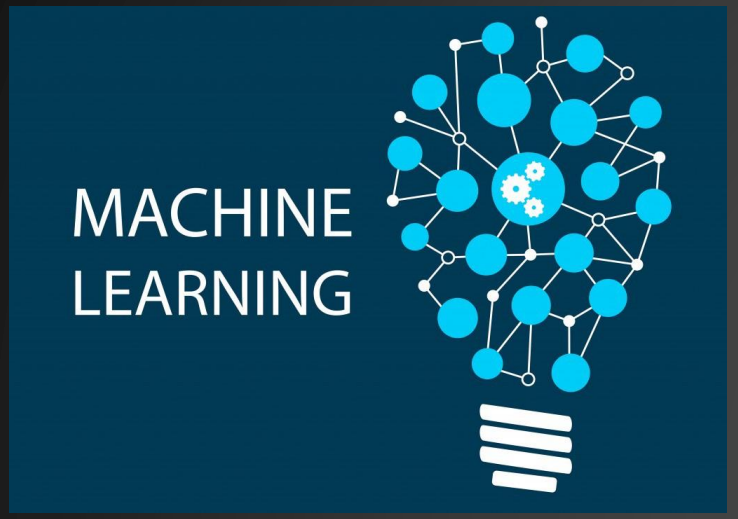
INTRODUCTION TO



FOR

USING

TUE VU

ADVANCED COMPUTING & DATA SCIENCE (ACDS)

CCIT\CITI

OUTLINES

1. Introduction to Machine Learning
2. Why R
3. Types of Machine Learning
4. Caret package
5. Supervised Learning
6. Unsupervised Learning

1. Introduction to Machine Learning

What is Machine Learning?

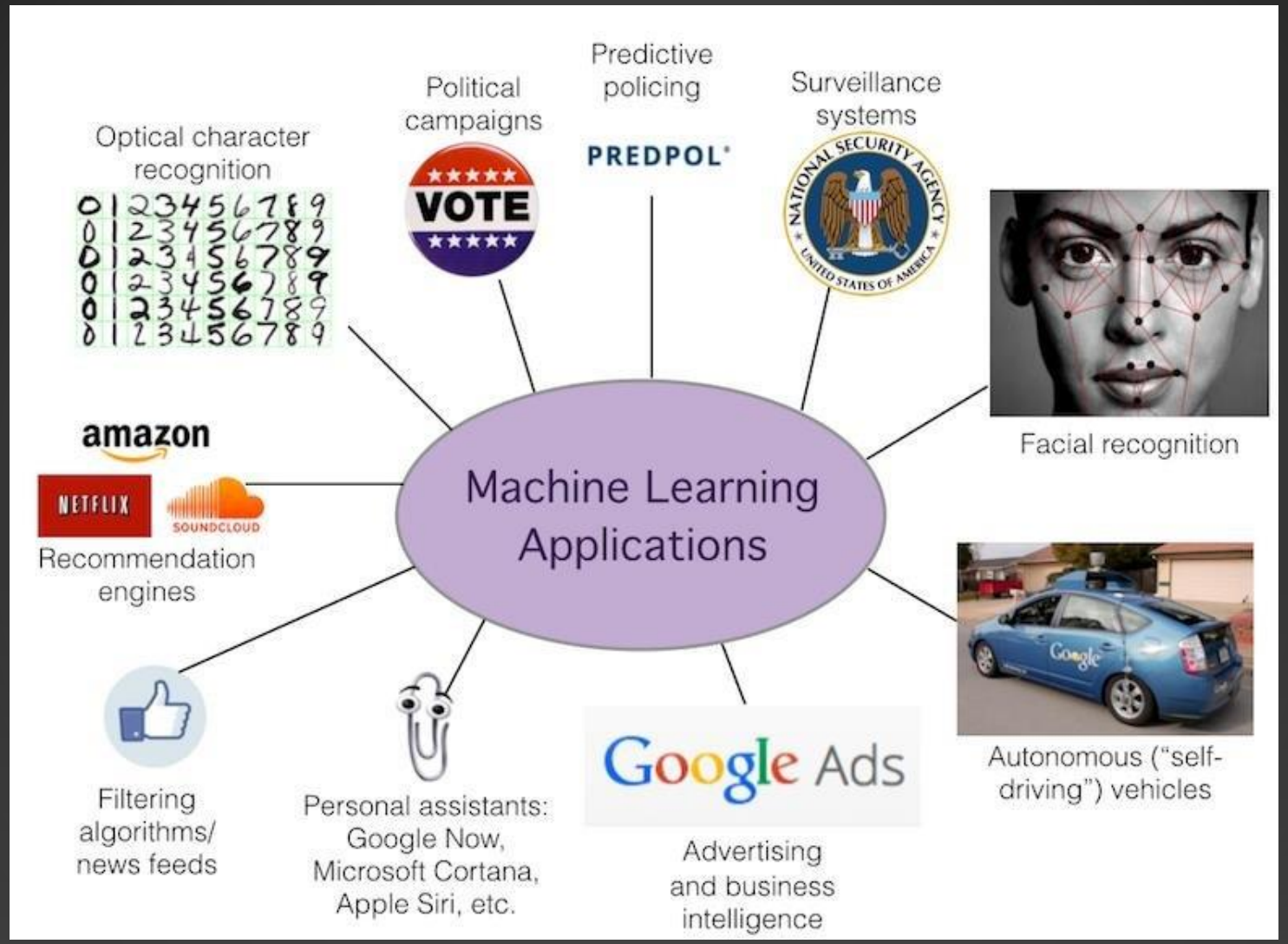
Arthur Samuel: Stanford

Field of study that gives computers the ability to learn without being explicitly programmed

Tom Mitchel: CMU

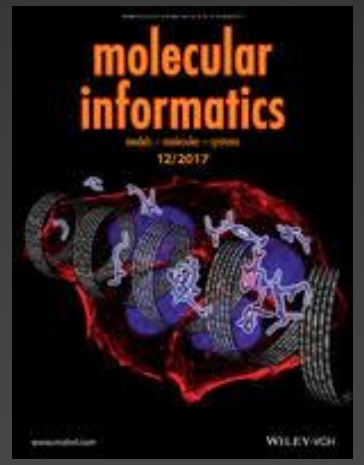
The field of Machine Learning seeks to answer the question:
How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

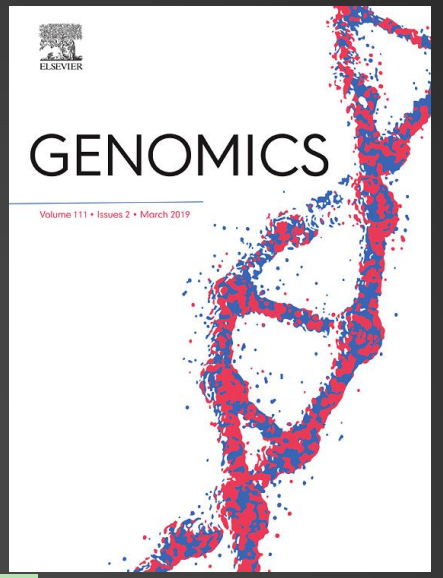
1. Introduction to Machine Learning

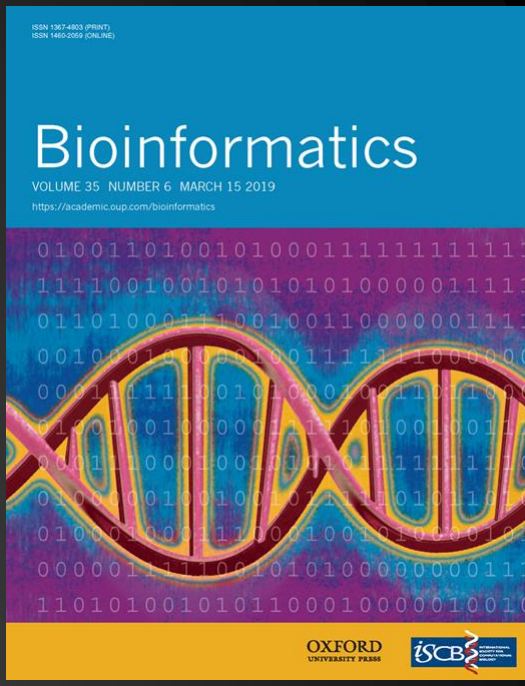


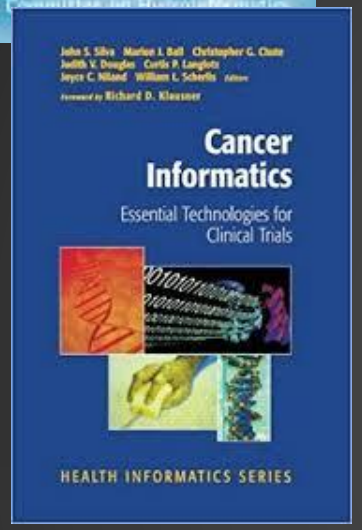
https://www.researchgate.net/publication/323108787_Introduction_to_Machine_Learning

