

Approaches for identifying & selecting the right data

Sandjai Bhulai
Vrije Universiteit Amsterdam

IARIA panel discussion

Tim vor der Brück
Gaia Ceresa
Gerald Fahner
Gregor Grambow



VRJE
UNIVERSITEIT
AMSTERDAM

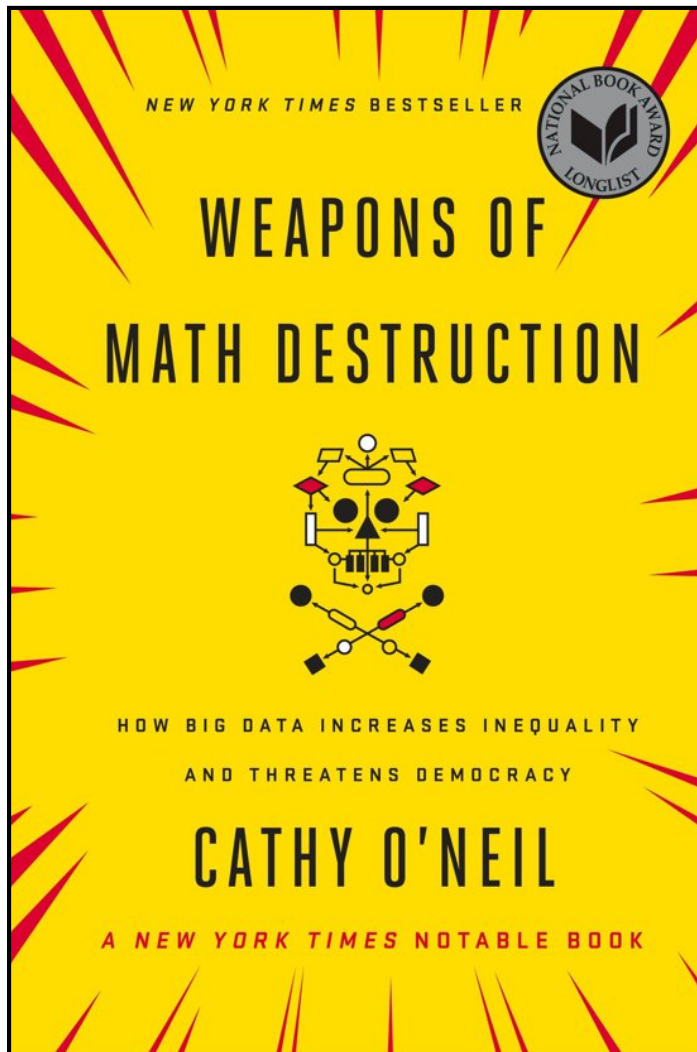
Faculty of Science

The era of data

2018 *This Is What Happens In An Internet Minute*

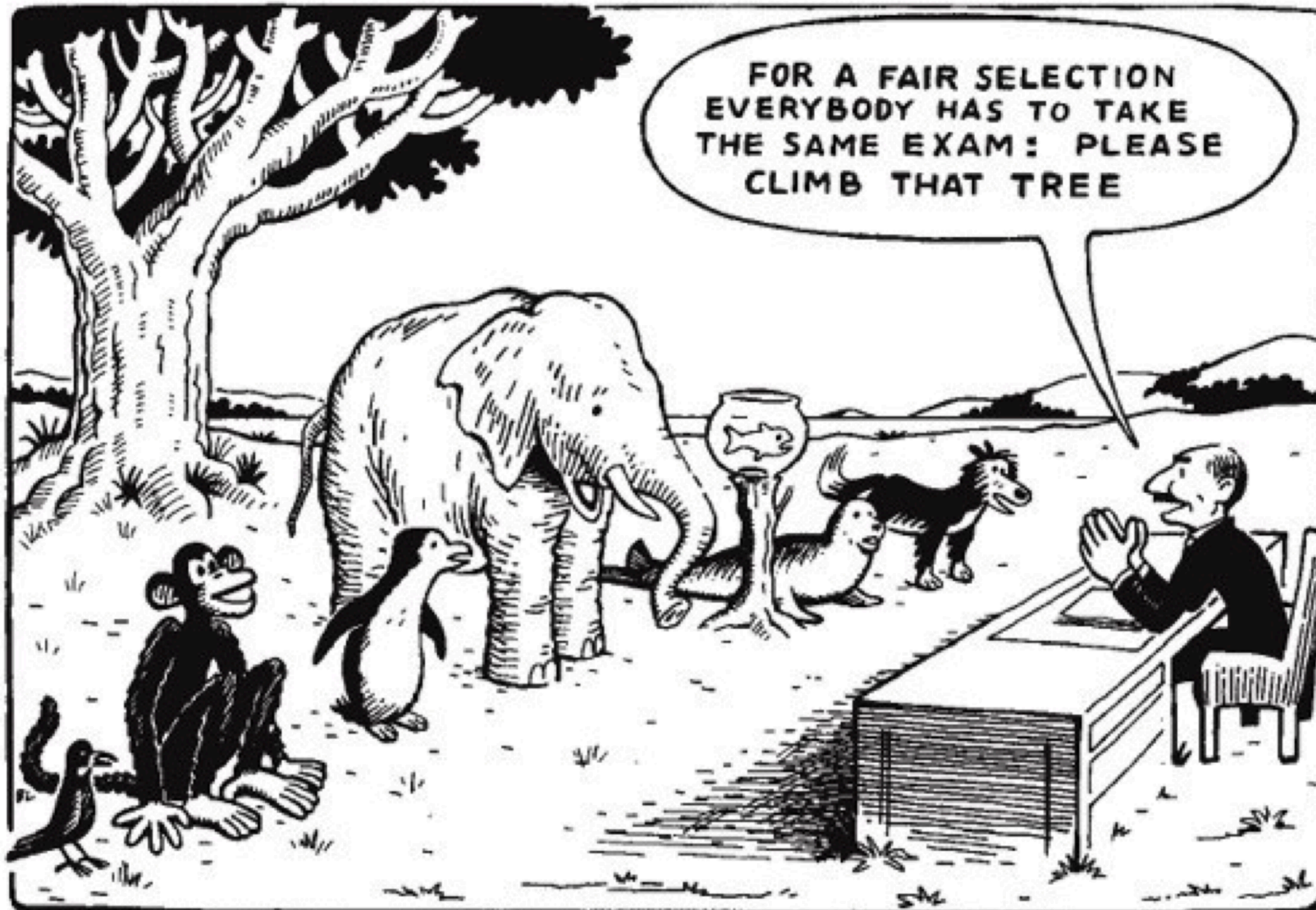


The dangers of models



- We live in the age of the algorithm
- Increasingly, the decisions that affect our lives are being made not by humans, but by mathematical models
- In theory, this should lead to greater fairness: everyone is judged according to the same rules, and bias is eliminated

The dangers of models



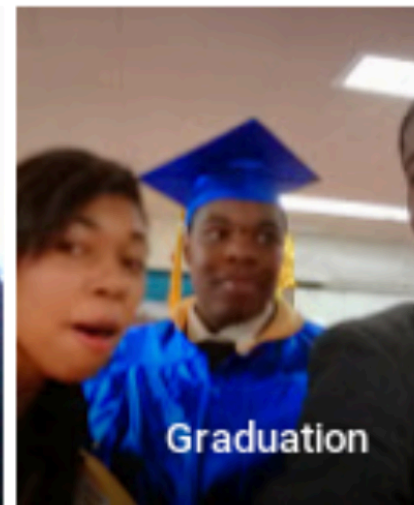
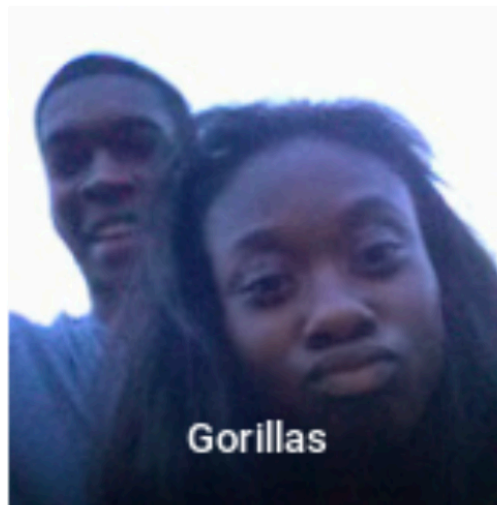
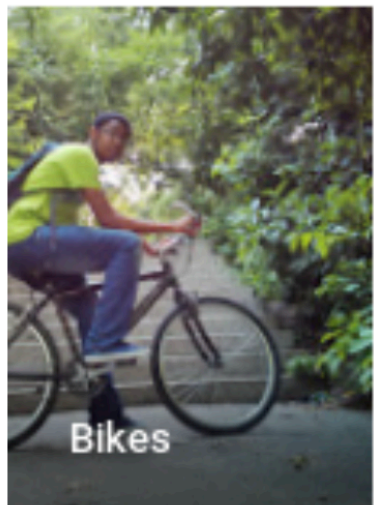
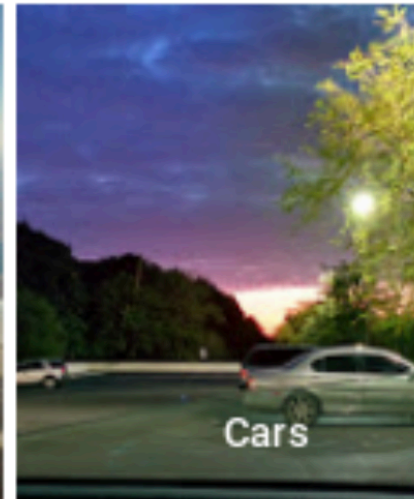
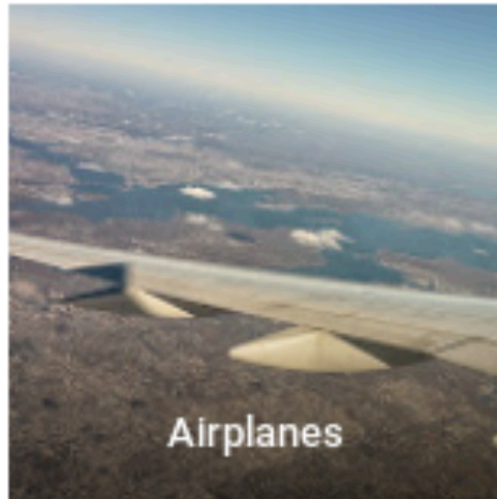
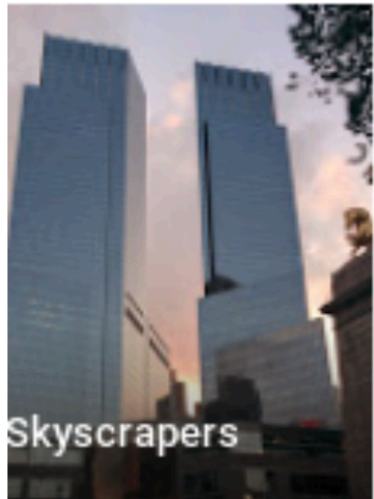
What are models doing?

