INTRODUCTION TO MACHINE LEARNING

UNIT-1

Topics Covered

What is Machine Learning?

Need for Machine Learning

Applications of Machine Learning

Machine Learning Techniques and Categorization

Supervised Learning

Unsupervised Learning

Semi supervised Learning

Reinforcement Learning

Training, Test, Validation data set

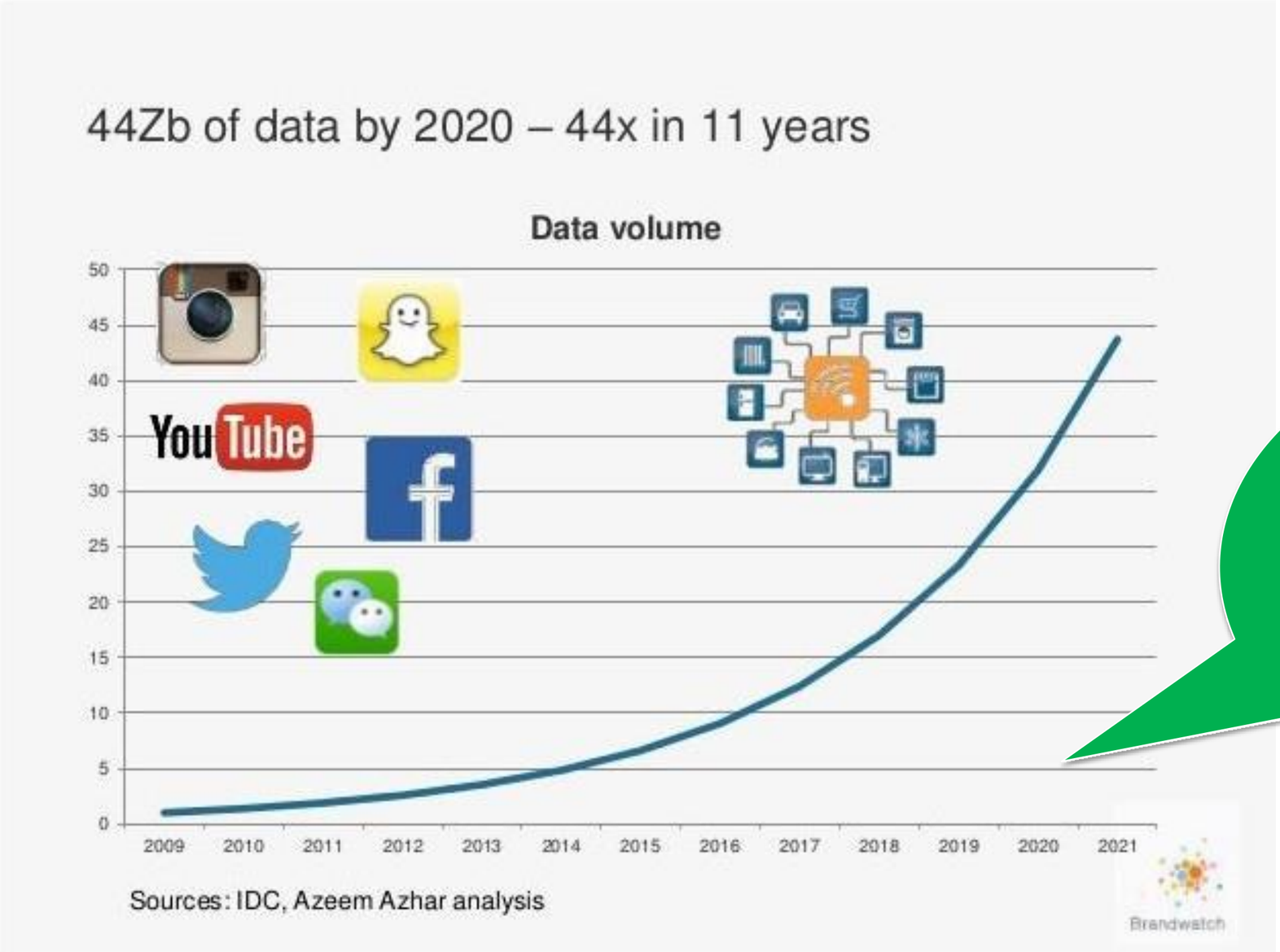
Disclaimer

The contents of this presentation are adapted from following sources:

1. Simplilearn.com
2. AnalyticsVidhya
3. Machine Learning Tom Mitchell McGraw hill education
4. Coursera.org
5. Online blogs and resources

NEED FOR MACHINE LEARNING

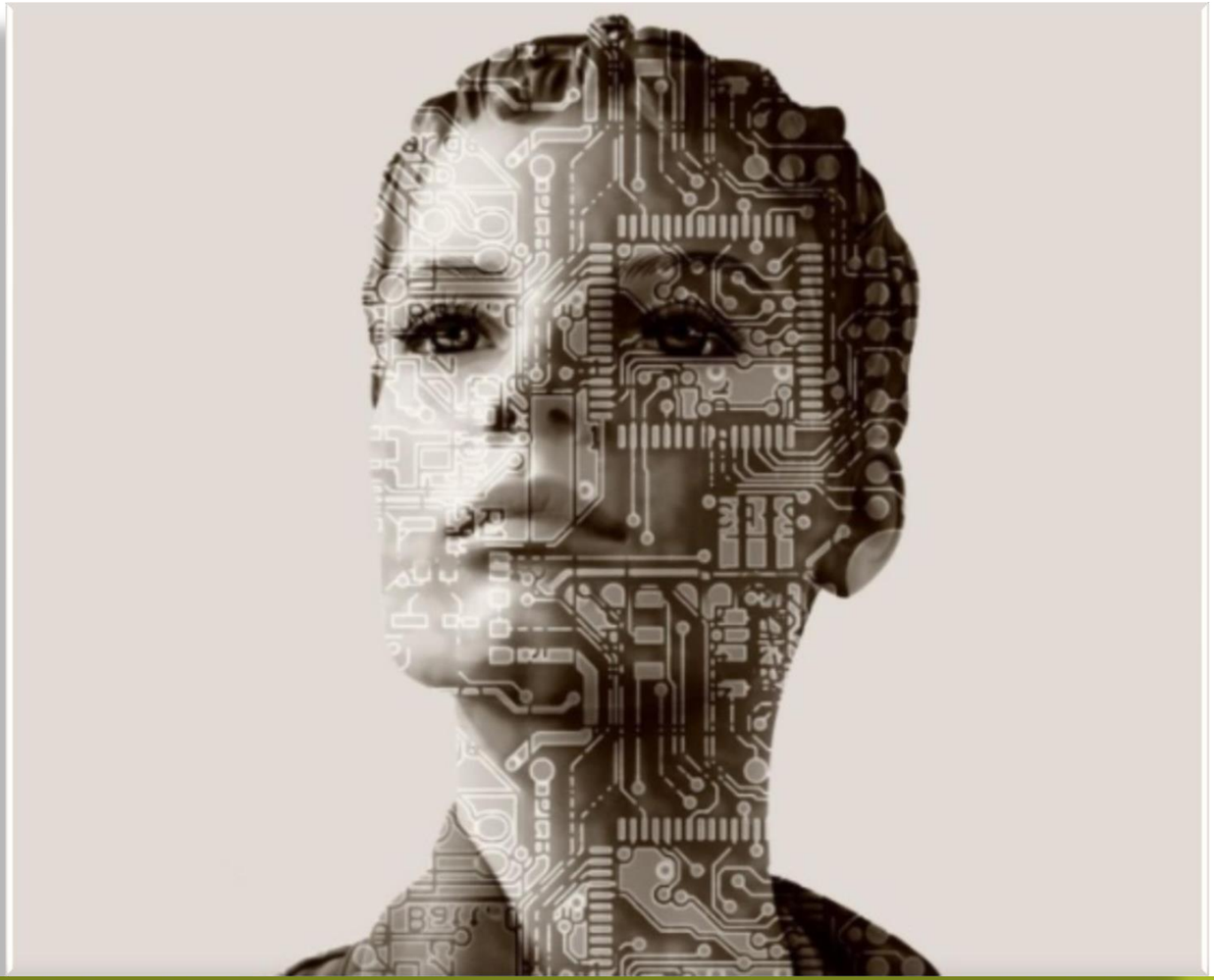
Data Economy



Explosion of data has given rise to a new economy and there is a constant battle for ownership of data between enterprises to derive benefits from it.

Emergence of Artificial Intelligence

Science associated with data is going toward a new paradigm where one can teach machines to learn from data and derive a variety of useful insights giving rise to Artificial Intelligence.



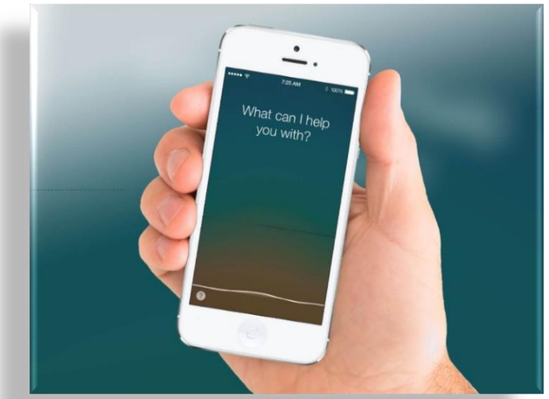
Definition of Artificial Intelligence

> Artificial Intelligence refers to intelligence displayed by machines that simulates human and animal intelligence.

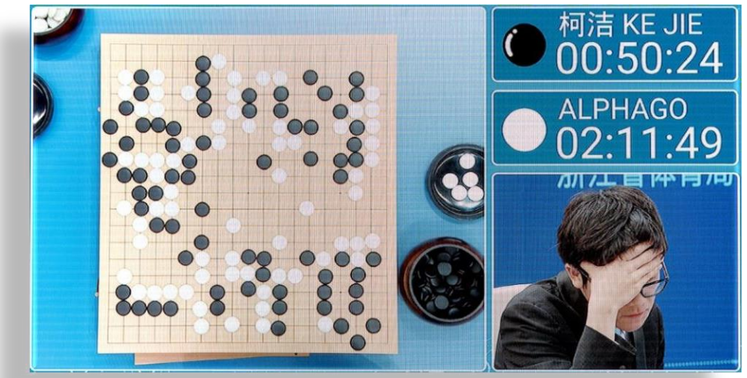
Artificial Intelligence in Practice

AI is redefining industries by providing greater personalization to users and automating processes.

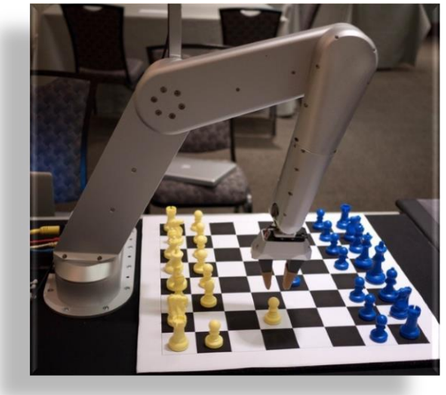
Self-driving cars



Siri(iPhone)



Google's AlphaGo



Chess

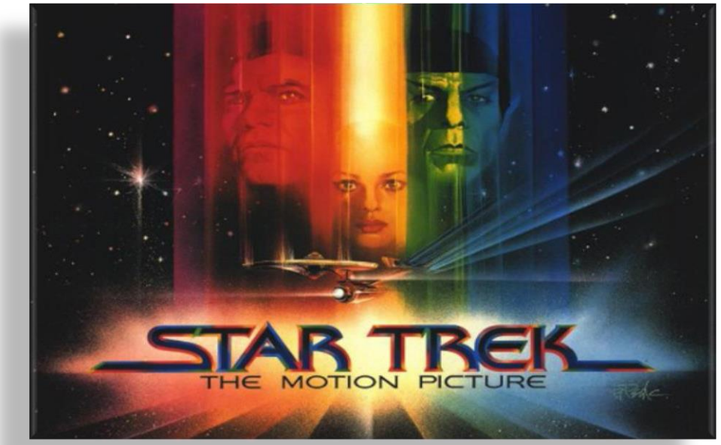
Amazon ECHO

Concierge robot from IBM Watson

Sources: documentarytube, wired, Quora

Sci-Fi Movies with the Concept of AI

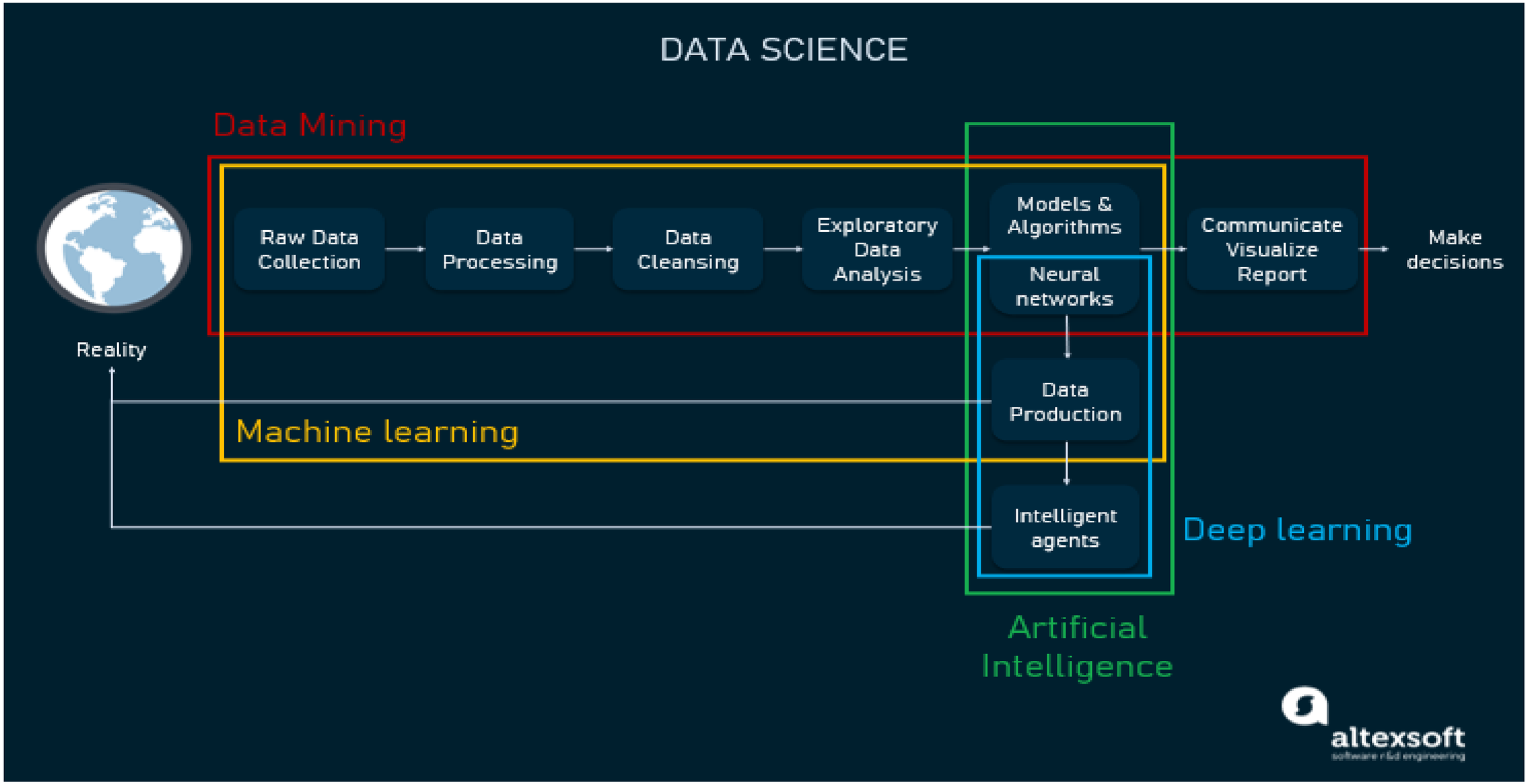
Few AI films spanned through the decades that reflect the everchanging spectrum of our emotions regarding the machines we have created



Data Facilitates in Recommendations

Amazon collects data from users and recommends the best products according to the user's buying/shopping pattern.

Relationship between Artificial Intelligence and Machine Learning



WHAT IS MACHINE LEARNING

Definition of Machine Learning

> The capability of Artificial Intelligence systems to learn by extracting patterns from data is known as Machine Learning.

Machine learning paradigms

Supervised learning

- Learn an input and output map
  - Classification – Categorical output
  - Regression – Continuous output

Unsupervised learning

- Discover patterns in the data
  - Clustering – Cohesive grouping
  - Association – Frequent co-occurrence

Reinforcement Learning

- Learning Control

Machine Learning Task

Classification

- Accuracy

Regression

- Error

Clustering

- Scatter/purity

Association

- Support/confidence

Reinforcement Learning

- Cost/reward

Source: Introduction to Machine Learning (NPTEL)](http://nptel.ac.in/courses/106106139/2)(NPTEL)

Challenges

How good a model?

How do I choose a model?

Do I have enough data?

Is the data of sufficient quality?

How confident can I be of the results?

Am I describing data correctly?

