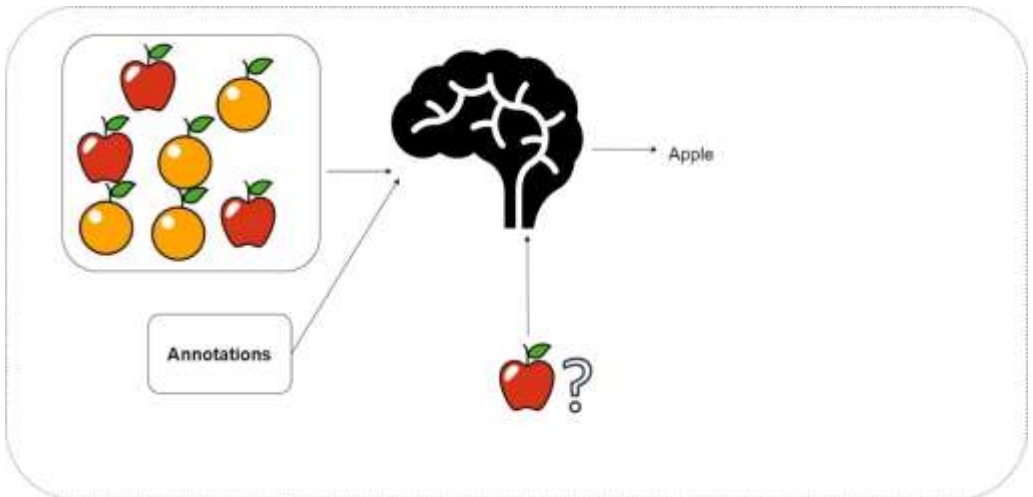
1

The Machine Learning Landscape

- Dr. Sifat Momen (SfM1)



Learning goals

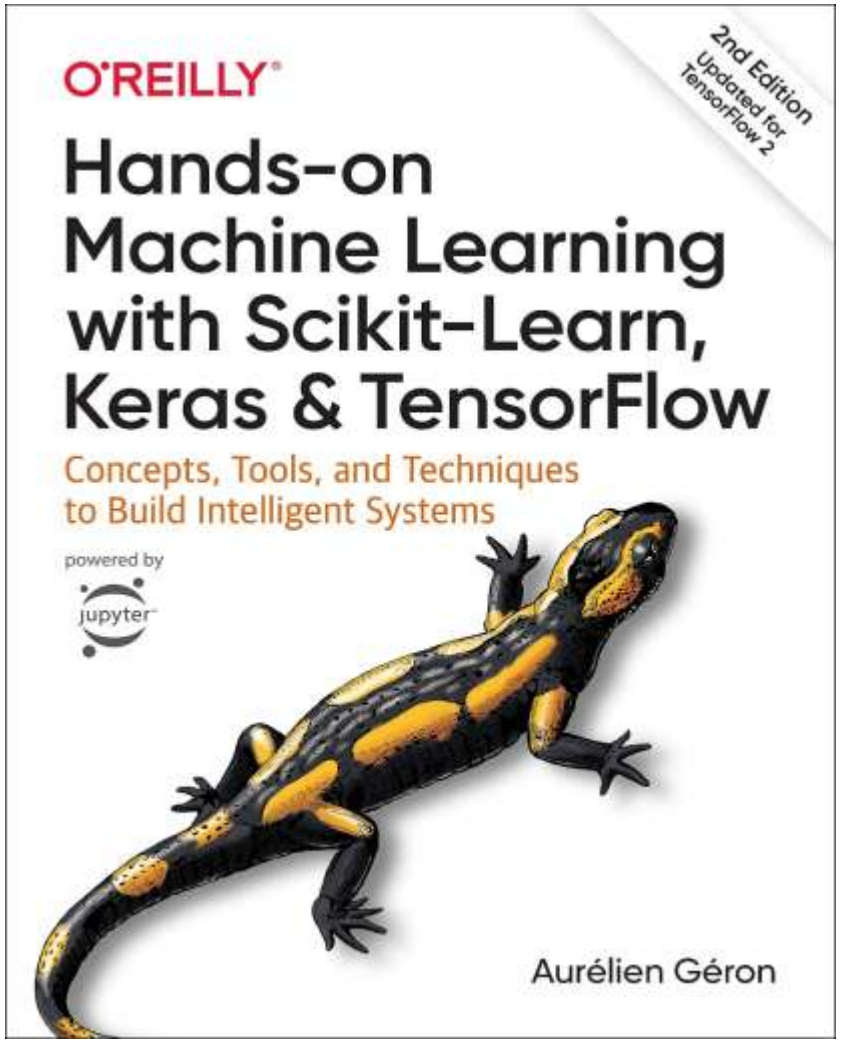
- After this presentation, you should be able to
  - Explain what machine learning is
  - Understand the different machine learning systems
  - Know different jargons related to machine learning
  - Understand the main challenges in machine learning
  - Understand the ethical concerns in machine learning

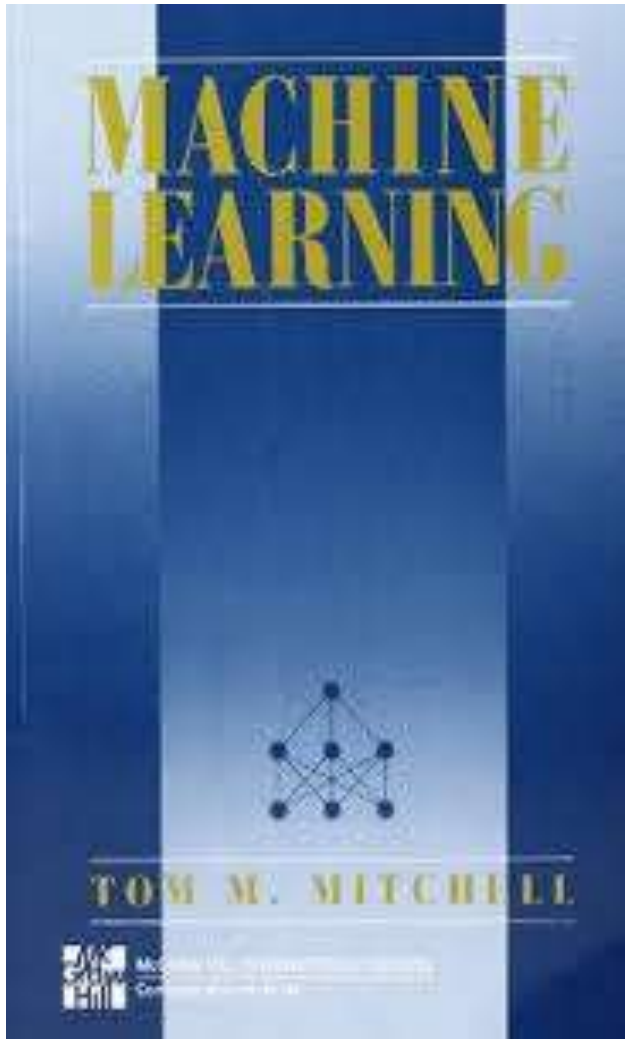
CSE 445: Machine Learning

- Welcome to CSE 445: Machine Learning – Spring 2025
- I am offering one section this semester
- Please note that you are **not** allowed to take classes (and specifically exams) in other sections
- This is a senior level course
  - It is expected that students should be able to demonstrate high degree of maturity
  - This includes reading and learning on own from existing literature, writing codes, conducting experiments, analyzing results and writing good technical paper(s).

Recommended Materials

- Lecture Materials
- Scholarly Papers





Project

- The objective of the project work is for you to
  - Experience a hands on skills in ML.
- Students will have to do innovative projects.
- I am not expecting the projects to have commercial value but it must address a realistic problem and should solve it.
- It should be ideally publishable in a reputed conference/ journal.

What is expected of you?

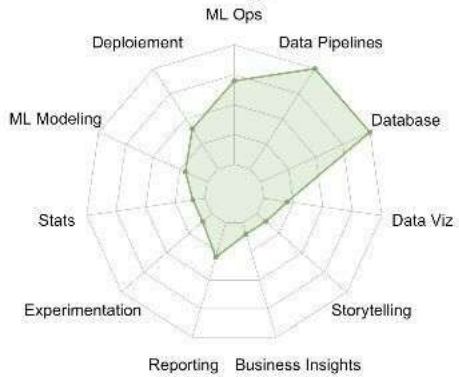
- Work hard
- Have an inquisitive mind
- Continued effort (You will not be able to do anything magically overnight)
- Consistency is the key to do well in this course

