Adatfeldolgozás és –elemzés (Data processing and analysis)

Dr. Gönczy László

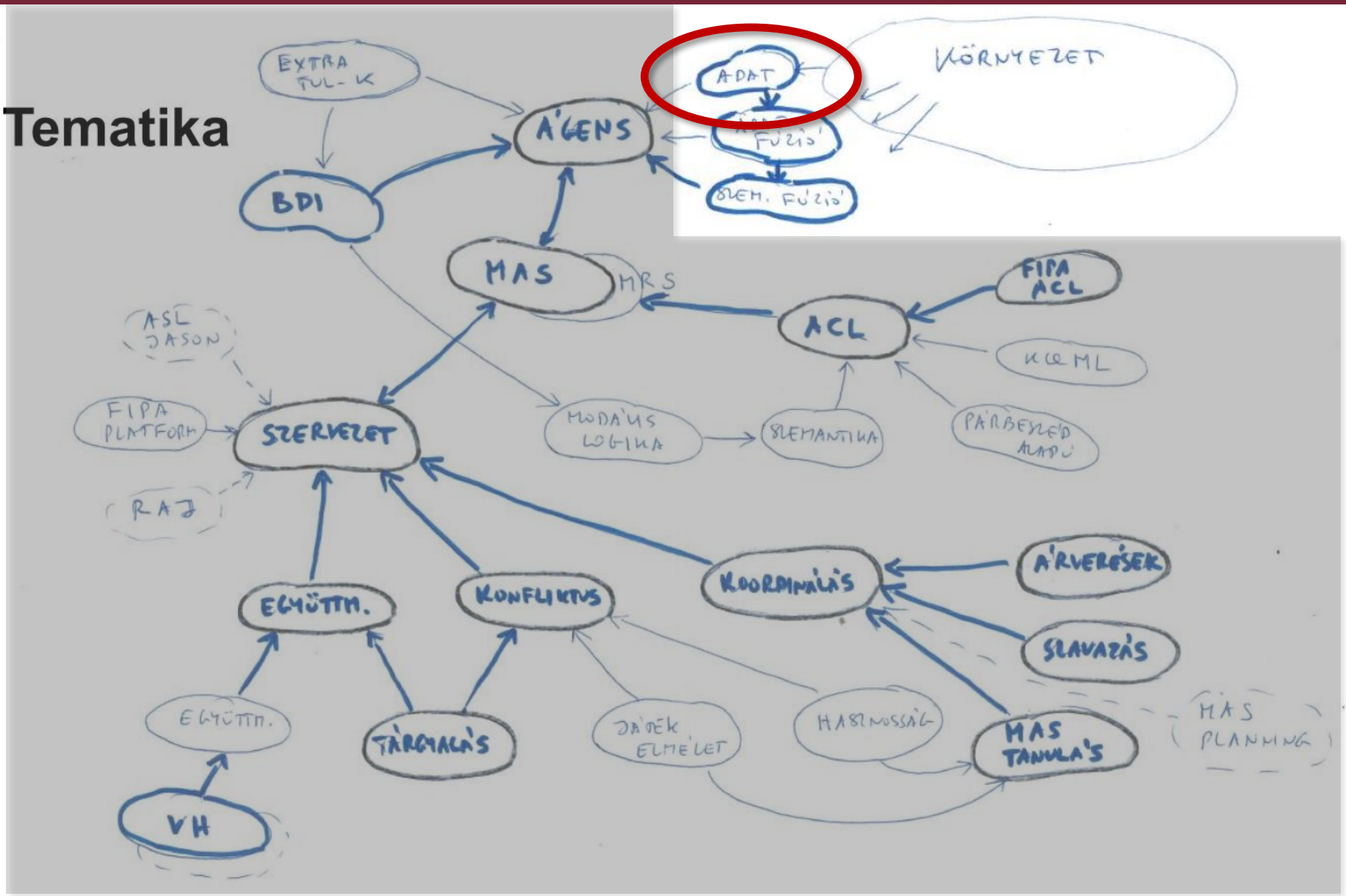
Intelligens Elosztott Rendszerek

http://www.mit.bme.hu/oktatas/targyak/vimiac02

Budapest University of Technology and Economics
Department of Measurement and Information Systems

ftsrg

Tematika


EXTRA TUL-K
ÁGENS
ADAT
KÖRNYEZET
FÚZIÓ
SZEM. FÚZIÓ
BDI
MAS
MRS
FIPA ACL
ACL
KQML
ASL JASON
FIPA PLATFORM
SZERVEZET
MODÁLIS LOGIKA
SZEMANTIKA
PÁRBESZÉD ALAPÚ
RAJ
KOORDINÁLÁS
ÁRVERÉSEK
SZAVAZÁS
ELKÖTTH.
KONFLIKTUS
ELKÖTTH.
TÁRGYALÁS
JÁTÉK ELMÉLET
HASZNOSSÁG
MAS TANULÁS
MAS PLANNING
VH

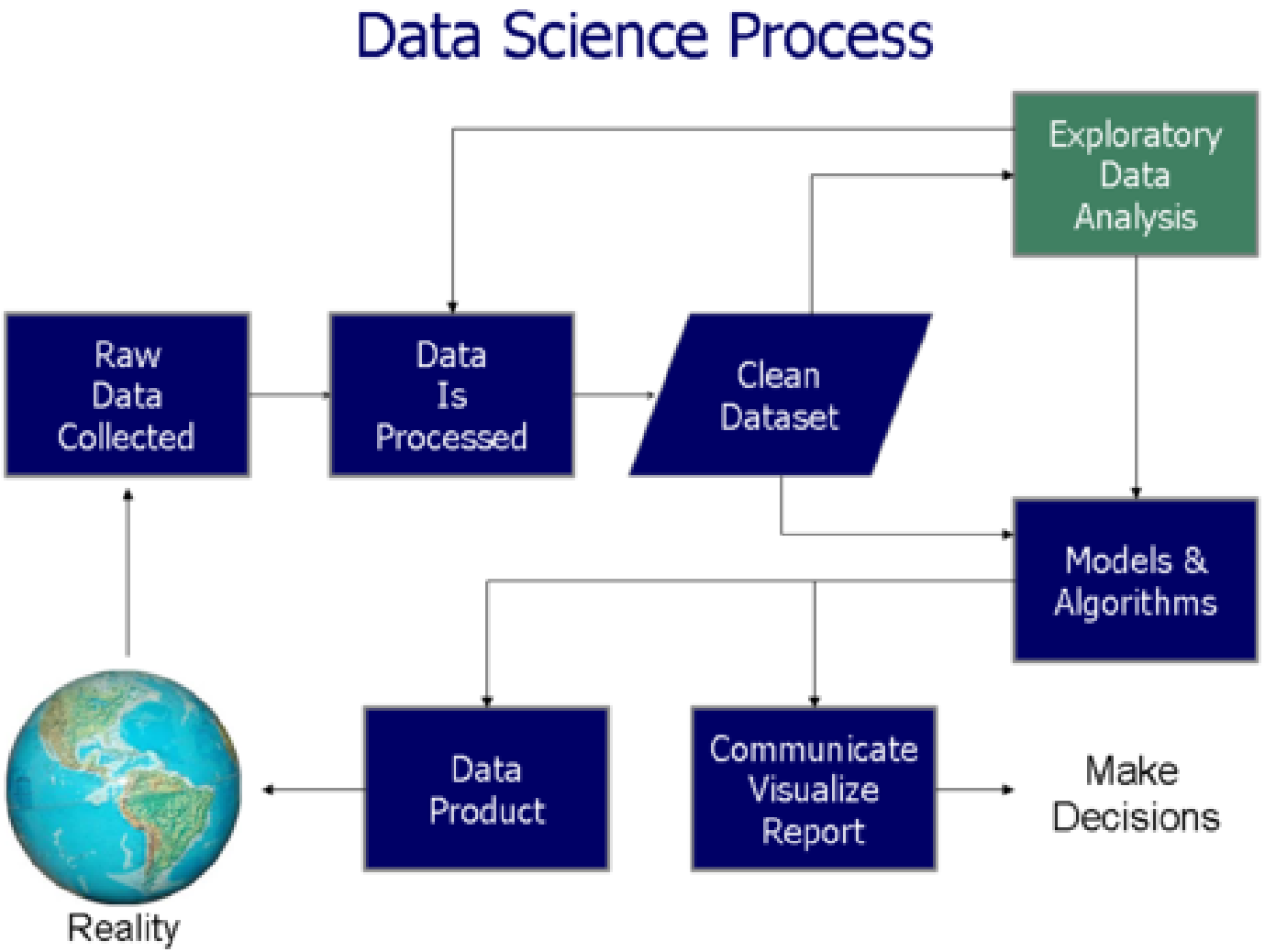
Outline

- Data collection
- Data processing
- ETL, workflow support
- Data format/representation
- Data storage

→Data analysis

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data. - @BigDataBorat Twitter

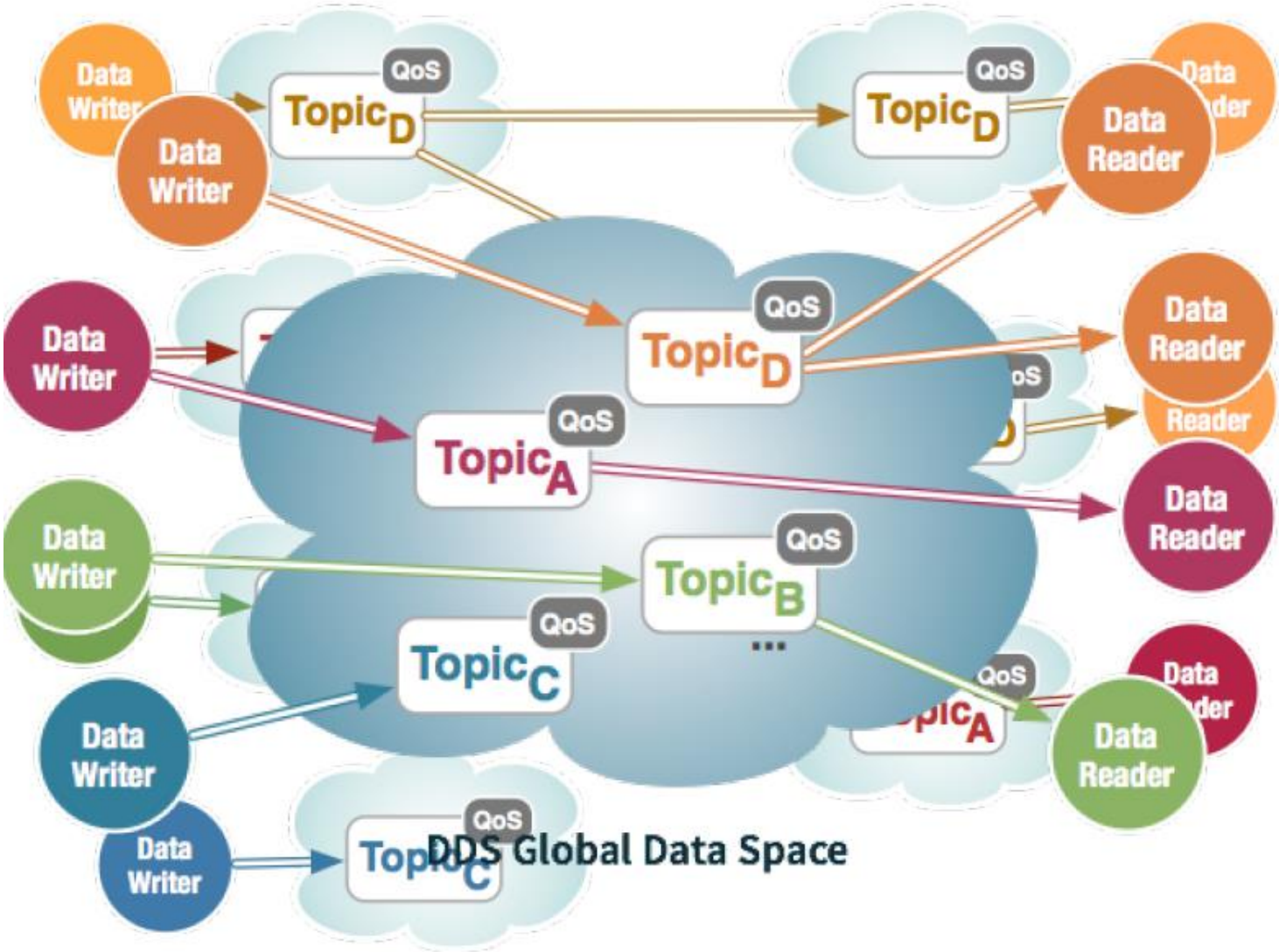
Data science „process”



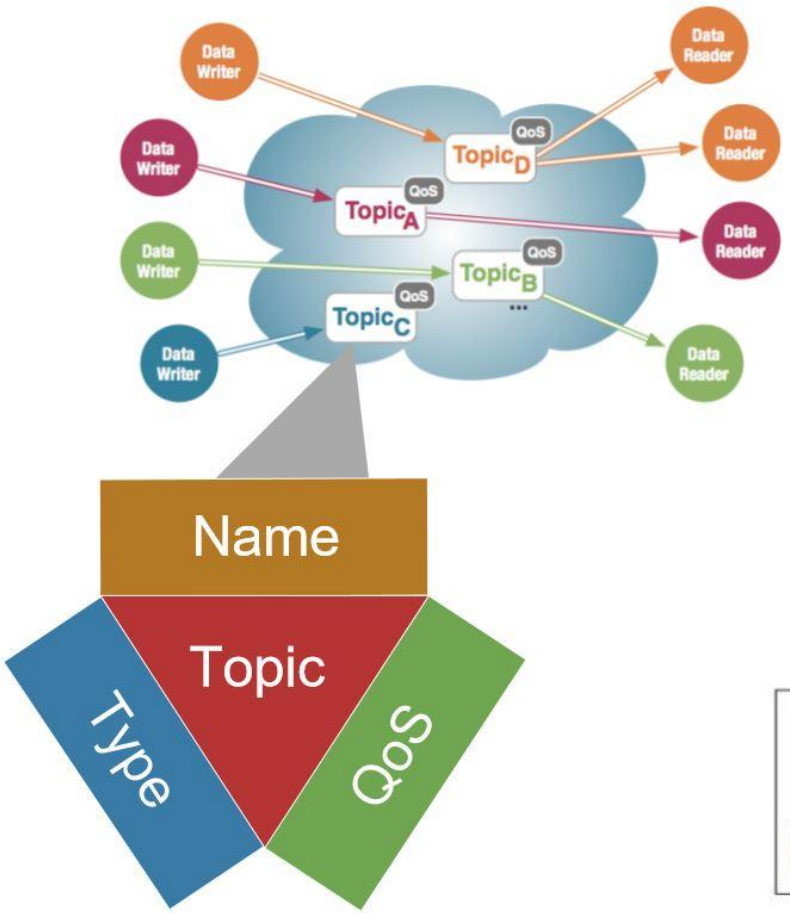
https://en.wikipedia.org/wiki/Data_science

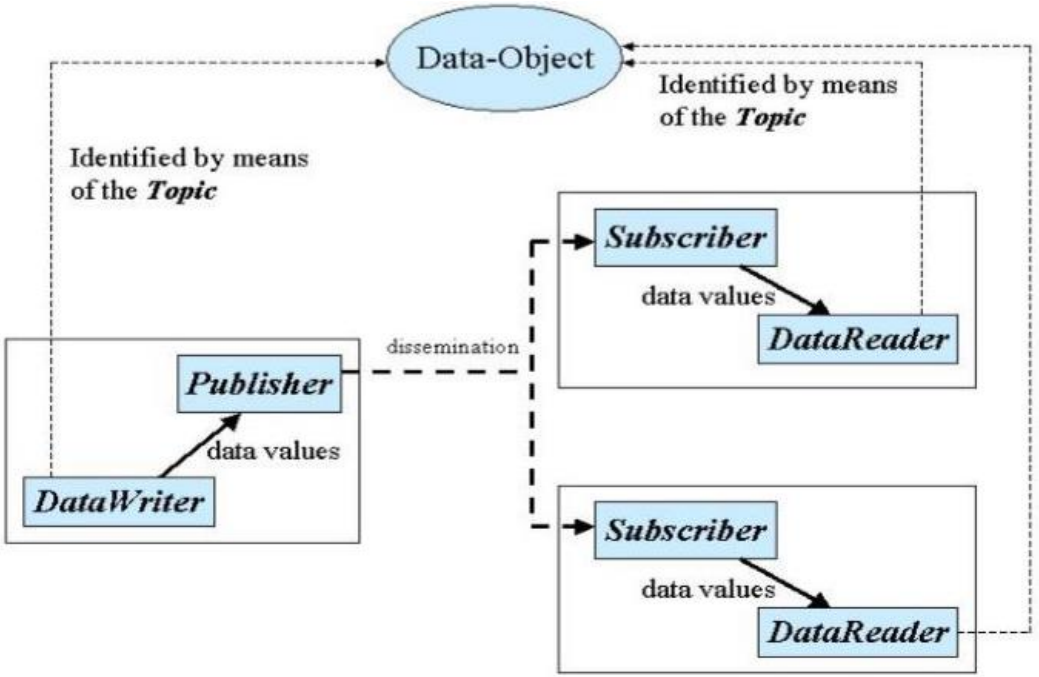
HOW TO GET THE DATA?

