

# Proyecciones Multidimensionales

## SECTION

1

una mirada hacia el descubrimiento de patrones y análisis de datos en alta dimensión

Prof. Dr. Diego Nascimento



DPTO. MATEMÁTICA | FACULTAD DE INGENIERÍA



# OVERVIEW

## MINI-COURSE SCHEDULE

### SECTION #1

#### Exploratory Analysis

- Visual Data Mining

### SECTION #2

#### Intro to Clustering in Data Modeling

- First steps to Overlook  
High-Dimensional Data

### SECTION #3

#### Intro to Causal Discovery

- Multivariate Time Series  
Structure Estimation

# WE ARE IN THE INFORMATION ERA!?!?!?

A lot of new data/observation  
...CONSTANTLY  
...ALL THE TIME

Nevertheless, it is NOT  
guaranteed that it is useful  
information and is important  
to be observed.

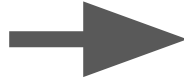
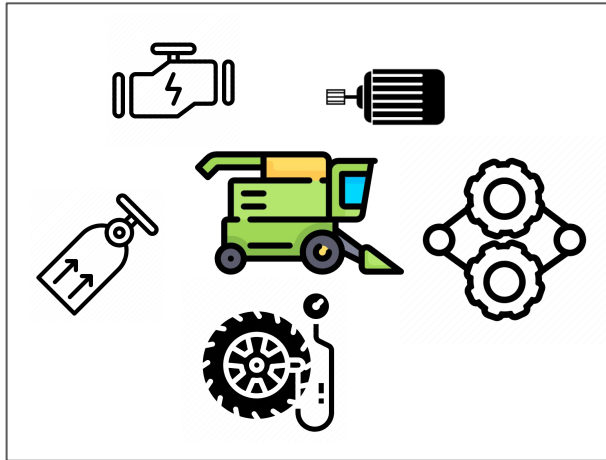


# THE WORLD NEEDS STATISTICS! WHY SO?

Statistics equals to Data Processed and Modeled, then **INFORMATION**.

- Unveiling the hidden pattern  
(i.e. lots of numbers -> transformed into a single number or a visual representation)

SIGNALS ARE RELATED MOST TIMES...



DOES THE  
MACHINE NEEDS  
A PREVENTIVE  
**MAINTENANCE** or  
**NOT?**

Conclusion are from a set of observations one draws conclusions  
(because is unfeasible to observe the entire population)

# IS ALL IMPORTANT INFORMATION COLLECTED?

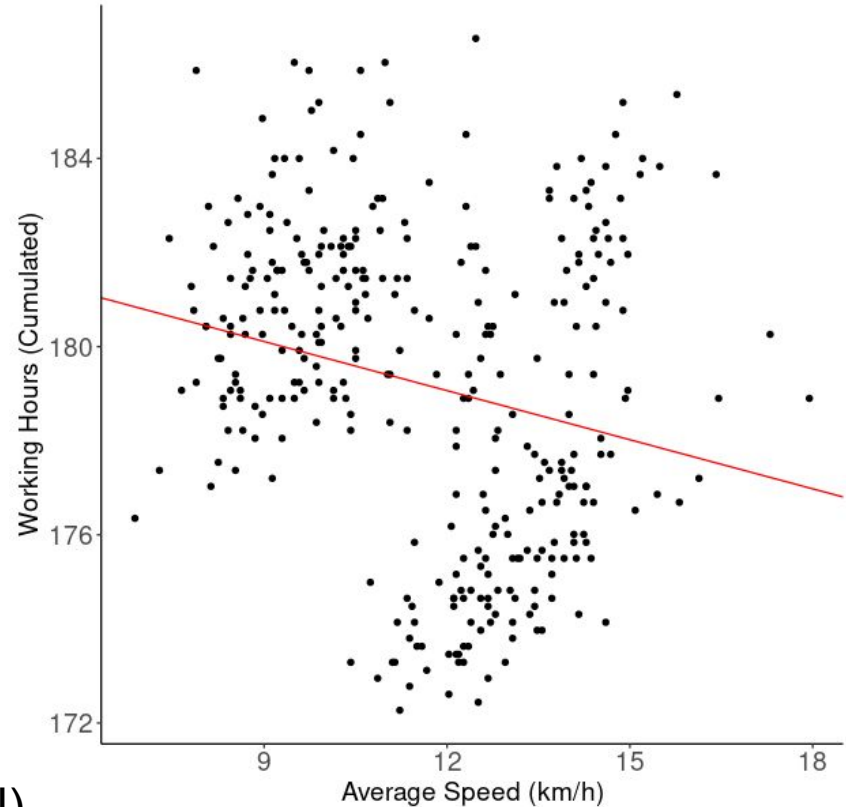
Let's see a Harvester set...

It seems that as  
more speed  
average is  
observed as  
more Maintenance will be needed!



- That is an inversely proportional relationship.

