Web Intelligence
Similarity Measures

Claudio Lucchese   claudio.lucchese@unive.it

Previously …

- Similarity is a key ingredient of Recommender Systems
  - And of several other Web Mining Tools …
- There are several similarity measures we can chose from
- The best depends on the specific task
  - We might need to design a new measure for new task
    - Are the vector pairs below equally similar ?

1 1 1 1 1 1 1 1 1 1 1 0		1 0 0 0 0 0 0 0 0 0 0 0
0 1 1 1 1 1 1 1 1 1 1 1	vs	0 0 0 0 0 0 0 0 0 0 0 1

- We will see that dimensionality is an issue

Minkowski distances

- Given two ***N***-dimensional objects X and Y, where $X = [x_0, \ldots, x_{N\text{-}1}]$ and $Y = [y_0, \ldots, y_{N\text{-}1}]$

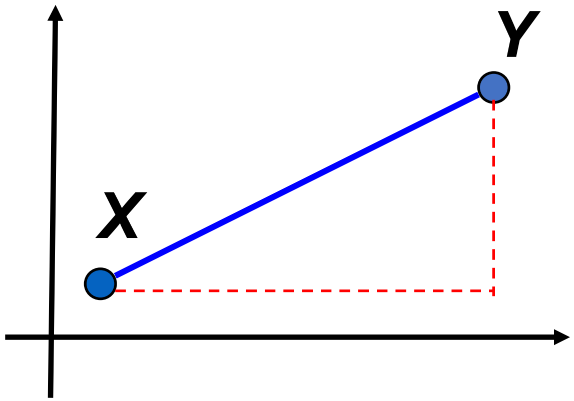
$$
\begin{aligned}
d(X,Y) &= \sqrt[q]{|x_0 - y_0|^q + |x_1 - y_1|^q + \ldots + |x_{N-1} - y_{N-1}|^q} \\
&= \sqrt[q]{\sum_{0 \le i < N} |x_i - y_i|^q}
\end{aligned}
$$

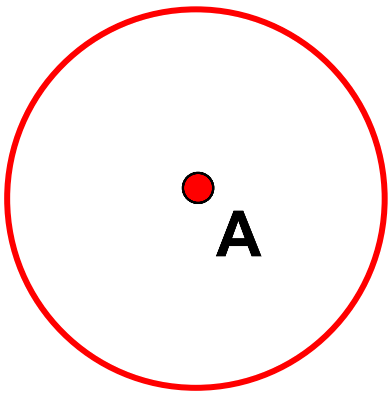
Euclidean Distance

- If ***q=2*** **,** ***L_2*** norm or ***Euclidean Distance***:

$$d(X,Y) = \sqrt[2]{|x_0 - y_0|^2 + |x_1 - y_1|^2 + \ldots + |x_{N-1} - y_{N-1}|^2}$$

- It defines a metric space:
  - d(X,Y) >= 0 *(Positivity)*
  - d(X,Y) = d(Y,X) *(Symmetry)*
  - d(X,Y) <= d(X,Z) + d(Z,Y) *(Triangular inequality)*



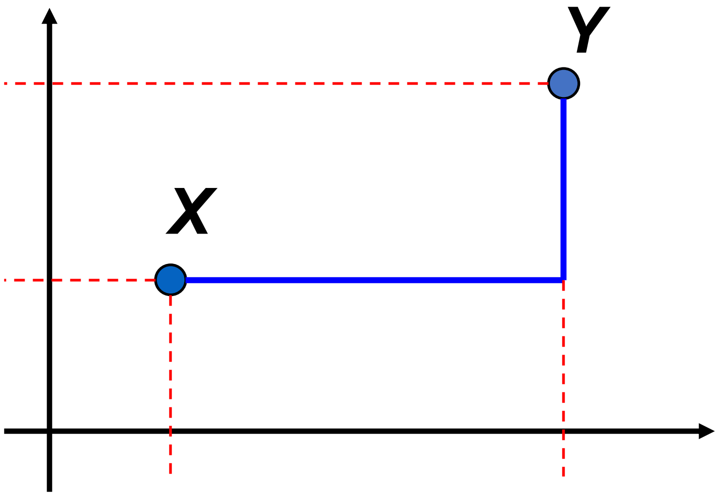


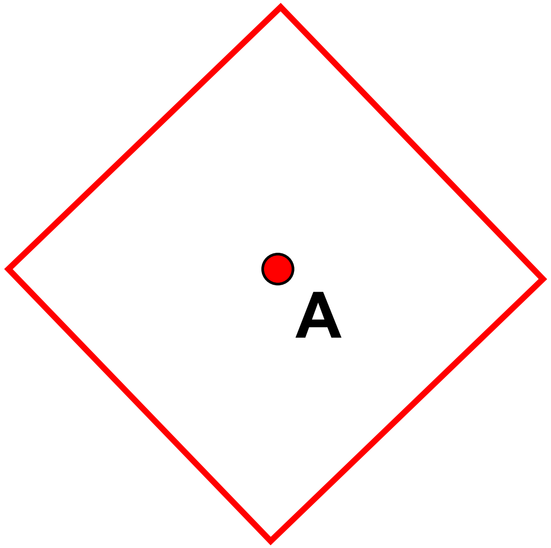
Same distance from A

Manhattan Distance

- If $q=1$, L_1 norm or ***Manhattan*** or ***City-Block***:

$$d(X,Y) = |x_0 - y_0| + |x_1 - y_1| + \ldots + |x_{N-1} - y_{N-1}|$$



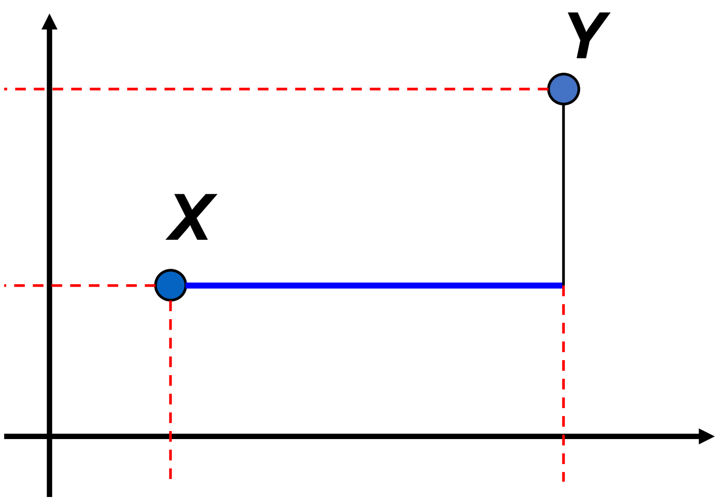


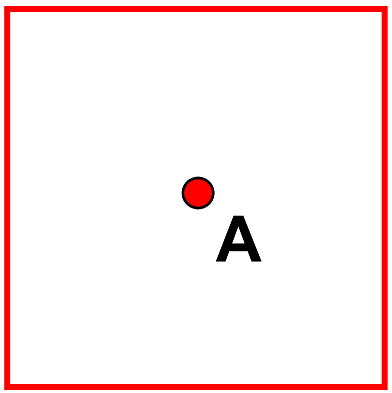
Same distance from A

Chebyshev Distance

- If ***q →∞***, ***L_∞*** norm or ***Chebyshev***, ***Chessboard Distance***:

$$d(X,Y) = \max_i |x_i - y_i|$$





same distance from A

Chessboard Distance

Binary Vectors

- Document representation
  - Document $X = [0,1,0,1,0,1,0,1,1,1]$
  - Document $Y = [1,0,1,1,1,0,1,0,1,1]$
  - $X[i] = 1$ *iff* the i-th term occurs in X
- Contingency table:

		X		
		1	0	*sum*
Y	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

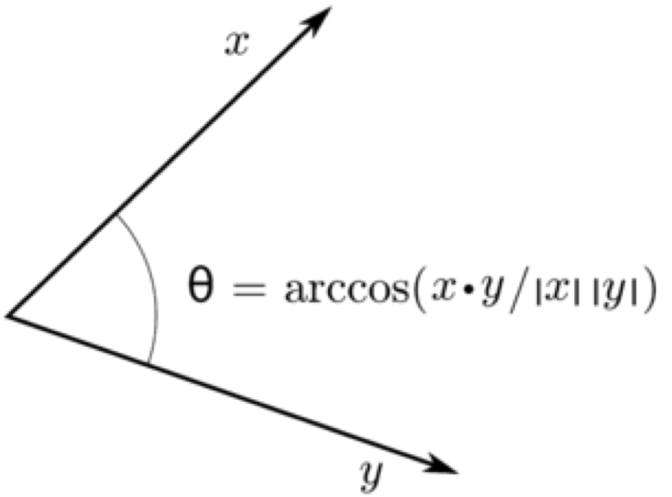
- ***Jaccard***: $X \cap Y/(X \cup Y)$ or $q/(q+r+s)$
- ***Simple matching***: $(q+t)/p$
- Entries of the contingency table can be weighted as needed

Cosine Similarity

- Document representation
  - Document $X = [0,0,0,3,0,5,0,14,7,9]$
  - Document $Y = [1,0,2,2,4,0,10,0,3,11]$
  - *X[i]* is the number of times the *i*-th term occurs in X

