Data Mining
(Mining Knowledge from Data)

Introduction to Data Mining and Visualization

Marcel Jiřina, Pavel Kordík

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FIT

Lecture

1) Subject organization
2) Introduction to data mining
3) Data and information visualization
4) Data matrix
5) RapidMiner

Subject organization

- Each week lecture and seminar
- No need to have deep theoretical foundations, often black-box approach is used
- Emphasis on practical application and interpretation of results
- Simple tasks, submitted to EDUX
- For those interested, there is the possibility of a deeper study of presented problems and algorithms
- Follow-up master's degree in Knowledge Engineering and PhD Studies

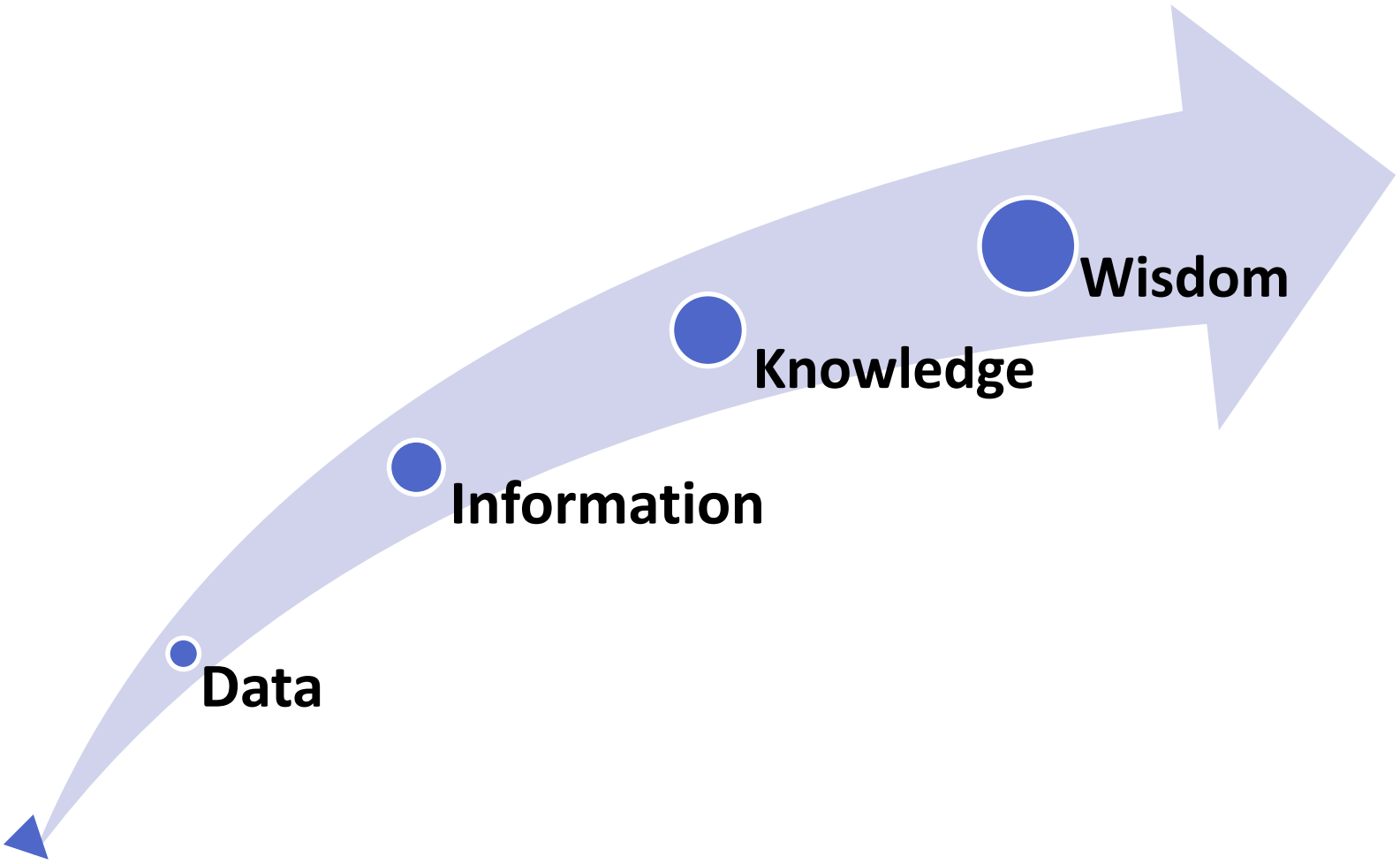
Evaluation/Ratings

- During the term, you can get 50 points from exercises . The test is for another 50 points.
- The mark consists of the sum of the points from exercises and the exam.
- During the semester, you can earn points for solving problems. Tasks are entered and checked during exercises.
- The minimum number of points from exercises is 25.
- The exam has a minimum score of 25 points.

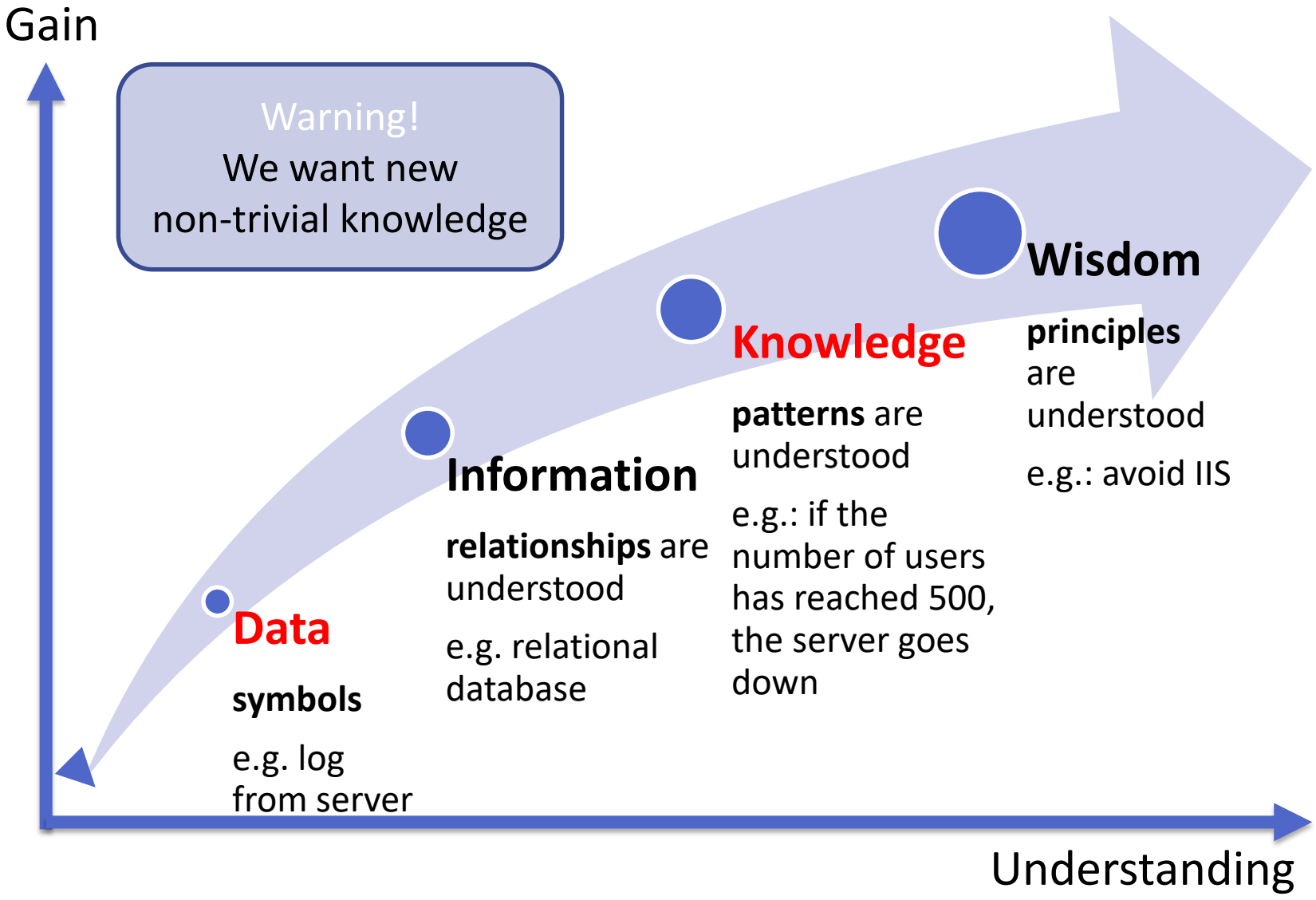
Data Mining

- **Data mining** is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
- It is an essential process where intelligent methods are applied to extract data patterns.
- It is an interdisciplinary subfield of computer science.
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data Mining (gaining knowledge from data)

