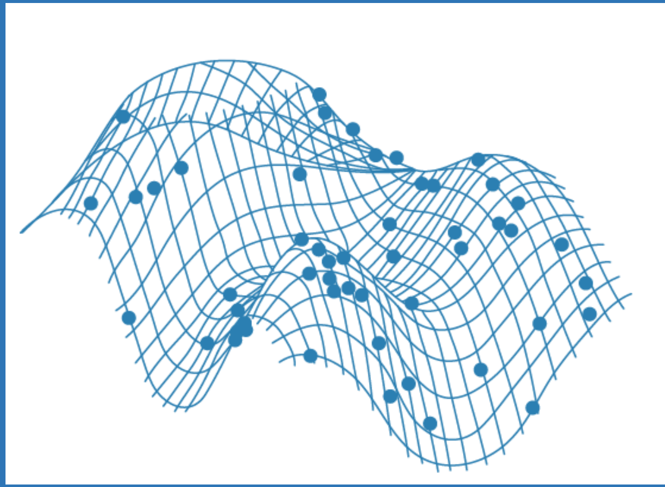
INTRODUCING MACHINE LEARNING FOR HEALTHCARE RESEARCH

Dr Stephen Weng
NIHR Research Fellow (School for Primary Care Research)
Primary Care Stratified Medicine (PRISM)
Division of Primary Care
School of Medicine
University of Nottingham



What is Machine Learning?

Machine learning teaches computers to do what comes naturally to humans and animals: learn from experience.

Machine learning algorithms use computation methods to “learn” information directly from data without relying on a predetermined equation to model.

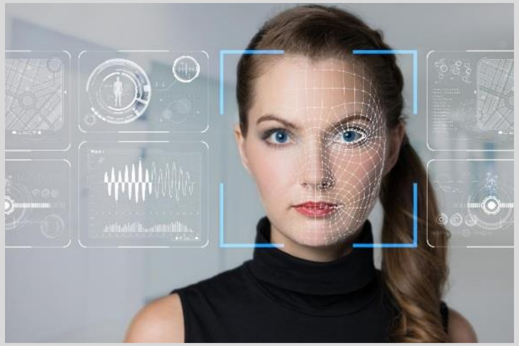
The algorithms adaptively improve their performance as the number of data samples available for learning increases.

When Should We Use Machine Learning?

Considerations:

- Complex task or problem
- Large amount of data
- Lots of variables
- No existing formula or equation
- Limited prior knowledge

Hand-written rules and equations are too complex – images, speech, linguistics

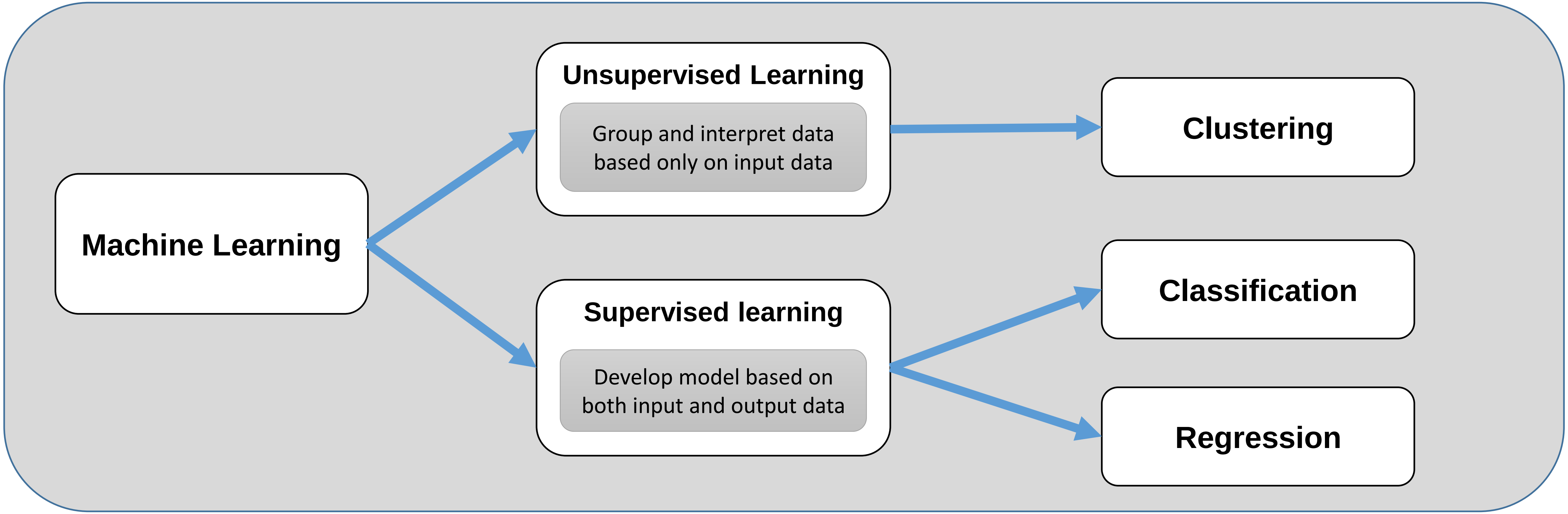


Rules of the task are dynamic – financial transactions

The nature of input and quantity of data keeps changing – hospital admissions, health care records

How Machine Learning Works

- ***Supervised learning,*** which trains a model on known inputs and output data to predict future outputs
- ***Unsupervised learning,*** which finds hidden patterns or intrinsic structures in the input data
- ***Semi-supervised learning,*** which uses a mixture of both techniques; some learning uses supervised data, some learning uses unsupervised learning



Supervised Learning

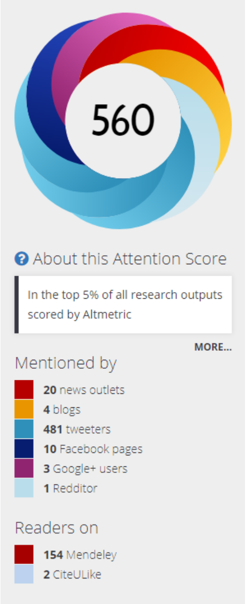
- To build a model that makes predictions based on evidence in the presence of uncertainty
- Takes a known set of input data and known responses to the data (output)
- Trains a model to generate reasonable predictions for the response to new data
  - **Classification:** predict discrete responses – for instance, whether an email is genuine or spam, or whether a tumour is cancerous or not
  - **Regression:** predict continuous response – for example, change in body mass index, cholesterol levels

Using supervised learning to predict cardiovascular disease

- Suppose we want to predict whether someone will have a heart attack in the future.
- We have data on previous patients characteristics, including biometrics, clinical history, lab tests results, co-morbidities, drug prescriptions
- Importantly, your data requires "the truth", whether or not the patient did in fact have a heart attack.