$$\cos(X,Y) = \frac{X \cdot Y}{\|X\|_2 \|Y\|_2}$$

$$= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$



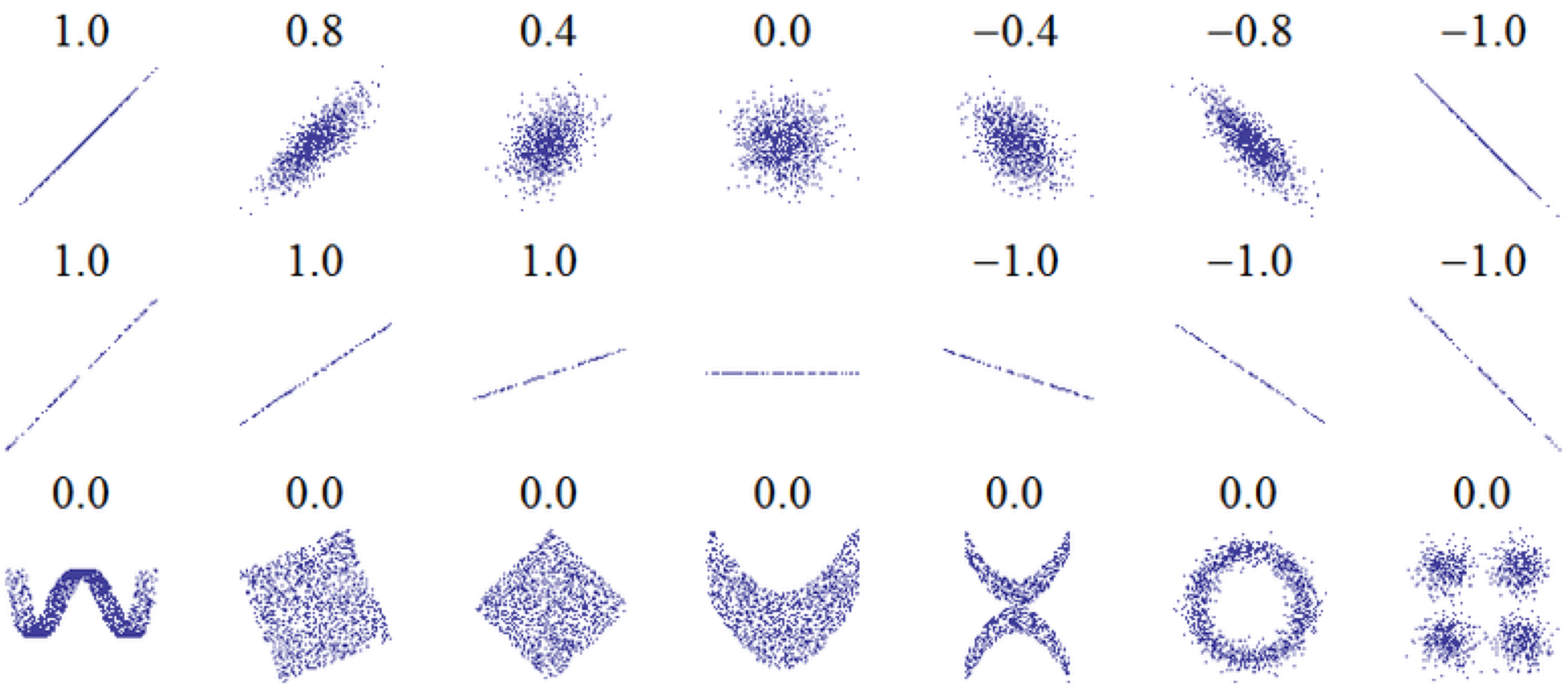
Pearson Correlation

- Linear dependency between variables
  - Does X increase when Y increases ?
  - Is there any correlation between income and degree ?
- Standardize and multiply:

$$X_i = \frac{x_i - \overline{X}}{\sqrt{\sum(x_i - \overline{X})^2}} \qquad Y_i = \frac{y_i - \overline{Y}}{\sqrt{\sum(y_i - \overline{Y})^2}}$$

$$\rho(X, Y) = \sum X_i Y_i$$

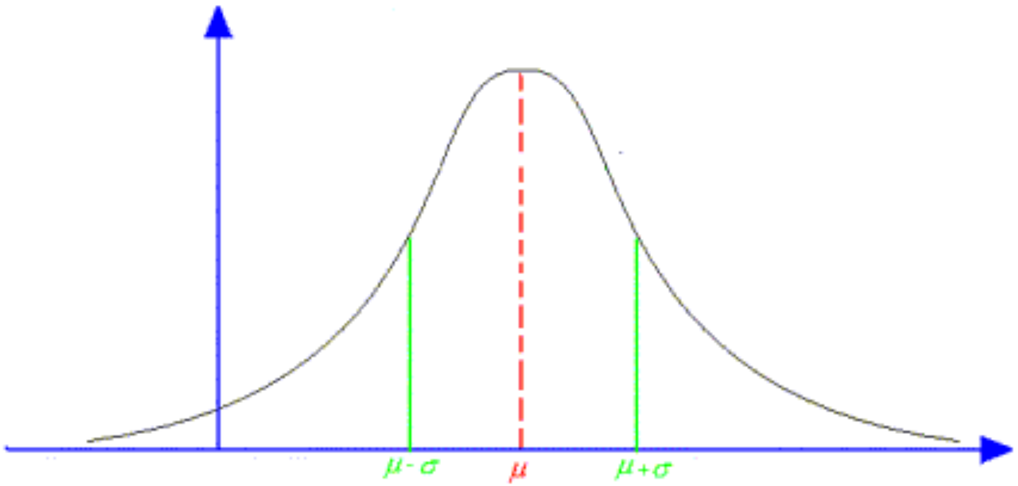
Correlation plots



Standardization ?

