

# Lecture 1: Basic digital knowledge

Isabel Casas

# Class content

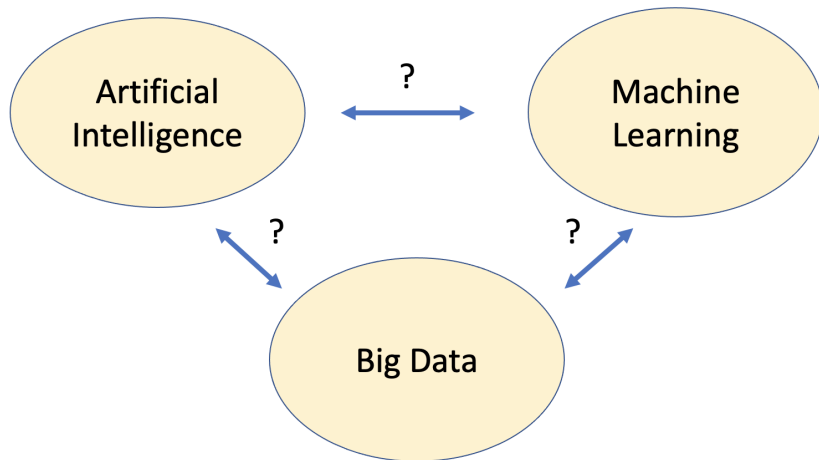
- Concepts on ML, AI and big data.
- Data repositories
- Download MPG data

# Section 1

## Concepts

- ➊ *App*: any application, integration, configuration, extension, update set or other material developed in a programming language
- ➋ *UI*: User interface, *GUI*: graphical user interface
- ➌ *Database*: Set of values saved in the form of a vector, matrix or tensor
- ➍ *API*: application programming interfaces. Set of functions to interact between two computing programs
- ➎ *Machine learning*:
- ➏ *Statistical modelling*:
- ➐ *Artificial intelligence*:
- ➑ *Big data*:

Are all these disciplines related?



# What is ...

- Can you give me some examples of artificial intelligence?
- Can you give me some examples of machine learning?

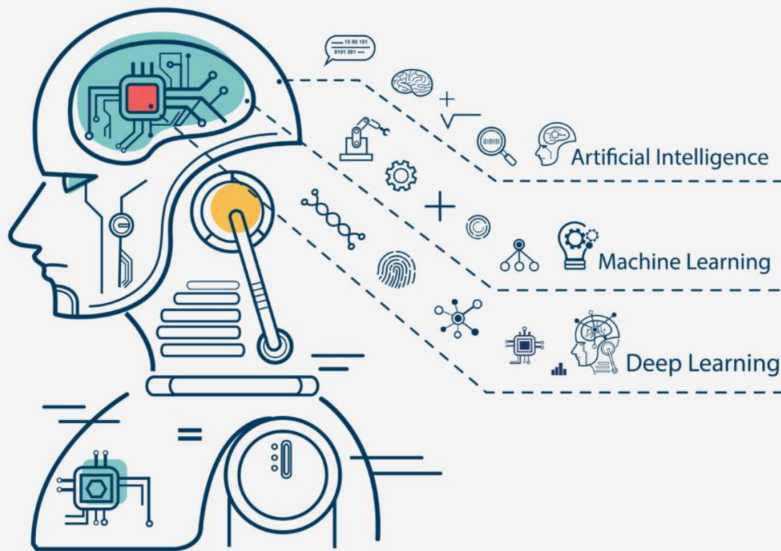
- The goal of AI is to create systems that can simulate human intelligence, understand natural language, perceive the environment, learn from experience, and make decisions autonomously.
- AI encompasses various techniques, including machine learning, natural language processing, robotics, expert systems, and more.

# Machine learning

- The goal of machine learning is to develop algorithms that improve automatically through experience and can generalize from past data to new, unseen data.
- Machine learning is a subset of AI that focuses on the development of algorithms and models that allow computers to learn from data and make predictions or decisions without being explicitly programmed for every task
- “A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.” Mitchell, 1997



# Machine learning



# Machine learning

- 1 Machine Learning is a newer field of study than statistics (for instance, Machine Learning was invented in 1959, whereas statistics originated in the 17th century)
- 2 Machine Learning focuses on prediction.
- 3 Machine Learning is a subfield of computer science and AI, and contributes to building systems that can learn from data
- 4 Machine Learning uses fewer assumptions than statistical modelling, or in other words, it is not interested in the generalisation of relationships, only on the ad hoc prediction performance.

