Session 1

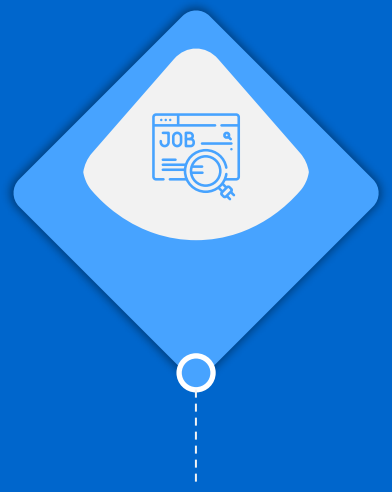
AI and Machine Learning

Hult International Business School

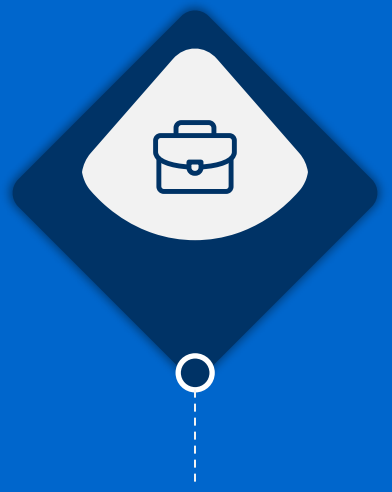
Michael de la Maza

Version 1.7

About you



Are you actively looking for a job?

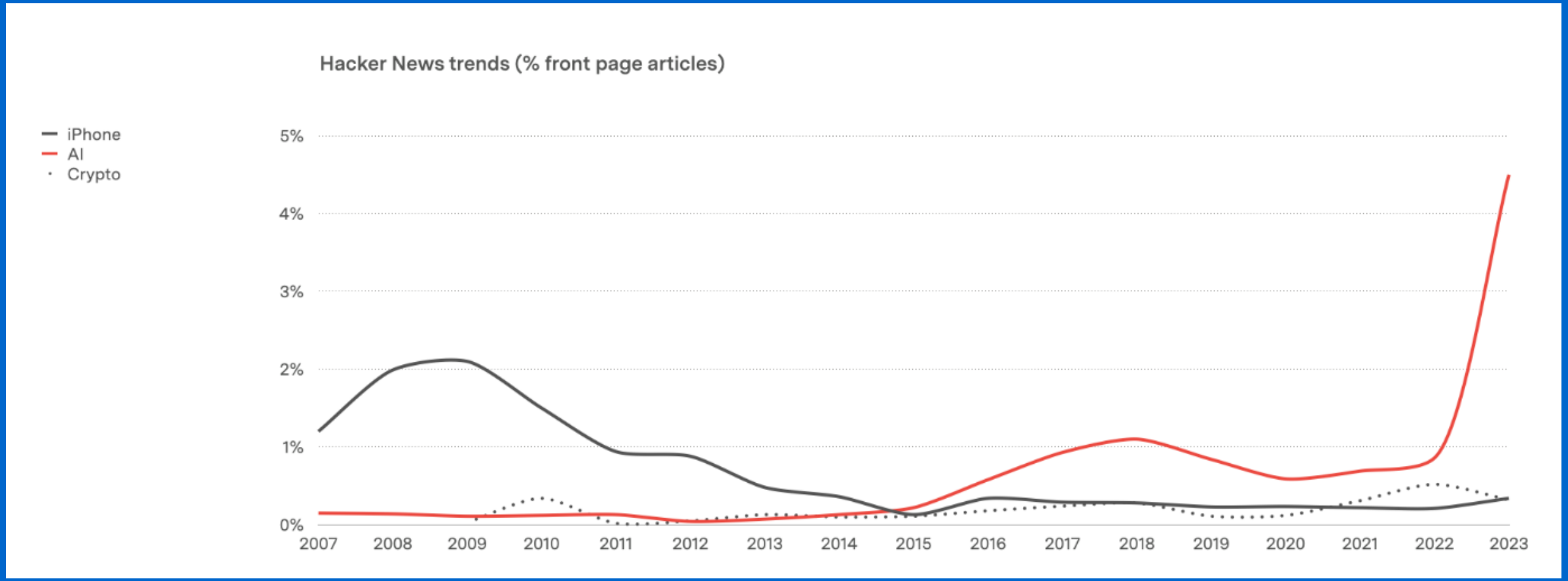


What is your ideal job? Job title?

In what country is your ideal job?