- Inaccurate predictions versus “accurate” (\neq perfect), but unethical
 - > discrimination by gender, race, religion
- Is it customer centric or organization centric?
 - > an average person versus each individual?
- Based on what data?
 - > Biases, noise, ...
- With help of what (big) data analytics?
 - > Can a chosen approach give an answer to the set optimization goal?

Example: Google ads

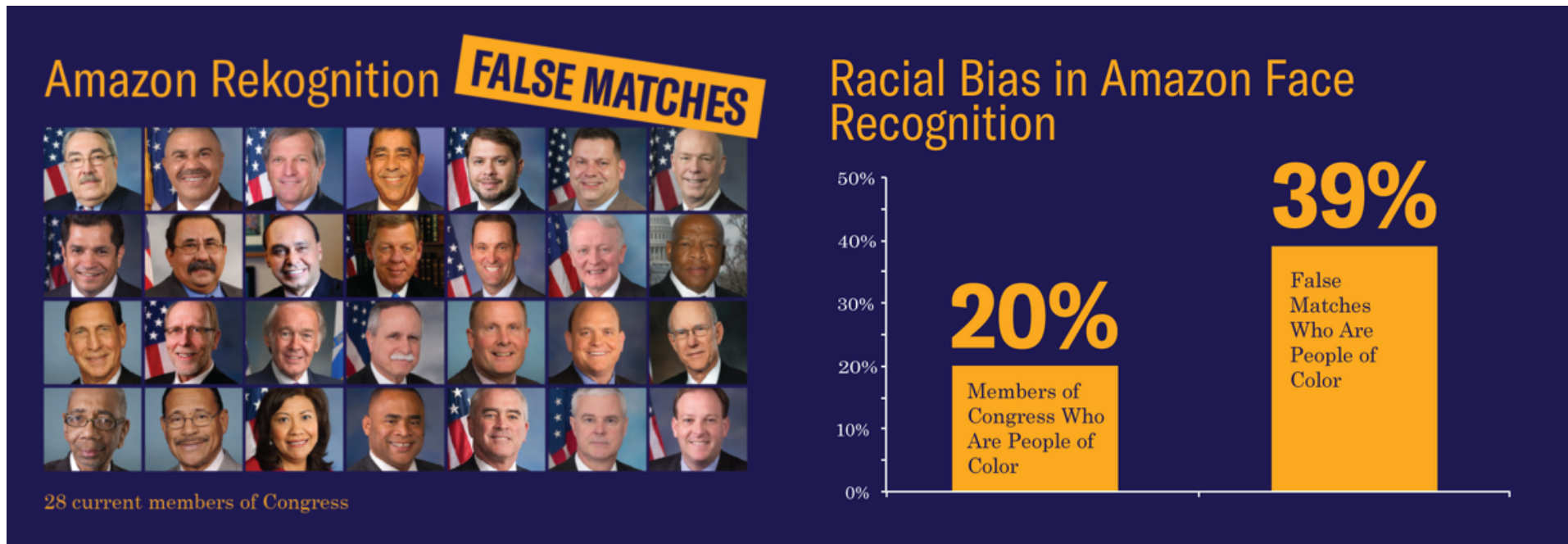
- “setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male”
Automated Experiments on Ad Privacy Settings
<http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf>
- ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity
<http://dataprivacylab.org/projects/onlineads/1071-1.pdf>
- target people who live in low-income neighborhoods with high-interest loans

Example: Google



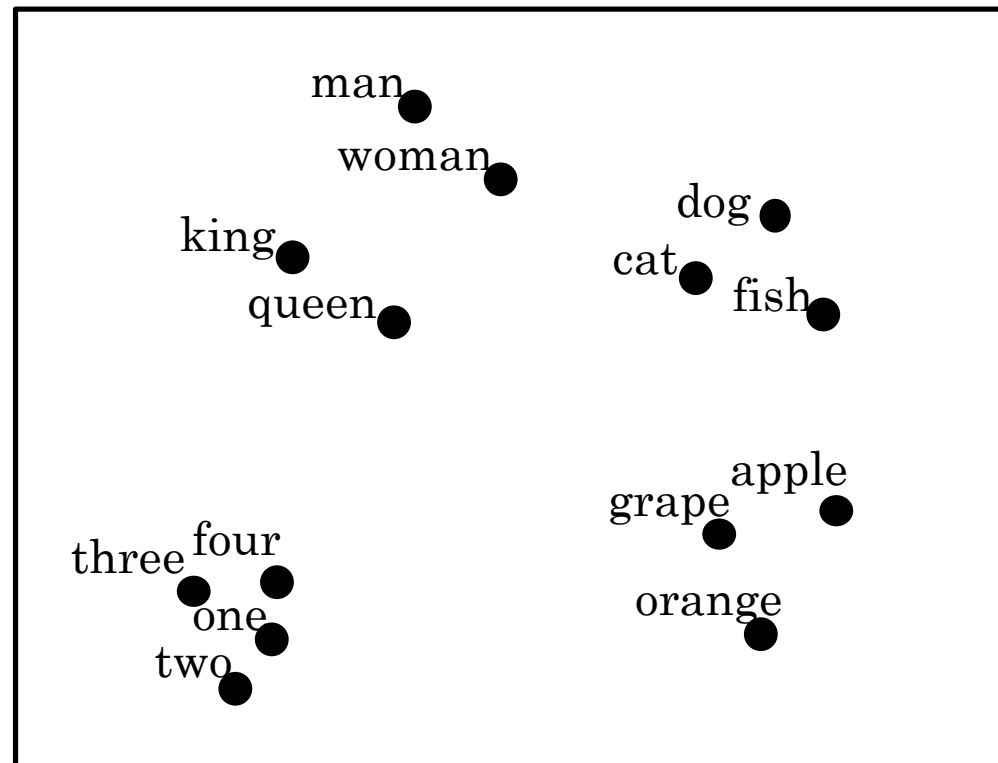
Example: Amazon recognition

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots



Example: word embeddings

- In natural language processing, it is common to learn embeddings from large text corpus (1-100B words) or download pre-trained embedding online



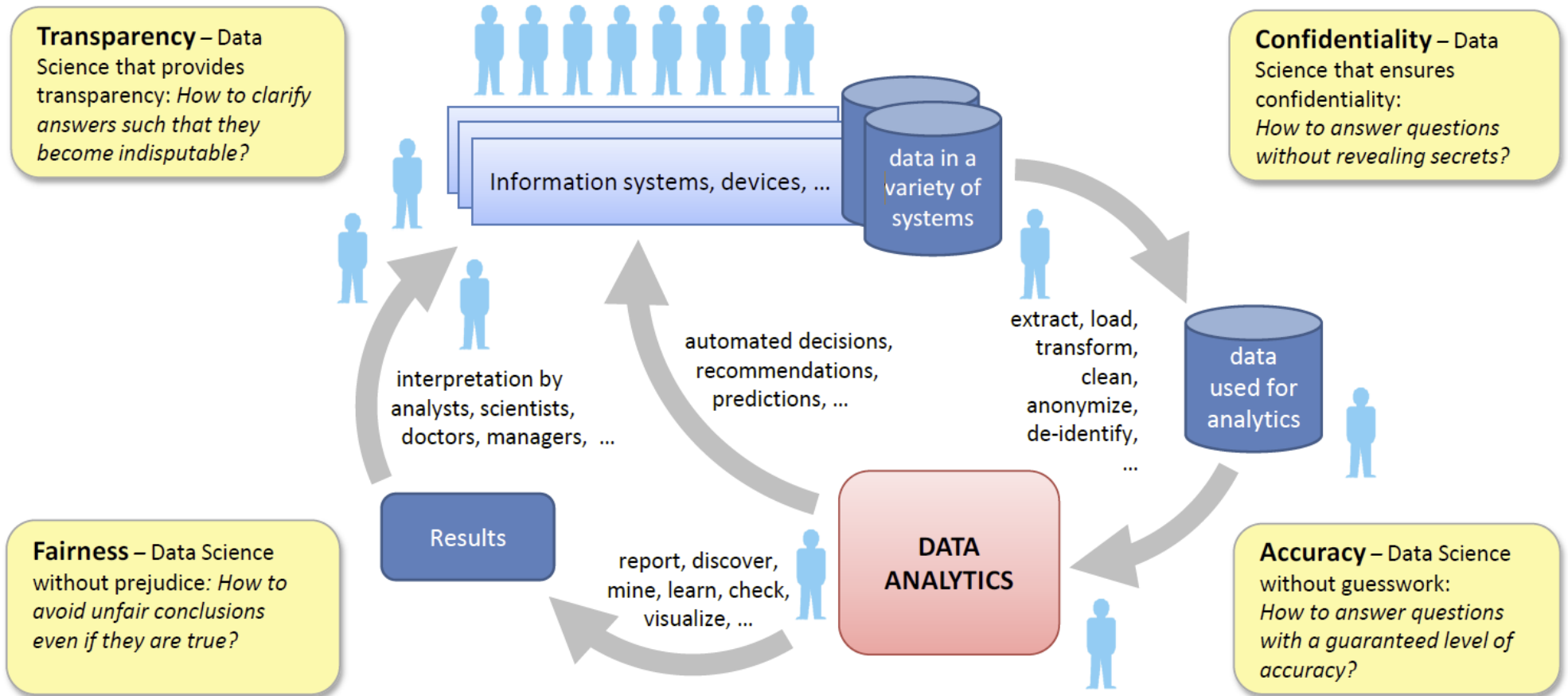
Example: word embeddings

- Reasoning with language:
 - > Man:Woman as Boy:Girl
 - > Ottawa:Canada as Nairobi:Kenya
 - > Big:Bigger as Tall:Taller
 - > Yen:Japan as Ruble:Russia
- But also:
 - > Man:Woman as King:Queen
 - > Man:Computer_Programmer as Woman:Homemaker
 - > Father:Doctor as Mother:Nurse

Impact of wrong data

- Police, security, intelligence – screening suspects
- Judges – deciding on pre-trial period of suspects
- eCommerce – cookie-based price adjustments
- Education – giving a (negative) study advice
- Medical diagnostics, personalized medicine, ...
- Mortgages, car insurances, CV screening, jobs, salaries, funding decisions, ...
- ...