(NEGATIVE ASSOCIATION)

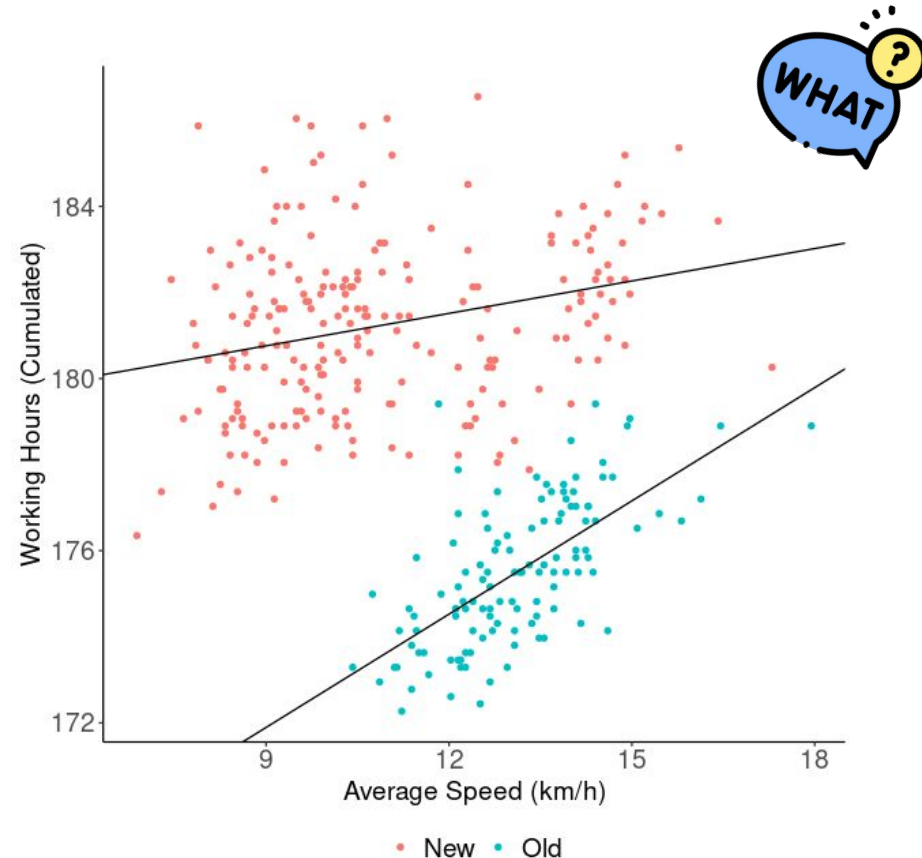


# What if the Harvester set is composed by TWO MACHINES?

...but by adding an extra information, that this data comes from two different machines with different ages,

- That proportional relationship becomes POSITIVE ASSOCIATION!

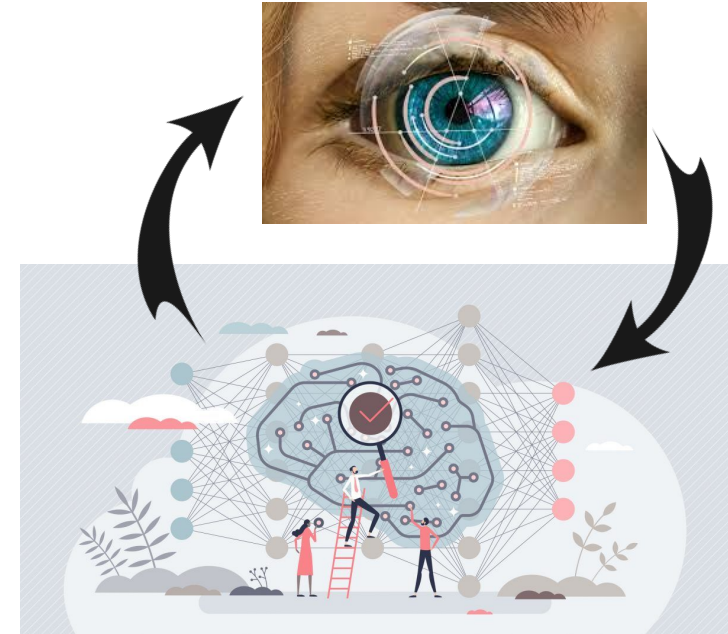
This phenomenon is called  
**SIMPSON'S PARADOX!**



# UNRAVELING PATTERN THROUGH VISUAL MINING

Approximately half of neural tissues are directly or indirectly linked to vision (the human cognitive system recognizes patterns in visual objects).

- The ability to identify, interpret, and extract features of interest from datasets displayed within **graphical elements** (HOFFMAN, 1998; WARD; GRINSTEIN; KEIM, 2010).
- Data complexity -> visual clutter produced (large volumes of information).
- In a limited environment, such as computer screens, there is a demand for system design geared towards visual exploration.



# DATA VISUAL MINING **field** (powered by R)



Translating Data Observation (Features) into patterns:

- Univariate Density Distribution (i.e. Boxplot, Histogram, Violin plot)
- Relational Distribution (i.e. Correlation)
- Ranking Distribution (Radar plot, Parallel Coordinates, Lollipop)
- Group comparison (Barplot, Tree, Pie chart, Dendrogram)
- Spatial or Time evolution
- Others like information Flow or Interactive graph



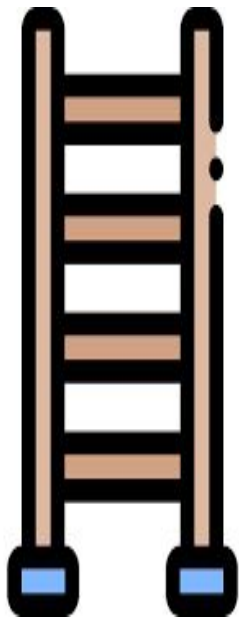
<https://r-graph-gallery.com/>

<http://www.sthda.com/english/wiki/data-visualization>



# TYPES OF FEATURES (level of Information)

MORE INFORMATIVE



LESS INFORMATIVE

QUANTITATIVE  
(NUMERICAL)

