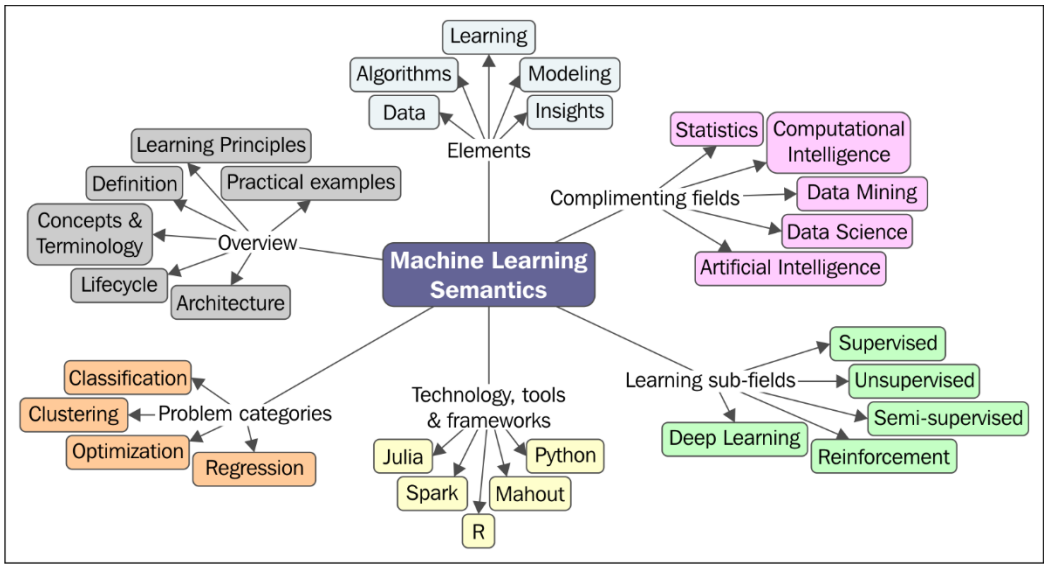
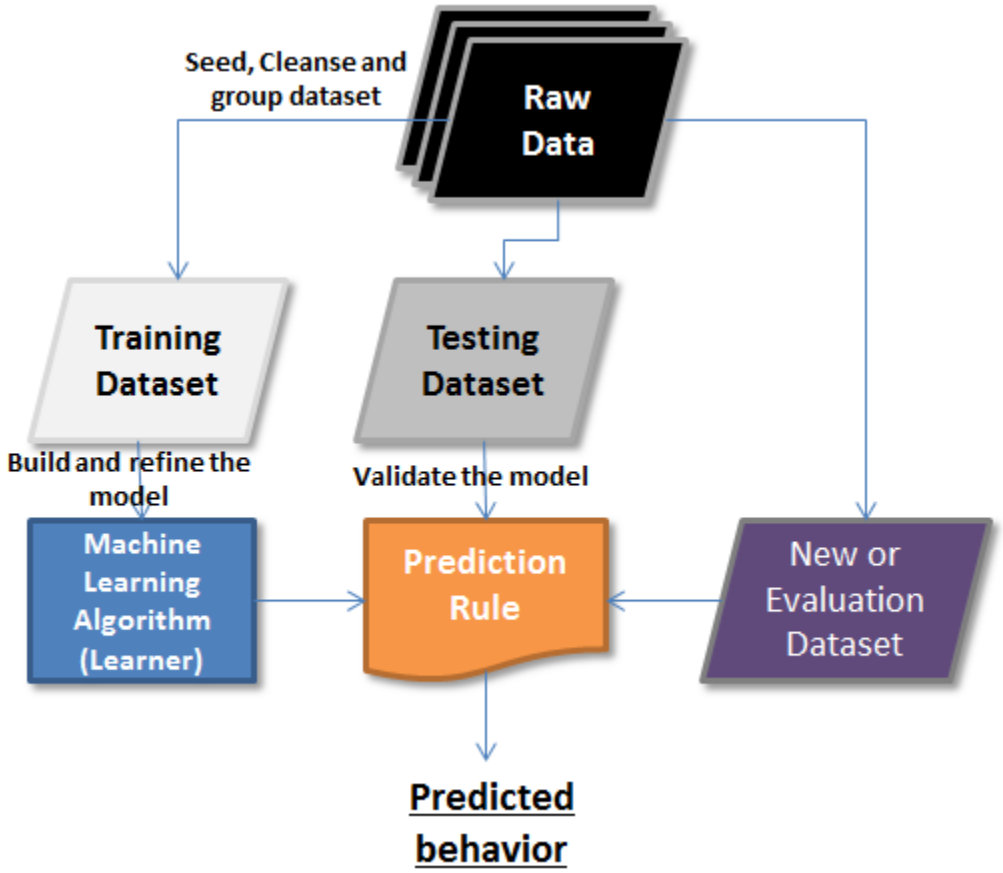
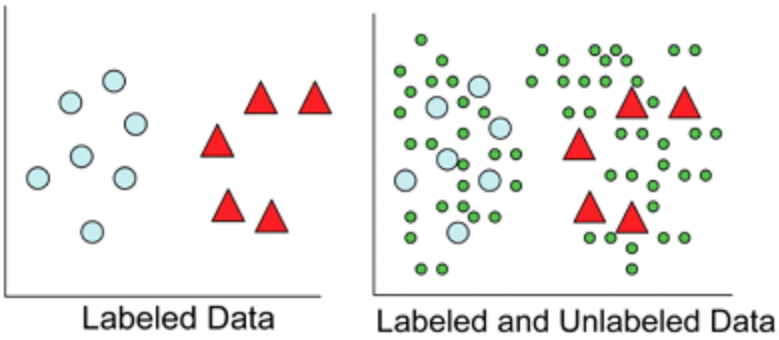
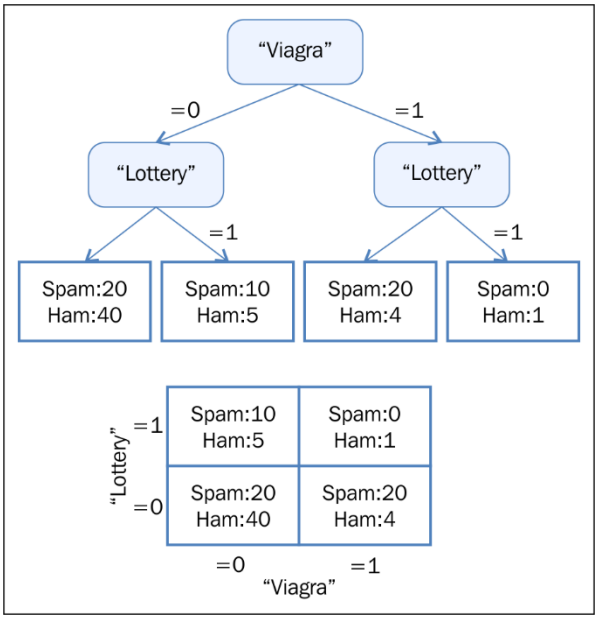
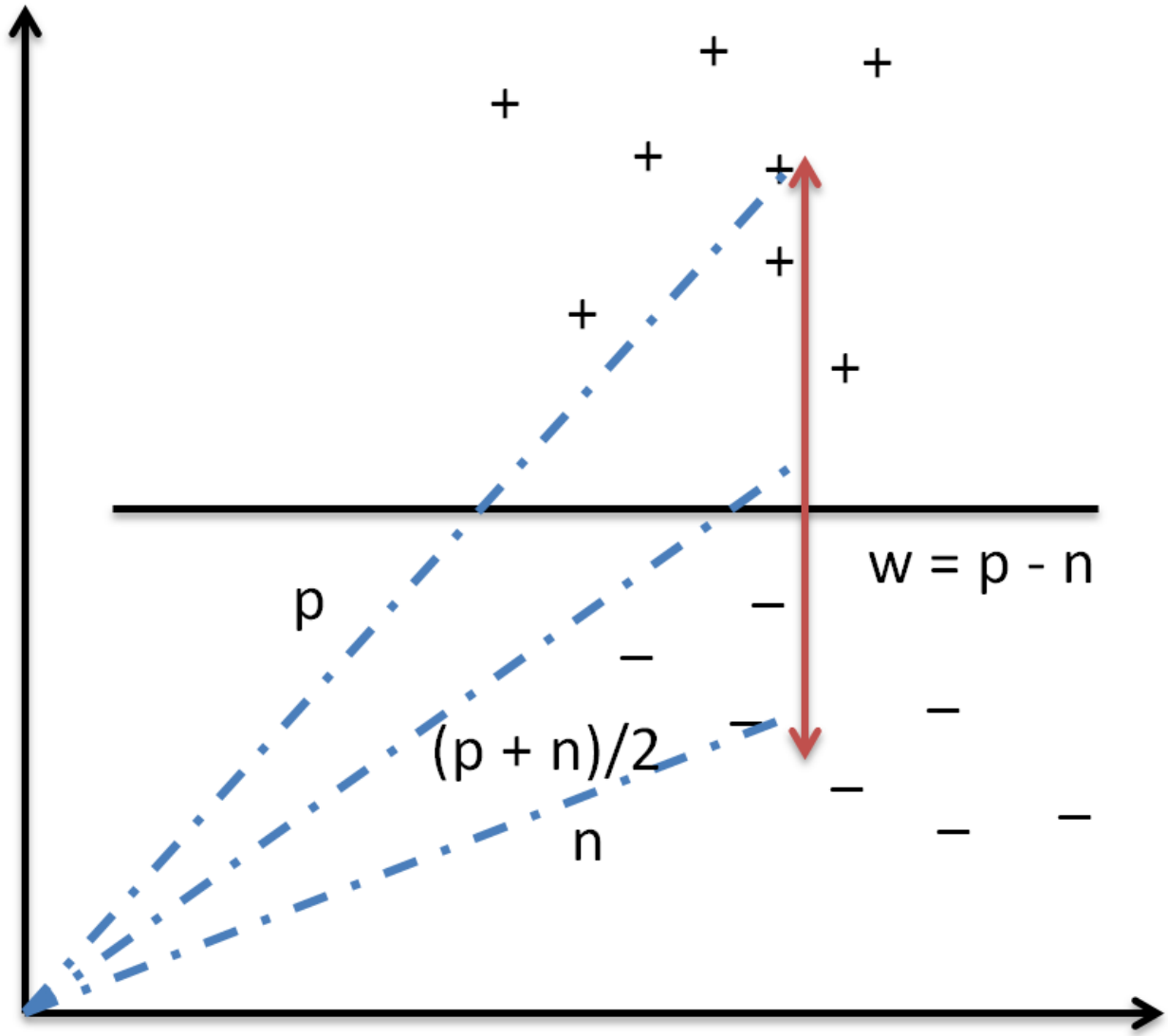
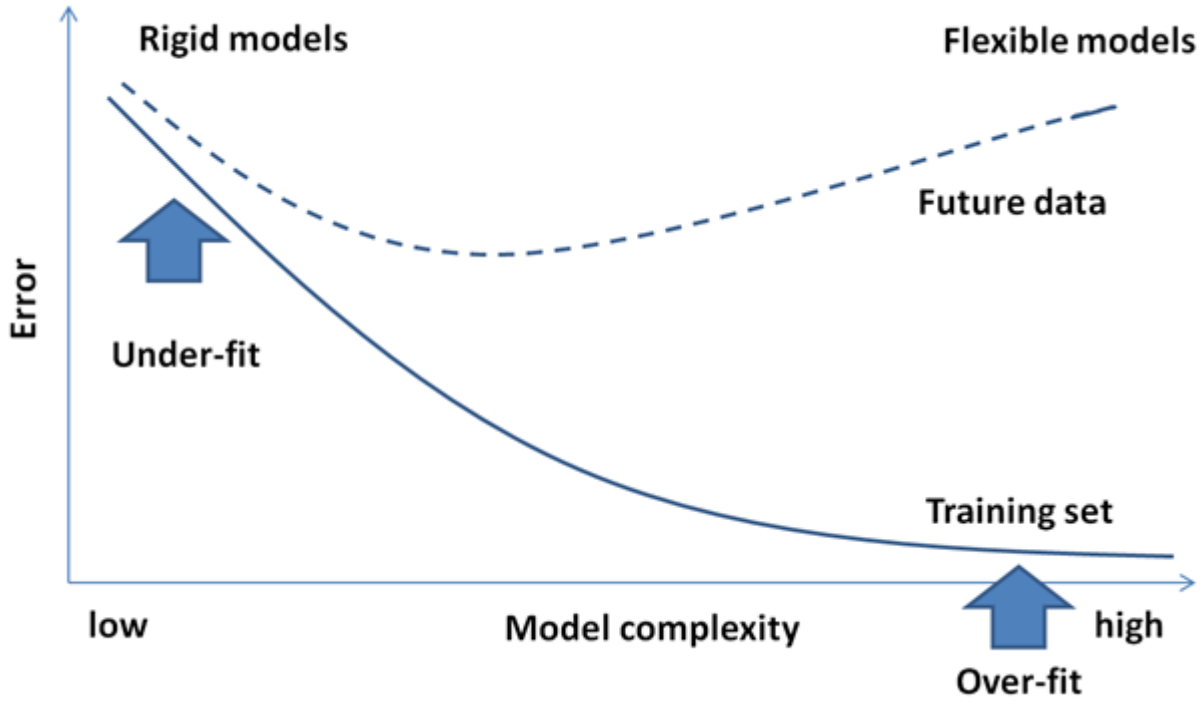

Machine Learning Semantics
Learning
Algorithms
Modeling
Data
Insights
Elements
Learning Principles
Definition
Practical examples
Concepts & Terminology
Overview
Lifecycle
Architecture
Statistics
Computational Intelligence
Complimenting fields
Data Mining
Data Science
Artificial Intelligence
Classification
Clustering
Problem categories
Optimization
Regression
Technology, tools & frameworks
Julia
Python
Spark
Mahout
R
Learning sub-fields
Supervised
Unsupervised
Semi-supervised
Deep Learning
Reinforcement

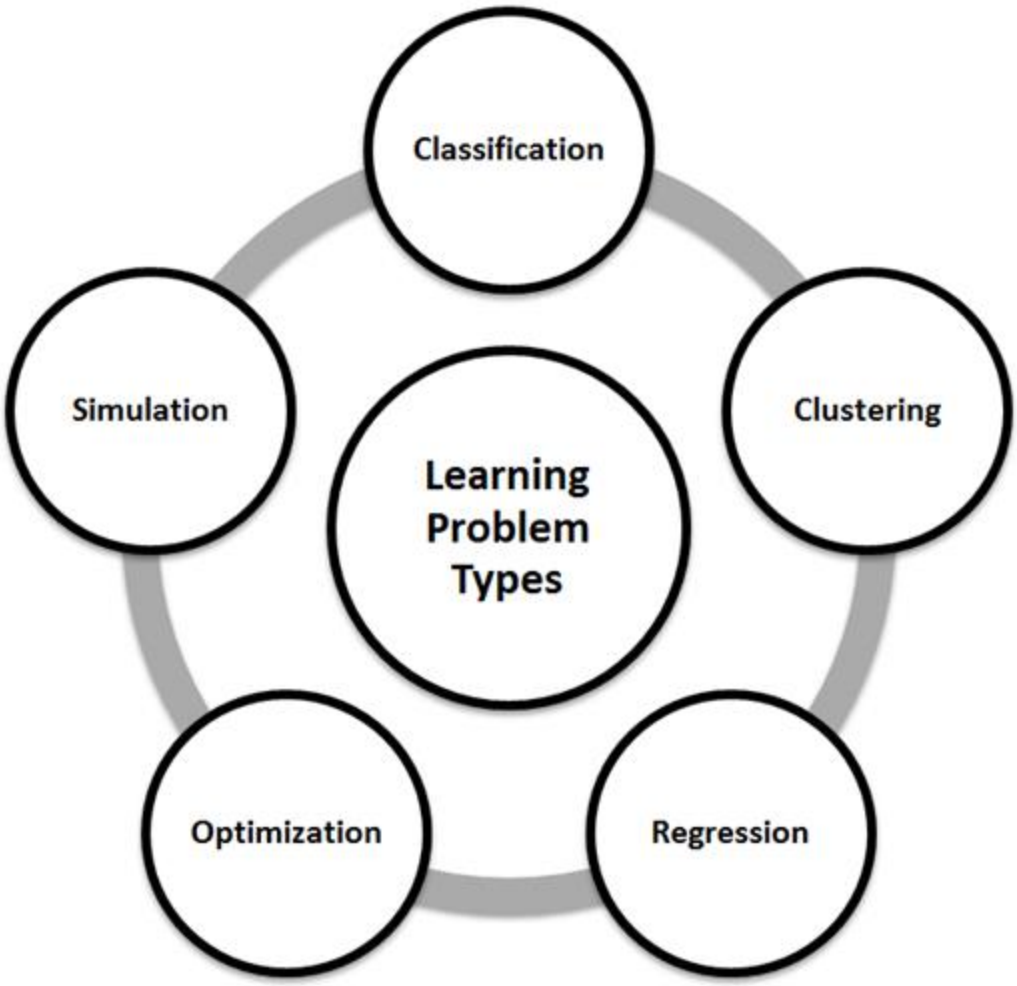

Seed, Cleanse and group dataset
Raw Data
Training Dataset
Testing Dataset
Build and refine the model
Validate the model
Machine Learning Algorithm (Learner)
Prediction Rule
New or Evaluation Dataset
Predicted behavior

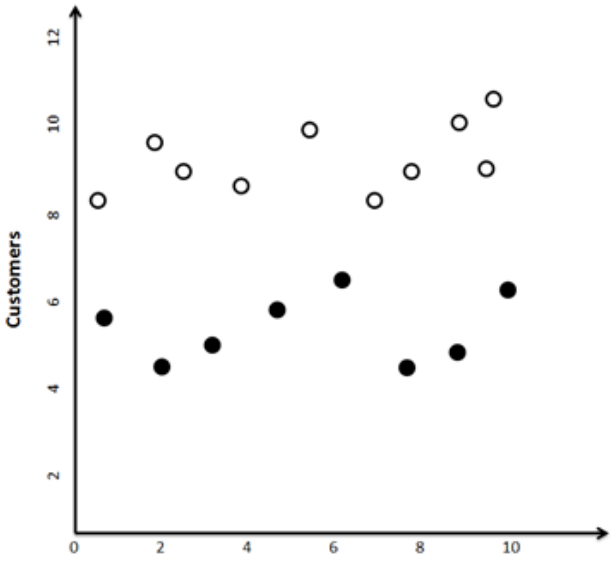

Labeled Data
Labeled and Unlabeled Data

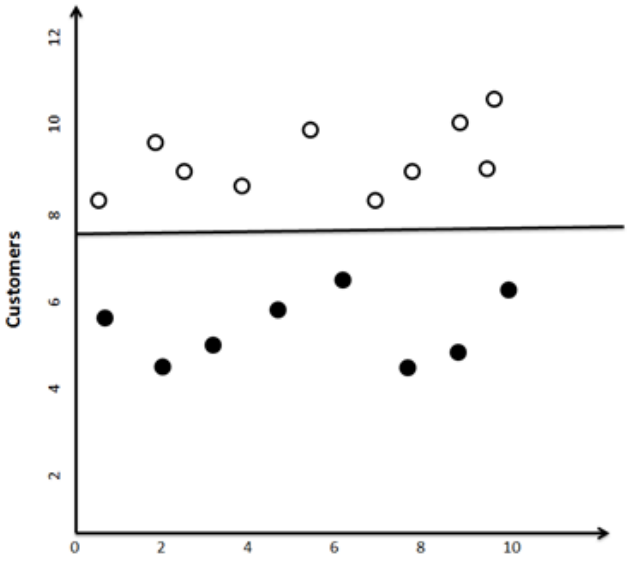

"Viagra"
=0
=1
"Lottery"
"Lottery"
=1
=1
Spam:20
Ham:40
Spam:10
Ham:5
Spam:20
Ham:4
Spam:0
Ham:1
"Lottery"
=1
=0
Spam:10
Ham:5
Spam:0
Ham:1
Spam:20
Ham:40
Spam:20
Ham:4
=0
=1
"Viagra"

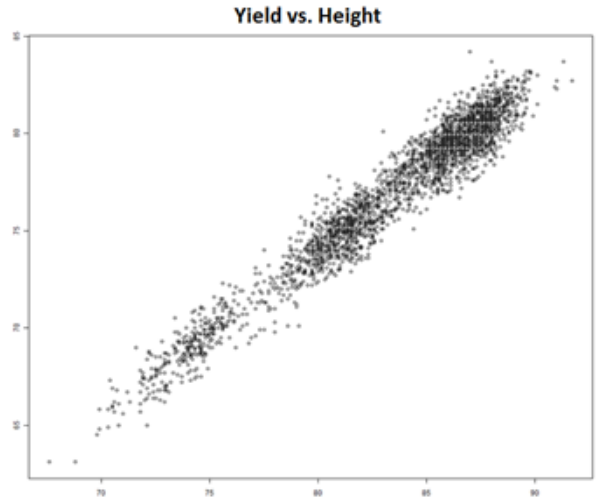

w = p - n
p
(p + n)/2
n

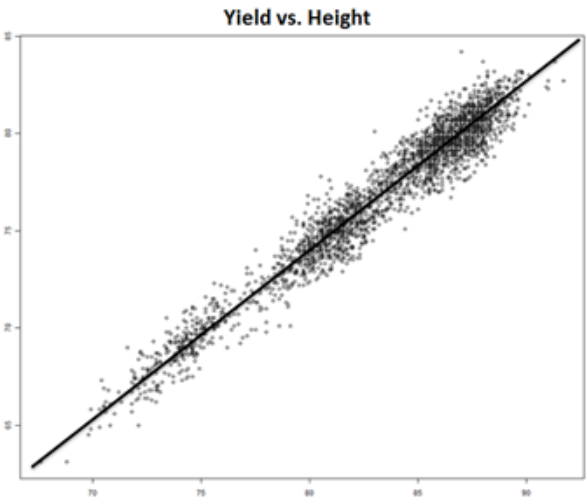

Rigid models
Flexible models
Future data
Error
Under-fit
Training set
low
Model complexity
high
Over-fit

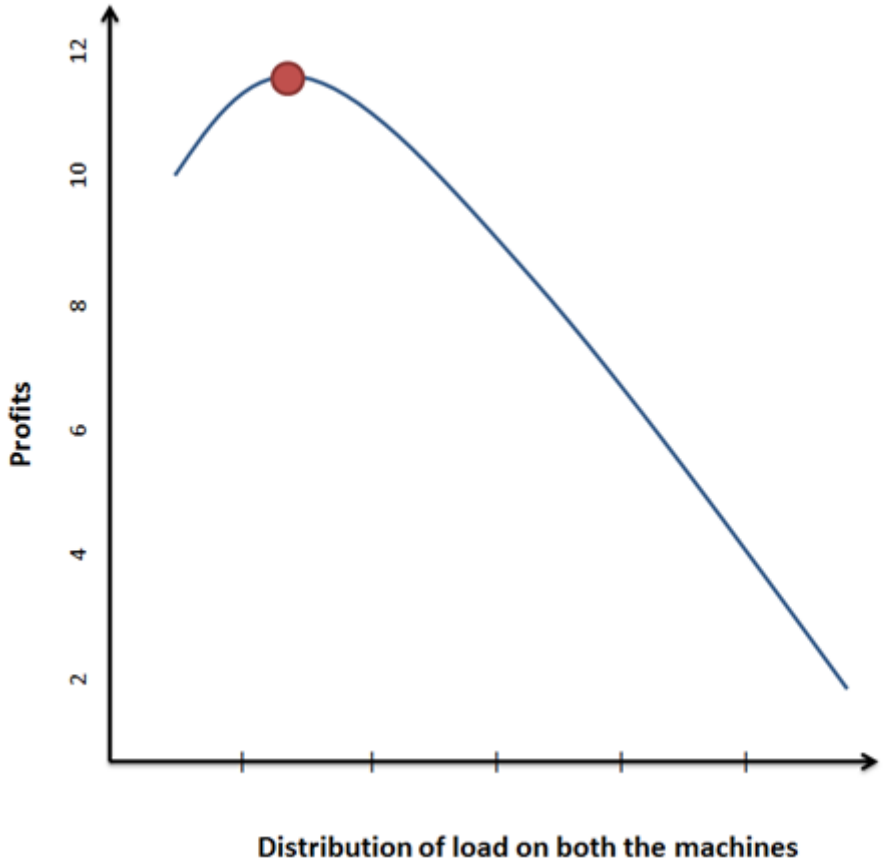

Classification
Simulation
Clustering
Learning Problem Types
Optimization
Regression

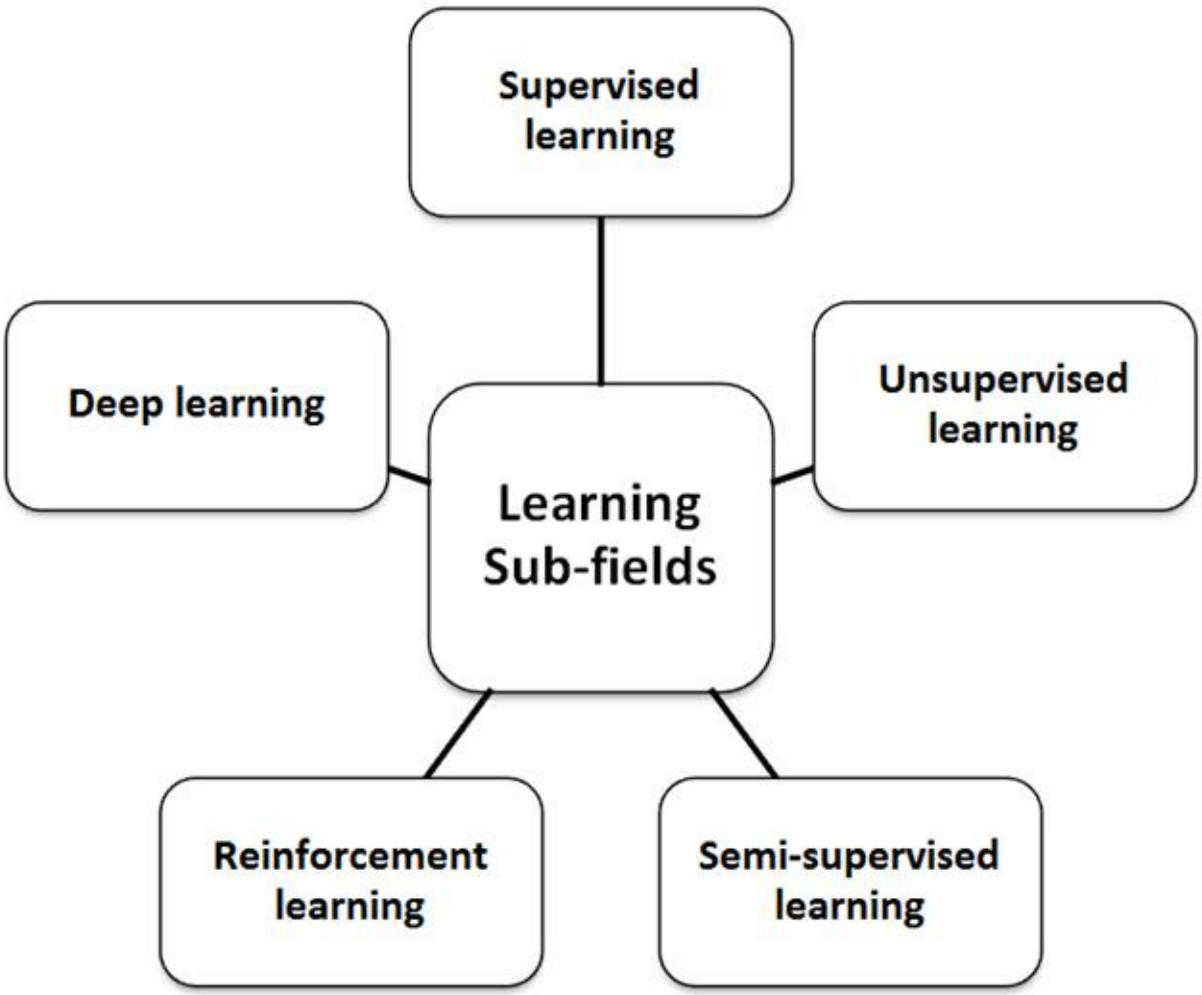

Customers
12
10
8
6
4
2
0
2
4
6
8
10

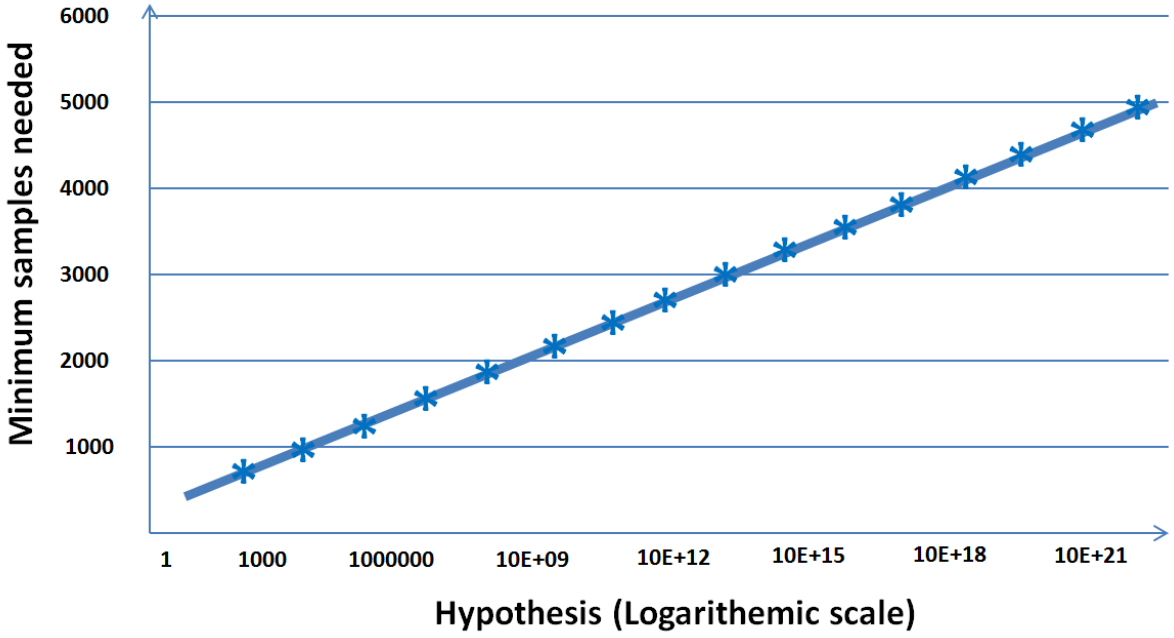

Customers
12
10
8
6
4
2
0
2
4
6
8
10


Yield vs. Height


Yield vs. Height


Profits
2
4
6
8
10
12
Distribution of load on both the machines

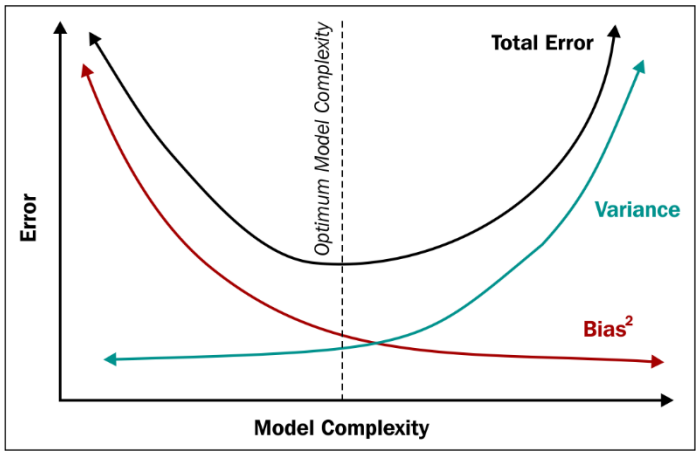

Supervised learning
Deep learning
Unsupervised learning
Learning Sub-fields
Reinforcement learning
Semi-supervised learning

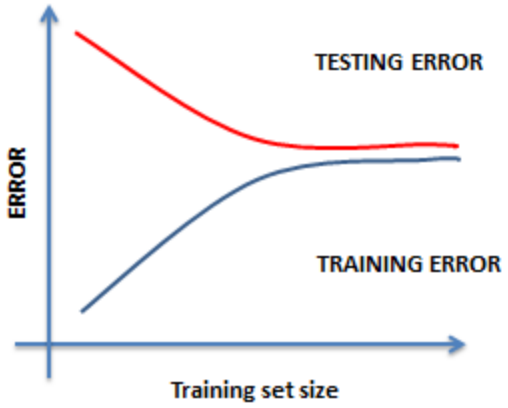


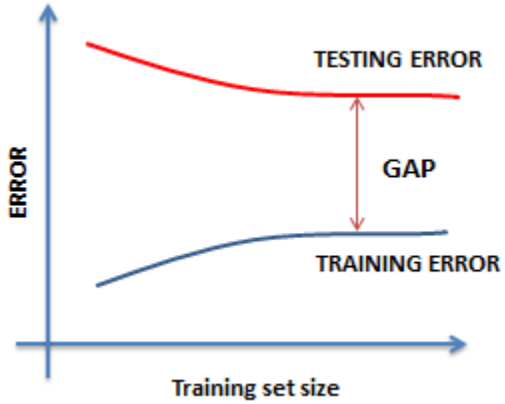

$$MSE = \frac{\sum_{i=1}^{n} \left(P_i - A_i\right)^2}{n}$$

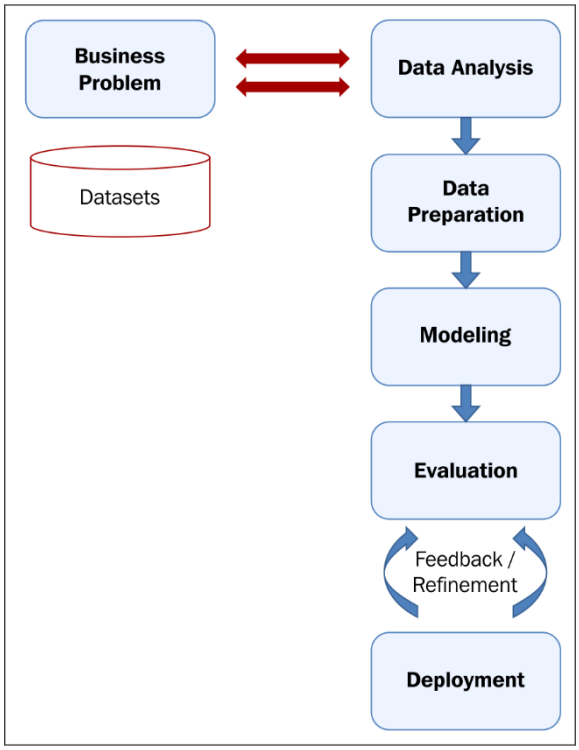
$$MAE = \frac{\sum_{i=1}^{n} \left|P_i - A_i\right|}{n}$$

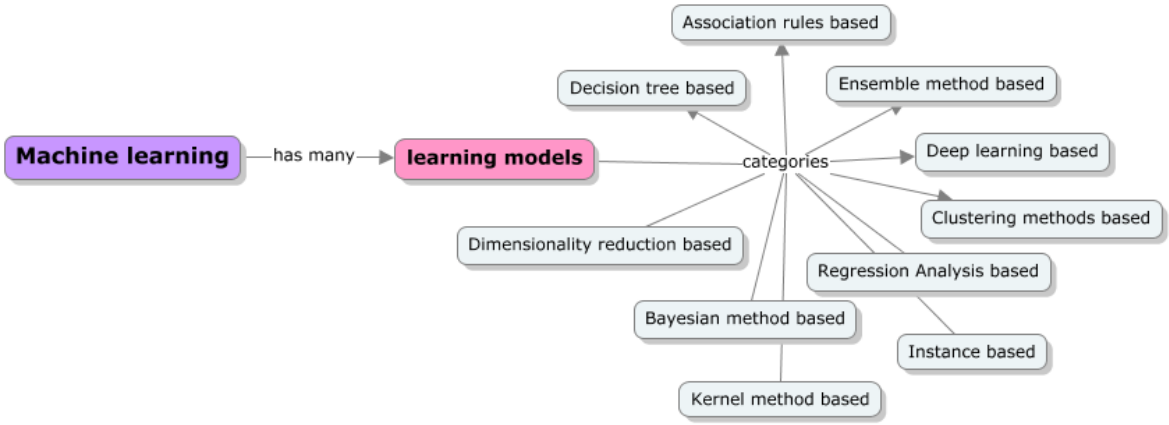
$$NMSE = \frac{MSE\ of\ developed\ model}{MSE\ of\ naive\ model}$$

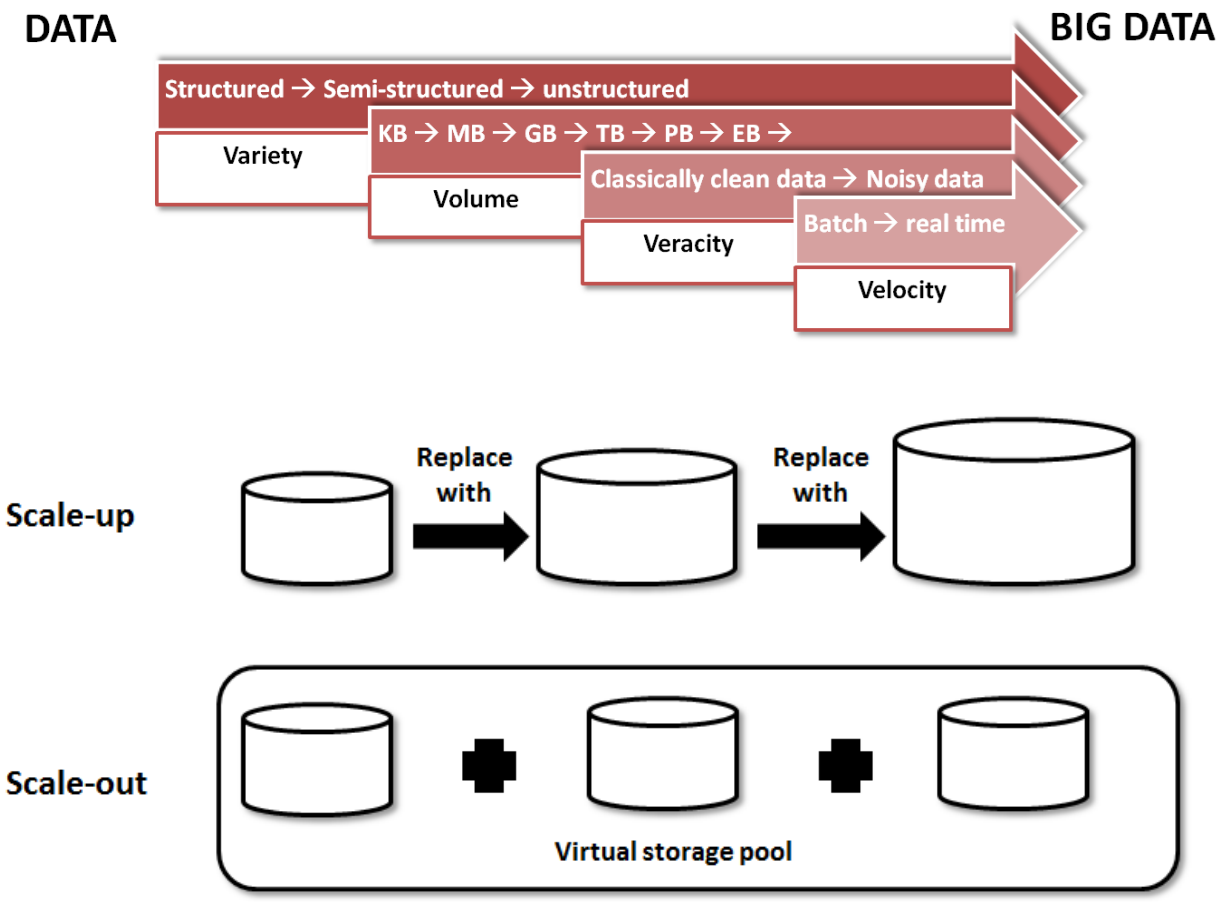


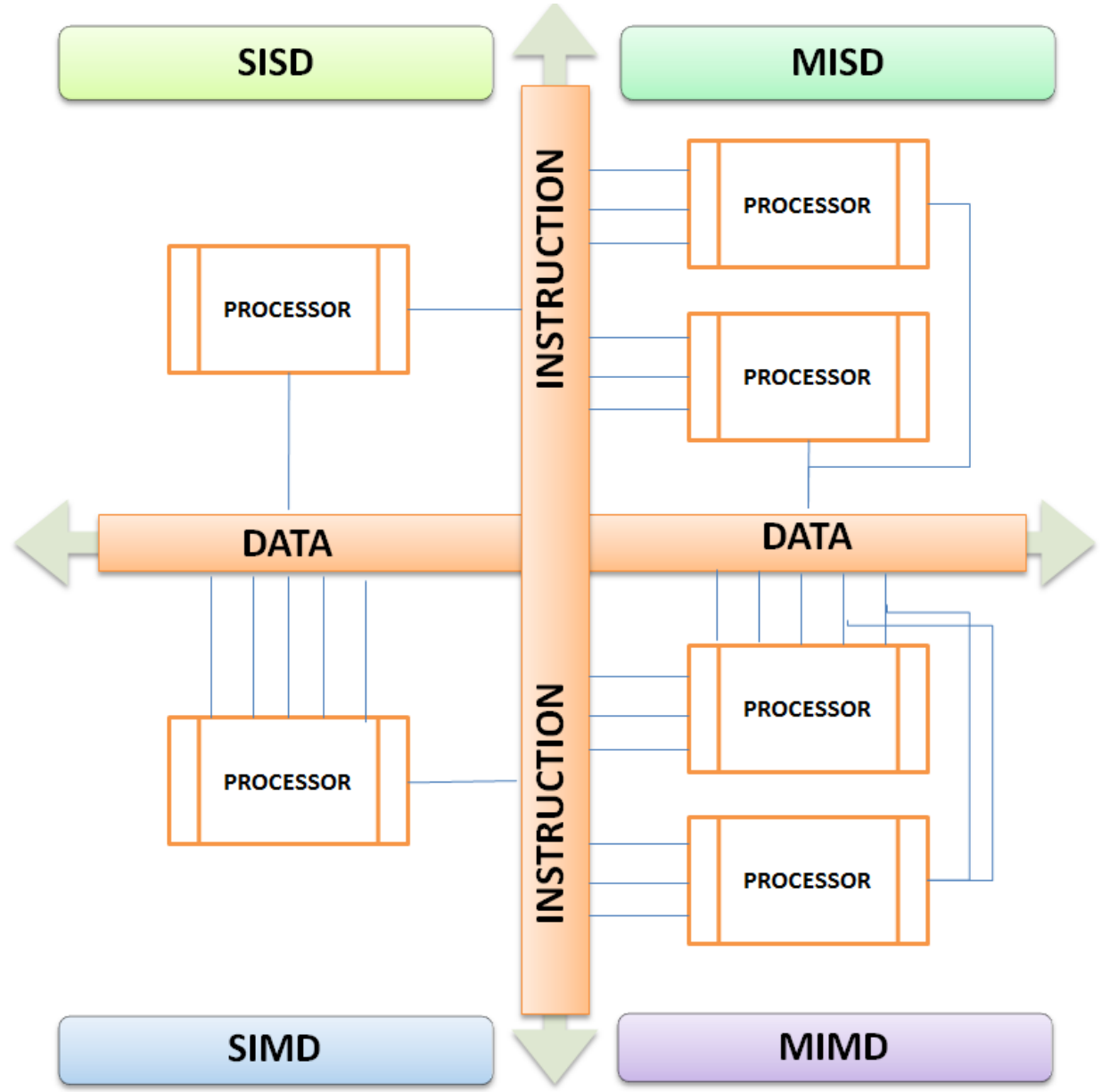

TESTING ERROR
ERROR
TRAINING ERROR
Training set size

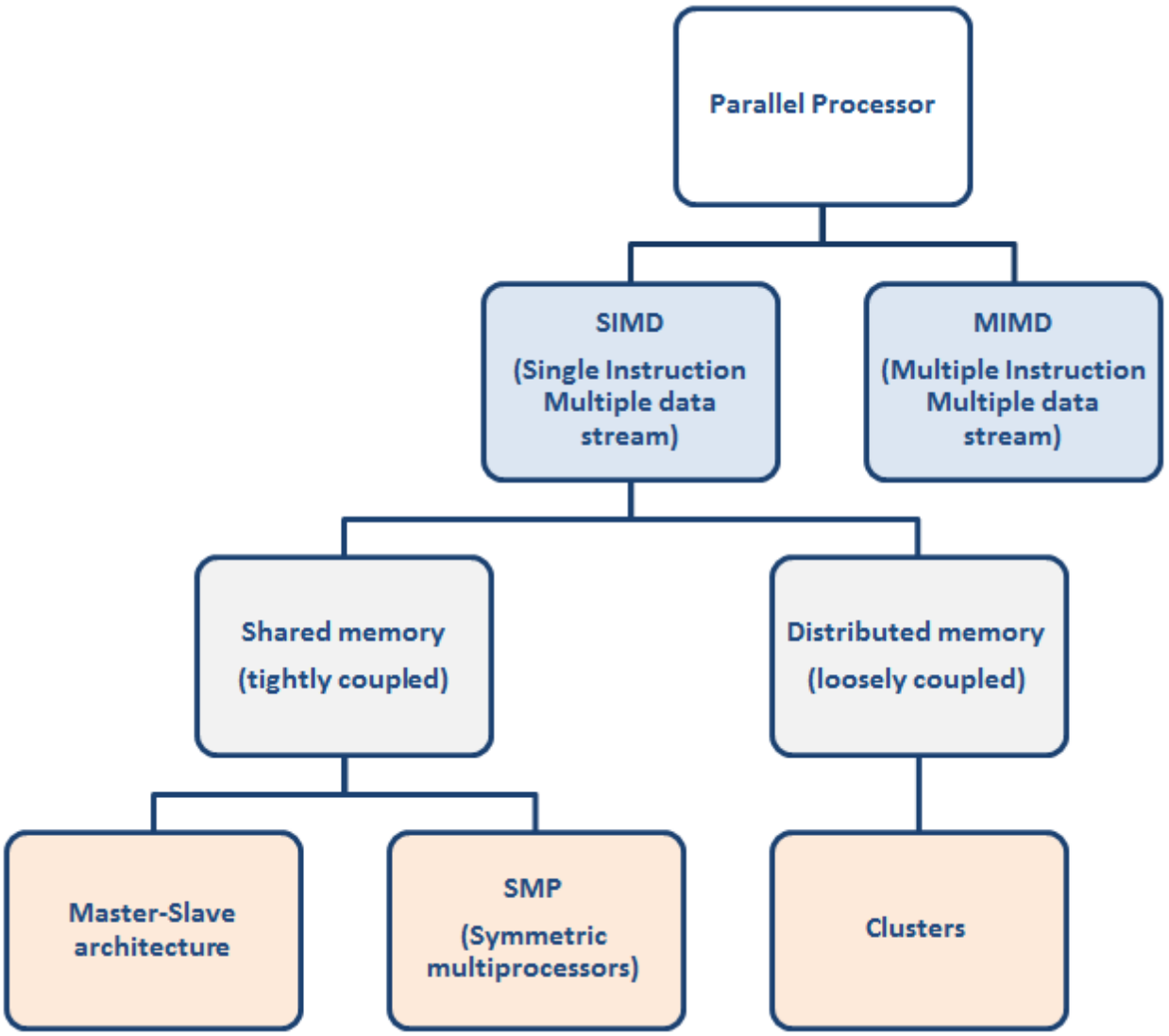

TESTING ERROR
GAP
ERROR
TRAINING ERROR
Training set size

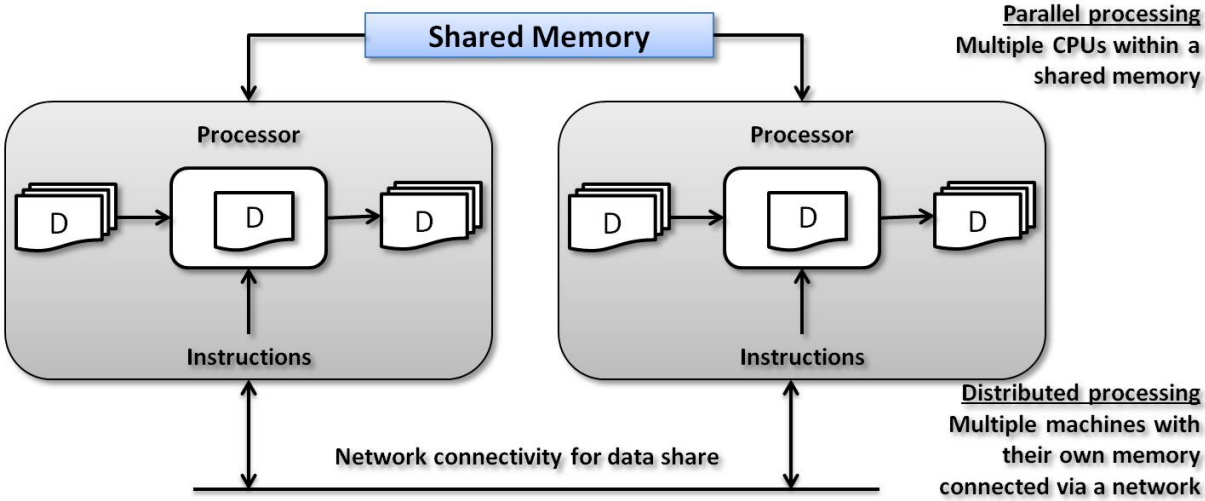

Business Problem
Data Analysis
Datasets
Data Preparation
Modeling
Evaluation
Feedback / Refinement
Deployment


Machine learning
has many
learning models
categories
Association rules based
Decision tree based
Ensemble method based
Deep learning based
Clustering methods based
Dimensionality reduction based
Regression Analysis based
Bayesian method based
Instance based
Kernel method based

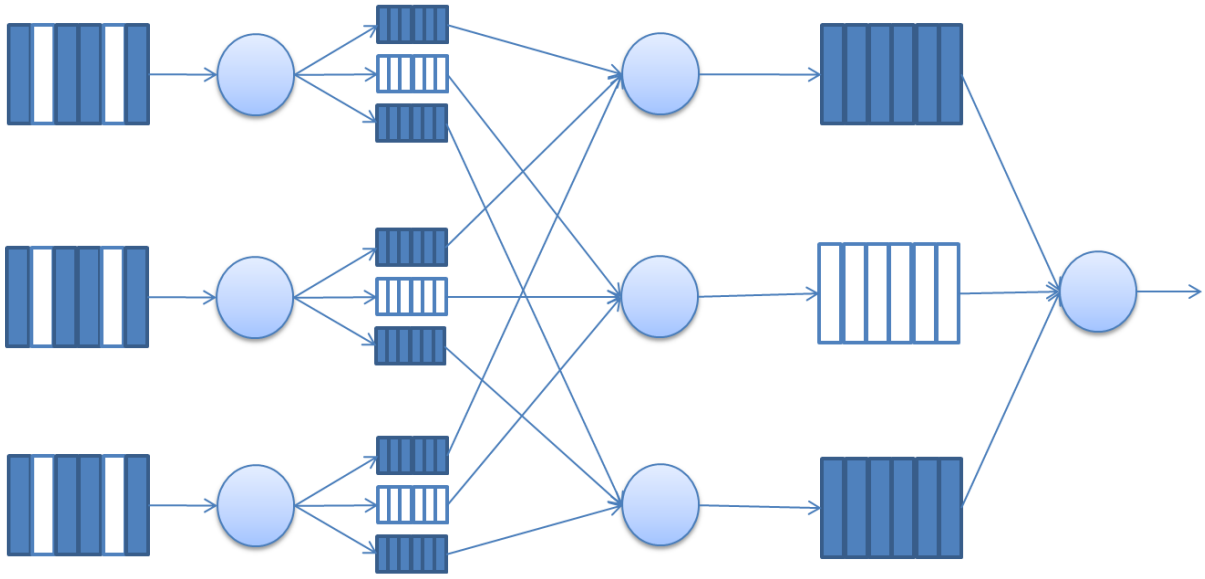

DATA
BIG DATA
Structured → Semi-structured → unstructured
KB → MB → GB → TB → PB → EB →
Classically clean data → Noisy data
Batch → real time
Variety
Volume
Veracity
Velocity
Scale-up
Replace with
Replace with
Scale-out
Virtual storage pool

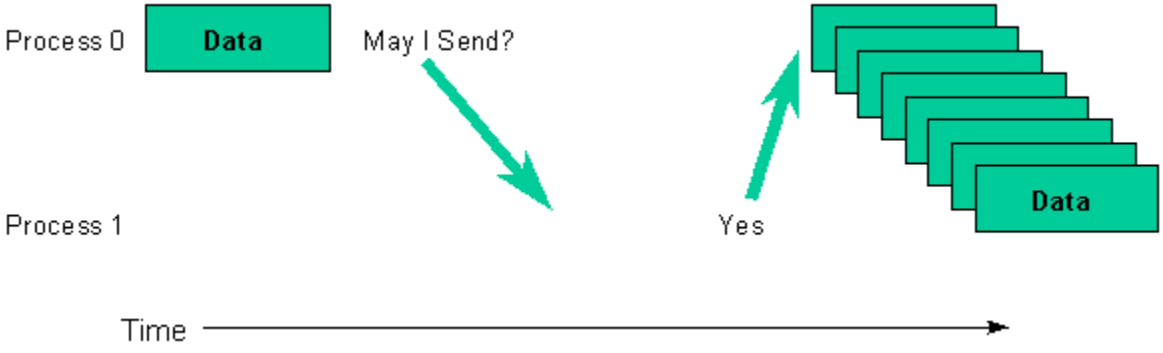

SISD
MISD
INSTRUCTION
PROCESSOR
PROCESSOR
PROCESSOR
DATA
DATA
PROCESSOR
PROCESSOR
PROCESSOR
INSTRUCTION
SIMD
MIMD

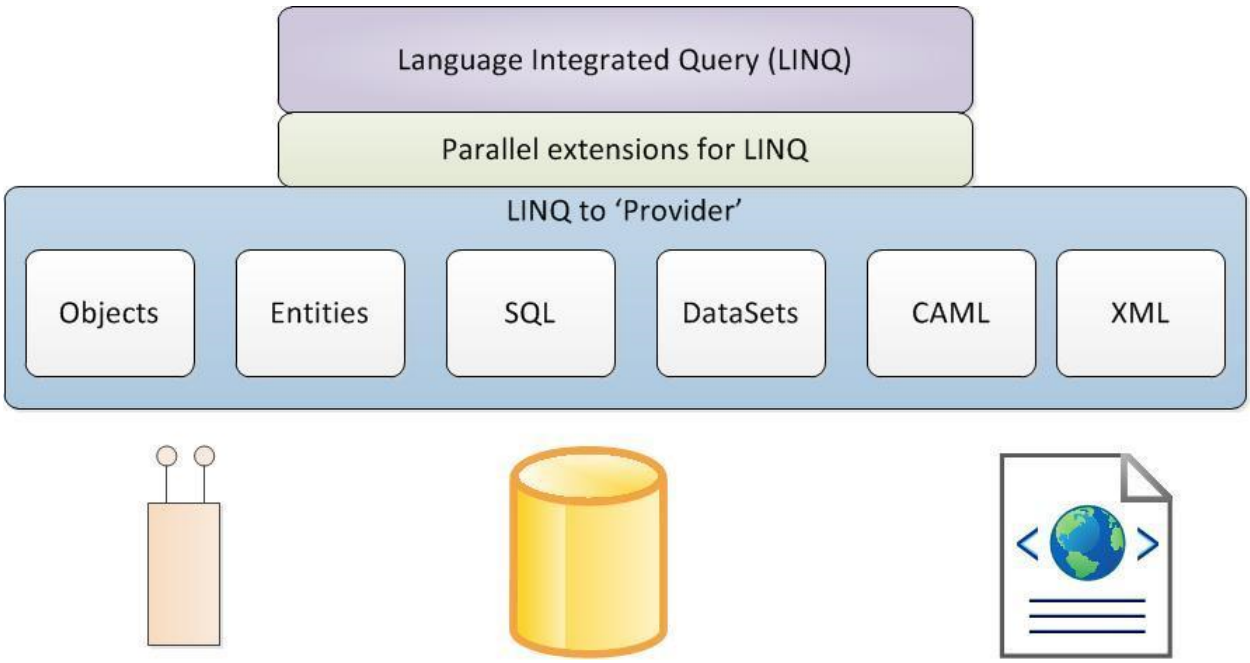

Parallel Processor
SIMD
(Single Instruction Multiple data stream)
MIMD
(Multiple Instruction Multiple data stream)
Shared memory
(tightly coupled)
Distributed memory
(loosely coupled)
Master-Slave architecture
SMP
(Symmetric multiprocessors)
Clusters

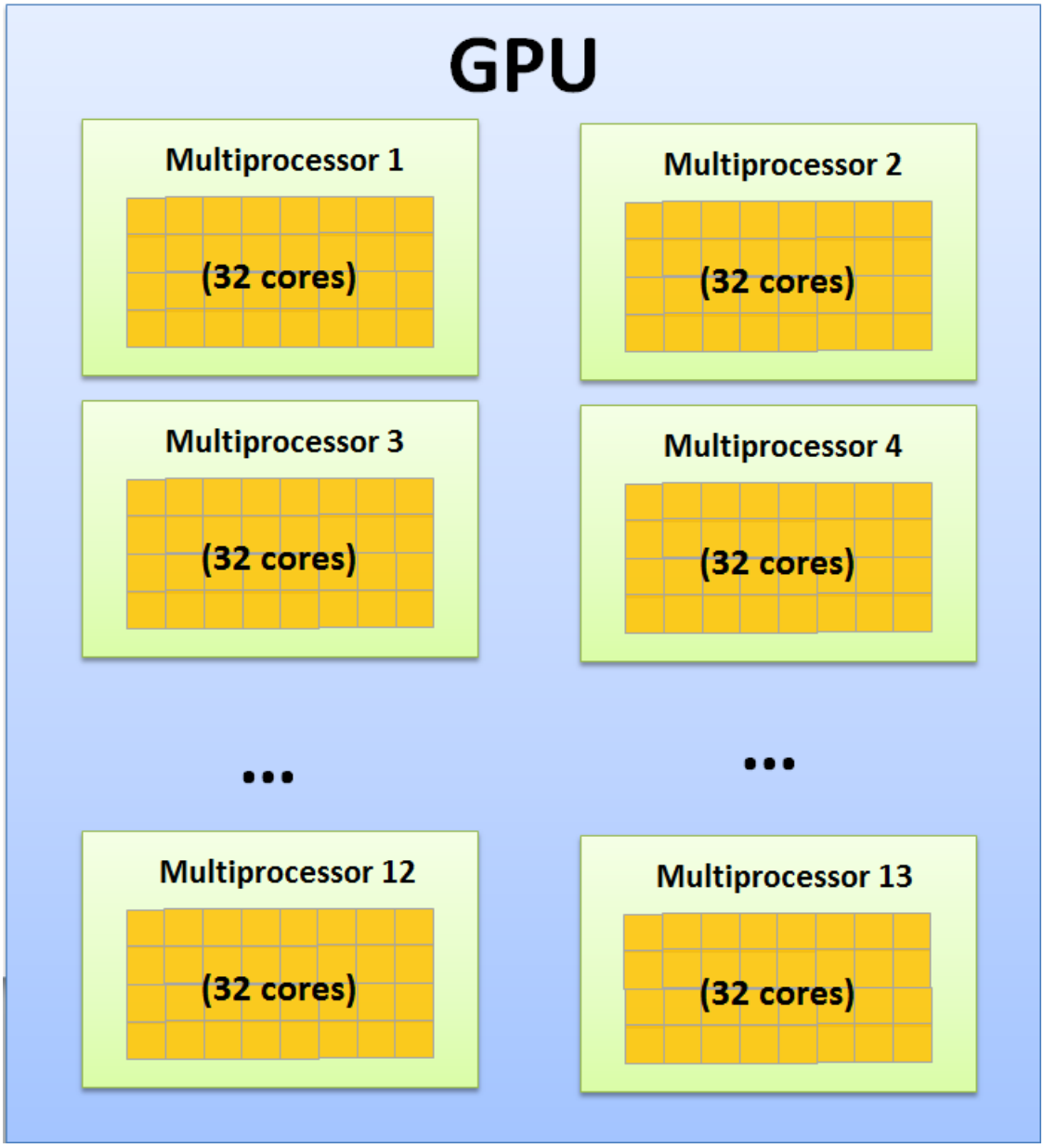

Shared Memory
Processor
D
D
D
Instructions
Processor
D
D
D
Instructions
Network connectivity for data share
Parallel processing
Multiple CPUs within a shared memory
Distributed processing
Multiple machines with their own memory connected via a network

