

# Data Mining M

## Machine Learning and Data Mining

### Introduction

---

Claudio Sartori

DISI

Department of Computer Science and Engineering – University of Bologna, Italy  
[claudio.sartori@unibo.it](mailto:claudio.sartori@unibo.it)

1	General information	2
2	Data in organisations	6
3	Data Mining and Big Data	31
4	Cloud Computing	56
5	Big Data	62
6	Data Mining	74

# Context

- Exam **B2125-Machine Learning and Data Mining M**  
*Master program Computer Engineering, second year*
- Module **95631-Machine Learning and Data Mining**  
part of Machine Learning and Deep Learning (i.c.)  
*Master program Artificial Intelligence, first year*
- Official course site
  - <https://www.unibo.it/sitoweb/claudio.sartori/didattica>
- E-learning site on <https://virtuale.unibo.it>
  - can be reached following the link from the official course site
  - students having the course in their study plan can self-enrol

# What's in this Course

## ● Part 1 – Data Mining

We will focus on the **data** side, studying the **enabling technologies** which have been developed for other purposes, but can positively influence the success of data mining processes

- Data Warehouse
- Data Lake
- Software Architectures for Data Mining processes

## ● Part 2 – Machine Learning

We will focus on the techniques that support **data–driven decisions**

- Data preparation
- Classification
- Regression
- Clustering
- Association rules

# Insight

Education is not the piling on of learning, information, data, facts, abilities or skills – that's training or instruction – but is rather making visible what is hidden as a seed

Thomas More<sup>1</sup>

---

1 Cited by Charu C. Aggarwal in his book “Data Mining – the Textbook”

1	General information	2
2	Data in organisations	6
3	Data Mining and Big Data	31
4	Cloud Computing	56
5	Big Data	62
6	Data Mining	74

# Data, Data Mining and Machine Learning

- Data **exists** independently from Data Mining and Machine Learning
  - but you **need** Data Mining and Machine Learning techniques to derive interesting and **actionable** insights
- Data Mining and Machine Learning were created long before the dramatic increase of the amount of data available
  - the increase of the amount of data **strengthen** DM and ML relevance and **economic impact**

# Big Data

A new player with Data Mining and Machine Learning

- Big Data exists independently from Data Mining and Machine Learning
  - but you need Data Mining and Machine Learning techniques to effectively analyse and use Big Data
- Data Mining and Machine Learning were created long before the existence of Big Data
  - but using them on Big Data greatly increase DM and ML relevance and economic impact

Data → Information → Knowledge ⇒ better, data driven, decisions

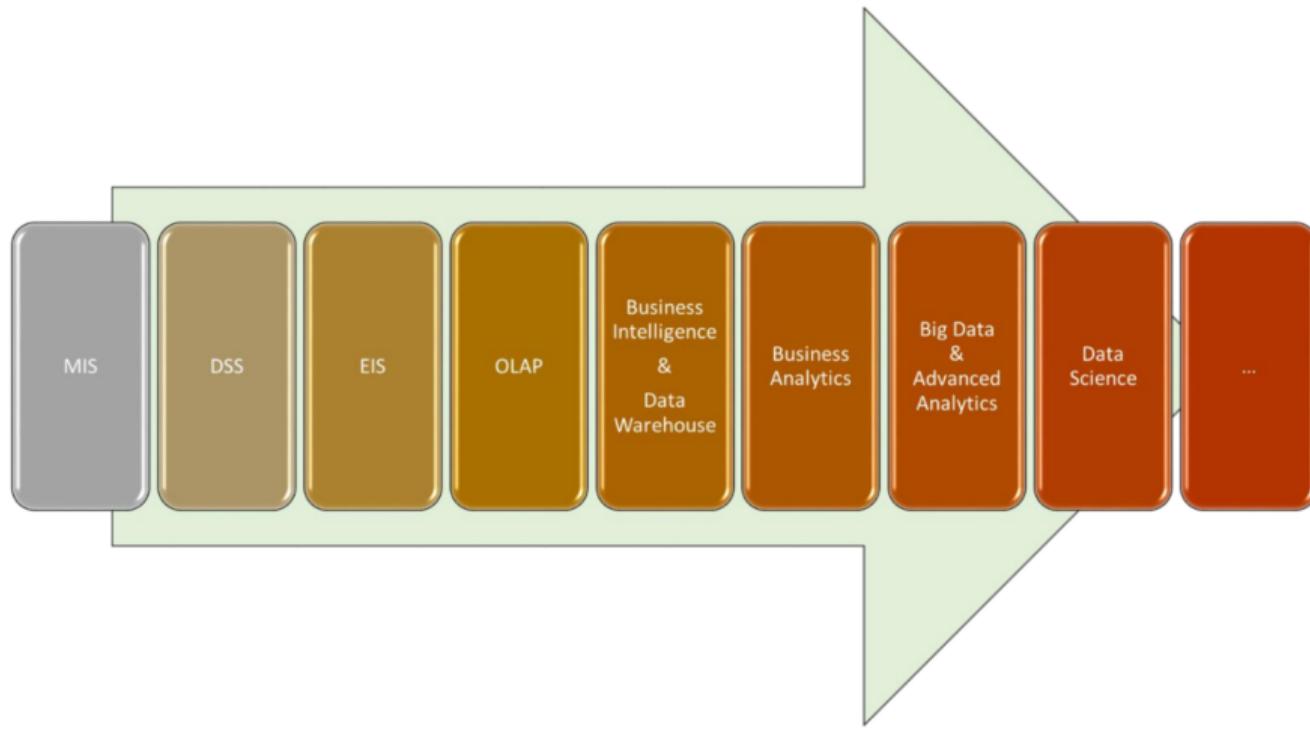
*Data:* a collection of raw value elements

*Information:* the result of collecting and organising data

- ⇒ relationships between data items
- ⇒ context
- ⇒ meaning

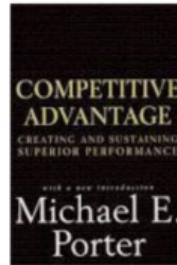
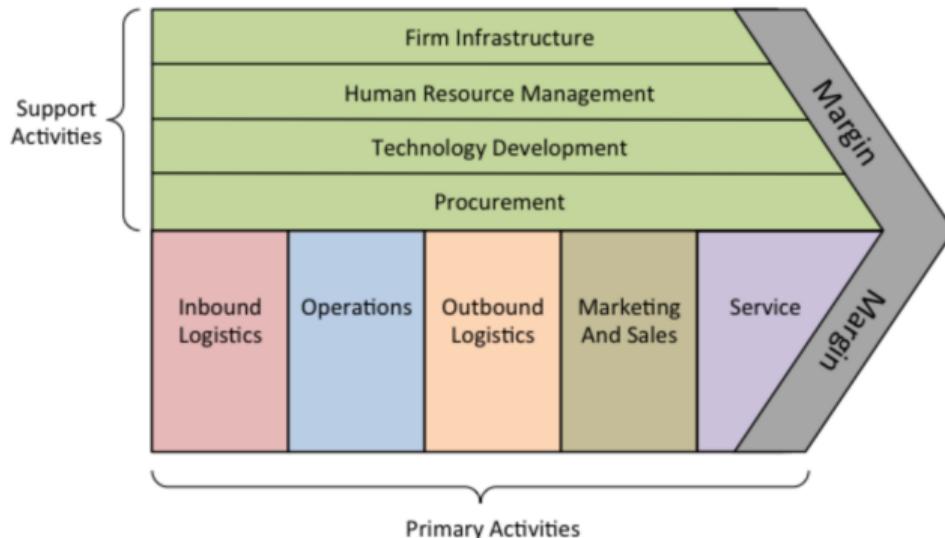
*Knowledge:* understanding information based on recognising patterns

# Increasing insights



# Where does *data* come from? 1/2

A *business process* is a set of activities that, once completed, will achieve an *organisational goal* (e.g. deliver your product to your customer)



# Where does *data* come from? 2/2

- When an event in the real world *changes the state* of the enterprise, one of the events below happens
  - a *transaction* is executed to reflect the corresponding change in the *database*
  - a signal is collected from the infrastructure and stored somewhere
- A *transaction* is a business event that generates or modifies data stored in an information system (database)
- A *signal* is the reading of a measure produced by a sensor
- Data may also be provided by *external subjects*

# OLTP (On-Line Transaction Processing)

- a class of software programs capable of supporting transaction-oriented applications and data storage
- designed to record the daily routine transactions necessary to run the business
- key goals: availability, speed, concurrency and recoverability

# ERP (Enterprise Resource Planning)

An *integrated system*

- can manage all the business processes of all departments within a single software product
- a common database supports all the applications
- operates in or near real time
- has a consistent look and feel across modules
- e.g. [SAP](#)



[https://en.wikipedia.org/wiki/Enterprise\\_resource\\_planning](https://en.wikipedia.org/wiki/Enterprise_resource_planning)

# MIS (Management Information Systems)

- standardized and fixed reporting systems built on existing OLTP
- support structured, operational decision making
  - decisions that can be described in detail before the decision is made
- used by both managers and employees
  - generate performance indicators

