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

Type of variables

•Numerical: Quantitative, measure

Categorization
Discretization

    continuous (real quantity):   *Weight, Height*

    discrete (natural quantity):   *Age, shoes size*

Mean/StDev
Histograms

•Categoric: Qualitative, adjective
    (evenctually codified)
     Ordinal (ordering over modalities):   *Socioecnonomic status*
    Binary (two modalities):   *wear glasses*
    Nominal (unordered modalitites):   *Hair color*

RECategori zation

Percentages
Tables
BarPlots

•Date: Special formats, only some softwares
•Other variables
(no standard, rarely used in standard data mining applications)
•Fuzzy variables

Loss information

•... variables
•... variables
•... variables
•Distributional variables
•Interval variables/Ratio variables (means, standard dev, dotplots)
•Textual data

Better avoid
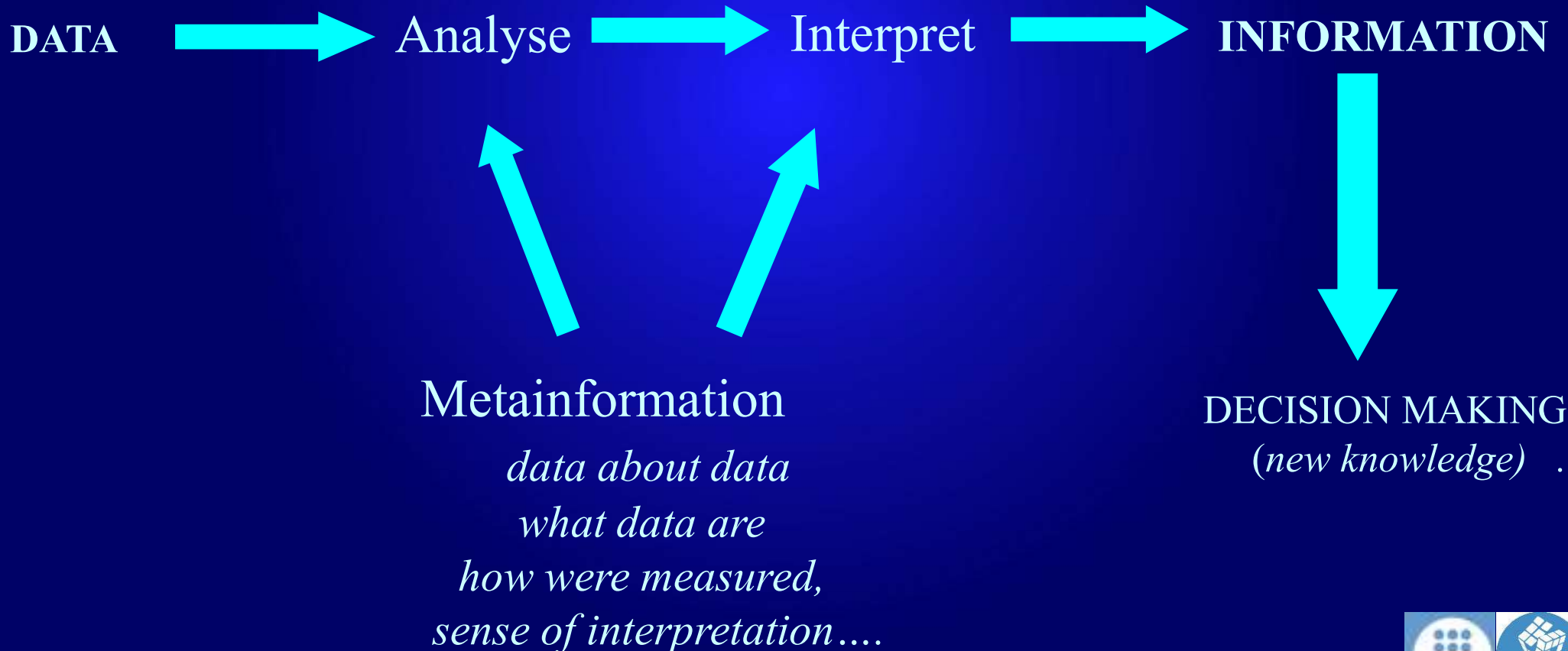
From Data to Decisional Knowledge

DATA **<>** **INFORMATION**

DATA ➡ Analyse ➡ Interpret ➡ **INFORMATION**

⬆ ⬆ ⬇

Metainformation

*data about data
what data are
how were measured,
sense of interpretation….*

DECISION MAKING
(*new knowledge*) .

©*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

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?...