Data Management and Visualization

INTRODUCTION TO THE COURSE

Academic Year 2020-2021

Before starting the COVID-compliant lecture

- Video: Politecnico youtube channel – Formazione/Training courses section https://www.youtube.com/c/PolitecnicodiTorinoChannel/videos
- only those people who have made regular **reservations** to attend lectures according to the procedure in force can be allowed to stay in the classroom.
- those people who are in the classroom but have not been authorized to may be inhibited from reserving and attending classes in the following weeks.
- it is possible to occupy only the **seats** indicated with appropriate signs, the other seats must be left free.
- it is obligatory to always maintain a minimum **safety distance of 2 m** from other people.
- it is mandatory to **always wear the mask** during your stay in the classroom and in general to circulate within the premises and common areas, while it is not mandatory to use it during breaks in the open spaces/areas of our University.
- it is necessary to frequently **sanitize your hands** with soap and water or with a suitable hydroalcoholic solution available in the classroom.
- if you need to **sneeze or cough**, you should do so in a handkerchief or in the fold of your elbow. The handkerchief will then immediately be placed in the basket of the undifferentiated collection in the classroom.
- the **teacher can remove your mask** during the lecture, by keeping a safe distance from all students

Data Management and Visualization

INTRODUCTION TO THE COURSE

Academic Year 2020-2021

Daniele Apiletti

What is data management? (1)

Data management is

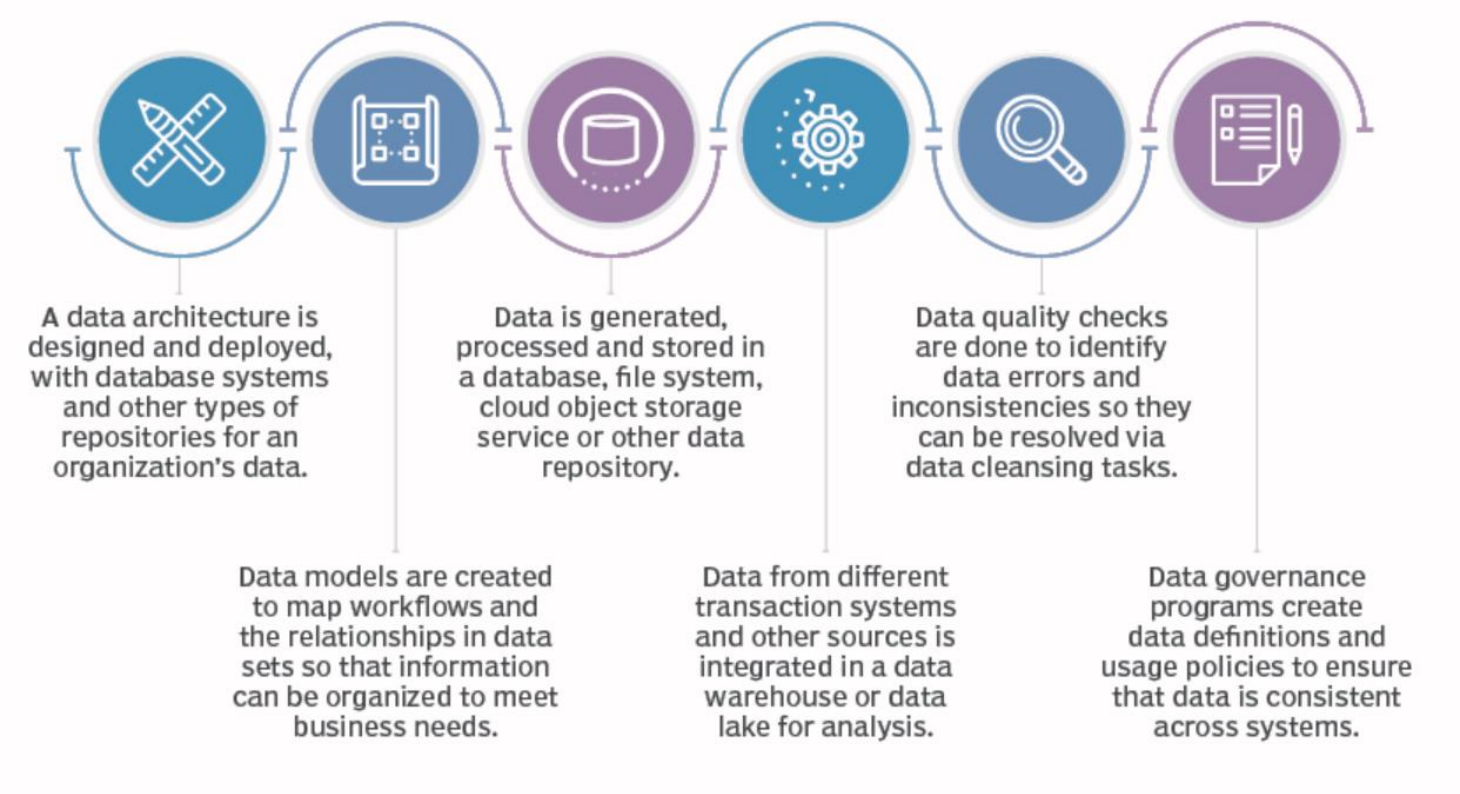
- a **business practice**
- used in **organizing** and maintaining **data processes**
- that meet ongoing **information lifecycle** needs
- within **every company**.