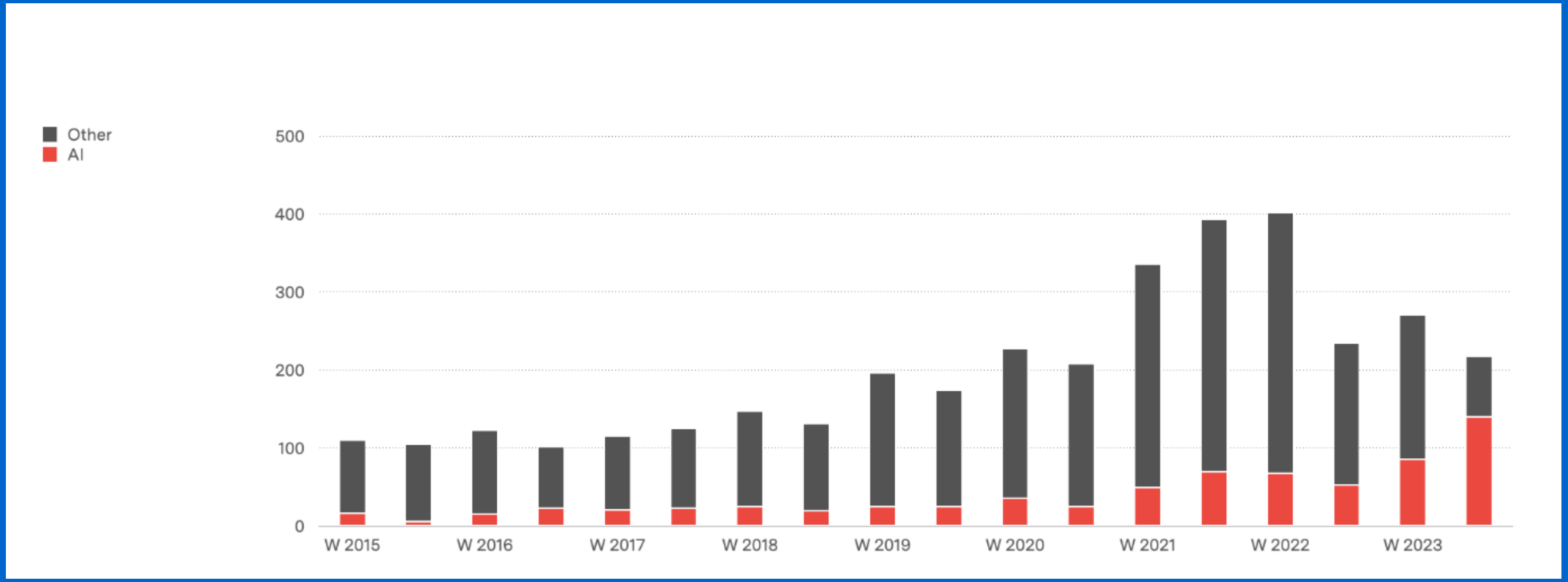
The Rise of AI

AI Dominates Hacker News



Adapted from Benedict Evans

Over half of YC startups are in AI

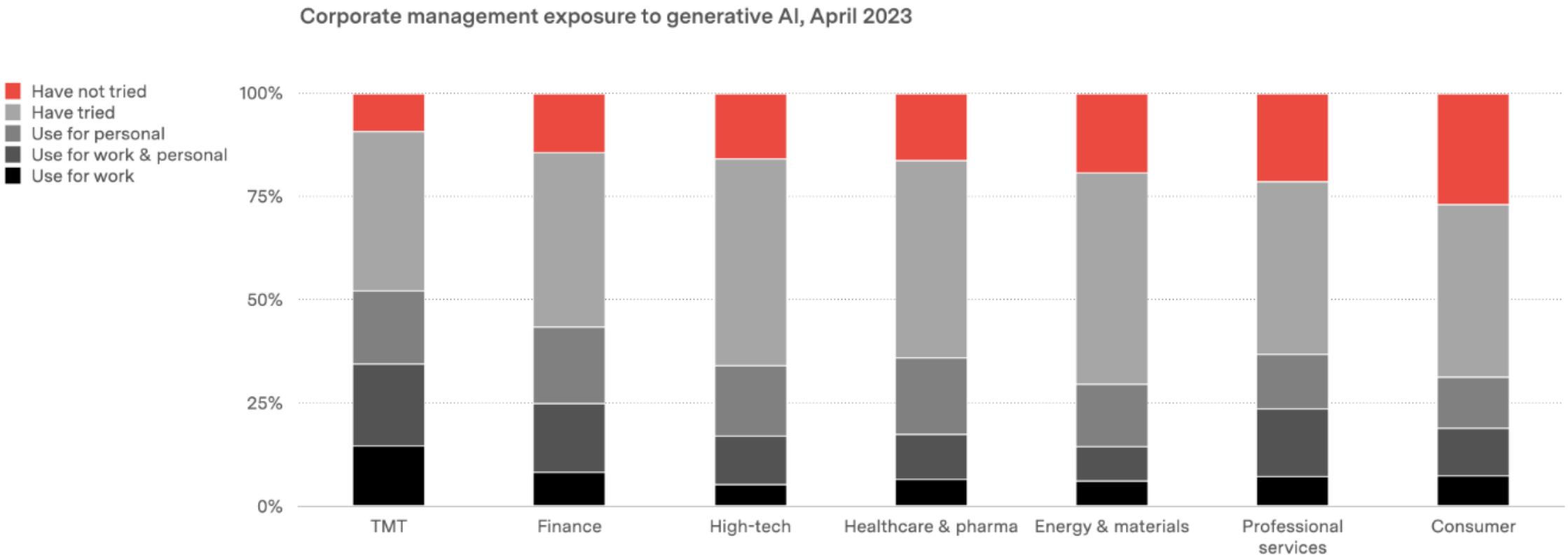


Adapted from Benedict Evans

NVIDIA is now a trillion dollar company

Up more than 10x in 5 years

Over 70% of execs have tried AI



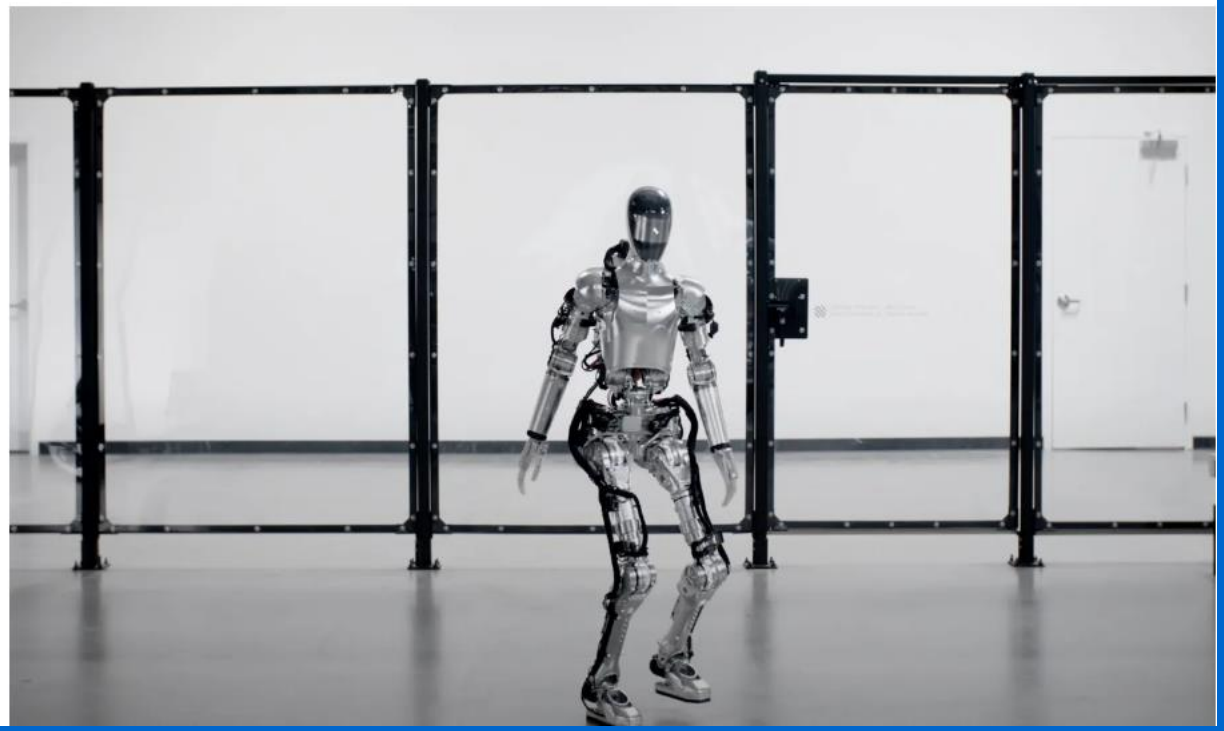
Adapted from Benedict Evans

Extreme advances are now commonplace…

Robotics

BMW will deploy Figure's humanoid robot at South Carolina plant

Brian Heater @bheater / 6:00 AM EST • January 18, 2024 Comment



…creating one person unicorn companies

NfX

Content Team Companies About Focus Areas

The 3-Person Unicorn Startup

James Currier · @jamescurrier · Jul 2023 · Generative AI

How generative AI is changing startups

With the next generation of AI tools, teams of three very talented people will be able to grow software-centric businesses to $100+ million in revenue with automated workflows.

They will be able to set up AI systems for each aspect of their business and let them run, like spinning plates, periodically revisiting them to improve them and keep them spinning.

Source: https://www.nfx.com/post/3-person-unicorn-startup

AI problems are also commonplace

CNN INVESTIGATES

A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning

By Curt Devine, Donie O'Sullivan and Sean Lyngaas, CNN

10 minute read · Updated 6:09 AM EST, Thu February 1, 2024

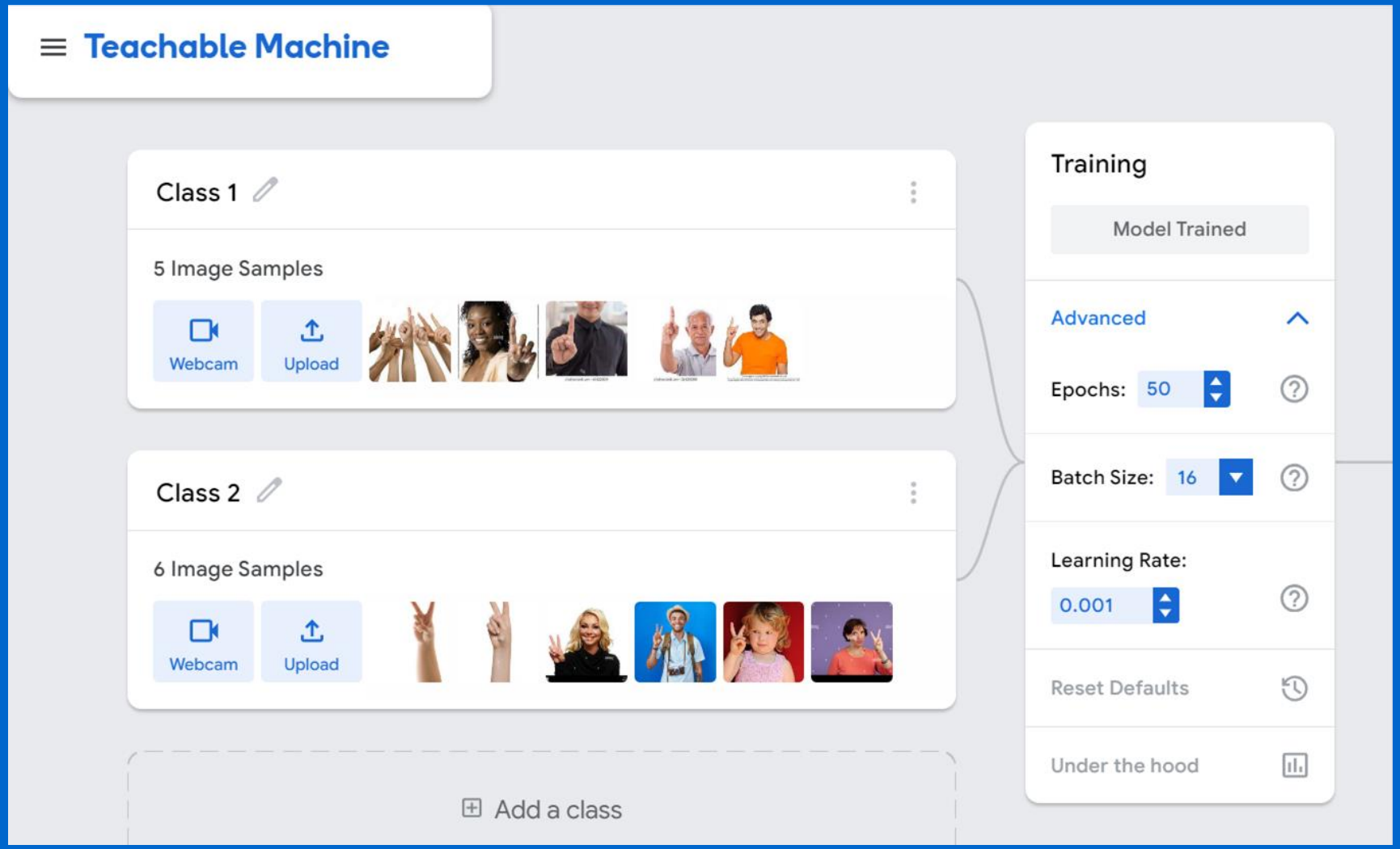
https://www.youtube.com/watch?v=Yd0yQ9yxSYY