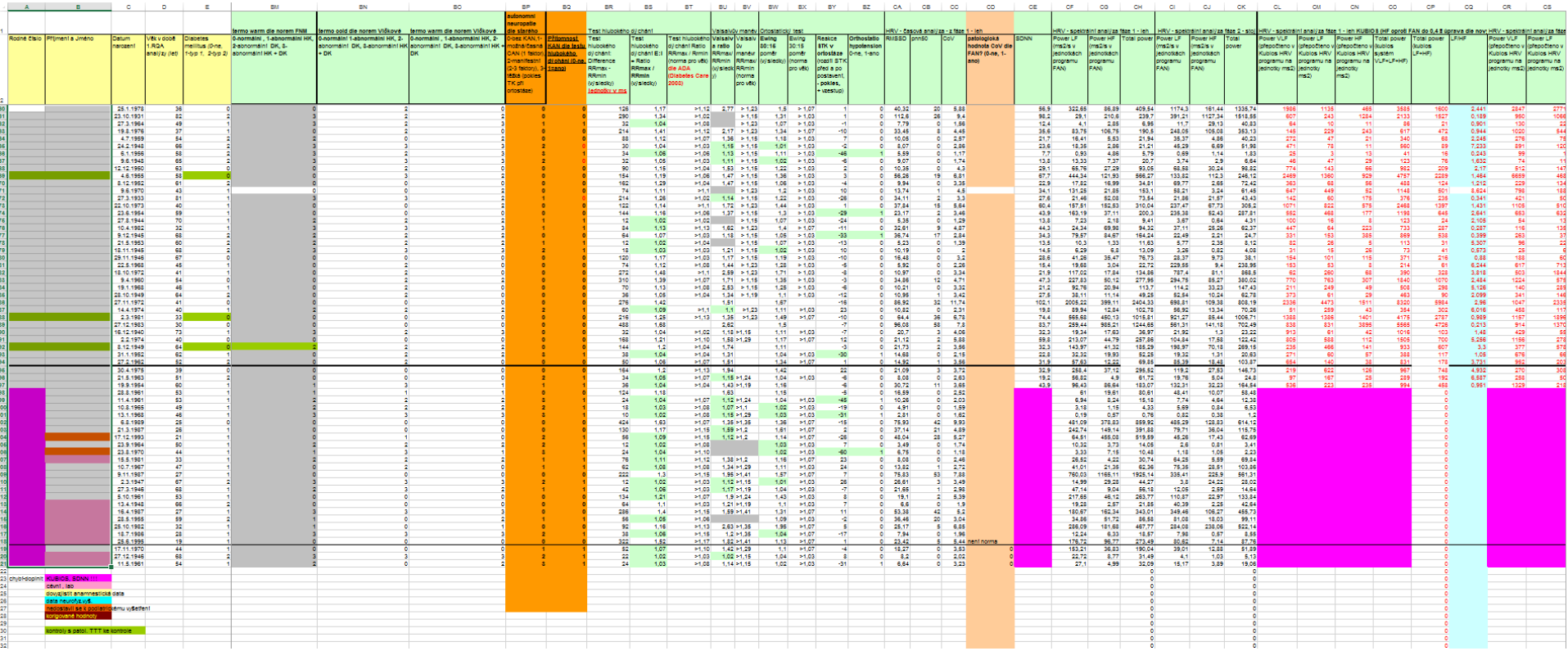


Will we be wise by mining the data?



Motivation

CAN detection – Motol Hospital – original data

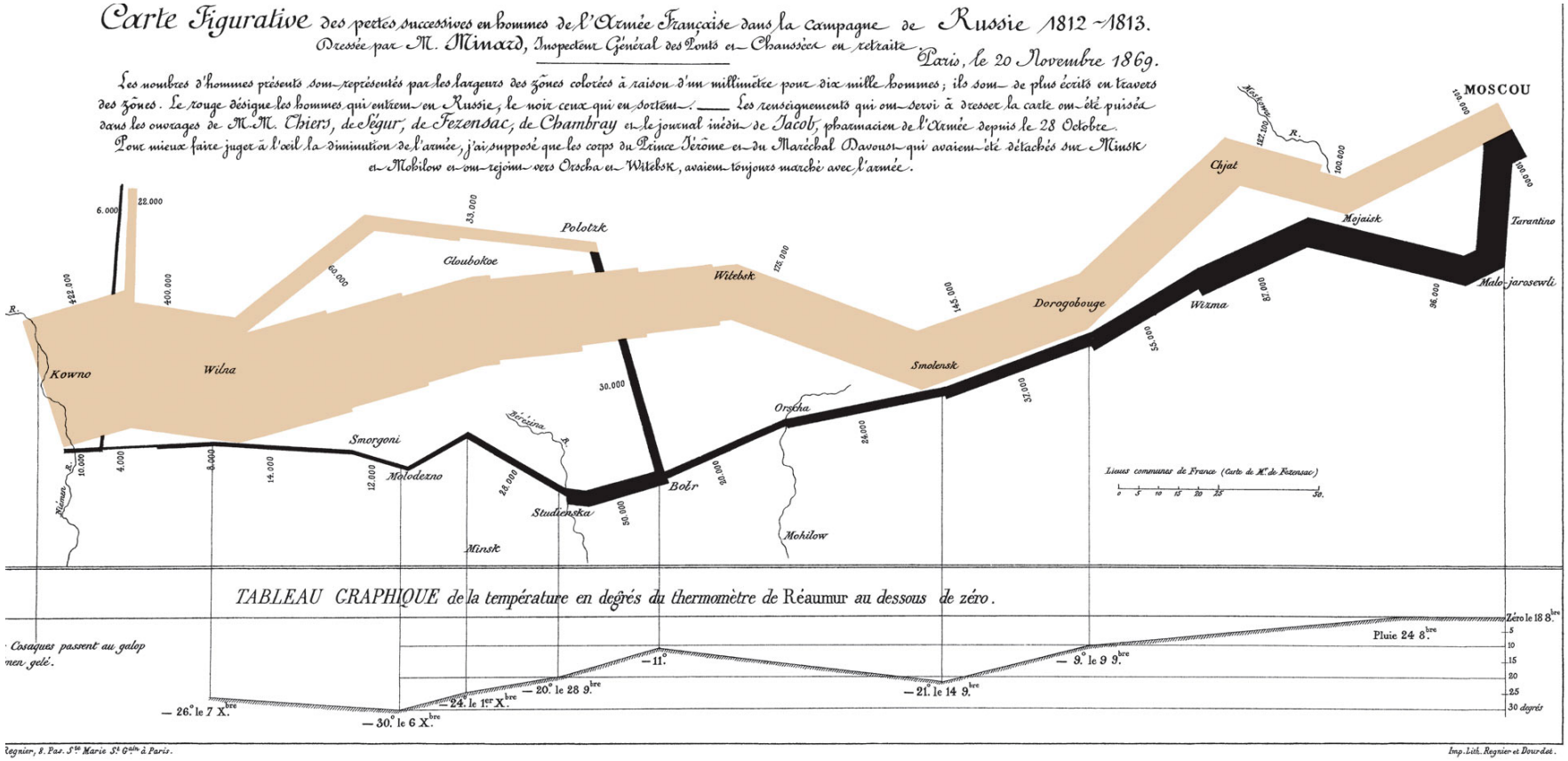


Visualization

- Very useful in all phases of data mining:

1. Preprocessing
   - Detection of missing values
   - Detection of outliers
   - Detection of non-normalized values
   - ... and many other problems ...
2. Search for patterns
3. Data representation
   - Often the best data representation for clients
   - Error detection

