# Introduction to Machine Learning


AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF THE UGC ACT, 1956

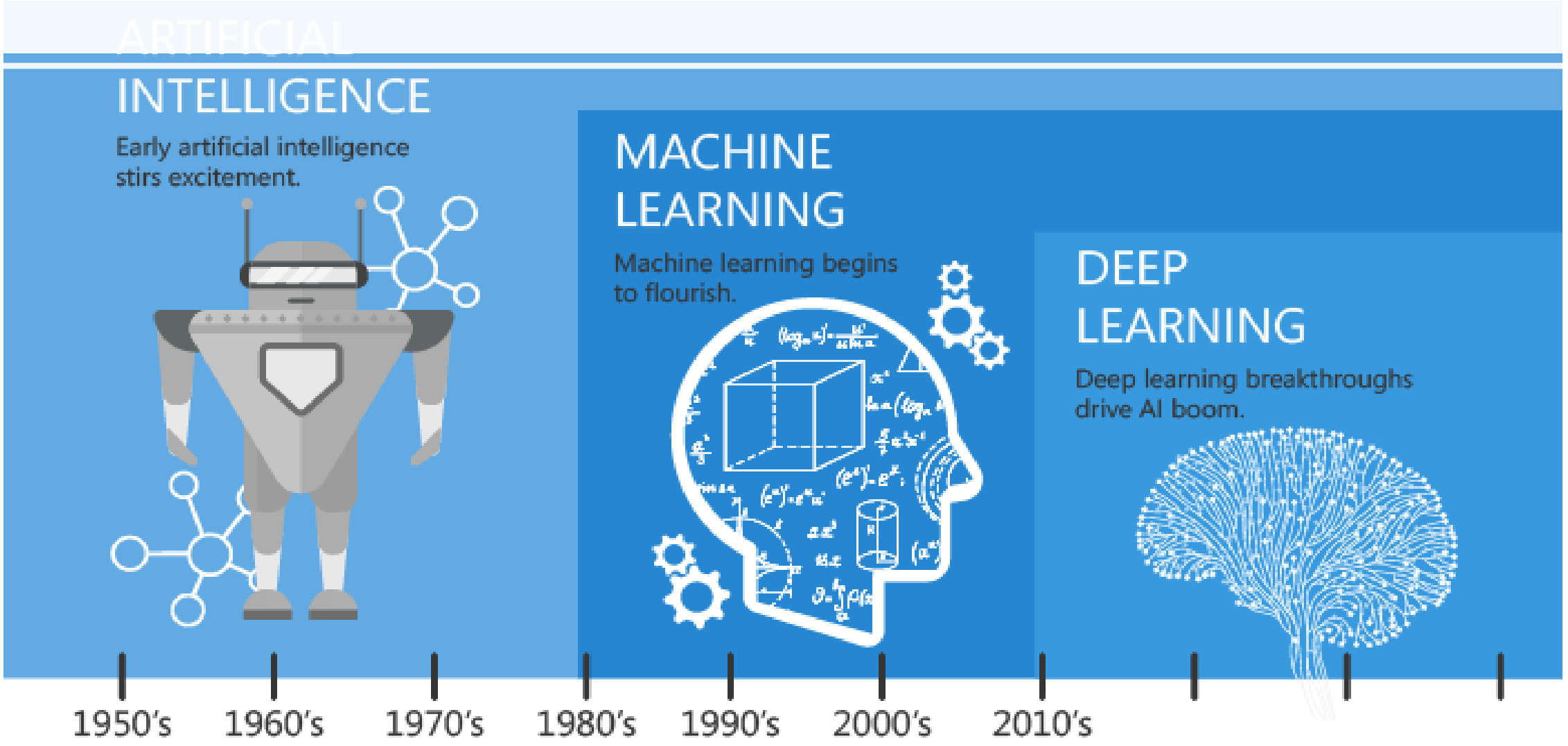
# Outline

- What is Machine Learning
- Why Machine Learning
- Formal Definition : T,E,P for Machine Learning
- Desired Characteristics of ML Algorithm
- Applications of Machine Learning
- Types of Machine Learning

# Machine Learning

- Subset of AI
  - AI is the simulation of human intelligence in machines
- Primary Goal
  - Training a machine to learn from its past experience
- Outcome
  - ML creates models that can improve its performance with minimal human support

# Machine Learning: A subset of Artificial Intelligence




Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

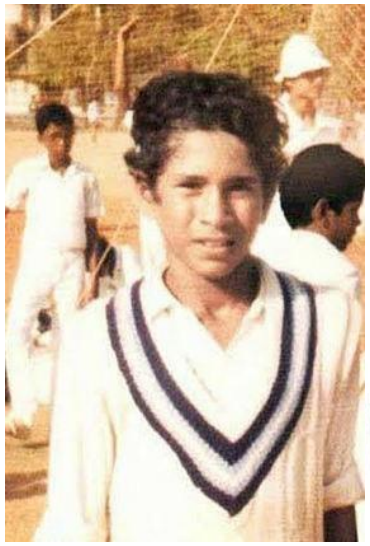
Source : https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08

# Conventional Algorithms / Programs

- Set of hard coded rules to solve a specific task.
- Works on programmer's logic and is task specific.
- Designing rules requires a deep understanding of the domain and is done by a human expert

# Why Machine Learning : A motivational example

Look at the images and understand/remember the patterns.

Model and generalization

- Machine learning algorithms generate model.
- Model is a mathematical structure that learns and generalize the data
- Not coded...(If curly hair then sachin)

# Why generate model ?

Guess : Is this Sachin ??

85% 5% 90% 95% 45%

From binary world to Probability world

Likelihood of an image to be of Sachin.

REDUCE GENERALIZATION ERROR !!!

# Definition

Definition 1 : Arthur Samuel(1959)

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

Definition 2 : Tom Mitchell (1998)

***A computer program is said to learn from Experience E with respect to some Task T and some Performance measure P if performance on T, as measured by P, improves with experience E.***

# Task, Experience, Performance

Ill posed problem :
Recognize Sachin Tendulkar in the given Photographs

Well posed Problem:

Task : Check if the face in an image is of Sachin or not
Experience : Various Images of Sachin
Performance : Out of 100 photographs, how many times was Sachin
identified with more than 90% confidence

# Ill posed versus Well posed Learning problems

- Ill posed problems :
  - No clear goal/Has simultaneous goals
  - No clear solution paths No unique solution

"Today is Sunday. Shreyas would like to watch a movie".

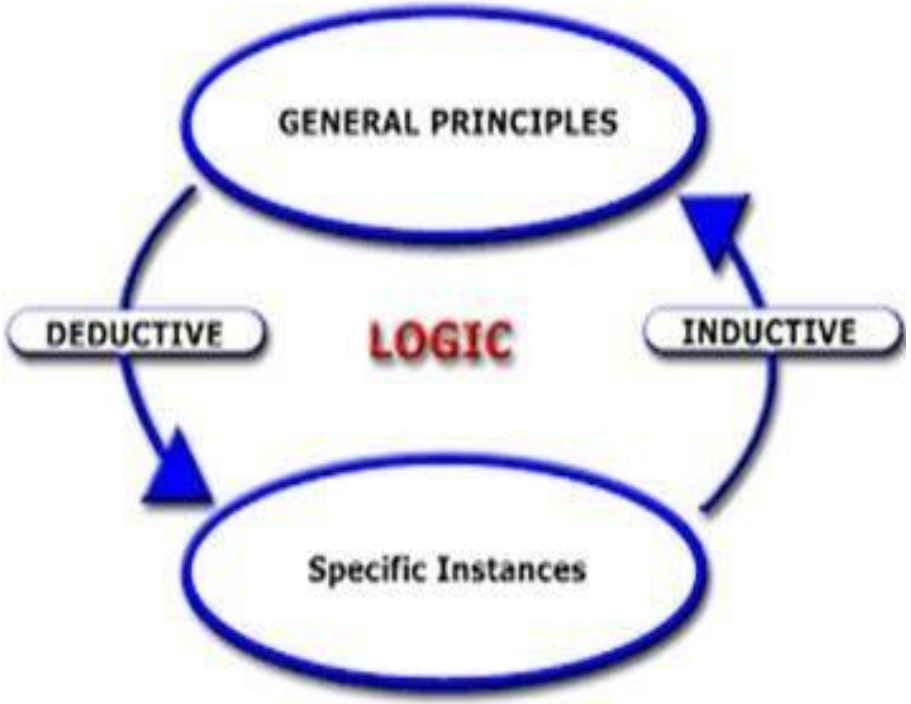
- Ill posed problems should be converted into a Well Posed problem in terms of Task, Experience, and Performance

- Well posed problem
  - Solution exists
  - The solution is unique

  The solution depends on the data/experience

"Task : Classify whether Krish movie will be liked by Shreyas"

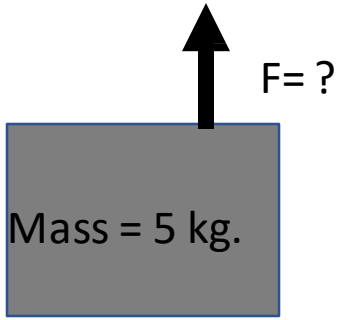
# Inductive versus Deductive Learning




t = ?

