TÉCNICO LISBOA

DATA EXPLORATION

Cláudia Antunes

2018 / 19

TÉCNICO LISBOA

DATA SCIENCE

- "The art of discovering what we don't know from data"

  Carlos Somohano, founder of Data Science London

- "A systematic study of a generalizable extraction of knowledge from data"

  Vasant Dhar, Professor at Center for Data Science at NYU

- "The nontrivial extraction of implicit, previously unknown, and potentially useful **information** from **data**."

  William J. Frawley, *AI Magazine* in **1995**

BASIC CONCEPTS

Data sets are made up of records.

Records

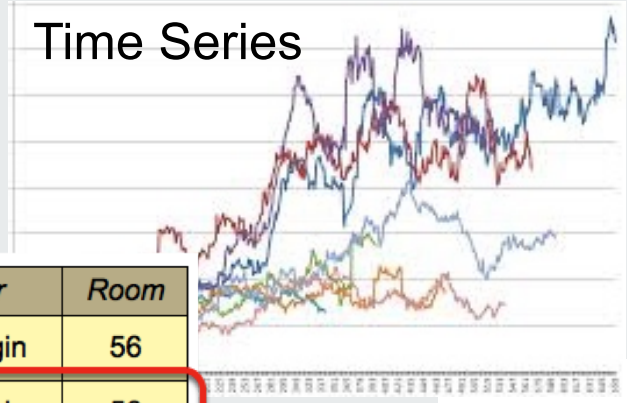
- also known as ***instances***, ***data objects*** or ***samples***
- the recorded facts that describe the state or behavior of some entity, in accordance with a set of attributes;

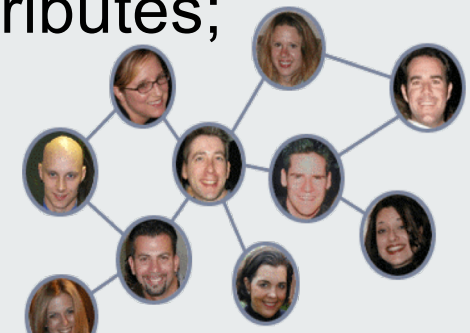
✦ Examples:

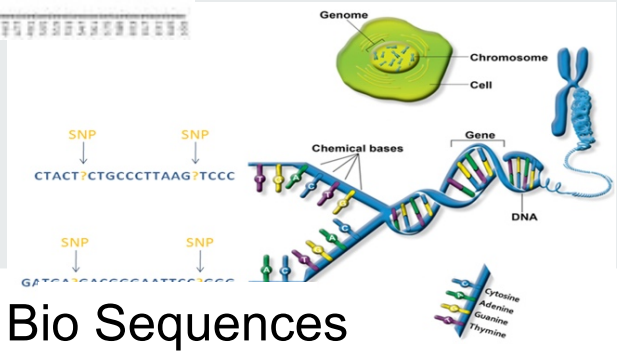
Tabular data

ID No.	Name	D.o.B.	Phone	Class	Tutor	Room
356	Jess	3 Mar 1995	7564356	5B	Mr Noggin	56
412	Hamad	12 Nov 1994	7465846	5B	Mr Noggin	56
459	Sita	9 Jan 1994	8565634	6Y	Ms Take	18
502	Hamad	3 Mar 1995	6554546	5B	Mr Noggin	56

One **Record**







Bio Sequences

BASIC CONCEPTS

Records are tuples of values, according to a set of attributes

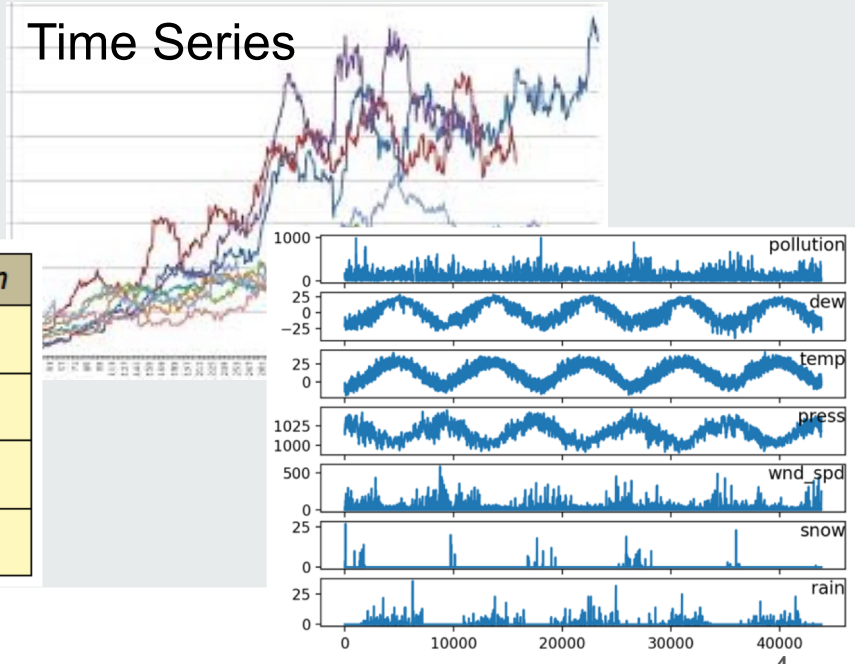
Attribute

- also known as ***fields, dimensions***, ***variables*** *or* ***features***
- **Attribute Domains**: values that can be taken by the attribute
- **Attribute Range**: the values that are actually taken

✦ Examples

Tabular data

ID No.	*Name*	*D.o.B.*	*Phone*	*Class*	*Tutor*	*Room*
356	Jess	3 Mar 1995	7564356	5B	Mr Noggin	56
412	Hamad	12 Nov 1994	7465846	5B	Mr Noggin	56
459	Sita	9 Jan 1994	8565634	6Y	Ms Take	18
502	Hamad	3 Mar 1995	6554546	5B	Mr Noggin	56

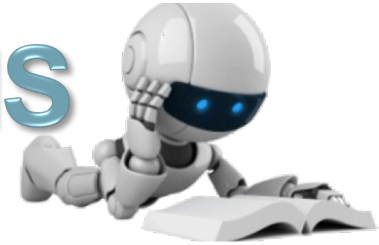


TABULAR DATA

- Data is usually represented as a ***n*x*d*** matrix, ***D***, with
  - ***n***, the number of records (lines) from x_1 to x_n
  - and ***d***, the number of attributes (columns) from A_1 to A_d

$$\mathbf{D} = \begin{array}{c|cccc} & A_1 & A_2 & \cdots & A_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

ATTRIBUTE DOMAINS



Numeric

Real-valued

- Quantities

Interval-based

- Measured on a scale of equal-sized units
- Values have order
- No true zero-point

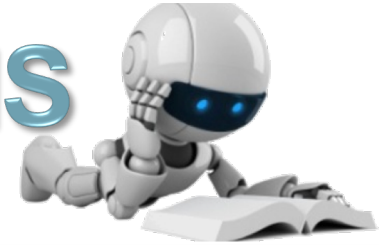
Ratio

- Inherent **zero-point**
- Known magnitude between values

Observations

- Real values can only be represented by a finite number of digits
- Typically represented as floating-point variables

ATTRIBUTE DOMAINS



Numeric

Real-valued
- Quantities

Interval-based
- Measured on a scale of equal-sized units
- Values have order
- No true zero-point

Ratio
- Inherent **zero-point**
- Known magnitude between values

Symbolic

Binary
- With only 2 values

Symmetric x Asymmetric

Nominal
- Set of discrete values
- Categories, states, "names of things"

Ordinal
- Values have a meaningful order (ranking)
- Unknown magnitude between values

INFORMATION

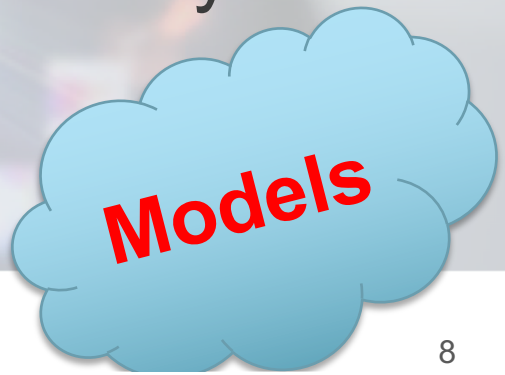
The **set of patterns** or expectations that underlie the data

Witten, 2000

From a mathematical point of view, the generation of information can be seen as **data compression**, since information can be seen as a **model** to represent several instances.

Addrians, 1996

As **the next step in the road of knowledge**, since it corresponds to an **abstraction** of the already known (recorded) data.



