Machine Learning to Deep Learning

Randil Pushpananda

University of Colombo School of Computing

No 35, Reid Avenue, Colombo 07

rpn@ucsc.cmb.ac.lk

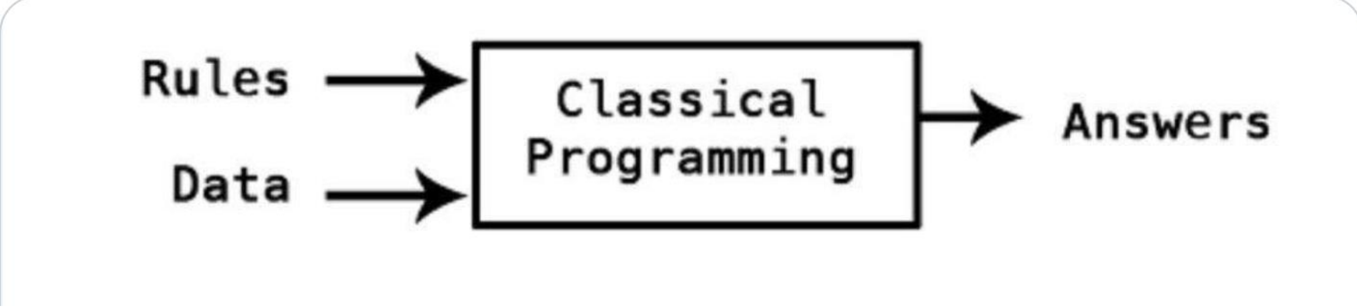
Natural Language Processing

Rule Based and Statistical Approaches

- Rule Based Approaches:

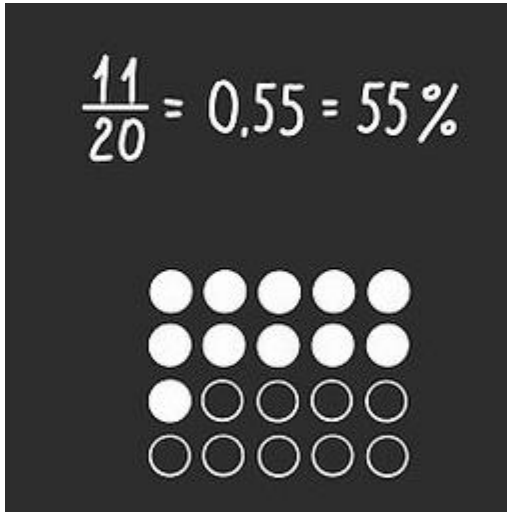
Expensive

Time Consuming



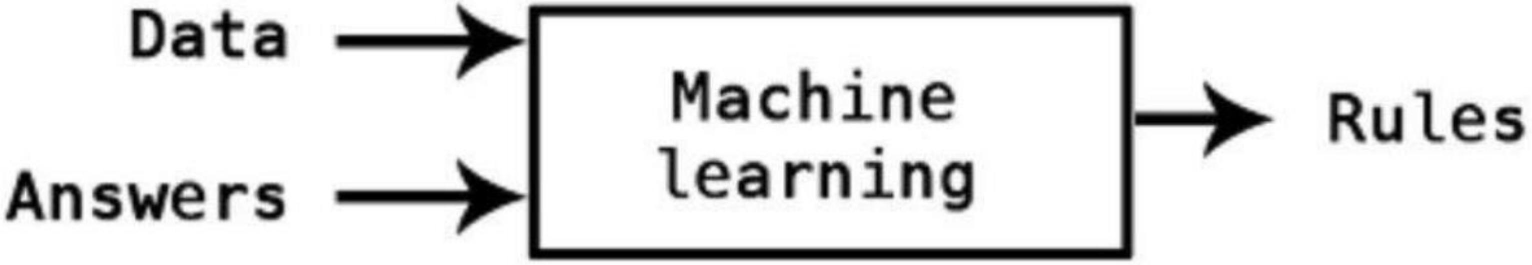
- Statistical Based Approaches:

Probabilities



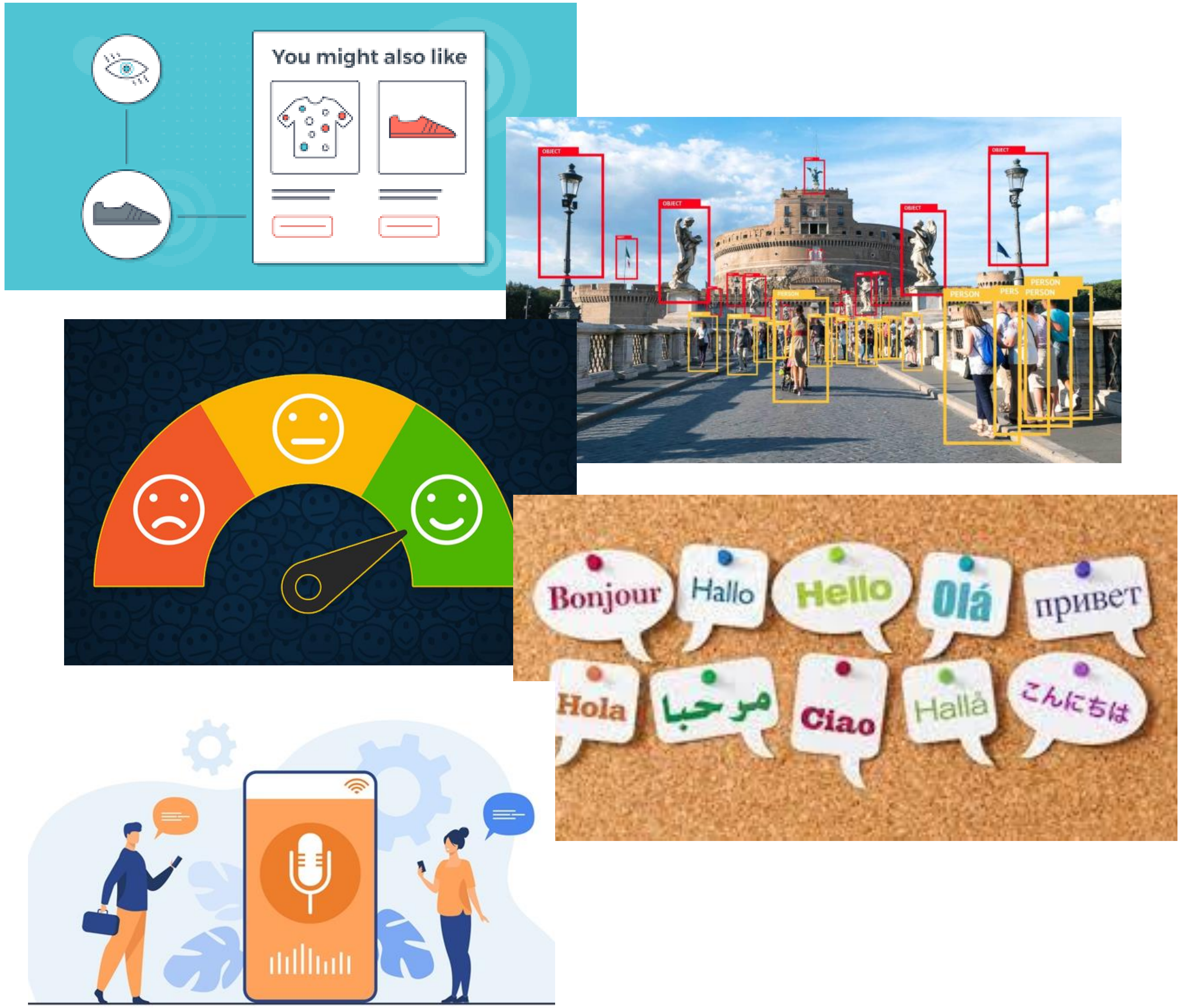
What is Machine Learning

- Machine learning is about extracting knowledge from data.



Applications of Machine Learning

- Product Recommendations
- Image Recognition
- Sentimental Analysis
- Language Translation
- Speech Recognition

