From Data Warehouse to Big Data

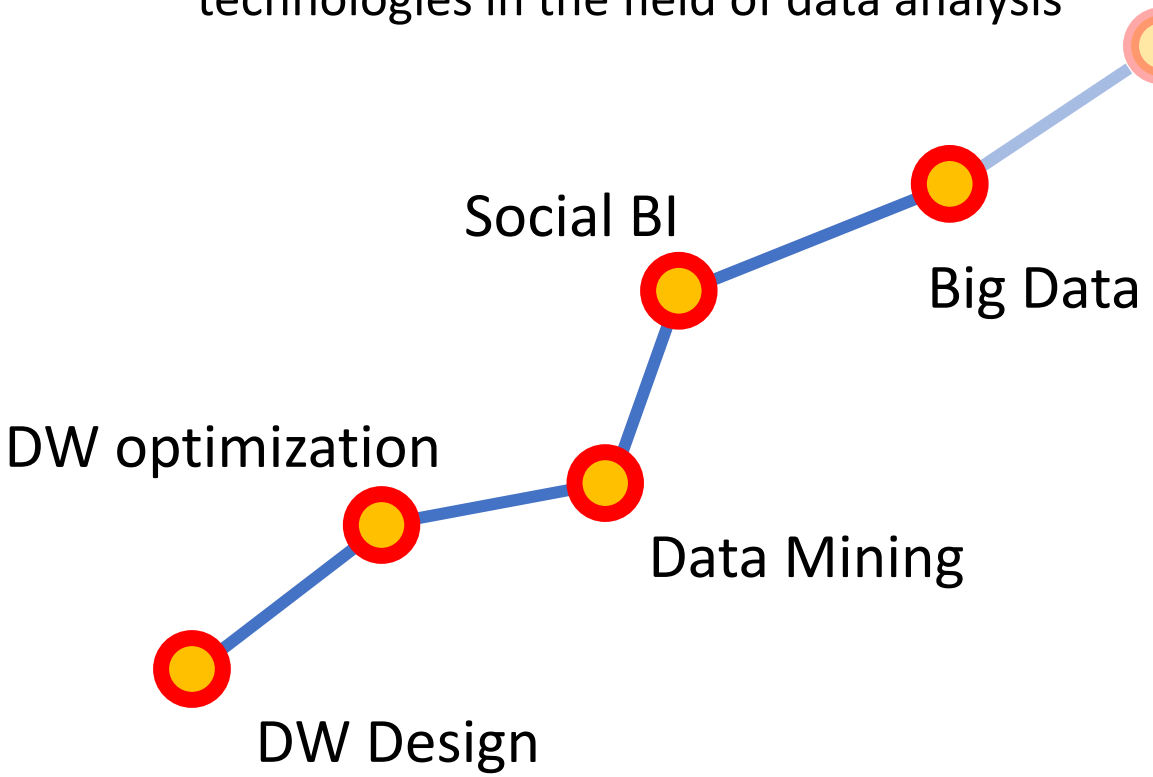
The turbulent evolution of data analysis

Matteo Francia <m.francia@unibo.it>

BBS
MULTIPLYING OPPORTUNITIES

The Business Intelligence Group

The Business Intelligence Group carries out researches on methodologies, techniques and technologies in the field of data analysis



BIG expertise

European funding

- *PANDA* (pattern management in DM)
- *ENPADASI* (EU Nutritional Phenotype Assessment and Data Sharing Initiative)
- *TOREADOR* (As-a-service Big Data Analytics)

Public funding

- *D2I* (integration and mining of heterogeneous DBs)
- *WISDOM* (ontology-enhanced web searching)
- *WebPolEU*(Comparing Social Media and Political Participation across EU)
- *GenData2020* (data-centric genomic computing)
- *DyNamiTE* (Digital fightiNg Tax Evasion)
- *MO.RE. Farming* (Big Data for Precision Farming)
- *INNOFRUVE* (Ricerca industriale ed innovazione nel comparto ortofrutta)
- *AgroBigDataScience* (Big Data for Precision Farming)

Private funding (2015-2021)

- *Data Mining in the Fashion Field* with Valentino
- *Set-up of a Social Business Intelligence framework* with Amadori s.p.a.
- *Feasibility study for a Social Business Intelligence system* with DOXA
- *Anomaly detection in the gas network* with HERA spa
- *Harnessing Wellness Knowledge* with Technogym
- *Methodological and Scientific Support to several Public bodies* With Ministry of Justice, Ministry of Economy and Finance
- *Vaccine monitoring* with Regione Veneto & ONIT
- *Intelligent Monitoring Systems for Critical Environments* with Leonardo-Finmeccanica
- *Data-driven budgetting* with Teddy
- *Digital Transformation* with BRT, PLT Energia

A mandatory premise: a module with multiple levels of understanding

Talking about technical topics to

- a non-technical audience is hard and sometimes frustrating
- a heterogeneous background audience is even harder and often frustrating

… listening is typically worse!

I promise to avoid all the unnecessary technicalities but… sometimes they are necessary!

Don't be afraid of technicalities

- If something is not clear but you believe can be useful to your profile, please ask!
- If something is not clear and useless to your profile, focus on the whole picture and don’t worry

Digital transformation

DT aims to improve the efficiency and effectiveness of companies by exploiting the possibilities offered by new technologies.

All public and private business sectors will be involved in this transformation, albeit with different times and methods

It is important to experiment and understand where and when to digitize

DT is not just a technological issue!

- It requires a long-term strategy and a step-by-step path
- It needs changes in people's mindsets and in the search for digital talent

Data revolution

Data is the main fuel that powers digital transformation and with digital transformation, more and more data are produced

Digitization began in the 1970s with the progressive spread of computers, giving way to the process of digitizing processes and information that continues to accelerate even today by changing its name but not its goal

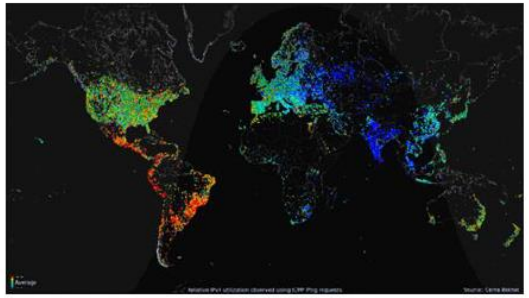
- Post-industrial society
- Information technology revolution
- Digital age

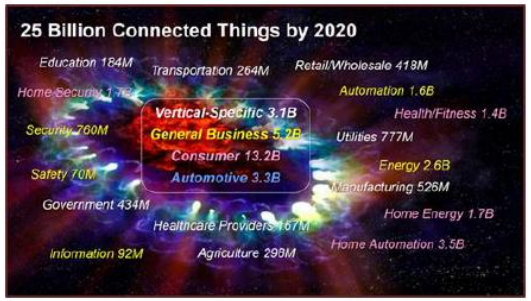
We can date the beginning of the Digital Age in 2002, when more digital than analog information was stored around the world

- At the end of the 1980s, less than 1% of information was in digital format
- in 2012 the percentage had risen to 99% with an annual increase of about 30%, which leads to a doubling of the information stored in less than 3 years.

Where does data come from?

Information systems are no longer limited to the data produced by business processes, but must be rethought to allow the exploitation of all the data useful to the company and to be able to support internal and external processes





Big data vs small data

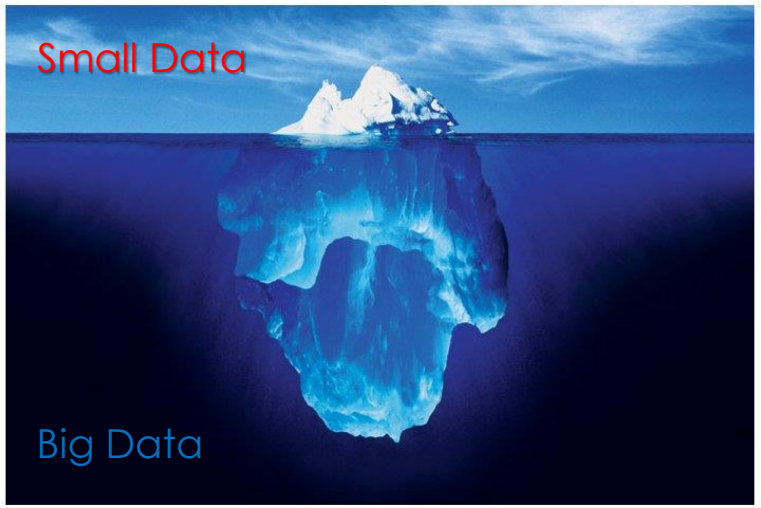
The progressive digitalization of services and systems generates an enormous mass of heterogeneous and real-time data

Big Data must be transformed into Small Data so that it can be exploited for decision-making purposes

Small data is data that is 'small' enough for human comprehension. It is data in a volume and format that makes it accessible, informative and actionable.