You need to have some backgrounds in

Linear Algebra Calculus Statistics Algorithms

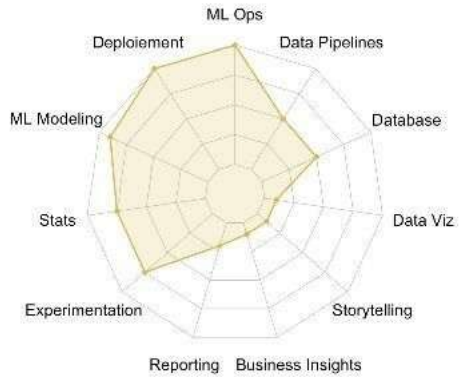
Types of Data Professionals

Data Engineer



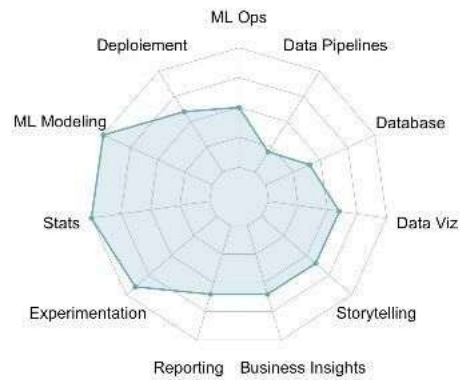
Data Scientists
ML Ops

ML Engineer

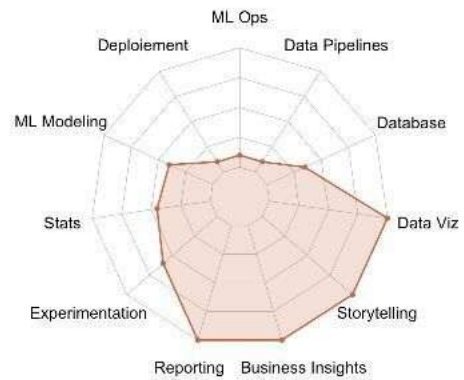


Data Analysts
ML Ops

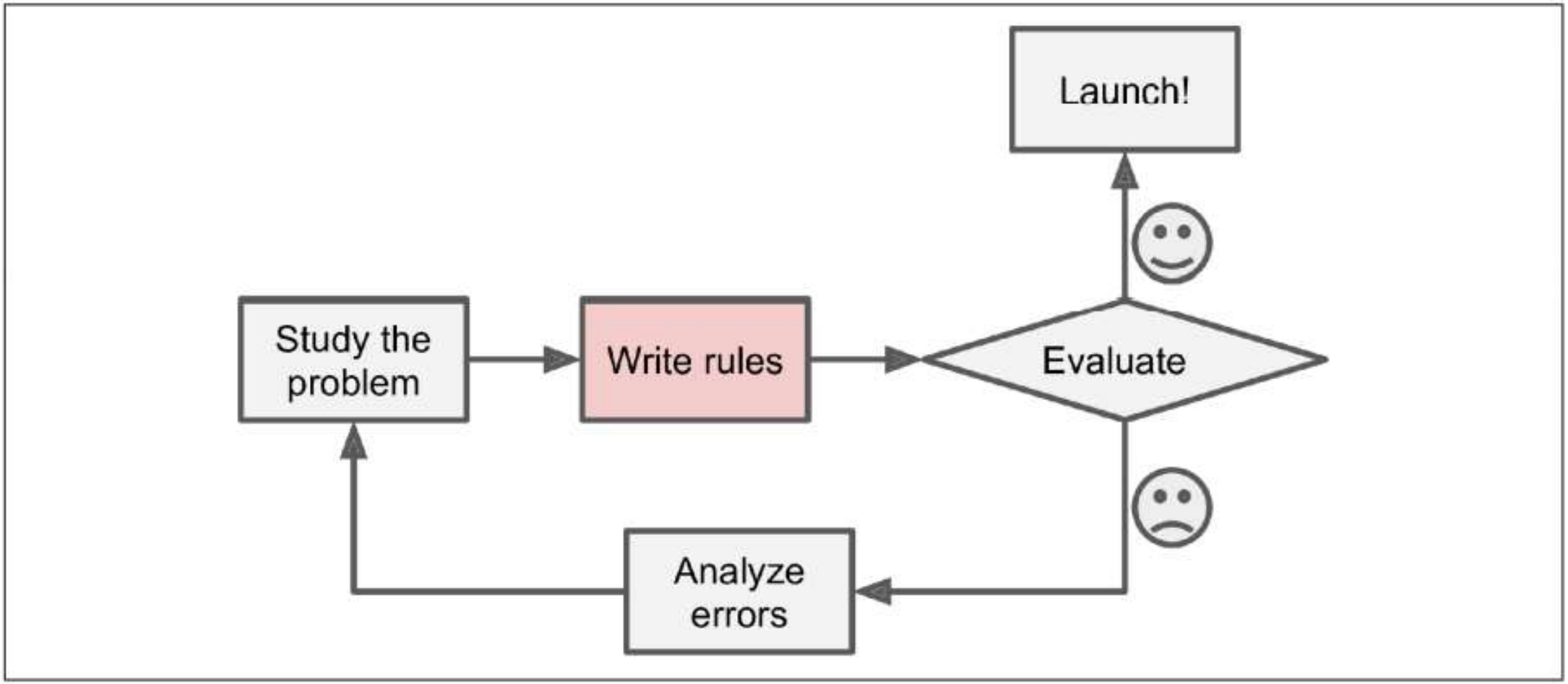
Data Scientist



Data Analyst



Traditional Approach to Solving Problems



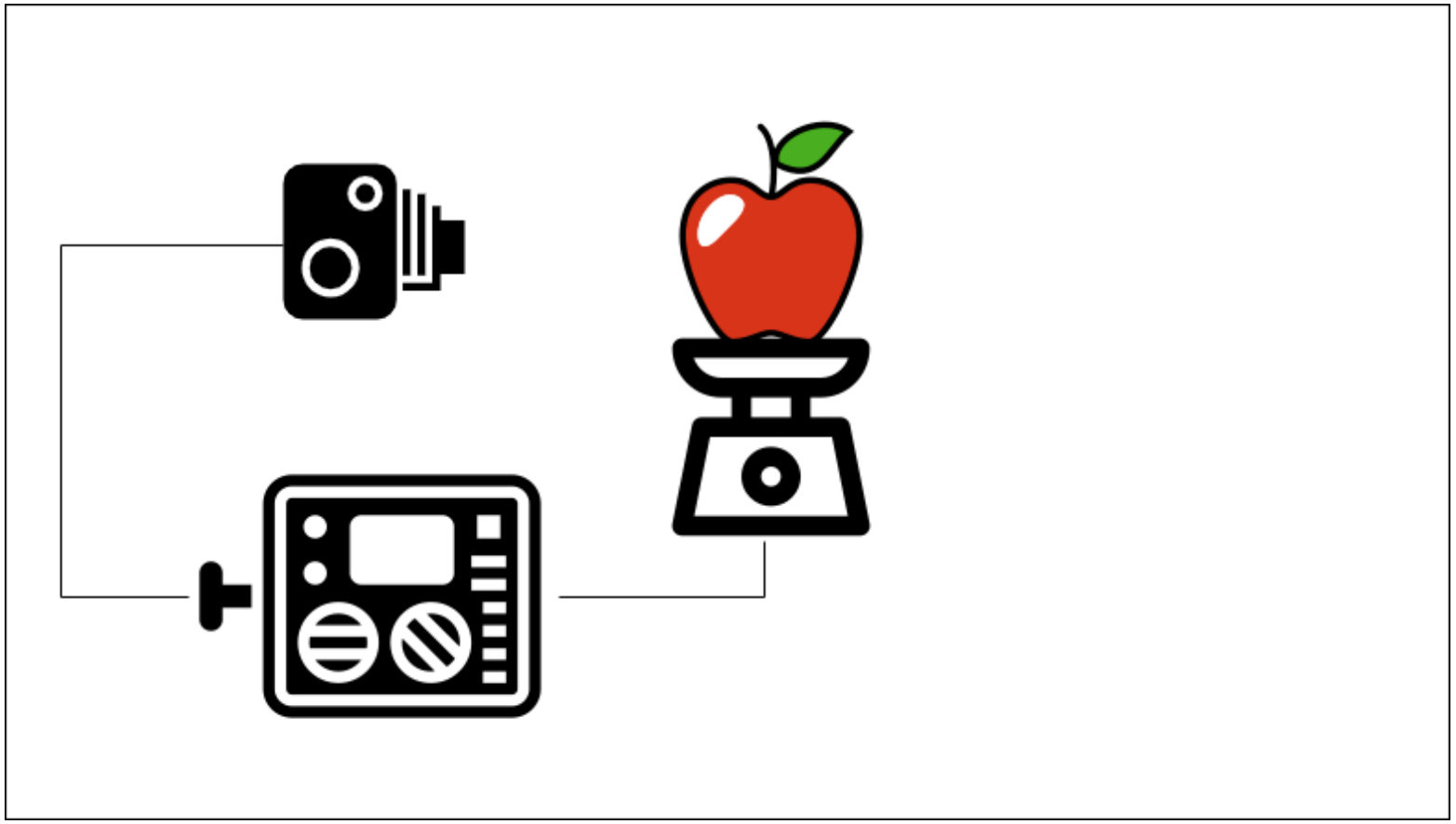
What is Machine Learning?

The best way to understand what a machine learning is and appreciate the notion of it is to start with an example.

A Thought Experiment

- Imagine we have a physical machine that can take the side-view picture of an object placed on it and measure the weight of the object.
- Object placed is either an orange or an apple
- The measurable features in this thought experiment are **color** and **weight**

A Thought Experiment



Now, you are tasked to write a program that determines if the object placed is an apple or an orange. How would you do that?

Real World

What about recognizing a cat from a picture?









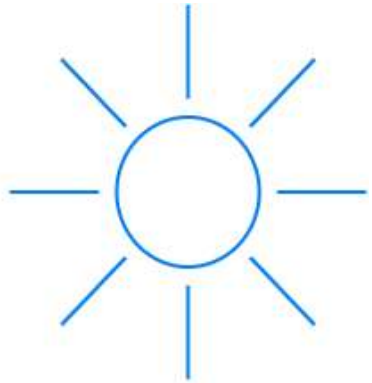

Real World Data

Noisy

Illumination

Occlusion

Camera angle



Exists a wide range of variation (in terms of features such as color, texture, weight etc..)

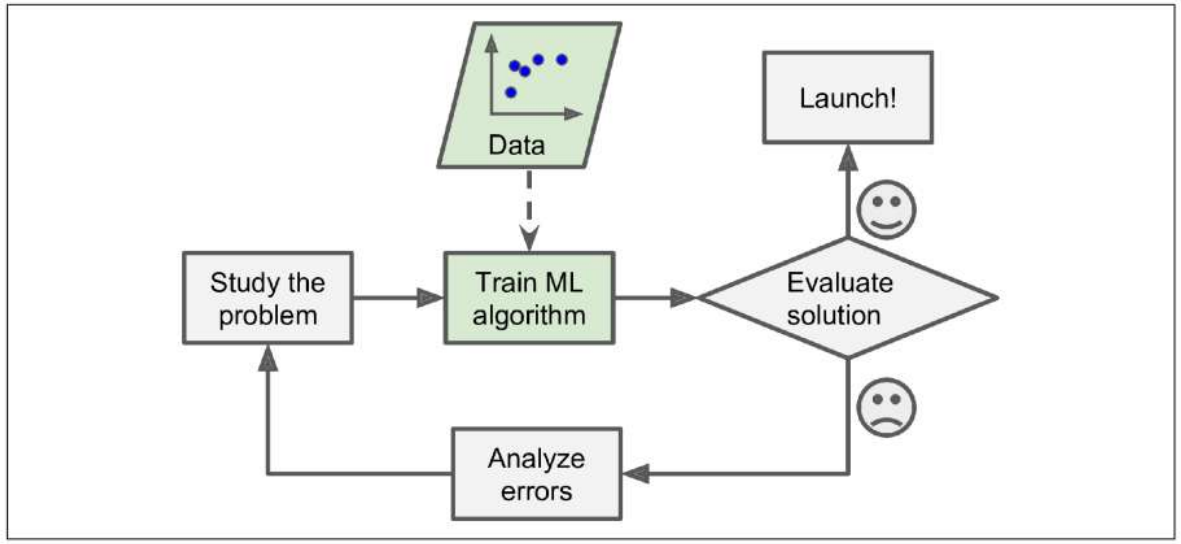
Real World Data (cont…)

