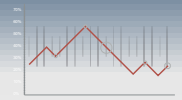
Data Science
The Big Data challenge

TANIA CERQUITELLI

Big data hype?

Emergency management

EARTH OBSERVATIONS

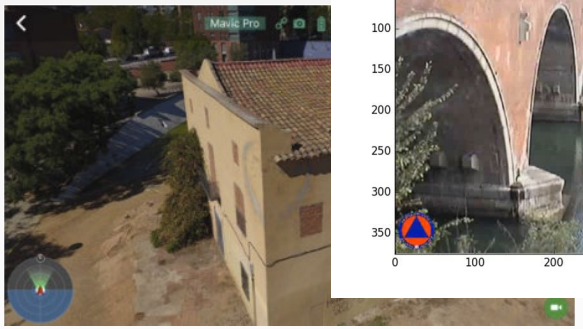


HISTORICAL DATA

SEASONAL
WEATHER FORECAST

SOCIAL MEDIA
DATA STREAMS

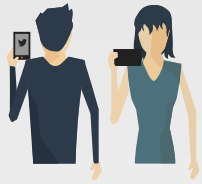
UNMANNED AERIAL VEHICLES



Improving Resilience to Emergencies
Through Advanced Cyber Technologies

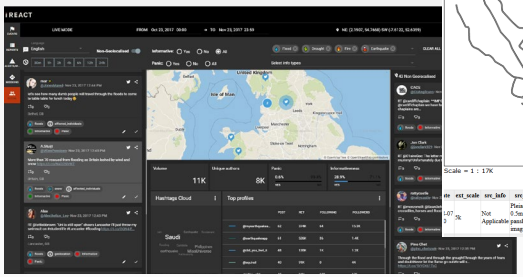
iREACT

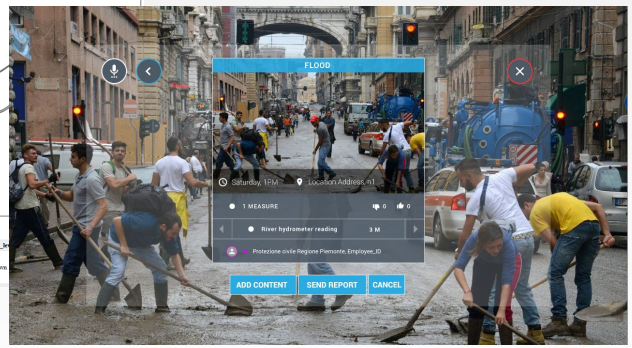
Emergency management



CITIZENS

FIRST RESPONDERS AND DECISION MAKERS





Improving Resilience to Emergencies Through Advanced Cyber Technologies

iREACT

User engagement

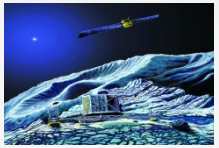
2005

2022

Who generates big data?

- User Generated Content (Web & Mobile)
  - E.g., Facebook, Instagram, Yelp, TripAdvisor, Twitter, YouTube

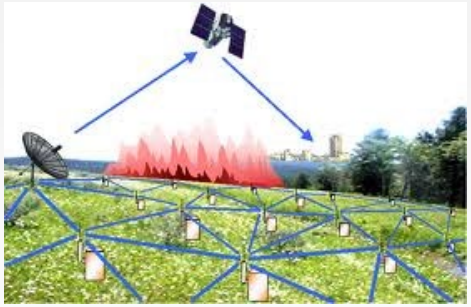
- Health and scientific computing



Who generates big data?

- Log files
  - Web server log files, machine syslog files

- Internet Of Things
  - Sensor networks, RFID, smart meters



What is big data?

- Many different definitions

"Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"

What is big data?

- Many different definitions

*"Data whose **scale**, **diversity** and **complexity** require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it"*

What is big data?

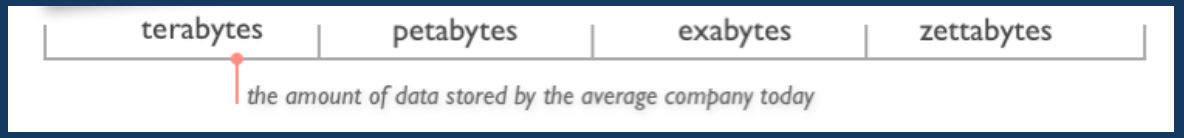
- Many different definitions

*"Data whose scale, diversity and complexity require new **architectures**, **techniques**, **algorithms** and **analytics** to manage it and extract value and hidden knowledge from it"*

What is big data?

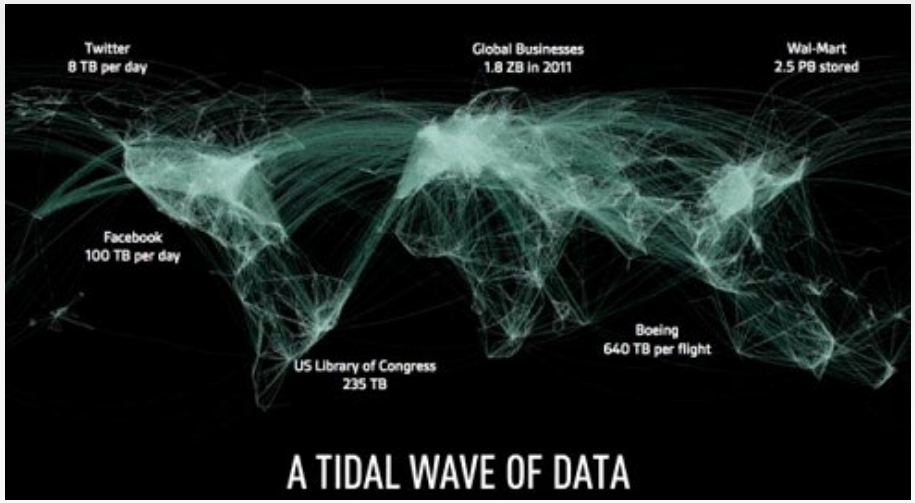
- Many different definitions

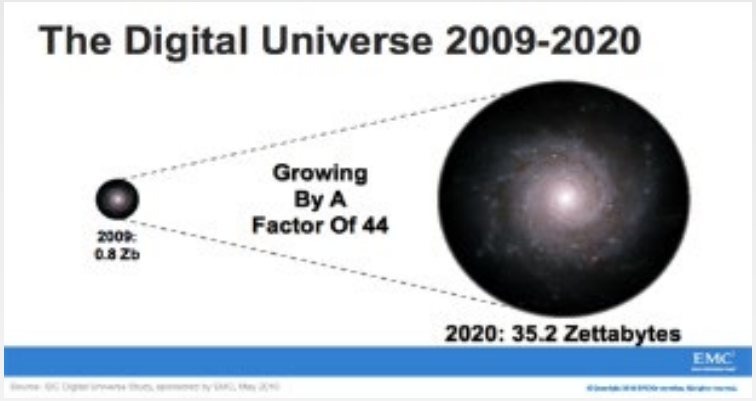
*"Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract **value** and hidden **knowledge** from it"*

