Data Management and Visualization

INTRODUCTION TO THE COURSE

Academic Year 2022-2023

Daniele Apiletti

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)

- **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/fall2022/01txasm

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).

---

piazza   01TXASM 5 ▾   **Q & A**   **Resources**   **Statistics**   **Manage Class**   **Daniele Apiletti** ▾ ⚙

📁 LIVE Q&A   📁 Drafts   📁 exam   datawarehousing   nosql   dataviz   lab1   lab2   lab3   lab4   lab5   lab6   lab7   lab8   lab9   lab10   lab11   other   oraclexesql   challenge

◀ | Unread   Updated   Unresolved   Following   i ▾ ⚙▾   Ban User Console · Note History: No history yet   disable history

📝 New Post   🔍 Search or add a post...

Show Actions

▾ PINNED 📌

■ Private **Search for Teammates!**   9/26/22   1

▾ YESTERDAY

■ Private **Introduce Piazza to your stu...**   9/26/22

■ Private **Get familiar with Piazza**   9/26/22   1

■ Private **Tips & Tricks for a successf...**   9/26/22   1

Welcome to Piazza!   9/26/22
Piazza is a Q&A platform designed to get you great answers from classmates and instructors fast. We've put together thi   1

📋 **private note @4** ⊕ ★ 🔒 ▾   2 views   Actions ▾

Introduce Piazza to your students

Post a Welcome Note!

In your first post on Piazza, welcome your students to their new class:

```
Students,

Welcome to Piazza! We'll be conducting all class-related discussion here this term. The quicker
you begin asking questions on Piazza (rather than via emails), the quicker you'll benefit from
the collective knowledge of your classmates and instructors. We encourage you to ask questions
when you're struggling to understand a concept—you can even do so anonymously.

-Daniele Apiletti
```

Add Post

Include this blurb in your syllabus

This term we will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from classmates, the TA, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. If you have any problems or feedback for the developers, email team@piazza.com.

Find our class page at: https://piazza.com/polito.it/fall2022/01txasm/home

Material

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

- 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/

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)

Pre-requisites: relational model + SQL

Please join the poll at

www.menti.com

Enter code **1106 1199**

What is data management? (1)
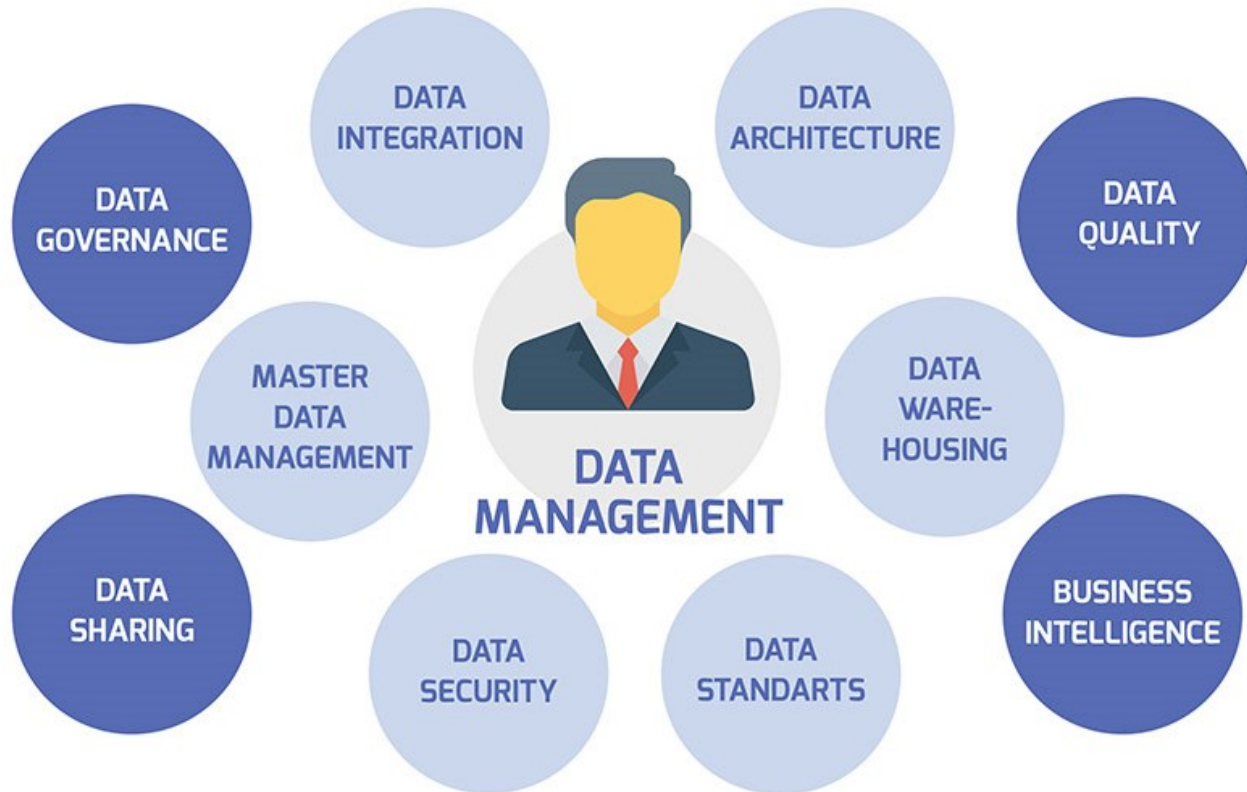
Data management is

