

# Data science

## *The Big Data challenge*

---

ELENA BARALIS

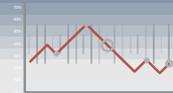
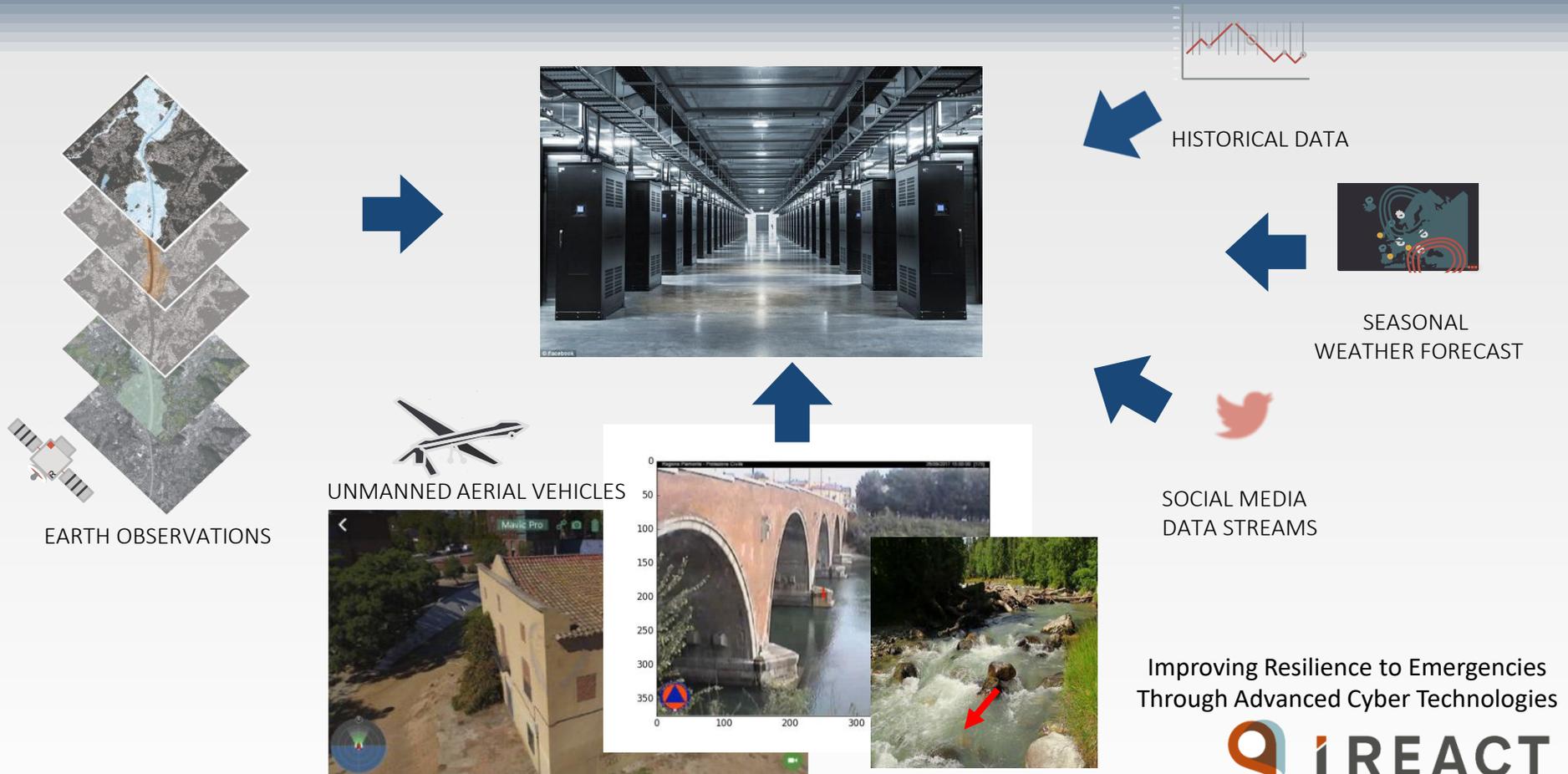
POLITECNICO DI TORINO