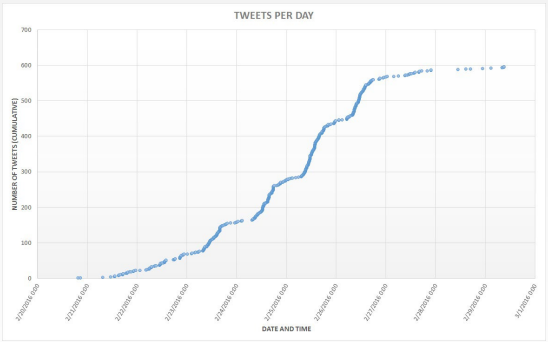


The Vs of big data: Volume

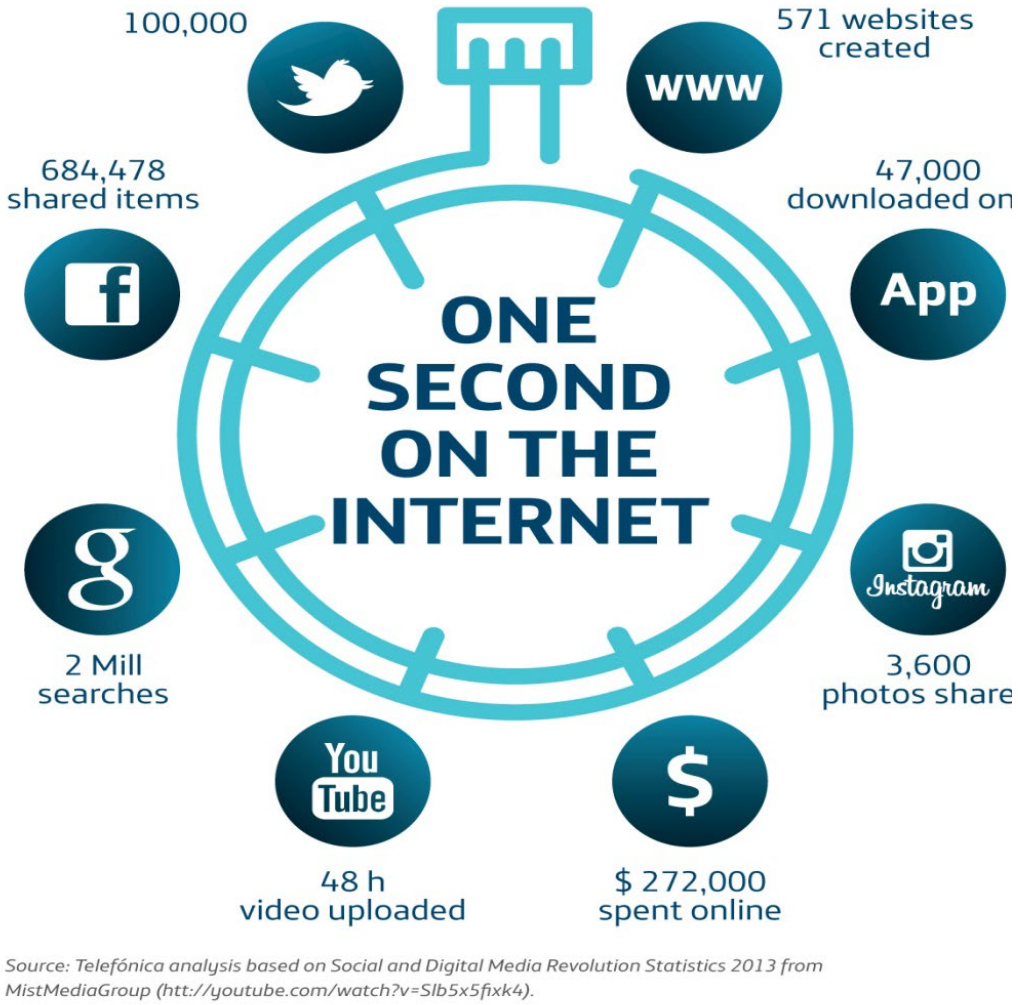
- Data volume increases exponentially over time
- 44x increase from 2009 to 2020
  - Digital data 35 ZB in 2020







On the Internet…



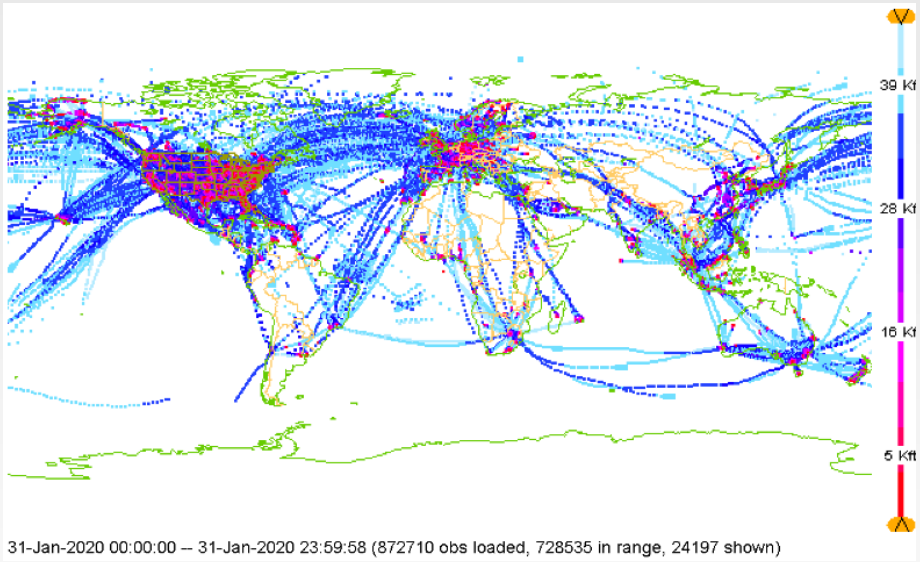
Source: Telefónica analysis based on Social and Digital Media Revolution Statistics 2013 from MistMediaGroup (htt://youtube.com/watch?v=Slb5x5fixk4).

