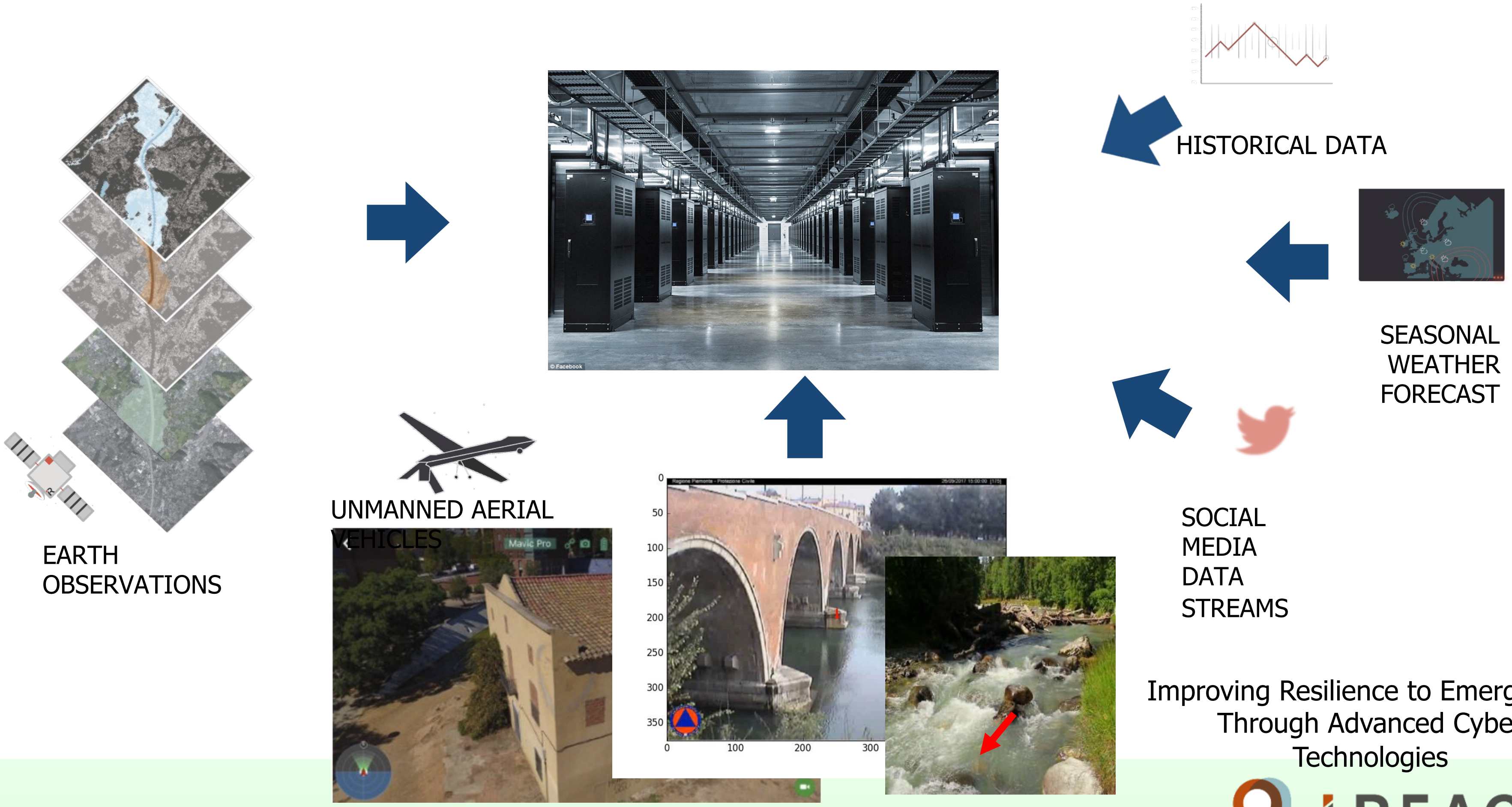
Data science
The Big Data challenge

Elena Baralis, Tania Cerquitelli, Eliana Pastor
Politecnico di Torino

Big data hype?

Emergency management



Improving Resilience to Emergencies Through Advanced Cyber Technologies

Emergency management

User engagement

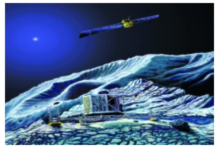
2005

Now

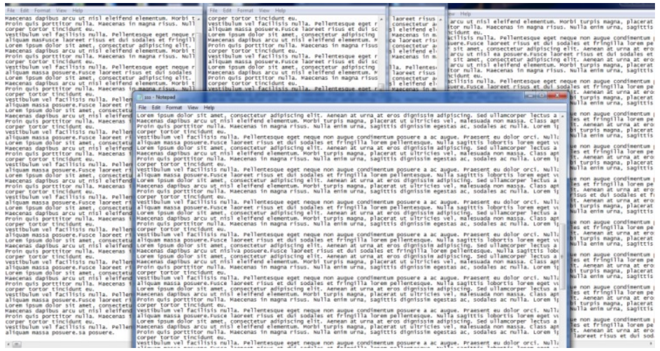
Who generates big data?

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

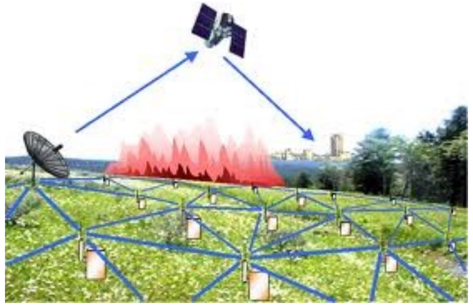
- Health and scientific computing

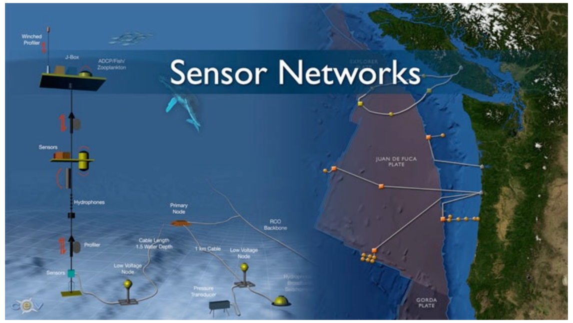


Who generates big data?



- 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"*

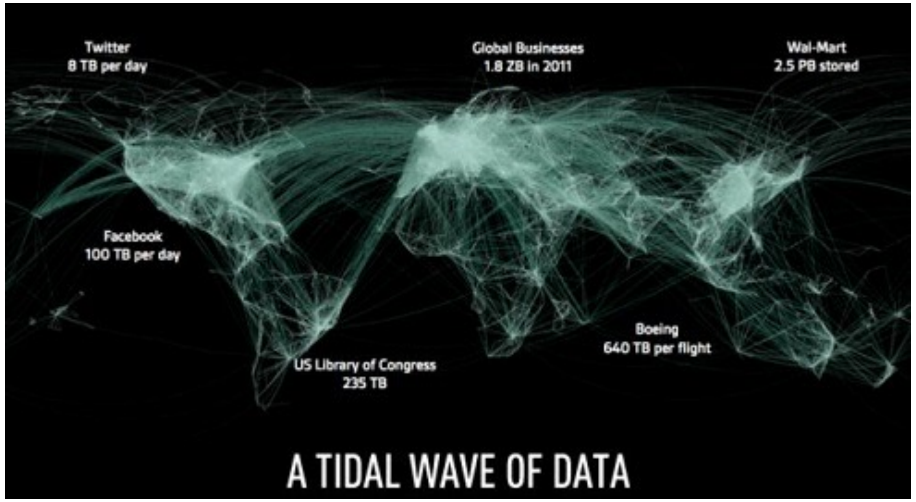
What is big data?

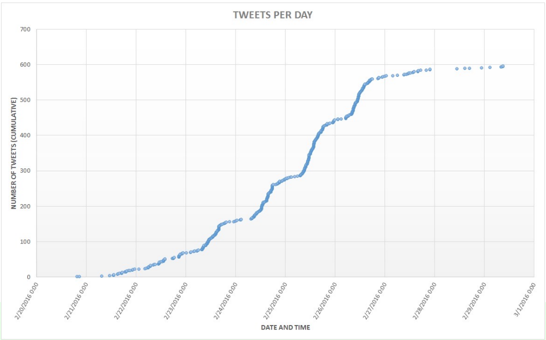
- Many different definitions

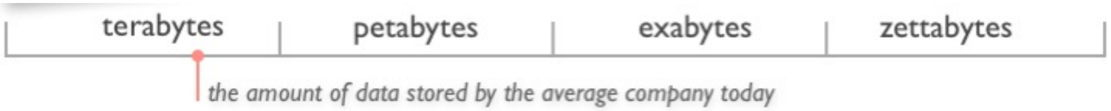
*"Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract **value** and hidden **knowledge** from it"*