Minard's chart



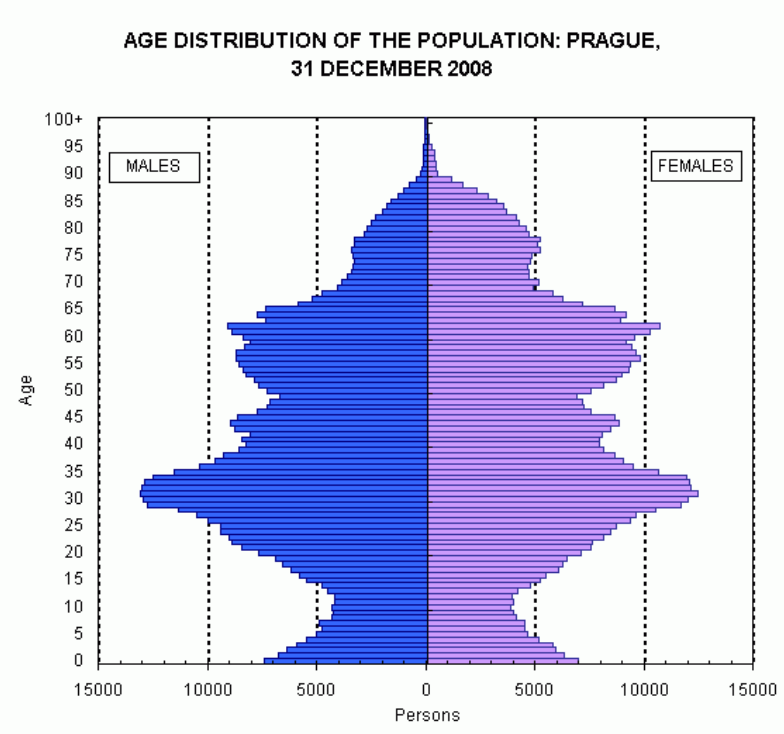
The famous illustrations (first infographics) of Napoleon's disastrous invasion of Russia in 1812.

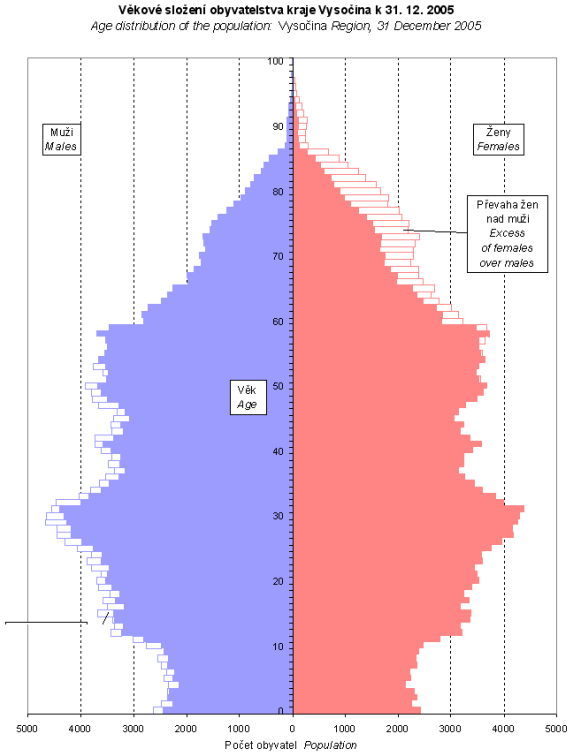
The cholera epidemic in London in 1854

- One of the first maps documenting epidemic.
- Dr. Snow discovered that cholera victims (dots) are close to a public water pump (crosses). Snow took this map at City Hall and on the next day the handle was dismounted from the pump.
- In the meantime, over 500 people died.

Basic charts for data visualization

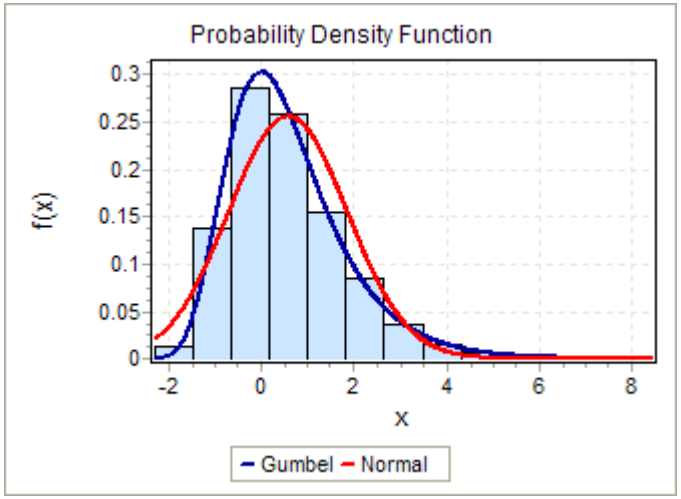
- Distribution charts, histogram
  - Distribution chart (probabilistic, density of probability)

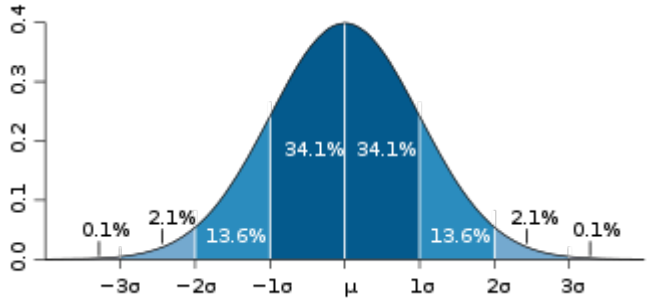




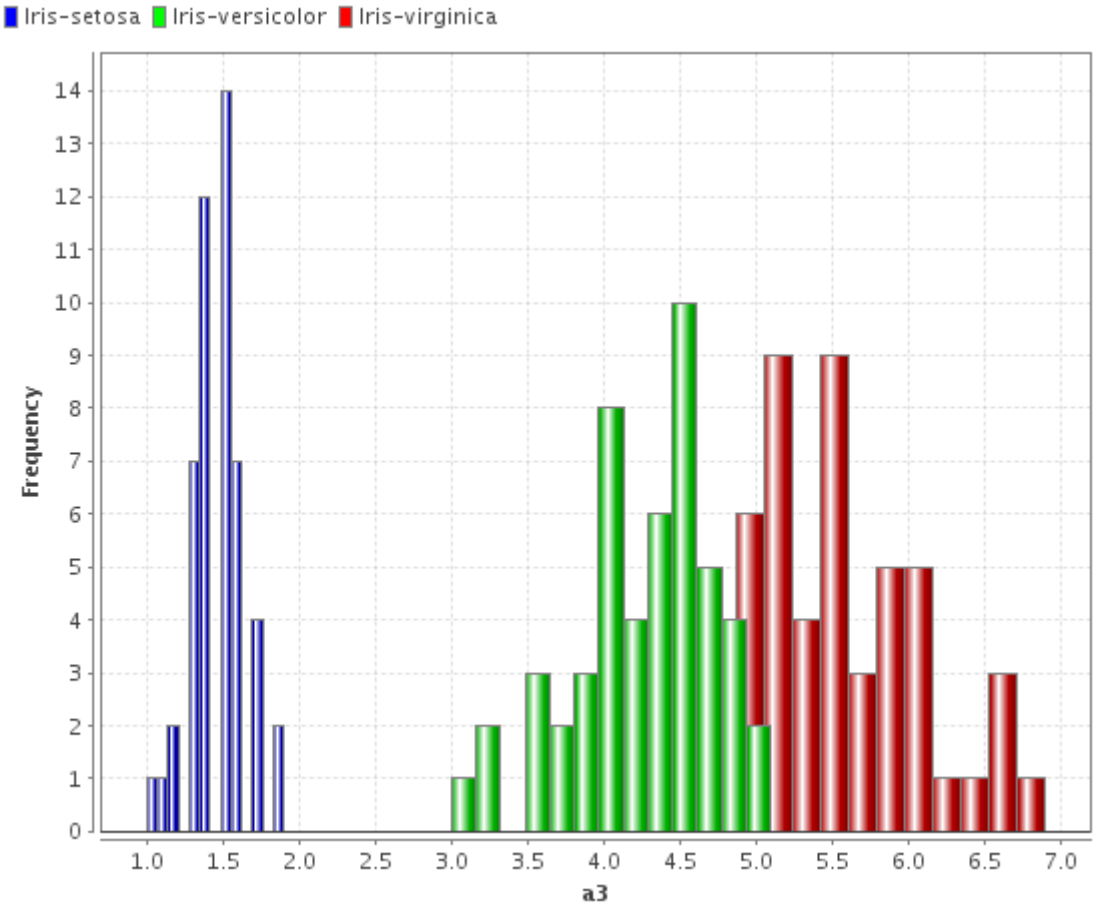
Basic charts for data visualization

- Distribution charts, histogram
  - Distribution chart (probabilistic, density of probability)





