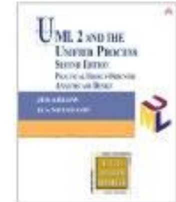


Lecture 5

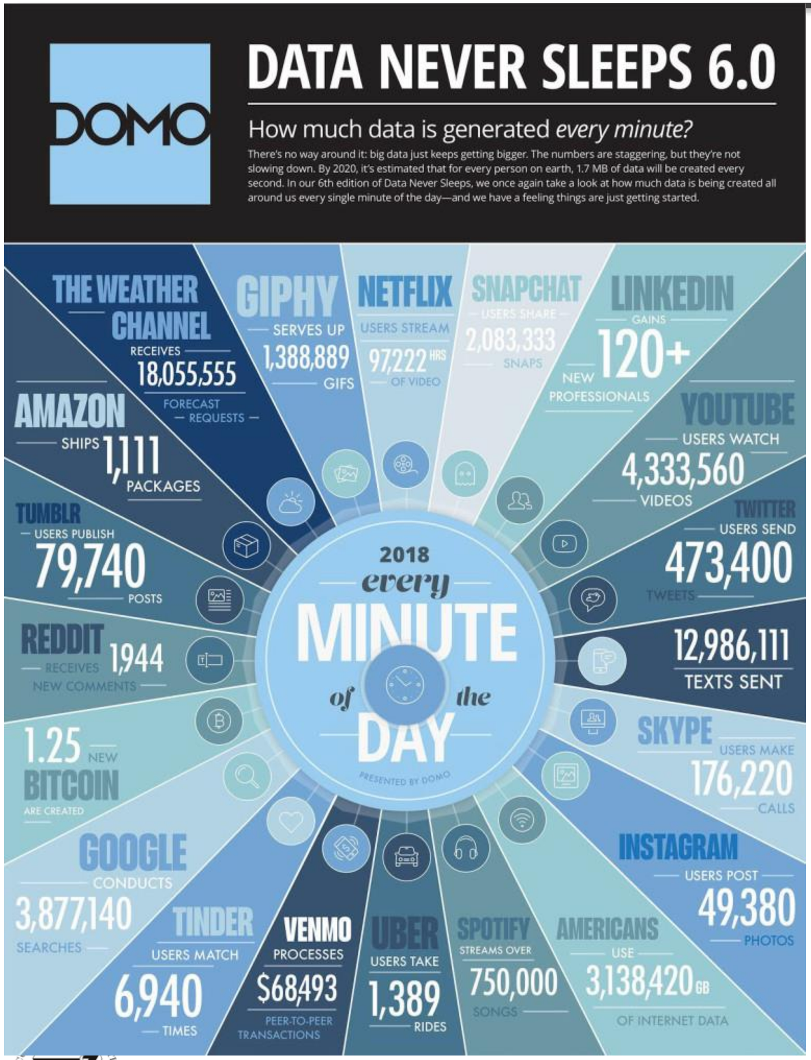
DATA MODELLING AND MANAGEMENT

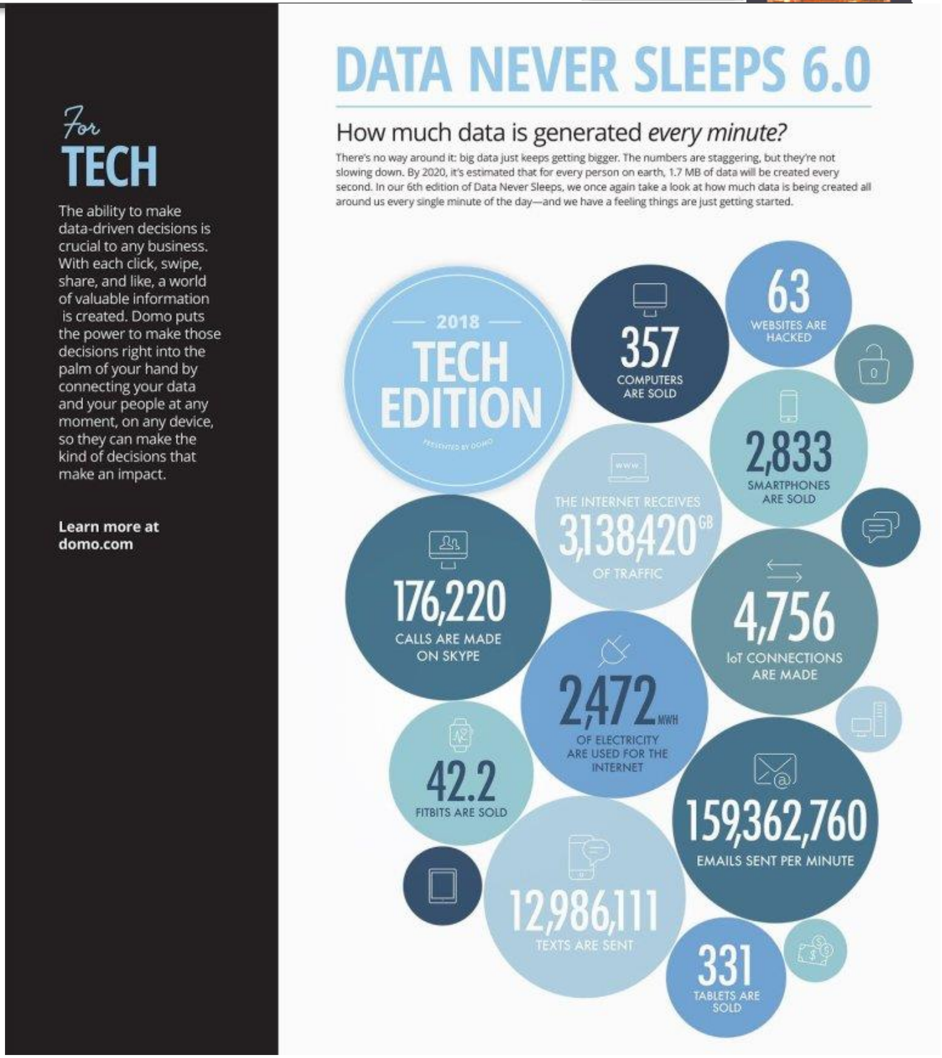
PB007 Software Engineering I
Faculty of Informatics, Masaryk University

Fall 2020

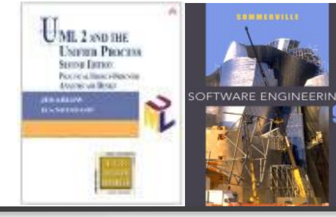
Motivation

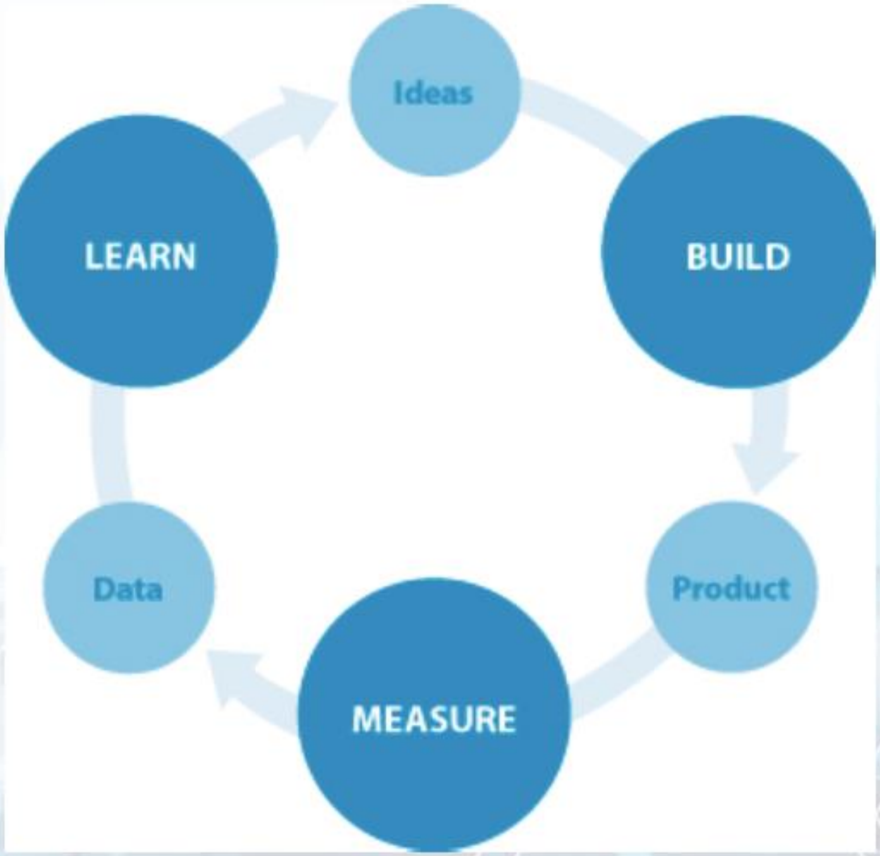

SOMMERVILLE
SOFTWARE ENGINEERING
9

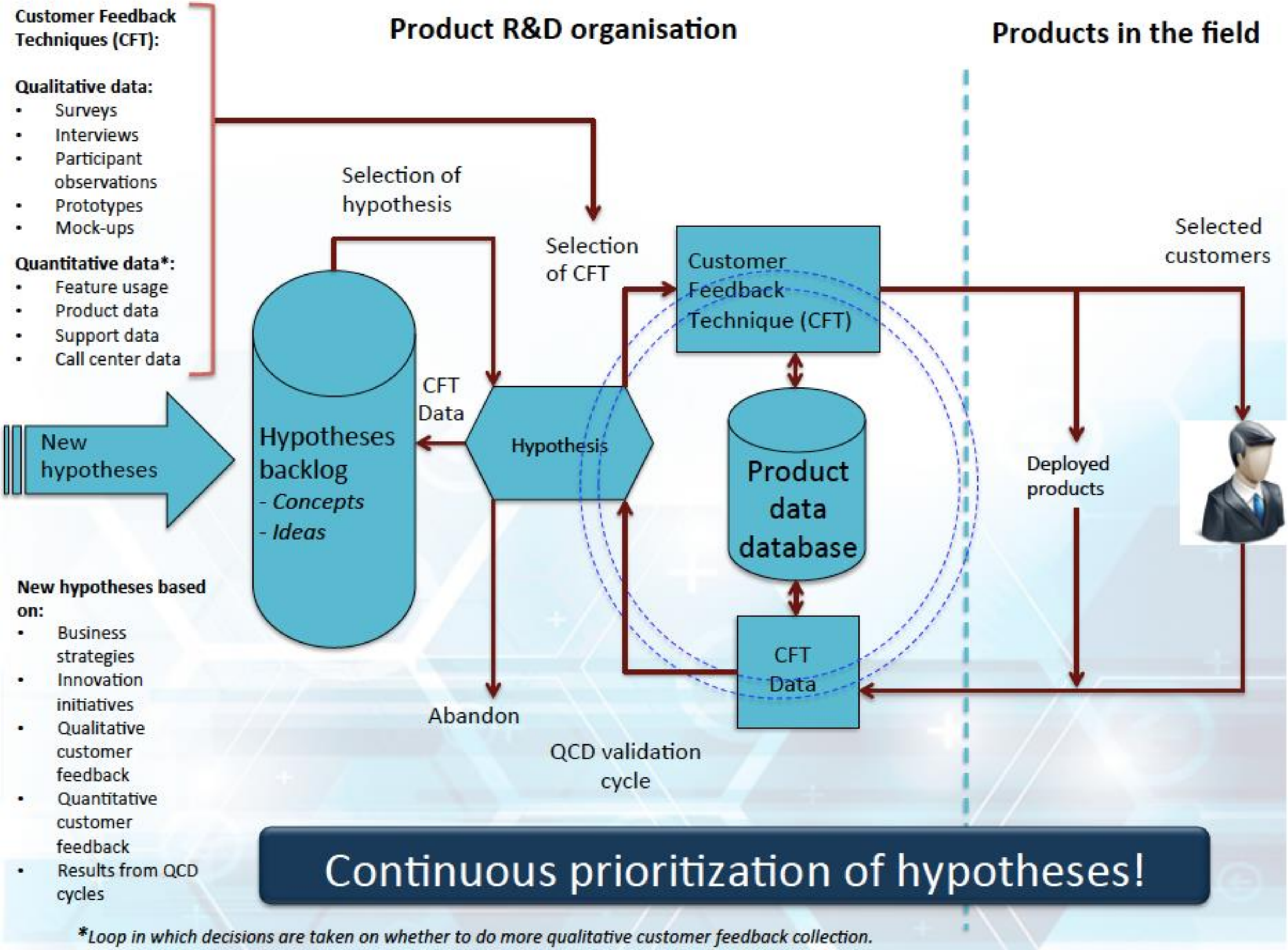

DOMO
DATA NEVER SLEEPS 6.0
How much data is generated every minute?
There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.
THE WEATHER CHANNEL RECEIVES 18,055,555 FORECAST REQUESTS
GIPHY SERVES UP 1,388,889 GIFS
NETFLIX USERS STREAM 97,222 HRS OF VIDEO
SNAPCHAT USERS SHARE 2,083,333 SNAPS
LINKEDIN GAINS 120+ NEW PROFESSIONALS
AMAZON SHIPS 1,111 PACKAGES
YOUTUBE USERS WATCH 4,333,560 VIDEOS
TUMBLR USERS PUBLISH 79,740 POSTS
TWITTER USERS SEND 473,400 TWEETS
REDDIT RECEIVES 1,944 NEW COMMENTS
12,986,111 TEXTS SENT
1.25 NEW BITCOIN ARE CREATED
SKYPE USERS MAKE 176,220 CALLS
2018 every MINUTE of the DAY
PRESENTED BY DOMO
GOOGLE CONDUCTS 3,877,140 SEARCHES
INSTAGRAM USERS POST 49,380 PHOTOS
TINDER USERS MATCH 6,940 TIMES
VENMO PROCESSES $68,493 PEER-TO-PEER TRANSACTIONS
UBER USERS TAKE 1,389 RIDES
SPOTIFY STREAMS OVER 750,000 SONGS
AMERICANS USE 3,138,420 GB OF INTERNET DATA


For TECH
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.
Learn more at domo.com
DATA NEVER SLEEPS 6.0
How much data is generated every minute?
There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.
2018 TECH EDITION
357 COMPUTERS ARE SOLD
63 WEBSITES ARE HACKED
2,833 SMARTPHONES ARE SOLD
THE INTERNET RECEIVES 3,138,420 GB OF TRAFFIC
176,220 CALLS ARE MADE ON SKYPE
4,756 IoT CONNECTIONS ARE MADE
2,472 MWH OF ELECTRICITY ARE USED FOR THE INTERNET
42.2 FITBITS ARE SOLD
159,362,760 EMAILS SENT PER MINUTE
12,986,111 TEXTS ARE SENT
331 TABLETS ARE SOLD

The cycle of innovation







Customer Feedback Techniques (CFT):
Qualitative data:
Surveys
Interviews
Participant observations
Prototypes
Mock-ups
Quantitative data*:
Feature usage
Product data
Support data
Call center data
Product R&D organisation
Products in the field
Selection of hypothesis
Selection of CFT
Customer Feedback Technique (CFT)
Selected customers
New hypotheses
Hypotheses backlog
- Concepts
- Ideas
CFT Data
Hypothesis
Product data database
Deployed products
New hypotheses based on:
Business strategies
Innovation initiatives
Qualitative customer feedback
Quantitative customer feedback
Results from QCD cycles
Abandon
CFT Data
QCD validation cycle
Continuous prioritization of hypotheses!
*Loop in which decisions are taken on whether to do more qualitative customer feedback collection.