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)

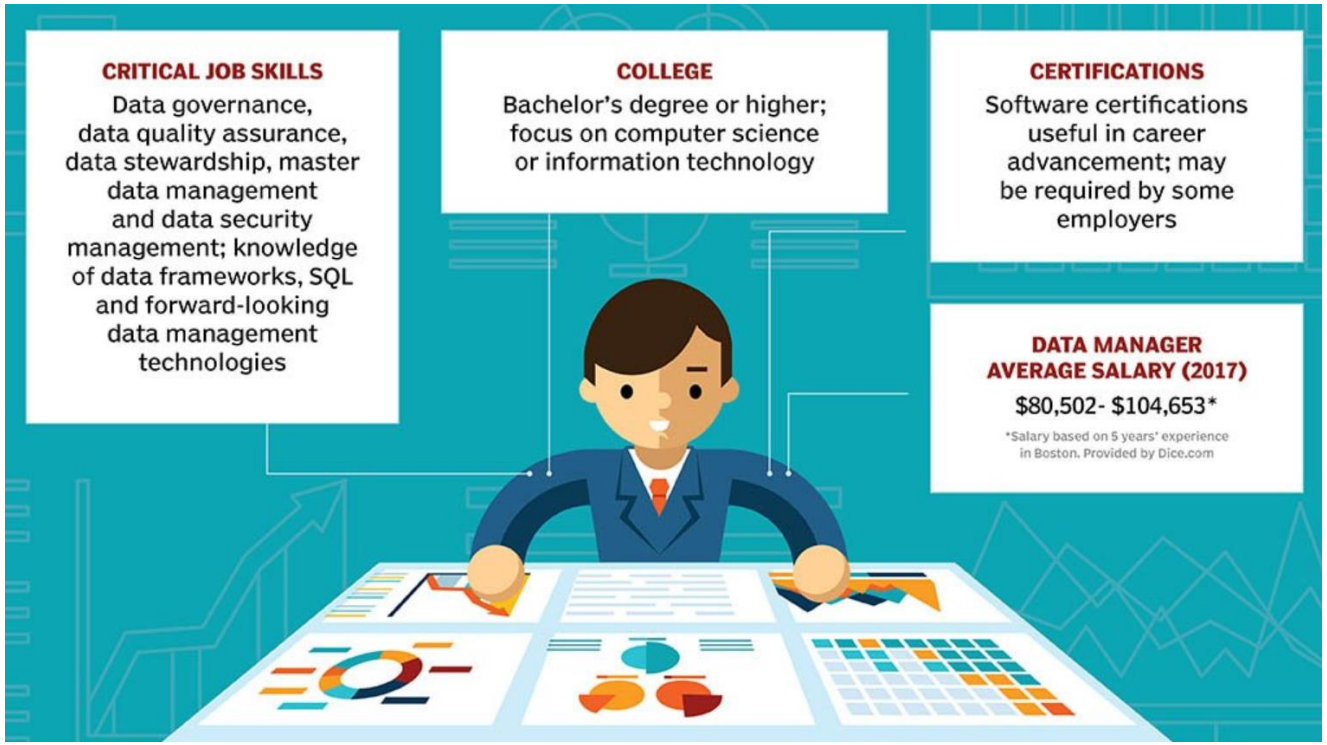
- […] in the digital age, **data is king**.