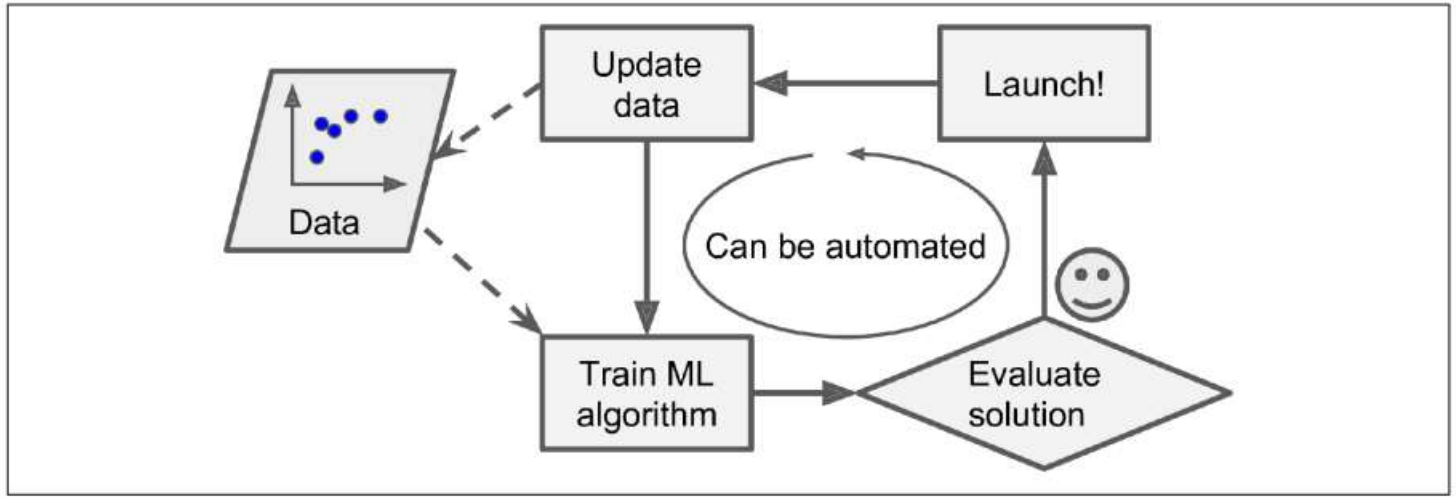
Consequently, it is impossible to write a rule based programming to recognize a cat or differentiate between fruits (such as orange and apple)

A shift in the paradigm is required

- Data driven Computing (Welcome to Machine Learning)
- We use data to create **model**
- Data is at the core of ML
- Without data, you will not be able to run a ML project.

Machine Learning Approach



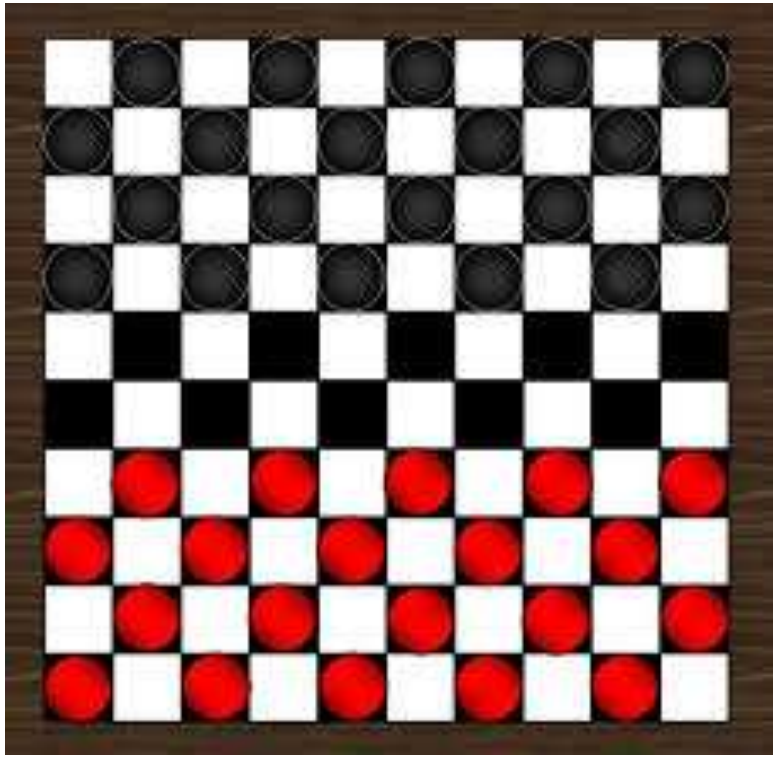


Well posed learning problems

- Machine Learning
  - It is a science that grew up from AI
  - The term, ML, was coined in 1959, by Arthur Samuel
- Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

- A computer program is said to learn from **experience *E***, with respect to some **task *T*** and **performance measure *P*** if its performance at **task *T***, as measured by ***P***, improves with **experience *E***. (Tom M. Mitchell)

Checkers learning problem



Task *T*: playing checkers

Performance measure, *P*: % of games won against opponents

Training Experience, *E*: Playing practice games against itself

Handwriting Recognition Problem

Task *T*: Recognizing and classifying handwritten words/characters within images

Performance measure *P*: % of words/characters correctly classified

Training Experience, *E*: Predicting the word/character from a collection of handwritten words/characters with given classifications

- MNIST Dataset (Yann Lecunn et al.)
- BanglaLekha Isolated (Nabeel Mohammed, Sifat Momen et al.)
  - https://data.mendeley.com/datasets/hf6sf8zrkc/2
  - Dataset comprises of 50 Bangla basic characters, 10 bangla numerals and 24 selected compound characters (total of 1,66,105 handwritten character images)

Robot driving learning problem

Task *T*: driving on public highways using vision sensors

Performance measure, *P*: Average distance travelled before an error (as judged by a human observer)

Training Experience, *E*: A sequence of images and steering commands recorded while observing a hums driver.

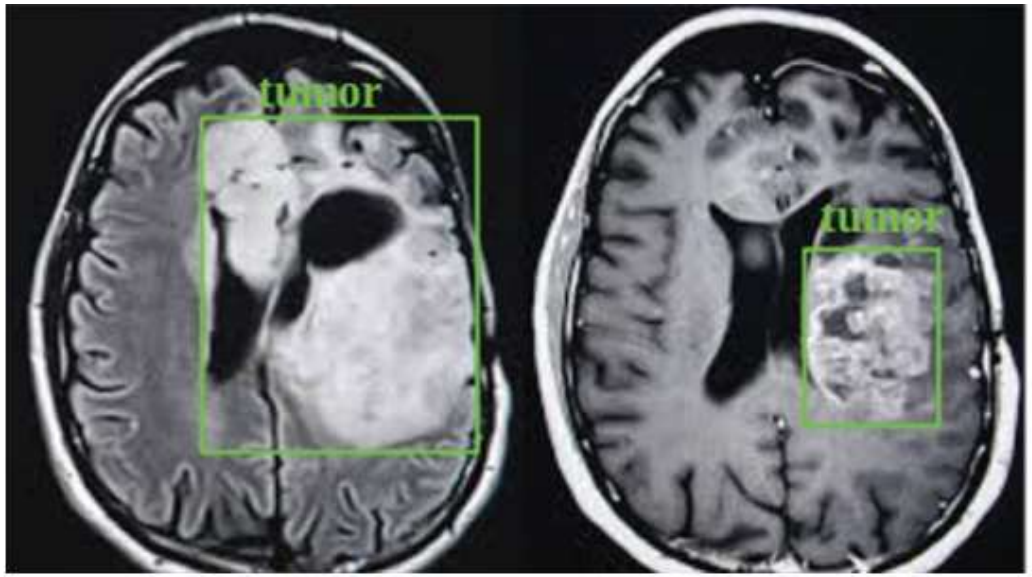
Classifying an incoming email as either a spam or not spam

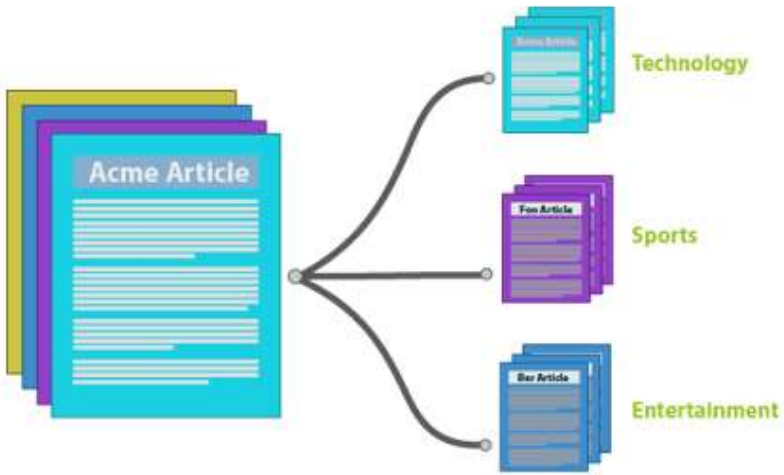
Task T: Correctly classify an email as either spam or not spam

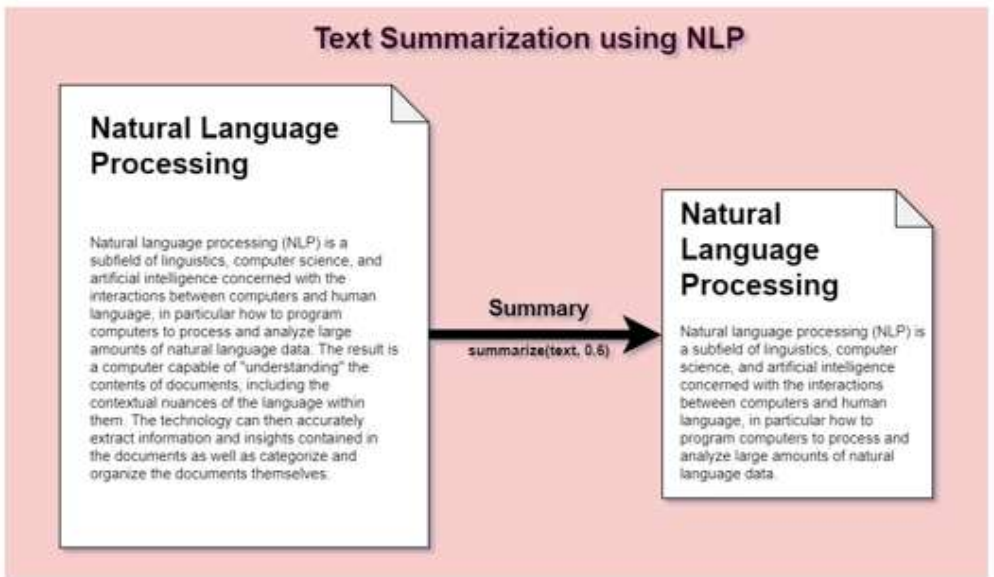
Performance measure, P: % of emails correctly classified

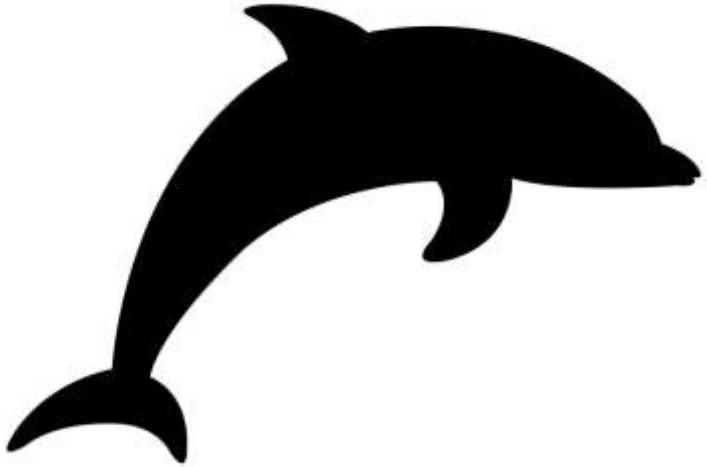
Training Experience, E: A repertoire of emails with proper labeling as either spam or not.