Predicting cardiovascular disease using electronic health records

- 681 UK General Practices
- 383,592 patients free from CVD registered 1st of January 2005 followed up for years
- Two-fold cross validation (similar to other epidemiological studies): n = 295,267 “training set”; n = 82,989 “validation set”
- 30 separate included features including biometrics, clinical history, lifestyle, test results, prescribing
- Four types of models: logistic, random forest, gradient boosting machines, and neural networks



Primary Care Stratified Medicine

University of Nottingham UK | CHINA | MALAYSIA

PLOS | ONE

RESEARCH ARTICLE

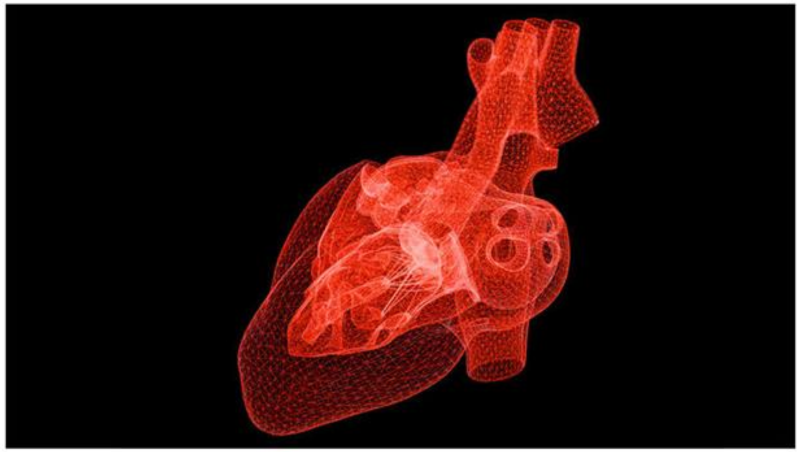
Can machine-learning improve cardiovascular risk prediction using routine clinical data?

Stephen F. Weng[1,2☯*], Jenna Reps[3,4☯], Joe Kai[1,2‡], Jonathan M. Garibaldi[3,4‡], Nadeem Qureshi[1,2‡]

1 NIHR School for Primary Care Research, University of Nottingham, Nottingham, United Kingdom, 2 Division of Primary Care, School of Medicine, University of Nottingham, Nottingham, United Kingdom, 3 Advanced Data Analysis Centre, University of Nottingham, Nottingham, United Kingdom, 4 School of Computer Science, University of Nottingham, Nottingham, United Kingdom

Science Home News Journals Topics Careers

SHARE 17K 276 4K



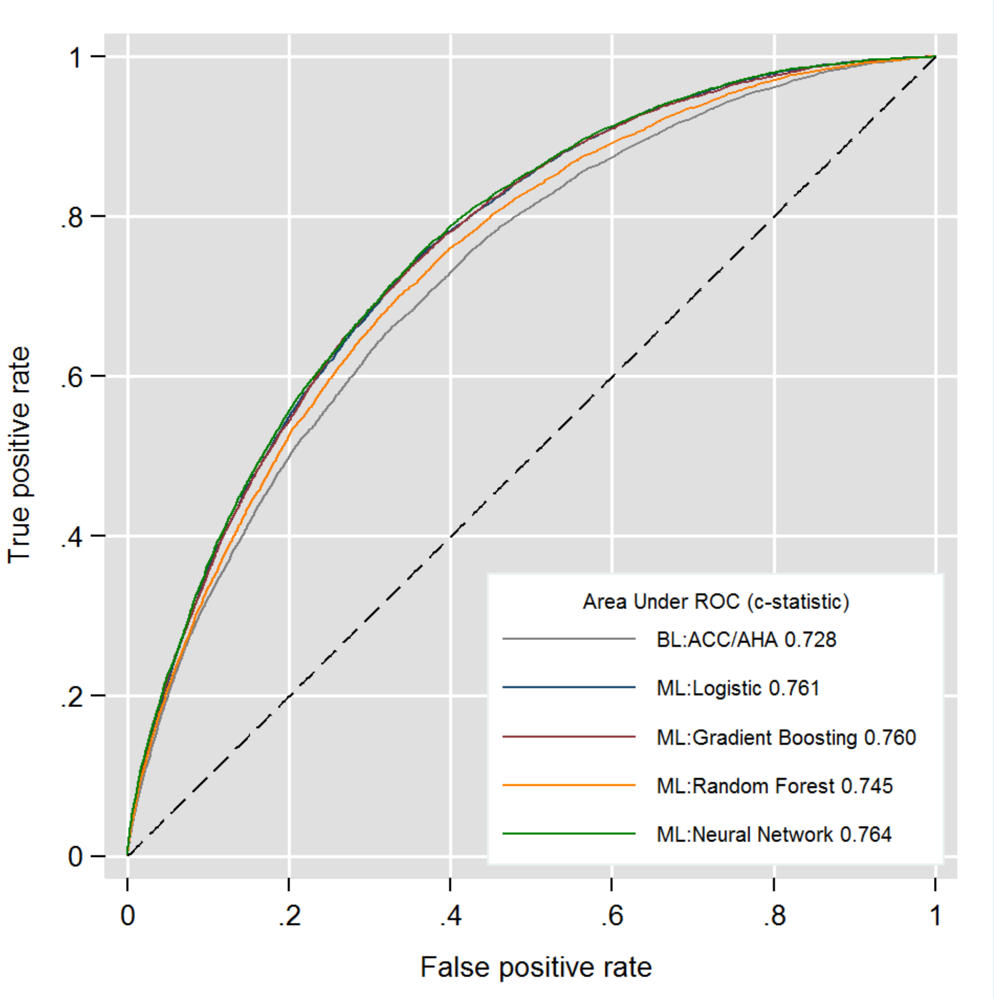
Artificial intelligence may help prevent heart failure. Devrimb/iStockphoto

Self-taught artificial intelligence beats doctors at predicting heart attacks

By Matthew Hutson | Apr. 14, 2017 , 3:30 PM

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLOS ONE 12(4): e0174944. https://doi.org/10.1371/journal.pone.0174944

