COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

Week 3

Relational Databases: Where Big Data is Typically Stored

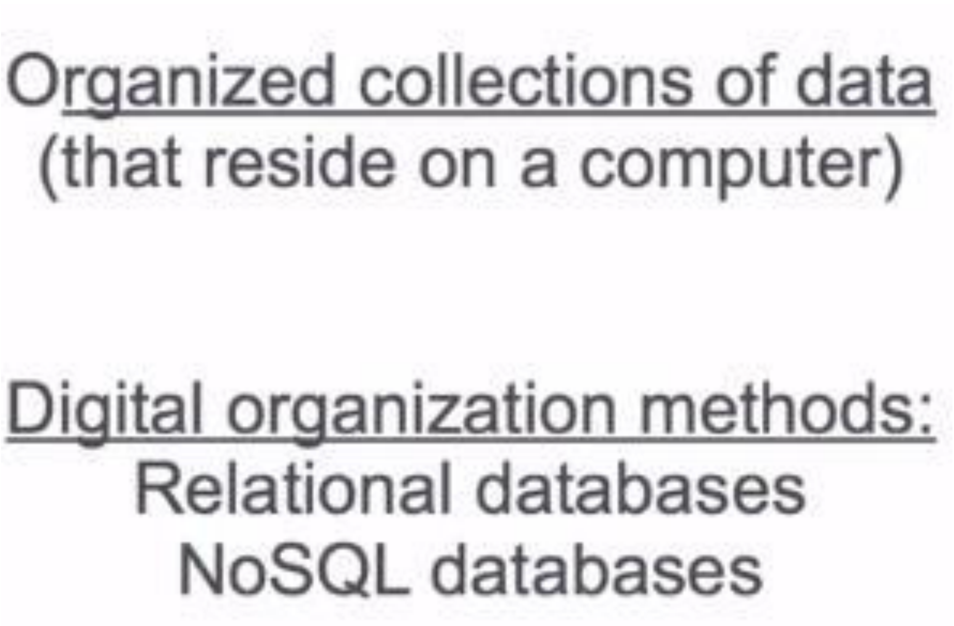
Applied Data Science

Columbia University - Columbia Engineering

Course Agenda

- Week 1: Python Basics: How to Translate Procedures into Codes
- Week 2: Intermediate Python — Data Structures for Your Analysis
- **Week 3: Relational Databases — Where Big Data is Typically Stored**
- Week 4: SQL — Ubiquitous Database Format/Language
- Week 5: Statistical Distributions — The Shape of Data
- Week 6: Sampling — When You Can’t or Won’t Have ALL the Data
- Week 7: Hypothesis Testing — Answering Questions About Your Data
- Week 8: Data Analysis and Visualization — Using Python’s NumPy for Analysis
- Week 9: Data Analysis and Visualization — Using Python’s Pandas for Data Wrangling
- Week 10: Text Mining — Automatic Understanding of Text
- Week 11: Machine Learning — Basic Regression and Classification
- Week 12: Machine Learning — Decision Trees and Clustering