Features of Machine Learning

01

It uses the data to *detect patterns* in a dataset and *adjust program actions accordingly*

It *focuses on* the *development of computer programs* that can teach themselves to *grow and change* when *exposed* to *new data*

02

03

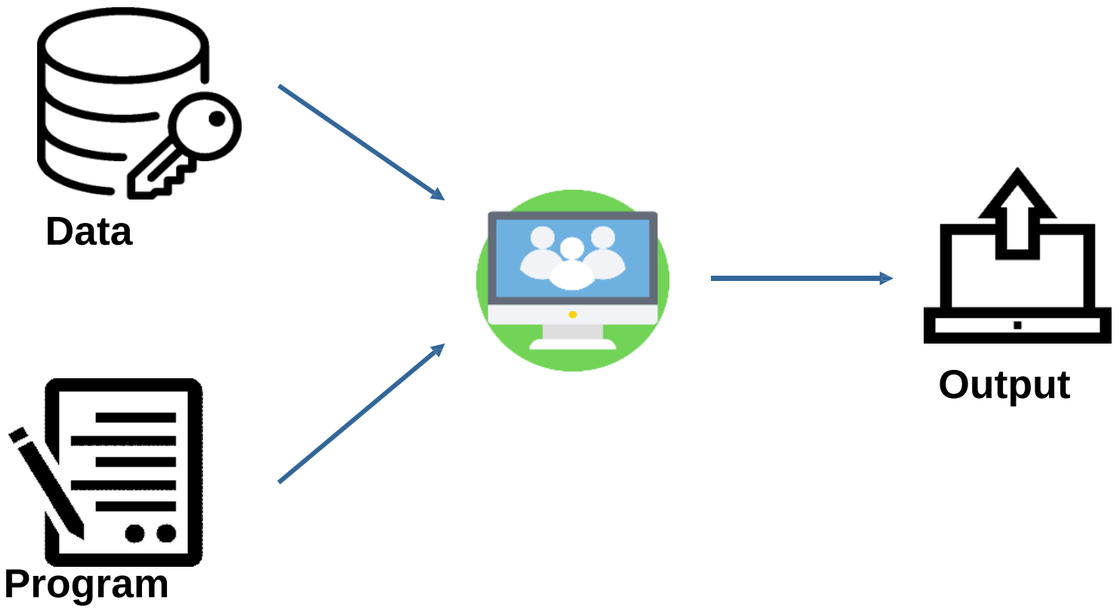
It enables computers to *find hidden insights using iterative algorithms without being explicitly programmed*

It *automates analytical model building*

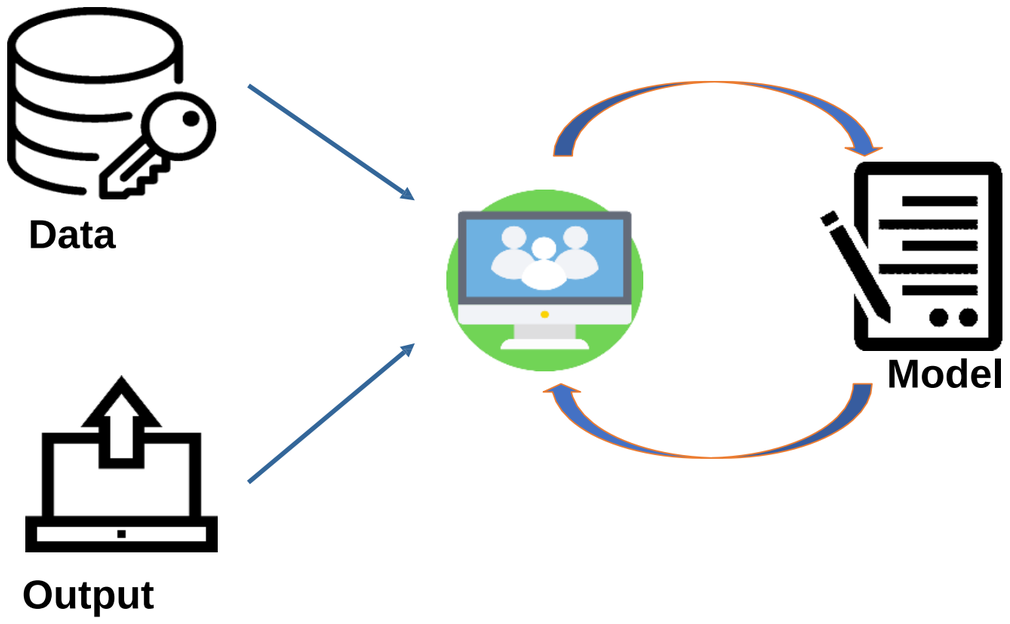
04

Traditional Approach vs. Machine Learning Approach

Traditional Programming: Data and program is run on the computer to produce the output

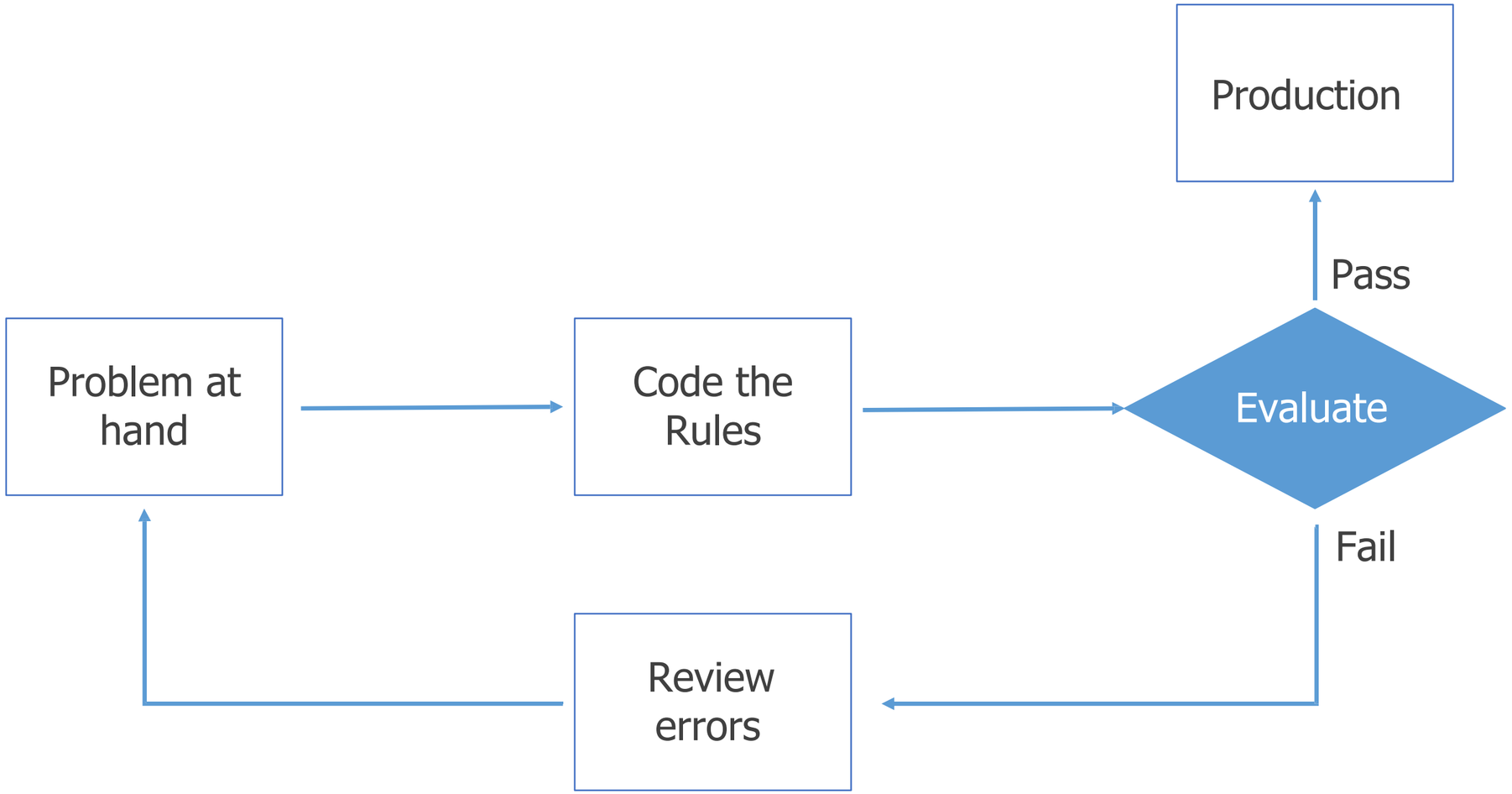


Machine Learning: Data and output is run on the computer to create a program



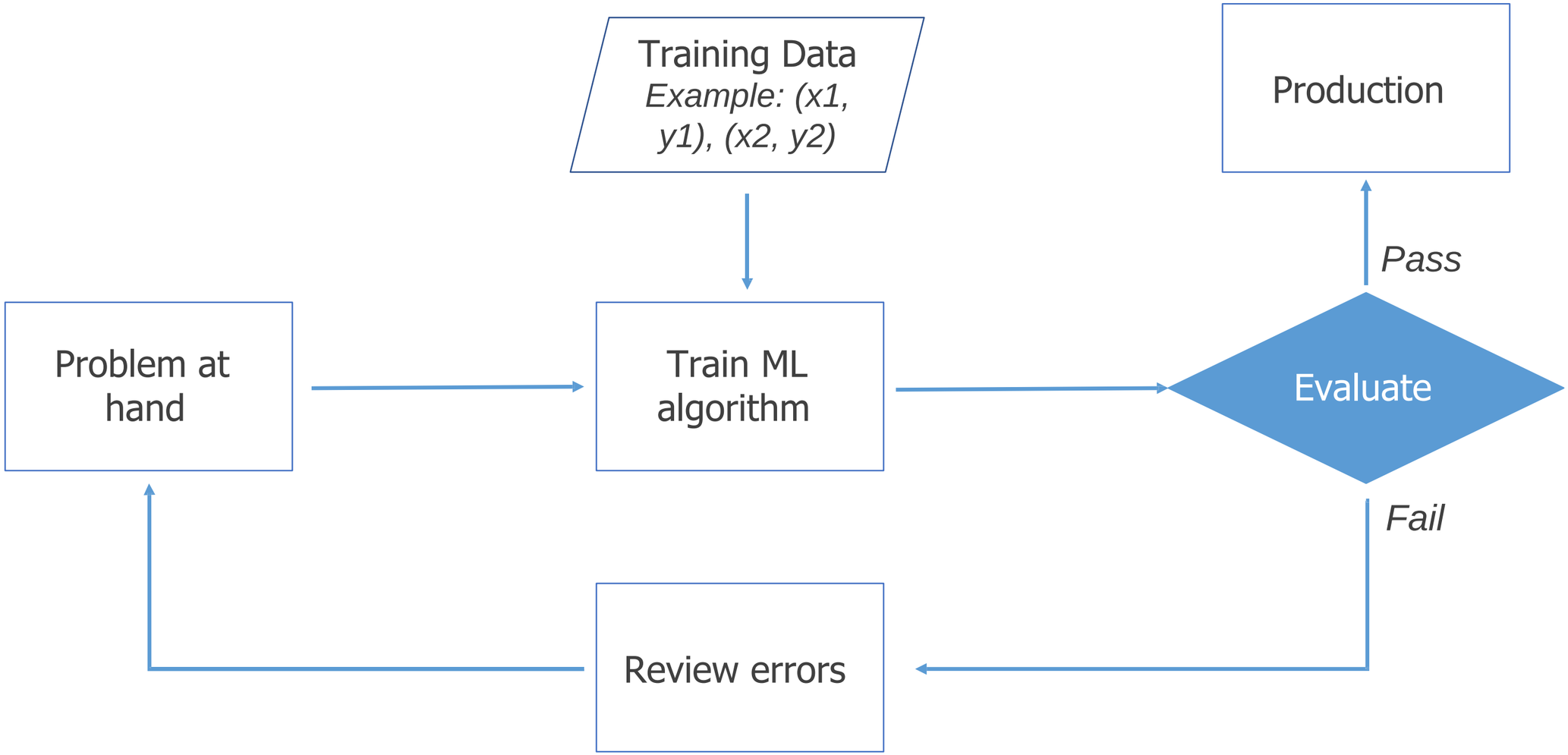
Traditional Approach

Traditional programming relies on hard-coded rules.



Machine Learning Approach

Machine Learning relies on learning patterns based on sample data.



Intelligence – Traditional vs Machine Learning

How to write a Spam Filter?

This Photo by Unknown Author is licensed under CC BY-NC

Source: Introduction to Machine Learning (CloudX Lab)

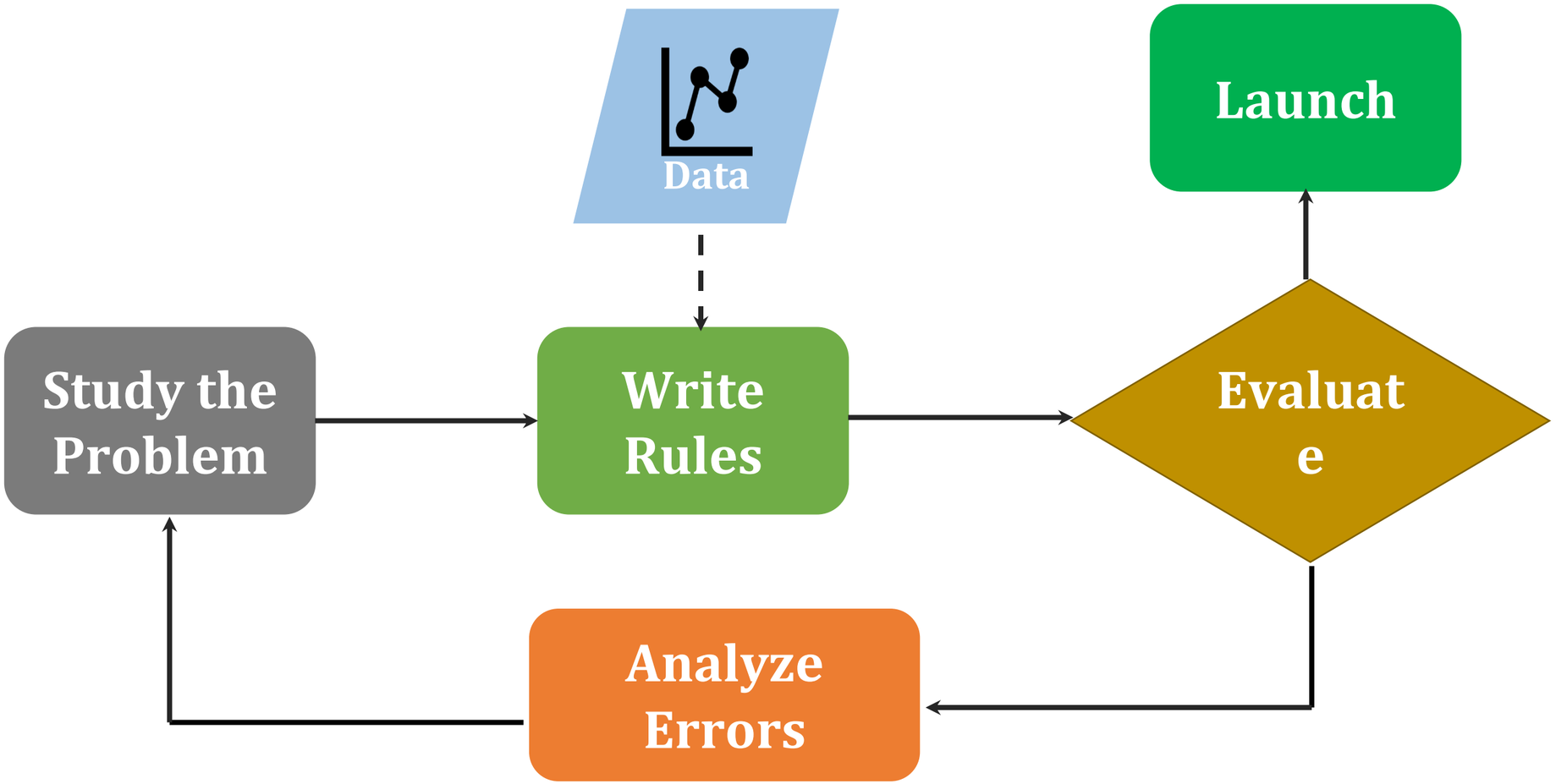
Problems?

- All their emails containing **“4U”** are blocked
- They might start writing **“For U”** instead
- If spammers keep working around spam filter, we will need to keep writing
- new rules forever

- Program will likely become a long list of complex rules
- Pretty hard to maintain

Source: Introduction to Machine Learning (CloudX Lab)

Intelligence – Spam Filter – Traditional Approach



Source: Introduction to Machine Learning (CloudX Lab)

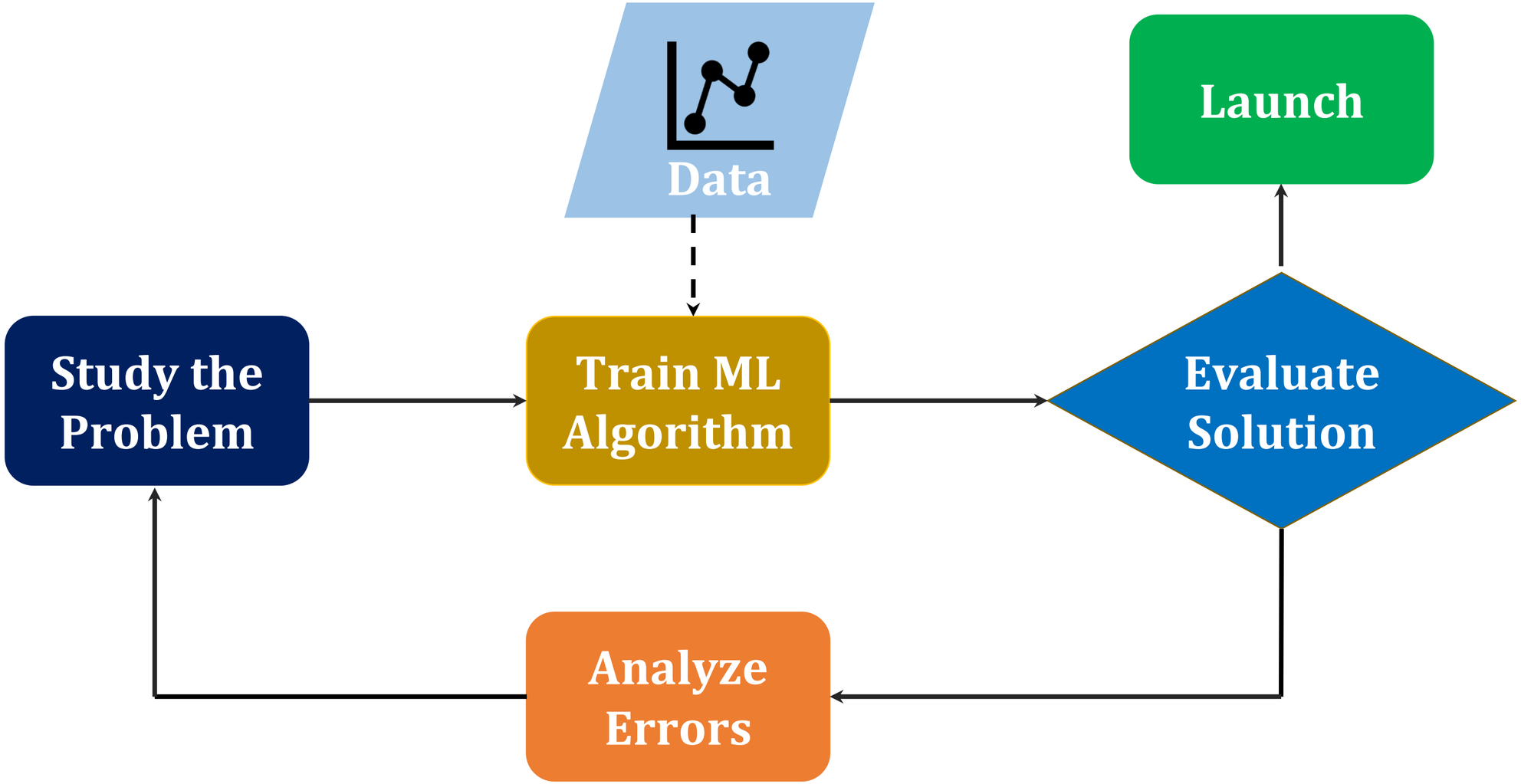
Intelligence - Spam Filter - ML Approach

