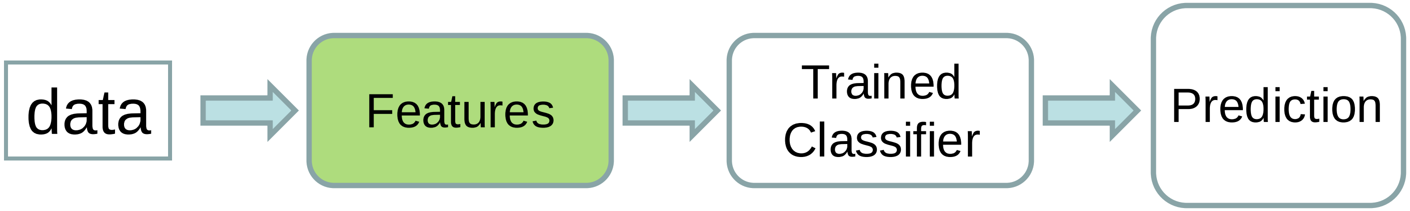
COMS20017 – Algorithms & Data

March 2025
Features

Majid Mirmehdi

Lecture MM-06

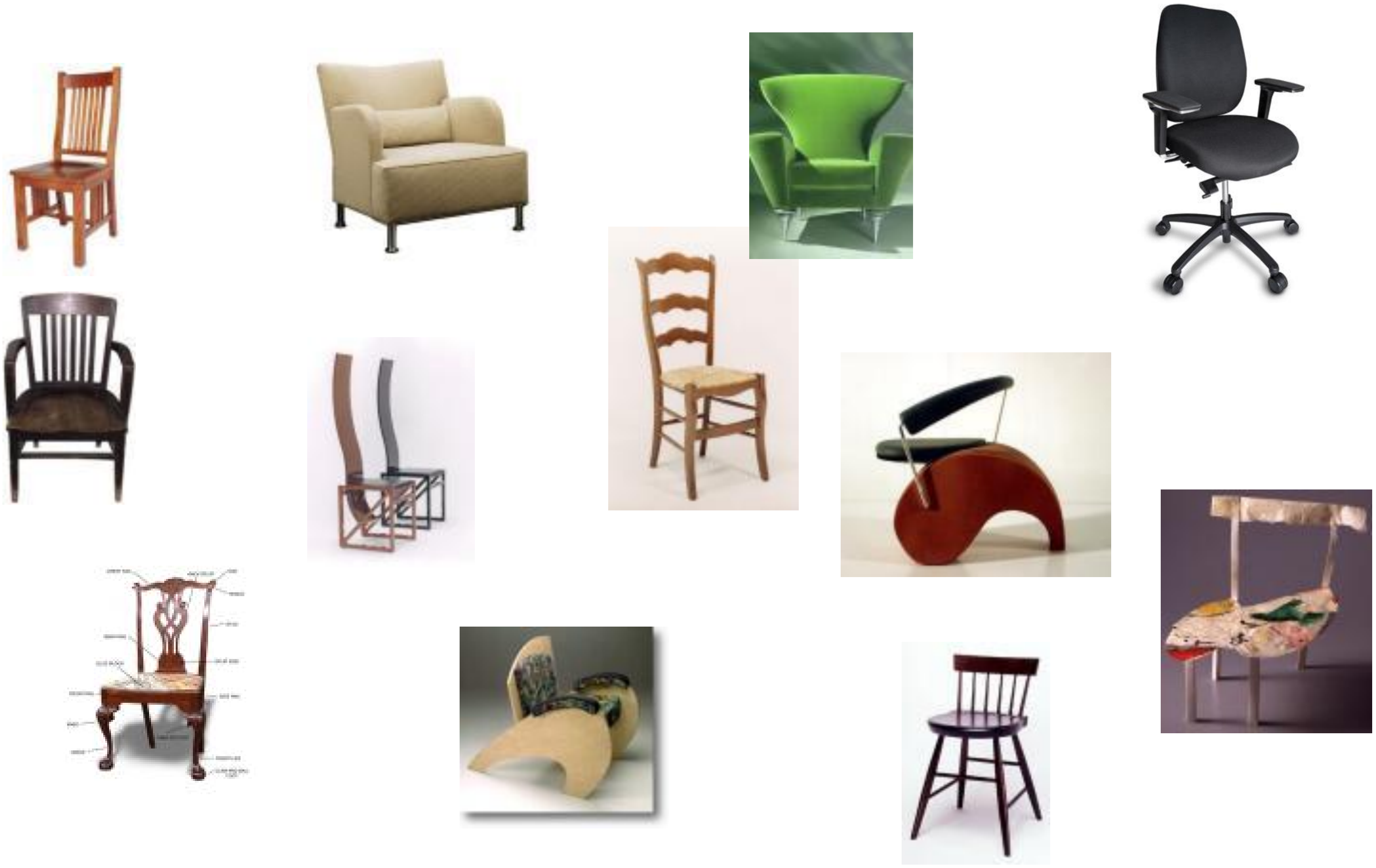
Next in DATA



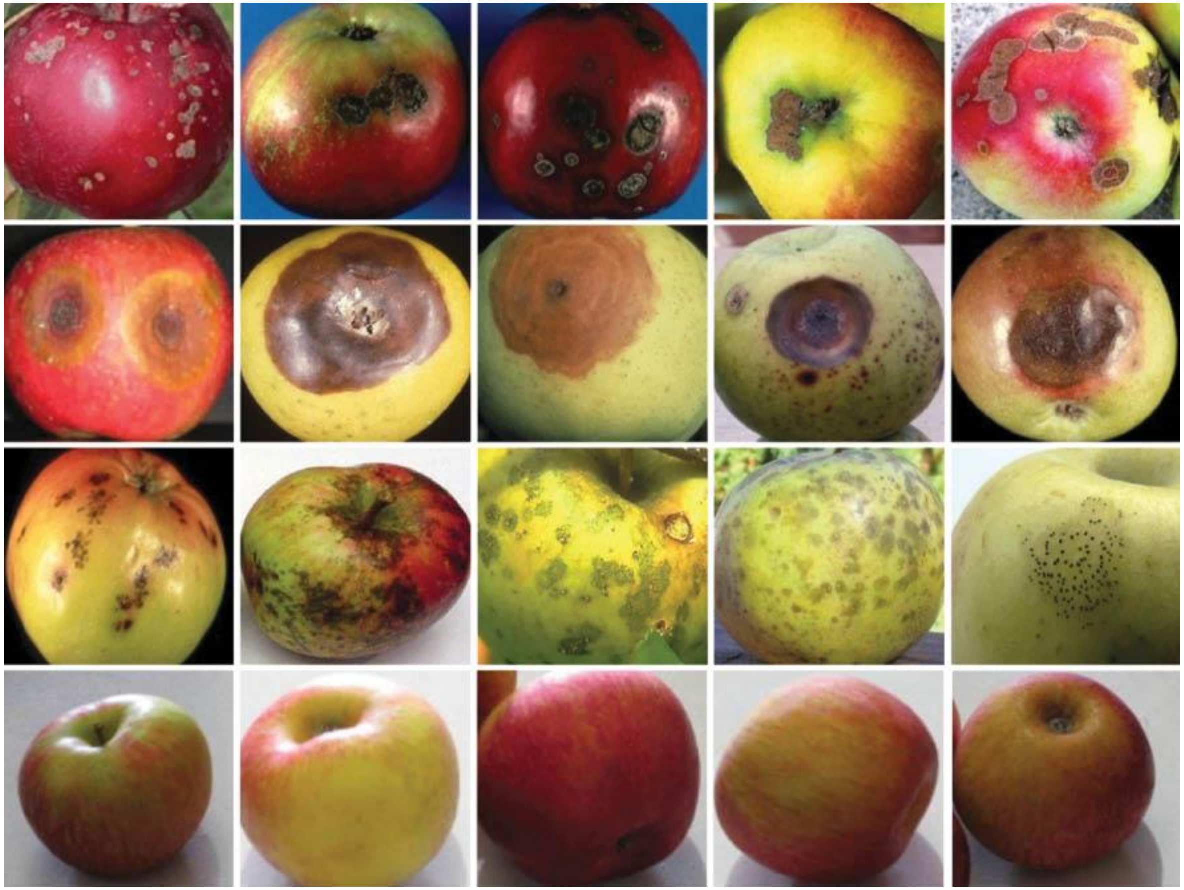
Feature Selection and Extraction

- Signal basics and Fourier Series
- 1D and 2D Fourier Transform
- **Another look at features**
- PCA for dimensionality reduction
- Convolutions

Features for recognizing a chair?



Features for recognizing good/bad apples?

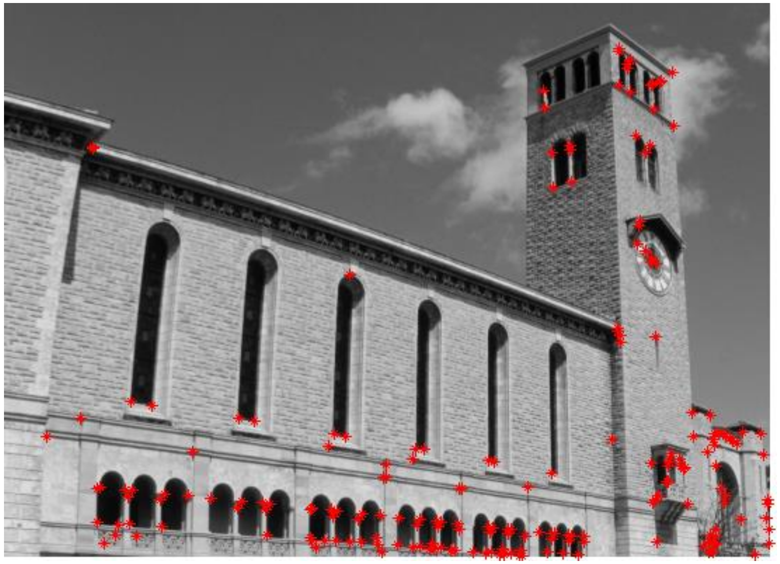


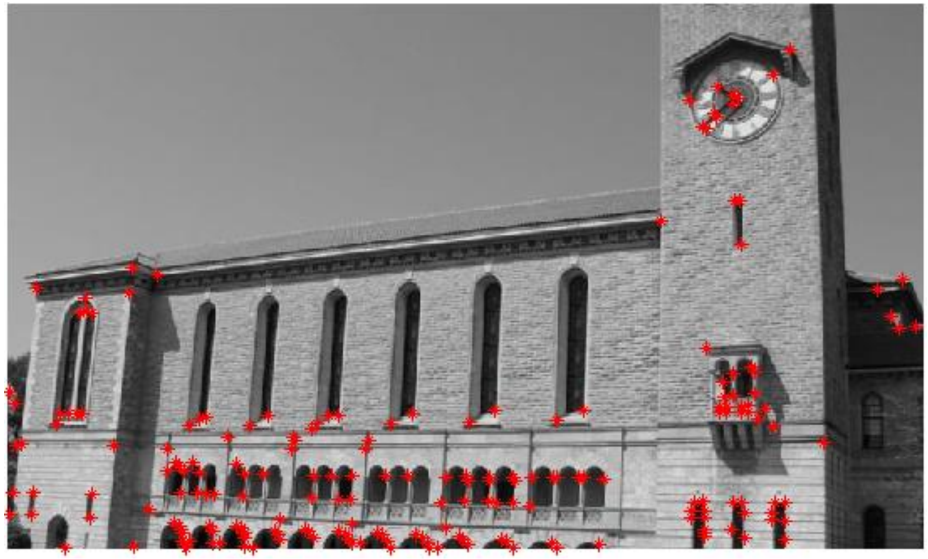
Features for recognizing masculine/feminine words in French?

• le fromage (cheese)	la salade (salad, lettuce)
• le monument (monument)	la télévision (television)
• le couteau (knife)	la culture (culture)
• le téléphone (telephone)	la situation (situation)
• le microscope (microscope)	la société (society)
• le romantisme (romanticism)	la différence (difference)

Image Features for Recognising Buildings?

Matching features (while also exploiting how they are arranged in the scene) would be much more efficient than matching all pixels





Common features between images allows us to perform tasks such as scene matching, face recognition, 3D model generation, and much more!

