Introduction to Machine Learning.

Siman Giri

Disclaimer!!!

- **Pre-requisites:**
  - None to be specific.
  - But basic understanding of python or any programming language might be helpful.
- **Assumptions:**
  - Complete no-voice in Machine Learning.
- **Objective:**
  - Introduce Machine Learning.
  - In brief explain various elements and components of Machine learning
  - Understand the basic of Machine Learning Workflow for any ML projects.

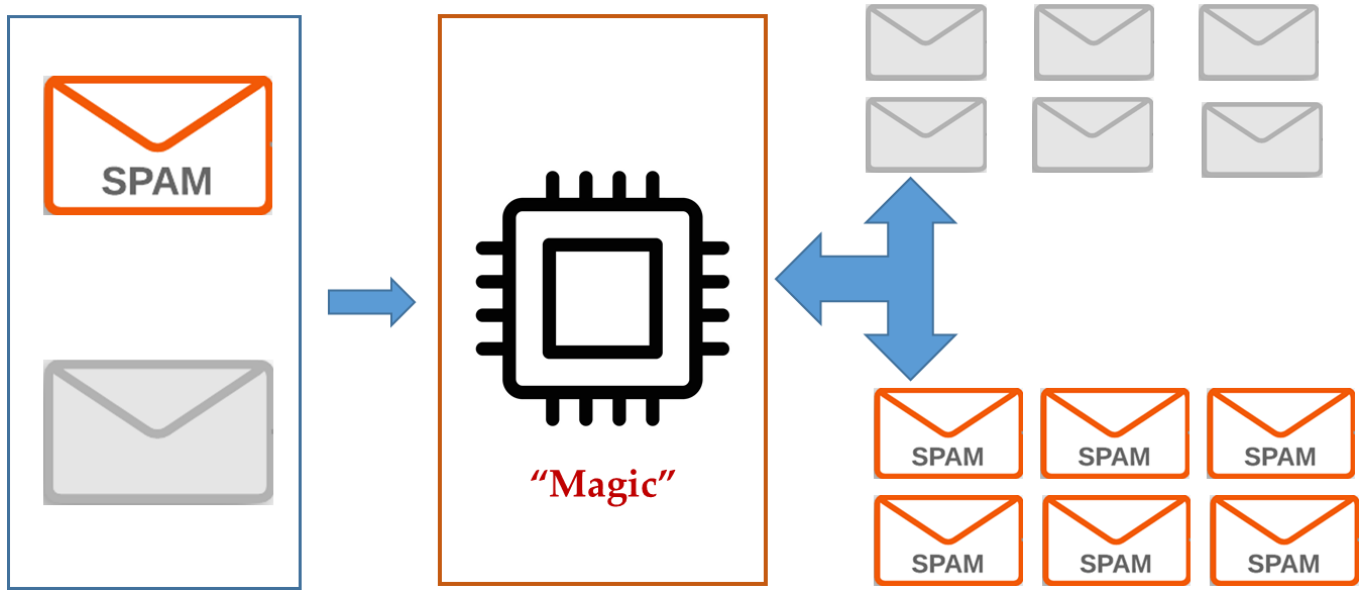
Learning → Artificial Intelligence.

1. What is Learning?

Background and Motivation for Machine Learning.

1.1 What is Learning? Intuition.

- **Example: Identify the spam emails!!!**
- **(Program a machine that learns how to filter spam emails.)**



1.1 What is Learning? Intuition.

- **Example: Identify the spam emails!!!**
- **(Program a machine that learns how to filter spam emails.)**
  - **Expert-System:**
  - In the early days of **"intelligent"** applications, many systems used **hand-coded rules** of **"if"** and **"else"** decisions to **process data** or **adjust to user input**.
  - A naïve solutions: machine can simply **make a array of all the words**, appearance of whose result in an **email being spam**, when a **new email arrives**, machine can check for those **blacklisted word from array**. If it matches one of them, it can be assigned as **spam** otherwise can be moved to **inbox**.
  - This would be an example of using an expert-designed rule system ("learning by memorization") to design an **"intelligent"** application.

1.2 Expert System: Challenges.

- **Example: Identify the spam emails!!!**
- **(Program a machine that learns how to filter spam emails.)**
  - **Expert-System ~ learning by memorizations:**
    - In our example - "learning by memorization" approach might work well but it lacks one important aspects of learning systems
      - – the ability to **label unseen email-messages** i.e**. email messages which may be spam but does not contain any of the word in the black-list(array) will be delivered to our inbox**.
    - **Manually crafting decision rules** is **feasible for some application**, but has following **two disadvantages**:
      - The **logic** required to **make a decision** is specific to a **single domain** and **task**. **Changing the task** even slightly might required to **rewrite** of the **whole system**.
      - **Designing rules** requires a **deep understanding** of how a **decision should be made by a human expert.**
      - **{We did not learn from the data we had, instead we memorize a features of data.}**

1.3 When do we need Machine Learning?

- A **successful learning system** must be able to **progress** **from individual examples** to **broader generalization**
  - – also referred as “inductive reasoning” or “inductive inference”.
- Example1: detect cat in an image.

- **Challenges with Expert System**:
  - way in which **pixels** (~ which make up an image in a computer) are “**perceived**” by the **computer** is very different from how **humans** **perceive** a **face**.
  - This difference in representation makes it basically **impossible for a human** to come up with a **good set of rules to describe** what **constitutes a cat in a digital image**.
  - Using machine learning, however, simply presenting a program with a large collection of images of faces is enough for an algorithm to determine what characteristics are needed to identify a face.
    - {learning from data ~ What does it means learning from data?}

1.4 What is Machine Learning?

- Some popular definition from legends of the field:
  - **"Learning is any process by which a system improves performance from experience". -- Herbert Simon**
  - **Definition by Tom Mitchel(1998):**
    - **Machine Learning is the study of algorithms that:**
      - **Improve their performance P**
      - **At some task T**
      - **With experience E**
    - **A well defined learning task is given by <P, T, E>.**
  - **"Field of study that gives computers the ability to learn without being explicitly programmed." - - Arthur Samuel ,1959 (an AI pioneer at IBM).**

1.5 Definition: Machine Learning?

- Machine learning is a sub-domain of artificial intelligence (AI) that utilizes **Statistics, Pattern recognition, knowledge discovery and data mining** to **automatically learn and improve with experiences** without **being explicitly programmed**.
- As machine-learning (ML) methods have improved in their capability and scope, ML has become the best way,
  - measured in terms of speed, human engineering time, and robustness, to make many applications.
- Great examples are face detection and speech recognition and many kinds of language-processing tasks.
- Almost any application that involves understanding data or signals that come from the real world can be best addressed using machine learning.

