Problem Understanding

Objectives

- Define common ML terms
- Describe examples of products that use ML and general methods of ML problem-solving used in each
- Identify whether to solve a problem with ML
- Compare and contrast ML to other programming methods
- Apply hypothesis testing and the scientific method to ML problems
- Have conversations about ML problem-solving methods

Common ML Problems

- Machine Learning is the process of training a software, usually called a **Model**
- using a set of data, usually called a **Training set/dataset**
- with the goal of making a good prediction to previously **unseen data**

Example

- Given 10,000,000 records of transaction in e-commerce, we train a model to make a prediction which products the customer are more likely to buy in the future.
- The predictions then can be used to make a personalized product recommendation to a customer.

2 Types of Training in Machine Learning

- With Supervision, usually called **Supervised Learning**
- Without Supervision, usually called **Unsupervised Learning**

Supervised Learning

- **Label is provided** when training the model

Study Case:

- Let’s say you are a teacher, you have been teaching for many years, and you have records of your students’ data.
- After Mid Exam, You want to predict which of your current students are likely to fail
- Given the predictions, you want to give special treatment to them

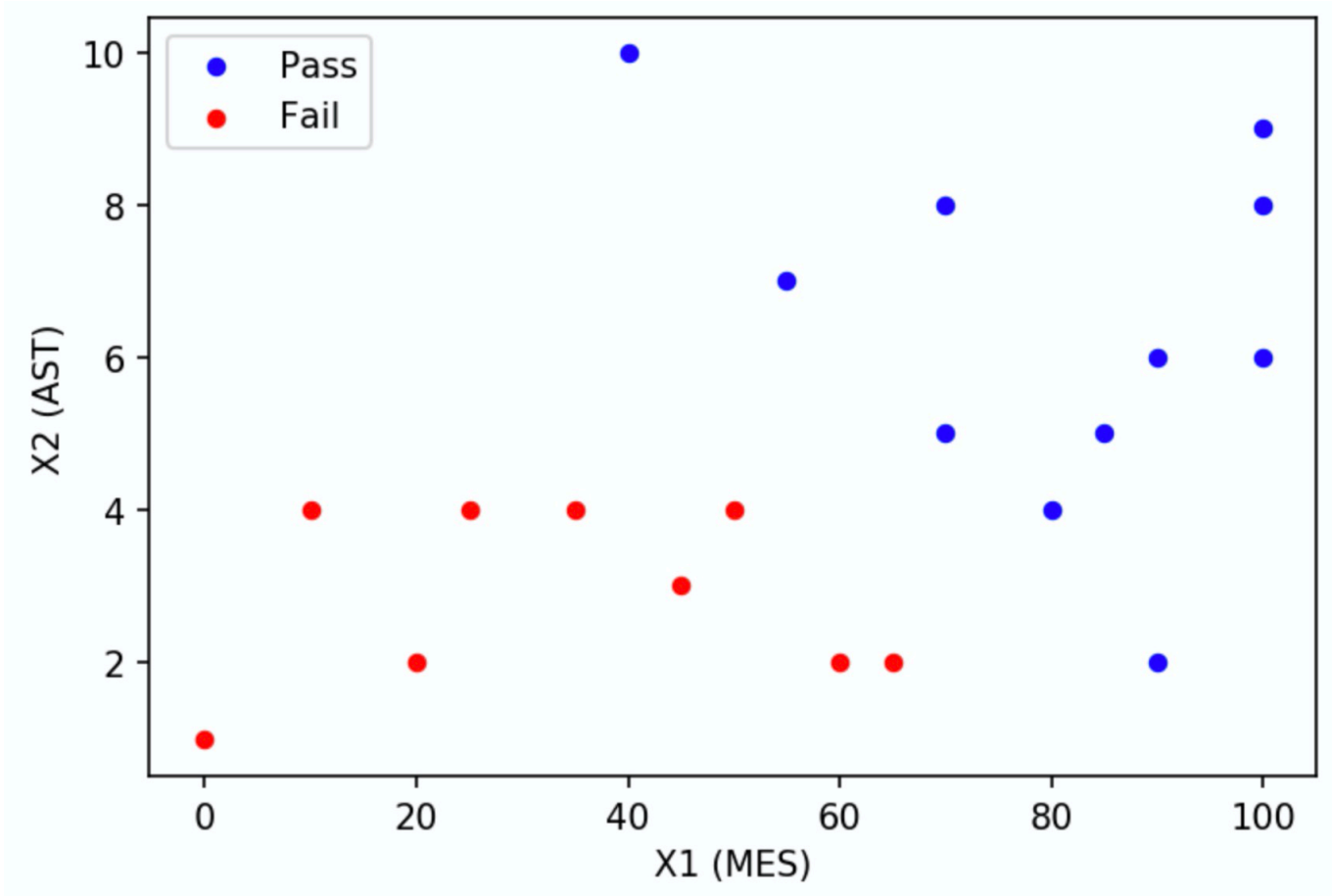
Supervised Learning

Mid Exam Score (MES)	Average Study Time Per Week (AST)	Pass?
80	4	Yes
40	10	Yes
50	4	No
60	2	No