Examples of Features

- Primitive features, e.g.: (or Discrete/Continuous or Quantitative)
  - weight, length, width, height, volume …
  - amplitude, frequency, phase, duration, roll-off, flux …
  - beats per minute, temperature, pressure,...
  - edges, corners, lines, curvature, …
  - mean RGB colour, colour histogram, …
- Semantic features, e.g.: (or Nominal or Qualitative)
  - colour layout (red, cyan, magenta,…)
  - texture descriptors (coarse, fine, rough, smooth,…)
  - shape descriptors (rectangular, circular, elliptical,…)
  - kind of day (warm, cold, sunny, rainy, …)
- Ordinal features, e.g.: (or Ordered)
  - Education Level (Higher, Secondary, Primary)
  - Body Mass Index (BMI): Underweight, normal weight, overweight, obese
  - Letter grades in the exam (A, B, C, D, etc.)
- Statistical features, e.g.:
  - mean, median, variance, percentiles, moments, ...
- ...

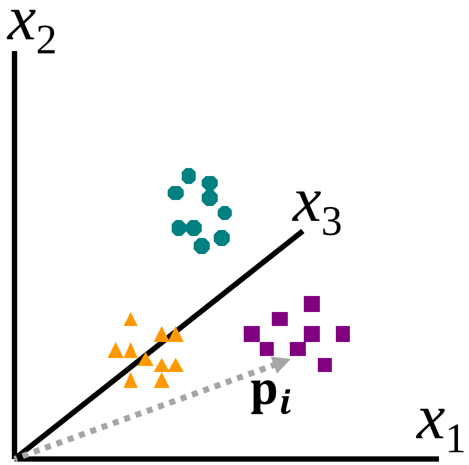
Quick Review: Features

$$\mathbf{p}_i = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_d \end{bmatrix}$$

- Features describe characteristics of our data.
- The combination of d features is represented as a d-dimensional column vector called a *feature vector*.
- The d-dimensional space defined by the feature vector is called the *feature space*.

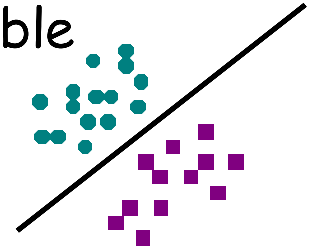
$\mathbf{p}_i \in X$ is a point in feature space X

Example: 3D feature space X

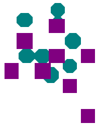


Feature Properties – *what makes a good feature vector?*

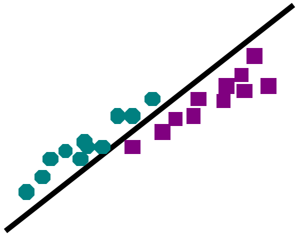
good, linearly-separable features



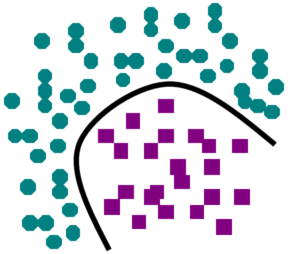
bad features



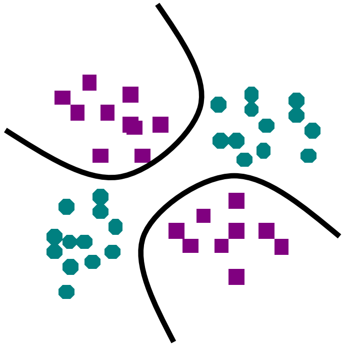
highly correlated features



nonlinearly-separable features

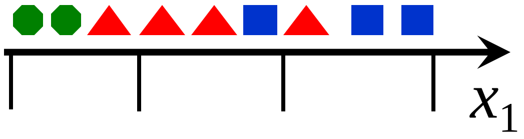


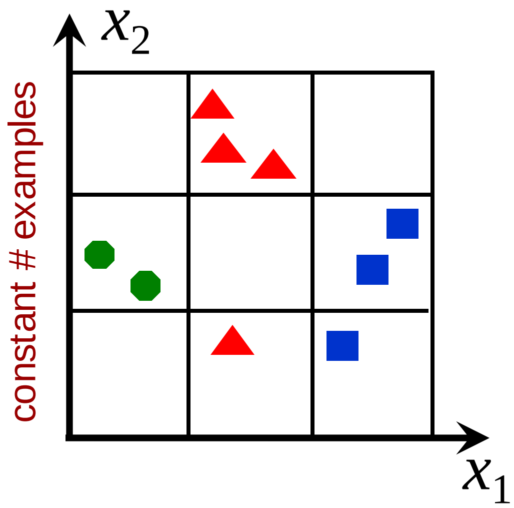
multimodal features

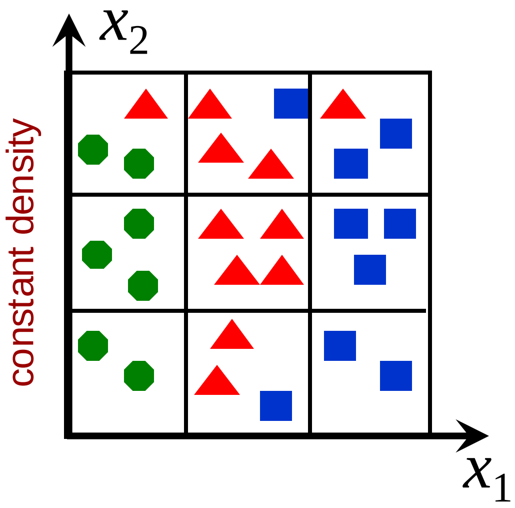


The Curse of Dimensionality

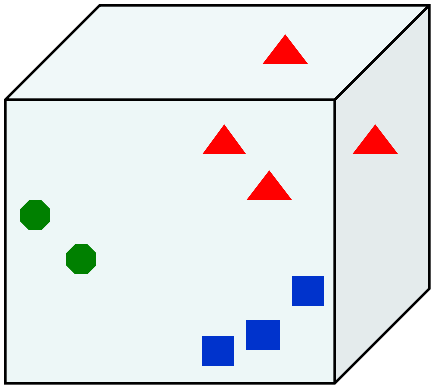
- Example: one feature, no. of bins = 3, but results not good enough, so incorporate another feature.
- But, the no. of bins → 3^2=9, so to maintain the density, increase the no. of examples from 9 to 27