http://www.internetlivestats.com/

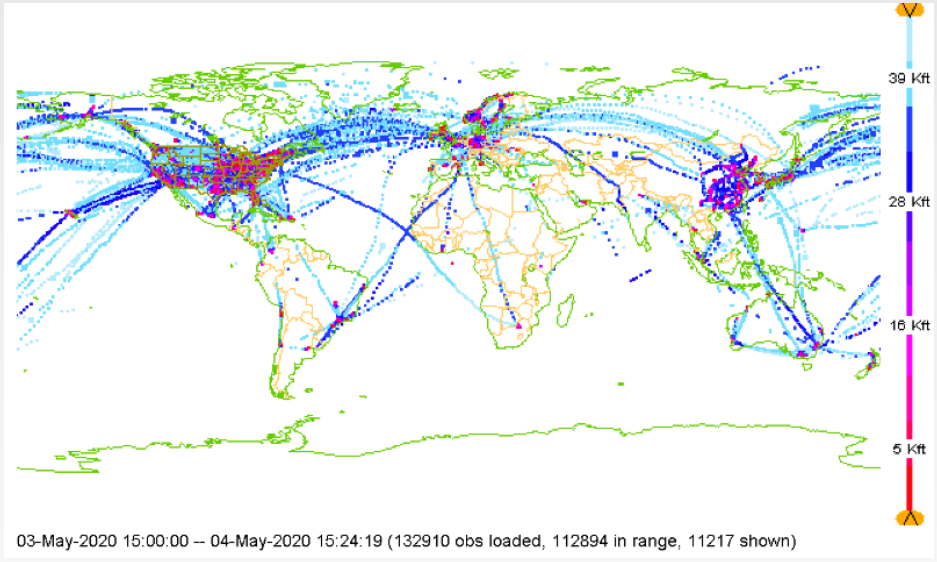
Weather forecast

WORLD METEOROLOGICAL ORGANIZATION

January 2020

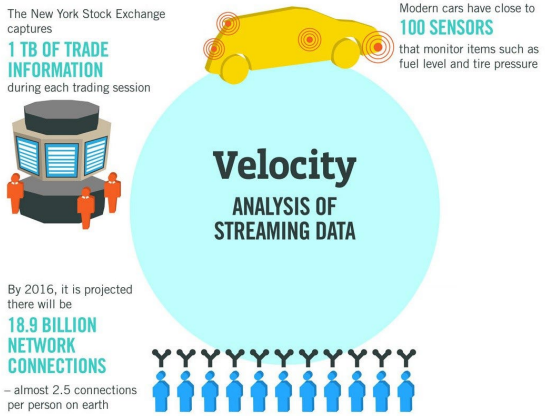


May 2020

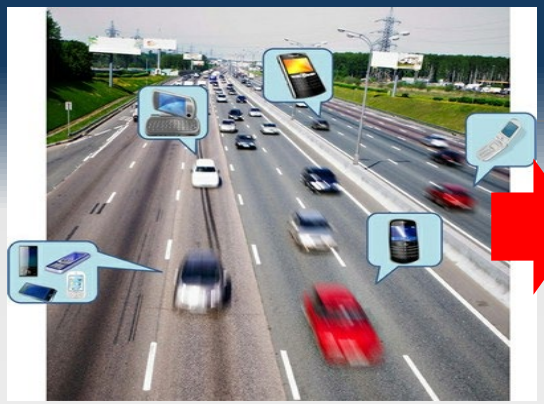


The Vs of big data: Velocity

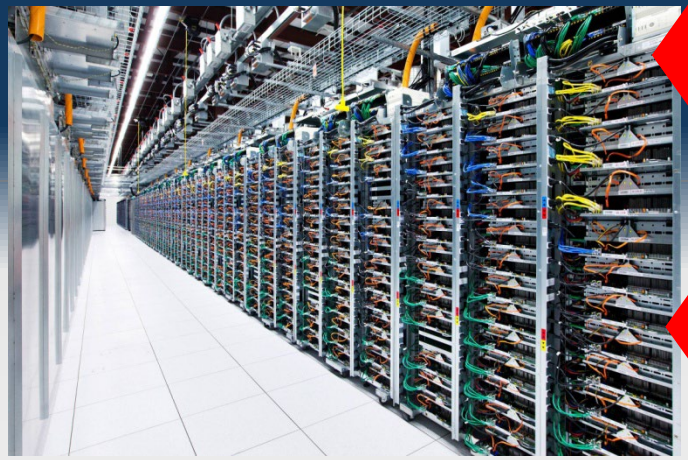
- Fast data generation rate
  - Streaming data
- Very fast data processing to ensure timeliness



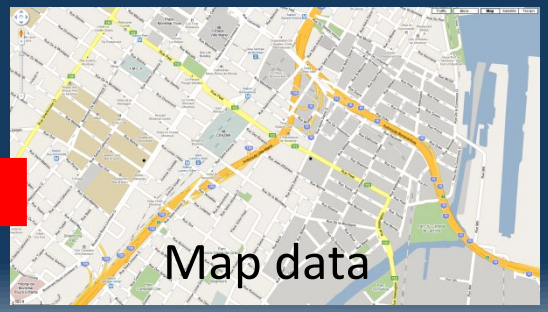
(Near) Real time processing

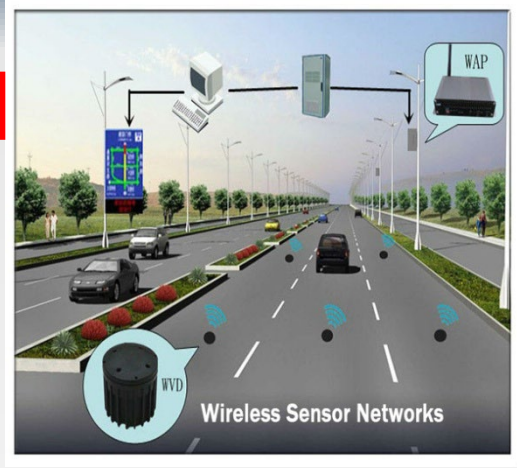


Crowdsourcing

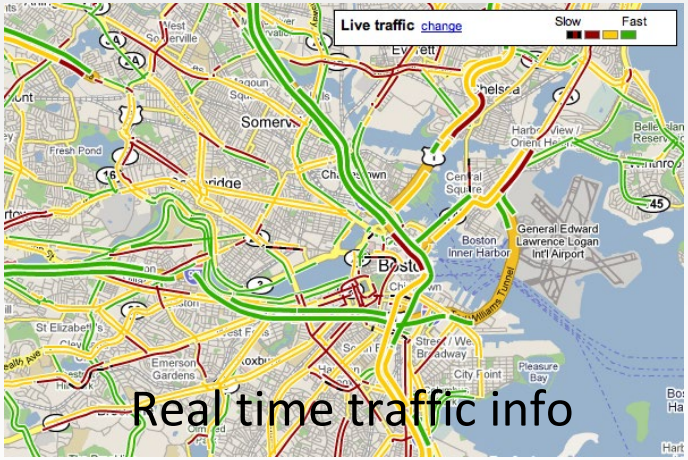


Computing



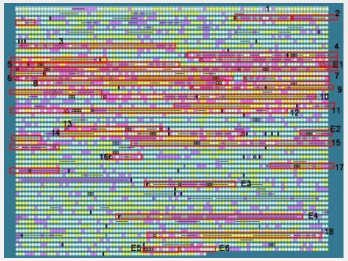


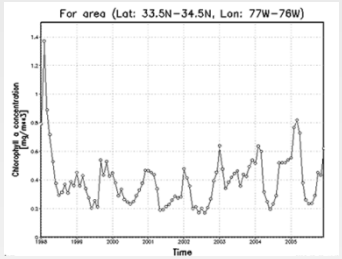
Sensing

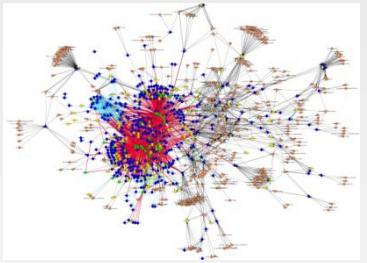


The Vs of big data: Variety

- Various formats, types and structures
  - Numerical data, image data, audio, video, text, time series







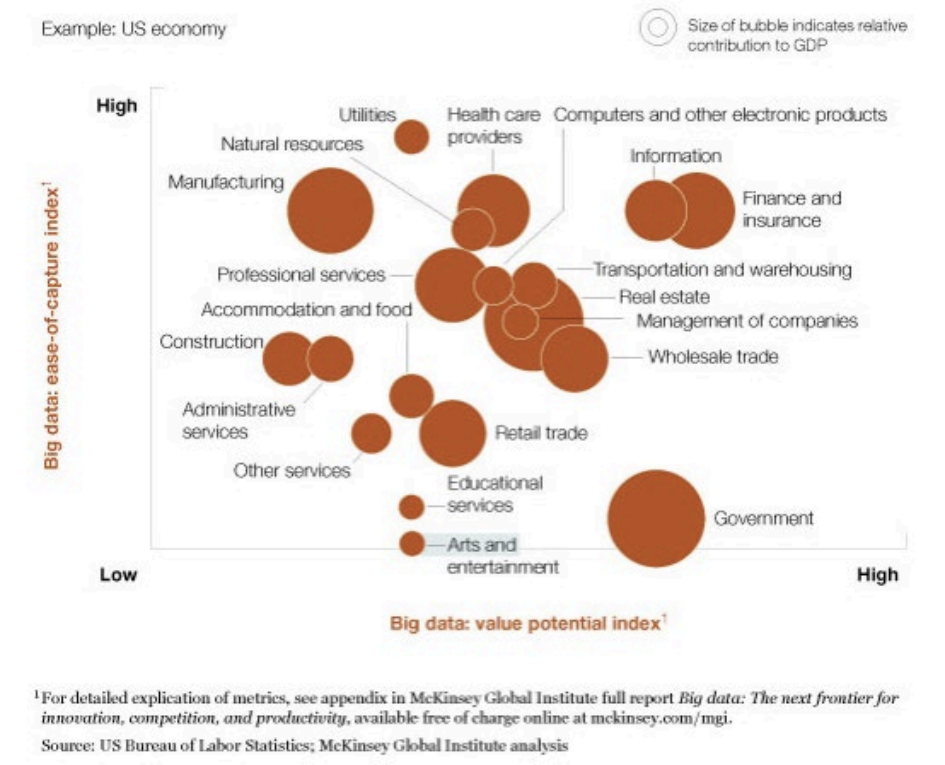
- A single application may generate many different formats

