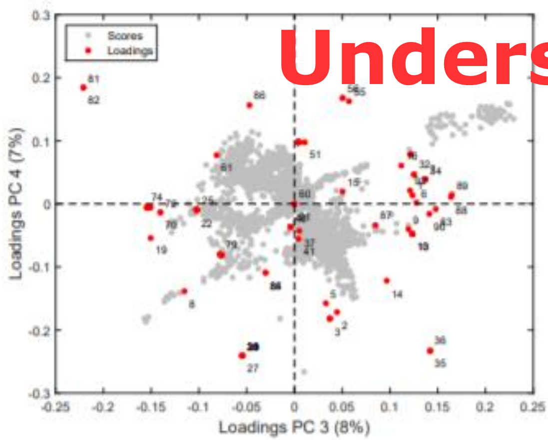
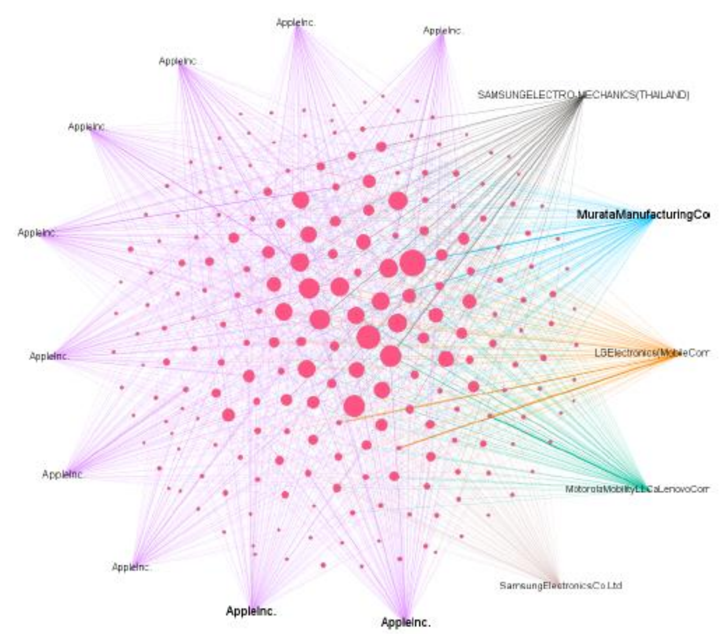
Multivariate Exploratory Data Analysis (MEDA):

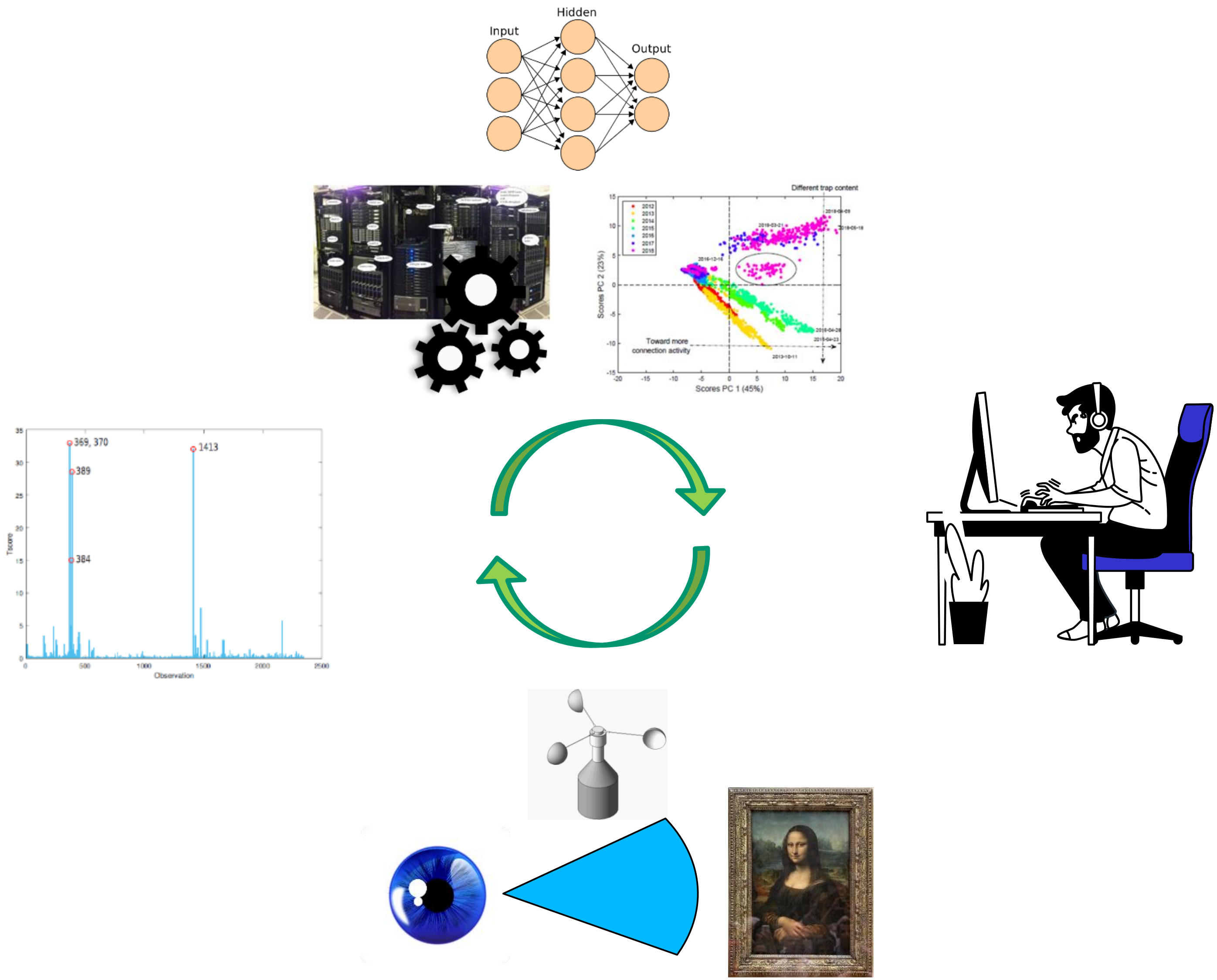
Understanding by looking at data





MEDA
University of Granada
José Camacho, Ph.D.

The Data-Driven approach

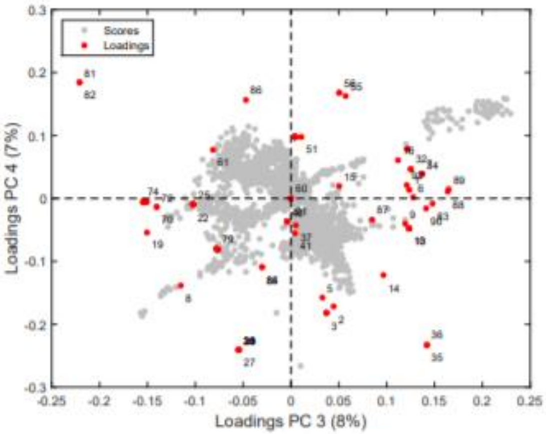


MEDA
University of Granada
José Camacho, Ph.D.

Disclaimer

I am a data analyst, specialized in multivariate analysis, with application to diverse domains:

- Biostatistics
- CyberSec
- Chemistry

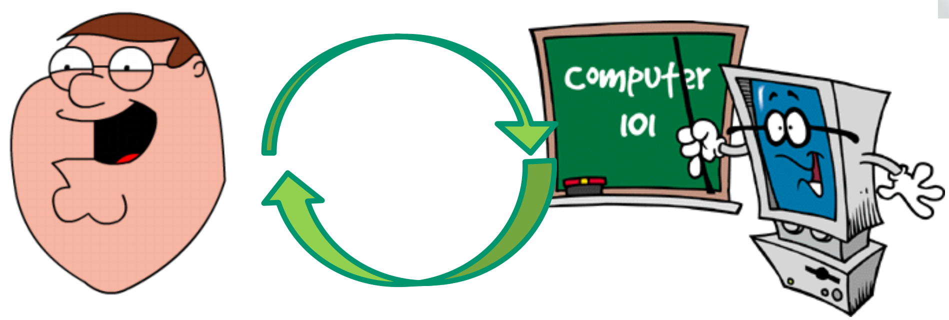


I am NOT a researcher on image perception, colorimetry, psychometry,…

- I train Ph.D. students in data exploration

Approaches to data analysis

Black Box Data mining / Machine learning

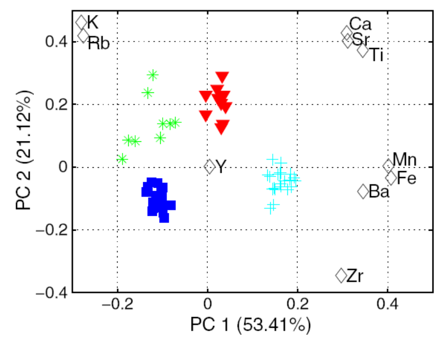


NETFLIX

amazon

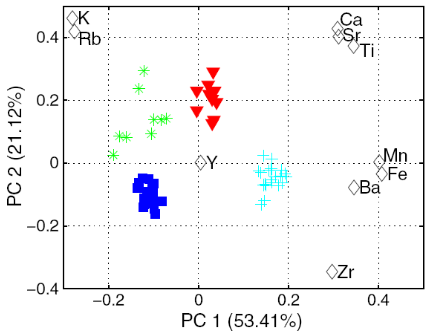
Google

Exploratory Data Analysis / Interpretable ML



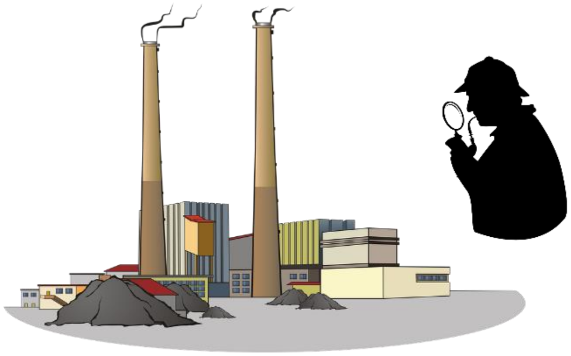
MEDA
University of Granada
José Camacho, Ph.D.