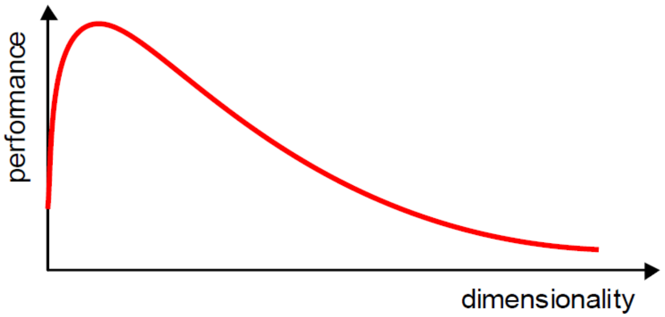
Moving to 3 features makes it even worse…



The amount of data to sustain a given spatial density increases exponentially with the dimensionality of the input space

Dimensionality Reduction

- Strive for compact representation of the *properties* of data, such that redundancy/irrelevancy is removed.
- The choice of features is very important, as it influences:
  - accuracy of classification
  - time needed for classification
  - difficulty in performing classification
  - no. of learning examples



Feature Selection/Extraction – given N features

- Feature selection: Choosing $d<N$ important features, ignoring the remaining $N - d$ features

  Subset selection algorithms

- Feature extraction: Project the original m_i , $i = 1,\ldots,N$ dimensions to new $d<N$ dimensions, x_j , $j = 1,\ldots,d$

  Principal Components Analysis (PCA),

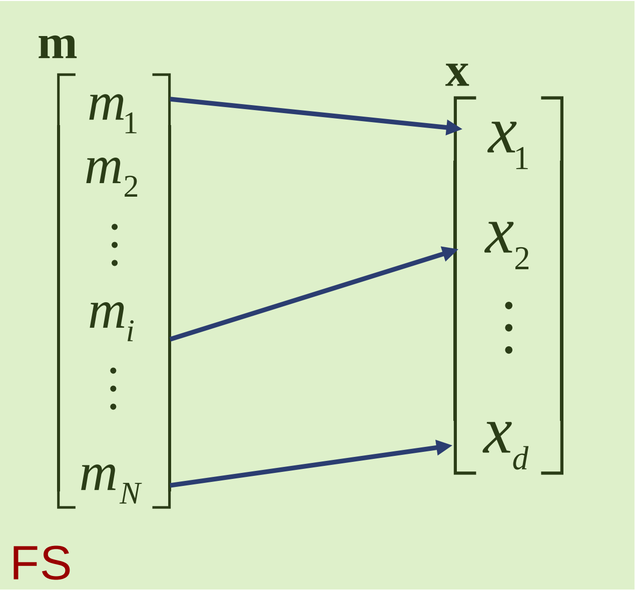
  Linear Discriminant Analysis (LDA), & many more…

- Feature construction: create new features based on old features: f=(.) where f is usually a non-linear function, e.g Support Vector Machines,…. (not within our scope)

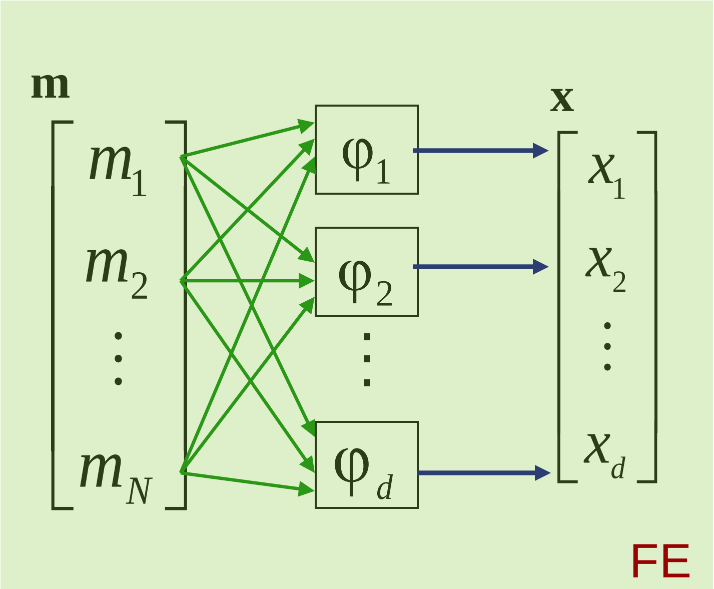
Selection or Extraction?

Two general approaches to dimensionality reduction:

- Feature Selection: Selecting a subset of the existing features without a transformation
- Feature Extraction: Transforming the existing features into a lower dimensional space



$d \ll N$



Exhaustive Feature Selection

- Feature Selection is necessary in a number of situations, e.g. there may be too many features or may be too expensive to obtain.

- Feature Selection involves a search strategy that may explore the space of all possible combinations of features.

Given a feature set $\{\boldsymbol{m}_i\}, i = 1, \dots, N$, find a subset $\mathbf{x}$ of size d with $d < N$, that optimizes an objective function $J(\mathbf{x})$, e.g. $P(correct\ classification)$.

