WeST People and Knowledge Networks | Institute for Web Science & Technologies

UNIVERSITÄT KOBLENZ · LANDAU

Machine Learning and Data Mining WS21/22

"1 Introduction"

Dr. Zeyd Boukhers

@ZBoukhers

Institute for Web Science and Technologies

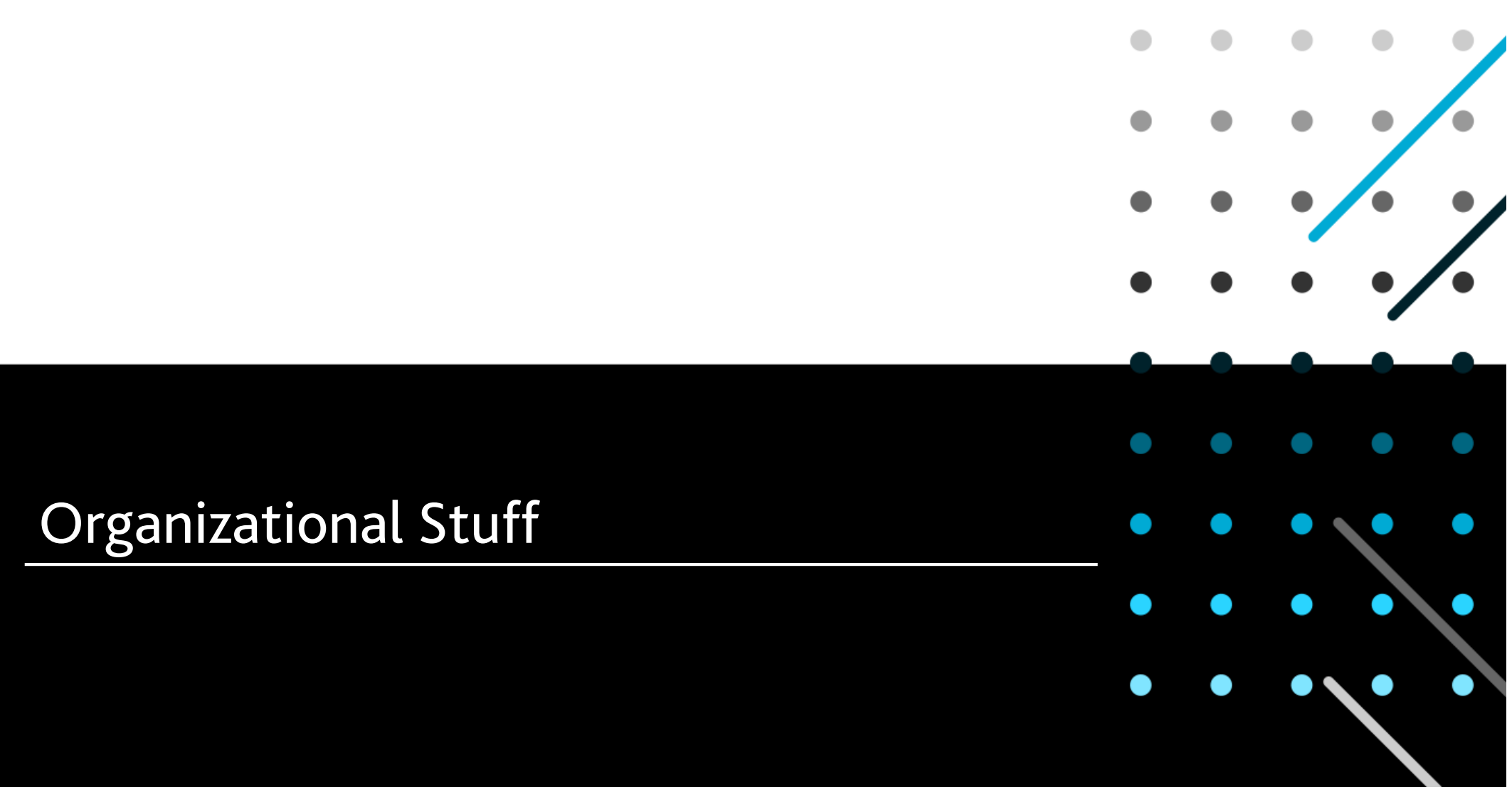
University of Koblenz-Landau

October 27, 2021

Agenda

- Organizational Stuff.
- What is MLDM?
- Predictive & Descriptive learning.
- Supervised & Unsupervised learning.
- Other learning approaches.

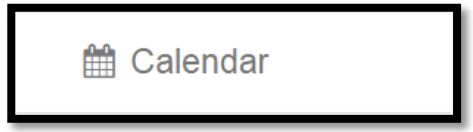
Organizational Stuff

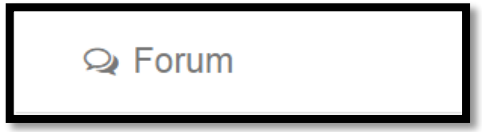


About the MLDM course

- Course ID:
  - 04IN2028.
- Aim:
  - understanding the fundamentals and basics of machine learning, data mining and related topics such as optimization.
- Who is this course for?
  - Master students in: 1) Web Science (Web & Data Science) , 2) Computer Science, 3) Computer Visualistic, 4) Mathematical Modelling, 5) etc.
  - Bachelor students
- **Credit points:**
  - 6 ECTS

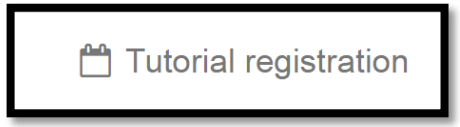
About the MLDM course

- **Where:**
  - Hybrid: Virtual and in-person.
- **When:**
  - **Lectures**: Exclusively video recordings that are released on Wednesdays at 12:00 pm (Noon)
  - **Tutorials**: In-person and live streaming on Thursdays at 02:00 pm (may be also at 04:00 pm)
    - They will be recorded.
  - See the full schedule on the calendar in Olat 

- **Consultation:**
  - Forum in Olat 

- **Material and further information:**
  - Olat: https://olat.vcrp.de/url/RepositoryEntry/3382739338
- **Klips Link:**
  - https://klips.uni-koblenz-landau.de/v/139109

In-person tutorial

- Regulation:
  - The distancing rules are not applied anymore.
  - It is mandatory to wear a mask during the entire session.
  - It is mandatory to possess a negative test result or proof of vaccination or recovery
  - The tutorial will be cancelled if the tutor shows Covid-19 symptoms.
- How it works:
  - For every tutorial, apply to attend at 

in Olat at 11:00 am on the tutorial day at the latest.
  - Every student who cannot attend in-person, must not apply or cancel his/her registration.
  - If the total number of registered students is 60 or less, only the tutorial at 14:00 will be held and the other one will be cancelled.
  - In all cases, only the tutorial at 14:00 will be live-streamed and recorded.

MLDM Team

Lecturer & Tutor

Zeyd Boukhers

Assistants

Akshaya Raja

Sowjanya Chennamaneni

Azeddine Bouabdallah

Assignments

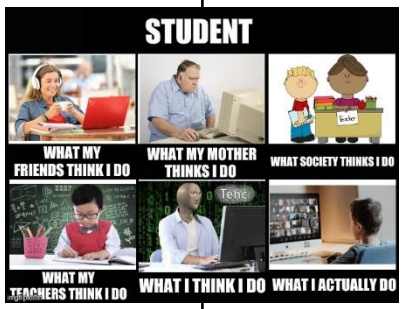
- An assignment is released every Wednesday at 02:00 pm.
- It is an online test on Olat.
- It contains three sections: *Knowledge Questions*, *Practice* and *Programming*.
  - An additional section "*Recall*" will be included in the first assignments
- Each section might contain several tasks.
- You have one week to work on it ***individually***.
- It must be submitted by the following Wednesday at 10:59 pm at the latest.
- Each assignment has 100 points.
- The corresponding tutorial will be the next following Thursday at 02:00 pm.

Structure

× 12

Lecture
Wed. at 12:00 pm

Assignment release
Wed. at 14:00 pm



One week

Assignment submission
Wed. at 22:59 pm

Next day

Tutorial
Thu. at 14:00 pm

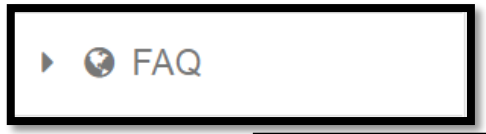
Exams

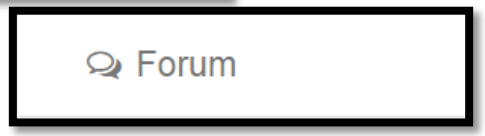
- Demo exam
  - **How?**: Online in Olat
  - **When?**: December 22 at 12:00 pm (Noon).
  - It covers the first eight lectures/assignments.
  - No eligibility is applied and no registration is required (Any one can take it).
- First exam (Official):
  - **How?**: not yet decided.
  - **When?**: February 23 at 10:00 am.
  - You must be eligible and register for the exam (see the calendar).
- Second exam (Official):
  - **How?**: similar to the first exam.
  - **When?**: March 23 at 10:00 am.
  - You must be eligible and register for the exam (see the calendar).

Exam eligibility

- To be eligible to write the exam, you need to:
  - Collect at least 720 points in 12 assignments.
  - Collect at least 40 points in each assignment.
    - This means that you need to submit all assignments in time.
  - If you were eligible in the WS 20/21, you may also write the exam without fulfilling the two previous conditions.

What else?

- Most of your questions are answered in ***FAQ*** in Olat. 

- If you do not find an answer, open a new topic in the Forum. 

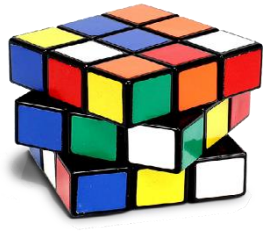
  - We will answer all the questions before every Tuesday and Friday at noon (This means that your question will be answered within 3 days at most.
  - Avoid sending emails.
- It is recommended to:
  - Add the schedule to your calendar.
  - Enable the notification on your calendar. 

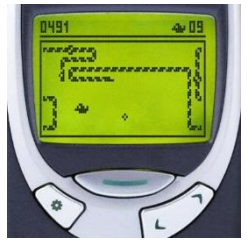
What do I need?

- The fundamental concepts of algebra
- The fundamental concepts of calculus
- The fundamental concepts of probability theory
- The fundamental concepts of statistics
- Programming skills (i.e. Python)

How to successfully pass the course?

- Watch the video lectures.
- Attend/watch the tutorials.
- Take notes.
- Solve the tasks of the assignments yourself.
- Do not forget to submit the assignment in time.
- Practice programming.
- Refresh and enhance your knowledge in Algebra and Probability theories.
- Prepare for the exam.
- Ask questions.
- Never plagiarize.





What is plagiarism?

- Turning in someone else's work as your own.
- Copying words or ideas from someone else without giving credit.
- Failing to put a quotation in quotation marks.
- Giving incorrect information about the source of a quotation.
- Changing words but copying the sentence structure of a source without giving credit.
- Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not (see our section on "fair use" rules).

