Machine Learning

Introduction

Eng. Gihan Weerasekara
Consultant/ Lecturer
National Institute of Business Management

Count Numbers from 1 – 1 000 000 000?

Human Evolution

Technology Disruptions

- Fire and Simple tools (2 million years back)
- Languages for Communication (50 000 – 200 000 years back)
- Agriculture (12 000 years back)
- Metal Tools (7 000 years back)
- Writing (5 000 years back)
- Industrial Revolution (1760 AD)
- Electronic Computer (1945 AD)
- Internet (1990 AD)
- Development of AI and Machine learning (2005 onwards)

What is Machine learning?

Definitions

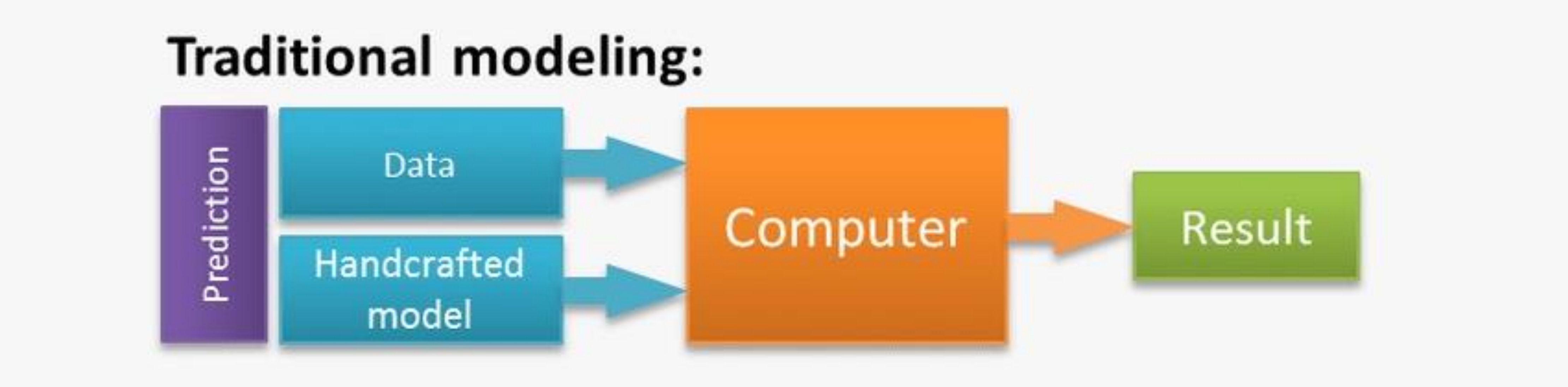
Arthur Samuel:

"Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed."

Tom Mitchell:

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

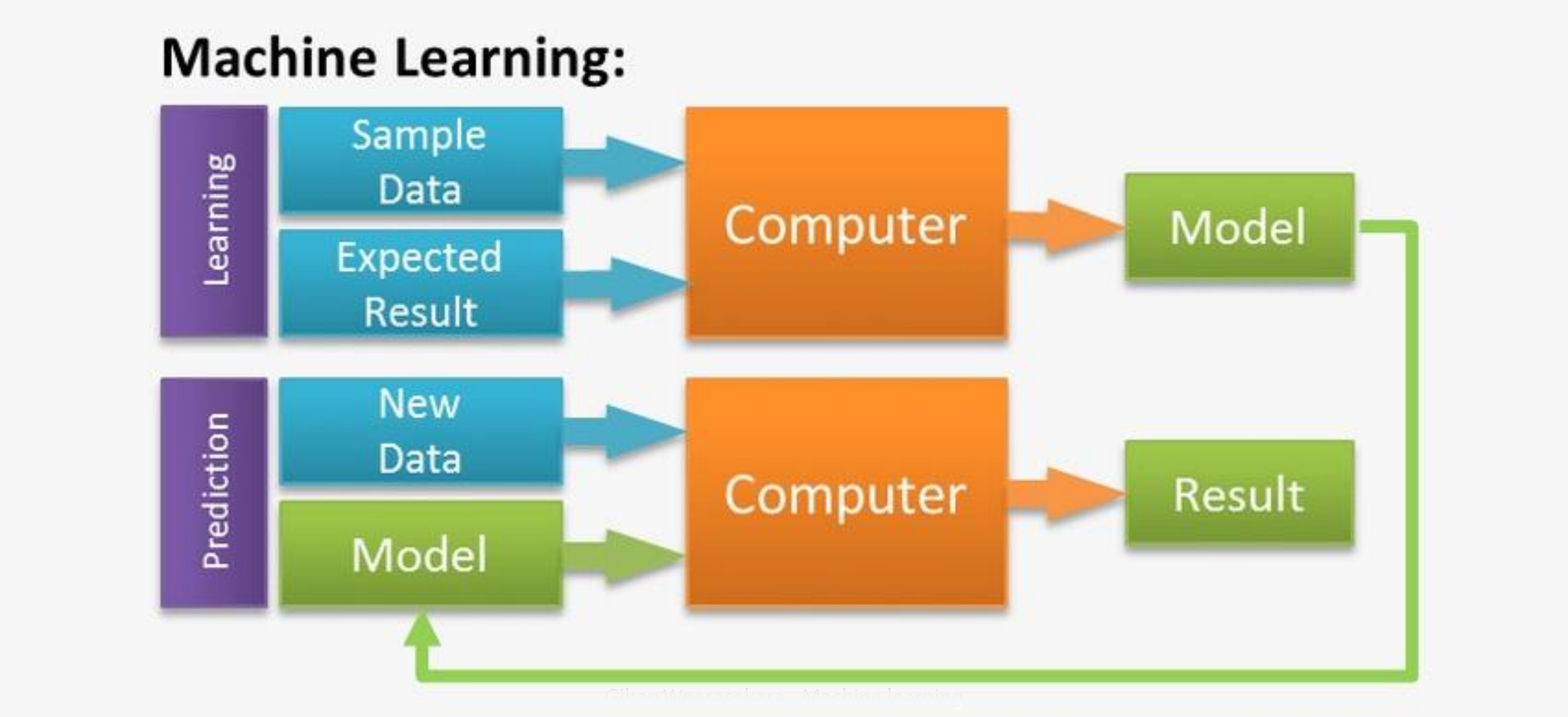
Traditional Approach



In **traditional programming,**

humans provide explicit instructions for the computer to follow

Machine Learning Approach



Machine Learning Approach

In **machine learning,**

the computer uses data and algorithms to automatically learn the rules or patterns without being explicitly programmed.

Learning from Data

ML is the process of using data to "train" a model to recognize patterns.

Similar to,

“how humans learn from past experiences to make decisions”

Key difference between traditional programming and ML is, instead of programming explicit rules, the machine learns rules from data on its own.

Learning from Data

Machine learning relies heavily on mathematical tools (Linear Algebra, Matrices etc), Calculus, statistical methods and algorithms to find the underlying patterns and relationships in the data.

Ex : Spam filters

Recommendation systems in Social media

(user specific content and advertising)

Generalization in Machine Learning

Generalization is,

a model's ability to perform well on unseen data

(not just the training data)

Example:

If you train a model to recognize different types of fruits from images, it should not just work on the specific images it was trained on, but also on new fruit images it has never seen before

Generalization in Machine Learning

train-test split

A dataset is often divided into two parts

- training data
- test data.

The model is trained on the training data and evaluated on the test data to check how well it generalizes.

Key Machine Learning Concepts

- Features and Labels
- Training and Testing Data
- Overfitting and Underfitting
- Bias-Variance Tradeoff

Features and Labels

Features

These are the input variables used to make predictions. They represent the attributes or properties of the data.

Labels

They are the outputs the model is trained to predict (also called targets).

In supervised learning, these are the known values used to teach the model.

Features and Labels

Example:

In a house price prediction task, features could be the size of the house, number of rooms, and location, while the label would be the house price.

Features are often denoted as **X** and labels as **Y**.

Training and Testing Data

The dataset is typically divided into two main parts:

training data and **testing data**.

Sometimes a third part, called **validation data,** is also used.

Training and Testing Data

Training Data:

This is the data used to train the model.

It contains both features and labels, allowing the model to learn the relationships between inputs and outputs.

Testing Data:

Once the model is trained, it is tested on unseen data to evaluate how well it generalizes.

The test data also includes labels but is not shown to the model during training.