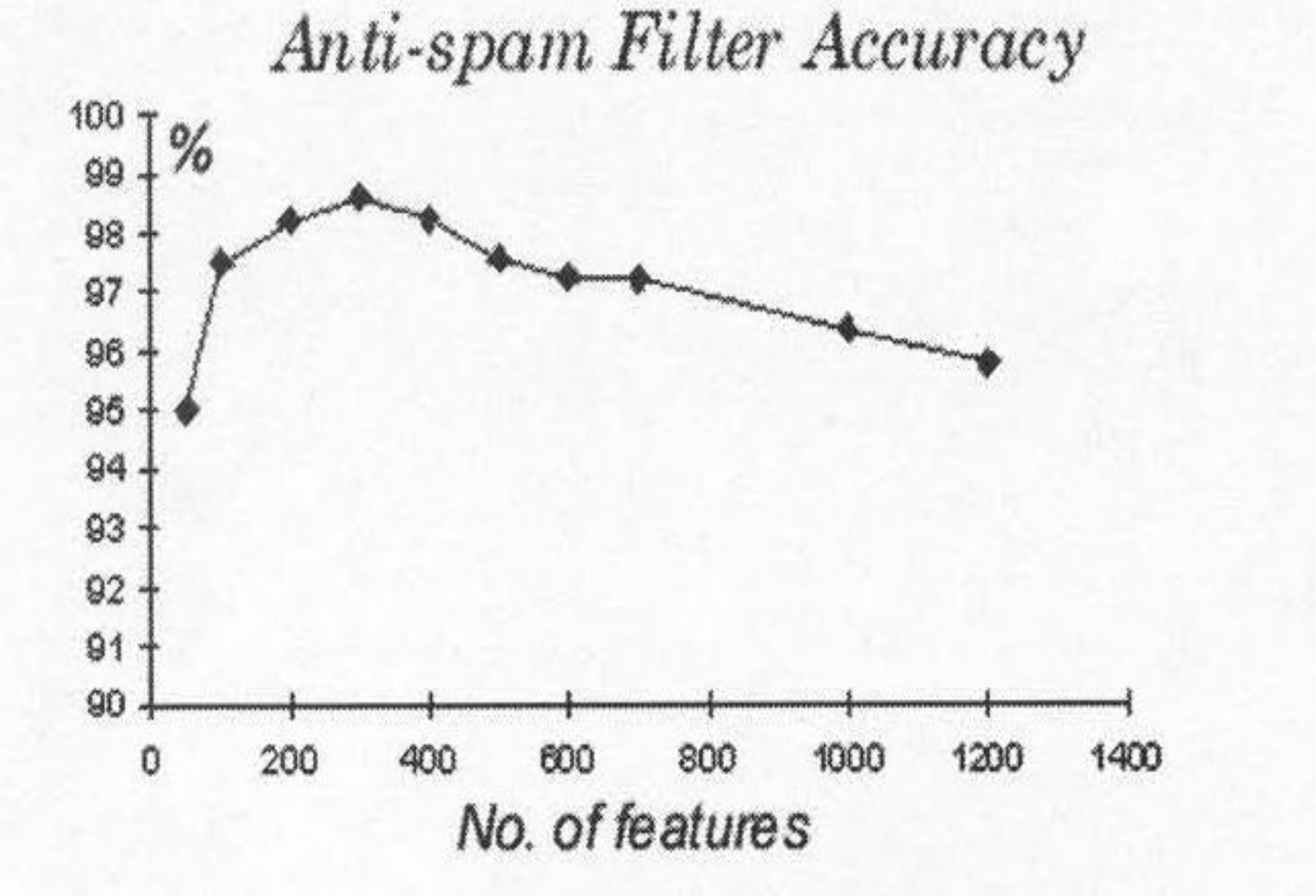
This function would have to be evaluated many times!

$$\frac{N!}{(N-d)!d!}$$

e.g. for 10 features out of 25 one would still have

to consider 3,268,760 feature subsets!

Feature Selection?



From http://elpub.scix.net/data/works/att/02-28.content.pdf

Sequential Feature Selection Methods

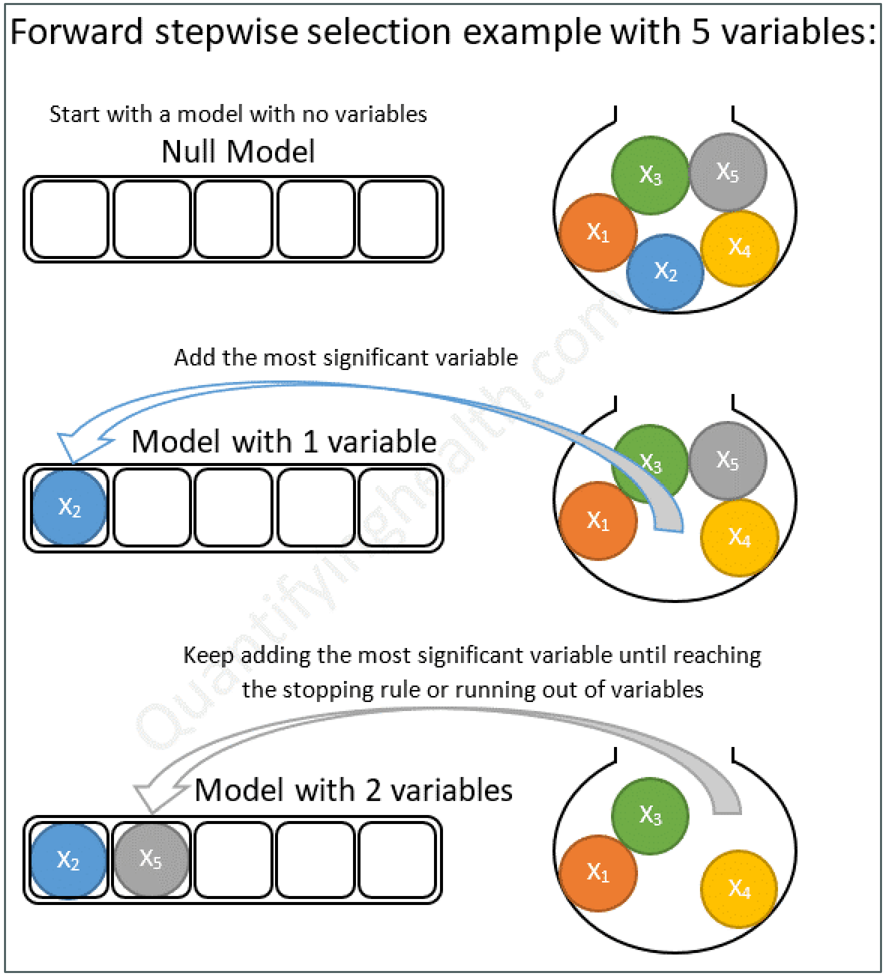
- Assume features are independent.
- Best single features can be chosen by significance tests.

Forward stepwise feature selection:
(bottom-up)

Build up d features incrementally,
starting with an empty set

The best single feature is picked first.

Then, next best feature conditioned to the first, ...


Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
X_3
X_5
X_1
X_4
X_2
Add the most significant variable
Model with 1 variable
X_2
X_3
X_5
X_1
X_4
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X_2
X_5
X_3
X_1
X_4
Quantifyinghealth.com

https://quantifyinghealth.com/stepwise-selection/

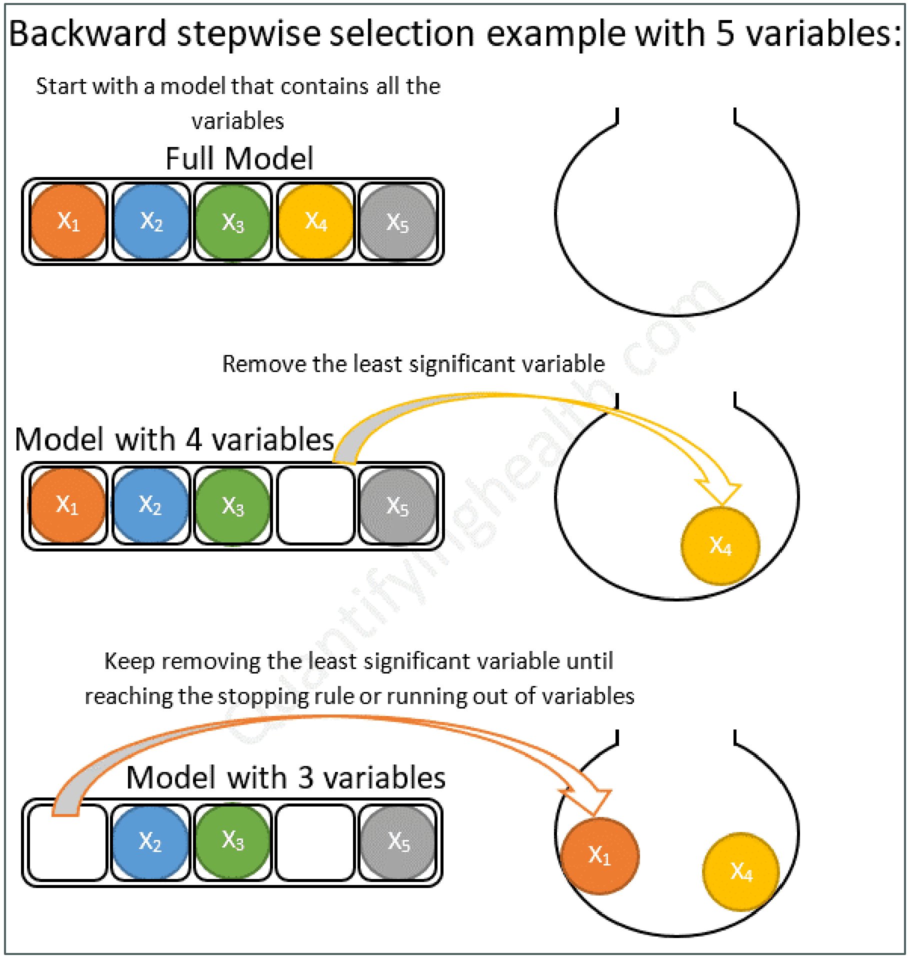
Sequential Feature Selection Methods

- Assume features are independent.
- Best single features can be chosen by significance tests.

Backward stepwise feature selection:
(top-down)

start with full set of features and remove redundant or least significant ones one after another

Until optimal d features reached

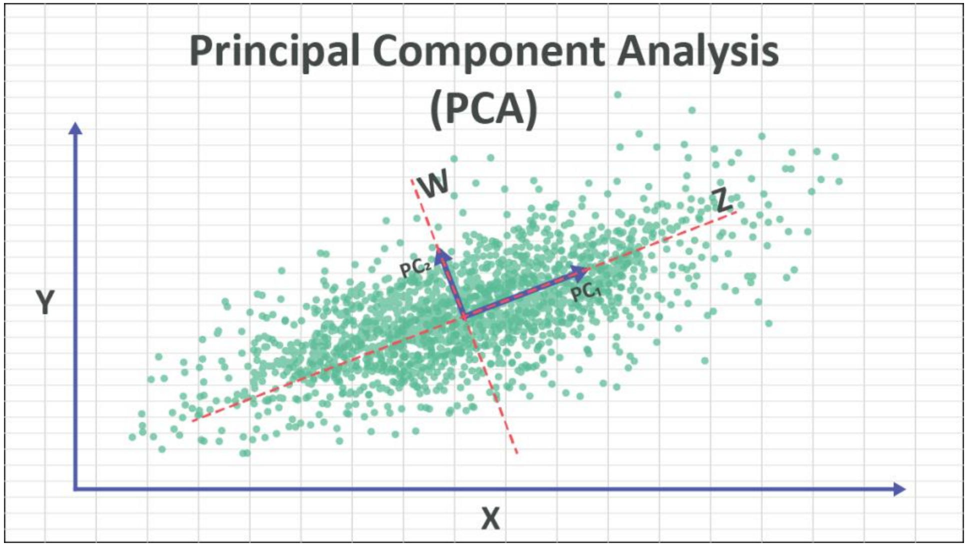

Backward stepwise selection example with 5 variables:
Start with a model that contains all the variables
Full Model
X1
X2
X3
X4
X5
Remove the least significant variable
Model with 4 variables
X1
X2
X3
X5
X4
Keep removing the least significant variable until reaching the stopping rule or running out of variables
Model with 3 variables
X2
X3
X5
X1
X4

https://quantifyinghealth.com/stepwise-selection/

Feature Extraction

- Linear or non-linear transformation of the original variables to a lower dimensional feature space → also known as *feature selection in the transformed space*.
- Given a feature space R^N with feature vectors $\mathbf{m}$, find a mapping $\mathbf{x} = \varphi(\mathbf{m}): RN \Rightarrow R^d, d < N$, such that the transformed feature vector $\mathbf{x} = \{x_i\} \epsilon\, R^d$ preserves (most of) the information or structure in R^N.

Principal Components Analysis (PCA) is an example of a transformed space for dimensionality reduction from which we can extract the most efficient features.



