NDORMS

NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

Medical Sciences Division

Big Data Methods

An Introduction to Machine Learning and Real-world Applications

Sara Khalid

NDORMS, University of Oxford

Outline

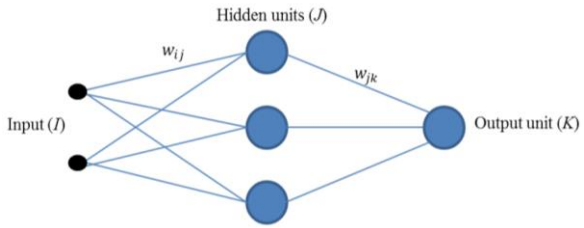
Part 1: Big Data and ML

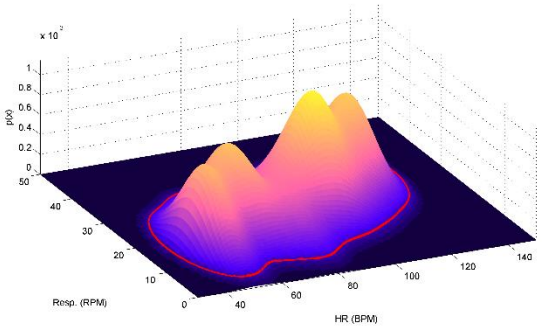
Part 2: ML Tasks

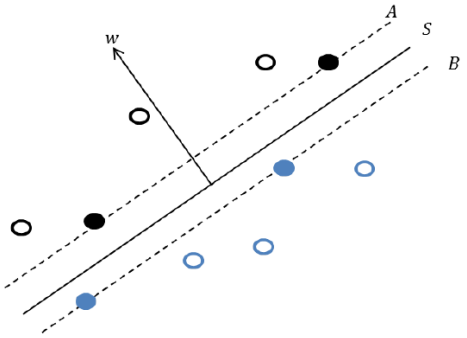
Part 3: ML Algorithms

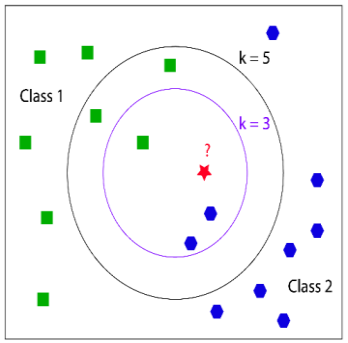
Part 4: Real World Applications

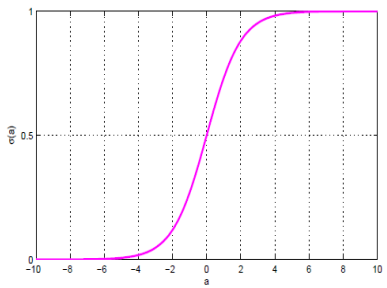
Part 1: Big Data and Machine Learning

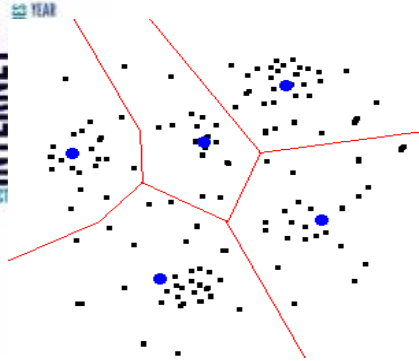












The Data Deluge

The Economist (2010)

- 2.5 billion GB of data created daily (2015)
- 200 ZB of data by 2025

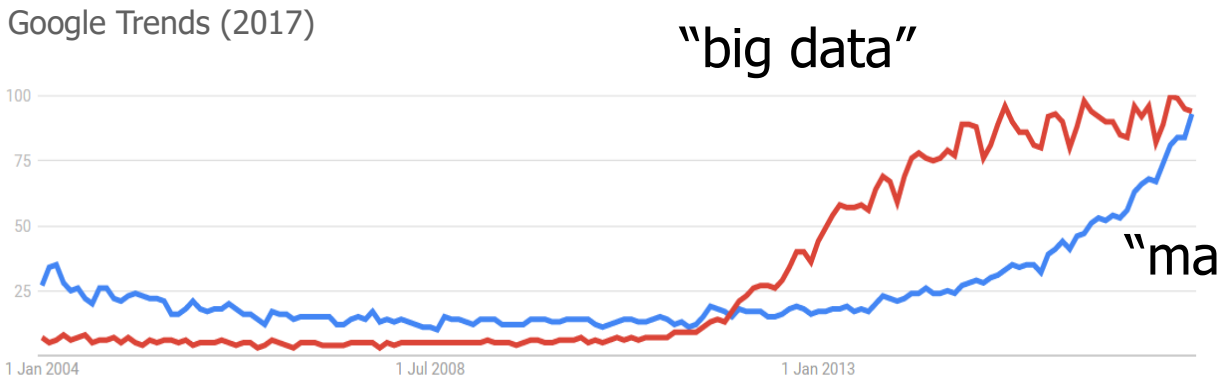
IBM (2015), Bloomberg (2017)

The World's Most Valuable Resource?

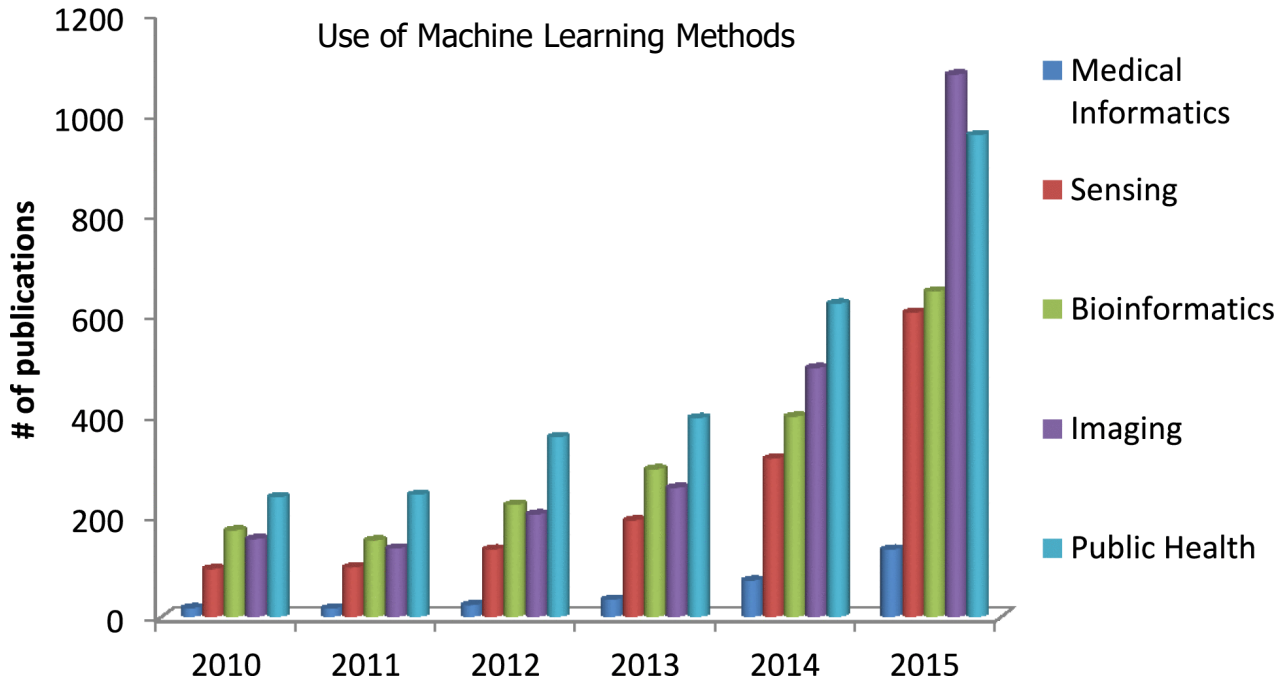
The Economist (2017)

- Data are cheap, knowledge is scarce
- Need for knowledge discovery methods
- Accurate, efficient

Big Data and Machine Learning



Big Data and Machine Learning in Healthcare



Ravi et al., IEEE Biomed & Health Informatics (2017)

What is Machine Learning?

"a field of study that gives (a) computer the ability without being explicitly programmed"

Arthur Samuel, created game of checkers that improved by playing by itself (1952)

Herbert Simon, created Logic Theory Machine, "first AI programme" (1956)

What is Machine Learning?

- A branch of **artificial intelligence**, concerned with design and development of algorithms that allow computers to evolve behaviors based on empirical data

- Routinely used commercially for speech recognition, cancer diagnosis, eye disease detection, etc.

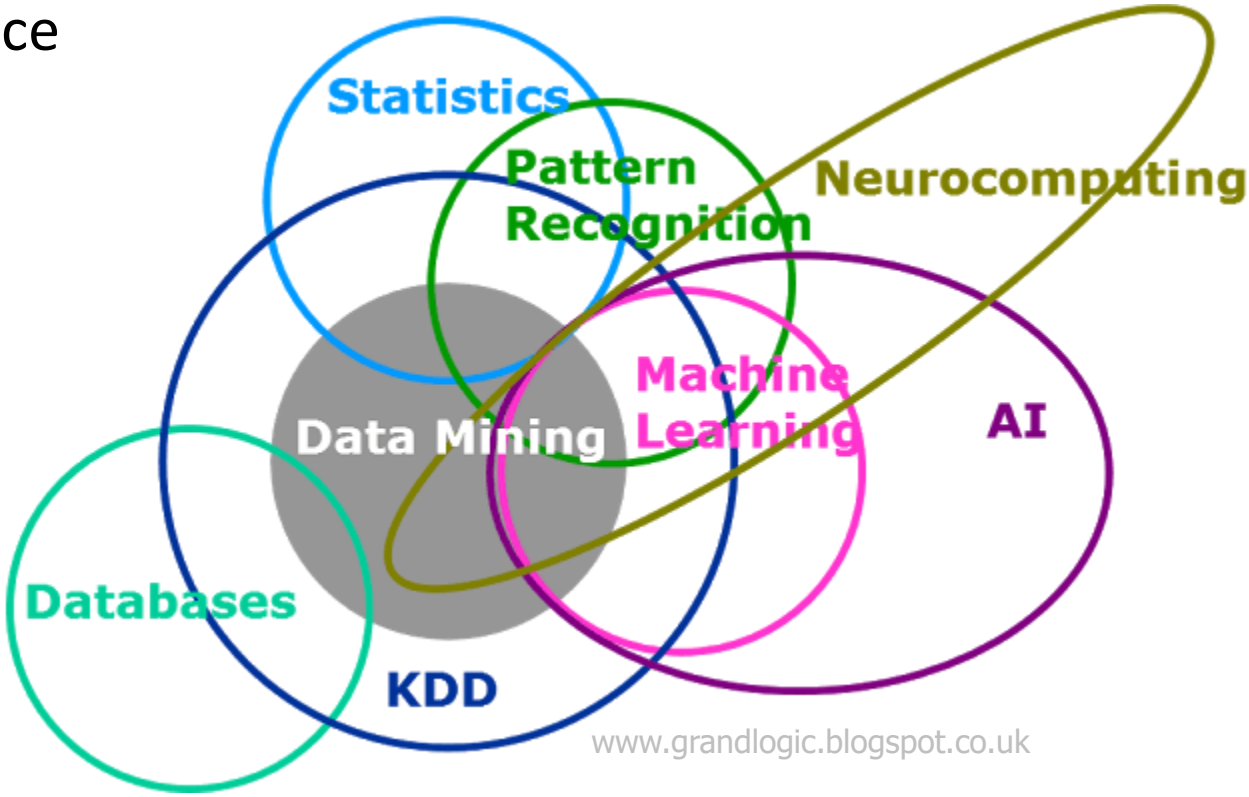
- Ability to acquire and process **big data**

- We call this paradigm shift **deep learning**

Mitchell, The Discipline of Machine Learning, 2006

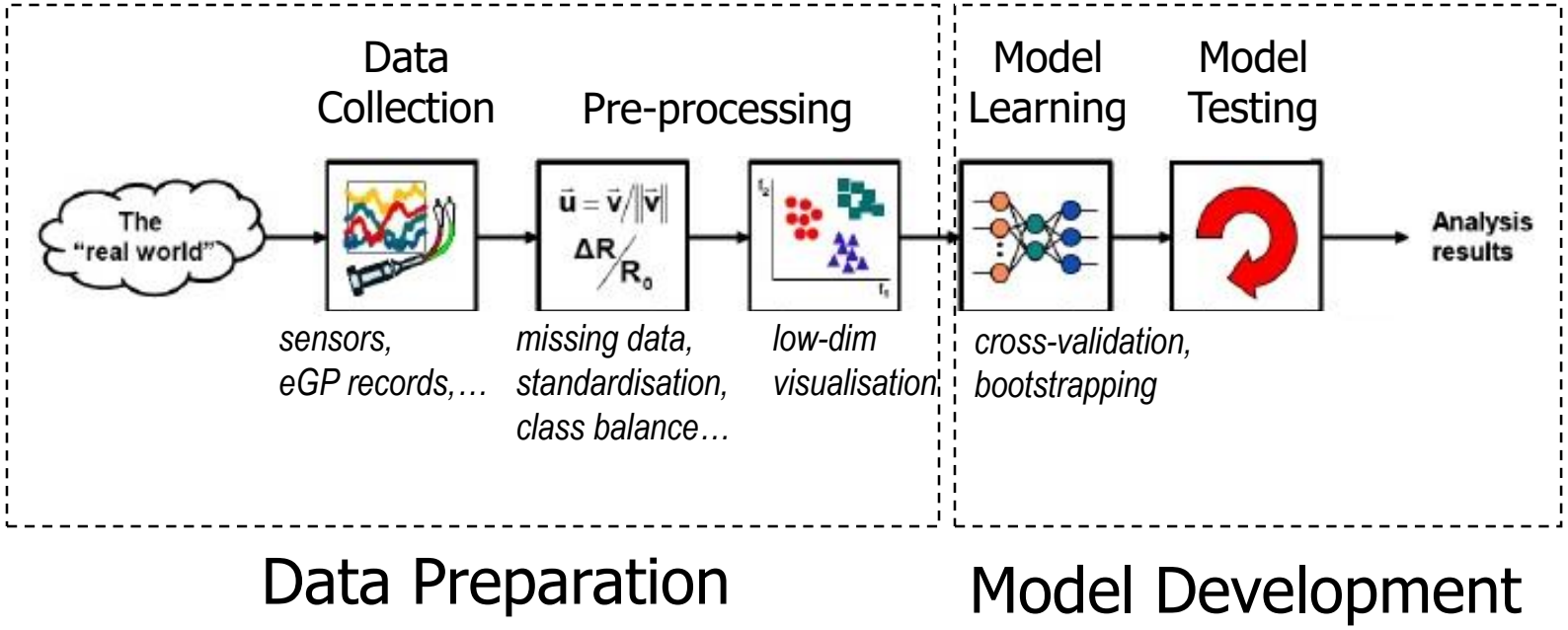
Roots and Links

- Statistics
- Neuroscience
- Artificial Intelligence
- Psychology



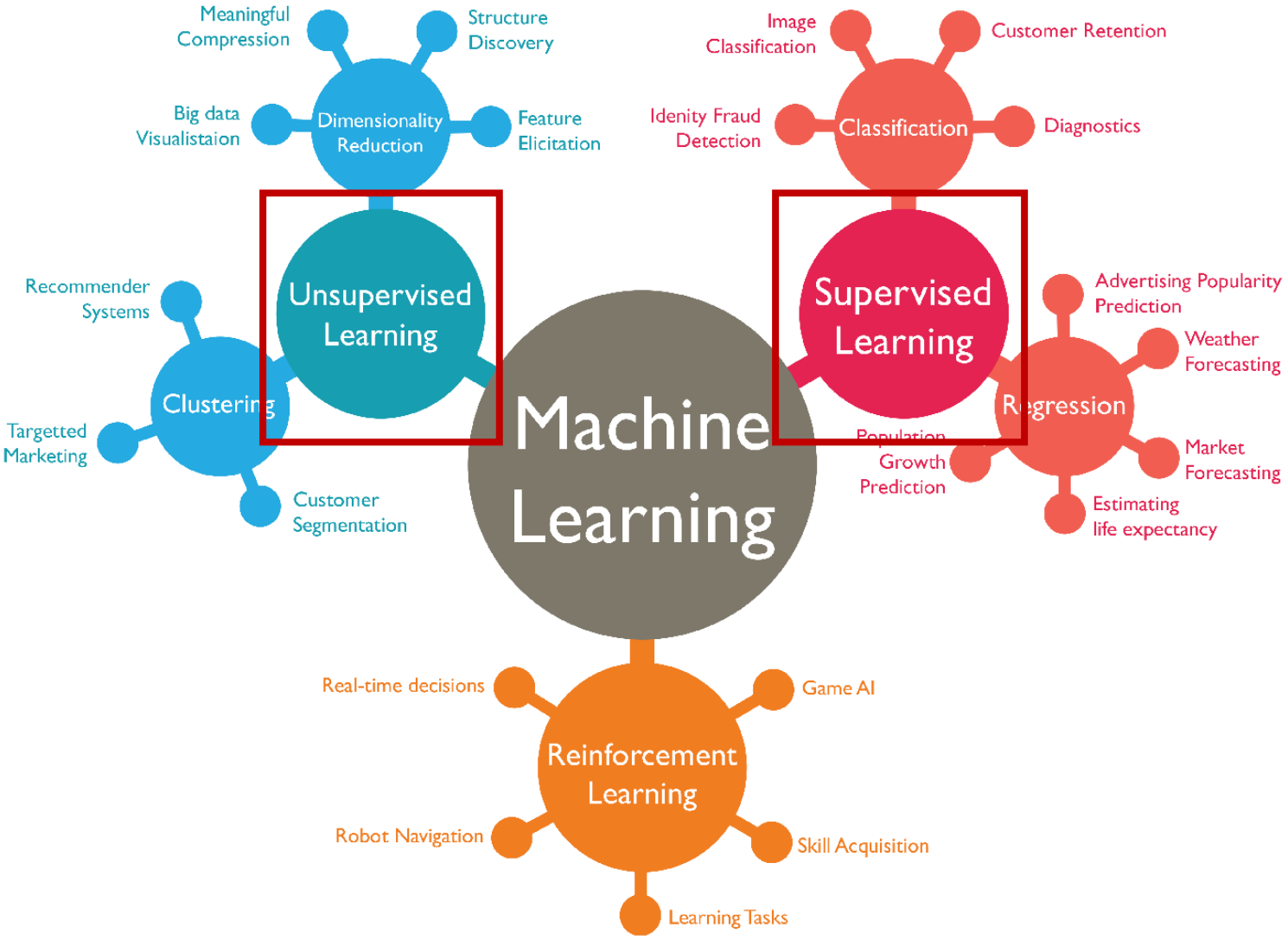
Part 2: What can Machine Learning do?

