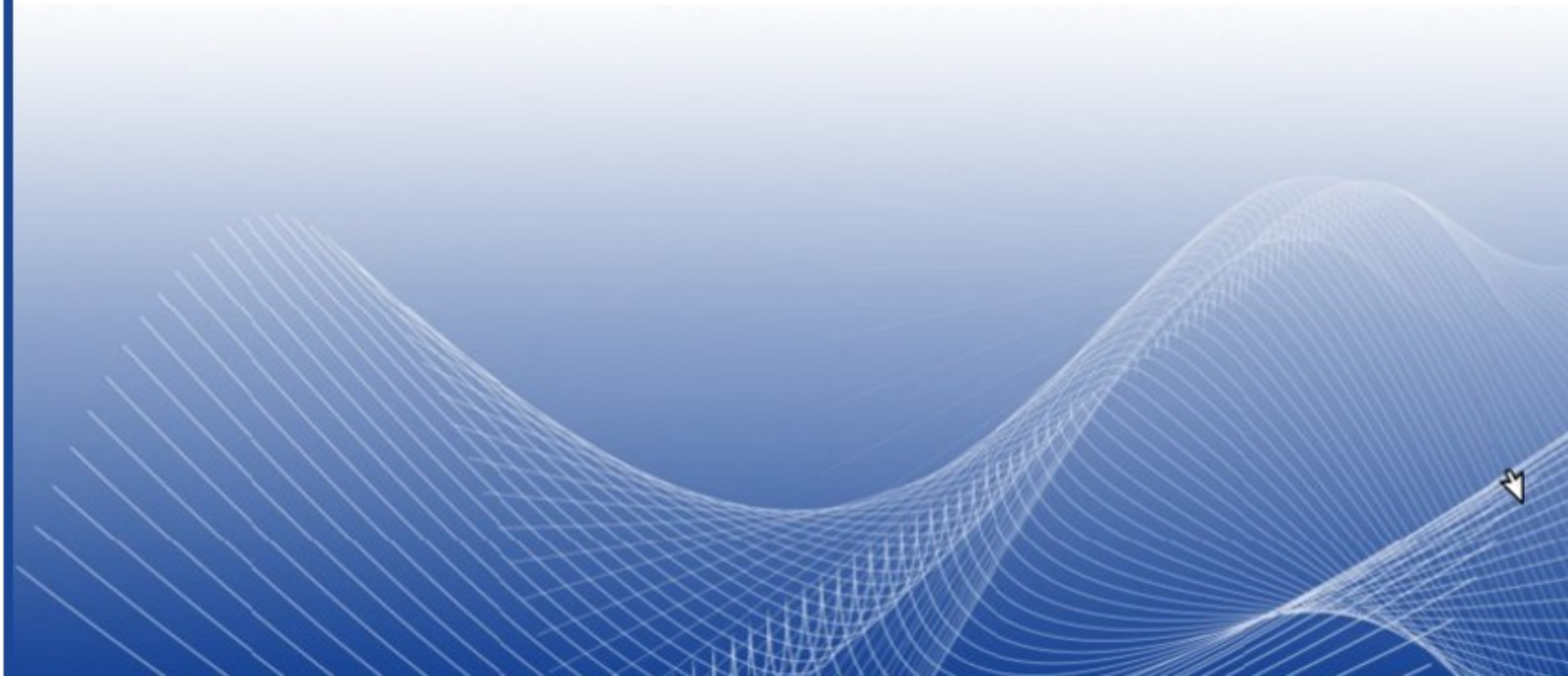
Network (Entities) Profiling

for classification and anomaly detection



universidade de aveiro

deti.ua.pt

Data Inputs

é possível introduzir dados em bruto, mas isso aumenta a complexidade do algoritmo de aprendizagem automática.

- Raw data inputs are possible, however its increases the complexity of the machine learning algorithm.
  Resultados piores, tempos de cálculo/resposta mais longos.
  - Worse results, longer calculation/response times.
- Input data should be the result of raw data processing (complexity reduction).
  - Observation features.
  - Statistical metrics, statistical functions, PCA, scale analysis metrics/descriptors, …
- Inputs should be normalized.
  - Usually

Variable Reduction

- An event/entity is many times described by multiple descriptors/metrics.
  - e.g., mean, variance, maximum, skewness, percentile x%, etc...
  - a.k.a. features.

$$e_i = [y_1, y_2, \ldots, y_m]$$

- The reduction of variables is mandatory to simplify classification.
- **Principal Components Analysis (PCA)**
  - Uses a transformation to convert a set of possibly correlated features into a set of values of uncorrelated variables called principal components.
  - The principal components of an event will be a linear combination of the that event features.

$$t_i = e_i W, W = [w_{ij}]_{i,j=1,\ldots,m}$$

  - The number of principal components is less than or equal to the number of original features.
    - Defined in such a way that the first principal component has the largest possible variance, and the m^{th} (last) component has the smallest variation.
    - The first *n* components can be chosen to describe the event.
    - *W* is a (*m* x *n*) matrix.

EX:
0.2 -> uploads
0.7 -> downloads
0.1 -> nº de pacotes

com estas 3 variaveis é suficiente para resolver a variabilidade dos resultados finais

Tiramos a correlação das variaveis
Depois vamos ao vetor proprio, vamos aos maiores e ficamos com esses, deitando vetores proprios para fora (nós assim perdemos info)

neste caso reduzimos de 3 para 2 ficando com DL e UL tirando o Nº de pacotes fora

o objetivo é reduzir de 100 para 20, sabemos quando parar quando o desempenho diminuir

Na pratica, vamos tirando variaveis e testando para ver se dá mais ou menos a mesma coisa, se der tiramos fora

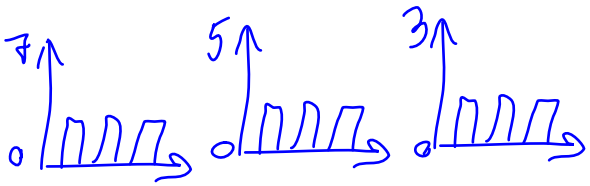
O PCA transforma os dados, meto lá os dados e indico quantos é que quero.

Data Normalization/Scaling

Se os dados tiverem grandezas mt diferentes
temos de ter cuidado com a normaliz

É preciso cuidado com a normalização, por exemplo

Solução:
Normalização logaritmica
por exemplo: 0-3; 0-5; 0-7



- Methods:
  - By maximum absolute value,
  - By min/max, scaling each feature to a given range,
  - Standard, removes the mean and scales to unit variance.
  - …
- Mandatory when variables/features have different orders of magnitude.
- Removes data bias from quantity, allow to focus on variable and time correlations on data.
  - e.g., YouTube traffic pattern correlation with video definition must be removed.

