Title: Introduction

Course: Data Mining

Instructor: Claudio Sartori

Master: Data Science and Business Analytics

Academic Year: 2022/2023

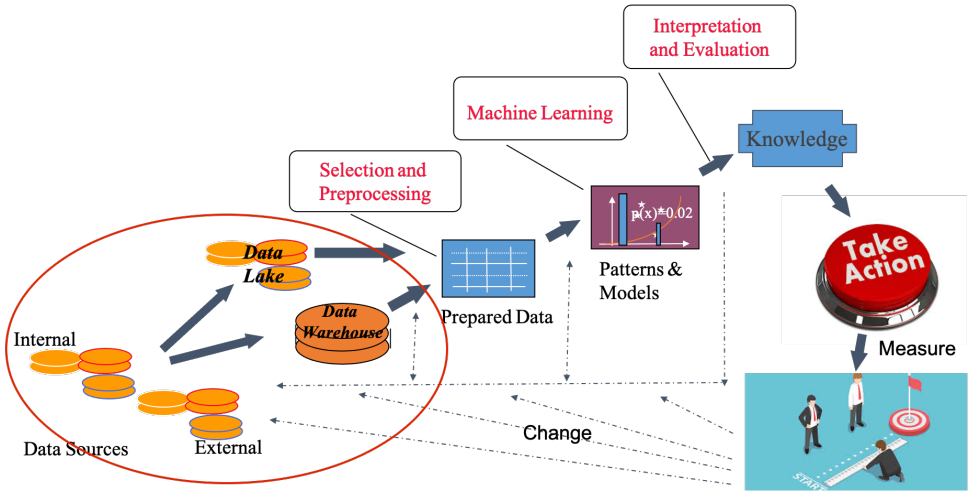
BOLOGNA BUSINESS SCHOOL

Alma Mater Studiorum Università di Bologna

Context

- Machine Learning
- Business Intelligence and Data Warehouse (DSBA only)
- Big Data For Industry
- Big Data Lab (DSBA only)
- Text Mining (DSBA only)
- Natural Language Processing and Applications (DTI and FFT only)

Where are we now? - The Data Mining Process



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

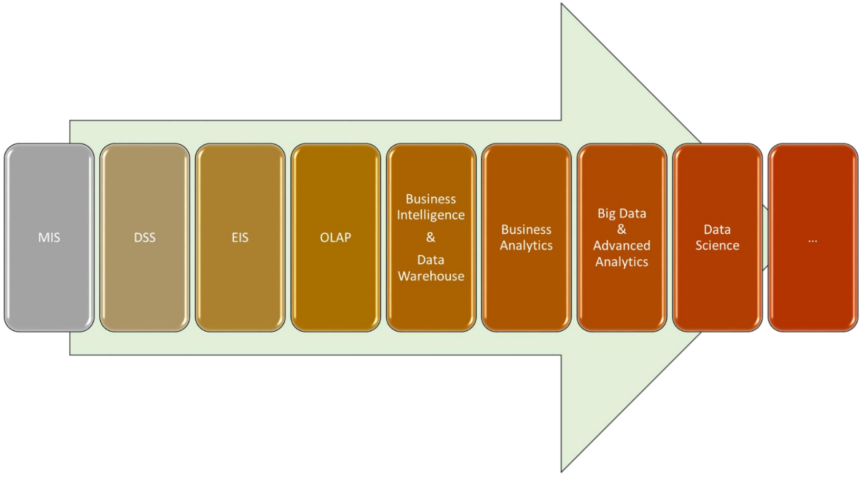
Data: a collection of raw value elements

Information: the result of collecting and organising data

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

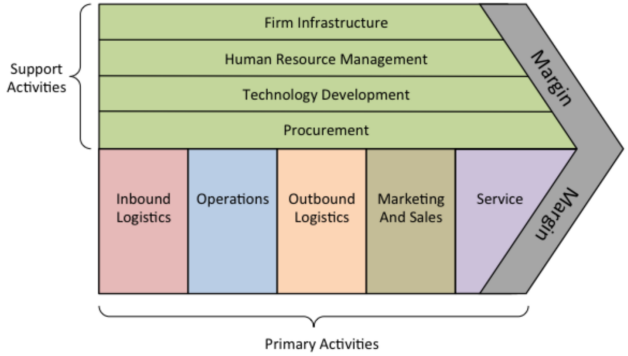
Knowledge: understanding information based on recognising patterns

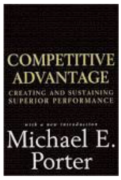
Increasing insights



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)





https://en.wikipedia.org/wiki/Value_chain

Where does *data* come from? 2/2

- When an event in the real world *changes the state* of the enterprise, one of 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)
- *signals* are produced by sensors
- Data may be made available by external organizations

Structured vs unstructured decisions

Structured		*Unstructured*	
Description	***Example***	***Description***	***Example***
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	Suddeen 0ne-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

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

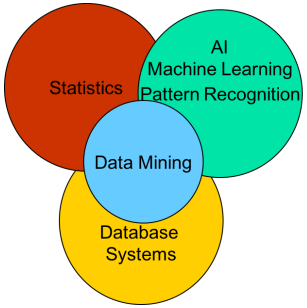
Thomas More[1]

Watch this: https://www.youtube.com/watch?v=EH3bp5335IU

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

Data Mining Origins

- The sizes of the circles to 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 this module we will use the topic names as follows

- Data mining is the discovery process described in page 5
- 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

What's in this course

- The CRISP-DM methodology for Data Mining processes (Sartori)
- Pre-processing: Data types and data conversions. Data transformations Data reduction. Feature creation (Sartori)
- Hands-on data mining and machine learning (Francia and Gallinucci)
- Hands-on Spark and OLAP (Francia and Gallinucci)