- 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

What is data management? (2)

A data architecture is designed and deployed, with database systems and other types of repositories for an organization's data.

Data is generated, processed and stored in a database, file system, cloud object storage service or other data repository.

Data quality checks are done to identify data errors and inconsistencies so they can be resolved via data cleansing tasks.

Data models are created to map workflows and the relationships in data sets so that information can be organized to meet business needs.

Data from different transaction systems and other sources is integrated in a data warehouse or data lake for analysis.

Data governance programs create data definitions and usage policies to ensure that data is consistent across systems.

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**.

Why data management? (1)
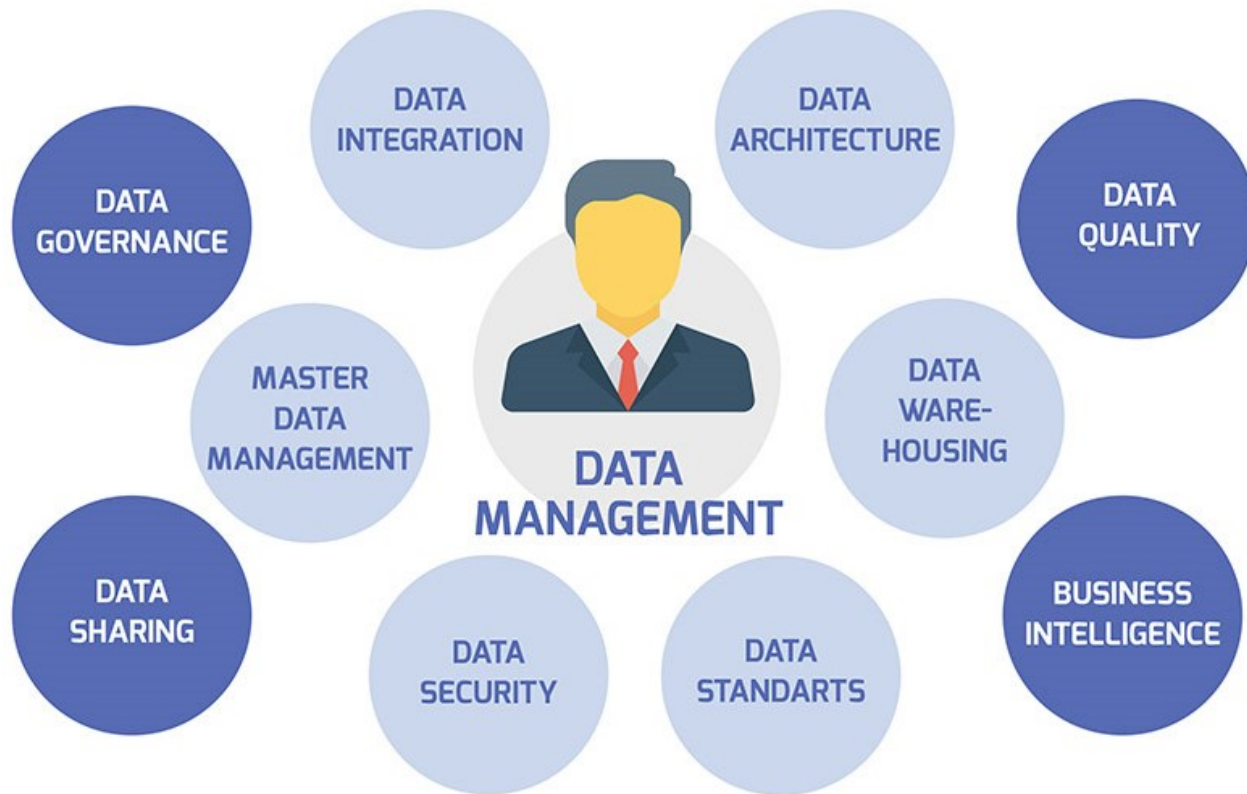
- [...] in the digital age, **data is king**.