RATIO

INTERVAL

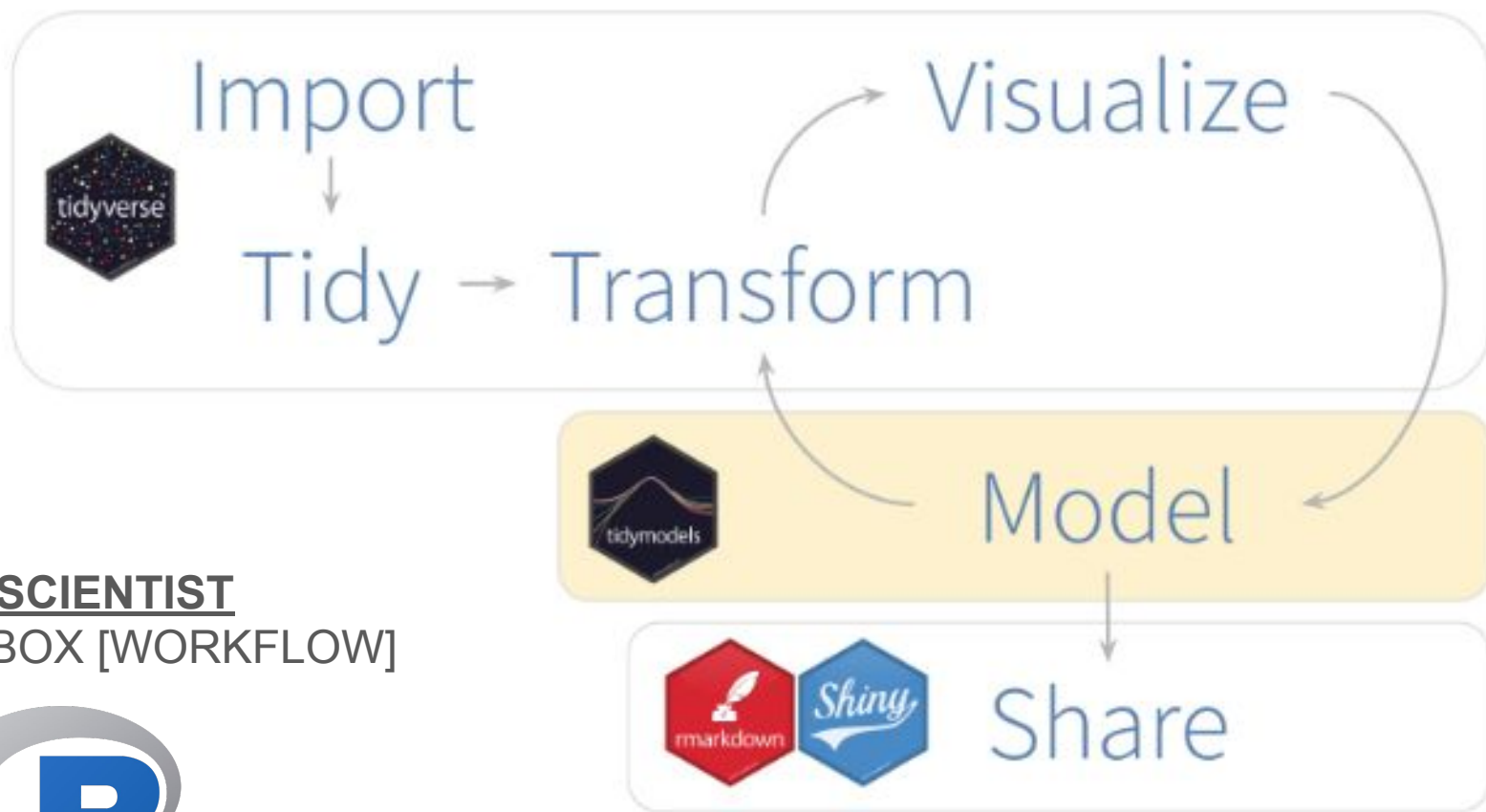
ORDINAL

QUALITATIVE  
(CATEGORICAL)