Some Applications of Machine Learning


tumor
tumor


Acme Article
Technology
Sports
Entertainment


Text Summarization using NLP
Natural Language Processing
Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.
Summary
summarize(text, 0.6)
Natural Language Processing
Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.



Write a story with dolphins and turtles as characters and include a moral in the story

Once upon a time, in a beautiful coral reef deep in the ocean, there lived a group of dolphins and a group of turtles. The dolphins were known for their intelligence and speed, while the turtles were known for their wisdom and long lives. The two groups lived separately and rarely interacted with each other. The dolphins were always busy swimming and playing, while the turtles were content to take their time and enjoy the peacefulness of the reef. One day, a group of humans came to the reef to fish. They used big nets to catch fish, not caring if they caught anything else in the process. In the process, they caught many of the reef's inhabitants, including the dolphins and turtles. The dolphins, being faster and more agile, managed to escape the nets, but the turtles were not as lucky. Many of them were caught and taken away. The dolphins felt guilty for not being able to help their friends and decided to come up with a plan to save the turtles.

They used their intelligence and speed to come up with a strategy to chase the fishing boat away and free the trapped turtles. The plan was successful and the boat was chased away, but not before many of the turtles were harmed. The dolphins and turtles were able to free the remaining turtles and bring them back to the reef. After that experience, the dolphins and turtles realized the importance of working together. They learned that each of them had unique strengths and weaknesses, and that by combining their abilities, they could protect the reef and all its inhabitants better. From then on, the dolphins and turtles worked together, and the reef was safe and prosperous once more. The moral of the story is that everyone is unique, and we all have different strengths and weaknesses, but when we work together, we can achieve so much more. Unity is strength and cooperation leads to protection of environment and preserving lives.

Write a rap about North South University

Verse 1:
Listen up y'all, I got a story to tell
About a uni that's doing well
North South University, the name you should know
Where education and knowledge both flow

Chorus:
North South, North South
We're the best, no doubt
North South, North South
We're reaching new heights, without a doubt

Verse 2:
From BBA to Engineering, we got it all
We're making sure you're ready for the call
Professors who care, and a campus so fine
We're creating leaders, one student at a time

Chorus:
North South, North South
We're the best, no doubt
North South, North South
We're reaching new heights, without a doubt

Verse 3:
Research and innovation, that's our game
We're putting Bangladesh on the map, with our name
We're not just a uni, we're a community
We're here to empower, and set you free

Chorus:
North South, North South
We're the best, no doubt
North South, North South
We're reaching new heights, without a doubt

Outro:
So come and check us out, and take a tour
North South University, where your future is secure.

Types of Machine Learning Systems

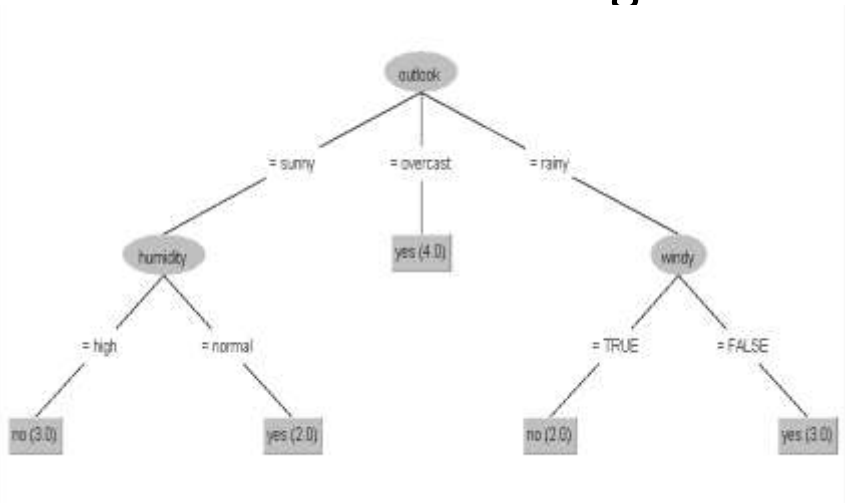
- Various Types of Machine Learning Systems based on the following criteria:
  - Whether or not they are trained with human supervision (supervised, unsupervised, semi-supervised, reinforcement learning)
  - Whether or not they can learn incrementally on the fly (online versus batch learning)
  - Whether they work by simply comparing new data points, or instead by detecting patterns in the training data and building a predictive model(instance-based versus model-based learning)

Supervised Learning

- In supervised learning, the algorithm builds a mathematical model from a set of data that contains inputs and labels (desired outputs) <X, y>.

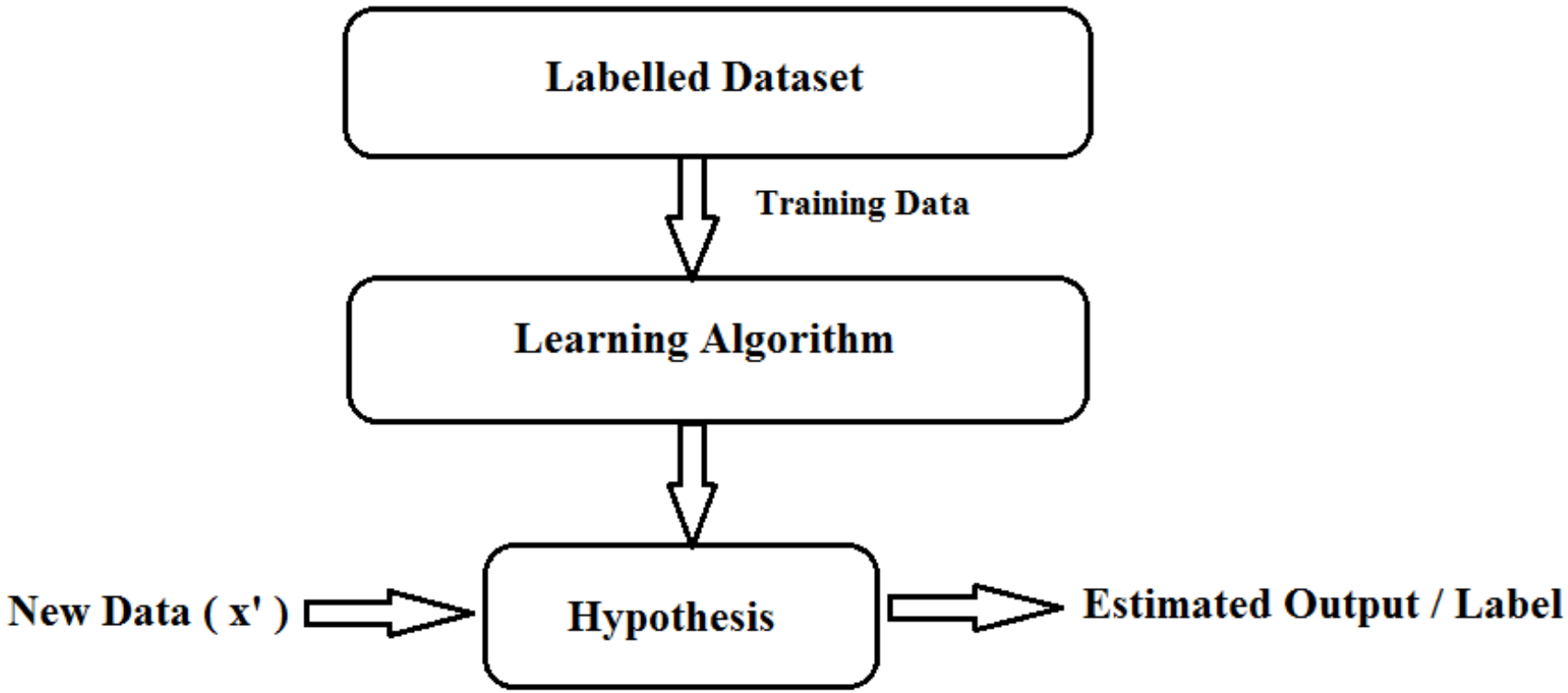
No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