- Most data has Gaussian Distribution:
  - Average height of people
  - Error in measurements
  - [Central Limit Theorem]

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



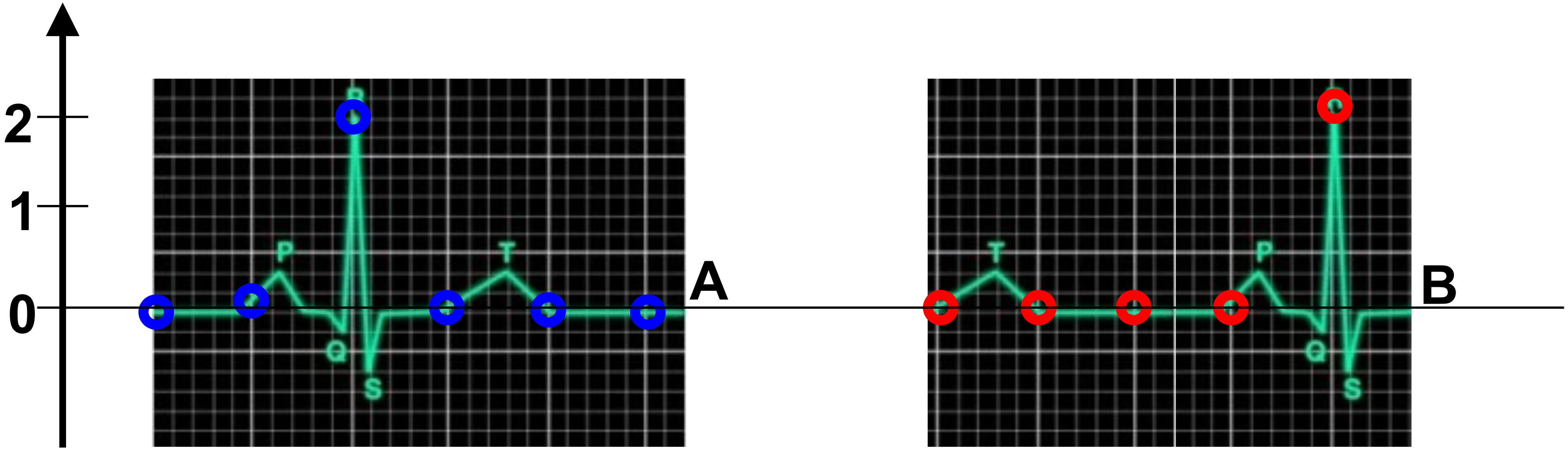
- **68.2%** of points is at distance at most **one** standard deviation from the mean
- **95,5%** of points is at distance at most **two** standard deviation from the mean

- Standardization is a normalization technique under Gaussian Distribution assumption
  - Set mean to 0, set variance to 1