The Data Analysis Process

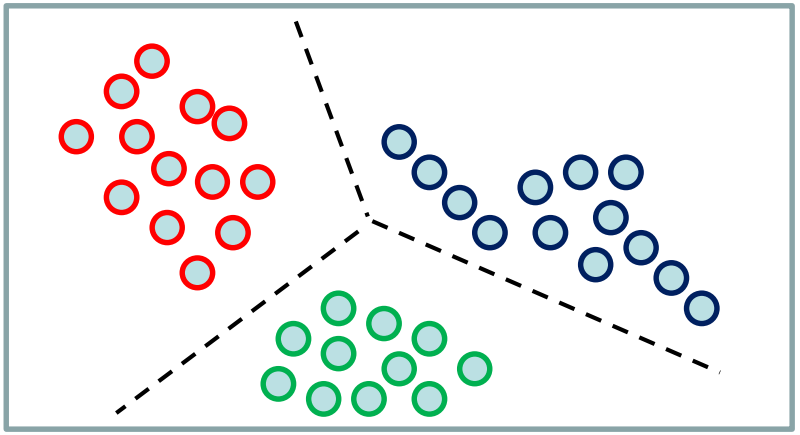


Adapted from Lior Rokach, 2017

Machine Learning Tasks



Supervised Learning



TASKS

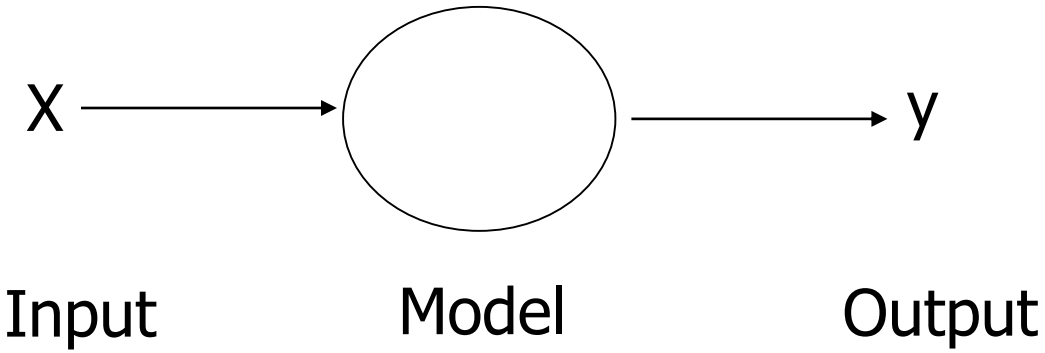
- Classification
  - predict categorical output variable given input variables
- Regression
  - predict a continuous output variable given input variables

ALGORITHMS

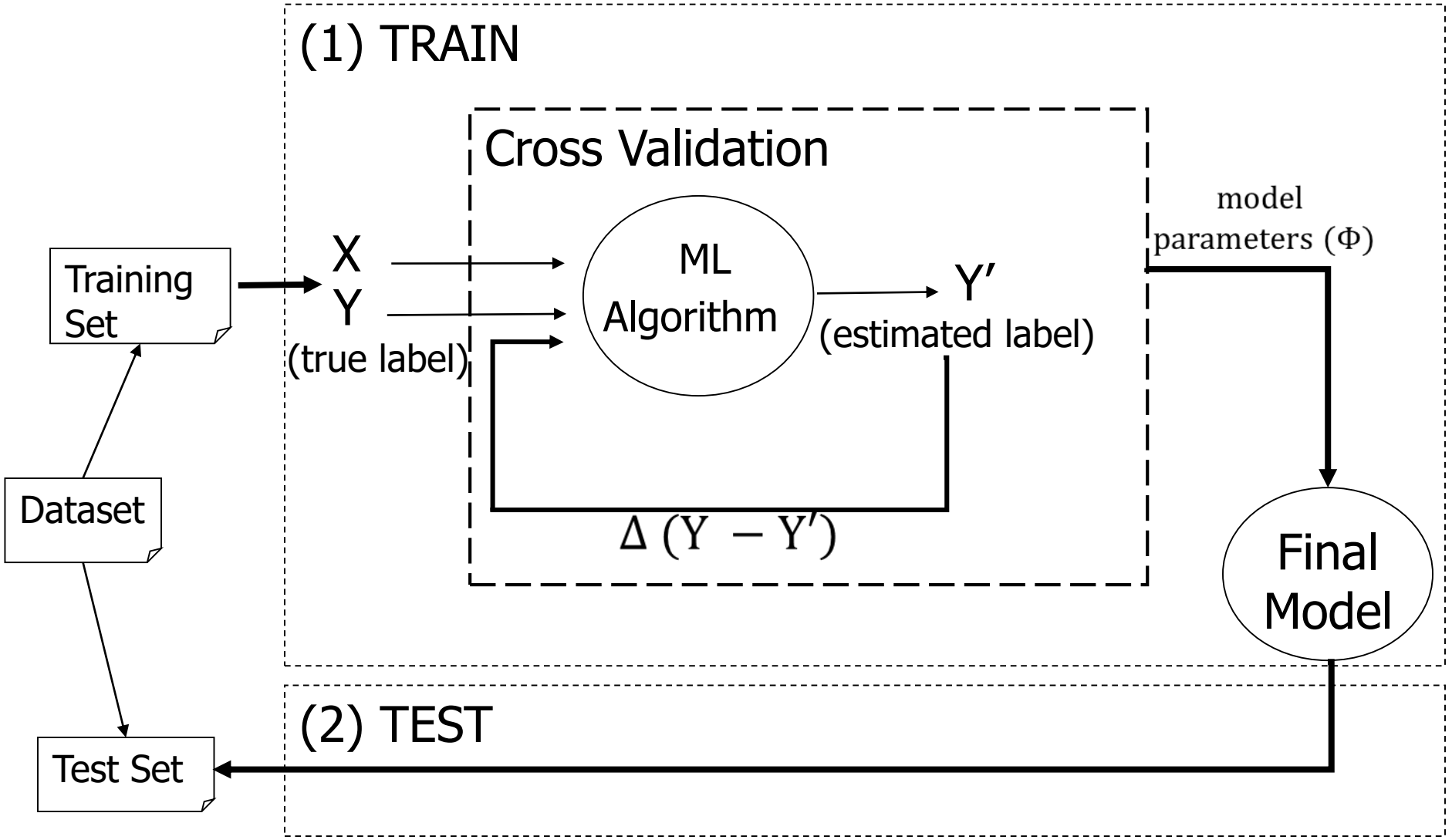
- Logistic Regression
- Neural Networks/ Deep Nets
- Support Vector Machines
- Trees and Random Forests
- Gaussian Processes
- …

The Process of Learning

- Hypothesis, data
- Training or learning
- Testing or generalization

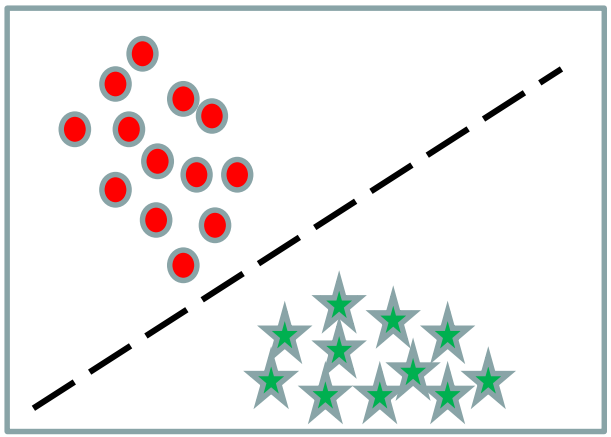


The Process of Learning – Supervised

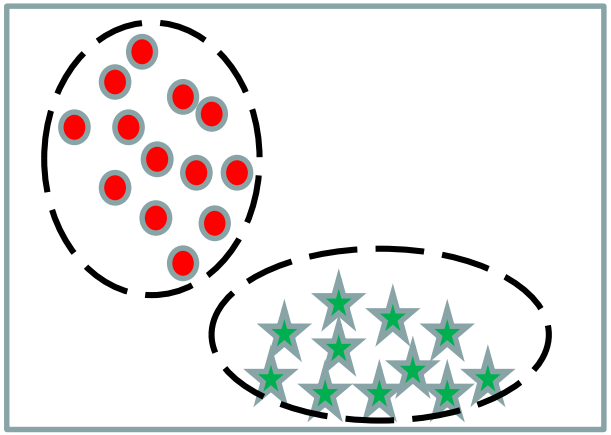


Supervised Learning

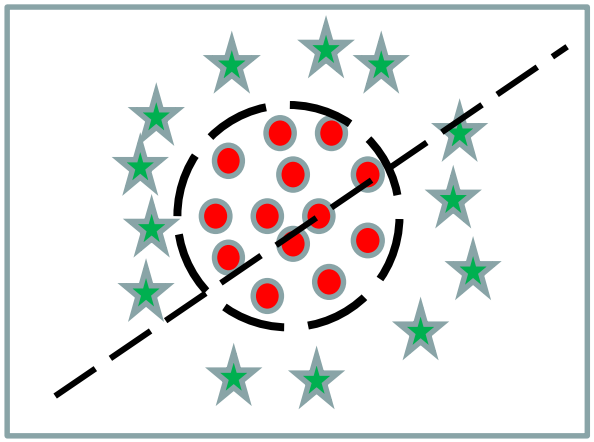
- Training: learning model parameters
- Testing: evaluating model performance, on independent data, from same underlying distribution
- Metrics: AUC, sensitivity/ specificity, PPV/ NPV, etc.
- No free lunch rule: no “best” algorithm, just “optimal” model



Linear classifier

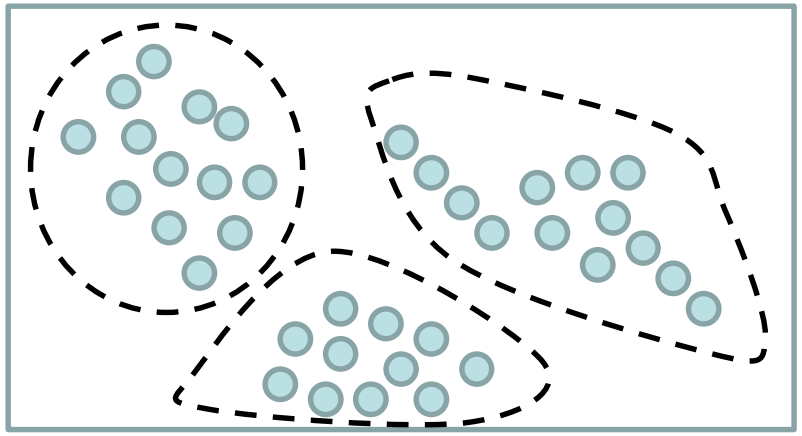


Non-linear classifier



Non-linear classifier

Unsupervised Learning



learn p(x)

no "label"

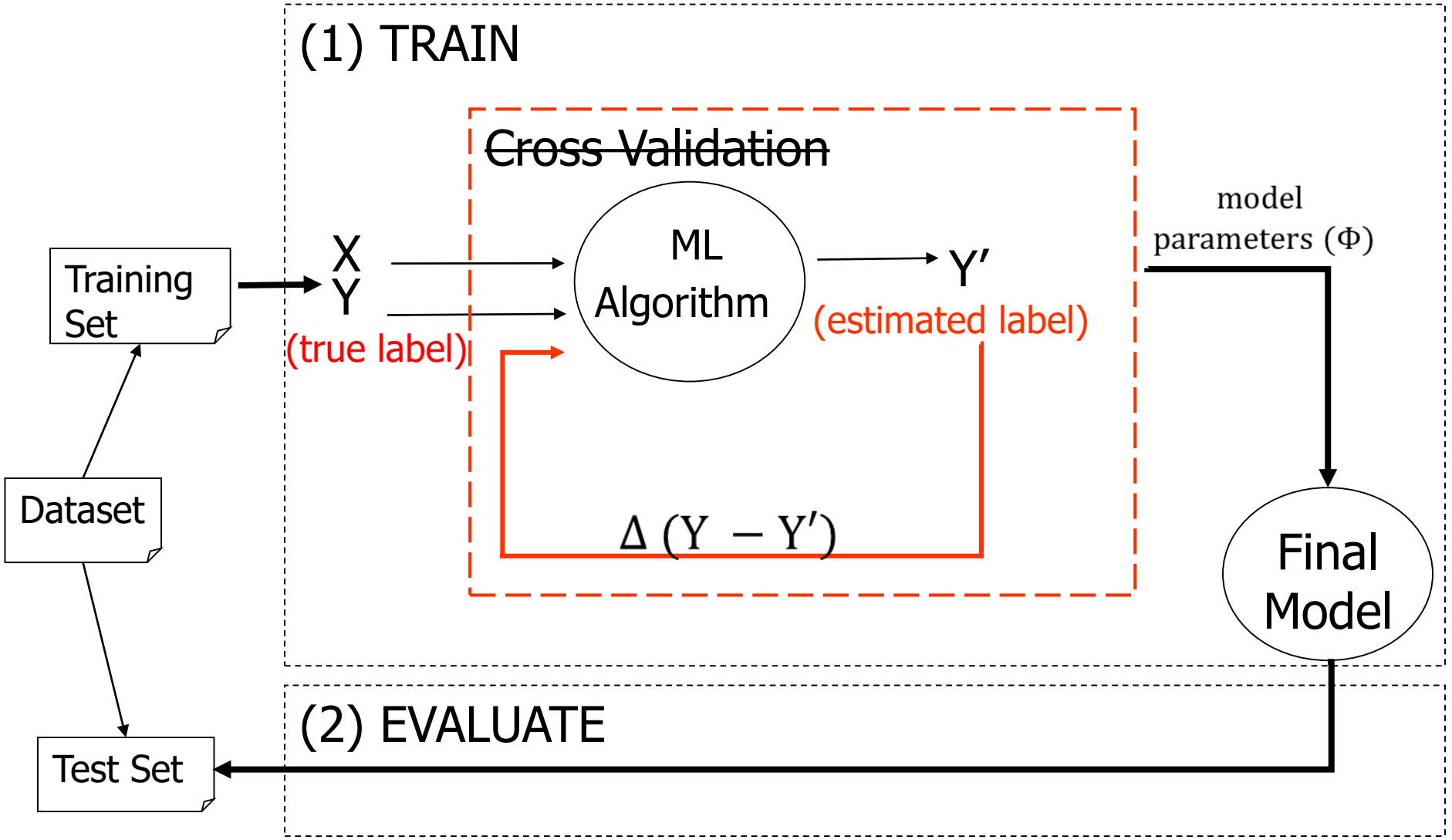
TASKS

- Clustering (group data into clusters using similarity measures)
- Density estimation
- Dimensionality reduction

