Data Management and Visualization

INTRODUCTION TO THE COURSE

Academic Year 2021-2022

Daniele Apiletti

What is data management? (1)

Data management is

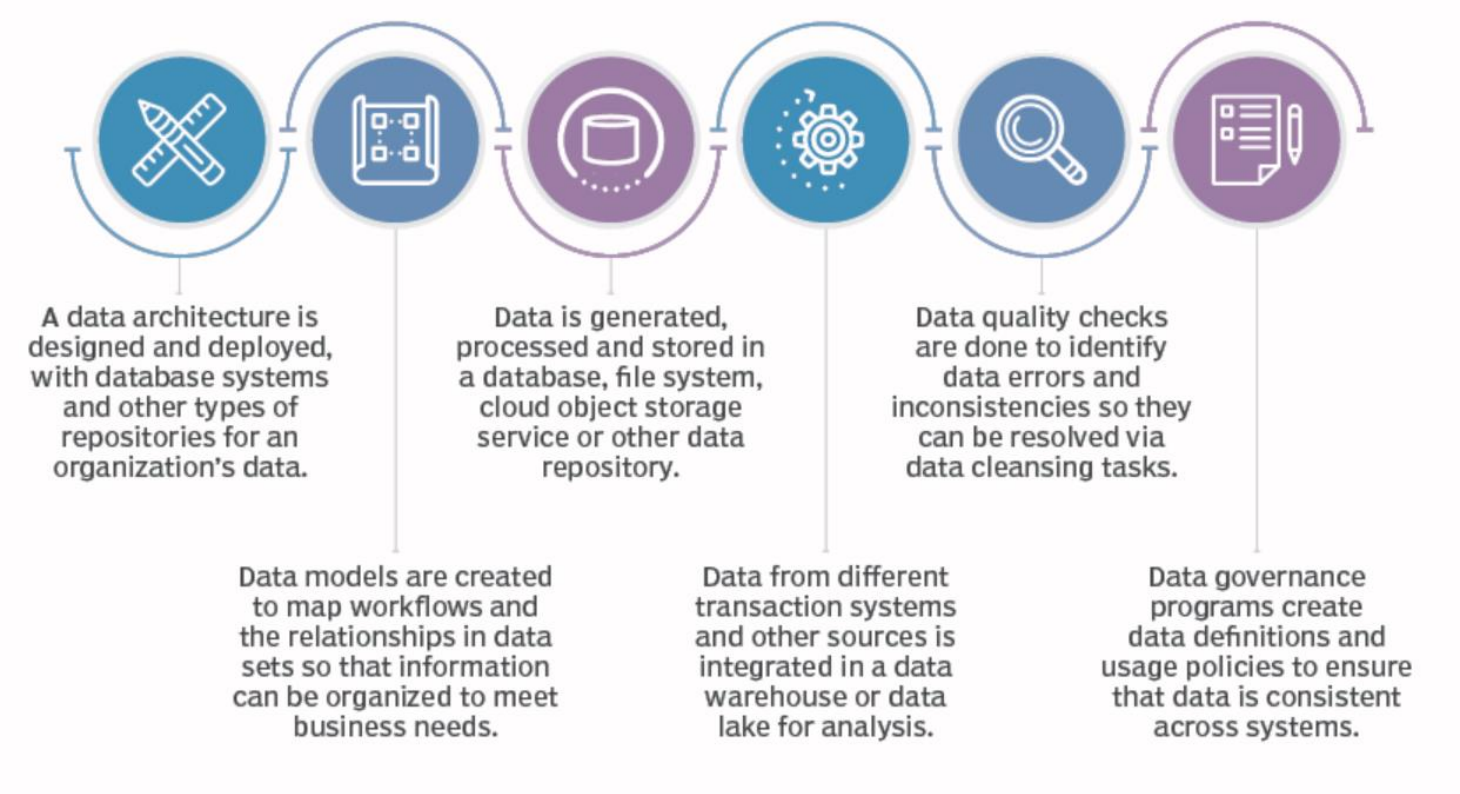
- a **business practice**
- used in **organizing** and maintaining **data processes**
- that meet ongoing **information lifecycle** needs
- within **every company**.

A global need for data management began with the electronics era or digital age of data processing […]

- […] **acquiring, storing, protecting**, and in-depth **processing** required data
- to ensure the required **accessibility**, **reliability**, and **timeliness** of all data for its users

https://www.dataentryoutsourced.com/blog/data-management-best-practices/

What is data management? (2)



Data management is

- the process of **ingesting**, **storing**, **organizing** and **maintaining** the data created and collected by an organization.
- […] deploying the **IT systems** that run business applications and provide **analytical information** to help drive operational **decision-making** and strategic planning by corporate executives, business managers and other end users.
- make sure that the data in corporate systems is **accurate**, **available** and **accessible**.

https://searchdatamanagement.techtarget.com/definition/data-management

Why data management? (1)

