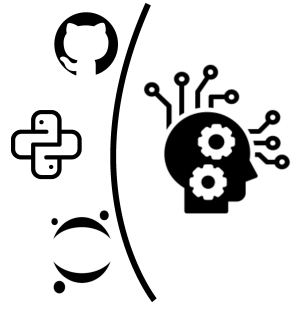


APPLIED MACHINE LEARNING
SDU

***

Lecture 1

Introduction
+
Course Setup

---

8 February 2022

Rahman Peimankar

Course Logistics

- **Room:** U174
- **Date and time:** Tuesday – 12:15-16:00
- **Office hours:** Monday – 15:00-16:00 (by appointment)
- **Zoom link:** https://syddanskuni.zoom.us/j/63660562254 (https://syddanskuni.zoom.us/j/63660562254)

We will talk about format of the course and grading in details later on today!

My Contact Information

- **Instructor:** Abdolrahman (Rahman) Peimankar
- **Affiliation:** Assistant Professor at The Maersk Mc-Kinney Moller Institute, University of Southern Denmark
- **Email:** abpe@mmmi.sdu.dk
- **Office location:** Ø8-700a-2

Work Experience:

Education:

Can you guess which one is Iran?

Please vote here! (https://PollEv.com/multiple_choice_polls/9VEt7DuctpaT67rpMXMp5/respond)

- We will use Poll Everywhere software for short quizzes and feedback in the course!

Now it is your turn! 😊

- Your name and education, and
- If you are an exchange or SDU student

Lecture 1 - Agenda

1. Course Overview
2. What is machine learning
3. Machine Learning Categories
4. Machine Learning Applications
5. Brief History and Machine Learning Ecosystem
6. Course setup

1. Course Overview

Part 1: Computational Foundation & Introduction

- Lecture 1: Course setup + Introduction
- Lecture 2: Python Basics and Packages

Part 2: Supervised Learning

- Lecture 3: Introduction to Supervised Learning
- Lecture 4: Preprocessing and Feature Transformation

Part 3: Linear Models

- Lecture 5: Linear Models for Regression
- Lecture 6: Linear Models for Classification

Part 4: Non-linear Models

- Lecture 7: Decision Trees, Random Forests, Ensemble

Part 5: Evaluation

- Lecture 8: Model Evaluation + Learning with Imbalanced Data

Part 6: Automate Machine Learning

- Lecture 9: Feature Selection + Parameter Tuning and Automated ML

Part 7: Clustering

- Lecture 10: Dimensionality Reduction + Clustring + Outlier Detection

Part 8: Neural Networks

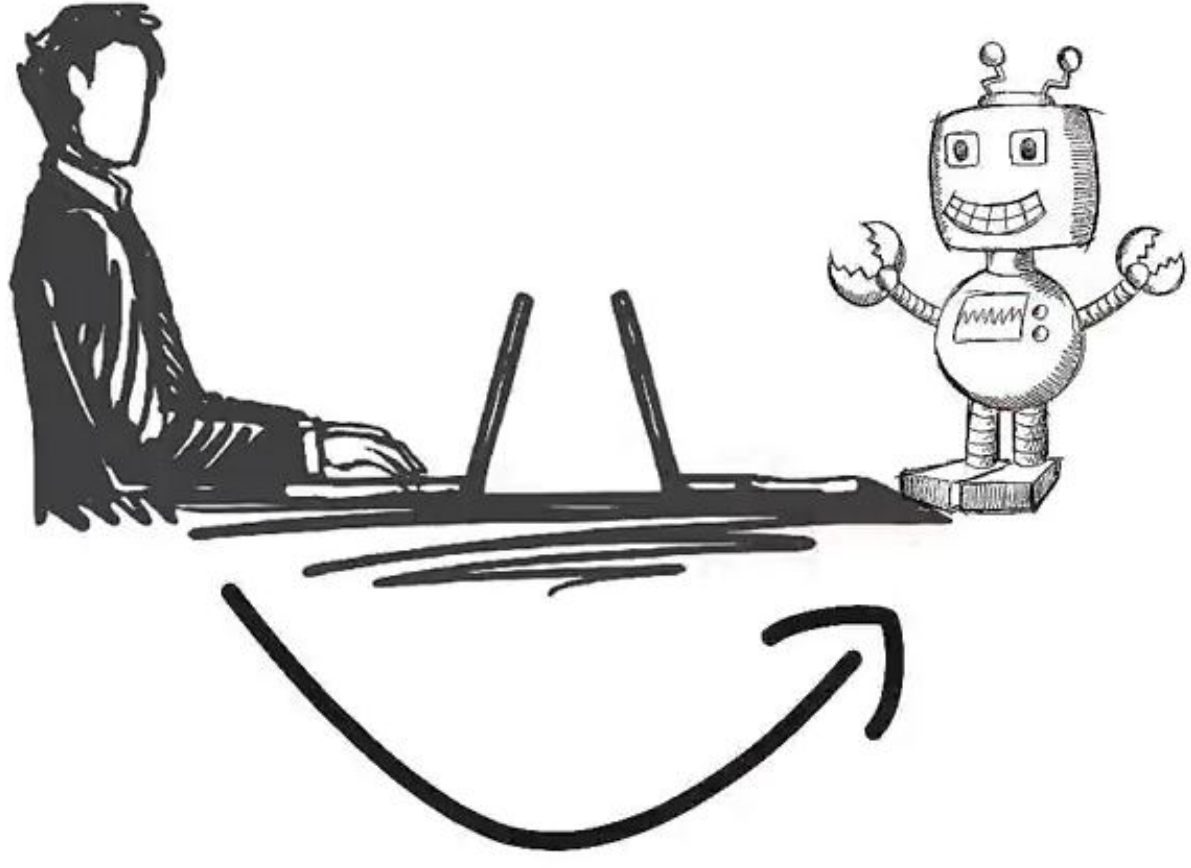
- Neural Networks + Keras and Deep Neural Networks