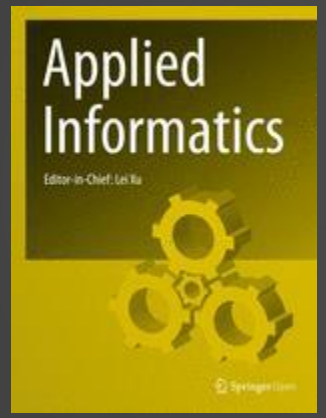
1. Introduction to Machine Learning

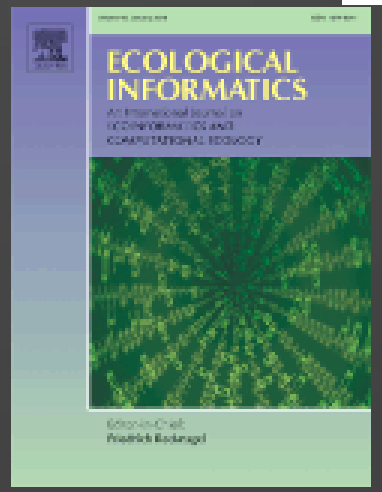


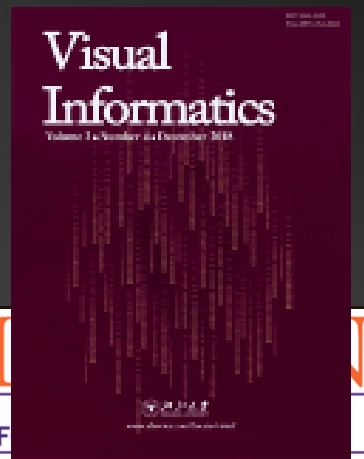


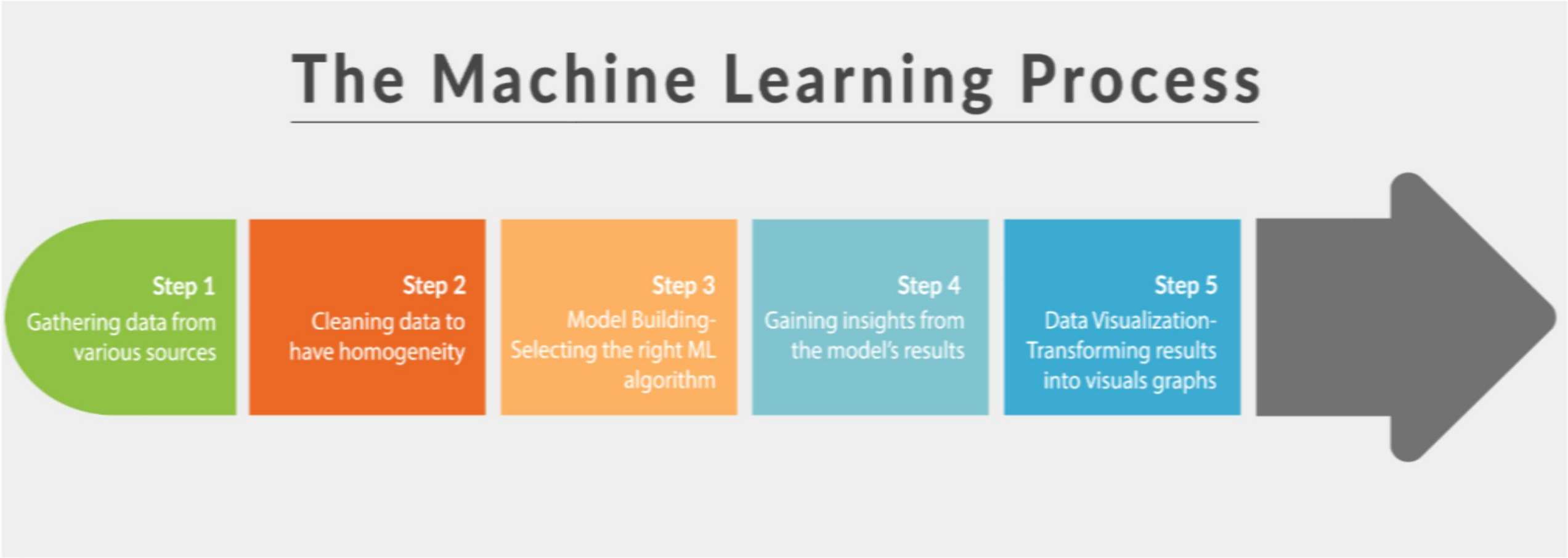




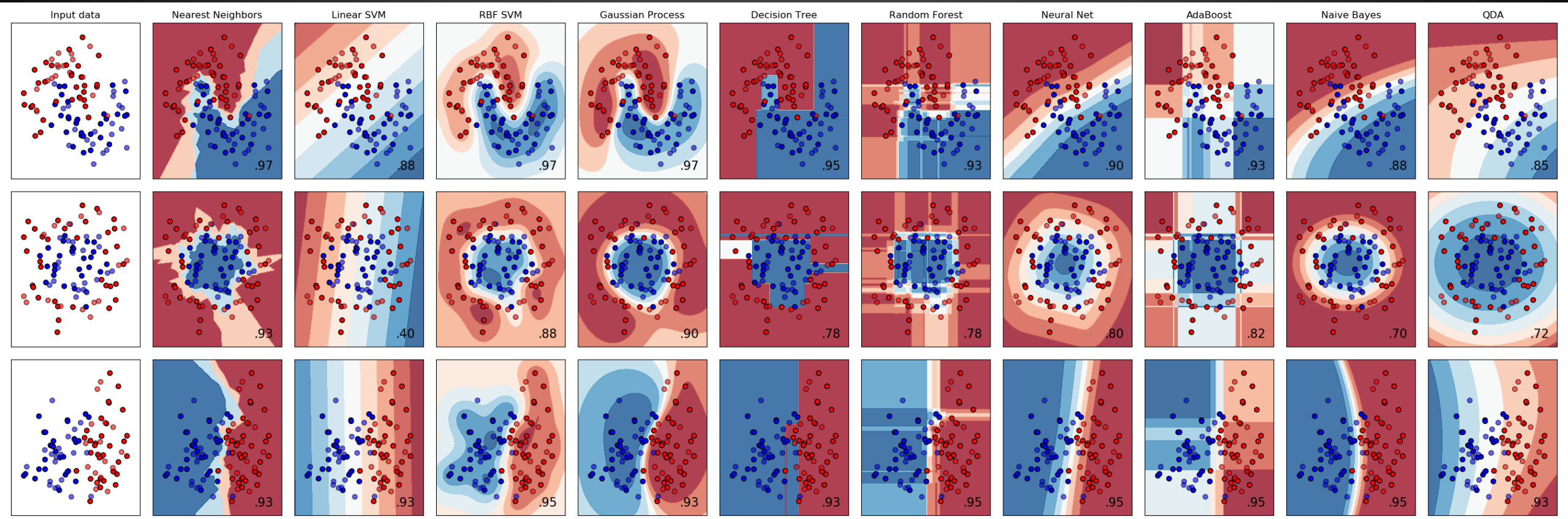






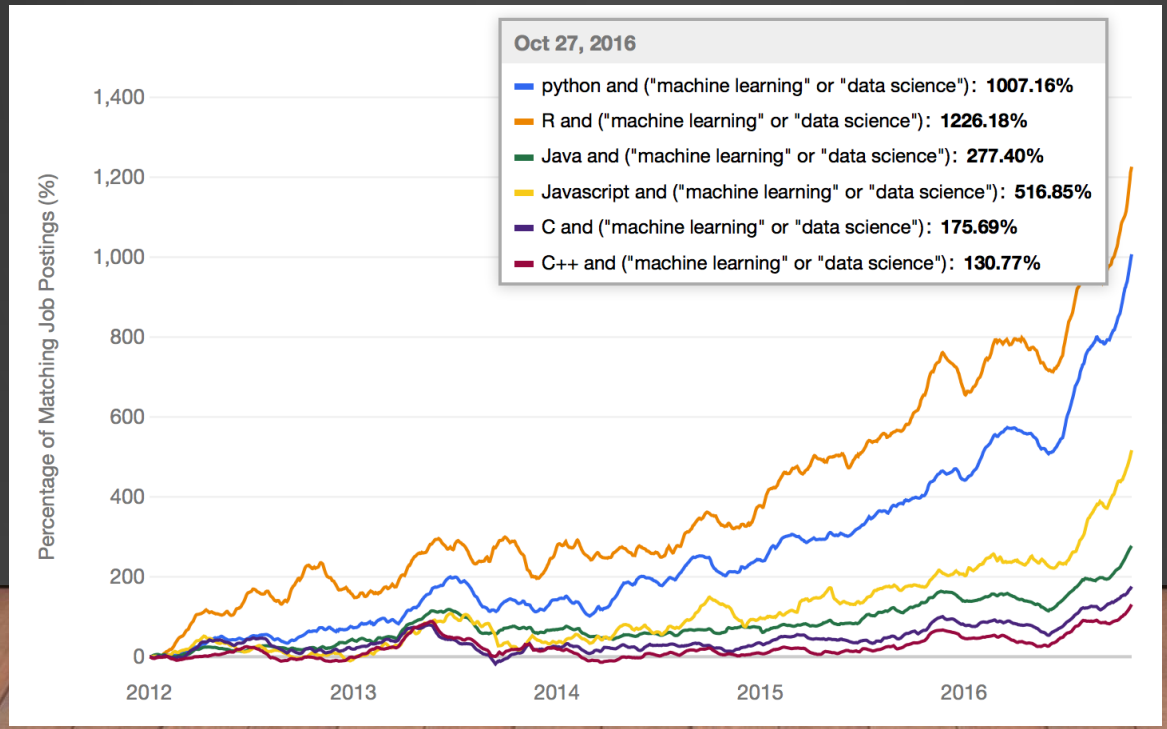

The Machine Learning Process
Step 1
Gathering data from various sources
Step 2
Cleaning data to have homogeneity
Step 3
Model Building- Selecting the right ML algorithm
Step 4
Gaining insights from the model's results
Step 5
Data Visualization- Transforming results into visuals graphs