Training and Testing Data

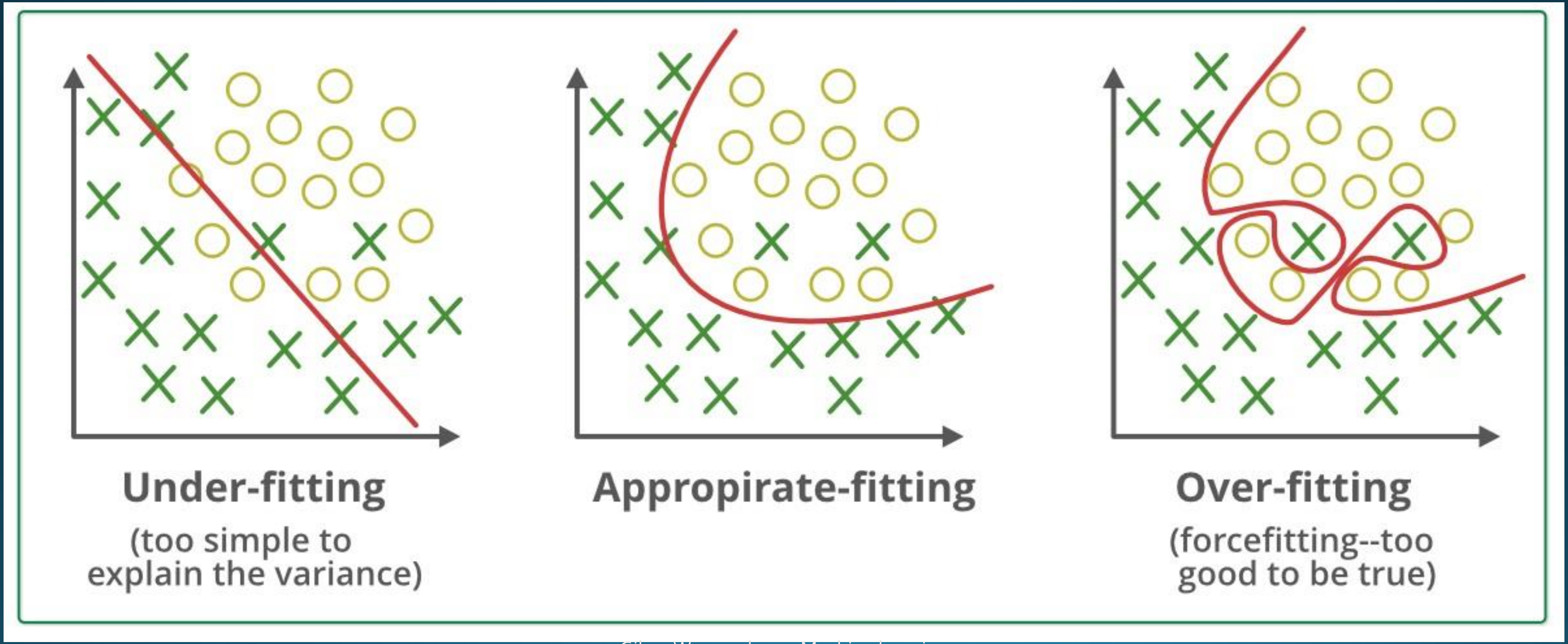
Validation Data:

Sometimes, a separate set of data is used to fine-tune hyperparameters, known as validation data.

Example:

For a model to predict house prices, 80% of the data might be used to train the model, while the remaining 20% is used to test its performance.

Overfitting and Underfitting


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Overfitting and Underfitting

Overfitting:

When a model performs very well on the training data but poorly on unseen data, it's said to overfit.

The model has learned not only the patterns in the data but also the noise or irrelevant details.

It becomes too complex and specific to the training data.

Overfitting and Underfitting

Underfitting:

On the other hand, underfitting occurs when the model is too simple and fails to capture the patterns in the data.

It performs poorly on both the training and test datasets.