t = a1 * a2

Force = Mass * Acceleration

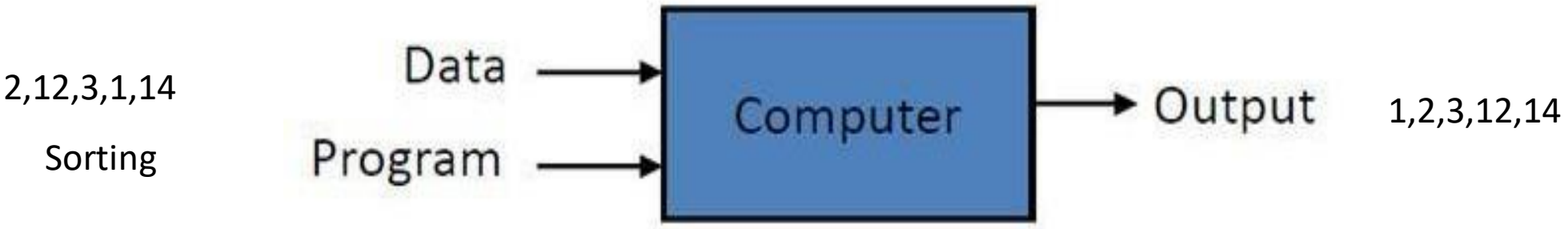



Acceleration upwards= 3.5 m/s^2

| a1 | a2 | t |
|---|---|---|
| 5 | 2 | 10 |
| 3 | 2 | 5.9 |
| 10 | 5 | 47 |
| 14 | 2 | 28 |

# Conventional versus ML algorithms

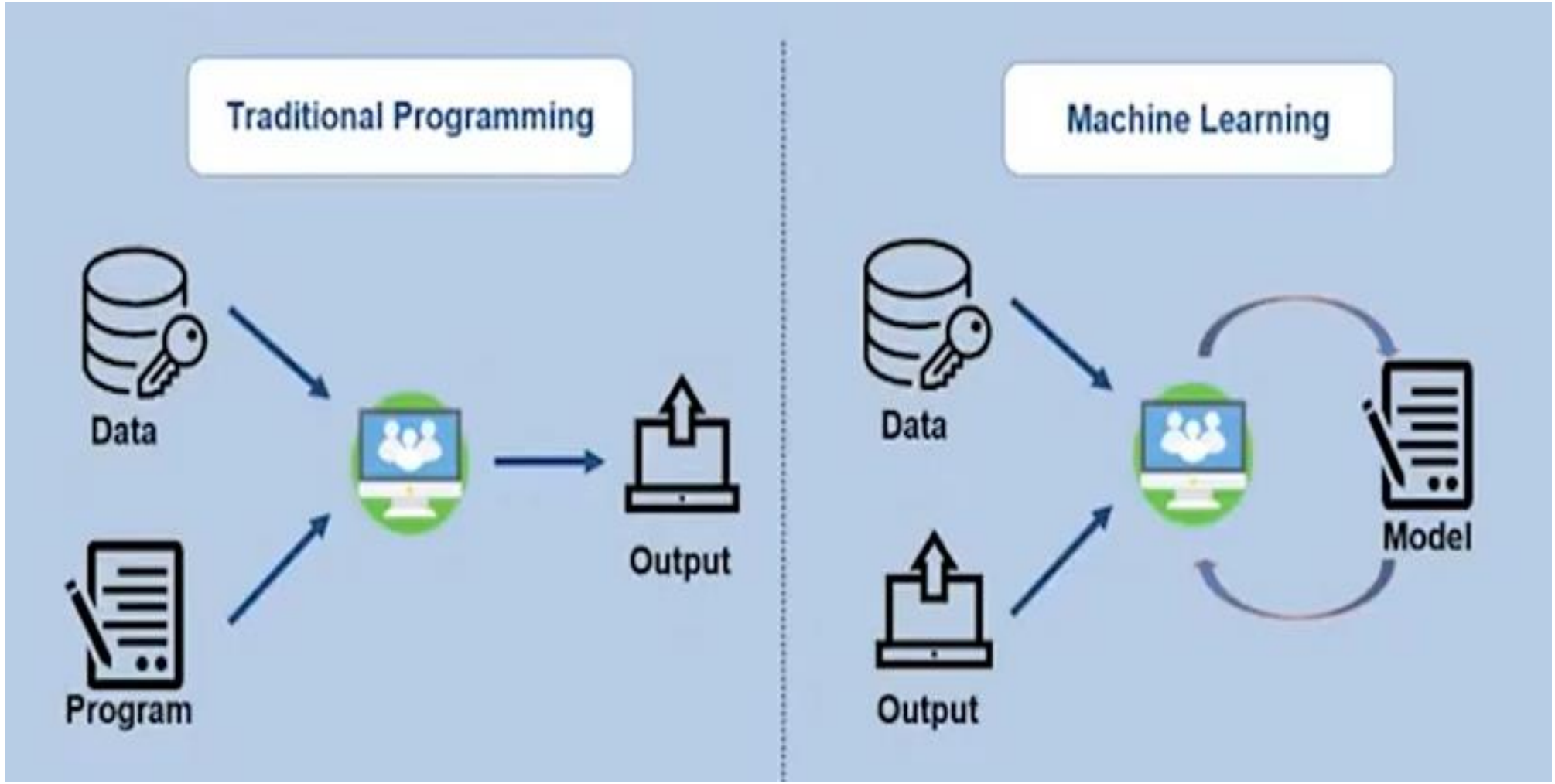
## Traditional Programming




## Machine Learning


Enable computers to learn from data without being explicitly programmed.

# Conventional versus ML algorithms

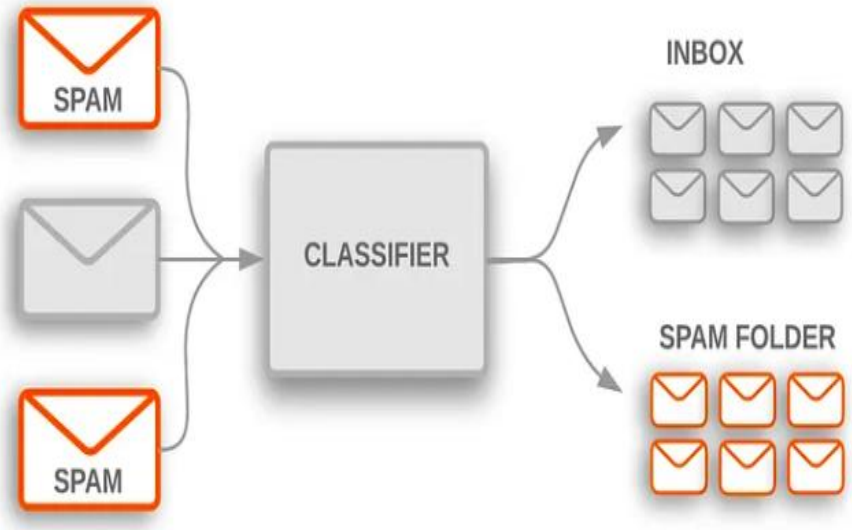


# Machine Learning Tasks

- Regression : Generalize to predict continuous valued output
  - What will be share price of SBI bank after a month ?


| | | | |
|---|---|---|---|
| Open | 991.88 | Div yield | - |
| High | 1,019.86 | Prev close | 940.67 |
| Low | 982.50 | 52-wk high | 1,019.86 |
| Mkt cap | 188.02B | 52-wk low | 207.51 |
| P/E ratio | - | | |

# Machine Learning Tasks




- Classification :

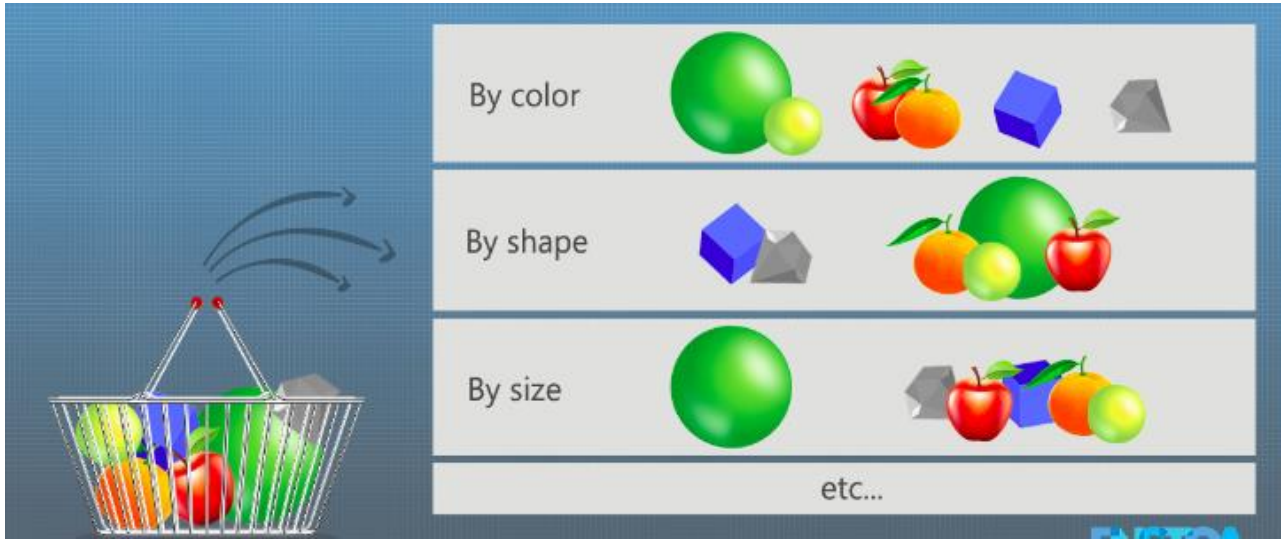
  Generalize to predict discrete valued output

