



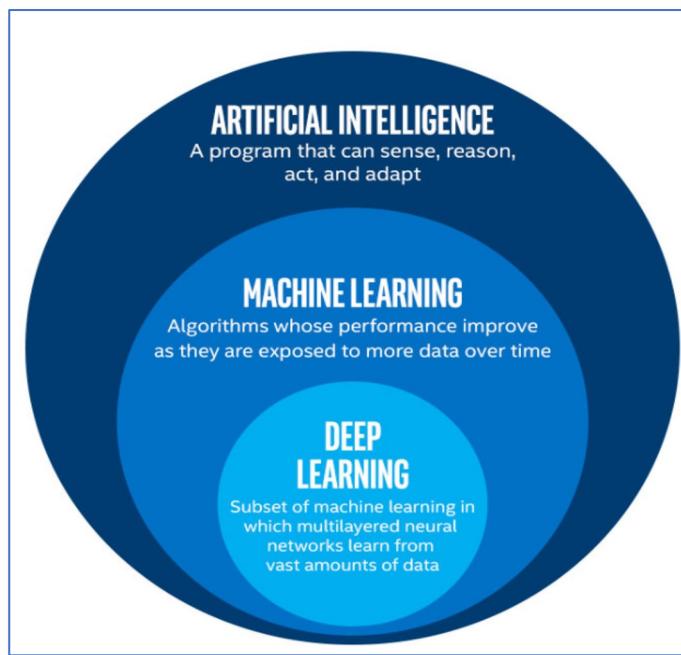
# Session 4: Machine Learning Basics

4/11/2022

NCME Data Science Training Workshop  
Instructors: Oren Livne & Jiangang Hao

# Machine Learning

A term coined by Arthur Samuel in 1959, refers to a set of methodologies that allow computers to “learn” the relationship among numerical representations of data without explicit instructions by human experts



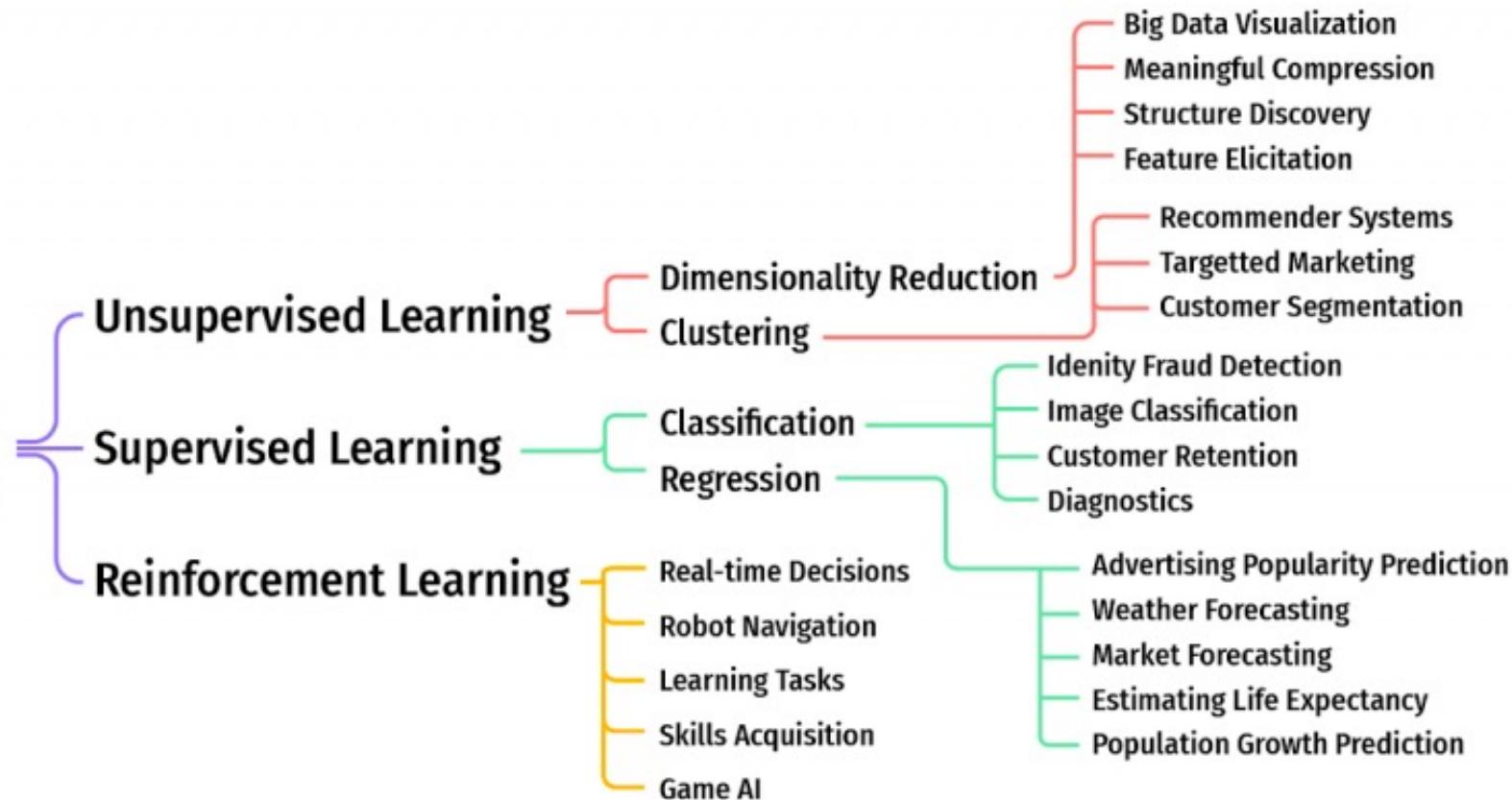
Mat Velloso  
@matvellos0

Difference between machine learning and AI:  
If it is written in Python, it's probably machine learning  
If it is written in PowerPoint, it's probably AI

22/11/18, 5:25 PM

3,514 Retweets 10.8K Likes

# Main Paradigms of Machine Learning

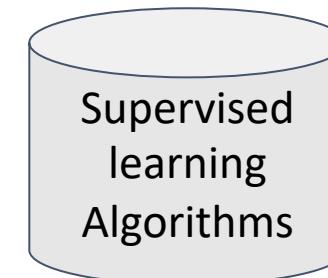


# Supervised learning

Machine learns the relationship (mapping) between **input** variables (features) and **output** variables (labels)