1.6 Machine Learning: Premises.

- **When and Why do we build Machine Learning System?**
  - There exists some **pattern/behavior** of interest:
    (Some Task to be solved)
  - The **pattern/behavior** is difficult to **describe:**
    (Encoding a rule to understand a behavior is difficult)
  - There is **data**
    (past experiences are in abundant)
  - Use data to **"learn"** the pattern

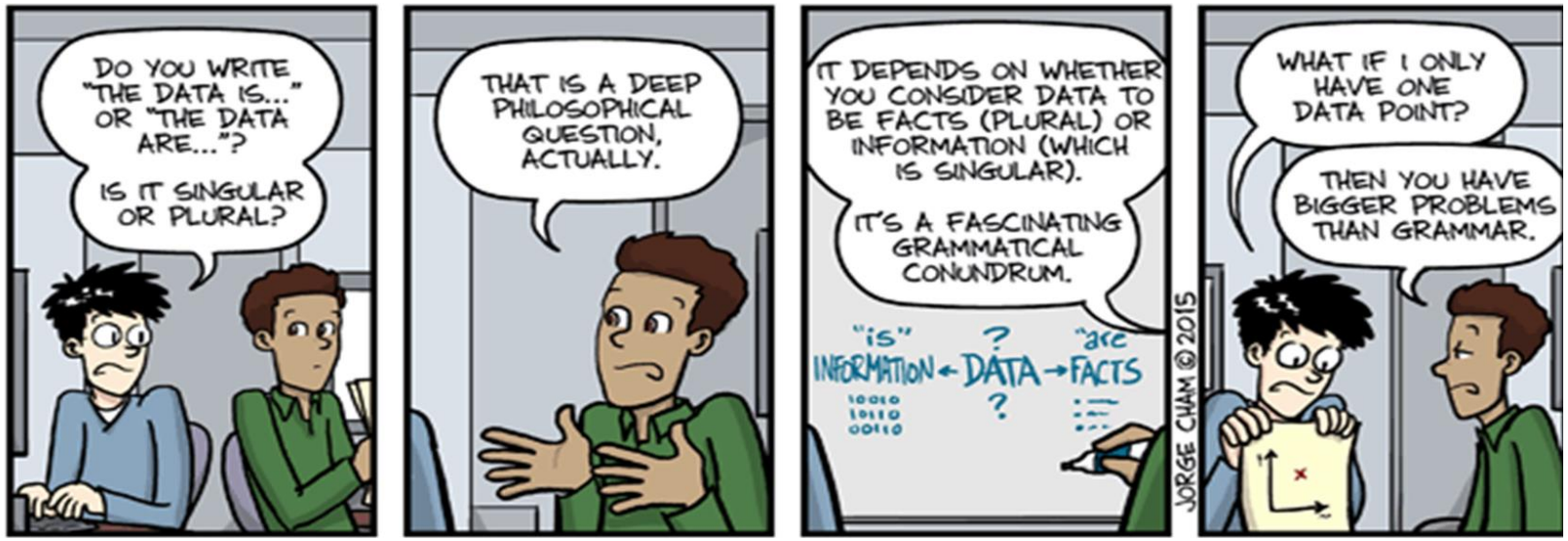
1.7 Machine Learning: Cautions!!!

- Machine learning is a very general and useful framework, but it is not **"magic"** and will not always work.
  - In order to better understand when it will and when it will not work, it is useful to **formalize** the **learning problem** more.
- **Some challenges of Machine Learning:**
  - Why do we think that previously seen data will help us predict the future?
  - **estimation:**
    - **When we have data that are noisy reflections of some underlying quantity of interest, we have to aggregate the data and make estimates or predictions about the quantity.**
      - **How do we deal with the fact that, for example, the same treatment may end up with different results on different trials?**
      - **How can we predict how well an estimate may compare to future results?**
  - **generalization:**
    - **How can we predict results of a situation or experiment that we have never encountered before in our data set?**

1.8 Components of Machine Learning.

2. Datasets for Machine Learning.

What is Data and Dataset?



2.1 Data-Basic Overview and Definitions.

- **"Data"** :a **collection of facts** about any **objects or phenomenon**.
  - Facts/Measurements can be of quantitative(numeric) or qualitative(descriptive) in nature.
  - **Variables** and **Measurements**
- Some similar definitions:
  - Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
  - Information in digital form that can be transmitted or processed
  - Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Cautions!!!!

Datum

A single piece of information, which can be treated as an observation

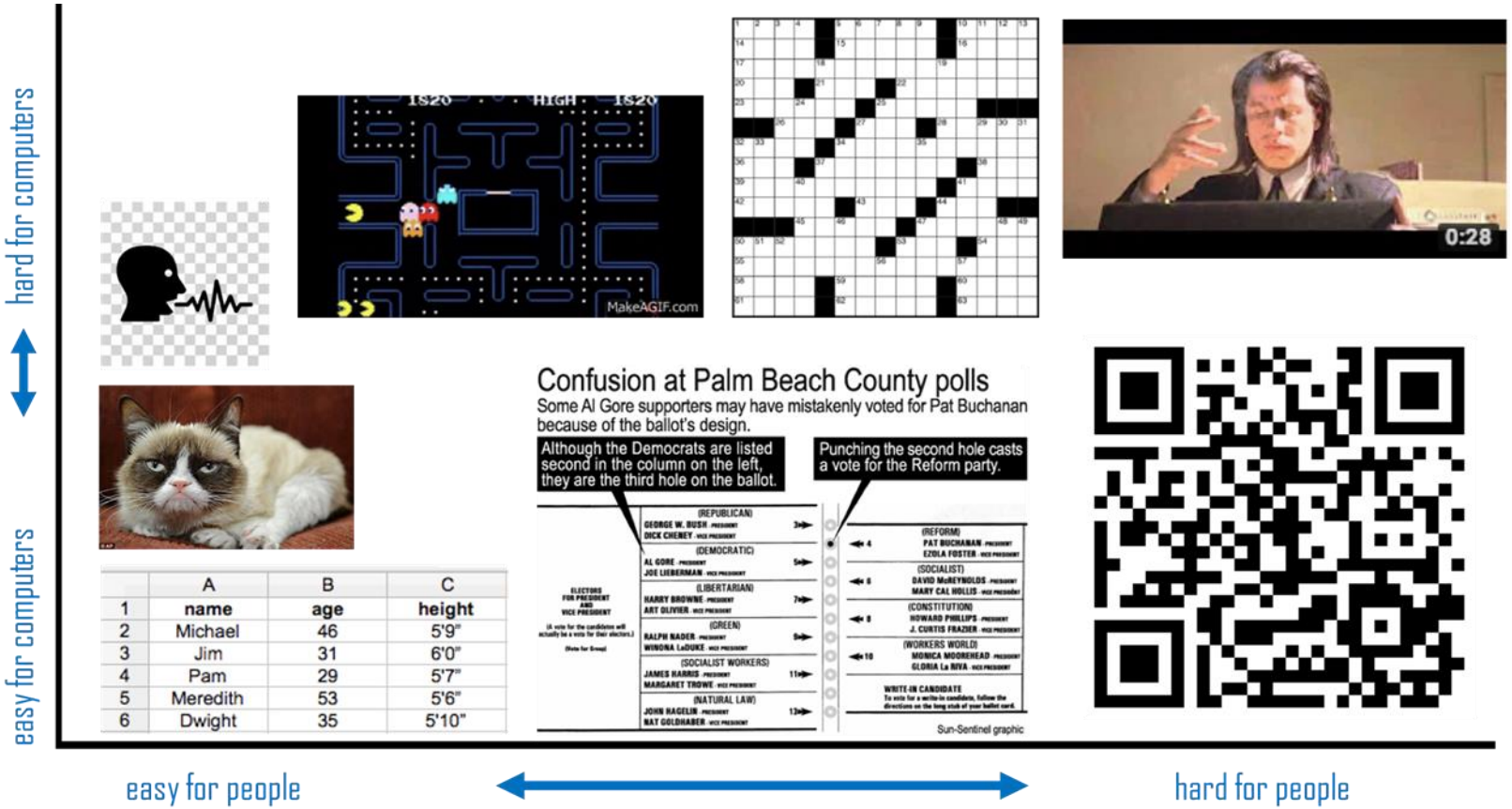
Data

The plural of datum; multiple observations

Dataset

A homogenous collection of data (each datum must have the same focus)

2.2 Dataset: Formats.



	A	B	C
1	name	age	height
2	Michael	46	5'9"
3	Jim	31	6'0"
4	Pam	29	5'7"
5	Meredith	53	5'6"
6	Dwight	35	5'10"

2.2 Dataset Formats for Machine Learning.

Plain Text:

Ends in .txt (generally)

No formatting, font type, font size, color, etc.