Outline

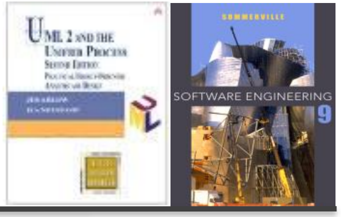
- Data management

- Data modelling
  - Entity relationship diagram (ERD)
- Relational database design
  - Normalization

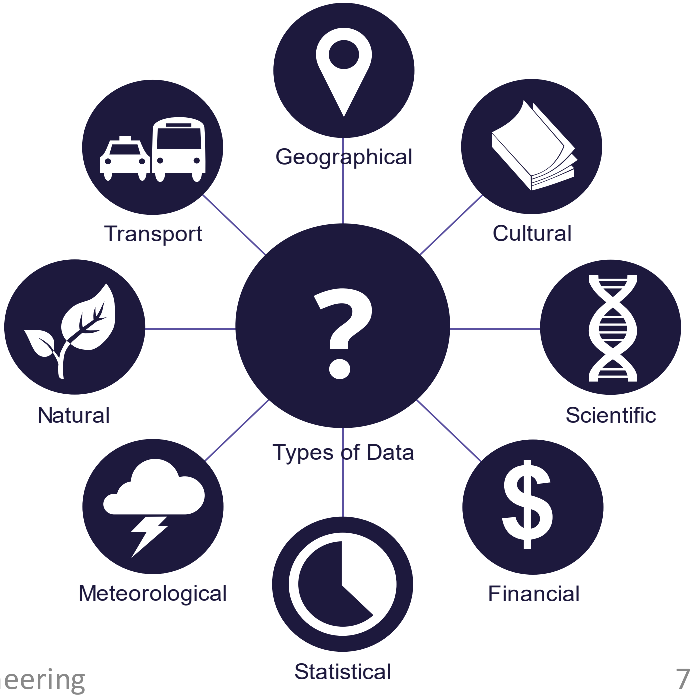
Data management

Lecture 5/Part 1

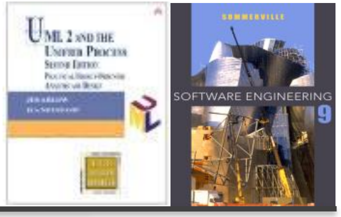
Data



- **Information** converted into binary digital form
  - information that has been translated into a form that is efficient for movement and processing
- It can be **created**, **processed**, **saved**, and **stored digitally**
  - This allows data to be transferred from one computer to another
- Digital information (i.e. data) in comparison to analog information **does not deteriorate** over time or lose quality after being used multiple times

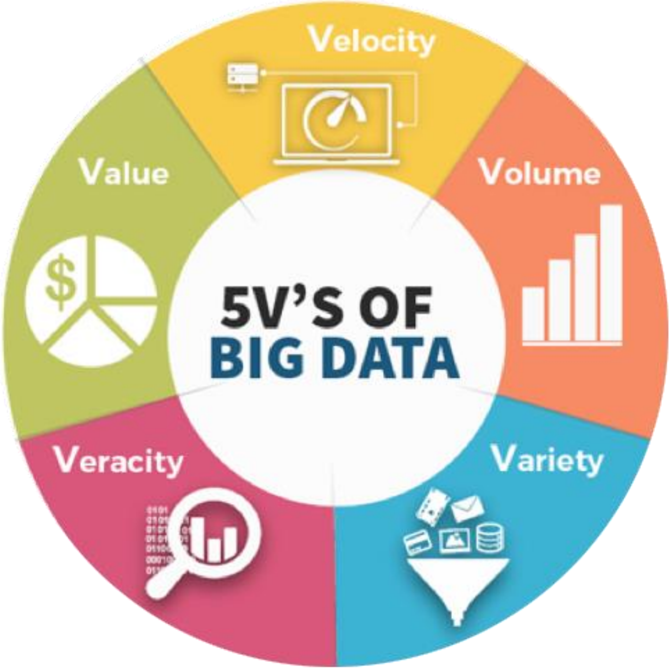


https://upload.wikimedia.org/wikipedia/commons/thumb/6/6d/Data_types_-_en.svg/300px-Data_types_-_en.svg.png



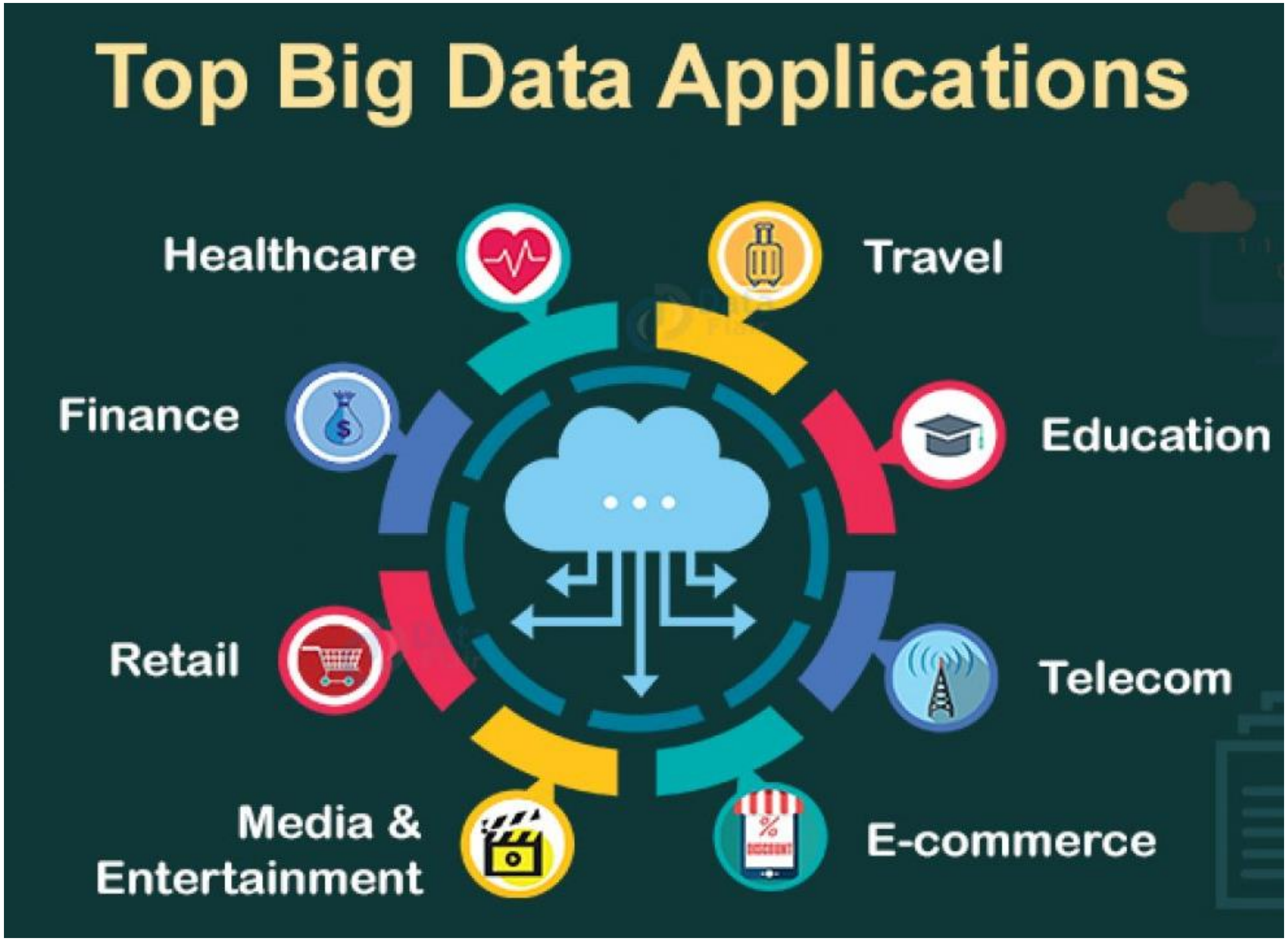
Big data

- ✧ **A collection of data** that is huge in size or growing exponentially with time
- ✧ It's difficult or impossible to process using traditional database and software tools
- ✧ Characteristics – 5 V's:
  - **Volume** – size of the data is enormous
  - **Variety** – various sources and format of data
  - **Variability** – data can be inconsistent and unpredictable
  - **Velocity** – data is generated very fast
  - **Veracity** – data is validated and verified



https://img.techentice.com/media/2019/04/5_V_Big_data-1.png

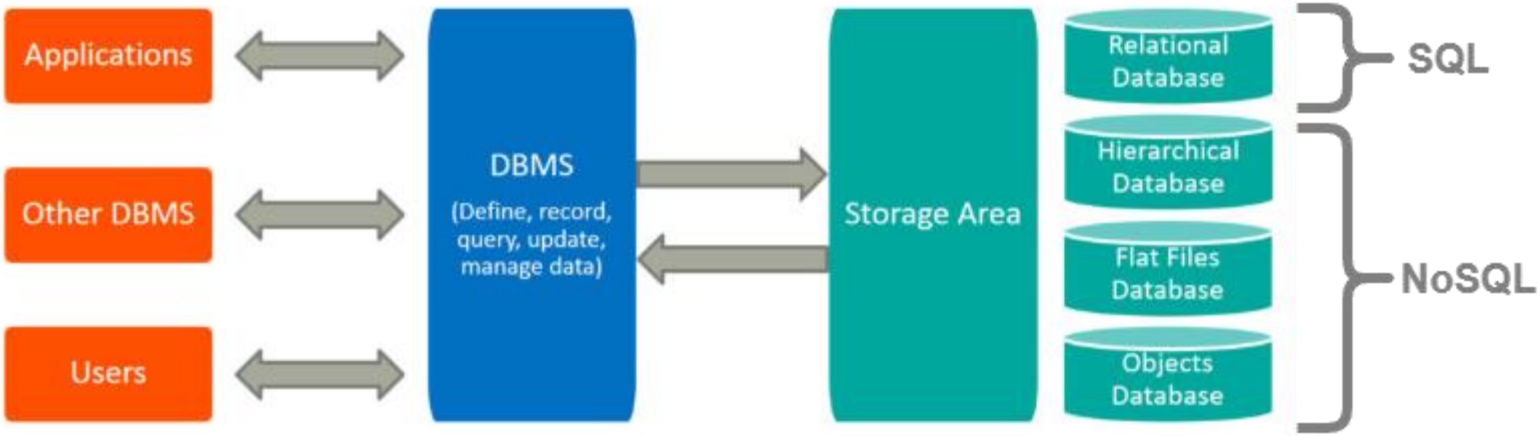
Big data



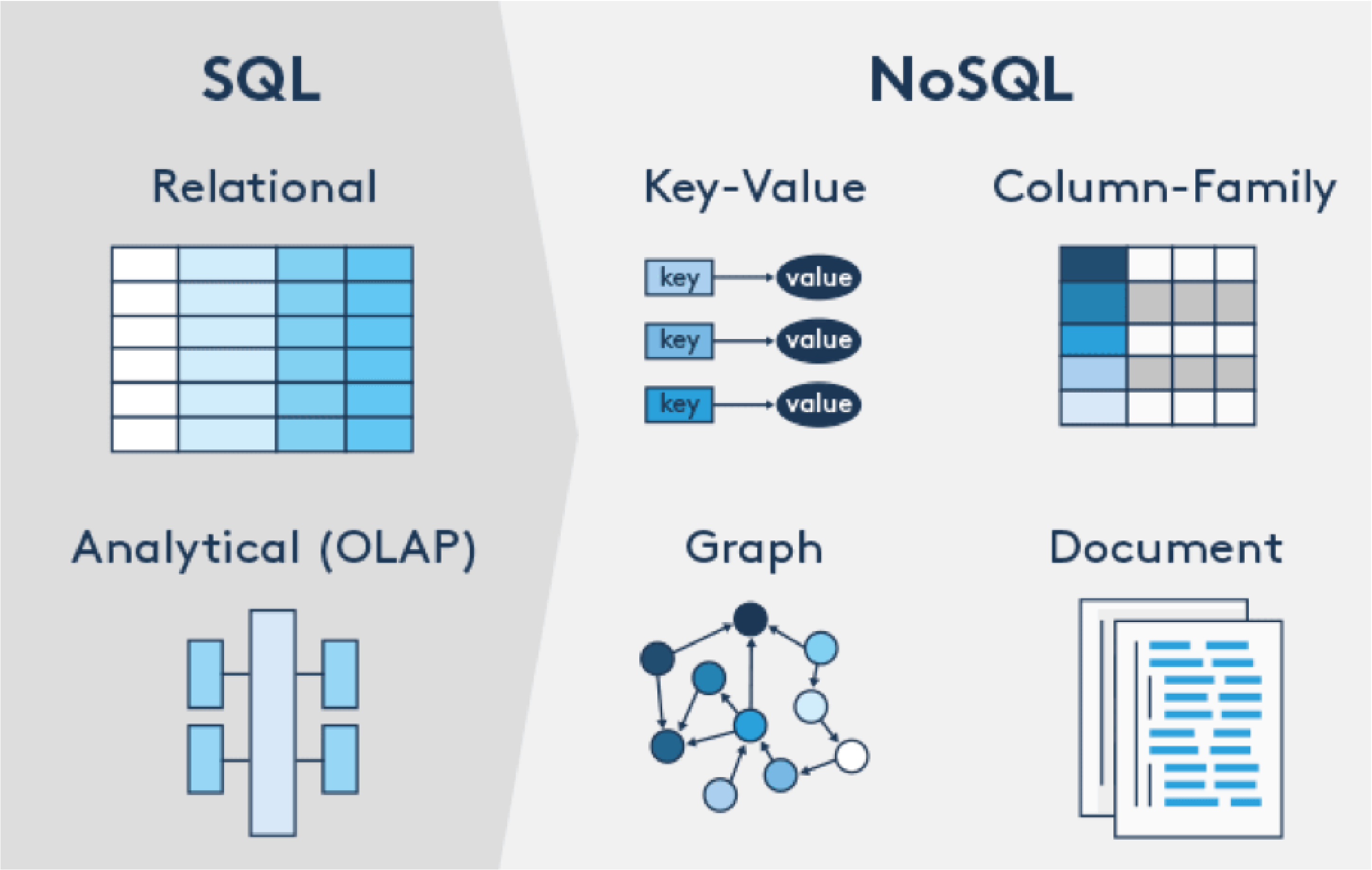
https://data-flair.training/blogs/wp-content/uploads/sites/2/2019/12/top-big-data-applications-2-1280x720.jpg

