Advanced Data Management

Academic year 2023/2024

9 CFU / 6 CFU

"You can have data without information, but you cannot have information without data."

(Daniel Keys Moran)

Who, when, and where?

Who is involved?

- Barbara Catania

- Giovanna Guerrini

- Ziad Janpih
  (help during the labs and the project)

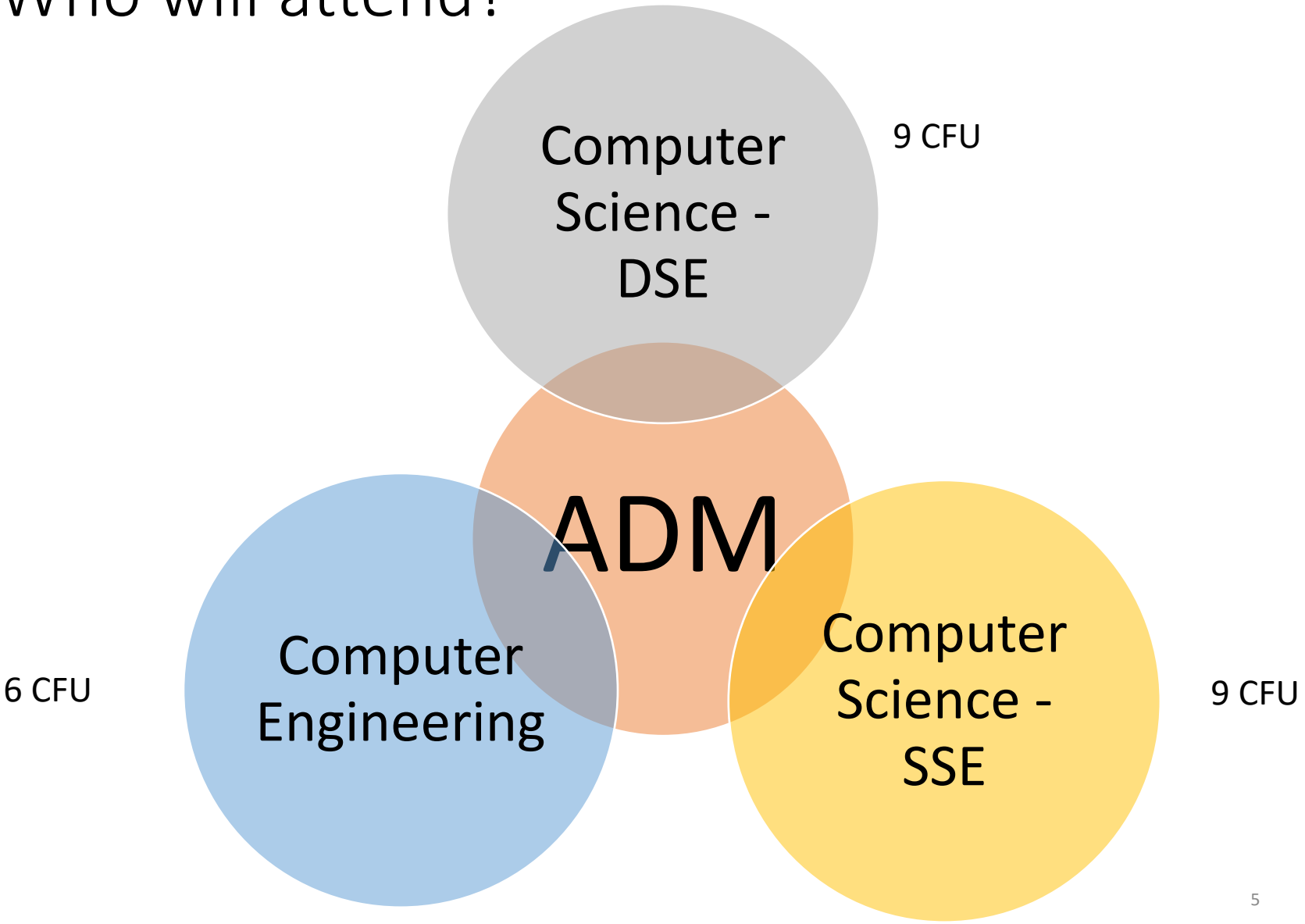
When and where will it take place?

- Monday, 11.00-13.00 [room 711, Valle Puggia]
- Wednesday, 11.00-13.00 [room 710, Valle Puggia]
- Thursday, 14.00-16.00 [room 711, Valle Puggia]

In presence, no streaming

Recording of lectures proposed in previous a. y. available on Aulaweb

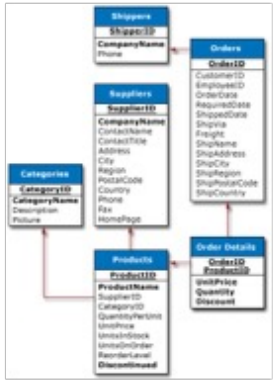
Who will attend?