The Vs of big data: Veracity

- Data quality

The most important V: Value

- Translate data into business advantage



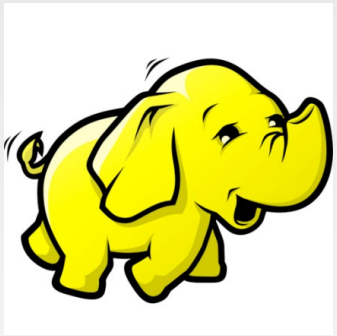
[1] For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Big data challenges

- Technology & infrastructure
  - New architectures, programming paradigms and techniques

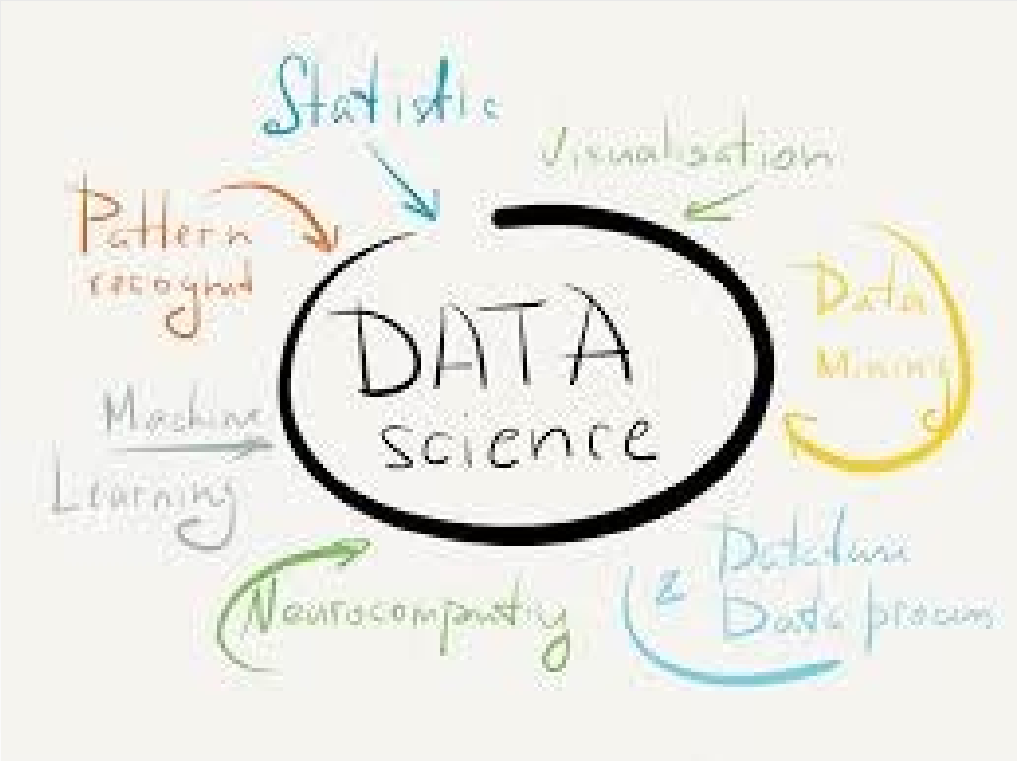
    Transfer the processing power to the data
  - Apache Hadoop/Spark ecosystem
- Data management & analysis
  - New emphasys on “data”



➡ ***Data science***

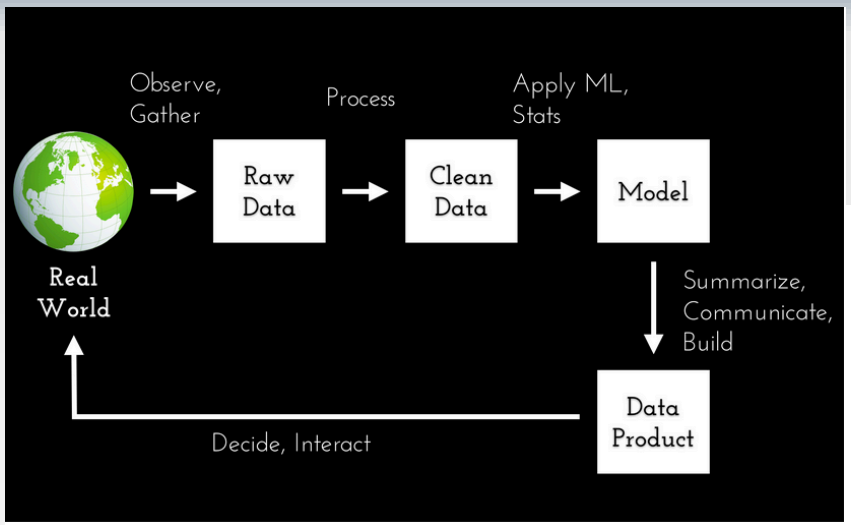
Data science

"Extracting meaning from very large quantities of data"



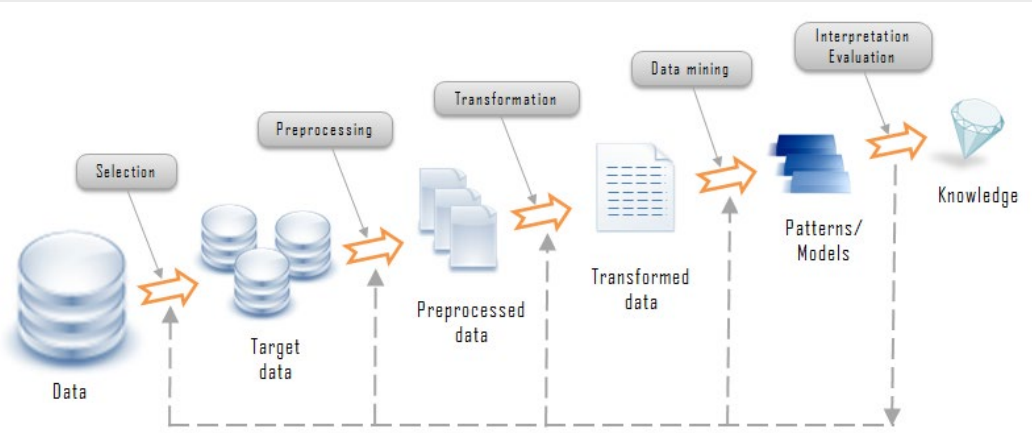
D.J. Patil coined the word *data scientist*