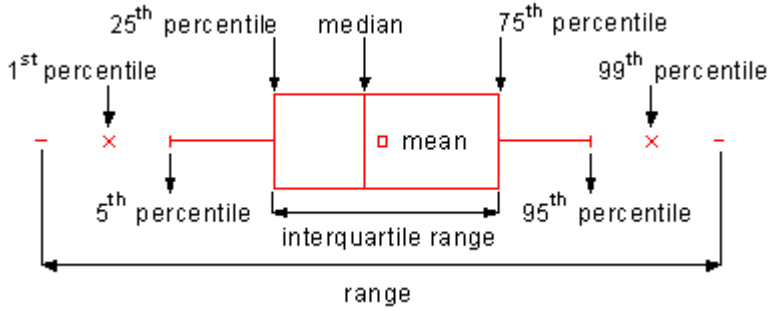
Histogram

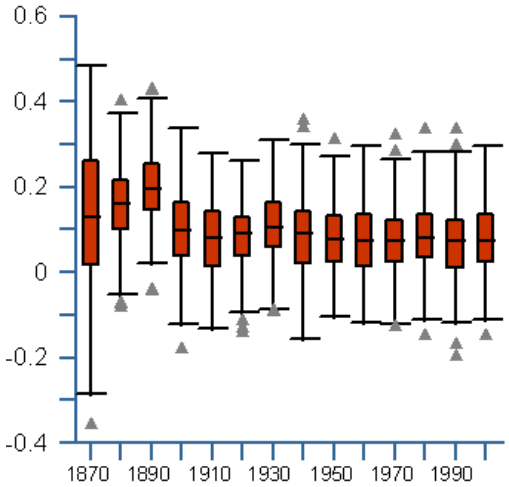


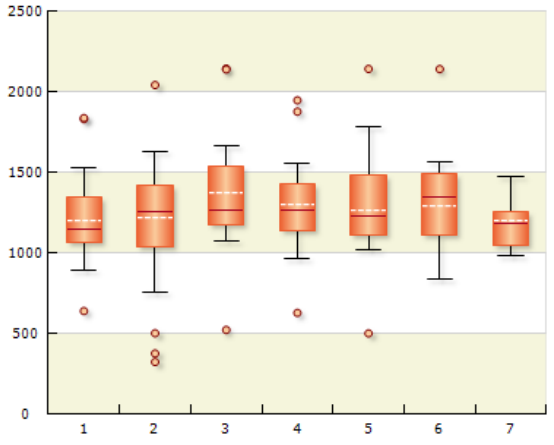
What is the ideal number of columns?