- A spam filter based on Machine Learning techniques automatically learns
  - Which words and phrases are good predictors of spam
  - By detecting unusually frequent patterns of words

- The program will be
  - Much shorter
  - Easier to maintain
  - Most likely more accurate than traditional approach

Source: Introduction to Machine Learning (CloudX Lab)

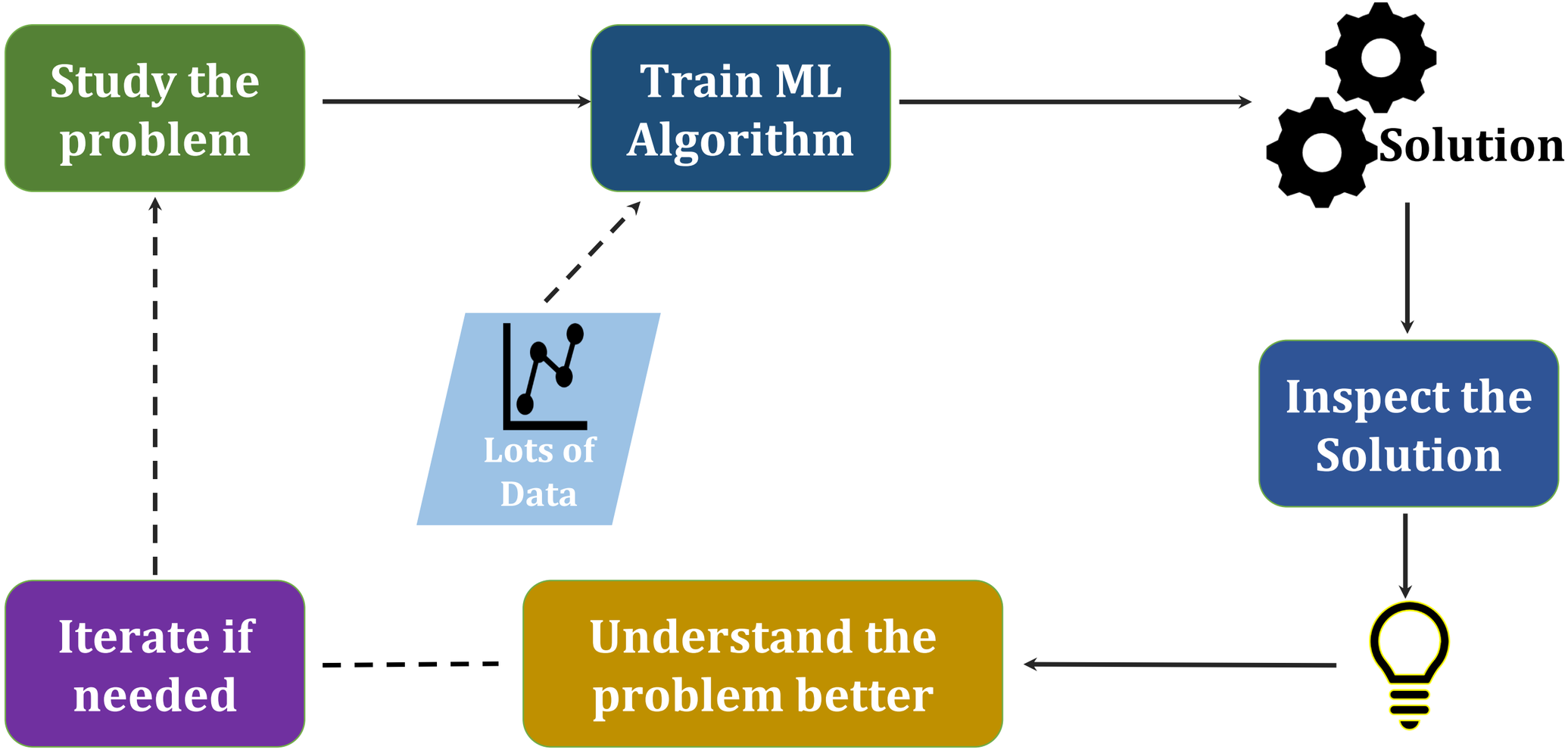
Intelligence – Spam Filter – ML Approach



Source: Introduction to Machine Learning (CloudX Lab)

Intelligence – Spam Filter – ML Approach

Can help humans learn

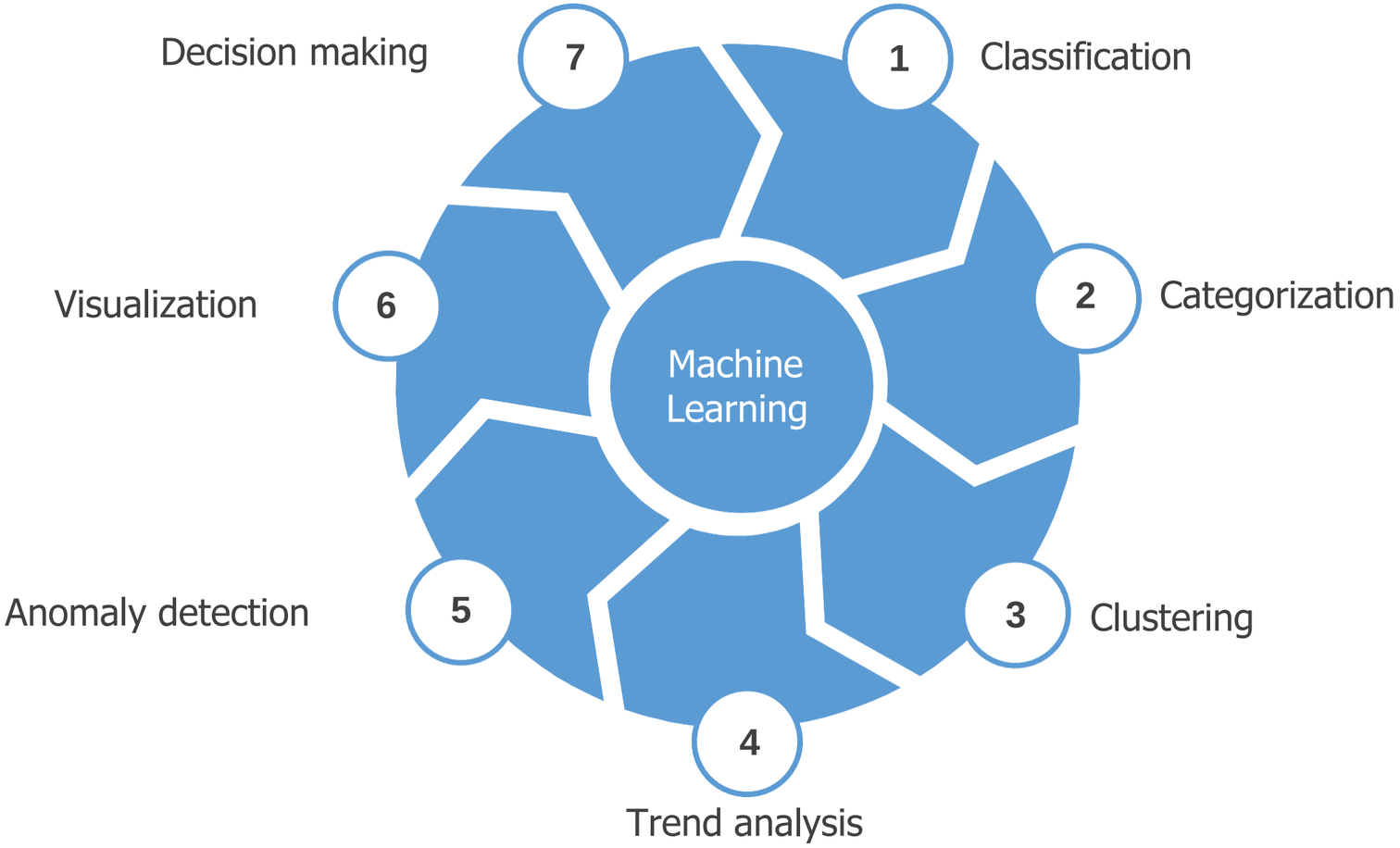


Source: Introduction to Machine Learning (CloudX Lab)

MACHINE LEARNING TECHNIQUES

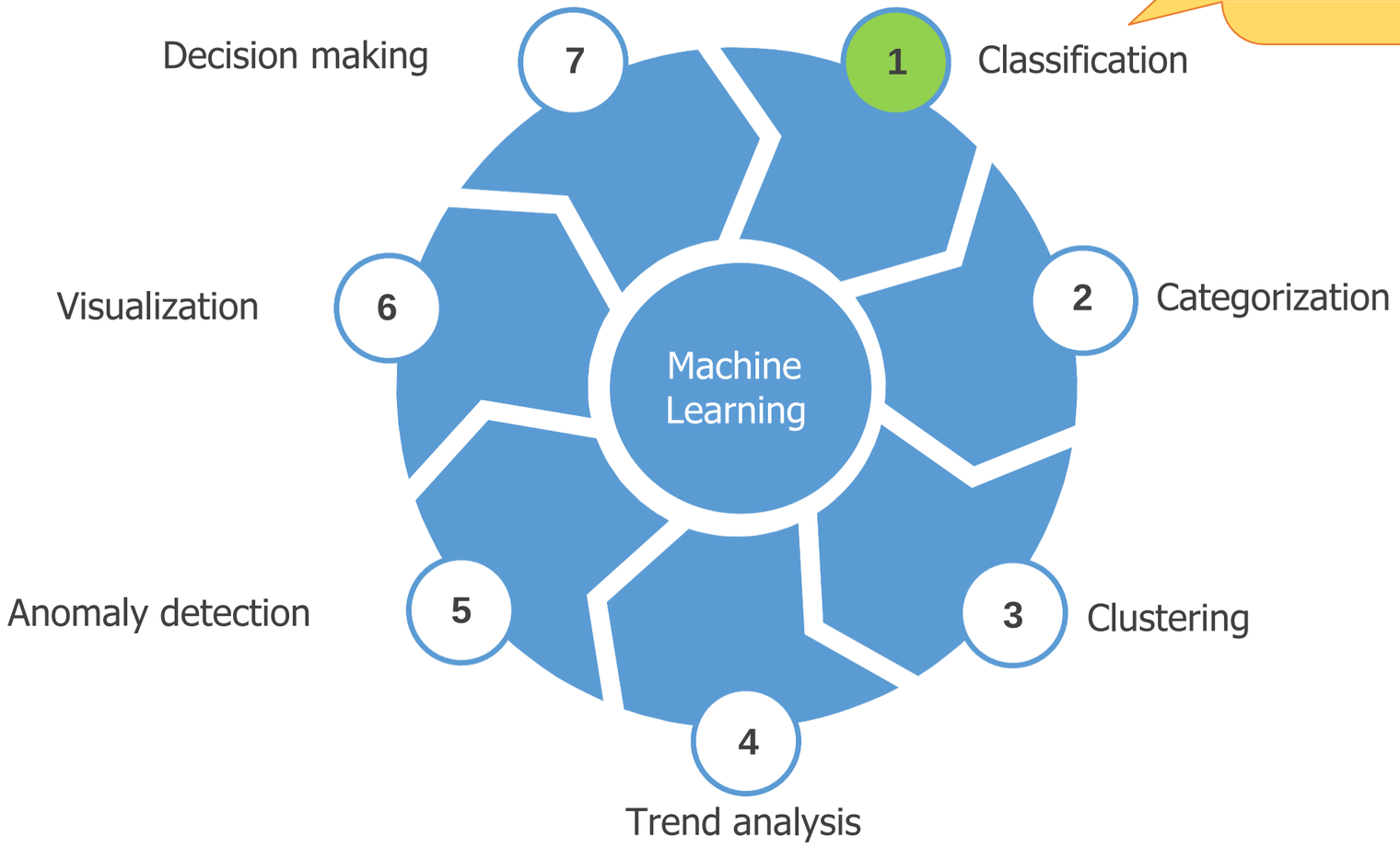
Machine Learning Techniques

Machine Learning uses a number of theories and techniques from Data Science:

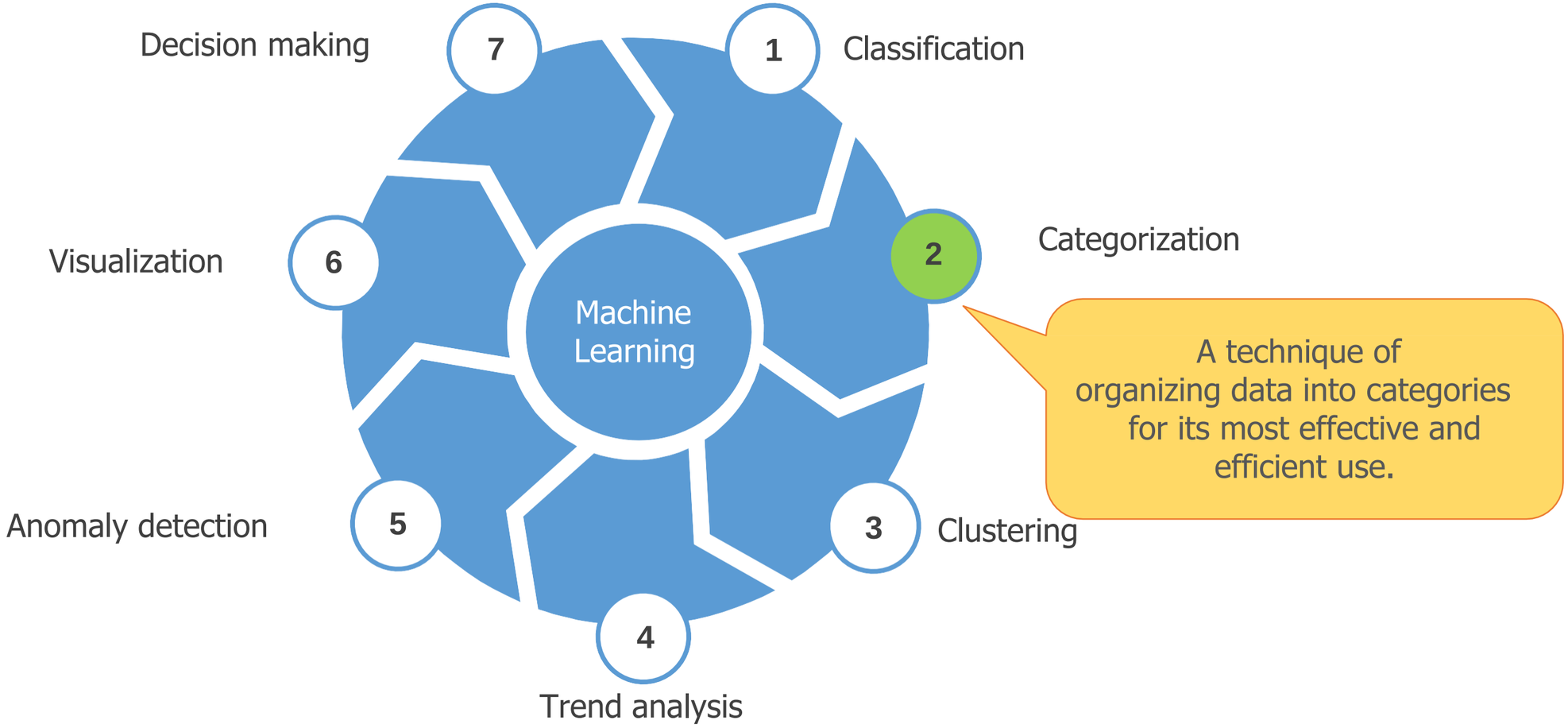


Machine Learning Techniques

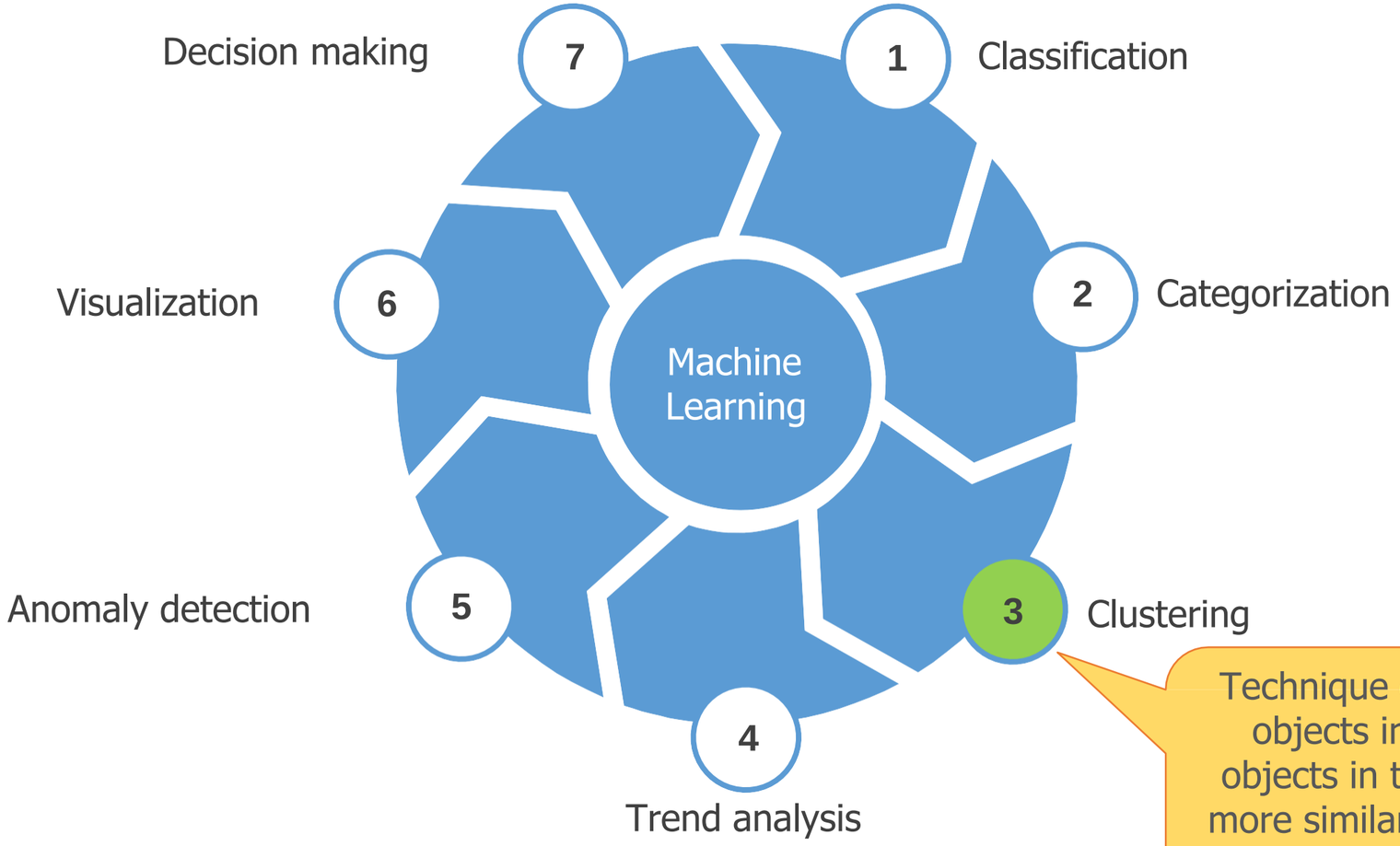
classification is a technique in which the computer program learns from the data input given to it and then uses this **learning** to **classify** new observation