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.

Why data management? (2)

CRITICAL JOB SKILLS
Data governance, data quality assurance, data stewardship, master data management and data security management; knowledge of data frameworks, SQL and forward-looking data management technologies

COLLEGE
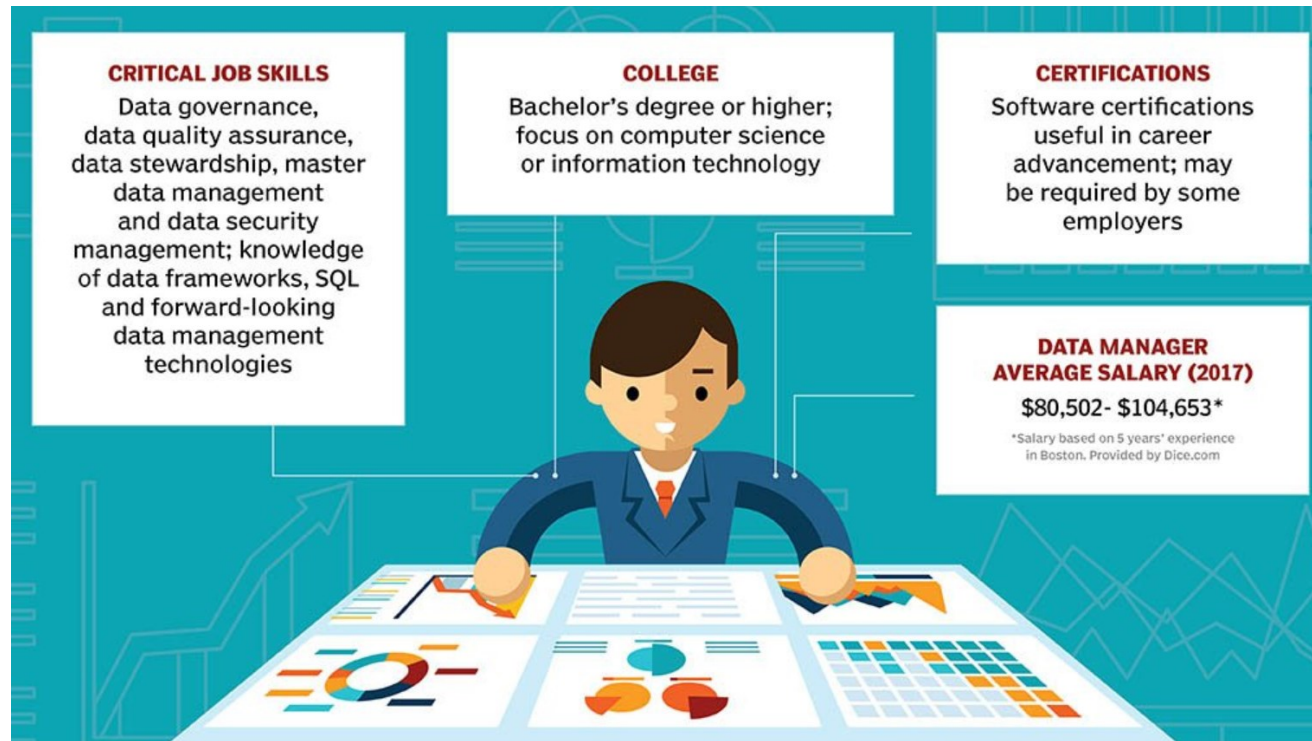Bachelor's degree or higher; focus on computer science or information technology