Classification vs. Anomaly Detection

- Classification à priori conhecemos as coisas, sabemos o que tamos à procura
  - Requires knowledge (historic data) on all patterns/classes.
  - Does not cope with pattern evolution and appearance of new patterns/classes.
- Anomaly Detection dão piores resultados, pq n sabemos o que tamos à procura
até podemos conhecer os vetores do atacante, as ferramentas, mas basta ele mudar o sleep, já n funciona o nosso detetor
  - Requires only knowledge (historic data) known of normal patterns/classes.
  - Does not require knowledge (historic data) anomalous patterns/classes.
  - Identify all significantly different patterns as anomalous.
  - Allows to identify never seen anomalies (zero-day detection).
  - May identify as anomalous licit patters that are evolving

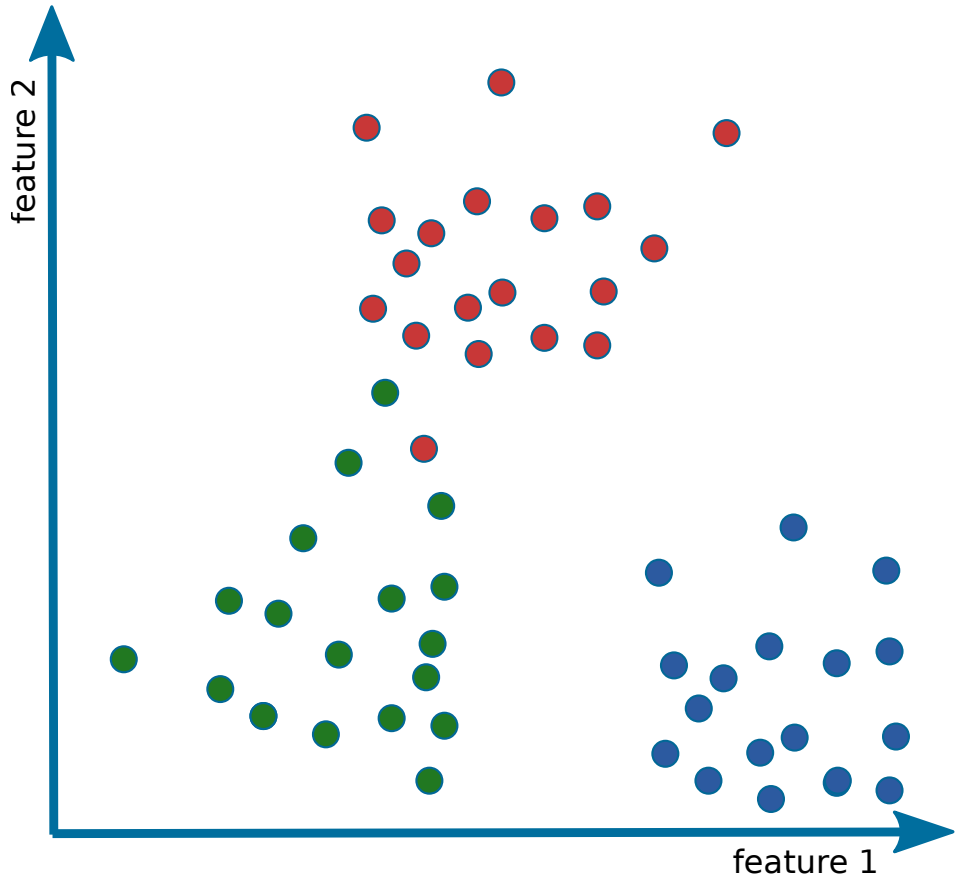
N ter labels à priori
Um exemplo de um label é quando temos uma sequencia de movimentos normais que termina com algo esperado, exemplo do Super Mario.

À posteriori, não sabemos distinguir "se a anomalia é boa ou se é má"

Exemplo, dos chineses a despultar os alarmes todos por terem varias vpns a ligarem quando se ligam à internet da ua, são deteções anormais mas não são más

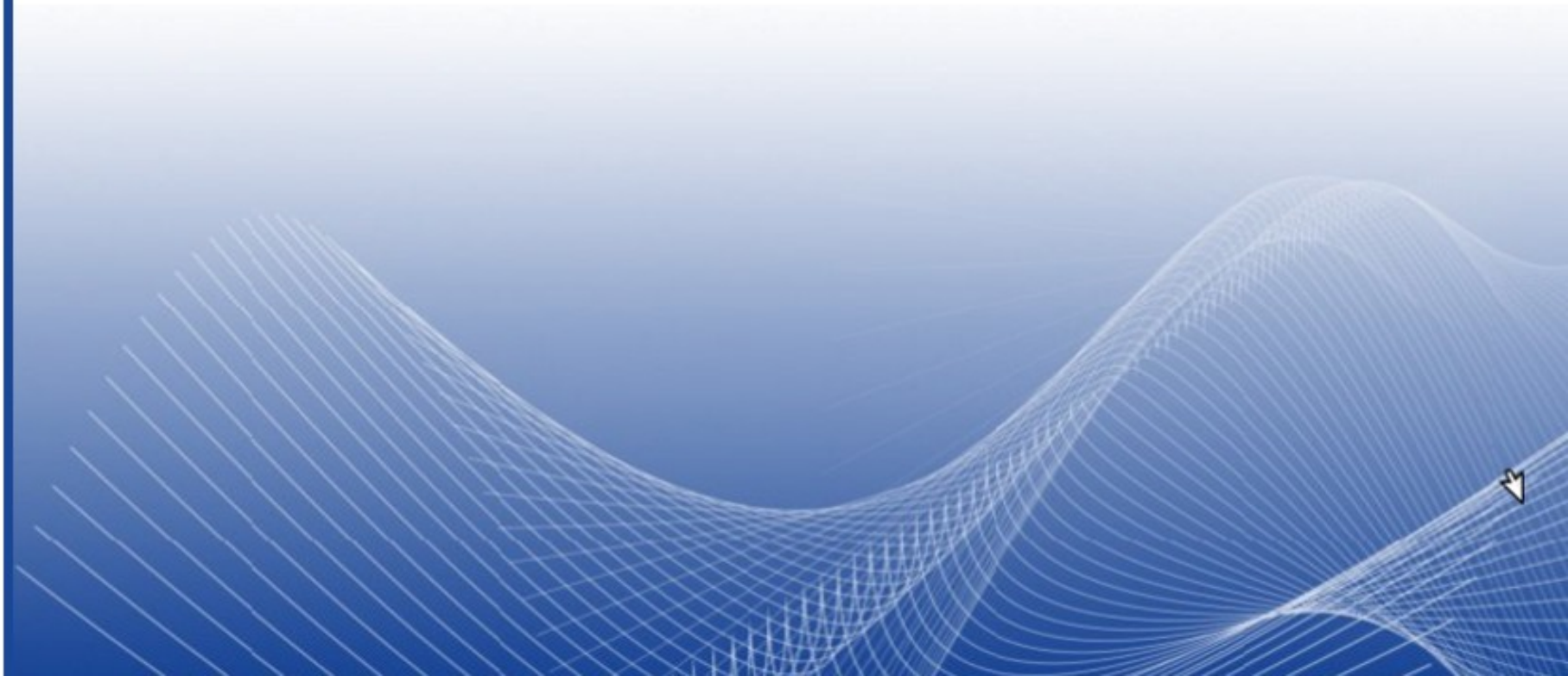
Profile as a N-Dimensional Euclidean Universe

- Each set of N features (reduced or not) in each observation can be seen as a point a N-dimensional Euclidean universe.
- Each point can be:
  - Pre-classified to identify know behaviors/activities.
  - Classified as an belong to a specific group
    - Short Euclidean distance from the known group points.
    - Short Euclidean distance from group points previously “grouped” (cluster).
  - Classified as an anomaly.
    - Large Euclidean distance from the other points.