PCA

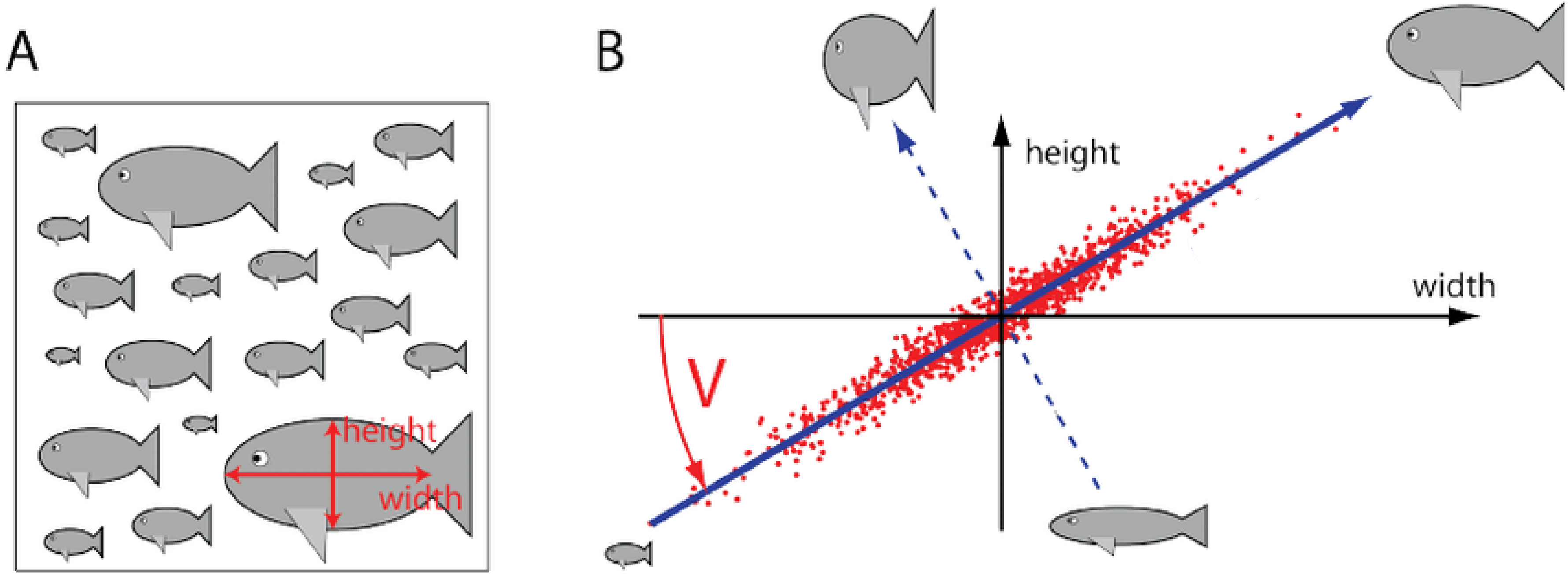


Image from Werner et al, PLoS ONE 9(11), 2014

Measuring Importance of Features (basic example)

- Simple scenario: binary classification problems
- Does feature f_i make the mean of samples from each class statistically significant?

- The t_score of any feature f_i is:

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

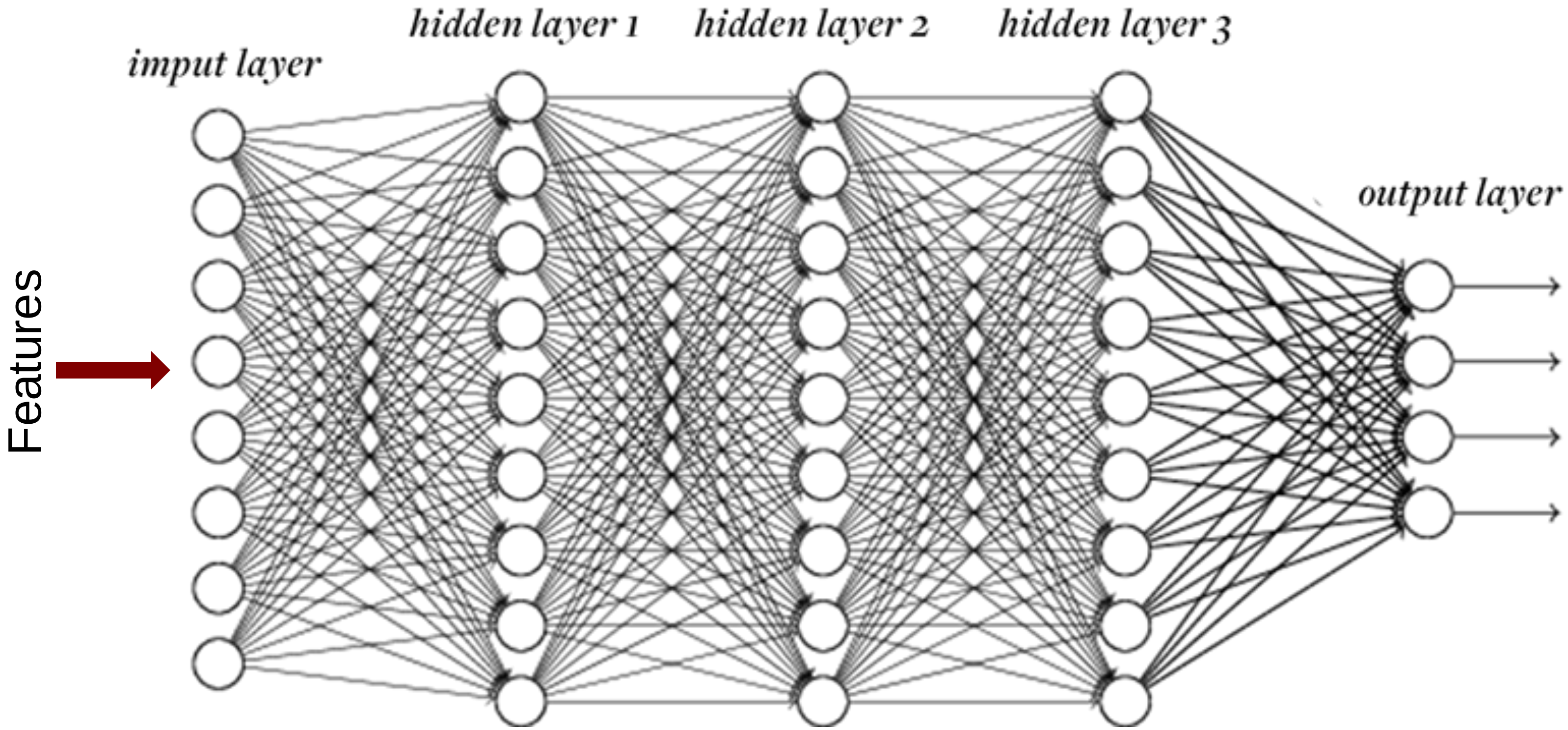
μ_1 and μ_2 are the means of samples from the two classes