Machine Learning Techniques

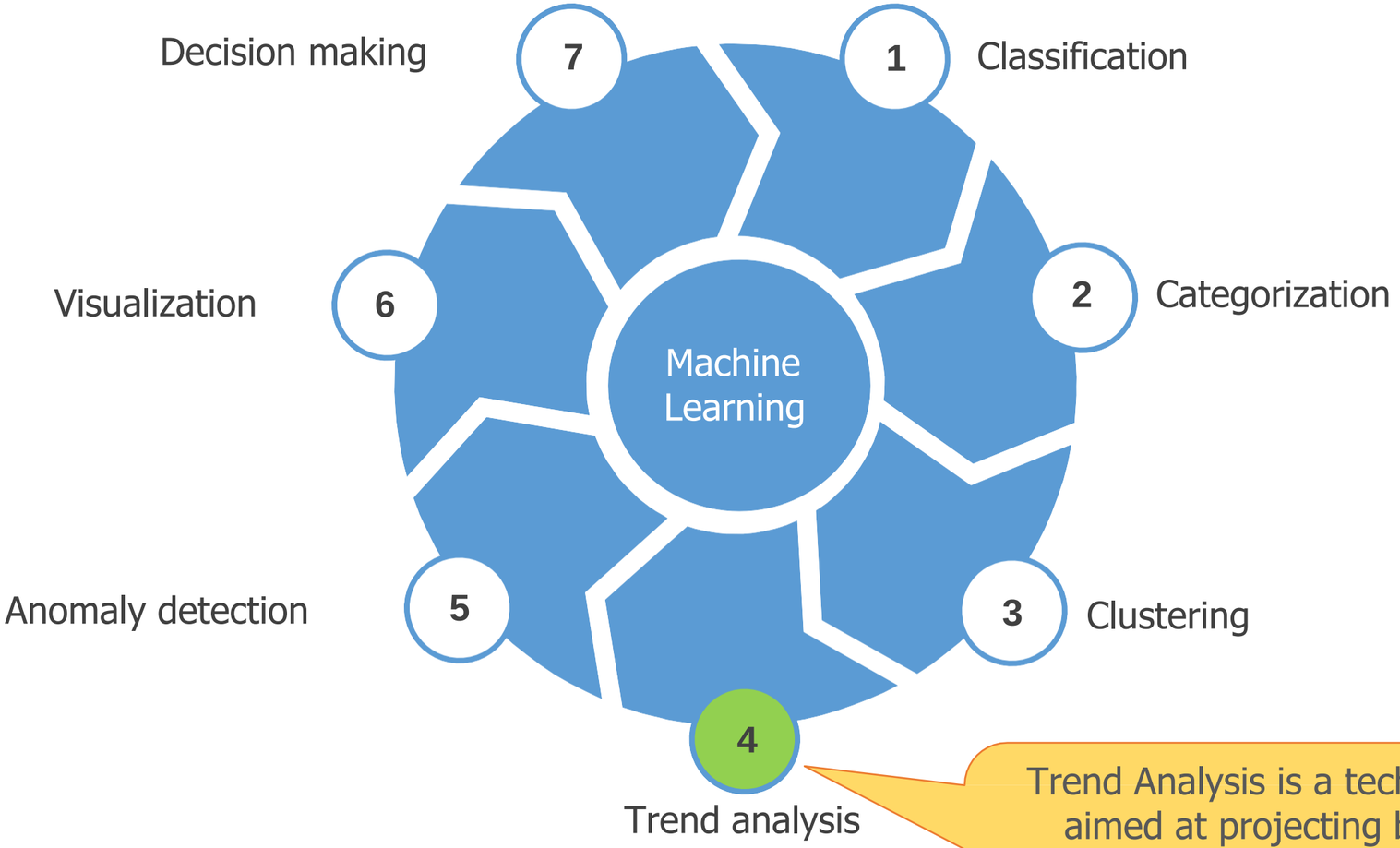


Machine Learning Techniques



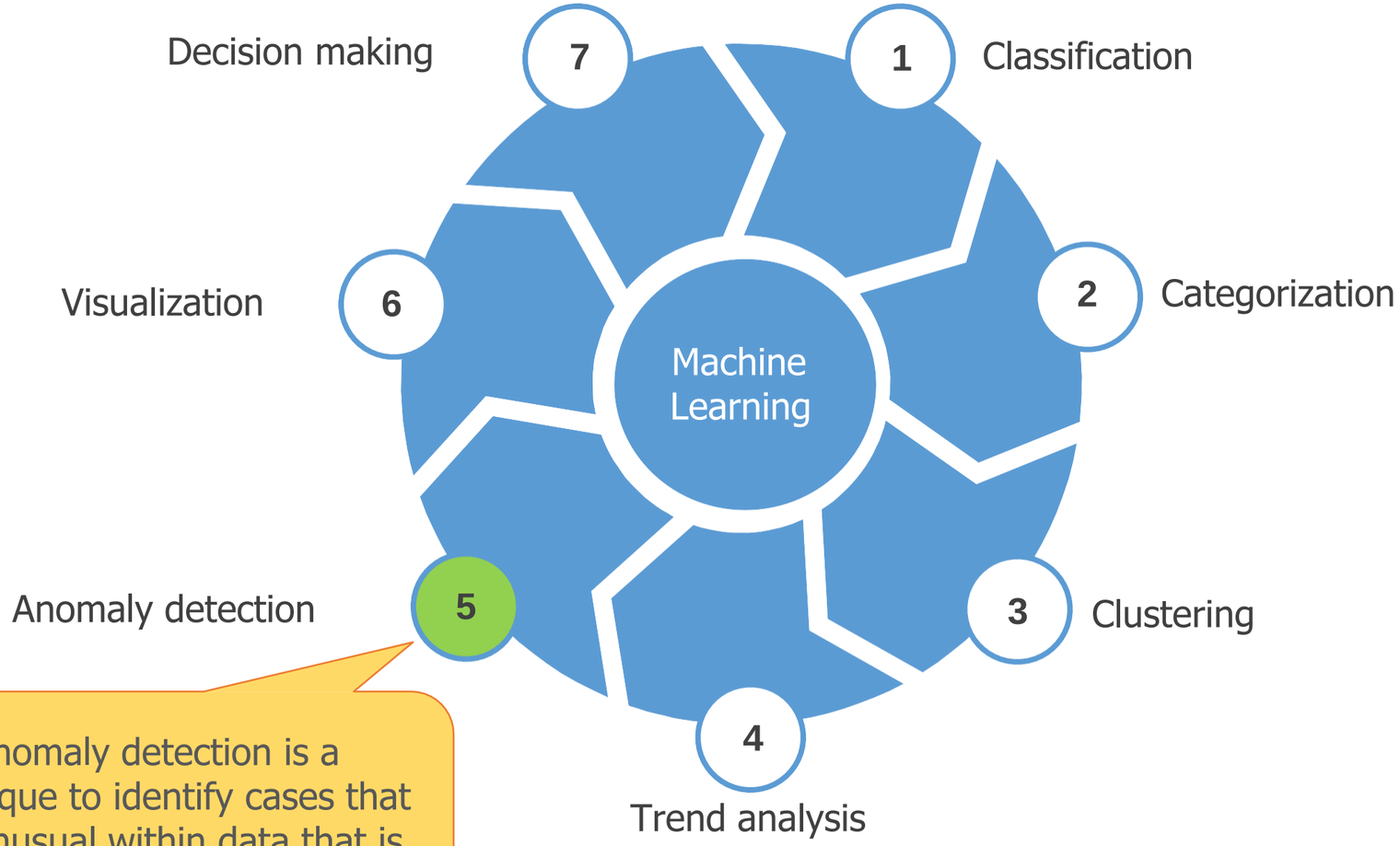
Technique of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups

Machine Learning Techniques



Trend Analysis is a technique aimed at projecting both current and future movement of events through use of time series data analysis

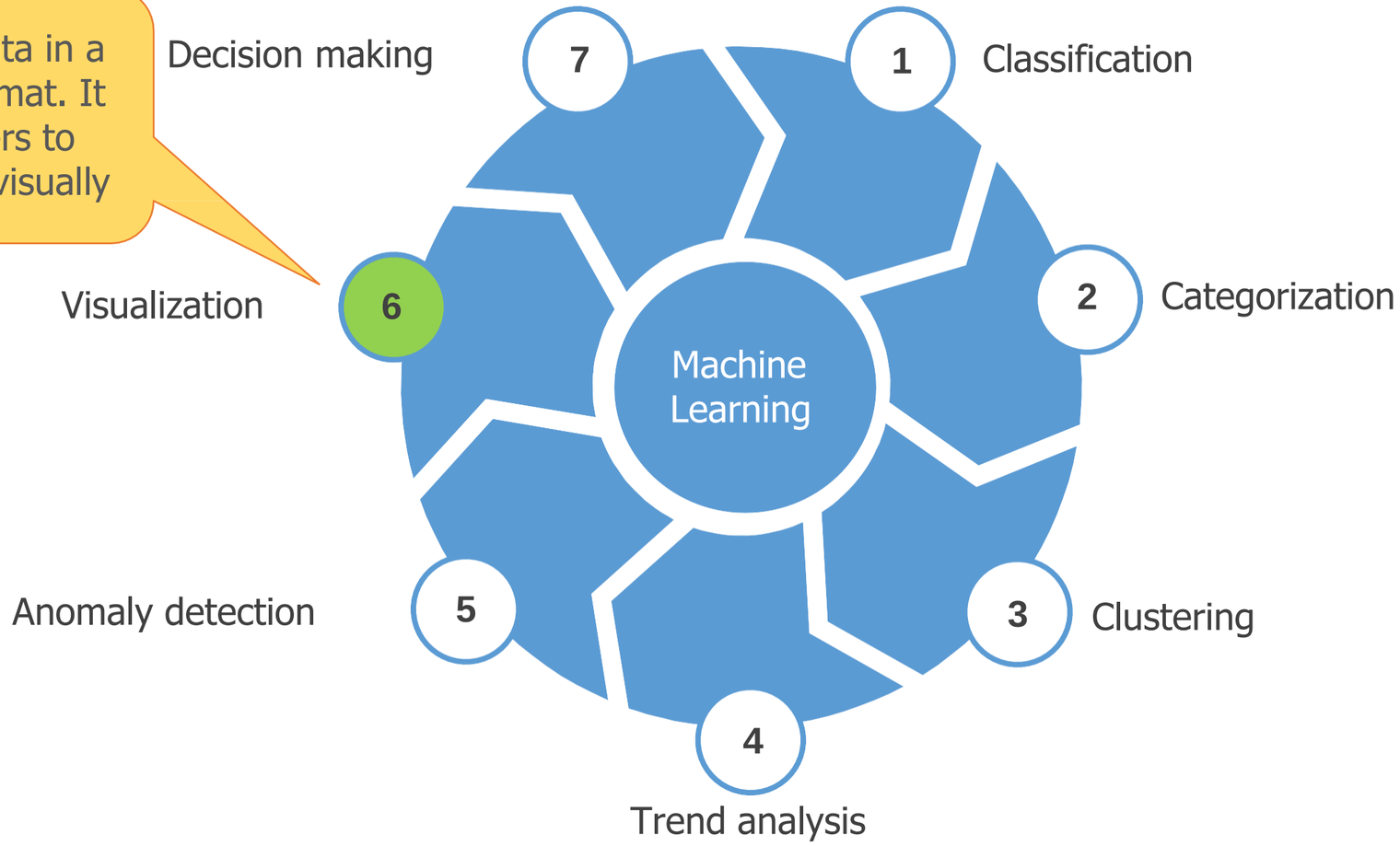
Machine Learning Techniques



Anomaly detection is a technique to identify cases that are unusual within data that is seemingly homogeneous

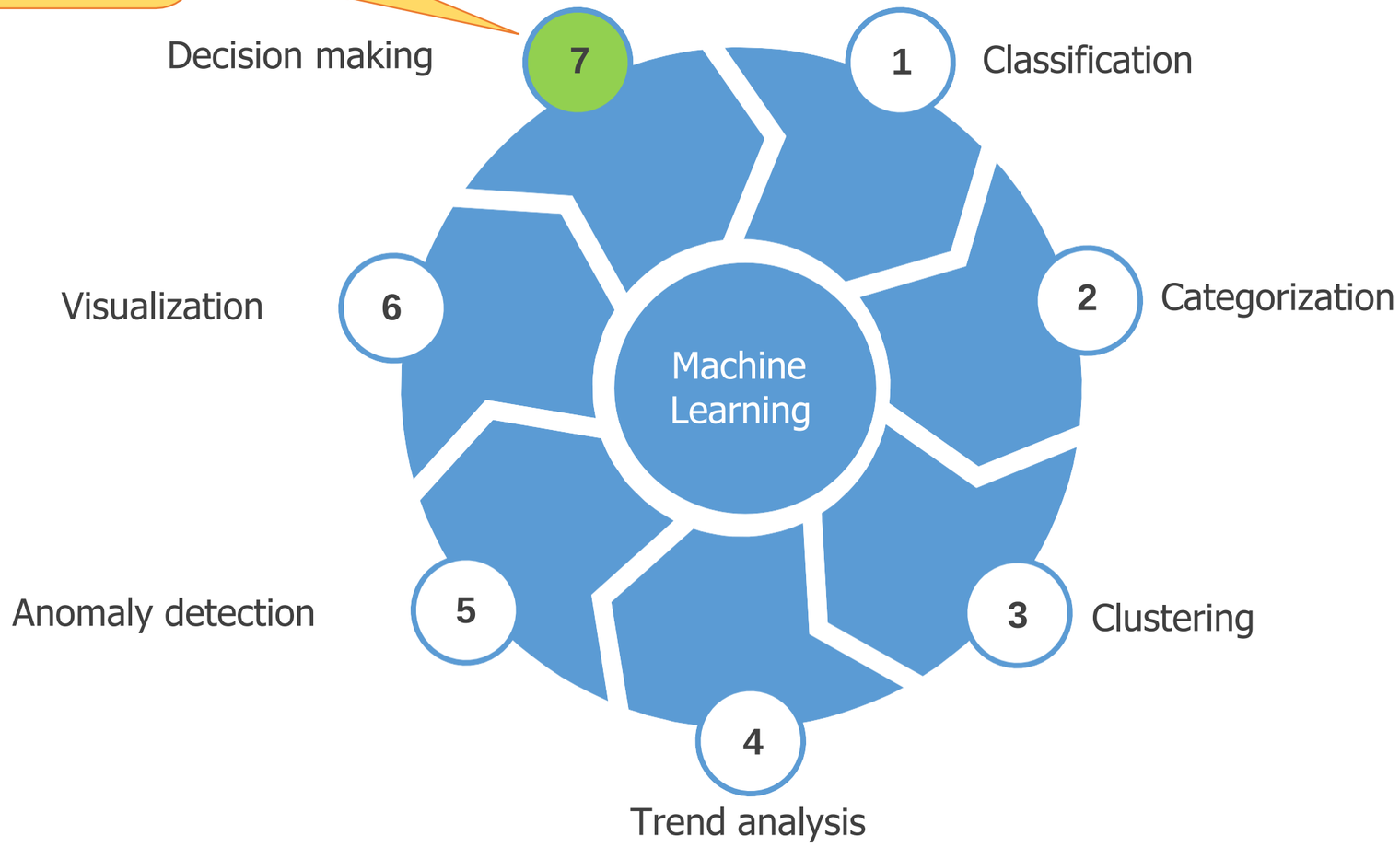
Machine Learning Techniques

Technique to present data in a pictorial or graphical format. It enables decision makers to see analytics presented visually