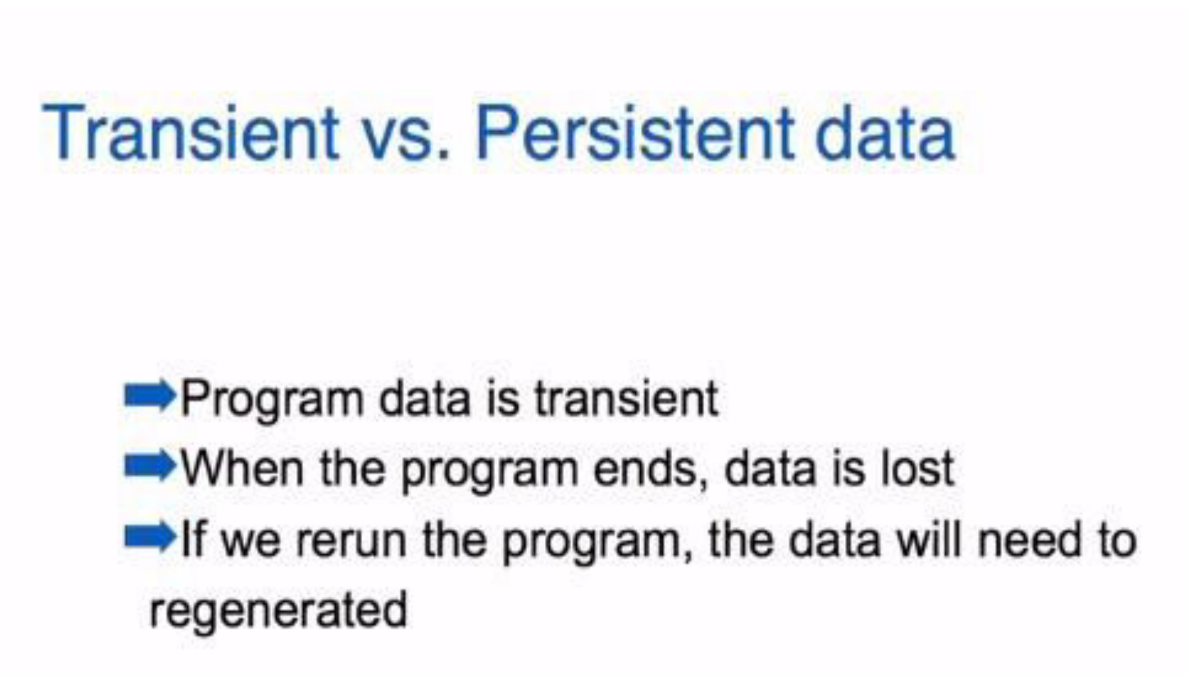
Databases: Definition and Data Types



Organized collections of data
(that reside on a computer)

