Chapter #1

Science of Information

* Knowledge ⇒ is also called response

Data Science

* management and analysis of data sets
* extraction of useful information
* understanding of the systems that produce data.

management: إدارة
the praers of deal with or controlling things

analysis: تحليل
detailed examination of the elements or structure of something

* The system can be a single unit formed by many interconnecting sub units.

e.g.
- a computer network
- a wireless sensor network

Chapter #1

Science of Information

* Knowledge ⇒ is also called response

Data Science

* management and analysis of data sets
* extraction of useful information
* understanding of the systems that produce data.

management: إدارة
the praers of deal with or controlling things

analysis: تحليل
detailed examination of the elements or structure of something

* The system can be a single unit formed by many interconnecting sub units.

e.g.
- a computer network
- a wireless sensor network

intrusion : دخل اندازی کرنا

* Some examples of systems:
  - Intrusion detection System
  - Climate Change Detection System
  - Public Space Intruder Detection System

Big Data:

These Real-world systems may produce massive amount of data, called big data, from many data sources that are highly complex, unstructured, and hard to manage, procesn and analyze.

* Big Data Classification → a research problem
  - Classification of different types of data
  - Extraction of useful information from massive and complex data sets

Big Data Characteristics

Volume | Velocity | Variety

Dilemma: a difficult situation or problem

Technological Dilemma:

* non-existance of a technology that can manage and analyze. dynamically growing massive data efficiently and extract useful information

* lack of intelligence approaches to select suitable techniques for solving big data problems.

- In this case, machine learning techniques like supervised learning and dimensionality reduction techniques are useful.

Technological Advancement:

* current technological advancement:
  - modern distributed file systems
  - distributed machine learning

* Hadoop → technology
  → faciliates distributed machine learning using external libraries like the scikit library to process big data.

* techniques based on classical models and algorithms may not be suitable for big data processing.

* techniques for big data classification:
  - Decision Tree learning
  - Deep learning

~~Big Data Paradigm~~

Distributed File System:

* Distributed File System (DFS) or network File system is any file system that allows access to files from multiple hosts sharing via a computer network

Distributed Machine learning:

* multi-node ML system that improve performance, increases accuracy and scales to larger input data sizes.

Paradigm: نموذج
a pattern or model

Big Data Paradigm:

Big Data Paradigm
- Big Data System
- Environment

* Goal of system → observe an environment and learn its characteristics to make accurate decisions.

e.g.
IDS is to learn traffic characteristics and detect intrusions to improve the security of a computer network.

* Environment → Generate ~~Facts~~ events
System → collect facts and statistics
↓ Transform
Knowledge
↓ Learn
event characteristics
↓ Predict
Environment characteristics

Facts and Statistics of a system:

Data:

* Hidden digital facts that the monitoring system collects

* Hidden digital facts → digitized facts that are not obvious to system without processing

* Requirements For Data:
  - Data Format i.e. mathematical or tabular
  - Data Type i.e labeled or unlabeled

Labeled Data:

* Digital facts not hidden

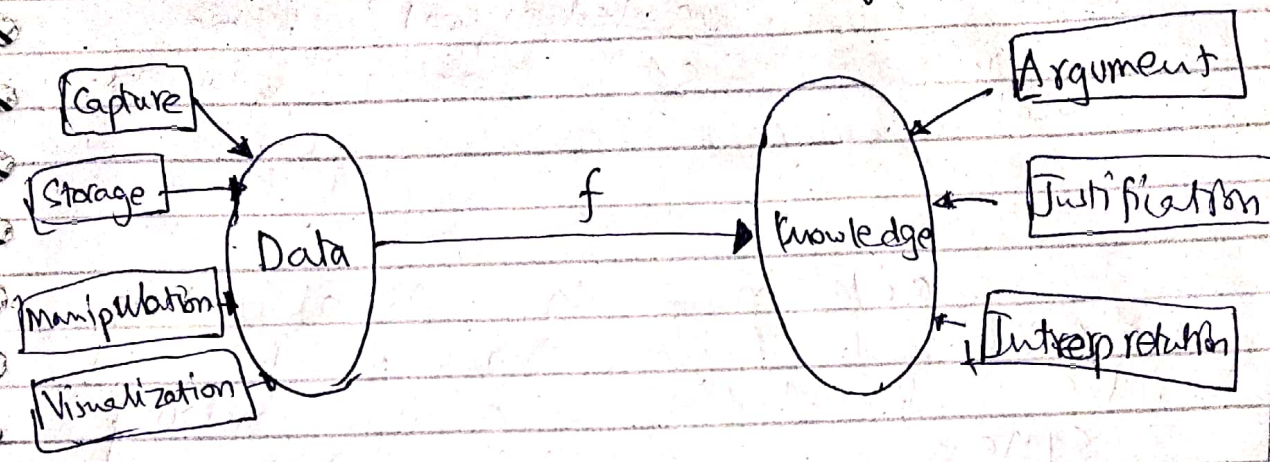
* can be used for training the machine learning techniques

Unlabeled Data:

* Digital facts are hidden

* can be used for testing or validation as a part of machine learning approach.

Transformation of Data into Knowledge



→ Justification:
The action of showing something to be right or reasonable

→ Argument:
a reason or set of reasons given in support of an idea, action or theory

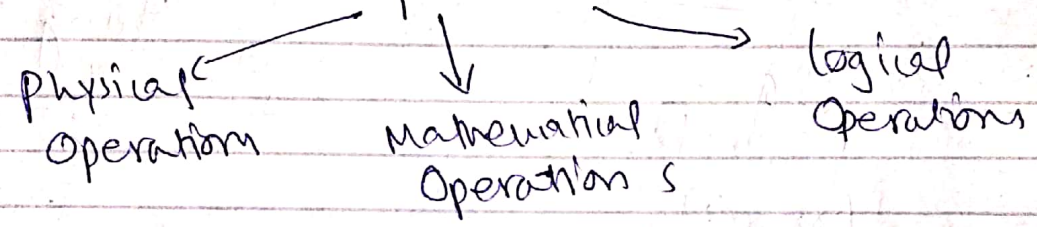
Knowledge:

* learned information from data

e.g.
- detailed detection of patterns in the data
- classification of variaties of patterns
- calculation of statistical unknown distribution

* form responses for system
* called "knowledge set" or "response set" or "labeled set"

⇒ In addition to these two elements (i.e., the data and the knowledge), a monitoring system needs three operations



Physical Operations:

* describe steps involve in process:
  - Data Capture
  - Data Storage
  - Data manipulation
  - Data Visualization

* Big Data cannot be solved using one file or single machine.
The indexing and distribution of data becomes necessary.

notes: a brief record of points or ideas written down as an aid to memory.

Sophisticated: having, revealing, developed to high degree of complexity

- Hadoop distributed file system
  ↳ uses Mapfedue framework
  ↳ helps generate sophisticated supervised learning models and algorithms for big data classification.

Mathematical Operations:

* describe the theory and applications of appropriate mathematical and statistical techniques
  +
  tools required for transformation of data into knowledge

* Transformation can be written as knowledge function:

$$f : D \rightarrow K$$

D → Data Domain
K → Knowledge

* ~~Iran~~ the size of data (structured) does not matter because they can be carried out using existing resories and tools.

