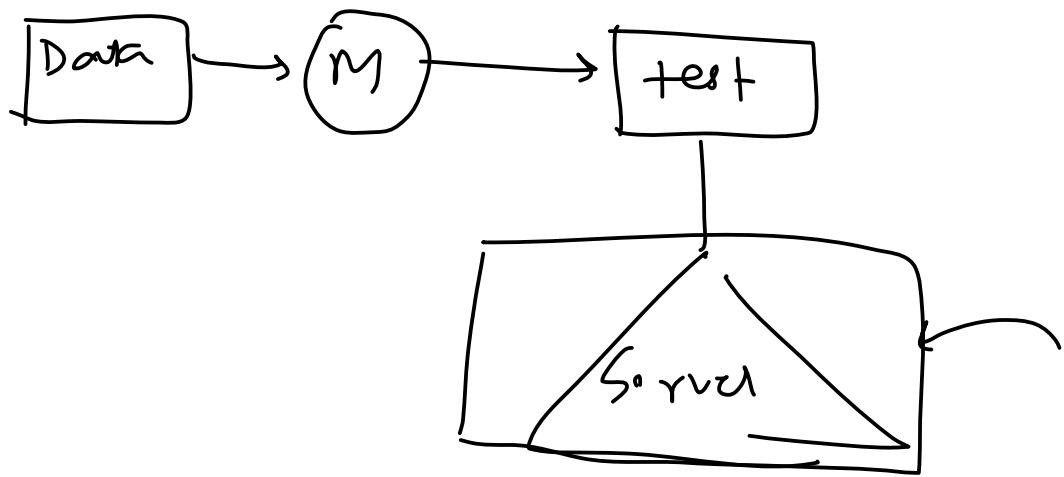


# 1. Batch Vs Online ML

Wednesday, March 17, 2021 5:30 PM

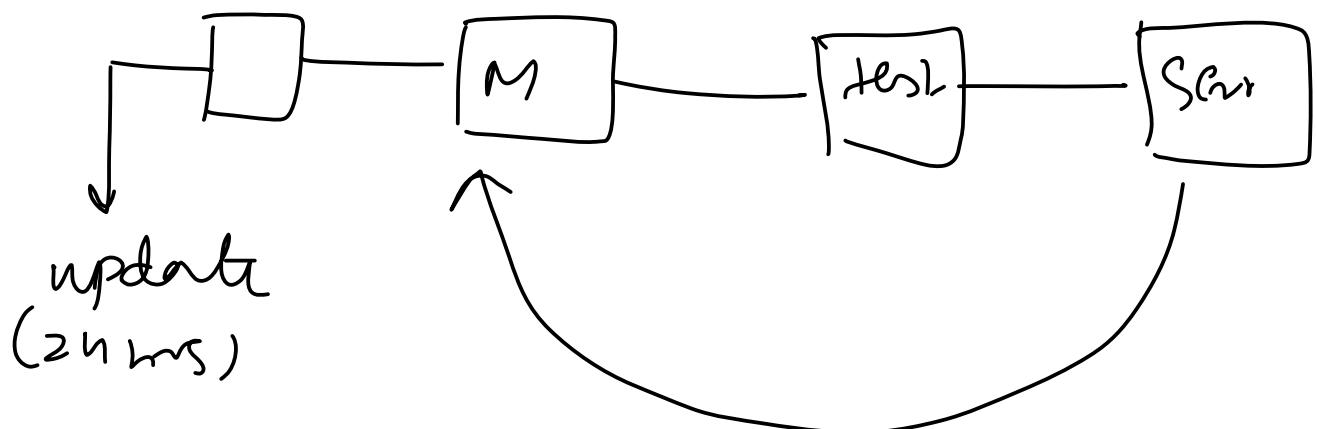
## 2. Batch/Offline ML

Wednesday, March 17, 2021 5:31 PM



### 3. The problem with Batch Learning

Wednesday, March 17, 2021 5:47 PM



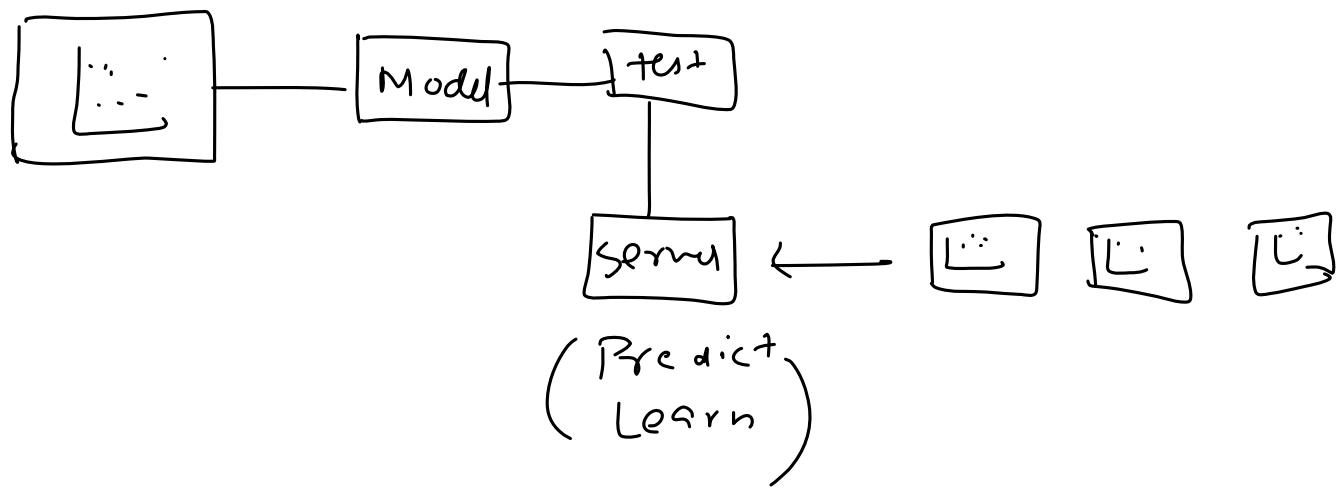
## 4. Disadvantages of Batch ML

Wednesday, March 17, 2021 5:32 PM

1. Lots of Data
2. Hardware Limitation
3. Availability

# 1. Online Machine Learning

Thursday, March 18, 2021 4:27 PM



## 2. When to use?

Thursday, March 18, 2021 4:33 PM

1. Where there is a concept drift
2. Cost Effective
3. Faster solution

### 3. How to implement?

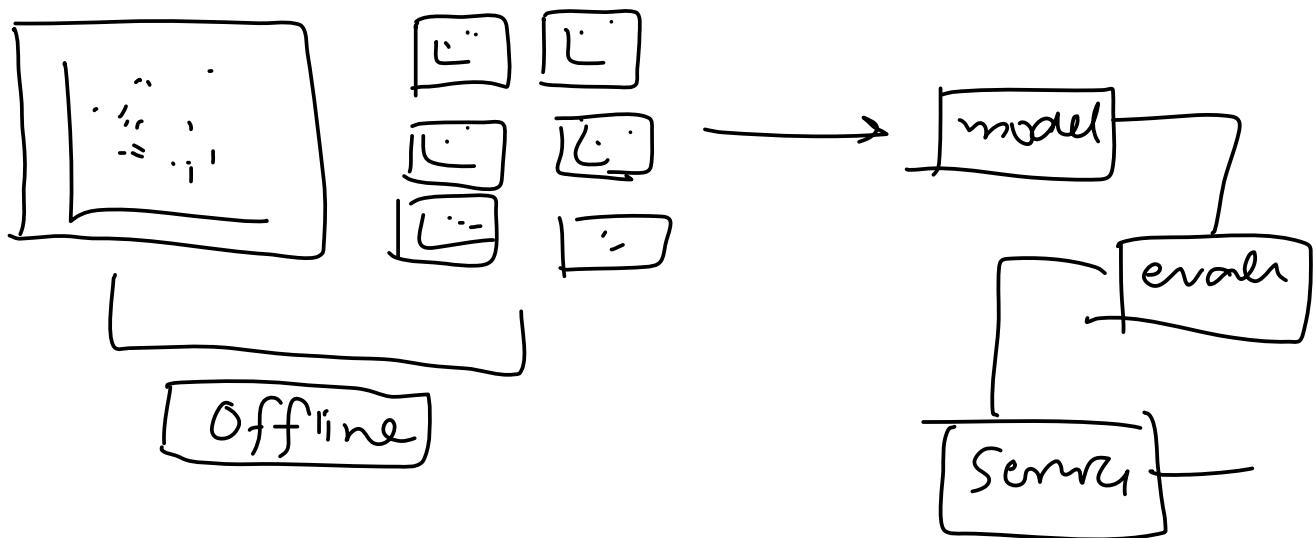
Thursday, March 18, 2021 4:28 PM

## 4. Learning Rate

Thursday, March 18, 2021 4:28 PM

## 5. Out of Core Learning

Thursday, March 18, 2021 4:28 PM



## 6. Disadvantage

Thursday, March 18, 2021 4:29 PM

1. Tricky to use
2. Risky

## 7. Batch Vs Online Learning

Thursday, March 18, 2021 4:29 PM

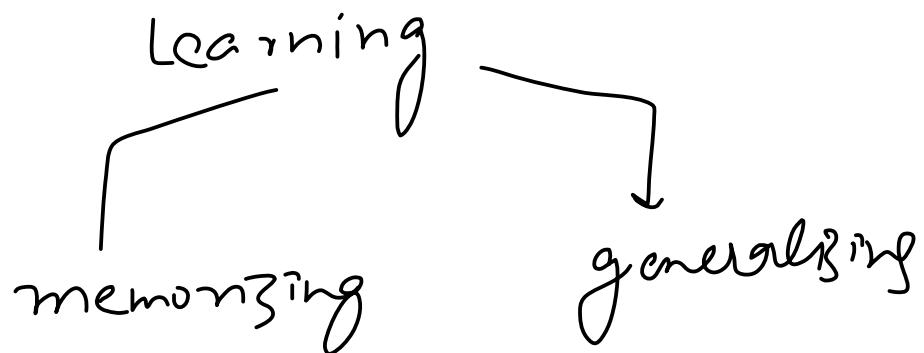
Offline Learning	Features	Online Learning
Less complex as model is constant	Complexity	Dynamic complexity as the model keeps evolving over time
Fewer computations, single time batch-based training	Computational Power	Continuous data ingestions result in consequent model refinement computations
Easier to implement	Use in Production	Difficult to implement and manage
Image Classification or anything related to Machine Learning - where data patterns remains constant without sudden concept drifts	Applications	Used in finance, economics, health where new data patterns are constantly emerging
Industry proven tools. E.g. Sci-kit, TensorFlow, Pytorch, Keras, Spark Mlib	Tools	Active research/New project tools: E.g. MOA, SAMOA, scikit-multiflow, streamDM



Image courtesy - <https://www.iunera.com/kraken/fabric/simple-introduction-to-online-learning-in-machine-learning/>

# 1. Instance Vs Model Based Learning

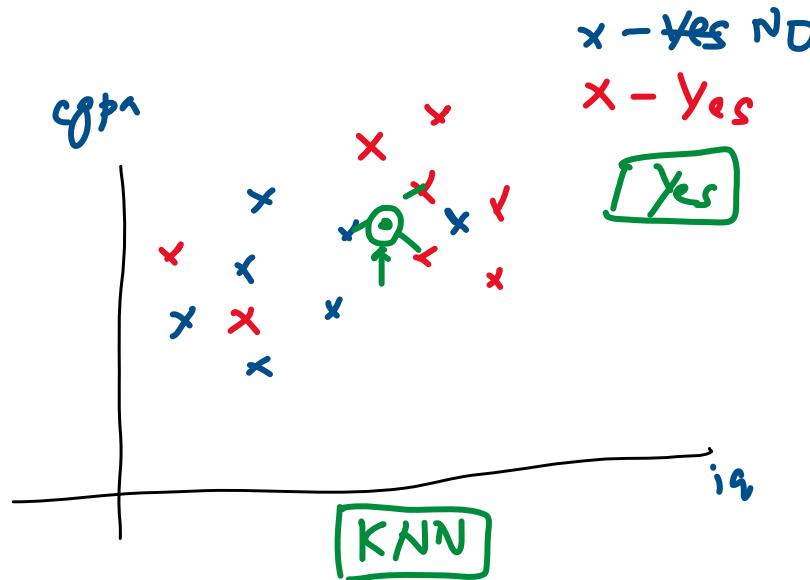
Friday, March 19, 2021 4:05 PM



## 2. Instance Based

Friday, March 19, 2021 4:06 PM

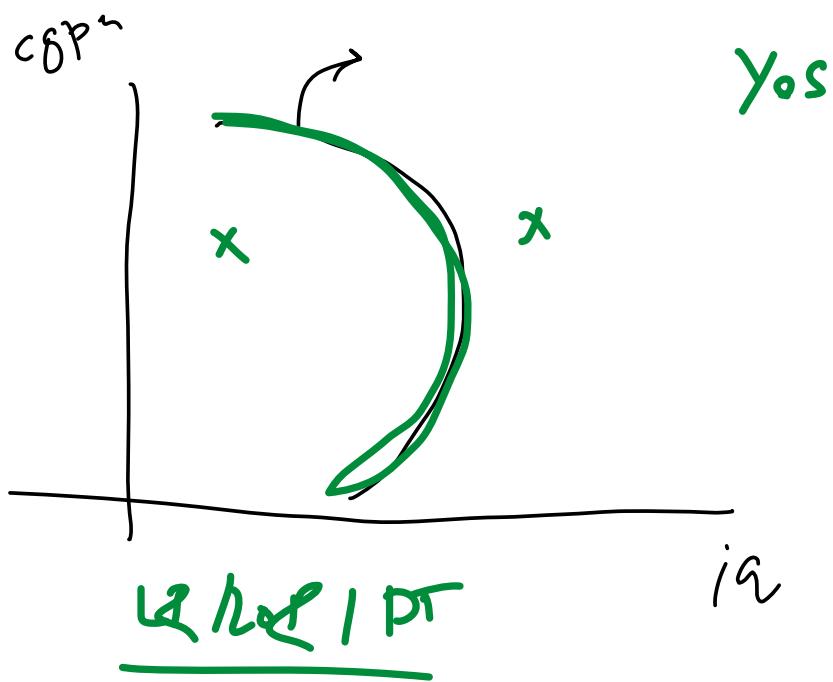
iq	cgp	Placement
80	8	y
70	7	n
7.5	103	



### 3. Model Based

Friday, March 19, 2021 4:06 PM

iq | cgpa | place



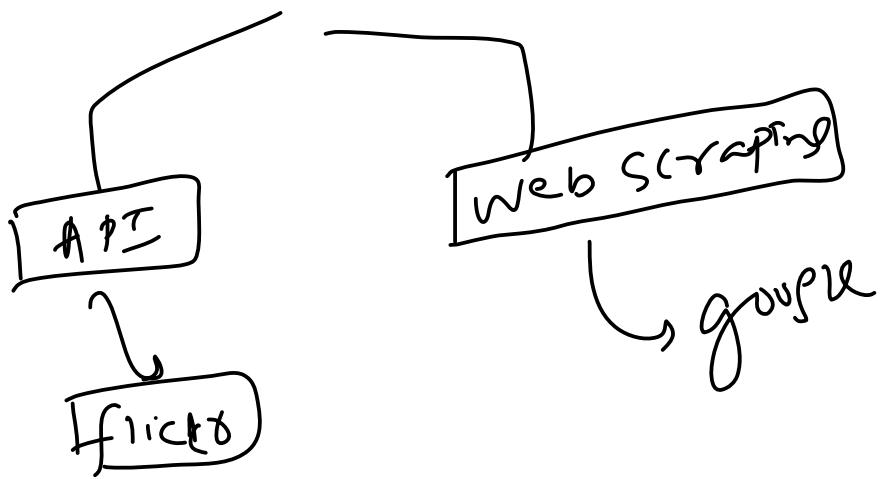
## 4. Differences

Friday, March 19, 2021 4:06 PM

<b>Usual/Conventional Machine Learning</b>	<b>Instance Based Learning</b>
Prepare the data for model training	Prepare the data for model training. No difference here
Train model from training data to estimate model parameters i.e. discover patterns	Do not train model. Pattern discovery postponed until scoring query received
Store the model in suitable form	There is no model to store
Generalize the rules in form of model, even before scoring instance is seen	No generalization before scoring. Only generalize for each scoring instance individually as and when seen
Predict for unseen scoring instance using model	Predict for unseen scoring instance using training data directly
Can throw away input/training data after model training	Input/training data must be kept since each query uses part or full set of training observations
Requires a known model form	May not have explicit model form
Storing models generally requires less storage	Storing training data generally requires more storage

# 1. Data Collection

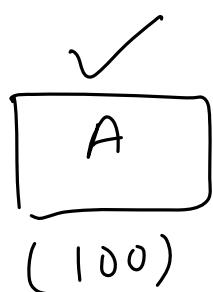
Saturday, March 20, 2021 5:59 PM



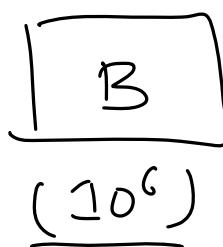
## 2. Insufficient Data/Labelled Data

Saturday, March 20, 2021 6:00 PM

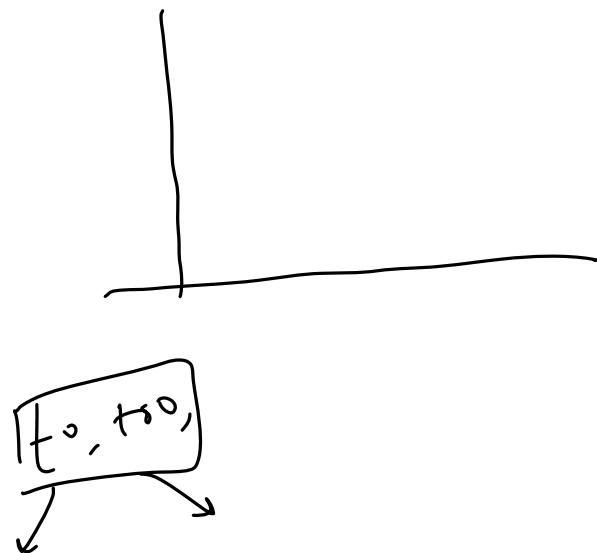
NLP



M1

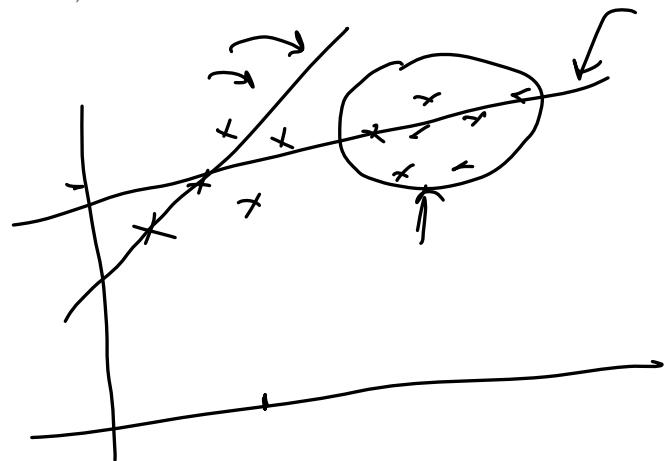


M2 ✓



### 3. Non Representative Data

Saturday, March 20, 2021 6:00 PM



Sampling noise  
Sampling bias

## 4. Poor Quality Data

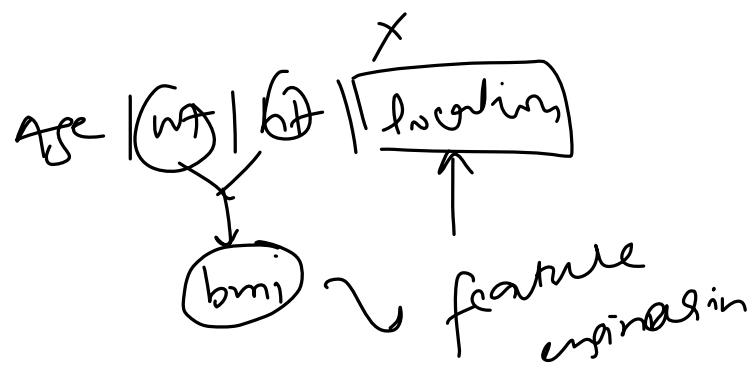
Saturday, March 20, 2021 6:00 PM

60 %

## 5. Irrelevant Features

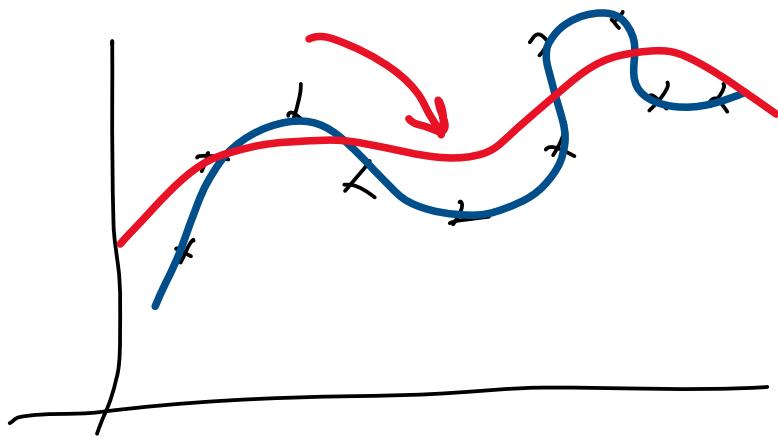
Saturday, March 20, 2021 6:00 PM

Garbage In  
Garbage Out



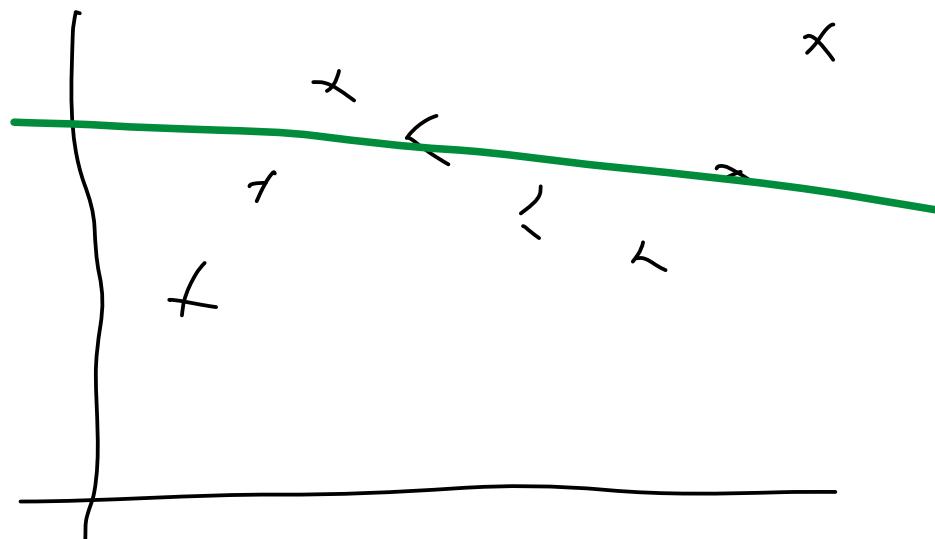
## 6. Overfitting

Saturday, March 20, 2021 6:01 PM



## 7. Underfitting

Saturday, March 20, 2021 6:01 PM

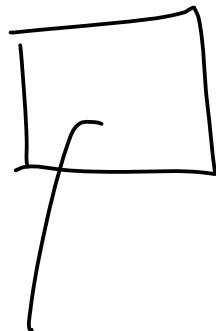


## 8. Software Integration

Saturday, March 20, 2021 6:01 PM

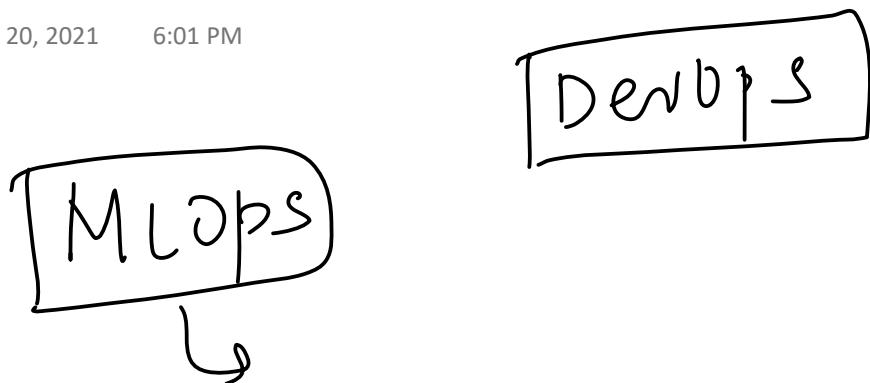
## 9. Offline Learning/ Deployment

Saturday, March 20, 2021 6:01 PM



## 10. Cost Involved

Saturday, March 20, 2021 6:01 PM



# 1. Retail - Amazon/Big Bazaar

Monday, March 22, 2021 6:07 PM



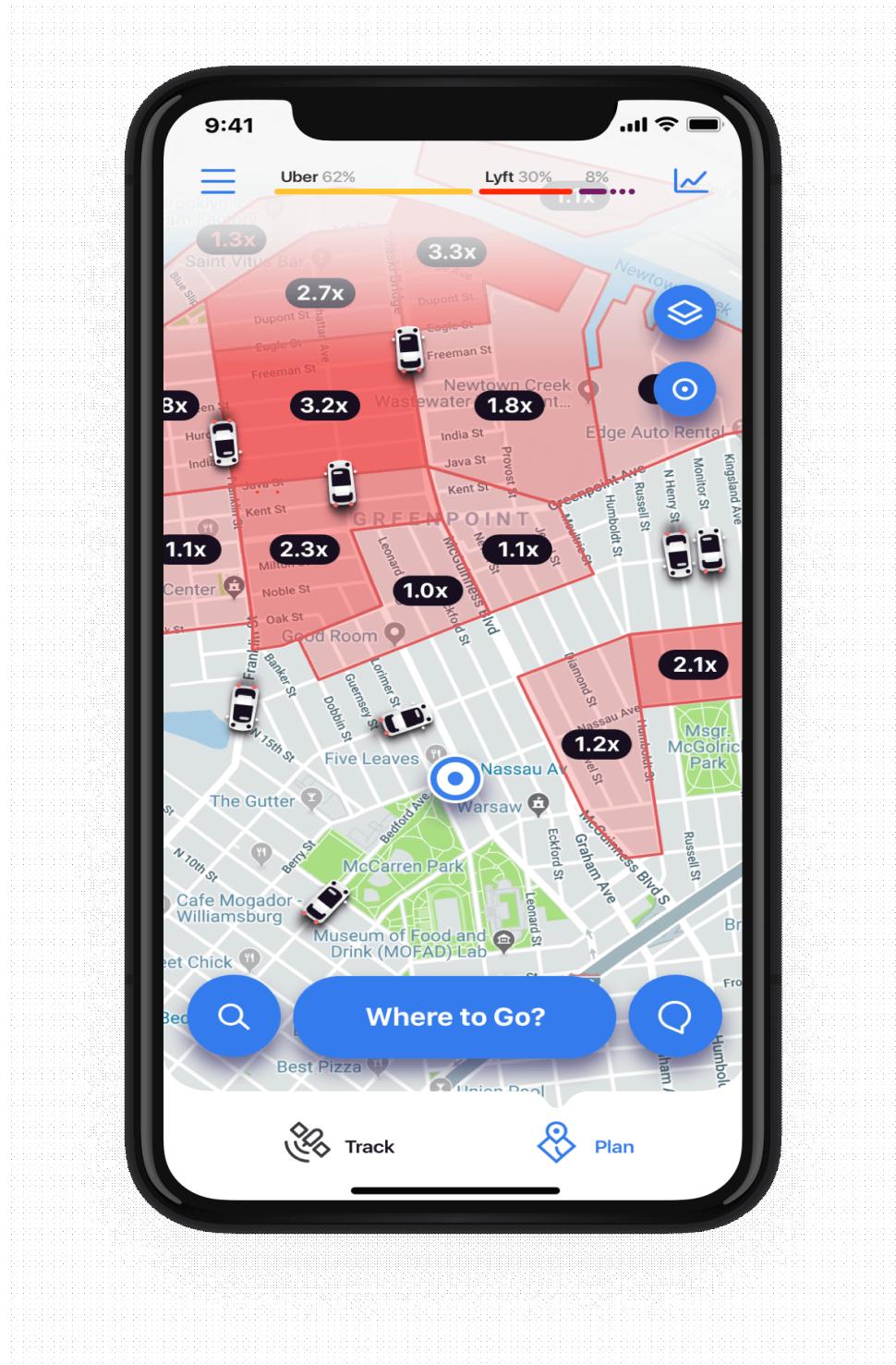
## 2. Banking and Finance

Monday, March 22, 2021 6:07 PM



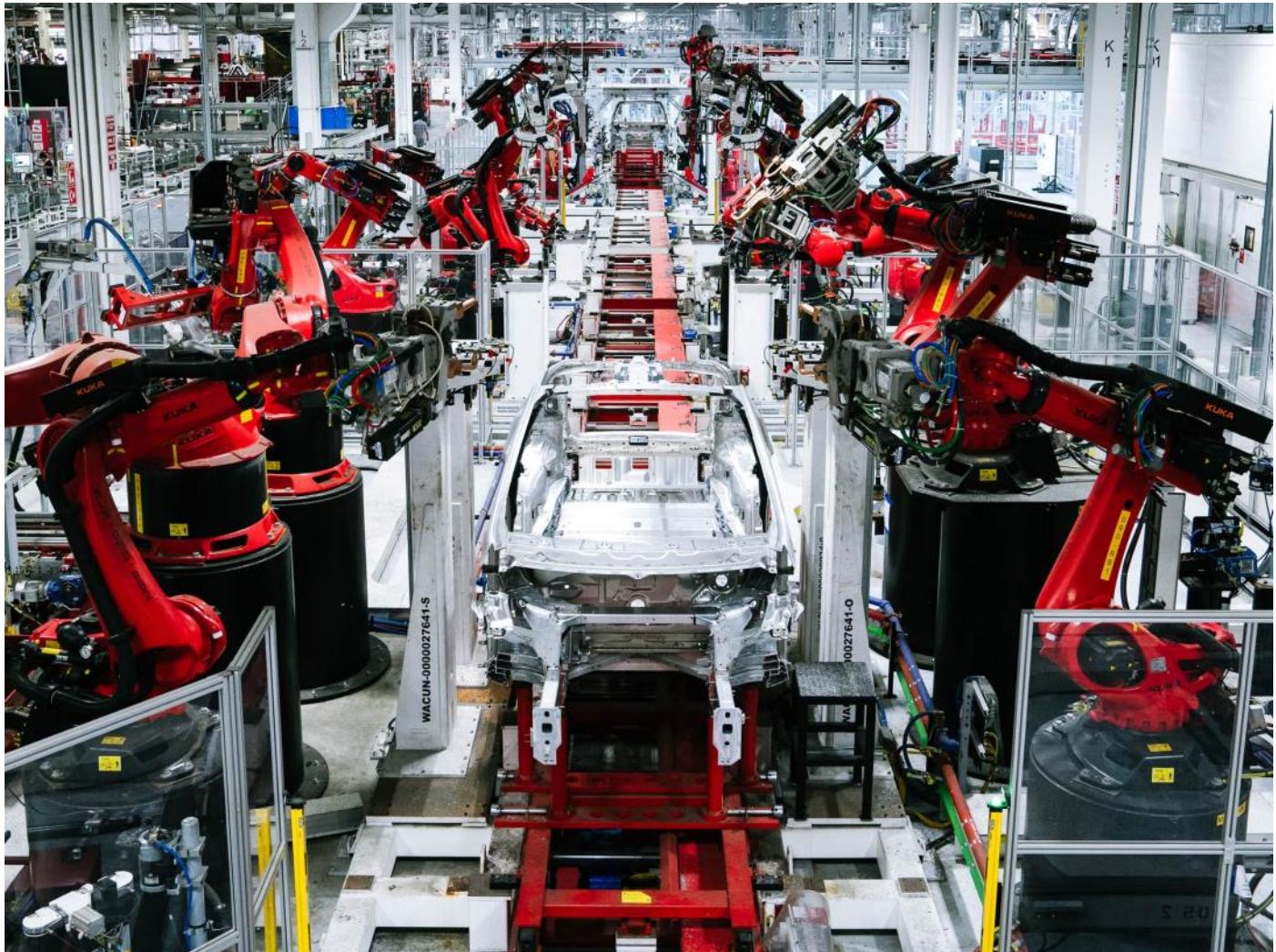
### 3. Transport - OLA

Monday, March 22, 2021 6:07 PM



## 4. Manufacturing - Tesla

Monday, March 22, 2021 6:08 PM



## 5. Consumer Internet - Twitter

Monday, March 22, 2021 6:08 PM



# Machine Learning Development Life Cycle(MLDLC/MLDC)

Tuesday, March 23, 2021 12:09 PM

SDLC

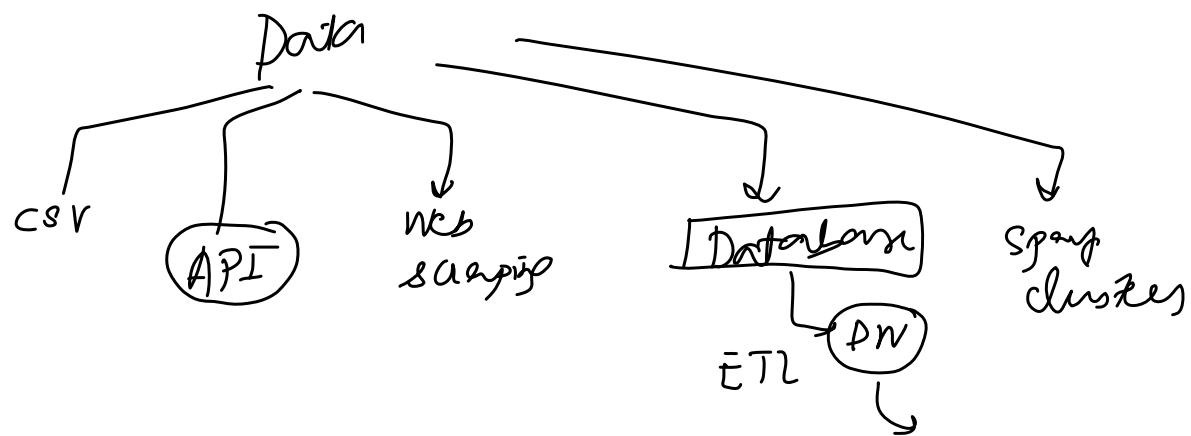
ML DLC

# 1. Frame the Problem

Tuesday, March 23, 2021 12:10 PM

## 2. Gathering Data

Tuesday, March 23, 2021 12:11 PM



### 3. Data Preprocessing

Tuesday, March 23, 2021 12:11 PM

- Remove duplicates
- Remove missing val
- Outliers
- Scale

## 4. Exploratory Data Analysis

Tuesday, March 23, 2021 12:11 PM

Vizs

Univariate/Bivariate

Outlier detection

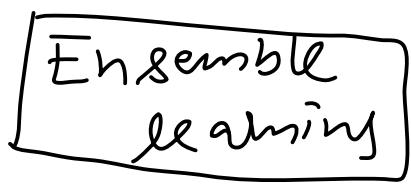
Imbalance →

## 5. Feature Engineering and Selection

Tuesday, March 23, 2021 12:12 PM

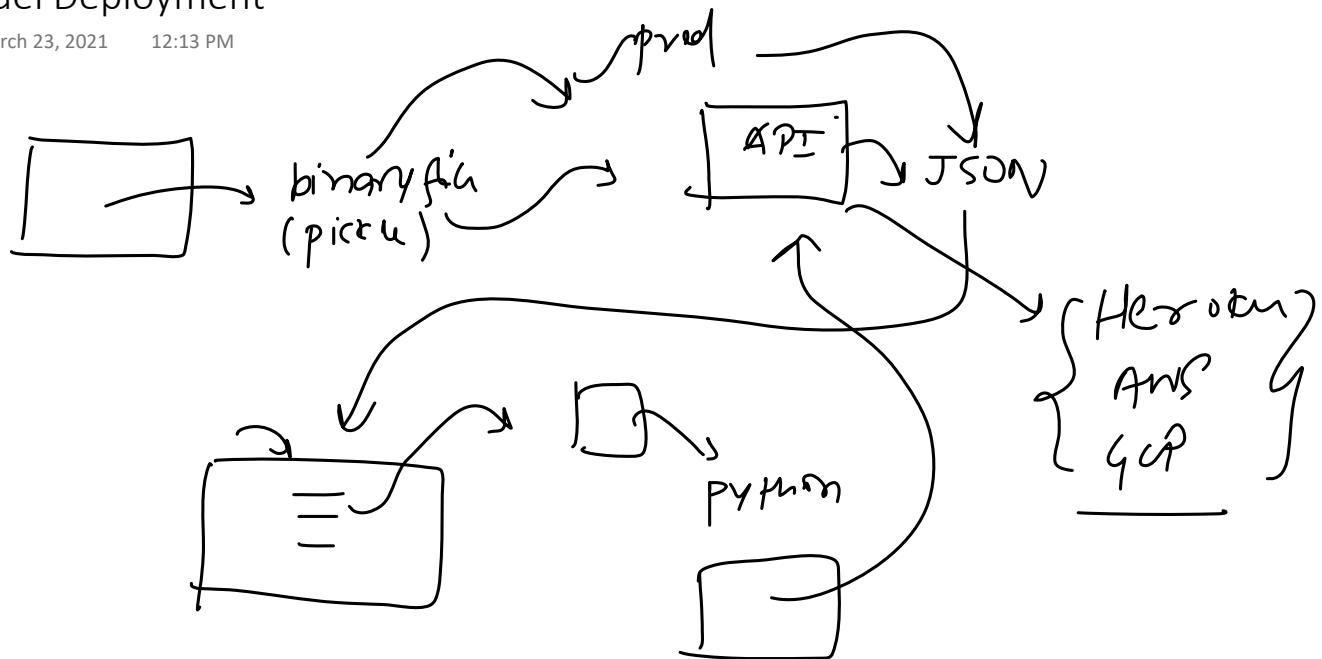
## 6. Model Training, Evaluation and Selection

Tuesday, March 23, 2021 12:12 PM



## 7. Model Deployment

Tuesday, March 23, 2021 12:13 PM



## 8. Testing

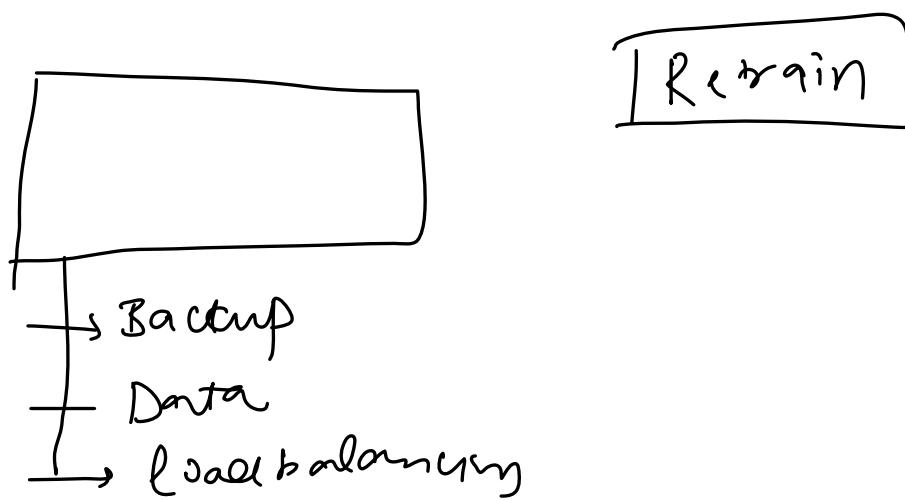
Tuesday, March 23, 2021 12:14 PM

A/B testing

## 9. Optimize

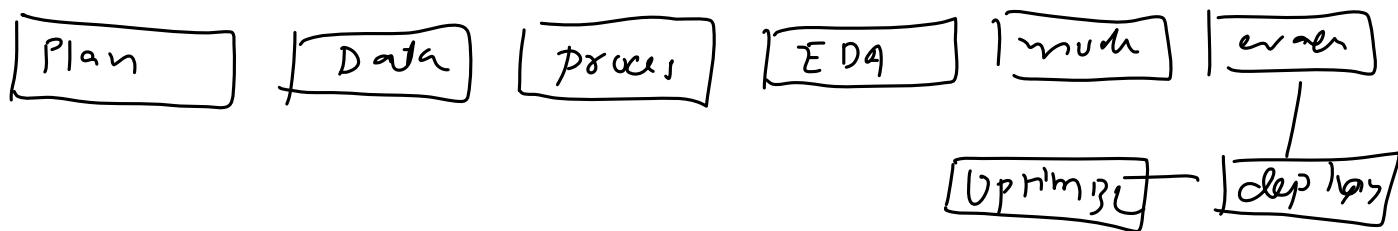
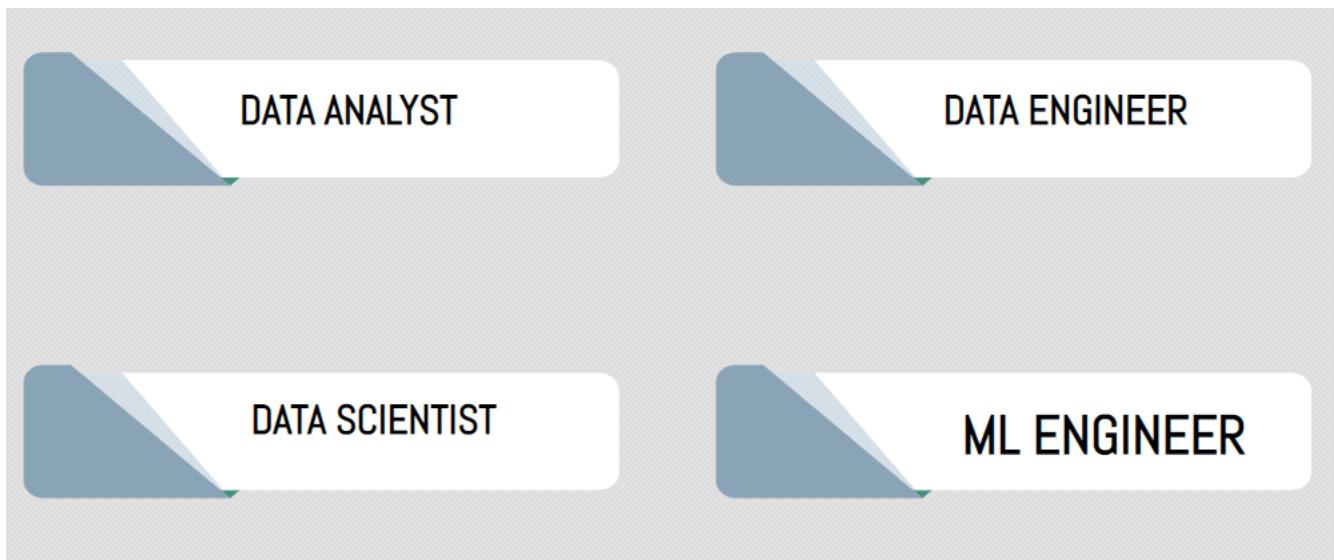
Tuesday, March 23, 2021 12:15 PM

1 Rotting



# 1. Various Data Based Job Roles

Wednesday, March 24, 2021 1:25 PM



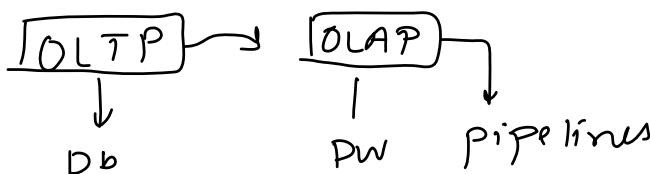
# 1. Data Engineer

Wednesday, March 24, 2021 1:25 PM

## Skills Required

### Job Roles

- Scrape Data from the given sources.
- Move/Store the data in optimal servers/warehouses.
- Build data pipelines/APIs for easy access to the data.
- Handle databases/data warehouses.



- Strong grasp of algorithms and data structures
- Programming Languages (Java/R/Python/Scala) and script writing
- Advanced DBMS's
- BIG DATA Tools (Apache Spark, Hadoop, Apache Kafka, Apache Hive)
- Cloud Platforms (Amazon Web Services, Google Cloud Platform)
- Distributed Systems
- Data Pipelines

## 2. Data Analyst

Wednesday, March 24, 2021 1:26 PM

### Skills

- *Statistical Programming*
- *Programming Languages (R/SAS/Python)*
- *Creative and Analytical Thinking*
- *Business Acumen — Medium to High preferred*
- *Strong Communication Skills.*
- *Data Mining, Cleaning, and Munging*
- *Data Visualization*
- *Data Story Telling*
- *SQL*
- *Advanced Microsoft Excel*

### Responsibilities of a Data Analyst

- *Cleaning and organizing Raw data.*
- *Analyzing data to derive insights.*
- *Creating data visualizations.*
- *Producing and maintaining reports.*
- *Collaborating with teams/colleagues based on the insight gained.*
- *Optimizing data collection procedures*

### 3. Data Scientist

Wednesday, March 24, 2021 1:26 PM

“A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician”.

## 4. ML Engineer

Wednesday, March 24, 2021 1:26 PM

### Responsibilities

- Deploying machine learning models to production ready environment
- Scaling and optimizing the model for production
- Monitoring and maintenance of deployed models

### Skills

- Mathematics
- Programming Languages (R/Python/Java/Scala mainly)
- Distributed Systems
- Data model and evaluation
- Machine Learning models
- Software Engineering & Systems design

## 5. Comparison

Wednesday, March 24, 2021 1:26 PM

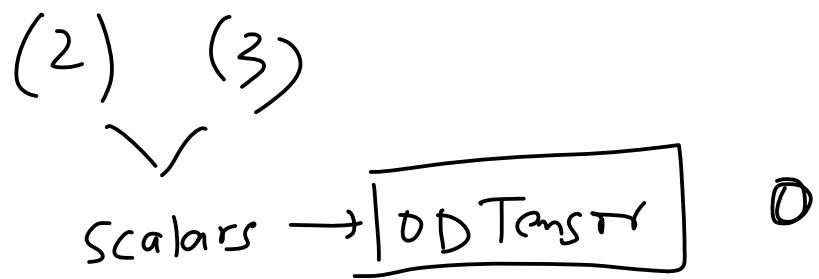
	ANALYTICAL SKILLS	BUSINESS ACUMEN	DATA STORYTELLING	SOFT SKILLS	SOFTWARE SKILLS
DATA ANALYST	HIGH	MEDIUM TO HIGH	HIGH	MEDIUM TO HIGH	MEDIUM
DATA ENGINEER	MEDIUM	LOW	LOW	MEDIUM	HIGH
DATA SCIENTIST	HIGH	HIGH	HIGH	HIGH	MEDIUM
ML ENGINEER	MEDIUM TO HIGH	MEDIUM	LOW	HIGH	HIGH

# 1. What are Tensors

Thursday, March 25, 2021 4:44 PM

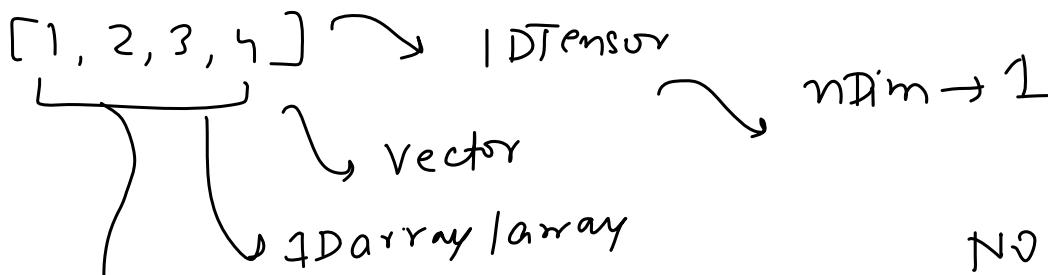
## 2. 0D Tensor/Scalar

Thursday, March 25, 2021 4:44 PM



### 3. 1D Tensor/Vector

Thursday, March 25, 2021 4:45 PM



$nDim \rightarrow 1$

Axis  
2 Dim

No. of axes = rank  
= dim

$[1, 2] \rightarrow$  vector (2)  
1D tensor

$[0, 1, 2, 3]$

4 scalars  $\rightarrow$  vector

## 4. 2D Tensor/Matrices

Thursday, March 25, 2021 4:45 PM

$$\begin{bmatrix} 1, 2, 3 \end{bmatrix} \quad \begin{bmatrix} 4, 5, 6 \end{bmatrix} \quad \begin{bmatrix} 7, 8, 9 \end{bmatrix}$$

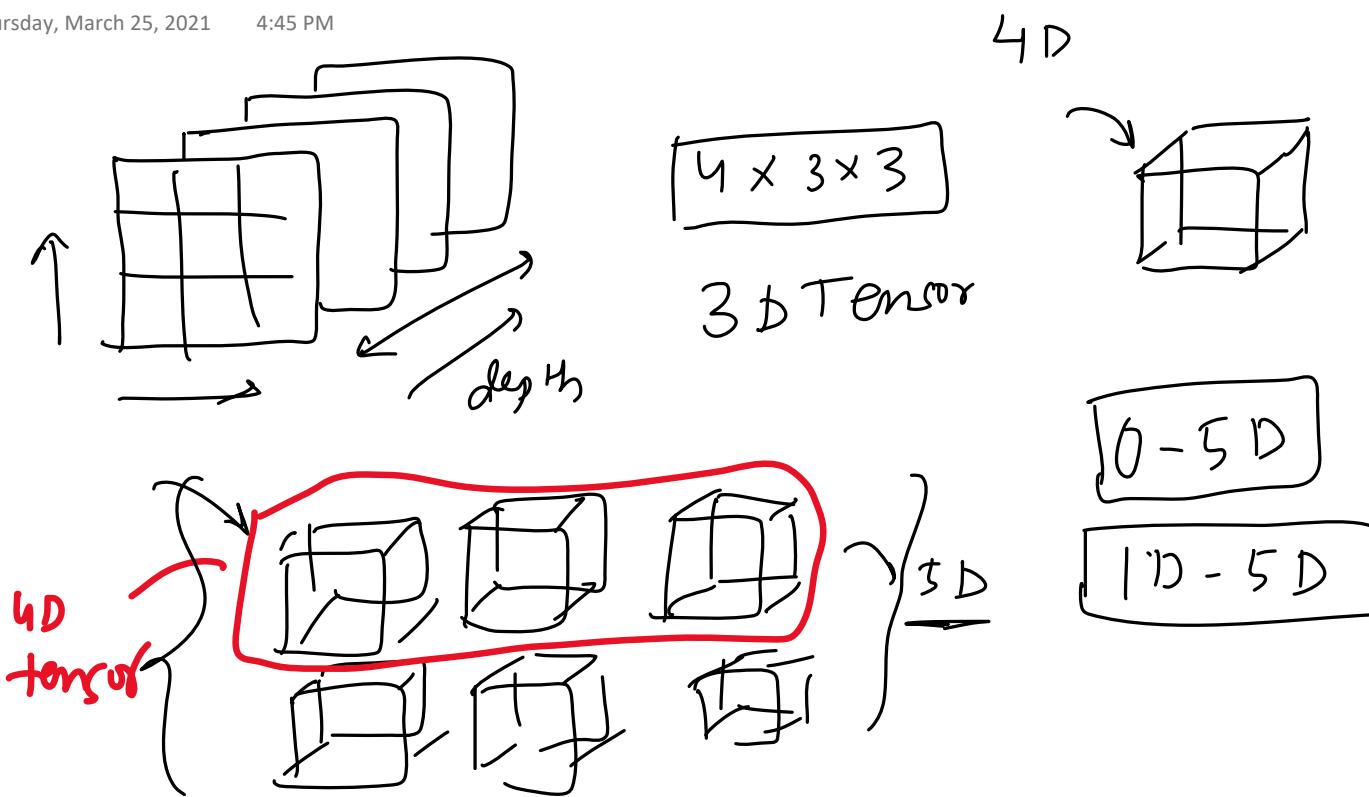
$\xrightarrow{\quad}$

$$\downarrow \left[ \begin{array}{c} [1, 2, 3] \\ [4, 5, 1] \\ [7, 8, 9] \end{array} \right] \curvearrowright 2^D$$

Rank = 2 = n^{dim}

## 5. ND Tensors

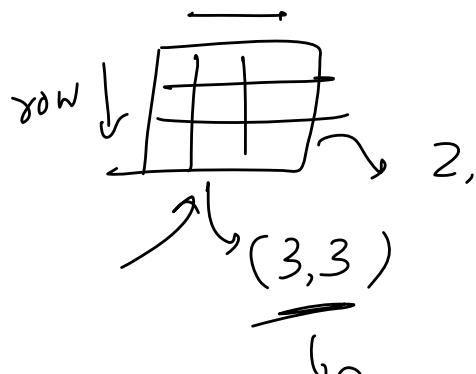
Thursday, March 25, 2021 4:45 PM



## 6. Rank, Axes and Shape

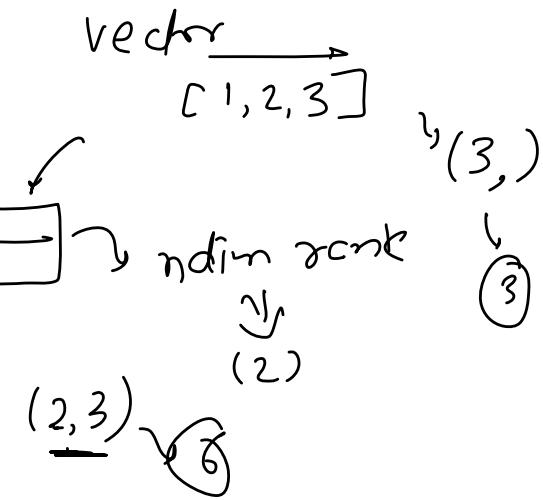
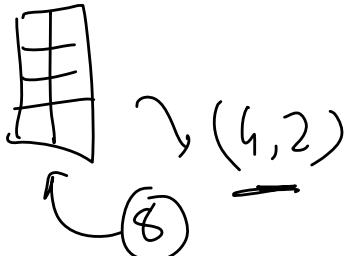
Thursday, March 25, 2021 4:45 PM

No. of axis = Rank = No. of dim



Size = 1

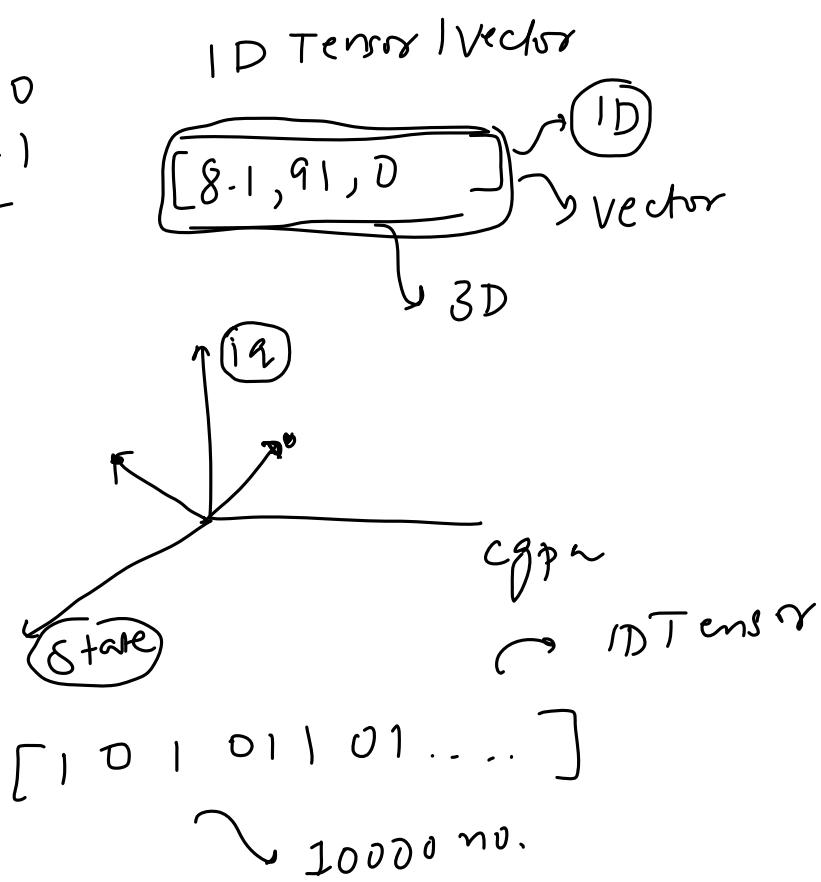
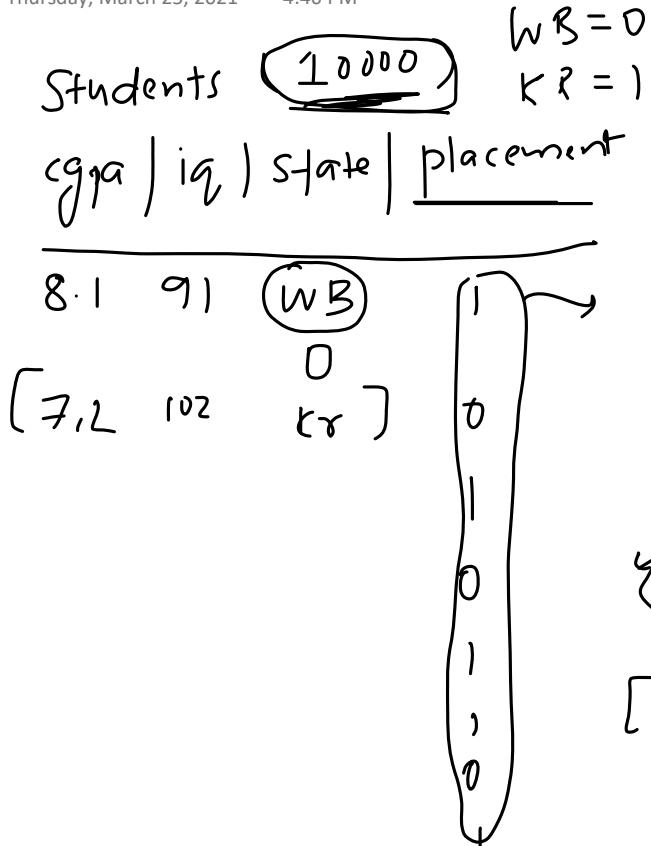
Shape