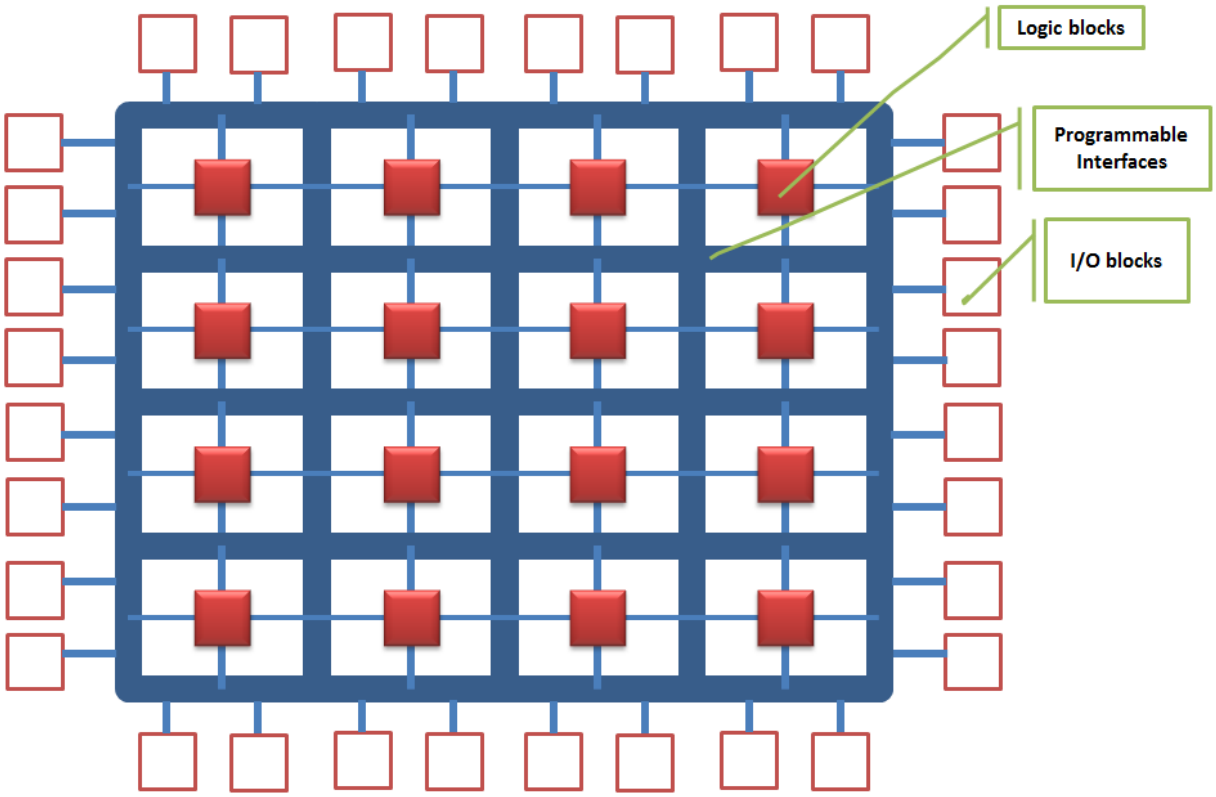
| ID# | Model | Year | Color | Dealer | Price |
|---|---|---|---|---|---|
| 4523 | Civic | 2002 | Blue | MN | $18,000 |
| 3476 | Corolla | 1999 | White | IL | $15,000 |
| 7623 | Camry | 2001 | Green | NY | $21,000 |
| 9834 | Prius | 2001 | Green | CA | $18,000 |
| 6734 | Civic | 2001 | White | OR | $17,000 |
| 5342 | Altima | 2001 | Green | FL | $19,000 |
| 3845 | Maxima | 2001 | Blue | NY | $22,000 |
| 8354 | Accord | 2000 | Green | VT | $18,000 |
| 4395 | Civic | 2001 | Red | CA | $17,000 |
| 7352 | Civic | 2002 | Red | WA | $18,000 |

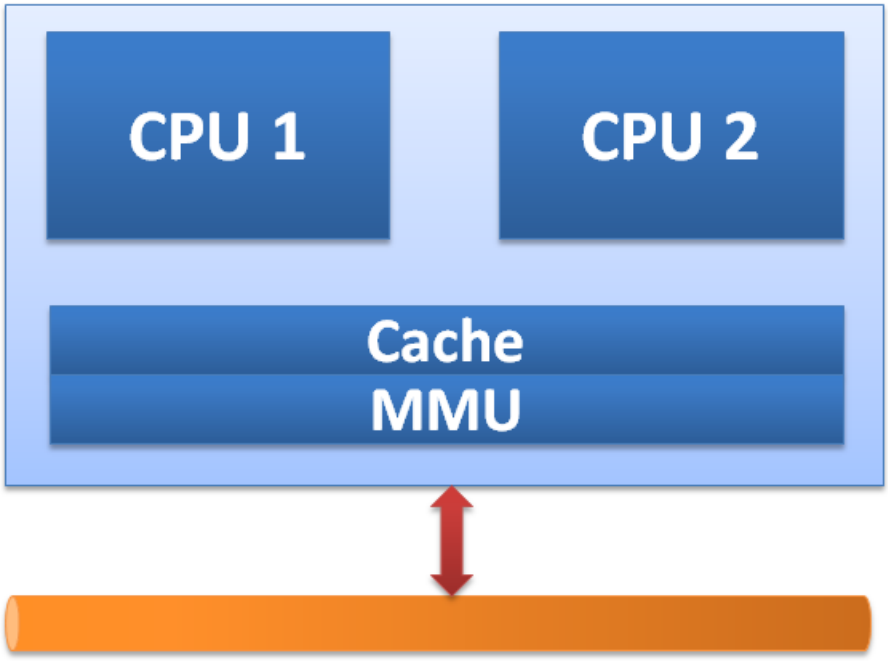


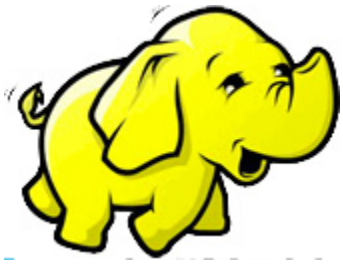




Language Integrated Query (LINQ)
Parallel extensions for LINQ
LINQ to 'Provider'
Objects
Entities
SQL
DataSets
CAML
XML

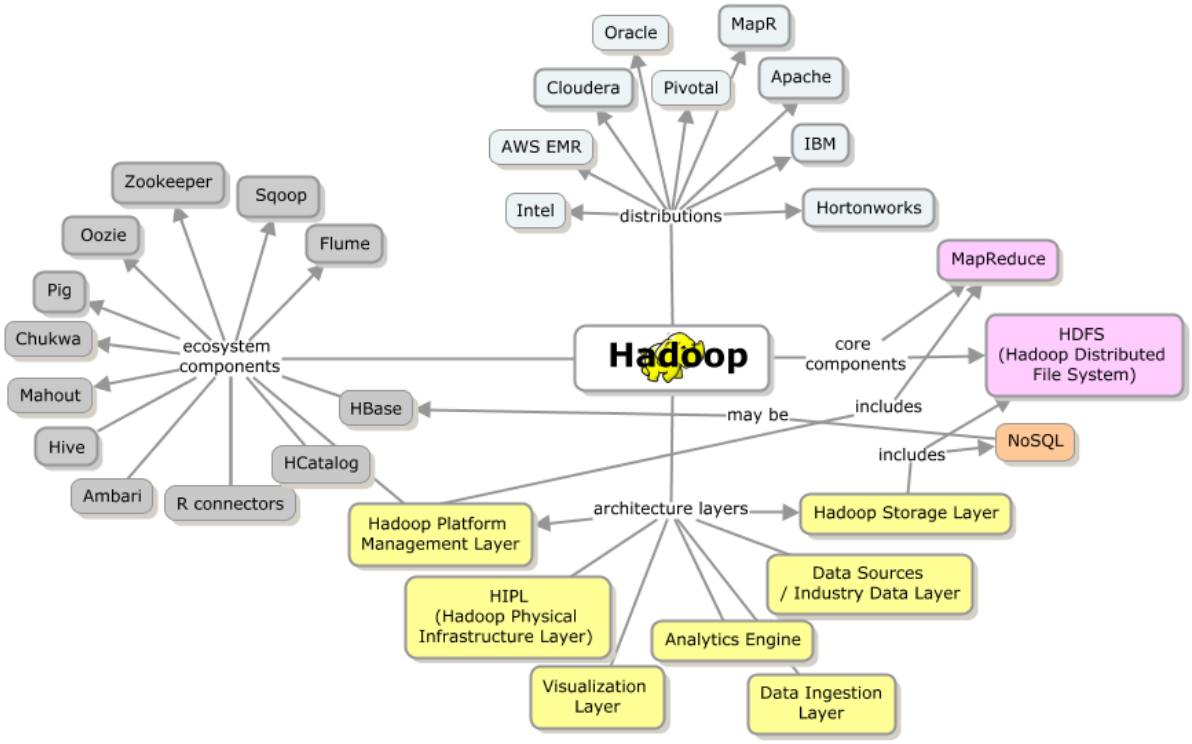

GPU
Multiprocessor 1
(32 cores)
Multiprocessor 2
(32 cores)
Multiprocessor 3
(32 cores)
Multiprocessor 4
(32 cores)
Multiprocessor 12
(32 cores)
Multiprocessor 13
(32 cores)

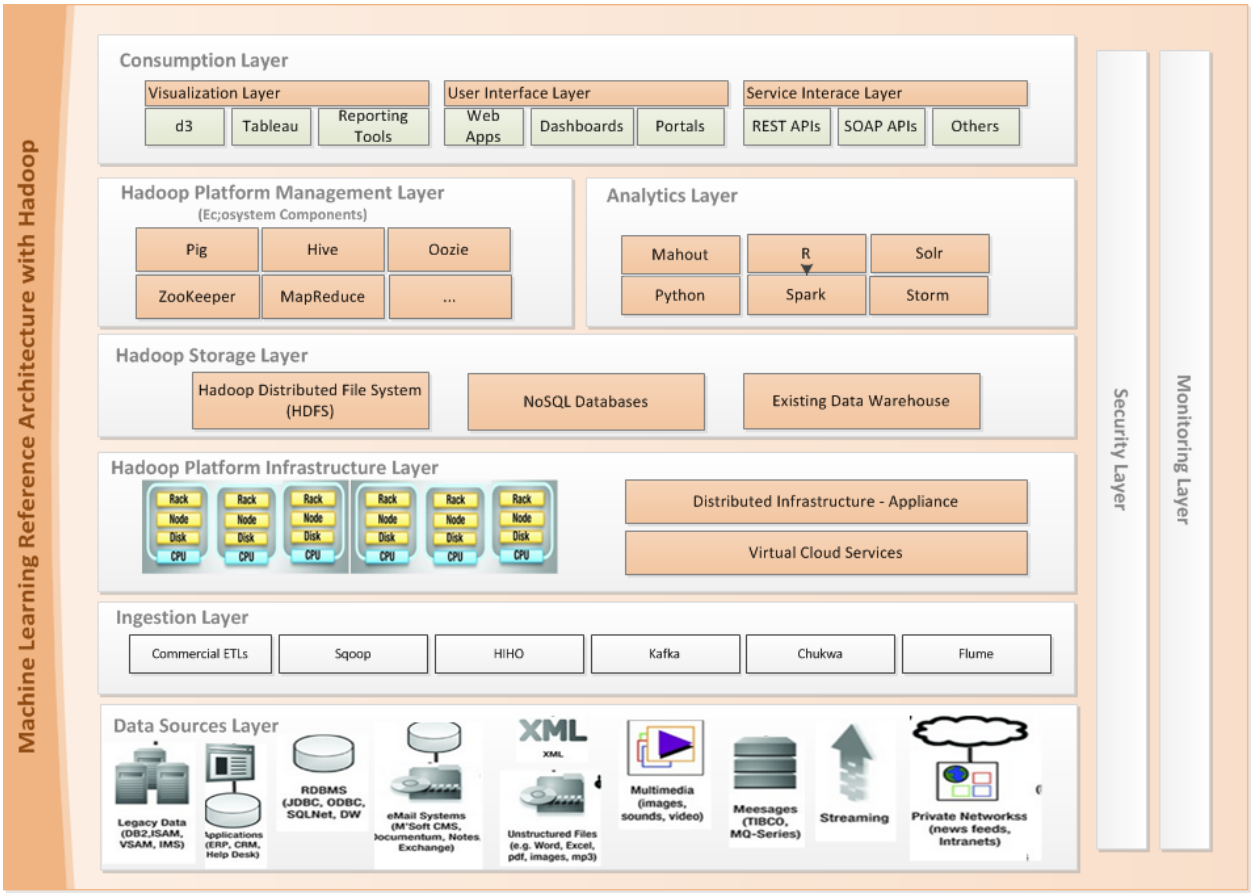

Logic blocks
Programmable Interfaces
I/O blocks

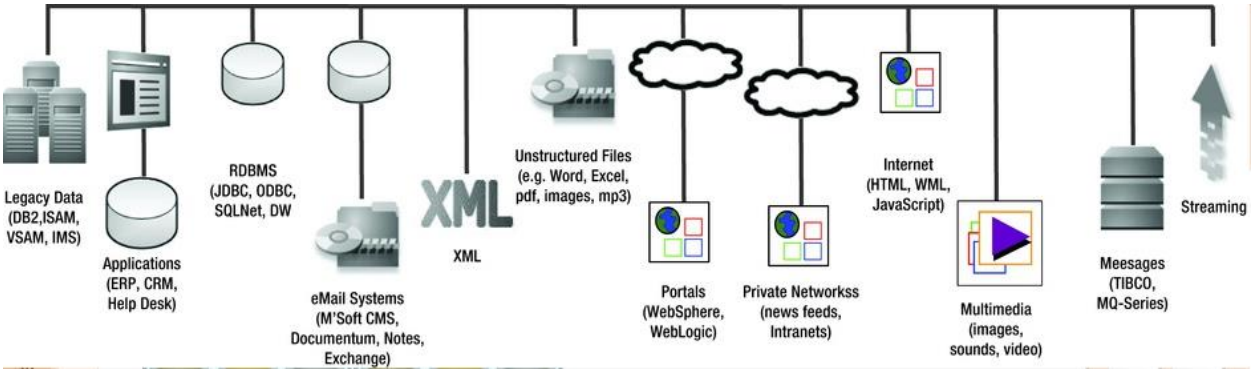

CPU 1
CPU 2
Cache
MMU

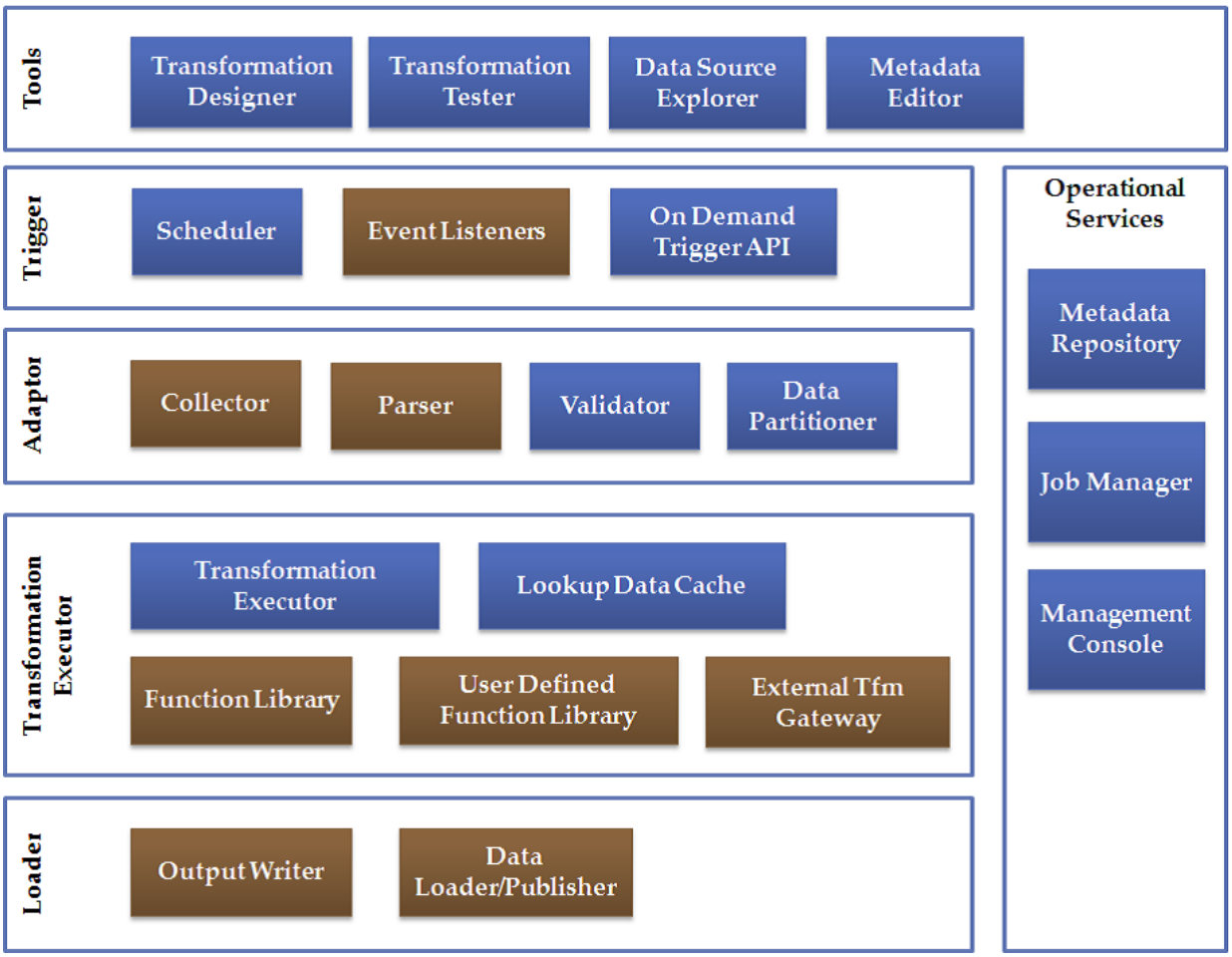


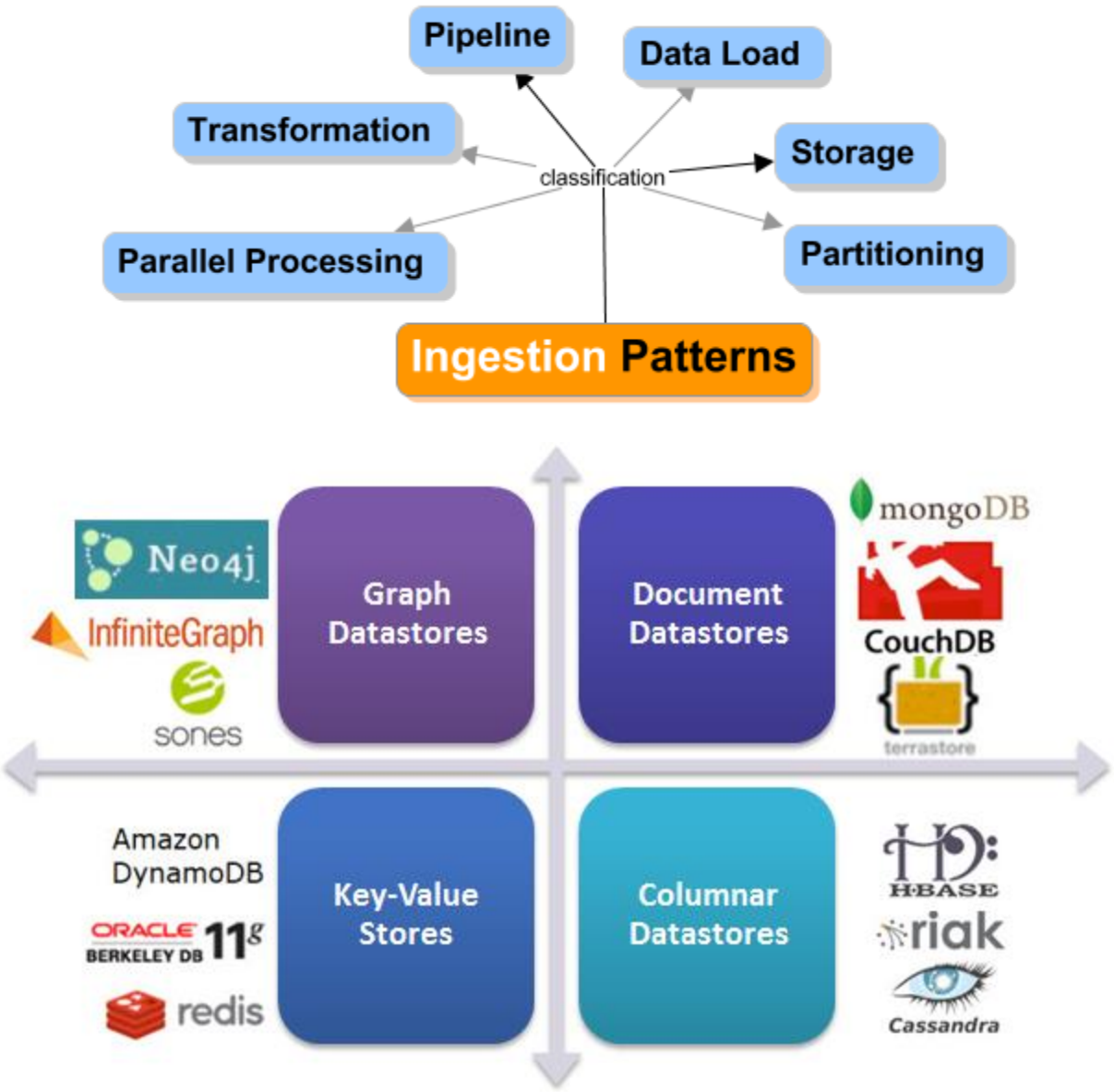
| Year | Evolution / Progress |
|---|---|
| 2002-2003 | Work on Nutch was started by Doug Cutting and Mike Cafarella |
| 2003 - 2004 | Google published work on GFS and MapReduce |
| 2004 | Doug Cutting added suport for GFS and MapReduce to Nutc |
| 2006 | Hadoop spins out of Nutch when Yahoo hired Doug Cutting |
| 2007 | NY Times converts 4TB of image archives over 100 EC2s |
| | Facebook launched Hive, an SQL support for Hadoop |
| 2008 | Fastest sort over 910 nodes taking 3.5 mins |
| | Cloudera founded |
| 2009 | First Hadoop Summit with 750 attendeed |
| | Doug Cutting joined Cloudera |


Oracle
MapR
Cloudera
Pivotal
Apache
AWS EMR
IBM
Intel
distributions
Hortonworks
Zookeeper
Sqoop
Oozie
Flume
Pig
Chukwa
ecosystem components
Mahout
Hive
HCatalog
Ambari
R connectors
Hadoop
core components
MapReduce
HDFS (Hadoop Distributed File System)
HBase
may be
includes
NoSQL
includes
Hadoop Platform Management Layer
architecture layers
Hadoop Storage Layer
HIPL (Hadoop Physical Infrastructure Layer)
Data Sources / Industry Data Layer
Analytics Engine
Visualization Layer
Data Ingestion Layer

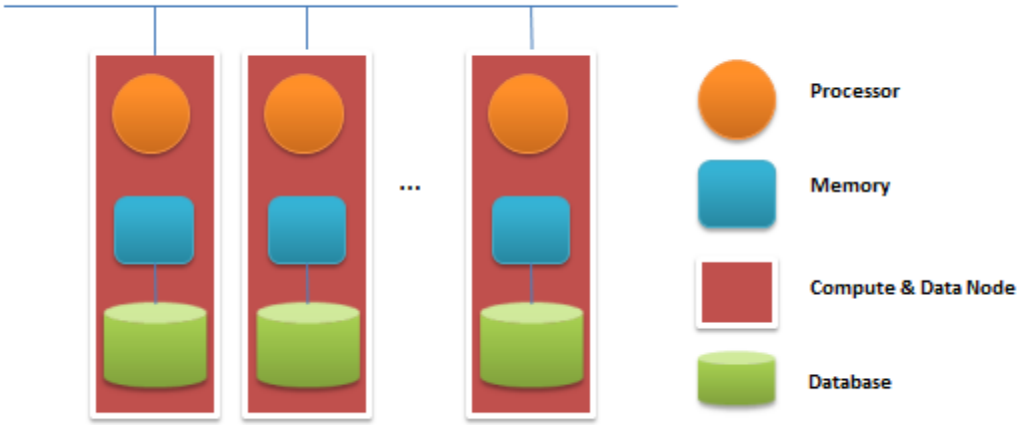

Machine Learning Reference Architecture with Hadoop
Consumption Layer
Visualization Layer
d3
Tableau
Reporting Tools
User Interface Layer
Web Apps
Dashboards
Portals
Service Interace Layer
REST APIs
SOAP APIs
Others
Hadoop Platform Management Layer
(Ec;osystem Components)
Pig
Hive
Oozie
ZooKeeper
MapReduce
...
Analytics Layer
Mahout
R
Solr
Python
Spark
Storm
Hadoop Storage Layer
Hadoop Distributed File System (HDFS)
NoSQL Databases
Existing Data Warehouse
Hadoop Platform Infrastructure Layer
Rack
Node
Disk
CPU
Distributed Infrastructure - Appliance
Virtual Cloud Services
Ingestion Layer
Commercial ETLs
Sqoop
HIHO
Kafka
Chukwa
Flume
Data Sources Layer
Legacy Data (DB2,ISAM, VSAM, IMS)
RDBMS (JDBC, ODBC, SQLNet, DW
eMail Systems (M'Soft CMS, Documentum, Notes Exchange)
XML
Unstructured Files (e.g. Word, Excel, pdf, images, mp3)
Multimedia (images, sounds, video)
Meesages (TIBCO, MQ-Series)
Streaming
Private Networkss (news feeds, Intranets)
Security Layer
Monitoring Layer

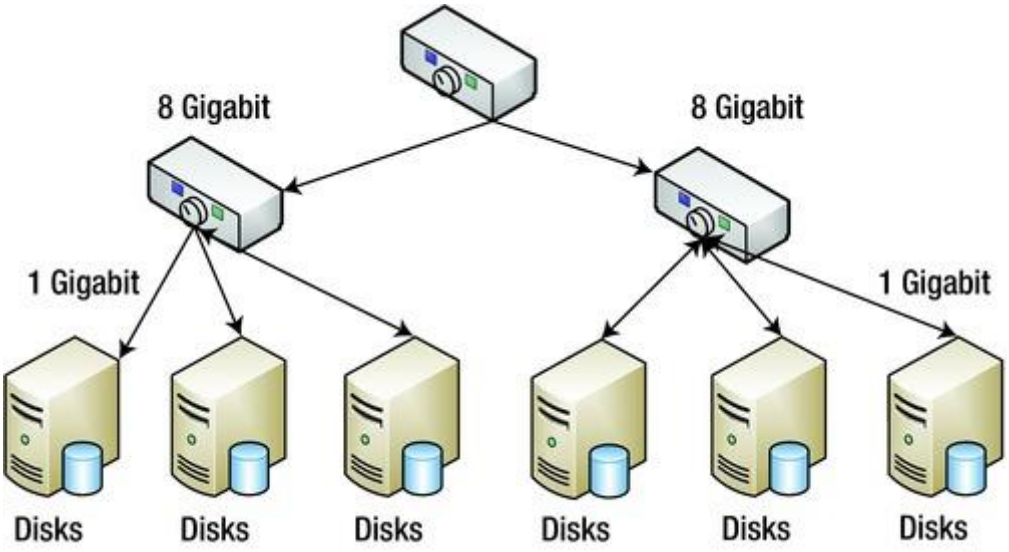

Legacy Data (DB2,ISAM, VSAM, IMS)
Applications (ERP, CRM, Help Desk)
RDBMS (JDBC, ODBC, SQLNet, DW
eMail Systems (M'Soft CMS, Documentum, Notes, Exchange)
XML
XML
Unstructured Files (e.g. Word, Excel, pdf, images, mp3)
Portals (WebSphere, WebLogic)
Private Networkss (news feeds, Intranets)
Internet (HTML, WML, JavaScript)
Multimedia (images, sounds, video)
Meesages (TIBCO, MQ-Series)
Streaming

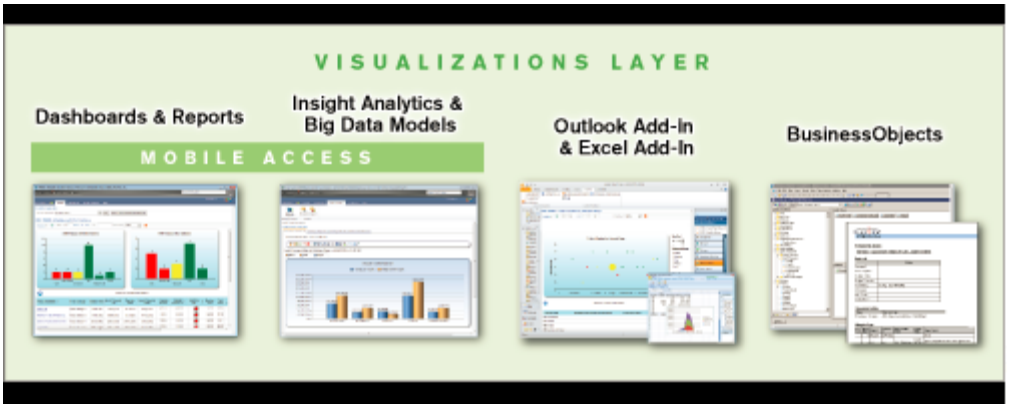

Tools
Transformation Designer
Transformation Tester
Data Source Explorer
Metadata Editor
Trigger
Scheduler
Event Listeners
On Demand Trigger API
Adaptor
Collector
Parser
Validator
Data Partitioner
Transformation Executor
Transformation Executor
Lookup Data Cache
Function Library
User Defined Function Library
External Tfm Gateway
Loader
Output Writer
Data Loader/Publisher
Operational Services
Metadata Repository
Job Manager
Management Console

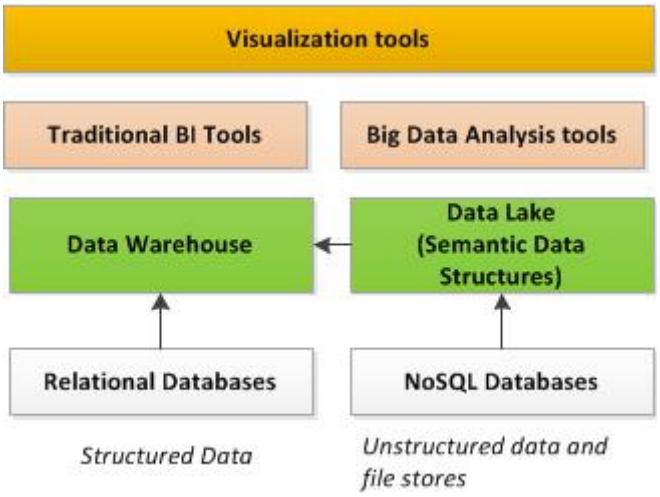

Pipeline
Data Load
Transformation
Storage
classification
Parallel Processing
Partitioning
Ingestion Patterns
Neo4j
InfiniteGraph
sones
Graph Datastores
Document Datastores
mongoDB
CouchDB
terrastore
Amazon DynamoDB
ORACLE BERKELEY DB 11g
redis
Key-Value Stores
Columnar Datastores
HBASE
riak
Cassandra

## Shared Nothing Data Architecture

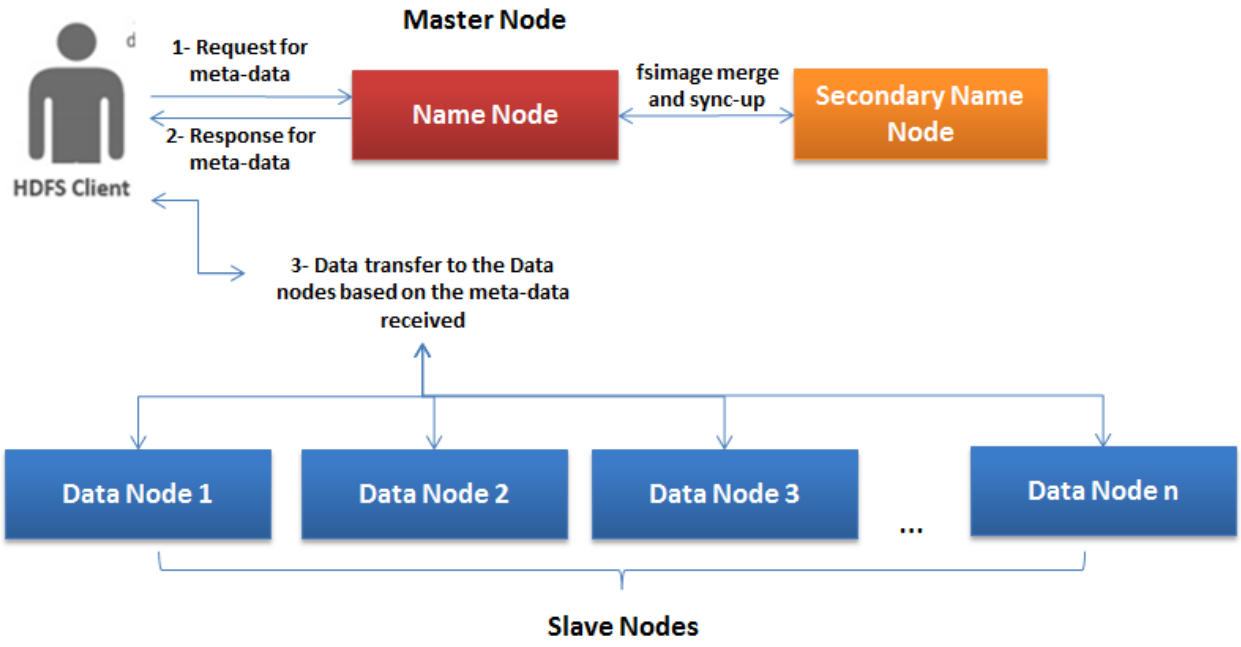


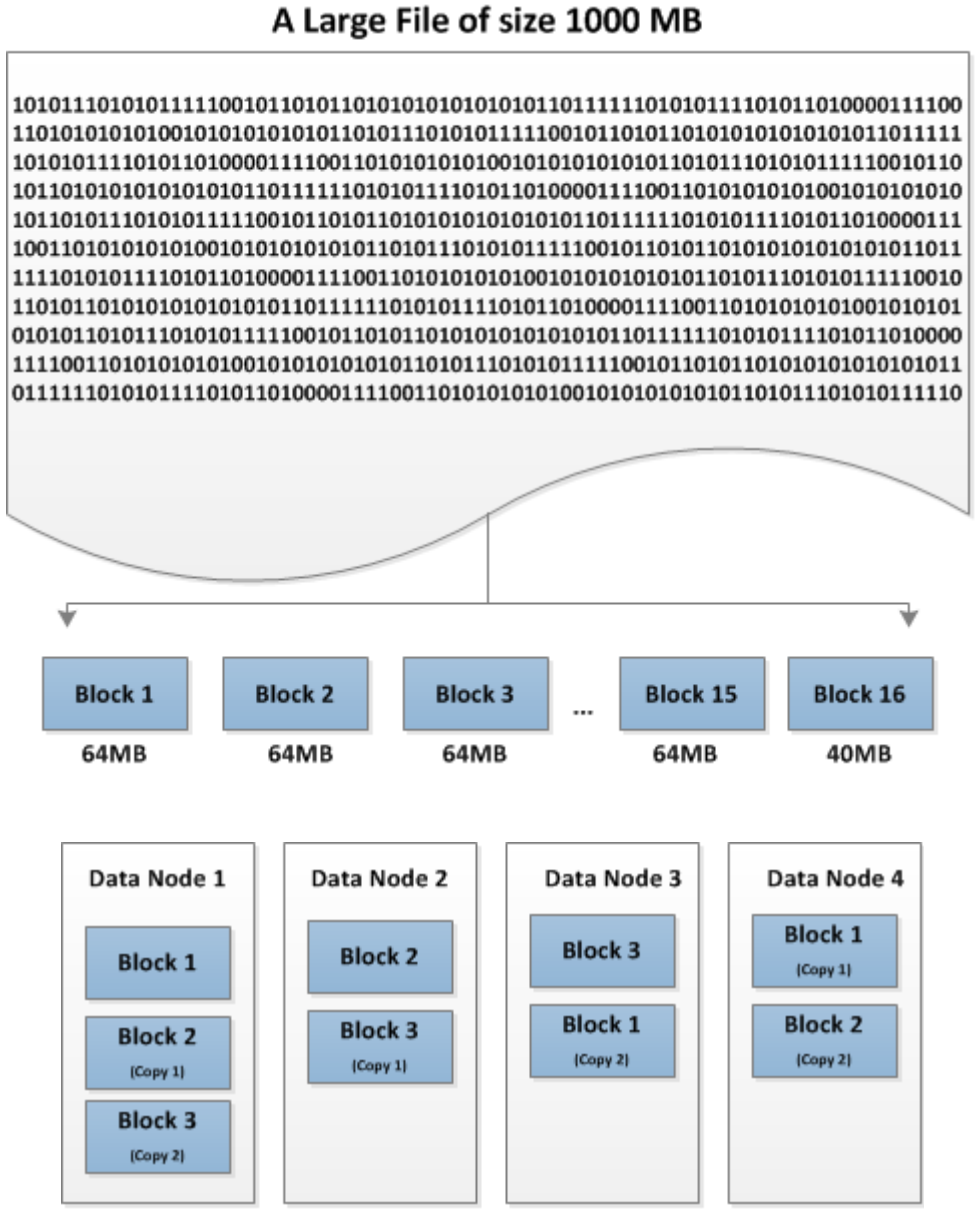



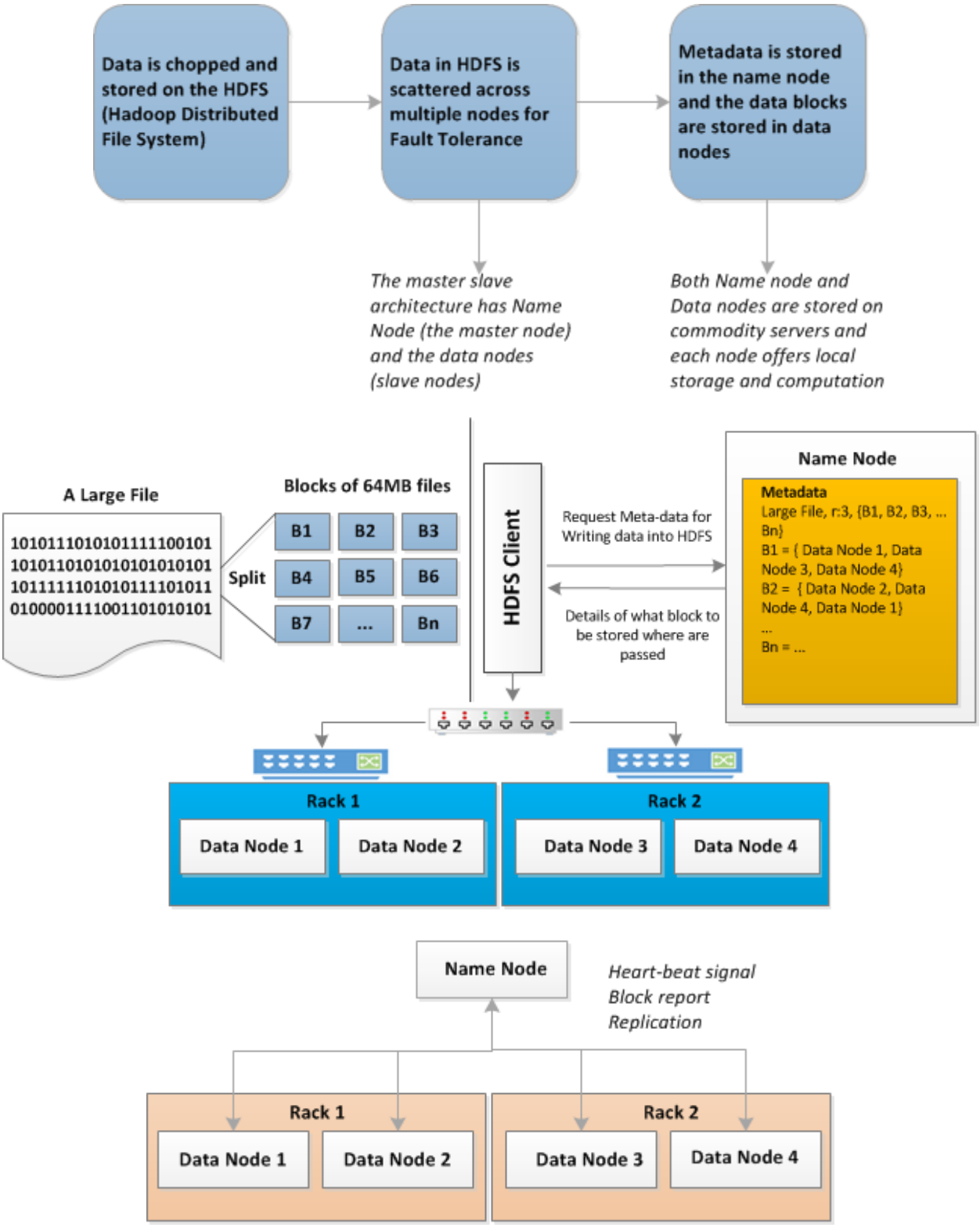

Visualization tools
Traditional BI Tools
Big Data Analysis tools
Data Warehouse
Data Lake
(Semantic Data
Structures)
Relational Databases
NoSQL Databases
Structured Data
Unstructured data and
file stores


hadoop
HDFS


Master Node
1- Request for meta-data
2- Response for meta-data
Name Node
fsimage merge and sync-up
Secondary Name Node
HDFS Client
3- Data transfer to the Data nodes based on the meta-data received
Data Node 1
Data Node 2
Data Node 3
...
Data Node n
Slave Nodes

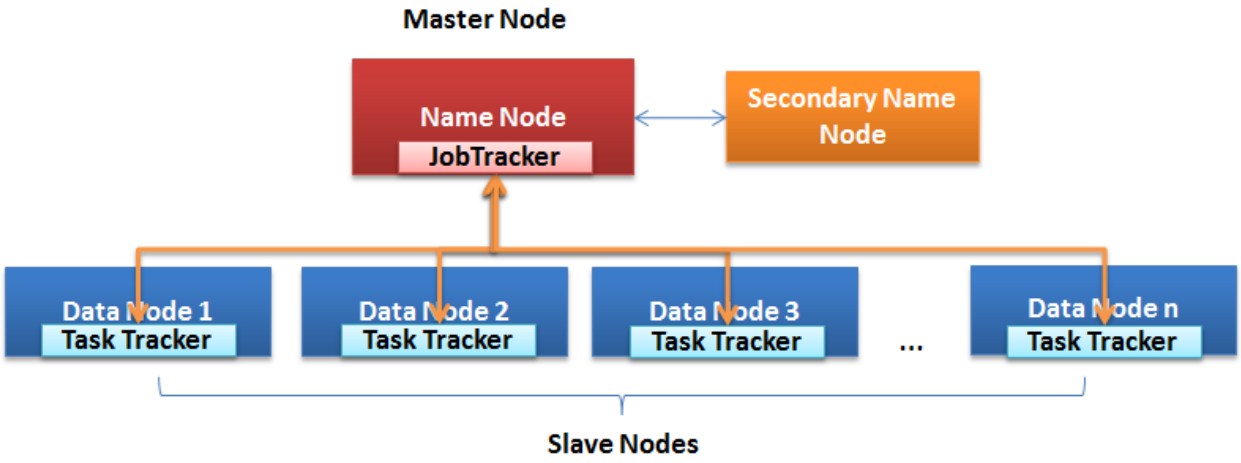

fsimage
log
Checkpoint
New fsimage

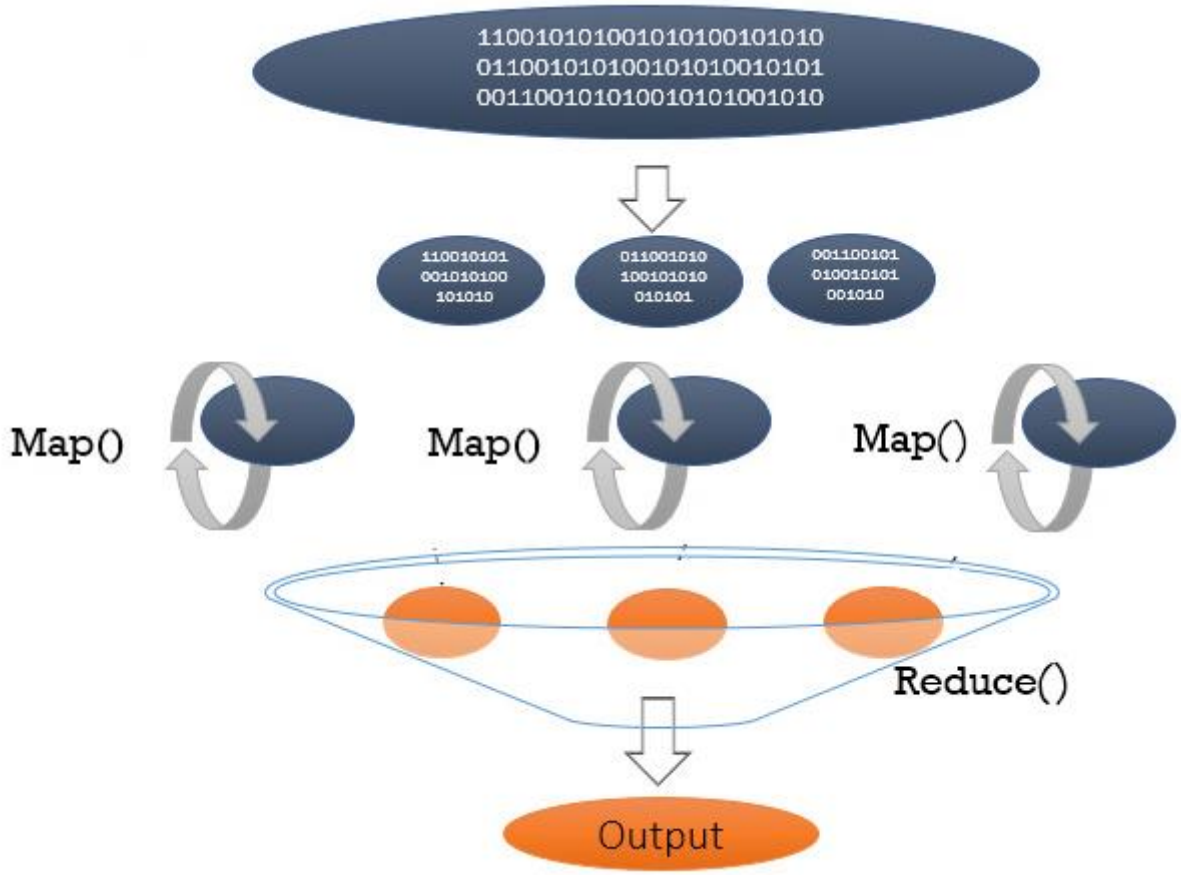

A Large File of size 1000 MB
Block 1
Block 2
Block 3
...
Block 15
Block 16
64MB
64MB
64MB
64MB
40MB
Data Node 1
Block 1
Block 2
(Copy 1)
Block 3
(Copy 2)
Data Node 2
Block 2
Block 3
(Copy 1)
Data Node 3
Block 3
Block 1
(Copy 2)
Data Node 4
Block 1
(Copy 1)
Block 2
(Copy 2)

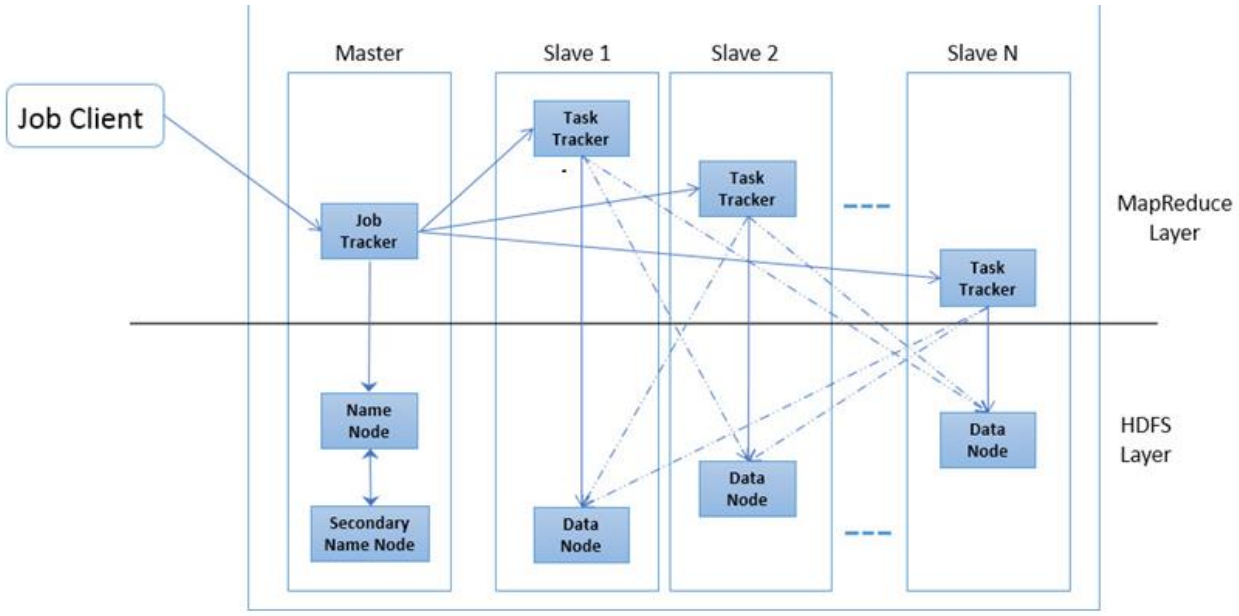

Data is chopped and stored on the HDFS (Hadoop Distributed File System)
Data in HDFS is scattered across multiple nodes for Fault Tolerance
Metadata is stored in the name node and the data blocks are stored in data nodes
The master slave architecture has Name Node (the master node) and the data nodes (slave nodes)
Both Name node and Data nodes are stored on commodity servers and each node offers local storage and computation
A Large File
10101110101011111001011010110101010101010110111111010101110110101
01000011110011010101
Blocks of 64MB files
Split
B1
B2
B3
B4
B5
B6
B7
...
Bn
HDFS Client
Request Meta-data for Writing data into HDFS
Details of what block to be stored where are passed
Name Node
Metadata
Large File, r:3, {B1, B2, B3, ... Bn}
B1 = { Data Node 1, Data Node 3, Data Node 4}
B2 = { Data Node 2, Data Node 4, Data Node 1}
...
Bn = ...
Rack 1
Data Node 1
Data Node 2
Rack 2
Data Node 3
Data Node 4
Name Node
Heart-beat signal
Block report
Replication
Rack 1
Data Node 1
Data Node 2
Rack 2
Data Node 3
Data Node 4

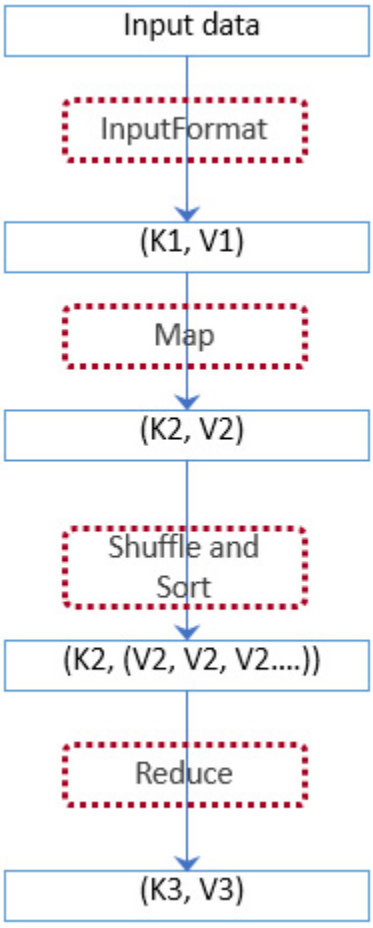
Hadoop Command Line

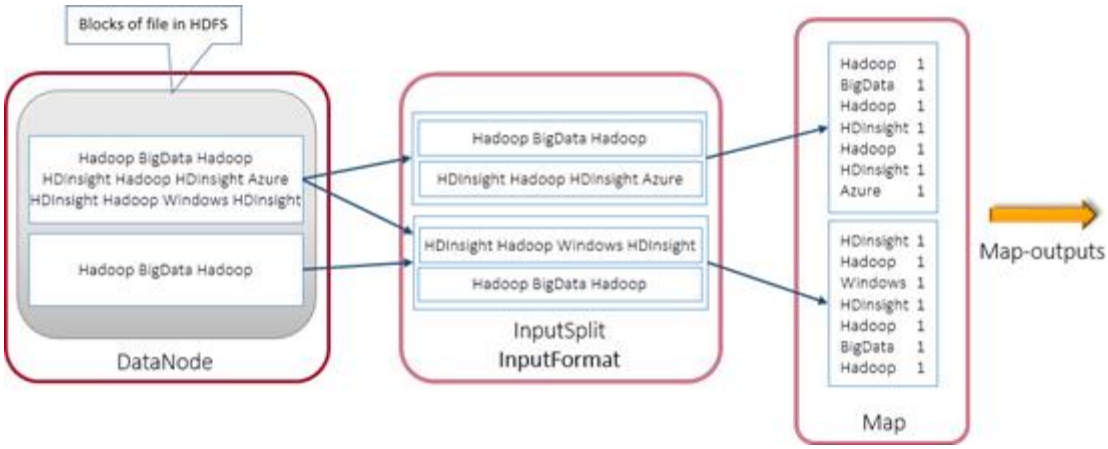
```
C:\apps\dist\hadoop-2.4.0.2.1.3.0-1948>hadoop fs
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] <localsrc> ... <dst>]
        [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] <path> ...]
        [-cp [-f] [-p] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] <path> ...]
        [-expunge]
        [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getmerge [-nl] <src> <localdst>]
        [-help [cmd ...]]
        [-ls [-d] [-h] [-R] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <pa
th>]]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touchz <path> ...]
        [-usage [cmd ...]]

Generic options supported are
-conf <configuration file>     specify an application configuration file
-D <property=value>            use value for given property
-fs <local|namenode:port>      specify a namenode
-jt <local|jobtracker:port>    specify a job tracker
-files <comma separated list of files>    specify comma separated files to be co
pied to the map reduce cluster
-libjars <comma separated list of jars>    specify comma separated jar files to
include in the classpath.
-archives <comma separated list of archives>    specify comma separated archives
 to be unarchived on the compute machines.

The general command line syntax is
bin/hadoop command [genericOptions] [commandOptions]


C:\apps\dist\hadoop-2.4.0.2.1.3.0-1948>_
```


Master Node
Name Node
JobTracker
Secondary Name Node
Data Node 1
Task Tracker
Data Node 2
Task Tracker
Data Node 3
Task Tracker
...
Data Node n
Task Tracker
Slave Nodes


11001010100101010010101
01100101010010101001010
00110010101001010100101
0
110010101
001010100
101010
011001010
100101010
010101
001100101
010010101
001010
Map()
Map()
Map()
Reduce()
Output


Master
Slave 1
Slave 2
Slave N
Job Client
Task Tracker
Task Tracker
Job Tracker
Task Tracker
MapReduce Layer
Name Node
Data Node
HDFS Layer
Secondary Name Node
Data Node
Data Node


Input data
InputFormat
(K1, V1)
Map
(K2, V2)
Shuffle and Sort
(K2, (V2, V2, V2....))
Reduce
(K3, V3)




# ARCHITECTURE COMPARISON

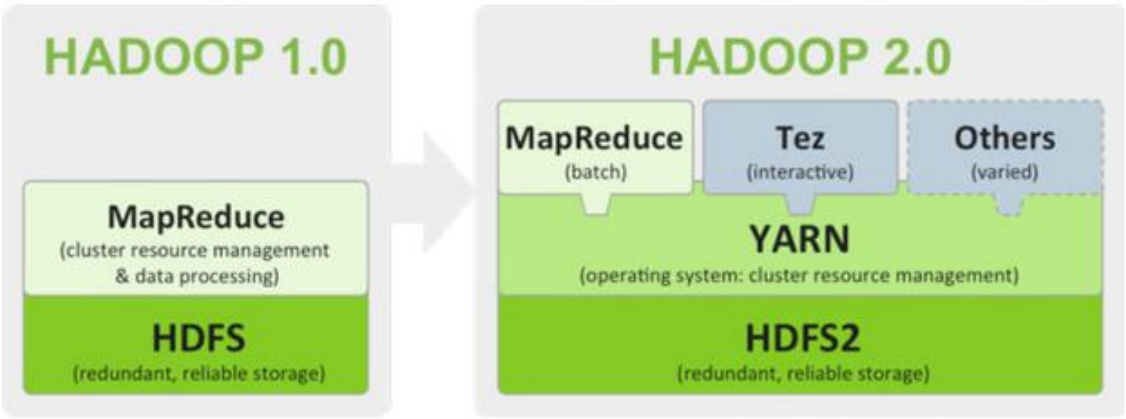
## Hadoop 1.0 vs. Hadoop 2.0.