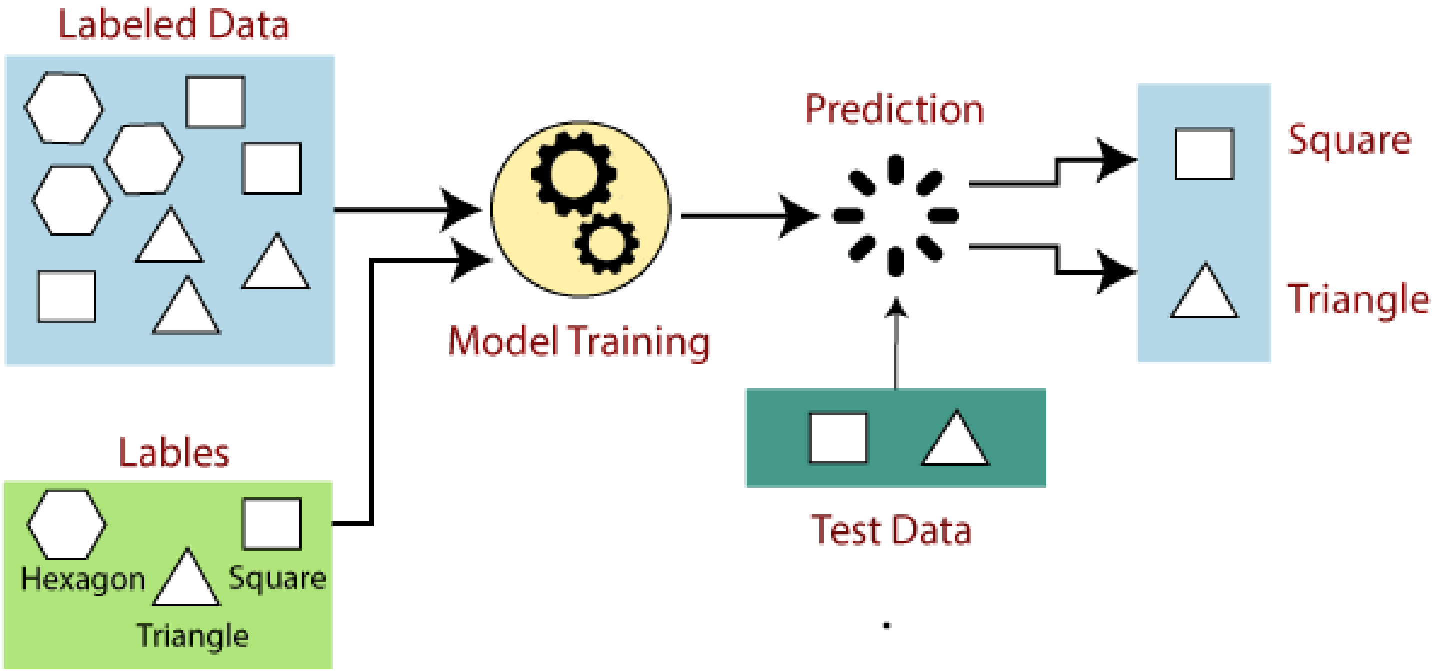


Problems Machine Learning Can Solve

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning

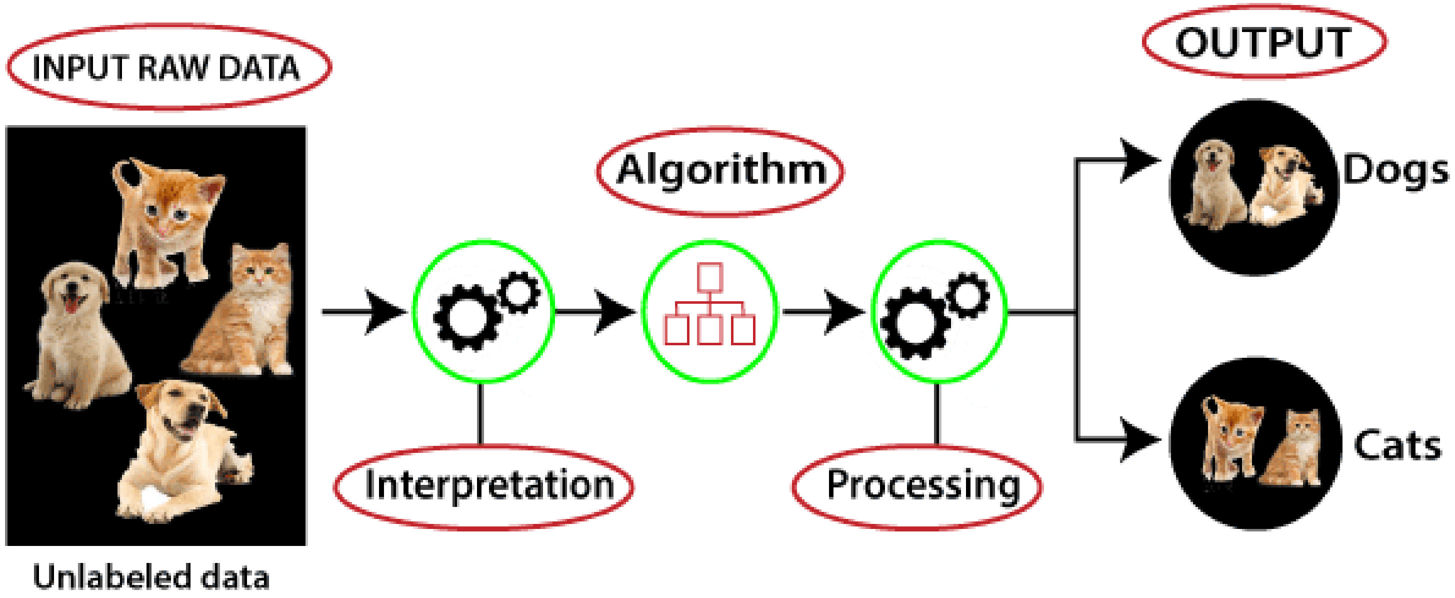
Supervised Learning:

- Machine learning algorithms to automate decision-making processes by analyzing known examples.



Unsupervised Learning

- Only the input data is known, and no known output data is given to the algorithm.
- Models itself find the hidden patterns and insights from the given data.