Responsible data analytics



Obtaining and combining domain specific corpora for creating word vectors

Tim vor der Brück

Obtaining and combining domain specific corpora for creating word vectors

Task we deal with: Creating word vectors using word2vec to analyze user answers in an online contest

Issues

- How to obtain a large Swiss German corpus where the participants expressed themselves in youth language?
- How to combine the domain specific corpus with a large general one like Wikipedia as input for Word2Vec?

How to obtain the domain-specific corpus

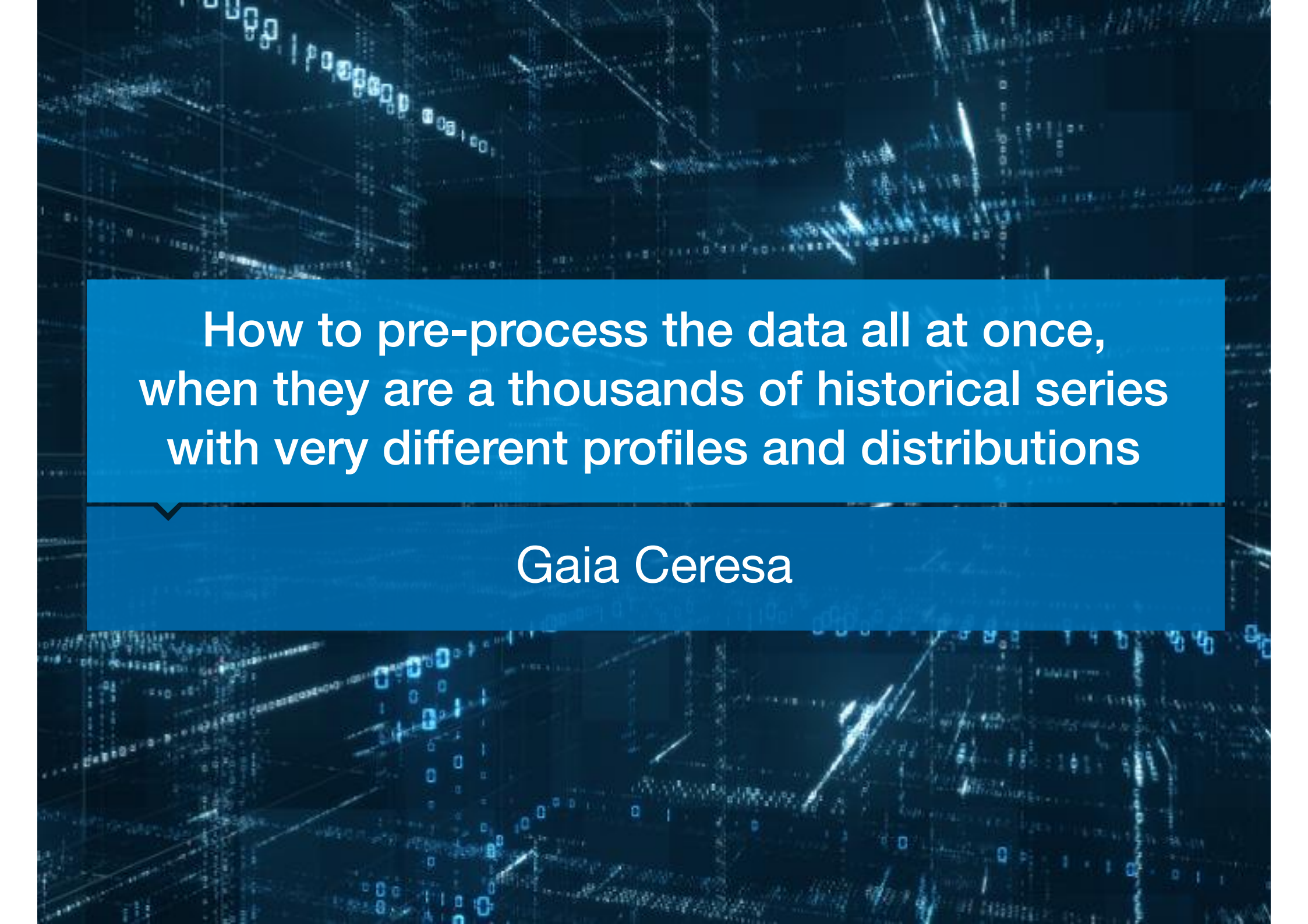
Possible solutions:

- Crawling the web (how to come up with suitable keywords?)
- Access domain-specific repositories (where to find good repositories?)
- Create artificial data employing a Generative Adversarial Network (GAN)
- Additional suggestions?

How to combine the domain-specific text corpus with a large general corpus before applying Word2Vec

Possible solutions:

- Concatenation of the corpora
- Oversampled concatenation (assign the small domain-specific corpus more weight)
- Create two different set of word vectors and combine them afterwards
- Additional suggestions?



How to pre-process the data all at once,
when they are a thousands of historical series
with very different profiles and distributions

Gaia Ceresa



The Seventh International Conference on
Data Analytics
DATA ANALYTICS 2018
November 18 to 22, 2018 - Athens, Greece



**How to pre-process the data all at once,
when they are a thousands of historical series
with very different profiles and distributions.**

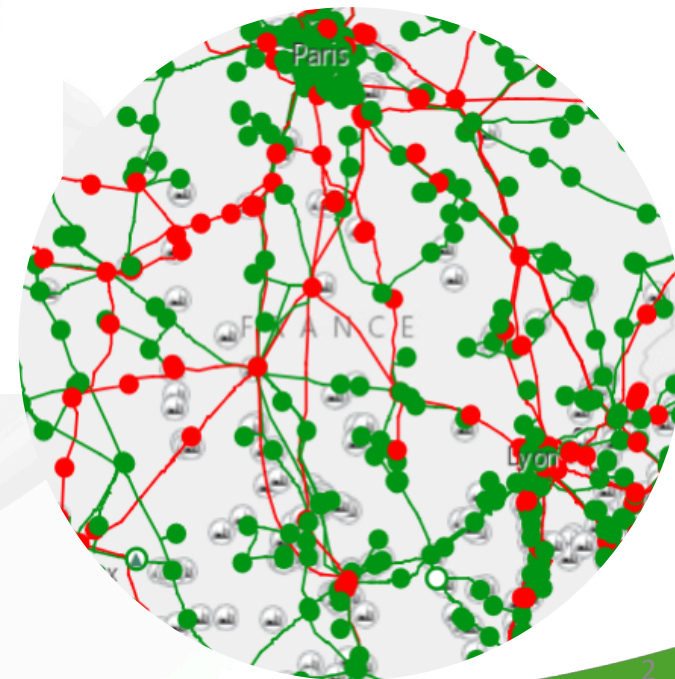
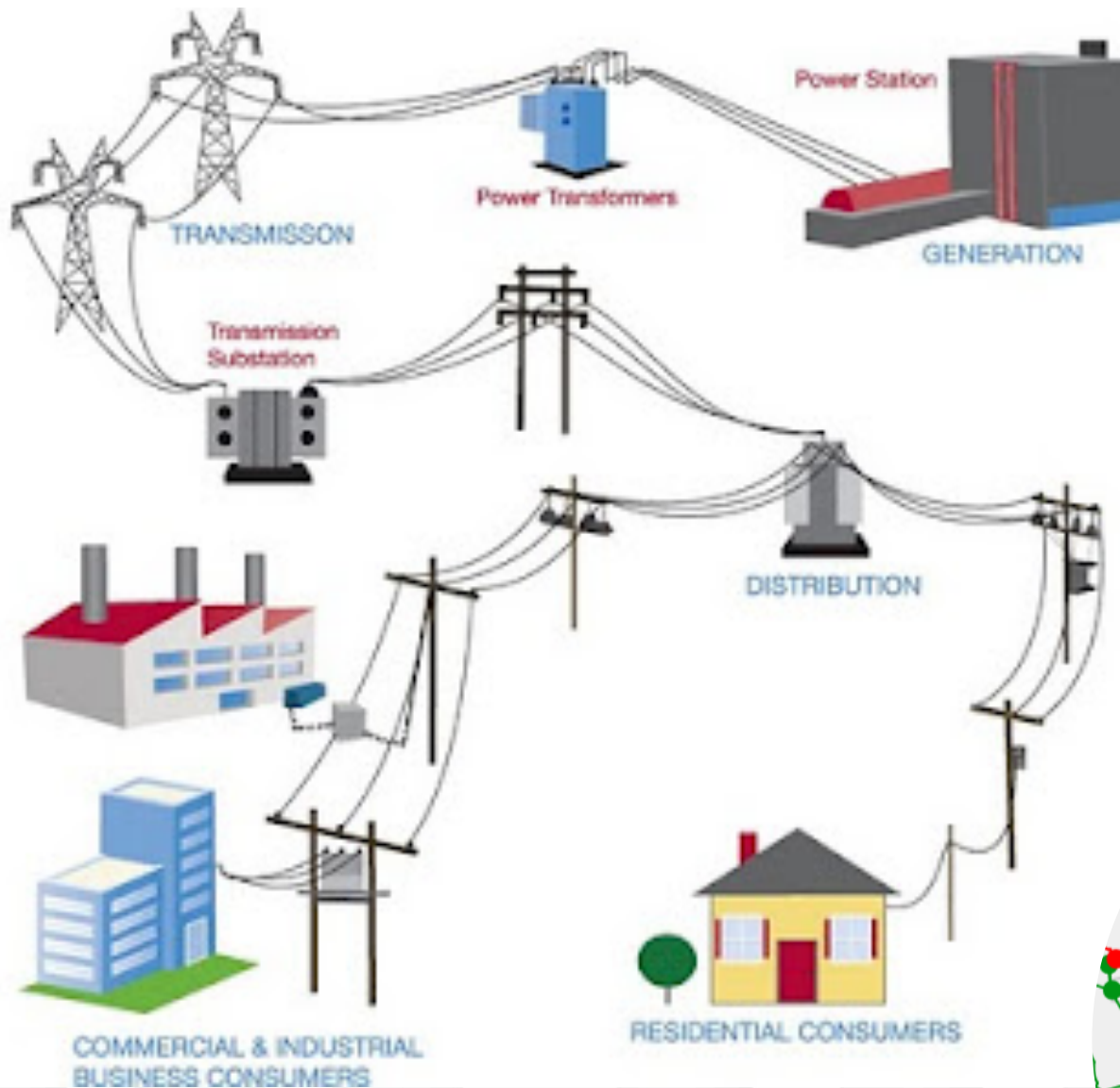
Gaia Ceresa

