Data, Metadata

K. Gibert

Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group at
Intelligent Data Science and Artificial Intelligence Specific Research Center

University Institute of Research on Science and Technology of Sustainability

Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona

Karina.gibert@upc.edu
https://www.eio.upc.edu/en/homepages/karina

©K. Gibert

Basic structure for analysis

The data matrix

	Weight	Height	Sex	Eyes
John	85	1.85	M	azul
	.	.	.	.
	.	.	.	.

Point cloud (video)

Rows: Individuals (study units) (i1….in)

Columns: Variables (characteristics of individuals) (X1..Xk)

Cells: Value of variables for individuals (xik)

https://www.youtube.com/watch?v=4pnQd6jnCWk

©K. Gibert

Type of variables

- Numerical: Quantitative, measure
  - continuous (real quantity): *Weight, Height*
  - discrete (natural quantity): *Age, shoes size*
- Categoric: Qualitative, adjective
  (evenctually codified)
  - Ordinal (ordering over modalities): *Socioecnonomic status*
  - Binary (two modalities): *wear glasses*
  - Nominal (unordered modalitites): *Hair color*
- Date: Special formats, only some softwares
- Other variables
  (no standard, rarely used in standard data mining applications)
  - Fuzzy variables
  - Compositional variables
  - Symbolic variables
  - Distributional variables
  - Interval variables/Ratio variables (means, standard dev, dotplots)
  - Textual data

Categorization
Discretization

Mean/StDev
Histograms

RECategorization

Percentages
Tables
BarPlots

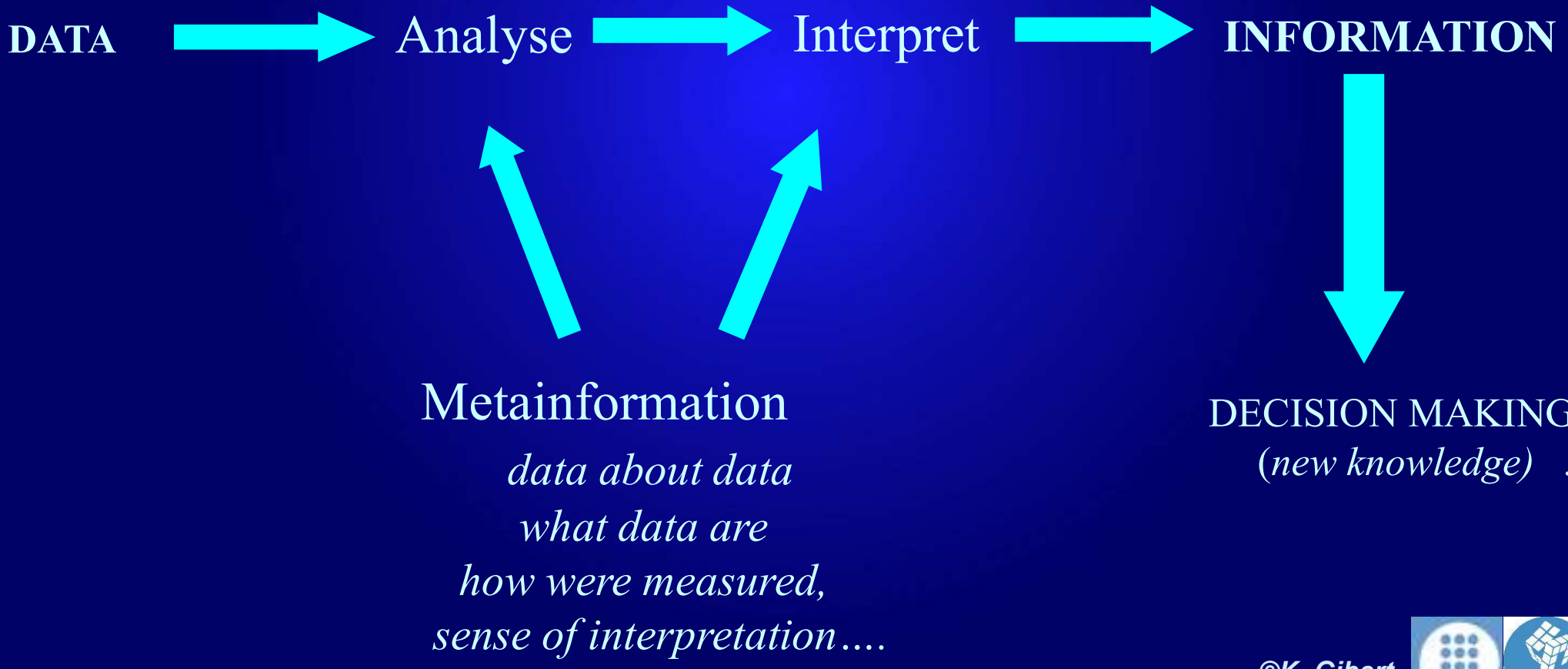
Loss information

Better avoid

©K. Gibert

From Data to Decisional Knowledge

DATA <> **INFORMATION**



©K. Gibert

Metadata File

url: *www.xxx.ssss.www*

Inclusion criteria: *People in [18,65] years, no hard attacks, no smoking, no cholesterol, married, with sons or daughters….*

n*: nro of rows*

K*: nro of columns*

Variable	Modalities	meaning	Type	Measuring unit	Missing code	Measuring procedure	Range	Role
Age		Age of marriage	Num	years	"*"		[1,105]	Explanatory
Sex		Gender	Quali		Unknown			Explanatory
	M	Male						
	H	Female						
FeC		Level of Iron in blood	Num	μg/dl	NA	Biochemical analysis on blood sample measuring transferrine ……	[30, 200]	Explanatoyr
Anemy		The person has anemy diagnosis	Boolean		Unknown	Levels of Fec<xxx and …….		Response

©K. Gibert

First insight to Data

- Look at Metadata
- Determine rows and columns to be kept for the analysis
- Basic descriptive analysis of remanining variables
  - Inspect anomalies, errors, missing data, outliers
- First report about data quality
- Preprocessing
- Verify after each processing step
- Final descriptive analysis *(report data improvements)*

©K. Gibert

Data, Metadata

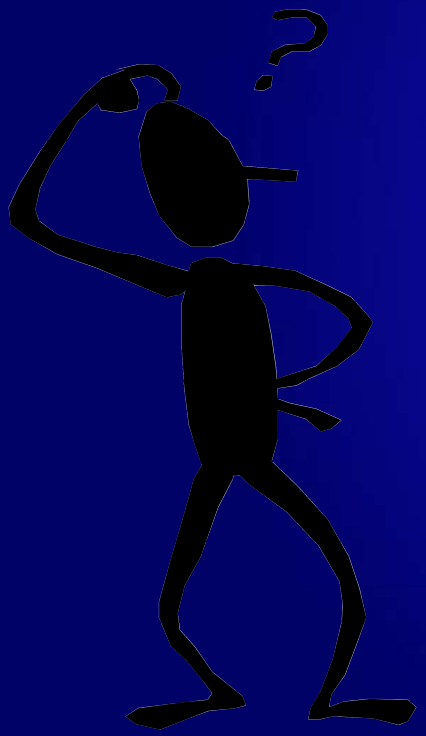
Karina Gibert

Dpt. Statistics and Operation Research

Knowledge Engineering and Machine Learning Research group at
Intelligent Data Science and Artificial Intelligence Specific Research Center

Institut Universitari de Recerca en Ciència y Tecnologia de la Sostenibilitat
Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu
www.eio.upc.edu/homepages/karina



Are there any questions?...

©K. Gibert