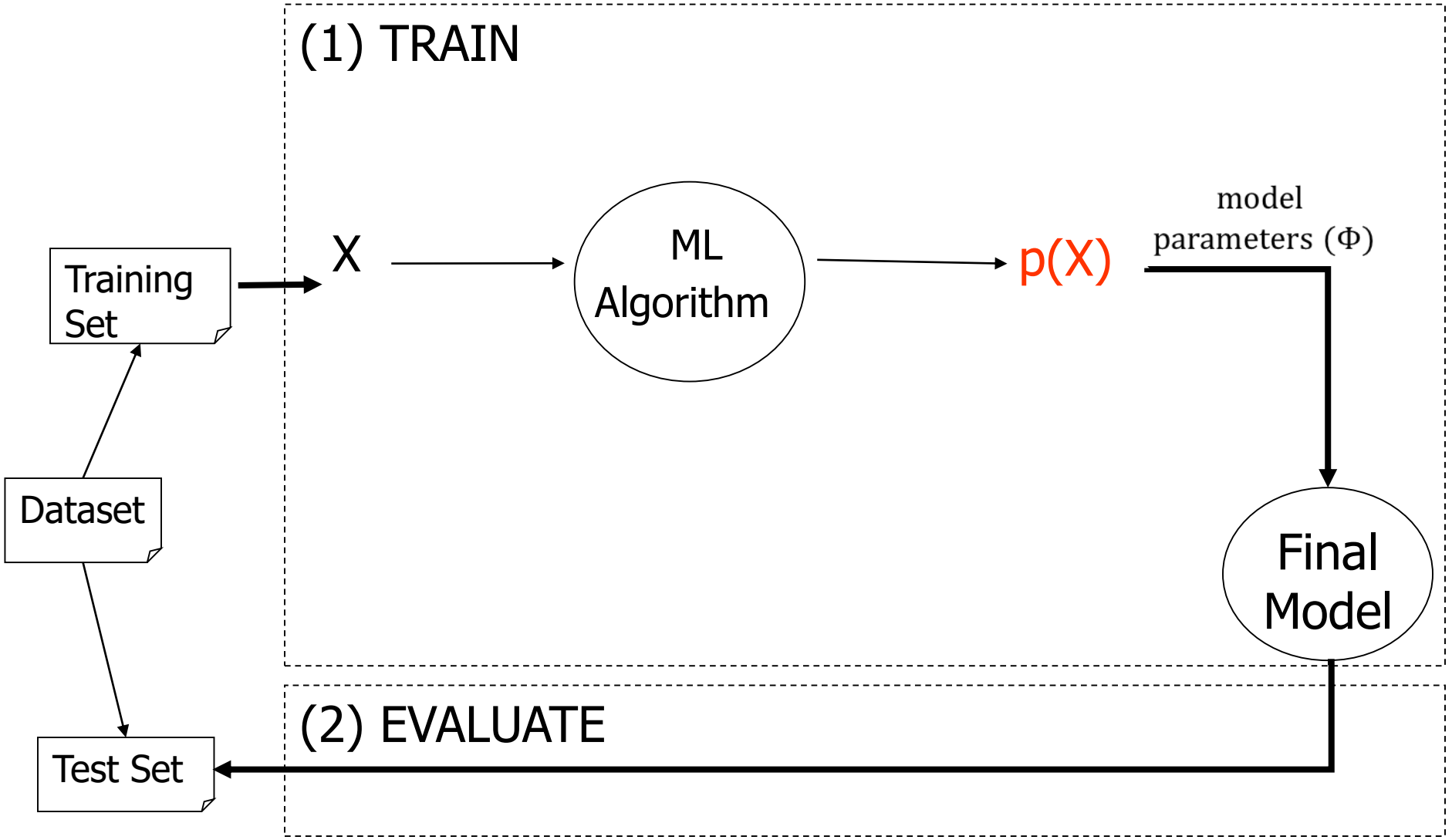
ALGORITHMS

- K-means
- Hierchical clustering
- Gaussian mixture models
- Self-organising maps
- Principal component analysis
- Factor Analysis
- Latent Class Analysis
- …

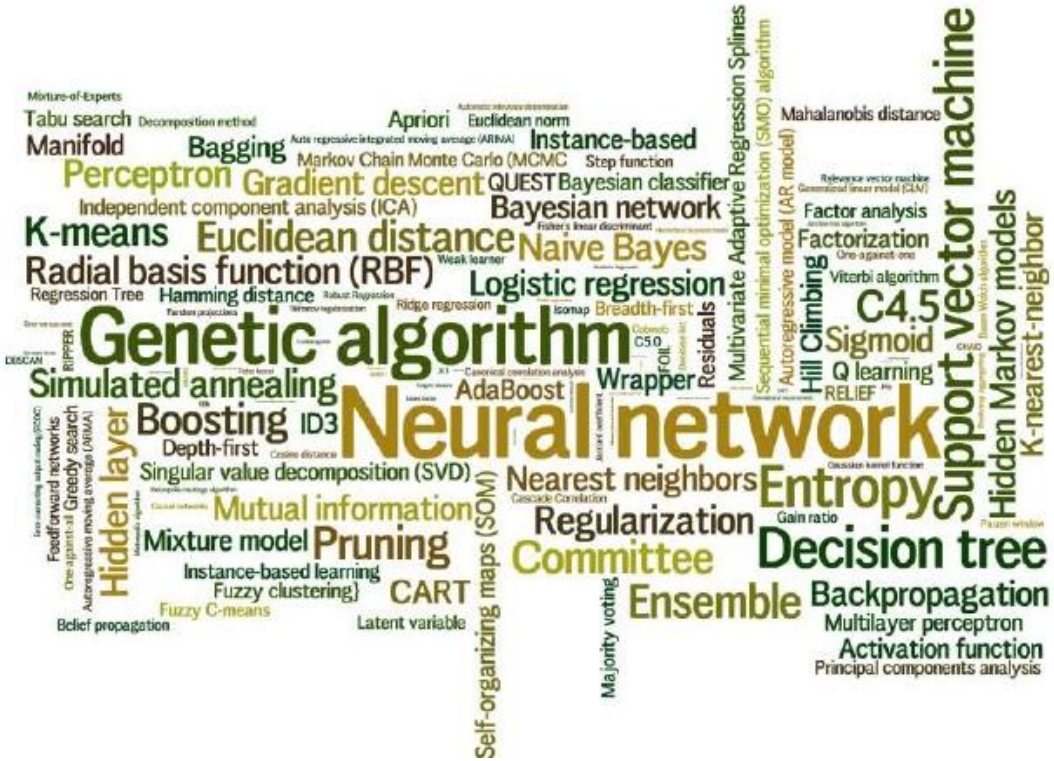
The Process of Learning – Unsupervised



The Process of Learning – Unsupervised



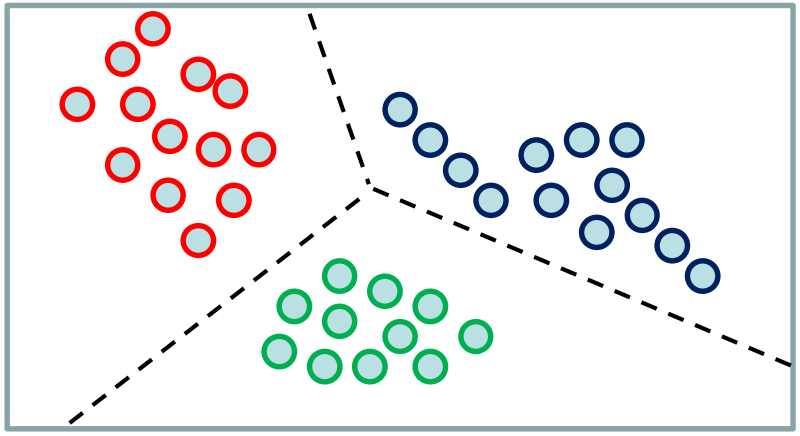
Part 3: Machine Learning Algorithms



Many Algorithms

Brownlee, Machine Learning Mastery (2017)

Supervised Learning



learn p(y|x)

TASK

- **Classification** (predict categorical output variable from input variables)