  - Classify mails into spam and ham

- Clustering :

  Generalize to form groups

  - Cluster the items in a basket



# Experience for Machine Learning : Data

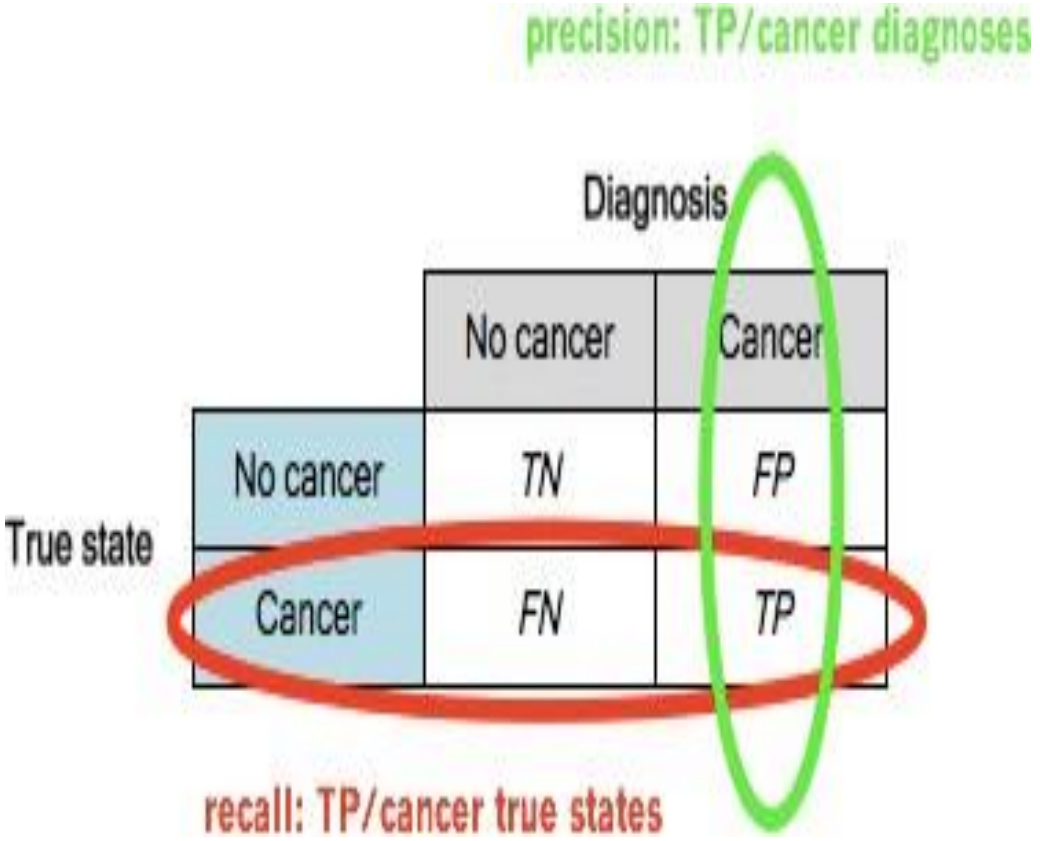
- Data : Qualitative or quantitative variables in organized/unorganized form
    - Images, Documents, Videos, Customers transactions
        - **DATA-> Information-> Knowledge**

- Information : Meaningful data
    - Customers buying grocery items
    - The documents are related to sports…

- Knowledge : Learned/Analyzed information
    - If milk is bought, then bread is also bought
    - Indian Cricketers are Sachin, Sehwag, Ganguly, Srinath, ….

# Performance measure

- Accuracy : How many of the predictions made by the model are correct
- Precision : Fraction of relevant instances among the retrieved instances
- Recall : Fraction of the total amount of relevant instances that were retrieved actually




Source : quora.com

# Confusion matrix

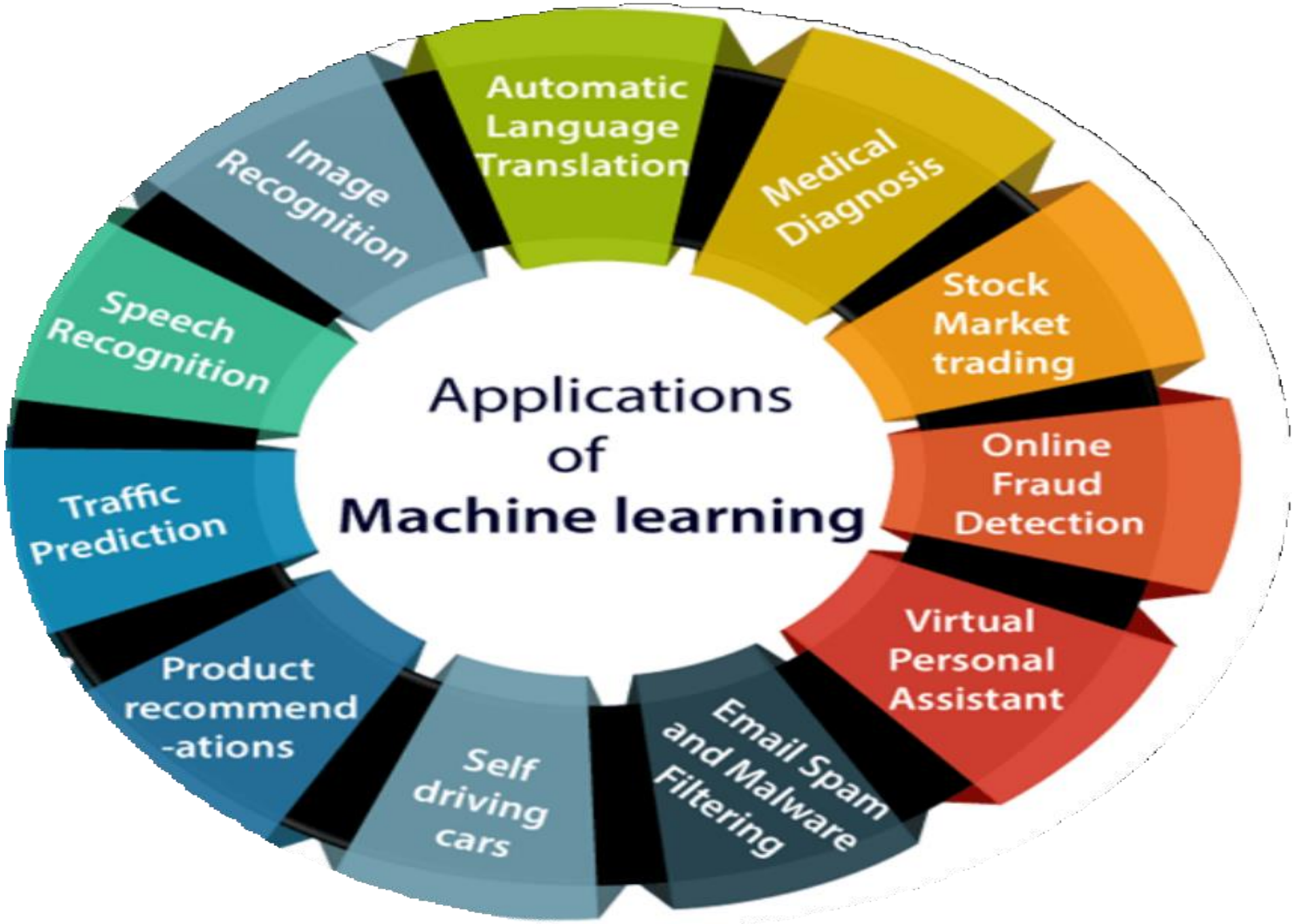
| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Actual | Dog | Dog | Dog | Not Dog | Dog | Not Dog | Dog | Dog | Not Dog | Not Dog |
| Predicted | Dog | Not Dog | Dog | Not Dog | Dog | Dog | Dog | Dog | Not Dog | Not Dog |
| Result | TP | FN | TP | TN | TP | FP | TP | TP | TN | TN |

| | | Actual | |
|---|---|---|---|
| | | Dog | Not Dog |
| Predicted | Dog | True Positive (TP =5) | False Positive (FP=1) |
| | Not Dog | False Negative (FN =1) | True Negative (TN=3) |