The data science process



AKA *KDD* process

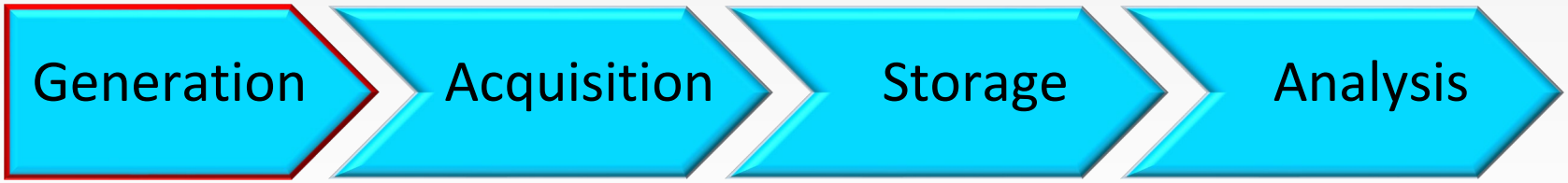
Knowledge Discovery in Databases



Generation → Acquisition → Storage → Analysis

Generation

- Passive recording
  - Typically structured data
  - Bank trading transactions, work hours, government sector archives
- Active generation
  - Semistructured or unstructured data
  - User-generated content, e.g., social networks
- Automatic production
  - Location-aware, context-dependent, highly mobile data
  - Sensor-based Internet-enabled devices (IoT)

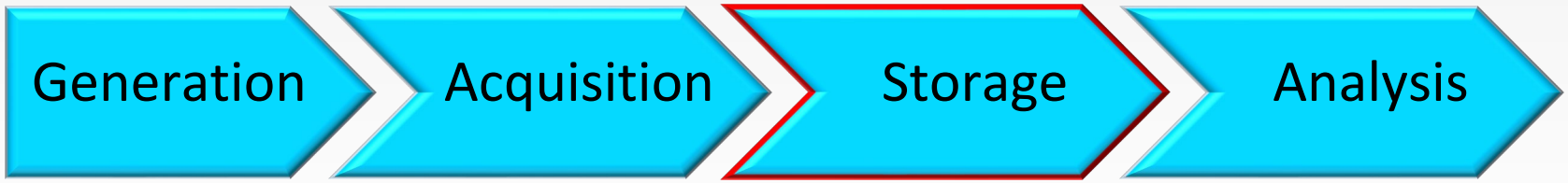


Acquisition

- Collection
  - Pull-based, e.g., web crawler
  - Push-based, e.g., video surveillance, click stream
- Transmission
  - Transfer to data center over high capacity links
- Preprocessing
  - Integration, cleaning, redundancy elimination

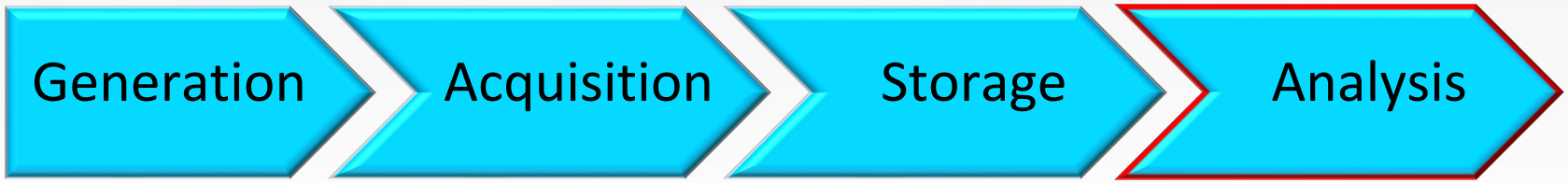
Storage

- Storage infrastructure
  - Storage technology, e.g., HDD, SSD
  - Networking architecture, e.g., DAS, NAS, SAN
- Data management
  - File systems (HDFS), key-value stores (Memcached), column-oriented databases (Cassandra), document databases (MongoDB)
- Programming models
  - Map reduce, stream processing, graph processing



Analysis

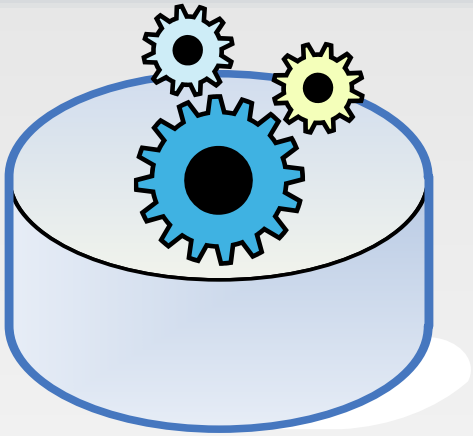
- Objectives
  - Descriptive analytics, predictive analytics, prescriptive analytics
- Methods
  - Statistical analysis, machine learning and data mining, text mining, network and graph data mining
  - Association analysis, classification and regression, clustering
- Diverse domains call for customized techniques



Data mining

- Non trivial extraction of
  - implicit
  - previously unknown
  - potentially useful

  information from available data



- Extraction is automatic
  - performed by appropriate algorithms
- Extracted information is represented by means of abstract models
  - denoted as *pattern*

Profiling: examples of data

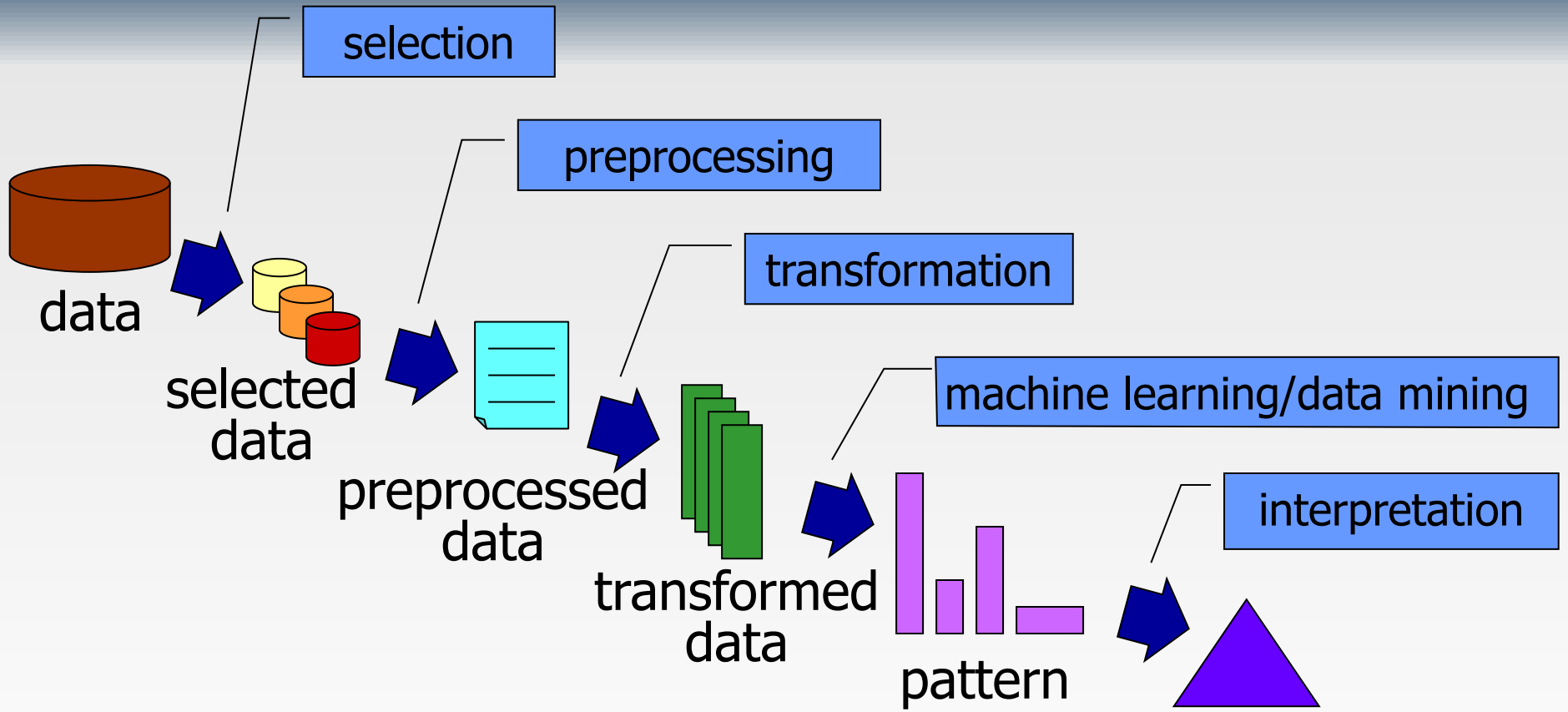
- Consumer behavior in e-commerce sites
  - Selected products, requested information, …
- Search engines and portals
  - Query keywords, searched topics and objects
- Social network data
  - Profiles (Facebook, Instagram, …)
  - Dynamic data: posts on blogs, FB, tweets
- Maps and georeferenced data
  - Localization, interesting locations for users