0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9

data



0
1
2
3
4
5
6
7
8
9

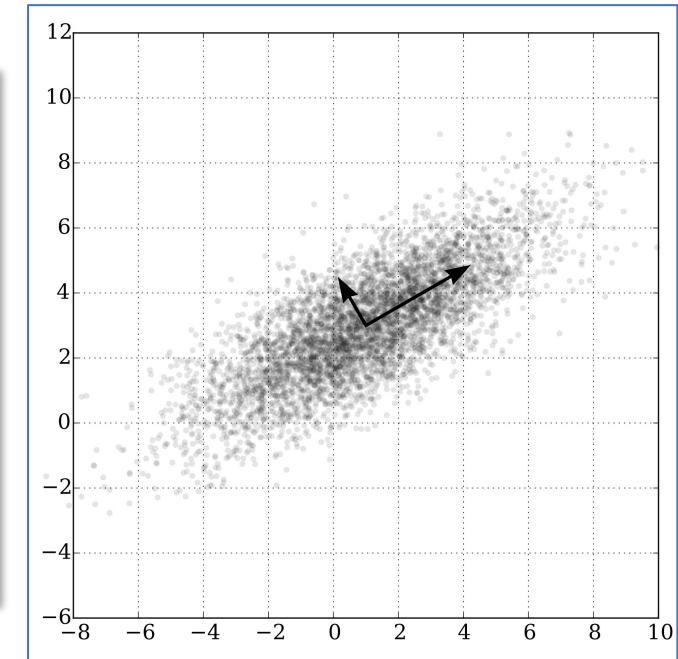
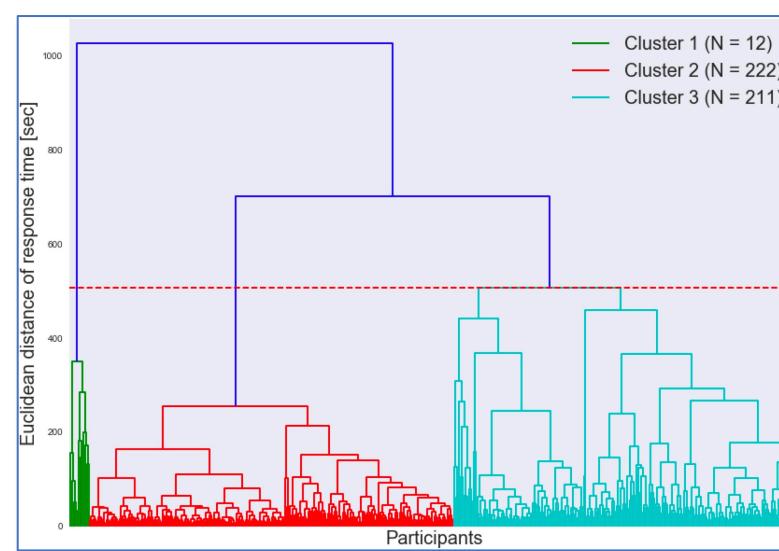
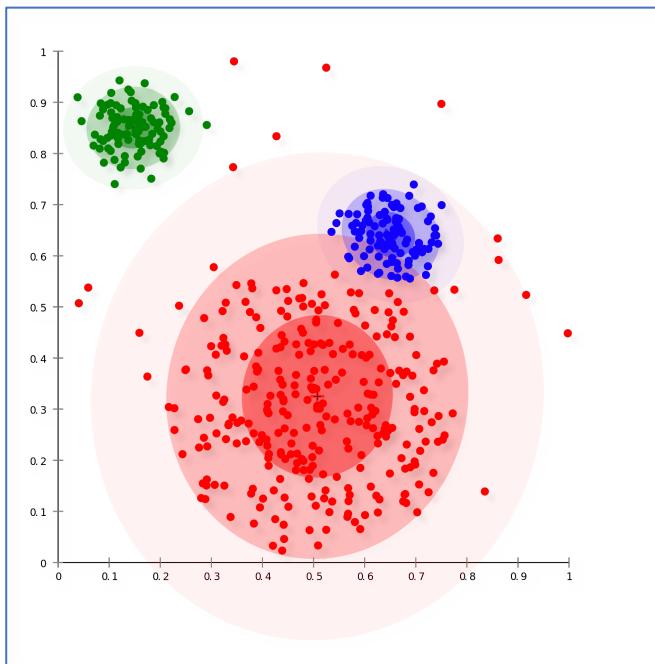
labels



# Unsupervised learning

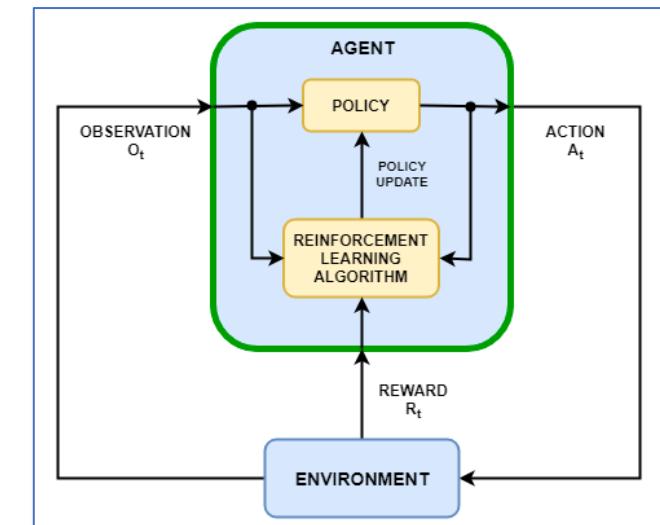
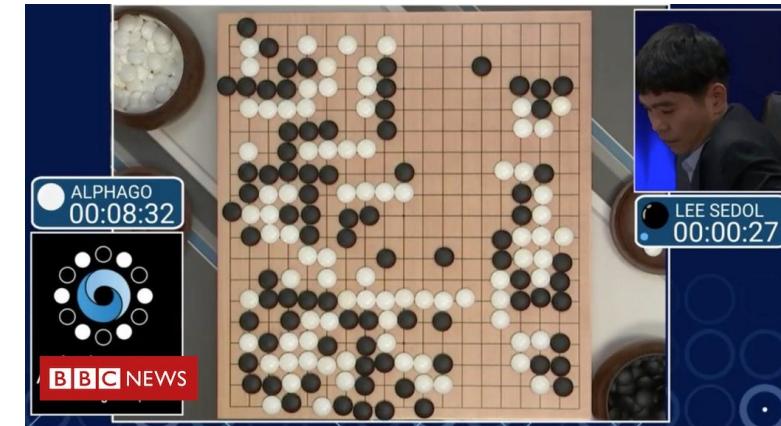
Methodology that discovers concentrations, associations, or correlations in data.

- Cluster analysis
- Dimensionality reduction



# Reinforcement Learning

- Machine learns optimal strategies by means of a goal-oriented exploration of a parameter or state-space to optimize a reward function (Sutton, & Barto, 1998).
- Key components
  - Agent
  - States
    - Hidden
    - Observable (observation)
  - Actions: what the agent can do
  - Policy/strategy: tell what the agent should do
  - Reward: what the agent gets after an action