- Mid Exam Score and Average Study Time Per Week are input attributes when training the model, usually called **Features**
- Pass? is a target attribute the model trained to correctly produce, usually called **Label**

Supervised Learning

- First Feature : X_1 (MES)
- Second Feature : X_2 (AST)

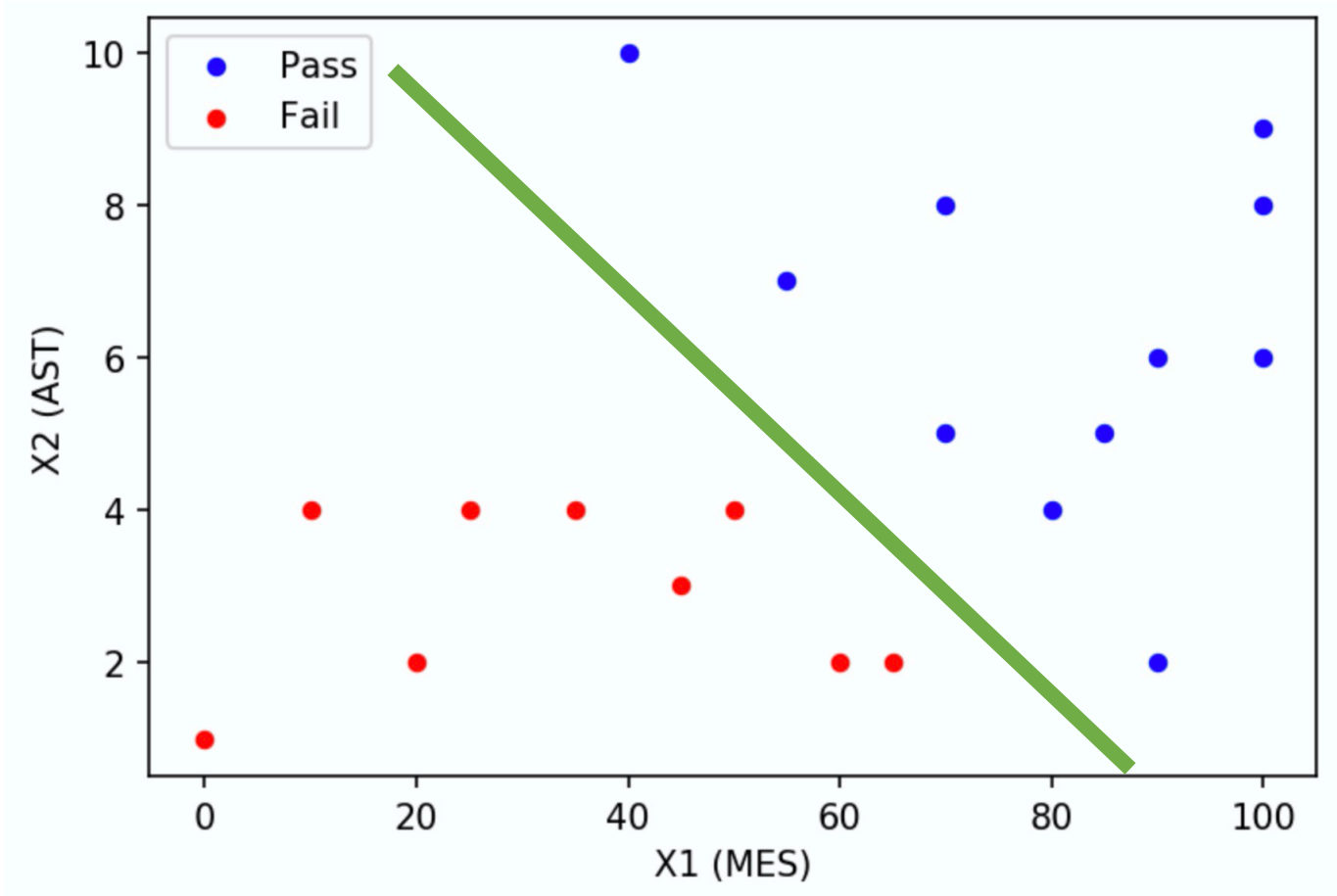


Training

- Feed the features and their corresponding labels into **an algorithm**
- The training is supervision because we are **telling** the algorithm the **expected label** of each input
- During training, the algorithm try to find the relationship between features and their labels
- The resulting relationship is called, **Model**