Supervised learning
Build an image classifier

Steps to build an image classifier

- **Collect instances of each class.**
  In this example, there are two classes: images of people holding up **one** finger and images of people holding up **two** fingers
- **Train the classifier (“fit”)**
- **Test the classifier by trying unseen instances (“predict”)**

Google Teachable Machine

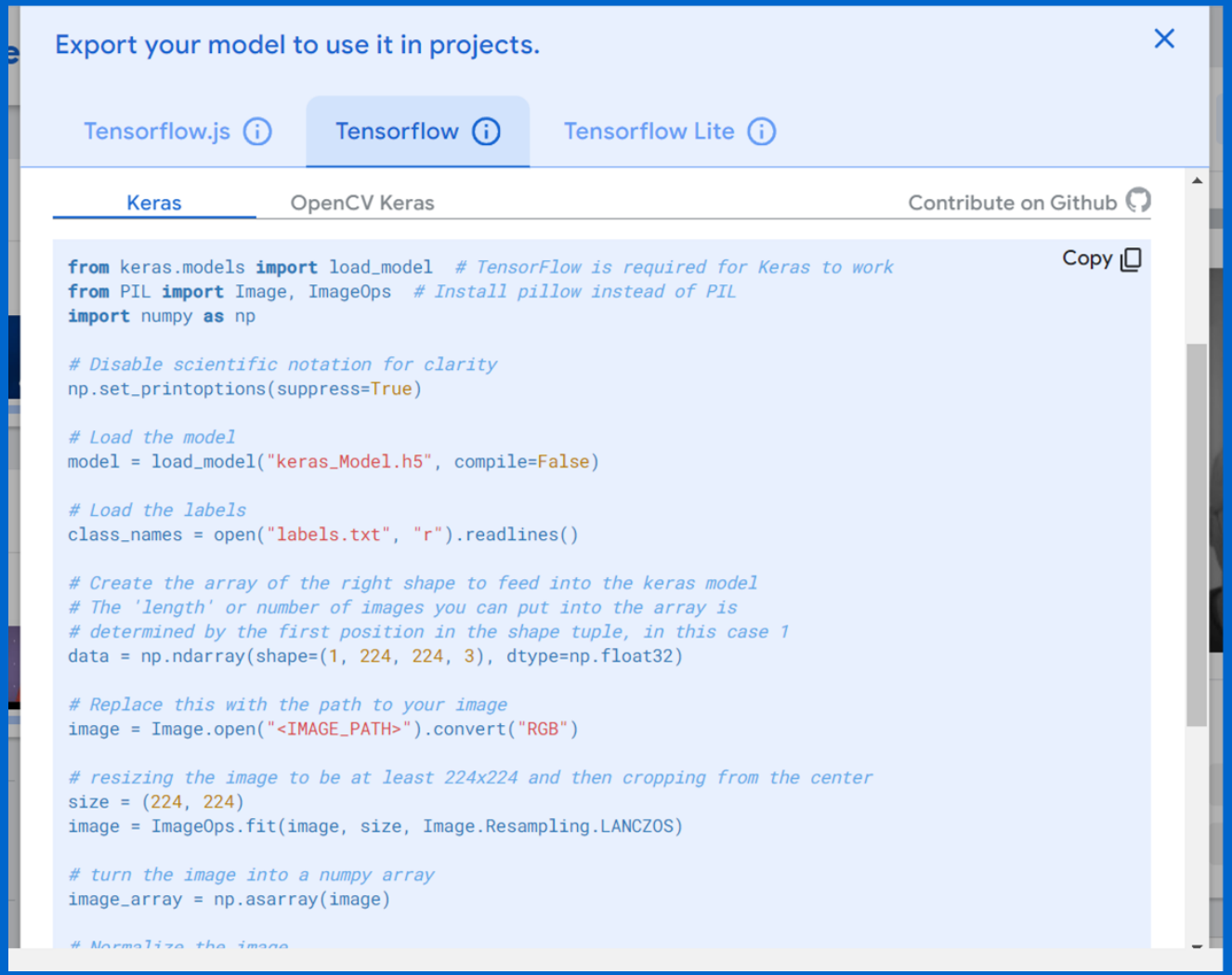


Glossary

Recommendation: Build out an AI/ML glossary using your own words.

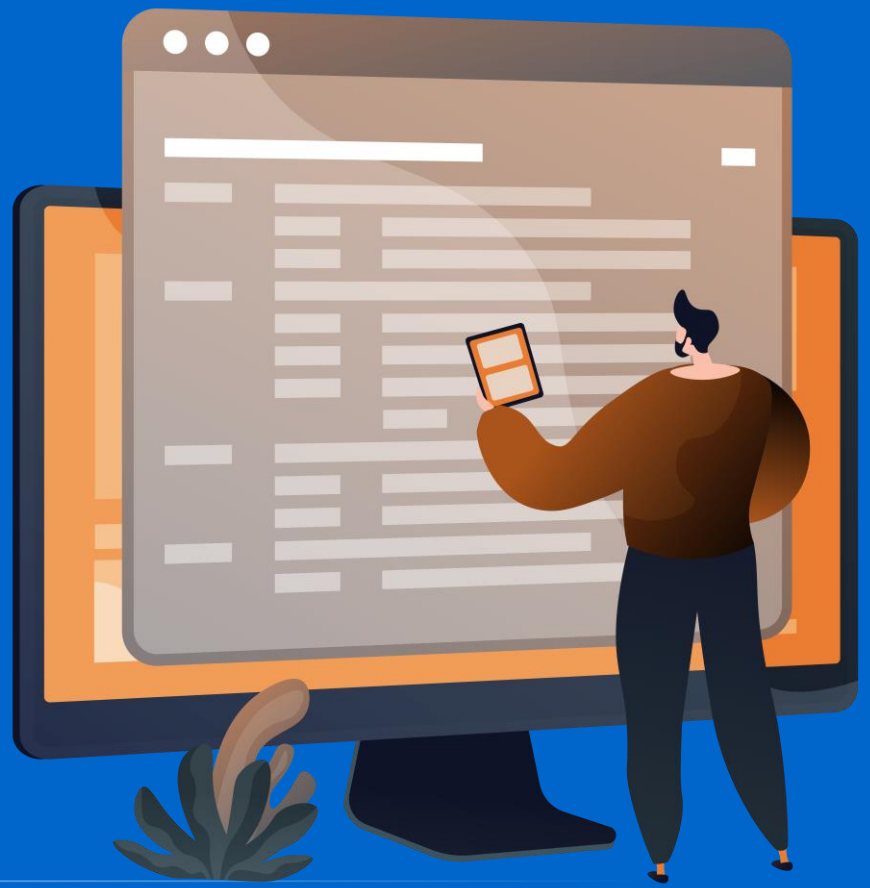
- Instances (or samples)
- Training instances
- Testing instances
- Epoch
- Batch
- Learning rate
- Overfit
- Underfit
- Accuracy
- Confusion matrix
- Loss
- Keras
- Unstructured data
- Structured data

Code simply loads the model and then predicts



Applications of supervised learning

- Is this email spam?
- Is this transaction fraudulent?
- Will this customer churn?
- Does this patient have pneumonia?
- How many people will rent bikes?
- What music does this customer like?