The data contains 14 instances and 4 attributes. The target



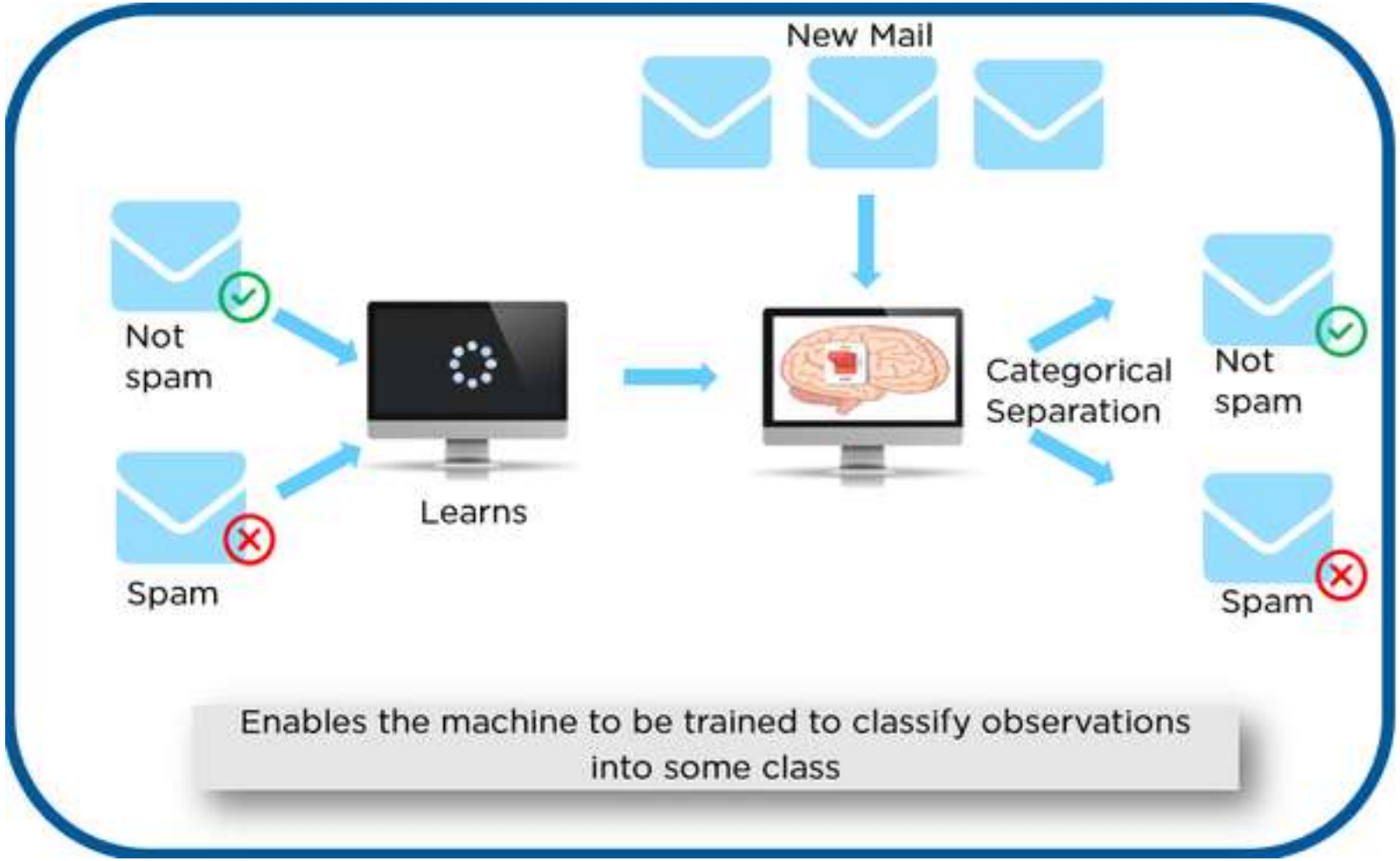
A Decision Tree Model

Supervised Learning

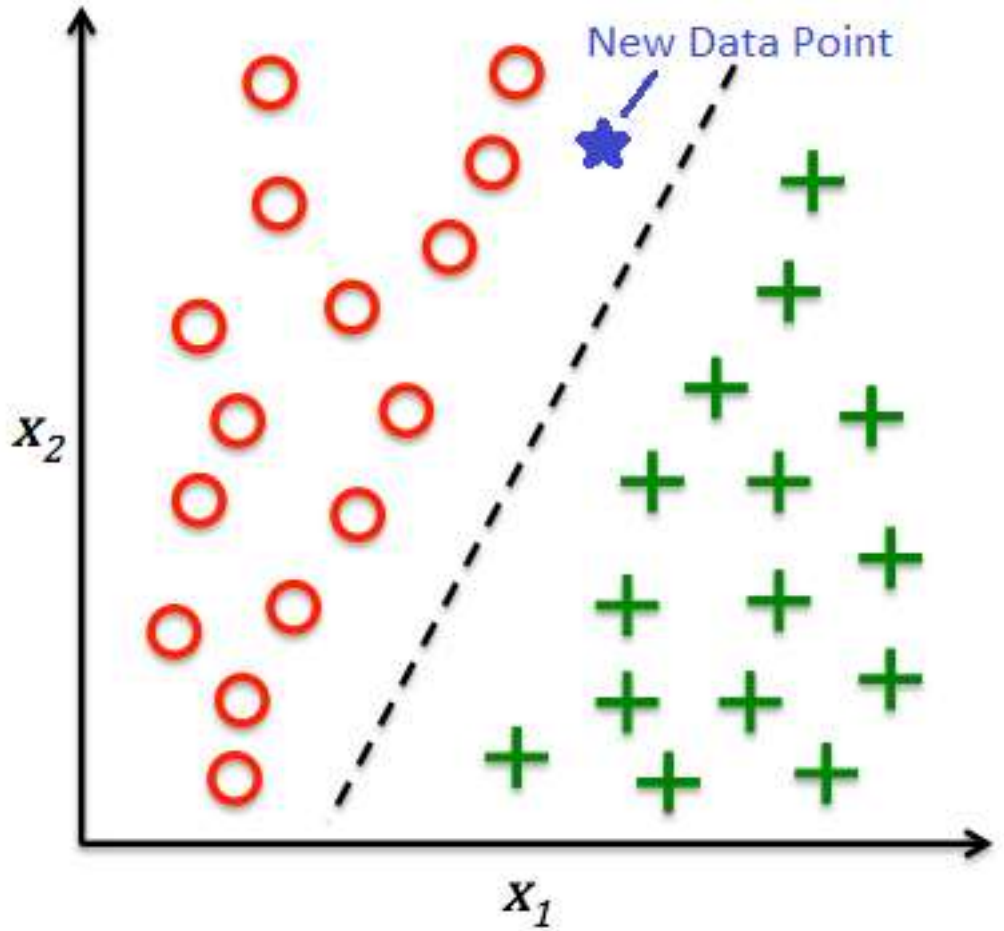


SUPERVISED LEARNING

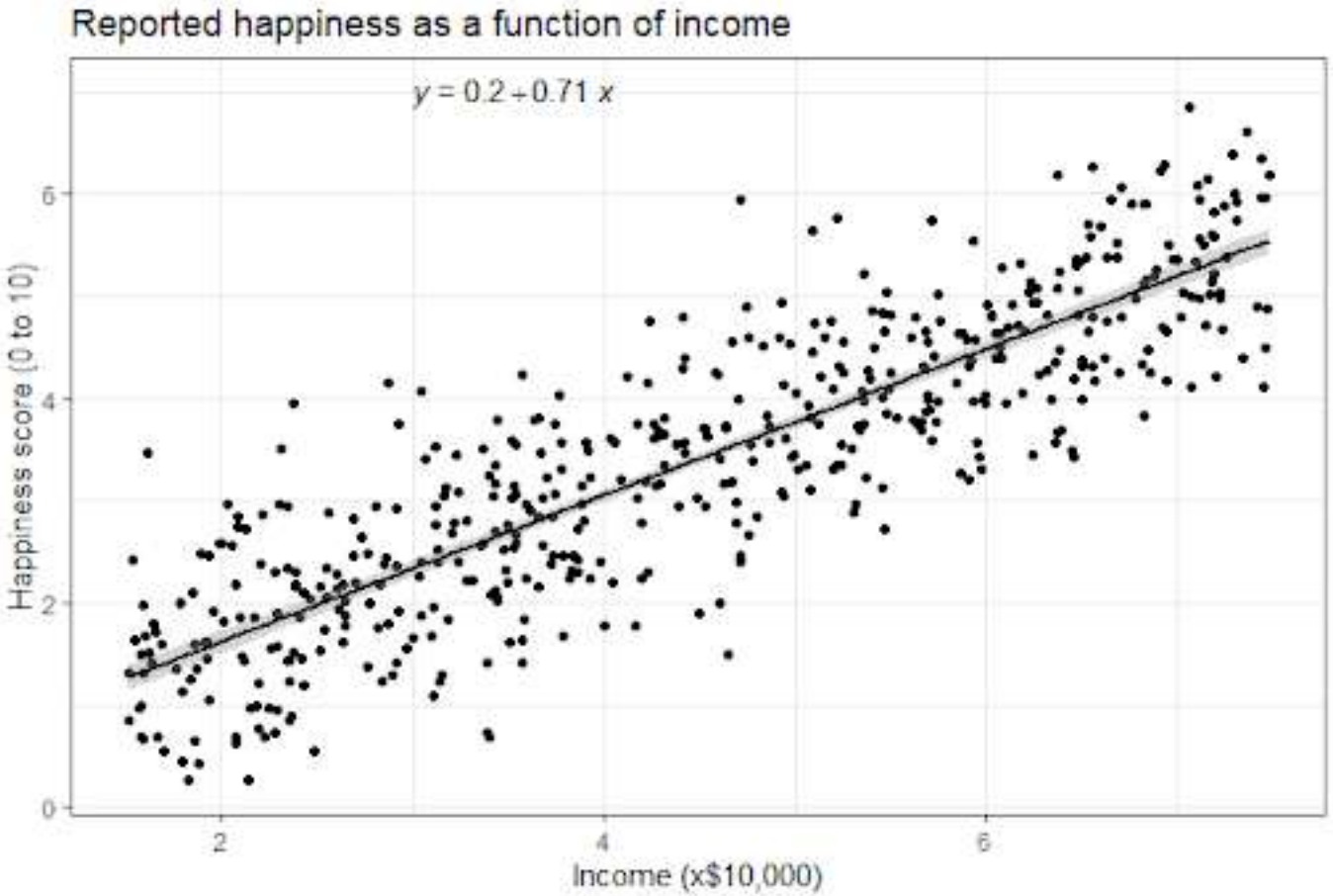
Supervised Learning - Classification

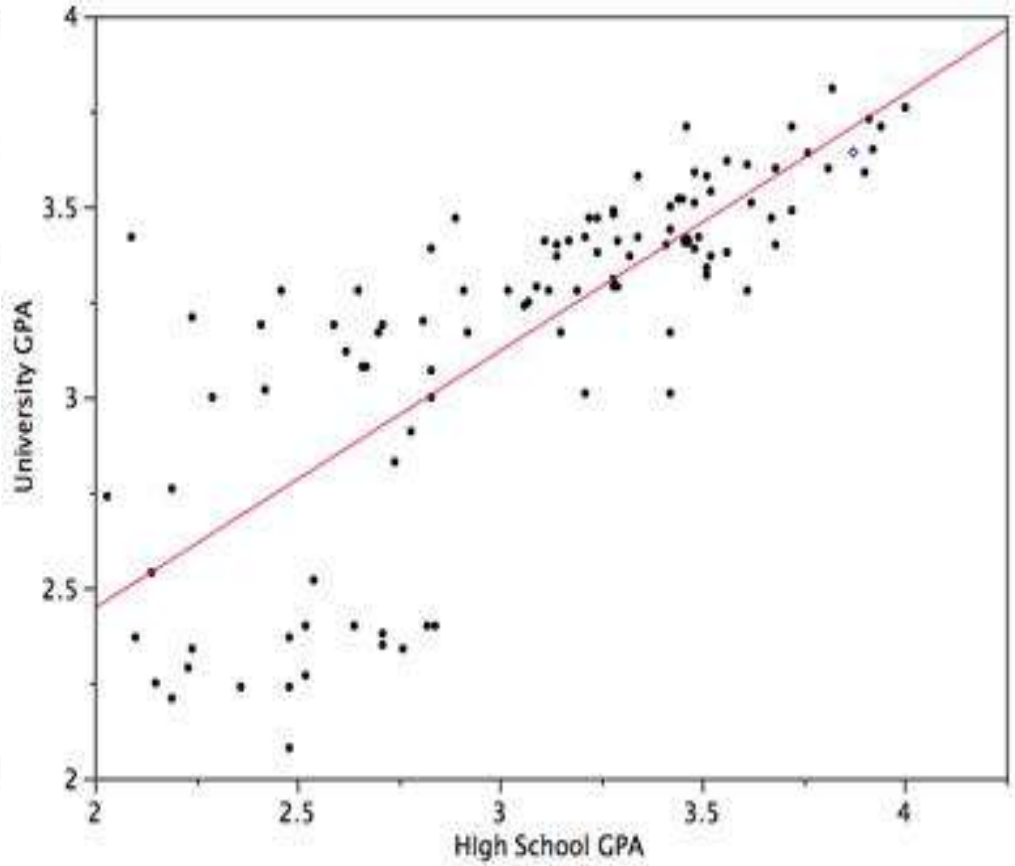


Supervised learning - Classification



Regression


Reported happiness as a function of income
y = 0.2 + 0.71 x
Happiness score (0 to 10)
Income (x$10,000)
0
2
4
6
2
4
6