1. Introduction to Machine Learning



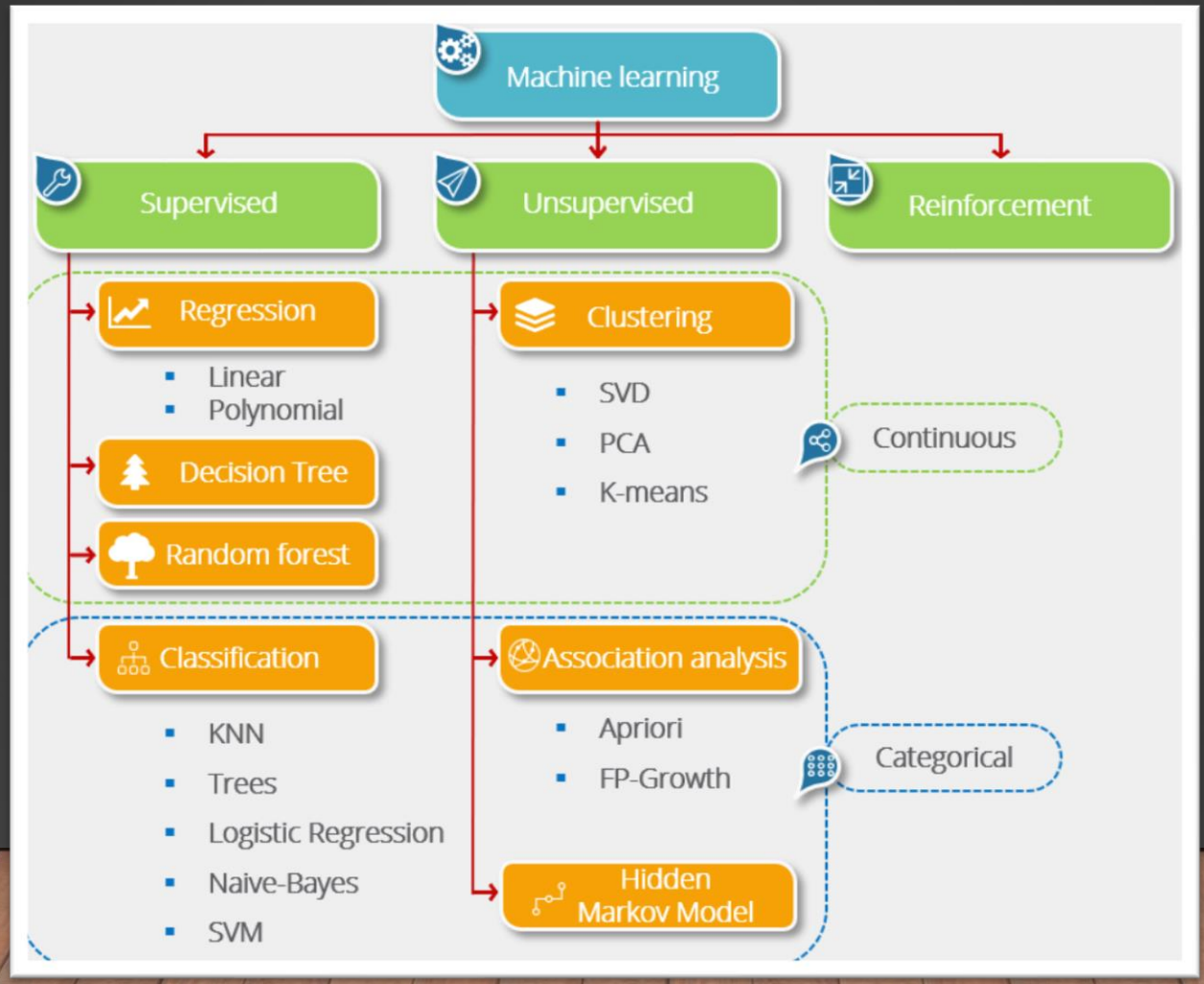
2. Why R?

- **R is used by the best data scientists in the world**. In surveys on Kaggle (the competitive machine learning platform), R is by far the most used machine learning tool.
- **R is powerful because of the breadth of techniques it offers**. The platform has more techniques than any other that you will come across.
- **R is state-of-the-art because it is used by academics**. One of the reasons why R has so many techniques is because academics that develop new algorithms are developing them in R and releasing them as R packages. This means that you can get access to state-of-the-art algorithms in R before other platforms.
- **R is free because it is open source software**. You can download it right now for free and it runs on any workstation platform you are likely to use.
- **R is a great tool for researcher.** PhD students and researchers need lots of statistics for their studies and publications



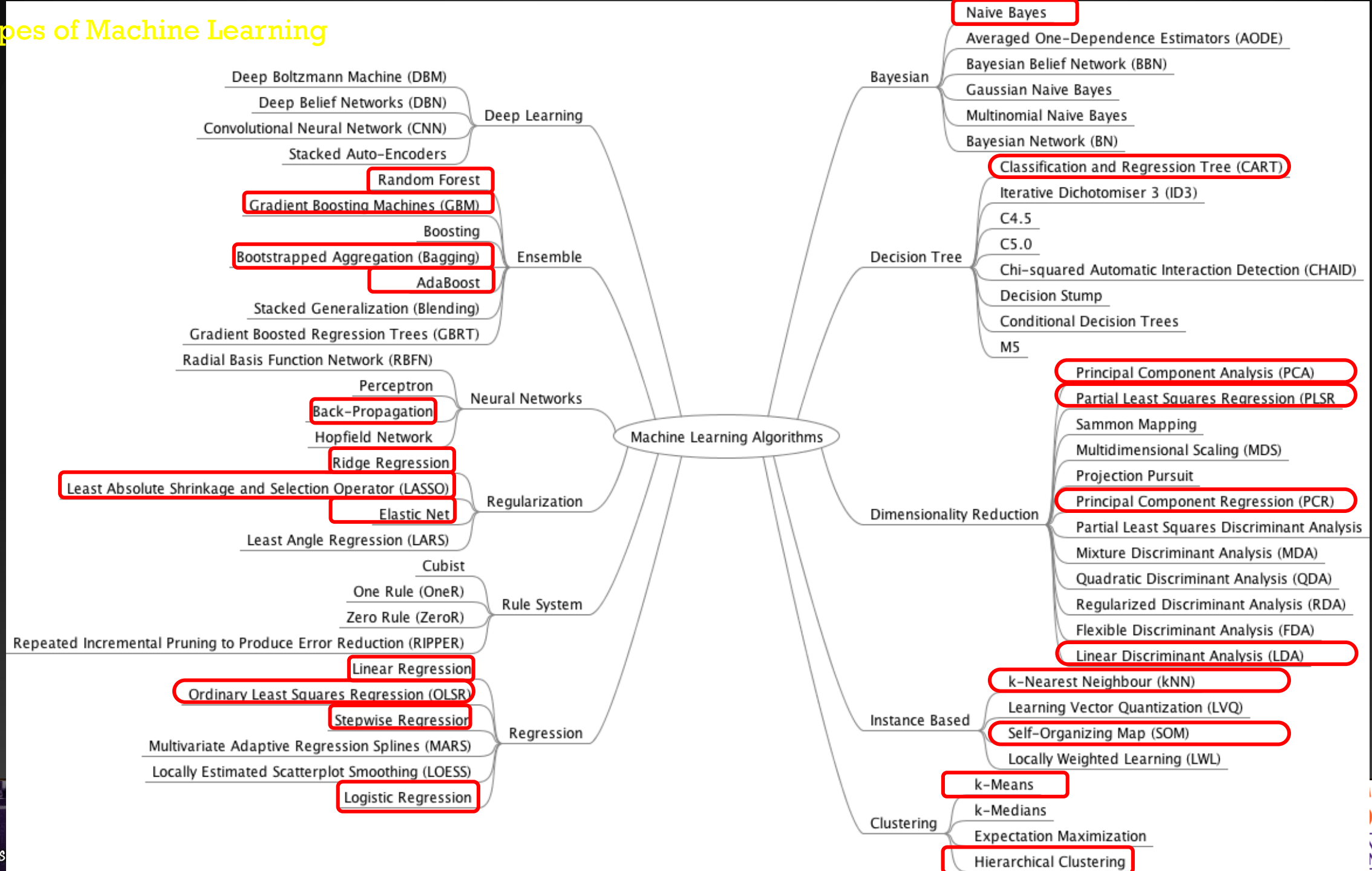
3. Types of Machine Learning

- Supervised Learning – Train Me! (target are dependent variables)
- Unsupervised Learning – I am self sufficient in learning
- Semi-supervised Learning: combination of both methods, when cost to label are high
- Reinforcement Learning – My life My rules! (Hit & Trial)