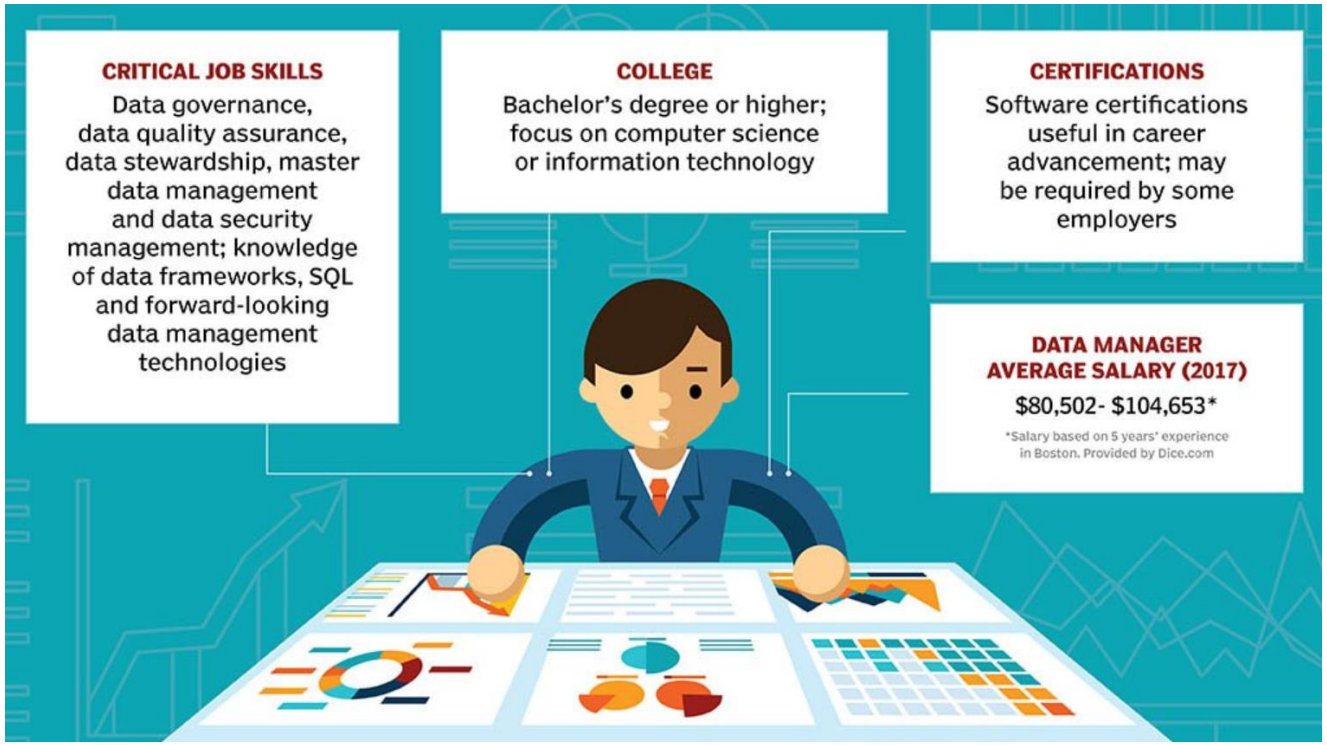
- […] in the digital age, **data is king**.

That is why it is seen as

- **one of the most important assets** of an organization;
- it is the foundation of information and the basis on which people **make decisions**.
- hence it would follow that if the data are **accurate**, **complete**, **organized** and **consistent**,
- it will contribute to the **growth** of the organization.

https://www.dataentryoutsourced.com/blog/data-management-best-practices/

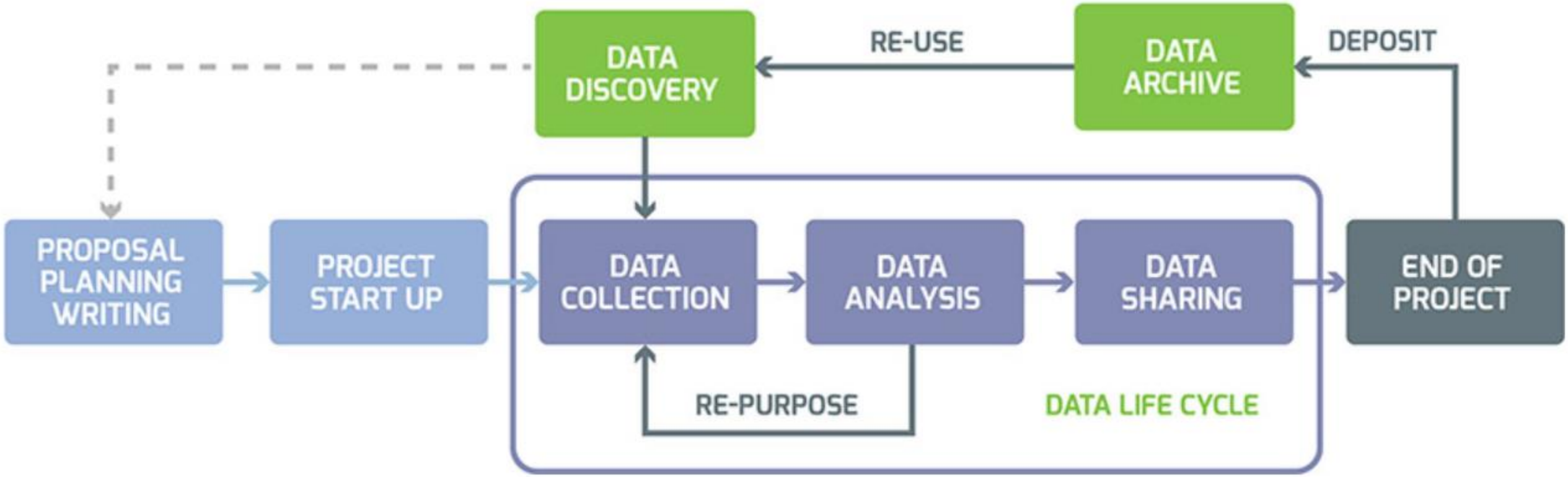
Why data management? (2)