$$X_i = \frac{x_i - \overline{X}}{\sqrt{\sum(x_i - \overline{X})^2}}$$

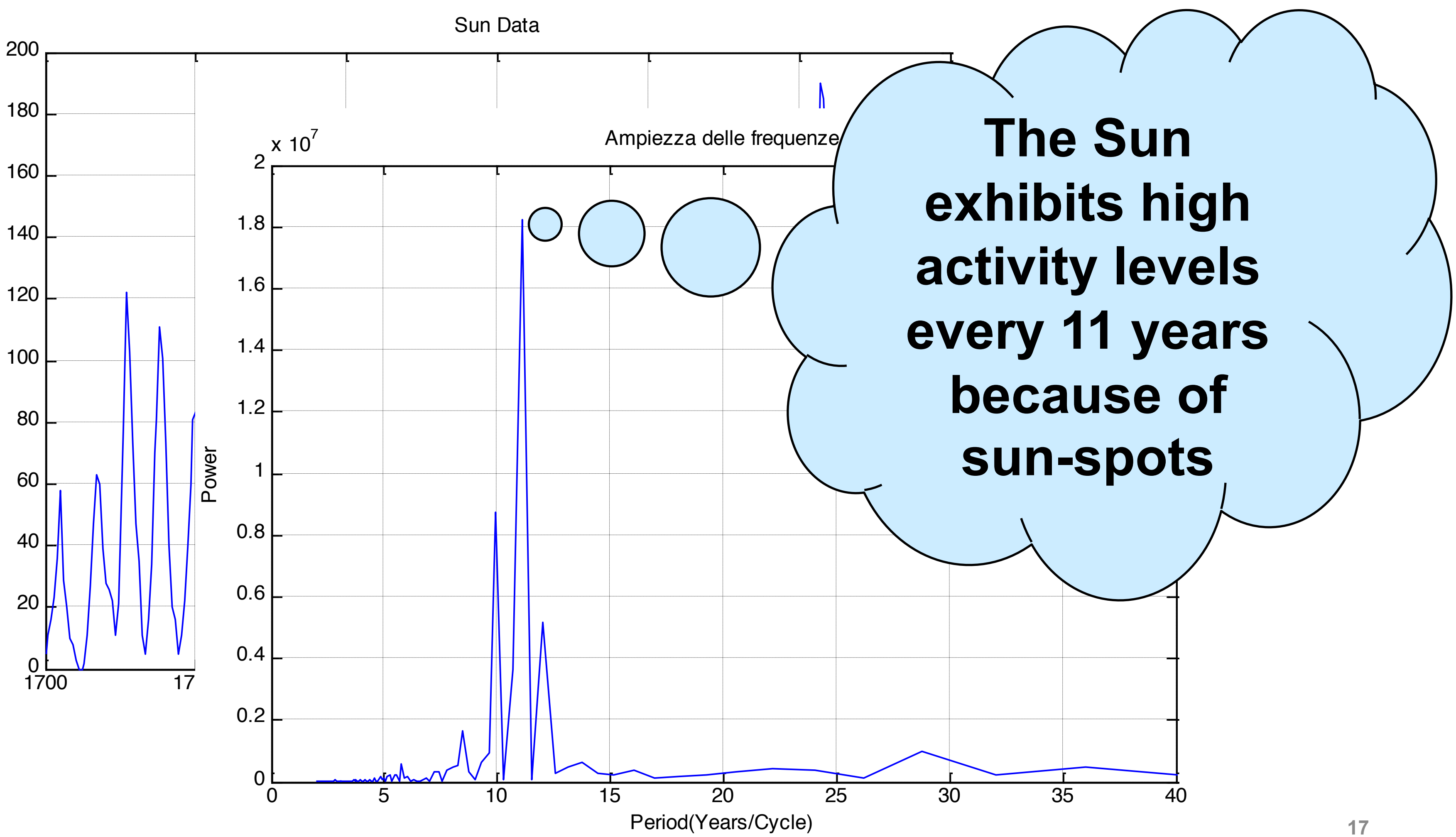
Similarity of time-series

- Examples:
  - ECG, stocks, temperatures, stars luminosity, etc.
- Euclidean Distance ?
  - Not robust against phase changes



$$d(A,B)=\sqrt{(0-0)^2+(0-0)^2+(2-0)^2+(0-0)^2+(0-2)^2}=\sqrt{4+4}=2.82$$

Periodic Distance



Periodic Distance

- Fourier Transform:
  - "Understands" the important ***frequencies*** in a signal, in terms of ***Amplitude*** and ***Phase***
    - *Amplitude_X = [100, 80, 70, 10, 0, 0, 0]*
    - *Amplitude_Y = [99, 80, 50, 20, 10, 0, 0]*
- Periodic Distance is Euclidean over amplitudes

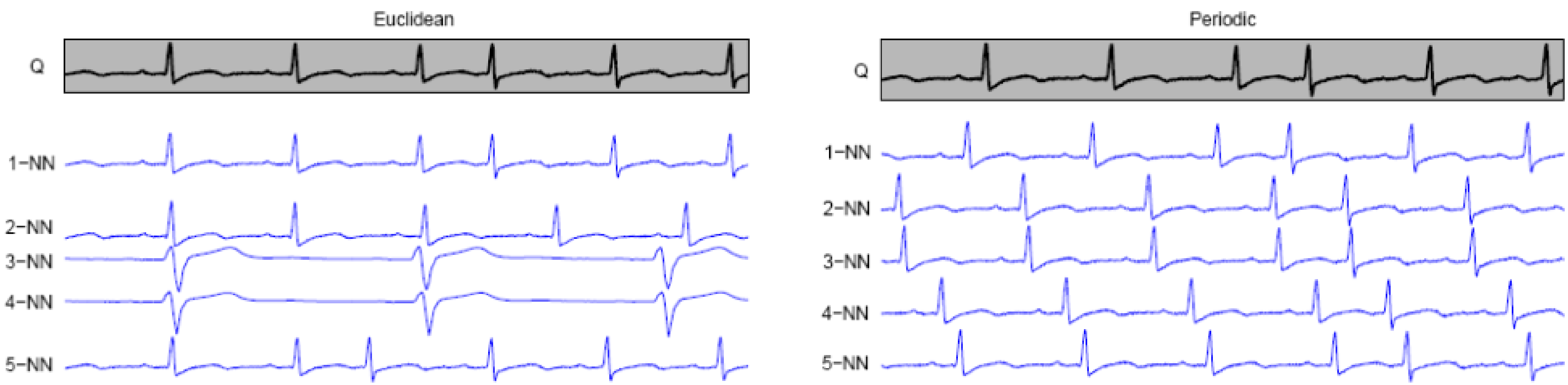


Fig. 1. 5-NN euclidean and periodic matches on an ECG dataset.

The curse of dimensionality

- Originally used to address optimization problems
  - Find the value of *x* that minimizes function *f*.

- Suppose you want to find the optimum value of $x \in \{1,2,3,4,5,6,7,8,9,10\}$.
  - Try every value and check the function to optimize.
- Suppose you have to variables $x,y \in \{1,2,3,4,5,6,7,8,9,10\}$.
  - You may need to try ***100*** cases:
    - $x=1$ & $y=1$, $x=1$ & $y=2$, $x=1$ & $y=3$, etc. etc.
- Suppose you have n such variables, the search space grows up to 10^n.