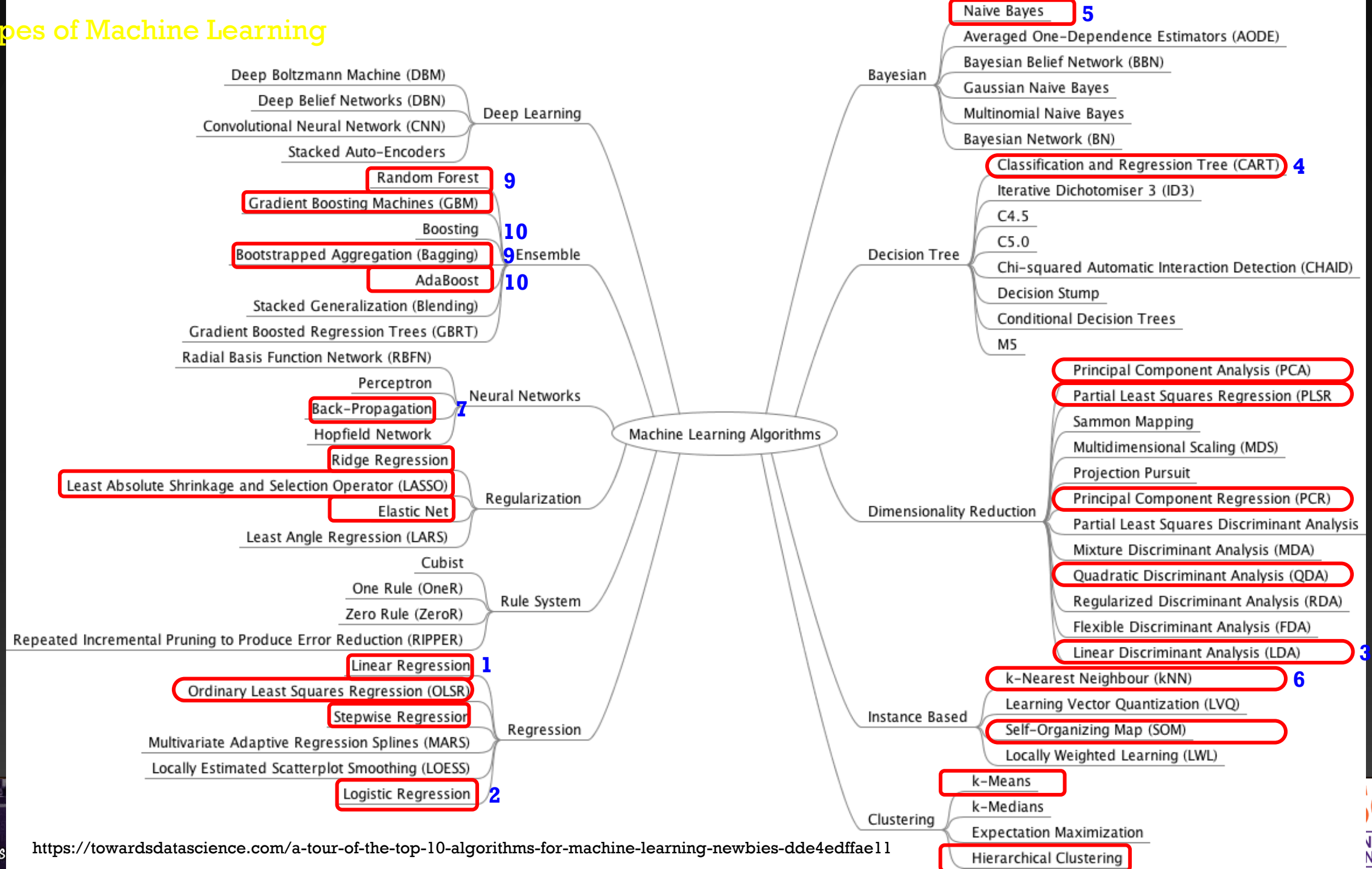


3. Types of Machine Learning

3. Types of Machine Learning



3. Types of Machine Learning

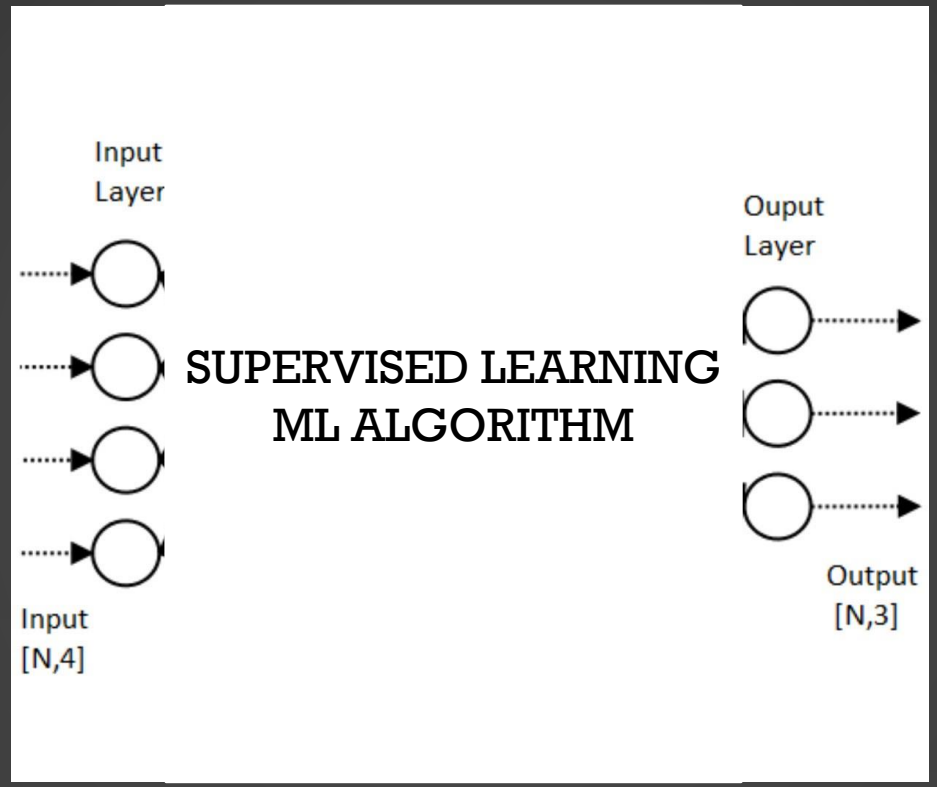


https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11

3. Types of Machine Learning

Terminology

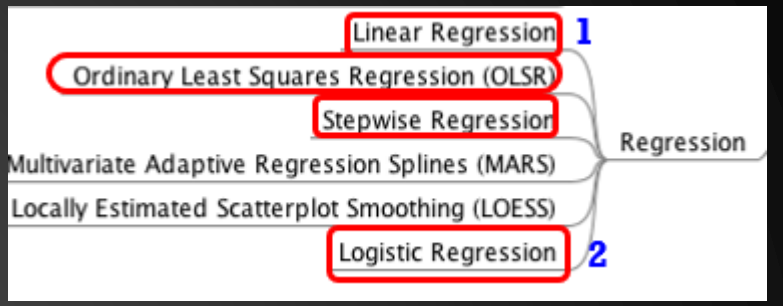
- Input variables
- Independent variables
- Predictors
- Features
- Input Field



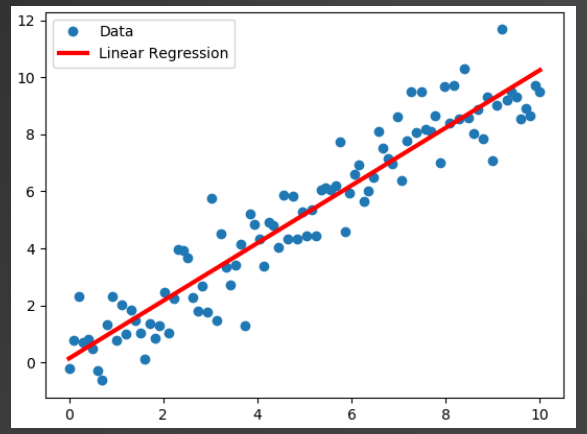
- Output variables
- Dependent variables
- Predictand
- Target variables
- Outcome Field

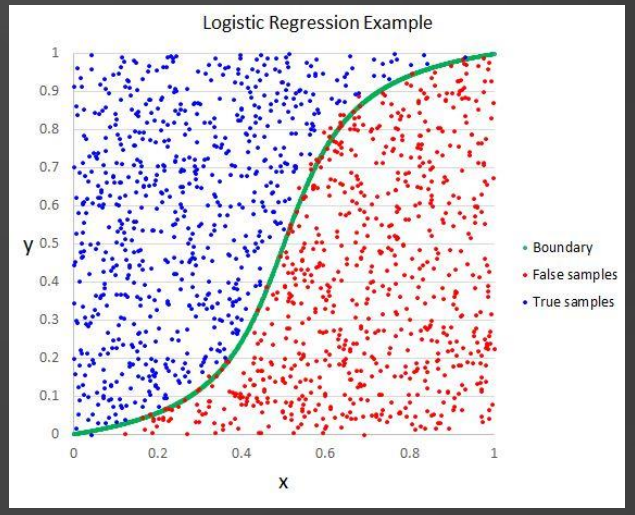
3. Types of Machine Learning

Regression based methods

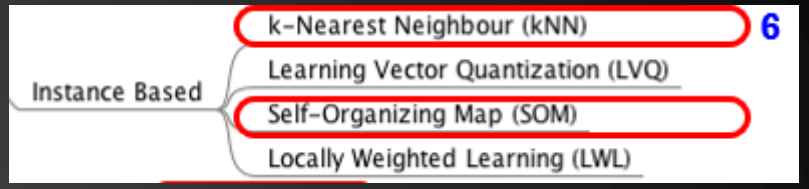


- Most popular & widely used in research for engineer
- Easy to explain and apply
- The relationship between dependent variable and set of independent variables is estimated by probabilistic method/error function minimization