Data Base and Data Mining Group of Politecnico di Torino



# Emergency management



HISTORICAL DATA



SEASONAL  
WEATHER FORECAST



SOCIAL MEDIA  
DATA STREAMS

Improving Resilience to Emergencies  
Through Advanced Cyber Technologies

**i REACT**

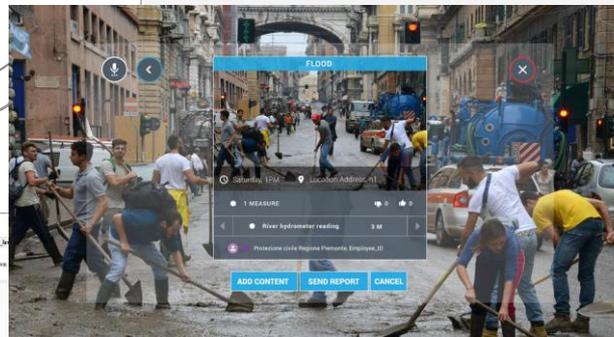
# Emergency management



FIRST RESPONDERS AND  
DECISION MAKERS



CITIZENS



Improving Resilience to Emergencies  
Through Advanced Cyber Technologies



# User engagement

2005



2013

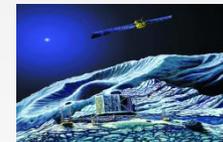


# Who generates big data?

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

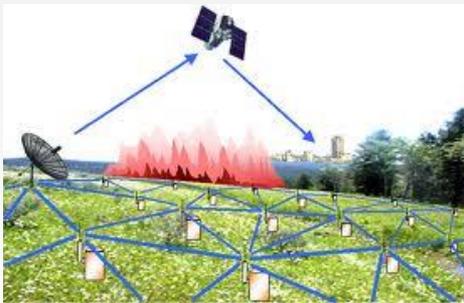


- Health and scientific computing



# 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?



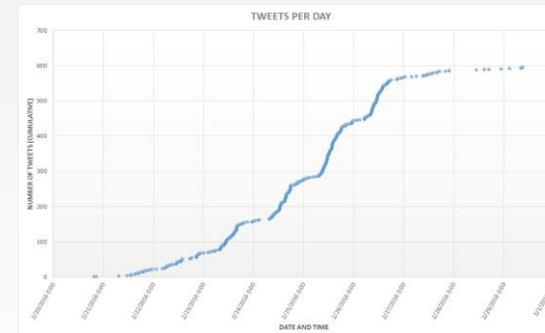
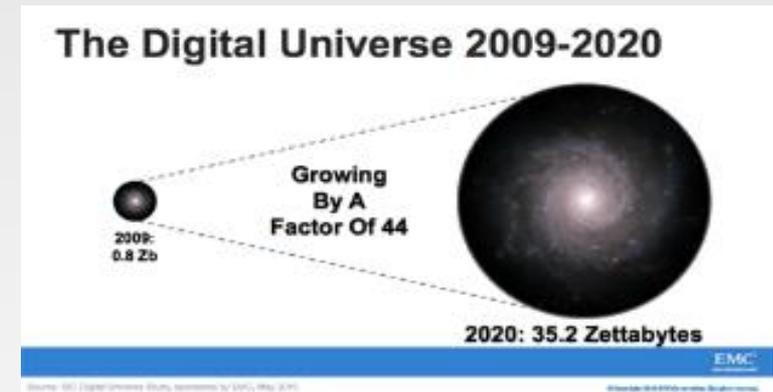
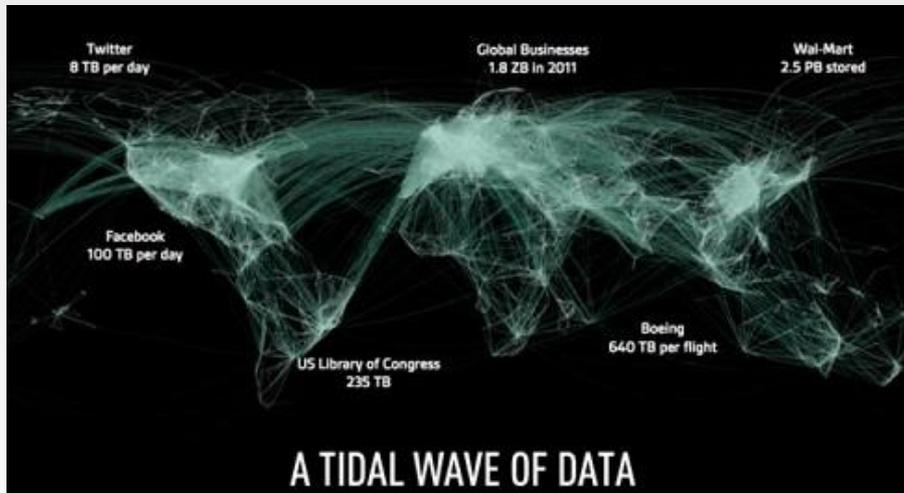
## □ Many different definitions