To manage the transformation we need:

- Ad hoc Technology (e.g., NO SQL DBMS)
- Computing power (e.g., cloud & cluster computing)
- Automated systems (e.g., artifical intelligence)
- Digital culture
- The right processes (i.e., digital ready processes)



The technology adoption cycle

The adoption of new technologies follows a standard path that involves (1) the maturation of one or more enabling technologies and (2) their diffusion

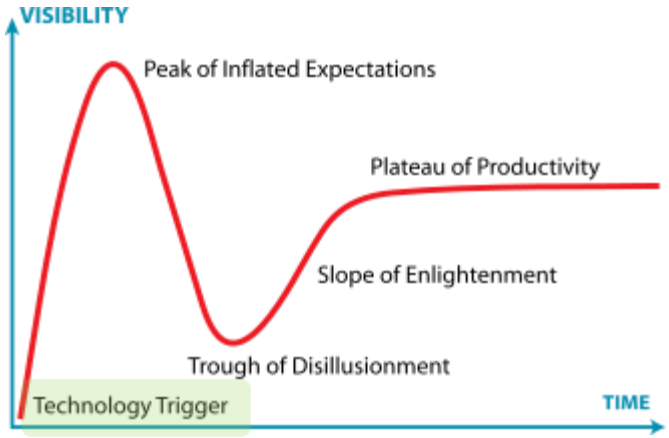
- The first one is driven by researchers and engineers
- The second from entrepreneurs
- The Gartner Hype cycle models such path

https://www.gartner.com/en/research/methodologies/gartner-hype-cycle
https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020

The technology adoption cycle

- The Gartner Hype cycle models such path

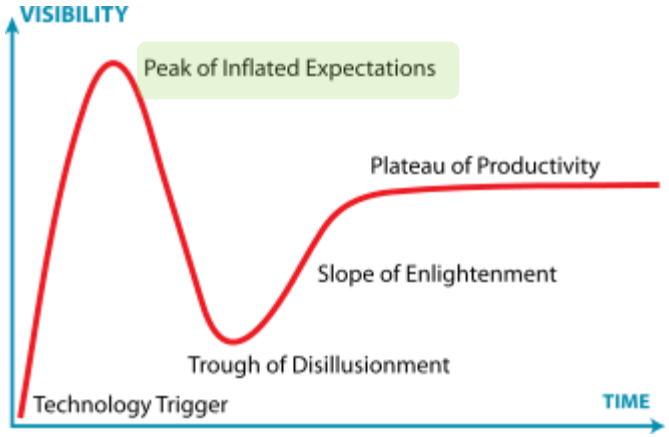
Innovation triggers: innovative subjects starts the adoption since they recognize the potential of the technology even in the absence of evidence of its usefulness



The technology adoption cycle

- The Gartner Hype cycle models such path

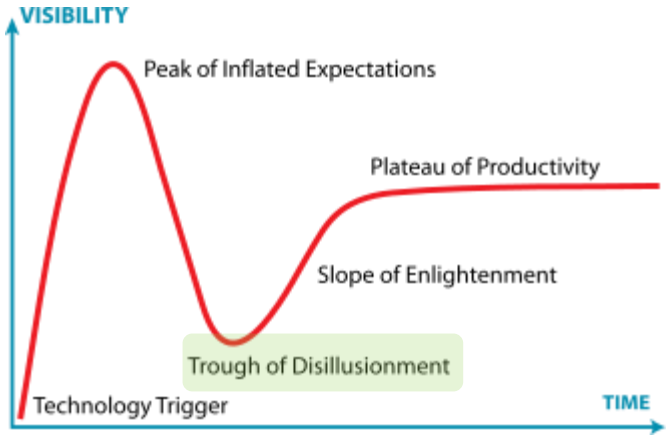
Peak of inflated expectations: media attention coupled with successful cases, often paired by many failed adoptions, lead to widespread use cases



The technology adoption cycle

- The Gartner Hype cycle models such path

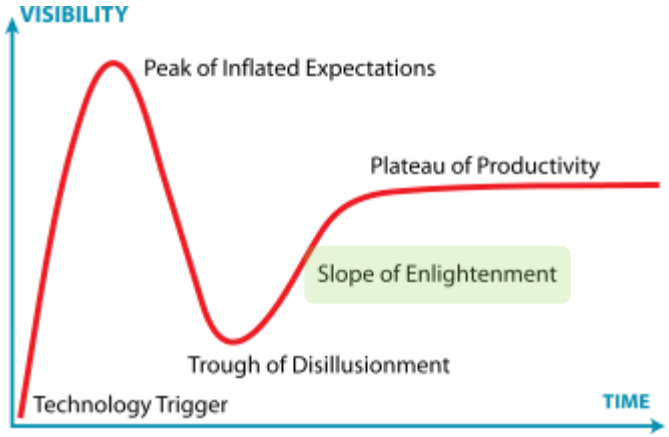
Trough of disillusions: the adoption of technology even in unsuitable contexts leads to an increase in failing cases



The technology adoption cycle

- The Gartner Hype cycle models such path

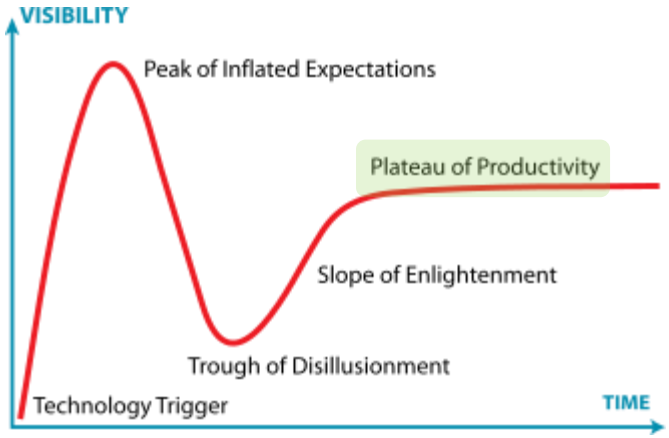
Slope of illumination: such a broad spectrum of applications allows to identify the fields of application in which the technology is effective and to make the technology itself evolve so that it can adapt to the contexts in which is actually useful



The technology adoption cycle

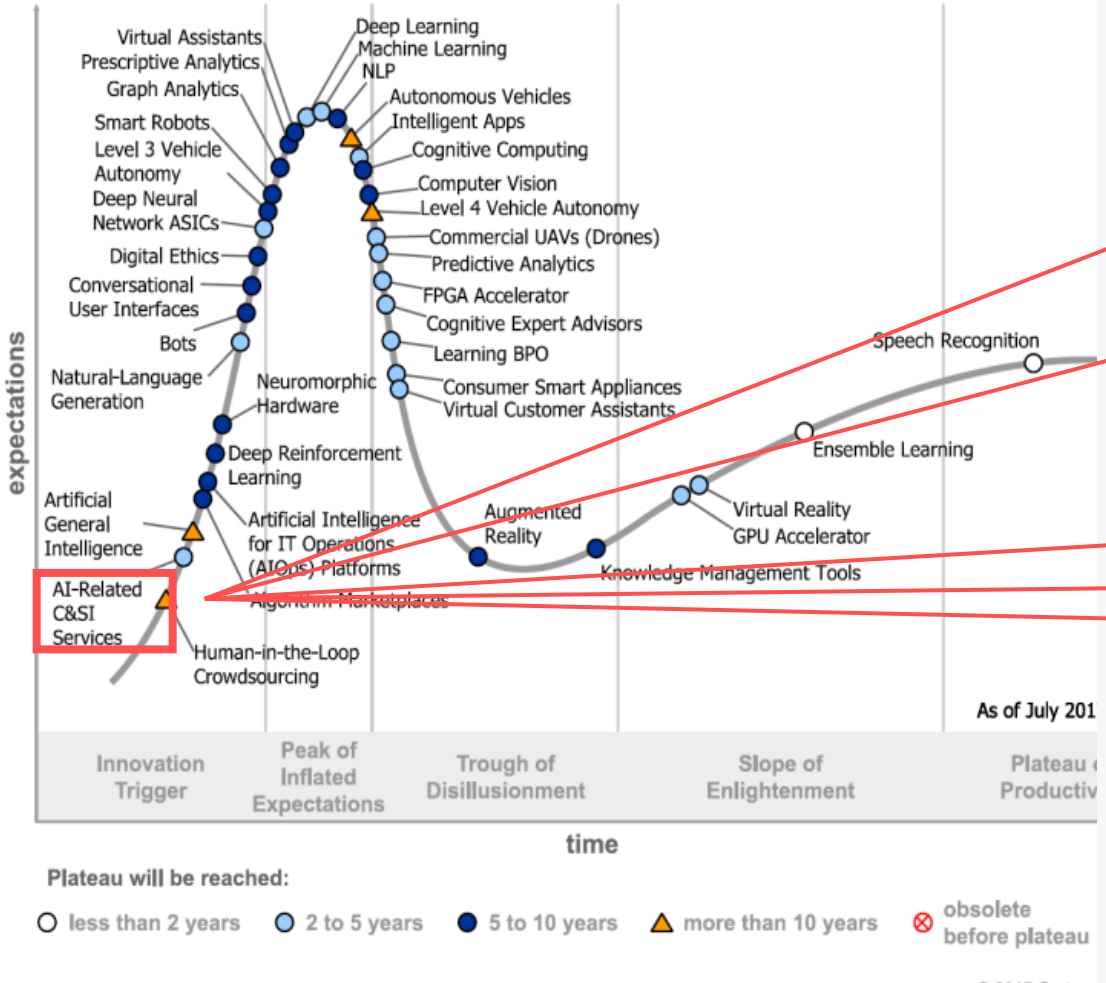
- The Gartner Hype cycle models such path