- Data are increasingly seen as a **corporate asset**
- used to make more-informed **business decisions**, […], **optimize** business operations and reduce costs, all with the goal of **increasing revenue and profits**.
- a lack of proper data management can saddle organizations with incompatible **data silos**, **inconsistent** data sets and data **quality** problems […] or, worse, lead to **faulty findings**.
- grown in importance as businesses are subjected to an increasing number of **regulatory compliance** requirements, e.g., data privacy and protection laws (GDPR)

https://searchdatamanagement.techtarget.com/definition/data-management

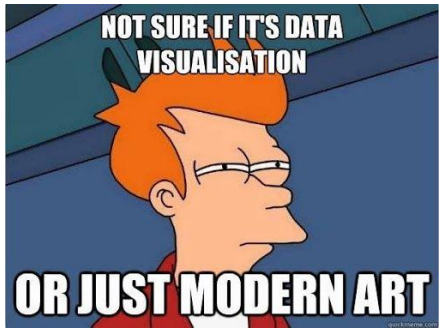
A sample data-management process



https://www.dataentryoutsourced.com/blog/data-management-best-practices/

Data Visualization

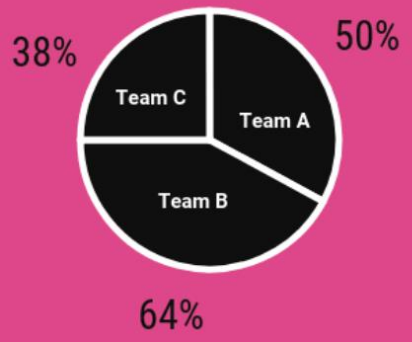
- Data visualization is the **visual presentation of data** or information.
- The goal of data visualization is to **communicate data** or information **clearly** and **effectively** to readers.
- It combines both **art** and **data science**: it should be creative, **pleasing** to look at, and **functional** in its visual communication of the data.

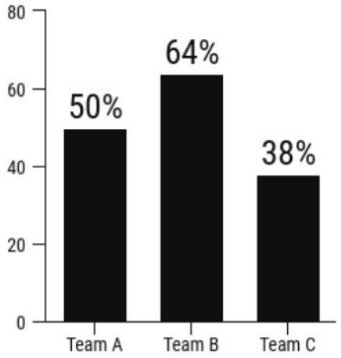


https://venngage.com/blog/data-visualization/

USING THE WRONG GRAPH

The type of graph you use should depend on the type of data you want to visualize. Using the wrong type of graph can skew the data. Writers will sometimes use the wrong type of graph on purpose.





MISLEADING

- Pie charts are used to compare parts of a whole, not the difference between groups
- A different type of graph should be used to compare the three teams

VS

ACCURATE

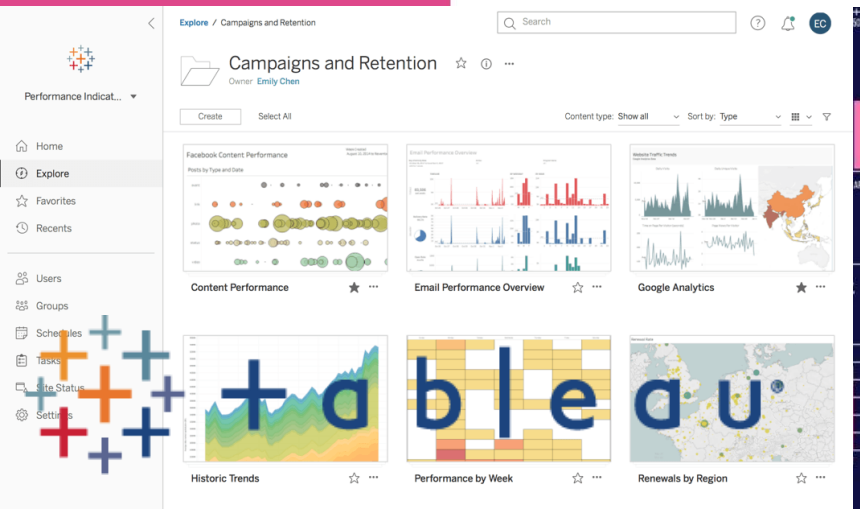
- Bar graphs are better for showing the differences between groups
- This chart is a better visualization of the data

Visuals are processed

60,000 X

faster than text.