Digital organization methods:
Relational databases
NoSQL databases

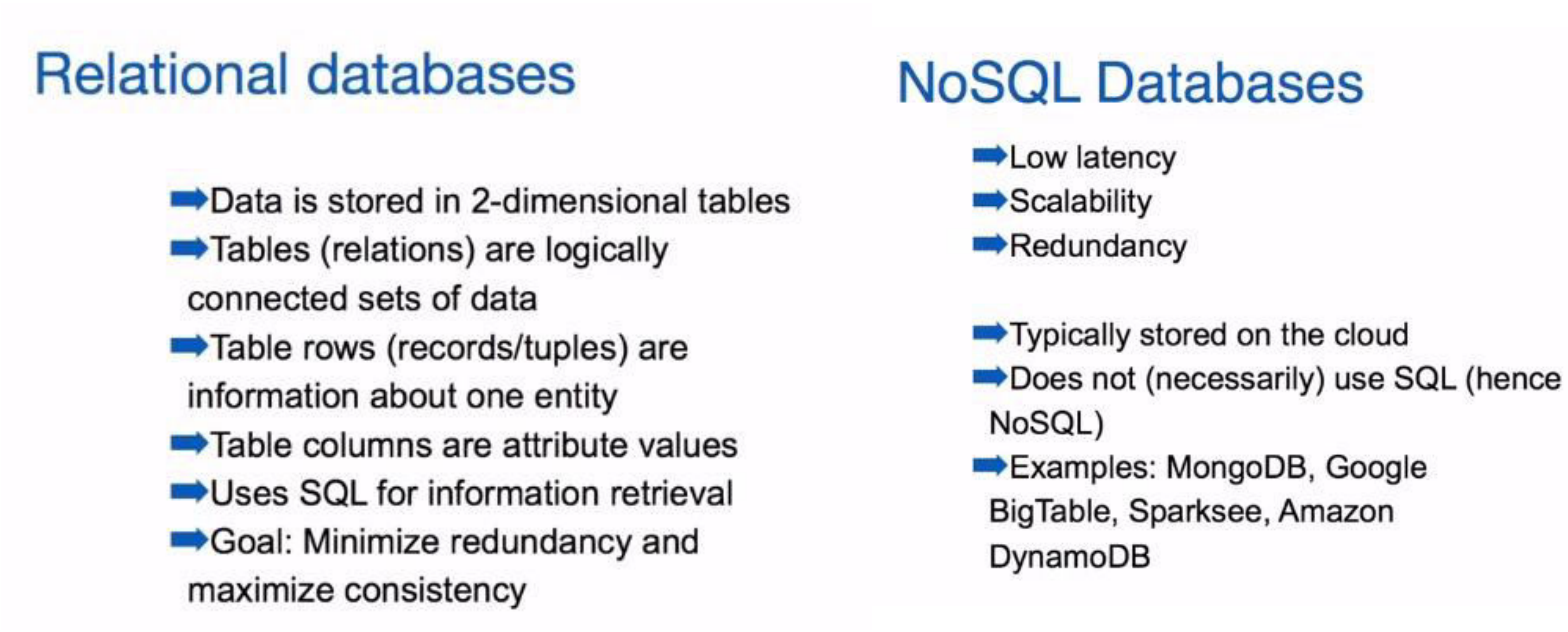
Types of Data



Transient vs. Persistent data

➡Program data is transient
➡When the program ends, data is lost
➡If we rerun the program, the data will need to regenerated

Types of Databases



Relational databases

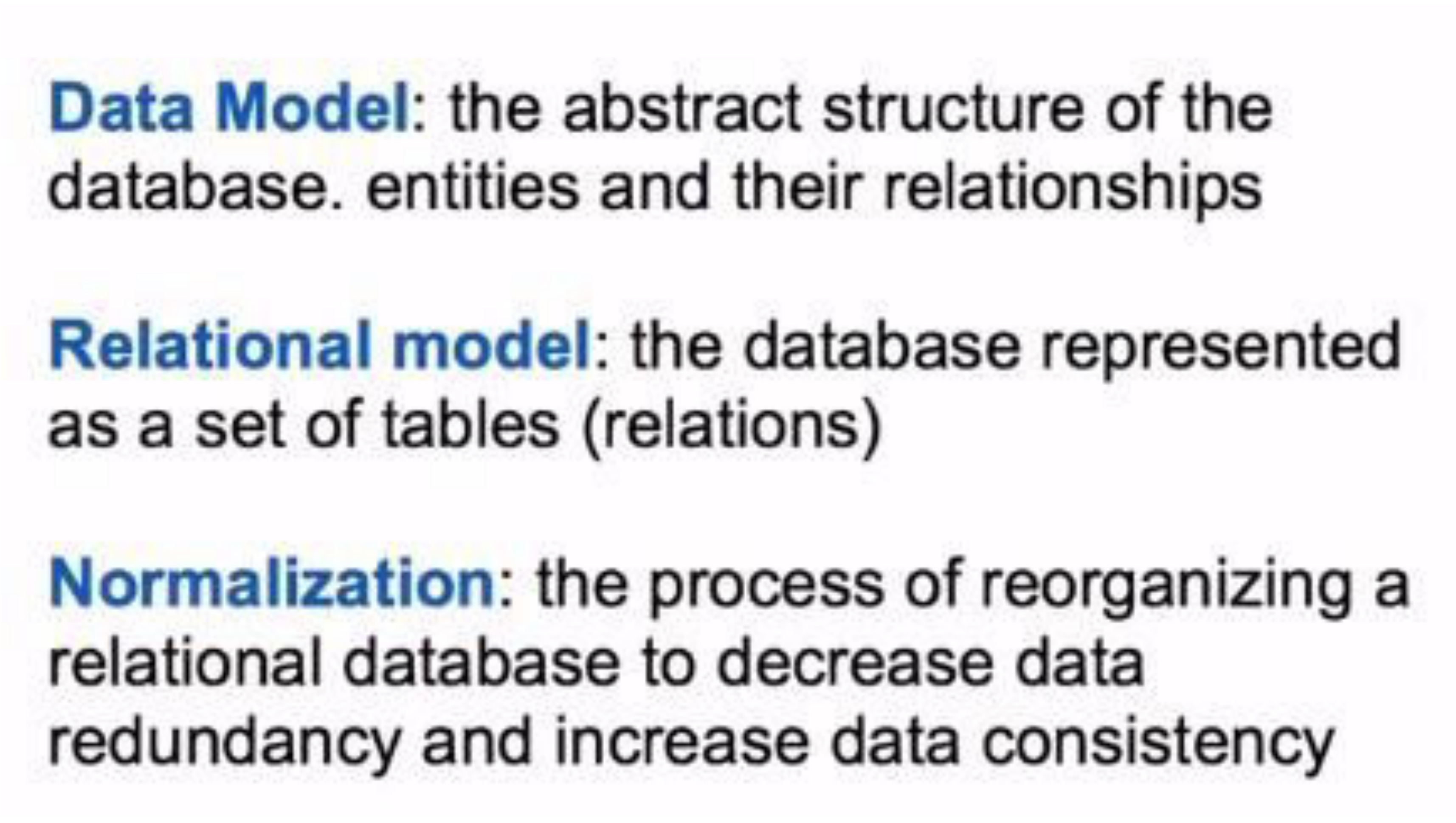
- Data is stored in 2-dimensional tables
- Tables (relations) are logically connected sets of data
- Table rows (records/tuples) are information about one entity
- Table columns are attribute values
- Uses SQL for information retrieval
- Goal: Minimize redundancy and maximize consistency