***Single Use System***

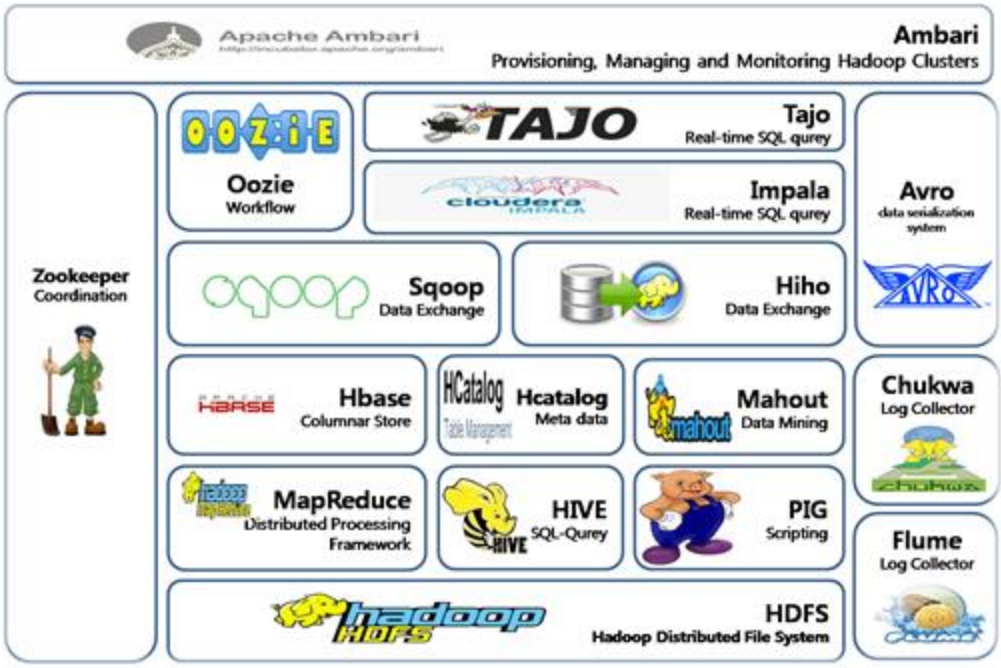
*Batch Apps*

***Multi Use Data Platform***

*Batch, Interactive, Online, Streaming, …*




SOURCE: HORTONWORKS




*Source: Internet*

```
HTTP request sent, awaiting response... 200 OK
Length: 96316511 (92M) [application/x-gzip]
Saving to: `jdk-7u25-linux-x64.tar.gz'

100%[======================================>] 96,316,511   311K/s   in 5m 3s

2013-10-24 14:34:22 (311 KB/s) - `jdk-7u25-linux-x64.tar.gz' saved [9631
6511/96316511]
```

```
master@Hadoopupgrade:~$ java -version
java version "1.7.0_45"
Java(TM) SE Runtime Environment (build 1.7.0_45-b18)
Java HotSpot(TM) 64-Bit Server VM (build 24.45-b08, mixed mode)
master@Hadoopupgrade:~$
```

```
master@Hadoopupgrade:~$ sudo addgroup hadoop
Adding group `hadoop' (GID 1001) ...
Done.
master@Hadoopupgrade:~$ sudo adduser --ingroup hadoop hduser
Adding user `hduser' ...
Adding new user `hduser' (1001) with group `hadoop' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hduser
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] Y
master@Hadoopupgrade:~$
```

```
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
     <name>yarn.nodemanager.aux-services</name>
     <value>mapreduce_shuffle</value>
</property>
 <property>
     <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
     <value>org.apache.hadoop.mapred.ShuffleHandler</value>
 </property>

</configuration>
```

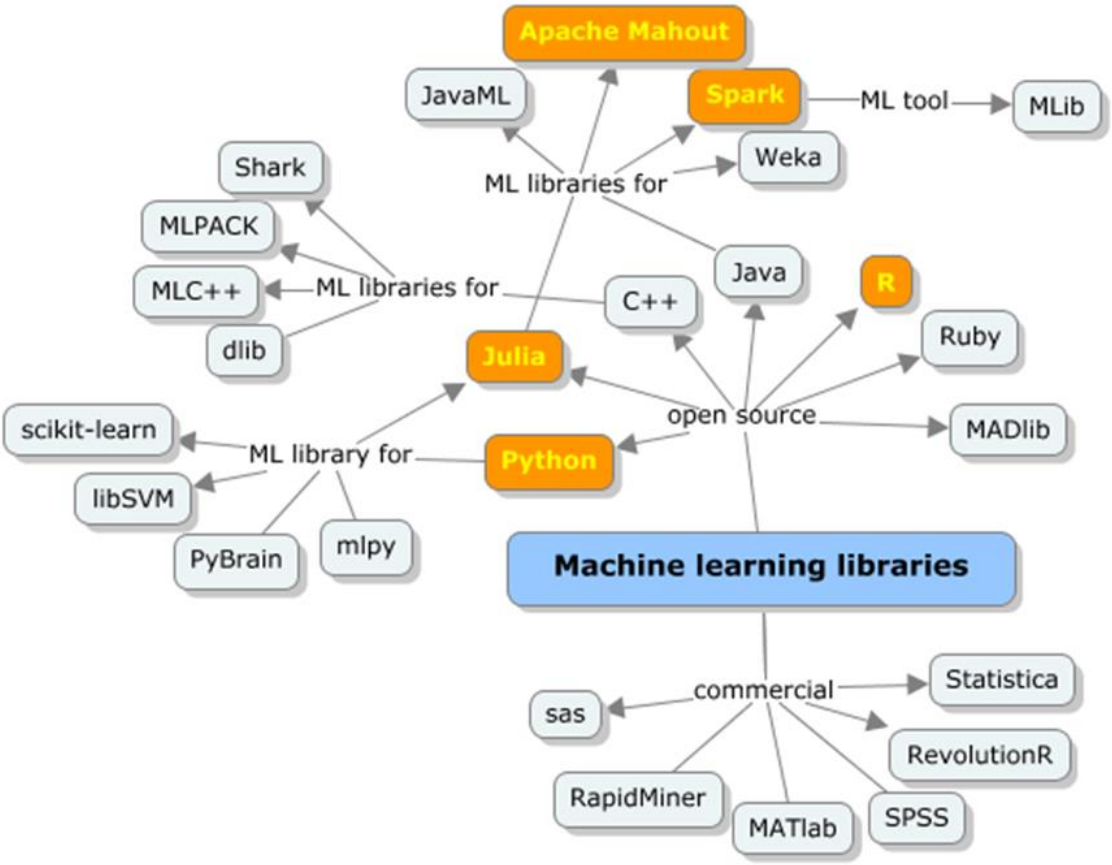
```
<configuration>
<property>
     <name>fs.default.name</name>
     <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

```
<configuration>
<property>
     <name>dfs.replication</name>
     <value>1</value>
  </property>
  <property>
     <name>dfs.namenode.name.dir</name>
     <value>file:/usr/local/hadoop/yarn_data/hdfs/namenode</value>
  </property>
<property>
     <name>dfs.datanode.data.dir</name>
     <value>file:/usr/local/hadoop/yarn_data/hdfs/datanode</value>
</property>
</configuration>
```

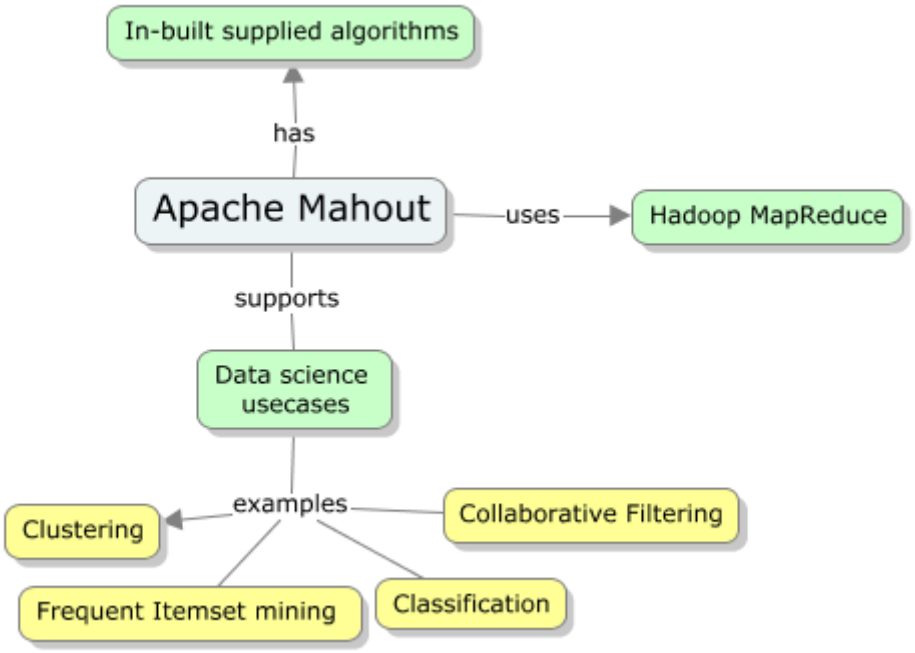
```
<configuration>
<property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
</property>
</configuration>
```

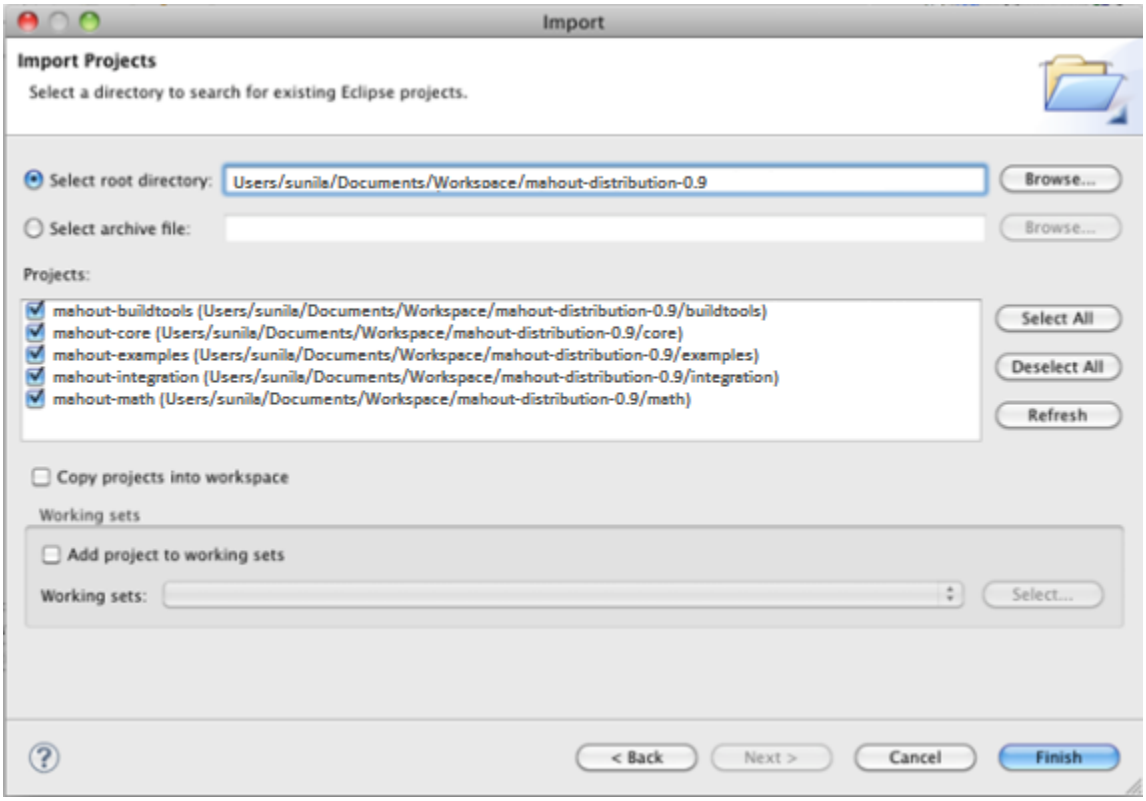
```
# Set Hadoop-related environment variables
export HADOOP_PREFIX='/usr/local/hadoop'
export HADOOP_HOME='/usr/local/hadoop'
export HADOOP_MAPRED_HOME=${HADOOP_HOME}
export HADOOP_COMMON_HOME=${HADOOP_HOME}
export HADOOP_HDFS_HOME=${HADOOP_HOME}
export YARN_HOME=${HADOOP_HOME}
export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop
# Native Path
export HADOOP_COMMON_LIB_NATIVE_DIR=${HADOOP_PREFIX}/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_PREFIX/lib"
#Java path
export JAVA_HOME='/usr/local/Java/jdk1.7.0_45'
# Add Hadoop bin/ directory to PATH
export PATH=$PATH:$HADOOP_HOME/bin:$JAVA_HOME/bin:$HADOOP_HOME/sbin
```

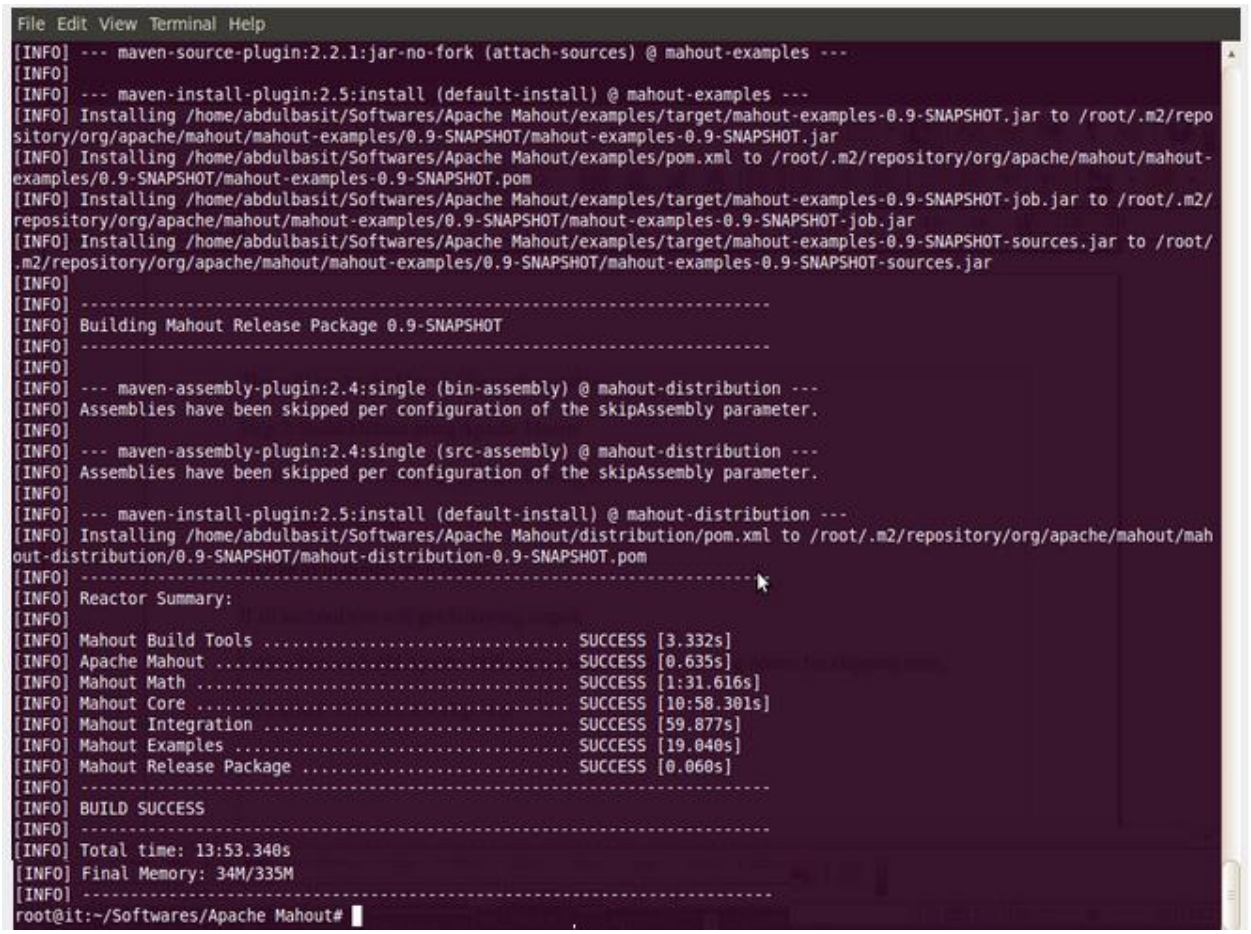
```
hduser@Hadoopupgrade:~$ hadoop-daemon.sh start namenode
starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-Hado
opupgrade.out
hduser@Hadoopupgrade:~$ jps
1244 NameNode
1280 Jps
hduser@Hadoopupgrade:~$ hadoop-daemon.sh start datanode
starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-Hado
opupgrade.out
hduser@Hadoopupgrade:~$ jps
1400 Jps
1244 NameNode
1332 DataNode
hduser@Hadoopupgrade:~$ yarn-daemon.sh start resourcemanager
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resource
manager-Hadoopupgrade.out
hduser@Hadoopupgrade:~$ jps
1474 Jps
1244 NameNode
1433 ResourceManager
1332 DataNode
```


Apache Mahout
JavaML
Spark
ML tool
MLib
Weka
ML libraries for
Shark
MLPACK
MLC++
ML libraries for
dlib
C++
Java
R
Ruby
Julia
open source
MADlib
scikit-learn
ML library for
Python
libSVM
PyBrain
mlpy
Machine learning libraries
commercial
Statistica
sas
RevolutionR
RapidMiner
MATlab
SPSS


mahout

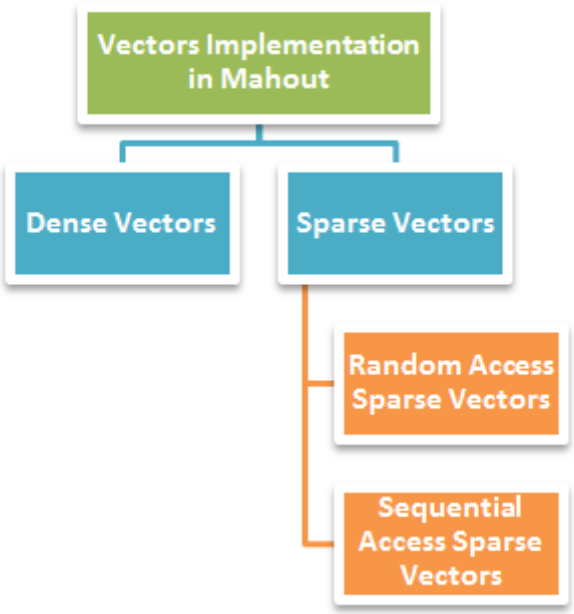

In-built supplied algorithms
has
Apache Mahout
uses
Hadoop MapReduce
supports
Data science usecases
examples
Clustering
Collaborative Filtering
Frequent Itemset mining
Classification

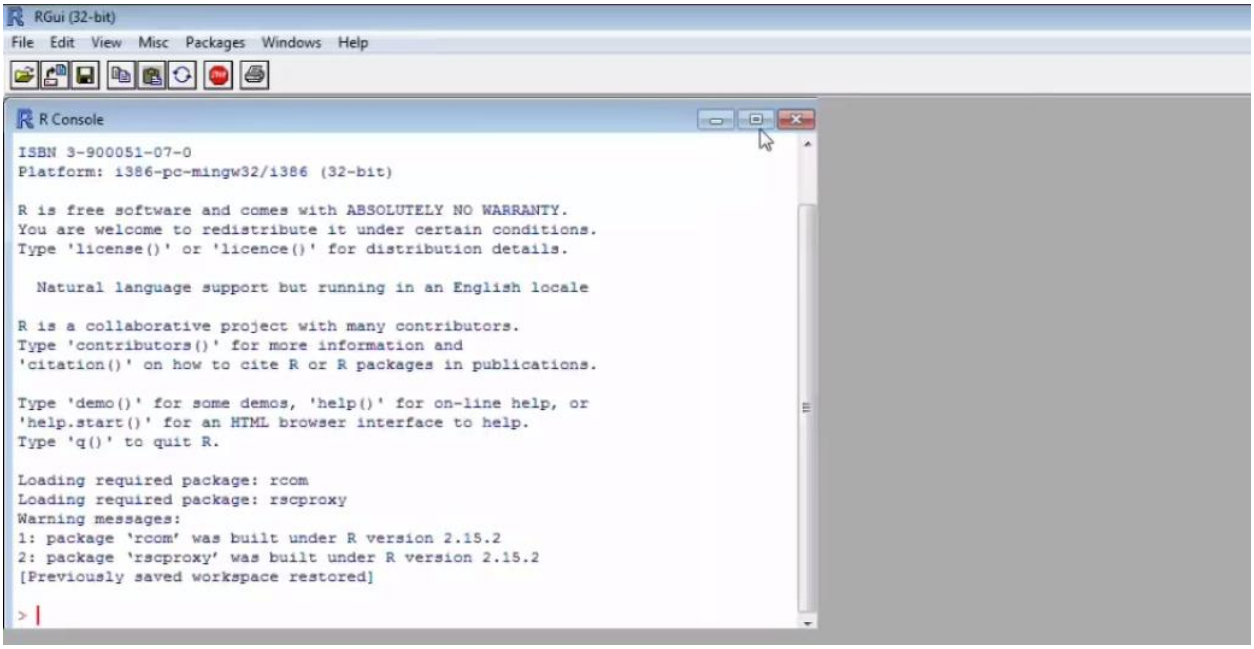

Import
Import Projects
Select a directory to search for existing Eclipse projects.
Select root directory:
Users/sunila/Documents/Workspace/mahout-distribution-0.9
Browse...
Select archive file:
Browse...
Projects:
mahout-buildtools (Users/sunila/Documents/Workspace/mahout-distribution-0.9/buildtools)
mahout-core (Users/sunila/Documents/Workspace/mahout-distribution-0.9/core)
mahout-examples (Users/sunila/Documents/Workspace/mahout-distribution-0.9/examples)
mahout-integration (Users/sunila/Documents/Workspace/mahout-distribution-0.9/integration)
mahout-math (Users/sunila/Documents/Workspace/mahout-distribution-0.9/math)
Select All
Deselect All
Refresh
Copy projects into workspace
Working sets
Add project to working sets
Working sets:
Select...
< Back
Next >
Cancel
Finish

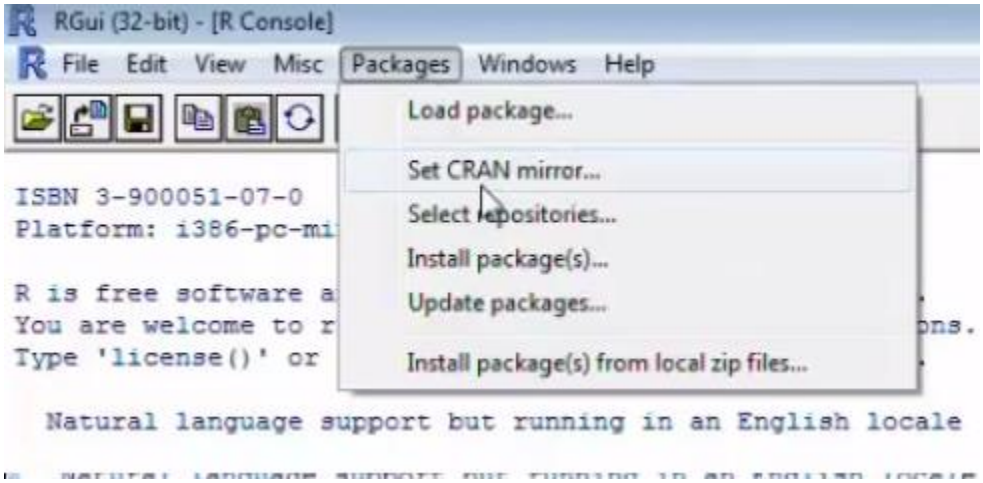


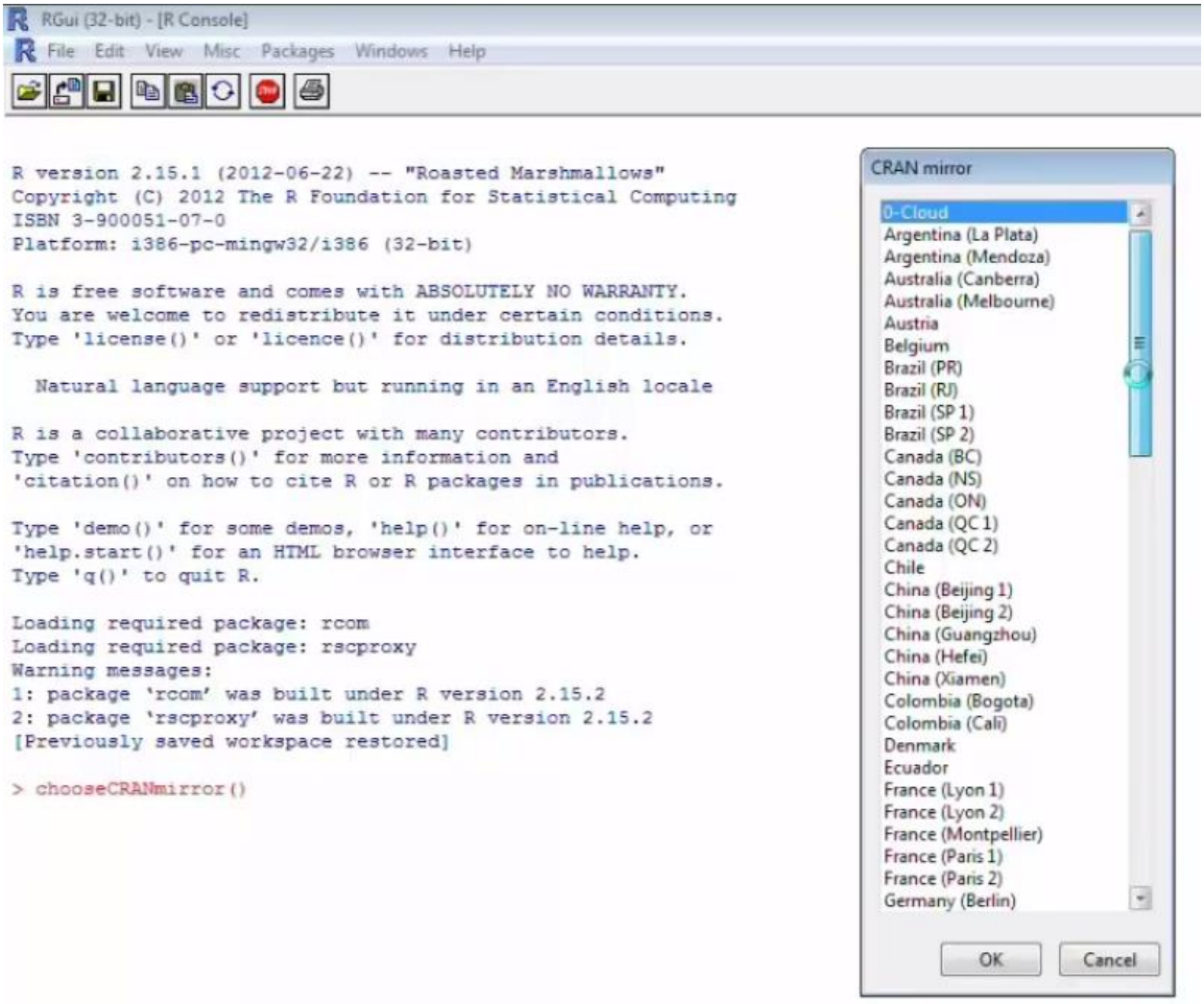

**Applications**

| Evolutionary Algorithms | Classification | Clustering | Recommenders | Regression | FPM | Dimension Reduction |
|---|---|---|---|---|---|---|
| Utilies<br>Lucene/Vectorizer | | Math<br>Vectors/ Matrics/SVD | | Collections<br>(Primitives) | | Hadoop |


Vectors Implementation in Mahout
Dense Vectors
Sparse Vectors
Random Access Sparse Vectors
Sequential Access Sparse Vectors


RGui (32-bit)
File Edit View Misc Packages Windows Help
R Console
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
Loading required package: rcom
Loading required package: rscproxy
Warning messages:
1: package 'rcom' was built under R version 2.15.2
2: package 'rscproxy' was built under R version 2.15.2
[Previously saved workspace restored]
>


RGui (32-bit) - [R Console]
File Edit View Misc Packages Windows Help
Load package...
Set CRAN mirror...
Select repositories...
Install package(s)...
Update packages...
Install package(s) from local zip files...
ISBN 3-900051-07-0
Platform: i386-pc-mi
R is free software a
You are welcome to r
Type 'license()' or
Natural language support but running in an English locale


RGui (32-bit) - [R Console]
File Edit View Misc Packages Windows Help
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
Loading required package: rcom
Loading required package: rscproxy
Warning messages:
1: package 'rcom' was built under R version 2.15.2
2: package 'rscproxy' was built under R version 2.15.2
[Previously saved workspace restored]
> chooseCRANmirror()
CRAN mirror
0-Cloud
Argentina (La Plata)
Argentina (Mendoza)
Australia (Canberra)
Australia (Melbourne)
Austria
Belgium
Brazil (PR)
Brazil (RJ)
Brazil (SP 1)
Brazil (SP 2)
Canada (BC)
Canada (NS)
Canada (ON)
Canada (QC 1)
Canada (QC 2)
Chile
China (Beijing 1)
China (Beijing 2)
China (Guangzhou)
China (Hefei)
China (Xiamen)
Colombia (Bogota)
Colombia (Cali)
Denmark
Ecuador
France (Lyon 1)
France (Lyon 2)
France (Montpellier)
France (Paris 1)
France (Paris 2)
Germany (Berlin)
OK
Cancel

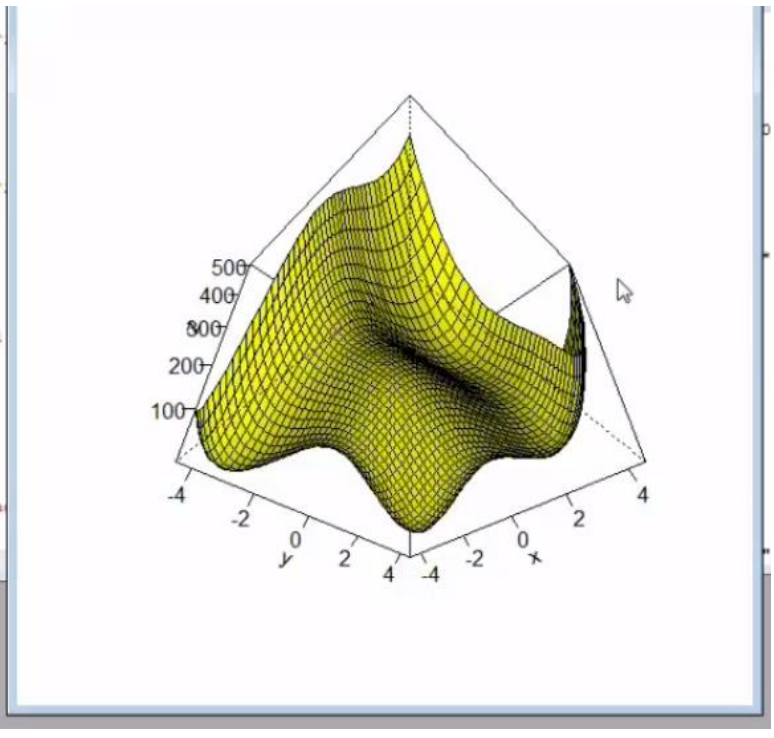
R Console

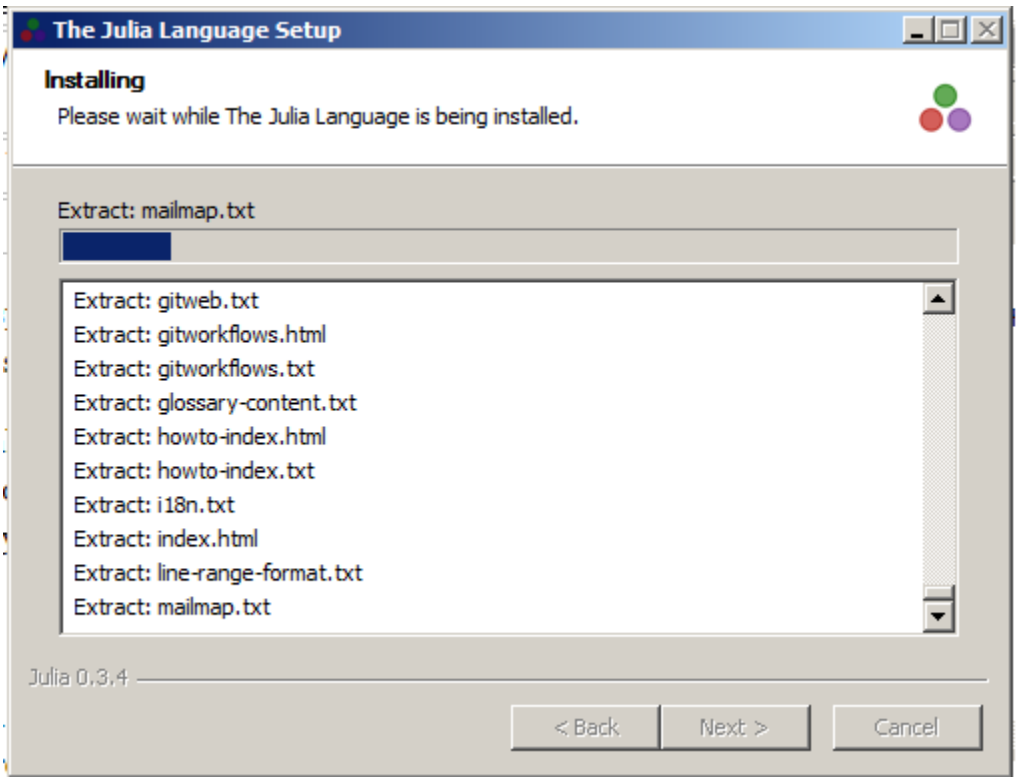
```
Warning messages:
1: package 'rcom' was built under R version 2.15.2
2: package 'rscproxy' was built under R version 2.15.2
[Previously saved workspace restored]

> chooseCRANmirror()
> 1+2
[1] 3
> log(10)
[1] 2.302585
> 1+2
[1] 3
> a = 2
> a
[1] 2
> a <-2
> a
[1] 2
> a <-10
> a
[1] 10
> b <-5
> b
[1] 5
>
```

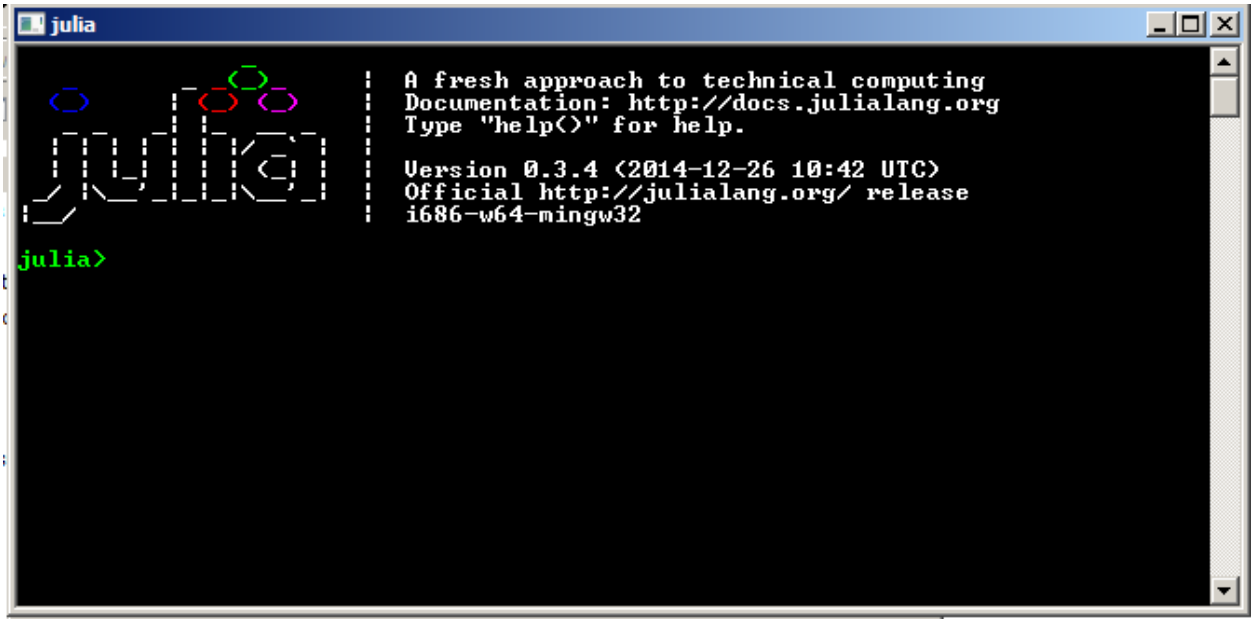
Untitled - R Editor

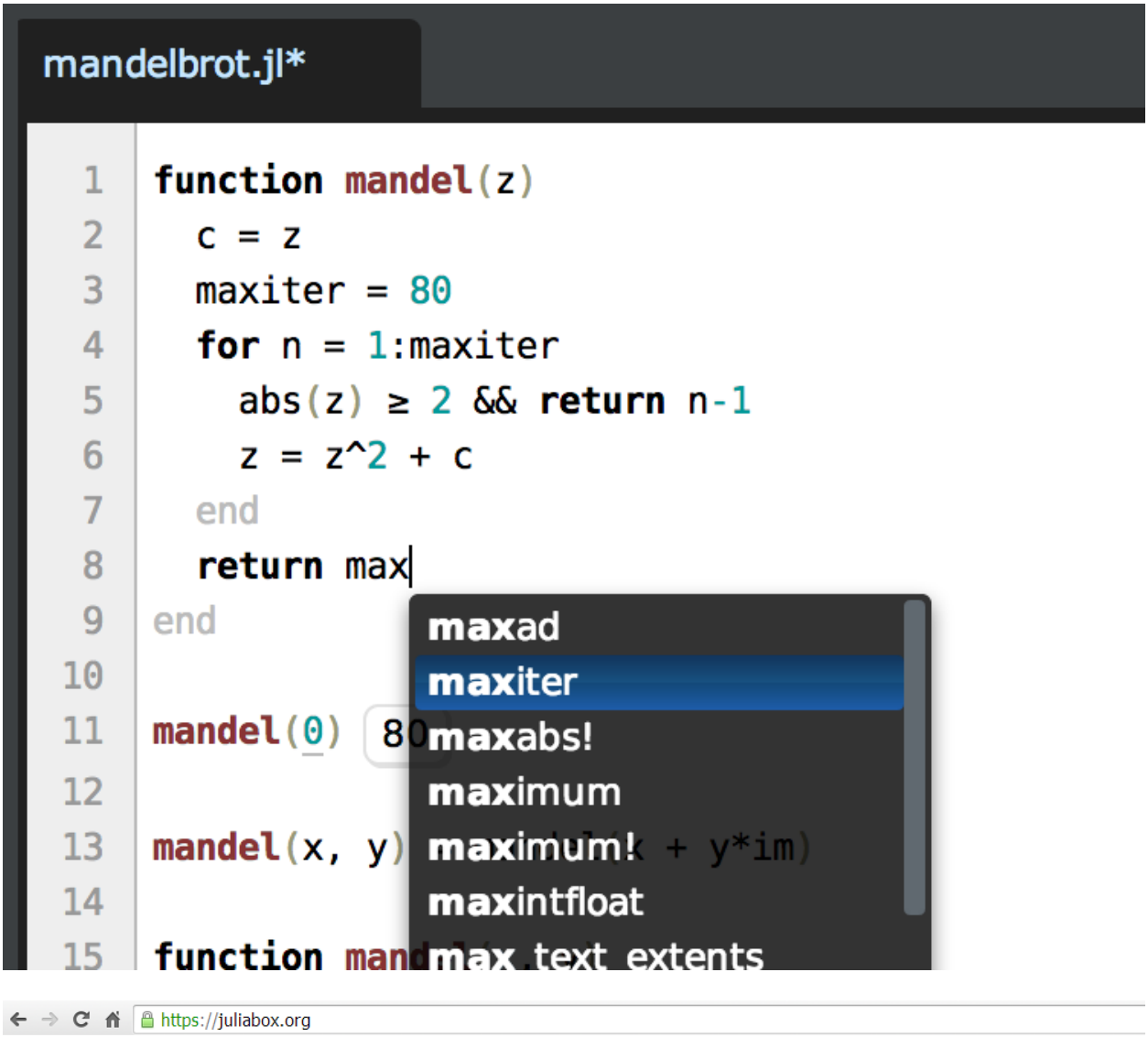
```
1+2
a = 2
a <-2
a <-10
b <-5
c<-a+
```

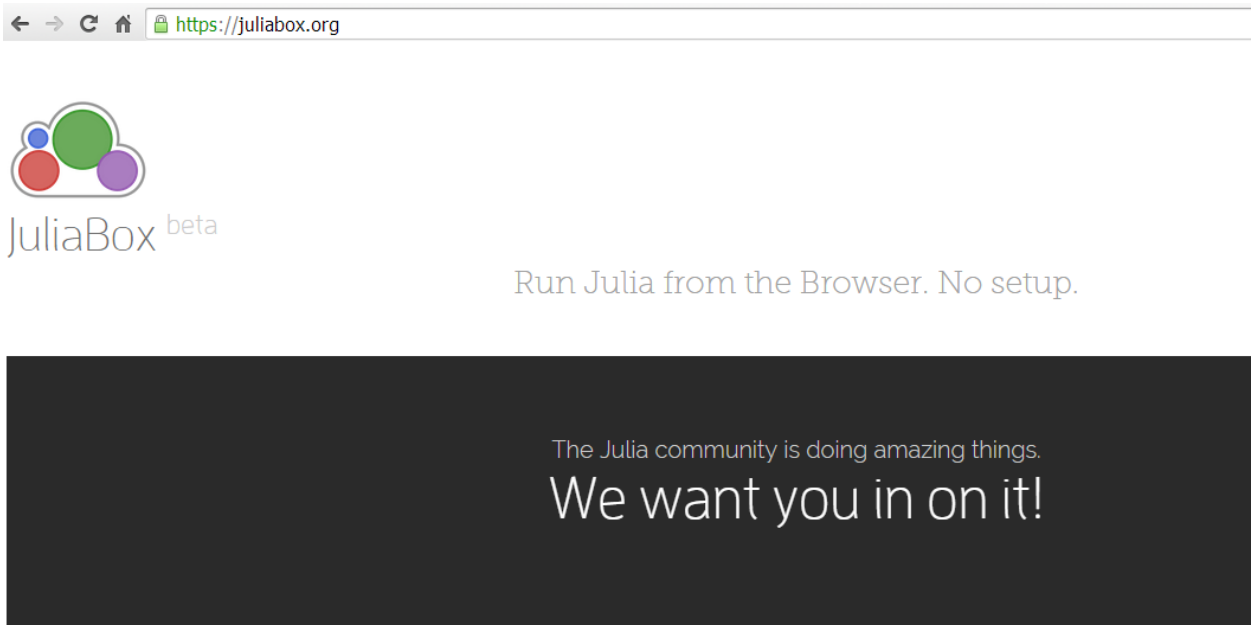
```
f <- function(x1,y1) (1-x1)^2 + 100*(y1 - x1^

optim( c(0,0), f )$par
ror in y1 - x1^2 : 'y1' is missing


f <- function(x) (1-x[1])^2 + 100*(x[2]-x[1]^

optim( c(0,0), f )$par
] 0.9999564 0.9999085


f <- function(x1,y1) (x1^2 + y1 - 11)^2 + (x1

x <- seq(-4.5,4.5,by=.2)

y <- seq(-4.5,4.5,by=.2)

z <- outer(x,y,f)