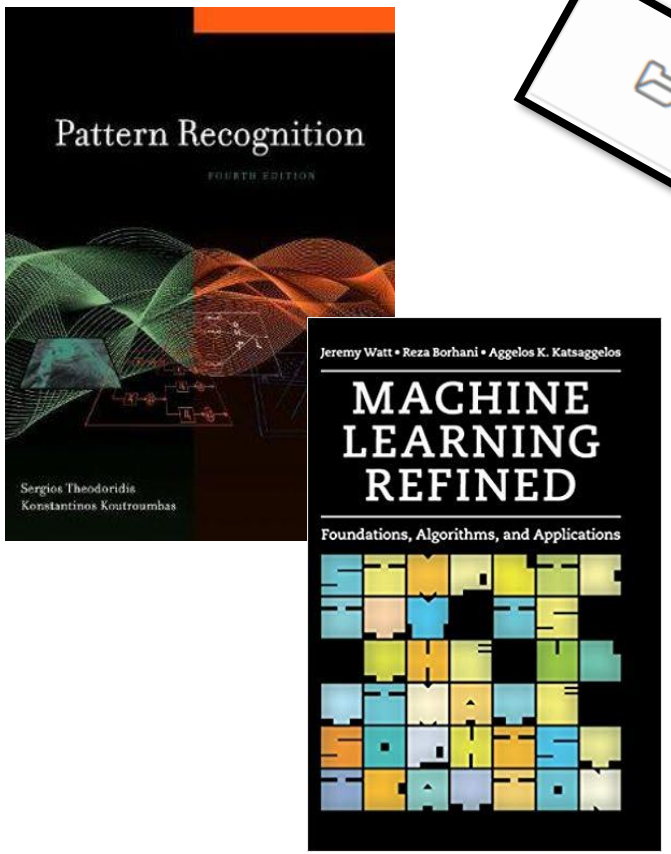
➢ Acknowledging that material has been borrowed and providing your audience with the information necessary to find that source is **usually** enough to prevent plagiarism.

Important to know

« **Any** sharing, copying or plagiarism of **any** exercise or assignment will lead to repeat the class next year for both students. »

We take this very seriously!

Books



In Olat

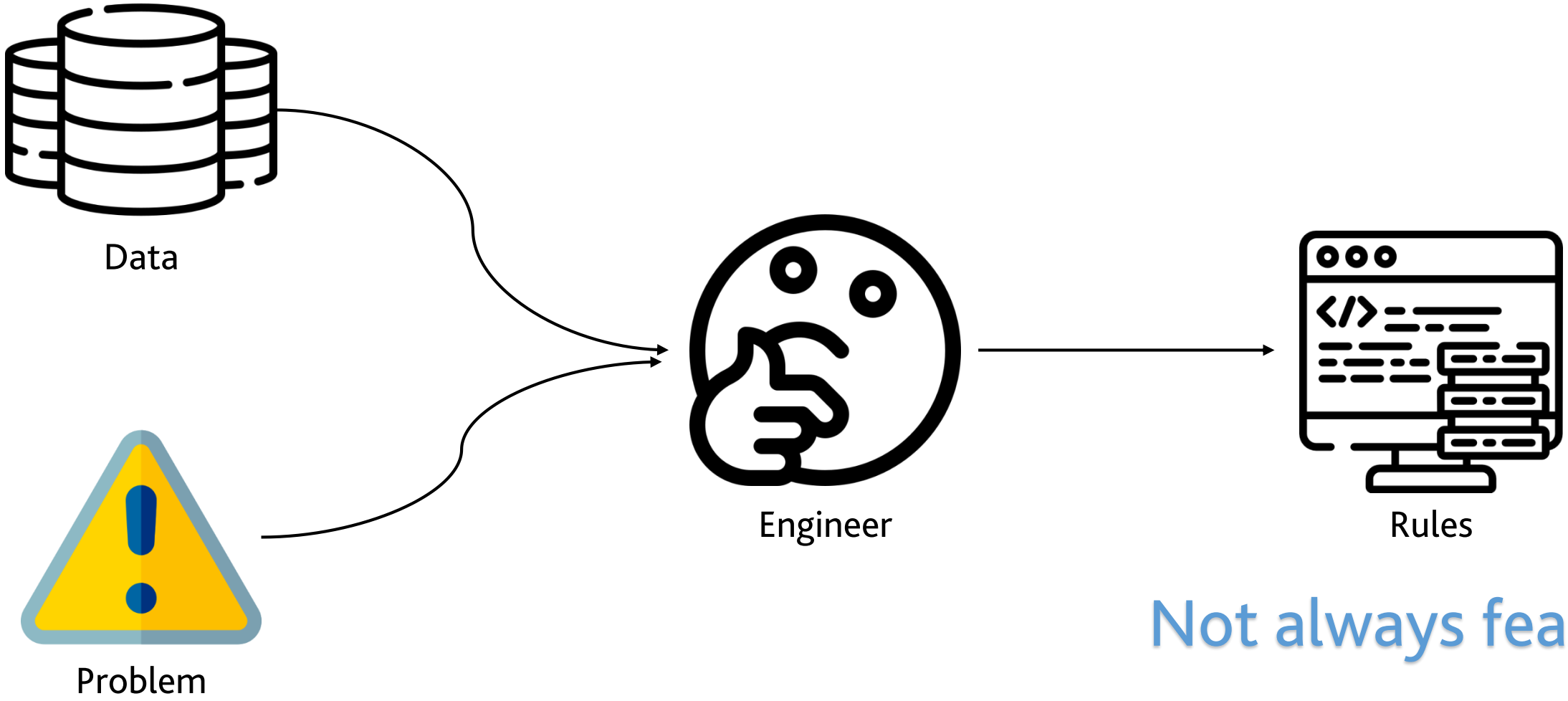
External materials

How to?

Screen Sharing

Let‘s get started!

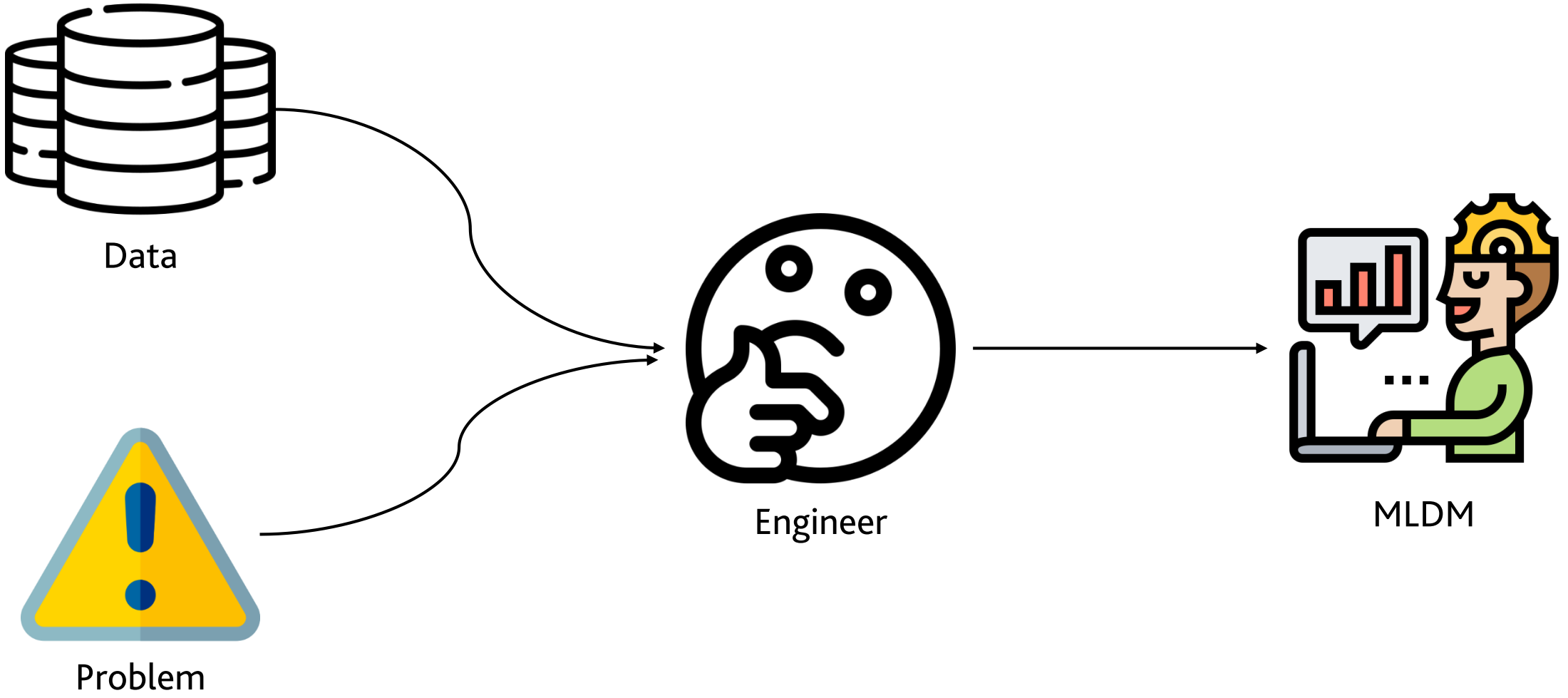
Why Machine Learning & Data Mining?



Not always feasible

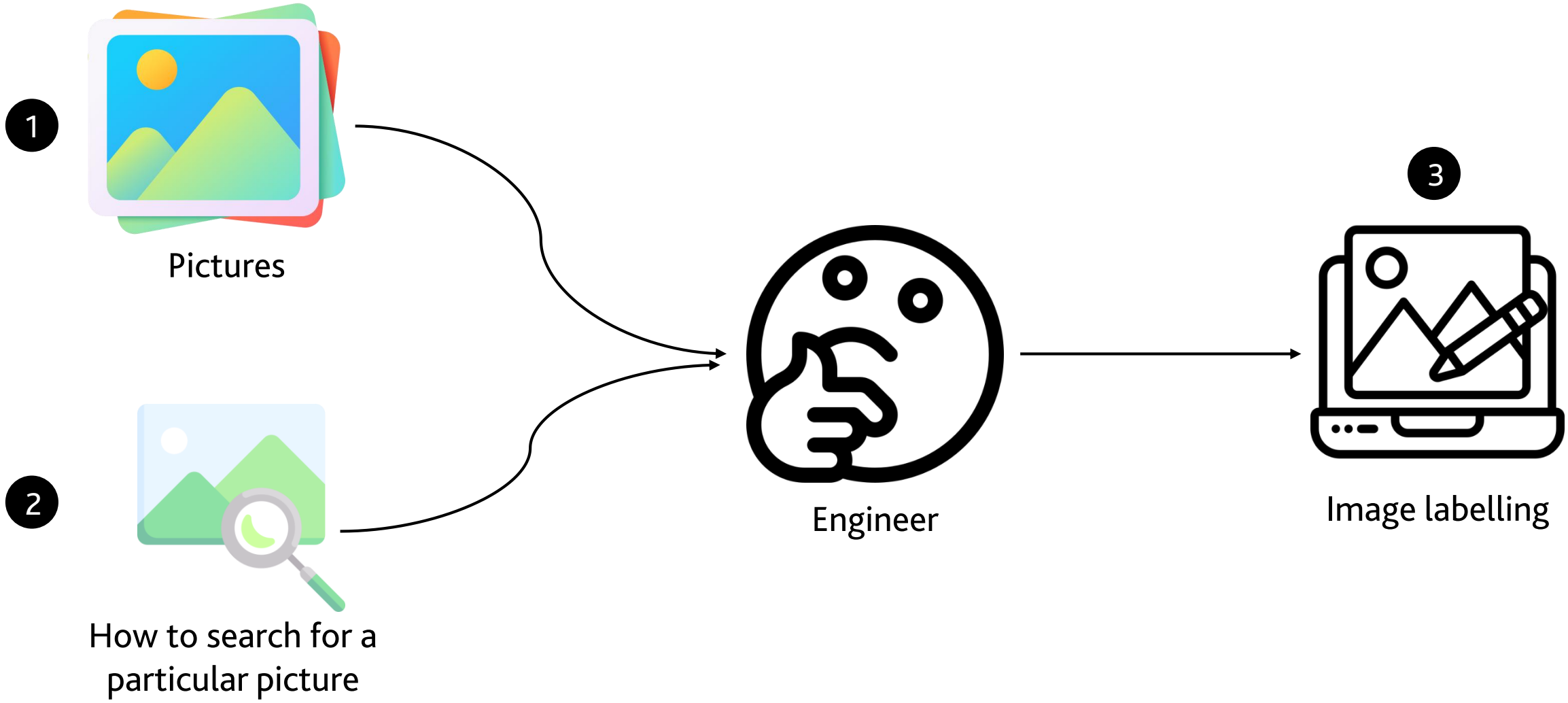
Icon vectors designed by Freepik

Why Machine Learning & Data Mining?



