# Understanding Machine Learning

Dr. Sunil Thomas T

College of Engineering Attingal

*suniltt@gmail.com*
http://brainstorms.in
https://www.facebook.com/vu2swx

August 20, 2016

# Buzz words !!

1. Machine Learning
2. Data Mining
3. Pattern Recognition
4. Big Data
5. Recommender systems

# Who uses it?

Lot of companies. In fact there is a growing requirement for ML/Data sciences professionals world wide.
Look at

Look at the ads. They are customized for you. The two web giants look at what you share or browse and learn about your habits, tastes and preferences. They then customize their advertisements based on this learning.

# This talk ?

There are two parts in this workshop/

- I will give an informal introduction to machine learning in the first part
- Then I will introduce you to R, a popular tool used by machine learning/data sciences community. We will then write some simple machine learning programs in R.

I don't expect you to be hard core programmers. But a basic familiarity with programming is assumed. I will give some code that you can try right now.

# Can Machines Learn ? Or can we teach Machines

# How do we learn ?

# How do we train kids ?

# How do we train kids ?

# How do we test our learning ?

# How do we test our learning ?

# What if we make errors ?

# Retraining

# Our Learning cycle

- Training
- Testing
- Retraining

# How do we learn?

What kind of learning problems we encounter

- Classification
- Regression

How do we perform the above tasks.

- We learn using features of objects.

# Feature ?

What are the features ?
Depends on the problem you are trying to solve.

- For example in the Bus vs Car problem we can have some of the following as features.
  - No of wheels
  - Size
  - No of windows
  - .....
  - Lot of other things

# Now let us teach machines.

We need thousands of examples.

**Features features .... and features**

For each class of objects we need to find out good features

# How do we represent features?

Table: Features for car

| | Size | Weight | No of Wheels | No of Doors |
|---|---|---|---|---|
| Alto | 8 | 700 | 4 | 4 |
| i10 | 7.5 | 800 | 4 | 4 |
| Honda City | 9 | 950 | 4 | 4 |
| Tata Nano | 5 | 600 | 4 | 4 |

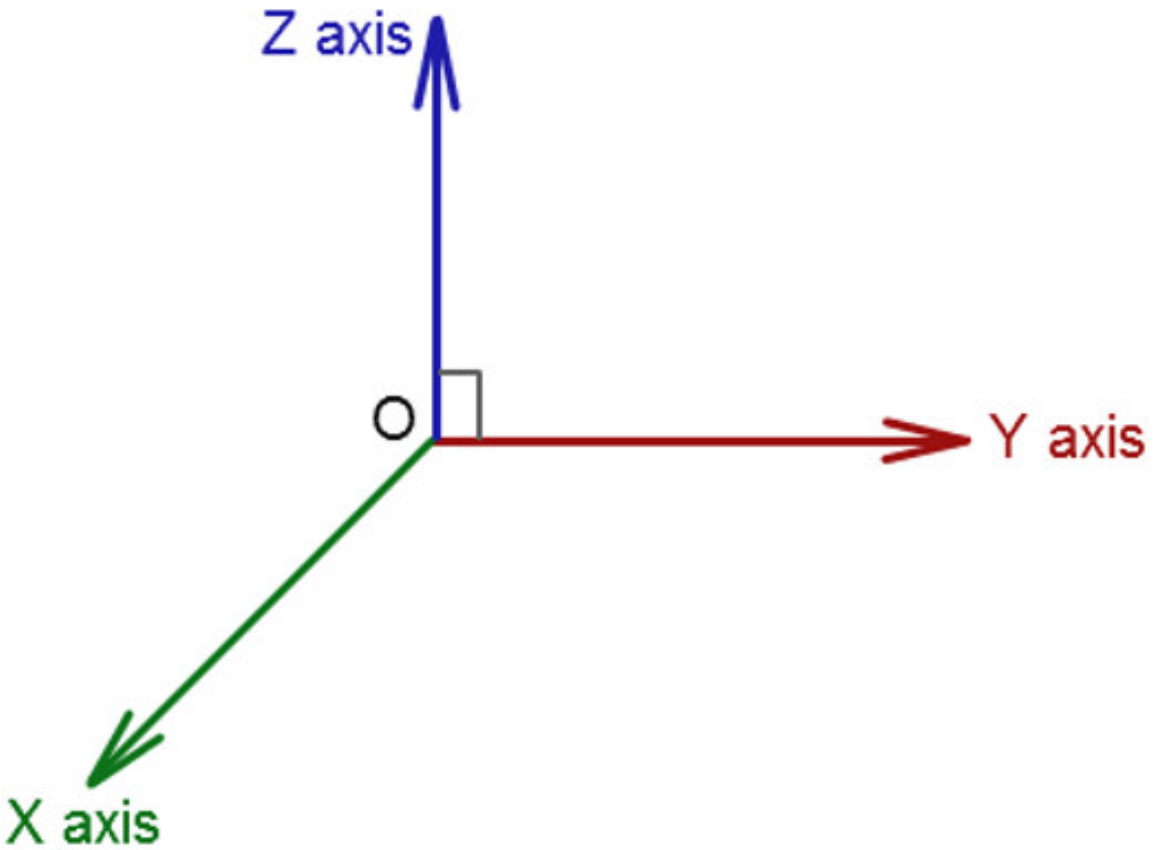
# How do we represent features?