Decentralised Data-Space

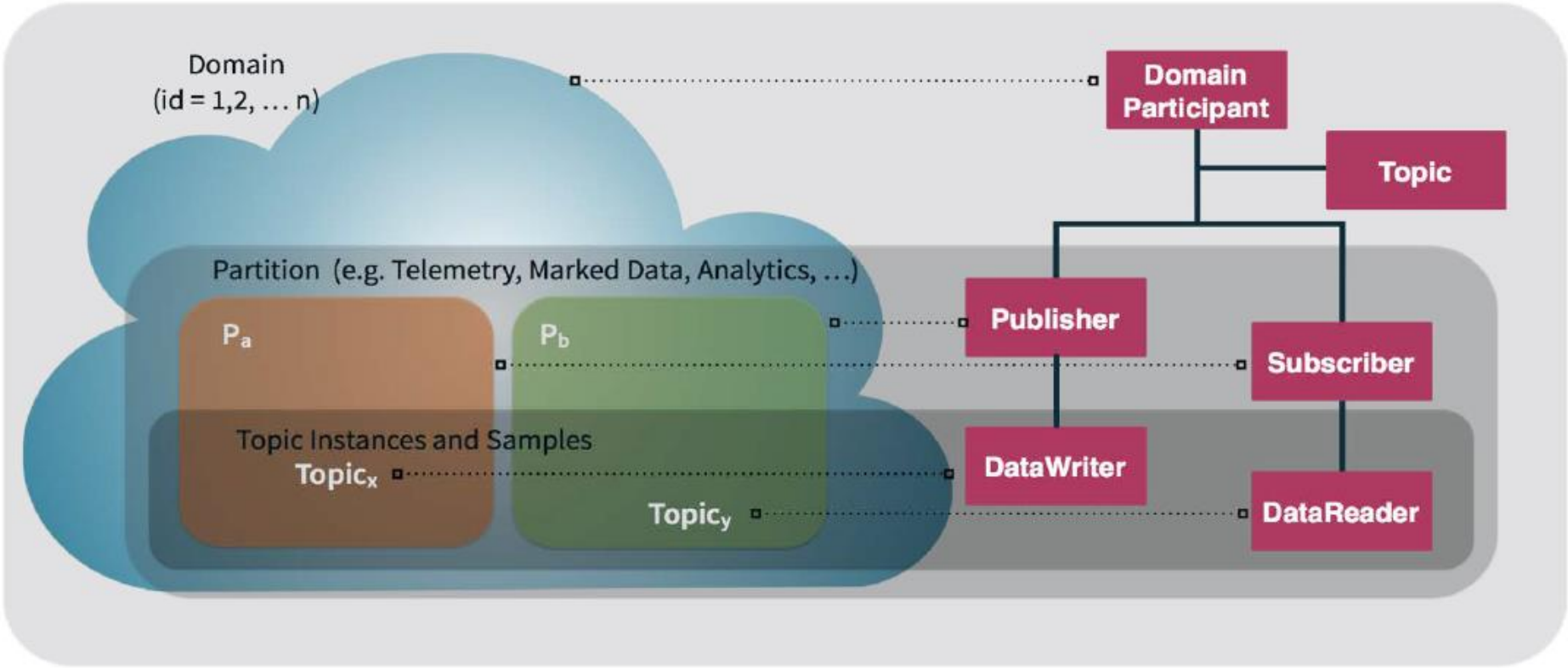


OMG DDS Core notions





Anatomy of a DDS Application



Technological aspects

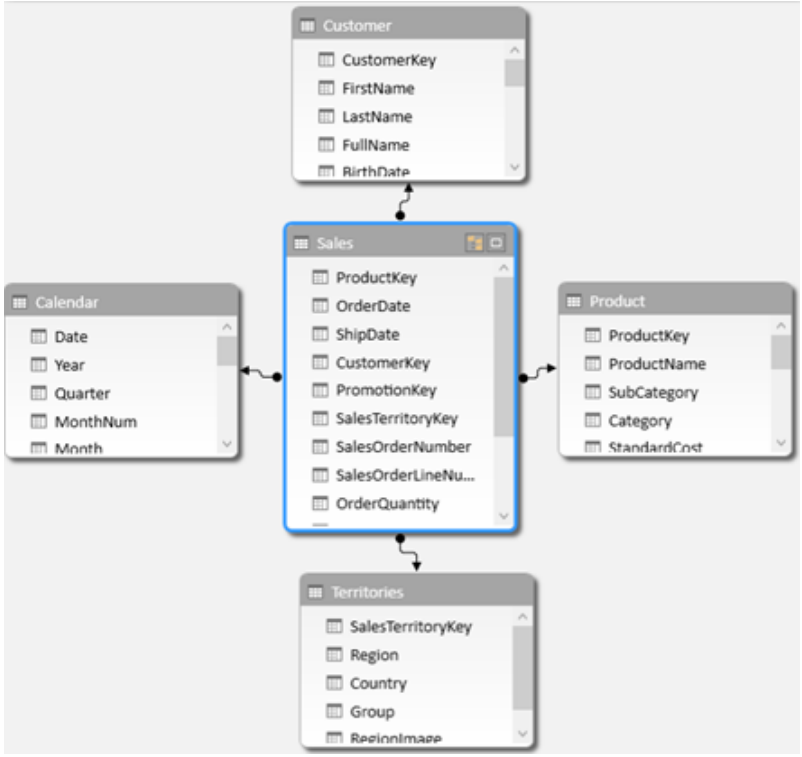
	Transport	Paradigm	Scope	Discovery	Content Awareness	Data Centricity	Security	Data Prioritisation	Fault Tolerance
AMQP	TCP/IP	Point-to-Point Message Exchange	D2D D2C C2C	No	None	Encoding	TLS	None	Impl. Specific
CoAP	UDP/IP	Request/ Reply (REST)	D2D	Yes	None	Encoding	DTLS	None	Decentralised
DDS	UDP/IP (unicast + mcast) TCP/IP	Publish/ Subscribe Request/ Reply	D2D D2C C2C	Yes	Content-Based Routing, Queries	Encoding, Declaration	TLS, DTLS, DDS Security	Tranport Priorities	Decentralised
MQTT	TCP/IP	Publish/ Subscribe	D2C	No	None	Undefined	TLS	None	Broker is the SPoF

[Ref: A Comparative Study of Data-Sharing Standards for the Internet of Things, Cutter Journal, Dec 2014

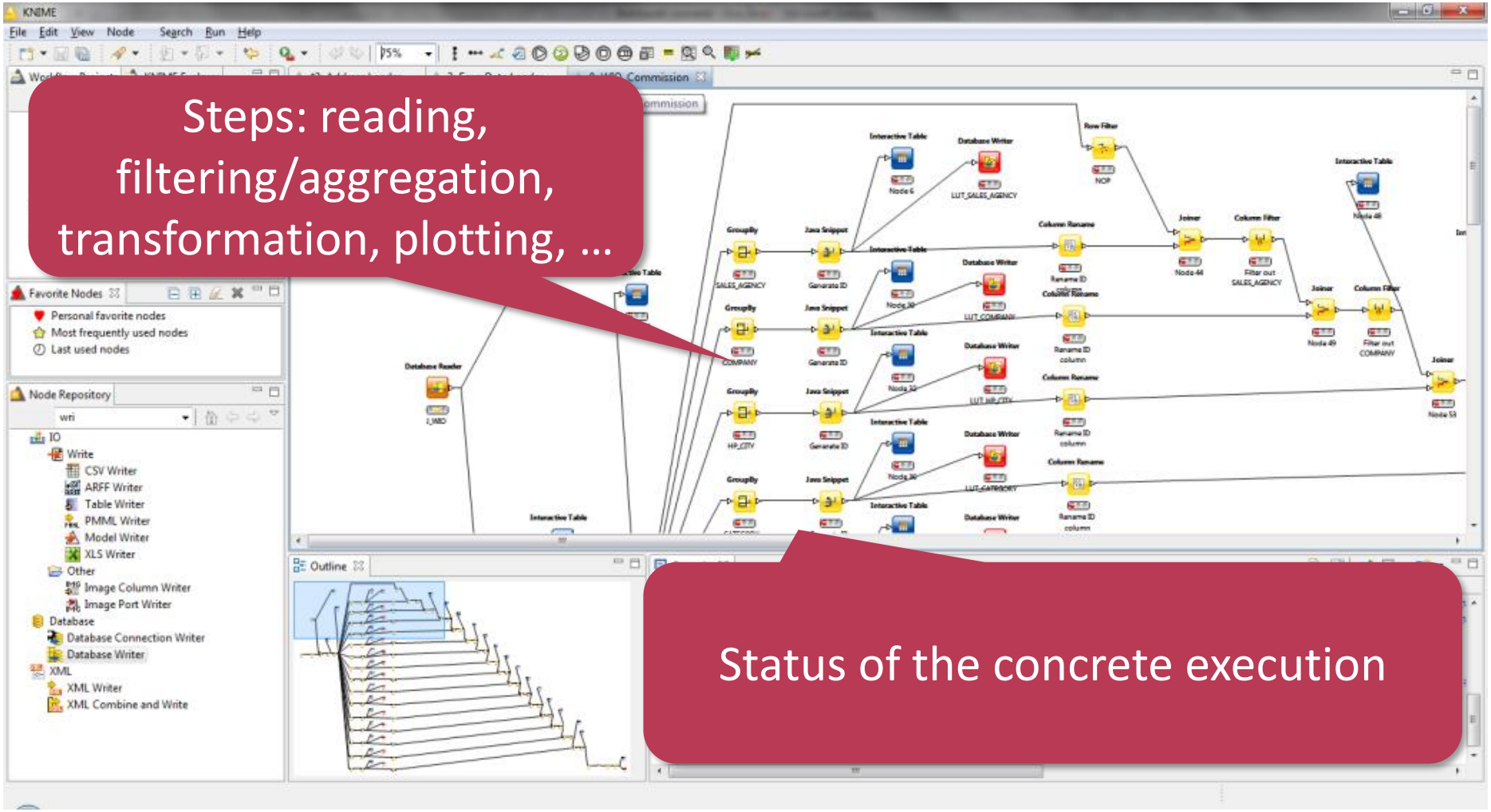
DATA PROCESING WORKFLOW & TOOLS

ETL

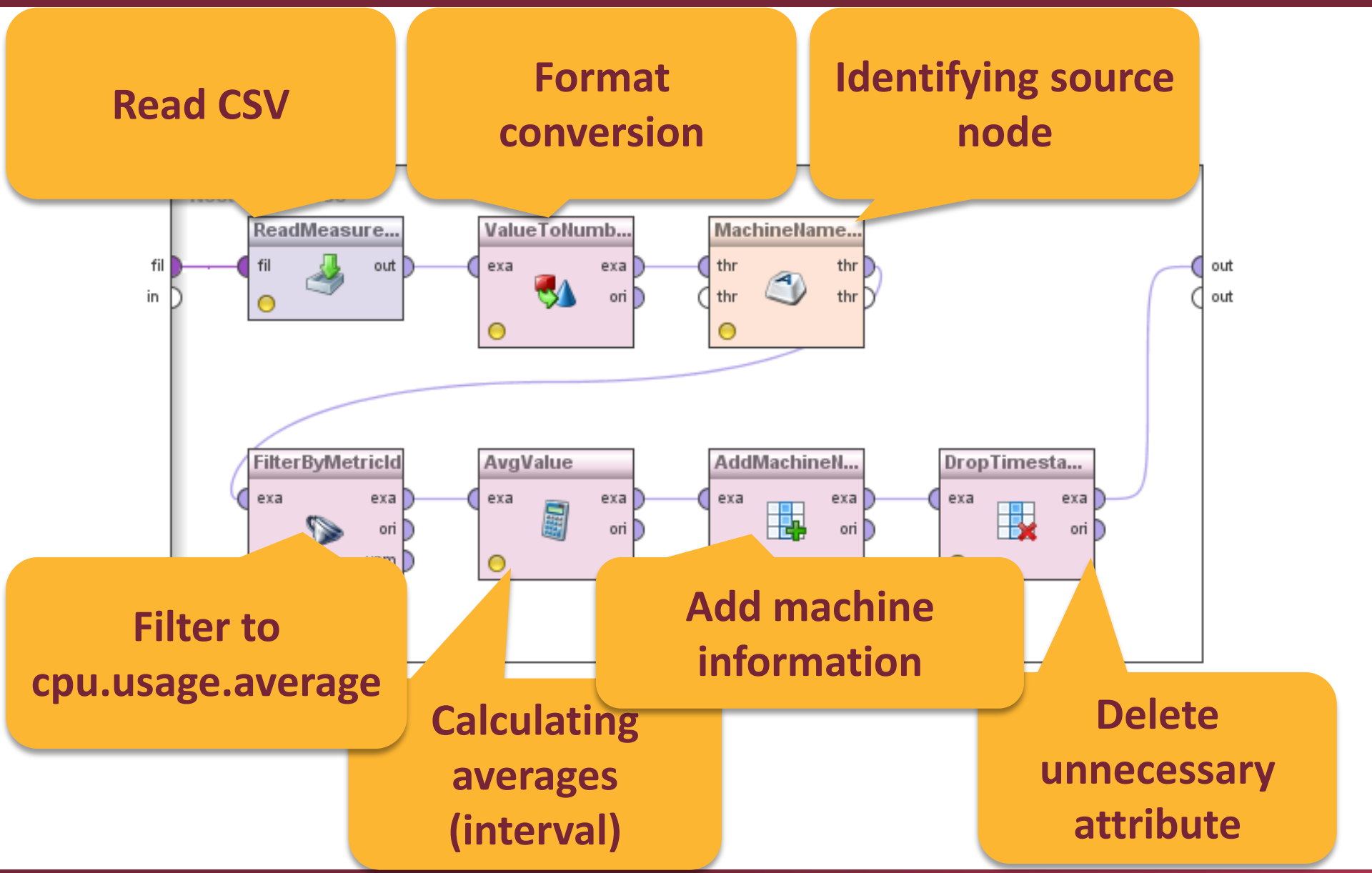
- „Extract-Transform-Load”
- Originally: to fill a snowflake/star schema
- In data science: create dataframes
- Cleaning tasks
  - Standardization
  - Normalization
  - Deduplication
  - Enrichment
  - Clear/fill NAs



Example data processing workflow (KNIME)