Logical Operations:

* Describe:
  - Logical arguments
  - justification
  - interpretations of knowledge

* used to derive ~~meaningful~~ meaningful facts

Big Data Versus Regular Data:

* n, p and t are parameter of system which determines (characteristics of data) whether data is Big Data or Regular Data

* An element of a monitoring system's data can also called an "observation" (or an event)

* observations generally depend on ~~feature~~ many independent variables called "features"
  - features forms feature space.

n → events

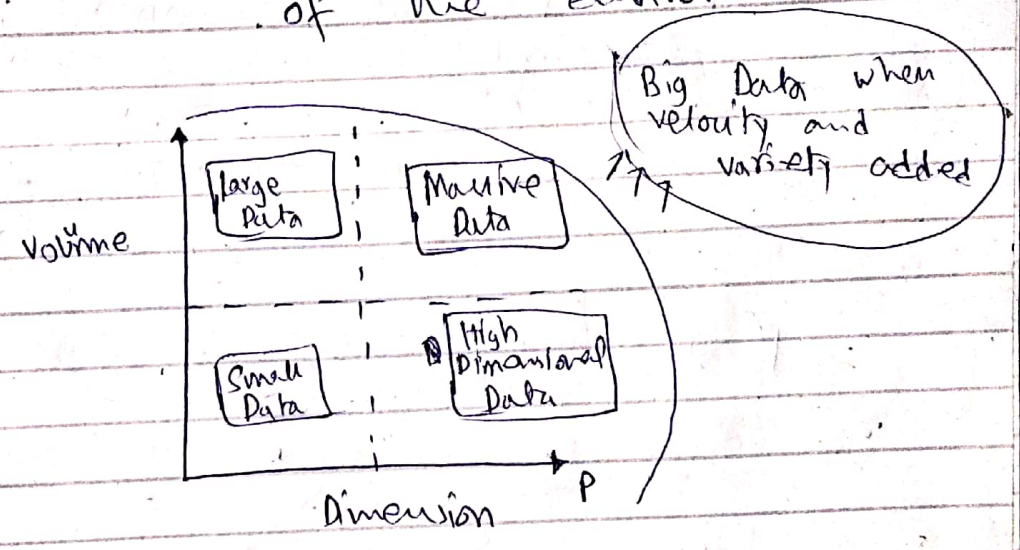
p → ~~no. of~~ features

t → time

dimensionality → no. of features of system

↘ Controls complexity of processing the data

features → ~~represents~~ characteristics of the environment



Big Data → Variety

↓ Velocity

↔ Volume

Dimensionality (no. of features)

Regular Data

↙ Volume

↓ Dimensionality (no. of features)

Class or Type of an event:

* changes in values of feature variables

* To determine correct class for an event, the event must be transformed into knowledge.

n → no. of observations captured by a system at time t
↓ determines
size of data set

p → no. of features
↓ determines
Dimensionality, Variety (no. of classes)

⇒ rate of velocity:

$$\frac{n}{t}$$

⇒ massive data becomes big data when velocity and variety added.

Data Representation:

* defined in mathematical or tabular form
* tabular form → visual

Machine Learning Paradigm:

Machine learning:

* exploration and development of mathematical models and algorithms to learn from data

focuses:

* classification objectives
* modeling an optimal mapping between data domain and knowledge set
* developing the learning algorithms

⇒ classification is also called supervised learning

⇒ requires:

- Training (labeled) data set
- validation data set
- Test Data set

Training data set:

→ help to find optimal parameters of a model.

Validation Data Set:

→ help avoid overfitting of model

Test Data Set:

→ helps determine the accuracy of the model

Supervised Machine learning:

* type of machine learning
* machine trained using labeled data set
* labeled data:
  Some input data is already tagged with correct output
* used for:
  Risk Assesment, Image Classification, Fraud detection, Spam filtering etc

Modeling and Algorithms:

* modeling refers to
  - mathematical modeling
  - Statistical modeling

Goal of Modeling:

* Developed parametrized mapping between the data domain and the response set

* Mapping could be parametrized function or parametrized process that learn the characteristics of a system from the input (labeled) data.