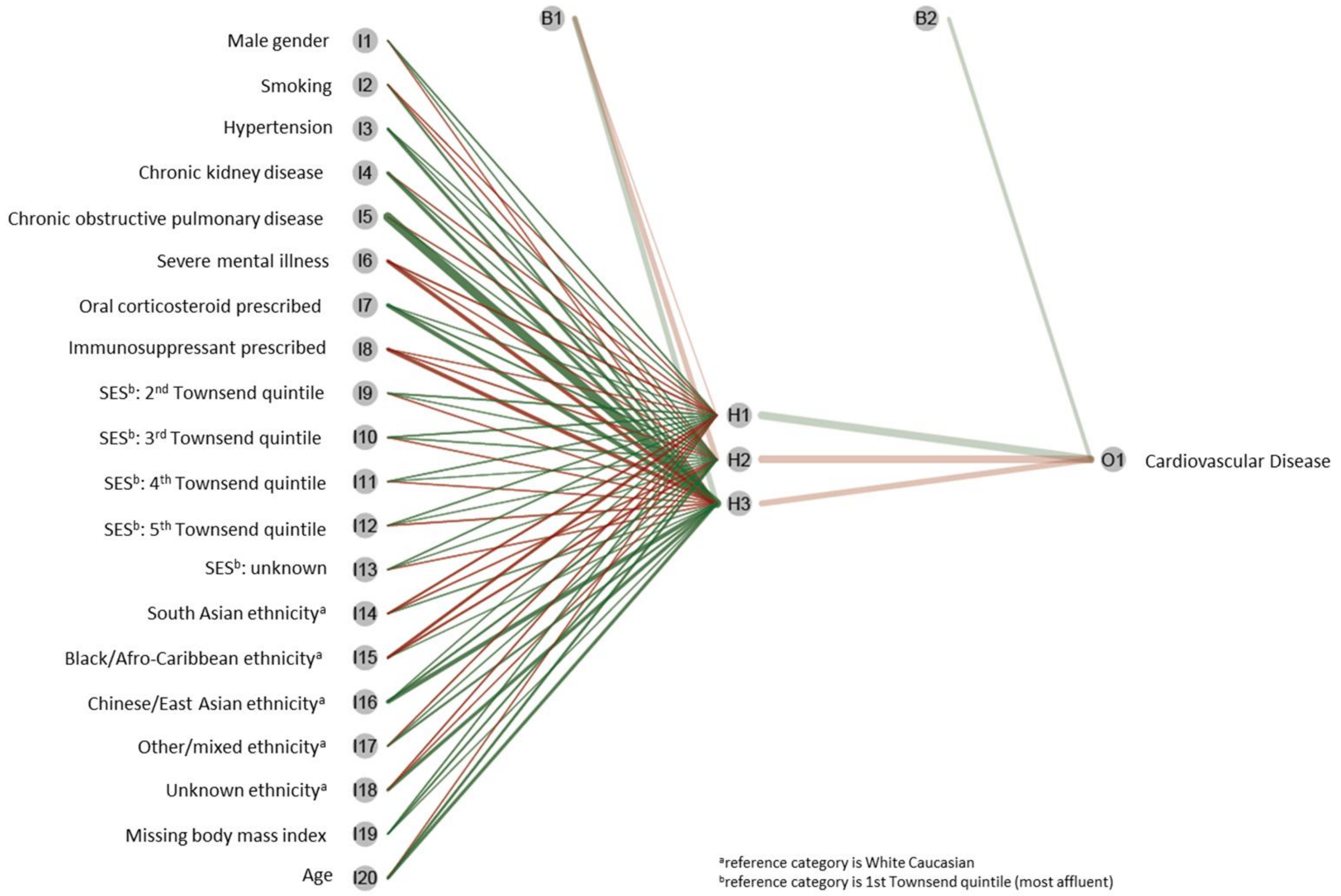
Predicting cardiovascular disease using electronic health records



Machine Learning Algorithms			
ML: Logistic Regression	**ML: Random Forest**	**ML: Gradient Boosting Machines**	**ML: Neural Networks**
Ethnicity	Age	Age	Atrial Fibrillation
Age	Gender	Gender	Ethnicity
SES: Townsend Deprivation Index	Ethnicity	Ethnicity	Oral Corticosteroid Prescribed
Gender	Smoking	Smoking	Age
Smoking	HDL cholesterol	HDL cholesterol	Severe Mental Illness
Atrial Fibrillation	HbA1c	Triglycerides	SES: Townsend Deprivation Index
Chronic Kidney Disease	Triglycerides	Total Cholesterol	Chronic Kidney Disease
Rheumatoid Arthritis	SES: Townsend Deprivation Index	HbA1c	BMI missing
Family history of premature CHD	BMI	Systolic Blood Pressure	Smoking
COPD	Total Cholesterol	SES: Townsend Deprivation Index	Gender

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLOS ONE 12(4): e0174944. https://doi.org/10.1371/journal.pone.0174944

Predicting cardiovascular disease using electronic health records



Green indicates positive weight

Red indicates negative weight

I1-I20 input variables, O1 outcome variable, H1-H3 hidden layers

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLOS ONE 12(4): e0174944. https://doi.org/10.1371/journal.pone.0174944

Unsupervised Learning

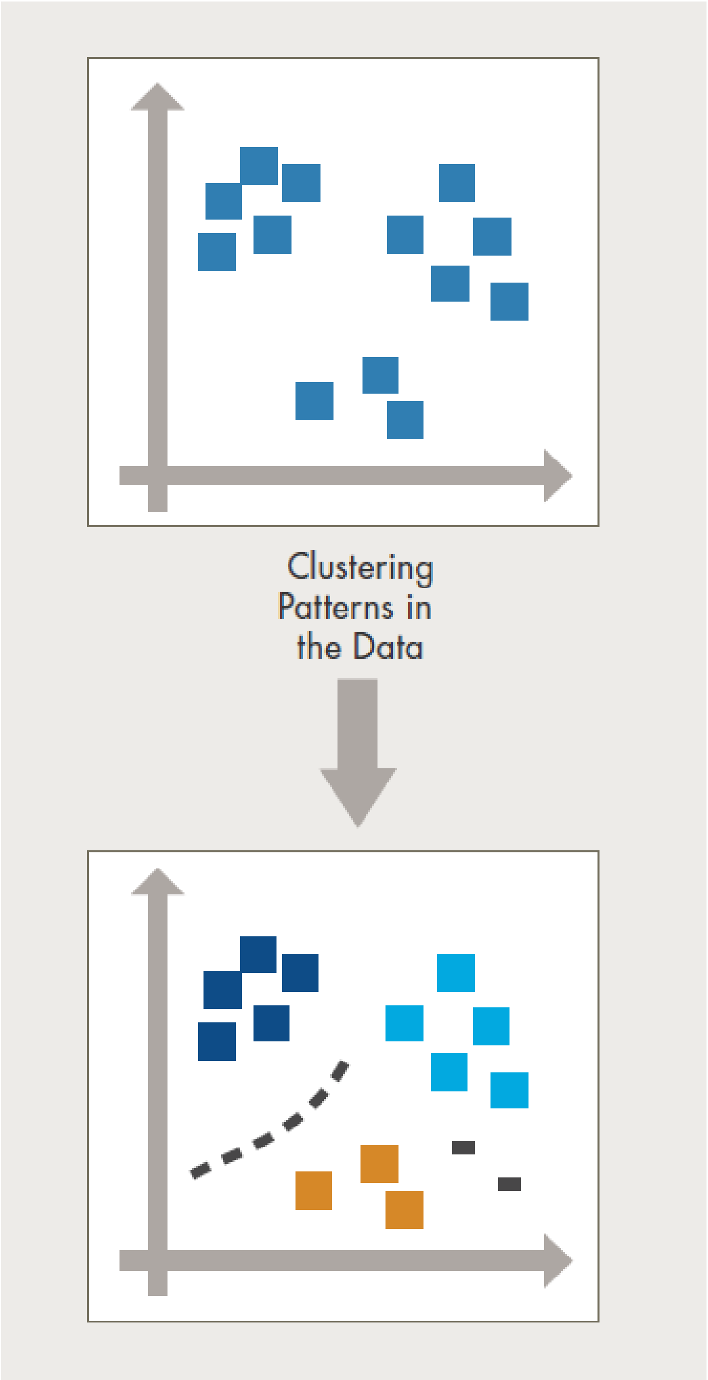
- To find hidden patterns or intrinsic structures in the data