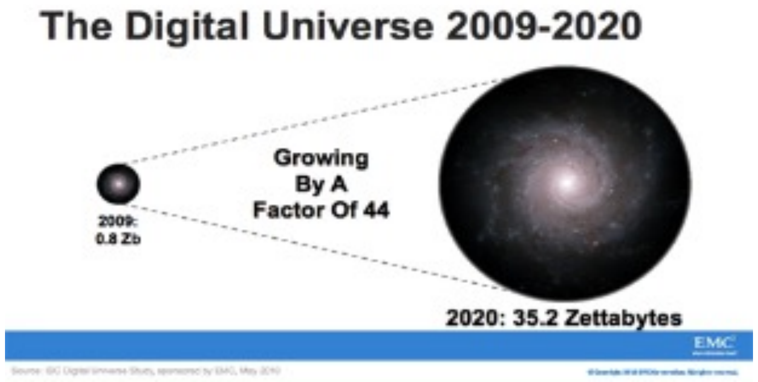
The Vs of big data: Volume

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

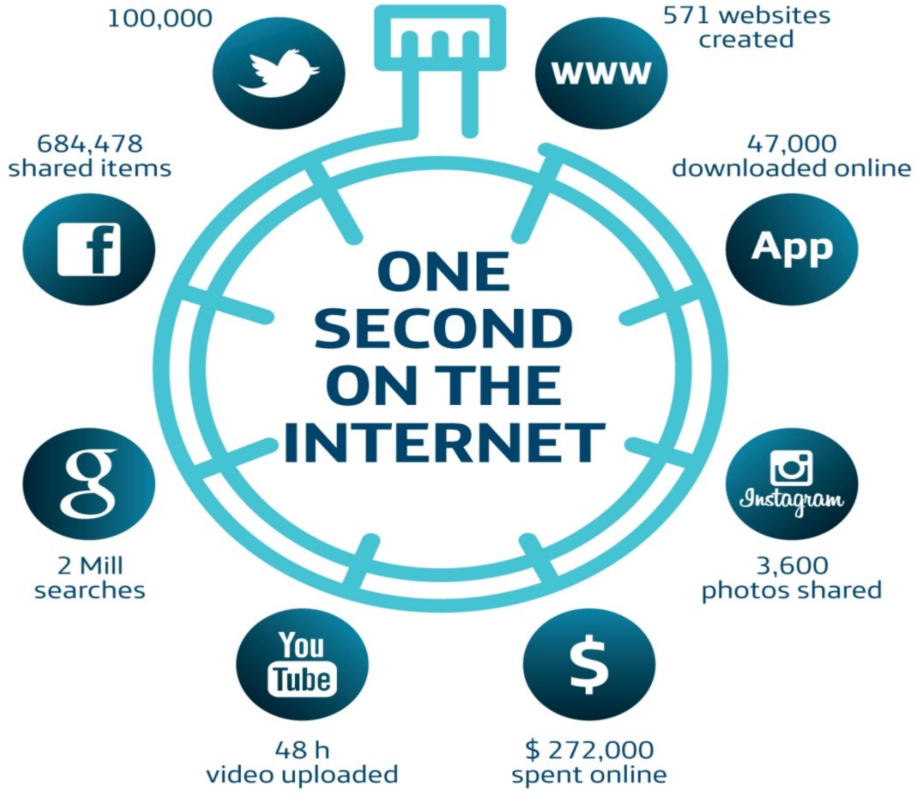








On the Internet…

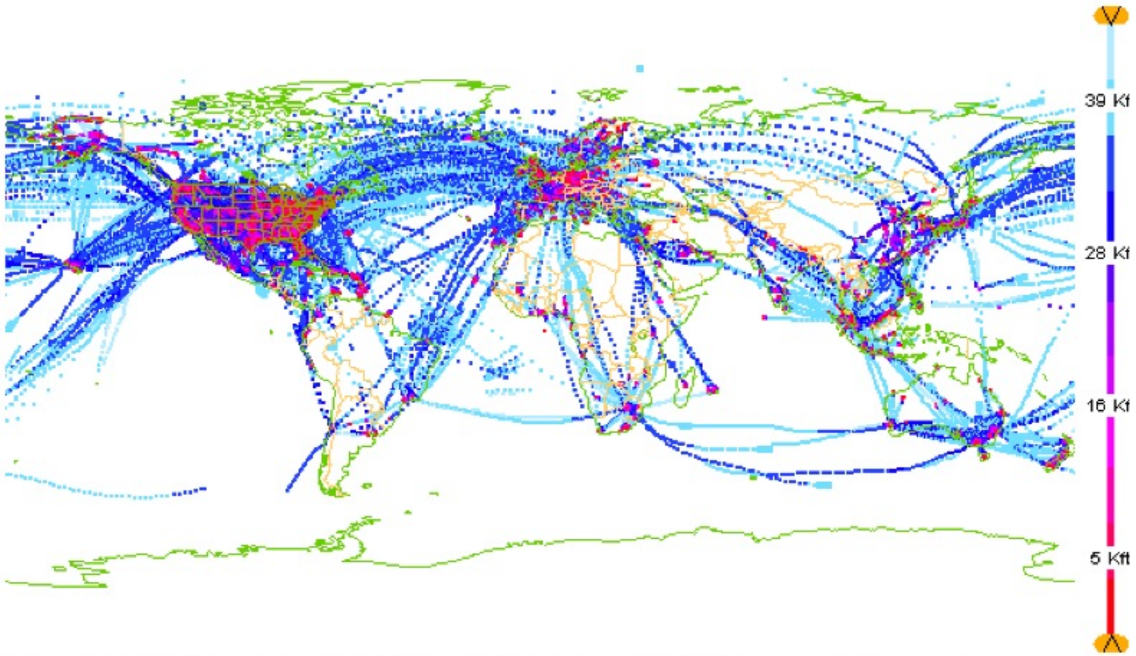


Source: Telefónica analysis based on Social and Digital Media Revolution Statistics 2013 from MistMediaGroup (htt://youtube.com/watch?v=Slb5x5fixk4).

- http://www.internetlivestats.com/

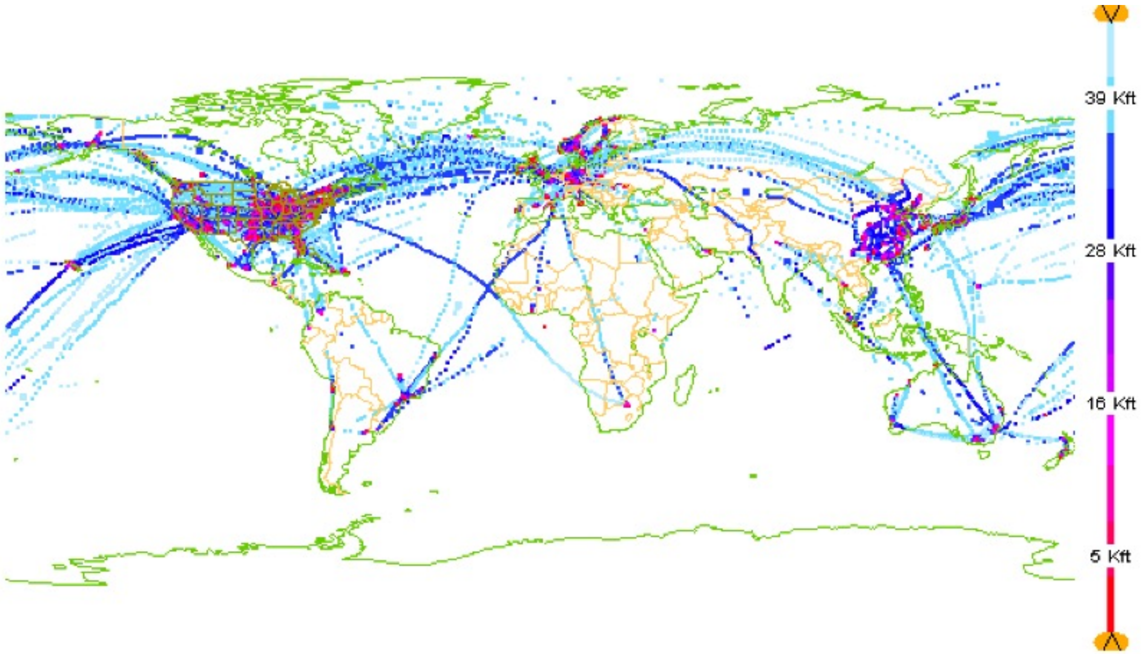
Weather forecast

January 2020



31-Jan-2020 00:00:00 -- 31-Jan-2020 23:59:58 (872710 obs loaded, 728535 in range, 24197 shown)