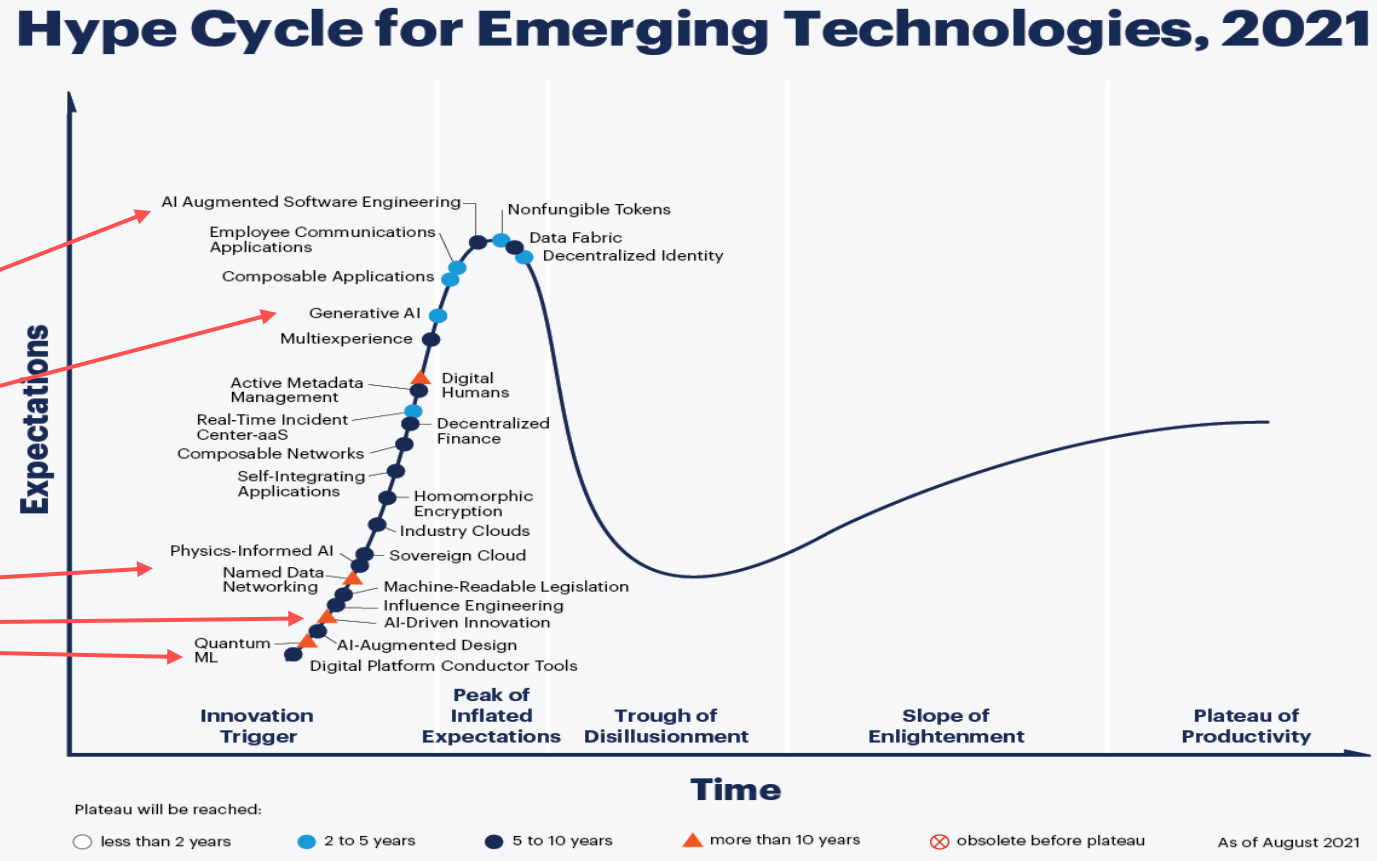
Plateau of productivity: … until it becomes mature, reliable and largely adopted



Hype cycle

Figure 1. Hype Cycle for Artificial Intelligence, 2017





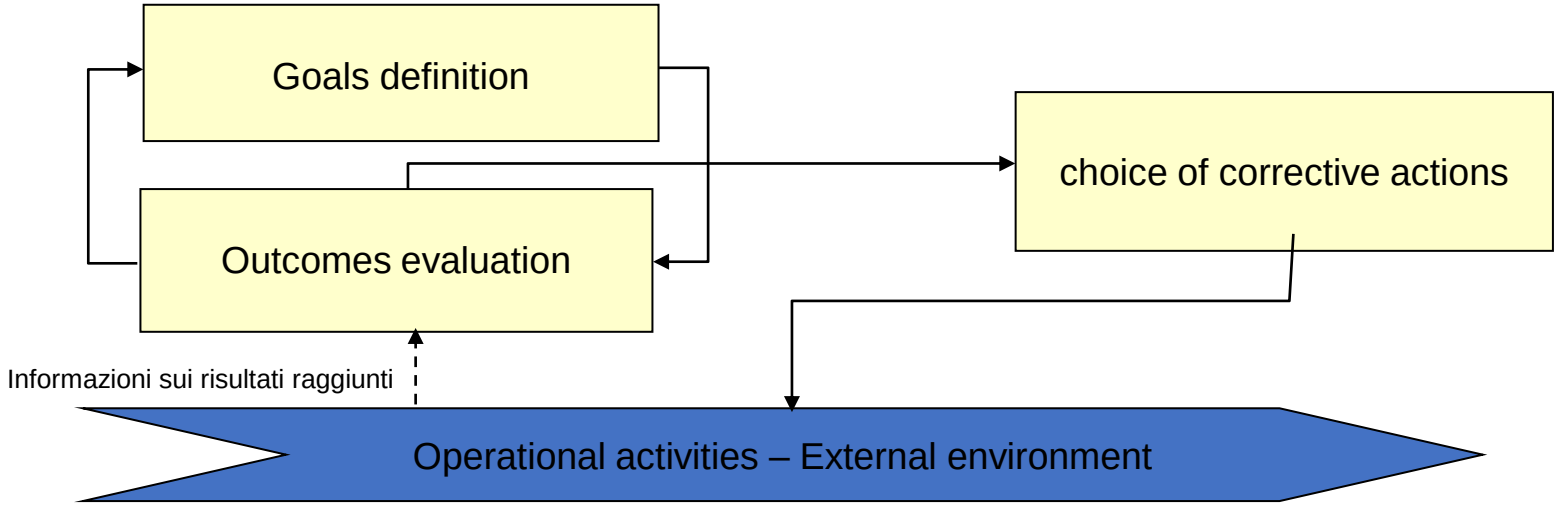
gartner.com

Gartner

Managerial and analytics information systems

Support the decisional process providing information to manager and knowledge worker

Their reference model is the control loop:

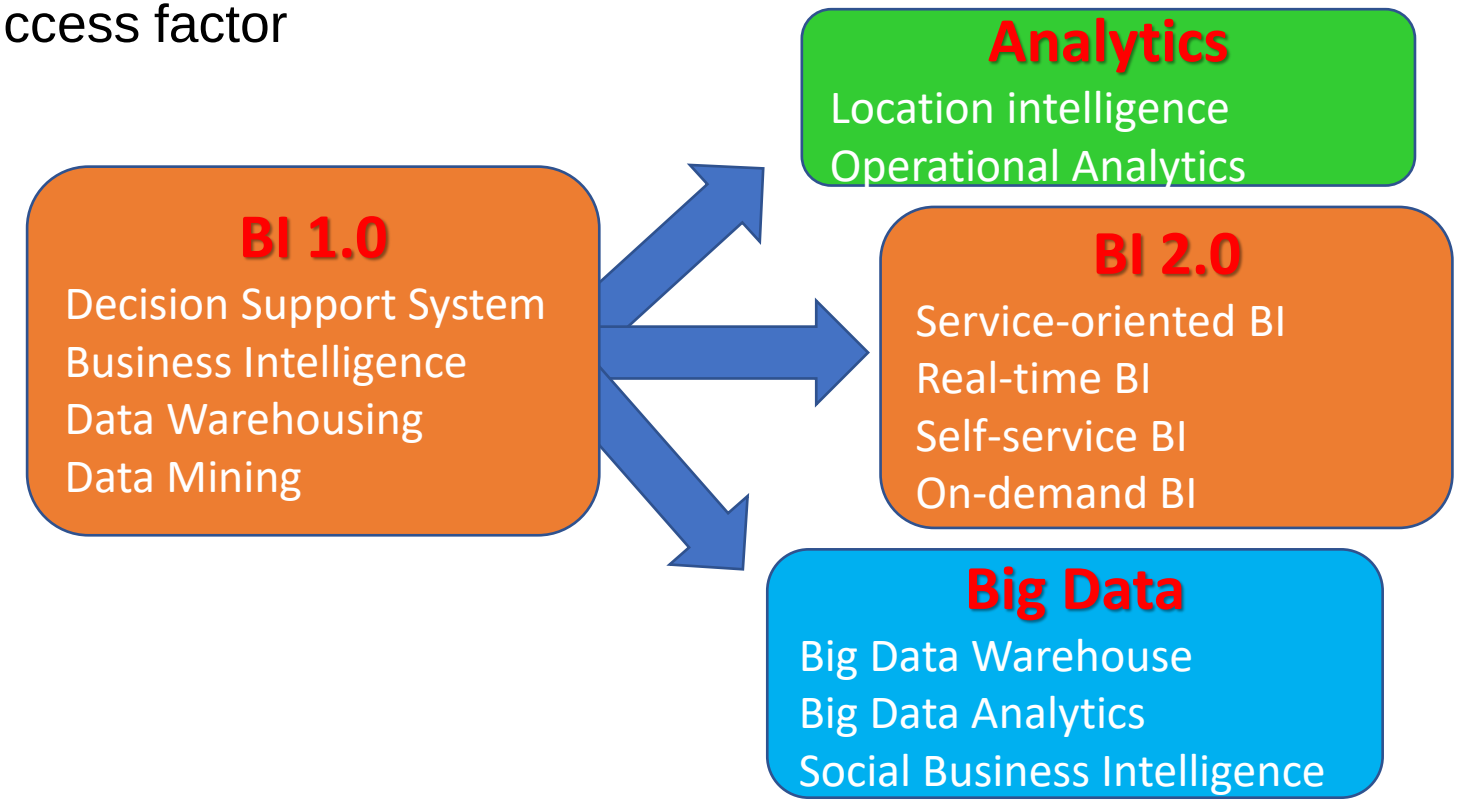


The managerial processes differ from the operational ones since:

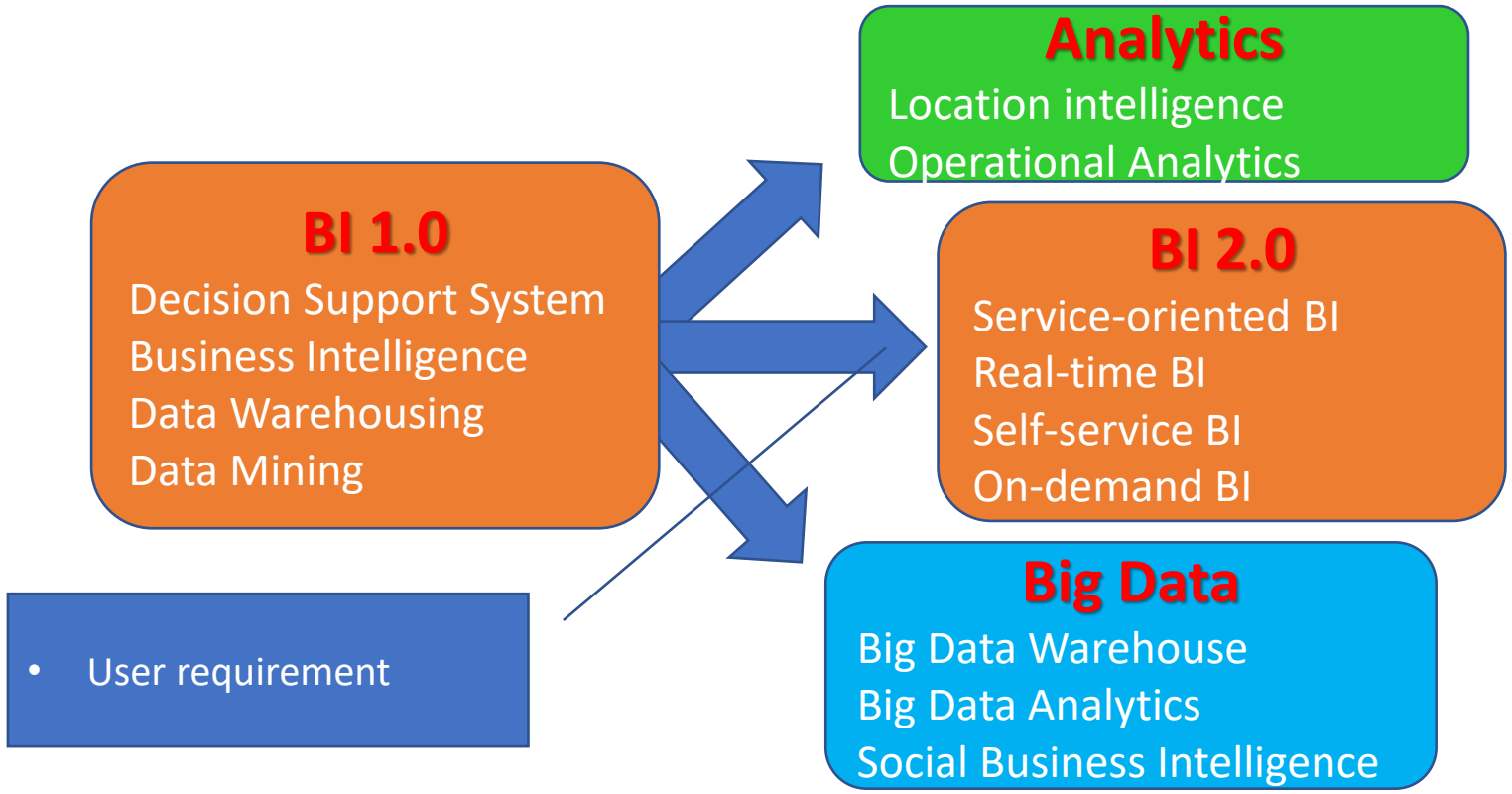
- Are based on indicators, that synthetic and aggregated
- Processing is periodical rather than continuous
- They rely on operational IS since they extract data from them

The evolution of analysis systems

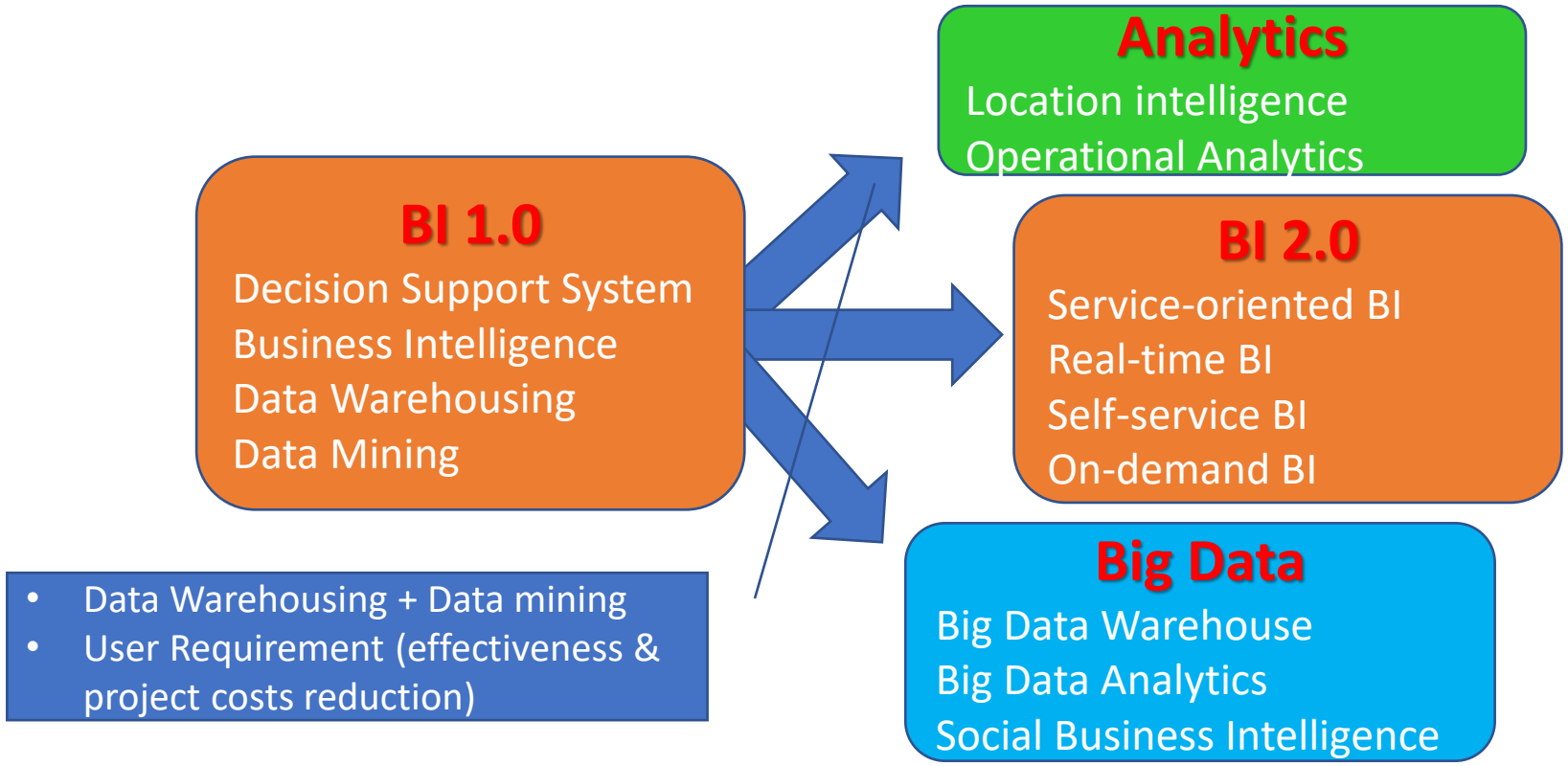
Since the introduction of computer in the industry at the end of ’70, the need of data analysis to support decision processes is increasing (sometimes slowly, sometimes very fast). This progressively changes the role of computer science in the companies making it a key success factor



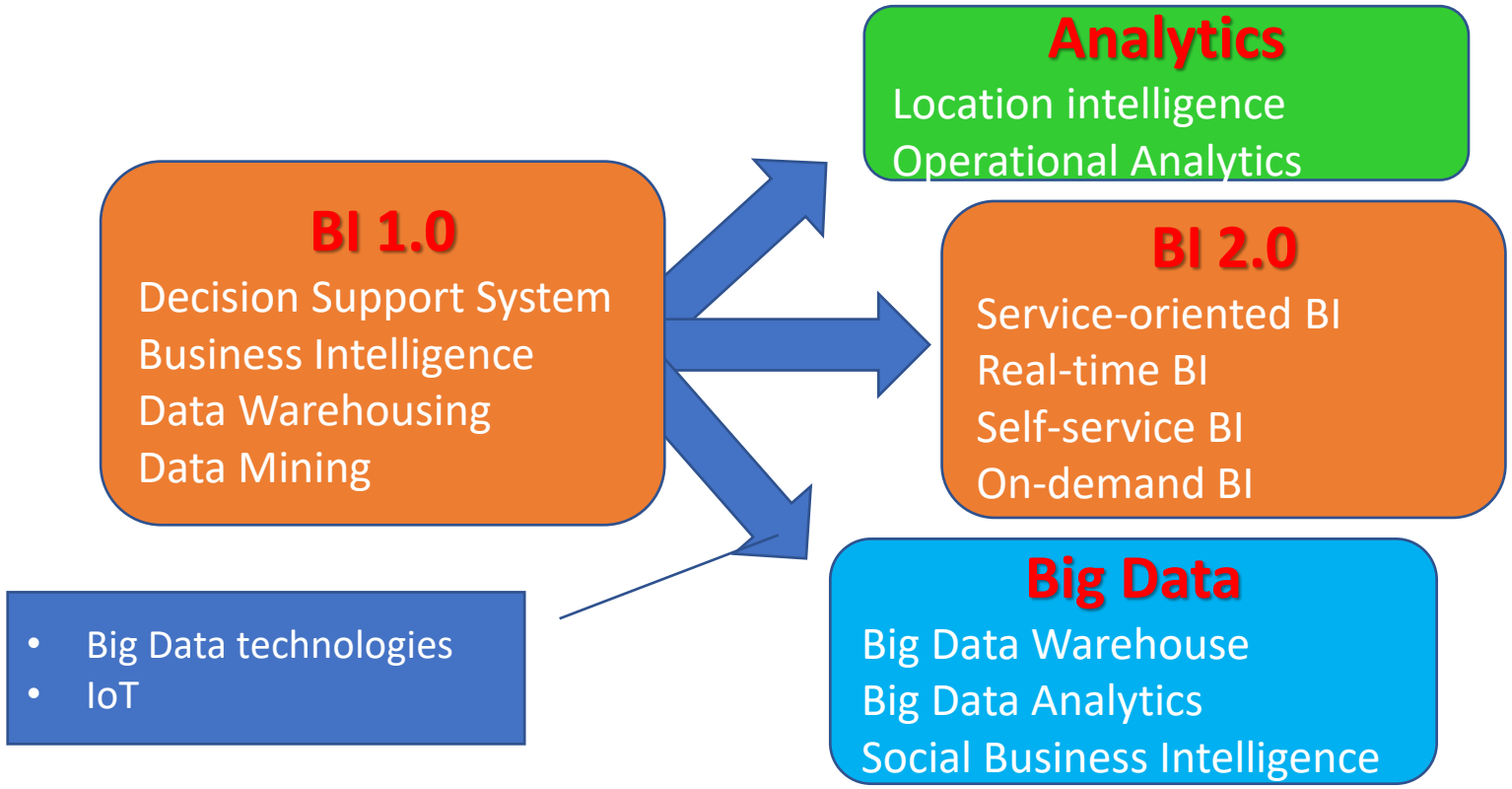
The evolution of analysis systems



The evolution of analysis systems



The evolution of analysis systems



The adoption path of BI

The adoption of BI solutions is incremental and rarely allows steps to be skipped