# DSS (Decision Support Systems)

- interactive and user-friendly analytical system
- provides support for complex and unstructured decisions
  - decisions that cannot be described in detail before the decision is made
- attempt to combine the use of models or analytic techniques with traditional data access and retrieval functions

# EIS (Executive Information Systems)

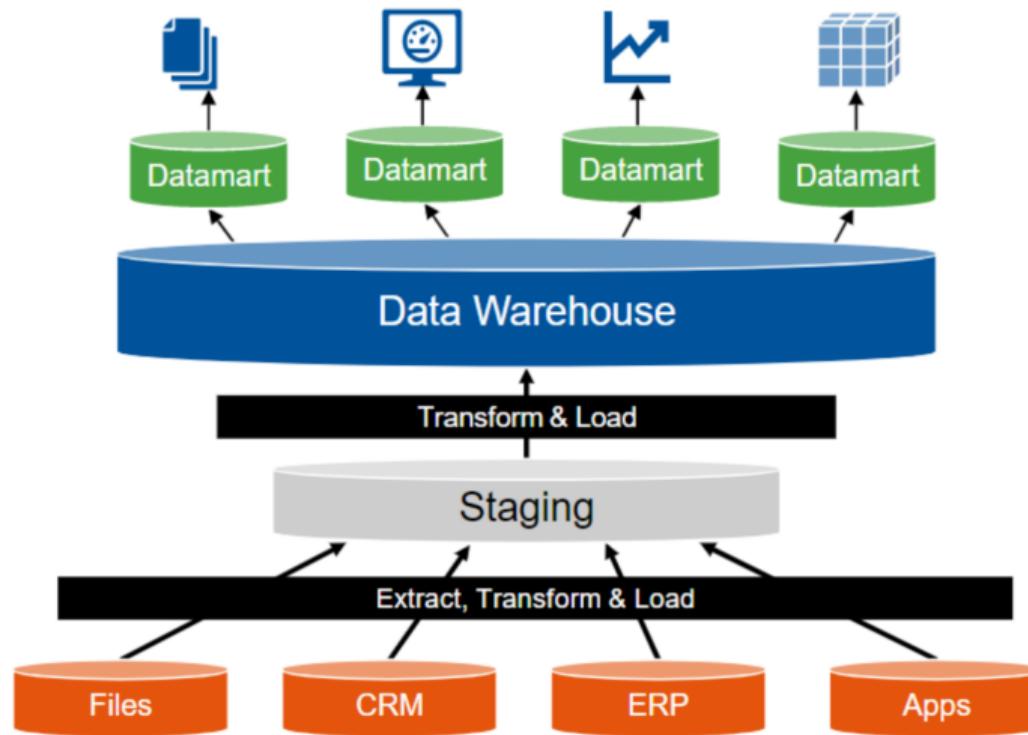
- support the executive level of management
- are used to formulate high level strategic decisions impacting the direction of the organisation
- usually have user friendly interfaces and the ability to extract summary data from internal and external systems

# BI (Business Intelligence)

## Several definitions

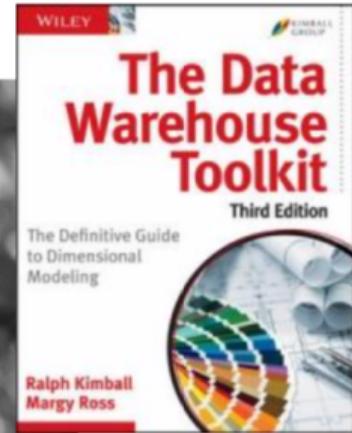
- “applications, infrastructure, tools and best practices that enable access to and analysis of information to improve and optimise decisions and performance” (Gartner)
- “a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making” (Forrester Research)

# BI Architecture



# Data Warehouse

Ralph Kimball (1996)  
“Data Warehouse is a copy of transaction data specifically structured for query and analysis”



# OLAP (OnLine Analytical Processing)

- Analyse multidimensional data interactively from multiple perspectives
- Roughly speaking: extensive usage of group by and summary functions with easy selections, projections, column exchanges, ...
- Uses algorithms and data structures specifically designed to ease these operations
  - e.g. *data cube*
- Some typical OLAP applications: business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting

# MIS vs OLAP

## Management Systems OLTP



## Analytical Systems OLAP



# Structured vs unstructured decisions

<b>Structured</b>		<b>Unstructured</b>	
<b>Description</b>	<b>Example</b>	<b>Description</b>	<b>Example</b>
Made under an established situation	Hiring a new employee	Made under an emergent situation	Fire breakout
Programmed	Start the monthly payment of salaries	Unplanned	Opportunity for financial investment
Fully understood	When a bank customer makes huge fund movements ask him the reason	Unclear or uncertain	Necessary to acquire information to understand which operation is to be performed
Routine task	Hiring new personnel in a given sector	Sudden One-shot situation	Dealing with a labor strike
Specified process	Manufacturing something	General processes	Managing security for IT equipment
Well defined methodology	Possible withdraw of funds from international accounts according to currency rates	Decisions relying on knowledge and/or expertise and on analysis of information	What new market segment could be targeted

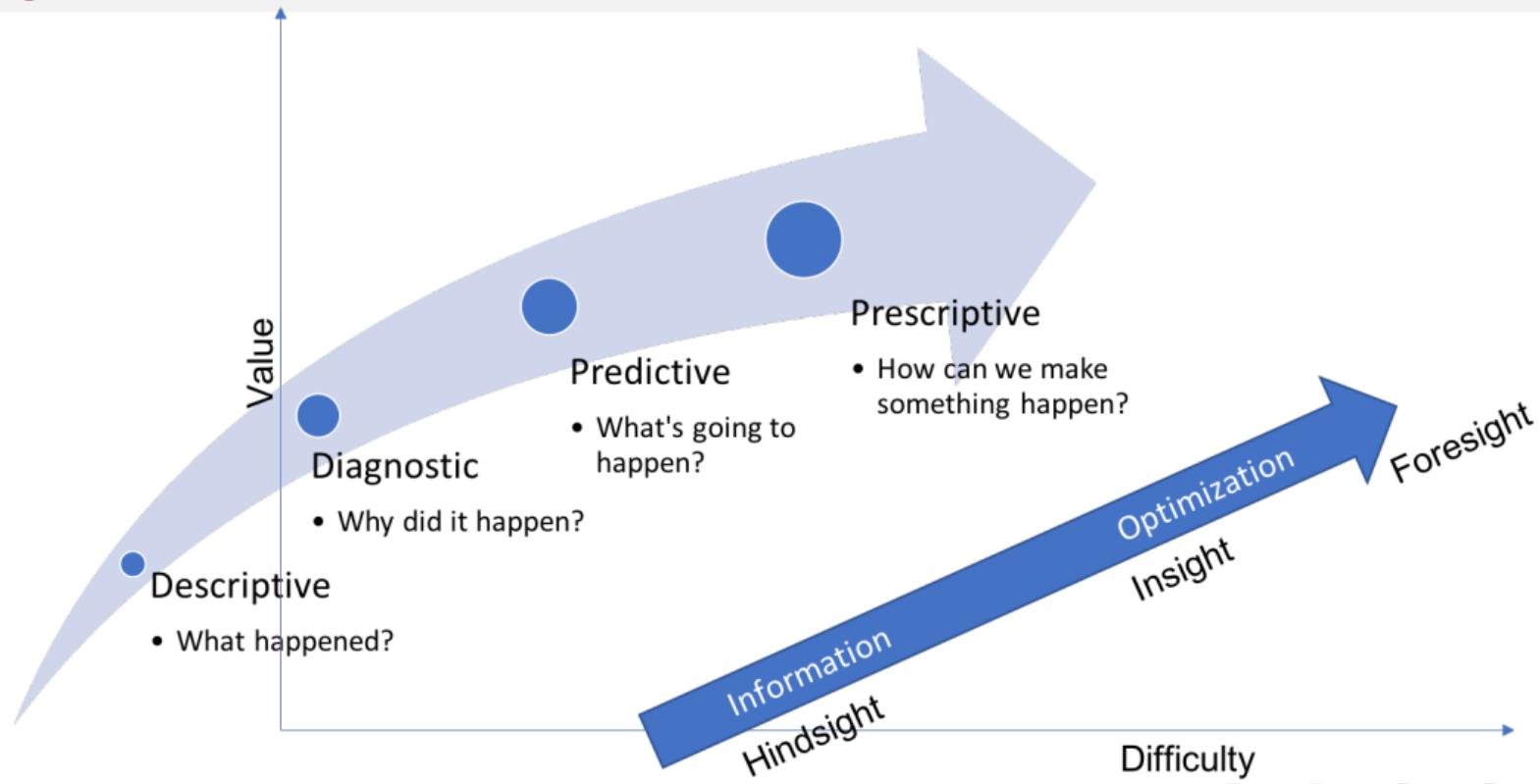
# Analytics vs Data Mining

**Analytics** – Structured decisions driven by data

**Data Mining** – Unstructured decisions driven by data

Sometimes they can provide insights in order to define a new structured decision

# Analytics



# Analytics

- descriptive
  - aggregate data with DB techniques, understand data, descriptive statistics and unsupervised machine learning
- diagnostic
  - descriptive + domain knowledge, understand causes
- predictive
  - calculate the most probable value of a variable in a future time, given the history of a set (sequence) of variables
- prescriptive
  - suggest actions to be taken to obtain the desired effect, choose among options and strategies, optimize

# Example of Descriptive Analytics

Summarize and understand historical data.