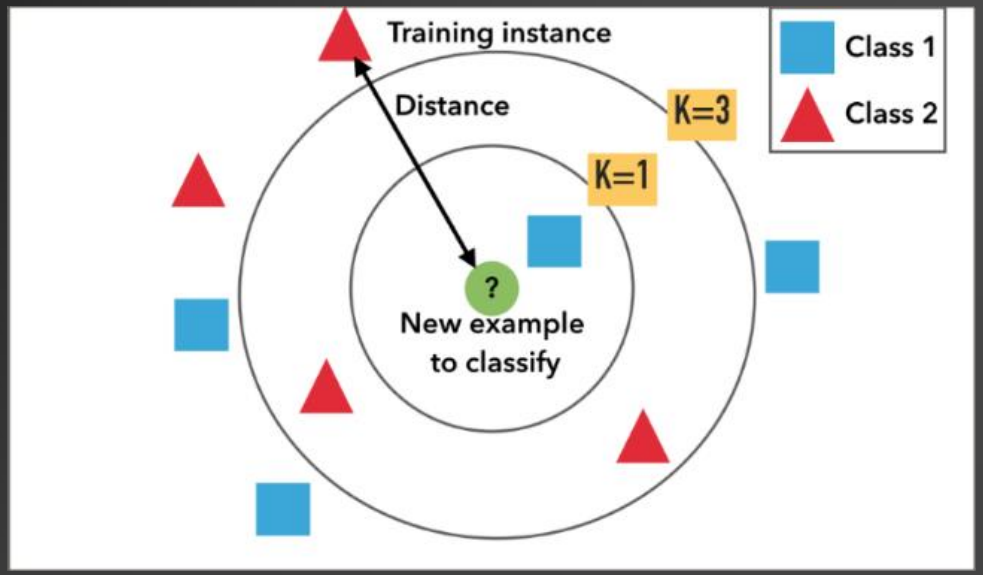


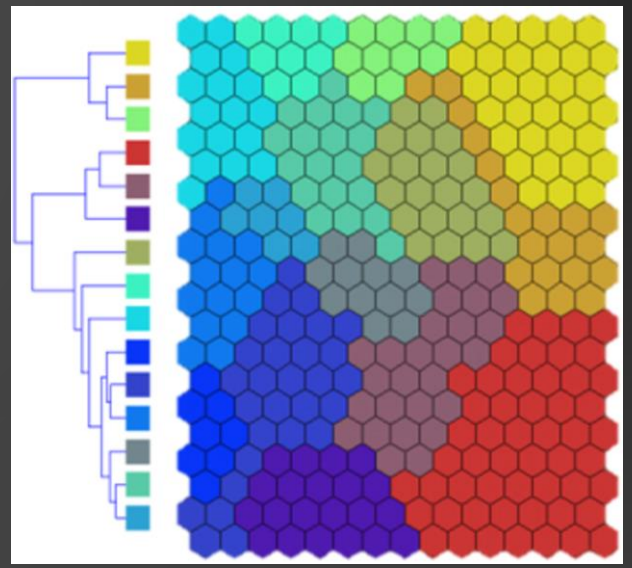
3. Types of Machine Learning



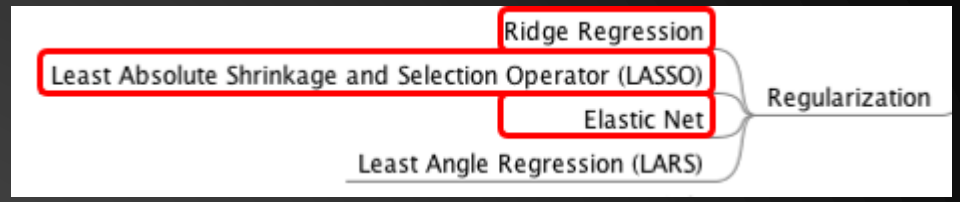
Instance based methods

- So called Distance-based, event-based or memory-based learning
- Self-learning and create a metric to identify whether an object belongs to the class of interest or not
- Learn from sets of events/instances captured in the data



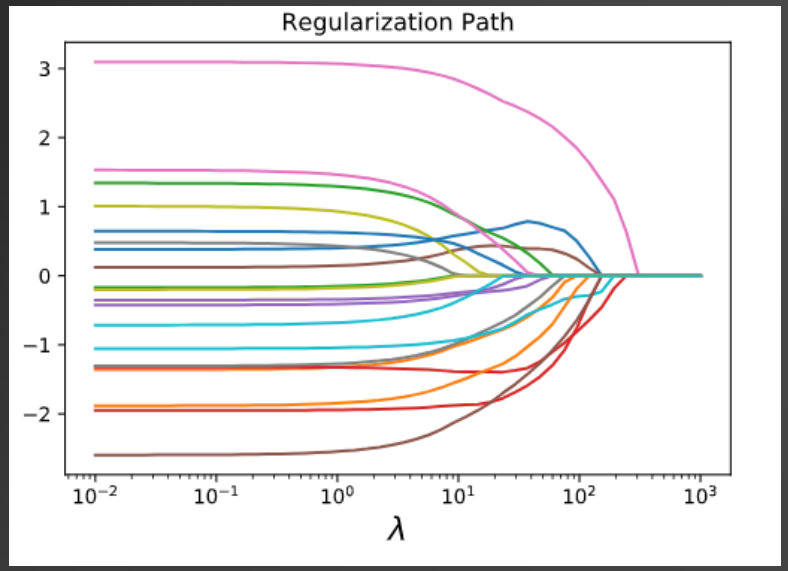


3. Types of Machine Learning



Regularization methods

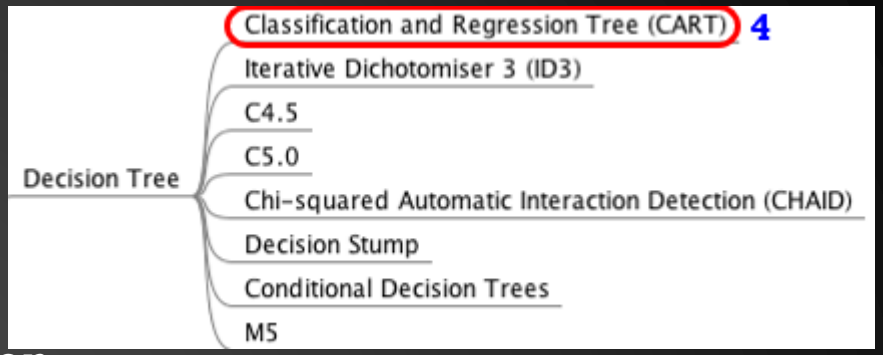
- An extension of regression methods
- Introduce a penalization term to the loss function for balancing between complexity of model and improvement in results
- Powerful dealing with large number of input dataset

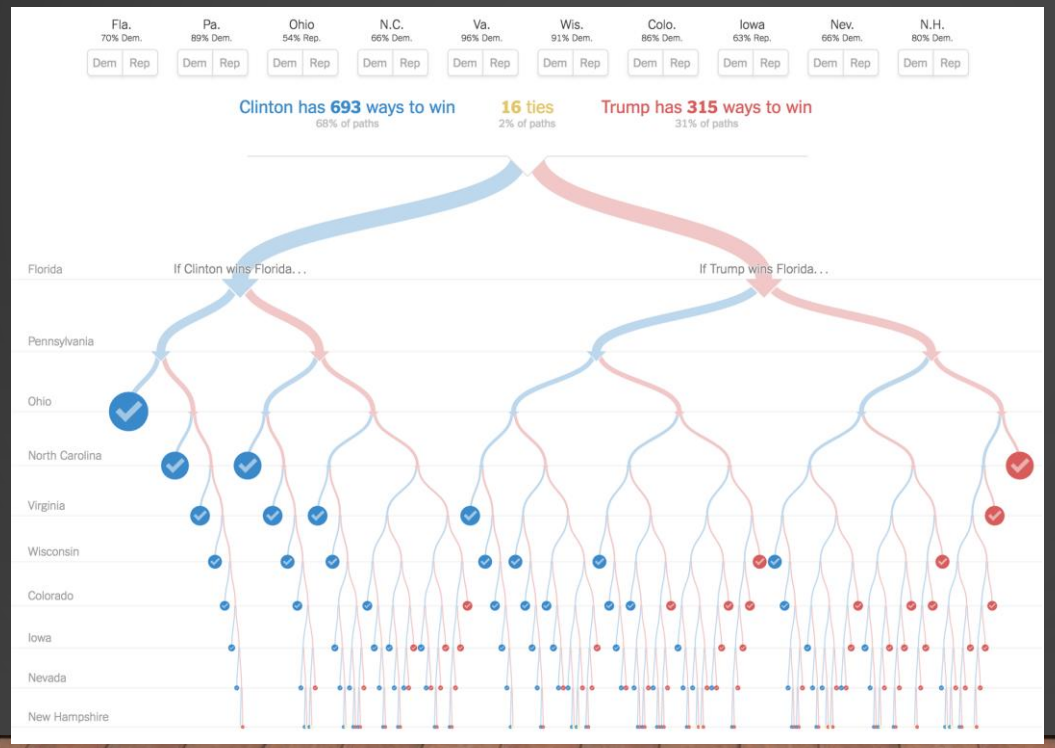


3. Types of Machine Learning

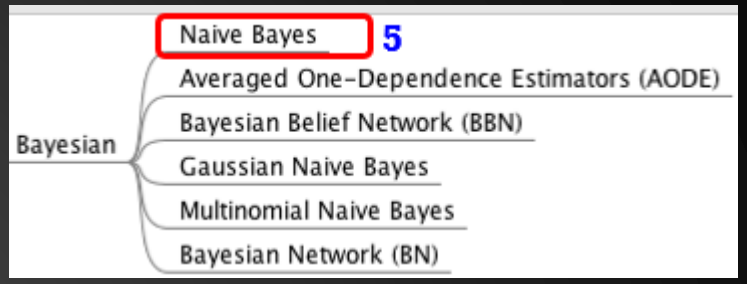
Tree-based algorithms

- Sequential conditional rules applied on the actual data
- Rules are applied serially and a classification decision is made when all conditions are met
- Fast and distributed algorithm