Table: Features for bus

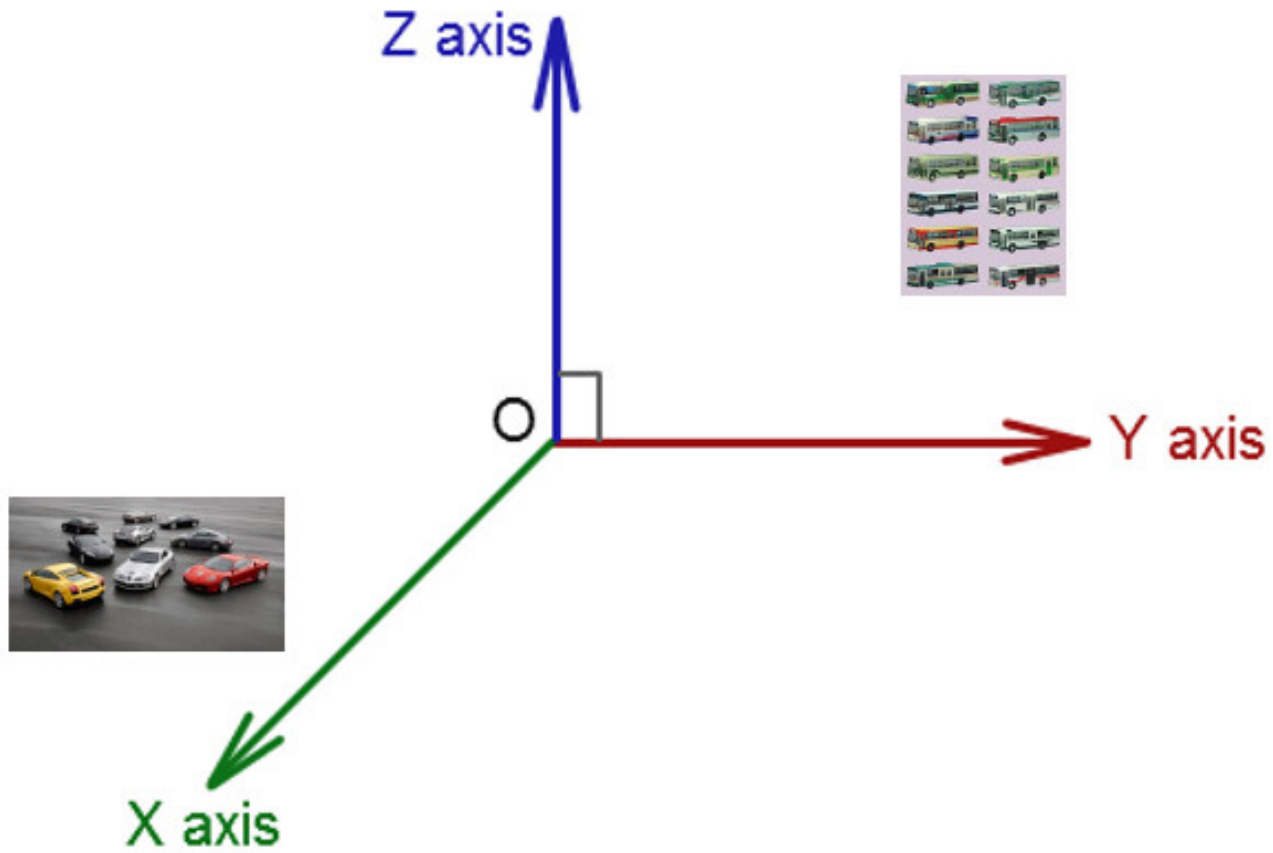
| | Size | Weight | No of Wheels | No of Doors |
|---|---|---|---|---|
| TATA | 20 | 1700 | 6 | 2 |
| Leyland | 16 | 1800 | 6 | 2 |
| Tata mini | 23 | 1950 | 6 | 3 |
| Volvo | 28 | 1600 | 6 | 3 |

