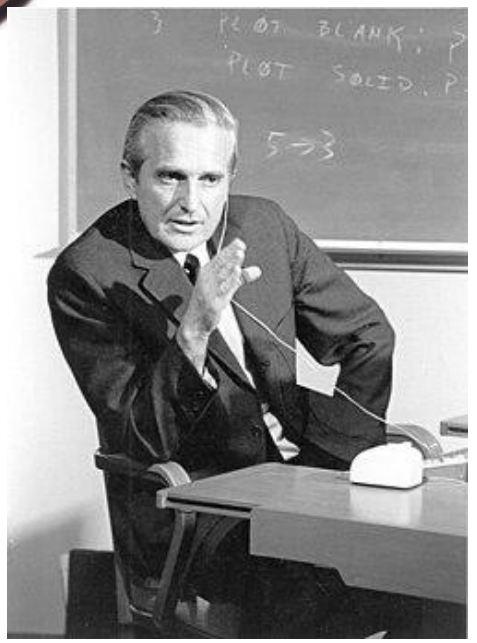
Bonus Slide

The Mother of All Demos





In 1968, Douglas Engelbart presented research to a group of engineers at a computer conference. This 1.5 hour presentation later became known as the Mother of All Demos because it introduced an astounding number of technologies that we use to this day.

The technologies Engelbart introduced in this live demo include:

- The computer mouse
- The GUI (Graphical User Interface)
- The WYSIWIG (What You See Is What You Get) text editor
- The concept of multiple windows
- Revision control
- Video conferencing
- Real-time collaborative editing





Watch the demo here: https://www.youtube.com/watch?v=yJDv-zdhzMY

Unit 2: Modern CS Techniques

Machine Learning Fundamentals

15-110 – Monday 10/26

Announcements

- Project1 was due today
- Project2 released
- Quizlet6 on Wednesday
- Oral2 sign-ups now open
  - Link: https://www.cs.cmu.edu/~110/oral/oral2.html
  - Discuss either Project1 or Project2
  - Bring your laptop and be ready to discuss your code

Learning Goals

- Describe how **training** is used to build a machine learning model
- Identify the three major categories of learning (**supervised**, **unsupervised**, and **reinforcement**) and the three major categories of reasoning (**classification**, **regression**, and **clustering**)
- Decide which combination of **learning** and **reasoning** categories are best used to solve a stated problem

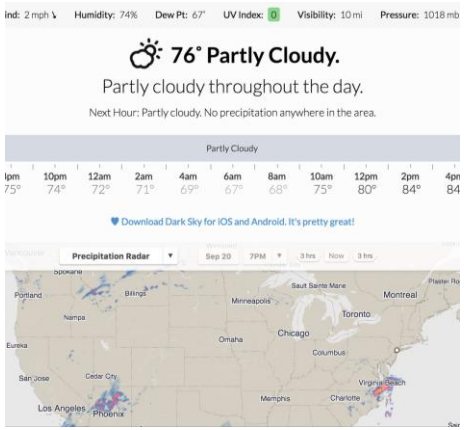
Machine Learning and Models

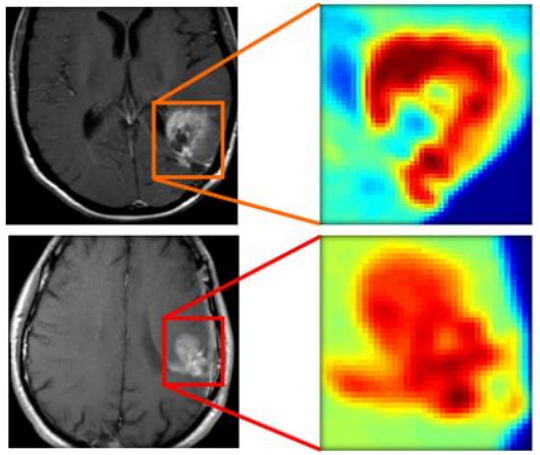
ML: What it is and uses

Machine Learning is a branch of computing in which systems learn patterns from data and improve at a task through experience, rather than being explicitly programmed with fixed rules for every case.