- Problems are considered intractable starting from $n=10$

Not only optimization problems

- Anytime you have objects with a large number of attributes (variables)

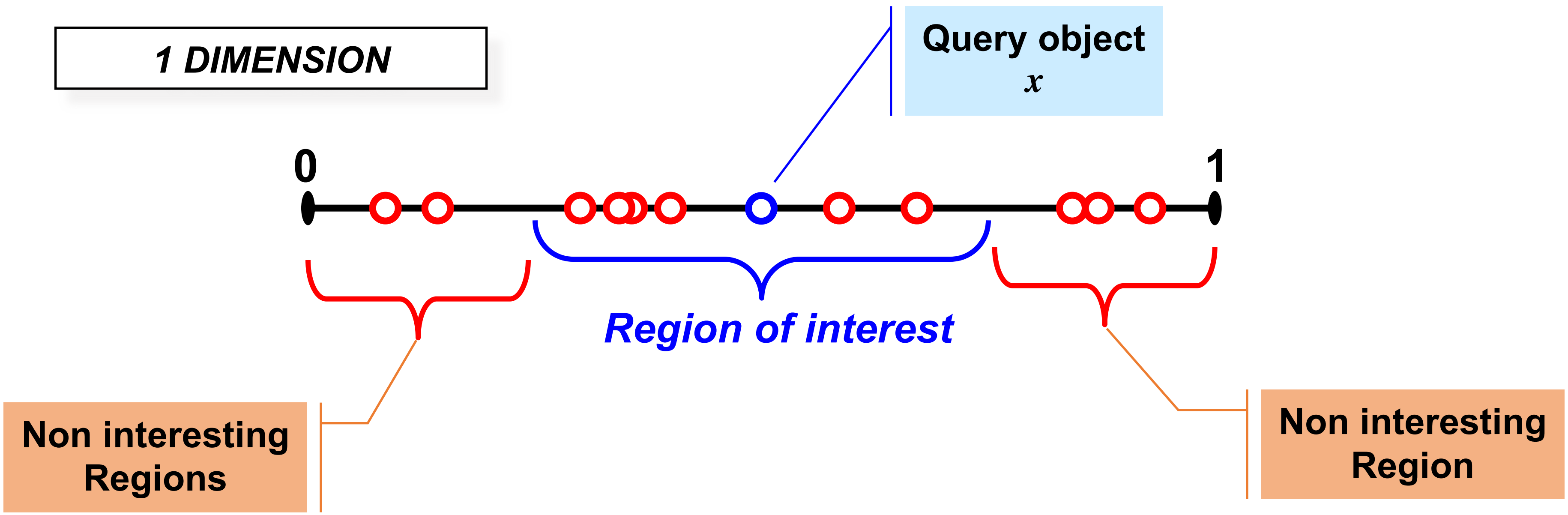
- In our case:
  - Objects are documents
  - Variables are term occurrence counts
  - Minimize similarity

Similarity

- Suppose objects are identically independently distributed at random in the (search) space
- Every dimension has values in the interval $[0, 1]$
- Find objects at distance <0.25 from x.

Similarity

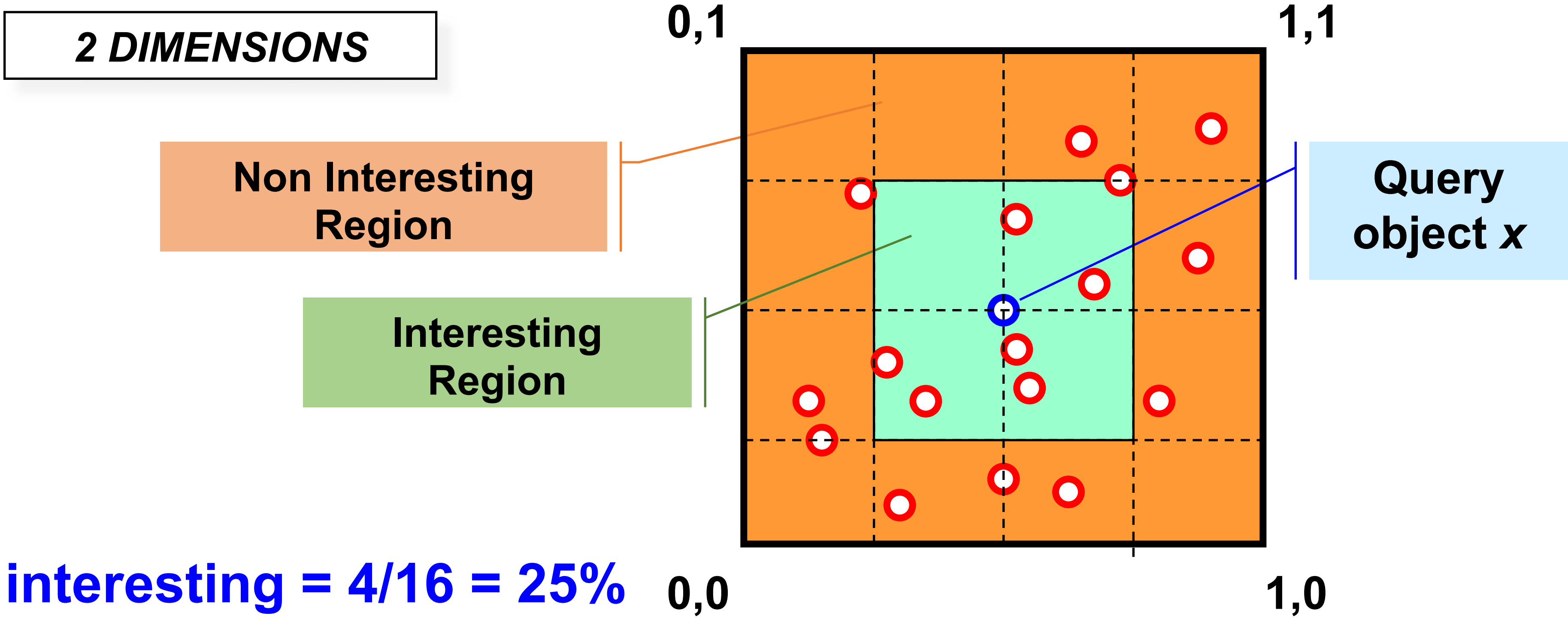
- Suppose objects are identically independently distributed at random in the (search) space
- Every dimension has values in the interval $[0, 1]$
- Find objects at distance <0.25 from x.



