

Machine Learning and Data Mining WS21/22

“1 Introduction”

Dr. Zeyd Boukchers

@ZBoukchers

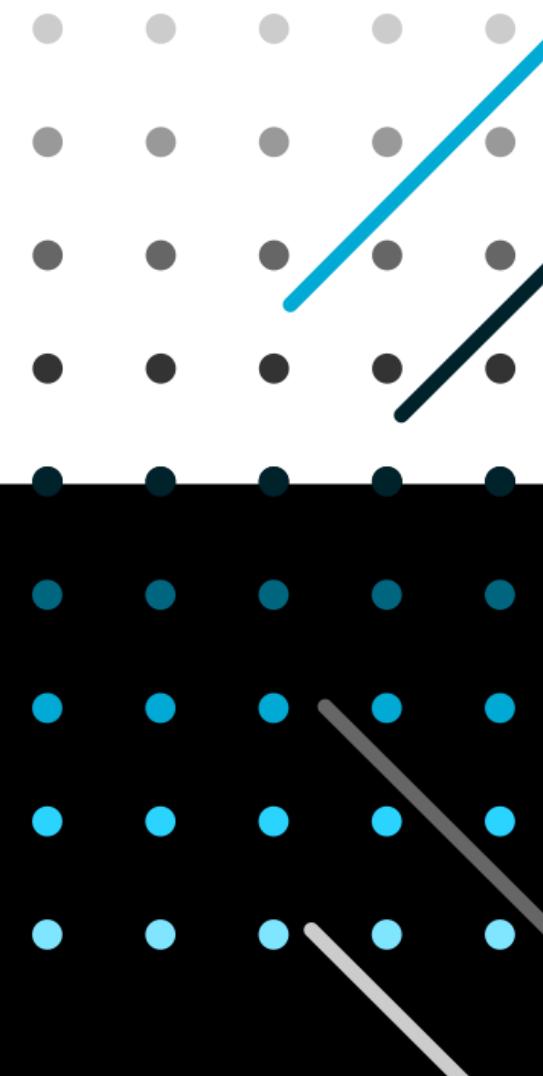
Institute for Web Science and Technologies

University of Koblenz-Landau

October 27, 2021

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

Organizational Stuff





- 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



- **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

 Calendar

- **Consultation:**

- Forum in Olat

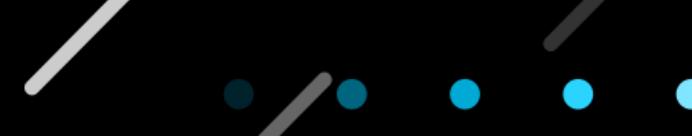
 Forum

- **Material and further information:**

- Olat: <https://olat.vcrp.de/url/RepositoryEntry/3382739338>

- **Klips Link:**

- <https://klips.uni-koblenz-landau.de/v/139109>

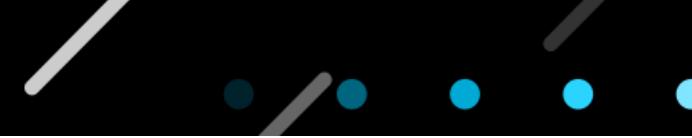


- Regulation:

- 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.



Lecturer & Tutor



Zeyd Boukher

Assistants



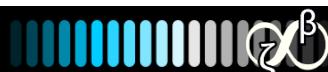
Akshaya Raja



Sowjanya Chennamaneni



Azeddine Bouabdallah



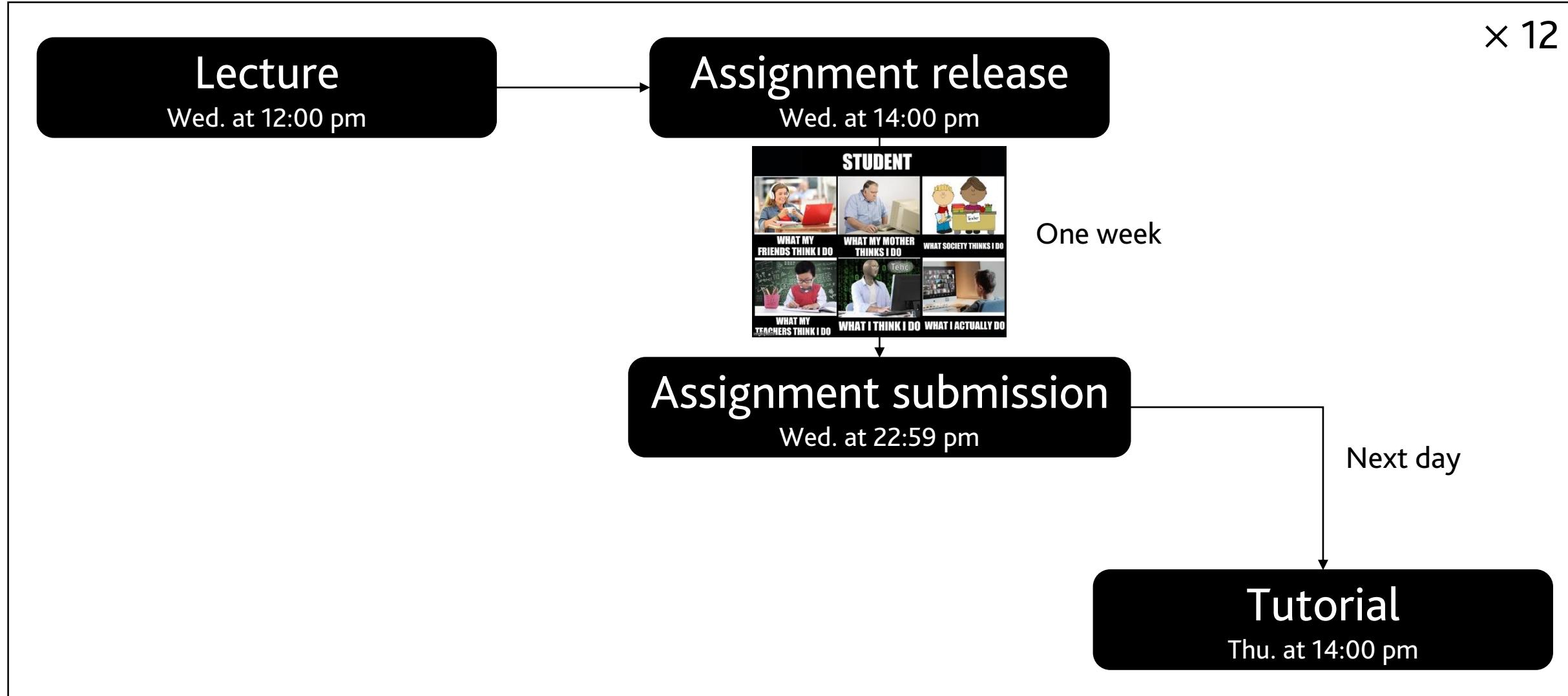
Assignments



- 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.

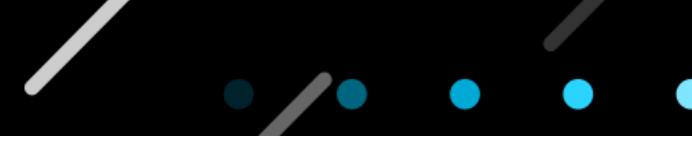


× 12



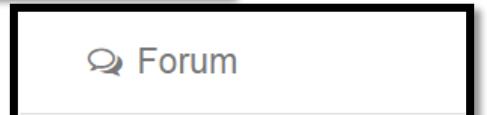


- Demo exam
 - How?: Online in Olat
 - When?: December 22 at 12:00 pm (Noon).
 - It covers the first eight lectures/assignments.
 - No eligibility is applied and no registration is required (Any one can take it).
- First exam (Official):
 - How?: not yet decided.
 - When?: February 23 at 10:00 am.
 - 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).



- 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.



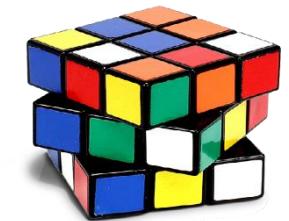
- Most of your questions are answered in *FAQ* in Olat. 
- If you do not find an answer, open a new topic in the Forum. 
 - We will answer all the questions before every Tuesday and Friday at noon (This means that your question will be answered within 3 days at most).
 - Avoid sending emails.
- It is 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?

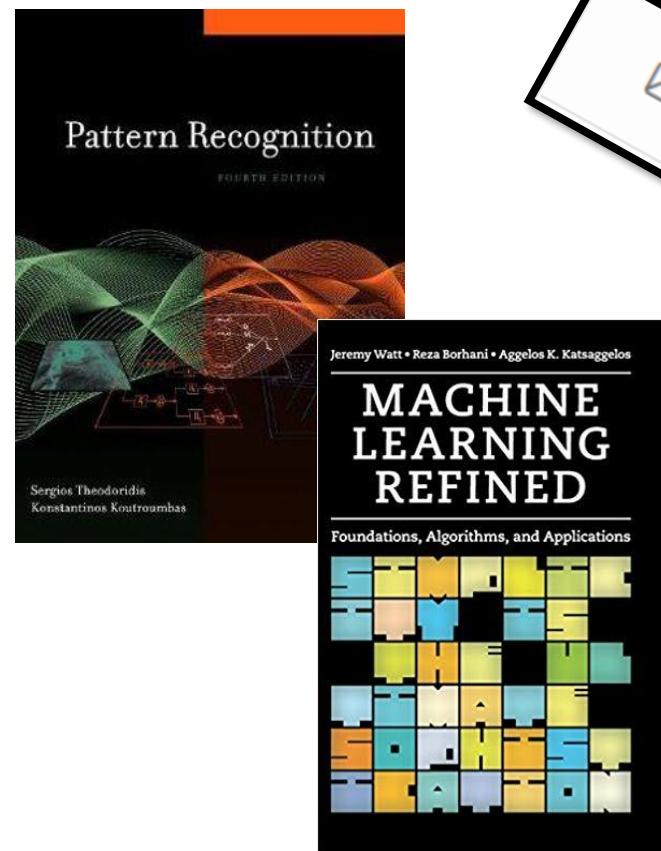
- Turning in someone else's work as your own.
 - Copying words or ideas from someone else without giving credit.
 - Failing to put a quotation in quotation marks.
 - Giving incorrect information about the source of a quotation.
 - Changing words but copying the sentence structure of a source without 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.



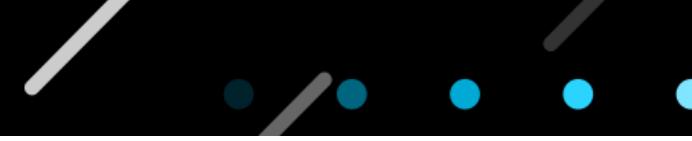
« Any sharing, copying or plagiarism of any exercise or assignment will lead to repeat the class next year for both students. »

We take this very seriously!