From MathWorks





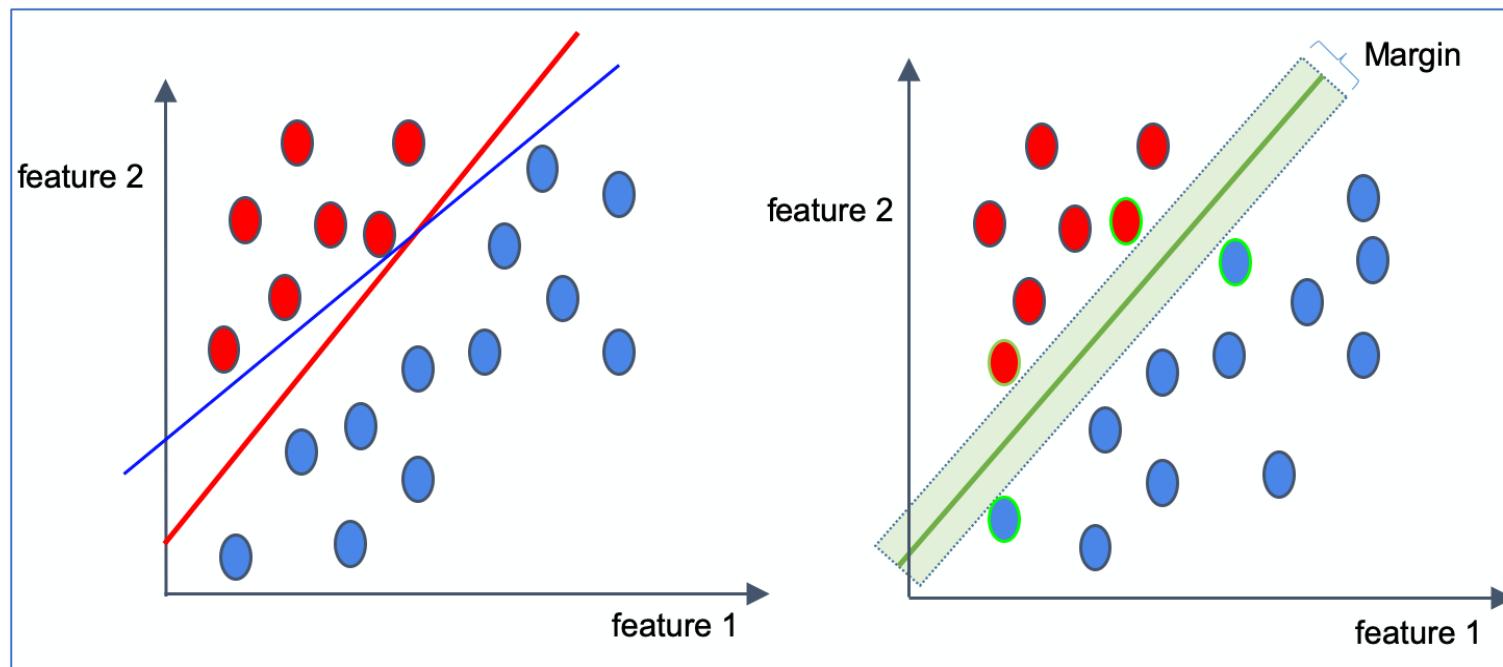
# Supervised Learning

# Two Upfront Questions

- How supervised learning is different from statistical inference, as they both aim at mapping the relationship between the independent and dependent variables?
  - Machine learning is more prediction-driven while statistical inference emphasizes both prediction and model parameter estimation.
  - Statistical inference emphasizes developing a probabilistic model to characterize the data and then estimating the model parameters based on some (usually well-studied) probability distribution functions, while machine learning emphasizes computation algorithms that can efficiently carry out the inference process by minimizing certain loss functions that do not necessarily have a probabilistic underpinning.
  - Statistical inference usually deals with data with a small number of variables obtained through planned *experiments* or quasi-experimental comparisons while machine learning handles sparse and high-dimensional data with a large number of variables (features), usually obtained through passive and uncontrolled *observations*.
- Which, if any, supervised learning methods are consistently superior to the others?
  - It is generally data and problem dependent
  - Some methods work well based on many explorations
    - SVM
    - Random Forest
    - XGBoost (Gradient Boosting Machine)
    - Deep Neural Network

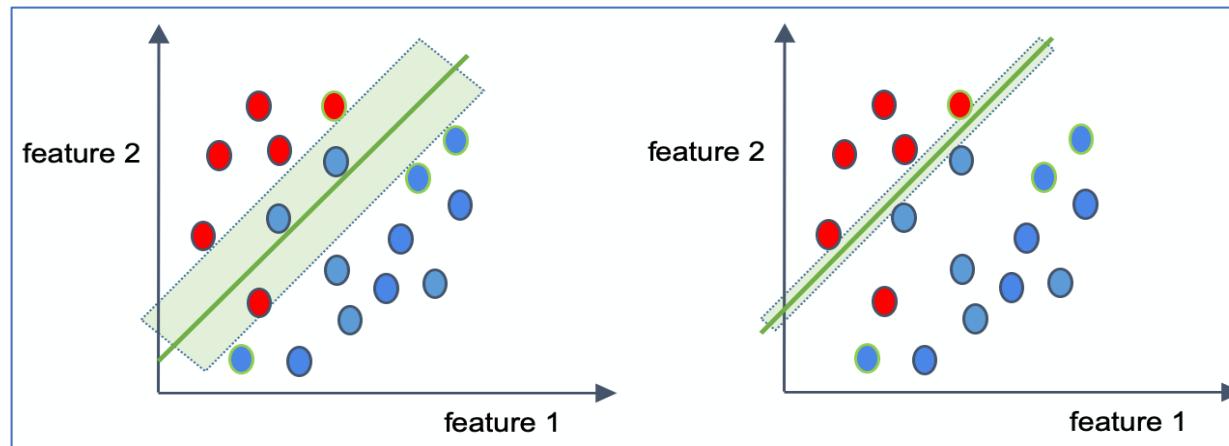
# Support Vector Machine

It aims at finding a decision hyperplane in the feature space so that data points can be separated into different categories with a maximum margin that is defined as the distance of the closest points (support vectors) from each category to the decision hyperplane (Vapnik, 1963).



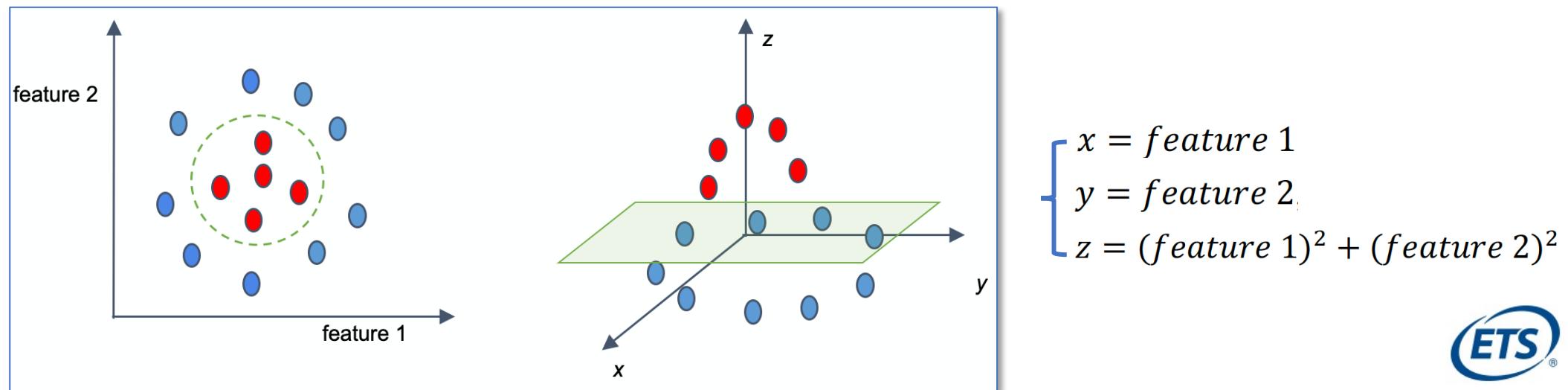
# Non-linearly Separable Case

- Choose between a decision hyperplane that has a larger margin but more misclassified data points, and the one with a narrower margin but fewer misclassified data points, as illustrated in the left and right panels
- A hyperplane with a narrower margin may lead to better classification accuracy based on the training data, but is more susceptible to overfitting, while a hyperplane with a larger margin has lower classification accuracy with respect to the training data, but it is more robust against overfitting.
- In SVM algorithm, the regularization hyper-parameter ( $C$ ), is used to control this balance. Choosing a larger  $C$  indicates that one favors a narrower margin with a lower tolerance of misclassified data points



# Kernel Trick

- The transformation of the original feature space into a higher dimensional space to make the data linearly separable is done through a procedure called kernel trick.
- The characteristic of data being more likely to be linearly separable when being projected into a higher dimensional space via some non-linear transformation is known as Cover's theorem (Cover, 1965).
- In SVM algorithms, one can choose the kernel type – another hyperparameter. Typical kernels are polynomial and RBF.