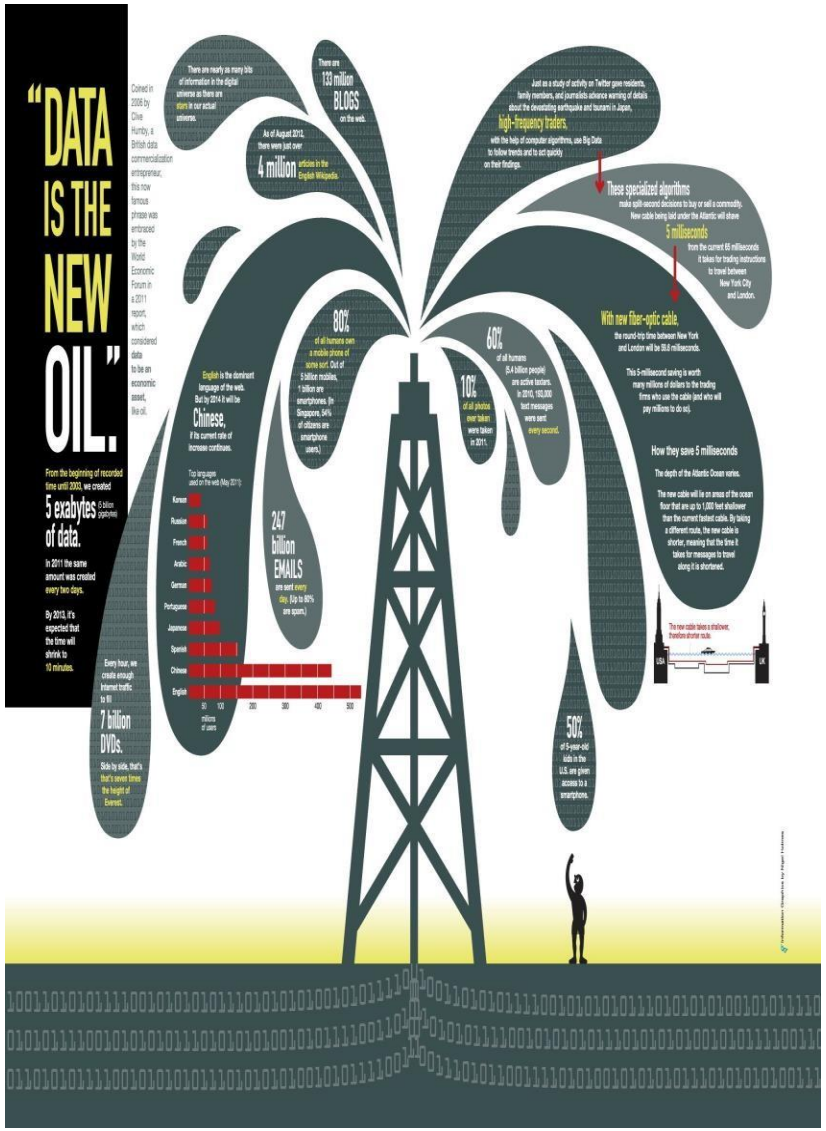
Source : geeksforgeeks.org

# Desired characteristics

- Learn and Generate models of high quality
- Generalize patterns with less error
- Learn from noisy and vague data
- Efficient and Scalable
- Deterministic


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend-ations
Traffic Prediction
Speech Recognition
Image Recognition

# Why Machine Learning hype?




- Data is the new Oil!!!
  - Enough Data but Scarce Knowledge

  - 1 trillion web pages – Web mining
  - One hour of video is uploaded to YouTube
    - every second amounting to 10 years of content every
  - 1M transactions per hour in Walmart

  - Science – Bioinformatics, Astrophysics... Genomes 1000s of people, each of which has a length of 3.8 × 109 base pairs sequenced by various labs;

# Why Machine Learning is inevitable ?

- Intellectually rich due to Mathematics and Algorithms

- Human design has limitations
  - NLP, Computation, , handwriting recognition,

- Dynamic Environment
  - Autonomous helicopter, self driving car

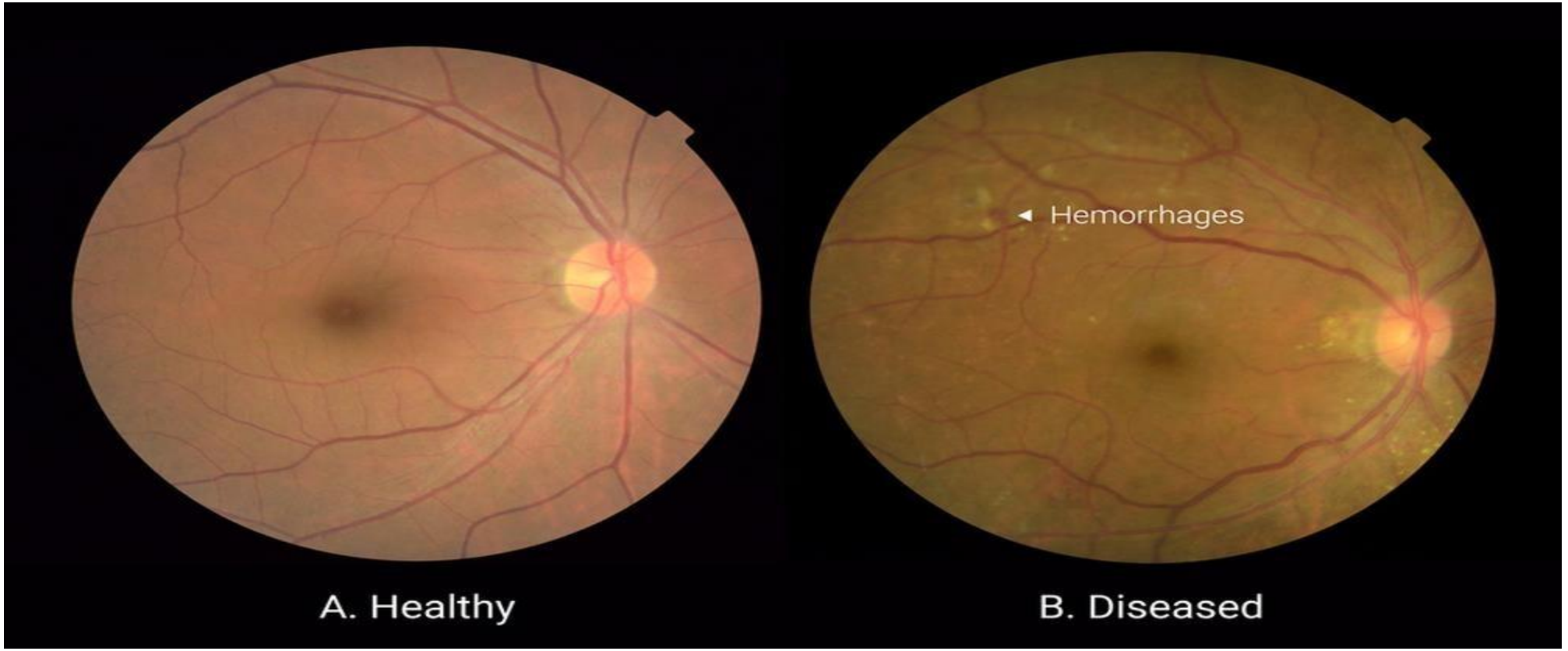
- New knowledge evolved and redesign is difficult
  - Self customizing programs
    - Amazon product recommender system.

      - DATA DOES THE TRICK!!!

# Machine Learning Applications

| Applications | Stakeholders |
| --- | --- |
| • Disease Diagnostics | • Doctors |
| • Recommendation | • Netflix, Amazon |
| • Image Tagging | • Social media like Facebook |
| • Advertising | • Web service providers |
| • Speech Processing | • Google assistant, Amazon Alexa |
| • Self-driving car | • Tesla |
| • Share Trading | • Financiers |

# Disease Diagnostics (Computer Vision)




Diagnose Diabetic Retinopathy

Source : https://blog.google/technology/ai/detecting-diabetic-eye-disease-machine-learning/

# T, E, P for Diabetic Retinopathy

- Task : Specific work
    - Diagnose if the patient has Diabetic Retinopathy or not (Classification)
- Experience : Data available
    - Eye images of various DR/non-DR patients
- Performance : Accurately perform work Correctly diagnosed DR patients