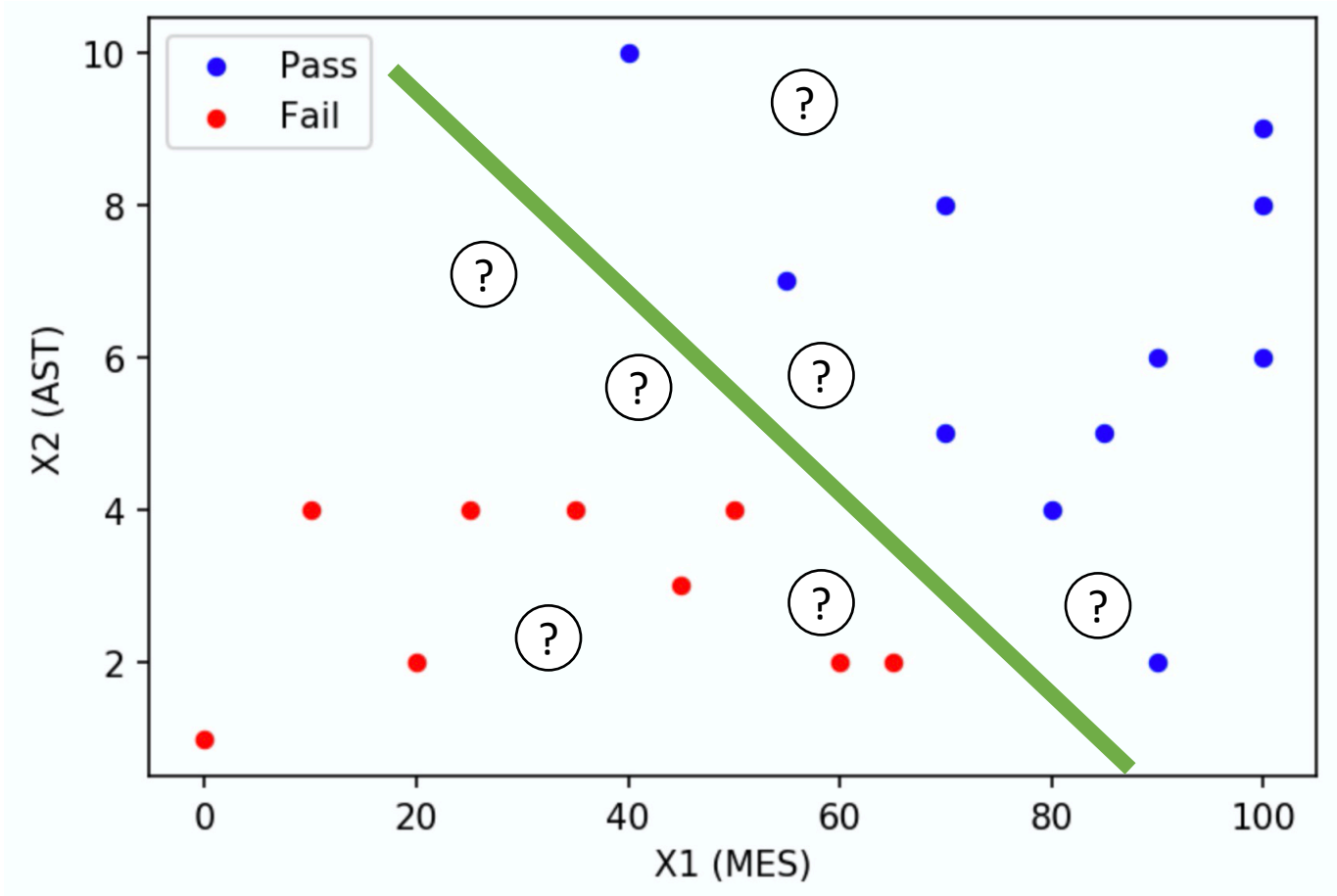
Model

- For example in this case the model can be represented as a line
- Usually, the model can be very complex (e.g non-linear)