TABULAR DATA

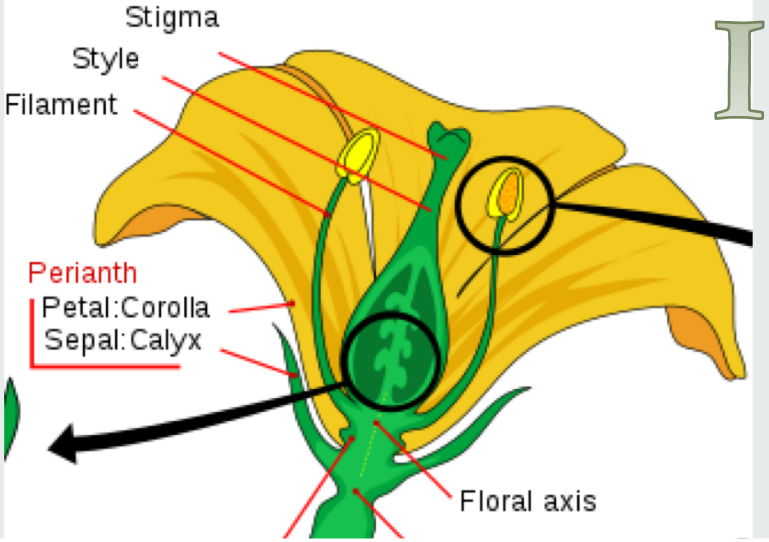
EXAMPLE

Versicolor

Verginica

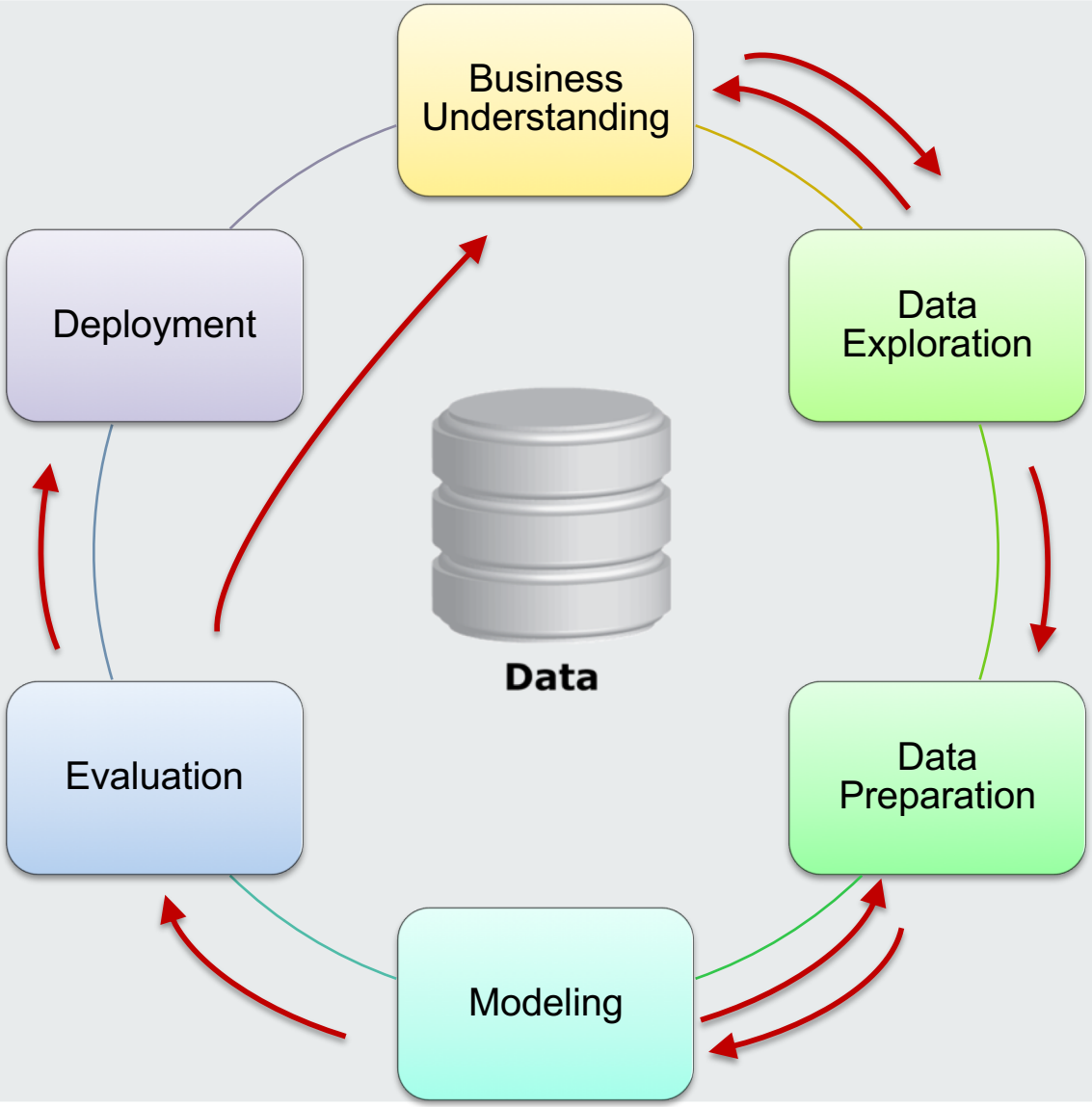
Setosa

Iris



	Sepal length	Sepal width	Petal length	Petal width	Class
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

DATA SCIENCE PROCESS



DATA EXPLORATION

- **Univariate analysis** focus at each attribute at a time

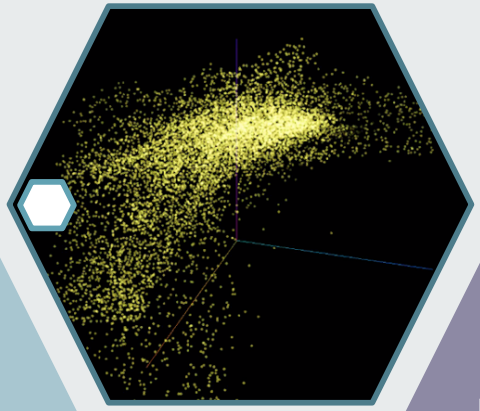
- **Multivariate analysis** focus at subsets of attributes at a time

$$\mathbf{D} = \begin{array}{c|c|c|c|c} & A_1 & A_2 & \cdots & A_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

DATA EXPLORATION

Multivariate

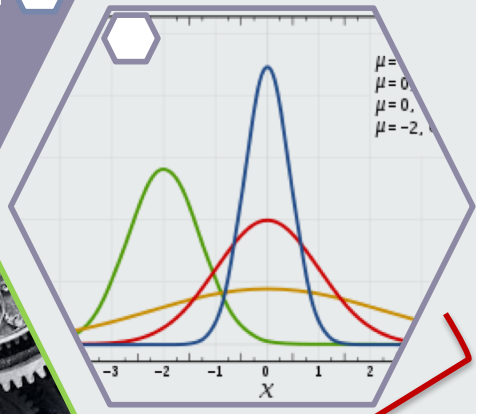
Curse of dimensionality

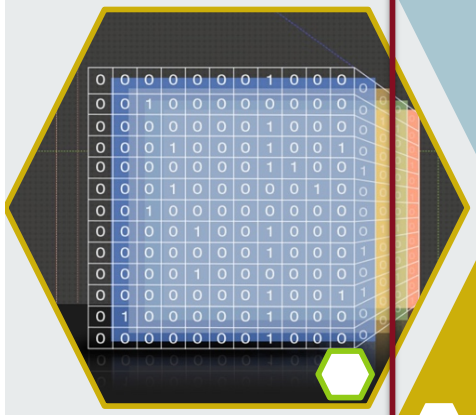


Dimension ality

Distribution

Centrality and dispersion





Granularity

Sparsity

Only presence counts

Patterns depend on the scale

Univariate

TÉCNICO LISBOA

DATA GRANULARITY

Analyzing each attribute individually

GRANULARITY

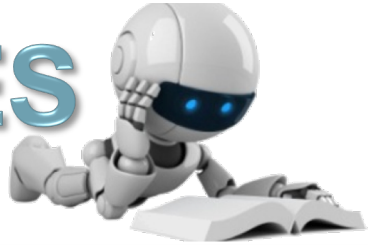
Granularity is the **data scale** or **level of detail**.

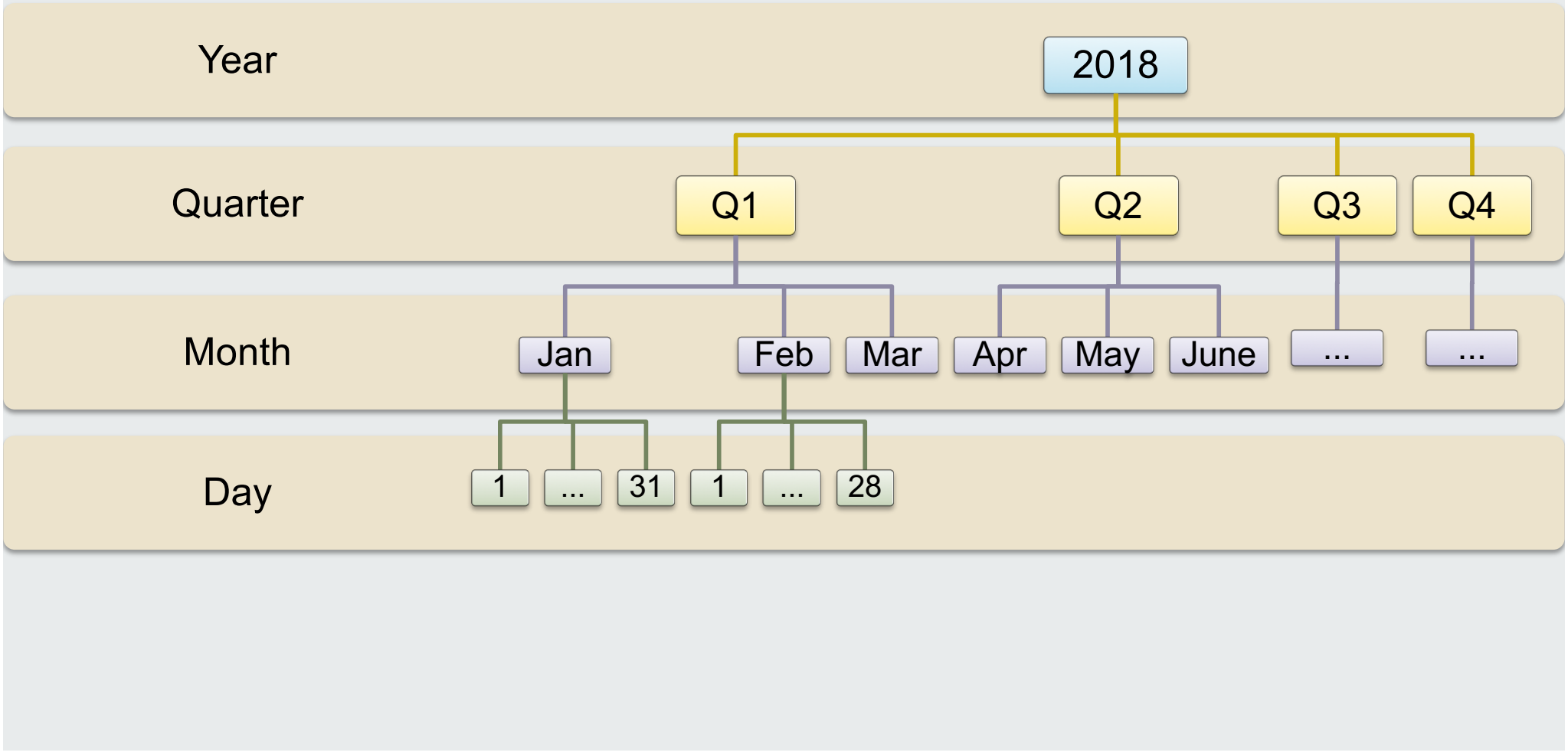
The finer the granularity,
the deeper the level of detail.

- Data at a finer granularity can be **aggregated** into a coarser one.

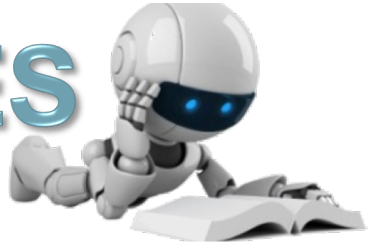
- Granularities are specified:
  - by **taxonomies** (concept hierarchies) for symbolic attributes
  - by **discretization** for numeric ones