# Feature Space

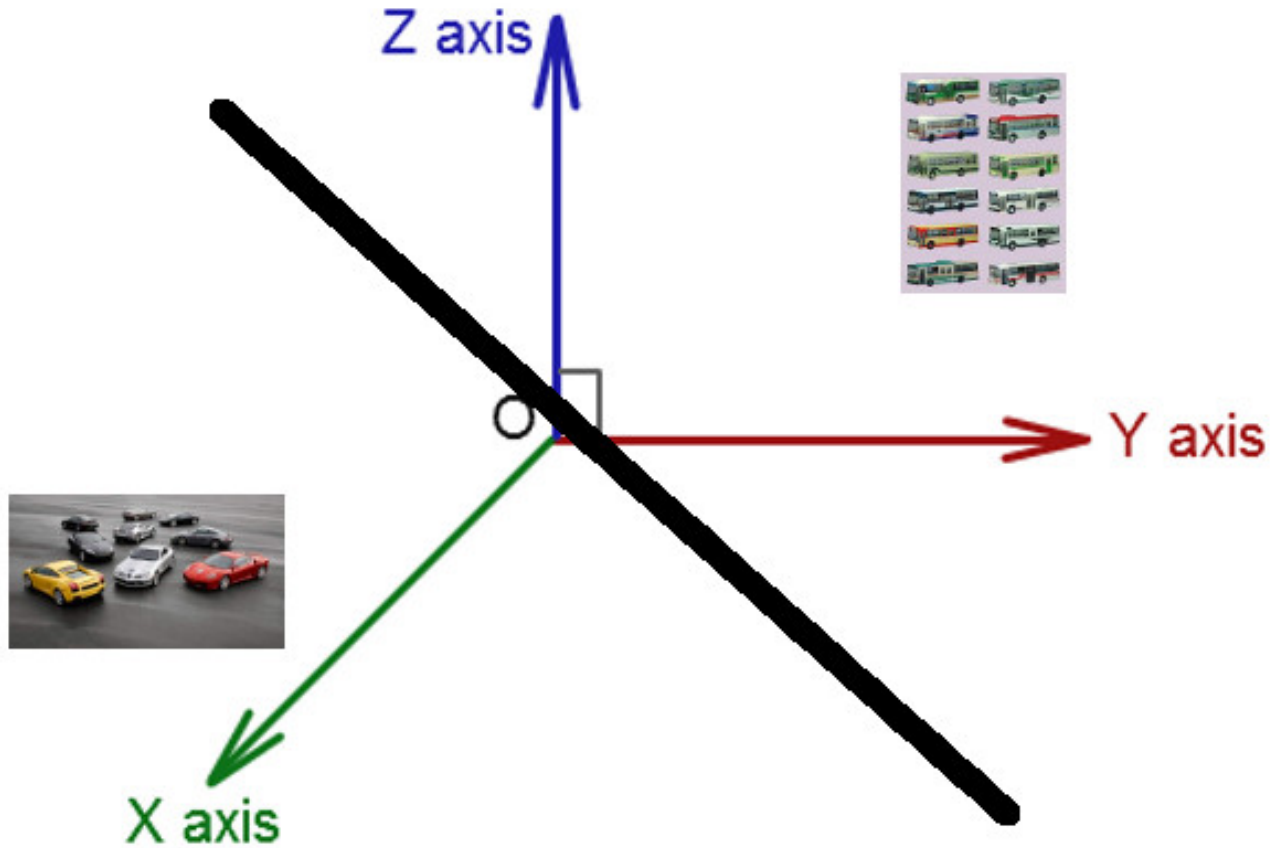
For simplicity I am showing a 3 dimensional feature space. But you can have as many dimensions as you want.