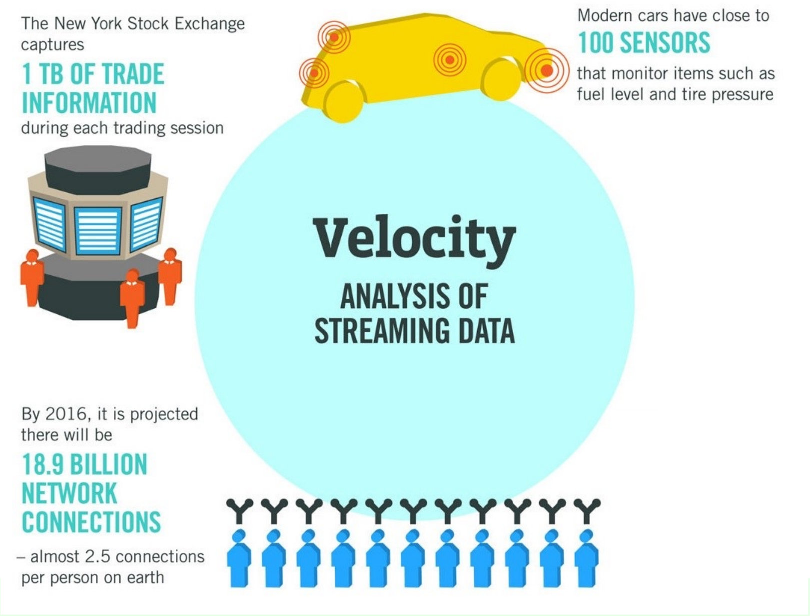
May 2020



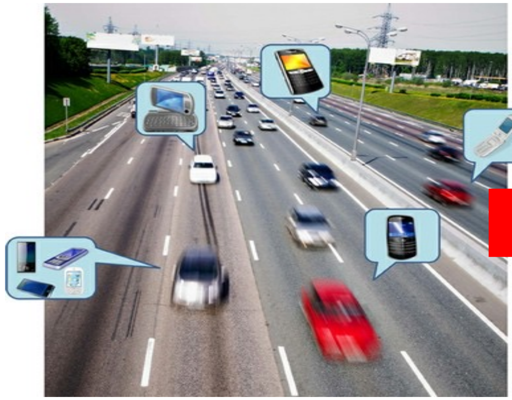
03-May-2020 15:00:00 -- 04-May-2020 15:24:19 (132910 obs loaded, 112894 in range, 11217 shown)

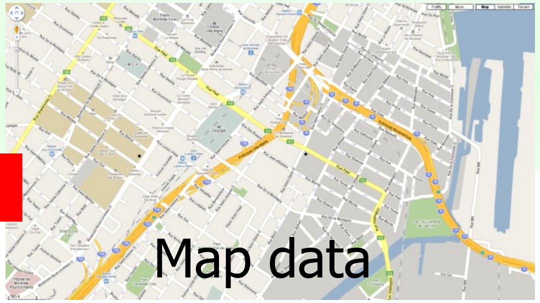
The Vs of big data: Velocity

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

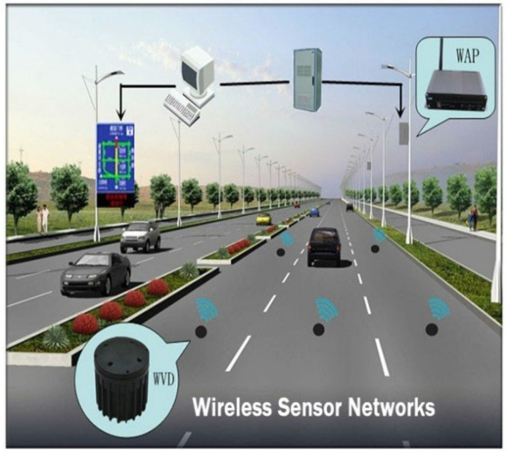


(Near) Real time processing



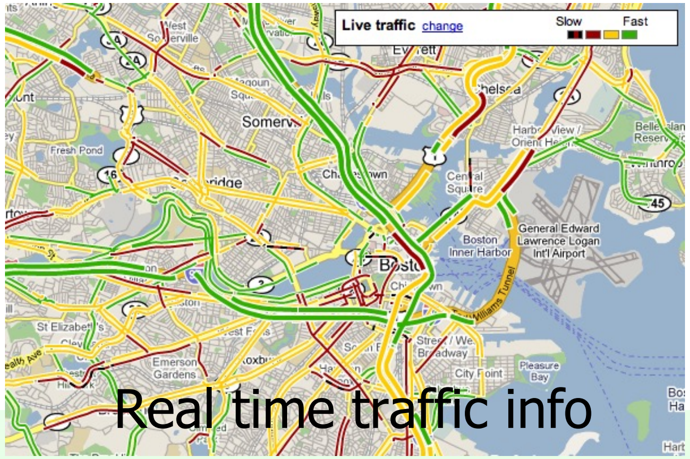


Map data



Sensing

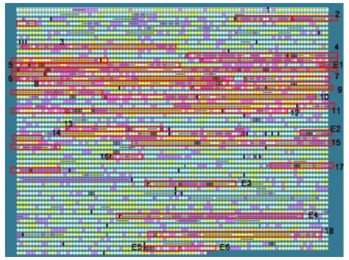
Computing

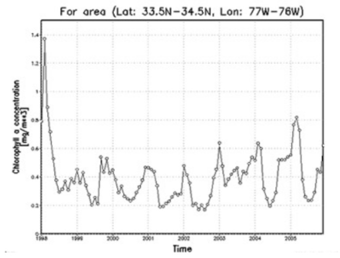


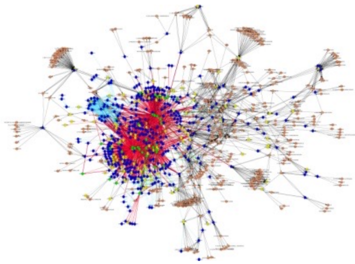
Real time traffic info

The Vs of big data: Variety

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







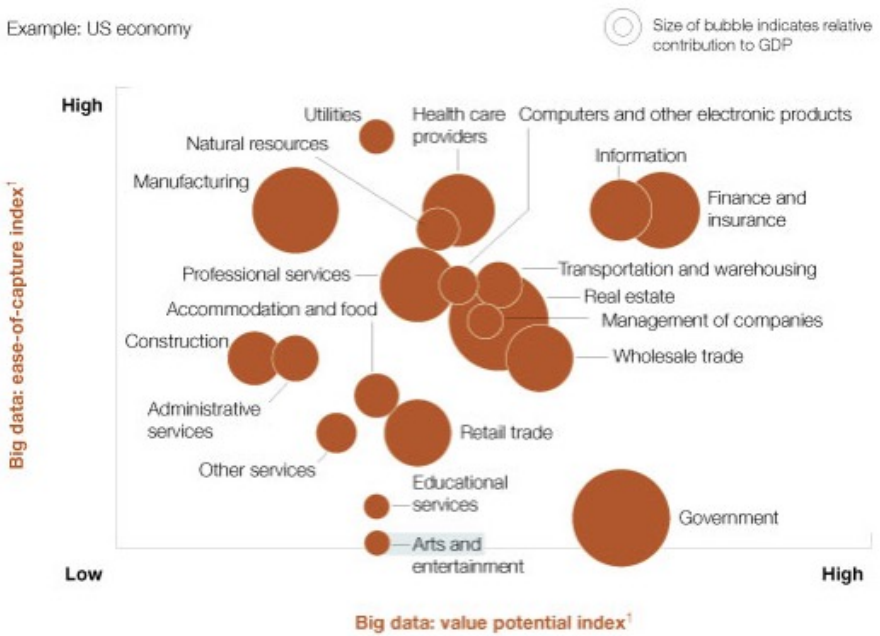
- A single application may generate many different formats

The Vs of big data: Veracity

- Data quality

The most important V: Value

- Translate data into business advantage



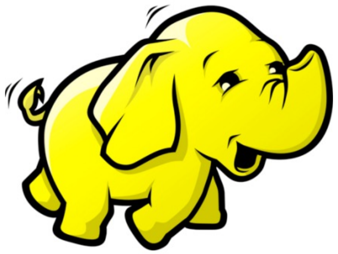
[1] For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

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"

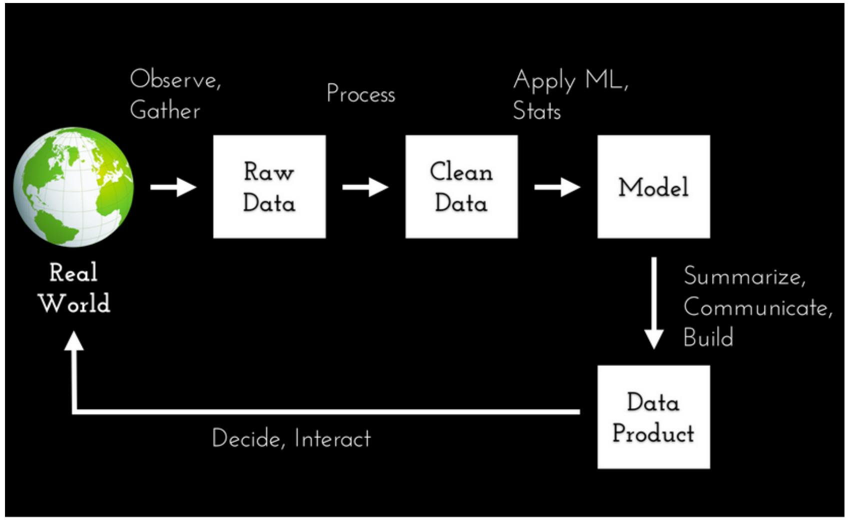


Data science

"Extracting meaning from very large quantities of data"

