Marie Skłodowska-Curie Actions
Innovative Training Networks (ITN)
European Joint doctorate (EJD)
H2020-2015

METHODS IN RESEARCH ON RESEARCH
Network Meeting & Training Event in Split
1-5 October 2018

Roser Rius
UPC

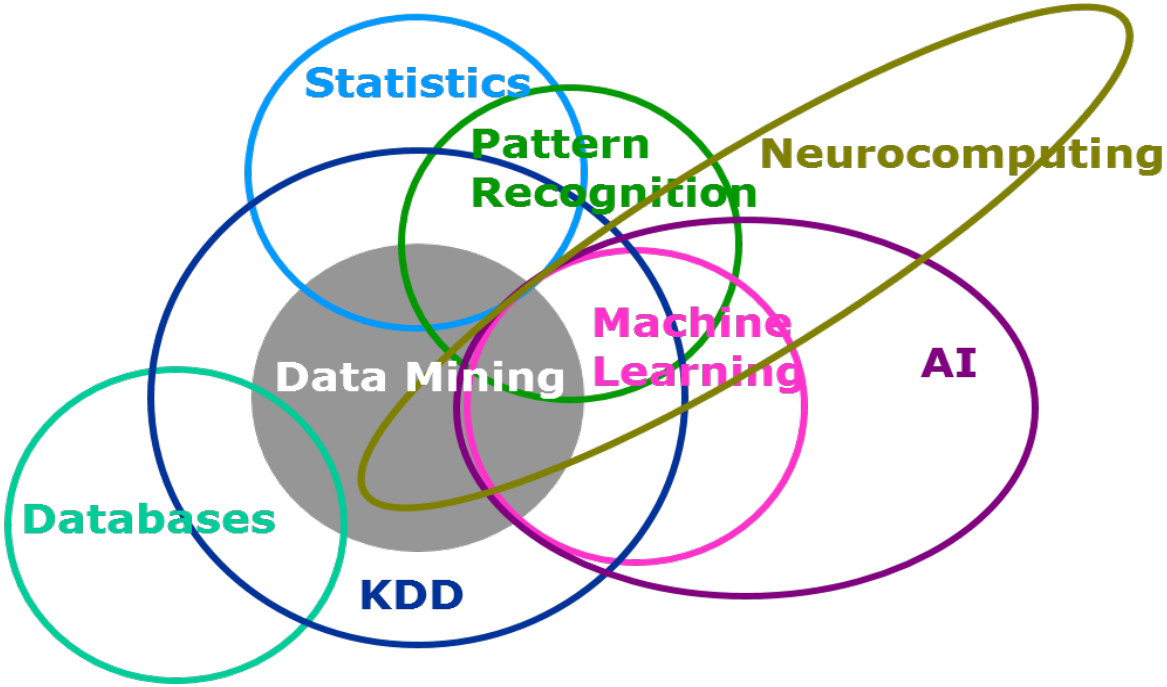
MiRoR
Methods in
Research on Research

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207

Content analysis: documents communication

Docs (text, video, audio,…)

Patterns Groups Relations Associations



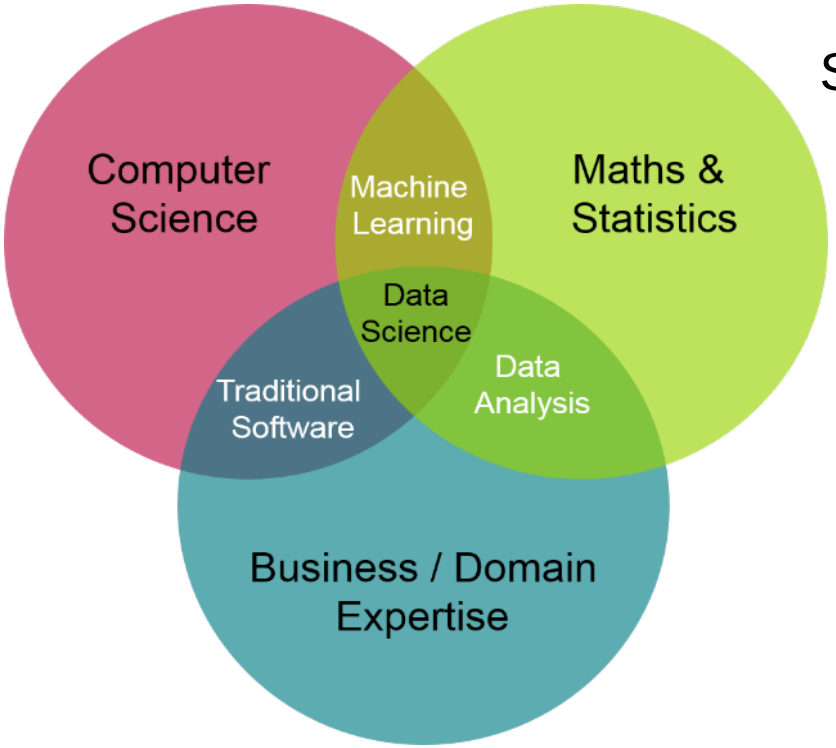
+ Big data, Text mining, Data Science, …
  Learning (supervised, unsupervised, deep) …