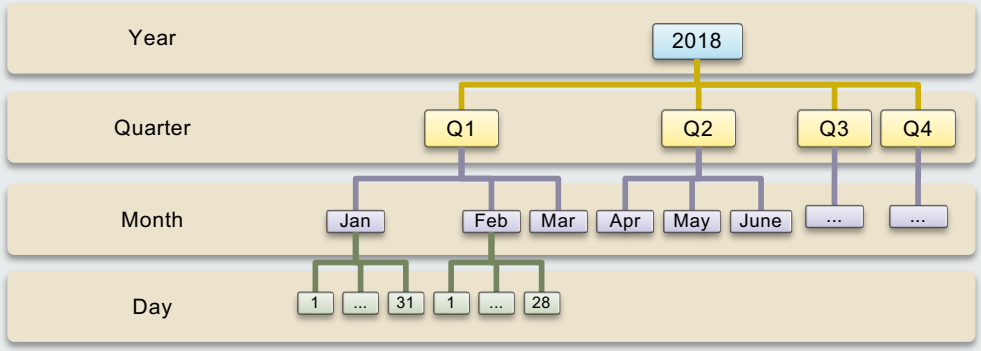
TIME GRANULARITIES

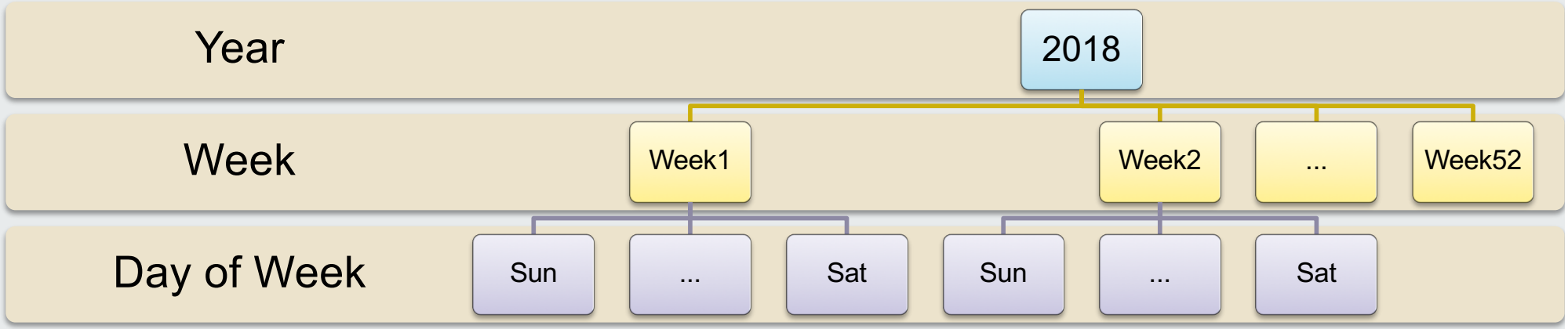




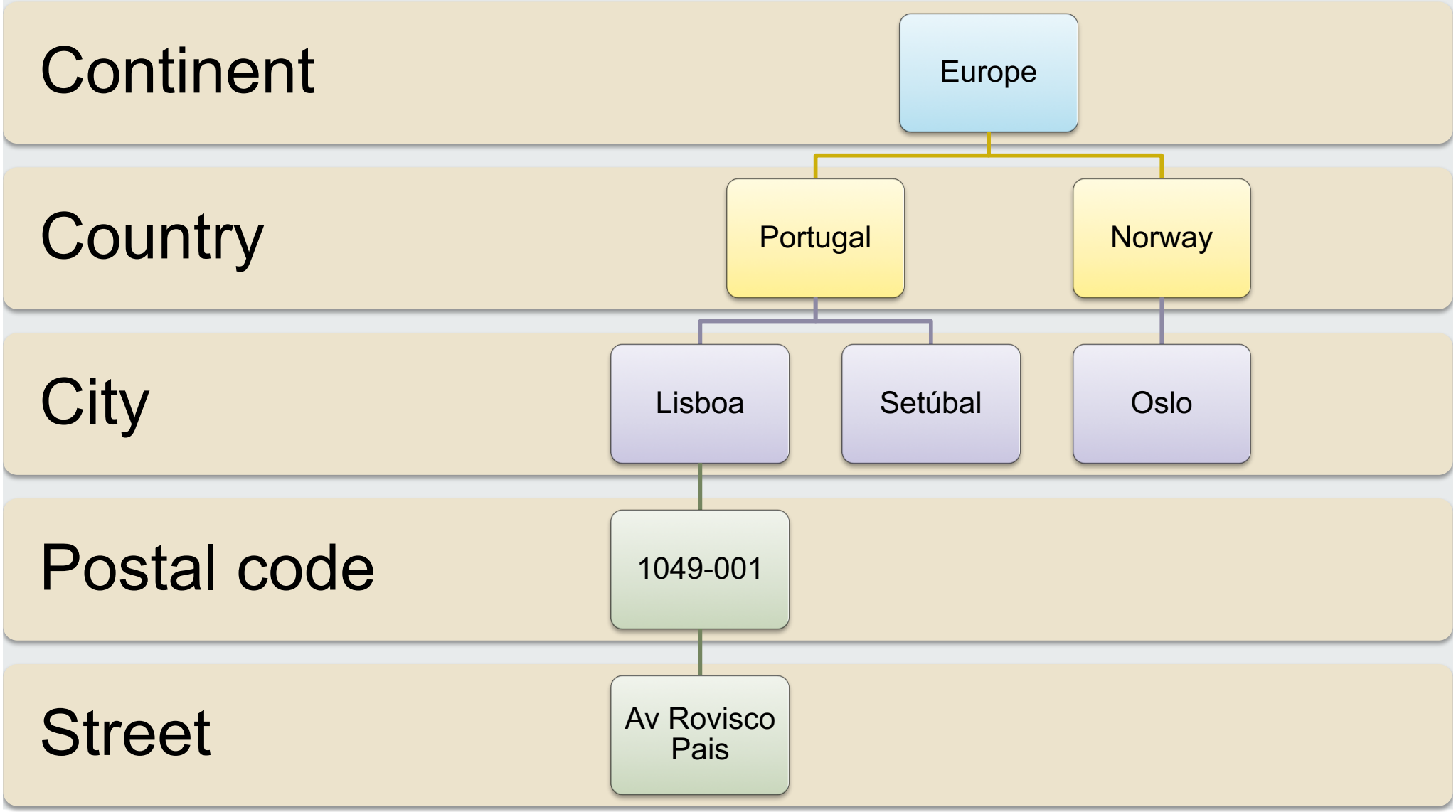
TIME GRANULARITIES







Local Granularities



TÉCNICO LISBOA

What about numerical data?

Discretization

Numeric Discretization

Transforms real-numbered into interval-based data

Equal-width

Partitions the range of A into k ***equal-width*** intervals.

Width corresponds to A's range divided by k

$$w = \frac{a_{max} - a_{min}}{k}$$

i^{th} interval boundary v_i is

$$v_i = a_{min} + iw, \forall i = 1, \dots, k-1$$

Equal-frequency

Divide the range of A into intervals containing approx. the same number of points, same **frequency**.

Each interval contains $1/k$ **of the probability mass**, and can be computed by the inverse cumulative distribution function $\hat{F}_A^{-1}$

i^{th} interval boundary v_i is

$$v_i = \hat{F}_A^{-1}\left(i/k\right), \forall i = 1, \dots, k-1$$

$$\hat{F}_A^{-1}(q) = \{a | P(A \leq a) \geq q\}, \quad for \ \ q \in [0,1]$$

TÉCNICO LISBOA

DATA DISTRIBUTION

Univariate analysis

Central Tendency

- **Mean** (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean
  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Median**:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Central Tendency

- **Mean** (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean
  - Trimmed mean: chopping extreme v

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Median**:
  - Middle value if
    middle tw
  - Estimated

$$med$$

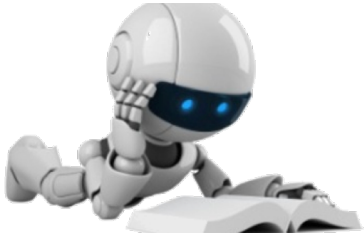
- **Mode**
  - Value that occur
  - Unimodal, bimoda
  - Empirical formula:

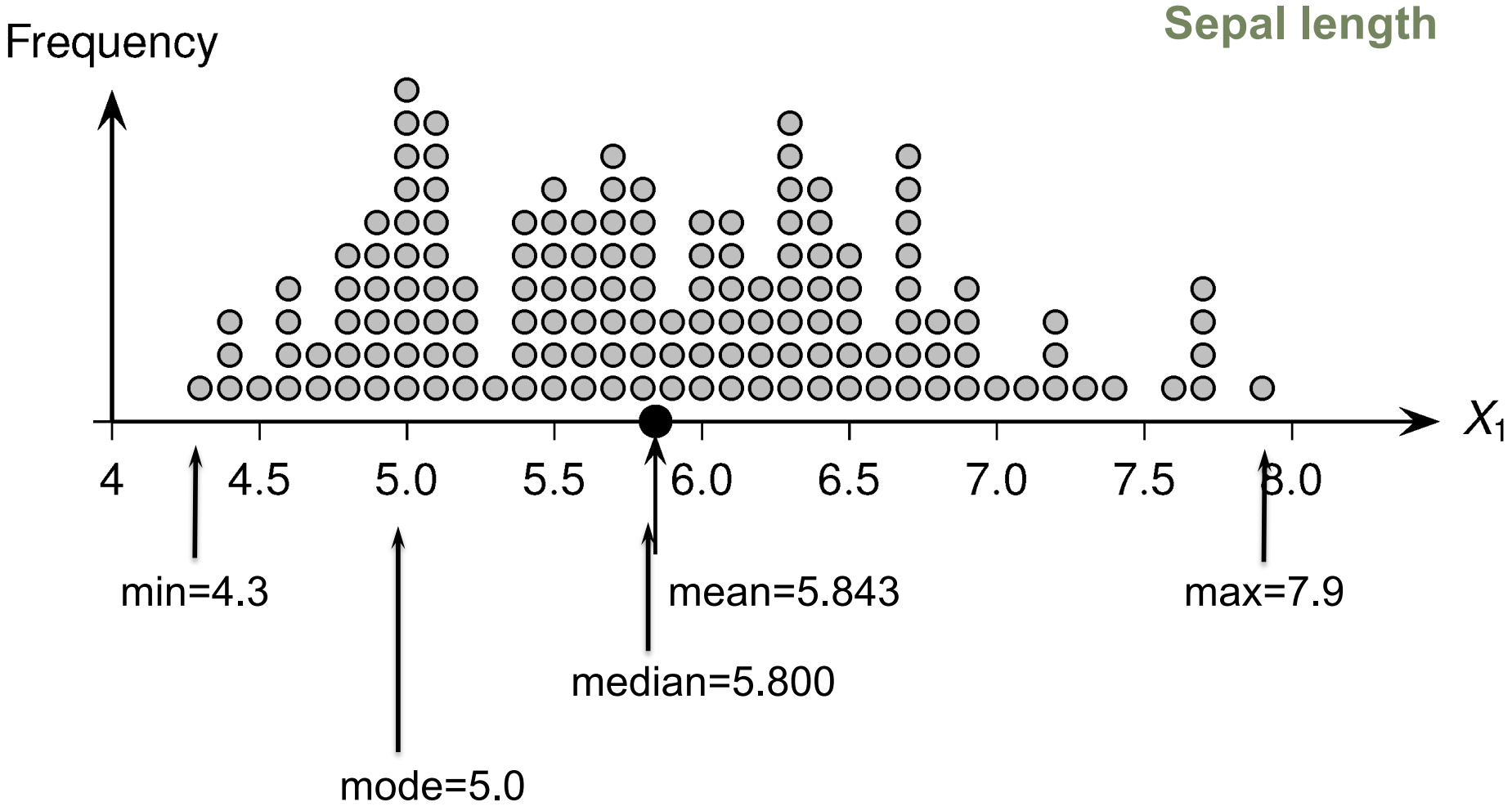
$$mean - mode = 3 \times (mean - median)$$

The median is robuster, because its value is not distorted by outliers

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

IRIS EXAMPLE





Measuring Dispersion

Variance and standard deviation

- (*sample: s, population: σ*)
- **Standard deviation** σ is the square root of variance σ^2

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\,2} - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