Database

- ✧ An **organised collection** of data
- ✧ Supports access, storage and manipulation of data
- ✧ Typically as rows and columns in a table
- ✧ Most used language is SQL (Structured Query Language)
- ✧ Controlled by database management system (DBMS)



Database

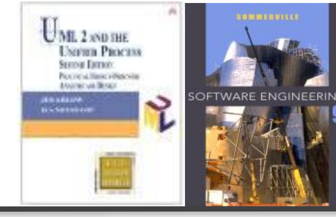


https://asesoftware.com/site/wp-content/uploads/2019/06/asesoftware-sql-nosql.png

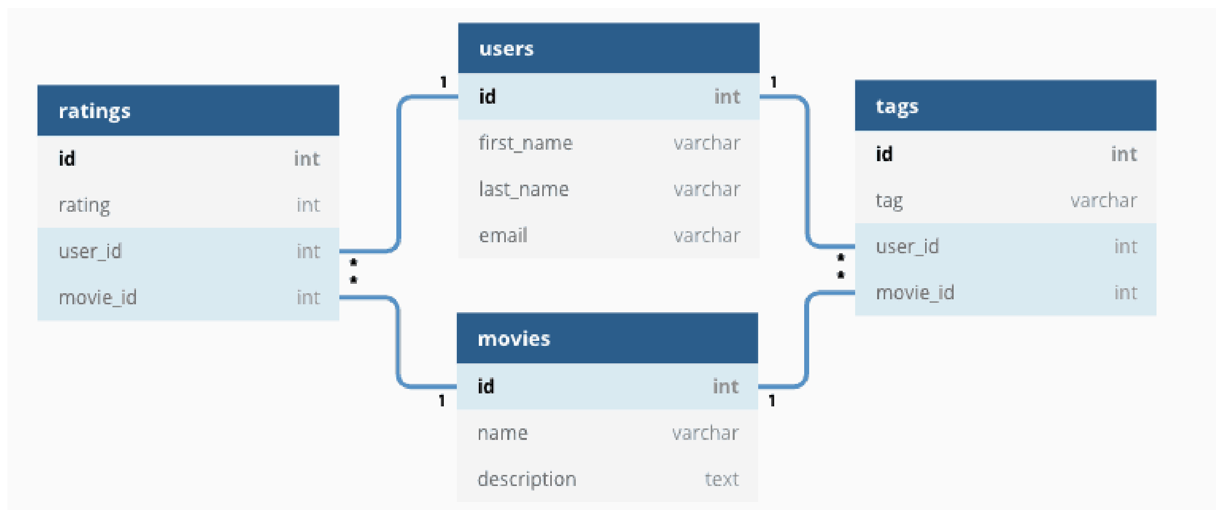
Relational database

- Stores data as a series of two-dimensional tables with rows and columns with pre-defined relationships between data
- Relationship between tables and field types is called a schema. The schema must be clearly defined before any information can be added
- Each table has its own columns, and every row in a table has the same set of columns and a unique ID called the key
- For querying uses structured query language (SQL)

Relational database

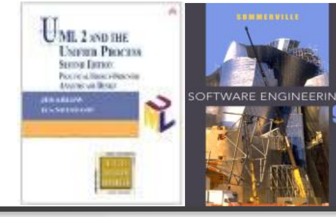


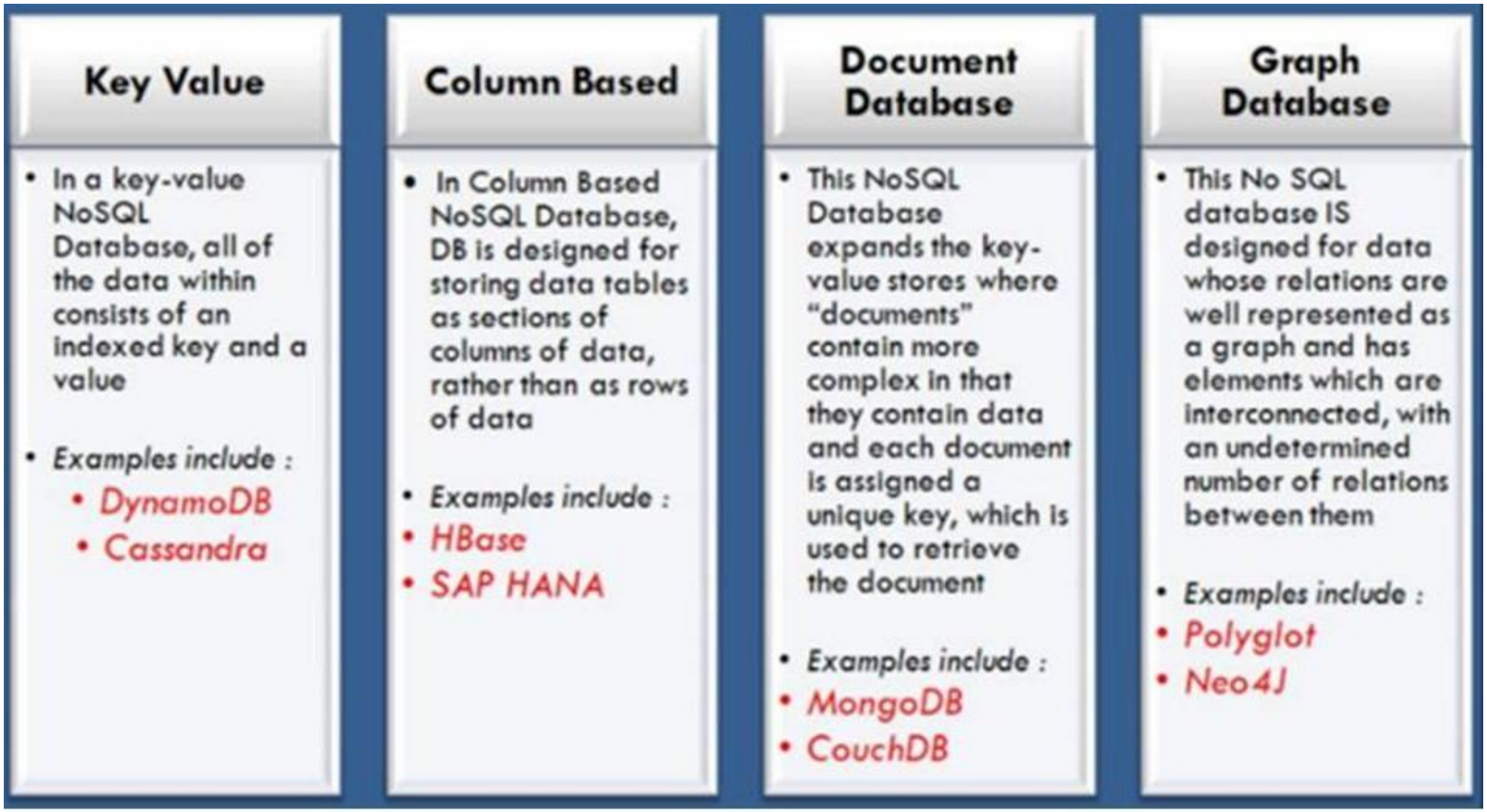
- Used in:
  - Transaction-oriented systems
  - Accounting software
  - Management tools
- Examples include:
  - PostgreSQL
  - MySQL



https://assets-global.website-files.com/5debb9b4f88fbc3f702d579e/5e3c1a71724a38245aa43b02_99bf70d46cc247be878de9d3a88f0c44.png

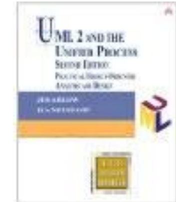
NoSQL databases





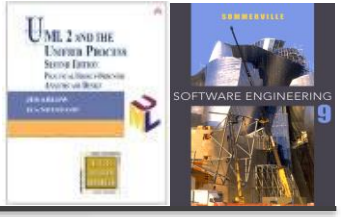
https://cdn.app.compendium.com/uploads/user/e7c690e8-6ff9-102a-ac6d-e4aebca50425/f4a5b21d-66fa-4885-92bf-c4e81c06d916/Image/b64beeec82803ee769516c731c52762e/2bigdata04.jpg

Key-value database

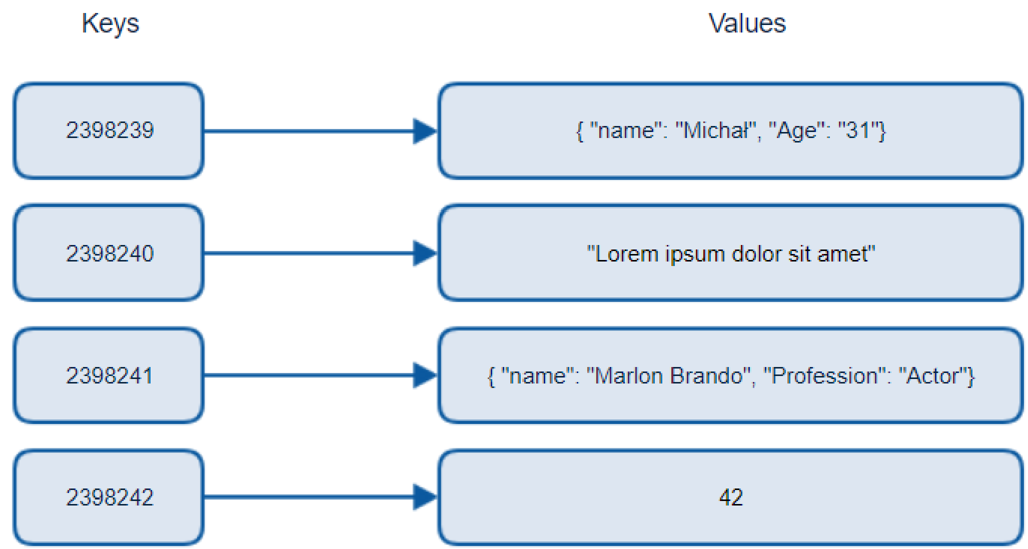


- Nonrelational database that uses a key-value method to store data
- Stores data as a collection of key-value pairs in which a key serves as a unique identifier
- A data structure more commonly known today as a dictionary or hash table
- Doesn’t have a query language; it provides a simple way to store, query and update data using get, put and delete commands – not optimized for querying by value

Key-value database



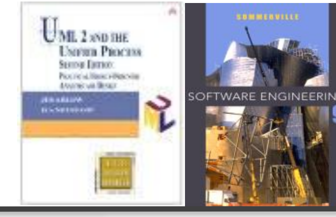
- Used in:
  - Shopping cart in e-shops
  - Caching
  - Multi-player games



https://www.michalbialecki.com/wp-content/uploads/2018/03/cosmos-db-key-value-schema.png

- Examples include:
  - Redis
  - Apache Cassandra
  - Amazon Dynamo DB
  - Microsoft Azure Cosmos DB

Column-family database



- Nonrelational database that stores data into rows and columns, conceptually similar to a relational database
- Columns are divided into groups known as column families. Each column family holds a set of columns that are logically related together and are typically retrieved or manipulated as a unit
- Suited for storing enormous, structured, volatile data because each row is not required to have the same columns

Column-family database

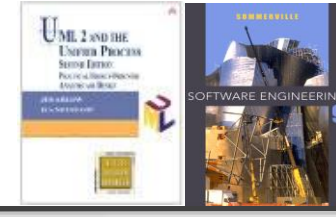
- ✧ Used in:
  - Internet of Things
  - Security analytics
  - Stock market
  - Bioinformatics
- ✧ Examples include
  - HBase
  - Apache Cassandra

CustomerID	Column Family: Identity
001	**First name:** Mu Bae **Last name:** Min
002	**First name:** Francisco **Last name:** Vila Nova **Suffix:** Jr.
003	**First name:** Lena **Last name:** Adamcyz **Title:** Dr.

