# Types of Machine Learning

APT 3025

SPRING 2022

# Overview

- Three different types of machine learning:
  - Supervised
  - Unsupervised
  - Reinforcement learning
- The difference between labeled and unlabeled data
- The difference between regression and classification, and how they are used

# Applications of machine learning

- Predicting house prices based on the house's size, number of rooms, and location
- Predicting today's stock market prices based on yesterday's prices and other factors of the market
- Detecting spam and non-spam emails based on the words in the e-mail and the sender
- Recognizing images based on the pixels in the image
- Processing long text documents and outputting a summary

# Applications of machine learning

- Recommending videos or movies to a user (e.g., on YouTube or Netflix)
- Building chatbots that interact with humans and answer questions
- Training self-driving cars to navigate a city by themselves
- Diagnosing patients as sick or healthy
- Segmenting the market into similar groups based on location, acquisitive power, and interests
- Playing games like chess or Go

# Data and features

- Data is simply information.
- Often information is organized as a table, where each row is a data point.
- Suppose we have a dataset of pets. In this case, each row represents a different pet.
- Each pet in the table is described by certain features of that pet.
- The features are the columns in the table.

# Labels

- The label is the target feature, that is the feature we want to predict in a given problem.
- The goal of a predictive machine learning model is to guess (that is, predict) the labels for new data.

| Age | Number of cars owned | Owns house | Number of children | Marital status | Owns a dog | Bought a boat |
|---|---|---|---|---|---|---|
| 66 | 1 | yes | 2 | widowed | no | yes |
| 52 | 2 | yes | 3 | married | no | yes |
| 22 | 0 | no | 0 | married | yes | no |
| 25 | 1 | no | 1 | single | no | no |
| 44 | 0 | no | 2 | divorced | yes | no |

Any one of the features can be the label (target) depending on the problem.

# Labeled and unlabeled data

- Labeled data is data that comes with labels. Unlabeled data is data that comes with no labels.
- An example of labeled data is a dataset of emails that comes with a column that records whether the emails are spam or ham, or a column that records whether the email is work related.
- An example of unlabeled data is a dataset of emails that has no particular column we are interested in predicting.

# Supervised and unsupervised learning

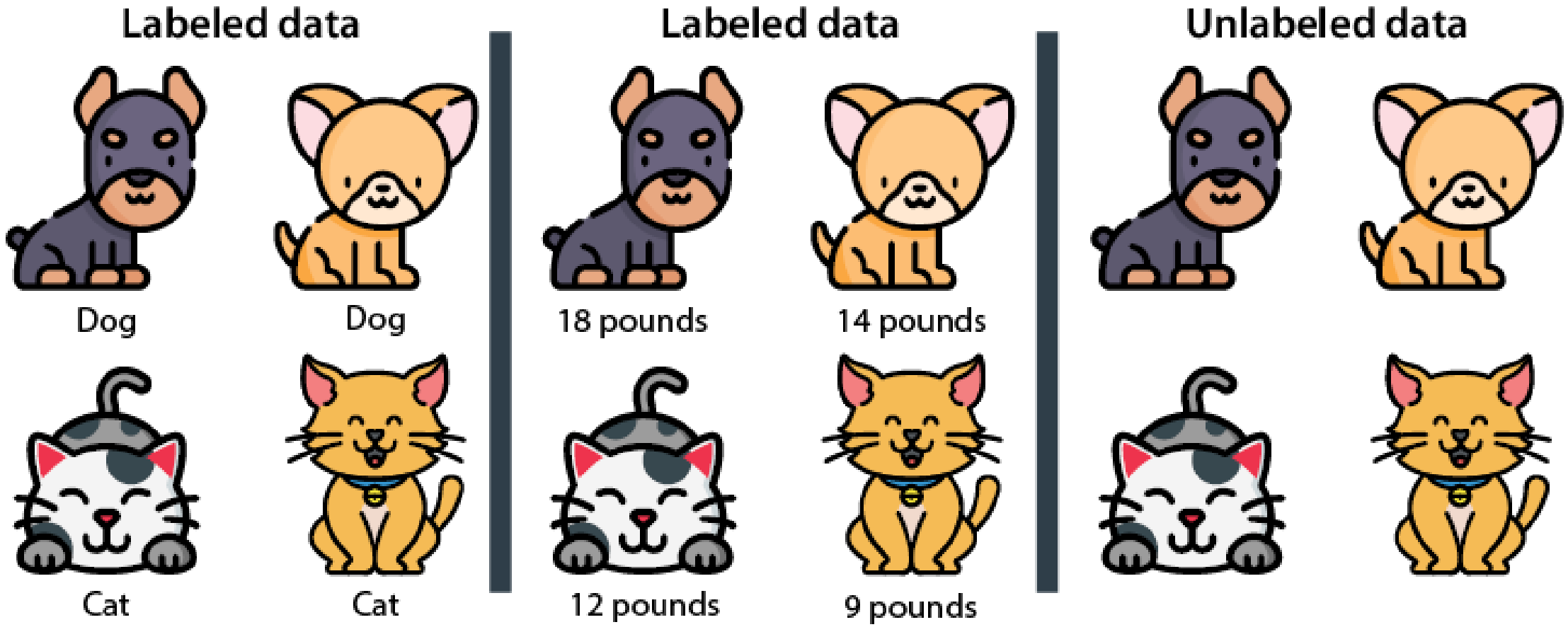
- Determining if data is labeled or unlabeled depends on the problem we are trying to solve.
- Labeled and unlabeled data yield two different branches of machine learning called supervised and unsupervised learning.
- Supervised learning works with labeled data. Unsupervised learning works with unlabeled data.