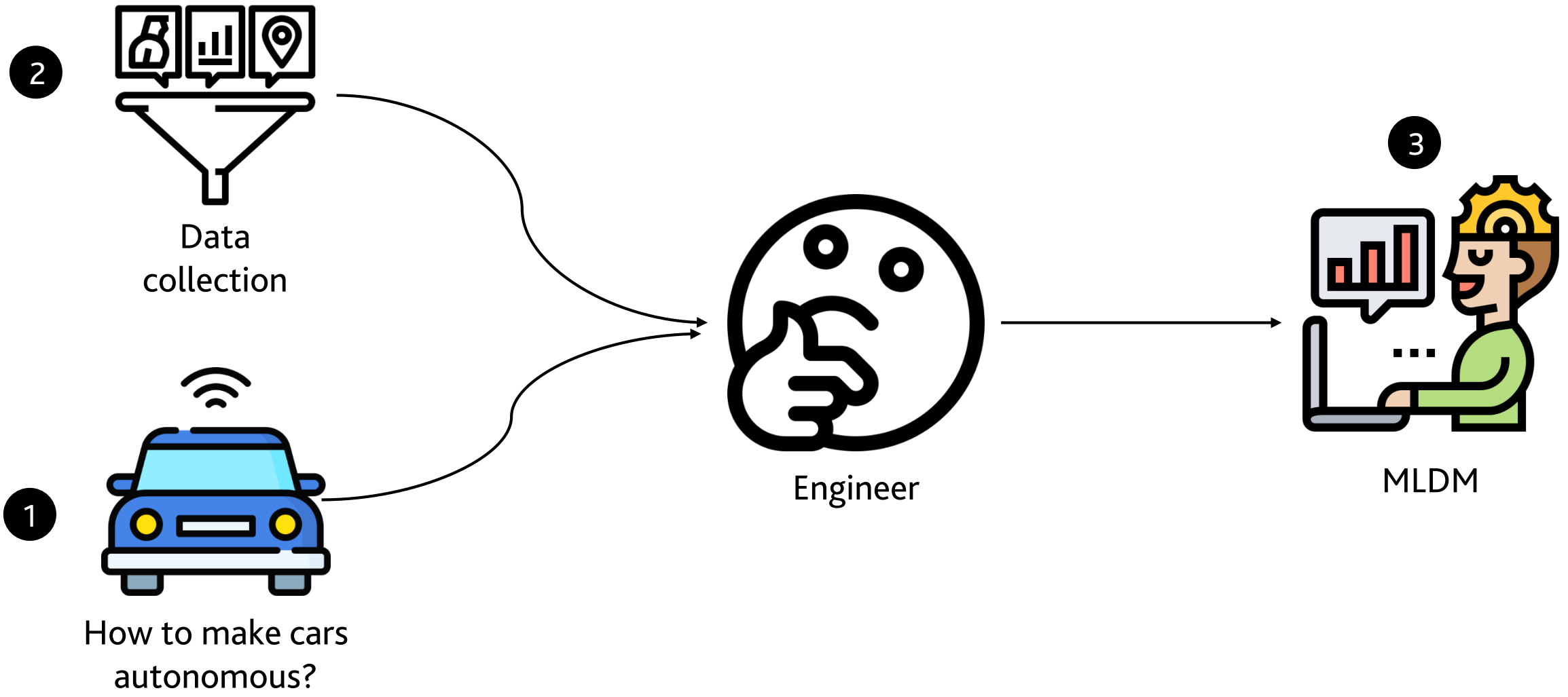
Icon vectors designed by Freepik

Why Machine Learning & Data Mining?



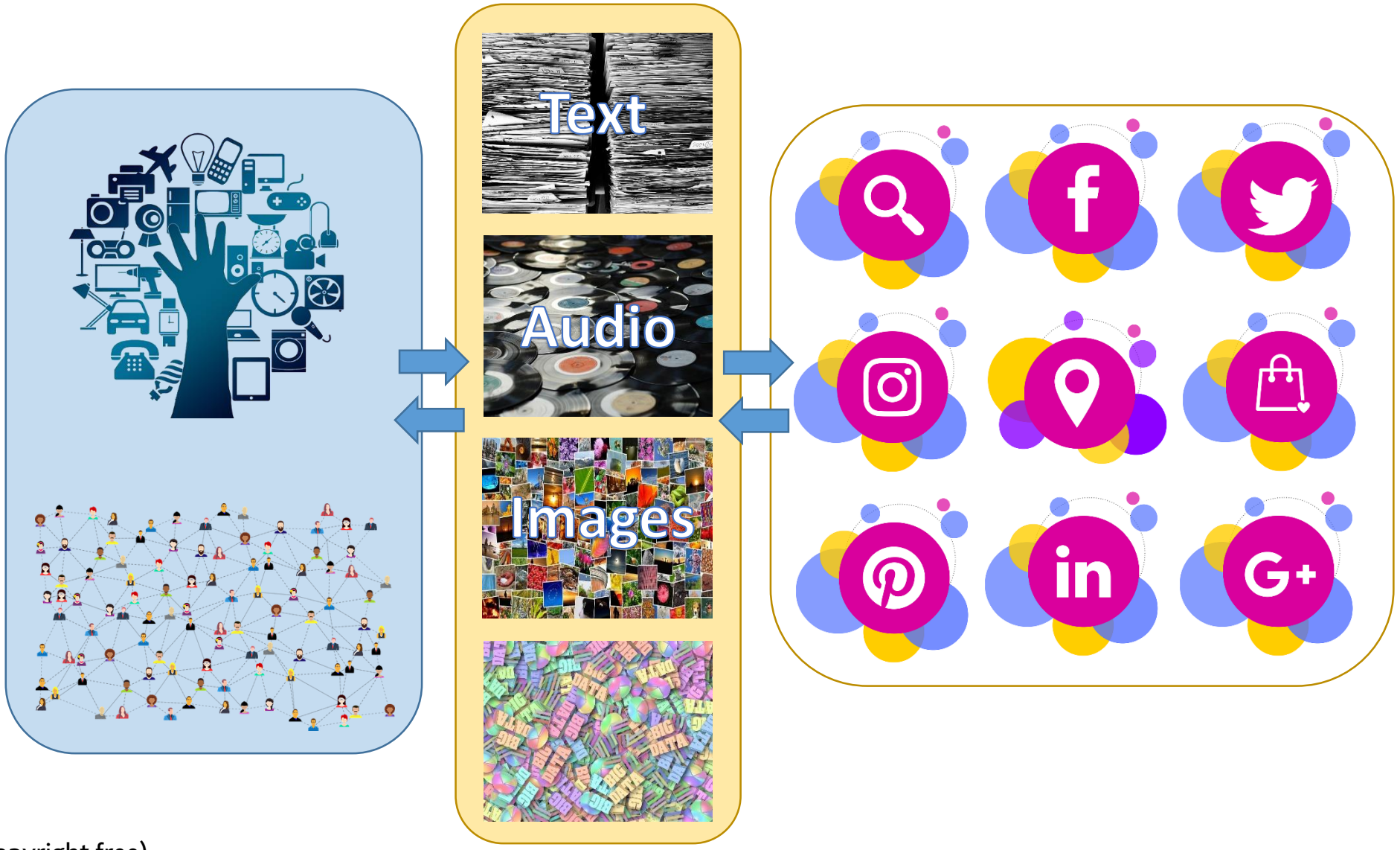
Icon vectors designed by Freepik

Why Machine Learning & Data Mining?



Icon vectors designed by Freepik

Why Machine Learning & Data Mining?



Images by Pixabay (copyright free)

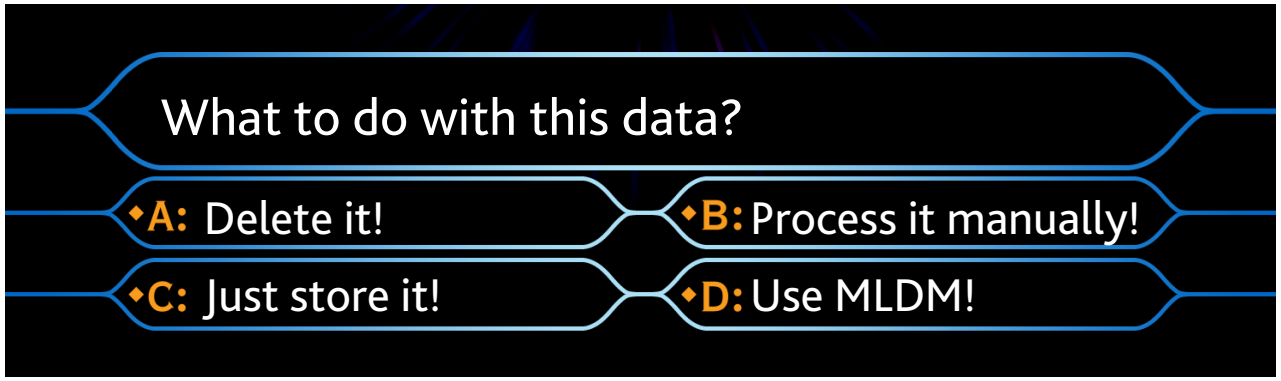
Why Machine Learning & Data Mining?

- We and our devices massively store and collect data in and from different platforms, e.g. Google, Facebook, amazon, etc.
  - 40,000 search queries are submitted every second on Google. [*]
  - 300 new hours of video show up on YouTube every minute. [*]
  - 100 terabytes of data are shared every day on Facebook. [*]
  - 5 quintillion bytes of data are produced every day by our smart devices. [*]

- Is that all?
  - **No**, There is also non-shared data (for example students' grades) and instantaneous/non-collected data (like videos in autonomous driving).

[*] https://hostingtribunal.com/blog/big-data-stats/

Why Machine Learning & Data Mining?



- Data Mining:
  - Helps human decision making by automatically extracting useful knowledge from a large amount of data.

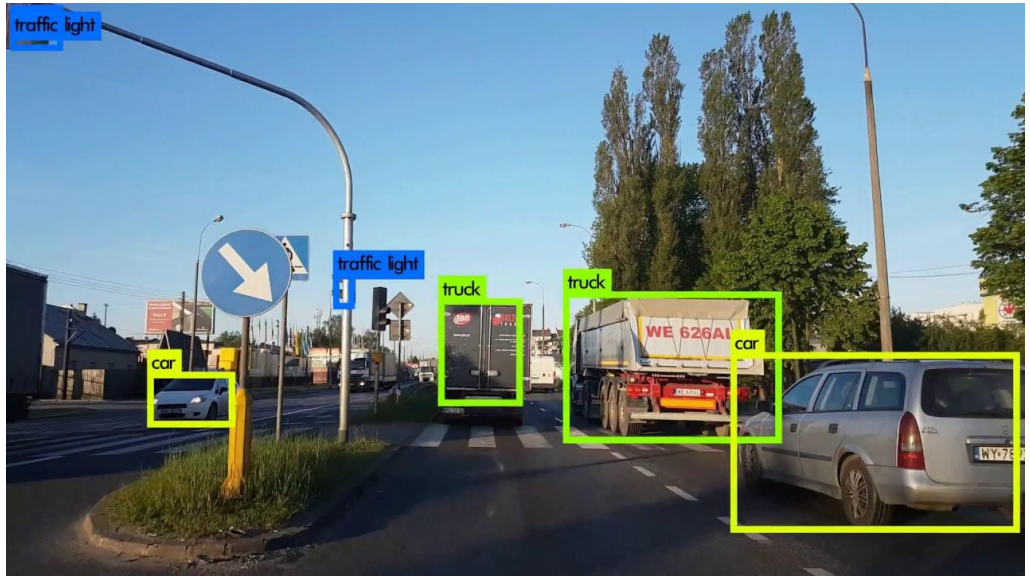
- Machine Learning:
  - Uses data to discover patterns and learn from them to make decisions that a human is capable for (e.g. autonomous driving) or that human cannot do (e.g. image generation, time series prediction).