Basic charts for data visualization

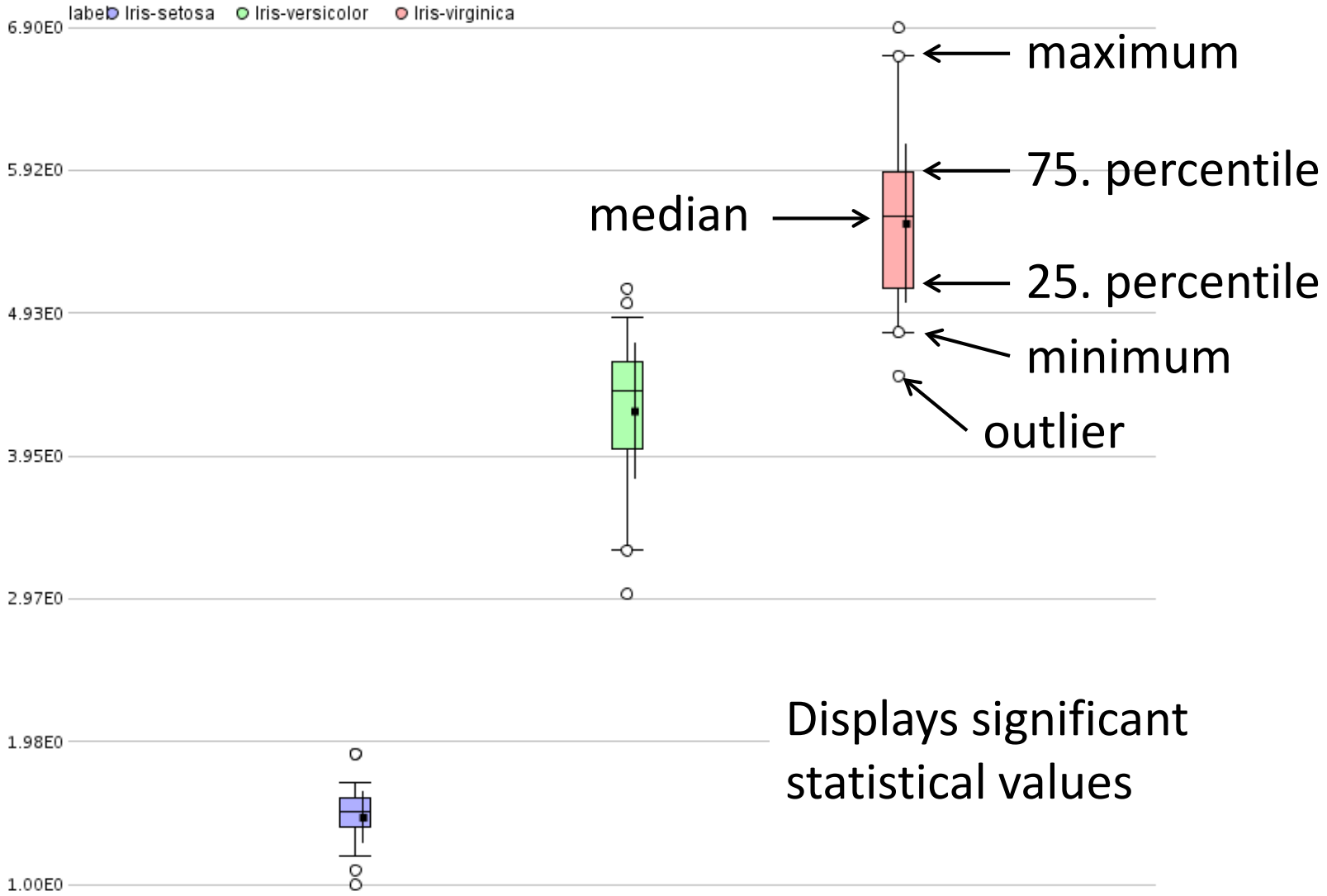
- Box and whiskers chart



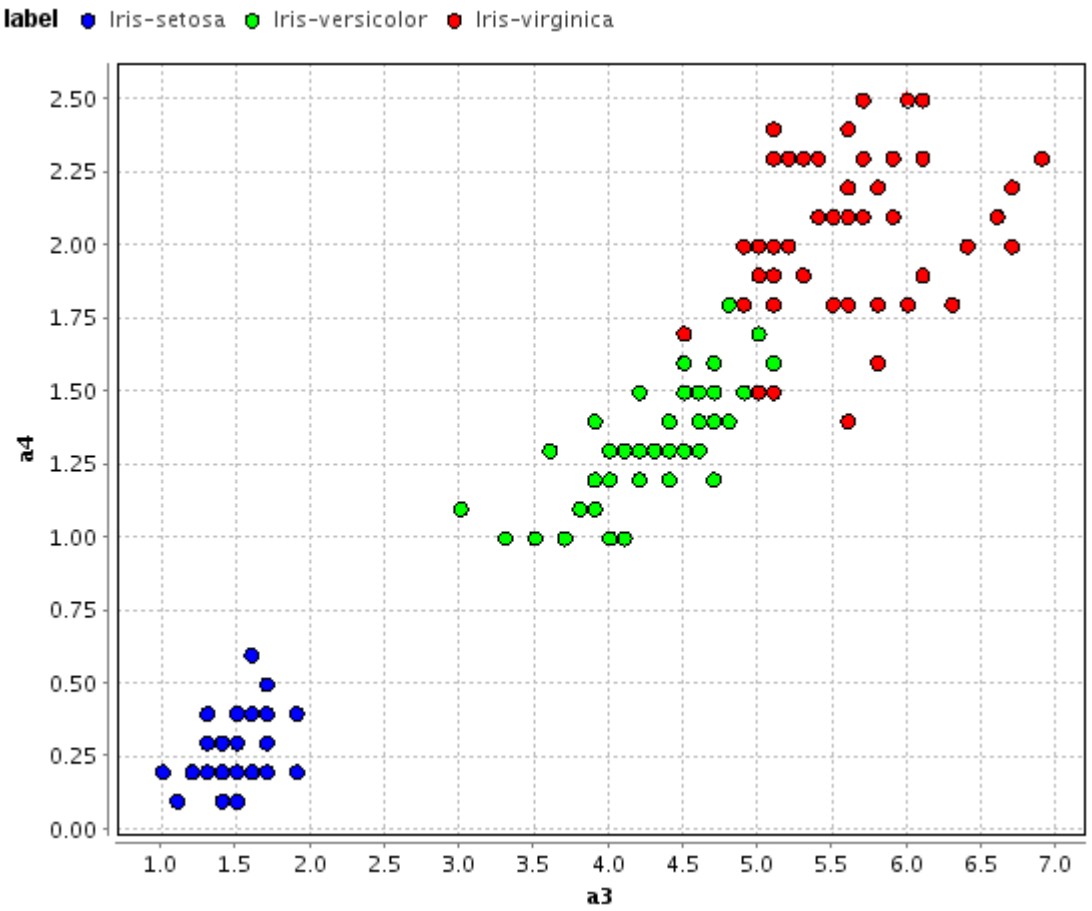




Box-and-whiskers plot

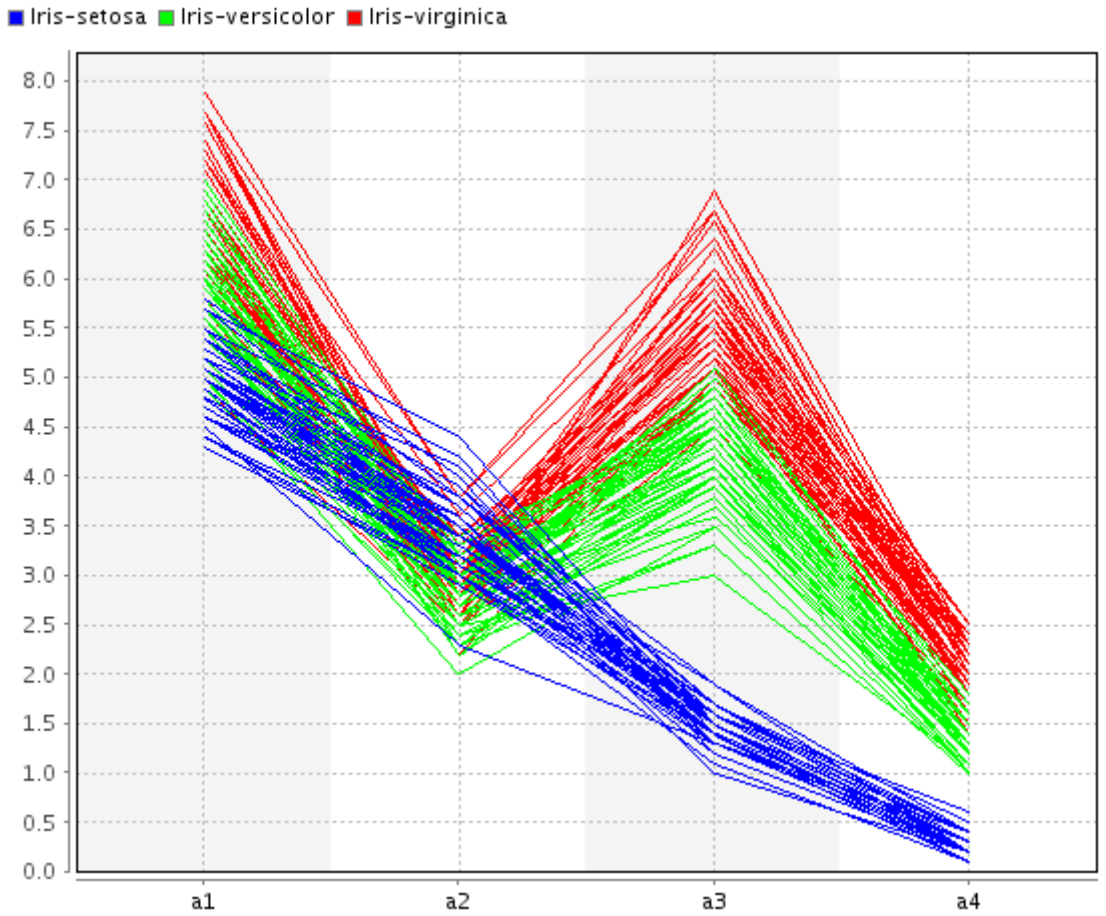


Scatter plot



What if we have more than two dimensions?

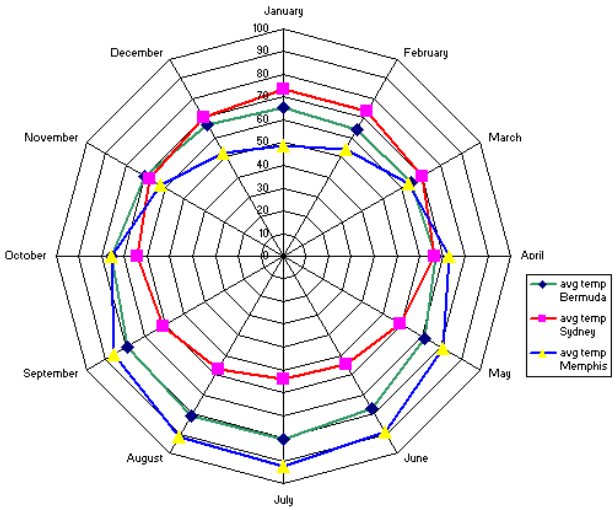
Chart of parallel coordinates

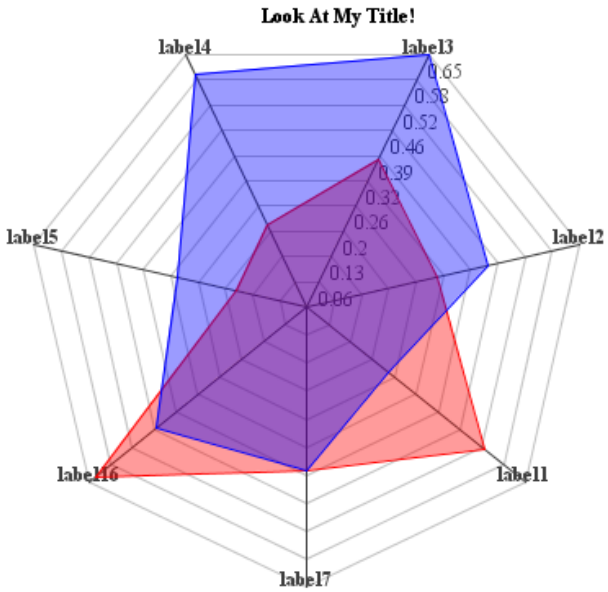


What attribute is the best for classification?