Interesting space = ½ = 50%

Similarity

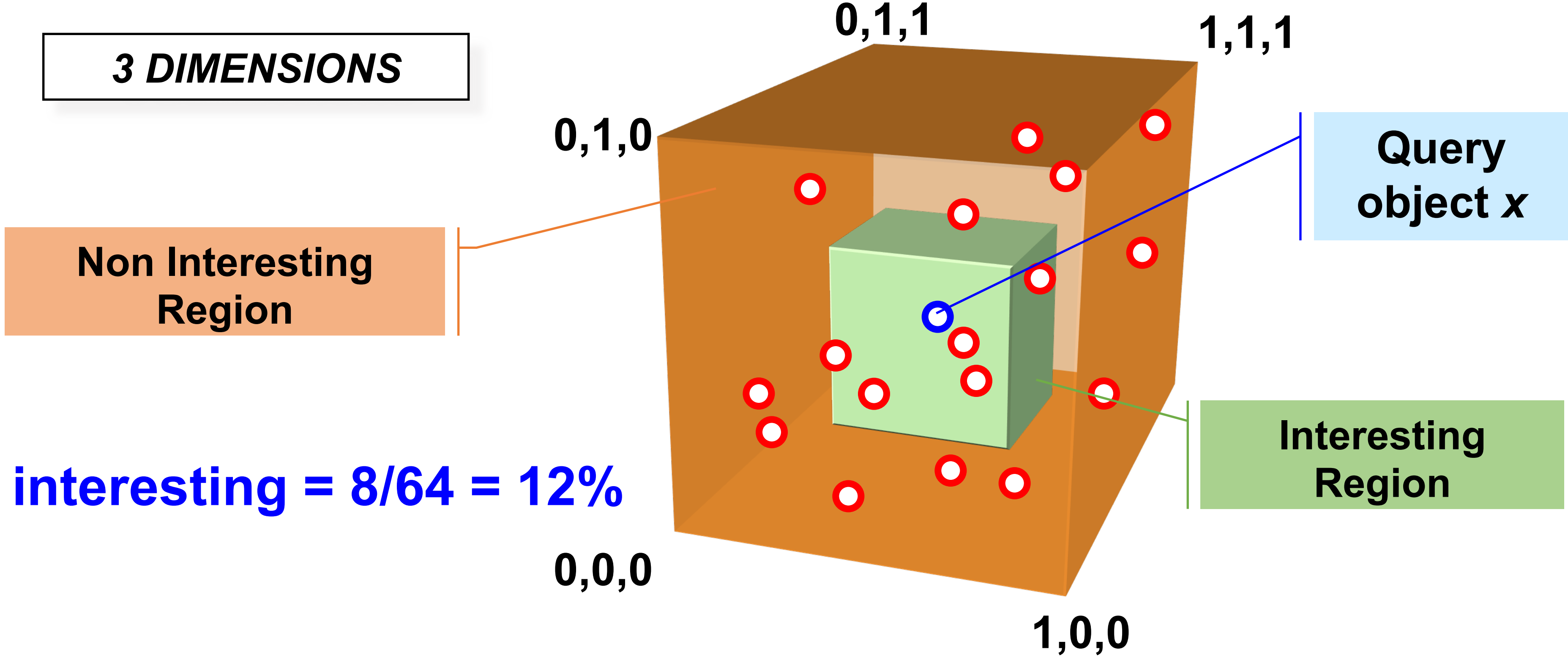
- Suppose objects are identically independently distributed at random in the (search) space
- Every dimension has values in the interval *[0, 1]*
- Find objects at distance <0.25 from x.



interesting = 4/16 = 25%

Similarity

- Suppose objects are identically independently distributed at random in the (search) space
- Every dimension has values in the interval *[0, 1]*
- Find objects at distance <0.25 from x.



What does it mean ?