NOMINAL

BEST MEASURES

CENTRAL	SPREAD
MEAN MEDIAN MODE	VARIANCE RANGE IQR
MEAN MEDIAN MODE	VARIANCE RANGE IQR
MODE MEDIAN	RANGE IQR
MODE	-



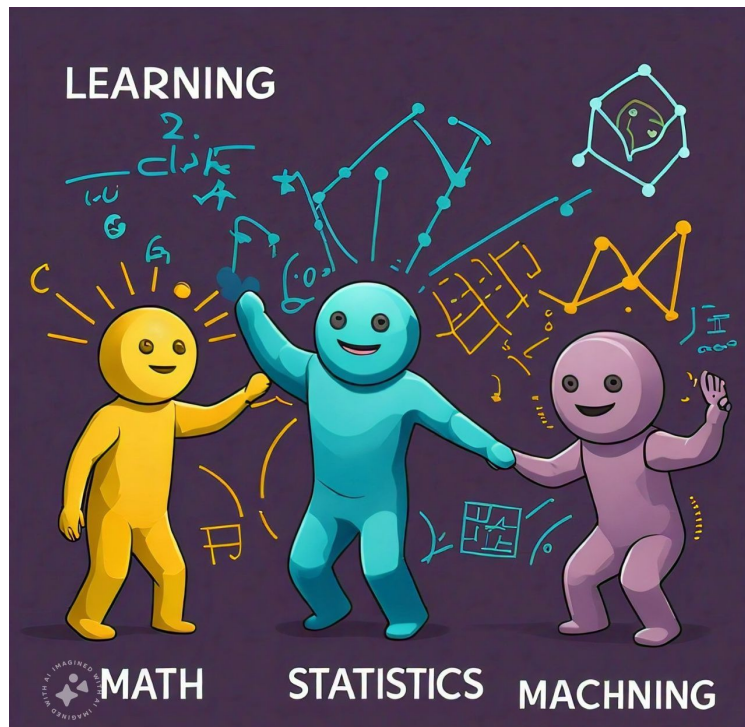
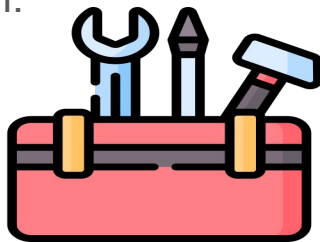
**DATA SCIENTIST**  
TOOLBOX [WORKFLOW]



# MACHINE LEARNING + STATISTICS + MATHEMATICS

The three areas tries to quantifies the observed world, through function that adopts constrains or other suppositions.

- **Statistics**, i.e. uses the Inferential and Probabilistic fields
- **Machine Learning**, i.e. same as statistics but focusing majorly in prediction
- **Mathematical** models adopts a deterministic approach.



## - What is Artificial Intelligence (AI)?

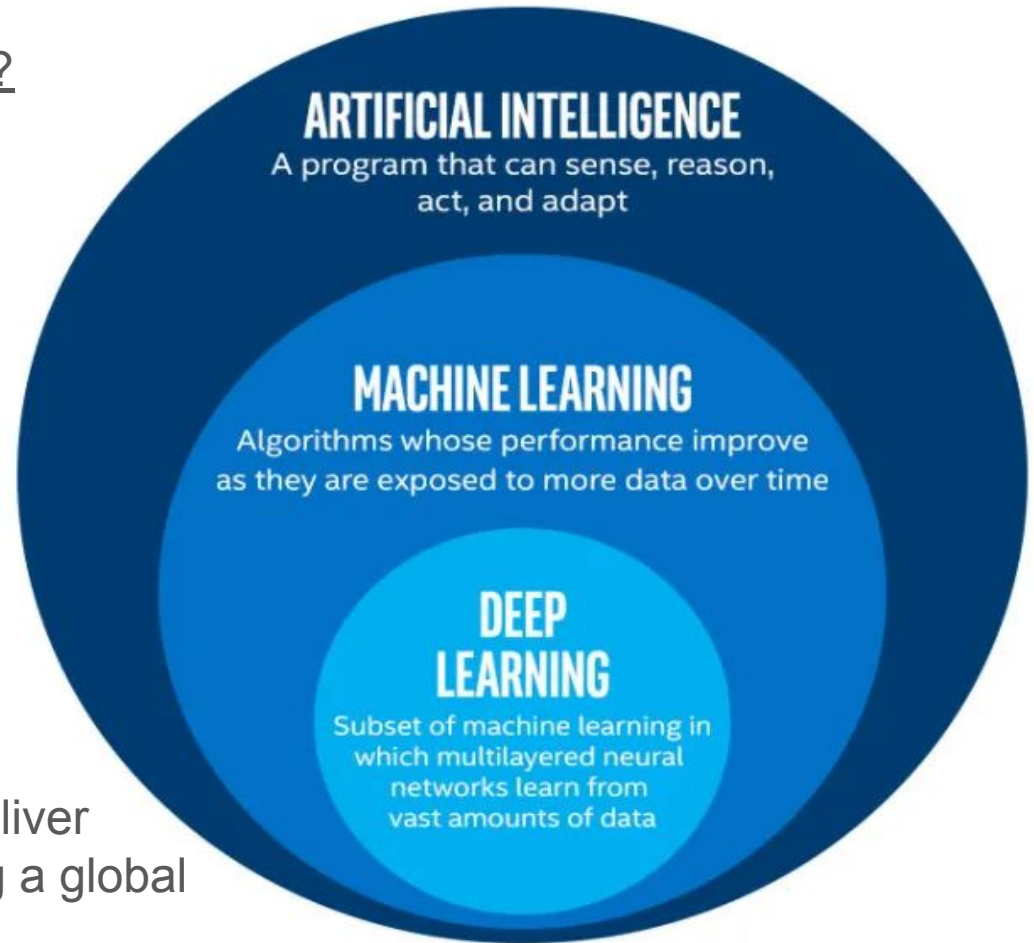
It's the art that machines learn to mimic humans tasks.

## - How do they do that?

Through a sequence of logical steps (instructions).

## - Is this rationality linear?

It all ends up to be a mathematical formal which given an input will deliver an output result (some times using a global pattern only other local + global).



