whoami

Matteo Francia

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

Research topics

- Big data / database
- Geo-spatial analytics

BUSINESS INTELLIGENCE GROUP

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

Data Strategy & Analytics (VI ed.)

Integrated Analytics Lab

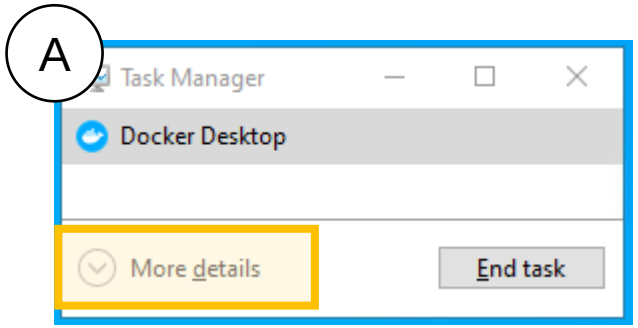
Before starting…

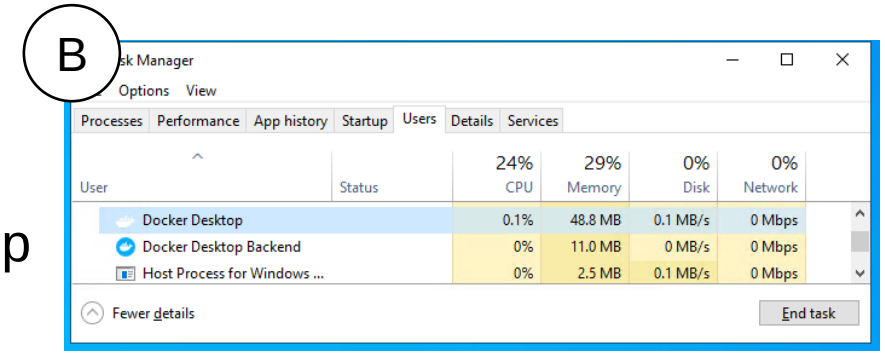
1. Turn on the virtual machine
2. Log in the virtual machine
   - *We will work on the virtual machine*

Before starting…

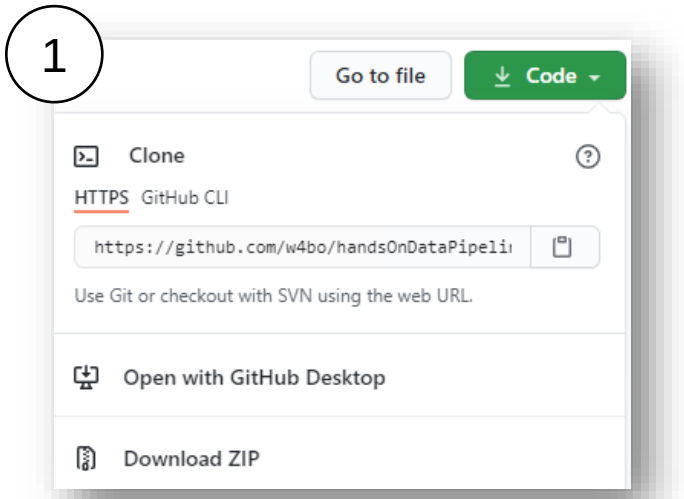
Run docker

1. In the Windows desktop, double click on Docker Desktop
2. After some minutes the Docker Engine will fail
3. Open the task manager (CTRL + SHIFT + ESC)
   A. Click on `More details`
   B. Click on `Users`
   C. Expand `bbsstudent`
   D. Click on `Docker Desktop`
   E. Click on `End Task`
4. In the Windows desktop, double click on Docker Desktop
5. Wait until the Docker icon becomes green