Text position is provided by whitespace characters (space, tab, return)

```
ALICE'S ADVENTURES IN WONDERLAND

Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired
of sitting by her sister on the bank,
and of having nothing to do: once or
twice she had peeped into the book her
sister was reading, but it had no
pictures or conversations in it, 'and
what is the use of a book,' thought
Alice 'without pictures or
conversations?'
```

.csv format

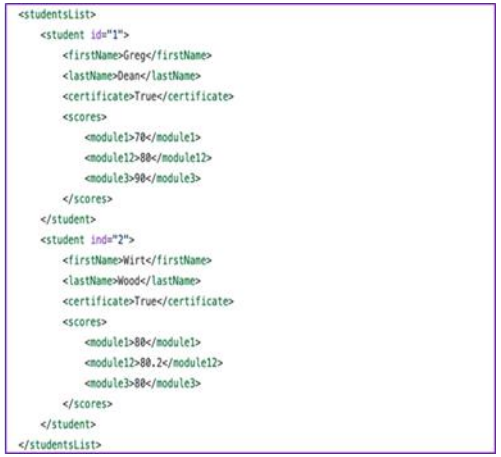
Tab-separated (.tsv/.csv)

Delimiter: character that seperates.

```
Bill #, Jack Reed (RI), Elizabeth Warren (MA)
Bill 27, Yay, Yay
Bill 28, Yay, Nay
Bill 30,, Nay
Bill 47, Nay, Nay
Bill 91, Nay, Nay
Bill 105, Yay,Yay
```

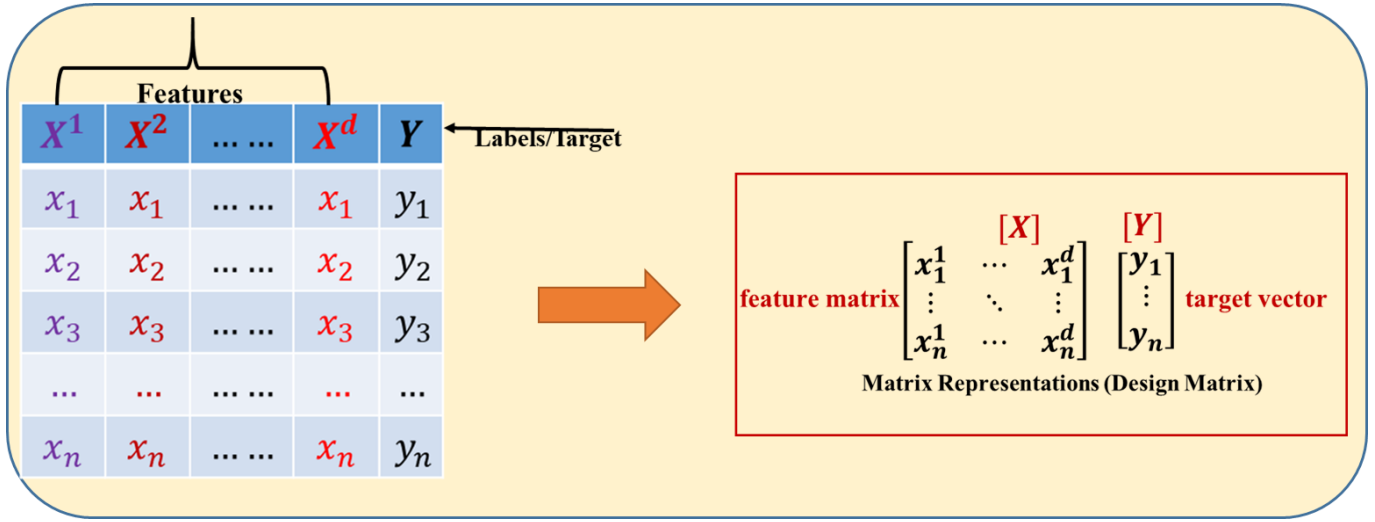
JSON

XML



2.2 Dataset Formats: In Practice.

- Some Terminology associated with dataset in practice:
- Variables:
  - **Target or output variables also referred as dependent variables.**
  - **Predictor, Feature or input variables also referred as independent variables**
- Notations:
  - Feature Variables: x or X.
  - Actual Target Variables: y or Y.
  - Predicted Target Variables:$\hat{y}$ or $\hat{Y}$.



2.3 Dataset pre-processing.

- **Major Tasks in Data (Pre)-processing:**
  - **Data cleaning**
    - Fill in missing values, smooth noisy data, **identify** or remove **outliers**, and resolve inconsistencies
  - **Data integration**
    - Integration of multiple databases, or with in same data set.
  - **Data transformation**
    - Normalization/Scaling (scaling to a specific range)
  - **Data reduction**
    - Feature Selection/Extraction.

2.3 Dataset pre-processing – Missing Data.

- **Data may not be always available.**
  - E.g., many variables have no recorded value for several attributes, such as customer income in sales data
  - **Missing data may be due to :**
    - data not entered due to misunderstanding
    - certain data may not be considered important at the time of entry
    - not register history or changes of the data
- **Missing data may need to be inferred.**

2.3 Dataset pre-processing – Missing Data.

- How to handle **missing Data**?
  - The process of handling missing data is known as Data Imputation. Some of the common approaches are:
    - Fill in the missing value **manually**: tedious + infeasible?
    - Use a **global constant** to fill in the missing value: e.g., "unknown", a new class?!
    - Use the **attribute mean** to fill in the missing value
    - Use the **attribute mean for all samples of the same class** to fill in the missing value: smarter
    - Use the most probable value to fill in the missing value
    - **inference**-based such as regression, Bayesian formula, decision tree

2.4 Dataset pre-processing – Data Transformation.

- Some common task in Data Transformation are:
  - **Scaling**.
  - **Encoding**.
  - Feature Selection and Feature Engineering (Dimensionality Reduction) .
  - Learned Embedding (often for text data).

2.4 Data Transformation – Scaling.

- **Goal: bring them all with the same range-scale.**
  - Use when different numeric features have different scales (different range of values).
  - Features with much higher values may overpower the others.
- Different methods exist. Most common techniques are:
  - **Standard Scaling**
    - Generally most useful and used, assumes data is normally distributed.
    - For every feature or attribute subtract the mean value and scale by standard deviation.
    - New feature has mean 0 and standard deviation 1.
      - $X_{new} = \frac{X-\mu}{\sigma}$
  - **Min-max Scaling**
    - Scales all features between a given min and max value.
    - Only used **when min and max has some sense** in data.
    - Sensitive to outliers.
      - $X_{new} = \frac{X-x_{min}}{x_{max}-x_{min}} . (max - min) + min.$

2.5 Data Transformation – Encoding.

- Many algorithms can only handle numeric features, so we need to encode the categorical ones.
  - **Ordinal Encoding**:
    - Assigns an integer value to each category in the order they are encountered.
    - Only useful for ordinal data types.

	boro	boro_ordinal	salary
0	Manhattan	2	103
1	Queens	3	89
2	Manhattan	2	142
3	Brooklyn	1	54
4	Brooklyn	1	63
5	Bronx	0	219

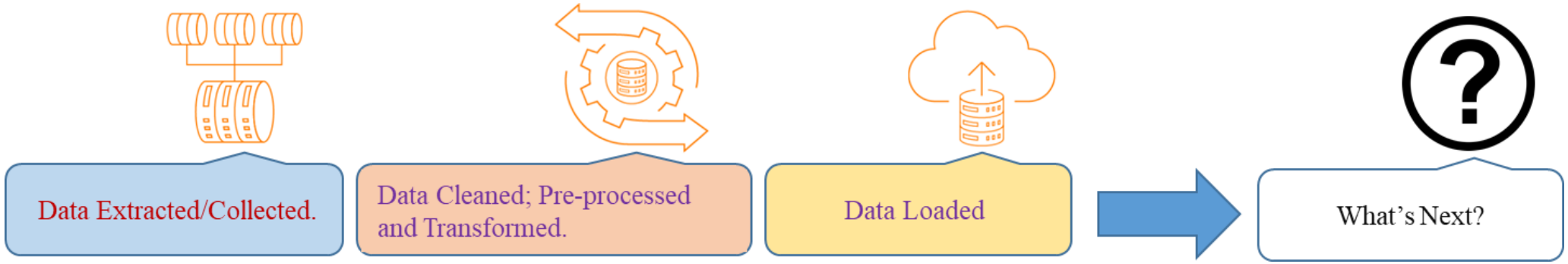
Ordinal Encoding

  - **One Hot Encoding**:
    - Also known as dummy encoding.
    - Adds a new features/attributes for every category in data.