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

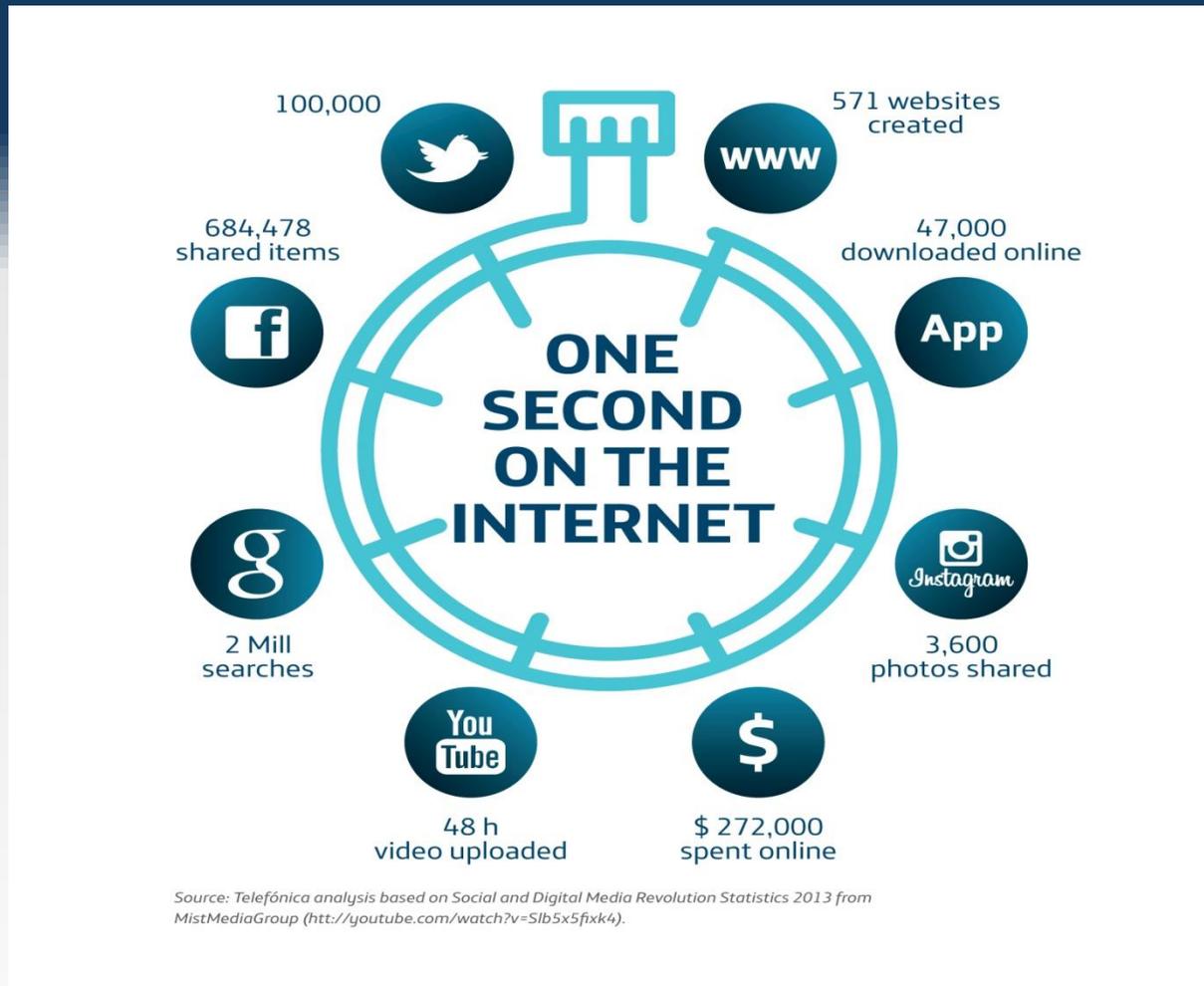


# The Vs of big data: Volume

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



# On the Internet...



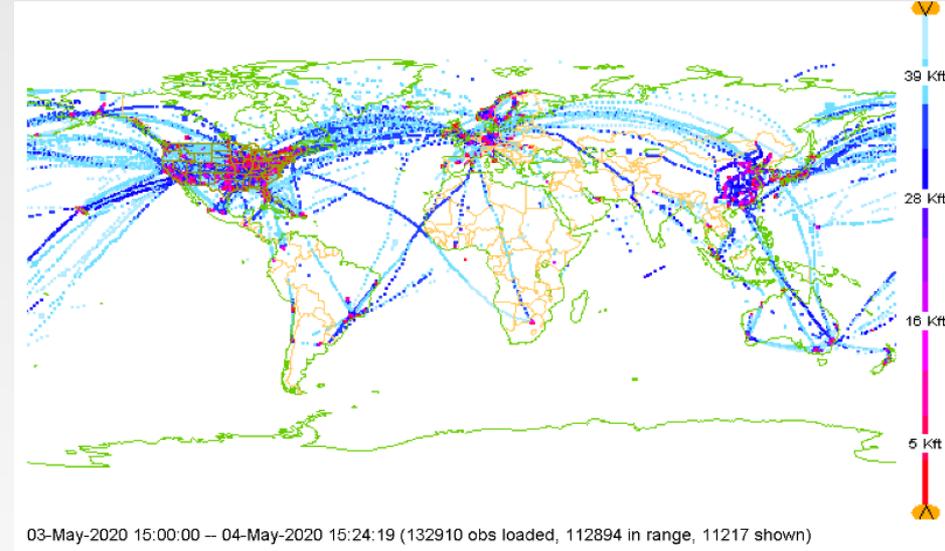
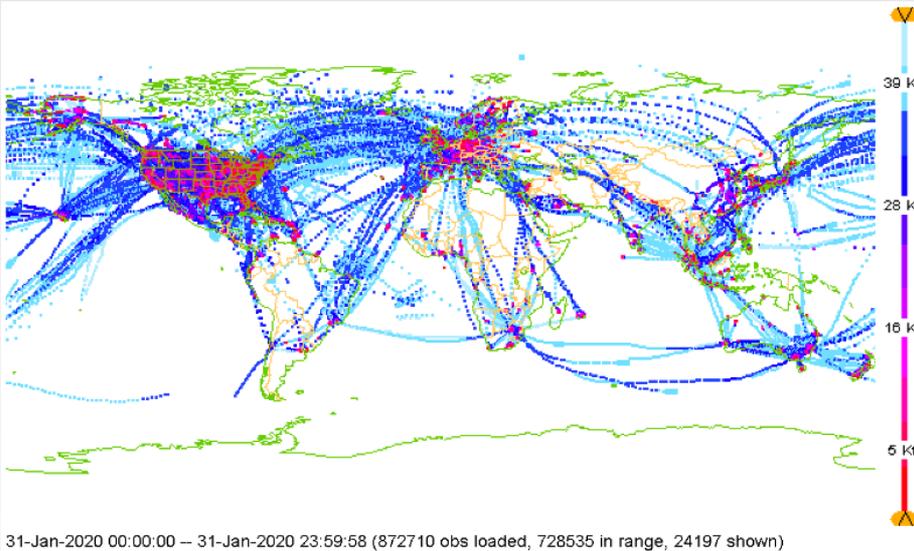
- <http://www.internetlivestats.com/>

# Weather forecast



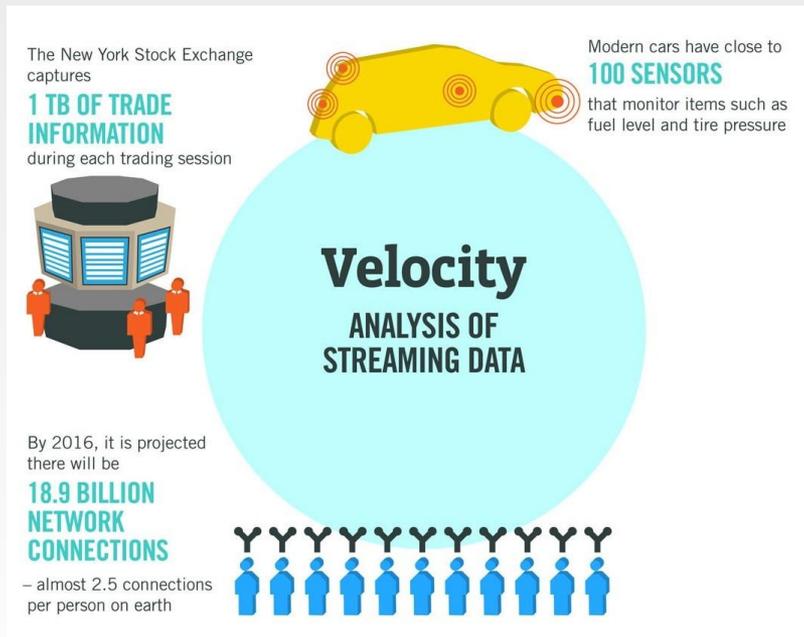
January 2020

May 2020

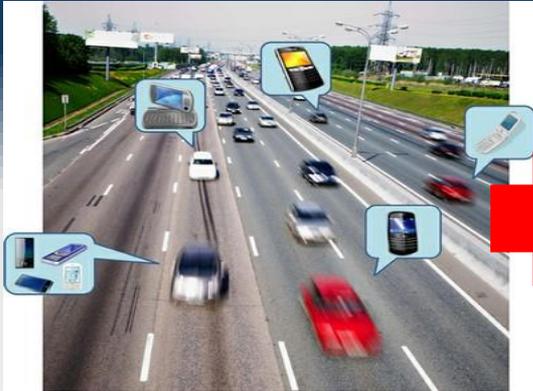


# The Vs of big data: **Velocity**

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



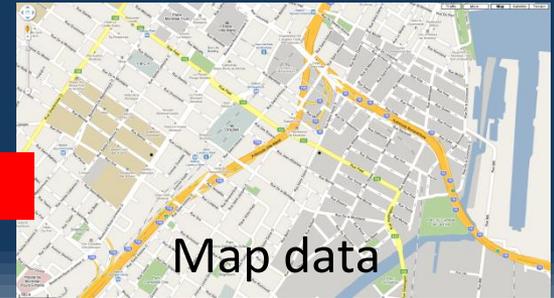
# (Near) Real time processing



Crowdsourcing



Computing

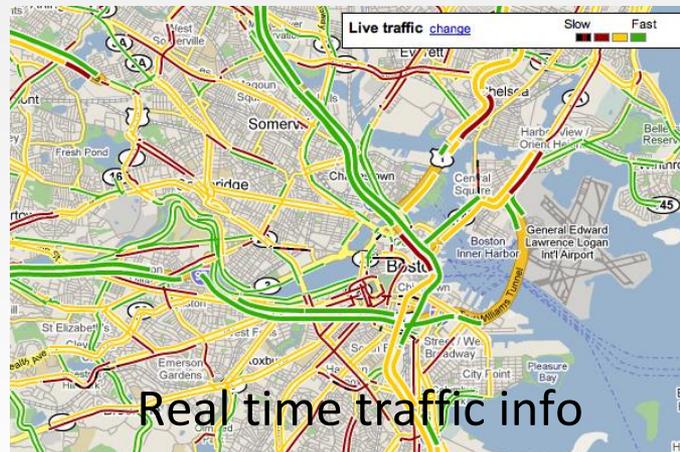


Map data



Wireless Sensor Networks

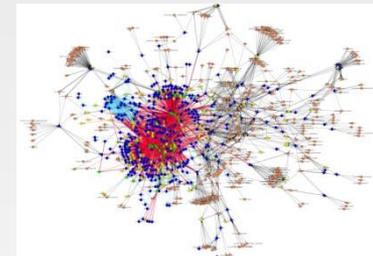
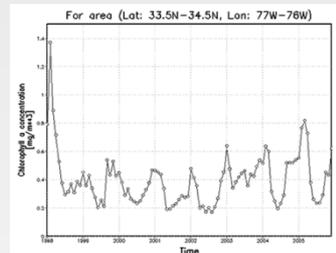
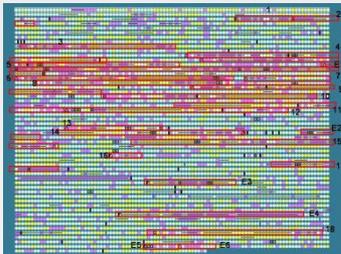
Sensing