2. What is Machine Learning

Humans learn from past experiences

Machines follow instructions given by humans

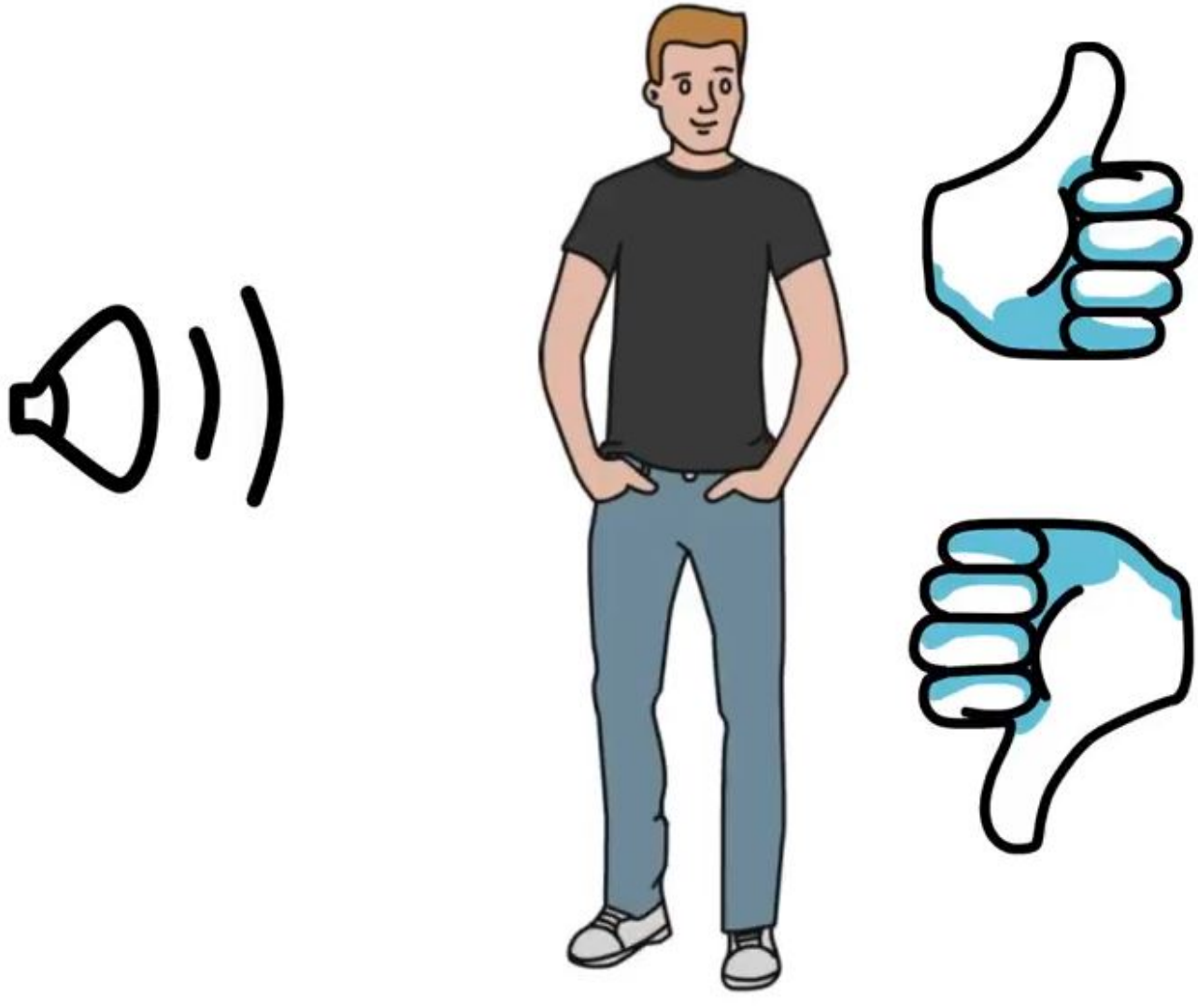
What if humans can train the machines ...



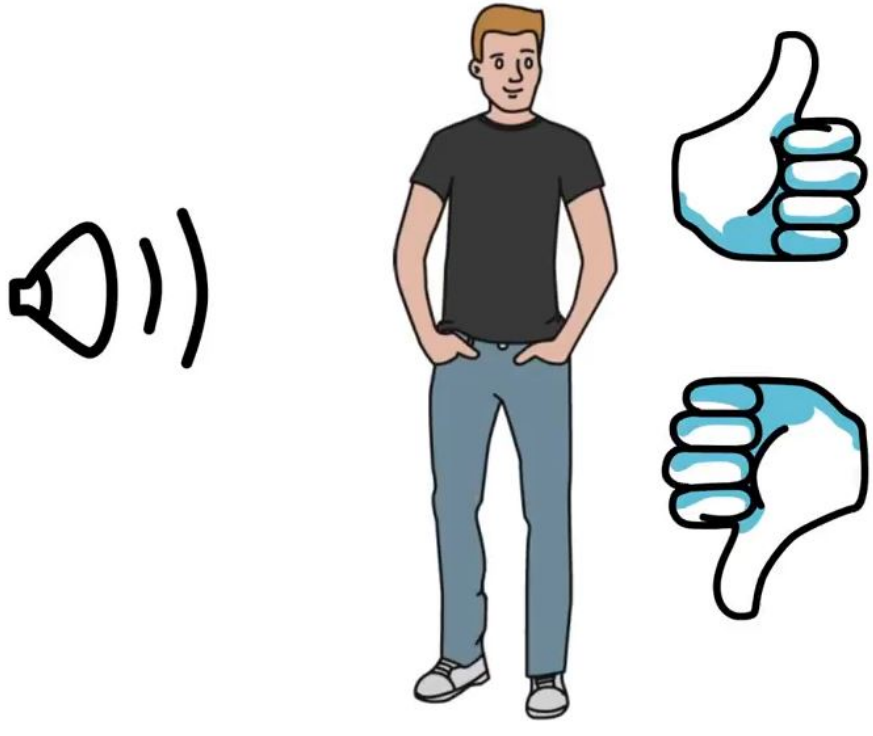
- This is what is called Machine Learning!
- But, it is more than just **learning**!
- It is also about **underestanding** and **reasoning**.

Basics of Machine Learning

- This is Paul.
- Suppose Paul is listening to songs ...

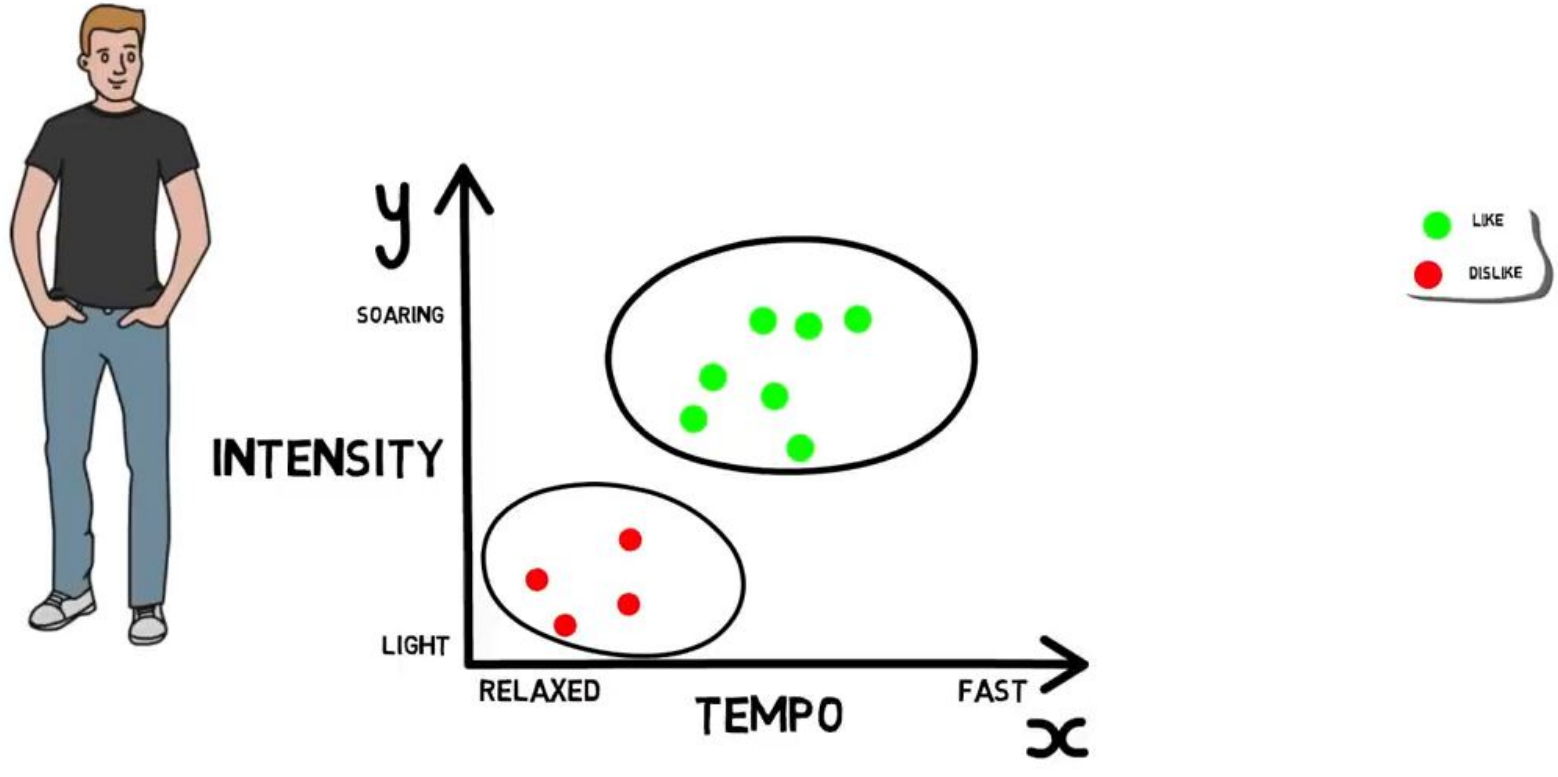


He decides based on ...