Why this course?

Data Management in the Big-Data era

Why this course?



Transactional data
(standard data, internal to one organization)

Social media and networks
(all of us are generating data)

Mobile devices
(tracking all objects all the time)

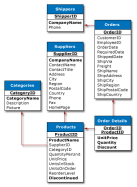
Scientific instruments
(collecting all sorts of data)

Sensor technology and networks
(measuring all kinds of data)

Why this course?

- big datasets (Volume)
- heterogeneous, often incomplete, and highly interconnected data (Variety and Veracity)
- data can be generated at a very high rate (very fast - Velocity)
- the volume of such data requires programming environments which exploit parallelism in order to cope with such huge volumes in an efficient way
- many previously unknown information can be extracted from them (high Value)

Why this course?

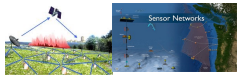


Transactional data
(standard data, internal to one organization)

Social media and networks
(all of us are generating da

Mobile devices
(tracking all objects all the time)

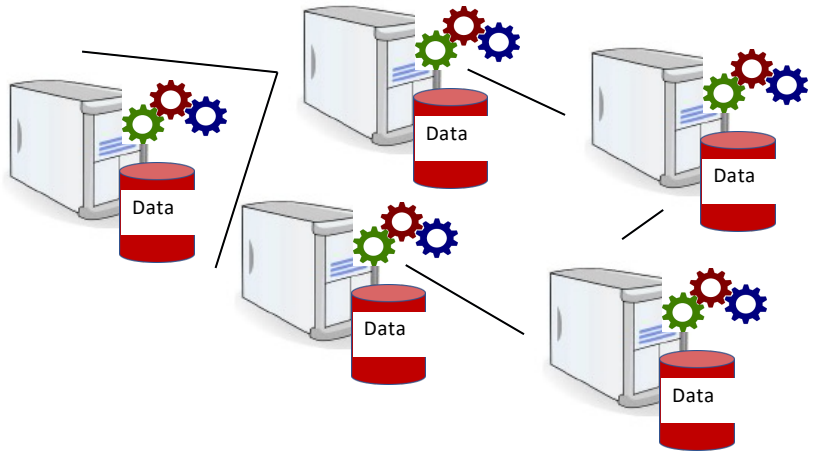
Scientific instruments
(collecting all sorts of data)



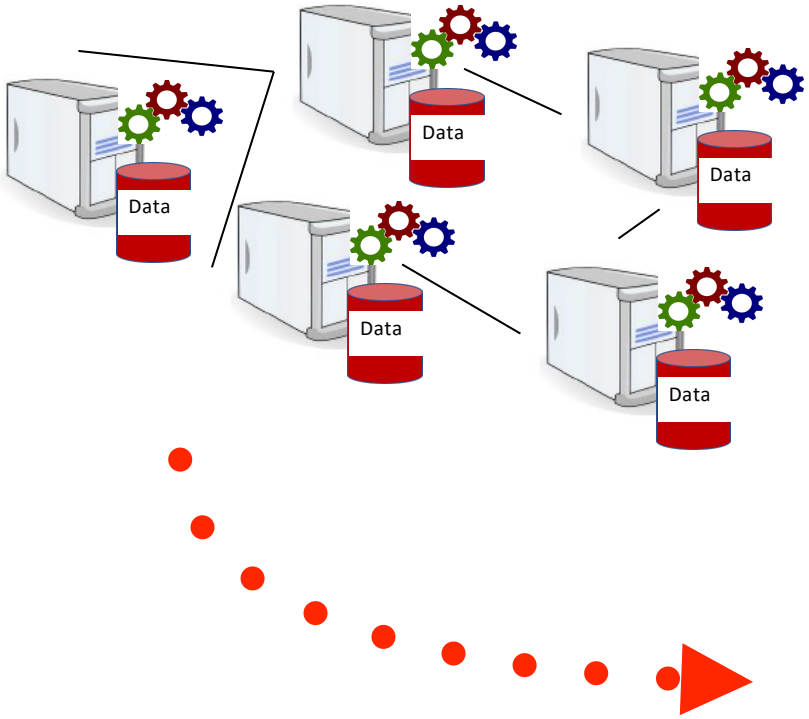
Sensor technology and networks
(measuring all kinds of data)