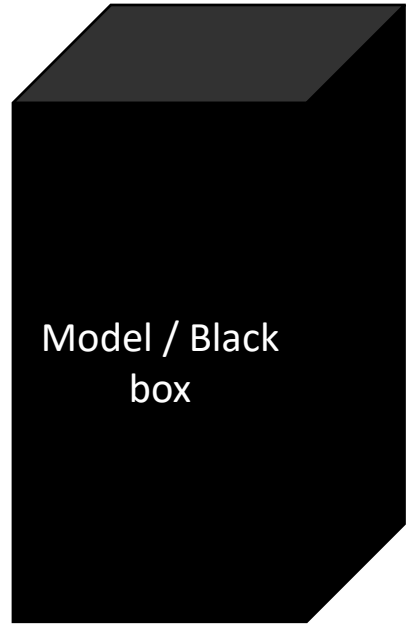
ML is used in thousands of contexts across the world, including speech recognition, weather prediction, and medical diagnosis.







Reminder from GenAI: Models



Training process:

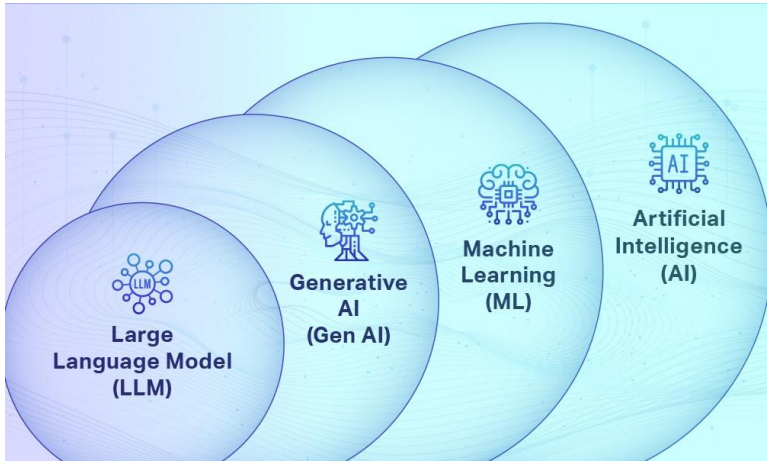
- Models are trained on massive datasets
- The model learns patterns across billions of examples

Generation process:

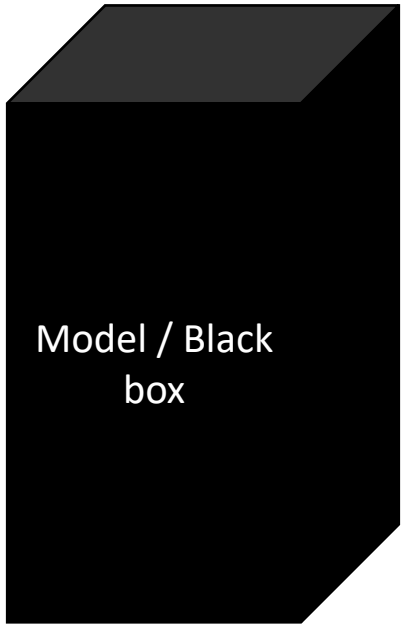
- Given a prompt, the model predicts the most likely next piece of content (word, pixel, code) based on what it learned

The same idea applies to ML,

especially because GenAI is a subset of ML!!



ML Models: What are they?



A trained function:
An ML model is a mathematical function that has been fit to data.

ML Models: How are they created?

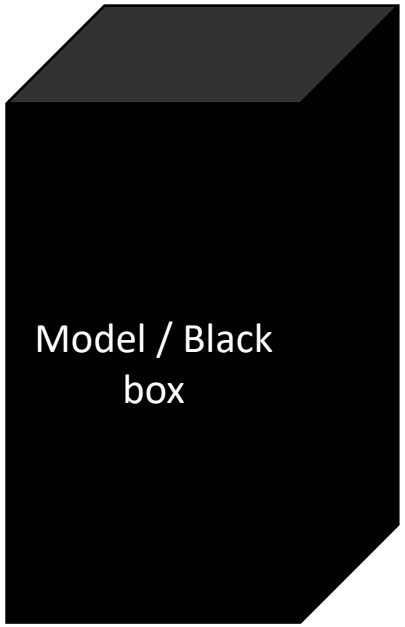
Models are learned, not hand-coded.