Definition

Arthur Samuel (1959): *"Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed."*

Some of the applications of MLDM

Image Processing: Object Detection, Recognition & Localisation



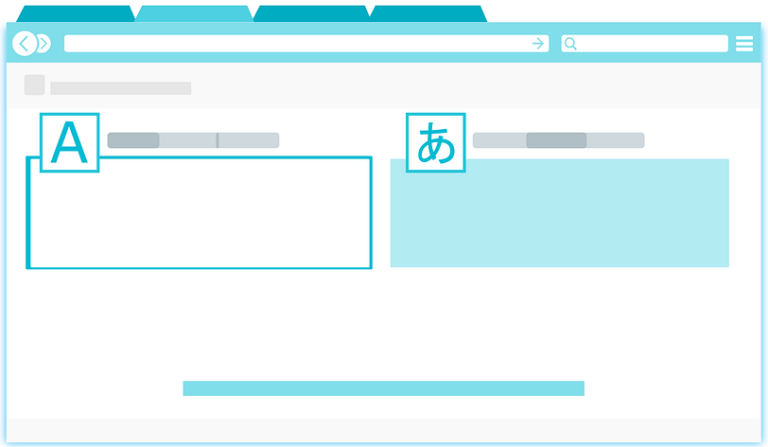
- Used in:
  - Autonomous driving
  - Image tagging and classification
  - Face recognition
  - Image annotation
  - Etc.

Natural Language Processing

- Used in:
  - Spam filtering.
  - Information extraction.
  - Text classification
  - Plagiarism detection.
  - And so many other applications

Images by Pixabay (copyright free)

Machine Translation



- Used because:
  - Real time.
  - Cheaper.
  - Easily accessible.

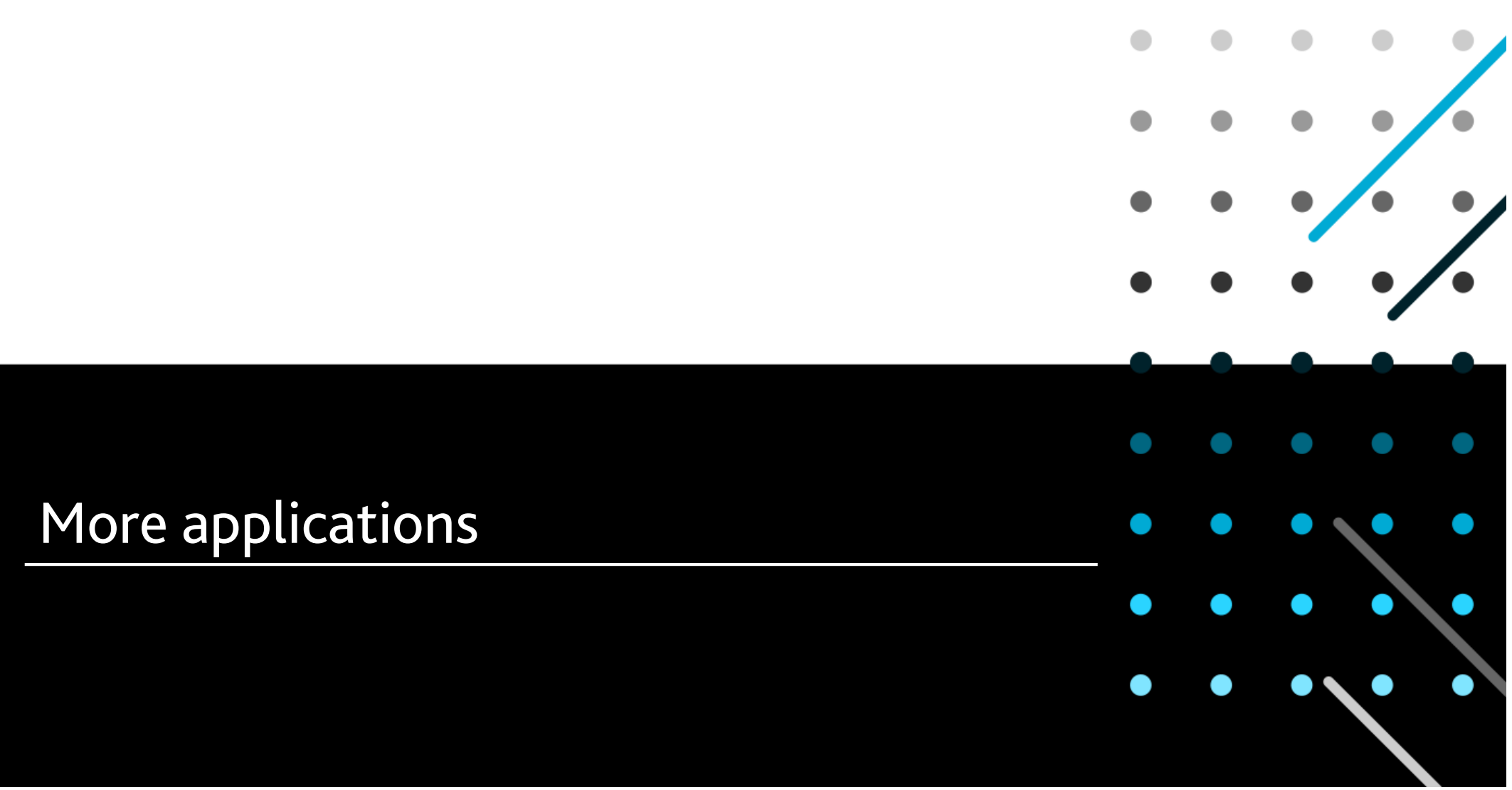
Images by Pixabay (copyright free)

Speech Recognition

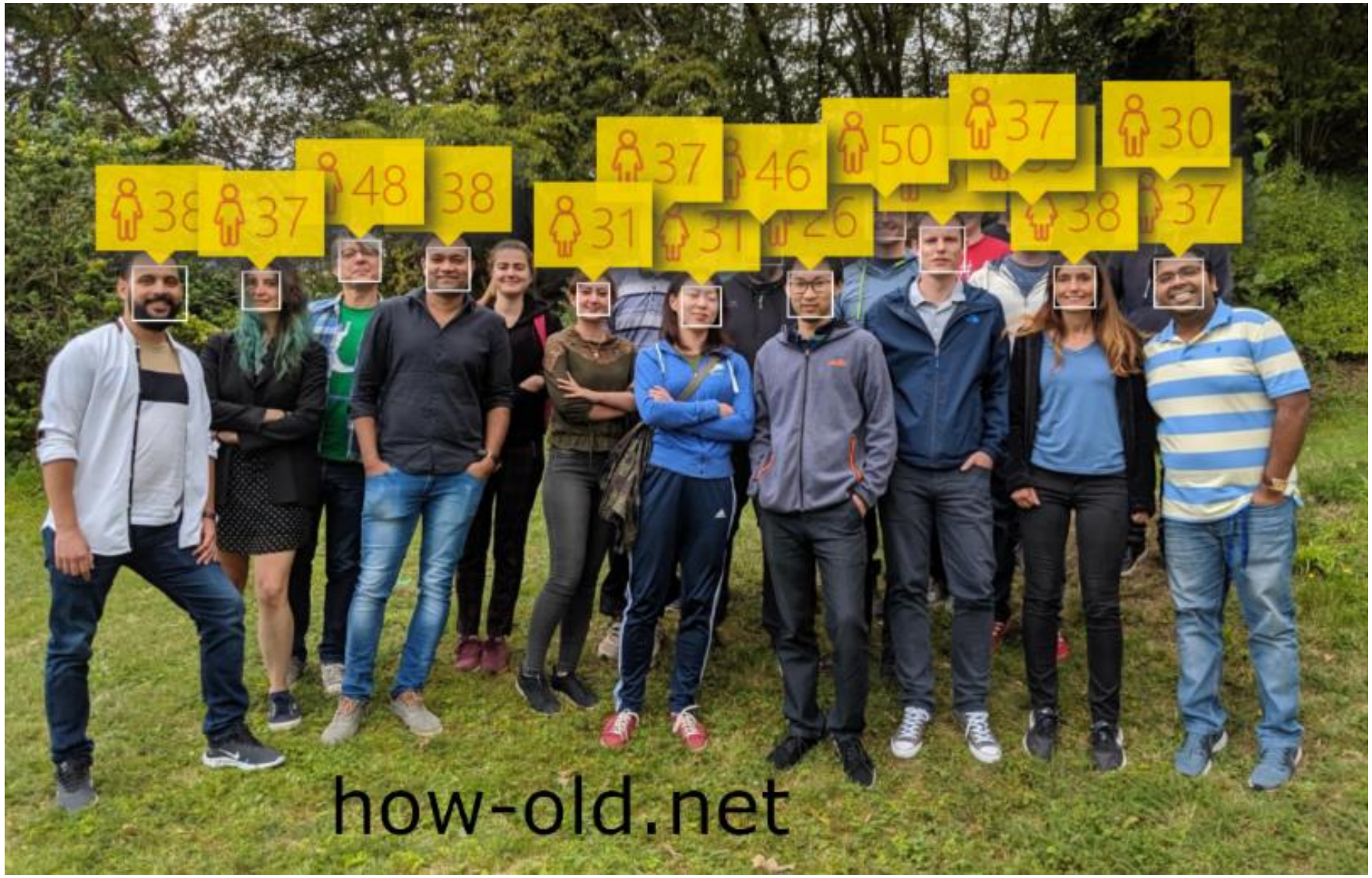
- Used in:
  - Easy and fast interaction with devices.
  - Important alternative for people with visual impairments.
  - Voice biometrics for identity recognition.
  - Effective communication with a big number of clients.
  - Etc.