Video: AlphaFold – The Making of a Scientific Breakthrough

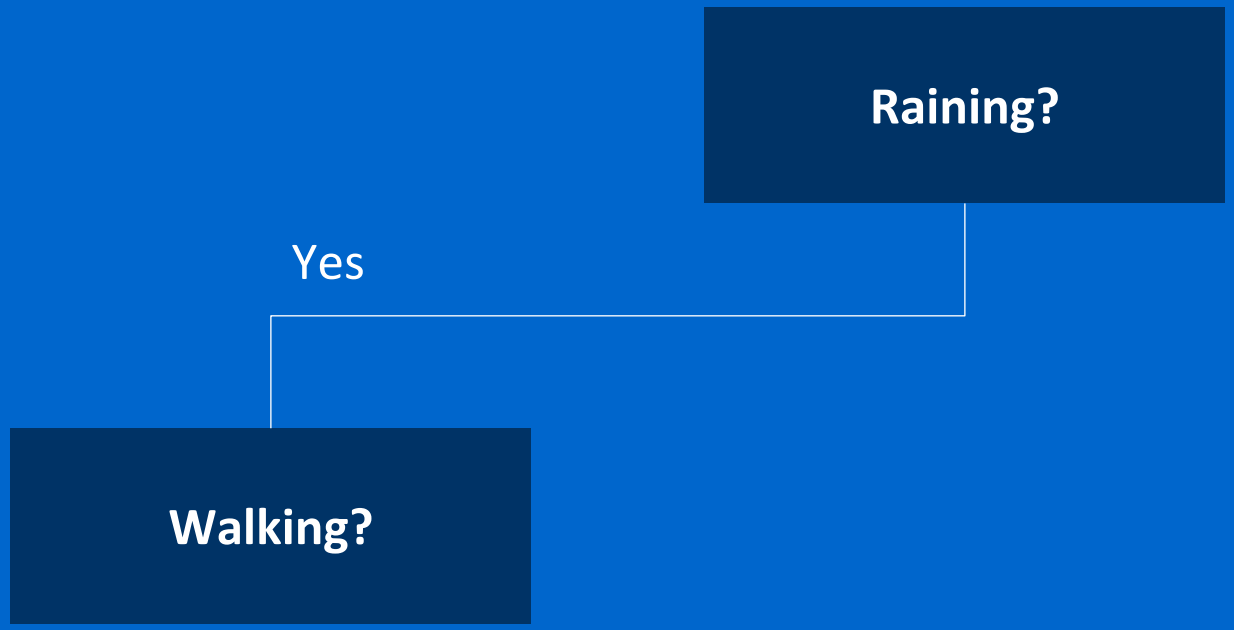
Decision Trees

How to read a decision tree: Wear a raincoat example

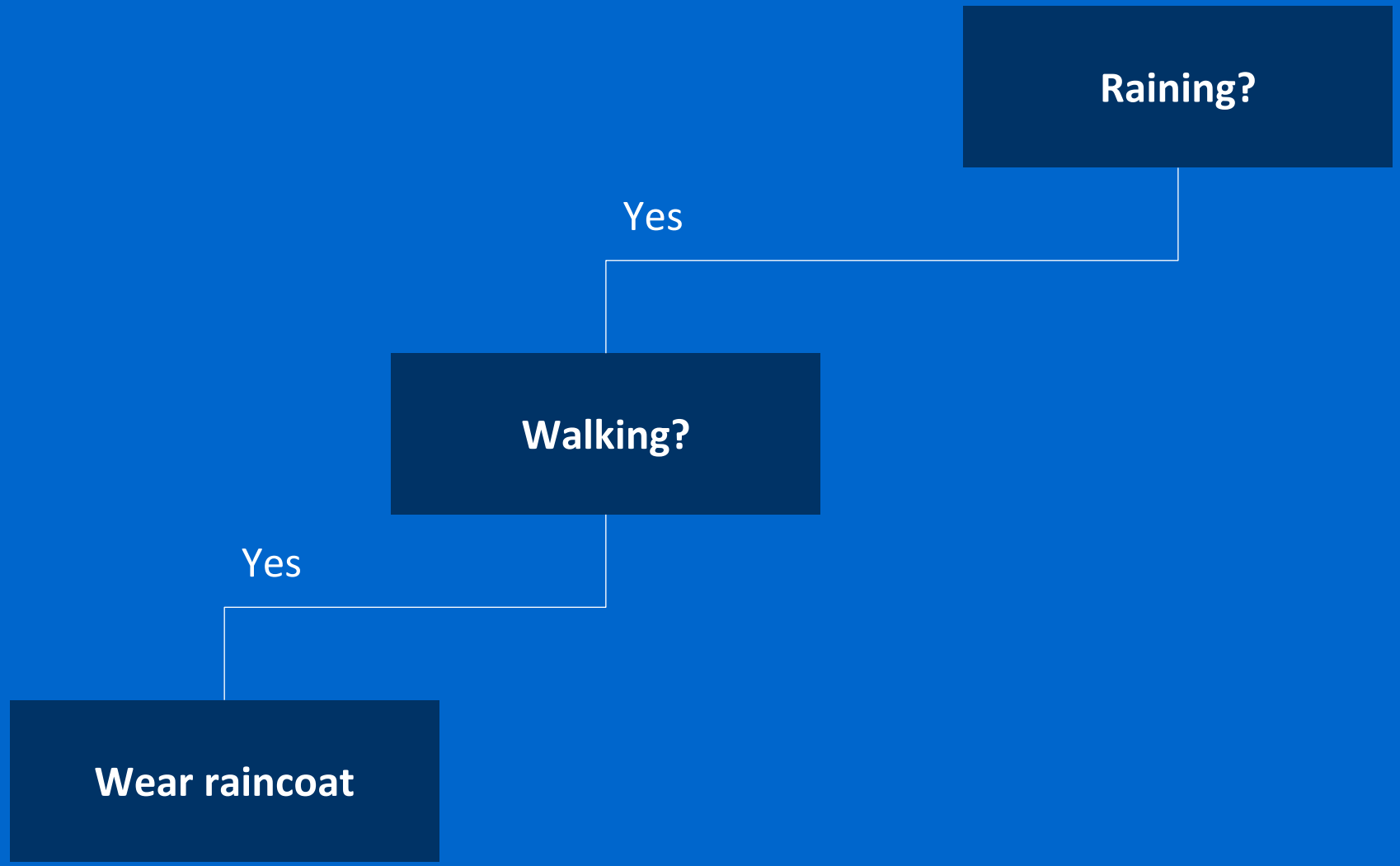
Is it raining?

Raining?

If it is raining, are you walking?

