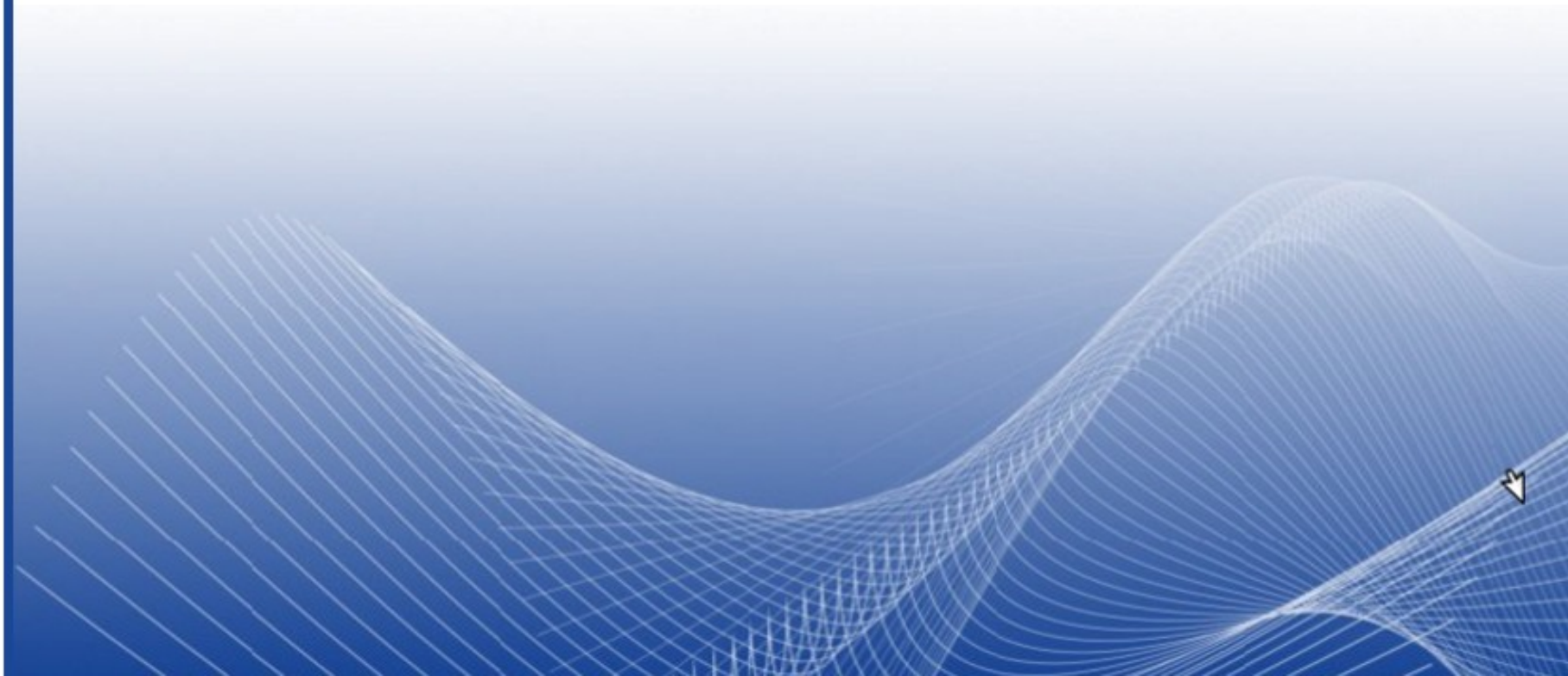
Network (Entities) Profiling and Classification

for differentiation, and anomaly detection

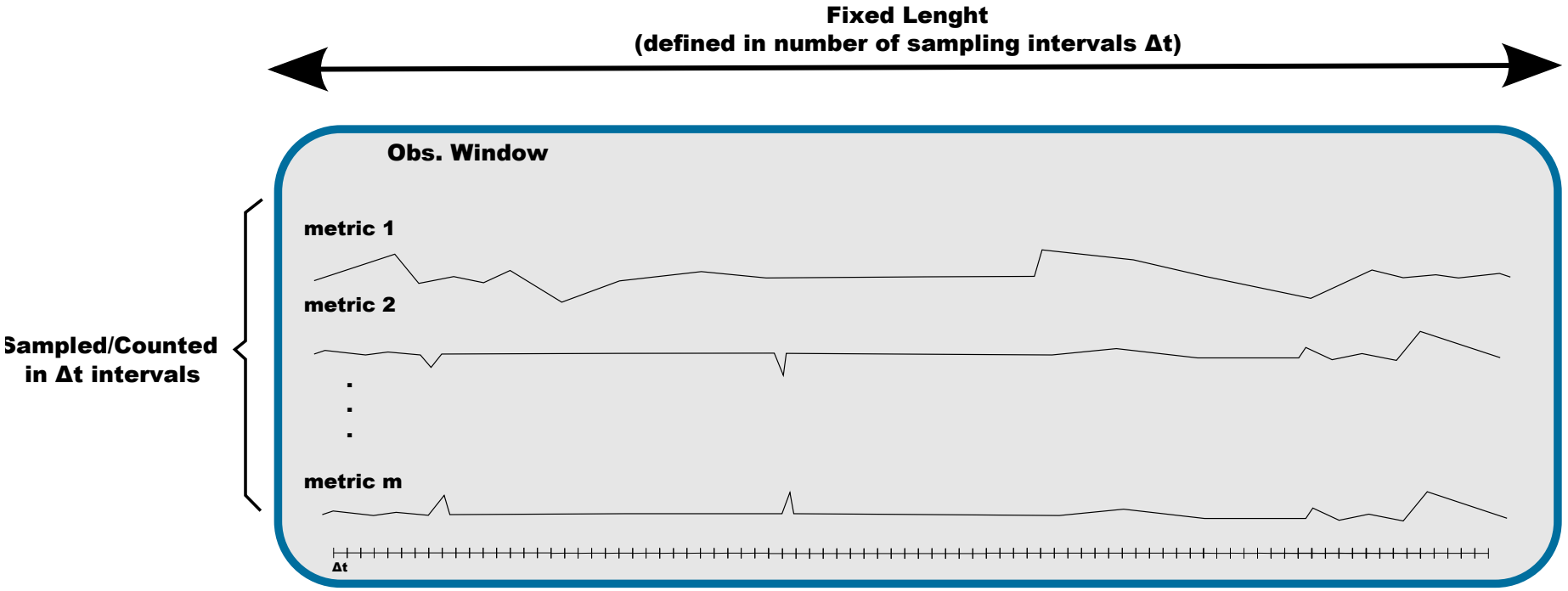


universidade de aveiro

deti.ua.pt

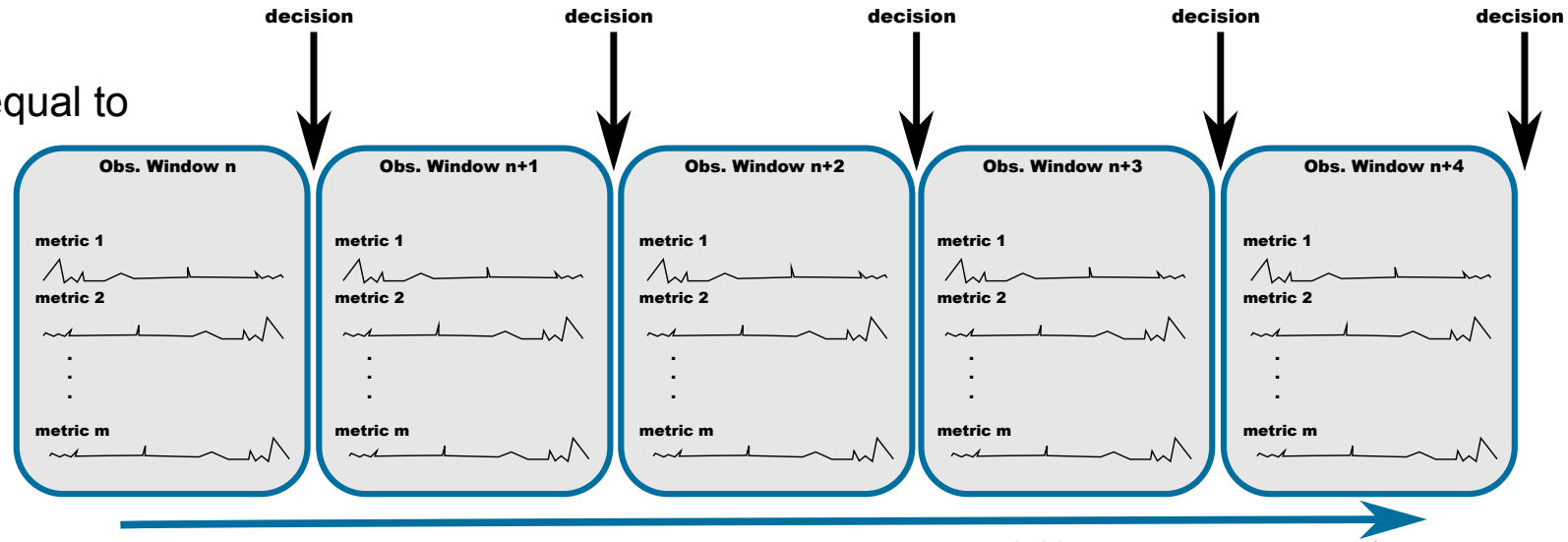
Observation Window (1)

- An observation is constructed based on multiple sampling/counting metrics.
- Sampling/counting metrics should quantify activity events:
  - Start/End of activity.
    - Traffic Flows, Calls, Service usage, etc…
  - Amount of activity.
    - Traffic per sampling interval, activity duration, actions per sampling interval, etc…
  - Activity targets
    - IP addresses contacted, UCP/TCP ports used, services user IDs, points of access, etc…