Use Cases for EDA / IML



- Research: Data ➔ Induction ➔ Hypothesis ➔ Testing ➔ New data

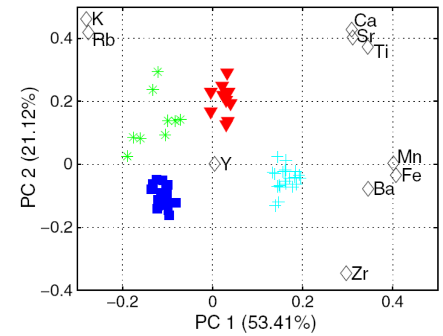
- Monitoring applications
  - (Cyber) Security
  - Industrial
  - Environment, …



Exploratory Data Analysis

Use Cases for EDA / IML

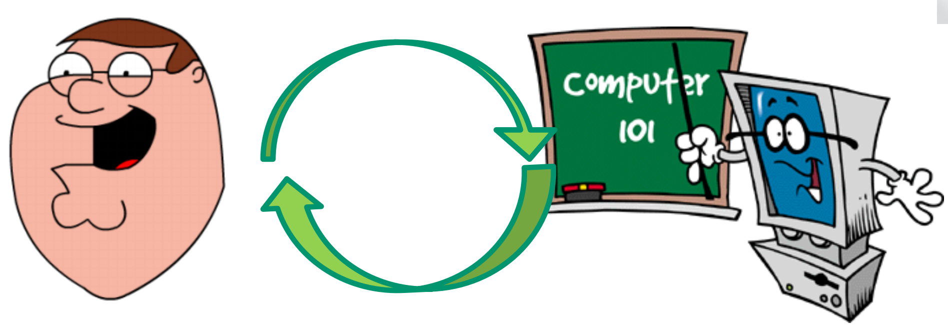
01



Pre-processing

Calibration

02



NETFLIX

amazon

Google

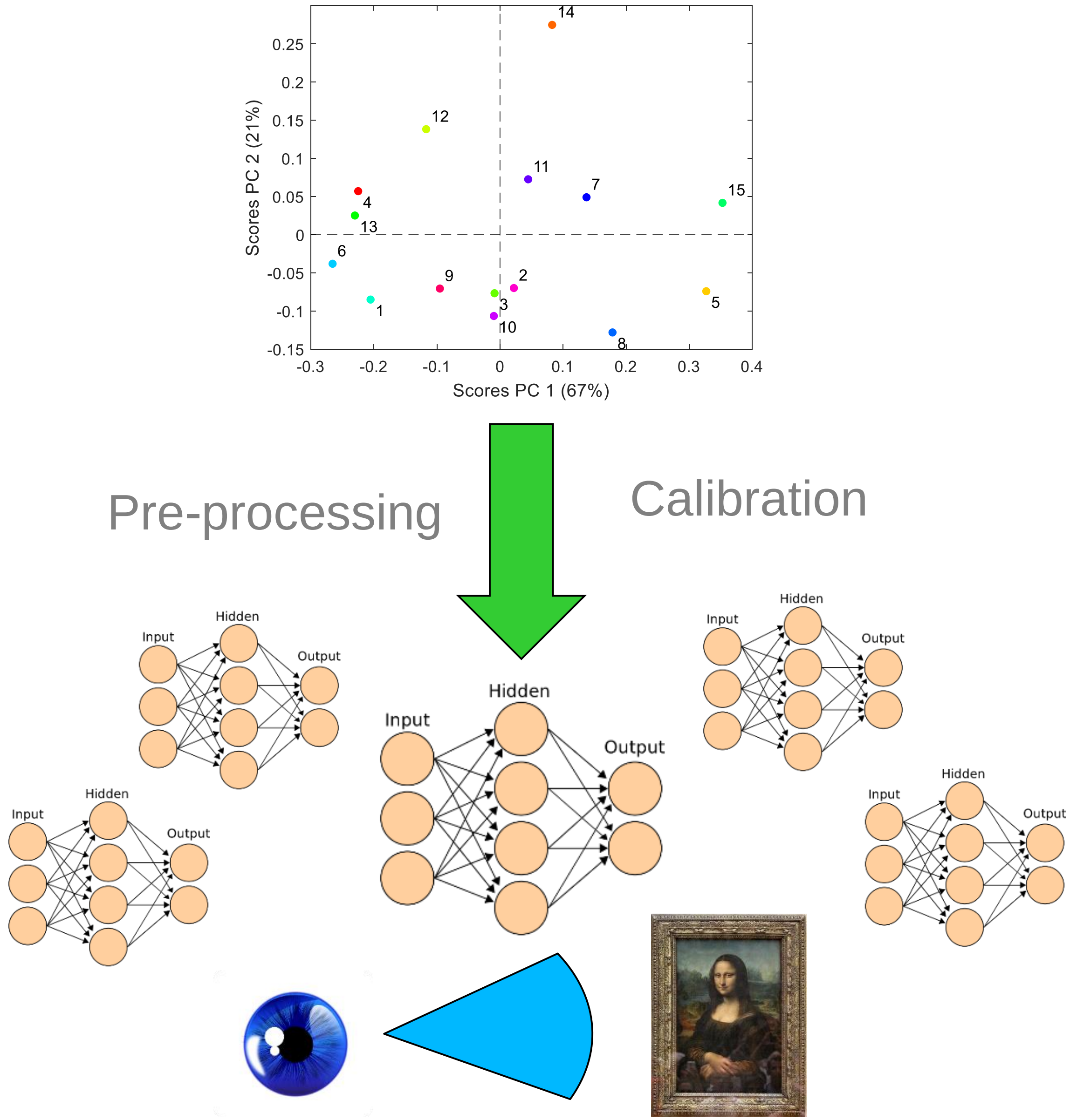
MEDA
University of Granada
José Camacho, Ph.D.

Exploratory Data Analysis

Use Cases for EDA / IML

01

02

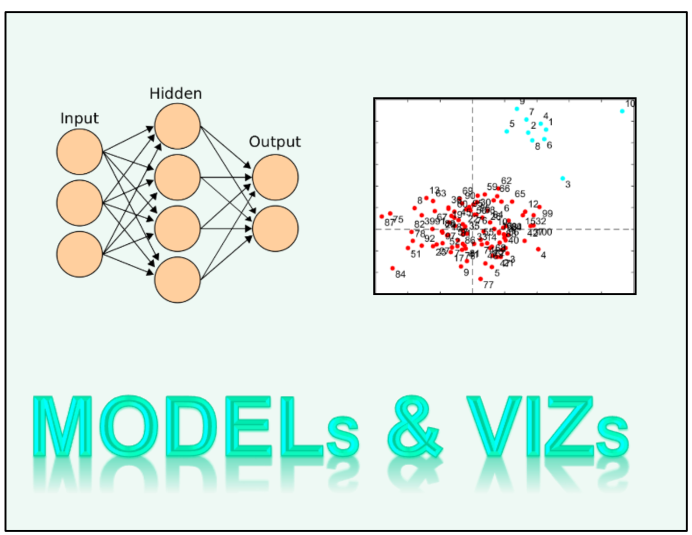


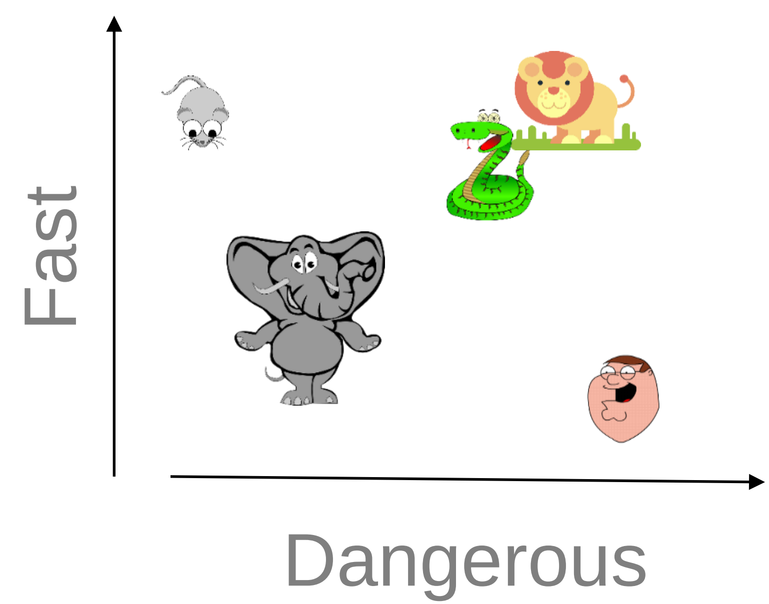
MEDA
University of Granada
José Camacho, Ph.D.

Exploratory Data Analysis

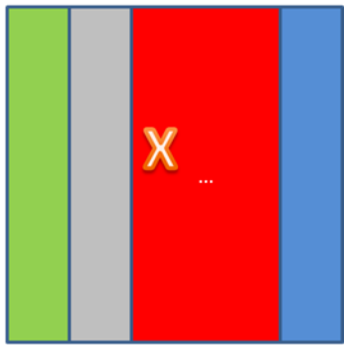
What is data understanding?