This is because it is risky, costly and useless to adopt advanced solutions before completely exploiting simple ones

Managers are not ready

- Not in the right mindset

Data are not ready

- Not of enough quality

Company processes are not ready

- Not defined to rely on and to be reactive to data

Beware of consultants and software vendors who offer advanced analytics if you barely exploit the corporate data warehouse

Turning your company in a data-driven one

The term data-driven company refers to companies where decisions and processes are supported by data

- Decisions are based on quantitative rather than qualitative knowledge
- Processes & Knowledge are an asset of the company and are not lost if managers change
- The gap between a data-driven decision and a good decision is a good manager

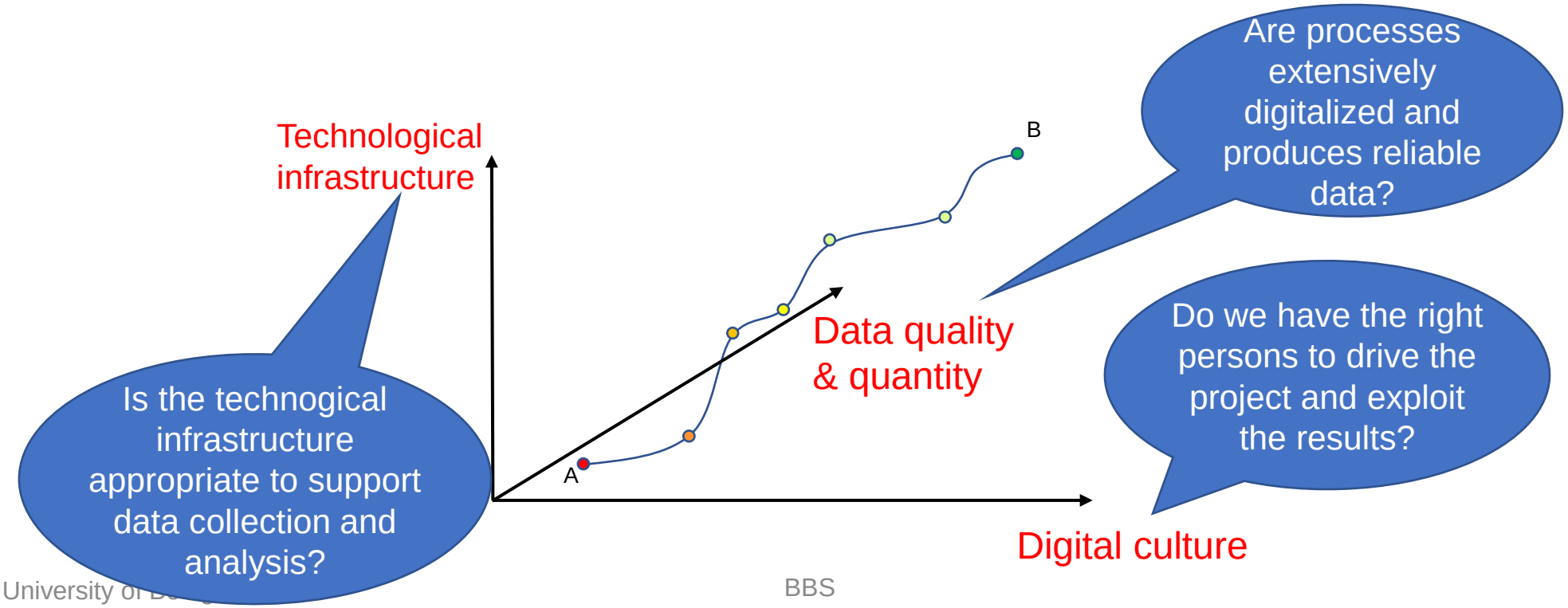
Adopting a data-driven mindset goes far beyond adopting a business intelligence solution and entails:

- Create a data culture
- Change the mindset of managers
- Change processes
- Improve the quality of all the data

Turning your company in a data-driven one

Digitalization is a journey that involves three main dimensions. Moving from A to B is a multi-year process made of intermediate goals each of which must be feasible

- Solves a company pain and brings value
- Can be accomplished in a limited time range (typically less than one year)
- Costs must be economically related to gains



