Data, Metadata

K. Gibert[1]

[1]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group
Universitat Politècnica de Catalunya, Barcelona

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)

Type of variables

- Numerical: Quantitative, measure
  - continuous (real quantity): *Weight, Height*
  - discrete (natural quantity): *Age, shoes size*

Categorization

Discretization

Mean/StDev
Histograms

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

REcategorization

Percentages
Tables
BarPlots

- Date: Special formats, only some softwares
- Other variables

  (no standard, rarely used in standard data mining applications)
  - Fuzzy variables
  - Hierarchical variables
  - Multivalued variables
  - Semantic variables
  - Distributional variables
  - Interval variables/Ratio variables (means, standard dev, dotplots)
  - Textual data

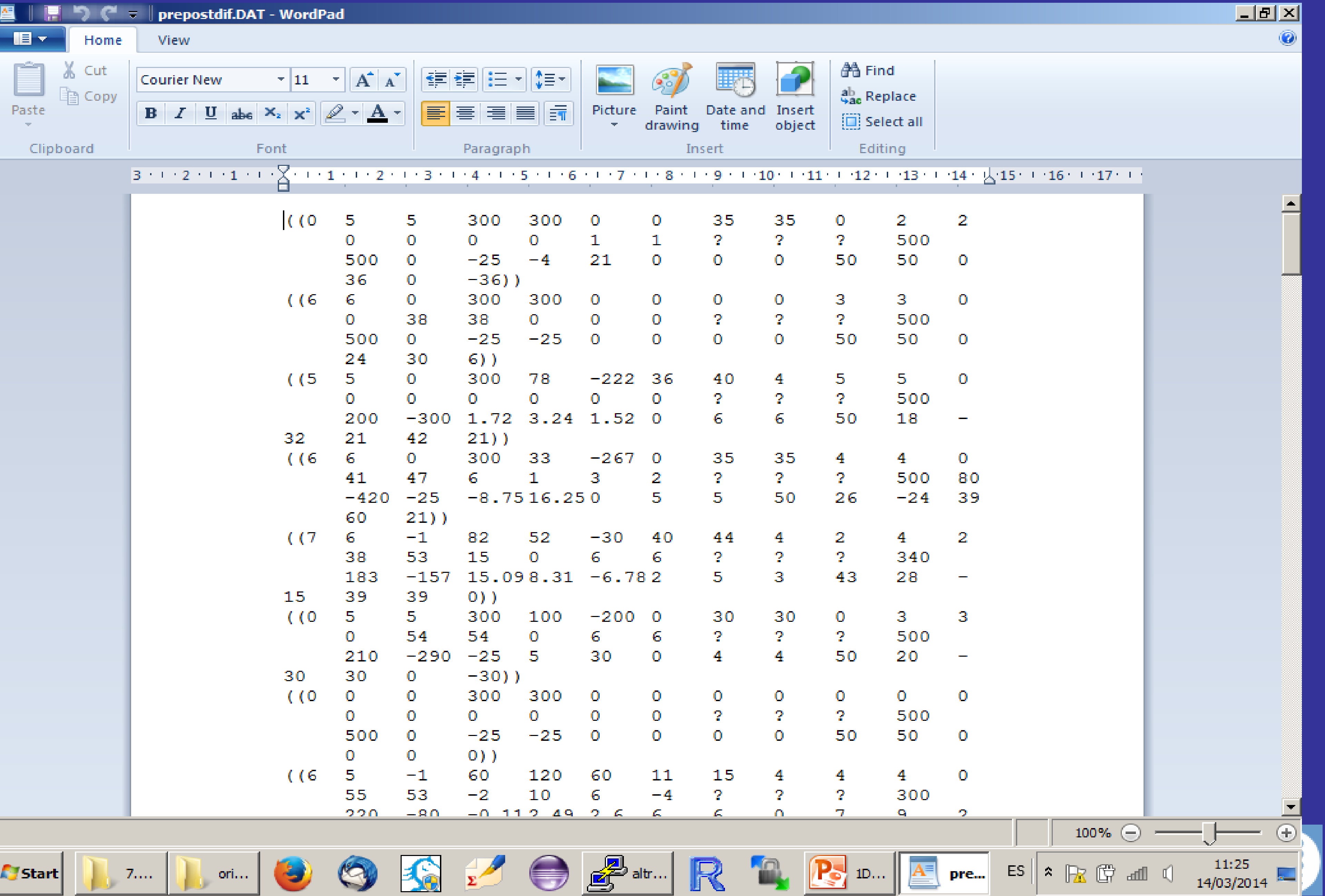
Loss information

Better avoid

©K. Gibert

From Data to Decisional Knowledge

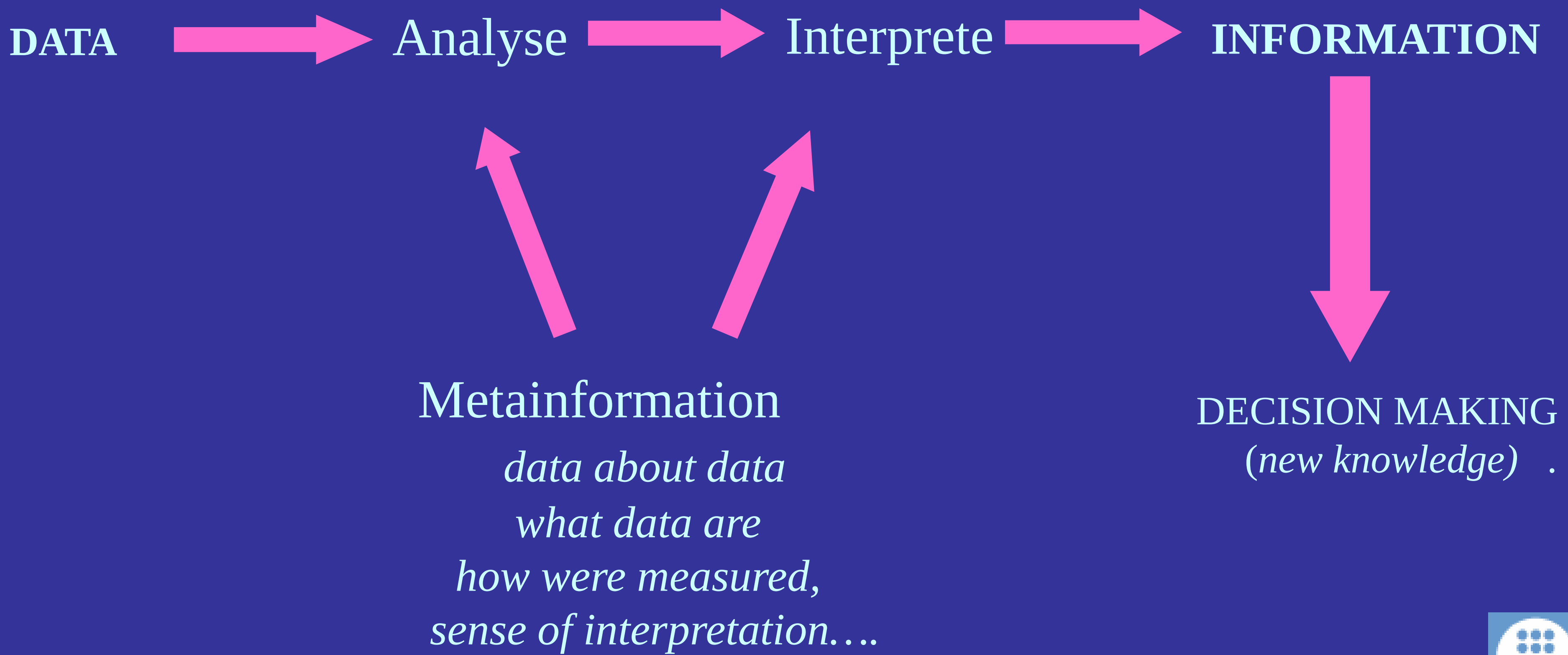
DATA <> INFORMATION



```
((0  5    5    300  300  0    0    35   35   0    2    2
     0    0    0    0    1    1    ?    ?    ?    500
     500  0    -25  -4   21   0    0    0    50   50   0
     36   0    -36))
((6  6    0    300  300  0    0    0    0    3    3    0
     0    38   38   0    0    0    ?    ?    ?    500
     500  0    -25  -25  0    0    0    0    50   50   0
     24   30   6))
((5  5    0    300  78   -222 36   40   4    5    5    0
     0    0    0    0    0    0    ?    ?    ?    500
     200  -300 1.72 3.24 1.52 0    6    6    50   18   -
32   21   42   21))
((6  6    0    300  33   -267 0    35   35   4    4    0
     41   47   6    1    3    2    ?    ?    ?    500  80