Images by Pixabay (copyright free)

More applications

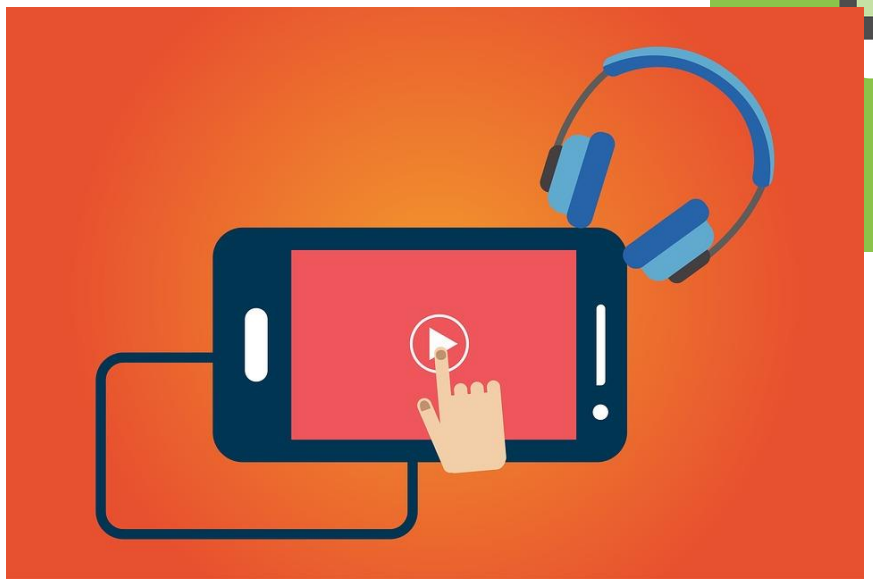


Age and gender detection



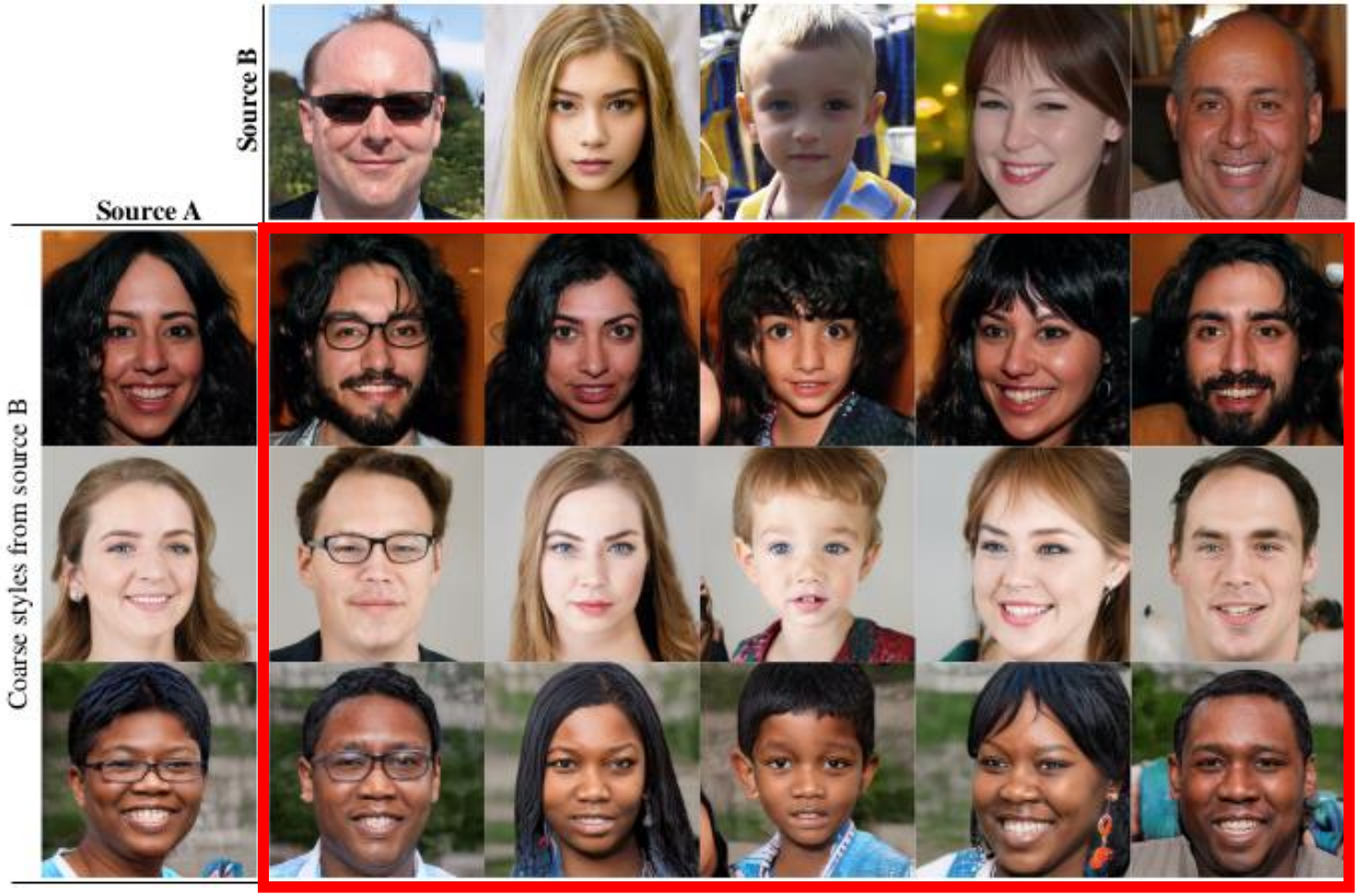
Recommendation system

- Ads
- Videos
- Products
- Etc.



Google Ads

Data Generation



FAKE!!

Karras, T., Laine, S., & Aila, T. "A style-based generator architecture for generative adversarial networks." In *CVPR, 2019.*

And there are even more in all domains . . .

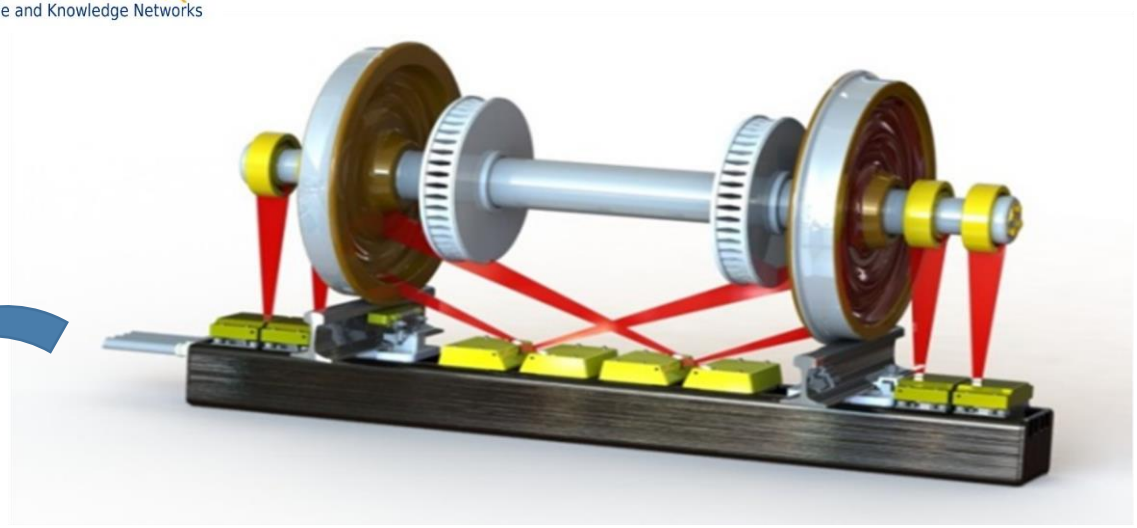
- Finance & Banking
- Security & Surveillance
- Transport & Traffic
- Healthcare & Well-being
- Media
- Commerce
- Sport
- Game & Entertainment
- Etc.

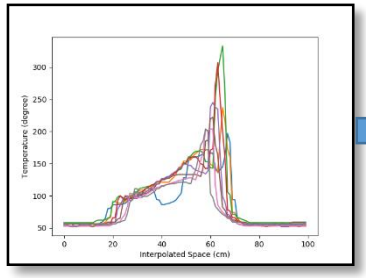
MLDM @ WeST

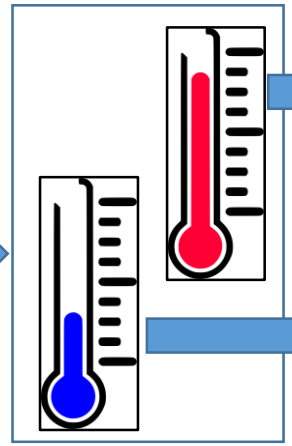
Example from past projects

Railway Monitoring System

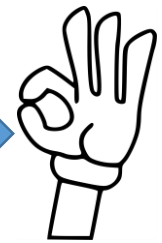
WeST
People and Knowledge Networks

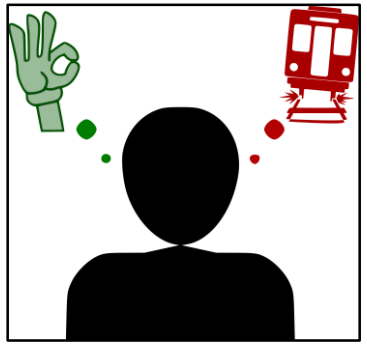




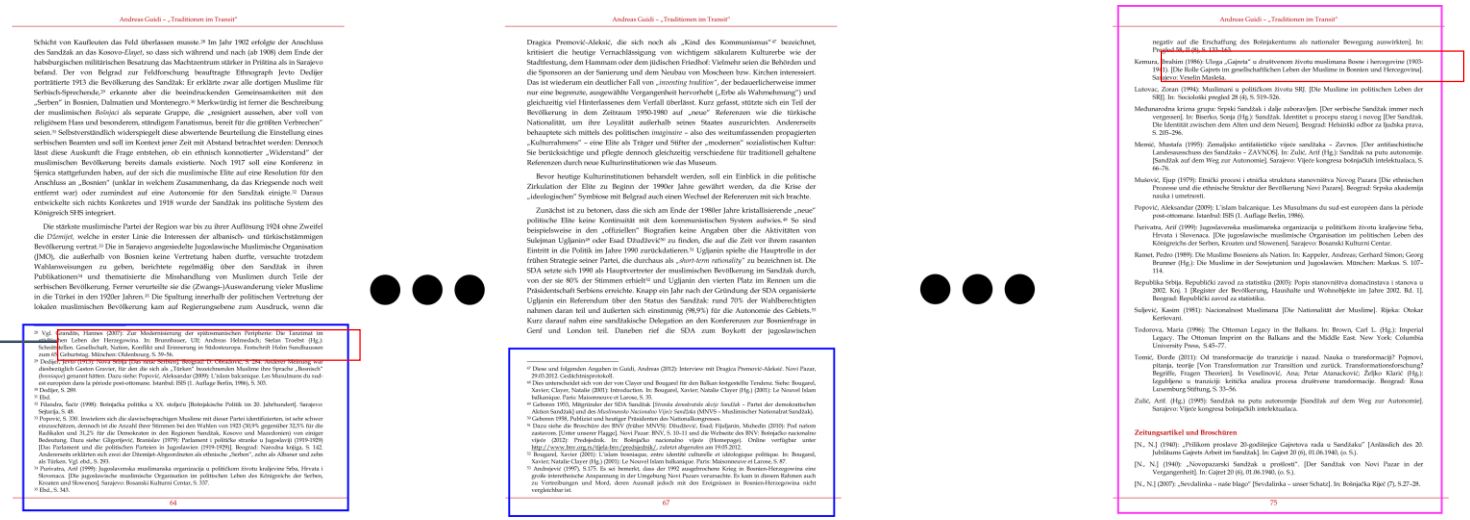


MLDM





Extracting and Parsing References



28 Vgl. Grandits, Hannes (2007): Zur Modernisierung der spätosmanischen Peripherie: Die Tanzimat im städtischen Leben der Herzegowina. In: Brunnbauer, Ulf; Andreas Helmedach; Stefan Troebst (Hg.): Schnittstellen. Gesellschaft, Nation, Konflikt und Erinnerung in Südosteuropa. Festschrift Holm Sundhaussen zum 65. Geburtstag. München: Oldenbourg. S. 39–56.