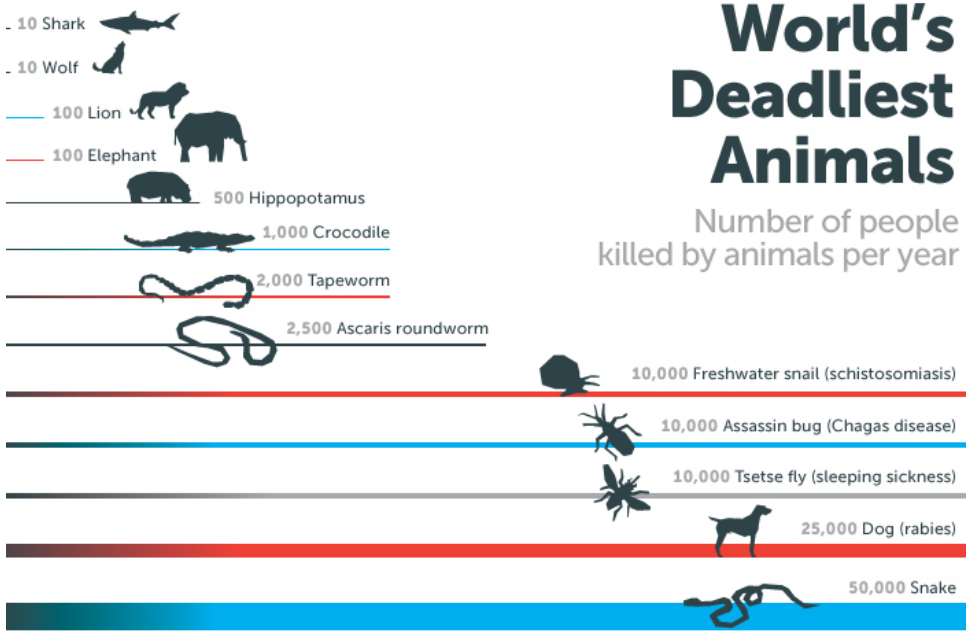
2

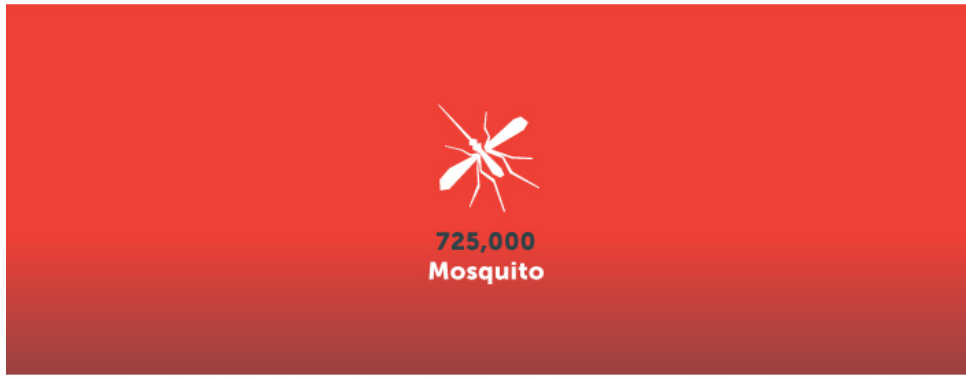


Number of deaths | Killer

World's Deadliest Animals

Number of people killed by animals per year



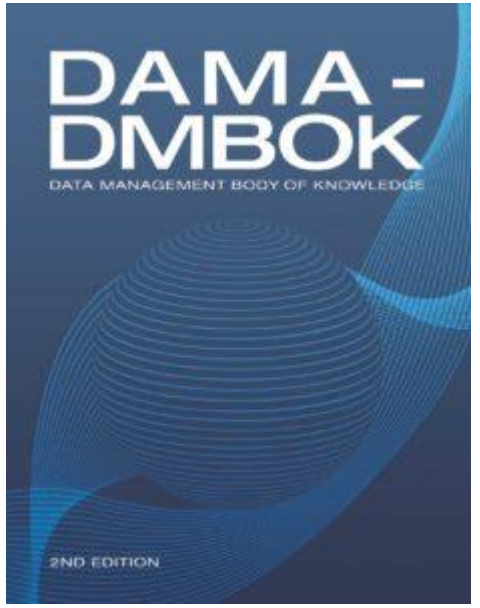


SOURCES: WHO; crocodile-attack.info; Kasturiratne et al. (doi.org/10.1371/journal.pmed.0050218); FAO (webcitation.org/6OgpS8SVO); Linnell et al. (webcitation.org/6ORL7DBUO); Packer et al. (doi.org/10.1038%2F436927a); Alessandro De Maddalena. All calculations have wide error margins.

Diving much deeper…

- **DAMA** International, the Data Governance Professionals Organization work to advance understanding of data management disciplines.
- They published DMBOK in **2009**, a reference book that attempts to define a standard view of data management functions and methods.
- The Data Management Body of Knowledge 2nd Edition, **2017**, (DMBOK2 for short)
- "Provides a functional framework for the implementation of enterprise data management practices; including widely adopted practices, methods and techniques, functions, roles, deliverables and metrics."

https://dama.org/content/body-knowledge



- Data Handling Ethics
- Data Governance
- **Data Architecture**
- **Data Modeling and Design**
- **Data Storage and Operations**
- Data Security
- **Data Integration & Interoperability**
- **Document and Content Management**
- Reference and Master Data
- **Data Warehousing** and Business Intelligence
- Metadata Management
- **Data Quality** Management
- **Big Data** and **Data Science**
- Data Management Maturity Assessment
- Data Management Organization and Role Expectations
- Data Management and Organizational Change Management

Course contents at a glance

Data Management

- OLAP (Online Analytical Processing), multi-dimensional analytical queries
- Data Warehousing
- Data retrieval (querying and indexing)
  - extended SQL
  - specific NoSQL language
- NoSQL data management
- Data modeling
- Distributed data management