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

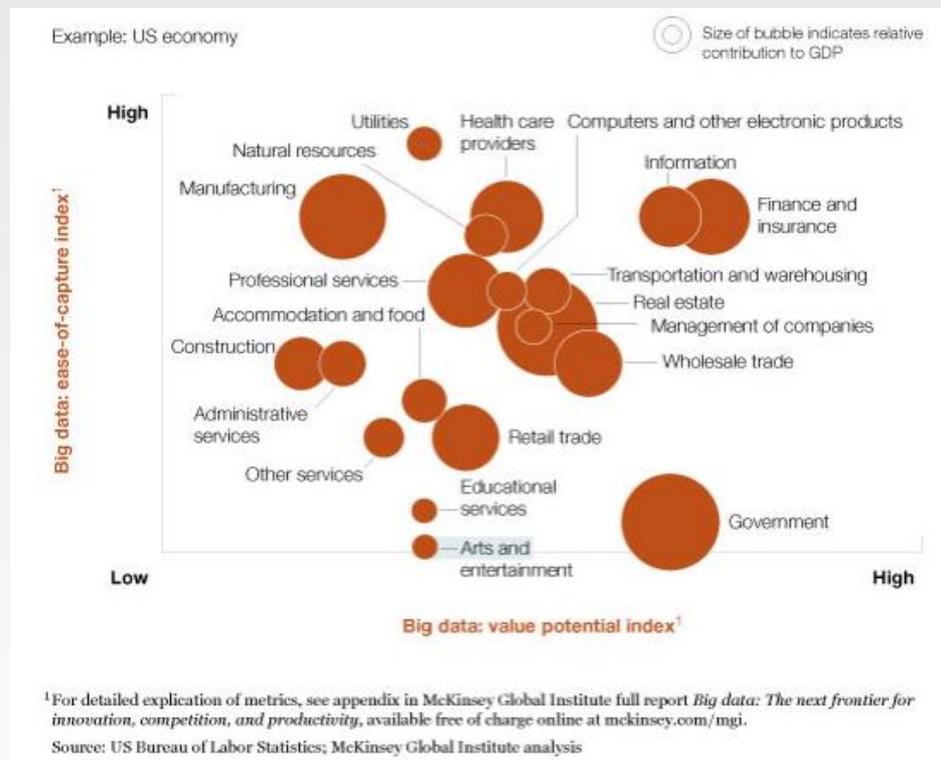
# The Vs of big data: **Veracity**

□ Data quality



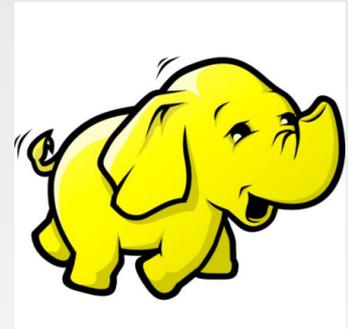
# The most important V: Value

Translate data into business advantage



# Big data challenges

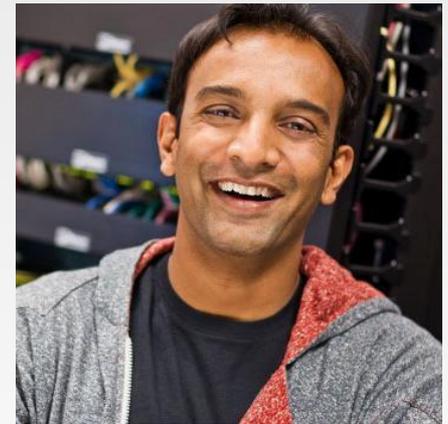
- ❑ Technology & infrastructure
  - ❑ New architectures, programming paradigms and techniques
    - Transfer the processing power to the data*
  - ❑ Apache Hadoop/Spark ecosystem
- ❑ Data management & analysis
  - ❑ New emphasis on “data”



**➔ *Data science***

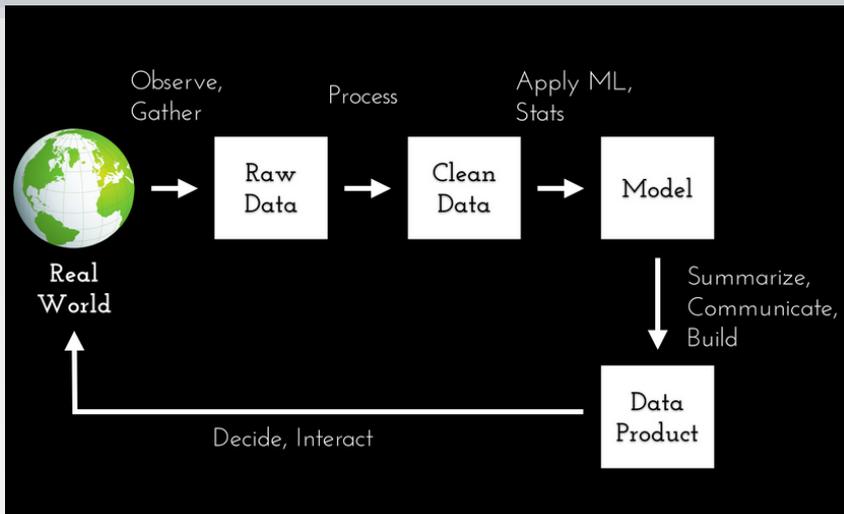
# Data science

“Extracting meaning from very large quantities of data”



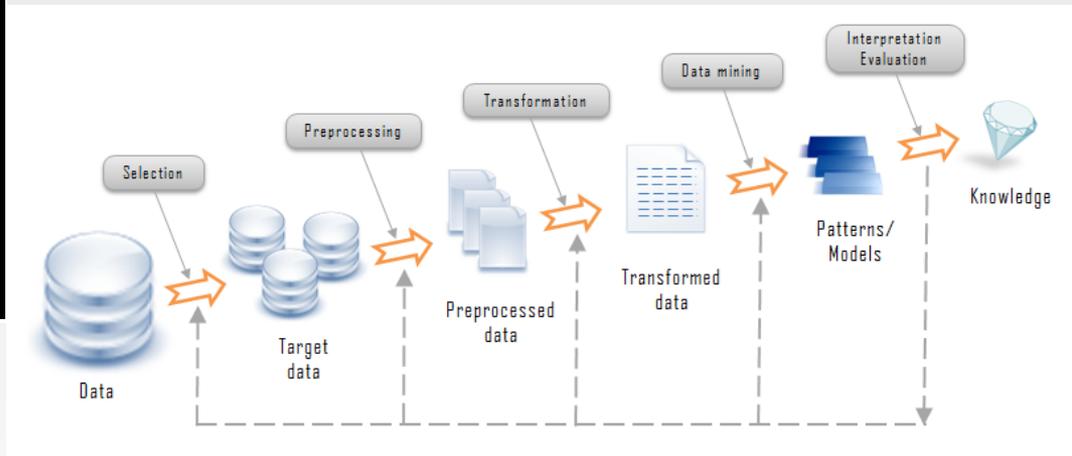
D.J. Patil coined the word *data scientist*