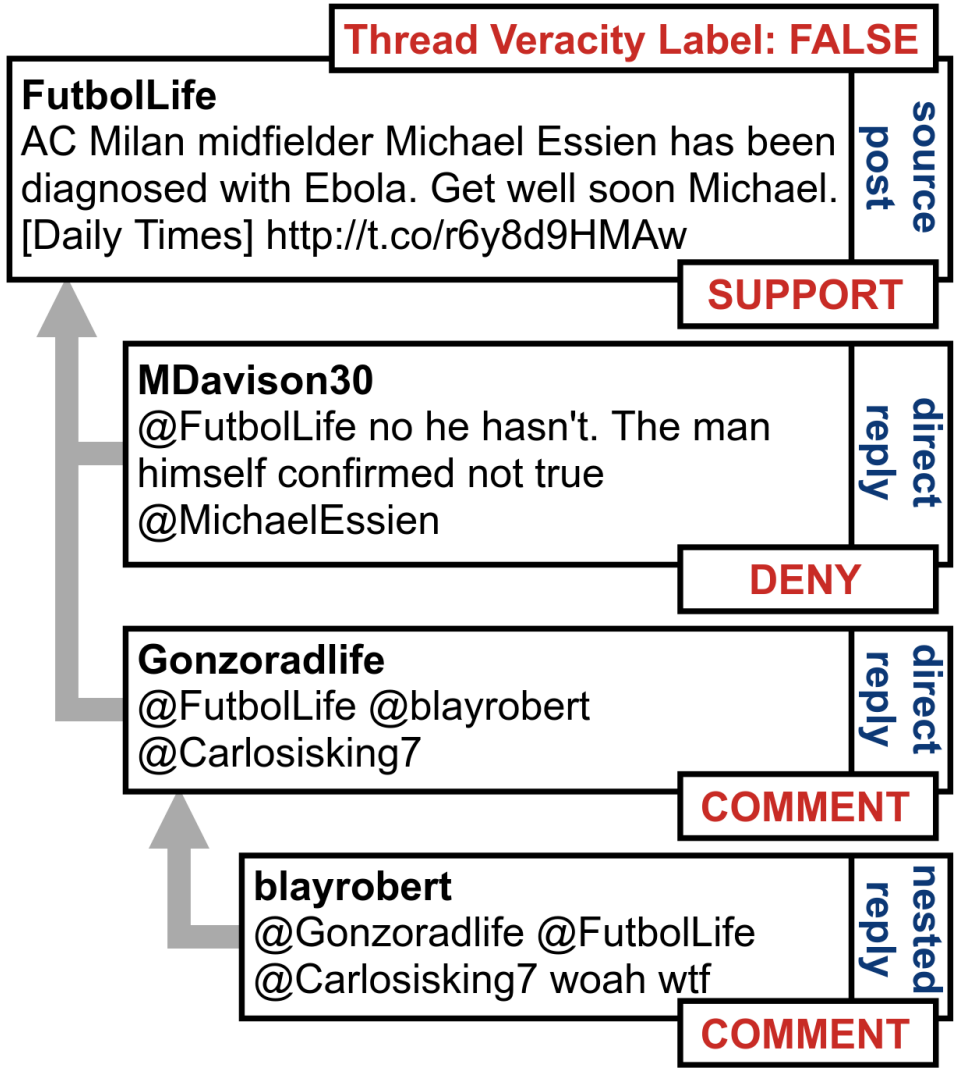
Author* Author** Title

19. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR01. (2001) I:511–518

Title Source Year Page range

Boukhers, Z., Ambhore, S., & Staab, S. (2019). An end-to-end approach for extracting and segmenting high-variance references from PDF documents. In *JCDL*

Rumour detection and stance classification



Baris, I., Schmelzeisen, L., & Staab, S. (2019). CLEARumor at SemEval-2019 Task 7: ConvoLving ELMo Against Rumors. *arXiv preprint arXiv:1904.03084*.

Predictive learning problems

Predictive learning problems

- "*They constitute the majority of tasks machine learning can be used to solve today.*"

Machine Learning

v.s.

Human Learning

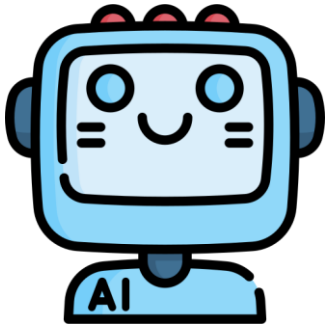
Classification

- **Human learning**

Teach a child how to distinguish between a “pear” and “apple”.

- **Machine learning**

Teach a computer to distinguish between a “pear image” and “apple image”.



Defining Task

Collecting Data

Designing Features

Training Model

Testing Model

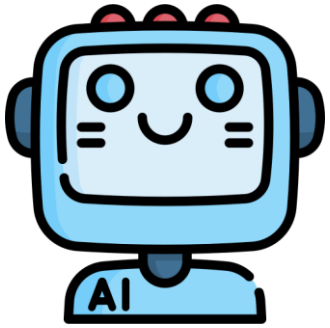
Evaluating Model

Icon vectors designed by Freepik

Classification

- **Human learning**
  - Give the child scientific definition « Genus: Malus, etc. »
  - the child is presented with one, few or many examples of what they are told are either “apples” or “pears” until he/she fully grasp the two concepts.

- **Machine learning**
  - Similarly, examples of “apples” and “pears” must be collected, labelled with their class names and presented to a computer.
  - The larger and more diverse the data set the better a computer can perform (applied also in human learning).



Defining Task

Collecting Data

Designing Features

Training Model

Testing Model

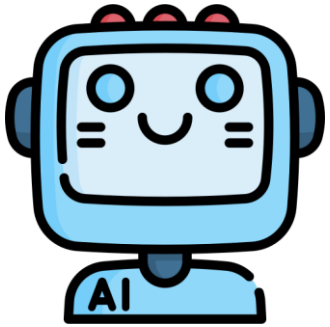
Evaluating Model

Icon vectors designed by Freepik

Classification

- **Human learning**
  - ~~Give the child scientific definition « Genus: Malus, etc. »~~
  - the child is presented with one, few or many examples of what they are told are either “apples” or “pears” until he/she fully grasp the two concepts.

- **Machine learning**
  - Similarly, examples of “apples” and “pears” must be collected, labelled with their class names and presented to a computer.
  - The larger and more diverse the data set the better a computer can perform (applied also in human learning).



Defining Task

Collecting Data

Designing Features

Training Model

Testing Model

Evaluating Model

Icon vectors designed by Freepik

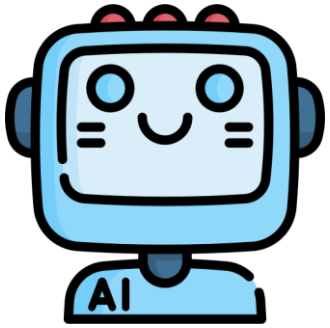
Classification

Images by Pixabay (copyright free)

Classification

- **Human learning**
  - How do we **visually** distinguish between an “apple” and “pear”?
  - Colour, shape, size, taste, etc.
  - Features must be discriminative.

- **Machine learning**
  - **Similarly**



Defining Task

Collecting Data

Designing Features

Training Model

Testing Model

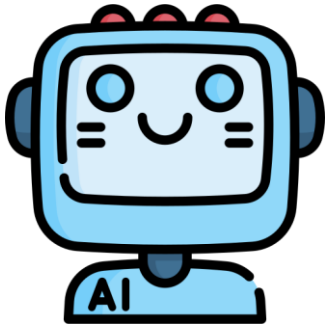
Evaluating Model

Icon vectors designed by Freepik

Classification

- **Human learning**
  - Show all examples to the child and explain to him the difference between the two classes.

- **Machine learning**
  - Find a line that separates between the two classes (Linear Classification). We will see it in details in the upcoming lectures.



- **Defining Task**
- **Collecting Data**
- **Designing Features**
- **Training Model**
- Testing Model
- Evaluating Model

Icon vectors designed by Freepik

Classification

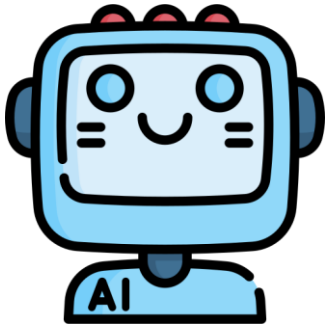
For a better visualisation, we consider only two features.

Images by Pixabay (copyright free)

Classification

- **Human learning**
  - Did the child understand the concept?
  - Show him/her **other** examples to confirm.
  - Don't give him/her the label of course!

- **Machine learning**
  - Let the trained model predict the classes of completely new example (not used in the training phase).

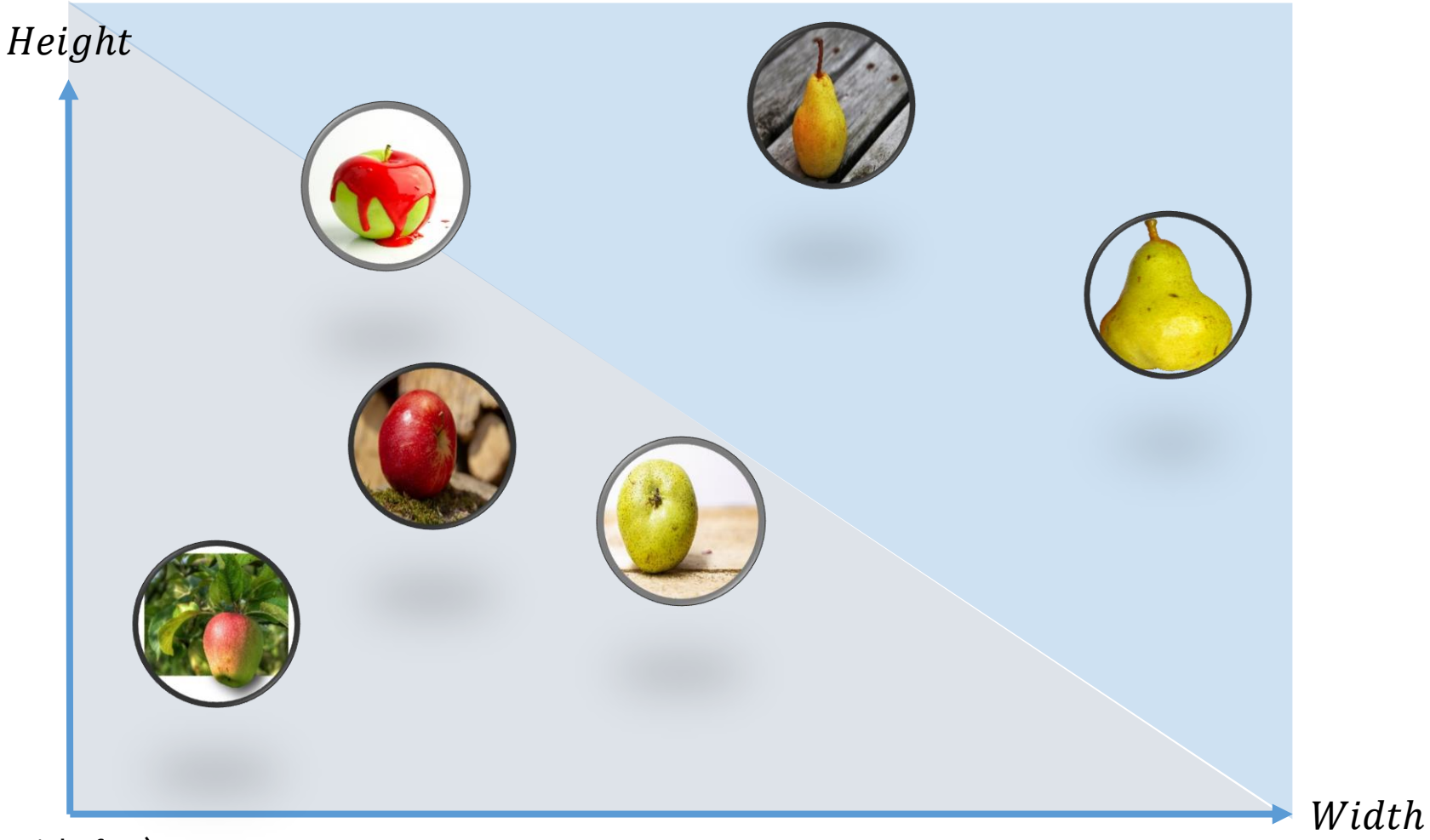


- Defining Task
- Collecting Data
- Designing Features
- Training Model
- Testing Model
- Evaluating Model

Icon vectors designed by Freepik

Classification

For a better visualisation, we consider only two features.