Iterative solutions

Computability

Mathematical solution

Significance



Interpretability. Results (poor, obvious, confirmatory)

INFERENTIAL STATISTICS	versus	DESCRIPTIVE STATISTICS
sample, hypothesis significance theoretical distribution		Univariate / Bivariate / Multivariate

MULTIVARIATE DESCRIPTIVE ANALYSIS
(MULTIDIMENSIONAL STATISTICS)

MULTIVARIATE DESCRIPTIVE ANALYSIS

"GEOMETRICAL" APPROACH mathematical solution singular value decomposition (PCA, MCA,…)	versus	ITERATIVE APPROACH iterative algorithms (Factor analysis, clustering…)
DIMENSIONALITY REDUCTION (PCA, Factor anaysis, …)	versus	"CLASSIFICATION" (Clustering, …)

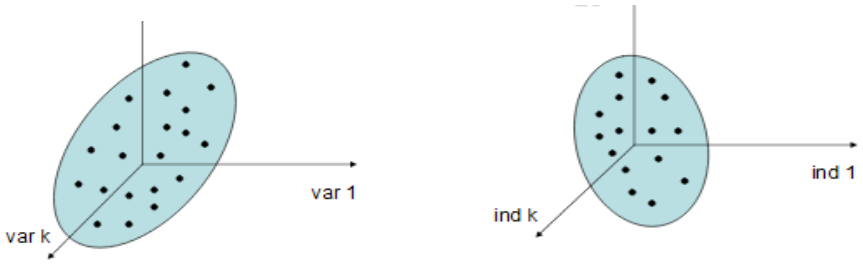
with
Complex Data (continous, categorical, specific, textual,…)
and
Complex results (graphical, partial, multifaceted, …)

⇩

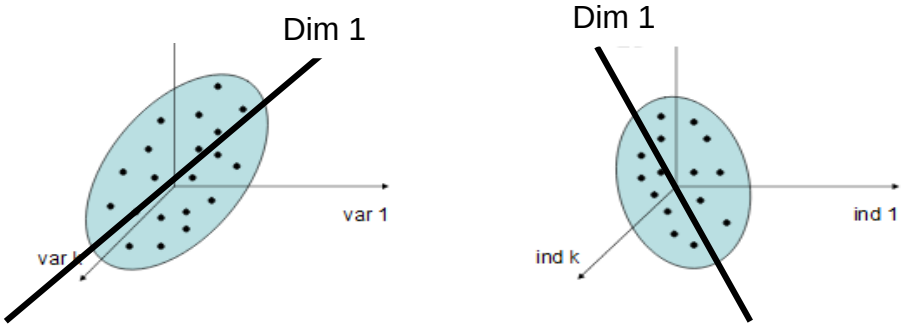
Combination of techniques and Preprocessing data