- Identify trends, averages, patterns of commonality or change

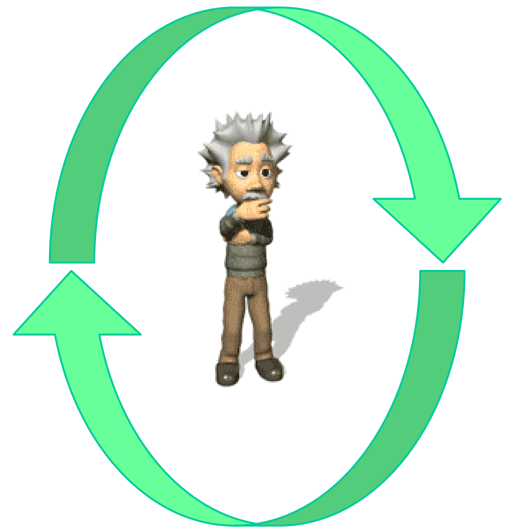


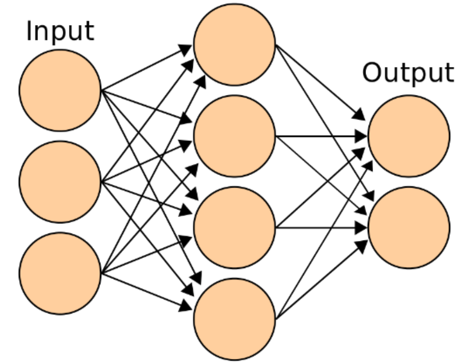


Exploratory Data Analysis



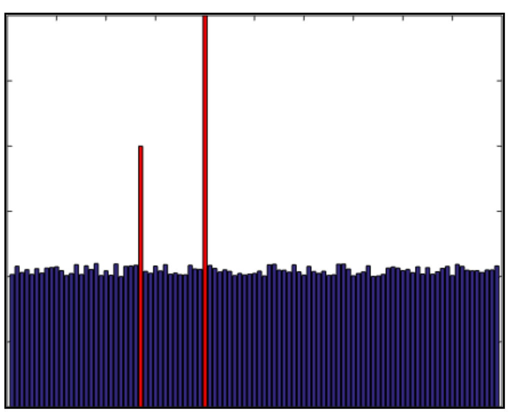
DATA



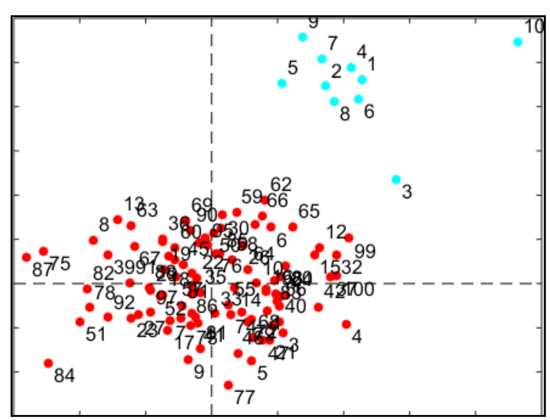


MODEL

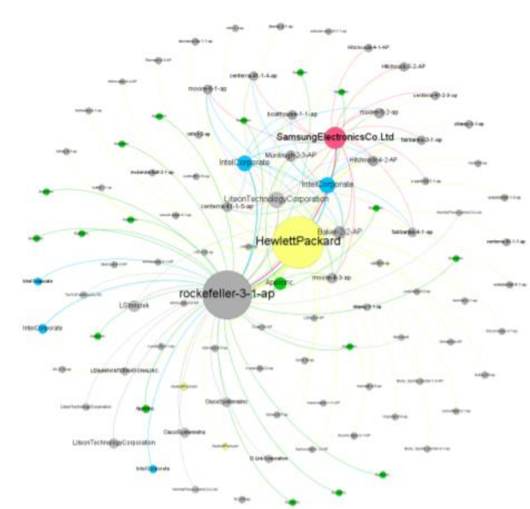
VISUALIZATION



Line plot



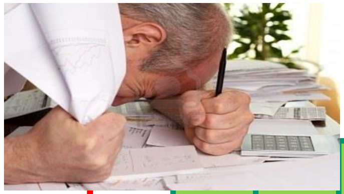
Scatter plot

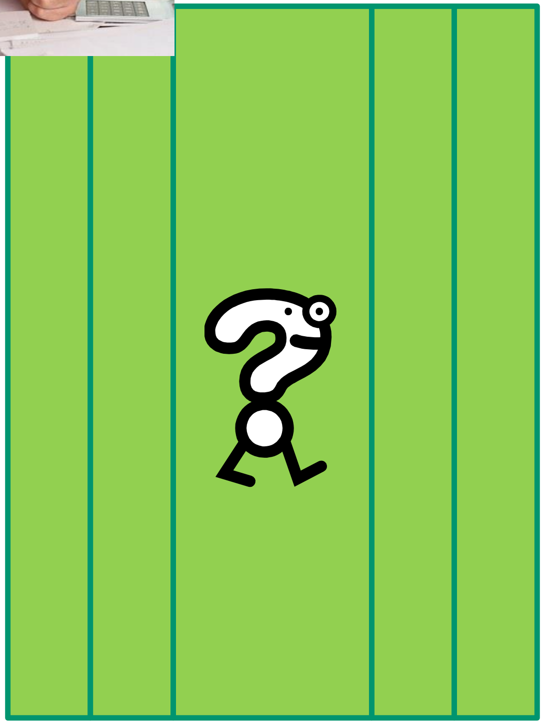


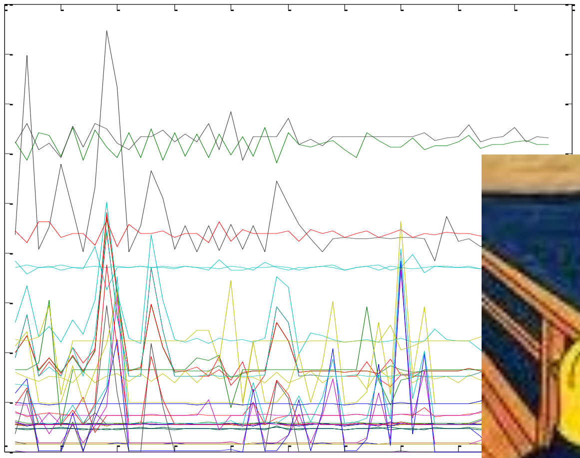
Networks

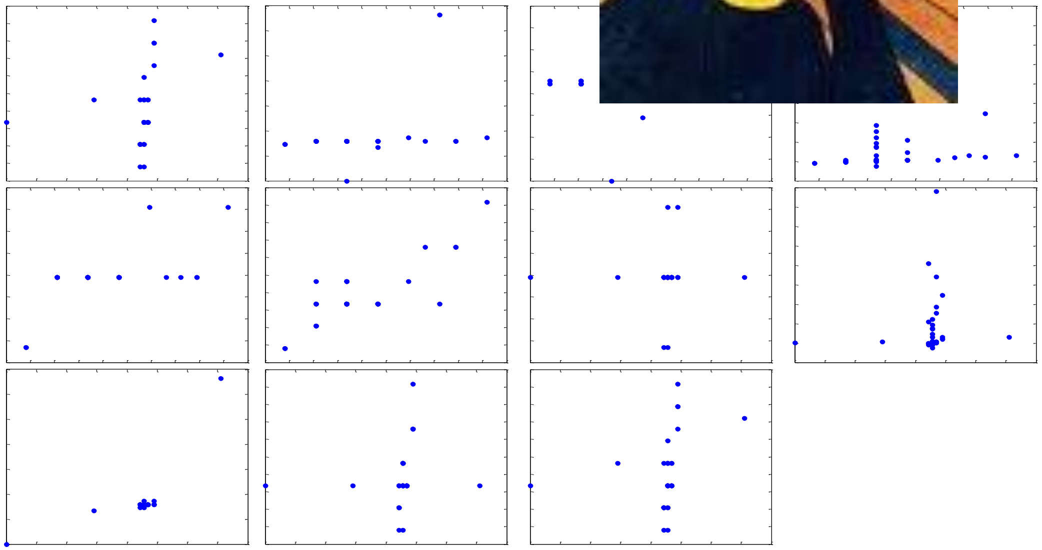
MEDA
University of Granada
José Camacho, Ph.D.

Data viz: How?





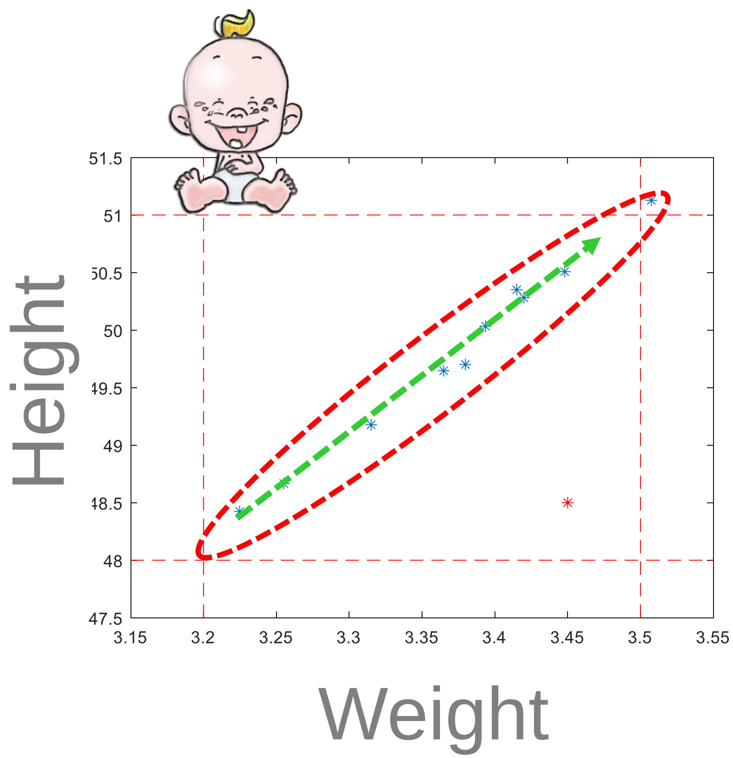


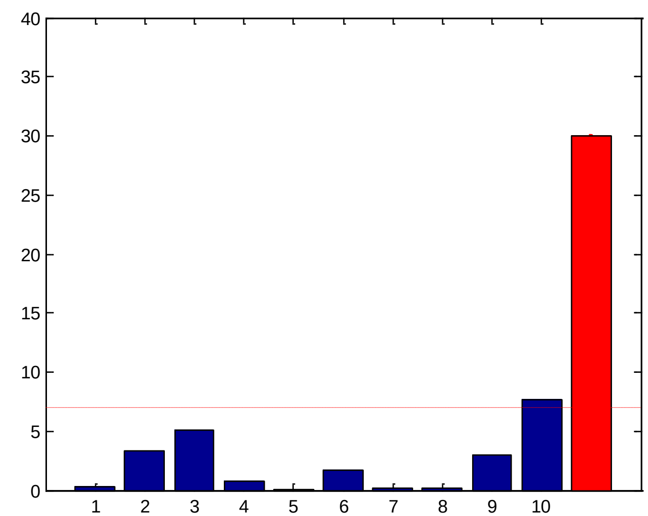


Multivariate approach

In a data set with many measured variables, the interesting information is contained in a (much lower) number of **latent variables**