Images by Pixabay (copyright free)

Classification

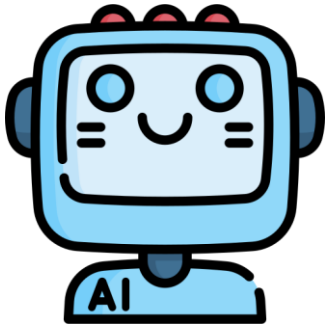
- **Human learning**

Bravo!

- **Machine learning**
  - How good is the model?

Confusion Matrix, Precision, Recall, Accuracy, F-Score, etc.

(we will see them in the upcoming lectures)



Defining Task

Collecting Data

Designing Features

Training Model

Testing Model

Evaluating Model

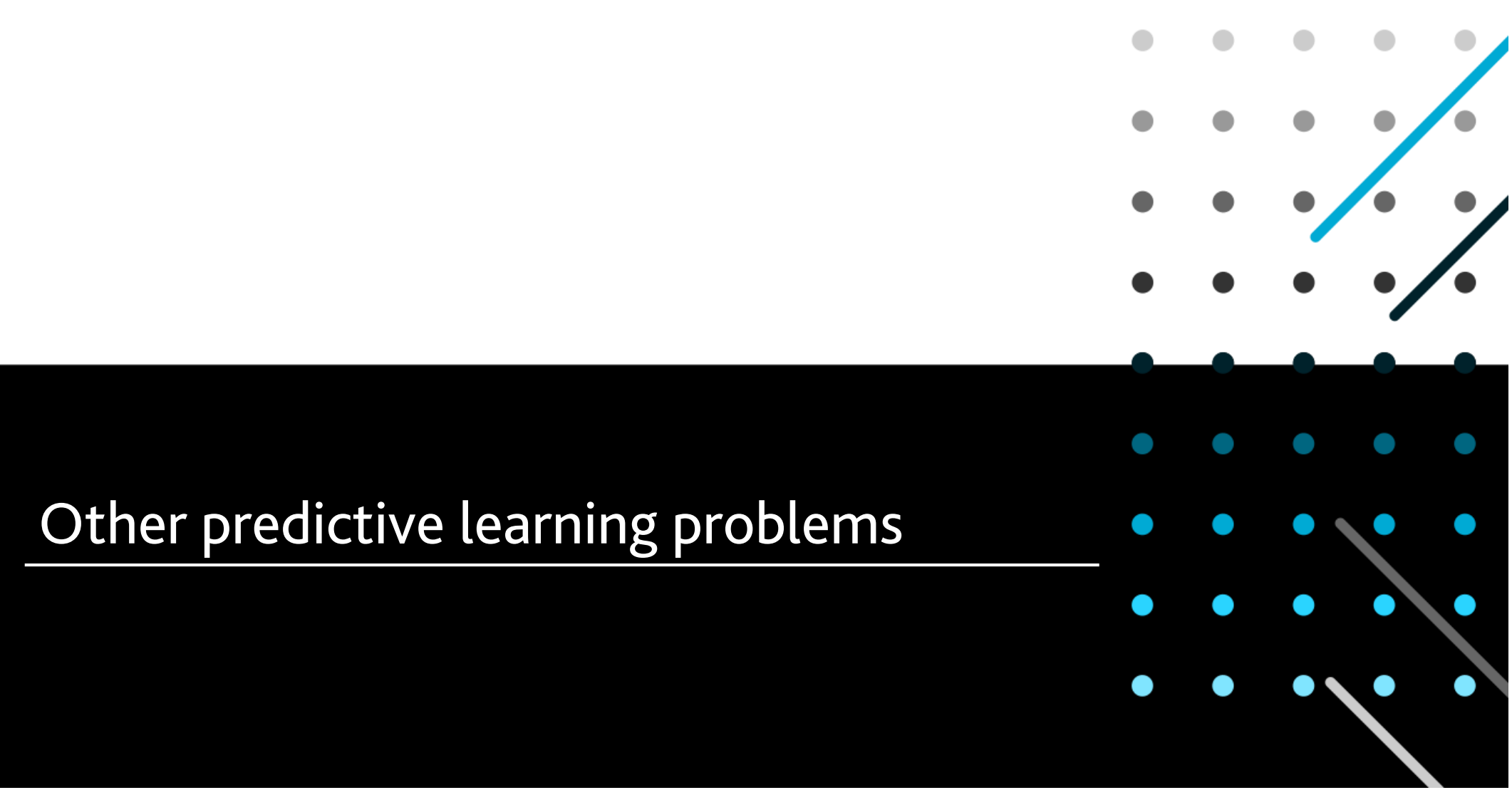
Icon vectors designed by Freepik

More examples of classification

- Is the skin mole a malignant melanoma?
  - False / True
    - Binary classification.
    - Two-class classification problem.
    - Categorical (discrete) classification.

- What is the topic of a given document?
  - Sport / Politics / Events / Society
    - Multi-class classification problem.
    - Categorical

Other predictive learning problems



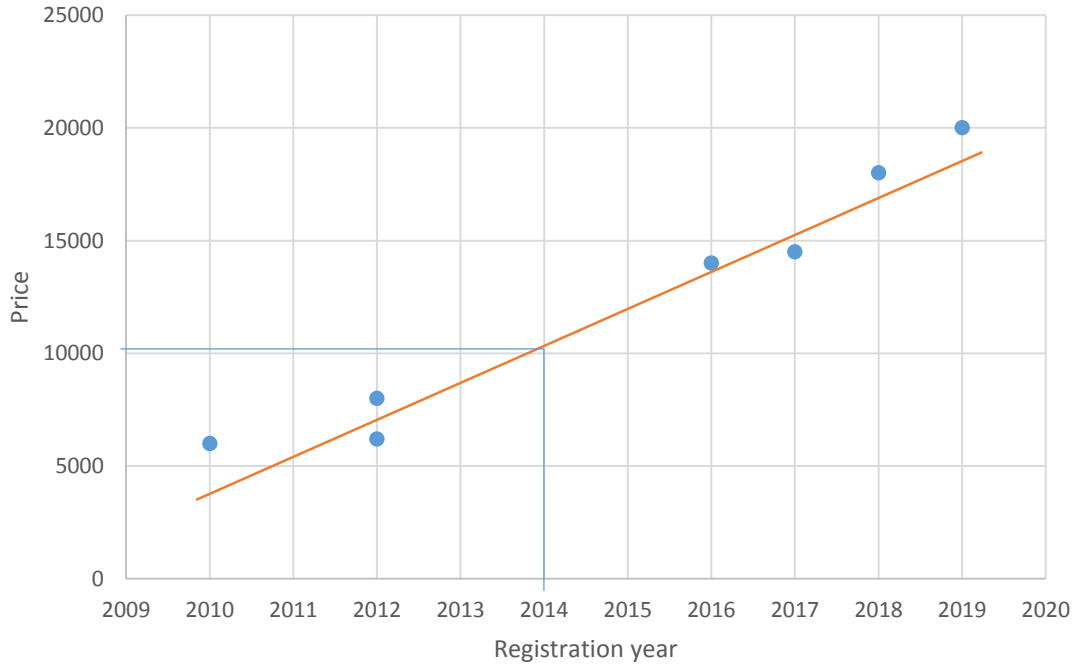
Regression

- We want to predict the price of our car based on the registration year, say 2014.
  - Obviously, all other properties (e.g. model, colour, etc.) are known.
  - We could obtain some examples of the same car from e-commerce websites:

How much is our car worth?

Registration Year	Price in €
2010	6,000
2012	7,200
2012	8,000
2016	14,000
2017	14,500
2018	18,000
2019	20,000

Regression

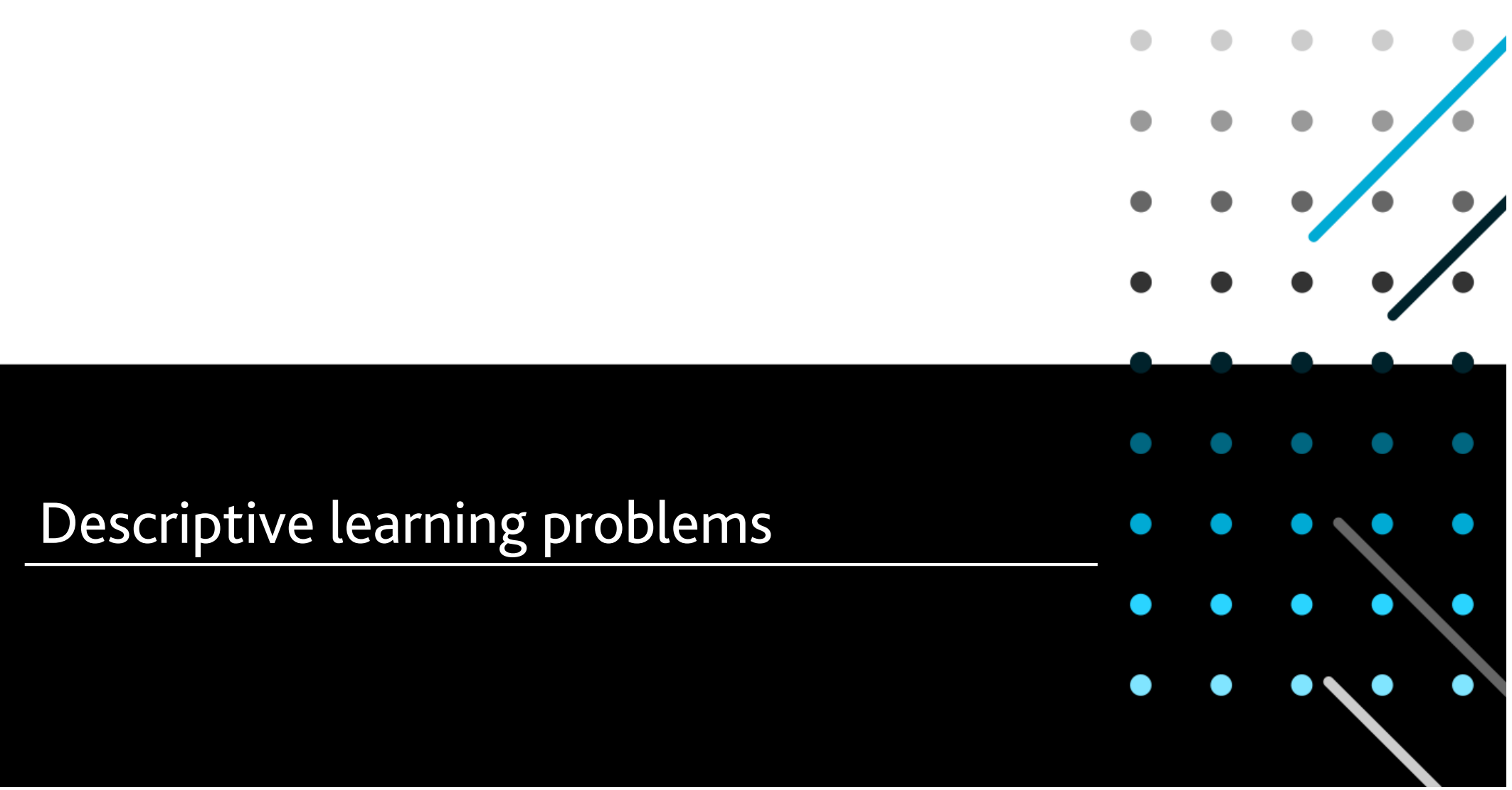


- Based on this data, our car should cost around 10,500 €.