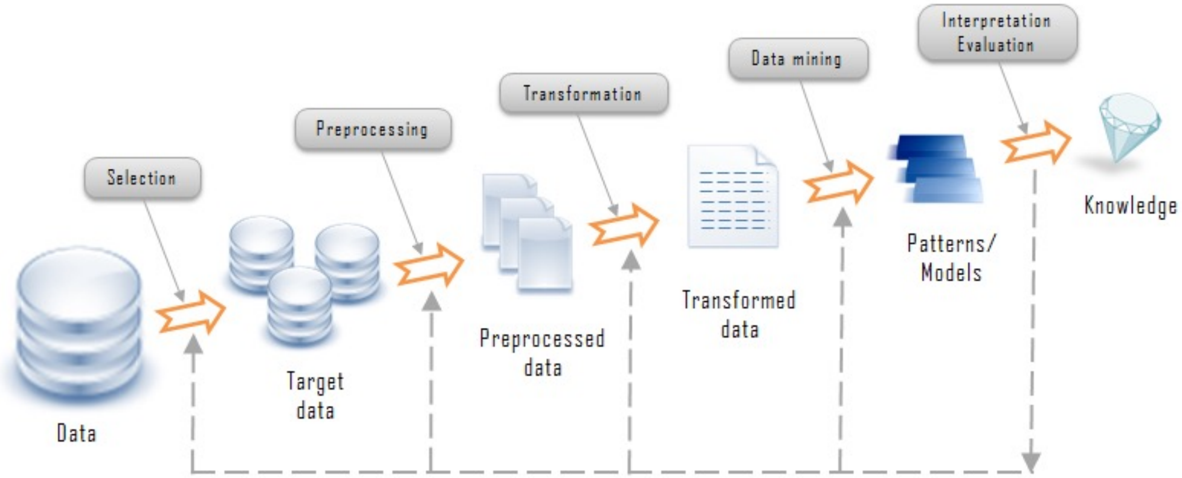
D.J. Patil coined the word *data scientist*

The data science process



AKA *KDD* process

Knowledge Discovery in Databases



Generation → Acquisition → Storage → Analysis

Generation

- Passive recording
  - Typically structured data
  - Bank trading transactions, shopping records, 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)

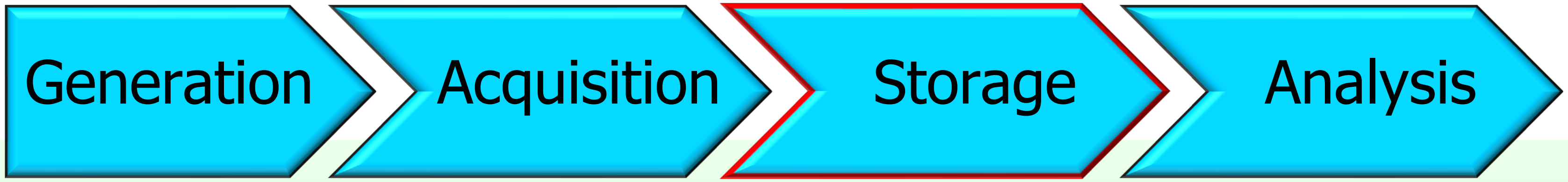
DBMG

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

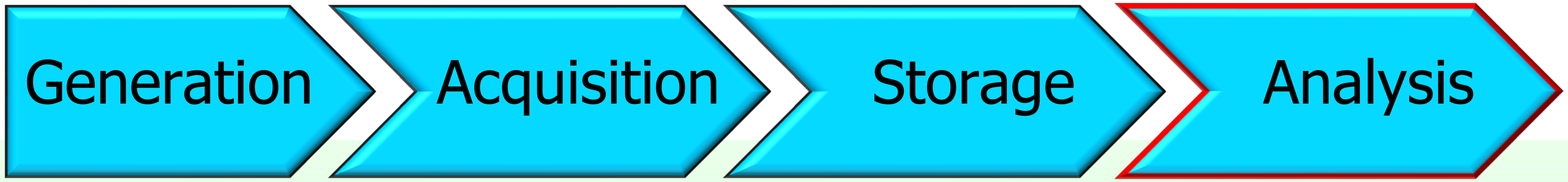
Storage

- 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

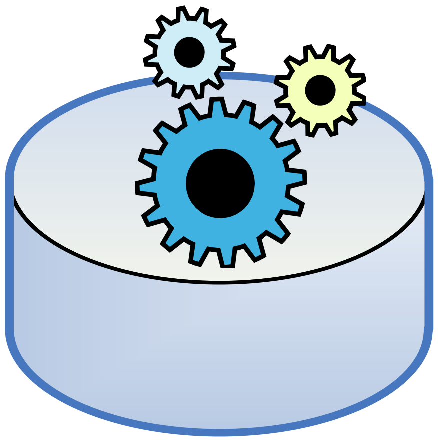
- Objectives
  - Descriptive analytics, predictive analytics, prescriptive analytics
- Methods
  - Statistical analysis, machine learning and data mining, text mining, network and graph data mining
  - Association analysis, classification and regression, clustering
- Diverse domains call for customized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



Machine learning and 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*



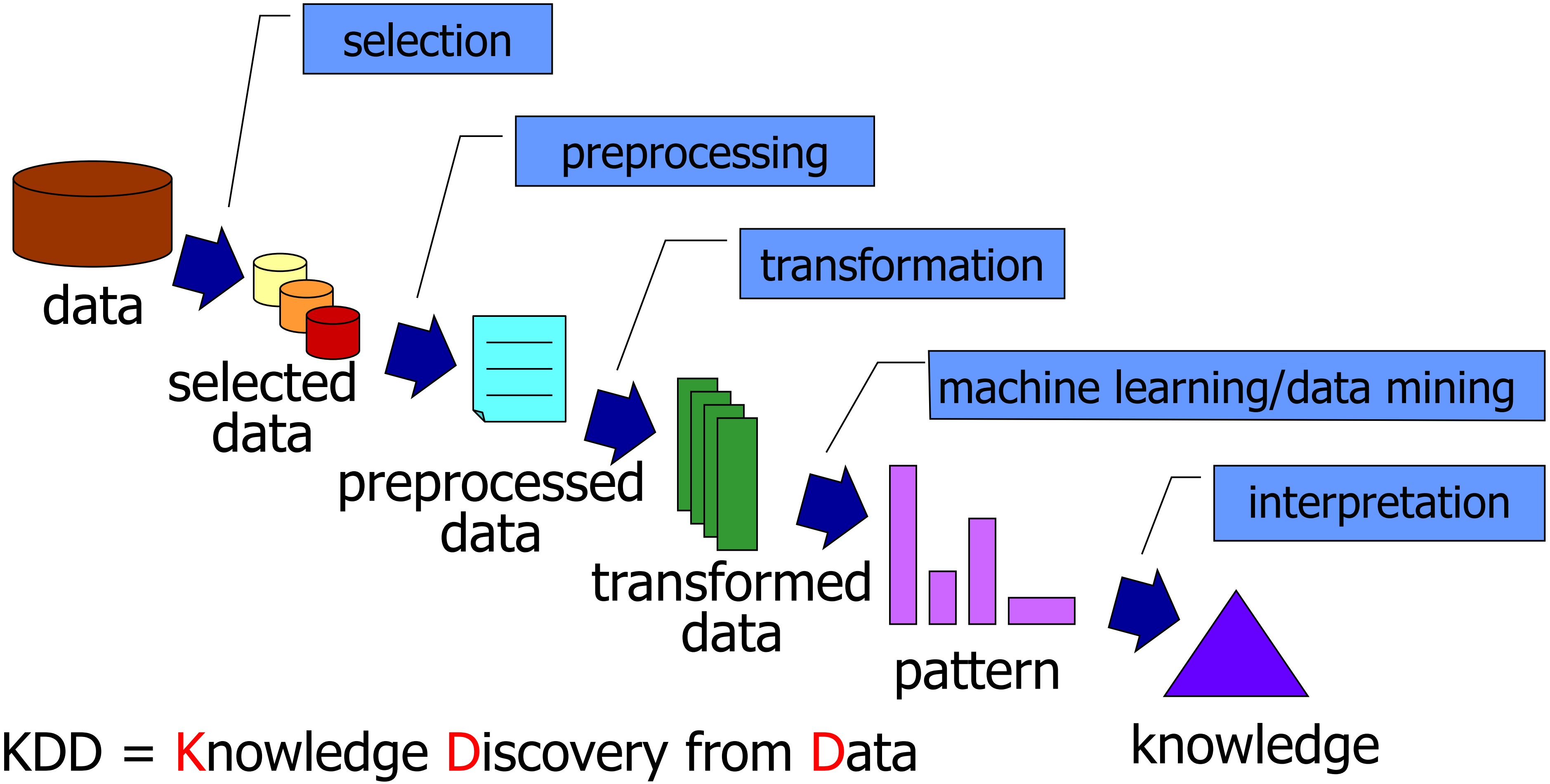
Example: profiling

- 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

Example: profiling

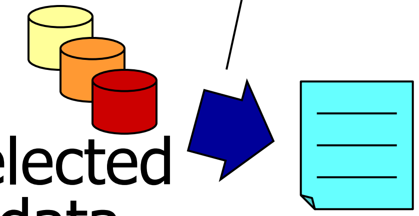
- User/service profiling
  - Recommendation systems, advertisements
- Market basket analysis
  - Correlated objects for cross selling
    - User registration, fidelity cards
- Context-aware data analysis
  - Integration of different dimensions
    - E.g., location, time of the day, user interest
- Text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