Books

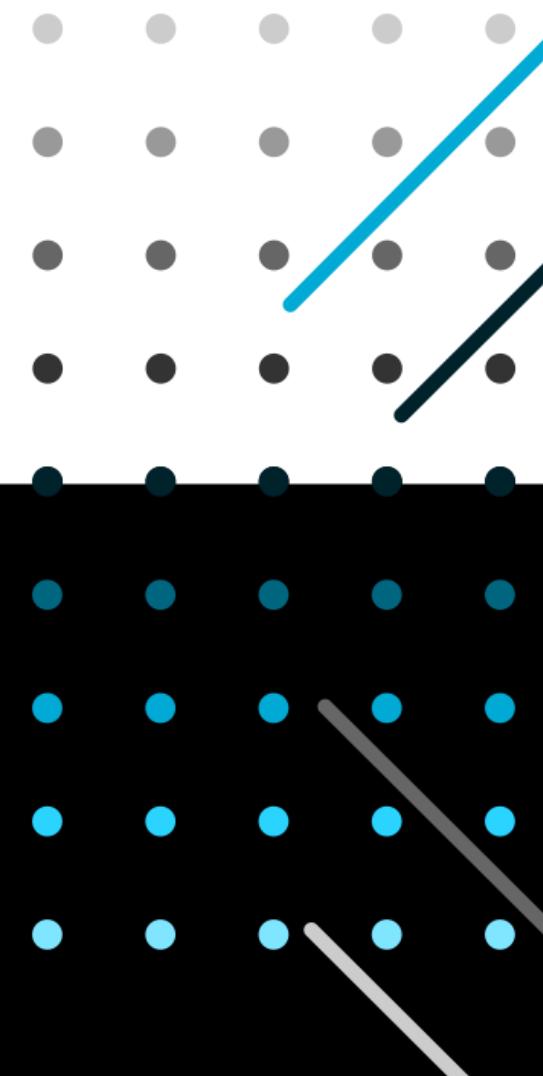


In Olat
External materials

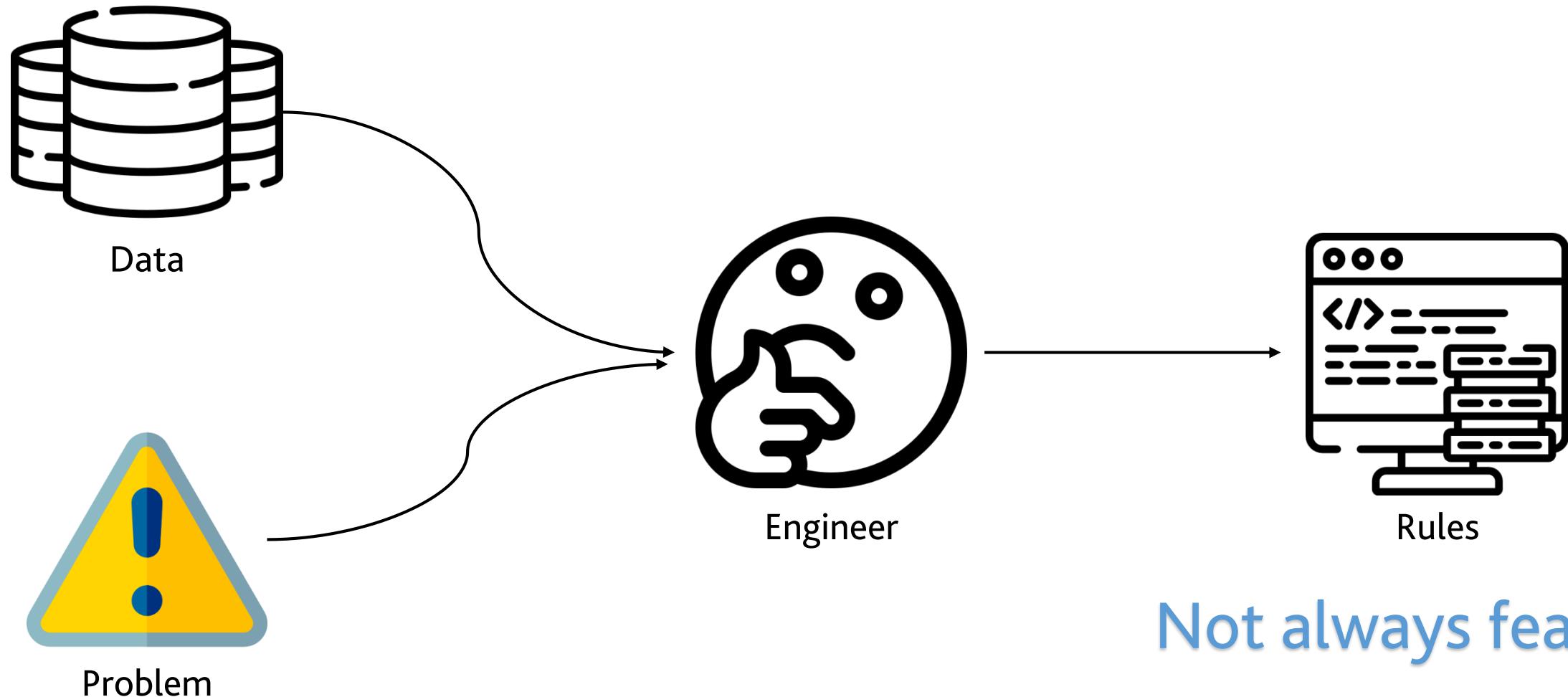


Screen Sharing

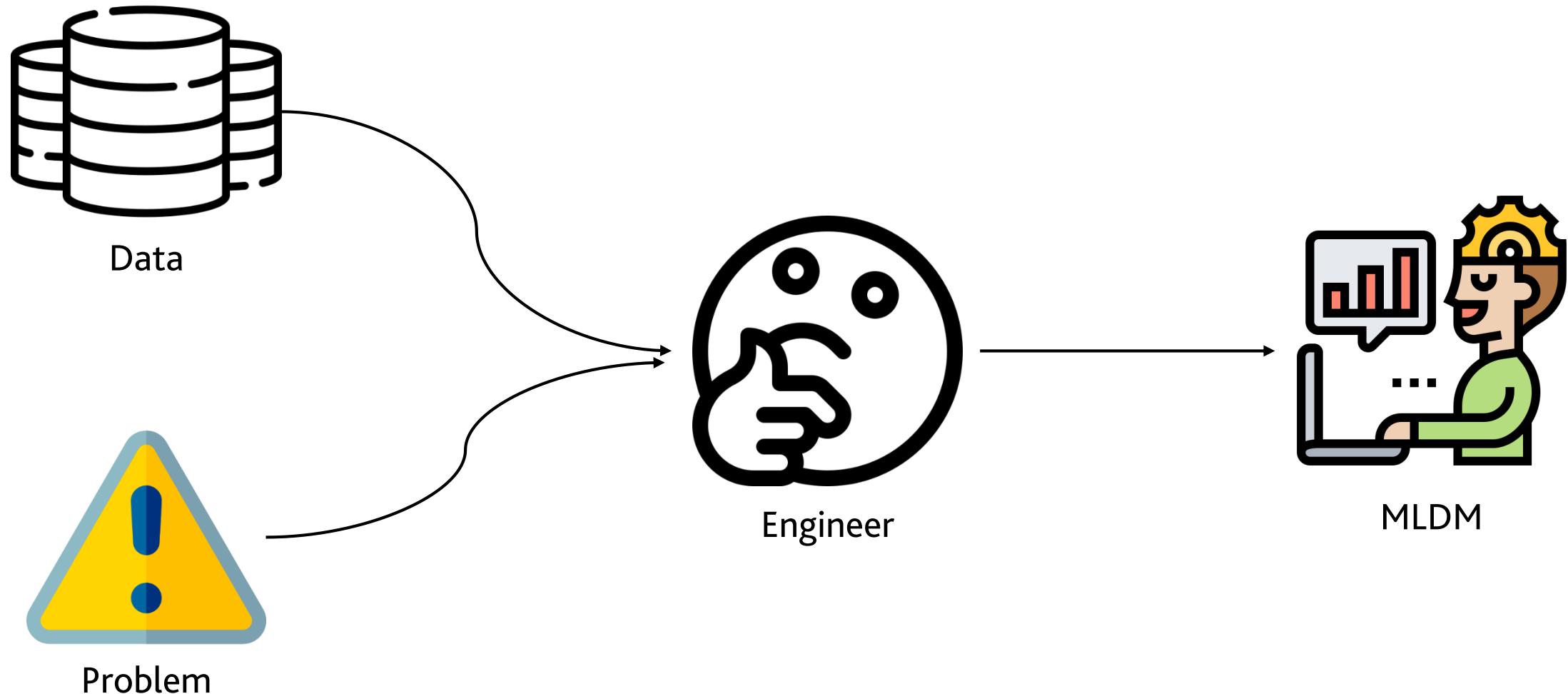
Let's get started!



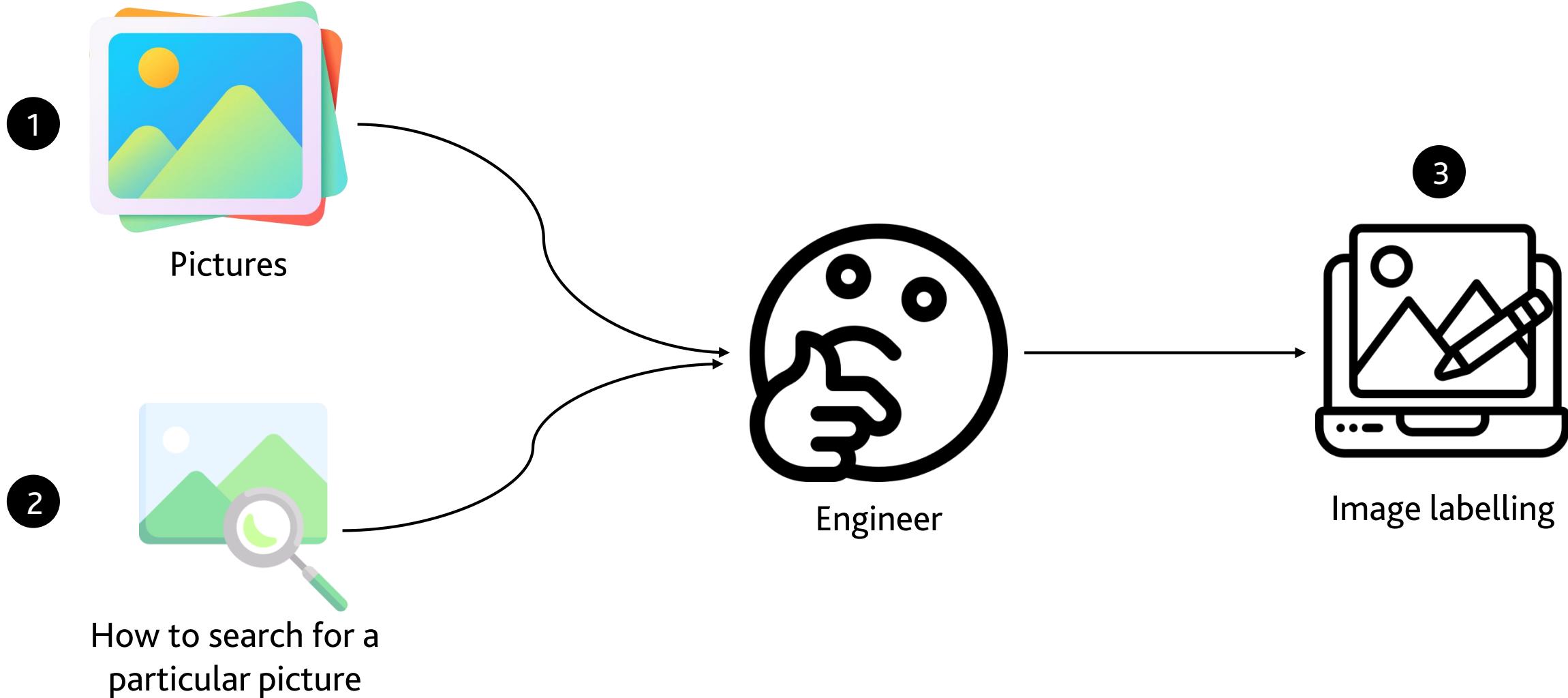
Why Machine Learning & Data Mining?