# Supervised and unsupervised learning

- Unsupervised learning may ask: Are there any clusters in the data? Find clusters provides useful insights but is probably not the end goal.
- Supervised learning may ask: What kind of person buys a boat?
- Supervised learning is what we will concern ourselves with for the rest of this course.

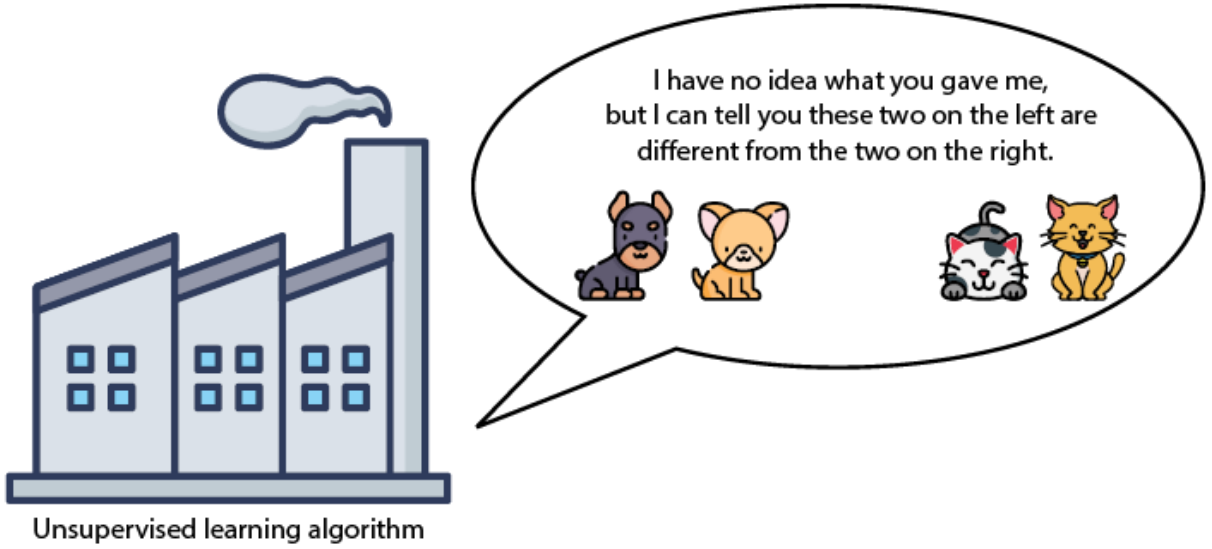
| Age | Number of cars owned | Owns house | Number of children | Marital status | Owns a dog | Bought a boat |
|---|---|---|---|---|---|---|
| 66 | 1 | yes | 2 | widowed | no | yes |
| 52 | 2 | yes | 3 | married | no | yes |
| 22 | 0 | no | 0 | married | yes | no |
| 25 | 1 | no | 1 | single | no | no |
| 44 | 0 | no | 2 | divorced | yes | no |