Size of tensor

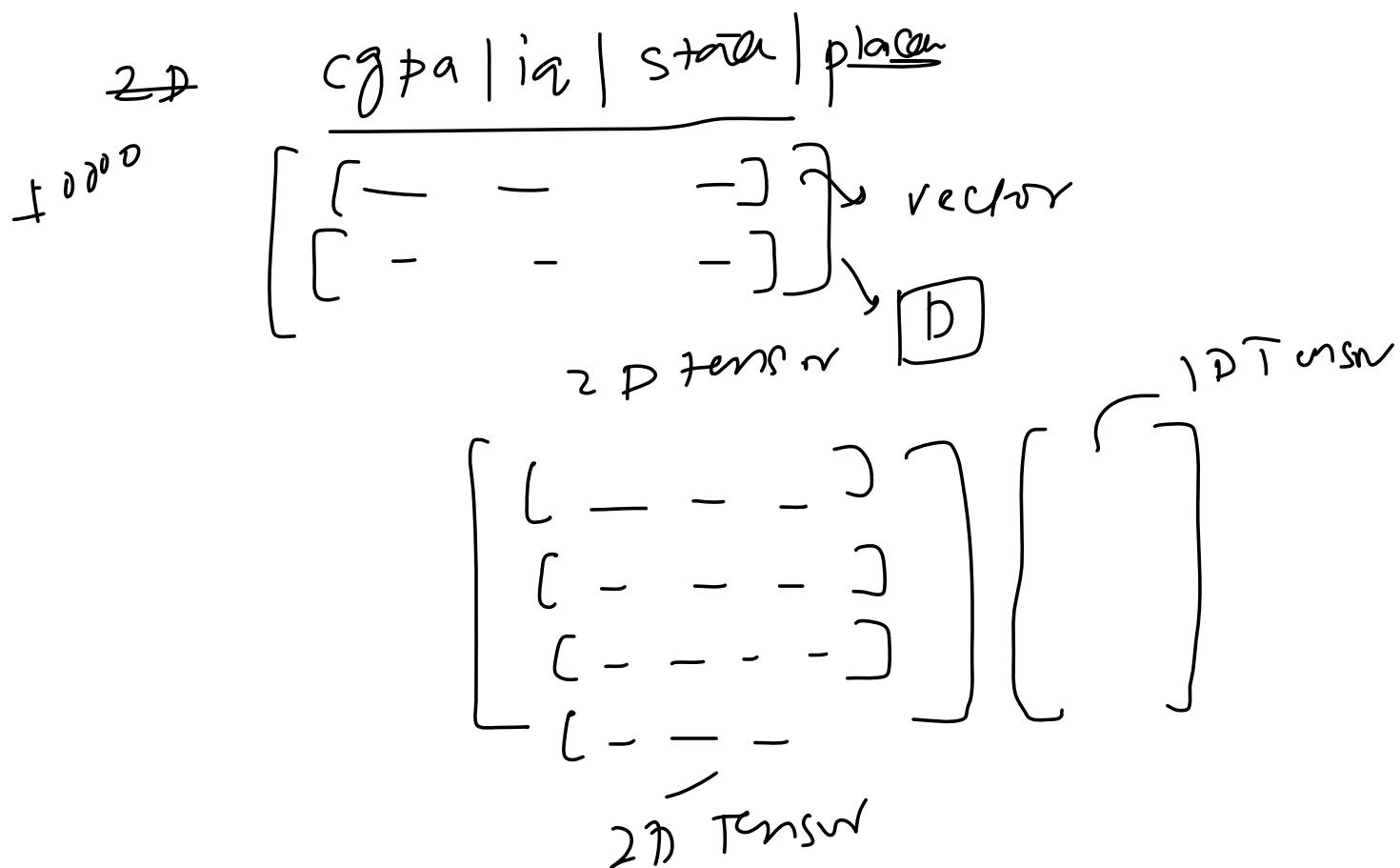
## 7. Example of 1D Tensors

Thursday, March 25, 2021 4:46 PM



## 8. Example of 2D Tensors

Thursday, March 25, 2021 4:46 PM



## 9. Example of 3D Tensors

Thursday, March 25, 2021 4:46 PM

<b>NLP</b>
<u>Hi Nitish</u>
<u>Hi Rahul</u>
<u>Hi Ankit</u>

Hi	Nitish	Rahul	Ankit
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

$$\begin{array}{c}
 \xrightarrow{\quad} [[1,0,0,0], [0,1,0,0]] \quad \xrightarrow{\text{2D}} (3,2,4) \\
 [[1,0,0,0], [0,0,1,0]] \quad \xrightarrow{\text{3D Tensor}} \\
 [[1,0,0,0], [0,0,0,1]] \quad \xrightarrow{\quad}
 \end{array}$$

Timeseries Data  $\xrightarrow{\quad} 70 \text{ years}$

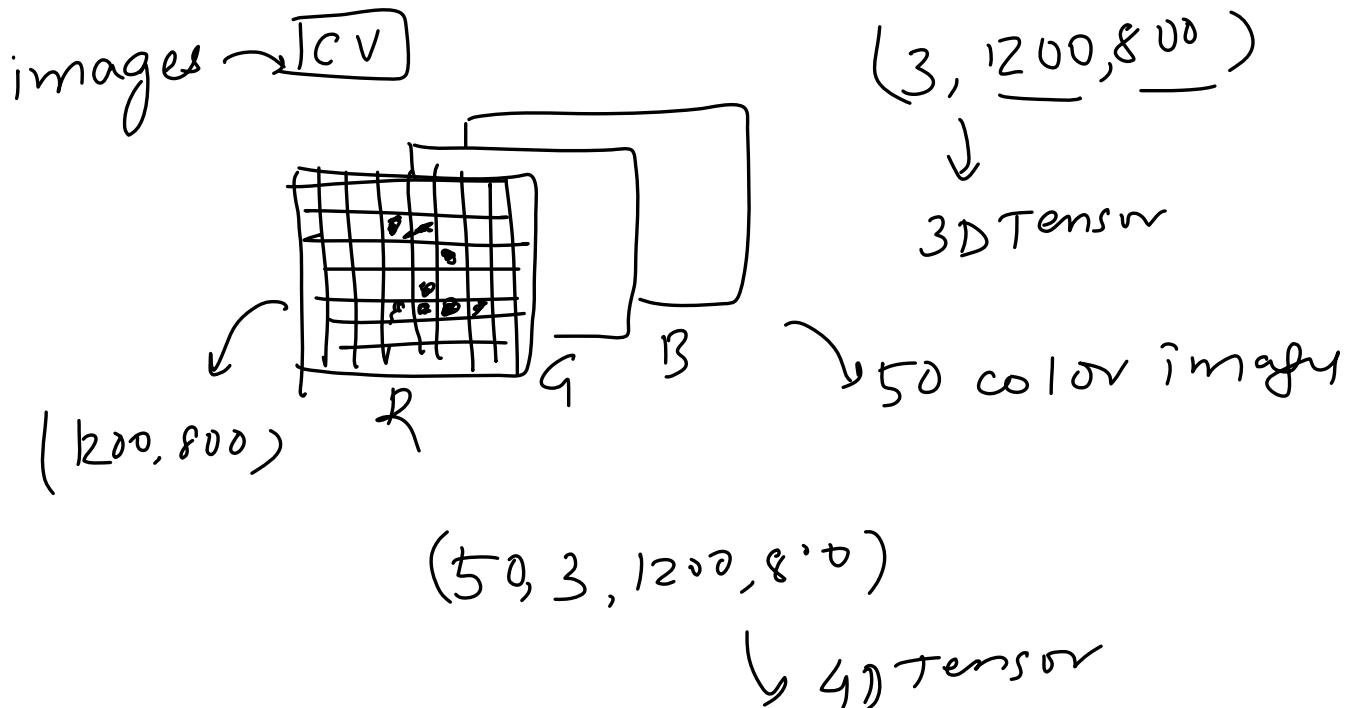
Highest	lowest	$(365, 2)$
---------	--------	------------

Day 1	.	—	—	
Day 2	—	—	—	$\rightarrow 2D \rightarrow 10$
Day 3	—	—	—	$(10, 365, 2)$
365				$\hookrightarrow 3D \text{ Tensor}$

time axis

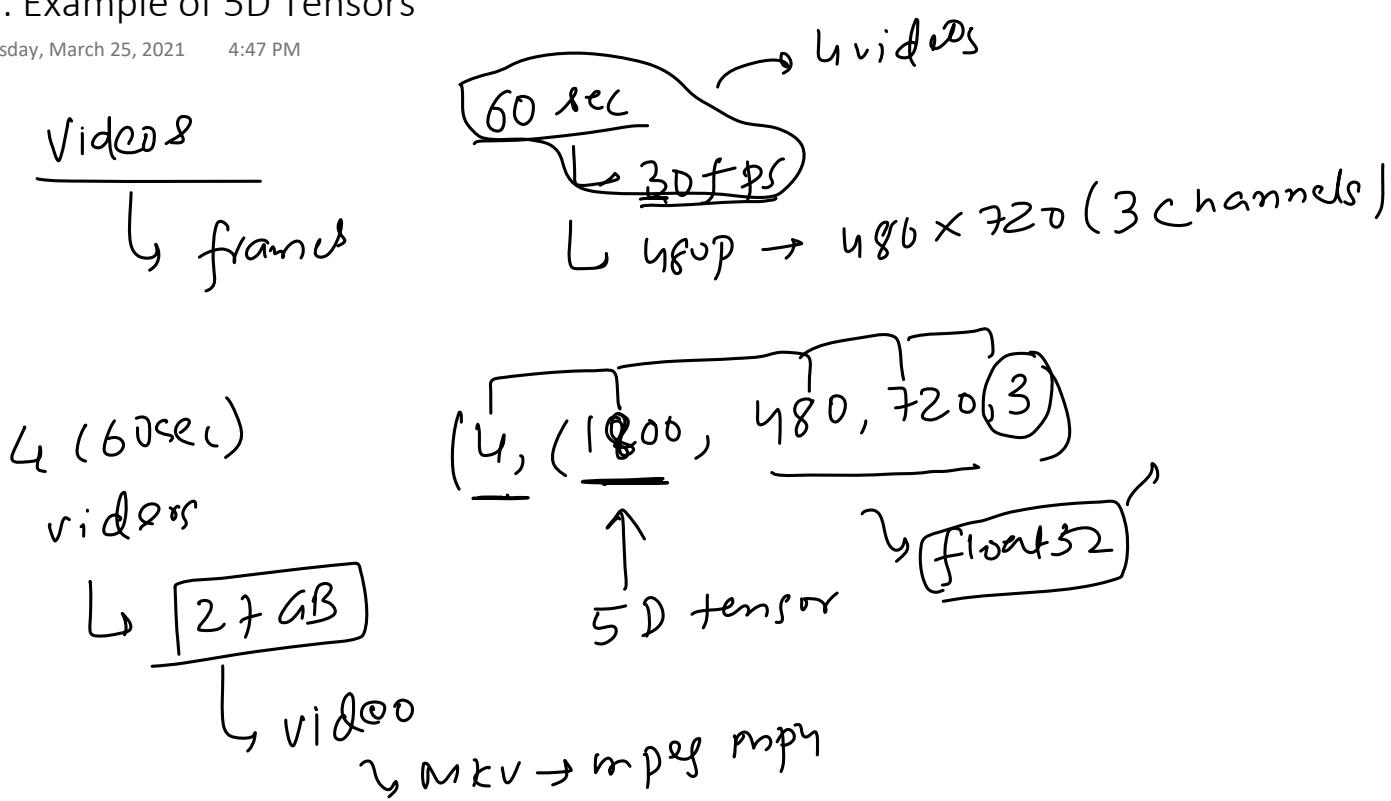
## 10. Example of 4D Tensors

Thursday, March 25, 2021 4:47 PM



## 12. Example of 5D Tensors

Thursday, March 25, 2021 4:47 PM



# 1. Installing Anaconda

Friday, March 26, 2021 5:40 PM

## 2. Jupyter Notebook Intro

Friday, March 26, 2021 5:40 PM

### 3. Virtual Env

Friday, March 26, 2021 5:40 PM

## 4. Using Kaggle

Friday, March 26, 2021 5:41 PM

## 5. Using Google Colab

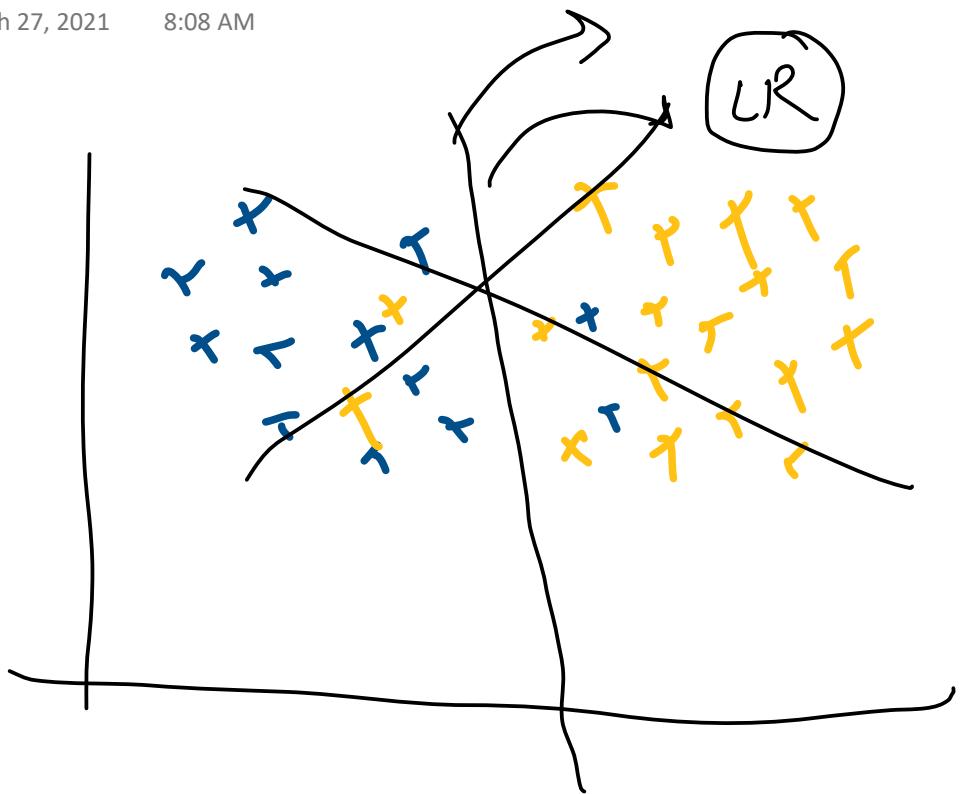
Friday, March 26, 2021 5:41 PM

## 6. Running Kaggle Data on Google Colab

Friday, March 26, 2021 5:41 PM

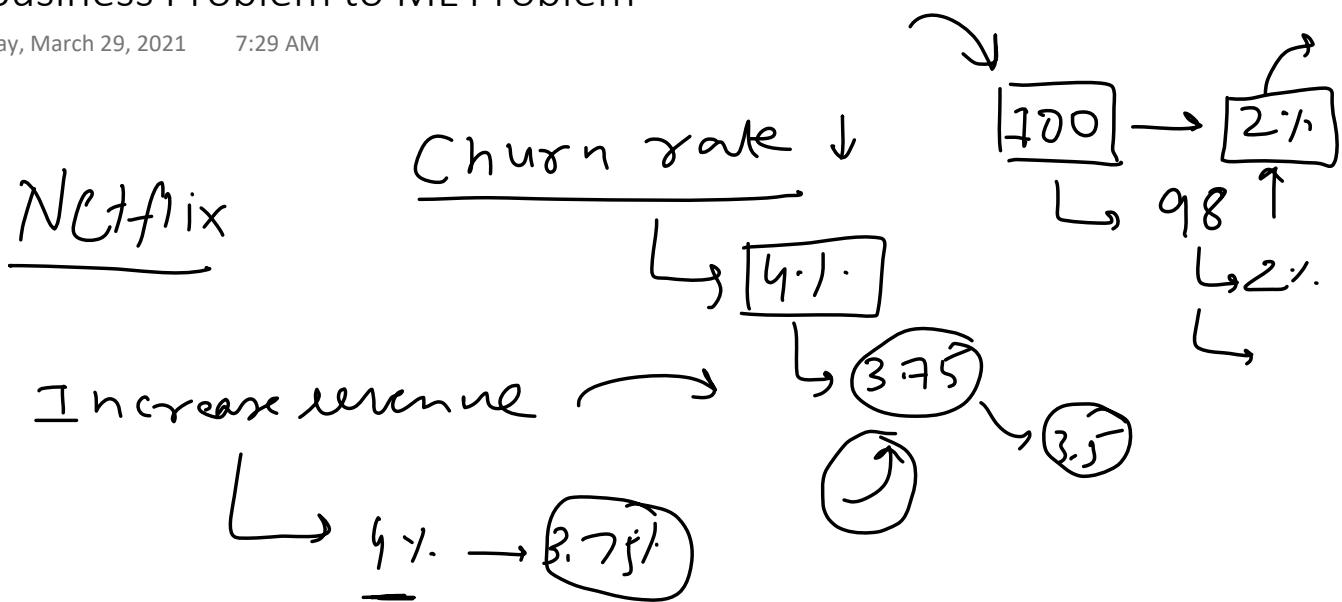
# End to End Example

Saturday, March 27, 2021 8:08 AM



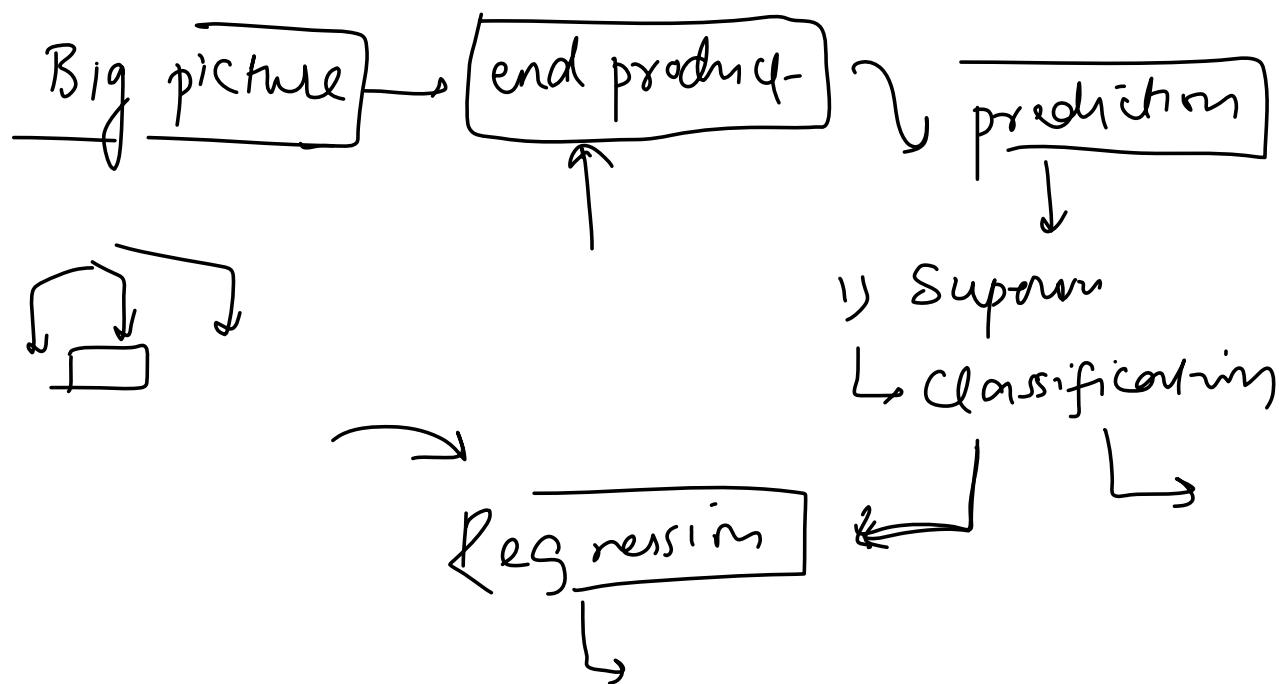
## 1. Business Problem to ML Problem

Monday, March 29, 2021 7:29 AM



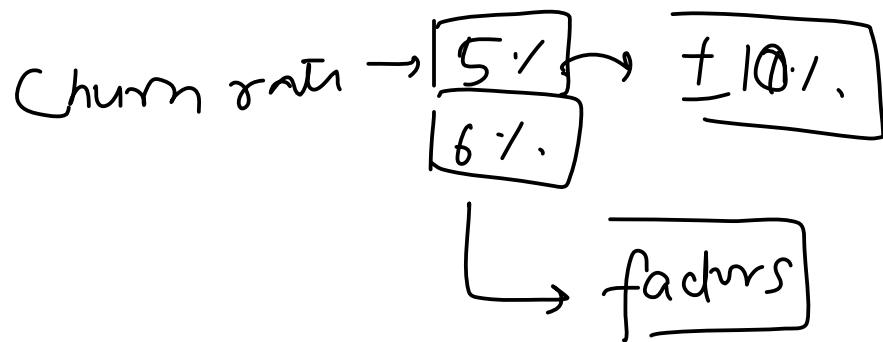
## 2. Type of Problem

Monday, March 29, 2021 7:30 AM



### 3. Current Solution

Monday, March 29, 2021 7:30 AM

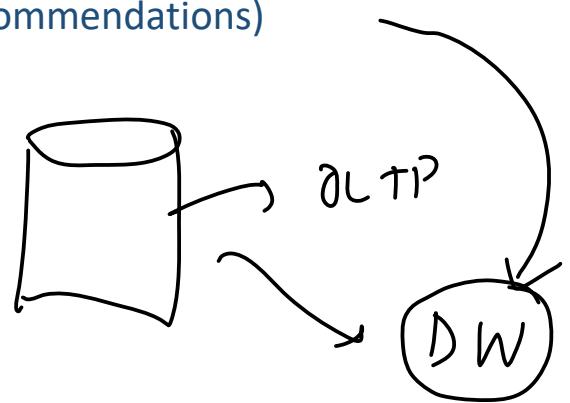


## 4. Getting Data

Monday, March 29, 2021 7:30 AM

1. Watch time
2. Search but did not find
3. Content left in the middle
4. Clicked on recommendations(order of recommendations)

Data engineer

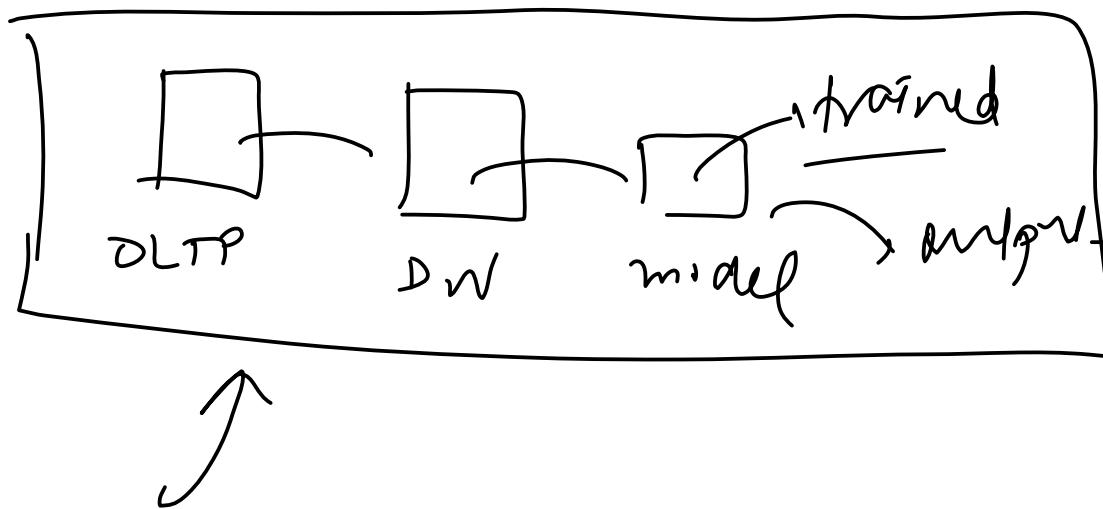


## 5. Metrics to measure

Monday, March 29, 2021 7:30 AM

## 6. Online Vs Batch?

Monday, March 29, 2021 7:31 AM

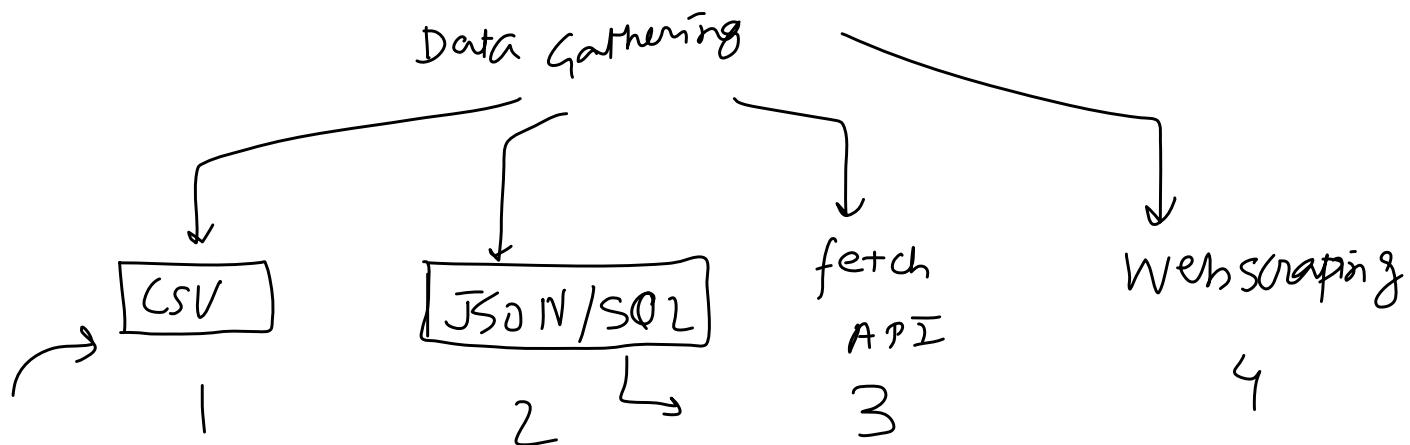


## 7. Check Assumptions

Monday, March 29, 2021 7:31 AM

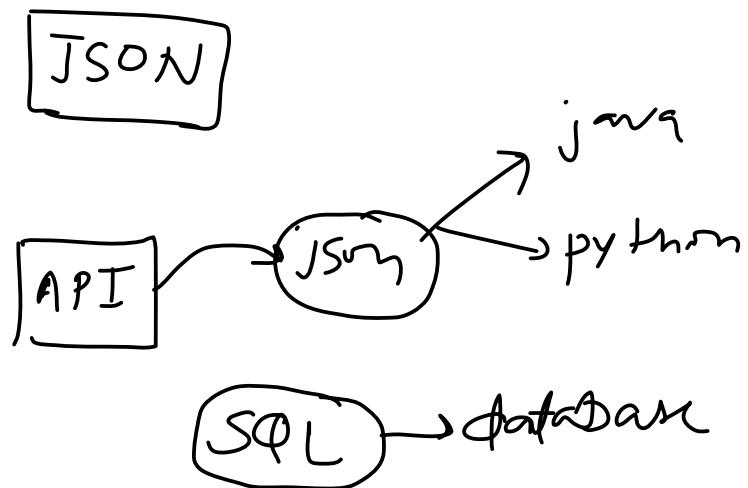
## Gathering Data

Tuesday, March 30, 2021 5:35 PM



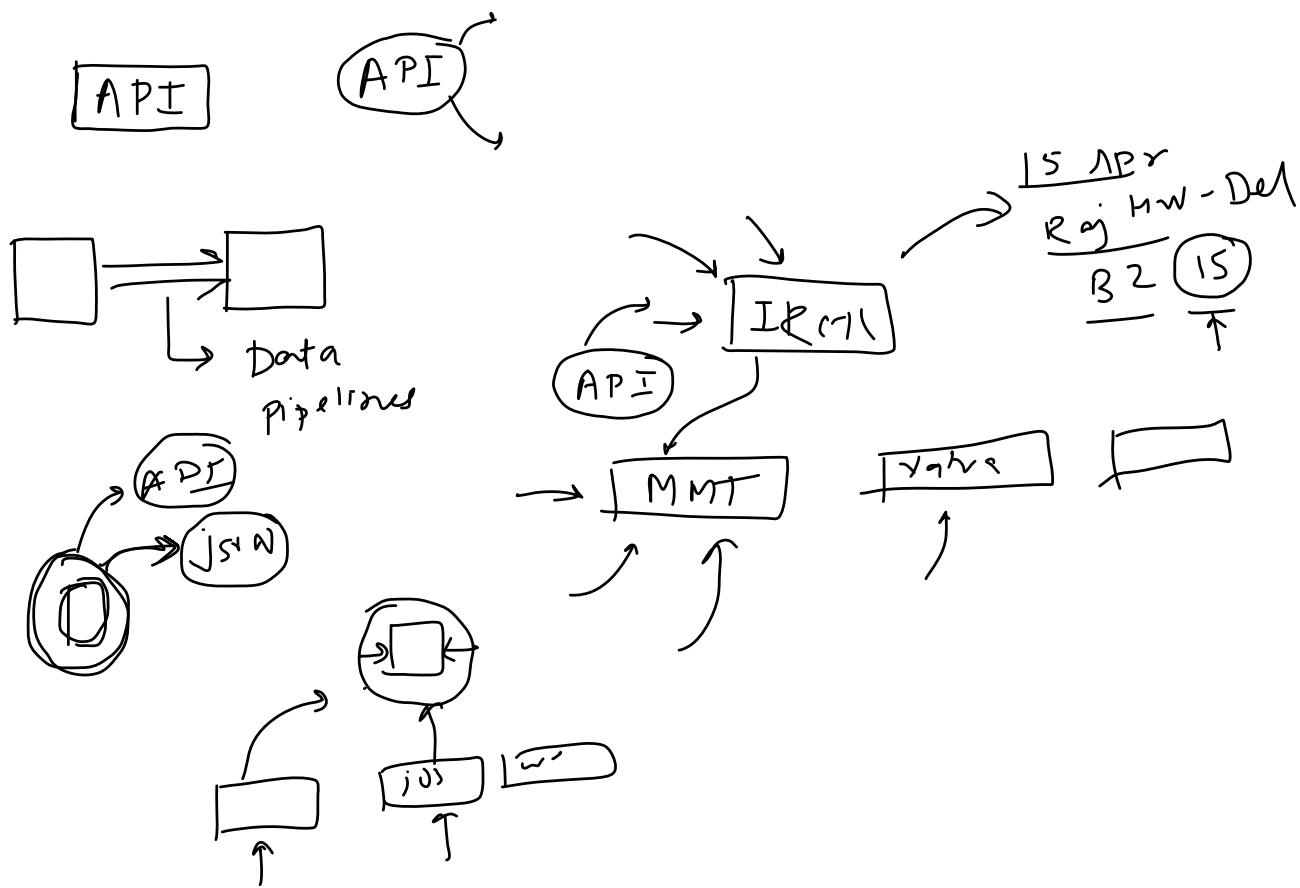
# JSON/SQL

Wednesday, March 31, 2021 8:10 AM



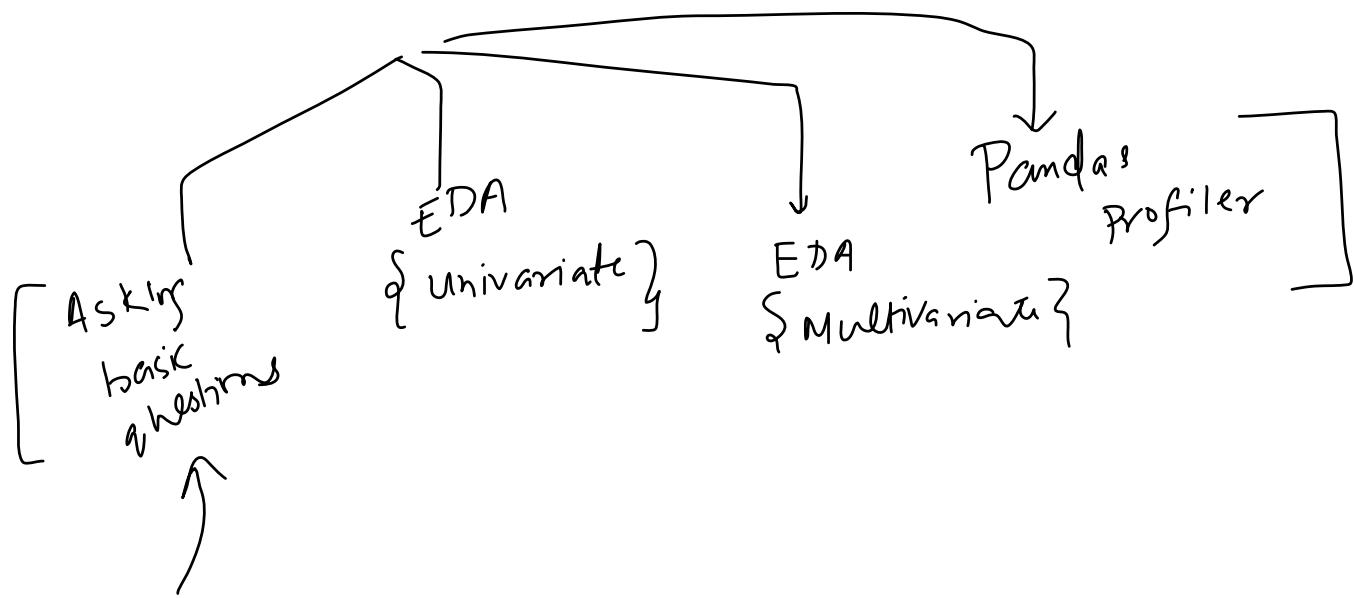
# What is an API

Thursday, April 1, 2021 6:55 AM



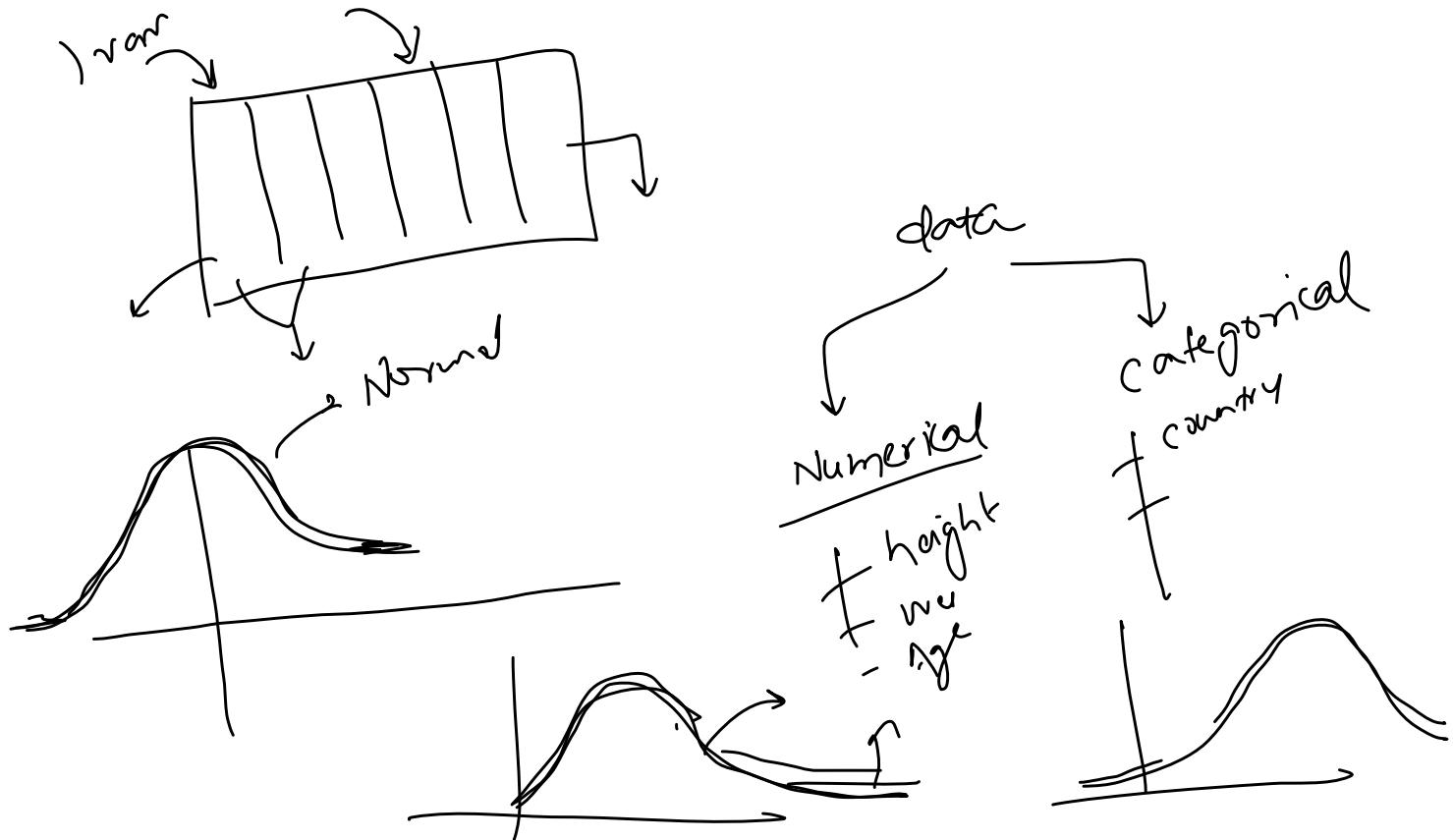
# Understanding Your Data

Saturday, April 3, 2021 9:07 AM



# Univariate Analysis

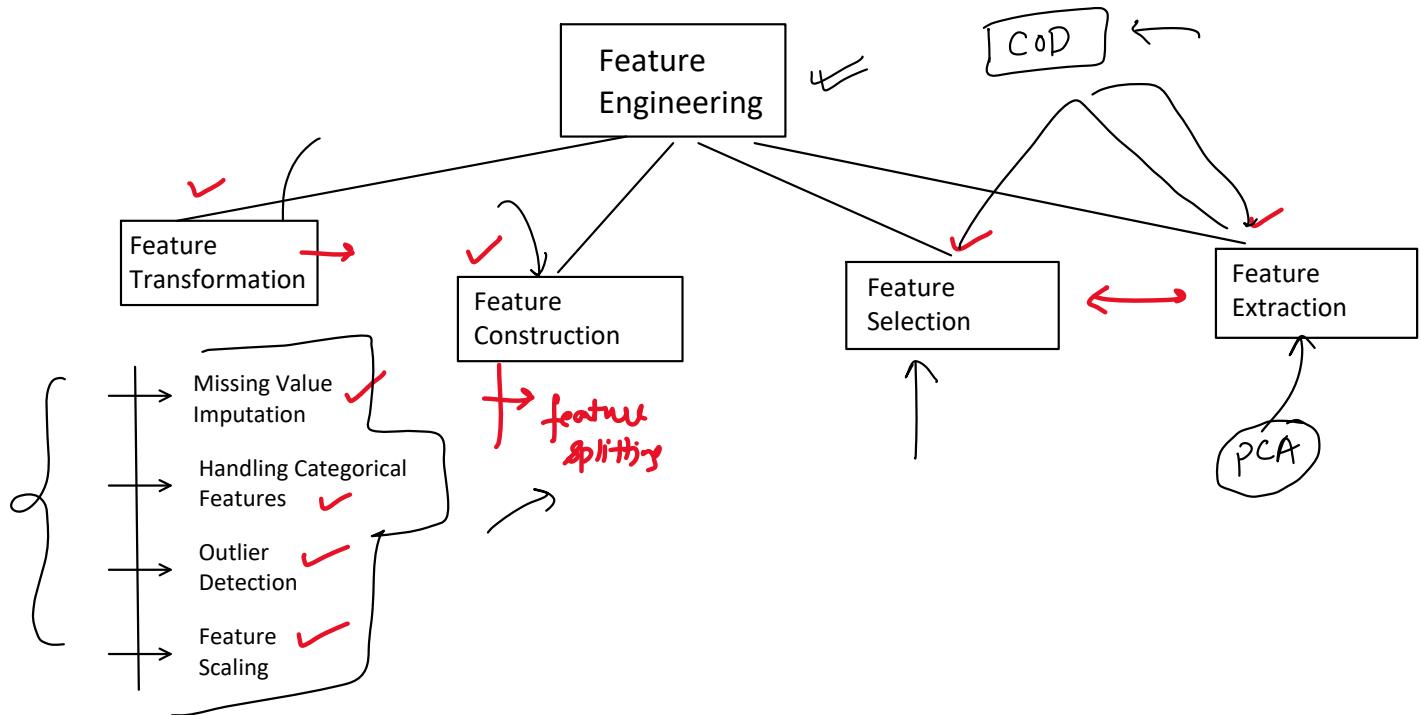
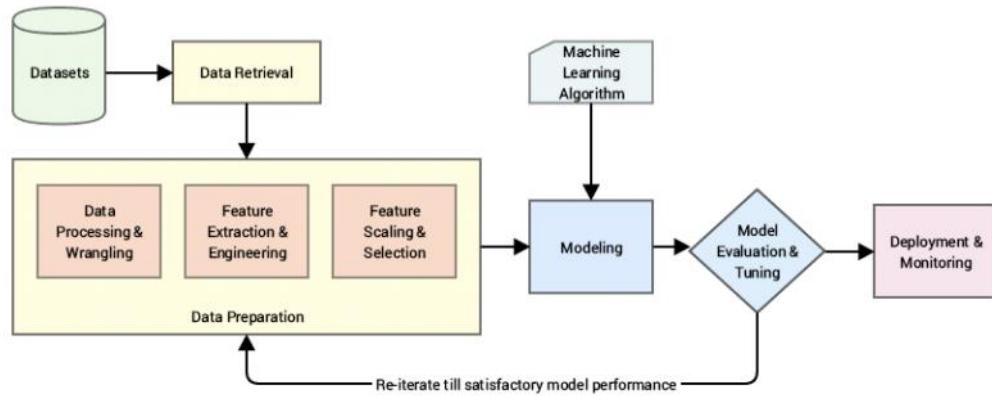
Sunday, April 4, 2021 5:39 PM



# Feature Engineering

Thursday, April 8, 2021 6:02 PM

Feature engineering is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning algorithms.



## 1.1 Missing Values Imputation

Thursday, April 8, 2021 7:05 PM

**Average\_Age = 26.0**

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1



ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

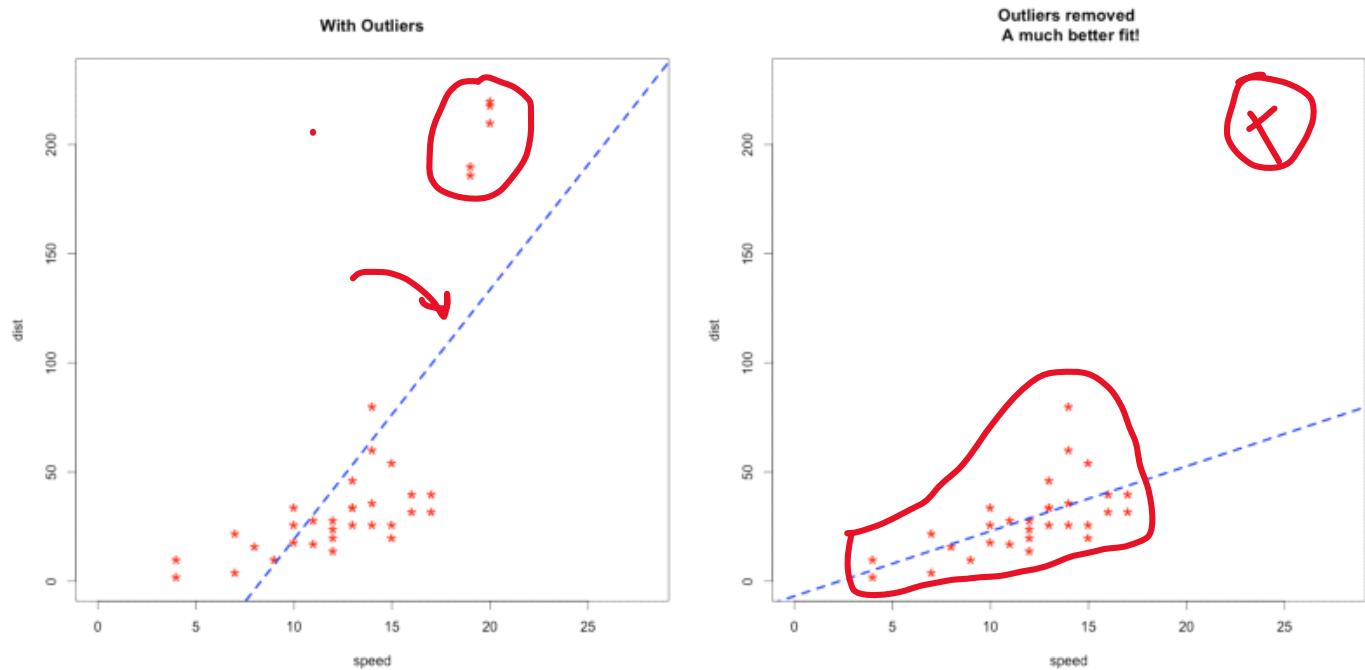
## 1.2 Handling Categorical Values

Thursday, April 8, 2021 7:06 PM

Index	Animal	Index	Dog	Cat	Sheep	Lion	Horse
0	Dog	0	1	0	0	0	0
1	Cat	1	0	1	0	0	0
2	Sheep	2	0	0	1	0	0
3	Horse	3	0	0	0	0	1
4	Lion	4	0	0	0	1	0

## 1.3 Outlier Detection

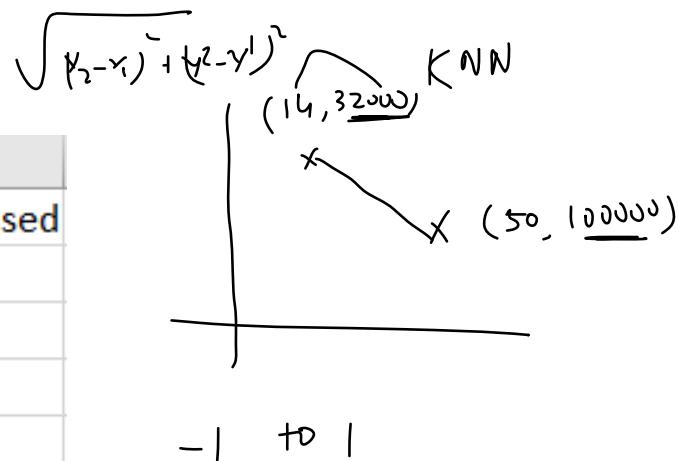
Thursday, April 8, 2021 7:07 PM



## 1.4 Feature Scaling

Thursday, April 8, 2021 7:08 PM

A	B		C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	No
3	Spain	27	48000	Yes
4	Germany	30	54000	No
5	Spain	38	61000	No
6	Germany	40		Yes
7	France	35	58000	Yes
8	Spain		52000	No
9	France	48	79000	Yes
10	Germany	50	83000	No
11	France	37	67000	Yes

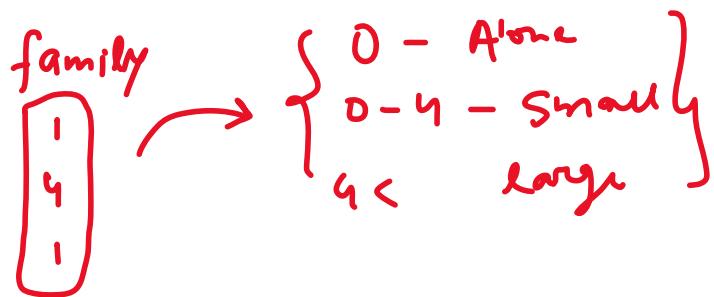


## 2. Feature Construction

Thursday, April 8, 2021 7:09 PM



P. Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Brigg	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmin	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C



### 3. Feature Selection ✓

Thursday, April 8, 2021 7:11 PM

→ MNIST

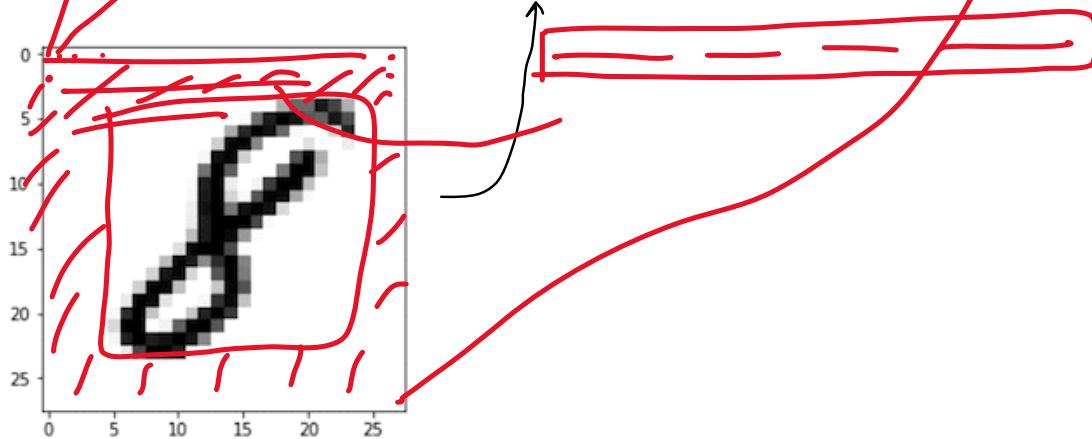
```
In [5]: train_df = pd.read_csv(f'{PATH}train.csv', header = 0)
```

```
In [19]: train_df
```

784 features

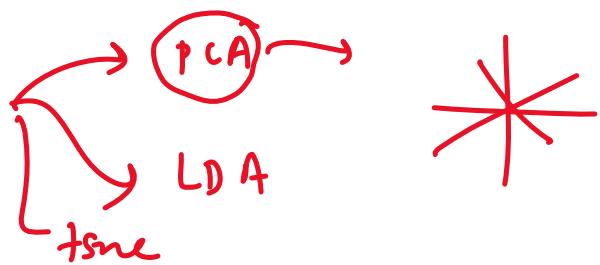
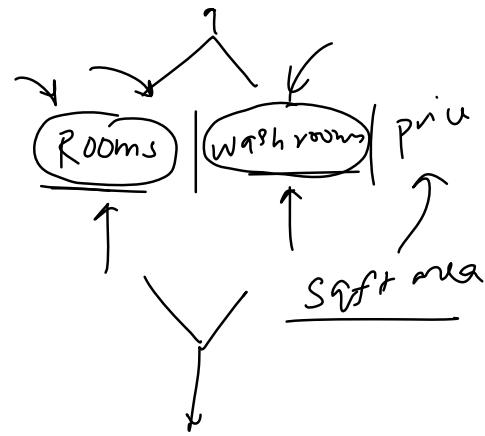
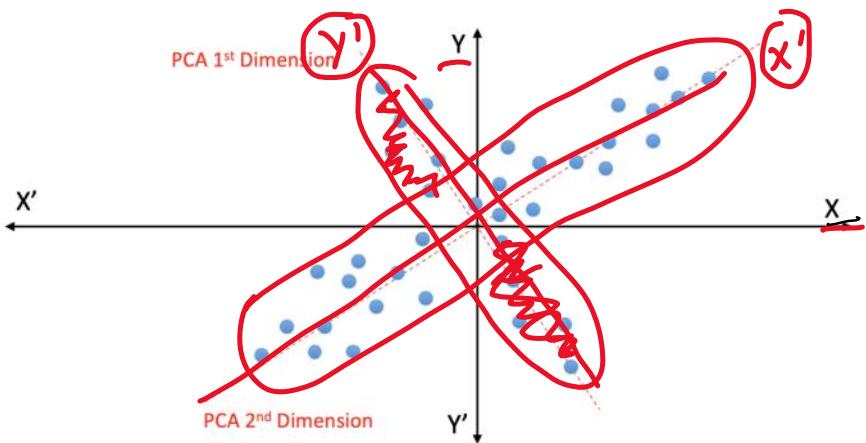
```
Out[19]:
```

	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel774
0	1	0	0	0	0	0	0	0	0	0	...	0
1	0	0	0	0	0	0	0	0	0	0	...	0
2	1	0	0	0	0	0	0	0	0	0	...	0
3	4	0	0	0	0	0	0	0	0	0	...	0