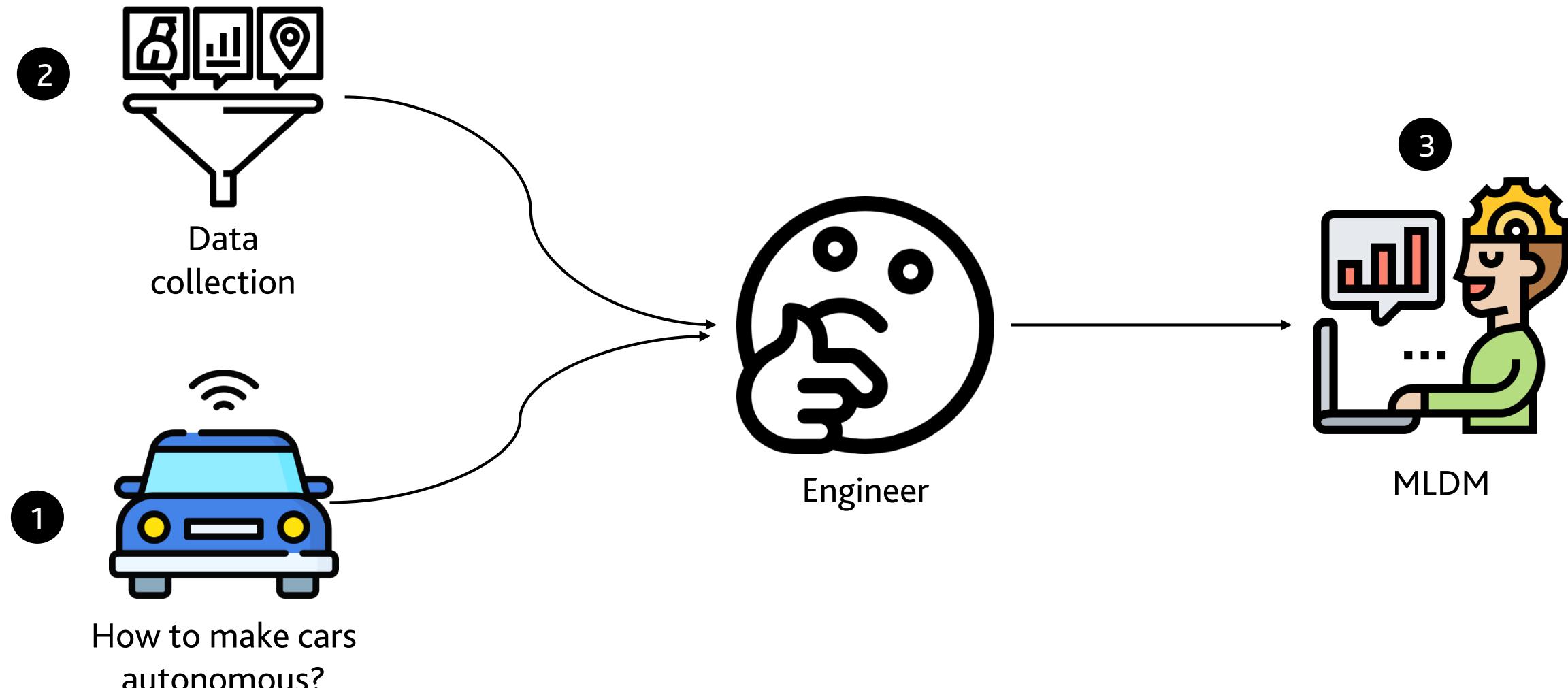
Why Machine Learning & Data Mining?



Why Machine Learning & Data Mining?

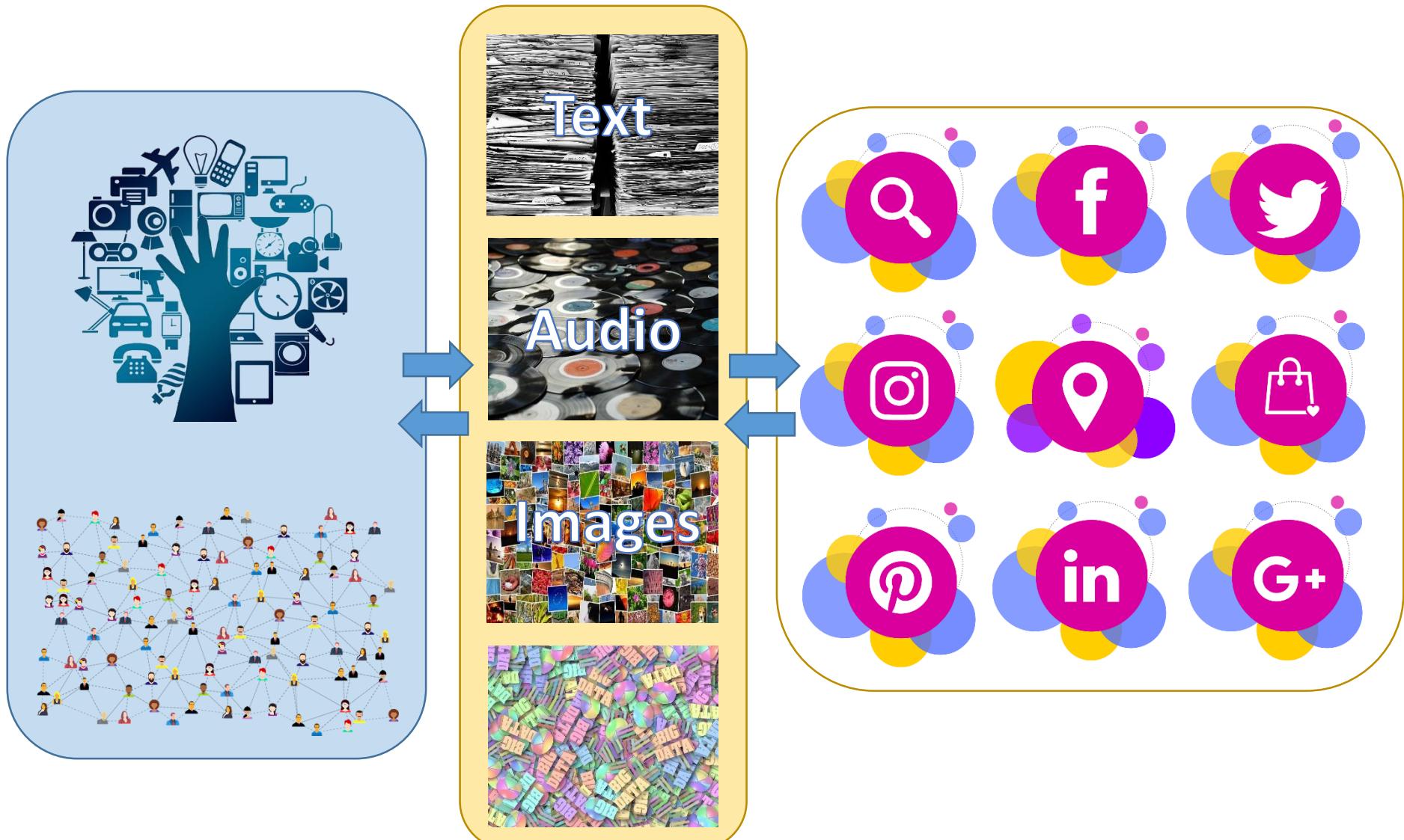


Why Machine Learning & Data Mining?



Icon vectors designed by [Freepik](#)

Why Machine Learning & Data Mining?



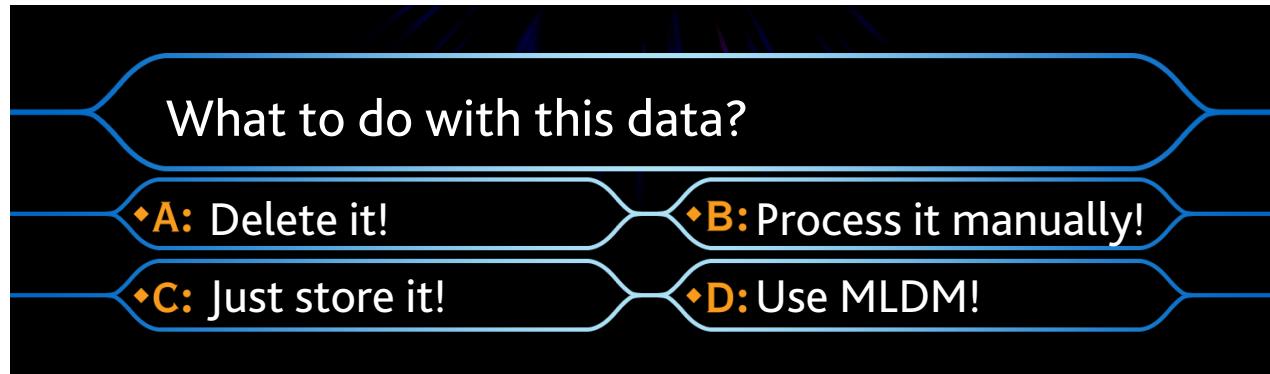
Images by [Pixabay](#) (copyright free)

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).

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

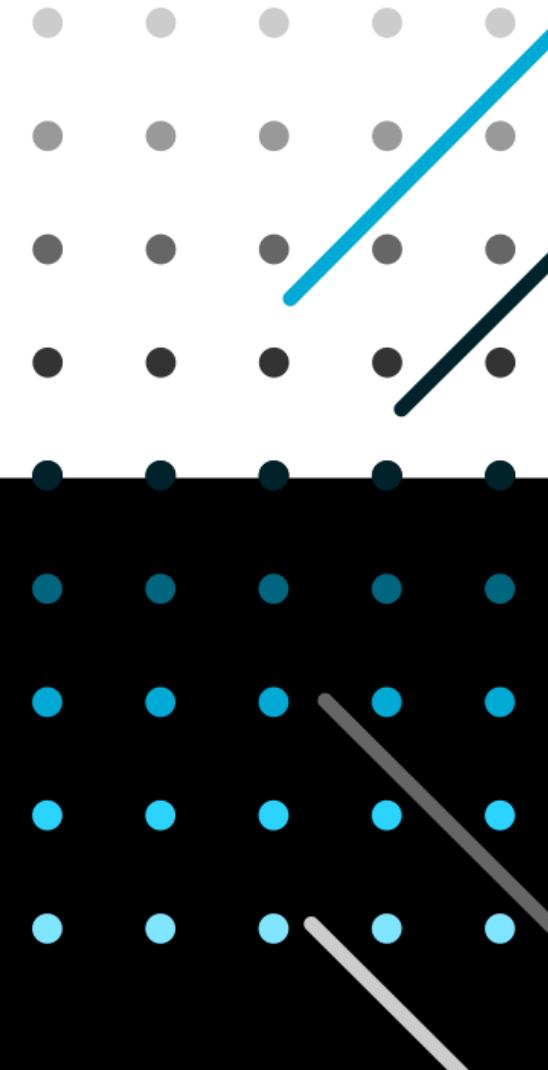
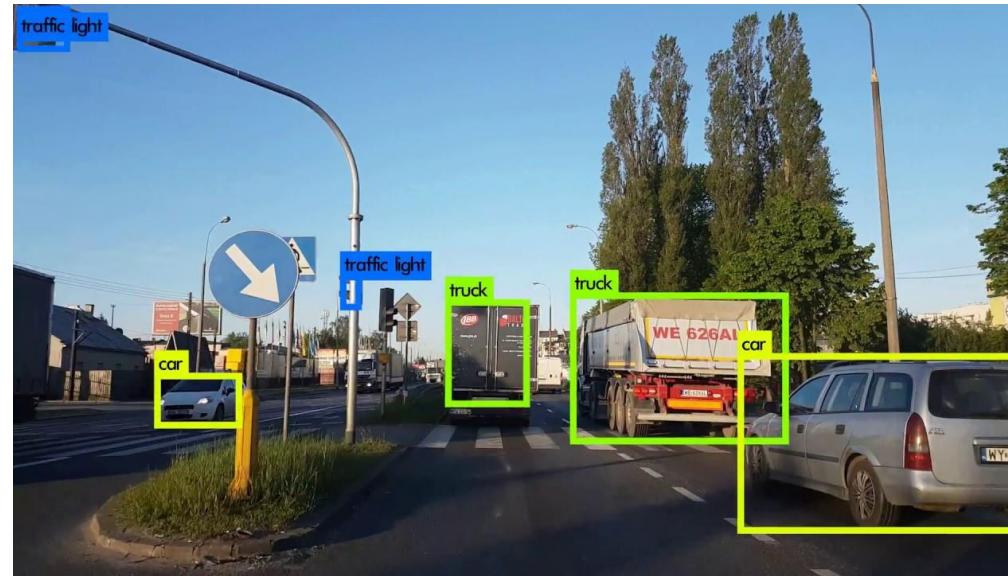
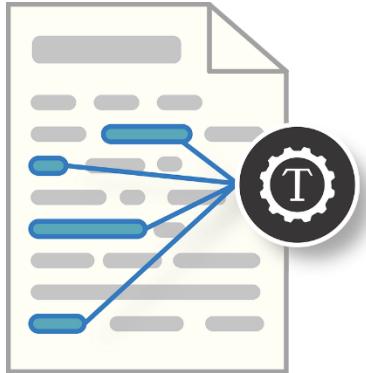
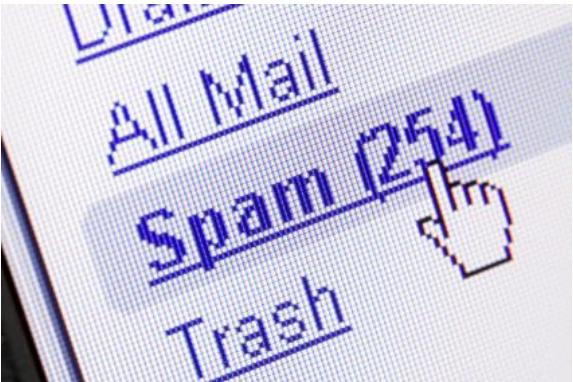