Agenda

Introduction to BI 1.0

- Data Warehousing
- OLAP

From BI 1.0 to BI 2.0

- Analytics
- Big data

Data Mining & Machine Learning

- Examples with Weka
- Case studies

Agenda

Introduction to BI 1.0

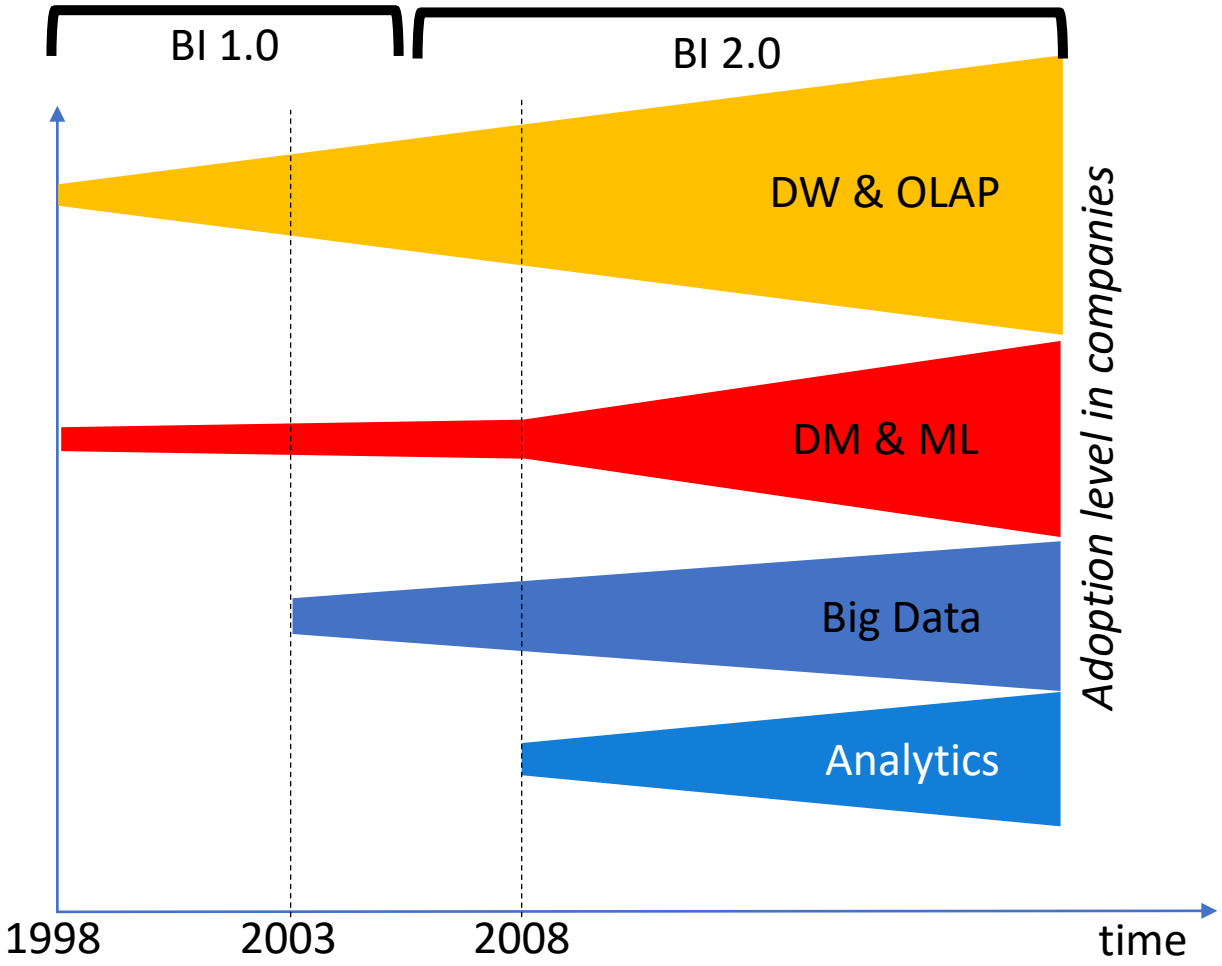
- Data Warehousing
- OLAP

From BI 1.0 to BI 2.0

- Analytics
- Big data

Data Mining & Machine Learning

- Examples with Weka
- Case studies



Agenda

Introduction to BI 1.0

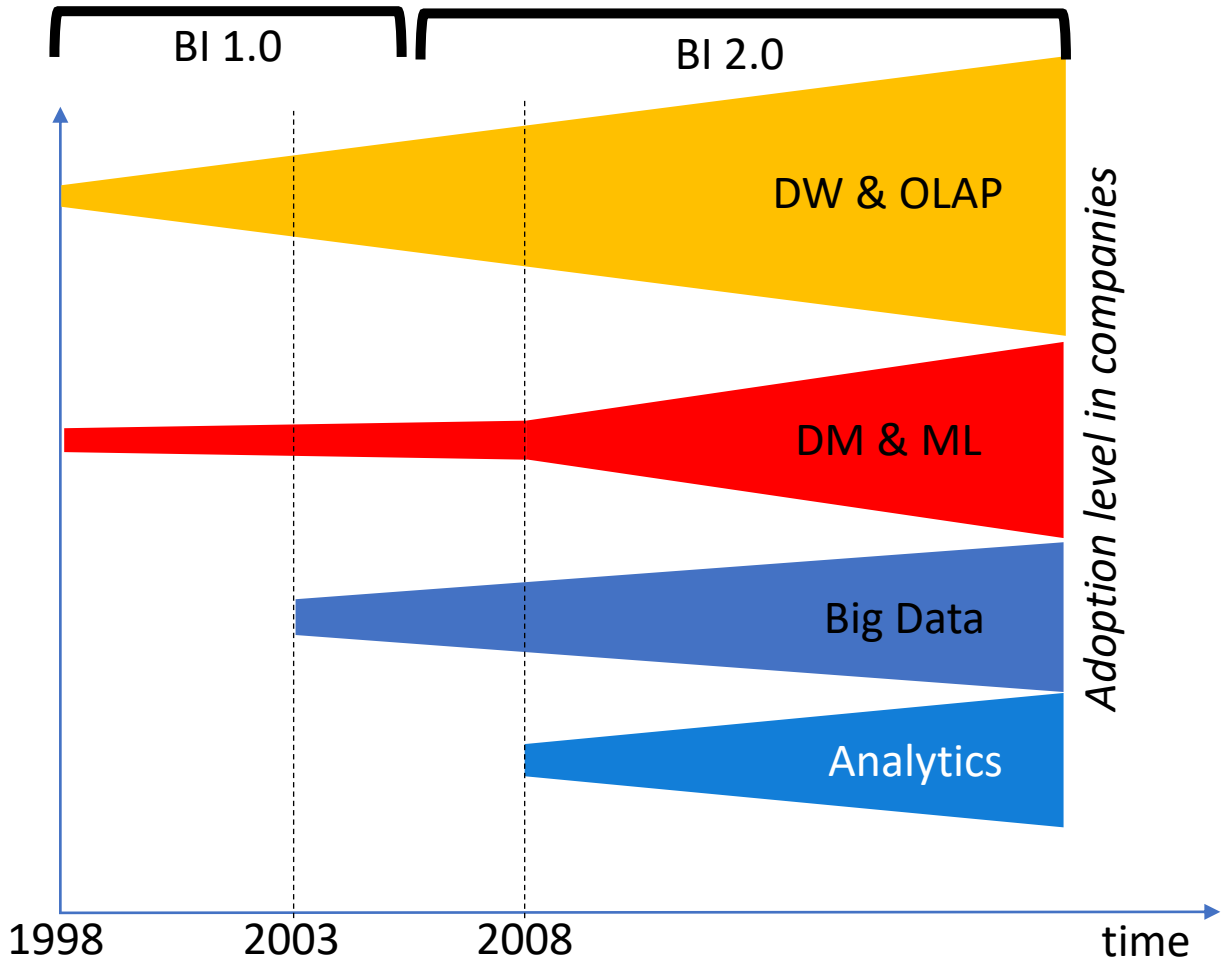
- Data Warehousing
- OLAP

From BI 1.0 to BI 2.0

- Analytics
- Big data

Data Mining & Machine Learning

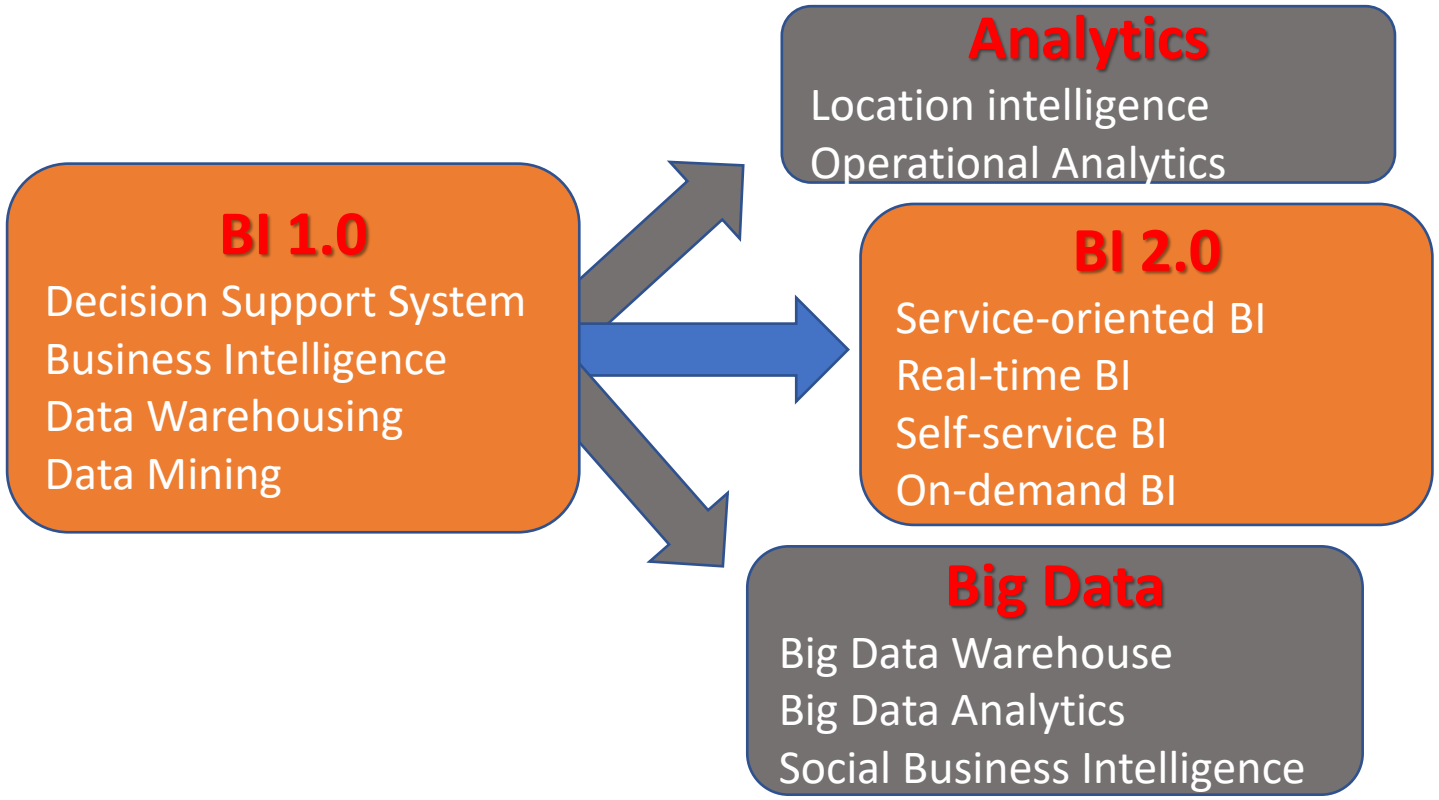
- Examples with Weka
- Case studies



Business Intelligence 1.0 Pros & Cons

Cons	Pros
Operates on structured data only	Best for analyzing the information coming from operational systems
Needs complex ETL processes expensive to build and edit	ETL guarantees high quality data
Implemented by IT staff and exploited by Business Users and data analysts	The business knowledge & the OLAP paradigm are sufficient to run it
Mainly implemented on Relational DBMS	IS based on mature technologies
It typically has a one-day loading interval	

The evolution of analysis systems



Service-Oriented BI

The BI platform is in the cloud and made available to the users through web services

Services can be made available according different models:

- Infrastructure-as-a-service, only the hardware is virtualized
- Platform-as-a-service, the cloud platform provides the core software (e.g., DBMS)
- Software-as-a-service, the whole BI software is part of the cloud platform (e.g., BIRST)
- Business process-as-a-service, part of the processes are outsourced too, this may introduce a dependence on the provider

Real-Time BI

Allow the analysis of the business operations that take place in *near* real time (from few minutes to few hours of latency)

- The DW architecture, based on periodical refreshes, must be modified
- Advanced solutions exploiting Trickle & Flip solutions are typically adopted

Areas of application are those in which:

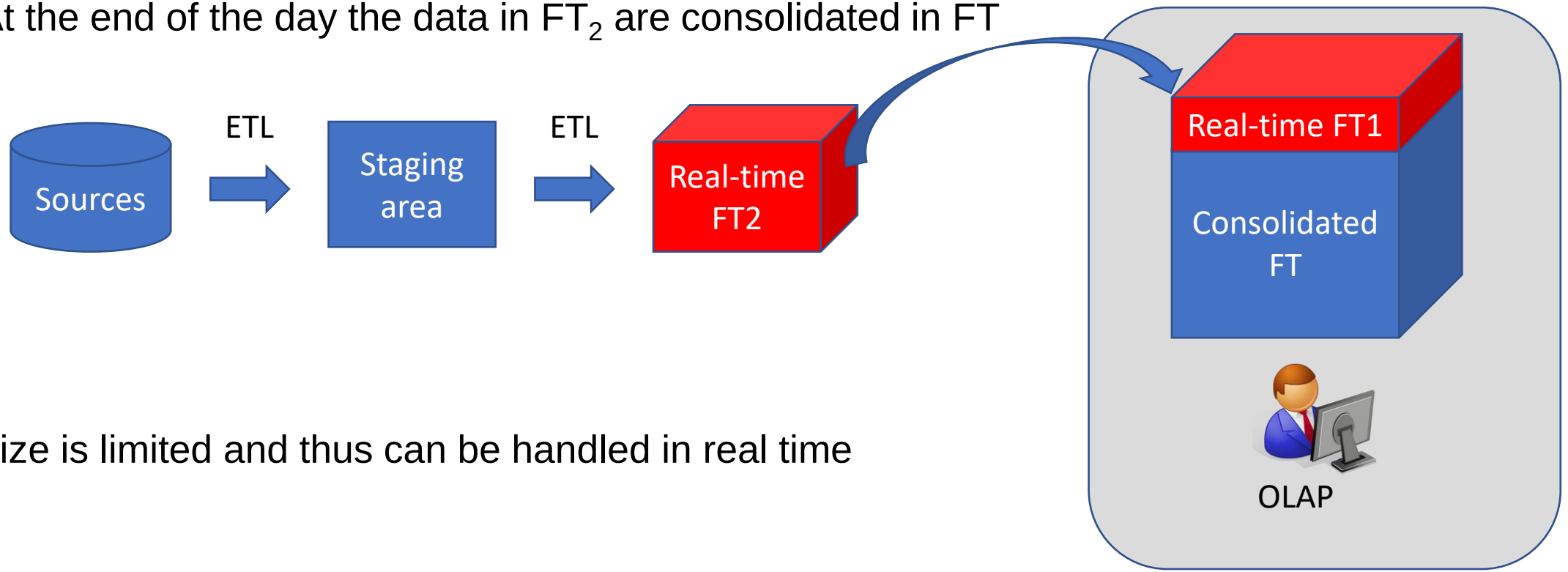
- The refresh interval is shorter than one day
- The queries require a mix of historical data (the majority) and up-to-date data