- Primarily used to draw inferences from datasets consisting of input data without labelled responses

- Exploratory data analysis to find hidden patterns or groupings in the data

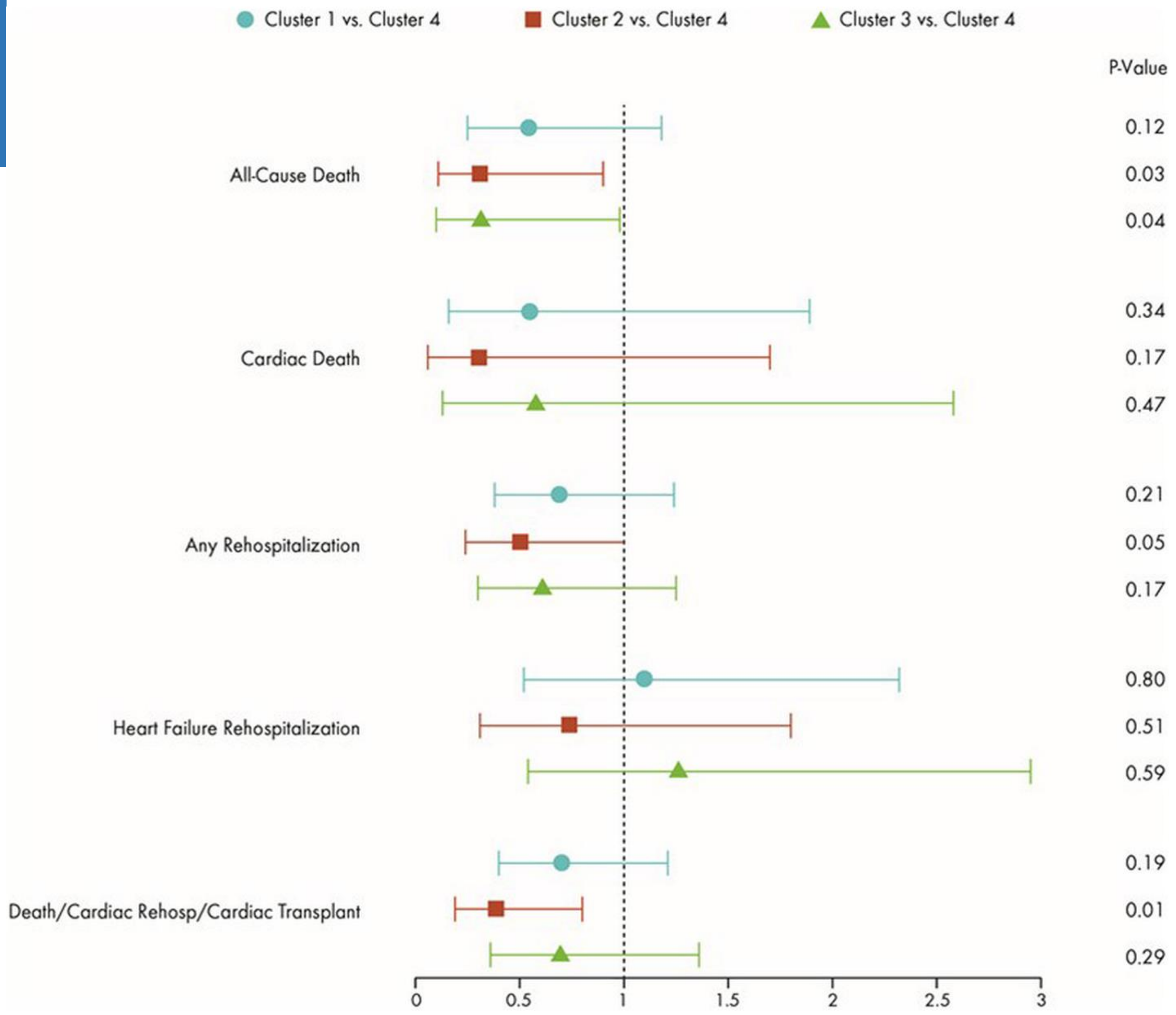
- **Clustering** is the most common unsupervised learning technique

  - Genomic sequence analysis
  - Market research
  - Objective recognition
  - Feature selection



Improving phenotyping of heart failure patients to improve therapeutic stratifies

- 172 patients hospitalised with acute decompensation heart failure from the ESCAPE trial
- Performed cluster analysis (hierarchical clustering) to determine similar patient groups based on combined measures characteristics
- Researchers conducing analysis had no knowledge of clinical outcomes for patients
- 14 candidate variables, including demographics, biometrics, cardiac biomarkers



Ahmad T, Desai N, Wilson F, Schulte P, Dunning A, et al. (2016) Clinical Implications of Cluster Analysis-Based Classification of Acute Decompensated Heart Failure and Correlation with Bedside Hemodynamic Profiles. PLOS ONE 11(2): e0145881. https://doi.org/10.1371/journal.pone.0145881

Improving phenotyping of heart failure patients to improve therapeutic stratifies

Characteristic	Cluster 1 (n = 75)	Cluster 2 (n = 33)	Cluster 3 (n = 29)	Cluster 4 (n = 25)	p-value[†]
Age, years	58 (46–67)	52 (44–59)	51 (42–57)	69 (59–79)	<0.001
Female, %	5	100	3	24	<0.001
Race					<0.001
White, %	87	52	0	76	
Minority, %	12	45	100	24	
Ischemic etiology, %	65	30	10	84	<0.001
LVEF, %	20 (15–23)	20 (15–25)	15 (13–18)	20 (19–25)	0.001
BMI, kg/m^2	29 (25–34)	26 (23–36)	28 (25–30)	24 (22–26)	0.013
Edema, %	72	56	79	60	0.145
Symptom score	40 (30–60)	44 (30–60)	35 (20–50)	50 (34–60)	0.295
MLHF score	78 (68–87)	76 (63–95)	83 (72–89)	74 (64–78)	0.212
Orthopnea, %	88	85	86	76	0.529
SBP, mmHg	100 (90–111)	109 (97–120)	110 (103–124)	100 (90–114)	0.005
DBP, mean	65 (60–70)	66 (56–70)	76 (68–85)	59 (55–70)	<0.001
Atrial fibrillation, %	44	15	7	24	<0.001
Angina pectoris, %	36	21	21	44	0.127
Prior CABG, %	32	15	7	64	<0.001
COPD, %	13	9	24	24	0.235
Depression, %	21	27	14	20	0.634
Diabetes, %	39	30	25	40	0.493
Hypertension, %	43	49	62	28	0.084
ICD, %	33	12	28	28	0.156
CVA, %	12	6	3.4	8	0.601
Peak VO_2, mL/kg/min	10.4 (8.0–11.9)	9.1 (7.3–10.6)	8.7 (7.6–9.3)	9.0 (7.6–10.4)	0.517
RAP, mmHg	13 (8–18)	11 (6–14)	17 (13–22)	14 (9–20)	0.005
PCWP, mmHg	27 (19–34)	22 (15–28)	32 (28–38)	23 (20–27)	<0.001
Cardiac index, $L/min/m^2$	1.9 (1.6–2.3)	2.0 (1.5–2.2)	1.6 (1.2–2.2)	1.8 (1.6–2.5)	0.120
Sodium, mEq/L	137 (134–139)	138 (136–139)	137 (136–139)	136 (134–138)	0.403
BUN, mg/dL	29 (20–41)	20 (12–26)	29 (23–41)	80 (47–98)	<0.001
Creatinine, mg/dL	1.4 (1.2–1.6)	0.9 (0.9–1.2)	1.4 (1.3–1.8)	2.5 (2.1–3.1)	<0.001
BNP, pg/mol	469 (174–963)	489 (183–860)	877 (89–1391)	1398 (518–4513)	0.001