DIMENSIONALITY REDUCTION

$X\ (n \times p)$ $\quad (\Re^p, M)$ $\quad (\Re^n, N)$



$$X = V \Lambda^{1/2} U^t \quad \text{with} \quad U^t M U = I_{r \times r} \quad \text{and} \quad V^t N V = I_{r \times r}$$



$$\Psi = X M U \qquad \Phi = X^t N V$$

DIMENSIONALITY REDUCTION

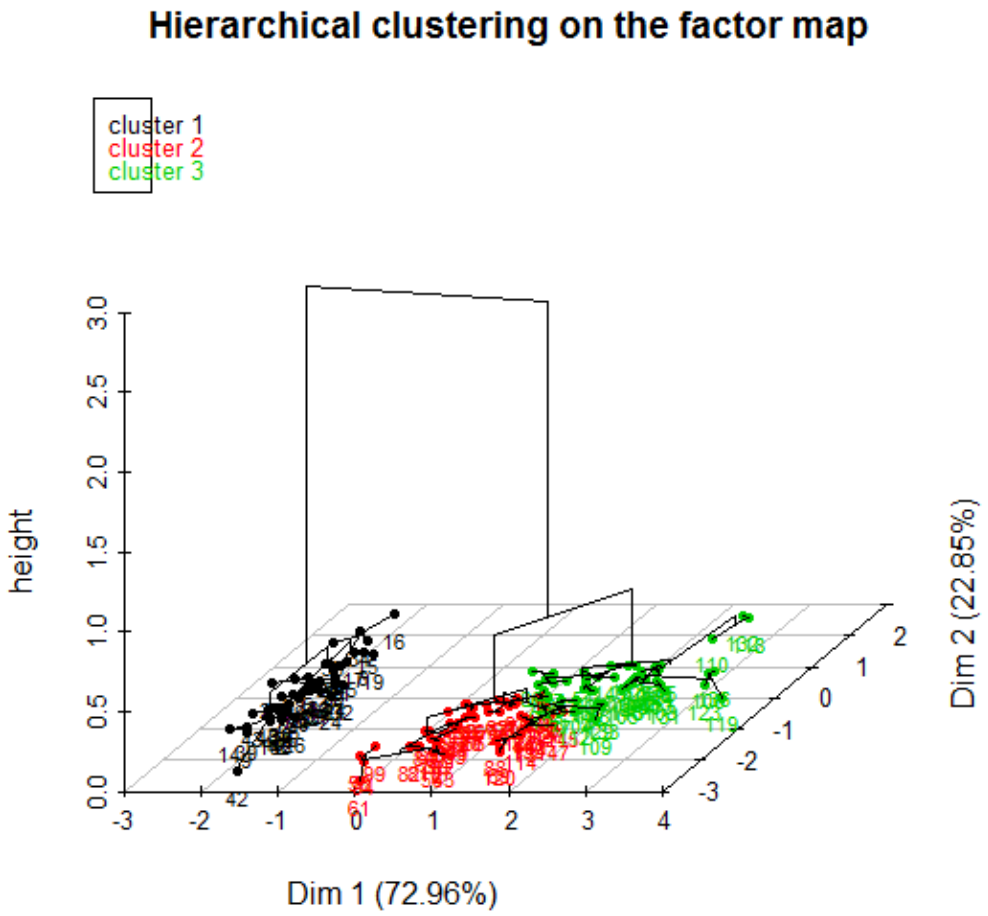
PCA

$$d^2(x_i\ ,\ x_{i\prime}) = \sum_{j=1}^{p} \left(\frac{x_{ij}}{s_j} - \frac{x_{i'j}}{s_j} \right)^2$$

CA

$$d^2_{\chi 2}(f_i\ ,\ f_{i\prime}) = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \frac{1}{f_{.j}} \qquad d^2_{\chi 2}(f_j\ ,\ f_{j\prime}) = \sum_{i=1}^{q} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij\prime}}{f_{.j\prime}} \right)^2 \frac{1}{f_{i.}}$$

CLUSTERING


Hierarchical clustering on the factor map
cluster 1
cluster 2
cluster 3
height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-3
-2
-1
0
1
2
3
4
Dim 1 (72.96%)
Dim 2 (22.85%)
2
1
0
-1
-2
-3

EXAMPLES in R (Rstudio)

Install packages:

- FactoMineR
- missMDA
- RColorBrewer
- SnowballC
- tm
- wordcloud

```
library(FactoMineR)      # PCA MCA
library(missMDA)         # missing data
library(RColorBrewer)    # color palettes
library(SnowballC)       # text stemming
library(tm)              # text mining
library(wordcloud)       # word-cloud generator
```

EXAMPLES in R (Rstudio)

PCA.R
PCA+HCPC.R

MCA.R

missMDA.R

wordcloud_Ihaveadream.R

YoungersIdentity.R