E.g. Babies height vs weight

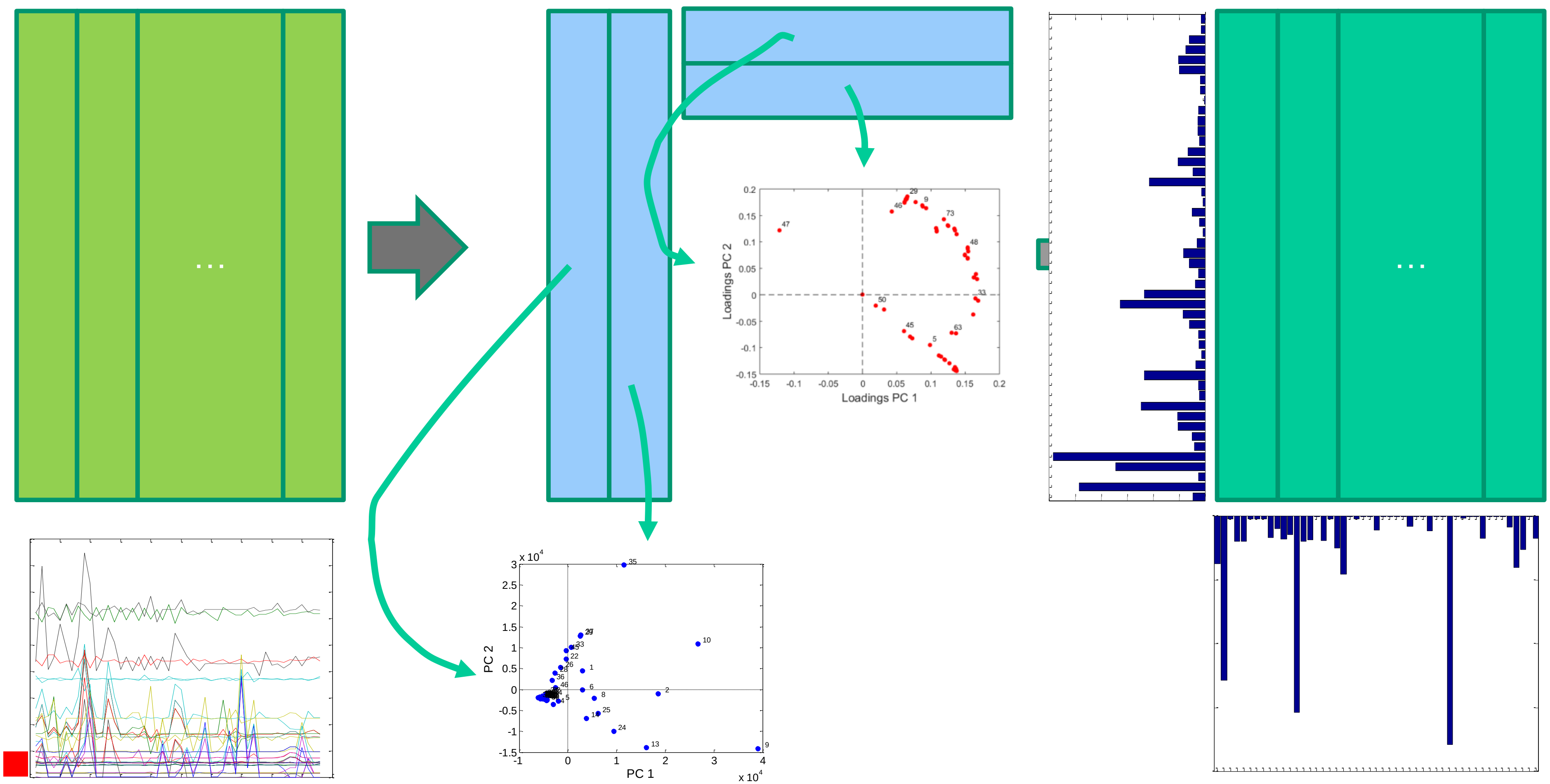




MEDA
University of Granada
José Camacho, Ph.D.

Multivariate Analysis

Matrix Factorization → Latent Variables



$$X = T \cdot P' + E$$

MEDA
University of Granada
José Camacho, Ph.D.

Multivariate Analysis

Matrix Factorization ➔ Latent Variables

MEDA
University of Granada
José Camacho, Ph.D.

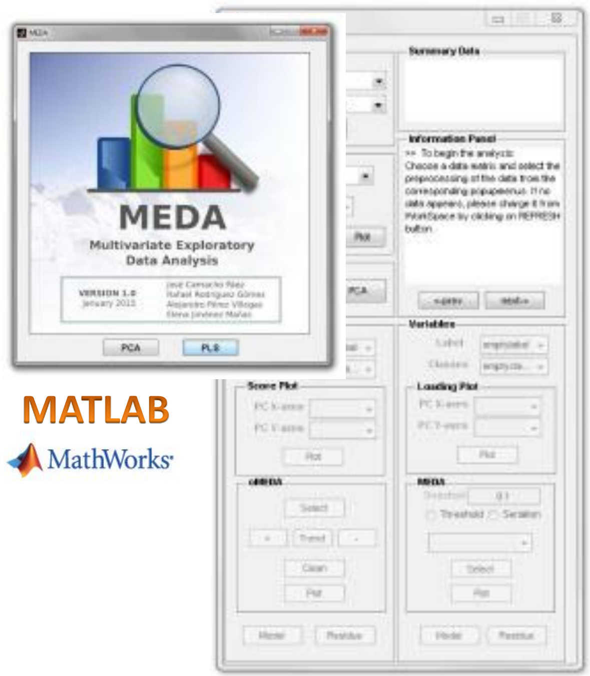
The MEDA Toolbox

Universidad de Granada

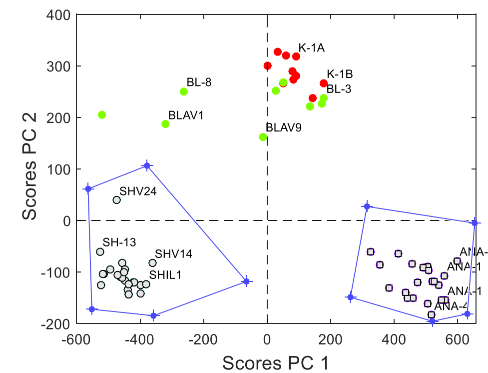
MEDA Toolbox

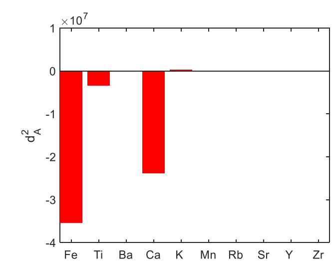
https://github.com/josecamachop/MEDA-Toolbox

- Models: PCA, PLS-DA, SPLS, GPCA, GPLS, ASCA, GASCA
- Dimensionality:
  - Scree plots
  - CV & D-CV
  - SVI Plots
- Structure in Variables:
  - Loading plots
  - MEDA plots
- Distribution of Observations
  - Score plots
  - MSPC: D-st, Q-st
  - Covariance MSPC: ADICOV
- Observations vs Variables
  - oMEDA plots
- Data simulation
  - simuleMV



MATLAB
MathWorks





ChemoLab, (2015) 143: 49

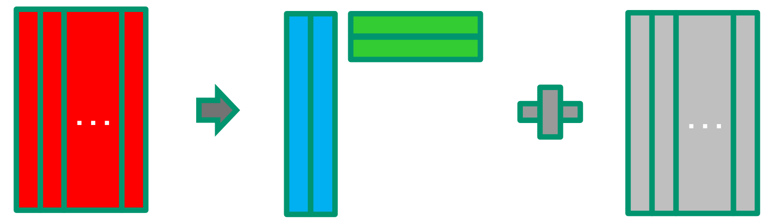
PCA Example: Wine dataset

	'Liquor'	'Wine '	'Beer '	'LifeEx'	'HeartD'
'France'	2.5000	63.5000	40.1000	78.0000	61.1000
'Italy '	0.9000	58.0000	25.1000	78.0000	94.1000
'Switz '	1.7000	46.0000	65.0000	78.0000	106.4000
'Austra'	1.2000	15.7000	102.1000	78.0000	173.0000
'Brit '	1.5000	12.2000	100.0000	77.0000	199.7000
'U.S.A.'	2.0000	8.9000	87.8000	76.0000	176.0000
'Russia'	3.8000	2.7000	17.1000	69.0000	373.6000
'Czech '	1.0000	1.7000	140.0000	73.0000	283.7000
'Japan '	2.1000	1.0000	55.0000	79.0000	34.7000
'Mexico'	0.8000	0.2000	50.4000	73.0000	36.4000

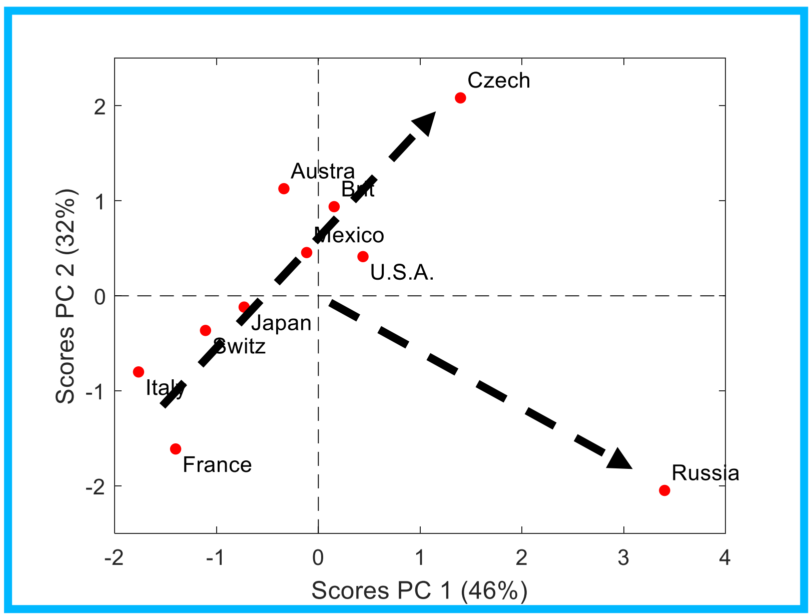
MEDA
University of Granada
José Camacho, Ph.D.