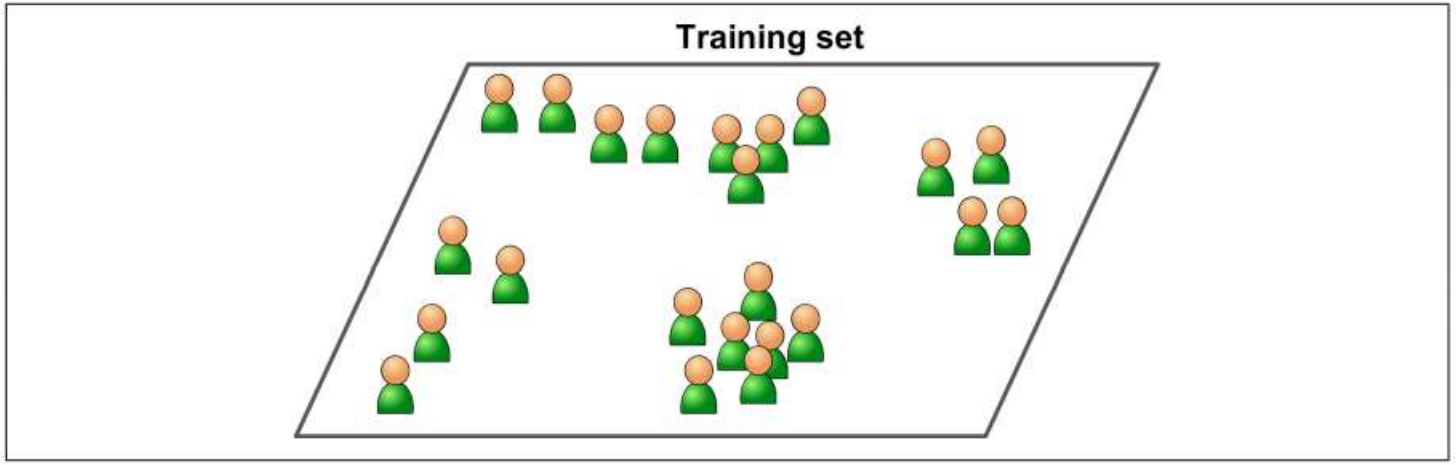

University GPA
High School GPA
2
2.5
3
3.5
4
2
2.5
3
3.5
4

Examples of Supervised Learning Algorithms

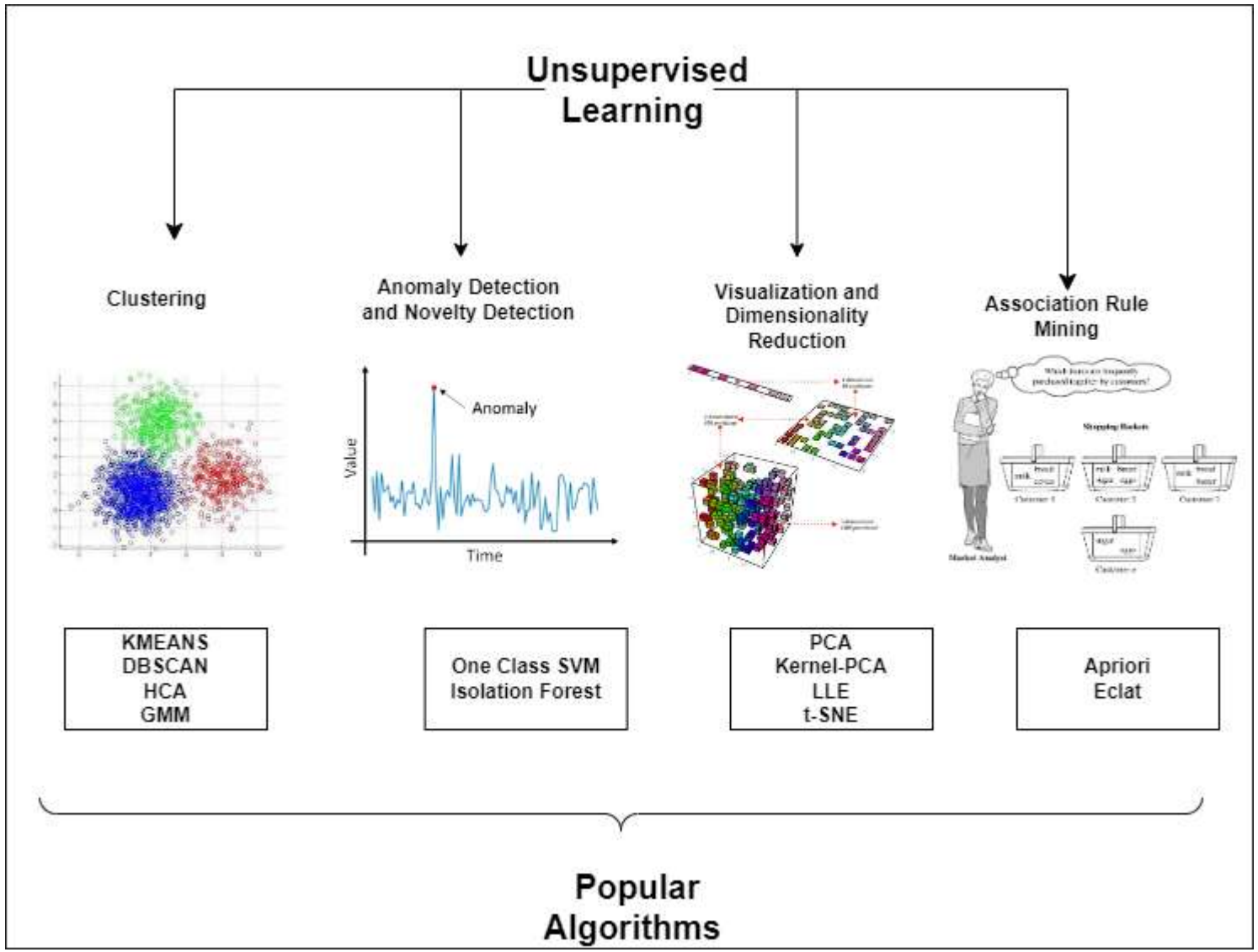
- K-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines
- Decision Trees
- Random Forests
- Neural Networks
- …

Unsupervised Learning

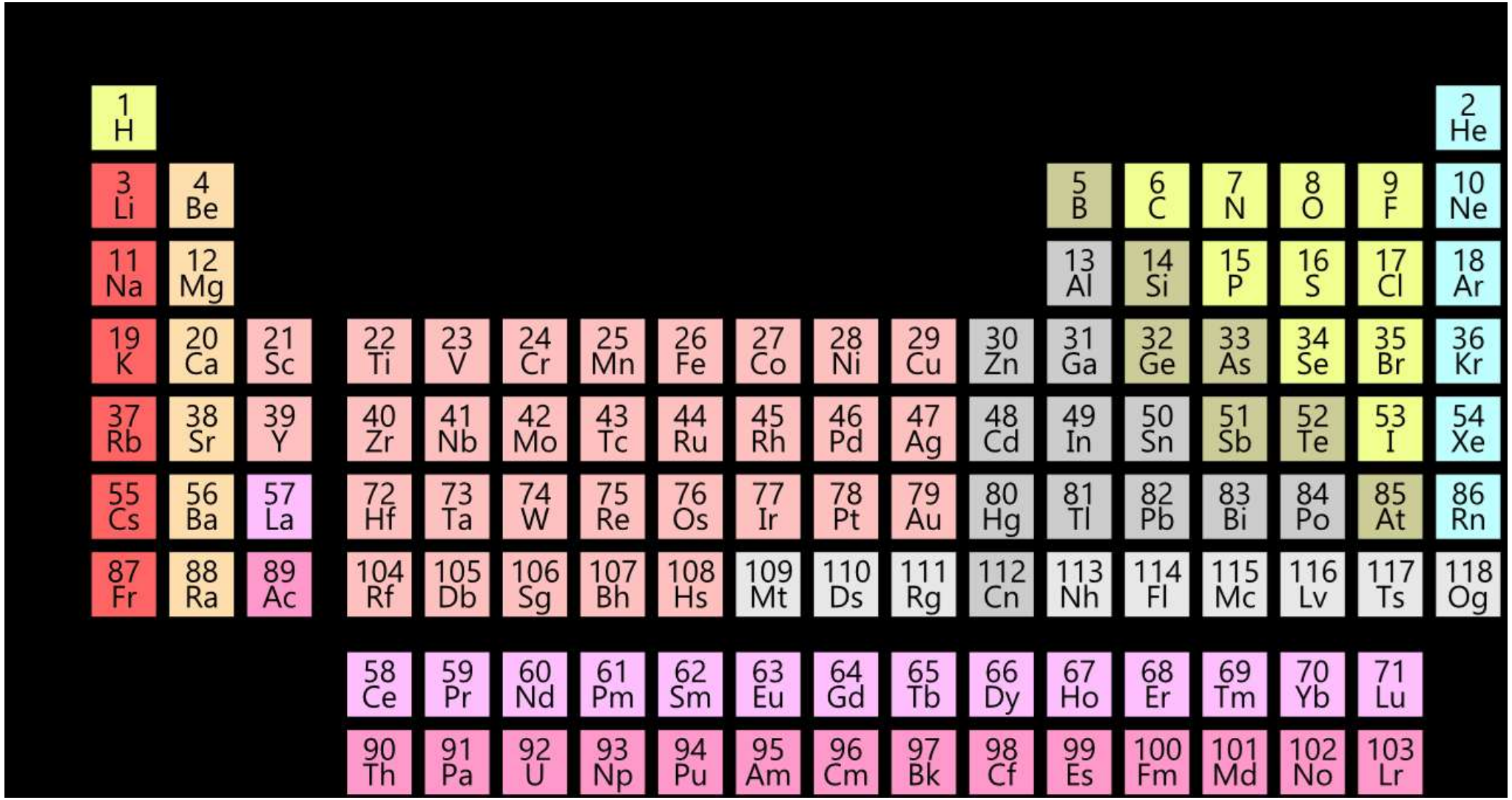
- In Unsupervised Learning, the training data is **unlabeled**. The system tries to learn without a teacher.



