Data science process

Prof. Marco Mellia

POLITECNICO DI TORINO

SmartData@PoliTO

SmartData

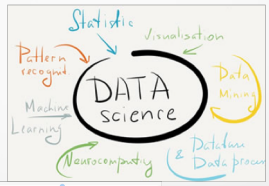
POLITECNICO DI TORINO

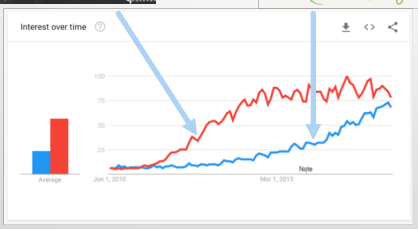
ICT4SS
Information and Communications Technologies for Smart Societies

Big data hype?

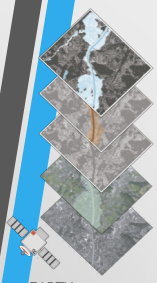
Big Data & Data Science

"Extracting **meaning** from very large quantities of **data**"





Emergency management

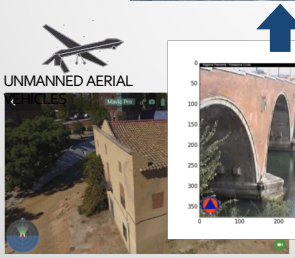


EARTH OBSERVATIONS

HISTORICAL DATA

SEASONAL WEATHER FORECAST

SOCIAL MEDIA DATA STREAMS



UNMANNED AERIAL VEHICLES

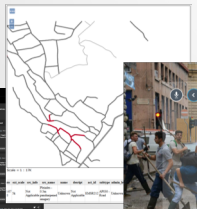
Improving Resilience to Emergencies Through Advanced Cyber Technologies

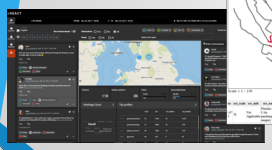
iREACT

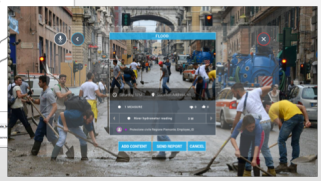
Emergency management

CITIZENS

FIRST RESPONDERS AND DECISION MAKERS







Improving Resilience to Emergencies Through Advanced Cyber Technologies

iREACT

User engagement

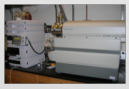
2005

2013

Who generates big data?

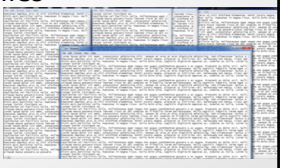
- User Generated Content (Web & Mobile)
  - E.g., Facebook, Instagram, Yelp, TripAdvisor, Twitter, YouTube

- Health and scientific computing

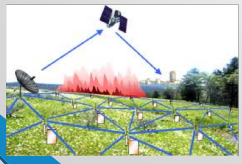


Who generates big data?

- 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?