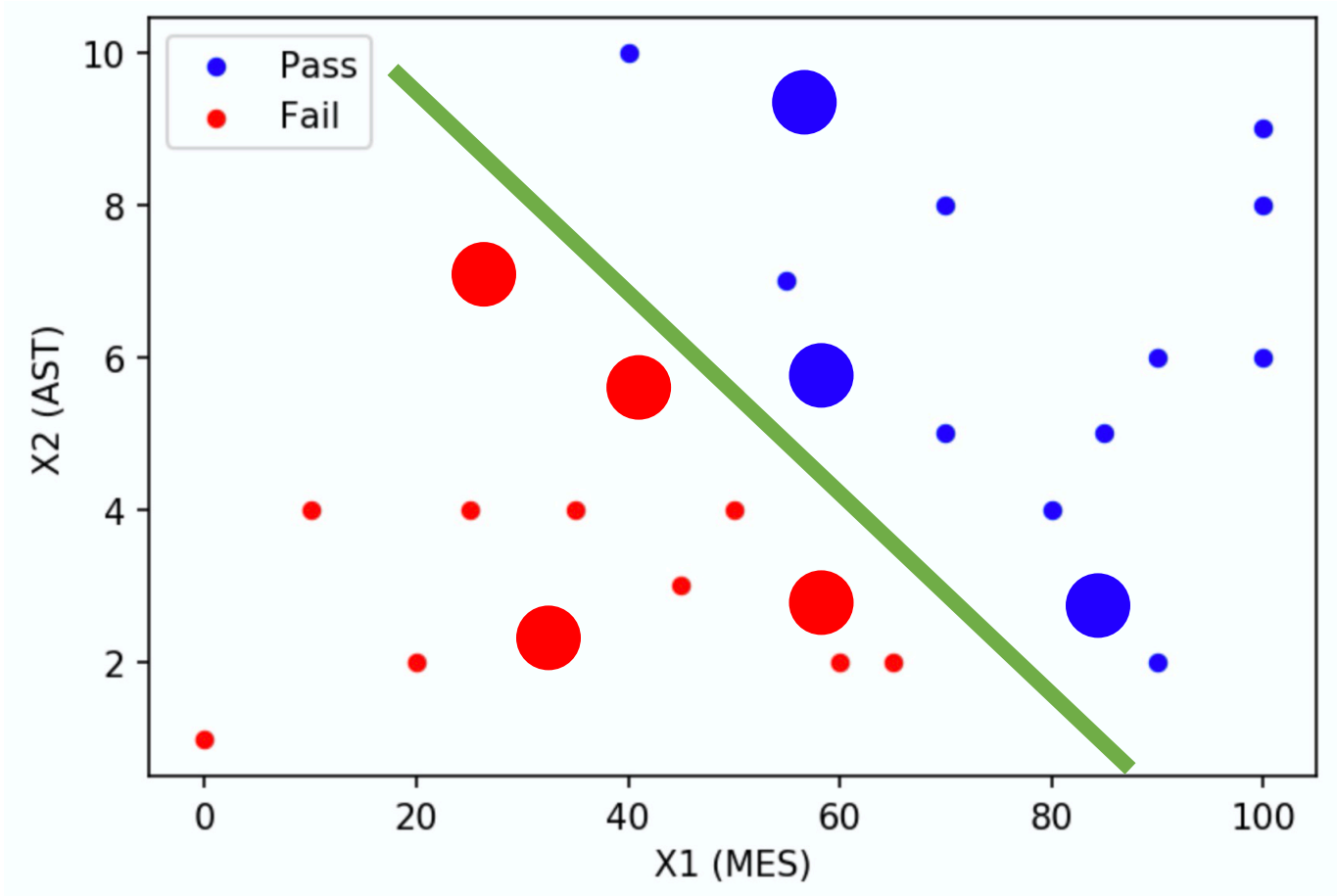
Testing

- During testing, we give **unseen** data to the model
- The model will give predictions



