ISA

Data Management and Storage

Lesson 01 - Welcome! Introduction.

U LISBOA | UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR DE AGRONOMIA

Welcome!

Overview of the course:

- Introduction to data management and data management systems
- Concept of relational databases
- Data quality principles, vocabularies and ontologies
- SQL language and application for creating and using databases
- NoSQL systems and their application
- Agro-environment data resources
- Introduction to the concept of large data sets

Assessment

Components:

- Tests (40%) or final exam

  Two short tests focused on operational knowledge

  Dates: 2024-11-15 (week 9) and 2024-12-20 (week 14)

- Project (40%) - group work

  Report on the design and implementation of a data management system

- Participation (20%)

  Weekly practice exercises

Project - work group

Goal

Design and implement a data management system integrating data from related domains

Components

- identification of data sources (actual or simulated), but including at least one web data source
- database schema design
- data pre-processing
- database implementation
- data use

Deliverable

- written report including goal, database diagram, documentation and use examples
- database backup dump with README.md instructions on how to implement/recover
- SQL script file with sample queries for data manipulation

Course management

Fenix: https://fenix.isa.ulisboa.pt/courses/gestadad-283463546571604

Github: https://github.com/isa-ulisboa/greends-dms-exercises

Moodle: https://elearning.ulisboa.pt/course/view.php?id=9101

What is data?

Examples of data:

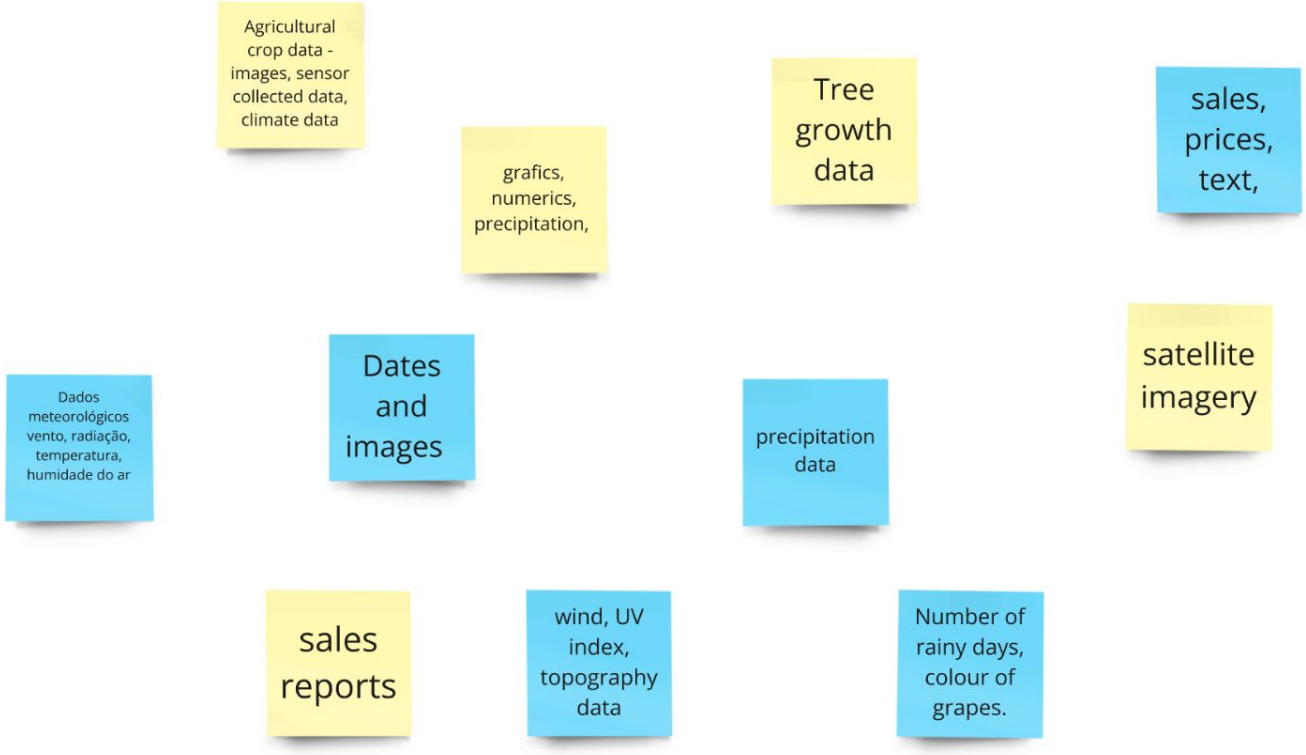
- identify types of data
- identify types of data sources