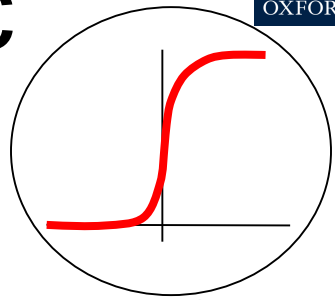
ALGORITHMS

1.Logistic Regression

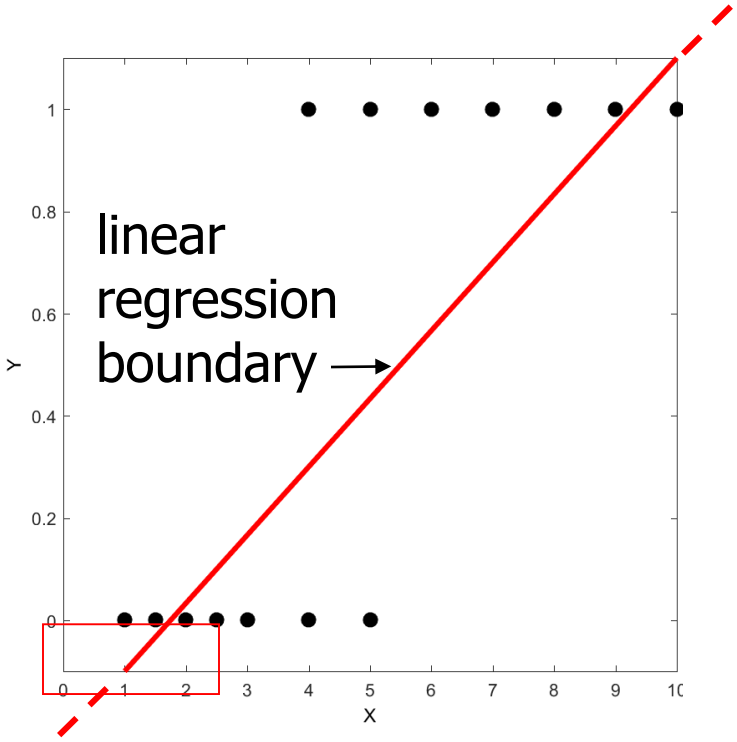
2.Neural Networks/ Deep Nets

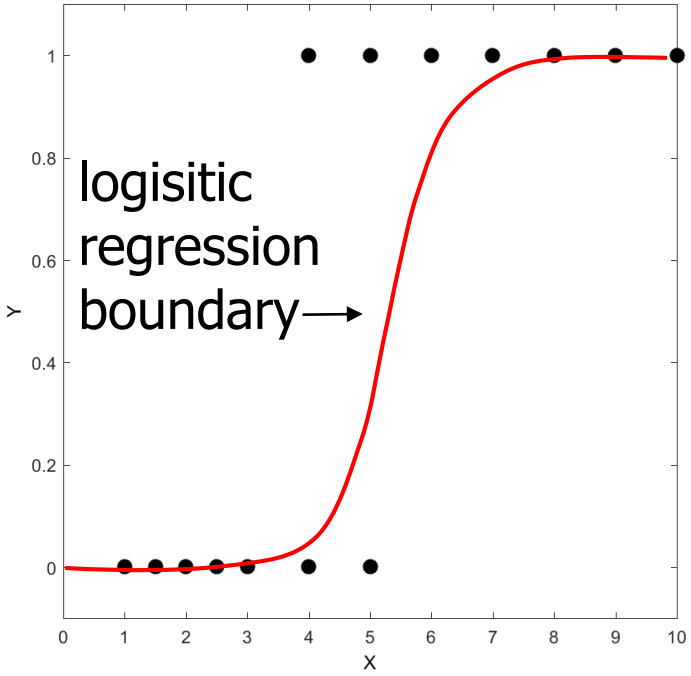
3.Support Vector Machines

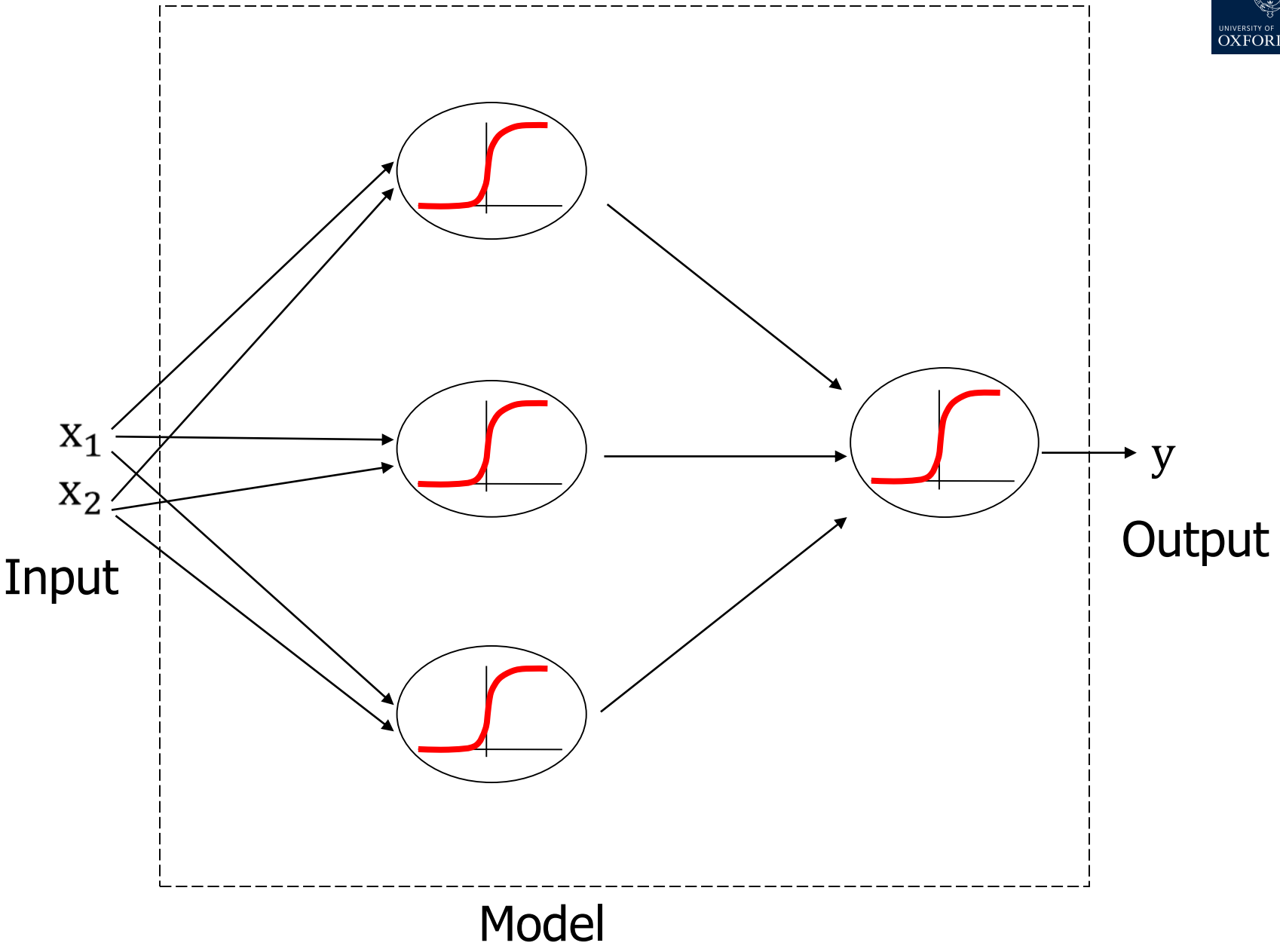
From Linear to Logistic Regression

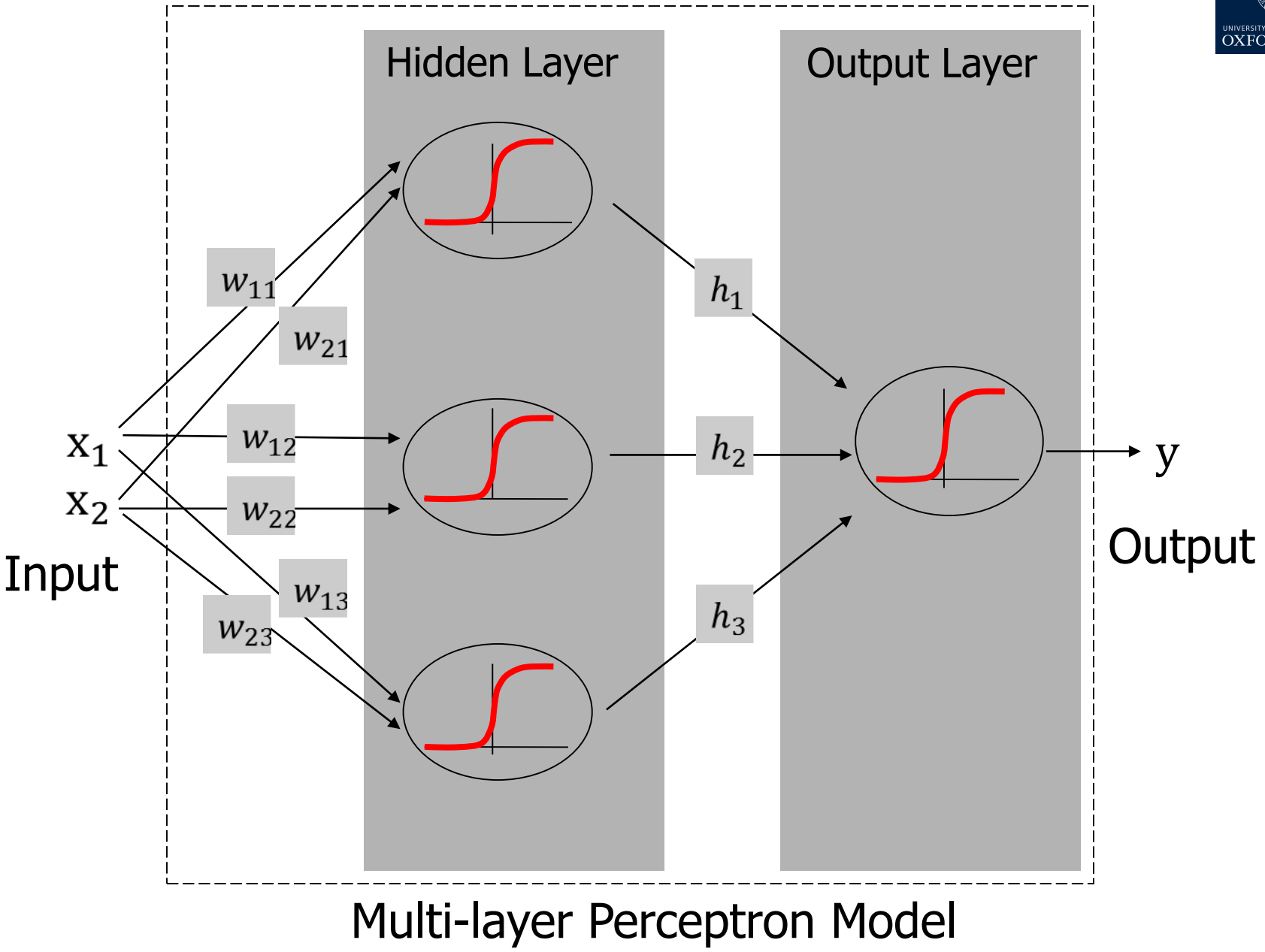


Sigmoid function






x_1
x_2
Input
y
Output
Model

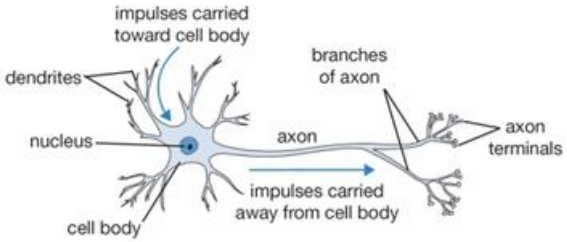


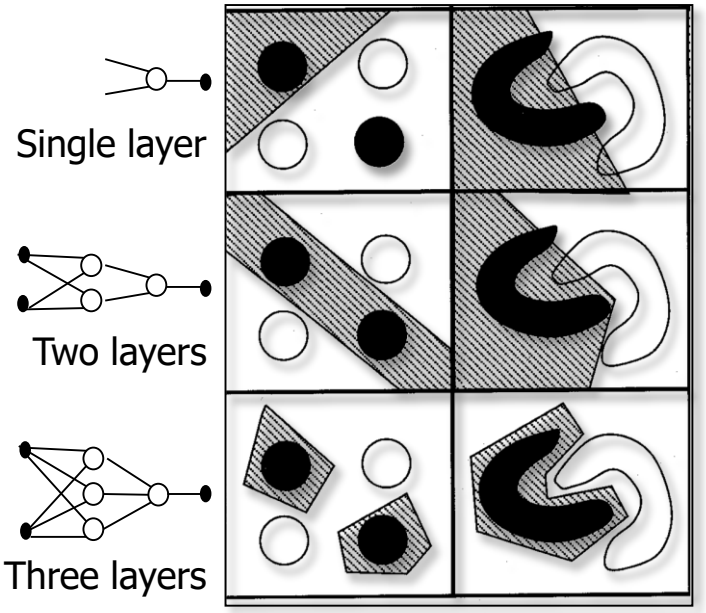
Multi-layer Perceptron Model

Multi-layer Perceptron

Key Concepts

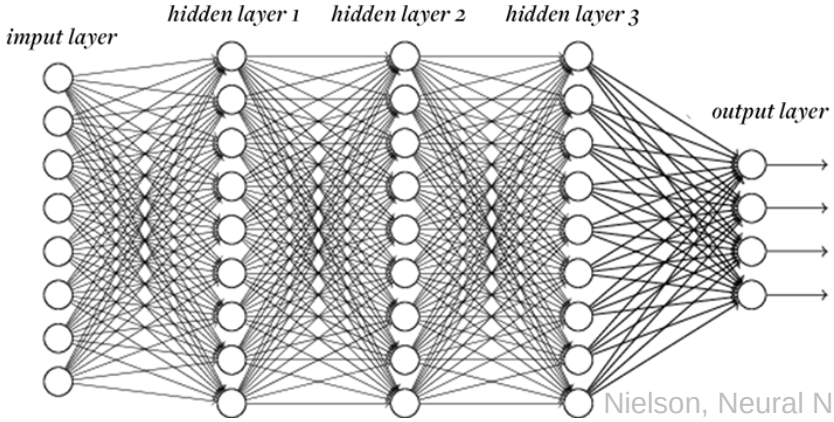
- Artificial neural network
- Modelled after neurons in human brain
- Consists of
  - Input Layer
  - Hidden Layer(s)
  - Output Layer
- Many variations
  - Multi-layer perceptron
  - Radial Basis Function
  - Kohonen Maps, etc.



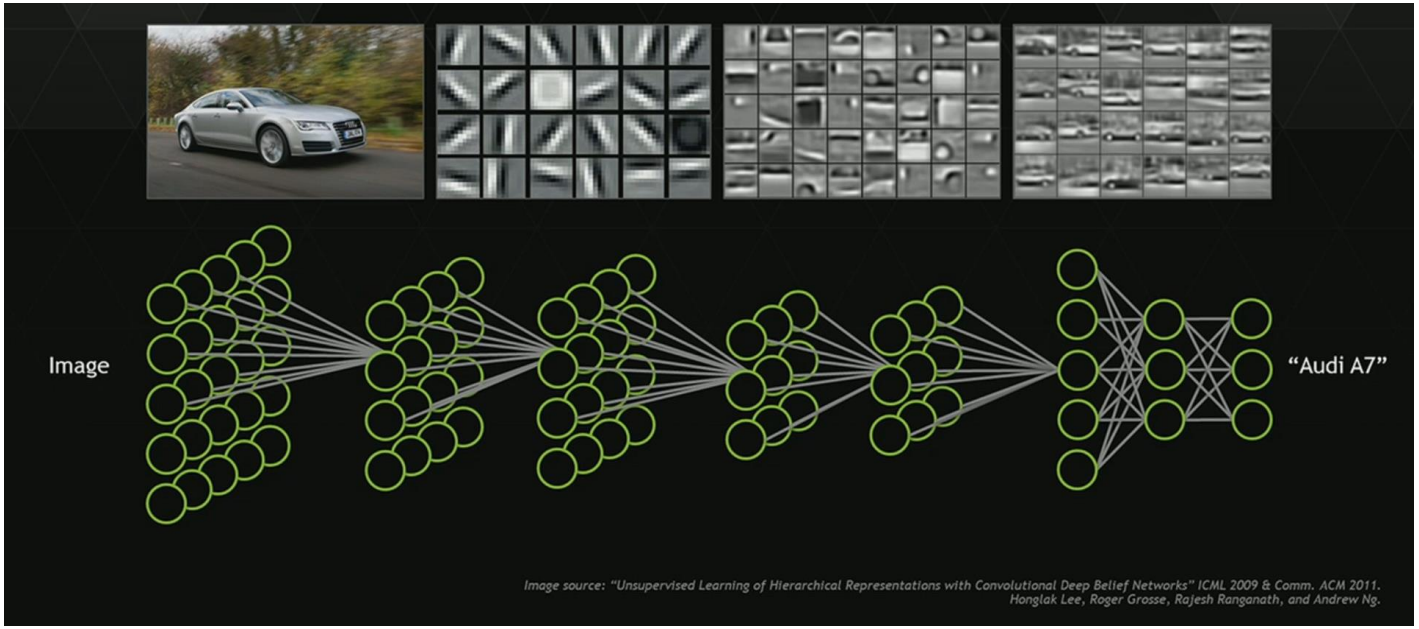


Adapted from Lippmann, IEEE ASSP ()1987

Deep Neural Networks and Deep Learning



Nielson, Neural Networks and Deep Learning (2017)



A Deep Learning Story

YouTube

In 2012, researchers at Google Brain (Lab X), created a network of 16,000 computer processers with >1 billion connections

Then let it browse YouTube for 3 days i.e. trained it by showing 10 million video thumbnails

From >20,000 different items,
it started to recognise two things…

A Deep Learning Story

Human Face

81.7% accuracy

Cat Face

74.8% accuracy

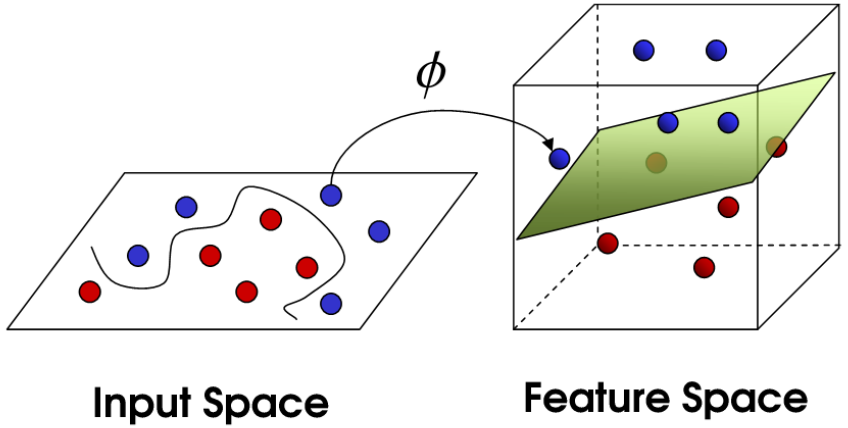
"…our experimental results reveal that it is possible to train a face detector **without having to label images** as containing a face or not"

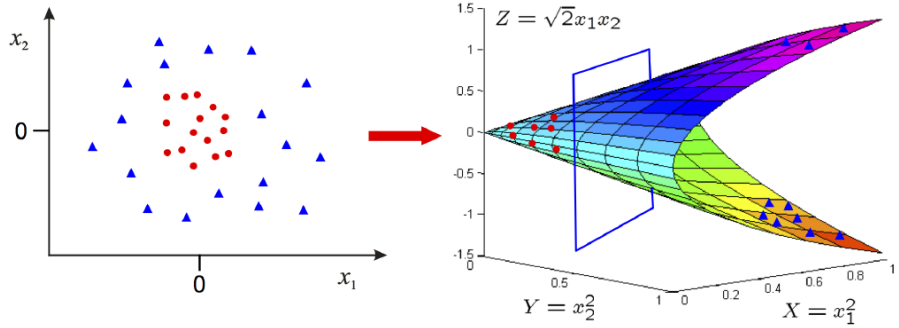
Quoc Le @ LabX, IEEE ICASSP (2013)

Support Vector Machine

Key Concepts

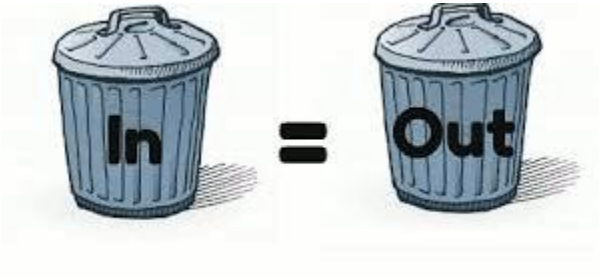
- Transform data to feature space
- Kernel function
- Non-linearly distributed data --> linearly separable
- Hyperplane that maximizes margin
- Support vectors lie on hyperplane, control flexibility

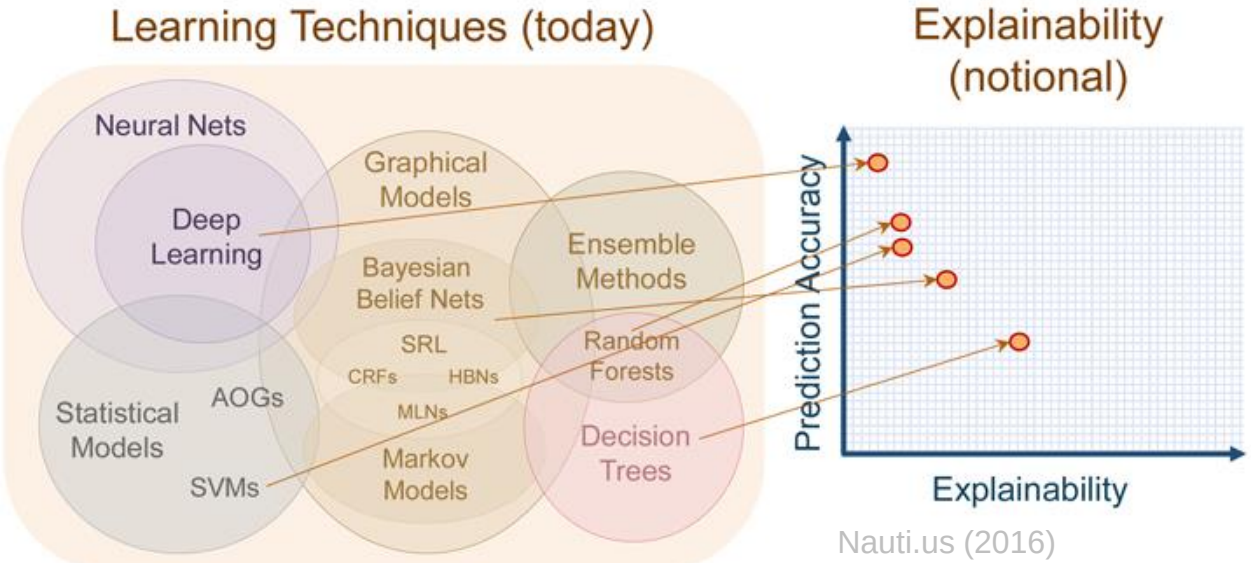




Considerations

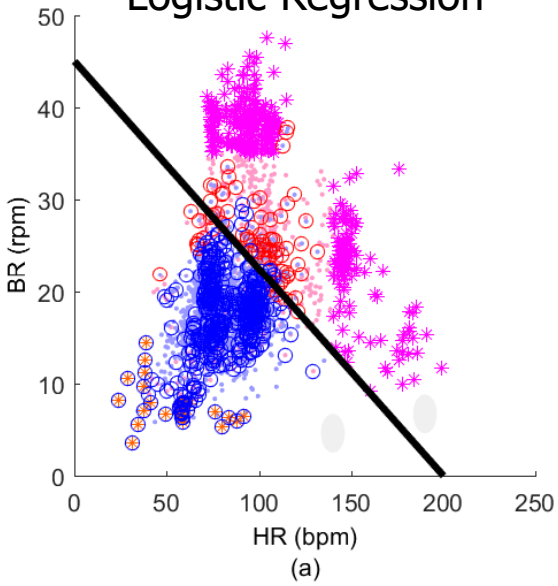
- Your model is as good as your (cleaned) data
- Requires lots of training data
- Explainability/ interpretability vs predictive performance



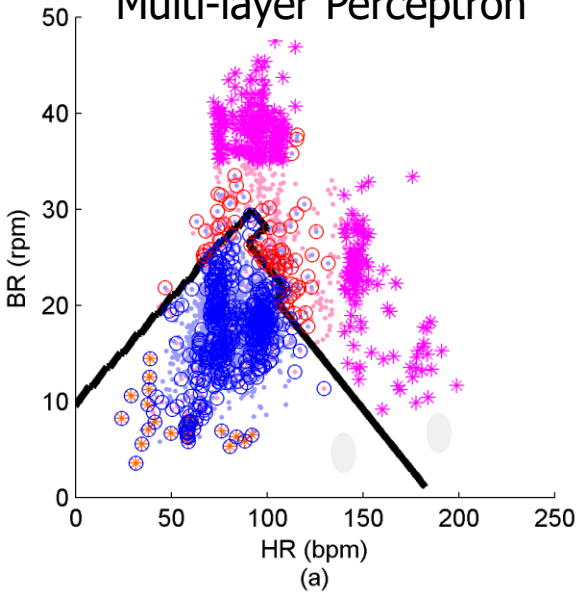


What might a classifier look like?