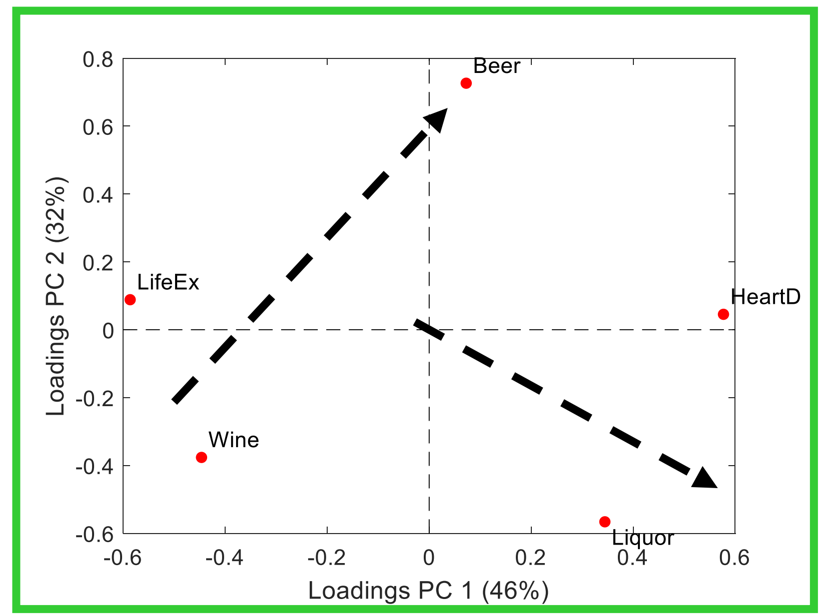
PCA: Wine dataset

	'Liquor'	'Wine '	'Beer '	'LifeEx'	'HeartD'
'France'	2.5000	63.5000	40.1000	78.0000	61.1000
'Italy '	0.9000	58.0000	25.1000	78.0000	94.1000
'Switz '	1.7000	46.0000	65.0000	78.0000	106.4000
'Austra'	1.2000	15.7000	102.1000	78.0000	173.0000
'Brit '	1.5000	12.2000	100.0000	77.0000	199.7000
'U.S.A.'	2.0000	8.9000	87.8000	76.0000	176.0000
'Russia'	3.8000	2.7000	17.1000	69.0000	373.6000
'Czech '	1.0000	1.7000	140.0000	73.0000	283.7000
'Japan '	2.1000	1.0000	55.0000	79.0000	34.7000
'Mexico'	0.8000	0.2000	50.4000	73.0000	36.4000



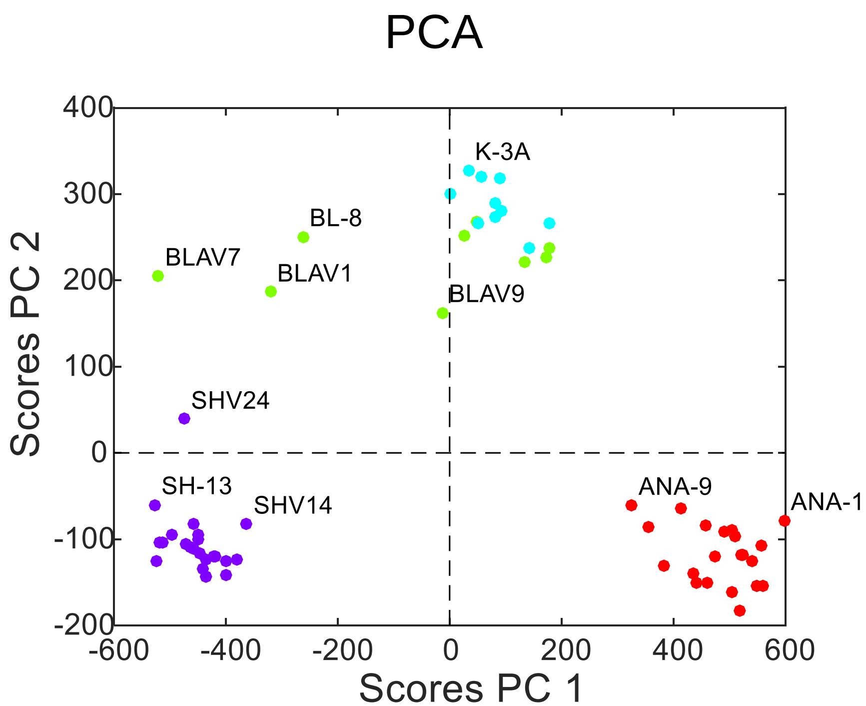
$$X = T \cdot P' + E$$



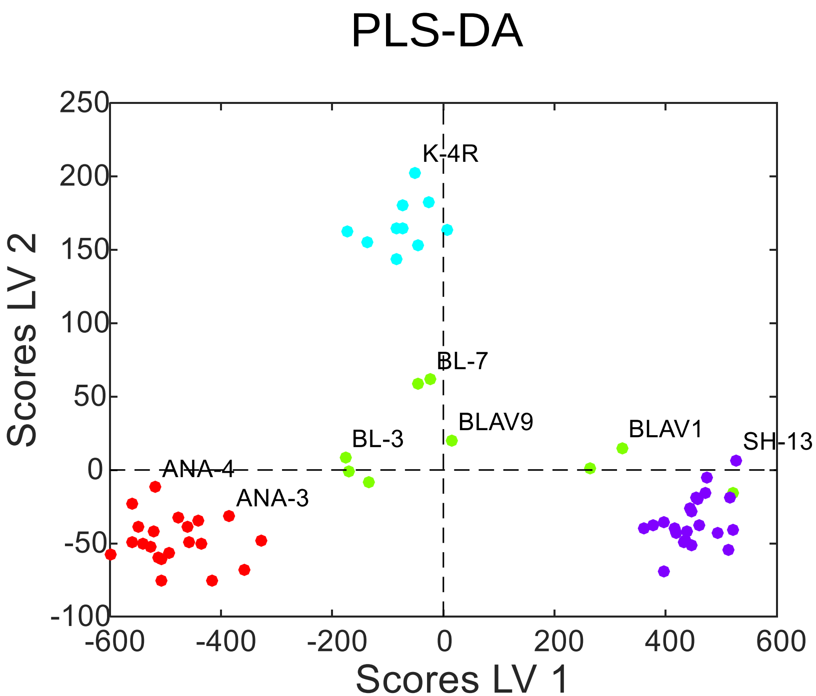


MEDA
University of Granada
José Camacho, Ph.D.

Extensions: PLS-DA



$$X = TP' + E$$



$$X = TP' + E$$

$$Y = TQ' + F$$

Example

Outdoor Experimental Comparison of Four Ad Hoc Routing Algorithms

Robert S. Gray[b]
robert.s.gray@dartmouth.edu

David Kotz[a]
dfk@cs.dartmouth.edu

Calvin Newport[a]
Calvin.Newport@alum.dartmouth.org

Nikita Dubrovsky[a]
Nikita.Dubrovsky@alum.dartmouth.org

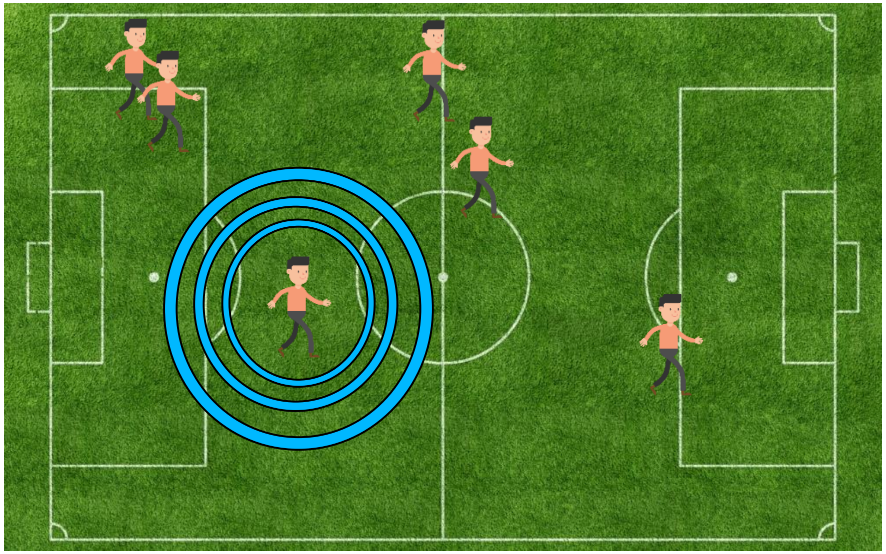
Aaron Fiske[a]
Aaron.Fiske@alum.dartmouth.org