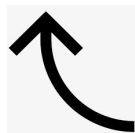


MultiLayer Perceptron (MLP)

Convolutional Neural Network

Recurrent Neural network

...Deep Learning



Artificial Neural  
Networks (ANN)

(Generalized) Linear Models  
Lasso & Ridge Regression  
eXtreme Gradient Boosting (XGBoost)

Regression

Reinforcement  
Learning

STATISTICAL  
LEARNING  
TECHNIQUES

Classification

Clustering

DIMENSIONALITY  
REDUCTION

K-means  
Fuzzy c-means (FCM)

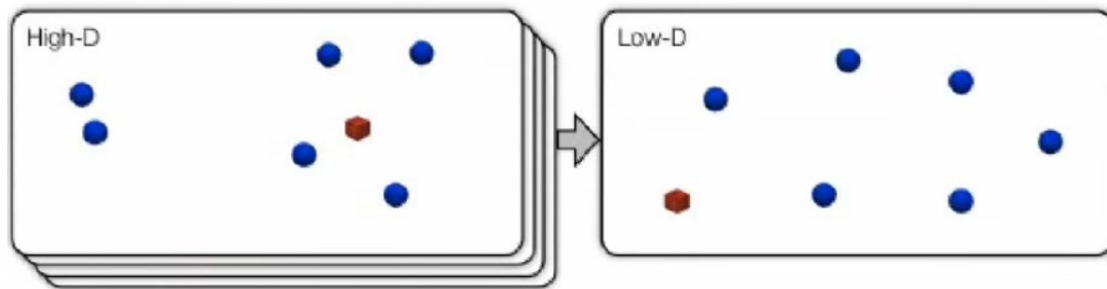
Multidimensional Project (MDP)  
Principal Component Analysis (PCA)

## CLASES DE FUNCIONES

Árbol de Decisión  
Regresión Logística  
Análisis Discriminante  
K-nearest neighbors (KNN)  
Support Vector Machine (SVM)

# BEFORE MODELING → DATA VISUALIZATION

- Build *map* in which distances between points reflect similarities in the data:



- Minimize some objective function that measures the *discrepancy* between similarities in the data and similarities in the map...

SCAN ME



## Getting over High-Dimensionality: How Multidimensional Projection Methods Can Assist Data Science

by Evandro S. Ortigossa <sup>1,†</sup>, Fábio Felix Dias <sup>1,†</sup> and Diego Carvalho do Nascimento <sup>2,†</sup>

<sup>1</sup> Institute of Mathematics and Computer Science, University of São Paulo, São Carlos 13566590, Brazil

<sup>2</sup> Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó 1530000, Chile

\* Author to whom correspondence should be addressed.

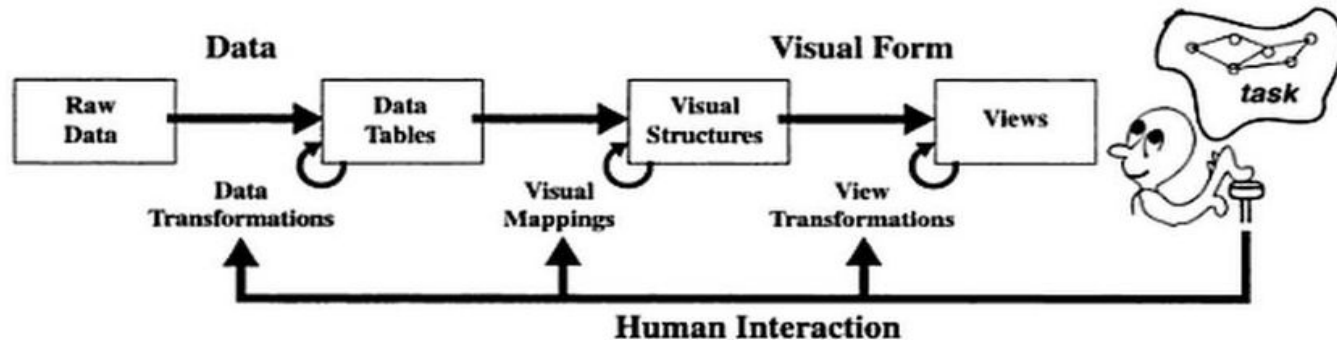
† These authors contributed equally to this work.



# DATA VISUAL MINING field

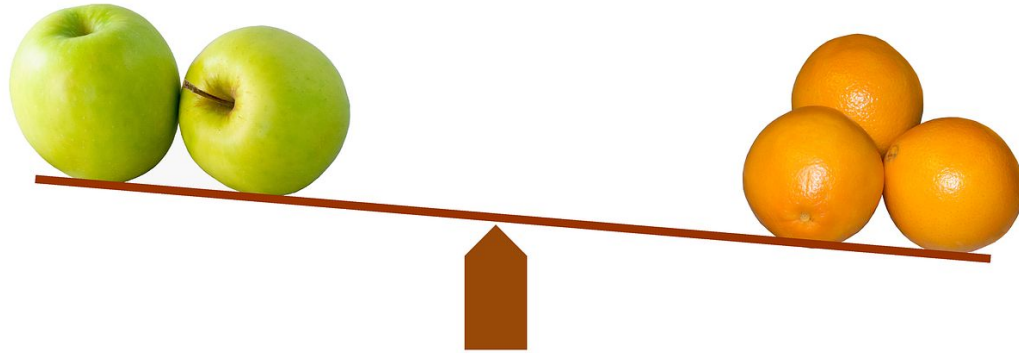
Visual data mining refers to the combination of traditional data mining techniques with data visualization tools to visually analyze patterns of interest.

- Information visualization techniques can be useful in solving problems.
- In visual data mining concerns depicting multidimensional data in a human-perceivable manner.
- An overview of methodologies for mining complex data types.



**SOURCE:** Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). Readings in information visualization: using vision to think. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.





HOW TO COMPARE OBSERVATIONS?

