Preparation :

" Big Data Analytics " :

Chapter #1 :

- Science of information -

Data Science :

we can define data science as the management & analysis of data sets & extraction of useful information from them & understanding the system that produces the data.

One of the research problem in big data is big data classification. This is caused by the data characteristics such as volume, velocity & variety.

Technological Dilemma :

→ non-existence of techology that can manage & analyze dynamically growing data.

→ lack of intelligent approaches that can select suitable technique from many design choices.

→ if we invest in expensive & modern technology assuming that data in our hand is big data & later we find out it is not big data then investment is basically lost.

Technological Advancements:

→ current advancement in technology includes the modern file systems and the distributed machine learning. One of such technology is called as **Hadoop**.

→ among several machine techniques in the libraries most of them are based on classical models & algorithms may not be suitable for big data.

Big Data Paradigm (example ﷺ):

→ the goal of a system is to observe an environment & learn its characteristics, to make accurate decisions. for example goal of a network intrusion detection is to learn traffic characteristics & detect intrusions to improve the security of a computer network.

→ Similarly goal of a wireless sensor network is to monitor changes in weather patterns for forecasting.

Facts & Statistics of a System:

we need clear definition for two important terms data & knowledge & for three operations, physical, mathematical & logical operations.

Data :

→ Data can be described as the hidden digital facts that the monitoring system collects.

→ Hidden digital facts are the digitalized that are not obvious to the system without further comprehensive processing.

→ One of the most important requirement for data is its format. Forexample data could be represented mathematicaly or in two-dimension tabular representation.

→ another requirement is type of data. Forexample data could be labeled or not labeled. In labeled data, the digital facts are not hidden & can be used for training the machine-learning techniques & in unlabeled data, the digital facts are hidden & can be used for testing or validation as part of machine learning approach.

Knowledge :

→ knowledge can be described as the learned information acquired from data.

→ Forexample the knowledge could be the detection of pattern in the data, the classification of varities of pattern in data, the calculation of unknown statistical distribution or computation of corelation of data.

→ It forms the response for the system & it is called the "knowledge set" or "response set" (sometimes called "labeled set")

Physical Operations :

→ Physical operation describes the steps involved in the process of data capture, data storage data manipulation & data visualization.

Mathematical Operations :

describe the theory and application of appropriate mathematical & statistical techniques & tools required for transformation of data into knowledge.

→ the transformation can be written as a knowledge function $f : D \Rightarrow K$ where the set D stands for data domain & set K stands for knowledge or response set.

→ if data is structured then the execution of these operations are not difficult. Even if structured data grows exponentialy these operations are not difficult because they can be carried out using existing resources & tools. Hence, size of data doesn't matters in case of structured data in general.

Logical Operations :

Logical operations describes the logical arguments, justification and interpretations of the knowledge, which can be used to derive the meaningful facts. Forexample the knowledge function

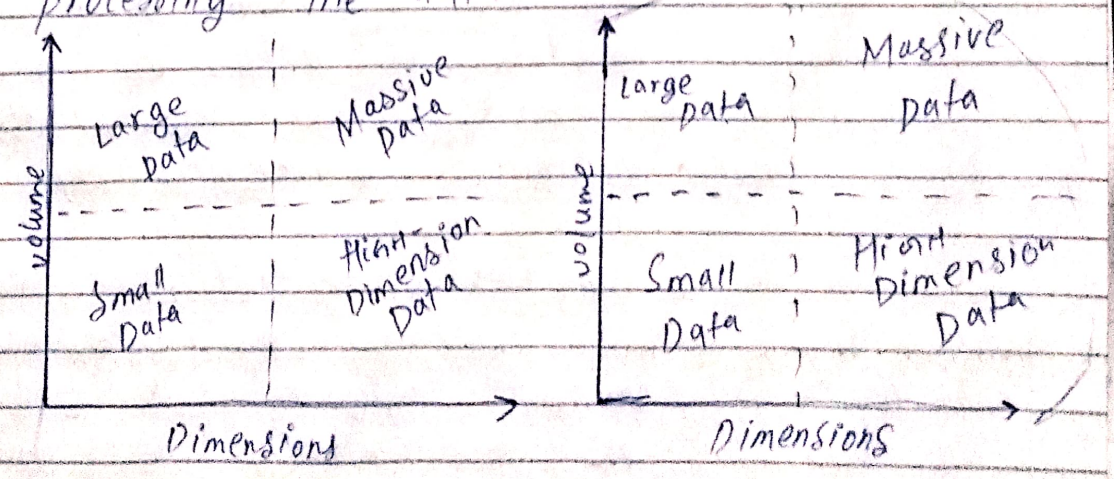
$$f : D \Rightarrow K$$

can divide the data domain & provide data patterns & then the logical operations & arguments must be used to justify & interpret the class types from the pattern.