persp(x,y,z,phi=-45,theta=45,col="yellow",sha
```




The Julia Language Setup
Installing
Please wait while The Julia Language is being installed.
Extract: mailmap.txt
Extract: gitweb.txt
Extract: gitworkflows.html
Extract: gitworkflows.txt
Extract: glossary-content.txt
Extract: howto-index.html
Extract: howto-index.txt
Extract: i18n.txt
Extract: index.html
Extract: line-range-format.txt
Extract: mailmap.txt
Julia 0.3.4
< Back
Next >
Cancel

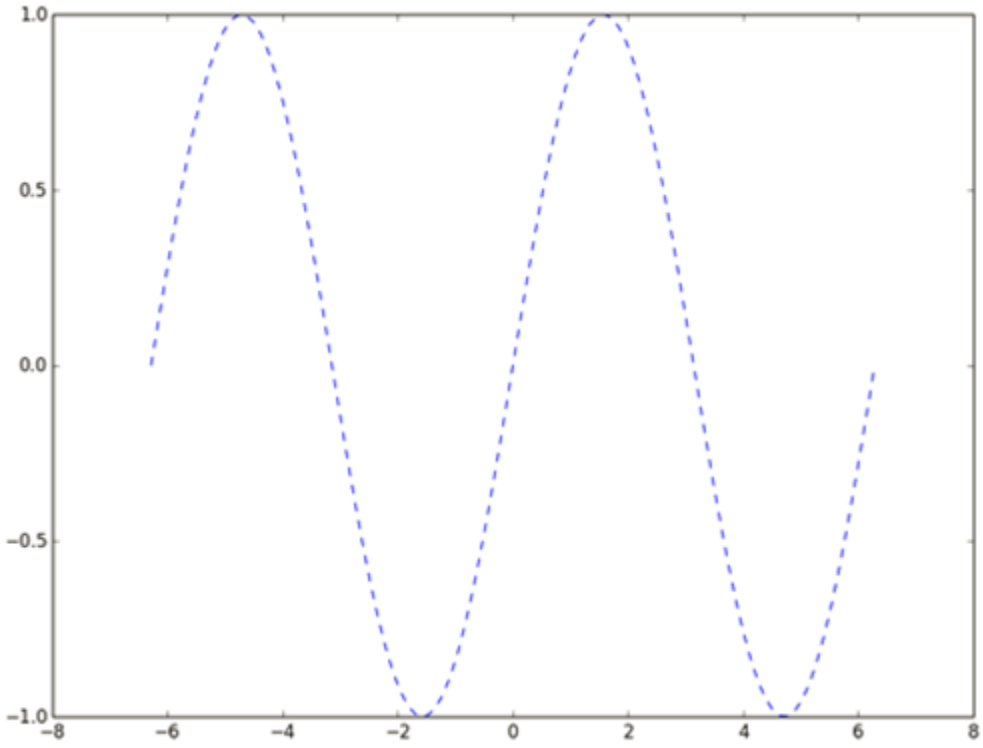

julia
A fresh approach to technical computing
Documentation: http://docs.julialang.org
Type "help()" for help.
Version 0.3.4 (2014-12-26 10:42 UTC)
Official http://julialang.org/ release
i686-w64-mingw32
julia>

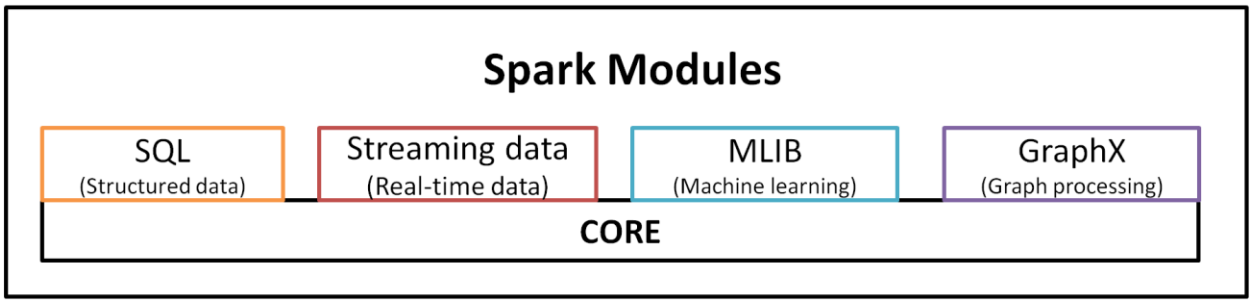


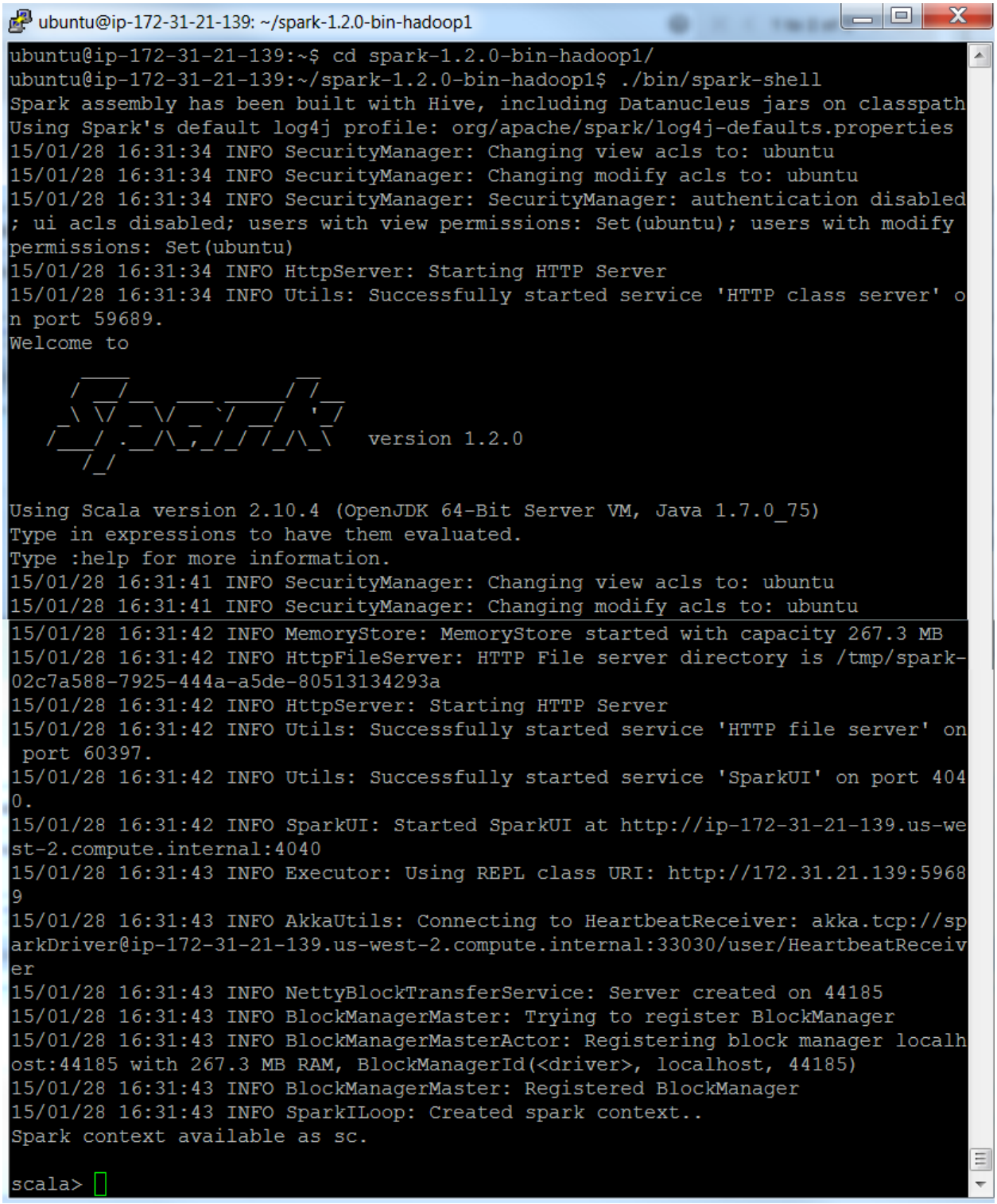




The Julia community is doing amazing things.

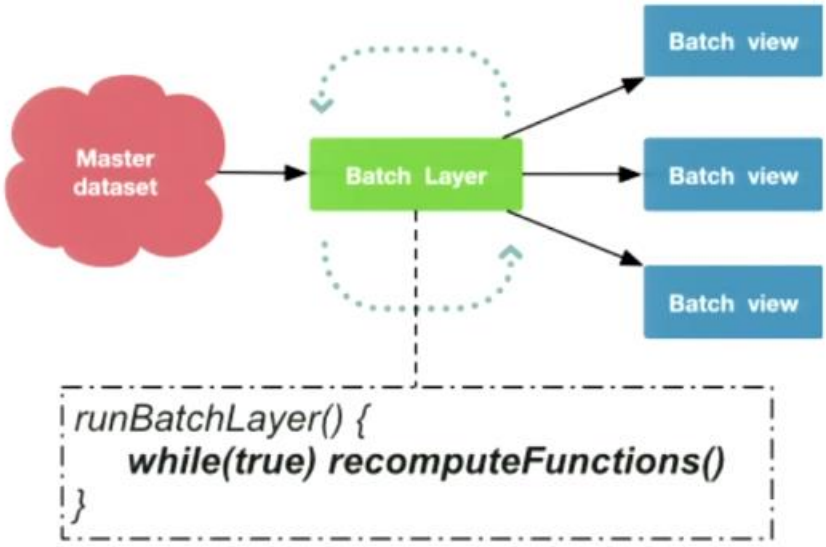
We want you in on it!


1.0
0.5
0.0
−0.5
−1.0
−8
−6
−4
−2
0
2
4
6
8

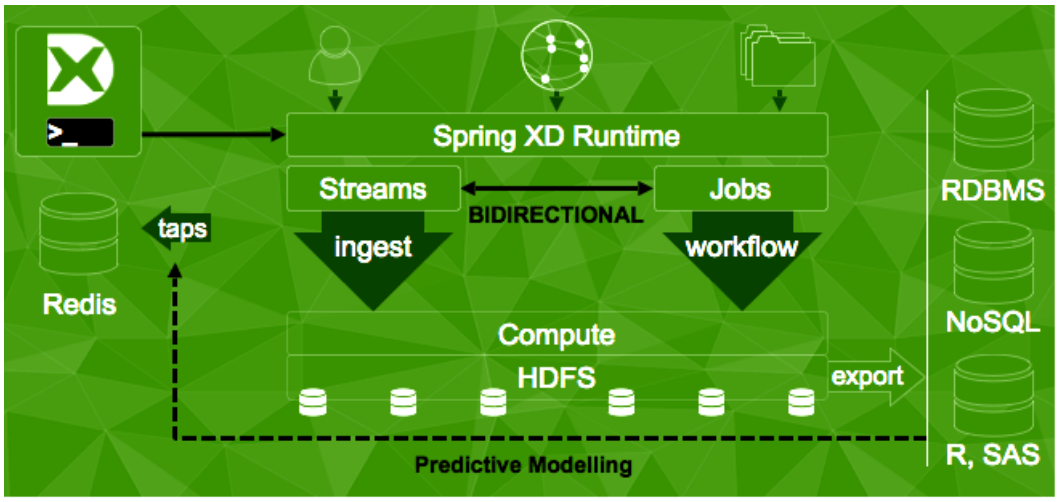

Spark Modules
SQL
(Structured data)
Streaming data
(Real-time data)
MLIB
(Machine learning)
GraphX
(Graph processing)
CORE


ubuntu@ip-172-31-21-139: ~/spark-1.2.0-bin-hadoop1
ubuntu@ip-172-31-21-139:~$ cd spark-1.2.0-bin-hadoop1/
ubuntu@ip-172-31-21-139:~/spark-1.2.0-bin-hadoop1$ ./bin/spark-shell
Spark assembly has been built with Hive, including Datanucleus jars on classpath
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
15/01/28 16:31:34 INFO SecurityManager: Changing view acls to: ubuntu
15/01/28 16:31:34 INFO SecurityManager: Changing modify acls to: ubuntu
15/01/28 16:31:34 INFO SecurityManager: SecurityManager: authentication disabled
; ui acls disabled; users with view permissions: Set(ubuntu); users with modify
permissions: Set(ubuntu)
15/01/28 16:31:34 INFO HttpServer: Starting HTTP Server
15/01/28 16:31:34 INFO Utils: Successfully started service 'HTTP class server' o
n port 59689.
Welcome to
version 1.2.0
Using Scala version 2.10.4 (OpenJDK 64-Bit Server VM, Java 1.7.0_75)
Type in expressions to have them evaluated.
Type :help for more information.
15/01/28 16:31:41 INFO SecurityManager: Changing view acls to: ubuntu
15/01/28 16:31:41 INFO SecurityManager: Changing modify acls to: ubuntu
15/01/28 16:31:42 INFO MemoryStore: MemoryStore started with capacity 267.3 MB
15/01/28 16:31:42 INFO HttpFileServer: HTTP File server directory is /tmp/spark-
02c7a588-7925-444a-a5de-80513134293a
15/01/28 16:31:42 INFO HttpServer: Starting HTTP Server
15/01/28 16:31:42 INFO Utils: Successfully started service 'HTTP file server' on
port 60397.
15/01/28 16:31:42 INFO Utils: Successfully started service 'SparkUI' on port 404
0.
15/01/28 16:31:42 INFO SparkUI: Started SparkUI at http://ip-172-31-21-139.us-we
st-2.compute.internal:4040
15/01/28 16:31:43 INFO Executor: Using REPL class URI: http://172.31.21.139:5968
9
15/01/28 16:31:43 INFO AkkaUtils: Connecting to HeartbeatReceiver: akka.tcp://sp
arkDriver@ip-172-31-21-139.us-west-2.compute.internal:33030/user/HeartbeatReceiv
er
15/01/28 16:31:43 INFO NettyBlockTransferService: Server created on 44185
15/01/28 16:31:43 INFO BlockManagerMaster: Trying to register BlockManager
15/01/28 16:31:43 INFO BlockManagerMasterActor: Registering block manager localh
ost:44185 with 267.3 MB RAM, BlockManagerId(<driver>, localhost, 44185)
15/01/28 16:31:43 INFO BlockManagerMaster: Registered BlockManager
15/01/28 16:31:43 INFO SparkILoop: Created spark context..
Spark context available as sc.
scala>

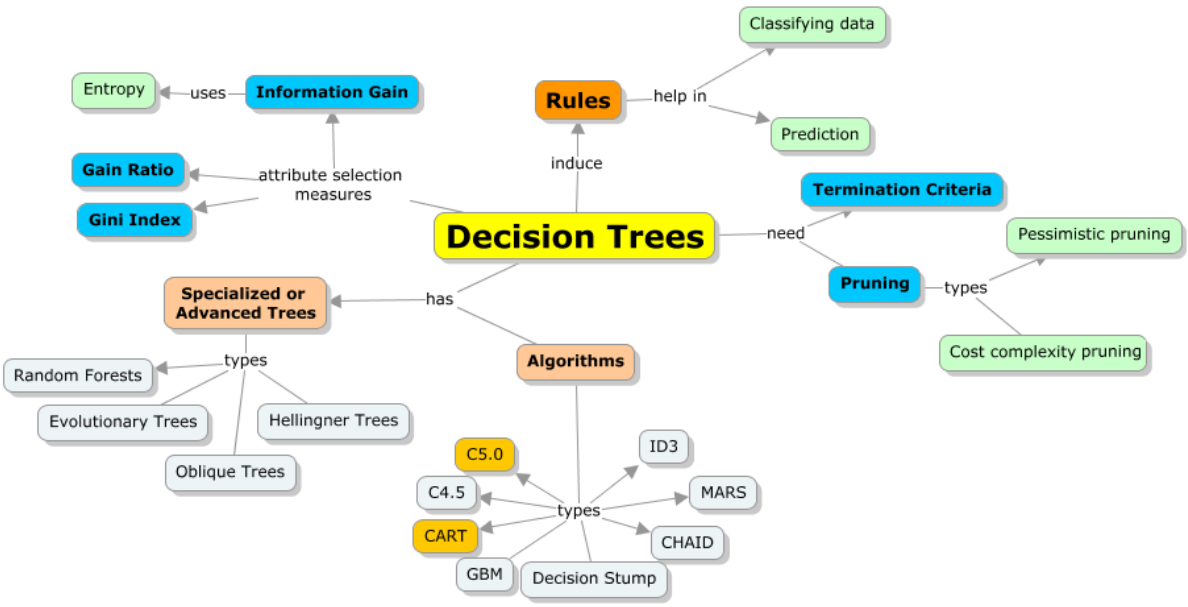

RealTime Function
RealTime View

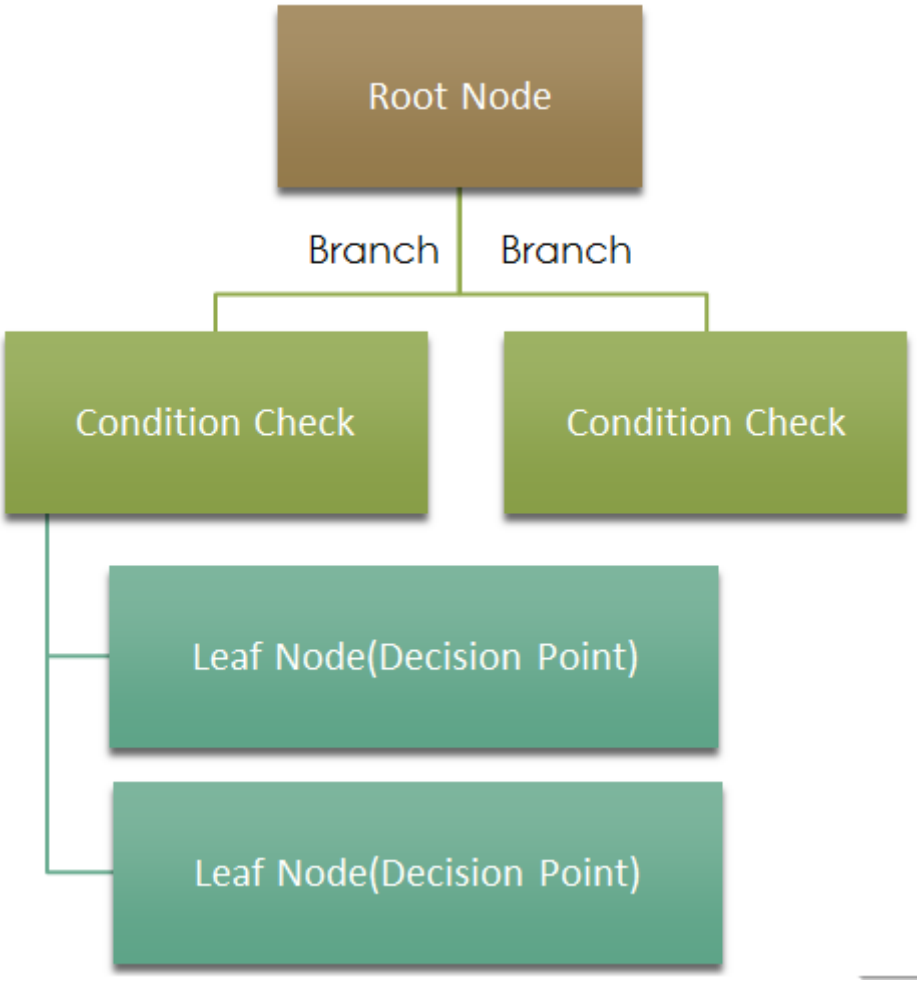

Batch view
Master dataset
Batch Layer
Batch view
Batch view
runBatchLayer() {
while(true) recomputeFunctions()
}

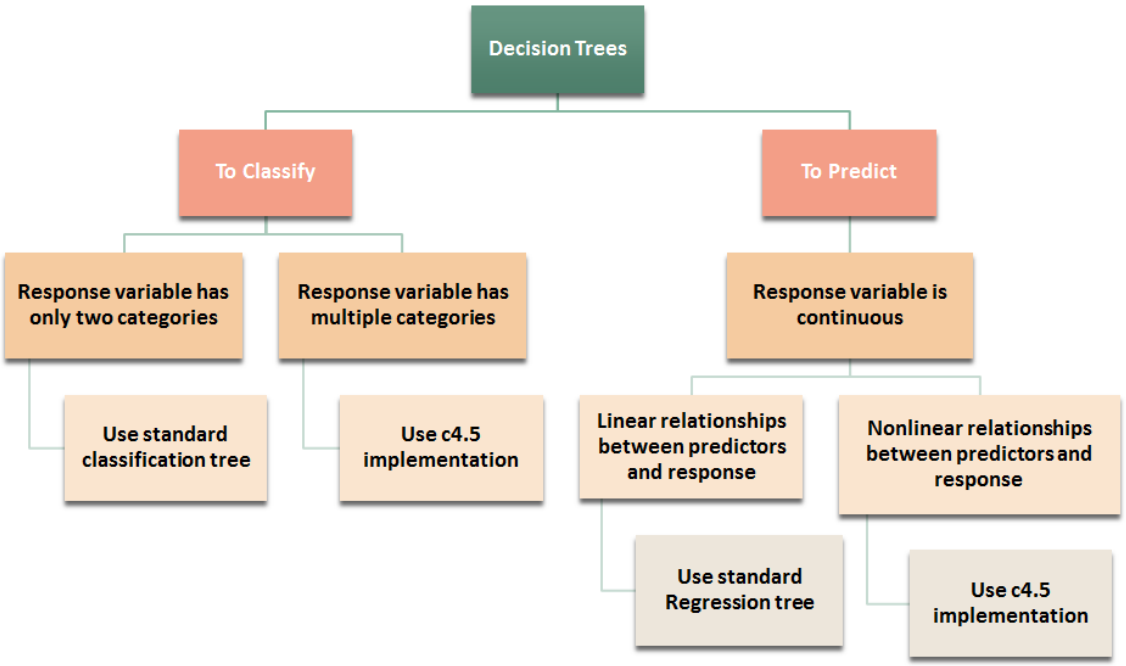

Batch view
Batch view


Spring XD Runtime
Streams
Jobs
BIDIRECTIONAL
ingest
workflow
taps
Redis
Compute
HDFS
export
Predictive Modelling
RDBMS
NoSQL
R, SAS

# CHAPTER 5

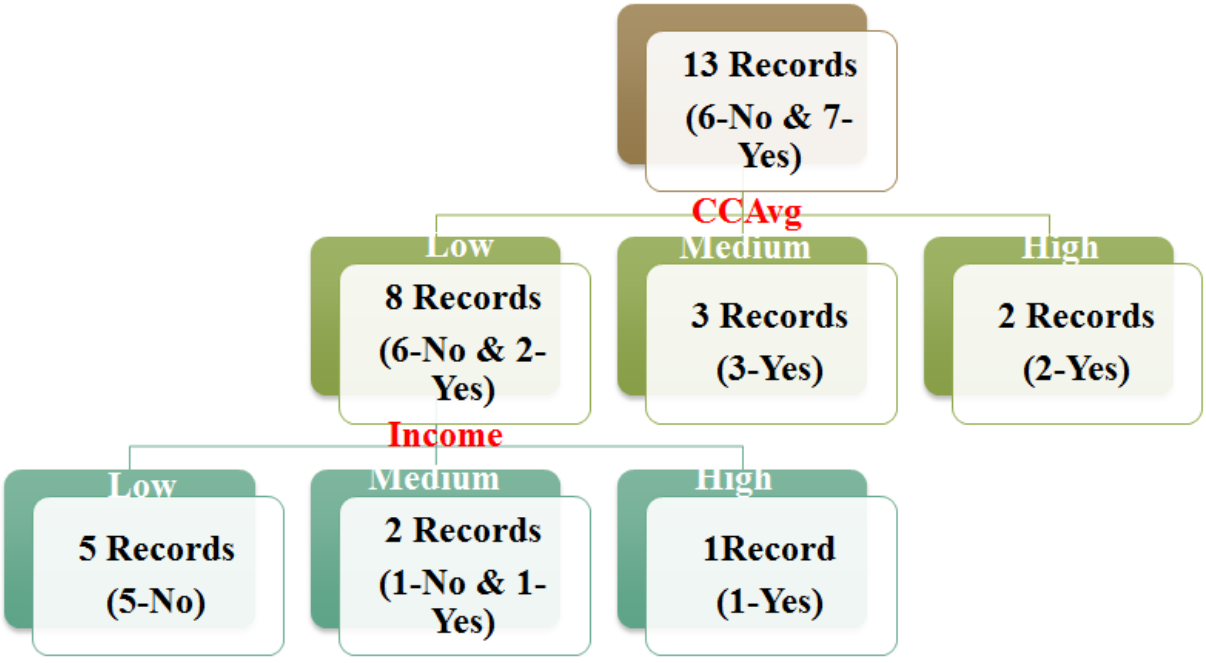

Classifying data
Entropy
uses
Information Gain
Rules
help in
Prediction
Gain Ratio
attribute selection measures
induce
Gini Index
Termination Criteria
Decision Trees
need
Pessimistic pruning
Pruning
types
Specialized or Advanced Trees
has
Cost complexity pruning
types
Algorithms
Random Forests
Evolutionary Trees
Hellingner Trees
Oblique Trees
C5.0
ID3
C4.5
MARS
types
CART
CHAID
GBM
Decision Stump

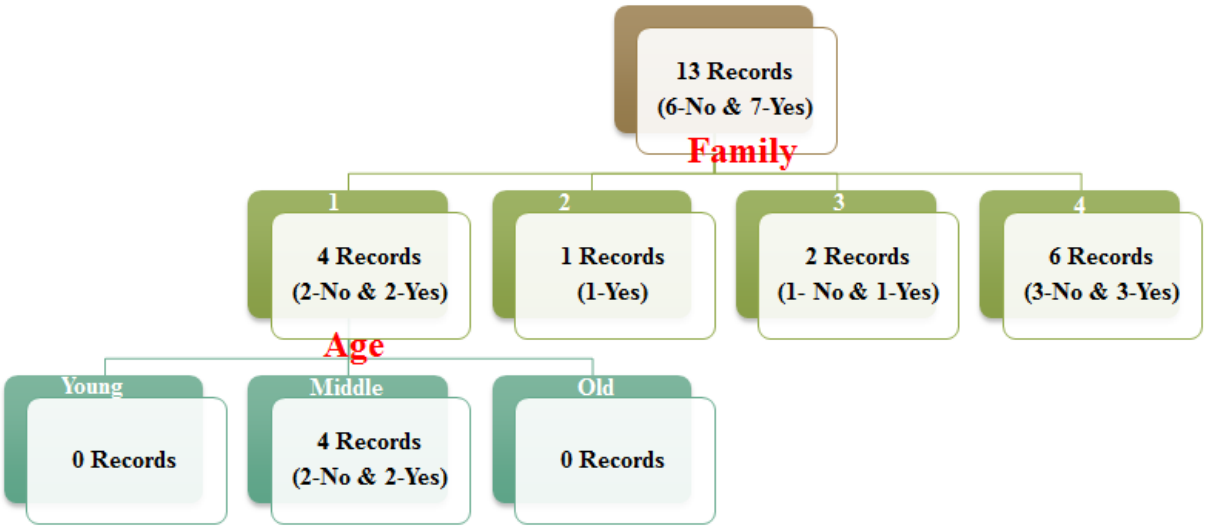


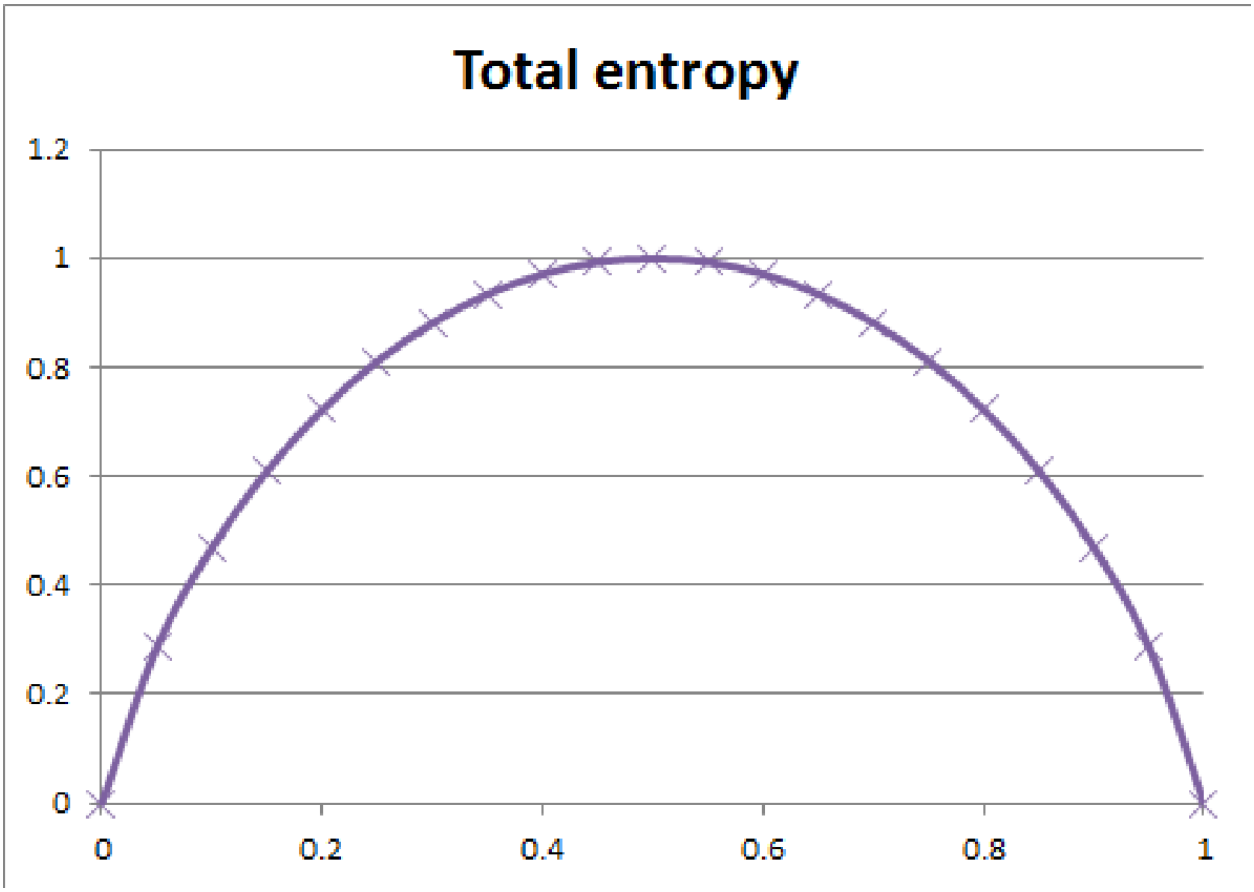




Source: http://www.simafore.com/blog/bid/62482/2-main-differences-between-classification-and-regression-trees

| ID | Age | Experience | Income | Family | CCAvg | Personal Loan |
|---|---|---|---|---|---|---|
| 1 | 25 | 1 | 49 | 4 | 1.60 | 0 |
| 2 | 45 | 19 | 34 | 3 | 1.50 | 0 |
| 3 | 39 | 15 | 11 | 1 | 1.00 | 0 |
| 4 | 35 | 9 | 100 | 1 | 2.70 | 0 |
| 5 | 35 | 8 | 45 | 4 | 1.00 | 0 |
| 6 | 37 | 13 | 29 | 4 | 0.40 | 0 |
| 10 | 34 | 9 | 180 | 1 | 8.90 | 1 |
| 17 | 38 | 14 | 130 | 4 | 4.70 | 1 |
| 19 | 46 | 21 | 193 | 2 | 8.10 | 1 |
| 30 | 38 | 13 | 119 | 1 | 3.30 | 1 |
| 39 | 42 | 18 | 141 | 3 | 5.00 | 1 |
| 43 | 32 | 7 | 132 | 4 | 1.10 | 1 |
| 48 | 37 | 12 | 194 | 4 | 0.20 | 1 |




13 Records
(6-No & 7-Yes)
Family
1
4 Records
(2-No & 2-Yes)
2
1 Records
(1-Yes)
3
2 Records
(1- No & 1-Yes)
4
6 Records
(3-No & 3-Yes)
Age
Young
0 Records
Middle
4 Records
(2-No & 2-Yes)
Old
0 Records


Total entropy
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

$$E = -\sum_{i=1}^{m} p_i log_2(p_i)$$

$$E_A = \sum_{i=1}^{v} \frac{D_i}{D} E(D_i)$$

| ID | Age | Income | Family | CCAvg | Personal Loan |
|---|---|---|---|---|---|
| 1 | Young | Low | 4 | Low | 0 |
| 2 | Old | Low | 3 | Low | 0 |
| 3 | Middle | Low | 1 | Low | 0 |
| 4 | Middle | Medium | 1 | Low | 0 |
| 5 | Middle | Low | 4 | Low | 0 |
| 6 | Middle | Low | 4 | Low | 0 |
| 10 | Middle | High | 1 | High | 1 |
| 17 | Middle | Medium | 4 | Medium | 1 |
| 19 | Old | High | 2 | High | 1 |
| 30 | Middle | Medium | 1 | Medium | 1 |
| 39 | Old | Medium | 3 | Medium | 1 |
| 43 | Young | Medium | 4 | Low | 1 |
| 48 | Middle | High | 4 | Low | 1 |

$$-\frac{6}{13} log_2\left(\frac{6}{13}\right) - \frac{7}{13} log_2\left(\frac{7}{13}\right) = 0.995727$$

| | | CCAvg | | |
|---|---|---|---|---|
| | Fraction | Loan=No | Loan=Yes | Entropy |
| Low | 0.615385 | 0.25 | 0.75 | 0.81 |
| Medium | 0.230769 | 0.00 | 1.00 | 0.00 |
| High | 0.153846 | 0.00 | 1.00 | 0.00 |
| Entropy | 0.499248 | | | |

$$Entropy_{CCAvg} = \frac{8}{13}E(6,2) + \frac{3}{13}E(0,3) + \frac{2}{13}E(0,2) = 0.499248$$

$$I_{CCAvg} = 0.995727 - 0.499248 = 0.496479$$
$$I_{Family} = 0.995727 - 0.923077 = 0.07265$$

$$\left(\frac{6}{13}\right)^2 + \left(\frac{7}{13}\right)^2$$

| | CCAvg | | | |
|---|---|---|---|---|
| | Fraction | Loan=No | Loan=Yes | Gini |
| Low | 0.615385 | 0.250000 | 0.750000 | 0.625 |
| Medium | 0.230769 | 0.000000 | 1.000000 | 1.000 |
| High | 0.153846 | 0.000000 | 1.000000 | 1.000 |
| Gini Index | | | 0.769231 | |

| | | Family | | |
|---|---|---|---|---|
| | Fraction | Loan=No | Loan=Yes | Entropy |
| 1 | 0.307692 | 0.5 | 0.5 | 0.5 |
| 2 | 0.076923 | 0 | 1 | 1 |
| 3 | 0.153846 | 0.50 | 0.50 | 0.5 |
| 4 | 0.461538 | 0.50 | 0.50 | 0.5 |
| Gini Index | | | 0.538462 | |

Men

| Number | Age | Married |
|---|---|---|
| Man 1 | 1 | No |
| Man 2 | 2 | No |
| ... | ... | ... |
| Man 40 | 40 | No |

Female

| Number | Age | Married |
|---|---|---|
| Woman 1 | 41 | Yes |
| Woman 2 | 42 | Yes |
| ... | ... | ... |
| Woman 59 | 99 | Yes |
| Woman 60 | 99 | No |

***Partition D into $D_{train}$ (training / "growing"), $D_{validation}$(validation / "pruning")***

***Build complete tree T on $D_{train}$***

***UNTIL accuracy on $D_{validation}$ decreases DO***

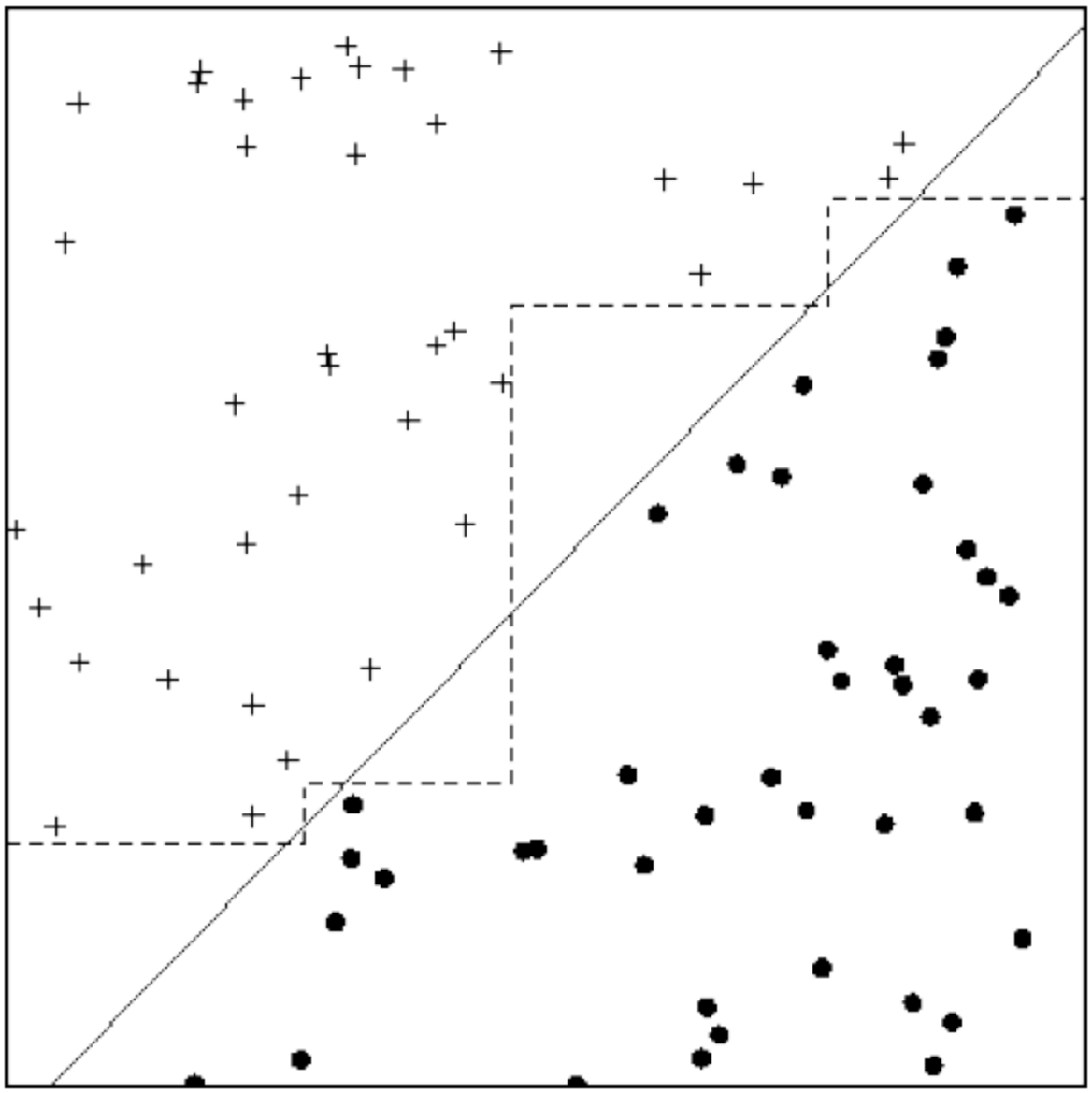
***FOR each non-leaf node candidate in T***

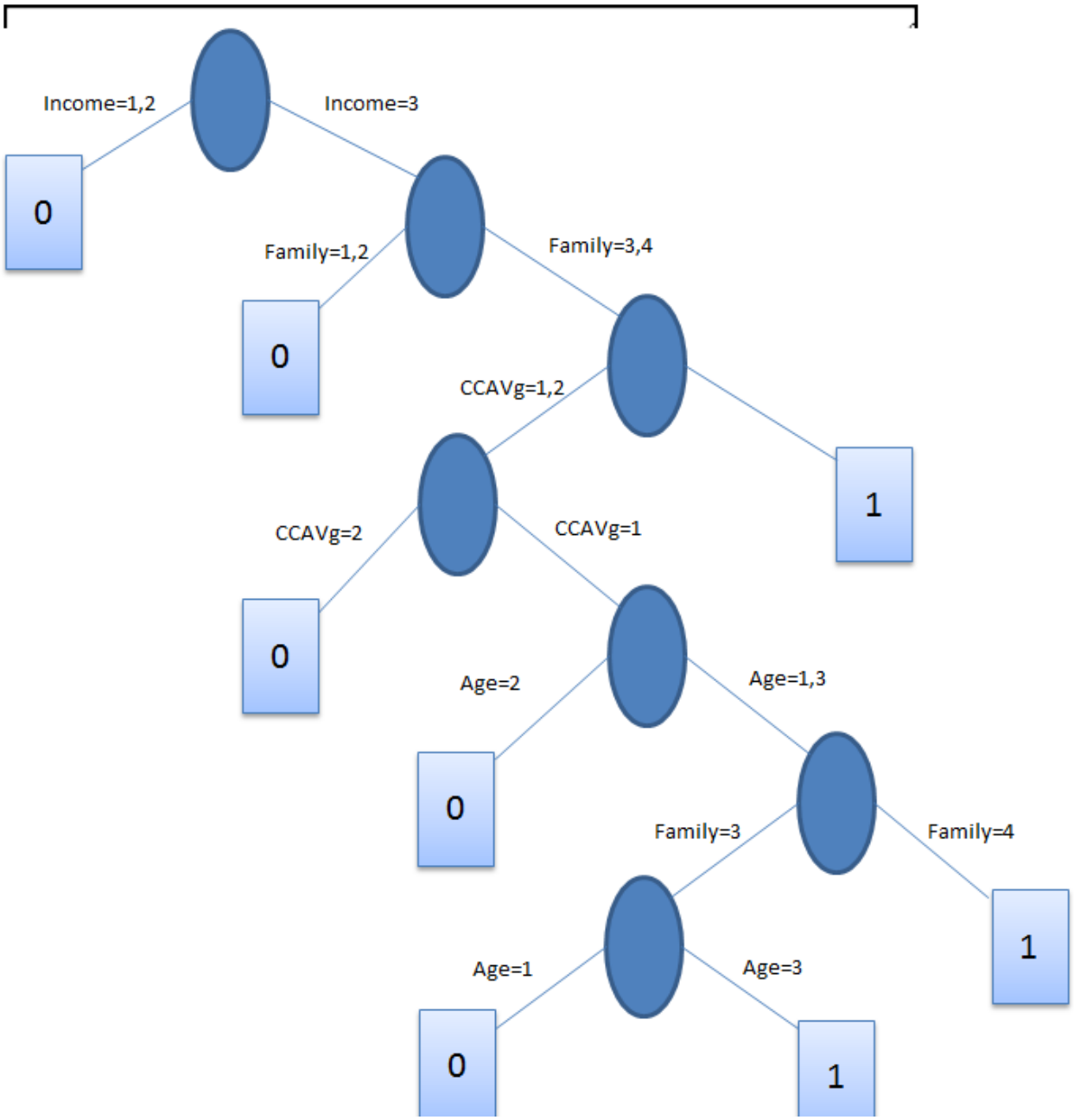
***Temp[candidate] ←Prune (T, candidate)***

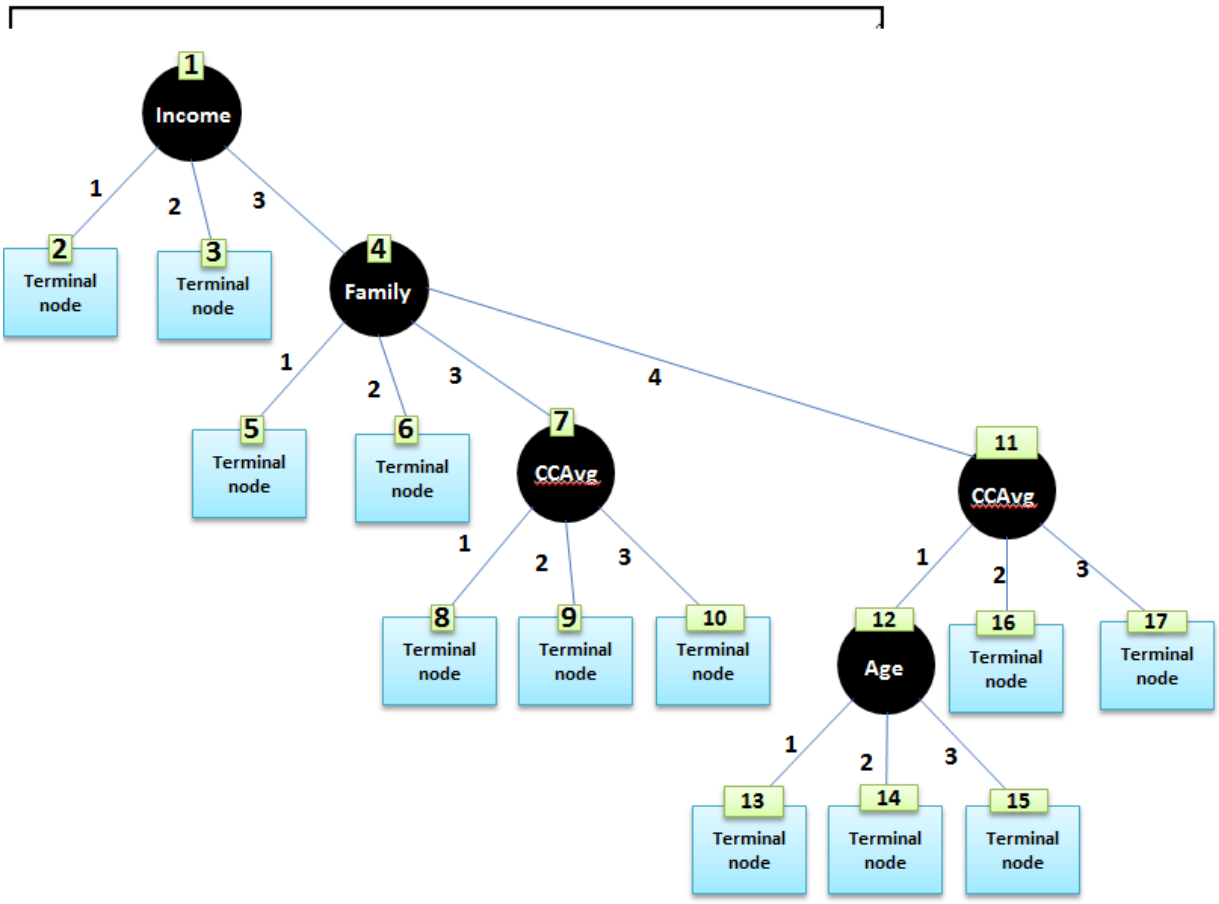
***Accuracy[candidate] ←Test (Temp[candidate], $D_{validation}$)***

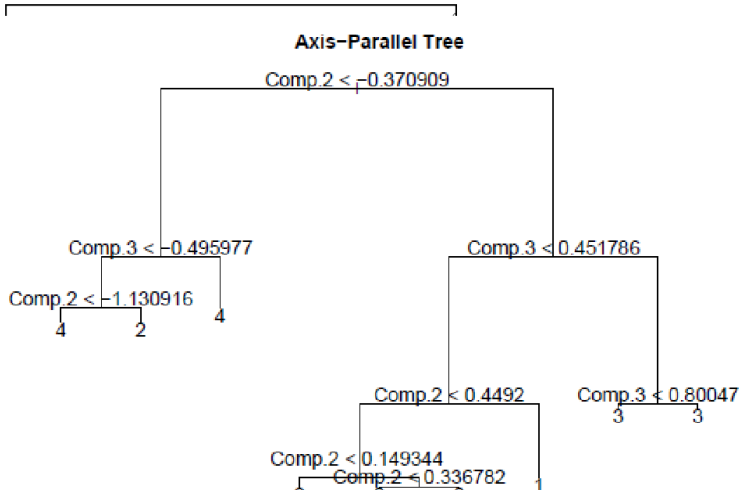
***T←T' ∈Temp with best value of Accuracy (best increase; greedy)***

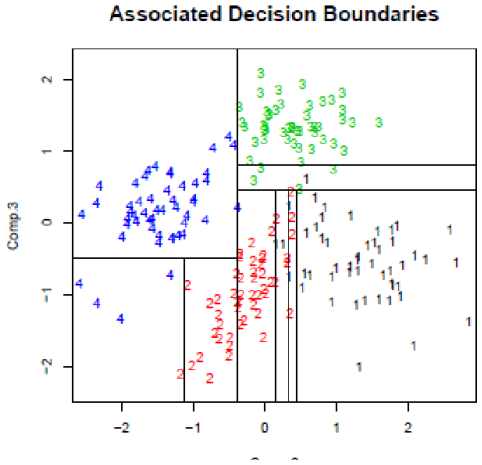
***RETURN (pruned)***

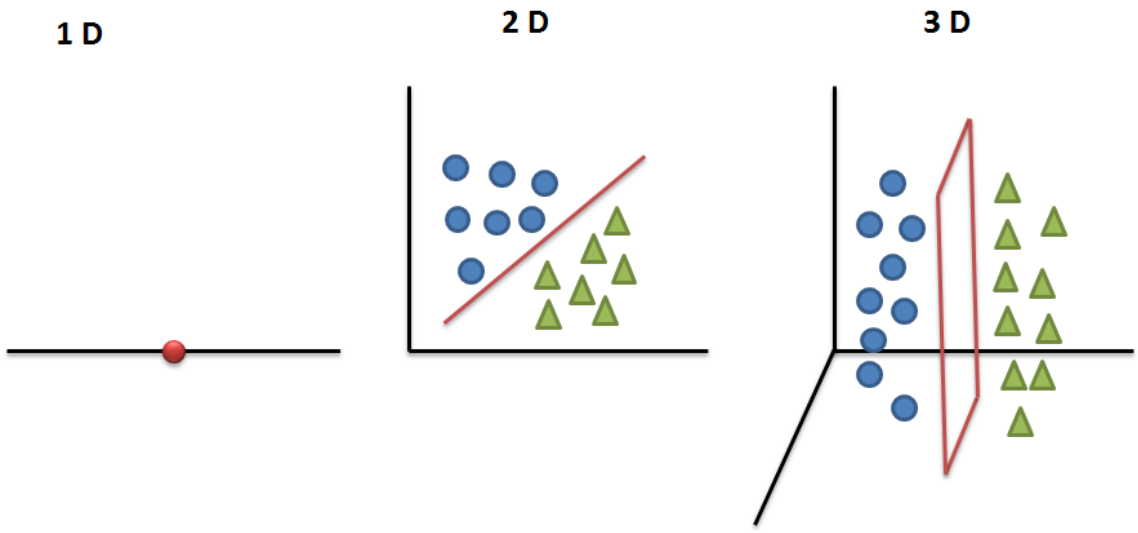


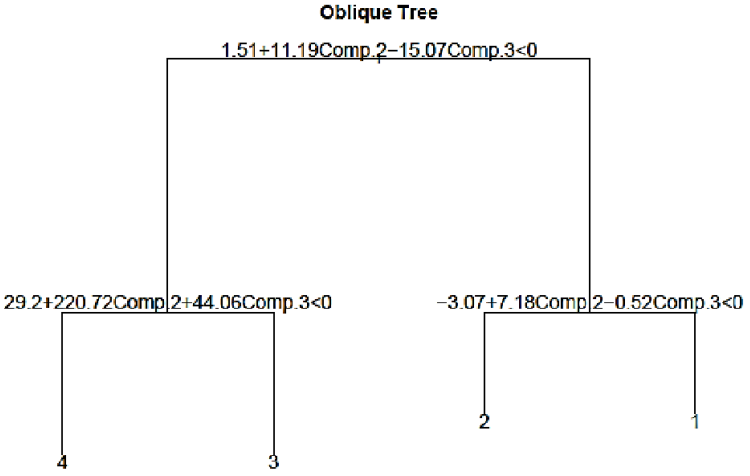

Income=1,2
Income=3
0
Family=1,2
Family=3,4
0
CCAVg=1,2
1
CCAVg=2
CCAVg=1
0
Age=2
Age=1,3
0
Family=3
Family=4
1
Age=1
Age=3
0
1

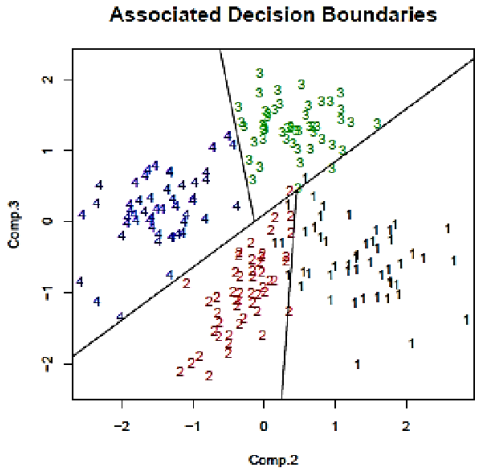

1
Income
1
2
3
2
Terminal node
3
Terminal node
4
Family
1
2
3
4
5
Terminal node
6
Terminal node
7
CCAvg
11
CCAvg
1
2
3
8
Terminal node
9
Terminal node
10
Terminal node
12
Age
16
Terminal node
17
Terminal node
1
2
3
13
Terminal node
14
Terminal node
15
Terminal node

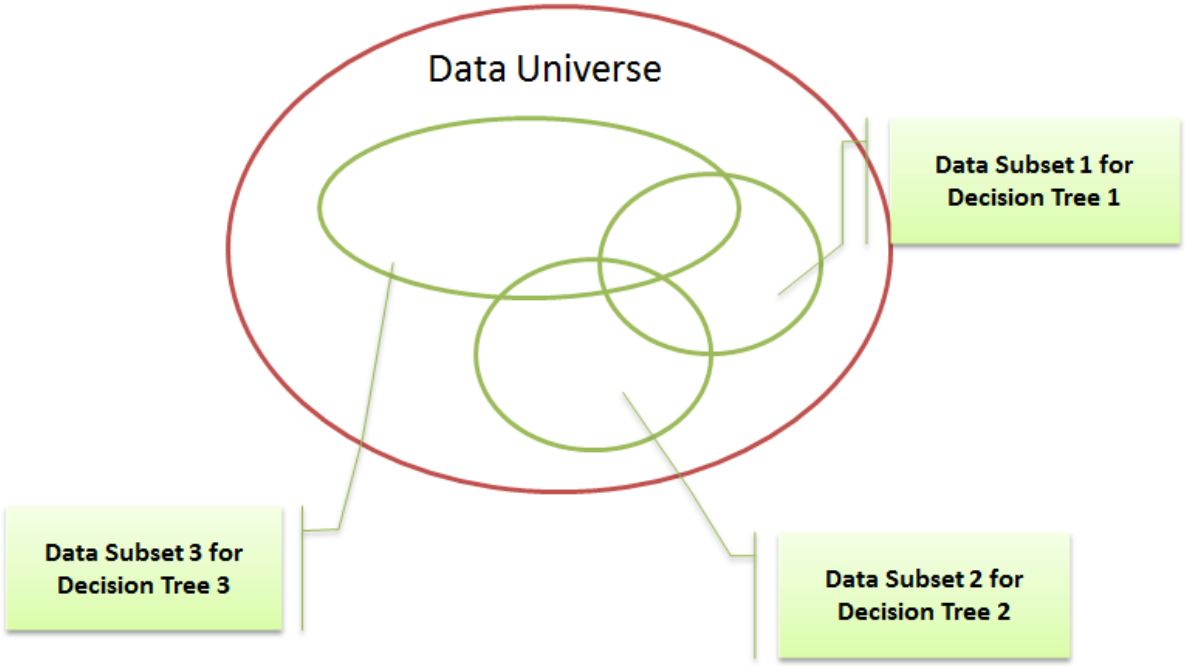

Axis-Parallel Tree
Comp.2 < −0.370909
Comp.3 < −0.495977
Comp.3 < 0.451786
Comp.2 < −1.130916
4
4
2
Comp.2 < 0.4492
Comp.3 < 0.80047
3
3
Comp.2 < 0.149344
Comp.2 < 0.336782


Associated Decision Boundaries
Comp.3
2
1
0
−1
−2
−2
−1
0
1
2


1 D
2 D
3 D


Oblique Tree
1.51+11.19Comp.2−15.07Comp.3<0
29.2+220.72Comp.2+44.06Comp.3<0
−3.07+7.18Comp.2−0.52Comp.3<0
4
3
2
1


Associated Decision Boundaries
Comp.3
Comp.2
−2
−1
0
1
2


Data Universe
Data Subset 1 for Decision Tree 1
Data Subset 3 for Decision Tree 3
Data Subset 2 for Decision Tree 2

| S No | Variable |
|---|---|
| 1 | X1 |
| 2 | X2 |
| 3 | X3 |
| 4 | X4 |
| 5 | X5 |
| ... | ... |
| N | Xn |

**X1, X2, X3 variable set for Decision Tree 1**

**X3, X4, X5 variable set for Decision Tree 2**

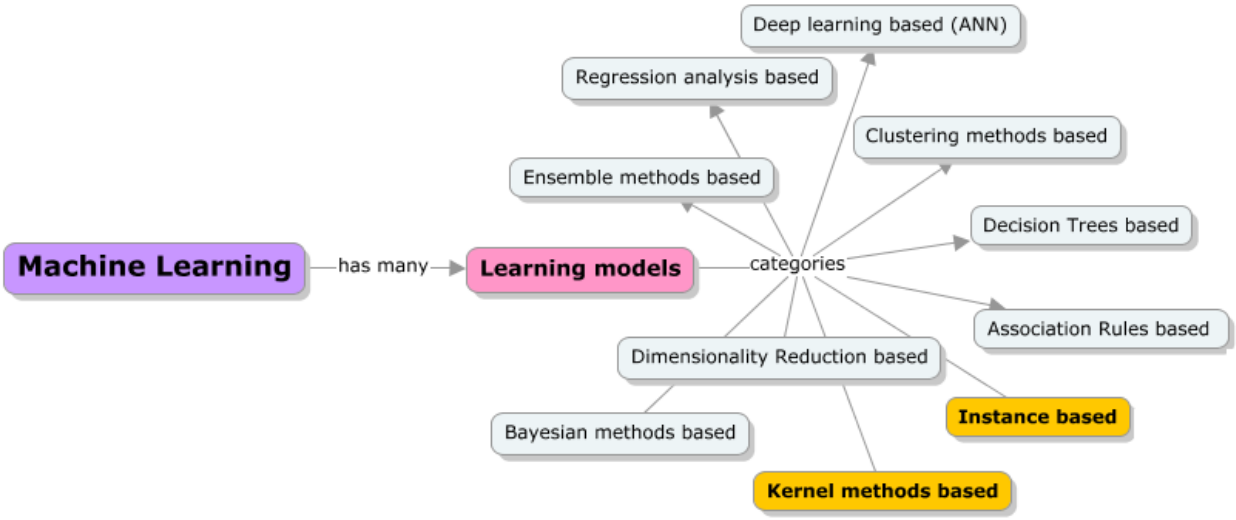
**Xa, Xb, Xn variable set for Decision Tree 3**

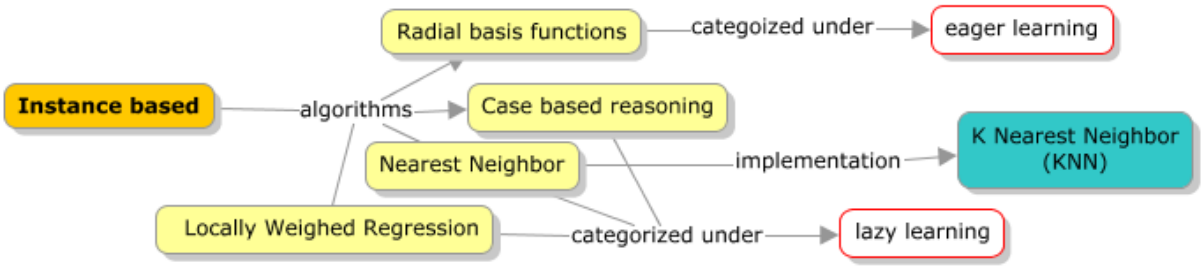
$$\mathrm{MSE}\left[\hat{f}_m(p)\right] = O\left(\frac{1}{m^{4/(D+4)}}\right)$$

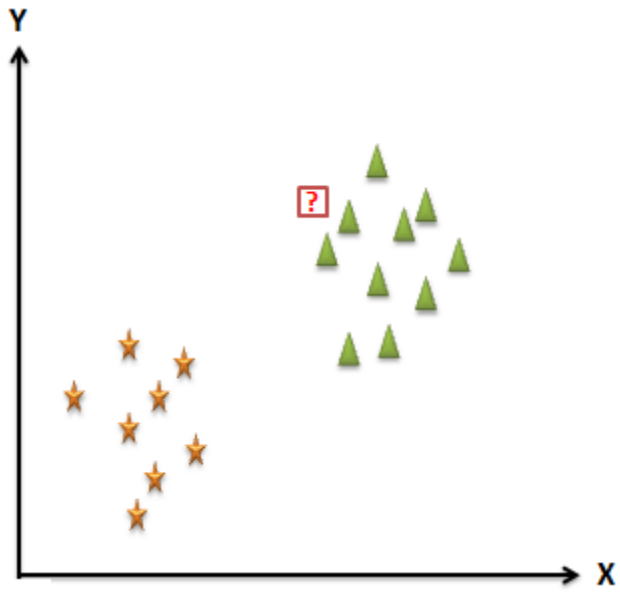
| Dimensionality | Required Sample Size |
|---|---|
| 1 | 4 |
| 2 | 19 |
| 5 | 786 |
| 7 | 10,700 |
| 10 | 842,000 |