MIRO Activity

https://tinyurl.com/greends-dms24-01

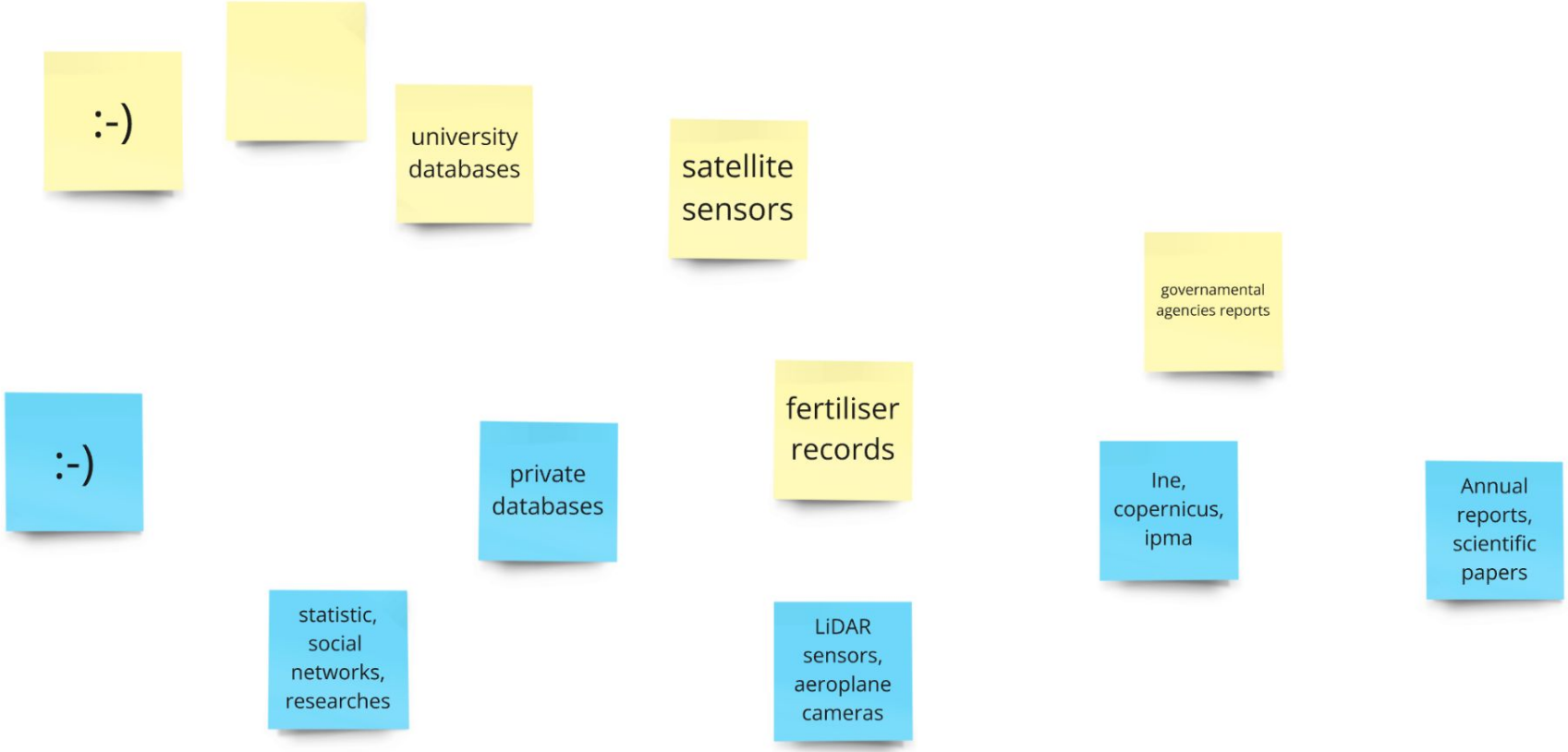
Give examples of types of data

Pick a post-it and reply in few words



Give examples of data sources

Pick a post-it and reply in few words



What is data?