Importance of Unsupervised Learning

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

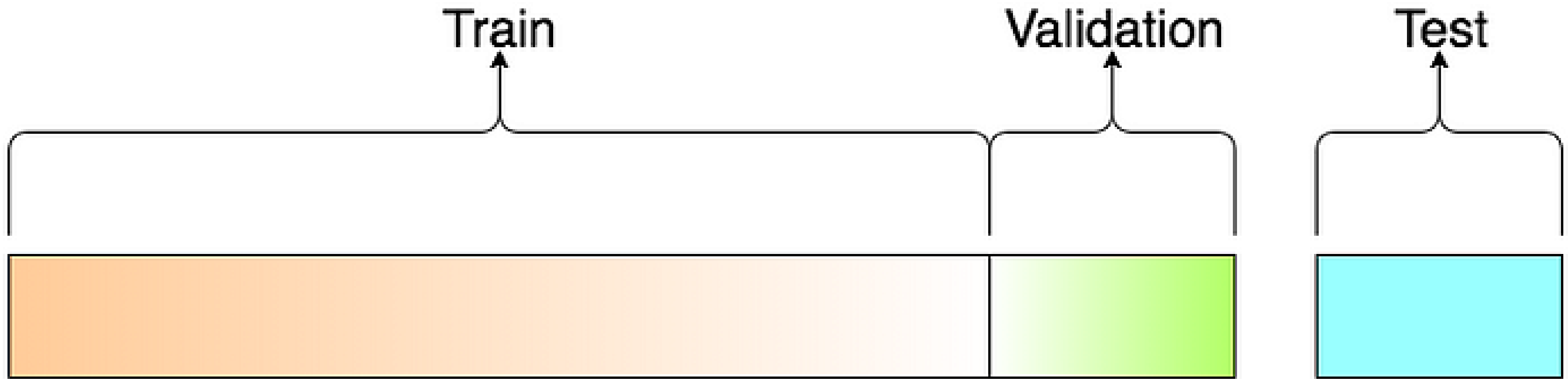
First Things First: Look at Your Data

REMEMBER

- We cannot use the data we used to build the model to evaluate it since the developed model always remembers the whole training dataset.
- Target is to generalize the model (Check whether the model performs well on new data)
- Split Dataset:
  - Training
  - Validation
  - Testing

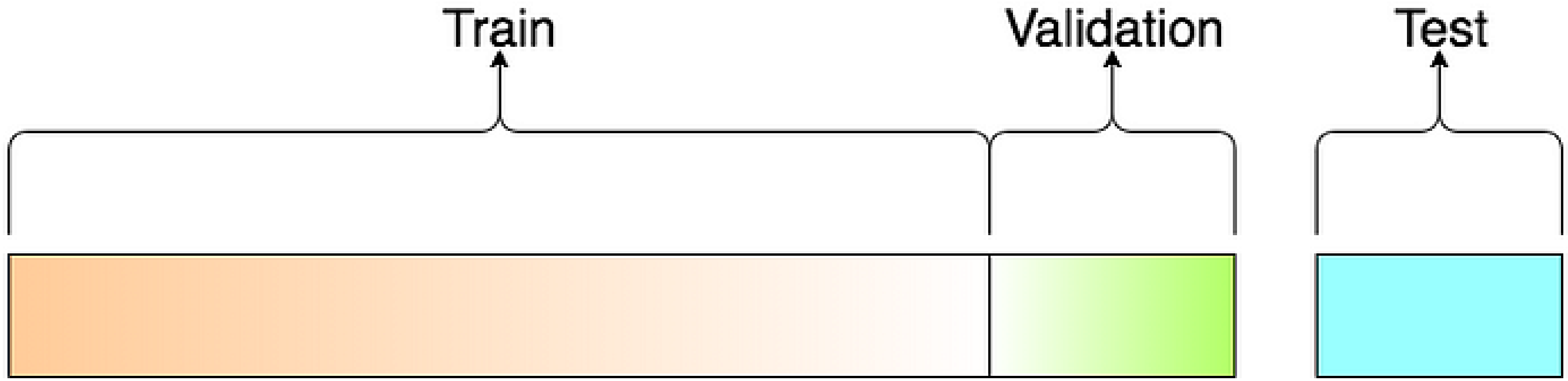
Training Dataset

- One part of the data is used to build our machine learning model
- The model *sees* and *learns* from this data.



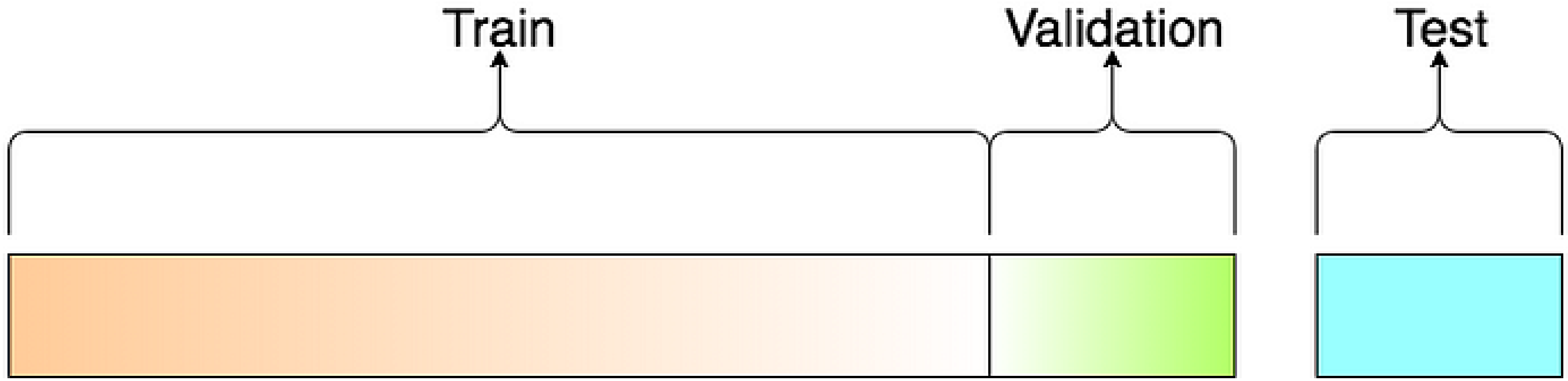
Validation Dataset

- The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
- Use this data to fine-tune the model hyperparameters.
- The model occasionally sees this data, but never does it “Learn” from this.