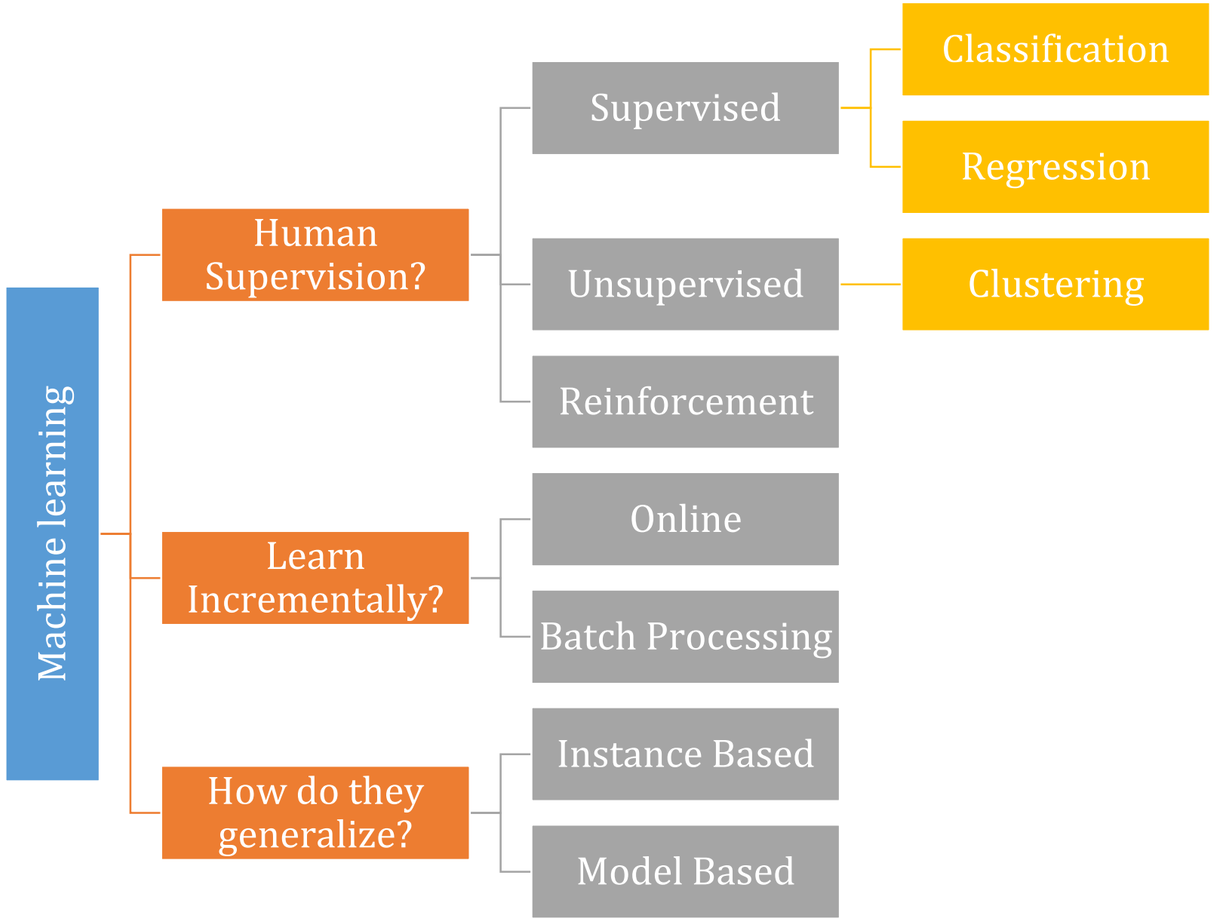


Machine Learning Techniques

A technique/skill which provides you with the ability to influence managerial decisions with data as evidence for those possibilities



Machine learning categories



Training Data

Food Item	Category	Fresh/Processed	Plant/Animal-Based	Healthy/Junk	Solid/Liquid
Apple	Fruit	Fresh	Plant-Based	Healthy	Solid
Rice	Grain	Fresh	Plant-Based	Healthy	Solid
Cheese	Dairy	Processed	Animal-Based	Healthy	Solid
Pizza	Fast Food	Processed	Mixed (Plant+Animal)	Junk	Solid
Fish	Protein	Fresh	Animal-Based	Healthy	Solid

Pre-defined rules

1. **Fruits** → Must be **fresh**, **plant-based**, and **solid** ☑ Apple ✕ Orange Juice (liquid, doesn't fit)
2. **Grains** → Must be **fresh**, **plant-based**, and **solid** ☑ Rice ✕ Bread (not in initial training, would be hard to classify)
3. **Dairy** → Must be **processed** and **animal-based** ☑ Cheese ✕ Almond Milk (plant-based)
4. **Fast Food** → Must be **processed** and **mixed (plant + animal-based)** ☑ Pizza ✕ Sushi (fresh, doesn't fit)
5. **Protein** → Must be **fresh** and **animal-based** ☑ Fish ✕ Peanut Butter (plant-based)

Machine Learning - Batch Learning

Offline learning

System is incapable of learning incrementally

- It must be trained offline using all the available data
- Takes lot of time and computing resources
- everytime training happens on the entire data

Once the system is trained, it gets

- Pushed to production
- Runs without learning anymore
- Just applies what it has learned offline

Source: Introduction to Machine Learning (CloudX Lab)

Challenges in batch learning

We have to train the new system

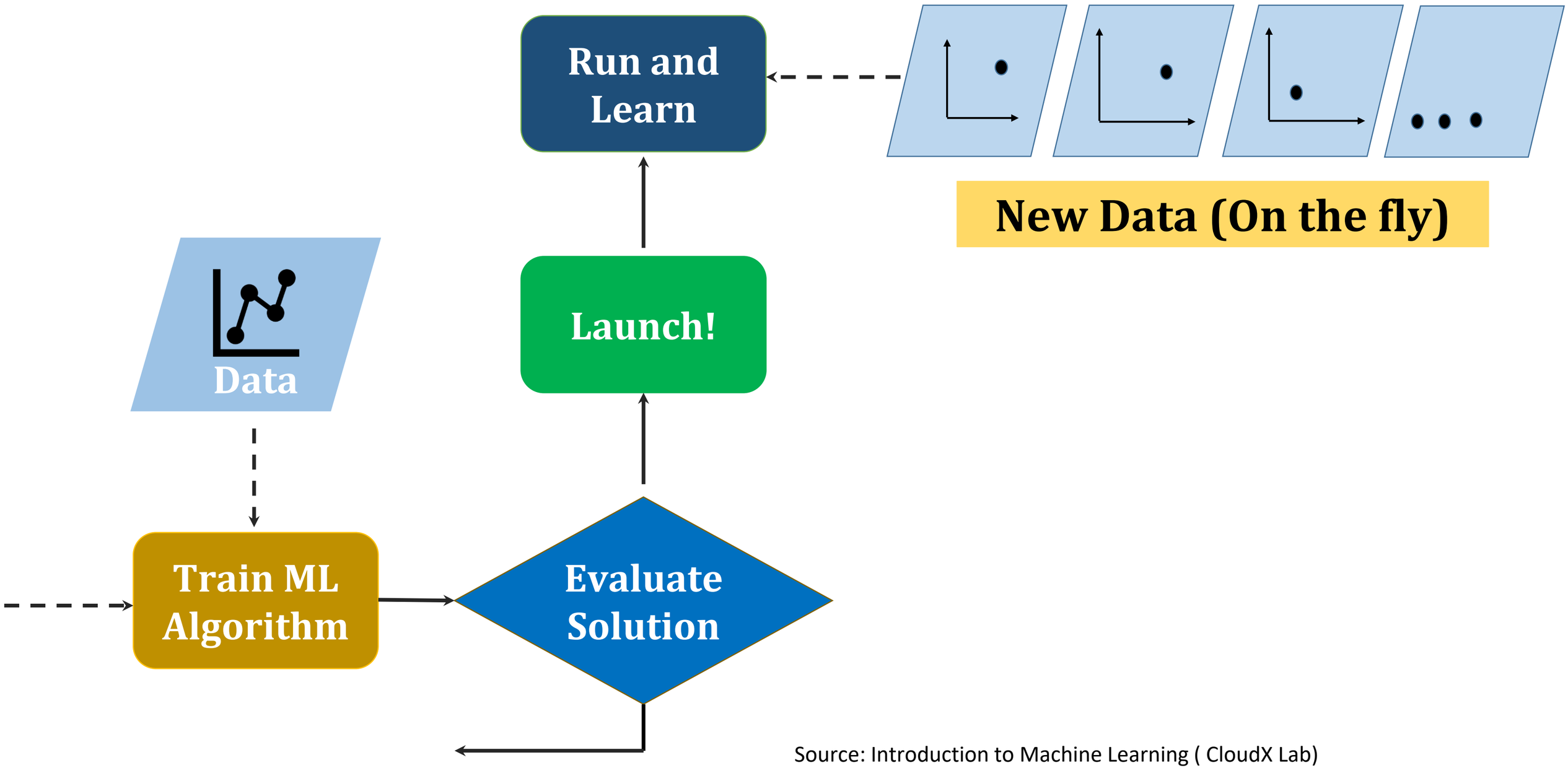
- From scratch on the full dataset
- Stop the old system and
- Replace it with the newly trained system

Limitations

- Since we have to train new system every time new data is available
- Takes lot of computing resources
- If the data is huge then
- Training from scratch is nearly impossible
- Becomes showstopper for lot of practical use cases

Source: Introduction to Machine Learning (CloudX Lab)

Machine Learning - Online Learning



Source: Introduction to Machine Learning (CloudX Lab)

Machine Learning - Online Learning

Train system incrementally

- By feeding new data sequentially
- Or in batches

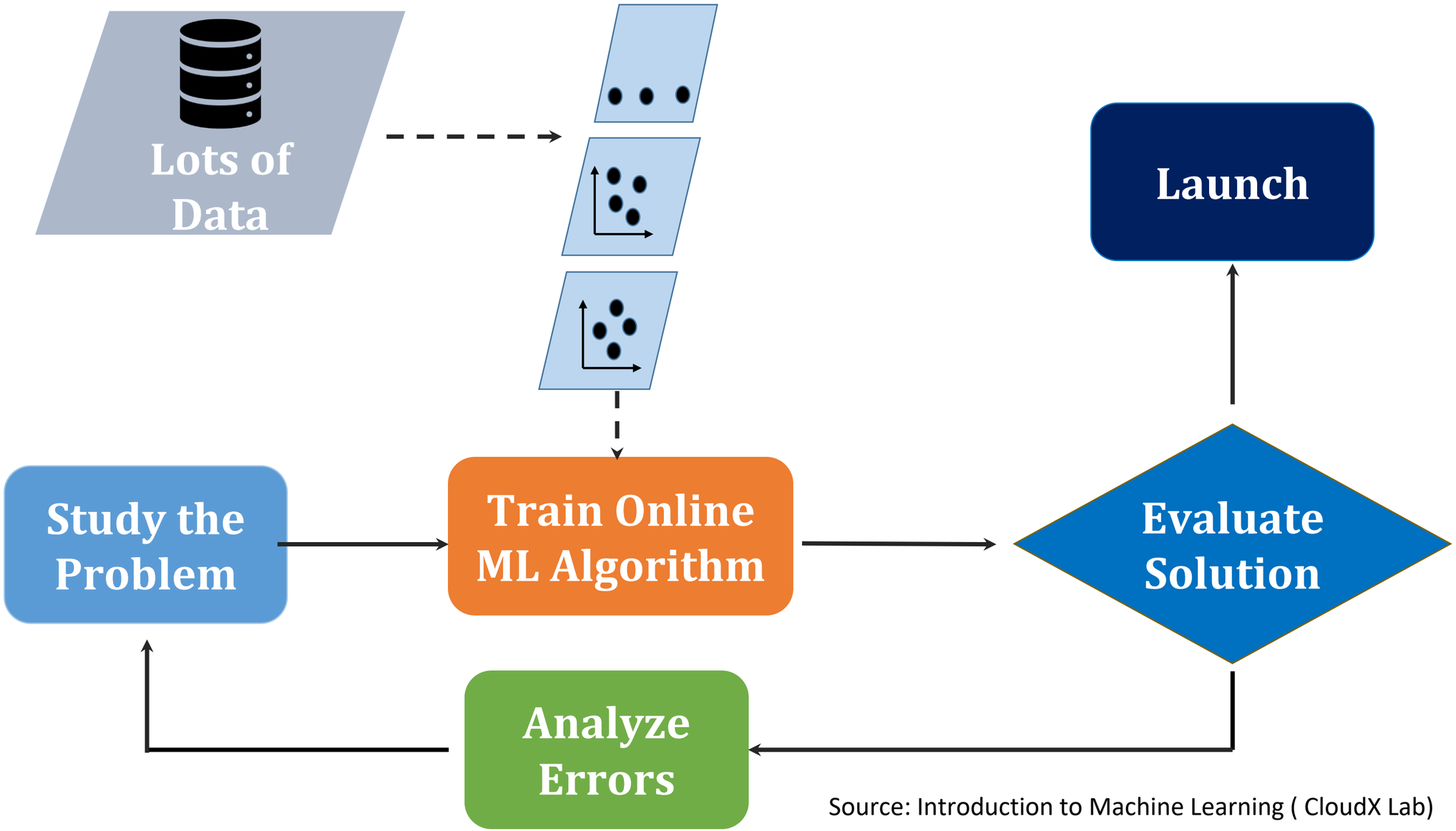
System can learn from new data on the fly

Good for systems where data is a continuous flow

- Stock prices

Source: Introduction to Machine Learning (CloudX Lab)

Machine Learning – Online Learning



Source: Introduction to Machine Learning (CloudX Lab)

Using online learning to handle huge datasets

Can be used to train huge datasets

- That can not be fit in one machine
- The training data gets divided into batches and
- System gets trained on each batch incrementally

Source: Introduction to Machine Learning (CloudX Lab)

Challenges in online learning

System's performance gradually declines

- If bad data is fed to the system
- Bad data can come from
- Malfunctioning sensor or robot
- Someone spamming your system

Closely monitor the system

- Turn off the learning if there is a performance drop
- Or monitor the input data and remove anomalies

Source: Introduction to Machine Learning (CloudX Lab)

Machine Learning - Instance & Model-Based Learning

Most Machine Learning tasks are about making predictions

Having a good performance in training data is good

But true goal is to perform well on new instances

System needs to be able to generalize to examples it has never seen before

Source: Introduction to Machine Learning (CloudX Lab)

Instance Based and Model Based Learning

Lets hear a Story

John vs Joseph

Generalisation (Model-based Learning)

Memorisation (Instance-based Learning)

John-Maths

Joseph- Science

- John explained all the in-depth concepts to Kevin and taught him about various scenarios and approaches for solving different kinds of problems.
- He also trained Kevin to solve many sample problems and made him understand topics and questions with similar contents and weightage in the examination.

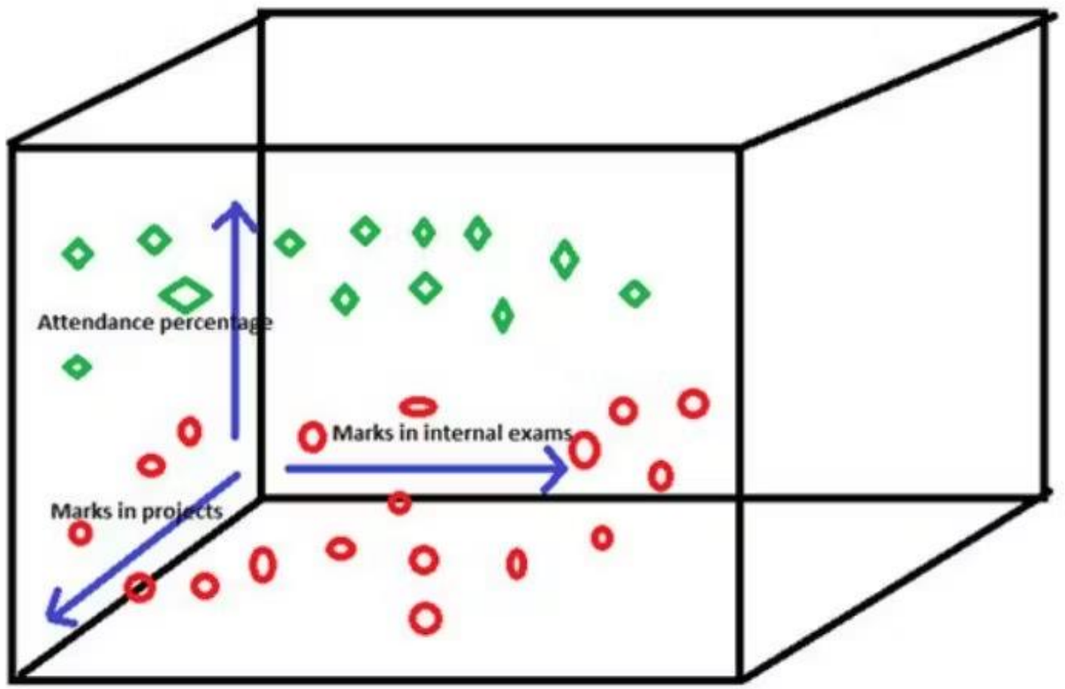
- Joseph asked Kevin if he wanted to understand the subject deeply or just memorize important questions for the exam.
- Being lazy, Kevin chose the easier way and asked for the questions.
- Joseph handed him a list of 50 important questions and told him to memorize them.