Ahmad T, Desai N, Wilson F, Schulte P, Dunning A, et al. (2016) Clinical Implications of Cluster Analysis-Based Classification of Acute Decompensated Heart Failure and Correlation with Bedside Hemodynamic Profiles. PLOS ONE 11(2): e0145881. https://doi.org/10.1371/journal.pone.0145881

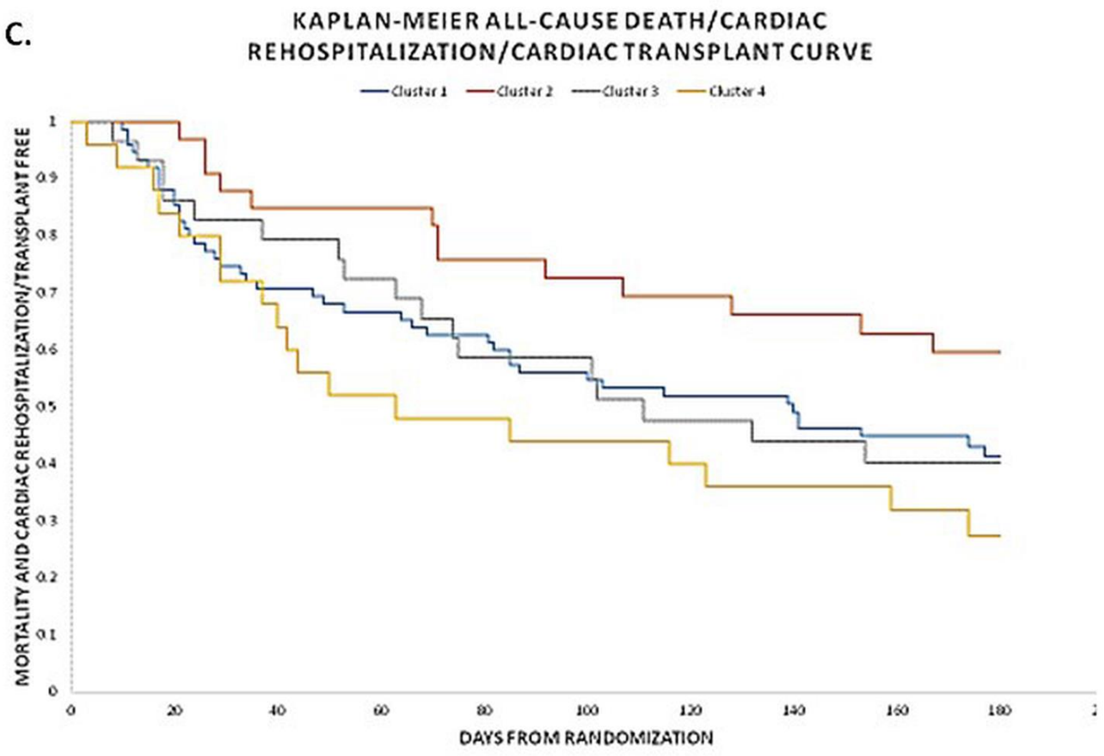
Improving phenotyping of heart failure patients to improve therapeutic stratifies

- Cluster 1: male Caucasians with ischemic cardiomyopathy, multiple comorbidities, lowest BNP levels

- Cluster 2: females with non-ischemic cardiomyopathy, few co-morbidities, most favourable hemodynamics, advanced disease

- Cluster 3: young African American males with non-ischemic cardiomyopathy, most adverse hemodynamics, advanced disease

- Cluster 4: older Caucasians with ischemic cardiomyopathy, concomitant renal insufficiency, highest BNP levels