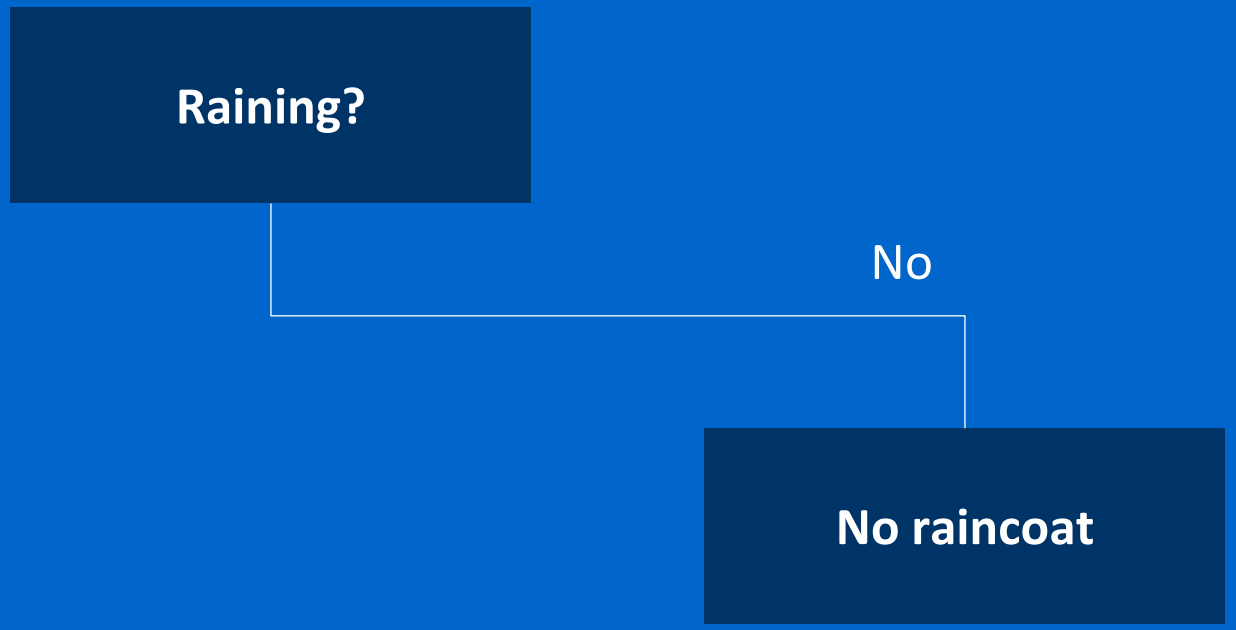


If it is raining and you are walking, wear raincoat



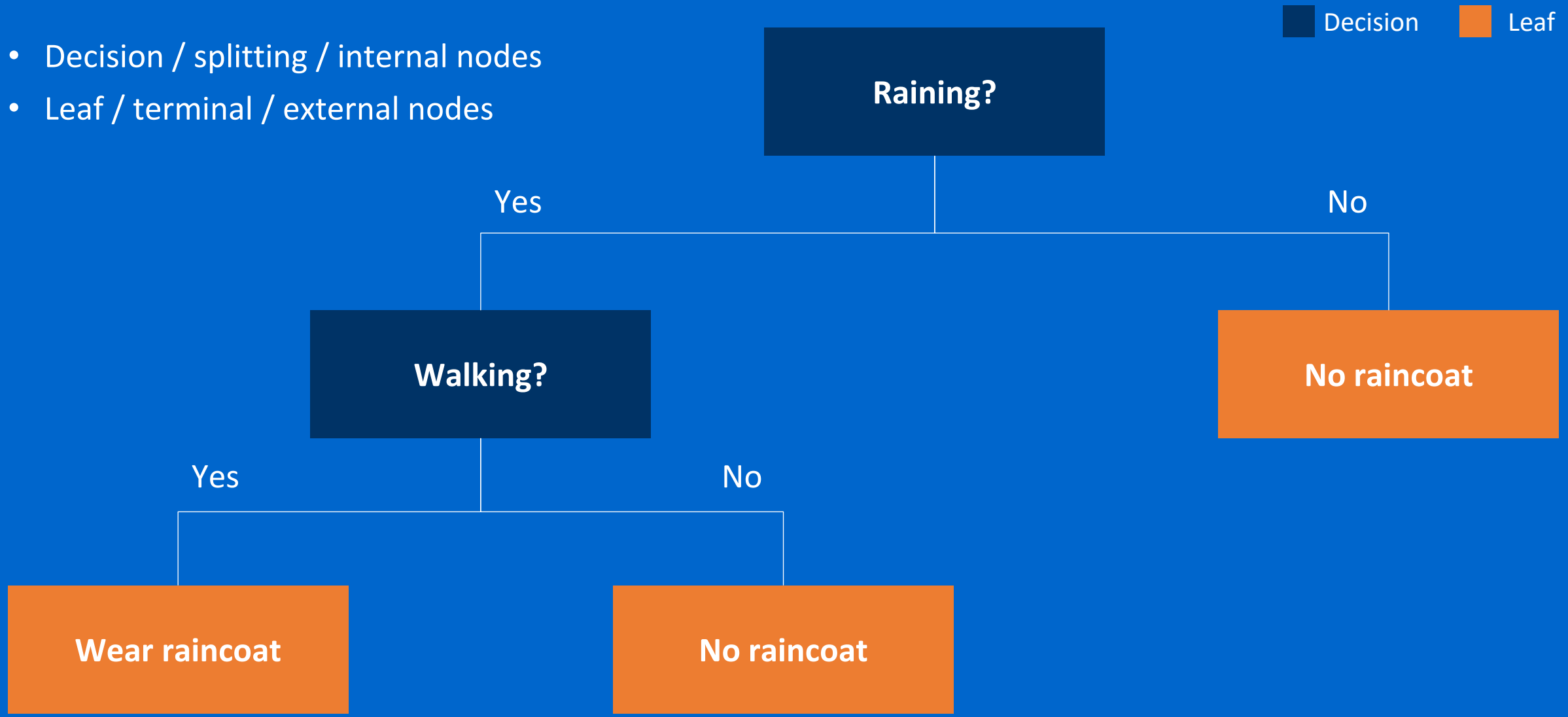
If it is raining and you are not walking, no raincoat

If it is not raining, no raincoat



Decision trees have two types of nodes: decision and leaf

- Decision / splitting / internal nodes
- Leaf / terminal / external nodes



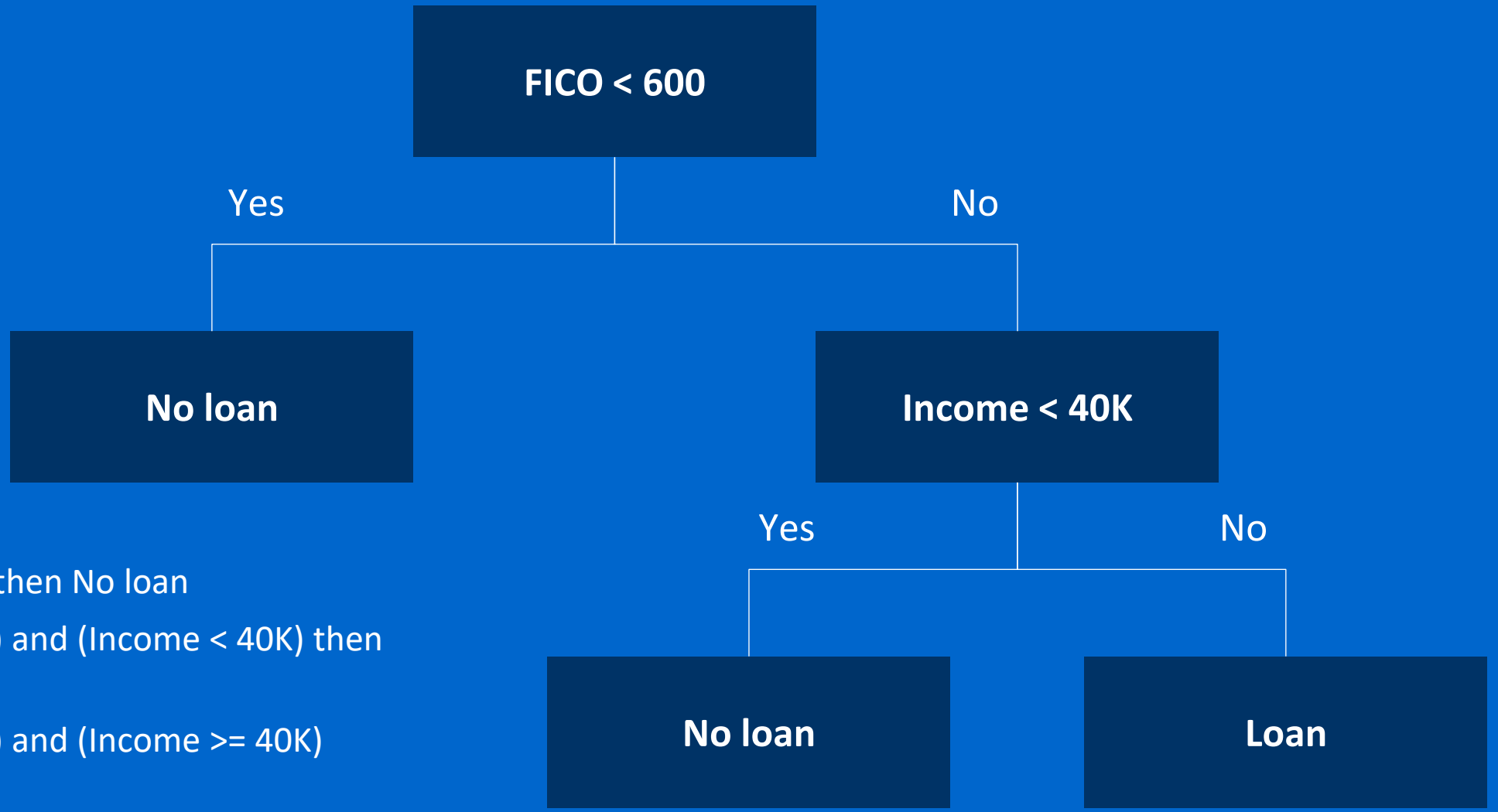
Two types of decision trees: Classifier and Regressor