RSE S.p.A. - Ricerca sul Sistema Energetico, Milano, Italy

Panel on Advances in Data Processing
Approaches for Identifying/Selecting the Right Data

NexTech 2018, November 18-22, 2018 - Athens, Greece

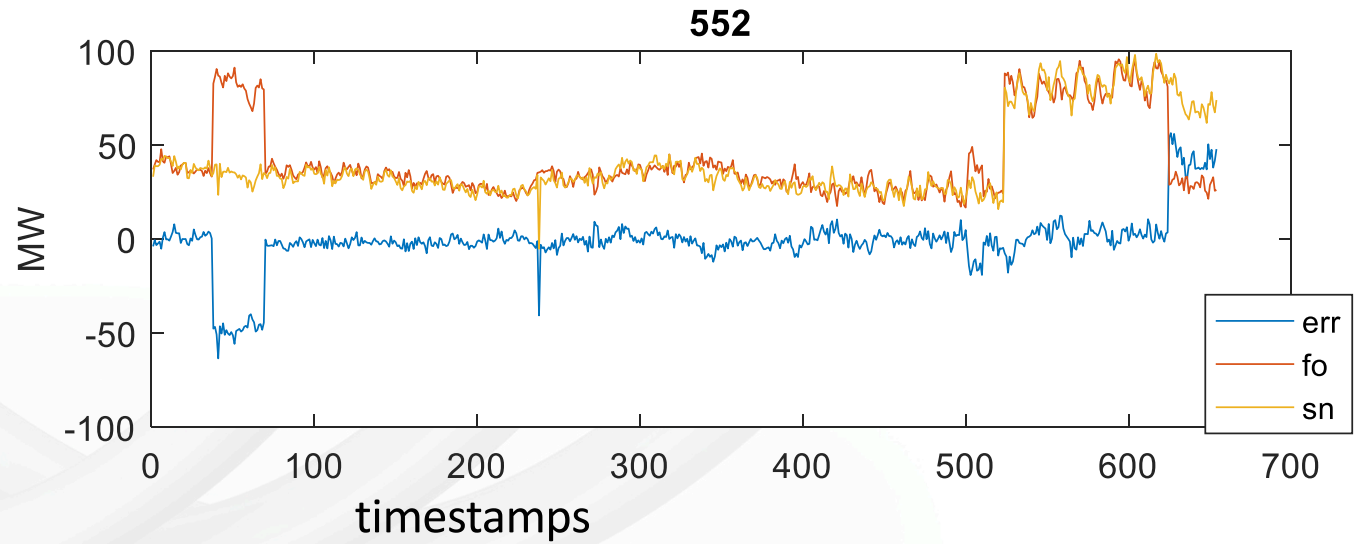
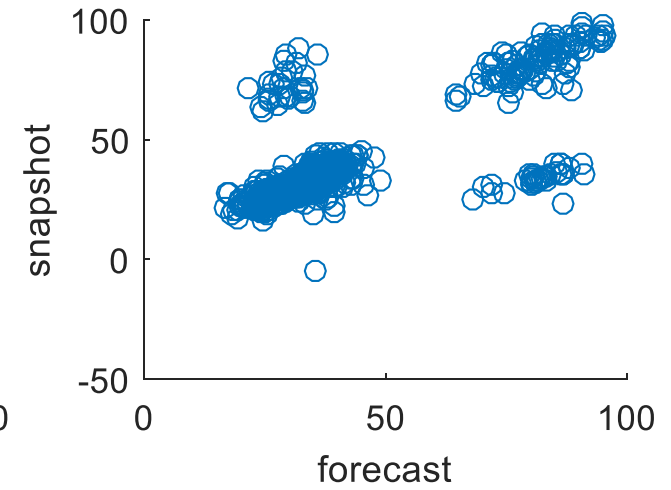
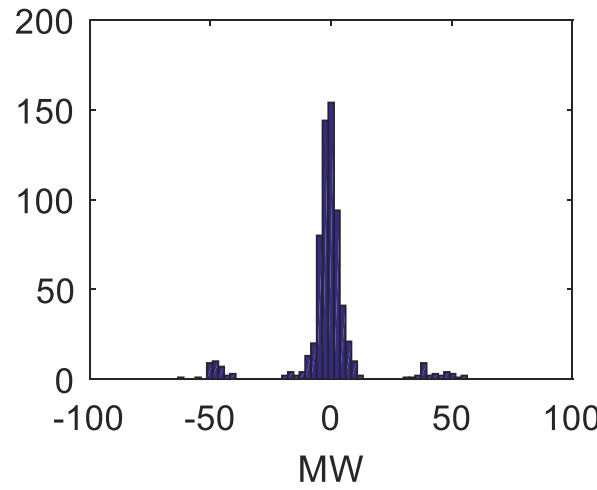
Electrical Power Grid HV and EHV



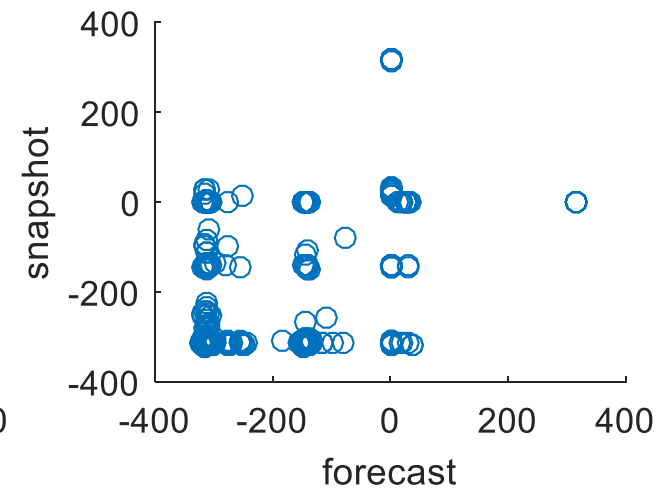
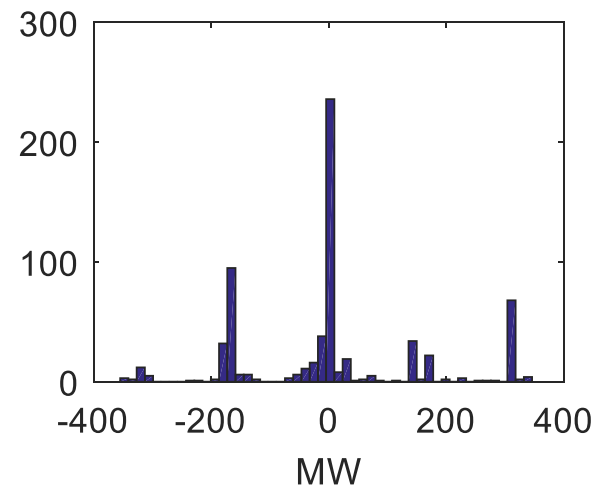
Tools:
Planning/operation
Expansion of the grid
Security assessment