Instead of a programmer designing a model directly, they apply a ML algorithm to data.

The programmer doesn't teach the algorithm patterns, they teach it *how to find* patterns.

The algorithm identifies patterns and generates the model automatically.

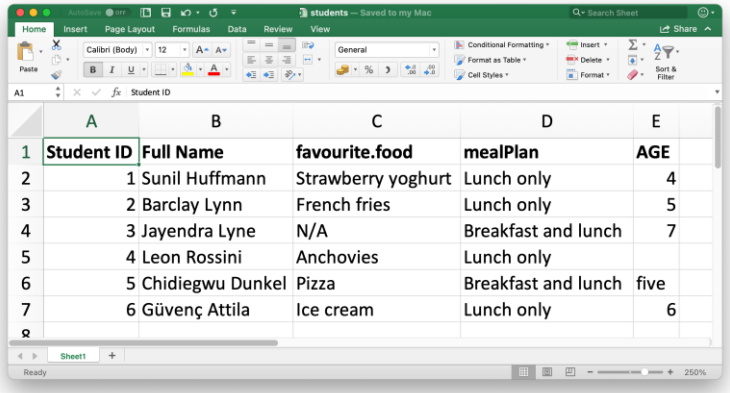
This happens during a process called **training**.



A trained function:
An ML model is a mathematical function that has been fit to data.

Models are trained on lots of data

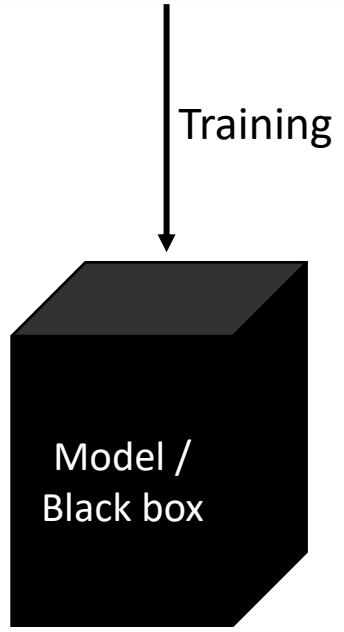
To train a model using machine learning, we must start with a large amount of **data**. The algorithm will try to find patterns and correlations in that data. A model is only as good as its dataset!



	A	B	C	D	E
1	**Student ID**	**Full Name**	**favourite.food**	**mealPlan**	**AGE**
2	1	Sunil Huffmann	Strawberry yoghurt	Lunch only	4
3	2	Barclay Lynn	French fries	Lunch only	5
4	3	Jayendra Lyne	N/A	Breakfast and lunch	7
5	4	Leon Rossini	Anchovies	Lunch only	
6	5	Chidiegwu Dunkel	Pizza	Breakfast and lunch	five
7	6	Güvenç Attila	Ice cream	Lunch only	6

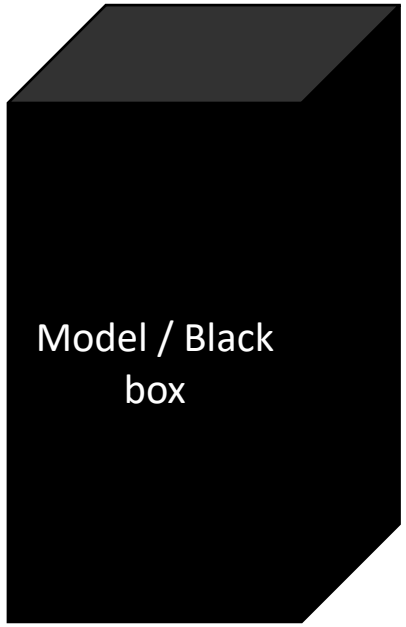
Each data entry must be broken down into a set of **features** so the machine learning algorithm can analyze features separately or in groups. If you have a table of data, the features would be the **columns**; if you have a set of text, the features would be **words** or **groups of words**.