Example

We have a classification problem to predict whether a student will pass or fail the examination. We have the following features as independent variables-

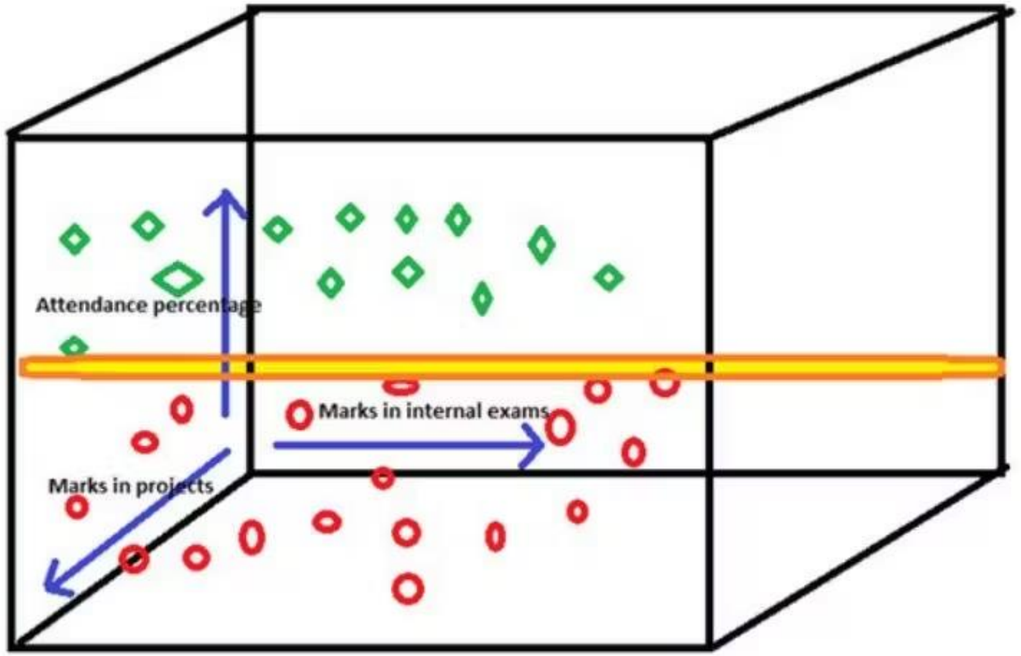
- Marks in internal exams
- Marks in projects
- Attendance percentage

So, these 3 independent variables become dimensions of a space like this-

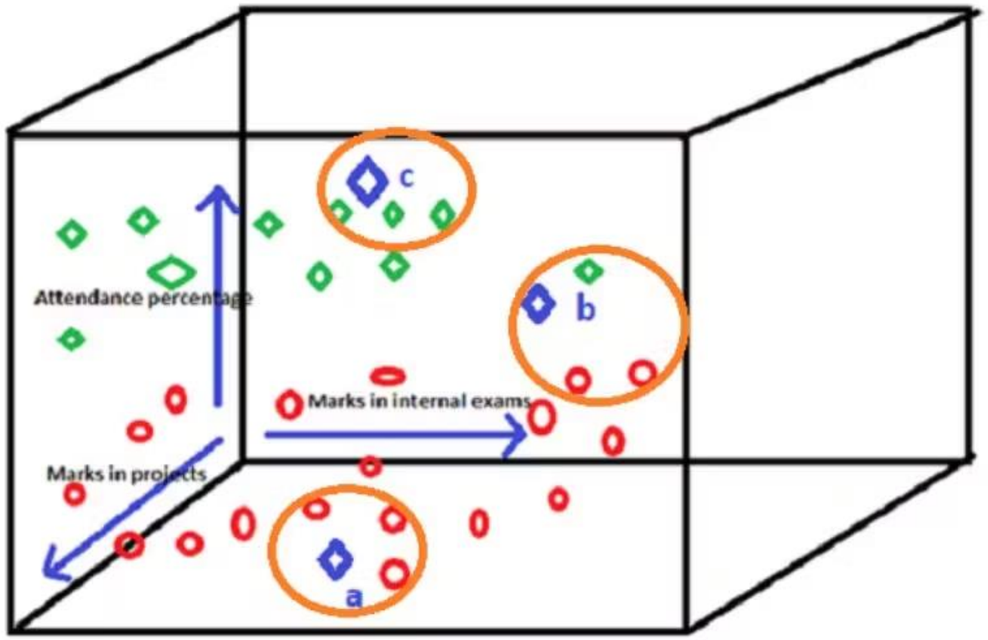


Original Source: https://hackernoon.com/the-notions-behind-model-based-and-instance-based-learning-in-ai-and-ml

Model- Based
Example: Support Vector Machine



Instance-Based
Example: K-Nearest Neighbour



Model-based vs Instance based

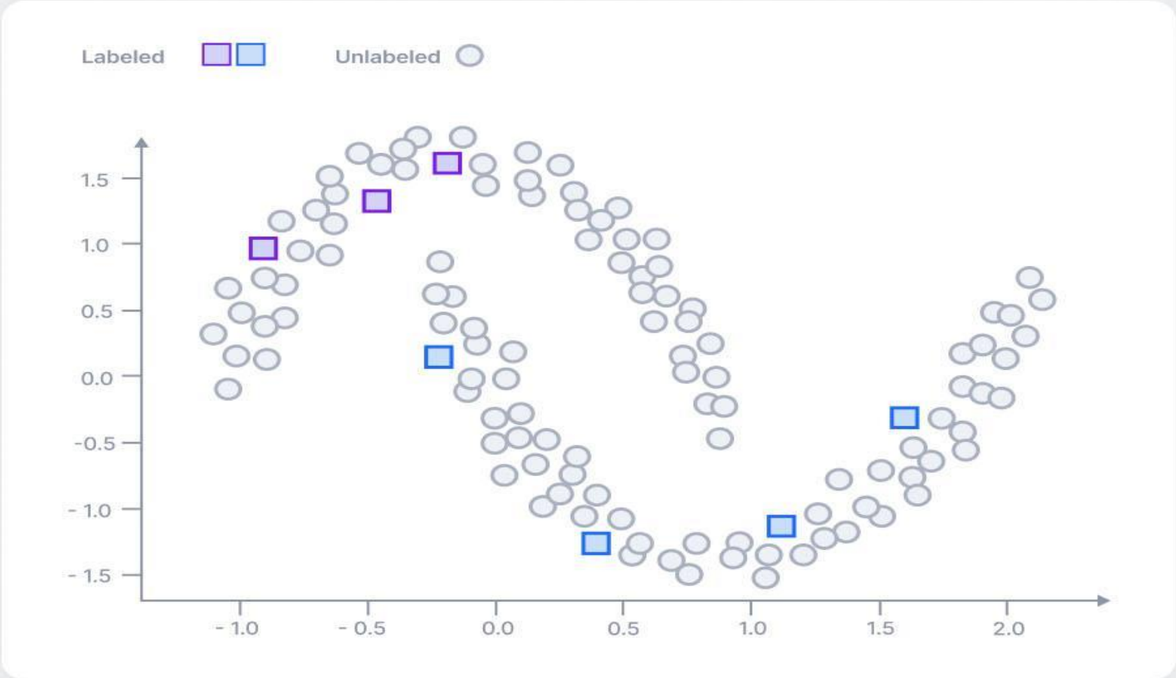
Aspect	Model-Based Learning	Instance-Based Learning
Training Approach	Learns patterns from the entire dataset and optimizes parameters.	Doesn't train the full dataset; predicts based on specific cases using predefined rules.
Data Retention	Training data can be removed after learning.	Training data must be retained for predictions.
Prediction Speed	Fast, as the model applies learned patterns directly.	Slow, as each new input is compared with stored data (lazy learning).
Feature Importance	Understands feature importance through optimization.	Can be misled by irrelevant features.
Handling Noisy Data	Struggles with noisy data; outliers are removed, potentially losing information.	Handles noisy data well, preserving all information.

SEMI-SUPERVISED MACHINE LEARNING

Original Source:

https://www.v7labs.com/blog/semi-supervised-learning-guide

Consider the Moons dataset

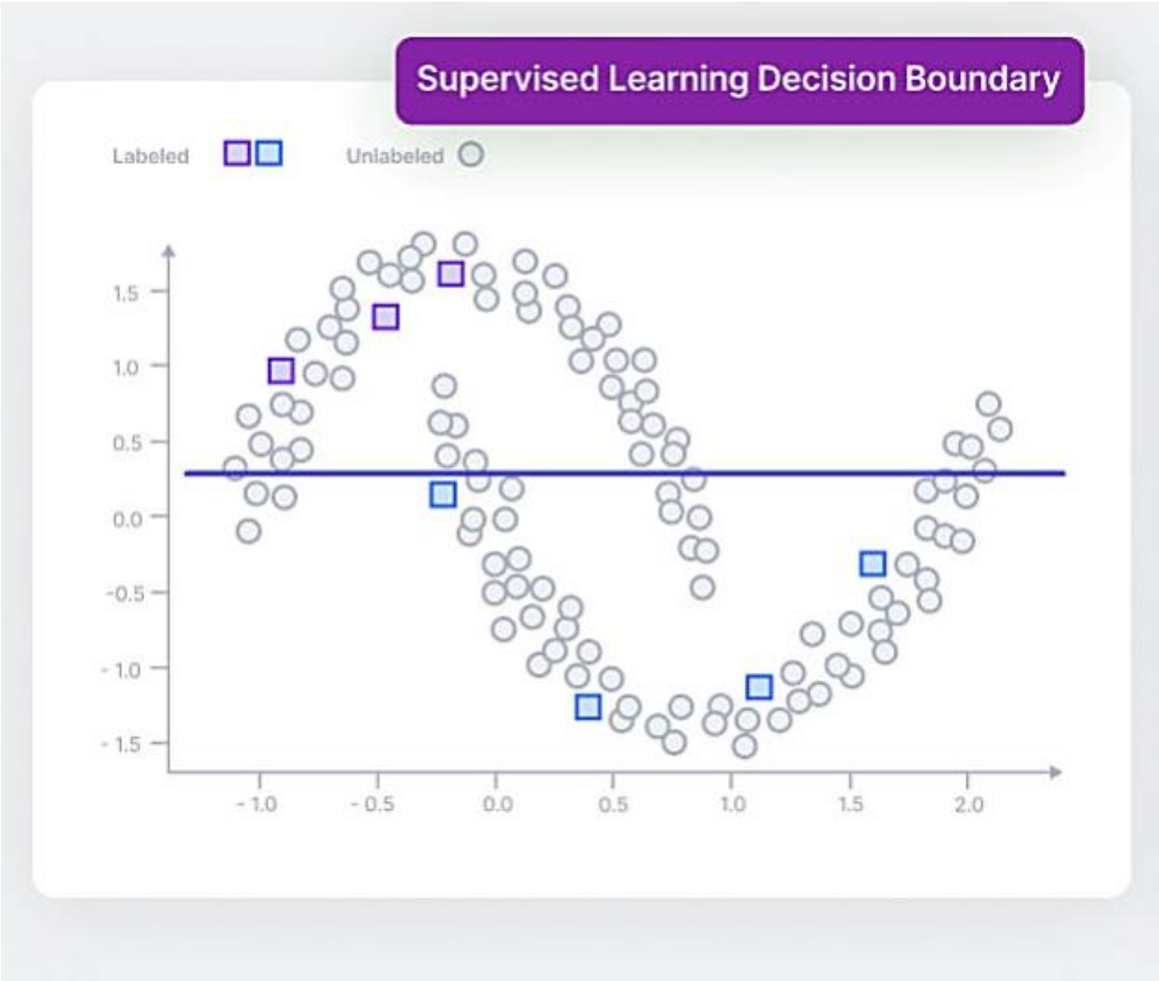


Moons dataset example for a binary classification problem

Source: https://www.v7labs.com/blog/semi-supervised-learning-guide

Supervised Approach?

- Supervised training updates model weights to minimize the average difference between predictions and labels.
- However, with limited labeled data, this might find a decision boundary that is valid for the labeled points but won't generalize to the whole distribution as in Figure.

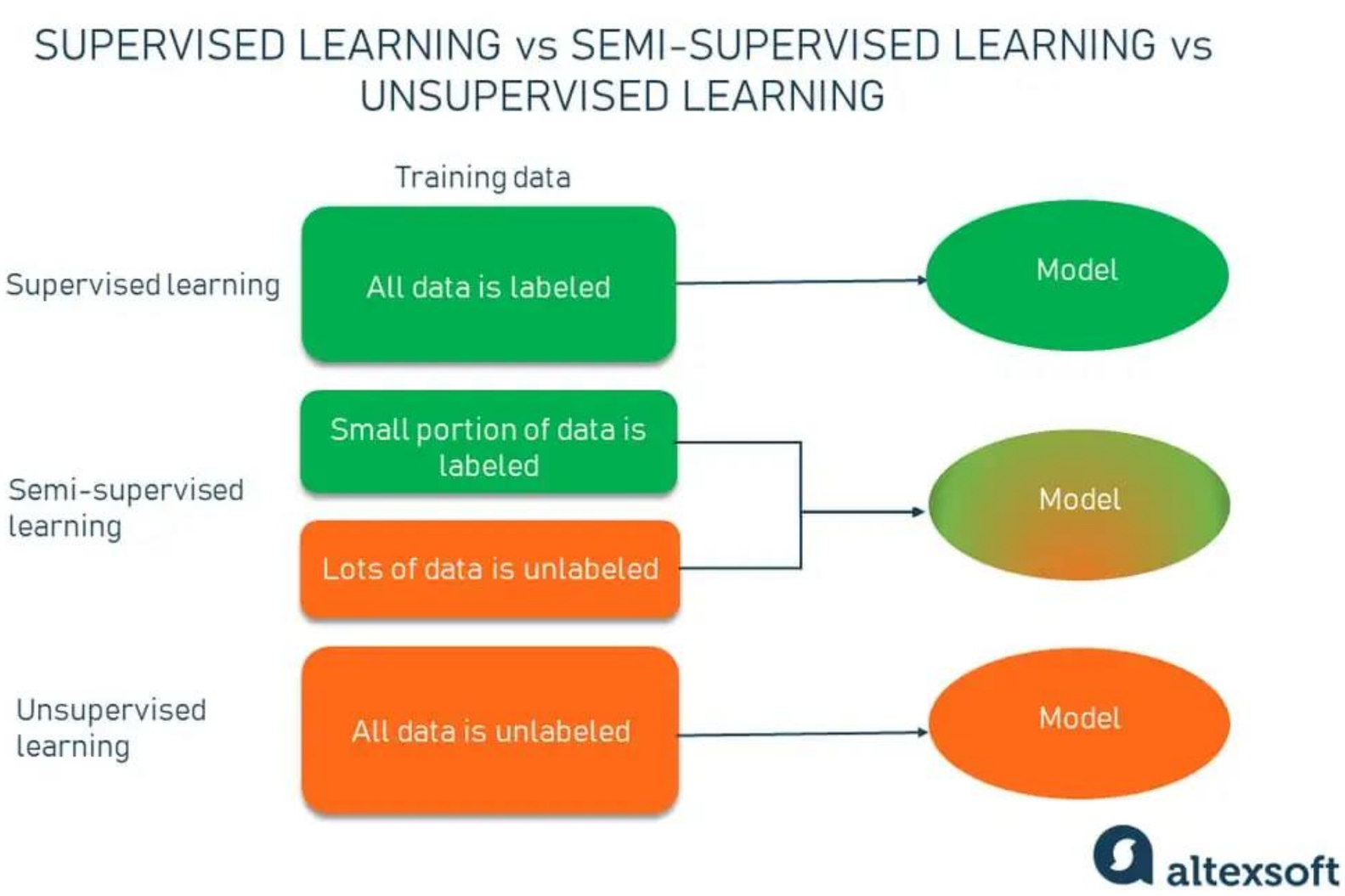


Source: https://www.v7labs.com/blog/semi-supervised-learning-guide

Un-Supervised Approach?

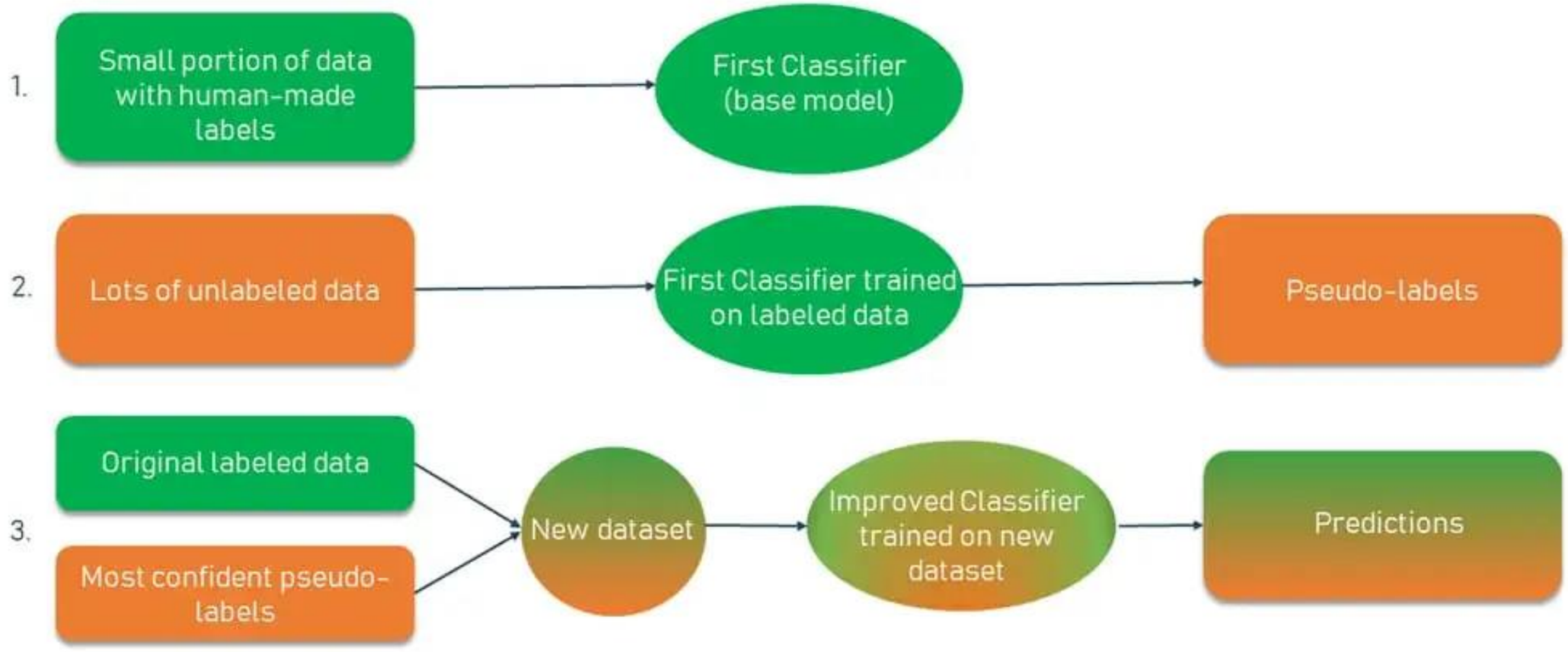
- Unsupervised learning, on the other hand, tries to cluster points together based on similarities in some feature-space.
- But, without labels to guide training, an unsupervised algorithm might find sub-optimal clusters.
- In Figure, for example, the discovered clusters incorrectly fit the true class distribution.