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/

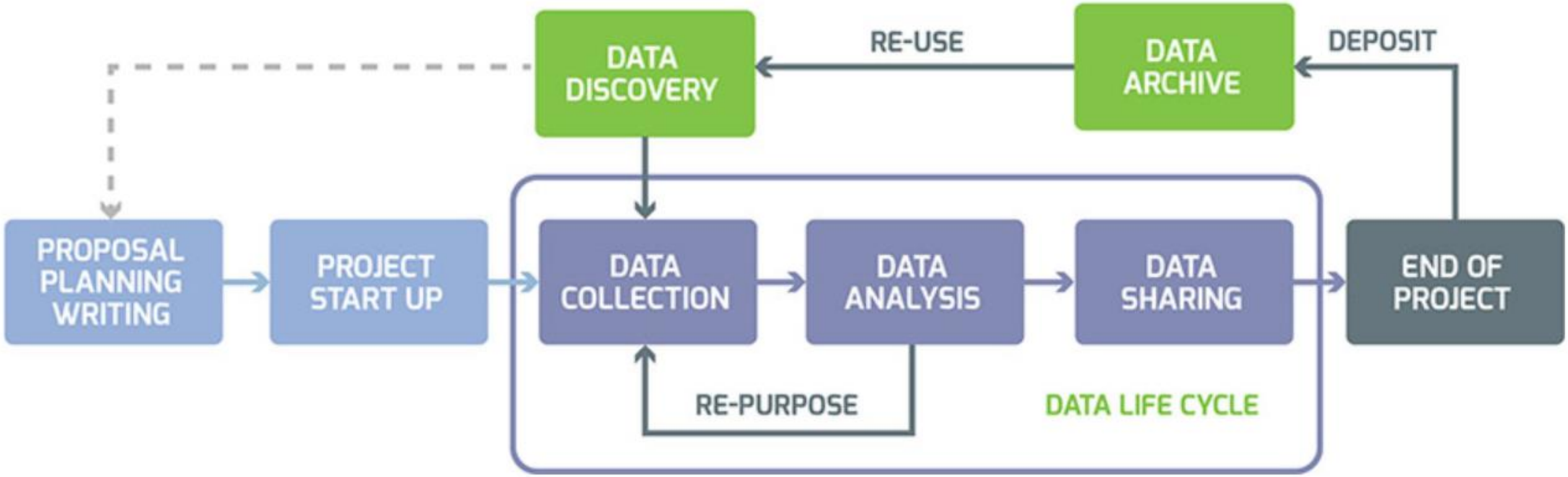
Why data management? (2)



- 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

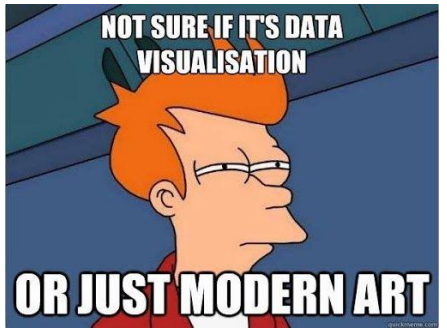
A sample data-management process



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

Data Visualization

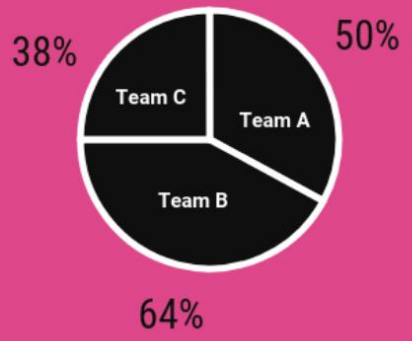
- Data visualization is the **visual presentation of data** or information.
- The goal of data visualization is to **communicate data** or information **clearly** and **effectively** to readers.
- It combines both **art** and **data science**: it should be creative, **pleasing** to look at, and **functional**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

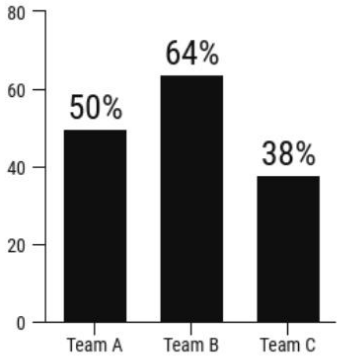


https://venngage.com/blog/data-visualization/

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.