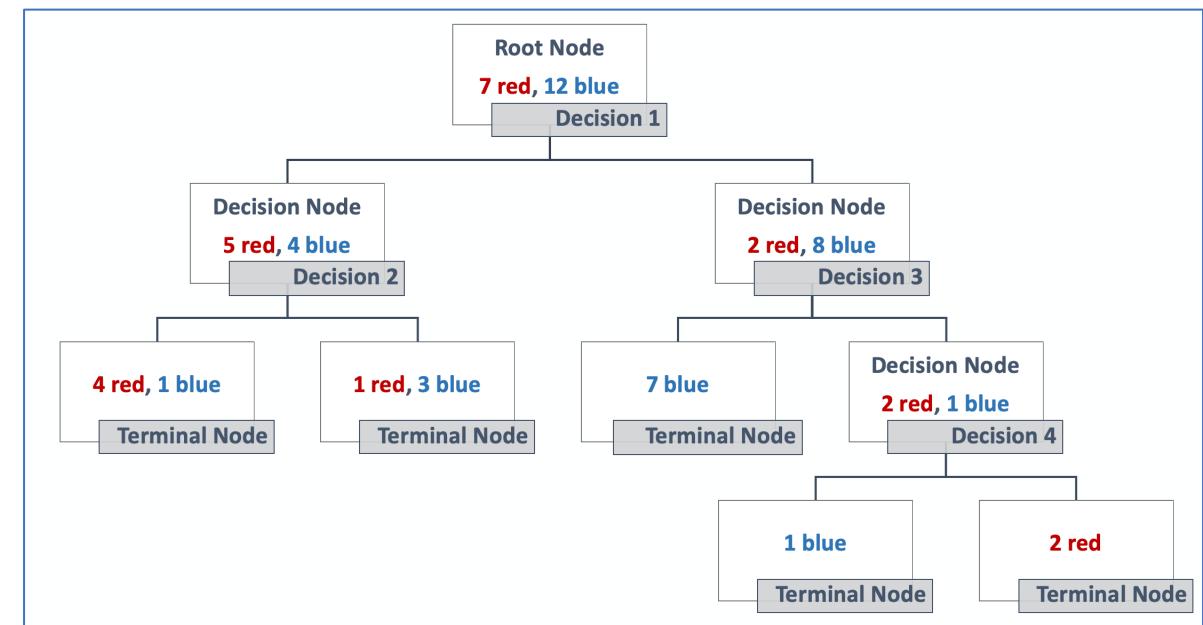
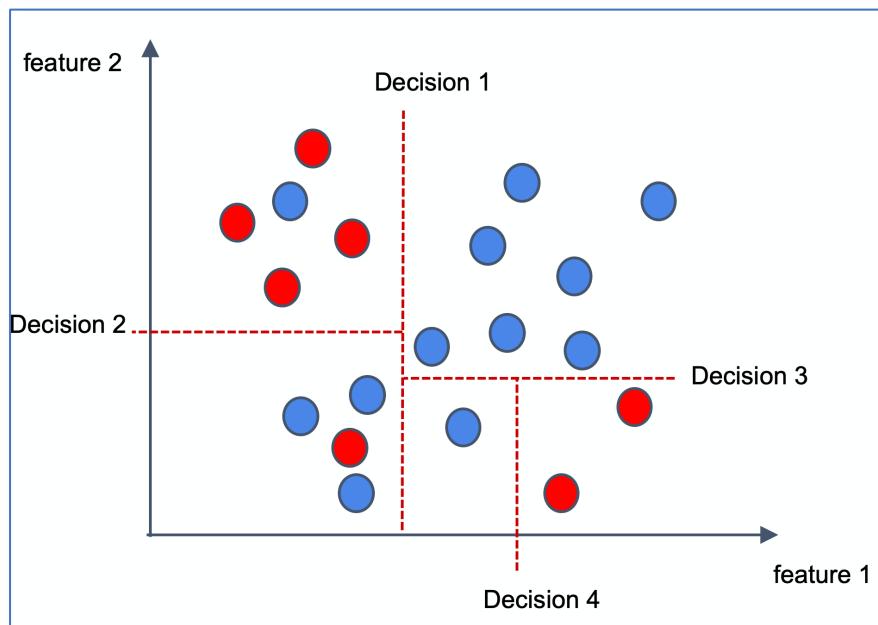
# Ensemble Learning

- Given there are many supervised learning methods, it is always tantalizing to think whether some sort of “average” from an ensemble of learners will lead to better predictive performance. Research studies in this direction leads to ensemble learning (Dietterich, 2000).
- Two leading ideas regarding how an ensemble of learners is constructed:
  - Bagging (bootstrap aggregating) (Breiman, 1996)
    - The ensemble is created through bootstrapping the dataset (Efron & Tibshirani, 1994)
    - The outputs from the learners from each sample will be combined (average or voting) to improve the predictive performance.
    - Example: Random Forest (Ho, 1995; Breiman, 2001)
  - Boosting (Friedman, 2001)
    - The ensemble is constructed sequentially, rather than in parallel, as in the bagging approach.
    - A learner is first trained on the full dataset and subsequent learners are added by fitting the residuals (after applying all preceding learners), by which new learners will assign more weights to the poorly predicted observations by the preceding learners.
    - Example: Gradient Boosting Machine



# Decision Tree

A decision tree works by forming decision rules on the features recursively to minimize some loss function of the classification.



# Decision Tree Cont'd

- Loss function
  - Gini index as used in the CART algorithm (Breiman et al., 1984).
  - Information Gain as used in the Iterative Dichotomiser 3 algorithm (Quinlan, 1986).
- Stopping rule
  - Condition under which no further splitting will happen.
  - E.g., minimum leaf size
- Pruning
  - A procedure to remove the sub-nodes from decision nodes performed after the training to avoid overfitting.
- Main drawback:
  - Instability of the results, as a small change in a dataset may lead to quite different decision rules.