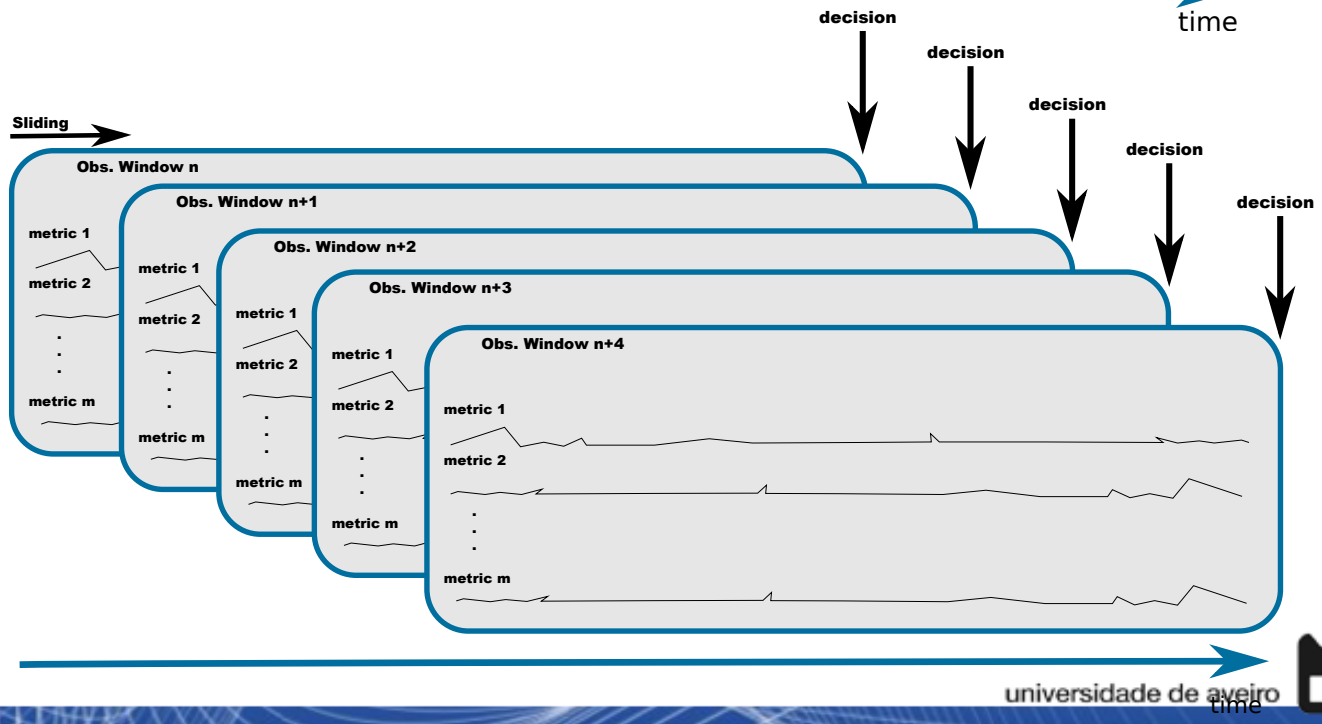


Observation Window (2)

- Sequential
  - Decision interval is equal to window size.

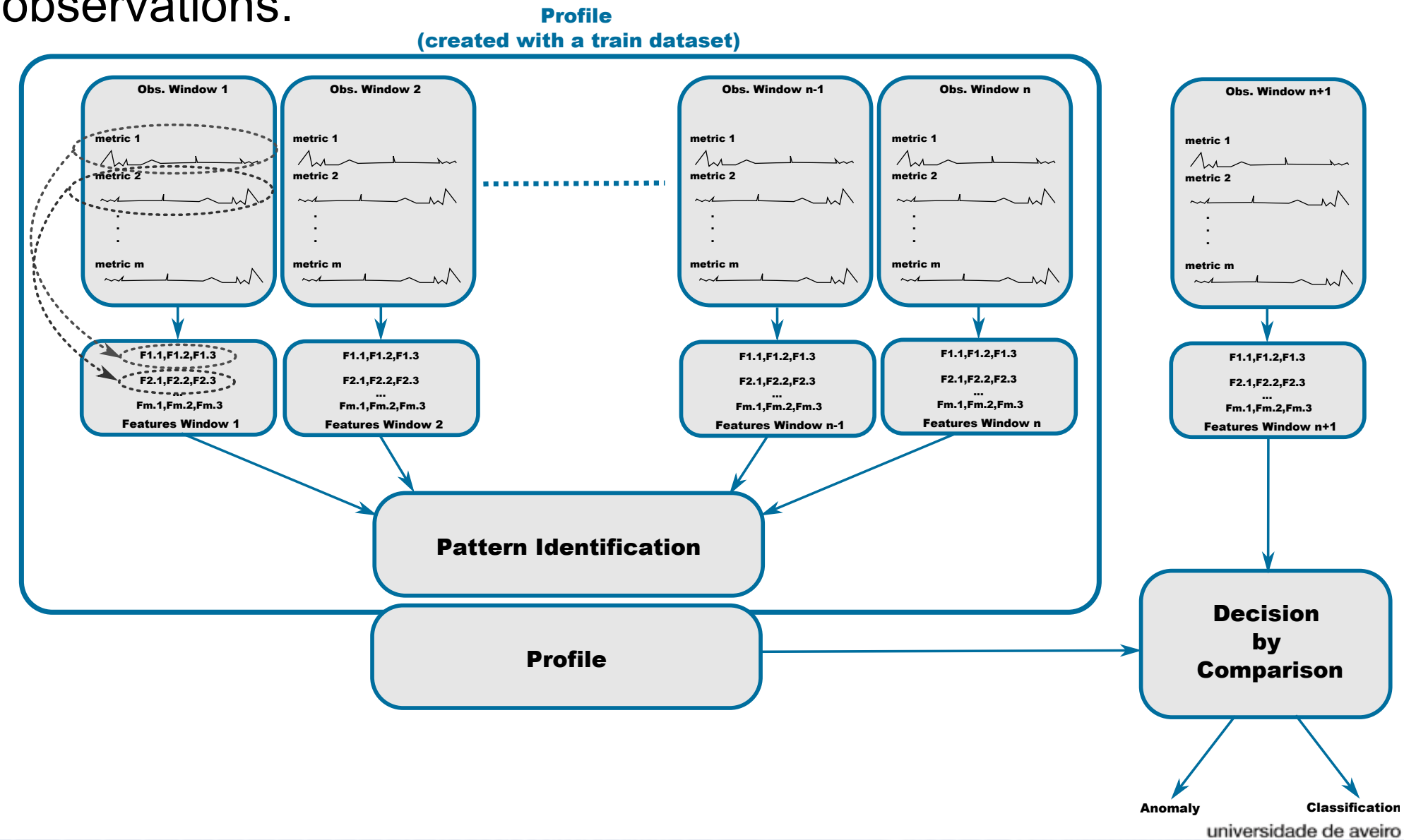


- Sliding
  - Allows for longer periods of observation, while maintaining a short period of decision.



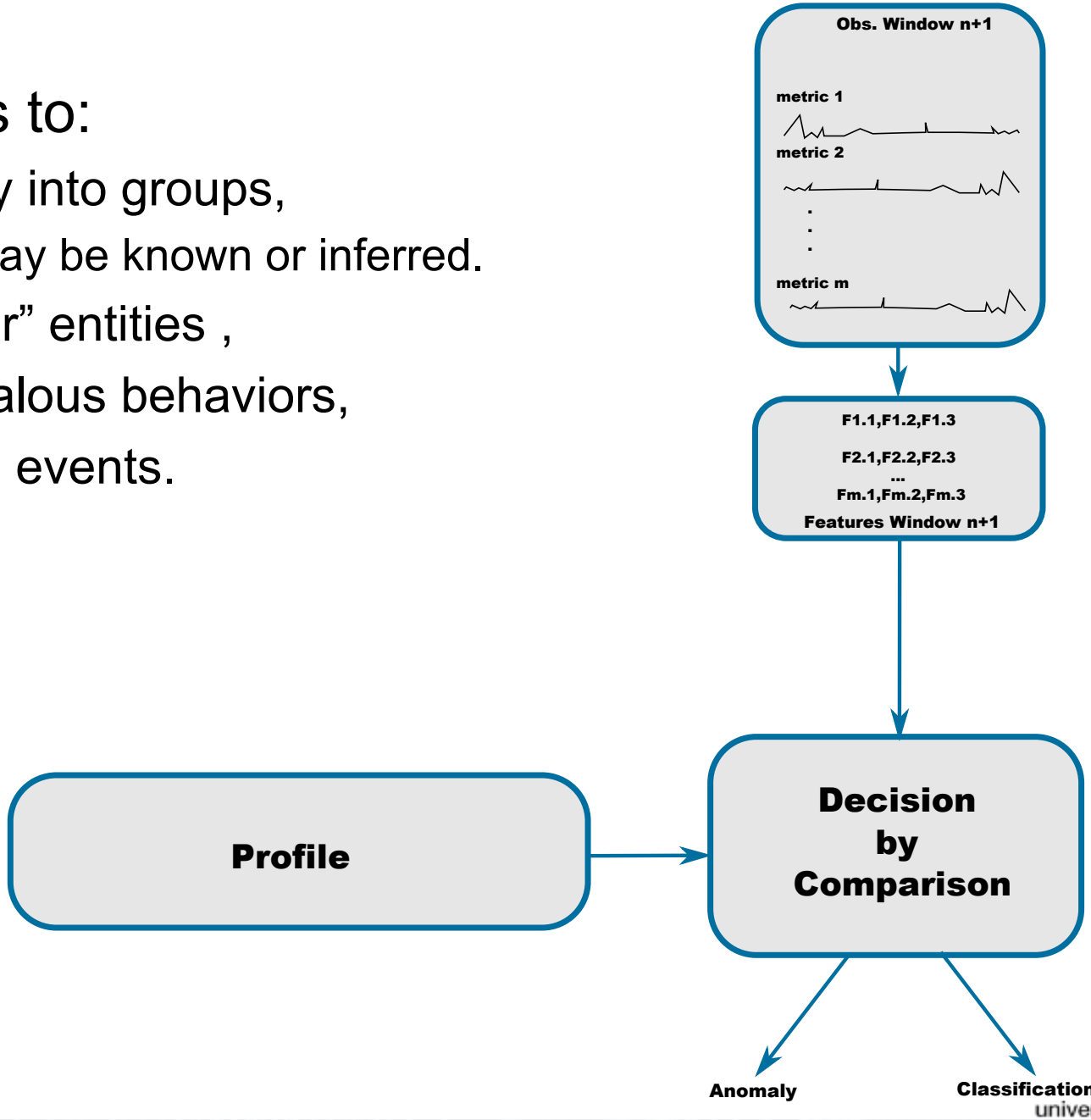
Entity Profiling

- Characterization of the observation windows after multiple observations.



Profile Comparison

- A profile allows to:
  - Classify entity into groups,
    - Groups may be known or inferred.
  - Group “similar” entities ,
  - Detect anomalous behaviors,
  - Predict future events.