Decision by Statistical Patterns

for differentiation, classification, and anomaly detection



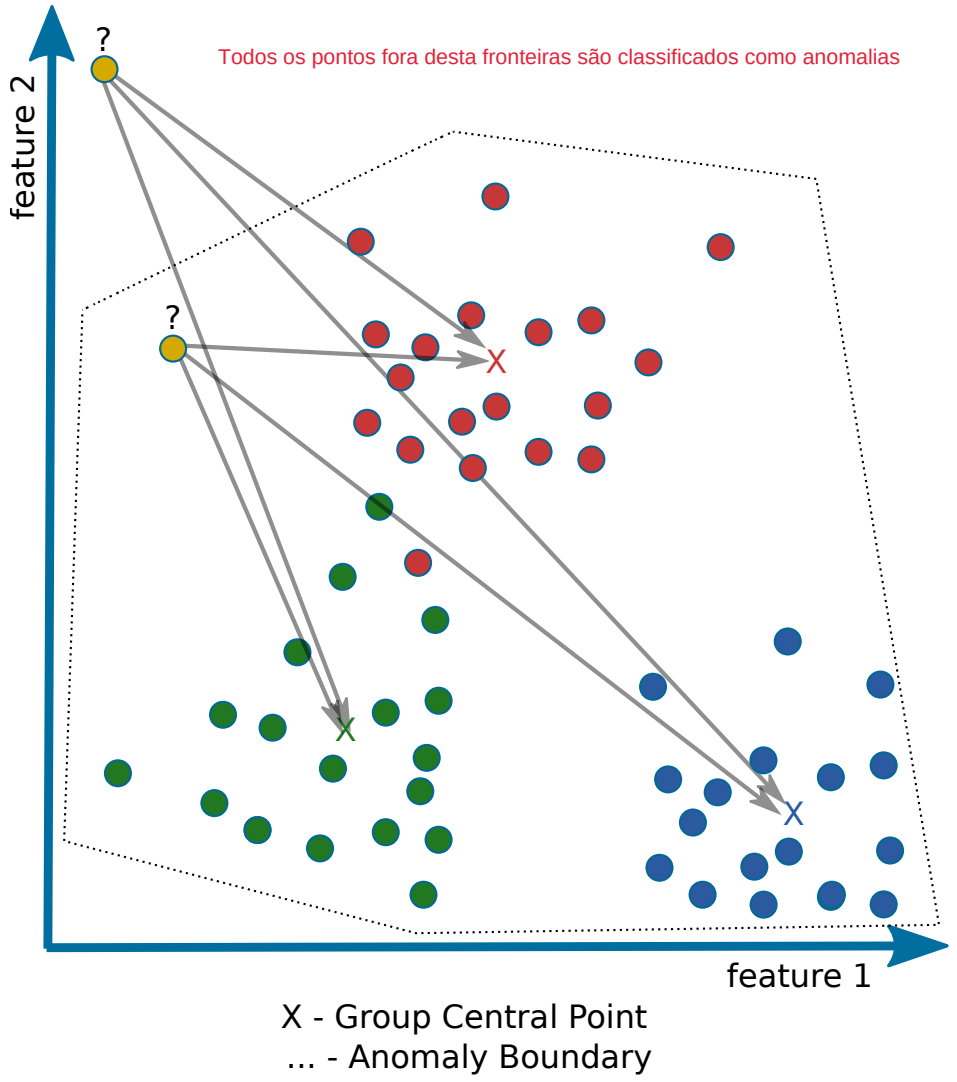
universidade de aveiro

deti.ua.pt

Distances to Central Point(s)

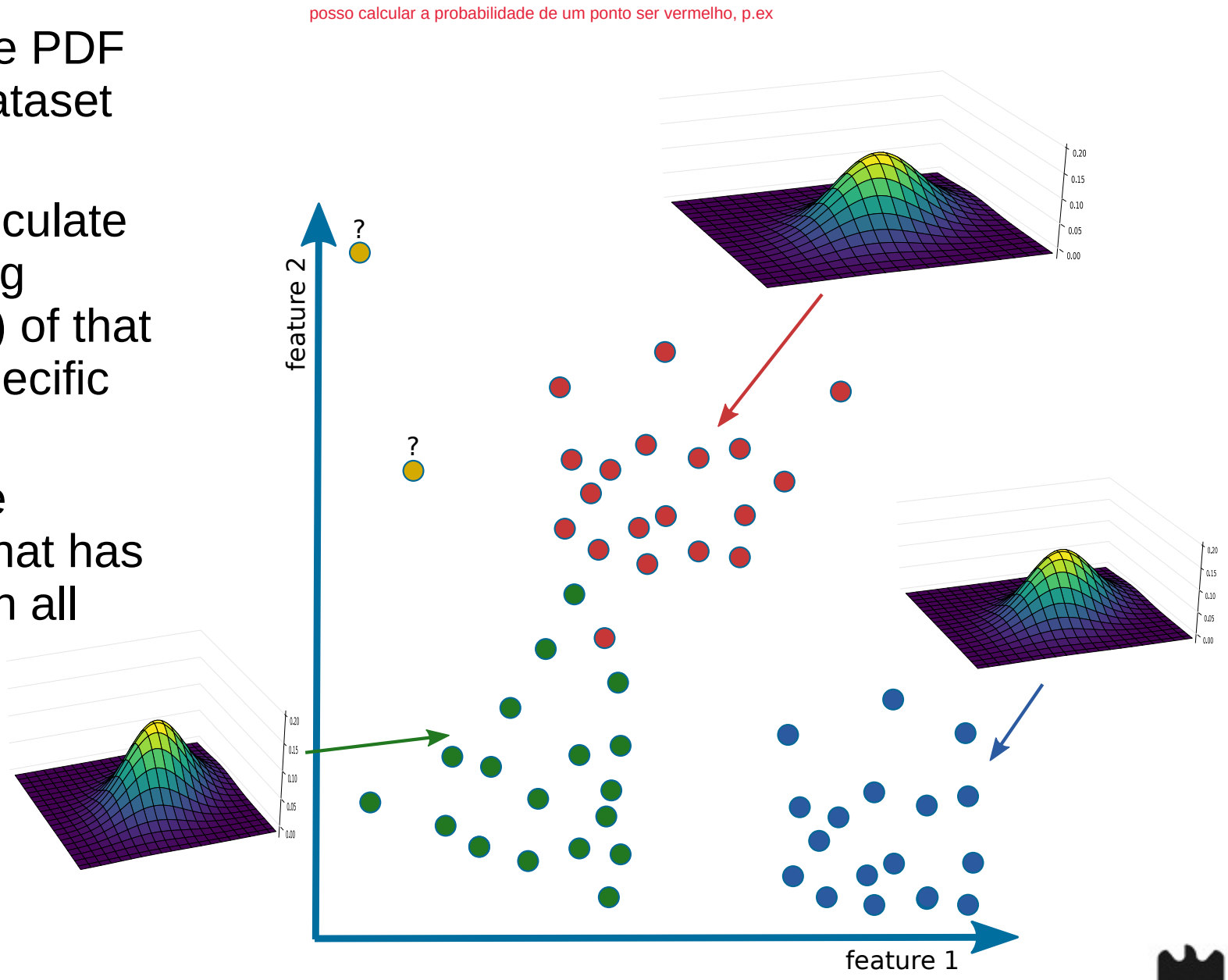
possivel até por com probabilidades

- Group dataset points
  - Use a single group (to detect anomalies),
  - By known classification,
  - By clustering algorithms.
- Find central point of each group.
- For each new dataset point:
  - Calculate Euclidean distances to each group central point,
  - Use distances to classify:
    - Shortest distance to group,
    - Probabilistic result based on the relative distances,
      - Ex: d1=10, d2=20, d3=30 → Group1 prob.=10/(10+20+30)=16.6%
    - Define as anomaly if distance(s) above predefined threshold.