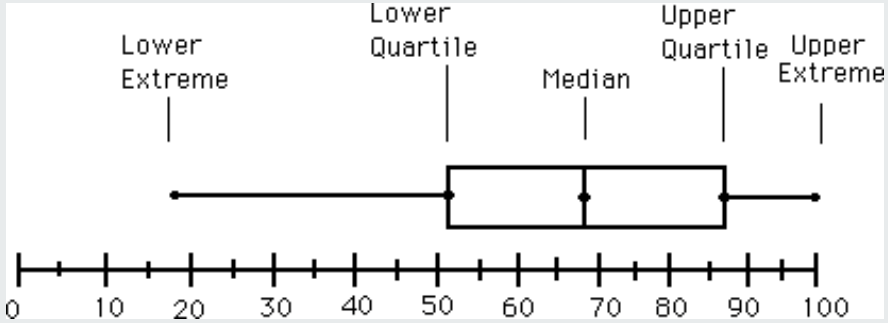
Quartiles

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: **$IQR = Q_3 - Q_1$**

BOXPLOT ANALYSIS

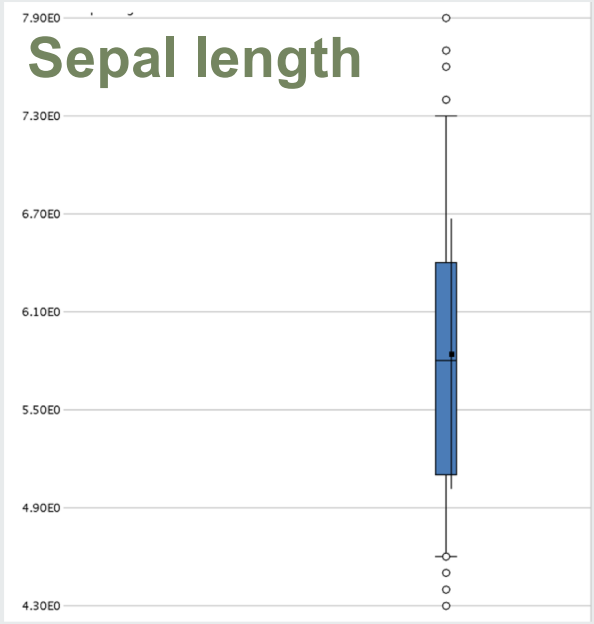
Five-number summary of a distribution

- Min, Q1, Median, Q3, Max

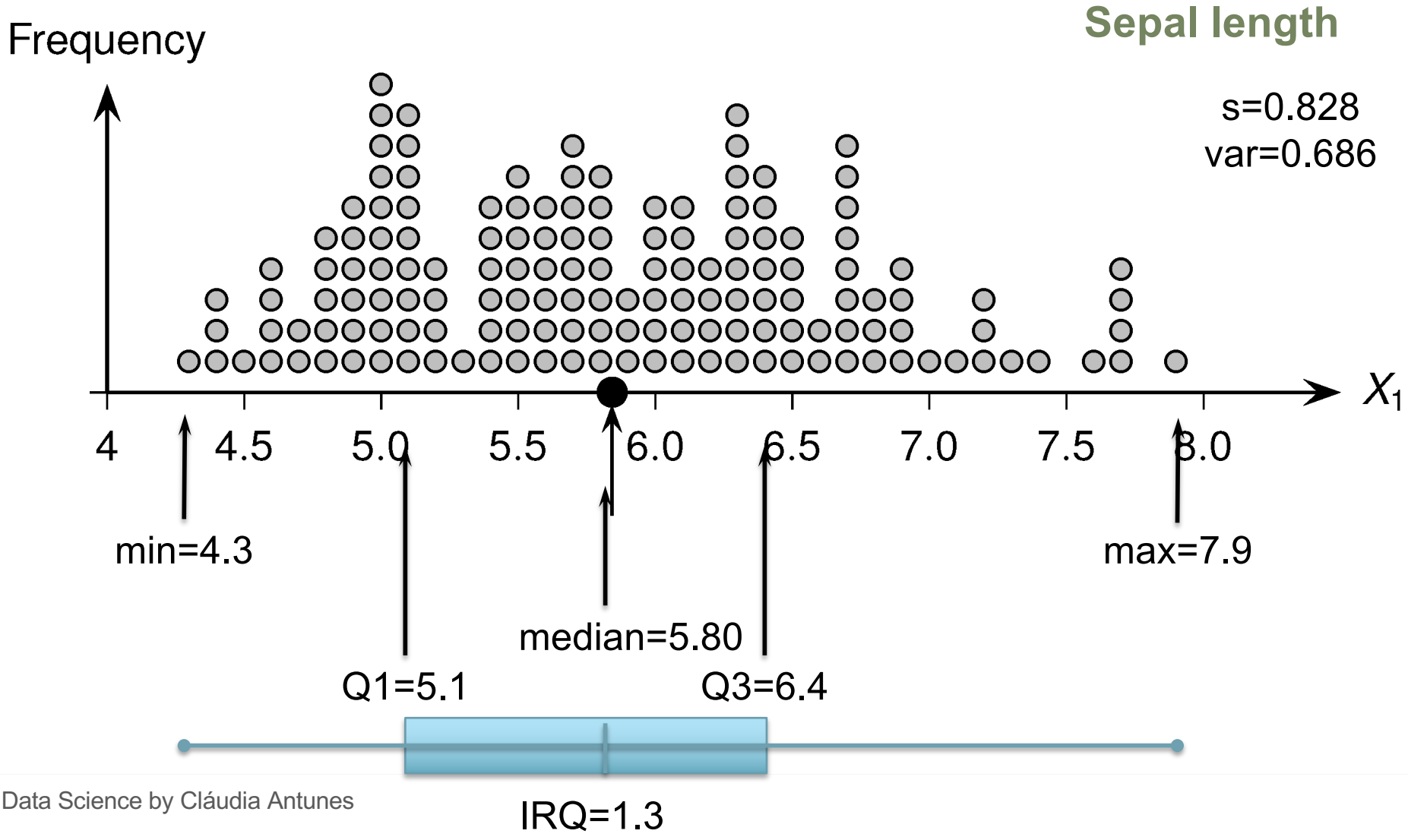


Boxplot

- Data is represented with a box
- The ends of the box are at **Q1** and **Q3**
- The **median** is marked by a line within the box
- Whiskers: two lines outside the box extended to **Min** and **Max**



IRIS EXAMPLE



OUTLIERS

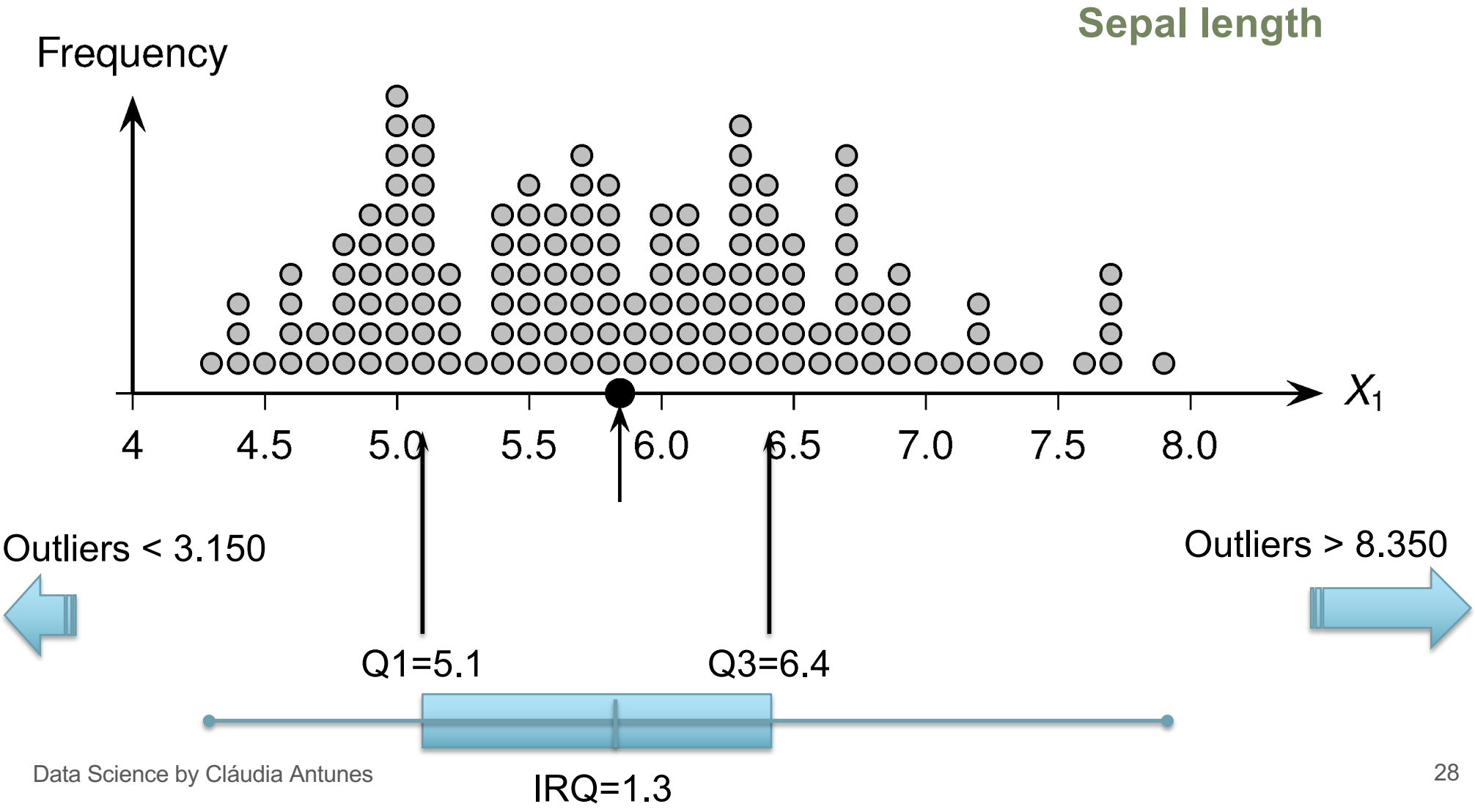
Outliers

- values that are very large or small and very uncommon;
- points beyond a specified threshold

If is a value lower than **Q1 - 1.5 x IQR**

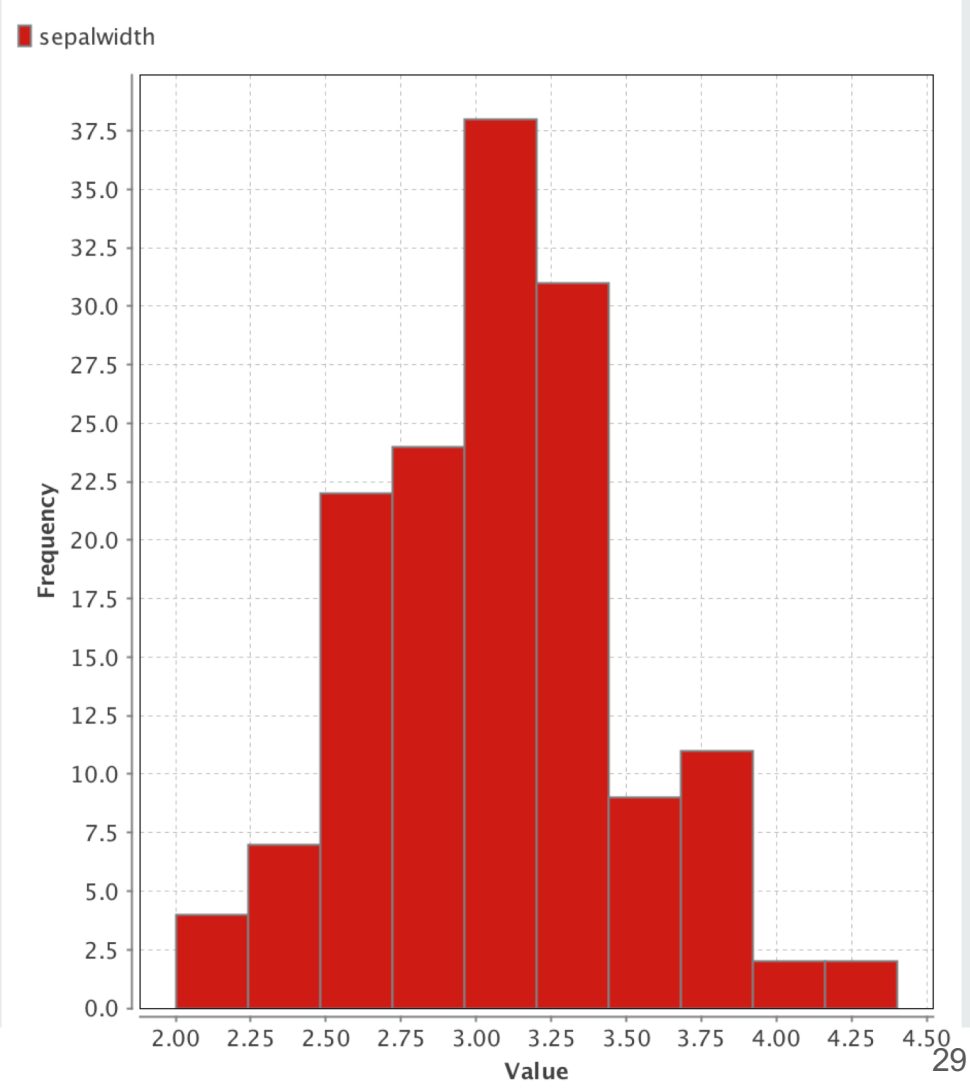
If is a value higher than **Q3 + 1.5 x IQR**