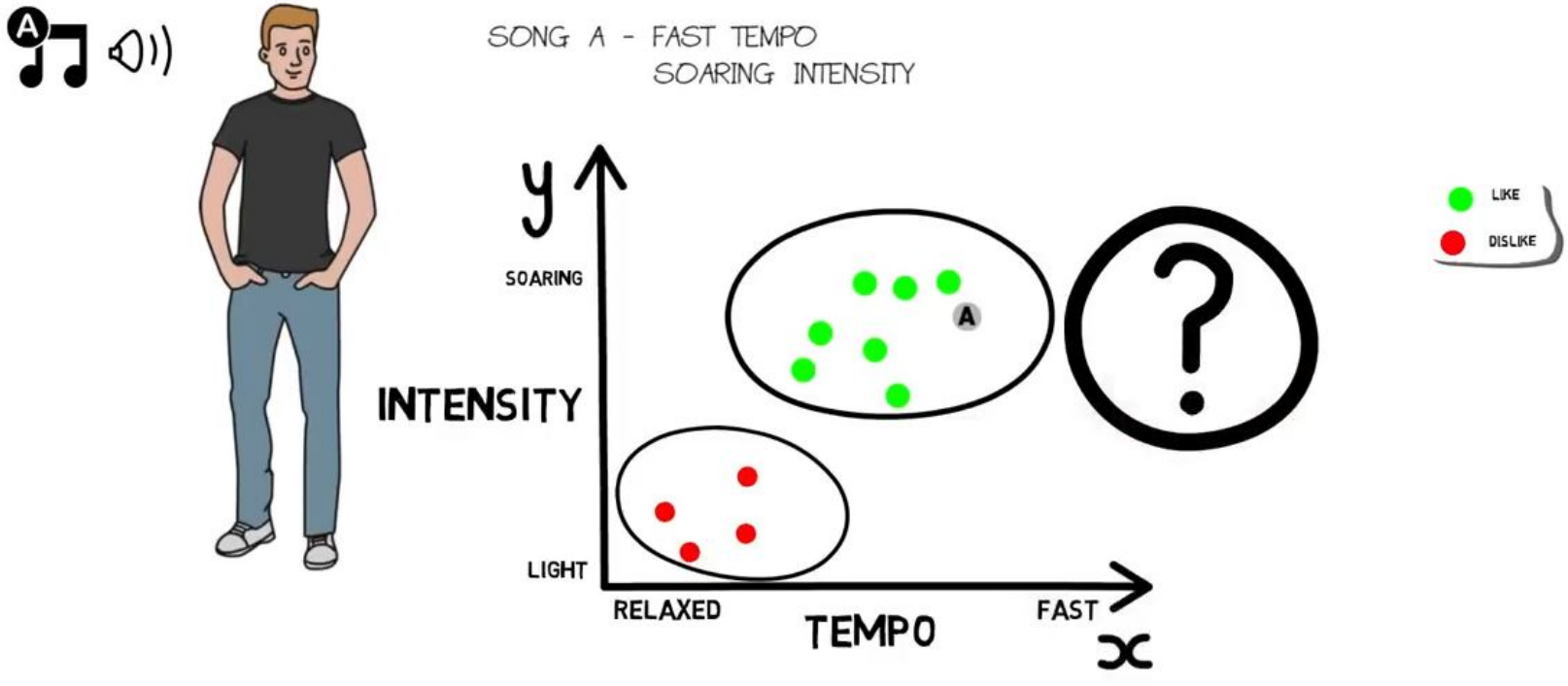
- TEMPO
- GENRE
- INTENSITY
- GENDER OF VOICE

Let's only look at the **tempo** and **intensity** ...



Now, we know Pual's choices!

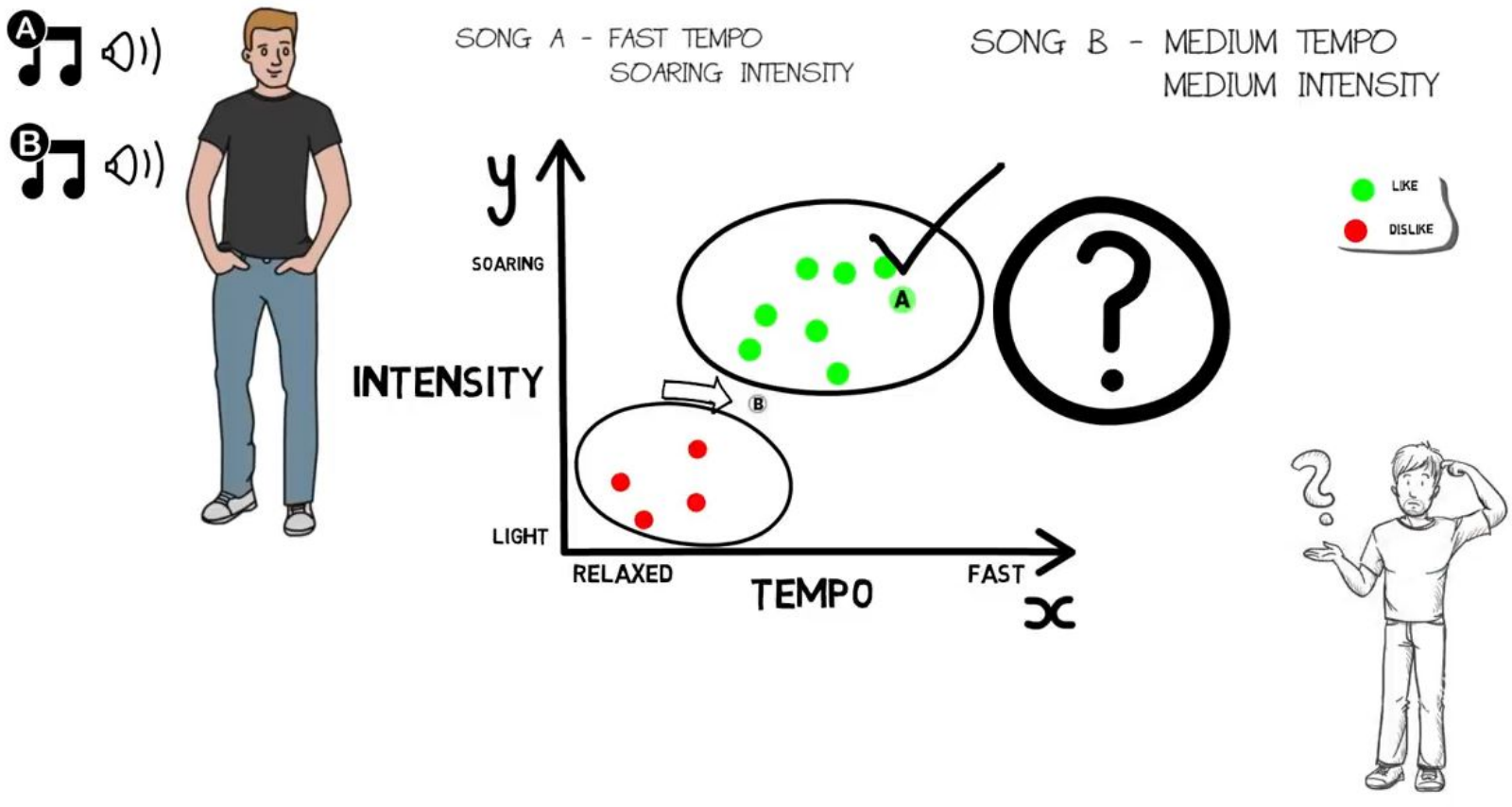
Let's say that Paul listens to a new song A ...