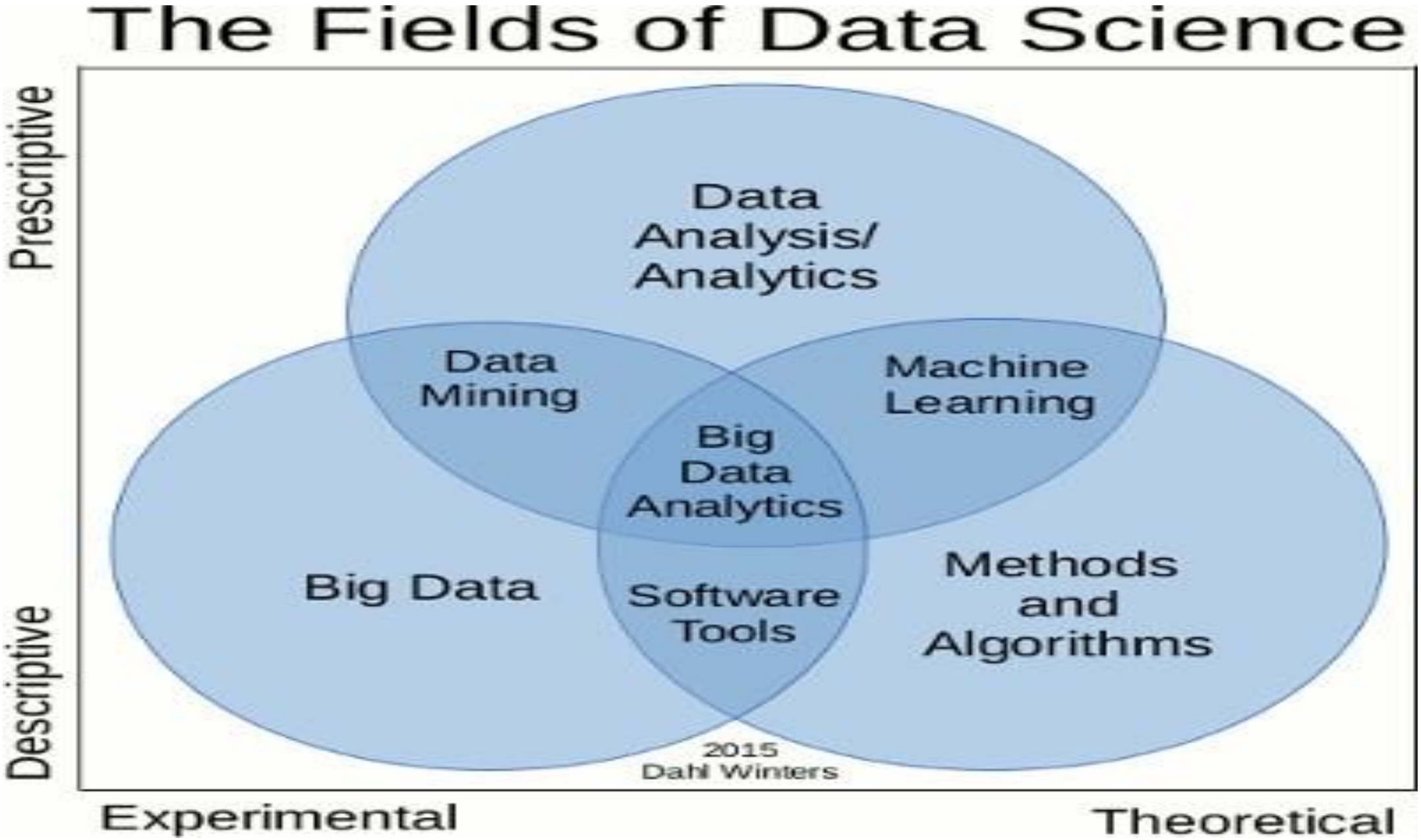
# Recommendation Systems

# Share Trading

# T, E, P for Share Trading

- Task : Specific work
    - What will be the share price of HDFC after a week (Regression)
- Experience : Data available
    - For each date, the opening price, closing, high, low buying etc. since 5 years
- Performance : Accurately perform work
    - How closer was the predicted price to the actual price.


ONLINE
ADVERTISING
AD
AFFILIATE
ACTION
LISTING
BEHAVIORAL
SALE
USUALLY
TIME
PER
WEBSITE
MARKETING
ADS
CONTENT
SYSTEM
LEAD
REVENUE
BANNER
USERS
COOKIES
MAY
COMMON
PAGE
CLICK
USER
COST
PAY
WEB
E-MAIL
LARGE
USE
SEMANTIC
TARGETING
ADVERTISERS
NUMBER
IMPRESSIONS
KNOWN
ADVERTISEMENTS
ENGAGEMENT
SEARCH
BASED
WEBSITES
ADVERTISEMENT
PRIVACY
SITES
MANY
ALSO
LIKE


The Fields of Data Science
Prescriptive
Descriptive
Experimental
Theoretical
Data Analysis/ Analytics
Data Mining
Machine Learning
Big Data Analytics
Big Data
Software Tools
Methods and Algorithms
2015
Dahl Winters

# Types of MachineLearning

# Data Representation

| | f1 | f2 | f3 | f4 | f5 | Y |
|---|---|---|---|---|---|---|
| X1 | | | | | | |
| X2 | | | | | | |
| X3 | | | | | | |
| X4 | | | | | | |
| X5 | | | | | | |
| X6 | | | | | | |
| X7 | | | | | | |

- Y : Target/Label/outcome/response/dependent variable
- (f1,….,fd) : Features/Attributes/Dimensions
    - Independent and Identically Distributed vectors

- (X1,X2,….,Xn) : Samples/Rows/Tuples/Instances/ Observations

- Time series (Dependent vectors)
- Images (Matrices)
- Variable size Non-vector data(Trees, Graphs, Text)
- Objects (Relational Schema)