IRIS EXAMPLE



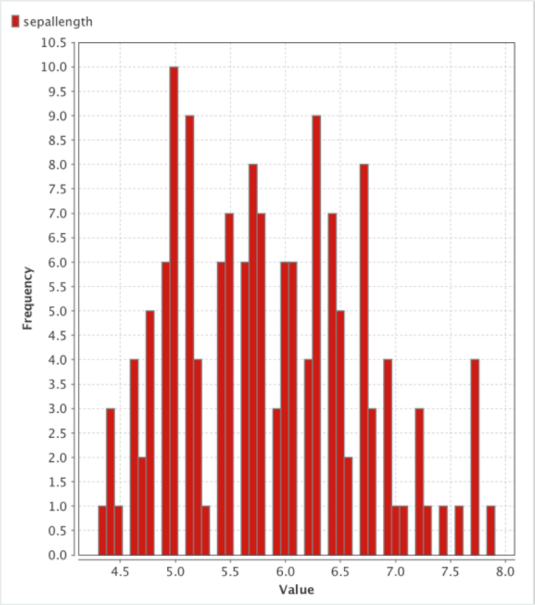
DISTRIBUTION

- A common visualization is the **frequency histogram**
  - It plots the relative frequencies of values in the distribution
- To construct a histogram
  - Divide the range between the **highest** and **lowest** values into **equal size bins**
  - Toss each value in the appropriate bin
  - Each bar in the histogram represents the number of values in the corresponding bin

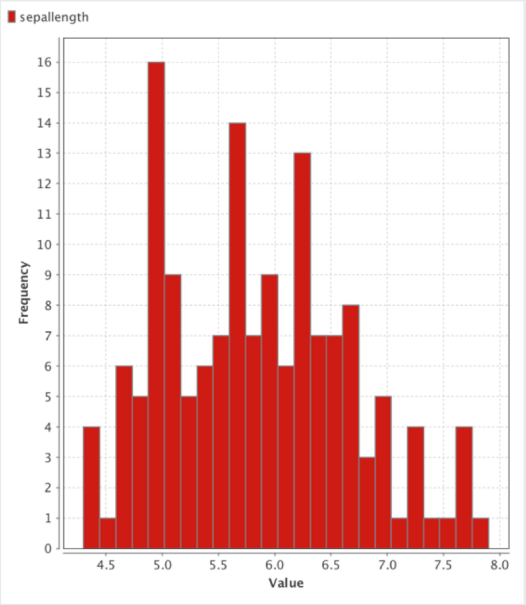


HISTOGRAMS

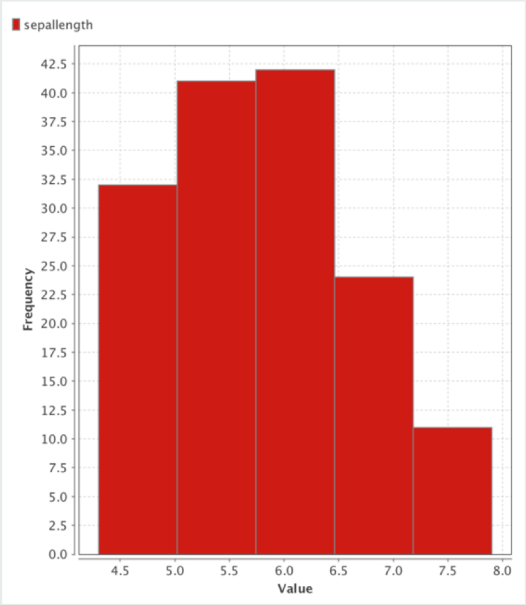
- The choice of bin size affects:
  - The details we see in the frequency histogram
    - Changing the bin size to a lower number illuminates things that were previously not seen
  - The one‘s perception of the shape of distribution



50 bins

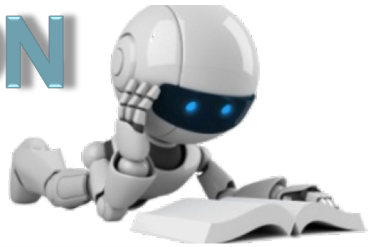


25 bins

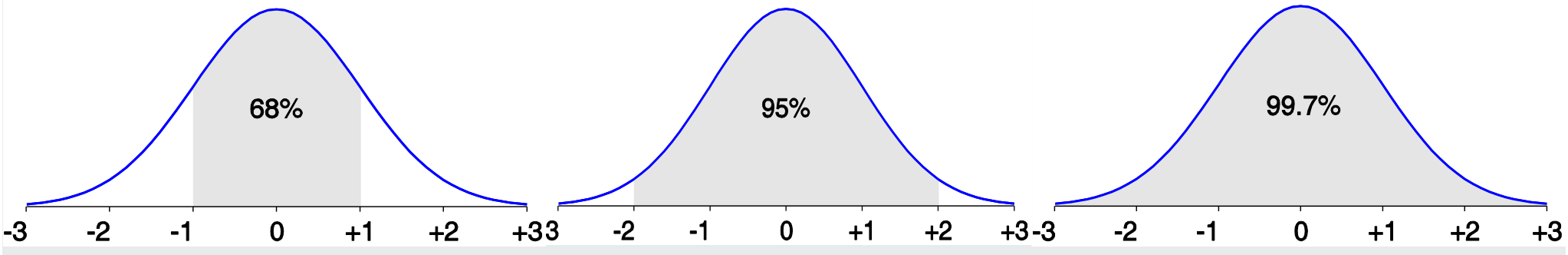


5 bins

Normal Distribution Properties

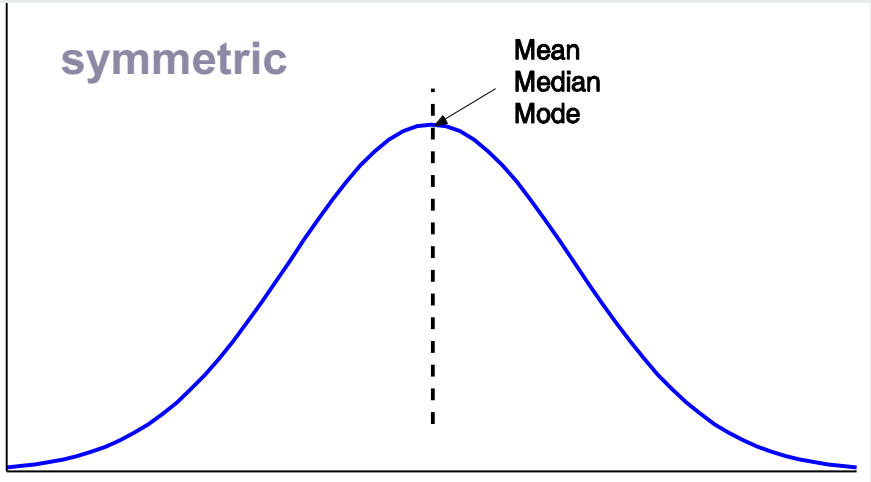


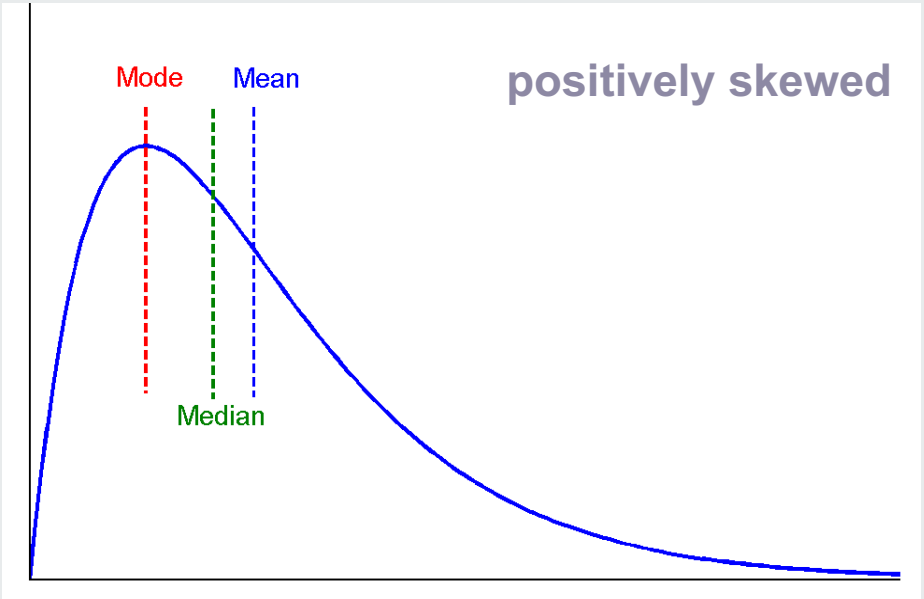
- The normal curve $N\sim(\mu, \sigma^2)$
  - **$[\mu-\sigma, \mu+\sigma]$** contains about 68% of the records
  - **$[\mu-2\sigma, \mu+2\sigma]$** contains about 95%
  - **$[\mu-3\sigma, \mu+3\sigma]$** contains about **99.7%**

