



Introduction Introduction to Supervised Algorithms



Why do we need Learning?

- Algorithms
- Super abilities of solving every task
- Experience
- AI

Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- Learning is used when:
 - Human expertise does not exist (navigating on Mars),
 - Humans are unable to explain their expertise (speech recognition)
 - Solution changes in time (routing on a computer network)
 - Solution needs to be adapted to particular cases (user biometrics)

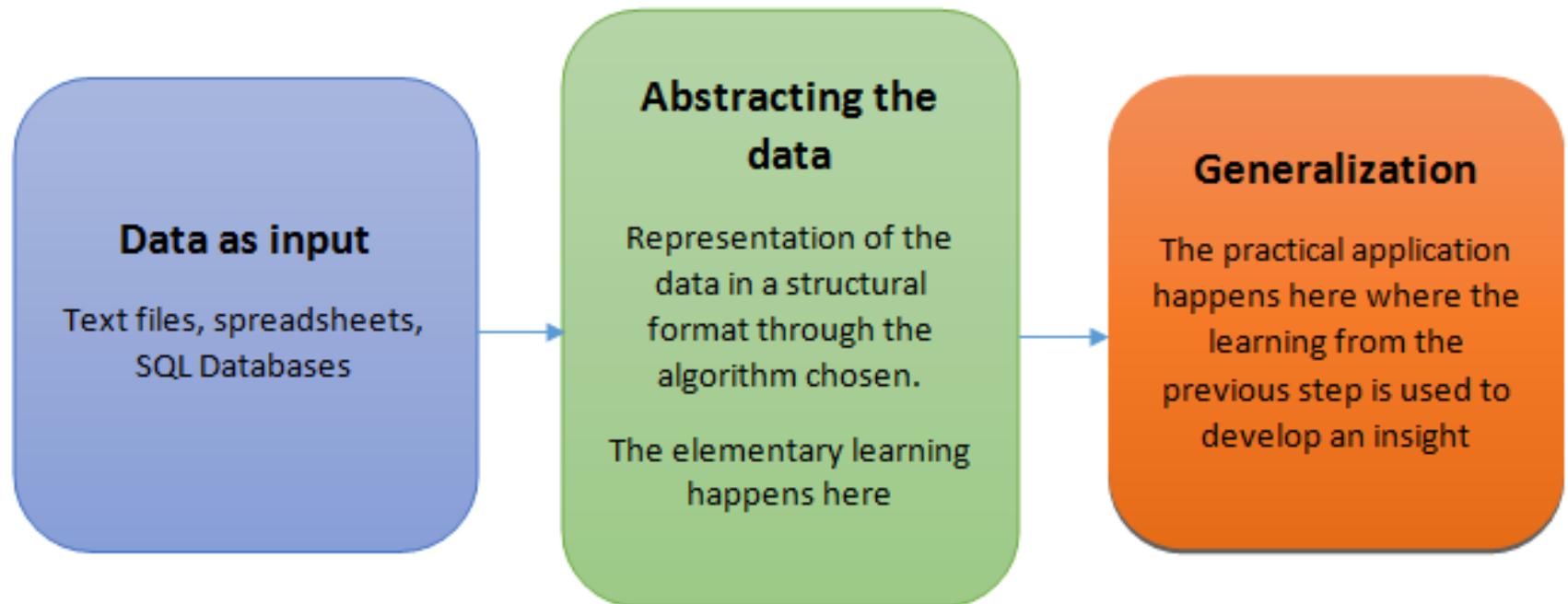
What We Talk About When We Talk About “Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

People who bought “Da Vinci Code” also bought “The Five People You Meet in Heaven” (www.amazon.com)

- Build a model that is *a good and useful approximation* to the data.

How exactly do we teach machines





Types of Learning

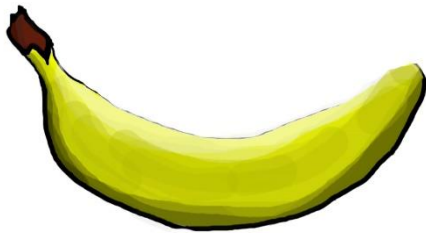
- **Supervised (inductive) learning**
 - Training data includes desired outputs
- **Unsupervised learning**
 - Training data does not include desired outputs
- **Semi-supervised learning**
 - Training data includes a few desired outputs
- **Reinforcement learning**
 - Rewards from sequence of actions

Supervised learning

suppose you are given an basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:



- If shape of object is rounded and depression at top having color **Red** then it will be labelled as – **Apple**.
- If shape of object is long curving cylinder having color **Green-Yellow** then it will be labelled as – **Banana**.