Data Visualization

- Motivation and history of Data Visualization
- Visual perception and reasoning
- Graph construction principles
- Data quality

Teaching staff

- **Silvia Chiusano**
  - Relational data management
  - OLAP, Data Warehousing
  - Lectures

- **Diego Monti**
  - Data Visualization
  - Lectures and practices

- **Daniele Apiletti**
  - NoSQL data management
  - Data Warehousing exercises
  - Lectures and… everything (any other business)

- **Eliana Pastor**
  - practice of OLAP, Data Warehousing

- **Alessandro Fiori**
  - practice of NoSQL (MongoDB)

- **Simone Monaco**
  - assistance on Data Warehousing

For private issues, you can contact us via **email** at: name.surname@polito.it

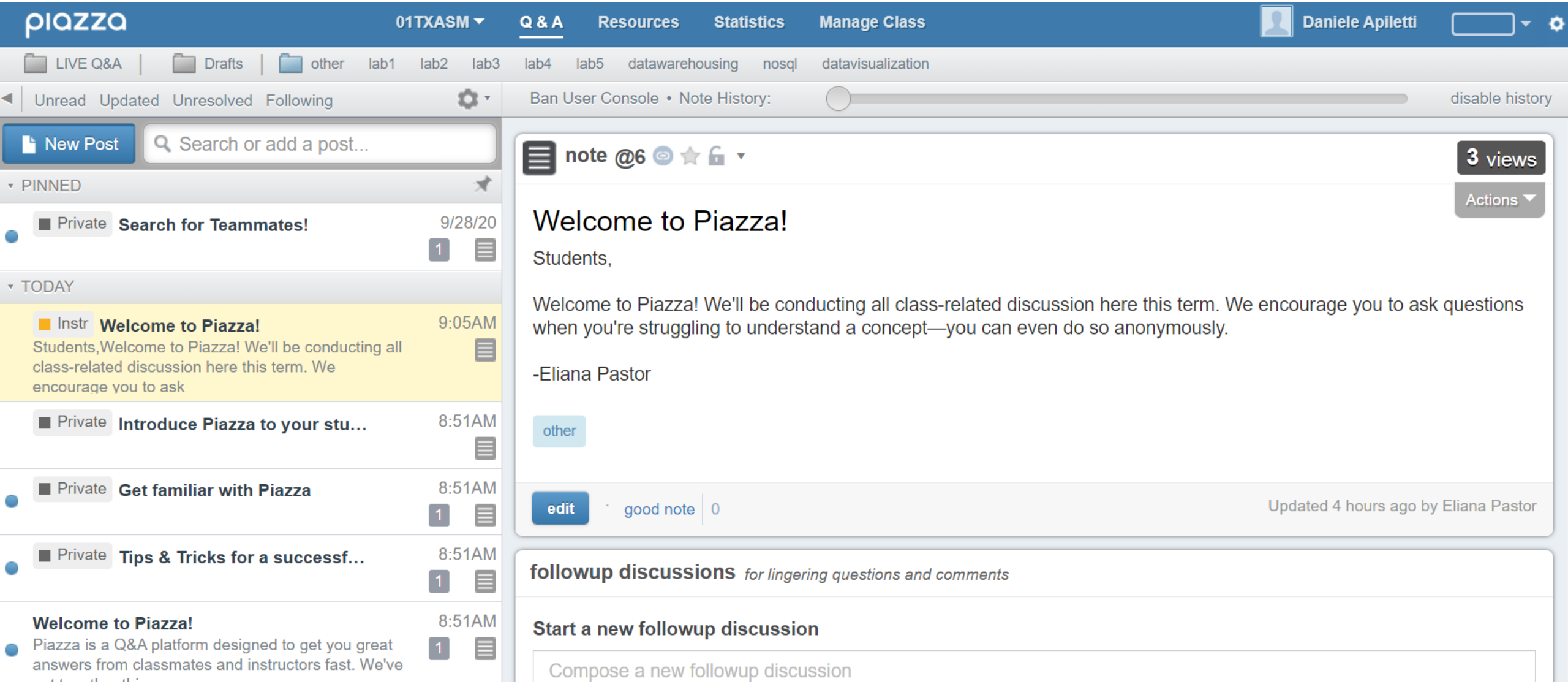
or better ask general course-related questions on **Piazza**:

https://piazza.com/polito.it/fall2021/01txasm

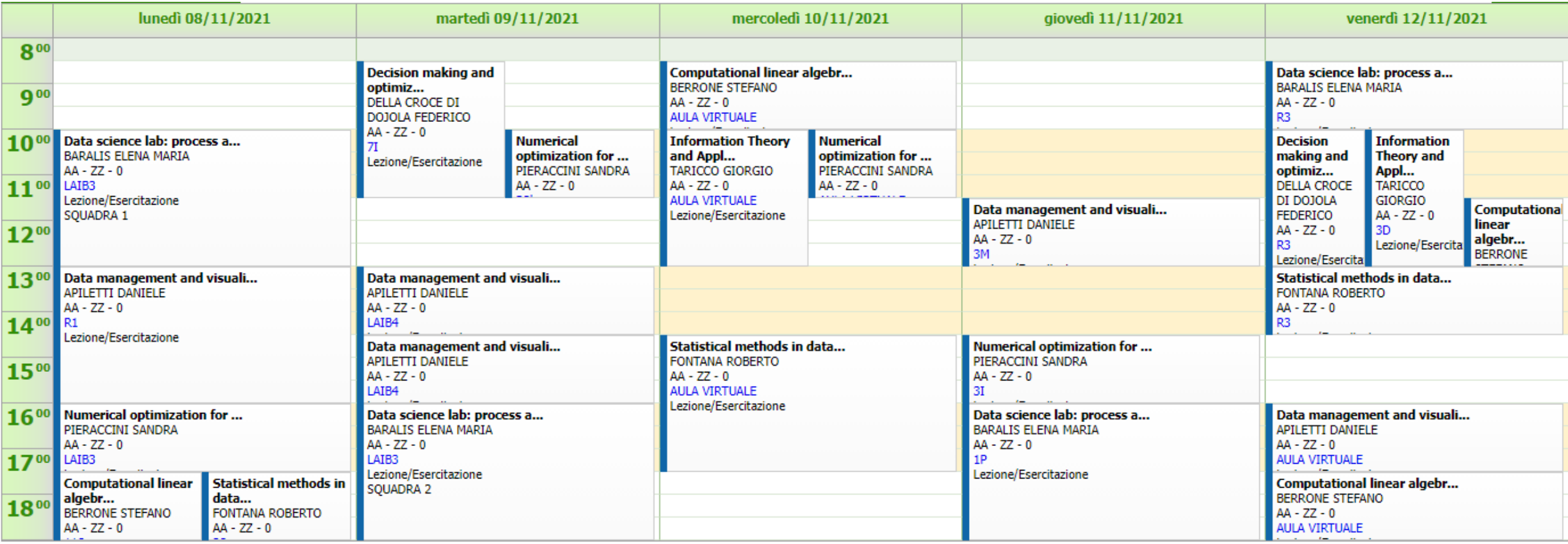
Piazza Q&A

We are using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from both classmates and teachers. Rather than emailing questions to the teaching staff, please post your questions on Piazza, even anonymously.

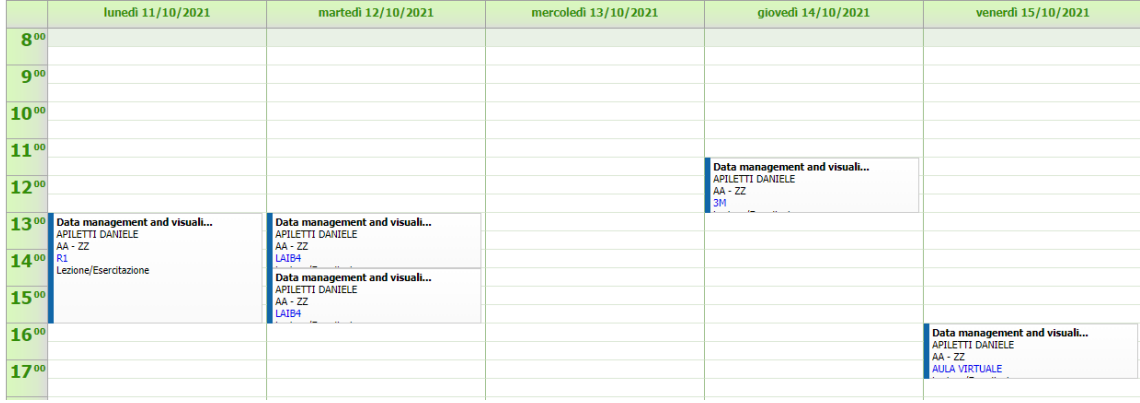
We might use Piazza for **announcements** in case of **failure** of either the Polito teaching portal or the Virtual Classroom services (or both).



Schedule


lunedì 08/11/2021
martedì 09/11/2021
mercoledì 10/11/2021
giovedì 11/11/2021
venerdì 12/11/2021
8 00
9 00
10 00
11 00
12 00
13 00
14 00
15 00
16 00
17 00
18 00
Data science lab: process a...
BARALIS ELENA MARIA
AA - ZZ - 0
LAIB3
Lezione/Esercitazione
SQUADRA 1
Data management and visuali...
APILETTI DANIELE
AA - ZZ - 0
R1
Lezione/Esercitazione
Numerical optimization for ...
PIERACCINI SANDRA
AA - ZZ - 0
LAIB3
Computational linear algebr...
BERRONE STEFANO
AA - ZZ - 0
Statistical methods in data...
FONTANA ROBERTO
AA - ZZ - 0
Decision making and optimiz...
DELLA CROCE DI DOJOLA FEDERICO
AA - ZZ - 0
7I
Lezione/Esercitazione
Numerical optimization for ...
PIERACCINI SANDRA
AA - ZZ - 0
Data management and visuali...
APILETTI DANIELE
AA - ZZ - 0
LAIB4
Data management and visuali...
APILETTI DANIELE
AA - ZZ - 0
LAIB4
Data science lab: process a...
BARALIS ELENA MARIA
AA - ZZ - 0
LAIB3
Lezione/Esercitazione
SQUADRA 2
Computational linear algebr...
BERRONE STEFANO
AA - ZZ - 0
AULA VIRTUALE
Information Theory and Appl...
TARICCO GIORGIO
AA - ZZ - 0
AULA VIRTUALE
Lezione/Esercitazione
Numerical optimization for ...
PIERACCINI SANDRA
AA - ZZ - 0
Statistical methods in data...
FONTANA ROBERTO
AA - ZZ - 0
AULA VIRTUALE
Lezione/Esercitazione
Data management and visuali...
APILETTI DANIELE
AA - ZZ - 0
3M
Numerical optimization for ...
PIERACCINI SANDRA
AA - ZZ - 0
3I
Data science lab: process a...
BARALIS ELENA MARIA
AA - ZZ - 0
1P
Lezione/Esercitazione
Data science lab: process a...
BARALIS ELENA MARIA
AA - ZZ - 0
R3
Decision making and optimiz...
DELLA CROCE DI DOJOLA FEDERICO
AA - ZZ - 0
R3
Lezione/Esercita
Information Theory and Appl...
TARICCO GIORGIO
AA - ZZ - 0
3D
Lezione/Esercita
Computational linear algebr...
BERRONE
Statistical methods in data...
FONTANA ROBERTO
AA - ZZ - 0
R3
Data management and visuali...
APILETTI DANIELE
AA - ZZ - 0
AULA VIRTUALE
Computational linear algebr...
BERRONE STEFANO
AA - ZZ - 0
AULA VIRTUALE