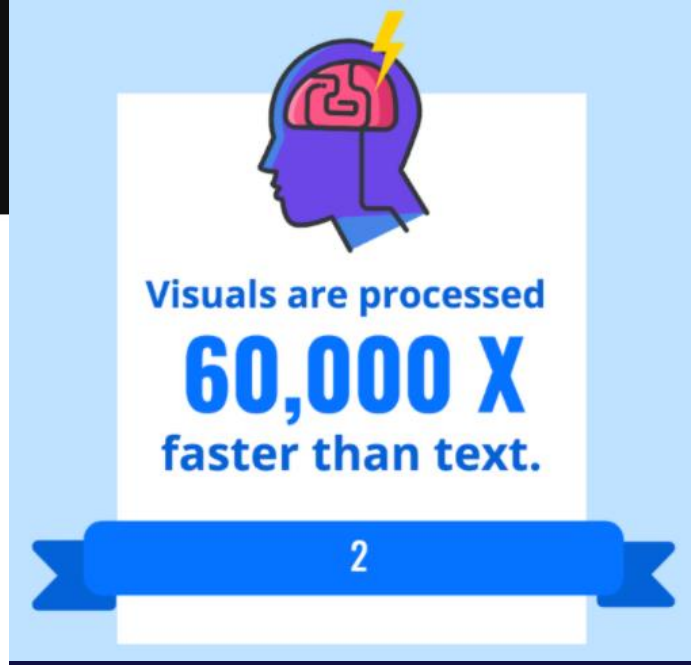
MISLEADING

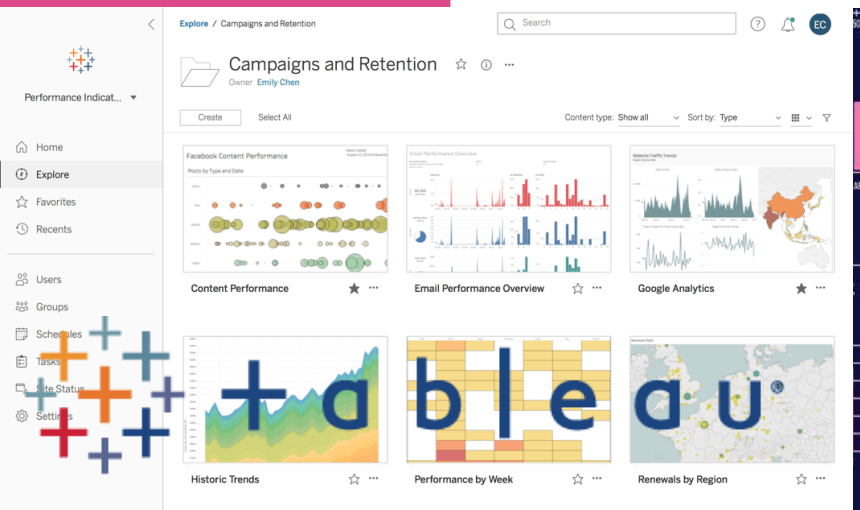
- Pie charts are used to compare parts of a whole, not the difference between groups
- A different type of graph should be used to compare the three teams