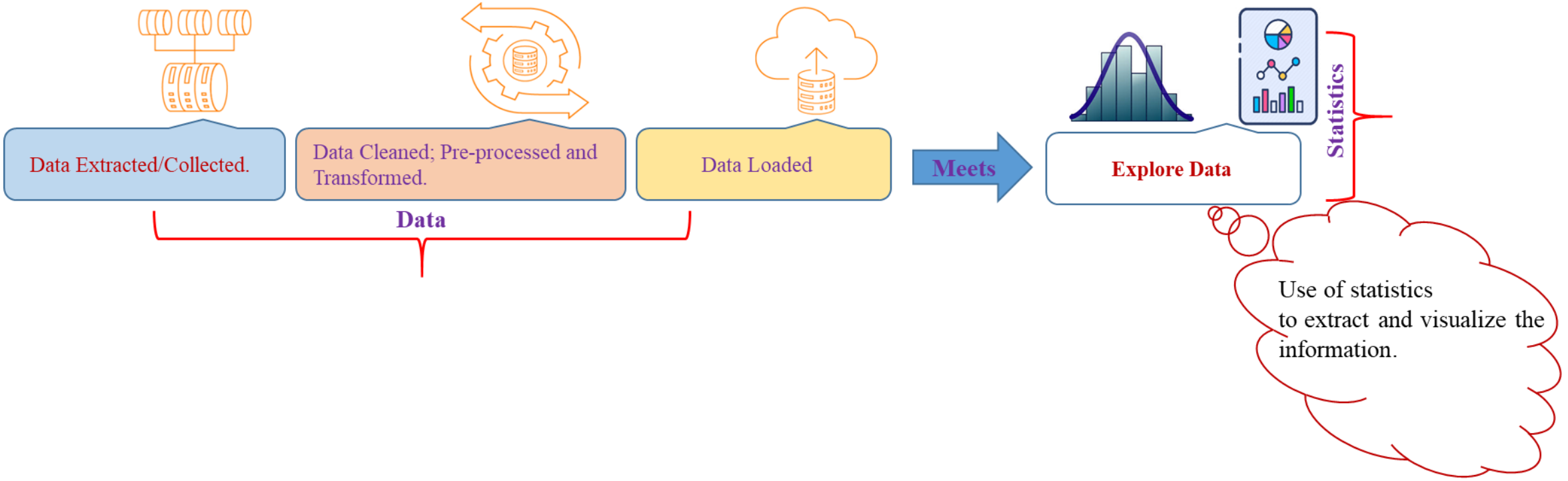
One Hot Encoding

	boro	boro_Bronx	boro_Brooklyn	boro_Manhattan	boro_Queens	salary
0	Manhattan	0	0	1	0	103
1	Queens	0	0	0	1	89
2	Manhattan	0	0	1	0	142
3	Brooklyn	0	1	0	0	54
4	Brooklyn	0	1	0	0	63
5	Bronx	1	0	0	0	219

Story So far……


Data Extracted/Collected.
Data Cleaned; Pre-processed and Transformed.
Data Loaded
What's Next?

Story So far……


Data Extracted/Collected.
Data Cleaned; Pre-processed and Transformed.
Data Loaded
Data
Meets
Explore Data
Statistics
Use of statistics to extract and visualize the information.

2.6 Data Exploration – Statistical Analysis.

- Data exploration, also known as **exploratory data analysis (EDA)**, is a process where users look at and understand their data with **statistical and visualization methods**.
  - This step helps **identifying patterns and problems in the dataset**, as well as **deciding which model or algorithm to use in subsequent steps.**
- Bad data will lead to bad results even with perfect model.
- Some common things that precedes any ML model building are (but not limited to):
  - **Outliers in Data:**
    - Observations that either is very large or very small value.
    - Boxplot may be used to detect the presence of such outliers.
  - **Distributions of Data:**
    - Most of the time in ML we assume normality of data.
    - Techniques like Histogram may be used to check for such normality.
  - **Collinearity:**
    - Determines the linearity of feature variables among themselves.
    - Can be deduce with covariance matrix or heatmap.

3. Elements of Learning Process.

Building a Machine Learning Models.

3.1 Framing a Learning Problem.

- Learning Problem(Tasks) in Machine Learning depends on type of the data we have:
- Datasets are made up of individual examples that contain features and a label.
  - Examples that contain both features and a label are called **labeled datasets**.
  - Examples that contain only features are called **unlabeled datasets**.

Features

Label

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23

Example

Labeled Dataset

Supervised Learning Techniques.

Un-Supervised Learning Techniques.

Features

Unlabeled Dataset

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure
2021-09-09	49.71N	82.16W	74	20	3	N	18.6
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94

Example

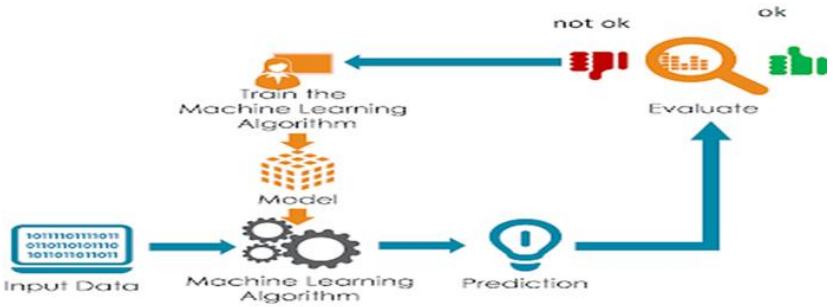
3.2 Supervised Machine Learning.

- Data in Supervised Learning:
  - For Supervised Learning Setup, **training data** comes in pairs of inputs **(x, y):** where $X \in R^d$ is the input instance and Y its label, which can be written as:
    - $D = \{(x_1, y_1) \dots (x_n, y_n)\} \subseteq R^d * C$
    - Where:
      - R^d: **d-dimensional feature space.**
      - x_i: **input vector of the i^{th} sample.**
      - y_i: **label of the i^{th} sample.**
      - C: **label space.**
- Tasks in Supervised Learning:
  - There can be multiple scenario for the label space c.

Binary Classification	$c = \{0 \text{ or } 1\}$	E.g.: An email is either spam or not a spam.
Multi Class Classification	$c = \{1,2,..k\}(k \geq 2)$	E.g.: Traffic sign Classification.
Regression	$c = \mathbb{R}$	E.g.: Height of the person.

Elements of Machine Learning.