# Labeled and unlabeled data



# Unsupervised learning with unlabeled data

- An unsupervised learning algorithm can group things together based on similarity, even without knowing what each group represents.
- This can be useful as a preprocessing step for supervised learning.

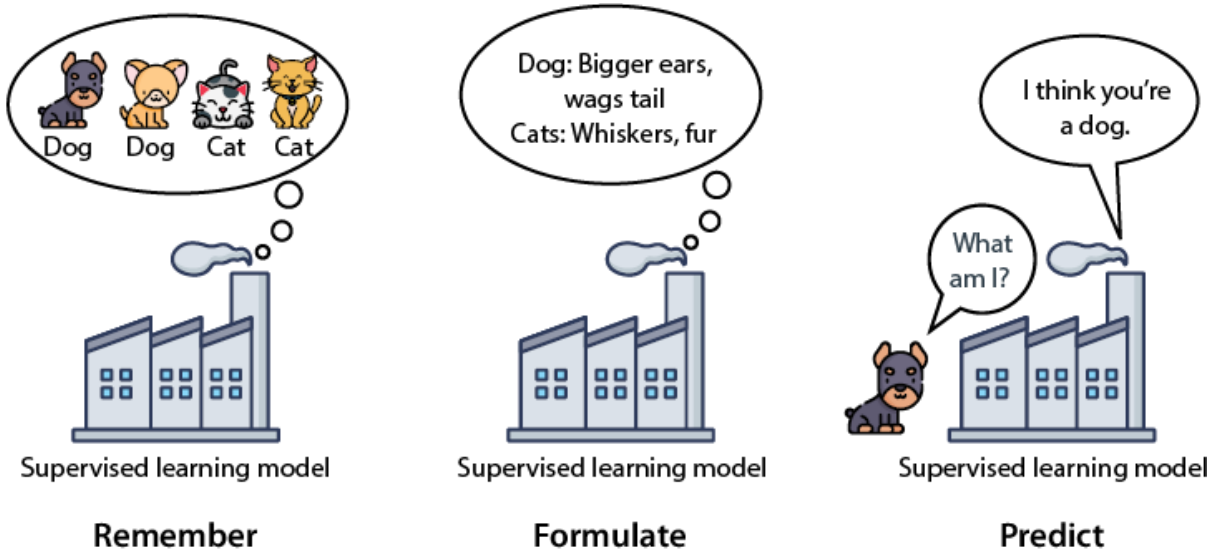


# Supervised learning with labeled data

- The dataset on the left contains images of dogs and cats, and the labels are “dog” and “cat.”
- For this dataset, the machine learning model would use previous data to predict the label of new data points.
- This means, if we bring in a new image without a label, the model will guess whether the image is of a dog or a cat, thus predicting the label of the data point.