Since machine has already learnt the things from previous data and this time have to use it wisely. It will first classify the fruit with its shape and color, and would confirm the fruit name as BANANA and put it in Banana category.

Steps Involved in Supervised Learning

- First Determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training **dataset**, **test dataset**, and **validation dataset**.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

Supervised learning classification

- **Classification:** A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.
- **Regression:** A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

It is called supervised learning because the process of an learning(from the training dataset) can be thought of as a teacher who is supervising the entire learning process. Thus, the “learning algorithm” iteratively makes predictions on the training data and is corrected by the “teacher”, and the learning stops when the algorithm achieves an acceptable level of performance(or the desired accuracy).

Supervised Learning: Uses

Supervised Learning algorithm learns from a known data-set(Training Data) which has labels to make predictions

- **Prediction of future cases:** Use the rule to predict the output for future inputs
- **Knowledge extraction:** The rule is easy to understand
- **Compression:** The rule is simpler than the data it explains
- **Outlier detection:** Exceptions that are not covered by the rule, e.g., fraud

Supervised Learning : Regression and Classification

Classification



Man

Woman

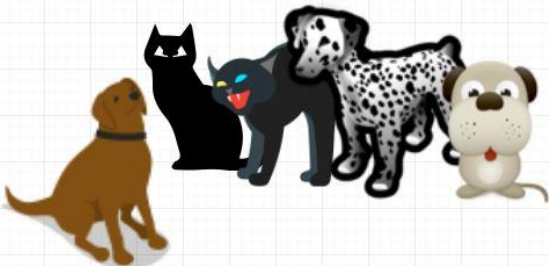


Regression

- Predict the price of house

Unsupervised learning

- Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
- the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.



machine has no any idea about the features of dogs and cat

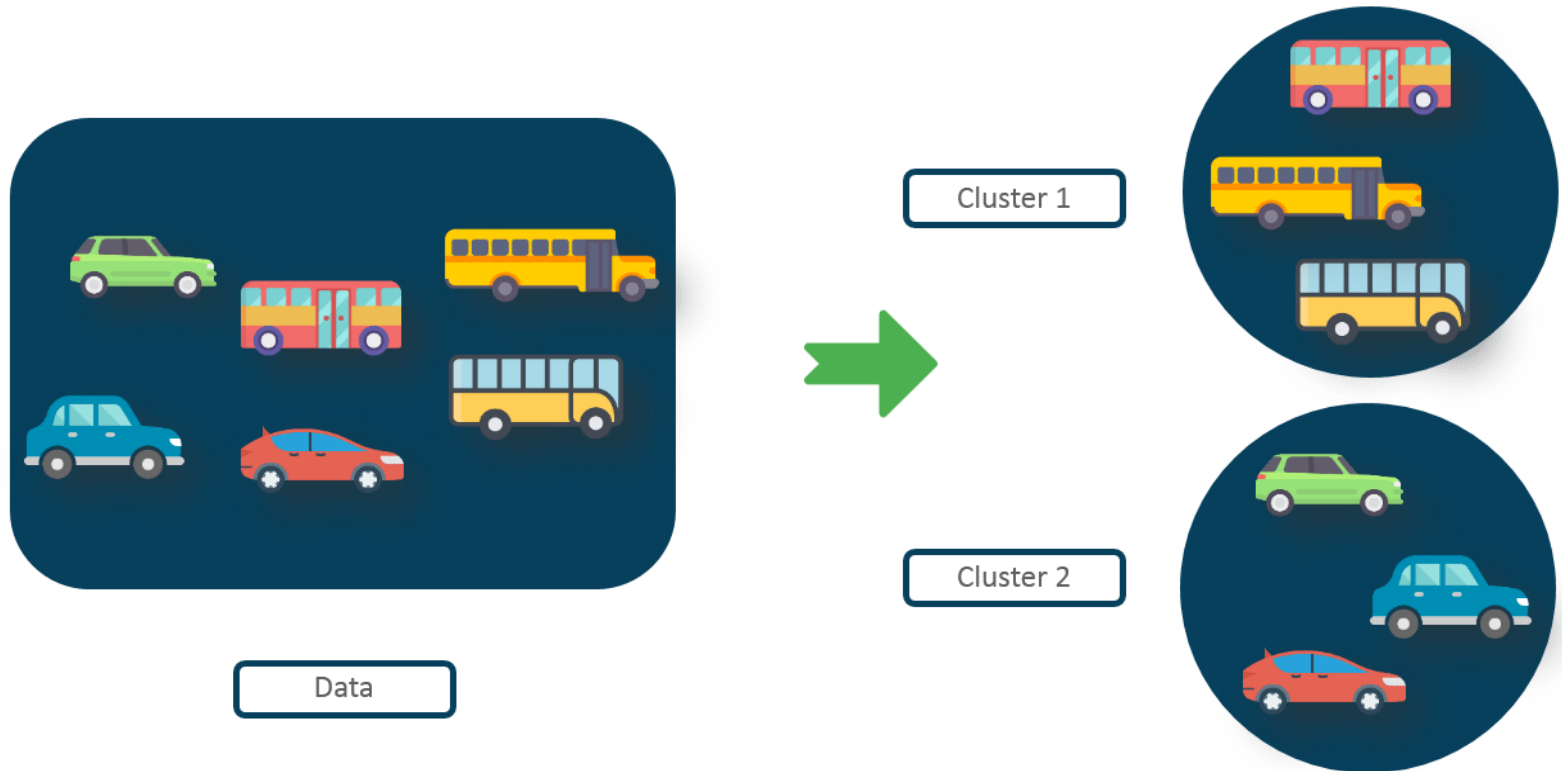
it can categorize them according to their similarities, patterns and differences

Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis
- Example applications
 - Customer segmentation in CRM
 - Image compression: Color quantization
 - Bioinformatics: Learning motifs

unsupervised learning

clustering





SUPERVISED LEARNING

UNSUPERVISED LEARNING

Input Data

Uses Known and
Labeled Data as input

Uses Unknown Data
as input

Computational
Complexity

Very Complex

Less Computational
Complexity

Real Time

Uses off-line analysis

Uses Real Time
Analysis of Data

Number of Classes

Number of Classes
are known

Number of Classes
are not known

Accuracy of Results

Accurate and Reliable
Results

Moderate Accurate
and Reliable Results

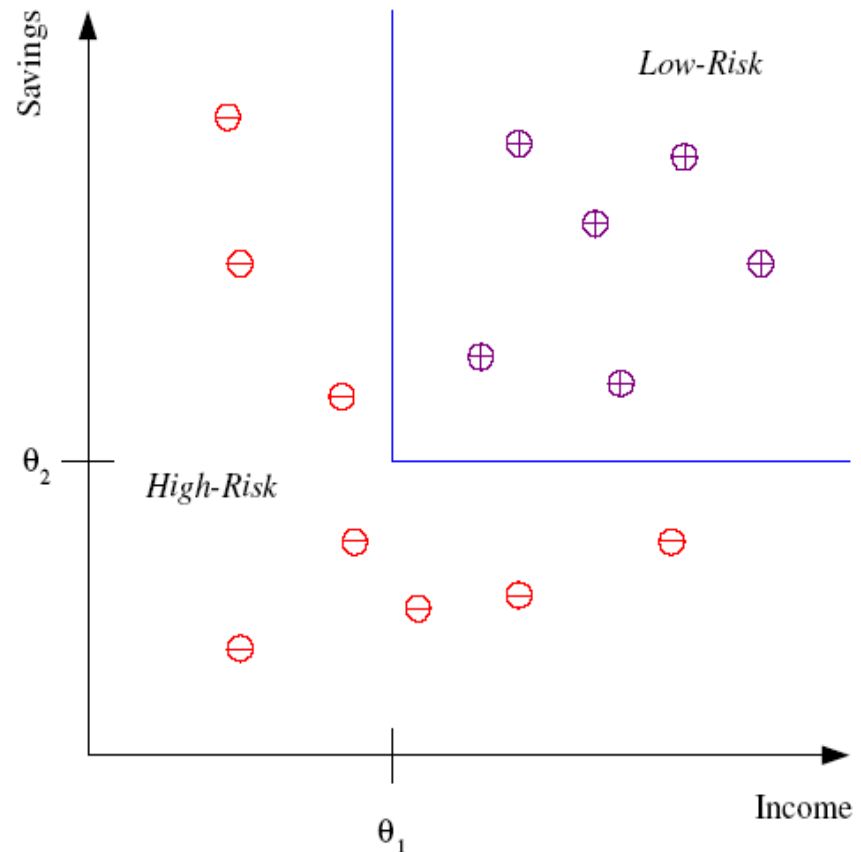


Applications

- Association
- Supervised Learning
 - Classification
 - Regression
- Unsupervised Learning
- Reinforcement Learning

Classification

- Example: Credit scoring