Observation Features

- Time-independent descriptive statistics.
  - Mean, variance, quantiles, etc...
- Time-dependent descriptive statistics.
  - Time-relations between metrics over time
    - E.g., length of silences [number of sampling slots with metric equal to zero], length of activity [number of sampling slots with metric greater than zero], etc…
- (Pseudo-)Periodicity components.
  - Time dependent.
    - Time multi-fractality (repetition of “similar events” in multiple time-scale).
  - Auto-correlation, FFT, CWT, DWT, and other spectral/frequency analysis.
- (Parameters of) Probabilistic functions/models.
  - Base function/model may be time independent or time dependent.

Descriptive Statistics (1)

- For a (equally) sampled-continuous time process:

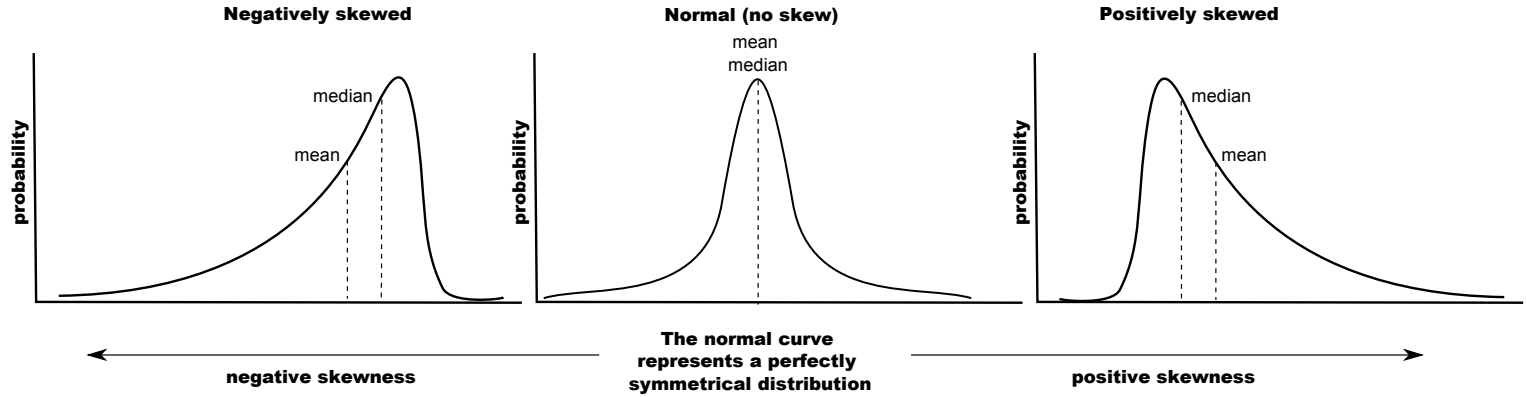
$$X = \{x'_t = x_k, T_0 + k\Delta t \leq t < T_0 + (k+1)\Delta t, k = 1, 2, \ldots, N\}$$

- **Mean**: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **Median**: $m_d = F^{-1}(0.5)$
- **Variance**: $Var(X) = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2$
- **nth Central Moment**: $m_n = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^n$
- **Quantiles/Percentiles** $Y = \{y_j\}_{1 \leq j \leq N} = \text{sorted}(\{x_k\}_{1 \leq k \leq N})$
  - 64th percentile (64%)=0.64 quantile
  - Quartiles: 25%, 50%, and 75%

$$\pi_p = \min(y_{j \geq pN})$$

Descriptive Statistics (2)

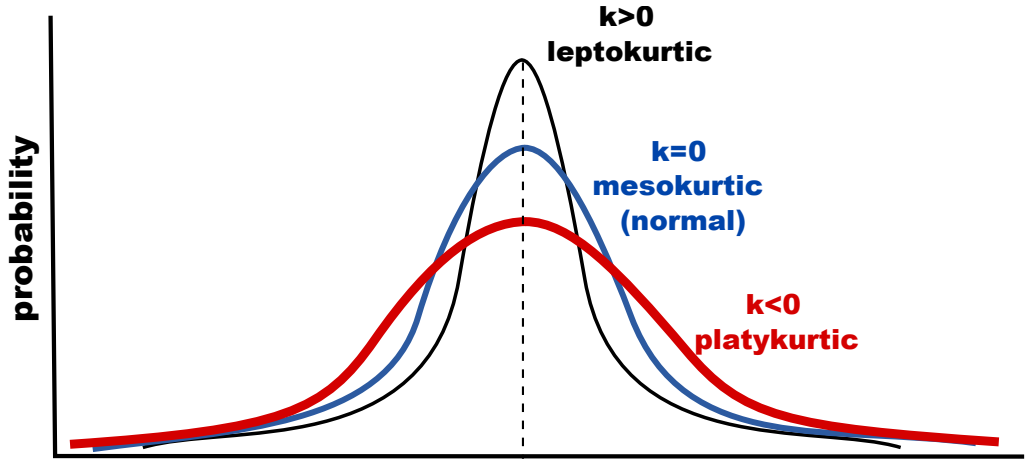
- **Skewness**:
  - Measure of the asymmetry of the probability distribution about its mean.



$$b_1 = \frac{m_3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2]^{3/2}}$$

- **Excess Kurtosis**:
  - Measure of the "tailedness" of the probability distribution.
    - "-3" constant is used to normalize kurtosis to zero for a normal distribution.

$$k = \frac{m_4}{\sigma^4} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2]^2} - 3$$



Descriptive Statistics (3)

- Covariance
  - Metric that quantifies how much two random variables have simultaneous variations:

$$\mathrm{Cov}_{X,Y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

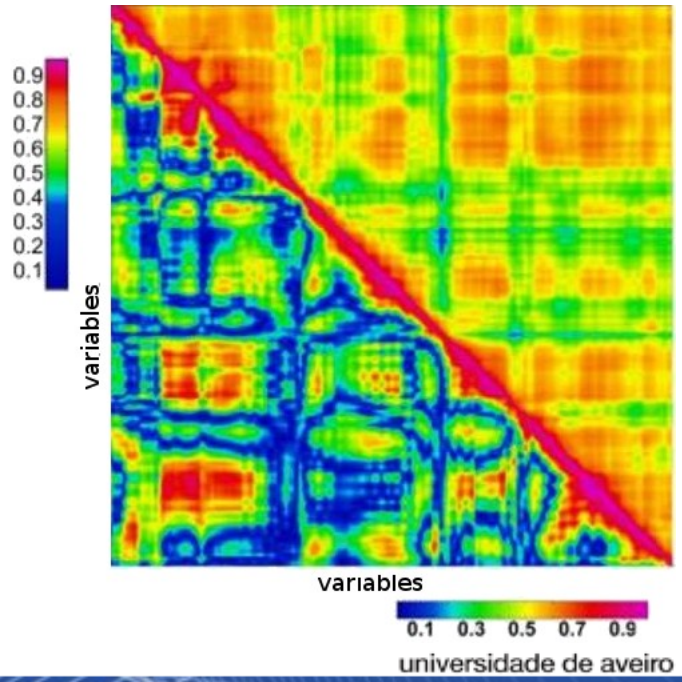
- Correlation coefficient
  - Normalized covariance, varies between -1 and 1:

$$\rho_{X,Y} = \frac{\mathrm{Cov}_{X,Y}}{\sigma_X \sigma_Y} \qquad \sigma_X = \sqrt{\mathrm{Var}(X)}$$

- Correlation matrix
  - Defined by a (MxM) matrix, to quantify the correlation between M variables X_i:

$$C = \{c_{i,j}\}, i, j = 1, \ldots, M$$

$$c_{i,j} = \rho_{X_i, X_j}$$

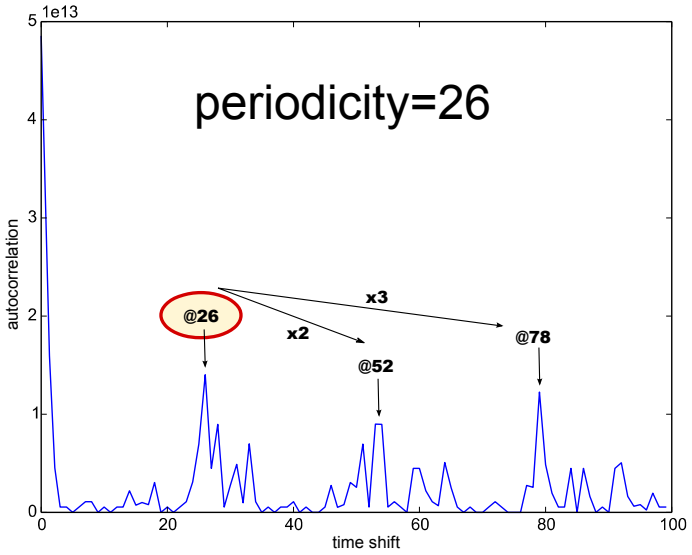


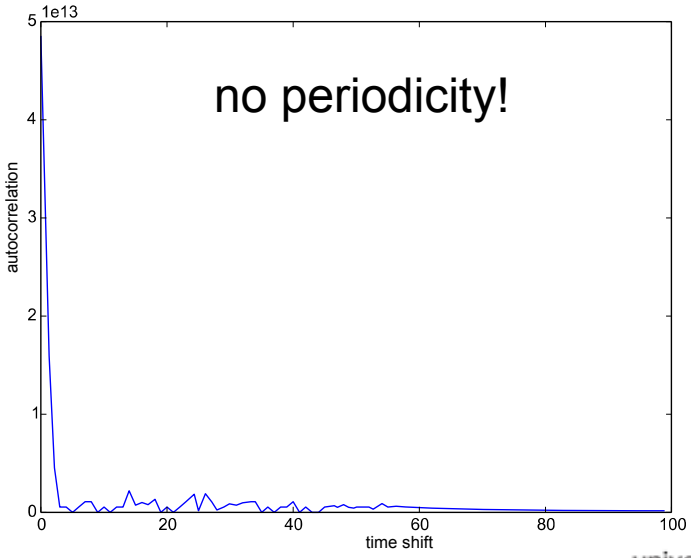
Periodicity Analysis (1)
Autocorrelation

- Autocorrelation
  - Correlation between the process and a shifted version (in time, by k samples) of the same process:

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i - \mu_X)(x_{i+k} - \mu_X)}{\sum_{i=1}^{N}(x_i - \mu_X)^2}$$

- Autocorrelation local maximums (peaks), reveal periodicity.
  - Differences between positions (k) of local maximums give periodicity.