Knowledge Discovery Process



KDD = Knowledge Discovery from Data

Preprocessing



selected data

preprocessed data

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

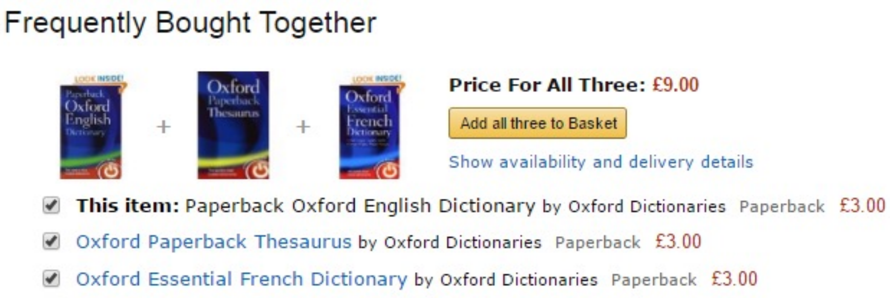
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk
…	…

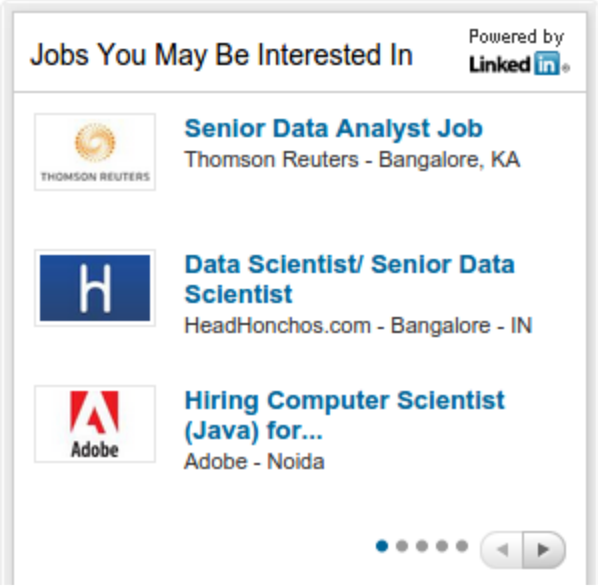
- Association rule

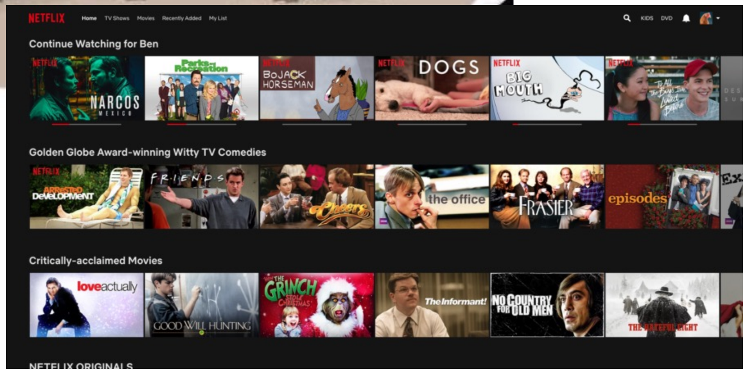
  diapers $\Rightarrow$ beer

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

Association rules

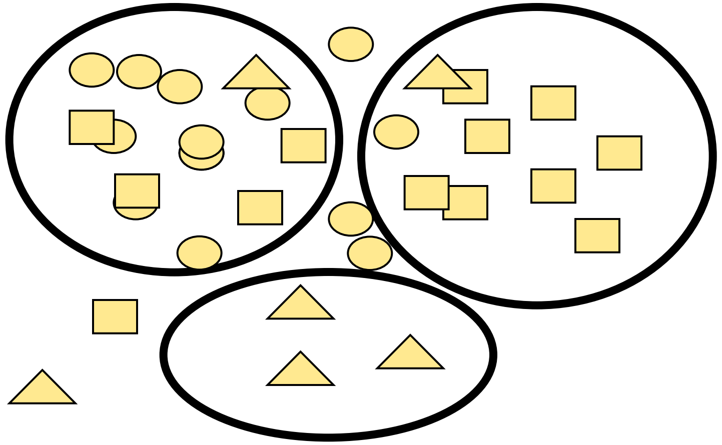






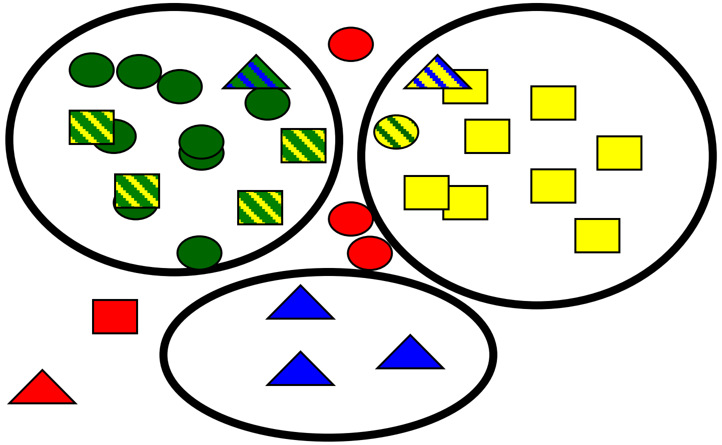
Clustering

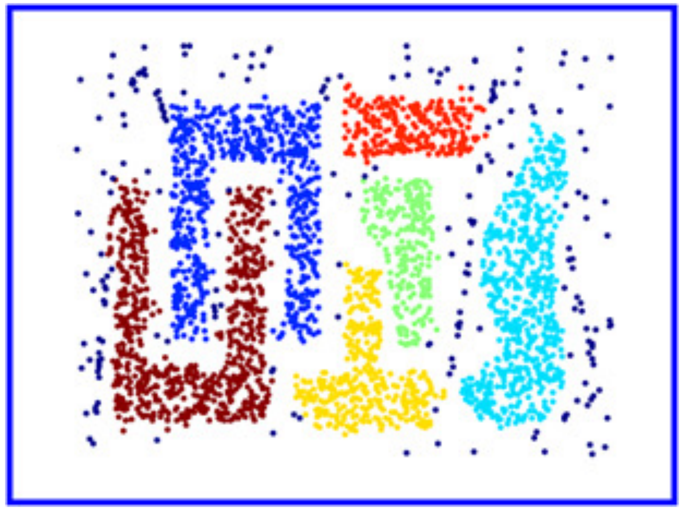
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers



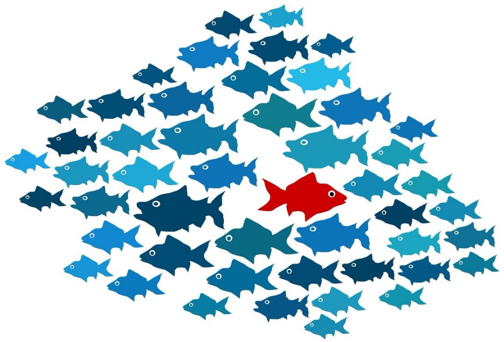
Clustering

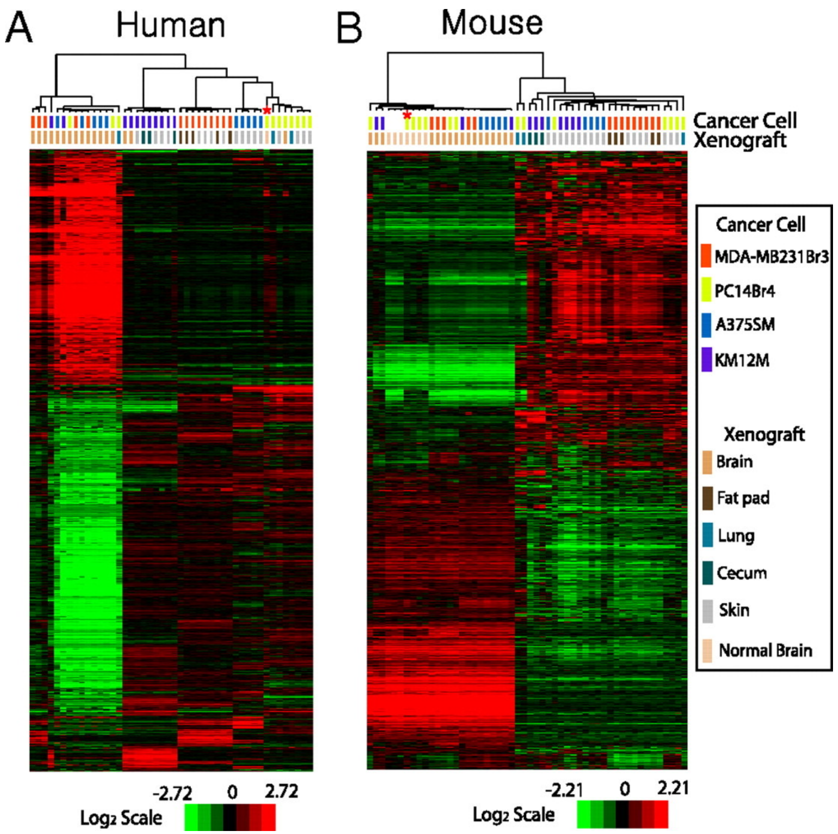
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers





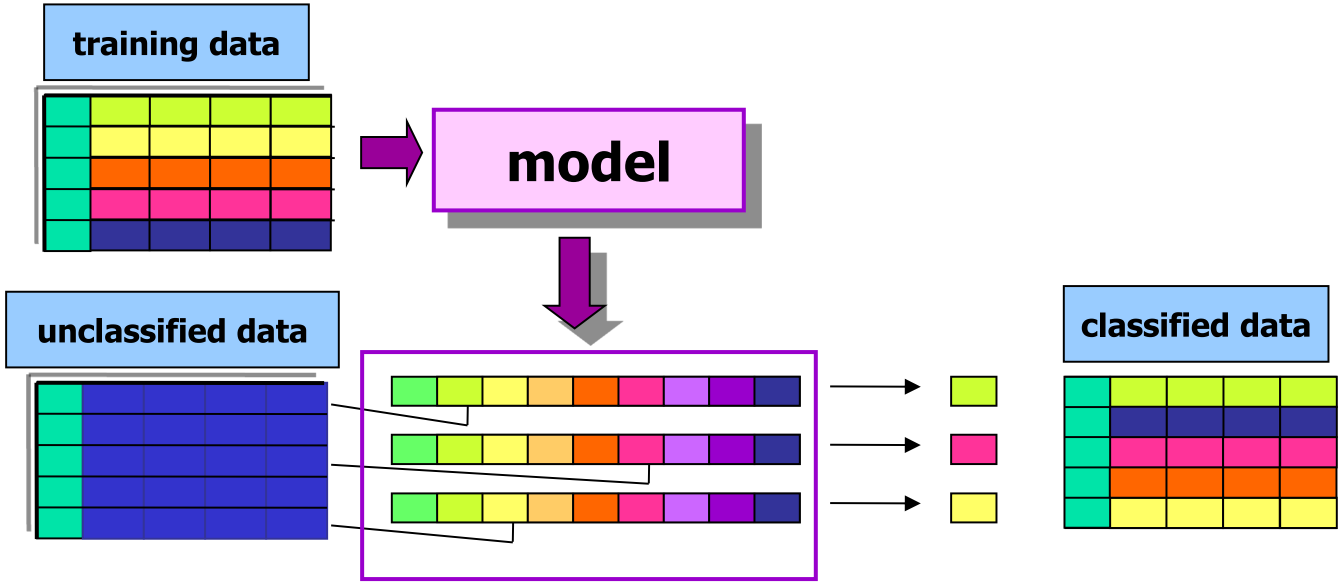
Clustering




A
Human
B
Mouse
Cancer Cell
Xenograft
Cancer Cell
MDA-MB231Br3
PC14Br4
A375SM
KM12M
Xenograft
Brain
Fat pad
Lung
Cecum
Skin
Normal Brain
-2.72
0
2.72
Log2 Scale
-2.21
0
2.21
Log2 Scale

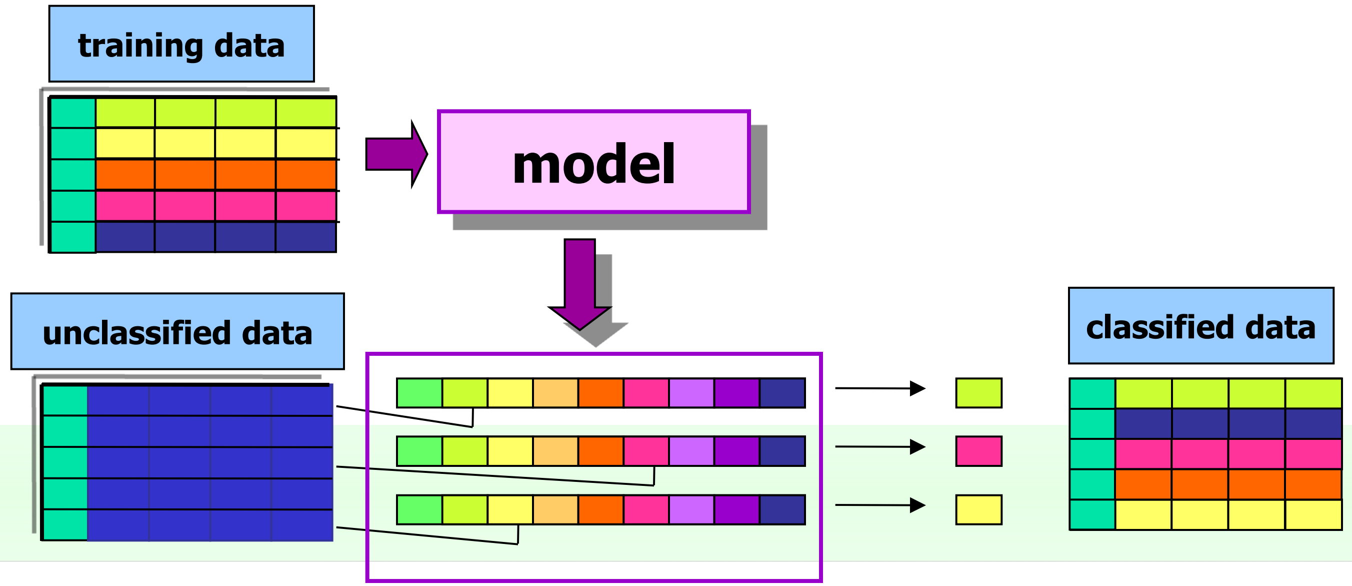
Classification

- Objectives
  - prediction of a class label
  - definition of an interpretable model of a given phenomenon



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



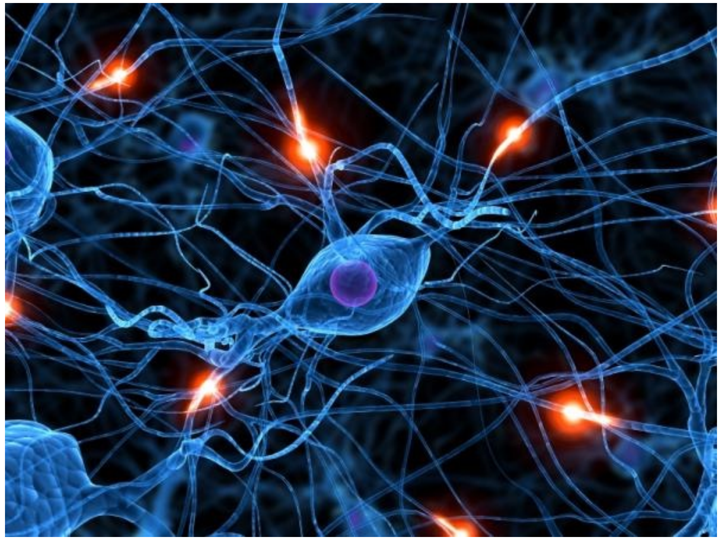
- A plethora of different algorithms
  - Decision trees
  - Classification rules
  - Association rules
  - Neural Networks
  - Naïve Bayes and Bayesian Networks
  - k-Nearest Neighbours (k-NN)
  - Support Vector Machines (SVM)
  - ...

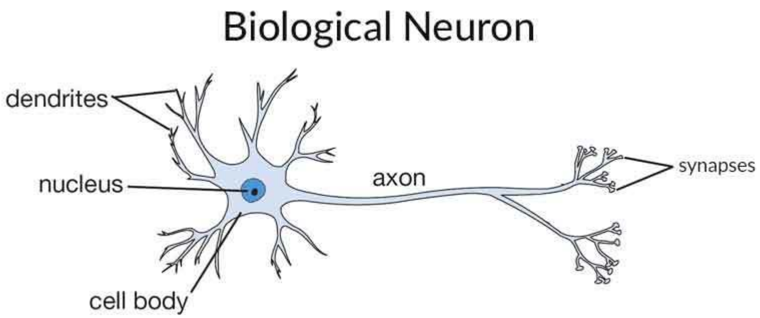
Evaluation dimensions

- 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

numerical vectors classification: feed forward NN (FFNN)