CustomerID	Column Family: Contact Info
001	**Phone number:** 555-0100 **Email:** someone@example.com
002	**Email:** vilanova@contoso.com
003	**Phone number:** 555-0120

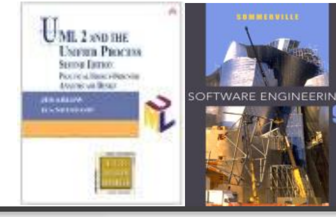
https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/images/column-family.png

Document database



- Nonrelational database that stores a collection of named fields and data (known as documents)
- Stored data can be encoded in different formats, e.g. XML, YAML, JSON, plain text
- Does not require that all documents have the same structure – provides flexibility for storing different data
- Allows querying and filtering documents by value of one or more fields and in-place modifying values without rewriting the whole document

Document database



- Used in:
  - User profiles
  - Real-time big data
  - Content management

Document 1

```
{
  "id": "1",
  “name": "John Smith",
  "isActive": true,
  "dob": "1964-30-08"
}
```

Document 2

```
{
  "id": "2",
  “fullName": "Sarah Jones",
  “isActive": false,
  "dob": "2002-02-18"
}
```

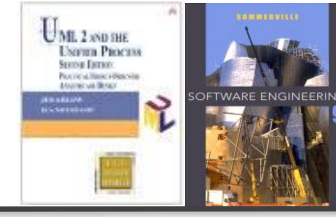
https://lennilobel.files.wordpress.com/2015/07/i4.png

Document 3

```
{
  "id": "3",
  “fullName":
  {
    "first": "Adam",
    "last": "Stark"
  },
  "isActive": true,
  "dob": "2015-04-19"
}
```

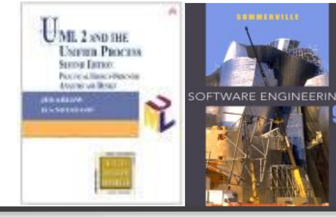
- Examples include
  - MongoDB
  - Google Cloud Firestore
  - Microsoft Azure Cosmos DB

Graph database



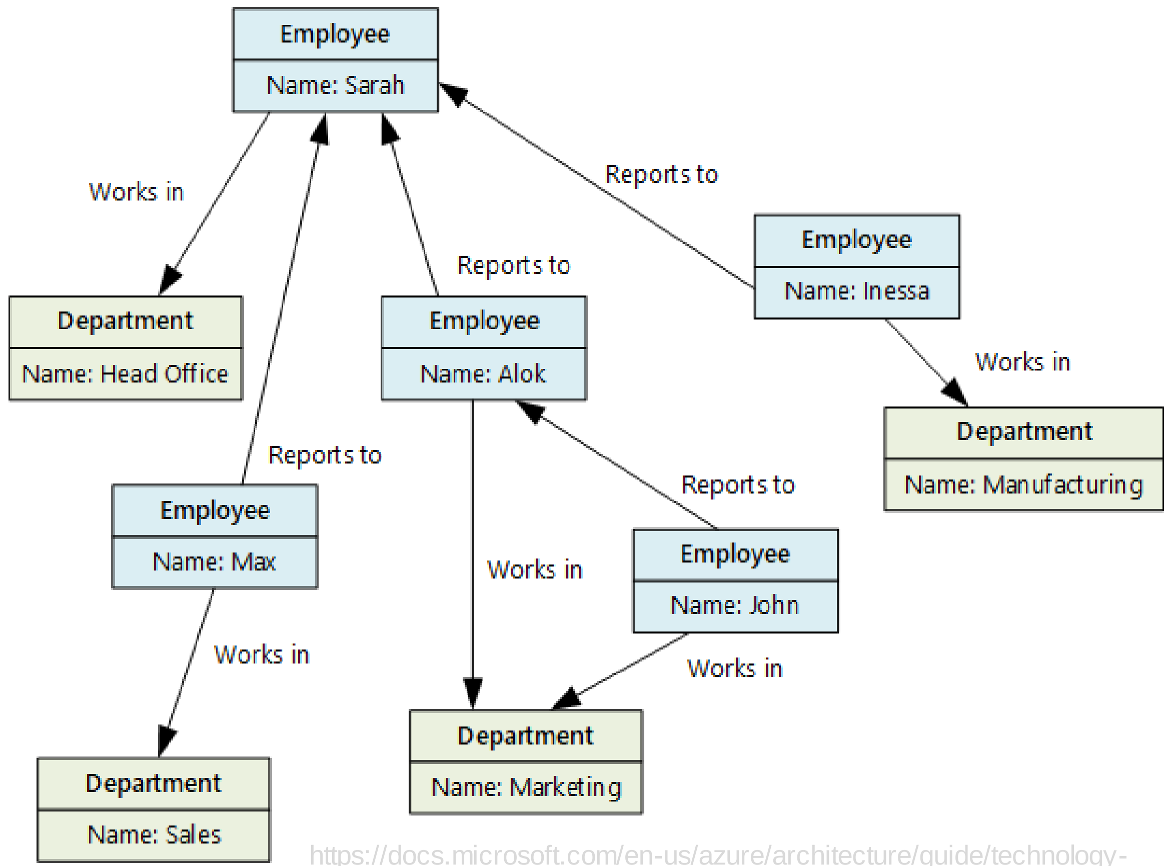
- ✧ Nonrelational database that stores two types of information, nodes and edges
- ✧ Nodes typically store information about people, places, and things while edges store information about the relationships between the nodes
- ✧ The relationships allow data in the database to be linked together directly and retrieved with one operation
- ✧ Provides a query language that can be used to traverse a network of relationships efficiently

Graph database



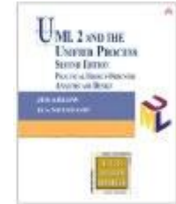
✧ Used in:
- Social networks
- Fraud detection
- Recommendation engines

✧ Examples include:
- Neo4j
- Amazon Neptune
- Apache Giraph



https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/images/graph.png

Data management



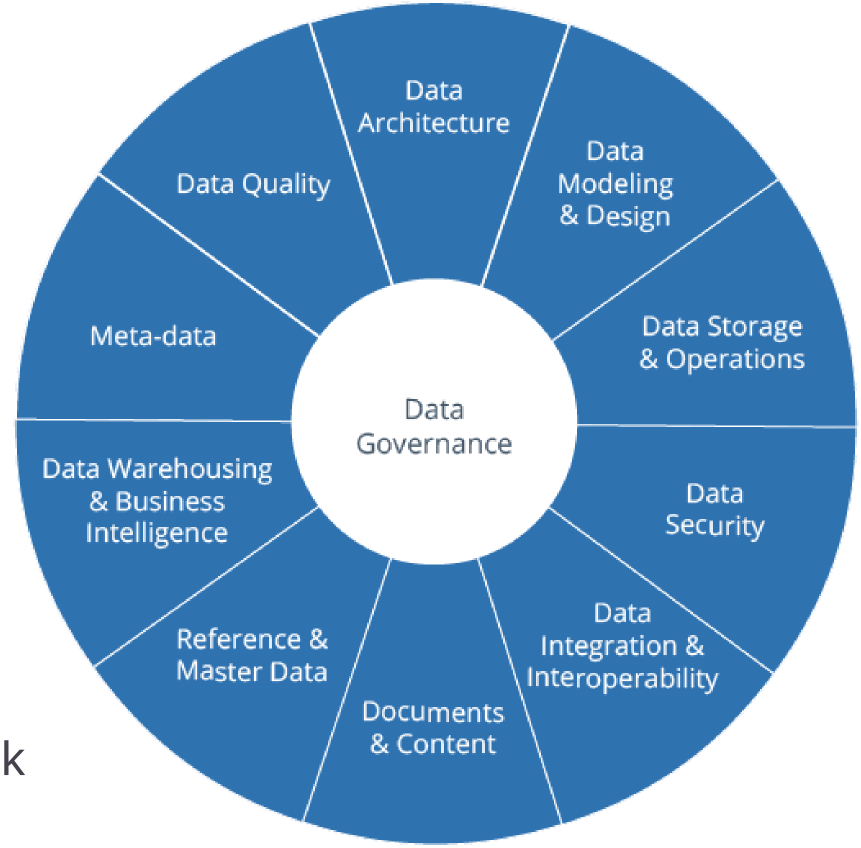
- Administrative process that includes **acquiring**, **validating**, **storing**, **protecting**, and **processing** required data to ensure the accessibility, reliability, and timeliness of the data for its users
- Encompasses the **entire lifecycle** of a data asset, from the very initial creation of the data to the final retirement of the data
- Some companies are good at collecting data, but they are not managing it well enough to **turn raw data into value**

Data governance

- A set of principles and practices that ensure **high quality** through the **complete lifecycle** of the data
- Includes the **people, processes and technologies** needed to manage and protect the company's data assets in order to guarantee generally **understandable, correct, complete, trustworthy, secure and discoverable** corporate data
- Most relevant in large enterprises to ensure security, compliance and improve business performance

Key goals of data governance

- Minimize risks
- Establish internal rules for data use
- Implement compliance requirements
- Improve internal and external communication
- Increase the value of data
- Facilitate the administration of the above
- Reduce costs
- Help to ensure the continued existence of the company through risk management and optimization



https://bi-survey.com/wp-content/uploads/2017/11/Data-Governance-topics.png

Data Lifecycle

- A high-level overview of the stages involved in successful management and preservation of data for use and reuse

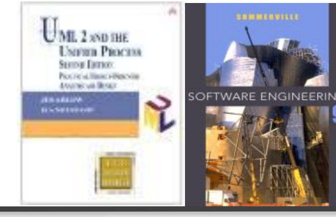
- **Data Capture / Creation**
- **Data Maintenance**
- **Data Usage**
- **Data Publication**
- **Data Archiving**
- **Data Purging / Destruction**

https://www.nexor.com/wp/wp-content/uploads/2017/03/Enabling-Secure-Information-Exchange-in-Cloud-environments-white-paper-diagram-1.png

Data modelling

Lecture 5/Part 2

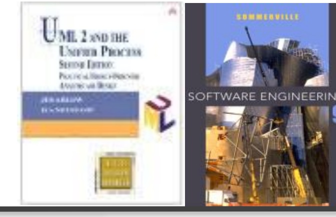
Data modeling



- ✧ Defines static data structure, relationships and attributes
- ✧ Complementary to the behavior model in structured analysis; models information not covered by DFDs
- ✧ More stable and essential information comparing to DFD

- ✧ **Entity-Relationship modeling**
  - Identify system entities – both abstract (lecture) and concrete (student)
  - For each entity examine – the purpose of the entity, its constituents (attributes) and relationships among entities
  - Check model consistency and include data details

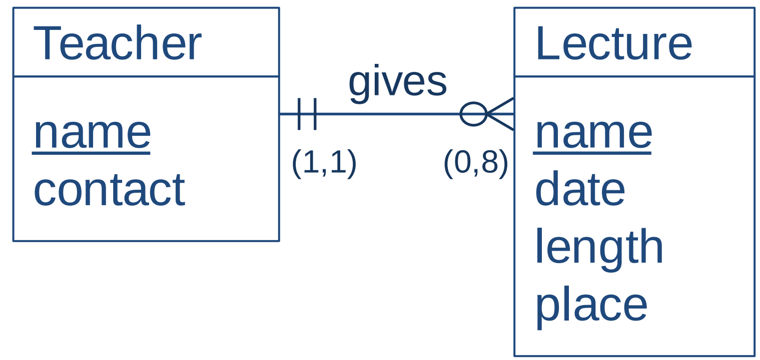
Entity Relationship Diagram (ERD)



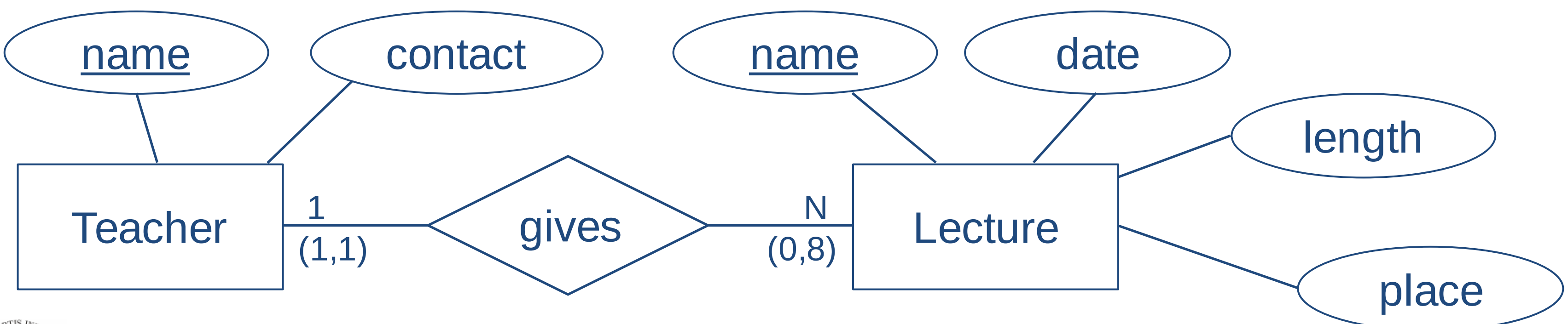
- ✧ **Entities** and their types
- ✧ **Relationships** and their types
- ✧ **Attributes** and their domains

<u>Not a UML diagram!</u>

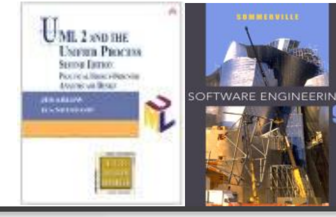
Crow's Foot notation
(implementation level descript.)



Chen's notation
(concept level description)



Entities and Entity types



✧ An **Entity** is anything about which we want to store data

- Identifiable – entities can be distinguished by their identity
- Needed – has significant role in the designed system
- Described by attributes shared by all entities of the same type

✧ An **Entity set** is a set of entities of the same **Entity type**.