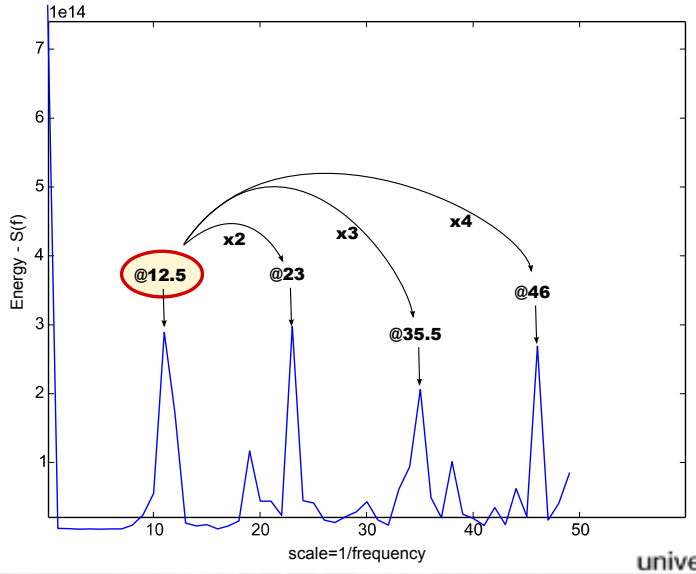


Periodicity Analysis (2)
Periodograms

- Periodogram
  - Frequency analysis → Spectral density estimation: Energy per frequency.
  - Given by the modulus squared of the discrete Fourier transform.
    - For a signal x_i sampled every Δt:

$$S(f) = \frac{\Delta t}{N} \left| \sum_{n=1}^{N} x_n e^{-j2\pi n f} \right|^2, -\frac{1}{2\Delta t} < t \leq \frac{1}{2\Delta t}$$

  - The inverse of the frequencies with higher energy give the different periods (of periodicity).

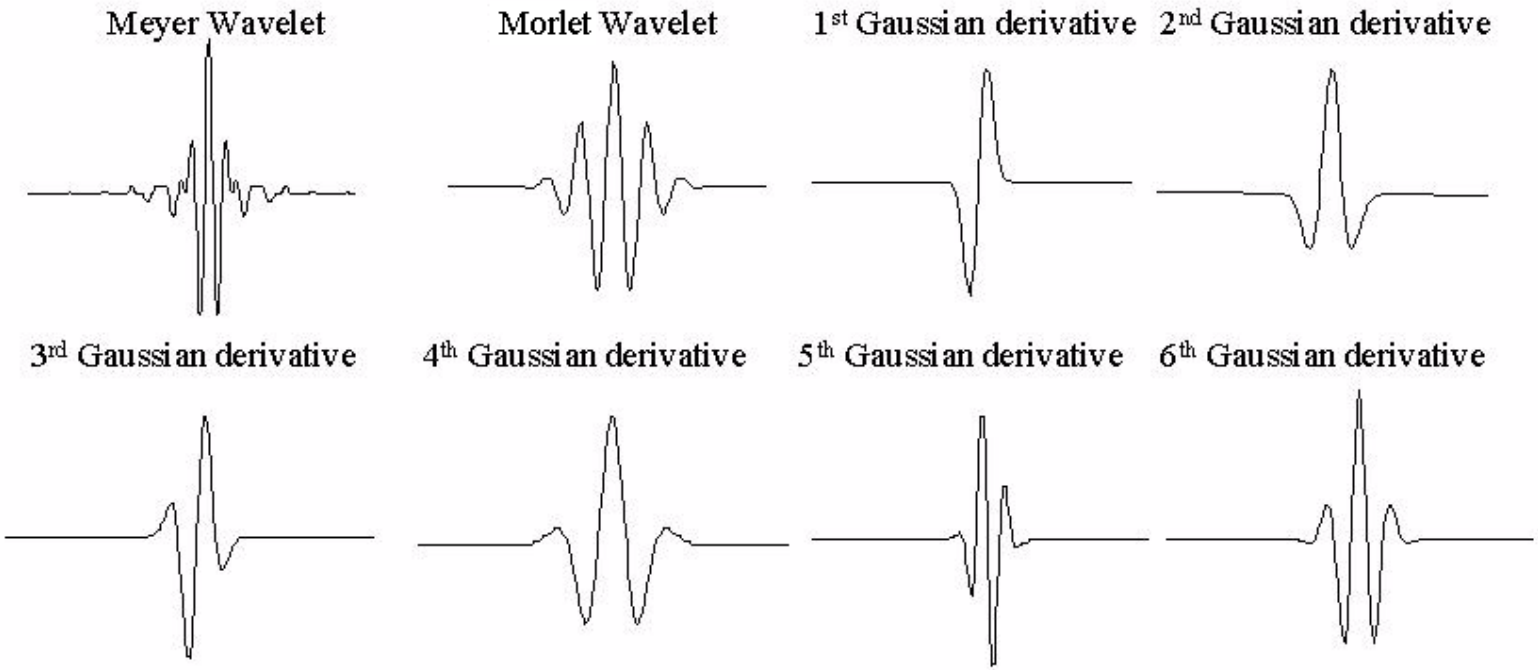


Periodicity Analysis (3)
Scalograms

- Scalogram
  - Joint Frequency/Time analysis → Wavelet Analysis
    - Energy per frequency/time.

$$\Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{+\infty}^{-\infty} x(t)\psi^*(\frac{t-\tau}{s})dt$$

Wavelet functions $\psi^*(t)$



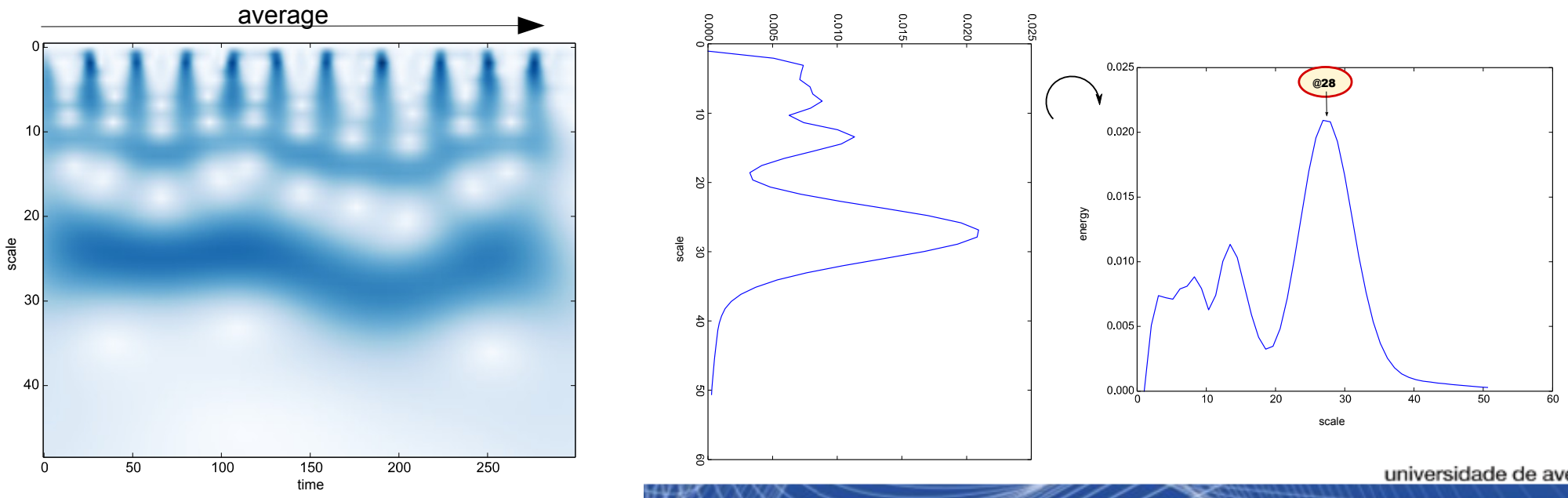
Periodicity Analysis (4)
Scalograms

- Given by the normalized modulus squared of the Wavelet transform.

$$\hat{E}_x(\tau, s) = \frac{\left|\Psi_x^{\psi}(\tau, s)\right|^2}{\sum_{\tau' \in \mathbf{T}} \sum_{s' \in \mathbf{S}} \left|\Psi_x^{\psi}(\tau', s')\right|^2}$$

  - Averaged over time.

$$\bar{e}_x(s) = \frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} \hat{E}_x(\tau, s), \forall s \in \mathbf{S}$$



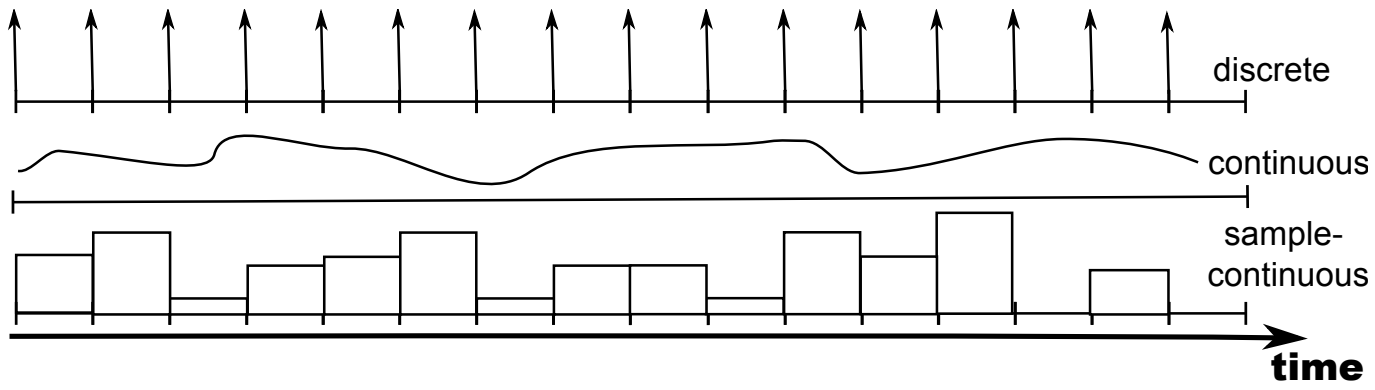
Stochastic Process

- A collection of variables indexed by a time variable, representing the evolution of some system over time.

$$X = \{x_t = a, t \in T\}$$

  - Discrete variables: $a \in A, A = \{\alpha_1, \alpha_2, \ldots, \alpha_S\}$
  - Continuous variables: $a \in \mathbb{R}$
  - Discrete time: $T = \{T_0 + k\Delta t, k \in \mathbb{N}_0\}$
  - Continuous time: $T = \mathbb{R}_0$
  - A continuous time process never exists in practice, what exists is a Sample-Continuous time process:

$$x_t = x'_{T_k}, t \in \mathbb{R}, T_k \leq t < T_{k+1}$$



Multivariate Stochastic Processes

- Variables belong to a multidimensional space of dimension N.