σ_1 and σ_2 are the standard deviations for samples from the two classes

The higher the t_score, the more important the feature is.

https://nautil.us/is-artificial-intelligence-permanently-inscrutable-236088/

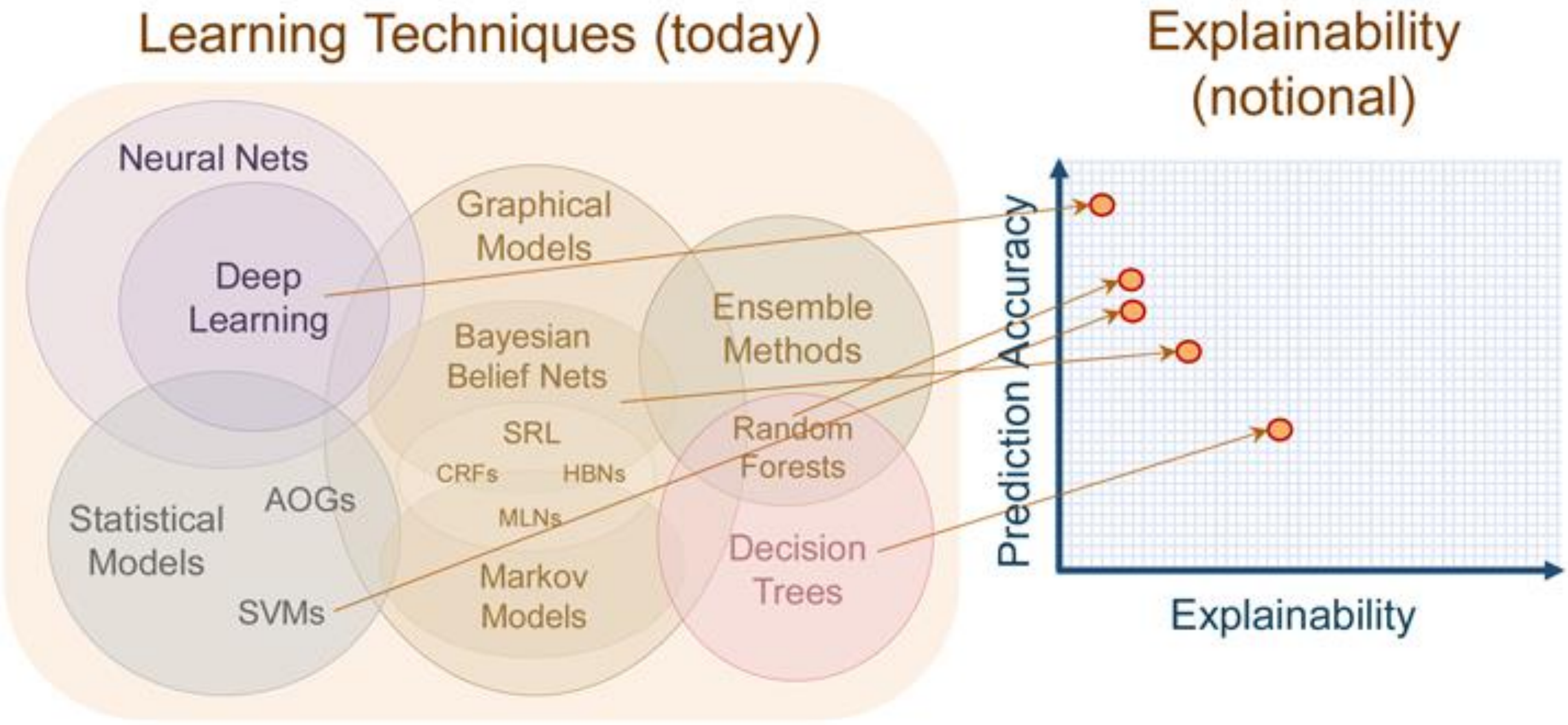
Learning Techniques and Explainability



Hidden meanings: In neural networks, data is passed from layer to layer, undergoing simple transformations at each step. Between the input and output layers are hidden layers, groups of nodes and connections that often bear no human-interpretable patterns or obvious connections to either input or output. “Deep” networks are those with many hidden layers.
Michael Nielsen / NeuralNetworksandDeepLearning.com

https://nautil.us/is-artificial-intelligence-permanently-inscrutable-236088/

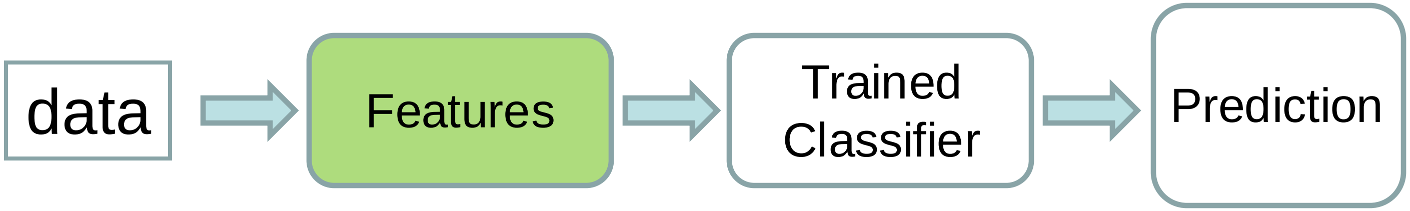
Learning Techniques and Explainability



Modern learning algorithms show a tradeoff between human interpretability, or explainability, and their accuracy.

https://nautil.us/is-artificial-intelligence-permanently-inscrutable-236088/

Next in DDCS



Feature Selection and Extraction

- Signal basics and Fourier Series
- 1D and 2D Fourier Transform
- Another look at features
- **PCA for dimensionality reduction**
- Convolutions