# Approaches for Learning

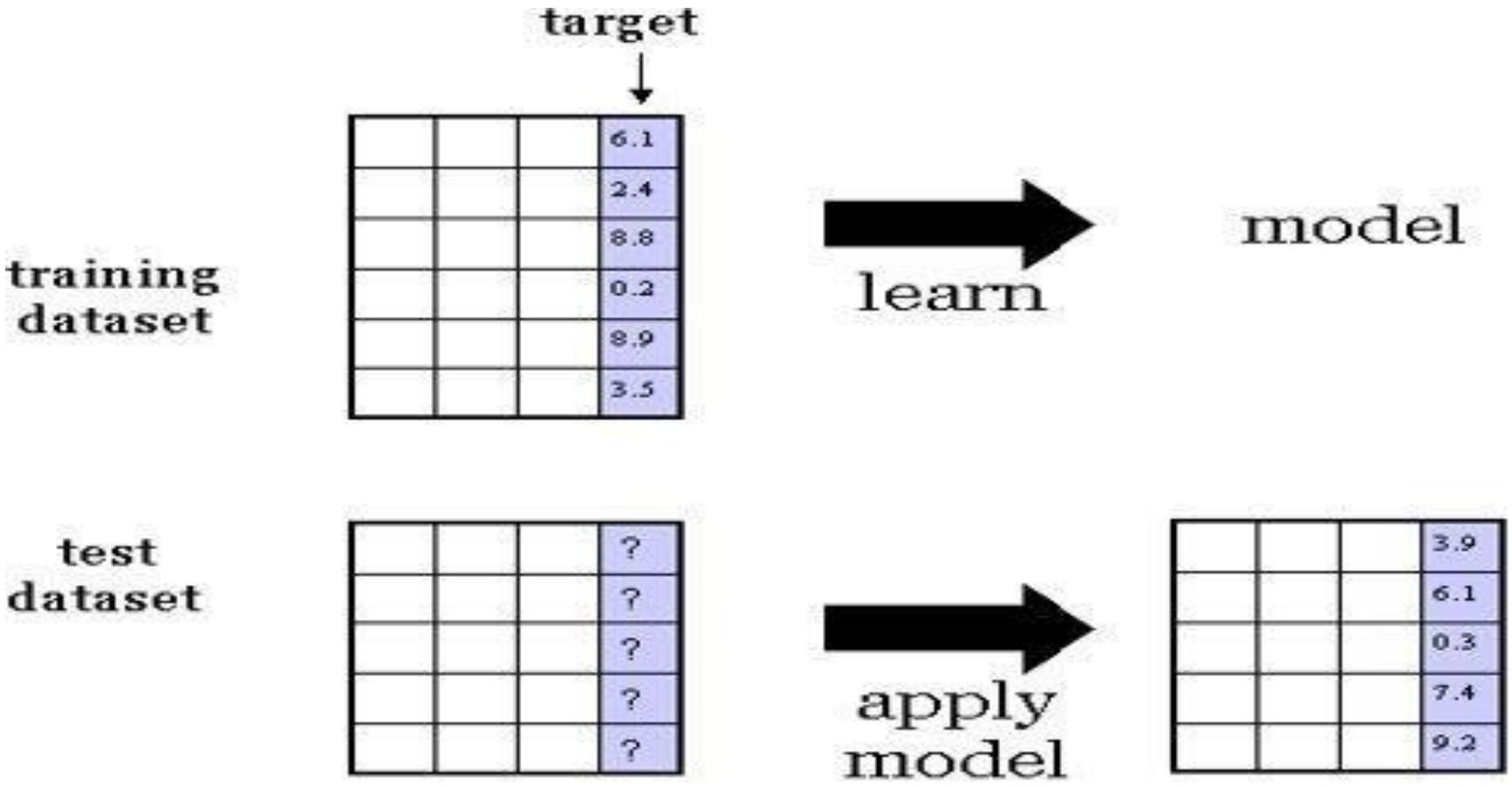
Supervised

UnSupervised

Reinforcement

# Supervised Learning from Data

- Predicting continuous target variable



# Task : Regression

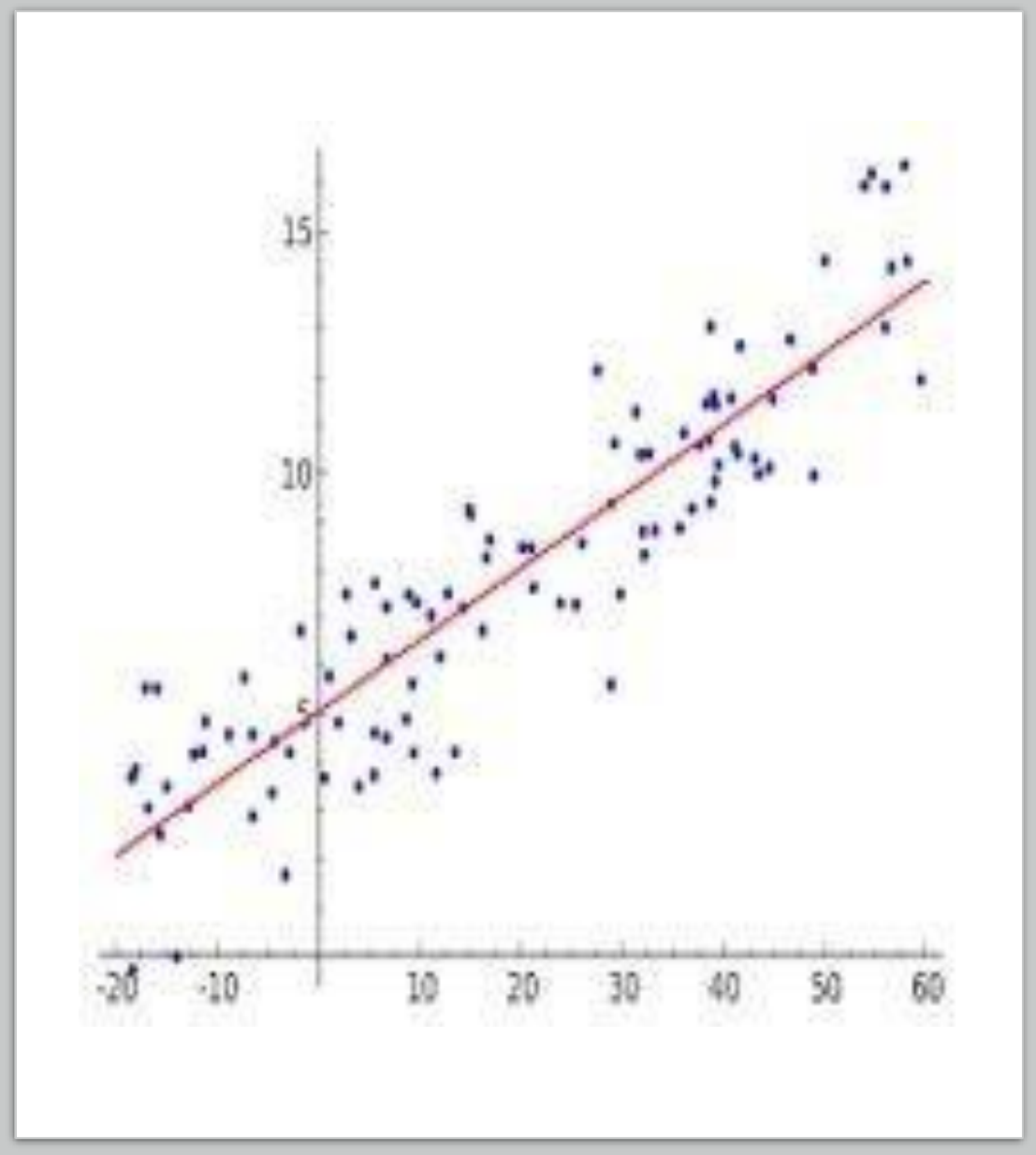
- Example : Stock Price Prediction

- Only one target variable ?

- How to take a combination of the features (linear regression)?

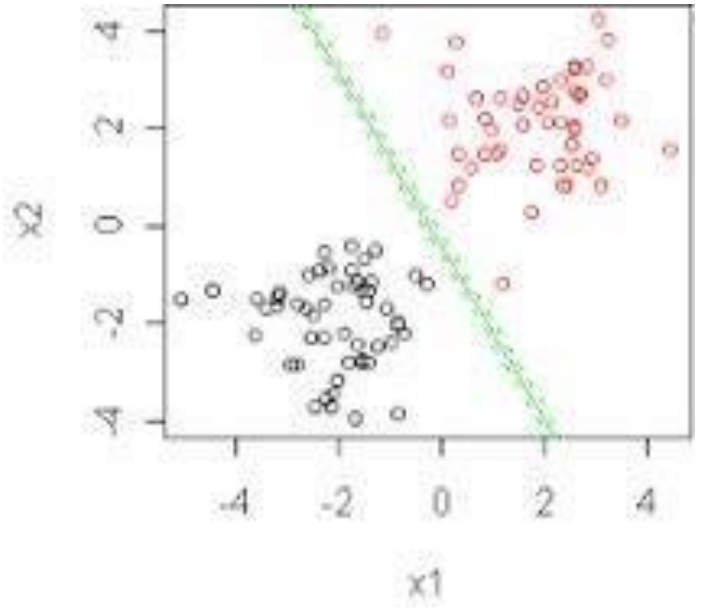
- Anything more general than that?

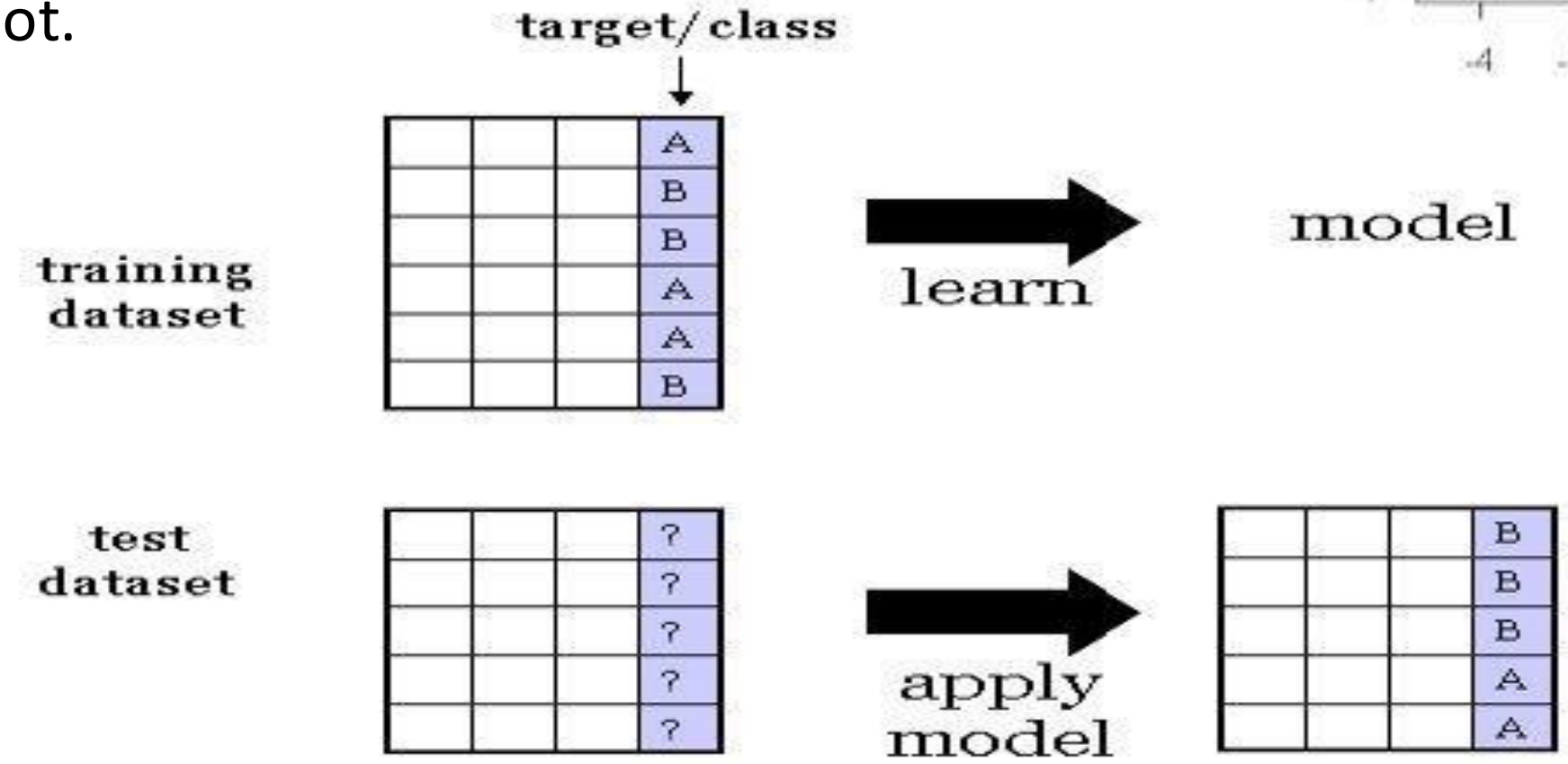
- Is local minimum error the global minimum error ?