- 1 Statistical Modelling is a subfield of Mathematics that deals with finding relationships between variables to predict outcomes, and also to formalise relationships between variables in the form of mathematical equations.
- 2 It deals with a small amount of data with fewer attributes and, as such, there is a good chance that over-fitting will occur.
- 3 Statistical models are the core of any machine learning algorithm.

# Comparison

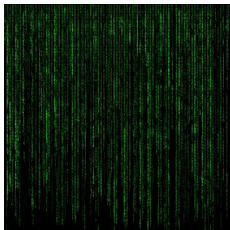
<b>Machine Learning</b>	<b>Statistical Modelling</b>
Network, Graphs	Model, Graphs
Weights, hyperparameters	Parameters
Learning	Fitting
Supervised Learning	Regression/Classification
Unsupervised Learning	Density Estimation/Clustering

We distinguish between supervised and unsupervised learning.

- ① In supervised learning, we have a set of training data, or labeled data, in which we know the structure and the outcome of it.
  - ▶ We take this data and train a machine learning model, so it can understand patterns in the data.
  - ▶ Once the model has been trained, we can use it to predict the results of out-of-sample data, or data in which the results are unknown.
- ② If we are given a set of data that is unstructured, then we can apply unsupervised machine learning models to find patterns that exist within that data.

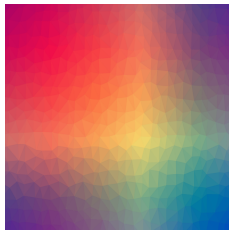
# Machine Learning

Unstructured data



Machine Learning

Pattern



# Big data

- **Big data** refers to datasets that are extremely large and complex, often beyond the capabilities of traditional data processing tools and techniques to efficiently analyze, manage, and process within a reasonable amount of time.
- The term “big data” is characterized by the volume, velocity, and variety of the data being collected and analyzed.
- Dealing with big data often requires specialized techniques and tools such as distributed computing frameworks (e.g., Hadoop, Spark), parallel processing, and advanced machine learning algorithms designed to handle large-scale datasets efficiently.
- In this course, we are going to see machine learning algorithms applied to datasets that are not considered big data

## Section 2

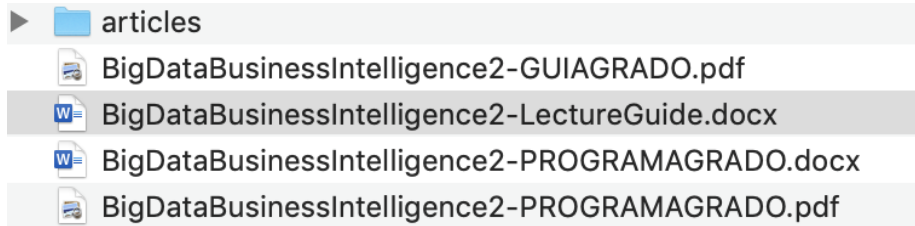
### Working with files and data



# Name and extension

## **BigDataBusinessIntelligence2-LectureGuide.docx**

Every file has a name and an extension. For example:



Questions: What does it mean each element in the graphic above?

# Activity 1 in groups of 3

## Clasify the following programs (applications):

vi, Excel, emacs, Gimp, macOS Edge, Photoshop, Access, SQL, R, C++, Word, Notepad, Matlab, Chrome, Windows 10, Python, Paint, Pages, Numbers, Javascript, HTML, Ubuntu, TextEdit, Opera, Notepad++, Fedora, Wordpad, SPSS, RStudio, Markdown

## into the following categories:

Text editor	Markup language	Spread sheet/ database	Programming language	Browser	Graphics editor	Operating system

Question: Explain each of the classification

Question: Which are free and which need a license?

## Activity 2 in groups

MPG is miles per gallon, a measure of efficiency of a car.

- 1 Go to this [data repository](#) and find dataset *Auto MPG*.
- 2 Download it
- 3 What type of files you got?
- 4 Open them using an R command.

# Data repositories

**UCI Machine Learning Repository**, including datasets for classification, regression, clustering, and other machine learning tasks.

**Kaggle**. You need to open an account. A collaborative community.

**Google's Dataset Search**. It is a search engine that provides access to a wide range of datasets from sources such as universities, research institutions, and government agencies.

**Microsoft's MSR Open Data**. It provides access to a collection of datasets related to software engineering, natural language processing, and other fields.

**Amazon's AWS Open Datasets**. You need to register and ask for your credit card.

## Activity 3 in groups

- 1 Choose a topic you would like to study and motivate why this is an important problem that needs study
- 2 Search for a dataset related to that topic in any of the dataset repositories mentioned above
- 3 Download the dataset into your computer and understand the variables in the dataset

- Log-in Posit Cloud with your Deusto account [login](#)
- Primers/The basics/Programming basics.