- There may be hundreds of machine learning algorithms, all of those algorithms must have following three attributes:
  - **A Decision Process (Representation/Model):**
    - Machine learning algorithms(Models) are used to make inference or estimate of an output based on input data – labeled or unlabeled.
  - **An Error Function (Evaluation):**
    - A performance metric used to evaluate the estimate of a model.
    - Metrics depends on types of learning (supervised or unsupervised) and types of task (Classification or Regression)
  - **An model Optimization Process:**
    - An automated algorithm or process used to update parameters of machine learning models until threshold or accepted evaluation metric has been achieved.
    - Examples:
      - Gradient Descent
      - ID3/CART



A Decision Process: Function Approximation.

- Machine learning is concerned with using the right features to build the right models that achieve the right tasks.
- For a given problem, what kind of function better approximates the relationship between input (feature) with output (target/label)
- The function broadly can be classified as:
  - **Numerical Function:**
    - **Linear Regression.**
    - **Support Vector Machines.**
  - **Symbolic Function (Logical or Rule Based):**
    - **Decision Tree and Random Forest**
  - **Instance Based:**
    - **Nearest Neighbor**
  - **Probabilistic Models:**
    - **Naïve Bayes**

Framework of a Learning Process.

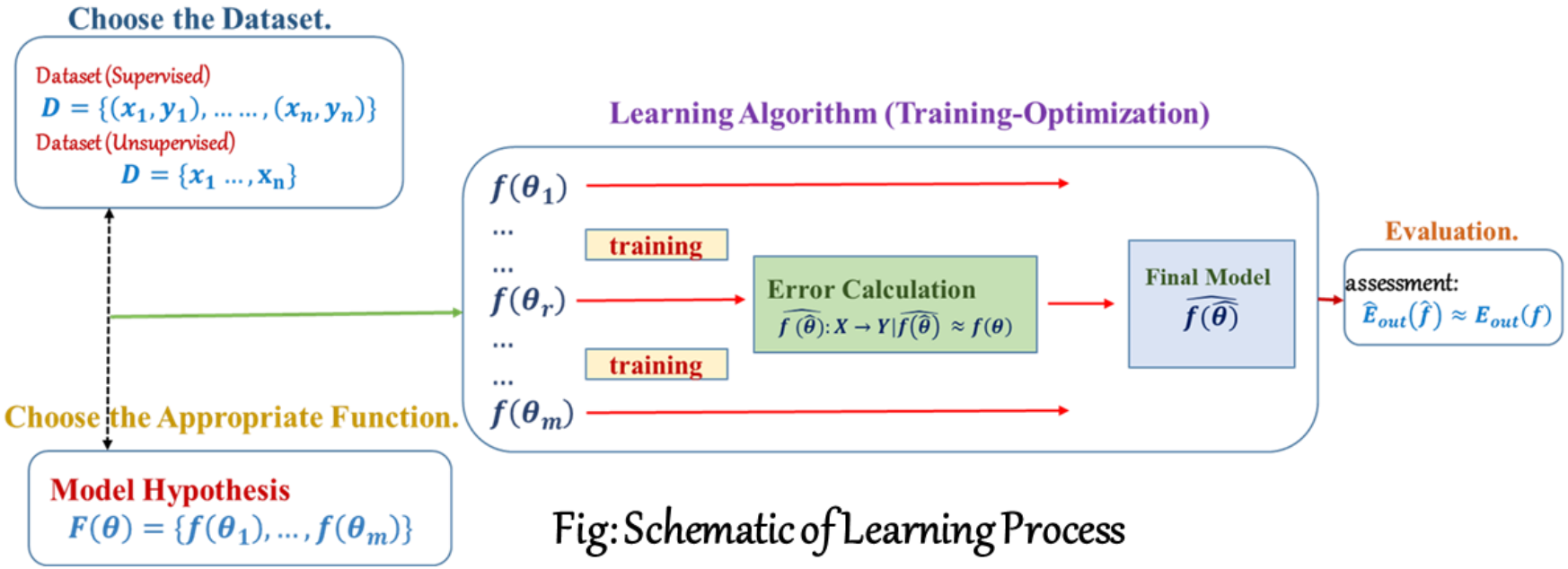


Fig: Schematic of Learning Process

3.2 Supervised Machine Learning.

- Data in Supervised Learning:
  - For Supervised Learning Setup, **training data** comes in pairs of inputs **(x, y):** where $X \in R^d$ is the input instance and Y its label, which can be written as:
    - $D = \{(x_1, y_1) \dots (x_n, y_n)\} \subseteq R^d * C$
    - Where:
      - R^d: **d-dimensional feature space.**
      - x_i: **input vector of the i^{th} sample.**
      - y_i: **label of the i^{th} sample.**
      - C: **label space.**
- Tasks in Supervised Learning:
  - There can be multiple scenario for the label space c.

Binary Classification	$c = \{0 \text{ or } 1\}$	E.g.: An email is either spam or not a spam.
Multi Class Classification	$c = \{1,2,..k\}(k \geq 2)$	E.g.: Traffic sign Classification.
Regression	$c = \mathbb{R}$	E.g.: Height of the person.

3.3 Supervised Machine Learning : Examples.

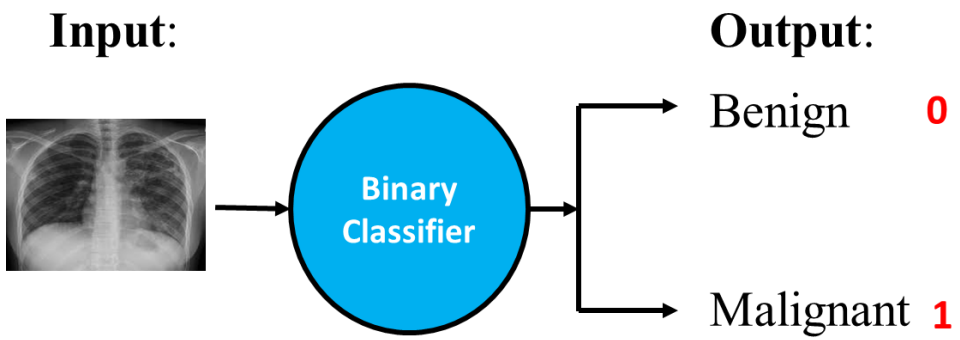
Regression

- House Price Prediction:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	Street	OverallCond	YearBuilt	YearRemo	MasVnrArea	TotalBsmt	Heating	CentralAir	BsmtFullB	FullBath	HalfBath	Bedroom/	SaleCondition	Price
	Pave	6	1961	1961	0	882	GasA	Y	0	1	0	2	Normal	11622
	Pave	6	1958	1958	108	1329	GasA	Y	0	1	1	3	Normal	14267
	Pave	5	1997	1998	0	928	GasA	Y	0	2	1	3	Normal	13830
	Pave	6	1998	1998	20	926	GasA	Y	0	2	1	3	Normal	9978
	Pave	5	1992	1992	0	1280	GasA	Y	0	2	0	2	Normal	5005
	Pave	5	1993	1994	0	763	GasA	Y	0	2	1	3	Normal	10000
	Pave	7	1992	2007	0	1168	GasA	Y	1	2	0	3	Normal	7980
	Pave	5	1998	1998	0	789	GasA	Y	0	2	1	3	Normal	8402
0	Pave	5	1990	1990	0	1300	GasA	Y	1	1	1	2	Normal	10176
1	Pave	5	1970	1970	0	882	GasA	Y	1	1	0	2	Normal	8400
2	Pave	5	1999	1999	0	1405	GasA	Y	1	2	0	2	Normal	5858
3	Pave	5	1971	1971	504	483	GasA	Y	0	1	1	2	Normal	1680
4	Pave	5	1971	1971	492	525	GasA	Y	0	1	1	3	Normal	1680
5	Pave	6	1975	1975	0	855	GasA	Y	0	2	1	3	Normal	2280
6	Pave	6	1975	1975	0	836	GasA	Y	0	1	0	2	Normal	2280
7	Pave	5	2009	2010	162	1590	GasA	Y	0	2	1	3	Partial	12858
8	Pave	5	2009	2010	256	1544	GasA	Y	0	2	0	3	Partial	12883
9	Pave	5	2005	2005	615	1698	GasA	Y	0	2	0	3	Normal	11520
0	Pave	5	2005	2006	240	1822	GasA	Y	0	2	0	3	Normal	14122
1	Pave	5	2003	2004	1095	2846	GasA	Y	1	2	1	3	Normal	14300
2	Pave	5	2002	2002	232	1671	GasA	Y	1	2	1	3	Normal	13650
3	Pave	5	2006	2006	178	1370	GasA	Y	0	2	0	2	Normal	7132
4	Pave	5	2005	2005	0	1324	GasA	Y	0	2	0	3	Normal	18494
5	Pave	5	2006	2006	14	1145	GasA	Y	0	2	0	2	Normal	3203
6	Pave	5	2004	2004	0	384	GasA	Y	1	2	1	3	Normal	13300