Data Management

- collect data
- integrate data
- clean data
- represent and store data
- query & process data

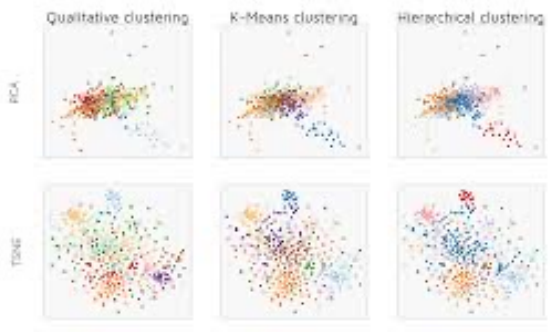


Why this course?



Data Analysis

- analyse your data
- interpret the obtained results
- take your decisions



Why this course?

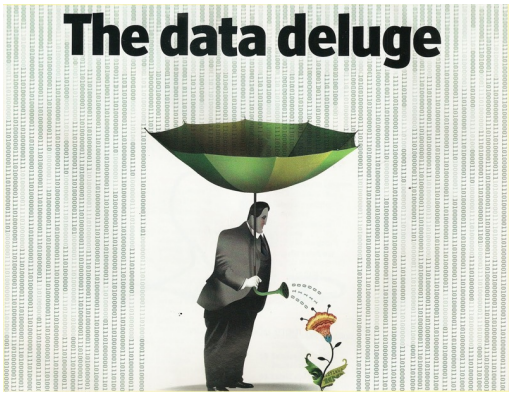
the way to the «value», through data analysis, could be very dangerous, be careful!

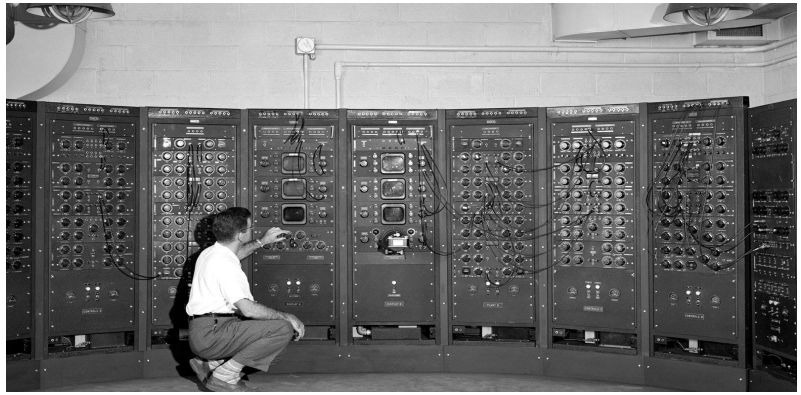
- performance, performance, performance!
- effectiveness
- heterogeneity
- flexibility

Why this course?

big data

big machines / big architectures





data management

Why this course?

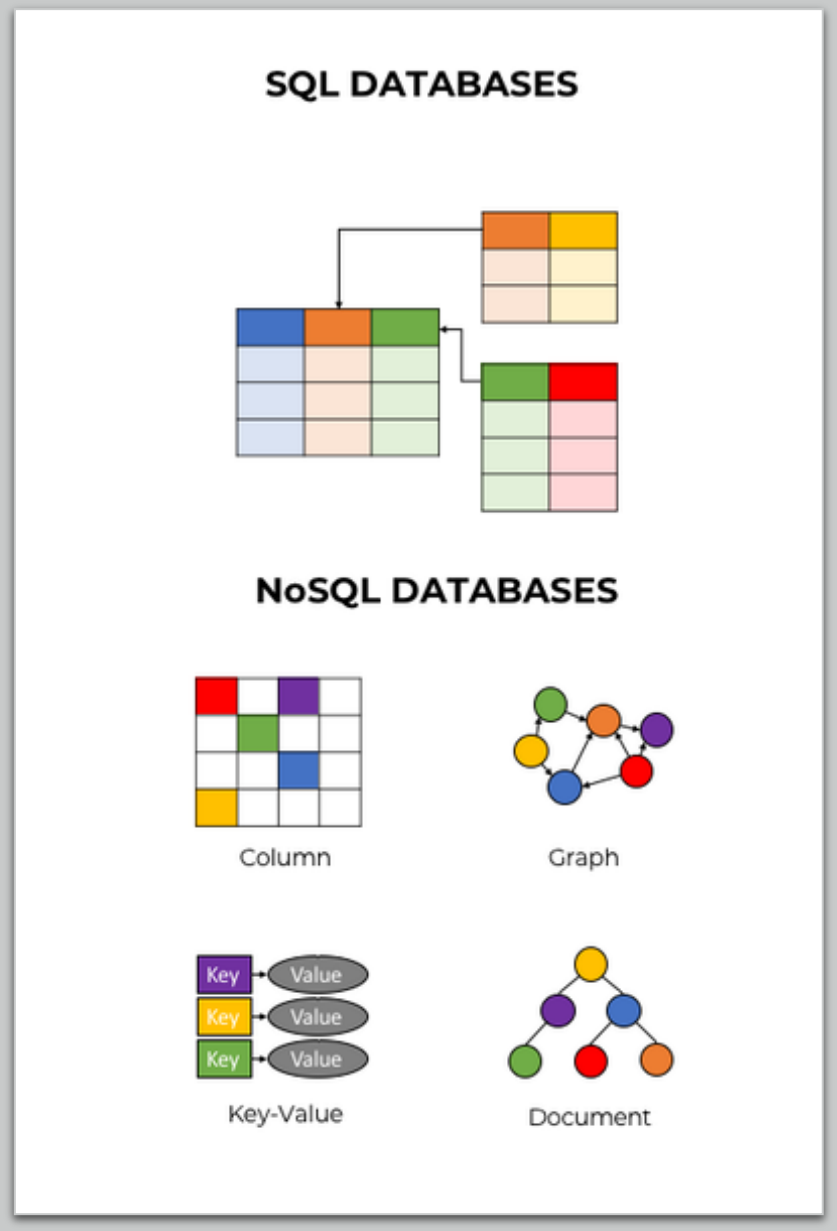
- unfortunately, traditional database management techniques do not scale to such huge datasets and do not effectively take into account issues raised by large-scale environments