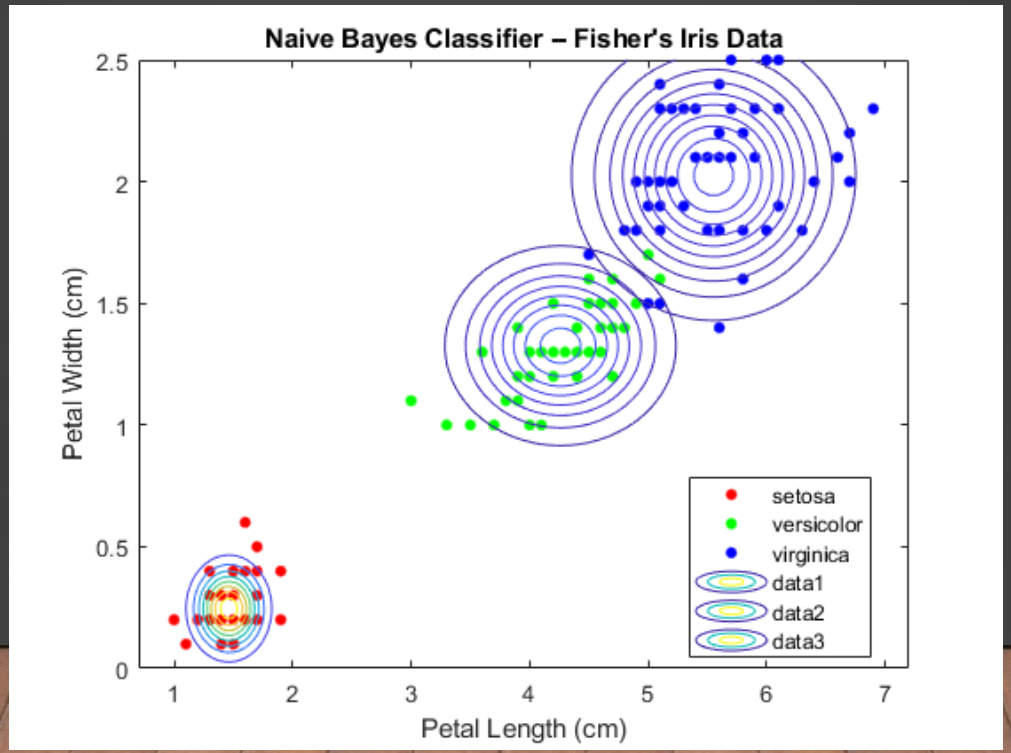


3. Types of Machine Learning

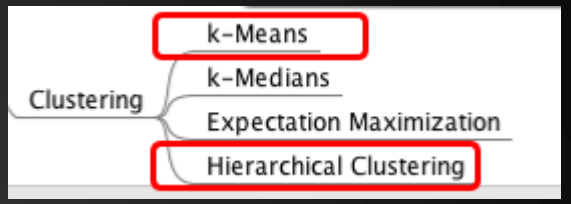


Bayesian Algorithms

- Work based on Bayes Theorem using prior and post distribution
- The machine does not learn from iterative process but using inference from distribution of variable
- Used in most classification and inference testing

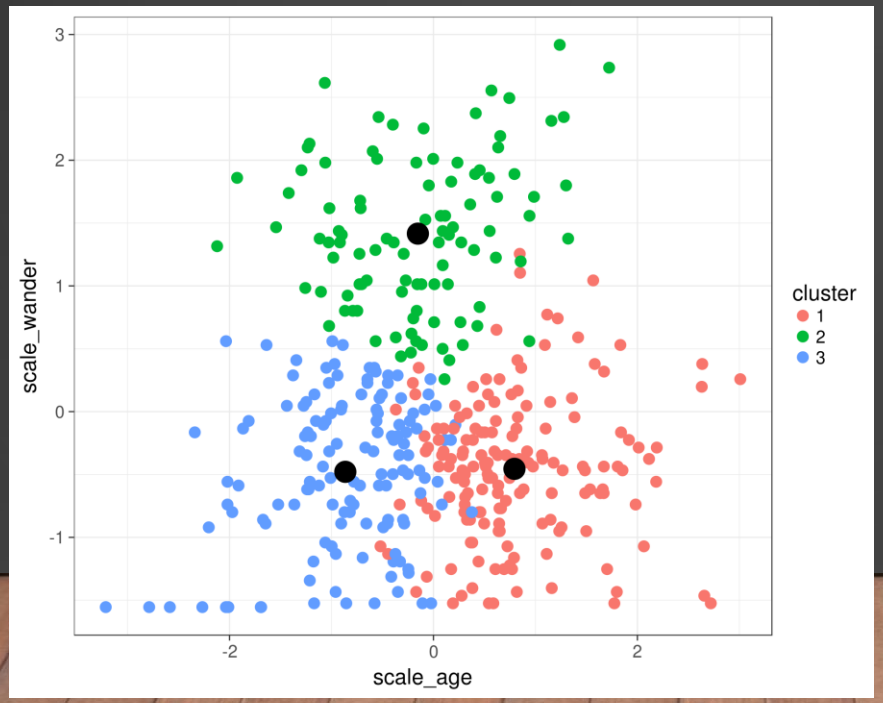


3. Types of Machine Learning

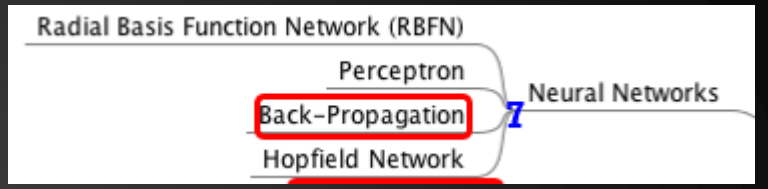


Clustering Algorithms

- Principle of maximization of intra-cluster similarities and minimization inter-cluster similarities
- The measure of similarities determines how the clusters need to be formed
- Unsupervised algorithm: group the data for maximum commonality

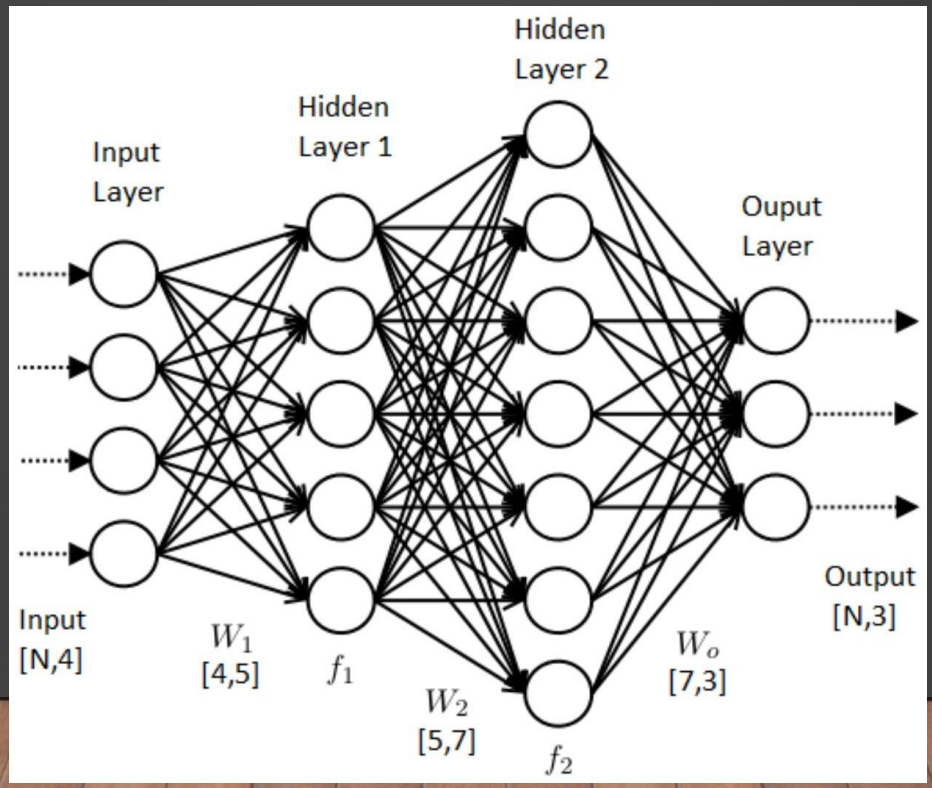


3. Types of Machine Learning



Artificial Neural Networks (ANN)

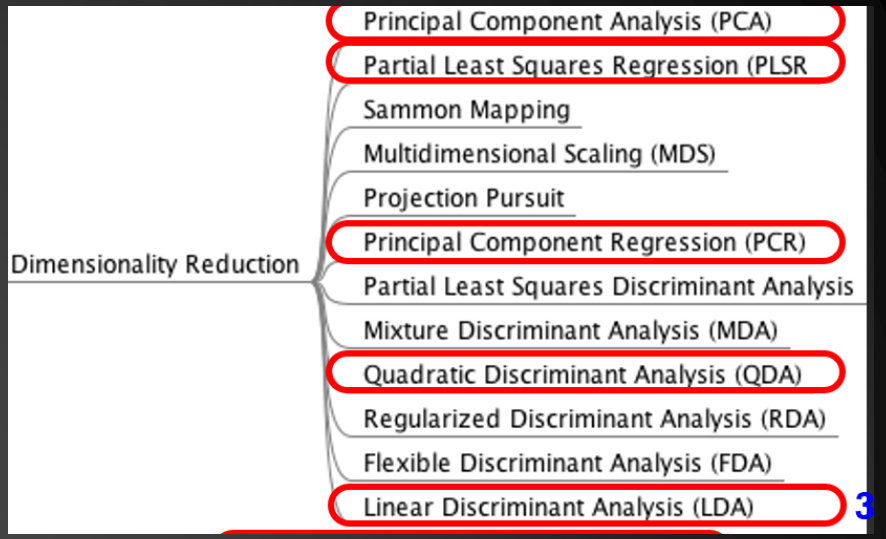
- Inspired by the biological neural networks
- Powerful to learn non-linear relationships
- Recognize higher order relationships among variables
- Used in both supervised/unsupervised learning