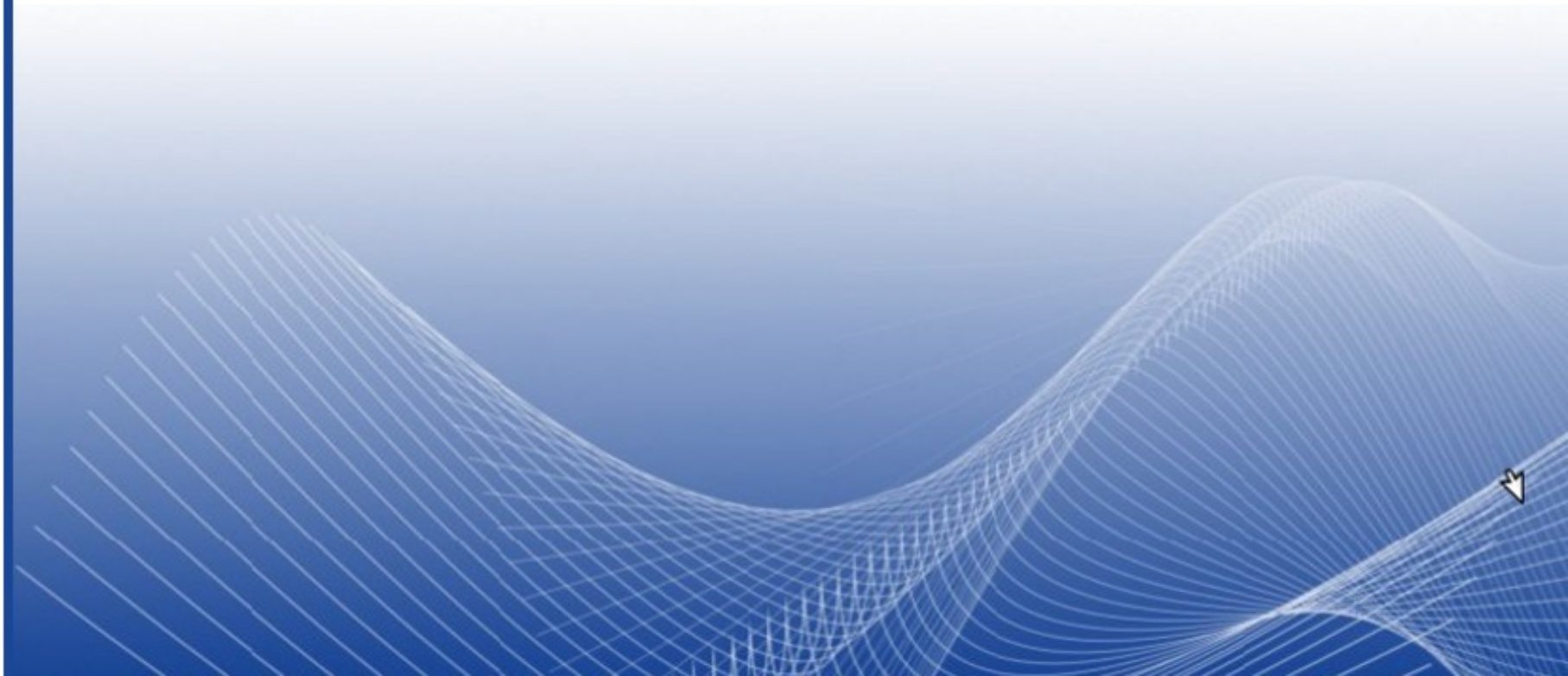
N-Dimensional Distributions

- Infer the multivariate PDF of each group of dataset points.
- For a new point, calculate the probability (using respective the PDF) of that point belong to a specific group.
- An anomaly may be defined as a point that has lower probabilities in all groups.



Decision by Machine Learning

for differentiation, classification, and anomaly detection



universidade de aveiro

deti.ua.pt

Categories

- Supervised learning
  - Inputs and outputs are given.
    - Outputs may be classification labels or system quantifiers.
  - Creates a general mapping rule between input and output.
- Unsupervised learning
  - Only inputs are given.
    - Algorithm must by structure in input data.
  - Post-classification based on known inputs and found data structure may be done to create a classifier.
- Reinforcement learning
  - Inputs are given, and “quality” of outputs is defined in terms of reward and penalization (cost functions) relative to the problem goal.

Se conseguirmos classificar os grupos, já é supervisionado

é mais dificil de ter supervised

Se sbemos identificar yt, netflix e spotify
detetamos sabendo logo à partida os grupos

Dar os dados sem labels, ou seja, dar sem conhecer nada
pedimos para dividir em grupos, pode ser em clustered (imaginado que não conhecemos os grupos)

se um milhão vi mil e era tudo netflix posso generalizar que um grupo é tudo quem vê netflix

depois pode acontecer, num grupo separar em mais grupos com base em dados apenas

ciclo de aprendizagem (erra, aprende, tenta de novo)

Approaches

Clusterins é unsupervised, que depois se pode tornar supervised

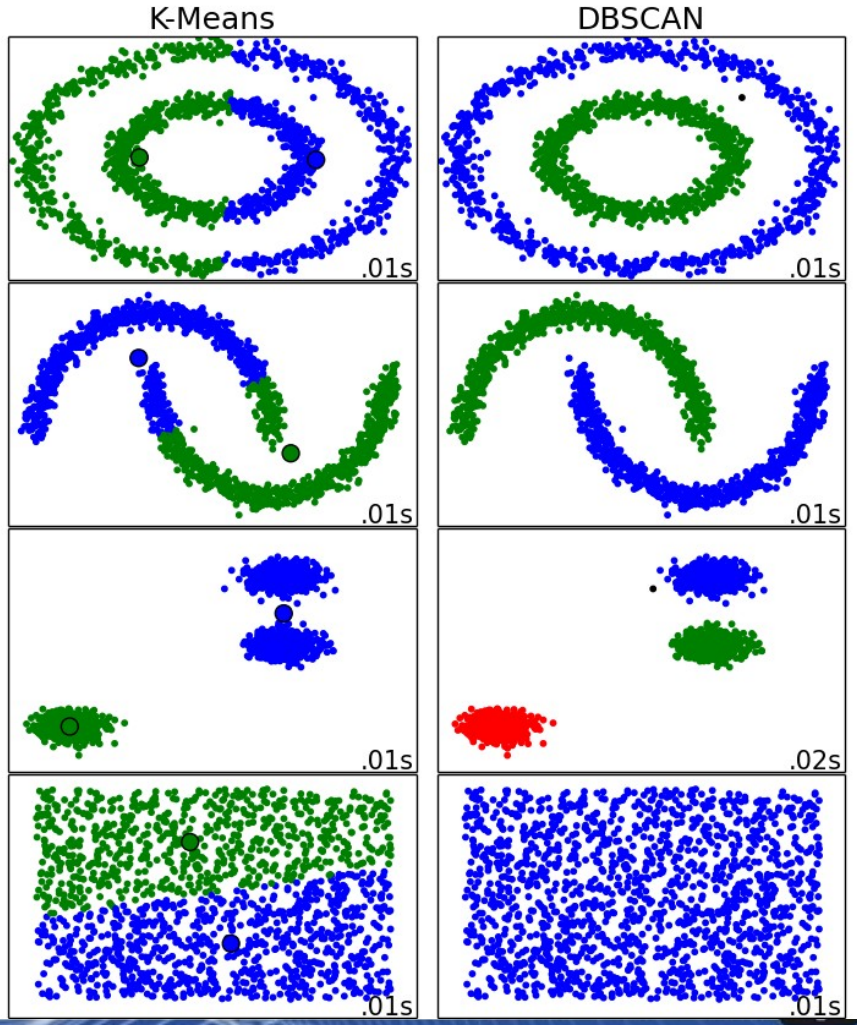
- Clustering

Pedimos para dividir em N grupos sem saber o N, dizzendo se tiver a uma distancia superior a 'x' guarda no grupo X, senao guarda no grupo Y etc