# Random Forest

- Suppose we have  $N$  observations and  $M$  features in a dataset, then a typical random forest classifier goes as follows:
  1. Create  $K$  samples of the dataset using the bootstrapping technique, and each sample has  $N$  observations
  2. For each of the  $K$  samples, a decision tree is applied. In each of the decision nodes of each of the trees, a decision is made based on the most discriminative one of a randomly chosen  $F$  features from the total  $M$  features – random subspace or feature bagging (Ho, 1998).
  3. The average/voting of multiple trees will be the final estimator
- Hyper-parameters
  - The number of trees ( $K$ )
  - The size of the feature subspace ( $F$ )



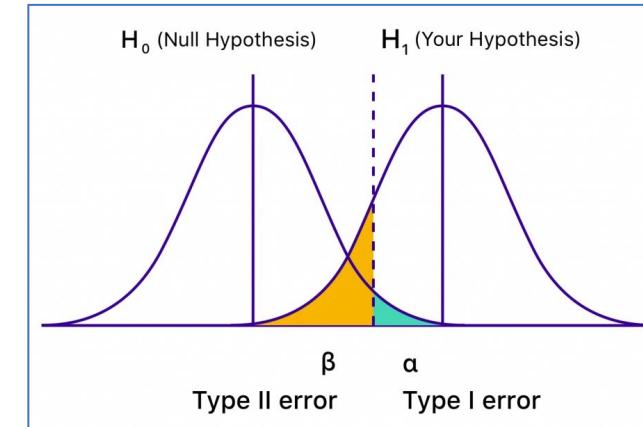
# Gradient Boosting Machine

- Assumption: the ideal mapping between the dependent and independent variables could be approximated by the sum of many functions belonging to the same parametric family.
- The parameters are fixed iteratively as following
  1. Fit the data  $(\mathbf{X}, y)$  with a base tree  $h_0(\mathbf{X}; \boldsymbol{\alpha}_0)$
  2. Add another tree  $h_1(\mathbf{X}; \boldsymbol{\alpha}_1)$  by fitting the residuals  $(\mathbf{X}, y - h_0(\mathbf{X}; \boldsymbol{\alpha}_0))$ .
  3. Repeat the procedure leads to  $F(\mathbf{X}) = \sum_{m=0}^M \beta_m h_m(\mathbf{X}; \boldsymbol{\alpha}_m)$
- Where is the “Gradient” from?
  - For loss function:  $L(y, F(\mathbf{X})) = \frac{1}{2}[y - F(\mathbf{X})]^2$
  - The Residual:  $y_i - F_m(\mathbf{X}_i) = -\left.\frac{\partial L(y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{X}_i)}\right|_{F(\mathbf{X})=F_{m-1}(\mathbf{X})} \equiv -g_m(\mathbf{X}_i)$

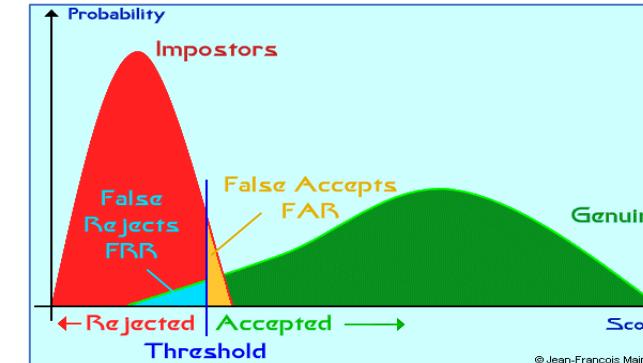


# Evaluation Metric for Binary Classification

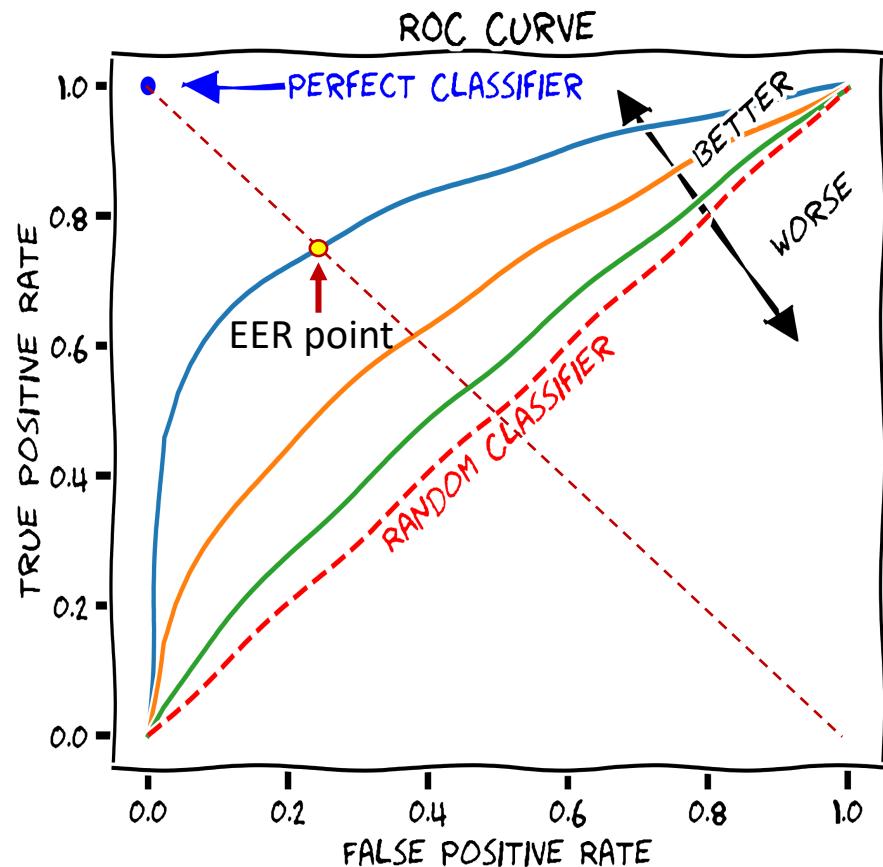
- The outcomes are in two categories: Yes/No, Positive/Negative, Accept/Reject, ...
- Example classifiers: SVM, ANN, RF, GBM
- Quantify the classification result
  - In statistical hypothesis testing
    - Null hypothesis/alternative hypothesis
    - Type I and Type II errors
  - In machine learning
    - Positive/negative
    - False Positive Rate and False Negative Rate
  - In Biometrics
    - Acceptance/rejection
    - False Acceptance/Match/ Rate and False Reject/Nonmatch Rate



		Predicted Class		
		Positive	Negative	Recall or True positive rate
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$ True negative rate
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$
		F-Score(Harmonic mean of precision and recall) = $\frac{(1+b)(PREC.REC)}{(bPREC+REC)}$ where b is commonly 0.5, 1, 2.	Positive Predicted value	



# Receiver Operating Characteristic - ROC



- Area under curve (AUC):
  - 0.5: non discriminative
  - 0.7 – 0.8: acceptable
  - 0.8 – 0.9: excellent
  - > 0.9: outstanding
- Equal Error Rate
  - Fingerprint: 0.2%
  - Signature: 2%
  - Voice: 2%
  - Keystroke: 4.7%

(Choi, Hao, Deane & Zhang, 2021)





# Unsupervised Learning

# Cluster Analysis

- Group similar things together
- What is “similar”
  - Similarity is measured by certain distance in a metric space

A **metric space** is an **ordered pair**  $(M, d)$  where  $M$  is a set and  $d$  is a **metric** on  $M$ , i.e., a **function**

$$d: M \times M \rightarrow \mathbb{R}$$

such that for any  $x, y, z \in M$ , the following holds:<sup>[2]</sup>

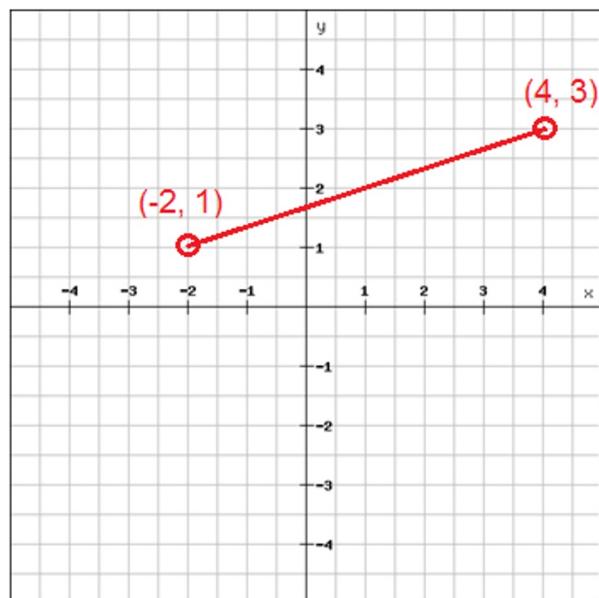
1.  $d(x, y) = 0 \Leftrightarrow x = y$       **identity of indiscernibles**
2.  $d(x, y) = d(y, x)$       **symmetry**
3.  $d(x, z) \leq d(x, y) + d(y, z)$  **subadditivity or triangle inequality**