- A retail company analyzing its sales data over the past year.
- The company generates reports showing:
  - Total sales by region.
  - Sales trends over time (monthly, quarterly, etc.).
  - Breakdown of sales by product category.
- Key output:
  - Visualizations like bar charts, line graphs, and pie charts summarizing the data.
  - Summary statistics like averages, percentages, and distributions.

# Example of Diagnostic Analytics

Diagnostic analytics focuses on understanding the reasons behind past outcomes.

- Example: A retail company investigating a sudden drop in sales during a specific quarter.
- The company conducts an analysis to determine:
  - Changes in customer behavior or preferences.
  - External factors such as economic downturns or competitive actions.
  - Internal issues like supply chain disruptions or marketing inefficiencies.
- Key output:
  - Correlation analysis to identify relationships between variables.
  - Drill-down reports that break down sales performance by location, product, or time.
  - Statistical tests to find significant causes of the decline.

# Example of Predictive Analytics

Predictive analytics uses historical data and statistical models to forecast future outcomes.

- Example: A retail company predicting future sales based on historical data.
- The company applies predictive models to:
  - Forecast sales for the upcoming quarters.
  - Identify which products are likely to see increased demand.
  - Estimate customer churn rates and predict high-value customers.
- Key output:
  - Time-series forecasts and regression models.
  - Probabilistic estimates for future sales growth.
  - Predictive scores and risk assessments for customer behavior.

# Example of Prescriptive Analytics

Suggest actions based on predictions to achieve desired outcomes.

- Example: A retail company optimizing inventory management for the upcoming holiday season.
- The company uses prescriptive analytics to:
  - Recommend optimal stock levels for each product category.
  - Suggest promotional strategies to maximize sales and reduce excess inventory.
  - Allocate resources (e.g., staff, storage) efficiently across different regions.
- Key output:
  - Decision models and optimization algorithms.
  - Actionable recommendations based on constraints (e.g., budget, capacity).
  - Simulations and what-if scenarios to explore the impact of different decisions.

1	General information	2
2	Data in organisations	6
3	<b>Data Mining and Big Data</b>	31
	● Data Sources	40
	● Technological Progress	50
4	Cloud Computing	56
5	Big Data	62
6	Data Mining	74

# Paradigm Shift

- It is not easy to manage and transform raw data from OLTP/ERP systems into information
- over time (from 1960's to today) the market found solutions
  - databases, SQL, Data Warehousing, OLAP, etc.
- now the world of data is facing a *paradigm shift*



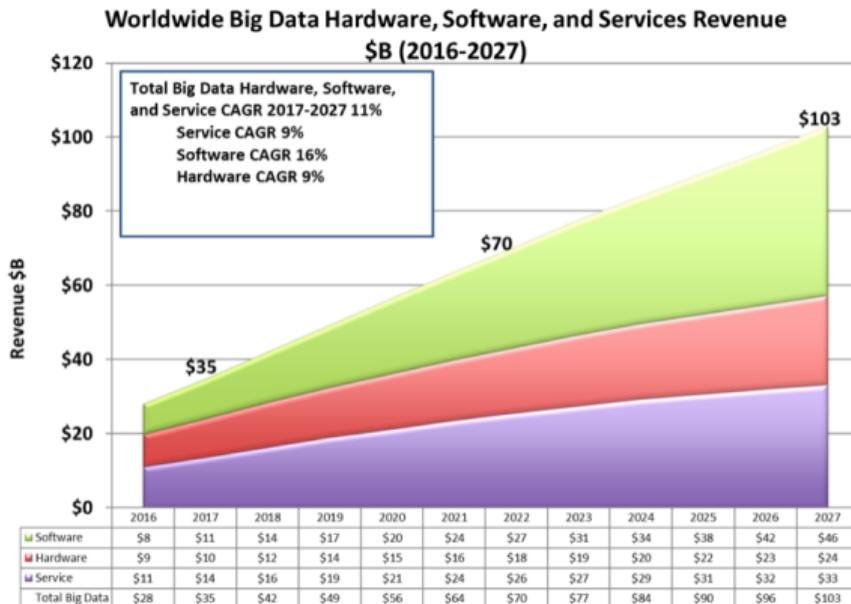
# Look at this



<https://www.youtube.com/watch?v=eVSfJhssXUA>

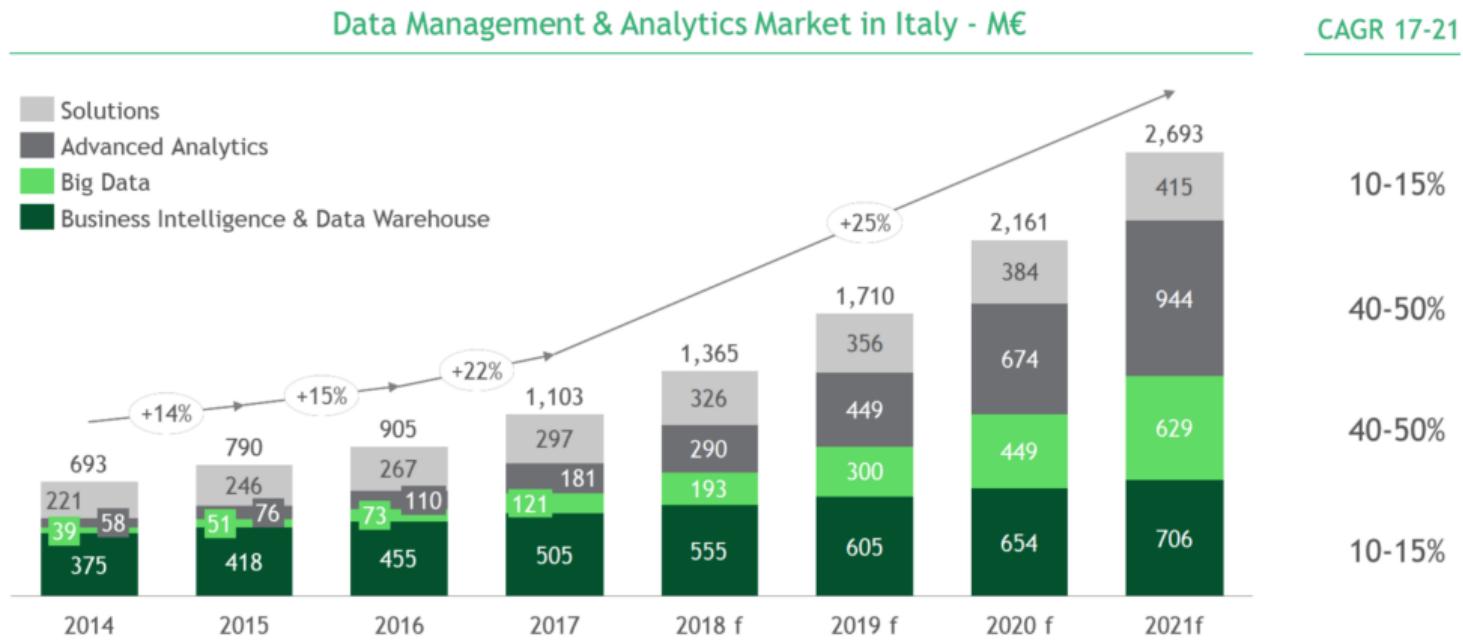
# Why are we talking about that?

- Worldwide IT spending is projected to total \$3.7 trillion in 2018, an increase of 4.5 percent from 2017, according to the latest forecast by Gartner
- Big Data has a significant share



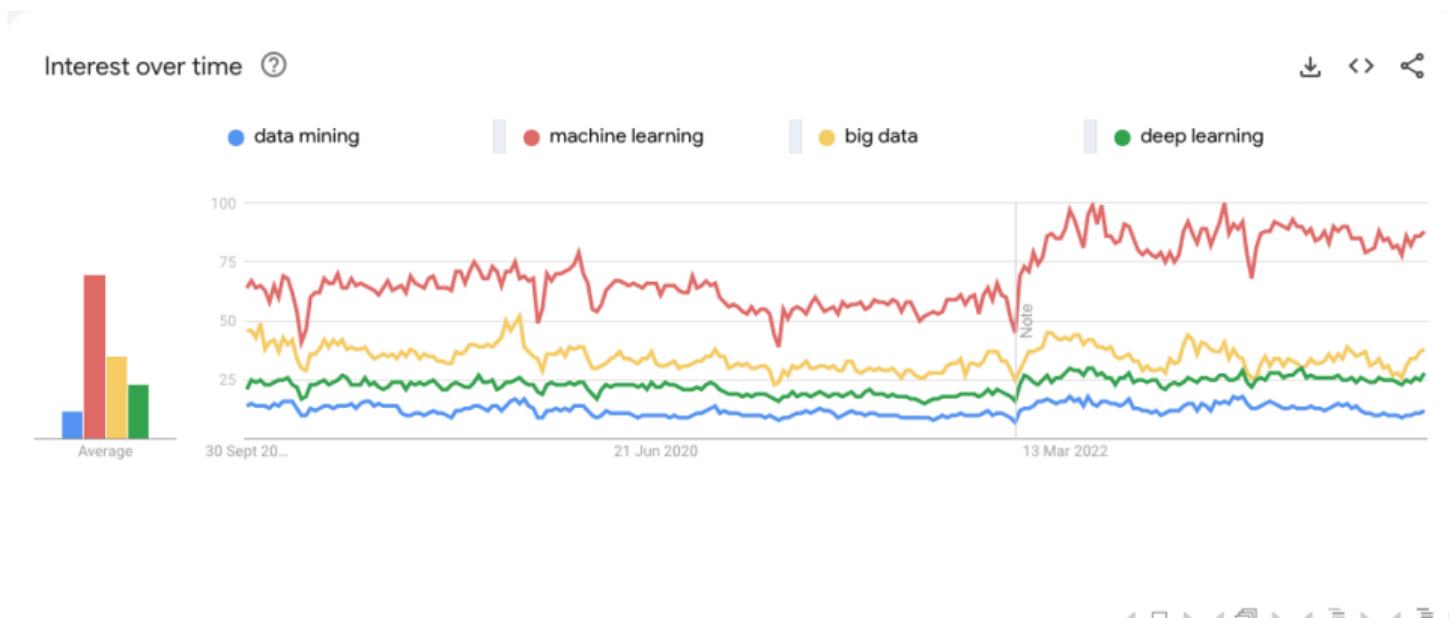
Source: [wikibon.com](http://wikibon.com)