Classification

- Tasks in Supervised Learning-Classification Task:
  - **Binary Classification.**
  - **Multi-Class Classification.**



3.4 Supervised Learning: Problem Formulation.

- **Data → Learning**.
- Our assumptions:
  - We belief that **datapoints (x_i, y_i)** are drawn from some (unknown) **distribution $P(X, Y)$.**
  - We would like to learn a function "h" such that for a new pair (x, y), we have
    - $h(x) = \hat{y}$ where $\hat{y} \approx y$.
- Learning:
  - **Step-I:** .From a hypothesis class **"H"** i.e. "$h \in H$"; we need to select the machine learning algorithm that we think is appropriate for the particular problem.
  - **Step-II:** Find the best function with in the class "$h \in H$"
    - **How can we find best function?**
      - For this we need be able to evaluate what it means for one function to be better than another. (Aka Loss/Error function.)

3.4 Supervised Learning: Error Function.

- **Error function($\mathbb{E}$):**
- Aka Loss Function($\mathbb{L}$).
- A loss function evaluates a hypothesis $h \in H$; on our training data and tells us how bad it is.
  - Higher the loss, the worse it is
  - A loss of zero means it makes **perfect predictions**.
- It is common practice to normalize the loss by the total number of training samples ; “n”, so that the output can interpreted as the average loss per sample (Cost Function).

3.4 Supervised Learning: Error Function.

- **Loss function(𝕃): Examples**.
- **Zero-One Loss**:
  - It counts how many mistakes and hypothesis function "h" makes on the training set, given by:
    - $\mathbb{L}_{0\backslash 1}(h) = \frac{1}{n}(\sum \delta_h(x_i) \neq y_i; \{\delta_h(x_i) \neq y_i = \begin{cases} 1, if\ h(x_i) \neq y\_i \\ 0, other\ wise. \end{cases}$
  - For every single example: if it is miss-predicted then loss is 1 and 0 otherwise.
  - The normalized zero-one loss returns the fraction of misclassified training samples.
  - Can be used to evaluate classifier but are seldom used because it is rarely is useful to guide **optimization procedures** because the function is non-differentiable and non-continuous.

3.4 Supervised Learning: Error Function.

- **Loss function($\mathbb{L}$): Examples.**
  - **Squared Loss:**
    - Typically used in regression settings. It iterates over all training sample and computes the loss as:
      - $\mathbb{L}_{sq}(h) = \frac{1}{n}(\sum_i^n (h(x_i) - y_i)^2)$.
  - **Absolute Loss:**
    - Expressed as:
      - $\mathbb{L}_{abs}(h) = \frac{1}{n}(\sum_i^n |h(x_i) - y_i|)$.

3.5 What after learning "h"?

- **Prediction:**
  - **Learned model(hypothesis) h(.)** is used to predict the label Y for data without label, the predicted label is represented as $\hat{Y}$.
- **Inference:**
  - Understanding the association between Y and X.
    - Which predictors are associated with the response?
    - What is the relationship between the response and predictor?
    - Can the relationship between Y and each predictor be adequately summarized using a linear equation?

3.6 (Supervised) Machine Learning: Conclusion.

- It is an attempt to find the function *"h"* that minimizes the selected loss such that:
  - $h = argmin_{h \in H} \mathbb{L}(h)$
- A big part of machine learning focuses on the question, how to do this minimization efficiently?
  - **Optimization Techniques.**
- If you find a function *h(.)* with low loss on your data *D*, how do you know whether it will still get examples right that are not in *D*?
  - **Generalization!!!**

4. Towards Generalization.

How to achieve Generalization in Machine Learning?

Train-Test Split?

4.1 How to achieve Generalization in Machine Learning?

- We want to build a model $\hat{f}(x) \approx f(x)$; such that it can make a best prediction on sampled data $\hat{y}_i = \hat{f}(x_i) | \mathrm{E_{in}}(\hat{y}, y) \approx 0$ and also on observed data $\hat{y_o} = \hat{f}(x_o) | E_o(\hat{y_o}, y_o) \approx 0$.
  - Sample data: data D_{in} used to train model.
  - Observed data: data D_o used to measure the predictive quality of model.
- E_{in} is typically defined as the average of *Pointwise errors* from data points in the sample data i.e.
  - $E_{in}(\hat{f}, f) = \frac{1}{n}\sum_i err_i$. [re-substitution error-> how well the model fits the learning data?]
- E_o is the *theoretical mean(expected value)* of the *Pointwise errors* over the entire *input space*:
  - $E_o(\hat{f}, f) = E_X\left[err\left(\hat{f}(x), f(x)\right)\right]$ [generalization error-> how well the model fits the data]
- The point x denotes a **general data** point in the *input space* X. And as we said, *the expectation is taken over the input space X*. This means that the nature of E_o **is highly theoretical**. In practice, we will **never** be able to **compute this quantity**.
- **What do we do in practice?**

4.2 Training vs. Test Error: Towards Generalization!!!

- **How to split? Holdout Method!!!**
- **Generalization: The Train-Test Split.**
  - **Training Data**: used to fit model (in-sample Data)
    - Error:

For a Dataset:

$$D_{train} = \{(x_1, y_1), \dots, (x_{n-a}, y_{n-a})\}.$$

$$E_{train}(\hat{f}, f) = \frac{1}{n-a} \sum_{i=1}^{n-a} err[\hat{f}(x_i), y_i];$$

Training error typically underestimates test error.

- **Test Data**: check generalization error(out-sample-data).
  - Error:

For a Dataset:

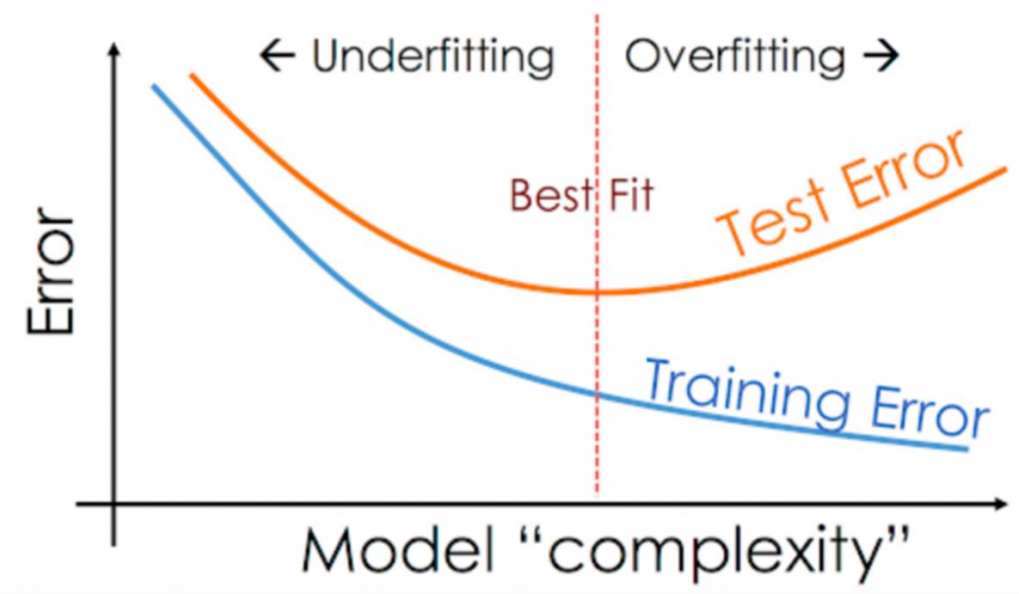
$$D_{test} = \{(x_1, y_1), \dots, (x_a, y_a)\}: a > 1.$$