- new solutions have therefore been devised

Why this course?

- high volume
- flexible data structure
- strong connection of data and applications
- moving away from using databases as integration points towards encapsulating databases within applications

- systems for large-scale data management, NoSQL systems



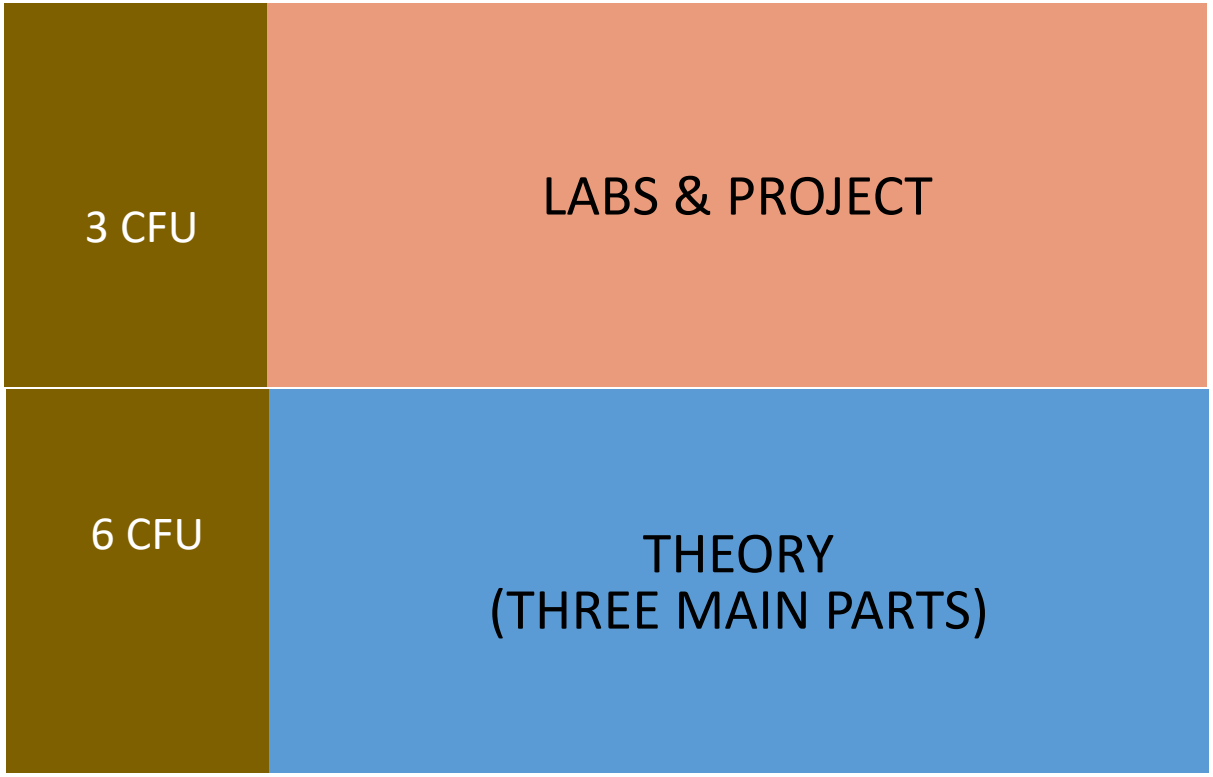
Why this course?