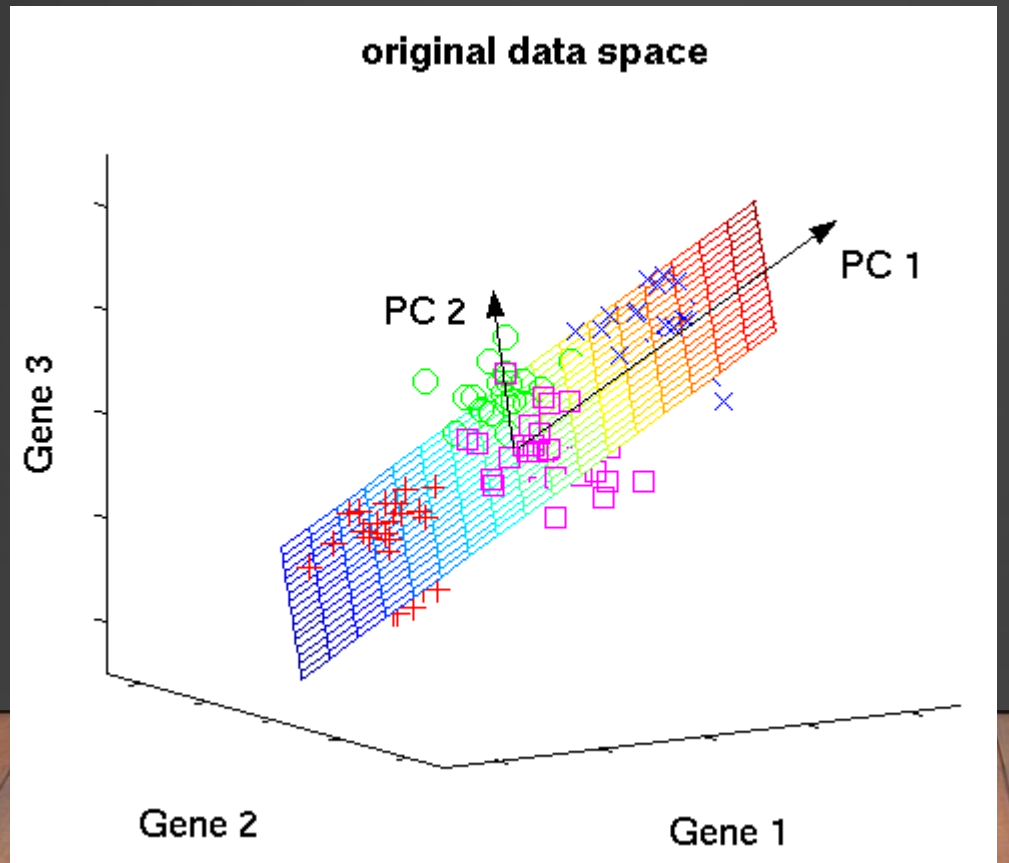


3. Types of Machine Learning

Dimensionality Reduction Algorithm

- Essential method to amplify the signal in data by various transformation
- Reduce number of independent variables (inputs)
- To be applied before modeling

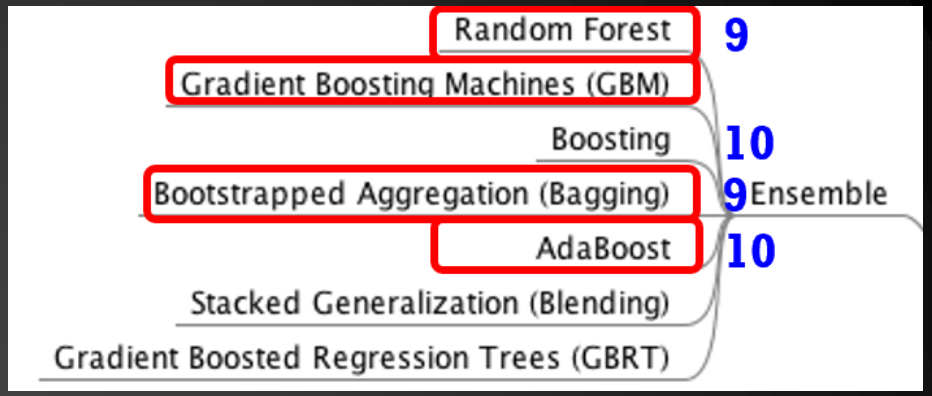




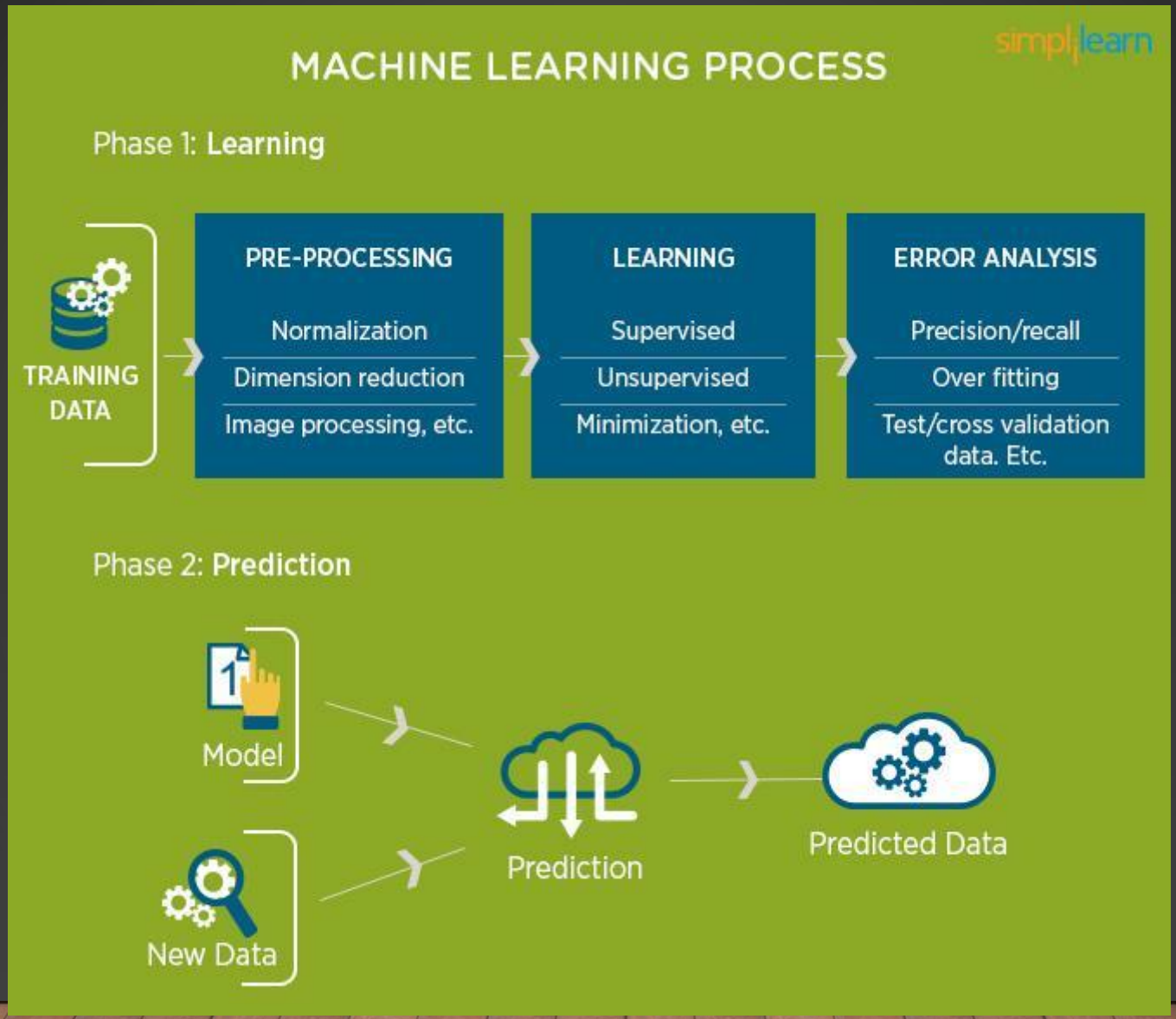
3. Types of Machine Learning

Ensemble Learning

- Combination of results from different ML approach
- Very popular as they have ability to provide superior results
- Possibility to break into independent model to train a distributed network



3. Types of Machine Learning



Caret: Classification And REgression Training

- R has so many ML algorithms, challenge to keep track, different syntax for different packages
- Possibly the biggest project in R
- All in one supervised learning problem
- Uniform interface
- Standard pre & post processing

Install.packages("caret")

4. Caret package

4.1 Data partition: training and testing

4.2 Descriptive statistics

4.3 Preprocessing with missing value

4.4 Preprocessing: transform data

4.5 Visualize important variables

4.6 Train and predict the model

4.7 Preprocessing argument

4.8 Evaluate test result

4. Caret package

1. Data partition: training and testing

- Splitting based on the outcome
- Splitting based on the Predictors using maximum Dissimilarity
- Splitting based on timeseries
- Splitting based on resampling set to different group: K-fold

4. Caret package

2. Descriptive statistics

3. Preprocessing function: *preProcess*

- Use in many operation on predictors
- Estimate the required parameters for each operation without recomputing the values
- "predict" function in preProcess used to apply to specific data set (testing)
- Can be interface by using train function

3. Preprocessing: missing value

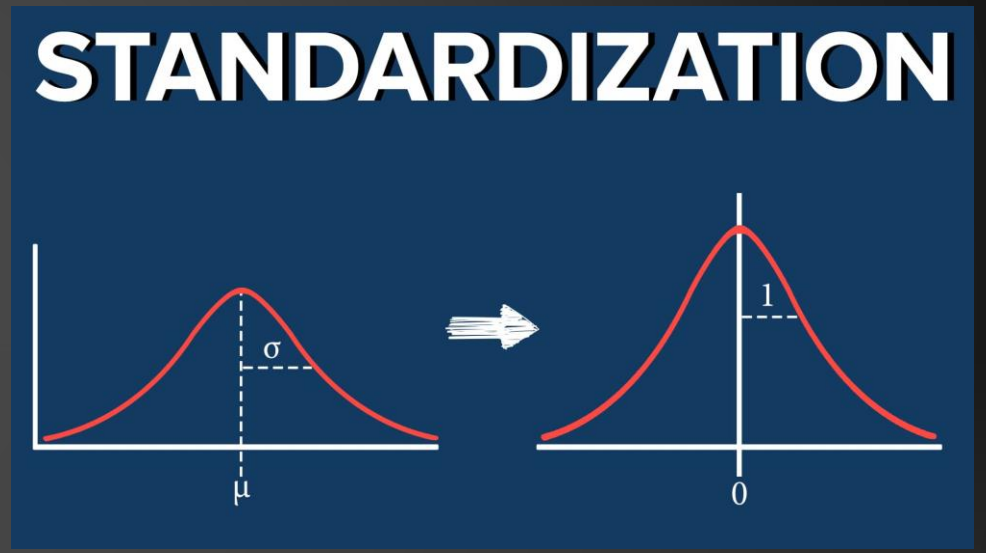
- Omit NA
- Set NA to some constant value (for example mean or 0)
- Impute using KNN and Bagging
- knnImpute: fill missing value with K-nearest neighbor technique and standardize
- bagImpute: fill missing value with Bagging technique: no standardize
- bagImpute is more powerful and computational cost than knnImpute