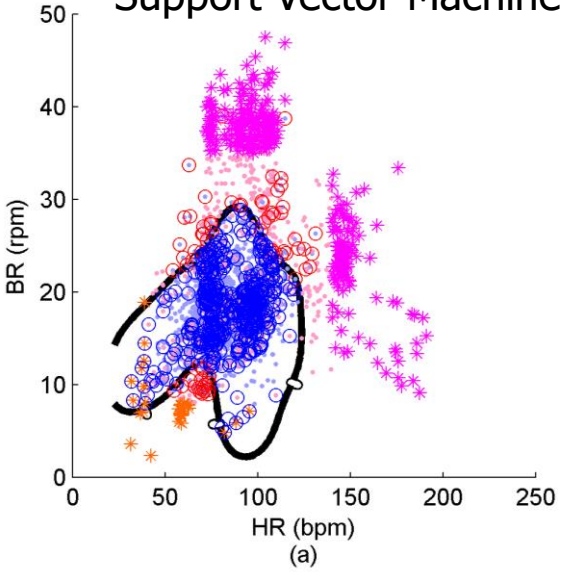
Logistic Regression



Multi-layer Perceptron

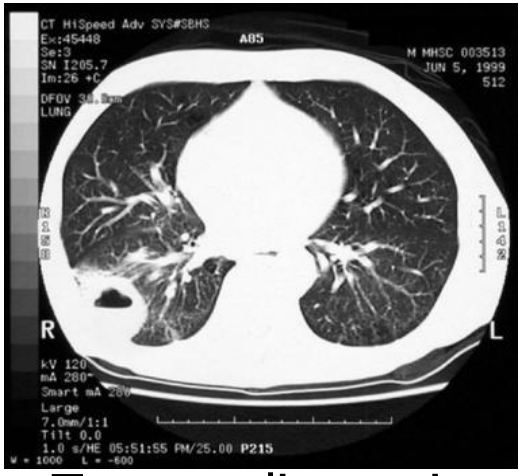


Support Vector Machine

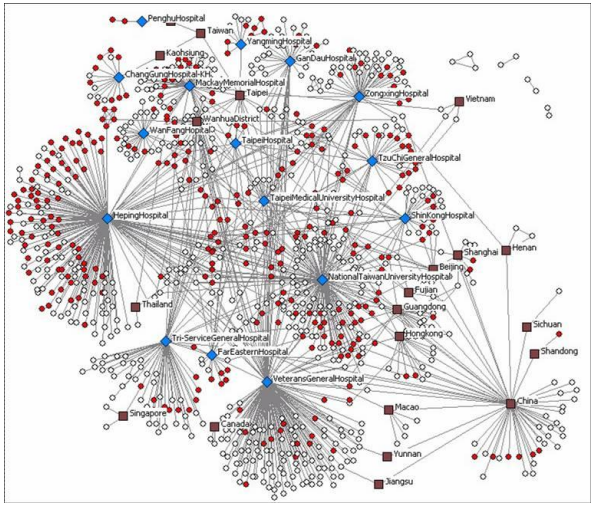


Applications

Eye disease detection

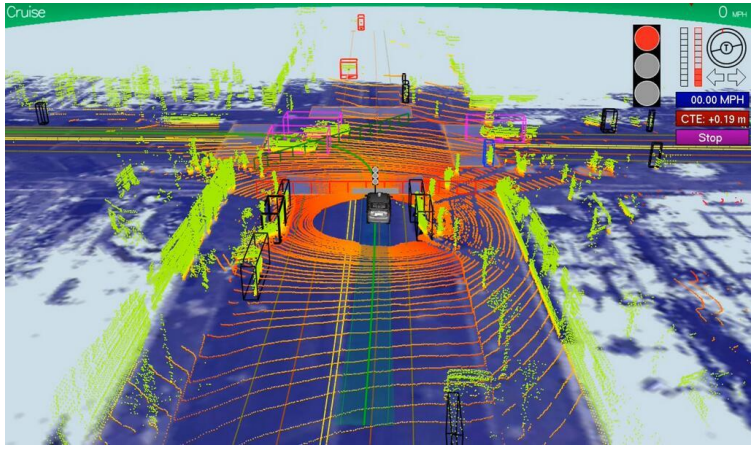


Tumour diagnosis



Disease mapping and tracking

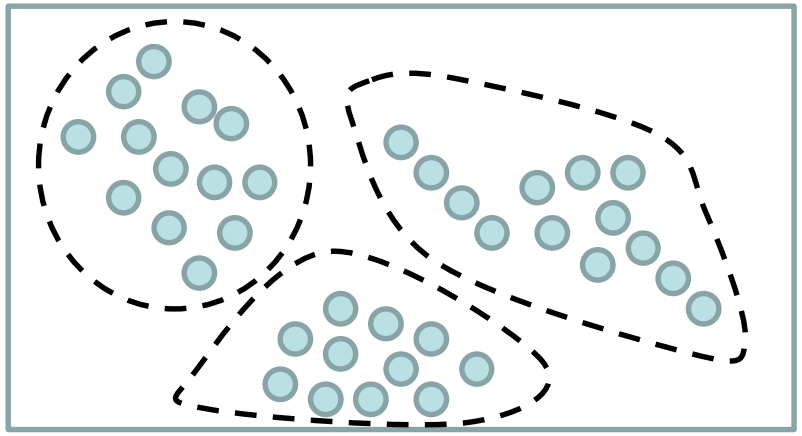
Bioinformatics



Autonomous vehicle navigation

AlphaGo

Unsupervised Learning



learn p(x)

TASK

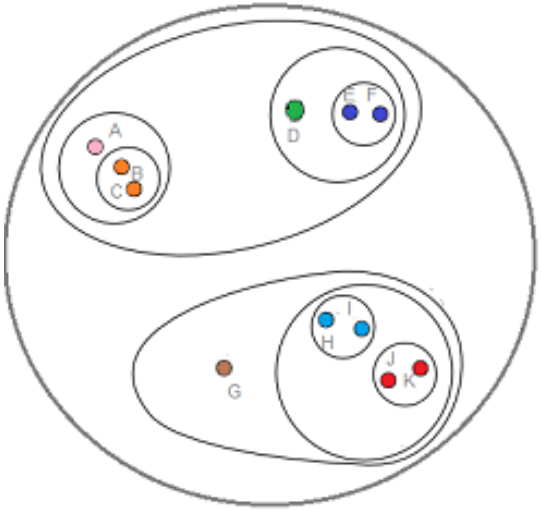
- **Clustering** (group data into clusters using similarity measure)

ALGORITHMS

1. *k*-means
2. Hierchical clustering

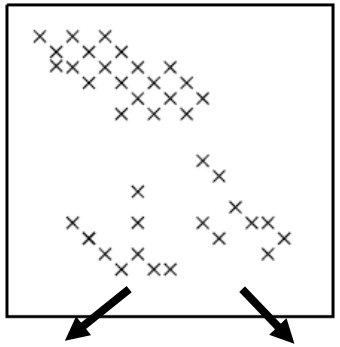
Hierarchical Clustering

- Multi-level hierarchy, tree-like
- Top-down (divisive) or bottom-up (agglomerative)
- How it works
  1. Each point is a cluster
  2. Merge “closest” clusters
  3. Repeat (2) until
     - no change in clusters, or
     - you decide to prune the tree
- Watch-out: will cluster outliers



(Tibshrani, 2003)

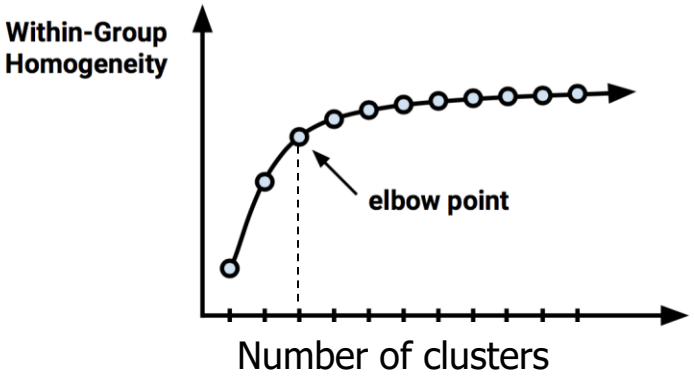
k-Means Clustering

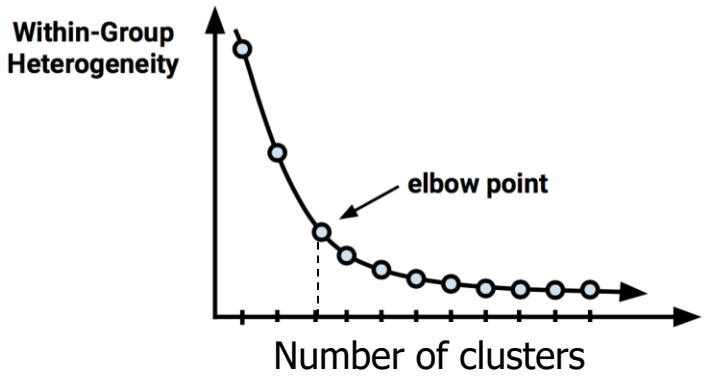


- How it works
  1. Assign *k* cluster centroids (randomly)
  2. Assign each point to “closest” centroid
  3. Re-calculate cluster centroid
  4. Re-assign points
  5. Repeat (3), (4) until no further changes
- Watch-out: Local minima trap
- Solution: repeat *n*>>1 times

(MacKay, 2003)

Cluster Evaluation

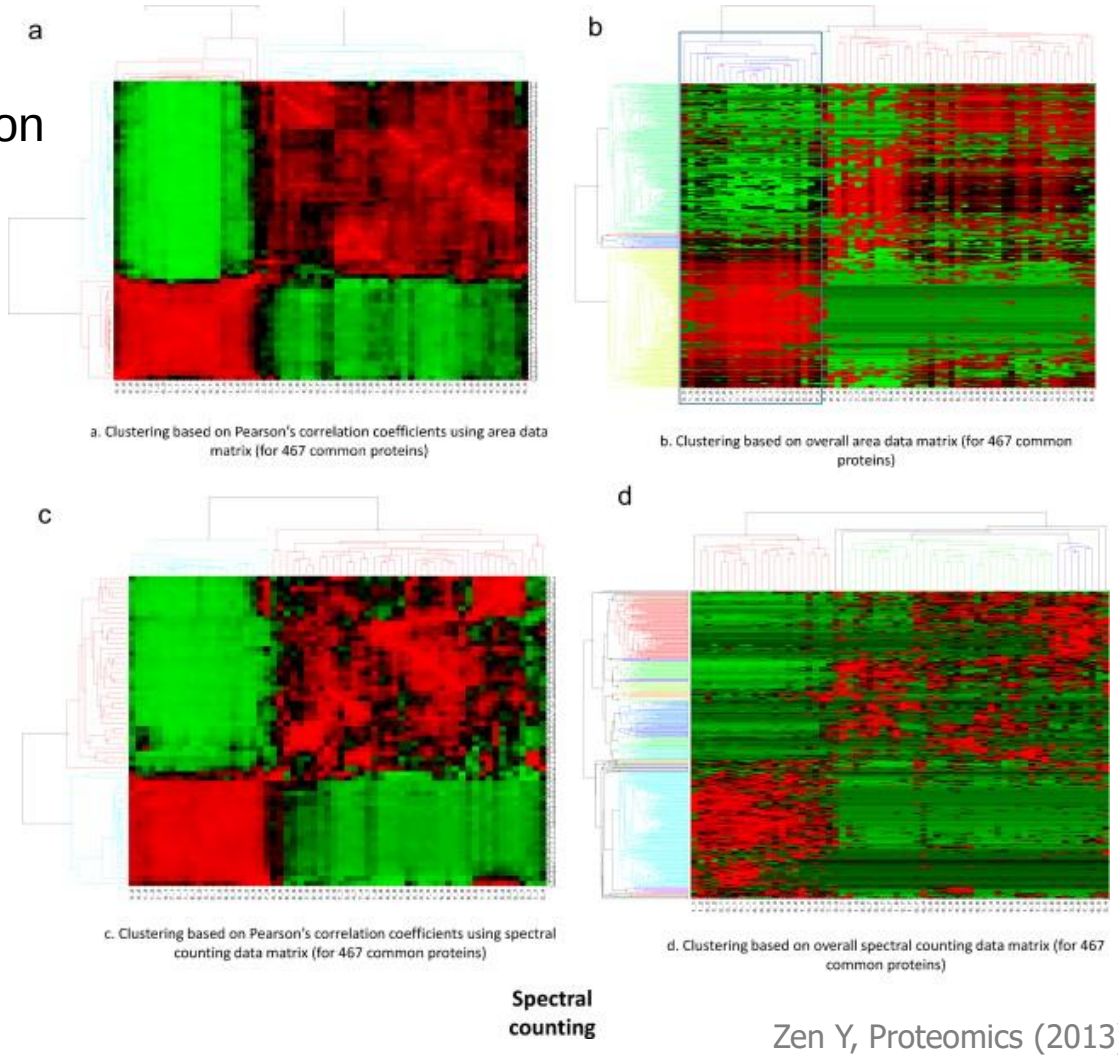




- Find a partition which
  (a) maximises **homogeneity**/ minimises **heterogeneity** within a cluster
  (b) maximises separation from other clusters.
- The "elbow" occurs at the most dramatic decrease in error measurement.
- Choose from different error measures, depending on the kind of data you are clustering.
- No single, definite way to decide optimal number of clusters.

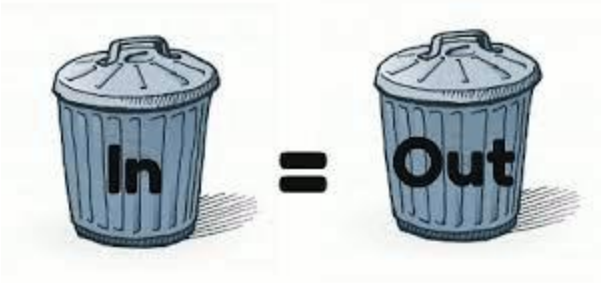
Cluster Analysis Applications

- Bioinformatics (gene expression clustering)
- Infection mapping in hospitals
- Investment banking
- Mapping voting behaviours & identifying stakeholders
- Text analysis



Zen Y, Proteomics (2013)

Considerations



- Needs good feature selection
- Outliers and small groups can throw off results (try density modelling, e.g. GMM)
- Try different similarity measures, different linkages
- Best guess at optimal clusters

All said and done

- The question
- The known (predictors/ features)
- The unknown (class/ response/ cluster)
- Measure of success

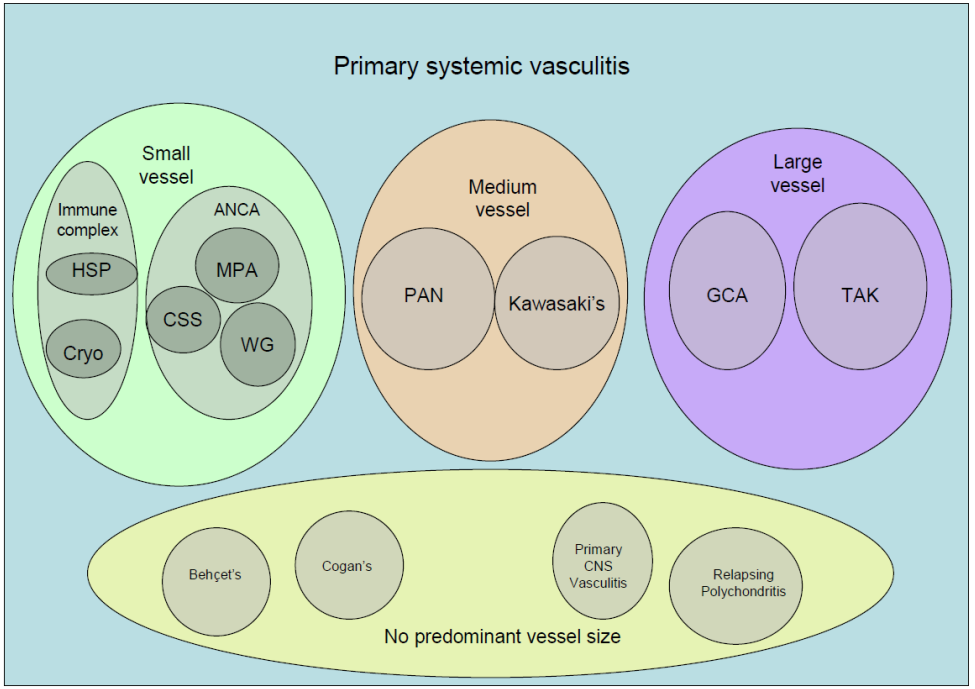
Part 4: Real-world examples

1. Clinical diagnostic tool (Classification)
2. Healthcare data mining (Clustering)

Example 1: Vasculitis Diagnosis

- 1500 patients
- >1000 symptoms
- Which ones predict GCA?

