Fundamentals of Computer Engineering

Module I - Unit 10
Data Foundations

Teachers: Moisés Martínez (1ºA English)

Year: 2022 - 2023

Universidad
Francisco de Vitoria
UFV Madrid

Grado en Ingeniería Informática
Escuela Politécnica Superior

Universidad
Francisco de Vitoria
UFV Madrid

Basic about Data Foundations

Grado en Ingeniería Informática
Escuela Politécnica Superior

Basic about Data Foundations

Data vs Information

Data is unorganised and unrefined **raw** facts.

Information is the **organization** and **interpretation** of those facts

Basic about Data Foundations

The information is classified based on how the data is stored and organized through some type of structure and/or labeling.

Basic about Data Foundations

The information is classified based on how the data is stored and organized through some type of structure and/or labeling.

Structured

Information is organized using some data model or schema. There is a precise definition of the meaning of each element.

There is a defined and identifiable structure.

Databases

Basic about Data Foundations

The information is classified based on how the data is stored and organized through some type of structure and/or labeling.

Structured

Information is organized using some data model or schema. There is a precise definition of the meaning of each element.

There is a defined and identifiable structure.

Databases

Semi-structured

Information does not obey the tabular structure of data models associated with relational databases or other forms of data tables.

There is tags or other markers which include some semantics.

XML, HTML, JSON

Basic about Data Foundations

The information is classified based on how the data is stored and organized through some type of structure and/or labeling.

Structured

Information is organized using some data model or schema. There is a precise definition of the meaning of each element.

There is a defined and identifiable structure.

Databases

Semi-structured

Information does not obey the tabular structure of data models associated with relational databases or other forms of data tables.

There is tags or other markers which include some semantics.

CSV, XML, HTML, JSON

Unstructured

Unstructured data is information that is not arranged according to a preset data model or schema.

Plain text

Basic about Data Foundations

There are different storage systems depending on the format we want to use to store the data.

Files

Plain text files with useful information.

They usually store raw data.

Information is messy and unstructured.

Basic about Data Foundations

There are different storage systems depending on the format we want to use to store the data.

Files	Distributed files
Plain text files with useful information. They usually store raw data. Information is messy and unstructured.	Mass storage in structured data files. They are stored in a distributed environment. Information has some structure.

Basic about Data Foundations

There are different storage systems depending on the format we want to use to store the data.

Files

Plain text files with useful information.

They usually store raw data.

Information is messy and unstructured.

Distributed files

Mass storage in structured data files.

They are stored in a distributed environment.

Information has some structure.

Databases

Storage in SQL and NOSQL databases.

Databases

Databases

A databases is a collection of information stored on a computer or computer system in a form that can be easily accessed, retrieved and modified.

- Facilitate the storage of large amounts of information.
- Facilitate the information retrieval in a fast and flexible way.
- Facilitate the organisation allowing to linking of different types of information
- Facilitate printing and distribution of information in a wide variety of forms.

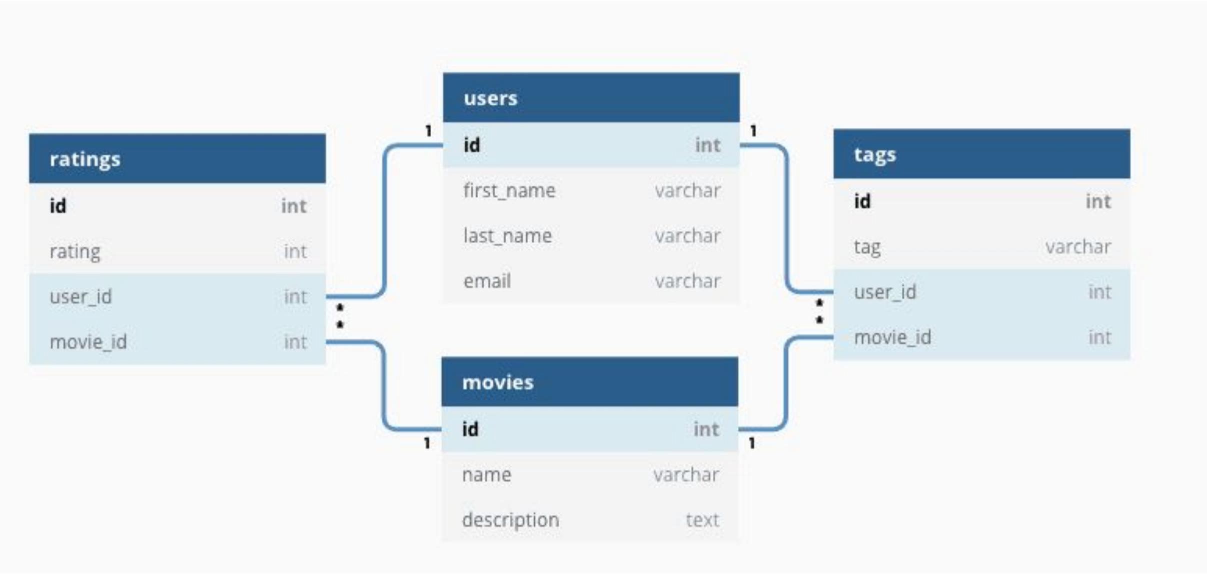
Databases

There are different types of databases, depending on

- How information is organized
  - SQL Databases
  - NoSQL (**N**ot **o**nly **SQL**) Databases
- How information is stored in the physical level
  - Centralised
  - Distributed
  - Cloud

Databases

A relational database (RDB) is a way of structuring information in tables, rows, and columns. An RDB has the ability to establish links—or relationships–between information by joining tables, which makes it easy to understand and gain insights about the relationship between various data points.