Entity	Entity type
You	Student
Your neighbor	Student
Me	Teacher
This PB007 lecture	Lecture

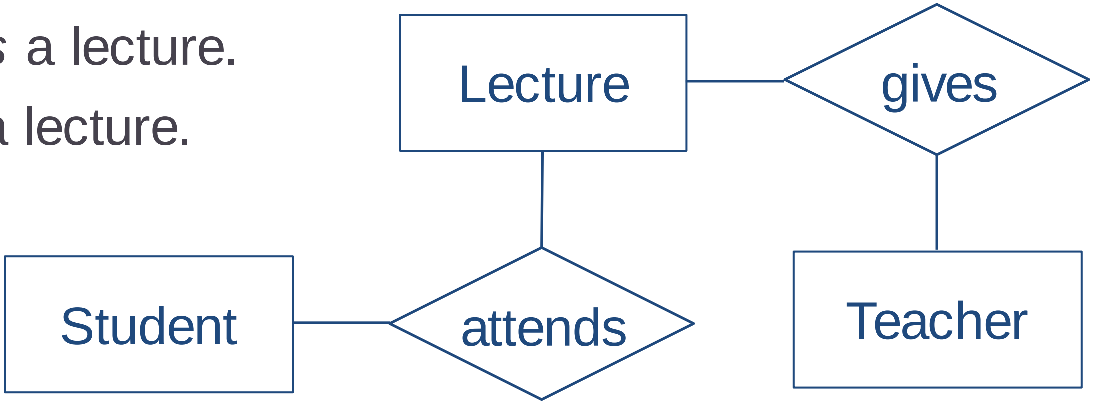
Student

Teacher

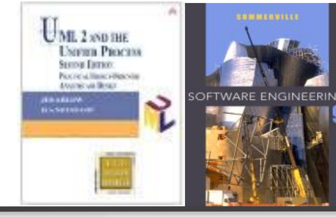
Lecture

Relationships and Relationship types

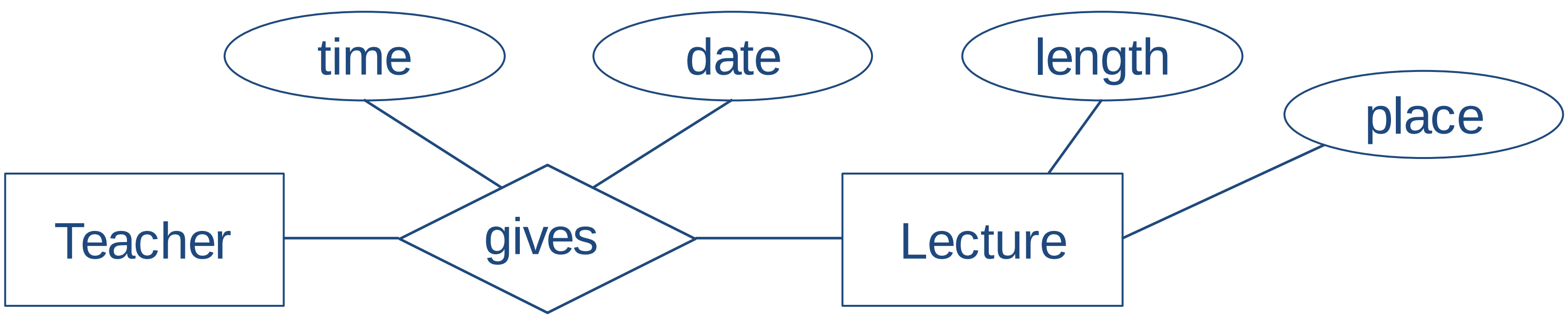
- Entities take part in **Relationships** (among possibly more than two entities), that can often be identified from verbs or verb phrases.
  - You are *attending* this PB007 lecture.
  - I am *giving* this PB007 lecture.
- A **Relationship set** is a set of relationships of the same **Relationship type**.
  - A student *attends* a lecture.
  - A teacher *gives* a lecture.



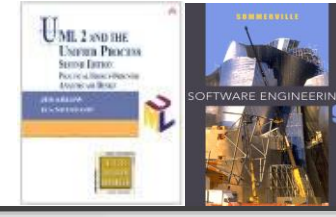
Attributes and Attribute domains



- An **Attribute** is a fact, aspect, property, or detail about either an entity type or a relationship type.
  - E.g. a lecture might have attributes: time, date, length, place.
- An **Attribute type** is a type domain of the attribute. If the domain is complex (domain of an attribute *address*), the attribute may be an entity type instead.

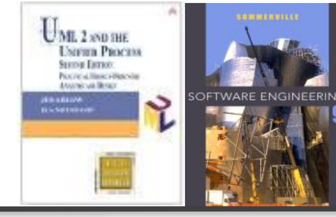


Attributes or entities?



- To decide whether a concept be modeled as an attribute or an entity type:
  - Do we wish to store any information about this concept (other than an identifying name)?
  - Is it single-valued?
  - E.g. *objectives* of a *course* – are they more than one? If just one, how complex information do we want to store about it?
- General guidelines:
  - Entities can have attributes but attributes have no smaller parts.
  - Entities can have relationships between them, but an attribute belongs to a single entity.

Relationship-type degree

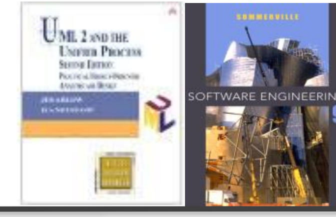


Every manager leads exactly one department.
Every department is led by exactly one manager.

Every edition plan contains one or more book titles.
Every book title is part of exactly one edition plan.

Every producer produces one or more products.
Every product is produced by one or more producers.

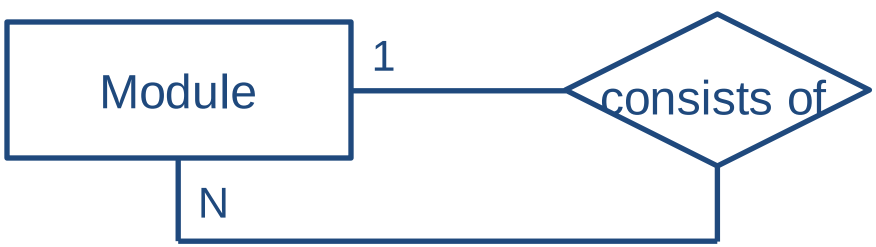
Relationship-type degree



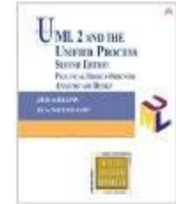
Mandatory relationship

Optional relationship

Recursive relationship



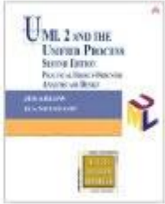
Cardinality ratio

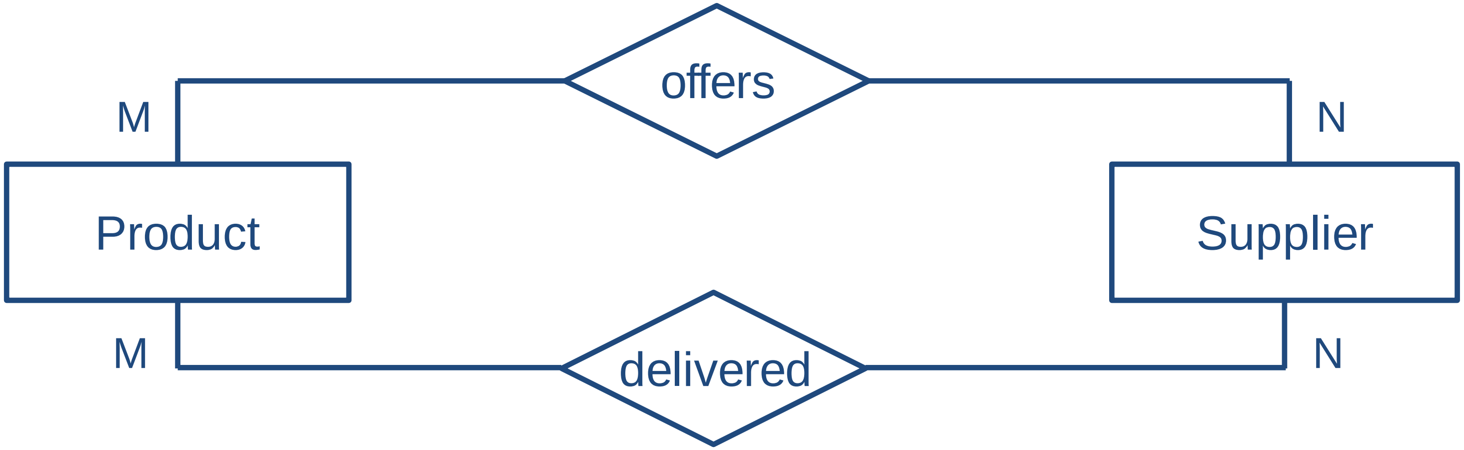


- **Cardinality ratio** of a relationship type describes the number of entities that can participate in the relationship.

- One to one 1:1
  - Each lecturer has a unique office.
- One to many 1:N
  - A lecturer may tutor many students, but each student has just one tutor.
- Many to many M:N
  - Each student takes several modules, and each module is taken by several students.

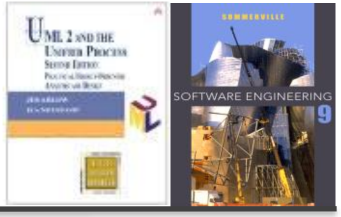
More relationships between two entities

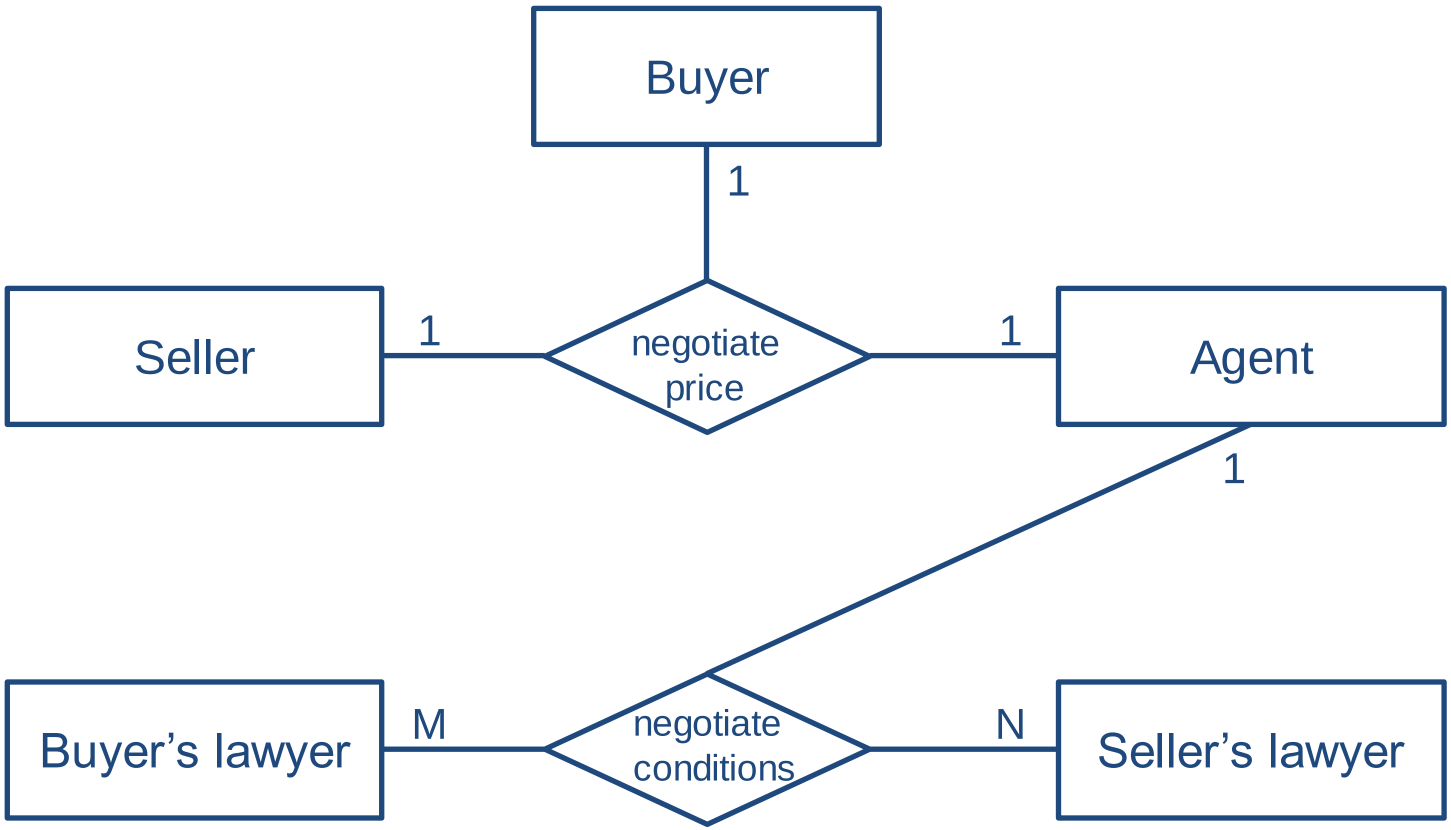




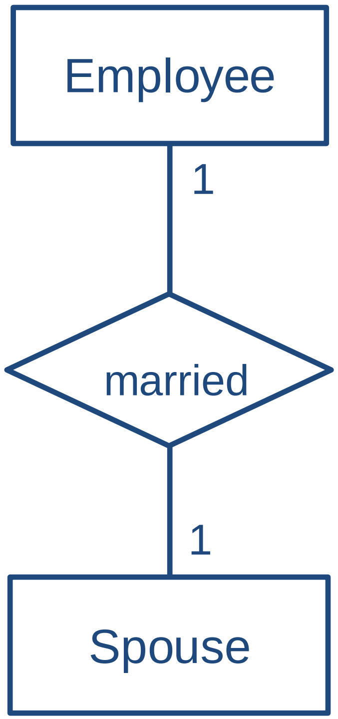
- Relationship *offers* has attributes:
  - *payment conditions, due date*.
- Relationship *delivered* has attributes:
  - *delivery note details*.

Relationships among more than two entities





Removal of unneeded (redundant) entities

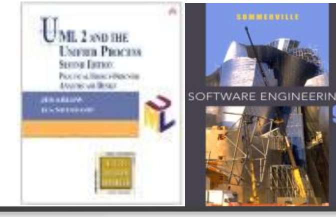


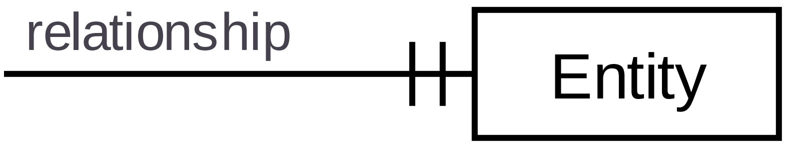
The *Spouse* entity is better suited as *Employee*’s attribute.