Real-Time DW differs from data stream analysis since:

- Data are progressively consolidated into the DW
- OLAP is the analysis paradigm
- Most of the data is consolidated

Trickle & Flip

1. The intra-day loading process updates FT_2 every t hours/minutes
   - FT_2 includes only data of the day
2. FT_2 is periodically substituted to FT_1 (the DW points to FT_2 instead on FT_1)
3. At the end of the day the data in FT_2 are consolidated in FT



FT_2 size is limited and thus can be handled in real time

Self-Service BI

An approach where business users can create analysis and reports proactively without mediating them with the IT staff

OLAP can be seen as a basic solution to self-service BI since it allows the users to freely query multidimensional cubes, more sophisticated applications are based on the sharing of business glossaries and meta-data glossaries and on the ability to dynamically alter and integrate useful data sources in decision-making

Systematization and assessment of business knowledge

DW is often underused because business users aren't aware that relevant information exist

- Often business users only know the portion of business processes they work on
- Departmental practices/dialects make terminology ambiguous
- With DW structure updates the set of available information get easily lost
- When a DW includes tens of cubes, hundreds of attributes and thousands of reports it is just impossible to be aware of which information can be jointly queried

Systematization of business knowledge requires:

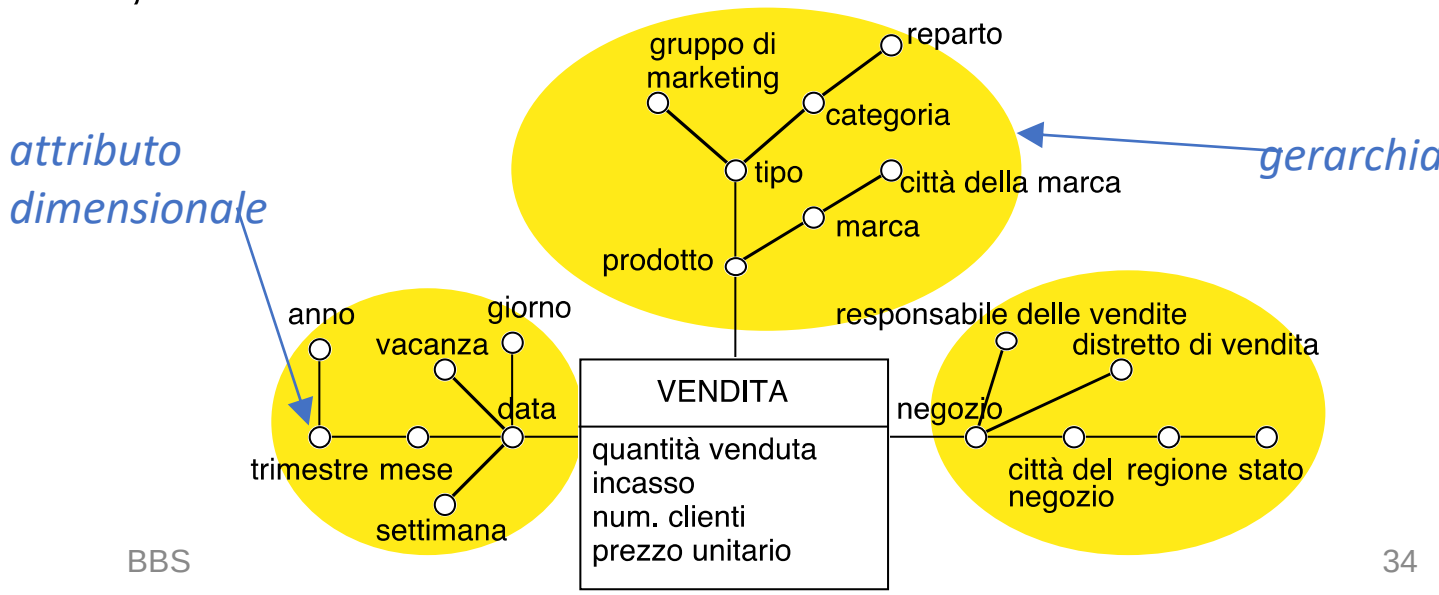
- A formal and user-friendly model for representing multidimensional cubes
- A tool specifically designed for knowledge visualization
- The possibility of integrating the knowledge into the OLAP tools

The Dimensional Fact Model & Indyco

The DFM is a graphical conceptual model for data mart design, devised to:

- lend effective support to conceptual design
- create an environment in which user queries may be formulated intuitively
- make communication possible between designers and end users
- build a stable platform for logical design (independently of the target logical model)
- provide clear and expressive design documentation

The conceptual representation generated by the DFM consists of a set of fact schemata that basically model facts, measures, dimensions, and hierarchies



Dimensional Fact Model & Indyco

Indyco is a tool commercialized by iConsulting that allows to:

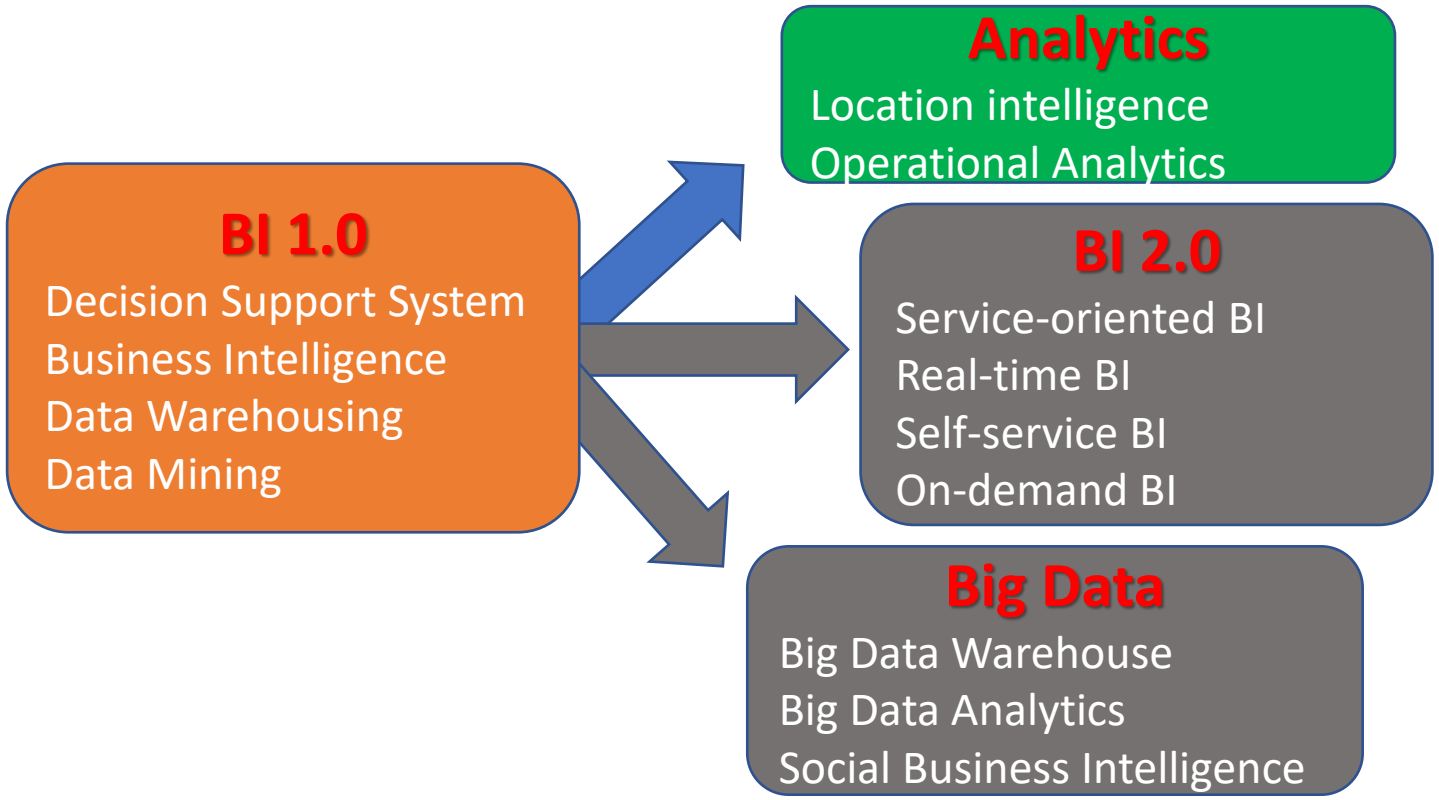
- Model business knowledge and creating glossaries
- Design and deploy multidimensional cubes on relational and big data platforms
- Track data from operational data sources to reports
- Integrate business knowledge up to business reports
- Carry out full text search of the business knowledge

Live time!

