you run it several times).

Hierarchical Clustering, HCL (Unsupervised)

Hierarchical clustering tends to be represented together with the heatmap. The heatmap only tells you the intensity of the values that were considered in the construction of the dendrogram. (The heatmap is not the HCL).



Performance Evaluation (Cluster Validation)

The quality of the clustering needs to be determined. We may use the following techniques:

Check cluster memberships with known class labels (clustering accuracy).

Evaluate performance on dummy data where no structure exists (false positives).

Determine reproducible results on other similar data (reproducibility test).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

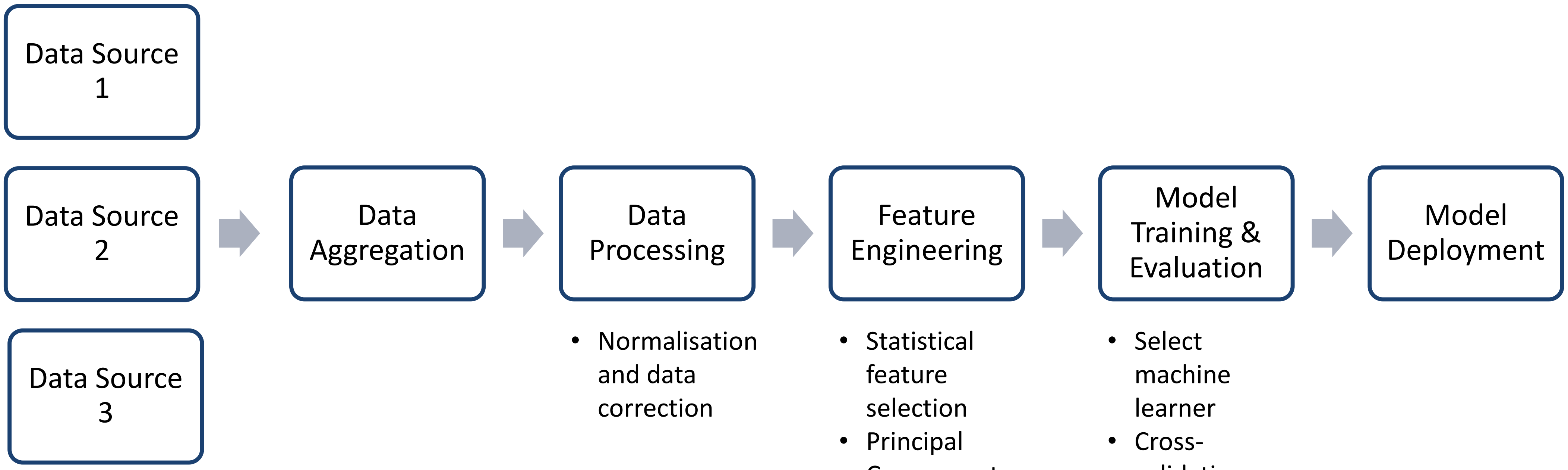
Putting it Together and Other Considerations

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

A Typical Analysis Pipeline

How a typical analysis pipeline incorporating ML can look like:



Data Processing:

- Normalisation and data correction

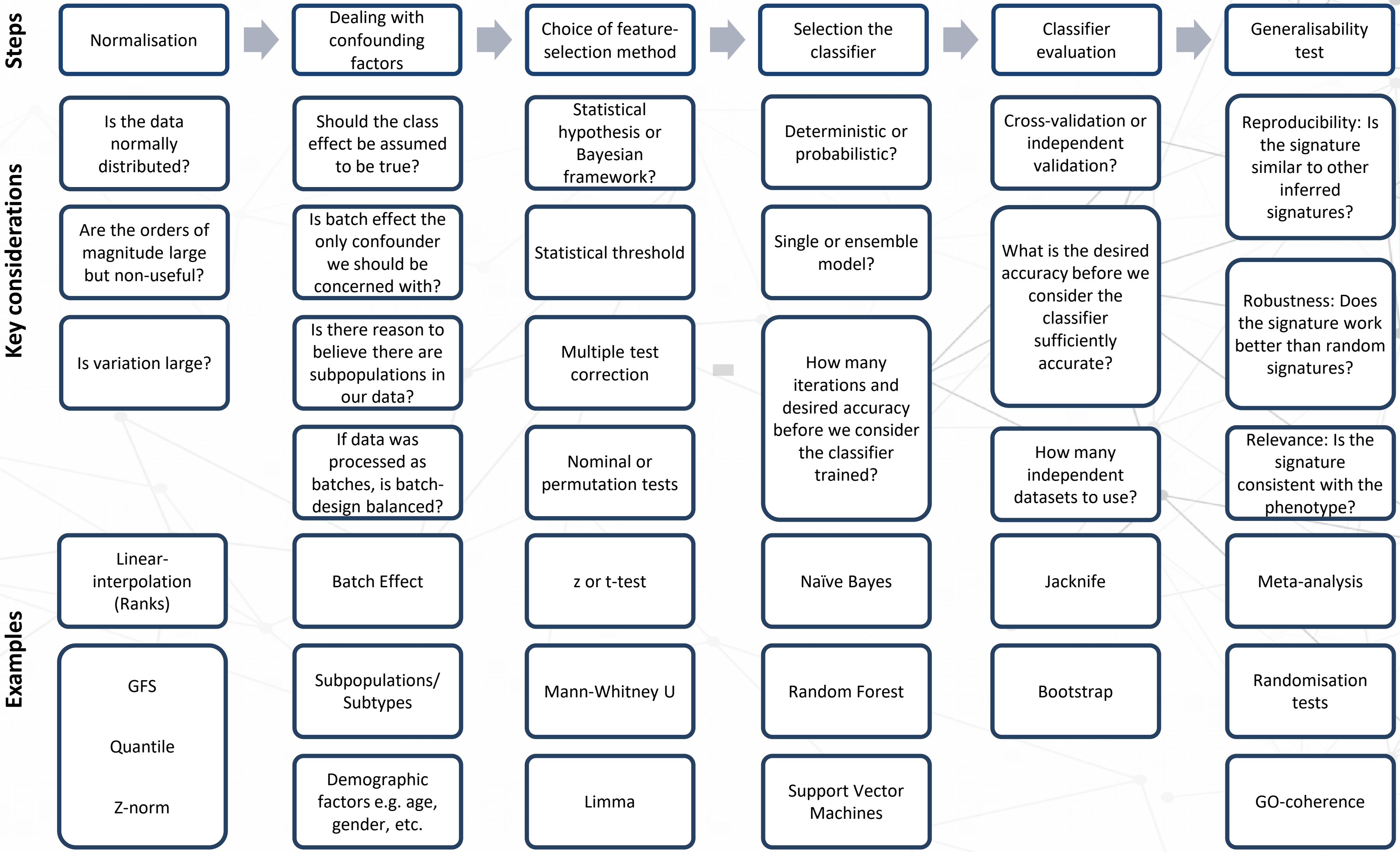
Feature Engineering:

- Statistical feature selection
- Principal Components Analysis
- Network scoring

Model Training & Evaluation:

- Select machine learner
- Cross-validation
- Independent-validation
- ROC curves

There is no One-size Fits all Strategy



Steps	Normalisation	Dealing with confounding factors	Choice of feature-selection method	Selection the classifier	Classifier evaluation	Generalisability test
Key considerations	Is the data normally distributed?	Should the class effect be assumed to be true?	Statistical hypothesis or Bayesian framework?	Deterministic or probabilistic?	Cross-validation or independent validation?	Reproducibility: Is the signature similar to other inferred signatures?
	Are the orders of magnitude large but non-useful?	Is batch effect the only confounder we should be concerned with?	Statistical threshold	Single or ensemble model?	What is the desired accuracy before we consider the classifier sufficiently accurate?	Robustness: Does the signature work better than random signatures?
	Is variation large?	Is there reason to believe there are subpopulations in our data?	Multiple test correction	How many iterations and desired accuracy before we consider the classifier trained?		
		If data was processed as batches, is batch-design balanced?	Nominal or permutation tests		How many independent datasets to use?	Relevance: Is the signature consistent with the phenotype?
Examples	Linear-interpolation (Ranks)	Batch Effect	z or t-test	Naïve Bayes	Jacknife	Meta-analysis
	GFS Quantile Z-norm	Subpopulations/ Subtypes	Mann-Whitney U	Random Forest	Bootstrap	Randomisation tests
		Demographic factors e.g. age, gender, etc.	Limma	Support Vector Machines		GO-coherence

The performance of the classifier is not independent of a large list of upstream considerations!