# Objects in Feature Space

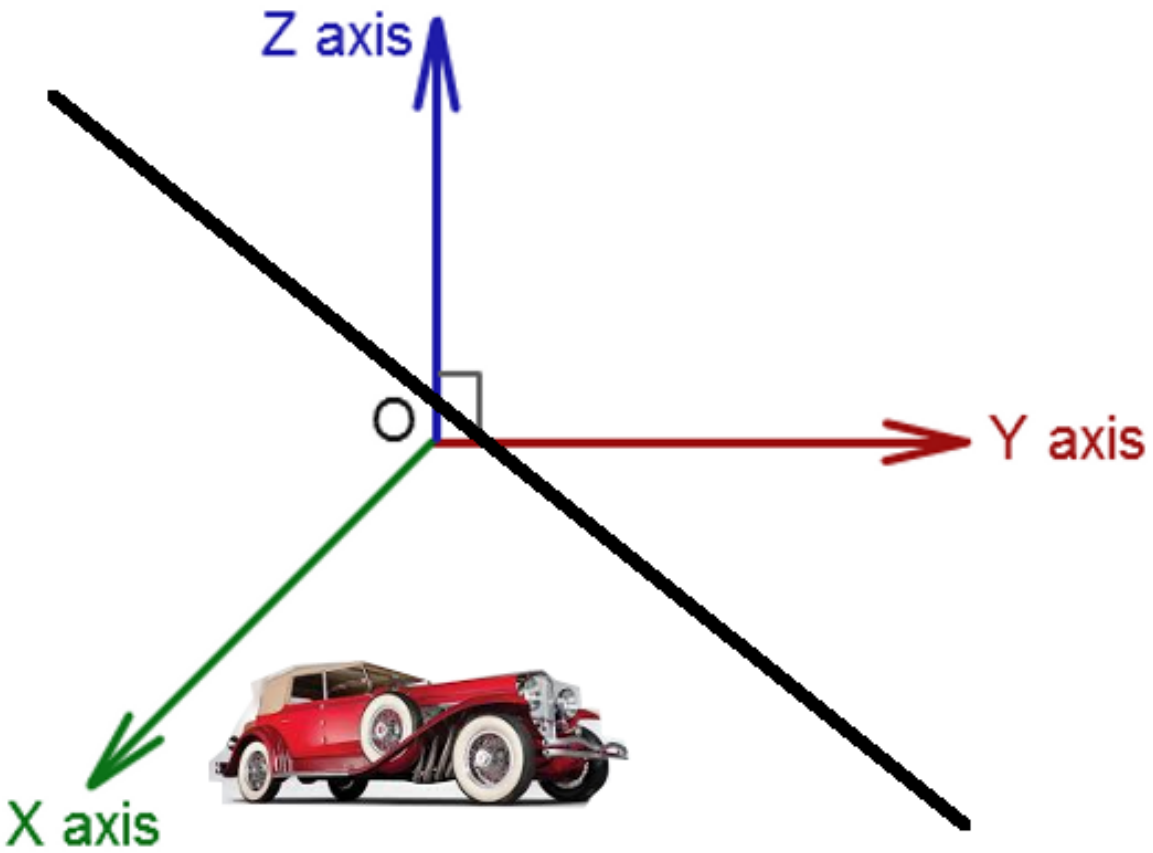


# Classification Problem

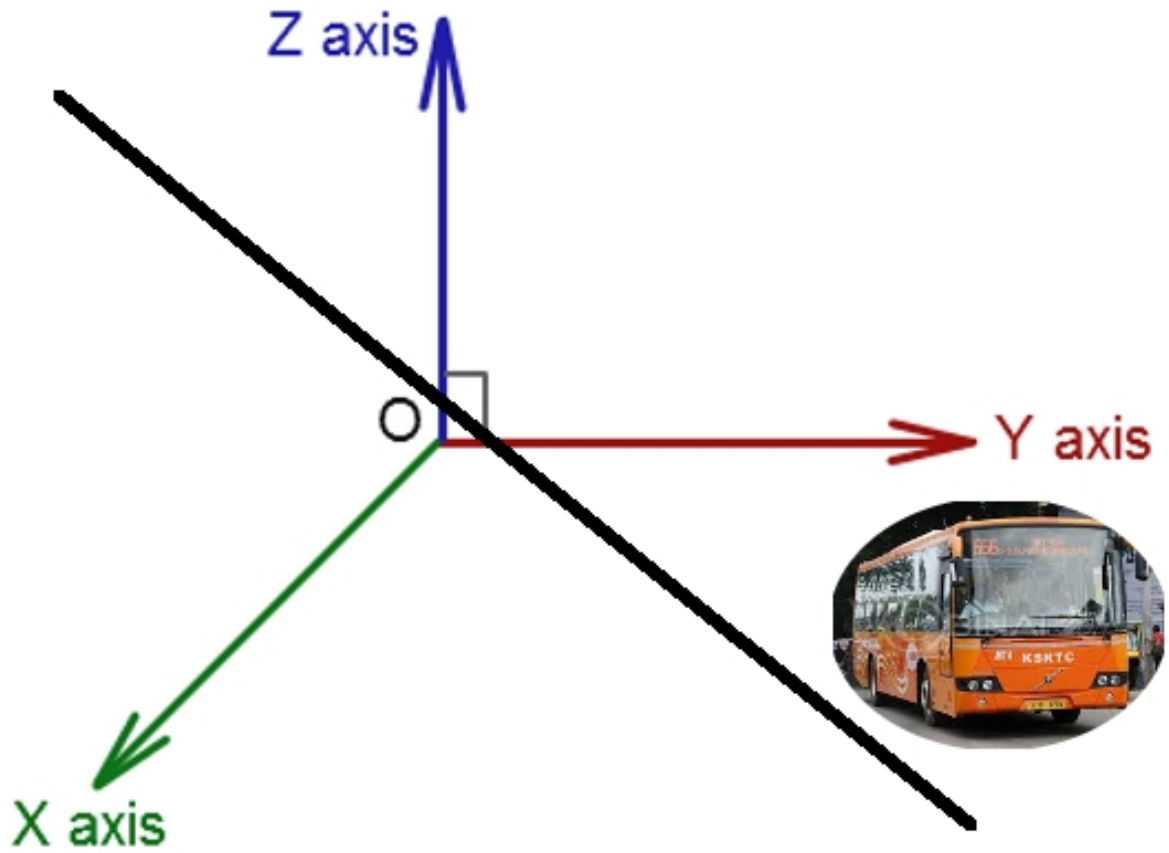



The black line above represents a plane splitting the feature space into two.