#### 4. Feature Extraction ✓

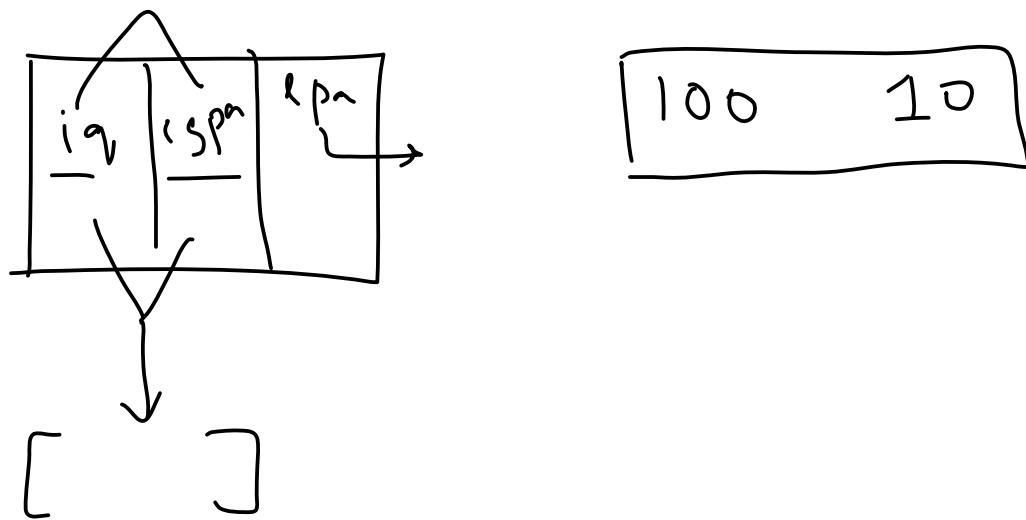
Thursday, April 8, 2021 7:13 PM



# What is Feature Scaling?

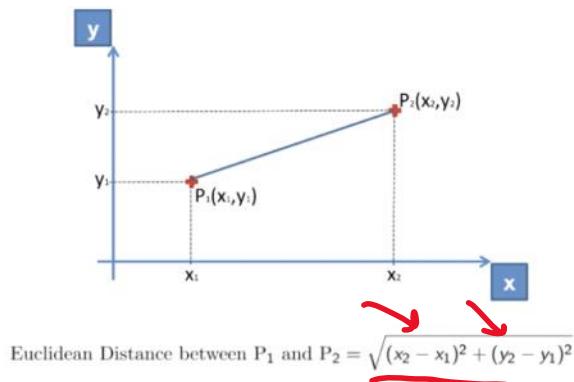
Friday, April 9, 2021 4:21 PM

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range



## Why do we need Feature Scaling?

Friday, April 9, 2021 4:27 PM



Let  $x$  be the no. of Salary and  $y$  be the no. of Age

Example:  $x_1 \& y_1$  are in row 2,  $x_2 \& y_2$  are in row 9

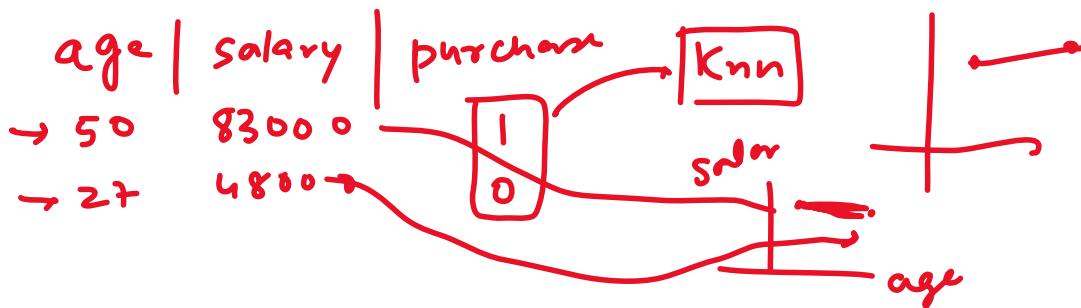
$$(x_2 - x_1)^2 = (83000 - 48000)^2$$

$$= 1225000000$$

$$(y_2 - y_1)^2 = (50 - 27)^2$$

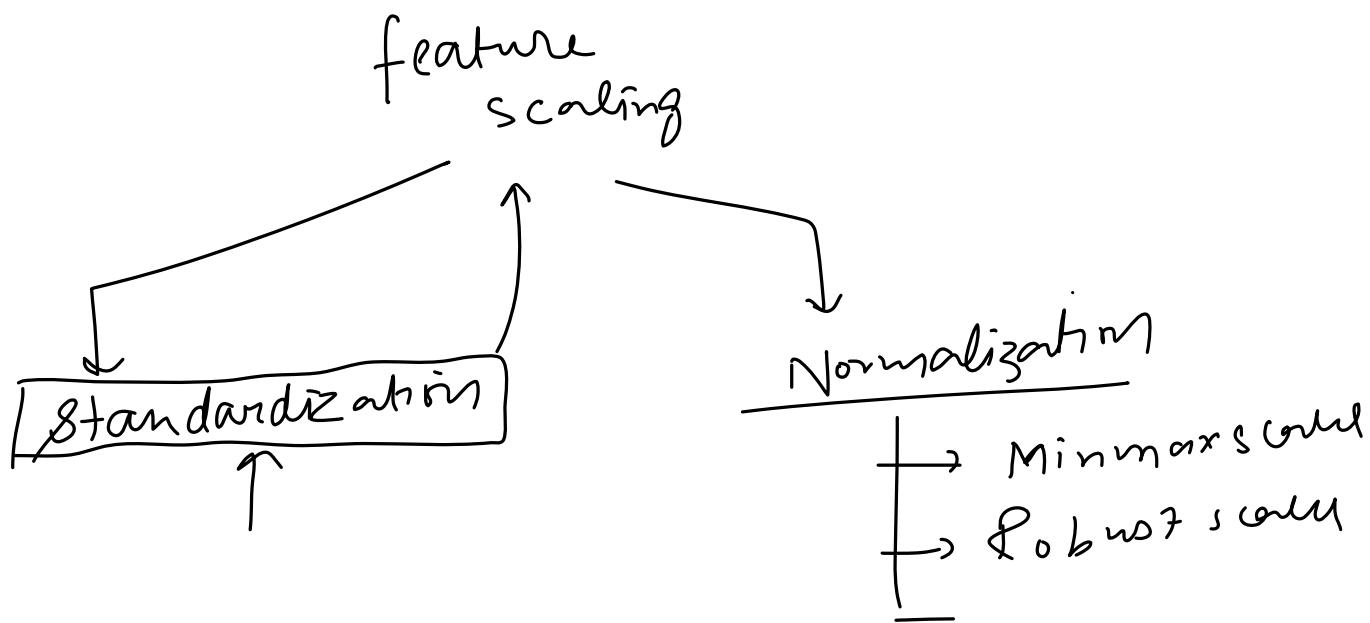
$$= 529$$

Knn



# Types of Feature Scaling

Friday, April 9, 2021 4:27 PM

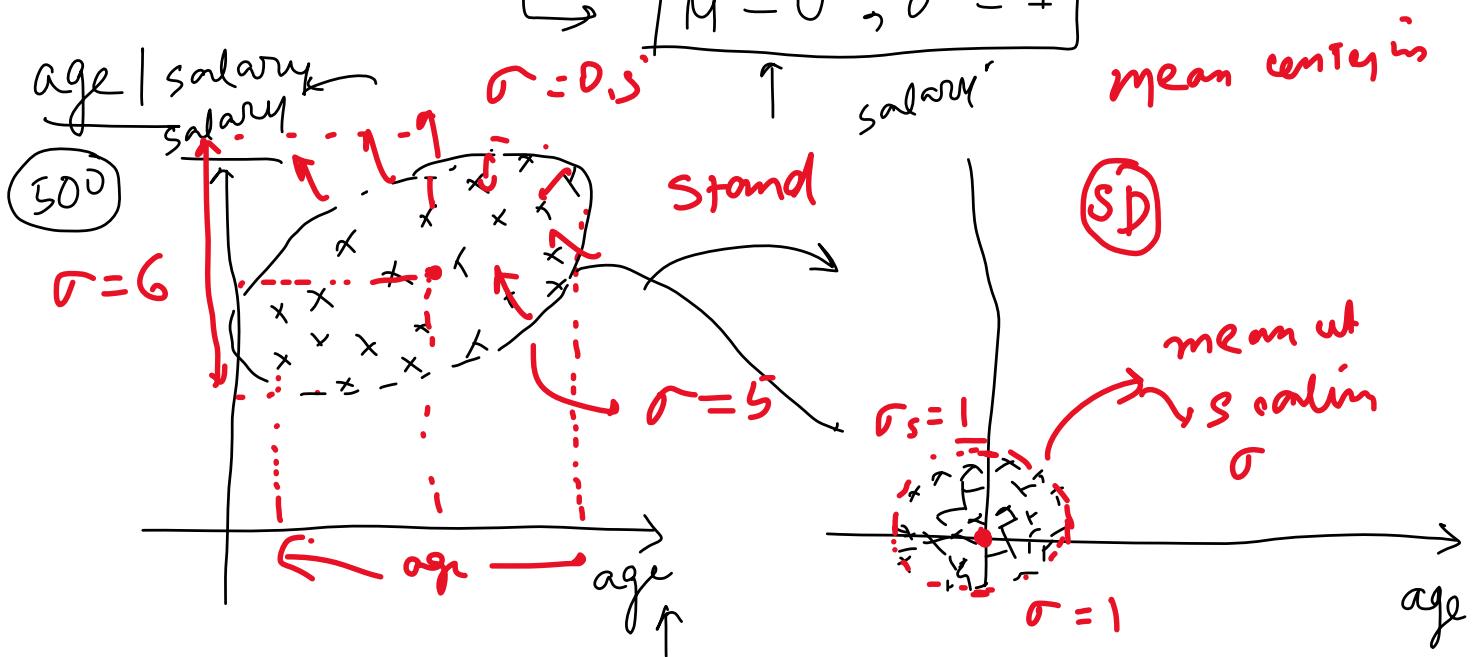
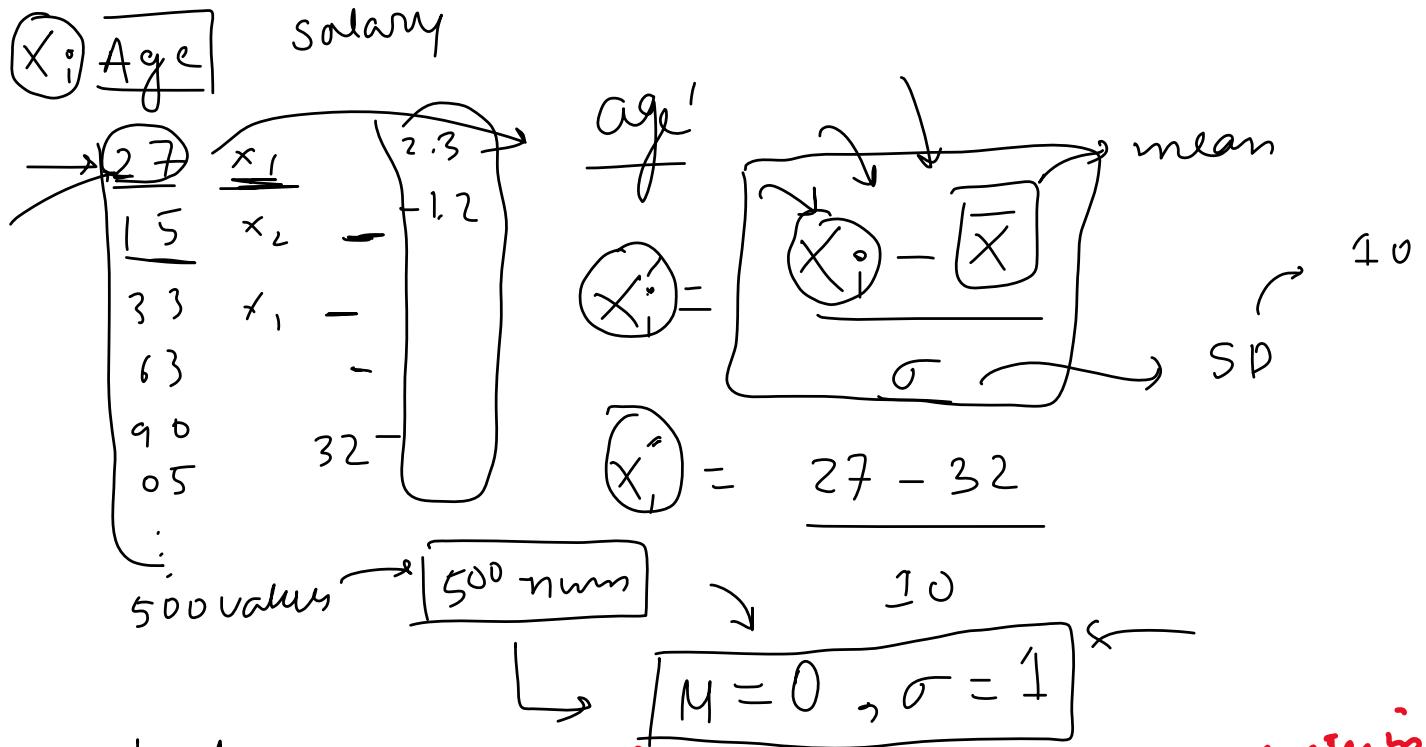


## Standardization - Intuition

Friday, April 9, 2021 4:27 PM

$$\frac{15 - 32}{10}$$

Also called as Z-score Normalization

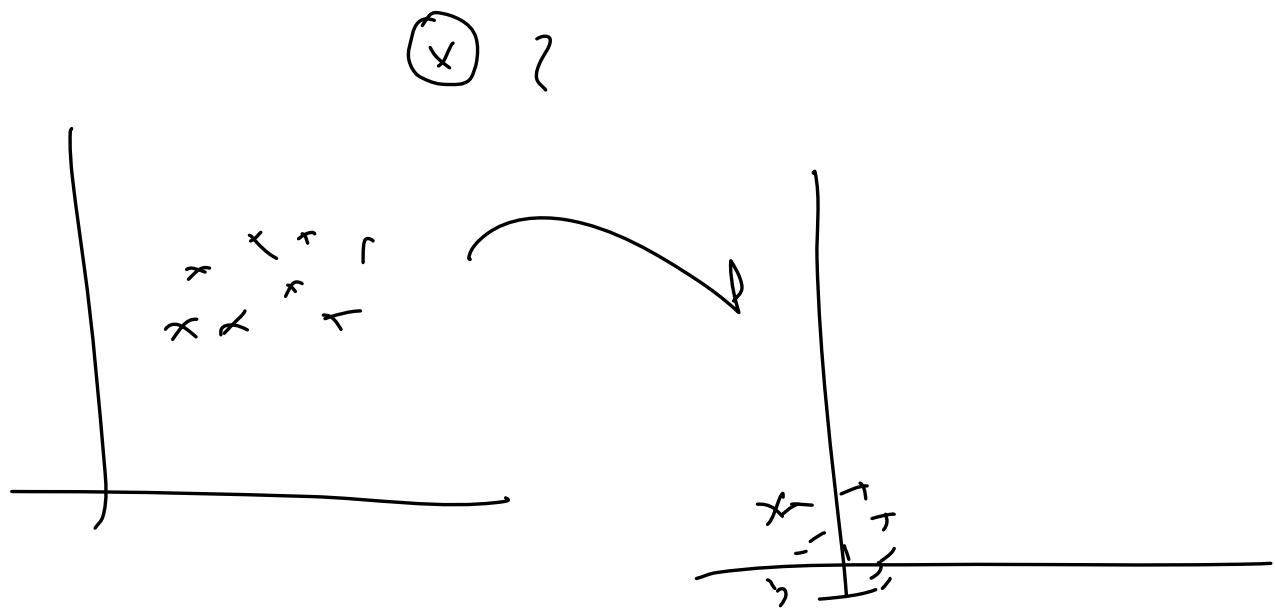


# Example

Friday, April 9, 2021 4:28 PM

# Impact of Outliers

Friday, April 9, 2021 4:28 PM



## When to use Standardization?

Friday, April 9, 2021 4:28 PM

$$a > b \quad 15 \cancel{>} 100$$

Algorithm(s)	Reason of applying feature scaling
1. K-Means	Use the Euclidean distance measure.
2. K-Nearest-Neighbours	Measure the distances between pairs of samples and these distances are influenced by the measurement units
3. Principal Component Analysis (PCA)	Try to get the feature with maximum variance
4. Artificial Neural Network	Apply Gradient Descent
5. Gradient Descent	Theta calculation becomes faster after feature scaling and the learning rate in the update equation of Stochastic Gradient Descent is the same for every parameter

Decision tree / Random Forest

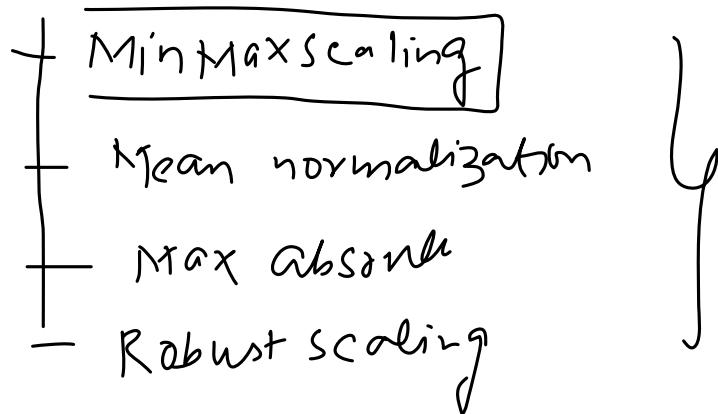
GB, XG BOOST

$$\#^{\frac{1}{2}}$$

# Normalization

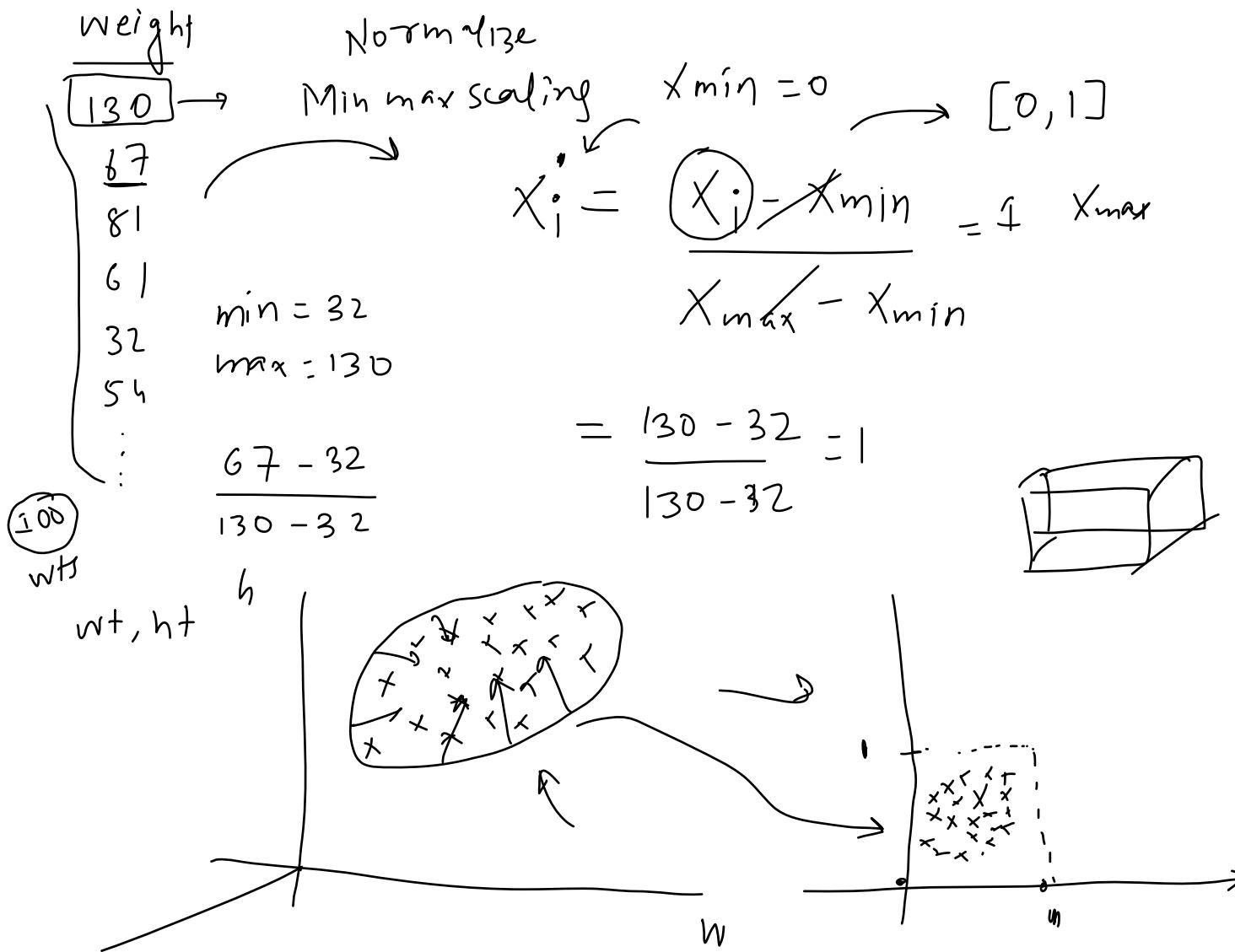
Saturday, April 10, 2021 1:15 PM

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information



## MinMaxScaling - Intuition

Saturday, April 10, 2021 1:16 PM



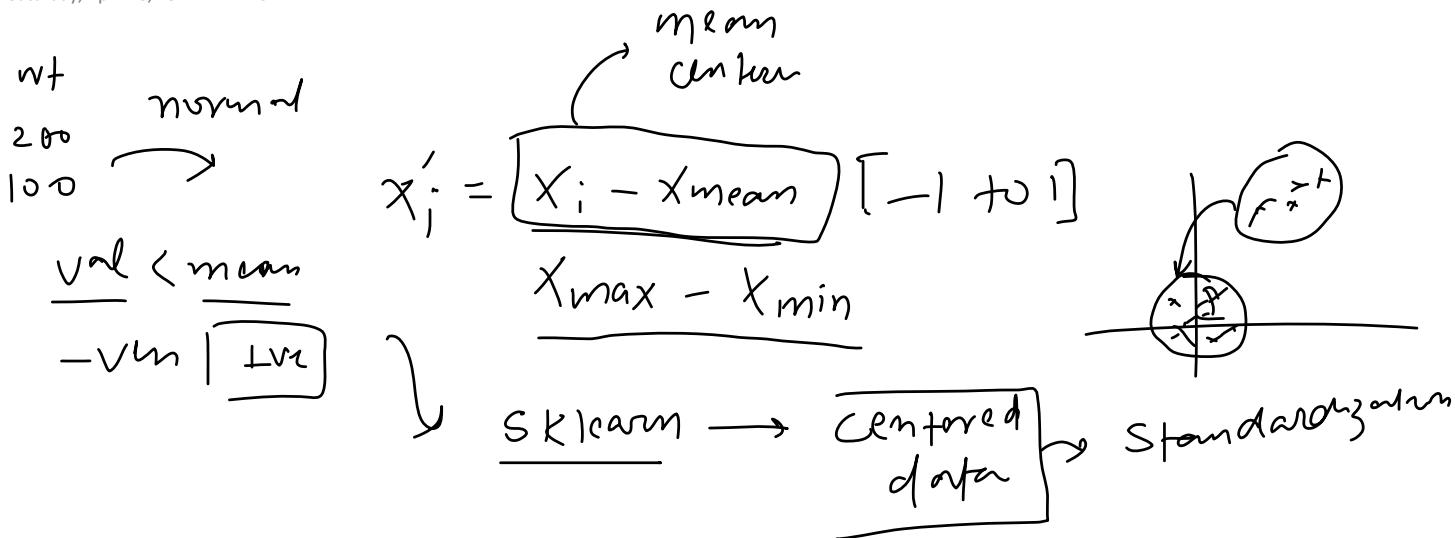
## Code Example

Saturday, April 10, 2021 1:17 PM

wine\_data →

## Mean Normalization

Saturday, April 10, 2021 1:16 PM



## MaxAbsScaling

Saturday, April 10, 2021 1:17 PM

The diagram illustrates the MaxAbsScaling formula and its application to sparse data. At the top left, there is a vertical list of values:  $Wt$ ,  $200$ ,  $100$ , and  $300$ . An arrow points from this list to a box containing the formula  $x'_i = \frac{x_i}{|x_{\max}|}$ . Inside the box,  $x_i$  is circled, and  $|x_{\max}|$  is underlined. From the bottom right corner of the box, an arrow points to the text "MaxAbs scaled". Below this, another arrow points to the text "sparse data" followed by a plus sign. A circle labeled "0's" is shown near the text "sparse data".

$$x'_i = \frac{x_i}{|x_{\max}|}$$

MaxAbs scaled

+ sparse data

0's

## Robust Scaling

Saturday, April 10, 2021 1:17 PM

wt

200

300

100

Robust scale

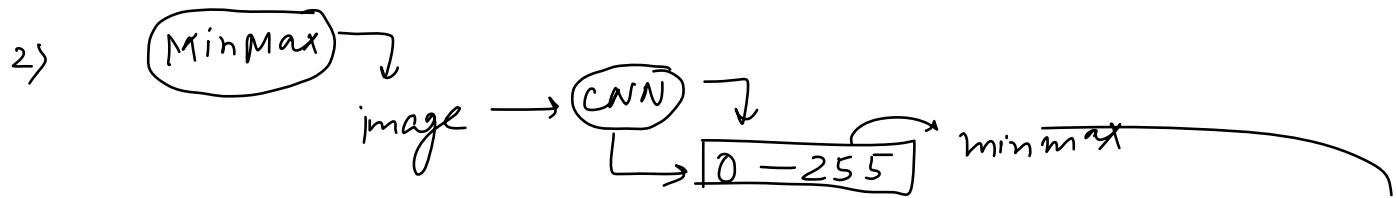
$$x_i' = \frac{x_i - \text{median}}{\text{IQR} \{ 75^{\text{mpg}} - 25^{\text{mpg}} \}}$$

→ Robust to Outliers

## Normalization Vs Standardization

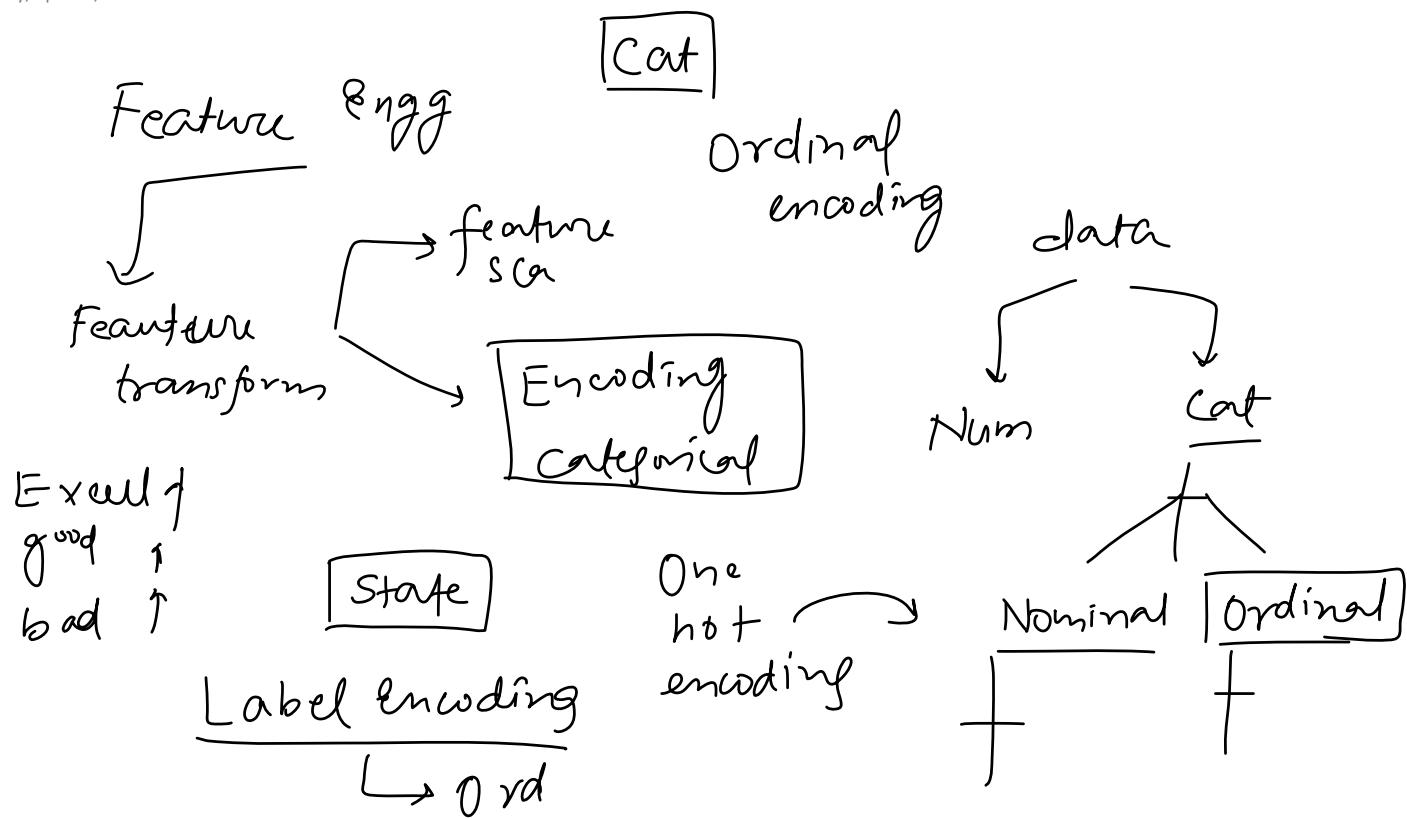
Saturday, April 10, 2021 1:17 PM

1) Is feature scaling required?



## Encoding Categorical Variables

Monday, April 12, 2021 3:53 PM



## Ordinal Encoding

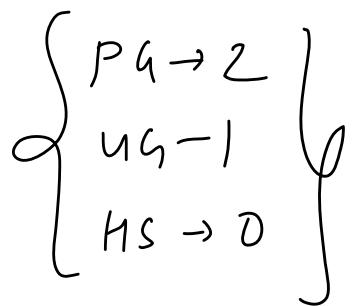
Monday, April 12, 2021 3:54 PM

Nominal OPE

<u>Education</u>	
HS	0
UG	1
PG	2
PG	2
UG	1
HS	0
UG	2

ordinal  
encoding

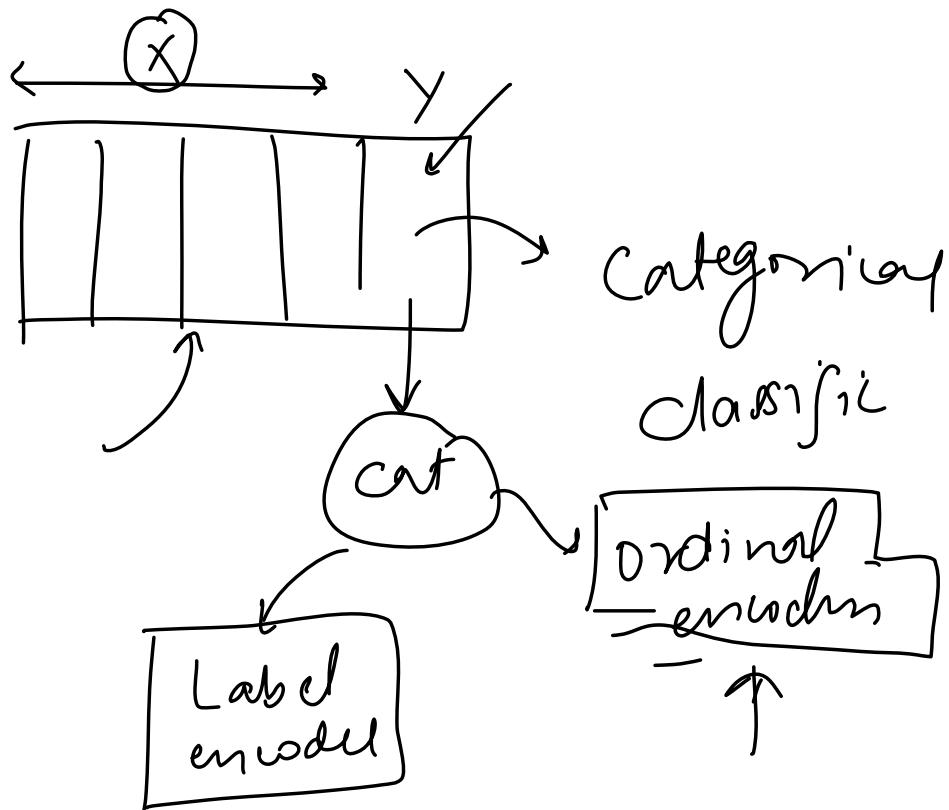
PG > UG > HS



# Label Encoding

Monday, April 12, 2021 3:54 PM

Ordinal  
encoder

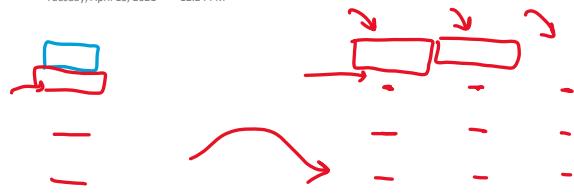


# Example

Monday, April 12, 2021 3:54 PM

## OneHotEncoding

Tuesday, April 13, 2021 12:24 PM



$[Y, B, R]$   
0 1 2

$[1, 0, 0] \rightarrow \text{Yellow}$   
 $[0, 1, 0] \rightarrow \text{Blue}$   
 $[0, 0, 1] \rightarrow \text{Red}$

50 diff

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color_Y	Color_B	Color_R	Target
1	0	0	0
1	0	0	1
0	1	0	1
1	0	0	1
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	0
1	0	0	1
0	1	0	0

One Hot Encoding



## Dummy Variable Trap

Tuesday, April 13, 2021 12:26 PM

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

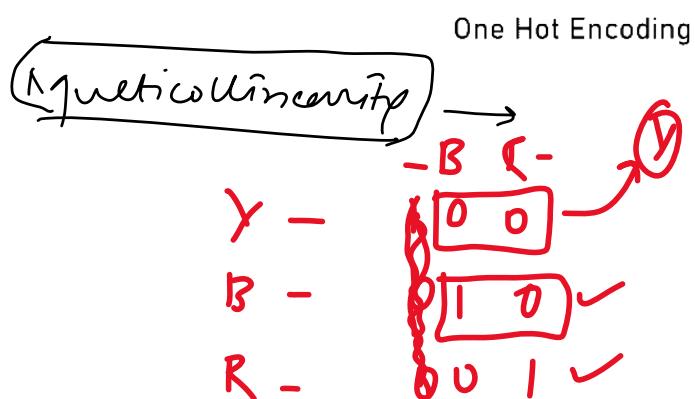
Multicat

(n-1) cols

Color	Target	Color_Y	Color_B	Color_R	Target
Yellow	0	1	0	0	0
Yellow	1	1	0	0	1
Blue	1	0	1	0	1
Yellow	1	1	0	0	1
Red	1	0	0	1	1
Yellow	0	1	0	0	0
Red	1	0	0	1	1
Red	0	0	0	1	0
Yellow	1	1	0	0	1
Blue	0	0	1	0	0

n cols

(n-1) cols



$\sum = 1$

linear model

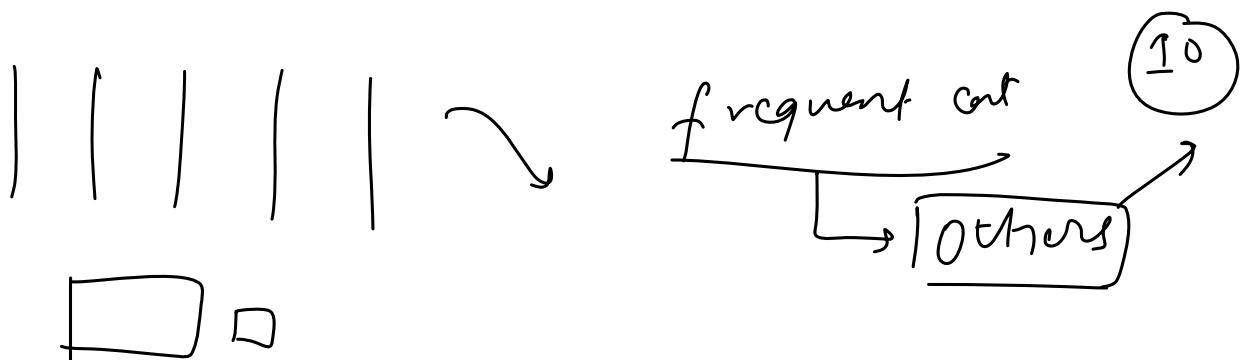
Linear Reg

Logistic

## OHE using most frequent variables

Tuesday, April 13, 2021 12:31 PM

brand → nominal → 40 brands  
cat → OHE



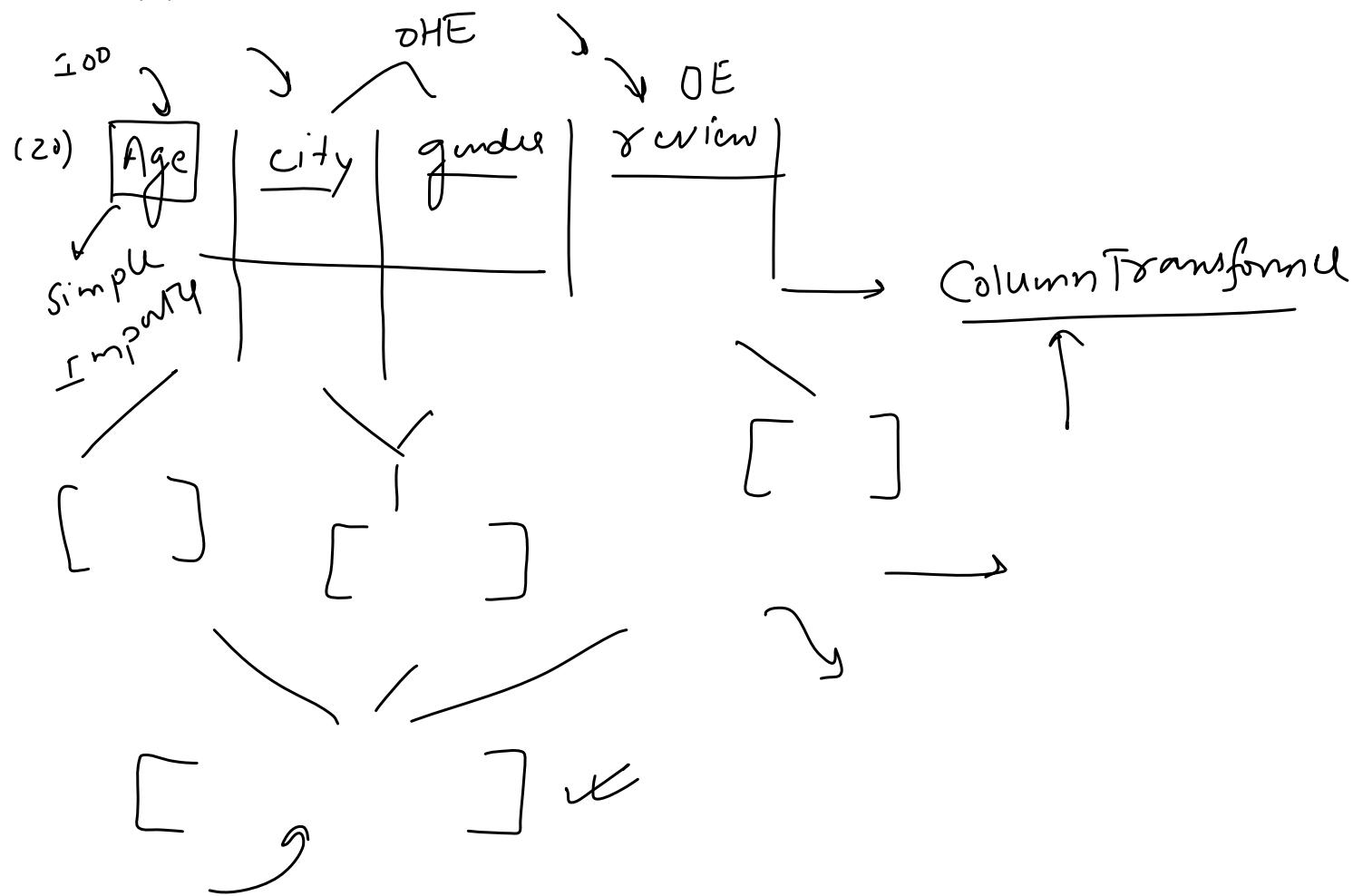
## Example

Tuesday, April 13, 2021 12:26 PM

	brand	km_driven	fuel	owner	selling_price
0	Maruti	145500	Diesel	First Owner	450000
1	Skoda	120000	Diesel	Second Owner	370000
2	Honda	140000	Petrol	Third Owner	158000
3	Hyundai	127000	Diesel	First Owner	225000
4	Maruti	120000	Petrol	First Owner	130000

## Column Transformer

Wednesday, April 14, 2021 5:16 PM



# Internship Project Discussion

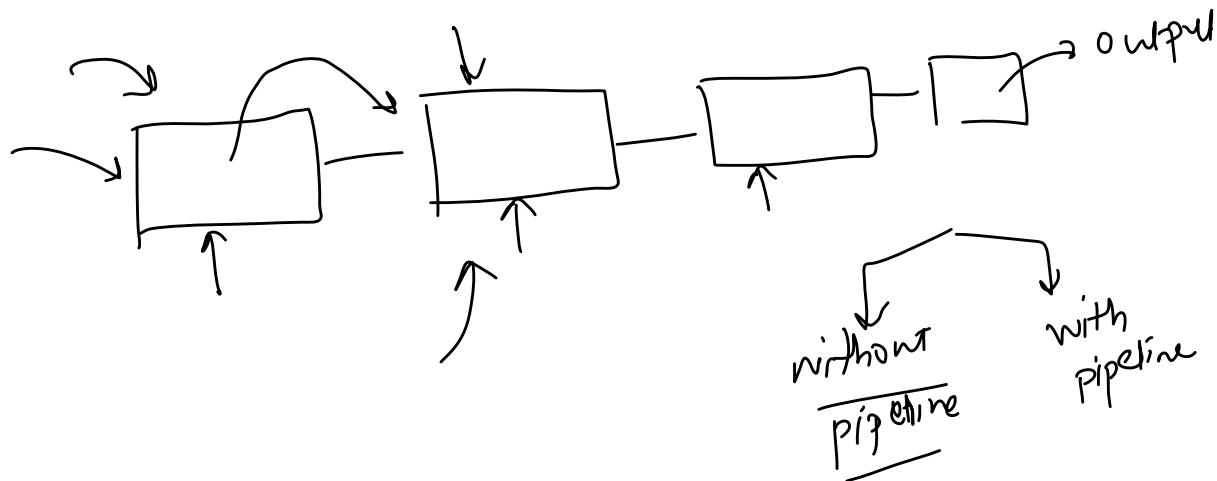
Wednesday, April 14, 2021 6:41 PM

## Scikit Learn Pipelines

Thursday, April 15, 2021 5:59 PM

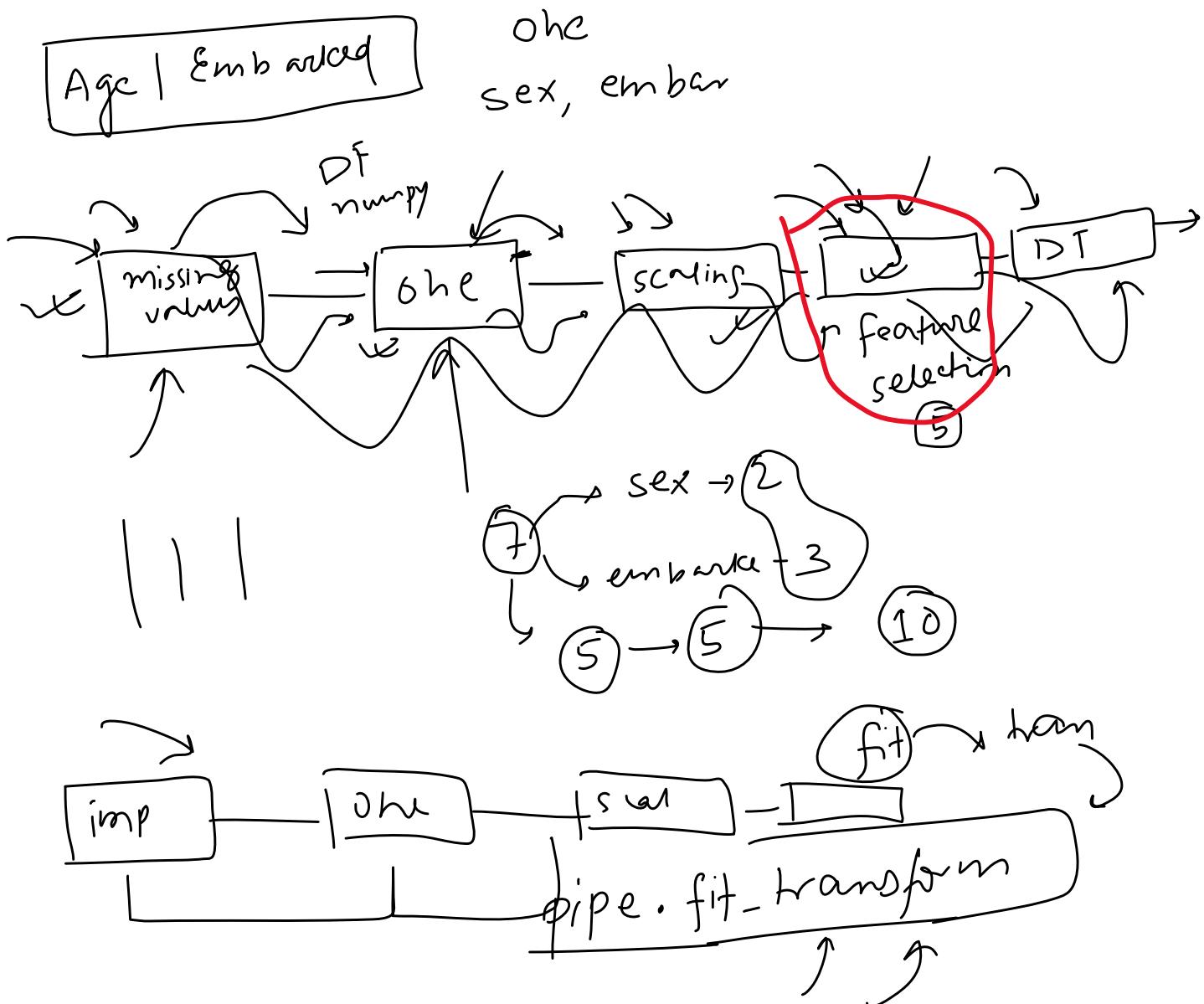
Pipelines chains together multiple steps so that the output of each step is used as input to the next step.

Pipelines makes it easy to apply the same preprocessing to train and test!



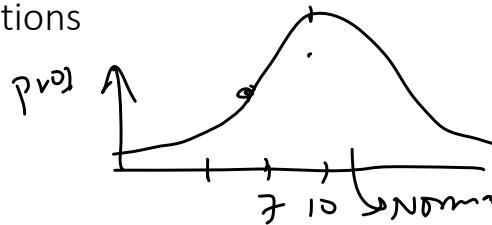
# Strategy

Thursday, April 15, 2021 6:58 PM



# Mathematical Transformations

Friday, April 16, 2021 5:21 PM



FE

feature  
transformations

mathematical  
trans

Normal  
distribution

1) log trans

2) Reciprocal

3) power ( $\sqrt{x}$  /  $x^{2/3}$ )

custom

1) Box-Cox

2) Yeo-Johnson

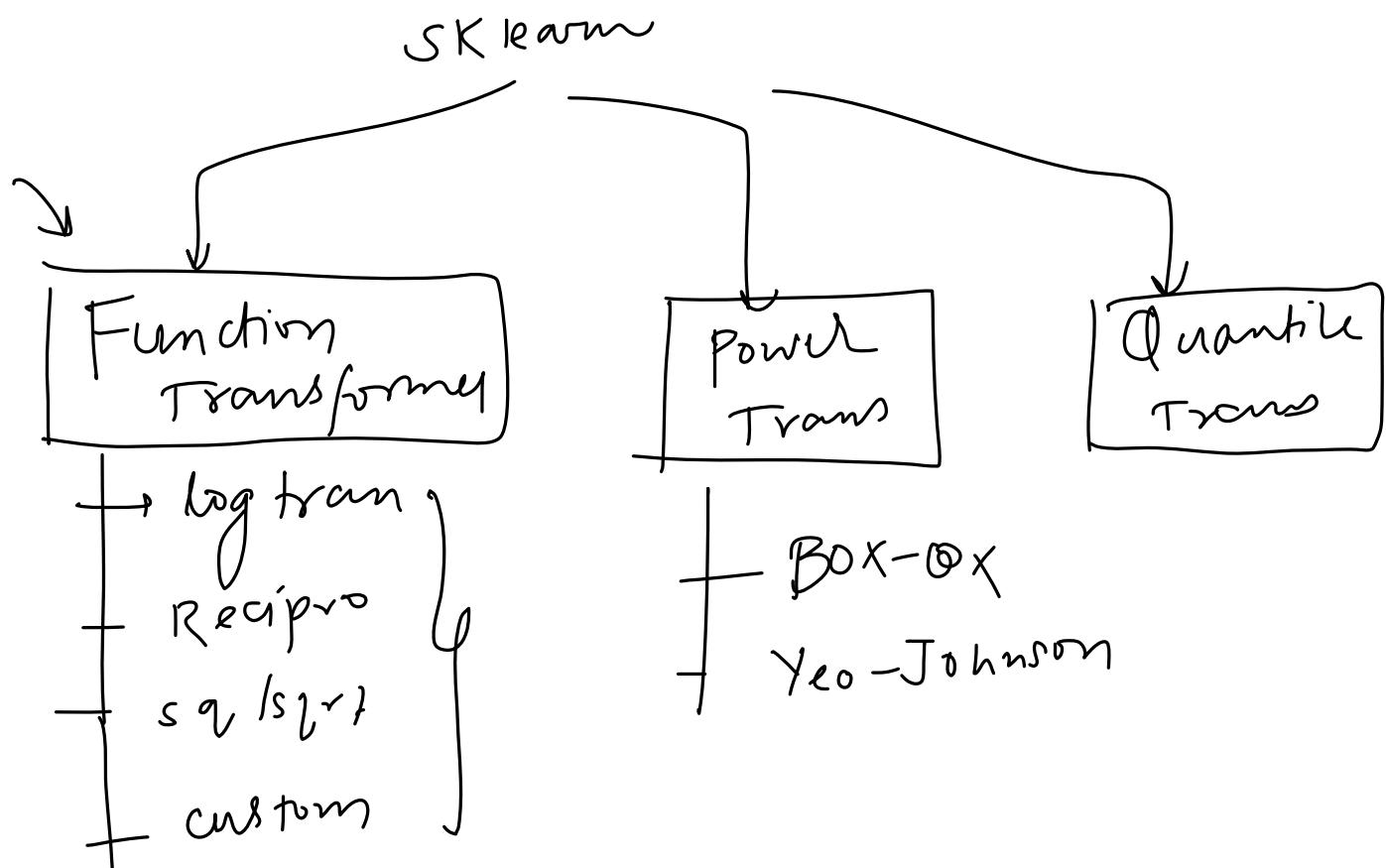
Statistic

(MD)

Linear Reg  
Logistic

# Function Transformer

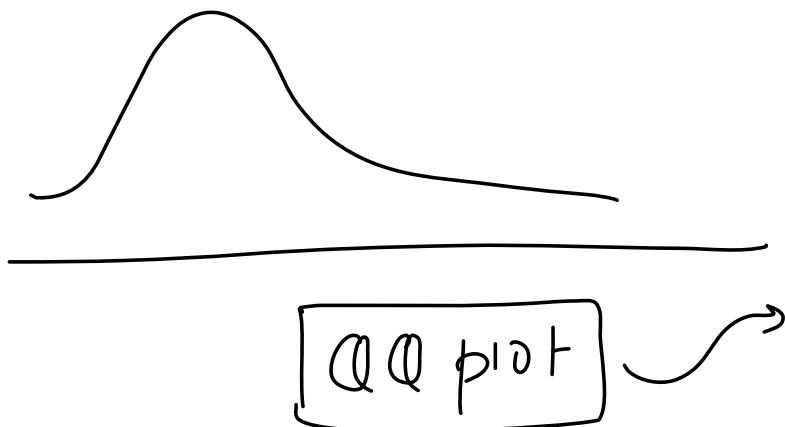
Friday, April 16, 2021 5:22 PM



## How to find if data is normal?

Friday, April 16, 2021 6:30 PM

`sns.distplot`



`pd.skew()` = 0

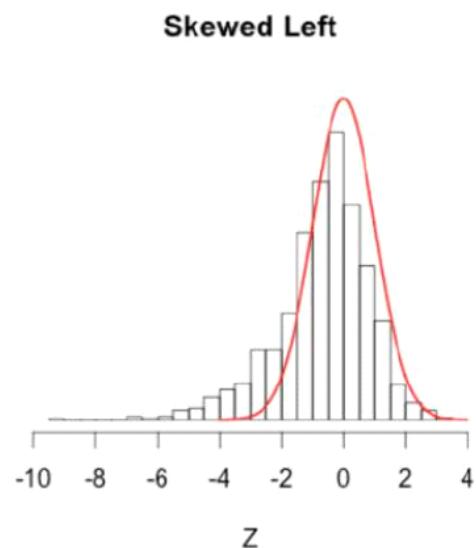
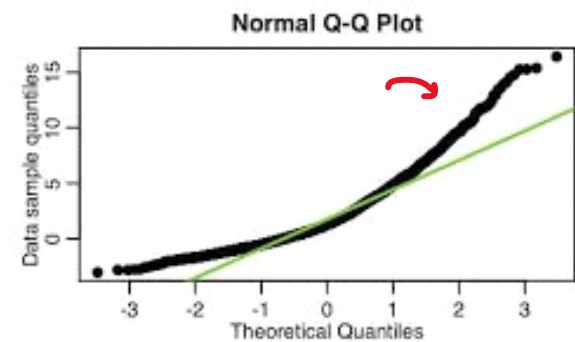
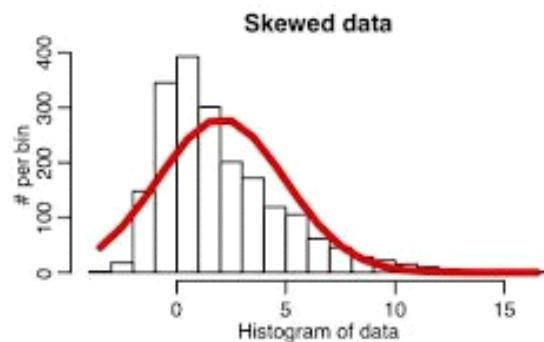
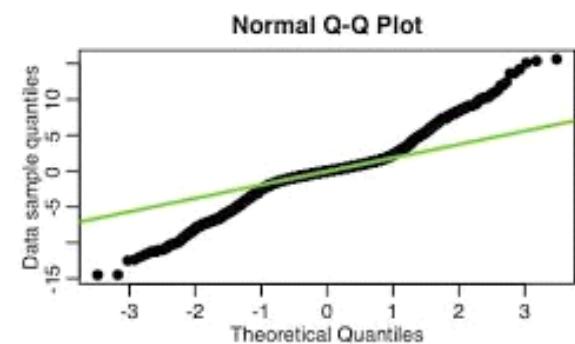
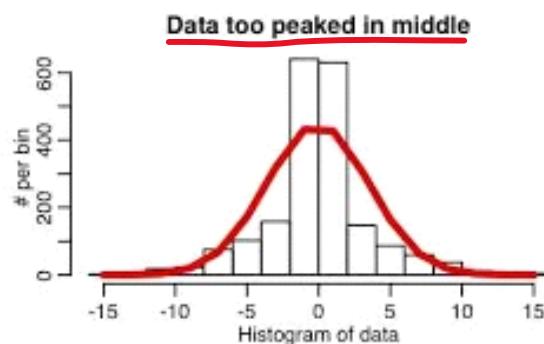
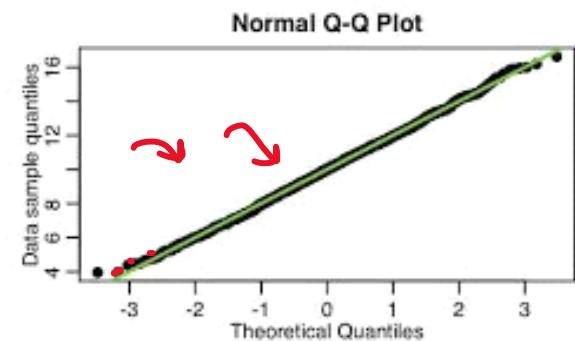
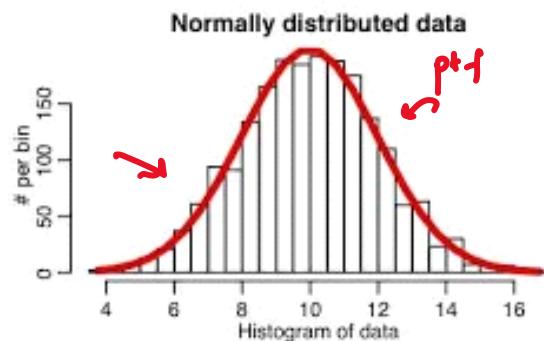
statistics