- Classifier predicts a class
  Example: churn or no churn; dog, cat, mouse; loan or no loan
- Regressor predicts a number
  Examples: bike sales; temperature
- For a classifier, the leaf nodes contain the class.
- For a regressor, the leaf nodes contain a number.

Each leaf node corresponds to a decision rule

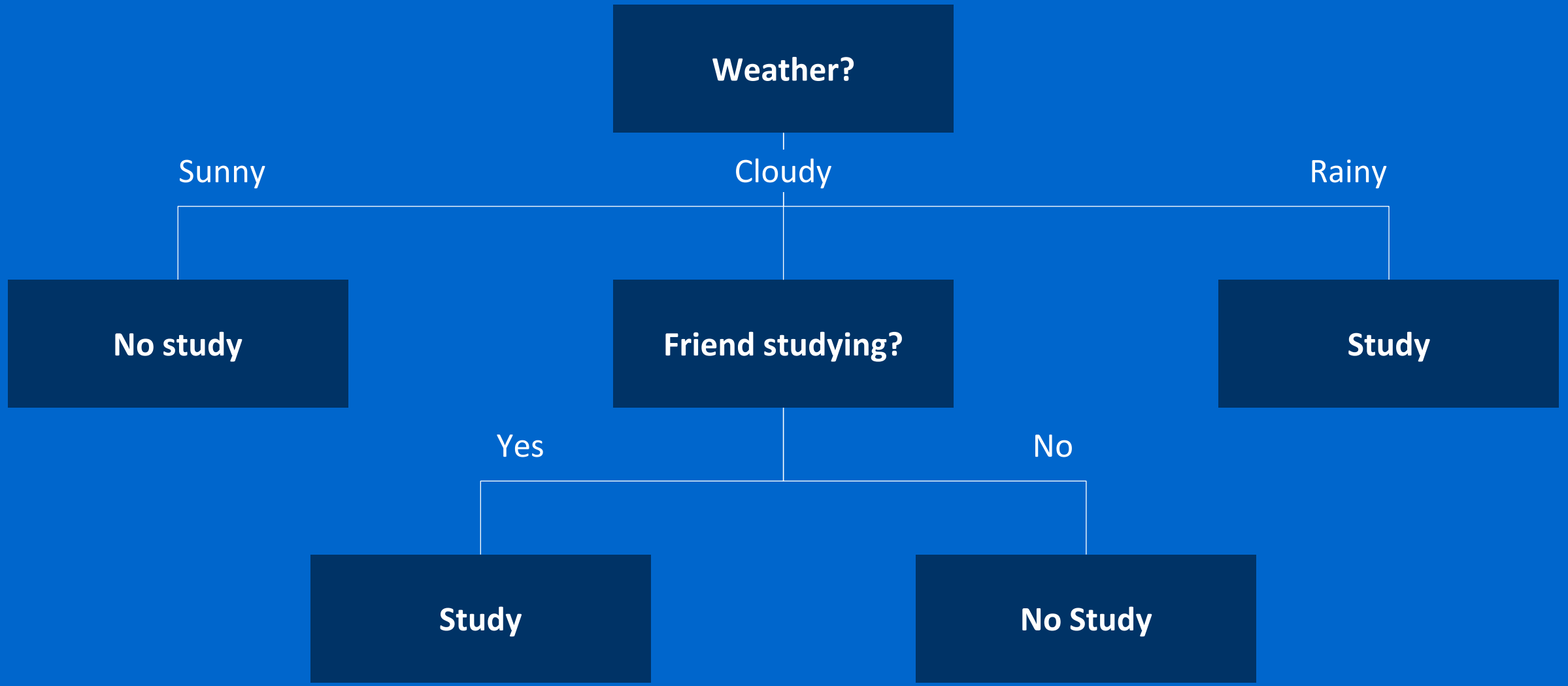
- Leaf node 1: If (Raining=Yes) and (Walking=Yes) then Wear raincoat
- Leaf node 2: If (Raining=Yes) and (Walking=No) then No raincoat
- Leaf node 3: If (Raining=No) then No raincoat

From decision rules to decision tree



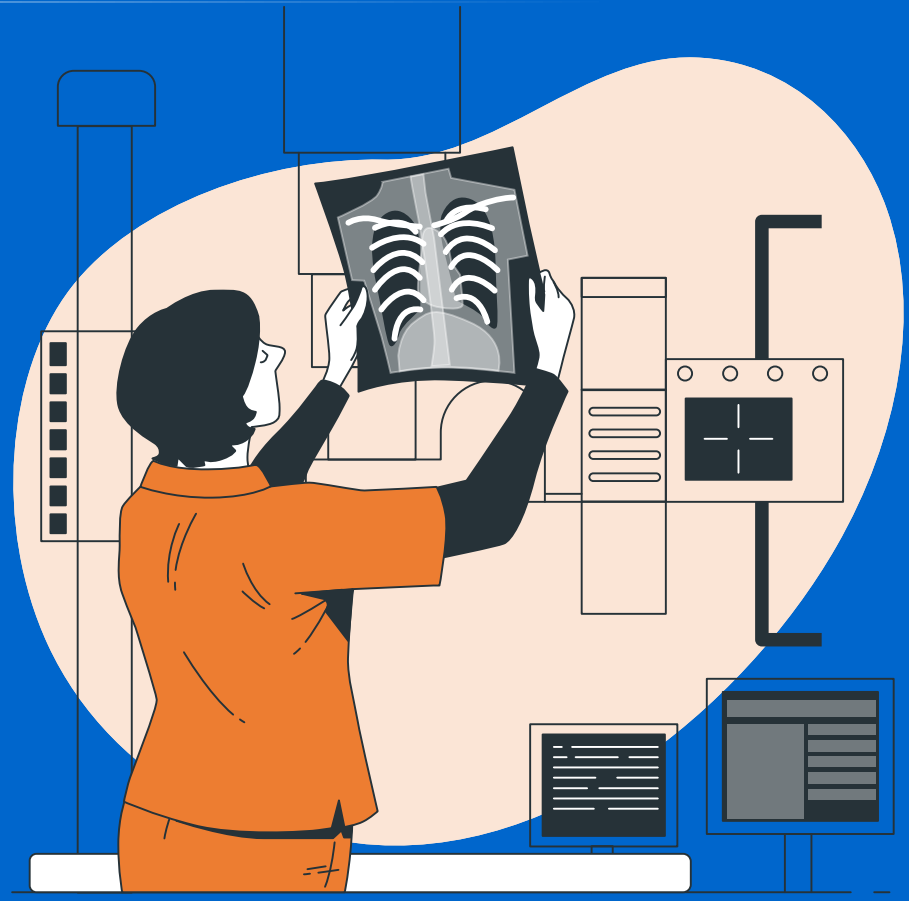
- If (FICO score < 600) then No loan
- If (FICO score >= 600) and (Income < 40K) then No loan
- If (FICO score >= 600) and (Income >= 40K) then Loan