NoSQL Databases

- Low latency
- Scalability
- Redundancy

- Typically stored on the cloud
- Does not (necessarily) use SQL (hence NoSQL)
- Examples: MongoDB, Google BigTable, Sparksee, Amazon DynamoDB

Relational Databases – Key Aspects

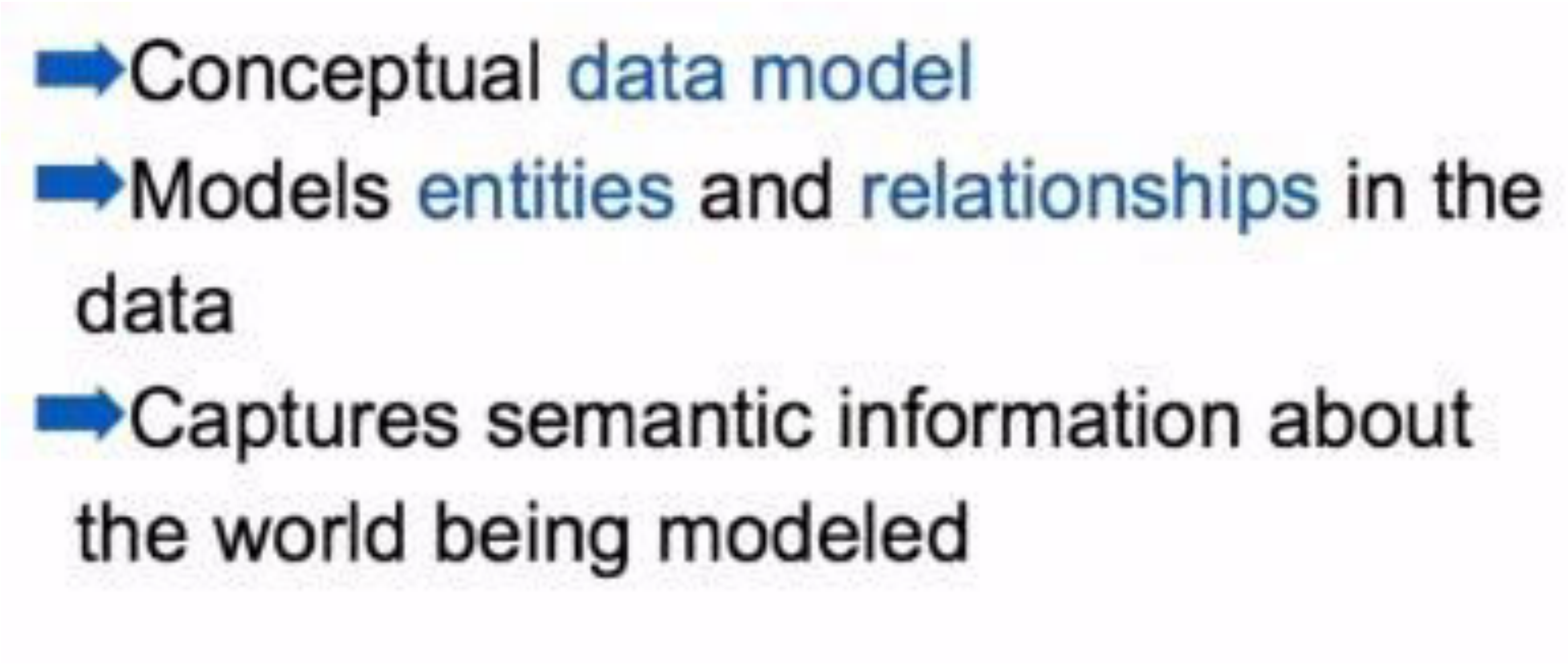


Data Model: the abstract structure of the database. entities and their relationships