looking at the data, can you guess whether Paul will like the song or not?

- Looking at the Paul's past choices, we were able to classify song A.

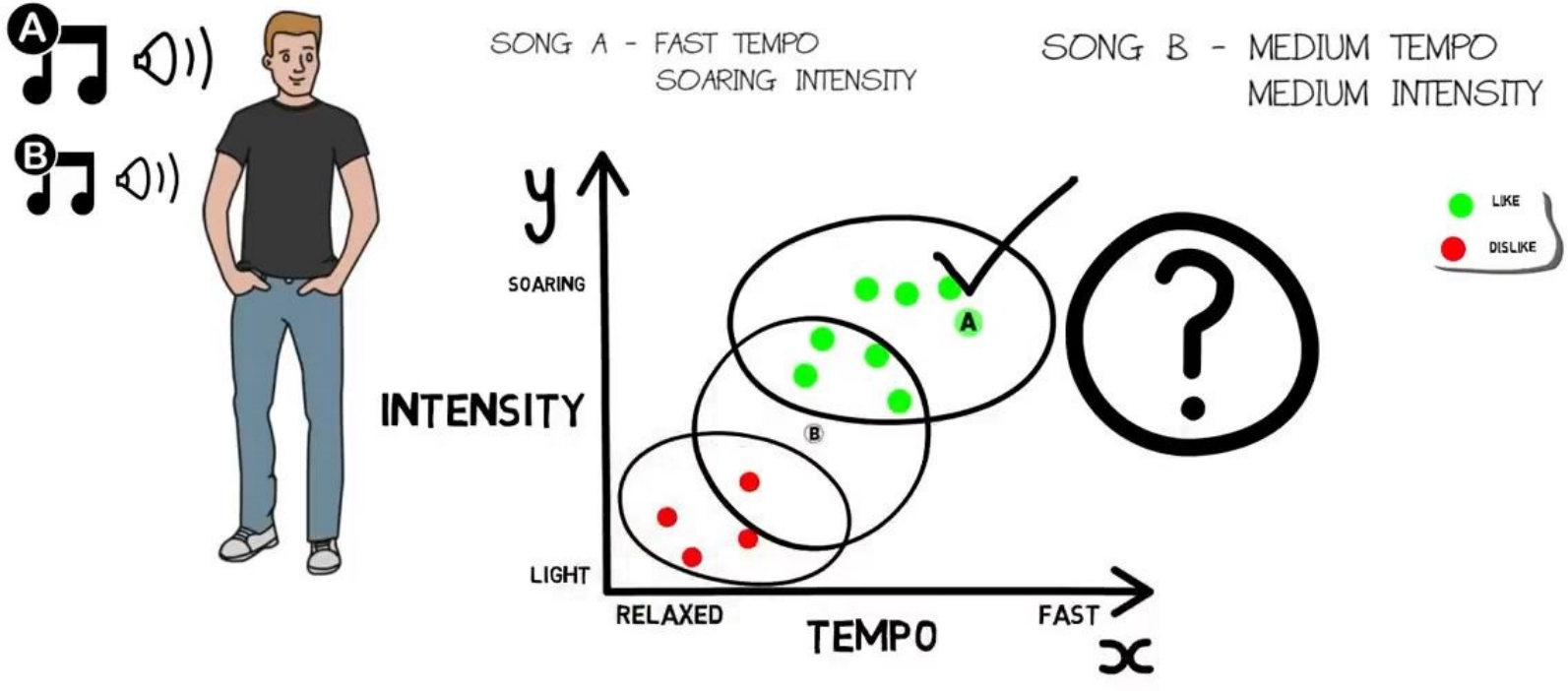
Let's look at another song B ...



Now, can you guess whether Paul likes Song B or not?

That is where Machine Learning comes in ...

What if we draw a circle around Song B?



- We see that there are 4 votes for like whereas there is only 1 vote for dislike!
- Based on the votes, we can see that Paul will definitely **like the song**.
- This is a simple example of a basic machine learning algorithm called **K-Nearest Neighbors** algorithm.

A Basic Rule of Thumb in Machine Learning

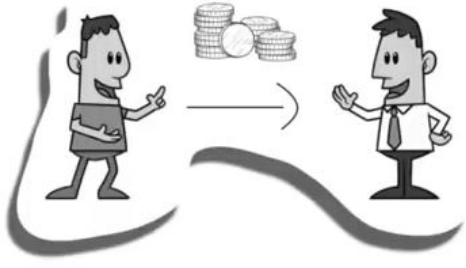
MORE DATA -> BETTER MODEL -> HIGHER ACCURACY

3. Machine Learning Categories

There are many ways in which a machine learns:

Supervised Learning

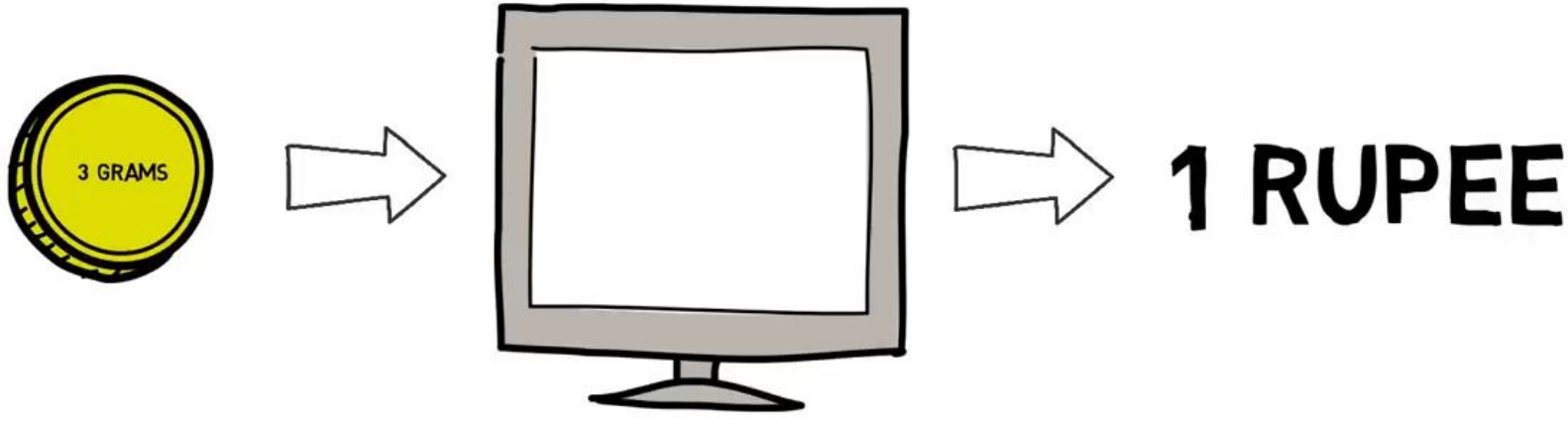
- Suppose your friend gives 1 million coins of 3 different currencies



Your model predicts the weight of each currency.