Row input data example

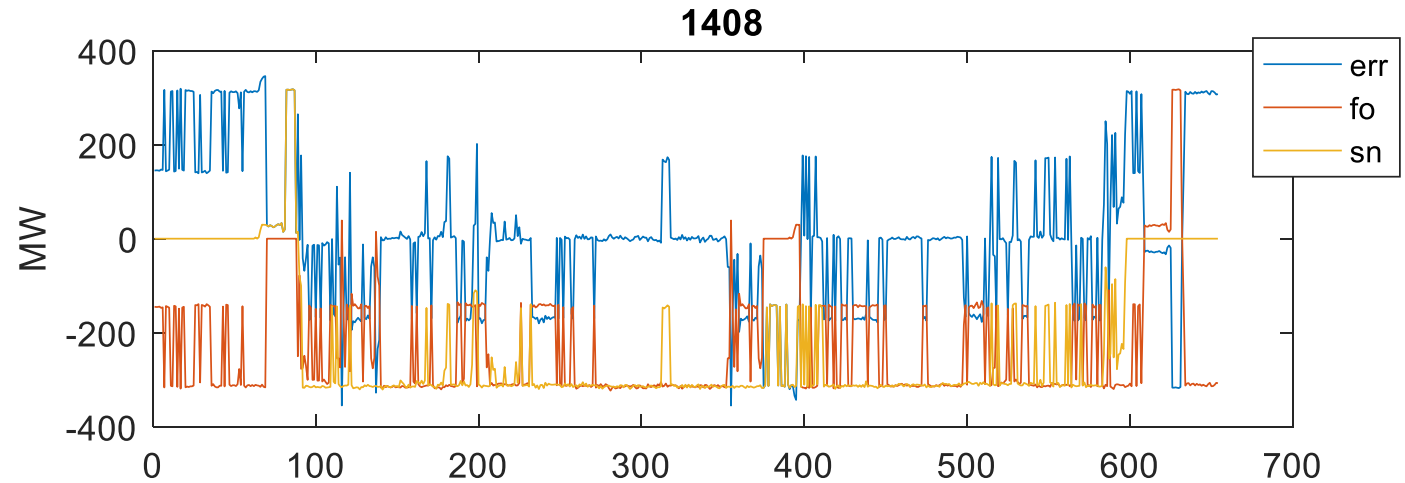
Transformer



Row input data example

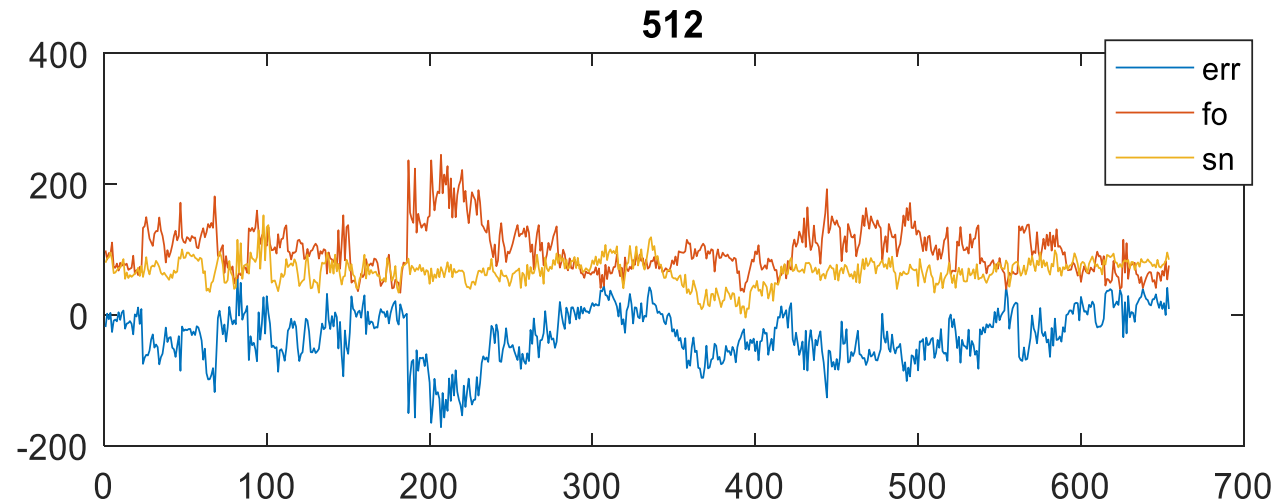
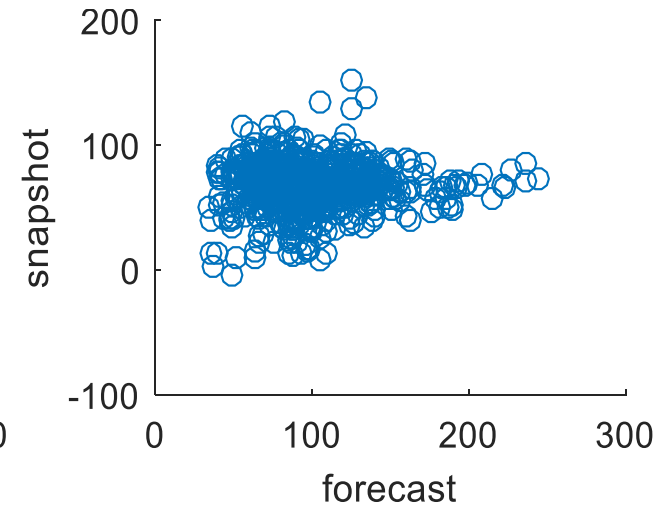
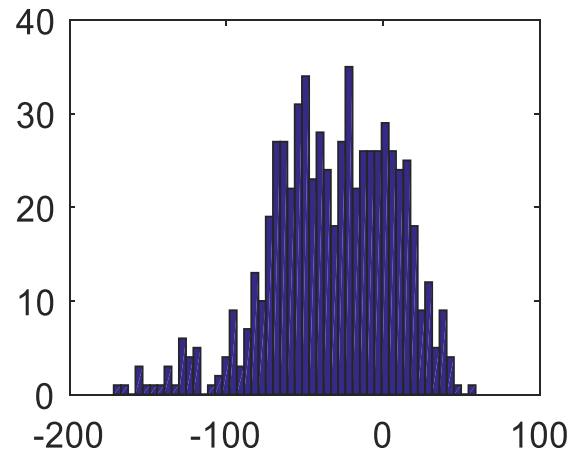


Transformer



Row input data example

Industrial Client



Significant series?

Missing Values and Zeros

Missing Values

- If more than 30% → variable removal
- If less than 30% and scattered in the series → replacement
- If less than 30% and concentrated in a block → replacement or removal

Many 0:

- Error of the measurement system → removal
- Real production → analysis

So:

If 0 more than 70% → removal

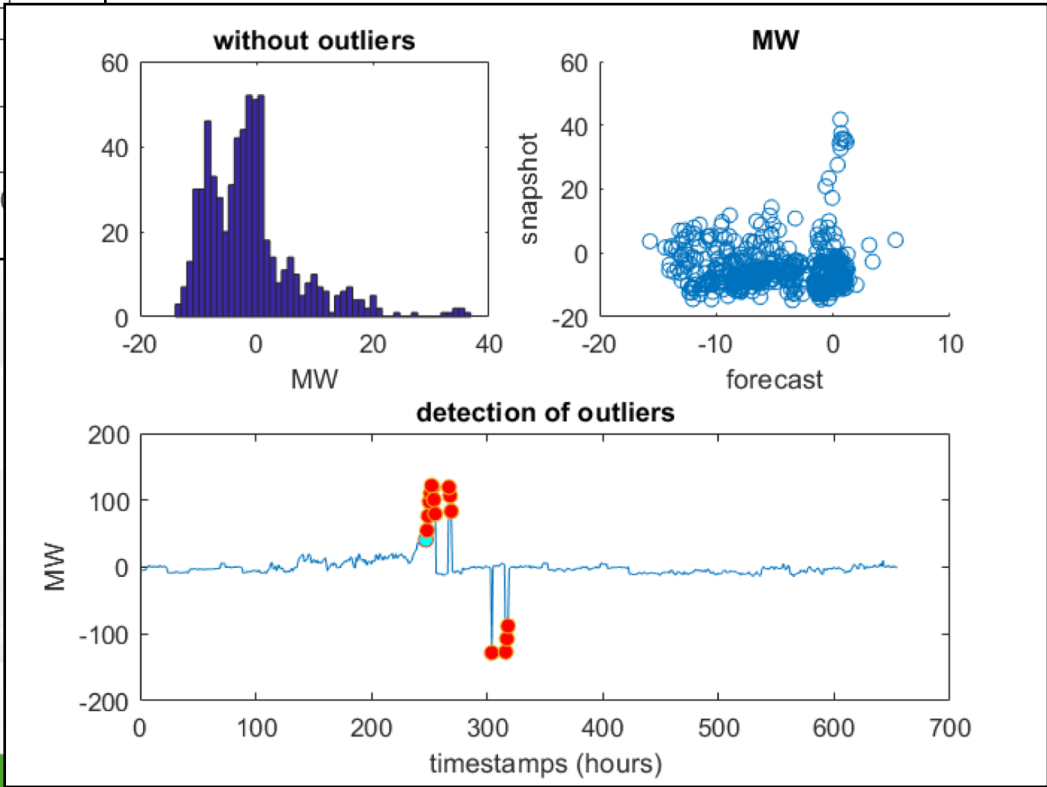
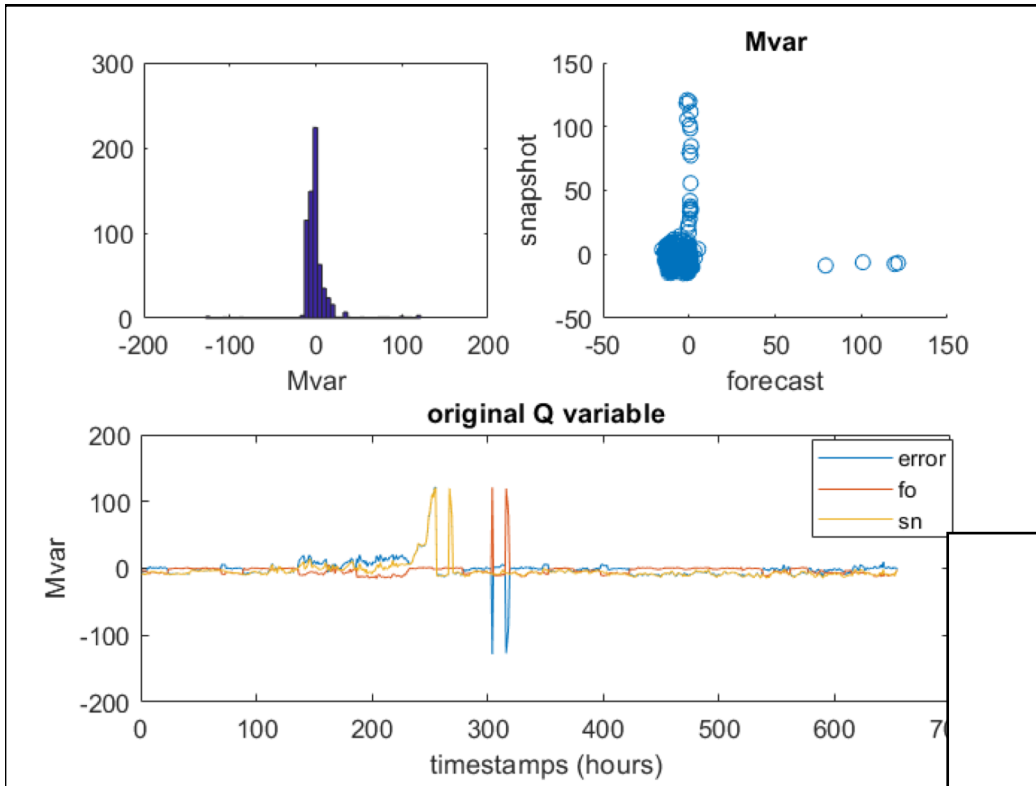
If 0 less than 70% → analysis o the variable