# Spending in *data* has constant and continuous growth



# Why are we talking about that?

Google Trends - Last 5 years



# Why are we talking about that?



“70% of the super-computation capacity in Italy is concentrated in Emilia-Romagna. On big data alone we have a network of 1,800 engineers, 230 foreign researchers on a permanent basis and 60 higher education courses. The future of manufacturing, sciences, medicine is based on the ability to manage huge amounts of data. We don't have to invent anything. We already have everything, it just needs to be put into the system”

Patrizio Bianchi (*former Councilor Emilia-Romagna Region*)

# Data Revolution

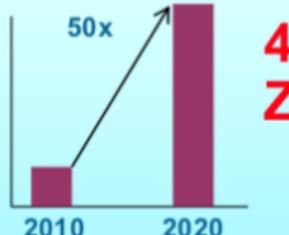
- Flood of Data from new sources
  - Social Networks, Internet of Things (sensors), smartphones, wearable devices, Industry 4.0, GPS, Smart Cities, Home automation, ...
  - ... more data has been created in the past two years than in the entire previous history of the human race!



# Let's get to the point

- Organisations struggled to collect data to improve the *decision process*
- Now they have *a lot of data, Big Data*
- It is not easy to extract value from such a Big Data

Cost efficiently processing the growing **Volume**



Responding to the increasing **Velocity**



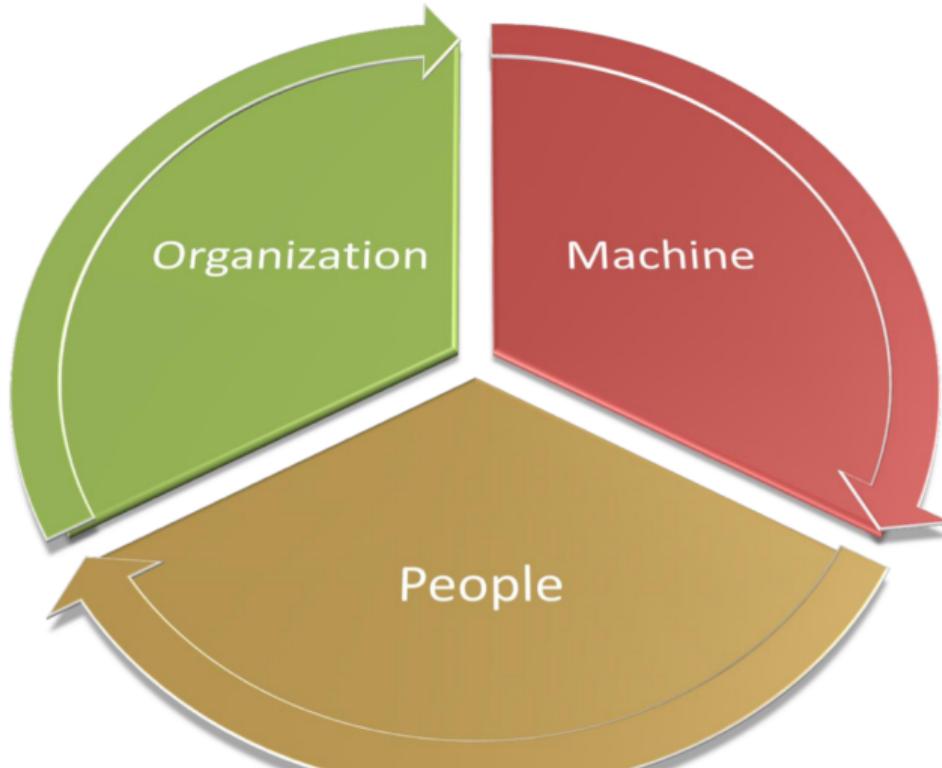
**30 Billion**  
sensors and counting

Collectively Analyzing the broadening **Variety**

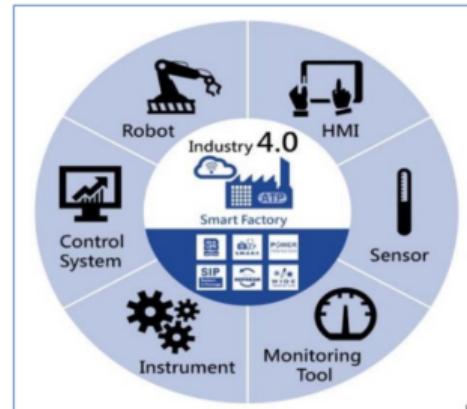
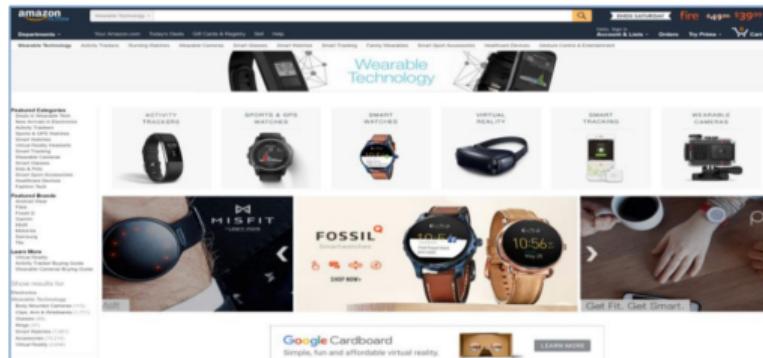
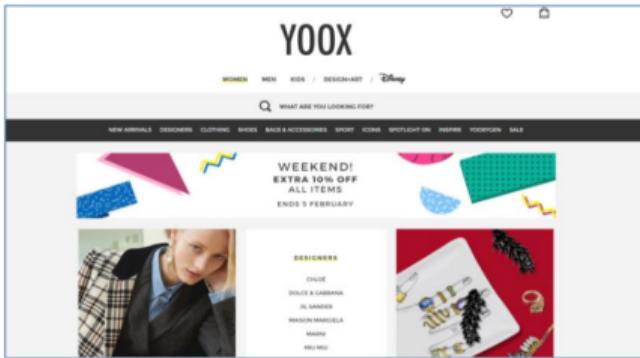
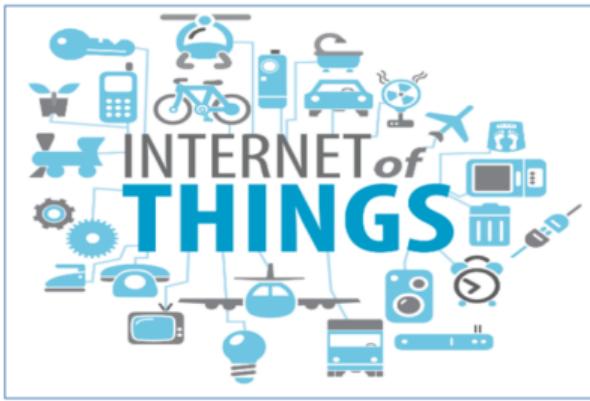


**80%** of the worlds data is unstructured

# Data Sources



# Machine-Generated Data



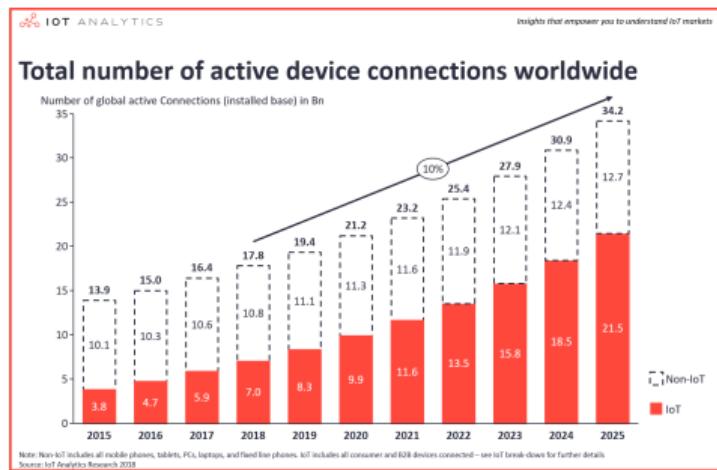
# IoT (Internet of Things)

Phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world

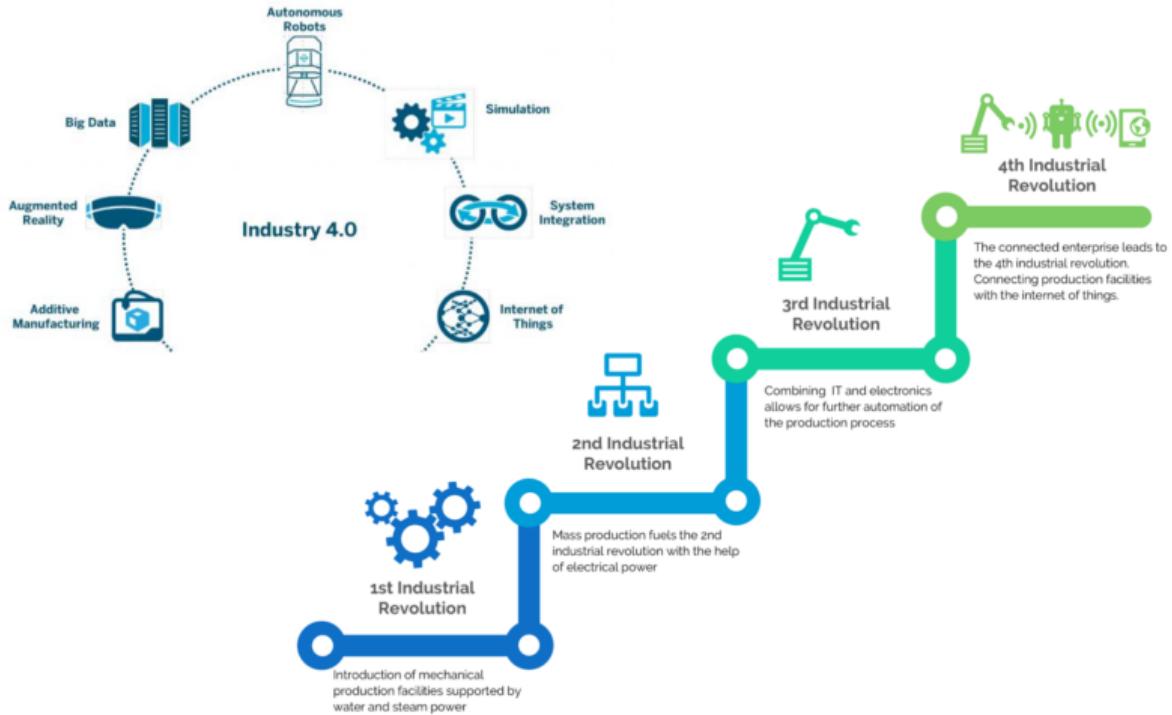
- Well-structured data, suitable for computer processing
- Size and speed beyond traditional approaches
- Data from sensors:
  - Fixed sensors
    - Traffic sensors/webcam, Home automation, Weather/pollution sensors, Scientific sensors, Security/surveillance videos/images, ...
  - Mobile sensors (tracking)
    - Mobile phone location, Cars, Satellite images, ...
  - Data from computer systems
    - Logs, Web logs, ...

# IoT – Sensors

- More than 20 billion objects in 2020 are interconnected and connected to the Internet
- They will be always connected, pervasively installed in environment, mixed with people and machines
- Are all the valuable assets well managed?



# Industry 4.0



# Industry 4.0

The screenshot shows a TEDx talk page. At the top left is the speaker's name, Olivier Scalabre. The main title is "The next manufacturing revolution is here". Below the title is a play button icon. To the right of the play button are four circular icons with text: "Add to list", "Favorite", "Download", and "Rate". On the left side of the video frame, there is a small thumbnail of the speaker and some text: "TEDxBIG Paris · 12:26 · Filmed May 2016", "24 subtitle languages", and "View interactive transcript". Below the video frame, there is a portrait of the speaker, Olivier Scalabre, and his bio: "Olivier Scalabre Industrial systems thinker BCG's Olivier Scalabre analyzes the evolution of large industrial companies' manufacturing footprint and operations. [Full bio](#)".

Economic growth has been slowing for the past 50 years, but relief might come from an unexpected place — a new form of manufacturing that is neither what you thought it was nor where you thought it was. Industrial systems thinker Olivier Scalabre details how a fourth manufacturing revolution will produce a macroeconomic shift and boost employment, productivity and growth.

[https://www.ted.com/talks/olivier\\_scalabre\\_the\\_next\\_manufacturing\\_revolution\\_is\\_here?language=en](https://www.ted.com/talks/olivier_scalabre_the_next_manufacturing_revolution_is_here?language=en)

# People-Generated Data



Company	Data Processed Daily
eBay	100 Petabytes (PB)
Google	100 PB
Facebook	30+ PB
Twitter	100 Terabytes(=.1PB)
Spotify	64 Terabytes



- Social media: Facebook, LinkedIn, Instagram, Youtube, ...
  - daily, huge amounts of new data
- Text-heavy, unstructured, no well-defined data model

# In 60 seconds . . .



# Organisation-Generated Data

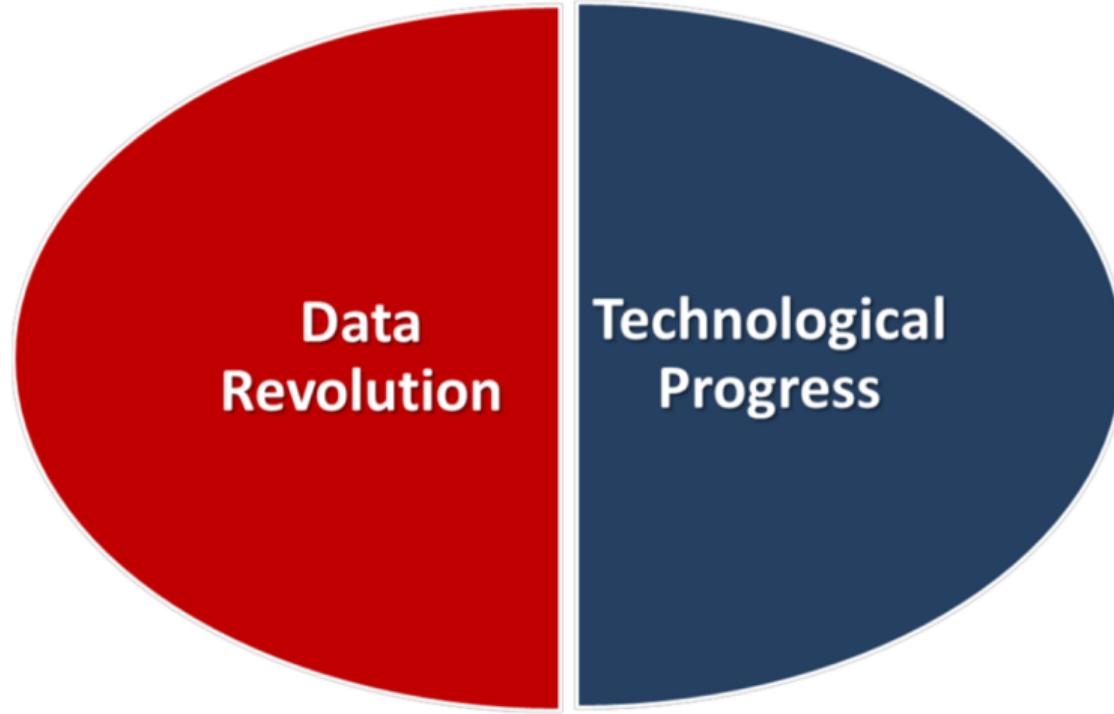
Common types of organisational big data:

- commercial transactions, credit cards, government institutions, e-commerce, banking or stock records, medical records, sensors, transactions, clicks,

...