Measurement processing: RapidMiner



DATA FORMAT

Tidy data

- 3 Simple rules to facilitate statistics and visualization
- One variable – one column
- One observation – one row
- Each type of observational unit – one table
- … seems to be trivial
- … not true in most practical cases
- … and even for statistical tools (e.g. output of R packages)

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23. https://github.com/hadley/tidy-data

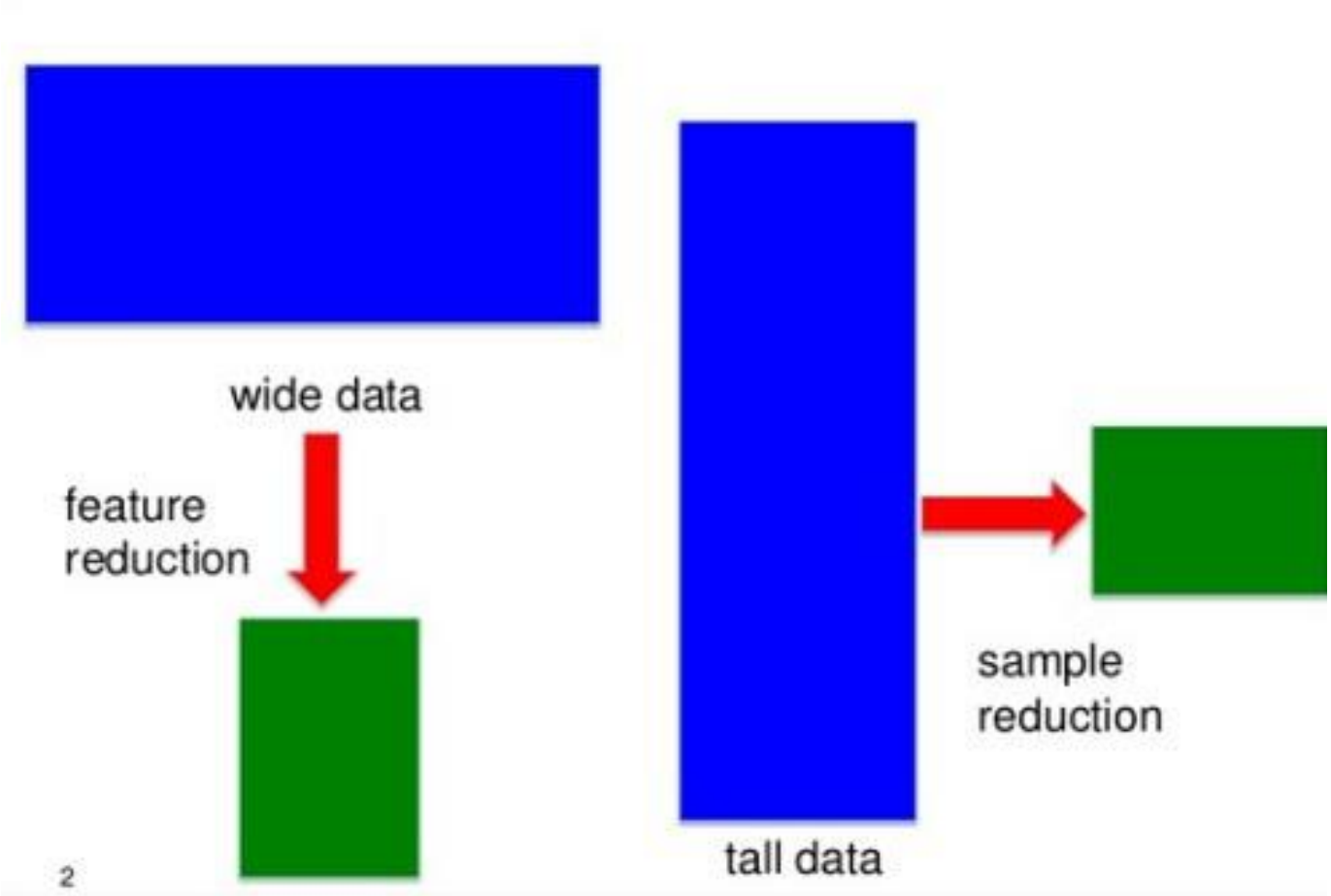
Data originally: long/wide

Person	Age	Weight
Bob	32	128
Alice	24	86
Steve	64	95

Person	Variable	Value
Bob	Age	32
Bob	Weight	128
Alice	Age	24
Alice	Weight	86
Steve	Age	64
Steve	Weight	95

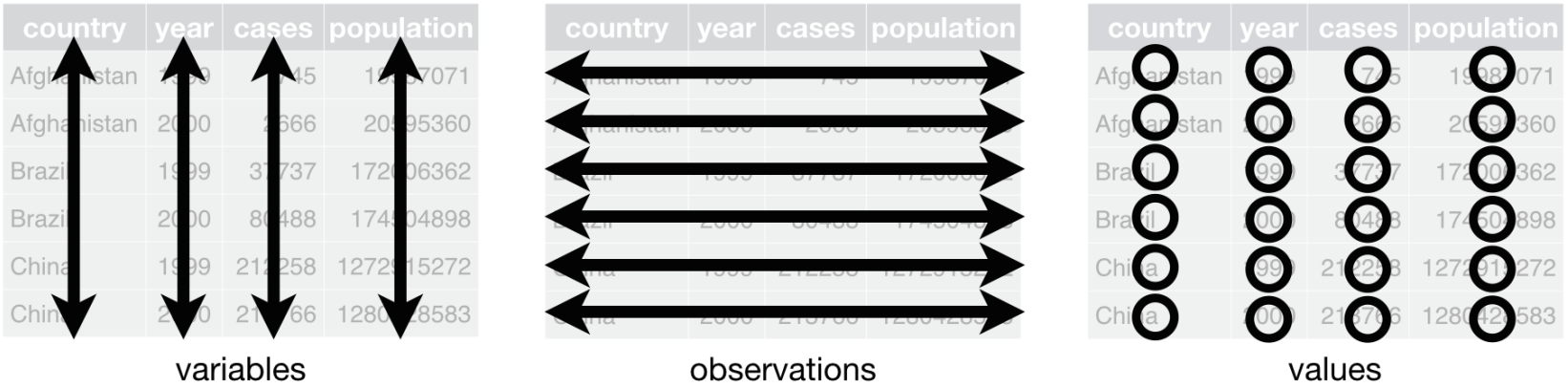
https://en.wikipedia.org/wiki/Wide_and_narrow_data

How to use these formats?

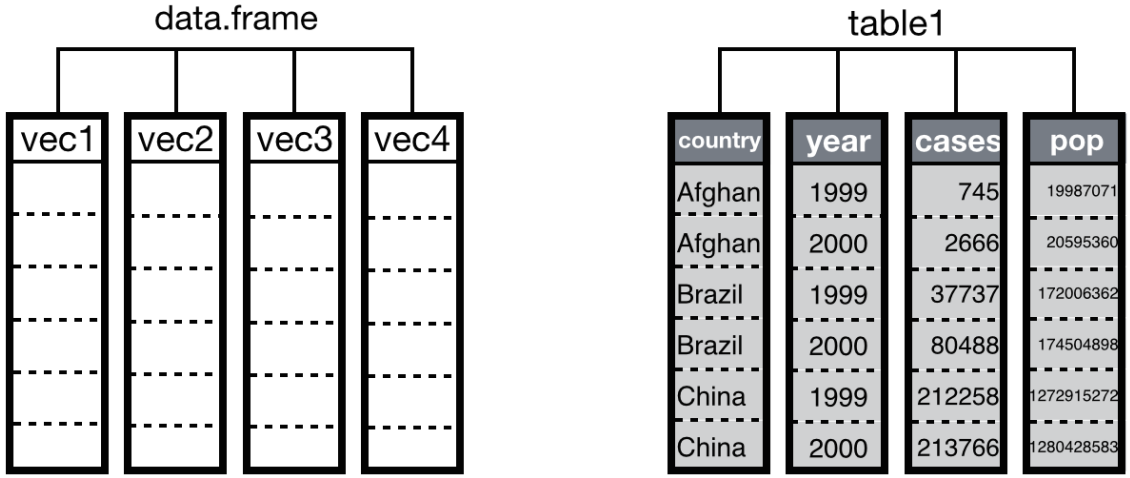


Sparse Screening for Exact Data Reduction. Jieping Ye, Arizona State University

Examples for tidy data



R dataframe representation:



http://garrettgman.github.io/tidying/

„tidying”

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2

R: spread(data,key,value)

http://garrettgman.github.io/tidying/

„tidying”

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

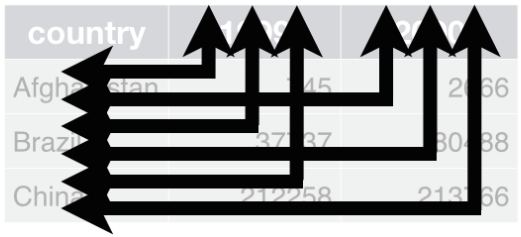
country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4 R: spread(data,key,value)

Generalization?

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

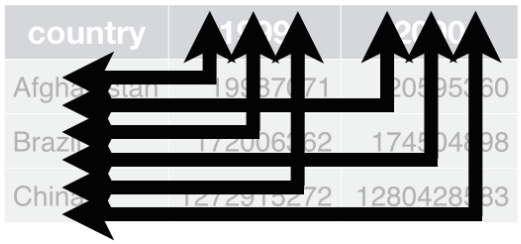
table4



country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

table5

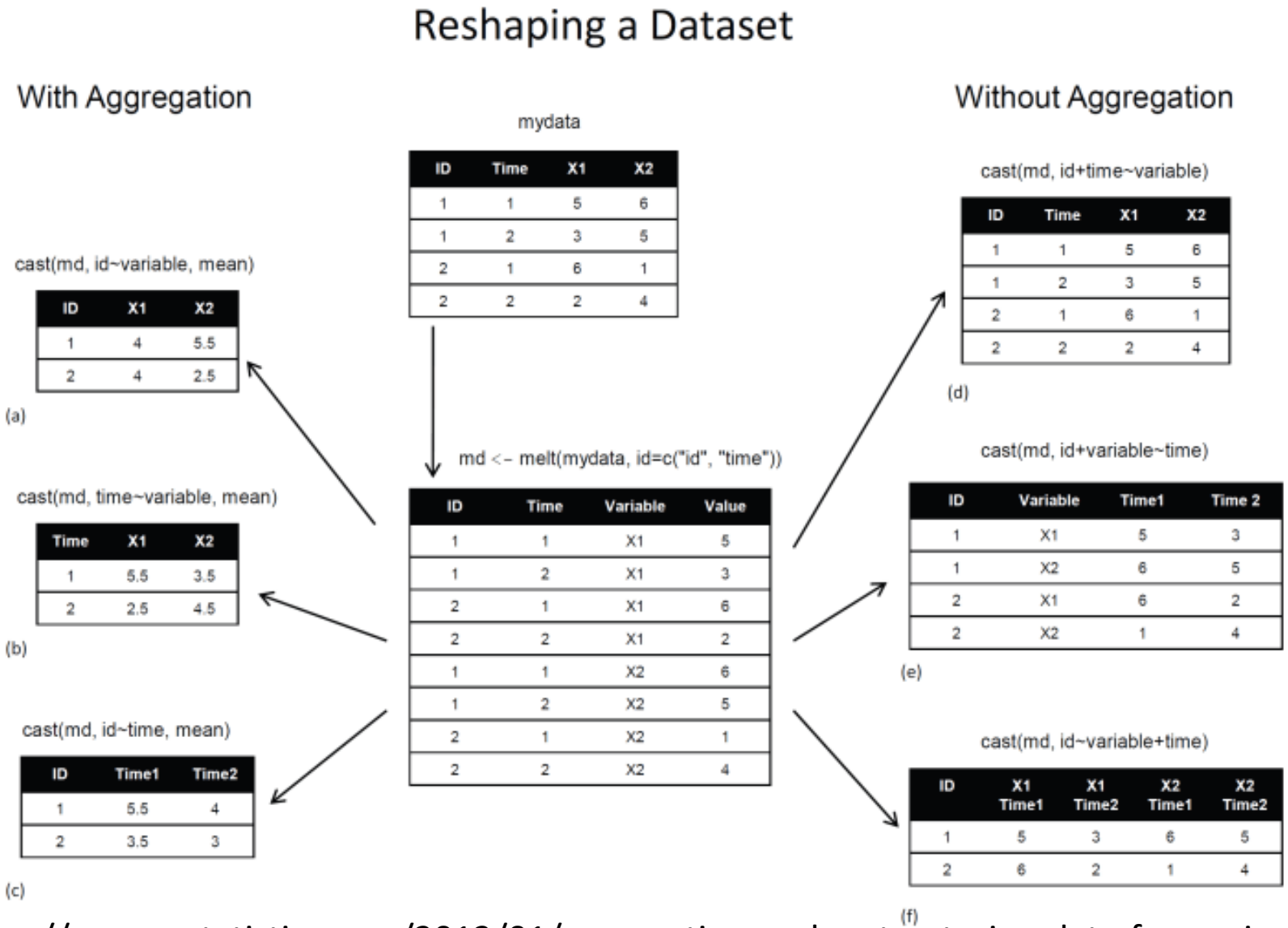
variables



observations

http://garrettgman.github.io/tidying/

Data restructuring examples (in R)



https://www.r-statistics.com/2012/01/aggregation-and-restructuring-data-from-r-in-action/

DATA STORAGE

Reminder: Tabular Representation

- **Rows of the table** = Model elements
- **Columns of the table** = Properties

Name	Type	Size (kB)	Last modified
Documents	directory		2016.02.02
Contracts.pdf	file	569	2015.11.09
Pictures	directory		2016.02.02
Logo.png	file	92	2015.03.06
Groundplot.jpg	file	1226	2016.02.02

- Data analysis languages (e.g. R, Python): **dataframe**
  - One row: one measurement/observation
  - Columns have their own **Types**

Common data storage techniques

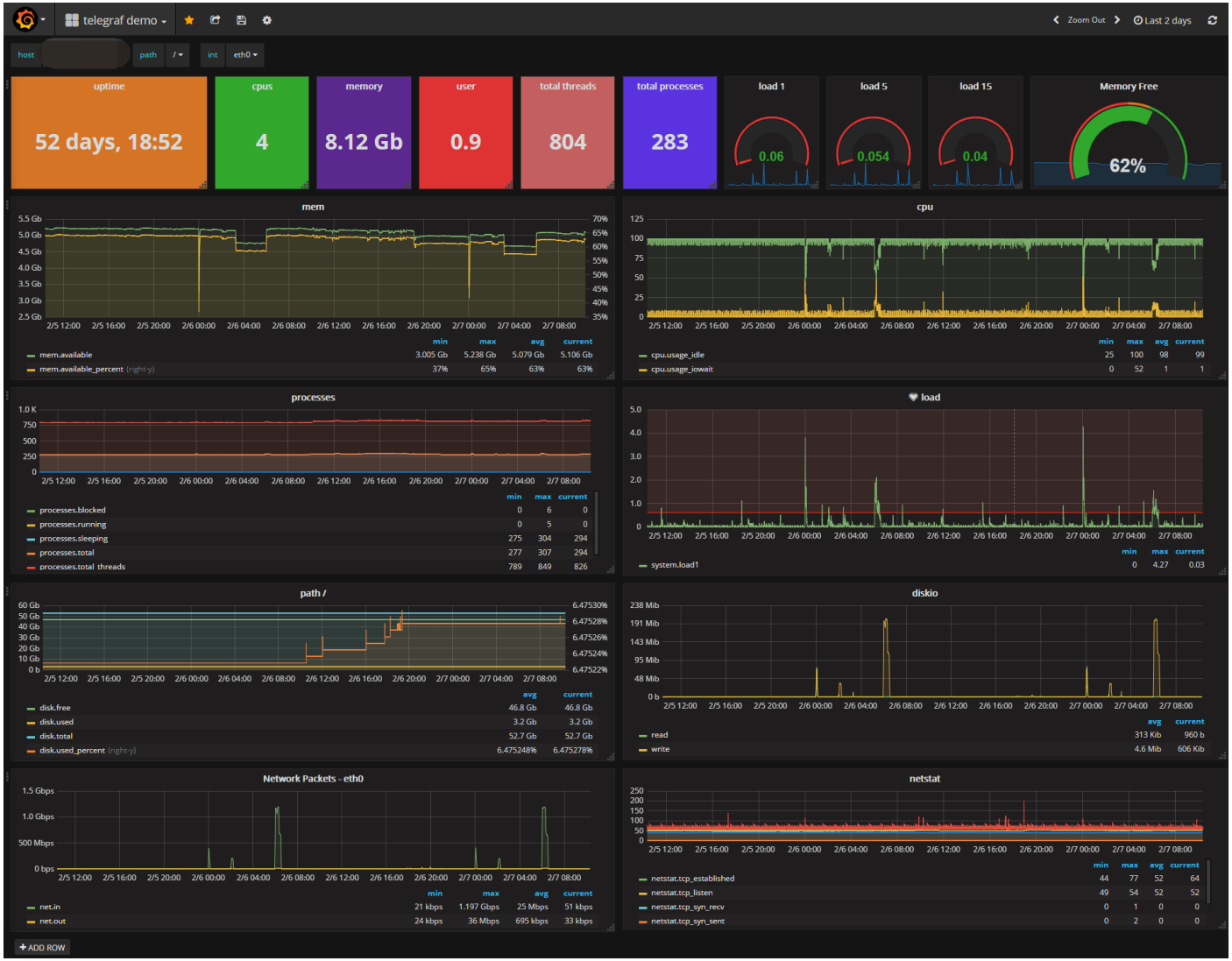
- .CSV
  - Majority of inputs
  - Length? Header? Encoding?
- DB with a schema (in memory?)
- Graph databases, ontologies, RDF…
- Key-value stores (redis)
- Time series databases (openTSDB, influxDB)
  - Time series + metadata
- „Data in motion”
  - Streams as input for processing/analysis