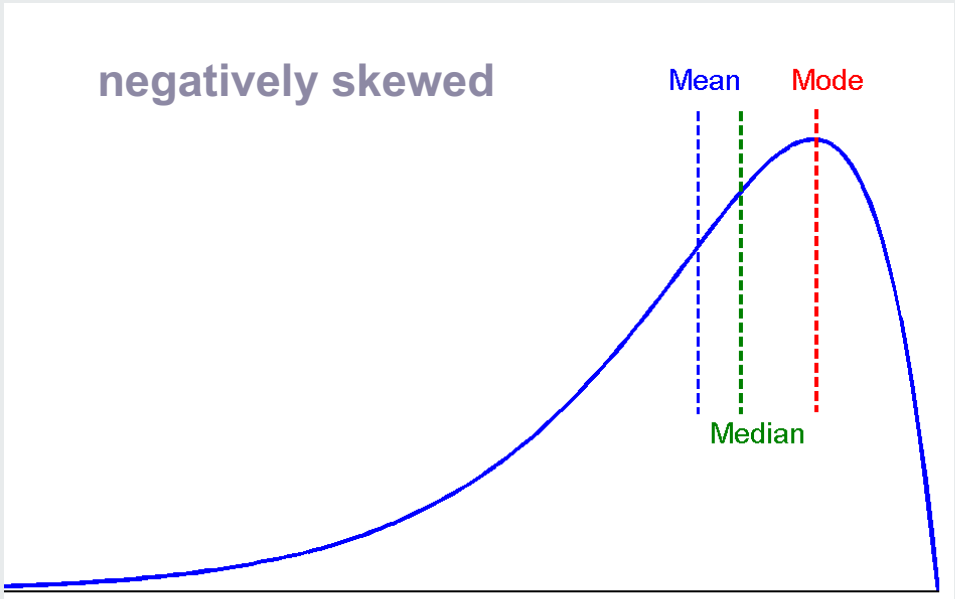


SYMMETRIC X SKEWED

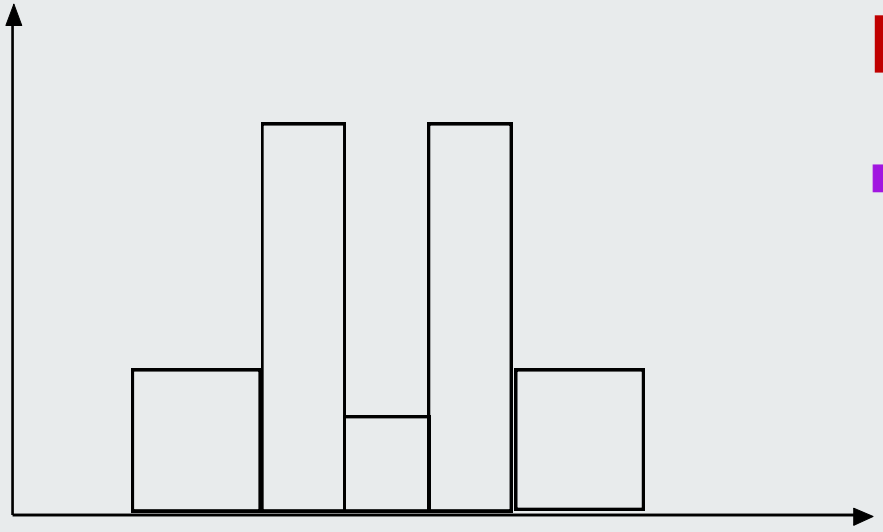
- Median, mean and mode of symmetric, positively and negatively skewed data

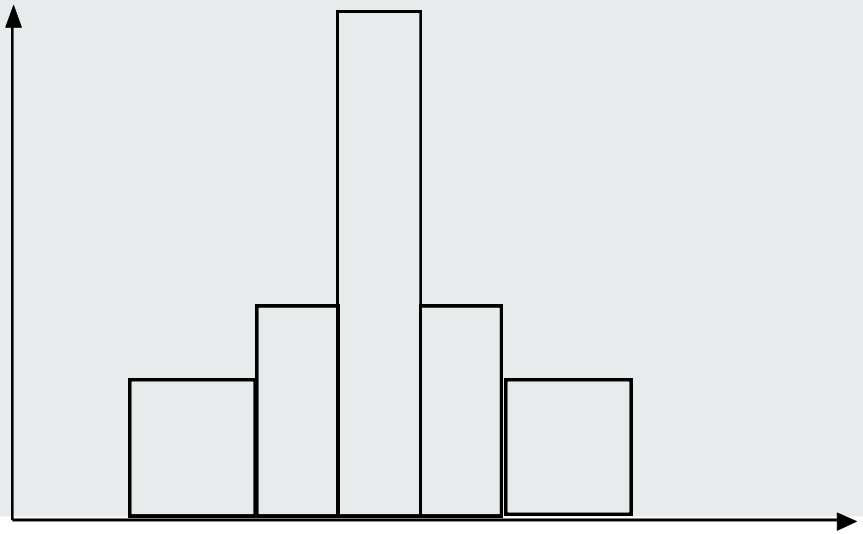






HISTOGRAMS X BOXPLOTS





BE CAREFUL:

- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

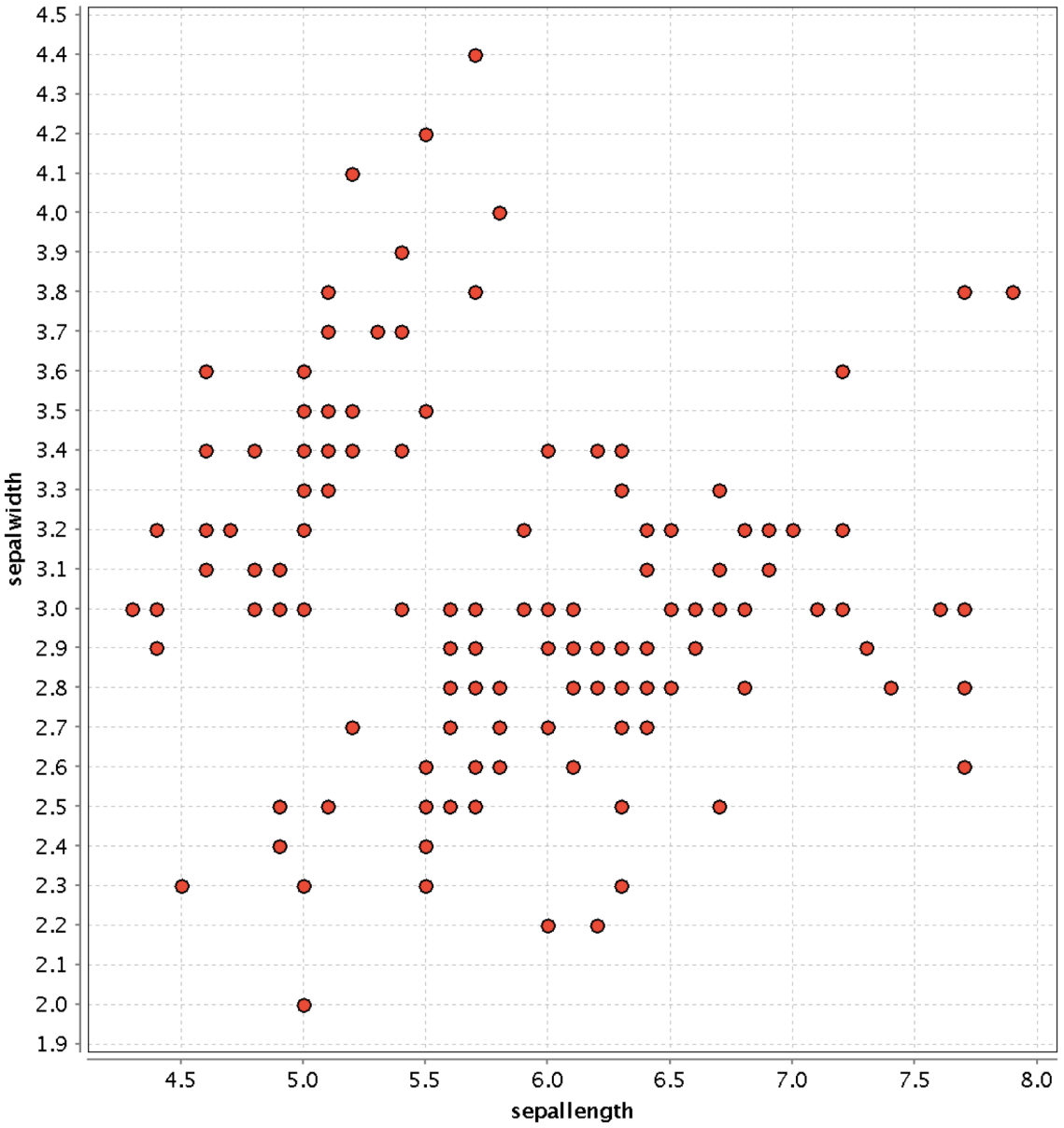
TÉCNICO LISBOA

DATA SPARSITY

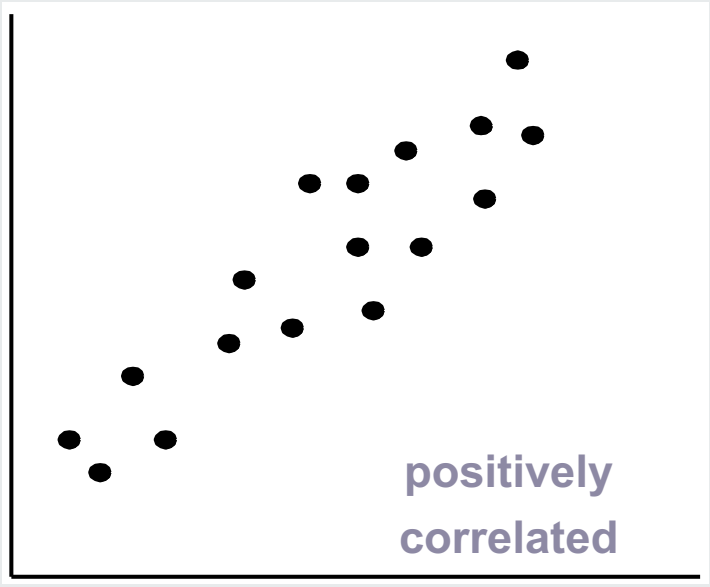
Multivariate analysis

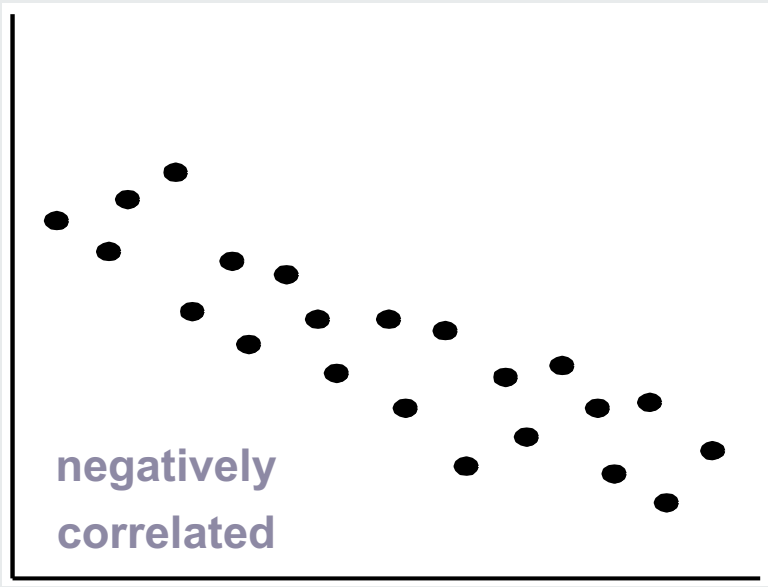
SCATTER PLOT

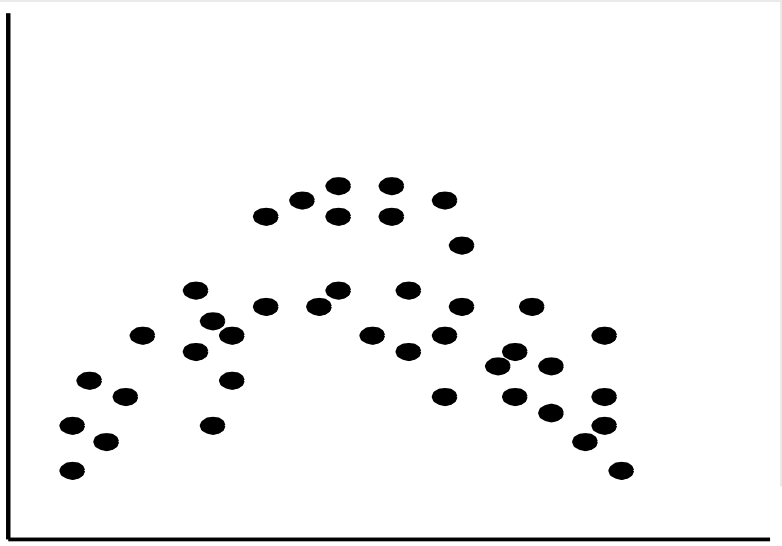
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