Profiling: examples of applications

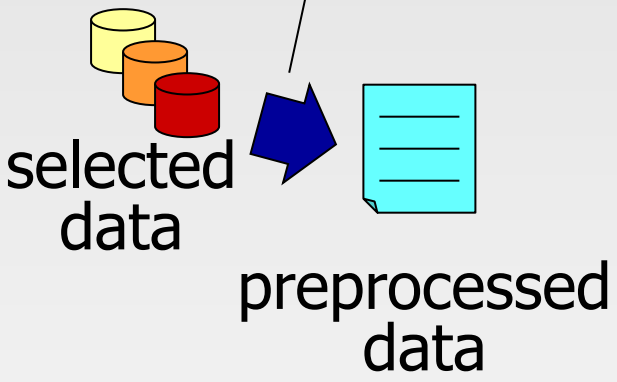
- User/service profiling
  - Recommendation systems, advertisements
- Market basket analysis
  - Correlated objects for cross selling
    - User registration, fidelity cards
- Context-aware data analysis
  - Integration of different dimensions
    - E.g., location, time of the day, user interest
- Text mining
  - Brand reputation, sentiment analysis, topic trends

Knowledge Discovery Process



KDD = Knowledge Discovery from Data

Preprocessing



preprocessing

data cleaning
- reduces the effect of noise
- identifies or removes outliers
- solves inconsistencies

data integration
- reconciles data extracted from different sources
- integrates metadata
- identifies and solves data value conflicts
- manages redundancy

Real world data is "dirty"

Without good quality data, no good quality pattern

A word from practitioners

- At least 80-90% of their work involves not machine learning, but
  - Working with experts to understand the domain, assumptions, questions
  - Trying to catalog and make sense of the data sources
  - Wrangling, extracting, and integrating the data
  - Cleaning the wrangled data

Association rules

- Objective
  - extraction of frequent correlations or pattern from a transactional database

Tickets at a supermarket counter

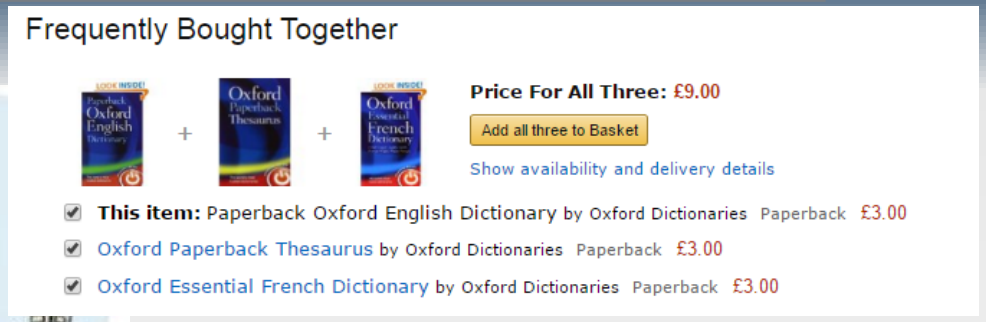
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk
…	…

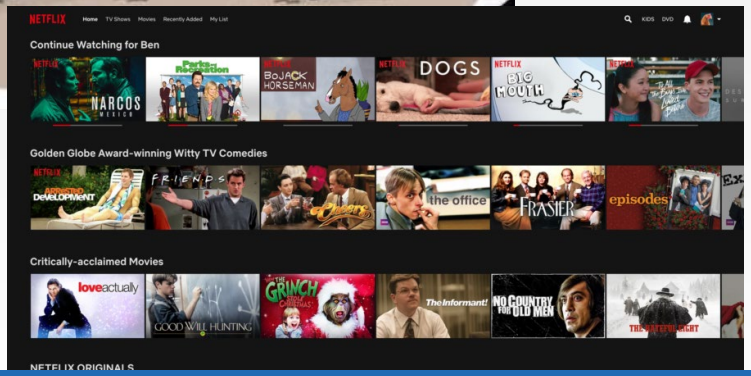
- Association rule

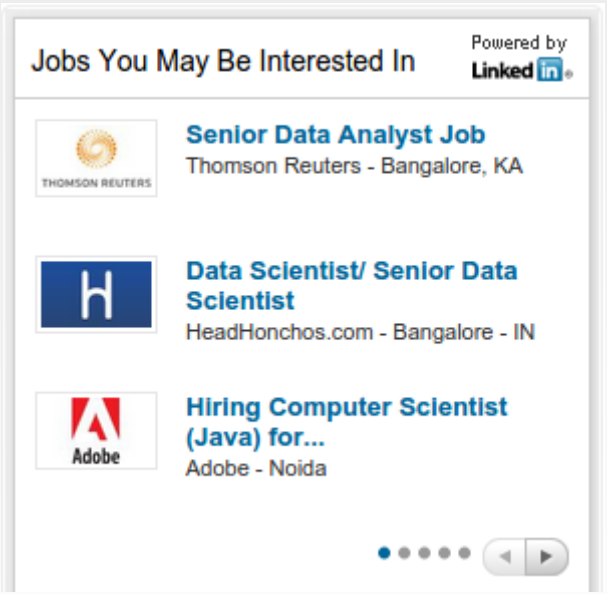
  $$\text{diapers} \Rightarrow \text{beer}$$

  - 2% of transactions contains both items
  - 30% of transactions containing diapers also contain beer