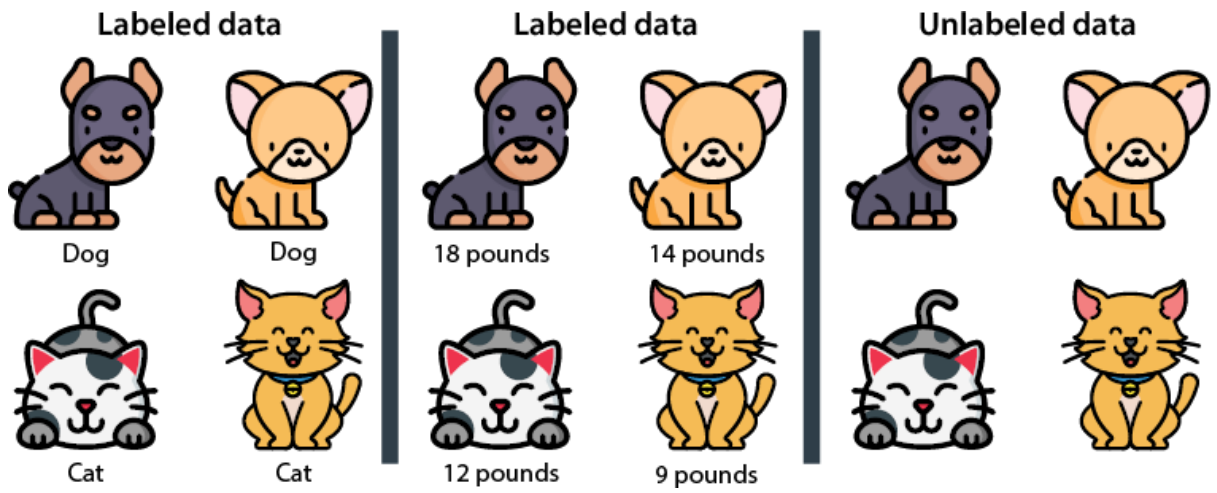
# Remember, formulate, predict

- Supervised learning follows the remember-formulate-predict pattern.
- The model first remembers the dataset of dogs and cats. Then it formulates a model, or a rule, for what it believes constitutes a dog and a cat.
- Finally, when a new image comes in, the model makes a prediction about what it thinks the label of the image is, namely, a dog or a cat.



# Numeric vs. categorical labels

- Notice in the database on the left, the label is a category. In the dataset in the middle, the label is a number.
- Numbers are infinite while categories are finite, and usually only a few.



# Regression and classification

- The above observation gives rise to two types of supervised learning models.
- Regression models are the types of models that predict numerical data, such as the weight of the animal.
- Classification models are the types of models that predict categorical data. The output of a classification model is a category, or a state, such as the type of animal (cat or dog).
- If Amazon predicts whether a potential customer will buy a particular book, is that regression or classification?
- What about if Amazon predicts how many books a customer is likely to buy?

# Regression Examples

- Stock market: predicting the price of a certain stock based on other stock prices and other market signals
- Medicine: predicting the expected life span of a patient or the expected recovery time, based on symptoms and the medical history of the patient
- Sales: predicting the expected amount of money a customer will spend, based on the client's demographics and past purchase behavior
- Video recommendations: predicting the expected amount of time a user will watch a video, based on the user's demographics and other videos they have watched

# Classification Examples

- The most common classification models predict a “yes” or a “no” (binary classification), but many other models use a larger set of states.
- Classifying handwritten digits based on the pixels in the images. Here there are ten classes, 0-9
- Social media: predicting whether a user will befriend or interact with another user, based on their demographics, history, and friends in common
- Video recommendations: predicting whether a user will watch a video, based on the user’s demographics and other videos they have watched
- Sentiment analysis: predicting whether a movie review is positive or negative, based on the words in the review

# Examples of Supervised Learning Tasks

- Identifying the postal code from handwritten digits on an envelope
- Here the input is a scan of the handwriting, and the desired output is the actual digits in the postal code.
- To create a dataset for building a machine learning model, you need to collect many envelopes. Then you can read the postal codes yourself and store the digits as your desired outcomes.

Input:

Output:

80322

# Examples of Supervised Learning Tasks

- Determining whether a tumor is benign based on a medical image
- Here the input is the image, and the output is whether the tumor is benign.
- To create a dataset for building a model, you need a database of medical images.
- You also need an expert opinion, so a doctor needs to look at all of those images and decide which tumors are benign and which are not.