DATA:

- Facts about things (text, numbers)
  - properties, measurements, descriptions, etc…
- Sounds, images, movies

INFORMATION:

- data with context (data organised and interpretable)

Algés
is the locality
where I live.

- data is a collection of facts, information puts them in context
- data not organised, information is
- data points are unrelated, information can provide links
- data does not have meaning *per se*. After being analysed, it may become information
- data does not depend on information. Information depends on data
- data does not support decision making, information can support it.

What is data?

Algés
is the locality
where I live.

- data is a collection of facts, information puts them in context
- data not organised, information is
- data points are unrelated, information can provide links
- data does not have meaning *per se*. After being analysed, it may become information
- data does not depend on information. Information depends on data
- data does not support decision making, information can support it.

Information systems

Data is the new oil!

The evolution from an industrial to an information and knowledge society is represented by the assessment of information as a factor in production. But different of material goods, because:

Representation

Processing

Combination

Age

Original

Vagueness

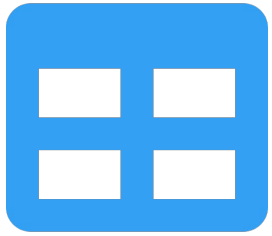
Medium

https://doi.org/10.3390/fi15020071

Types of datasets

Structured

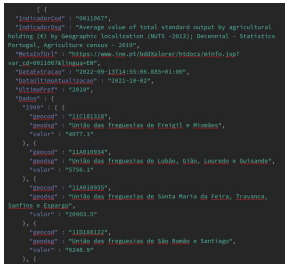
All records share a **common schema**, i.e., they all have values for the same attribute



Example: tabular data, csv, tsv

Semi-structured

Each record may have a **different schema** (also sometimes called heterogeneous data)



Example: email, json, xml, tree data, graph data

Unstructured

The **structure is not explicit**, as it is in structured and semistructured data (that is, no tags or markers to separate the records into attributes).

https://unsplash.com/photos/Wh9ZC4727e4

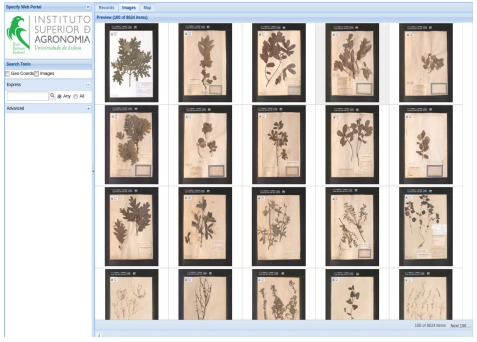
Example: news, social network post, etc.

Data store

Data store

Repository for persistently storing and managing collections of data. Can be based on

- databases
- files
- file systems
  - Hadoop Distributed File System (HDFS) - big data
- email storage systems
- distributed data stores



Data store

Data store - Can be organised in tree types

Data lake

Storage repository that holds vast amounts of raw, unstructured, semi-structured, and structured data in its **native format**.

- Google drive
- AWS S3
- OneDrive

Data Warehouses

Centralized repository that stores structured (or semi-structured), cleaned, and processed data from various sources.

- relational databases
- cloud databases

Data Platform

Unified system for efficiently managing and analyzing large datasets. It integrates components like databases, data lakes, and data warehouses to handle structured and / or unstructured data depending on the use cases

- Data lakes
- Data warehouses

Databases

Collection of data stored in a structured manner.