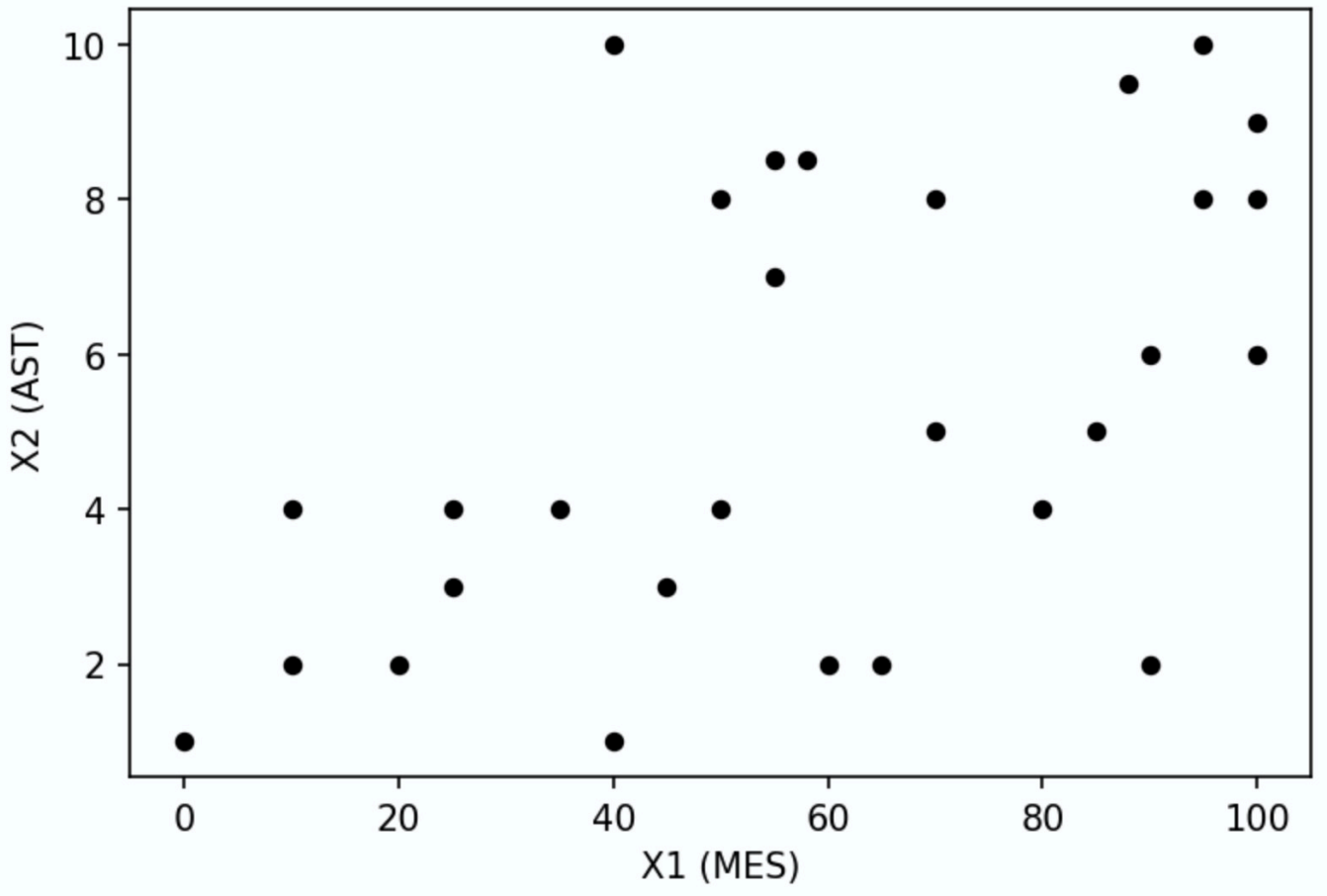
Testing

- During testing, we give **unseen** data to the model
- The model will give predictions