- Differentiating between **low-risk** and **high-risk** customers from their *income* and *savings*



Discriminant: IF $income > \theta_1$ AND $savings > \theta_2$
THEN **low-risk** ELSE **high-risk**

Classification: Applications

- Aka Pattern recognition
- **Face recognition:** Pose, lighting, occlusion (glasses, beard), make-up, hair style
- **Character recognition:** Different handwriting styles.
- **Speech recognition:** Temporal dependency.
 - Use of a dictionary or the syntax of the language.
 - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- **Medical diagnosis:** From symptoms to illnesses
- ...

Face Recognition

Training examples of a person



Test images



Unsupervised Learning

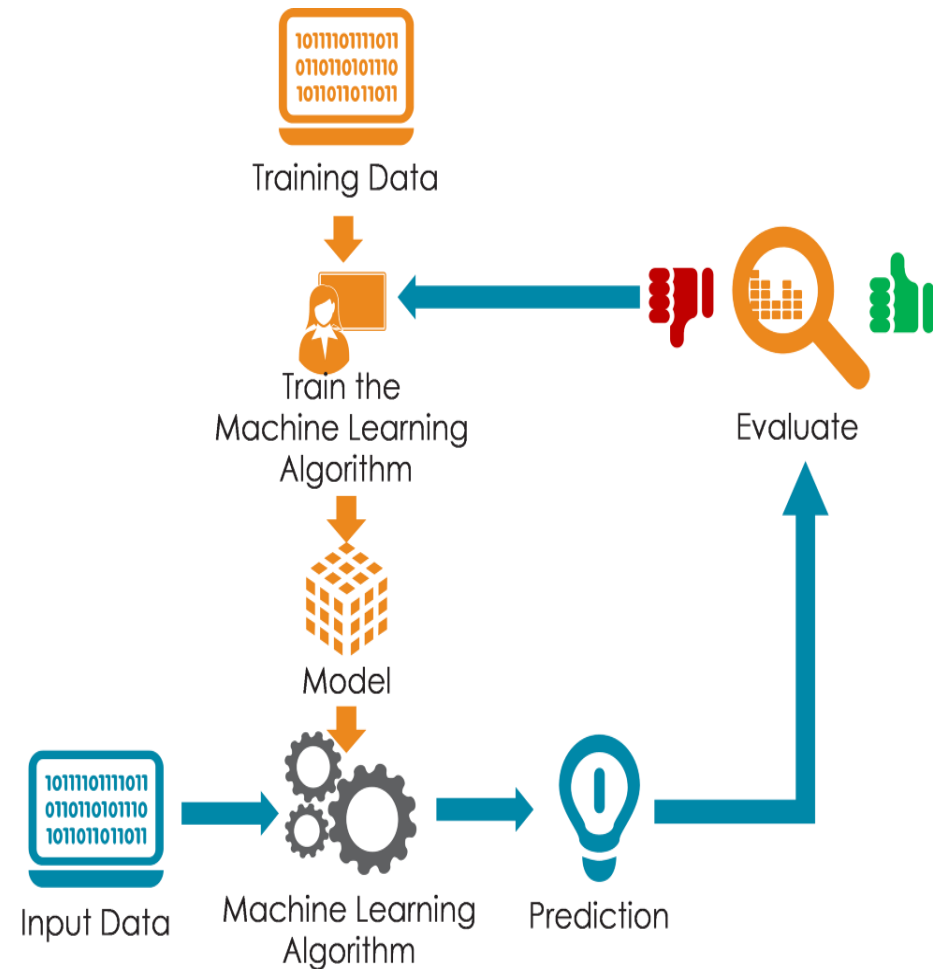
- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Example applications
 - Customer segmentation in CRM
 - Image compression: Color quantization
 - Bioinformatics: Learning motifs