$$E_{test}(\hat{f}, f) = \frac{1}{a} \sum_{l=1}^{a} err[\hat{f}(x_l), y_l];$$

E_{test} is unbiased estimate of $E_{out}(h)$.

4.3 Training vs. Test Error: Overfitting and Underfitting!!!

- **Errors is composed of:**
  - **MSE = Variance +Bias + Noise.**
- **Underfitting:**
  - ↑High Train Error ↑High Test Error.
  - High Bias Low Variance.
- **Overfitting:**
  - ↓Low Train Error ↑High Test Error
  - Low Bias High Variance.

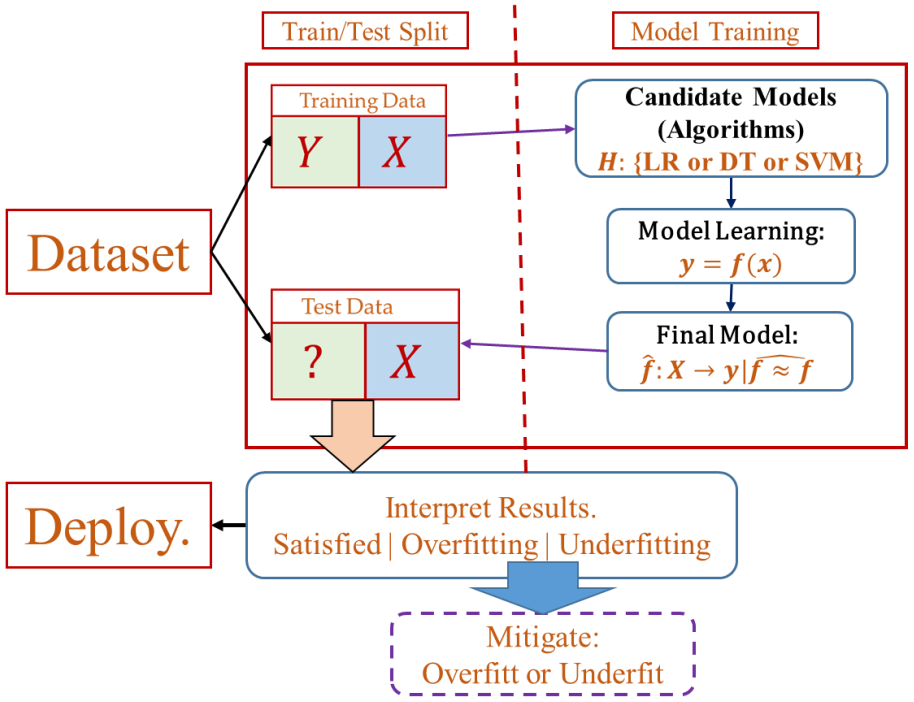


4.4 Mitigating Overfitting: Techniques.

- Some Techniques to avoid overfitting:
  - **Early stopping** ✓
  - **Train with more data** ✓
  - **Feature Selection or Data augmentation** ✓
  - **Cross-Validation** ✓
  - **Ensemble Methods** ✓
  - **Regularization** ✓

(Supervised) Machine Learning: Workflow.

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d from the distribution D.
  - Find $y = f(x) \in H$ that minimizes $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f, x_i, y_i)$
    - such that the expected loss is small:
      - $L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, xy)]$



5. Unsupervised Learning

5.1 Unsupervised Learning: Introduction.

- Unsupervised learning focuses on a set of **statistical tool/ methods** intended for setting in which we have only a **set of features** i.e.:
  - $\boldsymbol{D = \{(x_1) \dots (x_n)\} \subseteq R^d}$
  - Here:
    - $\boldsymbol{R^d}$: d-dimensional feature space.
    - $\boldsymbol{x_i}$: input vector of the $\boldsymbol{i^{th}}$ sample measured on **n observations**.
- In unsupervised learning we are **not interested in prediction**, because we do not have an associated response variable **Y**.
- Then, What can be the **goal/objective** of unsupervised learning?

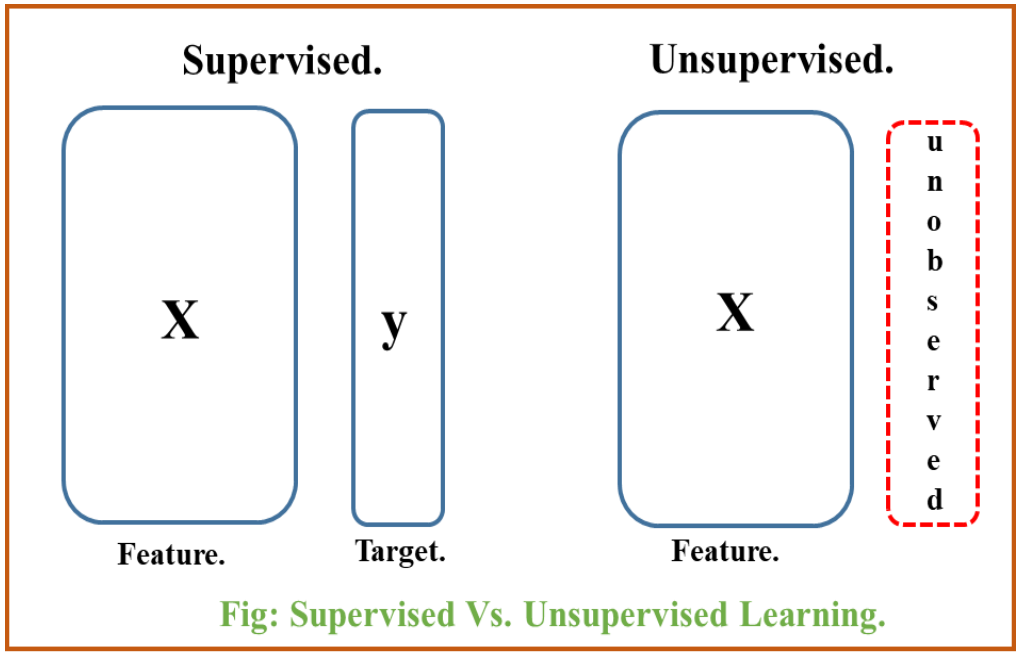


Fig: Supervised Vs. Unsupervised Learning.

5.2 Unsupervised Learning: Need.

- What can be the **goal/objective** of unsupervised learning?
  - Goal in unsupervised learning is to discover interesting things about the measurements/features. Thus tasks in unsupervised learnings are defined purely from exploratory perspective.
- Unsupervised learning explores the relationship between the feature variables and tries to answer the questions such as:
  - Is there an informative way to visualize the data?
  - Can we discover subgroups among the variables or among the observations?