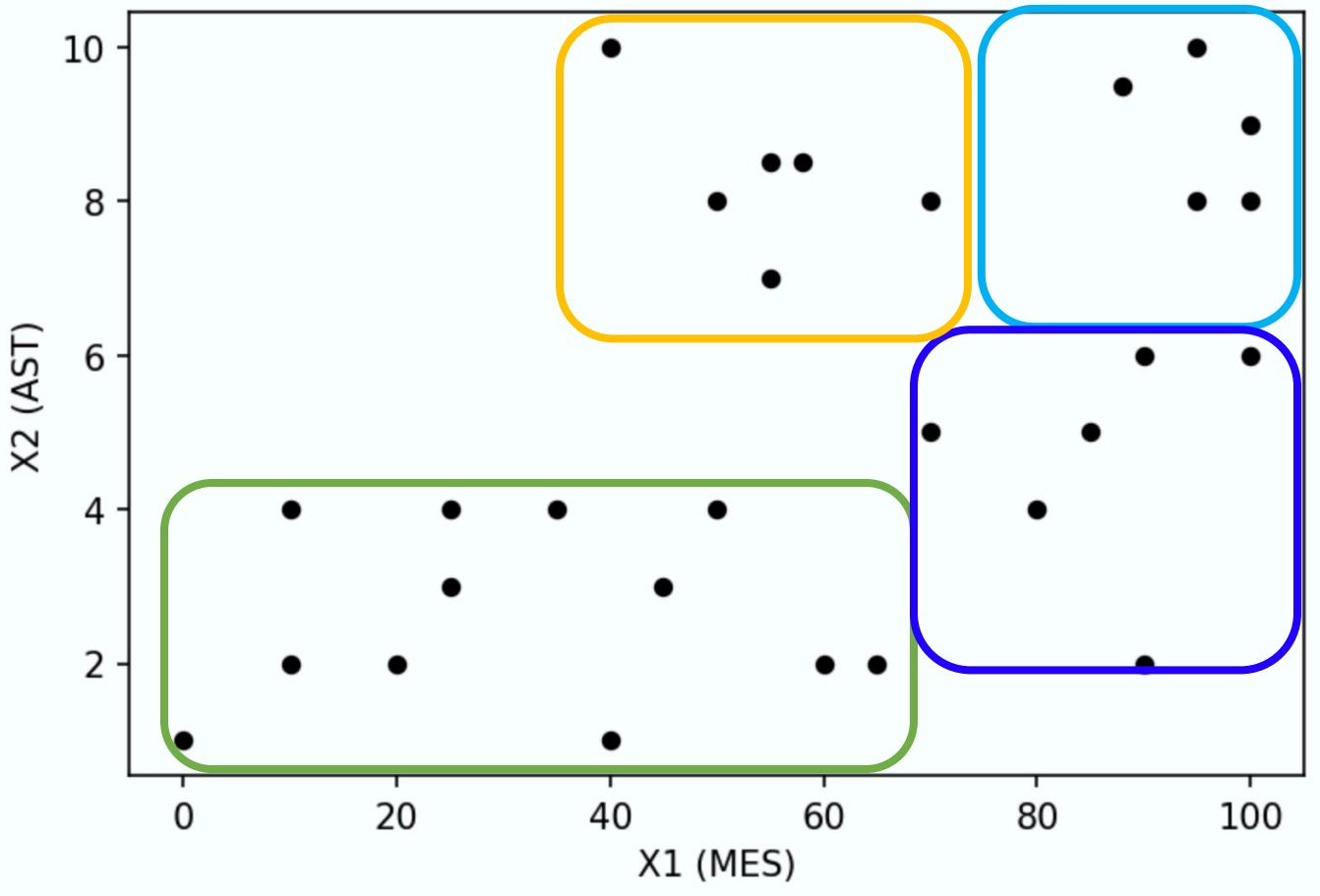
Unsupervised Learning

- The goal is to find a meaningful pattern in the data
- There is no specific label for each data