CERTIFICATIONS
Software certifications useful in career advancement; may be required by some employers

DATA MANAGER AVERAGE SALARY (2017)
$80,502- $104,653*
*Salary based on 5 years' experience in Boston. Provided by Dice.com

- 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)

# A sample data-management process
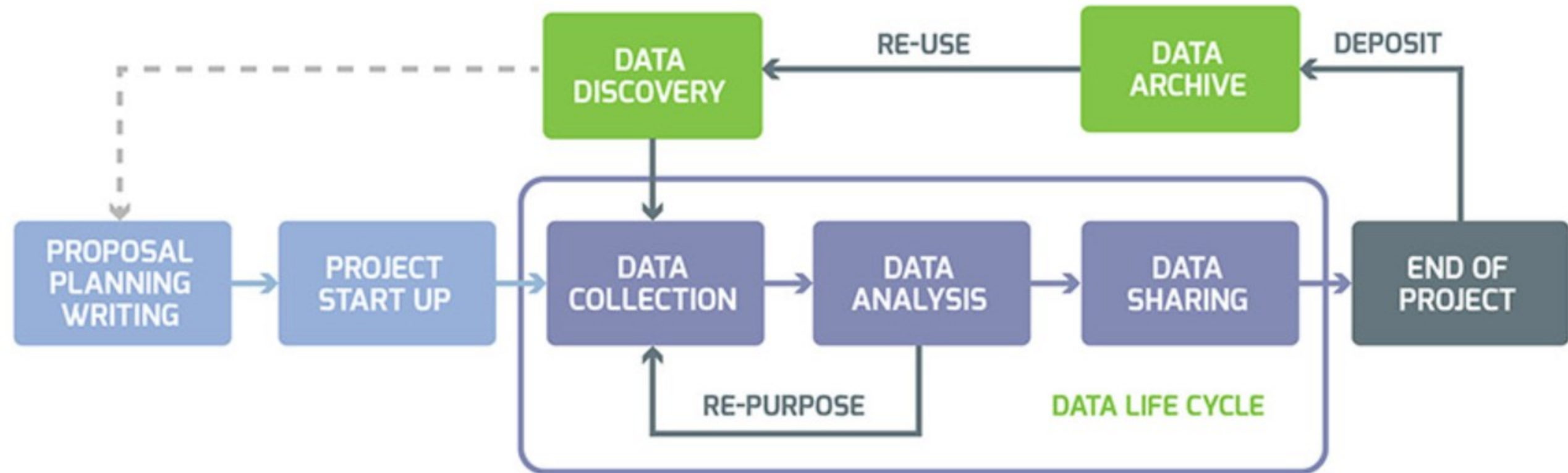
Data Visualization

- 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.
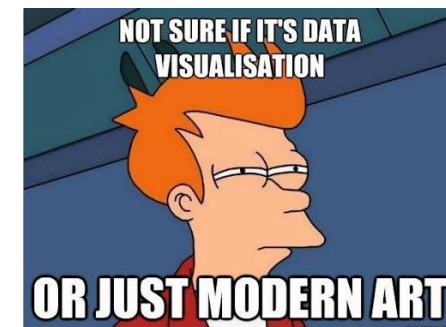
USING THE WRONG GRAPH

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.