- data exchanged between systems and applications
- heterogeneous, possibly unknown and mostly uncontrolled sources
- need to agree on the data meaning in order to avoid dangerous misunderstandings

- semantic elicitation and the availability of appropriate semantic metadata are the key to the meaningful use of information in modern distributed environments

What will you learn?

What will you learn?



What will you learn?

REFERENCE PART | THEORY | LABS

What will you learn?

PART I **Only 6 CFU**	Recap on large scale distributed architectures and data-intensive computing	

REFERENCE PART	THEORY	LABS

What will you learn?

Learning outcome - part I

- DESCRIBE the principles for data management in distributed systems, environments for large-scale data processing, systems for large-scale data management
- UNDERSTAND the differences between traditional data processing and management and large-scale (semantic) data processing and management

What will you learn?

PART II	Systems for large-scale data management (NoSQL systems)	Riak, Cassandra, MongoDB, Neo4J
PART I **Only 6 CFU**	Recap on large scale distributed architectures and data-intensive computing	

REFERENCE PART	THEORY	LABS

What will you learn?

PART III **Only 9 CFU**	Semantic data management	RDF, SPARQL, OWL
PART II	Systems for large-scale data management (NoSQL systems)	Riak, Cassandra, MongoDB, Neo4J
PART I **Only 6 CFU**	Recap on large scale distributed architectures and data-intensive computing	

REFERENCE PART	THEORY	LABS

What will you learn?

Learning outcome - parts II & III

- UNDERSTAND the differences between the presented approaches for large-scale (semantic) data management
- SELECT the system and the methodology for large-scale (semantic) data management, suitable in a given application context
- USE some of the presented systems for large-scale (semantic) data management, for solving simple problems
- USE at least one of the presented systems for large-scale (semantic) data management for solving non-trivial problems
- ANSWER questions related to large-scale (semantic) data management
- SOLVE exercizes related to the data design in some of the presented systems and the interaction with such systems, through the available languages

Prerequisites

Prerequisites

- basics on large-scale distributed systems and computing
  - *DSE-Computer Science students: Distributed Computing, I year*
  - *all the other students: in PART I of the course*
    - *partial overlap with Distributed Computing for SSE-Computer Science students*

- solid foundation in database design and querying (*see AulaWeb for references*)
  - you must pass the test for taking the exam
  - *first test on Monday, October 2*
  - *before any exam date*

How is the course organized?

Resources (see Aulaweb)

- books
- manuals
- scientific papers
- software links

- slides
- recording of lectures proposed in previous a.y.

Lectures and labs

- Lectures (in presence)
  - on the main theoretical and methodological issues
  - exercizes on the main theoretical and methodological issues
  - talks from prominent experts in the field
- labs on the main technologies presented in the course
  - each lab = one groupwork assignment (2 persons each)
  - we will communicate soon the proposed organization
- quiz (online) on the main concepts proposed in the course
- bonus based on labs and quiz
  - no oral exam (see later)
  - up to 2 points bonus

Project (for both 9 CFU and 6 CFU)

- groupwork (up to 2 persons)
- two options
  A. design and development
  B. research-based
     - *only for students that pass the assignments proposed during the course with a grade higher than a given threshold (more information later)*

Project (for both 9 CFU and 6 CFU)

A - Design & development project
(both 9 CFU and 6 CFU)

- choose one application domain
- write a requirement analysis document
- choose one technology, among those studied
- deliver a document explaining your choices
- design and develop your solution
- deliver it as soon as you want to take the exam

B- Research-based project
(both 9 CFU and 6 CFU)
only for students that pass the prerequisite test and the assignments proposed during the course with a grade higher than a given threshold (more information later)

- select one topic among a set of available ones
- review the literature on the selected topic
- prepare a report summarizing the investigated topic
- deliver it as soon as you want to take the exam

Exam modalities

- written exam (questions/exercizes)
- project delivery + video presentation
- oral exam [only for those that do not delivered the assignments]
  - theoretical questions and / or practices of the course topics
  - up to 2 points

- the project has to be submitted before the written exam
- the oral exam [if needed] will be scheduled after the written exam

First of all enroll to the course on Aulaweb!