$$X = \{x_t = \vec{a}, t \in T\}$$

  - Discrete variables:

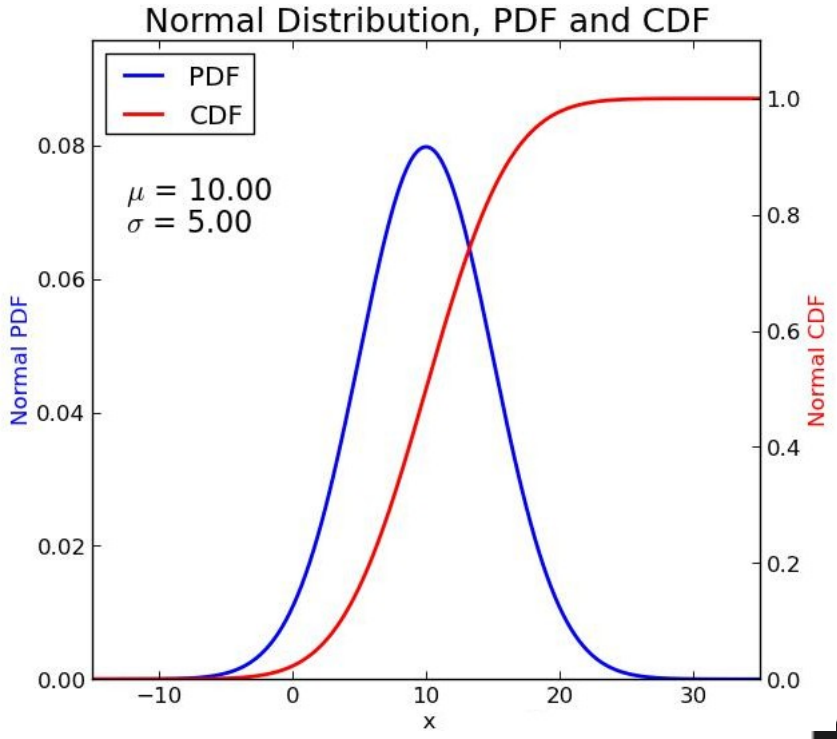
$$\vec{a} \in A, A = \{\vec{\alpha}_1, \vec{\alpha}_2, \ldots, \vec{\alpha}_S\}, \vec{\alpha}_i \in \mathbb{R}^N$$

  - Continuous variables:

$$\vec{a} \in \mathbb{R}^N$$

Probability Functions (1)

- Discrete
  - Probability Mass Function (PMF)
  - $\text{pmf}_X(a) = Pr[X = a], a \in A, A = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$
  - $\sum_{\forall a \in A} \text{pmf}_X(a) = 1$
- Continuous
  - Probability Density Function (PDF)
  - $f_X(a) = Pr[X = a], a \in \mathbb{R}$
  - $\int_{-\infty}^{+\infty} f_X(x)dx = 1$
  - Cumulative Density Function (CDF)
  - $F_X(a) = Pr[X \leq a] = \int_{-\infty}^{a} f_X(x)dx$



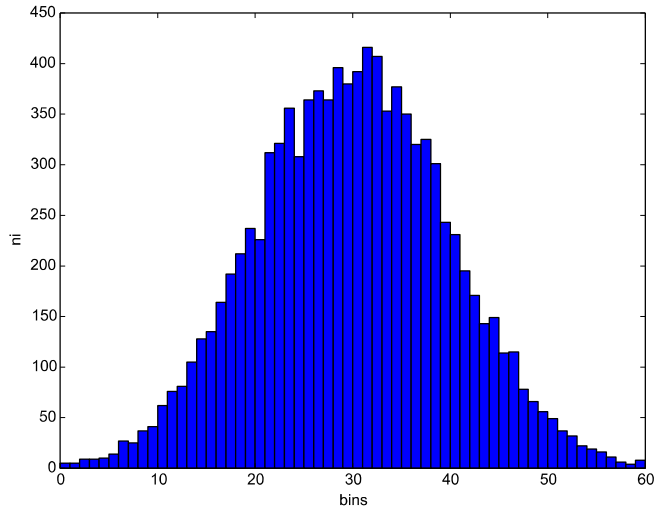
Probability Functions (2)

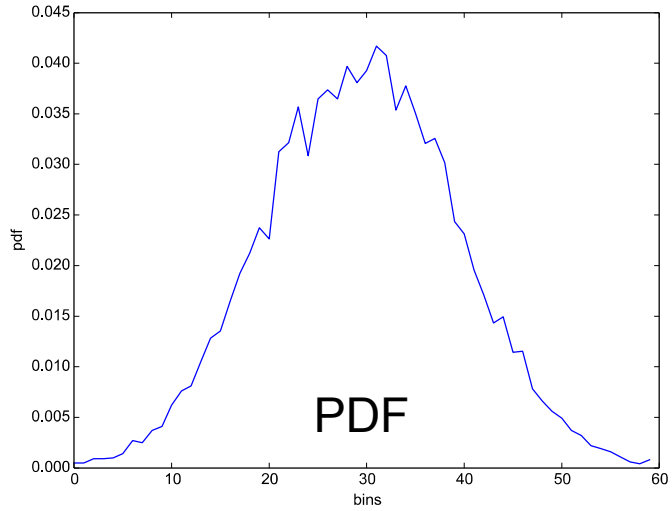
- Inference and interpretation
  - Histogram with bins $B = \{b_1, b_2, \ldots, b_{M+1}\}$

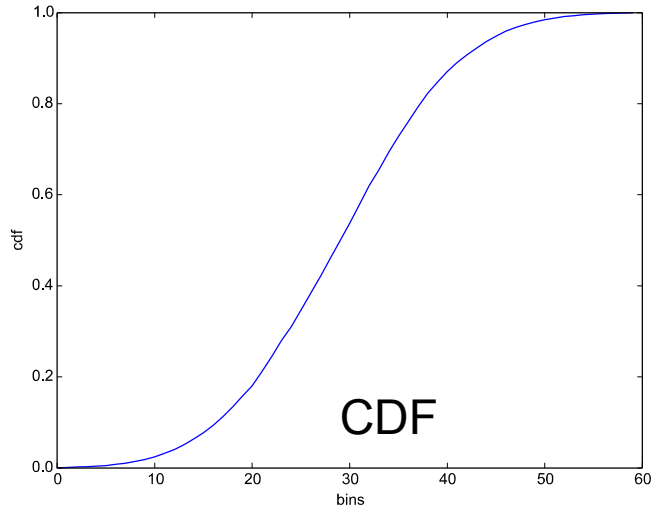
$$n_i = \text{count}(b_i \leq X < b_{i+1}), i = 1, 2, \ldots, M$$

$$f_X(a) = \frac{n_i}{N(b_{i+1} - b_i)}, \exists i, b_i \leq a < b_{i+1}$$

$$F_X(a) = \sum_{i=1}^{j} \frac{n_i}{N(b_{i+1} - b_i)}, \max_j : a < b_{j+1}$$

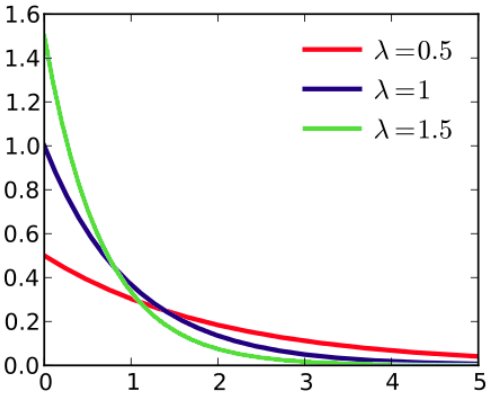


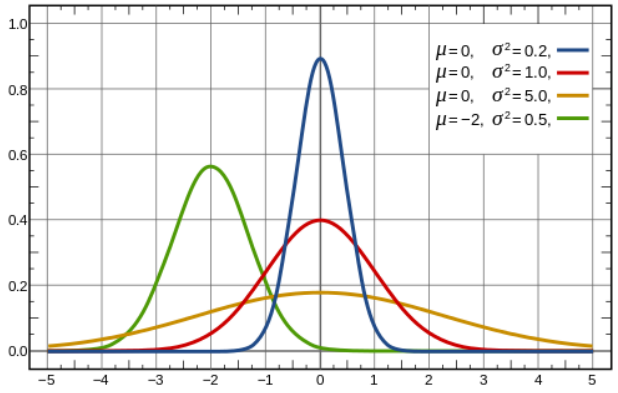




Statistical Univariate Distributions

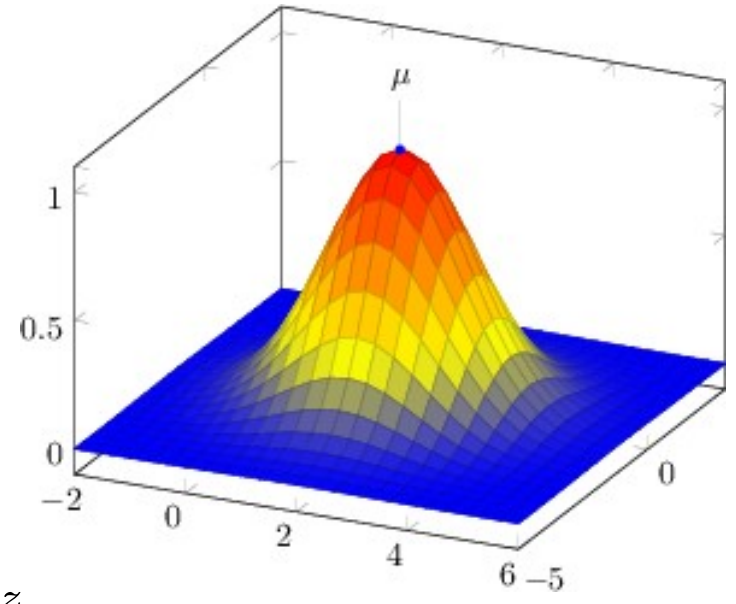
- Most commonly used distributions:
  - Discrete
    - Uniform: $\text{pmf}_X(a) = \begin{cases} \frac{1}{N}, a \in A \\ 0, a \notin A \end{cases}, A = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$
    - Poisson: $\text{pmf}_X(a) = \frac{\lambda^a e^{-\lambda}}{a!}, \lambda > 0$
  - Continuous
    - Uniform: $f_X(a) = \begin{cases} \frac{1}{a_{\max} - a_{\min}}, a \in [a_{\min}, a_{\max}] \\ 0, \text{otherwise} \end{cases}$
    - Normal/Gaussian: $f_X(a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$
    - Exponential: $f_X(a) = \lambda e^{-\lambda a}$