Basic charts for data visualization

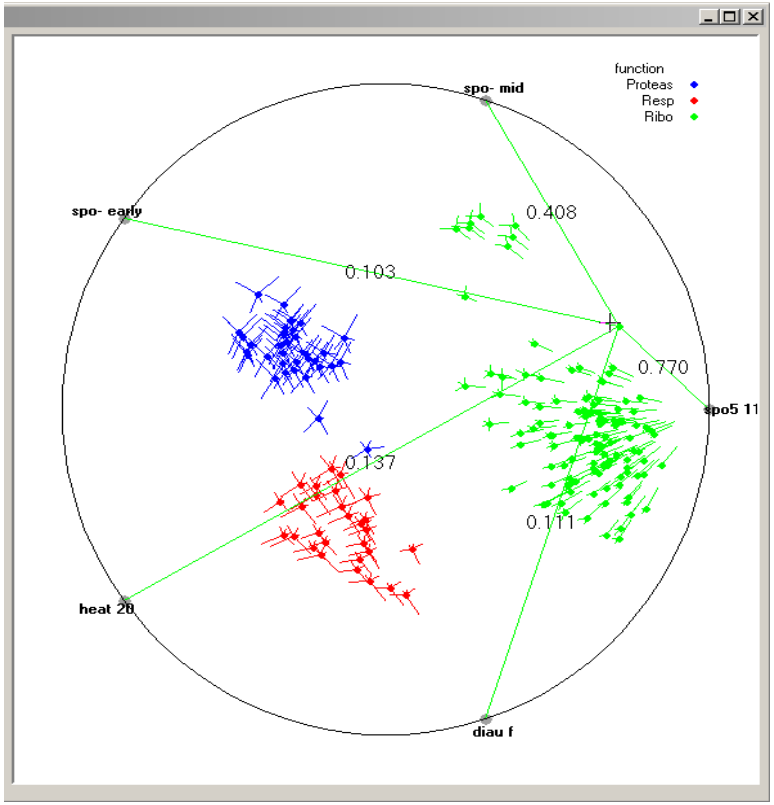
- Radar chart





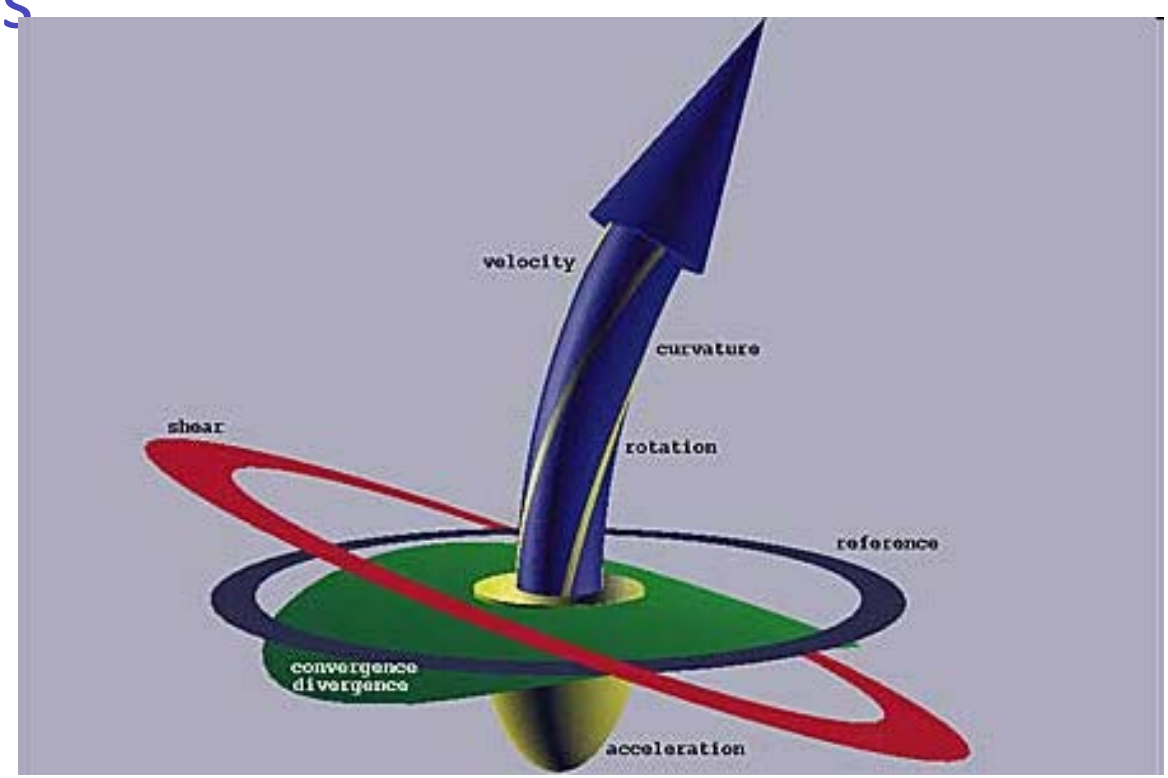
Visualization of high-dimensional data

- RadViz - Radial Coordinate Visualization
  - Value of feature = stiffness of spring
  - balanced spring system



