

Classification

What is Classification?

- Classification is a data mining technique used to **categorize data** into predefined ***classes or labels*** based on some **predefined properties**.

Examples:

1. A **bank loans officer** needs analysis of data to learn which loan applicants are “**safe**” and which are “**risky**” for the bank.
2. A **marketing manager** needs data analysis to help guess whether a customer with a given profile will **buy** a new computer.
3. A **medical researcher** wants to analyse cancer data to predict which one of three specific treatments a patient should receive.

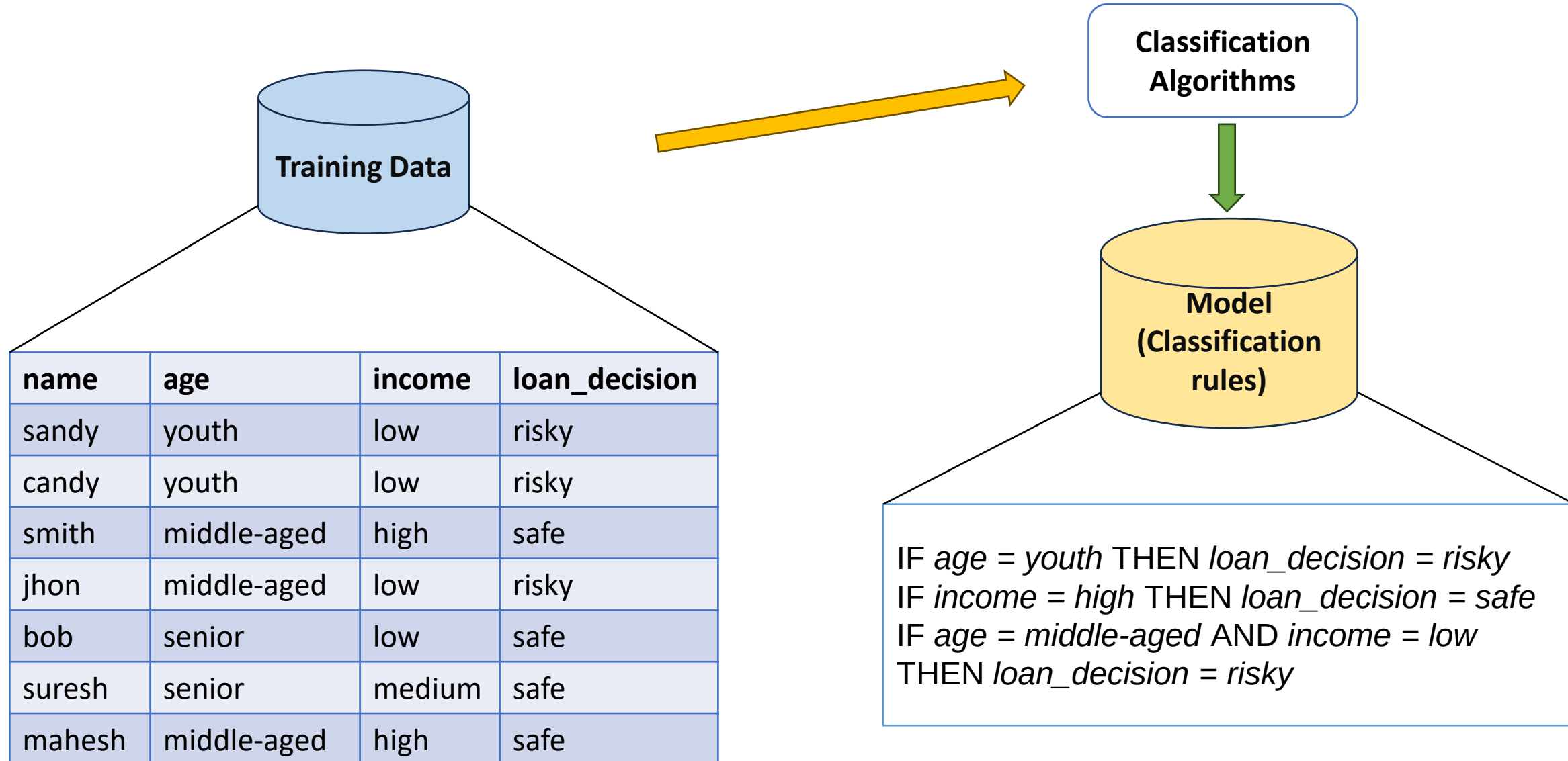
How does Classification Works?