Time series example: influxDB

- Data: measurement
  - Fields, tags, timestamp

AGGREGATIONS	SELECTORS	TRANSFORMATIONS
COUNT()	BOTTOM()	CEILING()
DISTINCT()	FIRST()	DERIVATIVE()
INTEGRAL()	LAST()	DIFFERENCE()
MEAN()	MAX()	FLOOR()
MEDIAN()	MIN()	HISTOGRAM()
SPREAD()	PERCENTILE()	NON_NEGATIVE_DERIVATIVE()
SUM()	TOP()	STDDEV()

Dashboards… (e.g. Grafana)



https://grafana.com/dashboards/1443

DATA ANALYSIS

Data mining „brickstones”

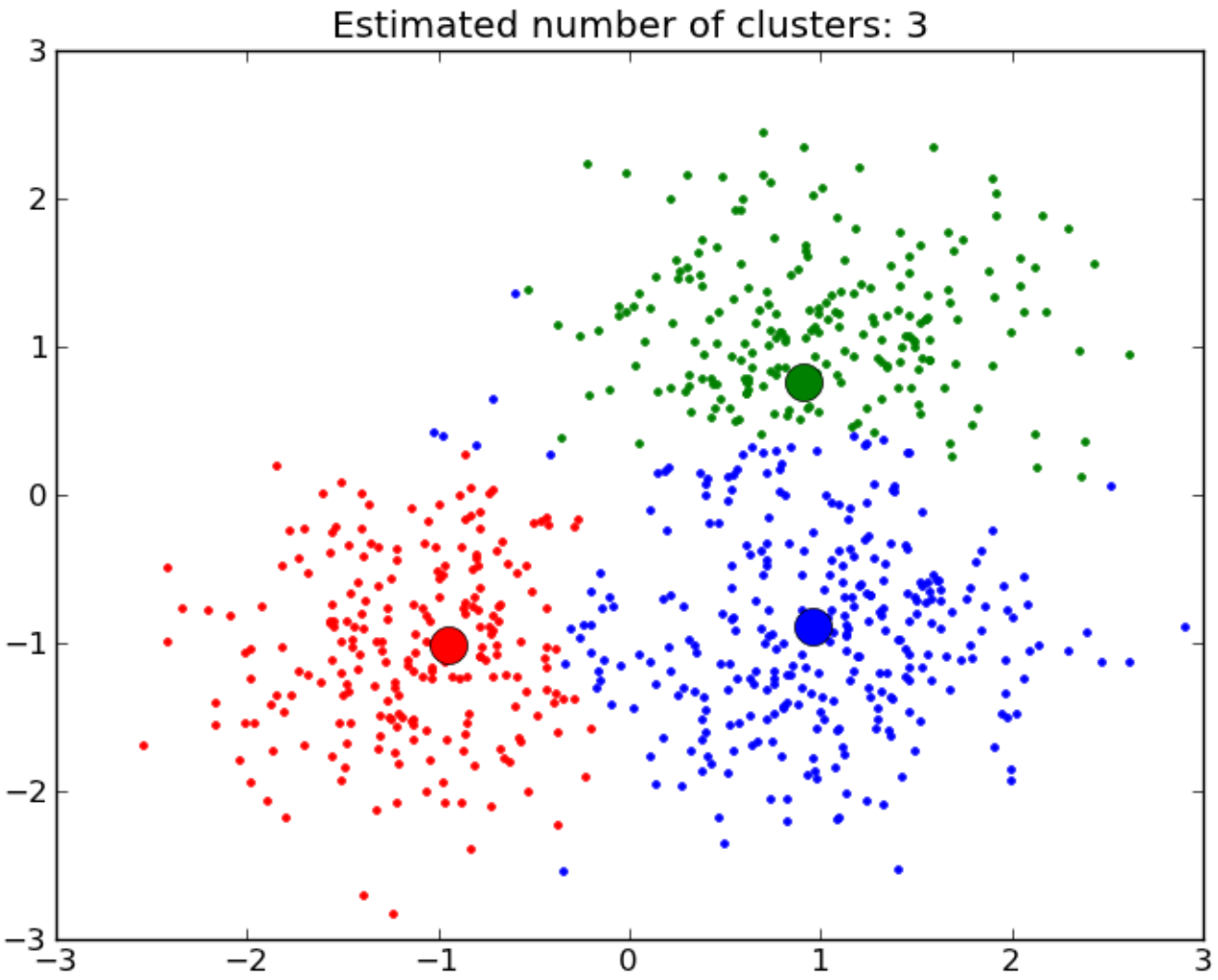
Clustering

Classification

Association rules

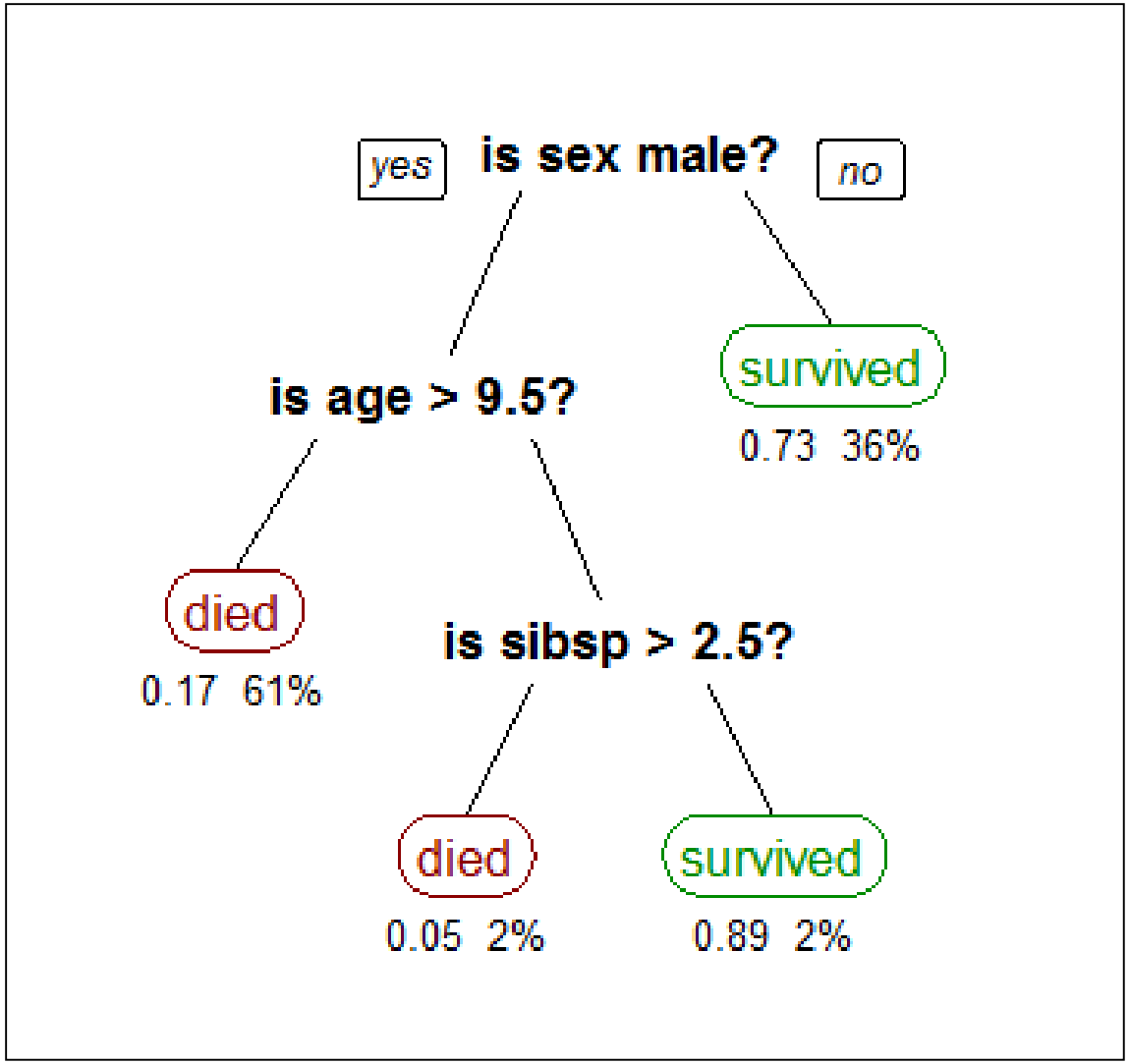
Regression

Clustering

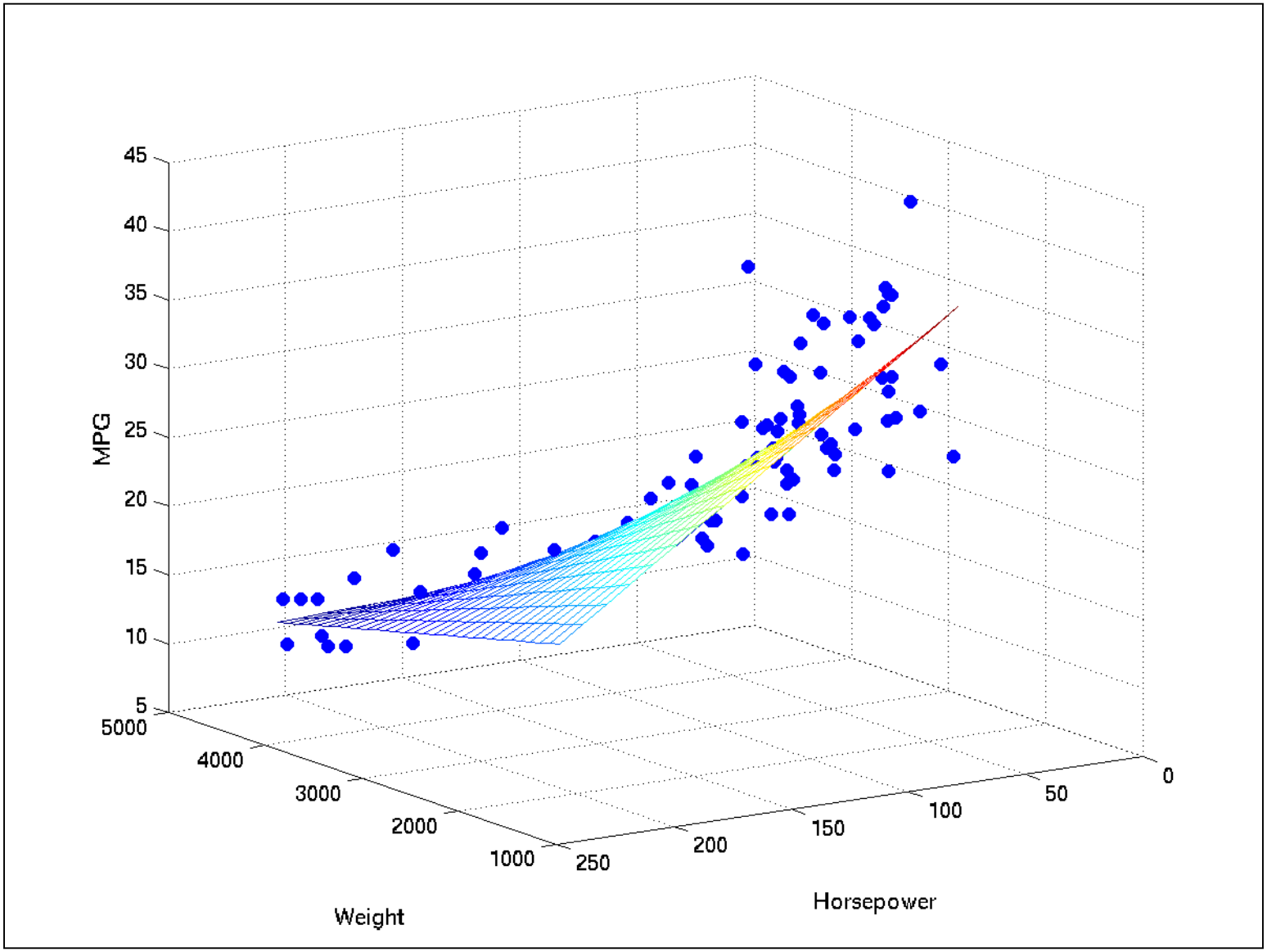


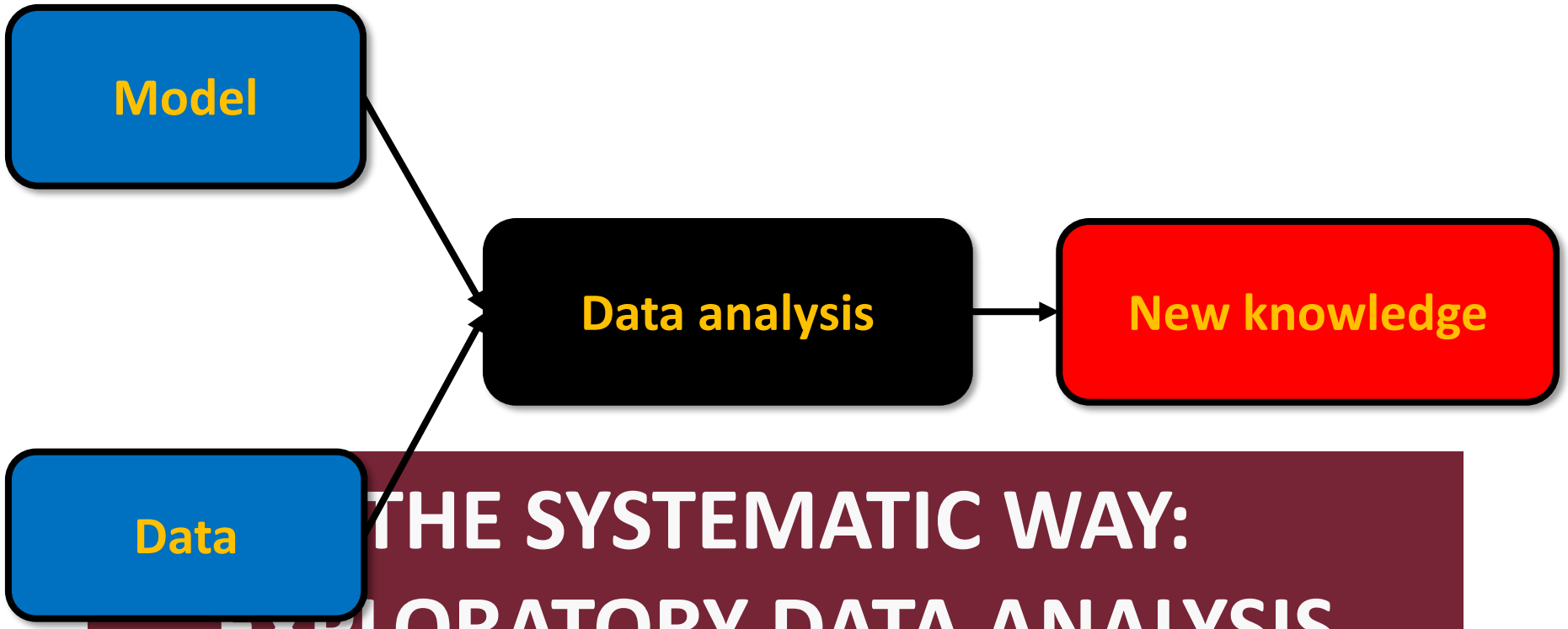
Association rules

Classification



Regression


MPG
45
40
35
30
25
20
15
10
5
Weight
5000
4000
3000
2000
1000
Horsepower
250
200
150
100
50
0



THE SYSTEMATIC WAY: EXPLORATORY DATA ANALYSIS

Look and see

Dr. Snow and 1854 London cholera epidemics

„About half of our sensory neurons are dedicated to vision, endowing us with a remarkable pattern-recognition ability.”