Relational Database Design

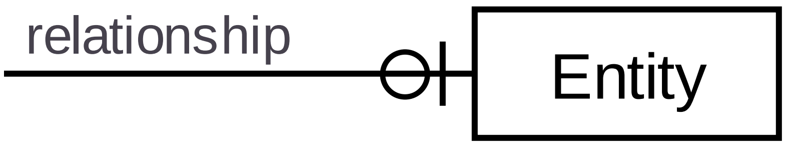
Lecture 5/Part 3

Crow's Foot notation

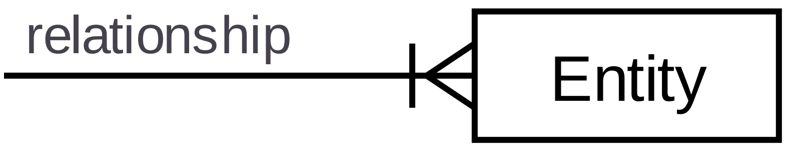




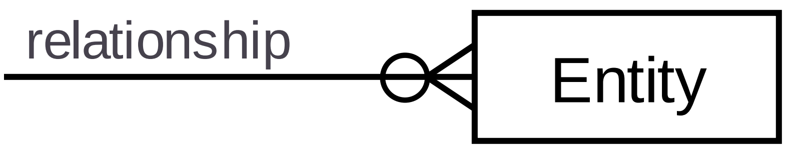
Exactly one occurrence



None or one occurrence

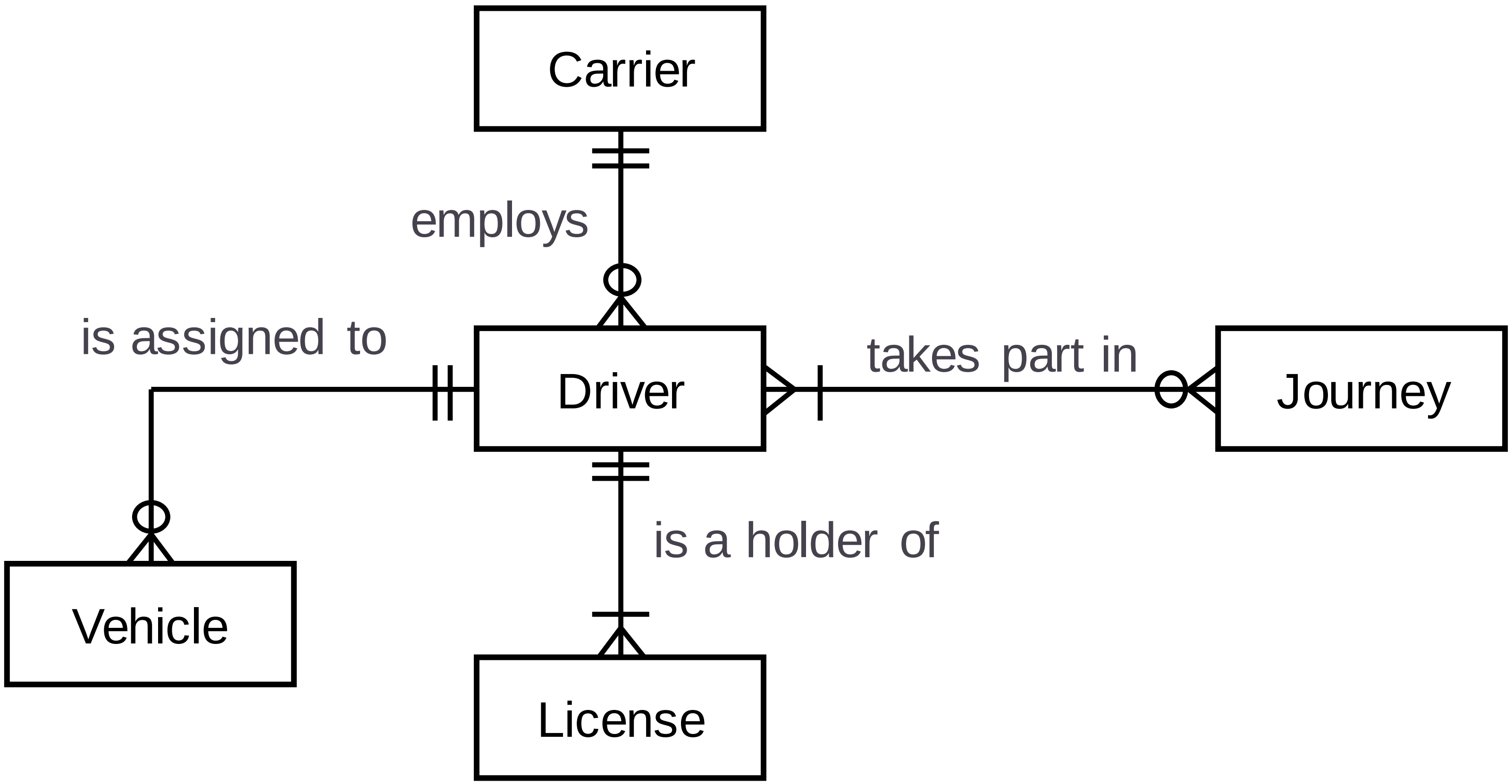


One or more occurrence

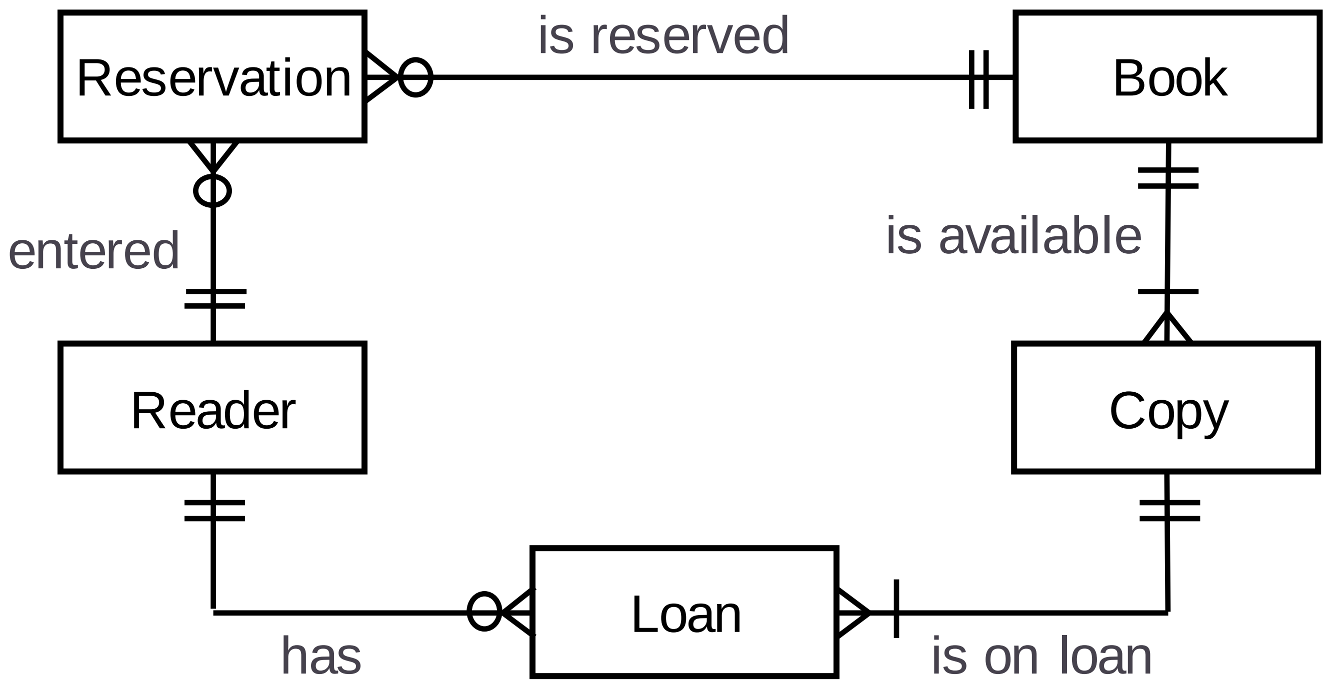


None or more occurrences

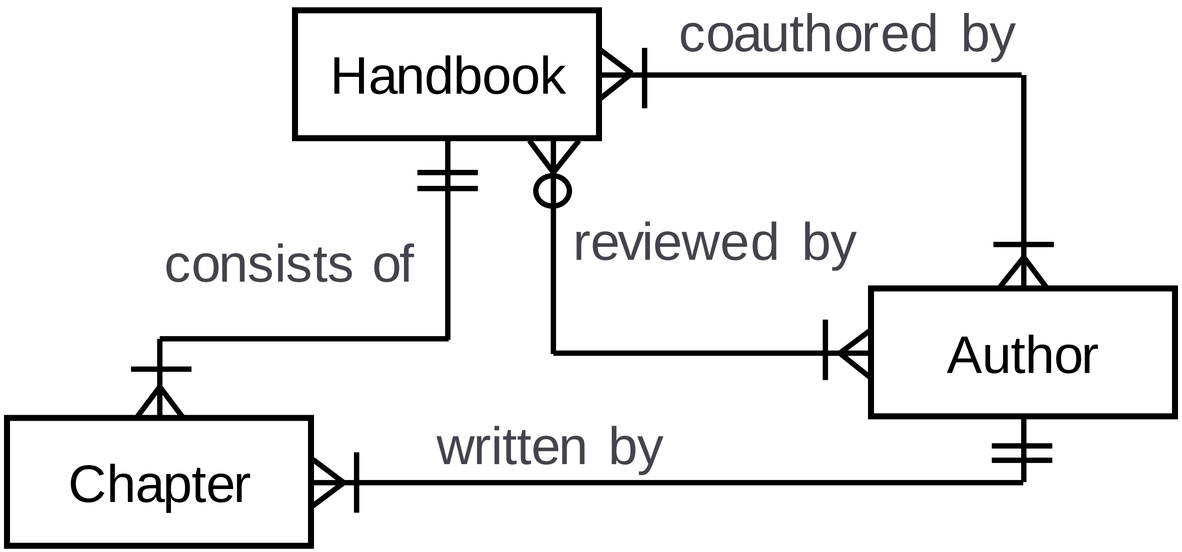
ERD example – Transport



ERD example – Library



ERD example – Book editing



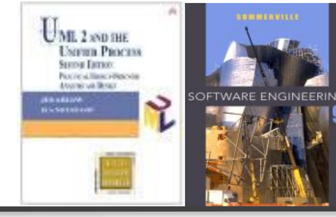
Relational database design based on ERDs

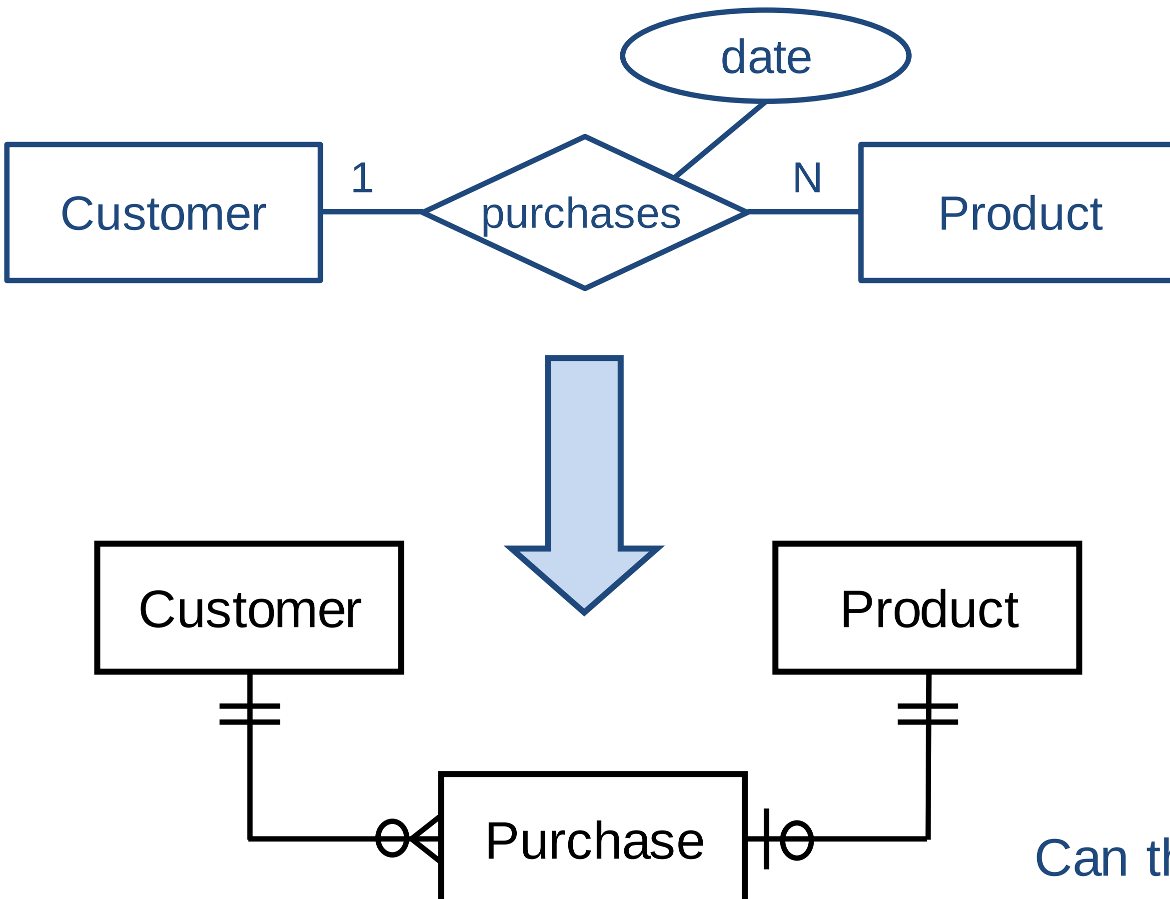
✧ Entity-relationship modeling is a first step towards database design.

Database design process:

1. **Determine the purpose of the database.**
2. **Find and organize the information required** - Create ERD model of the system. Each entity type becomes a table, attribute becomes a column, entity becomes a row in the table. Handle relationships with attributes, and M:N relationships.