You can create and add new features the same way you would add new columns in data analysis.



A trained function

ML Models: How are they used?



A trained function:
An ML model is a mathematical function that has been fit to data.

Once a model has been generated, we can apply it to **new data**.

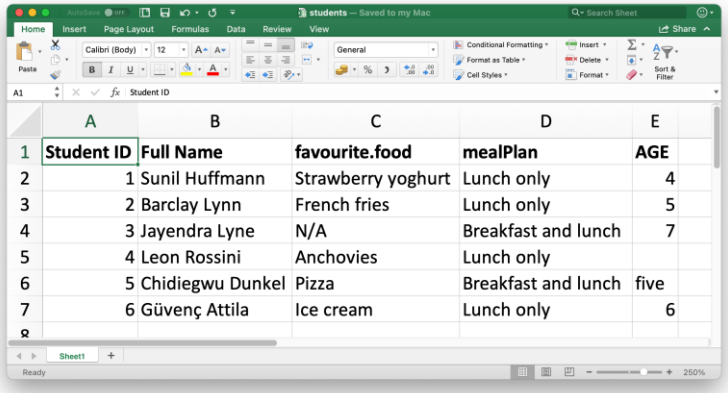
A model trained on market behavior could make a prediction about which products produced by a business might be most popular in an upcoming quarter.

We can also analyze models to gain insights about the original dataset. For example, we might build a model to group open-ended feedback provided in a country-wide survey, then use those groups to identify common trends among answers.

Summary: Training vs Using a model

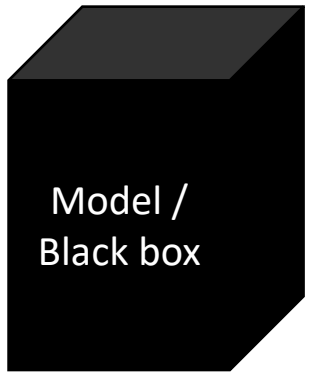
Training an ML model

Data (broken down into features)



	A	B	C	D	E
1	**Student ID**	**Full Name**	**favourite.food**	**mealPlan**	**AGE**
2	1	Sunil Huffmann	Strawberry yoghurt	Lunch only	4
3	2	Barclay Lynn	French fries	Lunch only	5
4	3	Jayendra Lyne	N/A	Breakfast and lunch	7
5	4	Leon Rossini	Anchovies	Lunch only	
6	5	Chidiegwu Dunkel	Pizza	Breakfast and lunch	five
7	6	Güvenç Attila	Ice cream	Lunch only	6

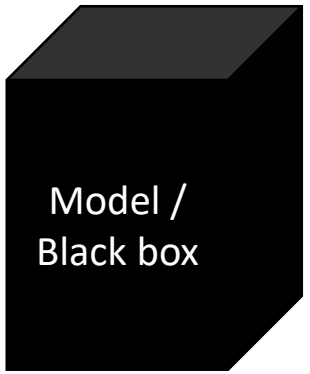
Training:
learns patterns
from the data



A trained function

Using an ML model

Input (new data)



A trained function

Output
(prediction, score, cluster, etc.)

Training a Model

What exactly happens during training?



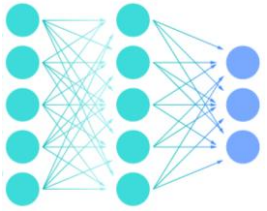
Training set

Features (X1, X2, X3, …) | Label (Y)

Has fur	Has pointy ears	Size (kg)	Barks	Y (label)
Yes	Yes	12	Yes	Dog
Yes	No	40	No	Not a dog (e.g. a cow)
No	No	0.5	No	Not a dog (e.g. a fish)
Yes	Yes	8	Yes	Dog

Fed into →

Training process



The learning algorithm searches for optimal values of parameters a, b, c, etc., in a function:

$f(x) = a \cdot X_1 + b \cdot X_2 + c \cdot X_3 + \ldots + d$

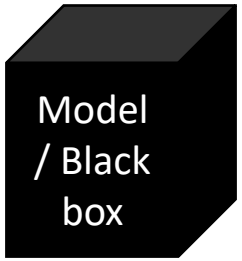
so that its output matches the values, *Y*.