Schedule



Lectures

- **Monday**
  13:00-14:30
  14:30-16:00
  classroom R1 + online
- **Thursday**
  11:30-13:00
  classroom 3M + online
- **Friday**
  16:00-17:30
  online only

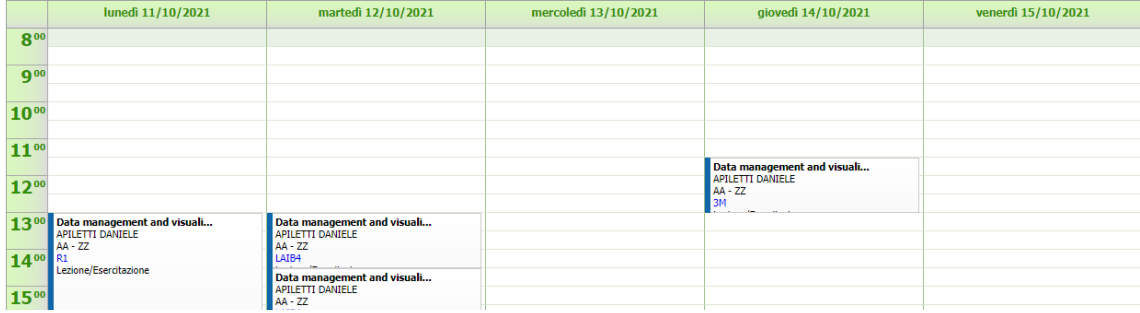
We will host some seminars toward the end of the course.

Schedule changes announced on the teaching portal.

Lab practice

- **Tuesday**
  13:00-14:30 team A (Lab practice)
  14:30-16:00 team B (Lab practice)
  Laboratory LAIB4 + online
- Starting on
  Tuesday, **October 19**, 2021
- Till the end of the course
- Each student will be assigned to a single Team (**either A or B**)
  - based on their surname
  - changing Team is allowed upon request

Schedule



Go to

www.menti.com

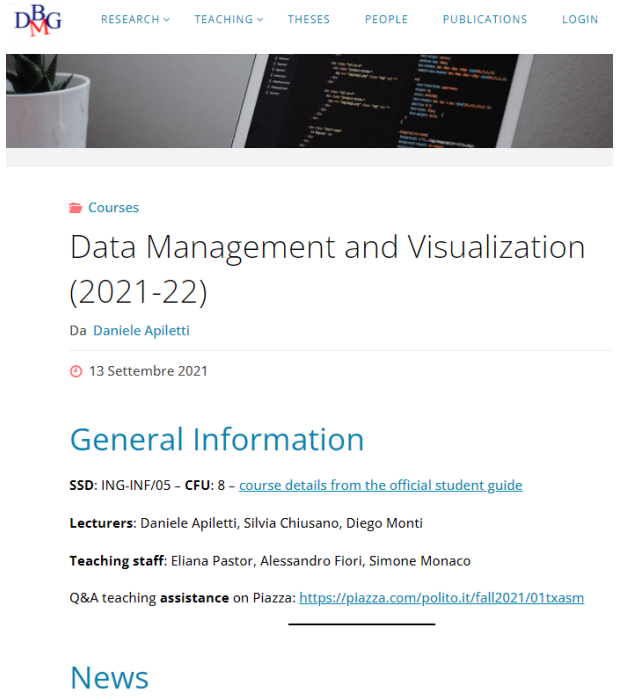
Enter the code

5670 3828

Material

- Announcements on the official «teaching portal» private page https://didattica.polito.it/