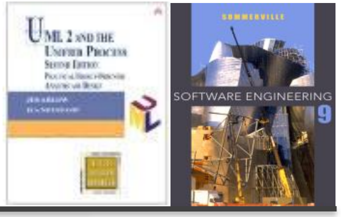
Relationships to entities

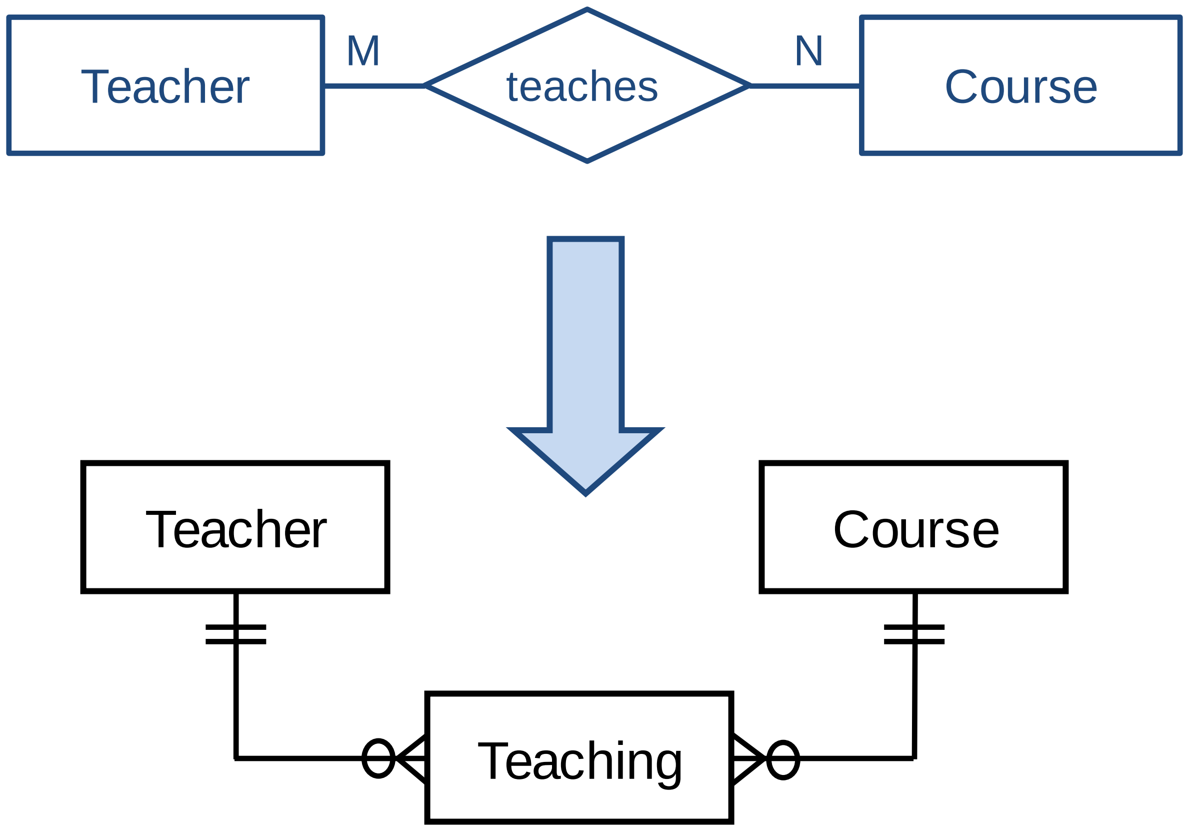




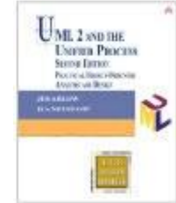
Can the purchase entity be omitted?

M:N relationships





Database design process (continued)

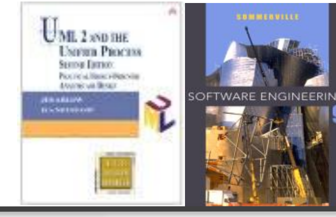


3. **Specify primary keys** - Choose each table's primary key. The primary key is a column that is used to uniquely identify each row. An example might be Product ID or Order ID.

4. **Apply the normalization rules** - Apply the data normalization rules to see if tables are structured correctly. Make adjustments to the tables.

5. **Refine the design** - Analyze the design for errors. Create tables and add a few records of sample data. Check if results come from the tables as expected. Make adjustments to the design, as needed.

Entities and keys



✧ Superkey

- A set of attributes that **uniquely identifies** each entity.

✧ Candidate key

- A **non-redundant** superkey, i.e. all items of a candidate key are necessary to identify an entity, no key attribute can be removed.
- There can be more combinations of entity attributes that can be used as candidate keys.

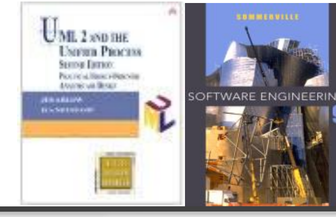
✧ Primary key

- The **selected candidate key**, marked with # symbol.

✧ Foreign key

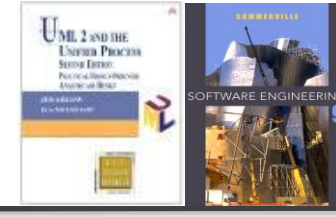
- A set of attributes in one entity that **uniquely identifies** (i.e. is a primary key in) **another entity**.

Data normalization goals by E.F. Codd

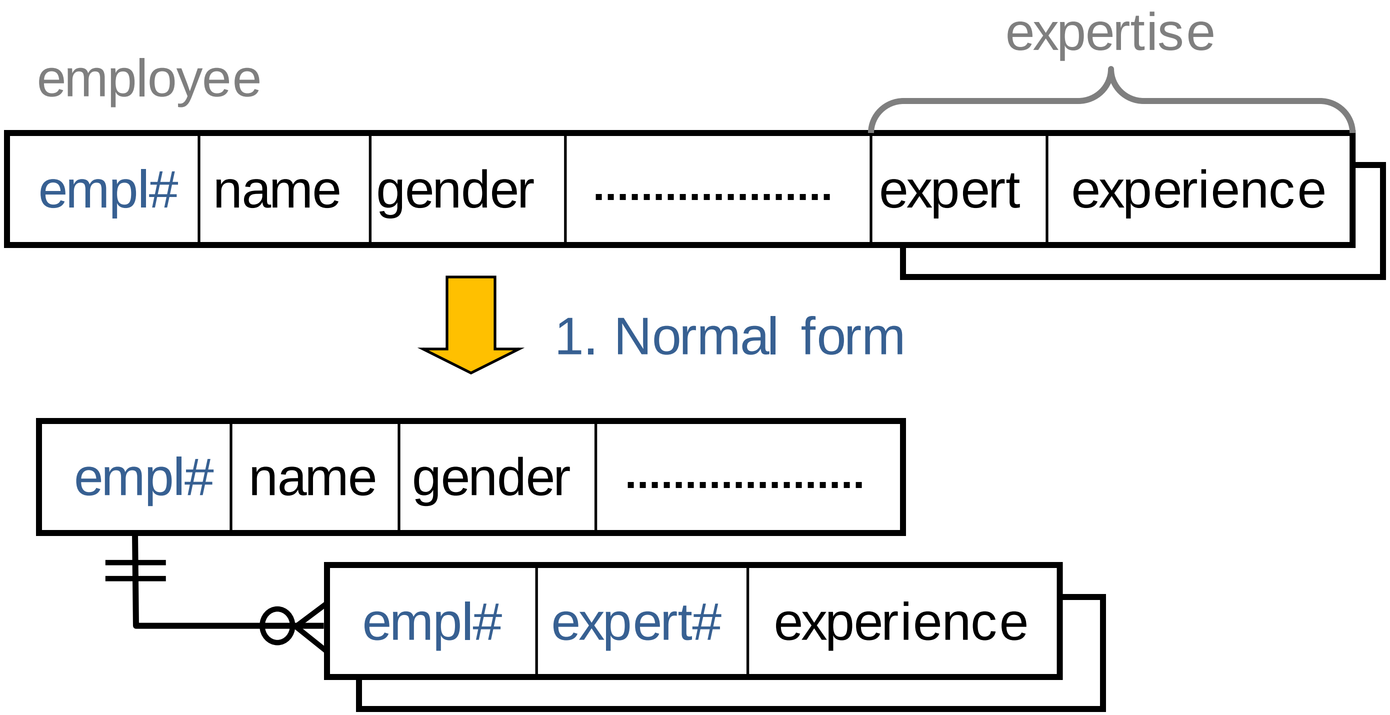


- ✧ Minimize **redundancy** and **dependency**
  - Minimize redesign when extending database structure
  - Make the data model more informative to users
- ✧ Free the database of modification anomalies
  - **Update anomaly** – the same information expressed on multiple rows → update resulting in logical inconsistencies.
  - **Insertion anomaly** – certain facts cannot be recorded, because of their binding with another information into one record.
  - **Deletion anomaly** – deletion of data representing certain facts necessitating deletion of unrelated data.
- ✧ Avoid bias towards any particular **pattern of querying**

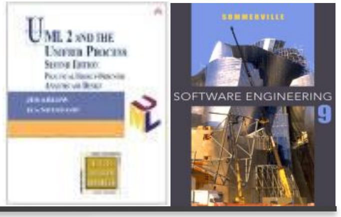
1. Normal form – no repeating groups

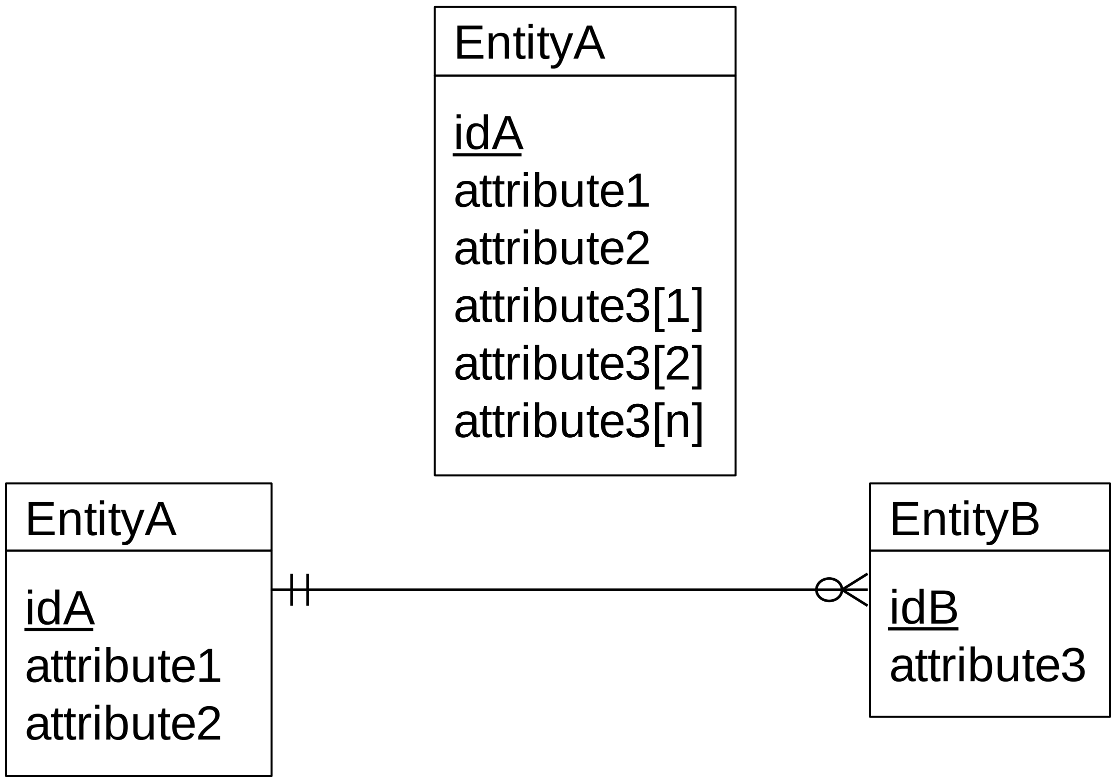


Def.1NF: A relation is in 1NF if the domain of each attribute contains only **atomic values**, and the value of each attribute contains only a **single value** from that domain.

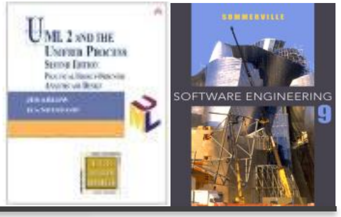


1. Normal form – normalization example



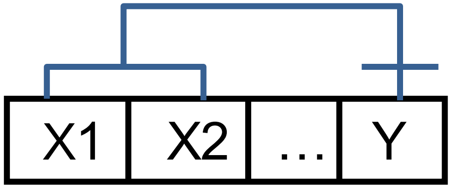


Functional dependency

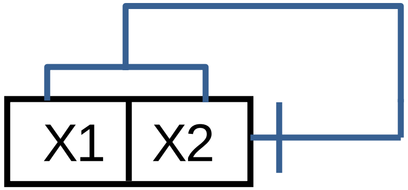


✧ Functional dependency

- In a given table, an attribute Y is said to have a functional dependency on a set of attributes X if and only if each X value is associated with precisely one Y value.

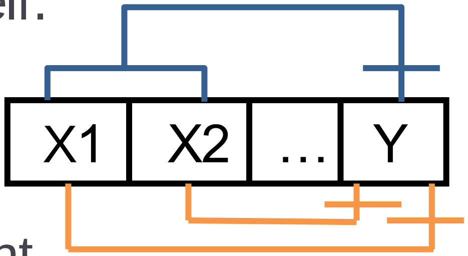


✧ Trivial functional dependency



- A trivial functional dependency is a functional dependency of an attribute on a superset of itself.

✧ Full functional dependency



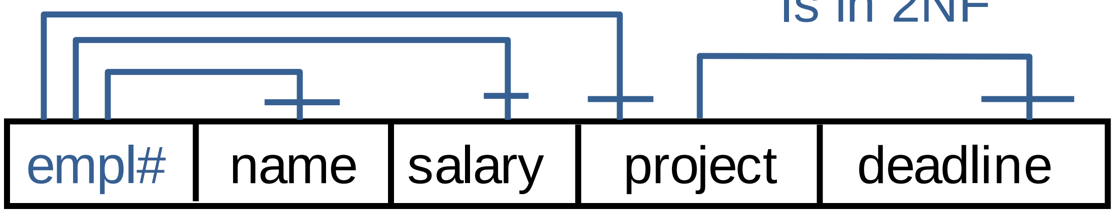
- An attribute is fully functionally dependent on a set of attributes X if it is: functionally dependent on X, and not functionally dependent on any proper subset of X.