Hierarchical Databases	Relational Databases	Non-Relational Databases
A single object (the “parent”) has one or more objects beneath it (the “child”). Child records can’t have more than one parent. Windows Registry	Also known as RDBMS. Store information in discrete tables, which can be joined together by fields known as foreign keys. Use SQL for operations. MariaDB MySQL ORACLE IBM DB2 SQL Server	NoSQL differ from RDBMS in that they can be schema-agnostic, allowing unstructured and semi-structured data to be readily stored and processed. JSON’ like | key : value | graph mongoDB. neo4j cassandra

RDBMS

Relational DataBase Management Systems

MariaDB MySQL ORACLE IBM DB2 SQL Server

Database

Data and relationships in tables
Metadata in system tables

+

Management System

Query and manipulation language SQL
Special functions (recovery, reorganization, security, data protection, etc.) with SQL

→

Relational Database System

Integrated systems for the consistent management of tables

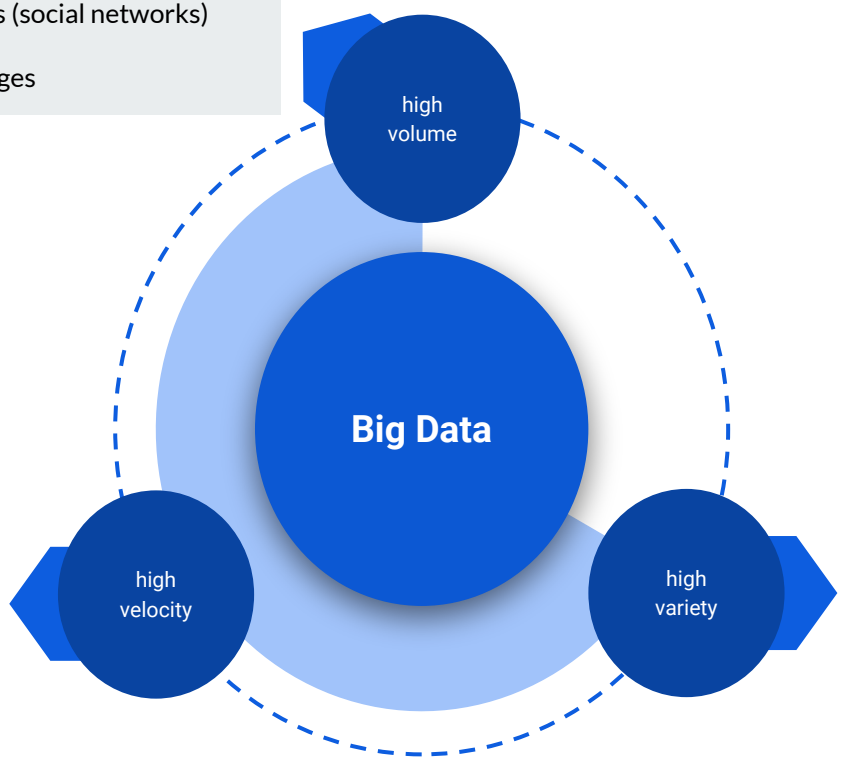
RDBMS

Relational DataBase Management Systems - properties

- **Model**: relational model, data represented in tables
- **Schema**: definition of tables and attributes
- **Language**: SQL data definition
- **Architecture**: data independence, data separated from applications
- **Multi-user operation**: several users
- **Consistency assurance**: data integrity
- **Data security and data protection**: protection from loss, destruction, unauthorized access