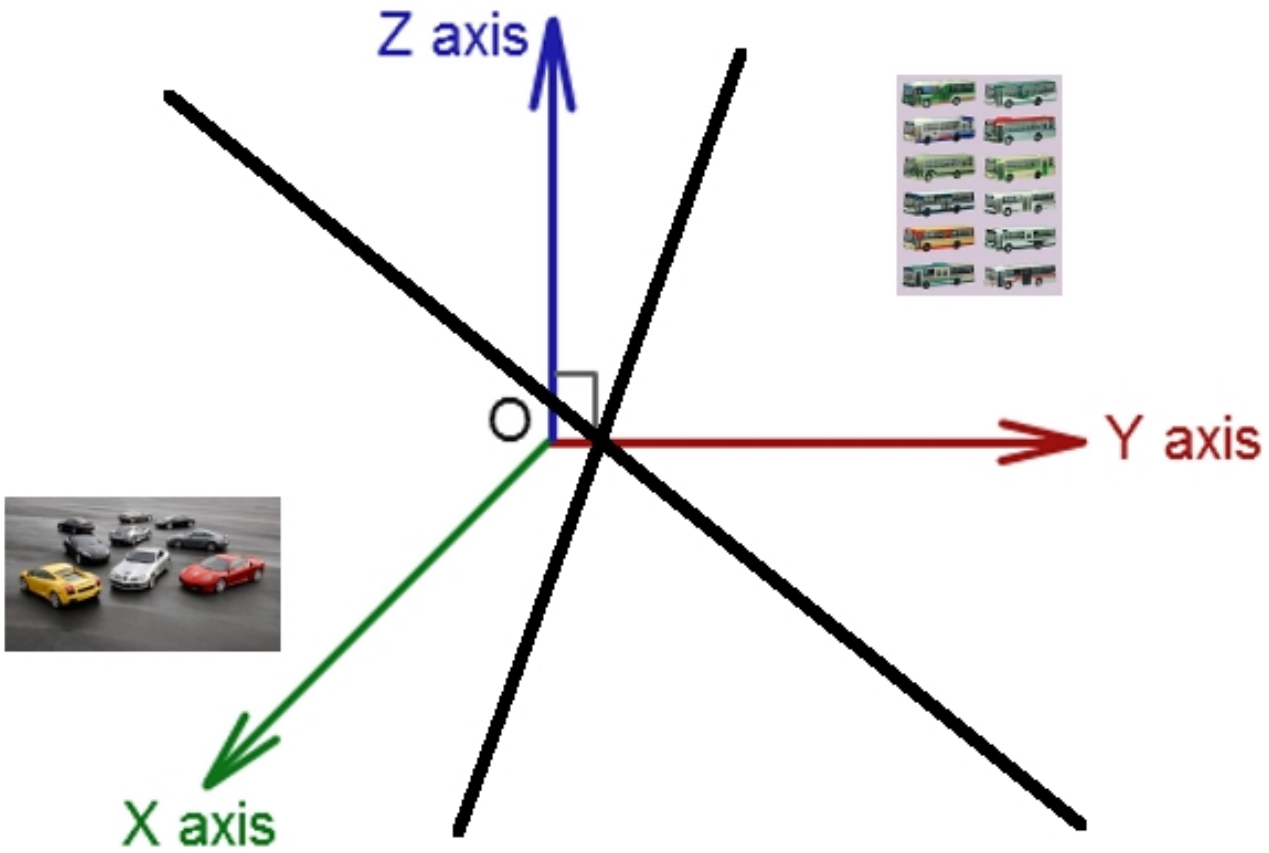
# Testing a classifier



# Testing a classifier



# Finding the best classifier-The Challenge



# How complicated is the math ?

A bit if you are working on theoretical side.
If you are planning to ML algorithms, to your problems,it is very easy. There are lot of libraries, open source codes and sample data sets.

# How complicated is the math ?

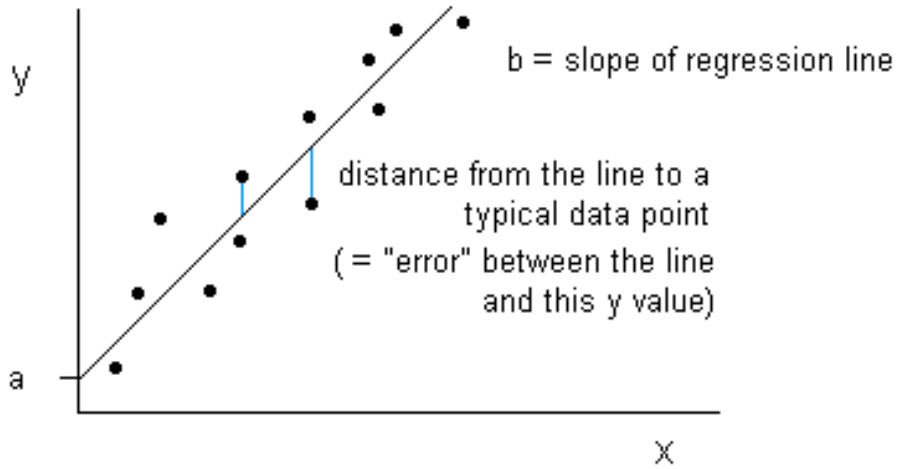
A bit if you are working on theoretical side.
If you are planning to ML algorithms, to your problems,it is very easy. There are lot of libraries, open source codes and sample data sets.