ratings	
id	int
rating	int
user_id	int
movie_id	int

users	
id	int
first_name	varchar
last_name	varchar
email	varchar

tags	
id	int
tag	varchar
user_id	int
movie_id	int

movies	
id	int
name	varchar
description	text

Often, a relational database can be referred to as a **SQL database.**

Databases

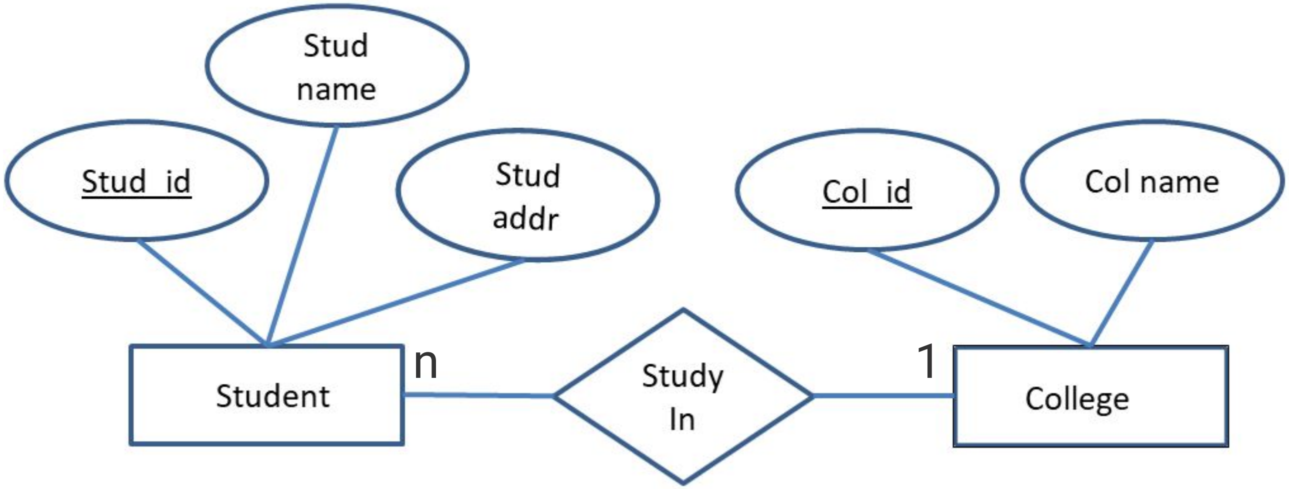
Relational databases are described using an Entity-Relationship diagram which is composed of three basic elements:

- Entity: thing, person, place, unit, object or any item which information is stored. For example, users in the previous diagram.
- Attribute: It is specific information (features) of an entity. They are the entity properties. For example, the first name of a user.
- Primary-key: It is a special attribute which identifies each record in a entity.
- Relationship: they are links or relations, as their name indicates, between entities. For example, one specific user can rate multiple movies according to the previous diagram.

Databases

Entity-Relationship diagrams have some rules when are defined:

- Entities are represented with rectangles.
- Attributes with ovals, those that constitute a primary key are underlined.
- Relationships between entities with diamonds. The ends of the relationship are labelled to indicate the type of the relationship:
  - A 1 at one indicates that the relationship is between a single entity at that side.
  - An N (or M) indicates that the relationship is established with more than one entity at that side.



Databases

There are different kind of relationship, **two** or three, between entities:

- A one-to-one (1:1) relationship: An entity A relates only to an entity B and vice-versa. This relationship is not very common, because often one of the entities is defined as an attribute of the other. For example, each car has a unique number plate and each number plate belongs to only one car.
- An One-to-many (1:N) relationship: An entity in A is related to zero or many entities in B. But an entity in B relates to only one entity in A. For example, a customer can place any number of orders. But each specific order is placed by only one customer.

There is another relationship called many-to-many (N:M) but this relationship is transformed into a intermedie entity where its attributes are the primary keys of the entities in the relationship.

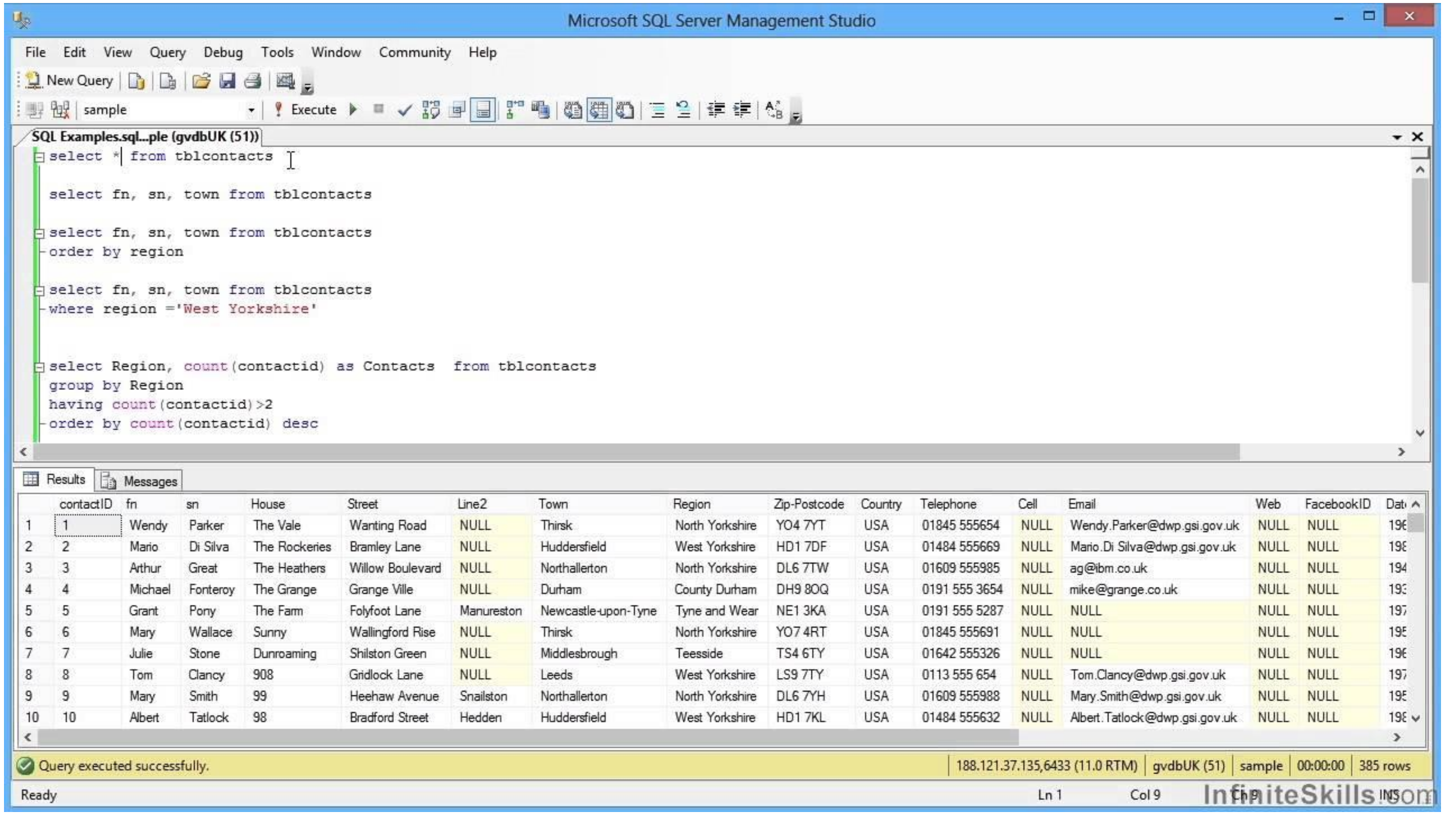
Databases

DataBase Management System (DBMS) is an application software designed to store, retrieve, query and manage data. DBMSs provide some functions that allow management of a database and its data:

- Definition: Creation, modification and removal of definitions that define the organization of the data.
- Manage and exploitation: Insertion, modification, query and deletion of the data.
- Administration: Registering and monitoring users, enforcing data security, monitoring performance, maintaining data integrity, dealing with concurrency control, and recovering information that has been corrupted by some event such as an unexpected system failure.

Databases

SQL (Structured Query Language) is a domain-specific programming language designed to manage and retrieve information from relational database management systems.



Microsoft SQL Server Management Studio
File Edit View Query Debug Tools Window Community Help
New Query
sample
Execute
SQL Examples.sql...ple (gvdbUK (51))
Results
Messages

```
select * from tblcontacts

select fn, sn, town from tblcontacts

select fn, sn, town from tblcontacts
order by region

select fn, sn, town from tblcontacts
where region ='West Yorkshire'


select Region, count(contactid) as Contacts  from tblcontacts
group by Region
having count(contactid)>2
order by count(contactid) desc
```

	contactID	fn	sn	House	Street	Line2	Town	Region	Zip-Postcode	Country	Telephone	Cell	Email	Web	FacebookID	Dat
1	1	Wendy	Parker	The Vale	Wanting Road	NULL	Thirsk	North Yorkshire	YO4 7YT	USA	01845 555654	NULL	Wendy.Parker@dwp.gsi.gov.uk	NULL	NULL	196
2	2	Mario	Di Silva	The Rockeries	Bramley Lane	NULL	Huddersfield	West Yorkshire	HD1 7DF	USA	01484 555669	NULL	Mario.Di Silva@dwp.gsi.gov.uk	NULL	NULL	198
3	3	Arthur	Great	The Heathers	Willow Boulevard	NULL	Northallerton	North Yorkshire	DL6 7TW	USA	01609 555985	NULL	ag@ibm.co.uk	NULL	NULL	194
4	4	Michael	Fonteroy	The Grange	Grange Ville	NULL	Durham	County Durham	DH9 8OQ	USA	0191 555 3654	NULL	mike@grange.co.uk	NULL	NULL	193
5	5	Grant	Pony	The Farm	Folyfoot Lane	Manureston	Newcastle-upon-Tyne	Tyne and Wear	NE1 3KA	USA	0191 555 5287	NULL	NULL	NULL	NULL	197
6	6	Mary	Wallace	Sunny	Wallingford Rise	NULL	Thirsk	North Yorkshire	YO7 4RT	USA	01845 555691	NULL	NULL	NULL	NULL	195
7	7	Julie	Stone	Dunroaming	Shilston Green	NULL	Middlesbrough	Teesside	TS4 6TY	USA	01642 555326	NULL	NULL	NULL	NULL	196
8	8	Tom	Clancy	908	Gridlock Lane	NULL	Leeds	West Yorkshire	LS9 7TY	USA	0113 555 654	NULL	Tom.Clancy@dwp.gsi.gov.uk	NULL	NULL	197
9	9	Mary	Smith	99	Heehaw Avenue	Snailston	Northallerton	North Yorkshire	DL6 7YH	USA	01609 555988	NULL	Mary.Smith@dwp.gsi.gov.uk	NULL	NULL	195
10	10	Albert	Tatlock	98	Bradford Street	Hedden	Huddersfield	West Yorkshire	HD1 7KL	USA	01484 555632	NULL	Albert.Tatlock@dwp.gsi.gov.uk	NULL	NULL	198

Query executed successfully. | 188.121.37.135,6433 (11.0 RTM) | gvdbUK (51) | sample | 00:00:00 | 385 rows

Ready Ln 1 Col 9 Ch 9 INS

Databases

SQL (Structured Query Language) is a domain-specific programming **language** designed to manage and retrieve information from relational database management systems.

```
CREATE TABLE shop (
 article INT(4) UNSIGNED ZEROFILL DEFAULT '0000' NOT NULL,
 dealer  CHAR(20)                 DEFAULT ''     NOT NULL,
 price   DOUBLE(16,2)             DEFAULT '0.00' NOT NULL,
 PRIMARY KEY(article, dealer));



INSERT INTO shop VALUES
(1,'A',3.45),(1,'B',3.99),(2,'A',10.99),(3,'B',1.45),
(3,'C',1.69),(3,'D',1.25),(4,'D',19.95);
```

```
SELECT * FROM shop;

+---------+--------+-------+
| article | dealer | price |
+---------+--------+-------+
|    0001 | A      |  3.45 |
|    0001 | B      |  3.99 |
|    0002 | A      | 10.99 |
|    0003 | B      |  1.45 |
|    0003 | C      |  1.69 |
|    0003 | D      |  1.25 |
|    0004 | D      | 19.95 |
+---------+--------+-------+
```

! OFF-EXAM CONTENT

We want to create a database for an online store.

Which entities do we need to store the basic data?

We want to create a database for an online store.

Which entities do we need to store the basic data?

Clients, Products and orders.

We want to create a database for an online store.

Which entities do we need to store the basic data?

Clients, Products and orders.

Which information do we need to store client data?

We want to create a database for an online store.

Which entities do we need to store the basic data?

Clients, Products and orders.

Which information do we need to store client data?

- Client: Name, Surname, DNI, billing address, shipping address and more …
- Product: Reference, name, description, price, available units and more …
- Order: Order_id, Date, client_id, product_id, state and more …

We want to create a database for an online store.

Which entities do we need to store the basic data?

Clients, Products and orders.

Which information do we need to store client data?