Source: https://www.v7labs.com/blog/semi-supervised-learning-guide



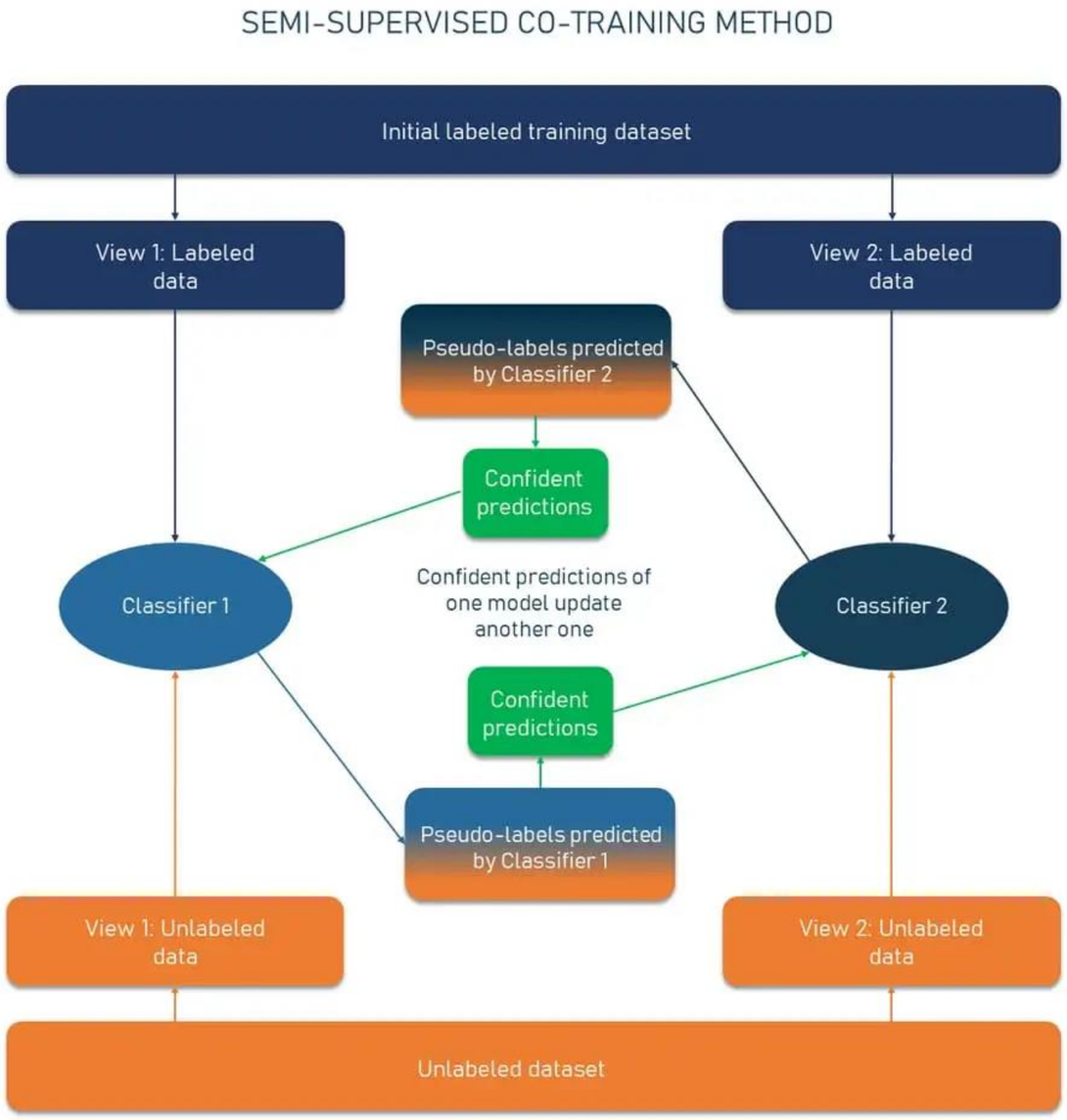
Source: https://www.altexsoft.com/blog/semi-supervised-learning/

SEMI-SUPERVISED SELF-TRAINING METHOD



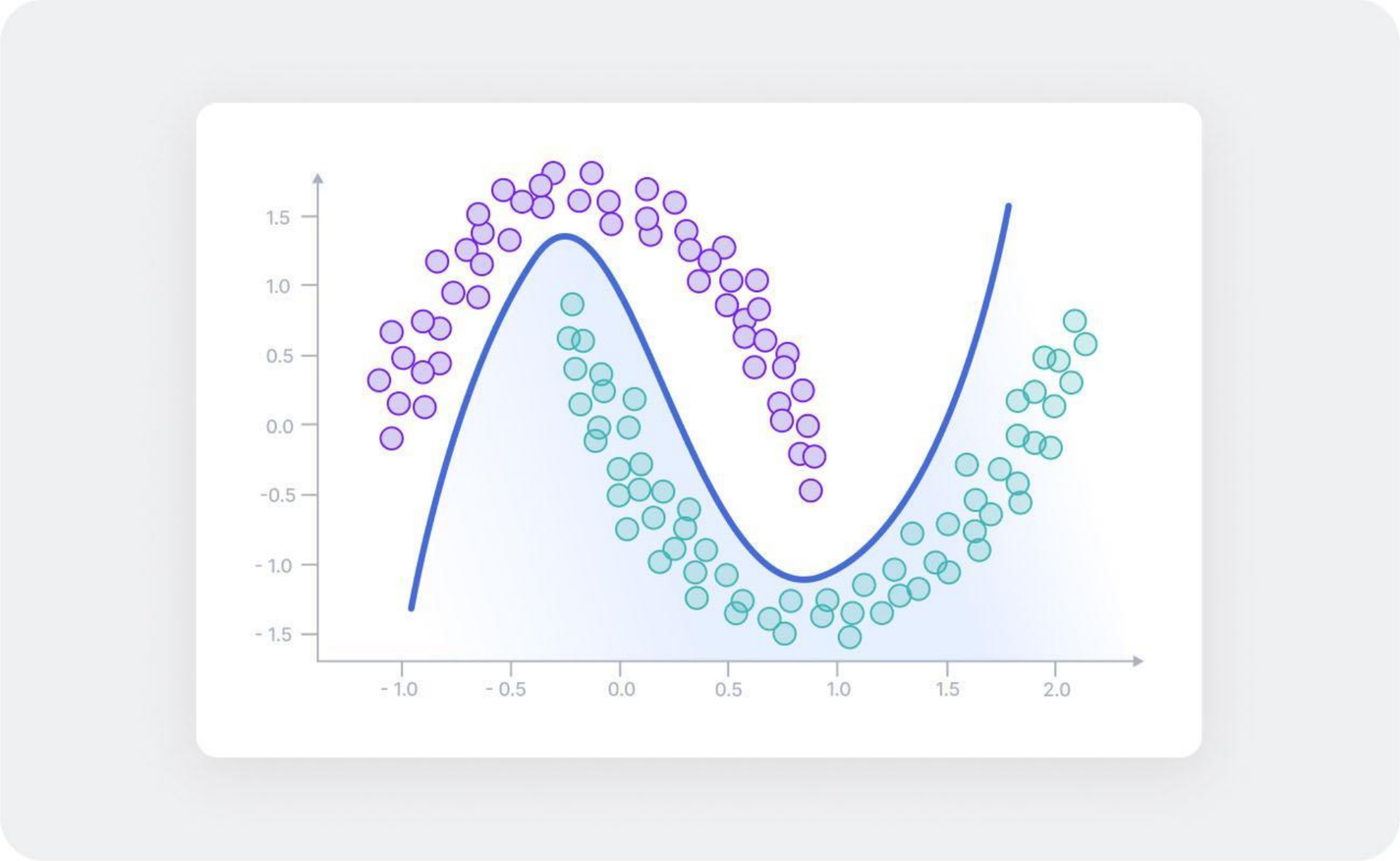
altexsoft

Source: https://www.altexsoft.com/blog/semi-supervised-learning/



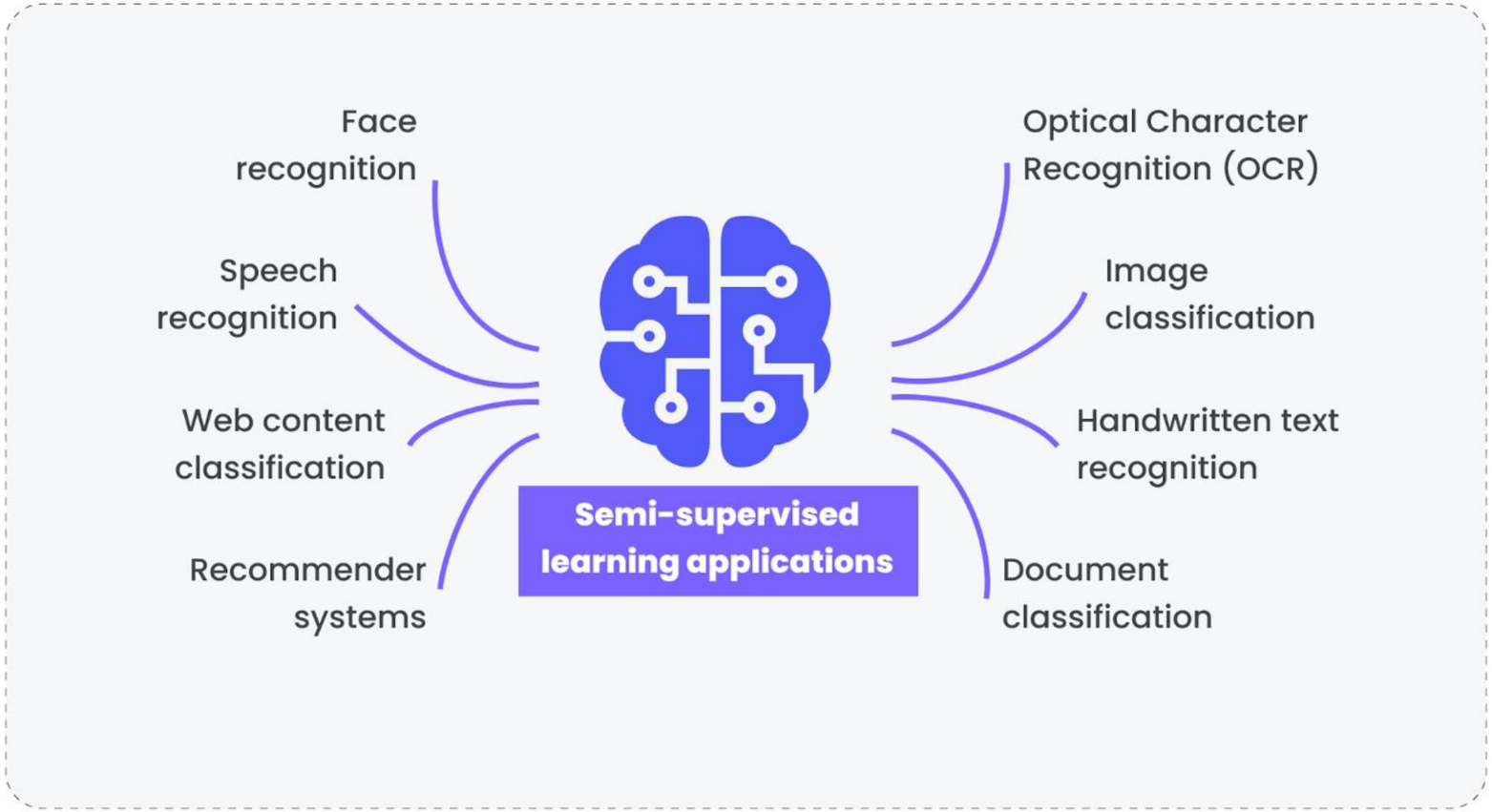
Source: https://www.altexsoft.com/blog/semi-supervised-learning/

Semi-Supervised Approach



Source: https://www.v7labs.com/blog/semi-supervised-learning-guide

Applications of Semi-Supervised Learning



https://maddevs.io/blog/semi-supervised-learning-explained/

Semi Supervised Learning

1. Companies like Google use semi-supervised learning for anomaly detection in network traffic.
2. Companies like PayPal use it for fraud detection in financial transactions.
3. Companies like Zebra Medical Vision use semi-supervised learning for symptom detection in medical diagnostics
4. Companies like Google DeepMind use it for protein structure prediction.
5. Companies like Boston Dynamics use it for robot navigation training.
6. Companies like Chevron analyze geological data to detect potential mineral or oil deposits.

CASE-STUDY

*Imagine a famous cooking school where students are training to become world-class chefs. The school has a **big competition** coming up, and the Head Chef, their mentor, must **train and evaluate** his students wisely.*

Phase-1: Training Set – Learning from Experience

- Head Chef picks **60% of the students** and teaches them **various cooking techniques**.
- They practice **chopping, frying, baking, and plating dishes** under his guidance.
- These students are **training models**, learning from labeled data (Head Chef 's expertise).

Phase-2: Validation Set – Fine-Tuning the Skills

- Now, Head Chef selects another **20% of students** and asks them to cook dishes **without his direct help**.
- He observes their cooking and **gives feedback** on seasoning, presentation, and timing.
- These students act as the **validation set**, helping Head Chef fine-tune his teaching style.

Phase 3: Test Set – The Final Challenge

- Finally, the last **20% of students** are chosen for the **ultimate test**: cooking a dish without any guidance.
- A panel of judges (external experts) **taste the dishes and give scores**.
- This is the **test set**, where performance is evaluated on unseen data.

TRAIN SET
VALIDATION SET
TEST SET

Source:

https://www.v7labs.com/blog/train-validation-test-set

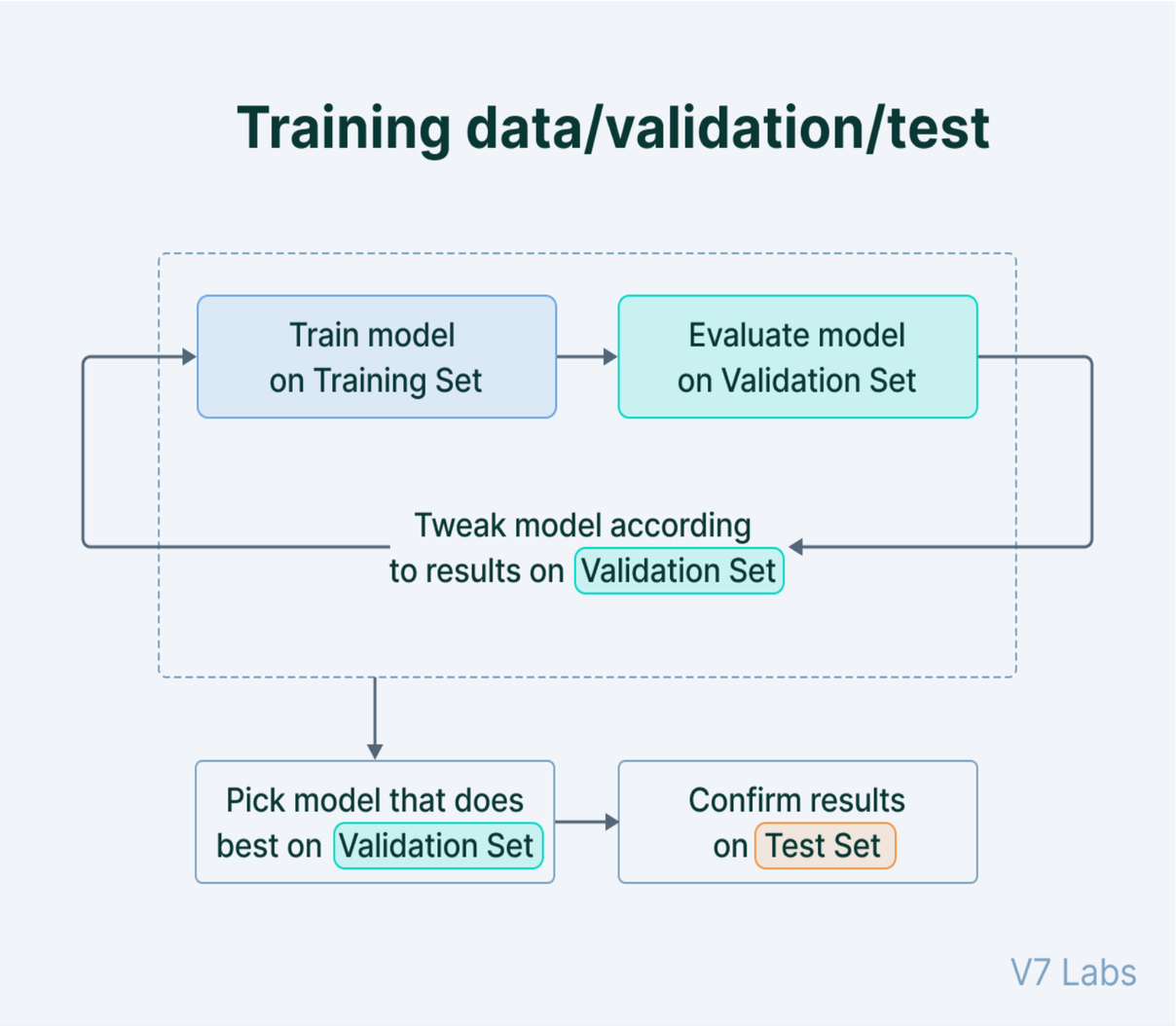
https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

First Rule of Machine Learning!

Don't use the same dataset for model training and model evaluation.

- If you want to build a *reliable* machine learning model, you need to split your dataset into the training, validation, and test sets.
- If you don't, your results will be biased, and you'll end up with a false impression of better model accuracy.

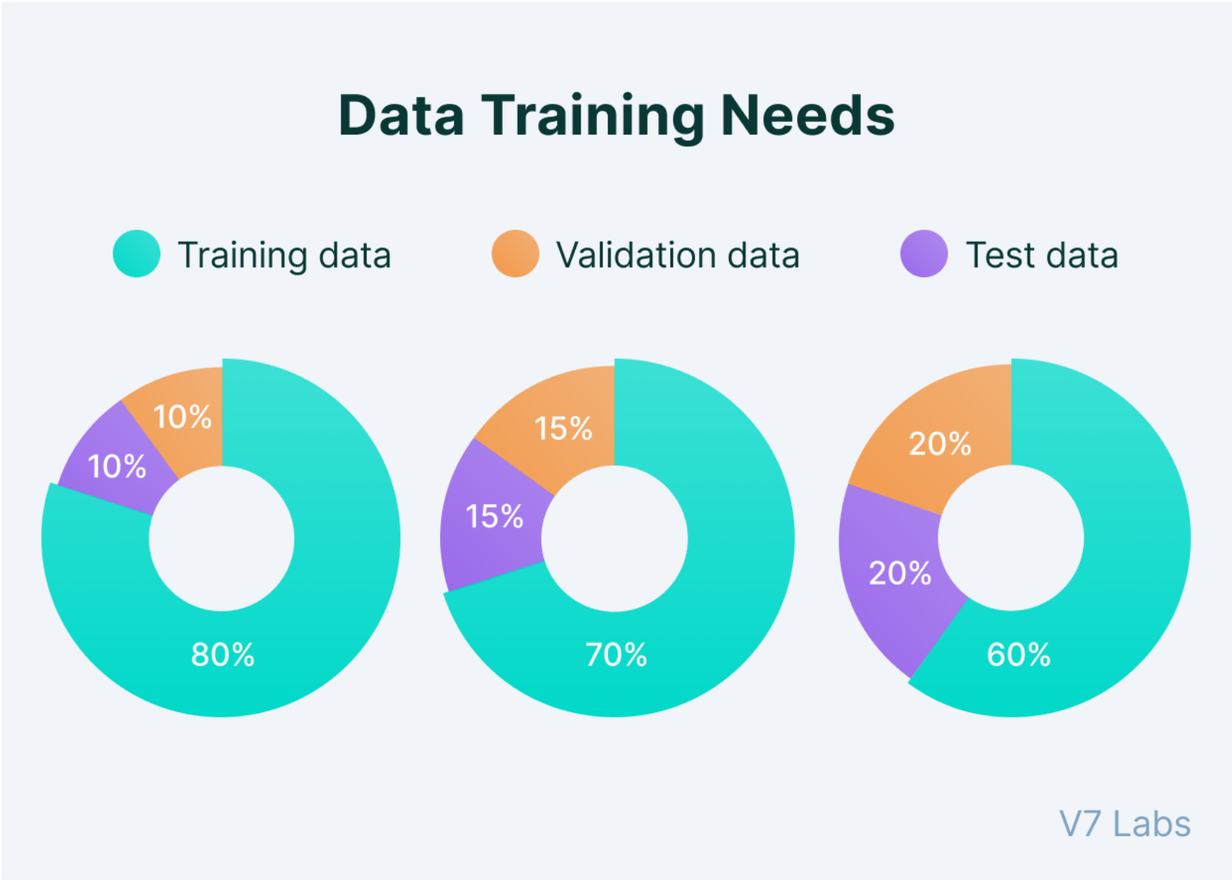
Train Set vs Validation Set vs Test Set



Source: https://www.v7labs.com/blog/train-validation-test-set

How to split your Machine Learning data?