TASK: classify GCA v. non-GCA



Robson et al. American College of Rheumatology EULAR (2016)

Data Preparation

Predictors (X) Outcome (Y)

Data

Training Set (70%)

ID	Age	Gender	Eye Redness	Hearing Loss	cANCA+	…	Diagnosis = GCA
1	62	M	No	No	Yes		No
2	56	M	No	Yes	No		Yes
3	71	F	Yes	Yes	No		No
…							

Use this training set to learn the model

Test Set (30%)

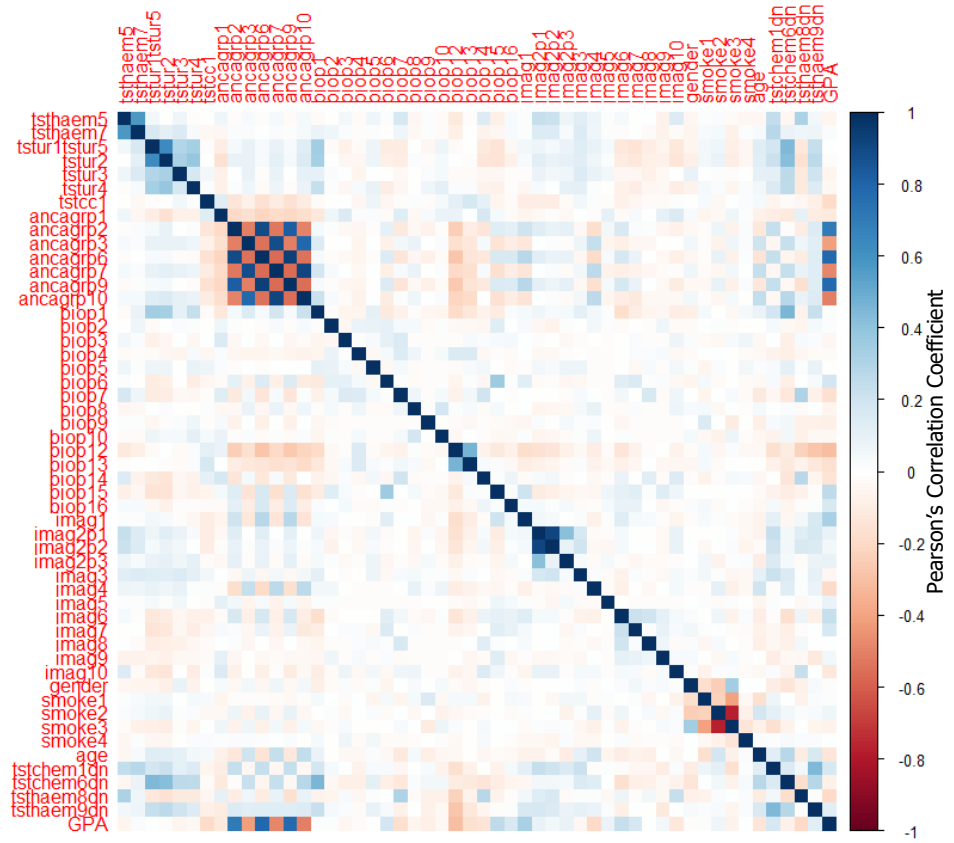
ID	Age	Gender	Eye Redness	Hearing Loss	cANCA+	…	Diagnosis = GCA
100	56	F	Yes	No	Yes		Yes
102	83	M	No	Yes	Yes		No
103	72	F	Yes	Yes	No		No
…							

Use this test set to measure model performance

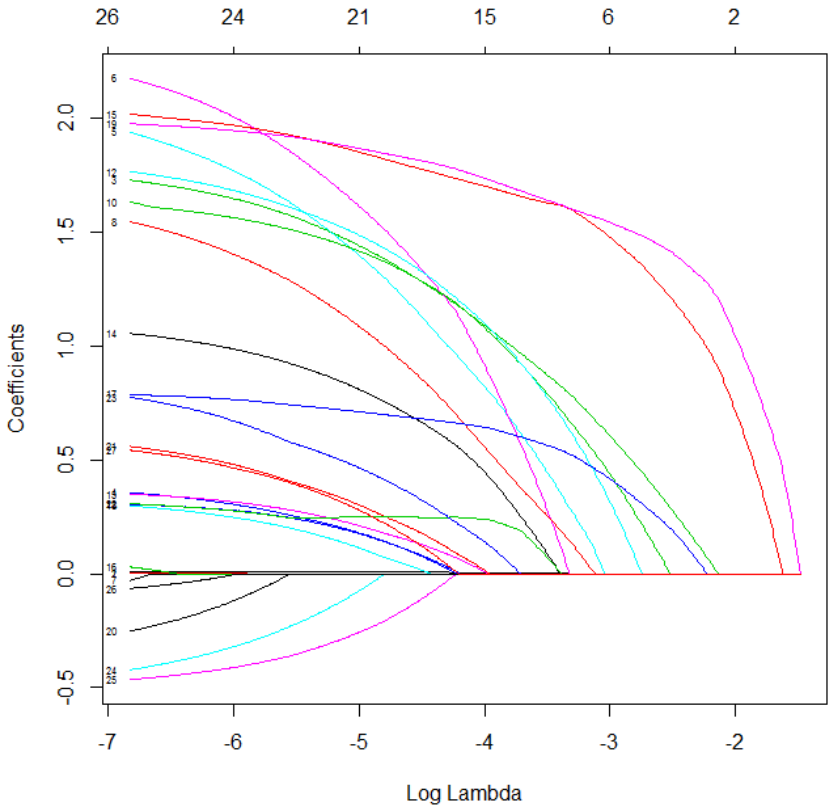
ID	Age	Gender	Eye Redness	Hearing Loss	cANCA+	…	Diagnosis = GCA
New patient	56	F	Yes	No	Yes		?

New patient walks into the clinic. Apply model to predict diagnosis.

Feature Selection

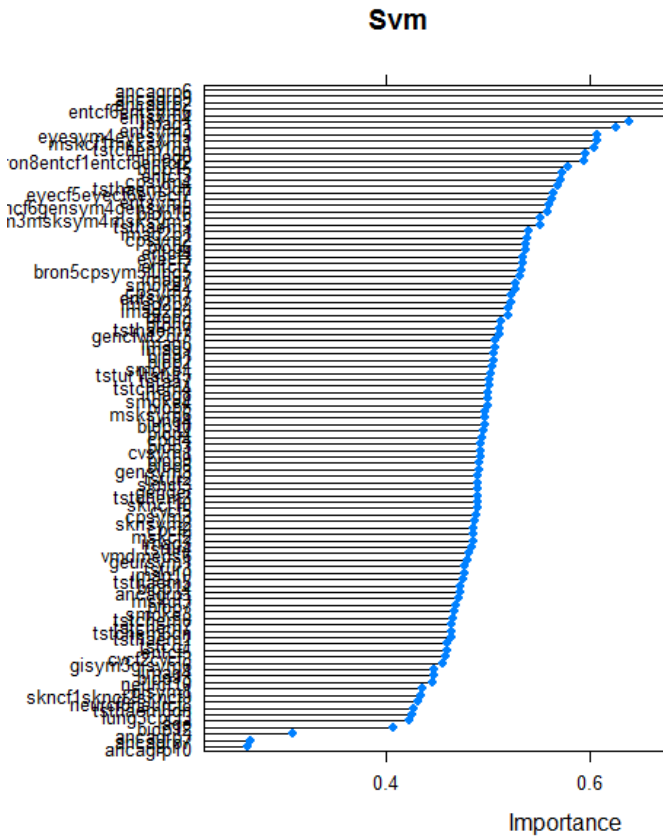


Look for highly correlated predictors

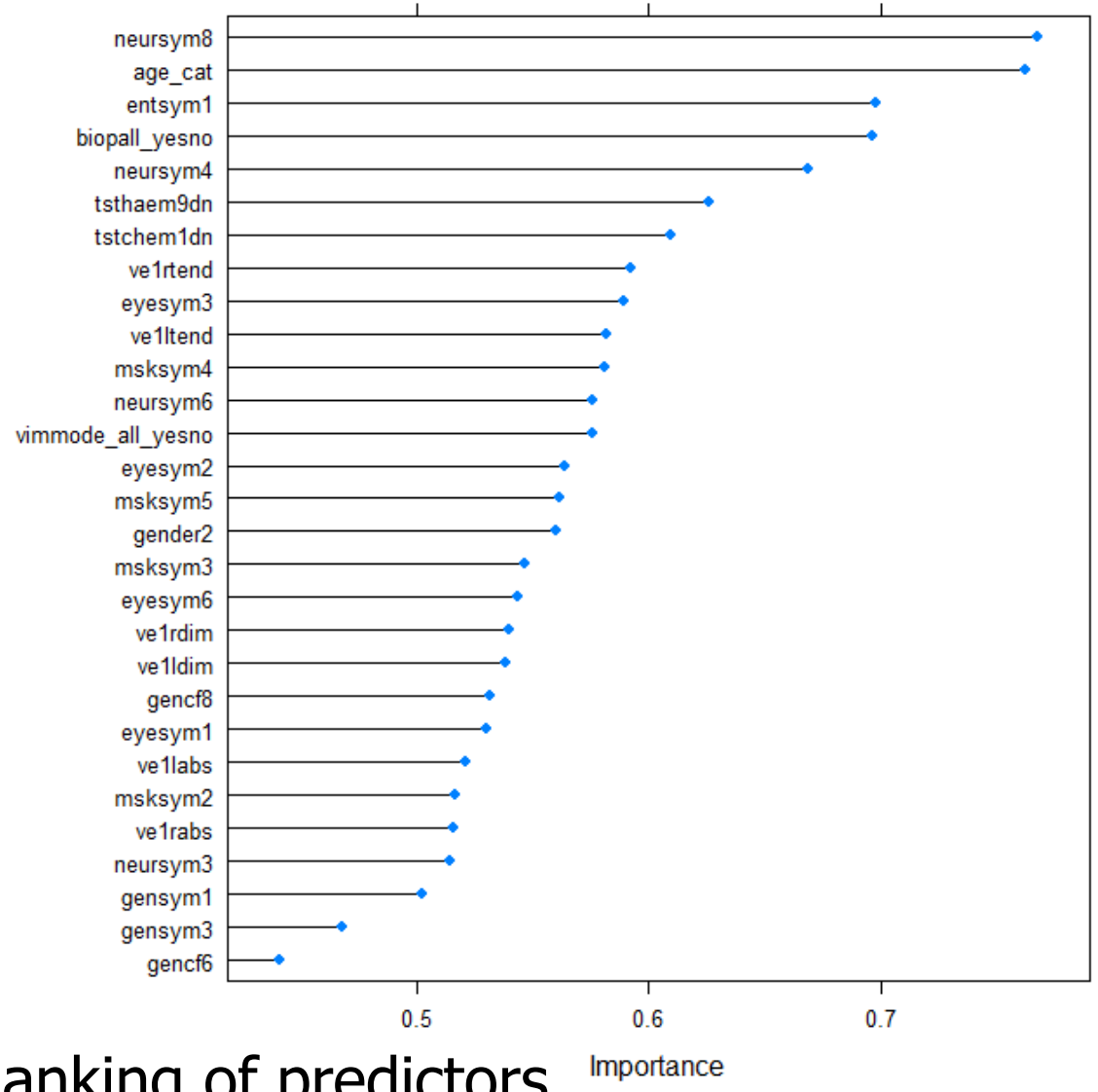


Lasso Regression – drop non-predictors

Model Development

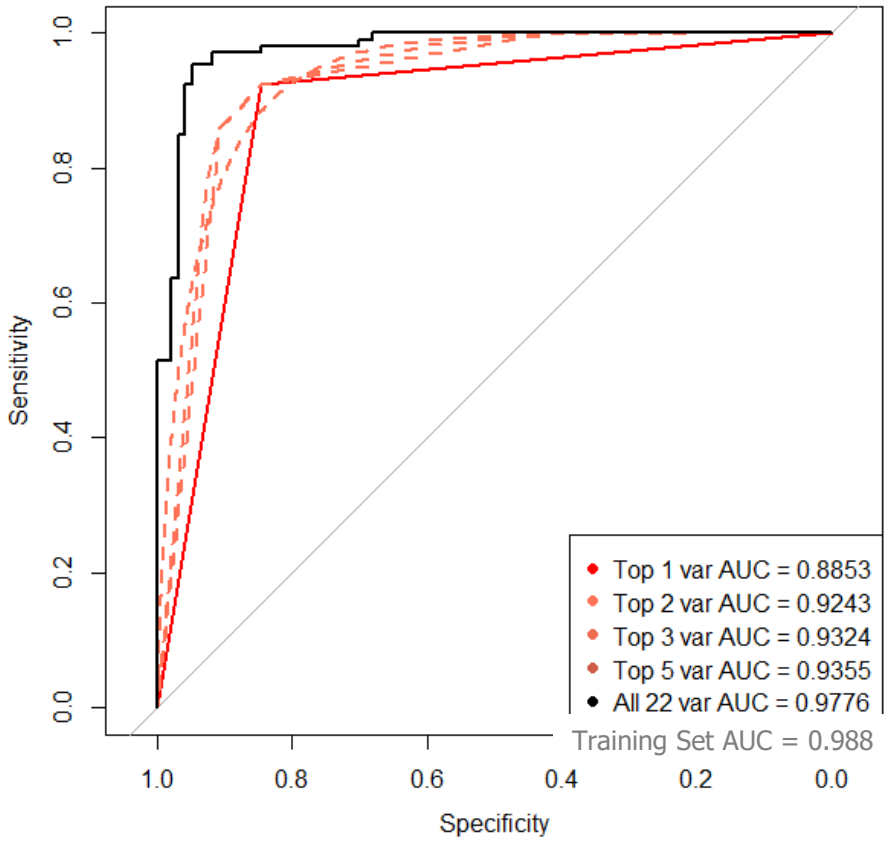


Train the model

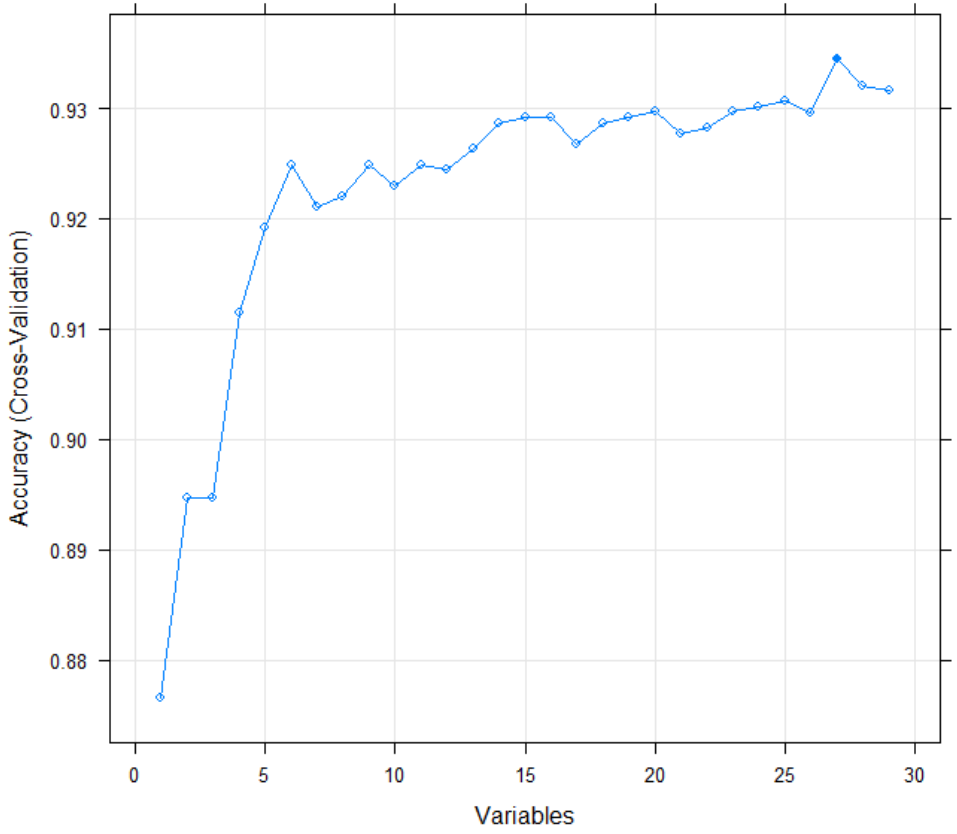


Ranking of predictors

Model Performance



Test Set AUC



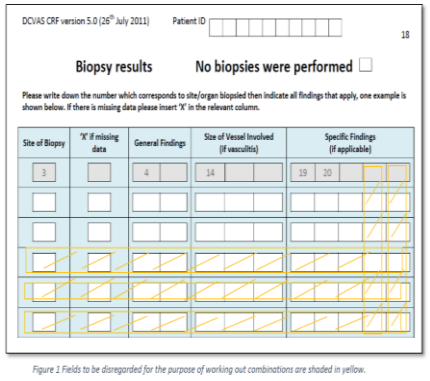
Accuracy
(True Classification Rate)