Prof. Alfred Inselberg

Exploratory data analysis (EDA)

- Summary of the main characteristics of a data set
  - Identification of outliers, trends, other patterns
  - Often with visual methods.
  - A statistical model can be used or not,
- „For seeing what the data can tell”
  - beyond formal modeling or hypothesis testing
  - hypotheses → new data collection and experiments

Approach-visual exploratory analytics

Resources

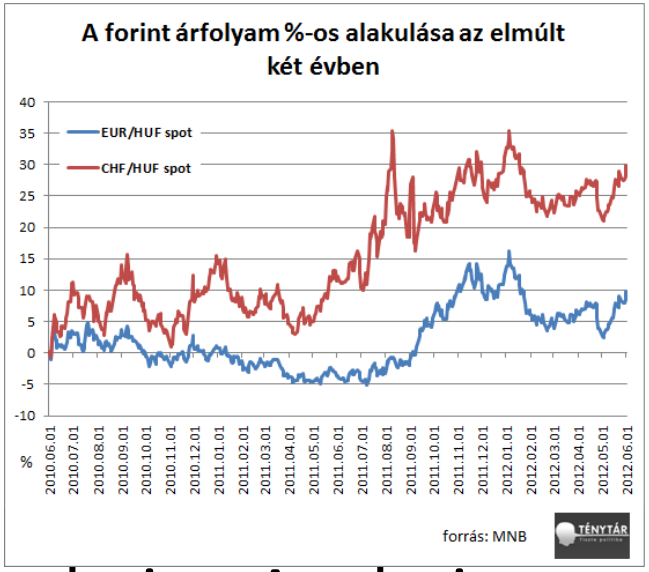
- 120.000.000 sensors
- 10^{10} processors

Process based on interactivity

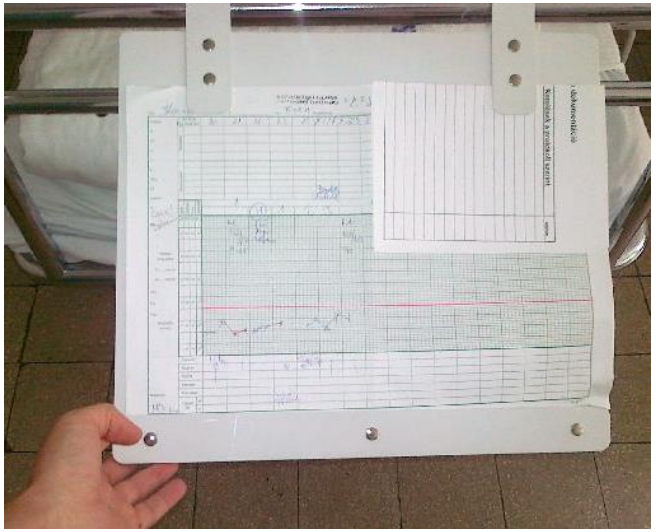
1. Graphical presentation of the data
   – multiple diagrams
2. Visual evaluation
   – exploiting human overview
3. Visual selection, manipulation – multiple diagrams
4. Interpretation, correlation with other models, evaluation (like architecture etc.)

Visualisation in Everyday Life

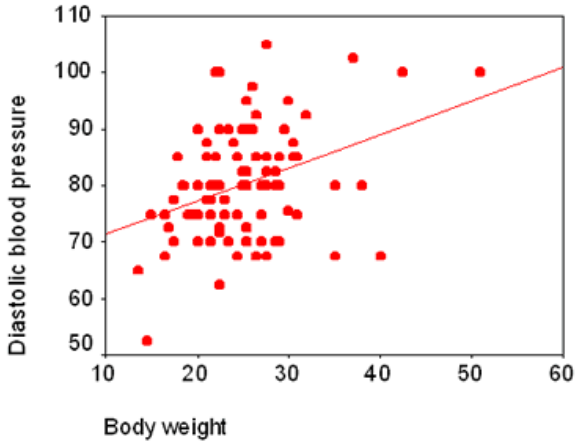
Trend Analysis and Forcast



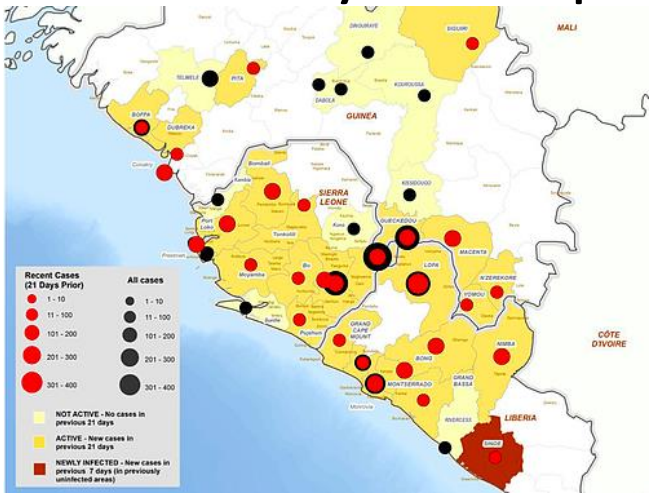
Time Series Analysis



Correlation Analysis



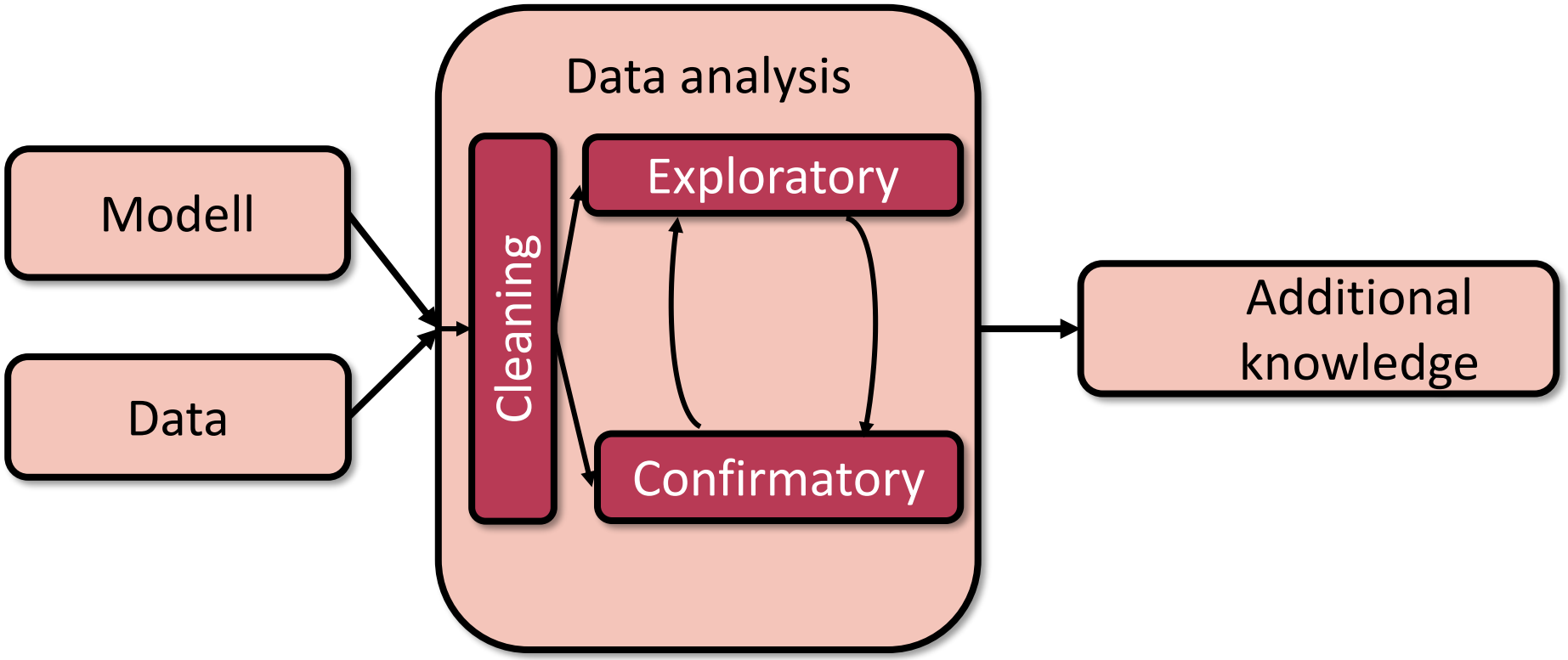
Analysis of Spatial Data



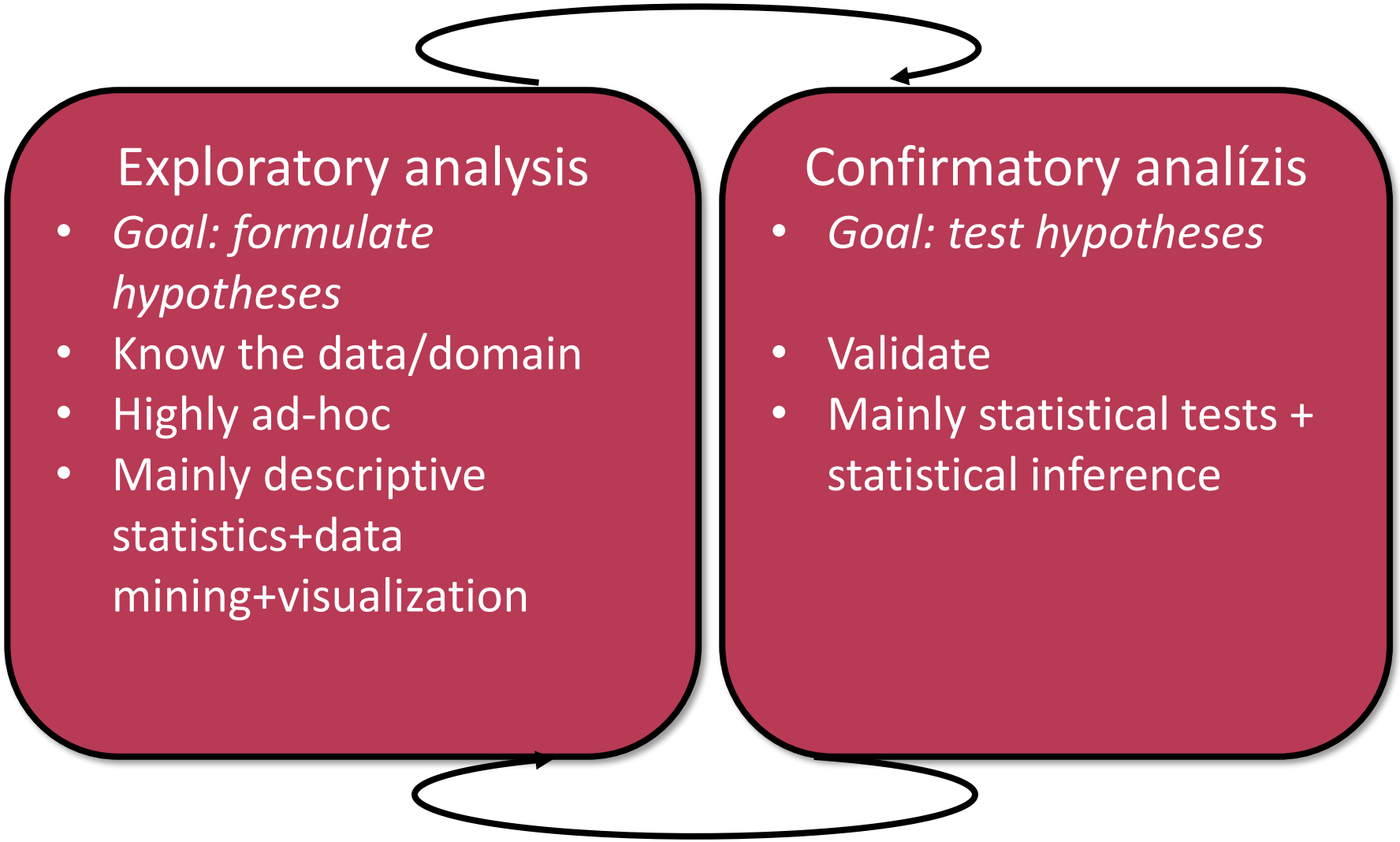
Additional knowledge

- „People buying coffee often buy milk”
- „There is a significant difference in salaries depending on gender”
- „The memory consumption of a software grows exponentially wrt. number of requests in queue”.
- „The population follows a $N(100, 15)$ disribution”
- „BME students fall into 3 main different groups (according to their grades)”