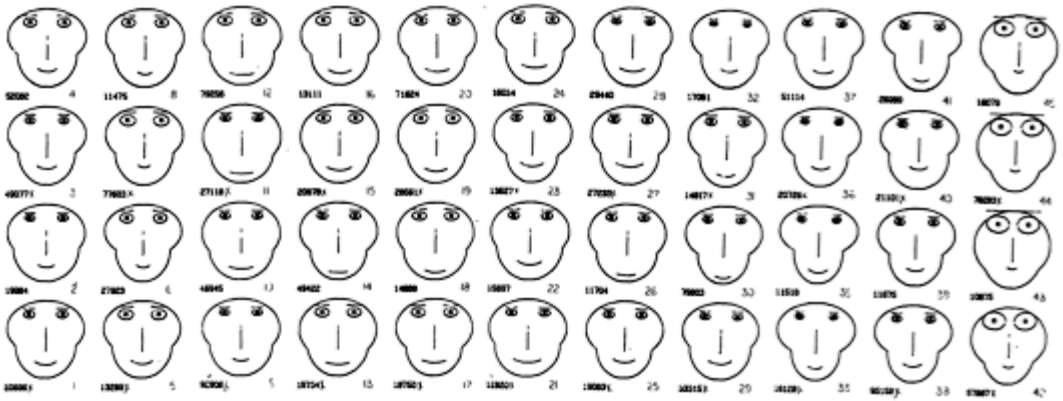
Visualization of high-dimensional data

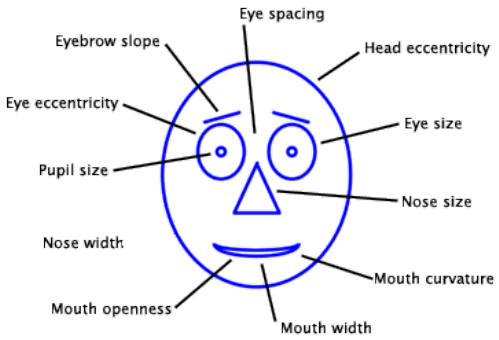
- Glyphs



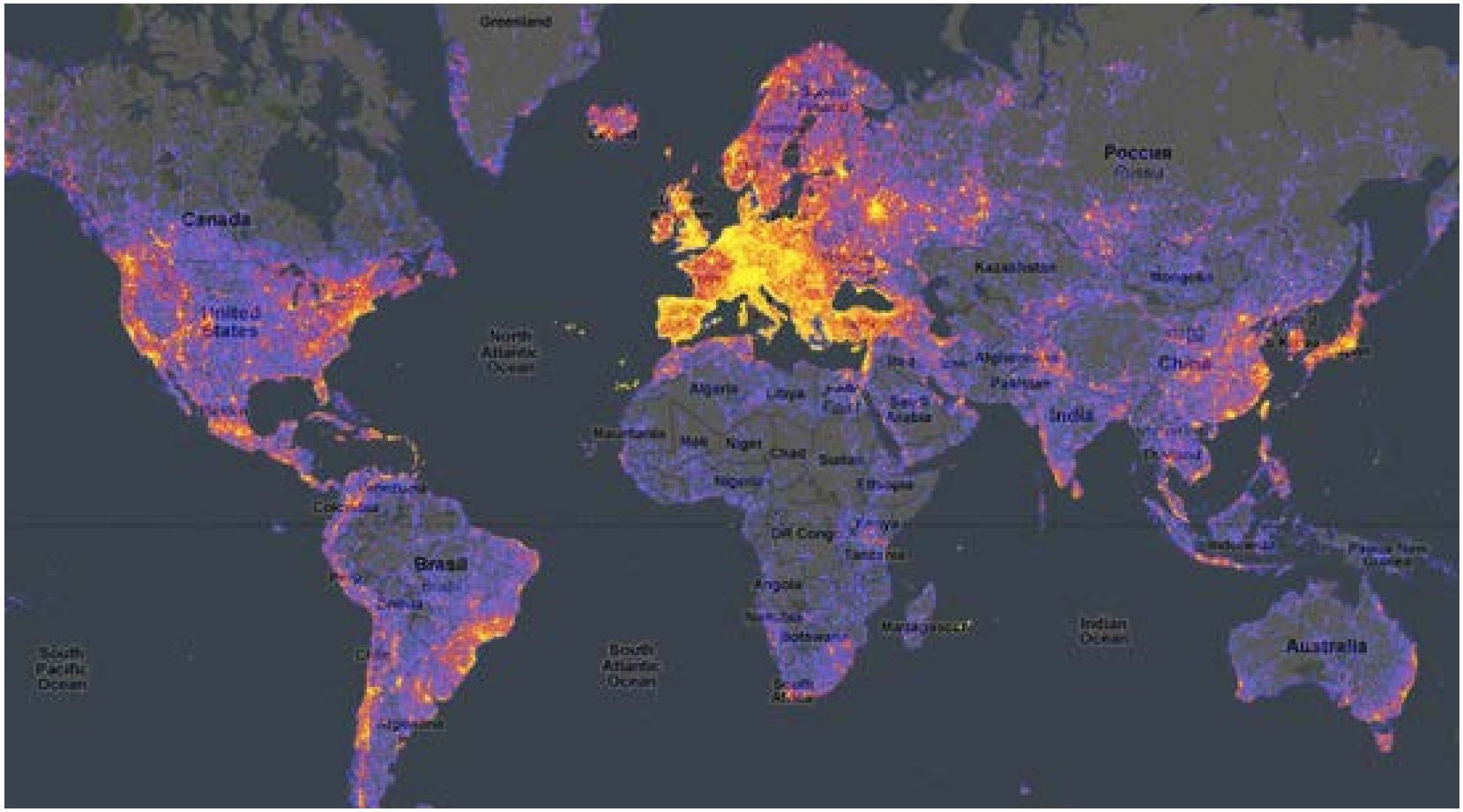
Visualization of high-dimensional data

- Glyphs





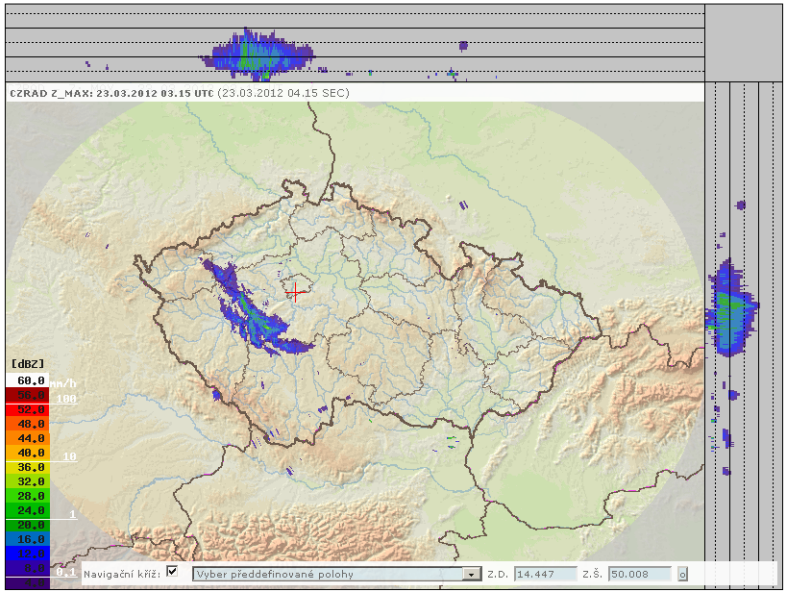
Heat map – tourist destinations

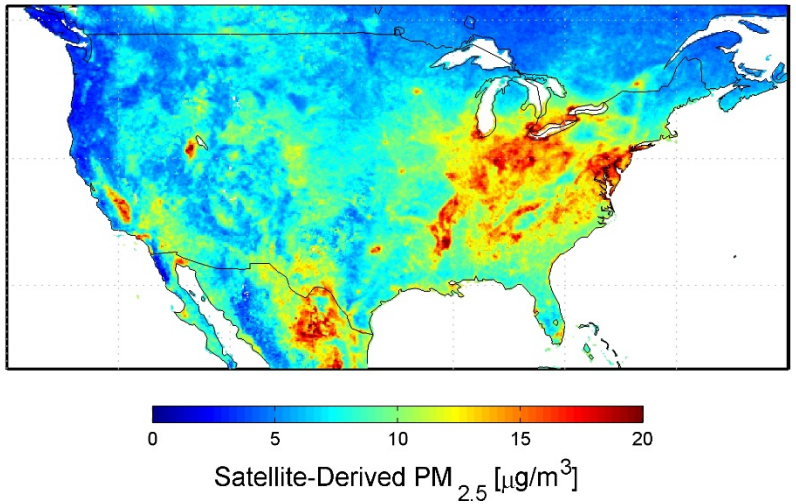


http://www.informationisbeautiful.net

Charts for data visualization

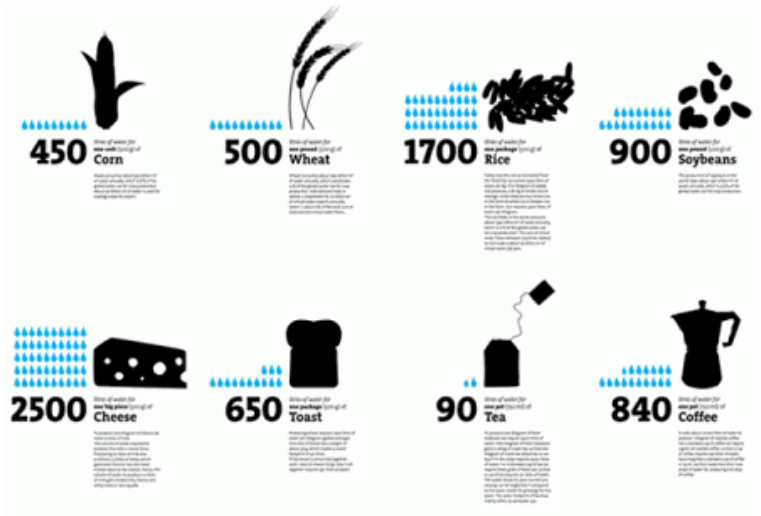
- Map Visualization

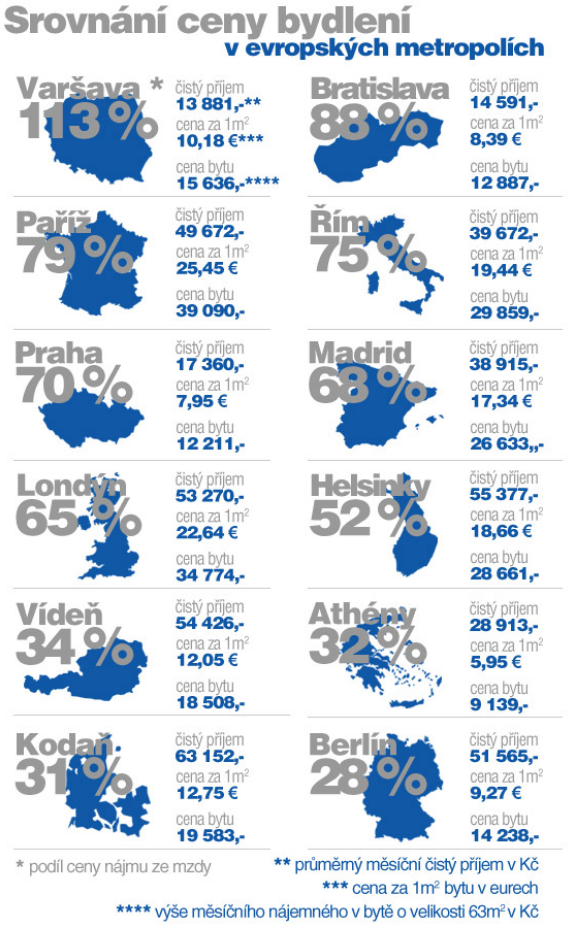




Charts for data visualization

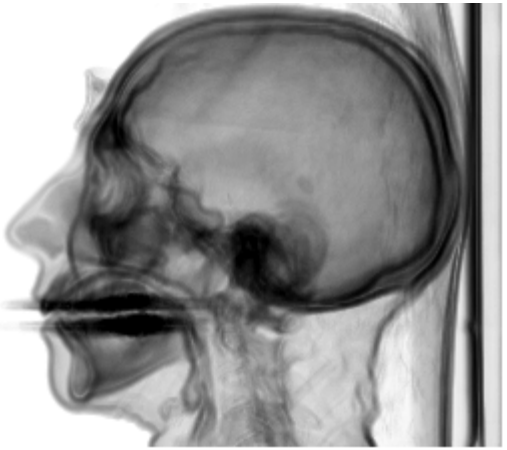
- Infographics

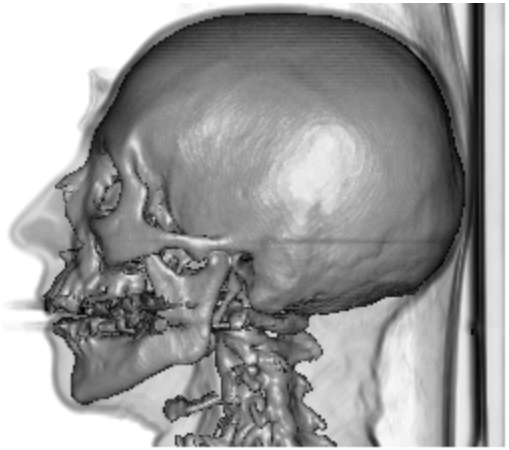




Visualization of volume data

- Photorealistic visualization
  - For clearness, not for perfect realistic visualization





Charts for data visualization

- Google charts

Distribution of variables

- Nominal
  - Text (identifiers, alphabetical values)
  - Ordinal
  - Binomial
- Numerical
  - Real
  - Integer
  - Binary
  - Polynomial

Nominal variables

- Nominal data (from the Latin *nomen,* name) items are items separated by their names.