The Complete Model

A Clinical Tool for Classification of GCA Vasculitis

Risk factor at baseline	Coeficient	Categories	Reference value	Risk score
cANCA or PR3	4.904	No	0	0
		Yes	1	5
Bloody nasal discharge OR Nasal ulcers/mucosal abnorms/crusting OR Sino nasal congestion or blockage	2.770	No	0	0
		Yes	1	3
Maximum eosinophil count ($x10^9/L$)		<1	0	0
	-2.564	>=1	1	-3
Nasal polyps	-3.696	No	0	0
		Yes	1	-4
Imaging of the chest/lungs with nodules OR mass/tumour OR cavitation	1.752	No	0	0
		Yes	1	2
Hearing loss or reduction	1.320	No	0	0
		Yes	1	1
Granuloma OR Extravascular granulomatous inflammation	2.559	No	0	0
		Yes	1	3
Red eye(s) OR Painful eye(s)	1.486	No	0	0
		Yes	1	1
Endobronchial involvement OR Inflamed ear or nose cartilage OR Hoarse voice / stridor OR Saddle nose deformity	1.705	No	0	0
		Yes	1	2

The Data Analysis Process


DCVAS CRF version 5.0 (26th July 2011) Patient ID

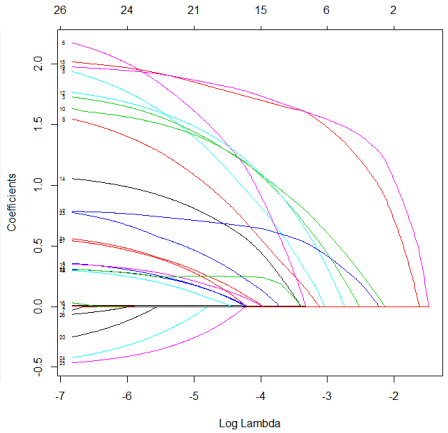
18

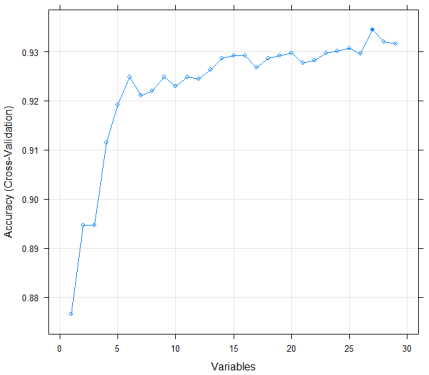
Biopsy results **No biopsies were performed** ☐

Please write down the number which corresponds to site/organ biopsied then indicate all findings that apply, one example is shown below. If there is missing data please insert 'X' in the relevant column.

Site of Biopsy	'X' if missing data	General Findings	Size of Vessel Involved (if vasculitis)	Specific Findings (if applicable)
3		4	14	19 20

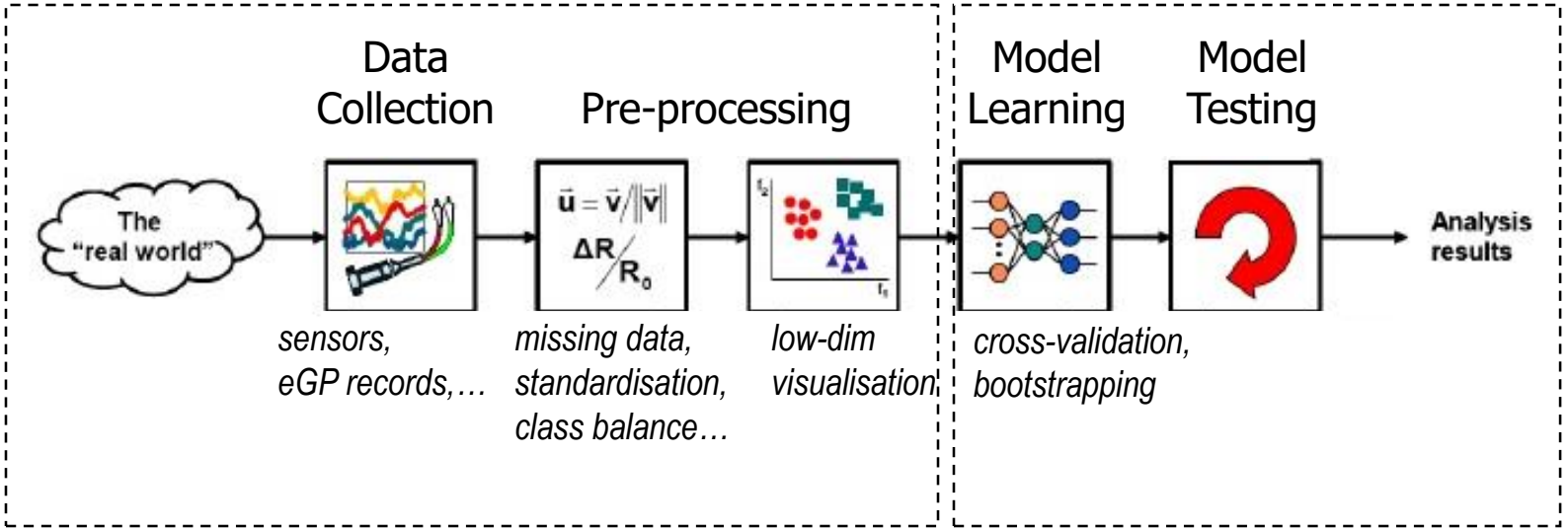
Figure 1 Fields to be disregarded for the purpose of working out combinations are shaded in yellow.





Risk factor at baseline	Coefficient	Categories	Reference value	Risk score
cANCA or PR3	4.904	No	0	0
		Yes	1	5
Bloody nasal discharge OR Nasal ulcers/mucosal abnorms/crusting OR Sino nasal congestion or blockage	2.770	No	0	0
		Yes	1	3
Maximum eosinophil count ($x10^9/L$)		<1	0	0
	-2.564	>=1	1	-3
Nasal polyps	-3.696	No	0	0
		Yes	1	-4
Imaging of the chest/lungs with nodules OR mass/tumour OR cavitation	1.752	No	0	0
		Yes	1	2
Hearing loss or reduction	1.320	No	0	0
		Yes	1	1
Granuloma OR Extravascular granulomatous inflammation	2.559	No	0	0
		Yes	1	3
Red eye(s) OR Painful eye(s)	1.486	No	0	0
		Yes	1	1
Endobronchial involvement OR Inflamed ear or nose cartilage OR Hoarse voice / stridor OR Saddle nose deformity	1.705	No	0	0
		Yes	1	2

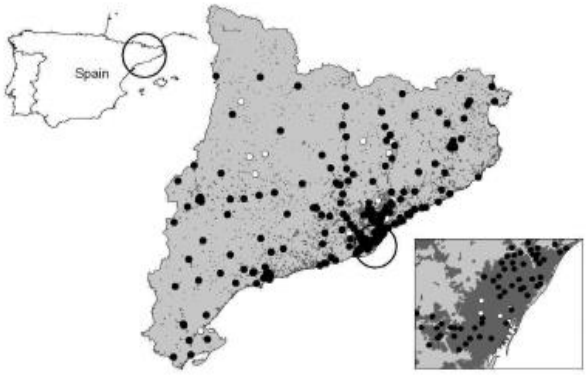
1 – Data Collection and Cleaning 2 - Feature Selection 3 – Model Development 4 – Final Model



Data Preparation Model Development

Example 2: Healthcare data mining

- >6 million people in Catalonia, Spain
- 150k anti-fracture medication users
- GP e-records
- Fracture risk factors
- What can we learn?
  - User “types”/ groups
  - Fracture risk
  - Should they be on medication
  - Costs



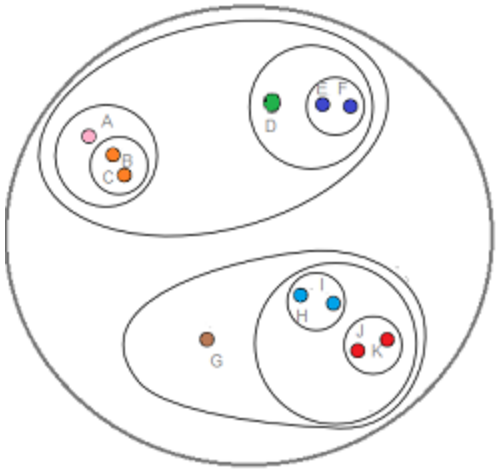
Khalid S. et al., World Congress on Osteoporosis, osteoarthritis, and musc., 2017

Cluster Analysis

Features (X)

ID	Age	Gender	Smoking	Drinking	Charlson>2	Steroid Use	Sedative Use	Previous Fracture
1	56	F	Yes	No	Yes	Yes	Yes	No
2	83	M	No	Yes	Yes	No	No	Yes
3	72	F	Yes	Yes	No	No	No	Yes
…	73	M	No	No	Yes	Yes	No	Yes
150K	62	F	Yes	Yes	No	No	No	Yes

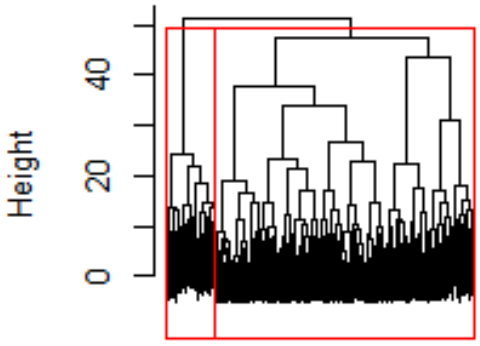
No label/ outcome (Y)



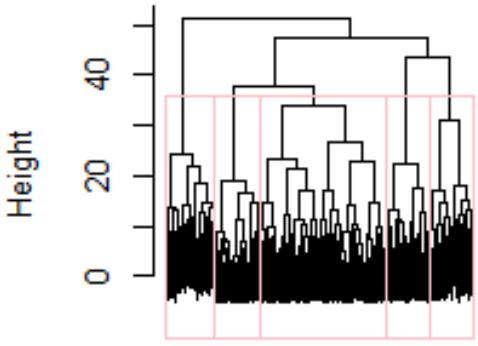
Hierarchical Clustering

Cluster Visualisation

Dendrogram

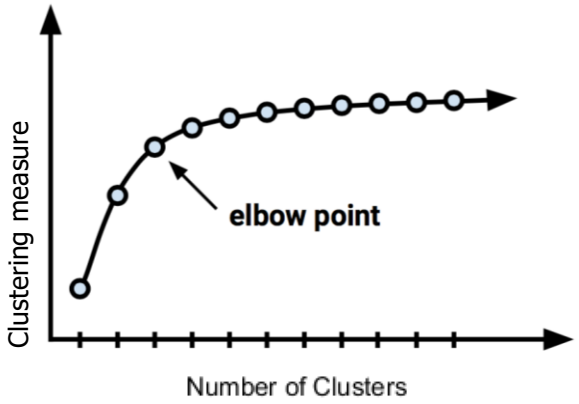


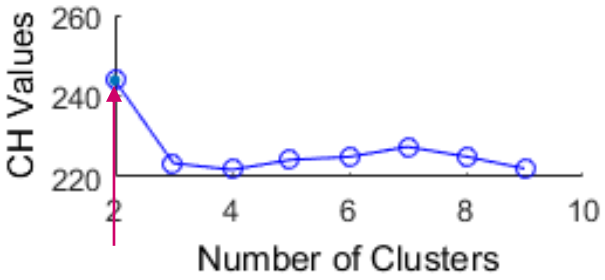
2 clusters



4 clusters

Cluster Evaluation



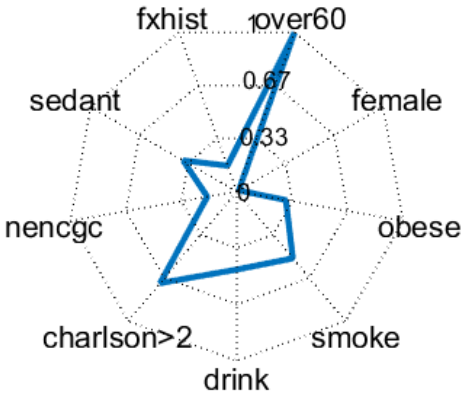


Optimal # clusters is at highest point

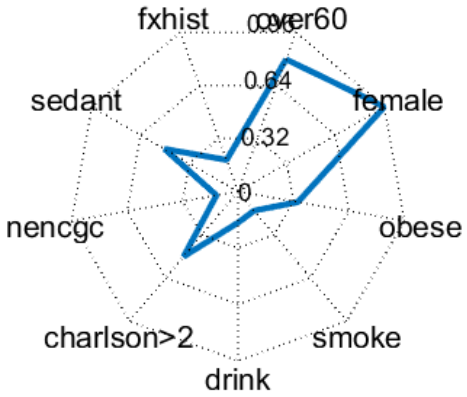
Cluster Visualisation

Number of clusters =2

cluster1, Npts =6661

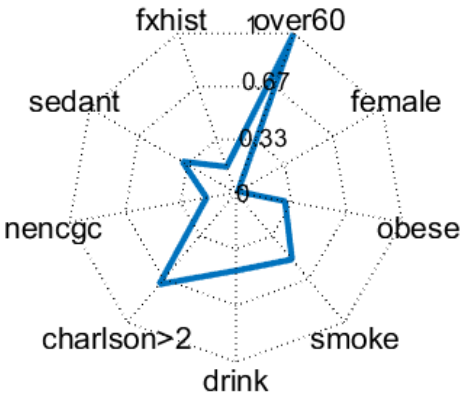


cluster2, Npts =31335

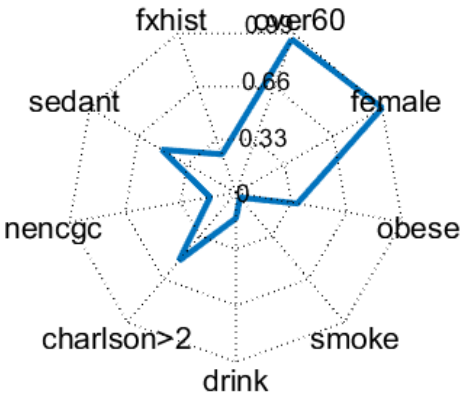


Number of clusters =3

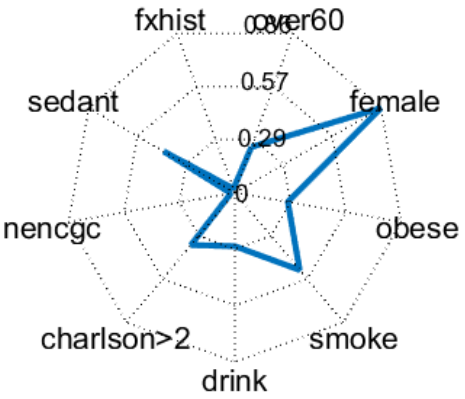
cluster1, Npts =6661



cluster2, Npts =24434

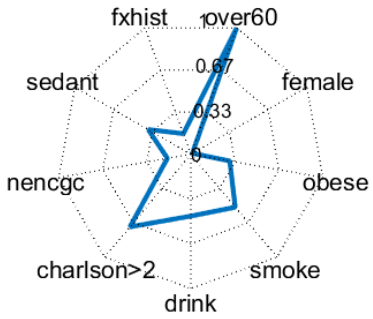


cluster3, Npts =6901

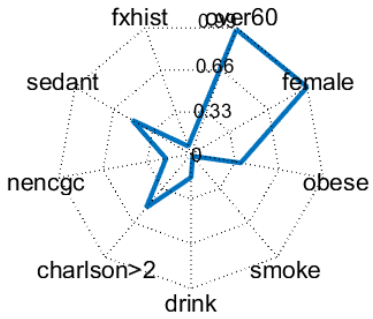


Number of clusters = 4

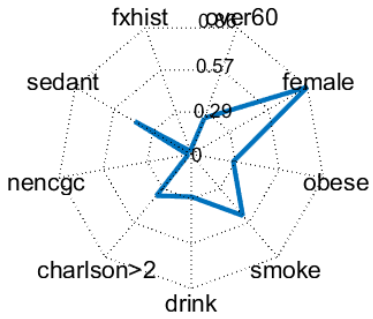
cluster1, Npts =6661



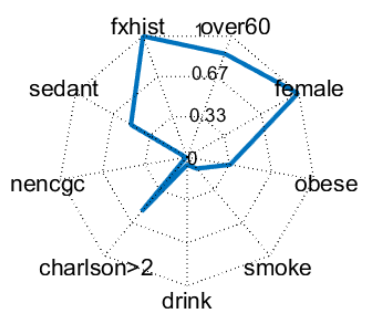
cluster2, Npts =19845



cluster3, Npts =6901

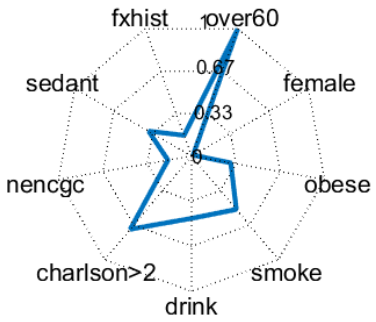


cluster4, Npts =4589



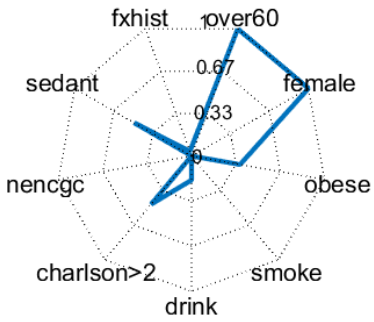
Number of clusters = 5

cluster1, Npts =6661



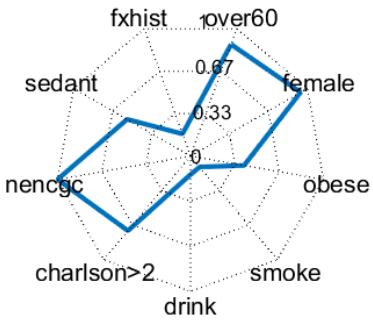
1. Elderly men, multi-morbidity, high prevalence of smoking and drinking