- Client: Name, Surname, DNI, billing address, shipping address and more …
- Product: Reference, name, description, price, available units and more …
- Order: Order_id, Date, client_id, product_id, state and more …

Which relationships do we need?

We want to create a database for an online store.

Which entities do we need to store the basic data?

Clients, Products and orders.

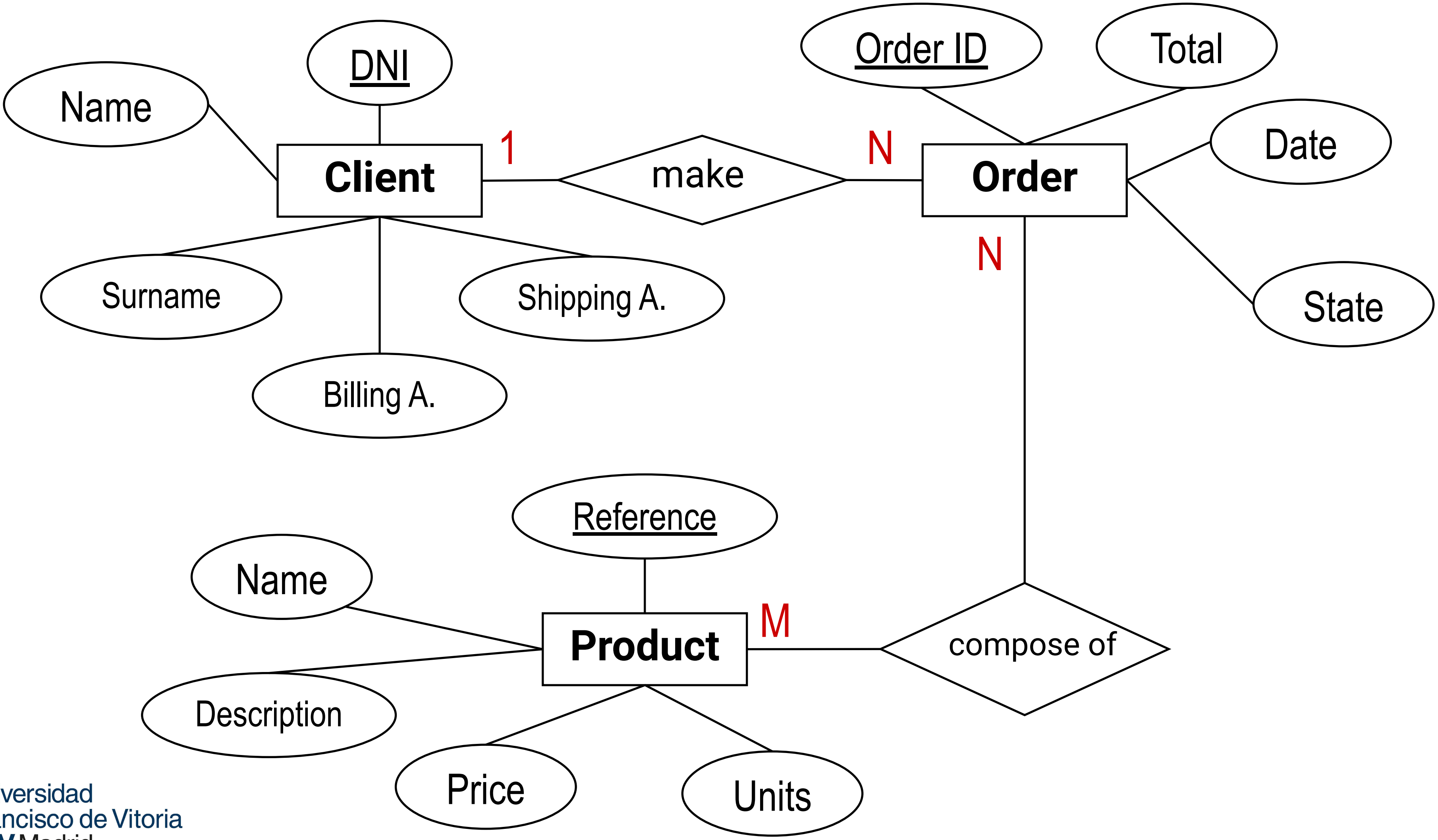
Which information do we need to store client data?

- Client: Name, Surname, DNI, billing address, shipping address and more …
- Product: Reference, name, description, price, available units and more …
- Order: Order_id, Date, client_id, product_id, state and more …

Which relationships do we need?

- 1:N → client:order
- N:M → product:order

We want to create a database for an online store.



We want to create a database for an online store.

Definition of the information

- The different entities are converted into tables, where each attribute is a field of the table.

Fields have a type: varchar, int, double, char, etc.

Each concrete element within the table (row) is called a record.

Primary keys are created using fields which must be unique to identify each concrete element.

We want to create a database for an online store.

Definition of the relationships' cardinalities

- The relationship (make) 1-to-many (1:N) is included to describe when a client have many orders but each specific order is only related to one client.

This relationship will include the primary key of the 1 entity side into the N entity side. The field DNI (primary key of client) is included as a field of the order table.

We want to create a database for an online store.

Definition of the relationships' cardinalities

- The relationship (compose-of) many-to-many (N:M) is included to describe when an order is composed of some product units.

This relationship will generate a new table into the DataBase. It can include other attributes like units, price, etc.

Order id	Reference
12324	364834034843784

This table will include the primary keys of the tables in the relationship.

Big Data

Big Data

https://www.youtube.com/watch?v=bAyrObl7TYE

❗ OFF-EXAM CONTENT

“Big Data” definition

The term "Big Data" describes the large volume of data, both structured and unstructured, that is created daily by today's society.

Big Data

“Big Data” definition

The term "Big Data" describes the large volume of data, both structured and unstructured, that is created daily by today's society.