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

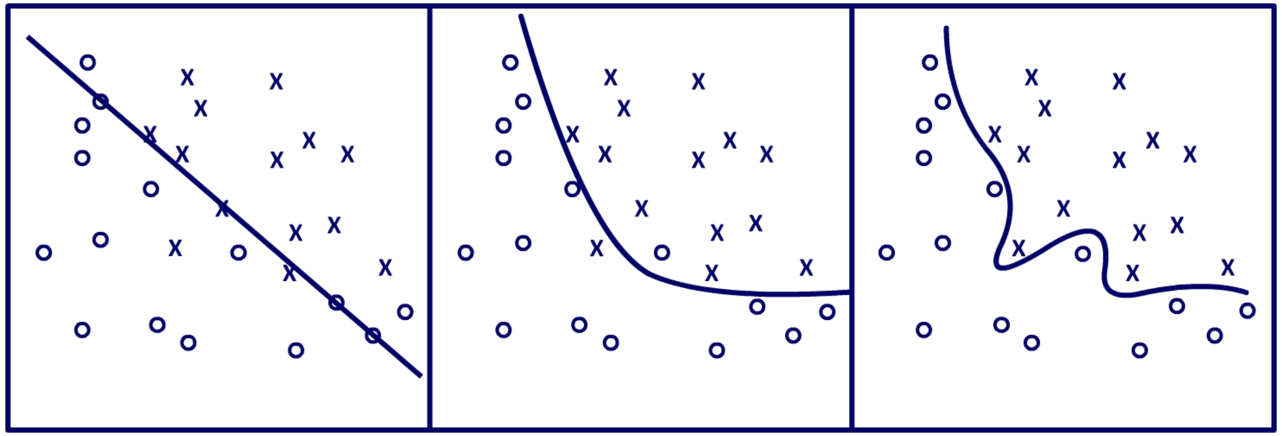
Overfitting

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Overfitting

“The most likely hypothesis is the simplest one consistent with the data.”



Inadequate Good Compromise Overfitting

- Overfitting occurs when the learner becomes so good at differentiating the test data classes to the point it fails to identify the correct generalised rules.
- Overfitting is more likely with non-parametric and non-linear models that have more flexibility when learning a target function. A classic example being random forests.
- Cross-validation and independent validation are evaluation approaches to identify and avoid overfitting.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

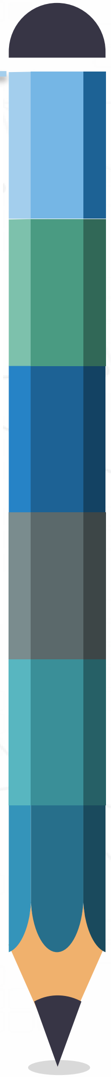
Summary

BS3033 Data Science for Biologists

Dr Wilson Goh

School of Biological Sciences

Summary



1. Machine learning methods can be broadly divided into supervised and unsupervised.

2. Decision trees are very comprehensive when variable size is small.

3. Hierarchical clustering builds clusters up iteratively based on maximising similarity.

4. Designing a machine learning analysis pipeline is very complex.

5. Always beware of overfitting.

Readings

[Machine learning] Witten, Ian. Data mining: practical machine learning tools and techniques 4th Ed. Cambridge, MA : Morgan Kaufmann, 2017. http://proquestcombo.safaribooksonline.com.ezlibproxy1.ntu.edu.sg/9780128042915 [Chapter 1, 3, 4, 5]