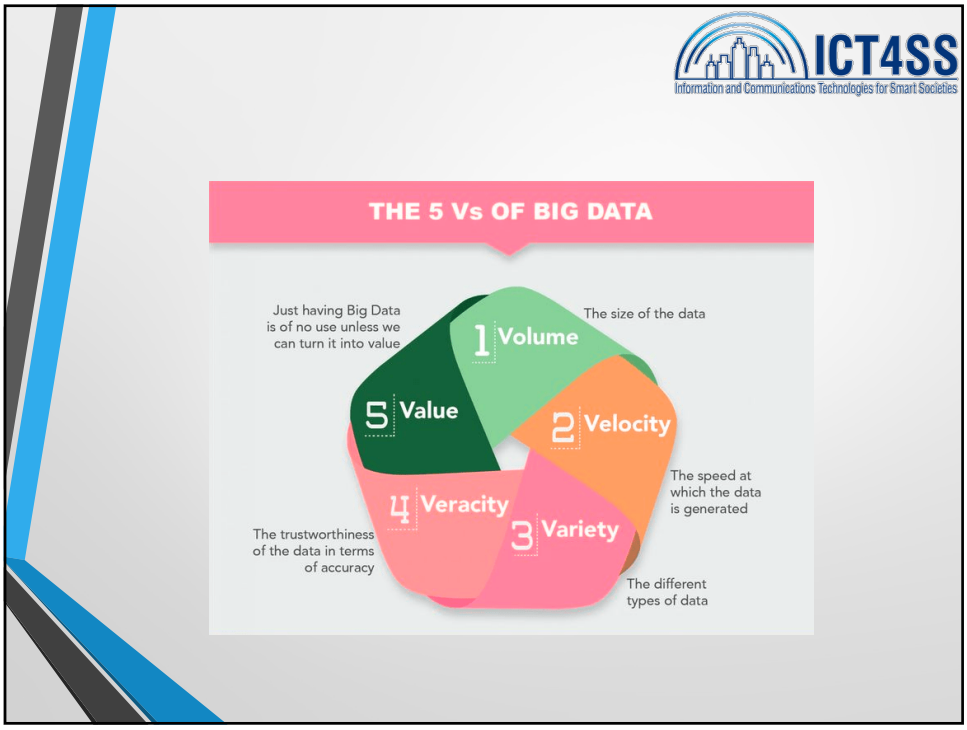
- Many different definitions

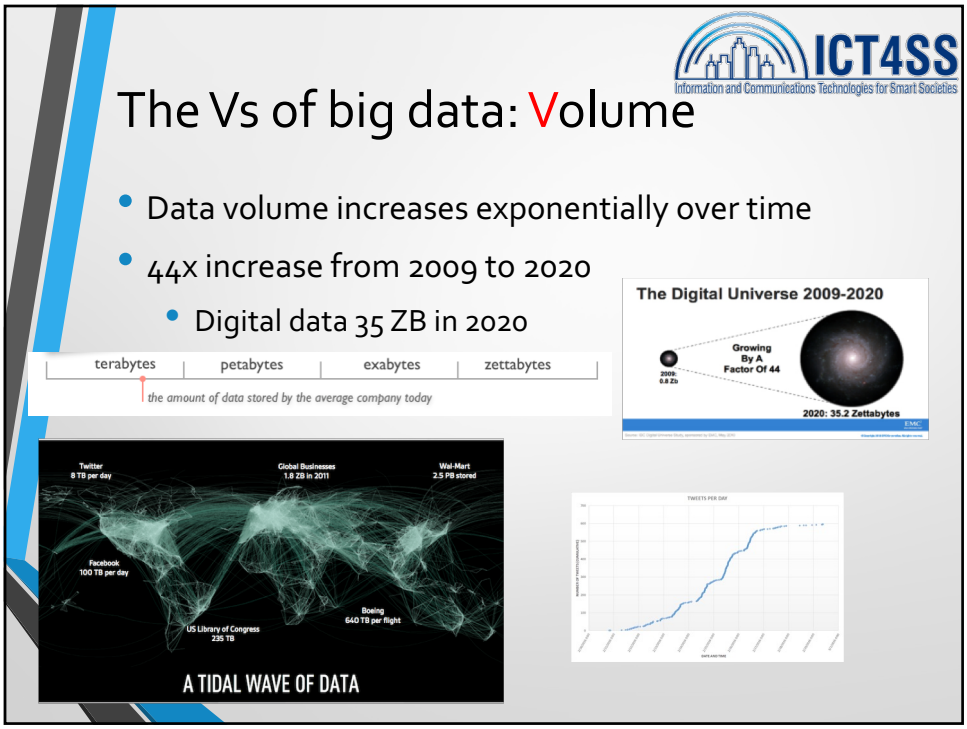
*"Data whose scale, diversity and complexity require new **architectures**, **techniques**, **algorithms** and **analytics** to manage it and extract value and hidden knowledge from it"*

What is big data?

- Many different definitions

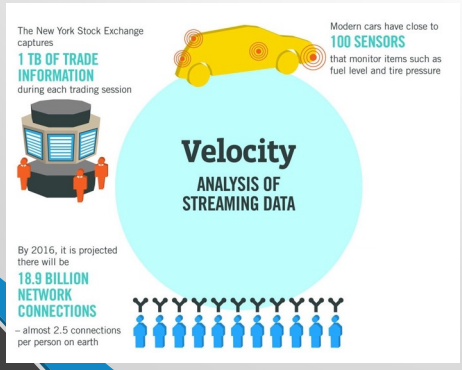
*"Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract **value** and hidden **knowledge** from it"*


ICT4SS
Information and Communications Technologies for Smart Societies
THE 5 Vs OF BIG DATA
Just having Big Data is of no use unless we can turn it into value
1 Volume
The size of the data
5 Value
2 Velocity
The speed at which the data is generated
4 Veracity
3 Variety
The trustworthiness of the data in terms of accuracy
The different types of data

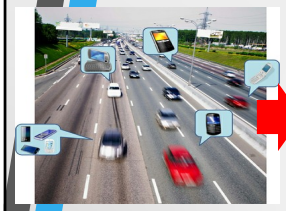

ICT4SS
Information and Communications Technologies for Smart Societies
The Vs of big data: Volume
Data volume increases exponentially over time
44x increase from 2009 to 2020
Digital data 35 ZB in 2020
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today
The Digital Universe 2009-2020
Growing By A Factor Of 44
2020: 35.2 Zettabytes
Twitter 8 TB per day
Global Businesses 1.8 ZB in 2011
Wal-Mart 2.5 PB stored
Facebook 100 TB per day
US Library of Congress 235 TB
Boeing 640 TB per flight
A TIDAL WAVE OF DATA

