Bukalapak

Natural Language Processing

Basic Class

Afif A. Iskandar

About Me

Name : Afif Akbar Iskandar

Role : AI Scientist

Company : Bukalapak

Specialization :

- Computer Vision
- Machine Learning
- Deep Learning
- Natural Language Processing

About Me

Educational Background :

- Bachelor of Mathematics at Universitas Indonesia (2011)
- Master of Computer Science at Universitas Indonesia (2015)

Working Experience :

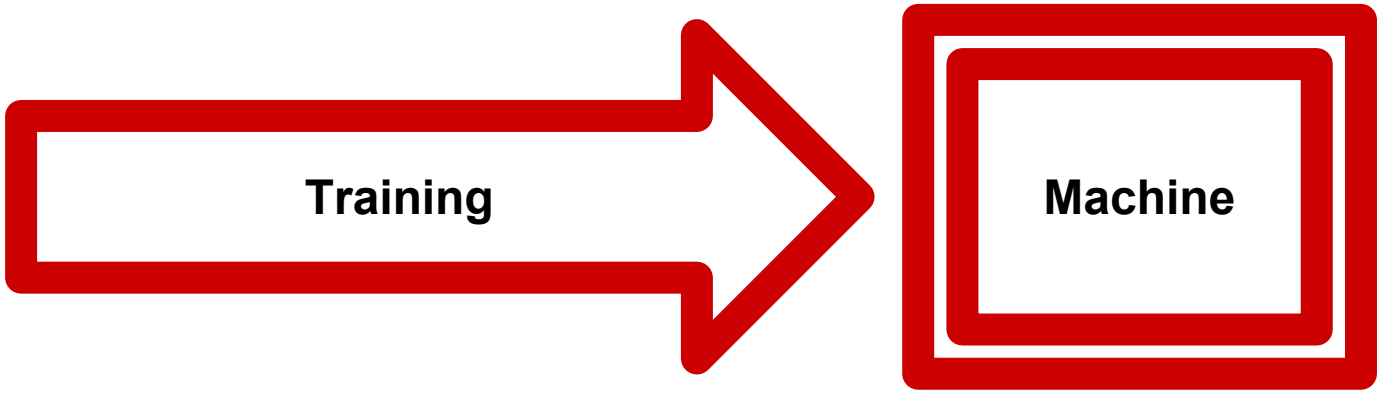
- Data Scientist (2015-Now)

OUTLINE

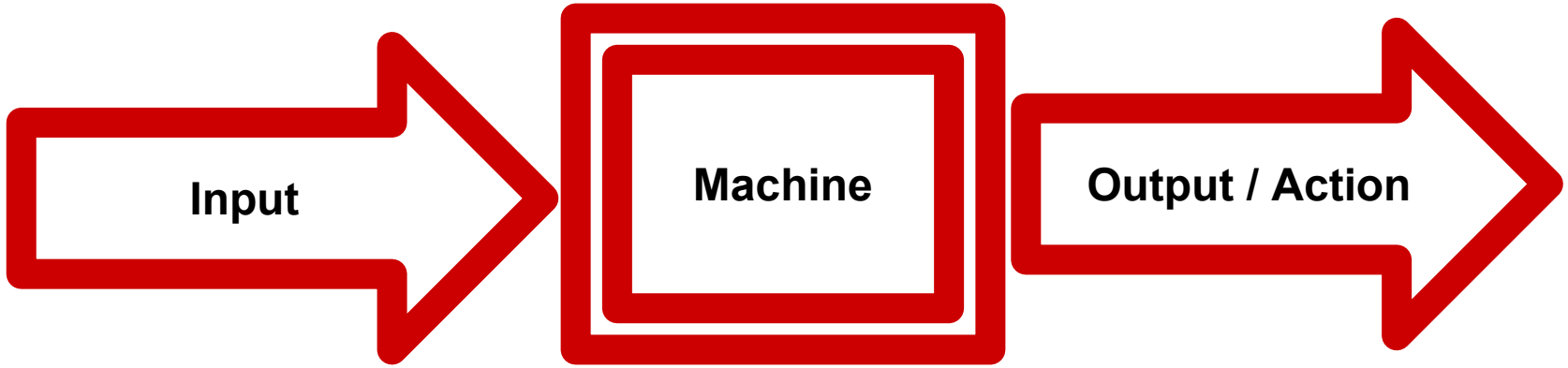
- Machine Learning Review
- Text Preprocessing
- Text Classification (Spam Detector)