- Hard clusters vs. fuzzy clusters
- Clusters are not unique, depending on the ways you look at or interpret the data

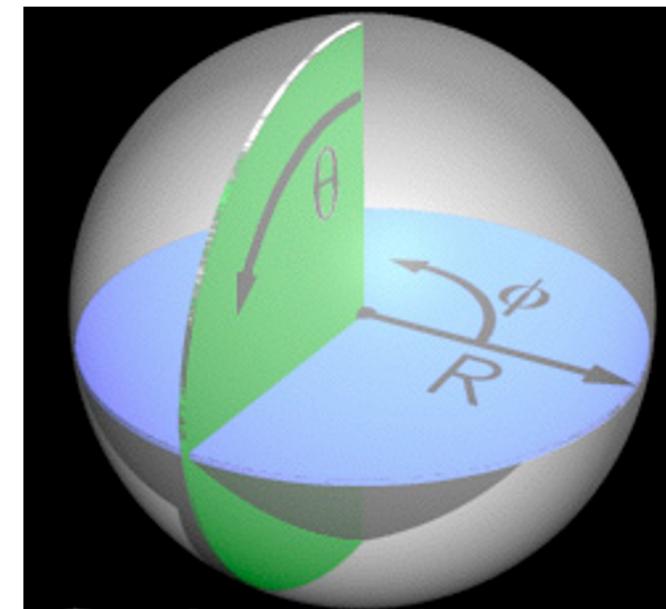


# Example Distance

- Euclidean Space



- Non-Euclidean Space



$$ds^2 = dx^2 + dy^2$$

$$ds^2 = R^2 d\theta^2 + R^2 \sin^2 \theta d\phi^2$$

# Popular Distances in ML

- Euclidean distance
- Minkowski distance  $D(X, Y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$
- Mahalanobis distance  $d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$
- Cosine distance  $1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$
- Edit distance: the minimum number of editing operations to bring two strings the same

*Analyzing Process Data from Game/Scenario-Based Tasks:  
An Edit Distance Approach (Hao, Shu, & von Davier, 2015)*

## Distances in SciPy package

Distance functions between two numeric vectors  $u$  and  $v$ . Computing distances over a large collection of vectors is inefficient for these functions. Use `pdist` for this purpose.

<code>braycurtis(u, v[, w])</code>	Compute the Bray-Curtis distance between two 1-D arrays.
<code>canberra(u, v[, w])</code>	Compute the Canberra distance between two 1-D arrays.
<code>chebyshev(u, v[, w])</code>	Compute the Chebyshev distance.
<code>cityblock(u, v[, w])</code>	Compute the City Block (Manhattan) distance.
<code>correlation(u, v[, w, centered])</code>	Compute the correlation distance between two 1-D arrays.
<code>cosine(u, v[, w])</code>	Compute the Cosine distance between 1-D arrays.
<code>euclidean(u, v[, w])</code>	Computes the Euclidean distance between two 1-D arrays.
<code>jensenshannon(p, q[, base])</code>	Compute the Jensen-Shannon distance (metric) between two 1-D probability arrays.
<code>mahalanobis(u, v, V)</code>	Compute the Mahalanobis distance between two 1-D arrays.
<code>minkowski(u, v[, p, w])</code>	Compute the Minkowski distance between two 1-D arrays.
<code>seuclidean(u, v, V)</code>	Return the standardized Euclidean distance between two 1-D arrays.
<code>squared_euclidean(u, v[, w])</code>	Compute the squared Euclidean distance between two 1-D arrays.
<code>wminkowski(u, v, p, w)</code>	Compute the weighted Minkowski distance between two 1-D arrays.

Distance functions between two boolean vectors (representing sets)  $u$  and  $v$ . As in the case of numerical vectors, `pdist` is more efficient for computing the distances between all pairs.

<code>dice(u, v[, w])</code>	Compute the Dice dissimilarity between two boolean 1-D arrays.
<code>hamming(u, v[, w])</code>	Compute the Hamming distance between two 1-D arrays.
<code>jaccard(u, v[, w])</code>	Compute the Jaccard-Needham dissimilarity between two boolean 1-D arrays.
<code>kulsinski(u, v[, w])</code>	Compute the Kulsinski dissimilarity between two boolean 1-D arrays.
<code>rogerstanimoto(u, v[, w])</code>	Compute the Rogers-Tanimoto dissimilarity between two boolean 1-D arrays.
<code>russellrao(u, v[, w])</code>	Compute the Russell-Rao dissimilarity between two boolean 1-D arrays.
<code>sokalmichener(u, v[, w])</code>	Compute the Sokal-Michener dissimilarity between two boolean 1-D arrays.
<code>sokalsneath(u, v[, w])</code>	Compute the Sokal-Sneath dissimilarity between two boolean 1-D arrays.
<code>yule(u, v[, w])</code>	Compute the Yule dissimilarity between two boolean 1-D arrays.

`hamming` also operates over discrete numerical vectors.



# High Dimensional Space

- High dimensional space is beyond our intuition
- Do not take things for granted in high dimensional space
- The curse of dimensionality

$$\lim_{d \rightarrow \infty} E \left( \frac{\text{dist}_{\max}(d) - \text{dist}_{\min}(d)}{\text{dist}_{\min}(d)} \right) \rightarrow 0$$

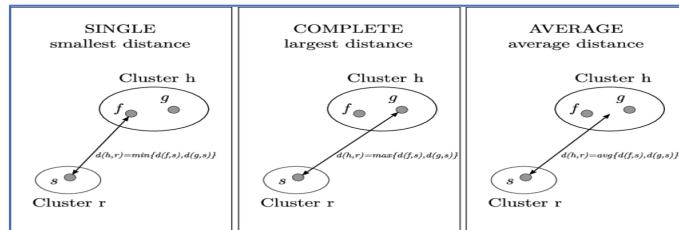
Beyer et al., 1997

- The blessing of dimensionality (Donoho, 2000)
  - Real world objects usually do not live in a super high dimensional space
  - Example: in a high dimensional space, the squared distance of random points to a selected point is, with high probability, close to the average (or median) squared distance.