Multivariate Distributions

- Joint probability of a multidimensional variable.
- Incorporates correlation (ρ) between dimensions.
- E.g., 2-Dimensions Gaussian:



$$f_X((a_1, a_2)) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{z}{2(1-\rho^2)}}$$

$$z = \frac{(a_1-\mu_1)^2}{\sigma_1^2} - \frac{2(a_1-\mu_1)(a_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(a_2-\mu_2)^2}{\sigma_2^2}$$

Variable/Features Reduction (1)

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

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

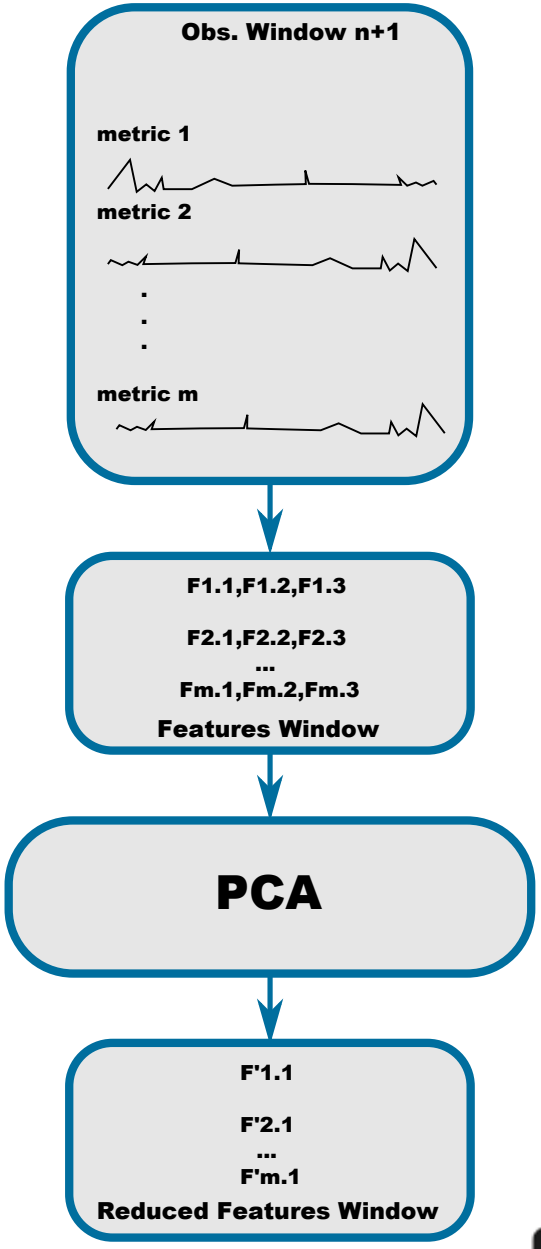
- The reduction of variables is mandatory to simplify classification.
- **Principal Components Analysis (PCA)**
  - Uses a transformation to convert a set of possibly correlated features into a set of values of uncorrelated variables called principal components.
  - The principal components of an event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

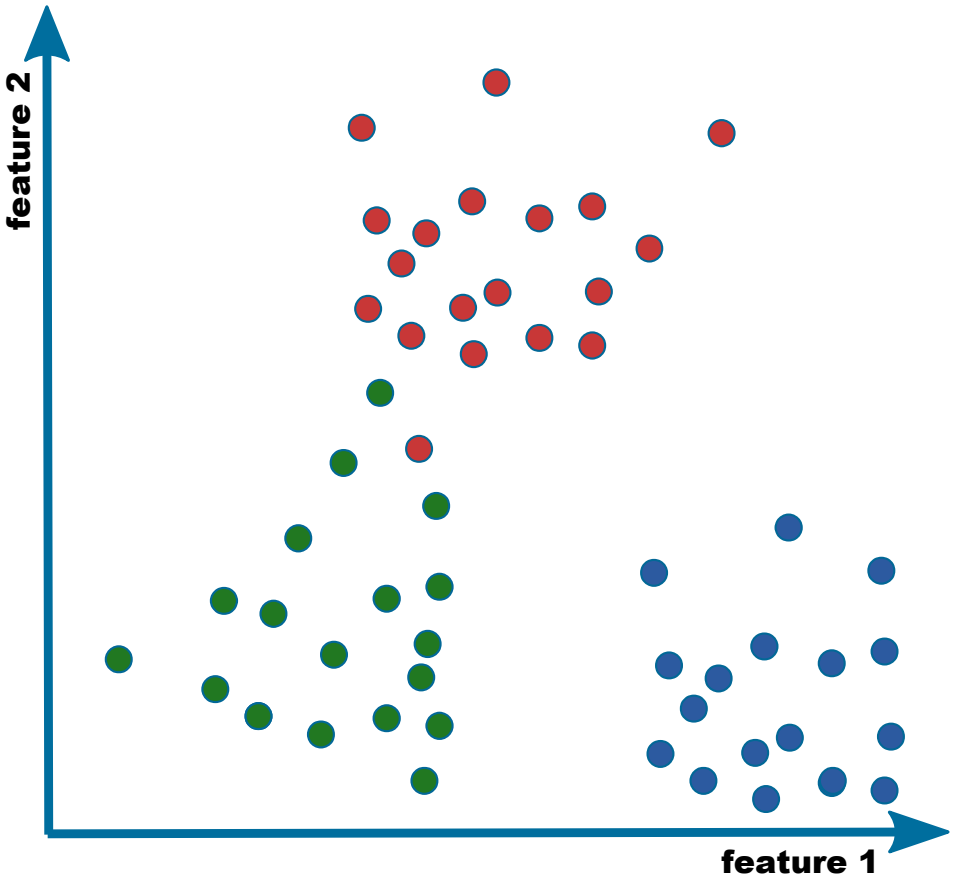
Variable/Features Reduction (2)

- PCA can be used to reduce the number of features.
  - Simplify Machine Learning or Statistical Analysis (input) complexity.
  - Removes correlated features!
  - Creates a linear combination of features to create uncorrelated new features.



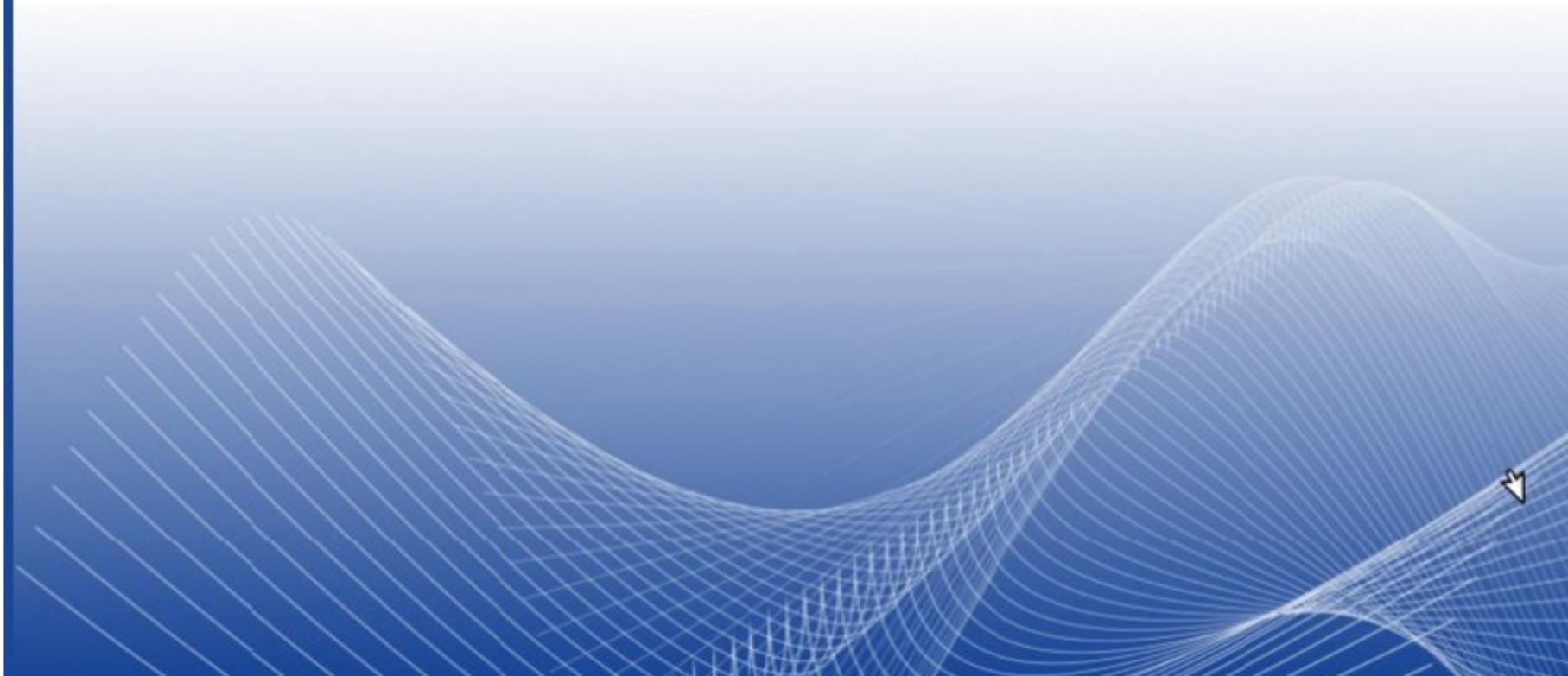
Profile as a N-Dimensional Euclidean Universe