38%
50%
Team C
Team A
Team B
64%

80
64%
60
50%
40
38%
20
0
Team A   Team B   Team C

☹ MISLEADING   VS   ACCURATE ☺

- 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

- 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
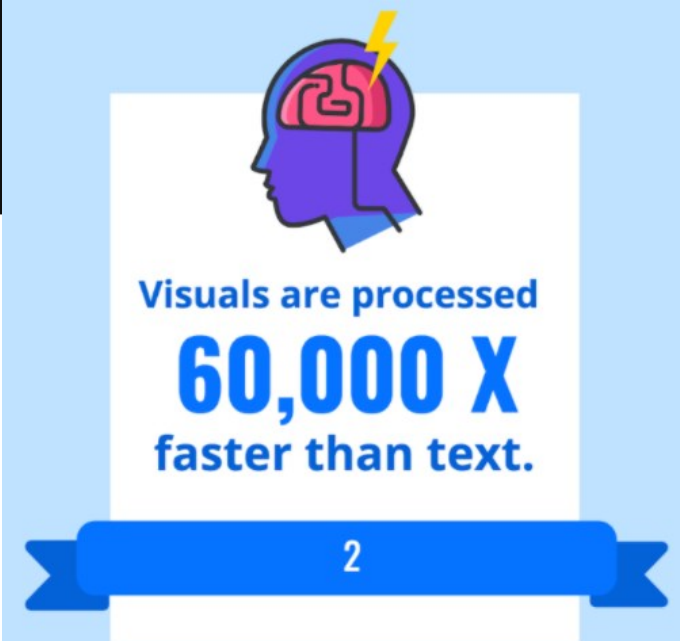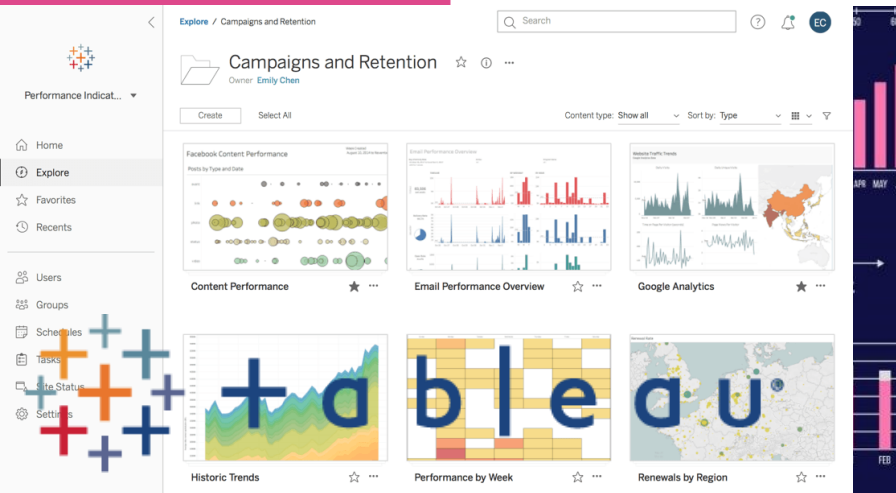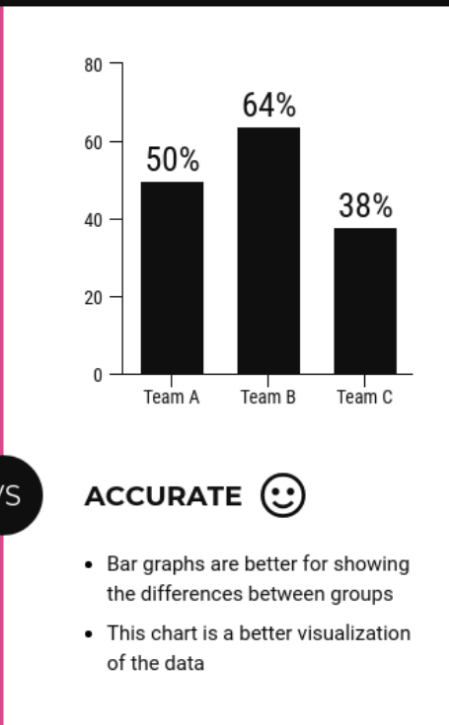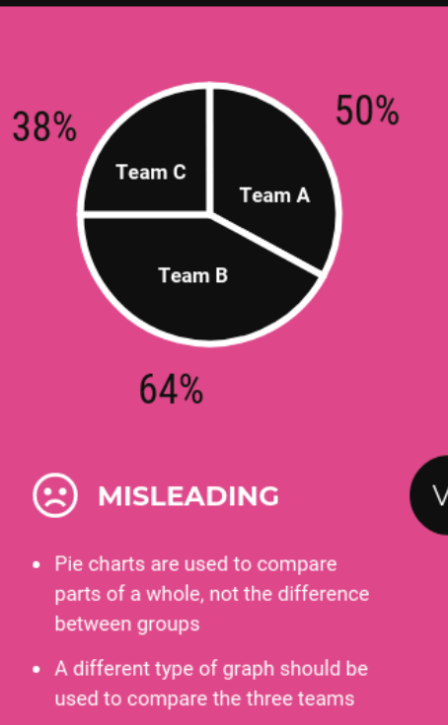faster than text.