Jason Liu[c]
jasonliu@mines.edu

Christopher Masone[b]
Christopher.Masone@dartmouth.edu

Susan McGrath[b]
smcgrath@ists.dartmouth.edu

Yougu Yuan[a]
yuanyg@cs.dartmouth.edu



Example

Outdoor Experimental Comparison of Four Ad Hoc Routing Algorithms

Robert S. Gray[b]
robert.s.gray@dartmouth.edu

David Kotz[a]
dfk@cs.dartmouth.edu

Calvin Newport[a]
Calvin.Newport@alum.dartmouth.org

Nikita Dubrovsky[a]
Nikita.Dubrovsky@alum.dartmouth.org

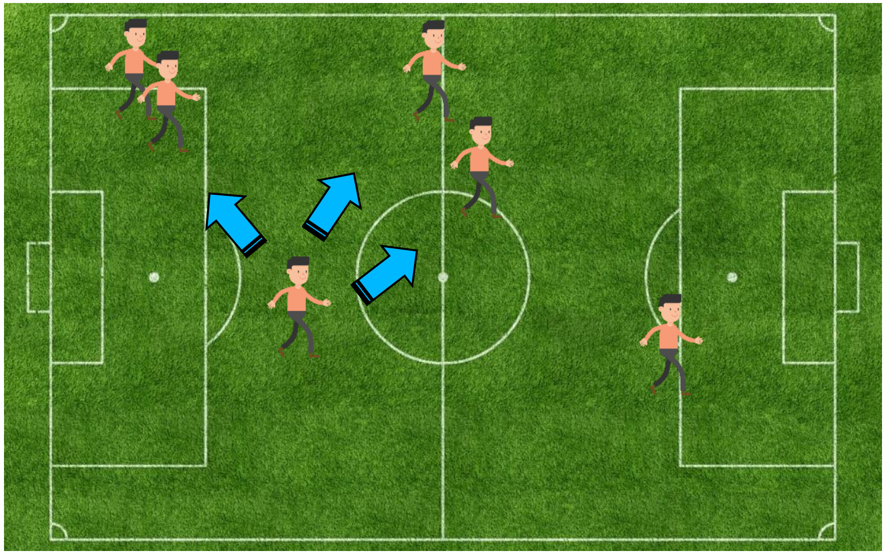
Aaron Fiske[a]
Aaron.Fiske@alum.dartmouth.org

Jason Liu[c]
jasonliu@mines.edu

Christopher Masone[b]
Christopher.Masone@dartmouth.edu

Susan McGrath[b]
smcgrath@ists.dartmouth.edu

Yougu Yuan[a]
yuanyg@cs.dartmouth.edu



Example

Outdoor Experimental Comparison of Four Ad Hoc Routing Algorithms

Robert S. Gray[b]
robert.s.gray@dartmouth.edu

David Kotz[a]
dfk@cs.dartmouth.edu

Calvin Newport[a]
Calvin.Newport@alum.dartmouth.org

Nikita Dubrovsky[a]
Nikita.Dubrovsky@alum.dartmouth.org

Aaron Fiske[a]
Aaron.Fiske@alum.dartmouth.org

Jason Liu[c]
jasonliu@mines.edu

Christopher Masone[b]
Christopher.Masone@dartmouth.edu

Susan McGrath[b]
smcgrath@ists.dartmouth.edu

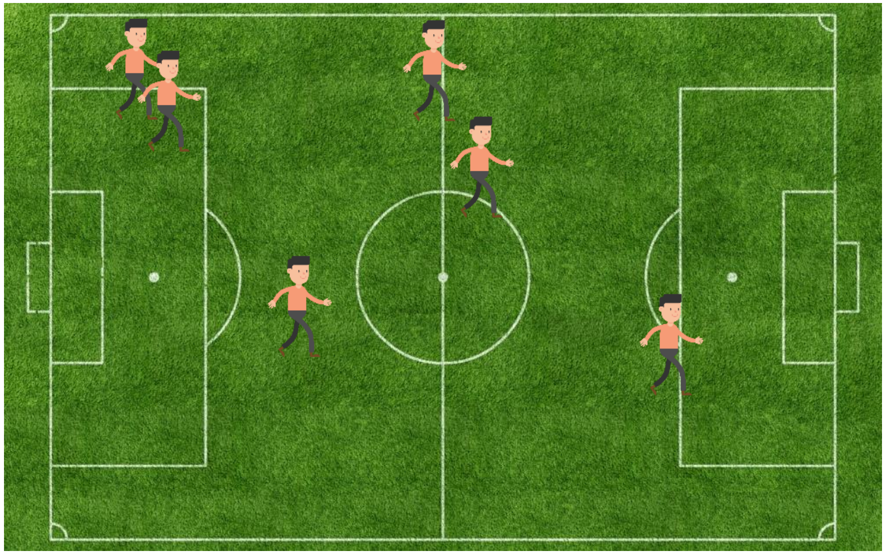
Yougu Yuan[a]
yuanyg@cs.dartmouth.edu

- 40 Laptops, 4 Routing algorithms

	Message Delivery Ratio
AODV	0.50
APRL	0.20
ODMRP	0.77
STARA-S	0.08

Are they really seeing the whole picture?

Example



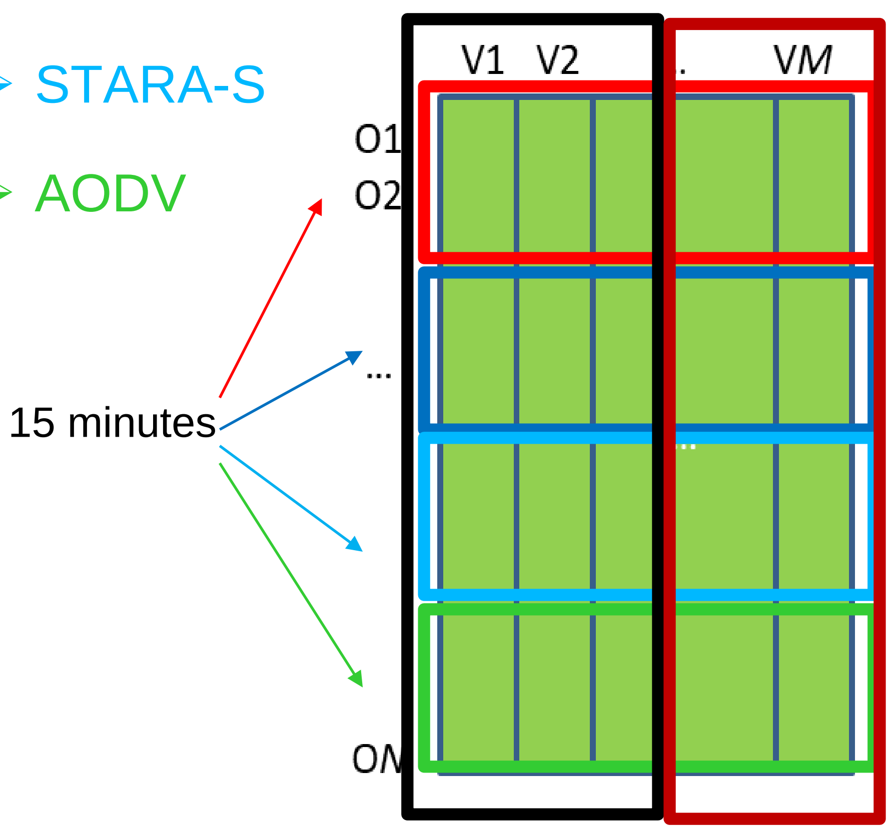
	Message Delivery Ratio
AODV	0.50
APRL	0.20
ODMRP	0.77
STARA-S	0.08

PD: Average of Distances
mM: mínimum of Maximum Distances
Mm: Maximum of mínimum Distances
cX: centroid X
cY: centroid Y
cZ: centroid Z
n1: # Users very close
n2: # Users close
n3: # Users far
n4: # Users very far

nTI: # TIN
nTO: # TOUT
nSI: # SIN
nSO: # SOUT
vTI: Vol TIN
vTO: Vol TOUT
vSI: Vol SIN
vSO: Vol SOUT

Example

- APRL
- ODMRP
- STARA-S
- AODV



PD: Average of Distances
mM: mínimum of Maximum Distances
Mm: Maximum of mínimum Distances
cX: centroid X
cY: centroid Y
cZ: centroid Z
n1: # Users very close
n2: # Users close
n3: # Users far
n4: # Users very far

nTI: # TIN
nTO: # TOUT
nSI: # SIN
nSO: # SOUT
vTI: Vol TIN
vTO: Vol TOUT
vSI: Vol SIN
vSO: Vol SOUT