Data analysis



Data analysis

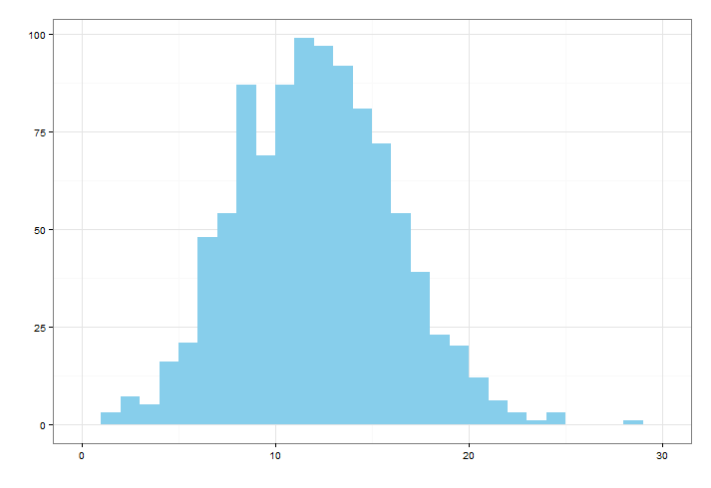


Data analysis

- E.g. distribution analysis

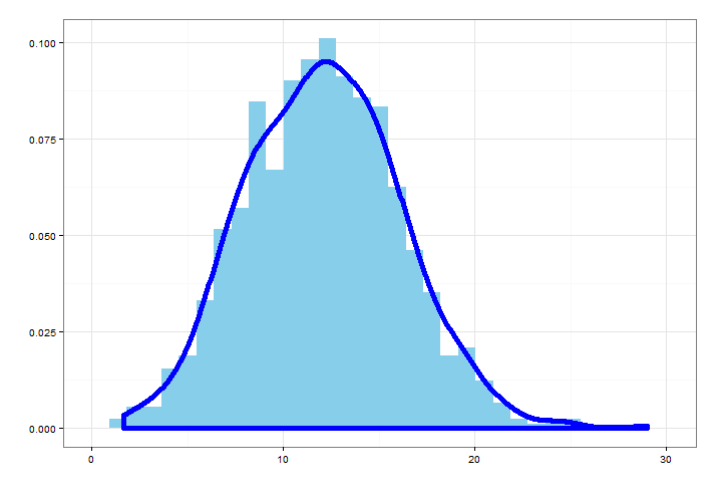
Exploratory

Hypothesis: variable x follows normal distribution



Confirmatory

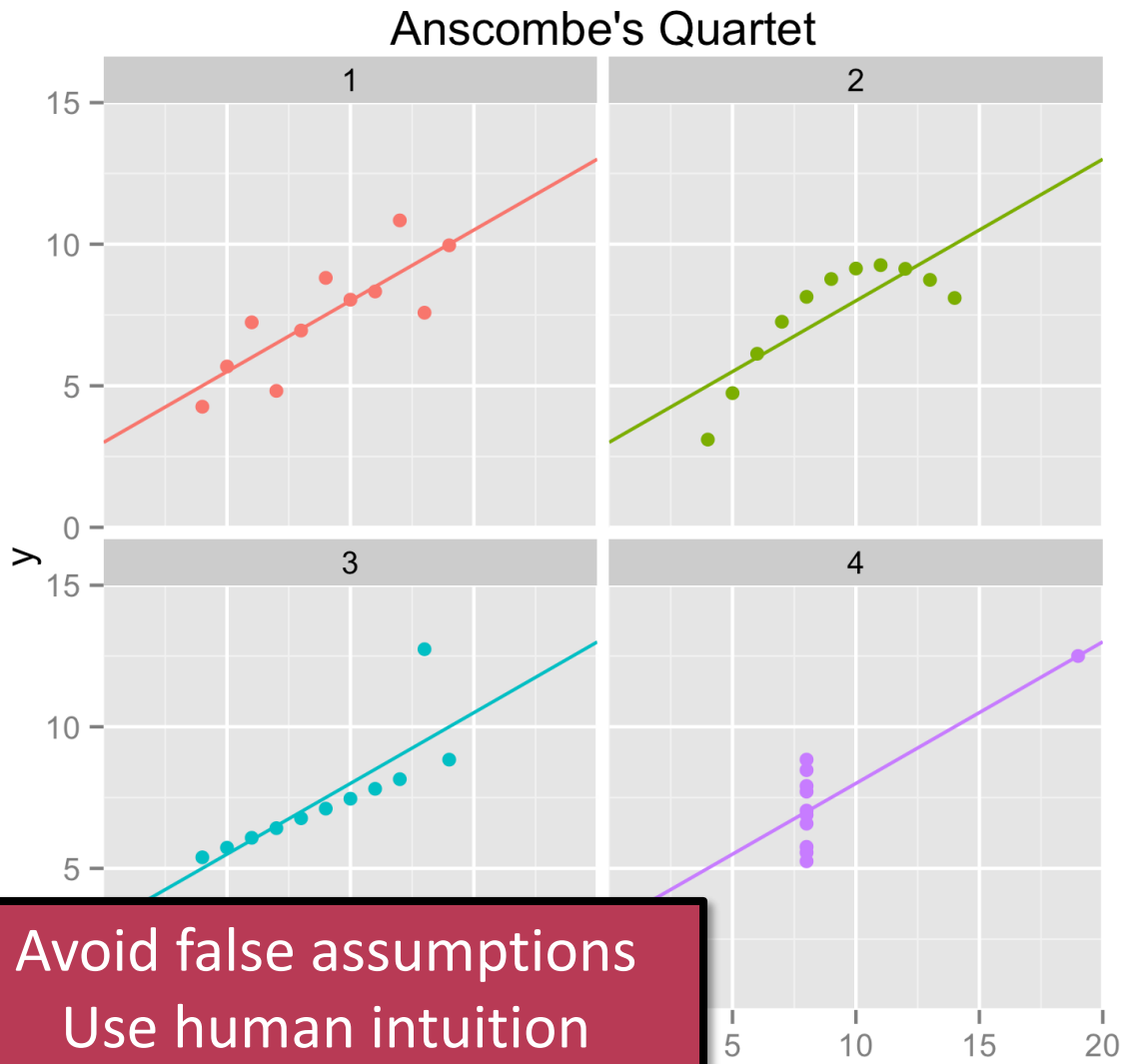
Variable x follows $N(12, 4)$ distribution



Exploratory Data Analysis

- Goal: hypothesis formulation
- Pattern recognition
- Early validation
- „Sensors of type X are sensitive to high temperature”
- „Application of Type Y is sensitive to CPU load”
- Interactive, human expert needed
- Later: automated support(IBM Watson Analytics)

Validation for automated methods



- For all cases:

Means:

$M[x] = 9$

$M[y] \sim 7.5$

Variance:

$\sigma[x] = 11$

$\sigma[y] \sim 4.12$

Correlation:

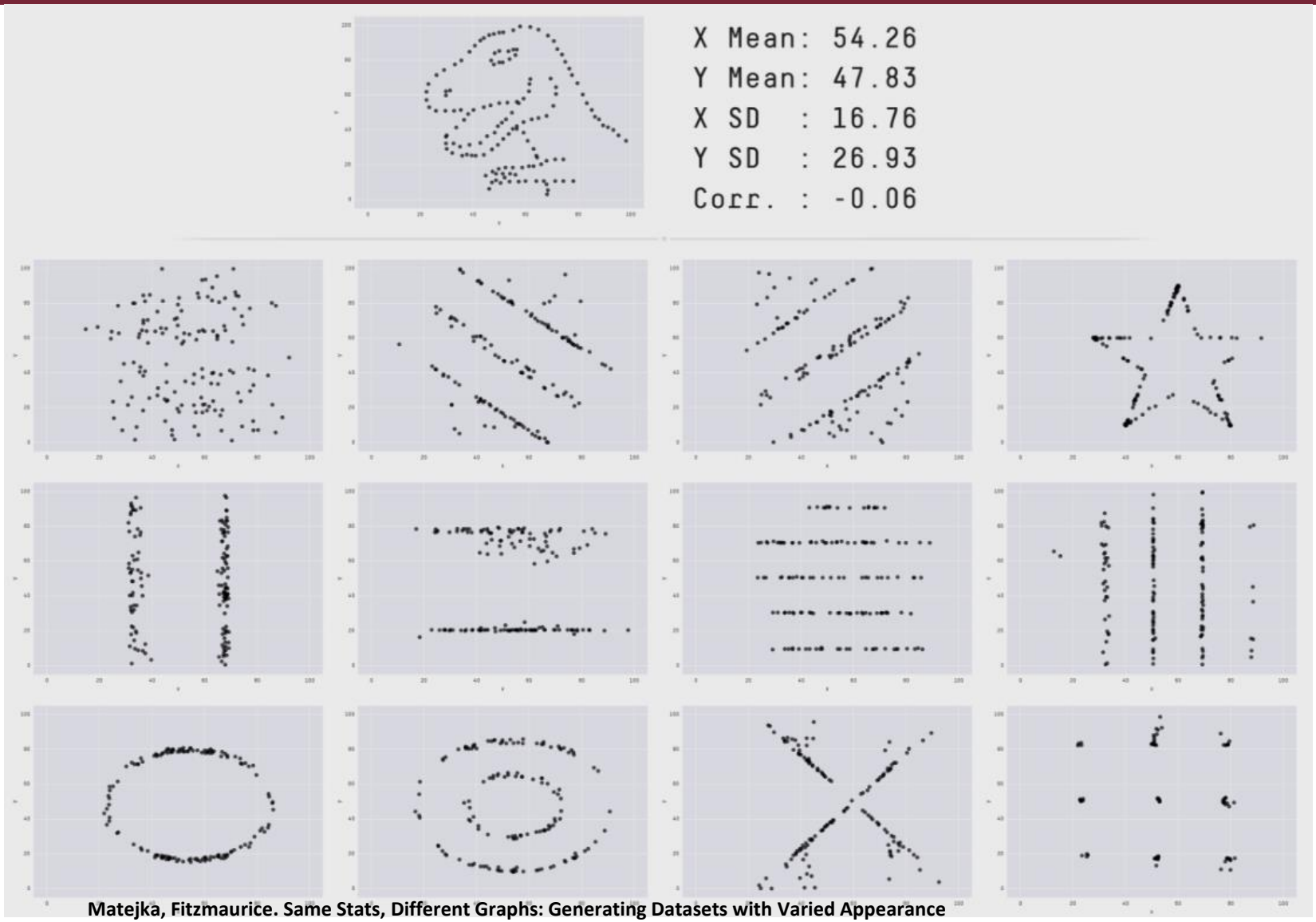
$C(x, y) \sim 0.816$

Regression:

$y \sim 3 + 0.5x$

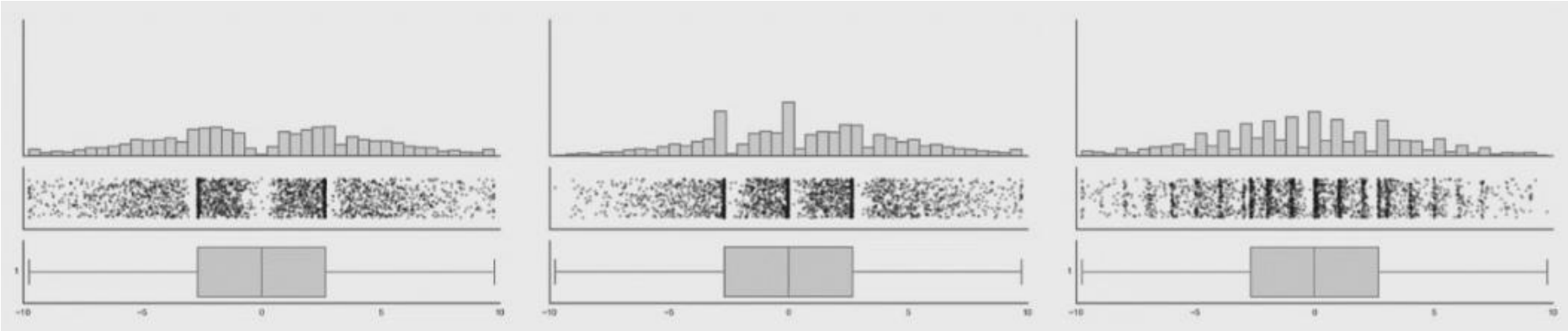
Avoid false assumptions
Use human intuition
(1973)

… and some more

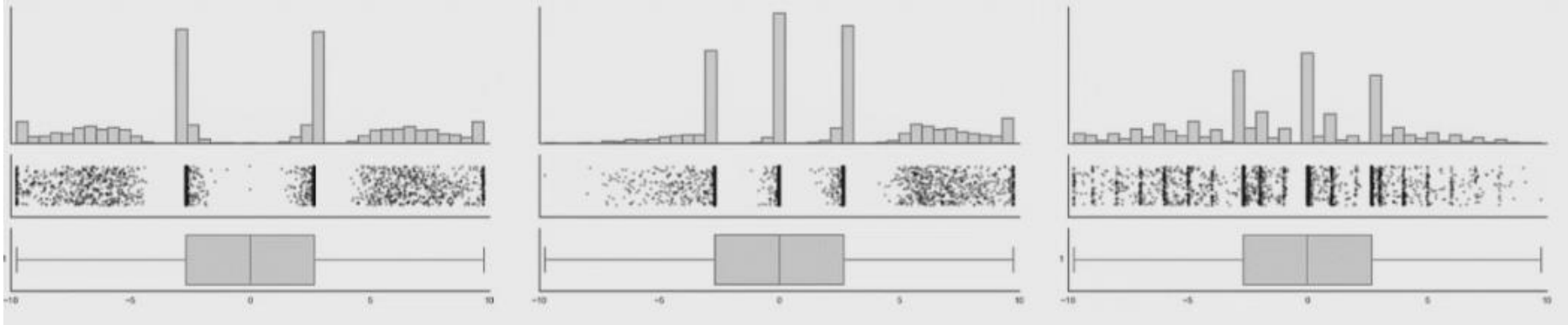


Matejka, Fitzmaurice. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. ACM SIGCHI Conference on Human Factors in Computing Systems
https://www.autodeskresearch.com/publications/samestats

Distribution vs summary



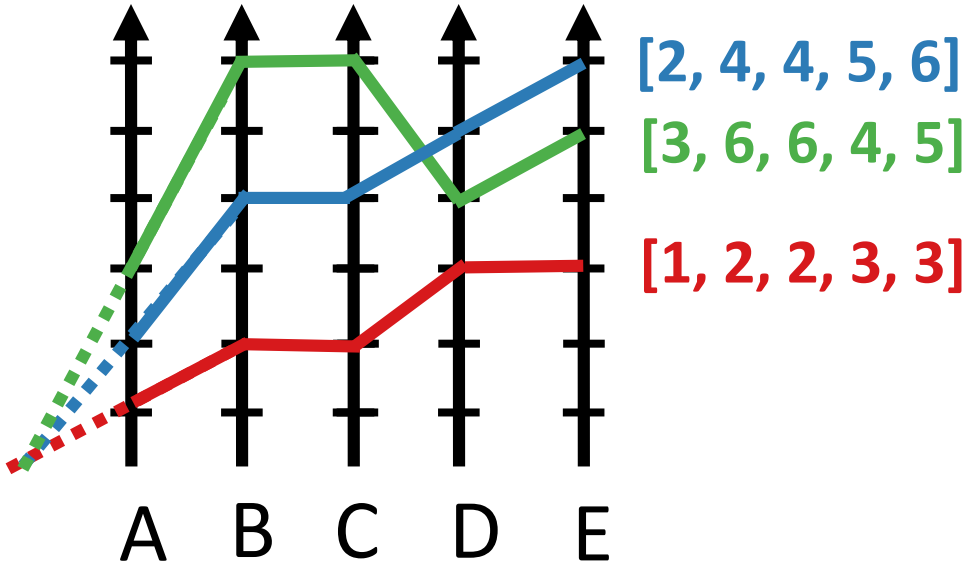
Same summary != same story

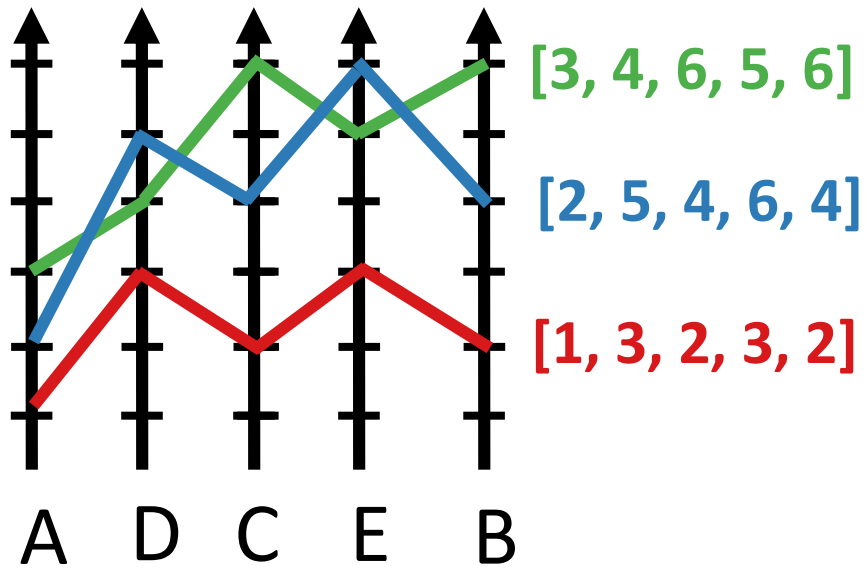


Matejka, Fitzmaurice. Same Stats, Different Graphs: Generating Datasets with Varied Appearance
and Identical Statistics through Simulated Annealing. ACM SIGCHI Conference on Human Factors in Computing Systems
https://www.autodeskresearch.com/publications/samestats