QQ plot

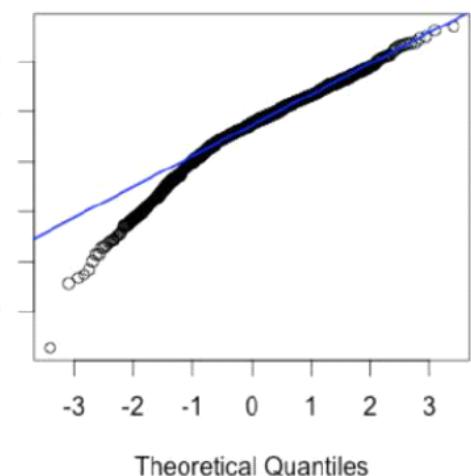


# QQ Plots

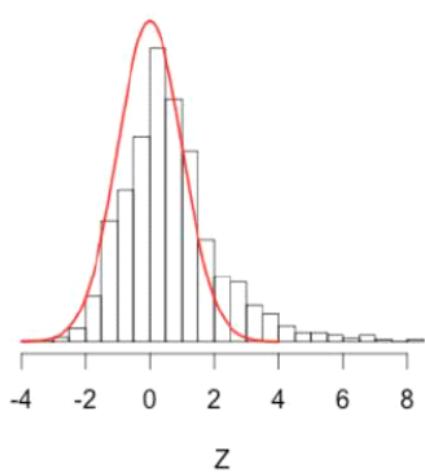
Friday, April 16, 2021 5:23 PM



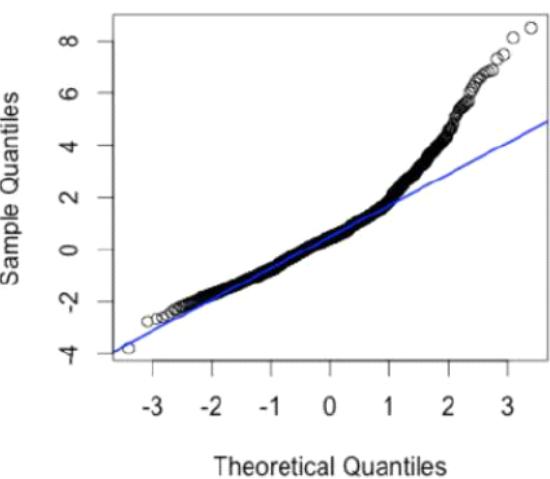
Sample Quantiles



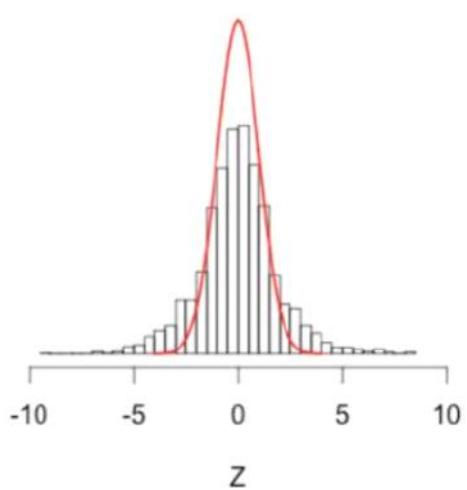
**Skewed Right**



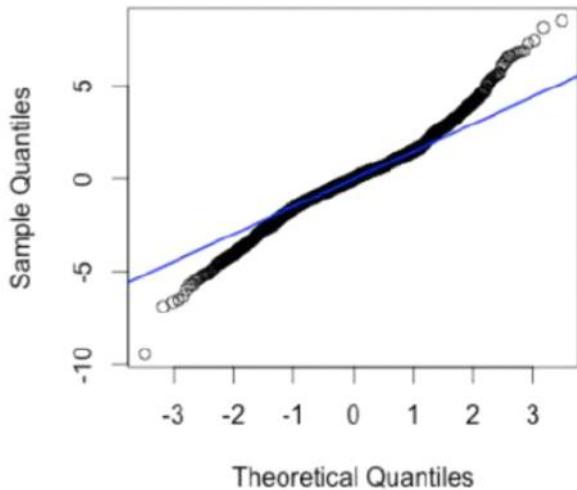
**Normal Q-Q Plot**



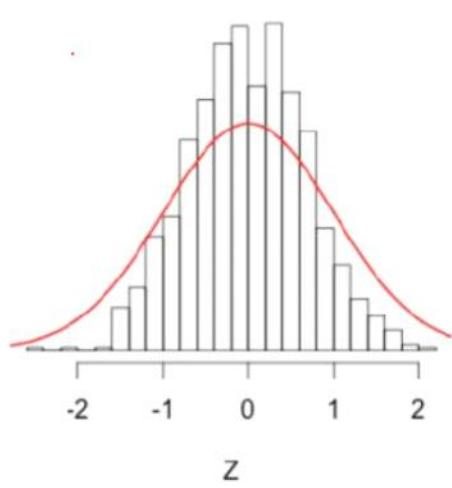
**Fat Tails**



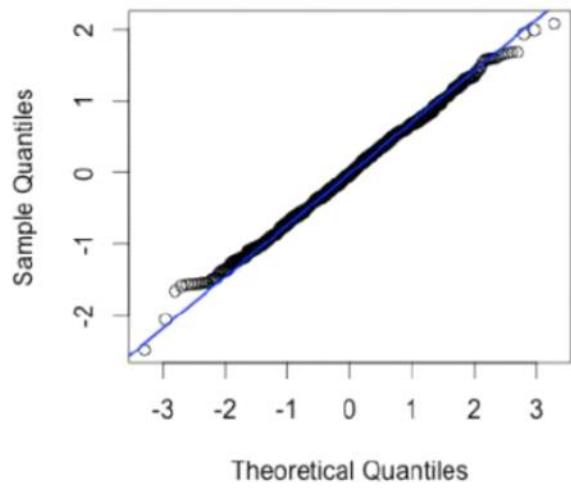
**Normal Q-Q Plot**



**Thin Tails**

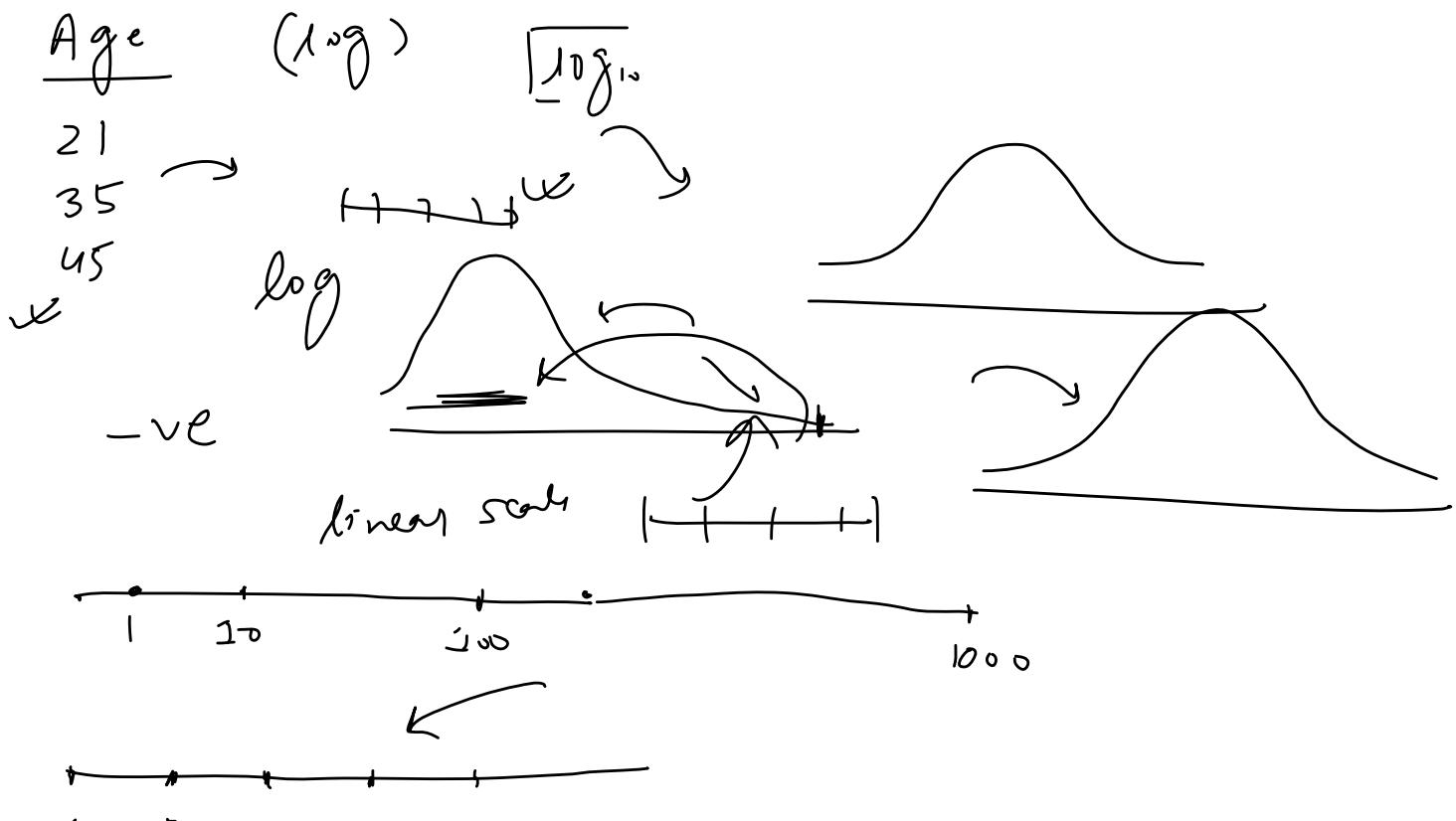


**Normal Q-Q Plot**



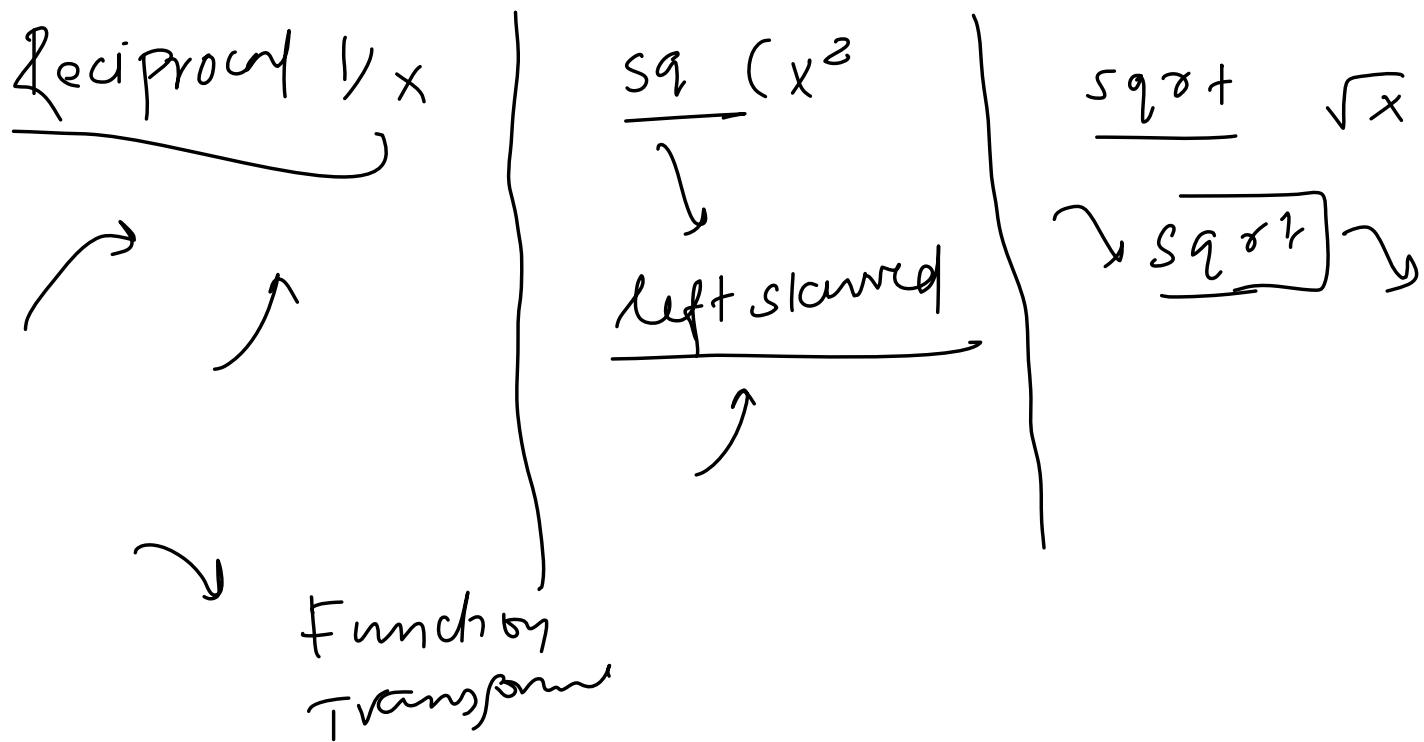
# Log Transform

Friday, April 16, 2021 5:22 PM



## Other Transforms

Friday, April 16, 2021 6:23 PM



## Example

Friday, April 16, 2021 5:22 PM

Titanic

age | fare | survived

$\eta_{P. \log}$

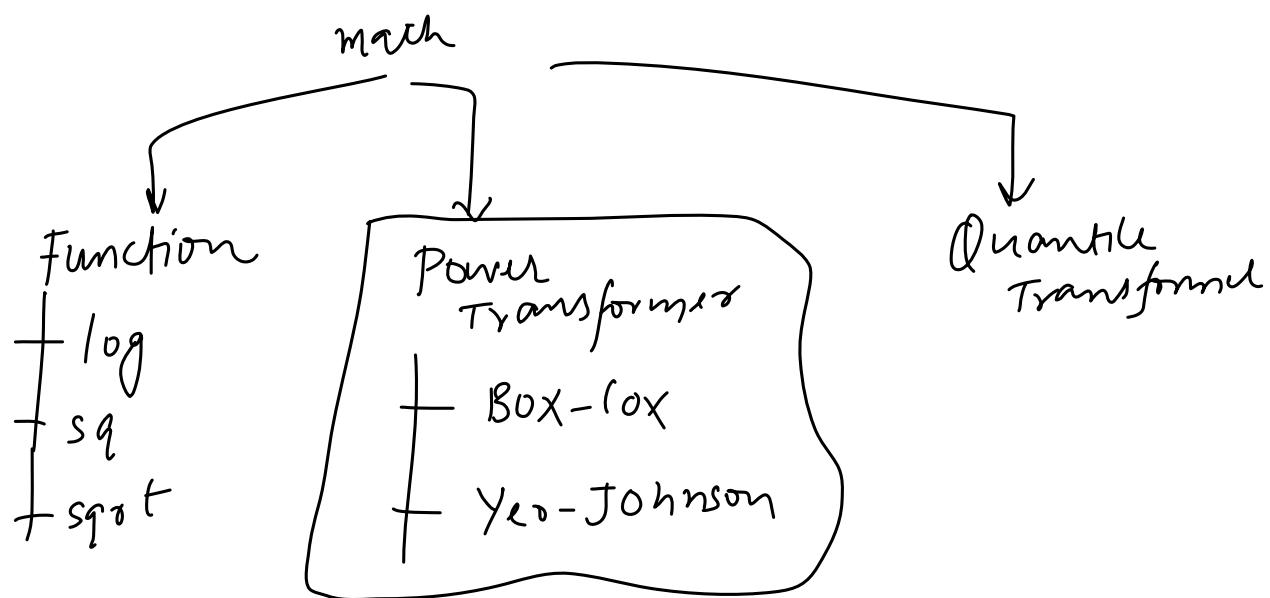
$\rightarrow 0$

$\eta_{P. \log^{-1} P}$

$(x^{-1})$

# Power Transformer

Saturday, April 17, 2021 4:55 PM



## Box Cox Transform

Saturday, April 17, 2021 4:56 PM

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x_i) & \text{if } \lambda = 0, \end{cases}$$

$\nabla > 0$  -ve

given dist.  
 ↴ Normal  
 dist

$x^2, x^3, x^{15}$

1.5, 11.

Max. likeliho<sup>od</sup>  
 ↳ Bayesian

The exponent here is a variable called lambda ( $\lambda$ ) that varies over the range of -5 to 5, and in the process of searching, we examine all values of  $\lambda$ . Finally, we choose the optimal value (resulting in the best approximation to a normal distribution) for your variable.

## Yeo - Johnson Transform

Saturday, April 17, 2021 4:56 PM

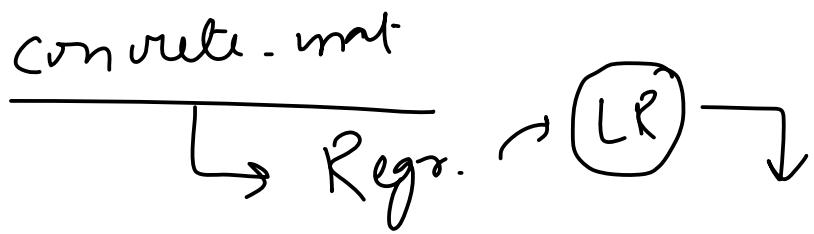
$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^\lambda - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i) + 1 & \text{if } \lambda = 0, x_i \geq 0 \\ -[(-x_i + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x_i < 0, \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases}$$

This transformation is somewhat of an adjustment to the Box-Cox transformation, by which we can apply it to negative numbers.

Power Transformer

## Example

Saturday, April 17, 2021 4:56 PM



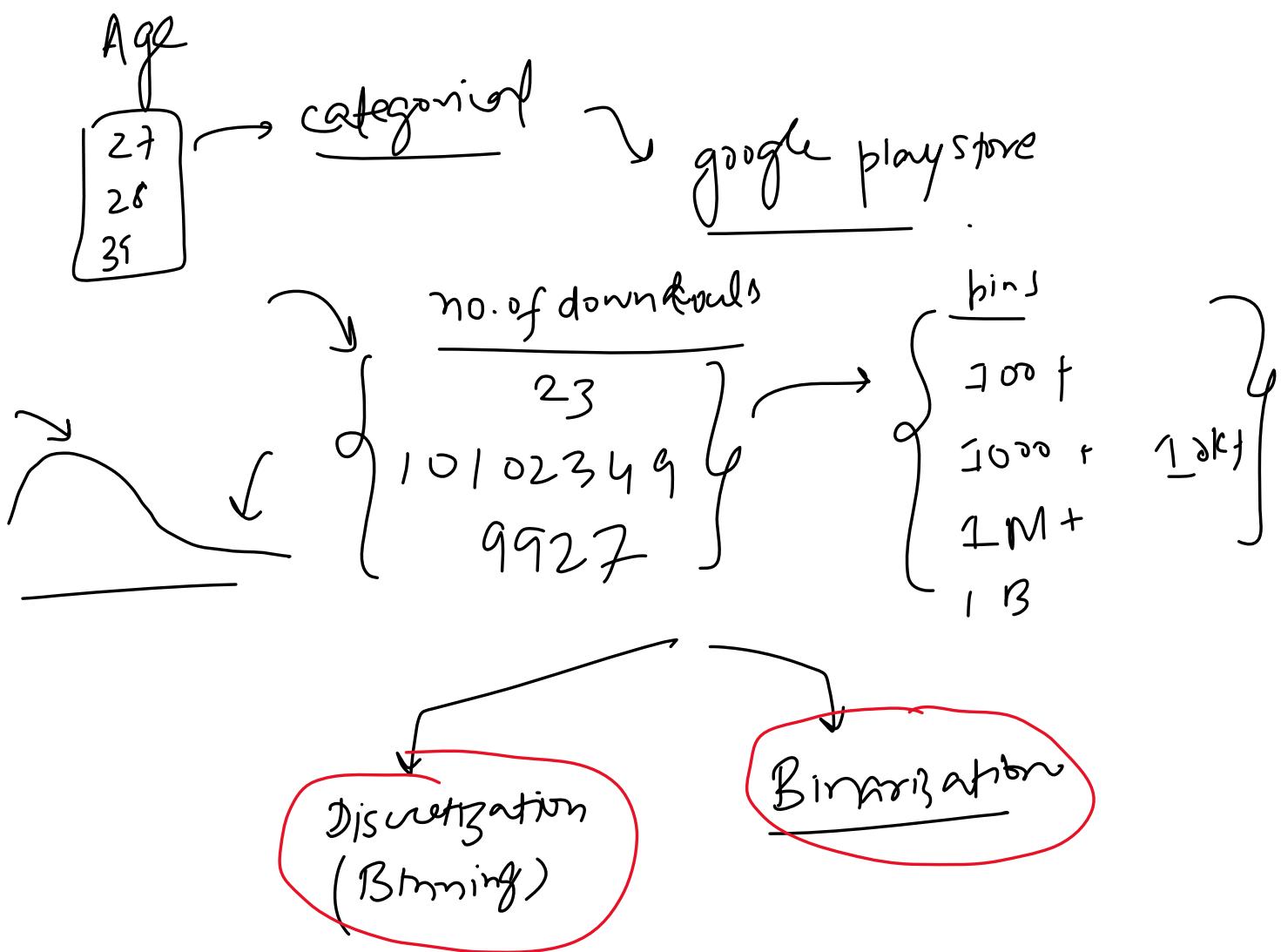
Cement

$$(540)^{0.177} \rightarrow$$

$$(30)^{(0.02)}$$

# 1. Encoding Numerical Features

Monday, April 19, 2021 4:18 PM



## 2. Discretization

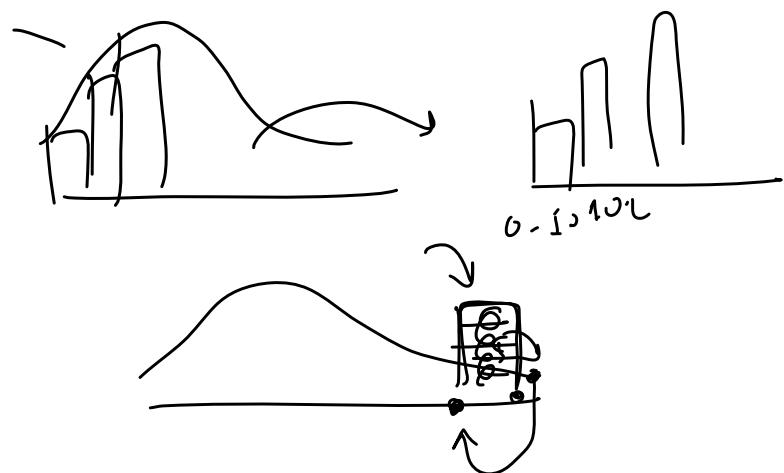
Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values. Discretization is also called binning, where bin is an alternative name for interval.

Why use Discretization:

1. To handle Outliers
2. To improve the value spread

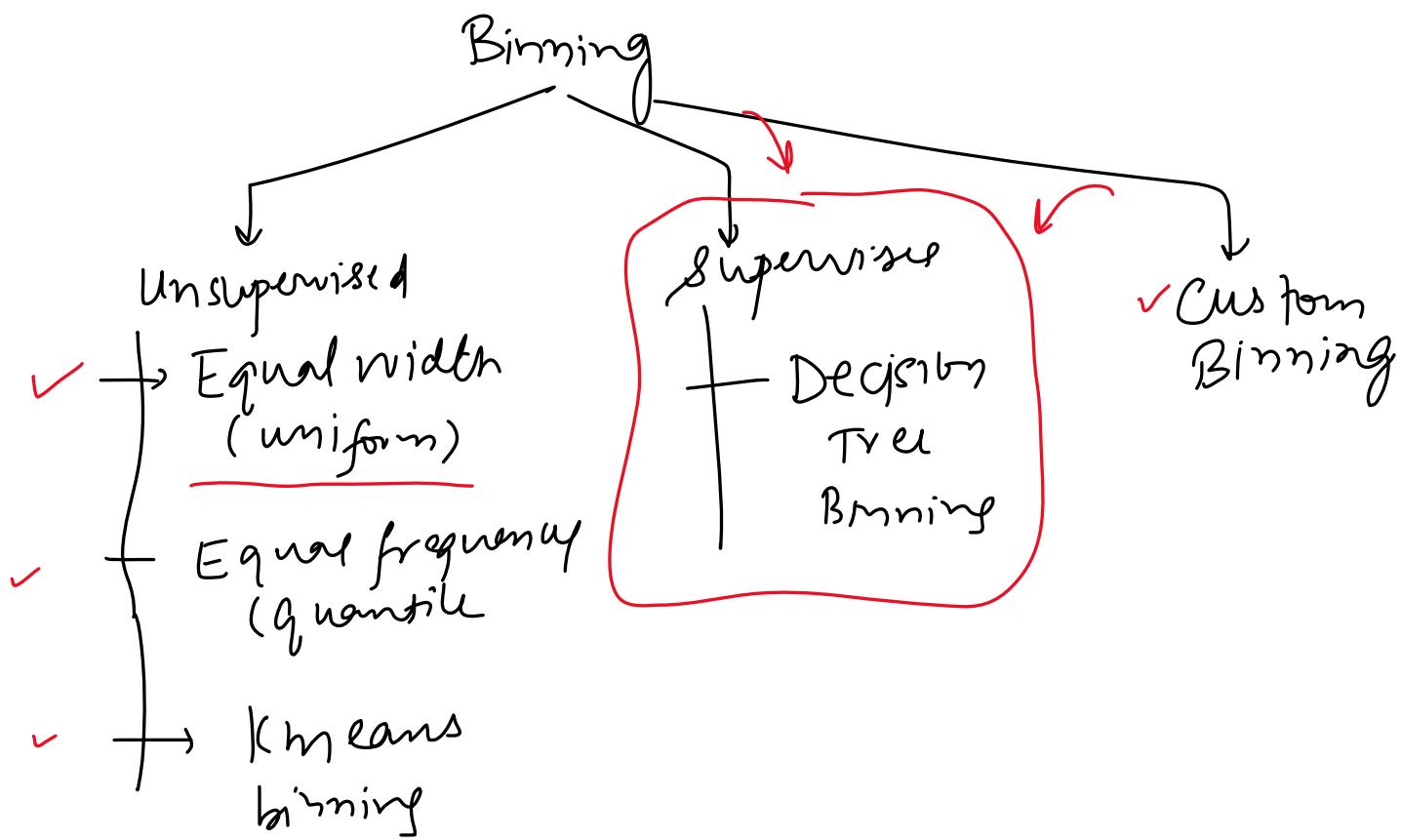
Age

23 42, 57 81. . . 100  
↓  
0-10, 10-20, 20-30 . . .  
5 6 10



### 3. Types of Discretization

Monday, April 19, 2021 3:56 PM



#### 4. Equal Width/Uniform Binning

Monday, April 19, 2021 3:56 PM

Age  
 23, 32, 84, 56, ...  
 Bins = 10

- 1) Outliers ↗
- 2) No change in spread

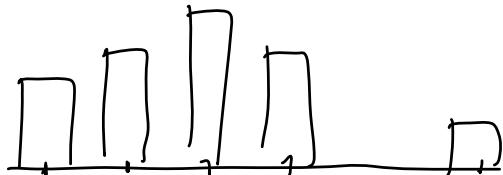
max 100

min 0

$$\frac{\text{max} - \text{min}}{\text{bins}} = \frac{100 - 0}{10} = 10$$

$$\frac{10}{5}, \frac{10}{16}, \frac{10}{17}, \dots, \frac{10}{5}$$

$\rightarrow$  10 bins

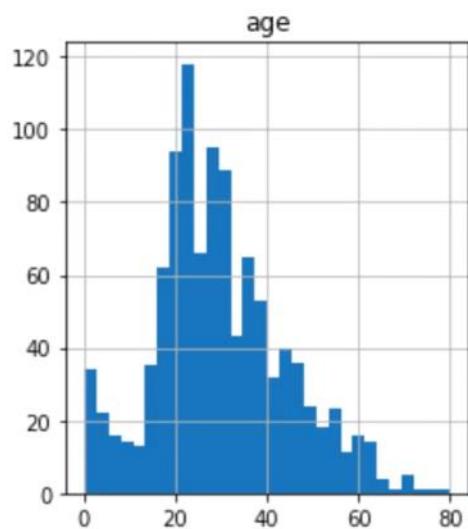


equal binning

	age	age_trf	age_labels
314	43.0	5.0	(40.21, 48.168]
523	44.0	5.0	(40.21, 48.168]
352	15.0	1.0	(8.378, 16.336]
534	30.0	3.0	(24.294, 32.252]
211	35.0	4.0	(32.252, 40.21]
530	2.0	0.0	(0.42, 8.378]
786	18.0	2.0	(16.336, 24.294]
827	1.0	0.0	(0.42, 8.378]
372	19.0	2.0	(16.336, 24.294]
518	36.0	4.0	(32.252, 40.21]

## 5. Equal Frequency/Quantile Binning

Monday, April 19, 2021 3:56 PM



- 1) Outliers
- 2) Value spread

0 → ?  
10<sup>th</sup> percent

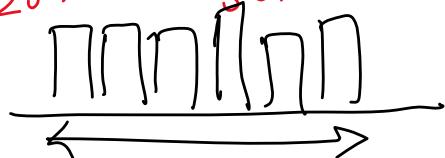
Intervals = 10

Each interval contains 10% of total observations

Intervals:

0-16; 16-20; 20-22; 22-25; ...  
50-74

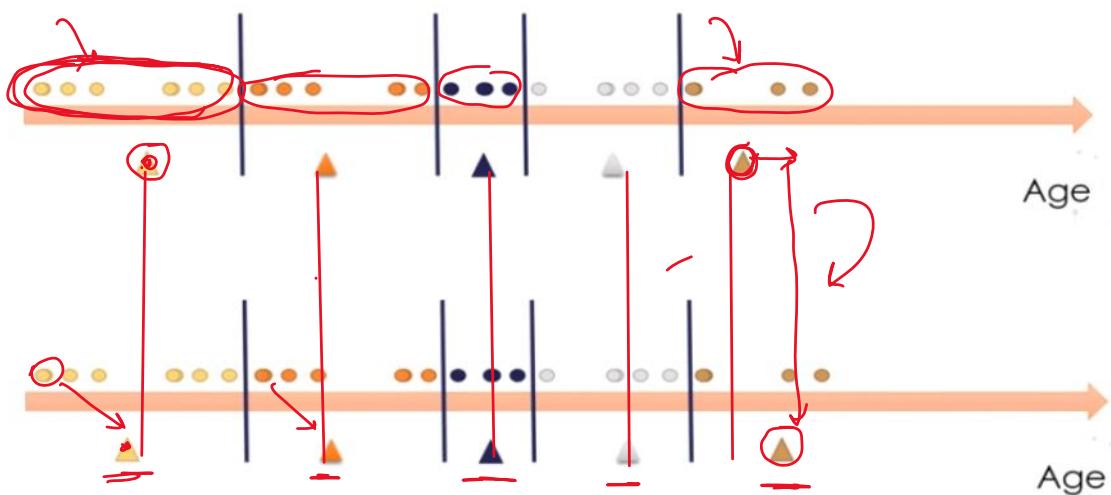
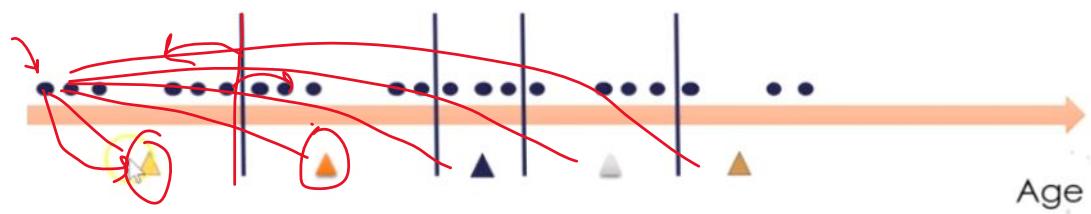
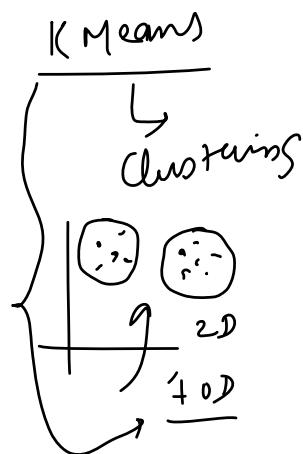
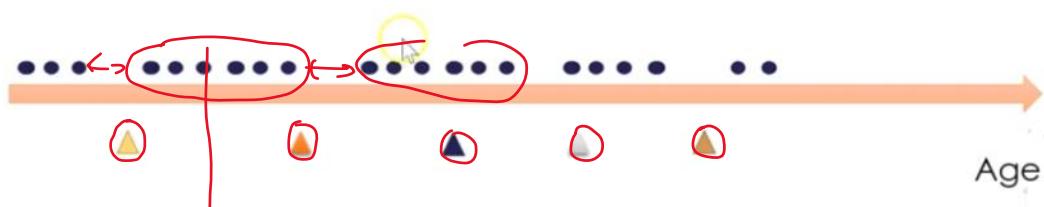
0-16      16-20      20-22  
10%      20%      30%



## 6. KMeans Binning

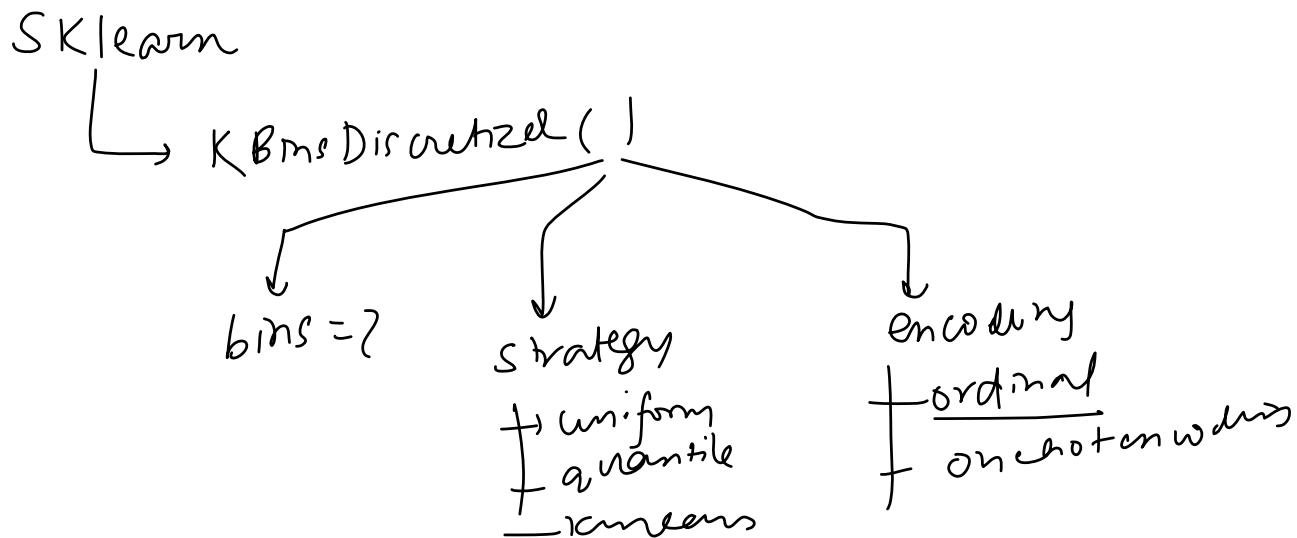
Monday, April 19, 2021 3:57 PM

centroid  
interval - 5



## 7. Encoding the discretized variable

Monday, April 19, 2021 3:58 PM



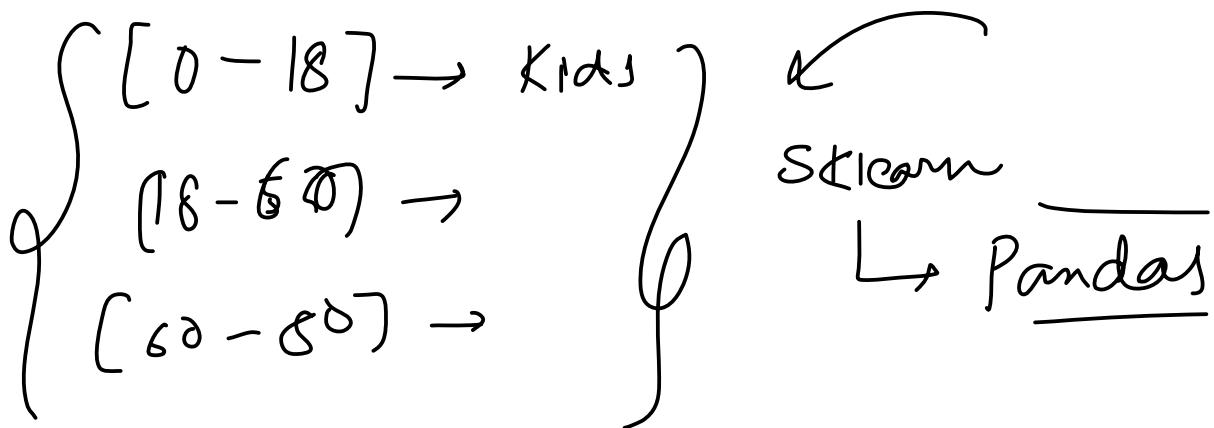
## 8. Example

Monday, April 19, 2021 3:58 PM

Titanic  
  └

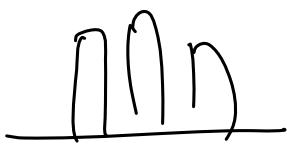
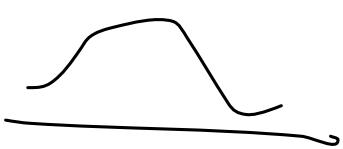
## 9. Custom/Domain Based Binning

Monday, April 19, 2021 3:58 PM

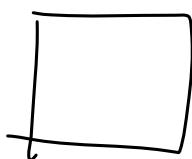


## 10. Binarization

Monday, April 19, 2021 4:17 PM



image



0, 1

0-255

color

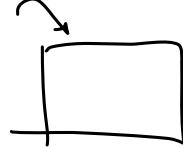
?

Annual income

61 <

>

127.5  
0      1



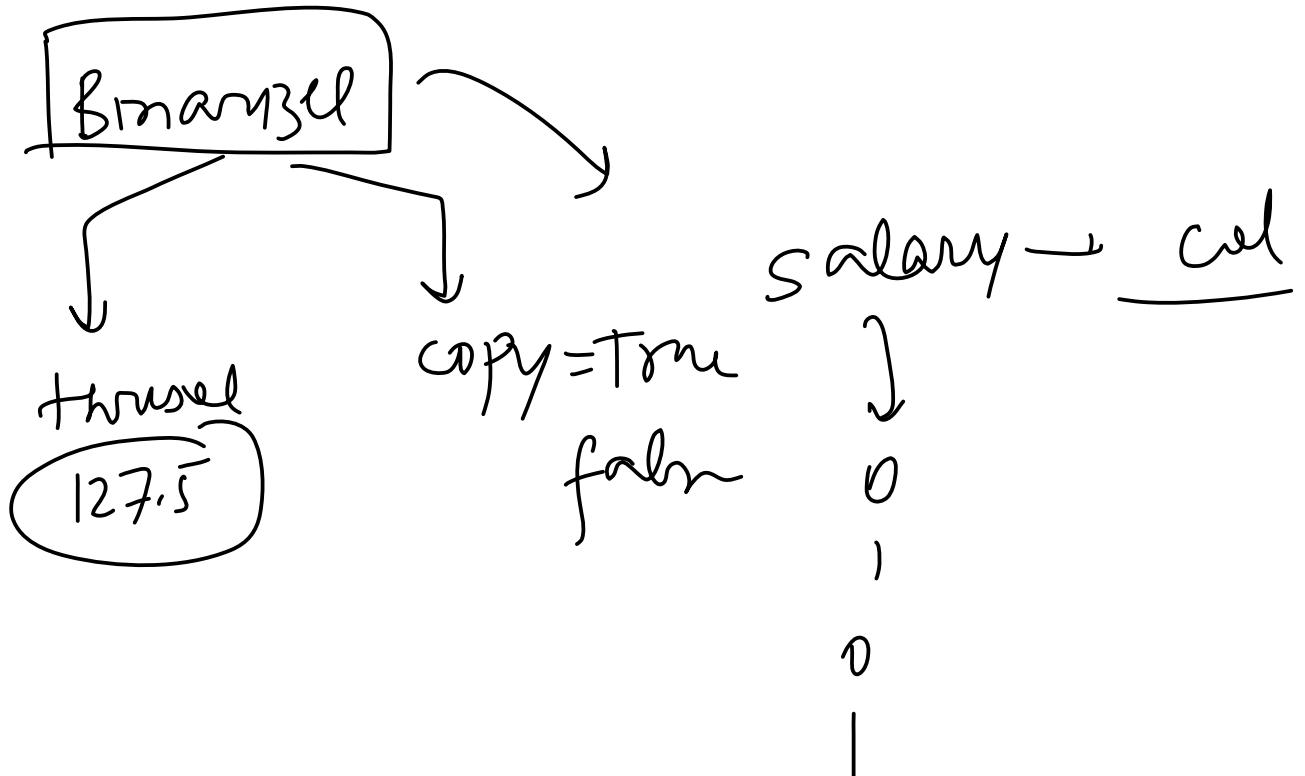
→



## 11. Example

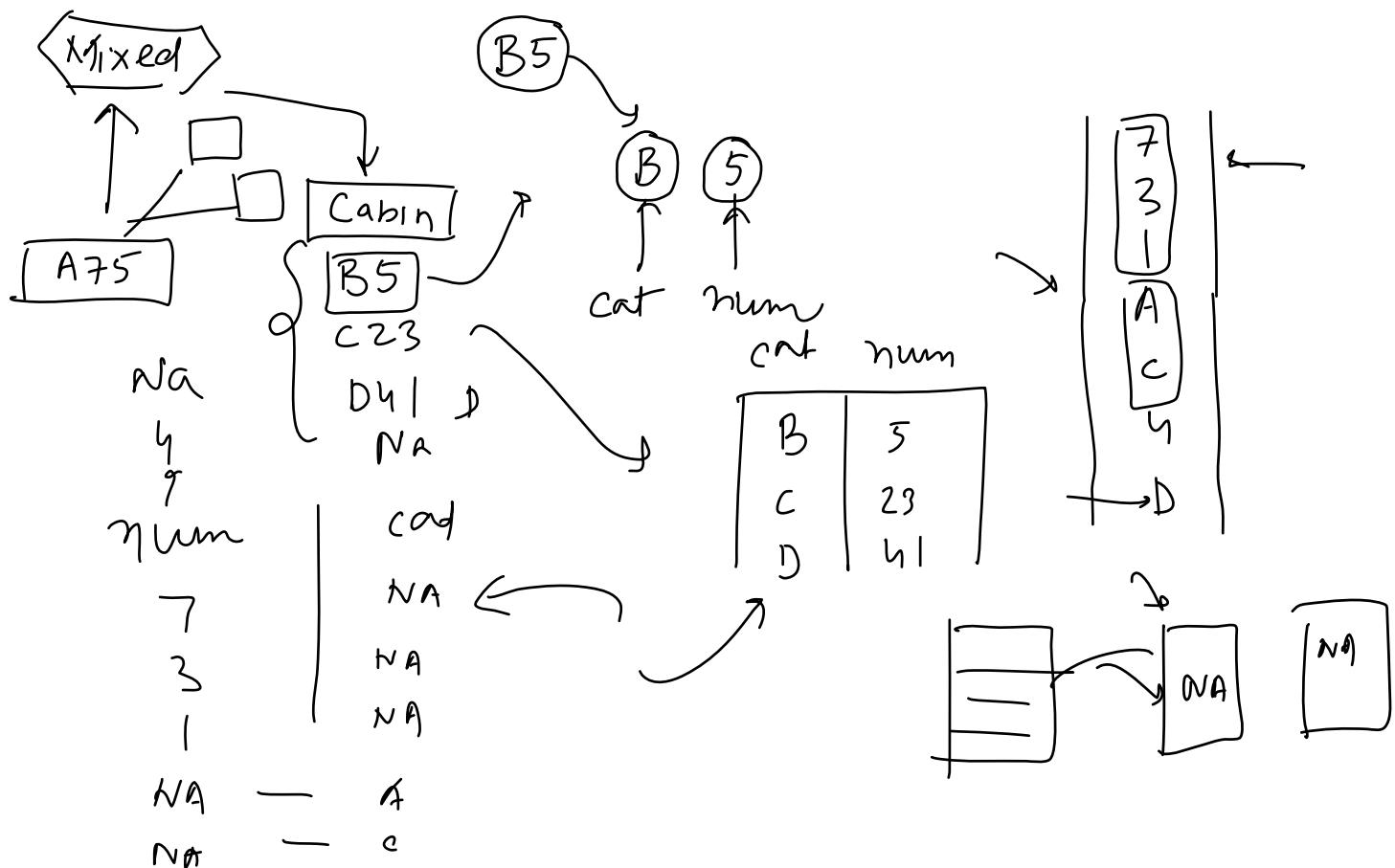
Monday, April 19, 2021

4:17 PM



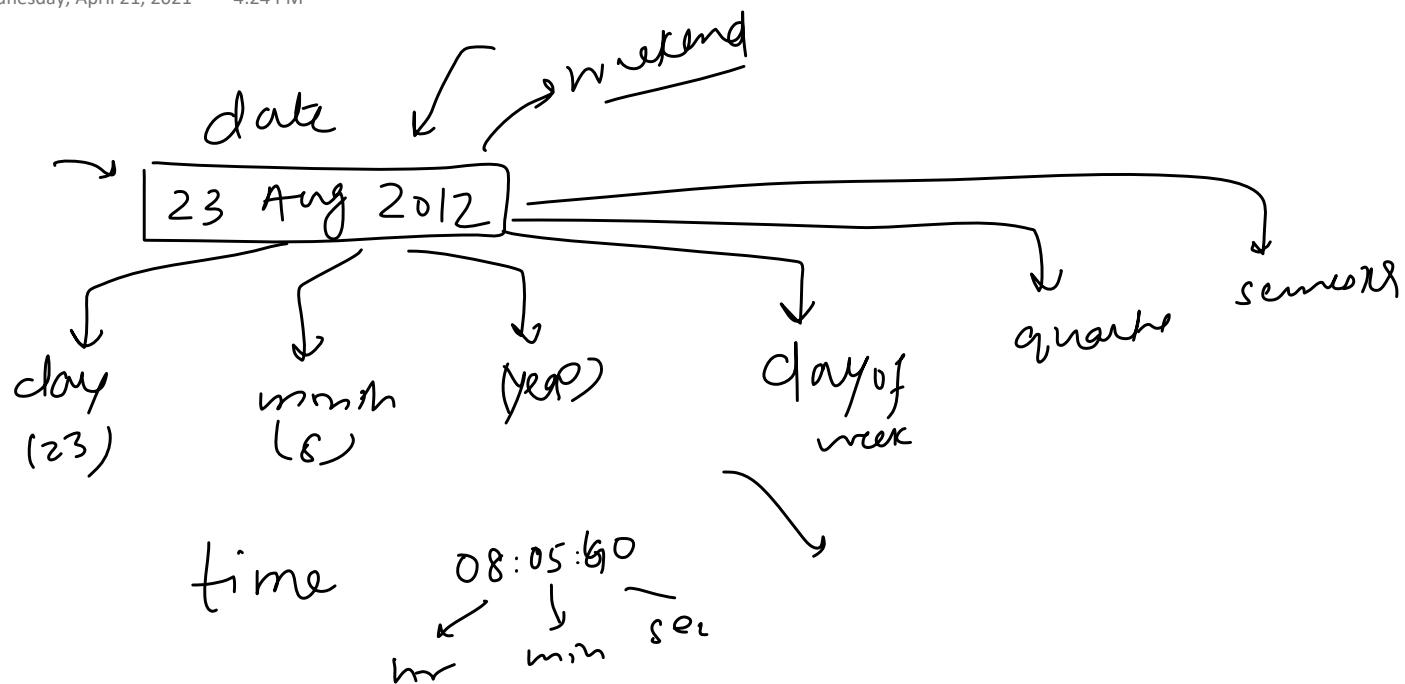
## Mixed Data

Tuesday, April 20, 2021 4:24 PM



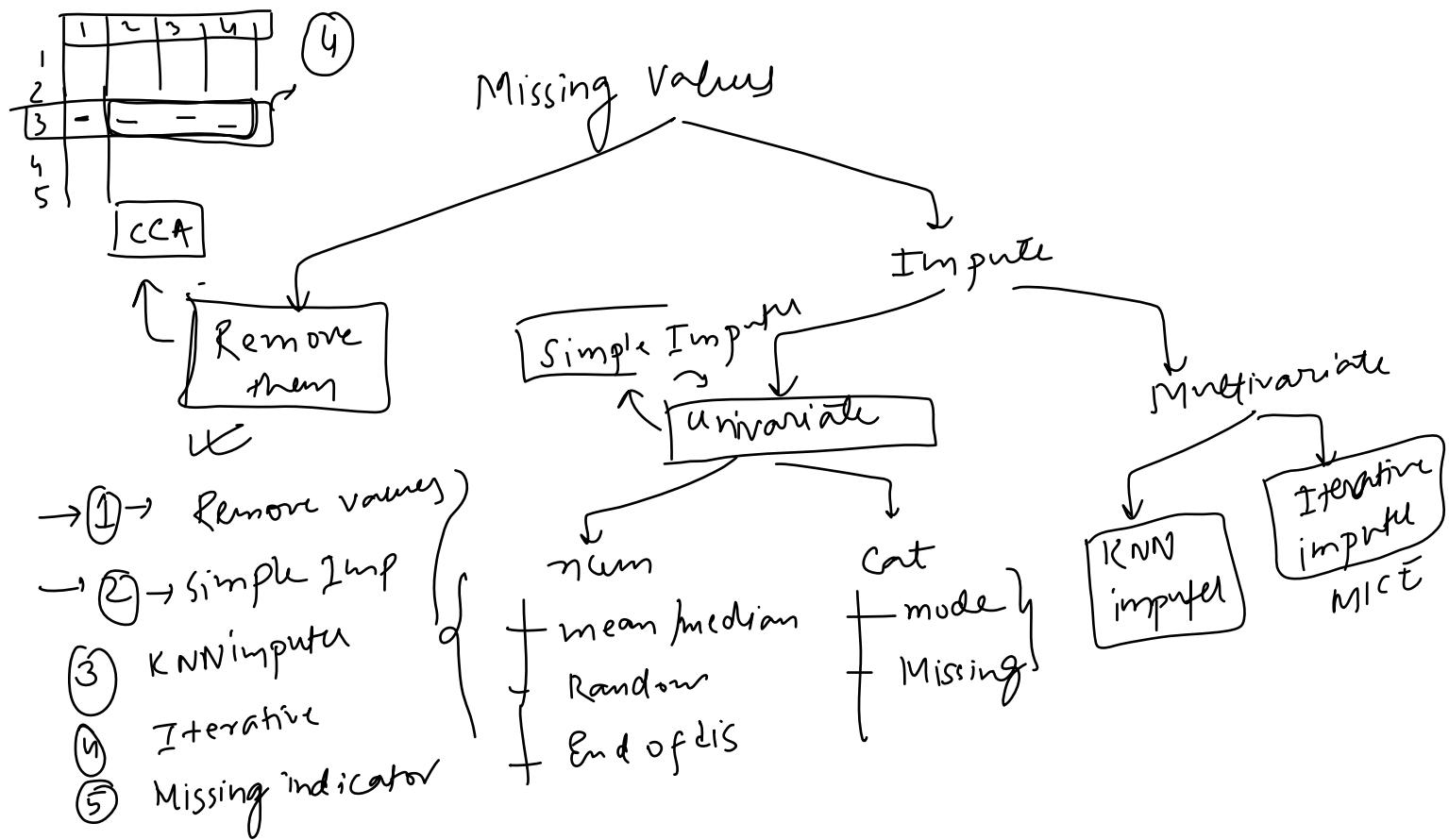
## Working with date and time

Wednesday, April 21, 2021 4:24 PM



# Handling Missing Data

Thursday, April 22, 2021 1:03 PM



## Complete Case Analysis

Thursday, April 22, 2021 12:55 PM

Complete-case analysis (CCA), also called "list-wise deletion" of cases, consists in discarding observations where values in any of the variables are missing.

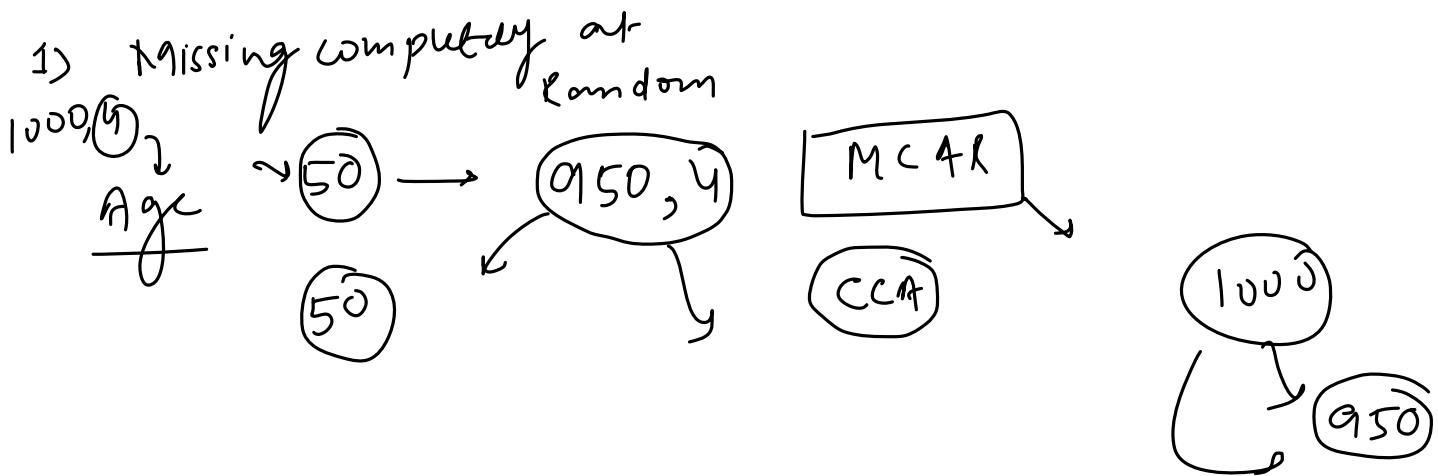
Y6W

CW1

Complete Case Analysis means literally analyzing only those observations for which there is information in **all** of the variables in the dataset.

## Assumption For CCA

Thursday, April 22, 2021 12:58 PM



# Advantage/Disadvantage

Thursday, April 22, 2021 12:59 PM

## Advantage

1. Easy to implement as no data manipulation required
2. Preserves variable distribution (if data is MCAR, then the distribution of the variables of the reduced dataset should match the distribution in the original dataset)

## Disadvantage

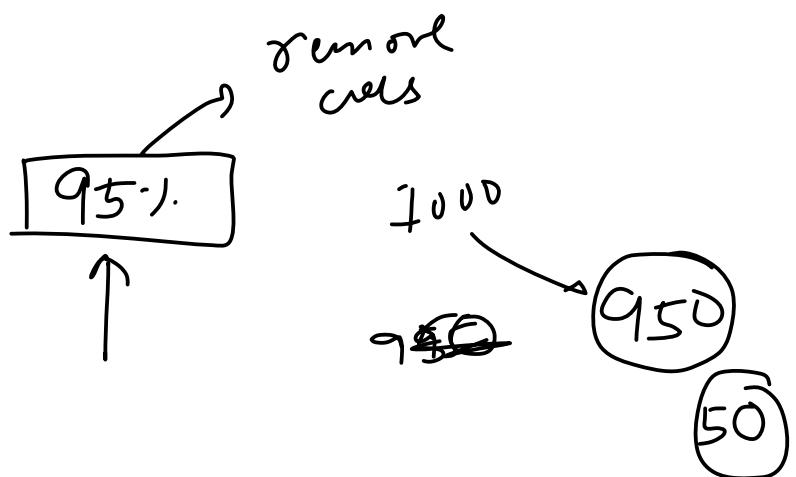
1. It can exclude a large fraction of the original dataset (if missing data is abundant)
2. Excluded observations could be informative for the analysis (if data is not missing at random)
3. When using our models in production, the model will not know how to handle missing data