Image Processing: Object Detection, Recognition & Localisation



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

Natural Language Processing

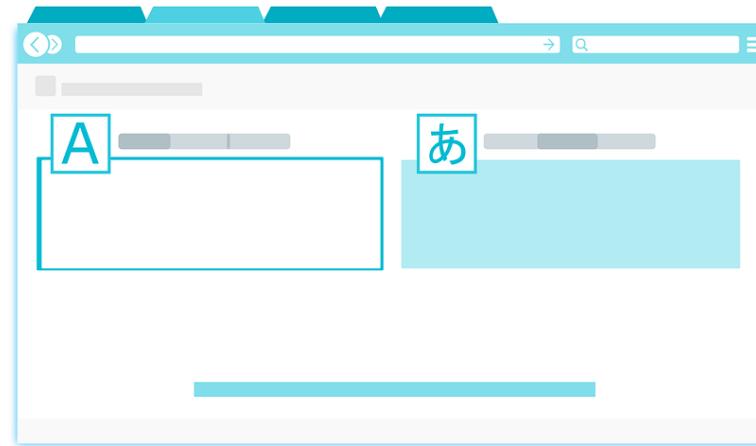
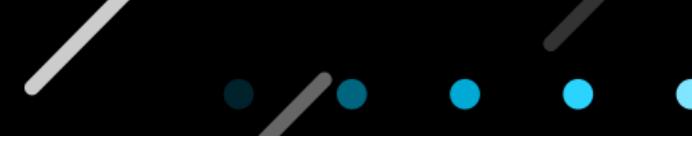


- Used in:
 - Spam filtering.
 - Information extraction.
 - Text classification
 - Plagiarism detection.
 - And so many other applications

Images by [Pixabay](#) (copyright free)

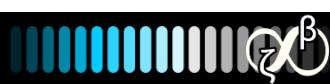


Machine Translation



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

Images by [Pixabay](#) (copyright free)



Speech Recognition

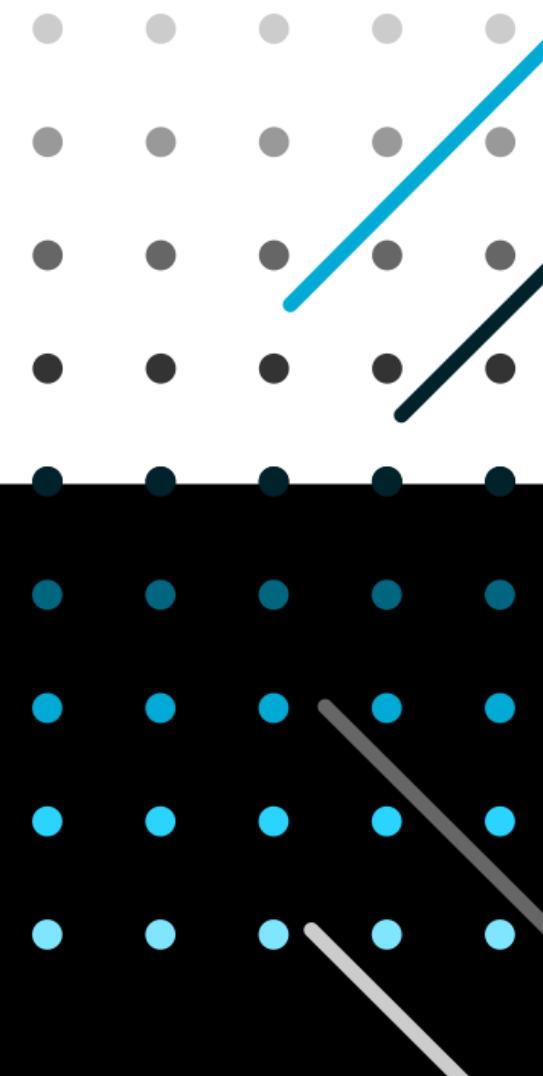


- Used in:
 - Easy and fast interaction with devices.
 - Important alternative for people with visual impairments.
 - Voice biometrics for identity recognition.
 - Effective communication with a big number of clients.
 - Etc.

Images by [Pixabay](#) (copyright free)



More applications



Age and gender detection



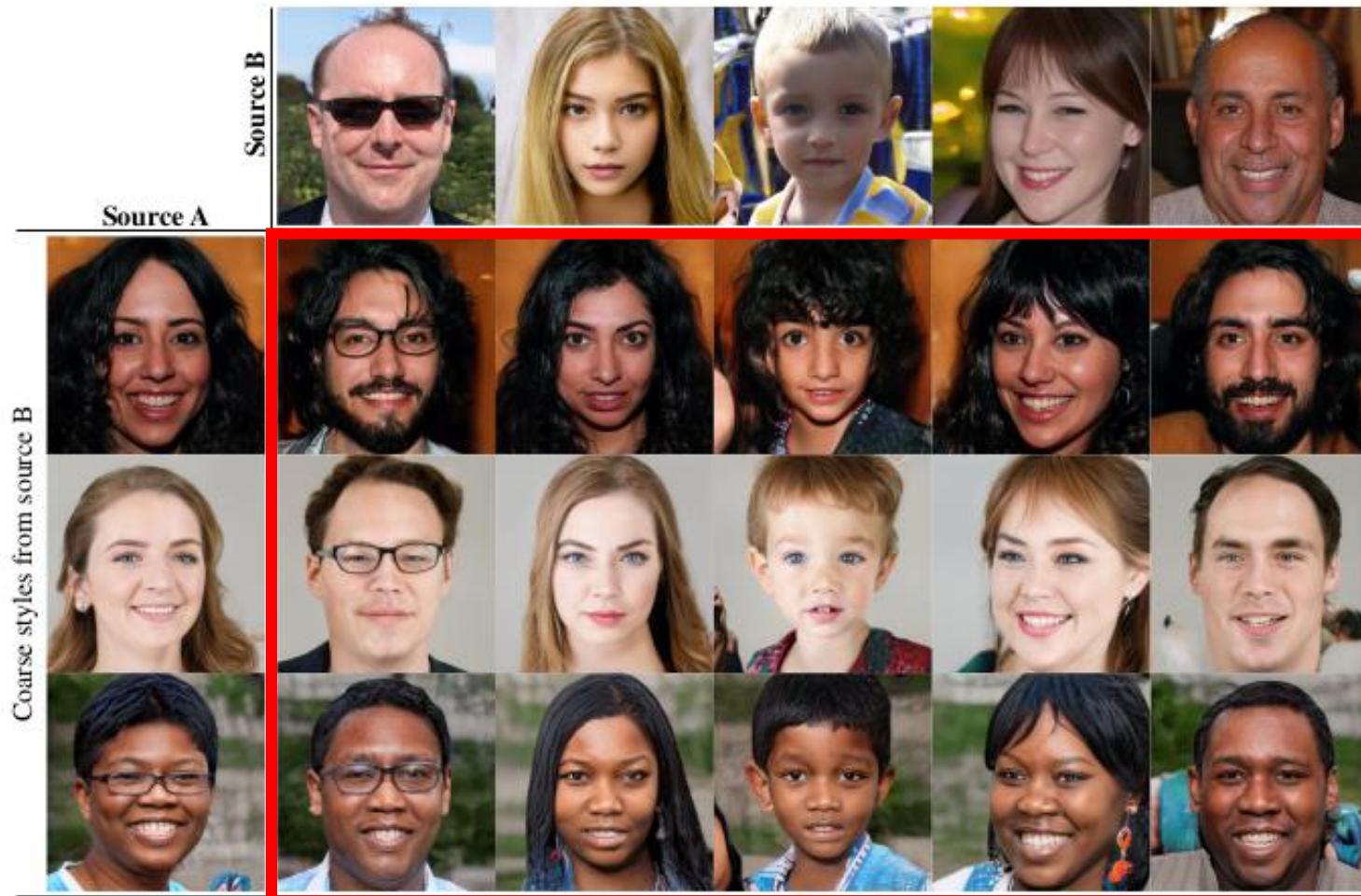
Recommendation system

- Ads
- Videos
- Products
- Etc.

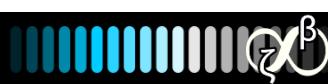


Google Ads

Data Generation



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

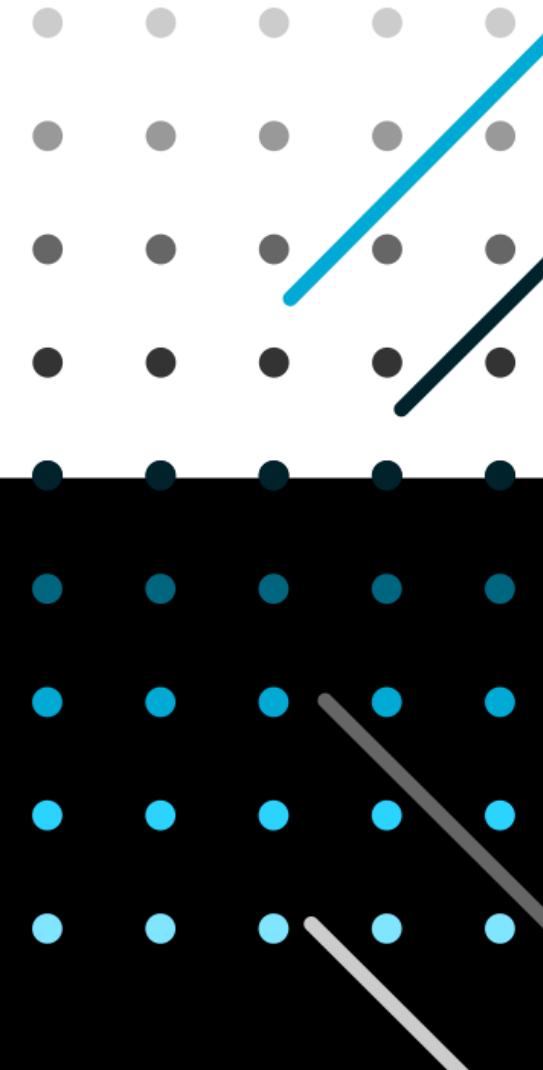


And there are even more in all domains . . .