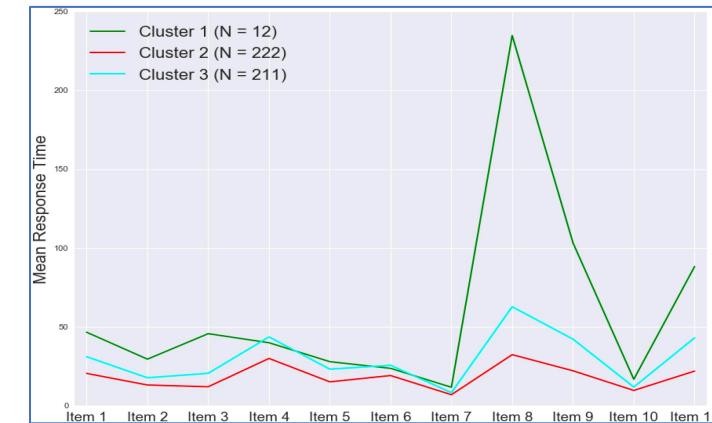
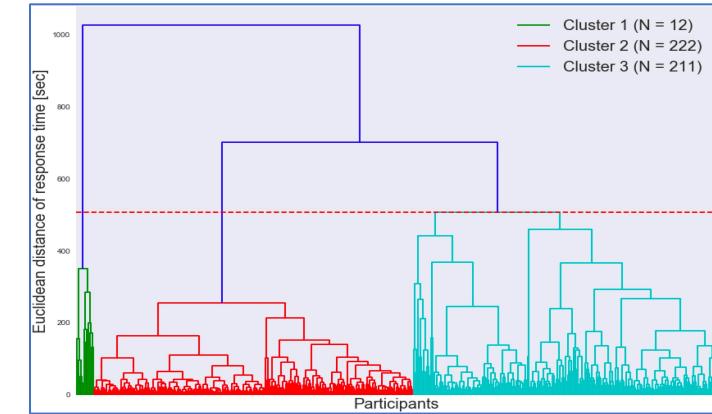
# Hierarchical Clustering

- Hierarchical clustering: method seeking to build a hierarchy of clusters.
- Two strategies:
  - Divisive – top down approach
  - Agglomerative – bottom up approach
- Distance: pairwise distance between the points
- Linkage: determines the distance between sets of points as a function of the pairwise distances between points.



- Ward's minimum variance criterion: minimizes the total within-cluster variance.

[https://en.wikipedia.org/wiki/Hierarchical\\_clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering)



Lee, et al., 2019

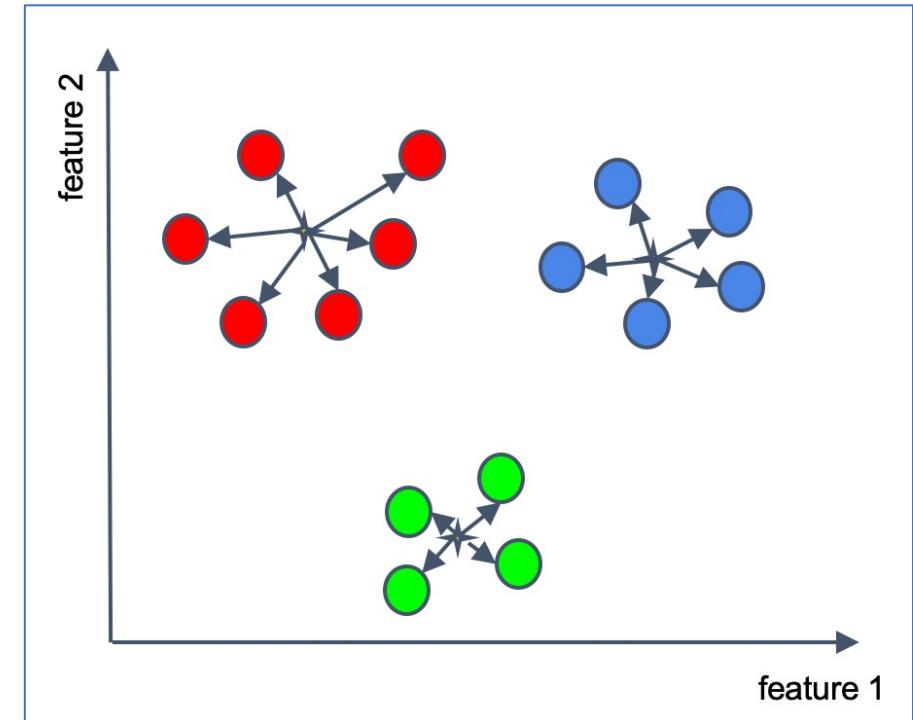


# K-means

- K-means (NP hard computation): divides a set of N samples  $X$  into K disjoint clusters  $C$ , each described by the mean  $\mu_j$  of the samples in the cluster. The means are commonly called the cluster “centroids”
- The K-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion:

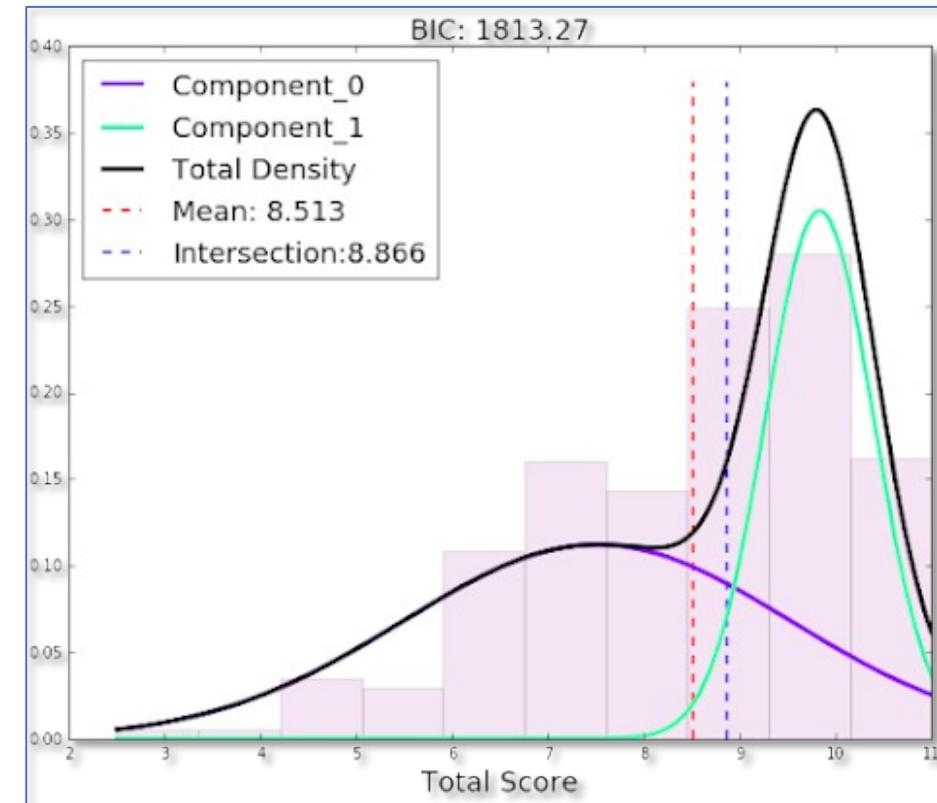
$$\sum_{i=0}^n \min_{\mu_j \in C} (\|x_i - \mu_j\|^2)$$

- Need to specify the number of clusters before running the algorithm



# Model-based clustering

- Gaussian mixture model: the underlying distribution is approximated as a weighted sum of gaussian distributions
- Using BIC/AIC to determine the number of mixtures





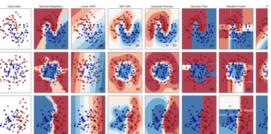
# Scikit-learn package in Python

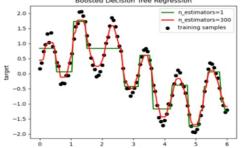
# Scikit-learn

scikit-learn  
Machine Learning in Python

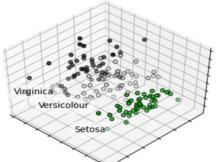
Getting Started Release Highlights for 0.23 GitHub

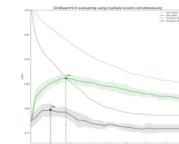
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

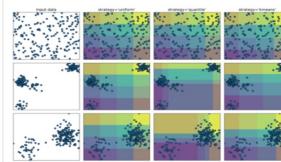
**Classification**  
Identifying which category an object belongs to.  
**Applications:** Spam detection, image recognition.  
**Algorithms:** SVM, nearest neighbors, random forest, and more...  
  