- Support vector machines
- Artificial neural networks
  - Composed of one input and one output layer, and at most one hidden layer in between.
- Deep learning
  - ANN with more than three layers (including input and output).
    - More than one hidden layer.
- Other
  - Bayesian networks
  - Decision tree learning
  - Genetic algorithms
  - ... random forest

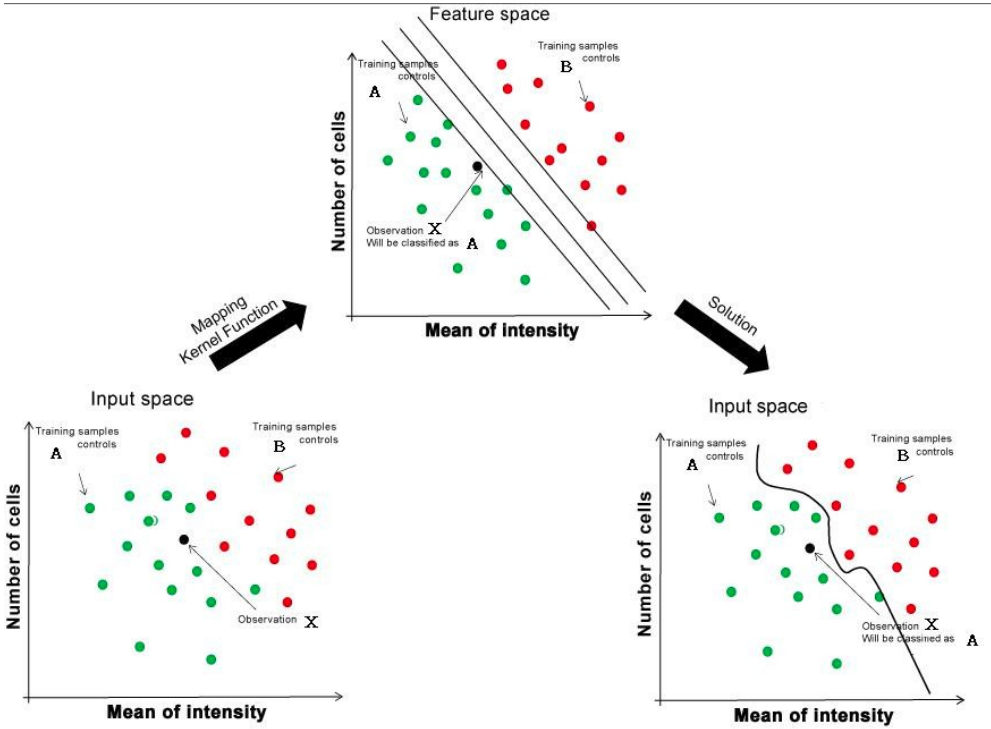
Classification / Clustering

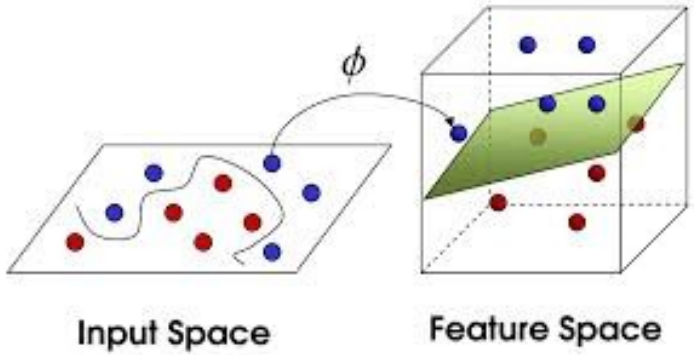
- Clustering is the process of grouping (classifying) a set of objects in such a way that objects in the same group (cluster) are more “similar” to each other than to those in other clusters.
- Algorithms:
  - K-Means
    - Requires the a priori knowledge of the number of clusters.
    - Uses the distances between points as metric.
  - DBSCAN
    - Requires the a priori definition of the neighborhood size.
    - Uses the distances between nearest points as metric.
  - Others…



Support Vector Machines (SVM)

- Classification defined by a separating hyper-plane-
- Optimal hyper-plane for linearly separable patterns.
- Kernel functions allow the separation of patterns that are not linearly separable by transformations of original data.
- Solutions found using a minimization problem.