# The data science process



AKA **KDD** process

**K**nowledge **D**iscovery in **D**atabases



# 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)



# 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

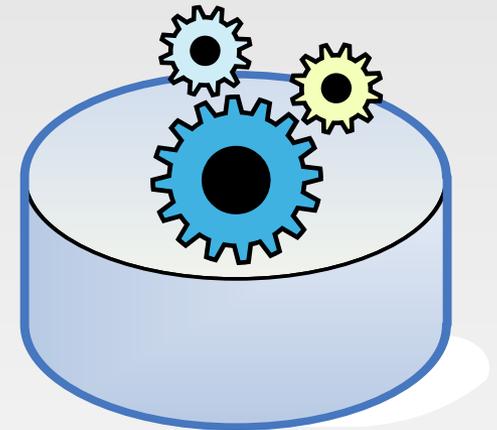
information from available data

- ❑ 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



twitter



- ❑ Maps and georeferenced data

- ❑ Localization, interesting locations for users



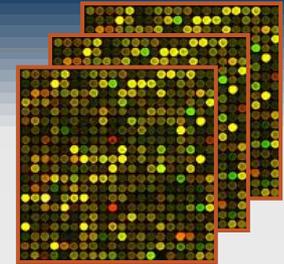
# Example: profiling

- ❑ 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

# Example: biological data

## Microarray

- expression level of genes in a cellular tissue
- various types (mRNA, DNA)



## Patient clinical records

- personal and demographic data
- exam results

CLID	PATIENT ID	shx013: 49A34	shv060: 45A9	shq077: 52A28	shx009: 4A34	shx014: 61A31	shq082: 99A6	shq083: 46A15	shx008: 41A31
IMAGE:74	ISG20    in	-1.02	-2.34	1.44	0.57	-0.13	0.12	0.34	-0.51
IMAGE:76	TNFSF13	-0.52	-4.06	-0.29	0.71	1.03	-0.67	0.22	-0.09
IMAGE:36	LOC93343	-0.25	-4.08	0.06	0.13	0.08	0.06	-0.08	-0.05
IMAGE:23	ITGA4    in	-1.375	-1.605	0.155	-0.015	0.035	-0.035	0.505	-0.865

## Textual data in public collections

- heterogeneous formats, different objectives
- scientific literature (PubMed)
- ontologies (Gene Ontology)



# Biological analysis objectives

## □ Clinical analysis

- detecting the causes of a pathology
- monitoring the effect of a therapy

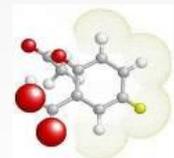
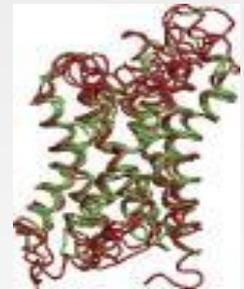
⇒ diagnosis improvement and definition of new specific therapies

## □ Bio-discovery

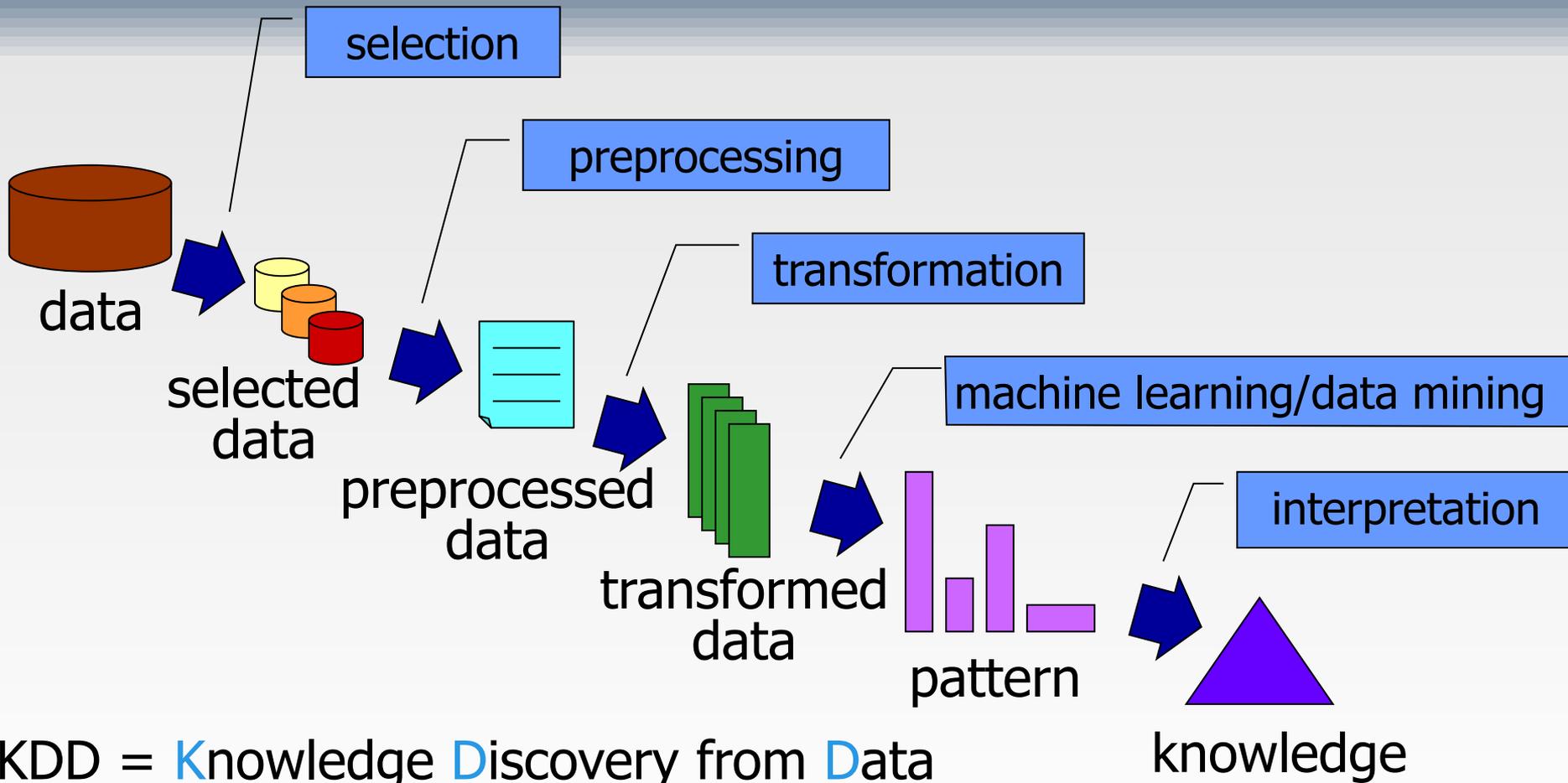
- gene network discovery
- analysis of multifactorial genetic pathologies