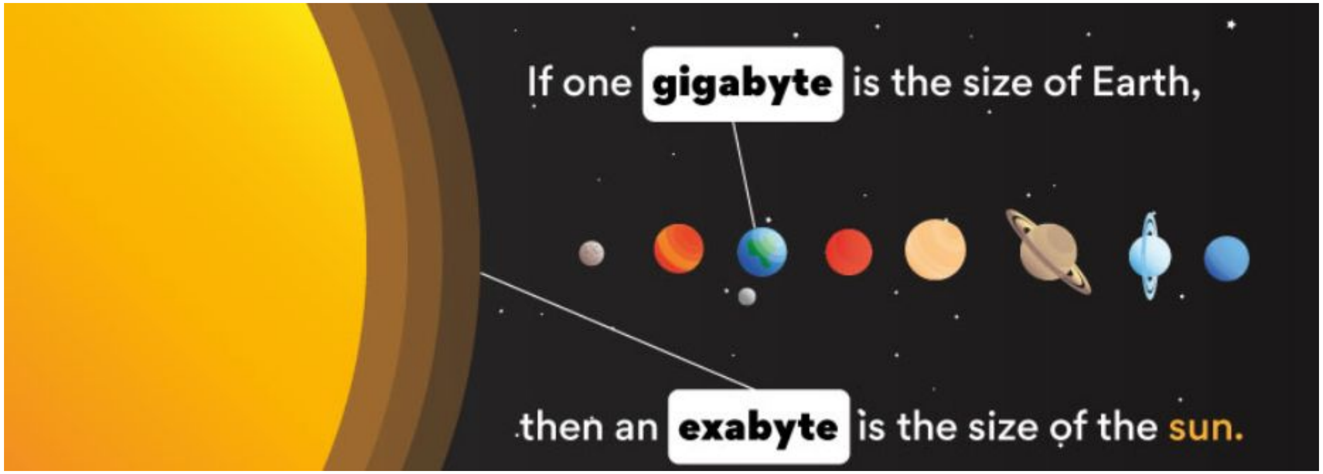
My “Big Data” definition

Process of collecting, storing and subsequent analysis and manipulation of data at a massive level in order to extract **value**.

❗ OFF-EXAM CONTENT

Big Data

From **cuneiform** writing, the oldest known writing system to date, to modern data centers, the human race has always collected information. Furthermore, it is predicted that by 2030 our civilization will generate several **yottabytes** of information per year.


If one gigabyte is the size of Earth,
then an exabyte is the size of the sun.

The **yottabyte** is currently the largest recognized unit of data storage for devices and cloud services.

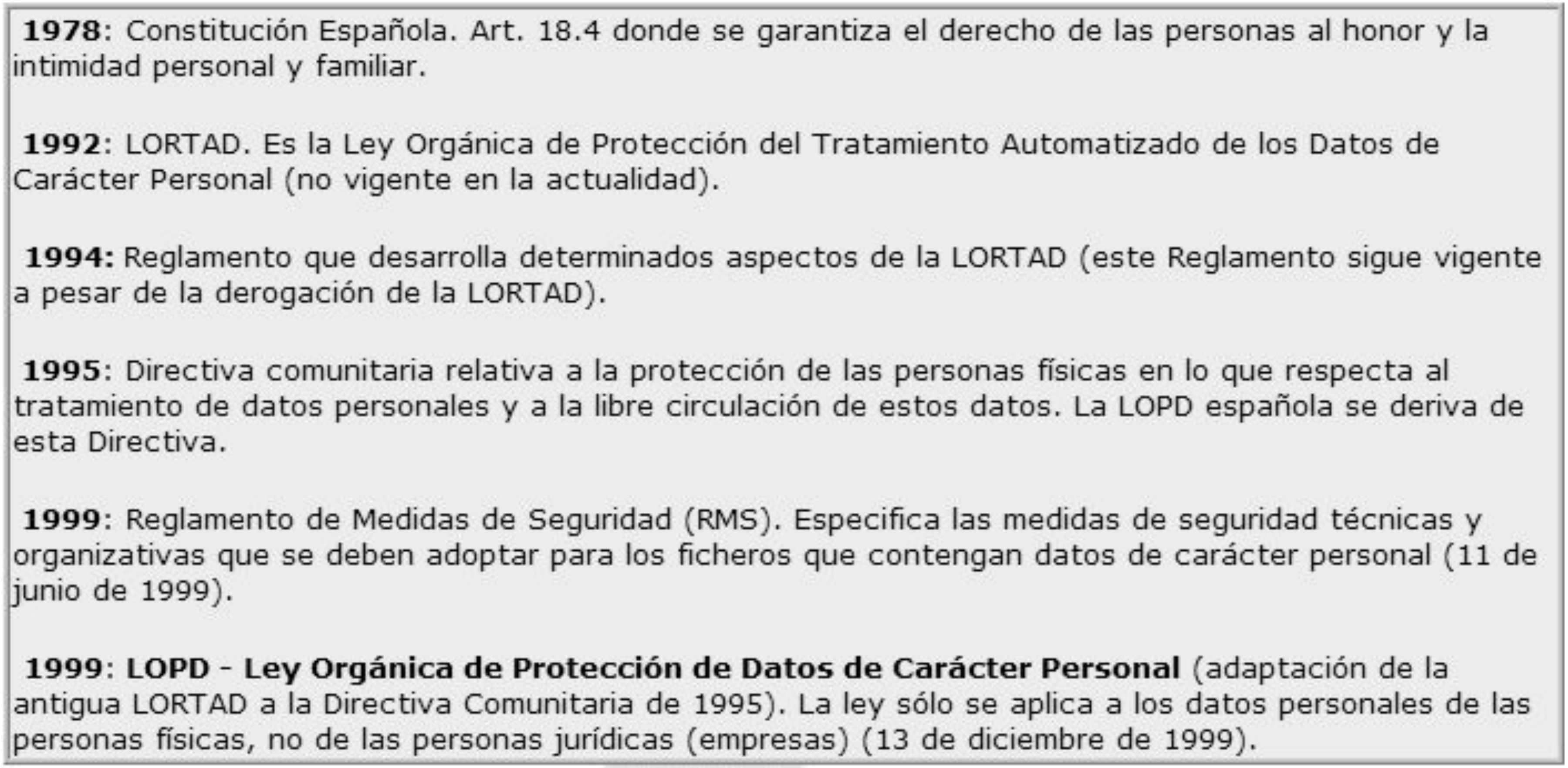
Universidad Francisco de Vitoria
UFV Madrid

! OFF-EXAM CONTENT

Laws and ethics

Laws and ethics

Evolution of the data laws in Spain



1978: Constitución Española. Art. 18.4 donde se garantiza el derecho de las personas al honor y la intimidad personal y familiar.

1992: LORTAD. Es la Ley Orgánica de Protección del Tratamiento Automatizado de los Datos de Carácter Personal (no vigente en la actualidad).

1994: Reglamento que desarrolla determinados aspectos de la LORTAD (este Reglamento sigue vigente a pesar de la derogación de la LORTAD).