Iteratively tries different values for a, b, c, etc., measures how far predictions are from actual results and adjusts parameters to reduce that error.

During training, it also identifies and favors features that reduce that error.

Produces →

Fitted Model



A trained, fitted function:

f(x) = aX1 + bX2 + cX3 + …

ML model == math function that has been fit to data

Example: Is It a Dog?

Let's use a very simplified example to demonstrate how a machine learning algorithm might learn which features best match a label.

We want to train a model to recognize pictures of dogs. What kinds of **features** might it identify?

(The features we describe here are far too high-level, but the general idea is about right).

Example: Is It a Dog?

Guess: 

Answer: 

Prior favored features:

???



Updated favored features:

- Has Pointy Ears
- Has Fur
- Screen in background

Example: Is It a Dog?

Guess:  Answer: 

Prior favored features :

- Has Pointy Ears
- Has Fur
- Screen in background

Updated favored features :

- Has Pointy Ears
- Has Fur
- Grey/Black Coloring



Example: Is It a Dog?

Guess:  Answer: 

Prior favored features :

- Has Pointy Ears
- Has Fur
- Grey/Black Coloring

Updated favored features :
(no updates needed,
we got it correct!)



Example: Is It a Dog?

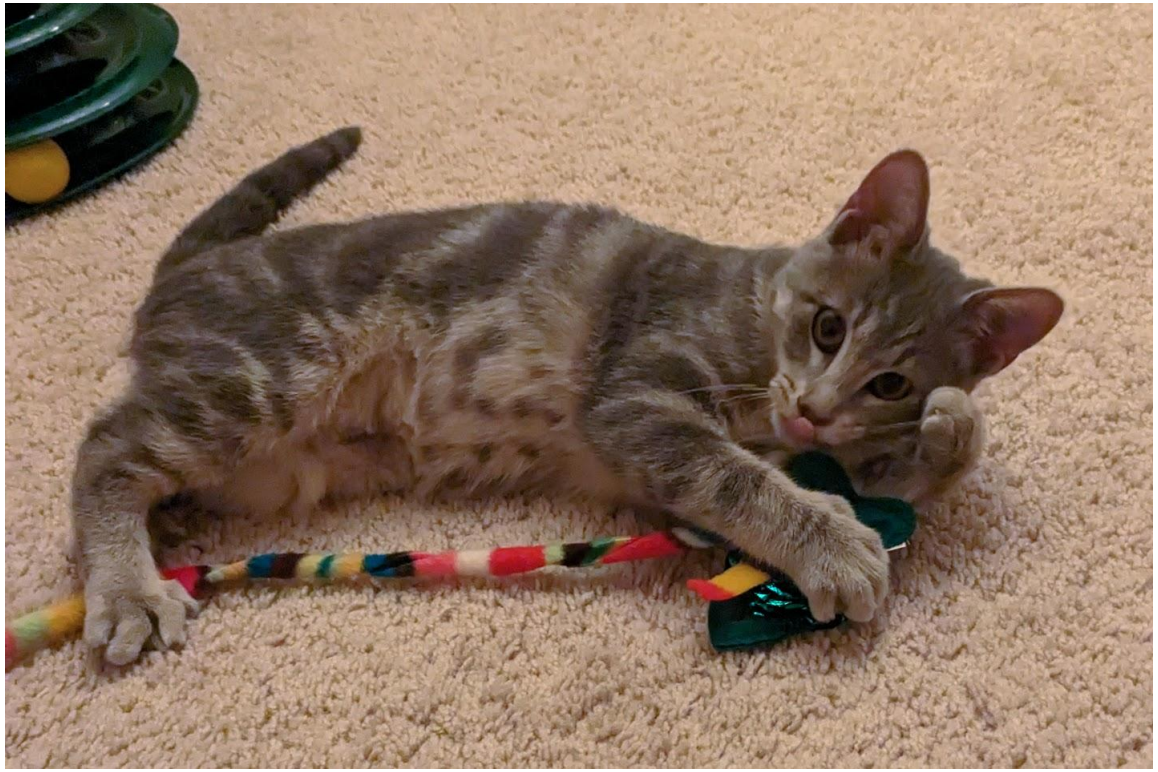
Guess:  Answer: 