The Vs of big data: Velocity

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

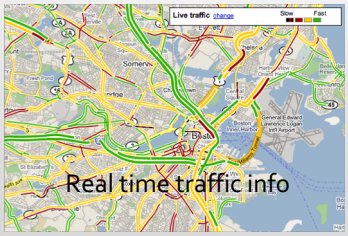


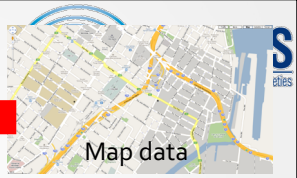
Real time processing



Computing

Crowdsourcing

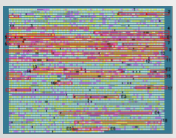


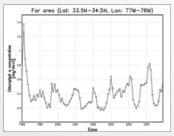


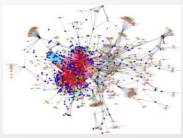
Sensing

The Vs of big data: Variety

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







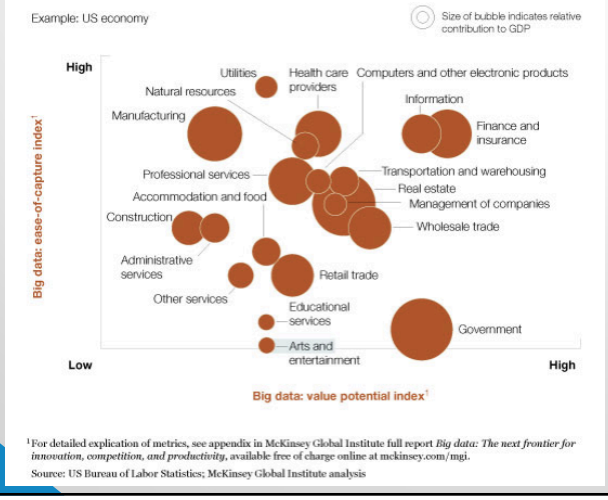
- A single application may generate many different formats

The Vs of big data: Veracity

- Data quality

The most important V: Value

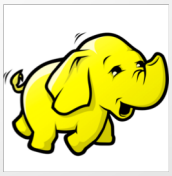
- Translate data into business advantage



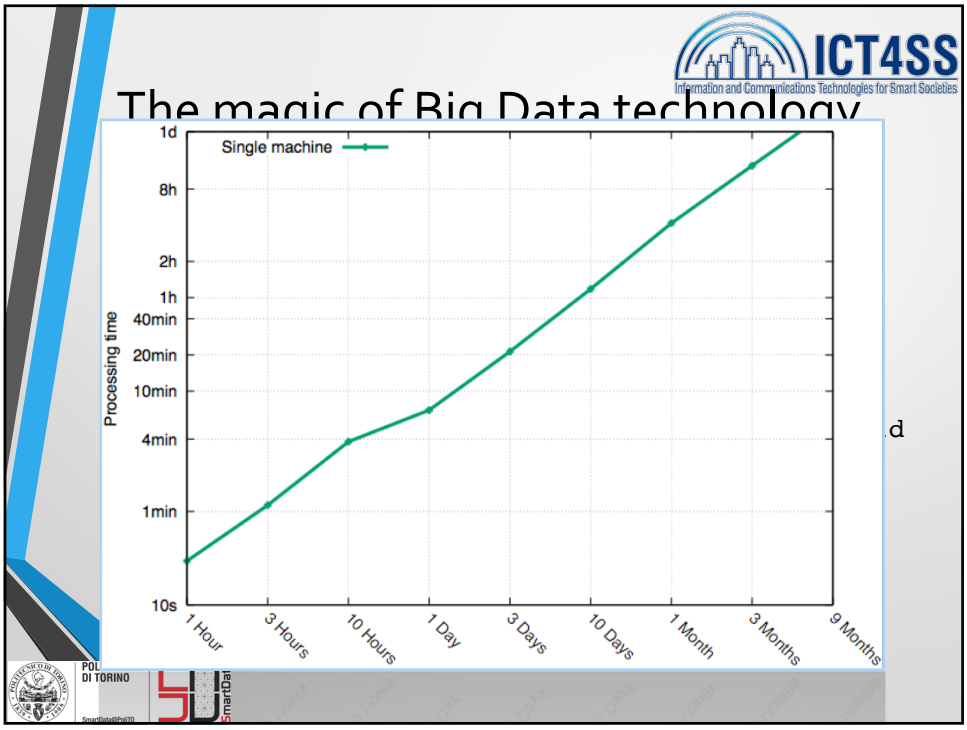
Big data challenges

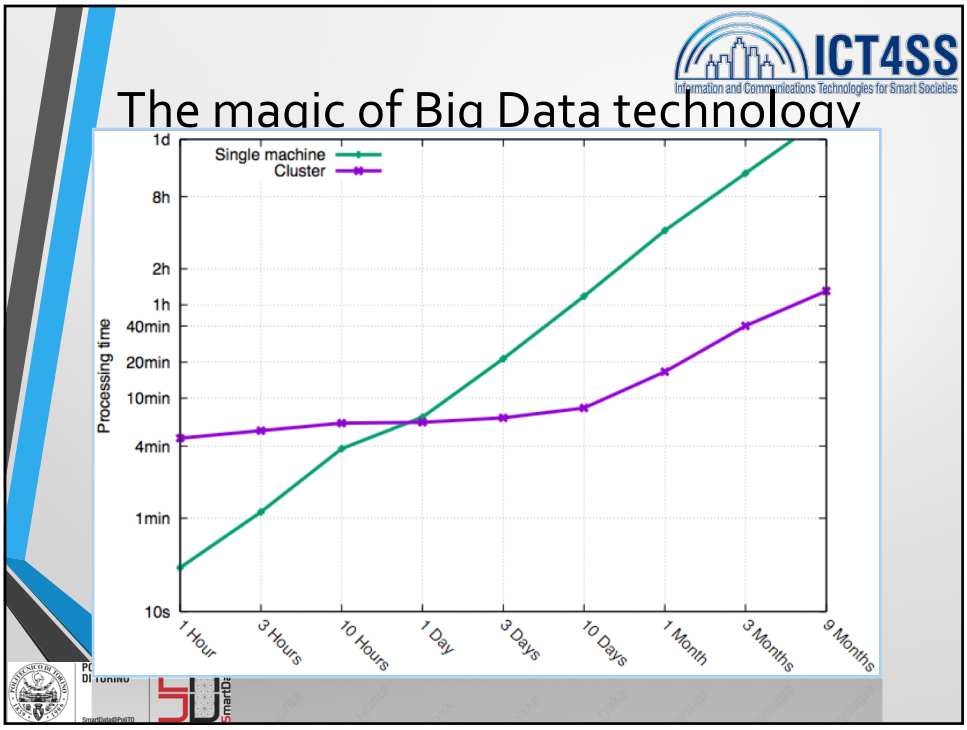
- Technology & infrastructure
  - New architectures, programming paradigms and techniques