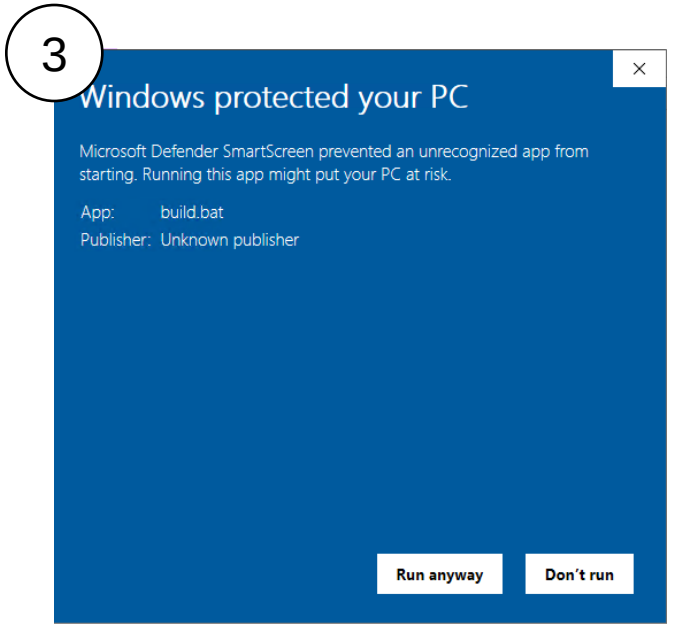




Before starting…



1. Download the content from https://github.com/w4bo/handsOnDataPipelines
   - Use Google Chrome (if possible)
   - Click `Code` and then `Download Zip`
   - Extract the zip in the `Downloads` folder
   - Make sure that the path does not contain any spaces
2. Enter the project directory
3. Double click on `build.bat`
   - Windows will complain, click on `More info` and then ` Run anyway`
   - This will take some minutes, let's switch to the slides



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 information from a particular domain (e.g., sales and supply chain)
- Analytics are based on the usage of statistics, machine learning, operational research, and advanced visualization techniques

Advanced Analytics

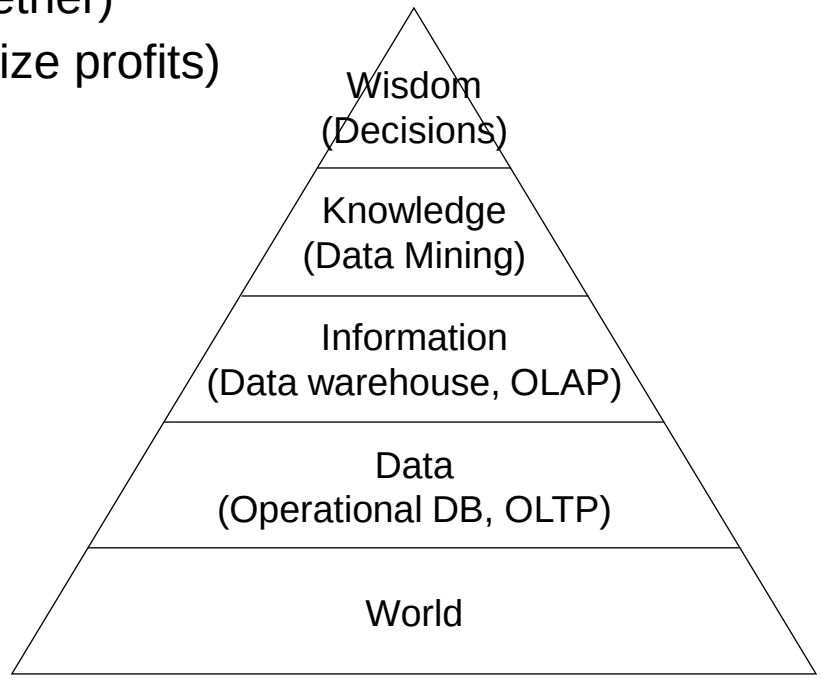
- (Semi-)Autonomous transformation of data or content using sophisticated techniques and tools, to discover deeper insights, make predictions, or generate recommendations

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)



[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)