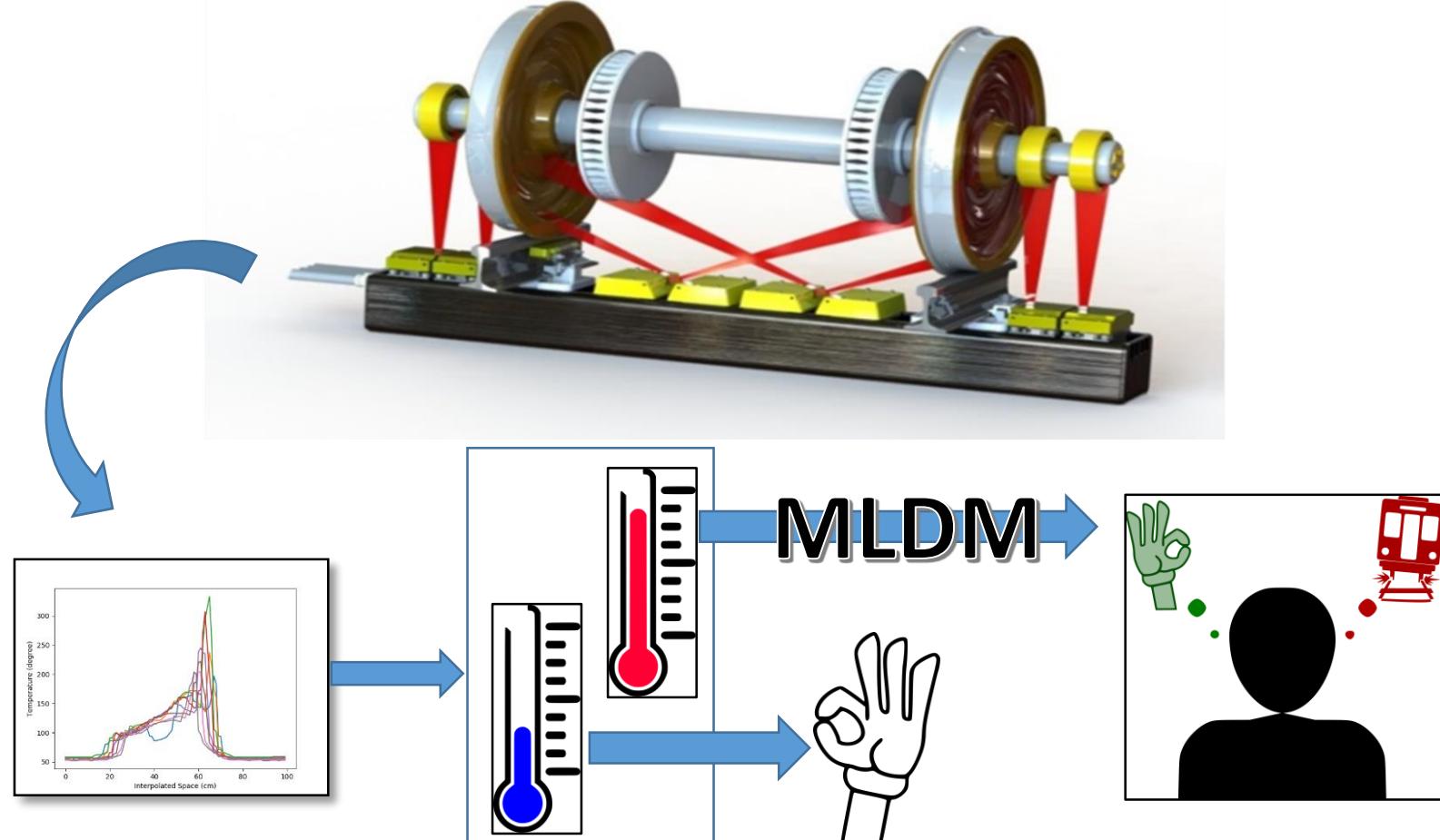
- Finance & Banking
- Security & Surveillance
- Transport & Traffic
- Healthcare & Well-being
- Media
- Commerce
- Sport
- Game & Entertainment
- Etc.

MLDM @ WeST

Example from past projects



Railway Monitoring System



Extracting and Parsing References



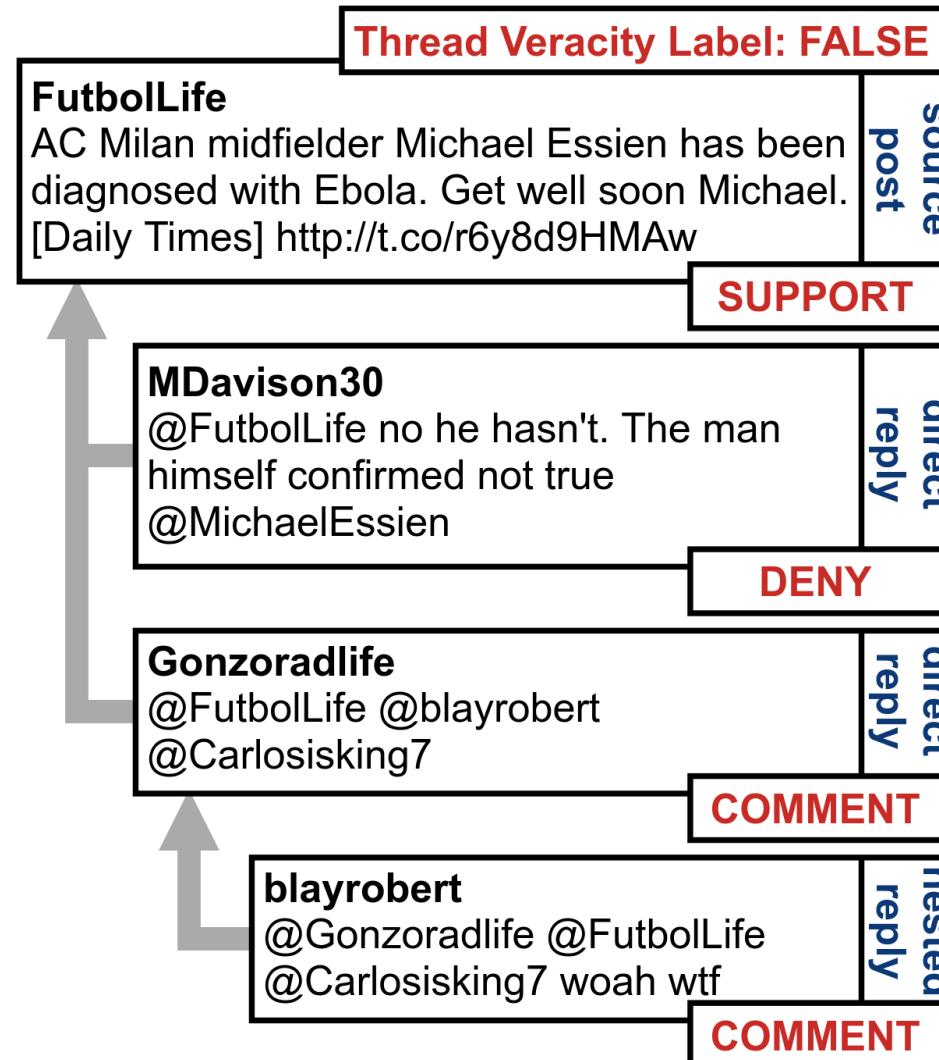
Andreas Gaudi - „Traditionen im Transit“	Andreas Gaudi - „Traditionen im Transit“	Andreas Gaudi - „Traditionen im Transit“
<p>Schicht von Kaufleuten das Feld überlassen musste.²⁸ Im Jahr 1902 erfolgte der Anschluss des Sandžak an das Kosovo-Etat, so dass sich während und nach (ab 1908) dem Ende der habsburgischen militärischen Bedrohung das Machtkontrum stärker in Prizren als in Sarajevo entfaltete. Der von Berlin zur Friedensherrschaft bestimmte Ethnograph Jevet Sefović (1873–1935) beschreibt die Bevölkerung als „ethnisch heterogen“ und „eher für Serbisch-Sprechende.“ erkante aber die bestimmkundigen Gemeinsamkeiten mit den „Serben“ in Bosnien, Dalmatien und Montenegro.²⁹ Merkwürdig ist ferner die Beschreibung der Bevölkerung als „ethnisch homogen“ und „eher für osmanische Sprache“. „Der Islam und ein religiöser Fanatismus, bereit für die größten Verbrechen“ sei „Selbstverständlich widergespiegelt“ diese abwegige Aussage.³⁰ Die Einstellung eines serbischsprachigen Autors wie Sefović kann man (heute bewundern). Dennoch lässt diese Auskunft die Frage entstehen, ob ein ethnisch konvertierter „Widmerding“ der muslimischen Bevölkerung bereits damals existierte. Noch 1921 konferenzierte eine Konferenz für den Zentralen Balkan-Konsortium in Sarajevo (die Konferenz war aufgrund der politischen Lage weit entfernt) oder zumindest auf eine Autonomie für den Sandžak einigte.³¹ Daraus entwickelte sich nichts. Konkurrenz und 1918 wurde der Sandžak aus dem Osmanischen Reich ausgeschlossen.</p>	<p>Dragica Presević-blaskić, die sich noch als „Kind des Kommunismus“³² bezeichnet, kritisiert die heutige Vernachlässigung von wichtigen sakularem Kulturerbe wie der Stadtbildung, dem Hafenamt oder dem jugoslawischen Friedhof. Vatneher seien die Behörden und die Gewerkschaften an der Sanierung und dem Neubau vom Museum und der Kirche interessiert. Das ist wiederum ein Zeichen der „Zersetzung“ des Sandžaks.³³ „Der Islam verhindert immer nur eine begrenzte, ausgewählte Vergangenheit bzw. Erbe („Erbe als Wahrnehmung“) und gleichzeitig wird Historizität dem Verfall überlassen. Kurz gefasst, stützt sich ein Teil der Bevölkerung auf die Zersetzung des Sandžaks.“³⁴ Der Islam verhindert nicht nur die Nationalität, um ihre Loyalität außerhalb seines Staates auszurichten. Anderswo beobachtete ich mittels des politischen Islam – also des weltumfassenden propagandistischen „Islam“ – die Trennung und Spaltung der moslemischen Gemeinden. Dennoch ist diese Beobachtung und prägt dennoch gleichzeitig verschärft die traditionell gehaltene Referenz durch neue KulturrinstitUTIONEN wie das Museum.</p>	<p>Bevor heutige KulturrinstitUTIONEN behandelt werden, soll ein Einblick in die politische Zugehörigkeit der Elite zu Beginn der 1990er Jahre gewährt werden, da die Krise der „Jugoslawien“ die politische Zugehörigkeit aufdeckte.</p> <p>Zuerst ist zu betonen, dass die sich am Ende der 1980er Jahre kristallisierte „neue“ politische Elite keine Kontinuität mit dem kommunalistischen System aufwies.³⁵ So sind beispielsweise in den „offiziellen“ Biografien keine Angaben über die Aktivitäten von jugoslawischen Parteifunktionären zu finden. Eine Ausnahme bildet der Politologe und Diplomat Ugljanin, der in die Politik im Jahre 1990 zurückkehrte.³⁶ Ugljanin spricht die Hauptlinie in der DRJA seitens der SDA als „short-term rationality“ zu bezeichnen ut. Die SDA setzte sich 1990 als Hauptverteidiger der moslemischen Bevölkerung ins Nachtschad durch, was 1991 zu einem gewalttätigen Aufmarsch der Ugljanin-Faktionen in Belgrad und der Präsidialstadt Sarajevo erfuhr. Knapp ein Jahr nach der Gründung der SDA organisierte Ugljanin ein Referendum über den Status des Sandžaks: rund 70% der Wahlberechtigten sahen sich in der „Bosnian“-Identität bestätigt.³⁷ Am 1. Januar 1992 wurde der Sandžak autonom. Kurz darauf nahm eine sandžakische Delegation an den Konferenzen zur Bosnienfrage in Genf und Lausanne teil. Daraufhin rief die SDA zum Boykott der jugoslawischen</p>
<p>Vgl. Grandits, Hannes (2007): Zur Modernisierung der spätmittelalterlichen 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.</p>	<p>²⁸ Diese und folgende Angaben in Gaudi, Andreas (2012): Interview mit Dragica Presević-blaskić. Novi Pazar, 29.03.2012. Gedächtnisprotokoll.</p> <p>²⁹ Diese unterscheidet sich von der von Cvetko und Božović für die Balkan kontingenzielle Tradition. Siehe Božović, Xaviera (2007): The Balkans in the 19th century. In: Božović, Xaviera; Čiler, Božidar (Hg.): (2007) The Nation is born. First Monographs on the Balkans. Ljubljana: Cankarjev dom.</p> <p>³⁰ Flaminio, Šarić (1998): Bosnjačka politika u XX. stoljeću [Bosnian political policy in the 20th century]. Sarajevo: Školski zavod.</p> <p>³¹ Vgl. Školski zavod za obrazovanje i kulturu. (2003): Inovacije u obrazovanju i kulturi: 2003. godina. Sarajevo: Školski zavod.</p> <p>³² Diese und folgenden Angaben in Gaudi, Andreas (2012): Interview mit Dragica Presević-blaskić. Novi Pazar, 29.03.2012. Gedächtnisprotokoll.</p> <p>³³ Diese sieht die Broschüre des BNV (Bosnian National Museum) Džudović, Eraljic; Mihalović (2010) Pod njenim naslovom „Sandžak: etno-geografski i kulturno-istorijski rezervat“ (Sandžak: ethnogeographical and cultural-historical reserve). Sarajevo: BNV.</p> <p>³⁴ Božović, Xaviera (2007): Ulaz bosnjanaca: etnički identitet i etnogenetika politike. In: Božović, Xaviera; Čiler, Božidar (Hg.): (2007) The Nation is born. First Monographs on the Balkans. Ljubljana: Cankarjev dom.</p> <p>³⁵ Andrićević (1995), S.175. Es ist beweislich, dass der 1992 abgeschlossene Krieg in Bosnien/Herzegowina eine gewalttätige Reaktion auf die politische Zugehörigkeit der Bevölkerung war. Siehe auch Božović, Xaviera (2007): The Balkans in the 19th century. In: Božović, Xaviera; Čiler, Božidar (Hg.): (2007) The Nation is born. First Monographs on the Balkans. Ljubljana: Cankarjev dom.</p> <p>³⁶ Božović, Xaviera (2007): The Balkans in the 19th century. In: Božović, Xaviera; Čiler, Božidar (Hg.): (2007) The Nation is born. First Monographs on the Balkans. Ljubljana: Cankarjev dom.</p> <p>³⁷ Božović, Xaviera (2007): The Balkans in the 19th century. In: Božović, Xaviera; Čiler, Božidar (Hg.): (2007) The Nation is born. First Monographs on the Balkans. Ljubljana: Cankarjev dom.</p>	<p>negativ auf die Errichtung des Beispiekaments als nationale Bewegung auswirken? In: Popović, M. (Ed.): (1998), 6–329–343.</p> <p>Kamra, Ibrahim (1986): Ulaga „Čapet“ u državnom životu muslimana Bosne i Hercegovine (1903–1914). In: Božović, Xaviera (Hg.): (1986) Muslimani u političkom životu Srbije i Bosne i Hercegovine. Sarajevo: Vodenik Matična.</p> <p>Lutovac, Žorža (1994): Muslimani u političkom životu Srbije i Bosne. In: Božović, Xaviera (Hg.): (1994) Muslimani u političkom životu Srbije i Bosne i Hercegovine. Sarajevo: Vodenik Matična.</p> <p>Medvedović, Krsto: Šepeć: Sandžak i duple obveznik. [Die serbische Sandžak immer noch vorgegen]. In: Božović, Xaviera (Hg.): Sandžak: Identitet u preprost smislu i novog (Der Sandžak: Identität im einfachen Sinn und neuer). Sarajevo: Vodenik Matična. S.206–211.</p> <p>Mehmed, Šefik (1998): Zemalj je autonomska vlast sandžaka – Zemalj. [Der autonome Landesrat des Sandžaks – ZAVNOŠ]. In: Zelci, Arif (Hg.): Sandžak na putu autonomske. Sarajevo: Vojno književno i naučno izdavaštvo.</p> <p>Milutović, Ilijas (1979): Etnički procesi i etnička struktura stanovništva Novog Pazaru. [Die ethnischen Prozesse und die ethnische Struktur der Bevölkerung Novi Pazar]. Beograd: Srpska akademija nauka i umjetnosti.</p> <p>Pavlović, Aleksandar (2009): Udan balkančan. Les Muslimanini su sud etiopijani da perioda post-otomanske. Istropol: ISTR.</p> <p>Pavlović, Arif (1998): [Die jugoslawische Muslimische Organisation im politischen Leben des Königreichs der Serben, Kroaten und Slowenen]. Sarajevo: Bosanski Kulturni Centar.</p> <p>Ramić, Dragan (1998): Sandžak: identitet i autonomska vlast. In: Božović, Xaviera; Čiler, Božidar; Georg Božović (Hg.): (1998) Muslimani in der Sonderform und Autonomie. Sarajevo: Vodenik Matična. S. 107–114.</p> <p>Republičko Školsko Predškolsko izdavaštvo. (2005): Počeci autonomske demokracije i status u 2002. Knj. 1: Register der Bevölkerung, Haushalte und Wohngebiete im Jahre 2002. Bd. 1. Beograd: Republičko izdavaštvo.</p> <p>Subject, Željko (1998): Naučenost Muslimana [Die Nationalität der Muslimen]. Rijeka: Oskar Kremers.</p> <p>Todorović, Maria (1996): The Ottoman Legacy in the Balkans. In: Brown, Carl L. (Hg.): Imperial Legacies: Ottoman, Habsburg, and Hellenic Empires on the Balkans and the Middle East. New York: Columbia University Press, S.65–77.</p> <p>Tomić, Đorđe (2011): Od monofundne do transnacionalne: Nacija o kojoj ne moguši? Popović, Željko (Hg.): (2011) Nacija: etnička identitetnost i politička transformacija. Sarajevo: Bogićević, Frane (Hg.): (2011) Sandžak na putu autonomske. Sarajevo: Vojno književno i naučno izdavaštvo. Zemalj, Vojko (Hg.) (2009): Sandžak na putu autonomske. Sarajevo: Vojno književno i naučno izdavaštvo.</p> <p>Todorović, Maria (1996): The Ottoman Legacy in the Balkans. In: Brown, Carl L. (Hg.): Imperial Legacies: Ottoman, Habsburg, and Hellenic Empires on the Balkans and the Middle East. New York: Columbia University Press, S.65–77.</p> <p>[N., N.] (1940): „Balkan predstavlja 20 političke Gajeteve rada u Sandžaku“ [Australisch des 20. Jhd. und jugoslawisch des 20. Jhd.]. In: Cvetko (2010), 60–68, 94ff. (n. 8).</p> <p>[N., N.] (1940): „Dnevnički Sandžak u prvoj svjetskoj vojni“ [Der Sandžak von Novi Pazar in der Vergangenheit]. In: Cvetko (2010), 61, 81–190, 204ff. (n. 5).</p> <p>[N., N.] (2007): „Sandžak – naše blago“ [Novi Pazar – unser Schatz]. In: Božović Rječ (7), 9–27.</p>