1995: Directiva comunitaria relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. La LOPD española se deriva de esta Directiva.

1999: Reglamento de Medidas de Seguridad (RMS). Especifica las medidas de seguridad técnicas y organizativas que se deben adoptar para los ficheros que contengan datos de carácter personal (11 de junio de 1999).

1999: **LOPD - Ley Orgánica de Protección de Datos de Carácter Personal** (adaptación de la antigua LORTAD a la Directiva Comunitaria de 1995). La ley sólo se aplica a los datos personales de las personas físicas, no de las personas jurídicas (empresas) (13 de diciembre de 1999).

! OFF-EXAM CONTENT

Laws and ethics

LOPD - Ley Orgánica de Protección de datos

The Organic Law 15/1999 of December 13 on Protection of Personal Data (Ley Orgánica de Protección de Datos de Carácter Personal, LOPD) was Spanish organic law that guaranteed and protected the processing of personal data, public liberties, and fundamental human rights, and especially of personal and family honor and privacy.

- Regulate the treatment of data and files, of a personal nature, regardless of the support in which they are treated.
- the rights of citizens over them and the obligations of those who create or treat them.

❗ OFF-EXAM CONTENT

LOPD in the European Union

The GRDP (General Regulation Data Protection) is a regulation in European Union (EU) law on data protection and privacy in the and the European Economic Area (EEA). This law is mandatory from May 25, 2018 in all member countries of the European Union.

The LOPD, as a national law (Spain), was integrated into the GRPD, which unified all European data protection regulations under the same legal umbrella. So that every European citizen has the same rights and guarantees regarding personal information.

LOPD == GRDP

LOPD in the European Union

The LOPD-GDD (Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales) is the new national law integrated into the GRPD, which unified all European data protection regulations under the same legal umbrella. This law extends the regulations defined into the GRPD.

! OFF-EXAM CONTENT

Universidad
Francisco de Vitoria
UFV Madrid

How to implement the new LOPD-GDD

in a company?

LOPD in the European Union

Any type of **company** or **business** that deals with **sensitive data of third parties**, must comply with each and every one of the requirements established in the new regulations of the Law on Protection of Personal Data and Guarantees of Digital Rights. The LOPD-GDD will be applied when the following treatments are given:

- Data processing of individual entrepreneurs and liberal professionals.
- Commercial operations.
- Use of surveillance systems.
- Advertising exclusion systems.
- Communication channels and complaints.
- Credit information systems.

OFF-EXAM CONTENT

Laws and ethics

LOPD in the European Union

There some important points in the new law:

Protection of minors: The consent of a minor will only be valid when he is over fourteen years of age, being necessary the authorization of the father, mother or guardian if it is not.

Control of personal data: To avoid the use of personal data for commercial use without prior consent, the LOPD-GDD establishes that the control of personal data falls directly on the user,always requiring their consent to use them.

Employee privacy: It is forbidden to take recordings in the areas intended for the rest of the workers, toilets and other places intended for leisure.

! OFF-EXAM CONTENT

LOPD in the European Union

There some important points in the new law:

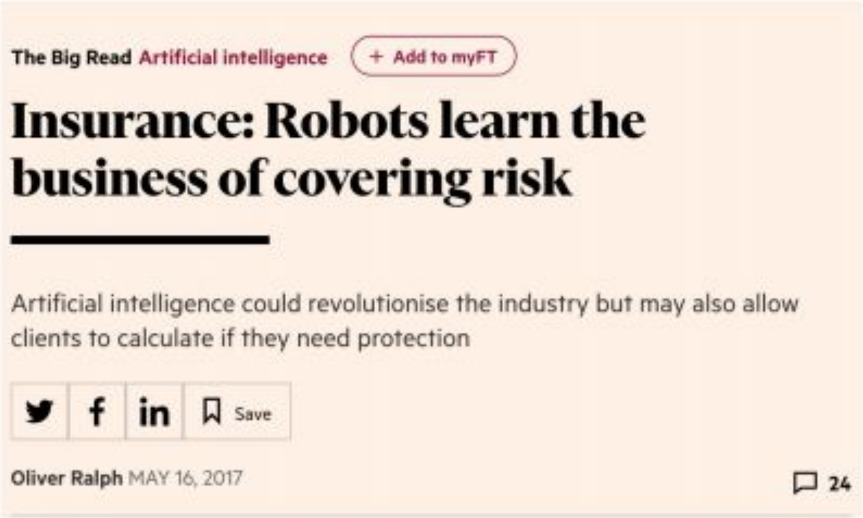
Right to be forgotten: it establishes the right to delete data on social networks and other equivalent services.

Data of deceased persons: In the event of death, any family member linked to the deceased person may request access, rectification or deletion of the shared data.

Clear information about the use of data: Companies must inform users in a clear, simple and concise way about the possible use of the personal data they have been given. **Companies could be fined up to EUR 20 million.**

What about ethics?