You need to know

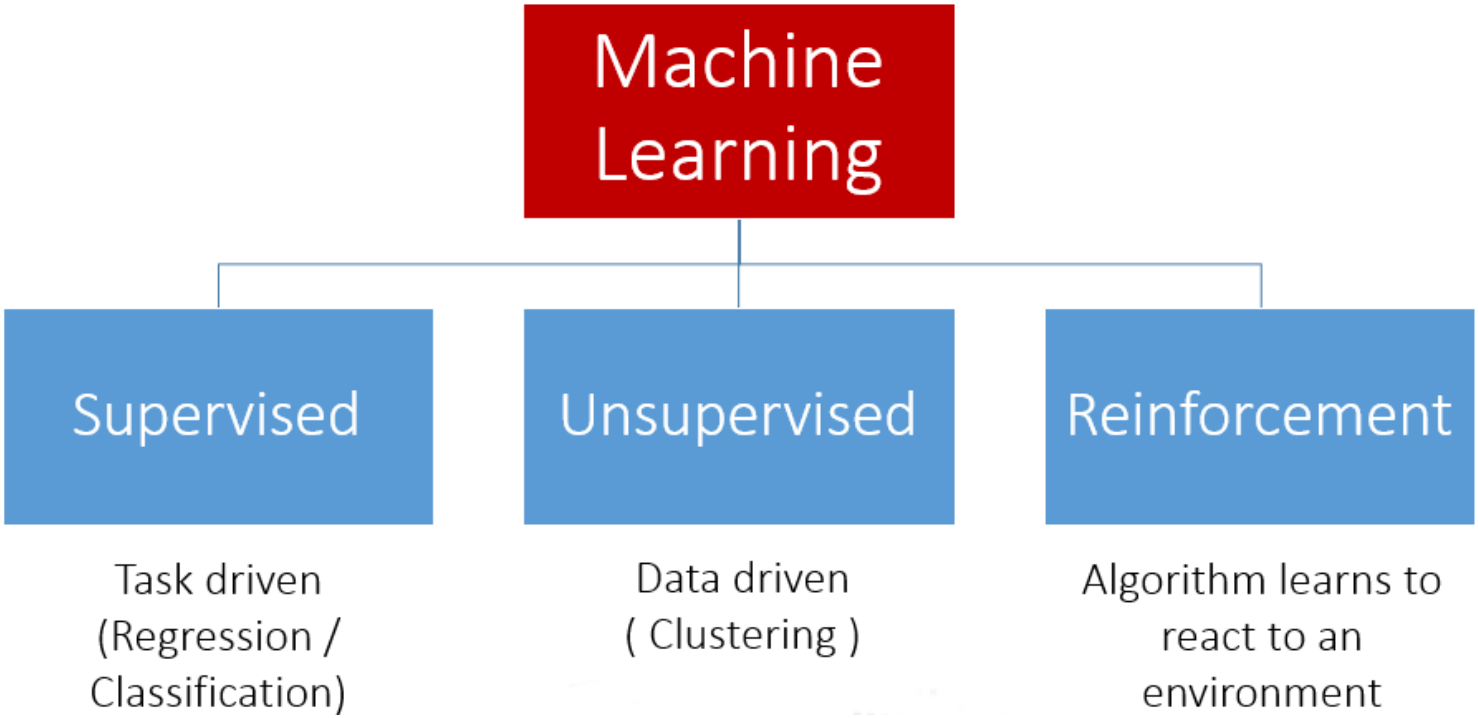
- Linear Algebra
- Probability Theory
- Optimization Theory

# Regression

You are give several data points. You have to find out a best fitting curve.



# Types of Machine Learning

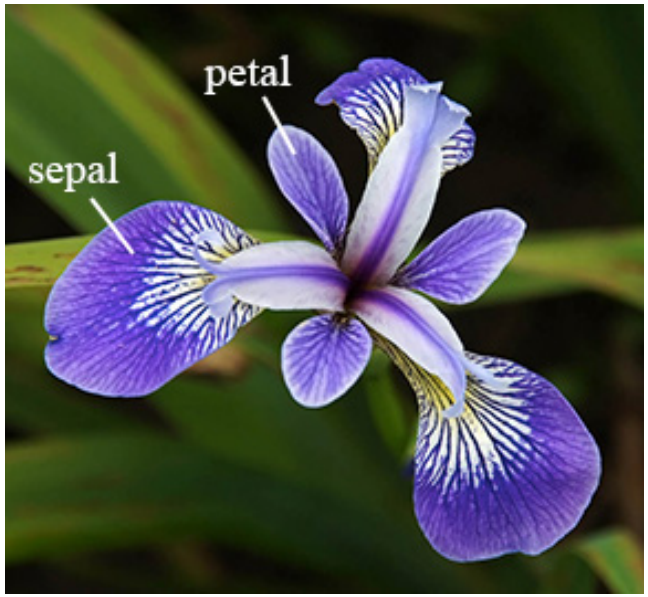


# Sample data sets

First let us look at some real data that ML community uses. There are several sources.
Public data sets for machine learning.
UC Irvine Machine Learning Repository
Middlebury CV data set
Lot more.