# Examples of Supervised Learning Tasks

- Detecting fraudulent activity in credit card transactions
- Here the input is a record of the credit card transactions, and the output is whether it is likely to be fraudulent or not.
- Collecting a dataset means storing all transactions and recording if a user reports any transaction as fraudulent.

# Data Collection Methods Vary

- While reading envelopes is laborious, it is easy and cheap.
- Obtaining medical imaging and diagnoses, on the other hand, requires not only expensive machinery but also rare and expensive expert knowledge, not to mention the ethical concerns and privacy issues.
- In the example of detecting credit card fraud, data collection is much simpler. Your customers will provide you with the desired output, as they will report fraud. All you have to do to obtain the input/output pairs of fraudulent and non fradulent activity is wait.

# Unsupervised Algorithms

- In unsupervised learning, only the input data is known, and no known output data is given to the algorithm.
- While there are many successful applications of these methods, they are usually harder to understand and evaluate.
- In this lecture we will briefly overview unsupervised learning; otherwise it is outside the scope of this course and will not be discussed further.

# Examples of Unsupervised Learning Tasks

- Identifying topics in a set of blog posts
- If you have a large collection of text data, you might want to summarize it and find prevalent themes in it. You might not know beforehand what these topics are, or how many topics there might be. Therefore, there are no known outputs.

# Examples of Unsupervised Learning Tasks

- Segmenting customers into groups with similar preferences
- Given a set of customer records, you might wan to identify which customers are similar, and whether there are groups of customers with similar preferences.
- For a shopping site, there might be "parents", "bookworms", or "gamers".
- Because you don't know in advance what these groups might be, or even how many there are, you have no known outputs.

# Examples of Unsupervised Learning Tasks

- Detecting abnormal access patterns to a website
- To identify abuse or bugs, it is often helpful to find access patterns that are different from the norm.
- Each abnormal pattern might be very different.
- Because you only observe traffic, and you don't know what constitutes normal and abnormal behavior, this is an unsupervised problem.

# Data Representation

- For both supervised and unsupervised learning tasks, it is important to have a representation of your input data that a computer can understand.
- Often it is helpful to think of your data as a table. Each point that you want to reason about (each email, each customer, each transaction) is a row, and each property that describes that data point (say, the age of a customer or the amount or location of a transaction) is a column.

# Data Representation

- You might describe users by their age, their gender, when they created an account, and how often they have bought from your online shop.
- You might describe the image of a tumor by the grayscale values of each pixel, or maybe by using the size, shape, and color of the tumor.
- Each entity or row here is known as a sample (or data point) in machine learning, while the columns--the properties that describe these entities--are called features.
- Building a good representation of your data, which is called feature extraction or feature engineering, is critical to the success of a machine learning system.

# Types of Unsupervised Learning

- The main branches of unsupervised learning are clustering, dimensionality reduction, and generative learning.
- Clustering algorithms - The algorithms that group data into clusters based on similarity
- Dimensionality reduction algorithms - The algorithms that simplify our data and faithfully describe it with fewer features
- Generative algorithms - The algorithms that can generate new data points that resemble the existing data

# Reinforcement Learning

- Reinforcement learning is a different type of machine learning in which no data is given, and we must get the computer to perform a task.
- Instead of data, the model receives an environment and an agent who is supposed to navigate in this environment.
- The agent has a goal or a set of goals. The environment has rewards and punishments that guide the agent to make the right decisions to reach its goal.

# Exercise

- For each of the following systems, state whether it uses supervised or unsupervised learning:
  - A recommendation system on a social network that recommends potential friends to a user
  - A system in a news site that divides the news into topics
  - The Google autocomplete feature for sentences
  - A recommendation system on an online retailer that recommends to users what to buy based on their past purchasing history
  - A system in a credit card company that captures fraudulent transactions
  - A system that predicts the number holidaymakers arriving at a resort in a given month.

# References


- Serrano, L.G. (2021). *Grokking Machine Learning*. Manning
- Müller, A.C. & Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.