# Key Enablers

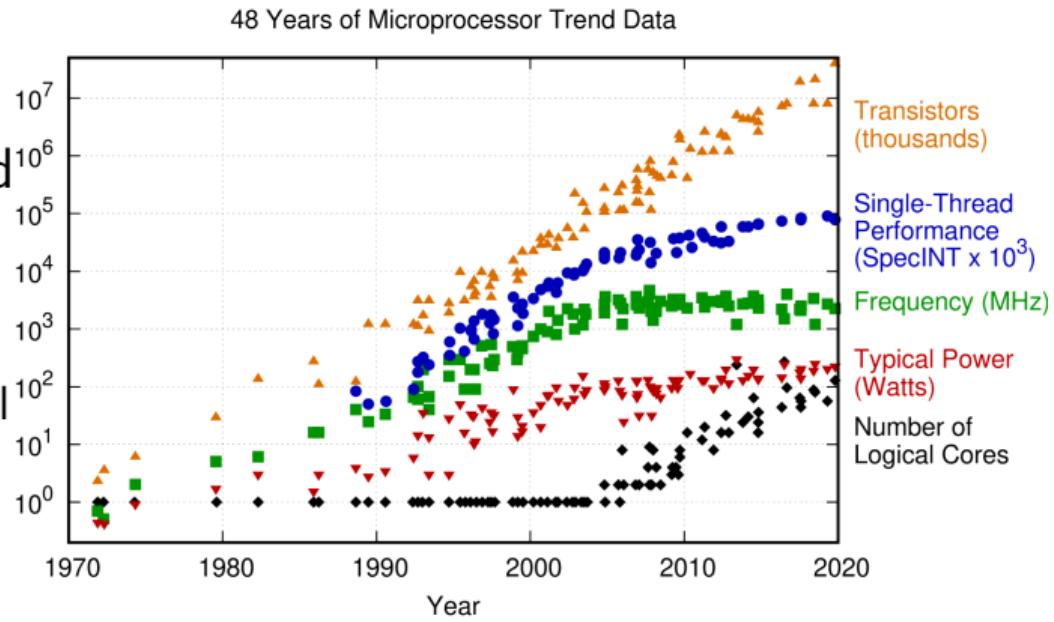


# Technological Progress



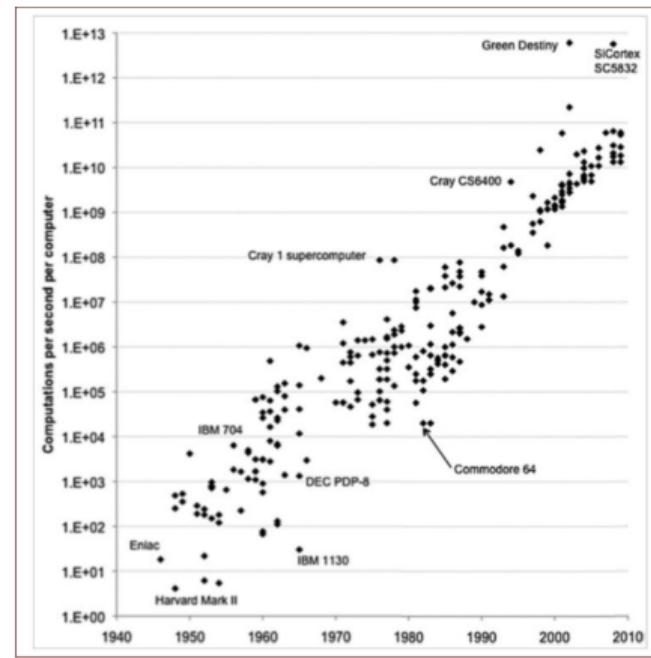
# Technological Progress

- The number of transistors on integrated circuits doubles approximately every 2 years?
  - Gordon E. Moore, Intel co-founder, 1965



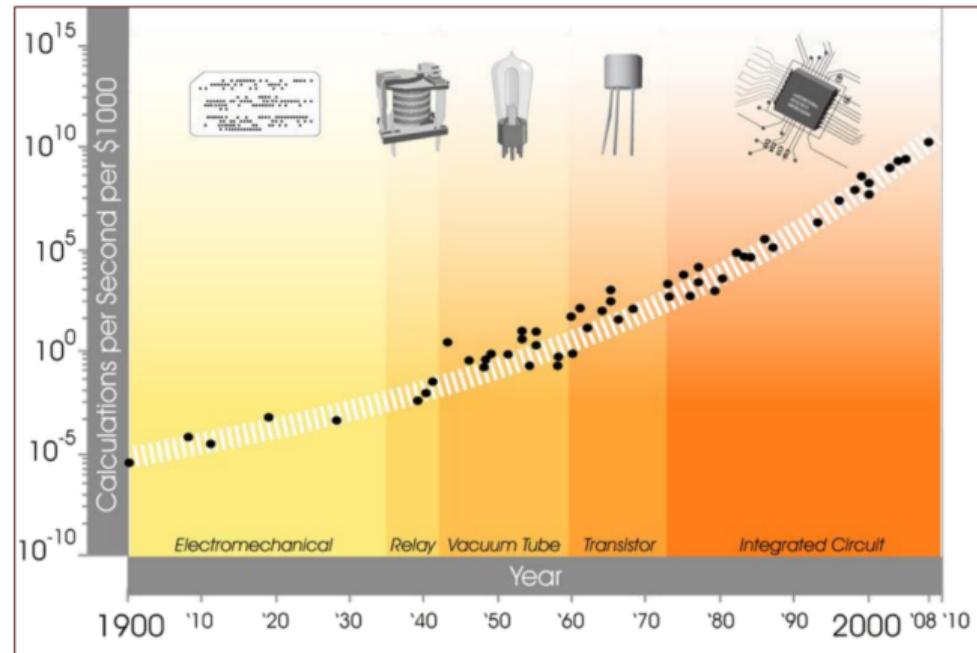
# Computational Capacity

- Computational capacity increases also thanks to architectural advances



# Calculations per second per \$1000

- Dramatic drop in the cost of computation
  - figure produced by Ray Kurzweil

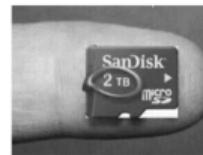


# Exponential Increase in Storage Capacity

... and Decrease in Storage Costs

- Changes in storage capacity and costs were subject to big leaps, rather than to linear laws
- In the last 40 years they have been even bigger than the changes in computational capacity

The IBM Model 350 disk file with a storage space of 5MB from 1956 and a Micro SD Card



# Technology keeps progressing

## 1 The accelerating pace of change ...

Agricultural Revolution → 8,000 years → Industrial Revolution → 120 years → Light-bulb → 90 years → Moon landing → 22 years → World Wide Web → 9 years → Human genome sequenced

## 2 ... and exponential growth in computing power ...

Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years

### COMPUTER RANKINGS

By calculations per second per \$1,000

**Analytical engine**  
Never fully built, Charles Babbage's invention was designed to solve computational and logical problems.



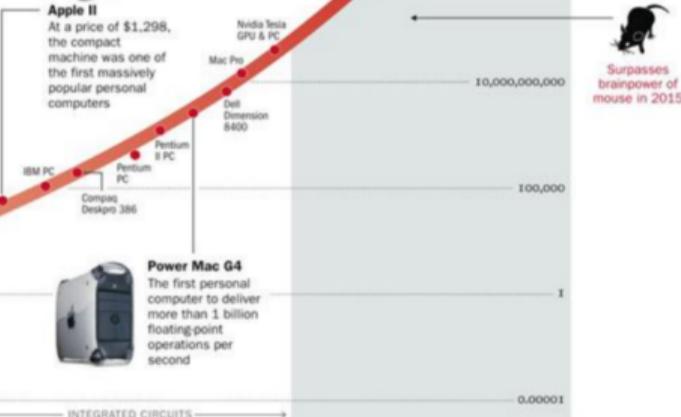
**Colossus**

The electronic computer, with 1,500 vacuum tubes, helped the British crack German codes during WW II



**UNIVAC I**

The first commercially marketed computer, used to tabulate the U.S. Census, occupied 943 cu. ft.

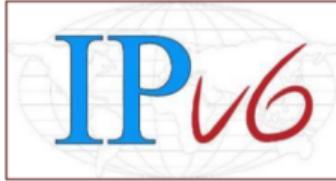


## 3 ... will lead to the Singularity

Surpasses brainpower equivalent to that of all human brains combined

Surpasses brainpower of human in 2023

Surpasses brainpower of mouse in 2015



1	General information	2
2	Data in organisations	6
3	Data Mining and Big Data	31
4	Cloud Computing	56
5	Big Data	62
6	Data Mining	74

# Cloud Computing

- Is a *delivery model*
- Access to a shared pool of configurable computing resources
  - servers, storage, databases, software, networks, ...
- On-demand
- Pay-per-Use
- Like a *utility*, e.g. electricity
- Fast provisioning and deploying

# Cloud Offerings by Service

More Structured

Software as a Service (SaaS)

Facebook, Salesforce.com, Gmail

Less Control



Platform as a Service (PaaS)

Google App Engine, Microsoft Azure



Infrastructure as a Service (IaaS)

Less Structured

3Tier, Amazon EC2, Rackspace, GoGrid

More Control

# Cloud Offerings by Service

**SaaS (Software as Service)**: Consumer uses provider's applications running on provider's cloud infrastructure

**PaaS (Platform as Service)**: Consumer can create custom applications using programming tools supported by the provider and deploy them onto the provider's cloud infrastructure

**IaaS (Infrastructure as a Service)**: Consumer can use computing resources within provider's infrastructure upon which they can deploy and run arbitrary software, including OS and applications.