* In machine learning:
  algorithm → learning algorithm

Learning Algorithm:

* train, validate and test the model
* find optimal value for the ~~path~~ parameters

Supervised and Unsupervised:

* Supervised:
  - classes are known
  - class boundaries are well defined
  - learning is done from these classes (class

* ~~eto~~ called classification

Unsupervised learning:

- classes or class boundaries are not known

- class labels themselves learned and classes are defined based on this

- class boundaries:
  - ↳ statistical
  - ↳ not sharply defined
  - ↳

→ called clustering

Classification:

* labeled data (classes) are available to generate rules (classifiers) that can help to assign a label to new data that does not have labels

* labeled data → evaluating and validating machine-learning technique.

evaluating: to determine or fix the value of

Inshall Allah!
I will be cyber Security Engineer!

validating → to make legally valid
(logically correct)

* This mathematical function help us to define suitable classifier for the classification of the data:

$$f: R^l \Rightarrow \{0, 1, 2, \dots, n\}$$

* well-known classification techniques:
  - Support Vector Machine
  - Decision Tree
  - Random Forest
  - Deep Learning

Clustering:

* Data sets are available to generate rules but they are not labeled.
* can only derive approximate values that can help to label new data that do not have labels.
* data may only be clustered not classified.
* clustering can be defined with Data set ~~to~~ can - to - knowledge approximation rule.

Transformation:

$$f: R^l \Rightarrow \{\hat{0}, \hat{1}, \hat{2}, \dots, \hat{n}\}$$

* Several Clustering Algorithms:

- K-Means Clustering
- Gaussian Mixture Clustering
- Hierarchical clustering

Chapter#2 Big Data Essentials

Big Data Applications:

* Business Intelligence
* Network Intrussion Detection
*
*

Big Data Controllers:

1- Class Characteristics
2- Feature Characteristics
3- Observation Characteristics

* Whether data set is structured or unstructured may be determined by proper understanding of the controllers.

* Size of dataset = no. of observations (events)

* Observation determine volume and velocity data.

* Observation controls the classification issues that resulted from the volume and velocity of the big data.

* Horizontal axis → independent variables

* Features → independent variables
  → Generate Events

* Features determines the volume & dimensionality of big data.

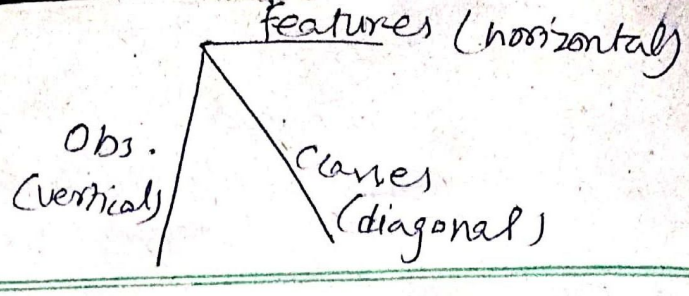
* no. of features = no. of dimensions of data set

* Features controls scalability of the data and the parameters.

* n (obs.) & p (feat.) determine characteristics of dimensionality.

* High Dimensional Data → $n < p$

orthogonally: statistically independent



* classes → types of events
* classes → determines the variety of big data
  ↳ helps to group data

Big Data Problems:

* Individualization and uncoordinated efforts of controllers can create problems in big Data.

individualization:
→ to make individual in character
intended for one person

uncoordinated:
→ no well organized
→ no able to move different parts of the body together well or easily

Classes:

* class (controller) contributes to the unpredictability of data.