## When to use CCA?

Thursday, April 22, 2021 1:02 PM

1) MCAR ✓

2)  $\boxed{5\%} <$

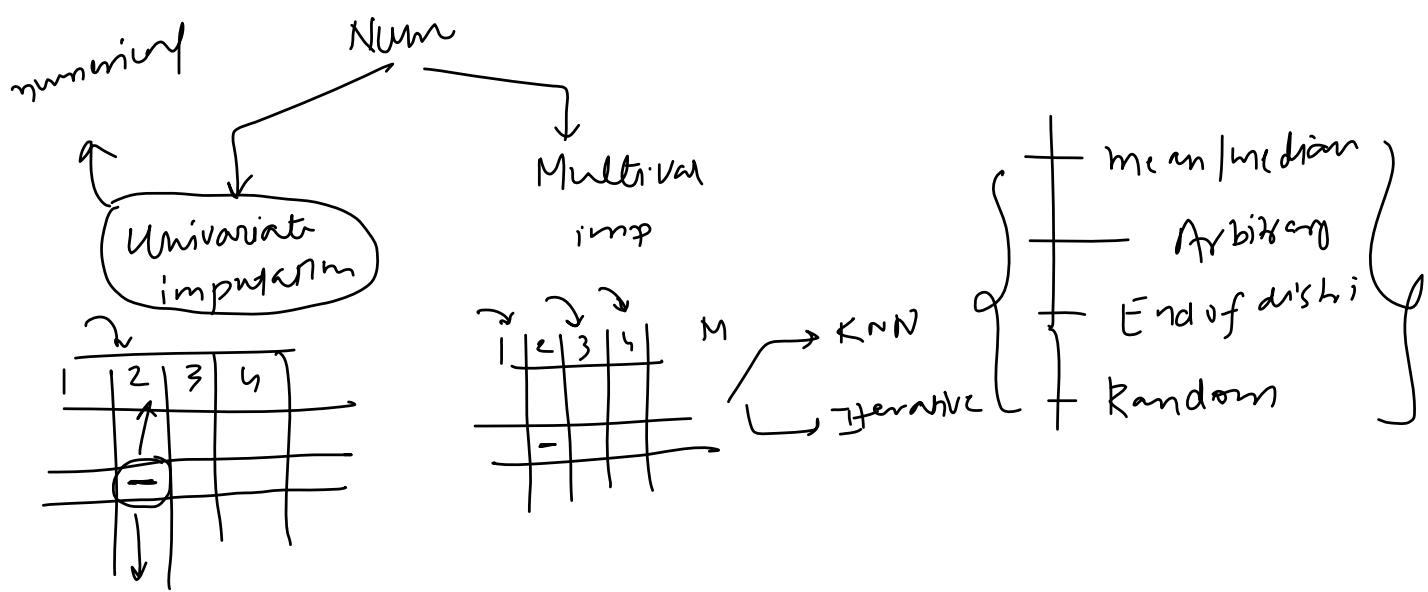


# Example

Thursday, April 22, 2021 1:03 PM

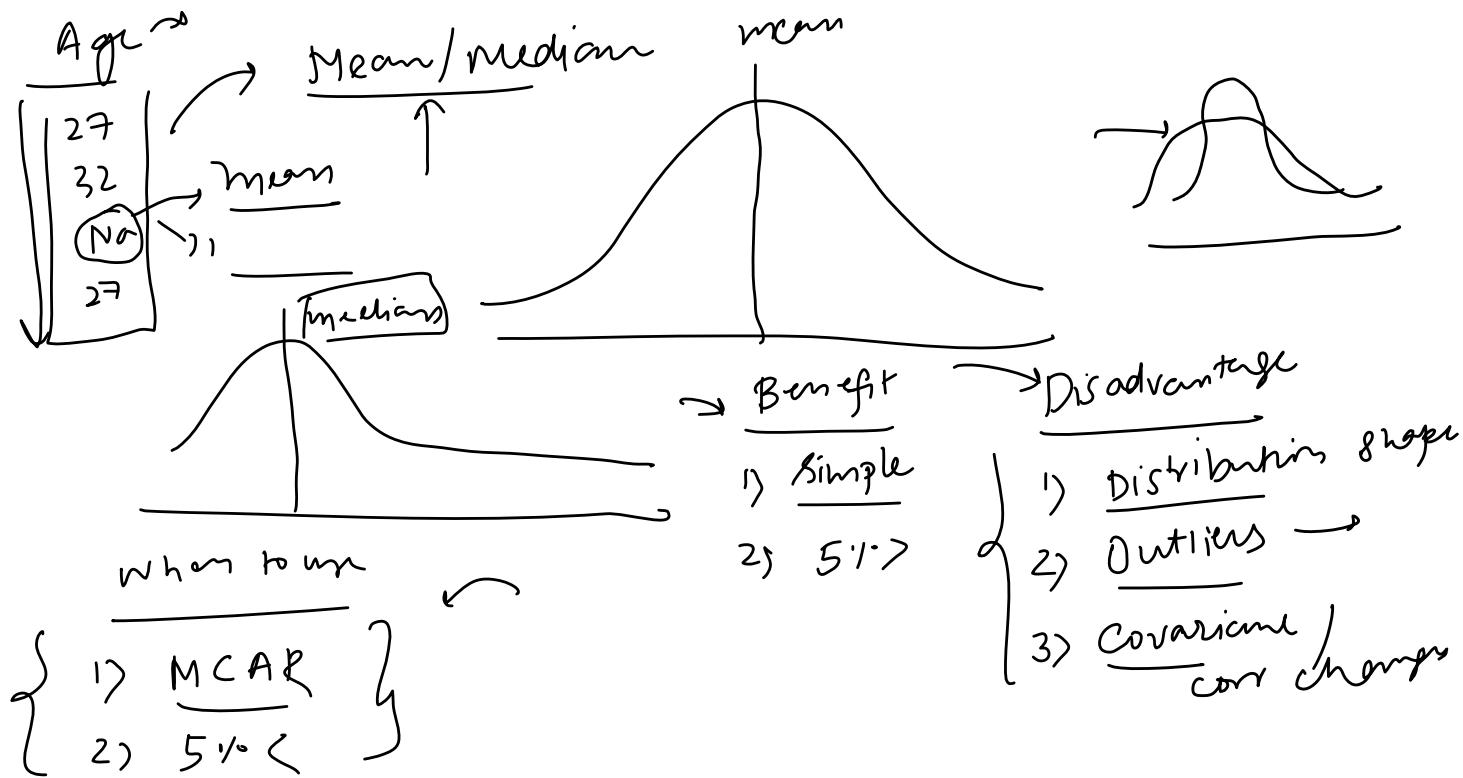
# 1. Handling Missing Numerical Data

Friday, April 23, 2021 11:28 AM



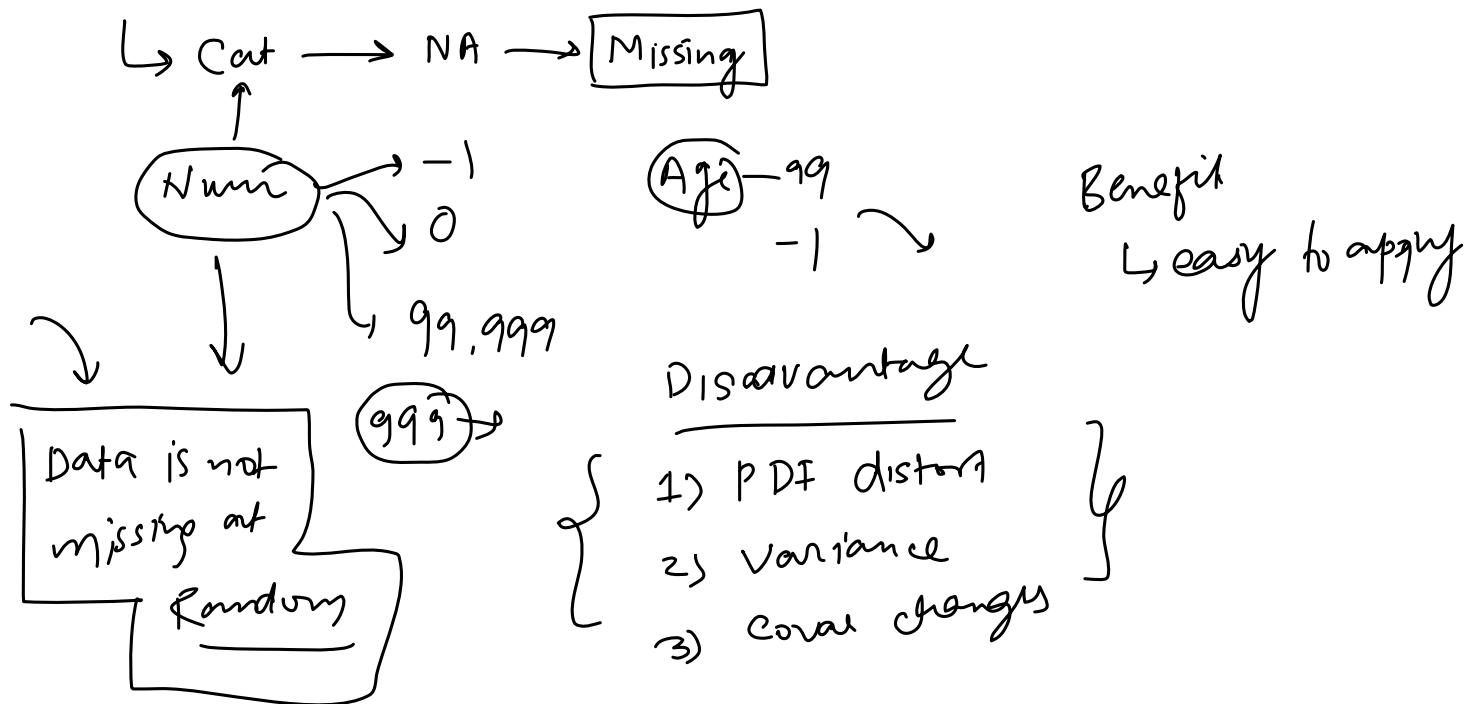
## 2. Mean/Median Imputation

Friday, April 23, 2021 11:31 AM



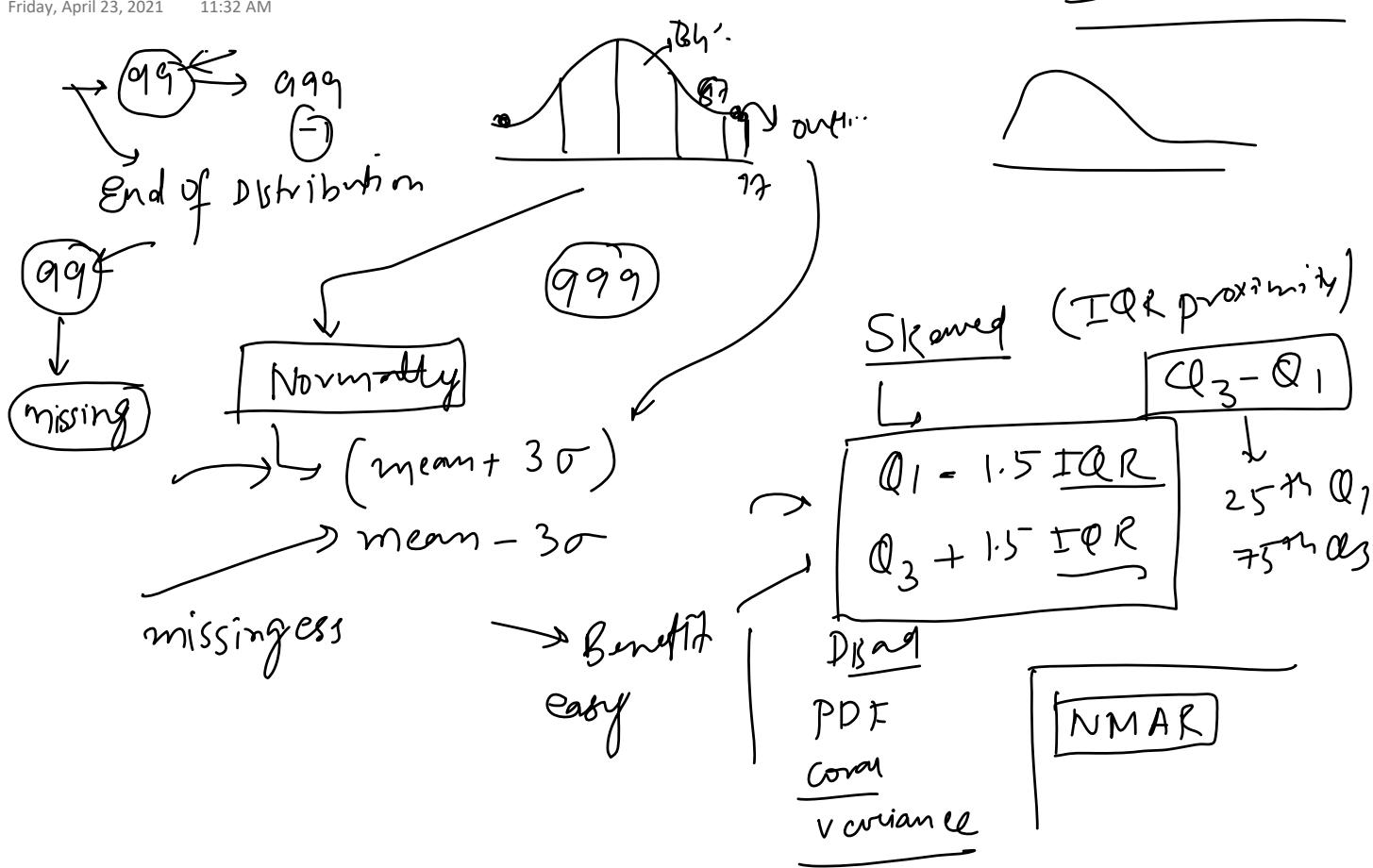
### 3. Arbitrary Value Imputation

Friday, April 23, 2021 11:32 AM



## 4. End of Distribution Imputation

Friday, April 23, 2021 11:32 AM



## 5. Random Sample Imputation

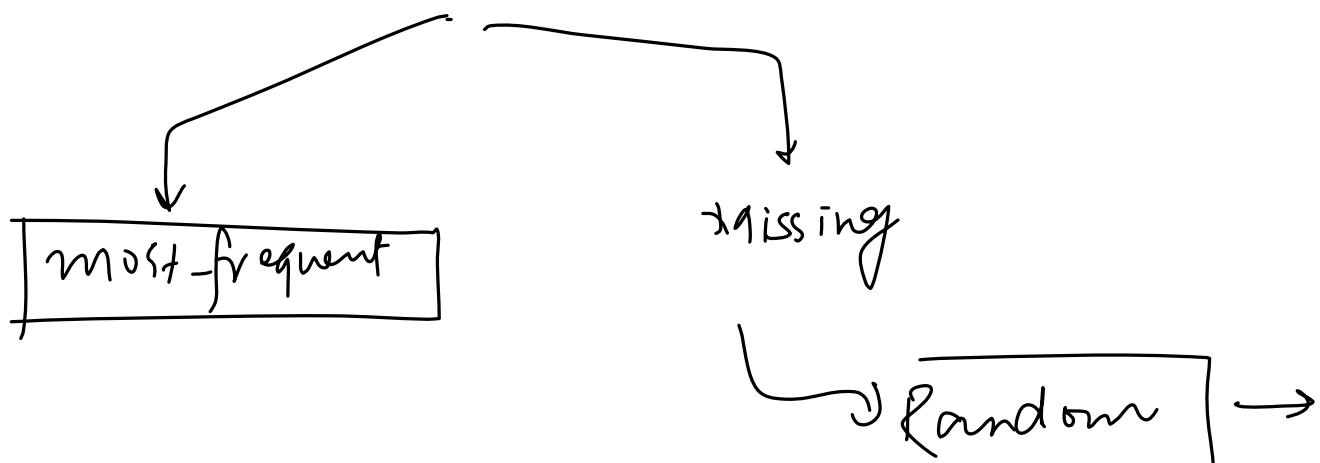
Friday, April 23, 2021 11:33 AM

## 6. Automatically select best imputation technique

Friday, April 23, 2021 11:32 AM

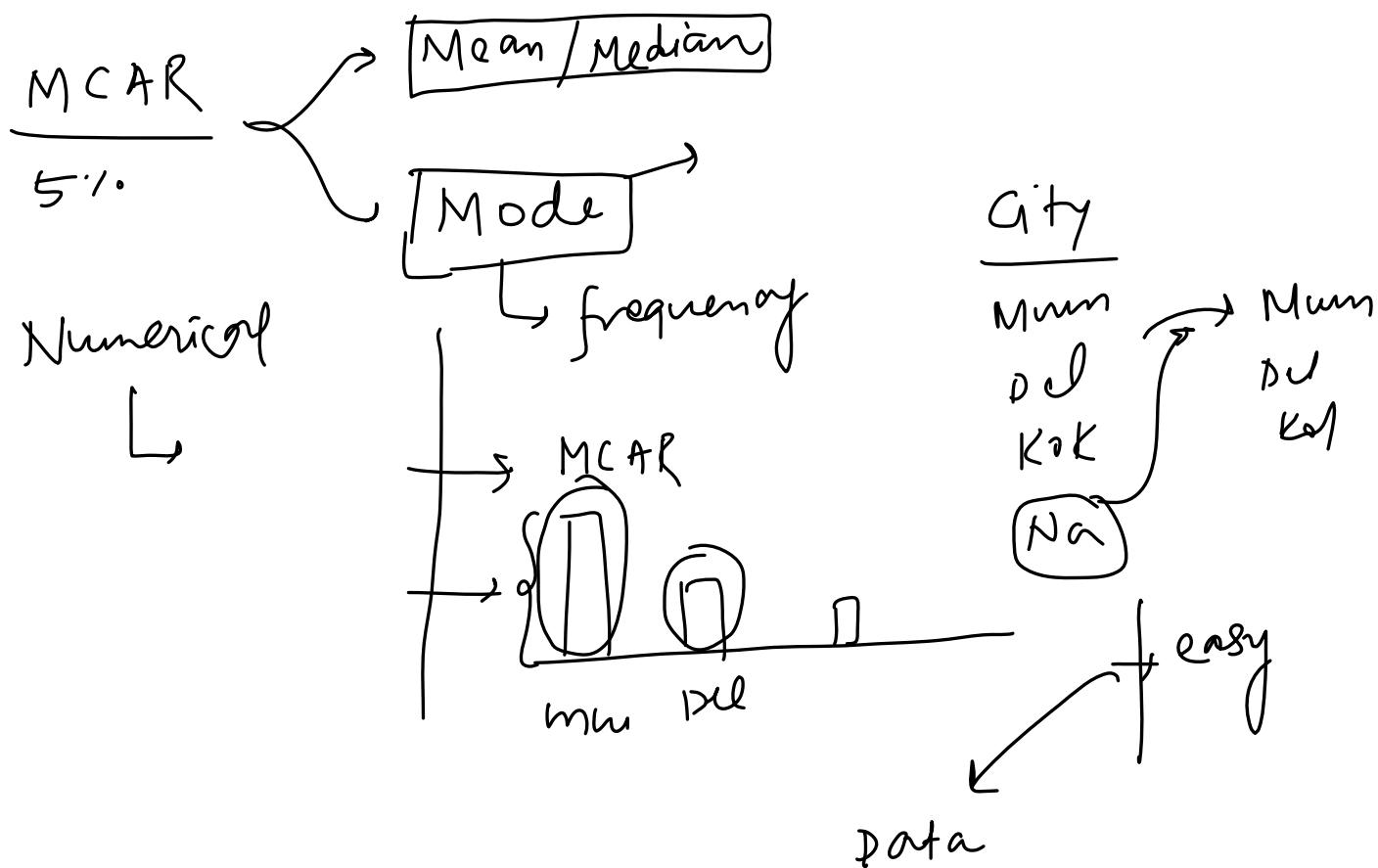
# Handling Categorical Missing Data

Saturday, April 24, 2021 5:16 PM



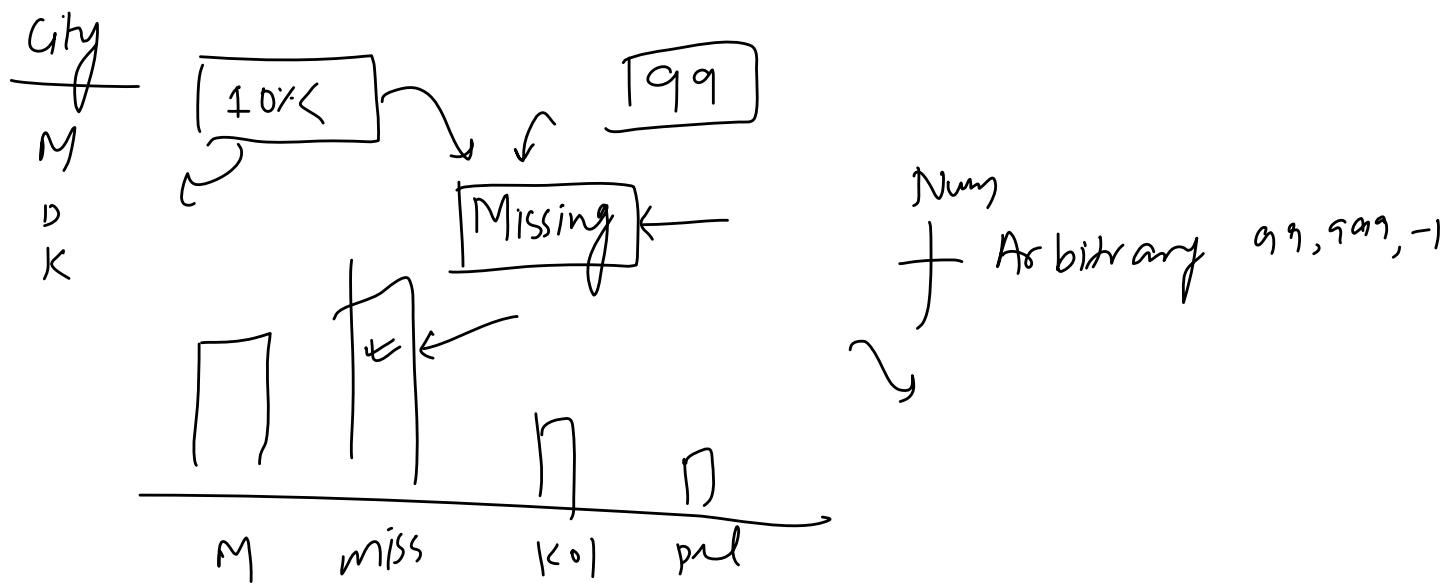
# Most Frequent Value Imputation

Saturday, April 24, 2021 5:17 PM



## Missing Category Imputation

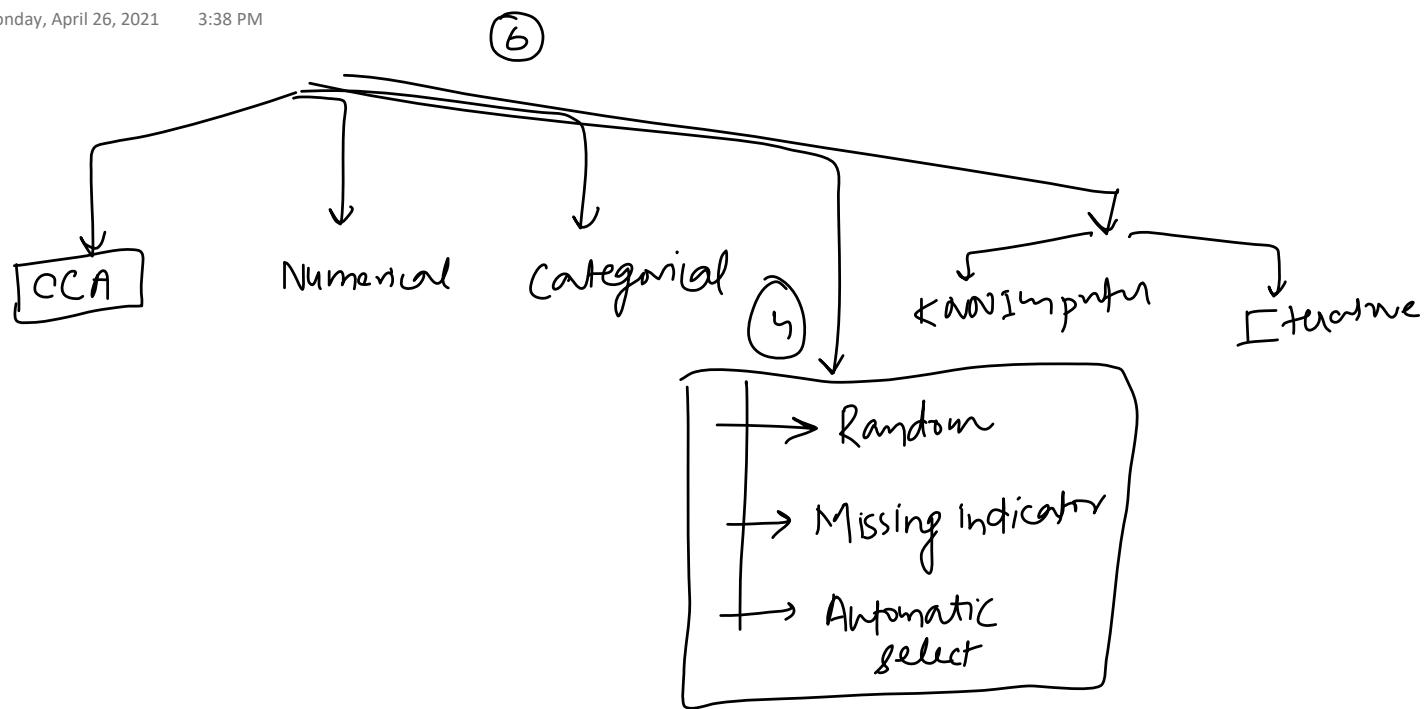
Saturday, April 24, 2021 5:17 PM



## Topics to be covered

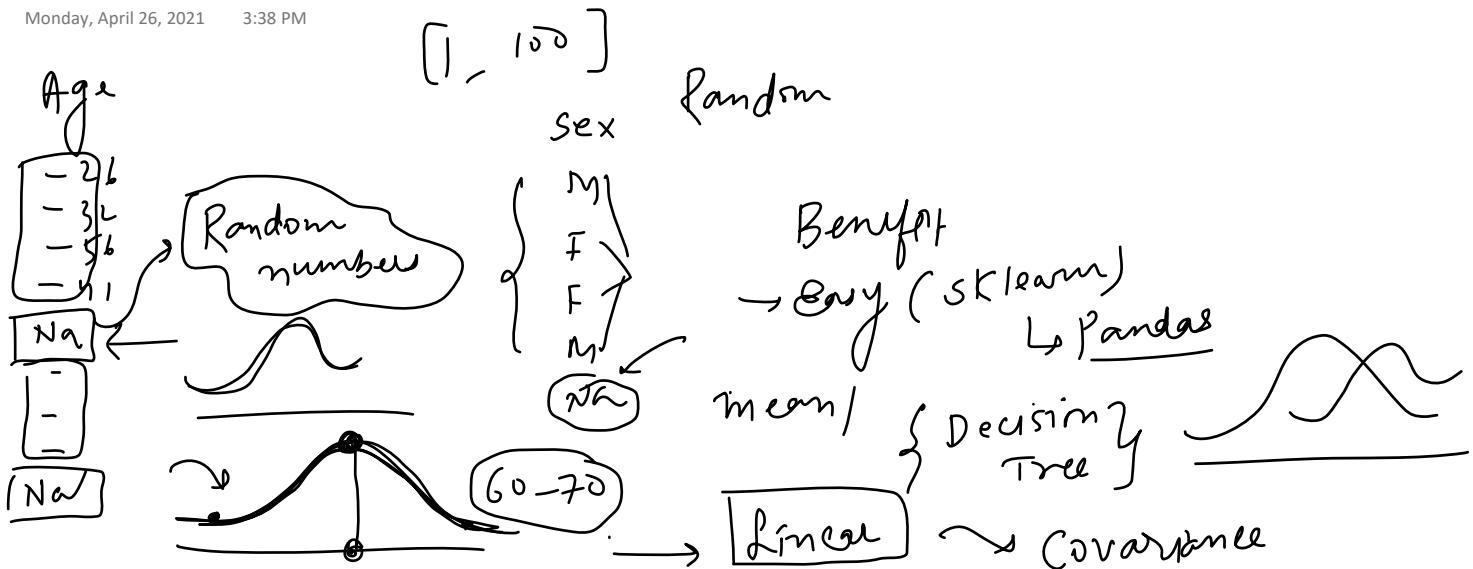
Monday, April 26, 2021 3:38 PM

(6)

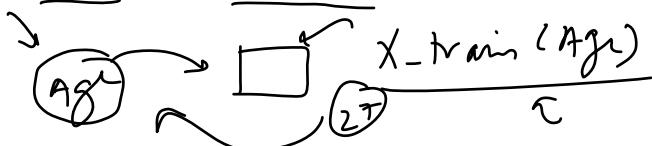


## Random Imputation

Monday, April 26, 2021 3:38 PM

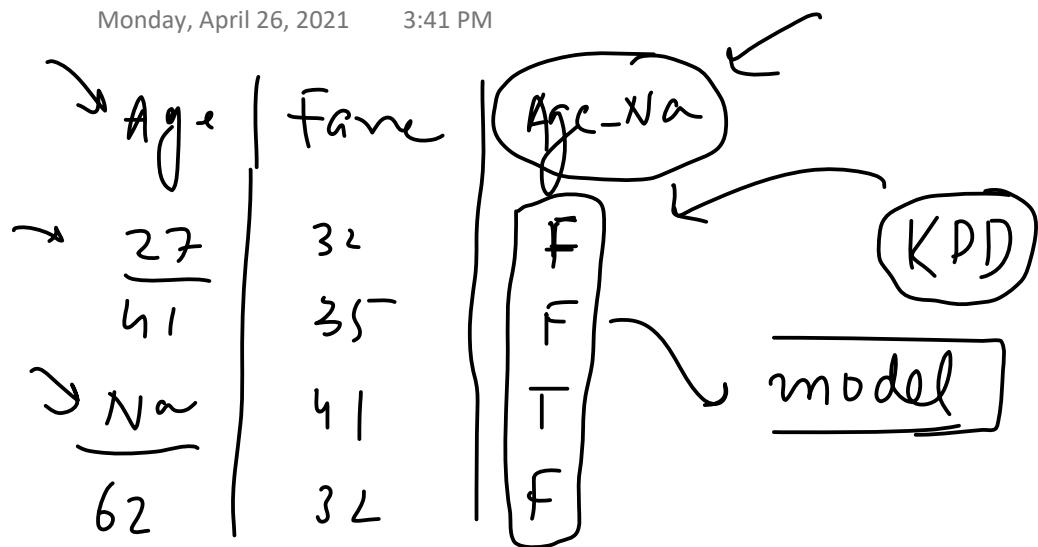


1. Preserves the variance of the variable (But Why?)
2. Memory heavy for deployment, as we need to store the original training set to extract values from and replace the NA in coming observations
3. Well suited for linear models as it does not distort the distribution, regardless of the % of NA



# Missing Indicator

Monday, April 26, 2021 3:41 PM



# Automatically select value for Imputation

Monday, April 26, 2021 3:41 PM

grid search cv

# KNN Imputer

Tuesday, April 27, 2021 12:19 PM

mean

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16



Iterative  
Impute

## Working

Tuesday, April 27, 2021 7:44 AM

K=2

S NO	Feature 1	Feature 2	Feature 3	Feature 4
1	33	-----	67	21
2	31.5	45	68	12
3	23	51	71	18
4	40	-----	81	-----
5	35	60	79	-----

Distance from pt 1

$$d = \sqrt{\frac{3}{2}((68-47)^2 + (12-21)^2)} \\ = \sqrt{1.5(41+81)} = \sqrt{127.5} = 11.29$$

Distance from pt 4

$$d = \sqrt{\frac{3}{2}((81-68)^2)} \\ = \sqrt{3/9} = \sqrt{27} = 5.19$$

Distance from pt 3

$$d = \sqrt{\frac{3}{2}((51-45)^2 + (71-68)^2 + (18-12)^2)} \\ = \sqrt{36 + 9 + 36} = 9$$

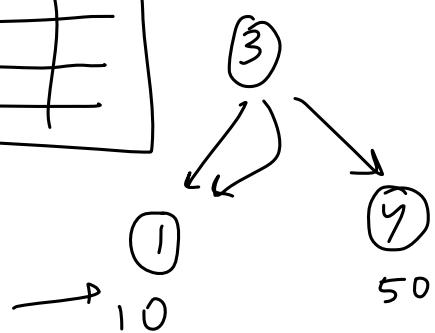
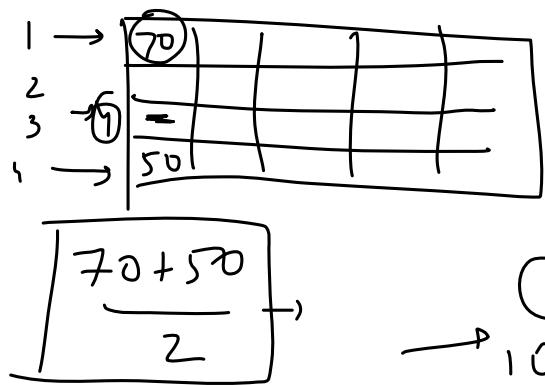
Distance from pt 5

$$d = \sqrt{\frac{3}{2}((60-45)^2 + (79-68)^2)} \\ = \sqrt{1.5(25+121)} = \sqrt{219} = 14.79$$

Advantage & Disadvantage

- 1) More accurate
- 2) More no. of calculations
- 3) Prediction  $\rightarrow$   $X_{train}$

$$K = 2$$



$$\frac{1}{10} \times \cancel{70} + \frac{1}{50} \times \cancel{50}$$

$$\frac{7+1}{2} = 4$$

distance

60

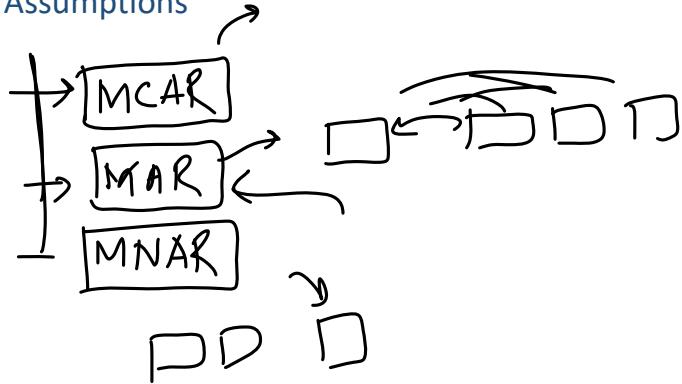
$\frac{1}{\text{dis}}$

# Iterative Imputer/MICE

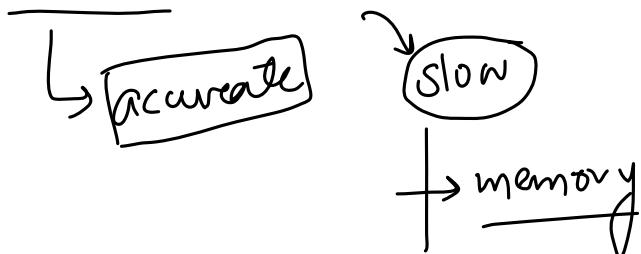
Wednesday, April 28, 2021 11:40 AM

MICE stands for Multivariate Imputation by Chained Equations

Assumptions



Advantages & Disadvantage



## How it works?

Wednesday, April 28, 2021 11:41 AM

### 1. Actual Dataset

	R&D Spend	Administration	Marketing Spend	Profit
21	8.0	15.0	30.0	11.0
37	4.0	5.0	20.0	9.0
2	15.0	10.0	41.0	19.0
14	12.0	16.0	26.0	13.0
44	2.0	15.0	3.0	7.0

### 2. Removing the target column

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	4.0	5.0	20.0
2	15.0	10.0	41.0
14	12.0	16.0	26.0
44	2.0	15.0	3.0

### 3. Introduced some fake nan values

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	NaN	5.0	20.0
2	15.0	10.0	41.0
14	12.0	NaN	26.0
44	2.0	15.0	NaN

Step 1 - Fill all the NaN values with mean of respective cols

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	9.25	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.25	26.00
44	2.00	15.00	29.25

Step 2 - Remove all col1 missing values

Step 3 - Predict the missing values of col1 using other cols

	R&D Spend	Administration	Marketing Spend		Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend	
21	8.00	15.00	30.00		21	15.00	30.00	21	8.00	15.00	30.00
37	Nan	5.00	20.00		37	23.14	5.00	37	23.14	5.00	20.00
2	15.00	10.00	41.00		2	10.00	41.00	2 <td>15.00</td> <td>10.00</td> <td>41.00</td>	15.00	10.00	41.00
14	12.00	11.25	26.00		14	11.25	26.00	14	12.00	11.25	26.00
44	2.00	15.00	29.25		44	15.00	29.25	44	2.00	15.00	29.25

Step 4 - Remove all col2 missing values

Step 5 - Predict the missing values of col2 using other cols

	R&D Spend	Administration	Marketing Spend		R&D Spend	Marketing Spend		R&D Spend	Administration	Marketing Spend	
21	8.00	15.00	30.00		21	8.00	30.00	21	8.00	15.00	30.00
37	23.14	5.00	20.00		37	23.14	20.00	37	23.14	5.00	20.00
2	15.00	10.00	41.00		2	15.00	41.00	2	15.00	10.00	41.00
14	12.00	NaN	26.00		14	12.00	26.00	14	12.00	11.06	26.00
44	2.00	15.00	29.25		44	2.00	29.25	44	2.00	15.00	29.25

Step 6 - Remove all col3 values

	R&D Spend	Administration	Marketing Spend		R&D Spend	Administration		R&D Spend	Administration	Marketing Spend	
21	8.00	15.00	30.00		21	8.00	15.00	21	8.00	15.00	30.00
37	23.14	5.00	20.00		37	23.14	5.00	37	23.14	5.00	20.00
2	15.00	10.00	41.00		2	15.00	10.00	2	15.00	10.00	41.00
14	12.00	11.06	26.00		14	12.00	11.06	14	12.00	11.06	26.00
44	2.00	15.00	29.25		44	2.00	15.00	44	2.00	15.00	31.56

Iteration 0

Iteration 1

Difference

	R&D Spend	Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend	
21	8.00	15.00	30.00		21	8.00	15.00	30.00	21	0.00	0.00	0.00
37	9.25	5.00	20.00		37	23.14	5.00	20.00	37	13.89	0.00	0.00
2	15.00	10.00	41.00		2	15.00	10.00	41.00	2	0.00	0.00	0.00
14	12.00	11.25	26.00		14	12.00	11.22	26.00	14	0.00	-0.19	0.00
44	2.00	15.00	29.25		44	2.00	15.00	31.56	44	0.00	0.00	2.31

Iteration 1

Iteration 2

Difference

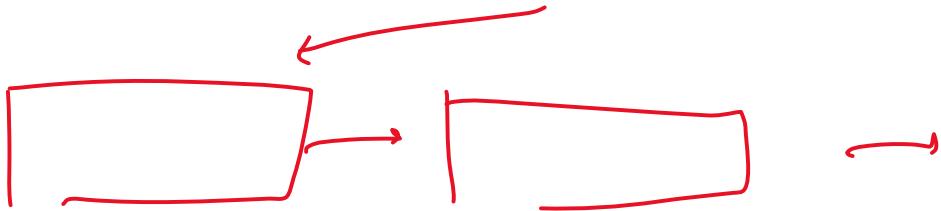
	R&D Spend	Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend	
21	8.00	15.00	30.00		21	8.00	15.00	30.00	21	0.00	0.00	0.00
37	23.14	5.00	20.00		37	23.78	5.00	20.00	37	0.64	0.00	0.00
2	15.00	10.00	41.00		2	15.00	10.00	41.00	2	0.00	0.00	0.00
14	12.00	11.06	26.00		14	12.00	11.22	26.00	14	0.00	0.16	0.00
44	2.00	15.00	31.56		44	2.00	15.00	31.56	44	0.00	0.00	0.00

Iteration 2

Iteration 3

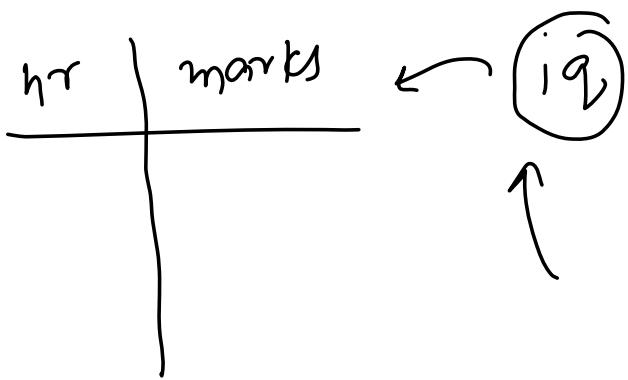
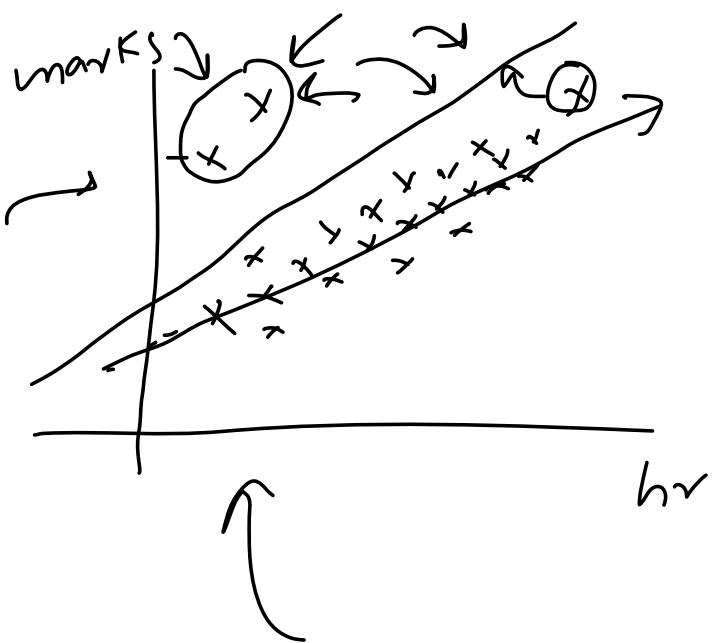
Difference

	R&D Spend	Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend		R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00		21	8.00	15.00	30.00	21	0.00	0.00
37	23.78	5.00	20.00		37	24.57	5.00	20.00	37	0.79	0.00
2	15.00	10.00	41.00		2	15.00	10.00	41.00	2	0.00	0.00
14	12.00	11.22	26.00		14	12.00	11.37	26.00	14	0.00	0.15
44	2.00	15.00	31.56		44	2.00	15.00	31.56	44	0.00	13.97



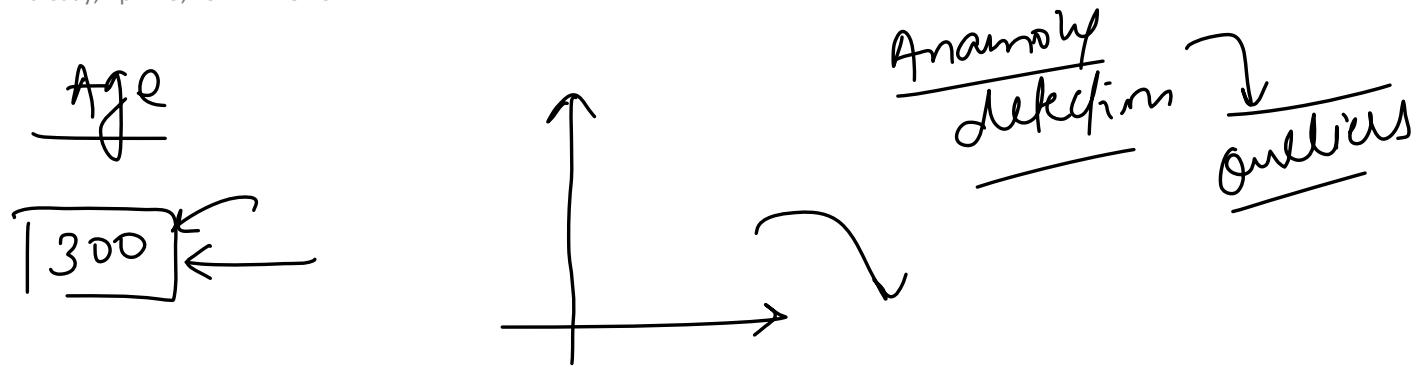
# What are Outliers?

Thursday, April 29, 2021 3:25 PM



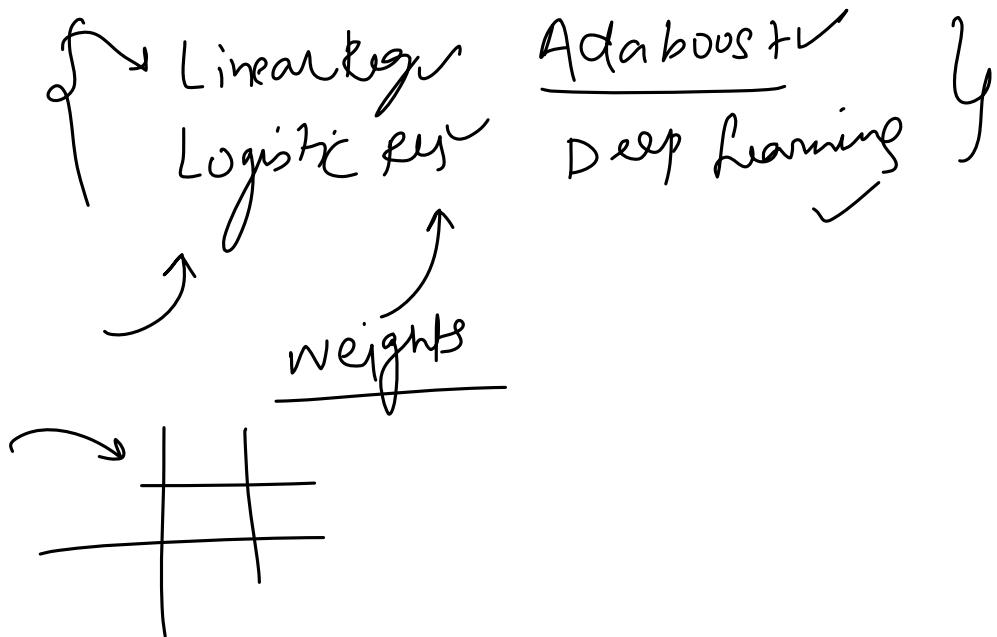
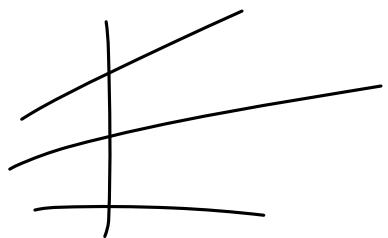
# When is outlier dangerous?

Thursday, April 29, 2021 3:26 PM



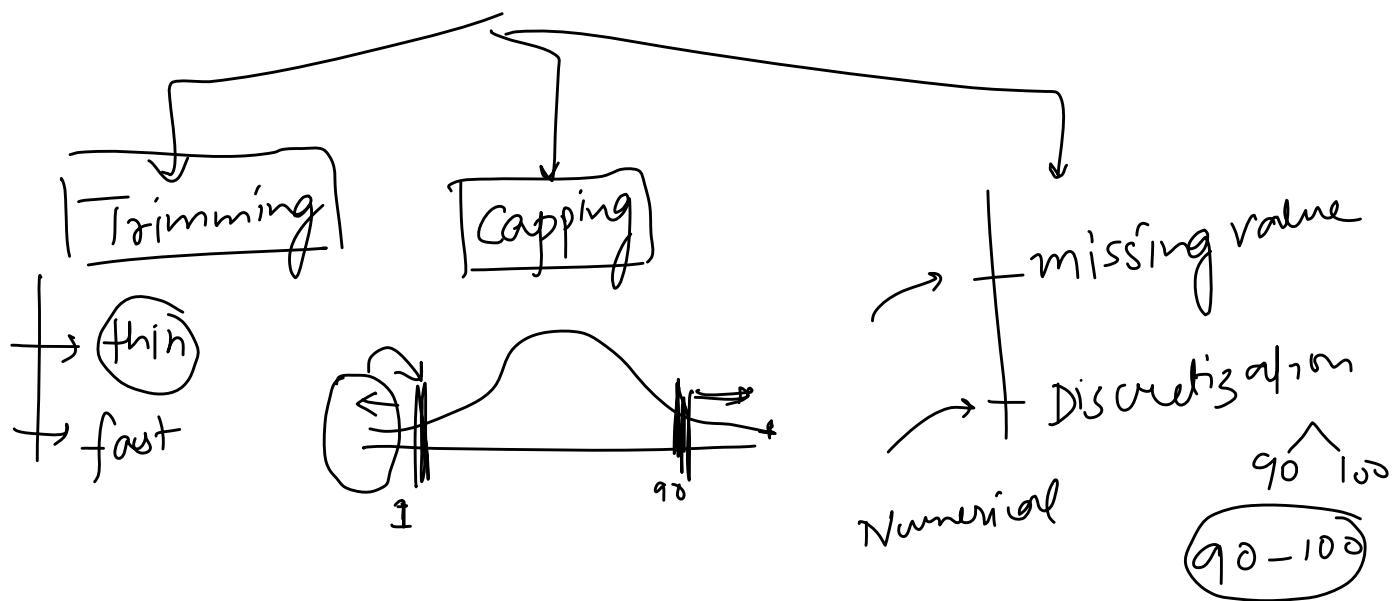
# Effect of Outliers on ML algorithms

Thursday, April 29, 2021 3:27 PM



## How to treat Outliers?

Thursday, April 29, 2021 3:27 PM



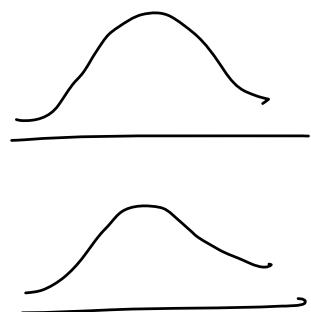
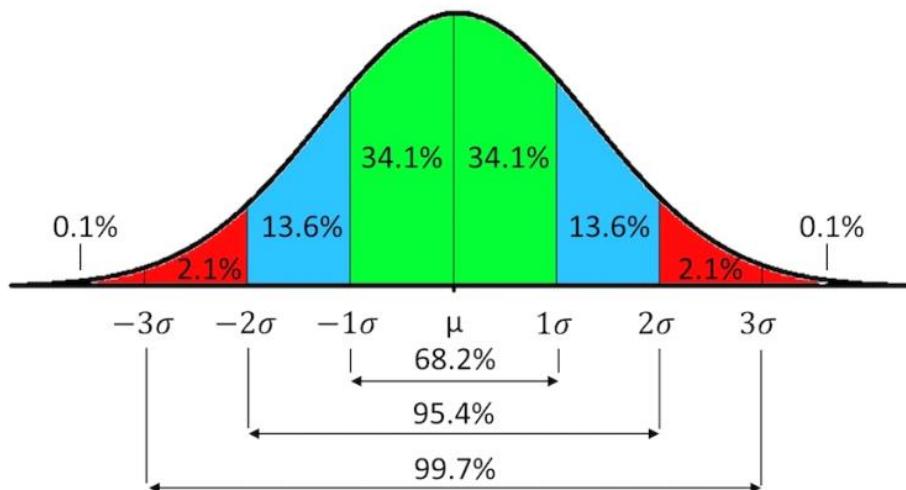
## How to detect Outliers?

Thursday, April 29, 2021 3:27 PM

### 1. Normal Distribution

$$\begin{aligned} (\mu + 3\sigma) &> \\ (\mu - 3\sigma) &< \end{aligned}$$

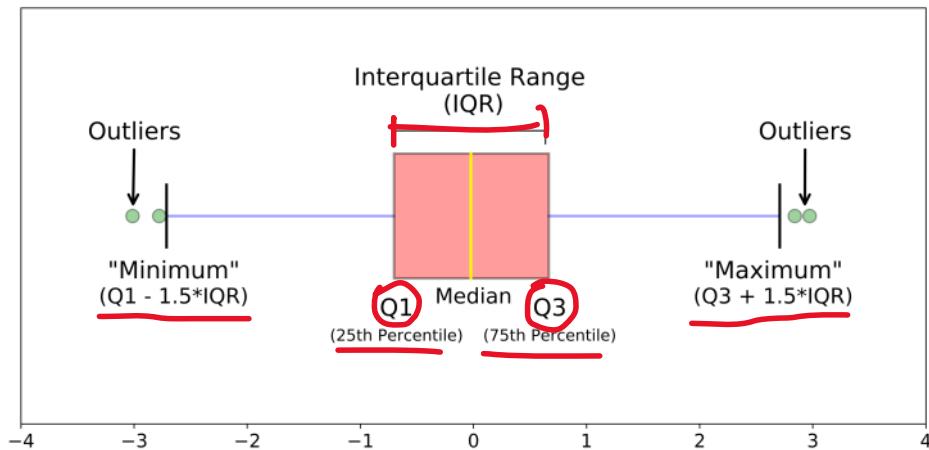
$\sim N$



### 2. Skewed Distribution

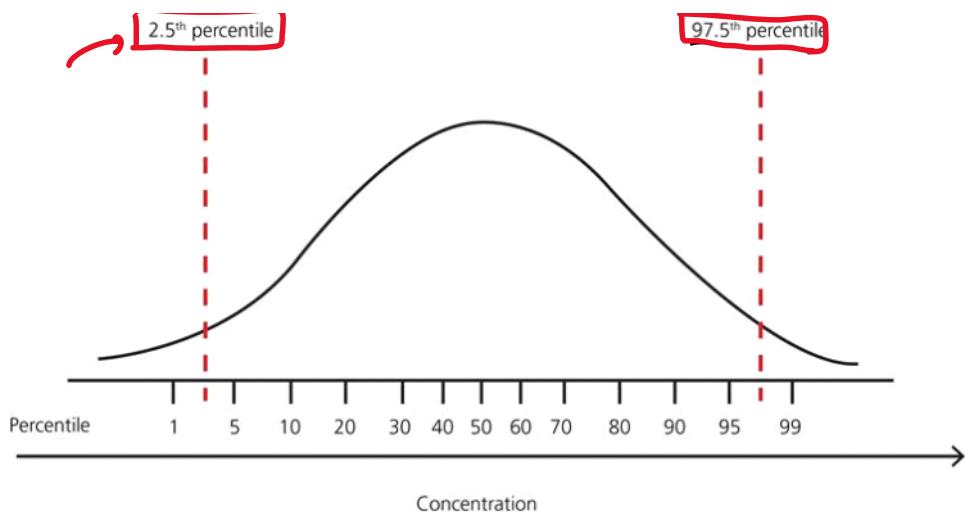
$$\min = \underline{Q_1 - 1.5 \text{ IQR}}$$

$$Q_3 + 1.5 \text{ IQR}$$



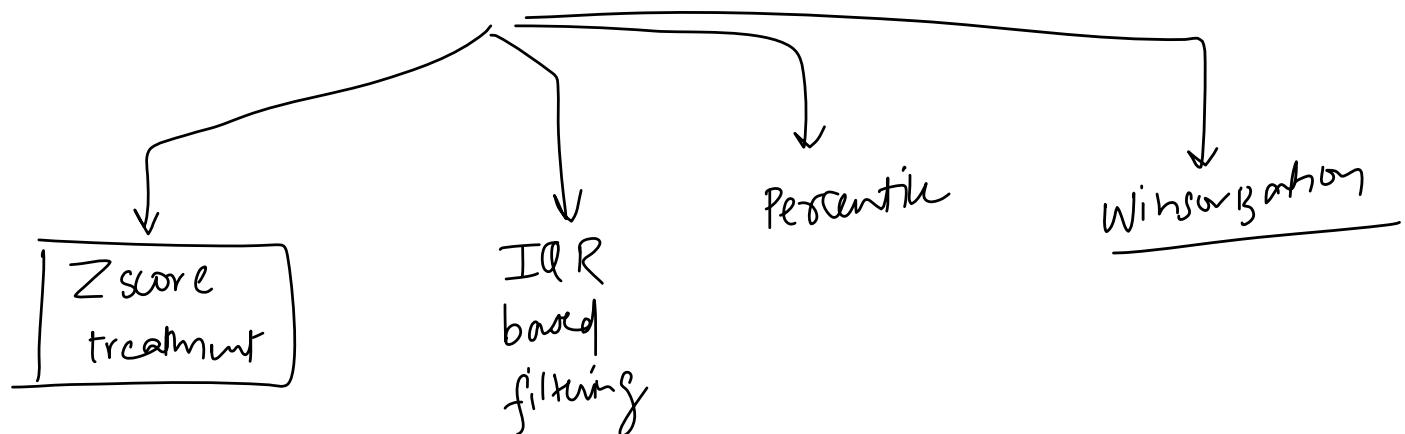
### 3. Other Distributions





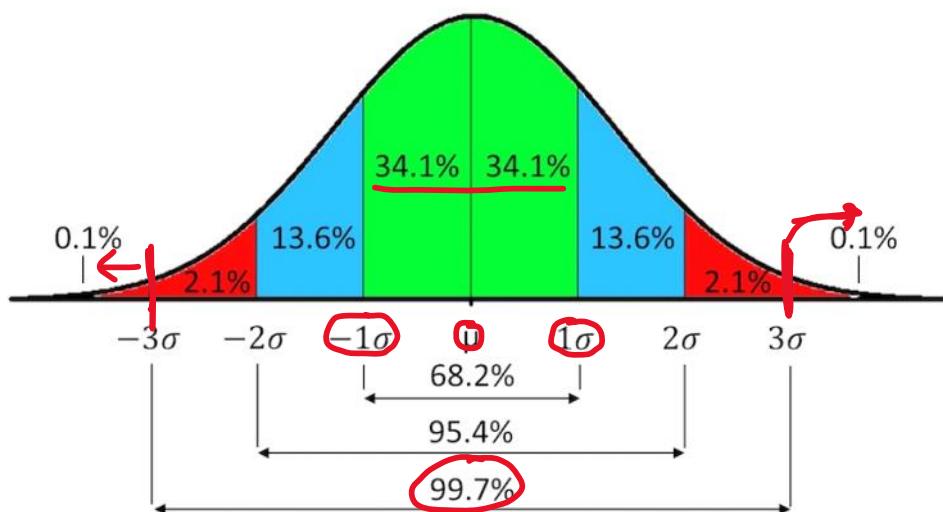
## Techniques for Outlier Detection and Removal

Thursday, April 29, 2021 3:35 PM

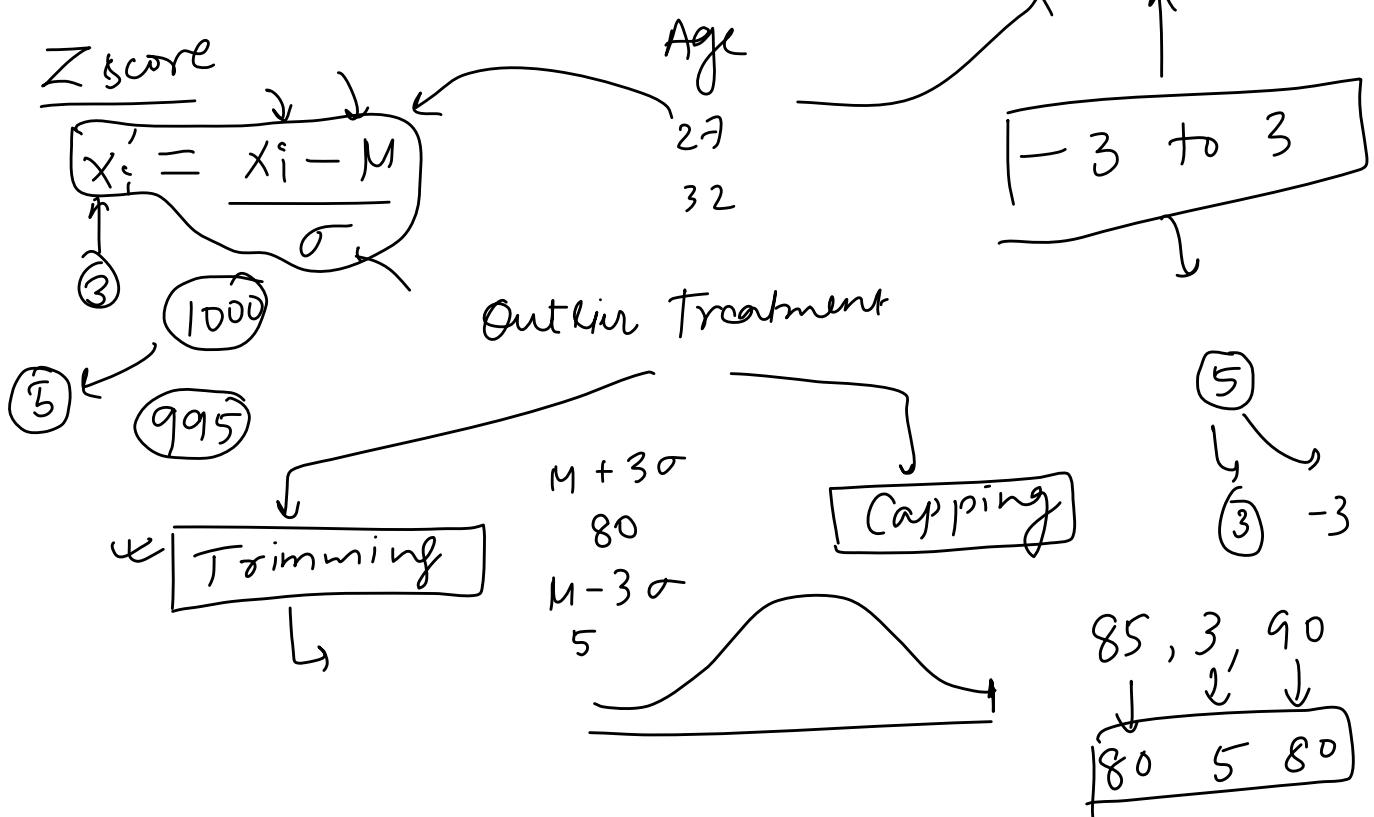


## Outlier removal using Z score

Friday, April 30, 2021 1:33 PM



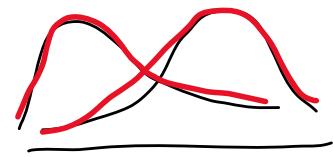
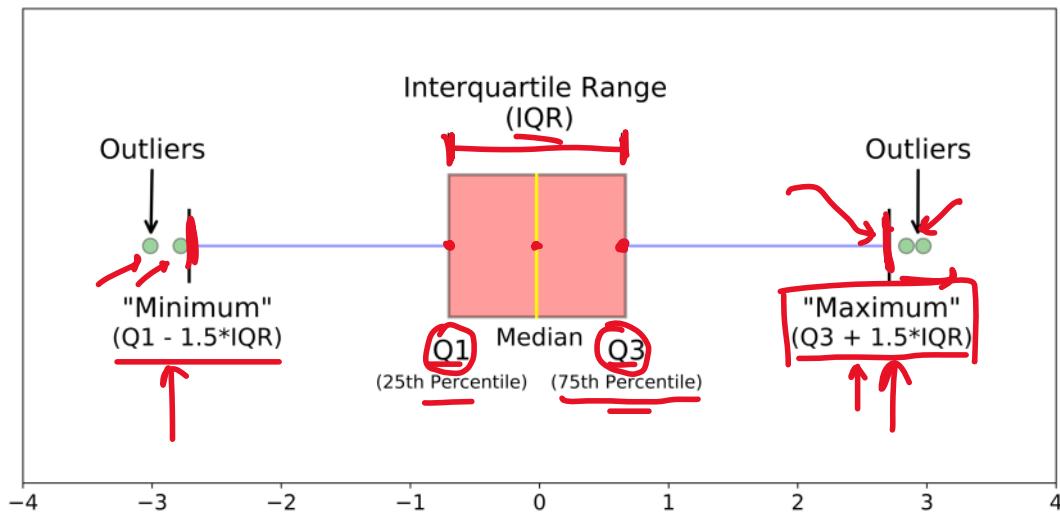
$$\begin{aligned} & \mu + 1\sigma \\ & \left\{ \begin{array}{l} \mu + \sigma \\ \mu - \sigma \end{array} \right\} 68\% \\ & \left\{ \begin{array}{l} \mu + 2\sigma \\ \mu - 2\sigma \end{array} \right\} 95\% \end{aligned}$$



# Outlier Detection using IQR

Friday, April 30, 2021 3:48 PM

**Boxplot  
IQR**



Percentiles

+ 25  
50 Median  
75  
100

Age       $\underline{100} \rightarrow \boxed{89} \leftarrow \max$       25<sup>th</sup>       $\left\{ \begin{array}{l} \pm IQR \\ \text{Proximity Rule} \end{array} \right.$