## □ Pharmacogenesis

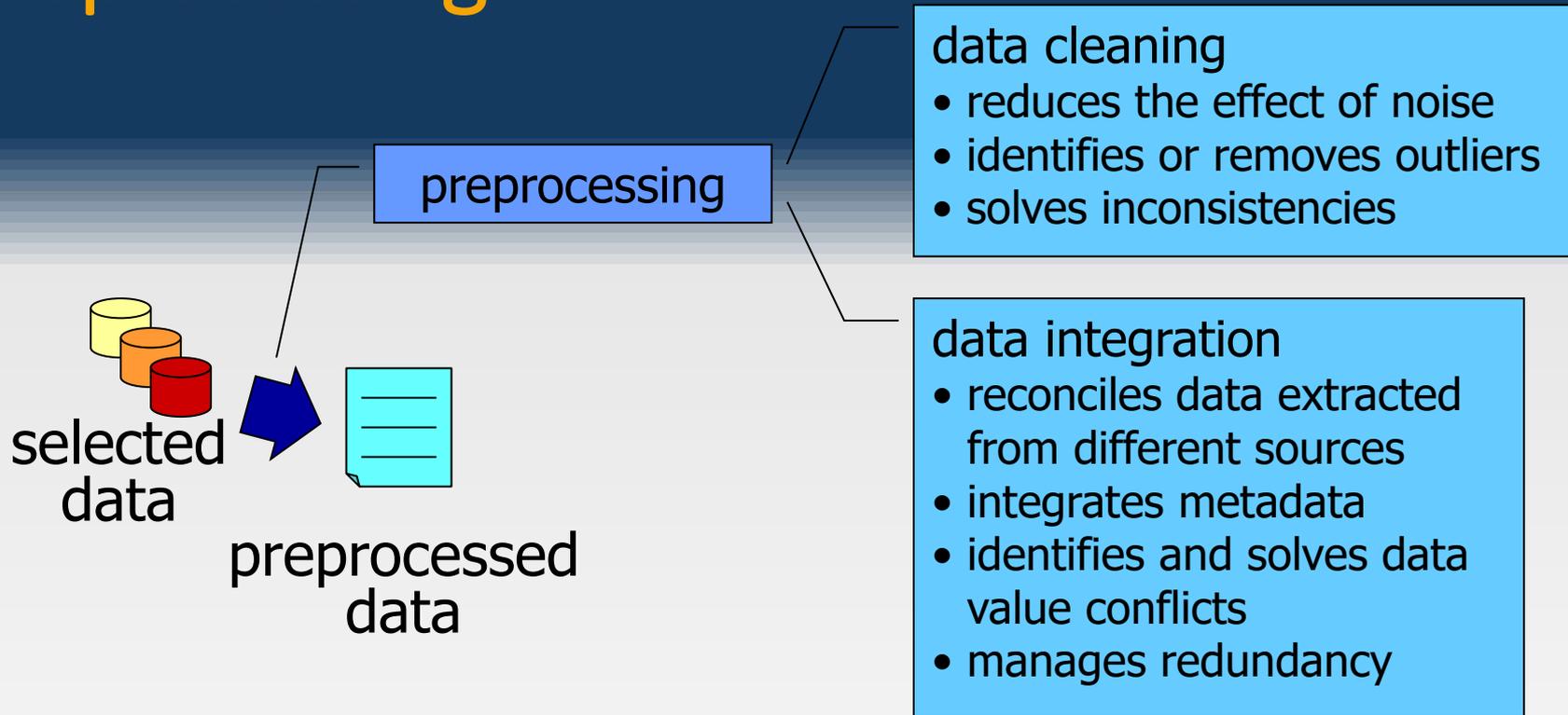
- lab design of new drugs for genic therapies



# Knowledge Discovery Process



# Preprocessing



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



## Frequently Bought Together



Price For All Three: £9.00

Add all three to Basket

Show availability and delivery details

- This item:** Paperback Oxford English Dictionary by Oxford Dictionaries Paperback £3.00
- Oxford Paperback Thesaurus by Oxford Dictionaries Paperback £3.00
- Oxford Essential French Dictionary by Oxford Dictionaries Paperback £3.00

## Jobs You May Be Interested In

Powered by  
LinkedIn



**Senior Data Analyst Job**  
Thomson Reuters - Bangalore, KA



**Data Scientist/ Senior Data Scientist**  
HeadHonchos.com - Bangalore - IN



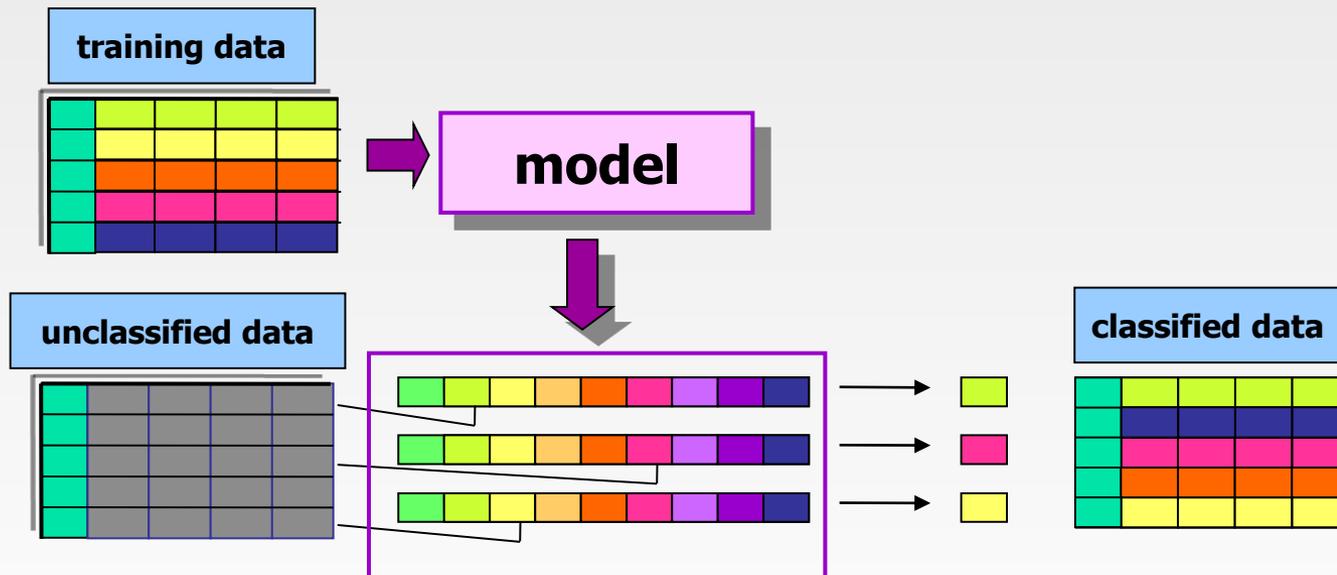
**Hiring Computer Scientist (Java) for...**  
Adobe - Noida