Scenario

Millions of events (n) (no. of observations)
time period (t)
number of features (p)

p controls the time complexity of processing the data.



Data Rate (velocity) : Ratio blw n & t

Data Representation:

A dataset may be in mathematical or tabular form. The tabular form is visual & can be easily understood by non-expert.

Machine Learning Paradigm:

Machine learning is about the exploration & development of a mathematical model & algorithms to learn from data. Its paradigm focuses on classification objectives & consist of modeling an optimal mapping b/w the data domain & knowledge set. & learning the developing algorithm. The classification is also called supervised learning which requires a training (labeled) data set & a validation data set & a test dataset.

Modeling & Algorithms:

→ The term modeling refers to both mathematical & statistical modeling of data. The goal of modeling is to develop a parametrized mapping between data domain & response set.

→ For a compuler scientist, the term algorithm means step by step systematic instruction for a computer to solve a problem. In ML the modeling itself, may have several algorithms to derive a model, however term algorithm refers to a learning algorithm.

Supervised & Unsupervised Learning:

In supervised learning, the classes are known & class boundaries are well defined in the given (training) dataset. & the learning is done using these classes. Hence, it is called classification.

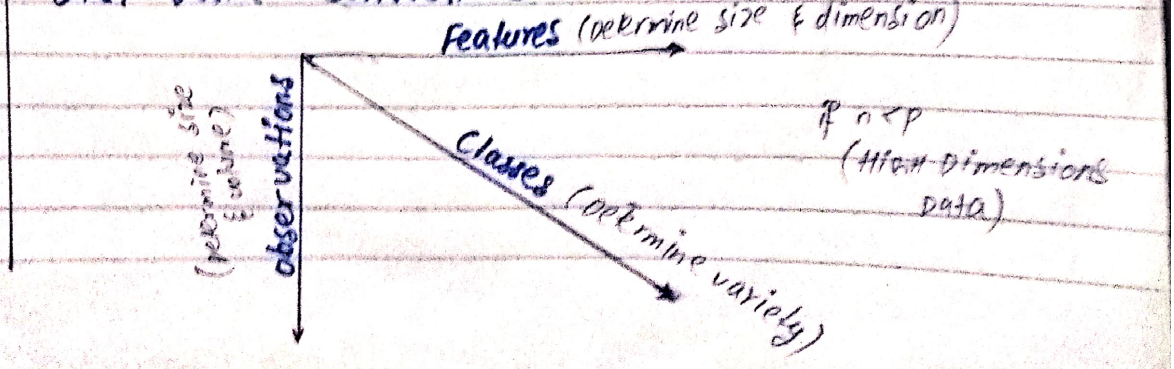
→ In unsupervised learning classes or class boundaries are not known, & hence these labels also learned & classes are defined on base of this, Hence class boundaries are statistical & not sharply defined & it is called clustering.

CHAPTER #02

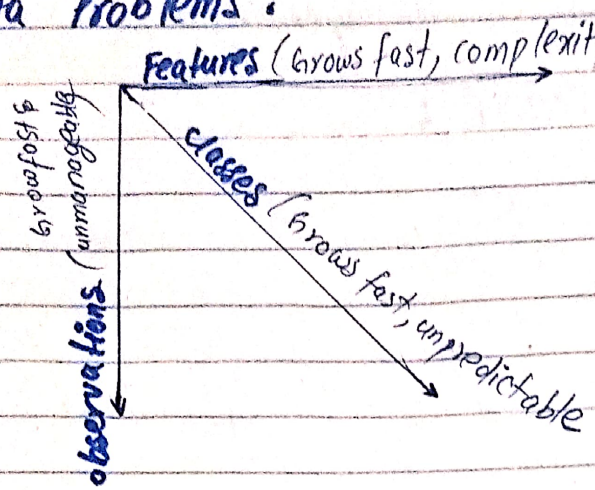
Big Data Essentials

"Philip Russom" defined the term big data analytics.