2

Explore / Campaigns and Retention

Performance Indicat...

Campaigns and Retention
Owner Emily Chen

Create   Select All   Content type: Show all   Sort by: Type

Home
Explore
Favorites
Recents

Users
Groups
Schedules
Tasks
Site Status
Settings

Facebook Content Performance
Email Performance Overview
Website Traffic Trends

Content Performance   Email Performance Overview   Google Analytics

Historic Trends   Performance by Week   Renewals by Region

Search

ONE   TWO

A   B
C

APR   MAY   JUN

FEB   MAR   APR   MAY   JUN   JUL

Number of deaths | Killer

10 Shark
10 Wolf
100 Lion
100 Elephant
500 Hippopotamus
1,000 Crocodile
2,000 Tapeworm
2,500 Ascaris roundworm
10,000 Freshwater snail (schistosomiasis)
10,000 Assassin bug (Chagas disease)
10,000 Tsetse fly (sleeping sickness)
25,000 Dog (rabies)
50,000 Snake

World's Deadliest Animals
Number of people killed by animals per year

475,000
Human

725,000
Mosquito

SOURCES: WHO; crocodile-attack.info; Kasturiratne et al. (doi/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."

- 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
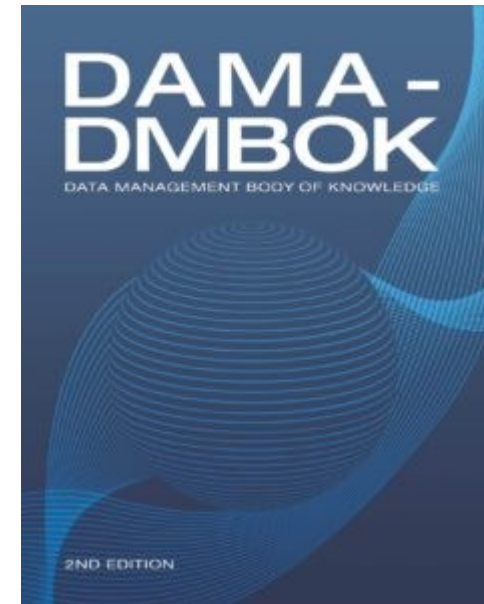
Course contents at a glance