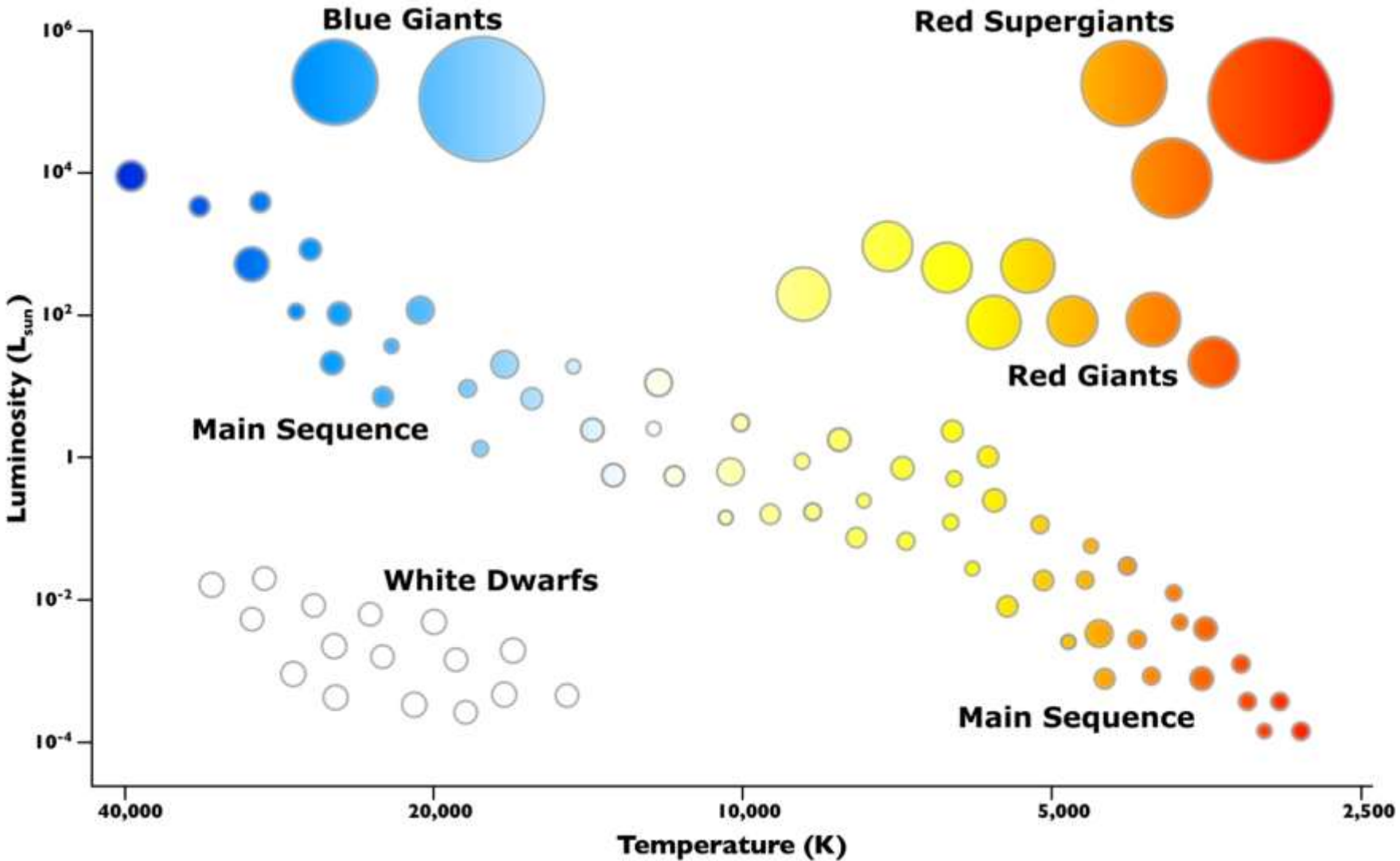
Unsupervised Learning



Unsupervised Learning



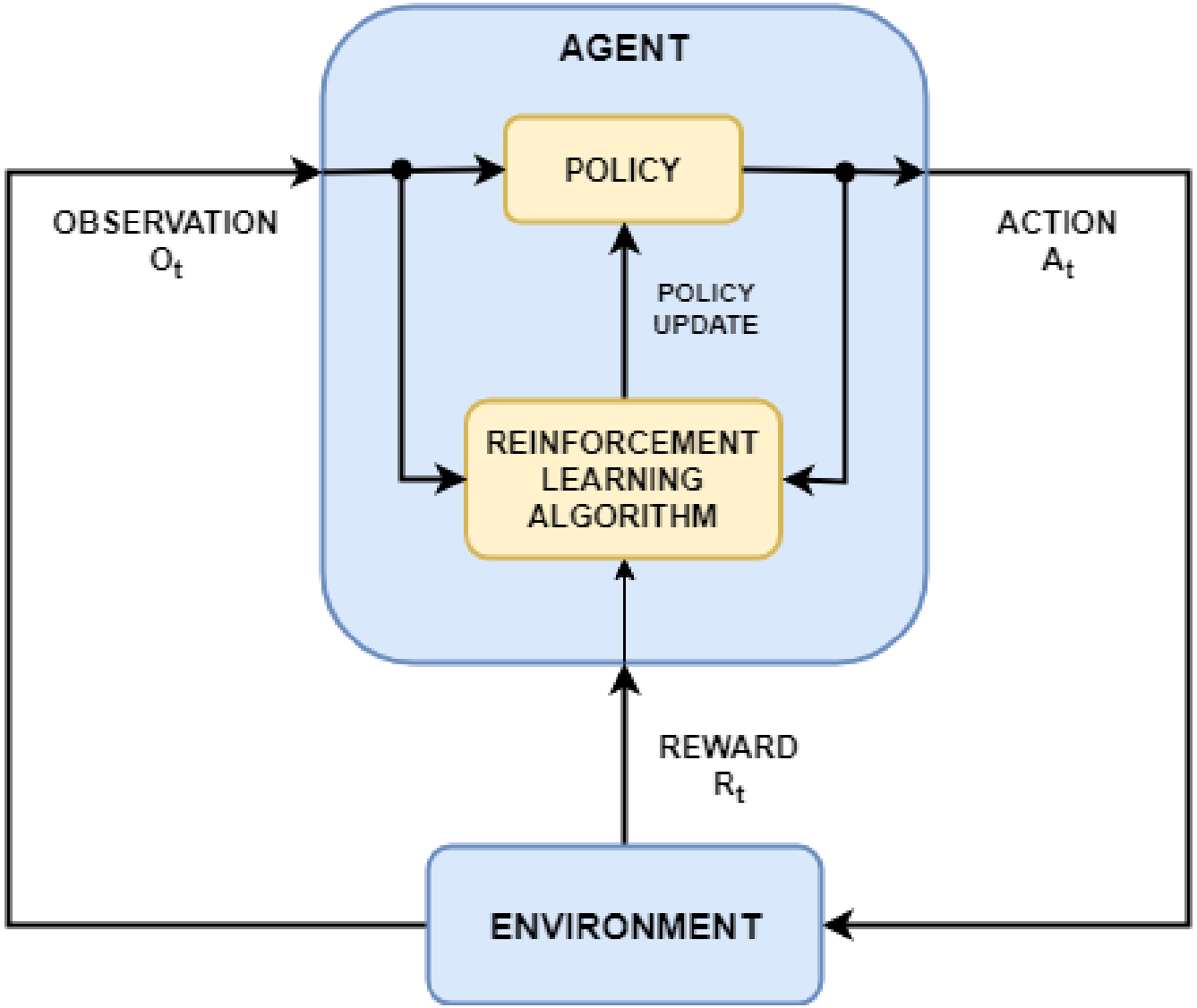
Unsupervised Learning



Semi-Supervised Learning

- Supervised learning works on **labeled data**
- Unsupervised learning works on **unlabeled data**
- In Semi-Supervised learning, the training data contains both **labeled and unlabeled data**
- **Scenario**
    - We have a large dataset. Manual labeling of the entire dataset is laborious and expensive
    - One solution could be to label a sample of the dataset and train the labeled portion to create a model. This would however mean that we are not fully utilizing the larger dataset we have and thus the model that we create may be less robust
    - A potential solution: (1) Label a sample of the large dataset, (2) train a model using this labeled portion, (3) use the model to predict the unlabeled portion (**pseudo-labeling**), (4) train using the entire dataset