Country	Assigned number
Austria	1
Ireland	2
Croatia	3

- Nominal elements may have assigned numbers, but it does not mean that Ireland is next to Croatia. Numbers only facilitate storage and processing.
- Some things therefore does not make sense to do for nominal data. For example, measure a diameter.

Binomial variables

- Nominal attribute, which takes only two nominal values
- For example, a coin toss:

Toss	Result	Assigned number
1	Back	1
2	Reverse	0
3	Reverse	0
4	Back	1

Ordinal

- The attribute can be *ordered* according to something, but the difference between values does not play any role
- The order is often expressed by a number or other sequence of symbols

Size	Assigned value
Small	0,1
Medium	0,5
Large	0,93

- The ordinal numbers can not perform arithmetic operations - only displays order

Interval

- At intervals the difference between the two values can be measured
- For example, the difference between 100 °C and 90 °C is the same as between 90 °C and 80 °C

Ratio

- The ratio has all the features of the interval, but in addition it has a clear definition of *zero*. Due to this feature, 4 kg is twice more than 2 kg, as well as 6 kg is twice more than 3 kg.
- Variables such as weight, length or temperature in Kelvins are relative quantities. But the temperature in °C has not this feature, because 0 °C does not mean temperature absence.

Comparison of variables

We can calculate	Nominal	Ordinal	Interval	Ratio
Distribution frequencies	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Addition and subtraction	No	No	Yes	Yes
Average, standard deviation	No	No	Yes	Yes
Ratio	No	No	No	Yes

Distribution may not always be clear. For example, the color. According to psychologists, it is a nominal variable. But according to physicists it is a ratio, because the color can be described a wavelength.

Numeric variables

- They can be continuous or discrete

Weight	# Coins	Length	# Eggs
12,32	1 250	120,6	12
18,00	1 360	10,2	2
6,50	800	13,9	4

Data matrix

- Attributes (features, variables, predictors) - columns
- Instances (cases) - rows

	x_1	x_2	…	x_n	d_1	d_2	…	d_m
s_1	0,84	0,96	…	0,14	0,99	0,53	…	1
s_2	0,51	0,04	…	0,12	0,78	0,23	…	0
s_3	0,62	0,21	…	0,87	0,25	0,57	…	1
…	…	…	…		…	…	…	…
s_N	0,37	0,83	…	0,17	0,64	0,09	…	1

Data matrix

- Attributes (features, variables, predictors) - columns
- Instances (cases) - rows

Data: Credit applications (6v by 60c)

	1 NNSET	2 HOME OWNER	3 AVG_INC	4 AGE	5 LOAN_VOL	6 STATUS
1	Verify	Yes	$53	40	15	1
2	Verify	Yes	$52	43	5	1
3	Train	No	$41	35	6	2
4	Verify	Yes	$70	27	6	1
5	Train	No	$28	36	3	2
6	Verify	Yes	$48	30	12	2
7	Verify	Yes	$38	41	13	1
8	Train	Yes	$42	35	4	1
9	Train	No	$42	27	7	1
10	Verify	Yes	$61	38	18	1
11	Verify	Yes	$56	38	20	1
12	Train	Yes	$59	41	11	1
13	Train	Yes	$38	36	4	2
14	Train	No	$33	29	11	2
15	Verify	Yes	$51	40	4	1
16	Train	No	$40	23	10	2

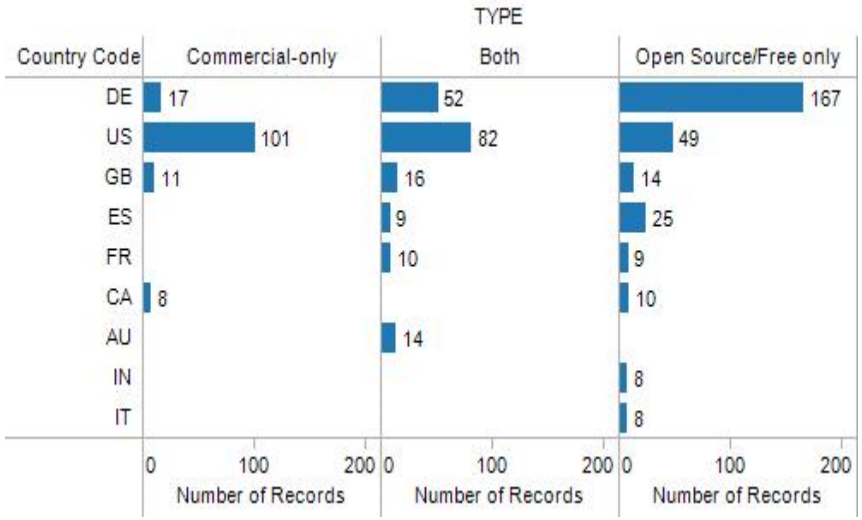
Data matrix

- Attributes (features, variables, predictors) - columns
- Instances (cases) - rows

?,C,A,08,00,?,S,?,000,?,?,G,?,COIL,0.700,0610.0,0000,?,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,Y,?,?,?,COIL,3.200,0610.0,0000,?,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,?,Y,?,B,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,SHEET,0.700,1300.0,0762,?,0000,?,3
?,C,A,00,60,T,?,?,000,?,?,G,?,?,?,?,M,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,COIL,2.801,0385.1,0000,?,0000,?,3
?,C,A,00,60,T,?,?,000,?,?,G,?,?,?,?,B,Y,?,?,?,Y,?,?,?,?,?,?,?,?,?,SHEET,0.801,0255.0,0269,?,0000,?,3
?,C,A,00,45,?,S,?,000,?,?,D,?,COIL,1.600,0610.0,0000,?,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,?,?,?,?,Y,?,?,?,?,?,?,?,?,?,?,?,?,?,?,SHEET,0.699,0610.0,4880,Y,0000,?,3
?,C,A,00,00,?,S,2,000,?,?,E,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,Y,?,?,?,COIL,3.300,0152.0,0000,?,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,Y,?,?,?,COIL,0.699,1320.0,0000,?,0000,?,3
?,C,A,00,00,?,S,3,000,N,?,E,?,SHEET,1.000,1320.0,0762,?,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,COIL,1.200,0610.0,0000,?,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,?,?,?,?,Y,?,?,?,?,?,?,?,?,?,?,?,?,?,?,SHEET,0.300,1320.0,4880,Y,0000,?,3
?,C,R,00,00,?,S,2,000,?,?,E,?,?,?,?,B,Y,?,?,?,Y,?,?,?,?,?,?,?,?,?,SHEET,1.200,0610.0,0150,?,0000,?,3
?,C,A,00,45,?,S,?,000,?,?,D,?,COIL,1.200,0609.9,0000,?,0000,?,3
?,C,A,00,00,?,S,2,000,?,?,F,?,?,Y,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,SHEET,0.600,1220.0,0761,?,0000,?,3
?,C,A,00,00,?,S,2,000,?,?,E,?,SHEET,4.000,1320.0,0762,?,0000,?,3
?,C,A,10,00,?,?,?,000,?,?,E,?,COIL,3.201,0600.0,0000,?,0000,?,U
?,C,A,00,80,T,?,?,000,?,?,G,?,SHEET,0.800,0610.0,4170,Y,0000,?,U
…

Ranking of DM tools

RapidMiner (345)	37.8%
R (272)	29.8%
Excel (222)	24.3%
KNIME (175)	19.2%
Your own code (168)	18.4%
Pentaho/Weka (131)	14.3%
SAS (110)	12.0%
MATLAB (84)	9.2%
IBM SPSS Statistics (72)	7.9%
Other free tools (67)	
IBM SPSS Modeler (former Clementine) (67)	
Microsoft SQL Server (63)	
Statsoft Statistica (57)	
Other commercial tools (56)	
SAS Enterprise Miner (50)	
Zementis (34)	
Orange (25)	



RapidMiner

- First open-source for datamining (1996)
- Over 100 algorithms
- Entangling architecture
- Data must typically fit into the entire memory

- Also open-source
- Can the same as Weka, plus a little extra
- Can process data "on the fly"
- Also written in Java, but it is more stable

https://rapidminer.com/