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

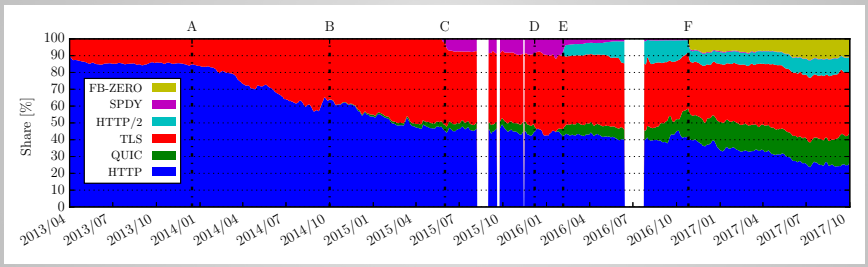


Data science


The magic of Big Data technology
ICT4SS
Information and Communications Technologies for Smart Societies
Single machine
Processing time
1d
8h
2h
1h
40min
20min
10min
4min
1min
10s
1 Hour
3 Hours
10 Hours
1 Day
3 Days
10 Days
1 Month
3 Months
9 Months

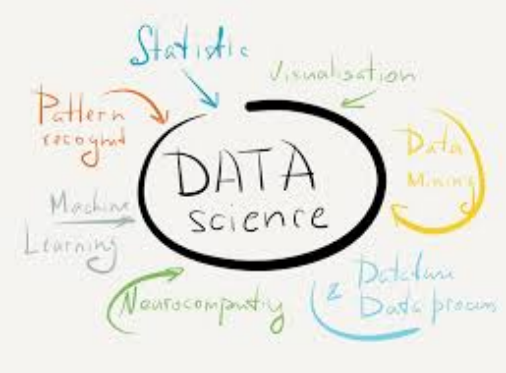

The magic of Big Data technology
ICT4SS
Information and Communications Technologies for Smart Societies
Single machine
Cluster
Processing time
1d
8h
2h
1h
40min
20min
10min
4min
1min
10s
1 Hour
3 Hours
10 Hours
1 Day
3 Days
10 Days
1 Month
3 Months
9 Months

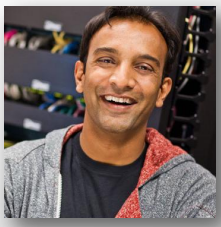
Protocol evolution in 5 years



Data science

"Extracting **meaning** from very large quantities of **data**"



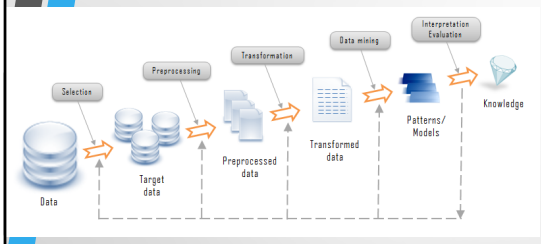


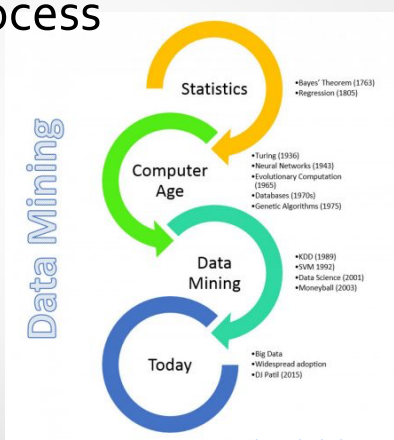
D.J. Patil coined the word *data scientist*

The data science process

AKA *KDD* process

Knowledge Discovery in Databases





https://hackerbits.com

Generation → Acquisition → Storage → Analysis

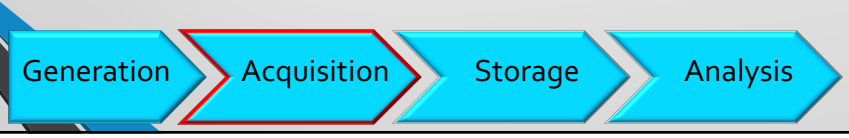
Generation

- 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)