Ethics

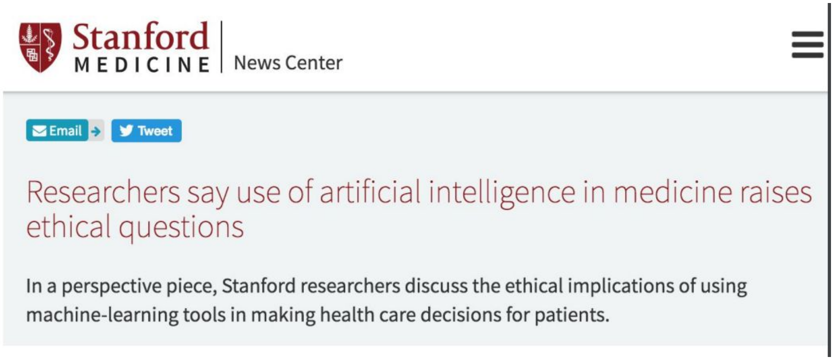


The Big Read Artificial intelligence

Insurance: Robots learn the business of covering risk

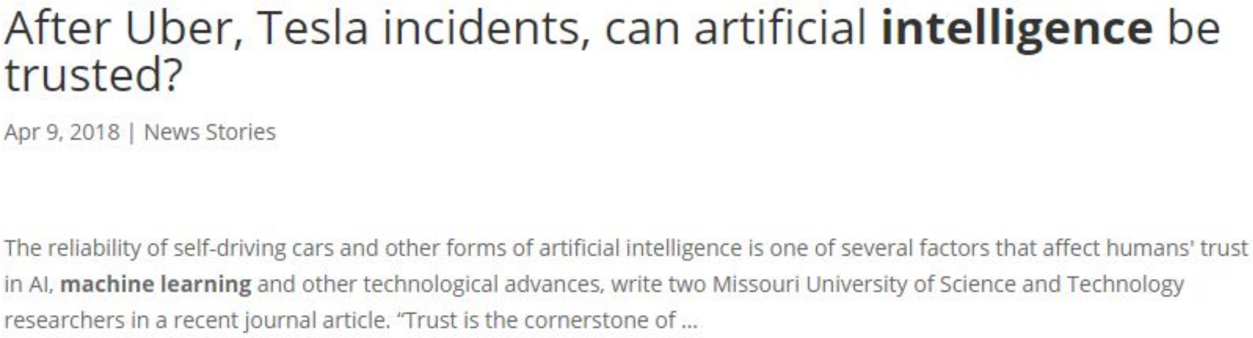
Artificial intelligence could revolutionise the industry but may also allow clients to calculate if they need protection

Oliver Ralph MAY 16, 2017



Researchers say use of artificial intelligence in medicine raises ethical questions

In a perspective piece, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients.



After Uber, Tesla incidents, can artificial **intelligence** be trusted?

Apr 9, 2018 | News Stories

The reliability of self-driving cars and other forms of artificial intelligence is one of several factors that affect humans' trust in AI, **machine learning** and other technological advances, write two Missouri University of Science and Technology researchers in a recent journal article. "Trust is the cornerstone of ...

OFF-EXAM CONTENT

Ethics

Ethics, also called moral philosophy, the discipline concerned with what is morally good and bad and morally right and wrong. The term is also applied to any system or theory of moral values or principles.

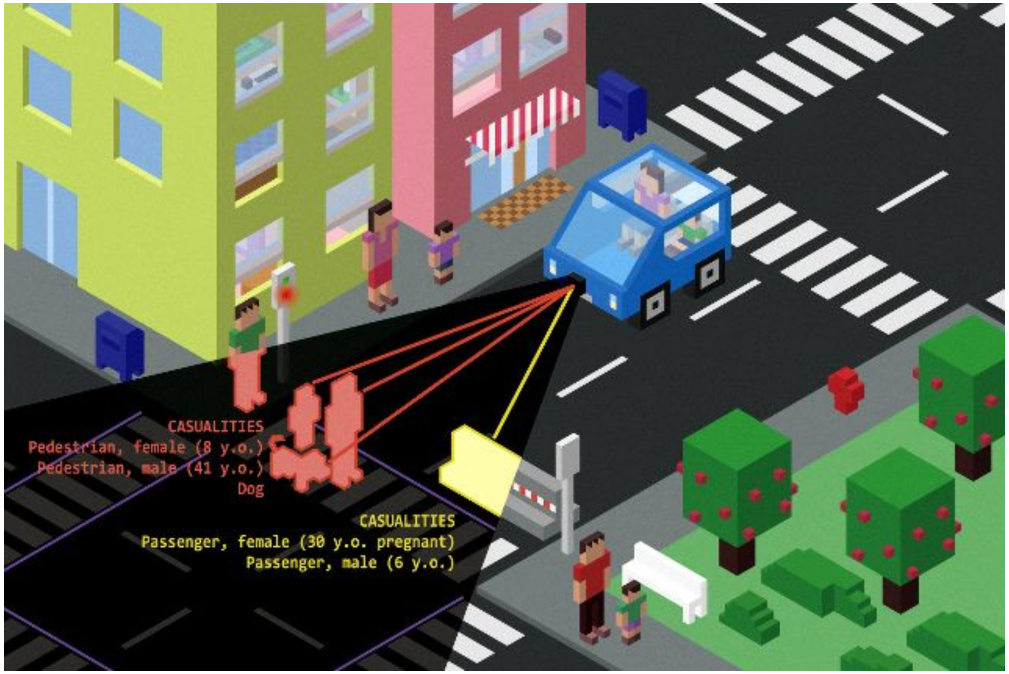
Why do we need ethics when we are creating software?

- Human have **biases** that we include in the information we create and analyze.
- During the design and development of software and data we contribute unconscious knowledge, we must ask ourselves if we have taken into account enough examples.
- Ideologies, expressions, validated technical information, natural/artificial light, atypical cases, etc.

❗ OFF-EXAM CONTENT

Ethics

https://www.moralmachine.net/


CASUALTIES
Pedestrian, female (8 y.o.)
Pedestrian, male (41 y.o.)
Dog
CASUALTIES
Passenger, female (30 y.o. pregnant)
Passenger, male (6 y.o.)

Option 1

In this case, the self-driving car with a sudden brake failure will swerve and drive through a crosswalk in the other lane. This will result in the death of an elderly woman, two athletes and a child.

Please note that affected pedestrians are complying with the law when crossing with the green signal

Option 2

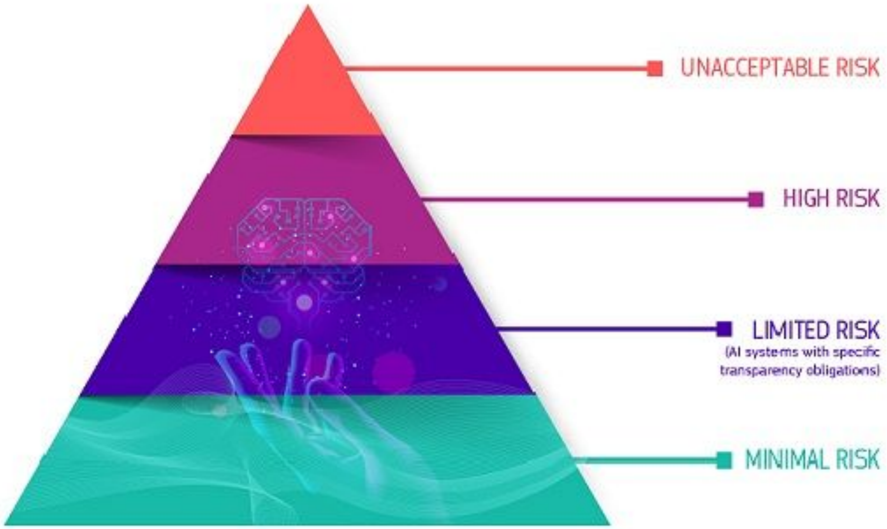
In this case, the self-driving car with sudden brake failure will continue forward and crash into a concrete barrier. This will result in the death of a child, a pregnant woman (passengers) and a dog.”

https://www.eeworldonline.com/this-mit-game-lets-you-choose-who-lives-and-dies-in-a-self-driving-car-wreck/

❗ OFF-EXAM CONTENT

How we are trying to create software or data ethically?