Relational model: the database represented as a set of tables (relations)

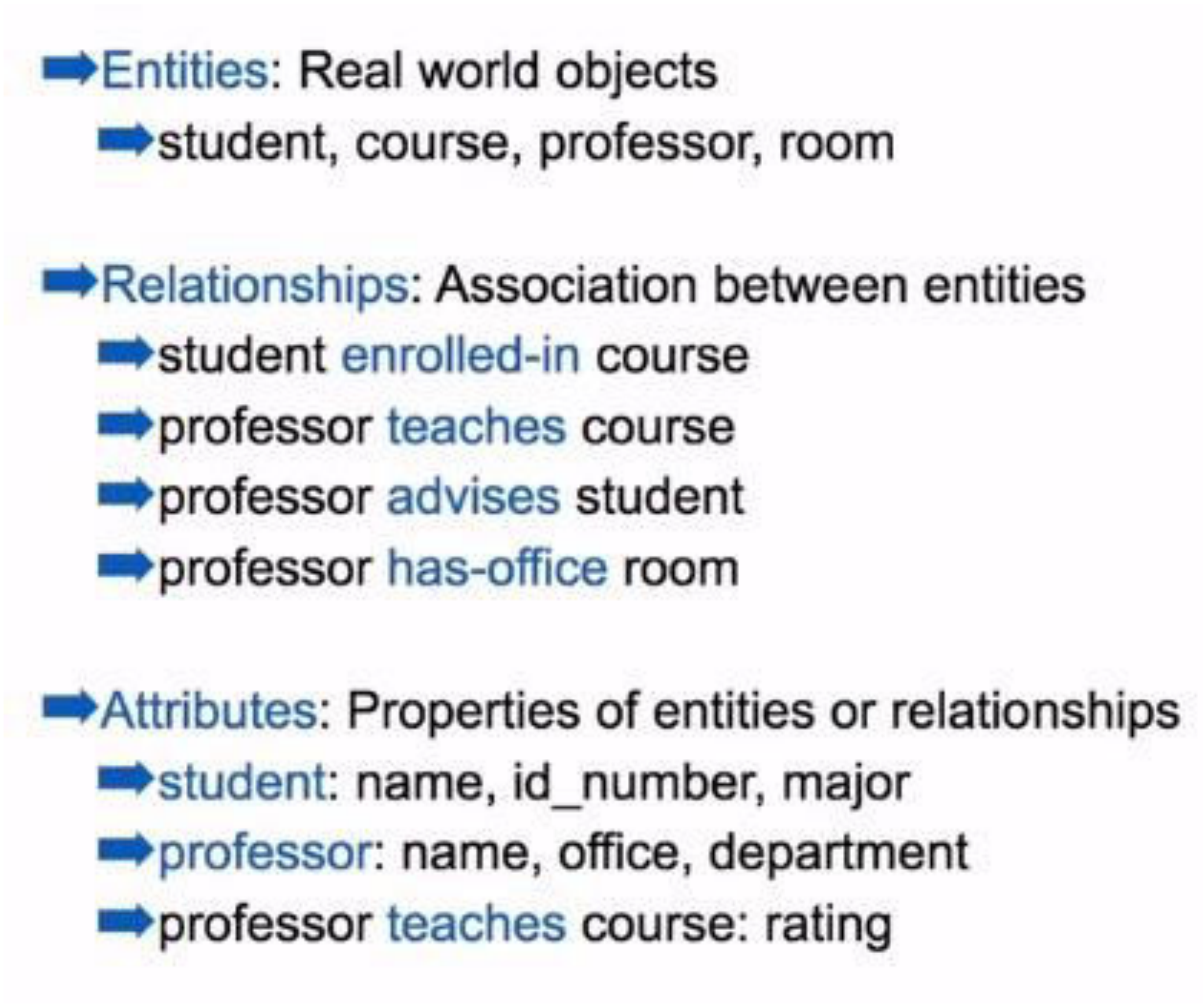
Normalization: the process of reorganizing a relational database to decrease data redundancy and increase data consistency