- ***Data classification*** is a **two-step** process:
 - ***Learning step:***
 - In this, a classification algorithm builds the ***classifier/model*** by analyzing or “***learning from***” a training dataset.
 - A trained model or classifier is represented in the form of ***classification rules***.
 - ***Classification step:***
 - In this, ***Test data*** are used to estimate the accuracy of the ***model*** (classification rules).
 - The classifier/model is used to ***predict class labels*** for the given new(unseen) data.

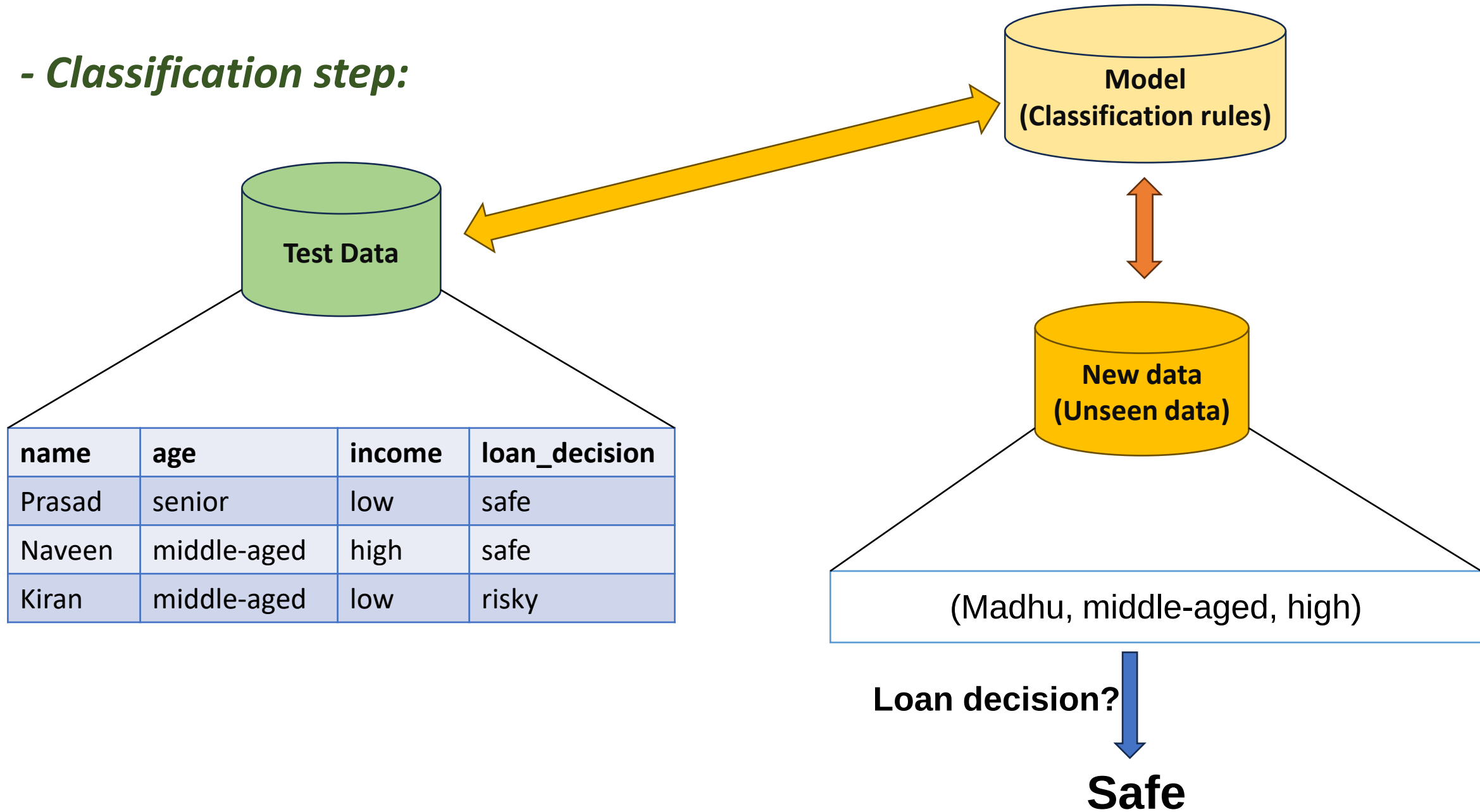
Example:

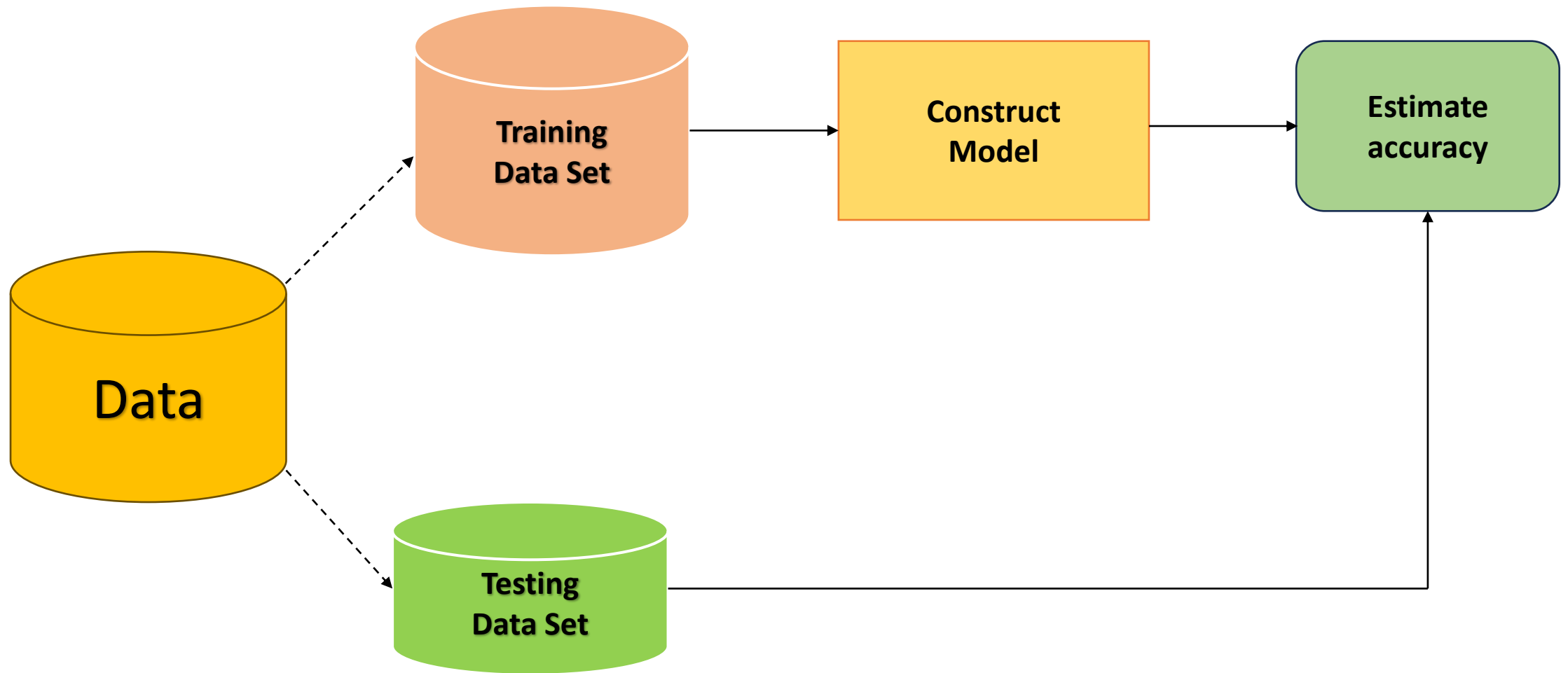
A **bank loans officer** needs analysis of data to learn which loan applicants are “**safe**” and which are “**risky**” for the bank.

- *Learning step:*



- Classification step:





The **accuracy** of a classifier on a given **test data set** is the percentage of test set tuples that are correctly classified by the classifier.