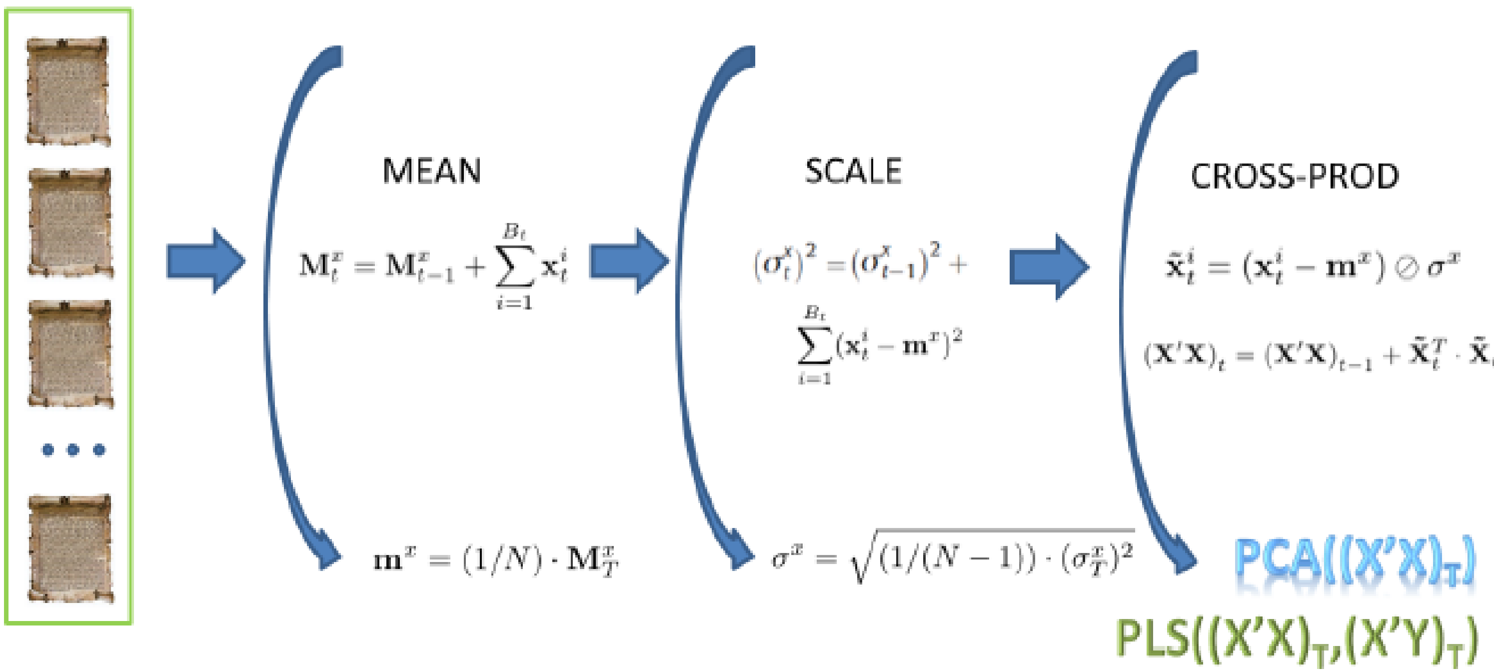
Extensions: BIG DATA

- The problem

OR

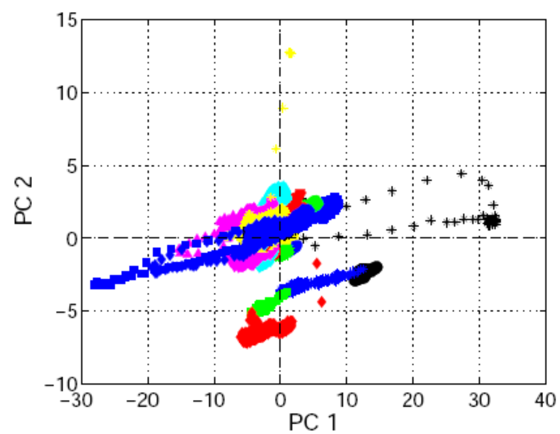
MEDA for BIG DATA

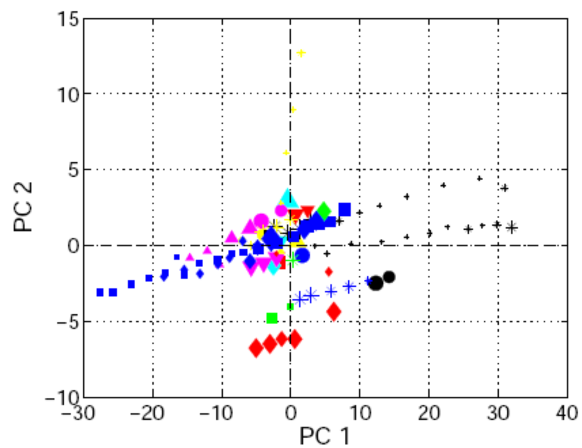
- Millions of observations (or more)
  - Problem 1: we cannot compute models
    - Solution: We can do it in batches



MEDA for BIG DATA

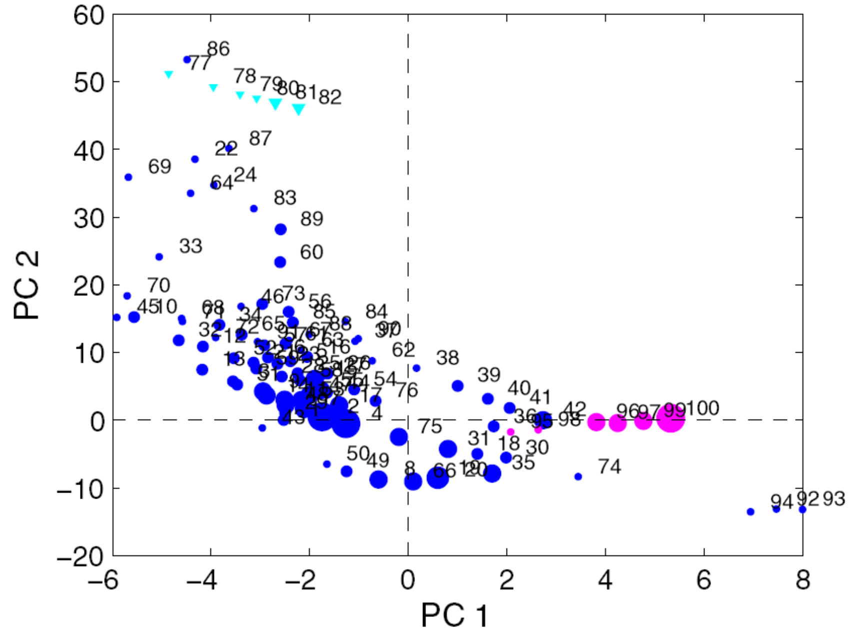
- Millions of observations (or more)
  - Problem 1: we cannot compute models
    - Solution: We can do it in batches
  - Problem 2: we cannot visualize millions of observations
    - Solution: Clustering

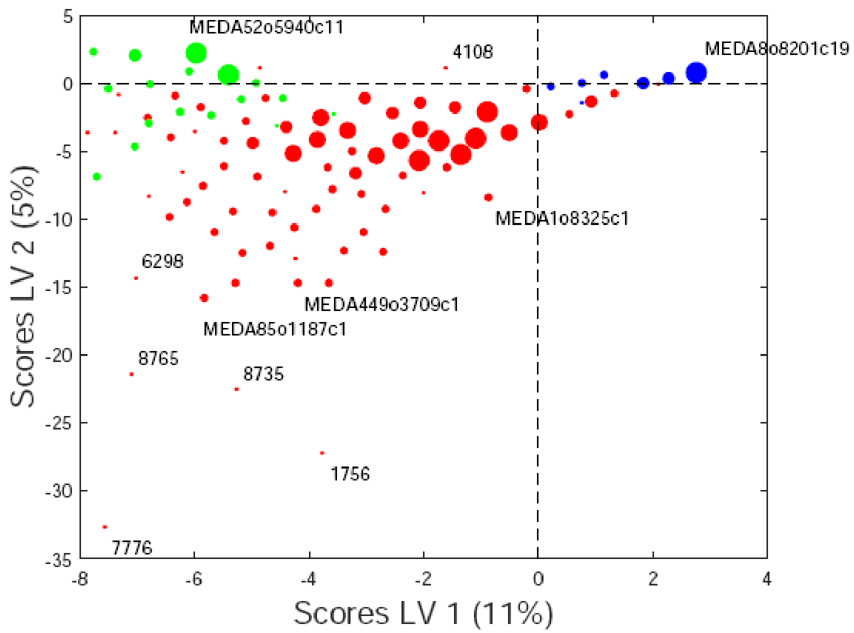




KDD Example

- Millions of variables
  - X = 5.000.000 x 100

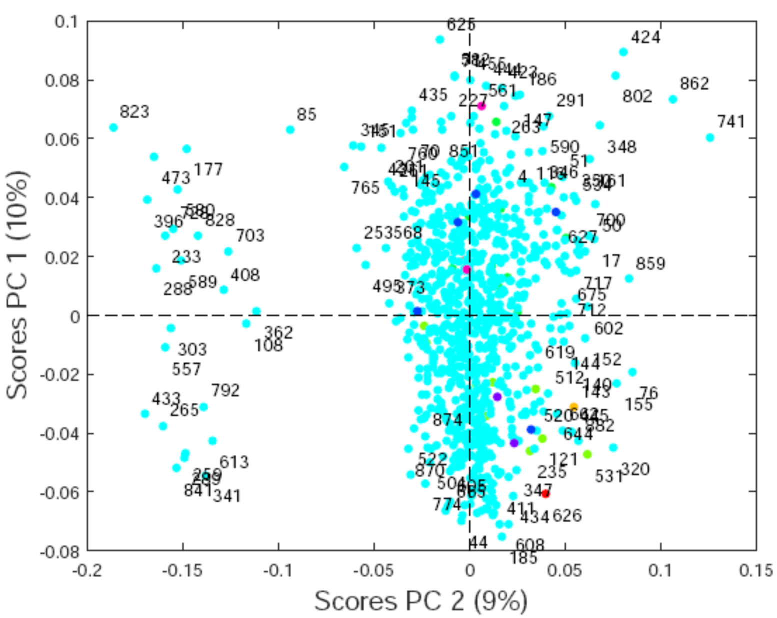


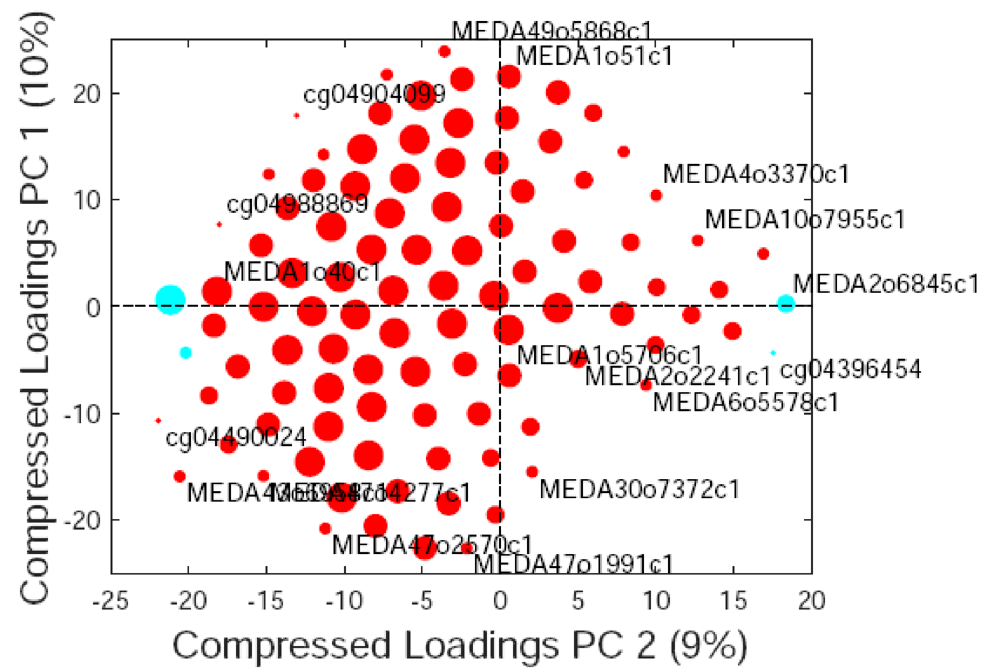


ChemoLab, (2014) 135: 110

Cancer Example

- Millions of variables
  - Same problem, just transposed
    - Example: BRCA. DNA Methylation: 900 x 500,000