Parallel Coordinates

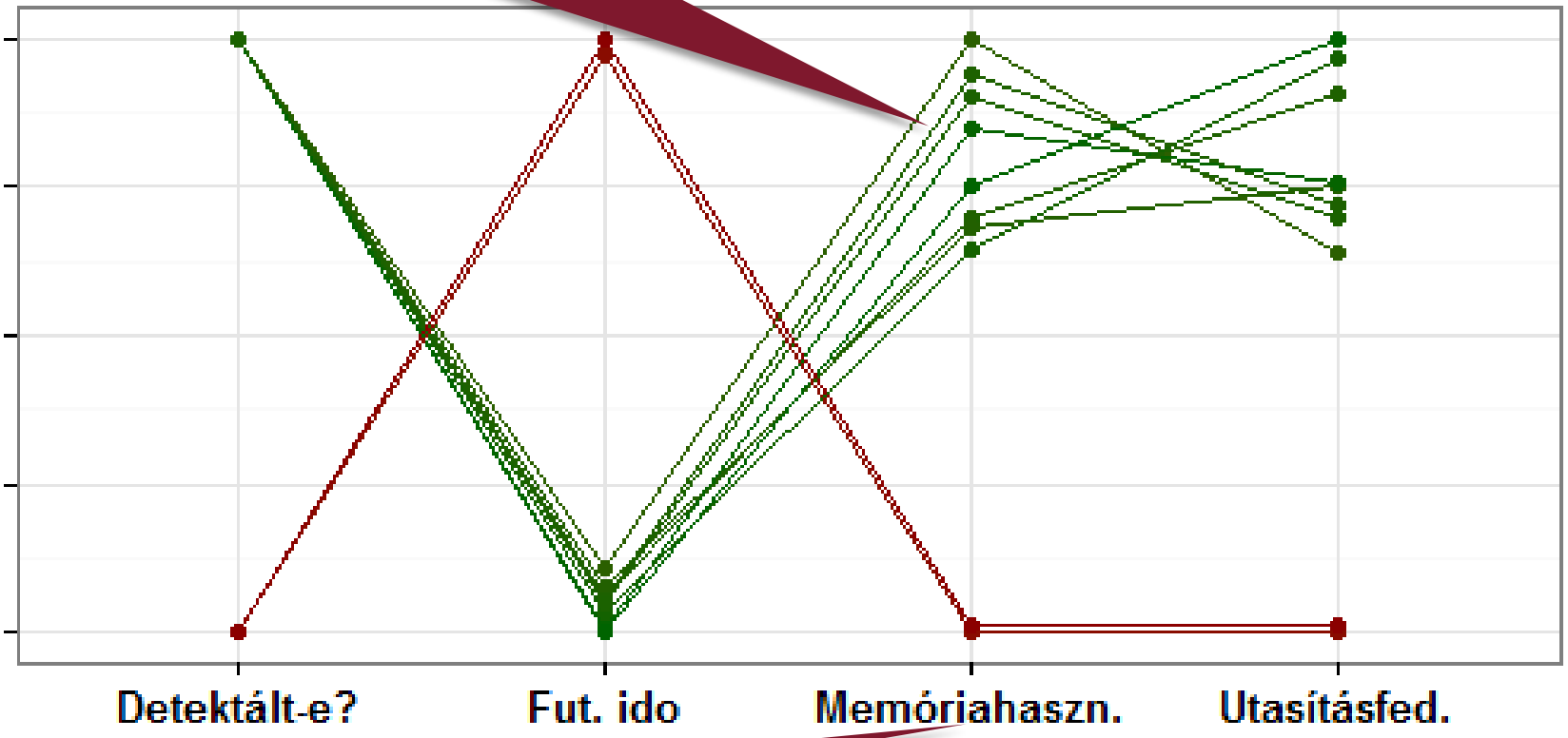
- Multi-dimensional visualization
- Compact, scalable
- Axis order?





Parallel Coordinates: Analysis of the Test Cases

1 test case: 1 broken line

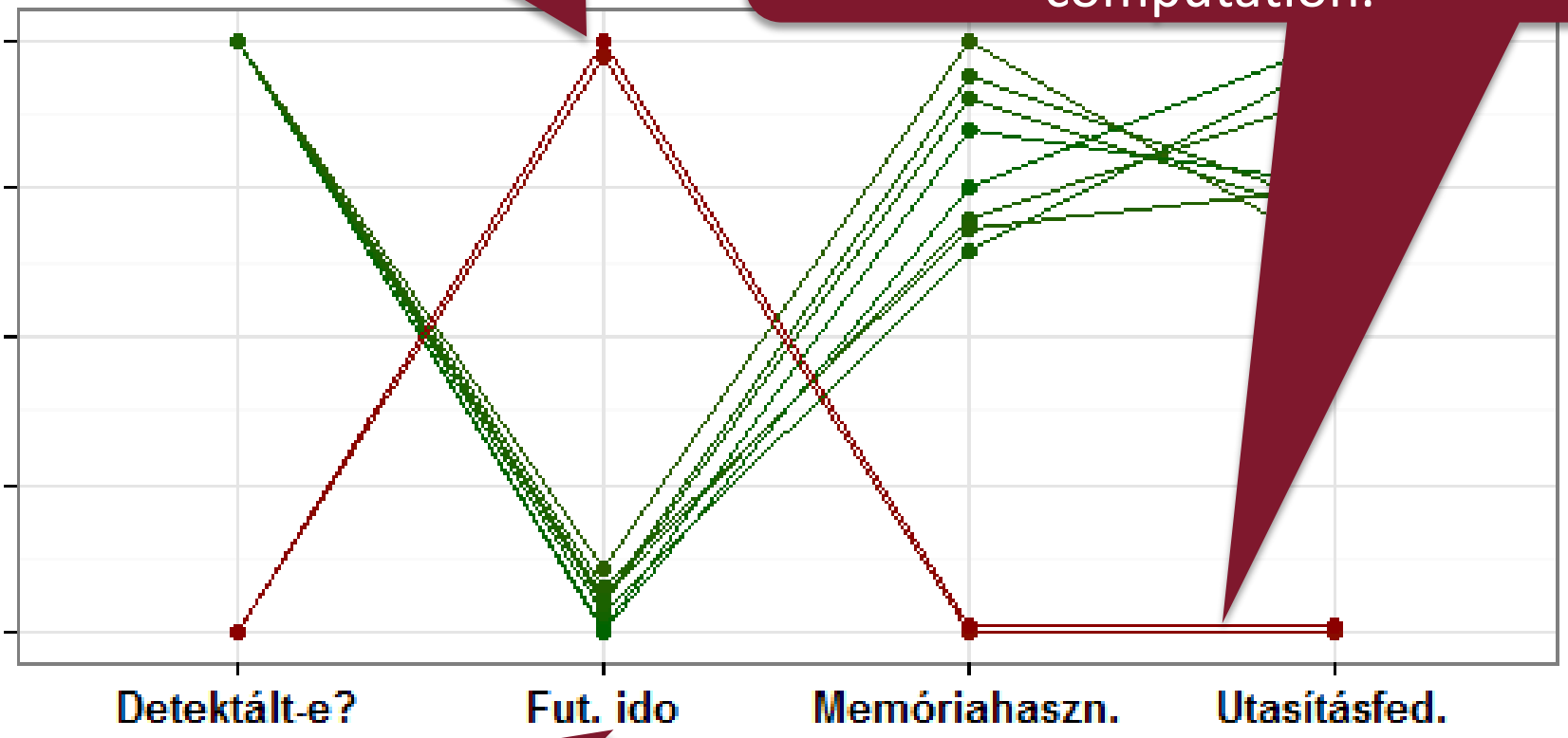


The variables appear on the x-axis

Parallel Coordinates: Analysis of the Test Cases

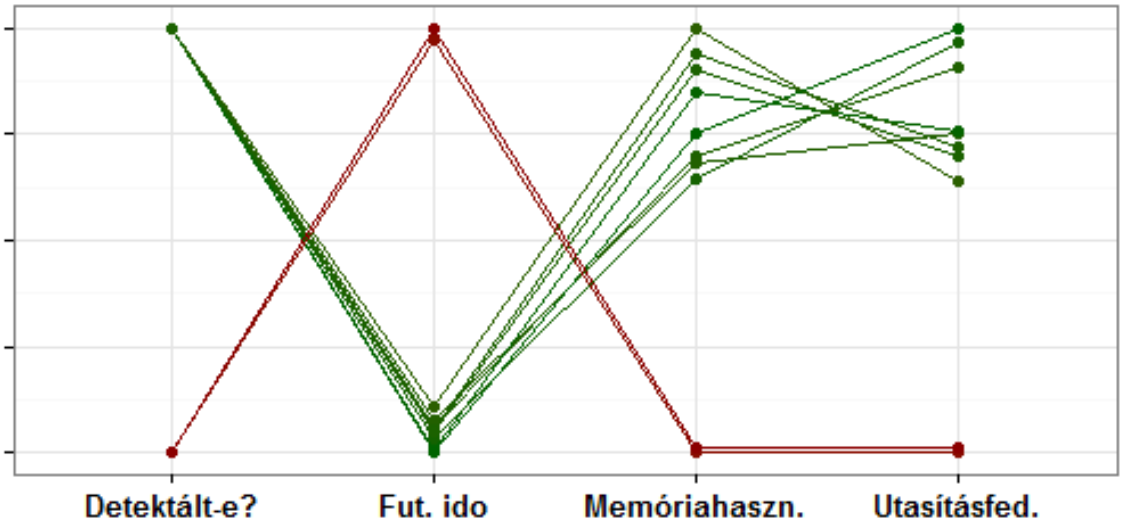
Timeout?

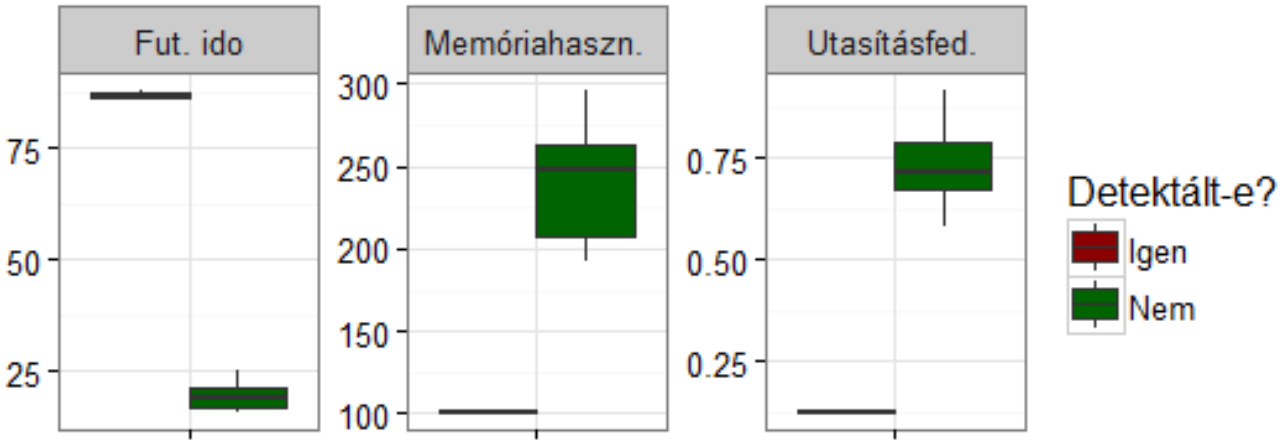
The ones detecting an error did not even came to the actual computation.



Run time and memory usage seem to be in a positive relation (if the test is successful)