Prior favored features :

- Has Pointy Ears
- Has Fur
- Grey/Black Coloring

Updated favored features :

- Has Fur (still useful, but not sufficient)
- Has Pointy Ears (too common across cats and dogs, favored less now)
- **Snout shape** (new feature favored, dogs tend to have longer snouts)



Types of Learning

Many Machine Learning Algorithms

There are a very large number of machine learning algorithms that computer scientists have designed which all generate models using different approaches. How do we choose which one to use?

We can break algorithms into groups based on how they **learn** and on how they perform **reasoning**.



Many Ways to Learn

There are several ways that machine learning algorithms learn patterns from data. Which approach you use depends on what type of problem you're trying to solve.

The three main approaches are **supervised learning**, **unsupervised learning**, and **reinforcement learning**.

Supervised Learning



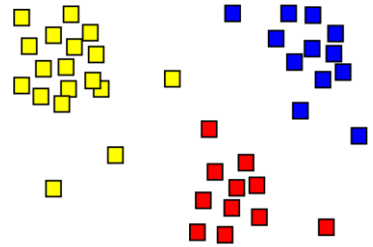
Supervised learning algorithms train models to predict outputs based on inputs. A data entry in the dataset is the input, and a **label** or **score** is the output.

Example: to train a model to predict a person's favorite ice cream based on other desert preferences, each data entry (set of desert preferences) must be accompanied by a label (that person's favorite ice cream).

Once the model is fully trained, it can be used to make predictions about unlabeled data.

This is like when you learn by completing a homework assignment. You're trying to predict a correct answer (label) which is known by the teacher. This means the teacher (ML algorithm) can check your work. Later, you'll apply your knowledge to new problems where the answer is unknown.

Unsupervised Learning



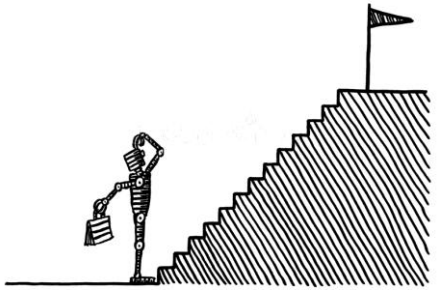
Unsupervised learning algorithms train models that infer the natural structure of a dataset. Which data entries are most similar to each other?

Example: identify common categories of favorite ice cream flavors by grouping together people who have similar desert preferences.

Unsupervised learning is often used when the data is **not** labeled – no true answer is known. That means that a human being will need to look over the model's results to see if they make sense or seem random.

This is like how researchers come up with theories to explain how the world works. Theories are based on data and may make sense, but might be changed of disproven later on.

Reinforcement Learning



Reinforcement learning algorithms train models that help an **artificial intelligence agent** (AI agent) approach an overall goal. Usually this involves taking a sequence of actions to bring the agent closer to the endpoint and giving feedback on those actions.

Example: train a robot to make ice cream by constantly changing the temperature and speed of an ice cream machine and measuring the consistency of the product until it's just right.

This is like playing a guessing game where the person with the secret word tells you if your guesses are hot or cold. You adjust your guesses and questions based on the person's feedback until you've gotten to the correct word.

Other Types of Learning

Semi-supervised learning: mixed dataset where some entries are labeled but most aren't.

Active learning: the algorithm asks an expert to label specific entries to help it learn.

Both are used when a model needs to be trained on a lot of unlabeled data, like with Large Language Models (LLMs).

