Data Strategy & Business Analytics

Integrated Analytics Lab

whoami

Matteo Francia
- Email: m.francia@unibo.it
- Research fellow @ UniBO
- Adjunct professor @ UniBO

Research topics
- Big data / database
- Geo-spatial analytics

https://big.csr.unibo.it/

Analytics

Business intelligence

- Strategies to transform raw data into decision-making insights

Analytics

- A catch-all term for a variety of different business intelligence and application-related initiatives
- The process of analyzing data from a particular domain (e.g., sales and supply chain)

Advanced Analytics

- (Semi-)Autonomous transformation of data using techniques and tools, to discover deeper insights, make predictions, or generate recommendations
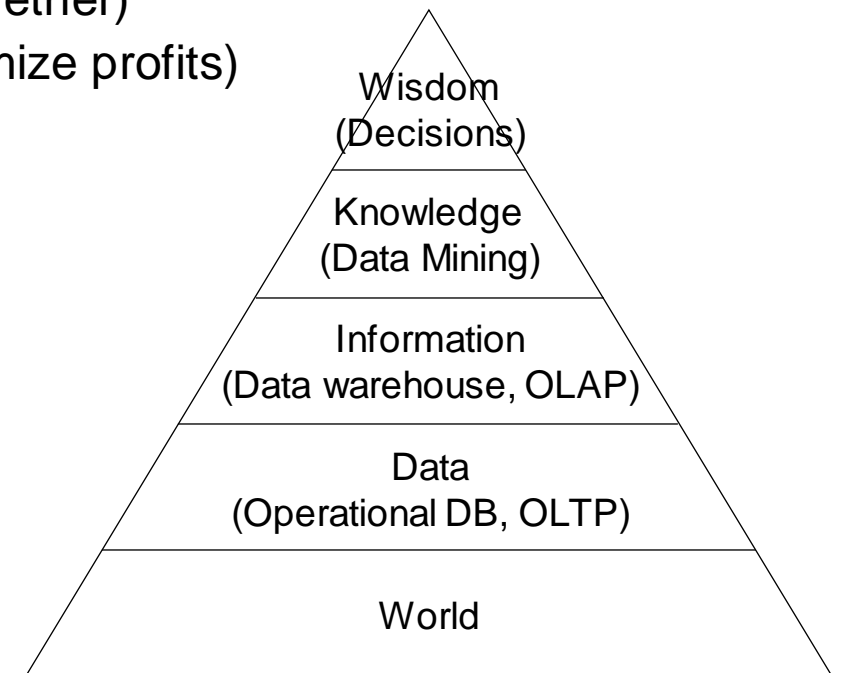
Integrated Analytics (Lab)

- Analytics are based on the usage of statistics, machine learning, operational research, and advanced visualization techniques

https://www.gartner.com/en/information-technology/glossary?glossarykeyword=analytics

The knowledge pyramid

Family of transformations are usually abstracted in the "knowledge pyramid"

- **Data:** symbols representing real-word objects (e.g., store product sales)
- **Information:** processed data (e.g., query the product with highest profit)
- **Knowledge:** understanding (e.g., mine products often sold together)
- **Wisdom:** knowledge in action (e.g., discount products to optimize profits)

Wisdom
(Decisions)

Knowledge
(Data Mining)

Information
(Data warehouse, OLAP)

Data
(Operational DB, OLTP)

World

[1] Jennifer E. Rowley: The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. 33(2): 163-180 (2007)
[2] Martin Frické: The knowledge pyramid: a critique of the DIKW hierarchy. J. Inf. Sci. 35(2): 131-142 (2009)

CRISP-DM

Data transformation requires a structured approach

- Choosing the best algorithm is only one of the success factors

Cross-industry standard process for data mining (CRISP-DM) is a model that describes common approaches used by data mining experts

CRISP-DM

CRISP-DM breaks the process of data mining into six major phases

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

The sequence of phases is not strict

- Arrows indicate the most important and frequent dependencies between phases
- The outer circle in the diagram symbolizes the cyclic nature of data mining itself

CRISP-DM

Understanding the domain

- Understanding project goals from the user's point of view, translate the user's problem into a data mining problem, and define a project plan

Understanding the data

- Preliminary data collection aimed at identifying quality problems and conducting preliminary analyzes to identify the salient characteristics

Data preparation

- Includes all the tasks needed to create the final dataset: selecting attributes and records, transforming and cleaning data

CRISP-DM

Model Creation

- Several data mining techniques are applied to the dataset also with different parameters in order to identify what makes the model more accurate

Evaluation of model and results

- The model(s) obtained from the previous phase are analyzed to verify that they are sufficiently precise and robust to respond adequately to the user's objectives

Deployment

- The built-in model and acquired knowledge must be made available to users. This phase can therefore simply lead to the creation of a report or may require implementation of a user-controlled data mining system

The full picture
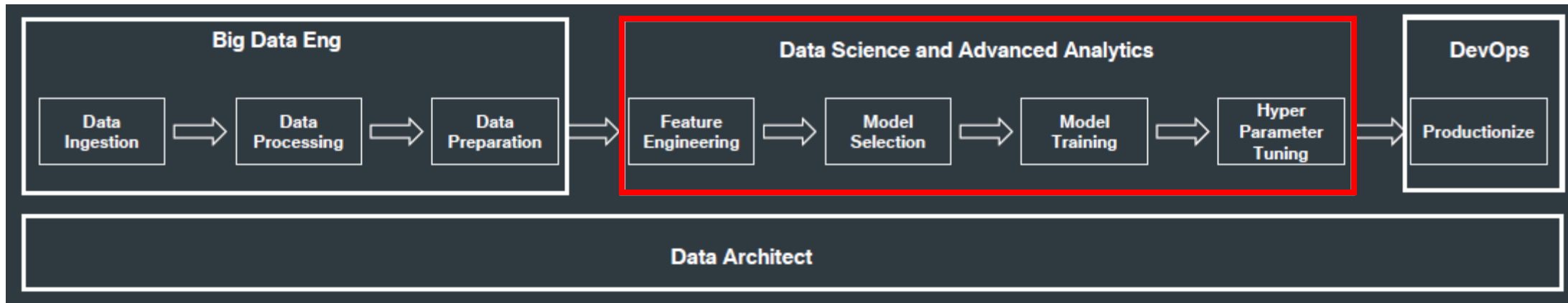
GOAL of this lab

Move through transformation phases

Integrated analytics lab

This checklist can help you while building your projects

- Frame the problem and look at the big picture
- Get the data
- Explore the data to gain insights
- Prepare the data
- Explore many different models and shortlist the best ones
- Fine-tune your models and combine them
- Present your solution
- Launch, monitor, and maintain your system

(Tentative) Time Schedule

Feel free to interrupt and ask questions

The time schedule can change

	Time	Activity
	Time	**Activity**
Day 1	14:15 – 15:30	Introduction to integrated analytics and Python
	15:30 – 15:40	Break
	15:40 – 16:30	Hands on data preprocessing
	16:30 – 16:40	Break
	16:40 – 17.45	Hands on machine learning
Day 2	9:15 – 10:30	Introduction to massive data processing
	10:30 – 10:40	Break
	10:40 - 11:30	Hands on big data
	11:30 – 11:40	Break
	11:40 – 13:00	Hands on OLAP and visualization