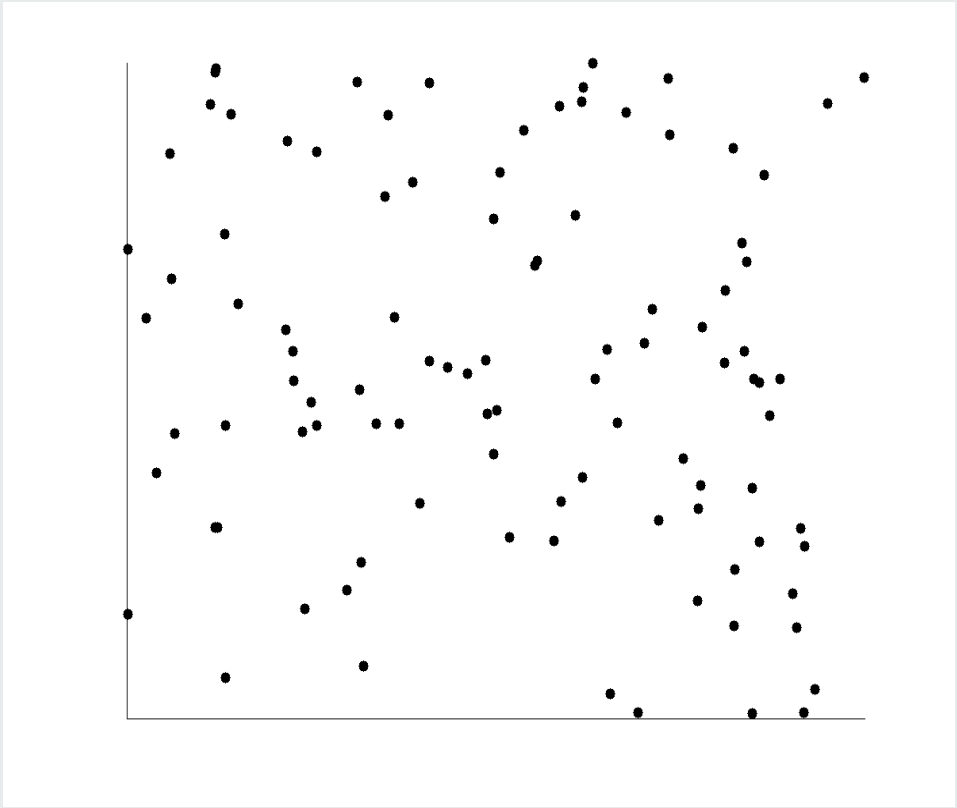
Positively and Negatively Correlated Data

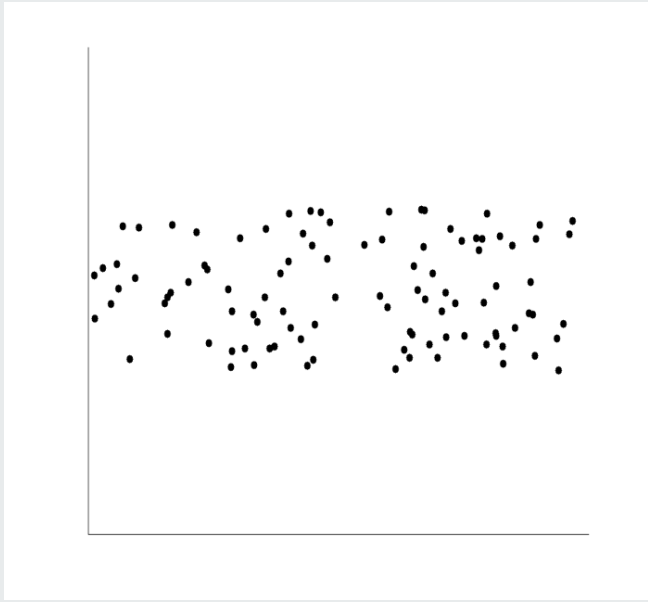


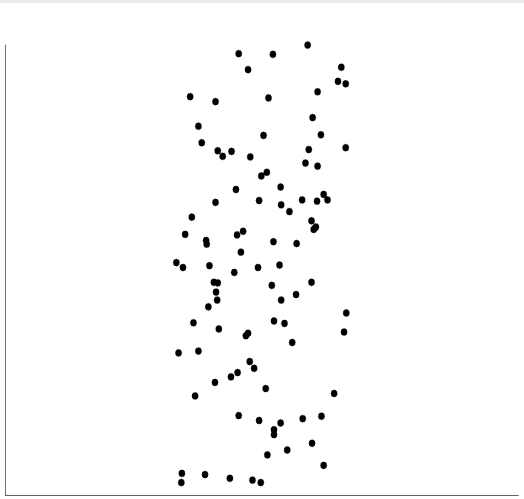


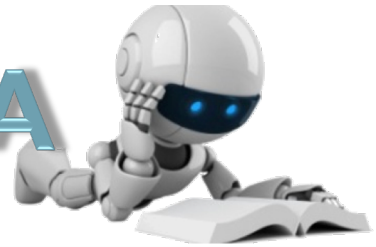


Uncorrelated Data









TÉCNICO LISBOA

DIMENSIONALITY

DIMENSIONALITY

- Usually, it corresponds to **the number of attributes** describing the data (***extrinsic dimensionality***)

$$dim(D)=d$$

- However, it can be lower, say k, when some of the attributes are not linearly independent

$$k<<d$$

Data Reduction

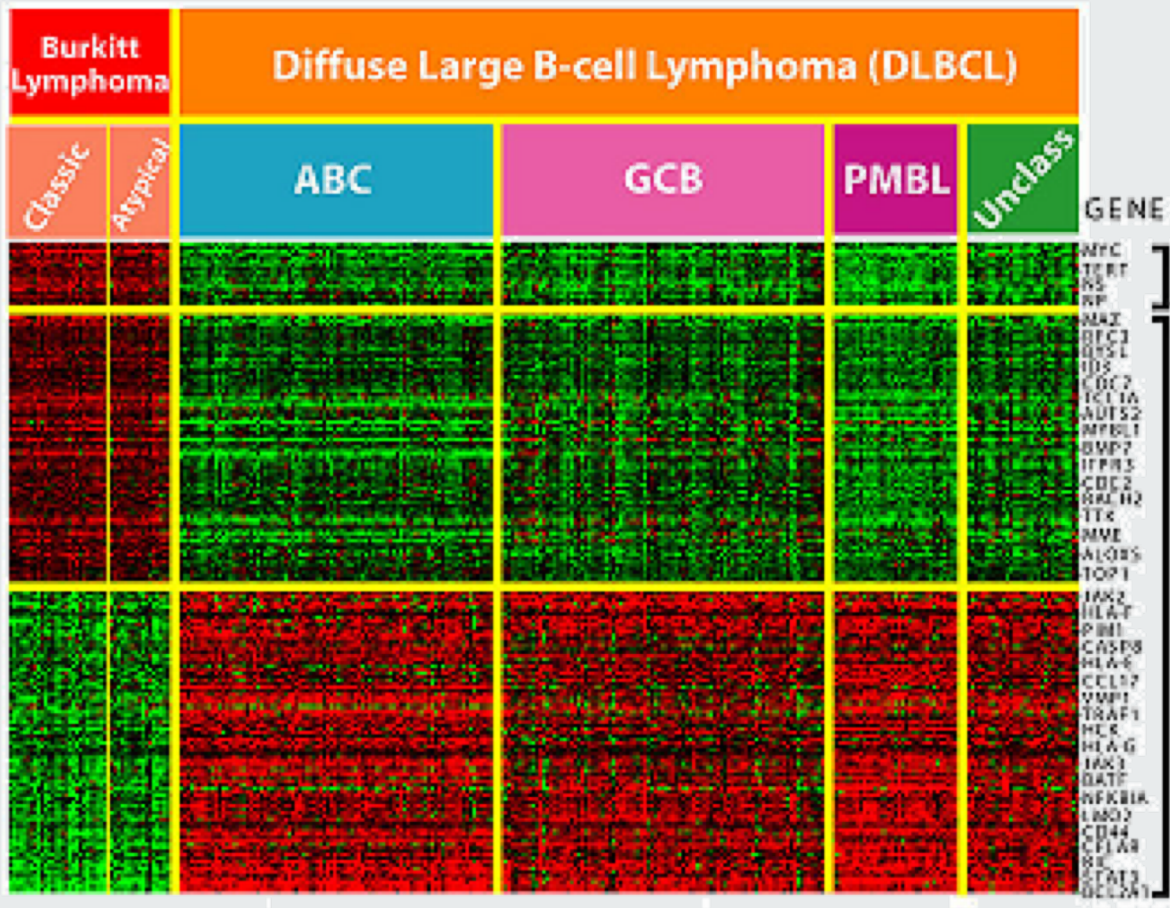
k is called the ***intrinsic dimensionality*** of the data.

Issues

High Dimensionality

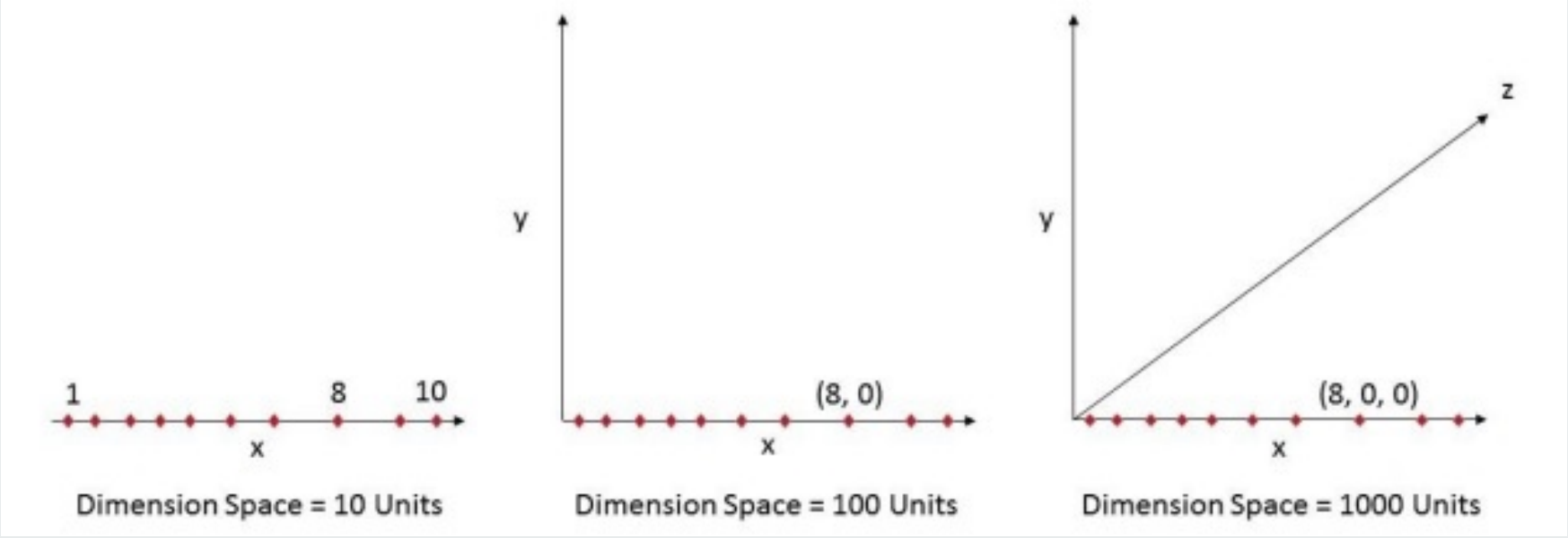
Sometimes, dimensionality d is much larger than the number of records n