- Cluster 2 least adverse outcomes, Cluster 4 worst outcomes

- Cluster 1-3 had 45-70% lower risk of all-cause mortality

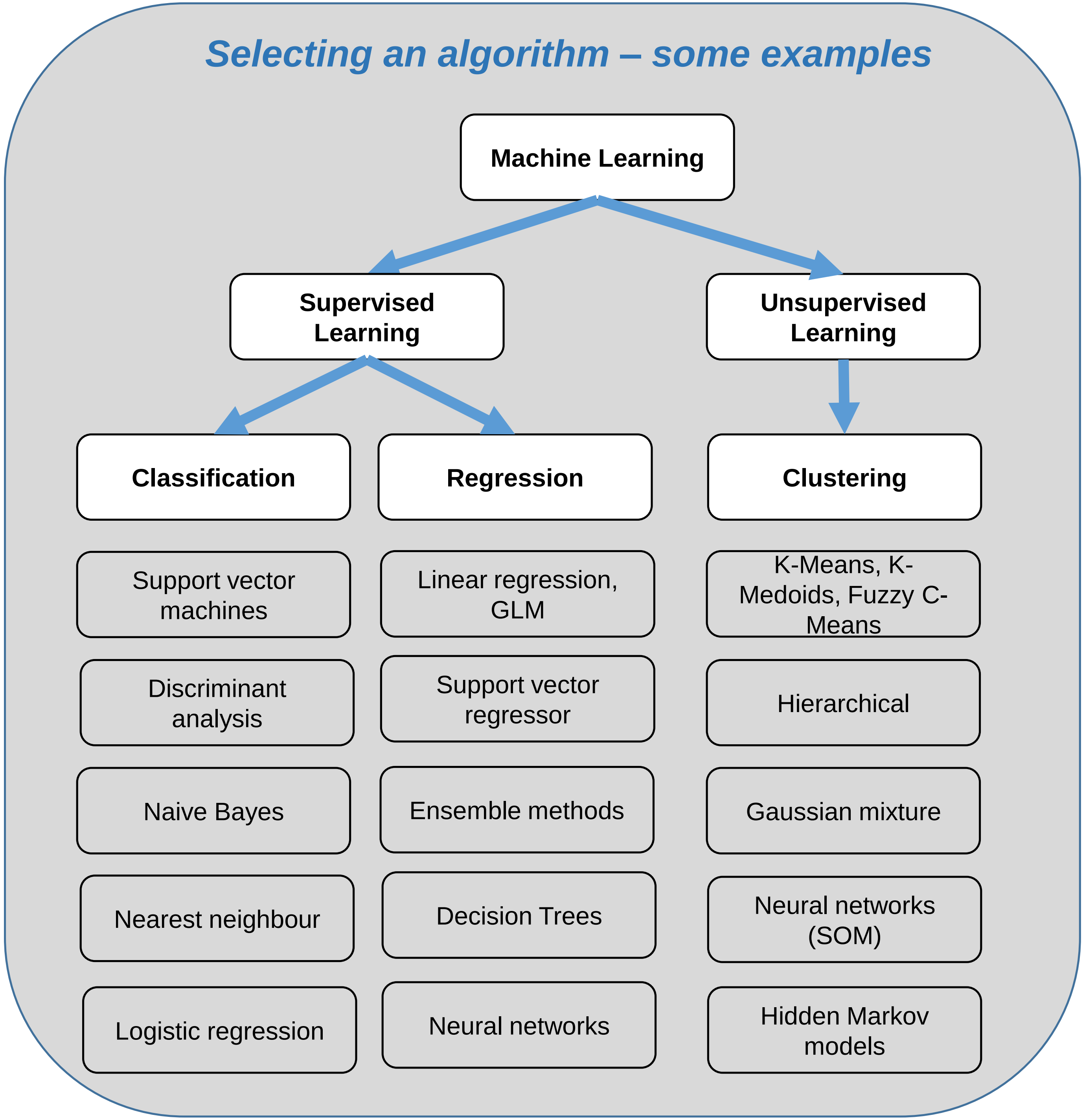
Ahmad T, Desai N, Wilson F, Schulte P, Dunning A, et al. (2016) Clinical Implications of Cluster Analysis-Based Classification of Acute Decompensated Heart Failure and Correlation with Bedside Hemodynamic Profiles. PLOS ONE 11(2): e0145881. https://doi.org/10.1371/journal.pone.0145881

How do you decide which algorithm to use?

Choosing the right algorithm can seem overwhelming – there are about a dozen supervised and unsupervised learning algorithms, each taking a different approach.

Considerations:

- ❖ There is no best method or one size fits all
- ❖ Trial and error
- ❖ Size and type of data
- ❖ The research question and purpose
- ❖ How will the outputs be used?



Supervised Learning

Supervised learning algorithm takes a known set of input data (the training set) and known responses to the data (output), and trains a model to generate reasonable predictions for the response to new input data.

Use supervised learning if you have existing data for the output you are trying to predict

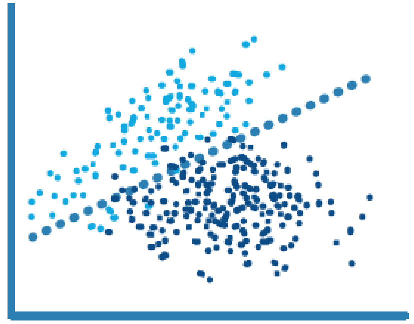
Using larger training datasets yield models that generalise better for new data

Common classification algorithms

Logistic regression

How it works
- Fits a model that can predict the probability of a binary response belonging to one class or the other
- Simple – commonly used a starting point for binary classification problems

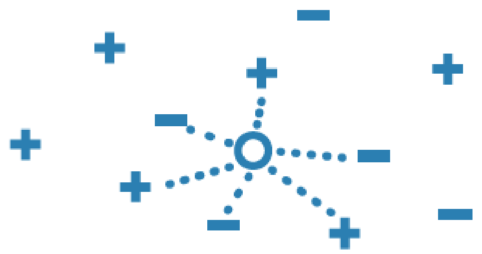
Best used…
- When data can be clearly separated by a single, linear boundary
- Baseline for evaluating more complex classification methods



k Nearest Neighbour (kNN)

How it works
- Categorises objects based on the classes of their nearest neighbours in the dataset
- Assume that objects near each other are similar
- Distance metrics used to determine nearness (e.g. Euclidean)

Best used…
- When you need a simple algorithm to establish benchmark learning rules
- When memory usage and prediction speed is a lesser concern



Common classification algorithms

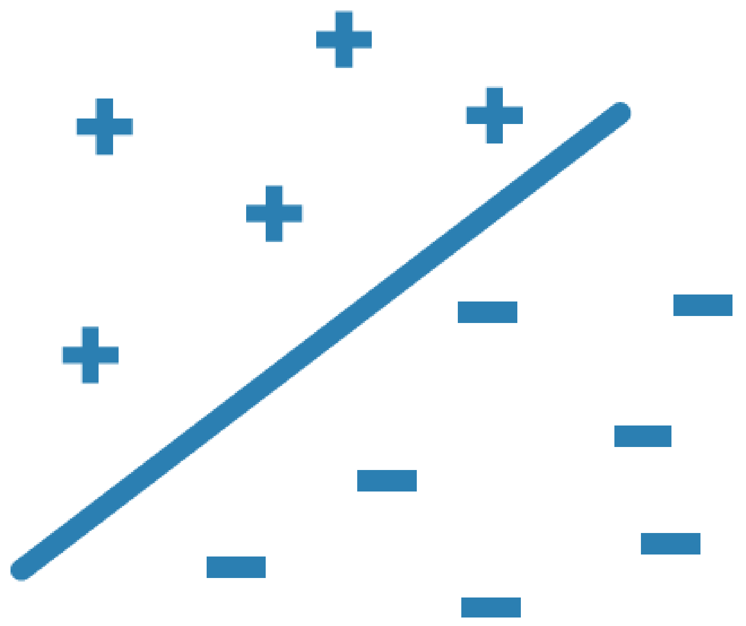
Support vector machine (SVM)

How it works

- Classifies data by finding the linear decision boundary (hyperplane) that separates all data points of on class from that of another class
- Points on the wrong side of the hyperplane is penalised using a loss function
- Uses a kernel transformation to transform non-linearly separable data into higher dimensions where a linear decision boundary can be found

Best used…

- Data that has exactly two classes (binary)
- High dimensional, non-linearly separable
- Need a classifier that's simple, easy to interpret, and accurate



Common classification algorithms

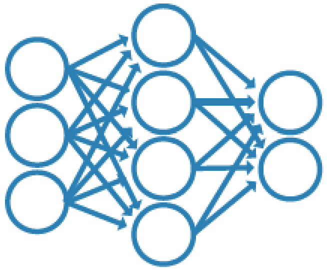
Neural Network

How it works

- Consists of highly connected networks of neurons that relate the inputs to the desire outputs
- Network is trained by iteratively modifying the strengths of the connections so that a given input maps to the correct responses