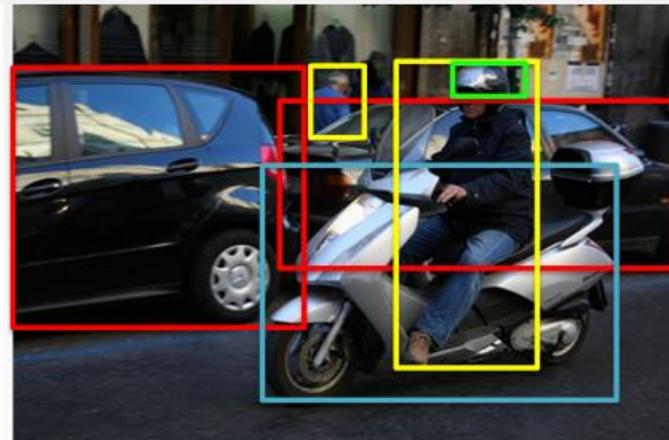
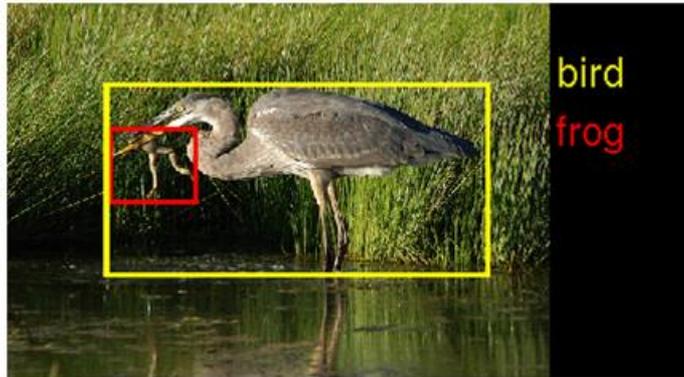
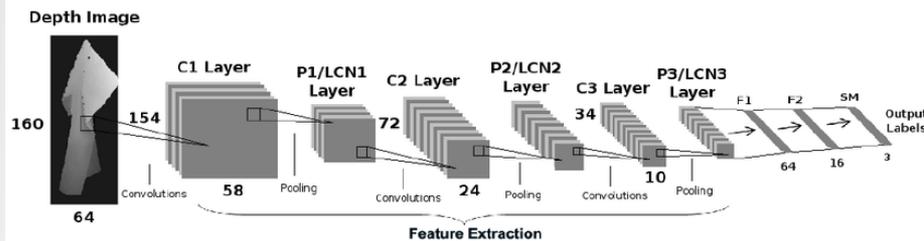
# Classification

## Objectives

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



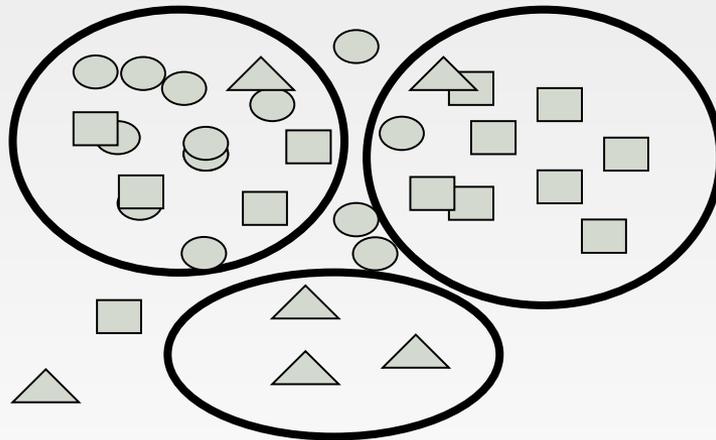
# Classification



Person  
Car  
Motorcycle  
Helmet

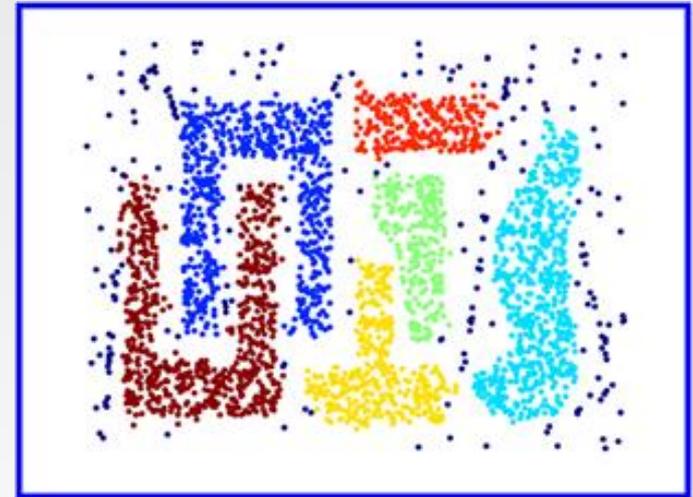
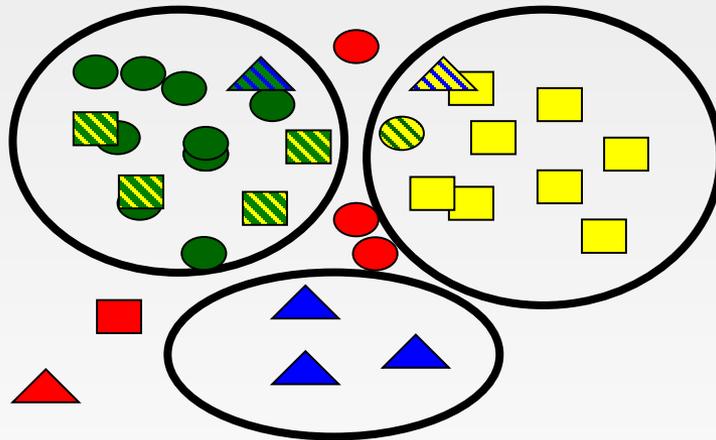
# Clustering

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

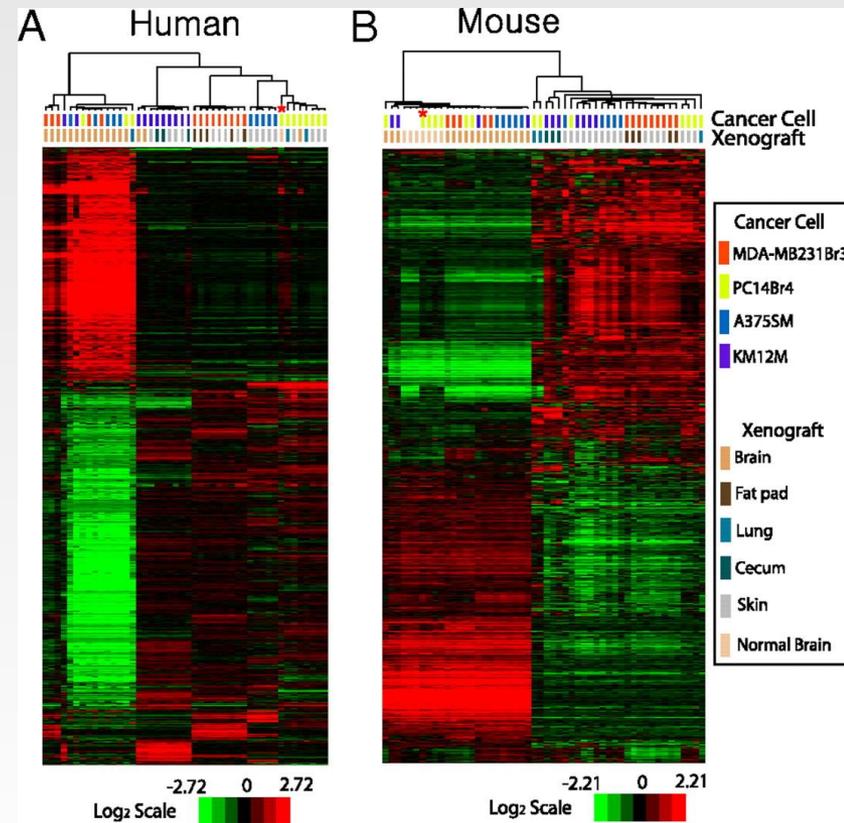
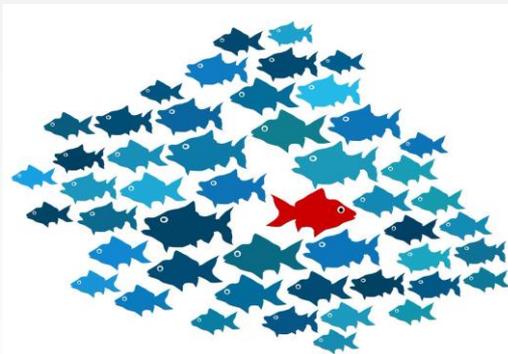