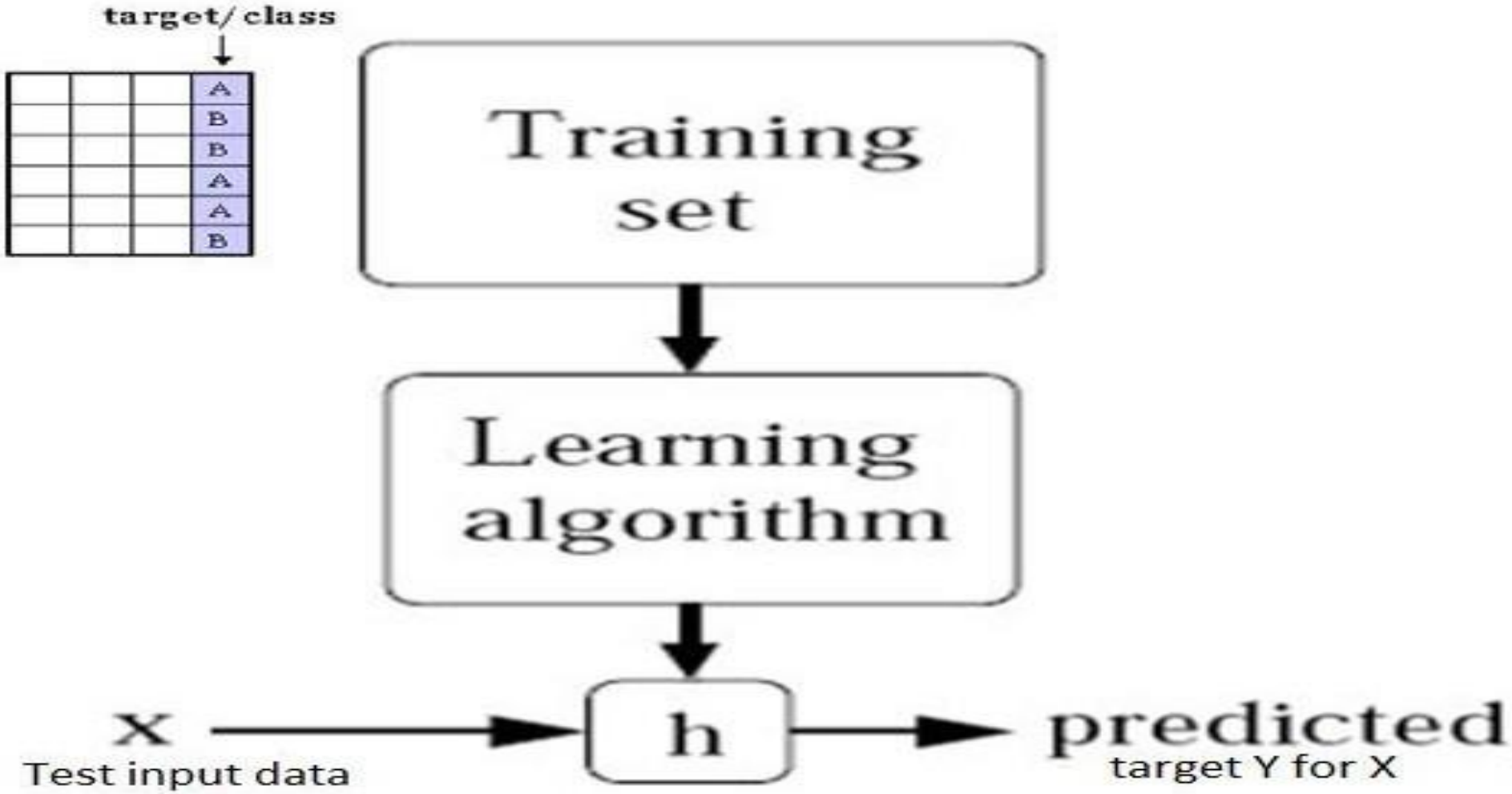
# Task : Classification

- Example : Classify whether the patient is suffering from cancer or not.