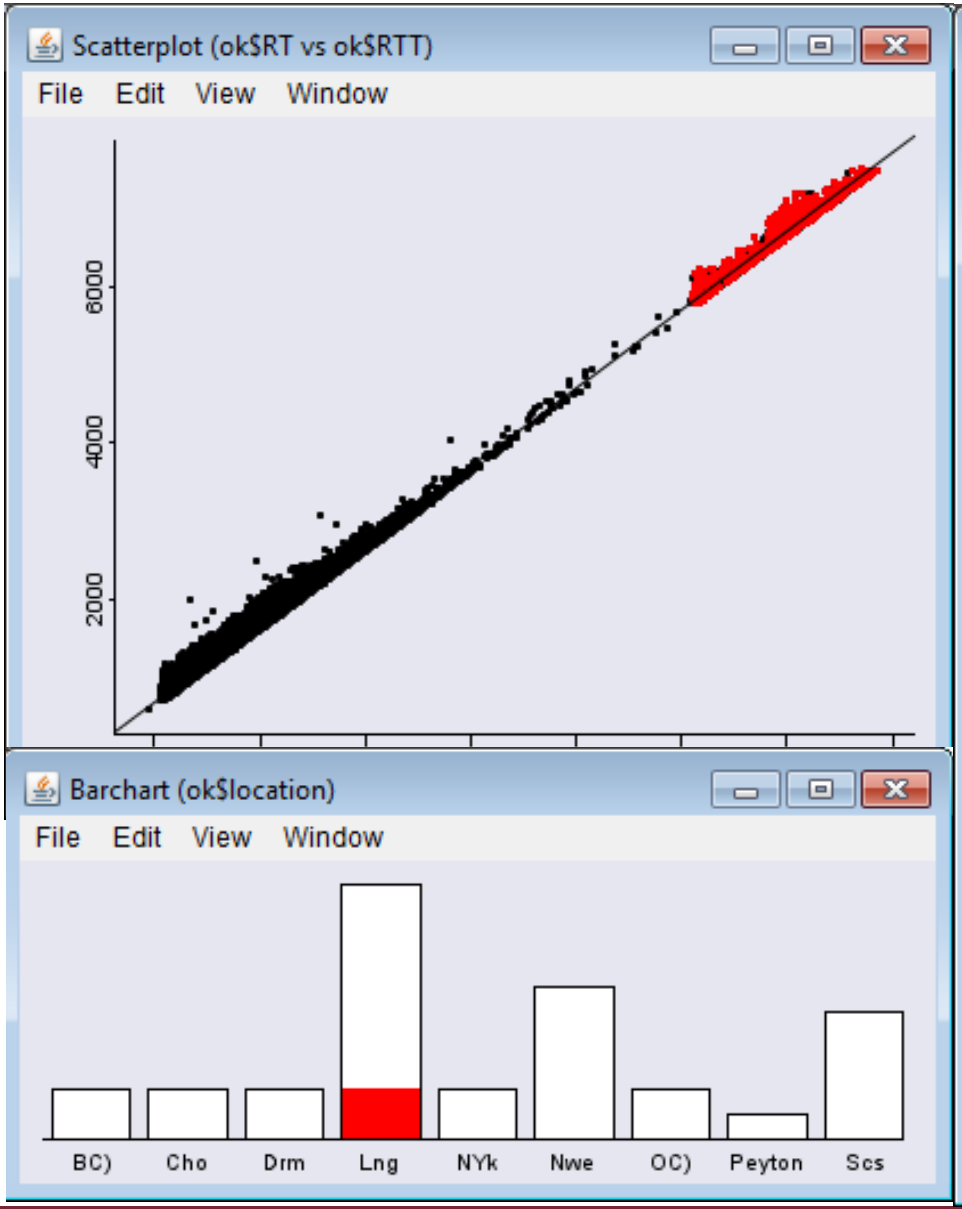
Parallel Coordinates: the Alternatives

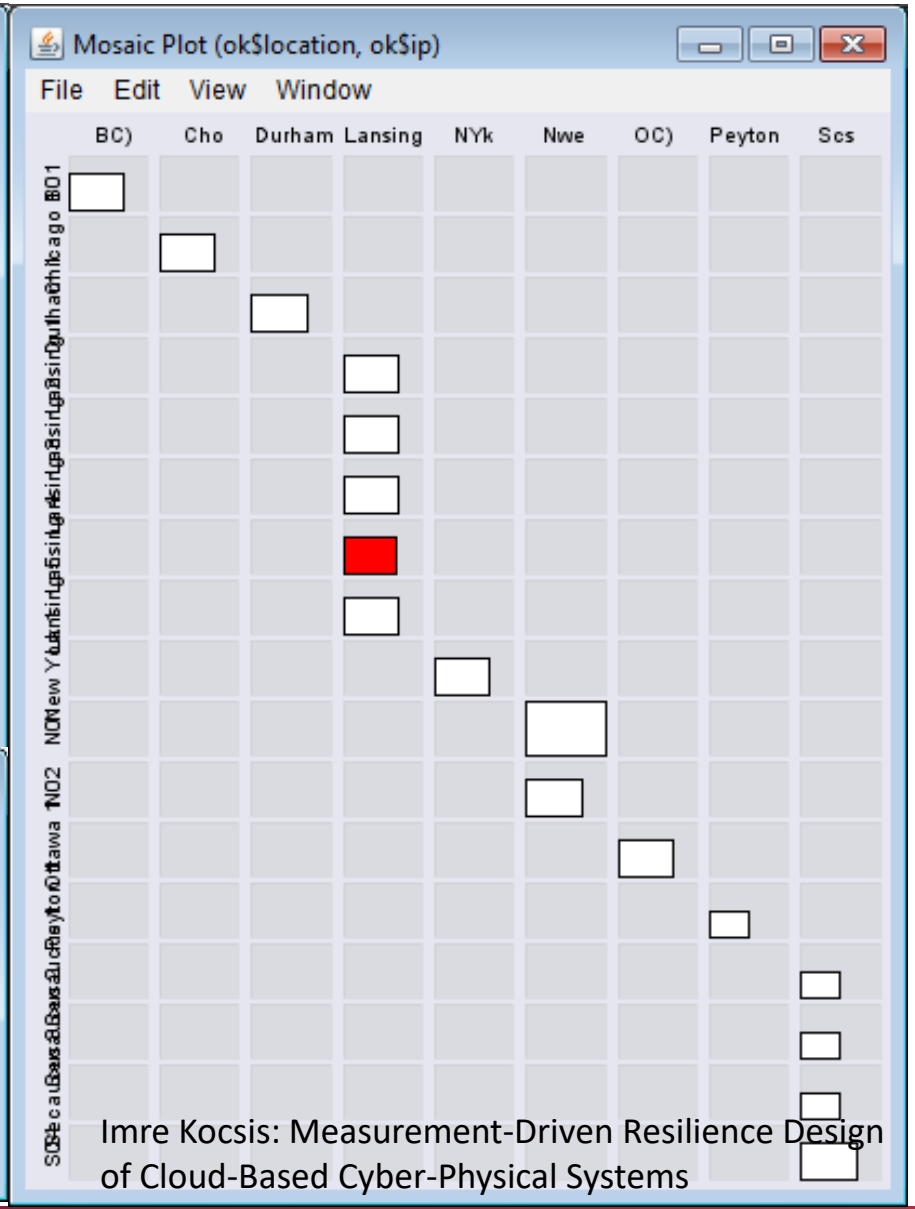




VISUAL EDA EXAMPLES

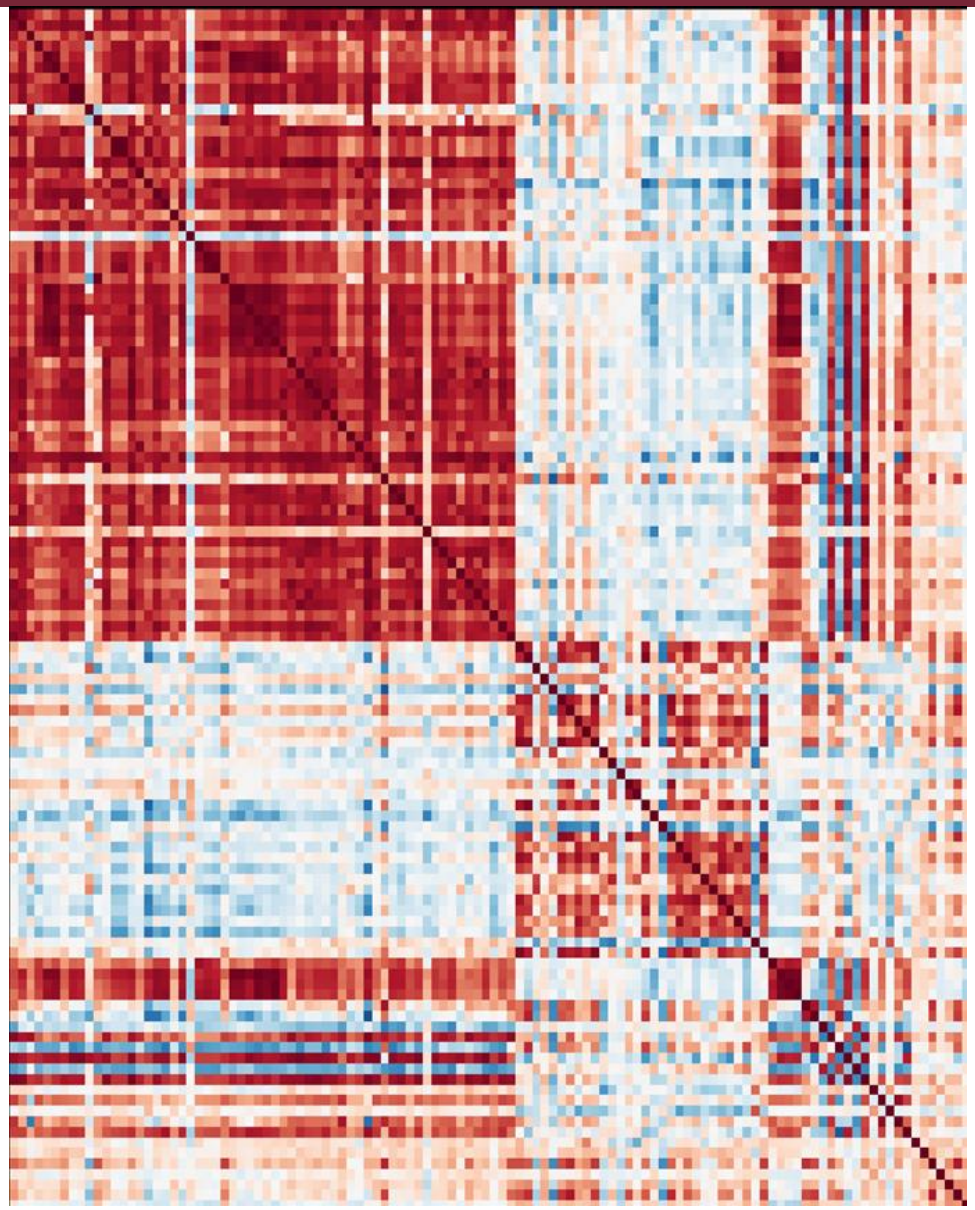
EDA example



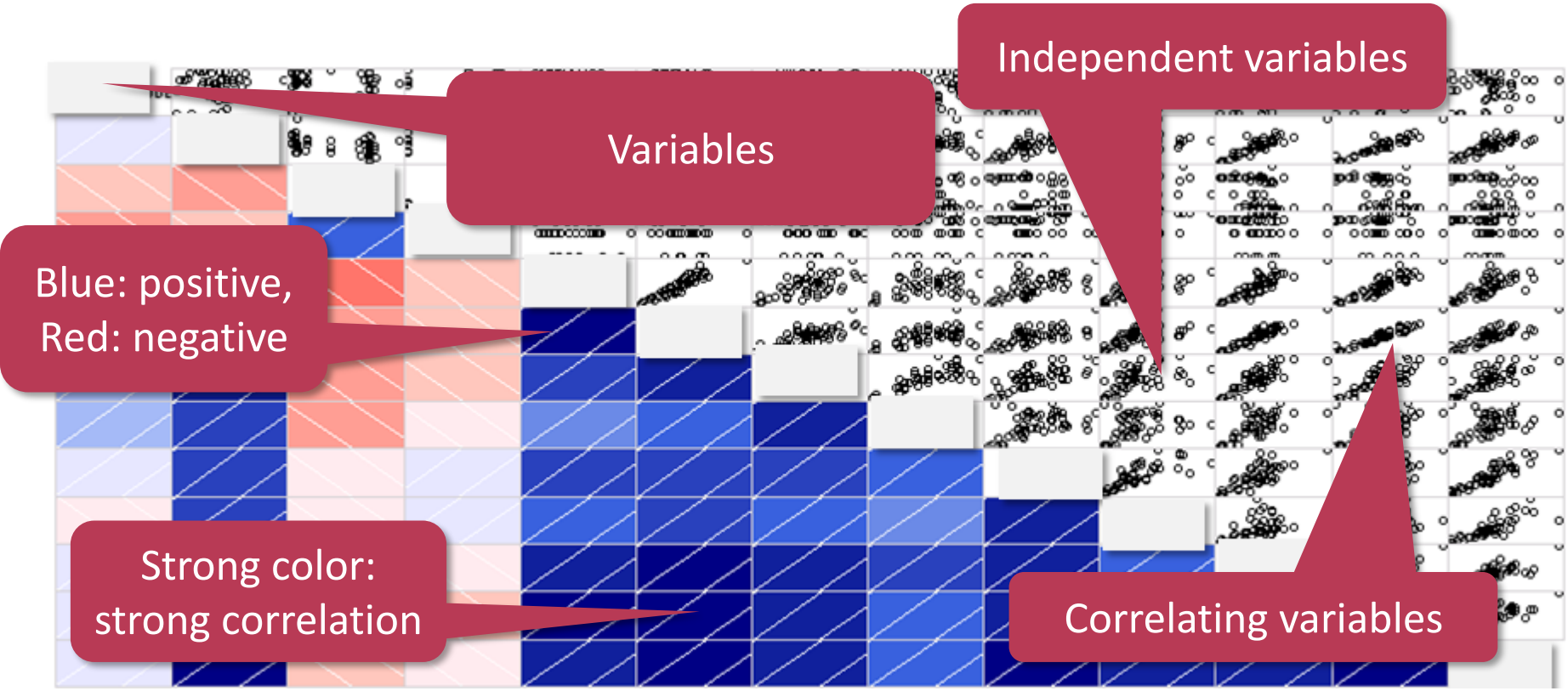


Imre Kocsis: Measurement-Driven Resilience Design of Cloud-Based Cyber-Physical Systems

EDA example2

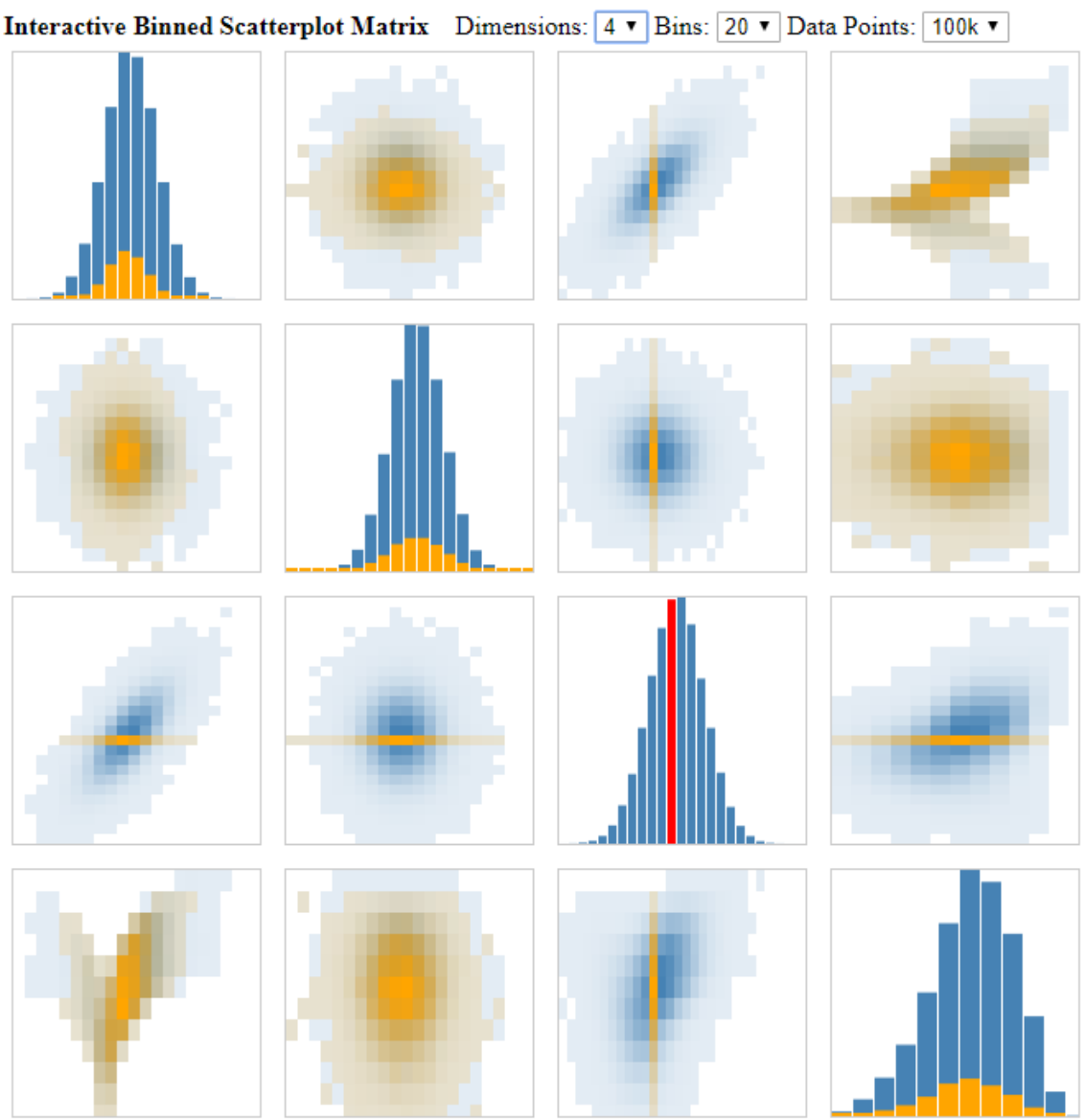


EDA example3: pairwise correlation



R „corrgram package”
Pearson linear correlation
Scatterplot matrix
Goal: identify correlating variables, outliers
→ Dimension reduction
→ Feature selection

Pairwise analysis



http://vis.stanford.edu/projects/datavore/splom/

EFFECTIVE MULTI-DIMENSIONAL DATA VISUALIZATION

https://github.com/dipanjanS/practical-machine-learning-with-python/tree/master/bonus%20content/effective%20data%20visualization

https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57

Analysis of wine characteristics

- Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems. 2009 Nov 1;47(4):547-53.

References

- Tukey, John W. "Exploratory data analysis." (1977): 2.
- Inselberg, Alfred, and Bernard Dimsdale. "Parallel coordinates for visualizing multi-dimensional geometry." *Computer Graphics 1987*. Springer, Tokyo, 1987. 25-44.
- Kocsis Imre. Vizuális analízis. In: Antal Péter, Antos András, Horváth Gábor, Hullám Gábor, Kocsis Imre, Marx Péter, Millinghoffer András, Pataricza András, Salánki Ágnes. Intelligens adatelemzés. 141 p. Budapest: Typotex Kiadó, 2014. pp. 58-69. (ISBN:978-963-2791-71-5)
- Garrett Grolemund, Hadley Wickham. „R for Data Science” (O’Reilly, 2017) http://r4ds.had.co.nz/