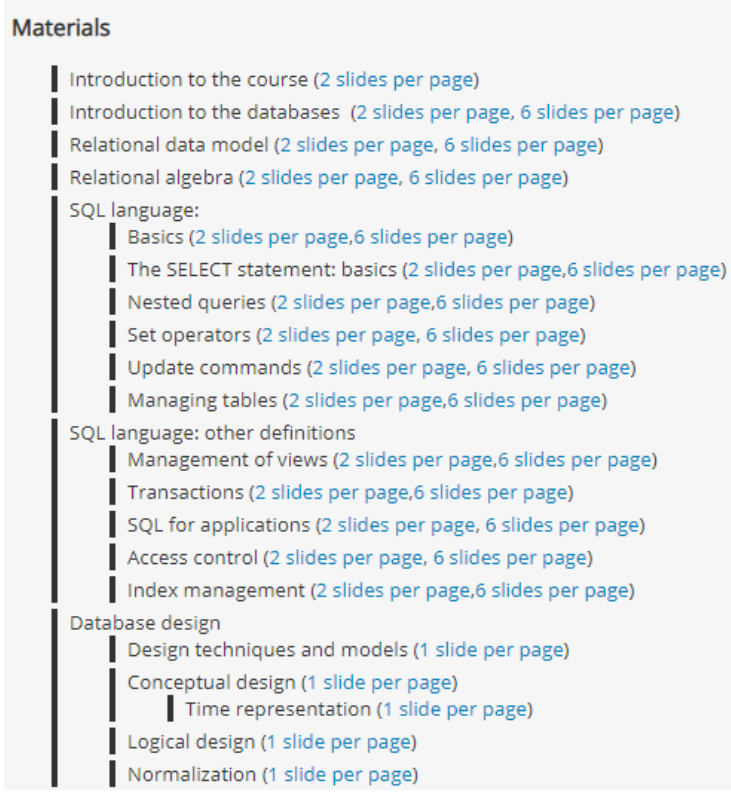
- Slides, texts of the practices, and all other materials are available on the **public page of the course** https://dbdmg.polito.it/



Pre-requisites: **relational model** + **SQL**

- «Introduction to databases» **videolectures** on the **portal**: https://didattica.polito.it/

- «Introduction to databases» **slides** on the public web **page**: https://dbdmg.polito.it/wordpress/teaching/databases/



Pre-requisites: **relational model** + **SQL**

- «Introduction to databases» **videolectures** on the **portal**: https://didattica.polito.it/

Go to **www.menti.com** and use the code **2651 7433**

Please join the poll here: https://www.menti.com/22zeyu73xj

- «Introduction to databases» **slides** on the public web **page**: https://dbdmg.polito.it/wordpress/teaching/databases/

Materials

- Introduction to the course (2 slides per page)
- Introduction to the databases (2 slides per page, 6 slides per page)
- Relational data model (2 slides per page, 6 slides per page)
- Relational algebra (2 slides per page, 6 slides per page)
- SQL language:
  - Basics (2 slides per page,6 slides per page)
  - The SELECT statement: basics (2 slides per page,6 slides per page)
  - Nested queries (2 slides per page,6 slides per page)
  - Set operators (2 slides per page, 6 slides per page)
  - Update commands (2 slides per page, 6 slides per page)
  - Managing tables (2 slides per page,6 slides per page)
- SQL language: other definitions
  - Management of views (2 slides per page,6 slides per page)
  - Transactions (2 slides per page,6 slides per page)
  - SQL for applications (2 slides per page, 6 slides per page)
  - Access control (2 slides per page, 6 slides per page)
  - Index management (2 slides per page,6 slides per page)
- Database design
  - Design techniques and models (1 slide per page)
  - Conceptual design (1 slide per page)
    - Time representation (1 slide per page)
  - Logical design (1 slide per page)
  - Normalization (1 slide per page)

Exam

- See exam policy on the official course web page on the teaching portal

- For students of the last academic year, mind the changelog
  - due to the Covid experience and the in-presence exam
  - Some improved topics:
    + NoSQL queries
    + extended NoSQL design patterns

Assessment and grading criteria for ONSITE exam	***Data management and visualization*** **Exam:** Written test;

Data management and visualization

The exam lasts 90 minutes and consists of theoretical questions and written exercises, as described in the following:

- [max 5 points] 3-6 multiple-choice questions on theoretical topics of the course, such as conceptual, logical, and physical data warehouse design, extended SQL language, technological characteristics of NoSQL databases and their usage, data management issues in distributed (non-relational) databases, data visualization techniques

- [max 12 points] exercises on data warehousing, including 2-4 open and/or multiple-choice questions on data warehouse design, and 2-3 queries for data access in extended SQL (open questions with answers to be provided in a text box)

- [max 9 points] 1-2 exercises on NoSQL database design and 1-2 queries for data access (open questions with answers to be provided in a text box)

- [max 5 points] 1 exercise on visualization analysis and design with open questions (answers to be provided in a text box)

Students are not allowed to use textbooks, notes, or electronic devices during the exam.

Exercises are evaluated according to the correctness of the proposed solution and to the appropriateness of the adopted resolution methodologies.
Specific points for each exercise are indicated in the exam text.
Multiple-choice questions have a penalty for wrong answers, whereas no-penalty no-points in case no answer is provided.

Learning objectives assessment.
The exam will assess:
- the knowledge of data warehouse architectures and of their design methodologies (conceptual, logical, and physical)
- the ability to design a data warehouse in a provided use case
- the ability to write extended SQL queries to extract data of interest from a data warehouse
- the knowledge of the main technological characteristics of NoSQL databases
- the ability to design NoSQL databases and to query NoSQL databases
- the ability to design dashboards and KPIs
- the knowledge of the basic principles of cognitive and perceptive aspects related to visualization, and of the main visualization techniques

Questions?

Data Management and Visualization

INTRODUCTION TO THE COURSE

Academic Year 2021-2022

Daniele Apiletti