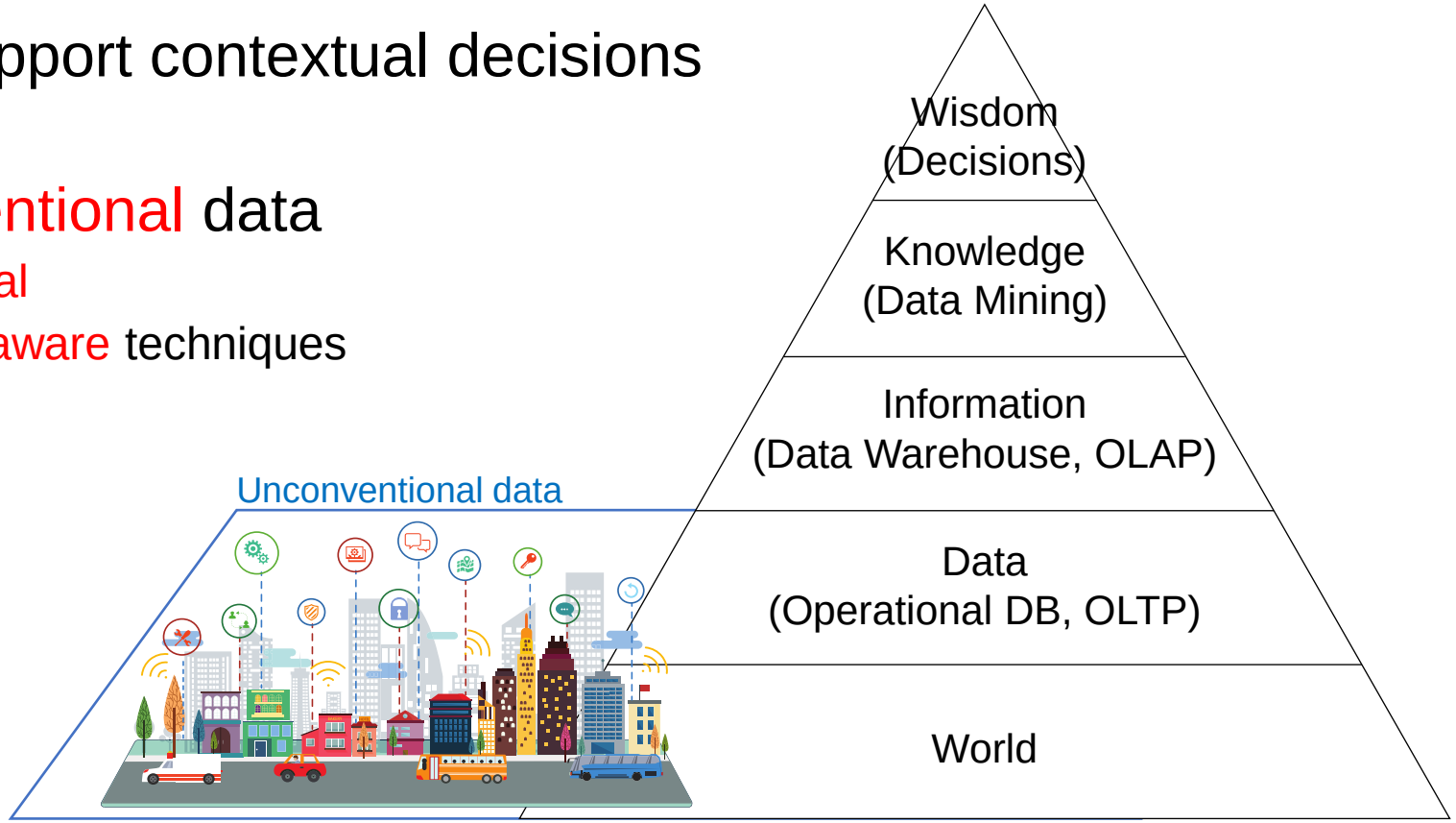
Challenges: unconventional data

Sensing provides data to support contextual decisions

- “World” and “Data” levels

New challenges on unconventional data

- Unstructured and non-relational
- Transformation requires type-aware techniques