72  
89  
15  
16       $0 \rightarrow \min \quad \circled{16}$

## Intro

Monday, May 3, 2021 11:49 AM

max -  $\textcircled{95} \rightarrow 100\%$

min - 10  $\rightarrow 0\%$

50%  $\rightarrow \underline{\text{median}}$

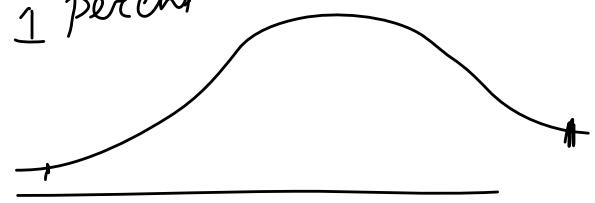
Age

27

35

72

1 Percent



99 $\textcircled{40}$   $\rightarrow \underline{1\text{-per}}$

$\textcircled{5}$

remove

95 per

5 per

$\textcircled{68}$

195 per

capping

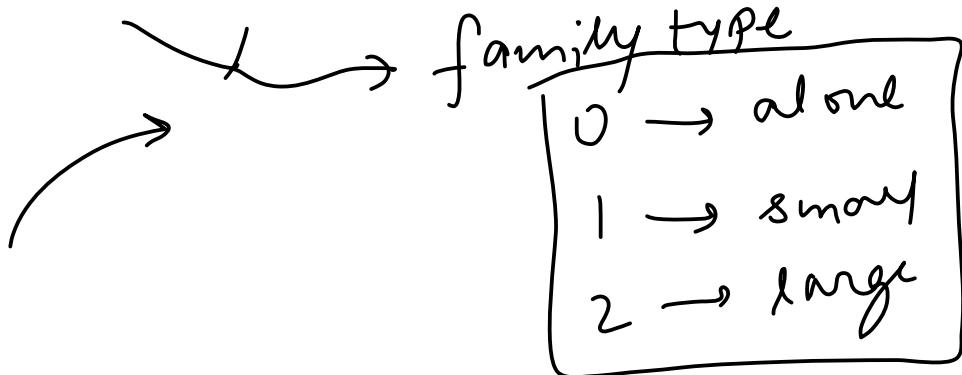
winsorization

# Feature Construction

Tuesday, May 4, 2021 11:13 AM

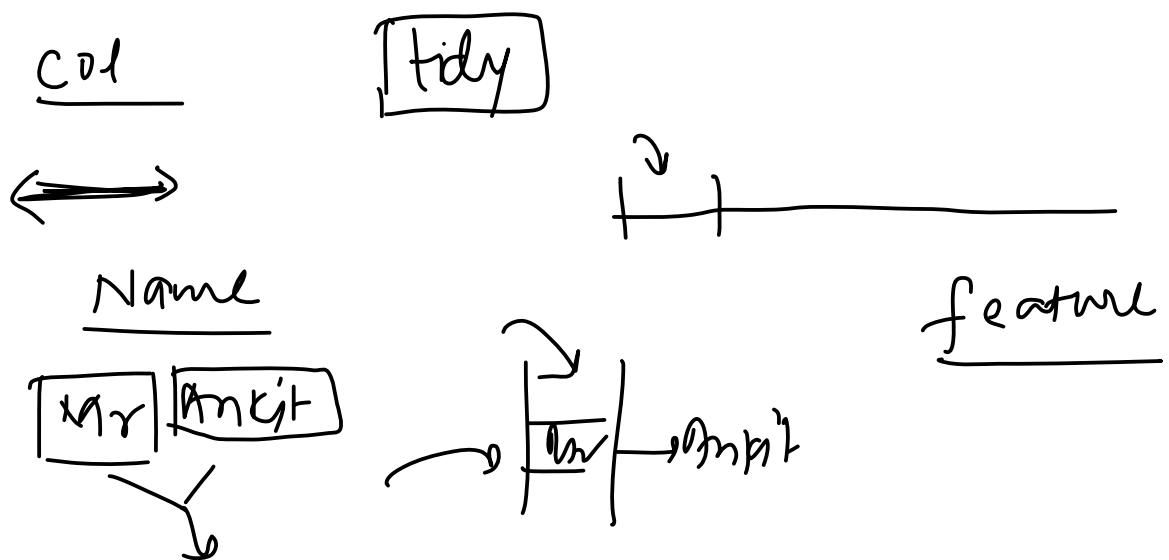
Titanic

SibSp | Parent



# Feature Splitting

Tuesday, May 4, 2021 11:13 AM



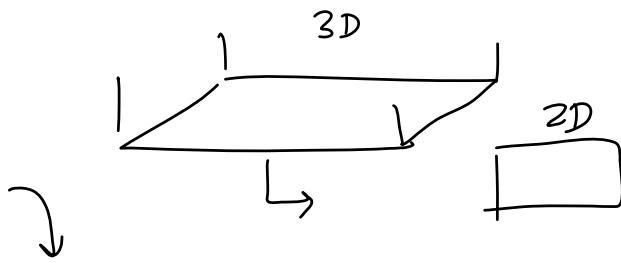
# Curse of Dimensionality

Wednesday, May 5, 2021 1:02 PM

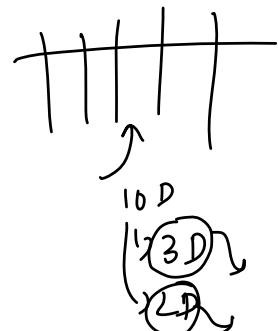
## 1. Introduction

Wednesday, May 5, 2021 1:02 PM

feature extraction → CoD ↓  
↳ PCA → f↓

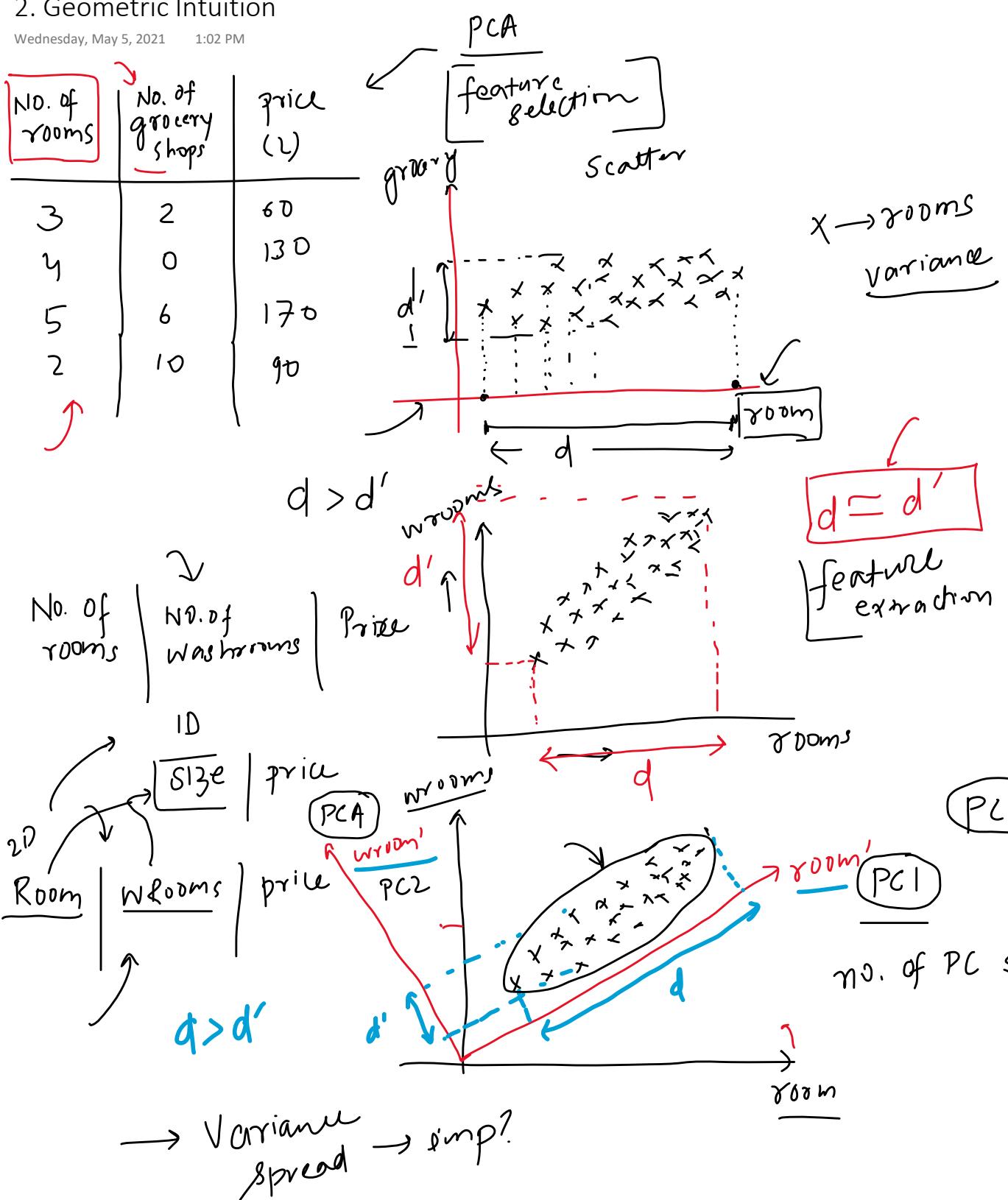


- Benefits
- 1) faster exec of algo
  - 2) Visualization



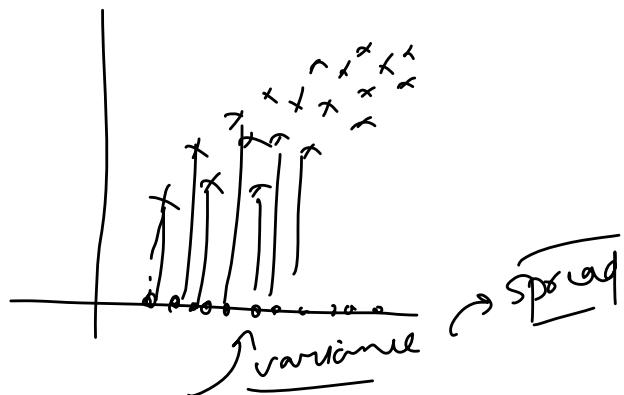
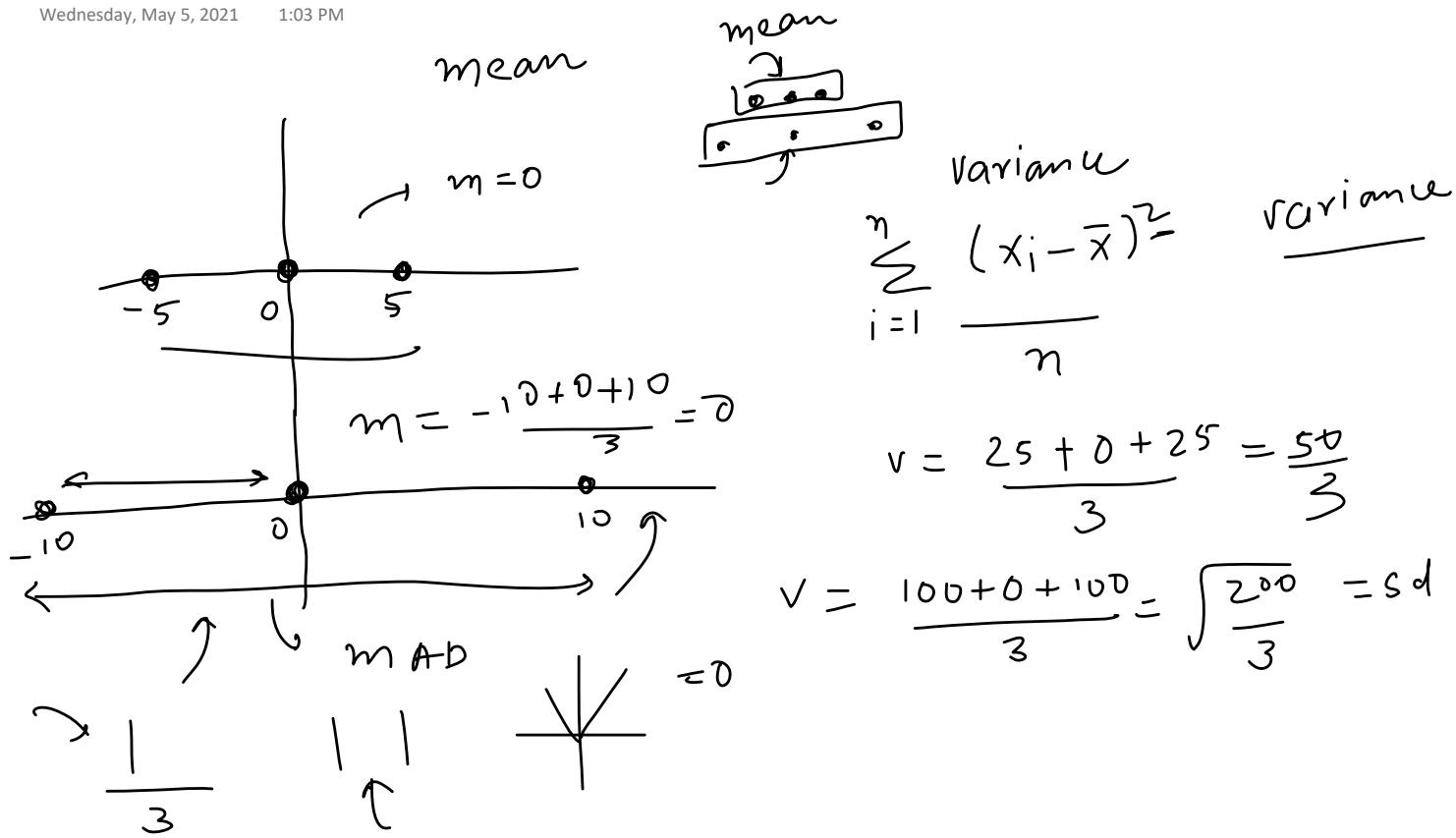
## 2. Geometric Intuition

Wednesday, May 5, 2021 1:02 PM



### 3. Why Variance is Important?

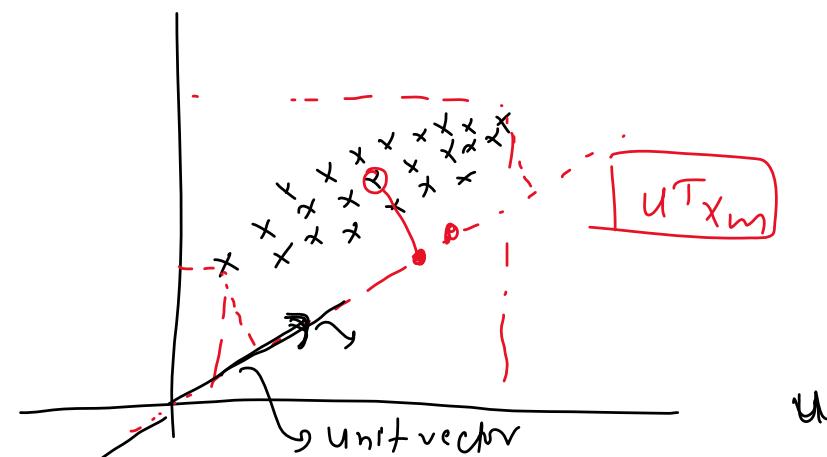
Wednesday, May 5, 2021 1:03 PM



## 4. Problem Formulation

Wednesday, May 5, 2021 1:03 PM

$2D \rightarrow 1D$



$$\frac{\vec{u} \cdot \vec{x}}{|u|} = \vec{u} \cdot \vec{x} = \boxed{u^T x}$$

$\begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

$\begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix}$

$= \boxed{x_1 x_2 + y_1 y_2} - \text{scalar}$

Variance  $u$  maxi

$\sum_{i=1}^n (x_i - \bar{x})^2$

$\sum_{i=1}^n (u^T x_i - \bar{u}^T \bar{x})^2$

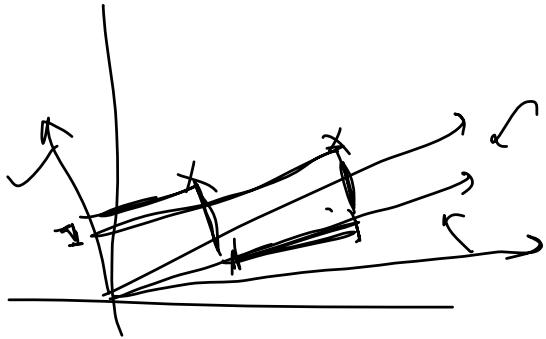
$\max = \text{variance } u$

Cov mat  $\bar{u}$

Optimiz ation

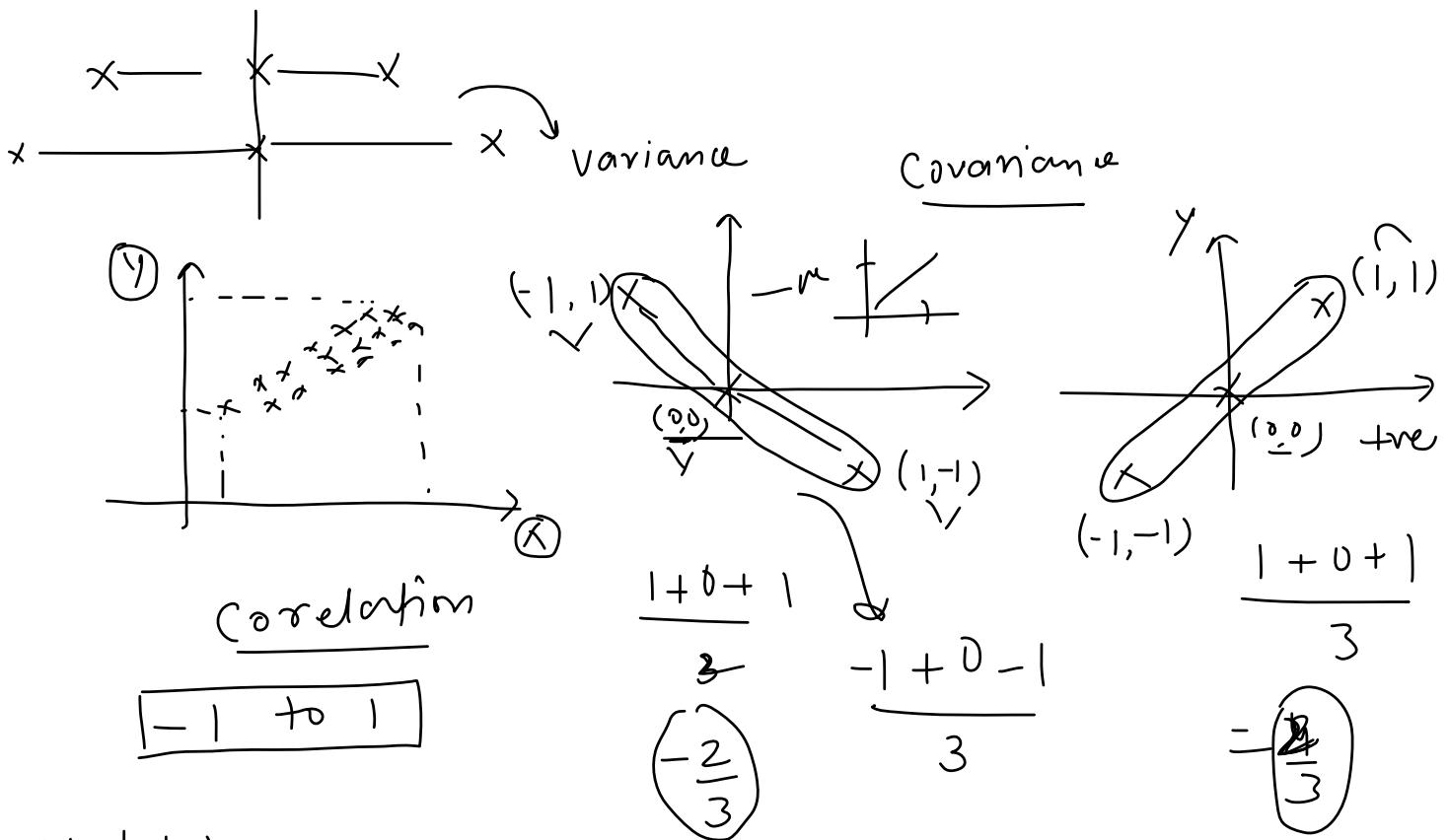
$\mathbf{L}$   $\eta$

$\text{Cov mat}$   $\mathbf{v}_1$   $\rightarrow$  eigen



# Covariance and Covariance Matrix

Thursday, May 6, 2021 7:02 AM



$$\frac{x_1 | x_2}{m} \quad \text{cov mat}$$

$$\begin{matrix} m \\ \hline x_1 & x_2 \end{matrix} \quad \begin{bmatrix} 2 \times 2 \end{bmatrix} \quad \begin{matrix} x_1 \\ x_2 \end{matrix} \quad \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_2, x_1) \\ \text{cov}(x_1, x_2) & \text{cov}(x_2, x_2) \end{bmatrix} \quad \checkmark$$

$\text{cov mat}$

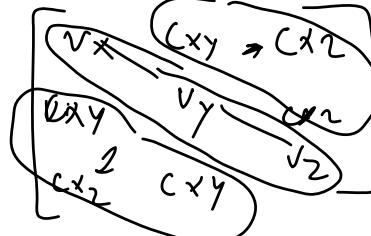
$$\text{cov}(x_1, x_1) \quad \text{var } x_1$$

$$\text{var}(x_1) \quad \text{cov}(x_2, x_1)$$

$$\text{cov}(x_1, x_2) \quad \text{var } x_2$$

square symmetric

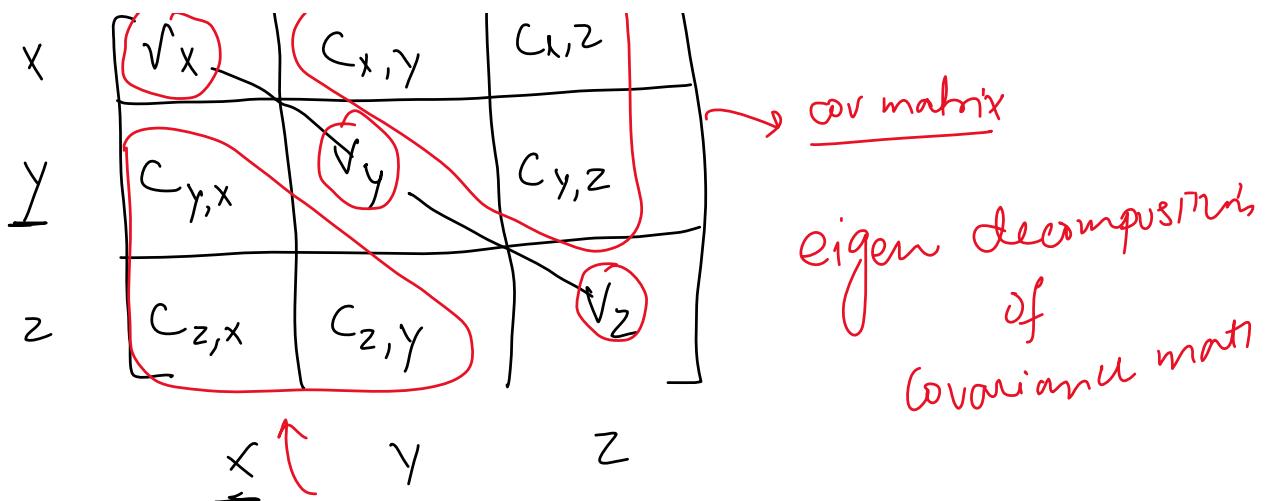
$$\text{cov}(a, b) = \text{cov}(b, a)$$



$$x | y | z$$

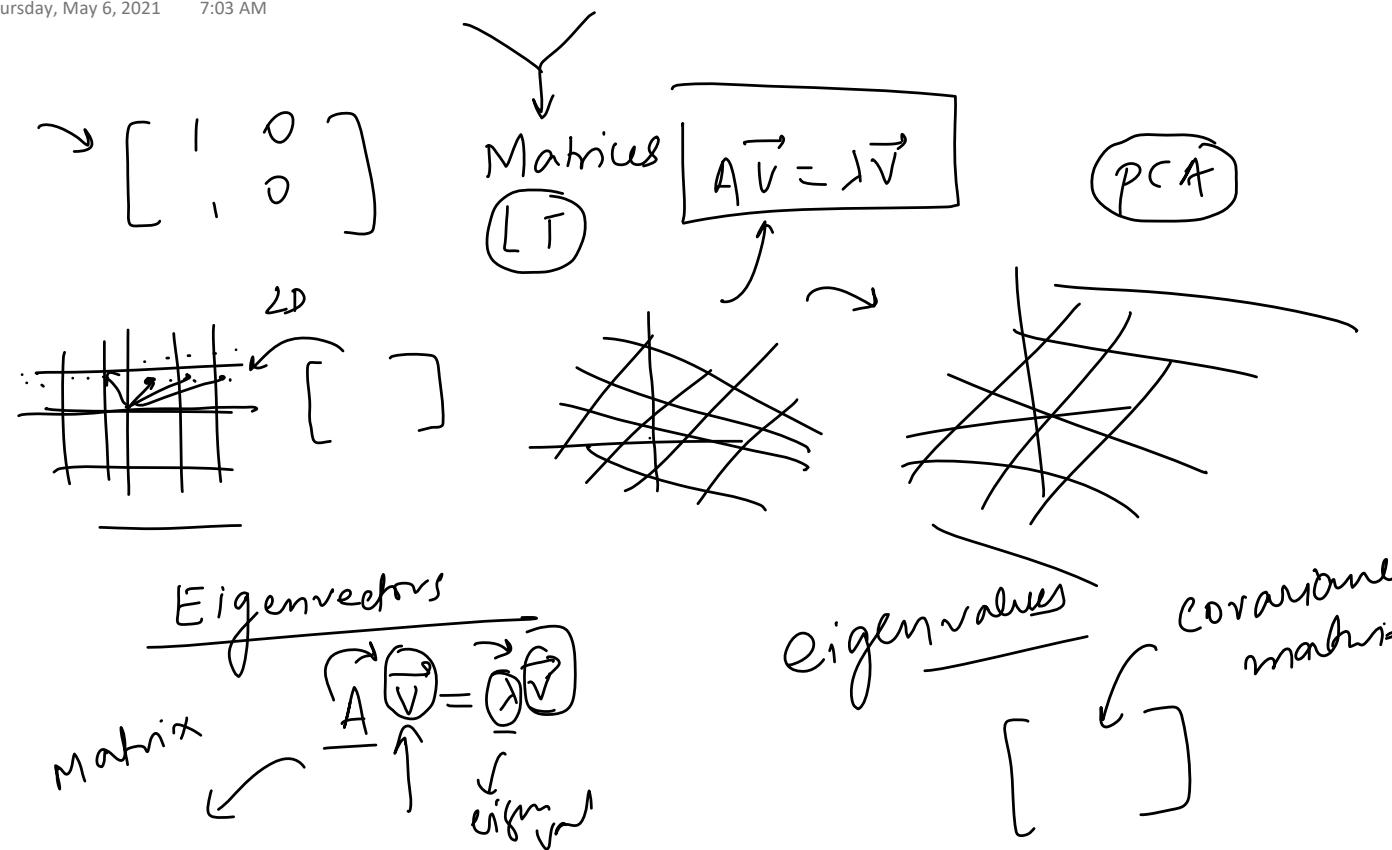
$$x \quad \boxed{\sqrt{x}} \quad | \quad C_{x,y} \quad | \quad C_{x,z}$$

our matrix



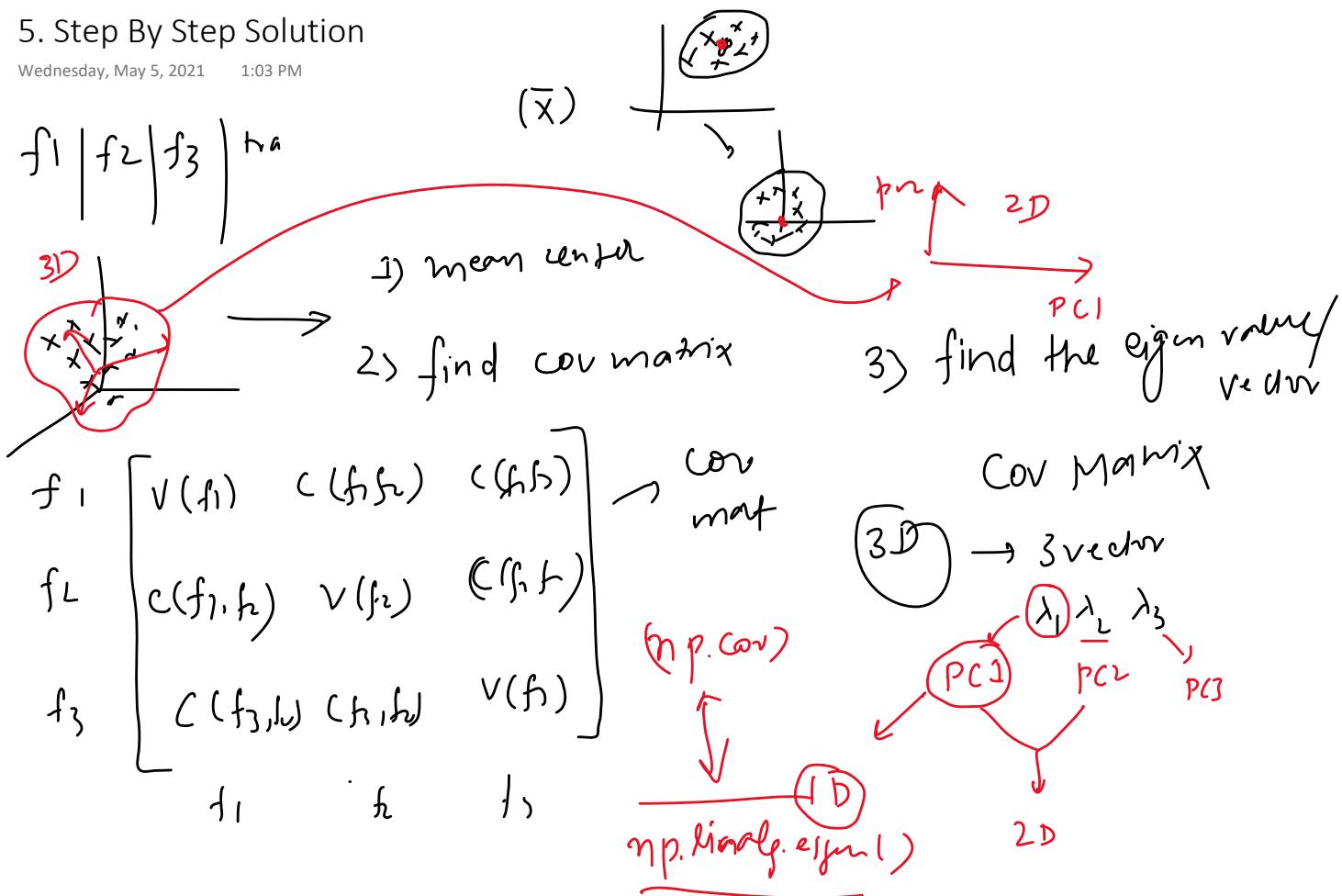
# Linear Transformation, Eigen Vectors and Eigen Values

Thursday, May 6, 2021 7:03 AM



## 5. Step By Step Solution

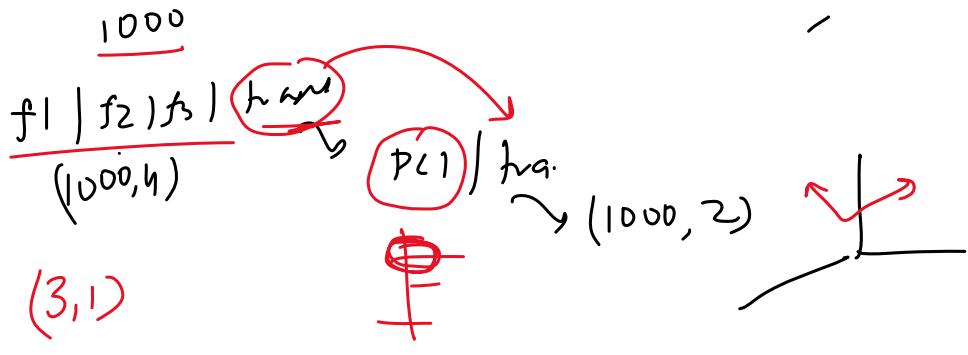
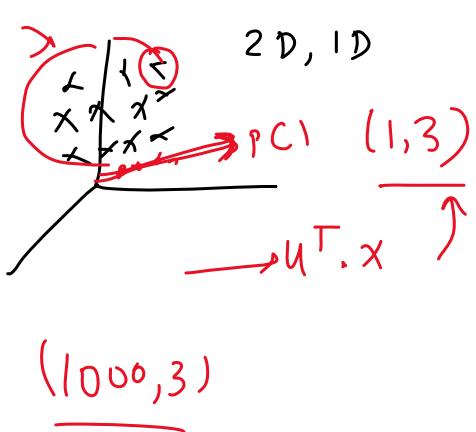
Wednesday, May 5, 2021 1:03 PM



## How to transform points?

Thursday, May 6, 2021

12:31 PM



$$(1000, 3) \cdot (3, 1) = (1000, 1)$$

$$\cancel{PCL} \mid \cancel{PC2} \mid \text{Target}$$

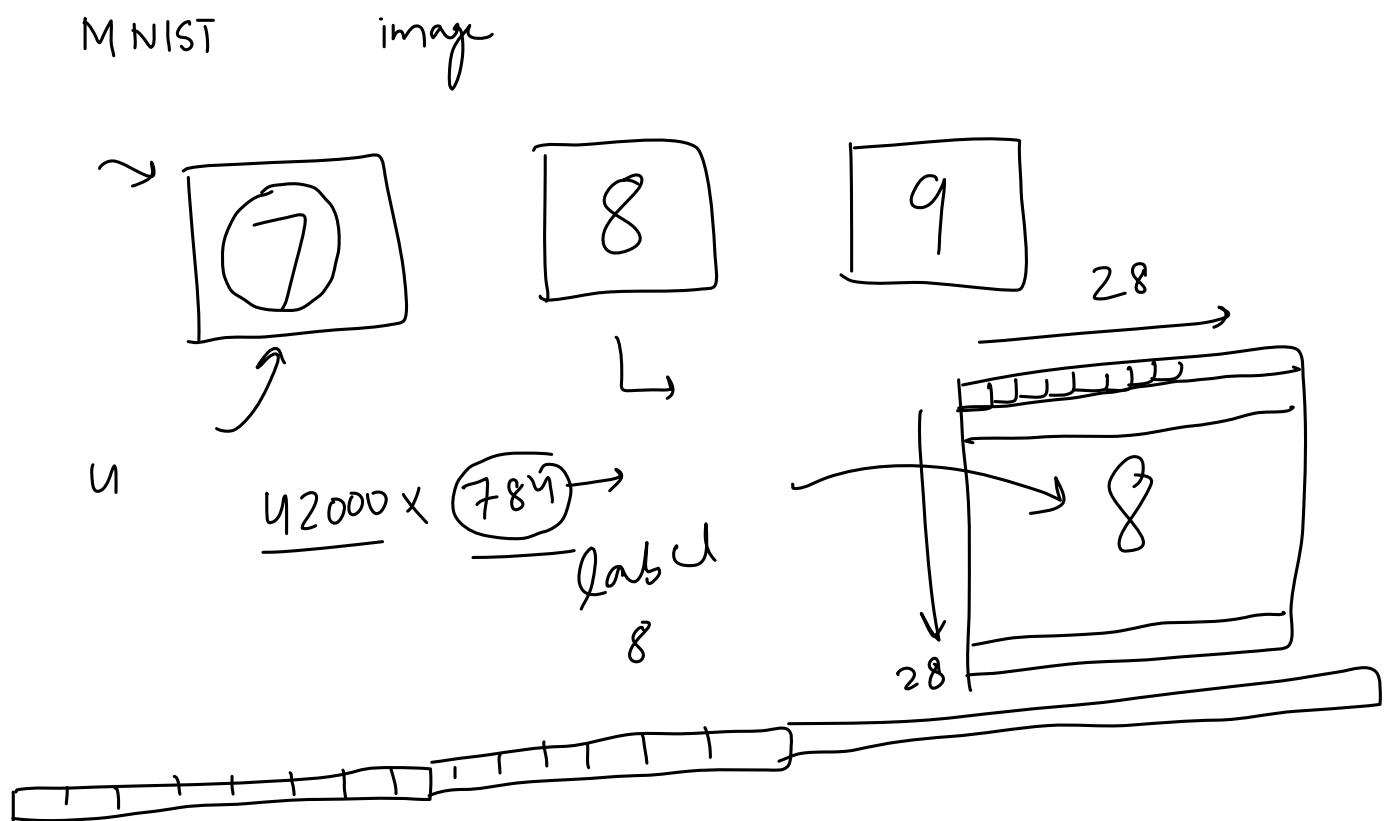
$$(1000, 3) \cdot (3, 2) = (1000, 2)$$

## 6. Coding the Steps

Wednesday, May 5, 2021 1:03 PM

## 7. Practical Example on MNIST Dataset

Wednesday, May 5, 2021 1:04 PM

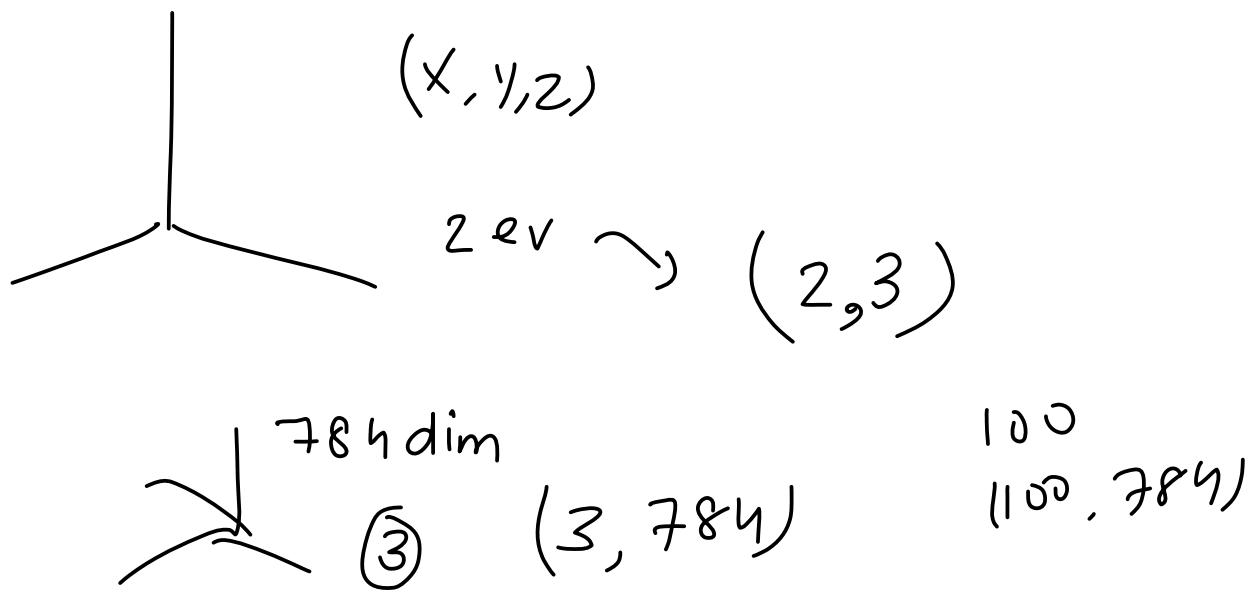


## 8. Visualization of MNIST Dataset

Wednesday, May 5, 2021 1:04 PM

## 9. Explained Variance

Wednesday, May 5, 2021 1:04 PM



## 10. Finding optimum number of Principle Components

Wednesday, May 5, 2021 1:05 PM

$$\text{EV} = \lambda$$

(784)  $\rightarrow \lambda_1, \lambda_2, \lambda_3, \dots, \lambda_{784}$

$\left( \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \dots + \lambda_{784}} \right) \times 100 \rightarrow \text{percentage}$

$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \dots, \lambda_{15}$

$\underline{30}, \underline{25}, \underline{15}, \underline{10}, \underline{5}, \dots, \lambda_{15}$

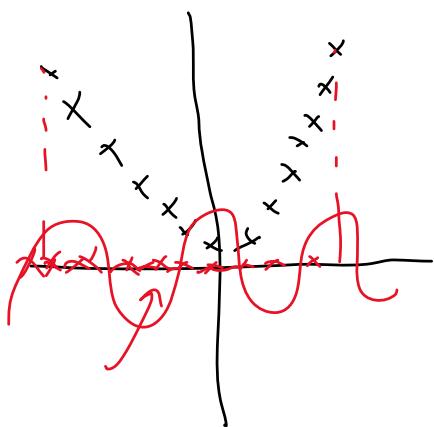
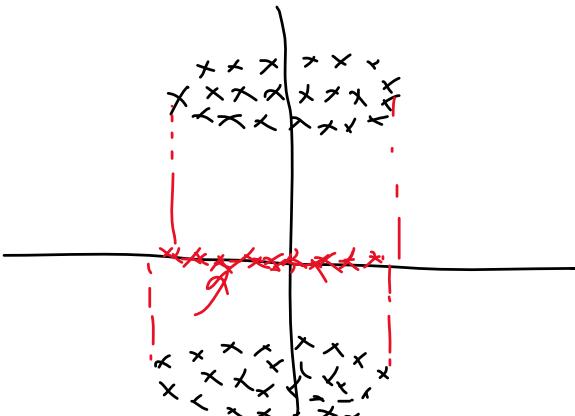
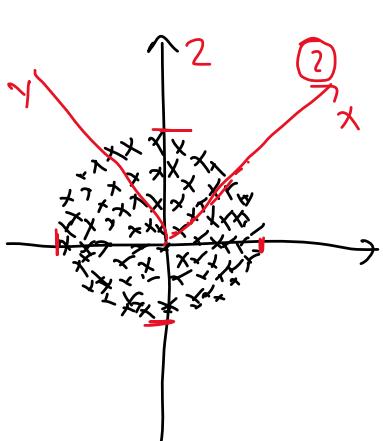
90%  $\rightarrow \lambda_1 - \lambda_{15}$

15

## 11. When PCA does not work

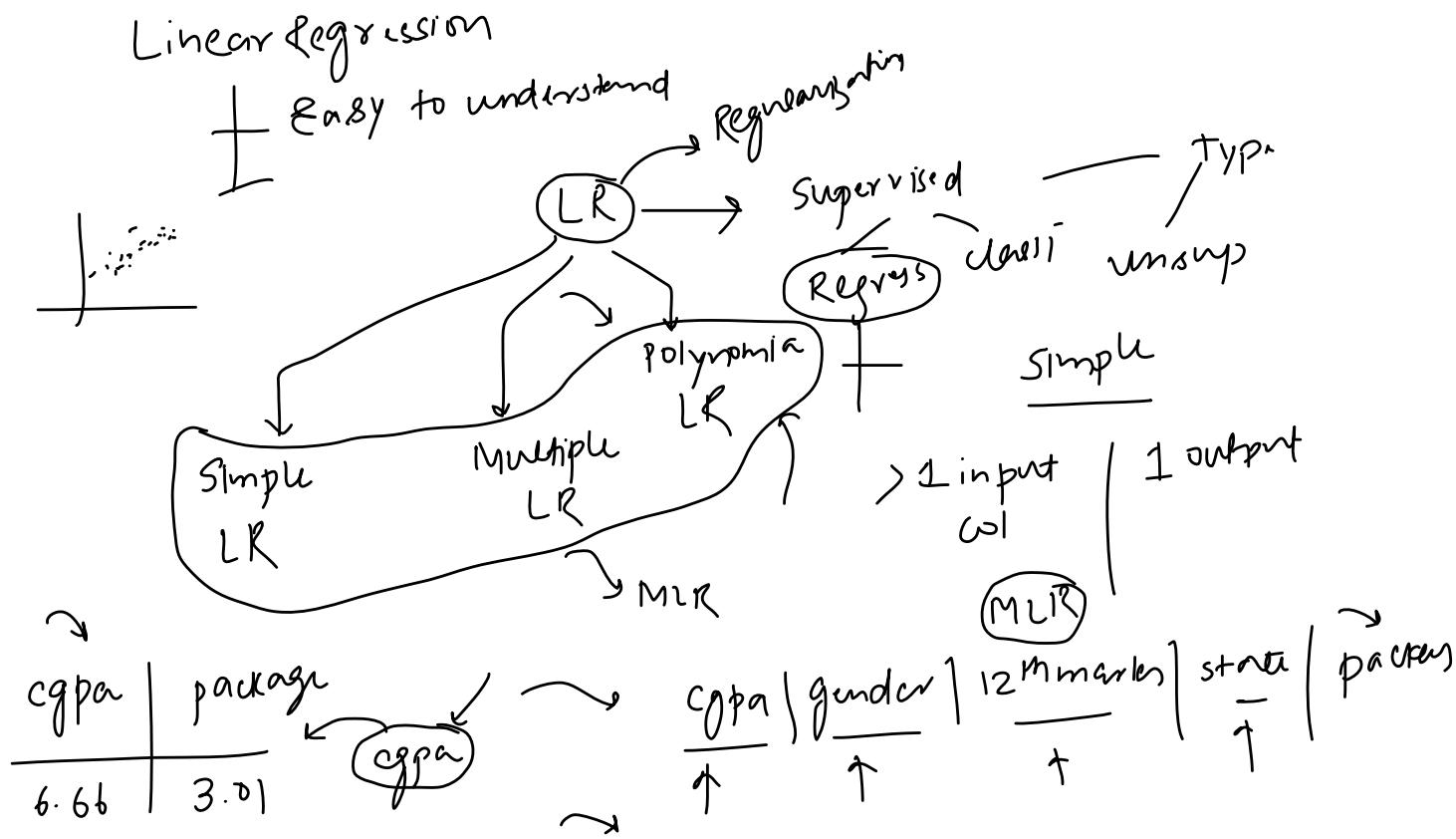
Thursday, May 6, 2021 7:03 AM

$$Y = X^2$$



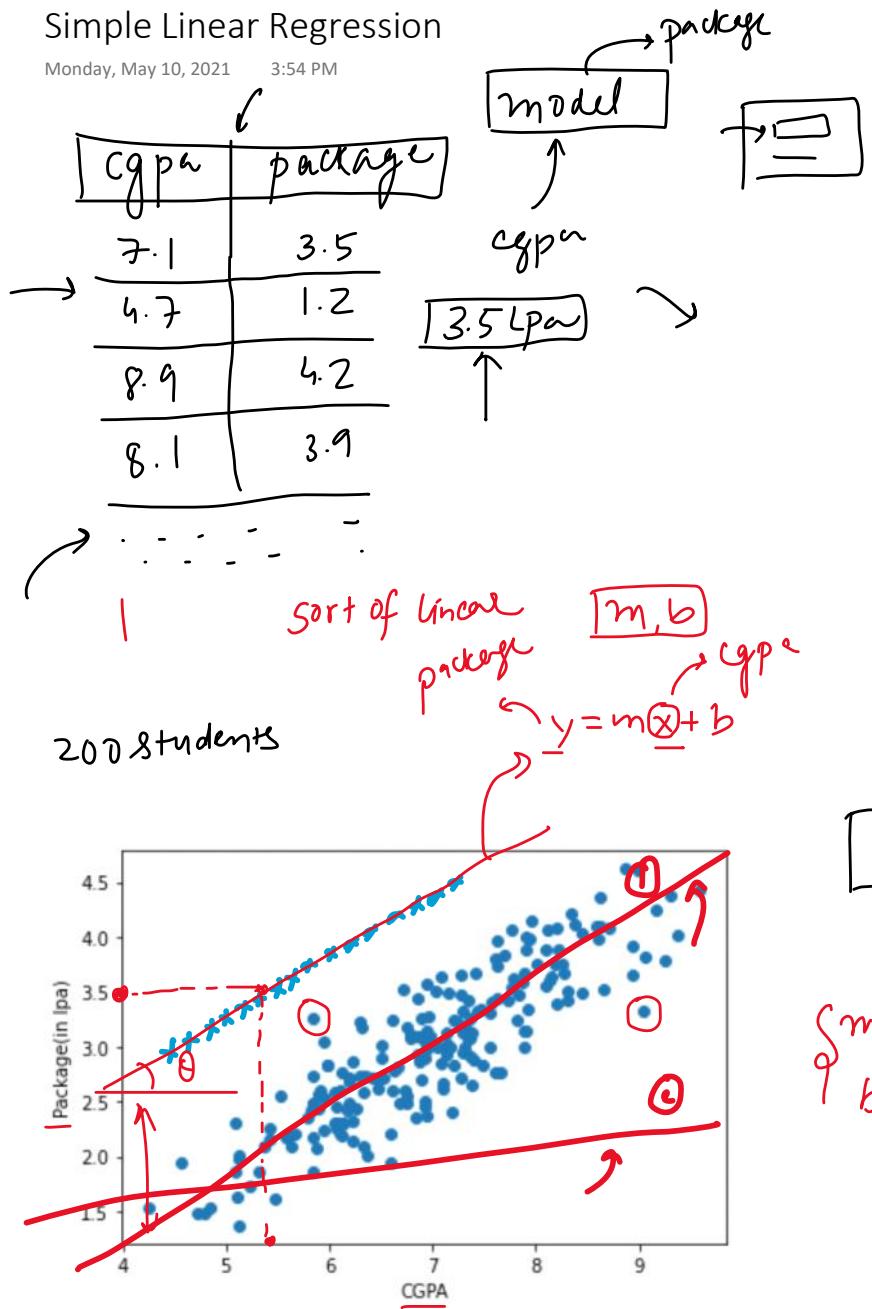
# Introduction

Monday, May 10, 2021 3:53 PM



# Simple Linear Regression

Monday, May 10, 2021 3:54 PM

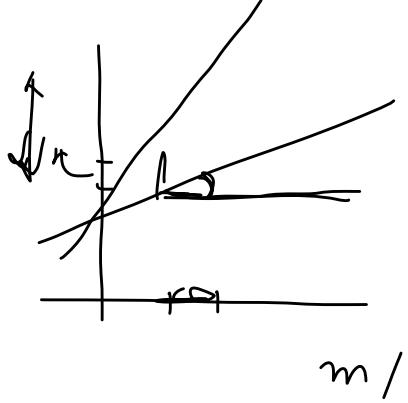


# Code Example

Monday, May 10, 2021 3:54 PM

## Intuition

Monday, May 10, 2021 3:54 PM



$$m/b$$

$m \rightarrow \text{weight}$

$$\begin{matrix} m \\ b \end{matrix} \uparrow$$

$$y = mx + b$$

package =  $mx \times \text{cgp}^a + b$

$b = 0$

$$\text{package} = mx \times \text{cgp}^a \quad \text{exp} \rightarrow 0$$

package =  $mx \times \text{exp}$  +  $b = 0$

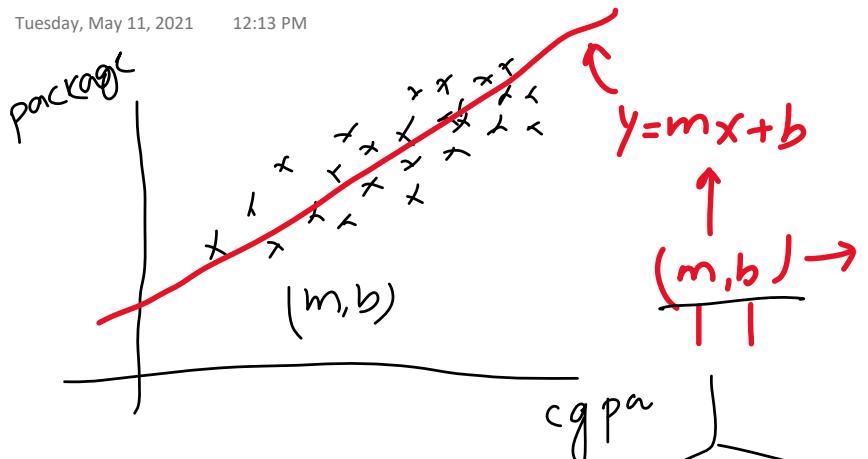
$p = mx \times \text{exp}$

$b = 0$

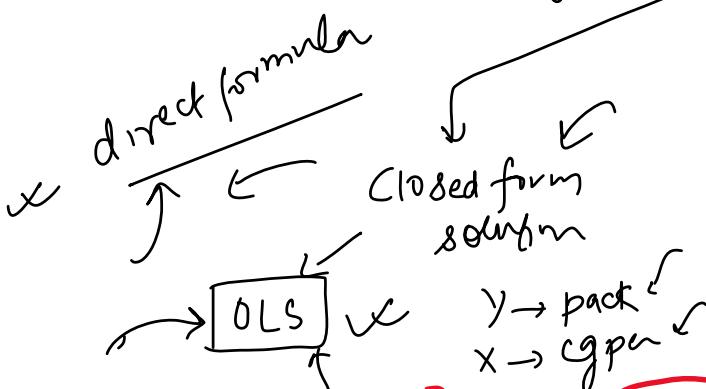
exp | package offset

## How to find m and b?

Tuesday, May 11, 2021 12:13 PM



Higher  $\rightarrow$  Linear  
SD Regressor

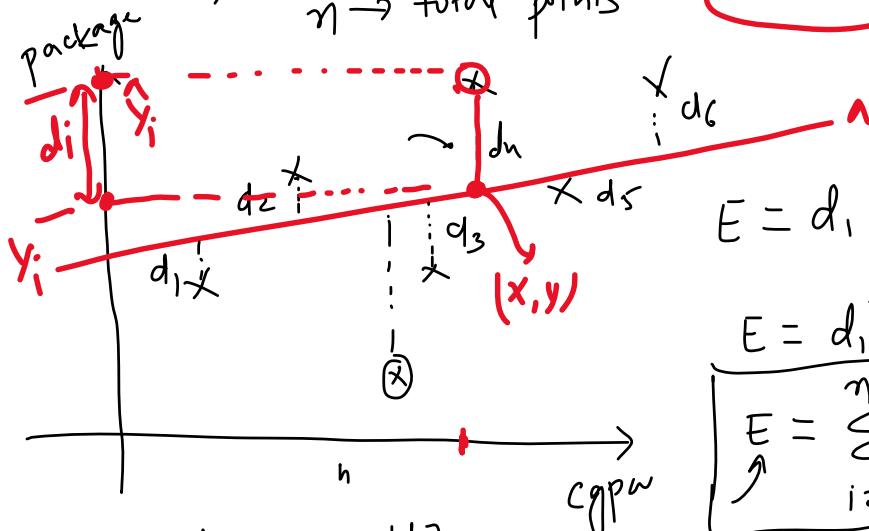


Non-closed form  
Gradient Descent

$$b = \bar{y} - m\bar{x}$$

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\bar{x}$  mean  $(m, b)$   
 $\bar{y}$   
 $n \rightarrow$  total points



$$E = d_1 + d_2 + d_3 + \dots + d_n$$

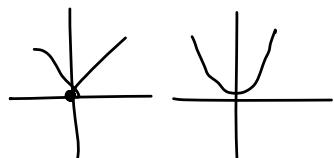
$$E = d_1^2 + d_2^2 + d_3^2 + \dots + d_n^2$$

$$E = \sum_{i=1}^n d_i^2$$

Error function  $\textcircled{J}$

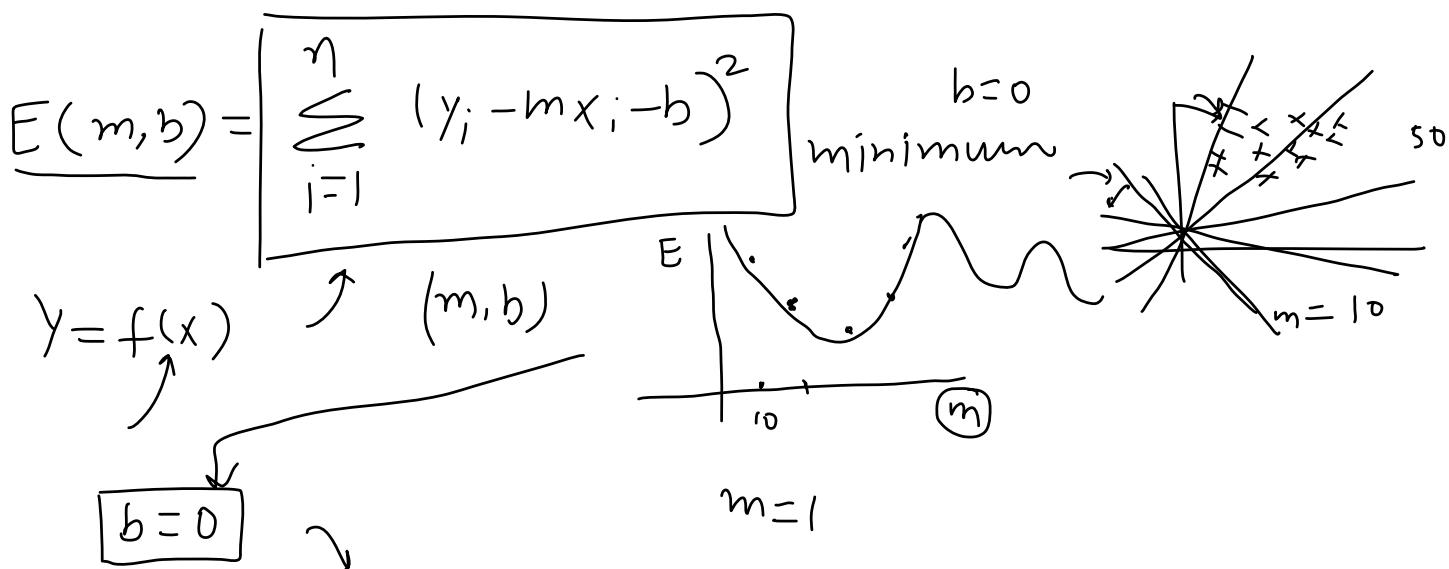
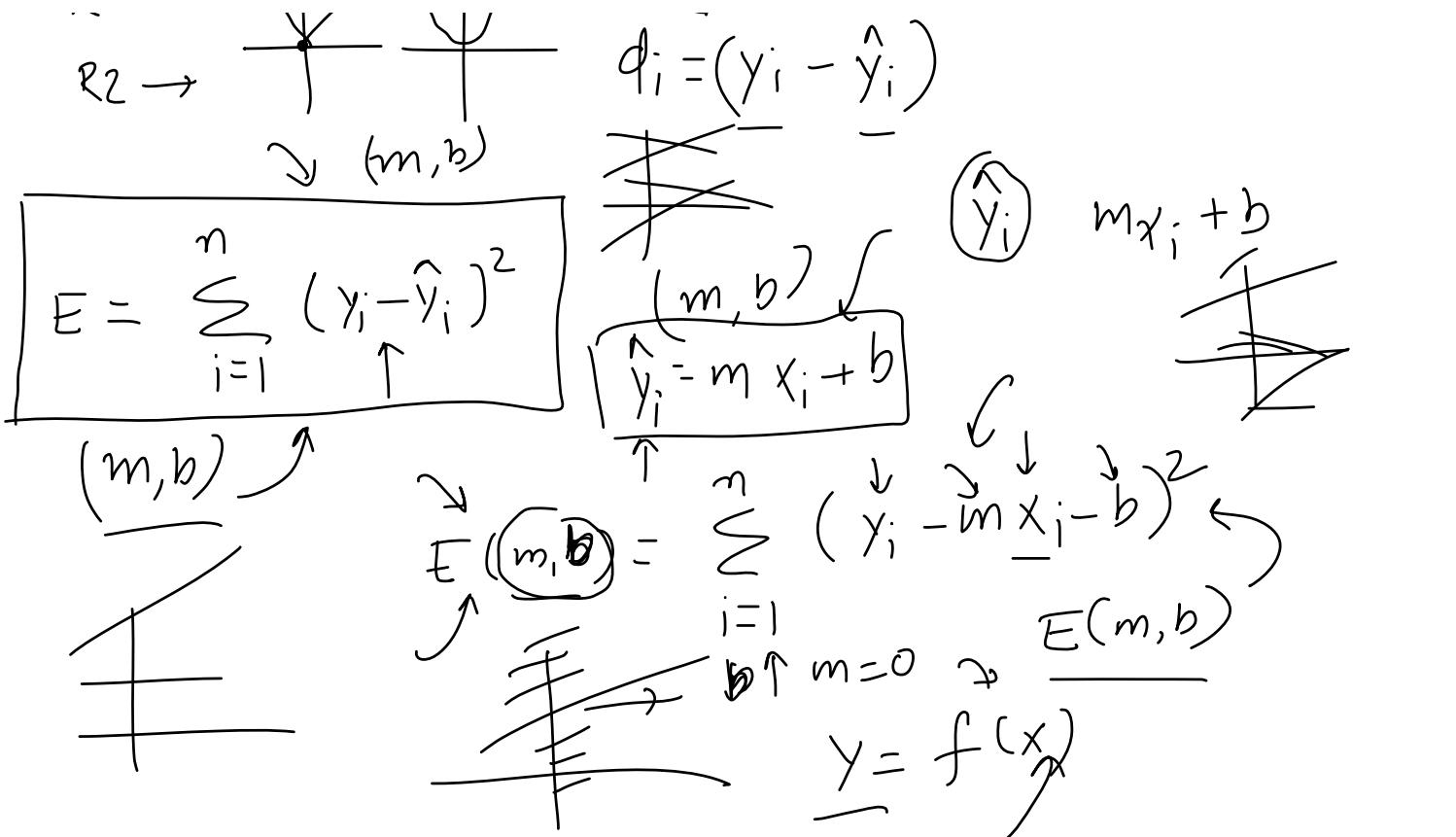
$$E = |d_1| + |d_2| + |d_3| + \dots$$