²⁸ Vgl. Grandits, Hannes (2007): Zur Modernisierung der spätmittelalterlichen 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.

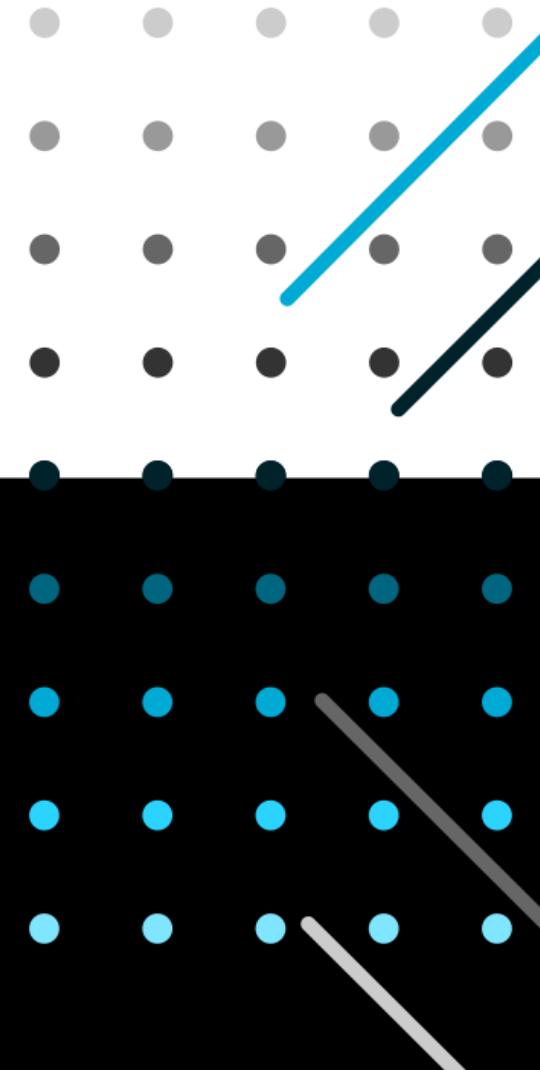
<i>Author*</i>	<i>Author**</i>	<i>Title</i>	
<i>Title</i>	<i>Source</i>	<i>Year</i>	<i>Page range</i>
19. Viola, P., Jones, M.:	Rapid object detection using a boosted cascade of simple features.	In: CVPR01.	I:511–518



Rumour detection and stance classification



Predictive learning problems





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

Machine Learning

v.s.

Human Learning

Defining Task

Collecting Data

Designing Features

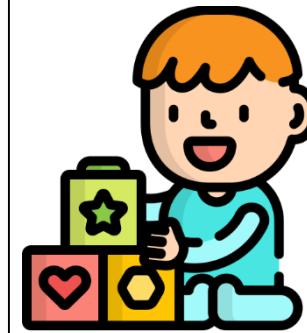
Training Model

Testing Model

Evaluating Model

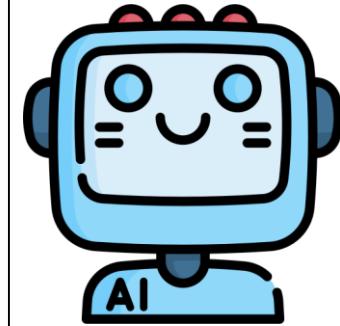
- **Human learning**

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”.





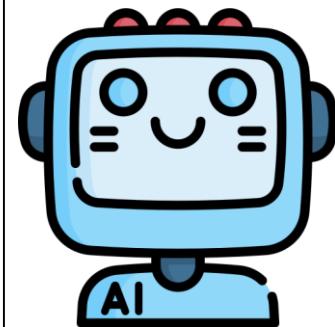
- **Human learning**

- 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



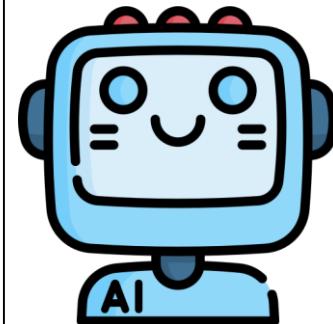
- **Human learning**

- ~~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

Classification

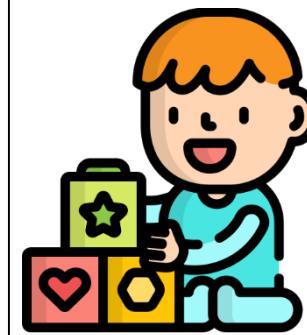


Images by [Pixabay](#) (copyright free)

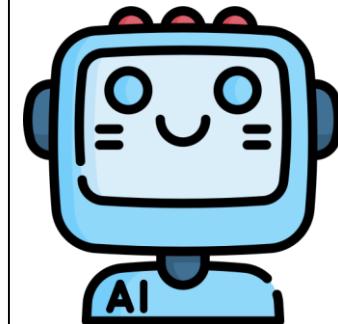


- **Human learning**

- 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

Defining Task

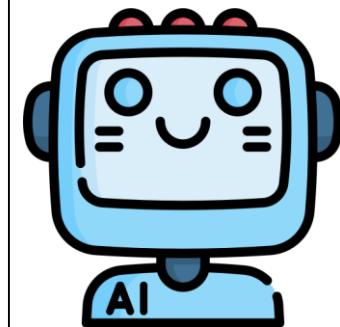
Collecting Data

Designing Features

Training Model

Testing Model

Evaluating Model



- **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.

Classification



For a better visualisation, we consider only two features.