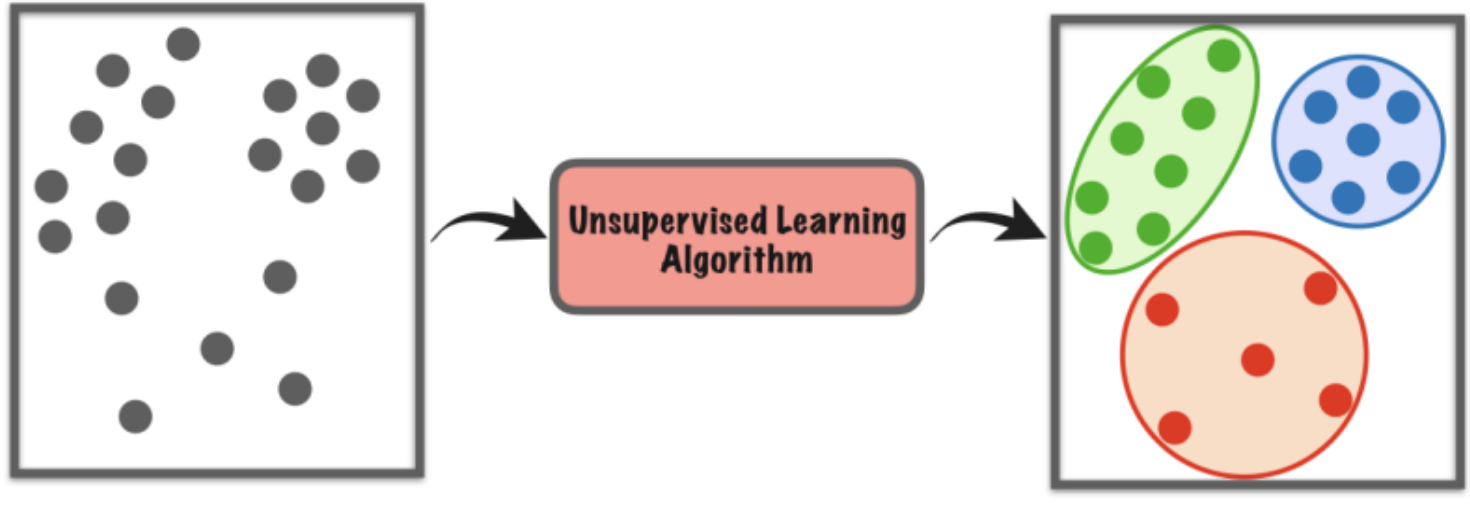
- Weight = Feature
- Currency = Label Machine Learning model **learns** from the **data** of which **feature** is associated with which **label**.

Let's give a new coin to the machine ...



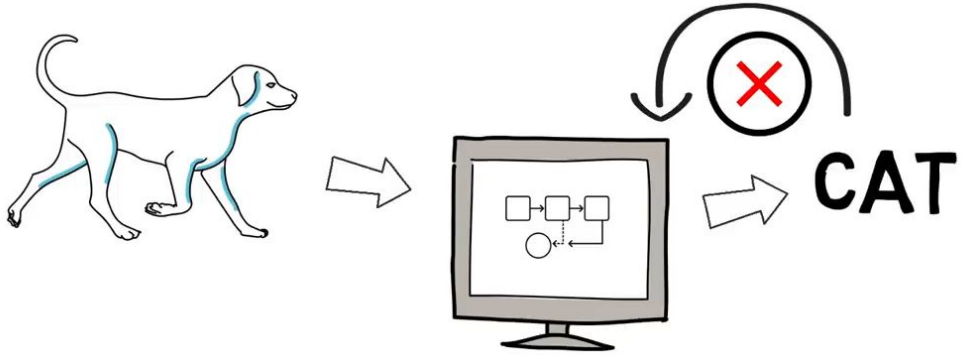
Unsupervised Learning

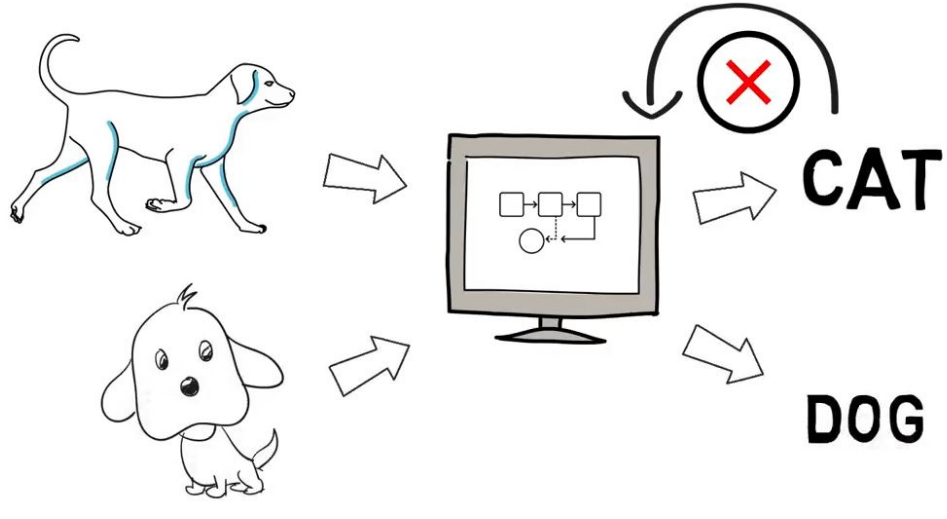
- There is no **Labeled Data**



Reinforcement Learning (Reward Based Learning)