Case Study

Assortment Analysis

The evolution of analysis systems



Analytics

Analytics refers to software used to discovery, understanding and sharing of relevant patterns in data

- Analytics are based on the concurrent use of statistics, machine learning and operational research techniques
- Analytics often exploit advanced visualization techniques

Analytics in BI 2.0 play the same role data mining played in BI 1.0

Data Mining solutions have spread much less than DW ones due to the:

- Complexity and costs
- Needs of an expert for results understanding
- Lack of certainty of meeting the project goals

Analytics

Analytics are encountering a greater success since:

- More data is available
- More computational is available
- Higher corporate culture and increasing competition
- Strong focus on user-experience
- Solutions are more industrialized and user-ready
  - Lower deployment costs
  - Fruition is easier
  - Strong focus on a specific business problem

Analytics – a methodological approach

In data analysis projects, reaching the goal remains uncertain

Before undertaking a data analysis project it is necessary to assess:

1) Does the scope of the project represent a very strong pain for the company?
2) The solution of the problem would lead to a strong competitive advantage/gain/savings?
3) Does simpler solutions exist?
4) Are the necessary data available?
   - If the answer is negative it is mandatory to work on a data collection process
     - VALENTINO – CUSTOMER PROFILING: strong investment in the data collection system (i.e., CRM) and prizes to the salesmen that properly collect data
     - WAYNET – car prognostic maintenance: project suspended until the introduction of a new generation of control units that can collect more car data
5) Which business processes should be modified to exploit project outcomes?

Location Intelligence

Location Intelligence is a set of tools that allow **a geographic dimension** to be integrated within a **BI platform**. The goal is to increase the **monitoring ability** and the capability of understanding **business events**. Location intelligence supports **data visualization and interaction with maps** in **BI contexts**

More than 80% of companies take decisions on the basis of information characterized by a spatial component

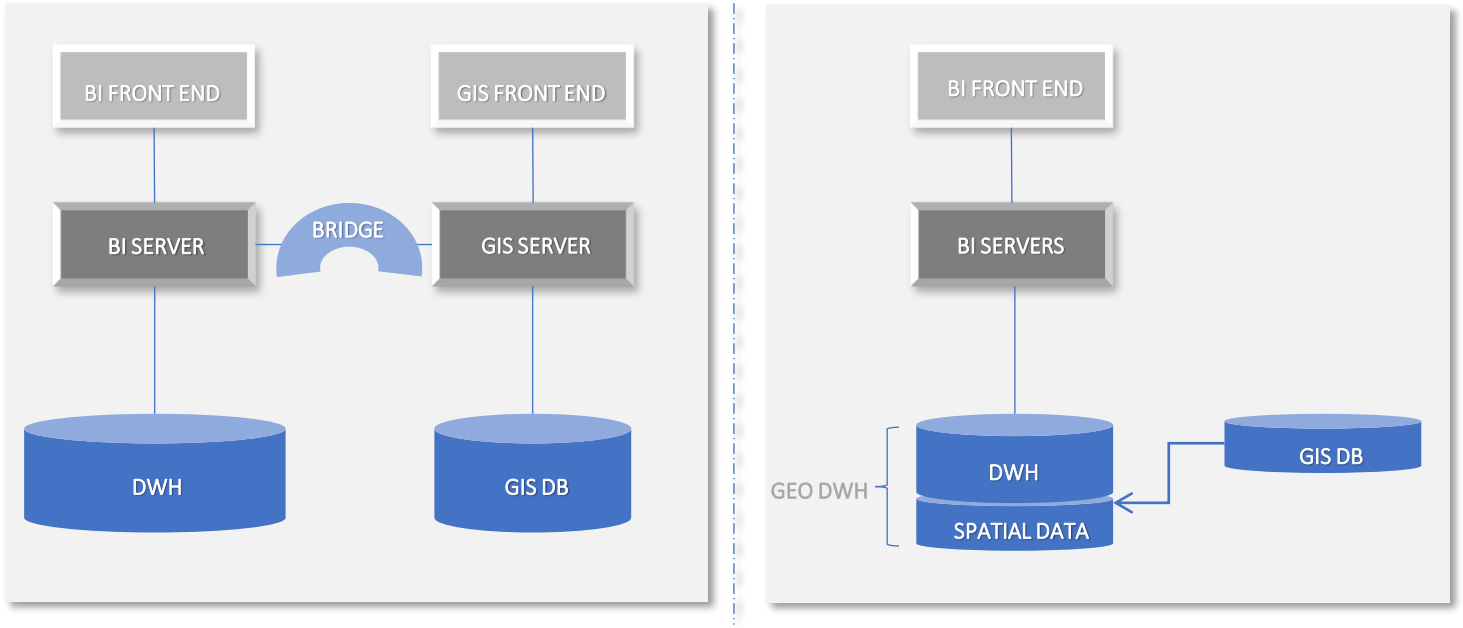
Architecture for Location Intelligence

Architecture classification

- Loosely-coupled: import-export-reformatting of data from GIS and OLAP.
- Semi-tightly coupled: GIS-centered solutions or OLAP-centered solutions
- Tightly-coupled: fully integrated solutions with GIS and OLAP technologies

Semi-tightly coupled

- Mixed queries are unfeasible
- Performance and data volume are limited
- 2 versions of reality

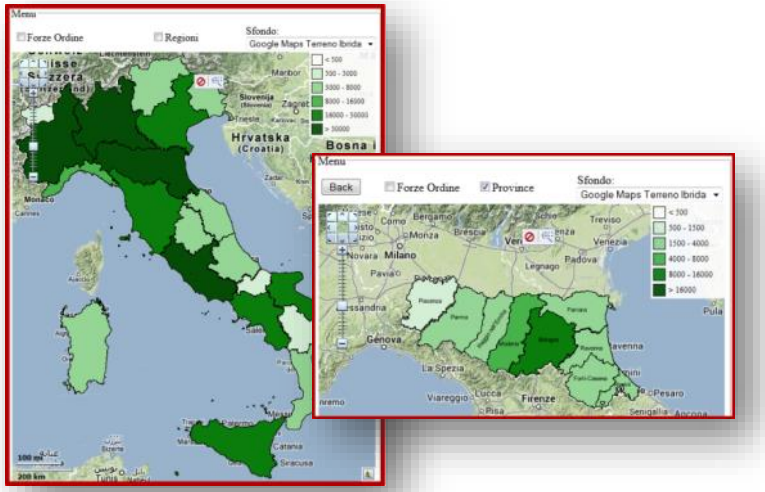


Tightly coupled

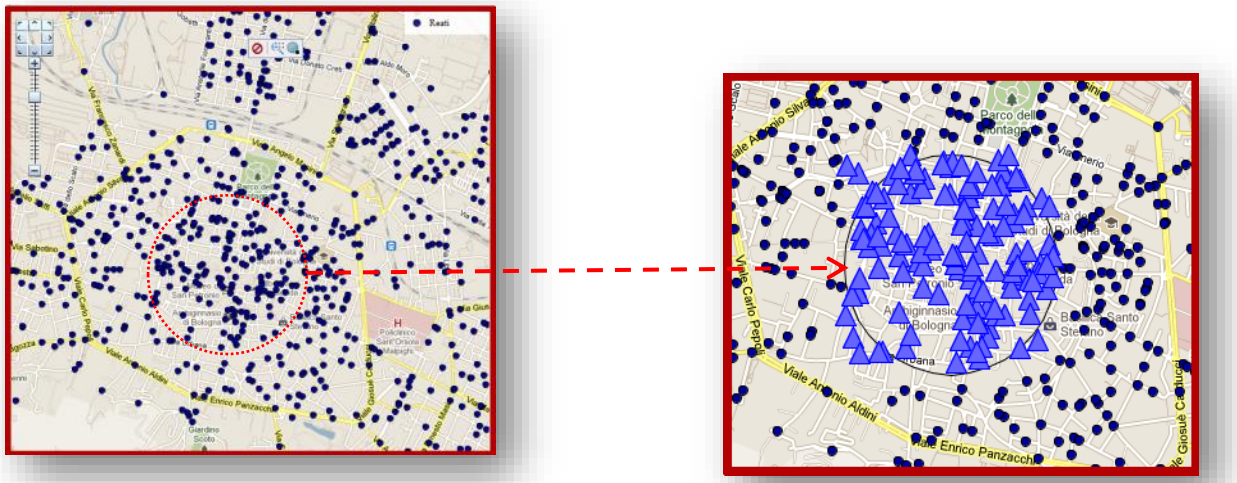
- Mixed queries
- Good performances even in presence of large quantity of data
- Integrated visualization

SOLAP analysis

Spatial drill-down

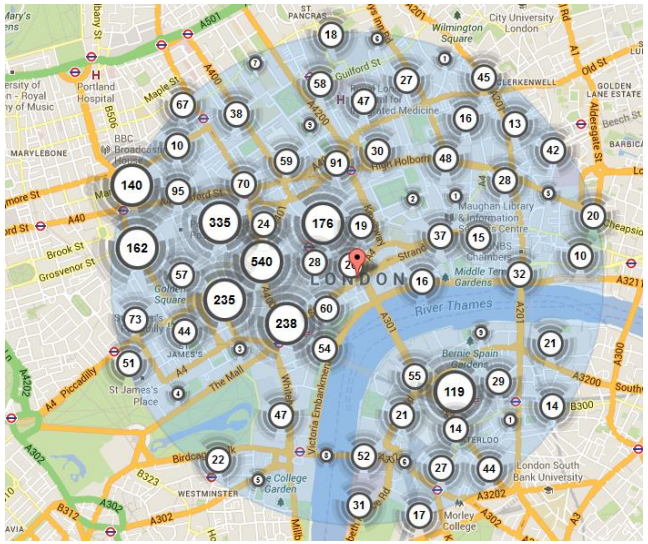


Spatial selection



Location Intelligence

Spatial roll-up/spatial clustering



Location Intelligence Demo

Live time!

Case Study

Social Habits Analysis