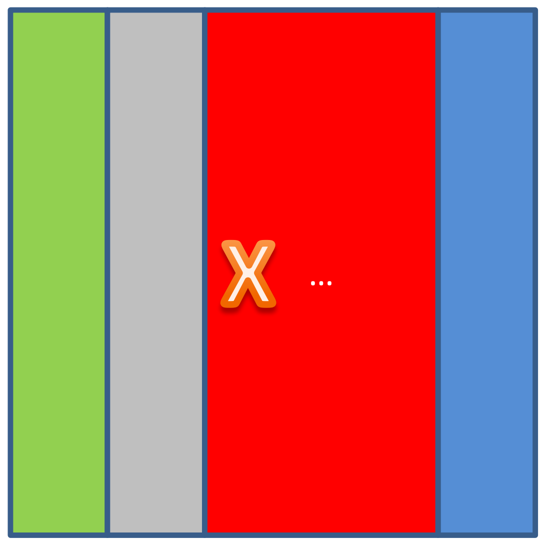




MEDA
University of Granada
José Camacho, Ph.D.

Feature & Observations Eng.

Definition of the features



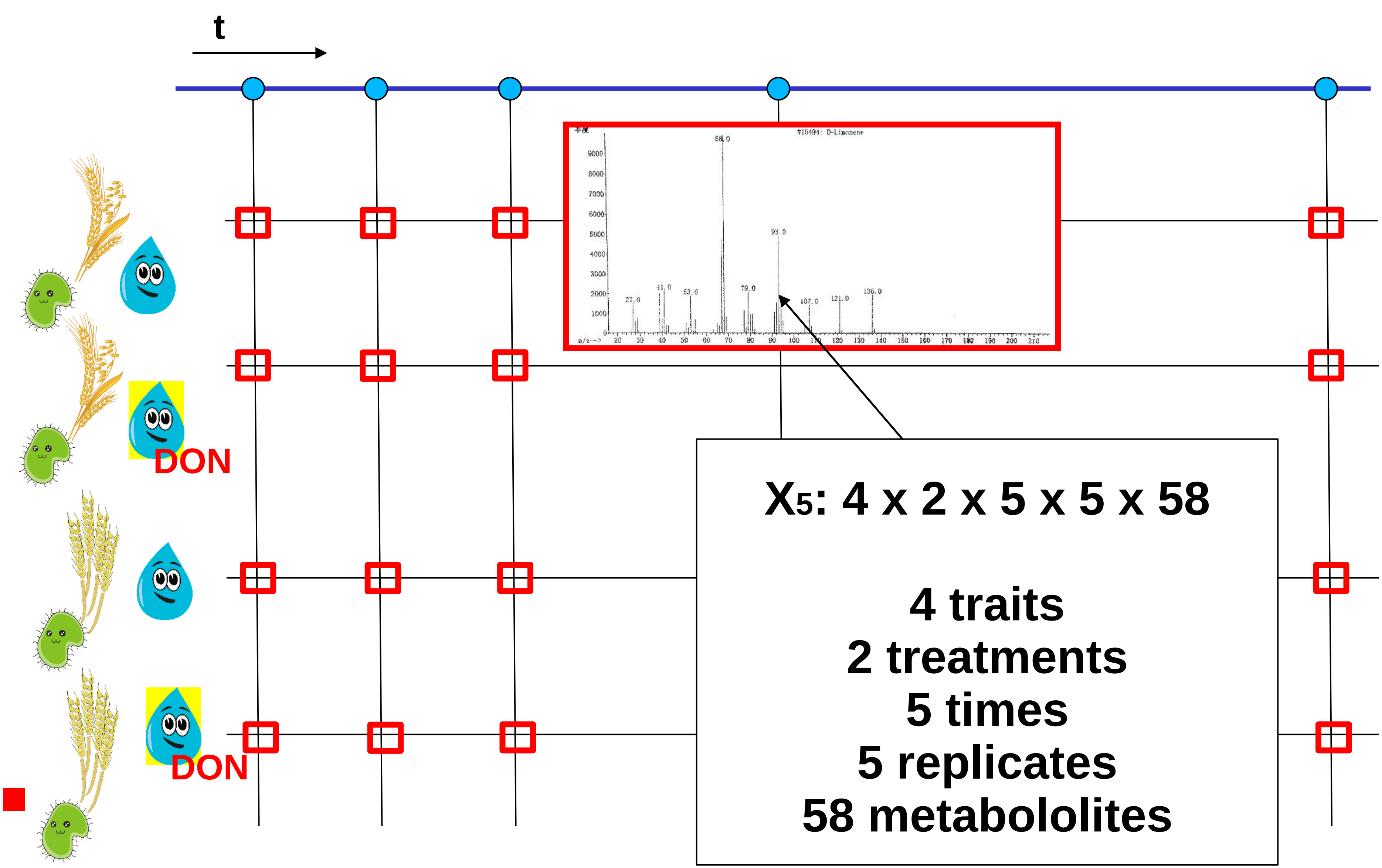
The **features** are the parameters that will be computed for the observations

Definition of the observations

The **observations** are the items or entities that are contrasted in terms of the value of their features.

At the end, it al boils down to the **specific questions** we pose

F&O Eng: Wheat data

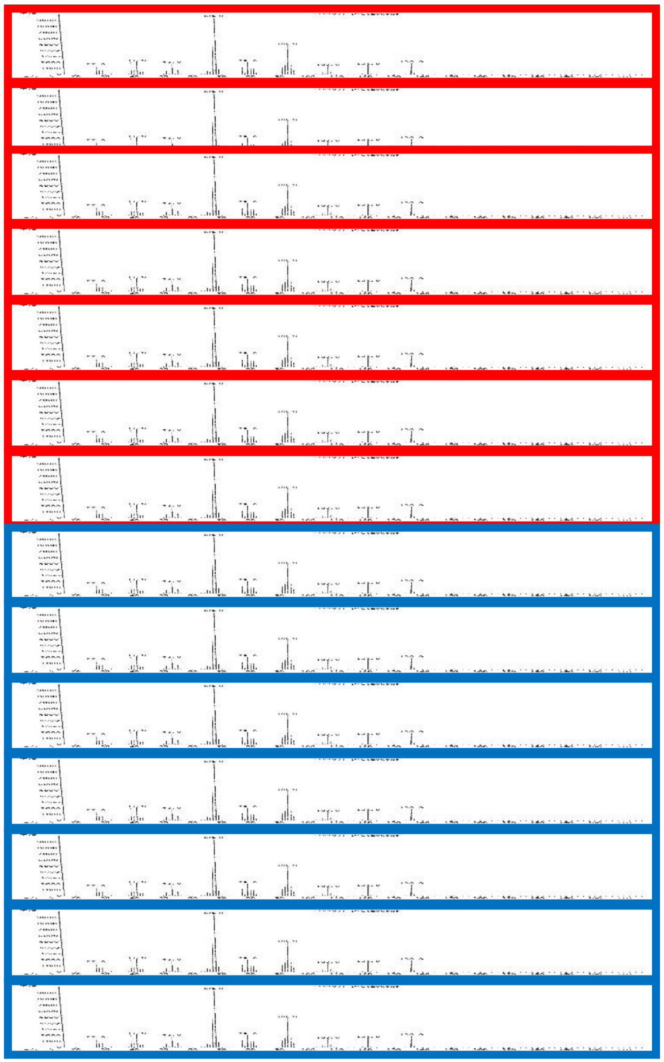


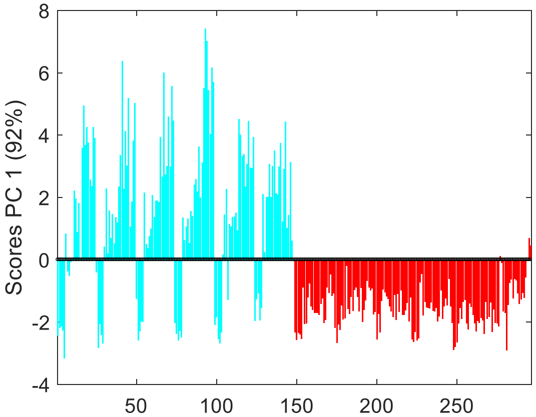
F&O Eng: Wheat data

X_4: 4 x 2 x 5 x 5 x 58

X_2: [4·2·5·5] x 58

0
12
24
48
96
…
…
96
0
12
24
48
96
…
…
96





MEDA
University of Granada
José Camacho, Ph.D.

F&O Eng: Wheat data

X_4: 4 x 2 x 5 x 5 x 58

X_2: [4·2·5] x [5·58]