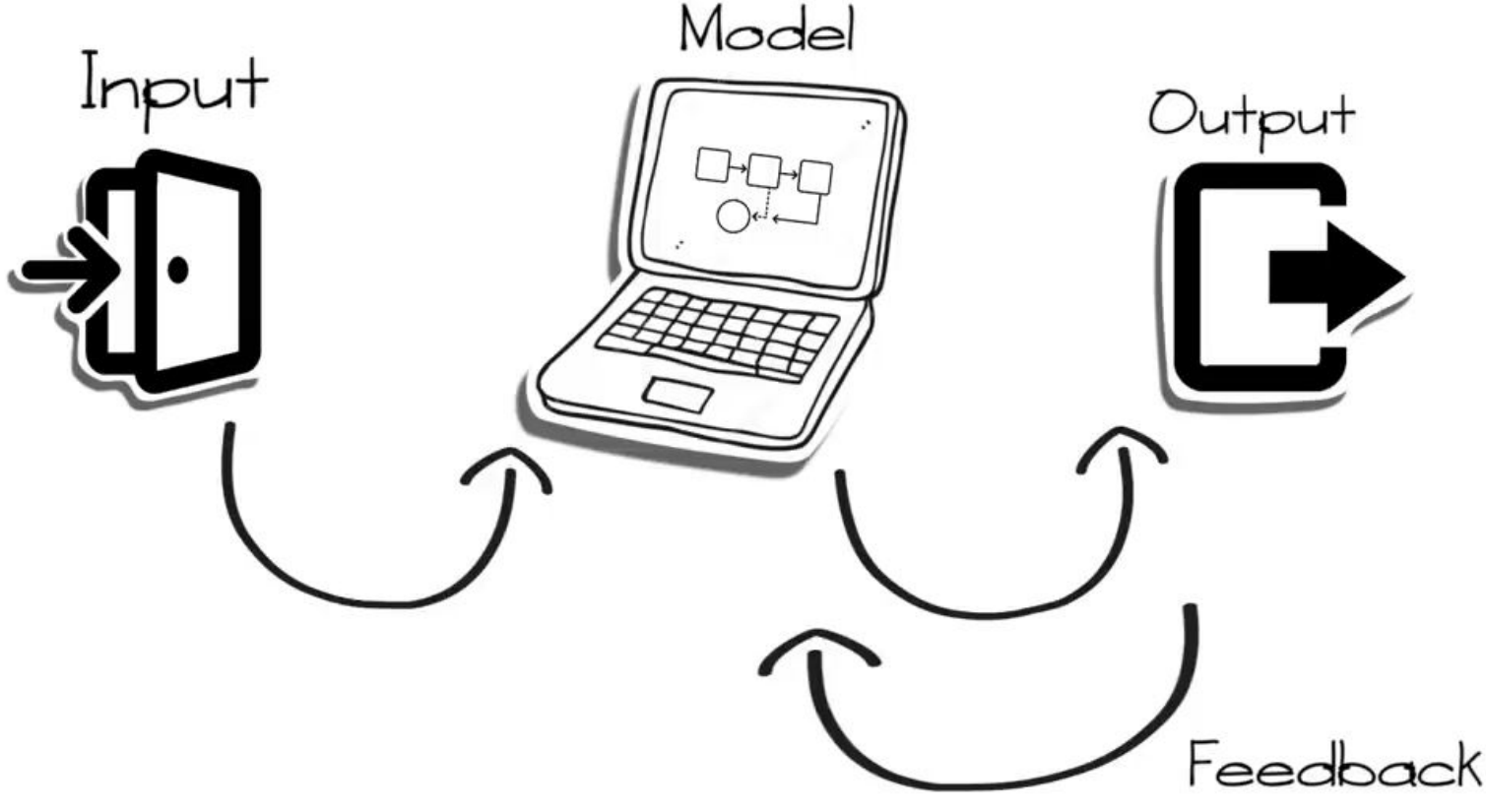
- Let's say that you provide the system with an image of a dog and ask if it can identify it ...
- If it identifies it as a **cat**, you give **negative feedback**.





Generalized Machine Learning Model/Workflow

1. Input is given to a machine learning model, which then gives an output.
2. If the output is right, we take the output as a final result.
3. Else, we provide feedback to the model and ask it to predict until it learns.



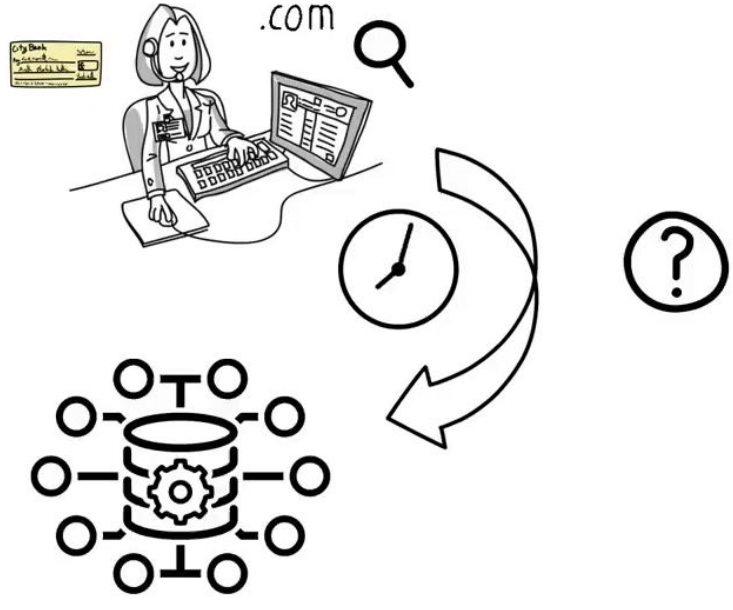
Quiz

Determine whether the below scenarios are **Supervised** or **Unsupervised**?

Scenario 1: Facebook recognizes your friends in a picture from an album and tagged photographs.
Scenario 2: Netflix recommends new movie based on someone's past movie choices.
Scenario 3: Analysis bank data for suspicious transactions and flagging flag transactions.

Why Machine Learning Is Possible Today?

- Everybody is online either using cellphones or just surfing the internet.
- That is generating a huge amount amount of data every minute.



In addition,

- The memory handling capabilities of computers have extensively increased.
- Computers has also now great computational power.

4. Machine Learning Applications

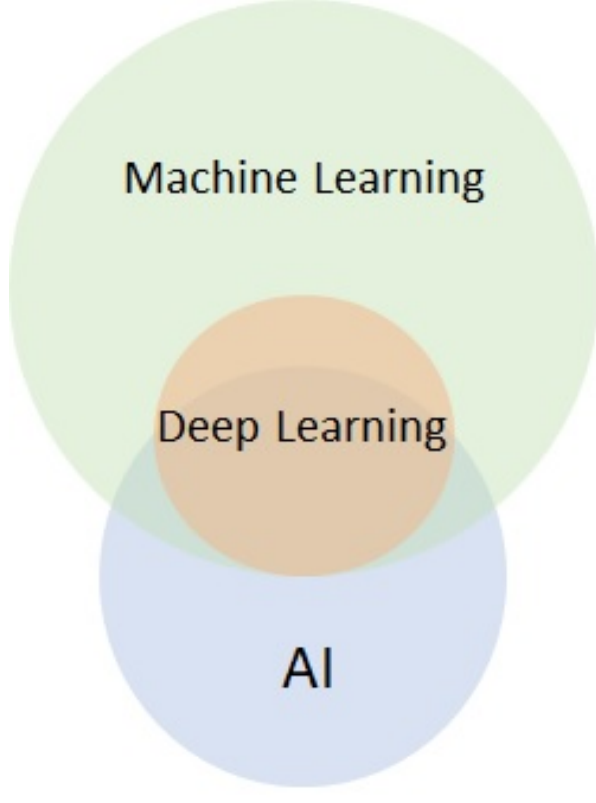
* 1
d
a
p
P
e
n
3
r
n
4
C
a
p
&
d
S
b

* **Healthcare** 1. Patient Risk Identification 2. Identifying diseases and diagnosis 3. Personalized medicine 4. Smart health records 5. Medical imaging diagnosis