2. Normal form – no partial dependency

Def. 2NF: In 1NF and no non-prime attribute in the table is functionally dependent on a proper subset of any candidate key.

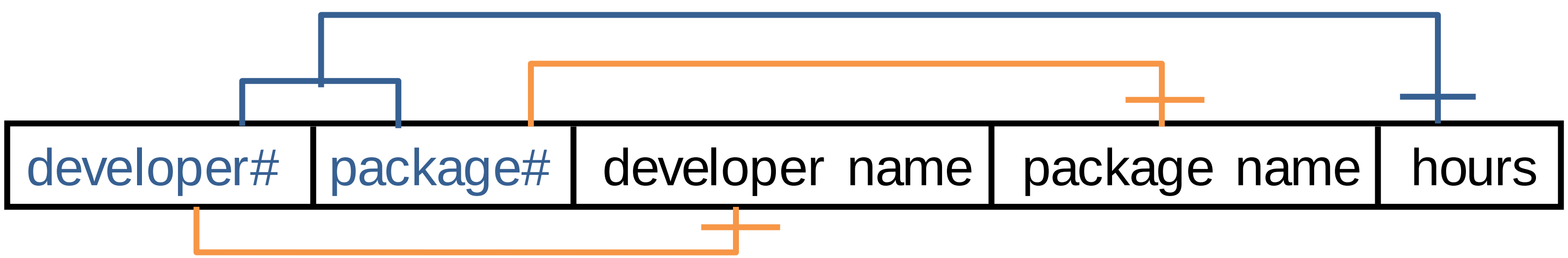
Example EMPLOYEE

Is in 2NF



Example PROGRAMING

Not in 2NF



What anomalies can you identify in this example?

2. Normal form – no partial dependency

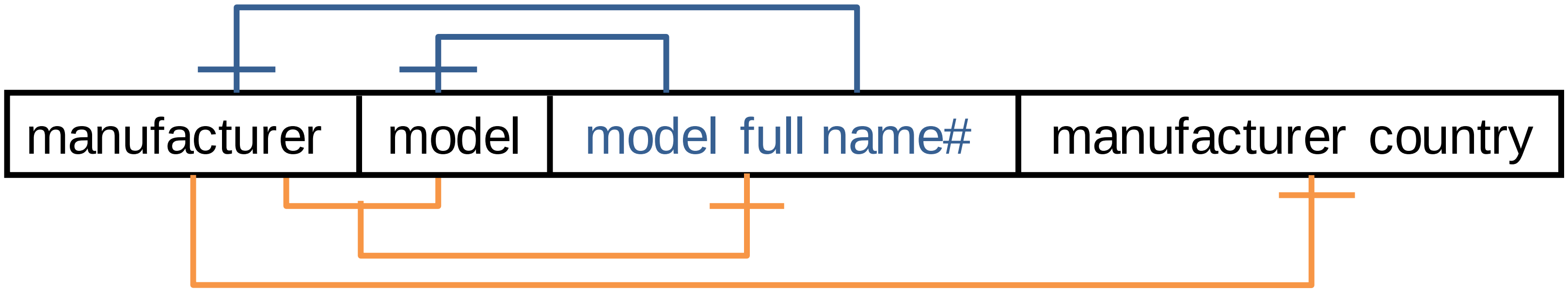
not part of any candidate key

Def. 2NF: In 1NF and no non-prime attribute in the table is functionally dependent on a proper subset of any candidate key.

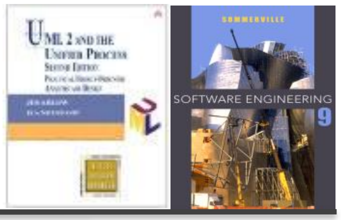
- Does the “candidate key” part of the definition make difference?
- When there is only one-item primary key, is 2NF guaranteed?

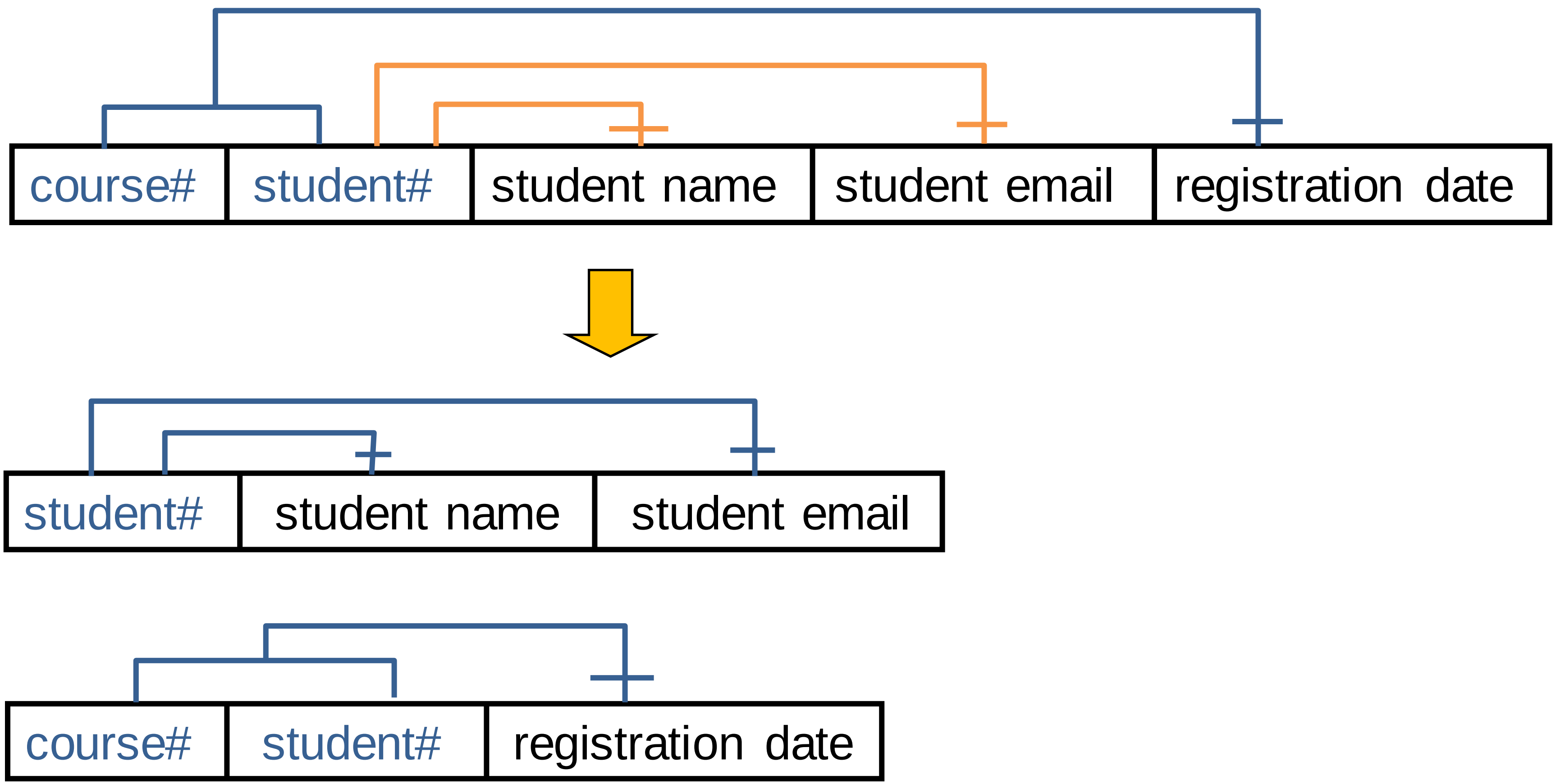
Example DISHWASHER MODELS

Not in 2NF



2. Normal form – normalization example

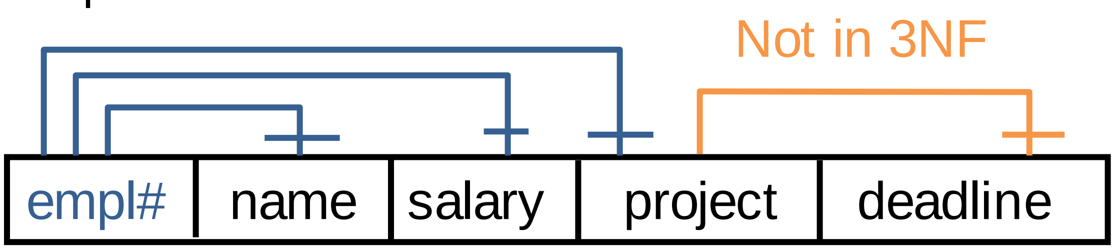




3. Normal form – no transitive dependency

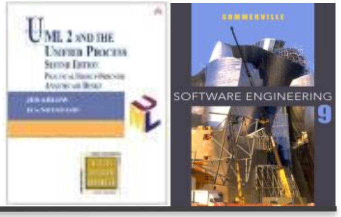
Def. 3NF: In 2NF and every non-prime attribute is non-transitively (i.e. only directly) dependent on every candidate key.

Example EMPLOYEE

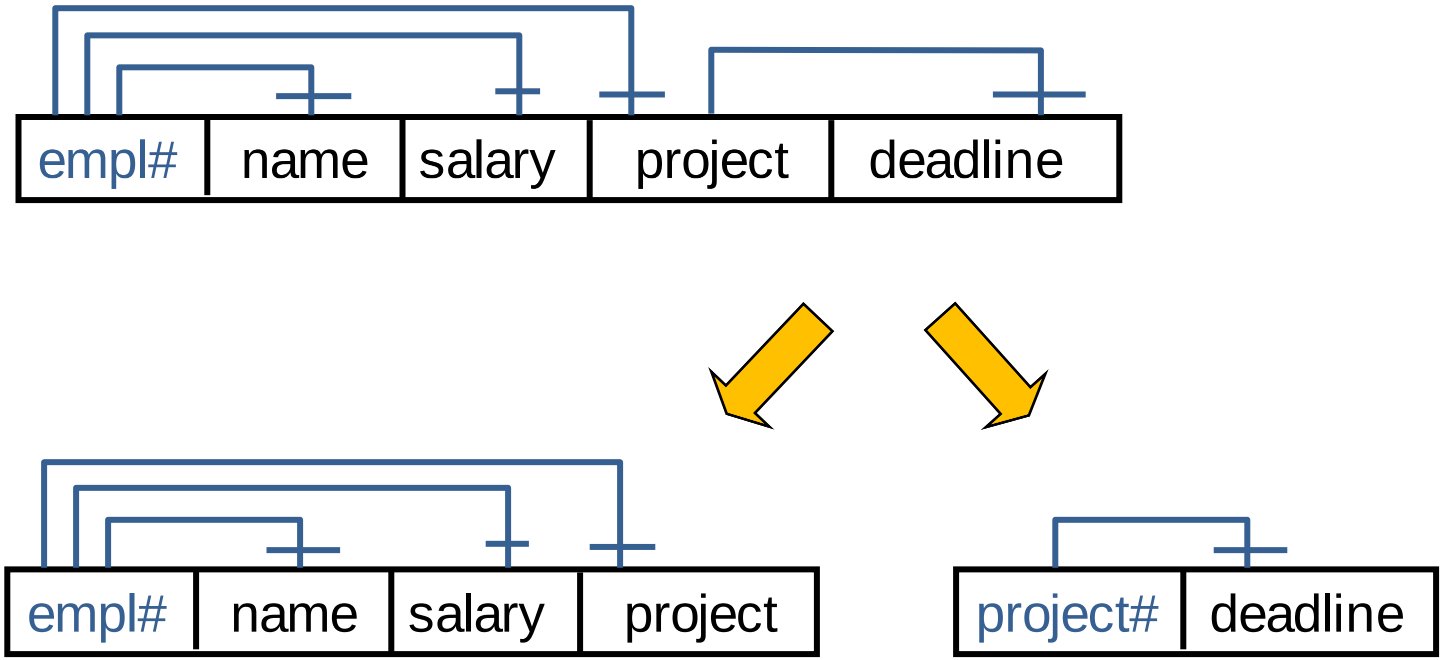


What anomalies can you identify in this example?

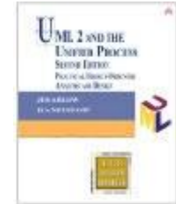
3. Normal form – normalization example



deadline is transitively dependent on empl#



ERD vs. UML Class Diagram



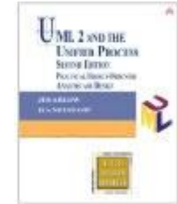
✧ Class diagrams

- model both **structural and behavior features** of a system (attribute and operations),
- contain **many different types of relationships** (association, aggregation, composition, dependency, generalization), and
- are more likely to **map into real-world objects**.

✧ Entity relationship models

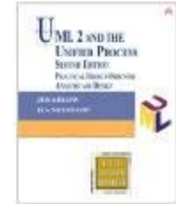
- model only **structural data view** with a low variety of relationships (simple relations and rarely generalization), and
- are more likely to **map into database tables** (repetitive records).
- They allow us to design **primary and foreign entity keys**, and used to be normalized to simplify data manipulation.

ERD vs. UML Class Diagram



- Although there can be one to one mapping between ERD and Class diagram, it is very common that
  - one class is mapped to more than one entity, or
  - more classes are mapped to a single entity.
- Furthermore, not all classes need to be persistent and hence reflected in the ERD model, which uses to be driven by the database design.
- **Summary:**
  - ERD is **data-oriented** and **persistence-specific**
  - Class diagram targets also **operations** and is **persistence independent**

Key points



- Data modeling, and ERD in particular, focuses on modeling data **entities, relationships and attributes**.
- Data normalization focuses on reducing **redundancy and dependency** in database design, and on avoiding bias towards a particular **pattern of querying**.
  - 1NF: no repeating groups
  - 2NF: no partial dependency
  - 3NF: no transitive dependency