5.3 Unsupervised Learning: Task.

- **Clustering**:
  - The method of dividing the objects into clusters which are similar between them and are dissimilar to the objects belonging to another cluster
    - K means Clustering,
- **Association**:
  - Rule based machine learning algorithms, that discovers the probability of the co-occurrence of items in a collection i.e. finding relationships between variables in a given dataset.
    - Apriori Algorithm, FP-Growth Algorithm
- **Dimensionality reduction**:
  - *Principal component analysis*

6. Concluding Remarks.

6.1 ML Process

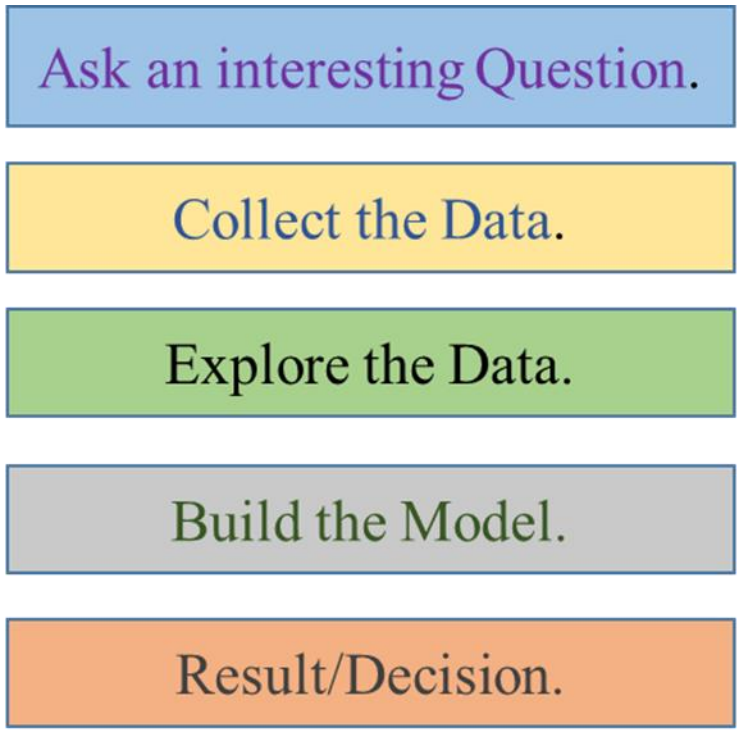
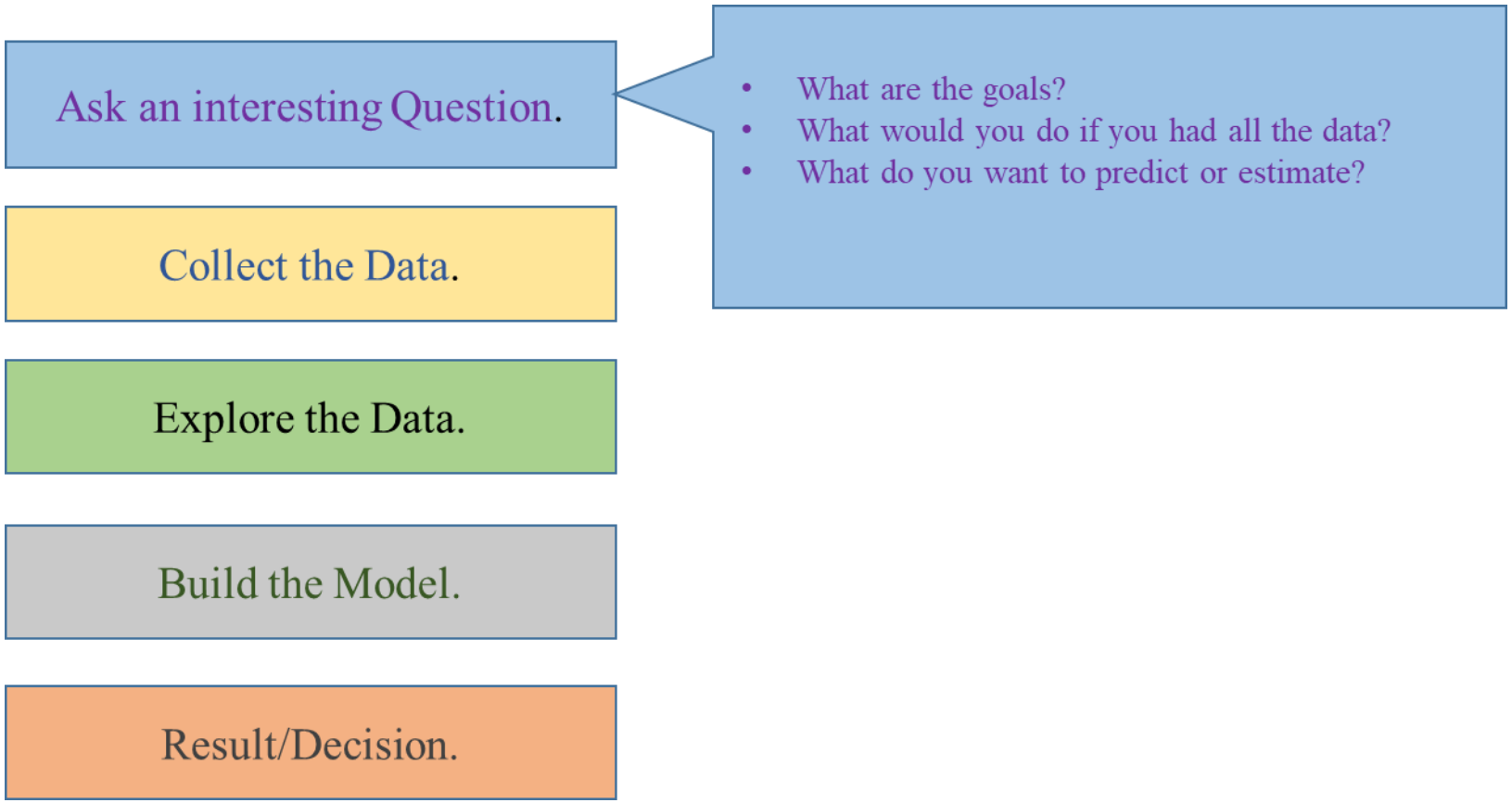


Fig: Flowchart of ML Process.

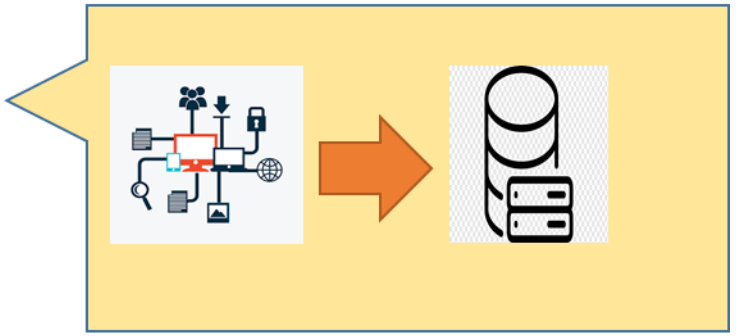
6.1 ML Process-1.



6.1 ML Process-2.

Ask an interesting Question.

Collect the Data.



Explore the Data.

Build the Model.

Result/Decision.

6.1 ML Process-3.

Ask an interesting Question.

Collect the Data.

Explore the Data.

Build the Model.

Result/Decision.

6.1 ML Process-4.

Ask an interesting Question.

Collect the Data.

Explore the Data.

Build the Model.

Result/Decision.

- Build/Train and Validate a Machine Learning Model.

6.1 ML Process-5.

Ask an interesting Question.

Collect the Data.

Explore the Data.

Build the Model.

Result/Decision.

Did it answered the question?

Thank You!!!
For Collaboration:
simangiri@heraldcollege.edu.np