Reinforcement Learning

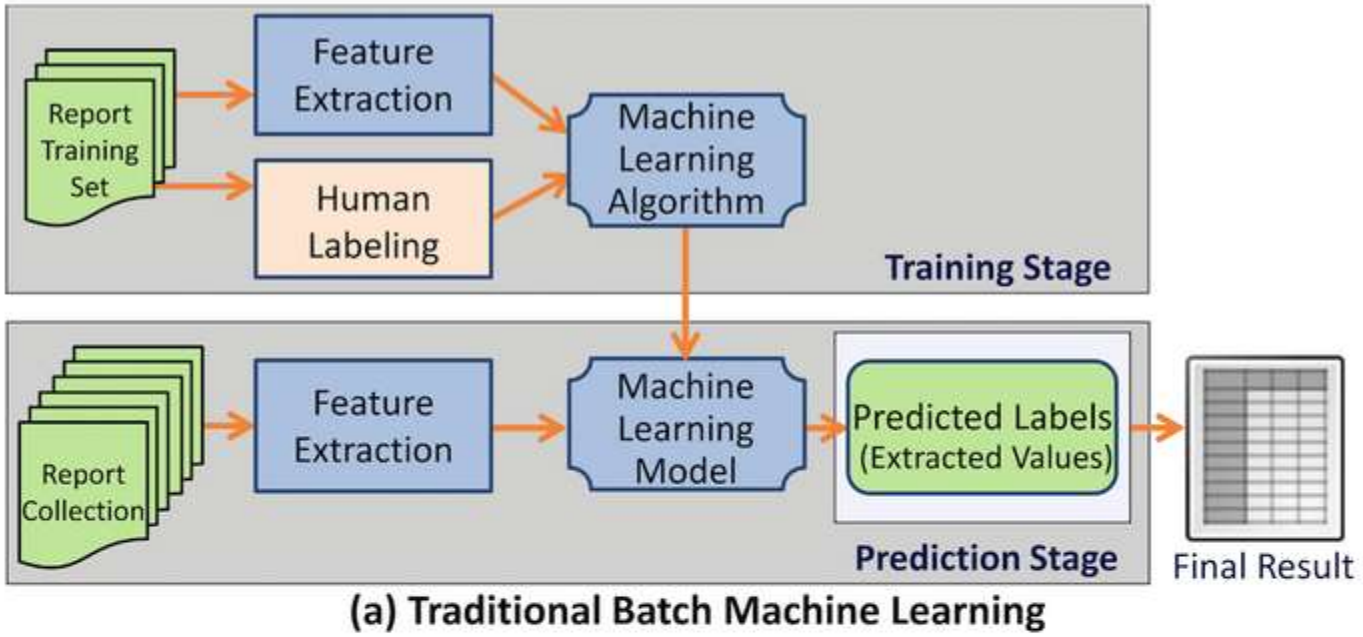


Batch and Online Learning

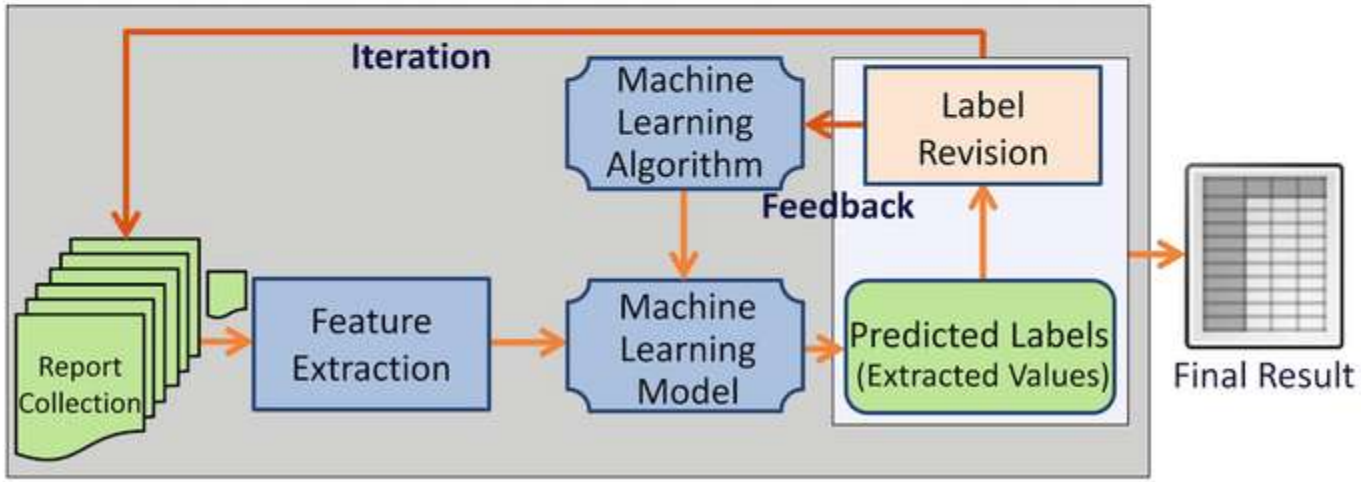
- Another criterion used to classify Machine Learning Systems is whether or not the system can learn incrementally from a stream of incoming data
- **Batch Learning**
  - Incapable of learning incrementally
  - It must be trained using all available data
  - Generally takes a lot of time and computing resources – so it is typically done offline
  - Hence, also known as offline learning

Batch Versus Online Learning

- Online Learning is great for systems that receive data as continuous flow (e.g. stock prices) and need to adapt autonomously.
- It is also a good option if you have limited computing resources – Once an online learning system has learned about new instances, it does not need them anymore, so you can discard them
- Online learning algorithms can also be used to train systems on huge datasets that cannot fit in one machine's main memory (this is called *out of core* learning). The algorithm loads part of the data, runs a training step on that data, and repeats the process until it has run all of the data

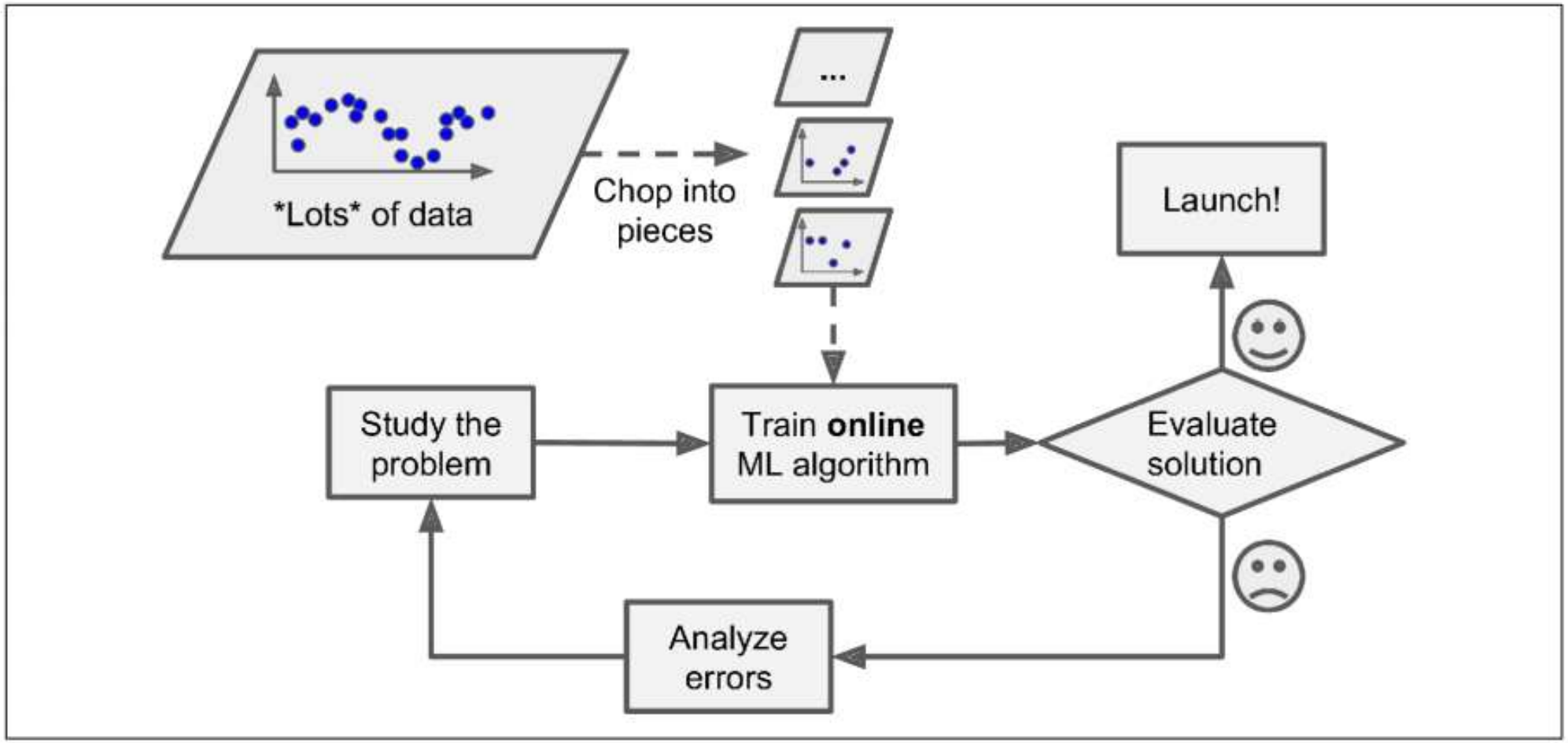


(a) Traditional Batch Machine Learning



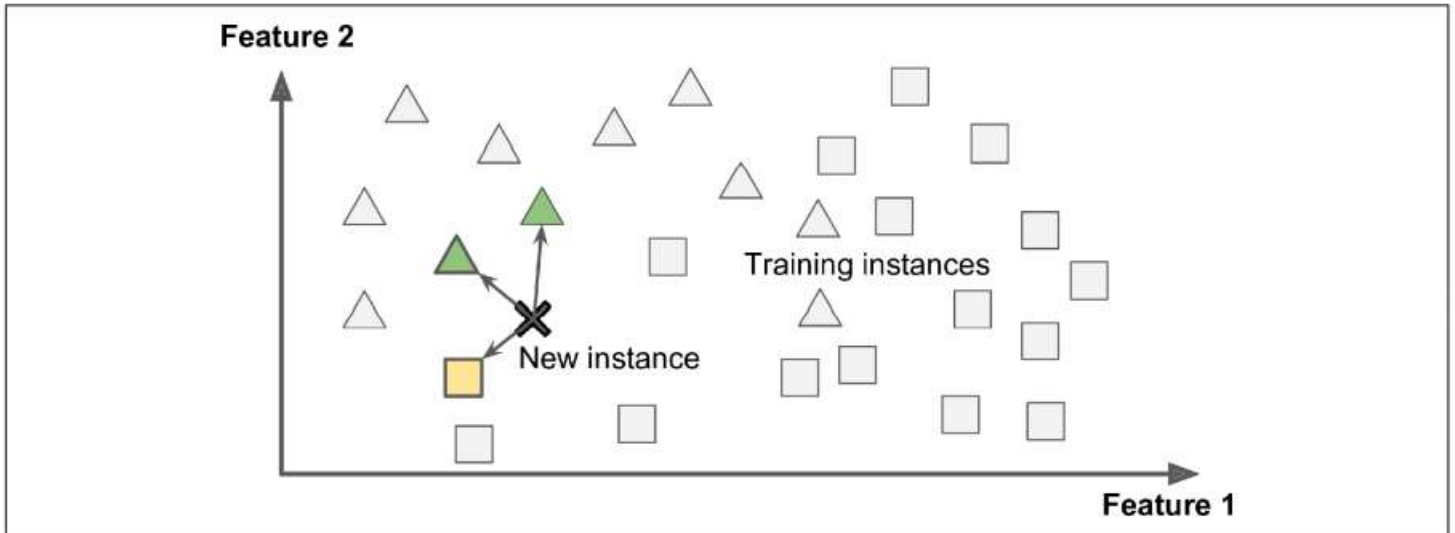
(b) Online Machine Learning

Online Learning to Handle Large Dataset



Instance Based Versus Model Based Learning

- One more way to classify machine learning algorithms is by how they *generalize*.
- **Instance Based Learning**
  - The most trivial form of learning is rote learning (to learn by heart)
  - System learns the examples by heart, then generalizes to new case by using a similarity measure to compare them to the learned examples (or a subset of them)



Instance Based Versus Model Based Learning

- Model Based Learning
  - Another way to generalize from a set of examples is to build a *model* of these examples and then use the model to make *predictions*