Generation → Acquisition → Storage → Analysis

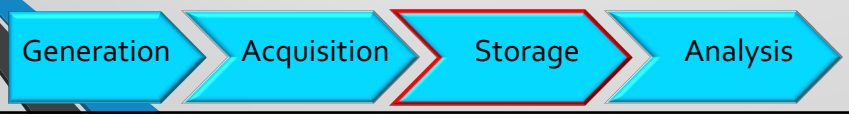
Acquisition

- 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

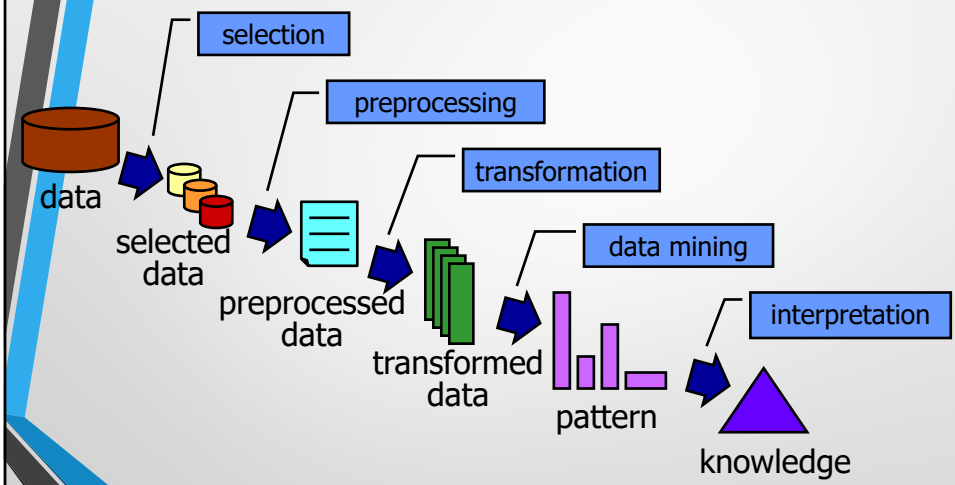


Analysis

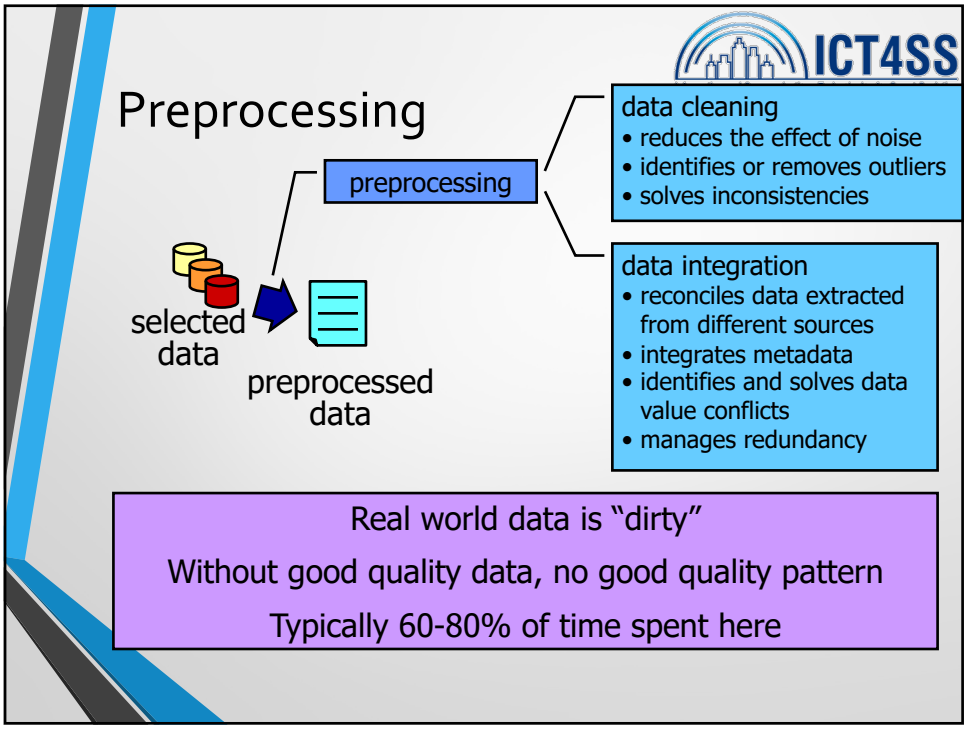
- Objectives
  - Descriptive analytics, predictive analytics, prescriptive analytics
- Methods
  - Statistical analysis, data mining, text mining, network and graph data mining
  - Association analysis, classification and regression, clustering
- Diverse domains call for customized techniques