     -420 -25  -8.7516.250    5    5    50   26   -24  39
     60   21))
((7  6    -1   82   52   -30  40   44   4    2    4    2
     38   53   15   0    6    6    ?    ?    ?    340
     183  -157 15.098.31 -6.782    5    3    43   28   -
15   39   39   0))
((0  5    5    300  100  -200 0    30   30   0    3    3
     0    54   54   0    6    6    ?    ?    ?    500
     210  -290 -25  5    30   0    4    4    50   20   -
30   30   0    -30))
((0  0    0    300  300  0    0    0    0    0    0    0
     0    0    0    0    0    0    ?    ?    ?    500
     500  0    -25  -25  0    0    0    0    50   50   0
     0    0    0))
((6  5    -1   60   120  60   11   15   4    4    4    0
     55   53   -2   10   6    -4   ?    ?    ?    300
     220  -80  -0.112.49 2.6  6    6    0    7    9    2
```

From Data to Decisional Knowledge

DATA <> INFORMATION



DATA Analyse Interprete **INFORMATION**

DECISION MAKING
(new knowledge) .

Metainformation

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

©K. Gibert

Metadata

- Data Origin: Secondary source/Primary
- Inclusion criteria: Representativity? Target?
- Size of data: nxK (n>10K?)
- All variables:
    - What is it measuring (Measuring tool or procedure)
    - Measuring unit
    - Representation of missing data
    - Meaning of variable
- Quantitative variables:
    - Range of possible values
- Qualitative variables:
    - Set of possible modalities
    - Representation of modalities
    - Meaning of modalities
- Role of variables: Response/Explanatory

- Software often do not support metadata:
    - Extra document individualy managed as project documentation
    - Relational Data Base for very complex Data Matrices *[Gibert, MMR 92]*

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