(it gives a bad behaviour if used in a training/testing set)

Outliers detection

Distribution transformer

Histogram, scatterplot, profile
of the raw variable



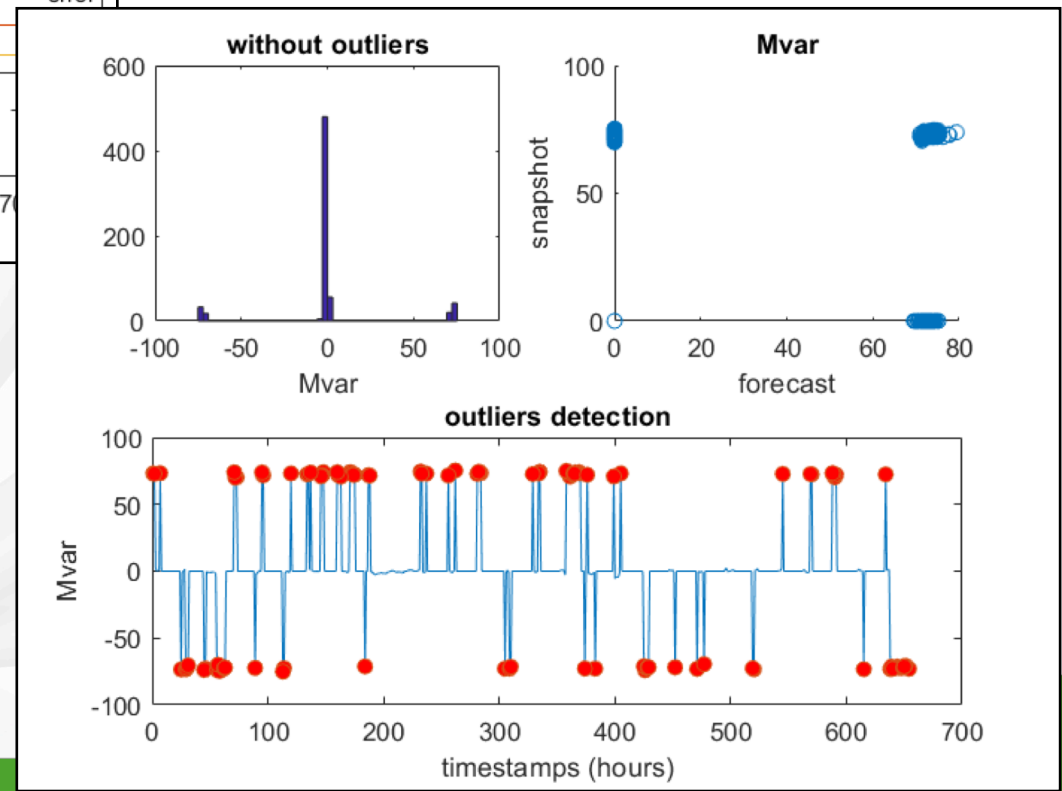
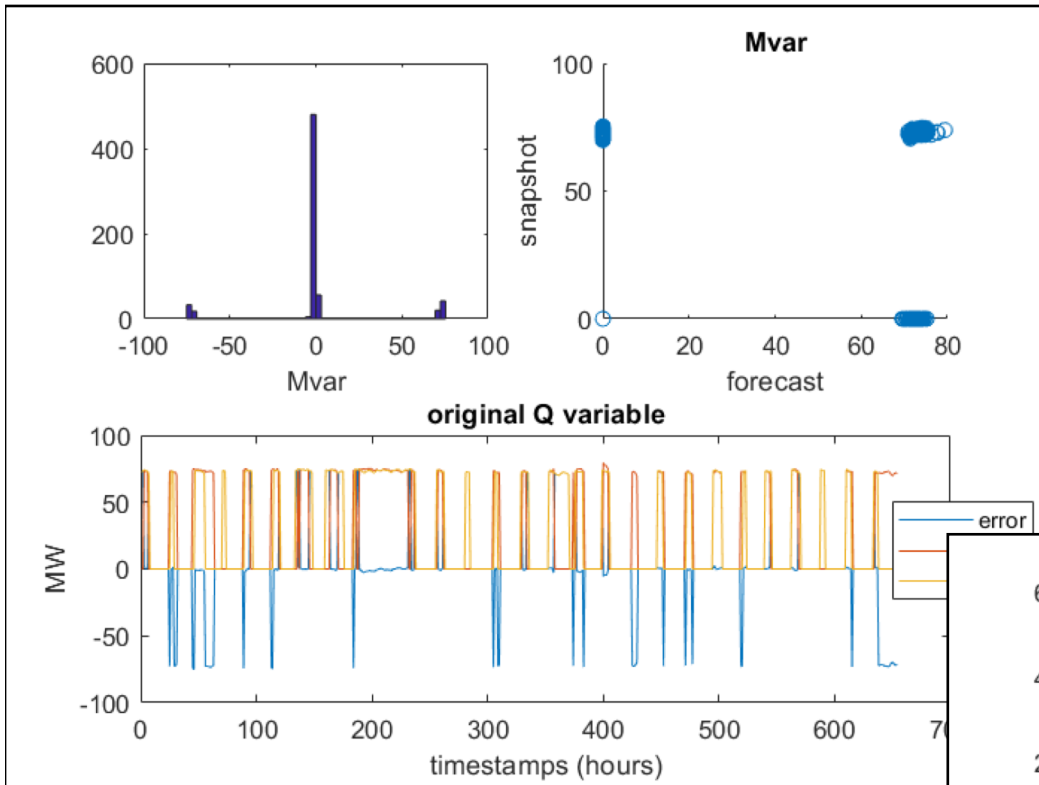
Detected outliers in the profile,
Histogram and scatterplot without
outliers



Outliers detection

Transformer to another distribution transformer

Histogram, scatterplot, profile of the raw variable



Detected outliers in the profile,
Histogram and scatterplot without
outliers



Outliers detection method

- **Outliers detection**

- a) Method based on the **Chebyshev inequality** $P(|X - \mu| \geq n\sigma) < \frac{1}{n^2}$.
Outliers stay out of the interval

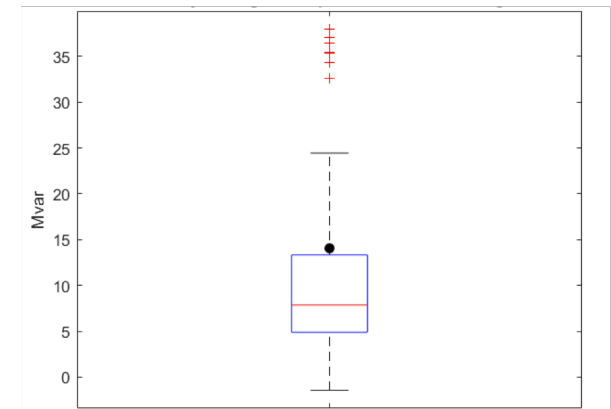
$$[\mu - n\sigma, \mu + n\sigma], n = 3$$

- b) Method based on **Quartiles**. Outliers stay out of the interval


$$[Q1 - n(Q3 - Q1), Q3 + n(Q3 - Q1)], n = 3$$

and pass the MAD test $\frac{|X_i - \text{median}(X_j)|}{\text{median}(|X_i - \text{median}(X_j)|)} > 5$

The largest set is selected, named *OUTL*.



- **Outliers elimination**

$|OUTL| \leq 7\%N$  $\begin{cases} \text{yes} & \rightarrow \text{removing outliers} \rightarrow \text{missing} \\ \text{no} & \rightarrow \text{preserving outliers} \end{cases}$

G. Ceresa, A. Pitto, D. Cirio, N.Omont

“Algorithm for Automatic Description of Historical Series of Forecast Error in Electrical Power Grid”.

Proceedings of 4TH Conference of the International Society for Nonparametric Statistics ISNPS 2018. Springer. Available in 2019.