Association rules

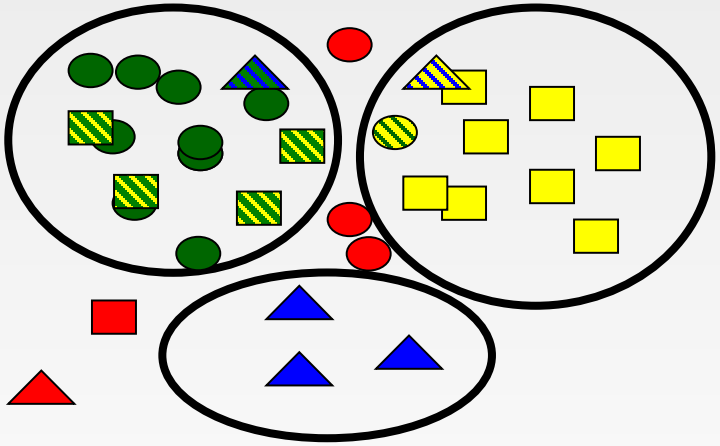


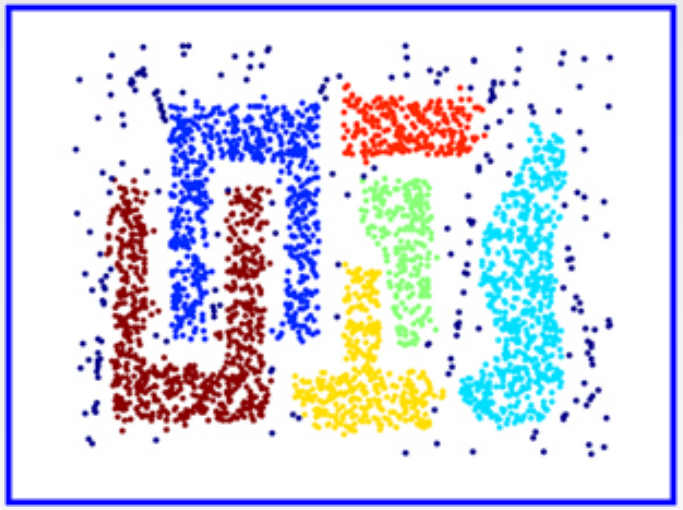




Clustering

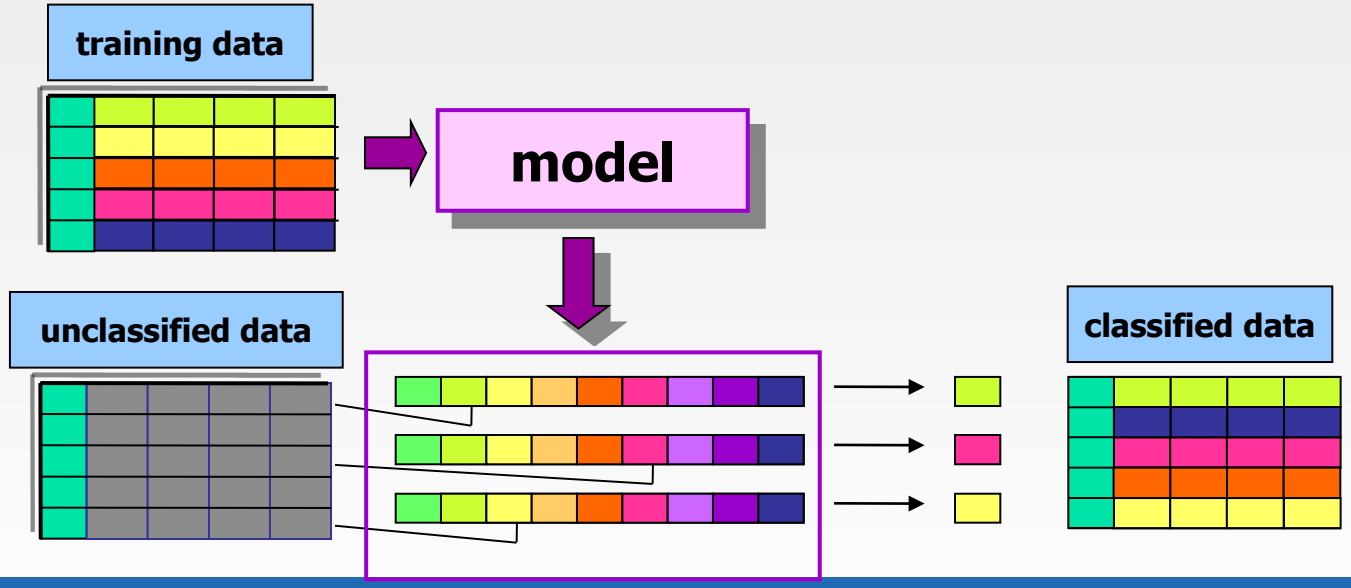
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers





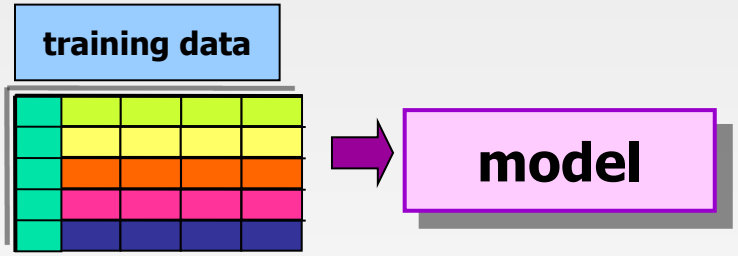
Classification

- Objectives
  - prediction of a class label
  - definition of an data-driven model (descriptive profile) of a given phenomenon, which will allow the assignment of unlabeled data objects to the appropriate class



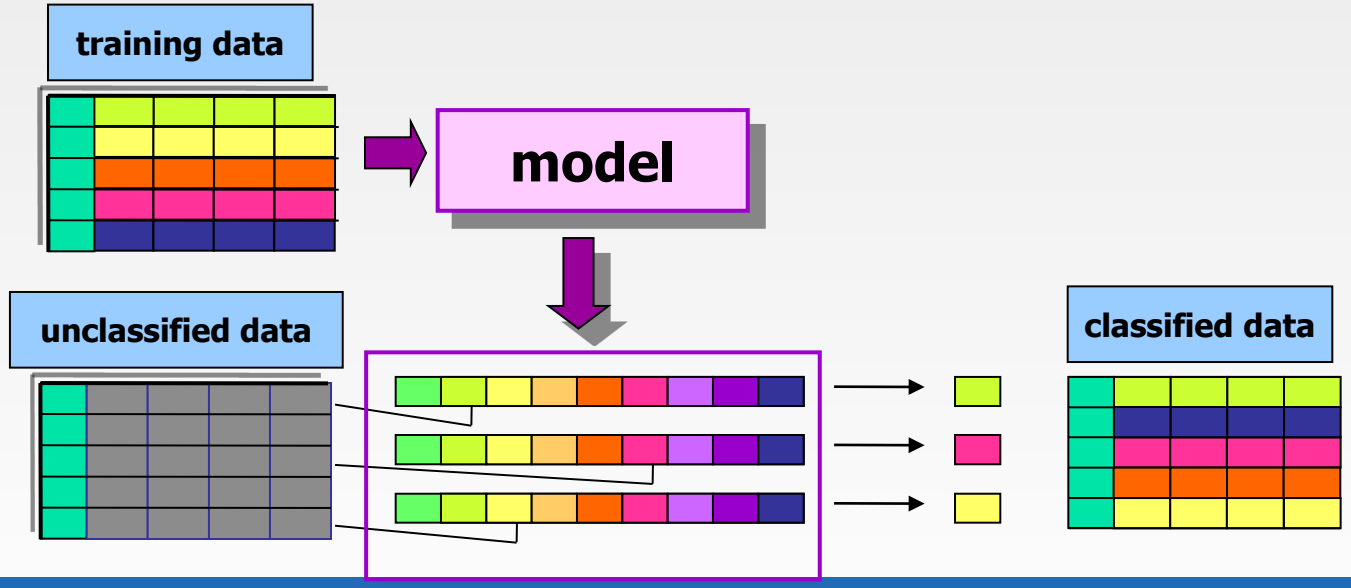
Classification

- Training set
  - Collection of labeled data objects used to learn the classification model



Classification

- Test set
  - Collection of labeled data objects used to validate the classification model
- New data with unknown class label
  - The data-driven model is exploited to predict the class label