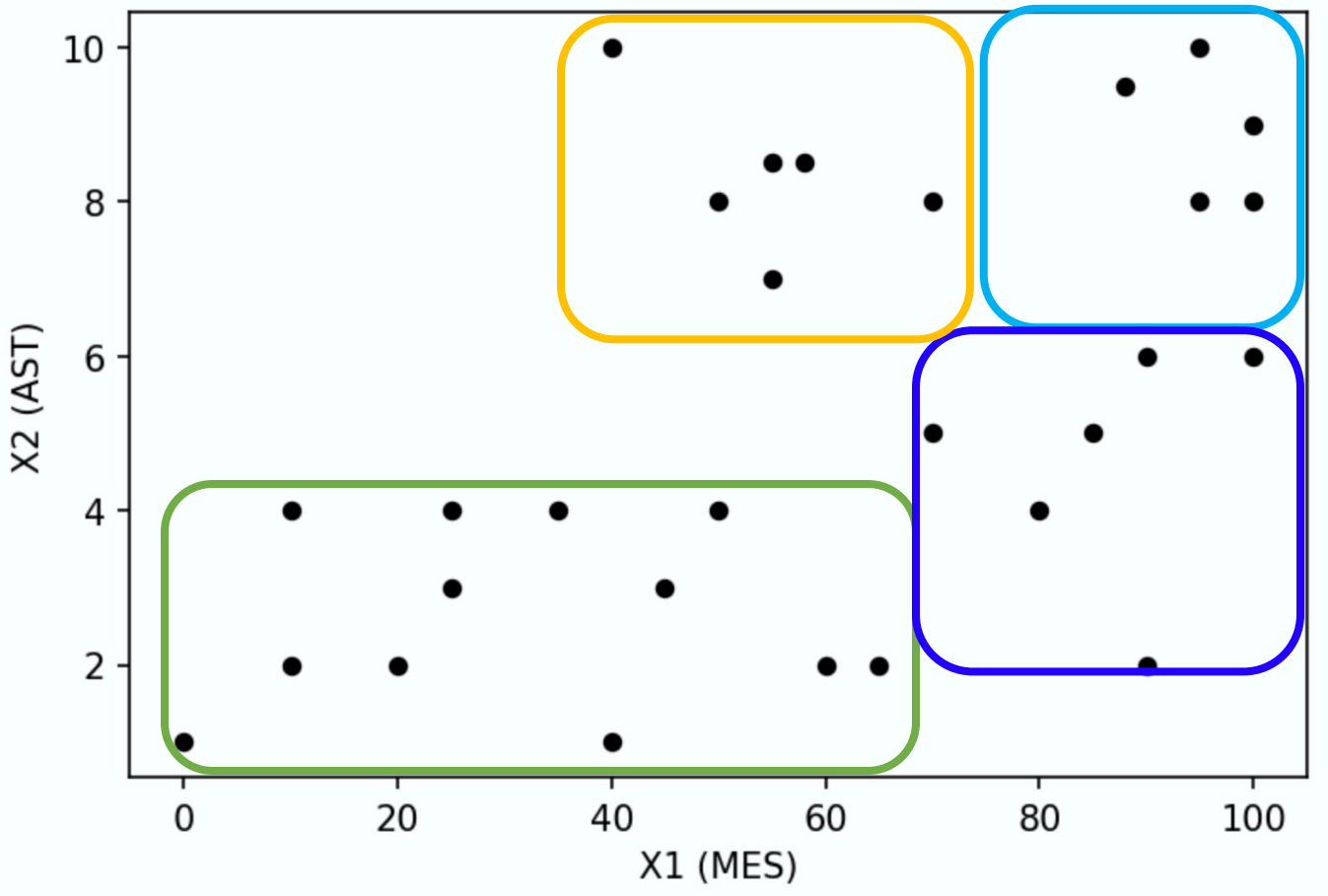
Unsupervised Learning

- The goal is to find a meaningful pattern in the data
- There is no specific label for each data
- Here we have try to group the data into 4 groups, usually called **Clusters**



Unsupervised Learning

- What do these clusters represent?
- Sometimes it can be described, but most of the time it's hard (lots of features).