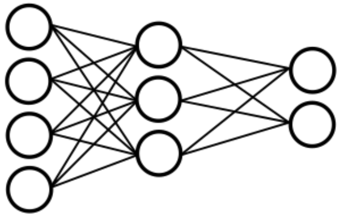
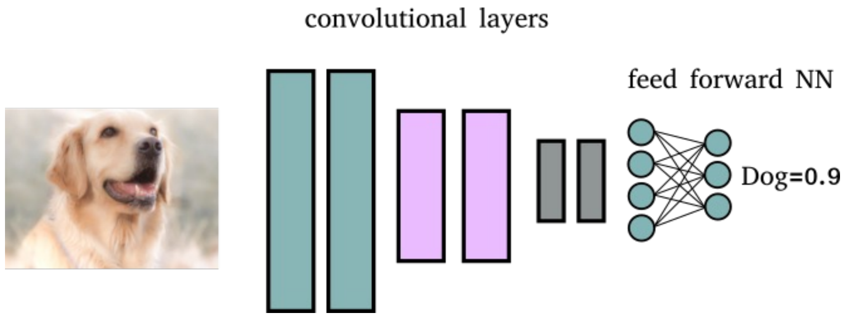
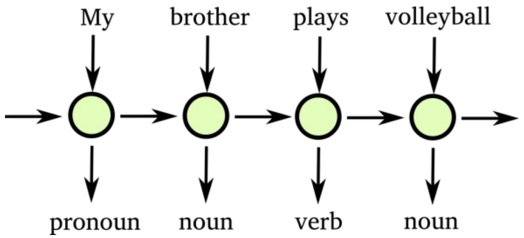


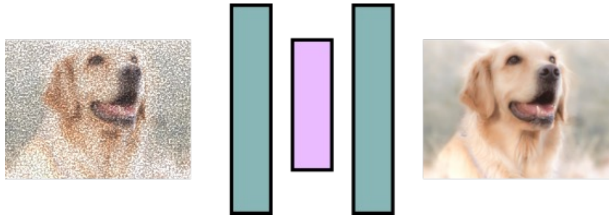
image understanding: convolutional NN (CNN)



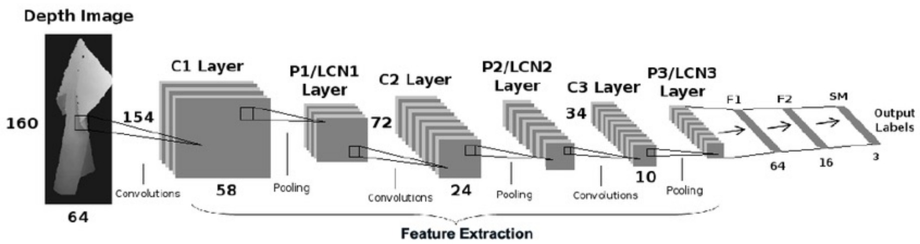
time series analysis: recurrent NN (RNN)

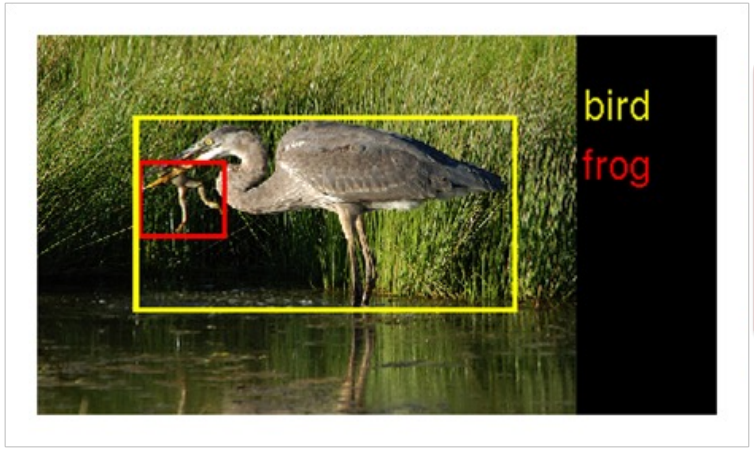


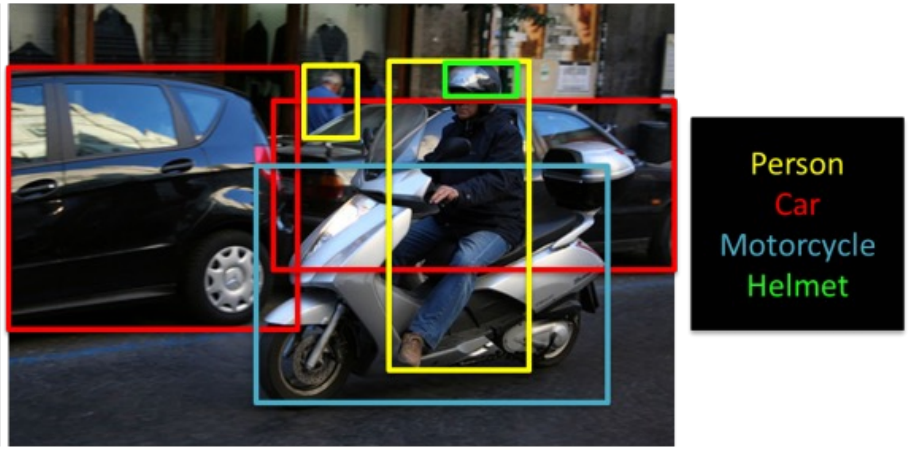
denoising: auto-encoders



Classification

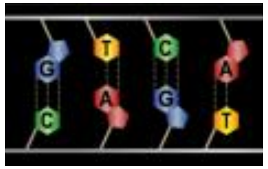




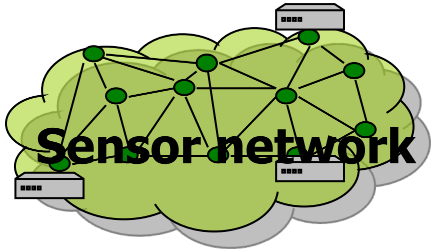


Other data mining techniques

- 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

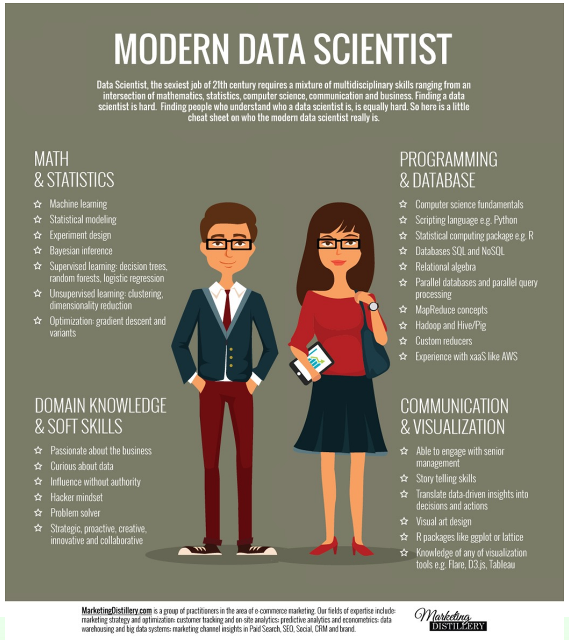
The data science process

- 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?

Content derived by material from the OpenDS4All project

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

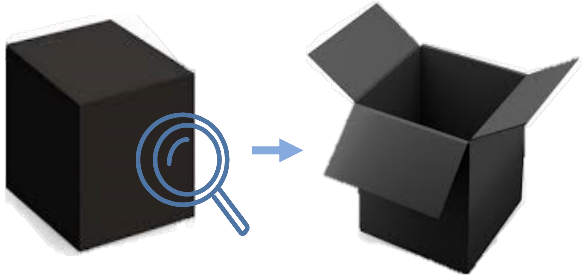


Open issues

- Social impact of analysis is very important
  - Interpretability and transparency of the analysis process
  - Bias in algorithms and data
  - Privacy preservation
- AI-based systems are often «black boxes»
  - It is unclear for humans why an AI system makes a certain decision based on some input data
  - Because of the opaqueness people cannot assess whether they were discriminated against on the basis of, e.g., racial origin

Interpretability in machine learning

"The ability to explain or to present in understandable terms to a human"



Open the black box

Trade-off Accuracy-Interpretability

- Model explanation: global understanding of how a model works
- Prediction explanation: local understanding of why a prediction is made
- Interpretable feature selection: incorporating interpretability-based criteria into the model design

Interpretability

- Learned decision rule in pneumonia patient dataset from USA hospital

  history of asthma → lower chance of dying from pneumonia
- MD consider asthma as a serious risk factor
- Analysis
  - asthmatics probably notice earlier the symptoms of pneumonia
  - a healthcare professional is going to provide earlier pneumonia diagnosis
  - as high-risk patients, they're going to get high-quality treatment sooner than other people
  - asthmatics actually have almost half the chance of dying than non asthmatics
- Using a neural network, this model issue would *never* have been uncovered

Algorithmic and data bias

- Task: predict likelihood of an individual committing a future crime
  - Risk scores used by US criminal justice system
- Scores computed from
  - Questions answered by the defendants
  - Information pulled by criminal records
- Race was not among the questions
  - … however other items may be correlated (e.g., poverty, joblessness)
- Software product flagged black defendants as future criminals more frequently than white defendants