Classification techniques

A plethora of different algorithms

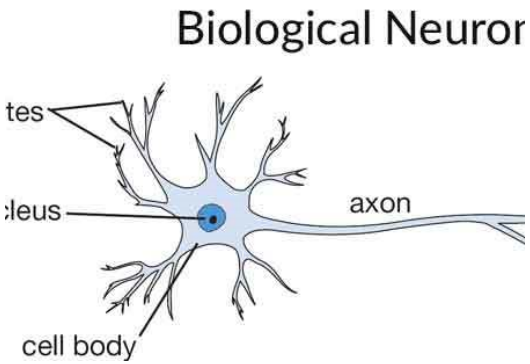
- Decision trees
- Classification rules
- Association rules
- Neural Networks
- Naïve Bayes and Bayesian Networks
- k-Nearest Neighbours (k-NN)
- Support Vector Machines (SVM)
- …

Evaluation dimensions

- **Accuracy**
  - quality of the prediction
- **Interpretability**
  - model interpretability
  - model compactness
- **Robustness**
  - noise, missing data
- **Incrementality**
  - model update in presence of newly labelled record
- **Efficiency**
  - model building time
  - classification time
- **Scalability**
  - training set size
  - attribute number

Artificial Neural Networks

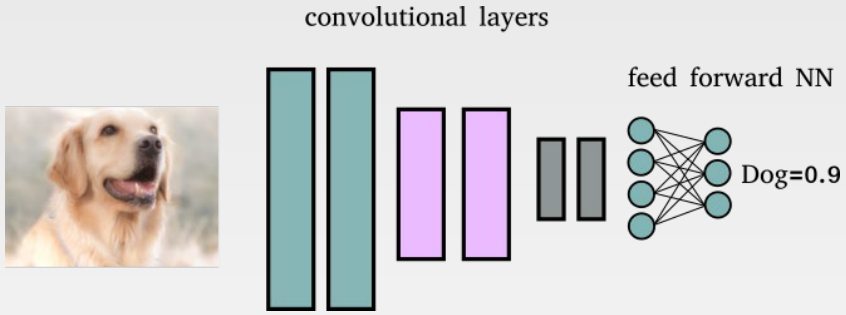
- Inspired to the structure of the human brain
  - Neurons as elaboration units
  - Synapses as connection network



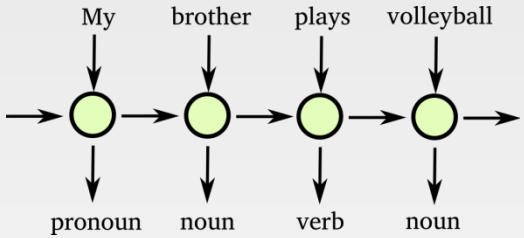
Artificial Neural Networks

- Different tasks, different architectures

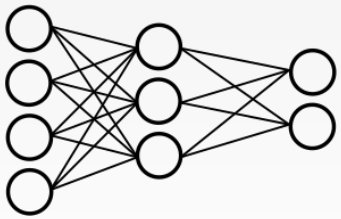
image understanding: convolutional NN (CNN)



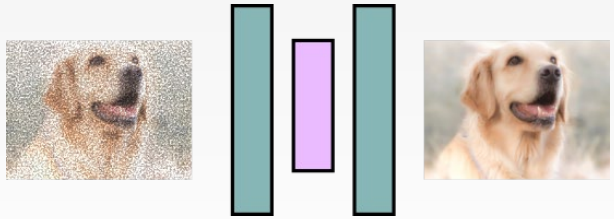
time series analysis: recurrent NN (RNN)



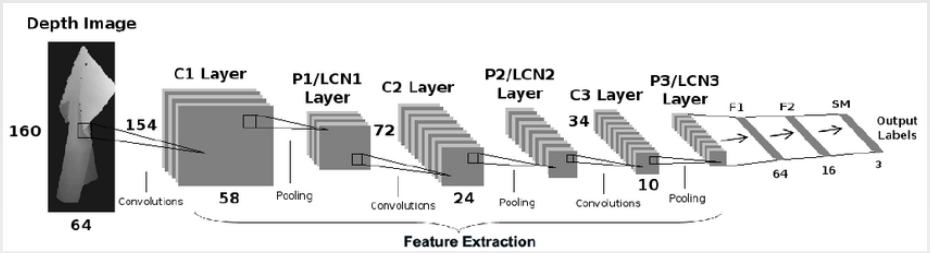
numerical vectors classification: feed forward NN (FFNN)

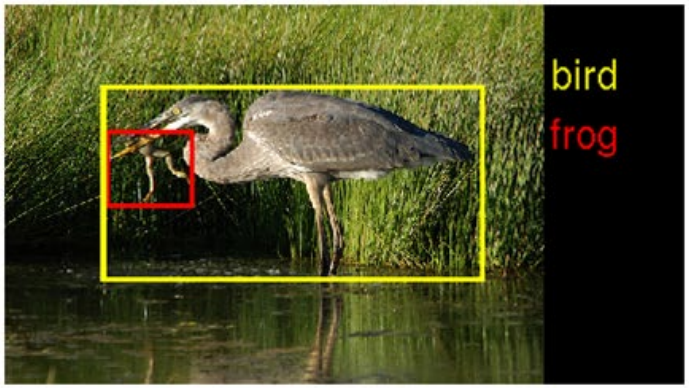


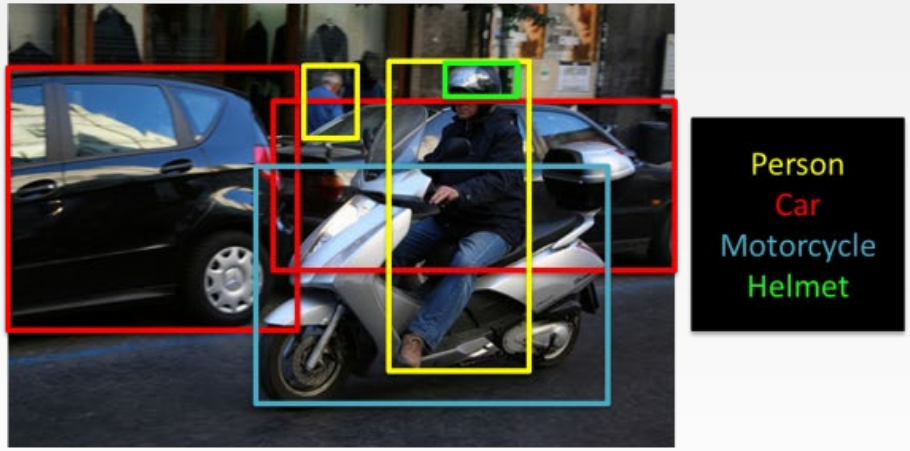
denoising: auto-encoders



Classification



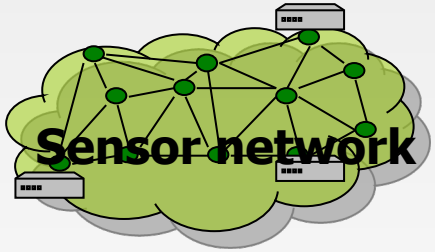




Other techniques

- Sequence mining
  - ordering criteria on analyzed data are taken into account
  - example: motif detection in proteins

- Time series and geospatial data
  - temporal and spatial information are considered
  - example: sensor network data



- Regression
  - prediction of a continuous value
  - example: prediction of stock quotes

- Outlier detection
  - example: intrusion detection in network traffic analysis

The data science process

What *question* are you answering?

What is the right *scope* of the project?

What *data* will you use?

What *techniques* are you going to try?

How will you *evaluate* your result?

What *maintenance* will be required?

The data science recipe

- Different ingredients needed
  - Data expert
    - Data processing, data structures
  - Data analyst
    - Data mining, statistics, machine learning
  - Visualization expert
    - Visual art design, storytelling skills
  - Domain expert
    - Provide understanding of the application domain
  - Business expert
    - Data driven decisions, new business models