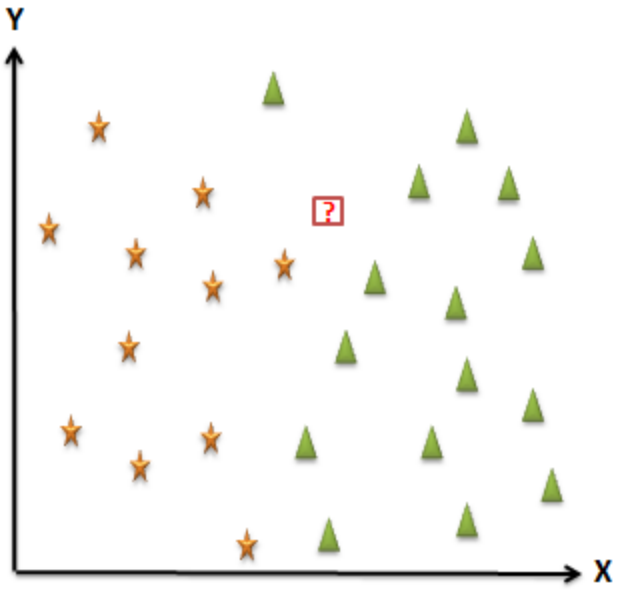
$$d_H\left(P(Y_+), P(Y_-)\right) = \sqrt{\sum_{i \in V}\left(\sqrt{P(Y_+ \mid X_i)} - \sqrt{P(Y_- \mid X_i)}\right)^2}$$

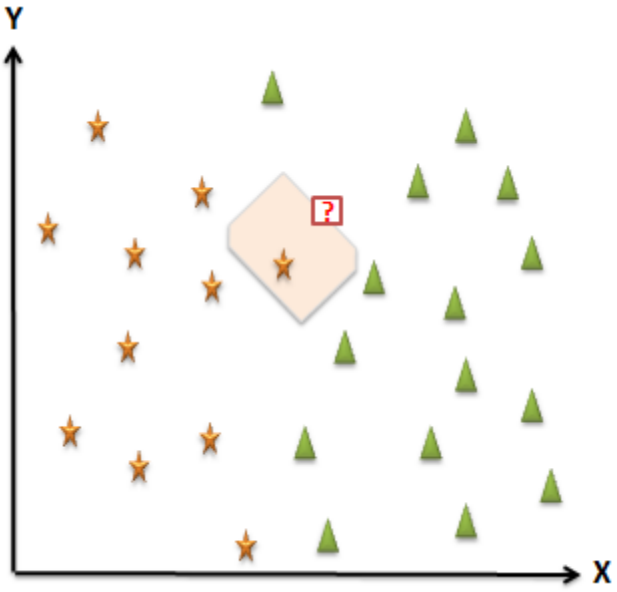
| | | |
|---|---|---|
| High | Low | 0 |
| High | High | 1 |

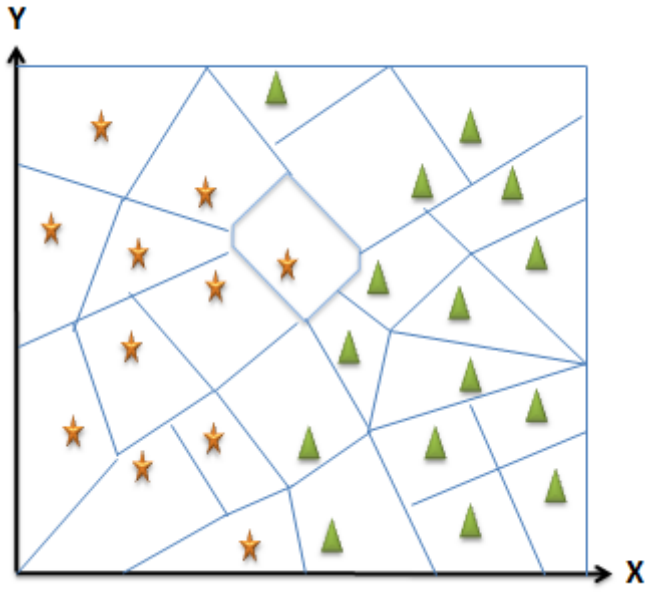

Machine Learning
has many
Learning models
categories
Deep learning based (ANN)
Regression analysis based
Clustering methods based
Ensemble methods based
Decision Trees based
Association Rules based
Dimensionality Reduction based
Instance based
Bayesian methods based
Kernel methods based

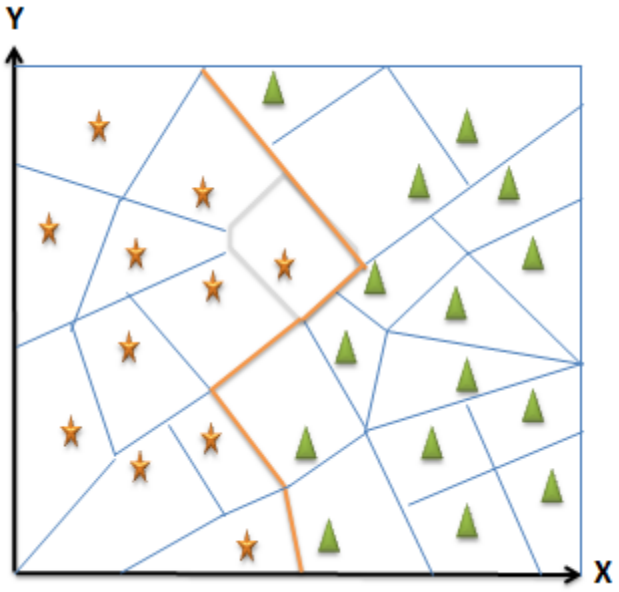

Instance based
algorithms
Radial basis functions
categoized under
eager learning
Case based reasoning
Nearest Neighbor
implementation
K Nearest Neighbor (KNN)
Locally Weighed Regression
categorized under
lazy learning

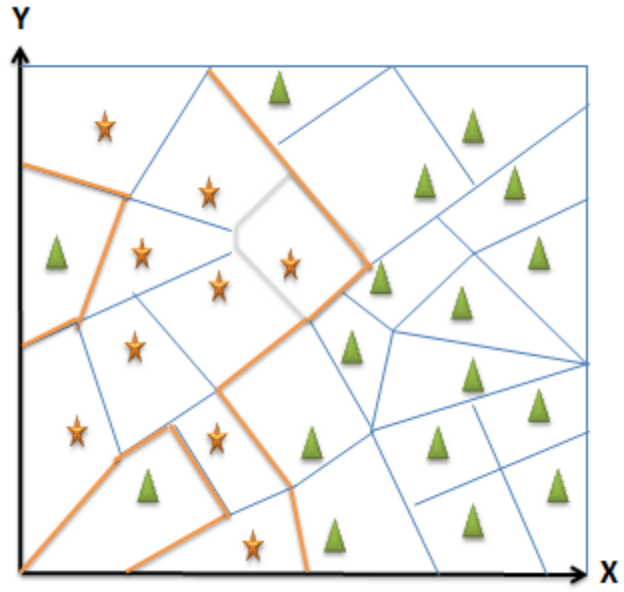

Y
?
X

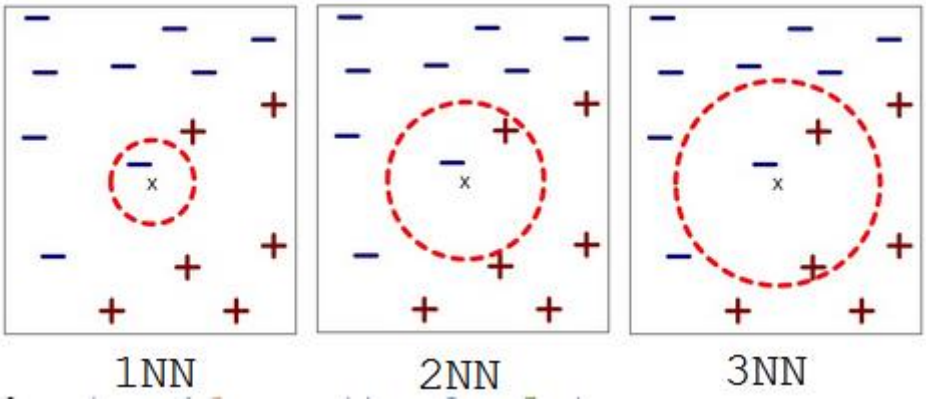

Y
?
X


Y
?
X


Y
X


Y
X

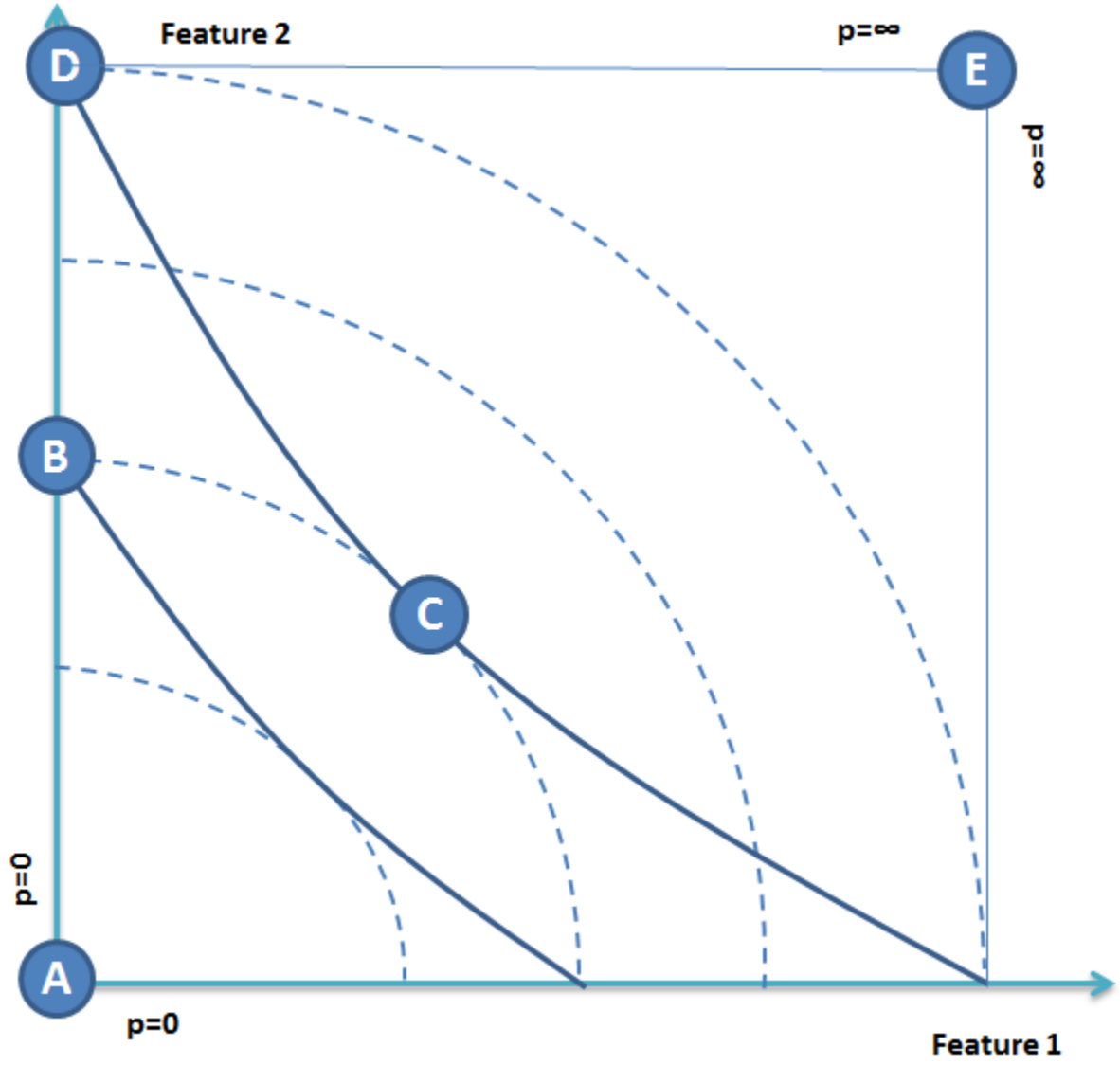

Y
X

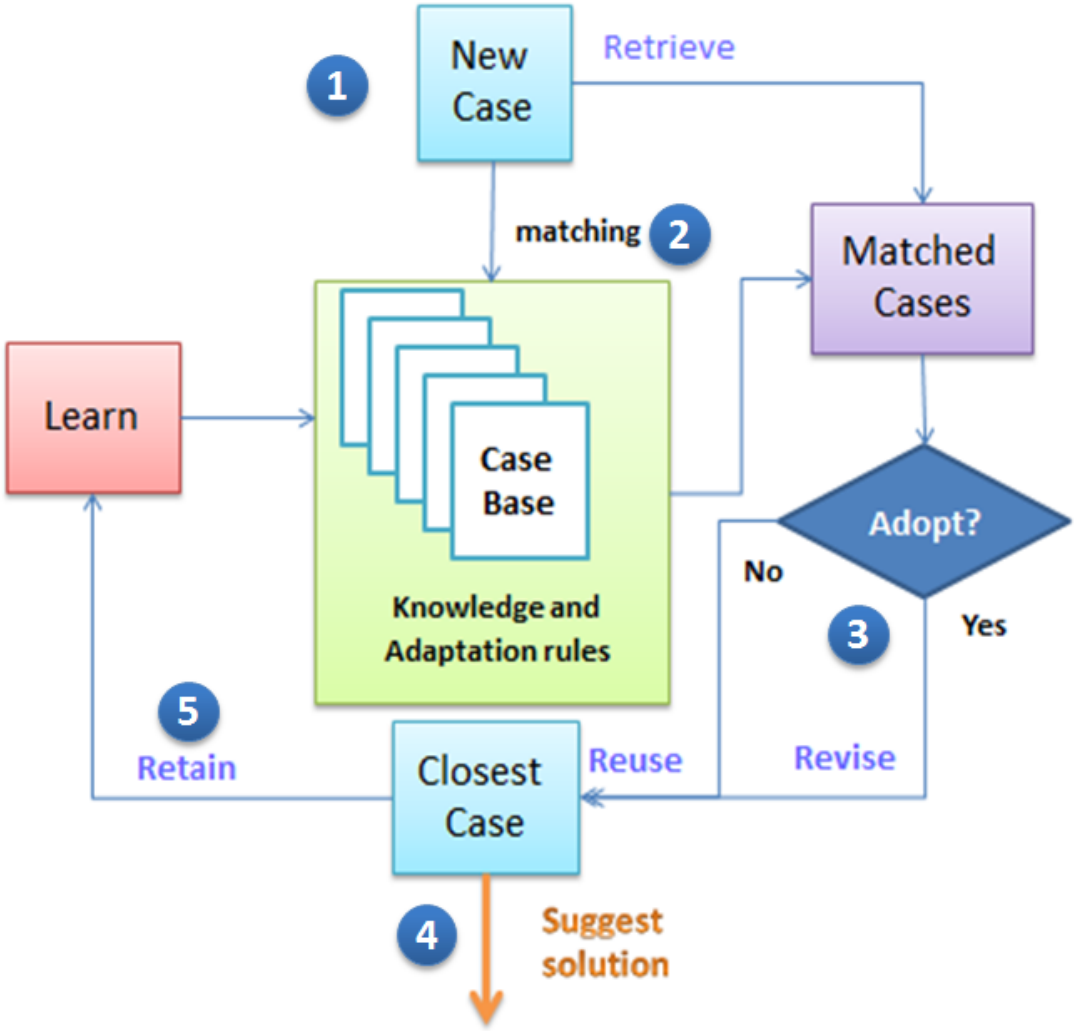

1NN
2NN
3NN

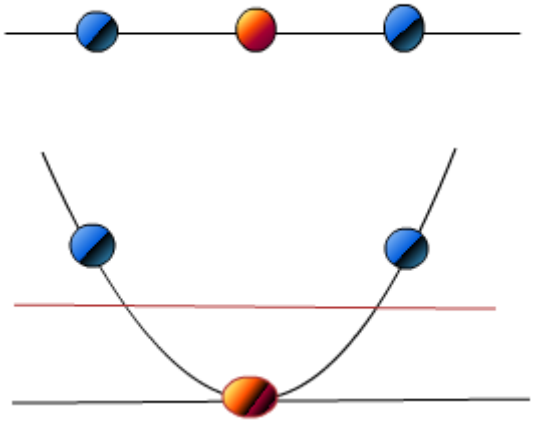
$$D(x,x') = \sqrt{\sum |x_d - x'_d|^2}$$

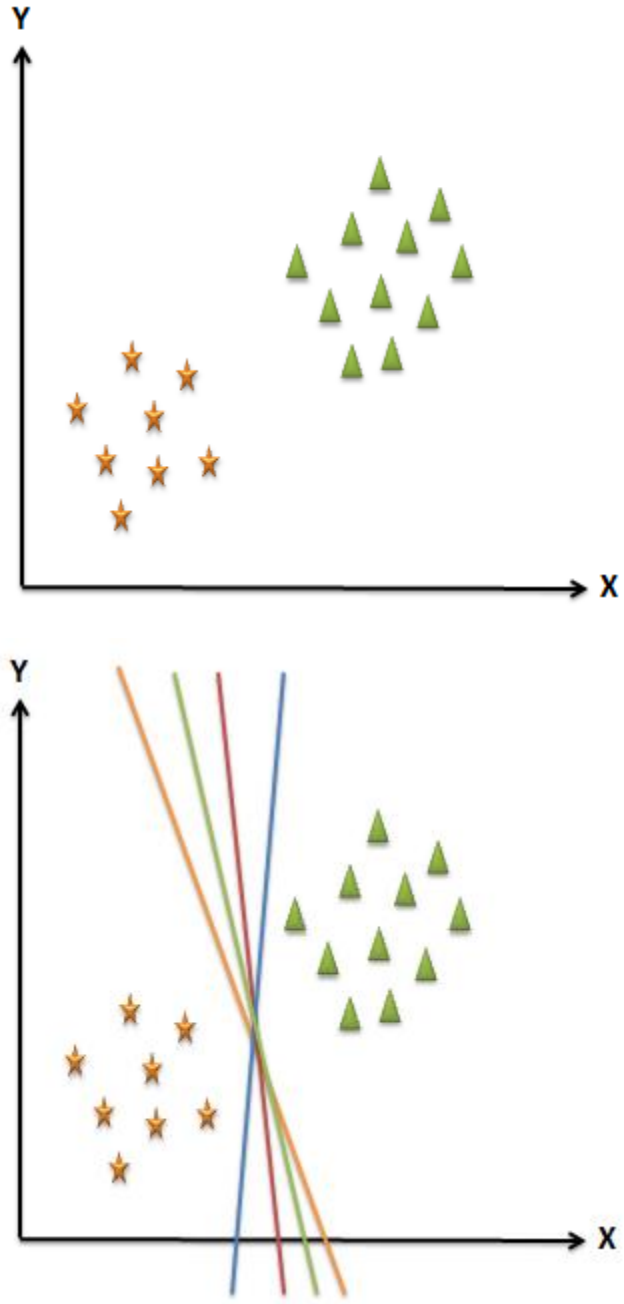
$$D(x,x') = \sum_d 1_{x_d \neq x'_d}$$

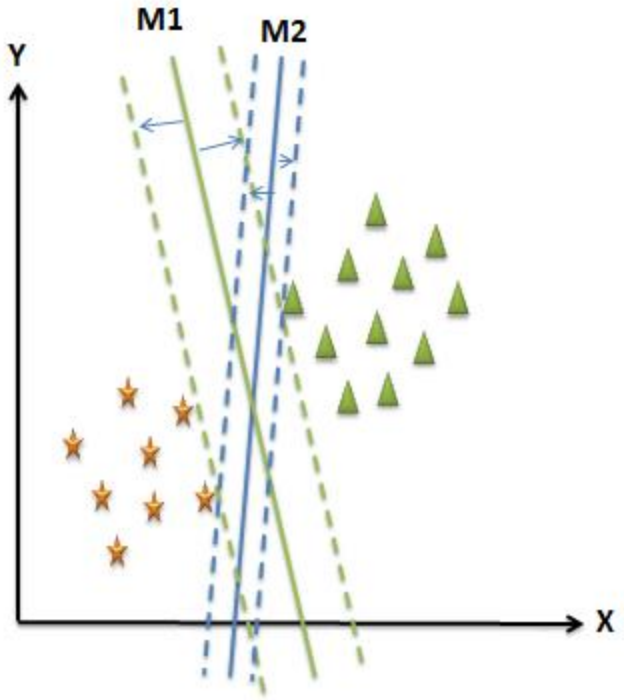
$$D(x,x') = \sqrt[p]{\sum |x_d - x'_d|^p}$$

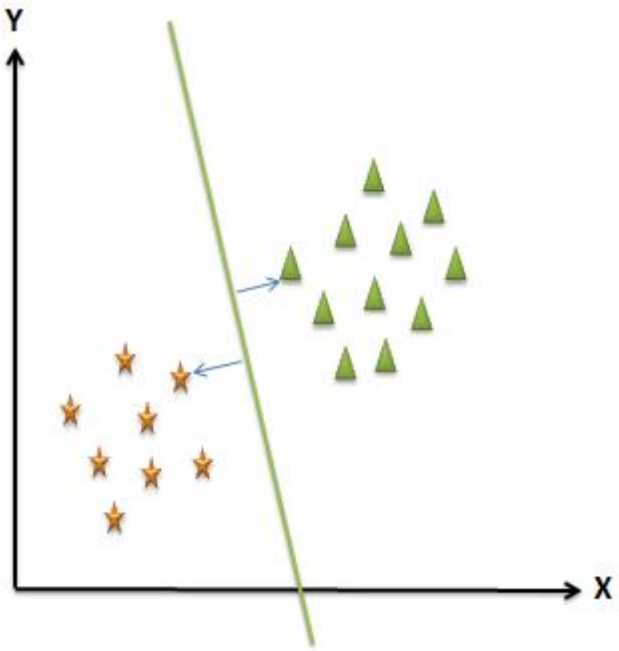



1
New Case
Retrieve
matching
2
Matched Cases
Learn
Case Base
Knowledge and Adaptation rules
Adopt?
No
Yes
3
5
Retain
Closest Case
Reuse
Revise
4
Suggest solution



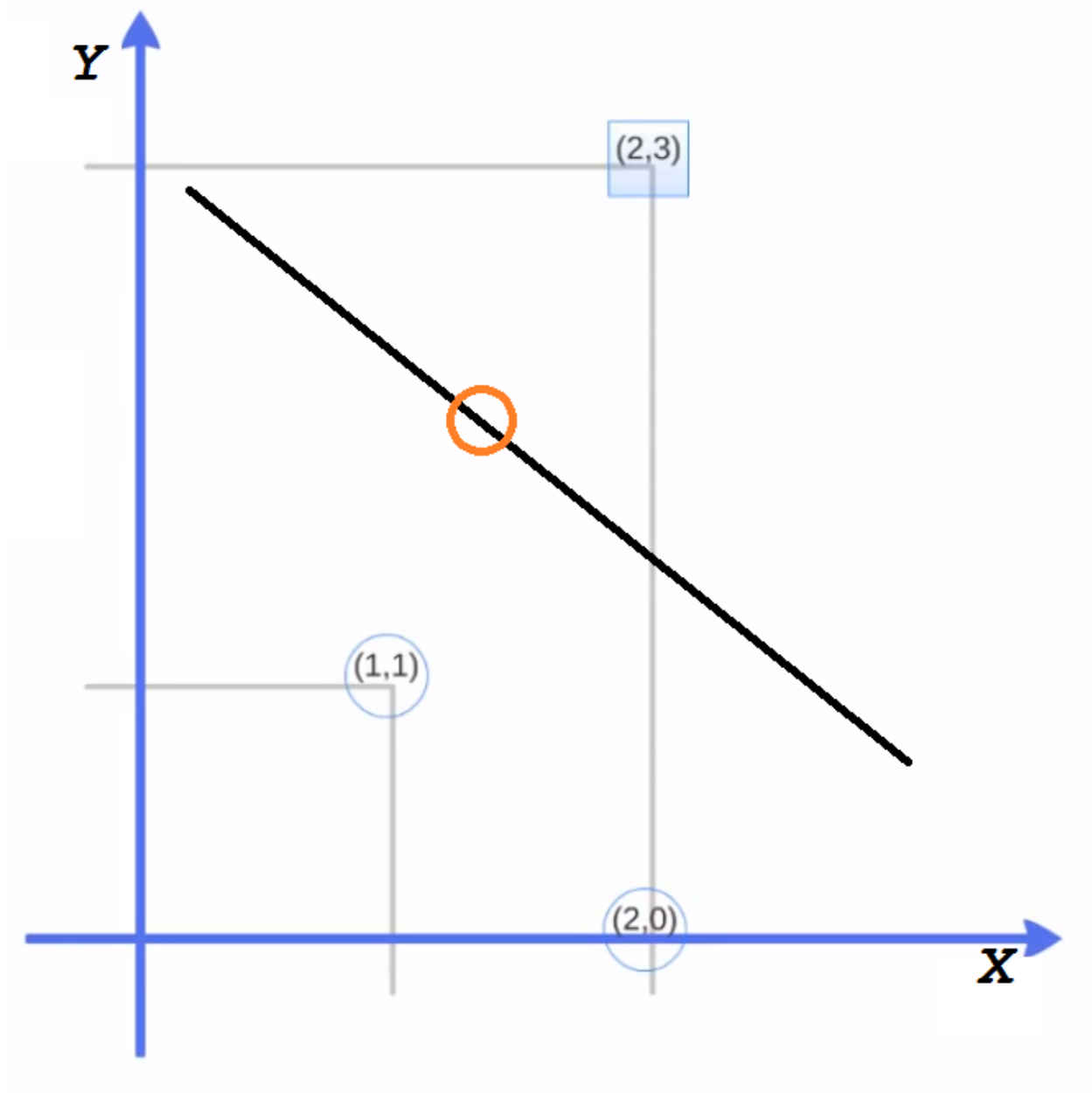

Y
X
Y
X


M1
M2
Y
X


Y
X

$$\vec{a} = \sum_{i=0}^{N} \lambda_i y_i \vec{x}_i$$

$$\sum_{i=0}^{N} \lambda_i y_i = 0$$




$$\omega_0 = 1 - 8a \quad 3a + 1 - 8a = -1$$

$$\vdots \quad 5a = 2$$

$$a = \frac{2}{5}$$

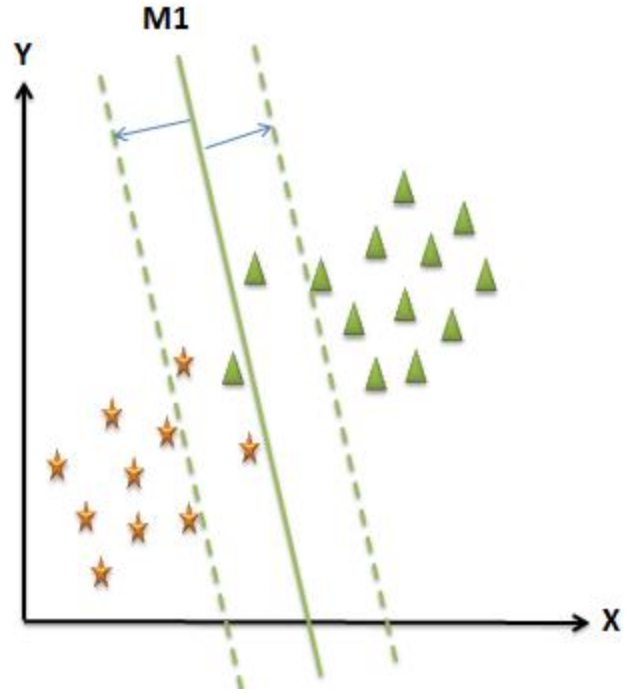
$$\omega_0 = 1 - 8\frac{2}{5} = \frac{5 - 16}{5}$$

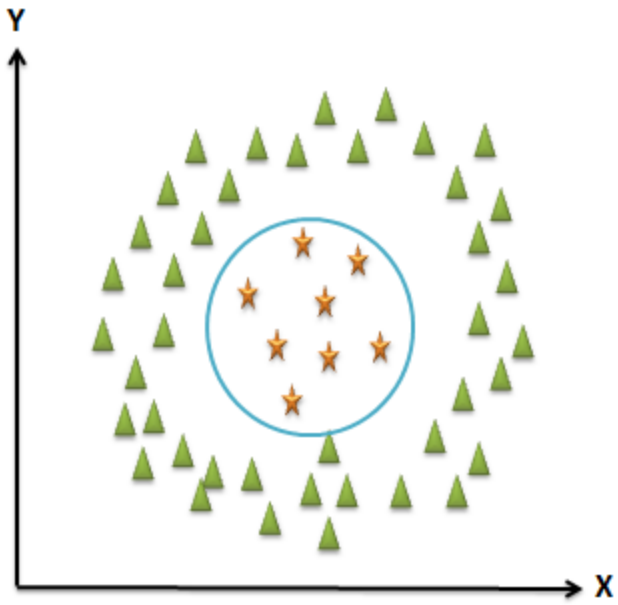
$$\omega_0 = -\frac{11}{5}$$

$$\vec{w} = \left( \frac{2}{5}, \frac{4}{5} \right)$$

$$g(\vec{x}) = \frac{2}{5} x_1 + \frac{4}{5} x_2 - \frac{11}{5}$$

$$g(\vec{x}) = x_1 + 2x_2 - 5.5$$



Minimize : $\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$

Subject to : $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i$

↓

Introduce slack variables $\xi_i \geq 0$

↓

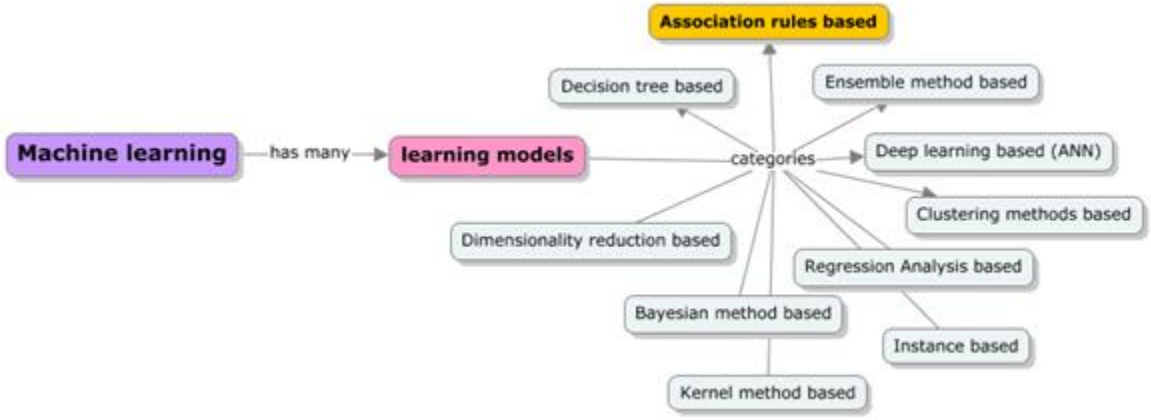
Minimize : $\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$

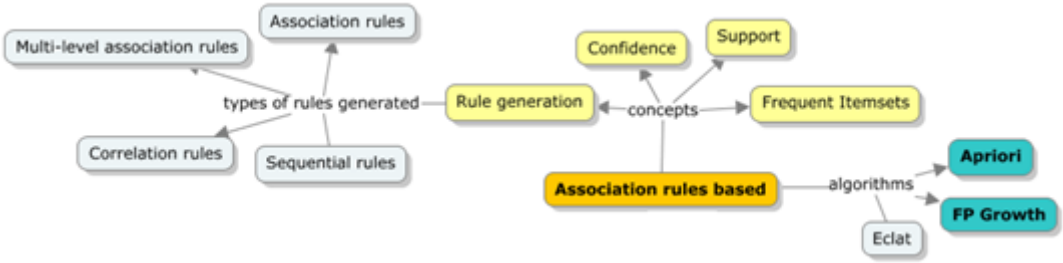
Subject to : $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i$

Polynomial: $K_p(\mathbf{X}, \mathbf{Y}) = (1 + \mathbf{X} \bullet \mathbf{Y})^p$

Radial Basis Function (RBF) or Gaussian: $K_r(\mathbf{X}, \mathbf{Y}) = e^{-\frac{1}{2\sigma^2}\|\mathbf{X}-\mathbf{Y}\|_2^2}$

Hyperbolic Tangent: $K_s(\mathbf{X}, \mathbf{Y}) = \tanh(\beta_0 \mathbf{X} \bullet \mathbf{Y} + \beta_1)$

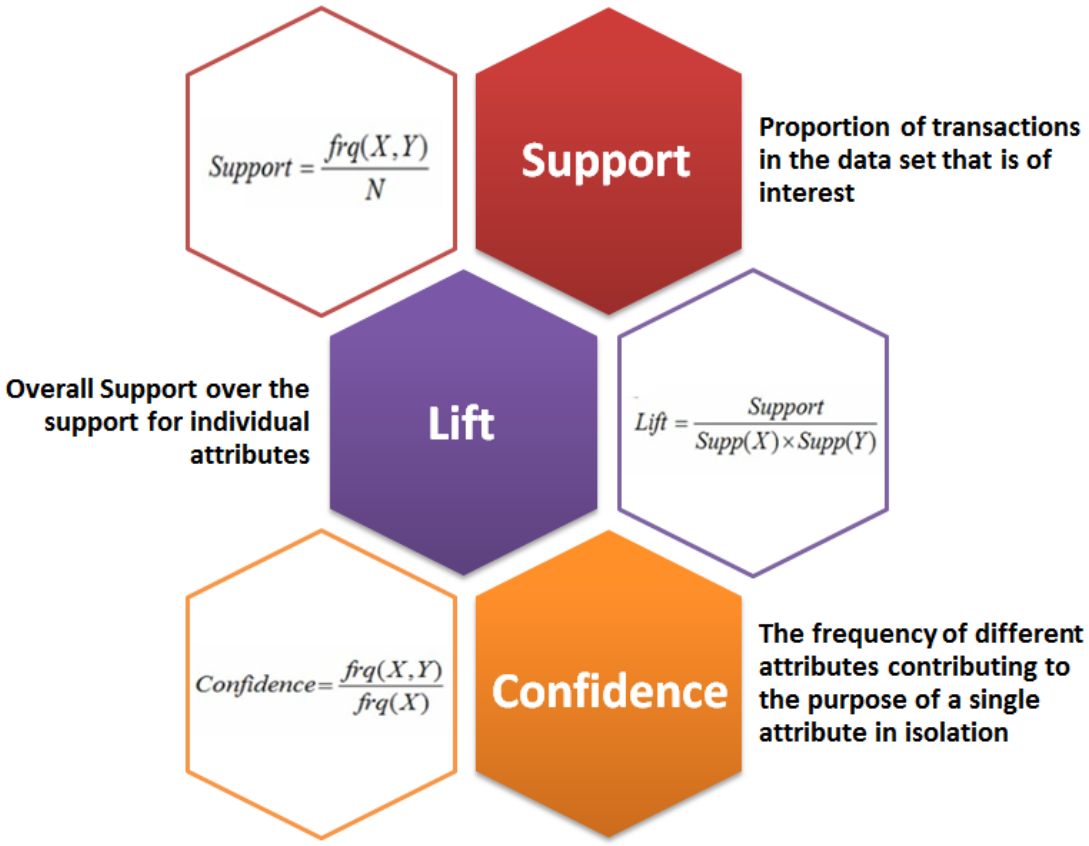


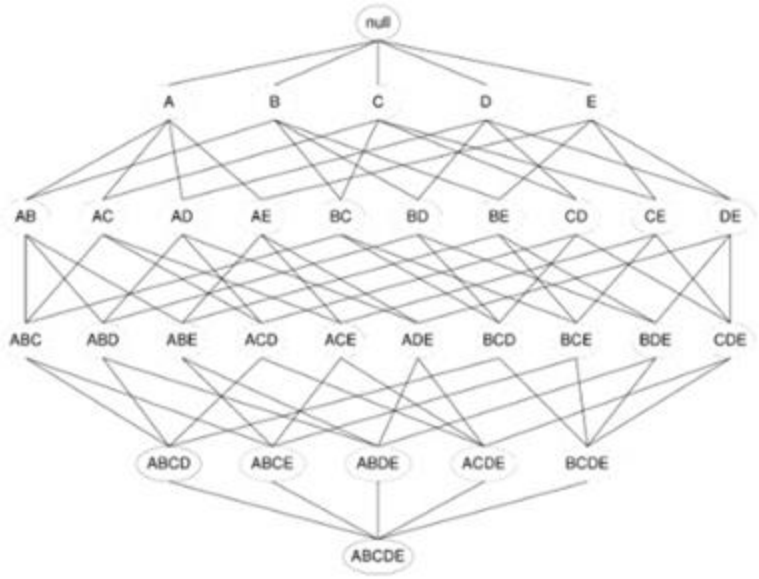

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

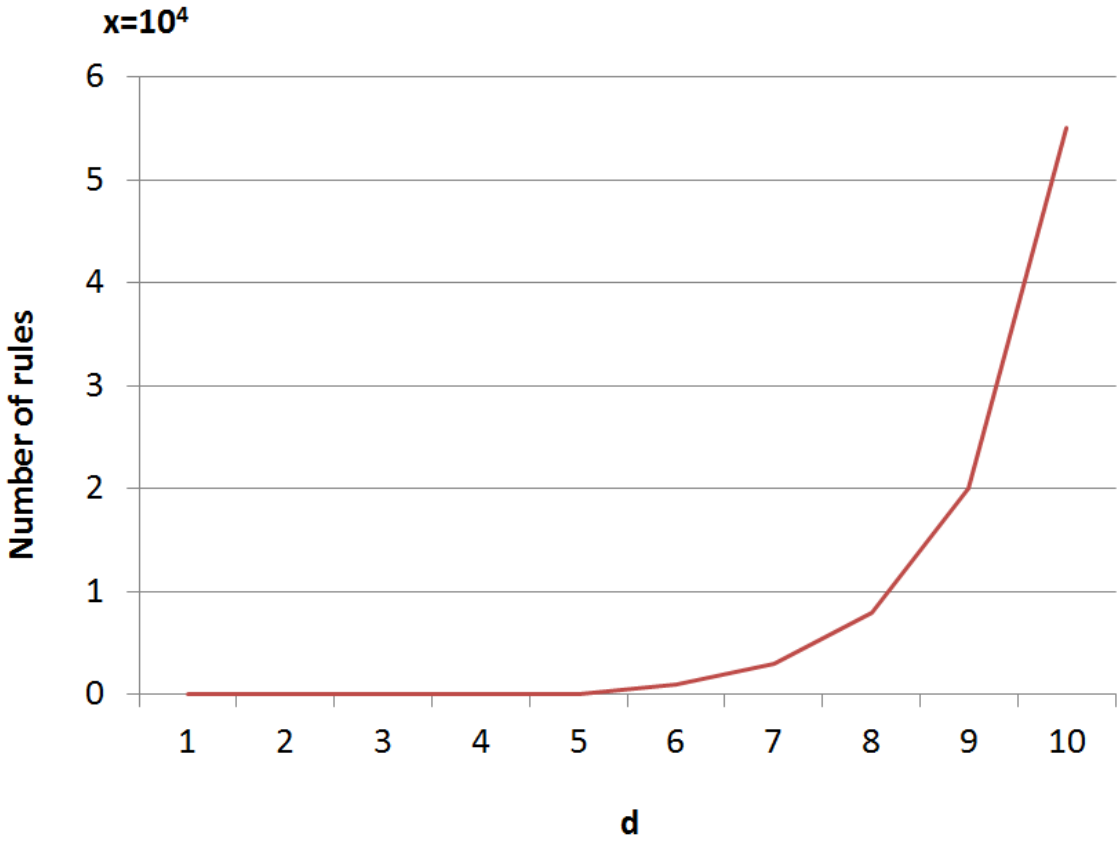
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$


$Support = \frac{frq(X,Y)}{N}$
Support
Proportion of transactions in the data set that is of interest
Overall Support over the support for individual attributes
Lift
$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$
$Confidence = \frac{frq(X,Y)}{frq(X)}$
Confidence
The frequency of different attributes contributing to the purpose of a single attribute in isolation


A B C
A C D
B C D
A D E
B C E

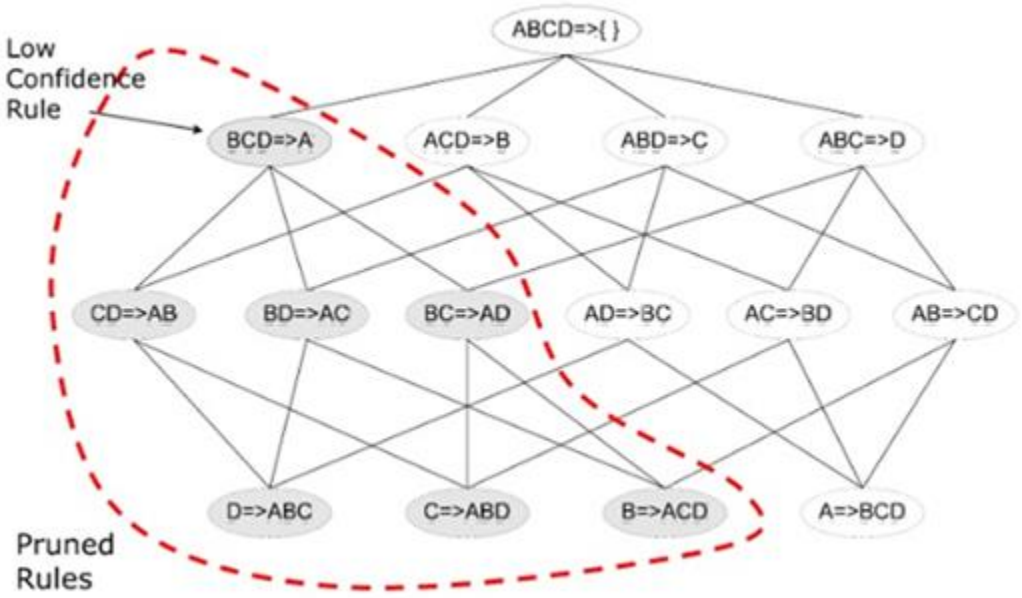


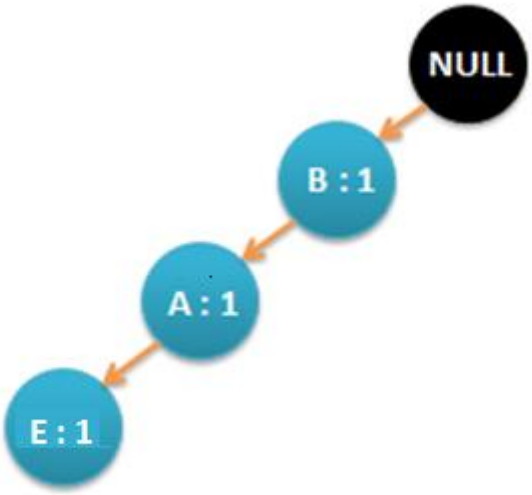

$$\sum_{k=1}^{d-1}\left[\binom{d}{k}\times\sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$=3^d-2^{d+1}+1$$

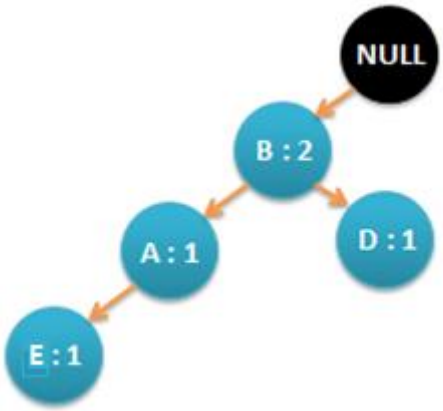


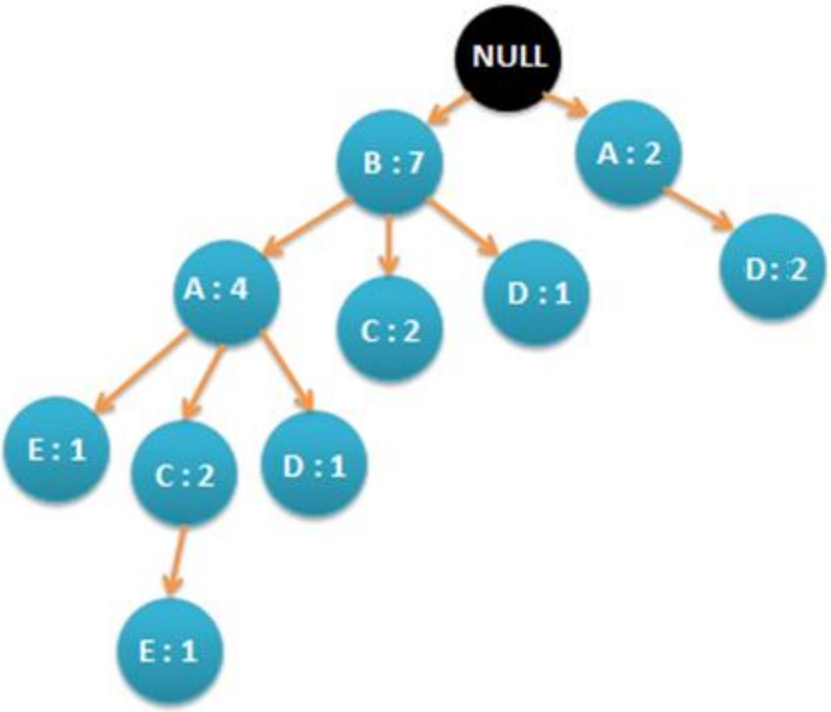
$C_k$: Candidate itemset of size k
$L_k$ : frequent itemset of size k

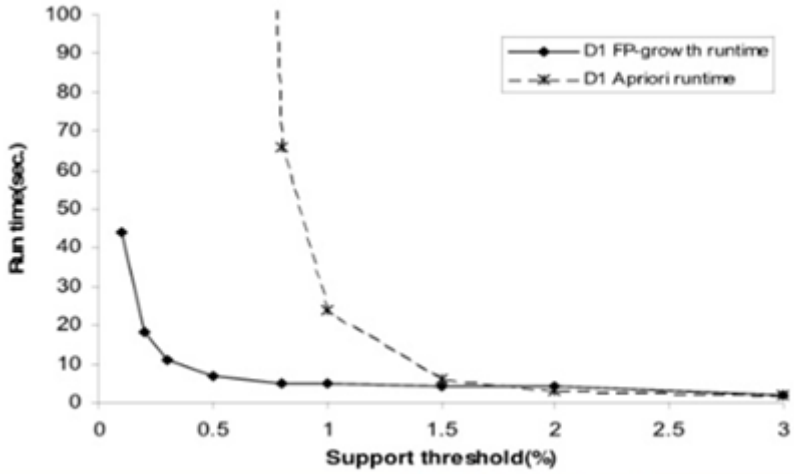
$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
  $C_{k+1}$ = candidates generated from $L_k$;
  for each transaction $t$ in database do
    increment the count of all candidates in $C_{k+1}$
    that are contained in $t$
  $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
  end
return $\cup_k L_k$;

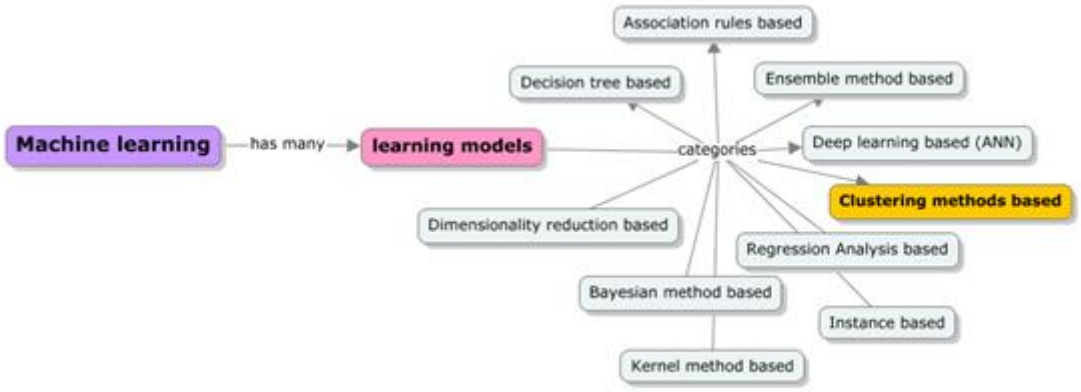


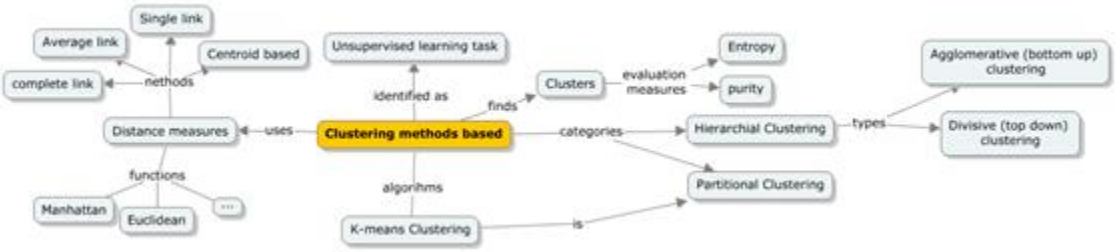



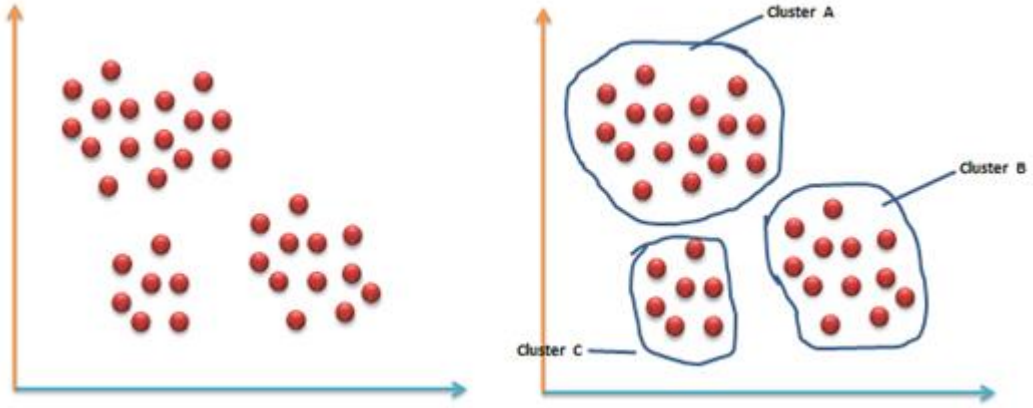

NULL
B : 2
A : 1
D : 1
E : 1

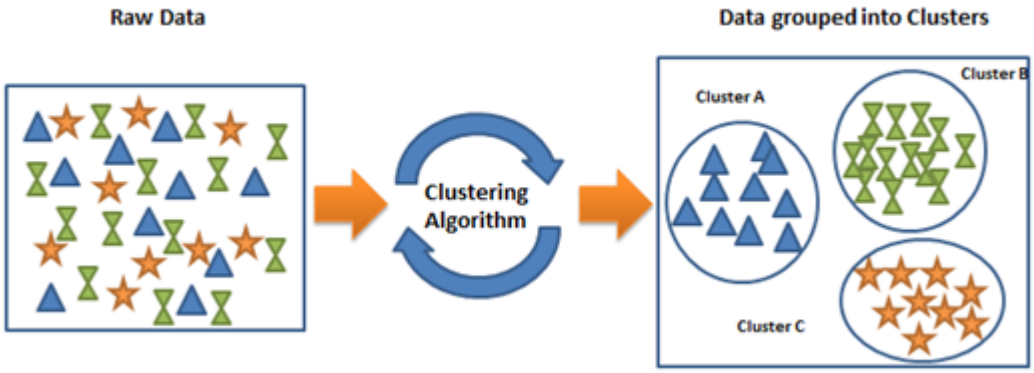

NULL
B : 7
A : 2
A : 4
C : 2
D : 1
D : 2
E : 1
C : 2
D : 1
E : 1

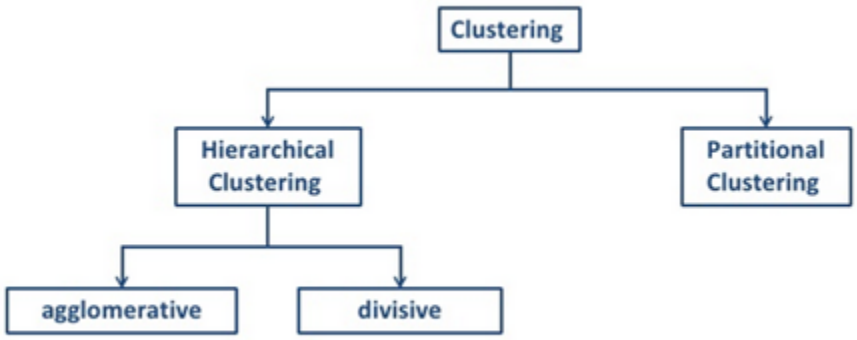

100
90
80
70
60
50
40
30
20
10
0
Run time(sec.)
0
0.5
1
1.5
2
2.5
3
Support threshold(%)
D1 FP-grow th runtime
D1 Apriori runtime

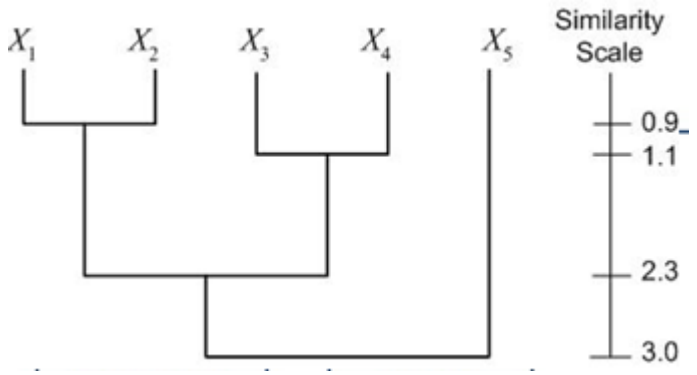

Machine learning
has many
learning models
categories
Association rules based
Decision tree based
Ensemble method based
Deep learning based (ANN)
Clustering methods based
Dimensionality reduction based
Regression Analysis based
Bayesian method based
Instance based
Kernel method based

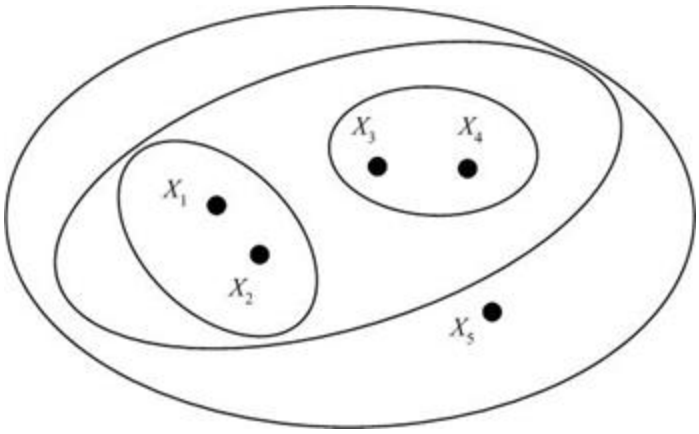

Single link
Average link
Centroid based
complete link
methods
Unsupervised learning task
identified as
Clusters
evaluation measures
Entropy
purity
finds
Agglomerative (bottom up) clustering
Distance measures
uses
Clustering methods based
categories
Hierarchial Clustering
types
Divisive (top down) clustering
functions
algorithms
Partitional Clustering
Manhattan
Euclidean
K-means Clustering
is

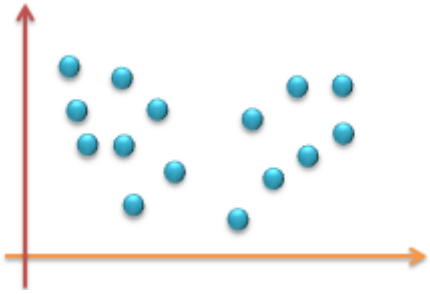

Cluster A
Cluster B
Cluster C

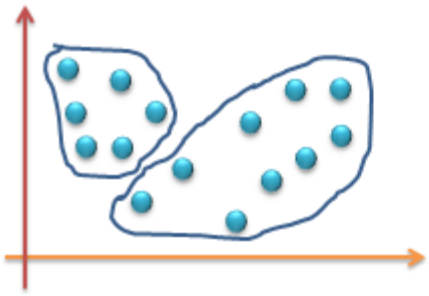

Raw Data
Data grouped into Clusters
Clustering Algorithm
Cluster A
Cluster B
Cluster C

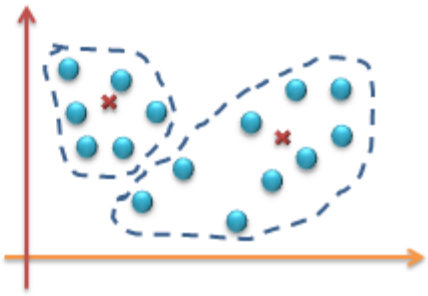

Clustering
Hierarchical Clustering
Partitional Clustering
agglomerative
divisive

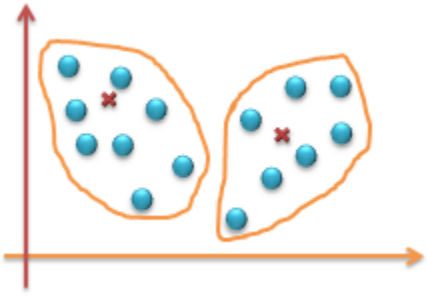

$X_1$
$X_2$
$X_3$
$X_4$
$X_5$
Similarity Scale
0.9
1.1
2.3
3.0

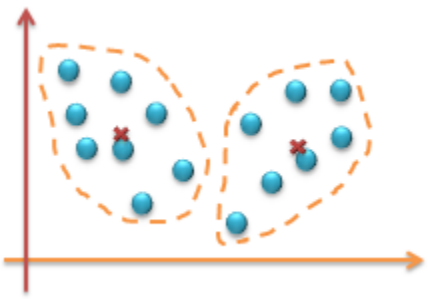

$X_1$
$X_2$
$X_3$
$X_4$
$X_5$

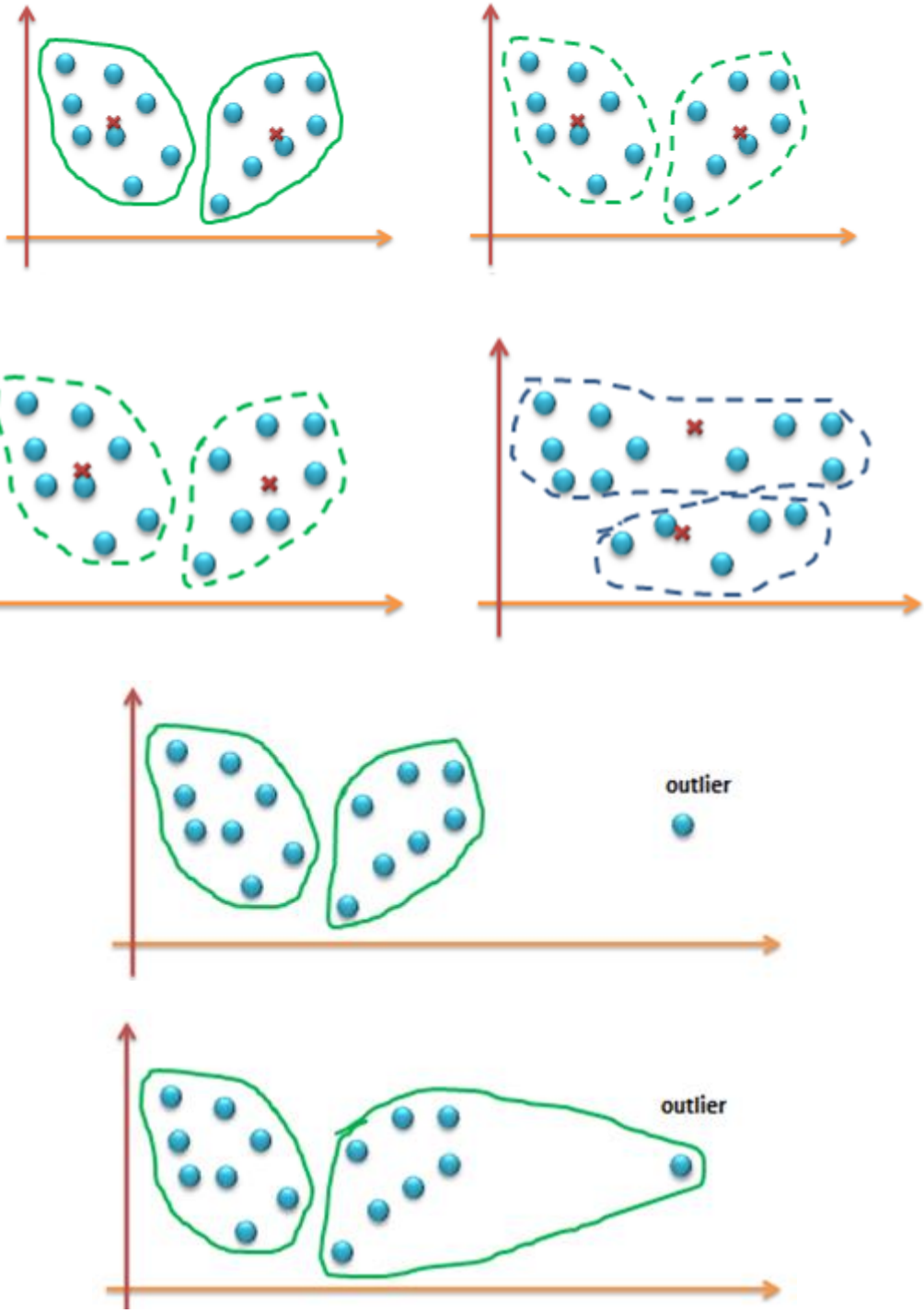


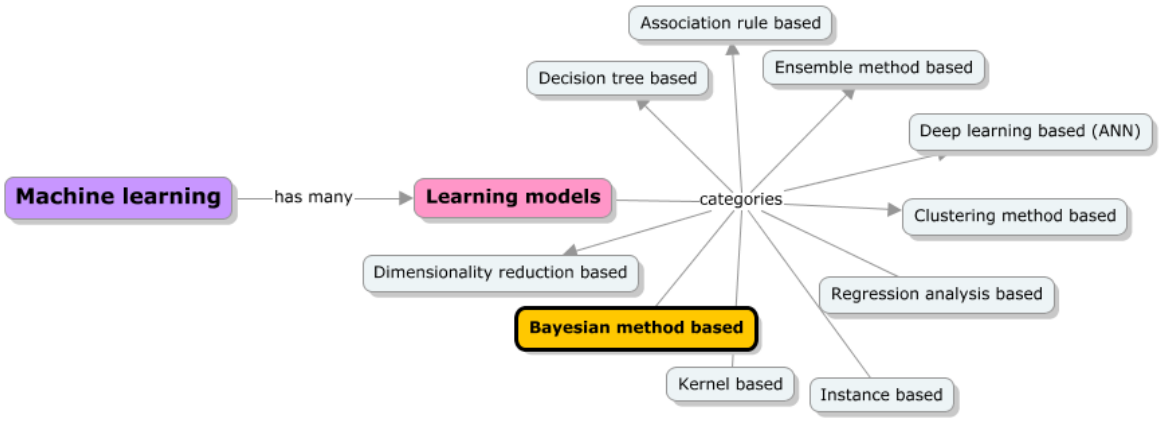




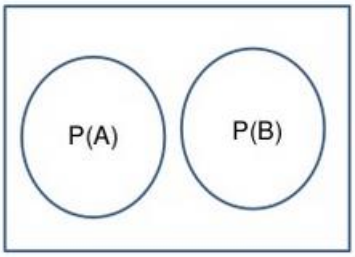


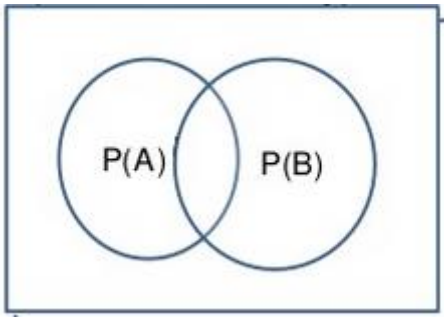



outlier
outlier




$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - \overline{x})^2}$$




$$\sum_x \sum_y P(X = x \; and \; Y = y) = 1$$

$$P(X = x) = \sum_y P(X = x, Y = y) = \sum_y P(X = x \mid Y = y) P(Y = y)$$

| X1 | X2 | Y (X1+X2) | Z (X1-X2) |
|---|---|---|---|
| 1 | 1 | 2 | 0 |
| 1 | 2 | 3 | -1 |
| 1 | 3 | 4 | -2 |
| 1 | 4 | 5 | -3 |
| 1 | 5 | 6 | -4 |
| 1 | 6 | 7 | -5 |
| 2 | 1 | 3 | 1 |
| 2 | 2 | 4 | 0 |
| 2 | 3 | 5 | -1 |
| 2 | 4 | 6 | -2 |
| 2 | 5 | 7 | -3 |
| 2 | 6 | 8 | -4 |
| 3 | 1 | 4 | 2 |
| 3 | 2 | 5 | 1 |
| 3 | 3 | 6 | 0 |
| 3 | 4 | 7 | -1 |
| 3 | 5 | 8 | -2 |
| 3 | 6 | 9 | -3 |
| 4 | 1 | 5 | 3 |
| 4 | 2 | 6 | 2 |
| 4 | 3 | 7 | 1 |
| 4 | 4 | 8 | 0 |
| 4 | 5 | 9 | -1 |
| 4 | 6 | 10 | -2 |
| 5 | 1 | 6 | 4 |
| 5 | 2 | 7 | 3 |
| 5 | 3 | 8 | 2 |
| 5 | 4 | 9 | 1 |
| 5 | 5 | 10 | 0 |
| 5 | 6 | 11 | -1 |
| 6 | 1 | 7 | 5 |
| 6 | 2 | 8 | 4 |
| 6 | 3 | 9 | 3 |
| 6 | 4 | 10 | 2 |
| 6 | 5 | 11 | 1 |
| 6 | 6 | 12 | 0 |

| | | Y | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Marginal Z |
| Z | -5 | 0 | 0 | 0 | 0 | 0 | 1/36 | 0 | 0 | 0 | 0 | 0 | 1/36 |
| | -4 | | | | | 1/36 | | 1/36 | | | | | 1/18 |
| | -3 | | | | 1/36 | | 1/36 | | 1/36 | | | | 1/12 |
| | -2 | | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | | 1/9 |
| | -1 | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | 5/36 |
| | 0 | 1/36 | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | 1/36 | 1/6 |
| | 1 | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | 5/36 |
| | 2 | | | 1/36 | | 1/36 | | 1/36 | | 1/36 | | | 1/9 |
| | 3 | | | | 1/36 | | 1/36 | | 1/36 | | | | 1/12 |
| | 4 | | | | | 1/36 | | 1/36 | | | | | 1/18 |
| | 5 | | | | | | 1/36 | | | | | | 1/36 |
| | | 1/36 | 1/18 | 1/12 | 1/9 | 5/36 | 1/6 | 5/36 | 1/9 | 1/12 | 1/18 | 1/36 | |

| **Discrete** | **Continuous** |
|---|---|
| Bernouli | Normal |
| Binomial | T distribution |
| Negative binomial | Gamma |
| Geometric | Chi Square |
| Poisson | Exponential |
| | Weibull |
| | F Distribution |

| Possible demand X | Number of days | Probability [P(X)] |
|---|---|---|
| 3 | 3 | 0.06 |
| 4 | 7 | 0.14 |
| 5 | 12 | 0.24 |
| 6 | 14 | 0.28 |
| 7 | 10 | 0.2 |
| 8 | 4 | 0.08 |

| Possible demand X | Probability [P(X)] | Weighted Value [XP(X)] |
|---|---|---|
| 3 | 0.06 | 0.18 |
| 4 | 0.14 | 0.56 |
| 5 | 0.24 | 1.2 |
| 6 | 0.28 | 1.68 |
| 7 | 0.2 | 1.4 |
| 8 | 0.08 | 0.64 |
| | **1** | **E(X) = 5.66** |

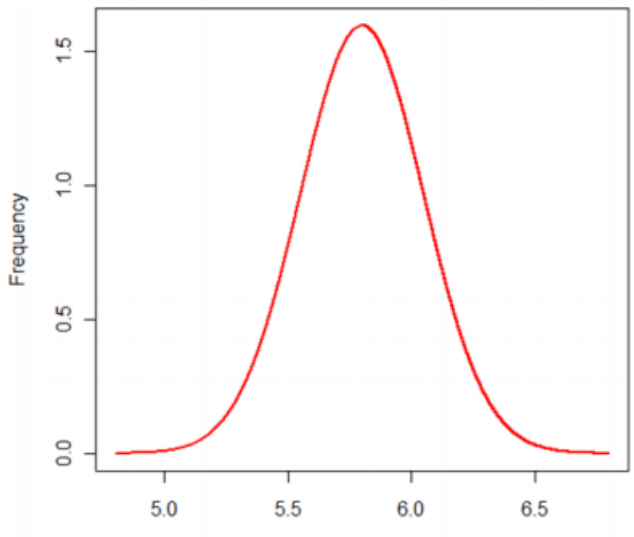
| Possible demand X | Probability [P(X)] | Weighted Value [XP(X)] | Squared demand (X2) | Weighted Square [X2P(X)] |
|---|---|---|---|---|
| 3 | 0.06 | 0.18 | 9 | 0.54 |
| 4 | 0.14 | 0.56 | 16 | 2.24 |
| 5 | 0.24 | 1.2 | 25 | 6 |
| 6 | 0.28 | 1.68 | 36 | 10.08 |
| 7 | 0.2 | 1.4 | 49 | 9.8 |
| 8 | 0.08 | 0.64 | 64 | 5.12 |
| | **1** | **E(X) = 5.66** | | **E(X2) = 33.78** |

$$\lim_{n\to\infty} c_r^n \left(\frac{\lambda}{n}\right)^r \left(1-\frac{\lambda}{n}\right)^{n-r}$$

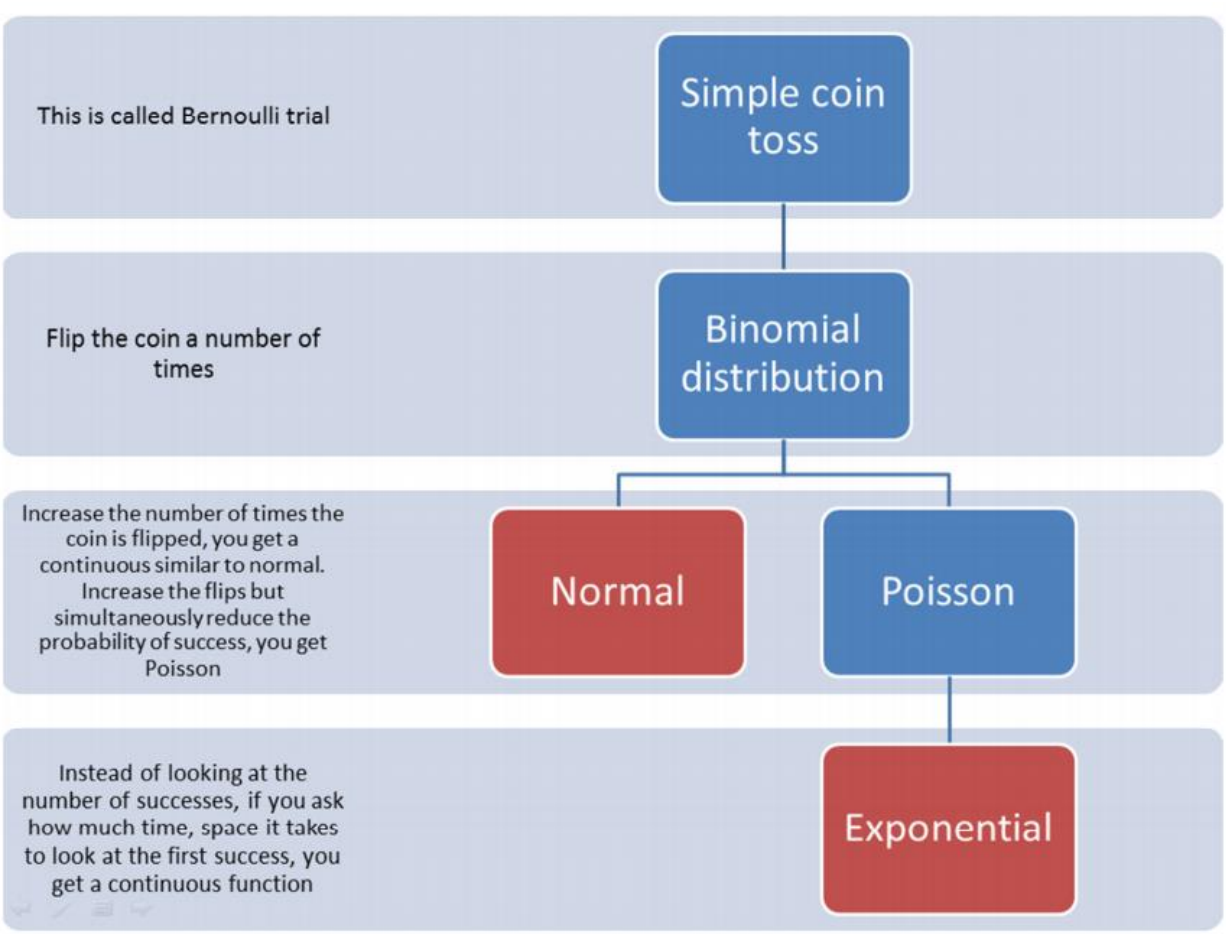
$$\lim_{n\to\infty} \frac{n!}{r!(n-r)!} \frac{\lambda^r}{n^r} \left(1-\frac{\lambda}{n}\right)^{n-r}$$

$$P(X=r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

$$P(X=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$$

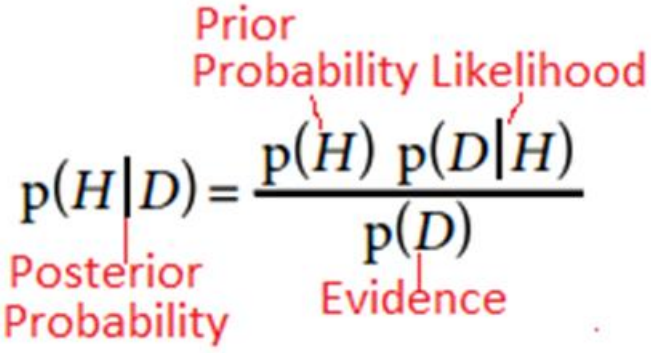



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



$$p(A \mid B) = \frac{p(A)\,p(B \mid A)}{p(B)}$$

$$p(H \mid D) = \frac{p(H)\,p(D \mid H)}{p(D)}$$




TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c / N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

APPLYMULTINOMIALNB($\mathbb{C}, V, prior, condprob, d$)
1 $W \leftarrow$ EXTRACTTOKENSFROMDOC($V, d$)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c] \mathrel{+}= \log condprob[t][c]$
6 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