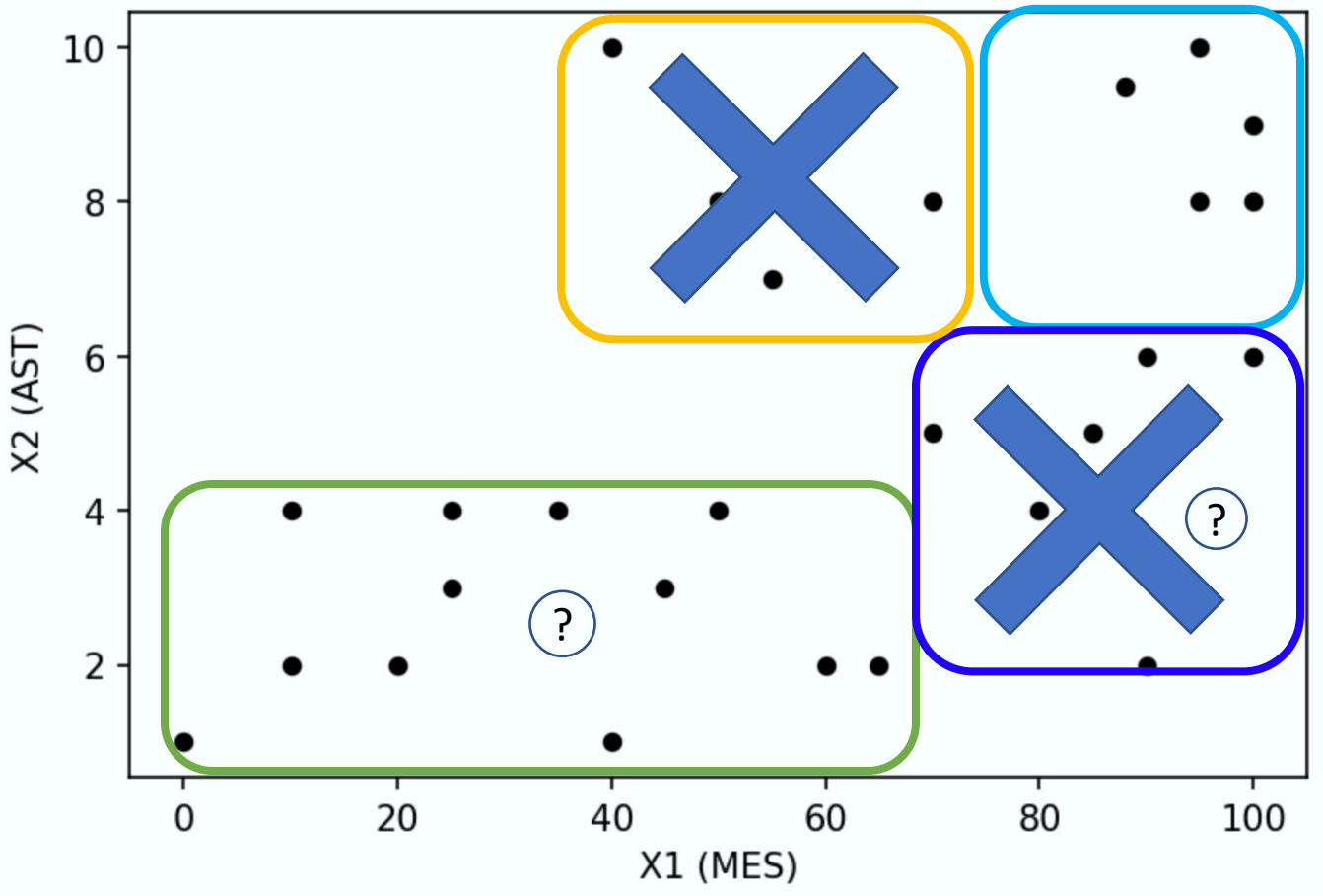
Unsupervised Learning

Study Case:

- You are a marketing manager, and you want to promote a new product, let’s say a brand new coffee.
- You cannot target all people, because it might not be effective.
- So, what you can do is try to find groups of people according to some characteristics, and test your product to some groups.

Unsupervised Learning

- Let’s say two groups do like your coffee
- Later, if you get new data (potential customers), you can identify it’s group, and decide whether you will promote your product or not



Types of Machine Learning Problems

Type of ML Problem	Description	Example
Classification	Pick one of N labels	Cat, dog, horse, or bear
Regression	Predict numerical values	Click-through rate
Clustering	Group similar examples	Most relevant documents (unsupervised)
Association rule learning	Infer likely association patterns in data	If you buy hamburger buns, you're likely to buy hamburgers (unsupervised)
Structured output	Create complex output	Natural language parse trees, image recognition bounding boxes
Ranking	Identify position on a scale or status	Search result ranking

Experimental Design

Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems?
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

From Problem to ML Solution

1. Articulate Your Problem Clearly
2. Identify Your Data Sources
3. Identify Potential Learning Problems
4. Think About Potential Bias and Ethics

1. Articulate Your Problem Clearly

Example:

Our problem is best framed as 3-class, single-label classification, which predicts whether a video will be in one of three classes—{very popular, somewhat popular, not popular}—28 days after being uploaded.

2. Identify Your Data Sources

- How much labeled data do you have?
  - Define your inputs and labels
  - ML Model usually needs lots of data to train (e.g. thousands for linear model, hundred thousands for Neural Network model)
- What is the source of your label?
  - Is it difficult to collect the data?
  - How much work needed to transform the raw data to the input format?
- Is your label closely connected to the decision you will be making?

2. Identify Your Data Sources

- Is your label closely connected to the decision you will be making?
  - Decision is something important and may impact business
  - Pick good prediction/label to support Decision Making

Prediction	Decision
What video the learner wants to watch next.	Show those videos in the recommendation bar.
Probability someone will click on a search result.	If P(click) > 0.12, prefetch the web page.
What fraction of a video ad the user will watch.	If a small fraction, don't show the user the ad.

3. Identify Potential Learning Problems

- The data set doesn't contain enough positive labels.
- The training data doesn't contain enough examples.
- The labels are too noisy.
- The system memorizes the training data, but has difficulty generalizing to new cases.

4. Think About Potential Bias and Ethics

- Stereotyping, prejudice or favoritism towards some things, people, or groups over others.
  - Will this ML problem has a potential to offend some groups? Using skin color as features for predicting criminal may not not appropriate.
- Systematic error introduced by a sampling or reporting procedure.
  - Does the population represented in the dataset match the population that the machine learning model is making predictions about?

References

- Google Developer Machine Learning Problem Framing

Any Question?