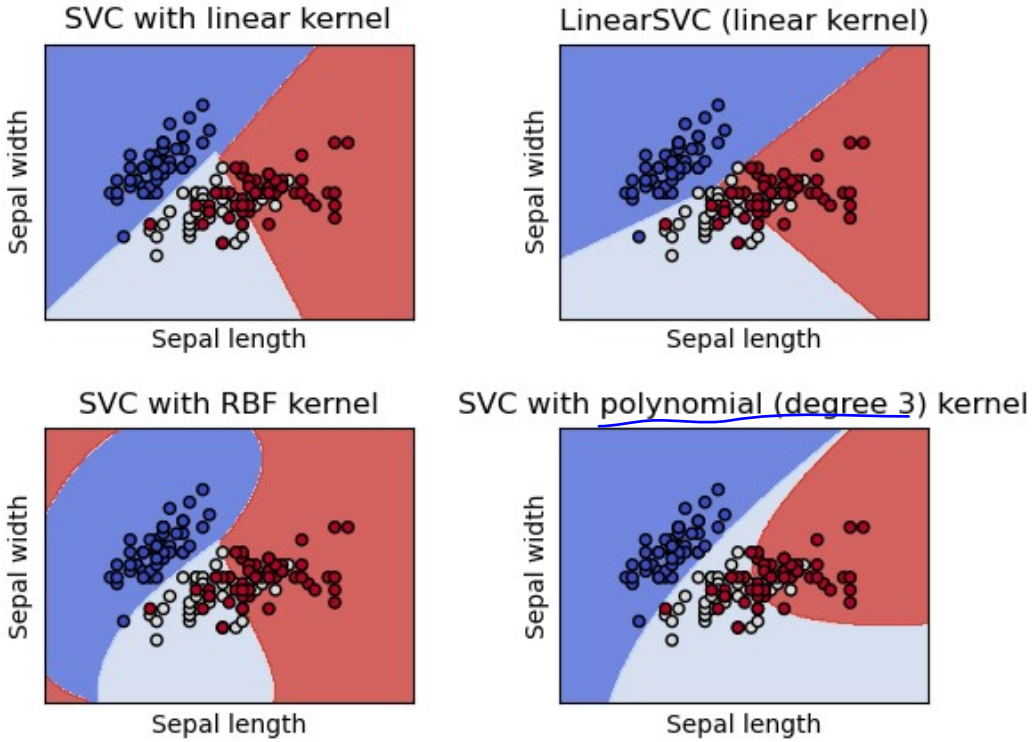




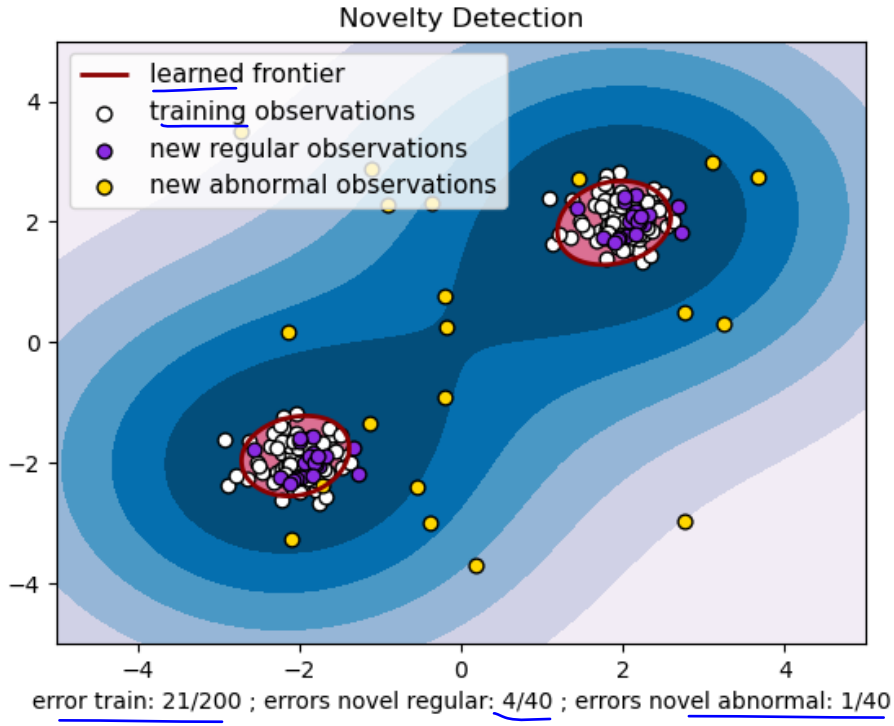
One-Class SVM vs. N-Class SVM

polinomio linear é uma reta
polinomia de 2 dim faz uma parabola
polinomia de 3 dim faz varias parabola

- N-Class SVM
  - Infers boundaries between each class.
- One-Class SVM
  - Infers “a boundary” that contains all known normal/licit traffic.



Os dados variam muito de observação para observação
Por isso é preciso ter cuidado ao minimizar a escala do erro



abnormal = normal mas fora do normal

Decision Trees

arvore de decisao -> Conj de parametros que ramifica em funçao dos parametros

cada ramificação é onde temos os dados --------> separamos : estes ramos sao anomalias/ estes ramos sao normais

- Data partitions by branching decisions based on features values.
- Decision based on:
  - Location of an observation on the decision tree;
  - Location of an observation on multiple decision trees (forest);
  - Number of partitions/branches required to isolate an observation.
- Variants:
  - Tree Regressor
    - Classification based on data partitions (over branches).
  - Isolation Forests
    - Detects anomalies based on the low number of branches (data partitions) required to isolate an observation.
  - Random Forests Testar varias coisas e ver qual a percentagem de coisas que funfam
    - Uses multiple tree classifiers on various random sub-samples of the dataset.
    - Averages the results.

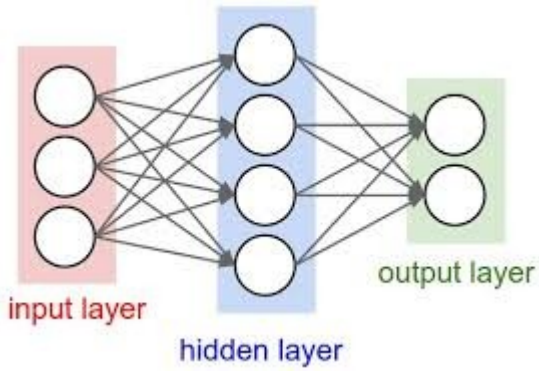
Artificial Neural Networks

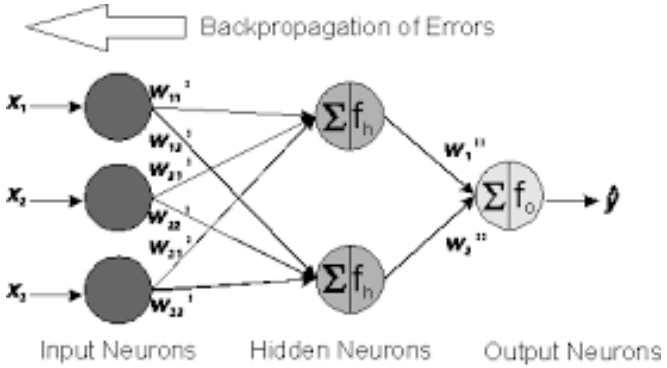
Ex: da key do euromilhoes
É preciso ter atenção ao histórico,
a rede neuronal (over feeding) replica o historico o que n funciona no euromilhoes pq n é baseado no historico

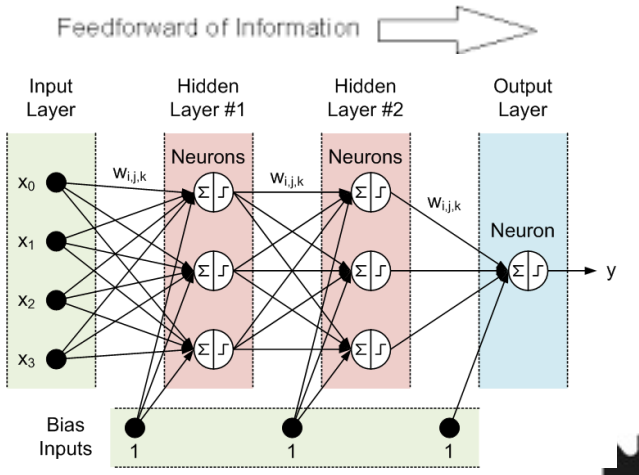
- Composed by input and output layers, and an optimal hidden layers
  - More than one hidden layer, becomes a deep learning NN.
- Hidden and output layers, perform a weighted sum of the values outputted by the nodes of the previous layer and applies an activation function.
  - Activation functions: linear, tanh, arctan, etc…
  - Weights define the NN, and must be inferred by a training algorithm.
    - Each node-node connection have a different weight.
- Training algorithms adjust connection weights to minimize the error between inputs and training outputs.

Variaveis com pesos e somatorios com tangentes hiperbolicas lá no meio que passam para outra camada que depois faz o mesmo até ter só uma saida

  - Back propagation of error.
  - Levenberg-Marquardt algorithm, Newton and quasi-Newton methods, Gradient descent, and Conjugate gradient.
- Some nodes/layers may have bias inputs to activate/deactivate and/or offset node outputs.