**Examples**

**Regression**  
Predicting a continuous-valued attribute associated with an object.  
**Applications:** Drug response, Stock prices.  
**Algorithms:** SVR, nearest neighbors, random forest, and more...  
  
**Examples**

**Clustering**  
Automatic grouping of similar objects into sets.  
**Applications:** Customer segmentation, Grouping experiment outcomes  
**Algorithms:** k-Means, spectral clustering, mean-shift, and more...  
  
**Examples**

**Dimensionality reduction**  
Reducing the number of random variables to consider.  
**Applications:** Visualization, Increased efficiency  
**Algorithms:** k-Means, feature selection, non-negative matrix factorization, and more...  
  
**Examples**

**Model selection**  
Comparing, validating and choosing parameters and models.  
**Applications:** Improved accuracy via parameter tuning  
**Algorithms:** grid search, cross validation, metrics, and more...  
  
**Examples**

**Preprocessing**  
Feature extraction and normalization.  
**Applications:** Transforming input data such as text for use with machine learning algorithms.  
**Algorithms:** preprocessing, feature extraction, and more...  
  
**Examples**

One of the main open-source ML package in Python

## Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language

Jiangang Hao, Tin Kam Ho

First Published February 20, 2019 | Review Article |  <https://doi.org/10.3102/1076998619832248>

Article information 

Altmetric 5 

### Abstract

Machine learning is a popular topic in data analysis and modeling. Many different machine learning algorithms have been developed and implemented in a variety of programming languages over the past 20 years. In this article, we first provide an overview of machine learning and clarify its difference from statistical inference. Then, we review *Scikit-learn*, a machine learning package in the Python programming language that is widely used in data science. The *Scikit-learn* package includes implementations of a comprehensive list of machine learning methods under unified data and modeling procedure conventions, making it a convenient toolkit for educational and behavior statisticians.

### Keywords

*machine learning, Python, Scikit-learn*

```
model = classifier(hyperparameters = something)
model.fit(X_train, y_train)
y_test = model.predict(X_test)
```



# Scikit-learn PCA Example

- Iris Data Set: 150 instances x 4 features (Fisher, 1950)
  - sepal length [cm]
  - sepal width [cm]
  - petal length [cm]
  - petal width [cm]
  - Class: Iris Setosa, Iris Versicolour, Iris Virginica



Notebook demo

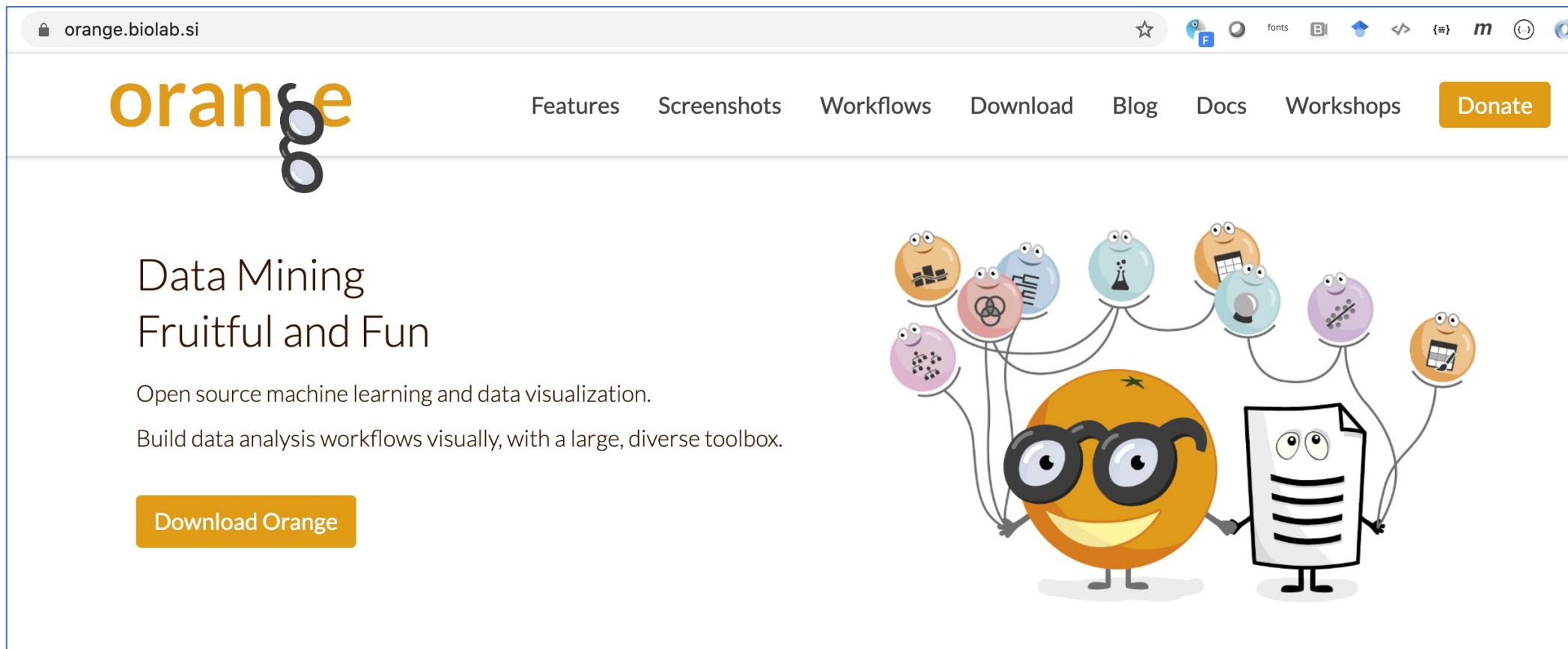
# Assignments

- Read the attached book chapter on supervised learning
- Practice the examples in the demo notebook



# Backup: Orange

# Interactive Machine Learning



The screenshot shows the homepage of the Orange Data Mining website. At the top, there is a navigation bar with links for Features, Screenshots, Workflows, Download, Blog, Docs, Workshops, and a yellow 'Donate' button. The main title 'orange' is displayed in a large, stylized font where the letter 'o' has a magnifying glass icon integrated into it. Below the title, the text 'Data Mining' and 'Fruitful and Fun' is shown. A subtext explains that Orange is an 'Open source machine learning and data visualization' tool that allows users to 'Build data analysis workflows visually, with a large, diverse toolbox.' To the right of the text is a cartoon illustration of a large orange character with glasses and a smile, holding hands with a white document character. They are surrounded by smaller circular characters, each representing a different machine learning or data mining component like a neural network, a database, or a plot.

orange.biolab.si

Features Screenshots Workflows Download Blog Docs Workshops Donate

orange

Data Mining  
Fruitful and Fun

Open source machine learning and data visualization.  
Build data analysis workflows visually, with a large, diverse toolbox.

Download Orange