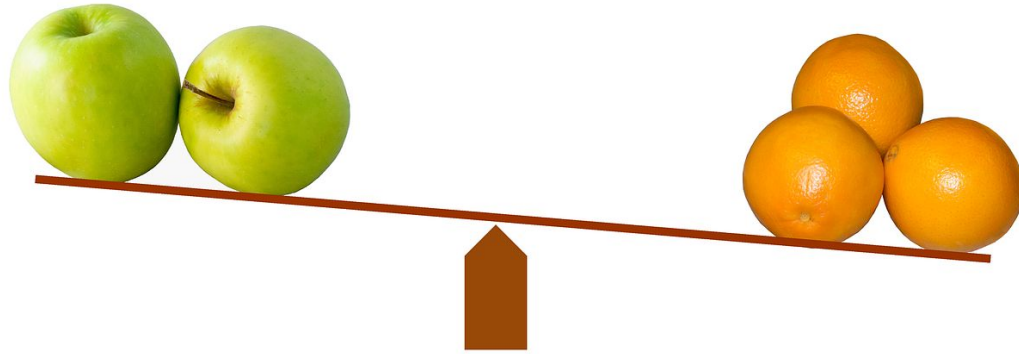
---



# SIMILARITY & DISSIMILARITY MEASURES

The idea is to calculate the “distance” between two points/observations:

ATTRIBUTE	DISSIMILARITY [SIMILARITY]	EXAMPLE
NOMINAL	$D(x,y) = 0$ if $x=y$ $D(x,y) = 1$ if $x \neq y$	<u>COLORS = {GREEN,RED,BLUE}</u>
	$[S(x,y) = 1$ if $x=y$ $S(x,y) = 0$ if $x \neq y]$	$D(\text{“GREEN”}, \text{“RED”}) = 1$ $S(\text{“RED”}, \text{“RED”}) = 1$
ORDINAL	$D(x,y) =  x-y  / (n-1)$	<u>5-POINT LIKERT SCALE</u> or {1,2,3,4,5} {Strong Dis., Disagree, Neutral, Agree, Strongly Agr.}
	$[S(x,y) = 1 - D(x,y)]$	$D(\text{Strongly Agree}, \text{Disagree}) =  5-2 /(4) = 0.75$ $S(\text{Strongly Agree}, \text{Disagree}) = 1 - 0.75 = 0.25$
INTERVAL or RATIO	$D(x,y) =  x-y $	<u>BODY TEMPERATURE = [35°, 42°]</u>
	$[S(x,y) = -D(x,y)$ or $S(x,y) = 1/(1+D(x,y))]$	$D(\text{“JOSE”}=\{37.5^\circ\}, \text{“JOHN”}=\{38.15^\circ\}) = 0.65^\circ$



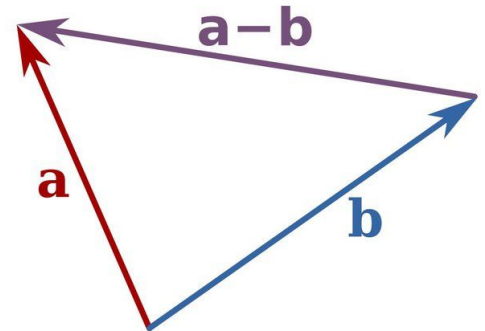
HOW TO COMPARE TWO OBJECTS USING MULTIPLE  
CHARACTERISTICS SIMULTANEOUSLY??

# GEOMETRIC CONCEPT (INTERVAL VALUES)

The number of dimensions is the observation space dimension, whereas each Vector in component form of a SINGLE OBSERVATION, in which the distance between TWO observation points is represented as  $d(\mathbf{a}, \mathbf{b})$  is the norm like,

$$||a - b|| = \sqrt{|a_x - b_x|^2 + |a_y - b_y|^2 + |a_z - b_z|^2 + \dots}$$

where  $\mathbf{a} = [a_x, a_y, a_z, \dots]$  and  $\mathbf{b} = [b_x, b_y, b_z, \dots]$



# DISSIMILARITY IDEA to feature vectors (INTERVAL VALUES)

MINKOWSKI DISTANCE (L-p norm)

$$d(\mathbf{X}_i, \mathbf{X}_j) = \|(\mathbf{X}_i - \mathbf{X}_j)\|_p = [ |X_{i,1} - X_{j,1}|^p + |X_{i,2} - X_{j,2}|^p + \cdots + |X_{i,n} - X_{j,n}|^p ]^{1/p}$$

MANHATTAN DISTANCE (L-1 norm)

$$d(\mathbf{X}_i, \mathbf{X}_j) = \|(\mathbf{X}_i - \mathbf{X}_j)\|_1 = [ |X_{i,1} - X_{j,1}| + |X_{i,2} - X_{j,2}| + \cdots + |X_{i,n} - X_{j,n}| ]$$