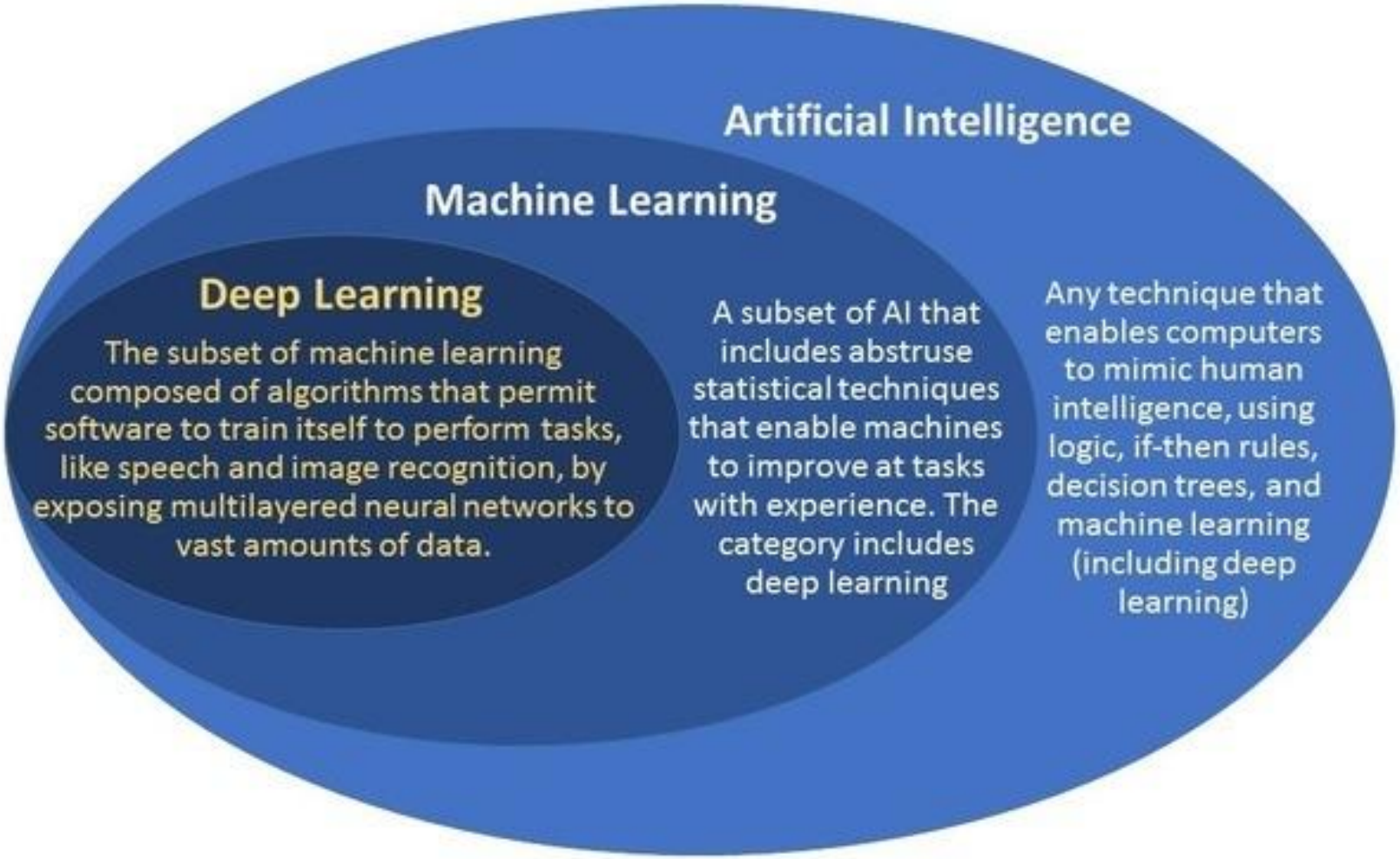
A very simplistic view of Machine Learning

Let’s get the machine to learn stuff, by training it thousands times, million times, billion times



...after it finished learning from the intensive training...




Artificial Intelligence
Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)
Machine Learning
A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning
Deep Learning
The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.