```
TRAINBERNOULLINB(ℂ, 𝔻)
1  V ← EXTRACTVOCABULARY(𝔻)
2  N ← COUNTDOCS(𝔻)
3  for each c ∈ ℂ
4  do N_c ← COUNTDOCSINCLASS(𝔻, c)
5     prior[c] ← N_c / N
6     for each t ∈ V
7     do N_ct ← COUNTDOCSINCLASSCONTAININGTERM(𝔻, c, t)
8        condprob[t][c] ← (N_ct + 1) / (N_c + 2)
9  return V, prior, condprob

APPLYBERNOULLINB(ℂ, V, prior, condprob, d)
1  V_d ← EXTRACTTERMSFROMDOC(V, d)
2  for each c ∈ ℂ
3  do score[c] ← log prior[c]
4     for each t ∈ V
5     do if t ∈ V_d
6           then score[c] += log condprob[t][c]
7           else score[c] += log(1 − condprob[t][c])
8  return arg max_{c∈ℂ} score[c]
```

$$d = \left\langle t_1, \ldots, t_k, \ldots, t_{n_d} \right\rangle, t_k \in V$$

$$d = \left\langle e_1, \ldots, e_i, \ldots, e_M \right\rangle,$$

$$e_i \in \{0,1\}$$

$$\hat{P}(X = t \mid c)$$

$$\hat{P}(U_i = e \mid c)$$

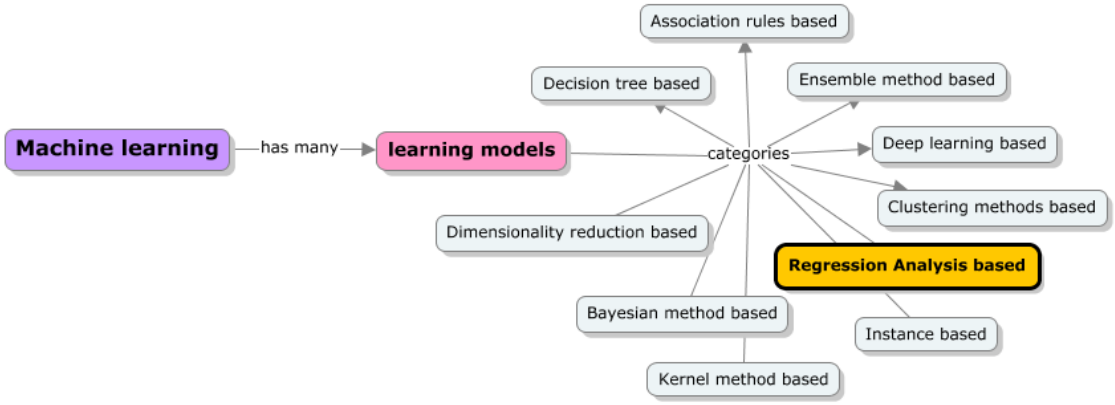
$$\hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(X = t_k \mid c)$$

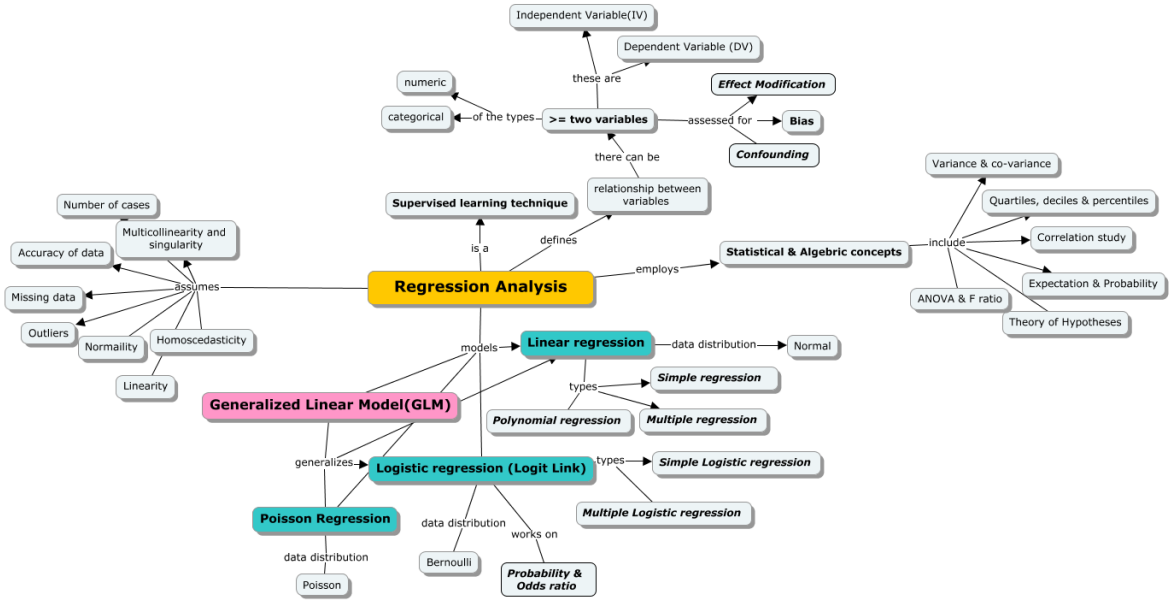
$$\hat{P}(c) \prod_{t_i \in V} \hat{P}(U_i = e_i \mid c)$$

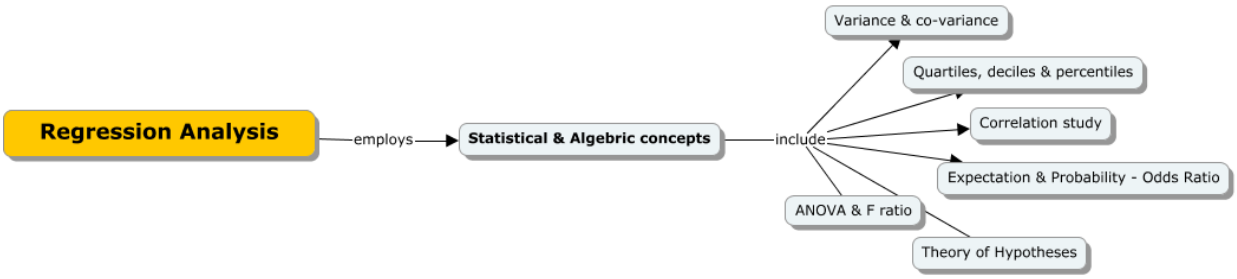
$$\hat{P}(X = the \mid c) \approx 0.05$$

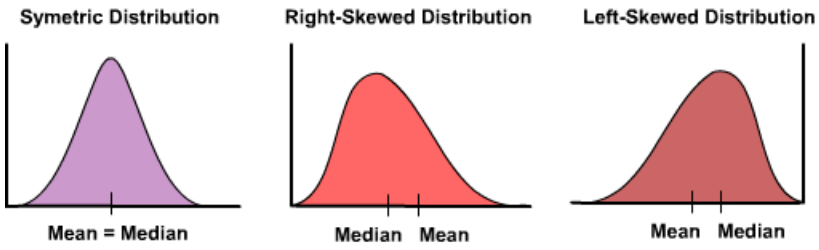
$$\hat{P}(U_{the} = 1 \mid c) \approx 1.0$$

## CHAPTER 10


Machine learning
has many
learning models
categories
Association rules based
Decision tree based
Ensemble method based
Deep learning based
Clustering methods based
Dimensionality reduction based
Regression Analysis based
Bayesian method based
Instance based
Kernel method based

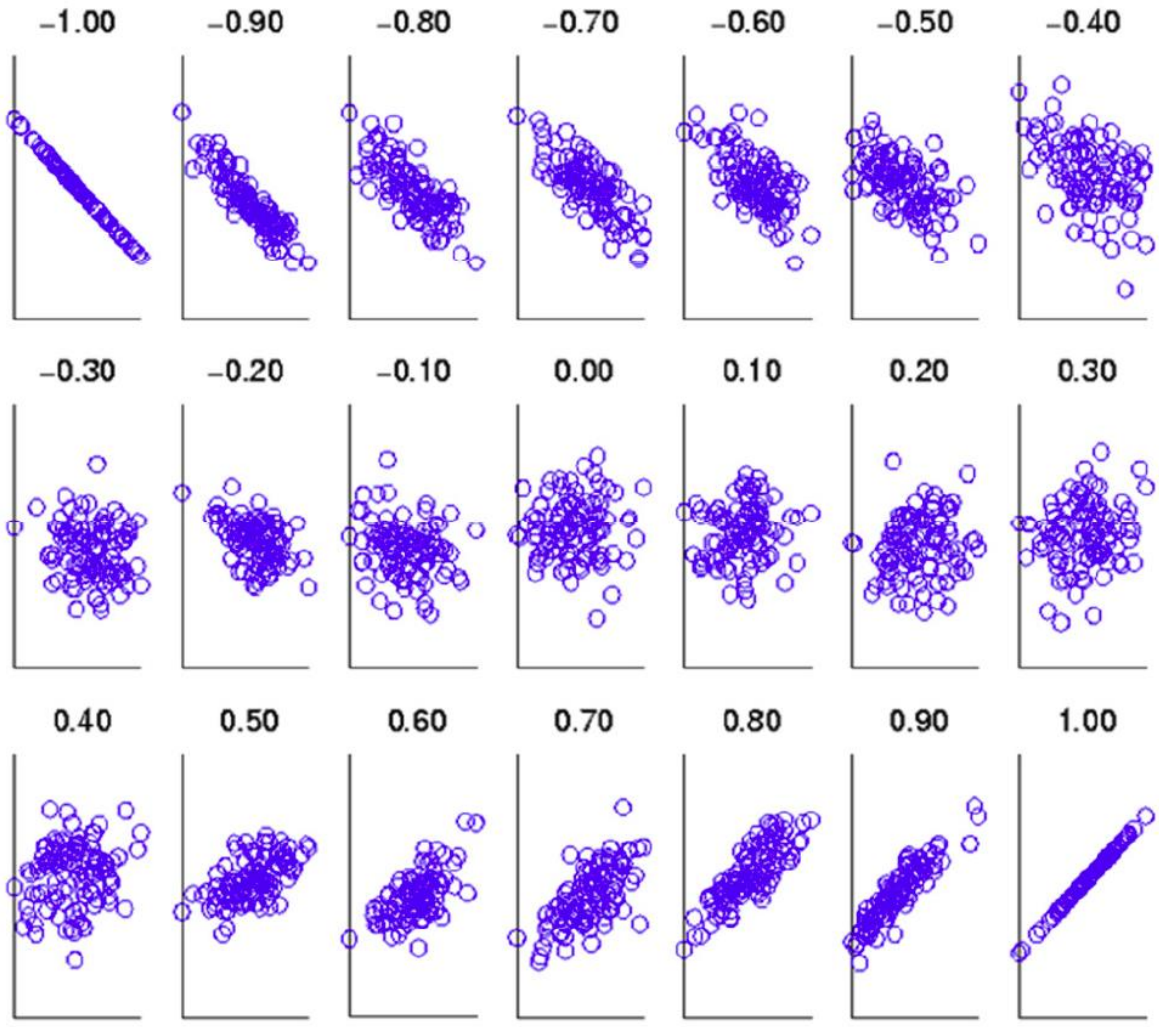

Independent Variable(IV)
Dependent Variable (DV)
these are
numeric
categorical
of the types
>= two variables
Effect Modification
assessed for
Bias
Confounding
there can be
relationship between variables
Supervised learning technique
is a
defines
Regression Analysis
employs
Statistical & Algebric concepts
include
Variance & co-variance
Quartiles, deciles & percentiles
Correlation study
Expectation & Probability
ANOVA & F ratio
Theory of Hypotheses
Number of cases
Multicollinearity and singularity
Accuracy of data
Missing data
Outliers
Normality
Homoscedasticity
Linearity
assumes
models
Linear regression
data distribution
Normal
types
Simple regression
Polynomial regression
Multiple regression
Generalized Linear Model(GLM)
generalizes
Logistic regression (Logit Link)
types
Simple Logistic regression
Multiple Logistic regression
Poisson Regression
data distribution
Poisson
data distribution
Bernoulli
works on
Probability & Odds ratio


Regression Analysis
employs
Statistical & Algebric concepts
include
Variance & co-variance
Quartiles, deciles & percentiles
Correlation study
Expectation & Probability - Odds Ratio
ANOVA & F ratio
Theory of Hypotheses




$$\rho_{xy}=\frac{\sum_1^n(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$Cor(x,y) = \rho_{xy}/\sigma_x\sigma_y$$




$$\text{Mean} = \mu = \int_a^b xP(x)dx$$

$$\text{Variance} = \sigma^2 = \int_a^b x^2P(x)dx$$

| Company A | Company B | Company C |
|---|---|---|
| 498.3 | 243.7 | 250.15 |
| 515.25 | 245.75 | 250.25 |
| 506.4 | 242.7 | 250.25 |
| 504.8 | 244.65 | 253.55 |
| 536.95 | 250.95 | 236.8 |
| 512.55 | 227.4 | 219.1 |
| 525.65 | 240.3 | 206.5 |
| 538.95 | 243.8 | 216.45 |
| 510.45 | 235.25 | 217.9 |
| 503 | 238.35 | 215.3 |
| 500.75 | 231.4 | 218.15 |
| 496.45 | 228.15 | 217.55 |
| 492.3 | 219.7 | 215.15 |
| 496.6 | 218 | 205.4 |

| Company A | Company B | Company C | A (returns) | B (returns) | C (returns) |
|---|---|---|---|---|---|
| 498.3 | 243.7 | 250.15 | | | |
| 515.25 | 245.75 | 250.25 | 0.0340157 | 0.008412 | 0.0004 |
| 506.4 | 242.7 | 250.25 | -0.017176 | -0.012411 | 0 |
| 504.8 | 244.65 | 253.55 | -0.00316 | 0.0080346 | 0.013187 |
| 536.95 | 250.95 | 236.8 | 0.0636886 | 0.0257511 | -0.06606 |
| 512.55 | 227.4 | 219.1 | -0.045442 | -0.0938434 | -0.07475 |
| 525.65 | 240.3 | 206.5 | 0.0255585 | 0.0567282 | -0.05751 |
| 538.95 | 243.8 | 216.45 | 0.025302 | 0.0145651 | 0.048184 |
| 510.45 | 235.25 | 217.9 | -0.052881 | -0.0350697 | 0.006699 |
| 503 | 238.35 | 215.3 | -0.014595 | 0.0131775 | -0.01193 |
| 500.75 | 231.4 | 218.15 | -0.004473 | -0.0291588 | 0.013237 |
| 496.45 | 228.15 | 217.55 | -0.008587 | -0.0140449 | -0.00275 |
| 492.3 | 219.7 | 215.15 | -0.008359 | -0.037037 | -0.01103 |
| 496.6 | 218 | 205.4 | 0.0087345 | -0.0077378 | -0.04532 |

| | A (returns) | B (returns) | C (returns) |
|---|---|---|---|
| | | | |
| | 0.0340157 | 0.008412 | 0.0004 |
| | -0.017176 | -0.012411 | 0 |
| | -0.00316 | 0.0080346 | 0.013187 |
| | 0.0636886 | 0.0257511 | -0.06606 |
| | -0.045442 | -0.0938434 | -0.07475 |
| | 0.0255585 | 0.0567282 | -0.05751 |
| | 0.025302 | 0.0145651 | 0.048184 |
| | -0.052881 | -0.0350697 | 0.006699 |
| | -0.014595 | 0.0131775 | -0.01193 |
| | -0.004473 | -0.0291588 | 0.013237 |
| | -0.008587 | -0.0140449 | -0.00275 |
| | -0.008359 | -0.037037 | -0.01103 |
| | 0.0087345 | -0.0077378 | -0.04532 |
| **Sum** | 0.0026265 | -0.1026342 | -0.18764 |
| **Mean** | 0.000202 | -0.0078949 | -0.01443 |

| | A (returns) | B (returns) | C (returns) |
|---|---|---|---|
| | -0.05288 | -0.09384 | -0.07475 |
| | -0.04544 | -0.03704 | -0.06606 |
| | -0.01718 | -0.03507 | -0.05751 |
| | -0.01459 | -0.02916 | -0.04532 |
| | -0.00859 | -0.01404 | -0.01193 |
| | -0.00836 | -0.01241 | -0.01103 |
| **Median** | -0.00447 | -0.00774 | -0.00275 |
| | -0.00316 | 0.008035 | 0 |
| | 0.008735 | 0.008412 | 0.0004 |
| | 0.025302 | 0.013177 | 0.006699 |
| | 0.025558 | 0.014565 | 0.013187 |
| | 0.034016 | 0.025751 | 0.013237 |
| | 0.063689 | 0.056728 | 0.048184 |

| A (returns) | B (returns) | C (returns) | Cov(A,B) | Cov(B,C) |
|---|---|---|---|---|
| -0.0528806 | -0.093843 | -0.07475 | 0.004562 | 0.005184 |
| -0.0454419 | -0.103097 | 0.00133 | 0.001505 | |
| -0.0171761 | -0.03507 | -0.05751 | 0.000472 | 0.001171 |
| -0.014595 | -0.029159 | -0.04532 | 0.000315 | 0.000657 |
| -0.0085871 | -0.014045 | -0.01193 | 5.41E-05 | -1.50E-05 |
| -0.0083594 | -0.012411 | -0.01103 | 3.87E-05 | -1.50E-05 |
| **A (returns)** | **B (returns)** | **C (returns)** | **Cov(A,B)** | **Cov(B,C)** |
| -0.0044732 | -0.007738 | -0.00275 | -7.30E-07 | -1.84E-06 |
| -0.0031596 | 0.0080346 | 0 | -5.40E-05 | 0.00023 |
| 0.0087345 | 0.008412 | 0.0004 | 0.000139 | 0.000242 |
| 0.025302 | 0.0131775 | 0.006699 | 0.000529 | 0.000445 |
| 0.0255585 | 0.0145651 | 0.013187 | 0.00057 | 0.00062 |
| 0.0340157 | 0.0257511 | 0.013237 | 0.001138 | 0.000931 |
| 0.0636886 | 0.0567282 | 0.048184 | 0.004103 | 0.004047 |
| | | **Covariance** | 0.001015 | 0.001154 |
| | | **Correlation** | 0.939734 | 0.942151 |

$$E\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i E(x_i)$$

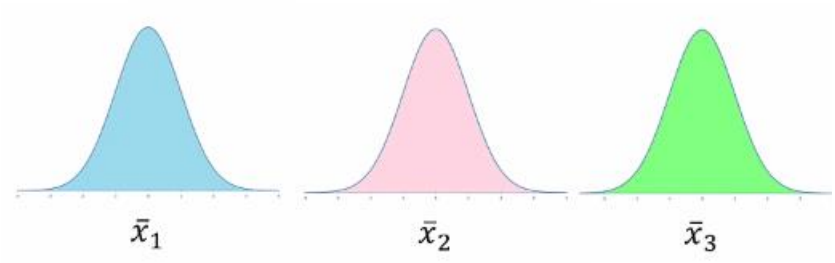
$$v \sum a_i X_i = \sum a_i^2 \, v\,(x_i) + 2 \sum\sum a_i a_j cov\,(x_i, x_j)$$

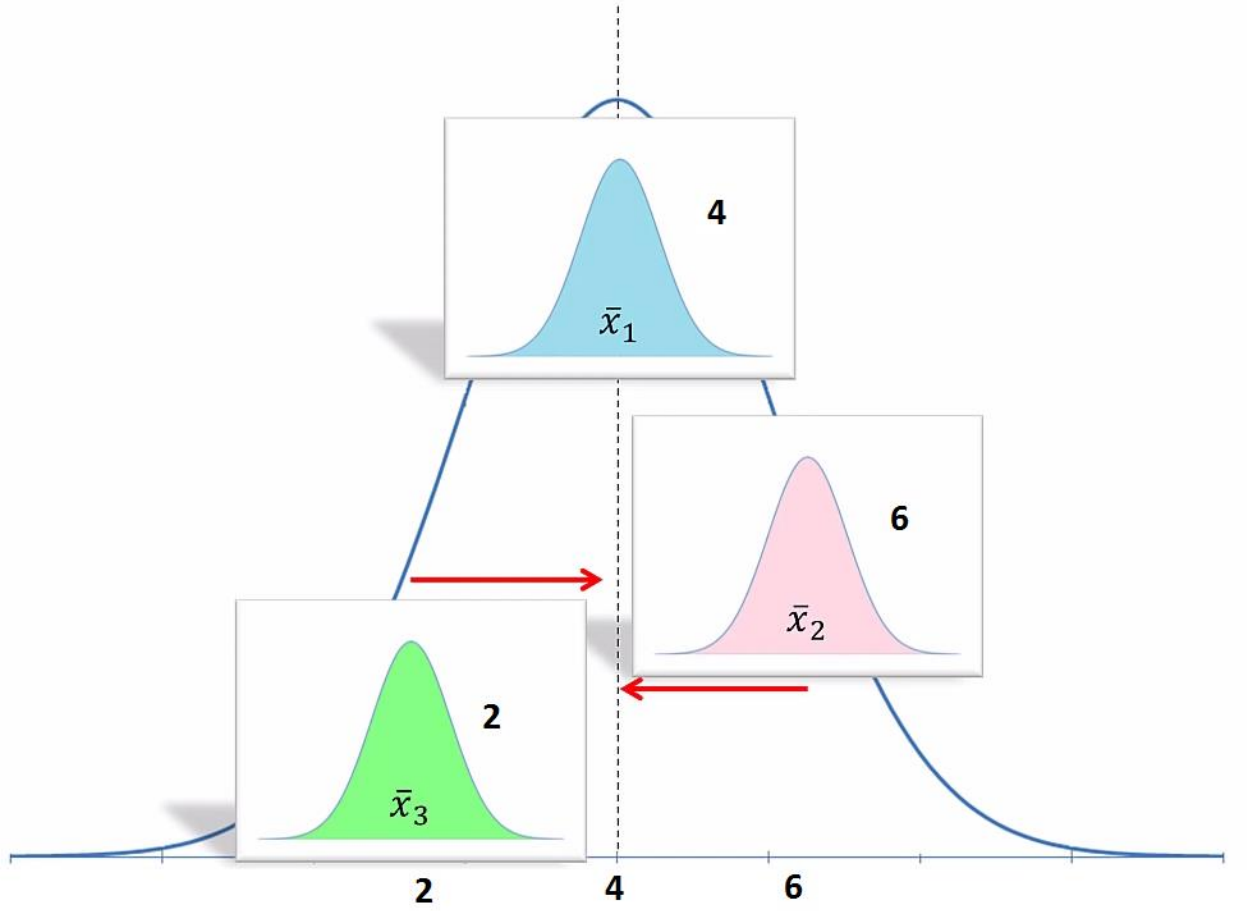
| | Gold | IT | Bank |
|---|---|---|---|
| **Returns** | 15 | 25 | 17 |
| **SD** | 5 | 15 | 10 |

| | Gold | IT | Bank |
|---|---|---|---|
| **Gold** | 1 | -1 | -0.5 |
| **IT** | -1 | 1 | 0.5 |
| **Bank** | -0.5 | 0.5 | 1 |

| | | | | |
|---|---|---|---|---|
| **Portfolio1** | 0 | 1 | 0 | 25 |
| **Portfolio2** | 0 | 0.9 | 0.1 | 24.2 |
| **Portfolio3** | 0 | 0.8 | 0.2 | 23.4 |
| **Portfolio4** | 0 | 0.7 | 0.3 | 22.6 |
| **Portfolio5** | 0 | 0.6 | 0.4 | 21.8 |

| wg | wi | wb | SD |
|---|---|---|---|
| 0 | 1 | 0 | 15 |
| 0 | 0.9 | 0.1 | 14.03 |
| 0 | 0.8 | 0.2 | 13.11 |
| 0 | 0.7 | 0.3 | 12.28 |
| 0 | 0.6 | 0.4 | 11.53 |

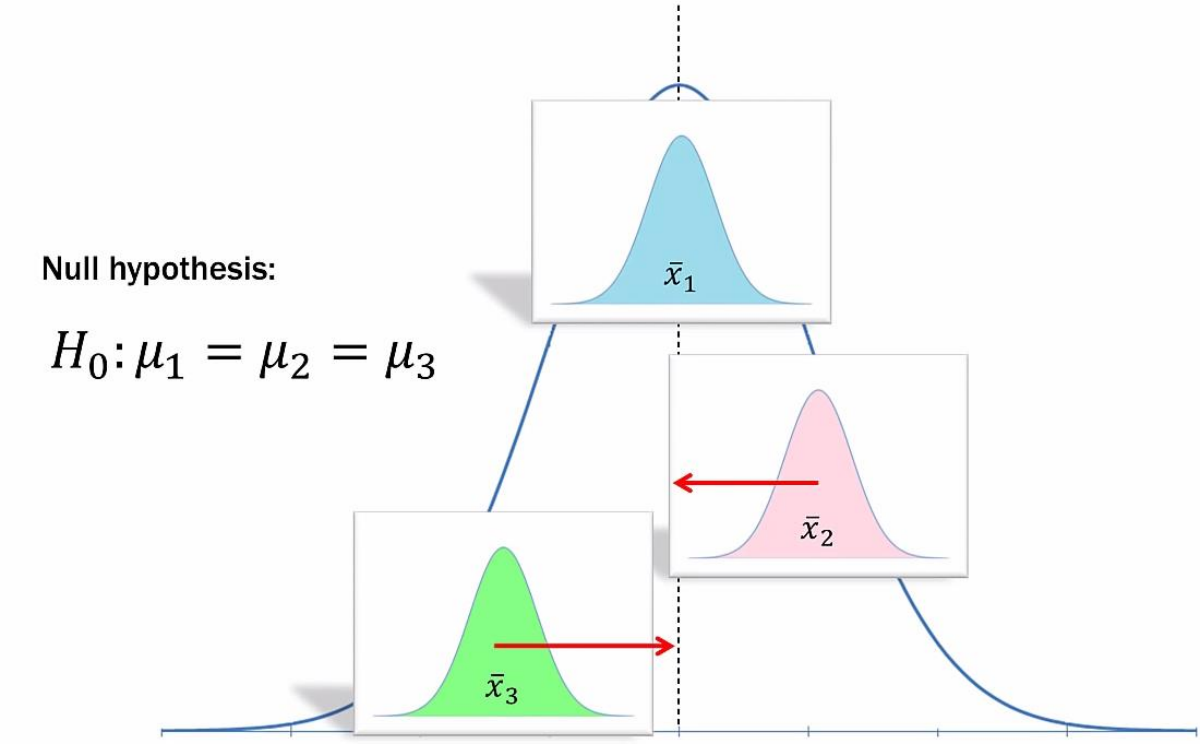


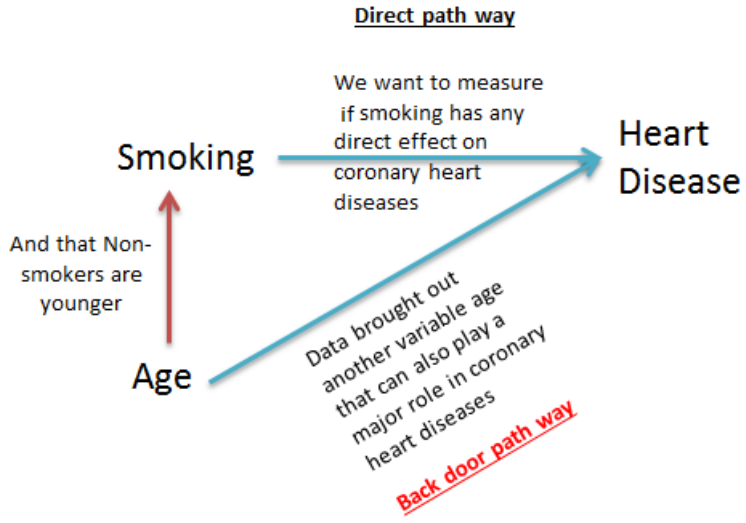



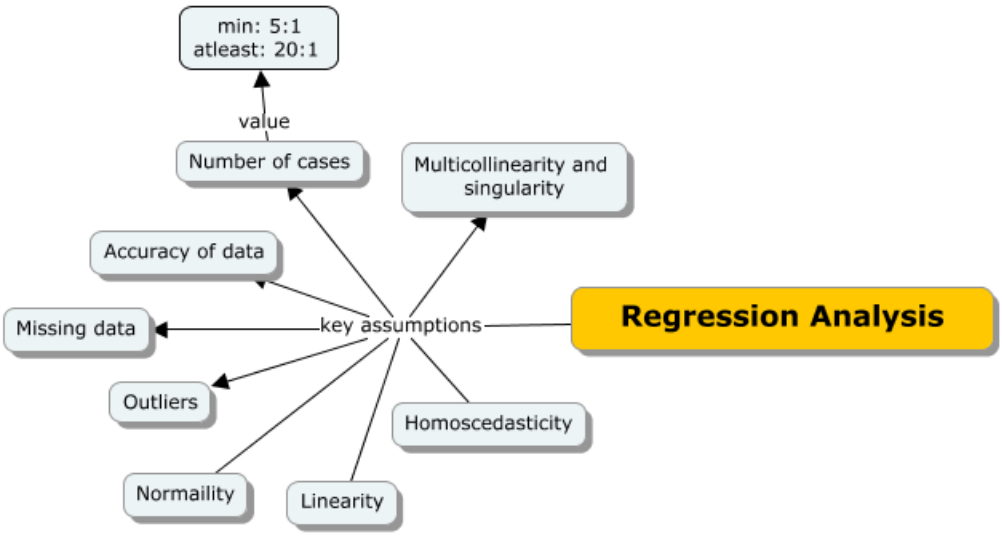
| | Variation | DOF |
|---|---|---|
| **Total** | 30 | 8(mn-1) |
| **Within** | 6 | 6(m(n-1) |
| **Between** | 24 | 2(m-1) |

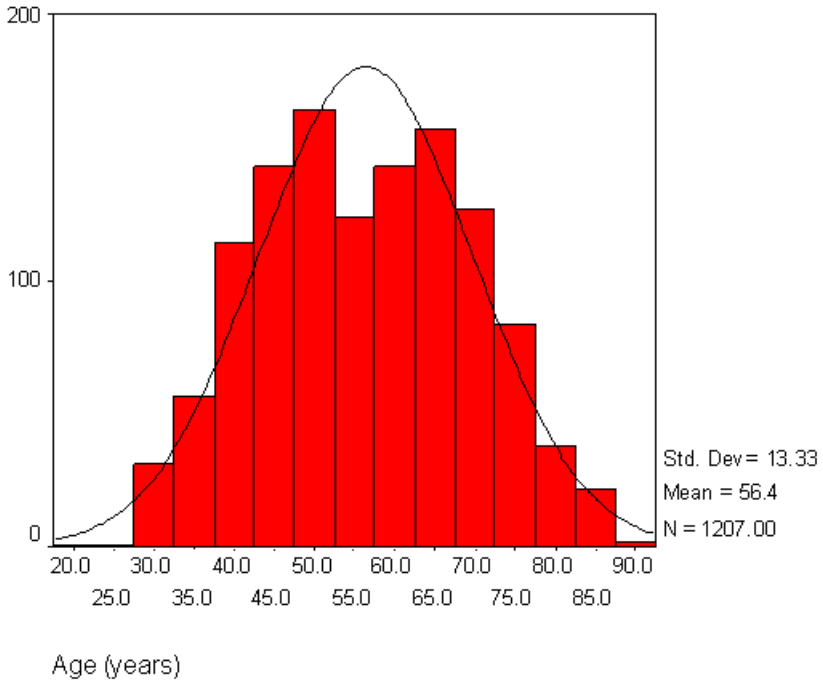
**Null hypothesis:**

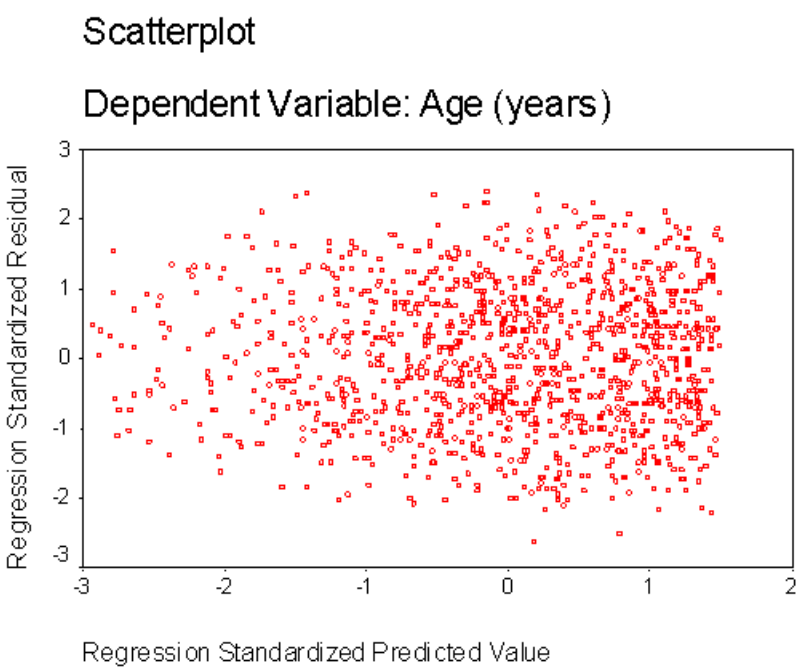
$$H_0: \mu_1 = \mu_2 = \mu_3$$

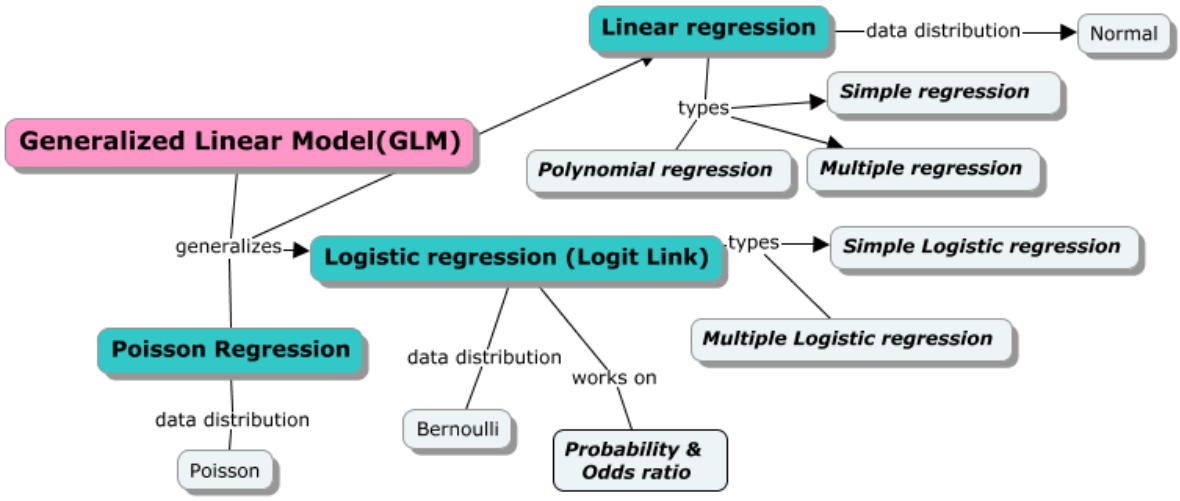


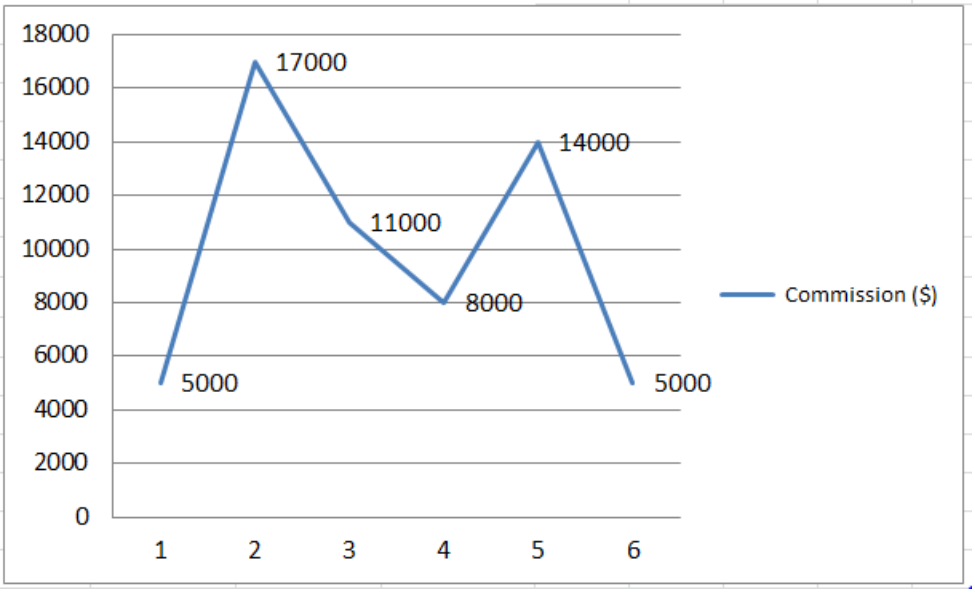

min: 5:1
atleast: 20:1
value
Number of cases
Multicollinearity and singularity
Accuracy of data
Missing data
key assumptions
Regression Analysis
Outliers
Homoscedasticity
Normaility
Linearity


200
100
0
20.0
25.0
30.0
35.0
40.0
45.0
50.0
55.0
60.0
65.0
70.0
75.0
80.0
85.0
90.0
Std. Dev = 13.33
Mean = 56.4
N = 1207.00
Age (years)

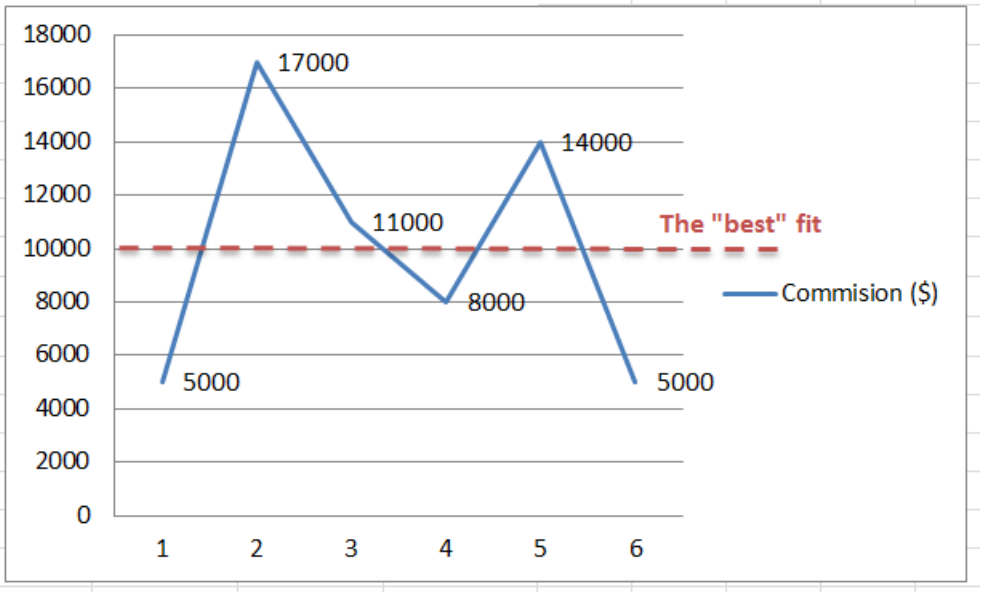


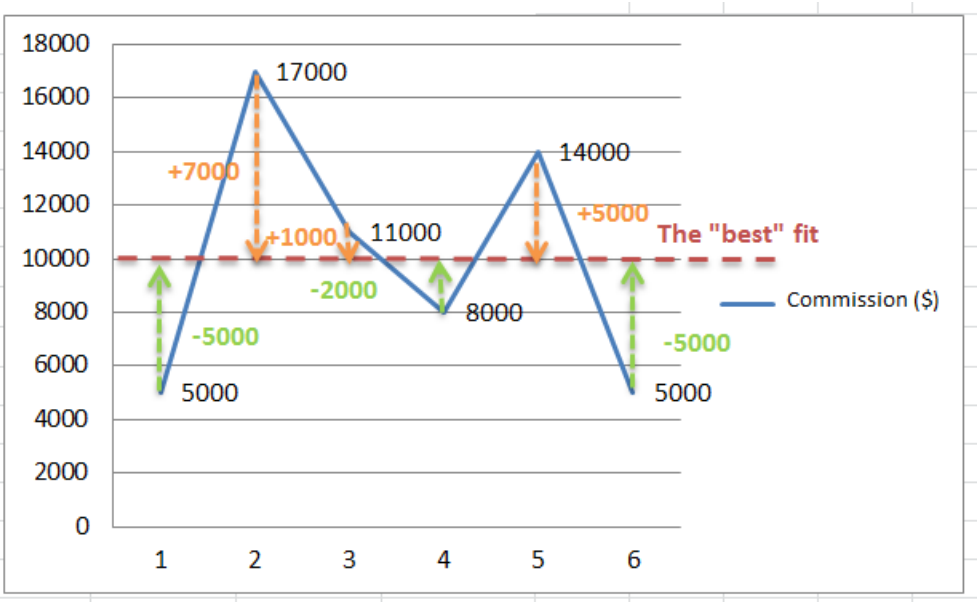

| Trns # | Commision ($) |
|---|---|
| 1 | 5000 |
| 2 | 17000 |
| 3 | 11000 |
| 4 | 8000 |
| 5 | 14000 |
| 6 | 5000 |




| Trns # | Commision ($) |
|---|---|
| 1 | 5000 |
| 2 | 17000 |
| 3 | 11000 |
| 4 | 8000 |
| 5 | 14000 |
| 6 | 5000 |
| **Mean** | **10000** |

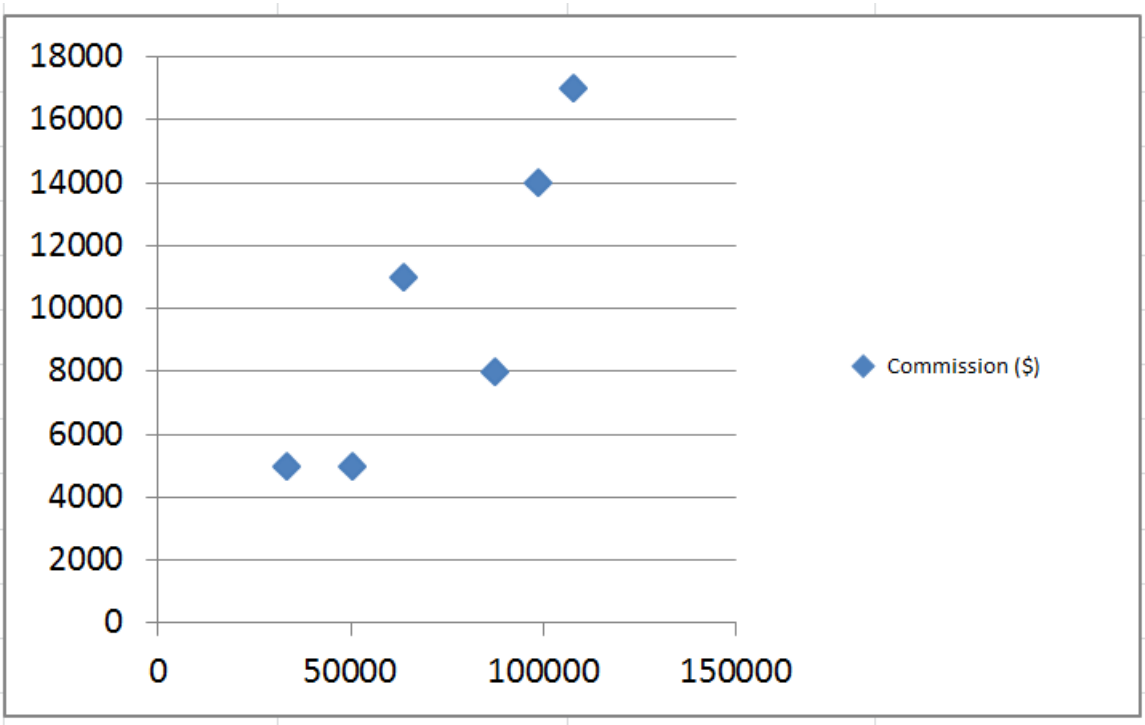


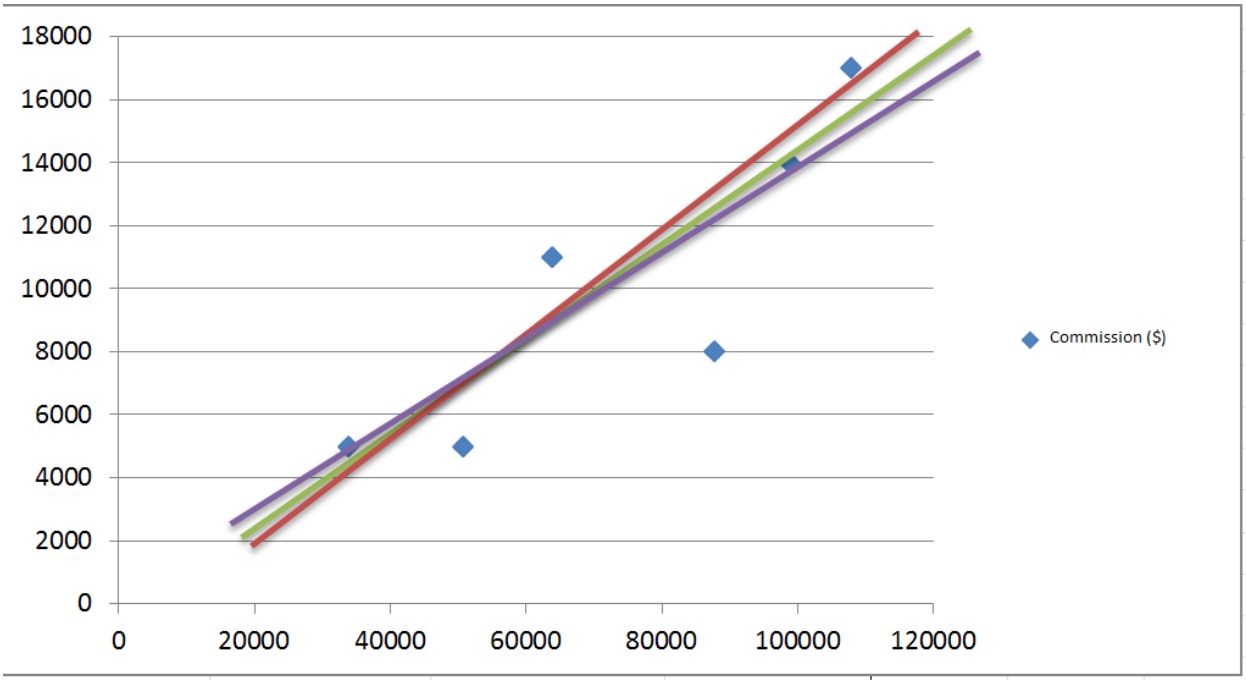
| Trns # | Commision ($) | error |
|---|---|---|
| 1 | 5000 | **-5000** |
| 2 | 17000 | **7000** |
| 3 | 11000 | **1000** |
| 4 | 8000 | **-2000** |
| 5 | 14000 | **4000** |
| 6 | 5000 | **-5000** |
| | **Total Error** | **0** |



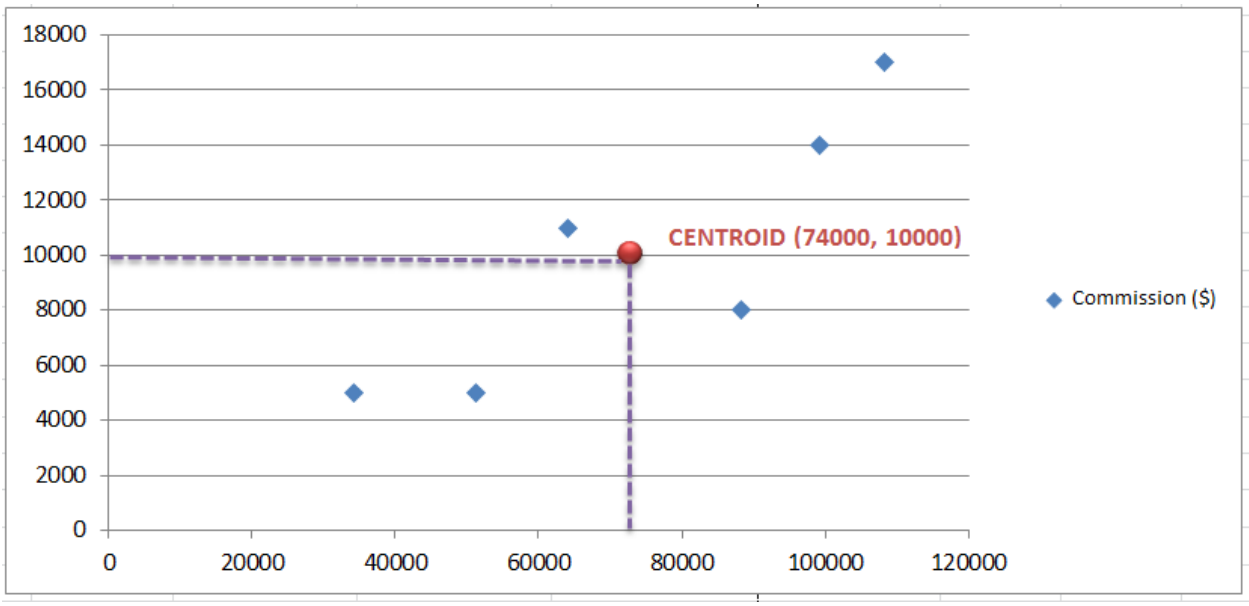

| Trns # | Commision ($) | error | error ^ 2 |
|---|---|---|---|
| 1 | 5000 | **-5000** | 25000000 |
| 2 | 17000 | **7000** | 49000000 |
| 3 | 11000 | **1000** | 1000000 |
| 4 | 8000 | **-2000** | 4000000 |
| 5 | 14000 | **4000** | 16000000 |
| 6 | 5000 | **-5000** | 25000000 |
| | | **SSE** | **120000000** |

| Transaction ($) | Commision ($) |
|---|---|
| 34000 | 5000 |
| 108000 | 17000 |
| 64000 | 11000 |
| 88000 | 8000 |
| 99000 | 14000 |
| 51000 | 5000 |






| | Transaction ($) | Commision ($) |
|---|---|---|
| | 34000 | 5000 |
| | 108000 | 17000 |
| | 64000 | 11000 |
| | 88000 | 8000 |
| | 99000 | 14000 |
| | 51000 | 5000 |
| **Mean** | 74000 | 10000 |



$$\hat{y}_i = b_0 + b_1 x_i$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$\bar{x} = mean\ of\ the\ independent\ variable$ $\quad x_i = value\ of\ independent\ variable$

$\bar{y} = mean\ of\ the\ dependent\ variable$ $\quad y_i = value\ of\ dependent\ variable$

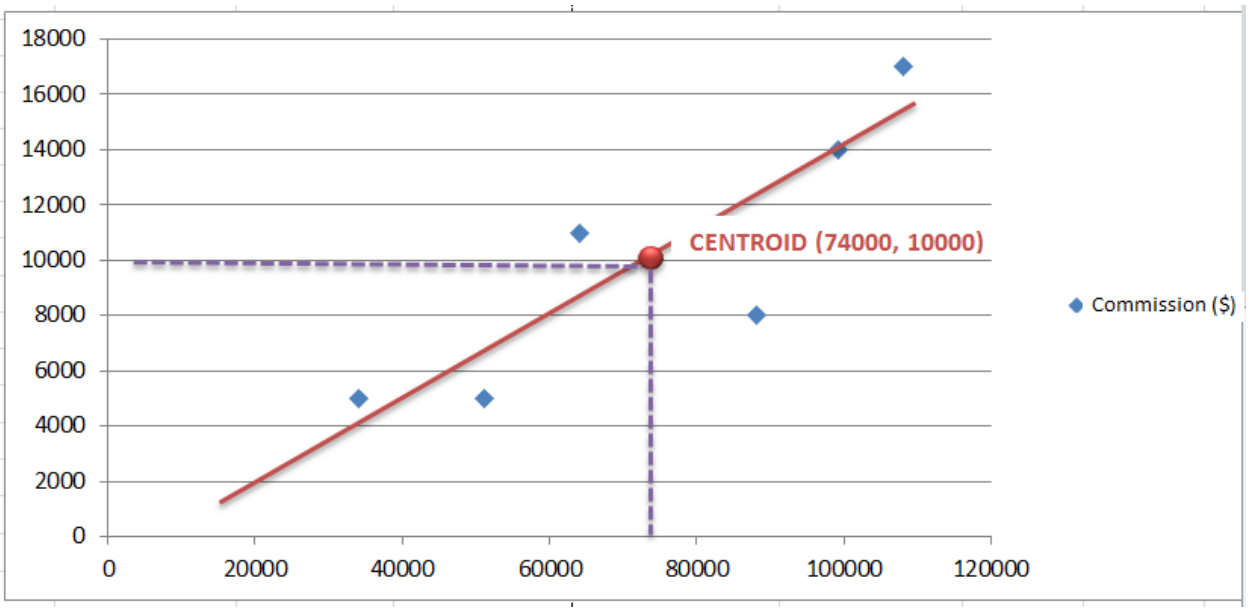
| | Transaction ($) | Commision ($) | Txn Deviation | Comm Deviation | Dev Product | Square Txn Dev |
|---|---|---|---|---|---|---|
| | 34000 | 5000 | -40000 | -5000 | 200000000 | 1600000000 |
| | 108000 | 17000 | 34000 | 7000 | 238000000 | 1156000000 |
| | 64000 | 11000 | -10000 | 1000 | -10000000 | 100000000 |
| | 88000 | 8000 | 14000 | -2000 | -28000000 | 196000000 |
| | 99000 | 14000 | 25000 | 4000 | 100000000 | 625000000 |
| | 51000 | 5000 | -23000 | -5000 | 115000000 | 529000000 |
| Mean | 74000 | 10000 | | Sum | 615000000 | 4206000000 |

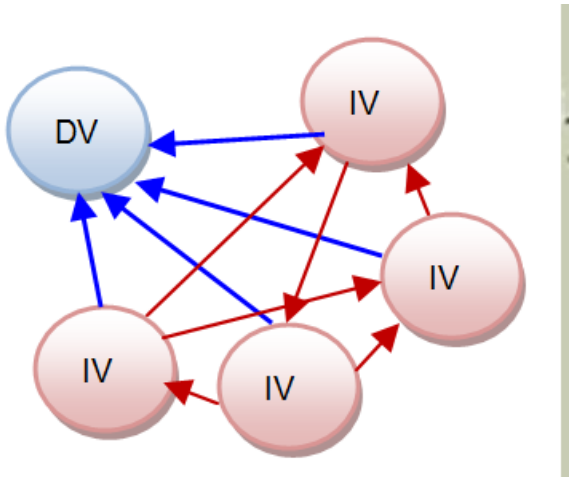
$\hat{y}_i = b_0 + b_1 x_i$ $\qquad b_0 = -0.8188$ intercept $\qquad b_1 = 0.1462$ slope

$$\hat{y}_i = -0.8188 + 0.1462x$$

OR

$$\hat{y}_i = 0.1462x - 0.8188$$


18000
16000
14000
12000
10000
8000
6000
4000
2000
0
0
20000
40000
60000
80000
100000
120000
CENTROID (74000, 10000)
Commission ($)


DV
IV
IV
IV
IV