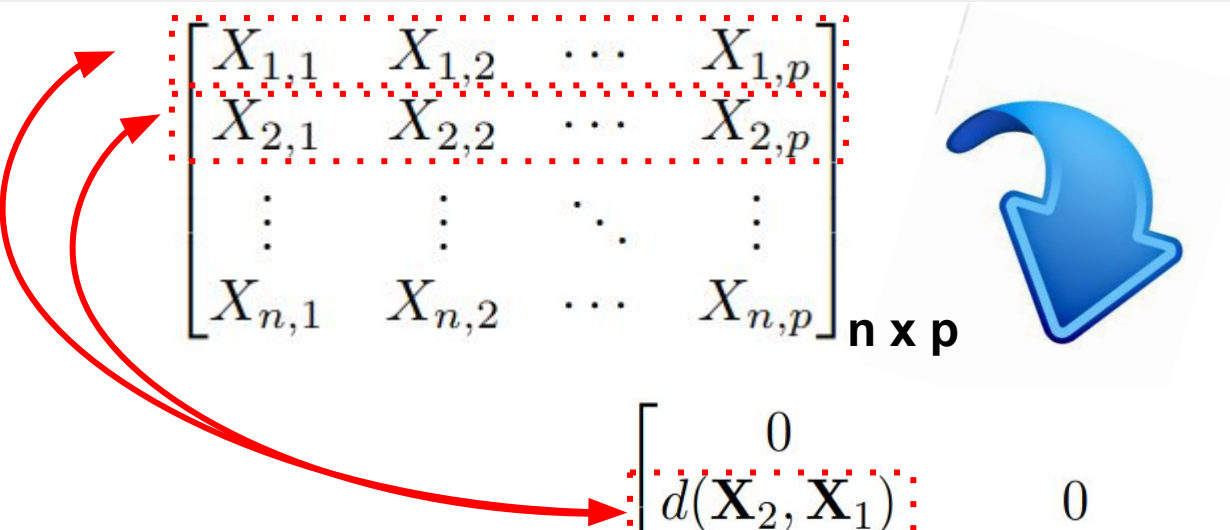
EUCLIDEAN DISTANCE (L-2 norm)

$$d(\mathbf{X}_i, \mathbf{X}_j) = \|(\mathbf{X}_i - \mathbf{X}_j)\|_2 = [ |X_{i,1} - X_{j,1}|^2 + |X_{i,2} - X_{j,2}|^2 + \cdots + |X_{i,n} - X_{j,n}|^2 ]^{1/2}$$

SUPREMUM DISTANCE (L-max norm)

$$d(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|_\infty = \max(|X_{i,1} - X_{j,1}|, |X_{i,2} - X_{j,2}|, \cdots, |X_{i,n} - X_{j,n}|)$$

# Transforming DATA MATRIX => DISTANCE MATRIX


$$\begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{bmatrix}_{n \times p}$$

$$\begin{bmatrix} 0 & & & & \\ d(\mathbf{X}_2, \mathbf{X}_1) & 0 & & & \\ d(\mathbf{X}_3, \mathbf{X}_2) & d(\mathbf{X}_3, \mathbf{X}_2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(\mathbf{X}_n, \mathbf{X}_1) & d(\mathbf{X}_n, \mathbf{X}_2) & d(\mathbf{X}_n, \mathbf{X}_3) & \cdots & 0 \end{bmatrix}_{n \times n}$$

# GENERIC DISTANCE METRIC (NOMINAL, ORDINAL or INTERVAL VALUES)

Jaccard INDEX:

It's a measure of SIMILARITY for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two populations.

- Jaccard Index = (the number in both sets) / (the number in either set) \* 100

$$S(X_i, X_j) = |X_i \cap X_j| / |X_i \cup X_j|$$

- Distance (dissimilarity) Jaccard Index

$$D(X_i, X_j) = 1 - S(X_i, X_j)$$

# JACCARD SIMILARITY –NOMINAL EXAMPLE–



Let's suppose one person has to go to the grocery store,

## Shopping LIST

- COOKIES
- CACAO
- MILK
- ORANGE

...but

## Effectively Bought

- COOKIES
- CACAO
- MILK
- BUBBLE GUM

How close (or likely) the bought items are from the original plan?

$$\text{JACCARD SIMILARITY} = \frac{3 \text{ COMMON ITEMS}}{5 \text{ TOTAL ITEMS}} = 0.6 \text{ or } 60\%$$

$$\text{DISTANCE} = 1 - \text{JACCARD INDEX} = 1 - 0.6 = 0.4$$

# MULTIDIMENSIONAL PROJECTION (MDP)

Since the visual space is limited to three dimensions, when the  $m$ -dimensionality increases, the complexity of representing and interpreting data also increases.

In this context, techniques devoted to dimensionality reduction aim to present multidimensional data from high-dimensional spaces as points in dimensionality-reduced spaces.

**Definition 1.** Let  $X$  be a set of  $n$  data objects in  $\mathbb{R}^m$ , with  $m > 3$  and  $\delta : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$  a distance measure between the instances in  $\mathbb{R}^m$ ;  $Y$  be a set of  $n$  points in  $\mathbb{R}^p$ , with  $p \in \{1, 2, 3\}$  and  $d : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$  a distance measure between points in  $\mathbb{R}^p$ . A multidimensional projection technique can be described as a function  $f : X \rightarrow Y$  that aims to make  $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$  as close as possible to zero,  $\forall x_i, x_j \in X$ .



# MDP APPROACHES

Multidimensional projection methods. The column “Complexity” presents the computational load related with the amount of observations ( $n$ ), feature-dimensions ( $m$ ), iterations ( $k$ ), samples ( $s$ ).