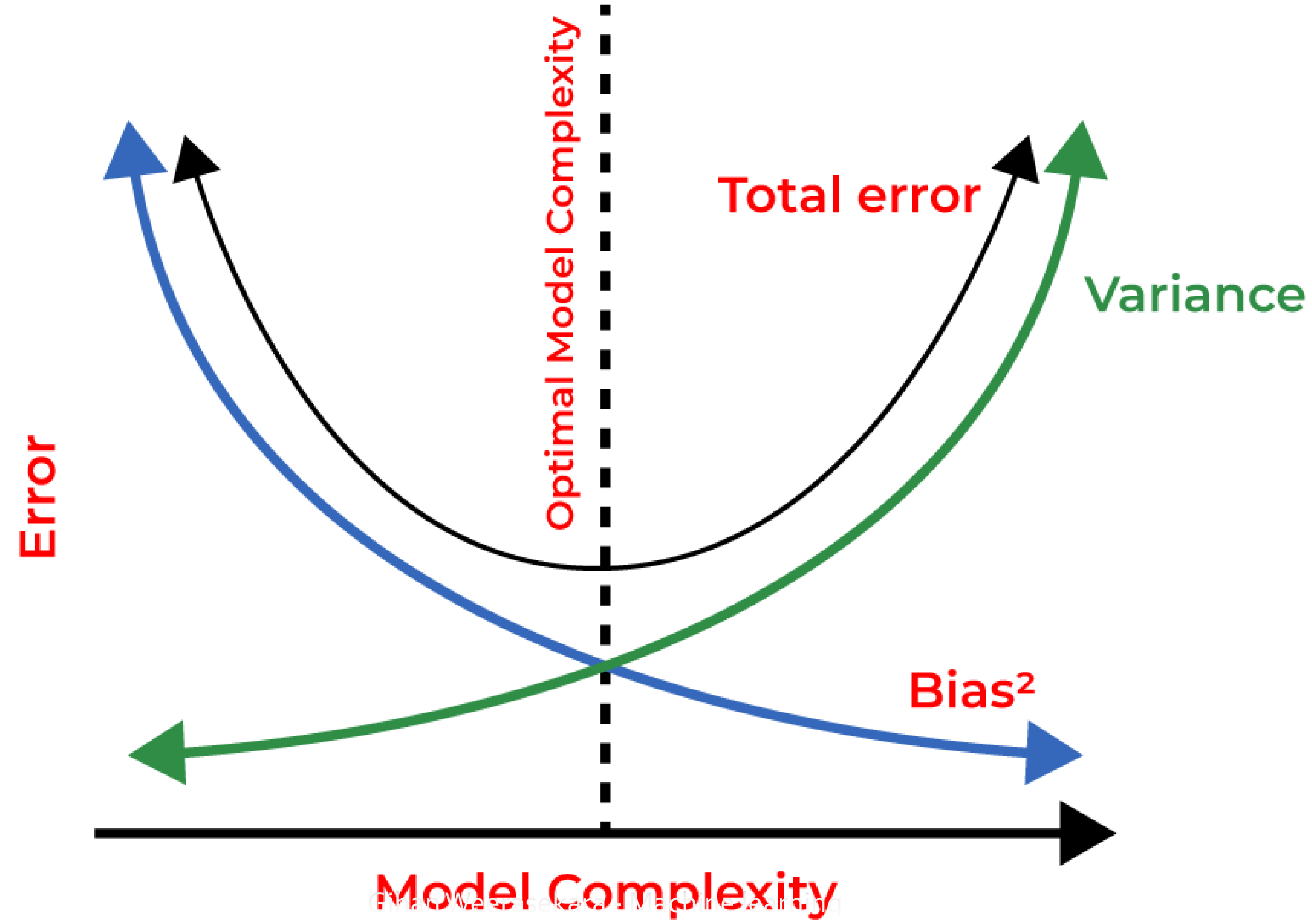
Overfitting and Underfitting

Example:

If you're predicting house prices, an underfitting model might only use a basic linear relationship between square footage and price, ignoring other important factors.

An overfitting model might memorize the training data, leading to poor predictions on new houses.

Bias-Variance Tradeoff



Bias-Variance Tradeoff

Bias:

Bias refers to the error due to overly simplistic models.

High bias can lead to underfitting, where the model cannot capture the complexity of the data.

Variance:

Variance refers to the error caused by too much complexity in the model.

High variance can lead to overfitting, where the model learns noise along with the actual patterns.

Bias-Variance Tradeoff

Tradeoff:

There is a tradeoff between bias and variance.

Increasing the complexity of the model reduces bias but increases variance, and vice versa.

The goal is to find the right balance.