More examples of regression

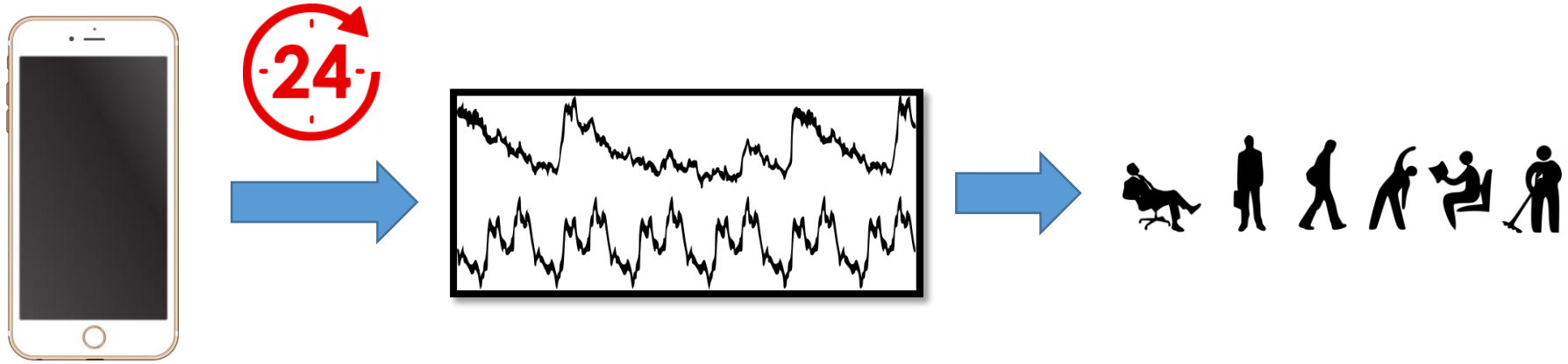
- What is the height given the age?
  - [45, 210] cm
    - Continuous prediction.
    - Called regression.

Descriptive learning problems



Clustering

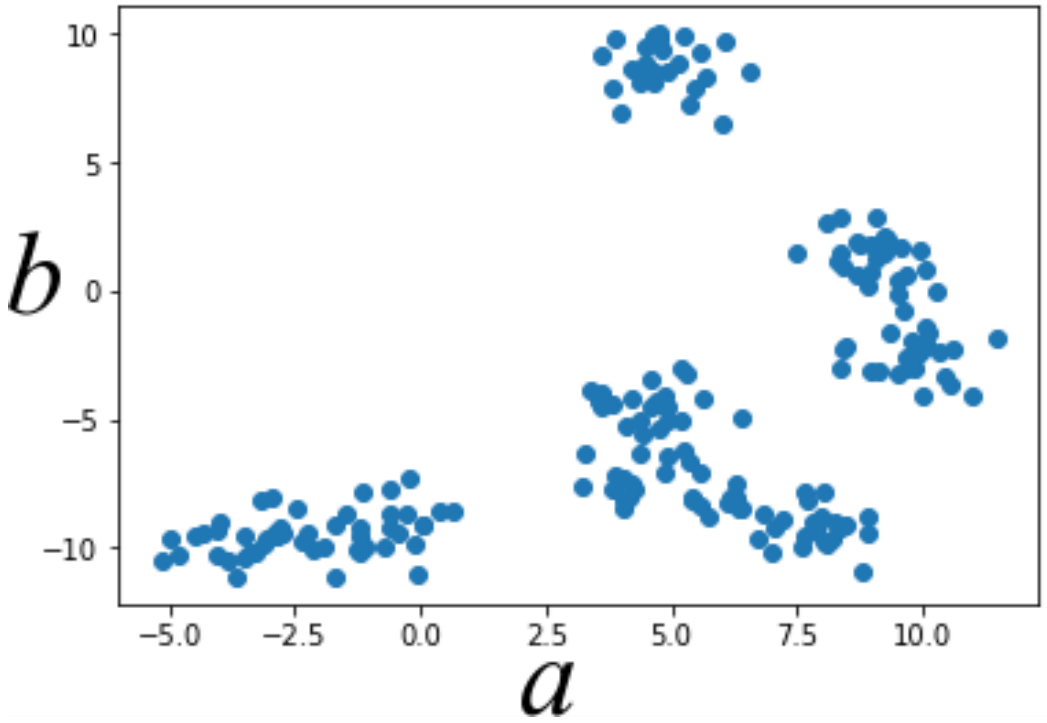
- Considering the following problem:
  - Extracting the different pattern activities that a person X performs in 24 hours using sensory data acquired from his/her smartphone per 15min.
  - Note that we don't have the labels of the activities.
  - We might as we might not know the number of activities.



Images by Pixabay (copyright free) + Icon vectors designed by Freepik

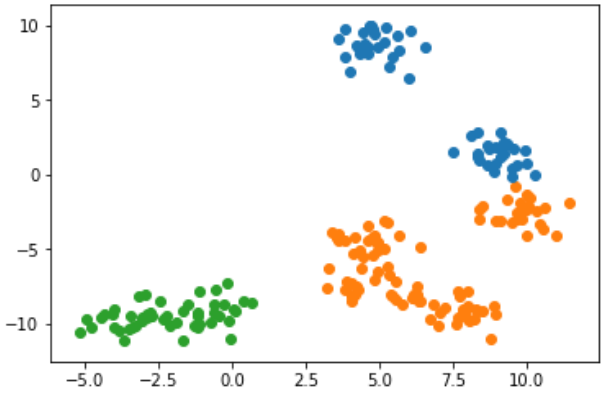
Clustering

- **For the sake of visualization**, the data (measurements) are represented by two features (a and b).

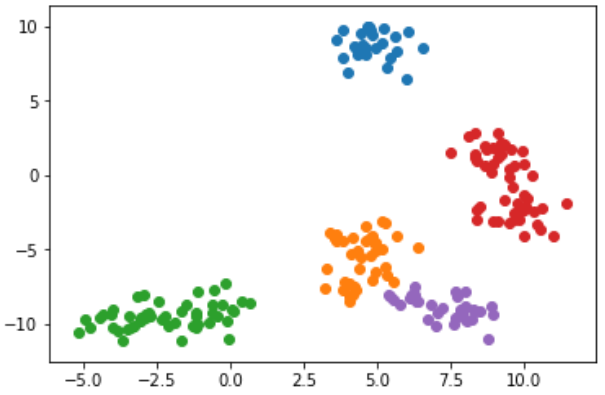


Clustering

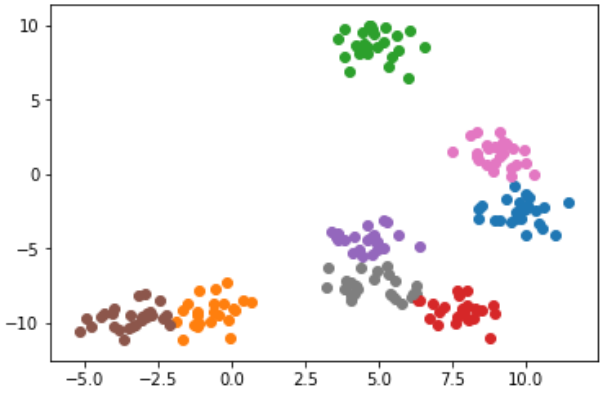
N. Clusters = 3



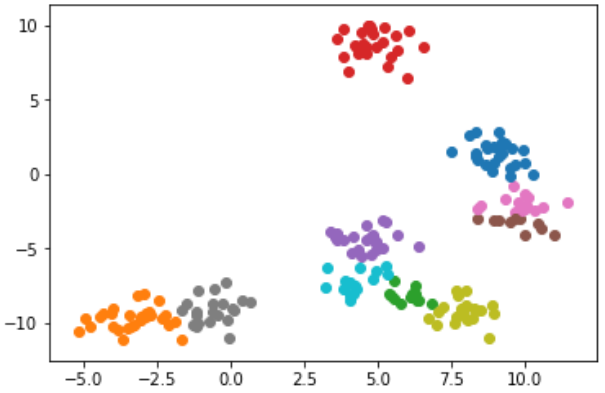
N. Clusters = 5



N. Clusters = 8



N. Clusters = 10



Some notions

Predictive vs. Descriptive learning

- **Predictive learning** is the task of learning from existing data to predict something about new data, such as:
  - Classifying it based on its classes.
  - Estimating the value of a propriety (regression).
  - Forecasting upcoming states.
  - Etc.
- **Descriptive learning** is the task of learning something about the data at hand, such as:
  - Discovering latent patterns (clustering).
  - Detecting anomalies.
  - Ranking.

Supervised vs. Unsupervised learning

- Supervised learning:
  - Labelled training data
  - Predict the label of new data
- Unsupervised learning:
  - Non-labelled data.
  - Capture the hidden structure in data.

- Although most of predictive tasks are supervised and most of the descriptive tasks are unsupervised,
  - Predictive learning ≠ Supervised learning
  - Descriptive learning ≠ Unsupervised learning

(Un-) Supervised vs. Predictive/Descriptive

	Predictive	Descriptive
Supervised	• The model uses past data to predict events or outcome in future data. • The model is trained with labelled data.	• The model extracts knowledge from the data at hand. • The model is trained with labelled data.
Unsupervised	• The model uses past data to predict events or outcome in future data. • The model is fed with unlabelled data only.	• The model extracts knowledge from the data at hand. • The model is fed with unlabelled data only.

Examples

	Predictive	Descriptive
Supervised	• Object classification • Object detection • Object extraction • Etc.	• I couldn't think of any example!
Unsupervised	• Anomaly detection (e.g. Bank transactions, Human activities, etc.)	• Image segmentation • Topic modelling • Clustering • Etc.

Other learning approaches

Semi-supervised learning

- It combines a small labelled data with a large unlabelled data.
- Commonly, adopted when collecting large labelled data is difficult and/or costly. For example:
  - Human activities
  - Speech analysis
  - Web pages classification
  - Biological sequence classification
  - Note that difficulty and cost are subjective.
- Example methods:
  - Positive Unlabelled (PU) learning
    - e.g. Suspicious activity detection.
  - Combining clustering & classification
    - Cluster → Label → Train/Classify
    - Label → Cluster → Train/Classify

Reinforcement learning

- It is goal-oriented that learns how to achieve a complex goal.
  - For every state, it takes an action.
  - The series of actions is supposed to reach the goal.
  - The data is unlabelled.
  - It learns by exploration and exploitation.
    - Exploration: try different sequence of actions.
    - Exploitation: take advantage of the actions/states that led good result.
- Examples of usage:
  - Games
  - Robotics
  - Trading
  - Etc.

Active Learning

- The model is trained with a small labelled data
- Then, It queries a user/expert or any information source to label some samples.
- It is mainly applied when:
  - The unlabelled data is abundant
  - But, the labelling is costly.
- Examples:
  - URL classification
  - Image classification.
  - Etc.

Summary

Summary

- Organization:
  - It is important to stick to the deadlines mentioned in the calendar.
  - Being registered in Olat & Microsoft teams and Panopto is all what you need for this course.

- MLDM:
  - MLDM is very important.
  - A lot of disciplines are benefiting from MLDM.
  - There are several tasks in MLDM: Supervised, Unsupervised, …
  - There are several techniques in MDLM: Predictive, descriptive, …

Thank you!

Zeyd Boukhers

E-mail: Boukhers@uni-koblenz.de
Phone: +49 (0) 261 287-2765
Web: Zeyd.Boukhers.com

University of Koblenz-Landau
Universitätsstr. 1
56070 Koblenz