Reinforcement Learning

- Learning a policy: A **sequence** of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Model

- a system for mapping inputs to outputs
- represents a theory about a problem
- to predict house prices, we could make a model that takes in the square footage of a house and outputs a price

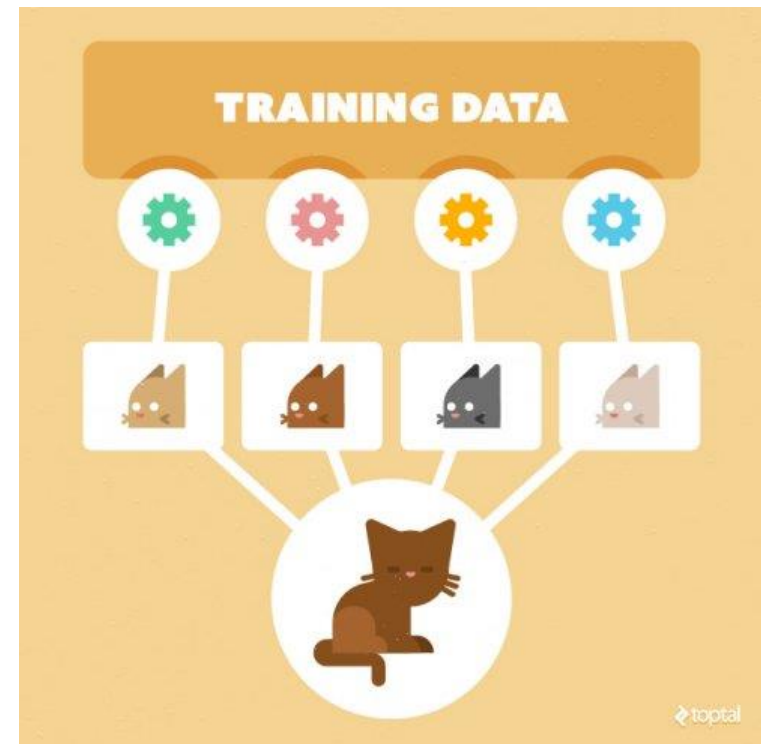



Model

- A model learns relationships between the inputs, called features, and outputs, called labels, from a training dataset
- A “model” in machine learning is the output of a machine learning algorithm run on data.
- A model represents what was learned by a machine learning algorithm.
- The model is the “thing” that is saved after running a machine learning algorithm on training data and represents the rules, numbers, and any other algorithm-specific data structures required to make predictions.
- **The best analogy is to think of the machine learning model as a “program.”**


Training Data

- The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.





Training data is used to help your machine learning model make predictions. It's the largest part of your dataset, forming at least 70-80% of the total data you'll use to build your model. This data is used exhaustively across multiple training cycles to improve the accuracy of your algorithm. Training data is different from validation and testing data in that its classes are often evenly distributed. Depending on your task, this might mean that the data doesn't accurately reflect its real-world use case.

- 
- **How Much Training Data Do I Need?**
 - **Why is it Difficult to Estimate Dataset Size?**
 1. Diversity of input
 2. Tolerance for errors
 3. **Complexity of model**
 4. Training method