Test Dataset

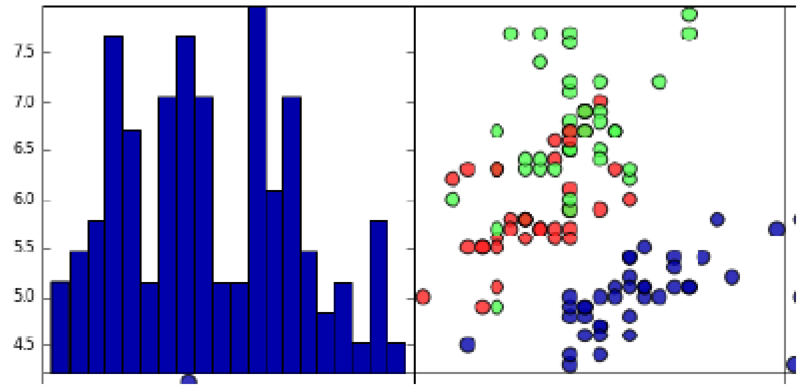
- Used to assess how well the model works.
- Used to provide an unbiased evaluation of a final model fit on the training dataset
- Only used once a model is completely trained(using the train and validation sets)



First Things First: Look at Your Data

REMEMBER:

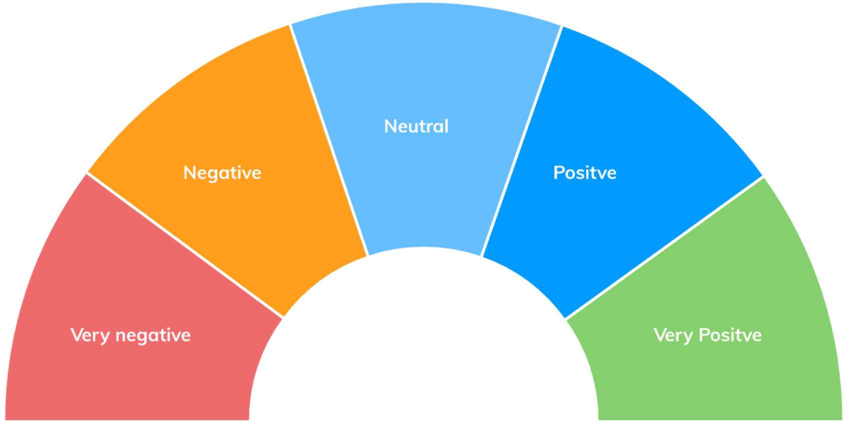
- Before building a machine learning model it is often a good idea to inspect the data, to see if the task is easily solvable without machine learning, or if the desired information might not be contained in the data.
- One of the best ways to inspect data is to visualize it.
    - Scatter Plots
    - Box and Whisker Plot for Large Data
    - Pie and Donut Charts



First Things First: Look at Your Data

Text Visualization/Analysis:

- Word Clouds
- Topic classification
- Sentiment Analysis/Aspect Based Sentiment Analysis



What kind of Data?

How much Data?

How to collect a Balanced Dataset?

Study the Dataset and Identify Features

How to divide Training, Validation and Testing Data?

What are the preprocessing steps required?

Features of a Dataset

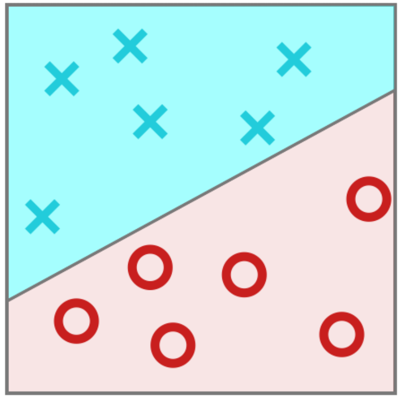
- garbage in → garbage out
- System will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones.
- A critical part of the success of a machine learning project is coming up with a good set of features to train on. This process, called *feature engineering*
  - Feature selection
  - Feature extraction
  - Creating new features by gathering new data

Machine learning is great for:

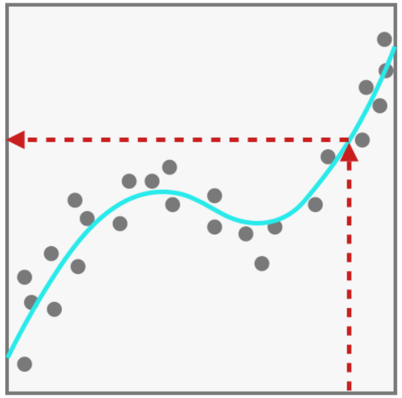
- Problems for which existing solutions require a lot of fine-tuning or long lists of rules (a machine learning model can often simplify code and perform better than the traditional approach)
- Complex problems for which using a traditional approach yields no good solution (the best machine learning techniques can perhaps find a solution)
- Fluctuating environments (a machine learning system can easily be retrained on new data, always keeping it up to date)
- Getting insights about complex problems and large amounts of data