R1

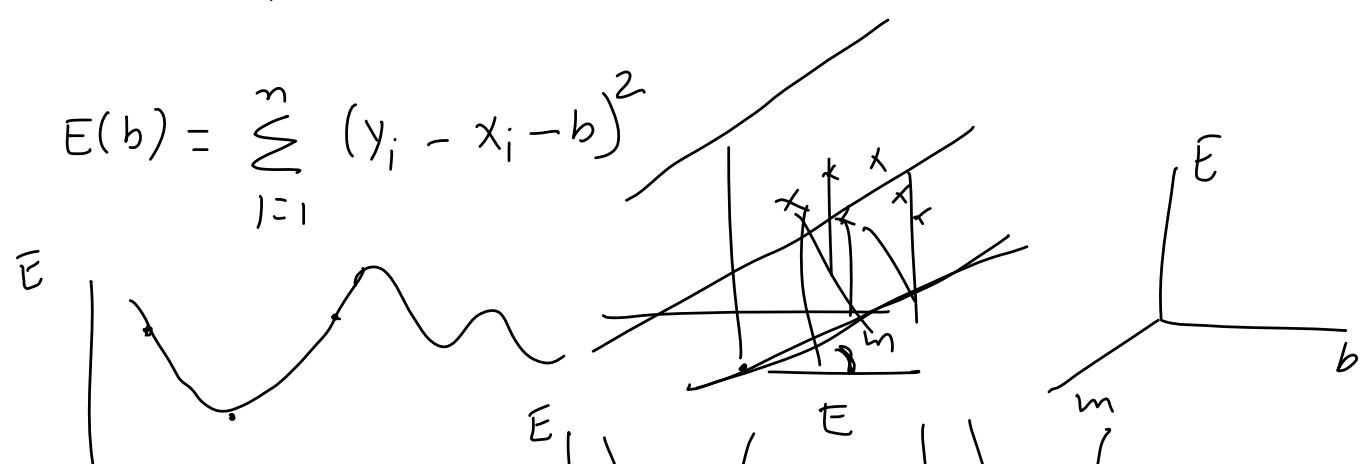


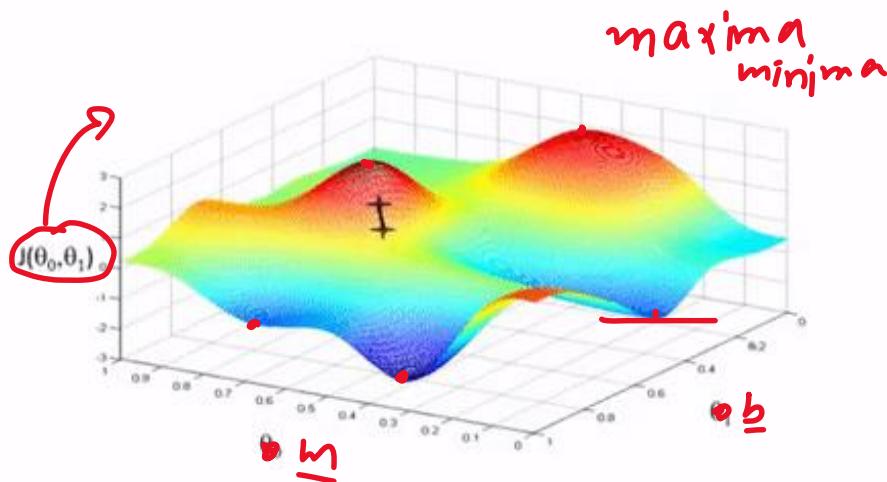
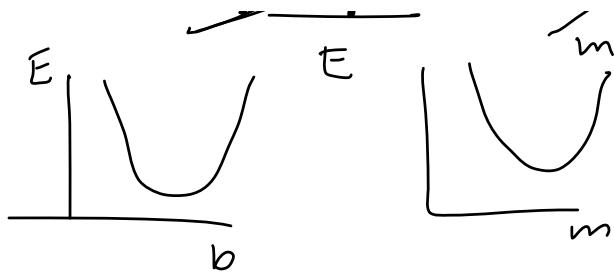
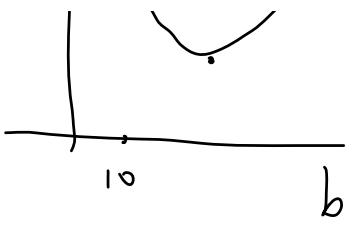
R2  $\rightarrow$

$$d_i = (y_i - \hat{y}_i)$$



$$E(m) = \sum_{i=1}^n (y_i - mx_i)^2$$





$\rightarrow E(x)$

$$\frac{\partial E}{\partial x} = 0$$

$$f(x, y)$$

$$\frac{\partial E}{\partial m} = 0, \frac{\partial E}{\partial b} = 0$$



Andrew Ng

$$(m, b)$$

$$\begin{aligned} \frac{\partial E}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - mx_i - b)^2 = 0 \\ &= \sum \frac{\partial}{\partial b} \left( \frac{y_i - mx_i - b}{n} \right)^2 = 0 \end{aligned}$$

$$\frac{\sum y_i}{n}$$

$$\sum mx_i - \frac{\sum b}{n} = 0$$

$$\bar{y} - \bar{mx} - \frac{\bar{b}}{n} = 0$$

$$\bar{y} - \bar{mx} = b$$

$$\boxed{b = \bar{y} - \bar{mx}}$$

$$\frac{b + b + b + b + \dots + b}{n} = nb$$

$$E = \sum (y_i - mx_i - \bar{y} + \bar{mx})^2$$

$$\frac{\partial E}{\partial m} = \sum \frac{\partial}{\partial m} \left( \frac{y_i - mx_i - \bar{y} + \bar{mx}}{n} \right)^2 = 0$$

$$\Rightarrow \sum 2(y_i - mx_i - \bar{y} + \bar{mx}) \left( -x_i + \bar{x} \right) = 0$$

$$= \sum -2(y_i - mx_i - \bar{y} + m\bar{x})(x_i - \bar{x}) = 0$$

$$= \sum \underbrace{(y_i - mx_i - \bar{y} + m\bar{x})}_{\leftarrow} (x_i - \bar{x}) = b$$

$$= \sum \left[ (y_i - \bar{y}) - m(x_i - \bar{x}) \right] (x_i - \bar{x}) = 0$$

$$= \sum \left[ (y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2 \right] = 0$$

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

$$\boxed{m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

# Code from scratch

Tuesday, May 11, 2021 12:14 PM

# Regression Metrics

Thursday, May 13, 2021 11:56 AM

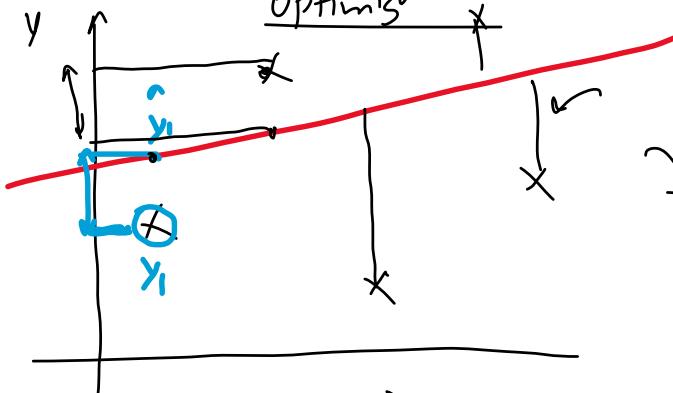
- 1) MAE
- 2) MSE
- 3) RMSE
- 4) R2 score
- 5) Adjusted R2 score

## MAE

Thursday, May 13, 2021

11:56 AM

$$\text{Optimization} \times \frac{\text{cgp} \times \text{pack (kpa)}}{=}$$



$$|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \dots + |y_n - \hat{y}_n|$$

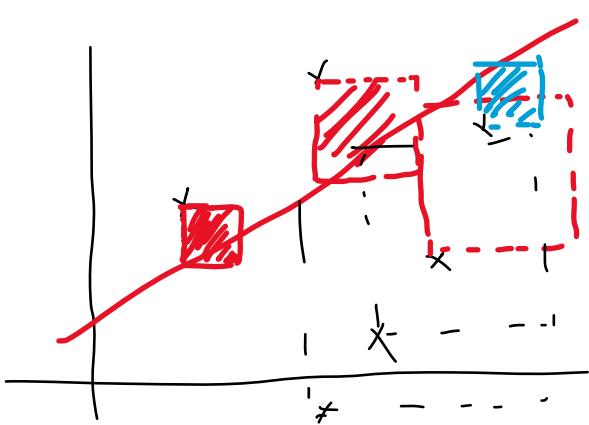
$$\boxed{mae = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}}$$

- Advantage
- 1) same unit
  - 2) Robust outliers
- Disadvantage
-

# MSE

Thursday, May 13, 2021 11:56 AM

mean squared error  $\text{mse} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$



$$\text{mse} = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

mse

$$\text{mse} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

$L_2$  function

$$(y_i - \hat{y}_i)^2$$

Advantage  
f<sup>2</sup> function

Disadvantage  
Robust  
to outliers

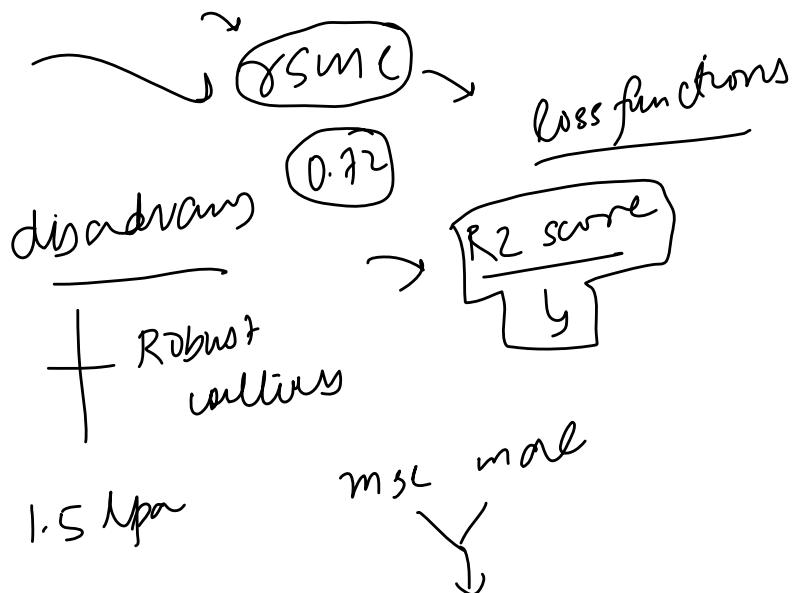
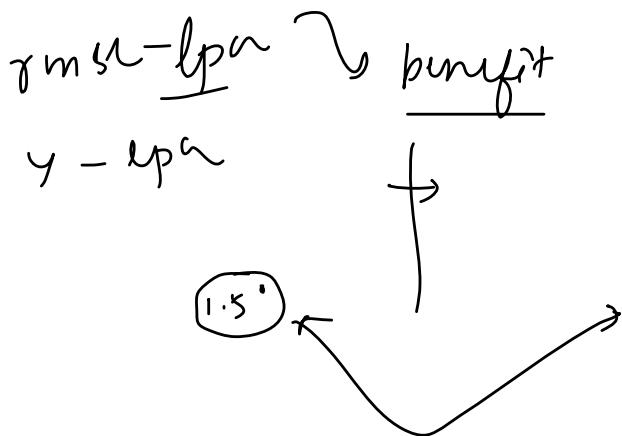
$$y - \hat{y}_i$$

$$\text{mse} = (\hat{y}_i)^2$$

## RMSE

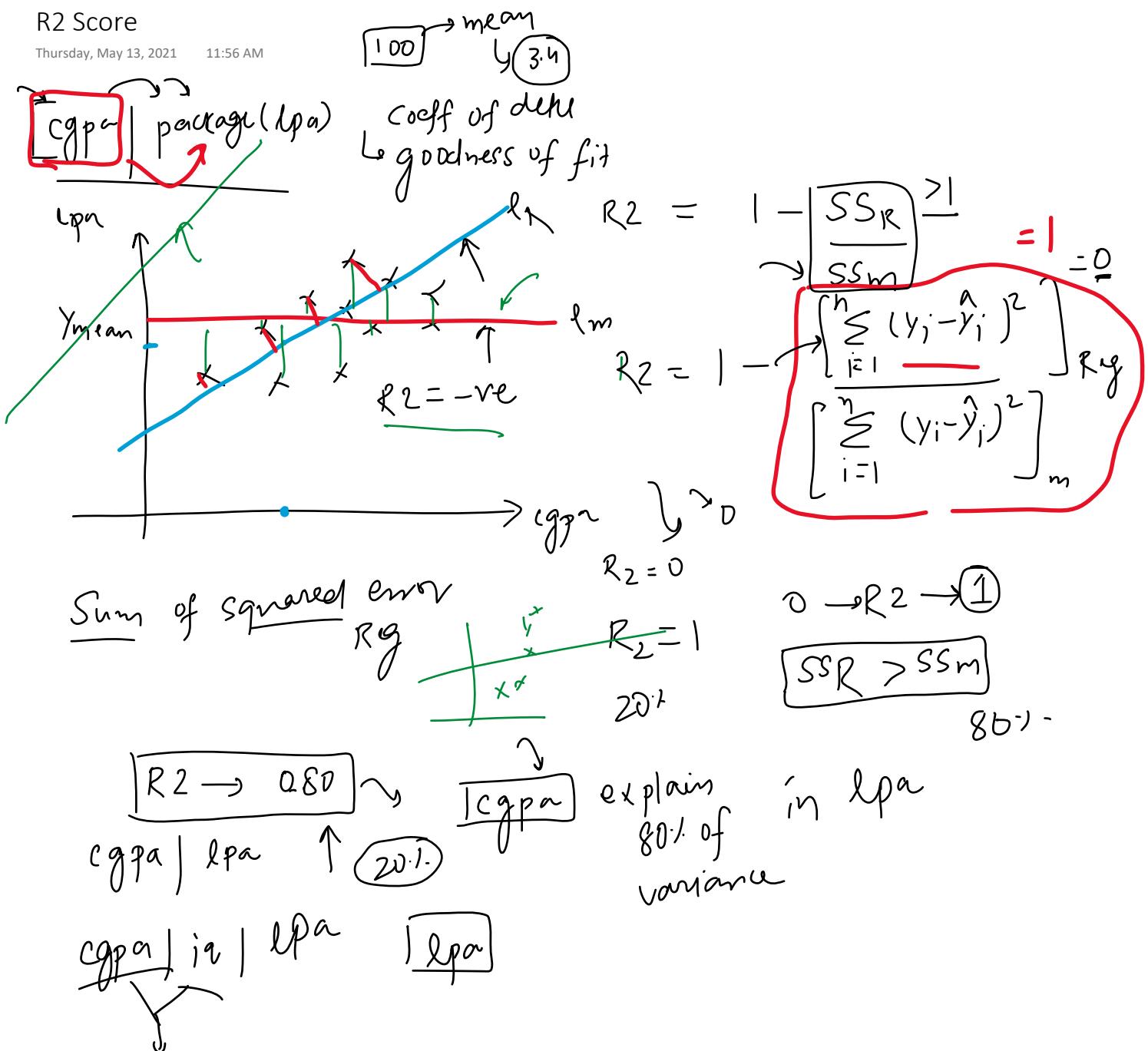
Thursday, May 13, 2021 11:56 AM

$$RMSE = \sqrt{MSE}$$
$$= \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$



## R2 Score

Thursday, May 13, 2021 11:56 AM



## Adjusted R2 score

Thursday, May 13, 2021 11:57 AM

Adjusted R<sup>2</sup> score

Thursday, May 13, 2021 11:57 AM

$R^2$  score

$R^2 \uparrow -$

$\boxed{R^2_{adj}} = 1 - \frac{(1-R^2) \frac{(n-1)}{(n-1-K)}}{(n-1-K)}$

$\boxed{\text{Multiple LR}}$

$R^2 \rightarrow$

$n \rightarrow \text{no. of rows}$

$K = \text{independent}$

$K=1, K=2, K=3$

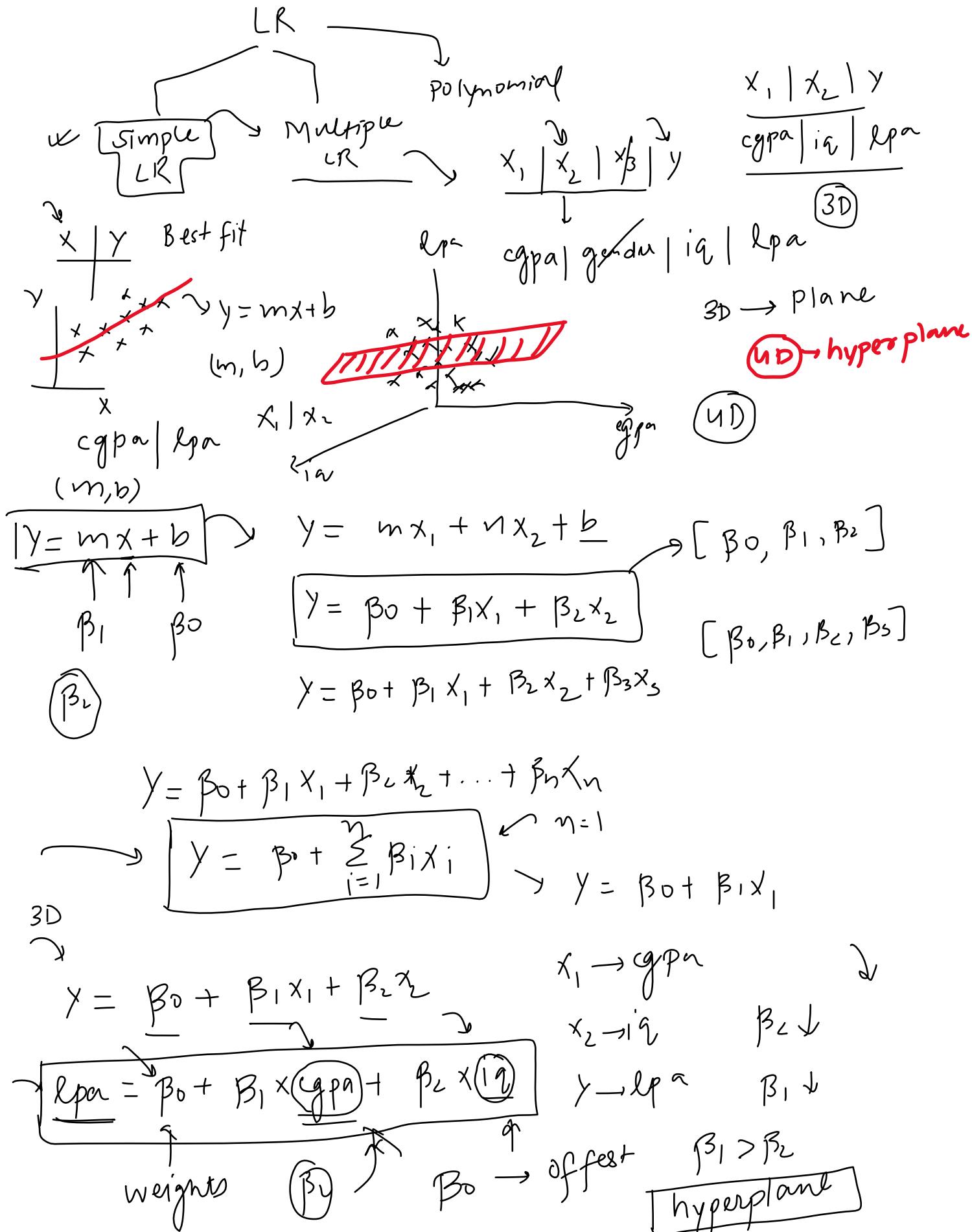
$\boxed{\text{temp}}$

$\boxed{R^2_{adj} \uparrow}$

$\boxed{\text{Adj. } R^2 \downarrow}$

# Multiple Linear Regression

Friday, May 14, 2021 4:31 PM

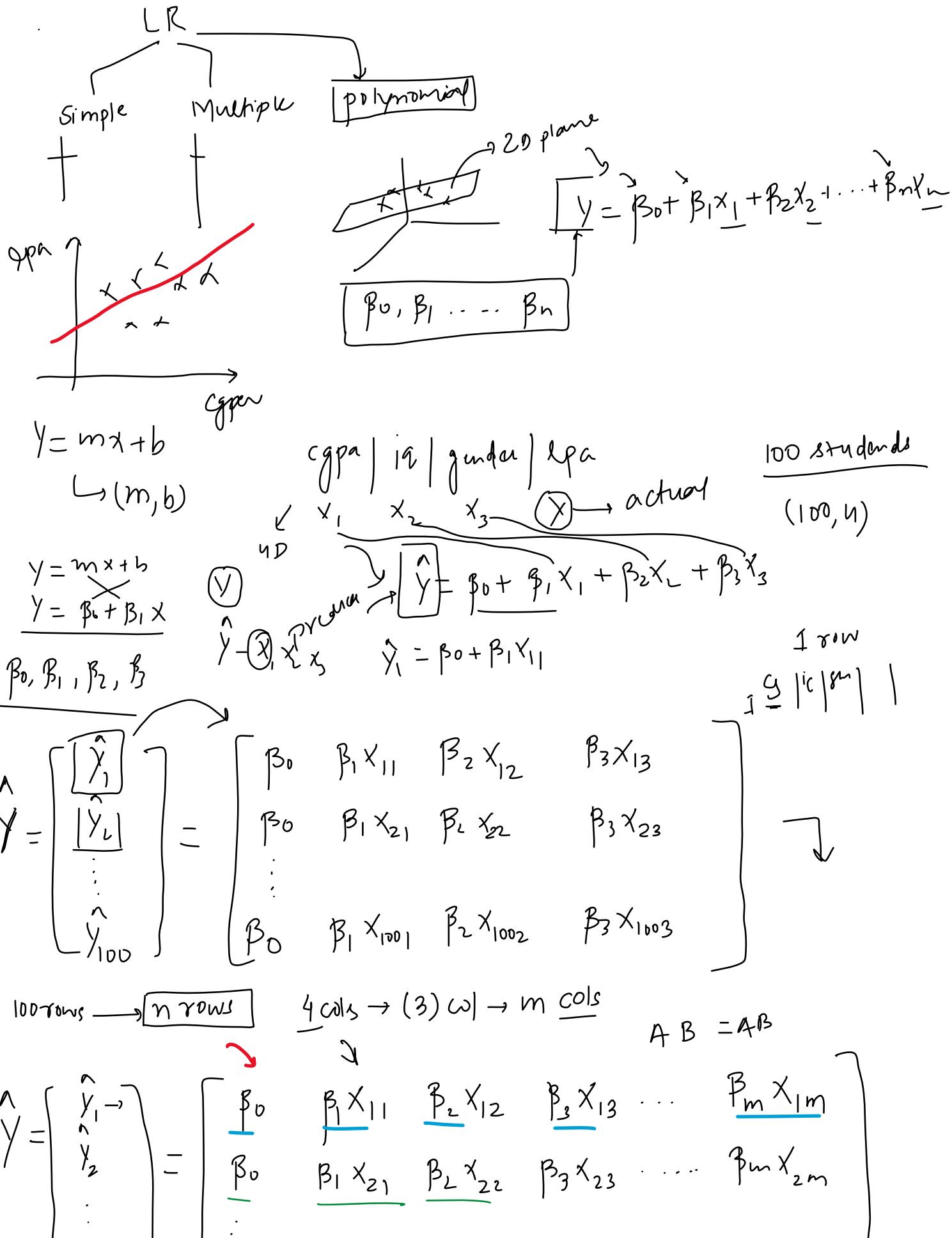


# Code Example

Friday, May 14, 2021 4:31 PM

# Mathematical Formulation

Saturday, May 15, 2021 4:02 PM



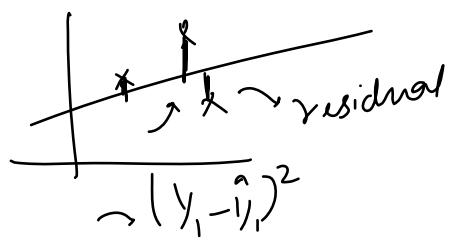
$$\begin{aligned}
 & \left[ \begin{array}{c} \vdots \\ x_n \end{array} \right] \left[ \begin{array}{c} 1 \\ \vdots \\ \beta_0 \\ \beta_1 x_{n1} \\ \beta_2 x_{n2} \\ \beta_3 x_{n3} \\ \dots \\ \beta_m x_{nm} \end{array} \right] \\
 = & \left[ \begin{array}{c} 1 \quad x_{11} \quad x_{12} \quad x_{13} \quad \dots \quad x_{1m} \\ | \quad x_{21} \quad x_{22} \quad x_{23} \quad \dots \quad x_{2m} \\ \vdots \\ 1 \quad x_{n1} \quad x_{n2} \quad x_{n3} \quad \dots \quad x_{nm} \end{array} \right] \left[ \begin{array}{c} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{array} \right] \\
 & \boxed{Y = X\beta} - ① \quad \boxed{X} \quad \boxed{\beta} \quad \boxed{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}} \quad \boxed{\begin{bmatrix} b \\ m \end{bmatrix}} \quad \boxed{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}
 \end{aligned}$$

cgp | gndu | iq  
 x<sub>1</sub> x<sub>2</sub> x<sub>3</sub>  
 |-----|-----|-----|  
x<sub>11</sub>	x<sub>12</sub>	x<sub>13</sub>
x<sub>21</sub>	x<sub>22</sub>	x<sub>23</sub>
-----	-----	-----
 x<sub>31</sub> | x<sub>32</sub> | x<sub>33</sub>

df → X(Y)

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad e = y - \hat{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$$C \left[ \begin{array}{c} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{array} \right]$$



Single LR

1

17

Simple LR

$$E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad E = \underline{e}^T \underline{e}$$

$$\begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \underline{e}$$

$$\begin{bmatrix} (y_1 - \hat{y}_1) & (y_2 - \hat{y}_2) & (y_3 - \hat{y}_3) & \dots & (y_n - \hat{y}_n) \end{bmatrix} \begin{bmatrix} (y_1 - \hat{y}_1) \\ (y_2 - \hat{y}_2) \\ (y_3 - \hat{y}_3) \\ \vdots \\ (y_n - \hat{y}_n) \end{bmatrix} \rightarrow \textcircled{1} \times 1$$

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \dots + (y_n - \hat{y}_n)^2 \quad (\underline{n} \times 1)$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (A + B)^T = A^T + B^T$$

$$(A - B)^T = A^T - B^T$$

$$E = \underline{e}^T \underline{e} = (y - \hat{y})^T (y - \hat{y})$$

$$= (y^T - \hat{y}^T)(y - \hat{y})$$

$$= [y^T - (X\beta)^T] (y - X\beta)$$

$$E = y^T y - \underbrace{y^T X \beta}_{\text{matrix diff}} - (X\beta)^T y + (X\beta)^T X \beta$$

$$E = y^T y - 2y^T X \beta + \beta^T X^T X \beta$$

matrix  
diff

$$\frac{dE}{d\beta} = \frac{d}{d\beta} \int [y^T y - 2y^T X \beta + \beta^T X^T X \beta] = 0$$

$$\frac{\partial \hat{y}}{\partial \beta} = \frac{y}{\beta} \left[ \frac{y^T y - 2 \underline{y^T X \beta}}{\beta^T X^T X \beta} + \beta' \wedge X P \right] - u$$

$$= 0 - 2 y^T X + \frac{d}{d \beta} \left[ \frac{\beta^T \underline{X^T X \beta}}{\beta^T X^T X \beta} \right] = 0$$

$\begin{array}{l} \frac{dy}{d\beta} = 2 \underline{X^T X} \\ y = \frac{A^T X A}{\uparrow} \end{array}$

$$= -2 y^T X + 2 X^T X \beta^T = 0$$

$$= \cancel{2 X^T X \beta^T} = \cancel{2 y^T X} \quad |$$

$\frac{1}{[X^T X]}$

$$\beta^T = y^T X \underline{(X^T X)^{-1}}$$

$$(\beta^T)^T = \underline{[y^T X \underline{(X^T X)^{-1}}]}^T$$

$$\beta = [(X^T X)^{-1}]^T \underline{(y^T X)}^T$$

$$\beta = \underline{[(X^T X)^{-1}]}^T X^T Y$$

$$\beta = \underline{(X^T X)^{-1}} X^T Y$$

$$\boxed{\beta = (X^T X)^{-1} X^T Y}$$

$$\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix}$$

$(m+1) \times 1$

$$[(m+1) \times (m+1)] [(m+1) \times n] [n \times 1]$$

$\Gamma.$   $\Gamma \Gamma$   $\Gamma \Gamma$

$$\boxed{X^T X} \quad m \rightarrow \text{cols}$$

$\Gamma - 1$

$$X = n \times (m+1)$$

$$(m+1) \times n \quad n \times (m+1)$$

$$(m+1) \times (m+1)$$

$$X \rightarrow 1 \quad X_{\text{train}}$$

$$Y = Y_{\text{train}}$$

$$(X^T X)^{-1}$$

$$n \times (m+1)$$

$$(m+1) \times n$$

$$\begin{array}{l}
 \beta_m = \\
 \underline{(m+1) \times 1} \\
 \uparrow \\
 \beta
 \end{array}
 \quad
 \begin{array}{c}
 \left[ \begin{matrix} (m+1) \times n \\ | \quad (n \times 1) \end{matrix} \right] \\
 \xrightarrow{\text{999}} \\
 \boxed{1 \times (m+1) \times 1} \\
 \xrightarrow{n=1000} \\
 \boxed{1000000000}
 \end{array}
 \quad
 \begin{array}{c}
 \boxed{X^T X} \xrightarrow{m \rightarrow \infty} \\
 (m+1) \times (m+1) \\
 \boxed{1000 \times 1000} \\
 \xrightarrow{n^3}
 \end{array}
 \quad
 \begin{array}{c}
 \left[ \begin{matrix} - \\ - \\ - \end{matrix} \right] \\
 n \times 1
 \end{array}$$

$$(AB)^T = \underline{B^T A^T} \quad Y = A \quad X\beta = B \quad A^T = A$$

$$\boxed{Y^T X\beta} = \underline{(X\beta)^T Y} \quad (A^T B)^T = \underline{\frac{B^T A}{RHS}} - \textcircled{11}$$

$$A^T B = \underline{\frac{B^T A}{RHS}} - \textcircled{1}$$

$$\overrightarrow{A^T B} = \underline{(A+B)^T} \rightarrow \overrightarrow{A^T B} = C$$

$$\boxed{C = C^T} \quad A^T B \quad C = \boxed{Y^T X\beta}$$

 $n$  students

$$Y = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

$$(n \times 1) \rightarrow (1 \times h)$$

$$\overrightarrow{(Y^T X\beta)^T} = Y^T X\beta$$

$$\overrightarrow{(1 \times n)} \quad \overrightarrow{(n \times (m+1))} \quad \overrightarrow{((m+1) \times 1)}$$

$$1 \times (m+1) \quad (m+1) \times 1$$

$$1 \times 1 = [?]^T$$

$$n \times (m+1)$$

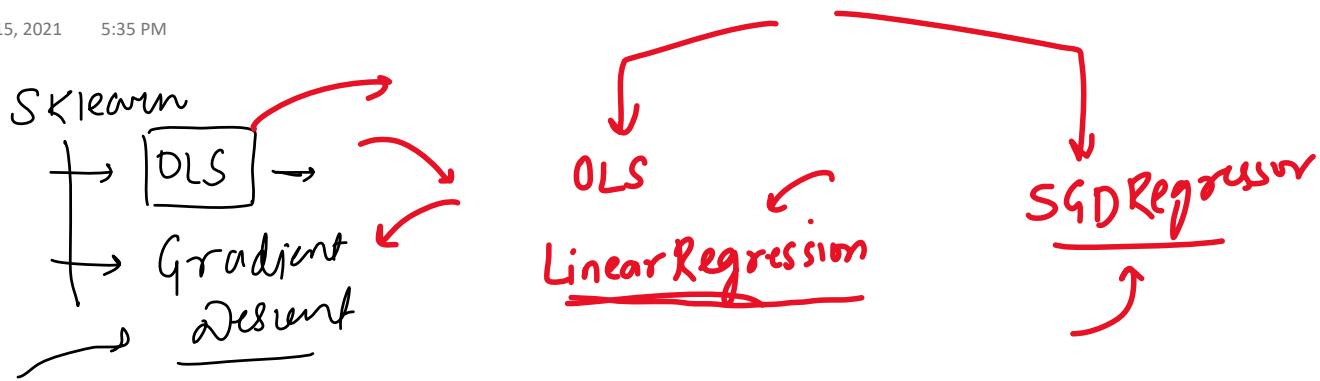
$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}^{m+1}$$

$$(m+1) \times 1$$

$$[?]$$

## Why Gradient Descent

Saturday, May 15, 2021 5:35 PM



## Code From Scratch

Monday, May 17, 2021 12:15 PM

$$\beta = (\underline{X}^T X)^{-1} X^T Y$$

$(100, 3)$      $(100)$   
 $\downarrow$   
 $100 \times (100, 1)$

$X \rightarrow \text{matrix}$

	cgpa	iq	gender	lpa
1	-	-	-	=
1	-	-	-	=

$X \rightarrow X_{\text{train}}$

$Y \rightarrow y_{\text{train}}$      $\xrightarrow{\text{diabetes}} \text{sklearn} \rightarrow R^2 - \text{same}$

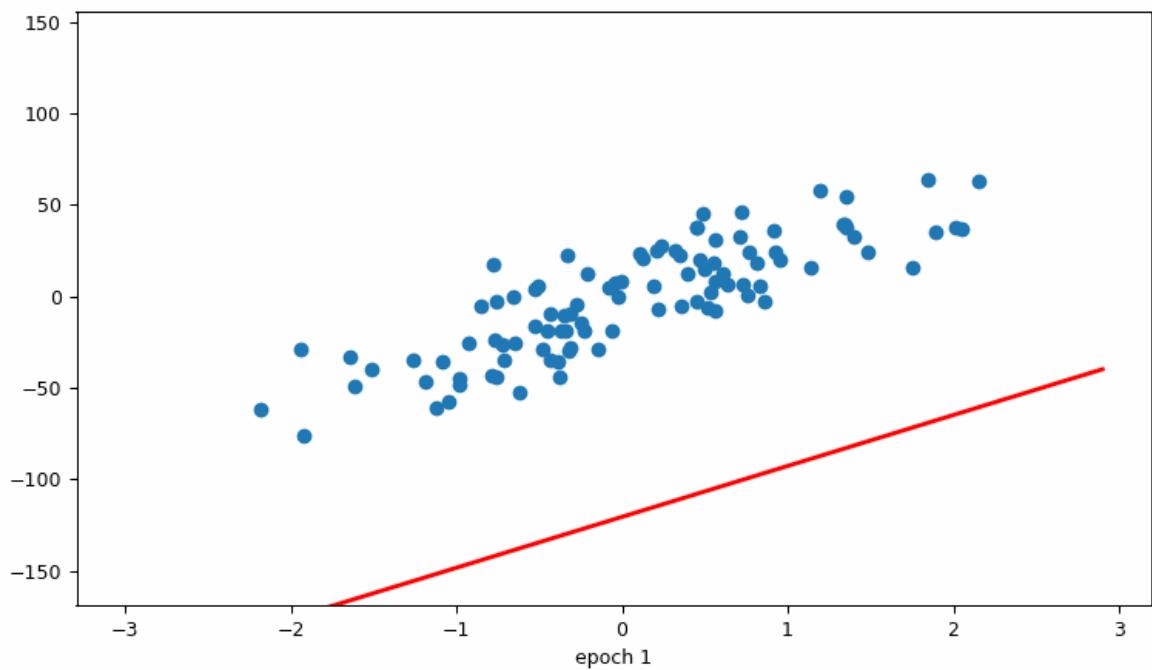
$$\underline{X_{\text{test}}} \quad \beta_0 \quad \beta_1 \rightarrow \beta_0$$

$\downarrow 3$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$= \beta_0 + \underbrace{\text{np.dot}(\beta, X_{\text{test}})}_{\text{in}}$$

$X_{\text{test}}$     (coeff)  
 $\underline{(89, 10)}$      $(10, 1)$   
 $(89, 1) + \beta_0$   
 $89 \rightarrow ① \rightarrow (89, 1)$   
 $\boxed{y_{\text{pred}}}$



$$\lesssim (y_i - mx_i - b)^2$$

$$\lesssim -2 (y_i - mx_i - b) x_i$$

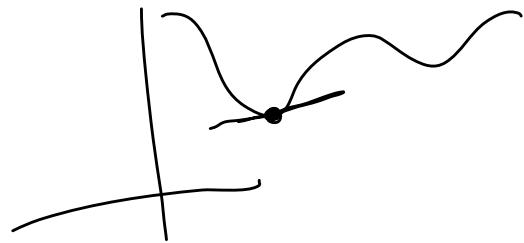
- 2

# What is Gradient Descent?

Thursday, May 20, 2021 1:46 PM

Linear Reg  
logistic Reg  
Tshu

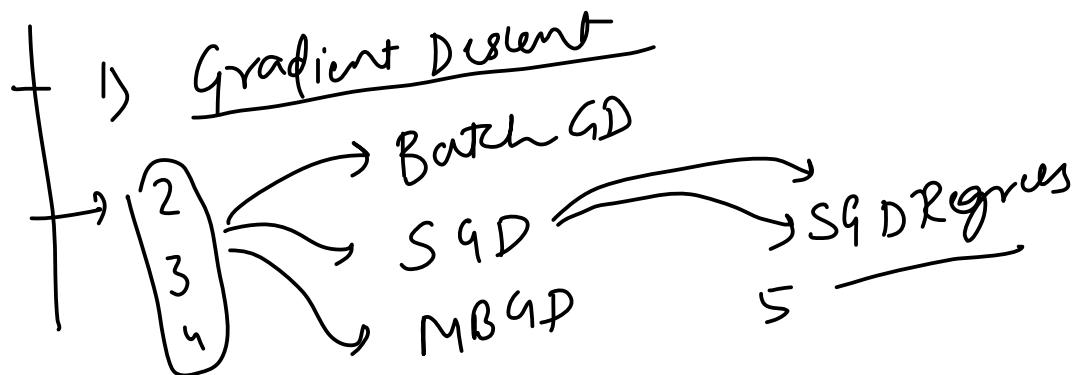
→ Deep Learning



# The Plan

Thursday, May 20, 2021 1:46 PM

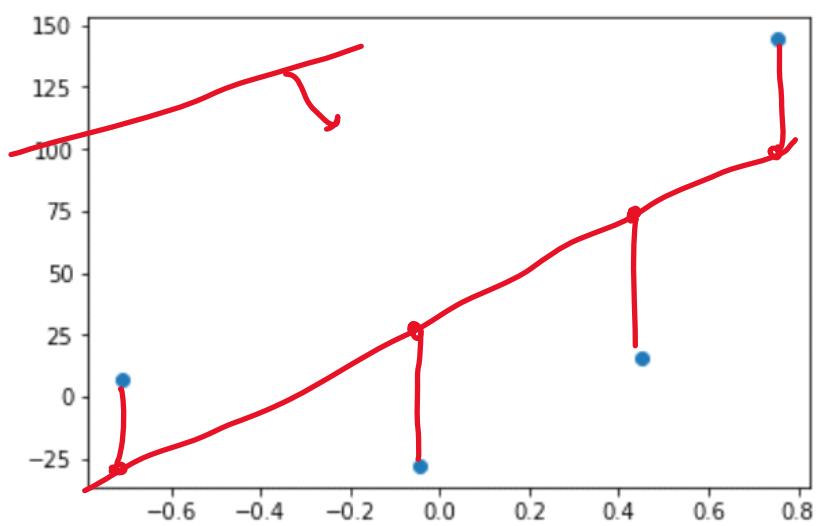
5 videos



# Intuition

Thursday, May 20, 2021 1:46 PM

2 cols  
4 rows



$$\text{cgpa} | \text{lpa}$$

$$\hat{y}_i = mx_i + b$$

$$m = 78.35$$

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$L = \sum_{i=1}^n (y_i - mx_i - b)^2$$

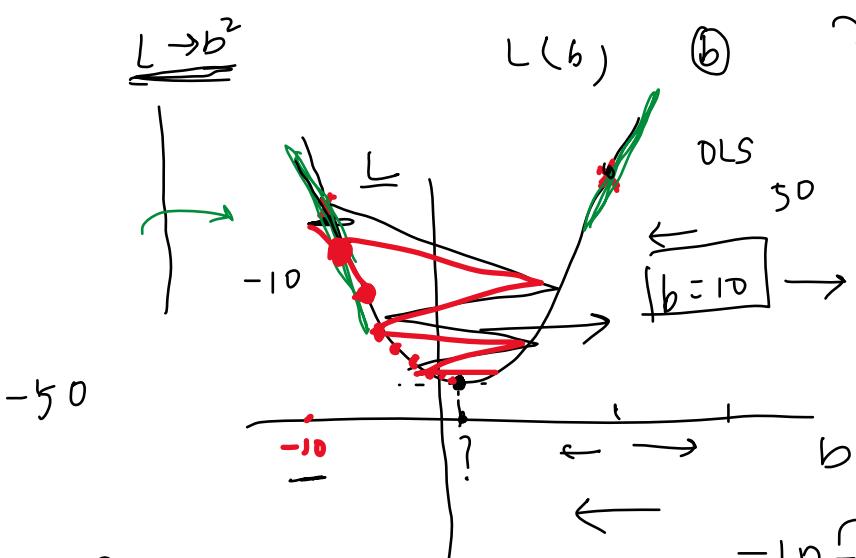
$$L \rightarrow b^2$$

$$L(b) \quad b$$

$$DLS \quad b = 10$$

$$L = \sum_{i=1}^n (y_i - 78.35 * x_i - b)^2$$

Step 1 - select a random  $b_{min}$   
 $b_{min}$   
 $b_{max}$



$$b = -10 \quad \text{Slope} = -ve$$

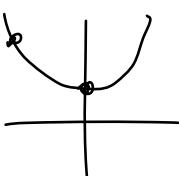
$$b_{new} = b_{old} - \eta \text{slope}$$

$$1 \quad b_{new} = b_{old} - \eta \text{slope}$$

$$b_{new} = -9.5 - (0.01 * -40) = -9.5 + 0.4 = -9.1$$

$$\text{When to stop}$$

$$| b_{new} - b_{old} | \Rightarrow 0.000$$



$$0.01, 0.0001$$

$$b_{new} = -10 + (0.01 * 50)$$

$$= -10 + 0.5 = -9.5$$

$$b_{new} - b_{old} = \frac{0.000}{\square}$$

$$\square$$

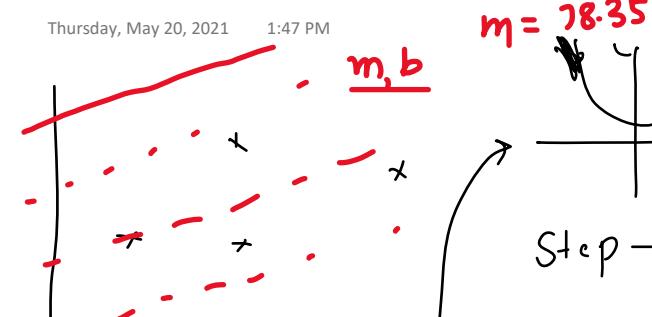
$\Rightarrow 0.000$

2) Iteration  $\rightarrow$  1000, 100,  
epochs

$$\boxed{b_{new} - b_{old} = 0}$$

## Mathematical Formulation

Thursday, May 20, 2021 1:47 PM



$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\frac{dL}{db} = \frac{d}{db} \left( \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) = \frac{d}{db} \sum_{i=1}^n (y_i - mx_i - b)^2$$

Slope =

Step → Start with a random  $b = b$   
for  $i$  in epochs:

$$b_{\text{new}} = \frac{b_{\text{old}}}{\uparrow} - \frac{\eta \times \text{slope}}{\downarrow} \quad (b = 0)$$

$$\begin{aligned} & \frac{d}{db} \sum_{i=1}^n (y_i - mx_i - b)^2 = 2 \sum_{i=1}^n (y_i - mx_i - b) (-1) \\ & \quad \boxed{-2 \sum_{i=1}^n (y_i - mx_i - b)} \quad b = 0 \\ & \quad = -2 \sum_{i=1}^n (y_i - 78.35x_i - 0) \end{aligned}$$

Epoch

$$b_{\text{new}} = b_{\text{old}} - \boxed{n \text{slope}_{b=0}}$$

i = 1

step SBC

# Example

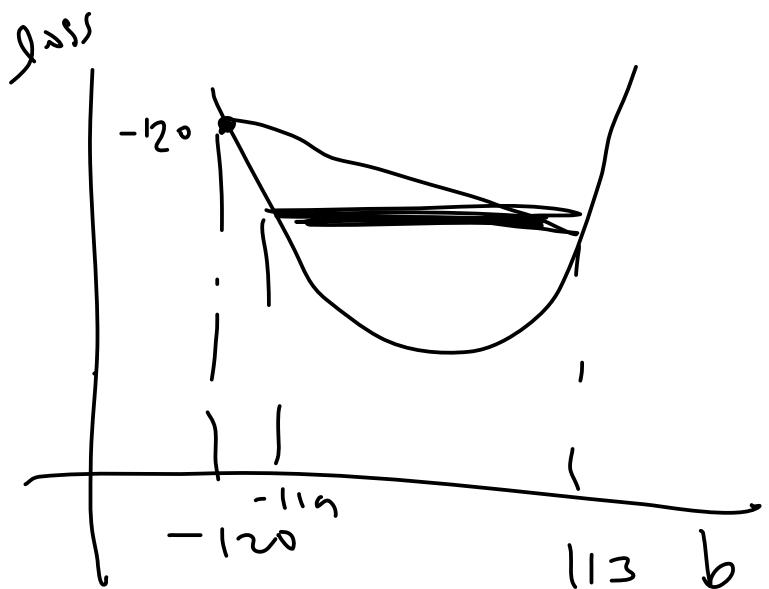
Thursday, May 20, 2021 1:47 PM

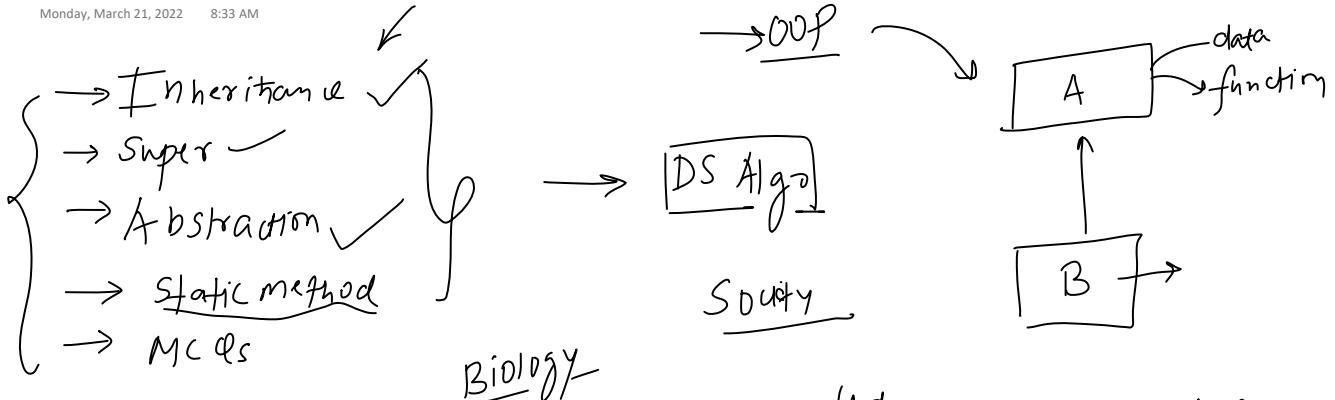
# Code from Scratch

Thursday, May 20, 2021 1:48 PM

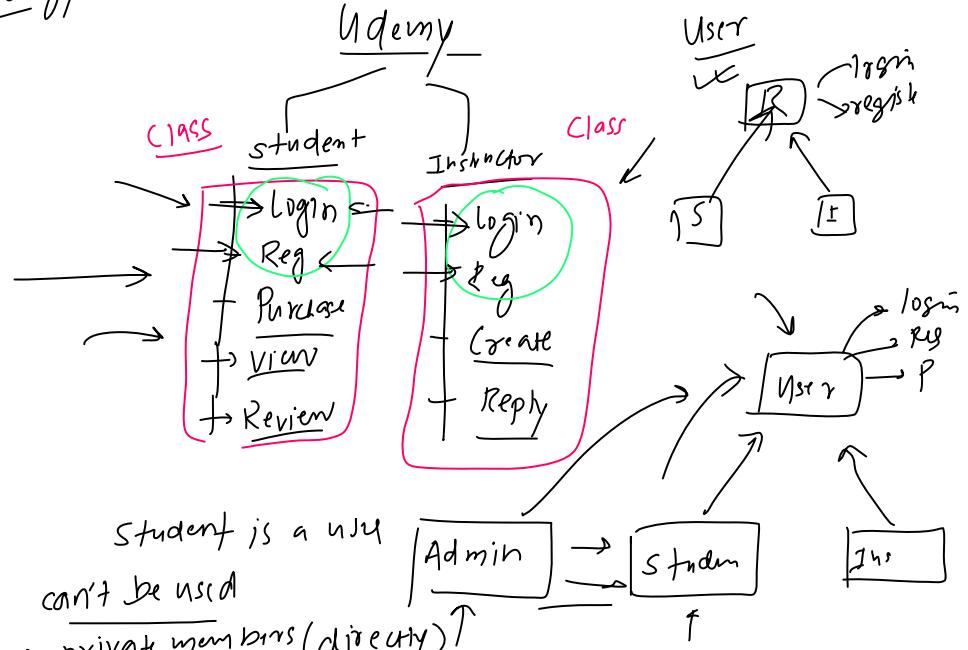
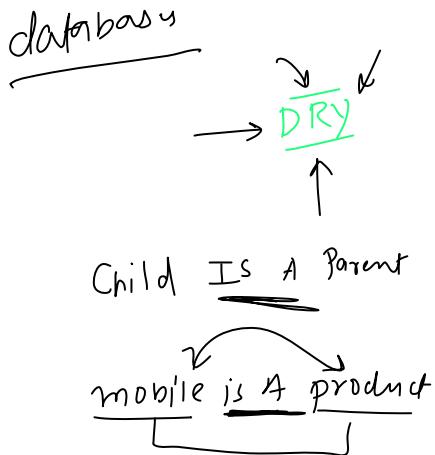
# Visualization 1

Thursday, May 20, 2021 1:52 PM





## Biology

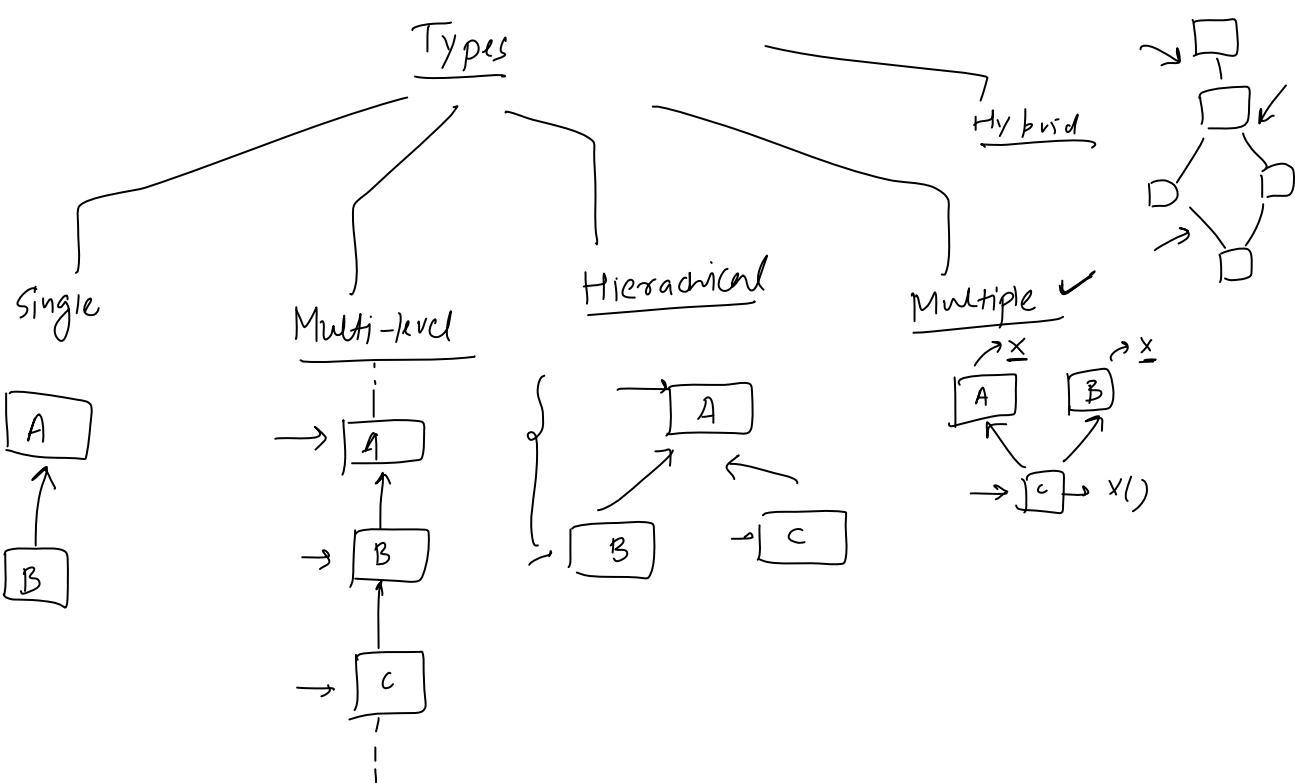


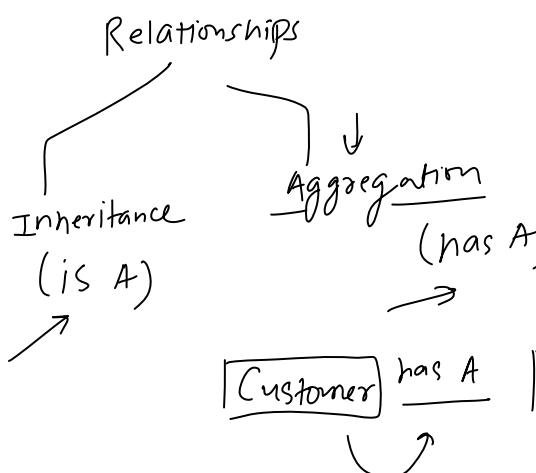
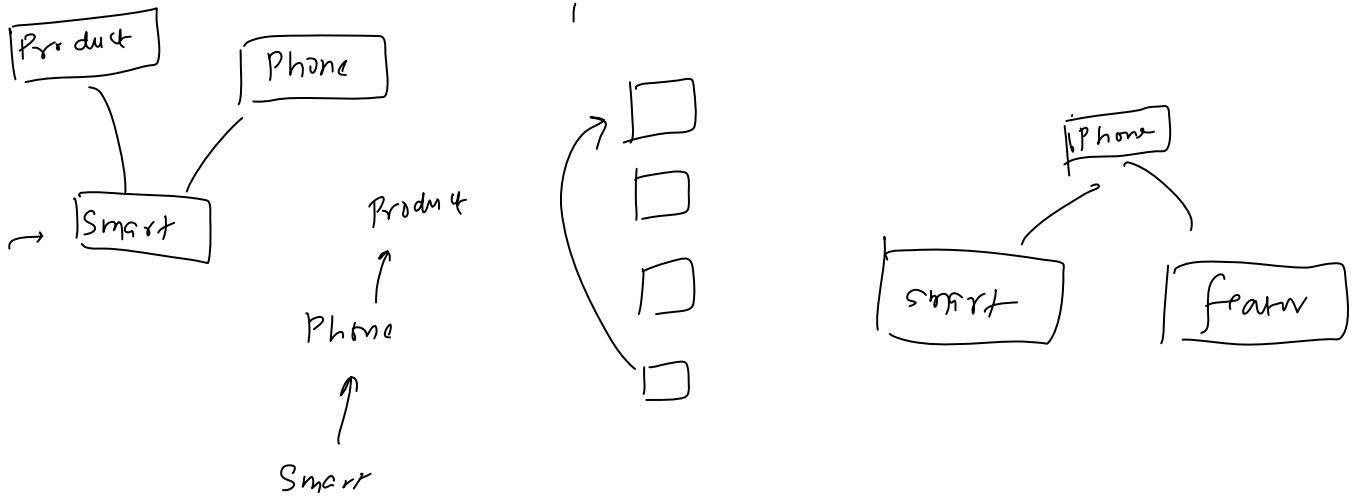
can be used

→ data (propertys)