Best used…

- Modelling highly non-linear systems
- Data is available incrementally and you wish to constantly update the model
- There may be unexpected changes in your input data
- When model interpretability is not a key concern



Naïve Bayes

How it works

- Assumes that the presence of a particular feature in a class is unrelated to the presence of another feature
- Data is classified on the highest probability of its belonging to a particular class

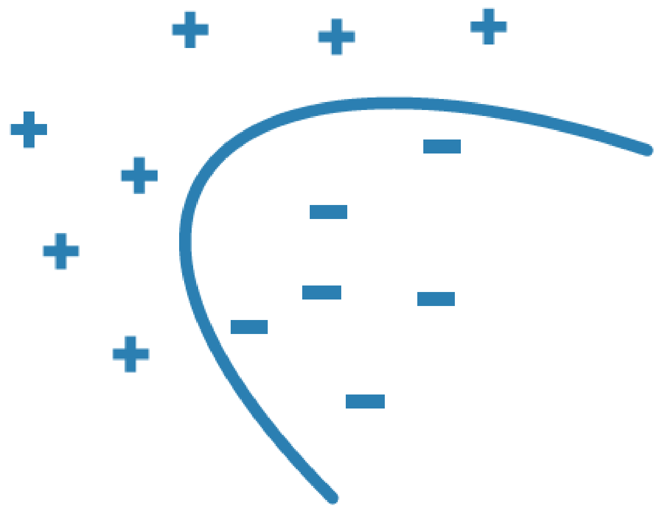
Best used…

- Small dataset containing many parameters
- Need a classifier that's easy to interpret
- Model will encounter scenarios that weren't in the training data

Common classification algorithms

Discriminant analysis

How it works
- Classifies data by finding linear combinations of features
- Assumes that different classes generate data based on Gaussian distributions
- Training involves finding the parameters for a Gaussian distribution for each class
- Distribution parameters used to calculate boundaries, which can be linear or quadratic functions
- The boundaries are used to determine new class of data



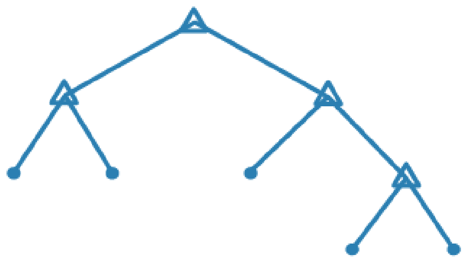
Best used…
- Easy to interpret and generates a simple model
- Efficient – memory usage and modelling speed is fast

Common classification algorithms

Decision Tree

How it works
- Predict responses to data by following the decisions in the tree from the root down to a leaf node
- Branching conditions where the value of a predictor is compared to a trainer weight
- The number of branches and values of the weights are determined in the training process

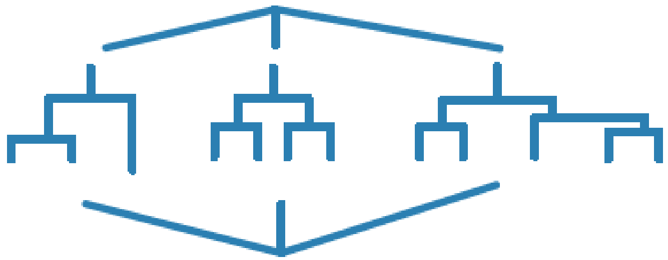
Best used…
- Need an algorithm that is easy to interpret and fast to fit
- Minimise memory usage
- High predictive accuracy is not a requirement



Bagged and Boosted Decision Tree (Ensemble)

How it works
- Several "weaker" decision trees are combined into a "stronger" ensemble
- Bagging – trees are trained independently on data that is bootstrapped from the input data
- Boosting – iteratively add "weak" learner models and adjusting weight of each weak learner to focus on misclassified examples

Best used…
- Predictors are categorical or behave non-linearly
- Time to train model is less concern

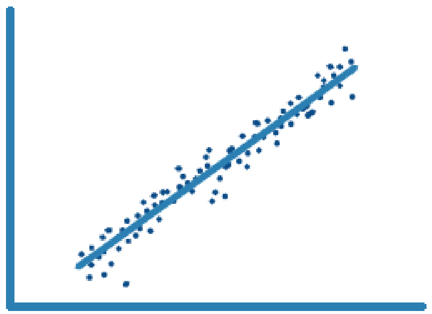


Common regression algorithms

Linear regression

How it works
- Used to describe a continuous response variable as a linear function of one or more predictor variables

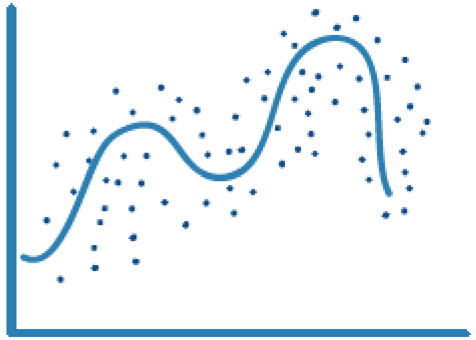
Best used…
- Easy to interpret and fast to fit
- Baseline for evaluating other, more complex regression models



Nonlinear regression

How it works
- Models described as a nonlinear equation
- Nonlinear refers to a fit function that is a nonlinear function of the parameters

Best used…
- Data has strong nonlinear trends and cannot be easily transformed into a linear space
- For fitting custom models to data

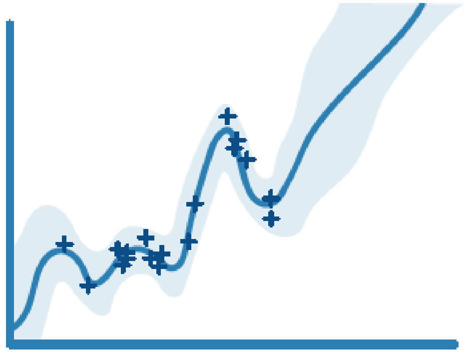


Common regression algorithms

Gaussian process regression model

How it works
- Nonparametric models used for predicting value of a continuous response variable
- Spatial analysis for interpolation in the presence of uncertainty

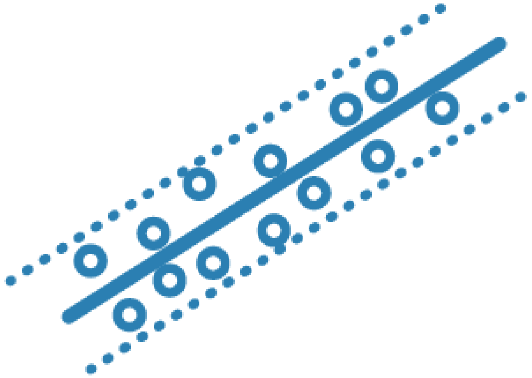
Best used…
- For interpolating spatial data
- Facilitate optimisation of complex systems/designs