Supervised Learning

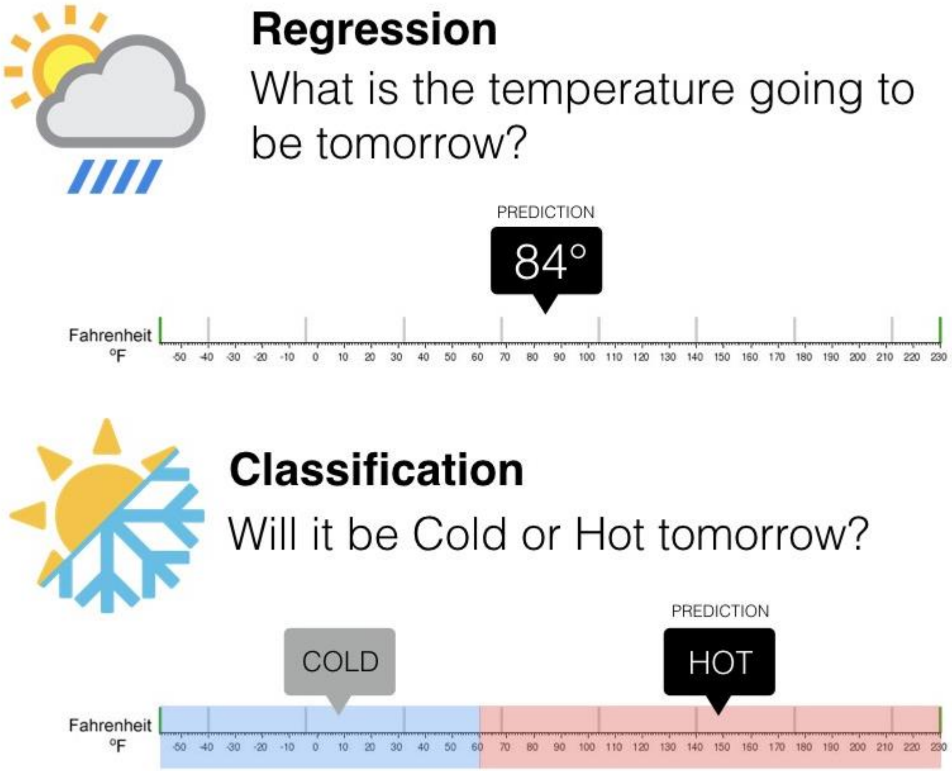
- There are two major types of supervised machine learning problems:
  - Classification
  - Regression.



Classification



Regression



Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

Classification

- Goal is to predict a class label, which is a choice from a predefined list of possibilities

Example -
“I really like the new design of your website!” → Positive
“The new design is awful!” → Negative

Anger
Happines
Surprise
Disgust
Sadness
Fear

Source: https://monkeylearn.com/

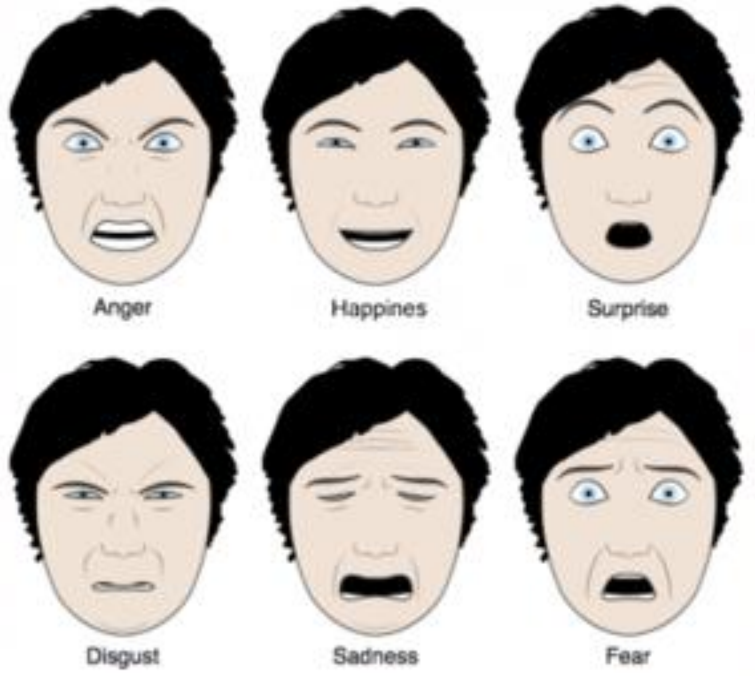
Classification

- Goal is to predict a class label, which is a choice from a predefined list of possibilities

Example -
“I really like the new design of your website!” → Positive
“The new design is awful!” → Negative

Binary Classification

Multi-Class Classification



Source: https://monkeylearn.com/

Regression

- Goal is to predict a continuous number, or a floating-point number in programming terms (or real number in mathematical terms).
- Example:
  - Predicting a person’s annual income from their education, their age, and where they live.
- An easy way to distinguish between classification and regression tasks is to ask whether there is some kind of continuity in the output. If there is continuity between possible outcomes, then the problem is a regression problem.

Generalization, Overfitting, and Underfitting

- Supervised Learning, build a model on the training data and then be able to make accurate predictions on new, unseen data.
- If a model is able to make accurate predictions on unseen data

generalize from the training set to the test set

Overfitting: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data

Underfitting: A statistical model is said to have underfitting when it neither performs well on the training data and new data (Test Data)

Overfitting - Reasons

- The training data size is too small and does not contain enough data samples to accurately represent all possible input data values.
- The training data contains large amounts of irrelevant information, called noisy data.
- The model trains for too long on a single sample set of data.
- Not used regularization technique.