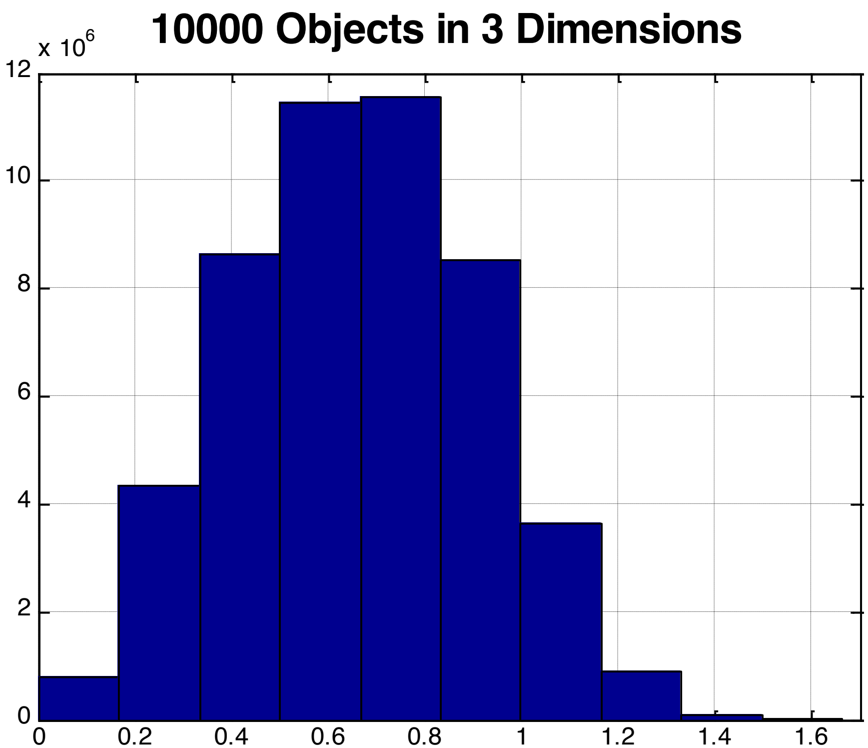
- The region of interest halves when increasing the number of dimensions
  - 50%, 25%, 12.5%, …

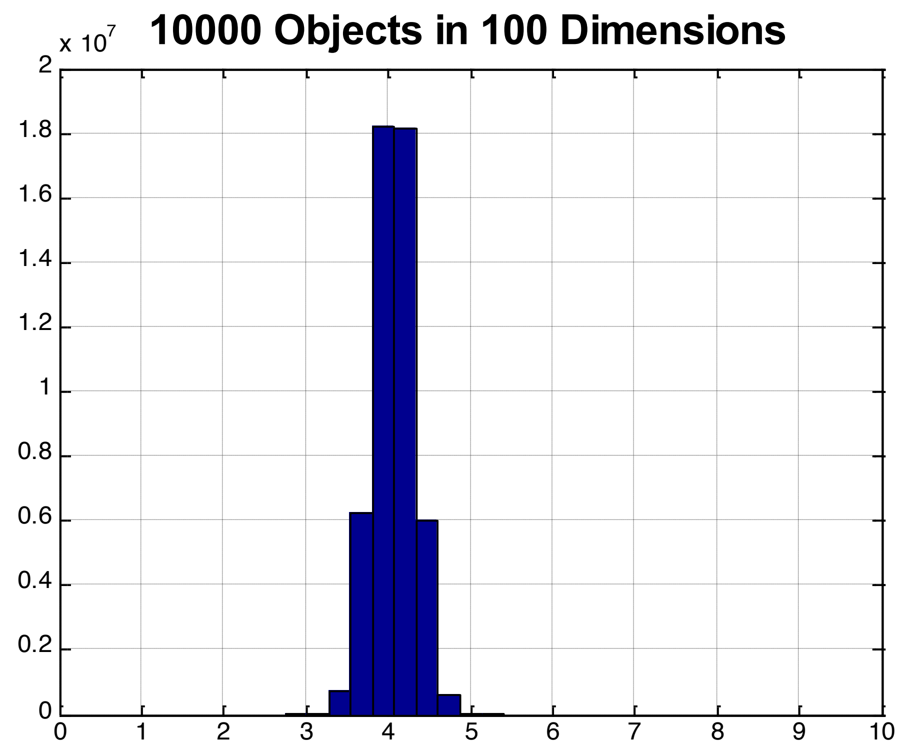
- Consequently, the number of interesting objects gets smaller and smaller

- For large values of n there will be no results, and for similar search radii
- You need to significantly increase the search radius to get some objects, but, you'll likely get everything !
- Anything is similar or un-similar to anything else

Curse of Dimentionality

- Everything is at the same distance.





How to overcome the dimensionality curse ?

- Try to understand what is useful, and what is not !

- ***Dimensionality reduction !***

- *In most cases it is worthwhile to first reduce the number of dimensions and then run any other analysis*

References

- **Data Mining Concepts and Techniques Third Edition**. Jiawei Han, Micheline Kamber Jian Pei. Morgan Kaufmann/Elsevier. Third Edition.
  - Section 2.4 Measuring Data Similarity and Dissimilarity
  - [optional] Chapter 2