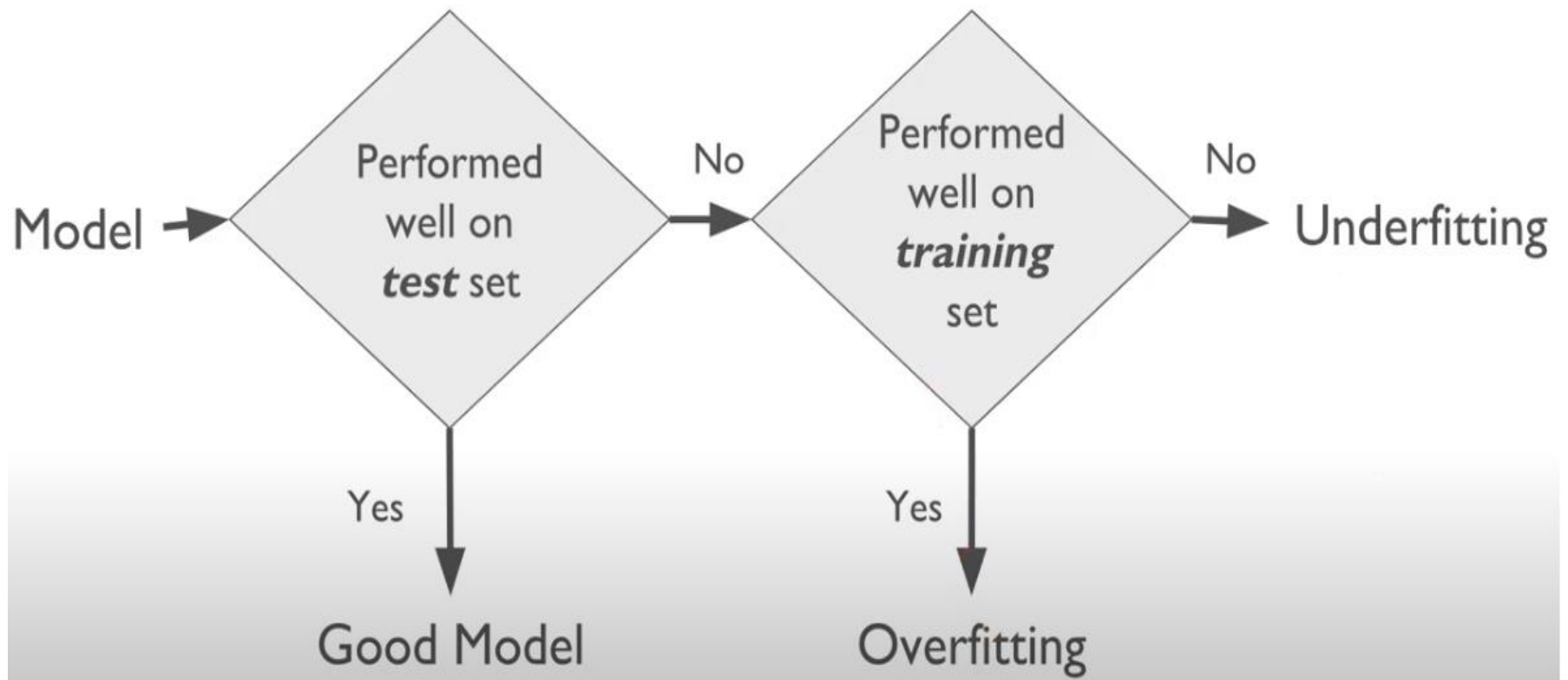
The deciding factor for how much data you'll need is your project's unique requirements and goals. Each project requires a unique balance of all of these influencing factors, which you'll have to figure out for yourself when coming up with that target dataset size. Keeping this in mind, let's now dive into some of the ways that you can begin to figure out your data needs.

What is Quality?

| CHARACTERISTIC | DEFINITION | ACTION ITEMS |
|--------------------------|---|---|
| Uniformity | All data points attribute values equally and come from comparable sources | Check for irregularities when pulling data from multiple internal or external sources |
| Consistency | All data points have the same | Ensure that classes are distributed |
| Comprehensiveness | Dataset has enough parameters to cover all of the model's use cases, including edge cases | Check that you have enough data; include examples of edge cases in an appropriate volume |
| Relevancy | Dataset contains only parameters which are useful to your model | Identify important parameters; consider asking a domain expert to perform analysis |
| Diversity | Dataset accurately reflects the model's user base | Perform user analysis to uncover hidden biases; consider pulling data from both internal and external sources; consider employing an expert for a third-party perspective |

What is Overfitting & Underfitting?

- **Overfitting** refers to the scenario where a machine learning model can't generalize or fit well on unseen dataset. A clear sign of machine learning overfitting is if its error on the testing or validation dataset is much greater than the error on training dataset.
- **Overfitting** is a term used in statistics that refers to a modeling error that occurs when a function corresponds too closely to a dataset. As a result, overfitting may fail to fit additional data, and this may affect the accuracy of predicting future observations.





Class Activity

- **Applications of Clustering in different fields**
- Justify the need of an algorithm