New regulatory framework on AI (European Union).



- Unacceptable risk: A very limited set of particularly harmful uses of AI that contravene EU values because they violate fundamental rights.
  - **Social scoring for governments (This is happening in China).**
- High Risk: A limited number of AI systems defined in the proposal, creating an adverse impact on people's safety or their fundamental rights (as protected by the EU Charter of Fundamental Rights) are considered to be high-risk.
  - Infrastructure.
  - Education.
  - Security.
  - Public services.
  - Inmigration or border line controls.

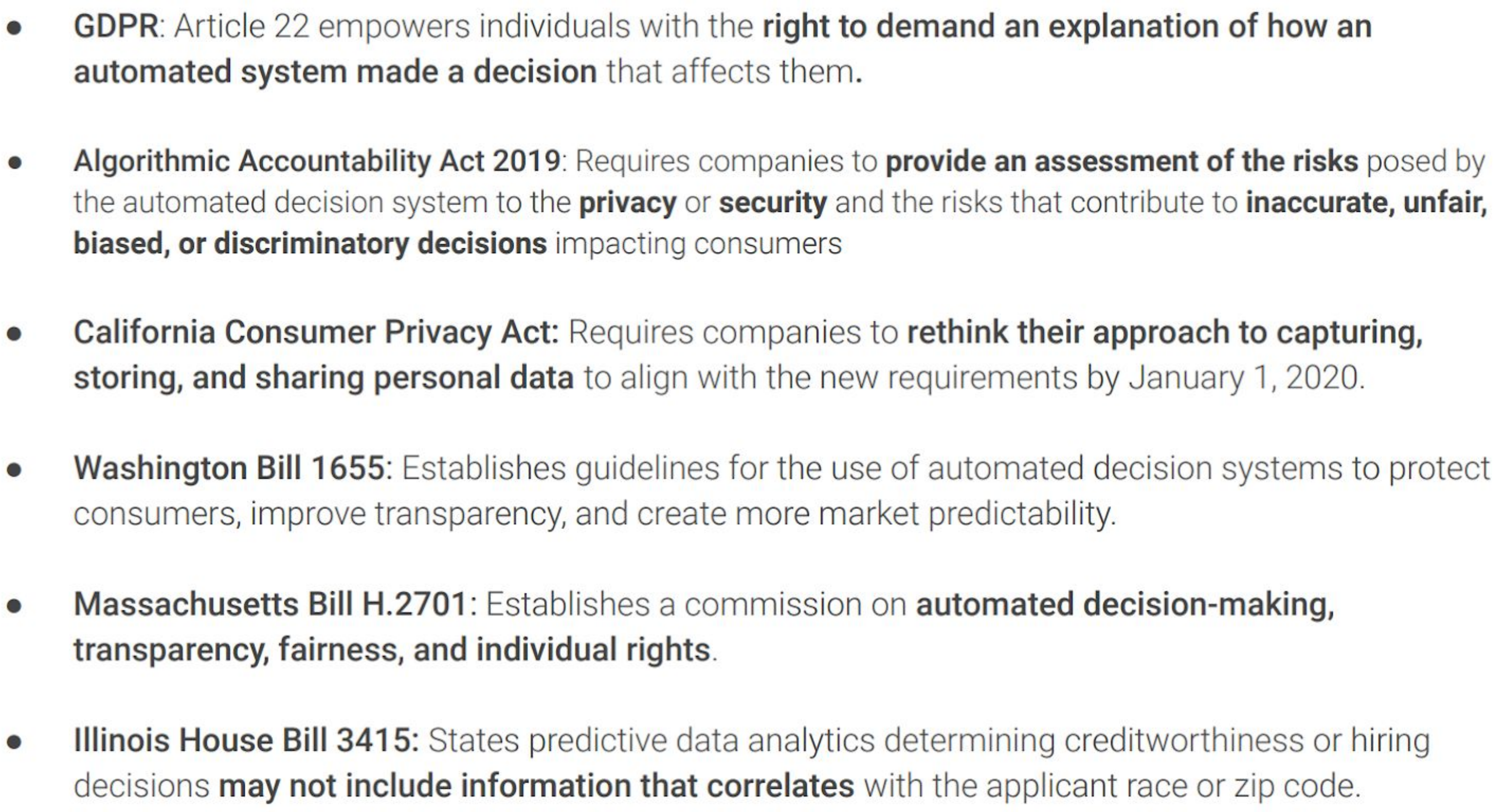
- Limited risk: For certain AI systems specific transparency requirements are imposed, for example where there is a clear risk of manipulation (e.g. via the use of chatbots). **Users should be aware that they are interacting with a machine.**

- Minimal risk: All other AI systems can be developed and used subject to the existing legislation without additional legal obligations.
  The vast majority of AI systems currently used in the EU fall into this category.

❗ OFF-EXAM CONTENT

Ethics

There are rules in other countries and states.



- **GDPR**: Article 22 empowers individuals with the **right to demand an explanation of how an automated system made a decision** that affects them.
- **Algorithmic Accountability Act 2019**: Requires companies to **provide an assessment of the risks** posed by the automated decision system to the **privacy** or **security** and the risks that contribute to **inaccurate, unfair, biased, or discriminatory decisions** impacting consumers
- **California Consumer Privacy Act:** Requires companies to **rethink their approach to capturing, storing, and sharing personal data** to align with the new requirements by January 1, 2020.
- **Washington Bill 1655:** Establishes guidelines for the use of automated decision systems to protect consumers, improve transparency, and create more market predictability.
- **Massachusetts Bill H.2701**: Establishes a commission on **automated decision-making, transparency, fairness, and individual rights**.
- **Illinois House Bill 3415:** States predictive data analytics determining creditworthiness or hiring decisions **may not include information that correlates** with the applicant race or zip code.

❗ OFF-EXAM CONTENT