# Cloud – Benefits I

**Scalability:** resource is available as and when the client needs it and, therefore, there are no delays in expanding capacity or the wastage of unused capacity

**No investment in hardware:** everything is set up and maintained by the cloud provider, saving the time and cost of doing so on the client side

**Pay for what you use:** if the service is only needed for a limited period then it is only paid for over that period and subscriptions can usually be halted at any time

# Cloud – Benefits II

**Updates are automated:** Updates will usually be free of charge and deployed automatically by the cloud provider

**Disaster Recovery and Security:** managed by the provider

**Flexibility and Scalability:** can easily increase on demand the amount of computational power/storage space

**Accessibility:** work from anywhere with web browser and pc/mobile device

1	General information	2
2	Data in organisations	6
3	Data Mining and Big Data	31
4	Cloud Computing	56
5	<b>Big Data</b>	62
6	Data Mining	74

# From data to Big Data

Warm-up quiz

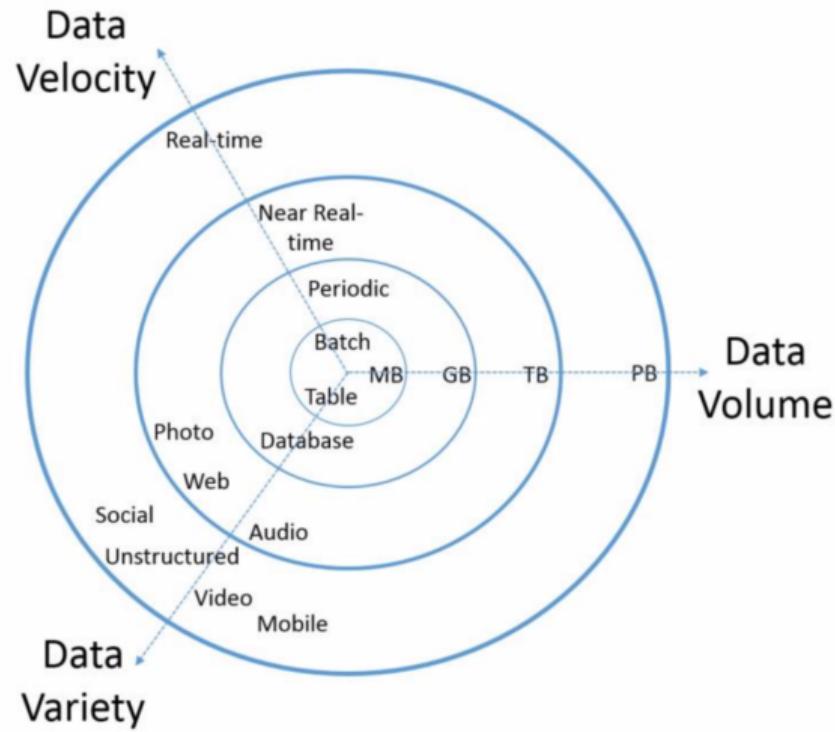
[https://wooclap.com/  
Data-to-big-data](https://wooclap.com/Data-to-big-data)

# Big Data

## Definition

a collection of data sets so *large* and/or *complex* and/or fast changing that they are difficult to process using traditional DBMSs or traditional data processing applications

# Big Data – The three V's



# Big Data is so . . . *Big!*

The total amount of DATA being captured and stored by industry doubles every 1.2 years

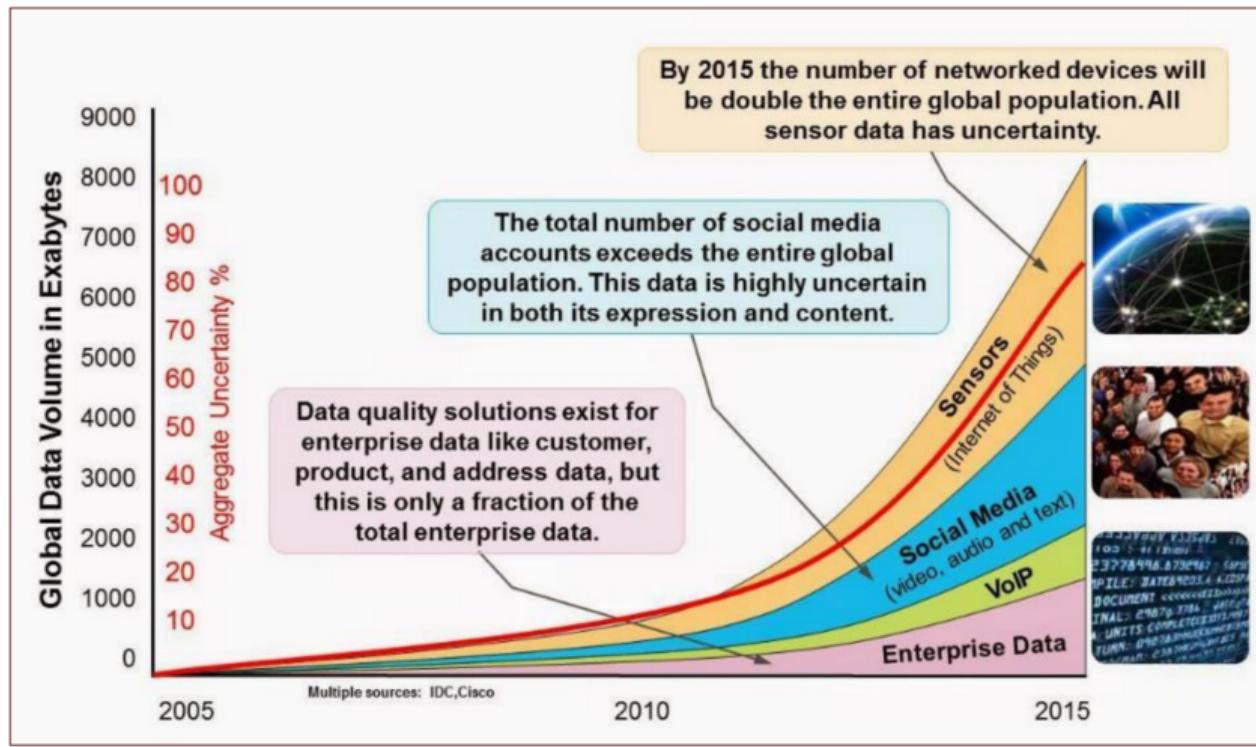


...more data has been created in the past two years than in the entire previous history of the human race

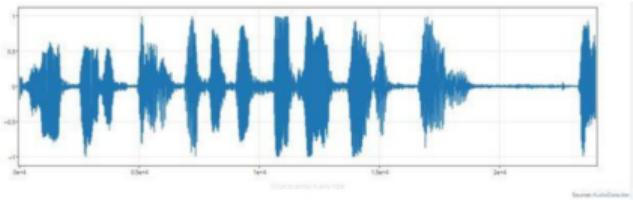
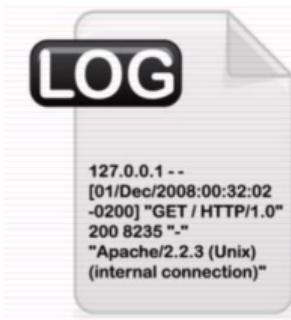
In 2017 over 1 trillion photos will be taken and billions of them will be shared online

In 5 years there over 50 billion smart connected devices in the world  
all developed to collect, analyze and share data

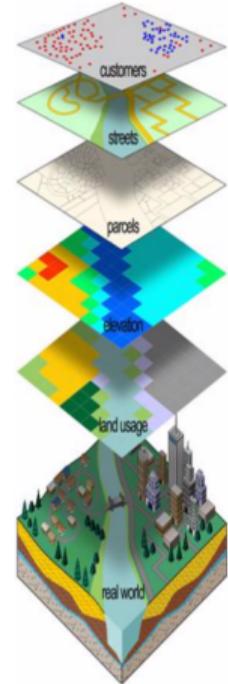
... and still *growing!*



... and *complex!*

variety... is the spice of life



# Big Data – A high level taxonomy

**Structured:** relational tables, spreadsheet (or data which could easily fit in them)

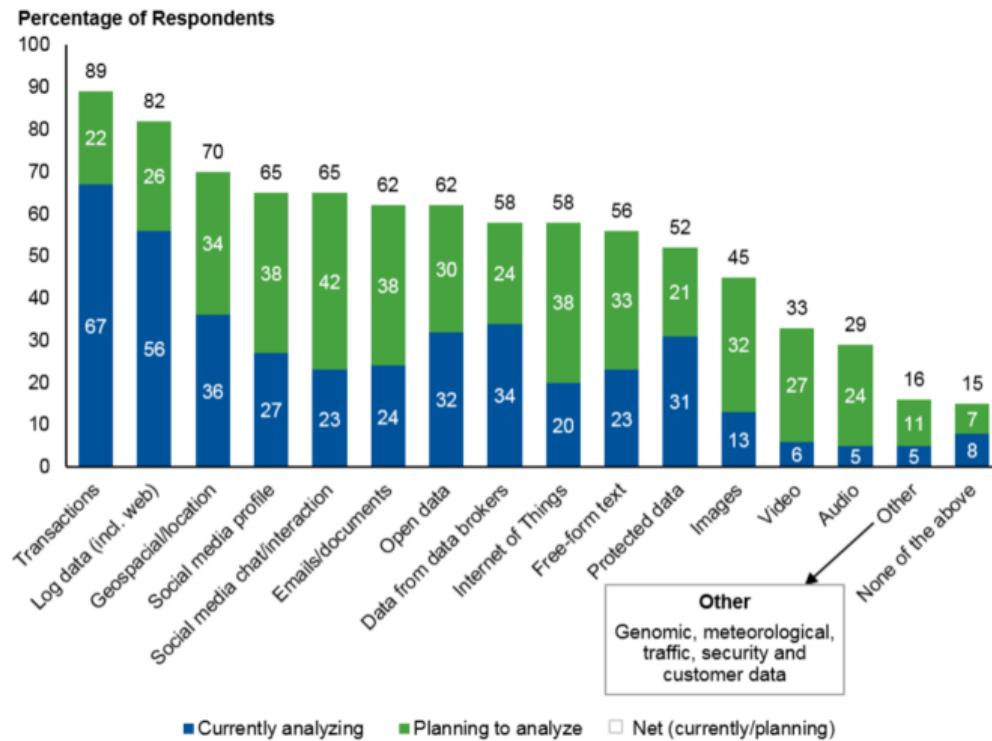
**Unstructured:** does not have an associated data model