Impact of Machine Learning Research in Health- and Energy-Informaics

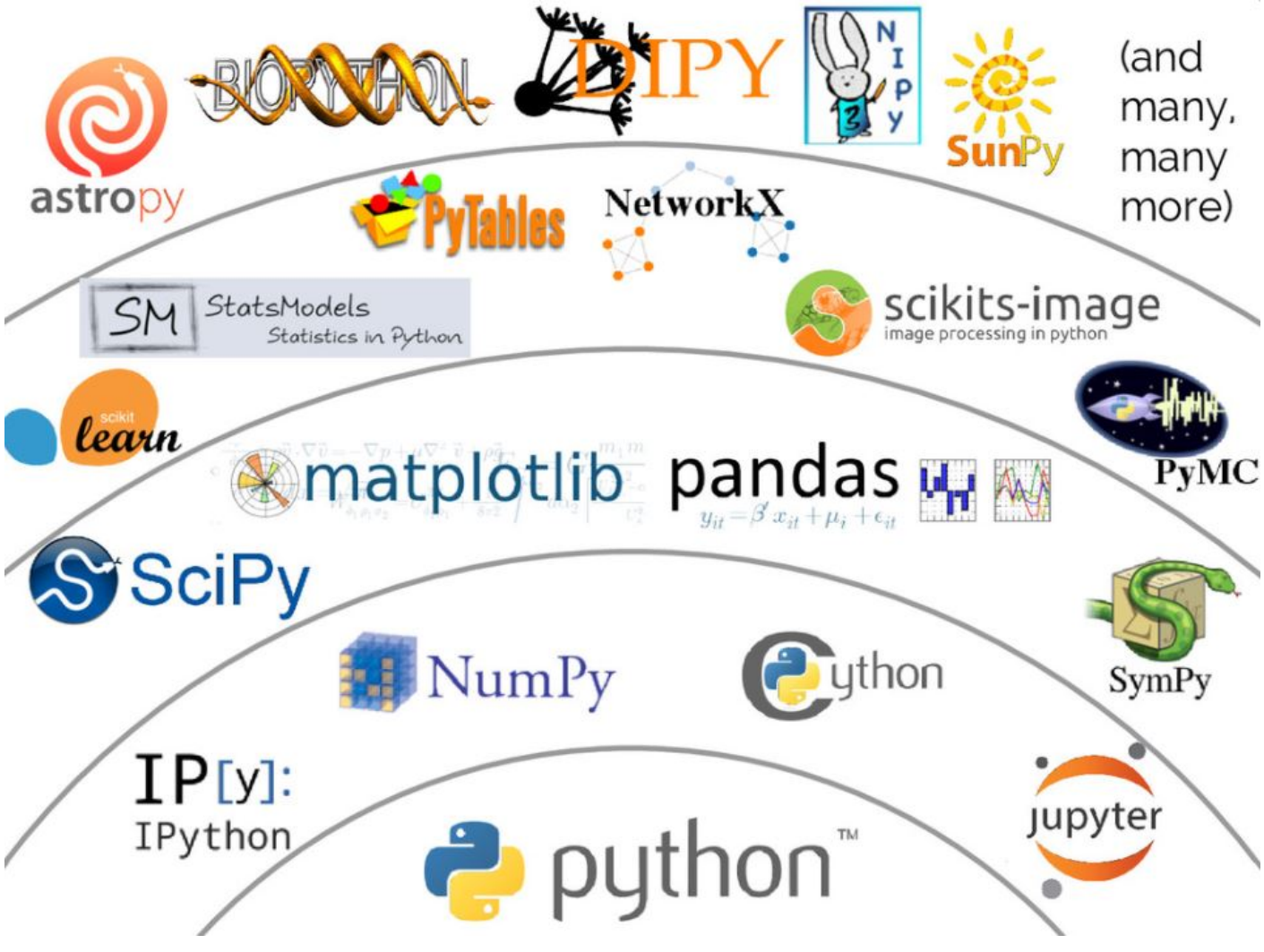
5. Machine Learning Ecosystem

Machine Learning, AI, and Deep Learning



- **AI:** A system that achieves intelligence through rules.
- **ML:** Algorithms that learn the rules and representations from data *automatically*.
- **DL:** Algorithms that learn the parameters of multilayer neural networks to extract the representation of data with multiple layers of abstracion.

Python and Machine Learning



https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote?slide=8 (https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote?slide=8)

6. Course Setup

6.1 Course Syllabus

- You can find full course syllabus and plan here (https://sdu.itslearning.com/ContentArea/ContentArea.aspx?LocationID=18416&LocationType=1).

6.2 Course goal

This course is intended to train students to:

- be able to apply ML algorithms and methods in practice
- have the skills to consider the pros and cons of different ML methods
- be able to choose appropriate ML methods for different applications/problems
- design and implement ML models in Python
- be able to document and present the obtained results using appropriate measures

6.3 Course materials:

- Introduction to machine learning with python (https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/), By Andreas C. Müller, Sarah Guido. O'Reilly Media, 2016. (we refer to this as “IntroML”)
- [Python Data Science Handbook], by Jake VanderPlas (free online book). (we refer to this as “PyDS”) (https://github.com/jakevdp/PythonDataScienceHandbook (https://github.com/jakevdp/PythonDataScienceHandbook))
- Whirlwind Tour of Python, by Jake VanderPlas (free online book). (we refer to this as “PyTour”) (https://jakevdp.github.io/WhirlwindTourOfPython/ (https://jakevdp.github.io/WhirlwindTourOfPython/))
- Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, by Aurélien Géron. O'Reilly Media, 2019. (we refer to this as “Hands-on ML”) (https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ (https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/))

NOTE: You will also need a laptop, which is capable of displaying and outputting graphics

6.4 Computational setup

Anaconda: Anaconda (https://www.anaconda.com/) can be downloaded and installed in order to use Jupyter Notebooks as a stand-alone solution on your machines/laptops.

- We will use Anaconda to to develop and data analysis.
- This will provide us an easy-to-use tool for writing text, code, and generating plots all in a single format called “notebook”. And many more!
- It is also free!

Colaboratory: You may also use Google Colaboratory (https://colab.research.google.com/notebooks/intro.ipynb).

- This will provide us an easy-to-use tool for writing text, code, and generating plots all in a single format called “notebook”.
- It is a free tool.
- In addition, Colaboratory runs all the codes on Google Cloud servers rather than your personal computer. This helps a lot to run the codes hassle-free.

Git & GitHub: In addition to itslearning, we will use GitHub (https://github.com/) to manage and share code and data.

- GitHub is a very efficient way of managing multiple versions of data and codes.
- You can share your codes with others easily.
- Again, is is free!
- Please create a GitHub account (free) **with your SDU email address** if you have not already done it.
- You will also need to download and install Git (https://git-scm.com/downloads) on your machines/laptops.

6.5 Zoom

- Lectures will be recorded and posted to itslearning so that all students can access or revisit the lectures.
- If you do not want to appear in the recording, or if you want me to exclude a specific comment or time interval from the recording, please let me know.
- Recordings of my office hours will not typically be posted online, with the possible exception of snippets that I think may be relevant to the entire class.

Reach out to me if you have requests or concerns!

6.6 Grading

Assignments

- Assignments contribute to the 5% of the final grade.
- All course assignments will be assigned a point value, added together, and converted to the nearest equivalent grade as follows:

(92-100) -> 12
(84-91) -> 10
(68-83) -> 7
(60-67) -> 4
(50-59) -> 02
(20-49) -> 00
(0-19) -> -3

Written report

- The written report contributes to the 10% of the final grade
- The report of each group will be assessed in the following format:

Abstract: 15 pts

- Is enough information provided get a clear idea about the subject matter?
- Are the main points of the report described succinctly?

Introduction: 15 pts

- Does the introduction cover the required background information to understand the work?
- Is the introduction well organized: it starts out general and becomes more specific towards the end?
- Is there a motivation explaining why this project is relevant, important, and/or interesting?

Related Work: 15 pts

- Is the similar and related work discussed adequately?
- Are references cited properly (here, but also throughout the whole paper)?
- Is the discussion or paragraph on comparing this project with other people's work adequate?

Proposed Method: 25 pts

- Are there any missing descriptions of symbols used in mathematical notations (if applicable)?
- Are the main algorithms described well enough so that they can be implemented by a knowledgeable reader?

Experiments: 25 pts

- Is the experimental setup and methodology described well enough so that it can be repeated?
- If datasets are used, are they referenced appropriately?