Demo: Supervised Learning

We can build an image-recognition model using supervised learning by labeling a bunch of pictures as belonging to specific classes.

Try it out here: https://teachablemachine.withgoogle.com/train/image

Types of Reasoning

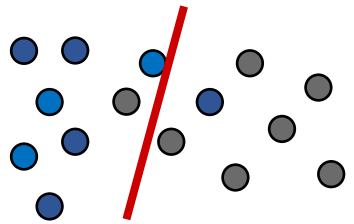
Many Ways to Reason

Once you've identified what type of learning you need to do, there are still lots of options for what type of algorithm you can use. One way to narrow the options down further is to ask what type of **reasoning** the model that is produced will need to do.

For supervised learning, common reasoning approaches include **classification** and **regression** algorithms.

For unsupervised learning, common reasoning approaches include **clustering**.

Classification Algorithms

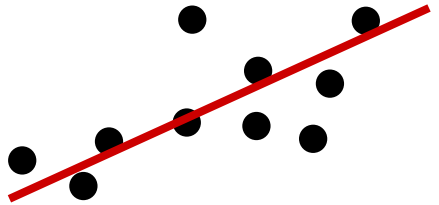


A **classification** algorithm takes labeled data of any type and produces a **model** (called a classifier) which determines which **category** a given data entry most likely belongs to. In other words, it produces a **categorical** (discrete) result.

For example, a classifier might take in data about an email (the sender, title, message content) and classify it as SPAM or NOT SPAM

Classification is commonly used in image recognition. Examples of classification algorithms include Naive Bayes, Decision Trees, Support Vector Machines (SVMs), and **logistic regression**, which we'll use in a future lecture.

Regression Algorithms

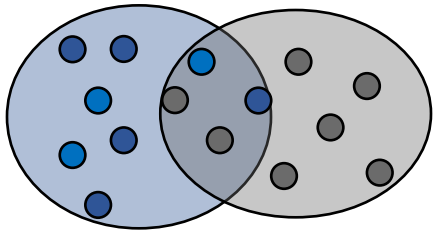


A **regression** algorithm takes numerically-labeled data and produces a function that maps data entries to **numerical** (continuous) results. Instead of predicting a specific category, it gives each data entry a **score**.

For example, a regression algorithm might take information about a house (GPS location, previous sell date, previous sell price, average price of nearby houses) and use it to predict its market value (a price).

Regression algorithms are used to predict things like stock market prices and weather temperatures. A linear regression is one kind of regression algorithm that predicts continuous values.