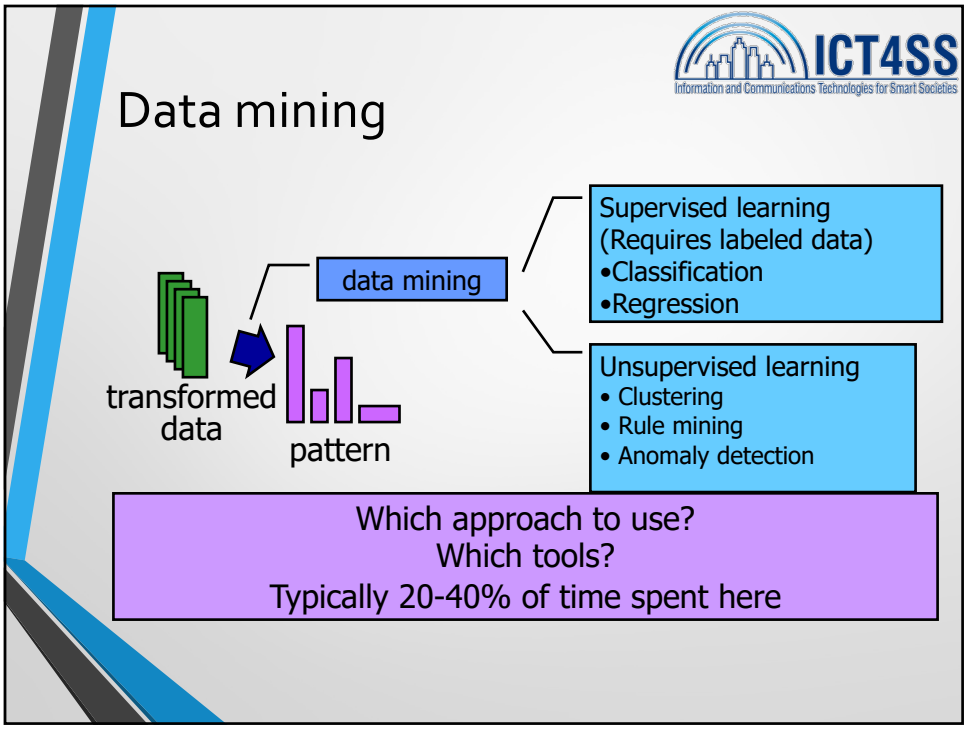
Generation → Acquisition → Storage → Analysis

Knowledge Discovery Process



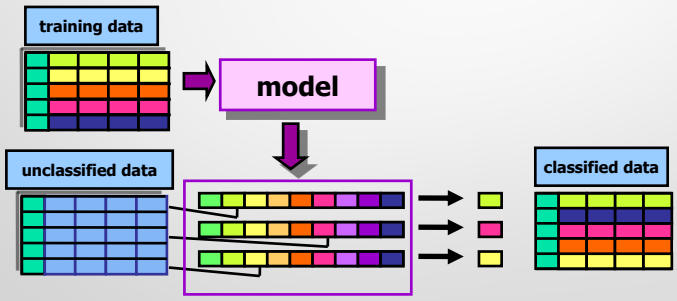
KDD = Knowledge Discovery from Data


ICT4SS
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
selected data
preprocessed data
Real world data is "dirty"
Without good quality data, no good quality pattern
Typically 60-80% of time spent here


ICT4SS
Information and Communications Technologies for Smart Societies
Data mining
data mining
Supervised learning
(Requires labeled data)
•Classification
•Regression
Unsupervised learning
• Clustering
• Rule mining
• Anomaly detection
transformed data
pattern
Which approach to use?
Which tools?
Typically 20-40% of time spent here

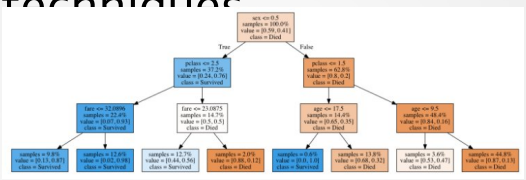
Classification

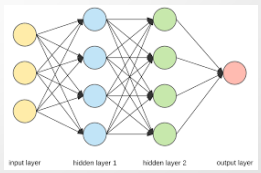
- Objectives
  - prediction of a class label
  - definition of an interpretable model of a given phenomenon

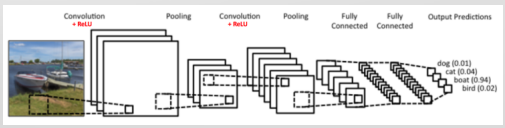


Classification techniques

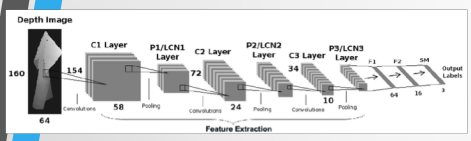
- Decision trees
- Classification rules
- Association rules
- Neural Networks
- Naïve Bayes and Bayesian Networks
- k-Nearest Neighbours (k-NN)
- Support Vector Machines (SVM)
- …

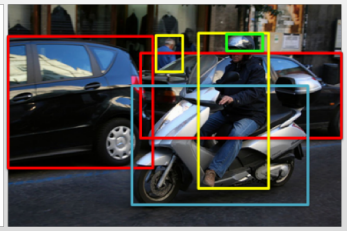






Classification





Person
Car
Motorcycle
Helmet

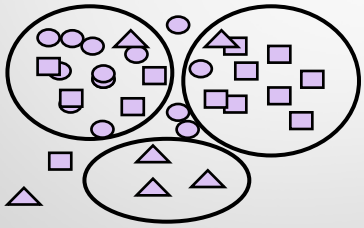
Evaluation of classification techniques

- Accuracy
  - quality of the prediction
- Interpretability
  - model interpretability
  - model compactness
- Robustness
  - noise, missing data

- Efficiency
  - model building time
  - classification time
- Scalability
  - training set size
  - attribute number

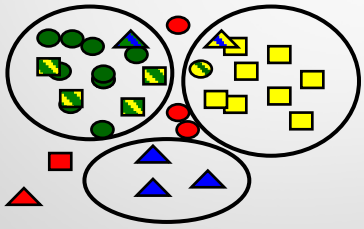
Clustering

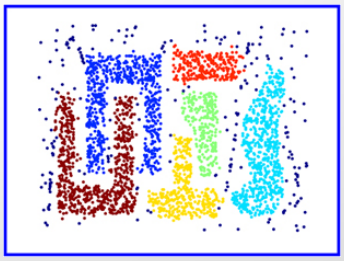
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers



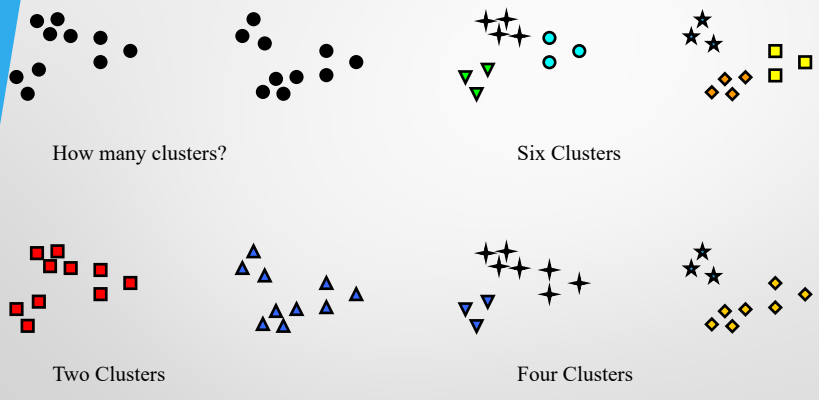
Clustering

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





Notion of a Cluster can be Ambiguous



How many clusters?

Six Clusters

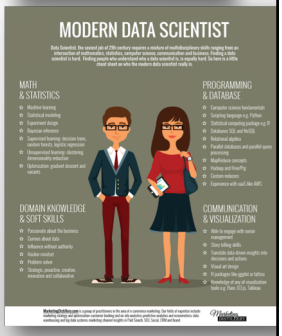
Two Clusters

Four Clusters

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

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 application domain
  - Business expert
    - Data driven decisions, new business models



Applications

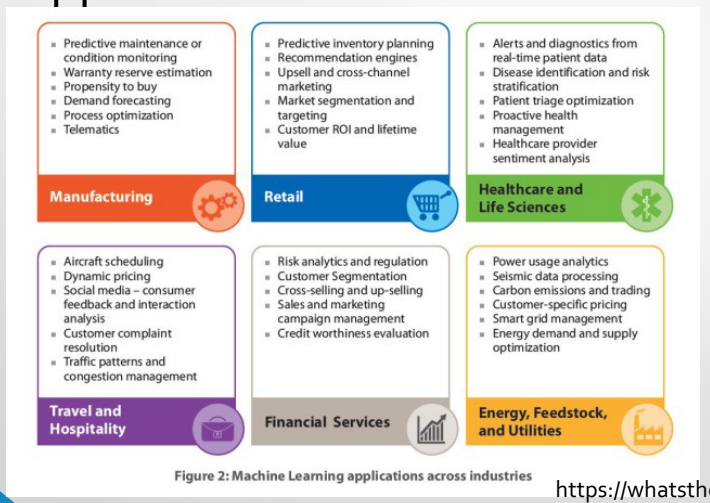
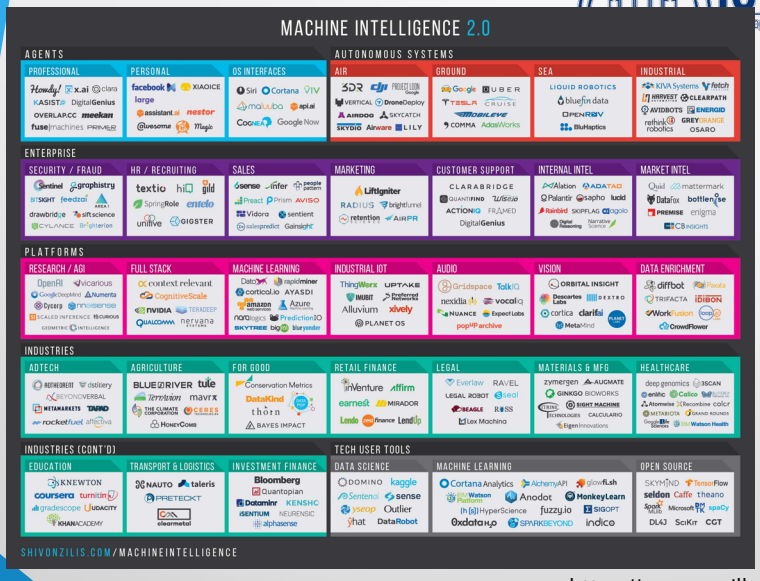


Figure 2: Machine Learning applications across industries

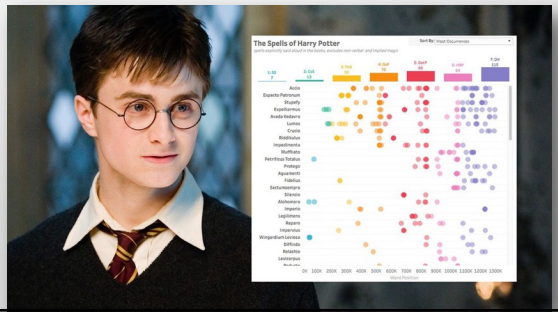
https://whatsthebigdata.com/



https://www.oreilly.com/

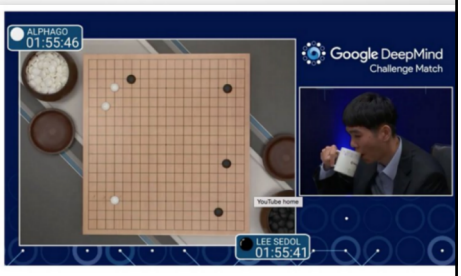
Conclusions and open issues

- We are leaving in a data deluge epoch
  - Getting data is easier and easier
- **Extracting value from the data is still a challenge**
  - **Don't blindly trust the magician – no Harry Potter stick here**