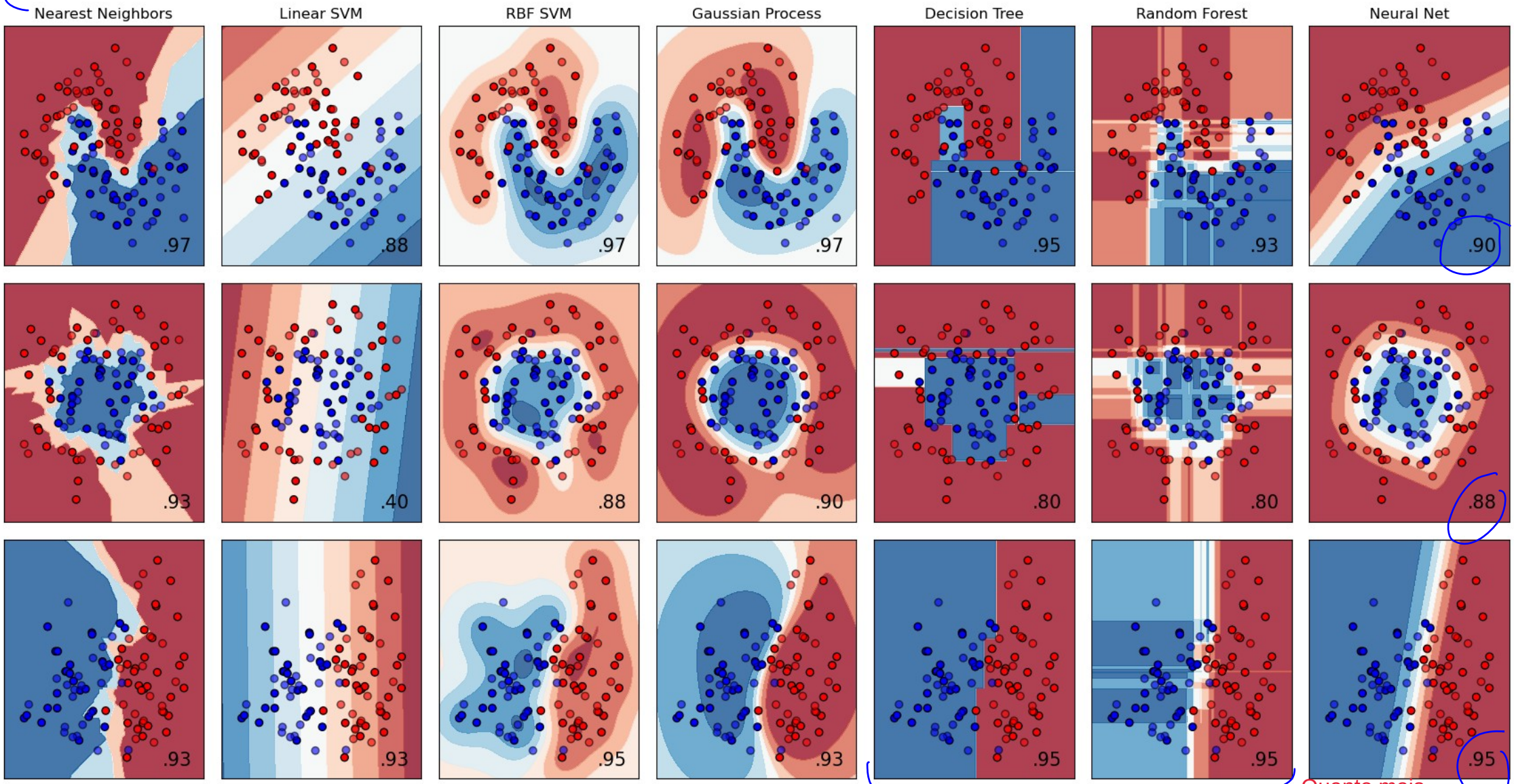


Overview

Redes neuronais funcionam melhor funcionam para minimizar os erros

algoritmo baseado em clustered de distancias

cria fronteiras para os azuis e vermelhos

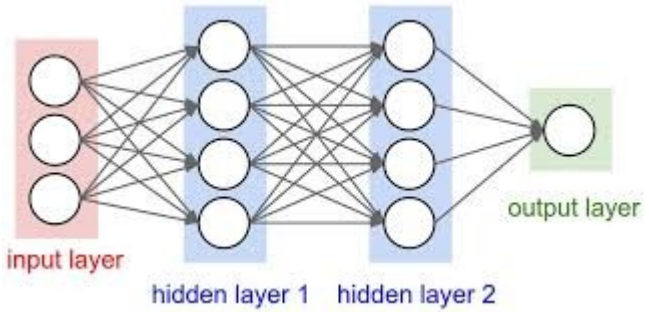


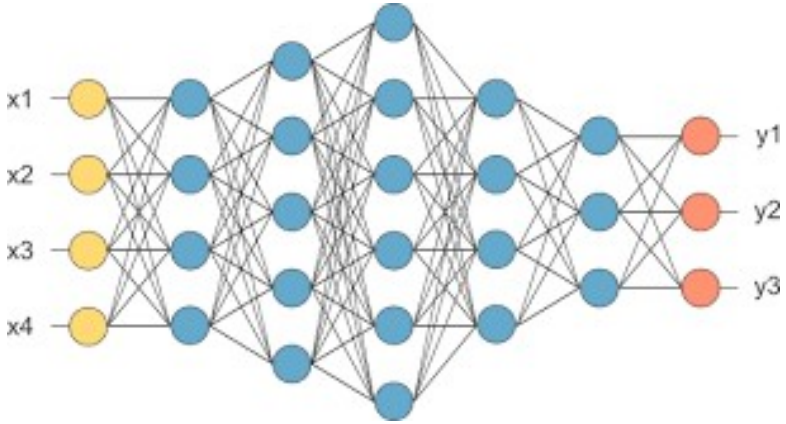
faz retas (n sao as ideais)

Quanto mais proximo de 1 melhorr

Deep Learning

- Supervised learning algorithms
  - Logistic Regression.
  - Multilayer perceptron.
  - Deep Convolutional Network.
- Unsupervised and semi-supervised learning algorithms
  - Auto Encoders
  - Denoising Autoencoders
  - Stacked Denoising Auto-Encoders
  - Restricted Boltzmann Machines
  - Deep Belief Networks





Abordou pouco este slide

Ensemble (1)