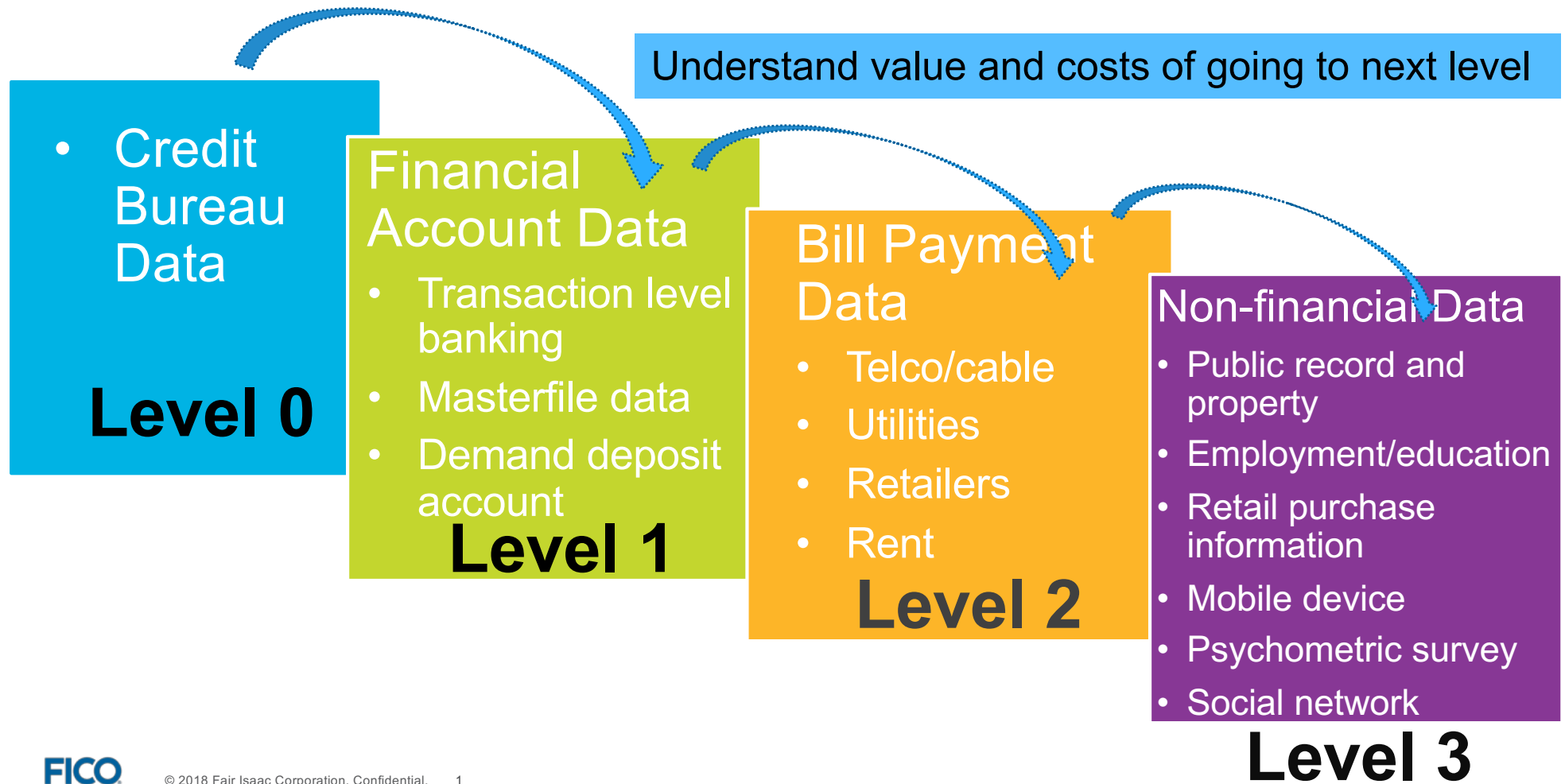
gaia.ceresa@rse-web.it



Data selection guidelines for credit scoring

Gerald Fahner

Hierarchy of Data Sources for Credit Scoring



FICO's Alternative Data Selection Guidelines

Regulatory Compliance	The data source must comply with all regulations governing consumer credit evaluation.
Depth of Information	Data sources that are deeper and contain greater detail are often of greater value.
Scope & Consistency of Coverage	A stable database covering a broad percentage of consumers can be favorable.
Accuracy	How reliable is the data? How is it reported? Is it self-reported? Are there verification processes in place?
Predictiveness	The data should predict future consumer repayment behavior.
Orthogonality	Useful data sources should be supplemental or complementary to what's captured by other data sources.



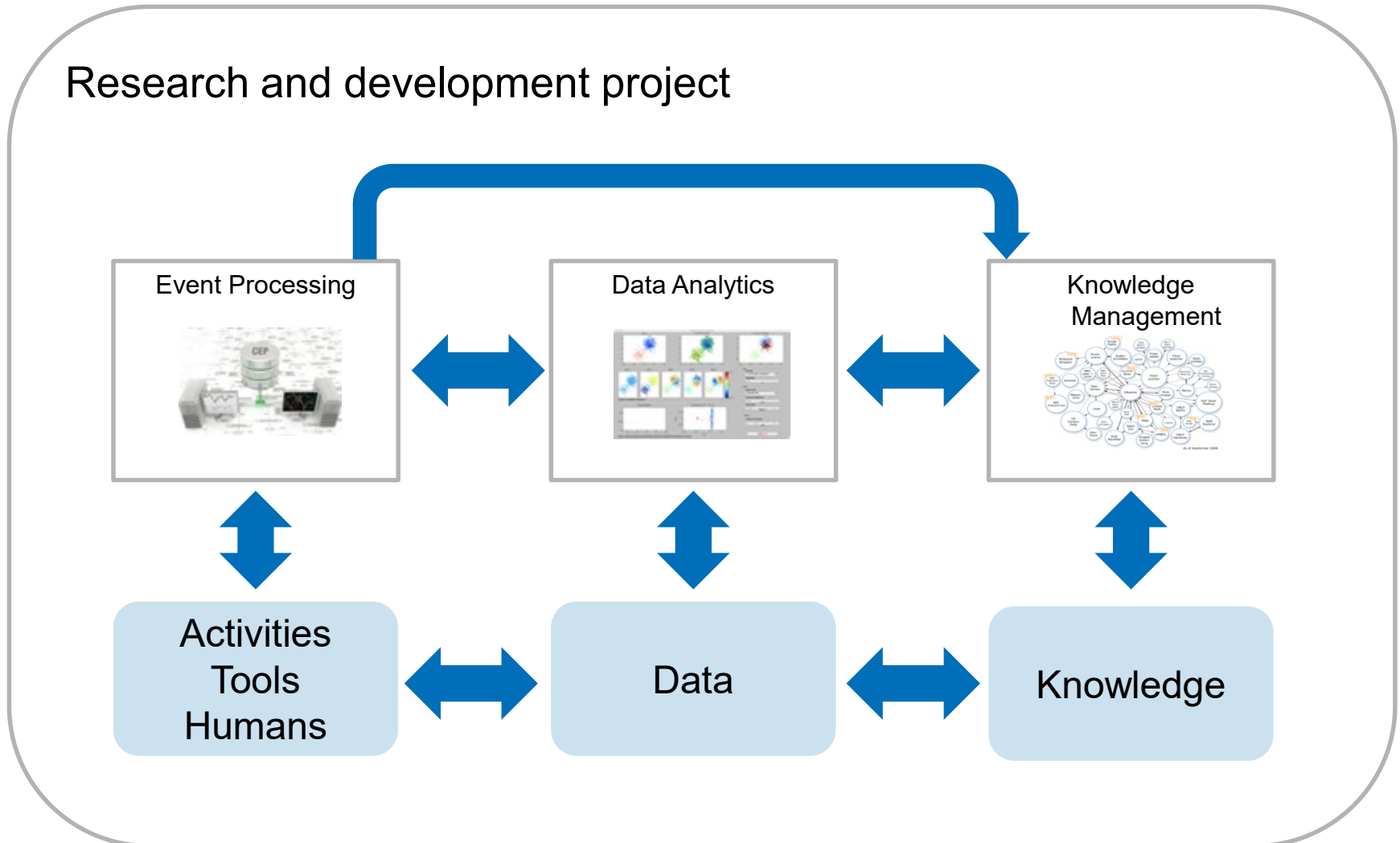
**Combining different approaches enabling a
consistent workflow for data analytics
including pre- and post-processing**

Gregor Grambow

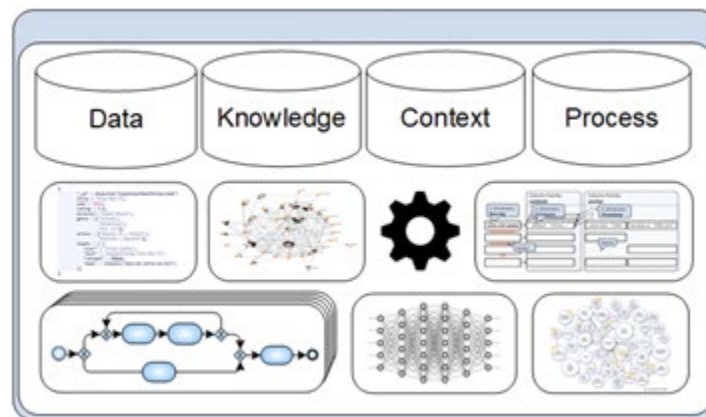
Combining different approaches enabling a consistent workflow for data analytics including pre- and post-processing