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

Schedule

	lunedì 21/11/2022	martedì 22/11/2022	mercoledì 23/11/2022	giovedì 24/11/2022	venerdì 25/11/2022
8:00	**Data management and visuali...** APILETTI DANIELE AA - ZZ - 0 12A Lezione/Esercitazione	**Data science lab: process a...** BARALIS ELENA MARIA AA - ZZ - 0 2P	**Computational linear algebr...** BERRONE STEFANO AA - ZZ - 0 10D Lezione/Esercitazione		**Data management and visuali...** APILETTI DANIELE AA - ZZ - 0 LAIB4 Lezione/Esercitazione
9:00					
10:00		**Data management and visuali...** APILETTI DANIELE AA - ZZ - 0 2P Lezione/Esercitazione	**Decision making and optimiz...** DELLA CROCE DI DOJOLA FEDERICO AA - ZZ - 0 1P Lezione/Esercitazione / **Numerical optimization for ...** PIERACCINI SANDRA AA - ZZ - 0	**Numerical optimization for ...** PIERACCINI SANDRA AA - ZZ - 0 LAIB3 Lezione/Esercitazione	**Data management and visuali...** APILETTI DANIELE AA - ZZ - 0 LAIB4
11:00					**Computational linear algebr...** BERRONE STEFANO AA - ZZ - 0 10D
12:00			**Information Theory for Data...** GARELLO ROBERTO AA - ZZ - 0	**Data science lab: process a...** BARALIS ELENA MARIA AA - ZZ - 0 LAIB3 Lezione/Esercitazione SQUADRA 2	
13:00	**Data science lab: process a...** BARALIS ELENA MARIA AA - ZZ - 0 LAIB3 Lezione/Esercitazione SQUADRA 1	**Numerical optimization for ...** PIERACCINI SANDRA AA - ZZ - 0 4D Lezione/Esercita	**Information Theory for Data...** GARELLO ROBERTO AA - ZZ - 0 4 Lezione/Esercita	**Decision making and optimiz...** DELLA CROCE DI DOJOLA FEDERICO AA - ZZ - 0 6N Lezione/Esercita	**Data science lab: process a...** BARALIS ELENA MARIA AA - ZZ - 0 R1 Lezione/Esercitazione	
14:00					
15:00					**Information Theory for Data...** GARELLO ROBERTO AA - ZZ - 0 3S Lezione/Esercitazione
16:00			**Statistical methods in data...** FONTANA ROBERTO AA - ZZ - 0 5I Lezione/Esercitazione	**Statistical methods in data...** FONTANA ROBERTO AA - ZZ - 0 5I Lezione/Esercitazione	
17:00					
18:00	**Computational linear algebr...** BERRONE STEFANO AA - ZZ - 0 12D				**Computational linear algebr...** BERRONE STEFANO AA - ZZ - 0 10D

Schedule

Lectures

- **Monday**
  08:30-11:30
  classroom 12A + online
- **Tuesday**
  10:00-13:00
  classroom 2P + online

Schedule changes announced on the teaching portal.

Lab practice

- **Friday**
  08:30-10:00 team A
  10:00-11:30 team B
  Laboratory LAIB4 (no online…)
- Starting on Friday, **October 14**, 2022
- Till the end of the course
  - no December 09th - no December 23rd
- Each student will be assigned to a single Team (**either A or B**)
  - based on their surname
  - changing Team is allowed upon request
- Simone Monaco, Alessandro Fiori, Diego Monti

Exam

- See exam policy on the [official course web page](#) on the teaching portal
- Exam + Lockdown browser
- Bring your own notebook + test everything in advance
  - WiFi, power plug, updates...

Exam: Computer-based written test in class using POLITO platform;

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

- [max 6 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 additional electronic devices during the exam, besides their own notebook with Lockdown/Respondus.

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

Daniele Apiletti