- video, audio, pictures,
- 80% of available data

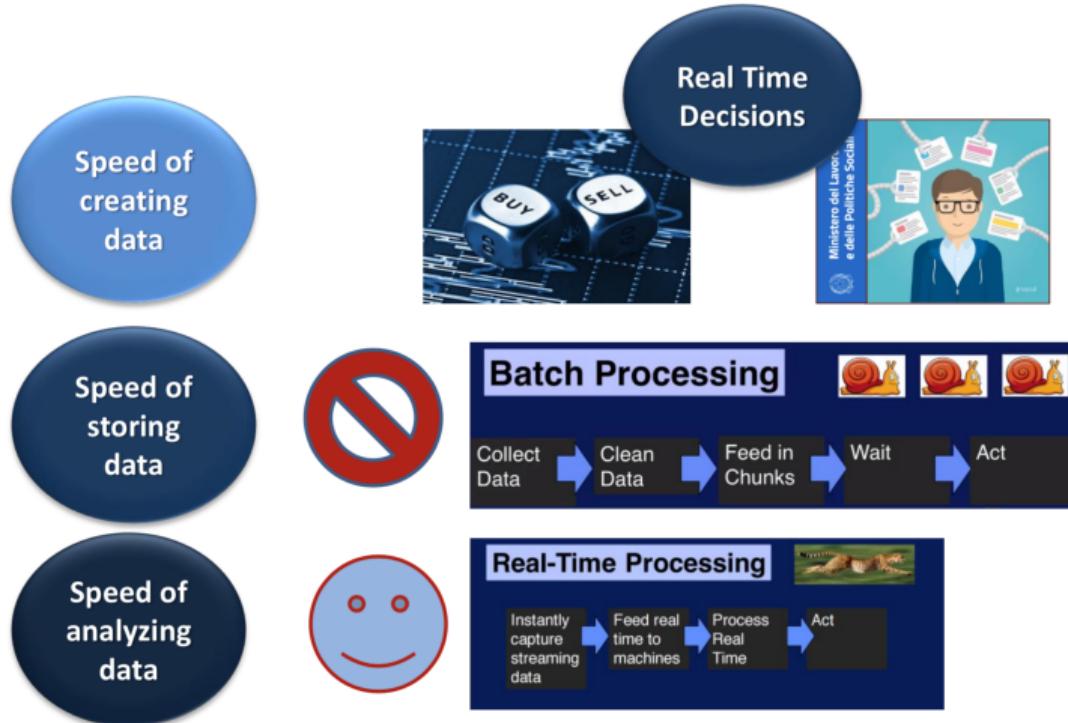
**Semi-structured:** there is some structure, perhaps data refer to different structures

- self describing data: XML, JSON, ...

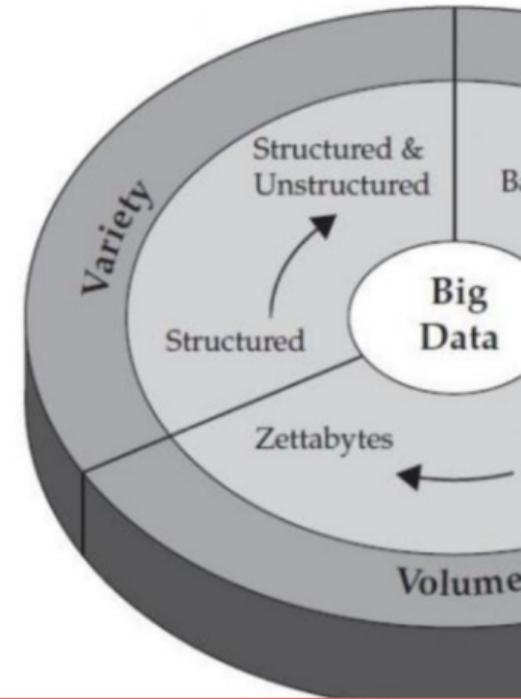
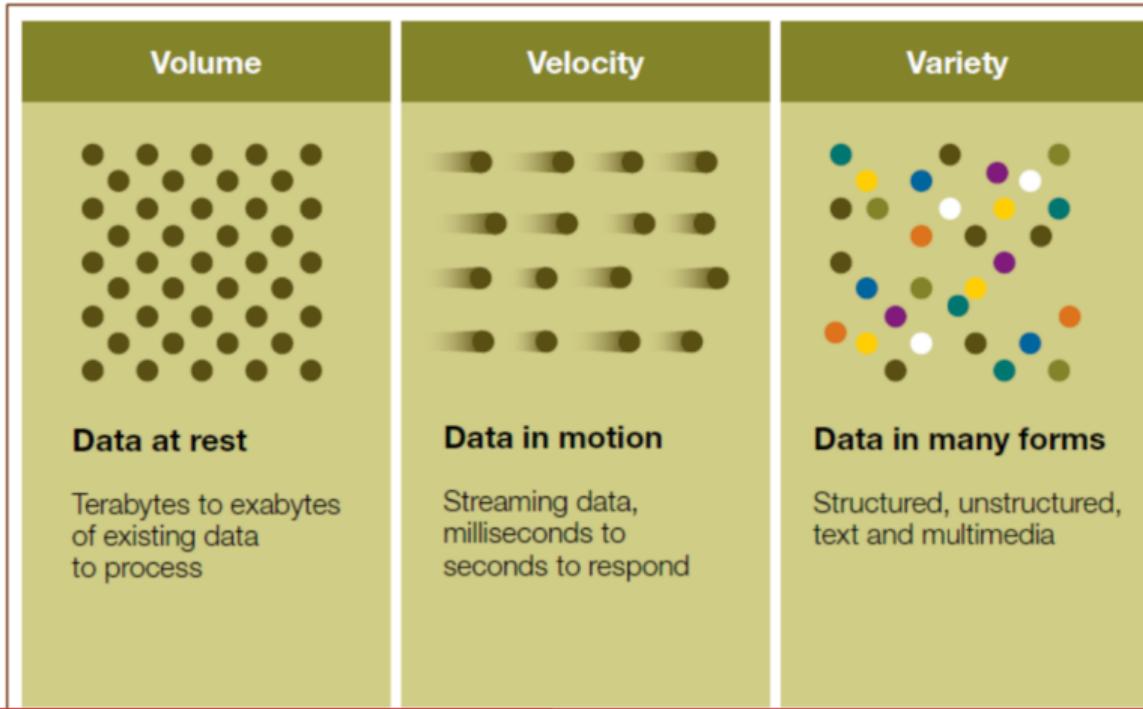
# Data Sources to be analysed



# Big Data is so . . . fast!



# The V of Big Data (again)



# What's the relationship between Big Data and Cloud Computing?

1. Cloud Computing is relevant also if you do not have Big Data
2. Big Data can also be processed without Cloud Computing
3. In many cases Cloud Computing is a key asset to be able to deal with Big Data

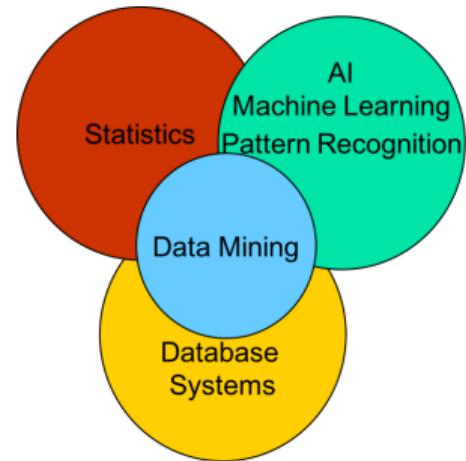
because in presence of one or more of the V it is difficult to use traditional databases or traditional data processing

*Can you imagine use cases for statements 1 and 2?*

1	General information	2
2	Data in organisations	6
3	Data Mining and Big Data	31
4	Cloud Computing	56
5	Big Data	62
6	Data Mining	74

# Data Mining Origins

- The sizes of the circles do not reflect the relative importance/size of the topics
- Many textbooks referring either to *machine learning* or *data mining* have a significant overlap, sometimes the separation between the two topics is a little *fuzzy*



# Data Mining $\rightleftharpoons$ Machine Learning

In the following we will use the topic names as follows

- **Data mining** is the discovery process described in page 77
- **Machine learning for data mining** is the core of learning models and algorithms which allow to extract actionable patterns from data

Looking at the literature

- Machine learning includes also other concepts and methods which are not used for data mining
- Data mining books frequently include also *learning models* which are not traditionally covered in machine learning literature
  - Look [here](#) for a comprehensive list of data mining topics

# The Data Mining Process