Type of Machine Learning

Supervised Learning

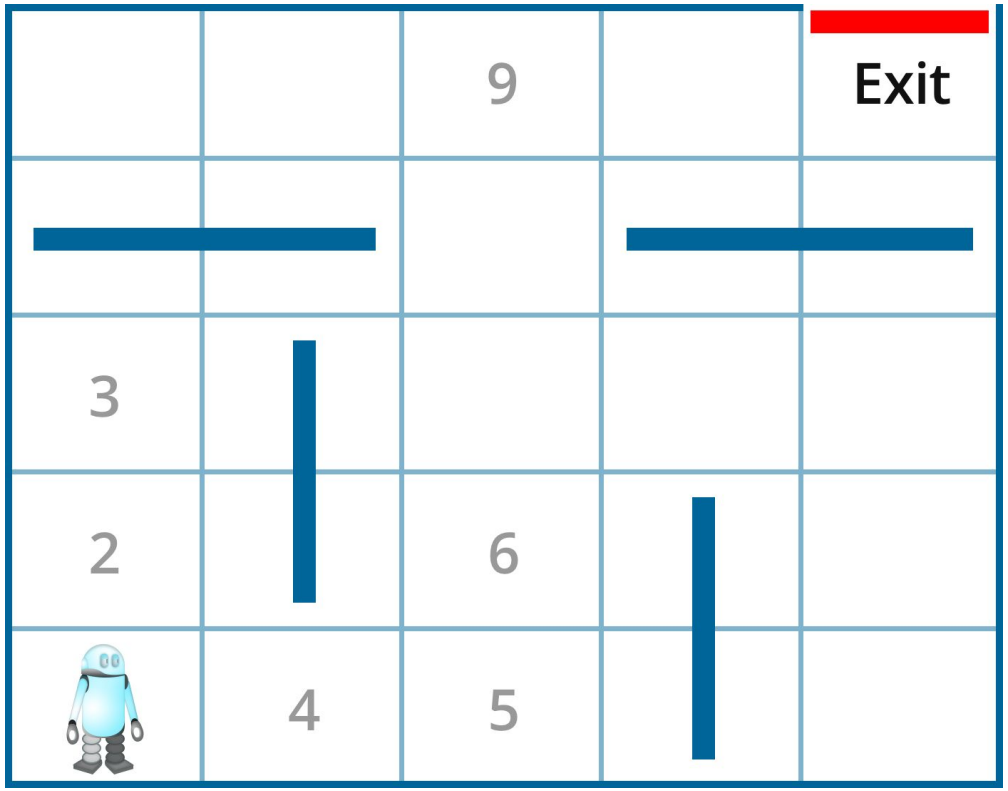
Laptop

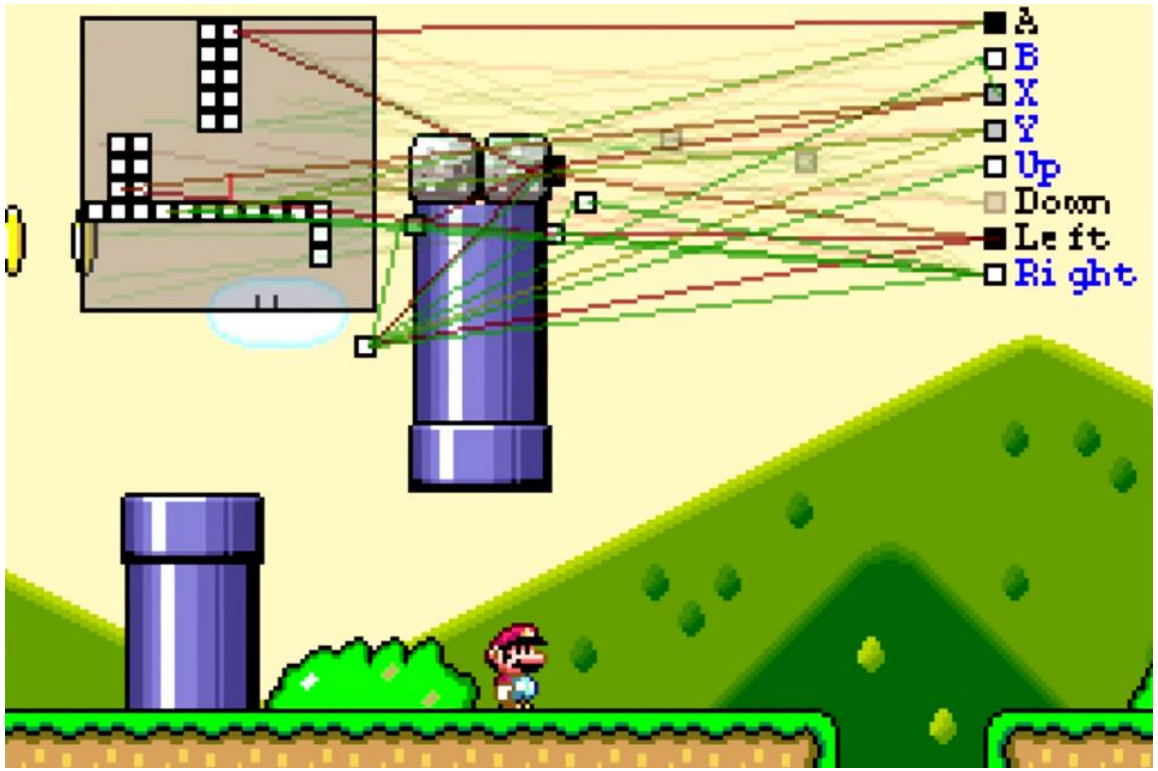
Tablet

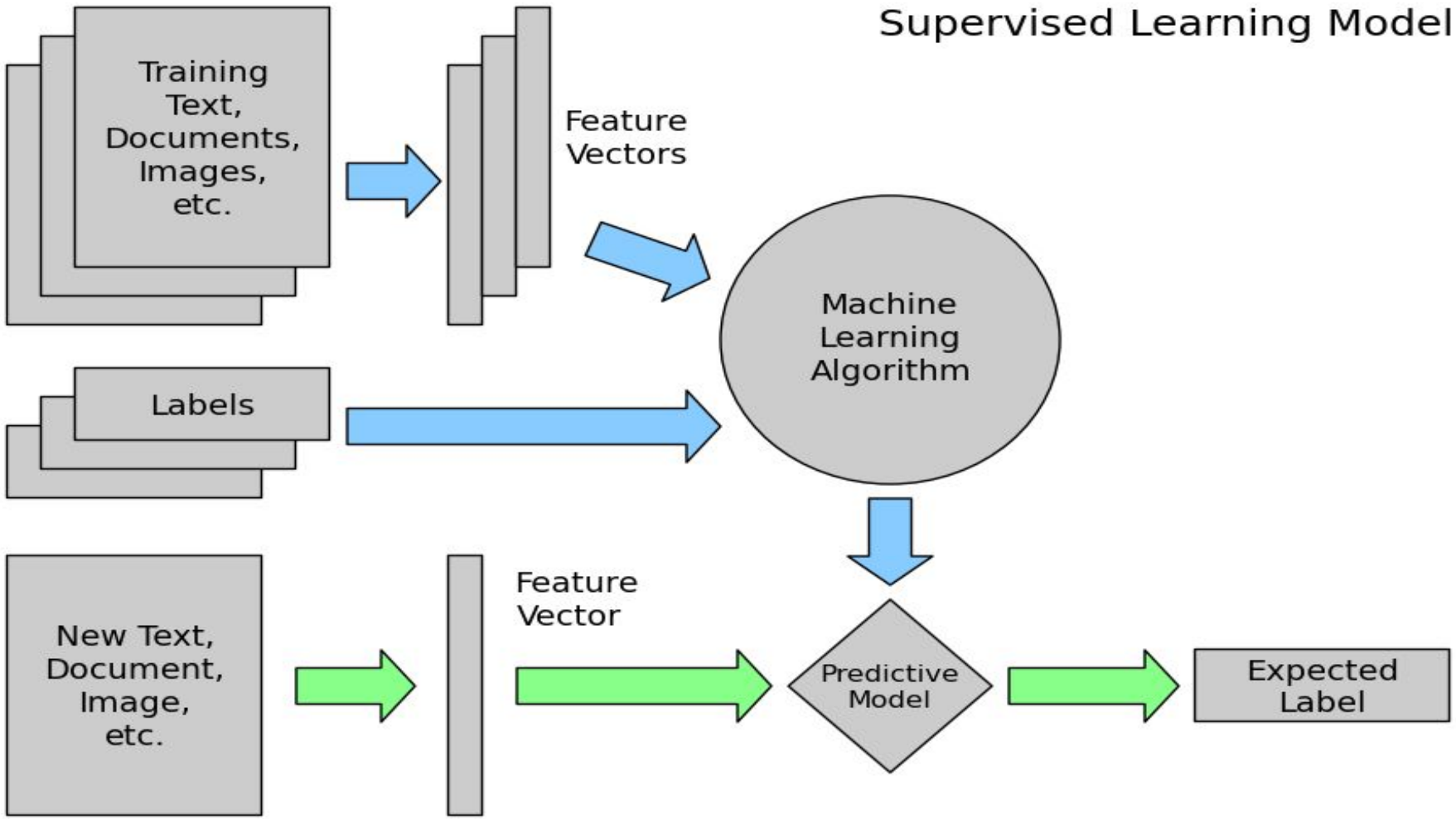
???

Unsupervised Learning

Reinforcement Learning






Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Extracting Features from Text

Bag of Words Model

- Count Vectorizer
- Term Frequency–Inverse Document Frequency

Tf-Idf

- TF: Term Frequency, which measures how frequently a term occurs in a document

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

- IDF: Inverse Document Frequency, which measures how important a term is

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

Spam Detector using Supervised Learning

Go to Jupyter Notebook

Thank You

BukaLapak