- The creation of different samples and splits in the dataset helps us judge the true model performance.
- The dataset split ratio depends on the number of samples present in the dataset and the model.
- If there are several hyperparameters to tune, the machine learning model requires a larger validation set to optimize the model performance.
- There is no optimal split percentage. One has to come to a split percentage that suits the requirements and meets the model's needs.

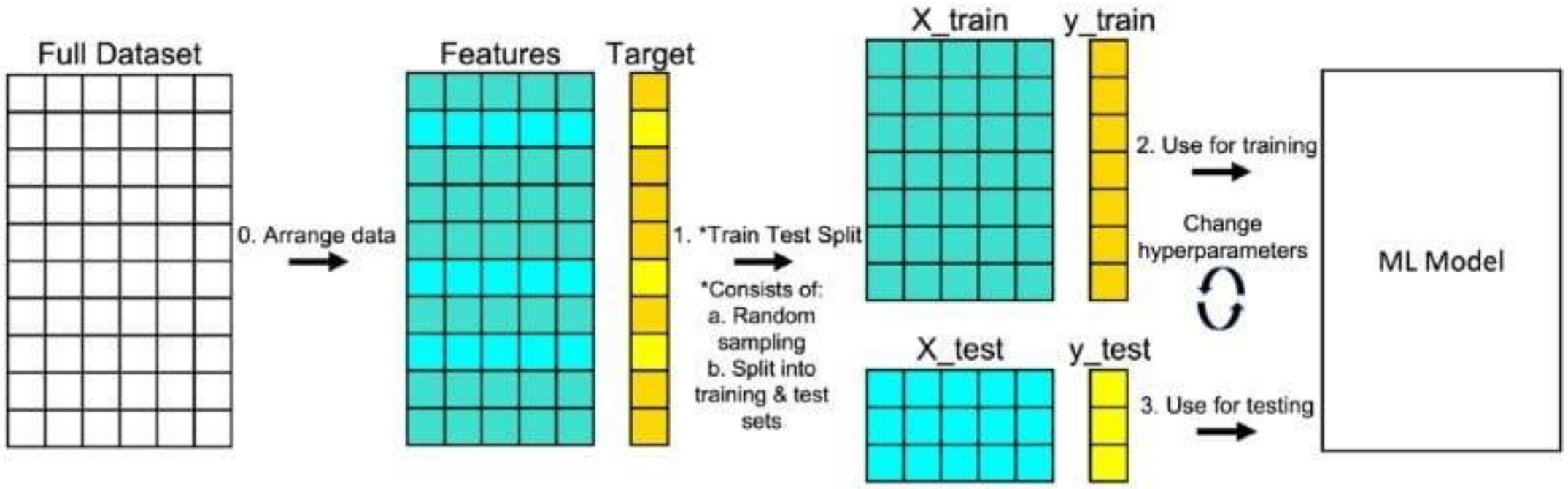


Advanced techniques for data splitting

1. Random
2. Stratify
3. K-fold cross validation
4. Stratified K-fold cross validation
5. Time series split

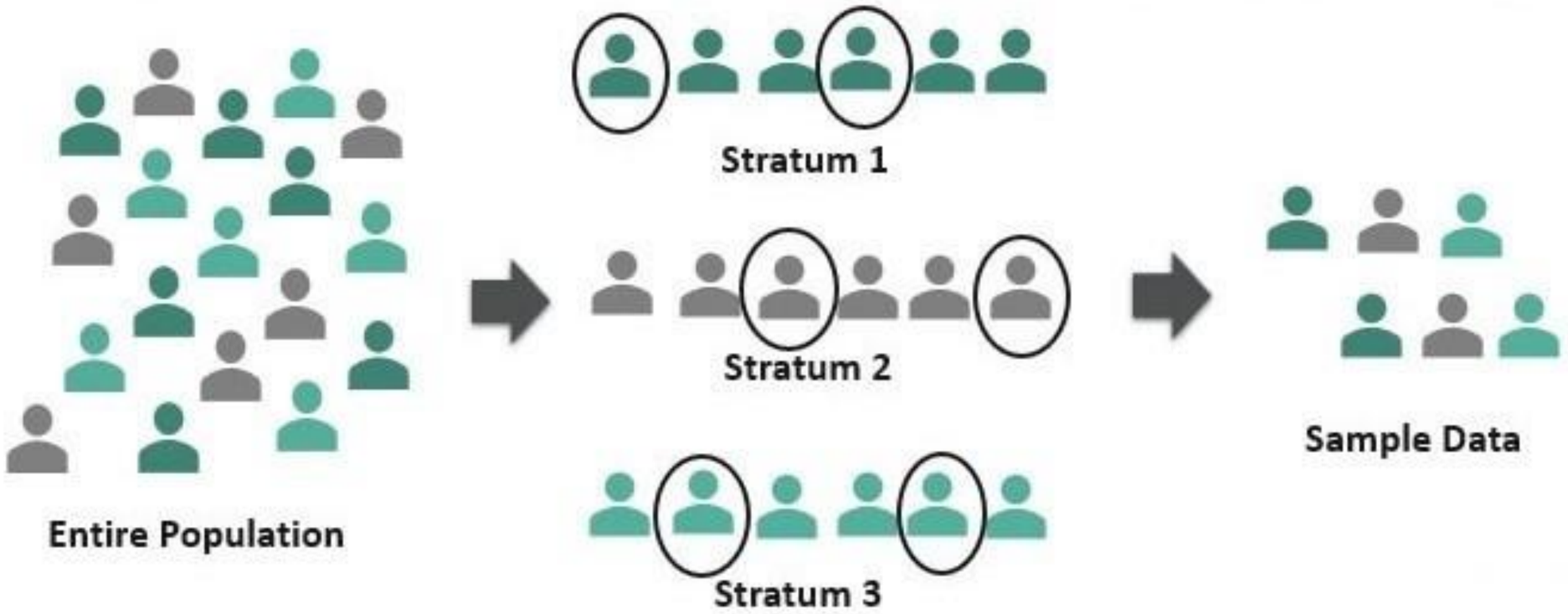
Original Source: https://www.linkedin.com/pulse/data-splitting-strategies-machine-learning-swapnil-sharma
https://medium.com/@mouadenna/time-series-splitting-techniques-ensuring-accurate-model-validation-5a3146db3088

Random Split



- randomly divide our dataset into a training set and a test set
- Random splitting may result in a biased dataset if the dataset is imbalanced, meaning that it contains more samples from one class than the others

Stratified Split



- ensures that the proportion of each class in the training and test sets is the same as that in the original dataset
- useful when dealing with imbalanced datasets
- ensure that the model is trained and evaluated on a representative sample of the data, and it can improve the model's overall performance

Stratify

```
    p1  p2 Output
0   45  53    Yes
1   75  35    Yes
2   55  78    Yes
3   74  68    Yes
4   79  56    Yes
5   55  87    Yes
6   78  68    Yes
7   57  87    Yes
8   35  67     No
9   45  46     No
10  45  53    Yes
11  75  35    Yes
12  55  78    Yes
13  74  68    Yes
14  79  56    Yes
15  55  87    Yes
16  78  68    Yes
17  57  87    Yes
18  35  67     No
19  45  46     No
---------
Yes    16
No      4
Name: Output, dtype: int64
```

Original Data

```
Yes    11
No      4
Name: Output, dtype: int64


Yes    5
Name: Output, dtype: int64
```

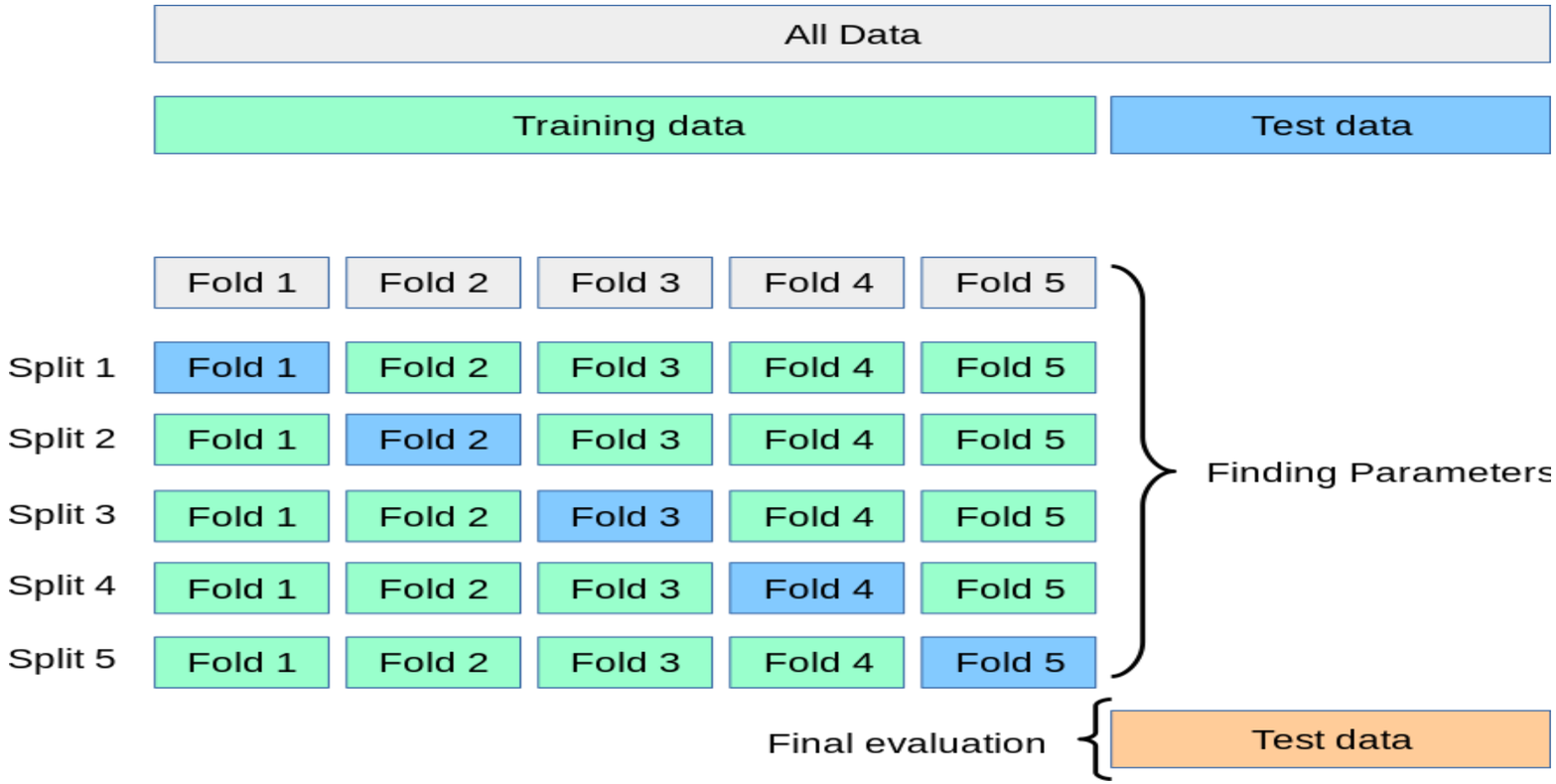
Without Stratify

```
Yes    12
No      3
Name: Output, dtype: int64


Yes    4
No     1
Name: Output, dtype: int64
```

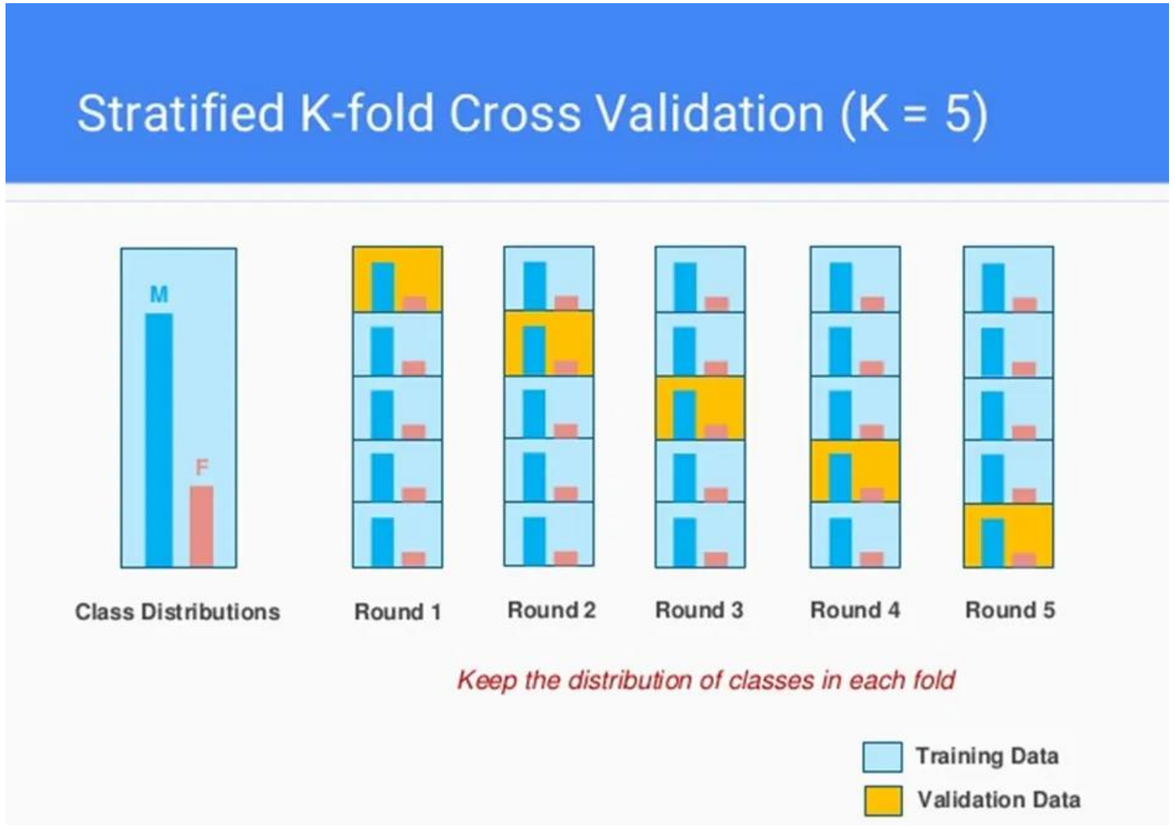
With Stratify

K-fold cross validation



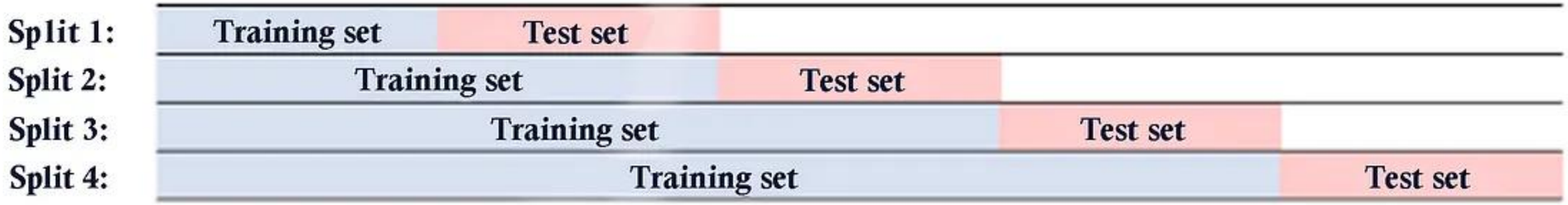
- splitting the data into k subsets or “folds” of roughly equal size
- The model is trained on k-1 of the folds and tested on the remaining fold.
- This process is repeated k times, with each fold being used as the test set once.
- The performance of the model is then averaged across all k iterations.

Stratified K-fold cross validation



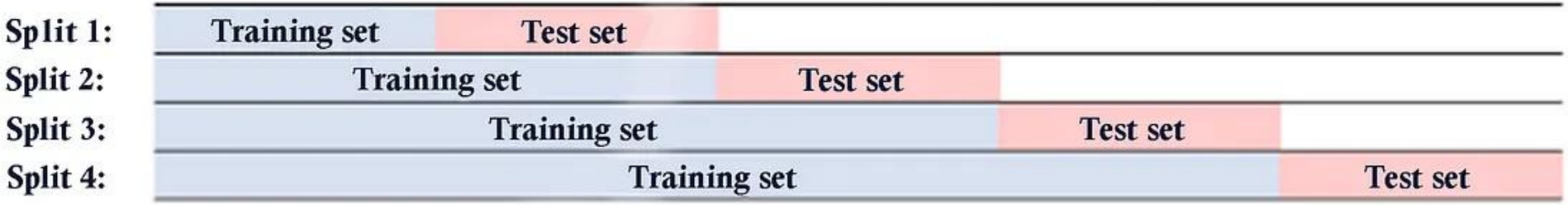
- the data is first divided into k folds based on the class distribution. each fold will have roughly the same proportion of samples for each class.
- Same process as k-fold cross validation is followed
- useful when the class distribution is imbalanced, meaning that there are different number of samples for each class
- helps to prevent bias in the evaluation process

Time Series Split



Option 1: TimeSeriesSplit

Option 2: Sliding/Rolling Window Split



Option 3: Expanding Window Split

Common mistakes during data training

1. Low-quality training data
2. Overfitting
3. Overemphasis on Validation and Test Set metrics

APPLICATIONS OF MACHINE LEARNING

https://builtin.com/artificial-intelligence/machine-learning-examples-applications

Thank You