5 minute exercise: Write the decision rules



5 minute exercise: Create the decision tree

- If (X-ray=white spots) then Pneumonia
- If (X-ray=clear) and (temp < 99) then Healthy
- If (X-ray=clear) and (temp > 99) then Fever



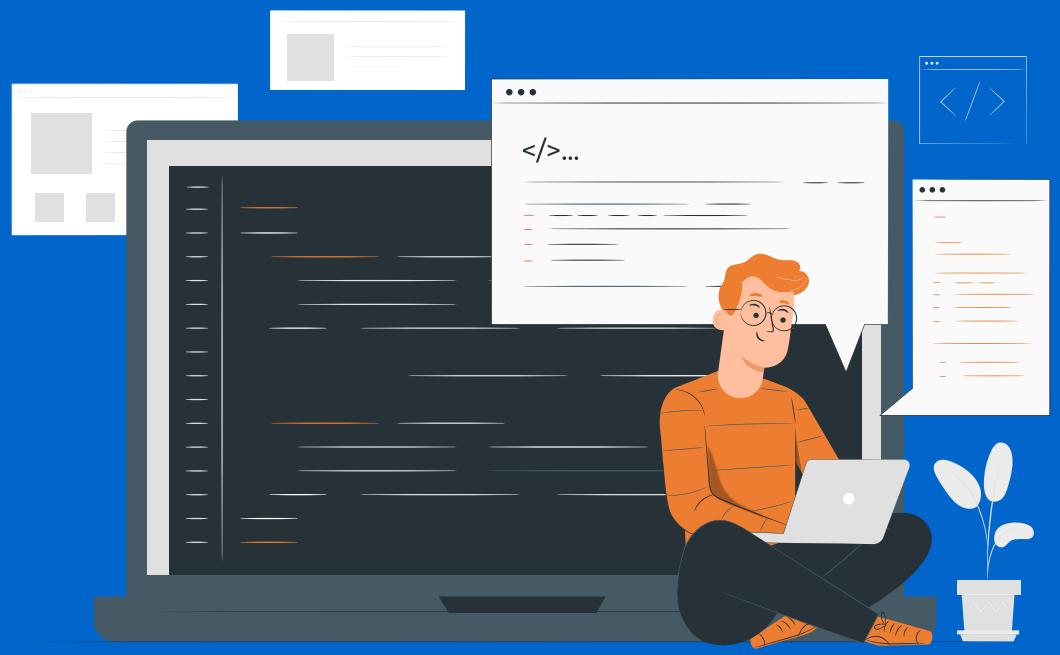
Decision tree algorithm

Intuition

- Find ‘best’ splits of data (by category or regression)
- Continue recursively splitting until a stopping criteria is met
  - When all instances are classified correctly
  - When error is below a threshold
  - When a maximum depth is reached
  - When the number of examples is below a threshold

Pseudocode for a decision tree

- Function DecisionTreeClassifier(n,A) // n: samples, A: attributes (all are binary)
- If all n are the same class
  - node_type = leaf
  - class = class of n
- else
  - node_type = decision
  - a_decision = bestAttribute(n,A)
  - LeftNode = DecisionTreeClassifier(n(a=1), A)
  - RightNode = DecisionTreeClassifier(n(a=0), A)



Decision tree Python code

Let’s look at some code!

- Three steps:
  - Data wrangling
  - Linear regression
  - Decision tree regressor
- We will use a housing data set.

Why try linear regression first?

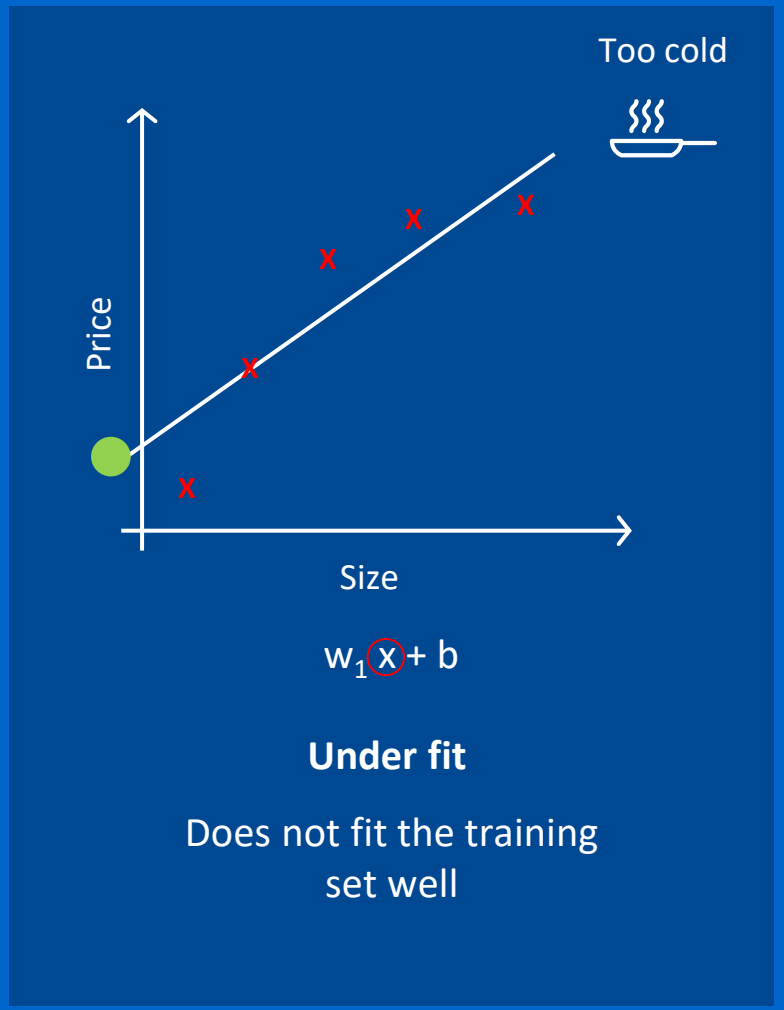
- Build intuition
- Establish a base level of performance
- Other ways to establish a base level of performance:
  - For classification algorithms, the frequency of the most common class
  - Human performance
  - Competing algorithms
  - Existing algorithm
  - Guess

Python libraries we will use frequently

- Pandas: Data handling
- Scikit-learn: Machine learning algorithms
- Matplotlib: Visualization

We will use many other libraries as well

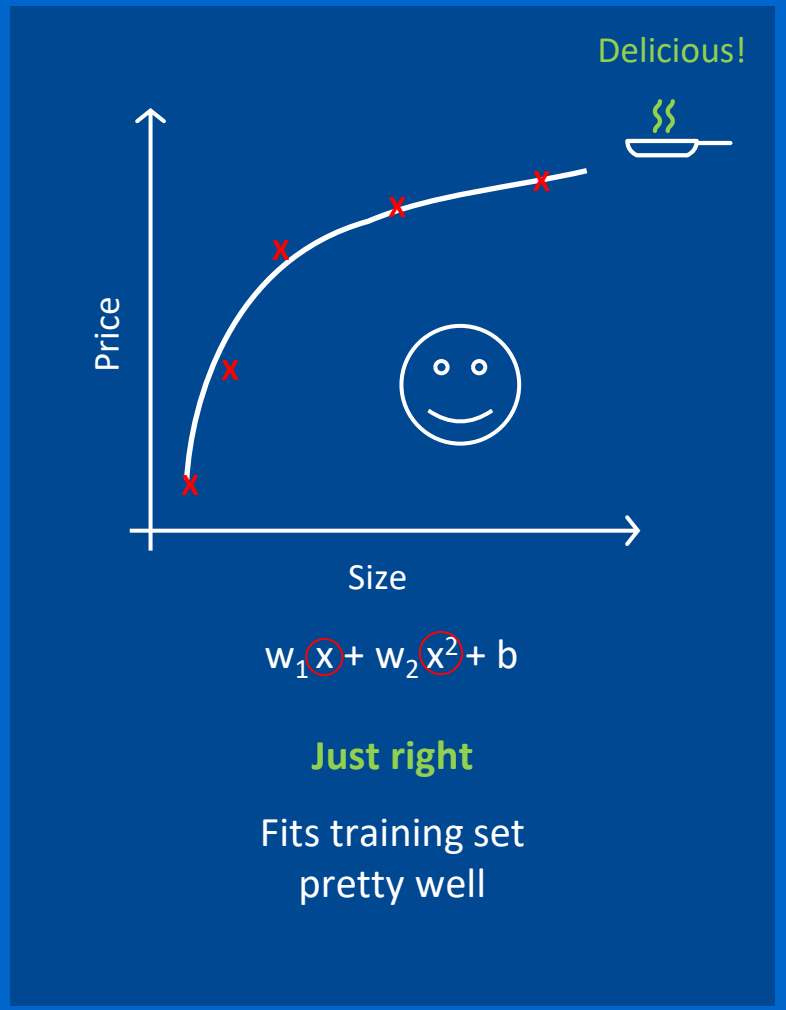
Overfitting is a key problem in machine learning
Regression example