# iris data set

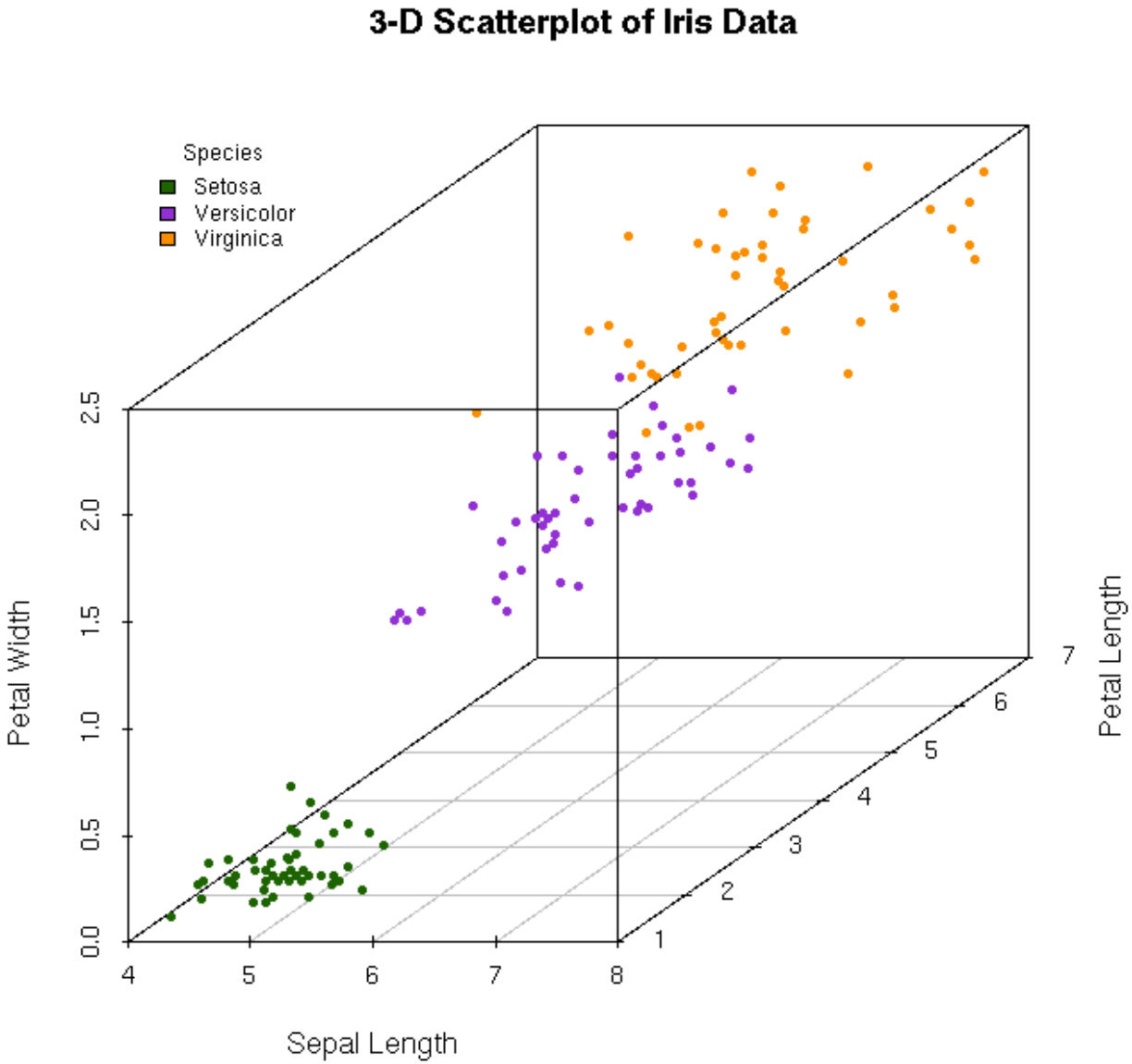
# iris data set



# iris data set

| sepal | | petal | | class |
|---|---|---|---|---|
| length | width | length | width | |
| 6.3 | 2.3 | 4.4 | 1.3 | versicolor |
| 6.2 | 3.4 | 5.4 | 2.3 | virginica |
| 5.2 | 3.4 | 1.4 | 0.2 | setosa |
| 6.9 | 3.1 | 5.4 | 2.1 | virginica |
| 5.7 | 4.4 | 1.5 | 0.4 | setosa |
| 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 5 | 3.3 | 1.4 | 0.2 | setosa |
| 6.4 | 2.8 | 5.6 | 2.1 | virginica |
| 6 | 3 | 4.8 | 1.8 | virginica |
| 5.5 | 2.5 | 4 | 1.3 | versicolor |