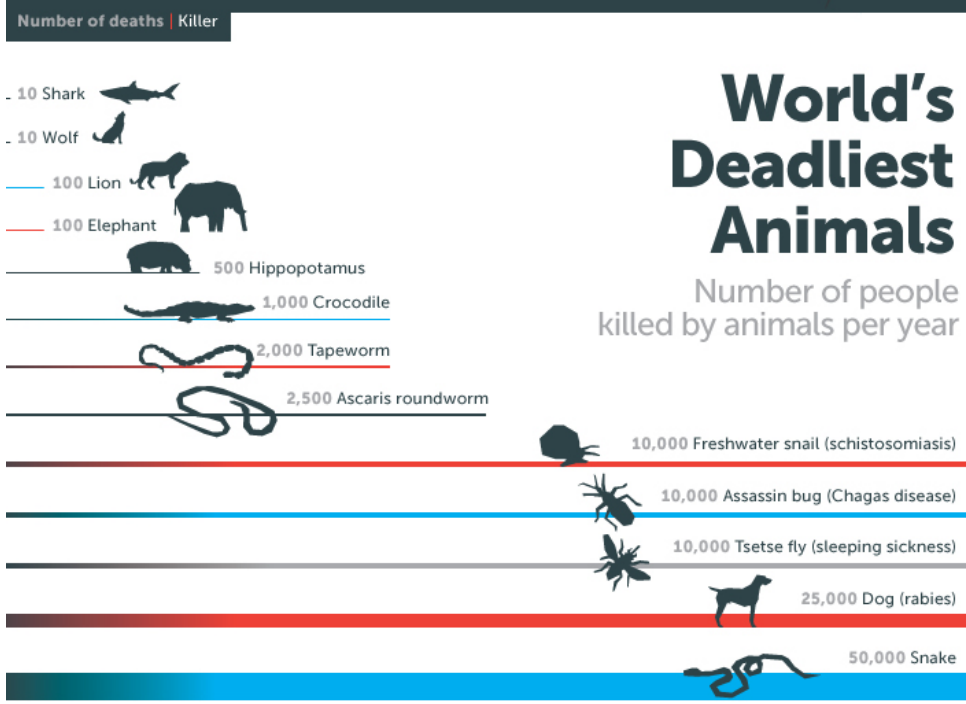
VS

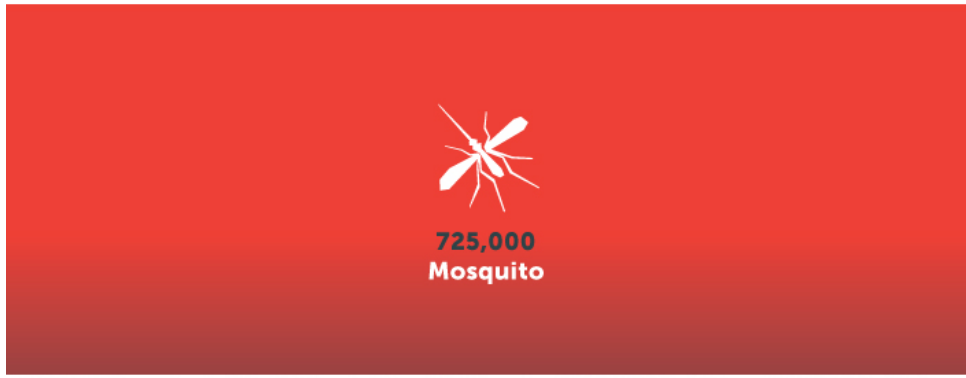
ACCURATE

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









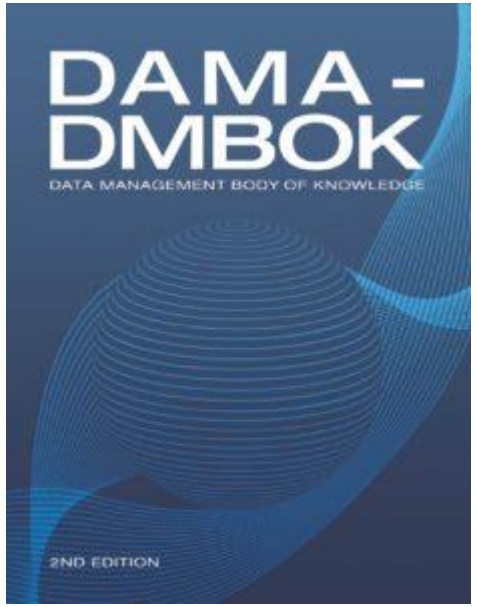
SOURCES: WHO; crocodile-attack.info; Kasturiratne et al. (doi.org/10.1371/journal.pmed.0050218); FAO (webcitation.org/6OgpS8SVO); Linnell et al. (webcitation.org/6ORL7DBUO); Packer et al. (doi.org/10.1038%2F436927a); Alessandro De Maddalena. All calculations have wide error margins.

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.”