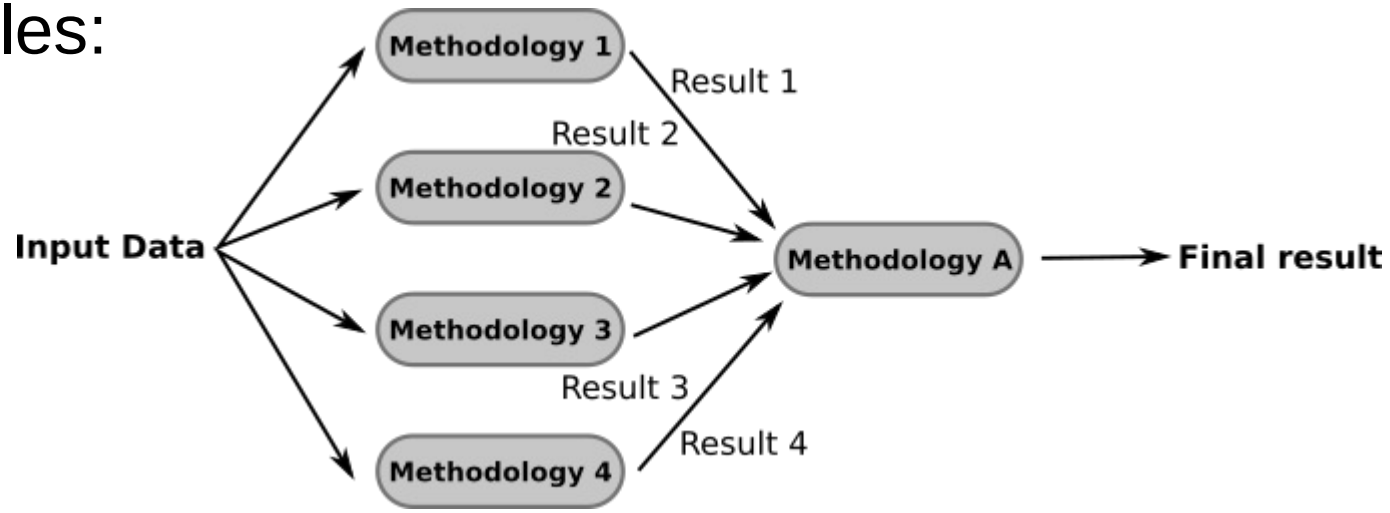
s

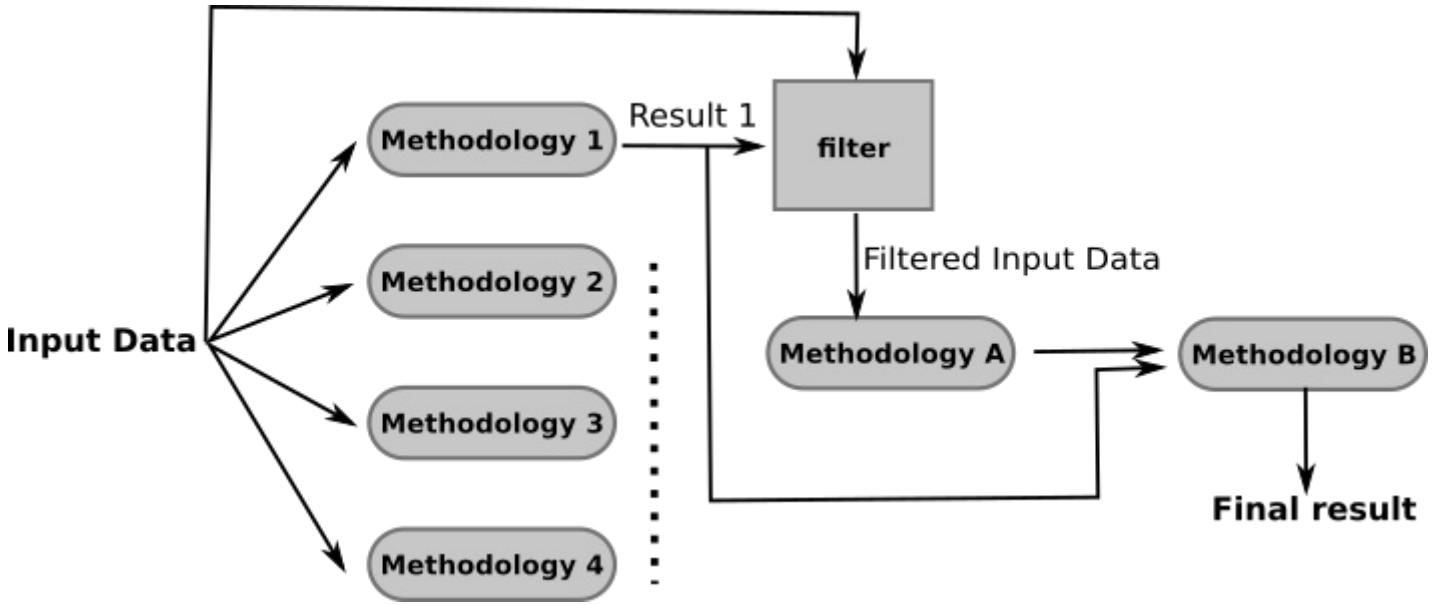
- Ensemble methods use multiple learning methodologies to obtain better that the individual methods.
- Methods:
  - Bayes optimal classifier
    - Final decision based on the probabilities given by each methodology
  - Bagging
    - Final decision based on the results given by each methodology with equal weight.
    - Input data may differ between methodologies
      - Aims to decrease final result variance.
  - Boosting
    - Final decision based on different methodologies applied in sequence (to correct wrong classifications by the previous methodology).
    - Previous results may be used to filter input data given to next level classification methodologies.

dividir para conquistar
Separar os dados complexos e depois aplicar diferentes metodologias

Ensemble (2)

- Examples:

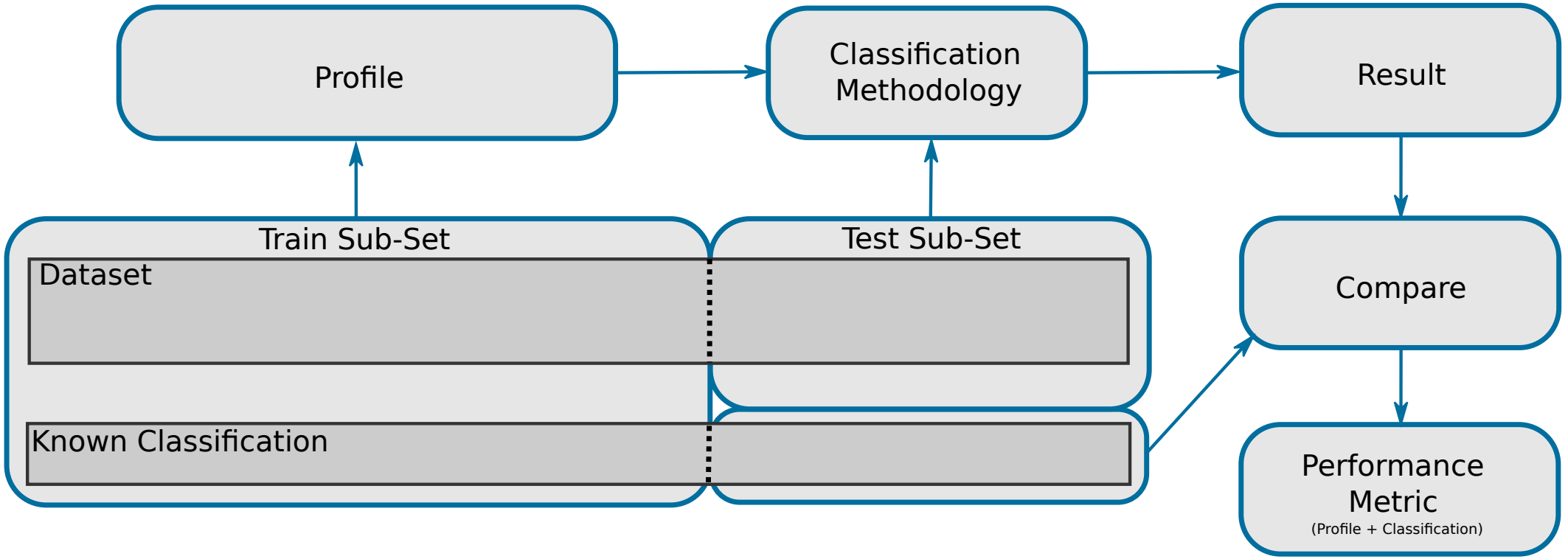




Performance Evaluation

universidade de aveiro
deti.ua.pt

Evaluation Process



Para testar é preciso

ter classificações
e os parametros

preferem um sistema de segurança c/ mts falsos alarmes (deteta td)
ou um sistema com uma taxa baixa de deteção (10%)?

Resposta é "depende", é subjetivo

Metrics

- True Positive (TP) - Correctly predicted positive
  - True Negative (TN) - Correctly predicted negative
- False Positive (FP) - Wrongly predicted as positive
  - False Negative (FN) - Wrongly predicted as negative
- Metrics
  - Accuracy=(TP+TN)/(TP+TN+FP+FN)
  - Precision=TP/(TP+FP)
  - Recall=TP/(TP+FN)
  - F1-Score== 2*(Recall * Precision) / (Recall + Precision)

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Confusion Matrix