Gregor Grambow
Computer Science Dept.
Aalen University

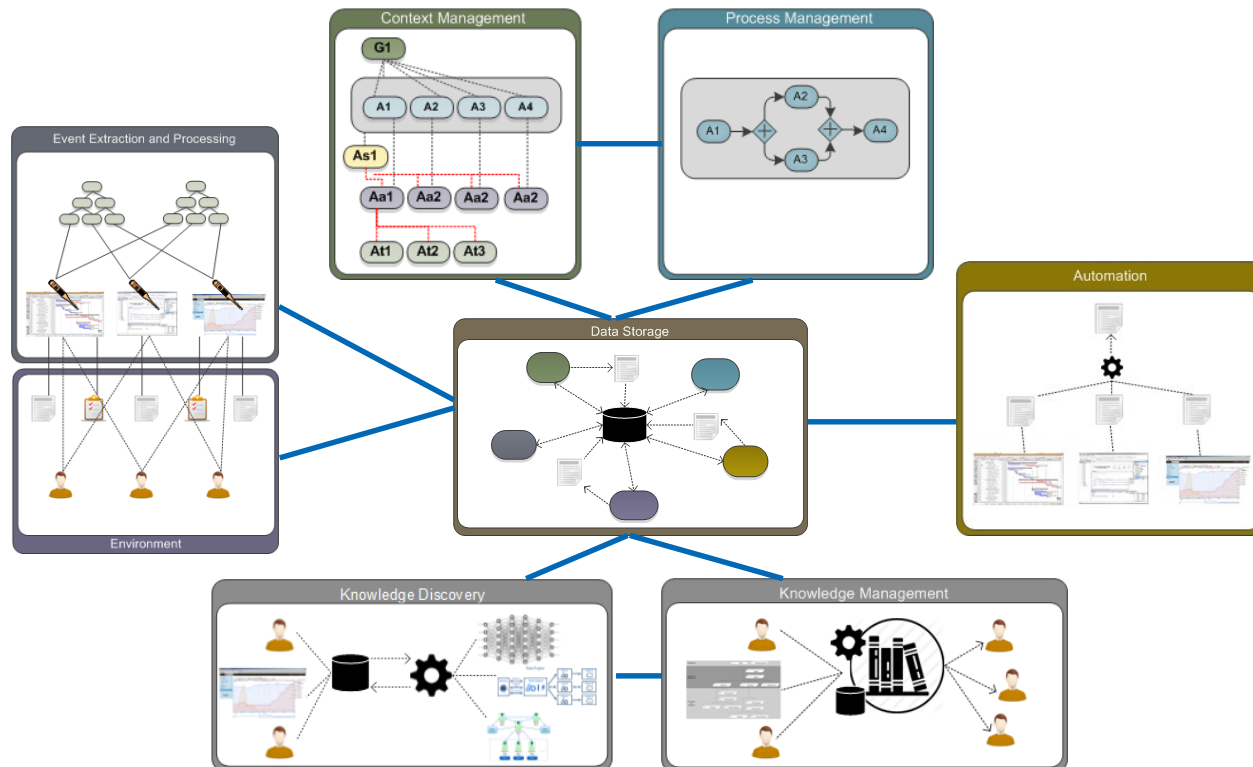
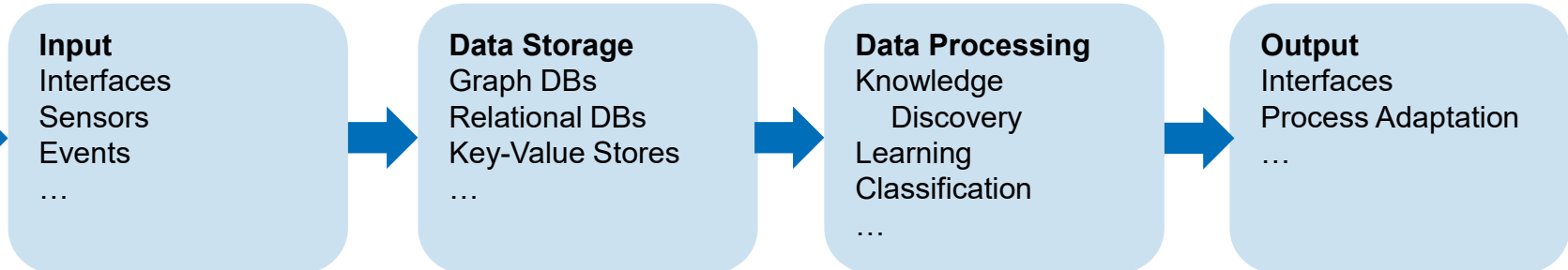
Data Science – holisitcally?



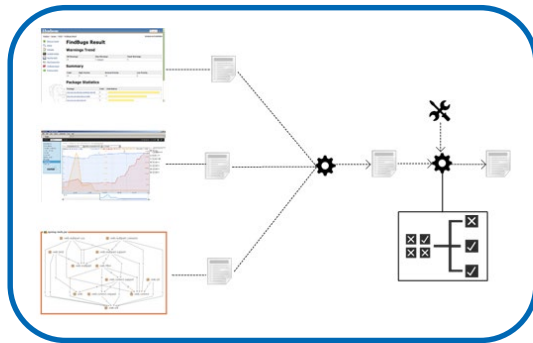
Research Activities



Research Activities



Research Framework



mongoDB

```

{
  "_id" : ObjectId("51a5d316d70bffe74ecc940")
  "title" : "Iron Man 3",
  "year" : 2013,
  "rating" : 7.6,
  "director" : "Shane Black",
  "genre" : ["action",
    "Adventure",
    "Sci-Fi"],
  "actors" : ["Downey Jr.", "Robert",
    "Paltrow", "Gwyneth"],
  "tweets" : [ {
    "user" : "Franz Kafka",
    "text" : "knows watching Iron Man 3",
    "retweet" : false,
    "date" : ISODate("2013-05-29T13:15:51Z")
  } ]
}
    
```

RAVENDB

Cassandra

HBASE

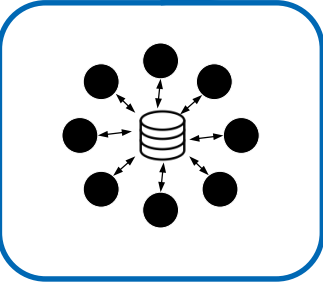
1. Dimension: Row Key
com.cnn.www

2. Dimension: CF Column
content: "chrisb-"

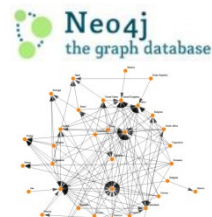
3. Dimension: Timestamp
cnn1.com: "CNN" my.look.co: "CNN.com"

Input
 GUI Interfaces
 Technical Interfaces
 Sensors
 Processes
 Event Processing
 Data Transformation
 ...

Data Storage
 Graph DBs
 Document DBs
 Key-Value-Stores
 Wide-Column-Stores
 Relational DBs
 ...



Neo4j
the graph database



InfiniteGraph
Powered by Objectivity

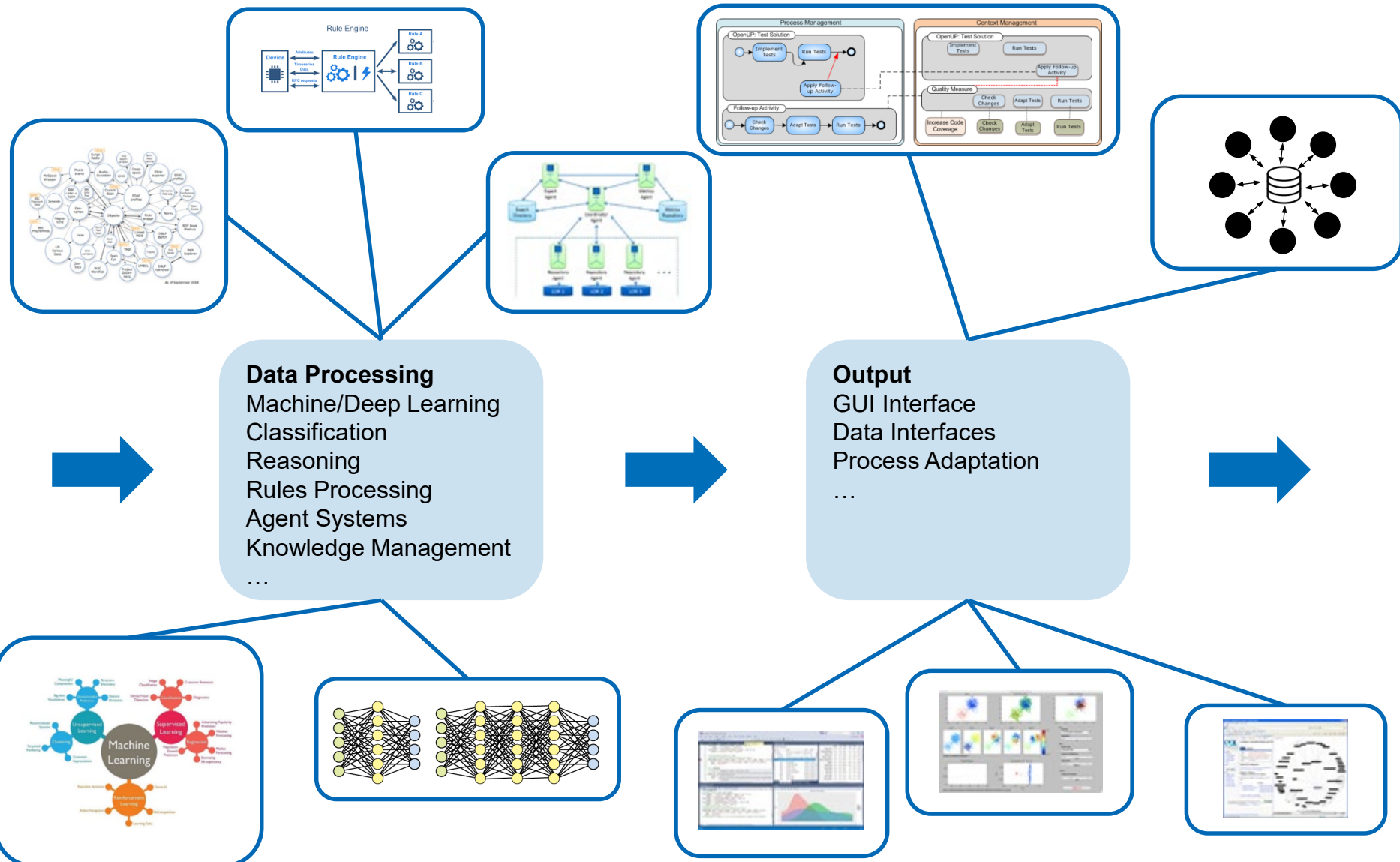
redis **riak**

- web:index → "html<head>..."
- users:2:friends → [23, 76, 233, 11]
- users:2:inbox → [234, 3466, 86, 55]
- users:2:settings → Theme → "dark", cookies → "false"
- top-posters → 466 → "2", 344 → "16"
- users:2:notifs → ["event: 'comment posted', time : ..."]

ORACLE **Microsoft SQL Server**



Research Framework



Special Topic: Parallelized Machine Learning?

- Hadoop provides capabilities to do machine learning in the cluster
 - SparkMLlib
 - Processes Keras / TensorFlow models
- Various advantages
 - Scalability and especially elasticity
 - Infrastructure for
 - Distributed data storage
 - Data ingestion
 - Various types of data processing
 - Query facilities