Big Data

data streams (social networks)
loggers, IoT
satellite images



data streams
loggers, IoT
satellite images

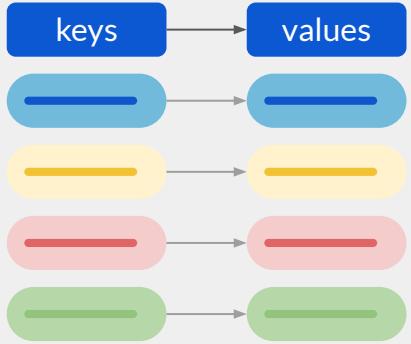
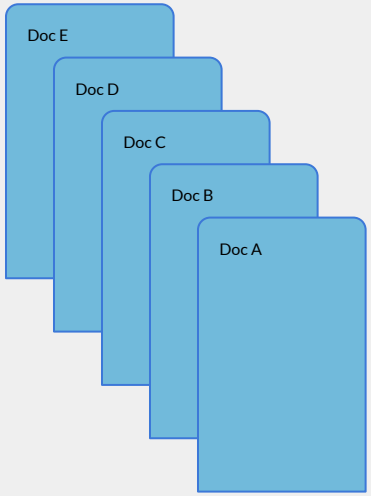
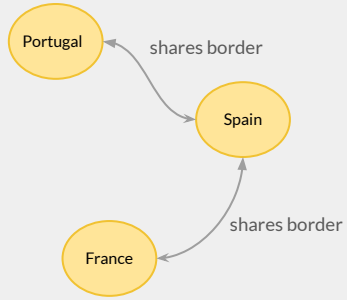
structured
semi-structured
unstructured (text, images, videos, sound)

Big Data - examples

01	Text	• continuous text • structured text • collection of texts, tags, etc
02	Graphics	• road maps • 3D graphics • technical drawing, etc
03	Image	• photographs • satellite images • x-ray image, ect
04	Audio	• language • music • sounds, animal sounds, etc
05	Video	• film • animation • ads, phone conferences, etc

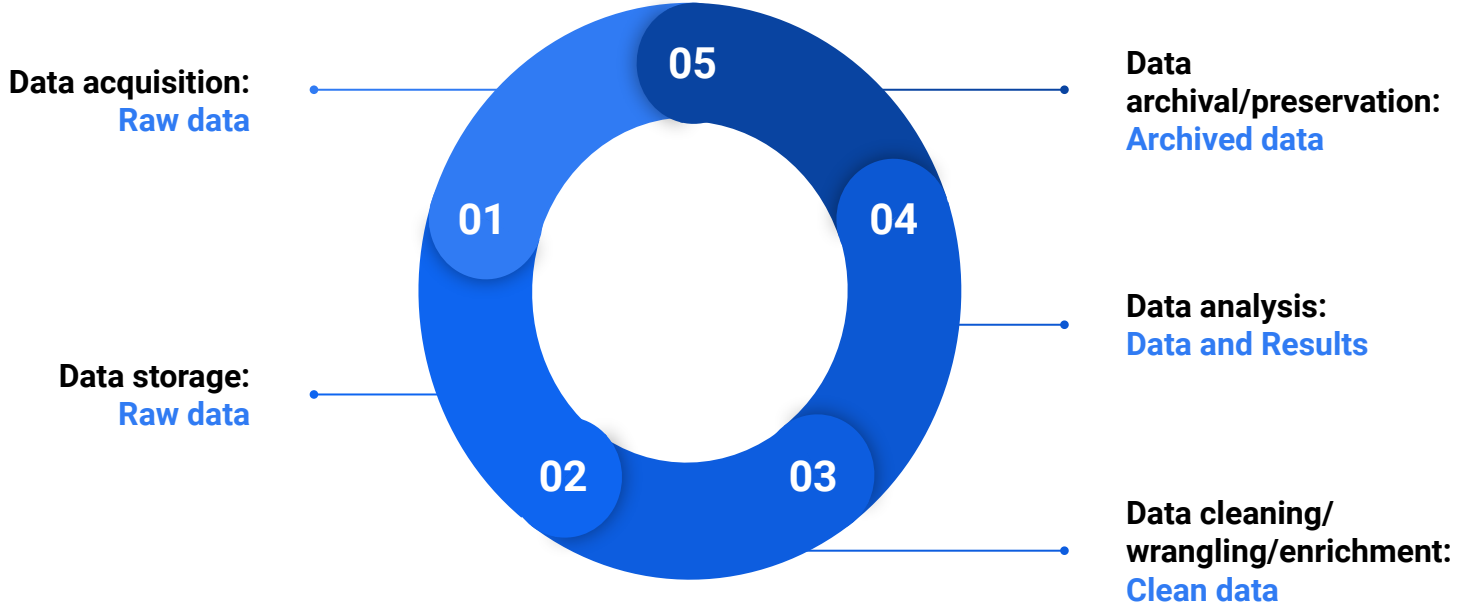
NoSQL Databases

Nonrelational data management approaches (in result of Web 2.0)