$w_1 x + b$

Under fit

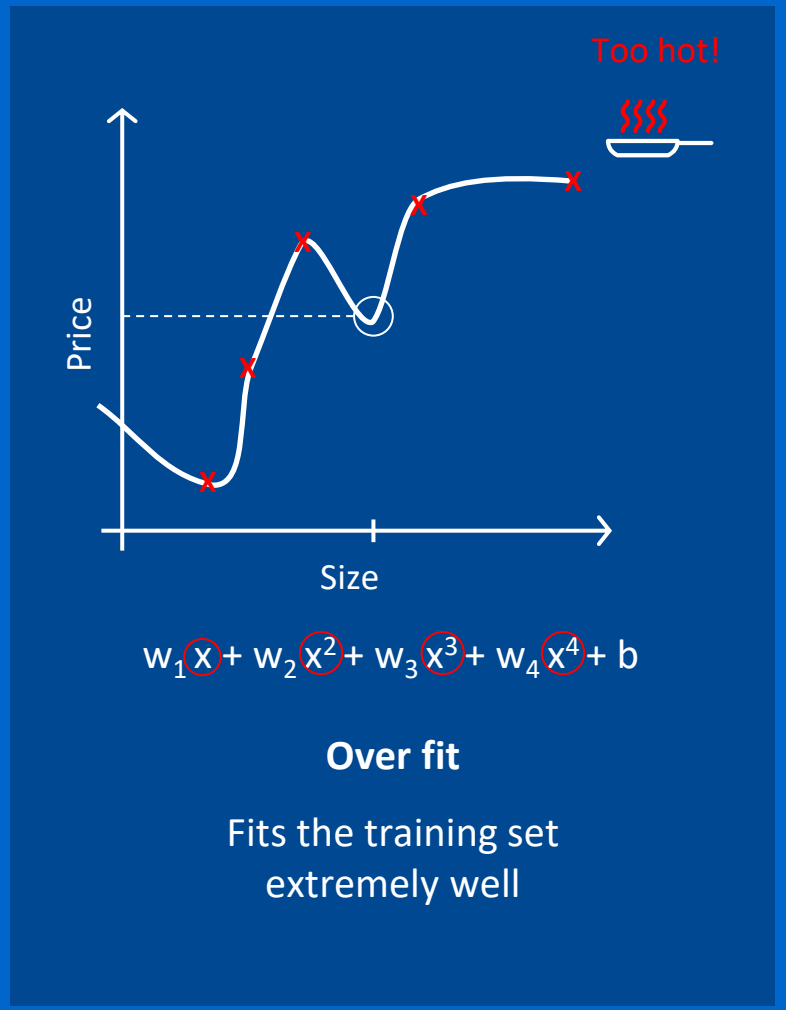
Does not fit the training set well



$w_1 x + w_2 x^2 + b$

Just right

Fits training set pretty well



$w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + b$

Over fit

Fits the training set extremely well

Adapted from Andrew Ng

Simple definition of overfitting

- Overfitting is taking place when:
  - Training error is decreasing while the testing error is increasing
  - Training accuracy is increasing while testing accuracy is decreasing

Advantages and disadvantages decision trees

- Advantages
  - Interpretable! Decision trees are easy to understand by humans. Neural networks are not.
  - Very fast to run
  - Memory efficient
  - Often the best algorithm for classification problems (XGBoost)
  - Do not need to normalize the variables (unlike with neural networks)
- Disadvantages
  - Sensitive to slight changes in the data (causes different splits)
  - Rarely best for regression (predicting numerical attributes)
  - Not recommended for unstructured data (images, text)

Course overview

Agenda

- Session 1: Decision trees
- Session 2: Advanced Decision Trees
- Session 3: Unsupervised learning: clustering
- Session 4: Metrics, Feature Engineering, and Ethics
- Session 5: Neural networks
- Session 6: Advanced neural networks
- Session 7: Review
- Session 8: Proctored, three hour individual assessment

Guest speakers

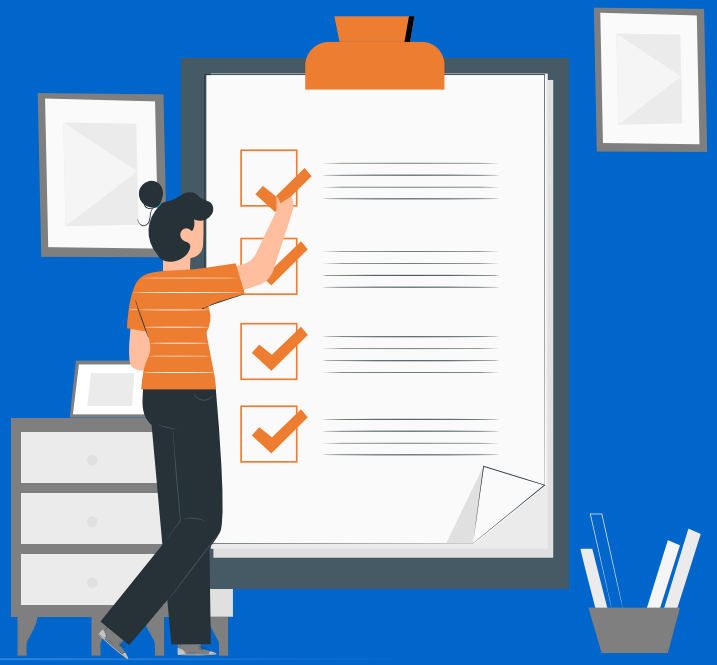
Grading

- Three knowledge checks: 10% each, 30% total
  - 10 question multiple choice; assume code works – no syntax errors or package issues
  - At the start of Sessions 2, 3, 4
  - Cumulative (e.g., Session 4 Knowledge Check covers Sessions 1-3)
  - Timing: 20 minutes in Session 2; 15 minutes in Session 3; 10 minutes in Session 4
- One team assessment: 30% total
  - Assigned: Session 5
  - Due: Session 7
- One individual assessment: 40% total
  - Three hour proctored knowledge check
  - Consists of multiple choice questions and short answers
  - Assume code works – no syntax errors or package issues

All knowledge checks and assessments are open book/Colab/AI.

AI Policy

- You can use AI!
  - Feel free to use on all knowledge checks, individual assessment, and group assessment.
  - No restrictions!
- Remember to cite your use of AI



Group work

- ***Always*** list all members of the group who did work in all files
  - In-class exercises
  - Python code
  - Group assessment
- For the group assessment, everyone in the group will receive the same grade

What does a grade mean?

- A: Excellent
- B: Very Good
- C: Acceptable

Suggestion: How to get an A

- Study 2-3 hours per day
  - Review all of the slides
  - Look at optional material
  - Run algorithms on additional datasets
- Attend office hours
- Ask questions
- Summarize what you have learned in your own words. Create a glossary.
- Follow Hult policy (late assignments, citations, academic honesty, etc.)
- *Bring your name card to class and complete attendance in first 5 minutes*

Suggestion: Organize algorithms and datasets in folders

- Algorithms folder: Python code
  - Linear Regression
  - Decision Tree Regressor
- Datasets folder
  - Housing sales
  - Bike rentals
- Content (organize by algorithm and datasets)
  - Algorithms
    - Decision trees
    - Neural networks
  - Datasets
    - Iris
    - Bike rentals
- Glossary

When you learn a new algorithm, run it against additional datasets