Entity-Relationship Model - Definition



- Conceptual data model
- Models entities and relationships in the data
- Captures semantic information about the world being modeled

Entity-Relationship Model – Components

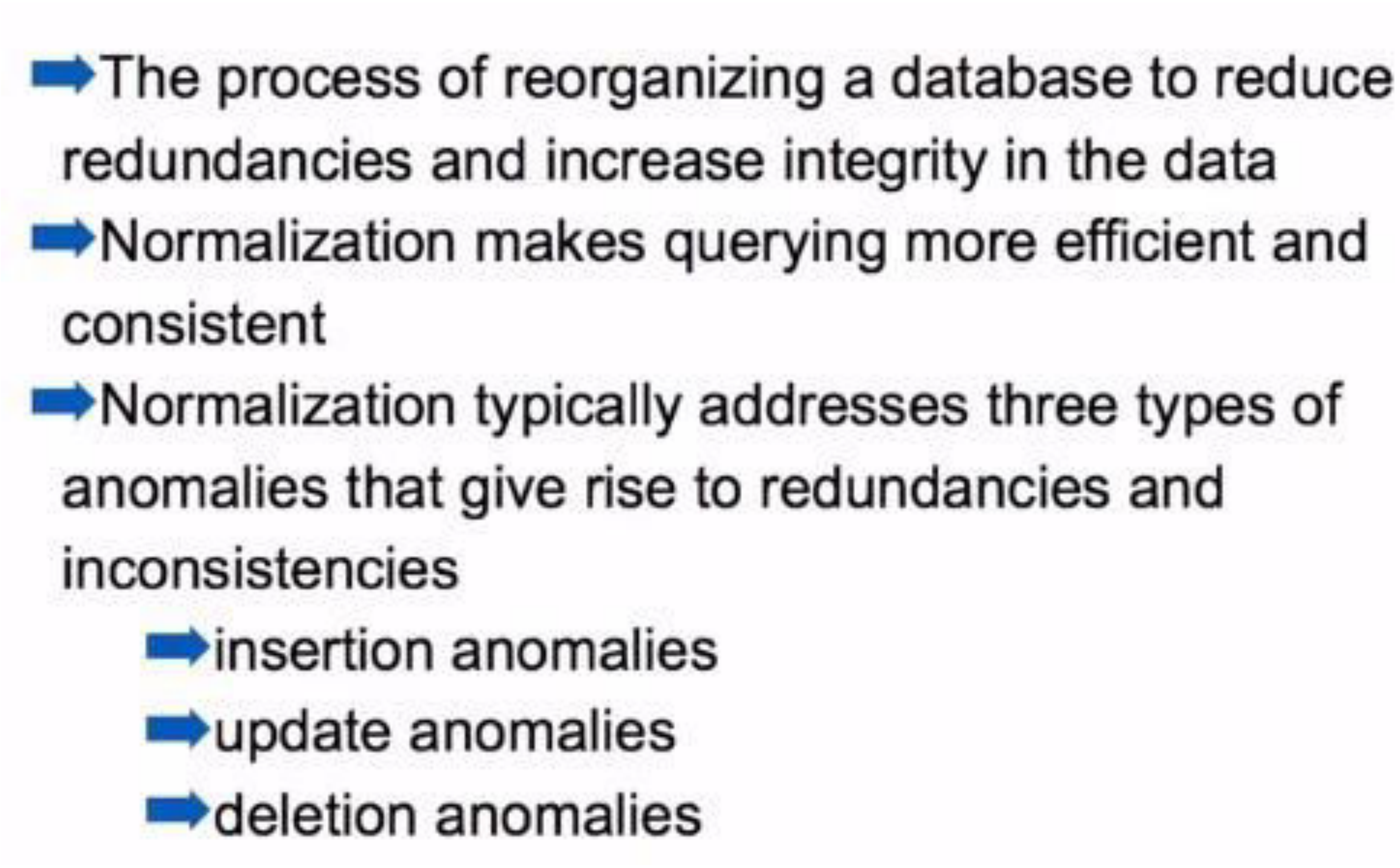


- Entities: Real world objects
  - student, course, professor, room

- Relationships: Association between entities
  - student enrolled-in course
  - professor teaches course
  - professor advises student
  - professor has-office room