$$n << d$$



Issues

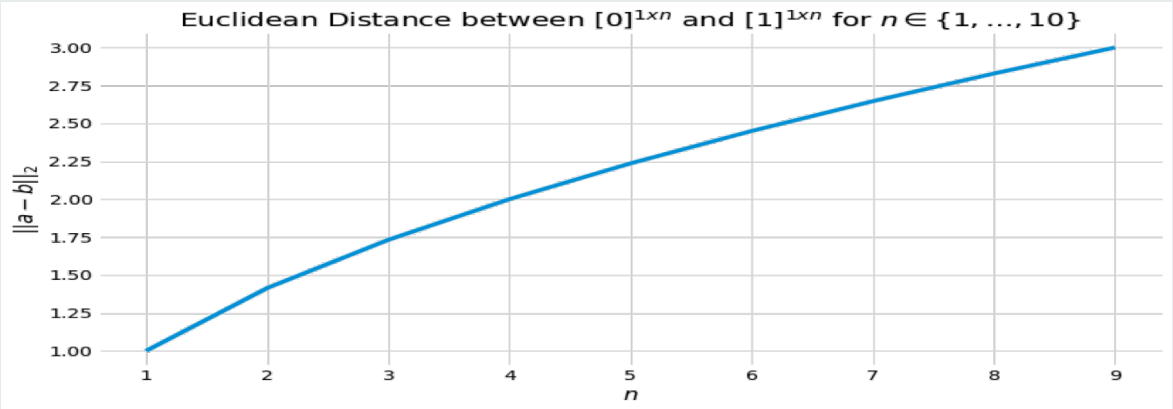
- Points in high-dimensional space are **highly sparse**



$$|1-0| = 1$$
$$\|(1,1)-(0,0)\| = \sqrt{2}$$
$$\|(1,1,1)-(0,0,0)\| = \sqrt{3}$$

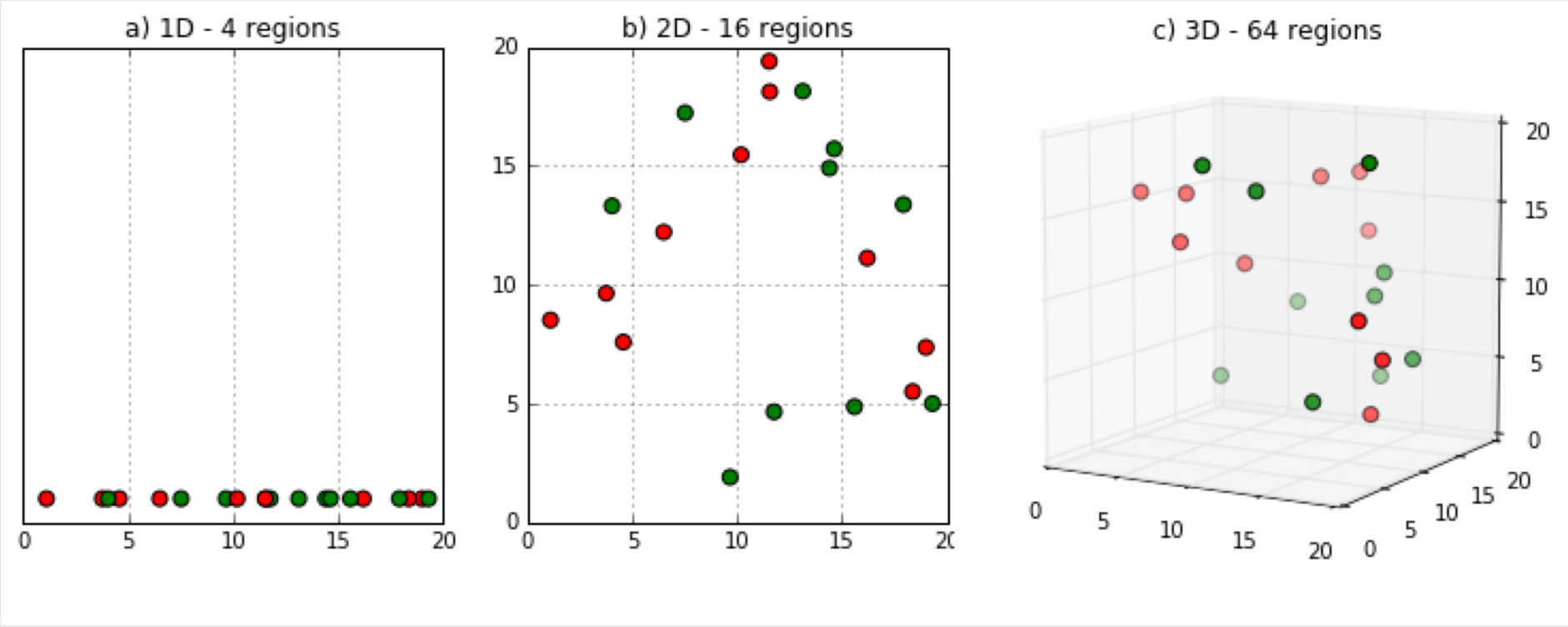
...

$$\|[\![1]\!]^{1\times 10} - [\![1]\!]^{1\times 10}\| = 3$$



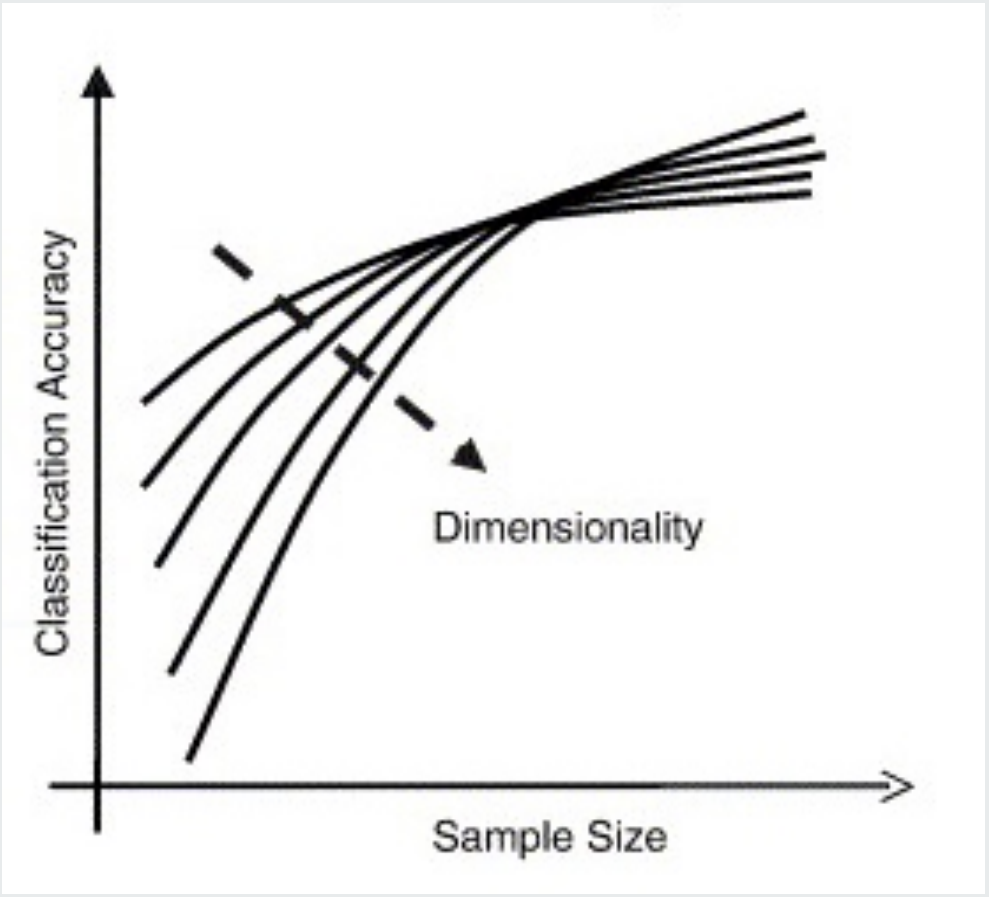
ISSUES

- *"As the number of attributes (dimensions) grows, the amount of data to constitute a sample grows exponentially."*



CURSE OF DIMENSIONALITY

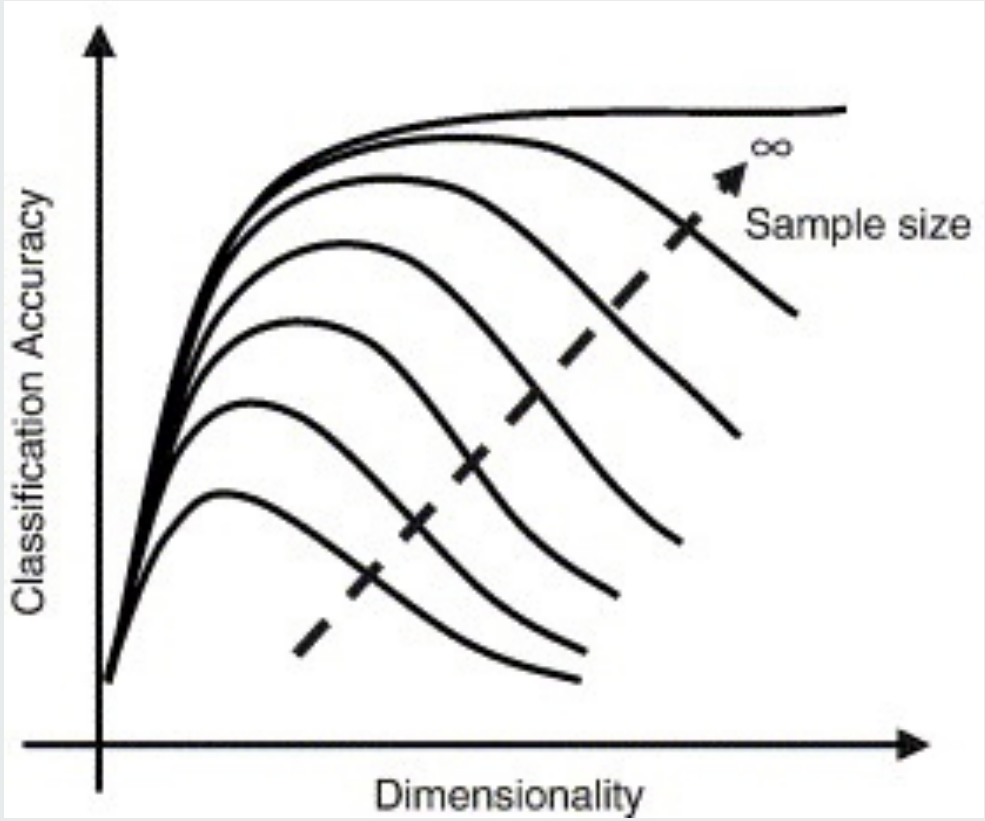
✦ *"As **the number of attributes (dimensions) grows**, the amount of data we need to generalize accurately **grows exponentially**."*



Larger the dimensionality ➔ larger the sample

HUGHES PHENOMENON

- Given a fixed number of records, **the predictive power** of a model **increases** with **the increase of the number of dimensions** until it reaches a maximum value, but then **it starts decreasing**



TÉCNICO LISBOA