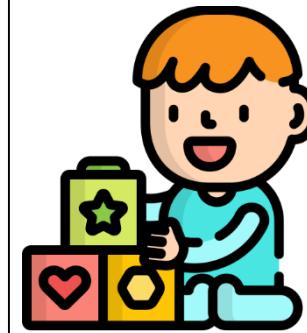
Images by [Pixabay](#) (copyright free)





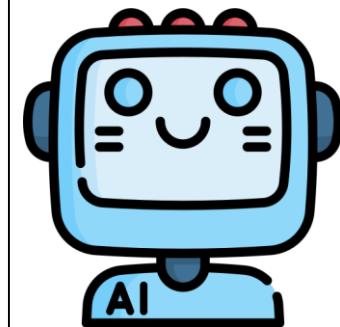
- **Human learning**

- 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

Classification



For a better visualisation, we consider only two features.



Images by [Pixabay](#) (copyright free)



Defining Task

Collecting Data

Designing Features

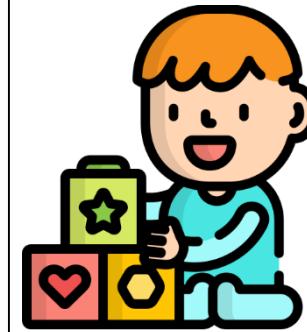
Training Model

Testing Model

Evaluating Model

- Human learning

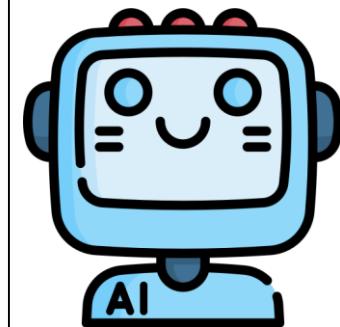
Bravo!



- Machine learning
 - How good is the model?

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

(we will see them in the upcoming
lectures)



More examples of classification

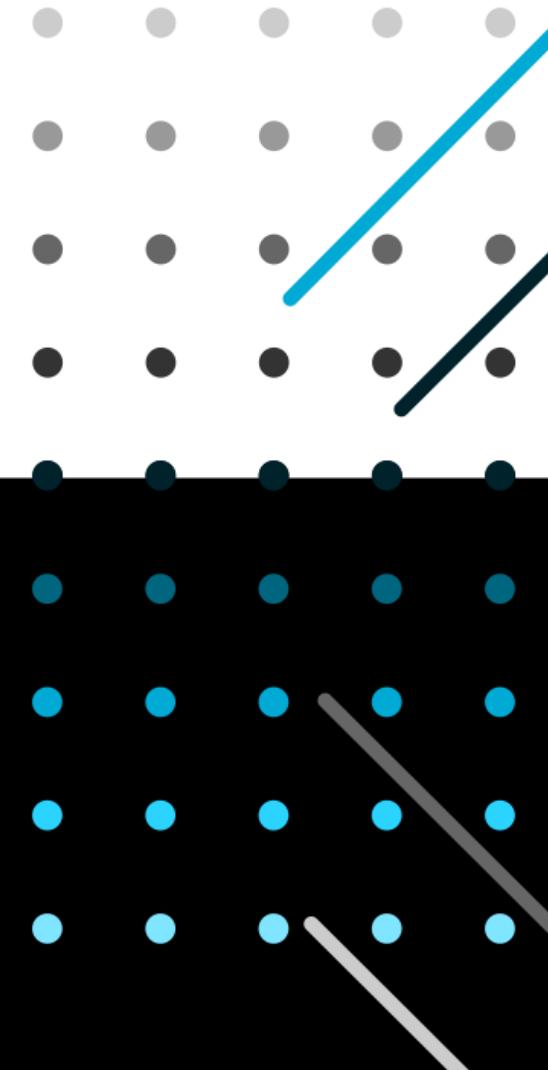
- Is the skin mole a malignant melanoma?

- 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

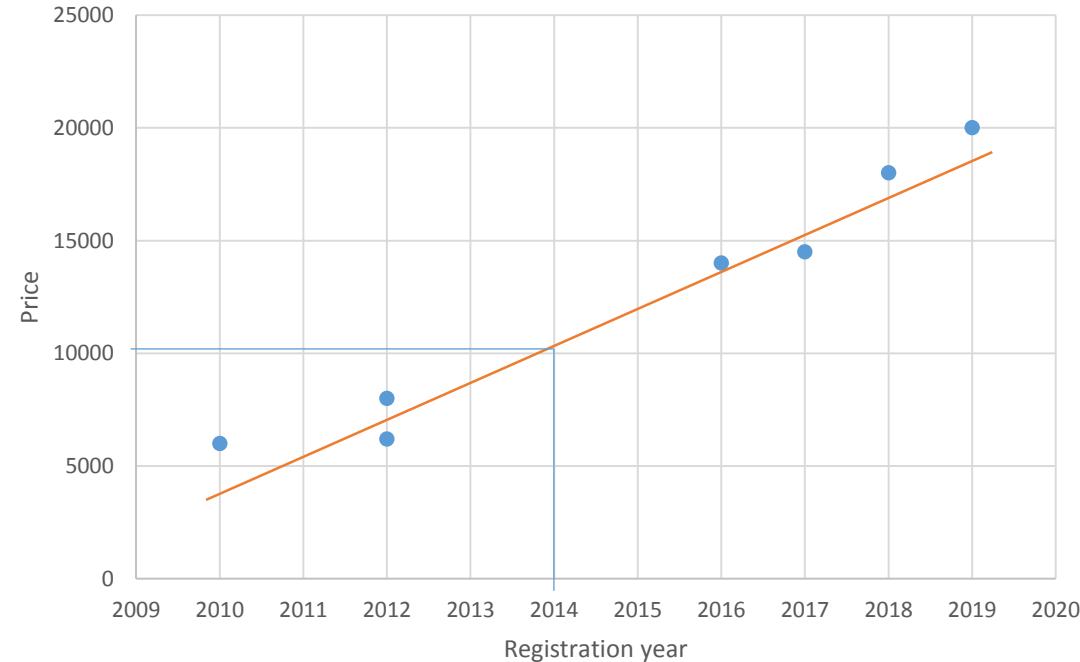


- We want to predict the price of our car based on the registration year, say 2014.
 - Obviously, all other properties (e.g. model, colour, etc.) are known.
 - We could obtain some examples of the same car from e-commerce websites:

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

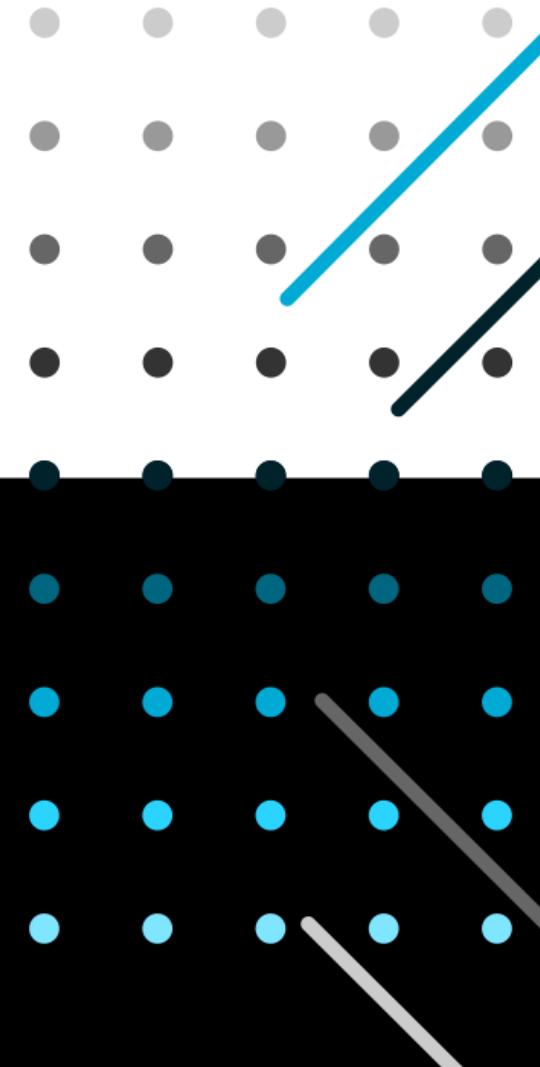


- Based on this data, our car should cost around 10,500 €.

More examples of regression

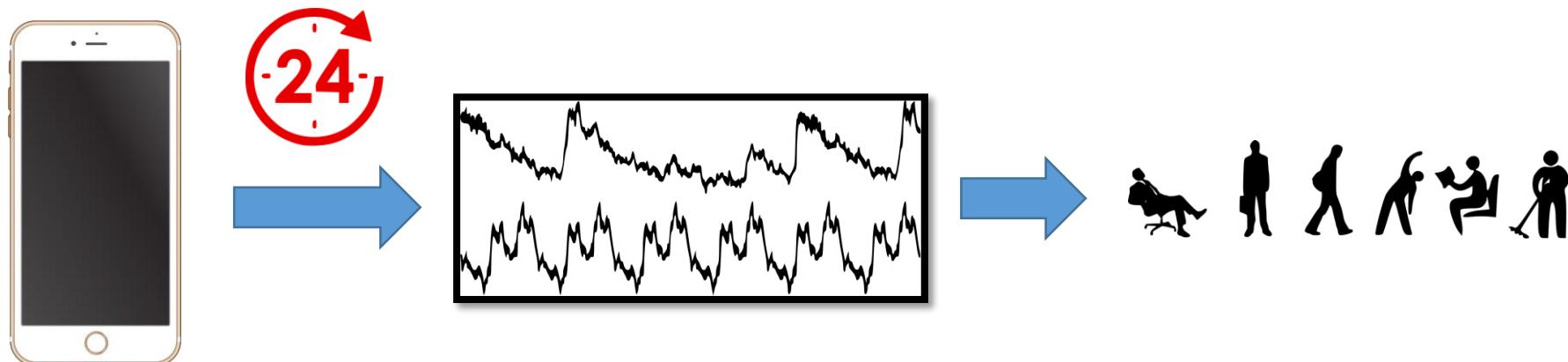
- What is the height given the age?
 - [45, 210] cm
 - Continuous prediction.
 - Called regression.

Descriptive learning problems



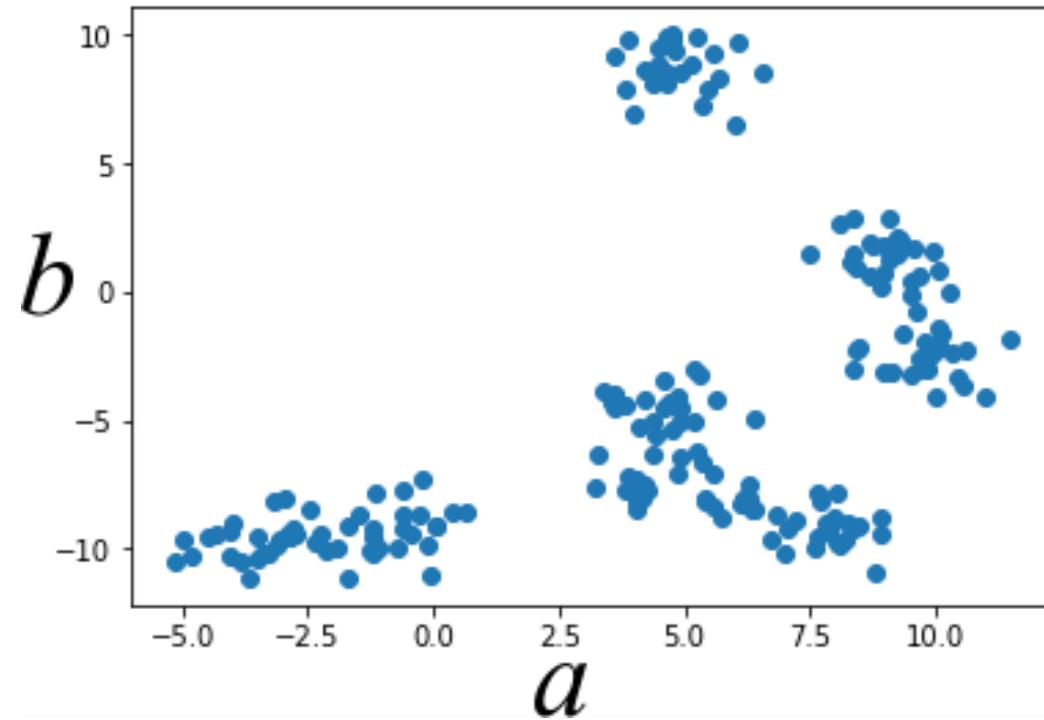
- Considering the following problem:

- 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.