Underfitting - Reasons

- The size of the training dataset used is not enough.
- The model is too simple.
- Training data is not cleaned and also contains noise in it.
- Need more training time/more epochs with more features

Learning Algorithms for Text Analytics

- Text Classification:
  - categorize text documents into predefined categories or classes

Machine Learning

- Naive Bayes
- Support Vector Machines (SVM)
- Decision Trees

Deep Learning

- Recurrent Neural Networks (RNNs)
- Convolutional Neural Networks (CNNs)

Learning Algorithms for Text Analytics

- Sentiment Analysis:
  - aims to determine the sentiment or emotional tone expressed in text

Machine Learning

- Logistic Regression
- Random Forests

Deep Learning

- Recurrent Neural Networks (RNN)
  - Long-Short Term Memory (LSTM)
- Transformers (e.g., BERT)

Learning Algorithms for Text Analytics

- **Named Entity Recognition (NER):**
  - identifying and classifying named entities (such as person names, organizations, locations) in text

Machine Learning

- Conditional Random Fields (CRF)
- Hidden Markov Models (HMM)

Deep Learning

- Recurrent Neural Networks (RNN)
- Transformers (e.g., BERT)

Learning Algorithms for Text Analytics

- Topic Modeling:
  - aims to uncover latent topics within a collection of documents.

Machine Learning

- Latent Dirichlet Allocation (LDA)

Deep Learning

- Variational Autoencoders (VAEs)
- Generative Adversarial Networks (GANs)

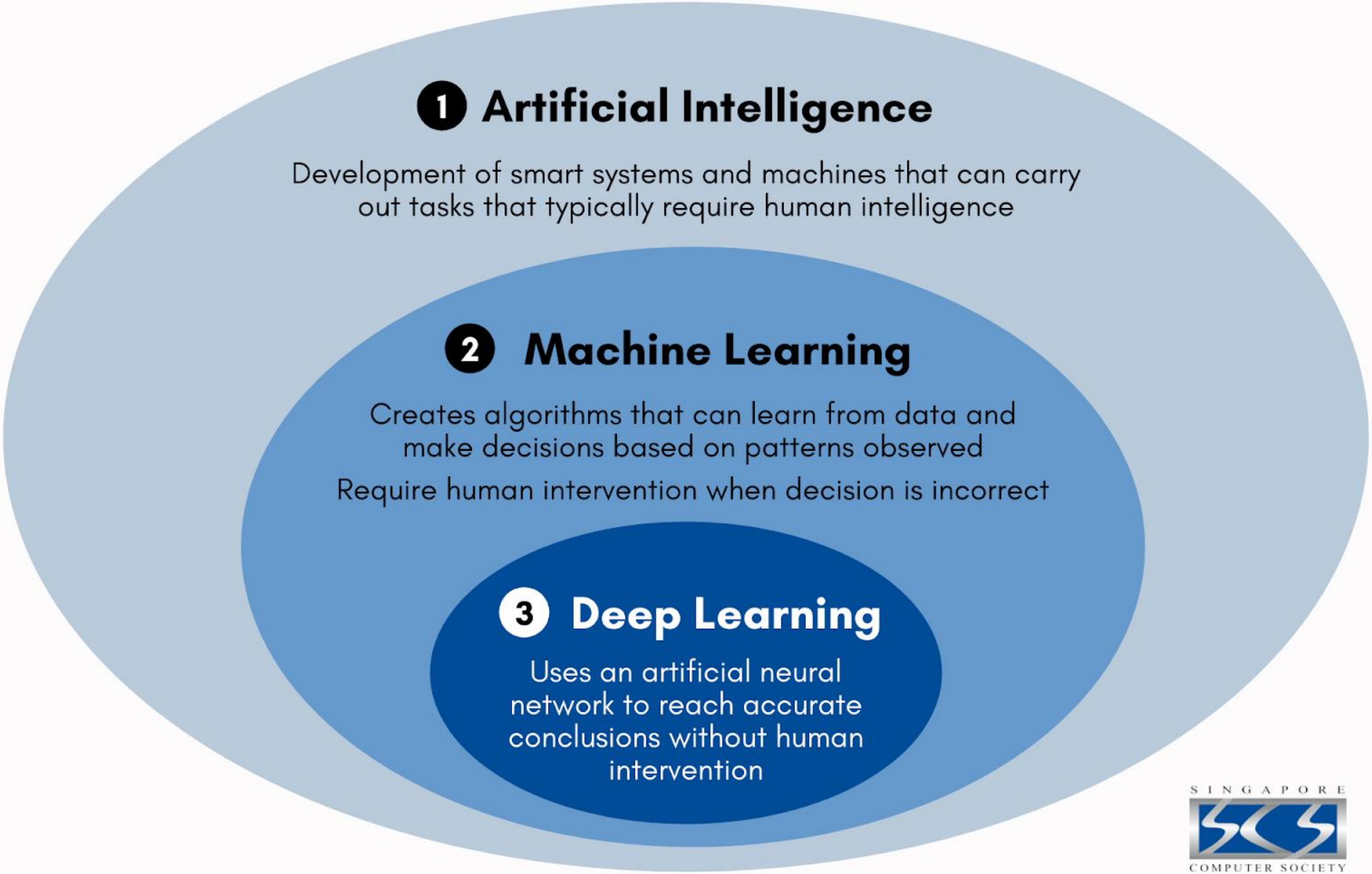
Learning Algorithms for Text Analytics

- **Text Generation:**
  - Text generation involves generating human-like text based on given prompts or contexts.