# Clustering

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



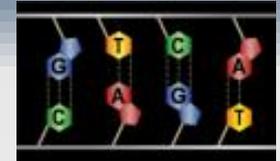
# Clustering



# Other data mining techniques

## Sequence mining

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



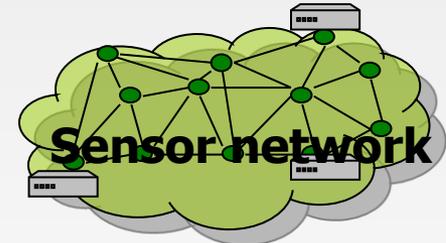
## Time series and geospatial data

- 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

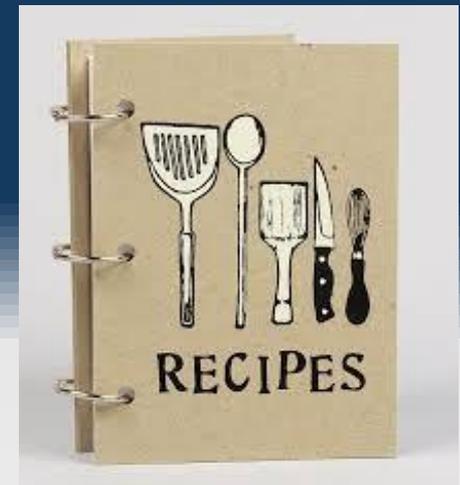


# The data science process

- 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?

# 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



## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

<b>MATH &amp; STATISTICS</b> <ul style="list-style-type: none"><li>☆ Machine learning</li><li>☆ Statistical modeling</li><li>☆ Experiment design</li><li>☆ Bayesian inference</li><li>☆ Supervised learning: decision trees, random forests, logistic regression</li><li>☆ Unsupervised learning: clustering, dimensionality reduction</li><li>☆ Optimization: gradient descent and variants</li></ul>	<b>PROGRAMMING &amp; DATABASE</b> <ul style="list-style-type: none"><li>☆ Computer science fundamentals</li><li>☆ Scripting language e.g. Python</li><li>☆ Statistical computing package e.g. R</li><li>☆ Databases: SQL and NoSQL</li><li>☆ Relational Algebra</li><li>☆ Parallel databases and parallel query processing</li><li>☆ MapReduce concepts</li><li>☆ Hadoop and Hive/Pig</li><li>☆ Custom reducers</li><li>☆ Experience with xaaS like AWS</li></ul>
<b>DOMAIN KNOWLEDGE &amp; SOFT SKILLS</b> <ul style="list-style-type: none"><li>☆ Passionate about the business</li><li>☆ Curious about data</li><li>☆ Influence without authority</li><li>☆ Hacker mindset</li><li>☆ Problem solver</li><li>☆ Strategic, proactive, creative, innovative and collaborative</li></ul>	<b>COMMUNICATION &amp; VISUALIZATION</b> <ul style="list-style-type: none"><li>☆ Able to engage with senior management</li><li>☆ Story telling skills</li><li>☆ Translate data-driven insights into decisions and actions</li><li>☆ Visual art design</li><li>☆ R packages like ggplot or lattice</li><li>☆ Knowledge of any of visualization tools e.g. D3.js, Tableau</li></ul>

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics and econometrics, data visualization and big data systems, marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

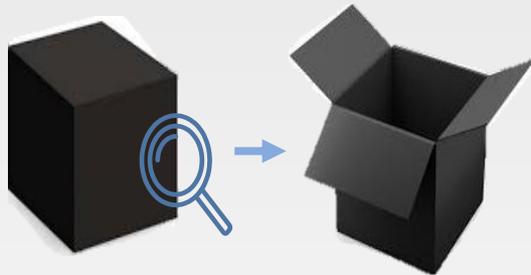
# Open issues

- ❑ Social impact of analysis is very important
  - ❑ Interpretability and transparency of the analysis process
  - ❑ Bias in algorithms and data
  - ❑ Privacy preservation



# 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 patients dataset from USA hospital
  - history of asthma → lower chance of dying from pneumonia*
- ❑ MD consider asthma as a serious risk factor for people who get pneumonia
- ❑ 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 of 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

# Privacy

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

The screenshot shows the Strava website interface. At the top, there's a navigation bar with 'STRAVA LABS' and links for 'Projects', 'Blog', 'Developers', 'Strava.com', and 'Careers'. Below this is a 'Global Heatmap' section with a 'Heatmap Color' dropdown and zoom controls. The main content area features a news article with the headline 'Fitness tracking app Strava gives away location of secret US army bases'. The article includes a sub-headline 'Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities' and a link to a 'Latest' update: 'Strava suggests military users 'opt out' of heatmap as row deepens'. Navigation tabs for 'Opinion', 'Sport', 'Culture', 'Lifestyle', and 'More' are visible, along with regional filters like 'IRAQ' and 'AFGHANISTAN'. A URL bar shows the Guardian article link: 'https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases'.

This screenshot shows a BBC News article. The top navigation bar includes the BBC logo, a search icon, and categories like 'Mark', 'News', 'Sport', 'Weather', 'iPlayer', and 'TV'. The article title is 'Fitness app Strava lights up staff at military bases', dated '29 January 2018'. Below the title are social media sharing icons for Facebook, Twitter, LinkedIn, Email, and a general 'Share' button.



The movements of soldiers within Bagram air base - the largest US military facility in Afghanistan

Security concerns have been raised after a fitness tracking firm showed the exercise routes of military personnel in bases around the world.

# Open issues

- ❑ Social impact of analysis is very important
  - ❑ Interpretability and transparency of the analysis process
  - ❑ Privacy preservation
- ❑ Many technical issues are not solved
  - ❑ Scalability to *huge* data volumes
  - ❑ Data dimensionality
  - ❑ Complex data structures, heterogeneous data formats
  - ❑ Data quality
  - ❑ Streaming data