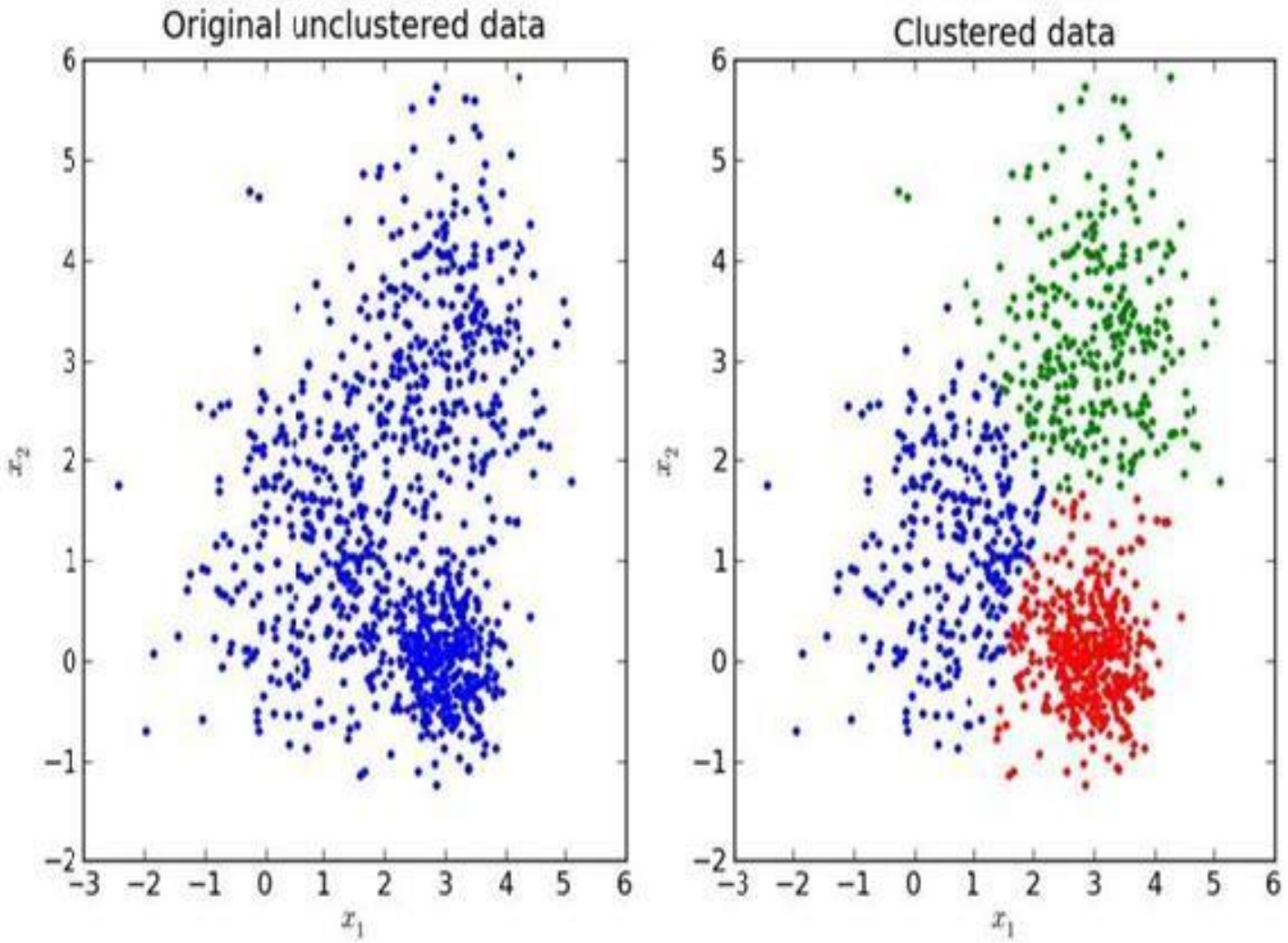



# Supervised Learning



# Unsupervised Learning

- Task : Clustering
- No Target variable



# Task : Association mining

| TID | Items |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

**{milk} → {bread}**
**{milk, cheese, eggs} → {bread}**
**{milk, diaper} → {beer}**

# Task : Dimensionality Reduction

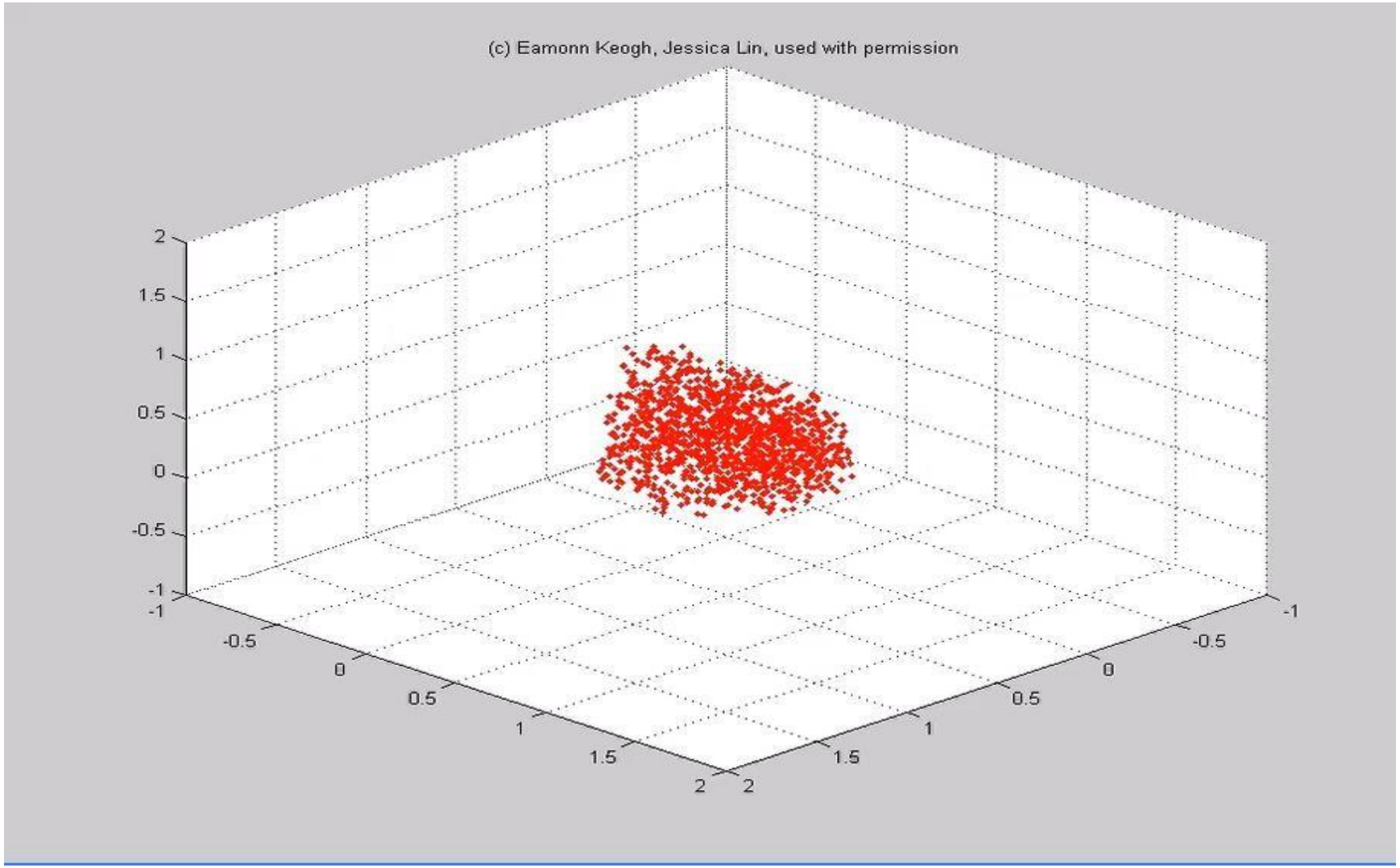
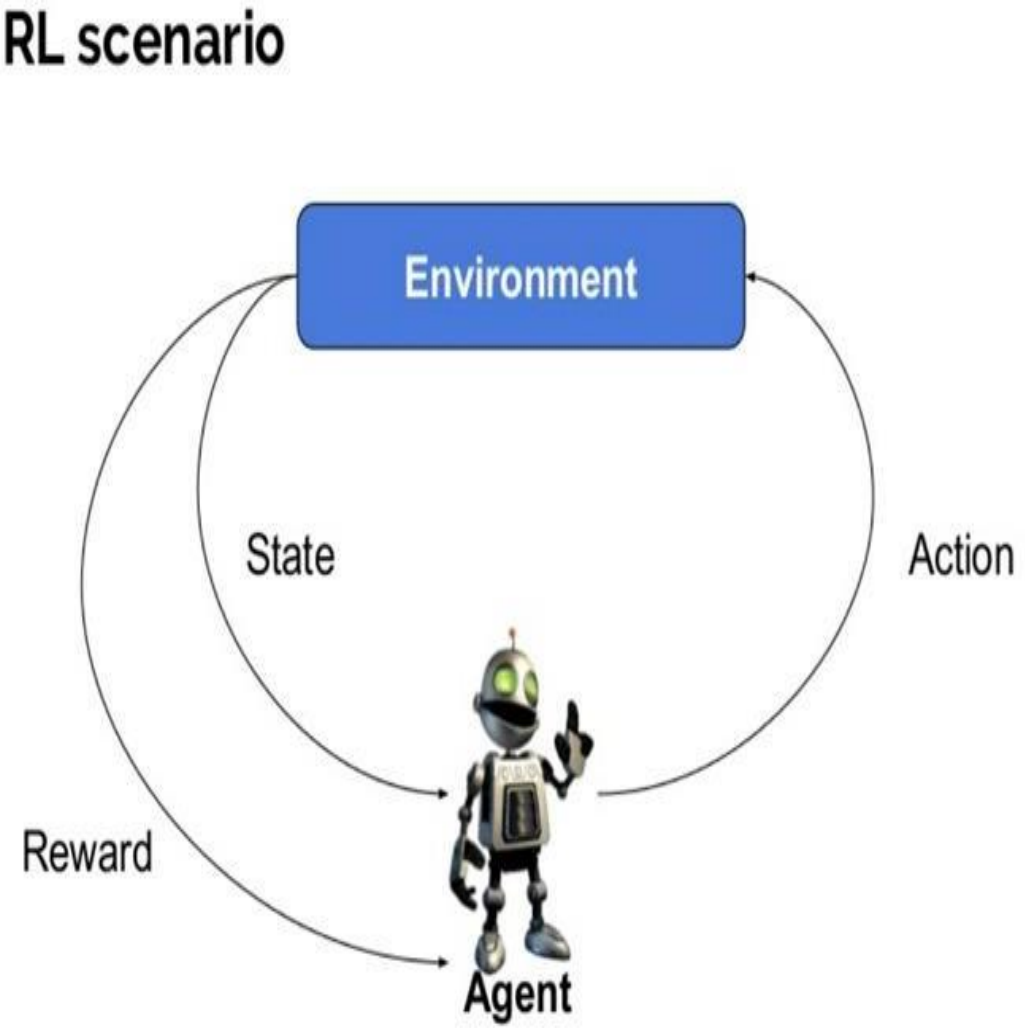



Image Source : https://cs.gmu.edu/~jessica/DimReducDanger.htm

# Reinforcement Learning



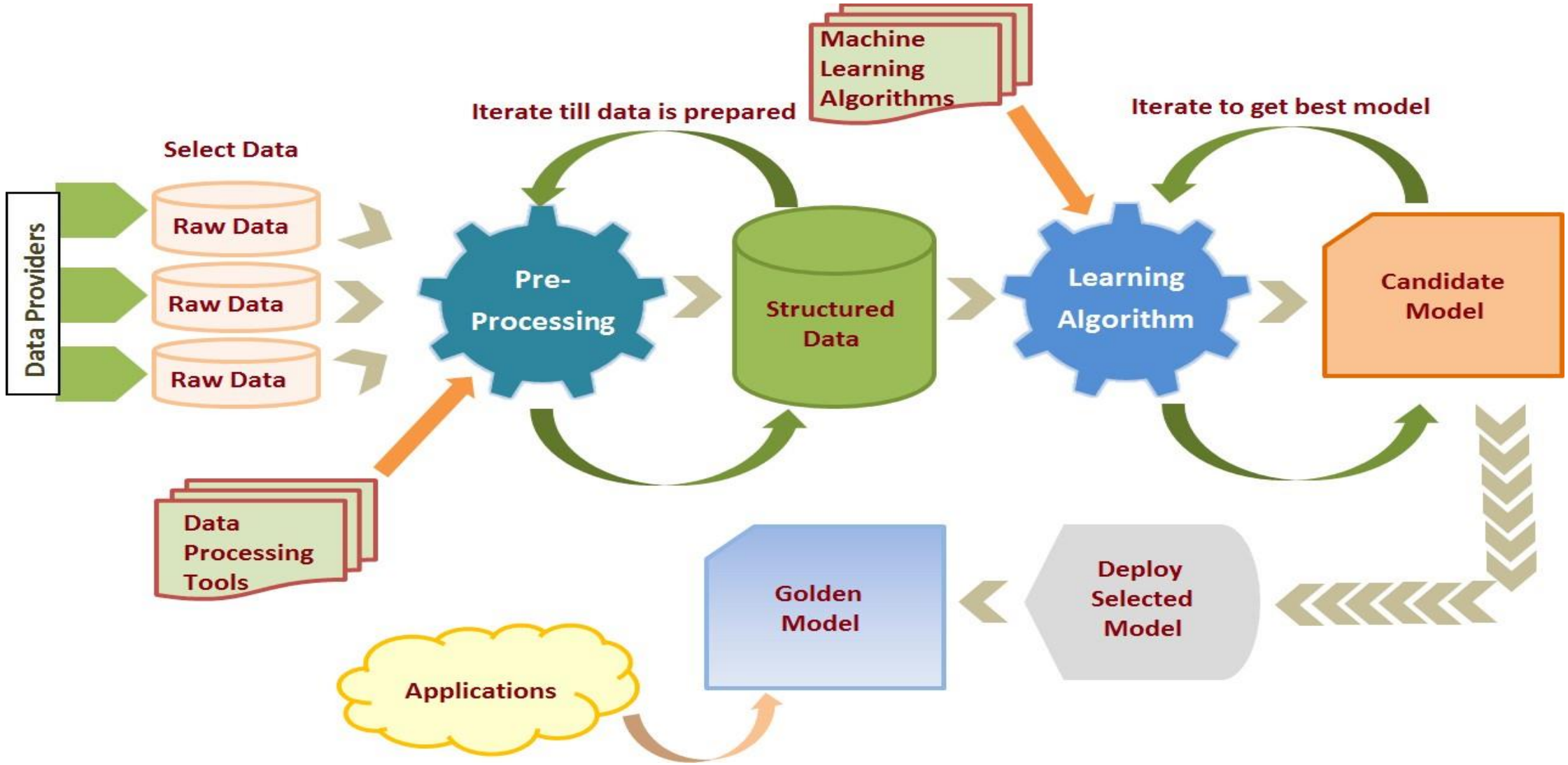

- Learn from interaction with the environment
- Transition from one state to another
- Rewards/punishments from sequence of actions

Image Source : https://www.guru99.com/reinforcement-learning-tutorial.html

# Applications

- Manufacturing
  - Pick a device and put it in right container
- Self-driving cars
  - Detect obstacles, proper routing, traffic signal
- Power systems
- Network routing

# Data Mining and Machine Learning Process



# References


- https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08
- https://www.houseofbots.com/news-detail/11973-1-clarifying-differences-between-data-analysis-data-mining-data-science-machine-learning,-and-big-data
- https://www.statisticshowto.com/well-posedness/
- https://www.quora.com/What-is-the-best-way-to-understand-the-_terms-precision-and-recall