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$$

linear parameters — error

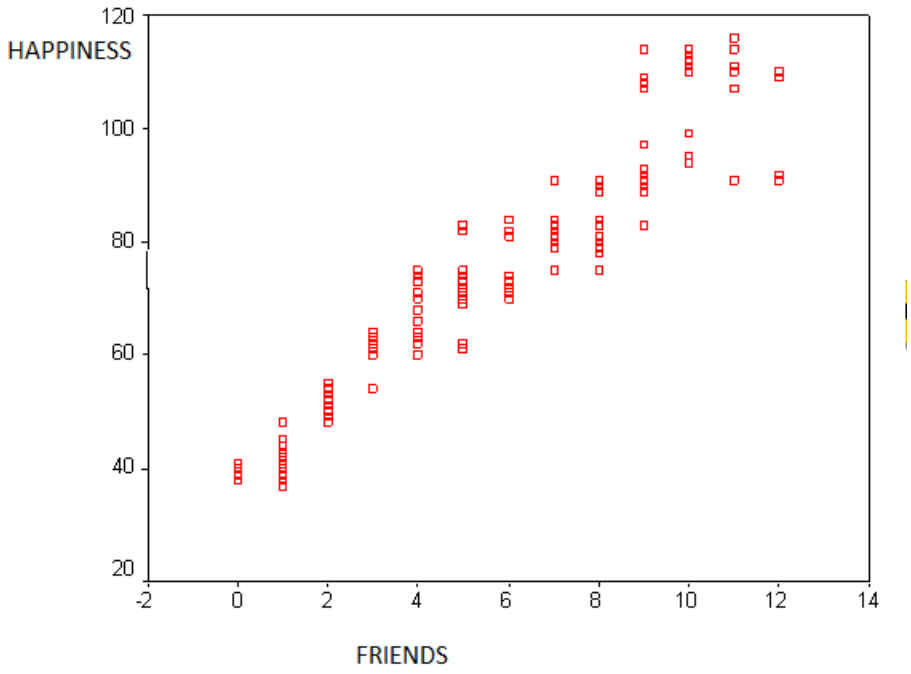
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

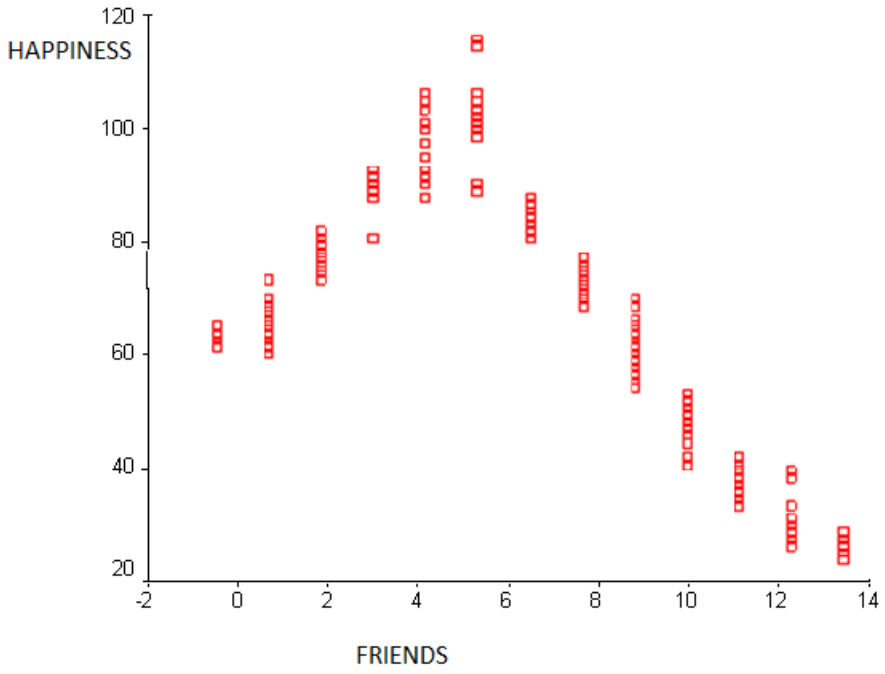
error term assumed to be zero

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_p x_p$$

$b_0, b_1, b_2, \ldots b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots \beta_p$

$\hat{y}$ = predicted value of the dependent variable



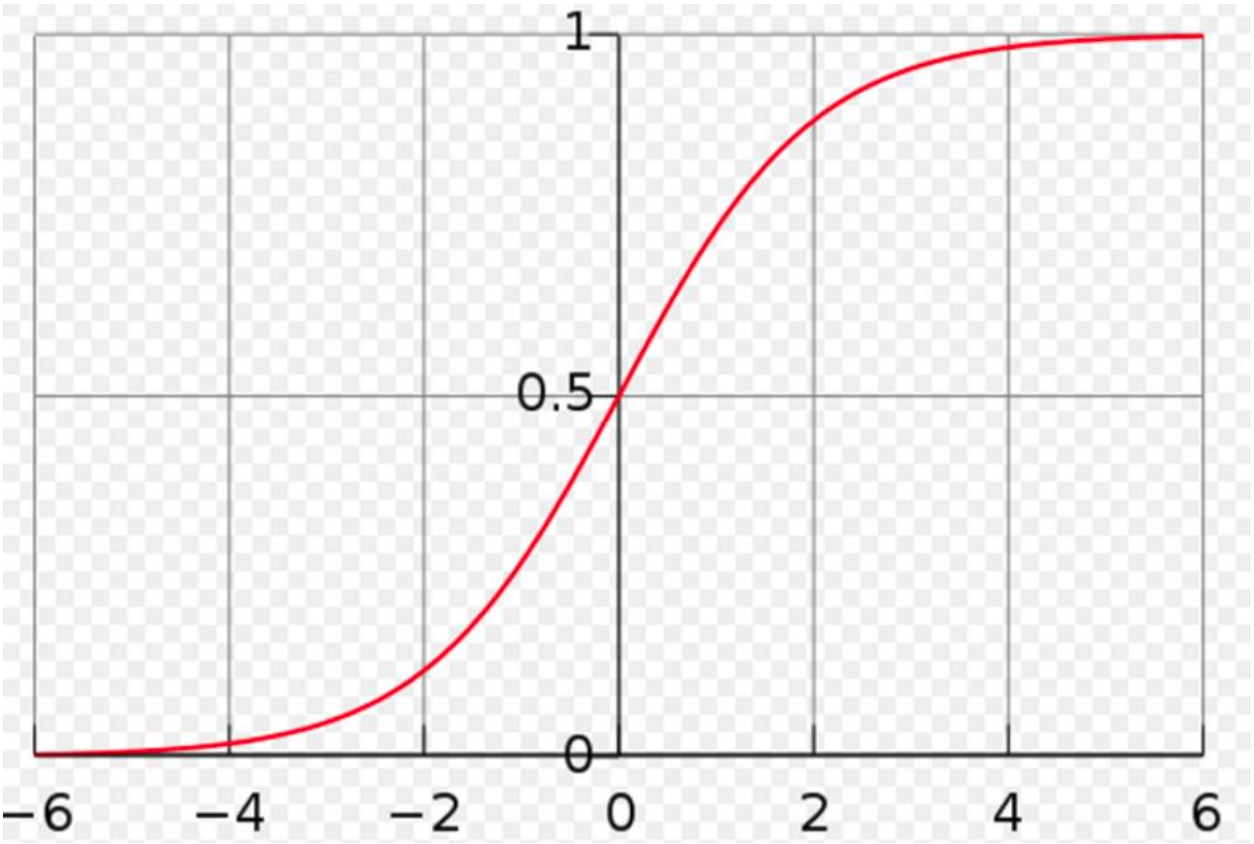


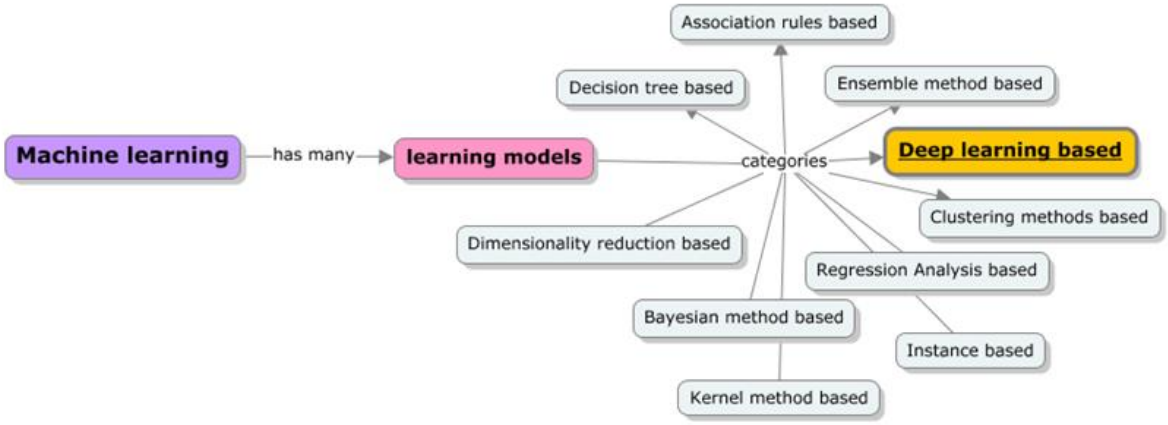

$p_i = \frac{1}{1+e^{-(a+bx_i)}}$ *(This is called the logistic response function)*

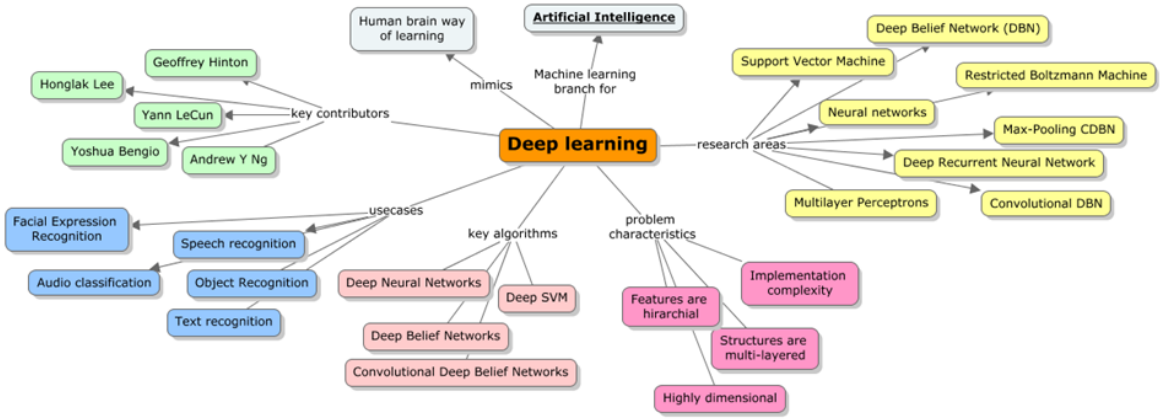
$$log\frac{p_i}{1-p_i} = a+bx_i$$

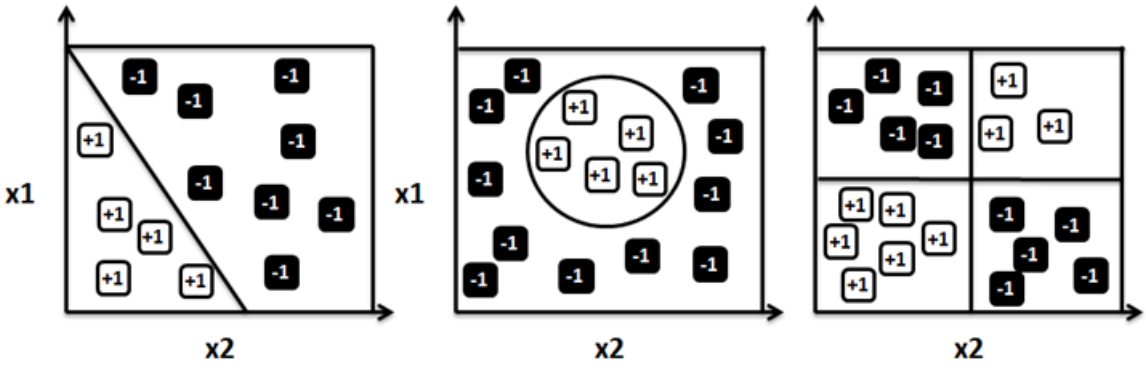
$$\frac{p_i}{1-p_i} = e^{\hat{a}+\hat{b}x_i}$$

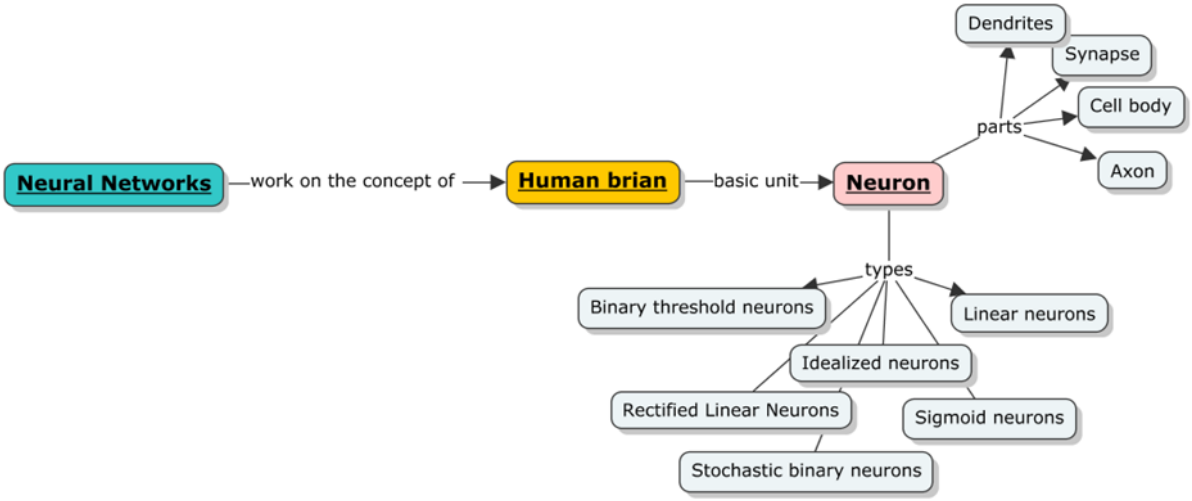
$$\widehat{p_i} = \frac{e^{\hat{a}+\hat{b}x_i}}{1+e^{\hat{a}+\hat{b}x_i}}$$

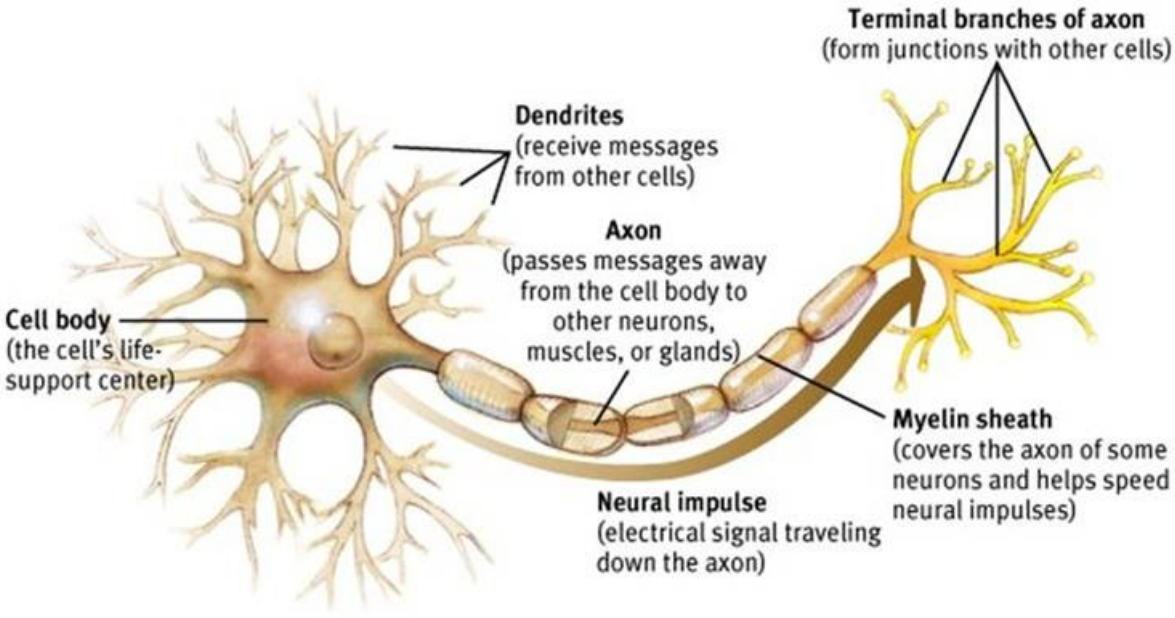

1
0.5
0
−6
−4
−2
0
2
4
6

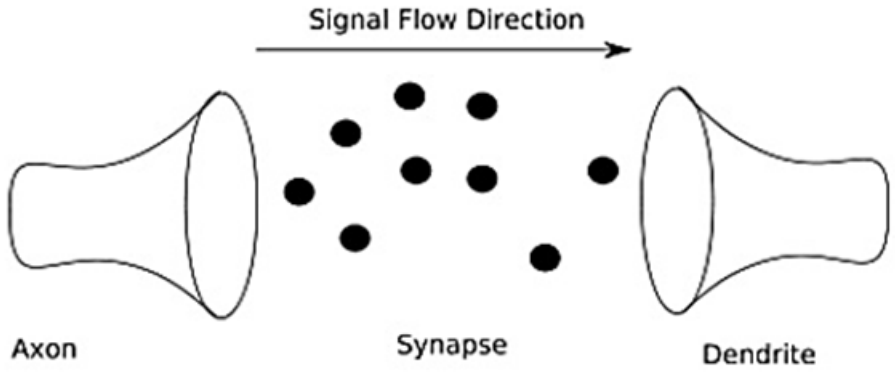

Machine learning
has many
learning models
categories
Association rules based
Decision tree based
Ensemble method based
Deep learning based
Clustering methods based
Dimensionality reduction based
Regression Analysis based
Bayesian method based
Instance based
Kernel method based

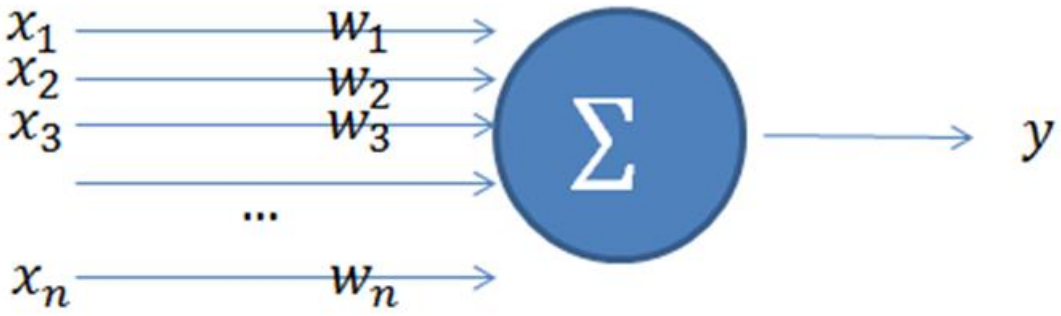

Deep learning
Artificial Intelligence
Human brain way of learning
mimics
Machine learning branch for
key contributors
Geoffrey Hinton
Honglak Lee
Yann LeCun
Yoshua Bengio
Andrew Y Ng
research areas
Deep Belief Network (DBN)
Support Vector Machine
Restricted Boltzmann Machine
Neural networks
Max-Pooling CDBN
Deep Recurrent Neural Network
Multilayer Perceptrons
Convolutional DBN
usecases
Facial Expression Recognition
Speech recognition
Audio classification
Object Recognition
Text recognition
key algorithms
Deep Neural Networks
Deep SVM
Deep Belief Networks
Convolutional Deep Belief Networks
problem characteristics
Implementation complexity
Features are hirarchial
Structures are multi-layered
Highly dimensional


x1
x2
x1
x2
x2

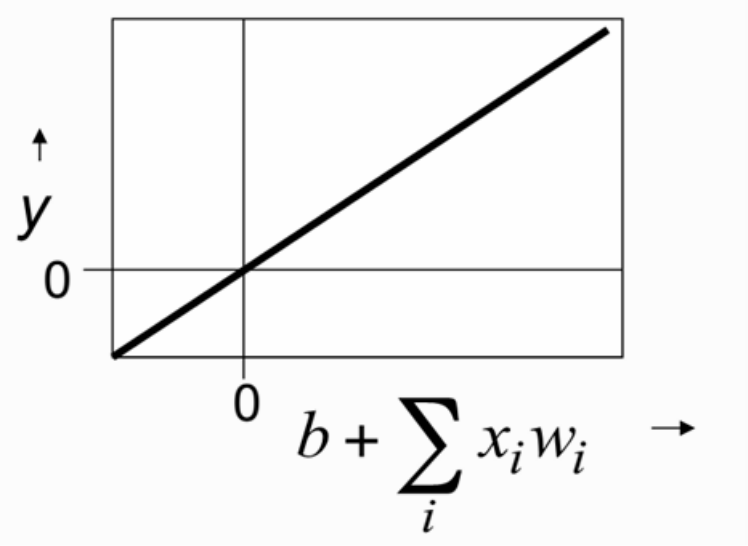

Neural Networks
work on the concept of
Human brian
basic unit
Neuron
parts
Dendrites
Synapse
Cell body
Axon
types
Binary threshold neurons
Linear neurons
Idealized neurons
Rectified Linear Neurons
Sigmoid neurons
Stochastic binary neurons


Terminal branches of axon
(form junctions with other cells)
Dendrites
(receive messages
from other cells)
Axon
(passes messages away
from the cell body to
other neurons,
muscles, or glands)
Cell body
(the cell's life-
support center)
Myelin sheath
(covers the axon of some
neurons and helps speed
neural impulses)
Neural impulse
(electrical signal traveling
down the axon)


Signal Flow Direction
Axon
Synapse
Dendrite

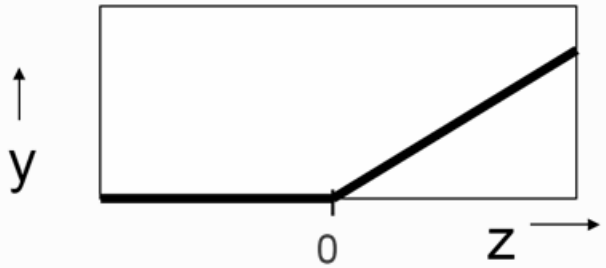

$x_1$
$x_2$
$x_3$
...
$x_n$
$w_1$
$w_2$
$w_3$
$w_n$
$\Sigma$
$y$

$$y = b + \sum_{i=1}^{n} (w_i x_i)$$



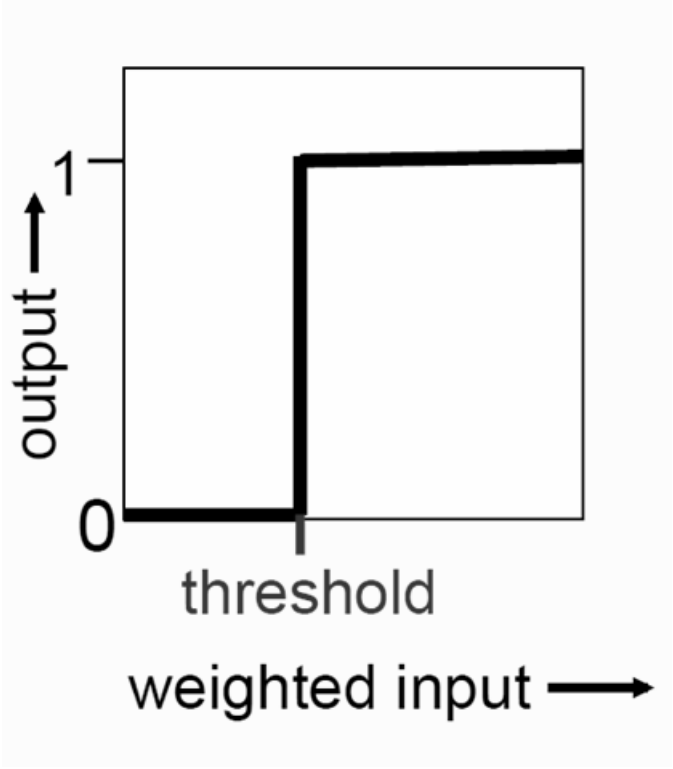

$$z = b + \sum_i x_i w_i$$

$$y = \begin{cases} z \text{ if } z > 0 \\ 0 \text{ otherwise} \end{cases}$$



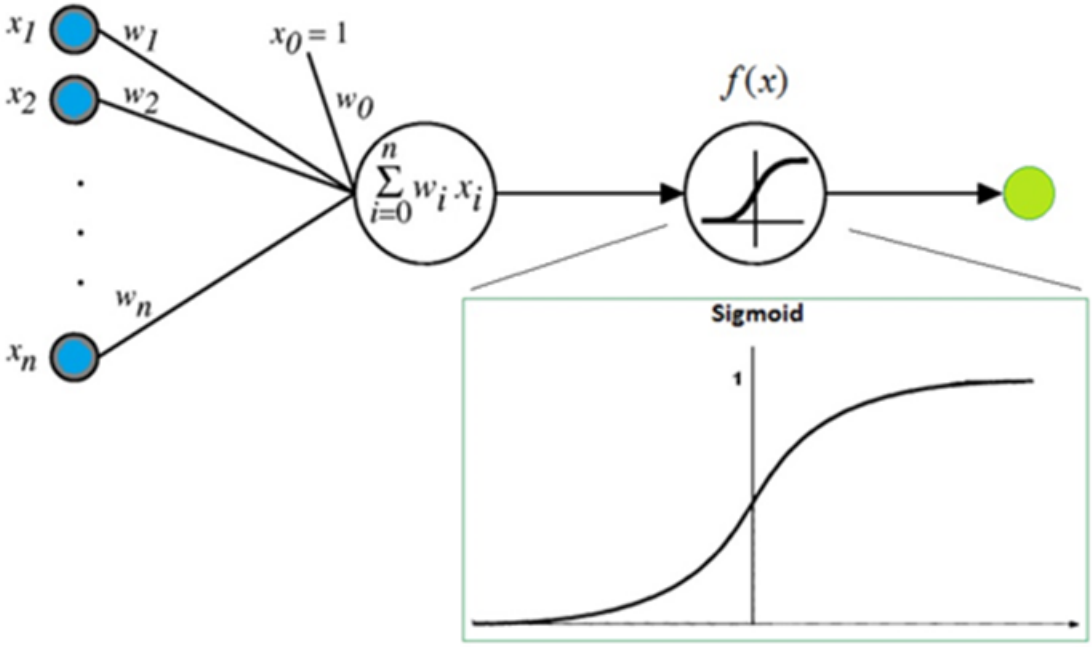

$$z = \sum_{i=1}^{n} (w_i x_i)$$

$$z = b + \sum_{i=1}^{n} (w_i x_i)$$



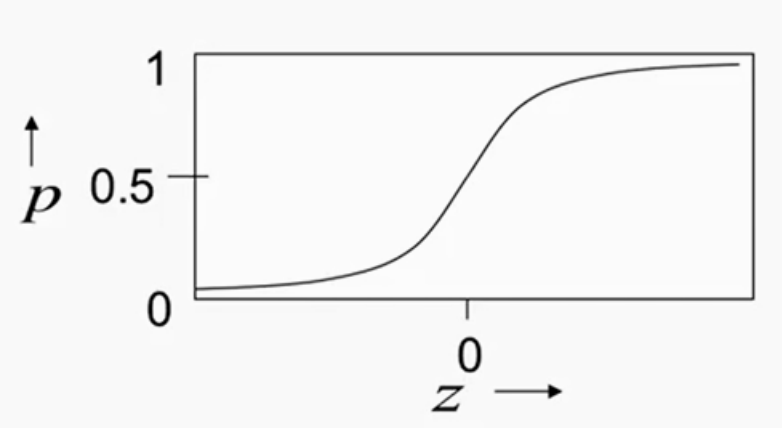

$$z = b + \sum_{i=1}^{n} \left( w_i x_i \right)$$

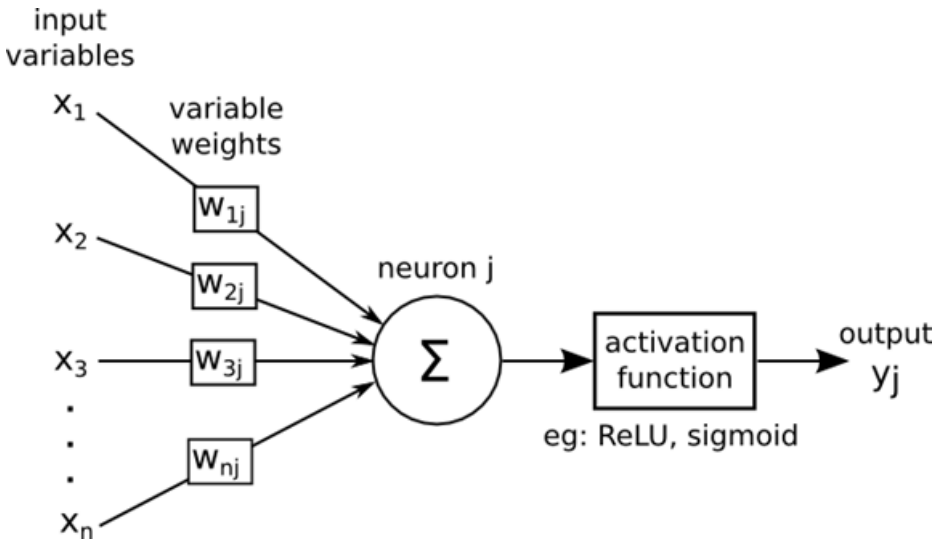
$$y = \frac{1}{1 + e^{-z}}$$

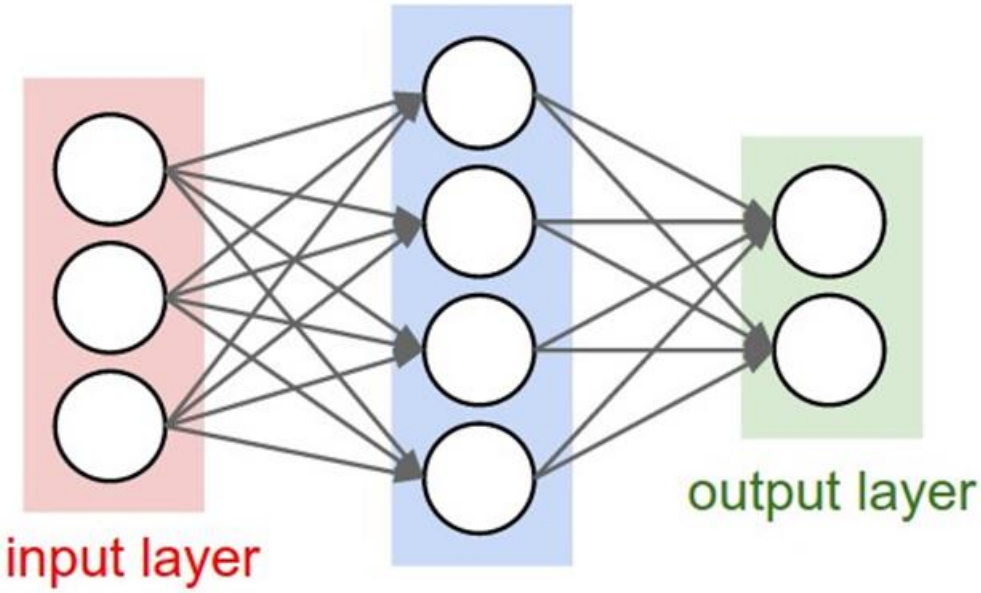


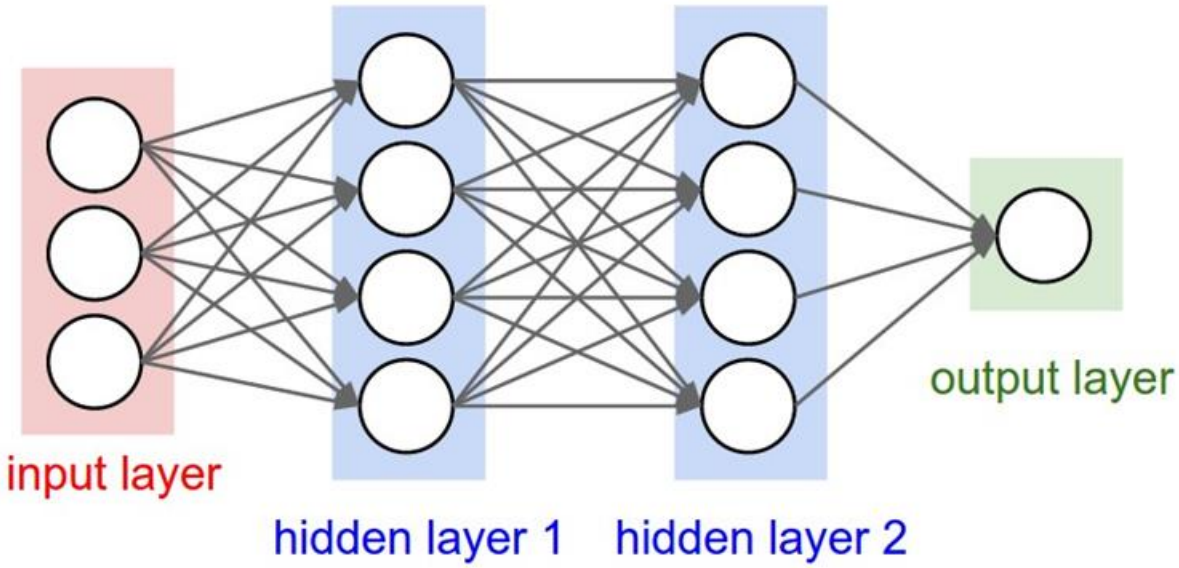
$$z = b + \sum_{i=1}^{n} \left( w_i x_i \right)$$

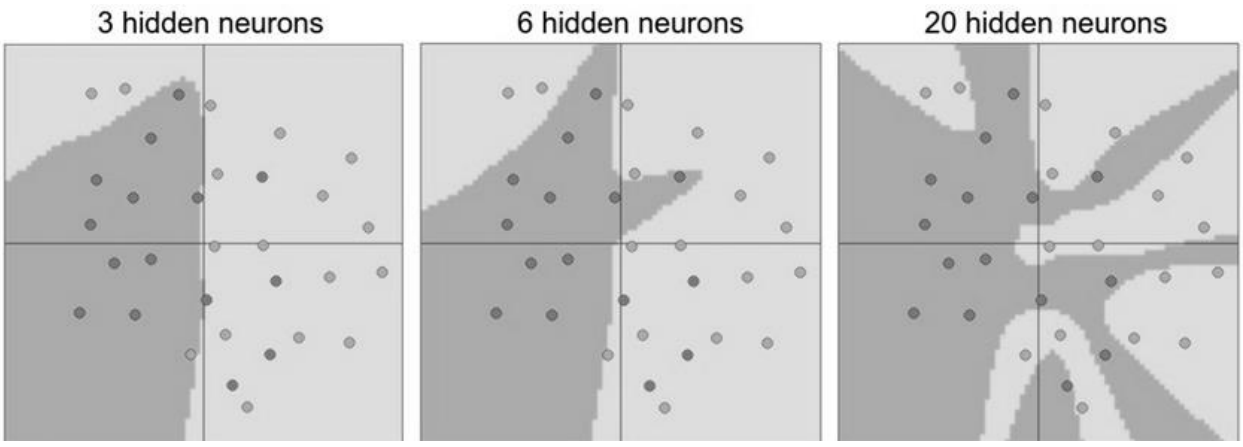
$$p\left( s = 1 \right) = \frac{1}{1 + e^{-z}}$$



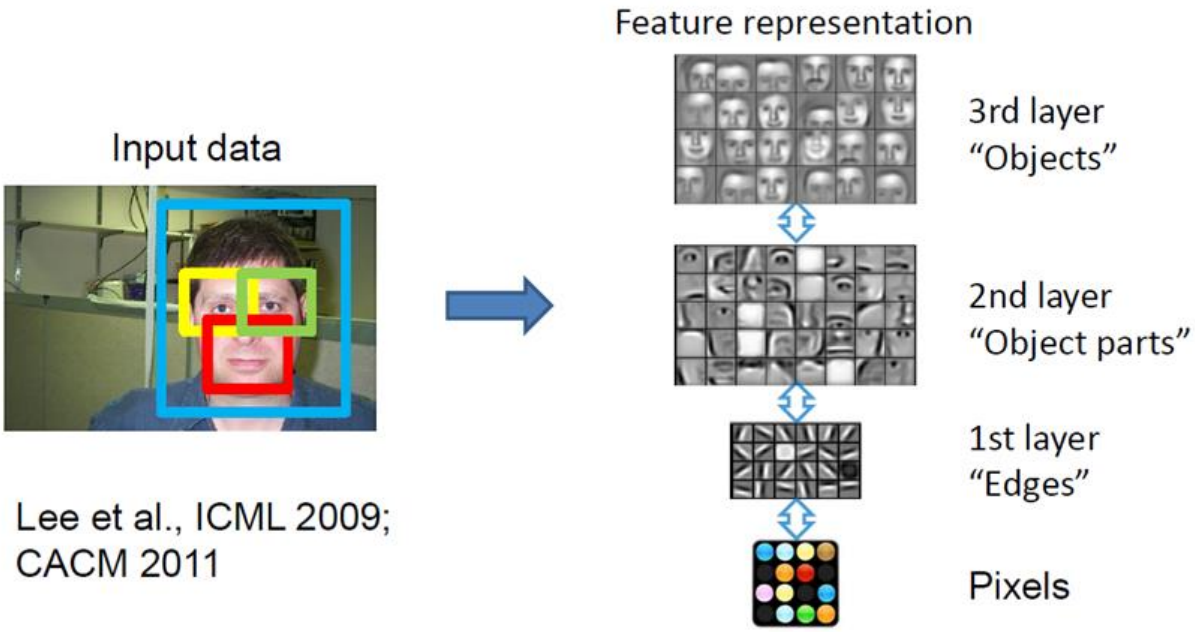



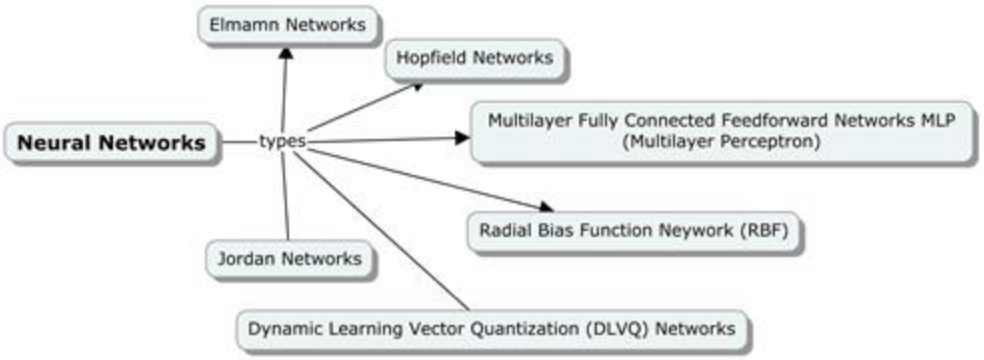

input layer
hidden layer
output layer

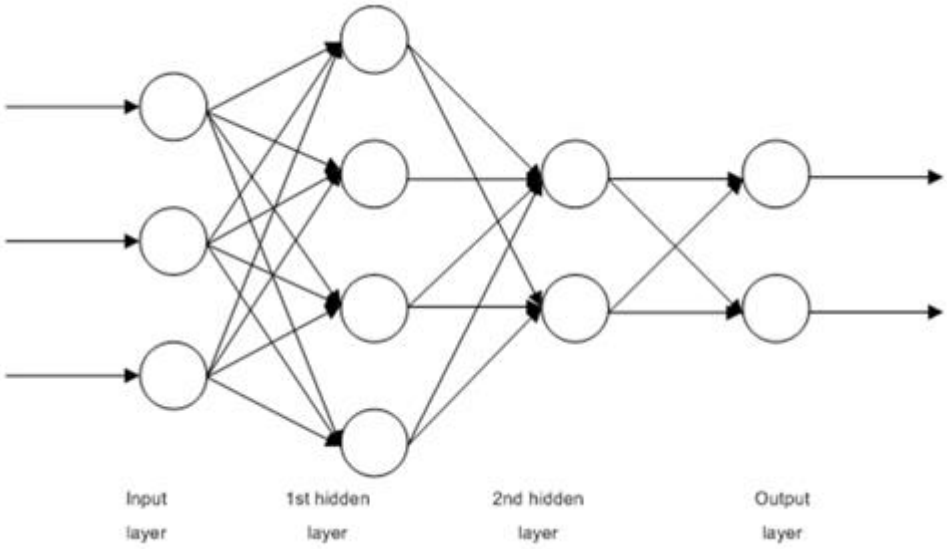

input layer
hidden layer 1
hidden layer 2
output layer


3 hidden neurons
6 hidden neurons
20 hidden neurons


Image/Video Pixels
Layer 1
Layer 2
Layer 3
Simple Classifier


Input data
Feature representation
3rd layer "Objects"
2nd layer "Object parts"
1st layer "Edges"
Pixels
Lee et al., ICML 2009; CACM 2011

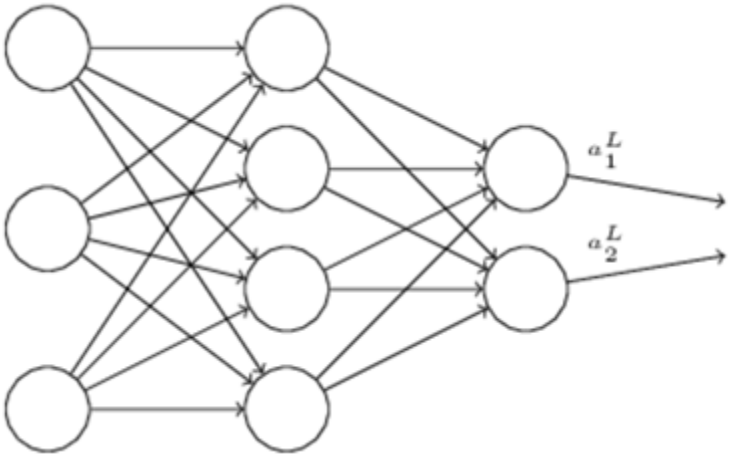



$$C = \frac{1}{2n} \sum_x \|y(x) - a^L(x)\|^2$$

$$C = \frac{1}{n} \sum_x C_x$$

$$C_x = \frac{1}{2} \|y - a^L\|^2$$




$$C = \frac{1}{2}\|y - a^L\|^2 = \frac{1}{2}\sum_j (y_j - a_j^L)^2$$

$$\partial C/\partial w_{jk}^l \text{ and } \partial C/\partial b_j^l$$

$$\delta_j^l \equiv \frac{\partial C}{\partial z_j^l}$$

Measures how the activation function changes at the current position in the network

$$\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L)$$

Measures how the cost function changes based on the jth activation output

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

Transpose of the weight matrix at l +1, this moves the error backwards through the network

Hadamard product, this moves the error backward through the activation layer l

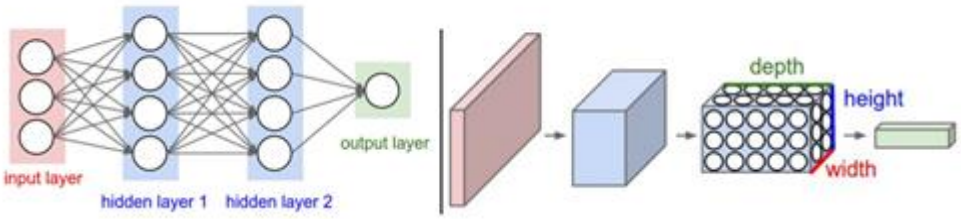
$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \odot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 * 3 \\ 2 * 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \end{bmatrix}$$

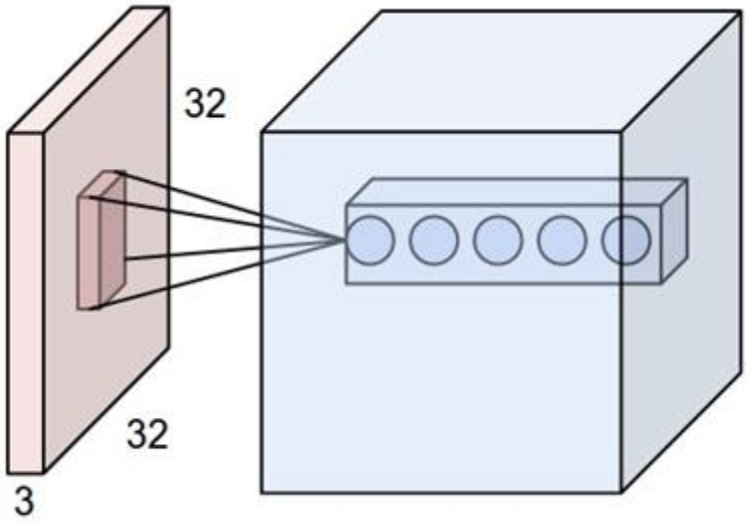
$$\frac{\partial C}{\partial b_j^l} = \delta_j^l$$

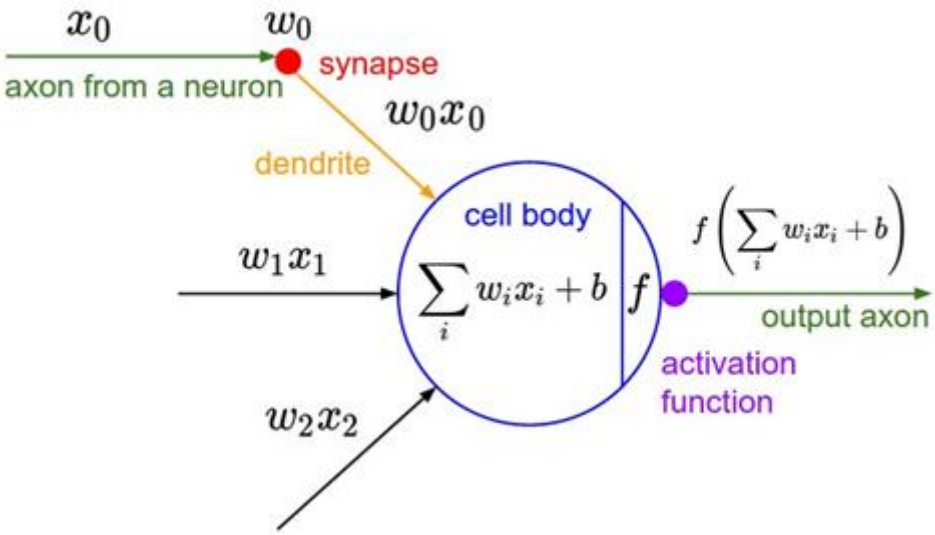
$$\frac{\partial C}{\partial b} = \delta$$

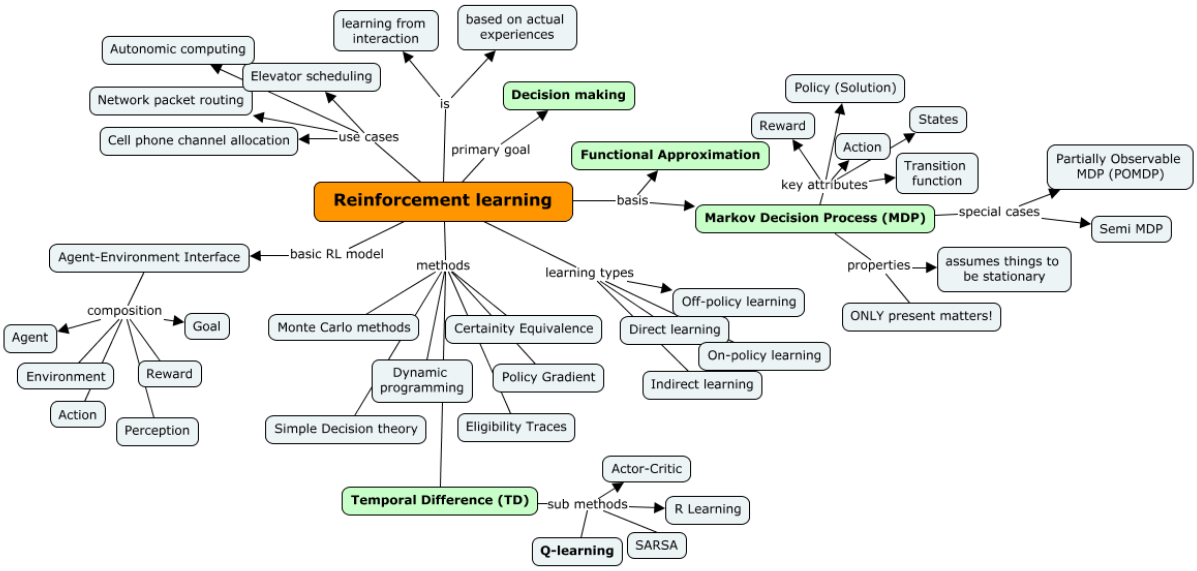
$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l$$

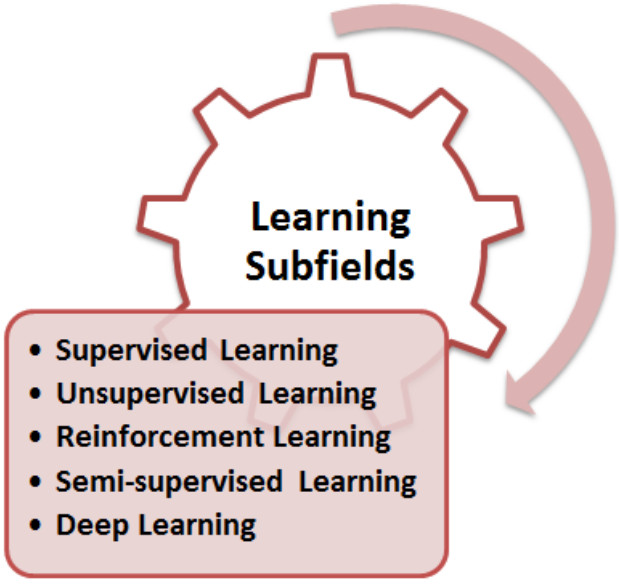
$$z^l = w^l a^{l-1} + b^l \text{ and } a^l = \sigma(z^l)$$

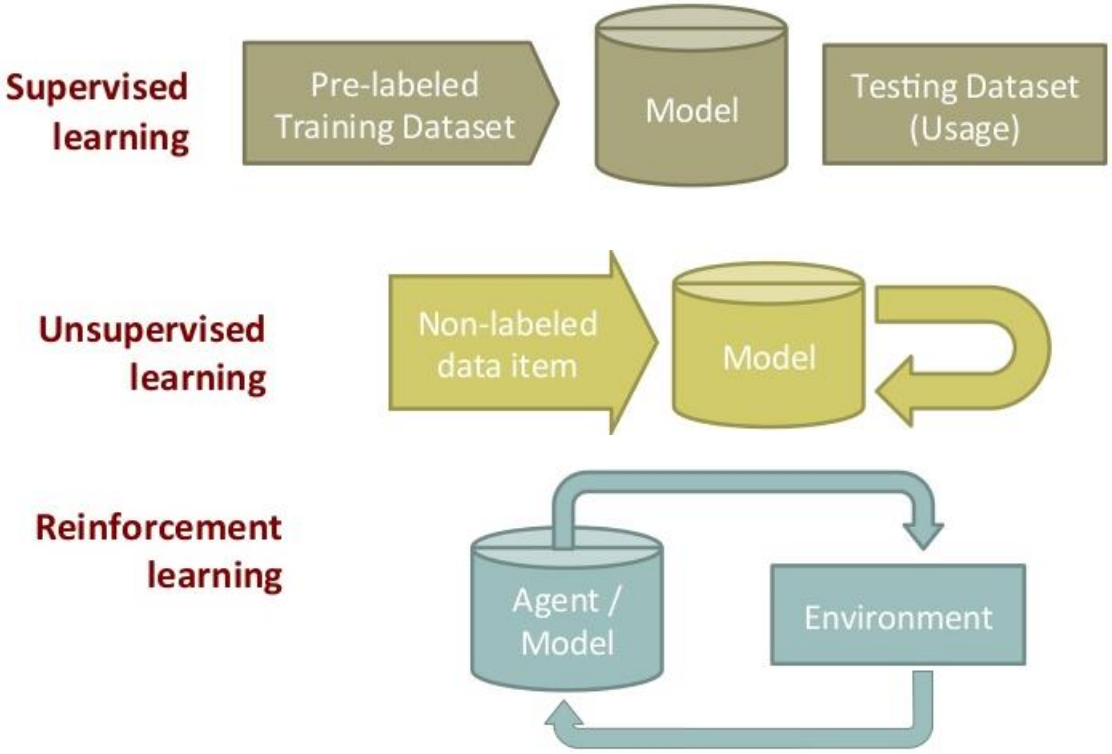


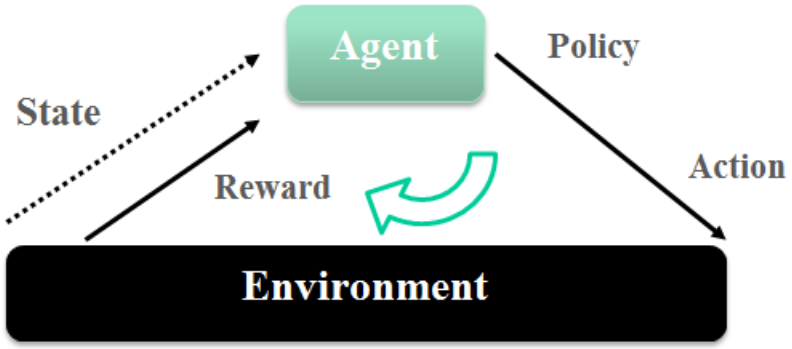

32
32
3


$x_0$
$w_0$
synapse
axon from a neuron
$w_0 x_0$
dendrite
cell body
$w_1 x_1$
$\sum_i w_i x_i + b$
$f$
$f\left(\sum_i w_i x_i + b\right)$
output axon
activation function
$w_2 x_2$

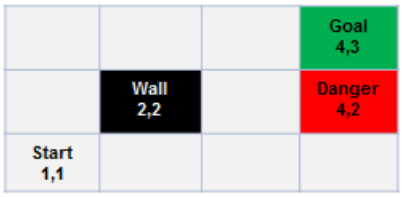

Reinforcement learning
is
learning from interaction
based on actual experiences
use cases
Autonomic computing
Elevator scheduling
Network packet routing
Cell phone channel allocation
primary goal
Decision making
basis
Functional Approximation
key attributes
Policy (Solution)
Reward
States
Action
Transition function
Markov Decision Process (MDP)
special cases
Partially Observable MDP (POMDP)
Semi MDP
properties
assumes things to be stationary
ONLY present matters!
basic RL model
Agent-Environment Interface
composition
Agent
Goal
Environment
Reward
Action
Perception
methods
Monte Carlo methods
Certainity Equivalence
Dynamic programming
Policy Gradient
Simple Decision theory
Eligibility Traces
learning types
Off-policy learning
Direct learning
On-policy learning
Indirect learning
Temporal Difference (TD)
sub methods
Actor-Critic
R Learning
Q-learning
SARSA

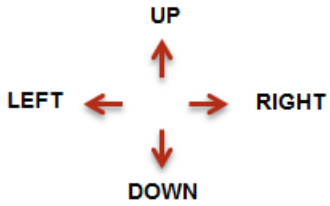

Learning Subfields
• Supervised Learning
• Unsupervised Learning
• Reinforcement Learning
• Semi-supervised Learning
• Deep Learning

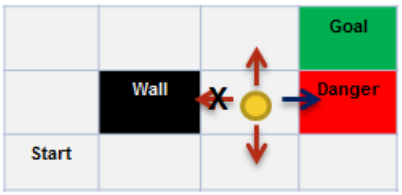


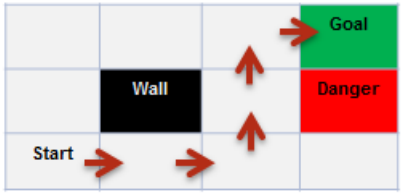

| | | | |
|---|---|---|---|
| 1,3 | 2,3 | 3,3 | 4,3 |
| 1,2 | 2,2 | 3,2 | 4,2 |
| 1,1 | 2,1 | 3,1 | 4,1 |


Goal
4,3
Wall
2,2
Danger
4,2
Start
1,1

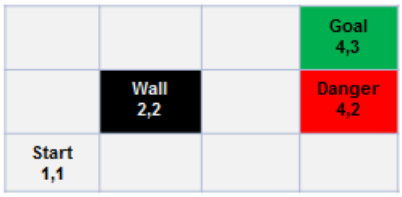

UP
LEFT
RIGHT
DOWN

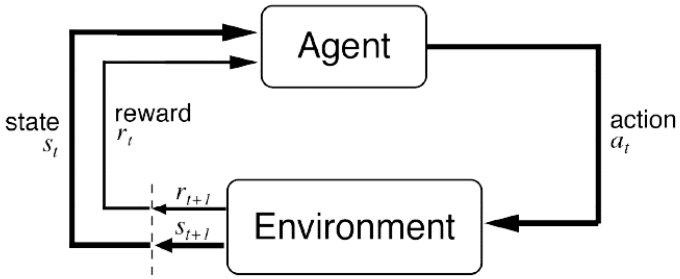

Goal
Wall
Danger
Start


Goal
4,3
Wall
2,2
Danger
4,2
Start
1,1


Goal
Wall
Danger
Start


Goal
Wall
Danger
Start



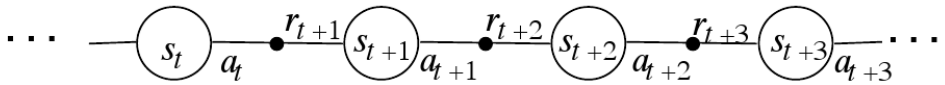

Agent and environment interact at discrete time steps : $t = 0, 1, 2, \ldots$

Agent observes state at step $t$: $s_t \in S$

produces action at step $t$: $a_t \in A(s_t)$

gets resulting reward : $r_{t+1} \in \Re$

and resulting next state : $s_{t+1}$




$$\pi * = argmax\, \pi \; (E\,[\sum_{t=0}^{\infty} r^t \mathrm{R}(s_t)/\pi]$$

$$\mathrm{R(s)} \neq \mathrm{U^{\pi}(s)} = E\,[\sum_{t=0}^{\infty} r^t \mathrm{R}(s_t)/\pi\,, s_0 = s]$$

$$\pi * = argmax\pi\, E\,[\sum_{s1} \mathrm{T(s,a,s1)}U(s1)$$

$$U(s) = R(s) + \gamma\, max\pi\, E\,[\sum_{s1} \mathrm{T(s,a,s1)}U(s1)]$$

$$
\begin{aligned}
V^*(s) &= \max_{a \in \mathcal{A}(s)} Q^{\pi^*}(s,a) \\
&= \max_a E_{\pi^*}\left\{ R_t \;\middle|\; s_t = s, a_t = a \right\} \\
&= \max_a E_{\pi^*}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; s_t = s, a_t = a \right\} \\
&= \max_a E_{\pi^*}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \;\middle|\; s_t = s, a_t = a \right\} \\
&= \max_a E\left\{ r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a \right\} \\
&= \max_{a \in \mathcal{A}(s)} \sum_{s'} \mathcal{P}^a_{ss'} \left[ \mathcal{R}^a_{ss'} + \gamma V^*(s') \right].
\end{aligned}
$$

1. Initialization
   $V(s) \in \Re$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

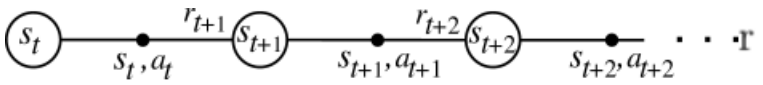
2. Policy Evaluation
   Repeat
   $\Delta \leftarrow 0$
   For each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s'} \mathcal{P}^{\pi(s)}_{ss'} \left[ \mathcal{R}^{\pi(s)}_{ss'} + \gamma V(s') \right]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number)

3. Policy Improvement
   *policy-stable* $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   $b \leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s'} \mathcal{P}^a_{ss'} \left[ \mathcal{R}^a_{ss'} + \gamma V(s') \right]$
   If $b \neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop; else go to 2

Initialize $V$ arbitrarily, e.g.,$V(s) = 0$, for all $s \in \mathcal{S}^+$

Repeat
  $\Delta \leftarrow 0$
  For each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V(s')\right]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, $\pi$, such that
  $\pi(s) = \arg\max_a \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V(s')\right]$

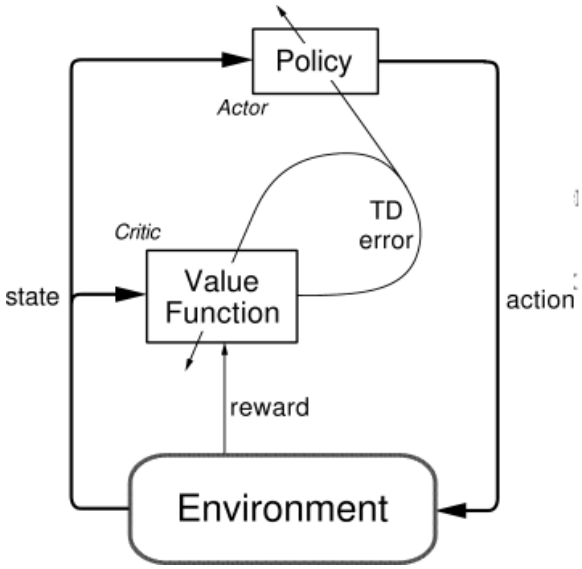



$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Big[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\Big].$$

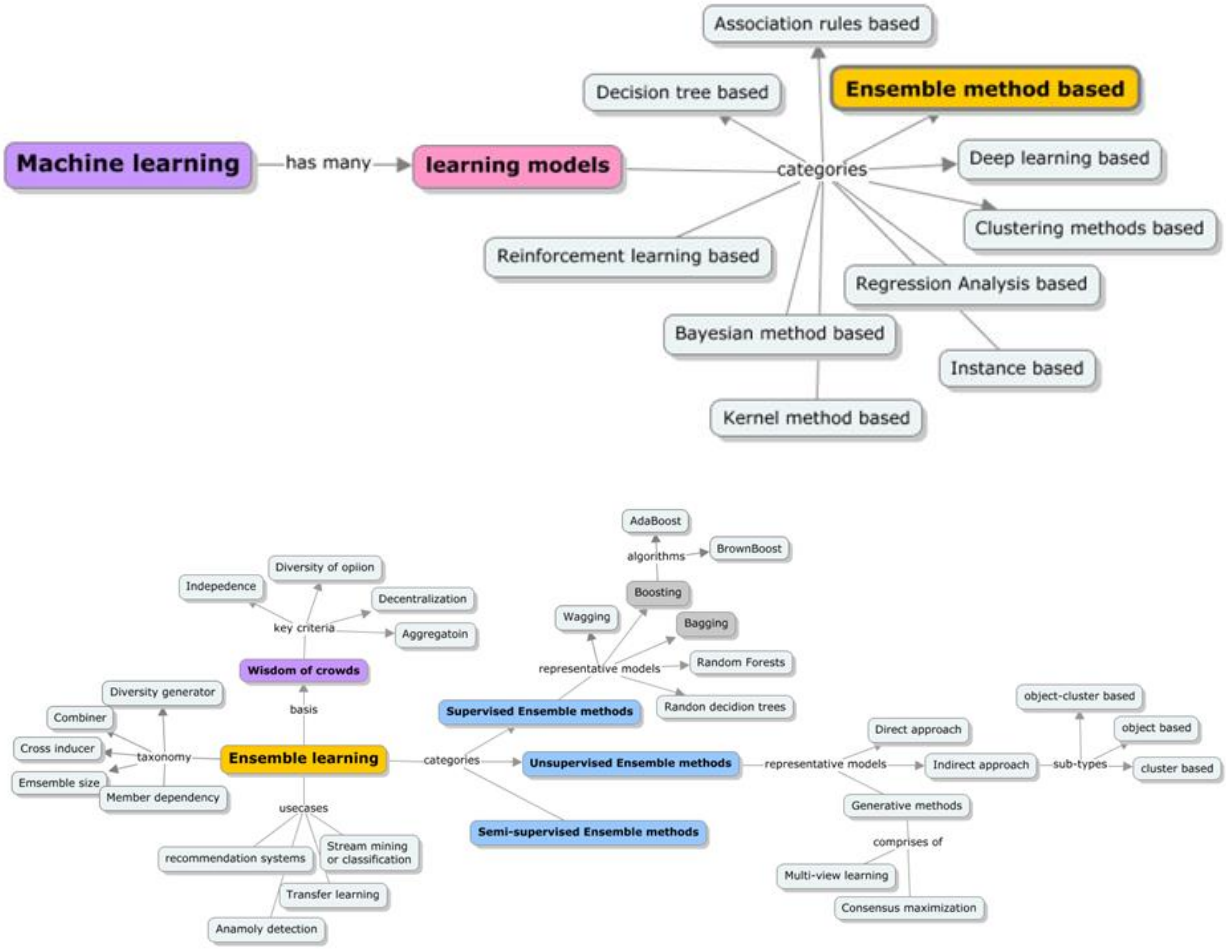
Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize $s$
  Choose $a$ from $s$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
  Repeat (for each step of episode):
    Take action $a$, observe $r$, $s'$
    Choose $a'$ from $s'$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    $Q(s, a) \leftarrow Q(s, a) + \alpha \big[r + \gamma Q(s', a') - Q(s, a)\big]$
    $s \leftarrow s'$; $a \leftarrow a'$;
  until $s$ is terminal

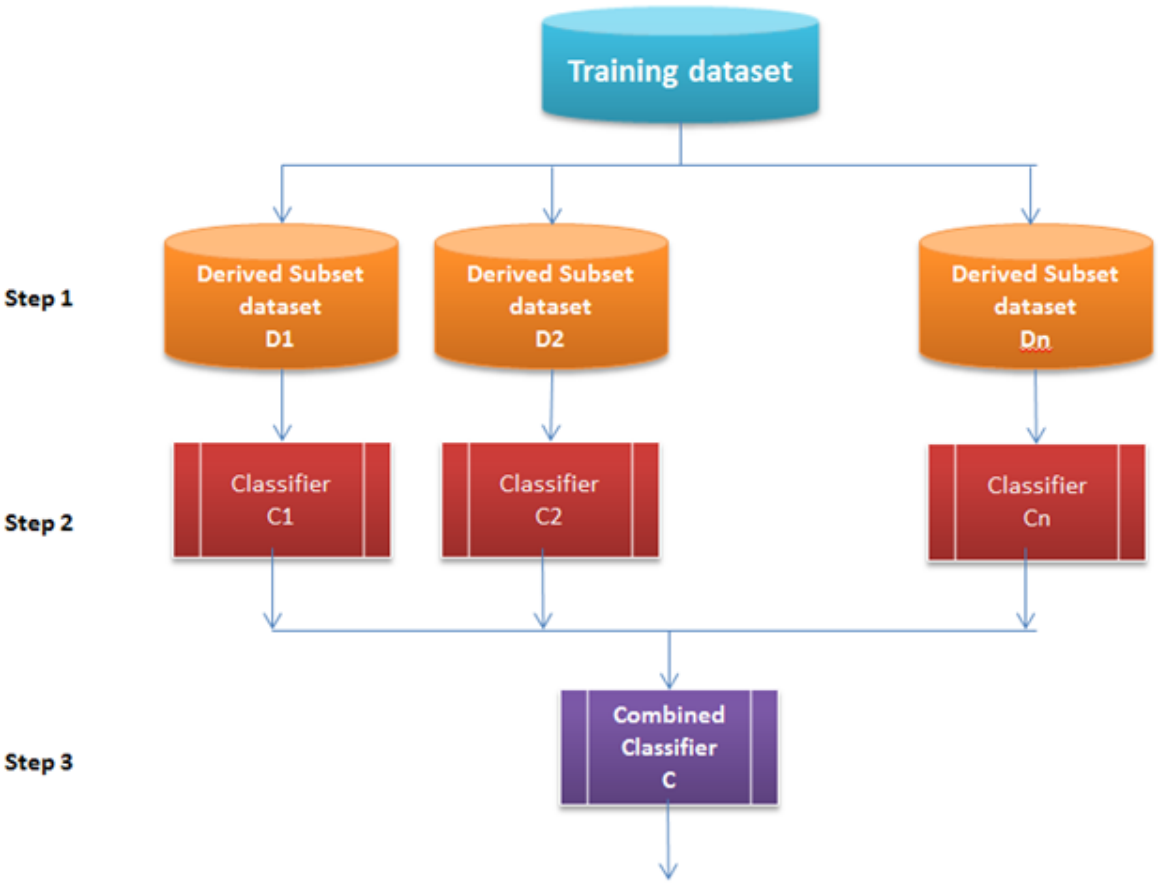
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Big[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\Big].$$

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
  Initialize $s$
  Repeat (for each step of episode):
    Choose $a$ from $s$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    Take action $a$, observe $r$, $s'$
    $Q(s,a) \leftarrow Q(s,a) + \alpha\big[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\big]$
    $s \leftarrow s'$;
  until $s$ is terminal



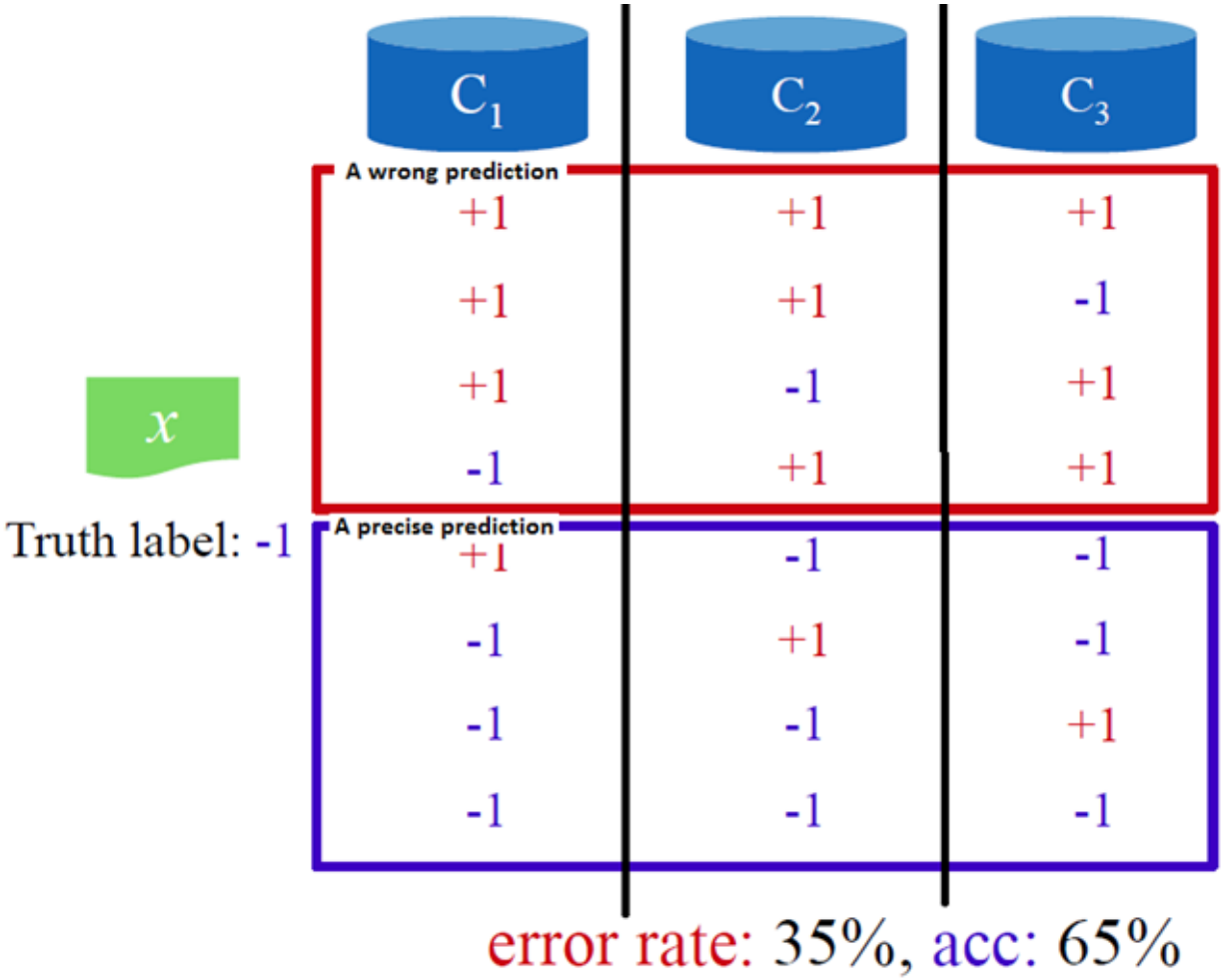

Initialize $\rho$ and $Q(s,a)$, for all $s, a$, arbitrarily
Repeat forever:
  $s \leftarrow$ current state
  Choose action $a$ in $s$ using behavior policy (e.g., $\varepsilon$-greedy)
  Take action $a$, observe $r$, $s'$
  $Q(s,a) \leftarrow Q(s,a) + \alpha\ \big[r - \rho + \max_{a'} Q(s',a') - Q(s,a)\big]$
  If $Q(s,a) = \max_a Q(s,a)$, then:
    $\rho \leftarrow \rho + \beta\ \big[r - \rho + \max_{a'} Q(s',a') - \max_a Q(s,a)\big]$


Machine learning
has many
learning models
categories
Association rules based
Decision tree based
Ensemble method based
Deep learning based
Clustering methods based
Reinforcement learning based
Regression Analysis based
Bayesian method based
Instance based
Kernel method based
Ensemble learning
basis
Wisdom of crowds
key criteria
Indepedence
Diversity of opiion
Decentralization
Aggregatoin
taxonomy
Diversity generator
Combiner
Cross inducer
Emsemble size
Member dependency
usecases
recommendation systems
Stream mining or classification
Transfer learning
Anamoly detection
categories
Supervised Ensemble methods
representative models
Boosting
Wagging
Bagging
Random Forests
Randon decidion trees
algorithms
AdaBoost
BrownBoost
Unsupervised Ensemble methods
representative models
Direct approach
Indirect approach
sub-types
object-cluster based
object based
cluster based
Generative methods
comprises of
Multi-view learning
Consensus maximization
Semi-supervised Ensemble methods



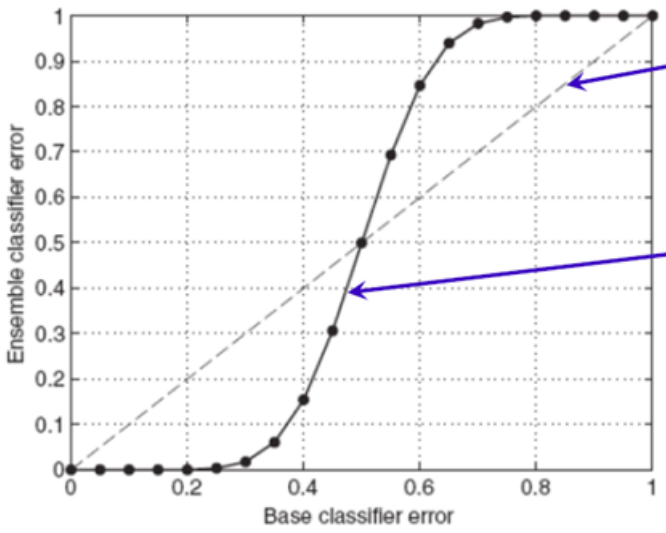

Base classifiers: $C_1$ $C_2$ $C_3$

A test instance: $x$



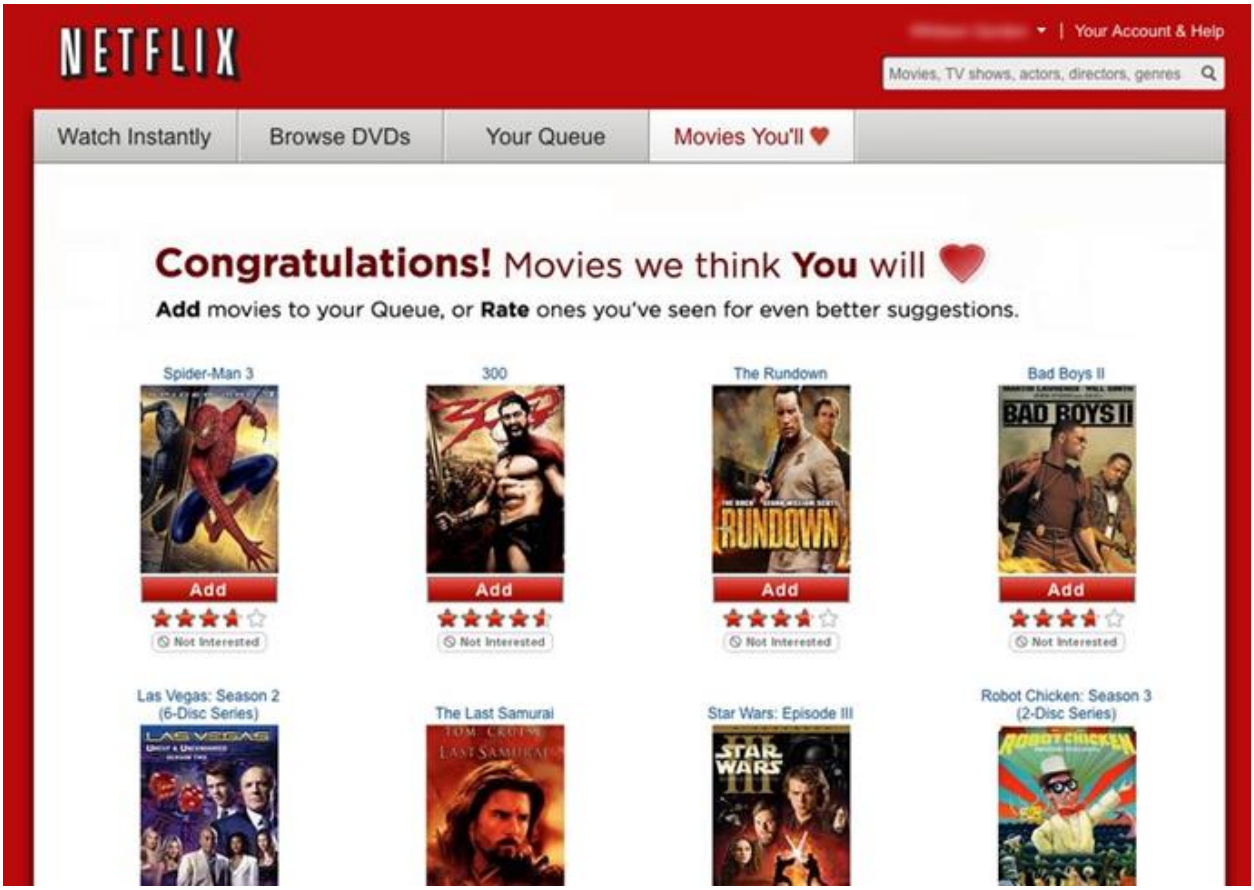

$$\sum_{i=2}^{3}\binom{3}{i}\varepsilon^{i}(1-\varepsilon)^{3-i} = 3\times 0.35^{2}\times 0.65 + 1\times 0.35^{3}\times 1 = 0.2817$$

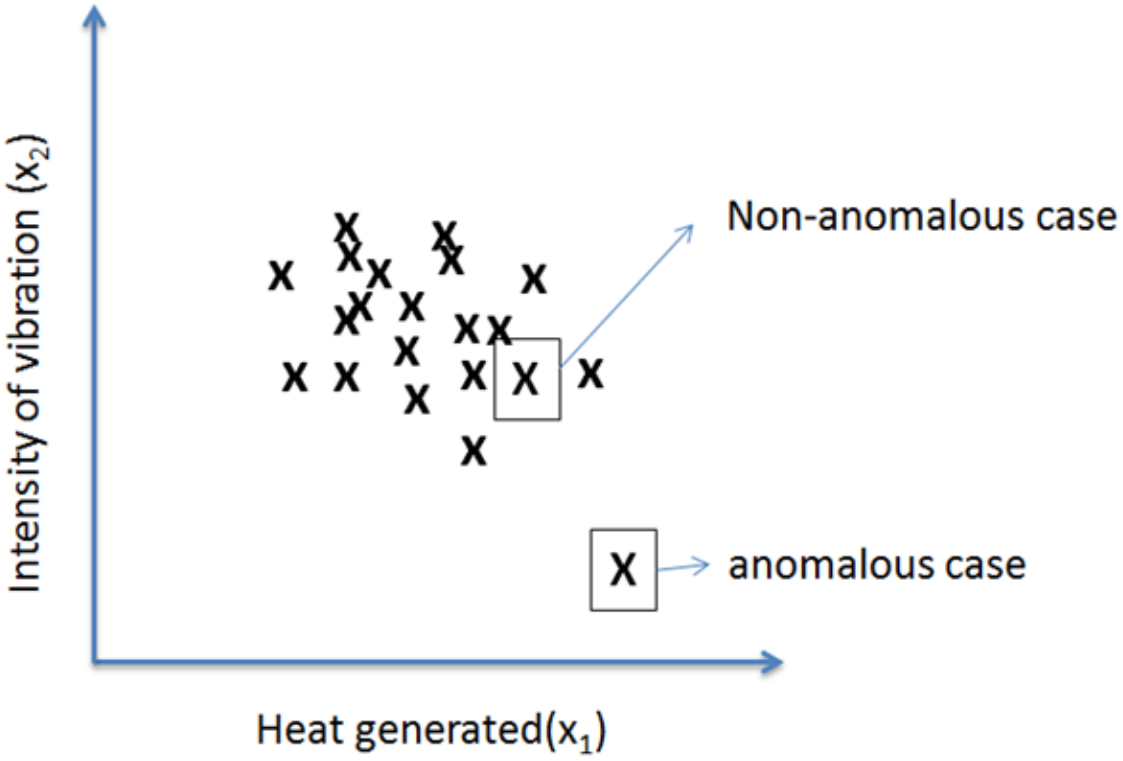
$$\sum_{i=13}^{25}\binom{25}{i}\varepsilon^{i}(1-\varepsilon)^{25-i} = 0.06$$

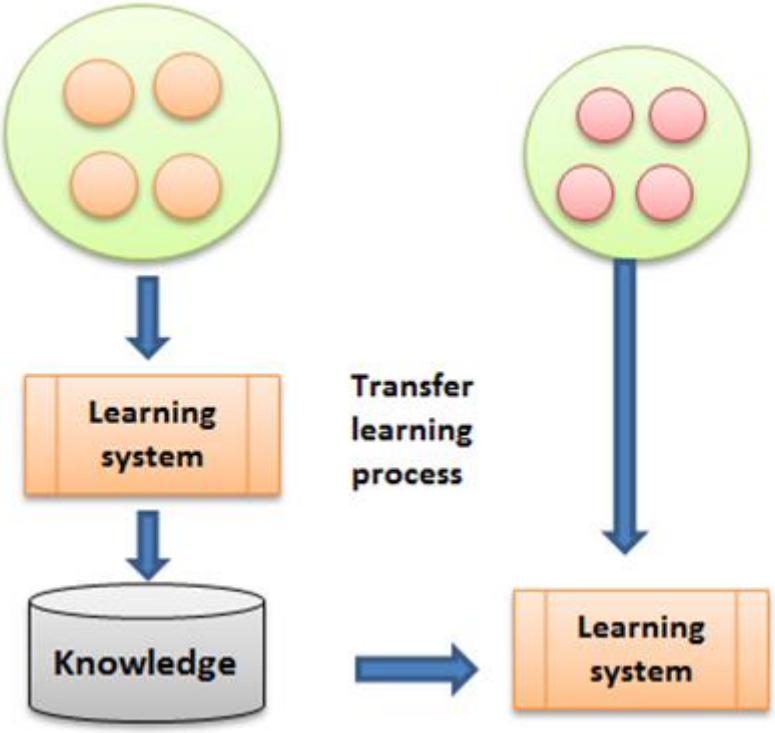


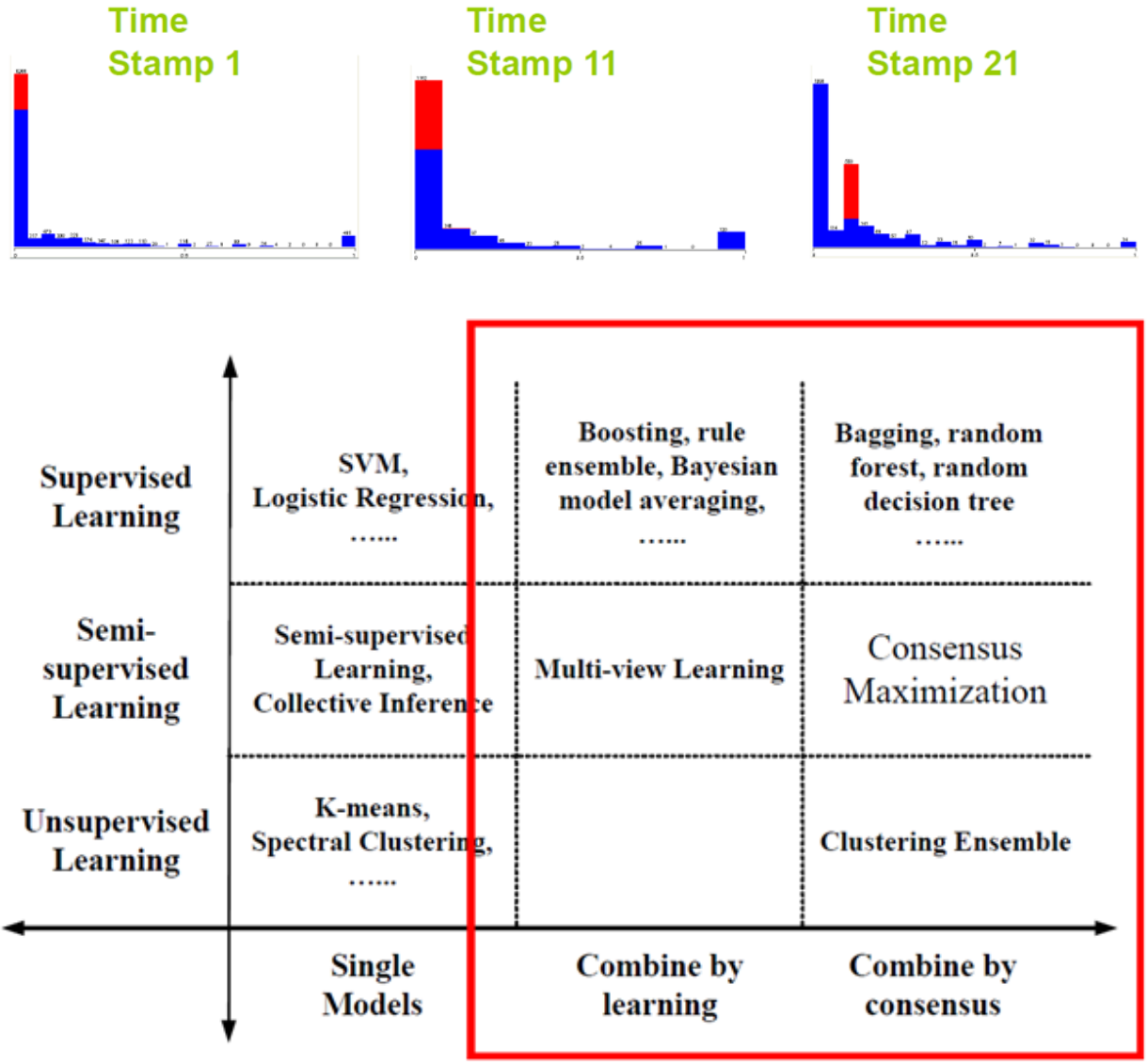

The base classifiers are identical (perfectly correlated)

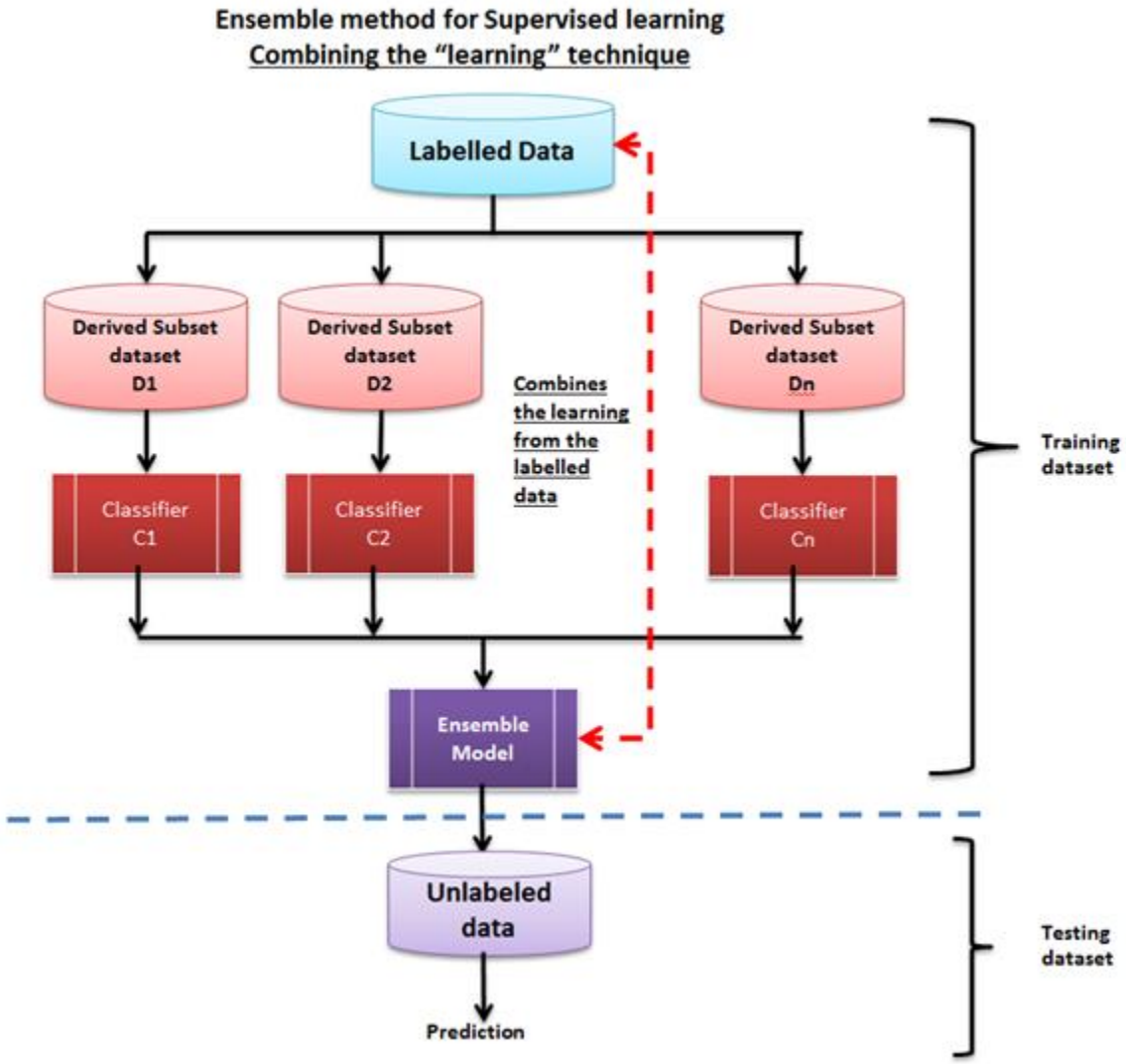
The base classifiers are independent

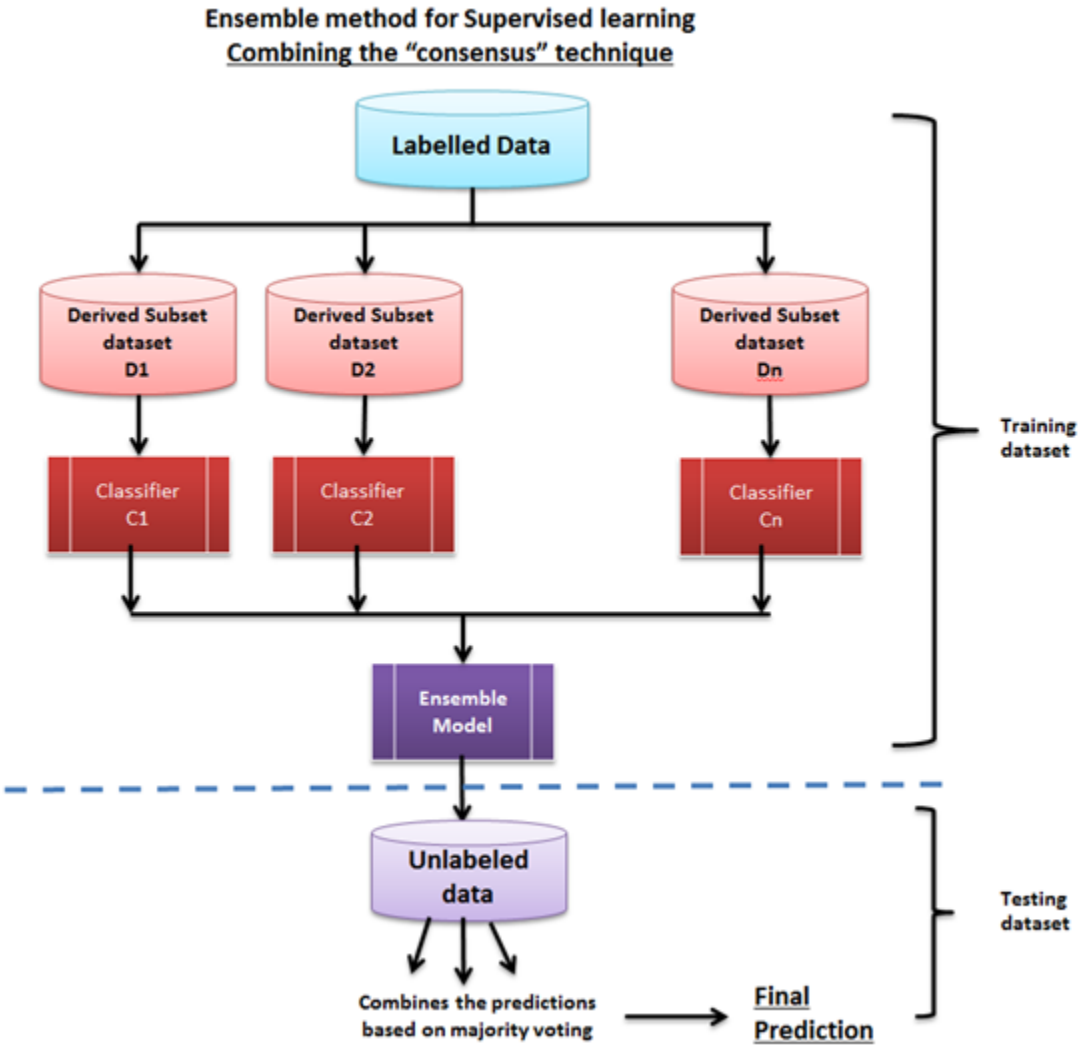


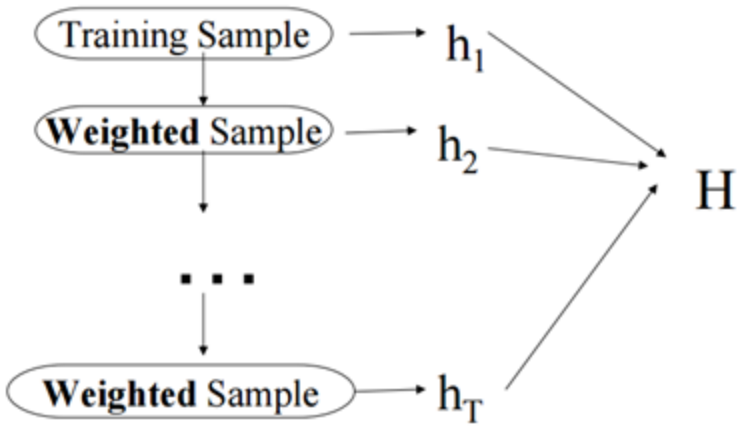

Intensity of vibration ($x_2$)
Heat generated($x_1$)
Non-anomalous case
anomalous case


Learning system
Knowledge
Transfer learning process
Learning system


Time Stamp 1
Time Stamp 11
Time Stamp 21
Supervised Learning
Semi-supervised Learning
Unsupervised Learning
SVM, Logistic Regression, ……
Boosting, rule ensemble, Bayesian model averaging, ……
Bagging, random forest, random decision tree ……
Semi-supervised Learning, Collective Inference
Multi-view Learning
Consensus Maximization
K-means, Spectral Clustering, ……
Clustering Ensemble
Single Models
Combine by learning
Combine by consensus


Ensemble method for Supervised learning
Combining the "learning" technique
Labelled Data
Derived Subset dataset D1
Derived Subset dataset D2
Derived Subset dataset Dn
Combines the learning from the labelled data
Classifier C1
Classifier C2
Classifier Cn
Ensemble Model
Training dataset
Unlabeled data
Prediction
Testing dataset


Ensemble method for Supervised learning
Combining the "consensus" technique
Labelled Data
Derived Subset dataset D1
Derived Subset dataset D2
Derived Subset dataset Dn
Classifier C1
Classifier C2
Classifier Cn
Ensemble Model
Training dataset
Unlabeled data
Testing dataset
Combines the predictions based on majority voting
Final Prediction


Training Sample
Weighted Sample
Weighted Sample
h1
h2
hT
H

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

**Given** $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \epsilon X$, $y_i \epsilon \{-1, +1\}$

**Initialise** weights $D_1(i) = 1/m$

**Iterate** $t=1, \ldots, T$:

- Train weak learner using distribution ***Dt***
- Get weak classifier: $h_t : X \rightarrow R$
- Choose $\alpha_t \varepsilon R$
- Update: $D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_t h_t(x_t))}{Z_t}$
  - where Zt is a normalization factor (chosen so that $D_{t+1}$ will be a distribution), and $\alpha_t$:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$$

**Output** – the final classifier

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

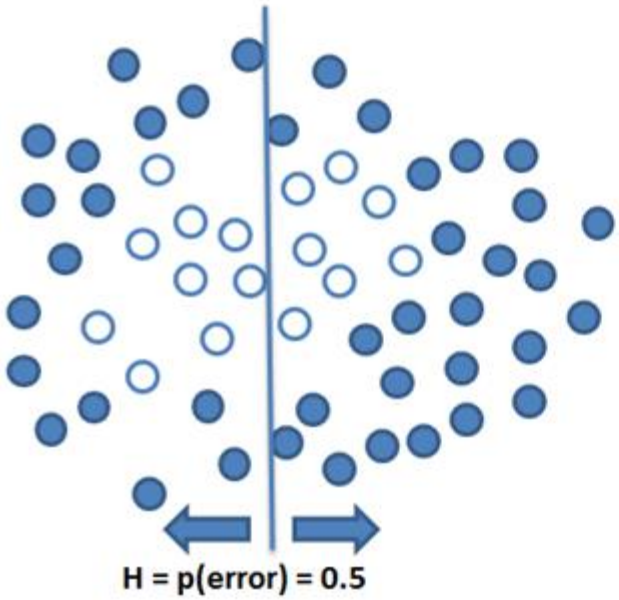
**Data points class label**

**$y_t = +1$**

**$y_t = -1$**

**Data points weight value**

**$w_t = 1$**



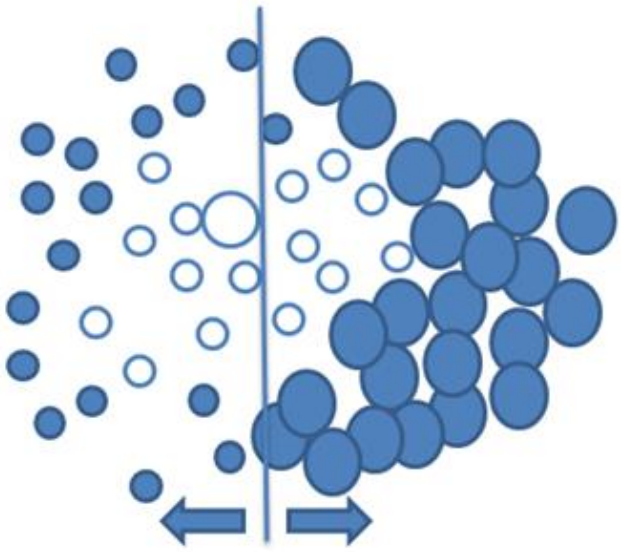

Data points class label

$y_t = +1$

$y_t = -1$

Data points weight value

$w_t = 1$



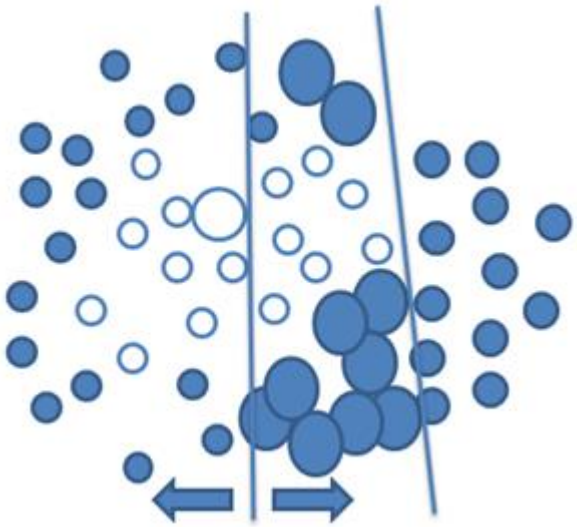
Data points class label

$y_t = +1$

$y_t = -1$

Weight value updated

$w_t = w_t \exp\{-y_t, H\}$



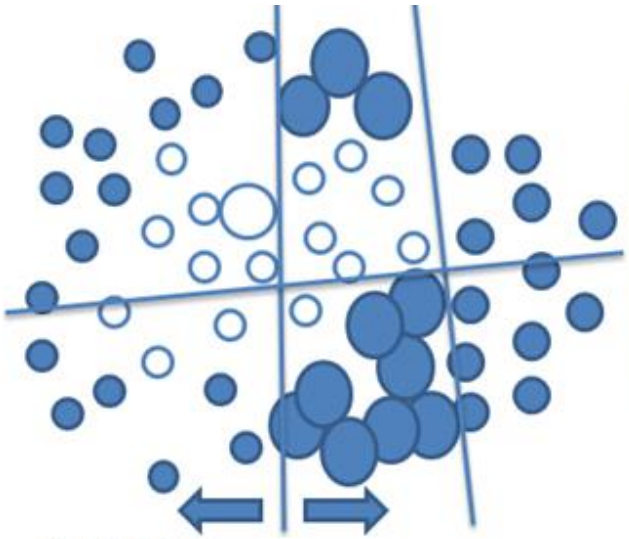
Data points class label

$y_t = +1$

$y_t = -1$

Weight value updated

$w_t = w_t \exp\{-y_t, H\}$



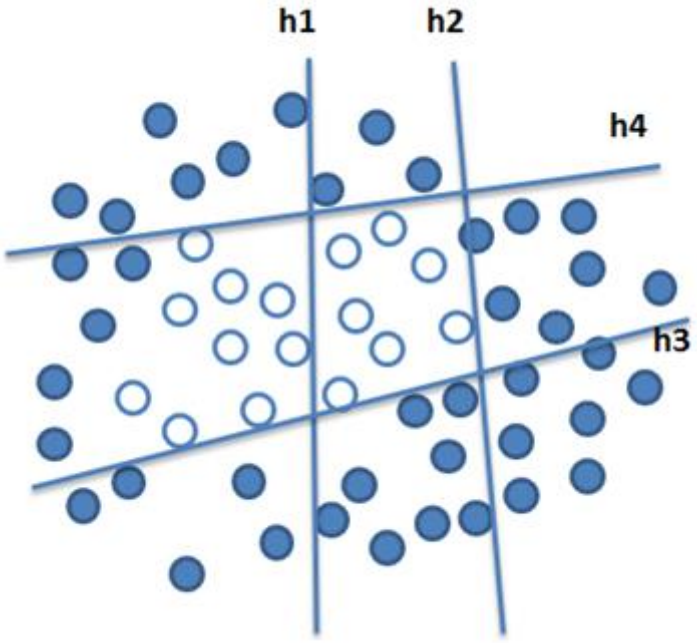
Data points class label

$y_t = +1$

$y_t = -1$

Weight value updated

$w_t = w_t \exp\{-y_t, H\}$




BAGGING

**Training phase**

1. Initialize the parameters
   - $\mathcal{D} = \emptyset$, the ensemble.
   - $L$, the number of classifiers to train.
2. For $k = 1, \ldots, L$
   - Take a bootstrap sample $S_k$ from $\mathbf{Z}$.
   - Build a classifier $D_k$ using $S_k$ as the training set.
   - Add the classifier to the current ensemble, $\mathcal{D} = \mathcal{D} \cup D_k$.
3. Return $\mathcal{D}$.

**Classification phase**

4. Run $D_1, \ldots, D_L$ on the input $\mathbf{x}$.
5. The class with the maximum number of votes is chosen as the label for $\mathbf{x}$.

the expected discrepancy between the estimated and true function

$$E\left[\left(\hat{f}(X) - E[f(X)]\right)^2\right]$$

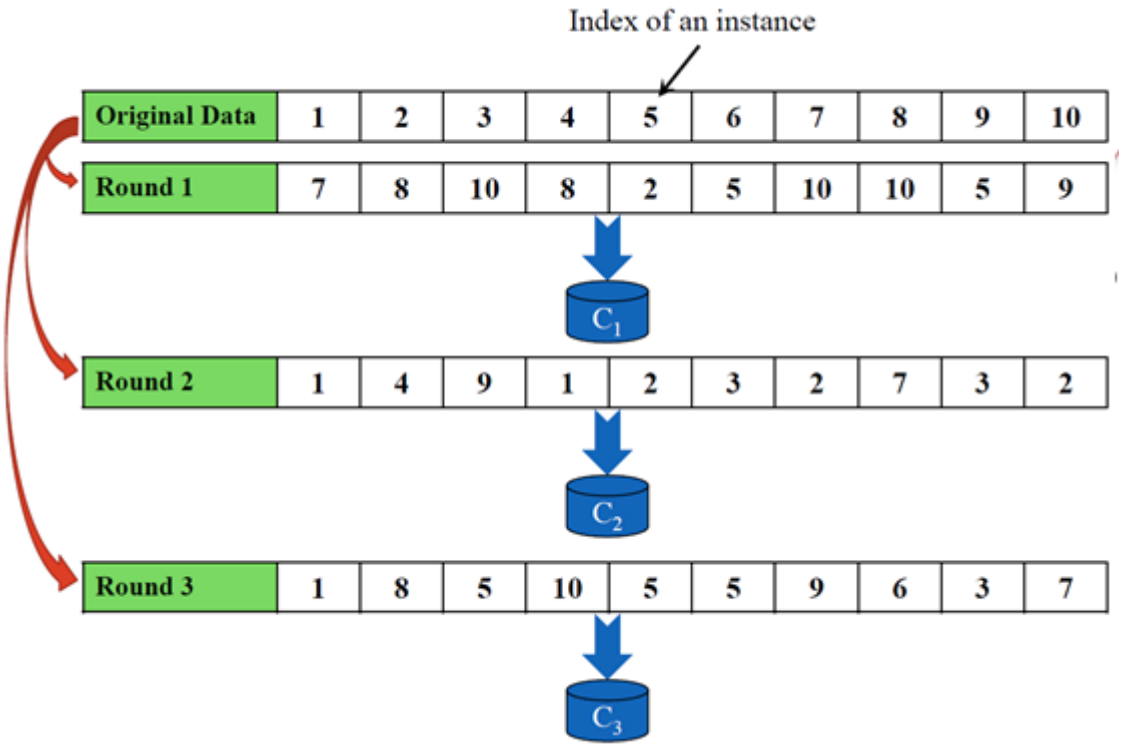
# Expected Error = Variance + Bias

is squared discrepancy between averaged estimated and true function

$$E\left[\left(\hat{f}(X) - E\left[\hat{f}(X)\right]\right)^2\right]$$

is expected divergence of the estimated function vs. its average value

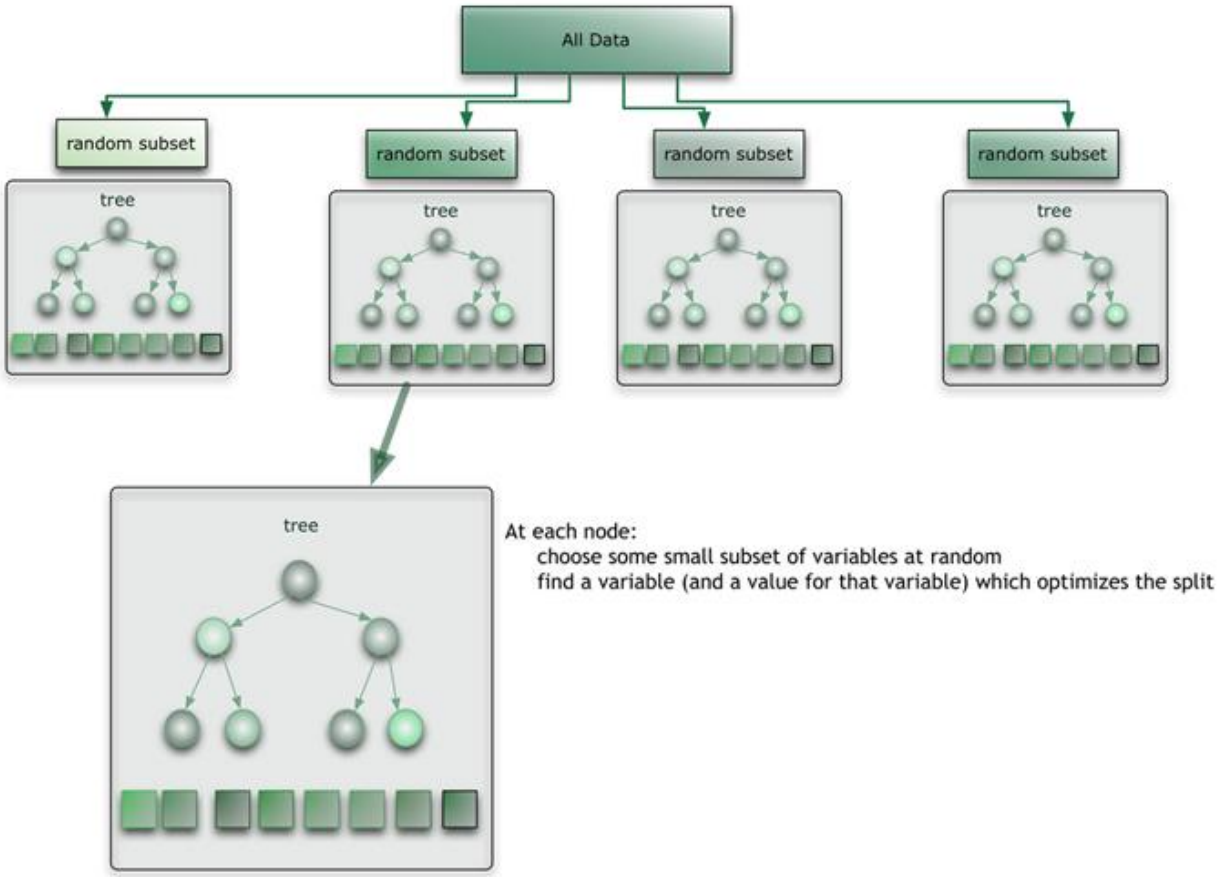
$$\left(E\left[\hat{f}(X)\right] - E[f(X)]\right)^2$$



**Require:** $I$ (an inducer), $T$ (the number of iterations), $S$ (the training set), $d$ (weighting distribution).

**Ensure:** $M_t; t = 1, \ldots, T$

1: $t \leftarrow 1$
2: **repeat**
3: $S_t \leftarrow S$ with random weights drawn from $d$.
4: Build classifier $M_t$ using $I$ on $S_t$
5: $t++$
6: **until** $t > T$



**Algorithm 1** Friedman's Gradient Boost algorithm

**Inputs:**

- input data $(x, y)_{i=1}^{N}$
- number of iterations $M$
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

**Algorithm:**

1: initialize $\widehat{f}_0$ with a constant
2: **for** $t = 1$ to $M$ **do**
3: compute the negative gradient $g_t(x)$
4: fit a new base-learner function $h(x, \theta_t)$
5: find the best gradient descent step-size $\rho_t$:

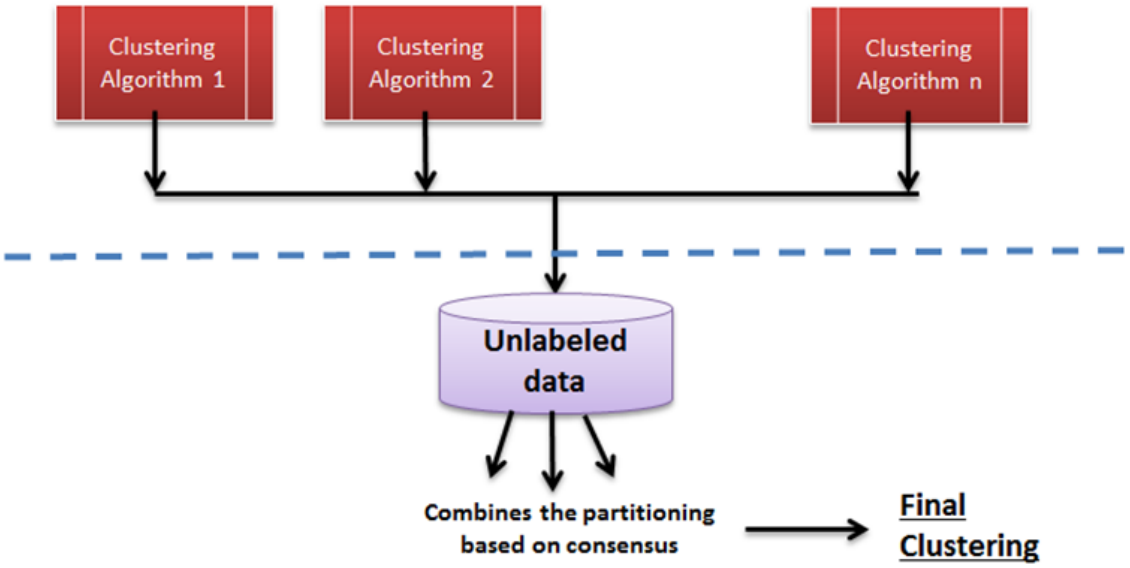
$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi\left[y_i, \widehat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right]$$

6: update the function estimate:

$$\widehat{f}_t \leftarrow \widehat{f}_{t-1} + \rho_t h(x, \theta_t)$$

7: **end for**

**Ensemble method for Unsupervised learning**
**Combining the "consensus" technique**




base clustering models

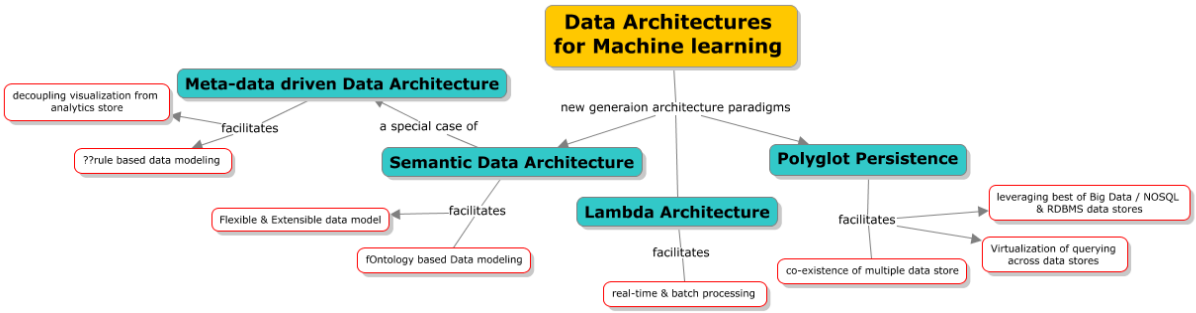
objects

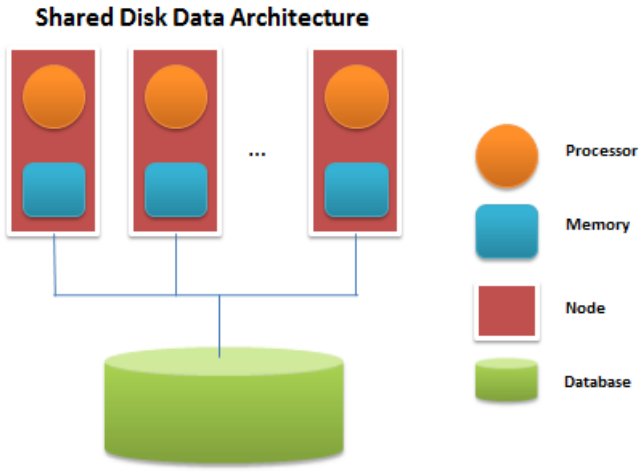
| | $\mathcal{C}_1$ | $\mathcal{C}_2$ | $\mathcal{C}_3$ | $\mathcal{C}$ |
|---|---|---|---|---|
| $v_1$ | 1 | 1 | 1 | 1 |
| $v_2$ | 1 | 2 | 2 | 2 |
| $v_3$ | 2 | 1 | 1 | 1 |
| $v_4$ | 2 | 2 | 2 | 2 |
| $v_5$ | 3 | 3 | 3 | 3 |
| $v_6$ | 3 | 4 | 3 | 3 |

they may not represent the same cluster!

The goal: get the consensus clustering

# CHAPTER 14




**Archives**
Scanned documents, statements, medical records, e-mails etc..

**Media**
Images, video, audio etc.

**Data Storages**
RDBMS, NoSQL, Hadoop, file systems etc.

**Docs**
XLS, PDF, CSV, HTML, JSON etc.

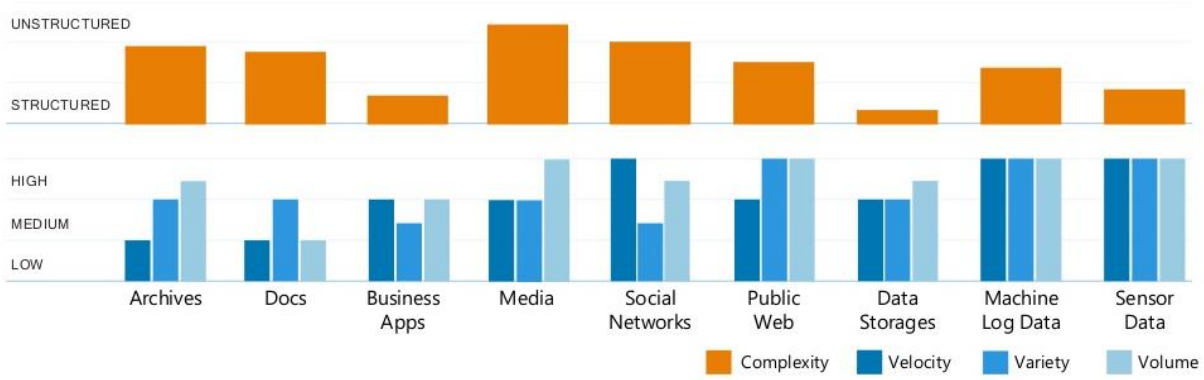
**Social Networks**
Twitter, Facebook, Google+, LinkedIn etc.

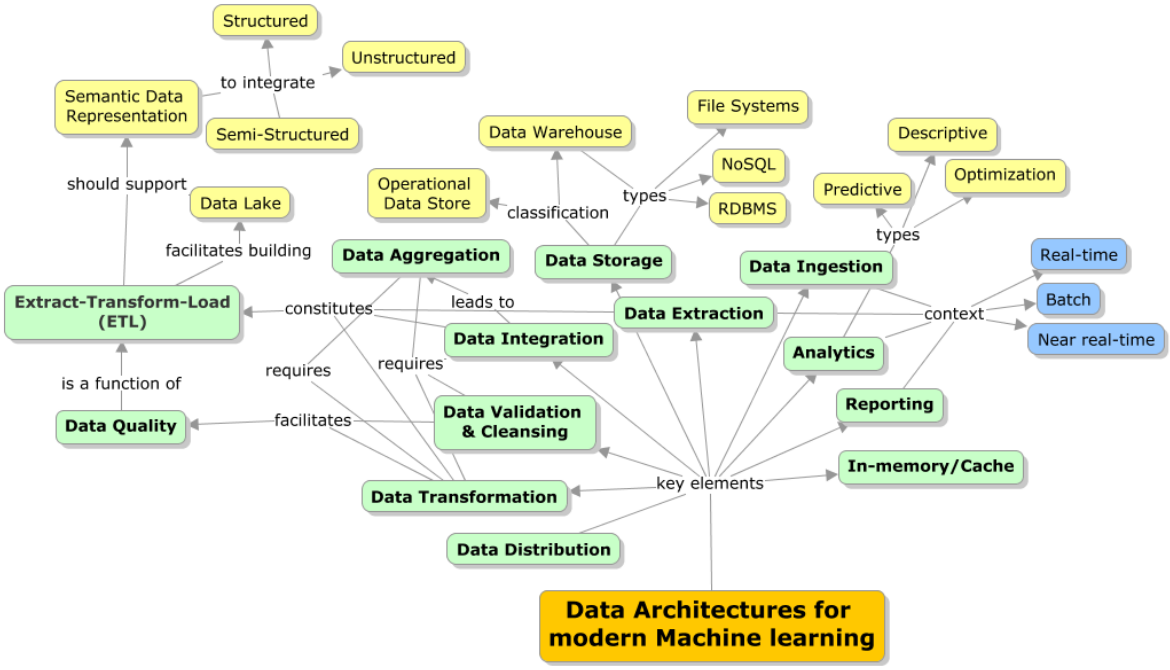
**Machine Log Data**
Application logs, event logs, server data, CDRs, clickstream data etc.

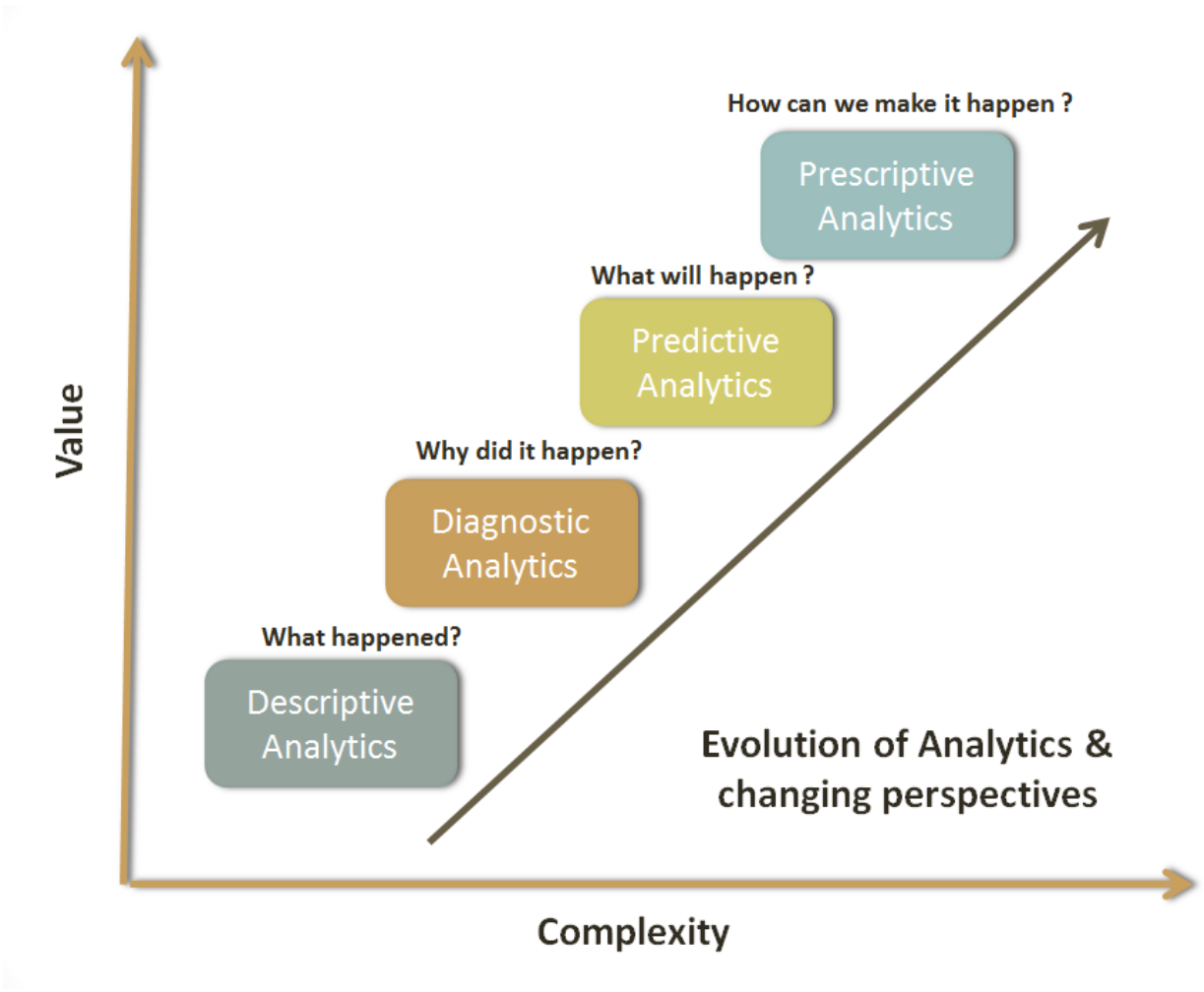
**Business Apps**
CRM, ERP systems, HR, project management etc.

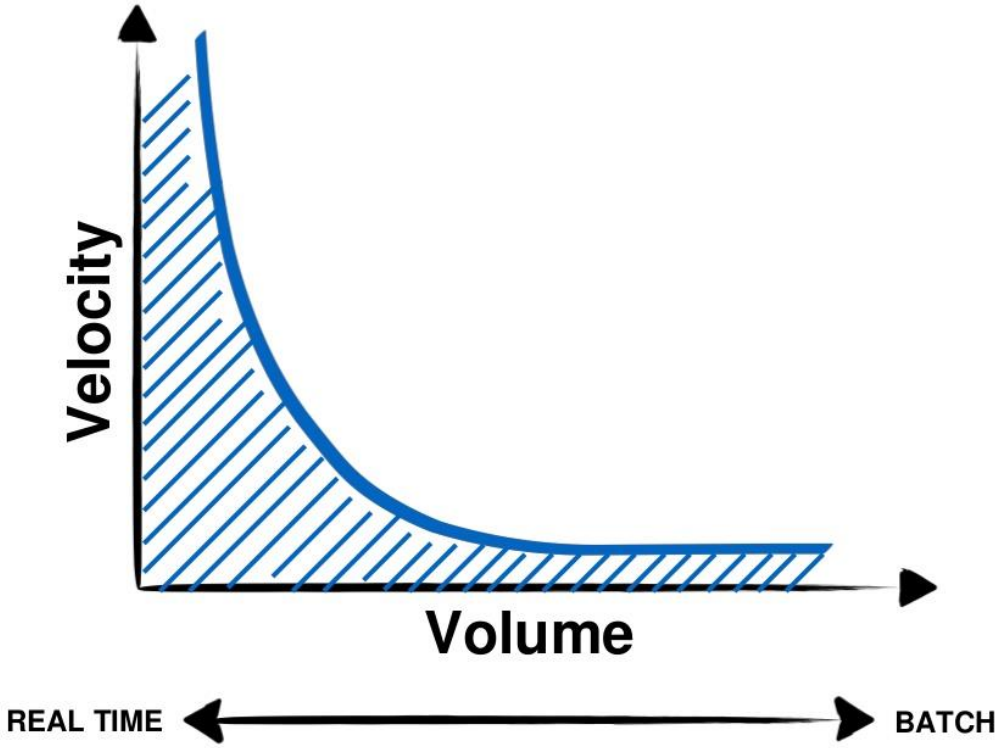
**Public Web**
Wikipedia, news, weather, public finance etc

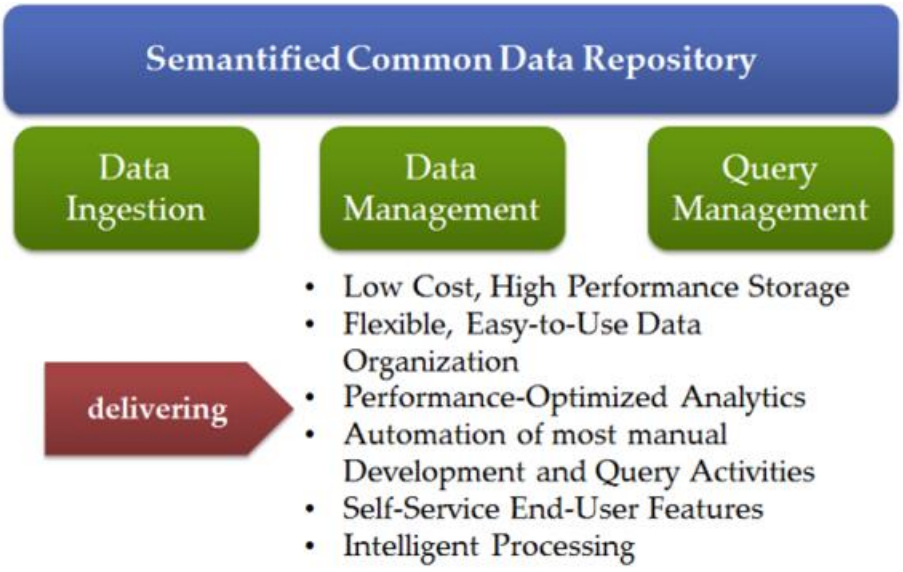
**Sensor Data**
Smart electric meters, medical devices, car sensors, road cameras etc.

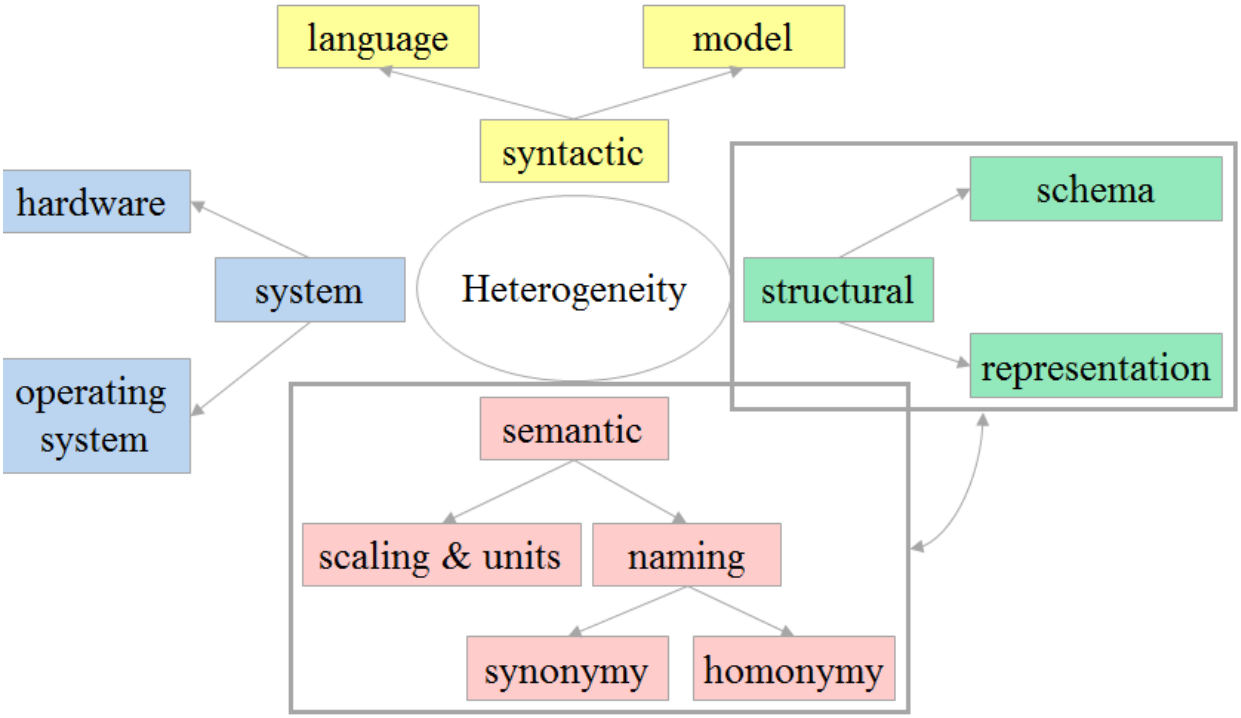

UNSTRUCTURED
STRUCTURED
HIGH
MEDIUM
LOW
Archives
Docs
Business Apps
Media
Social Networks
Public Web
Data Storages
Machine Log Data
Sensor Data
Complexity
Velocity
Variety
Volume

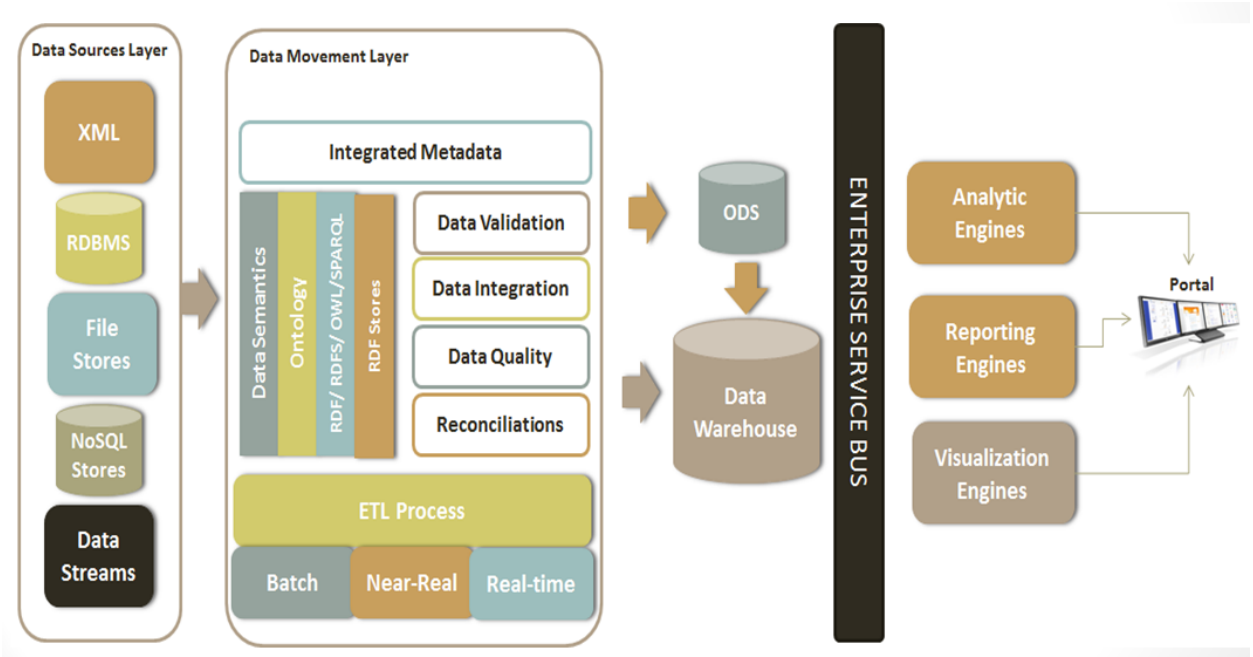

Structured
Unstructured
to integrate
Semantic Data Representation
Semi-Structured
should support
Data Lake
facilitates building
Extract-Transform-Load (ETL)
is a function of
Data Quality
Data Warehouse
File Systems
NoSQL
RDBMS
types
Operational Data Store
classification
Descriptive
Predictive
Optimization
types
Data Aggregation
Data Storage
Data Ingestion
Real-time
Batch
Near real-time
context
constitutes
leads to
Data Extraction
Data Integration
Analytics
Reporting
requires
requires
facilitates
Data Validation & Cleansing
In-memory/Cache
key elements
Data Transformation
Data Distribution
Data Architectures for modern Machine learning

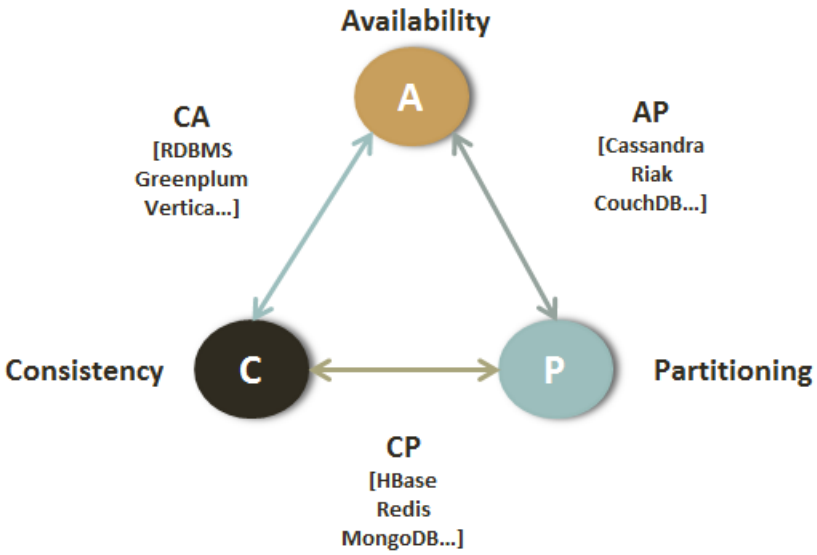

How can we make it happen ?
Prescriptive Analytics
What will happen ?
Predictive Analytics
Why did it happen?
Diagnostic Analytics
What happened?
Descriptive Analytics
Value
Complexity
Evolution of Analytics & changing perspectives


Velocity
Volume
REAL TIME
BATCH

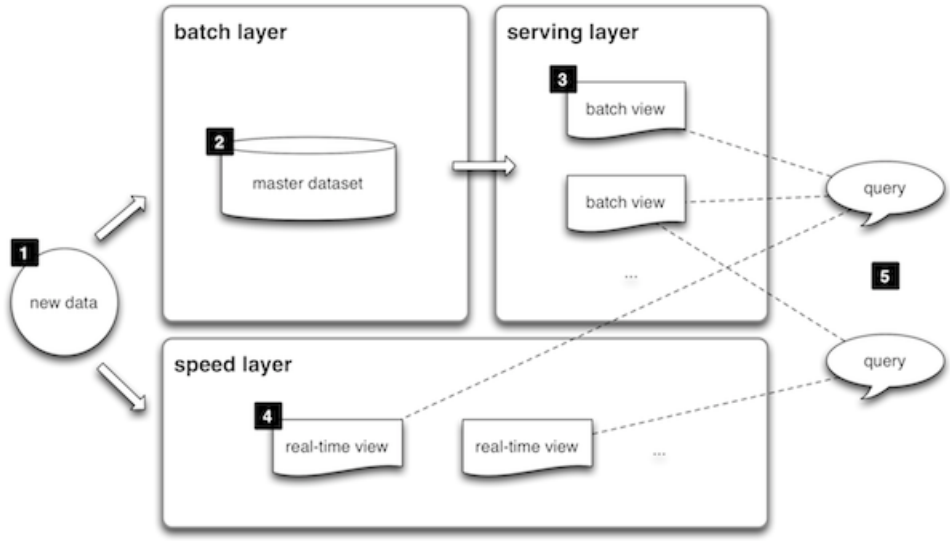

Semantified Common Data Repository
Data Ingestion
Data Management
Query Management
delivering
Low Cost, High Performance Storage
Flexible, Easy-to-Use Data Organization
Performance-Optimized Analytics
Automation of most manual Development and Query Activities
Self-Service End-User Features
Intelligent Processing


language
model
syntactic
hardware
system
operating system
Heterogeneity
schema
structural
representation
semantic
scaling & units
naming
synonymy
homonymy


Data Sources Layer
XML
RDBMS
File Stores
NoSQL Stores
Data Streams
Data Movement Layer
Integrated Metadata
Data Semantics
Ontology
RDF/ RDFS/ OWL/SPARQL
RDF Stores
Data Validation
Data Integration
Data Quality
Reconciliations
ETL Process
Batch
Near-Real
Real-time
ODS
Data Warehouse
ENTERPRISE SERVICE BUS
Analytic Engines
Reporting Engines
Visualization Engines
Portal


Availability
A
CA
[RDBMS
Greenplum
Vertica...]
AP
[Cassandra
Riak
CouchDB...]
Consistency
C
P
Partitioning
CP
[HBase
Redis
MongoDB...]


Speculative Retailers Web Application
User sessions
Redis
Financial Data
RDBMS
Shopping Cart
Riak
Recommendations
Neo4J
Product Catalog
MongoDB
Reporting
RDBMS
Analytics
Cassandra
User activity logs
Cassandra


batch layer
serving layer
speed layer
1
new data
2
master dataset
3
batch view
batch view
...
4
real-time view
real-time view
...
5
query
query