Technique	Transformation	Nature	Complexity
PCA	Linear	Global	$O(m^3)$
Classical MDS	Linear	Global	$O(n^3)$
Kruskal	Nonlinear	Global	$O(kn^2)$
Sammon's	Nonlinear	Global	$O(kn^2)$
FastMap	Nonlinear	Global	$O(n)$
Chalmers	Nonlinear	Local	$O(n^2)$
Pekalska	Nonlinear	Global	$O(s^3 + sn)$
Isomap	Nonlinear	Global	$O(n^3)$
Chalmers Hybrid	Nonlinear	Local	$O(n\sqrt{n})$
L-Isomap	Nonlinear	Global	$O(sn \log n)$

Multidimensional projection methods. The column “Complexity” presents the computational load related with the amount of observations ( $n$ ), feature-dimensions ( $m$ ), iterations ( $k$ ), samples ( $s$ ), and graph edges ( $E$ ).

SNE	Nonlinear	Local	$O(n^2)$
Force Scheme	Nonlinear	Global	$O(n^2)$
LMDS	Nonlinear	Global	$O(s^3 + sn)$
LSP	Nonlinear	Global	$O(n^3)$
HiPP	Nonlinear	Global	$O(n\sqrt{n})$
t-SNE	Nonlinear	Local	$O(n^2)$
Glimmer	Nonlinear	Global	$O(n^2)$
PLMP	Partially linear	Global	$O(n^3)$
PLP	Nonlinear	Local	$O(n\sqrt{n})$
LAMP	Nonlinear	Local-Hybrid	$O(sn)$
LoCH	Nonlinear	Local	$O(n\sqrt{n})$
UMAP	Nonlinear	Local	$O(n^{1.14})$
TopoMap	Nonlinear	Global	$O(n \log n)$
GRMP	Nonlinear	Local	$O(n +  E )$

Source: elaborated by the authors.



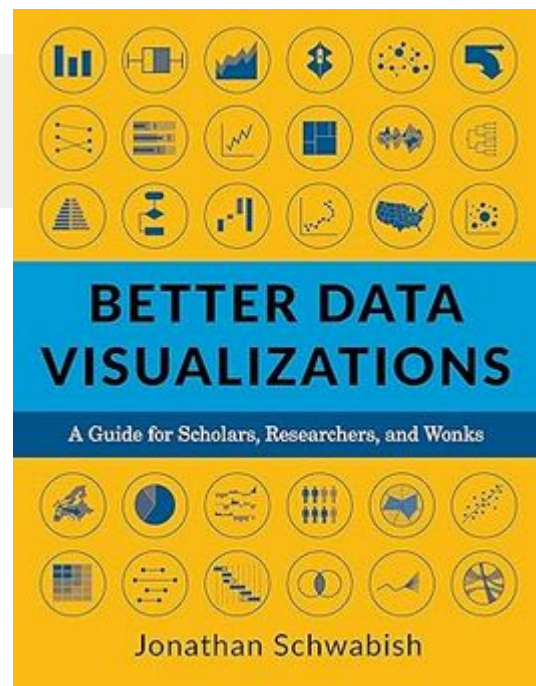
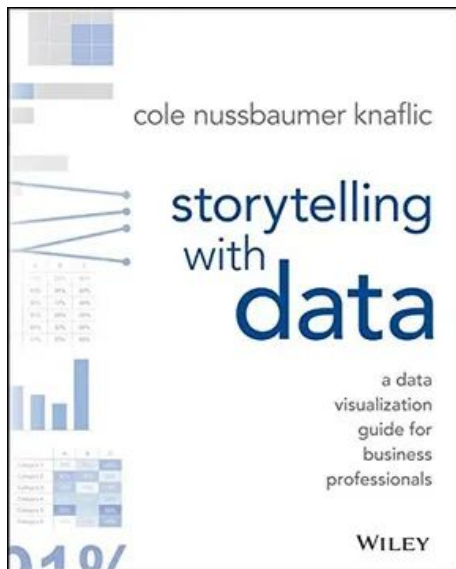
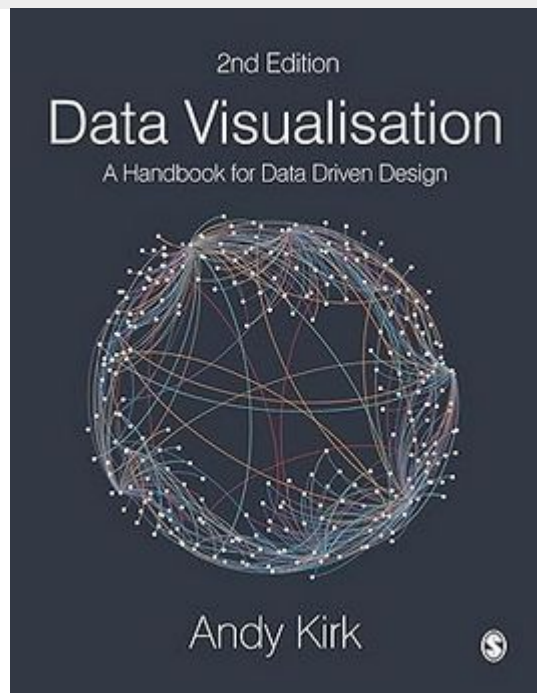
## EXEMPLIFICATION:

- Data Visual Mining

<https://github.com/ProfNascimento/ECU>

Ortigossa E.S., Dias F.F., Nascimento D.C. (2022) Getting over high-dimensionality: How Multidimensional Projection methods can assist Data Science? *Applied Sciences*, 12(13), 6799.  
<https://github.com/ProfNascimento/MP>

# FURTHER INVESTIGATION





UNIVERSIDAD  
**DE ATACAMA**

**Diego Nascimento**  
diego.nascimento@uda.cl  
Ph.D. in Statistics (USP/UFSCar)

[illegible]