DAMA International® The Global Data Management Community

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
- Extended SQL
- NoSQL data management
- Data modeling
- Data retrieval (querying and indexing)
- Distributed data management

Data Visualization

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

Teachers

- **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… miscellaneous (any other business)

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

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

- **Isabeau Oliveri**
  - practice of Data Visualization

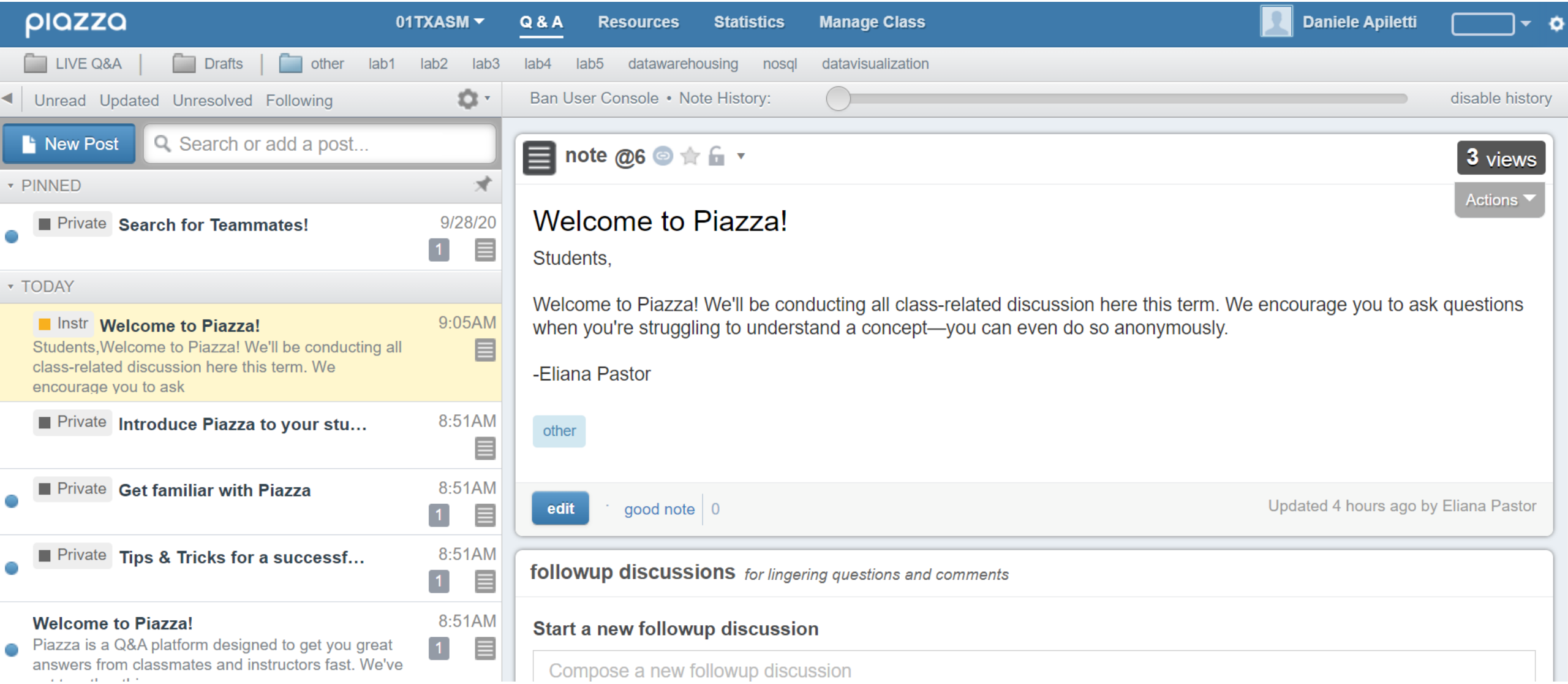
For private issues, you can contact us via **email** at: name.surname@polito.it

or better ask general questions on **Piazza**:

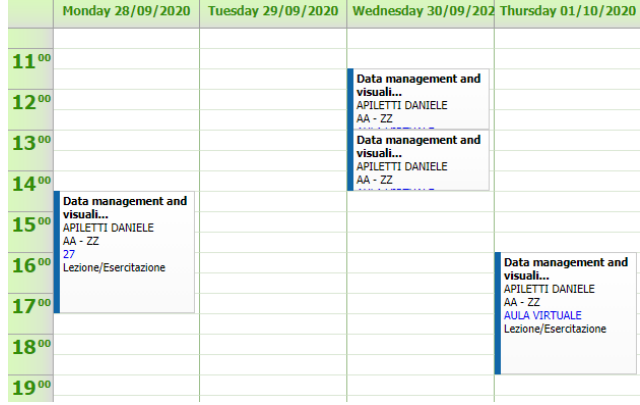
https://piazza.com/polito.it/fall2020/01txasm/home