Key-value store	Document store	Graph database
		

Data Life Cycle

Data life cycle - transformations applied to data and to the states that data goes through as a result of these transformations



Data Life Cycle

Acquisition	Storage	Cleaning/ enrichment	Analysis	Archival/ Preservation
• acquire, collect, gather • sample of the population • opportunistic (easy to access) • observational - uncontrolled settings	• extract, transform, load (ETL) • EDA • metadata (augmentation) • visualization techniques	• inconsistencies • missing, entry errors • outlier detection • duplicates • augmentation • transformation (scaling, normalization, binarization, etc.)	• dimension reduction • supervised (regression, classification) • unsupervised (clustering, kNN)	• deleted • archived • published (metadata)

Data management in organisations

Roles	Tasks	Tools
Data architecture	- Creation and maintenance of data architecture - Definition of data protection rules	- Data analysis and design methodology - tools for information modeling
Data administration	- Management of data and methods using standardisation guidelines and international standards - Consultation of developers and end users	- Data dictionary systems - Tools for cross-reference and usage lists

Data management in organisations

Roles	Tasks	Tools
Data technology	- Installation, reorganisation and secure data content - Definition of the distribution concept - Disaster prevention and recovery	- Database system services - Tools for performance optimisation - Tools for recovery/restart
Data utilisation	- Data analysis and interpretation - Knowledge creation - Pattern detection - Modelling	- Analysis tools - Report generation - Data mining tools - Visualization methods

Data management in organisations

Data specialists

Data architects:

- access adjusted to business model
- distribution
- fragmentation
- replication

Data specialists:

- database technology
- manage infrastructure
- security

Data scientists:

- analysis
- modelling
- interpretation
- prediction

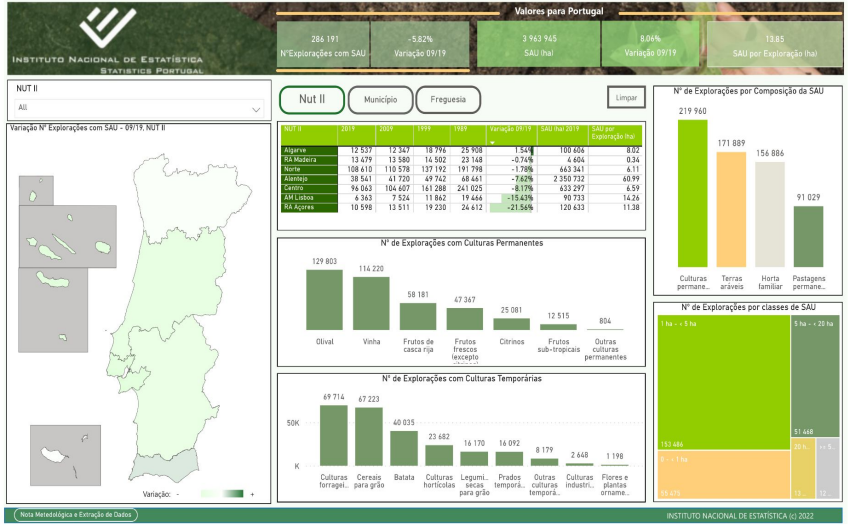
Use case

Design and implement an extended database for the Agricultural Census 2019 in Portugal:

link

Objectives

- be able to respond to selected questions based on the Census report
- extend with data from additional domains of knowledge



https://www.ine.pt/scripts/db_ra_2019.html

Additional reading

Instituto Nacional de Estatística - Recenseamento Agrícola. Análise dos principais resultados : 2019. Lisboa : INE, 2021. Disponível em https://www.ine.pt/xurl/pub/437178558. ISBN 978-989-25-0562-6