* classes dependent but independent on system user's knowledge to the and the experience

confront: Lio Lu

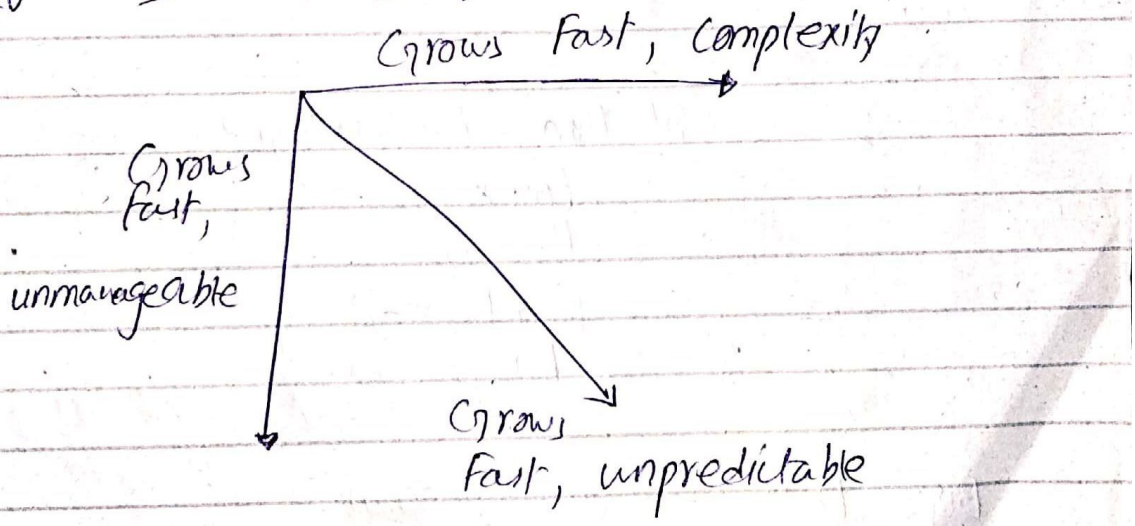
Features:

* features (controller) contributes to complexity of big data
  ↳ major contributor to scalability problems

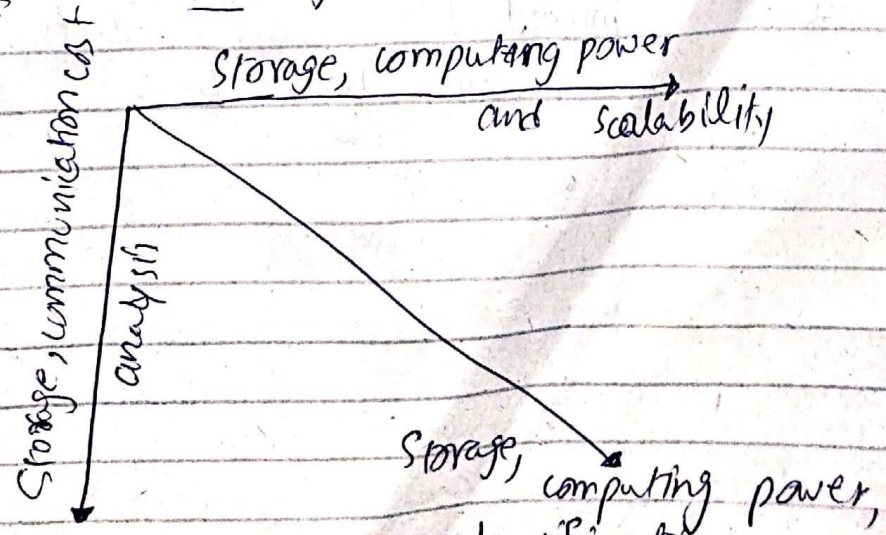
Observations:

* contributes to difficulties of managing, processing and analyzing the data

~~Big Data Challenges.~~



Big Data Challenges:



Big Data Solutions

⇒ Solution Technique
Feature Hashing

Stochastic Gradient Descent

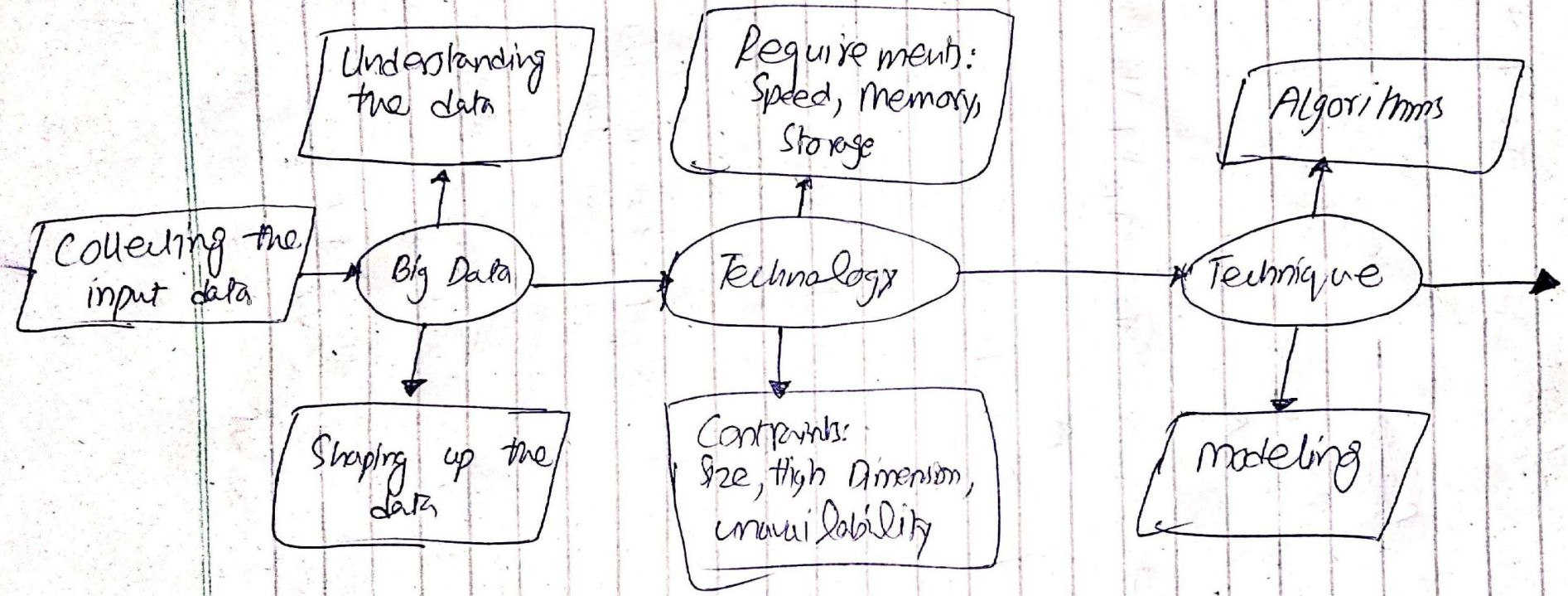
Machine learning

⇒ Solution Technologies
Linux
Hadoop
Map Reduce
Programming

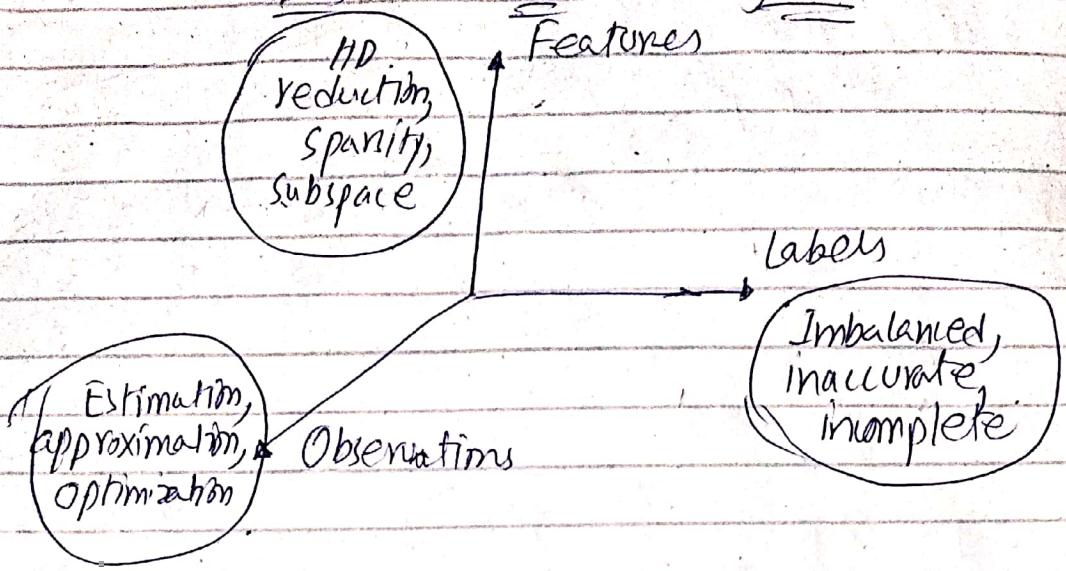
Big Data Classification:

* important and difficult problem in big Data Analytics

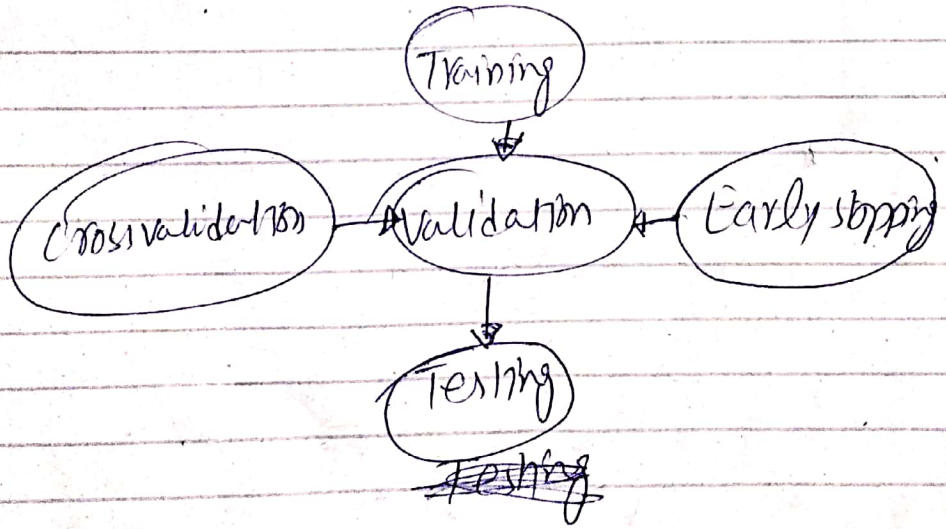
Classification Process of Big Data



Data versus Big Data



Machine Learning



* Early Stopping.

to avoid over fitting problem

Representation learning:

* useful for understanding and shaping the data
* These techniques involve statistical measures and processes.
* Statistical measure:
  - help pattern detection numerically
  - involves: mean, standard deviation, covariance
* graphical tool:
  - help in understanding patterns
  - involve: pie charts, Histograms, scatter plots.
* Statistical Processes,
  - manipulate data to extract and understand patterns
  - involves: normalization, standardization

* Focus on feature selection (controlled)

Goal

* Take data dynamically changing data characteristics into consideration.

* applicate applied to understand data but do not incorporate domain-division (class-separate) objectives.

* Cross-domain representation learning
  - Framework proposed by Tu and Sun
  - useful to understand the data for big data analytics.

. Distributed File System:

* suitable for big data management processing and analysis.