# iris data set



# Algorithms

1. k-Nearest Neighbors (kNN)
2. Support Vector Machines (SVM)
3. Expectation Maximization (EM)
4. Random forest
5. Naive Bayes
6. And many more

We will try out a few later

# What are the opportunities

## Research

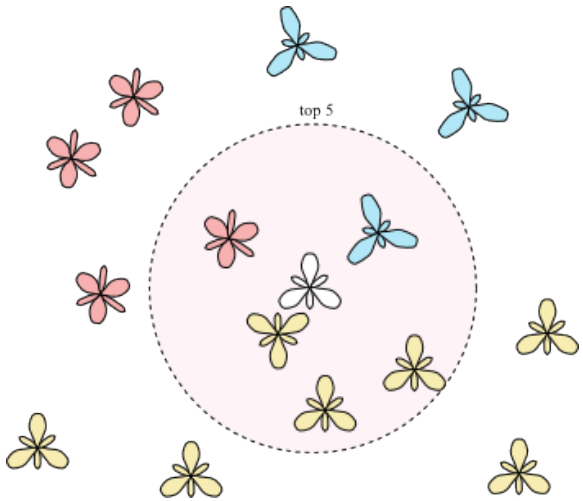
## Commercial

# Let us look at some algorithms

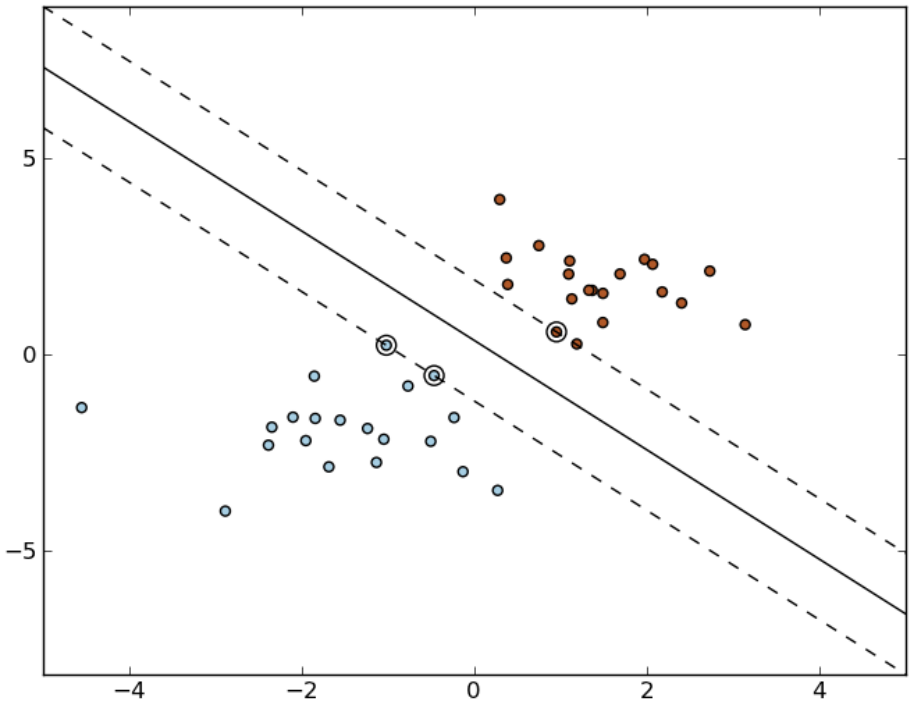
k-Nearest Neighbor

Support vector machines

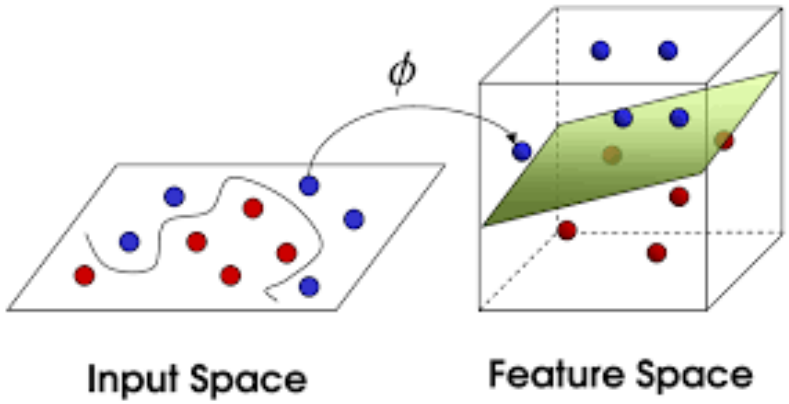
# k-NN



# Support vector machine



# SVM kernel



# Tools

1. weka
2. rapid miner
3. tensor flow
4. scikit learn
5. Many more

# I guess tea and snacks are waiting

We can have a short break