Results and Discussion: 30 pts

- Are the results described clearly?
- Is the data analyzed well, and are the results logical?
- Are the figures clear and have no missing labels?
- Do the figure captions have sufficient information to understand the figure?
- Is each figure referenced in the text?
- Is the discussion critical/honest, and are potential weaknesses/shortcomings are discussed as well?

Conclusions: 15 pts

- Do the authors describe whether the initial motivation/task was accomplished or not based on the results?
- Is it discussed adequately how the results relate to previous work?
- If applicable, are potential future directions given?

Contributions: 10 pts

- Are all contributions listed clearly?
- Did each member contribute approximately equally to the project?

The report will be assigned a point value and converted to the nearest equivalent grade as follows:

(131-150) -> 12
(105-130) -> 10
(80-104) -> 7
(60-79) -> 4
(40-59) -> 02
(20-39) -> 00
(0-19) -> -3

Midterm project progress presentation

- The midterm project progress presentation contributes to the 5% of the final grade.
- On Week 15, each group present their project progress (even though it is not completed yet) to the class.
- The presentation should cover the following:
    1. *introduce the project and the topic to the class.*
    2. *discuss the main method*
    3. *present the results of the analysis*

- Each presentation should be maximum 8 minutes, and there will be 2 minutes for questions and answers.
- All the group members should participate in the presentation.

- There will be also three categories:
    1. *Oral presentation*
    2. *Visualization*
    3. *Creative approach*

- The winner(s) will be determined by other students' votes.
- The voting will be conducted using PollEverywhere software (will be introduced later on).
- Each student votes as follows:
    - Group 1: (Best oral presentation)/10, (Best visualization)/10, (Most creative approach)/10
    - Group 2: (Best oral presentation)/10, (Best visualization)/10, (Most creative approach)/10
    - Group 3: (Best oral presentation)/10, (Best visualization)/10, (Most creative approach)/10
    - ...

NOTE: Each vote will provide 2.5 bonus points for your group. This means that if all the group members vote for all the presentations, your group project receives 10 bonus point.

Final exam (presentation)

- Final exam contributes to the 80% of the final grade.
- Each student (individually) will give a presentation based on the report.
- Afterwards, there will be questions based on their presentation, report, and the whole curriculum.

Final Grade Calculation

The weighted average grade will be calculated and will be rounded to the highest possible grade. The final grade will be calculated as:

$$\begin{aligned}&\text{Final grade}\\&= \text{ceiling}([0.05 \times \text{assignments} + 0.05 \times \text{midterm presentation} + 0.1 \times \text{report}\\&+ 0.8 \times \text{final exam}])\end{aligned}$$

For example:

$$\text{ceiling}([0.05 \times 7 + 0.05 \times 12 + 0.1 \times 7 + 0.8 \times 10]) = 10$$

NOTE: In case a student fails the final exam (0 or -3), the 20% for assignments, project report, and midterm presentation grades will NOT be considered in the final grade. This means that s/he will get 0 or -3 as her/his final grade.

6.7. Group Project and Report Template

- The students will team up in groups of 2-4 people to do their project and write the report at the beginning of the semester.
- One of the team members should send the list of team members to me **no later than 15th of February**.
- Otherwise, you will be randomly assigned to a group.
- Each group will choose a dataset from the list **no later than 22nd of February**.

Energy:

UCI Data Sets

- Energy Efficiency Data Set (https://archive.ics.uci.edu/ml/datasets/Energy+efficiency
- Appliances Energy Prediction Data Set (https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction)
- Condition Monitoring of Hydraulic Systems Data Set (https://archive.ics.uci.edu/ml/datasets/Condition+monitoring+of+hydraulic+system
- Electrical Grid Stability Simulated Data Data Set (https://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+)
- Condition Based Maintenance of Naval Propulsion Plants Data Set (https://archive.ics.uci.edu/ml/datasets/Condition+Based+Maintenance+of+Naval+P
- Gas Turbine CO and NOx Emission Data Set Data Set (https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data
- SML2010 Data Set-Indoor Temperature Forecasting (https://archive.ics.uci.edu/ml/datasets/SML2010)

Kaggle

- NASA Turbofan Jet Engine Data Set (https://www.kaggle.com/behrad3d/nasa-cmaps)
- Railway Track Fault Detection (https://www.kaggle.com/salmaneunus/railway-track-fault-detection)
- Appliances Energy Prediction (https://www.kaggle.com/loveall/appliances-energy-prediction)
- Power Grid Fault Detection Data (https://www.kaggle.com/himanshu36/time-series-classification/notebooks)
- Solar Power Generation Data (https://www.kaggle.com/anikannal/solar-power-generation-data)
- LBNL Automated Fault Detection for Buildings Data (https://www.kaggle.com/claytonmiller/lbnl-automated-fault-detection-for-buildings-data)

Health:

UCI Data Sets

- Breast Cancer Coimbra Data Set (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra)
- Z-Alizadeh Sani Data Set-Coronary Artery Disease (https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani)
- Heart Failure Clinical Records Data Set (https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records)

PhysioNet

- MIT-BIH Atrial Fibrillation Database (https://www.physionet.org/content/afdb/1.0.0/)
- MIT-BIH Arrhythmia Database (https://www.physionet.org/content/mitdb/1.0.0/)

Kaggle

- Disease Symptom Prediction (https://www.kaggle.com/itachi9604/disease-symptom-description-dataset)
- Healthcare cost (https://www.kaggle.com/ravichaubey1506/healthcare-cost)

- The deadline for submitting the detailed final project report will be on **17th of May at 23:00**.
- Remember that you should **submit both the report (PDF and .tex files) and the Python codes** you used for this project via itslearning.
- Also, only one member per team needs to submit the project material.
- The project report should be **maximum 20 pages long (not counting references)** and should contain the sections that are already provided in the LaTeX project template on Overleaf (https://www.overleaf.com/read/crtctkcswfcq).
- Please use Overleaf (https://www.overleaf.com/) to write your report.
- Overleaf is an online and collaborative LaTeX editor so that all the team members can see and edit the report.
- You may need to register to use Overleaf if you do not have an account already. **Please use your SDU email address to open an account**.

NOTE: Please read the template thoroughly. There are more details regarding how to use it in there.

6.8 Late Submission Policy

Assignments and projects that are submitted late will be considered as follows:

- If it is submitted within 12 hours of the deadline (late), there will be 10% deduction from the points.
- If it is submitted within 12 and 24 hours of the deadline (late), there will be 20% deduction from the points.
- If it is submitted more than 24 hours of the deadline (late), there will be no points (zero point).

6.9 Proper Academic Practice

I expect you to abide by SDU's Proper Academic Practice (https://mitsdu.dk/-/media/files/information_til/studerende_ved_sdu/eksamen/1proper+aca +november+2016.pdf) at all times.

I encourage you to discuss your assignments and projects with your classmates. However, it is expected that these should be completed by you.

Furthermore, you cannot re-use projects from other sources without modifying them. You should simply submit your own assignments and projects, even if you discussed them with others.

6.10 Scheduling conflicts

I expect you to take part in the course and attend the lectures, in-class discussions, give presentations, and complete assignments/exercises and tutorials.

However, I also understand that in some special circumstances and fixed-schedule activities, you may miss the class.

Please contact me before the end of Week 6 (February 13) to discuss this, if you have any scheduling conflicts.

Feedback during the semester.

Please provide your anonymous feedback here (https://PollEv.com/free_text_polls/bQ5SF66UyZs6fq7yZyVQQ/respond)

Thank you!