Deep Learning

- Recurrent Neural Networks (RNN)
  - LSTM
  - GRU (Gated Recurrent Unit)
- Transformers
  - GPT (Generative Pre-trained Transformer)

ARTIFICIAL INTELLIGENCE VS MACHINE LEARNING VS DEEP LEARNING



1 Artificial Intelligence

Development of smart systems and machines that can carry out tasks that typically require human intelligence

2 Machine Learning

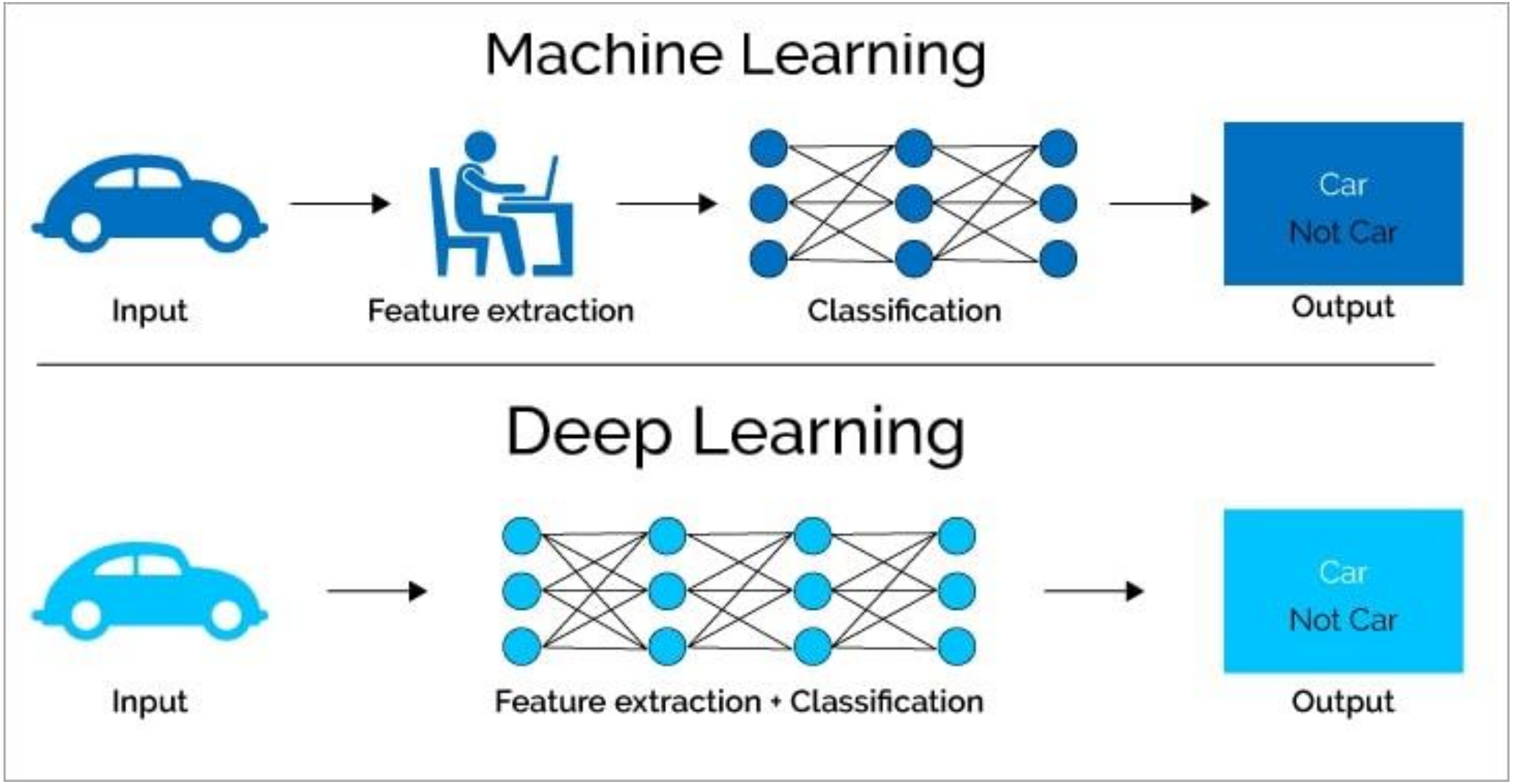
Creates algorithms that can learn from data and make decisions based on patterns observed

Require human intervention when decision is incorrect

3 Deep Learning

Uses an artificial neural network to reach accurate conclusions without human intervention

SINGAPORE SCS COMPUTER SOCIETY


Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction + Classification
Car
Not Car
Output