4. Caret package

4. Preprocessing: transform data

Standardization: **Centering and Scaling**

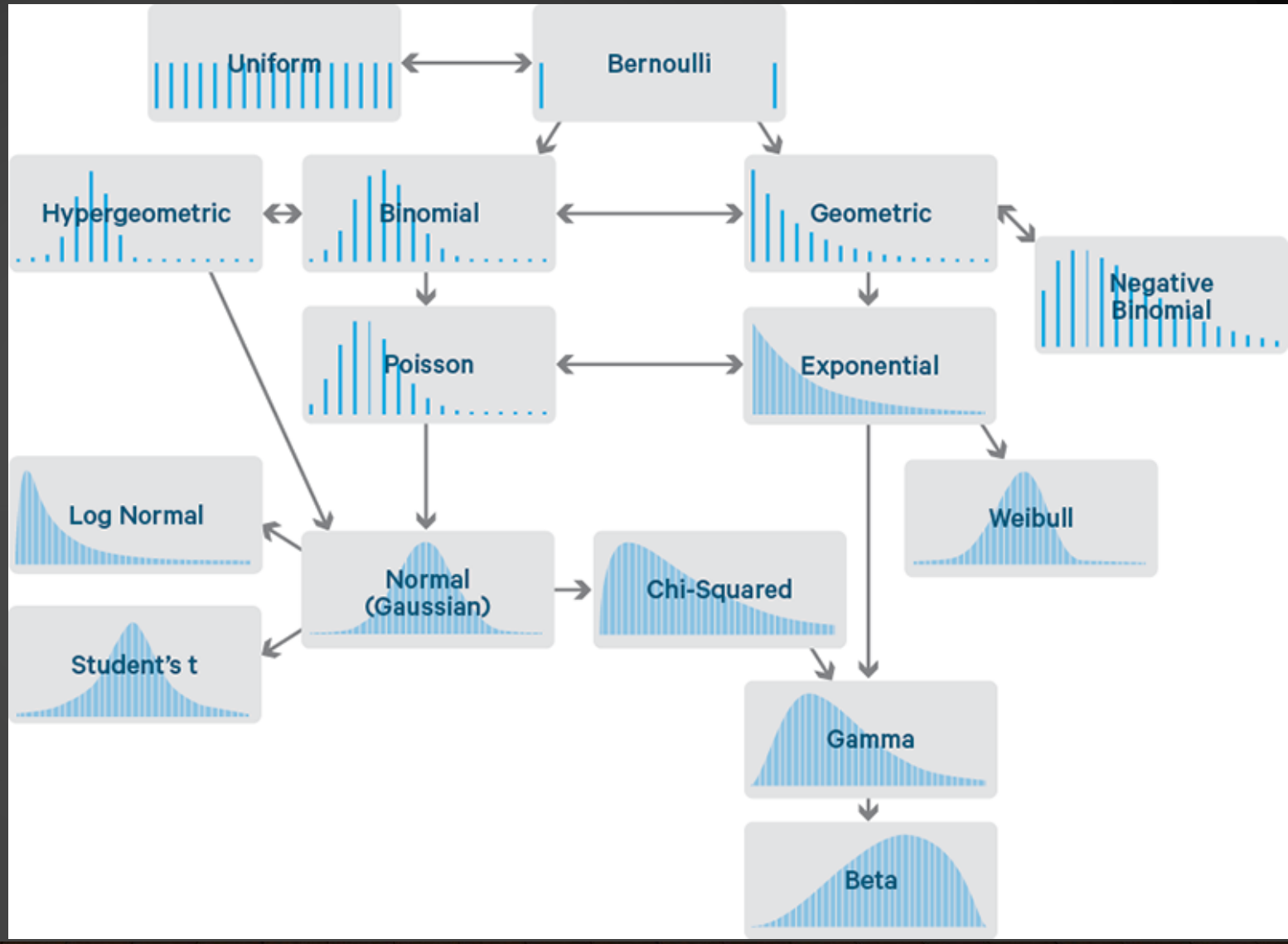
- Convert all independent variables into the same scale (mean=0, std=1)



4. Preprocessing: transform data

Linear Transformation

- Perform math operation on each piece of original data
- Most popular are linear transformation (currency exchange, etc.)
- Linear transformation do not change shape of distribution, especially to normal shape

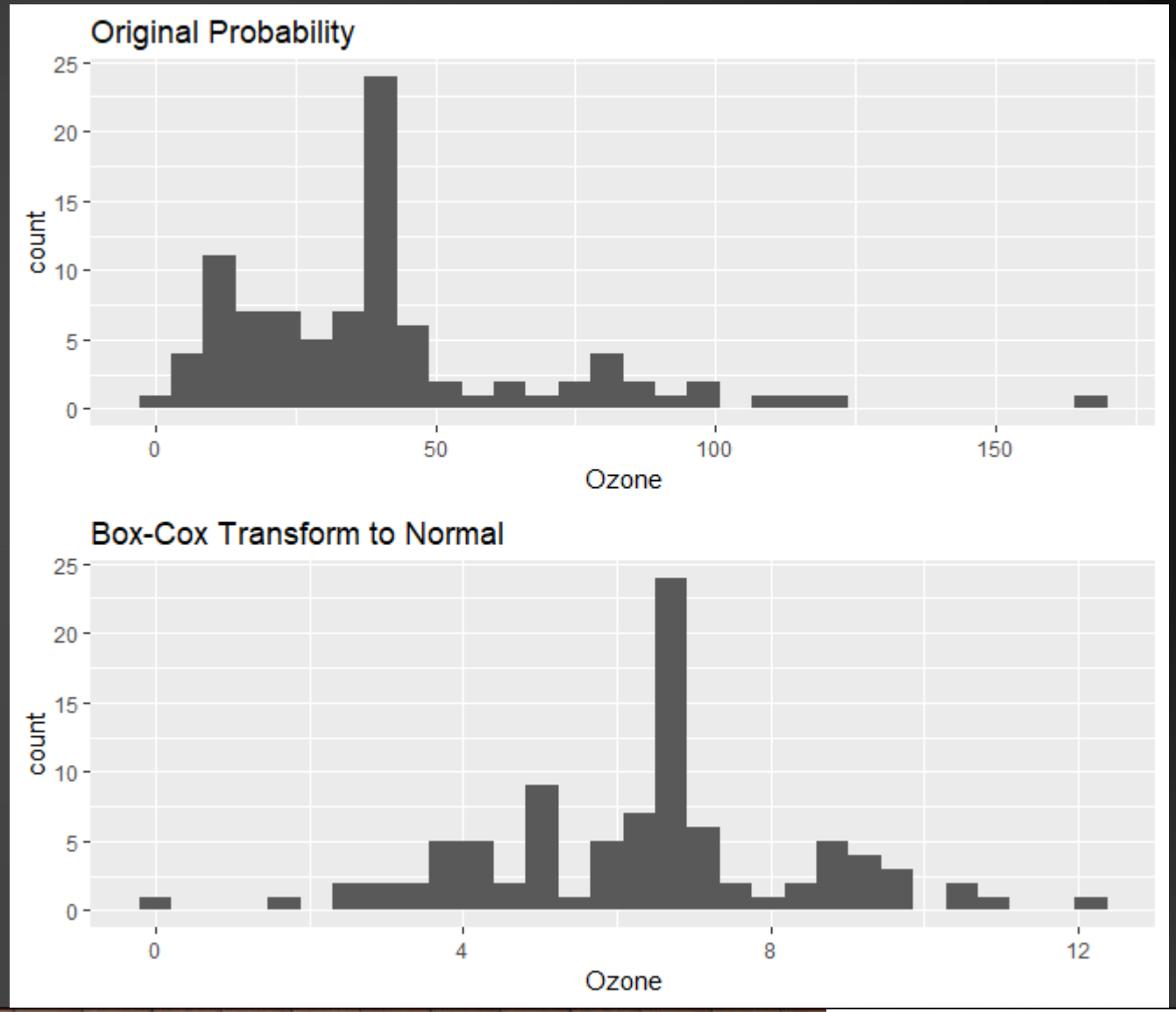


4. Preprocessing: transform data

Box-Cox Transformation

- Transform data to normal shape (data > 0)
- Parameter to be changed is λ
- $-5 < \lambda < 5$
- Optimization to find λ.
- So called power transform

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda}$$



4. Caret package

5. Preprocessing: transform data

Other Transformation

- BoxCox; YeoJohnson, expoTrans, center, scale, range, knnImpute, bagImpute, etc.

4. Caret package

6. Visualize important variables

6. Train and predict model

```
ModelFit <- train(type,data=training, method="ML model")
Prediction<- predict(ModelFit,testing)
```

```
ModelFit <- train(type,data=training,
                  preprocess=c("center","scale"), method="ML model")
Prediction<- predict(ModelFit,testing)
```

7. PreProcess Argument

8. Evaluate test results

For regression/continuous results:
Cor()
Cor.test()
postResample()

For discreet/classification results
confusionMatrix()

For probability results
twoClassSummary()