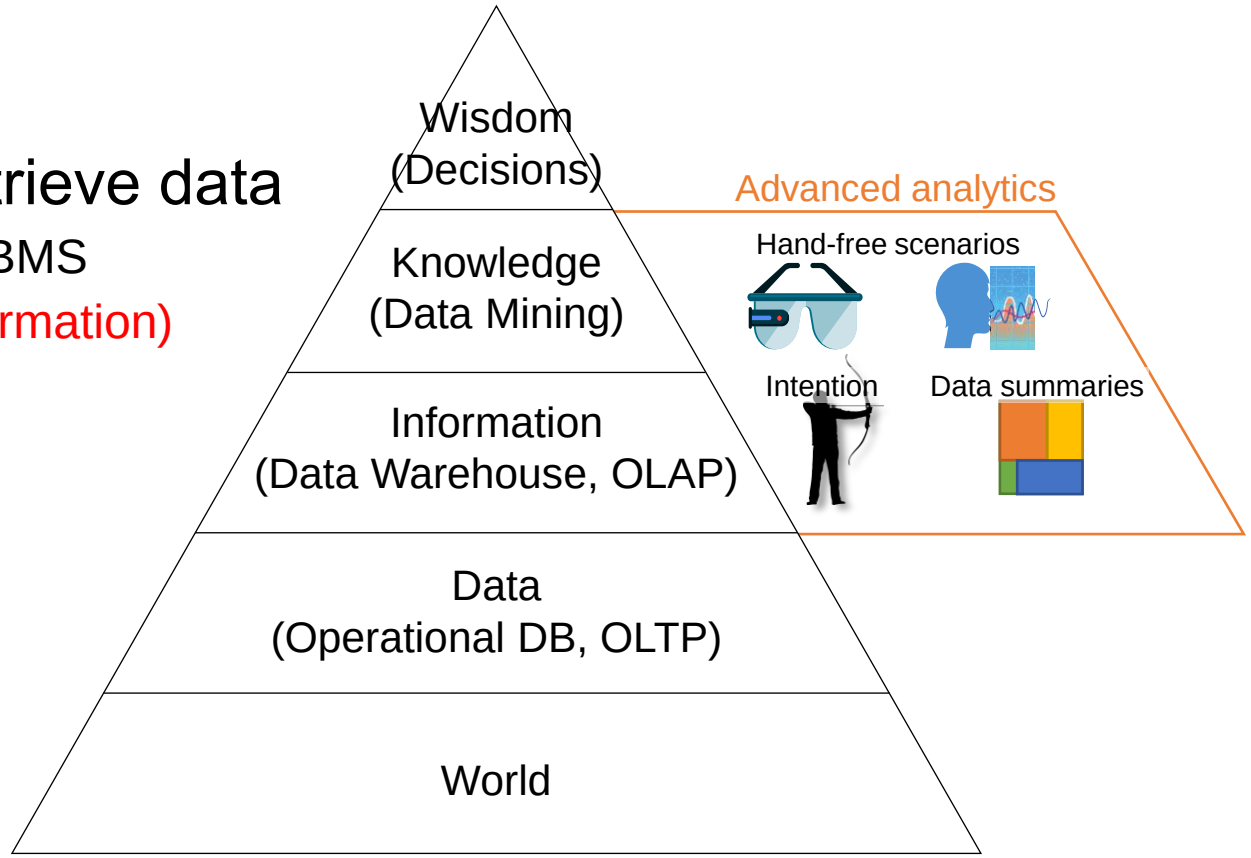
Challenges: advanced analytics

High availability and accessibility attract new data scientists

- High competence in business domain
- Low competence in computer science

Since the ’70s, relational queries to retrieve data

- Comprehension of formal languages and DBMS
- Advanced analytics (semi-automatic transformation)



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

GOAL of this lab

Move through transformation phases

https://forms.gle/EP4VJ9nvwJ2BpGTq8

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 into a great solution
- 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
9:15 – 10:30	Introduction to integrated analytics
10:30 – 10:40	Break
10:40 - 11:30	Hands on data preprocessing
11:30 – 11:40	Break
11:40 – 13:00	Hands on machine learning
13:00 – 14:00	Launch break
14:15 – 15:30	Introduction to massive data processing
15:30 – 15:40	Break
15:40 – 16:30	Hands on big data
16:30 – 16:40	Break
16:40 – 17.45	Hands on OLAP and visualization