- 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

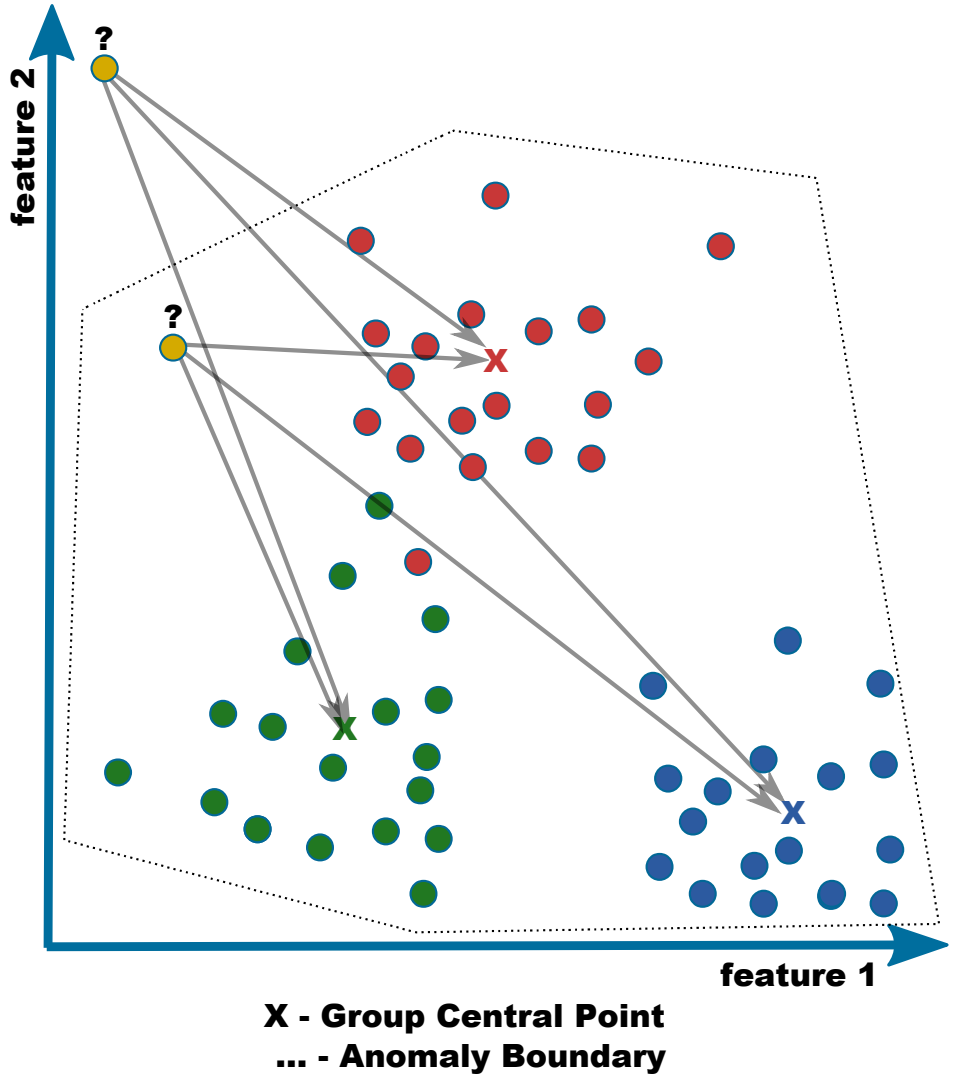


universidade de aveiro

deti.ua.pt

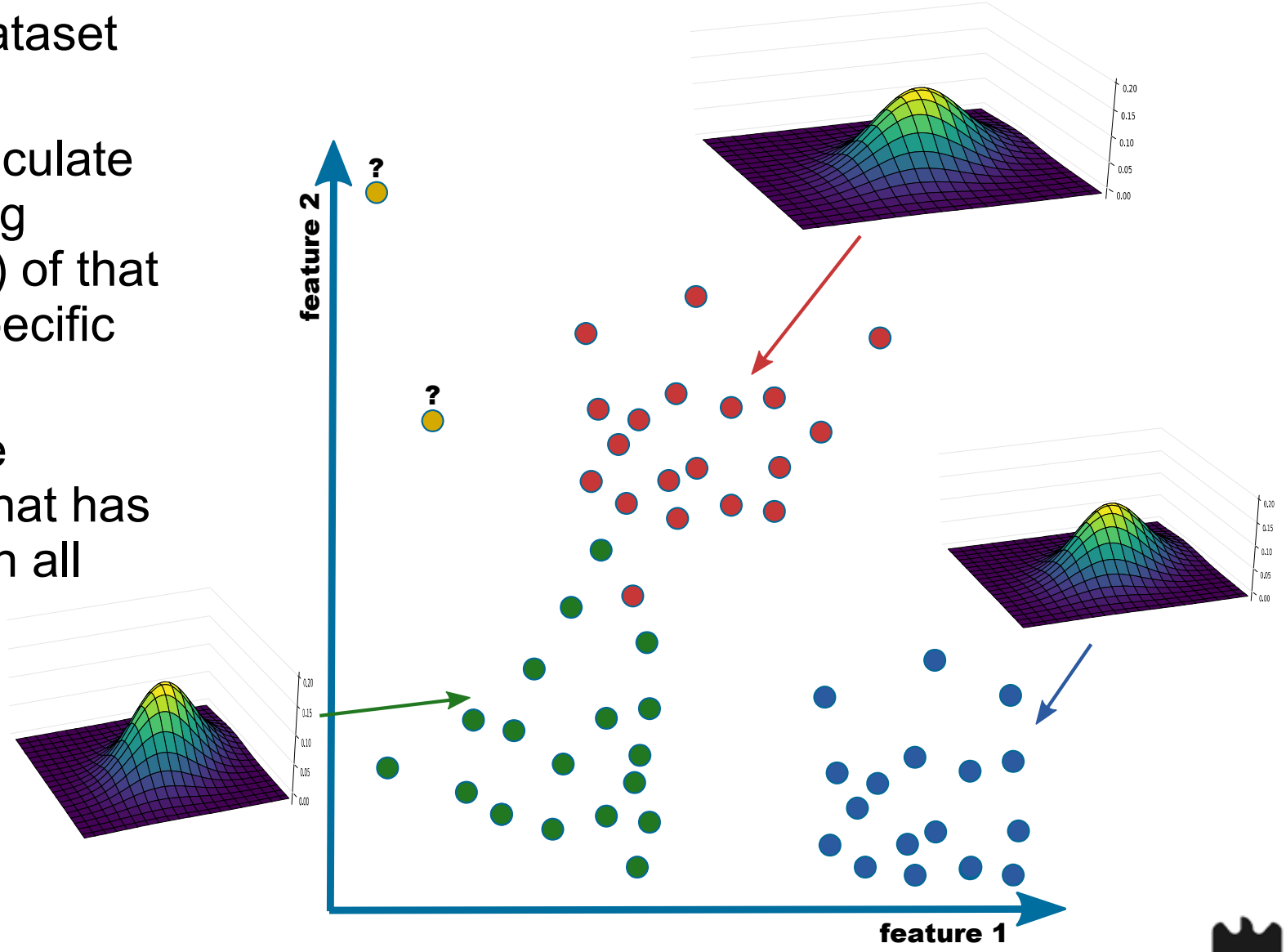
Distances to Central Point(s)

- 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.



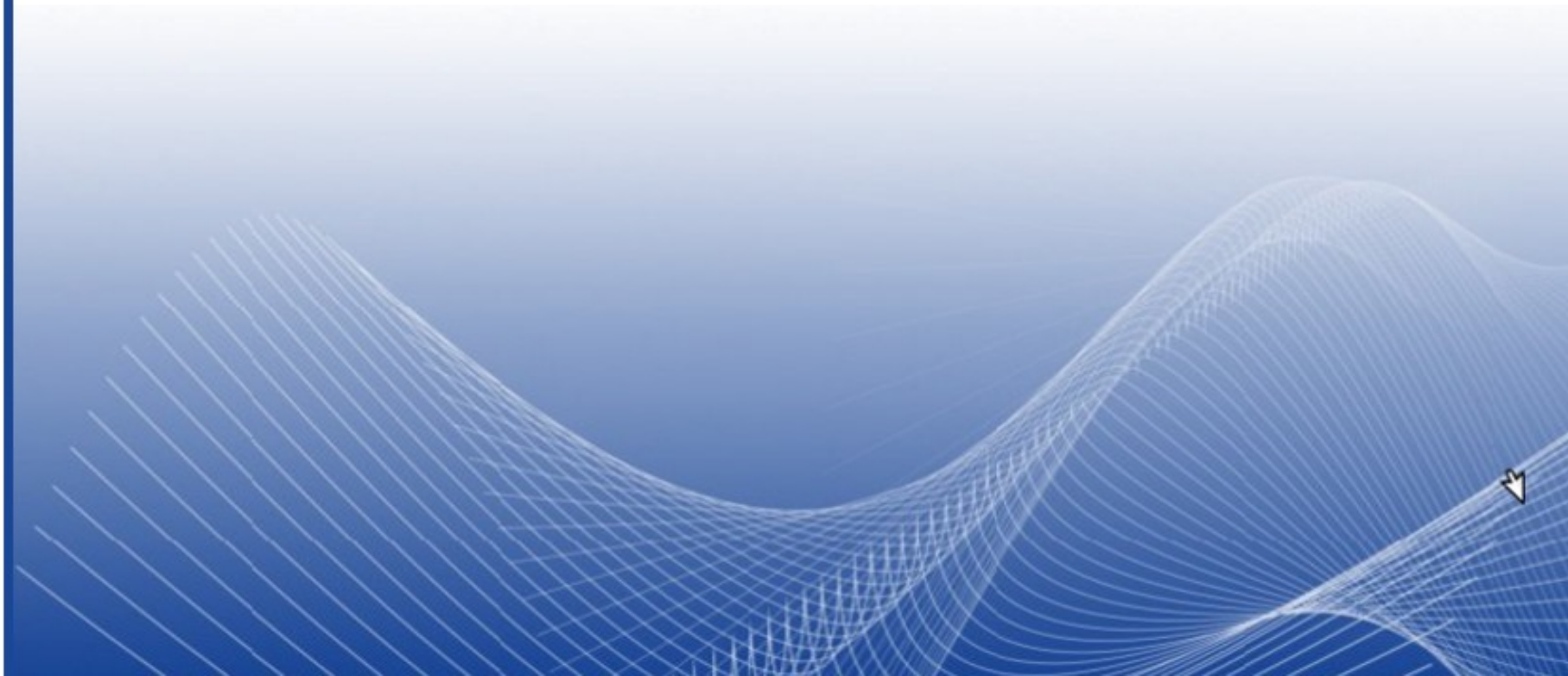
N-Dimensional Distributions

- 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.