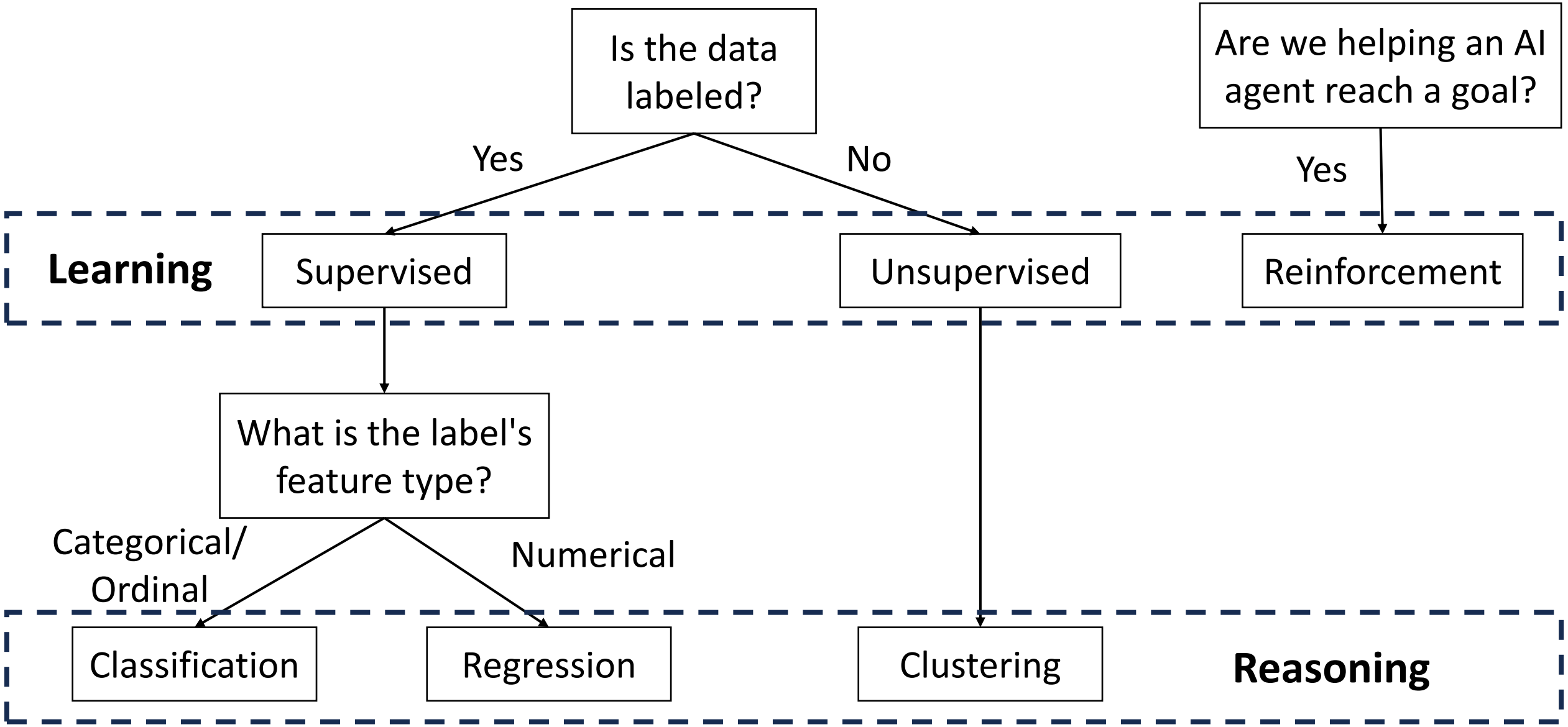
Clustering Algorithms



A **clustering** algorithm takes a set of data and produces a model that breaks the data into some number of **clusters** by identifying data entries that are similar to each other. We don't actually know the true categories or scores of the provided data, so it's hard to 'test' the resulting model, as we aren't sure what the actual result should look like.

For example, a clustering algorithm could recommend a new set of plant species by finding similarities between plant features in a given region.

Clustering is used in consumer analysis, data compression, and bioinformatics. K-means clustering and affinity propagation are examples of clustering algorithms.


Is the data labeled?
Are we helping an AI agent reach a goal?
Yes
No
Yes
Learning
Supervised
Unsupervised
Reinforcement
What is the label's feature type?
Categorical/ Ordinal
Numerical
Classification
Regression
Clustering
Reasoning

Activity: Choose How to Reason

You do: you want to use machine learning to solve a problem. What combination of learning category + reasoning category would you use?

- You have a dataset of the restaurants that people have visited in the past three months. How would you create new customer-profile groupings of people based on this data?

- You have a dataset of survey results, including participant's favorite TV shows, sports, and hobbies. How would you predict someone's favorite sport based on the rest of the survey data?

- You have a dataset of exercise logs (time spent in a variety of exercise activities). How would you estimate the amount of time a person spends running on a treadmill based on other activities?

Many Other Algorithms

Classification, regression, and clustering are just the most common types of reasoning algorithms. There are plenty of others that we don't have time to discuss in this class.

To make a model, though, you just need one algorithm: the best fit for the problem you're trying to solve. We'll learn how to create that algorithm next time.

Learning Goals

- Describe how **training** is used to build a machine learning model
- Identify the three major categories of learning (**supervised**, **unsupervised**, and **reinforcement**) and the three major categories of reasoning (**classification**, **regression**, and **clustering**)
- Decide which combination of **learning** and **reasoning** categories are best used to solve a stated problem