Q&A on Piazza

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



Schedule



Lectures

- **Monday**
  14:30-17:30
  classroom 27 + Virtual Classroom
- **Wednesday** *
  11:30-13:00
  Virtual Classroom only
- **Thursday** *
  17:30-19:00
  Virtual Classroom only

During the first 2-3 weeks, the schedule might slightly change since we possibly use the «practice» slots on Wednesday and Thursday to recover lectures (e.g., due to technical issues of the Virtual Classroom system…); **we will keep you updated on the actual schedule on a weekly basis**; we expect such changes to have a low impact, since those lectures are virtual only and recorded.

Practices

- Starting on
  Wednesday, **October 21**, 2020
- Till the end of the course
- Weekly schedule
  - Team **A** (Virt. Classroom and Labinf)
    Wednesday from 13:00 to 14:30
  - Team **B** (Virtual Classroom only)
    Thursday from 16:00 to 17:30
- Each student will be assigned to a single Team (**either A or B**)
  - Moving to another Team is allowed upon request via email to Eliana
- Labinf: waiting for access policy…

Material

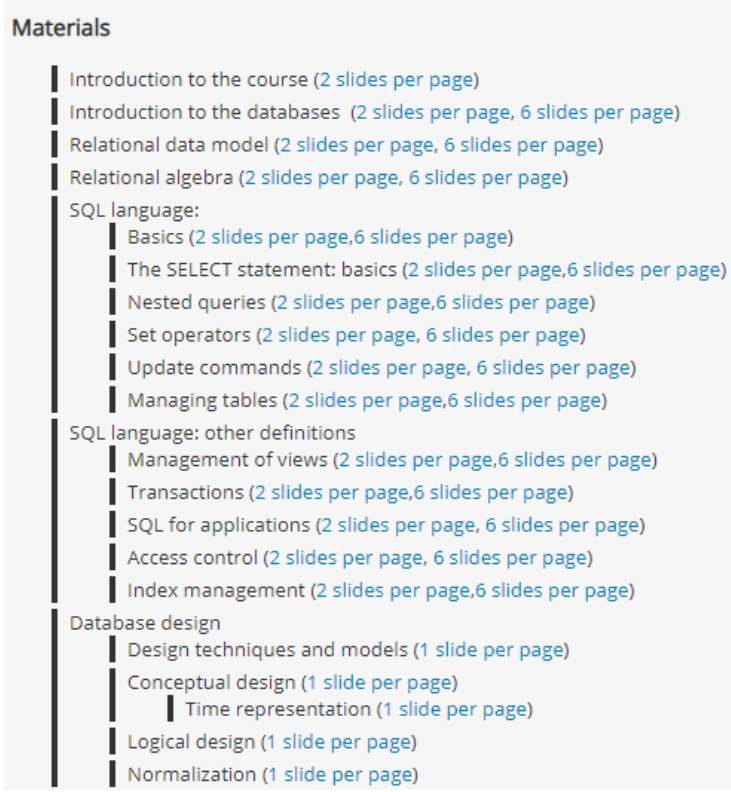
- Announcements on the official «teaching portal» private page https://didattica.polito.it/

- 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/



Exam

- See exam policy on the course public web page https://dbdmg.polito.it/

- For students of the last academic year, mind the changelog
  - due to the Covid experience and the online-only exam
  - New topics:
    + theory
    + NoSQL queries
    + extended NoSQL design patterns

POLITECNICO DI TORINO

Data Management and Visualization

Exam policy (A.Y. 2020-2021)

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

- **Theory [5 points]**
  - at least 3 multiple-choice questions on theoretical topics of the course, including all the topics presented by the teachers during the lectures and the related material (slides), such as the following:
    - 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

- **DW [12 points]**, exercises on data warehousing, including:
  - at least 2 multiple-choice questions and at least 1 open-text question on data warehouse design (respectively on conceptual schema and logical design)
  - at least 2 queries for data access in extended SQL (open text-box questions)

- **NoSQL [9 points]**
  - at least 1 exercise on NoSQL database design
  - at least 1 query for data access in MongoDB
  - both are open questions with answers to be provided in a text box

- **Data Visualization [5 points]**
  - 1 exercise on visualization analysis and design with many different open questions
  - answers to be provided in a text box

Questions?