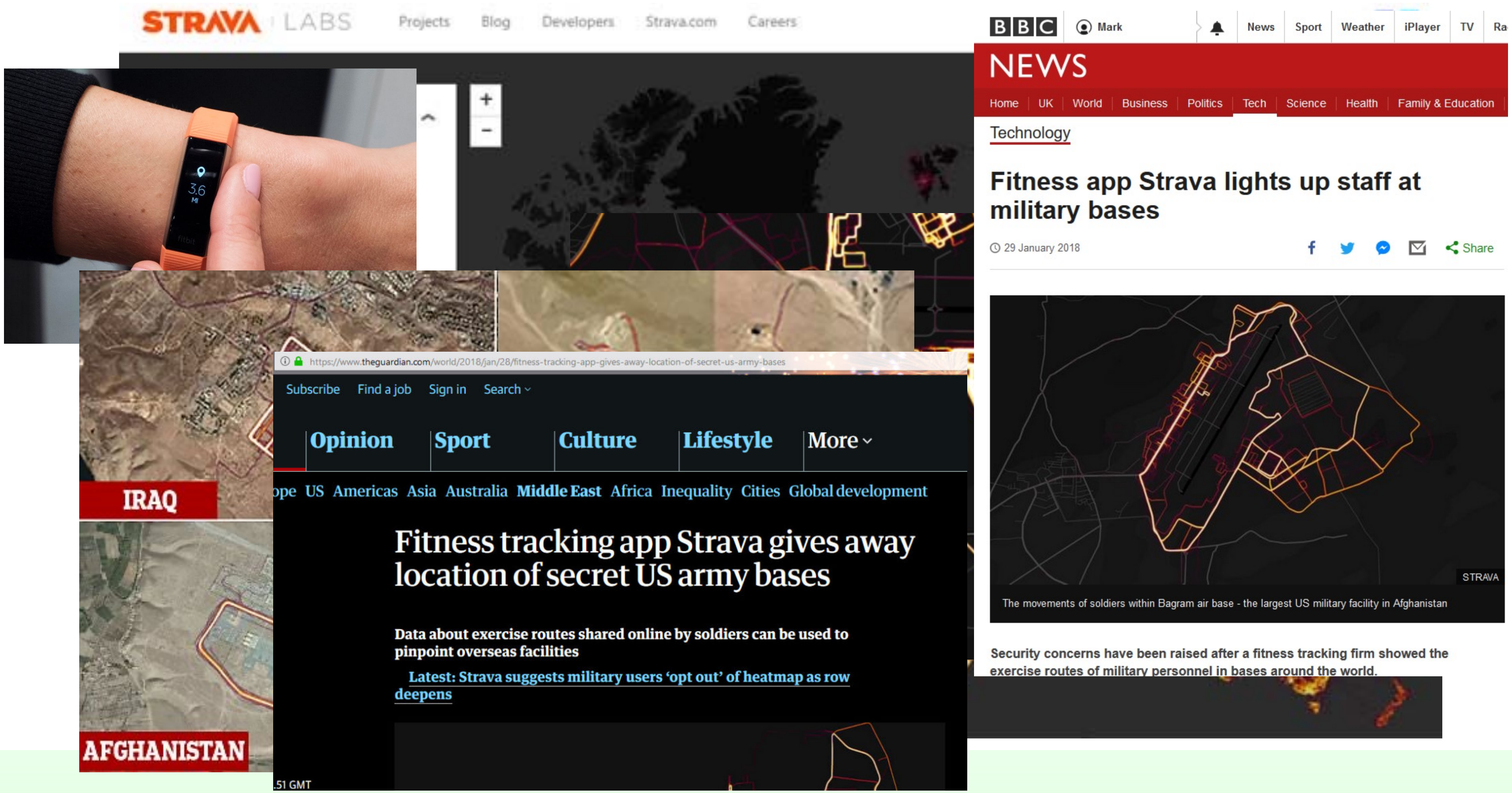
⇨ Training data was biased by a larger black defendant population

CV-scanning tool

- In 2014, Amazon's data scientists simplified employee recruitment
  - an AI algorithm to automatically identify the most qualified candidates from a vast pool of resumes.
- Issue: the algorithm discriminated against women.
  - The data-driven model was derived from analysis of resumes submitted in the past, which were dominated by male applicants
  - The algorithm learned that men would be better applicants than women

Privacy

Strava released their global heatmap. 13 trillion GPS points from their users



How AI can lead to discrimination

- **Definition of the label to be predicted**
  - Objective: Selection of the best employees of a company
  - Method: What criteria are used to define a good employee
  - Issue: It is easy to discriminate against protected categories (even if this is done unintentionally)

How AI can lead to discrimination

- **The data used to train the model contains biases**
  - The data model created by an AI algorithm reflects the biases in the data
  - Examples: Datasets with only male resumes, datasets with only crimes committed by foreign nationals

How AI can lead to discrimination

- **Attributes used to create the data-driven model**
  - Objective: Automatic selection of the best resumes for specific leadership positions
  - Interesting attributes: University Name, Disciplines, Graduation grade
  - Issue: The company could consider individuals who have studied at famous and prestigious (expensive) universities
  - This would discriminate against individuals with strong backgrounds who have not studied at famous universities.

How AI can lead to discrimination

- **Proxies**
  - Variables that are 'neutral' and not directly discriminatory (e.g., zip code)
  - These variables may be indirectly correlated with a minority category (e.g., zip code only for certain geographic areas)

Responsible Artificial Intelligence

- Ethical principles
  - Mandatory for fully-integrating AI systems in our society
  - Enforced throughout the
    - development
    - implementation
    - operation stages
  - of new AI solutions
- Companies need to adopt clear processes and practices that ensure AI systems comply with strict responsible AI principles

Responsible AI

- Fairness
  - AI systems must be designed in ways that maximize fairness, non-discrimination and accessibility.
  - All AI designs should promote inclusivity by correcting both unwanted data biases and unwanted algorithmic biases.
- Reliability, Safety, and Security
  - AI systems should cause no direct harm and always aim to minimize indirect harmful behavior.
  - AI systems must be reliable in that they should always perform as from unauthorized parties.
- Privacy
  - By design, AI systems must respect privacy by providing individuals with agency over their data and the decisions made with it.
  - AI systems must also respect the integrity of the data they use.

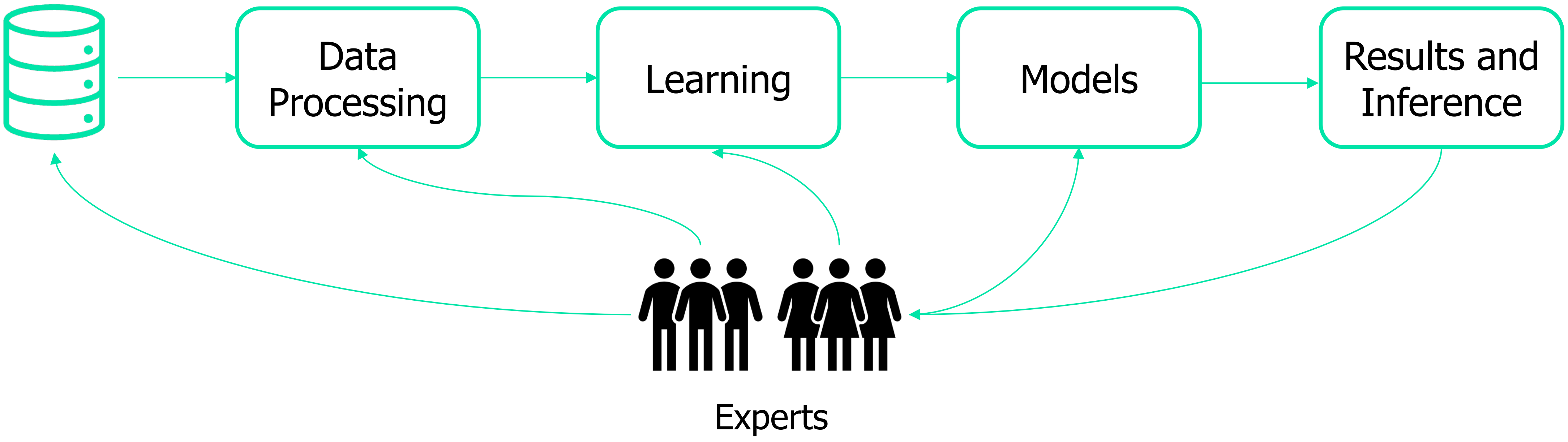
Content derived by material from Nokia's 6 Pillars of Responsible AI

Responsible AI

- Transparency
  - AI-based systems must be explainable and understandable.
  - AI systems should produce outputs that are easily comprehensible to the stakeholder
- Sustainability
  - AI-based systems should attempt to be societally sustainable by empowering society and democracy
  - environmentally sustainable, by reducing the amount of power required to train and run them
- Accountability
  - AI systems should be developed and deployed through consultation and collaboration with all stakeholders such that true accountability becomes possible.
  - The long-term effects of any AI application should be understandable by all stakeholders
  - If an AI system deviates from its intended results, then we need to have policies in place to ensure those deviations are detected, reported and remedied.

Content derived by material from Nokia's 6 Pillars of Responsible AI

Humans in the loop (HITL)



Open issues

- Social impact of analysis is very important
  - Towards responsible AI systems

- Many technical issues are not solved
  - Data dimensionality
  - Complex data structures, heterogeneous data formats
  - Data quality