cluster2, Npts =16124



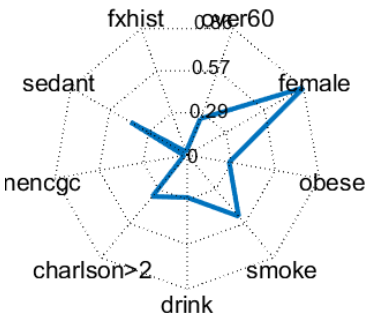
2. Elderly women, high co-morbidity

cluster3, Npts =3721



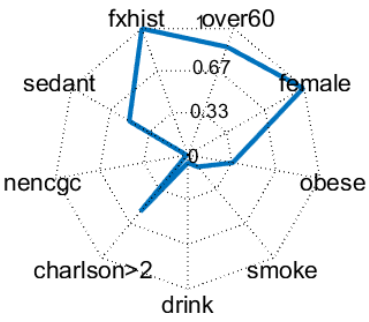
3. Elderly women, systemic steroid users

cluster4, Npts =6901



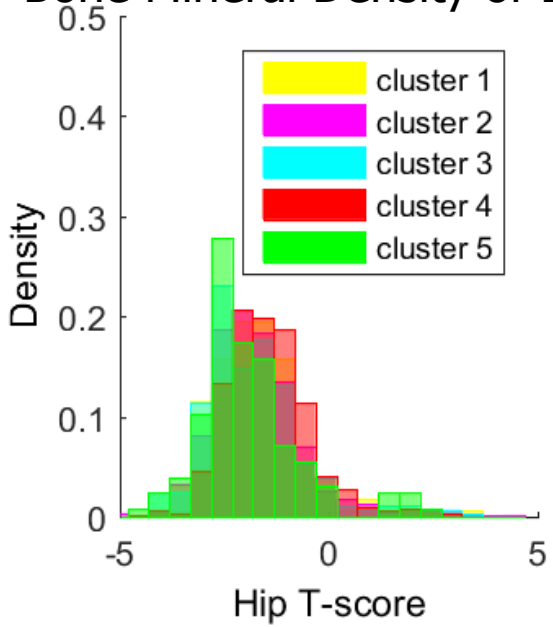
4. Younger women (early post-menopausal), low-medium co-morbidity

cluster5, Npts =4589

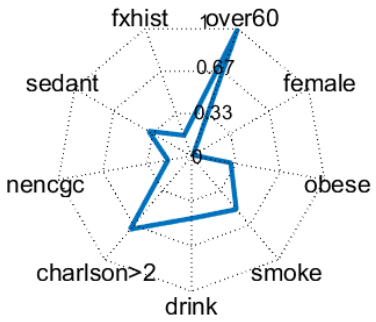


5. Secondary prevention women (previous fracture history)

Bone Mineral Density of Each Cluster

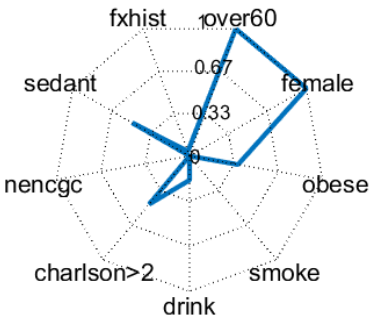


cluster1, Npts =6661



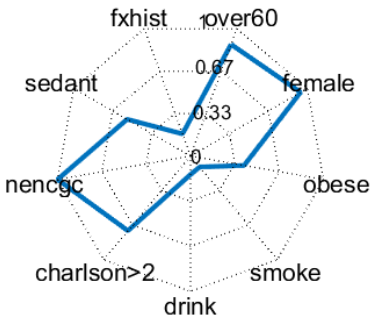
1. Elderly men, multi-morbidity, high prevalence of smoking and drinking

cluster2, Npts =16124



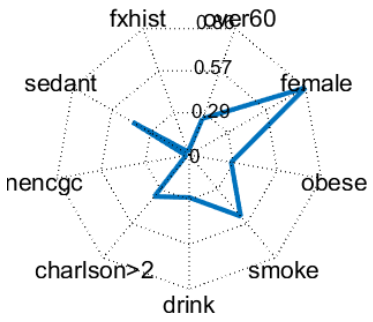
2. Elderly women, high co-morbidity

cluster3, Npts =3721



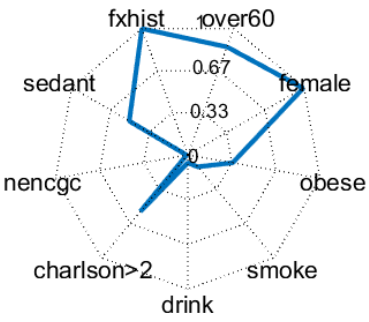
3. Elderly women, systemic steroid users

cluster4, Npts =6901



4. Younger women (early post-menopausal), low-medium co-morbidity

cluster5, Npts =4589

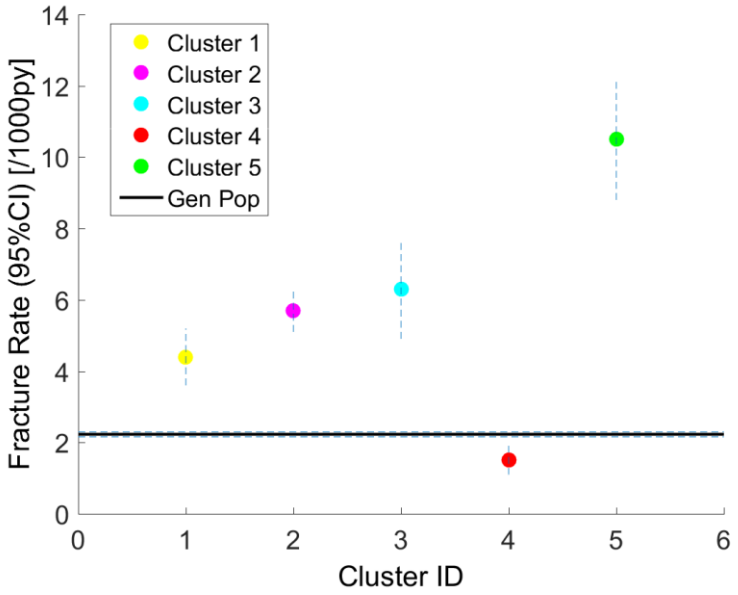


5. Secondary prevention women (previous fracture history)

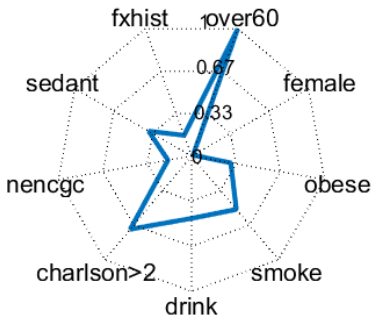
Fracture Risk of Each Cluster

What can we learn?

- User "types"/ groups
- Fracture risk
- Should they be on medication
- Costs

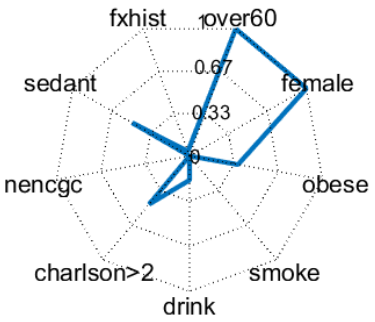


cluster1, Npts =6661



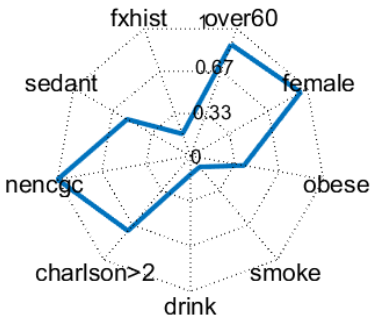
1. Elderly men, multi-morbidity, high prevalence of smoking and drinking

cluster2, Npts =16124



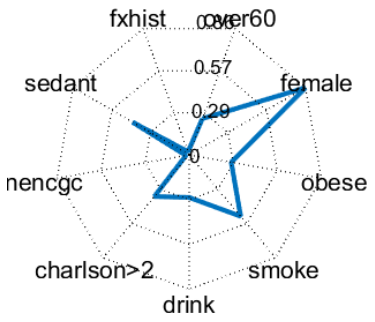
2. Elderly women, high co-morbidity

cluster3, Npts =3721



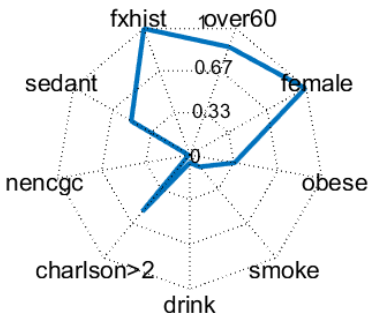
3. Elderly women, systemic steroid users

cluster4, Npts =6901



4. Younger women (early post-menopausal), low-medium co-morbidity

cluster5, Npts =4589



5. Secondary prevention women (previous fracture history)

Handle with Care

Top scientists call for caution over artificial intelligence

Artificial intelligence has the potential to eradicate disease and poverty, say world's top scientists, but researchers must not create something which cannot be controlled

Artificial intelligence must be carefully considered, say scientists Photo: REX

When to do What and How

When	What	How
Select most predictive features from n>>1 features	Feature selection	• Recursive feature elimination • Lasso • SVM,…
Predict disease A versus disease B/ other outcome	Supervised classification	• LR • MLP • SVM • RForests • *k*NN, ….
Don't know if there are sub-types of disease	Unsupervised learning	• *k*-Means • Hierchical • GMM,…
Can I visualise disease sub-types, and tell if they are the same as for other diseases?	High-dimensional visualisation	• PCA • Neuroscale map • Sammon's map,…

ML in Epidemiology Literature

American Journal of Epidemiology

Vol. 177, No. 5
DOI: 10.1093/aje/kws241
Advance Access publication: January 29, 2013

Practice of Epidemiology

Mortality Risk Score Prediction in an Elderly Population Using Machine Learning

Sherri Rose*

* Correspondence to Dr. Sherri Rose, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205 (e-mail: srose@jhsph.edu).

Deep Learning for Health Informatics

Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang, *Fellow, IEEE*

***Abstract*—With a massive influx of multimodality data, the role of data analytics in health informatics has grown rapidly in the last decade. This has also prompted increasing interests in the generation of analytical, data driven models based on machine learning in health informatics. Deep learning, a technique with its foundation in artificial neural networks, is emerging in recent years as a powerful tool for machine learning, promising to reshape the future of artificial intelligence. Rapid improvements in computational power, fast data storage, and parallelization have also contributed to the rapid uptake of the technology in addition to its predictive power and ability to generate automatically optimized high-level features and semantic interpretation from the input data. This article presents a comprehensive up-to-date review of research employing deep learning in health informatics, providing a critical analysis of the relative merit, and potential pitfalls of the technique as well as its future outlook. The paper mainly focuses on key applications of deep learning in the fields of translational bioinformatics, medical imaging, pervasive sensing, medical informatics, and public health.**

***Index Terms*—Bioinformatics, deep learning, health informatics, machine learning, medical imaging, public health, wearable devices.**

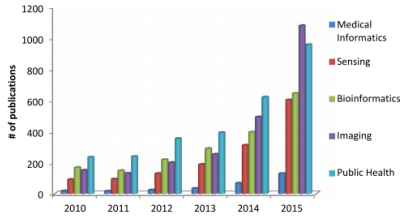


Fig. 1. Distribution of published papers that use deep learning in subareas of health informatics. Publication statistics are obtained from Google Scholar; the search phrase is defined as the subfield name with the exact phrase *deep learning* and at least one of *medical* or *health* appearing, e.g., "public health" "deep learning" medical OR health.

an automatic feature set, which otherwise would have required hand-crafted or bespoke features.

In domains such as health informatics, the generation of this

Statistics in Medicine

Research Article

Received 4 April 2009, Accepted 8 October 2009 Published online 3 December 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3782

Improving propensity score weighting using machine learning

Brian K. Lee,[a*†] Justin Lessler[b] and Elizabeth A. Stuart[c,d]

ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 398–407

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes

Peter C. Austin[a,b,c,*], Jack V. Tu[a,b,d], Jennifer E. Ho[e,f,g], Daniel Levy[e,f,h], Douglas S. Lee[a,b,i]

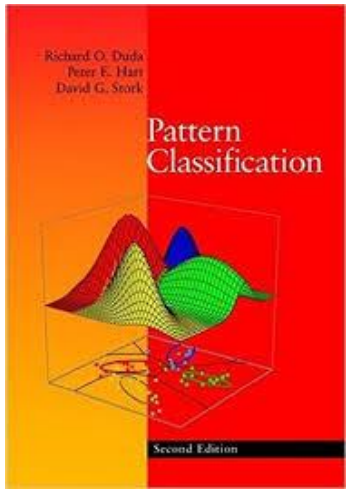
Software, Online Resources, and Books

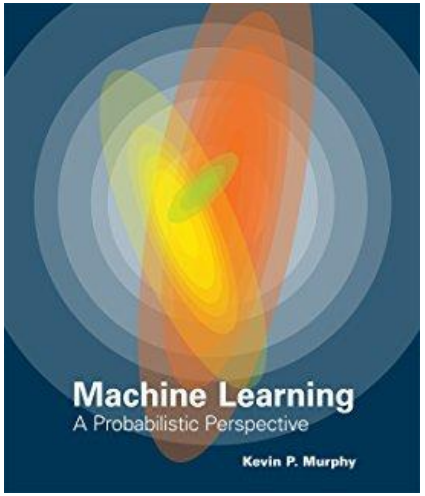
Tutorials and courses:

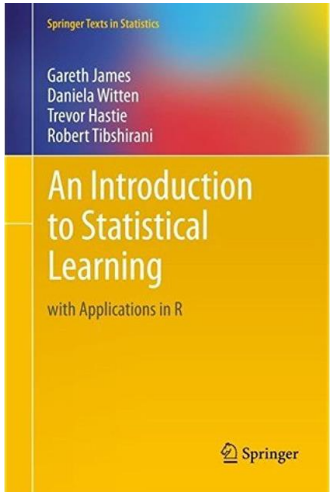
- coursera.org
- datacamp.org
- bigML.com
- Kaggle.com

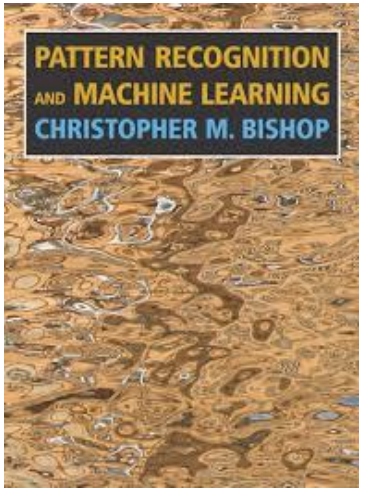
Data repositories:

- www.kdnuggets.com
- Mathworks.com (proprietary)









NDORMS

NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

Medical Sciences Division

Big Data Methods

An Introduction to Machine Learning and Real-world Applications

Sara Khalid

NDORMS, University of Oxford