Bias-Variance Tradeoff

Example:

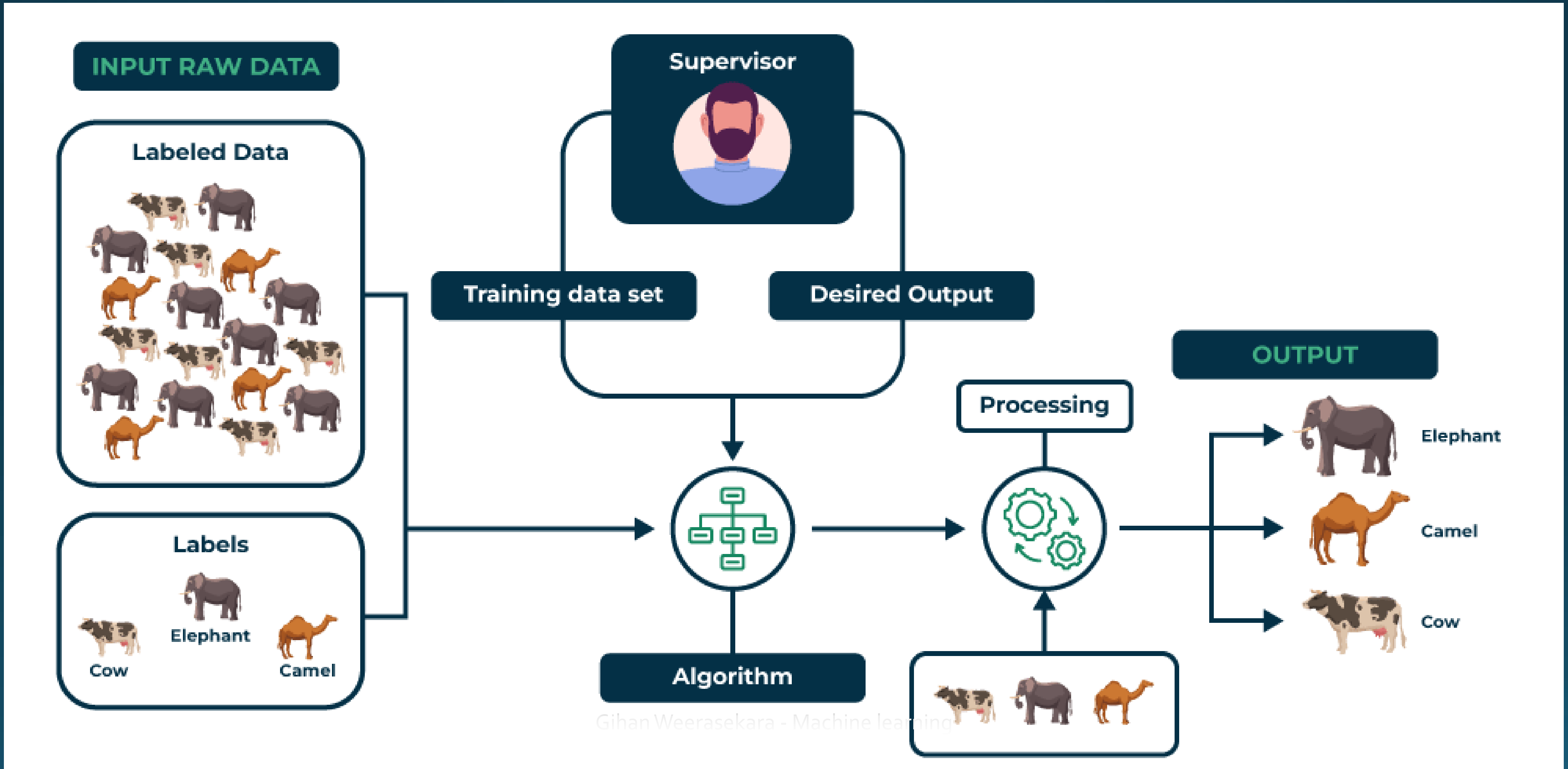
A linear regression model might have high bias and underfit, while a deep neural network with too many parameters might have high variance and overfit.

The tradeoff is about finding a model complexity that minimizes both.

Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning



Supervised Learning

The most commonly used type of machine learning, where the algorithm learns from labeled data.

The model is trained on a labeled dataset,

which means that each training example is paired with the correct output (label).

The goal is for the model to learn a mapping from input to output and make accurate predictions for unseen data.

Supervised Learning

Example:

Consider a dataset of house features (size, number of bedrooms, location) with the corresponding house prices. The model learns to predict the price of a house given its features.

Algorithms:

Common supervised learning algorithms include linear regression, decision trees, support vector machines (SVMs), and neural networks.

Supervised Learning

Real-World Applications:

- Spam email classification
- Image recognition
- Predictive analytics in finance (e.g., stock price predictions).