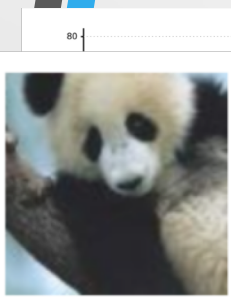
Is it for free?

AlphaGo

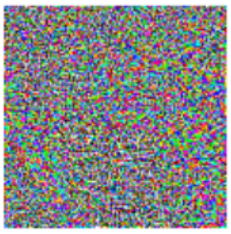


AlphaGO	Lee Se-dol
1202 CPUs, 176 GPUs, 100+ Scientists.	1 Human Brain 1 Coffee.
1 MW	30W

better than humans?



$+.007 \times$



$=$

x

"panda"
57.7% confidence

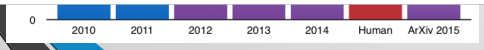
$\text{sign}(\nabla_x J(\theta, x, y))$

"nematode"
8.2% confidence

$x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$

"gibbon"
99.3 % confidence

Adversarial images can cause major problems



Google Flu trends

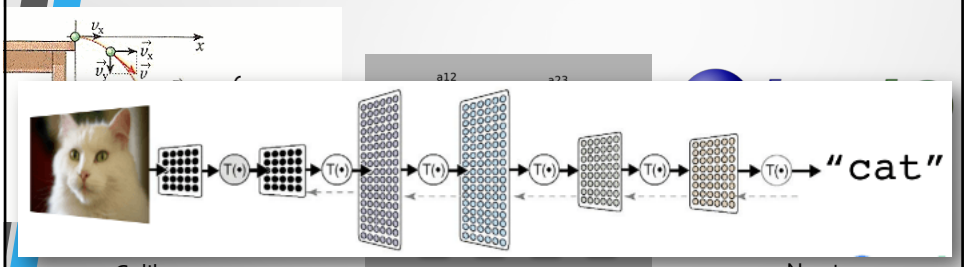
- February 2010
  - detected flu outbreak two weeks ahead of CDC data

Discontinued!

Conclusions and open issues

- We are leaving in a data deluge epoch
  - Getting data is easier and easier
- Extracting value from the data is still a challenge
  - Don't blindly trust the magician – no Harry Potter stick here
- **Humans need to understand a process**

Causality and models



Galileo

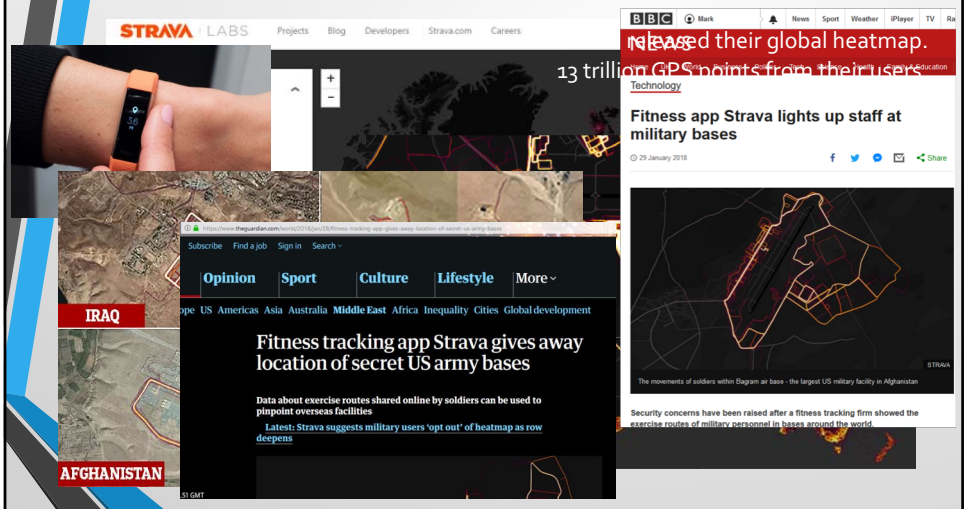
Newton

Hidden Markov Model

Conclusions and open issues

- We are leaving in a data deluge epoch
  - Getting data is easier and easier
- Extracting value from the data is still a challenge
  - Don't blindly trust the magician – no Harry Potter stick here
- Humans need to understand a process
- **Social impact of analysis is very important**
  - Interpretability and transparency of the analysis process
  - Privacy preservation
  - Modification of users'habits

Privacy



Impact on our habits

- Listen to this
- 10 most popular questic
  - People cannot read a n
- People become harsh
  - OK google, take me hc
    - Ok, here is your pat
  - Hey Siri, call mom
    - Ok, calling mom
  - Alexa, buy my pills,
    - Ok, done
- And Kids ?