Data Inputs

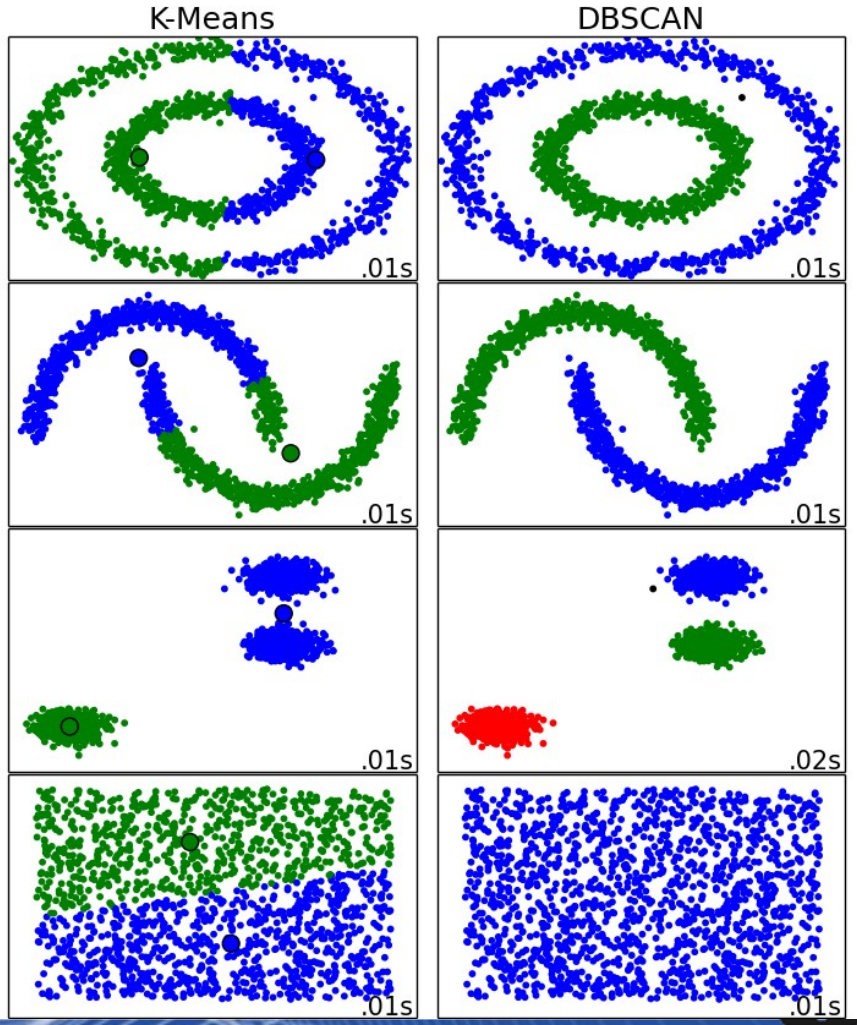
- Raw data inputs are possible, however its increases the complexity of the machine learning algorithm.
  - Worse results, longer calculation/response times.
- Input data should be the result of raw data processing (complexity reduction).
  - Observation features.
  - Statistical metrics, statistical functions, PCA, scale analysis metrics/descriptors, …
- Inputs should be normalized.
  - Usually to mean zero, variance one!

Approaches

- Clustering
- 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
  - ...

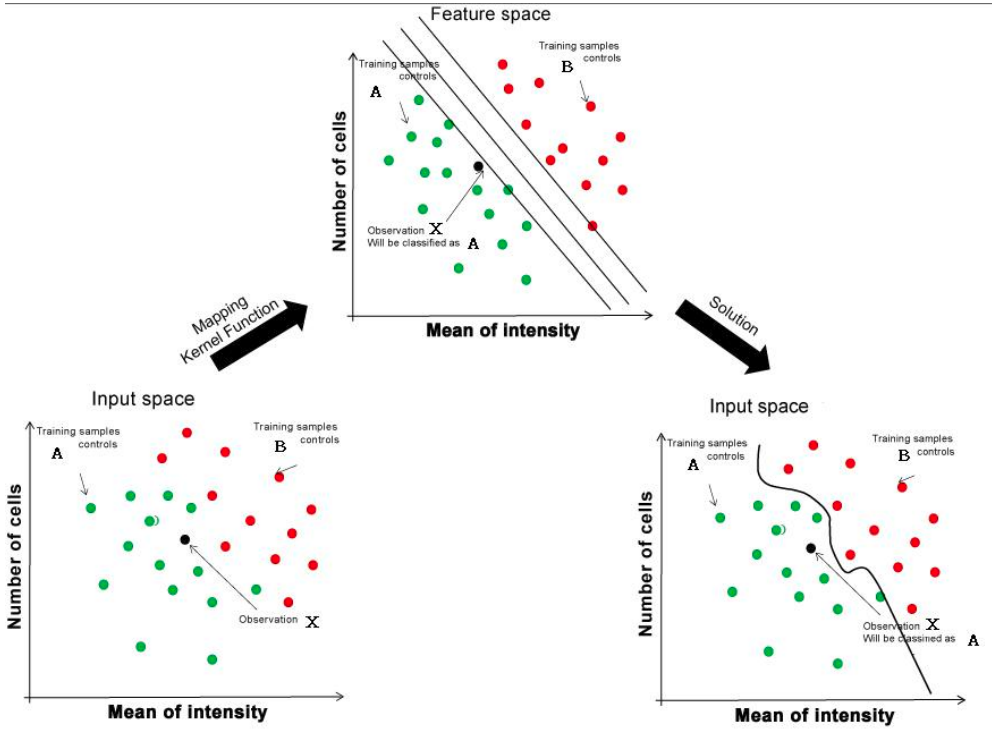
Classification / Clustering

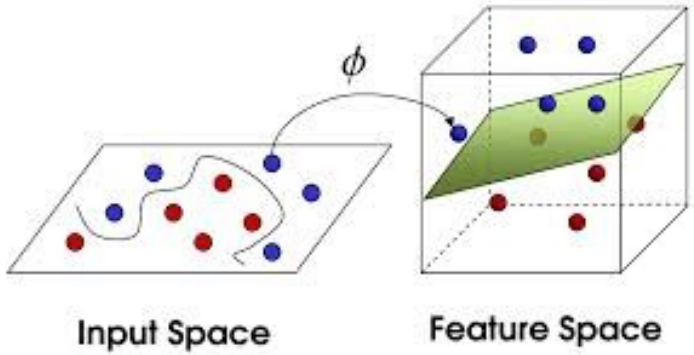
- Clustering is the process of grouping (classifying) a set of objects in such a way that objects in the same group (cluster) are more “similar” to each other than to those in other clusters.
- Algorithms:
  - K-Means
    - Requires the a priori knowledge of the number of clusters.
    - Uses the distances 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

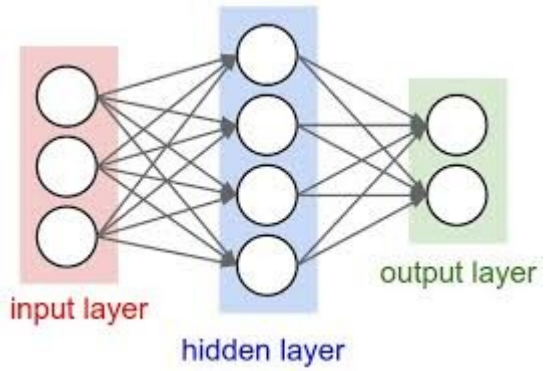
- Classification defined by a separating hyper-plane-
- Optimal hyper-plane for linearly separable patterns.
- Kernel functions allow the separation of patterns that are not linearly separable by transformations of original data.
- Solutions found using a minimization problem.

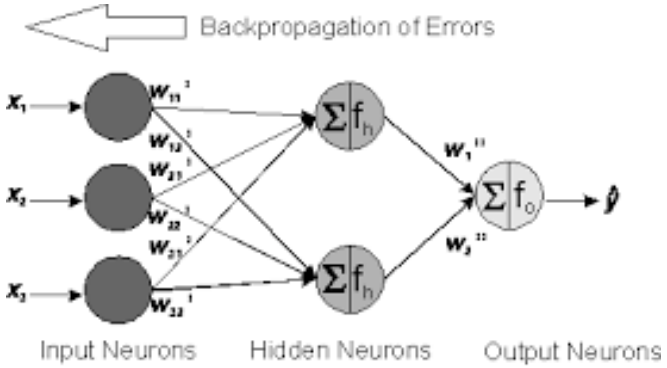


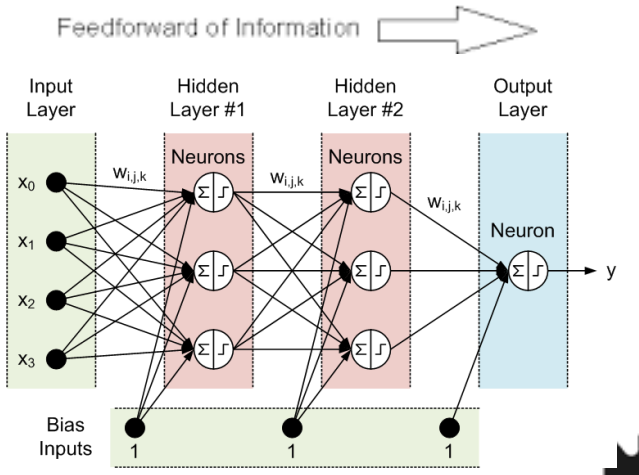


Artificial Neural Networks

- 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.
  - 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.

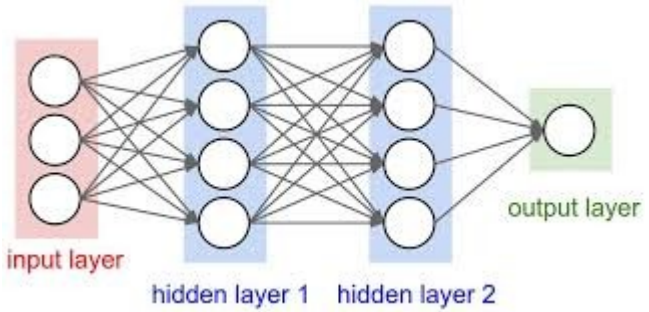


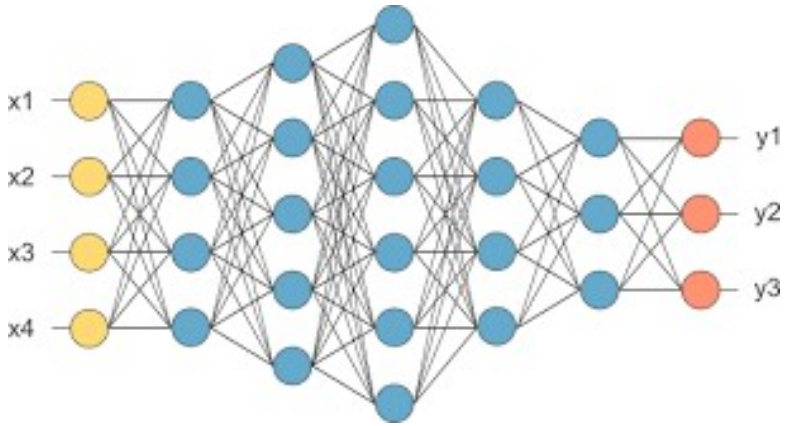




Deep Learning

- 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





Performance Evaluation

universidade de aveiro
deti.ua.pt

Evaluation Process