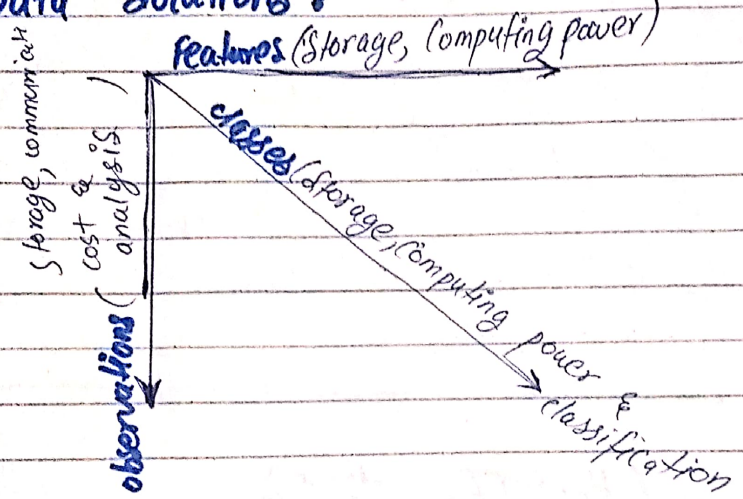
Big Data Controllers:



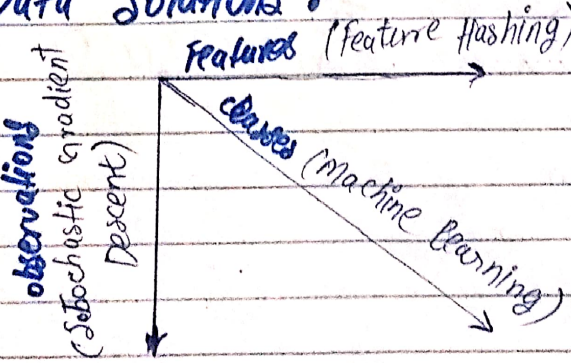
Big Data Problems:



Big Data Solutions:

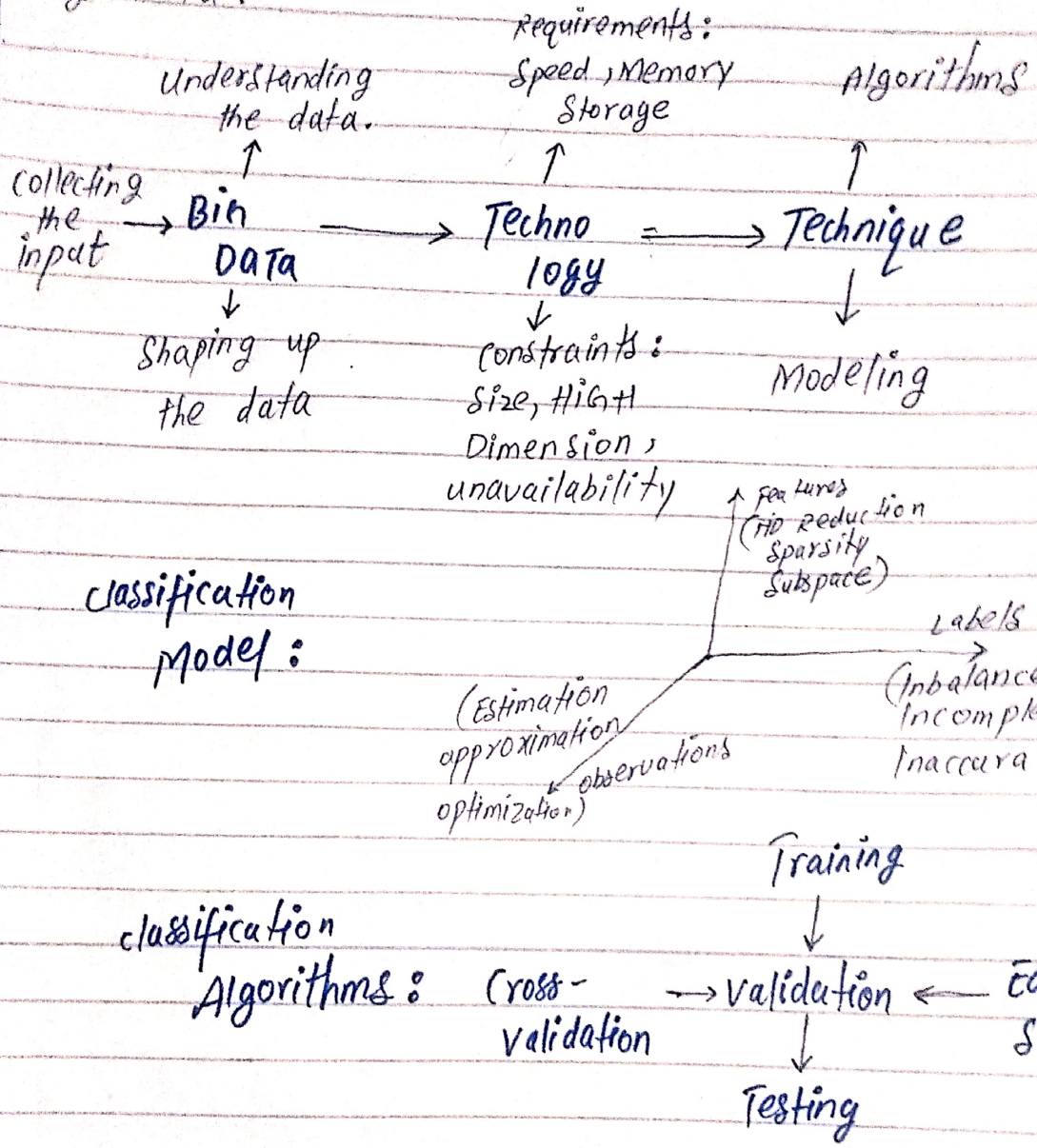


Big Data Solutions:



Big Data Classification:

Big data classification is the process of classifying big data under the problems & challenges under the problems & challenges introduced by the controller of big data.



Representation Learning:

are useful for understanding and shaping the data. These techniques requires statistical measures & processes.

2 helps in dimensionality reduction objectives in machine learning.

Distributed File System:
are suitable for big data management, processing & analysis
Hadoop

- Approximation
- Estimation
- optimation

↑

Error Characteristics

↑

Data Domain → Mapping f → Response set

↓ Domain Characteristics

- Dimensionality
- Sparsity
- Subspace

↓ class Characteristics

- Imbalance
- Incomplete
- Inaccurate

Classification Modeling:

Class Characteristics:

Inbalanced, Incomplete & inaccurate can be defined in response set portion of modeling objective. These values are influenced by the big data controllers.

Inbalanced :

if we have more observations in one class than other then it is said to be inbalanced.

(Balance)

(Inbalanced)

Incomplete :

if a class is missing information & is incomplete.

(complete)

(Incomplete)

inaccurate :

if class observations are labeled incorrectly.

(inaccurate)

Error Characteristics :

can be defined in mapping portion of modeling process.

- **Estimation error :**

error between true model & model we assumed is called estimation error. (impacts accuracy of classification model)

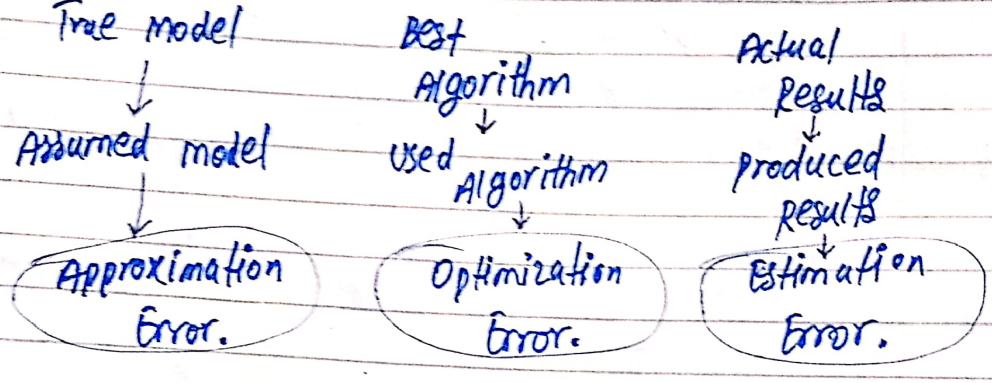
- **Optimization error :**

error between best algorithm & algorith we supposed is called optimization error.

(")

Estimation Error :

if we use best algorithm & true model we get actual results but if our assumed model produces different results. This error is called as the estimation error.



Domain Characteristics :

Dimensionality :

The number of features determine the dimensionality of data.

Subspace :

Some features may not be relevant & they may not contribute to pattern in data. This can lead to dimensionality reduction & new space with fewer features is called subspace.

Sparsity : ( )

presence of null values in the features.

(2, 0, 0), (0, 8, 0), (0, 0, 3)

Classification Algorithm :

Training :

The training phase provides an algorithm to train the model. The parameters of a machine learning model are estimated, approximated & optimized using labeled data set, class characteristics & domain characteristics. The data used in training is labeled dataset.

Validation :

Validation phase provides an algorithm to validate the effectiveness of the model using another dataset which was not used in training phase.

→ Data set is called validation set.

→ Quantitive measure plays important role in validation phase. (entropy & root-mean-squared error)

Testing :

→ simple phase

→ provides algorithm to test if trained & cross validated models works using another data set which was not used in taining & validation.

→ Several Qualitative measures are available for this purpose

(Accuracy, sensitivity, specificity, precision)