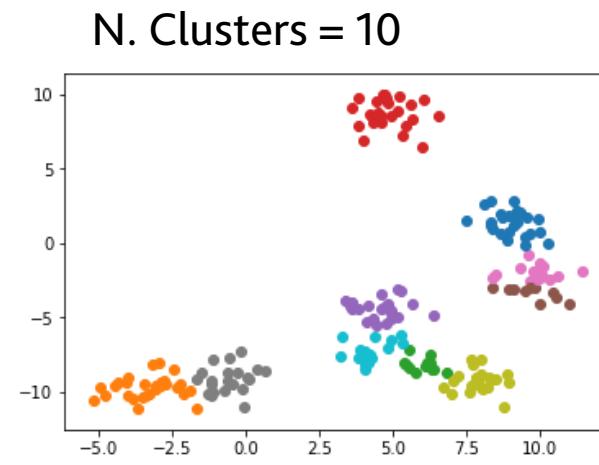
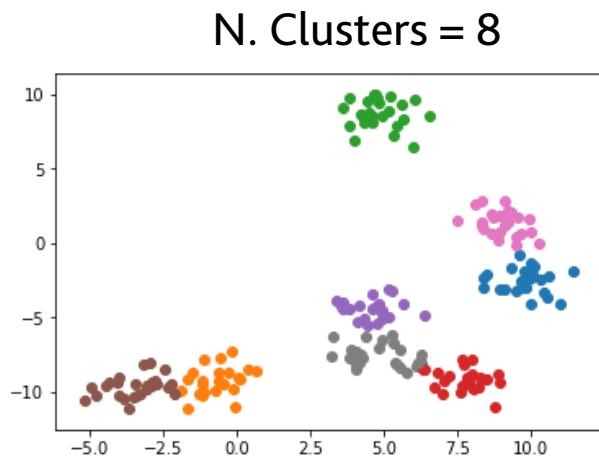
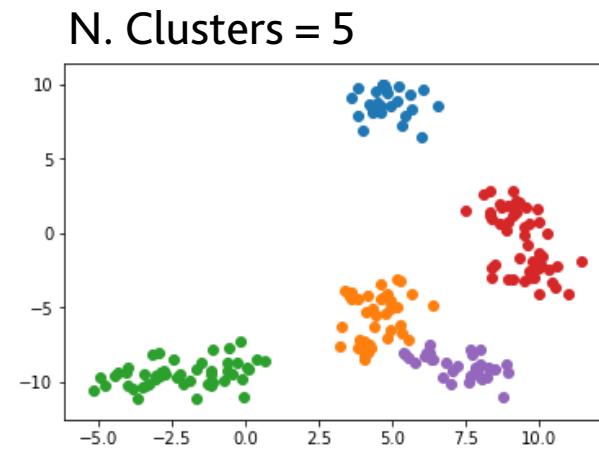
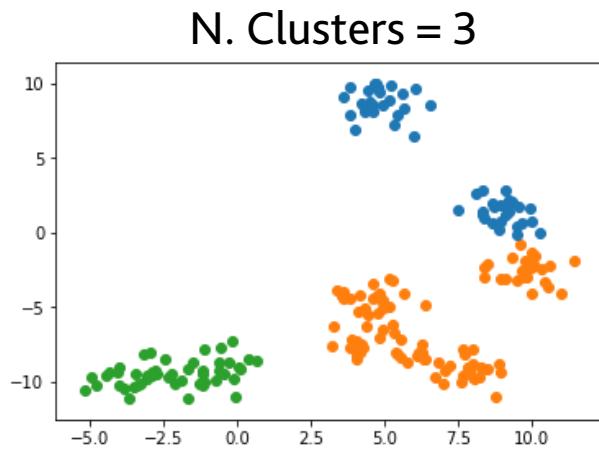


Images by [Pixabay](#) (copyright free) + Icon vectors designed by [Freepik](#)

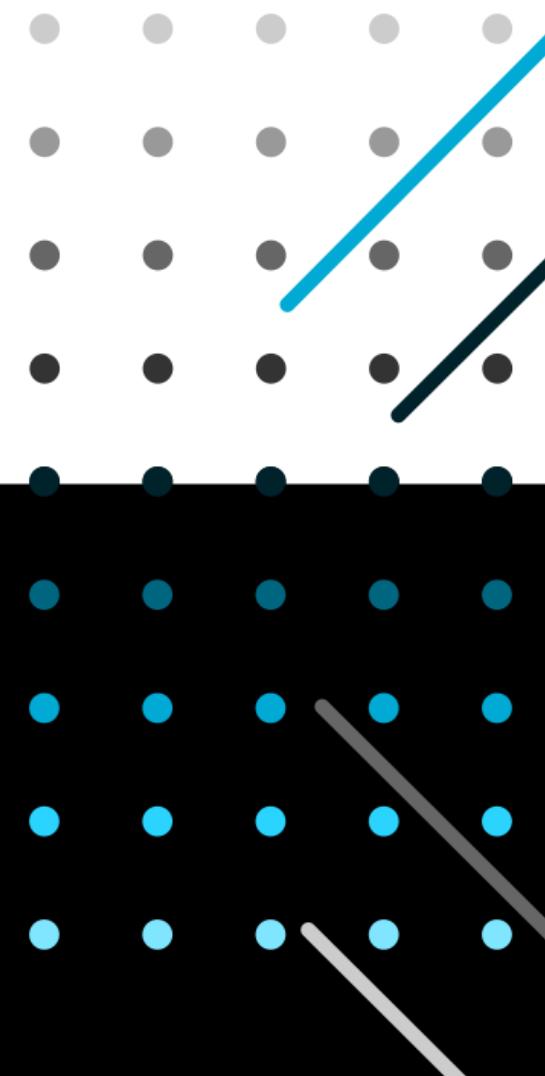
- For the sake of visualization, the data (measurements) are represented by two features (a and b).



Clustering



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 property (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 \neq Supervised learning
 - Descriptive learning \neq Unsupervised learning

(Un-) Supervised vs. Predictive/Descriptive



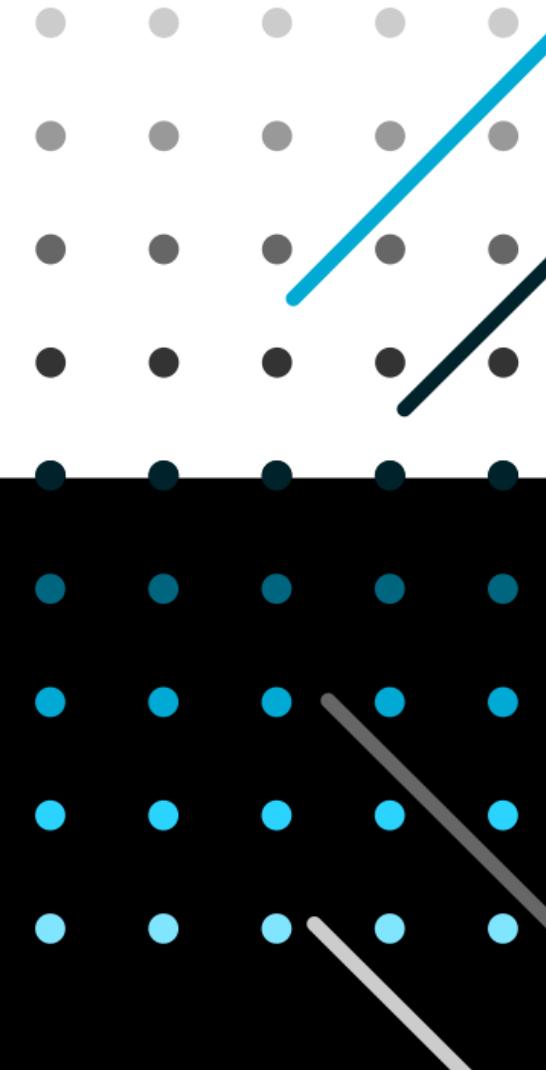
	Predictive	Descriptive
Supervised	<ul style="list-style-type: none">The model uses past data to predict events or outcome in future data.The model is trained with labelled data.	<ul style="list-style-type: none">The model extracts knowledge from the data at hand.The model is trained with labelled data.
Unsupervised	<ul style="list-style-type: none">The model uses past data to predict events or outcome in future data.The model is fed with unlabelled data only.	<ul style="list-style-type: none">The model extracts knowledge from the data at hand.The model is fed with unlabelled data only.

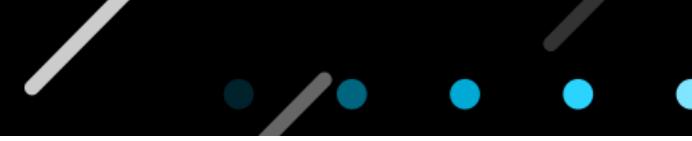
Examples



	Predictive	Descriptive
Supervised	<ul style="list-style-type: none">• Object classification• Object detection• Object extraction• Etc.	<ul style="list-style-type: none">• I couldn't think of any example!
Unsupervised	<ul style="list-style-type: none">• Anomaly detection (e.g. Bank transactions, Human activities, etc.)	<ul style="list-style-type: none">• Image segmentation• Topic modelling• Clustering• Etc.

Other learning approaches





- 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

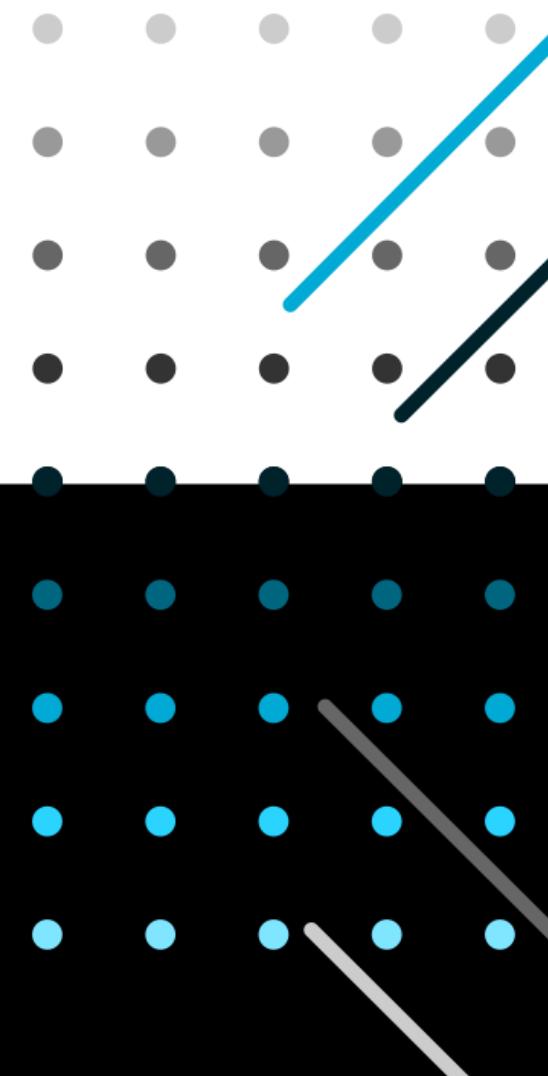


- It is goal-oriented that learns how to achieve a complex goal.
 - For every state, it takes an action.
 - The series of actions is supposed to reach the goal.
 - The data is unlabelled.
 - It learns by exploration and exploitation.
 - Exploration: try different sequence of actions.
 - Exploitation: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



- 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



- 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 MLDM: Predictive, descriptive, ...

Thank you!



Zeyd Boukher

E-mail: Boukher@uni-koblenz.de

Phone: +49 (0) 261 287-2765

Web: Zeyd.Boukher.com

University of Koblenz-Landau
Universitätsstr. 1
56070 Koblenz