Support vector regressor

How it works
- Similar to support vector for classification but are modified to be able to predict continuous response
- Does not fit a hyperplane but rather a model that deviates from the measure data by no greater than a small amount (error)

Best used…
- High dimensional data (where there is a large number of predictor variables)



Common regression algorithms

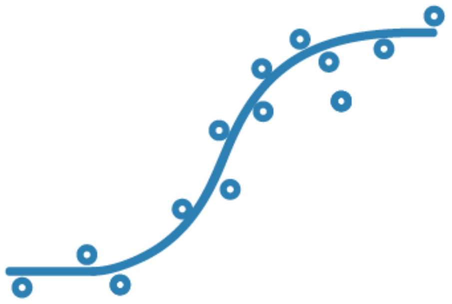
Generalised linear model

How it works

- Special case of a nonlinear model that uses linear methods
- Involves fitting a linear combination of the inputs to a non-linear function (link function) of the outputs

Best used…

- When the response variables have non-normal distributions, such as a response variable that is always expected to be positive



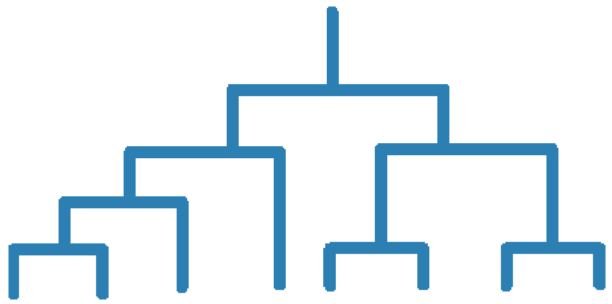
Regression tree

How it works

- Decision trees for regression are similar to decision trees for classification, but modified to be able to predict continuous responses

Best used…

- Predictors are categorical (discrete) or behave nonlinearly



Unsupervised Learning

Unsupervised learning is useful when you want to explore your data but don't yet have a specific goal or are not sure what information the data contains.

It's a good way to reduce the dimensionality of your data

Clustering algorithms call into two broad groups:

Hard clustering: each data point only belongs to one group

Soft clustering: each data point can belong to more than one group

Common *hard* clustering algorithms

***k* Means**

How it works

- Partitions data into *k* number of mutually exclusive clusters
- Determined by distance from particular point to the cluster's centre

Best used…

- When the number of clusters is known
- For fast clustering of large datasets

***k* Medoids**

How it works

- Similar to *k* Means but with requirement that the cluster centres coincide with the points in the data

Best used…

- When the number of clusters is known
- For fast clustering of categorical data
- Large datasets

Common *hard* clustering algorithms

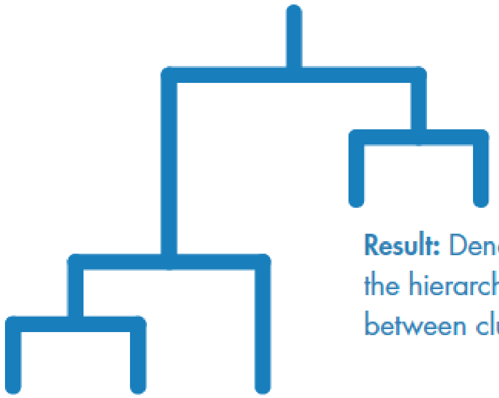
Hierarchical clustering

How it works

- Produces nested sets of clusters by analysing similarities between pairs of points
- Grouping objects into a binary hierarchical tree

Best used…

- When you don't know how many clusters are in your data
- You want to visualisation to guide your selection



Result: Dendrogram showing the hierarchical relationship between clusters

Self organising map

How it works

- Neural network based clustering that transform a dataset into a topology-preserving 2D heat map

Best used…

- To visualise high-dimensional data in 2D or 3D
- To reduce to dimensionality of the data

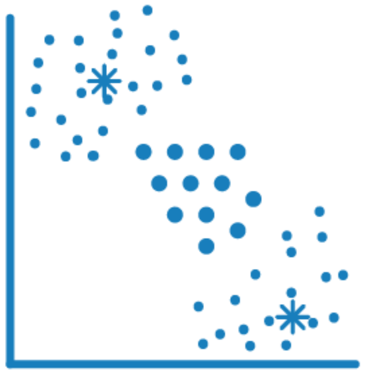
Result: Lower-dimensional (typically 2D) representation

Common *soft* clustering algorithms

Fuzzy c-Means

How it works
- Partition-based clustering when data points may belong to more than one cluster

Best used…
- When the number of clusters is known
- For pattern recognition
- When clusters overlap



Result: Cluster centers (similar to k-means) but with fuzziness so that points may belong to more than one cluster

Gaussian mixture model

How it works
- Partition-based clustering where data points come from different multivariate normal distributions with certain probabilities

Best used…
- When a data point might belong to more than one cluster
- When clusters have difference sizes and correlation structures within them

Result: A model of Gaussian distributions that give probabilities of a point being in a cluster

Key challenges for healthcare data

Most challenges come from handling your data and finding the “right” model

- **Data comes in all shapes and sizes:** Real-world datasets are messy, incomplete, and come in a variety of formats
- **Pre-processing your data requires clinical knowledge and the right tools:** For example to select the correct features (variables) and codes to use in primary care datasets, you’ll need clinical verification and knowledge of NHS coding and content expertise
- **Can your question be answered without ML:** many research questions don’t actually require ML. For instance, accurate risk prediction models can be developed stepwise regression models.
- **Choosing the “right” model:** Highly flexible models tend to over-fit while simple models make too many assumptions. Trial and error is at the core of machine learning
- **Understand the limitations:** Not recommended for causal inferences, interpretation of results can be difficult

Simplified workflow

1. ACCESS: format and load the data

2. PREPROCESS: data management, cleaning, coding, organising

3. DERIVE: features (variables) using the cleaned data

5. TRAINING: select algorithm, train models using derived features

6. ITERATE: different algorithms to find the best model

7. VALIDATE: trained model on separate dataset

8. INTERPRETATION: clinical verification and interpretation of outputs

9. DISSEMINATION: integrate into production system/publish in journals

Popular Programmes

https://www.r-project.org/

http://workspace.nottingham.ac.uk/display/Software/Matlab

https://www.rstudio.com/

https://azure.microsoft.com/en-gb/pricing/

https://www.python.org/

https://spark.apache.org/

https://anaconda.org/anaconda/python

Open Source Training

Follow these tutorial for Deep Learning:
http://rstudio.github.io/sparklyr/articles/guides-h2o.html (simple)
- Uses in built R library dataset ‘mtcars’

https://shiring.github.io/machine_learning/2017/02/27/h2o (advanced)
- Download external open access dataset from https://archive.ics.uci.edu/ml/datasets/arrhythmia

Follow this tutorial for Neural Networks:
https://datascienceplus.com/fitting-neural-network-in-r/
- Uses in built R library dataset ‘MASS’

Follow this tutorial for Hierarchical Clustering:
http://uc-r.github.io/hc_clustering
- Uses in built R library dataset ‘USArrests’