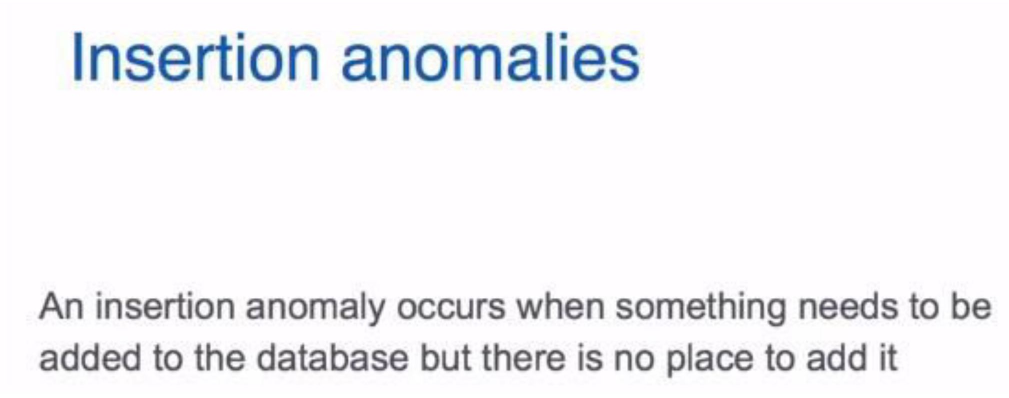
- Attributes: Properties of entities or relationships
  - student: name, id_number, major
  - professor: name, office, department
  - professor teaches course: rating

Normalization



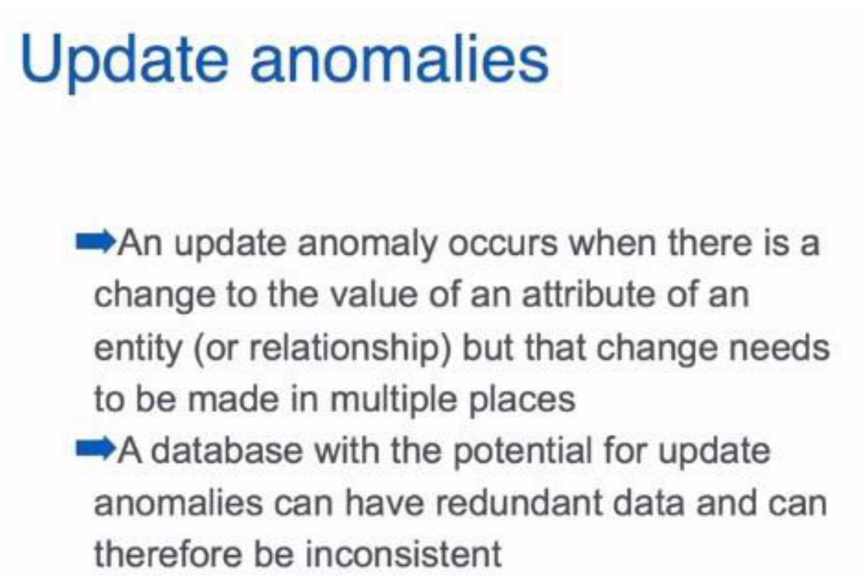
- The process of reorganizing a database to reduce redundancies and increase integrity in the data
- Normalization makes querying more efficient and consistent
- Normalization typically addresses three types of anomalies that give rise to redundancies and inconsistencies
    - insertion anomalies
    - update anomalies
    - deletion anomalies

Database Anomalies



Insertion anomalies

An insertion anomaly occurs when something needs to be added to the database but there is no place to add it



Update anomalies

- An update anomaly occurs when there is a change to the value of an attribute of an entity (or relationship) but that change needs to be made in multiple places
- A database with the potential for update anomalies can have redundant data and can therefore be inconsistent

Deletion anomalies

- A deletion anomaly occurs when deleting something from the database results in some or the other, most important, fact being deleted as well
- A database with the potential for deletion anomalies can lose data

GLOBAL IVY EDUCATION

www.emeritus.org