* can be customized
* must be configured

real-time data → big data on demand

optimization: act of making something as fully perfect as possible

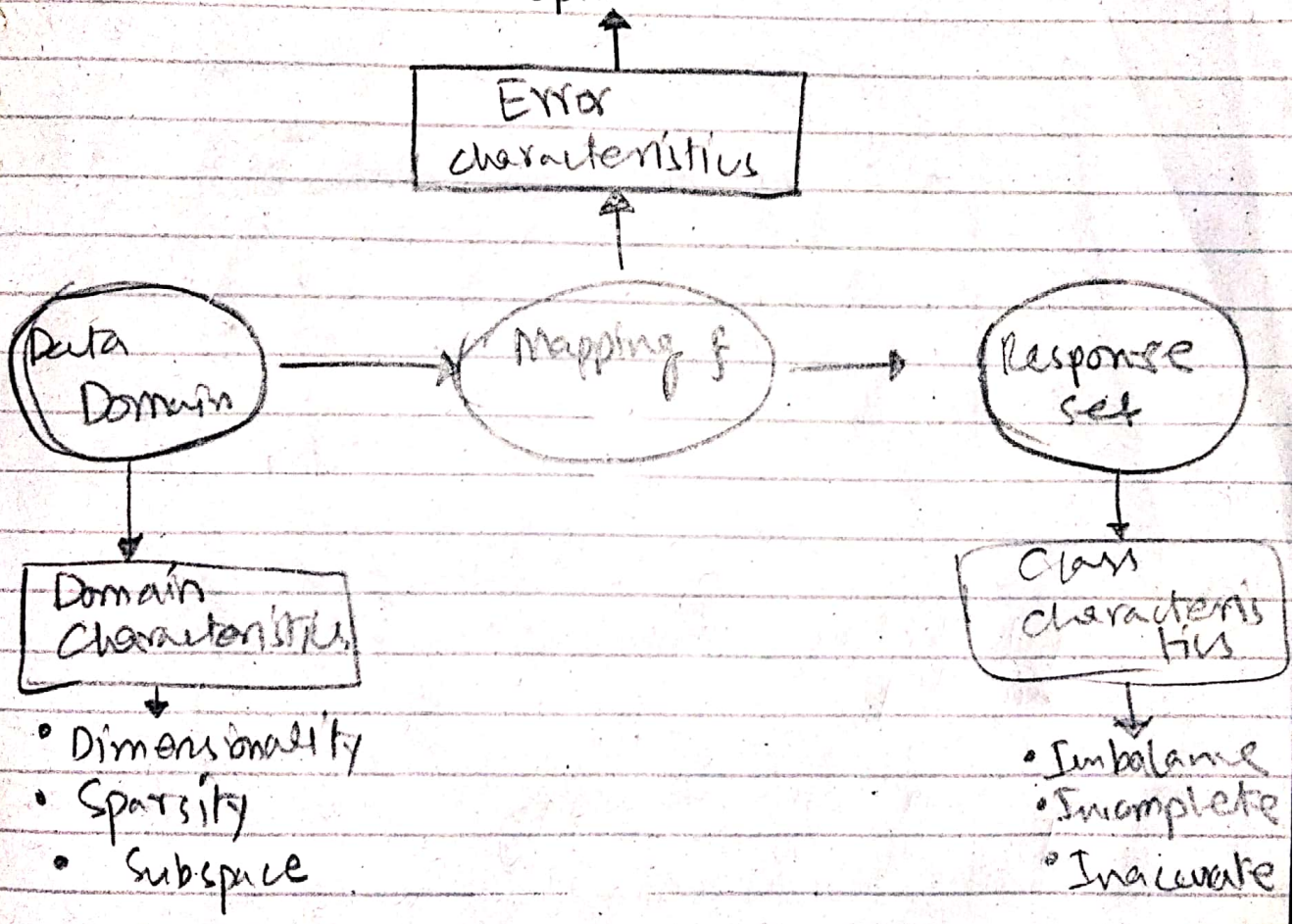
estimation: to determine roughly the size, extent or nature of

approximation: quantity that is close in value to but not the same

Hadoop → distributed File system

Classification Modeling:

- Approximation
- Estimation
- Optimization



Class Characteristics