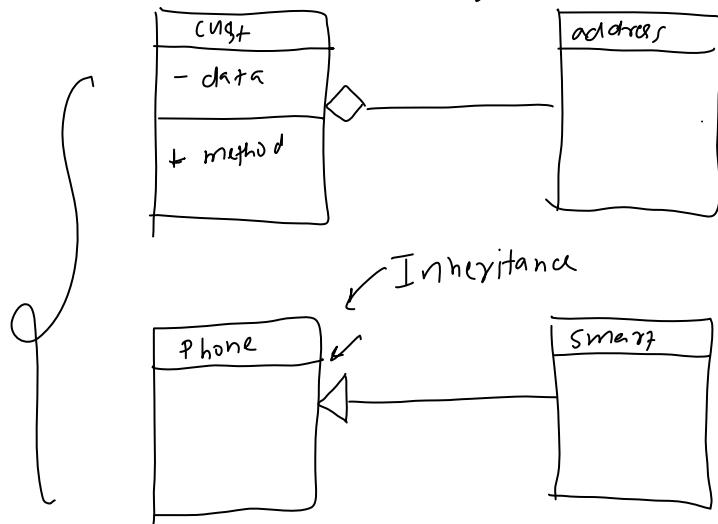
→ function

→ constructor

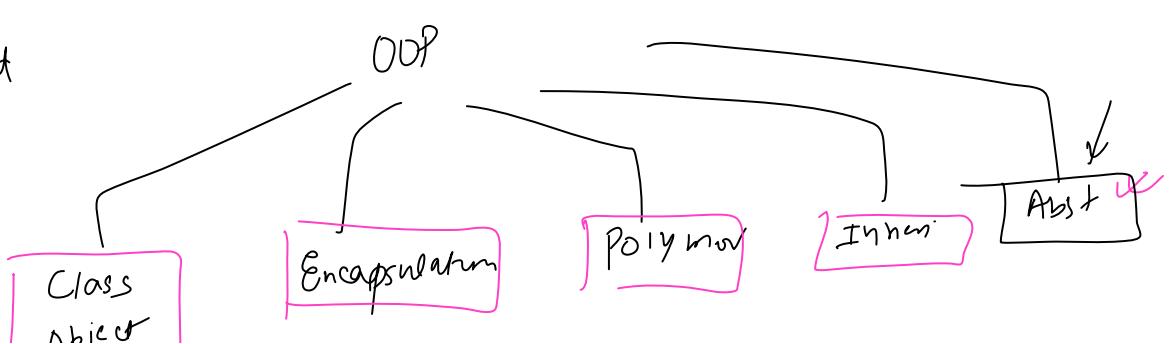




Class diagram



{  
Construct  
static method  
inst vs sm  
super



Class  
Object

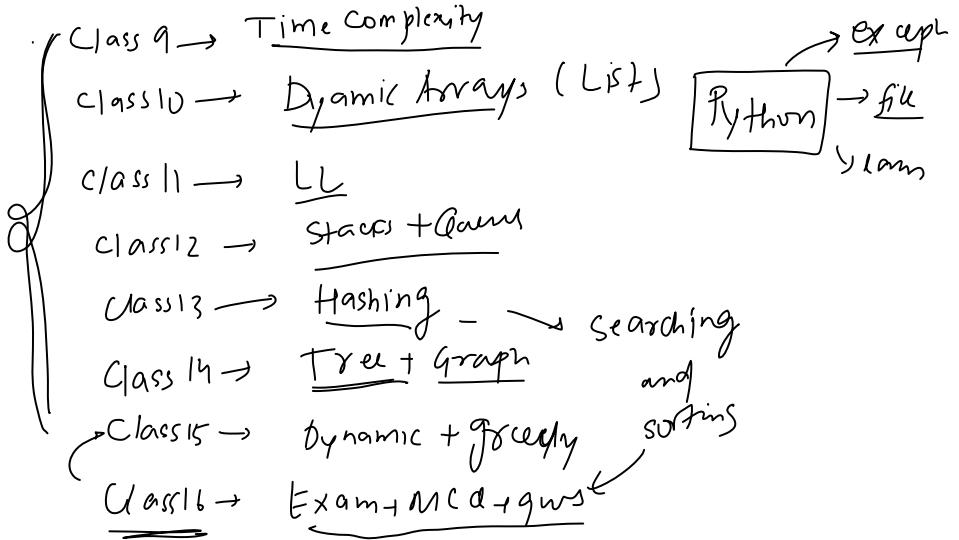
Encapsulation

Time complexity

Time complexity

(20) → DSAlg → ⑧ class

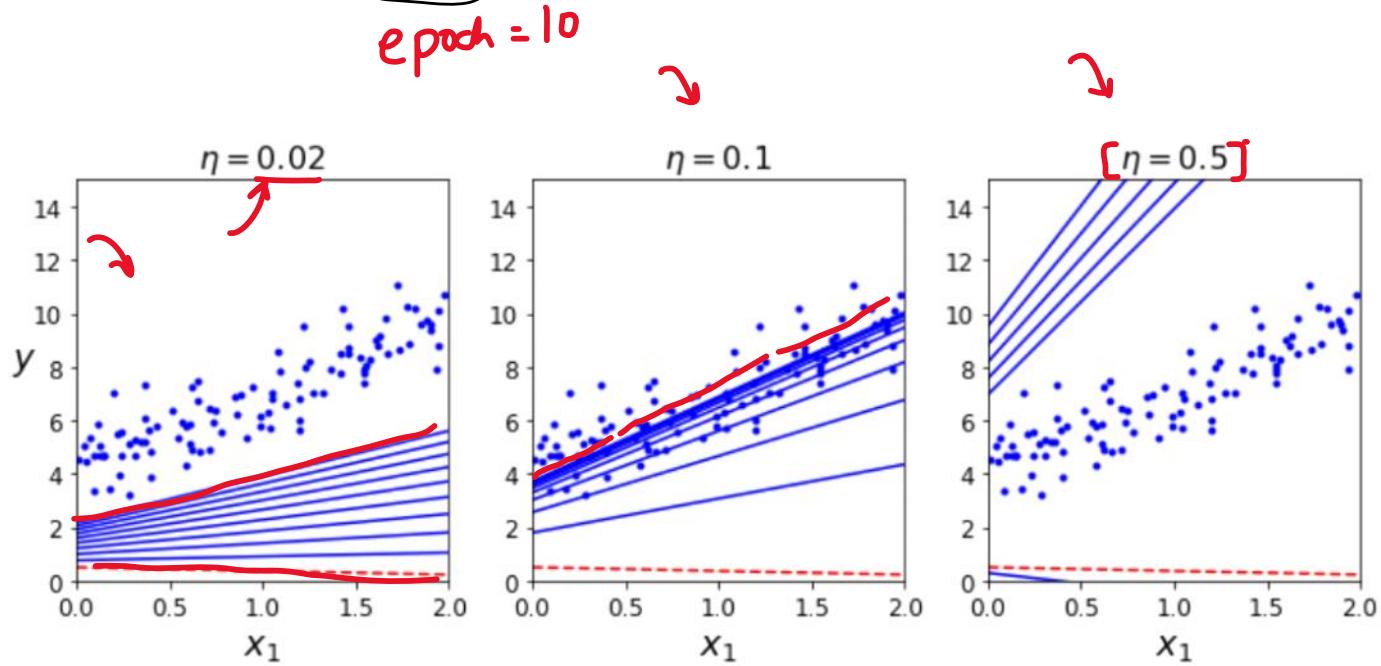
{  
17 18 19 20 }  
SOL



## Few Discussions

Thursday, May 20, 2021 4:47 PM

1. Effect of Learning rate
2. The universality of Gradient Descent



$$\hat{b} = 0$$

$$b = b_{\text{old}} - \eta \underbrace{\text{Slope}}_{\frac{dL}{db}}$$

$$\frac{dL}{db} = \left[ \sum (y_i - \hat{y}_i)^2 \right] \quad \text{LR} \rightarrow \text{Deep} \rightarrow \text{LDR} \rightarrow \underline{\text{functions}}$$

## Adding m into the mix

Thursday, May 20, 2021 1:48 PM

### 8 steps

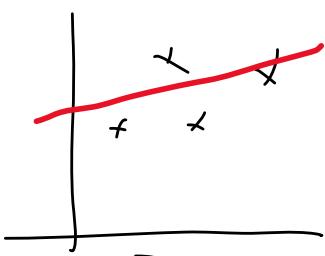
1) init random vals for  $m$  and  $b$   
 $m = 1$  and  $b = 0$

2) epochs = 100,  $\eta = 0.01$

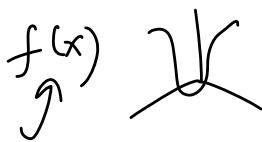
for i in epochs:

$$b = b - \eta \boxed{\text{slope}}$$

$$m = m - \eta \boxed{\text{slope}}$$



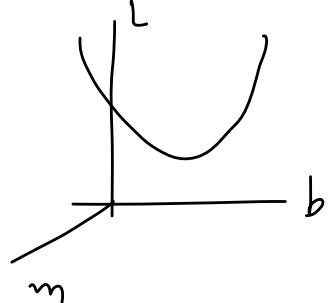
$$z = f(x, y)$$



$$b = 0$$

$$\begin{aligned} L(m, b) &= \sum (y_i - \hat{y}_i)^2 \\ b\text{-slope} &= \boxed{\frac{\partial L}{\partial b}} \end{aligned}$$

$$m\text{-slope} = \frac{\partial L}{\partial m} = \sum (y_i - mx_i - b)^2$$



2D gradient path ( $m, b$ )

$$\begin{aligned} \frac{\partial L}{\partial b} &= -2 \sum (y_i - mx_i - b) \\ &= -2 \sum (y_i - mx_i - b) \\ &= \text{slope}_b \text{ at } b = 0 \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial m} &= 2 \sum (y_i - mx_i - b) \\ &= -2 \sum (y_i - mx_i - b) x_i \\ &\quad \text{slope}_m \text{ at } \boxed{m = 1} \end{aligned}$$

# Code

Thursday, May 20, 2021 1:48 PM

# Visualization 2

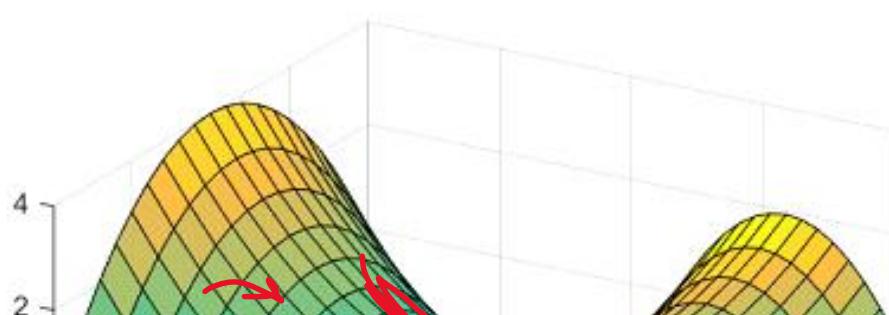
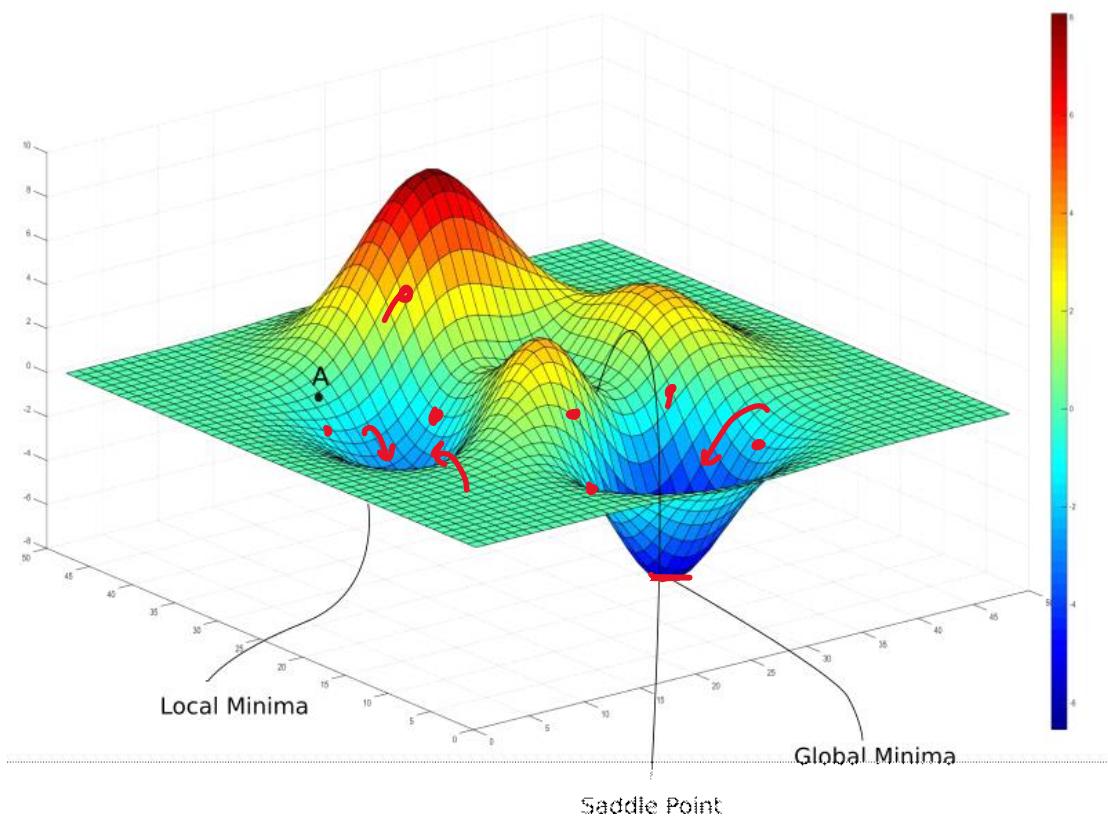
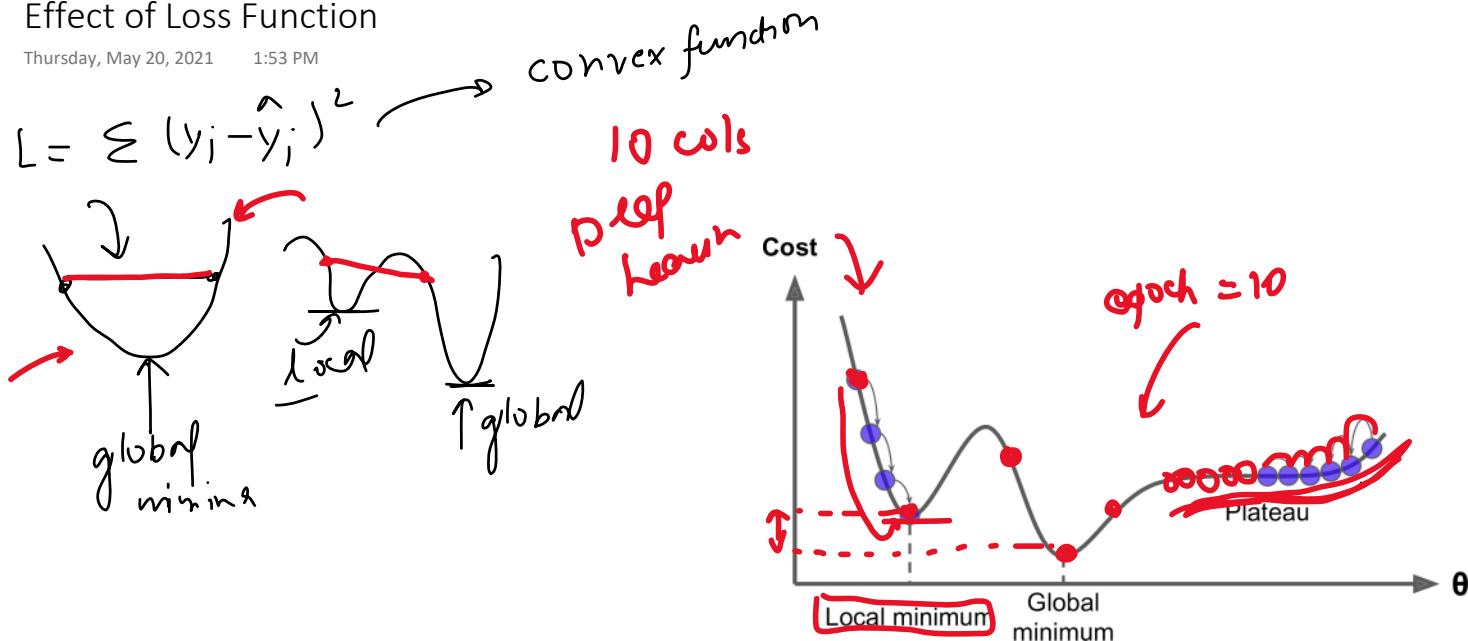
Thursday, May 20, 2021 1:52 PM

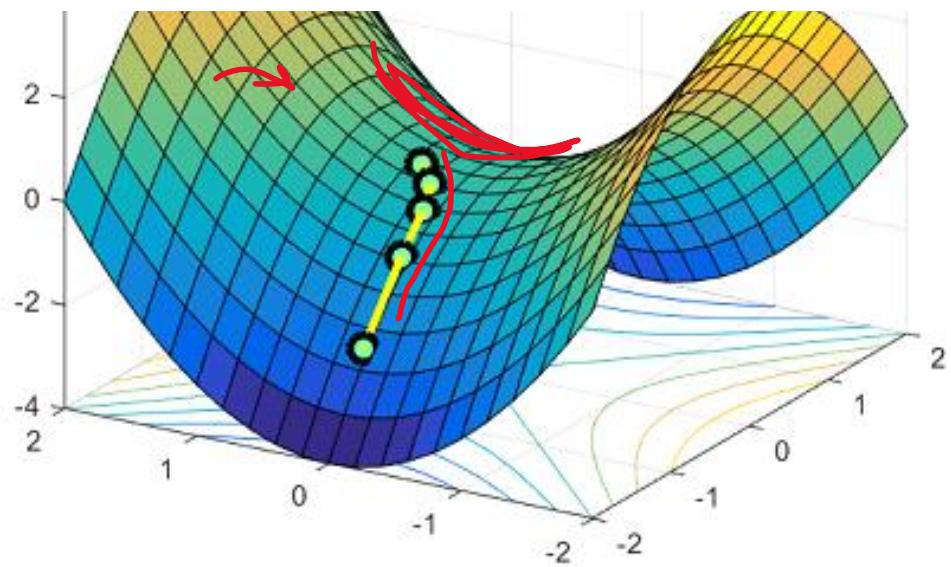
# Effect of Learning Data

Thursday, May 20, 2021 1:53 PM

## Effect of Loss Function

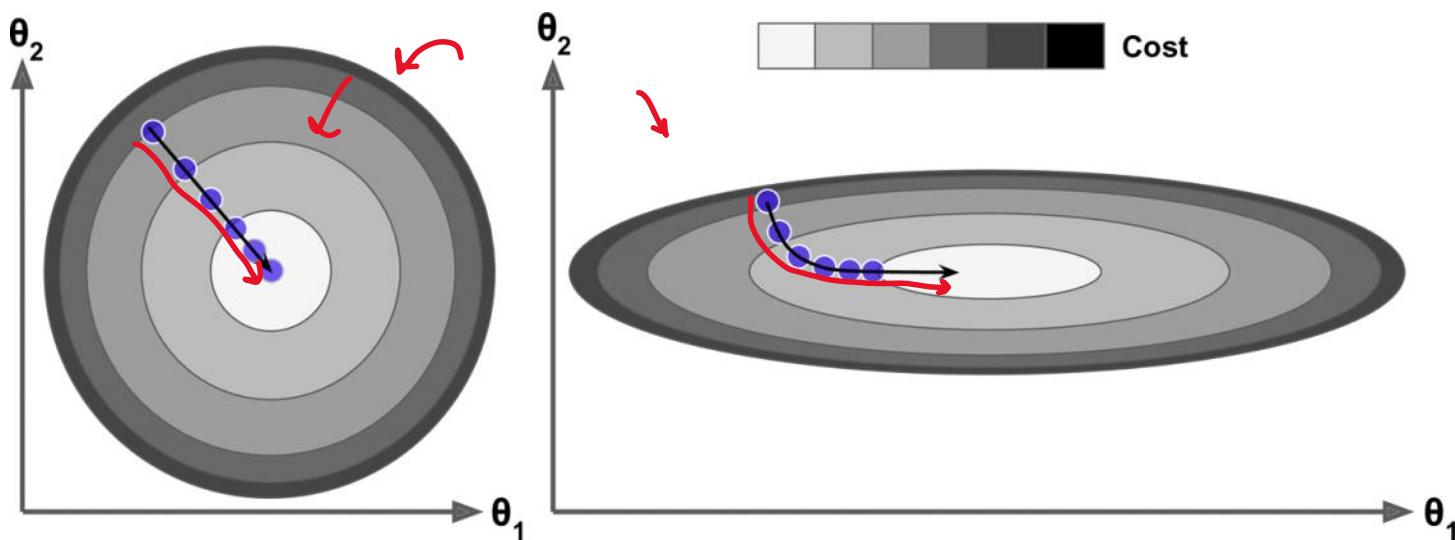
Thursday, May 20, 2021 1:53 PM





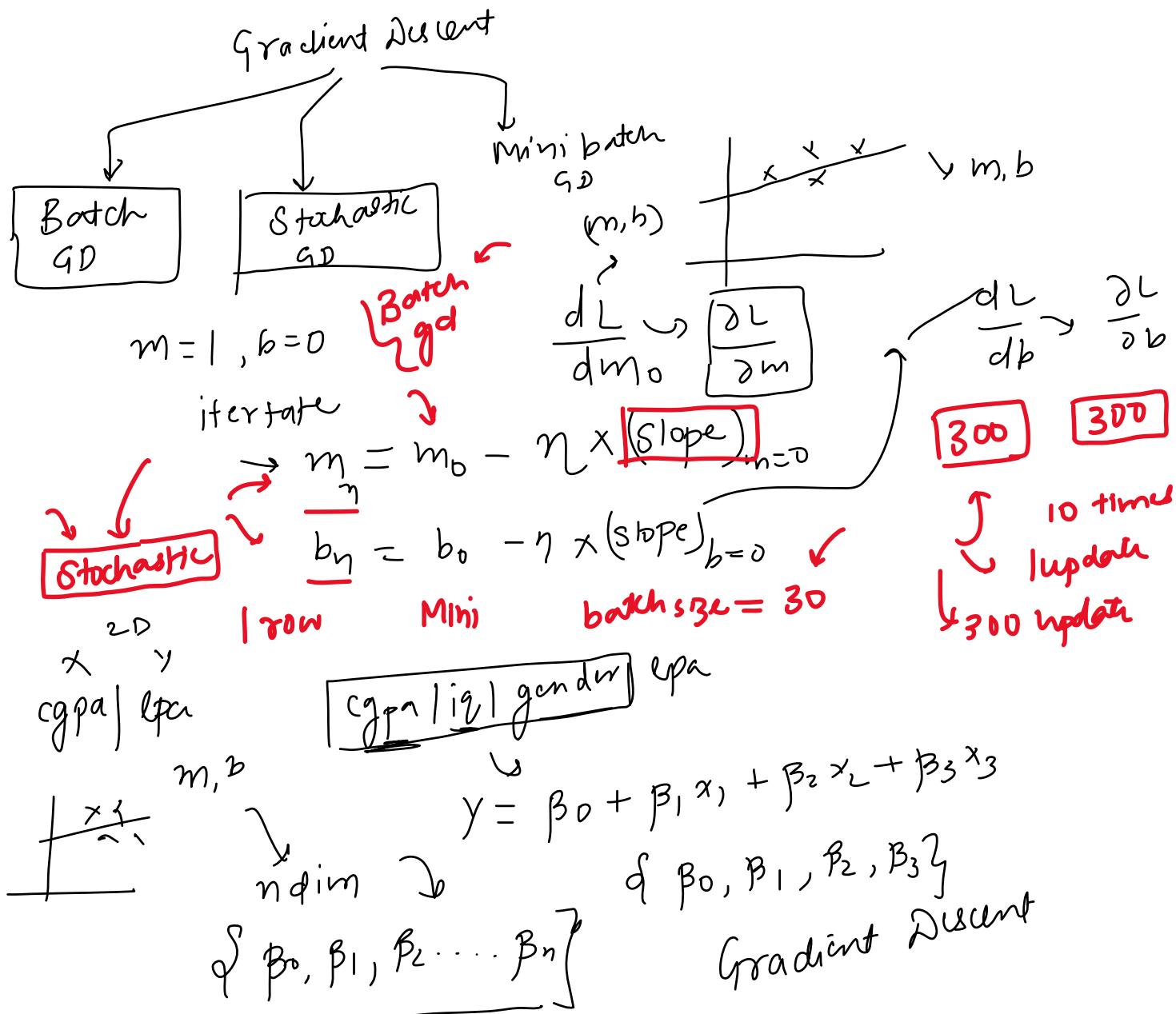
## Effect of Data

Thursday, May 20, 2021 1:53 PM



## Types of Gradient Descent

Saturday, May 22, 2021 1:30 PM



$n$ -dim dataset 3-cols

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

(lpa) (cgpa) (iq)

$$\{m, b\}$$

$$\{\beta_0, \beta_1, \beta_2\}$$

1) Random values

$$\beta_0 = 0, \beta_1, \beta_2 = 1$$

2) epoch = 100, lr = 0.1

$$\begin{cases} \beta_0 = \beta_0 - \eta \text{ slope} \\ \beta_1 = \beta_1 - \eta \text{ slope} \\ \beta_2 = \beta_2 - \eta \text{ slope} \end{cases}$$

MSE

$$L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

{ row = 2, cols = 2+1 }

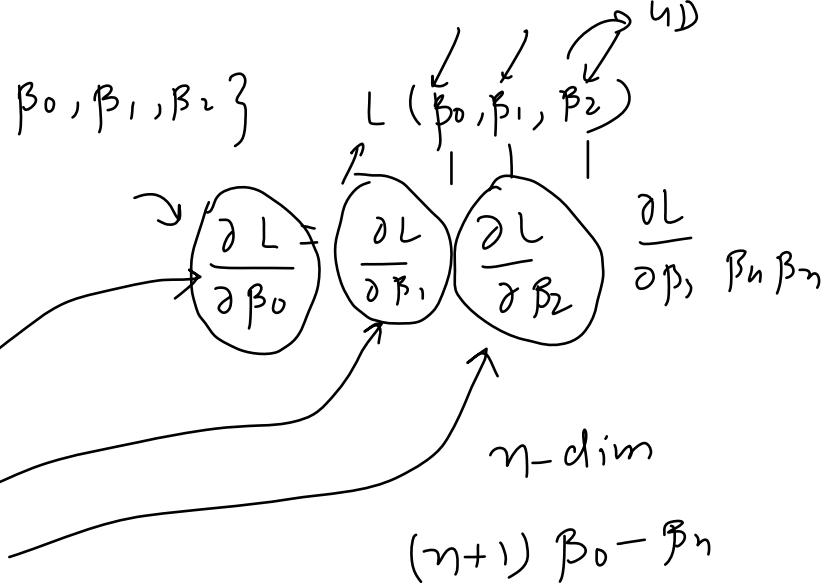
$$= \frac{1}{2} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2]$$

$$L = \frac{1}{2} [ (y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 ]$$

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{2} [ 2(y_1 - \hat{y}_1)(-1) + 2(y_2 - \hat{y}_2)(-1) ]$$

$$\frac{\partial L}{\partial \beta_0} = -\frac{1}{2} [ (y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) ]$$

$$= -\frac{2}{n} [ \frac{(y_1 - \hat{y}_1)}{m} + \frac{(y_2 - \hat{y}_2)}{m} + \dots + \frac{(y_n - \hat{y}_n)}{m} ]$$



$$\hat{y}_i = \beta_0 +$$

	$x_1$	$x_2$	$y$
1	8.1	9.3	3.2
2	7.5	9.5	3.5

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12}$$

$$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22}$$

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

$$= -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$= -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \frac{\partial L}{\partial \beta_0}$$

353 →  $y_i$   
353 →  $\beta_0$  mod  
 $x_{\text{train}}$

$$L = \frac{1}{2} \sum_{i=1}^2 (y_i - \hat{y}_i)^2 \left[ \frac{1}{n} \sum_{i=1}^n x_i \right] - \beta_1 x_{11} = -x_{11}$$

$$L = \frac{1}{2} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2]$$

$$L = \frac{1}{2} \left[ (y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 \right]$$

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{2} \left[ 2(y_1 - \hat{y}_1)(-x_{11}) + 2(y_2 - \hat{y}_2)(-x_{21}) \right] \quad \text{for } i=1, 2, 3, \dots, n$$

$$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n} \left[ (y_1 - \hat{y}_1)(x_{11}) + (y_2 - \hat{y}_2)(x_{21}) + (y_3 - \hat{y}_3)(x_{31}) + \dots + (y_n - \hat{y}_n)(x_{n1}) \right]$$

$$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{y}_i) x_{i1}$$

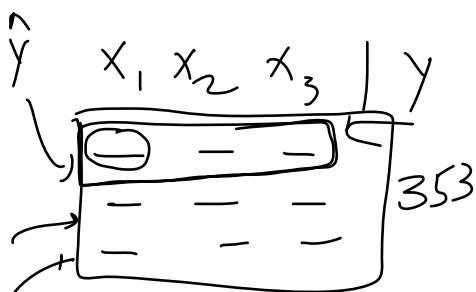
$x_{ij}$  → 1 col data  
 $\beta_1$  → values of 1 col.

$$\frac{\partial L}{\partial \beta_2} = -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{y}_i) x_{i2}$$

$m$  cols  
 $\beta_0 - \beta_m$

$$\frac{\partial L}{\partial \beta_m} = -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{y}_i) x_{im}$$

code



$$\hat{y}_i = \frac{\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}}{\beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23}}$$

$$\beta_0 + \text{np.dot}(x_{\text{train}}, \underline{\text{coef}}) \stackrel{x}{=} \beta_0 + [x_{11} \ x_{12} \ x_{13}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\beta_0 + (353, 1) \rightarrow$$

$\downarrow \quad \beta_{53, 1}$

$$\hat{y} = \text{np.dot}(\text{coef} \rightarrow x_{\text{train}}) + \beta_0$$

$$\begin{array}{c|cc|c|c|c} & x_1 & x_2 & y & \hat{y} \\ \hline 1 & 1 & 2 & 5 & 6 \\ 2 & 3 & 4 & 7 & 8 \end{array}$$

$$\hat{y} - y = [-1 \ -1]$$

$$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^n (y_i - \hat{y}_i) x_{i1} = -\frac{2}{n} \begin{bmatrix} [-1 \ -1] & \begin{bmatrix} 1 \\ 3 \end{bmatrix} \end{bmatrix}$$

$$\begin{aligned} \frac{\partial L}{\partial \beta_2} &= [y - \hat{y}] \begin{bmatrix} 2 \\ 4 \end{bmatrix} & \begin{bmatrix} -4 & -4 \\ -2 & -8 \end{bmatrix} &= 8 & (1, 2) \\ &= [-1 \ -1] \begin{bmatrix} 2 \\ 4 \end{bmatrix} \times -\frac{2}{n} & \begin{bmatrix} (1, 2) & \cancel{(2, 2)} \\ \cancel{(1, 2)} & (2, 2) \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times -\frac{2}{n} \end{aligned}$$

$$\frac{\partial L}{\partial \beta_1}, \dots, \frac{\partial L}{\partial \beta_{53}} = \left[ \boxed{(y_i - \hat{y}_i) \times \text{train}} \right] \times -\frac{2}{n}$$

$y_{\text{train}}$

$$y_i = 353$$

$\downarrow$

$$(353, 1) \rightarrow (1, 353)$$

$$(1, 353) \rightarrow (353, 1)$$

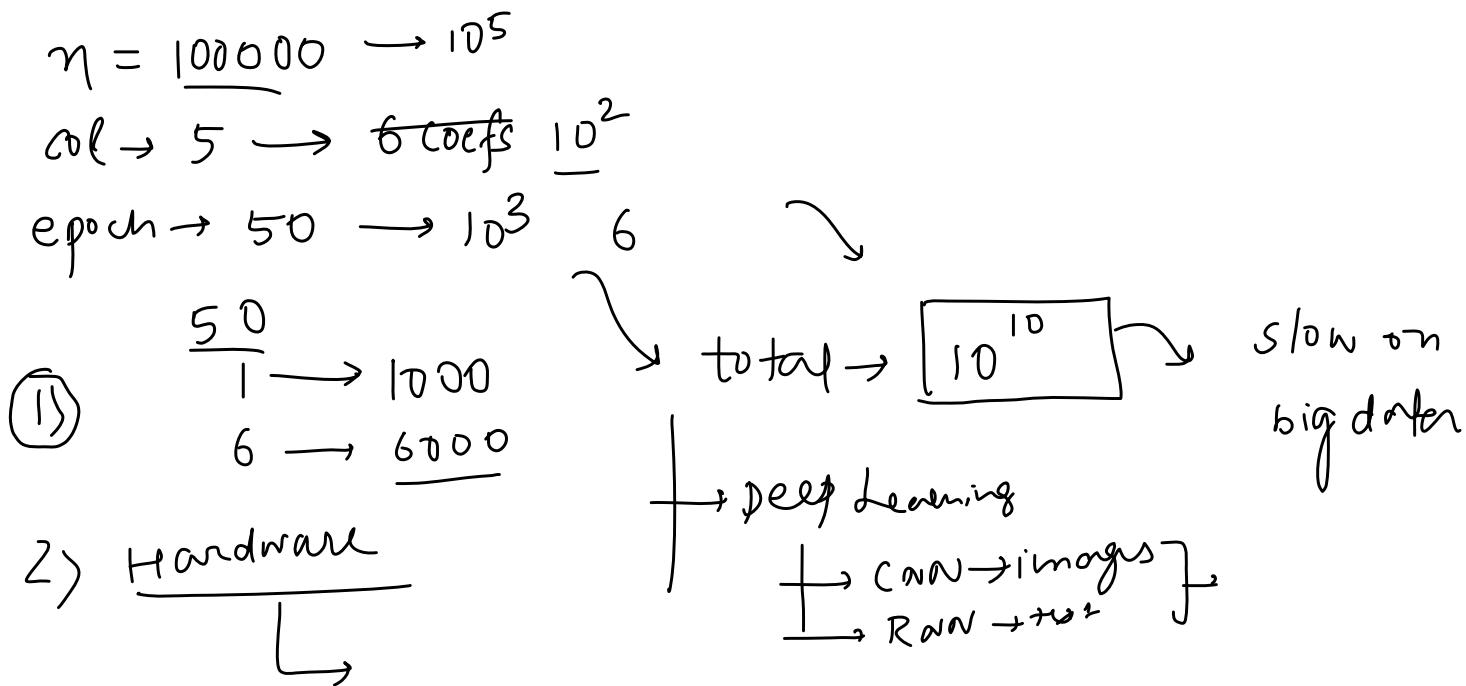
$$\begin{bmatrix} -2 \\ n \end{bmatrix}$$

coeff-clm

$$(1, 10) \times \begin{bmatrix} -2 \\ n \end{bmatrix} = (1, 10)$$

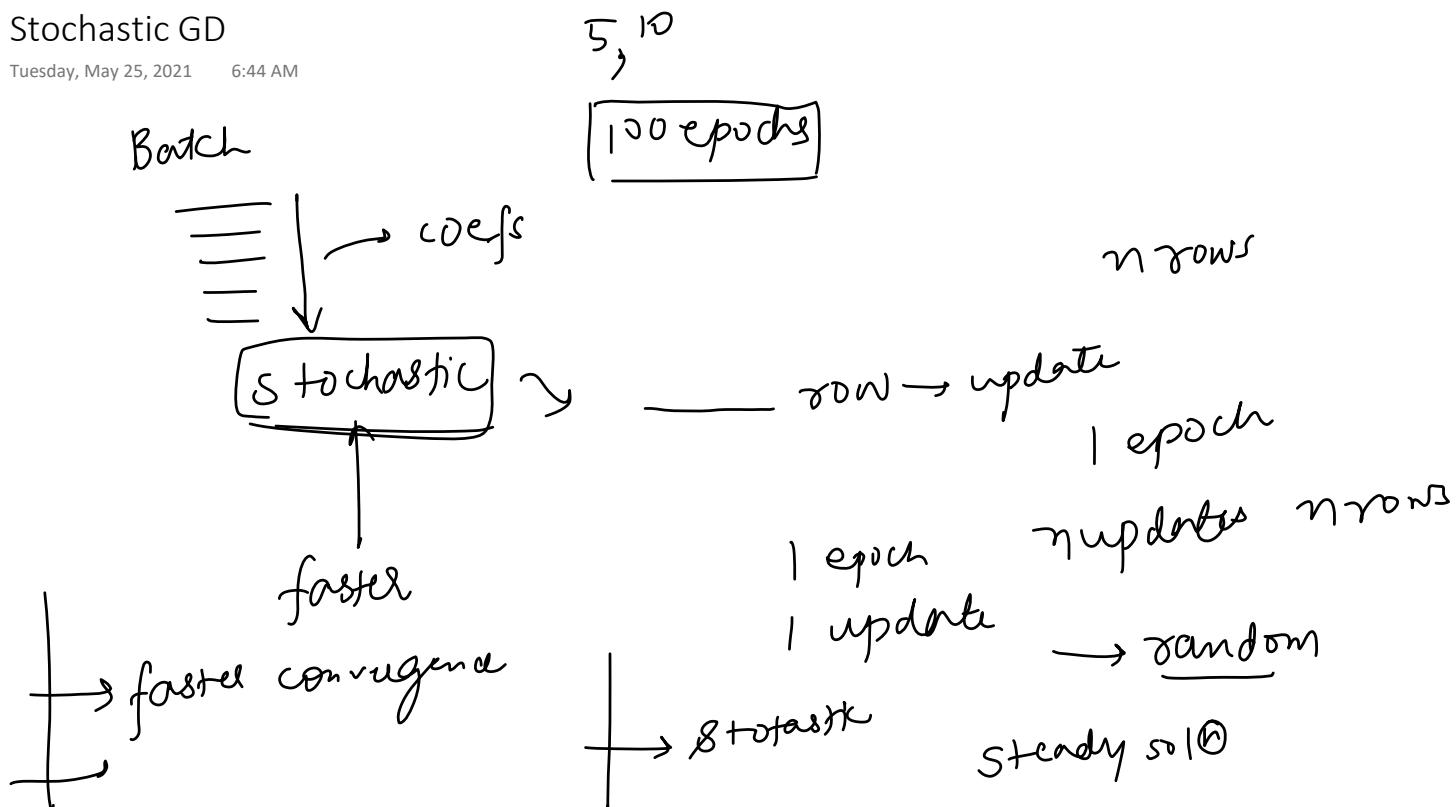
## The Problem with Batch GD

Tuesday, May 25, 2021 6:43 AM



## Stochastic GD

Tuesday, May 25, 2021 6:44 AM



# Code

Tuesday, May 25, 2021

6:44 AM

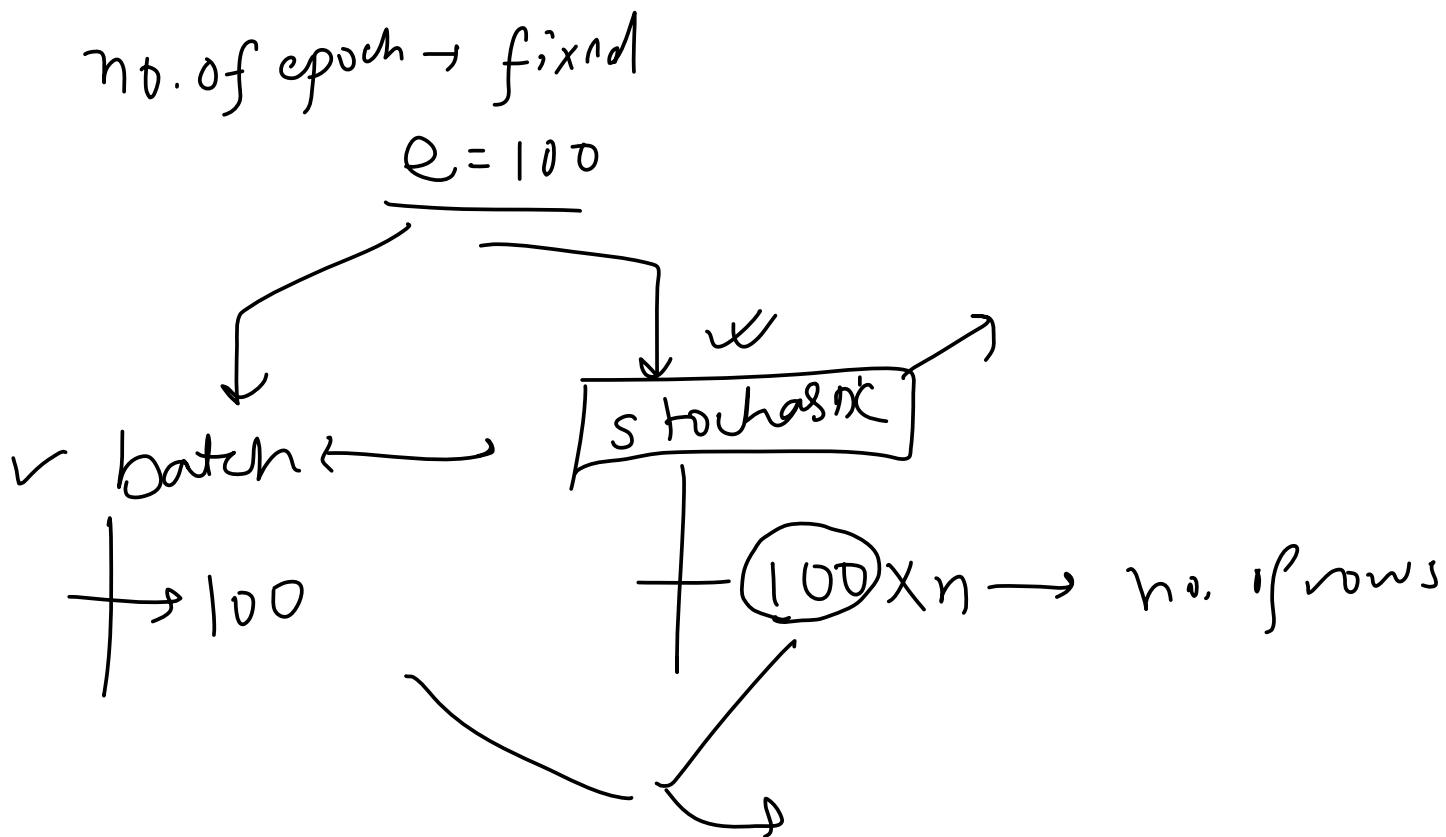
$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^m (y_i - \hat{y}_i)$$

$i = id X$

$$= -2 \underline{(y_i - \hat{y}_i)}$$

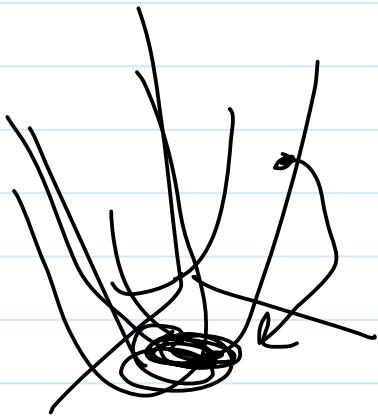
# Time Comparison

Tuesday, May 25, 2021 12:39 PM



# Visualizations

Tuesday, May 25, 2021 6:44 AM

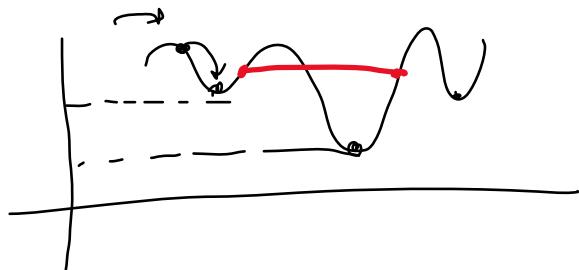


## When to use Stochastic GD

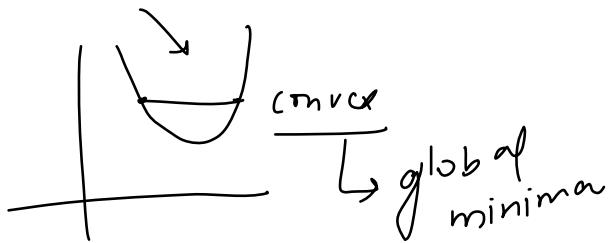
Tuesday, May 25, 2021 6:44 AM

1) Big data  $\rightarrow$  SGD

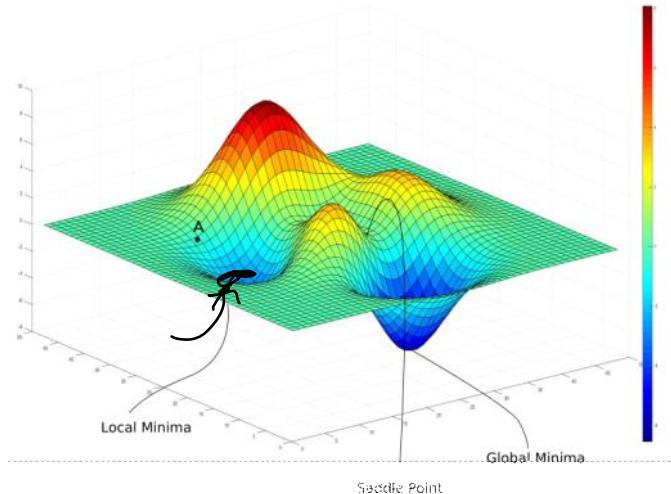
2) Non convex function



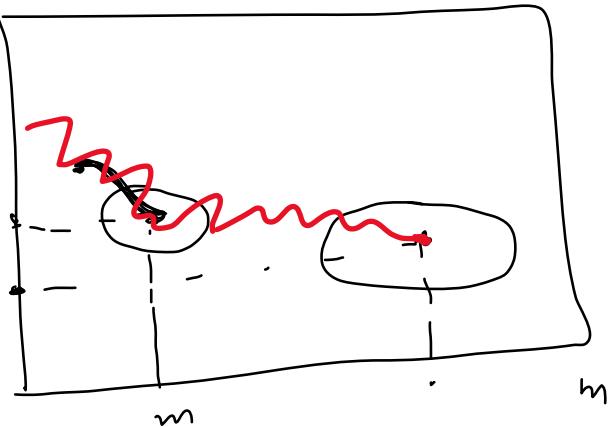
non convex



learning



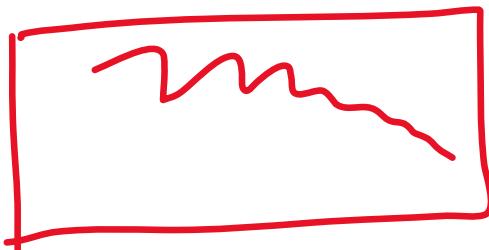
b



# Learning Schedules

Tuesday, May 25, 2021 7:00 AM

DL



$$\eta = 100$$
$$\text{epoch} = 1$$

$$\rightarrow$$

$$lr = 0.1$$

$$lr = 0.03$$

```
t0, t1 = 5.50
def learning_rate(t):
    return t0/(t + t1)

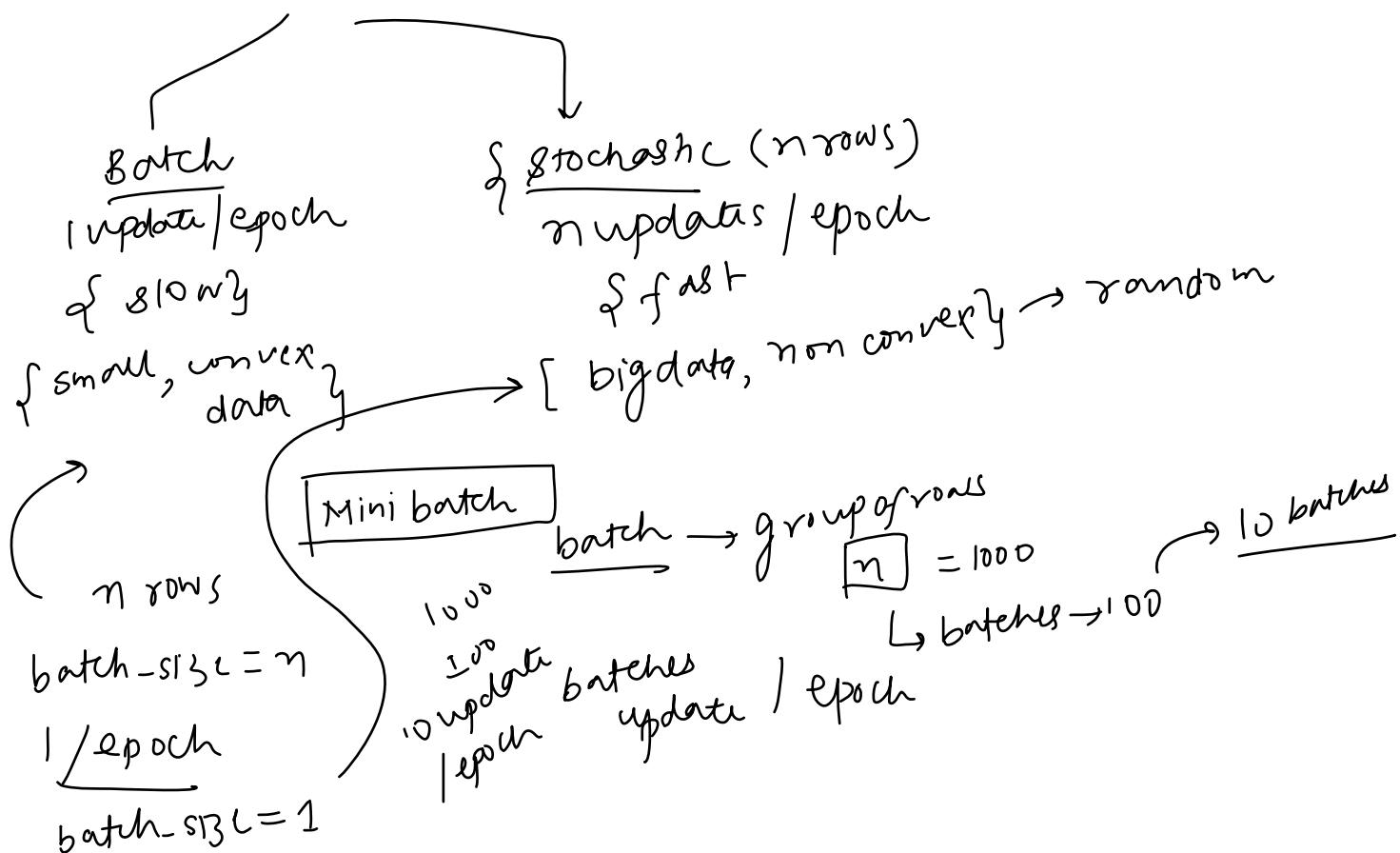
for i in range(epochs):
    for j in range(X.shape[0]):
        lr = learning_rate(i * X.shape[0] + j)
```

# Sklearn Implementation

Tuesday, May 25, 2021 2:16 PM

## Mini-Batch Gradient Descent

Wednesday, May 26, 2021 4:47 PM



# Code

Wednesday, May 26, 2021 4:47 PM

# Visualization

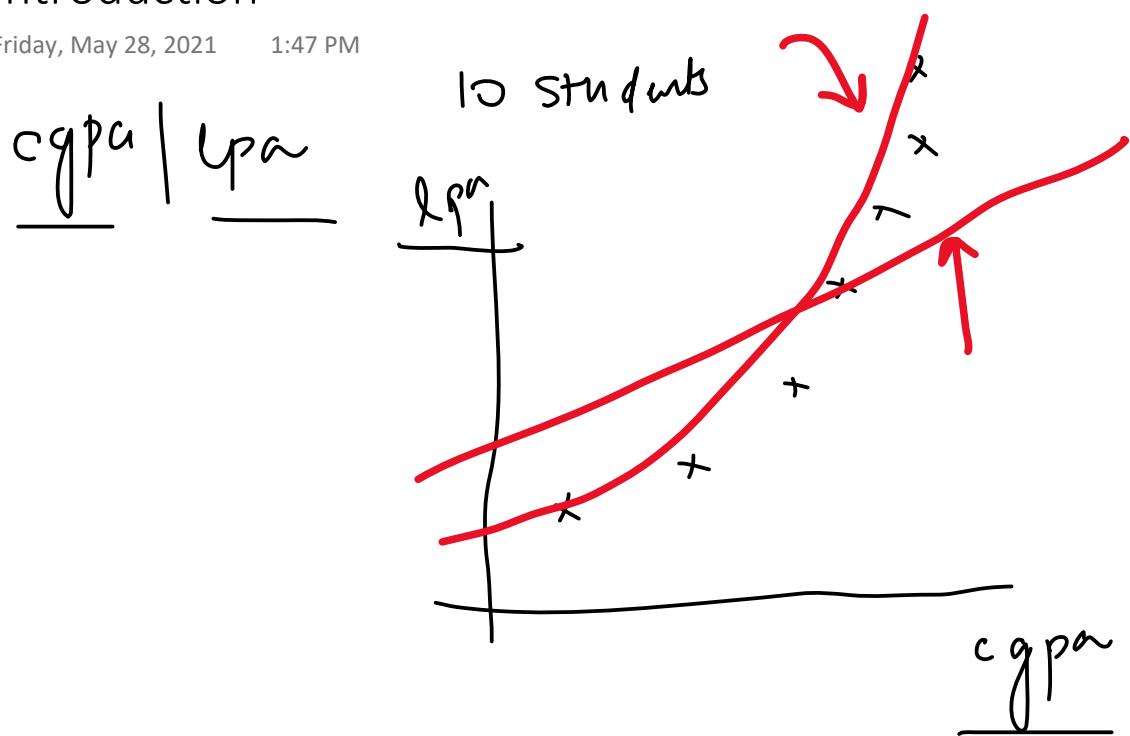
Wednesday, May 26, 2021 4:47 PM

# Sklearn Implementation

Wednesday, May 26, 2021 4:47 PM

# Introduction

Friday, May 28, 2021 1:47 PM

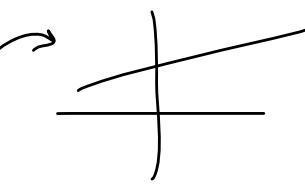


# Intuition

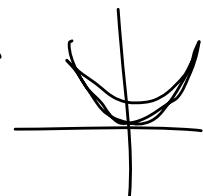
Friday, May 28, 2021 1:47 PM

polynomials

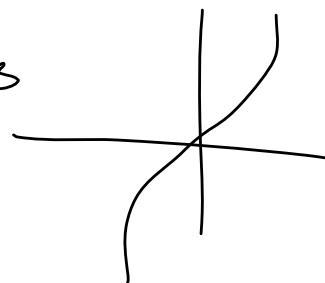
$$y = mx + b$$



$$y = x^2$$



$$y = x^3$$



$$x^3 + x^2$$

$$y = \textcircled{a}x^4 + \textcircled{b}x^3 + \textcircled{c}x^2 + dx + e$$

↑      ↑      ↑      ↑      ↓  
degree

$$\begin{array}{r} 1 \\ -x \\ \hline 2 \\ \end{array} \quad \begin{array}{r} y \\ \curvearrowright \\ \end{array} \quad \rightarrow \quad y = mx + b$$

↑      ↓  
transform polynomial      degree = 2

$$\begin{array}{r} 3 \\ x^0 \quad | \quad x^1 \quad | \quad x^2 \\ \hline 1 \quad | \quad 2 \quad | \quad 4 \\ \hline 1 \quad | \quad 3 \quad | \quad 9 \end{array}$$

$$Y = \beta_0 + \beta_1 x^0 + \beta_2 x^2$$

# Code

Friday, May 28, 2021 1:48 PM

# Multiple Polynomial Regression

Friday, May 28, 2021 1:48 PM

# Why is it linear

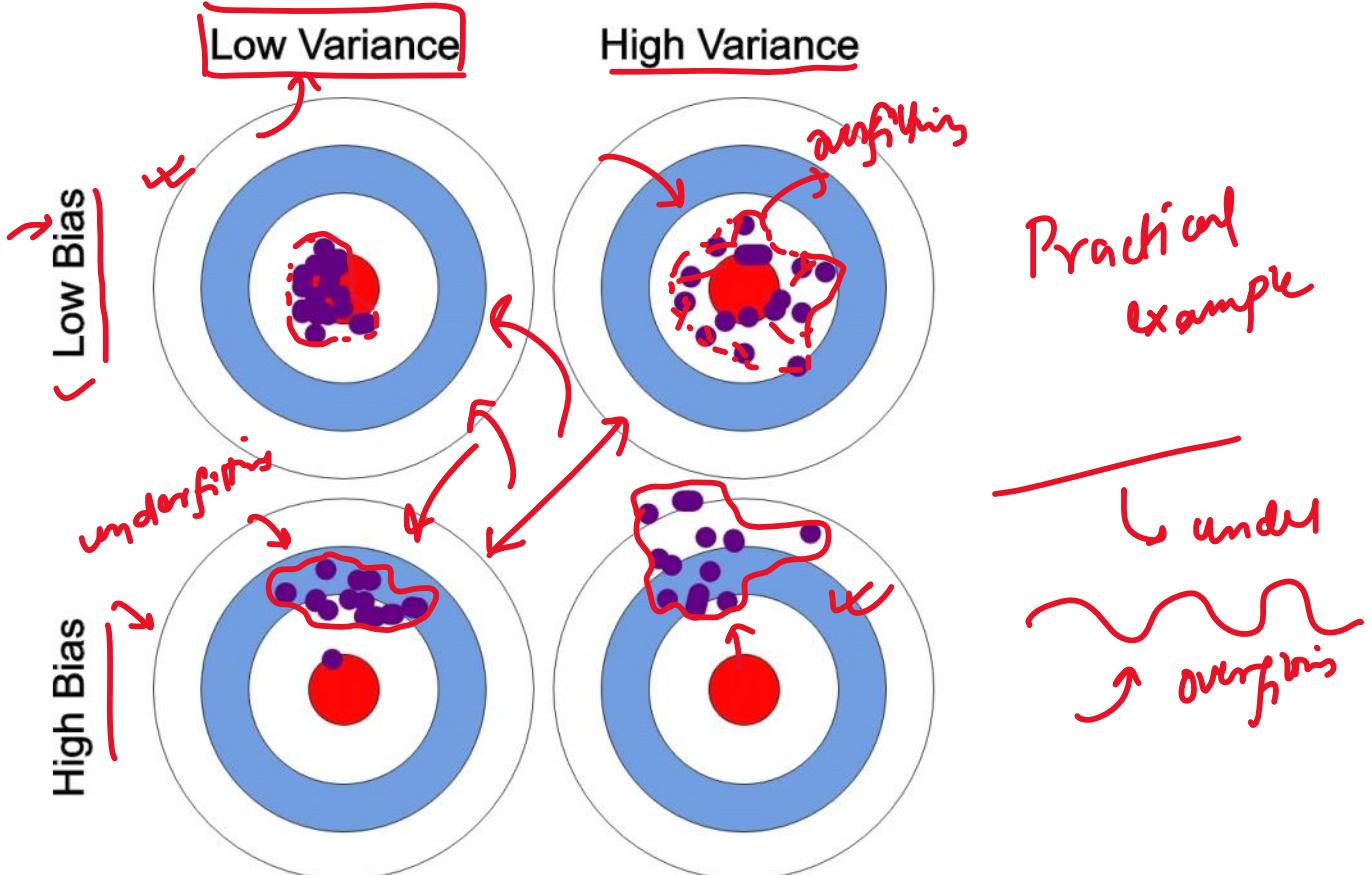
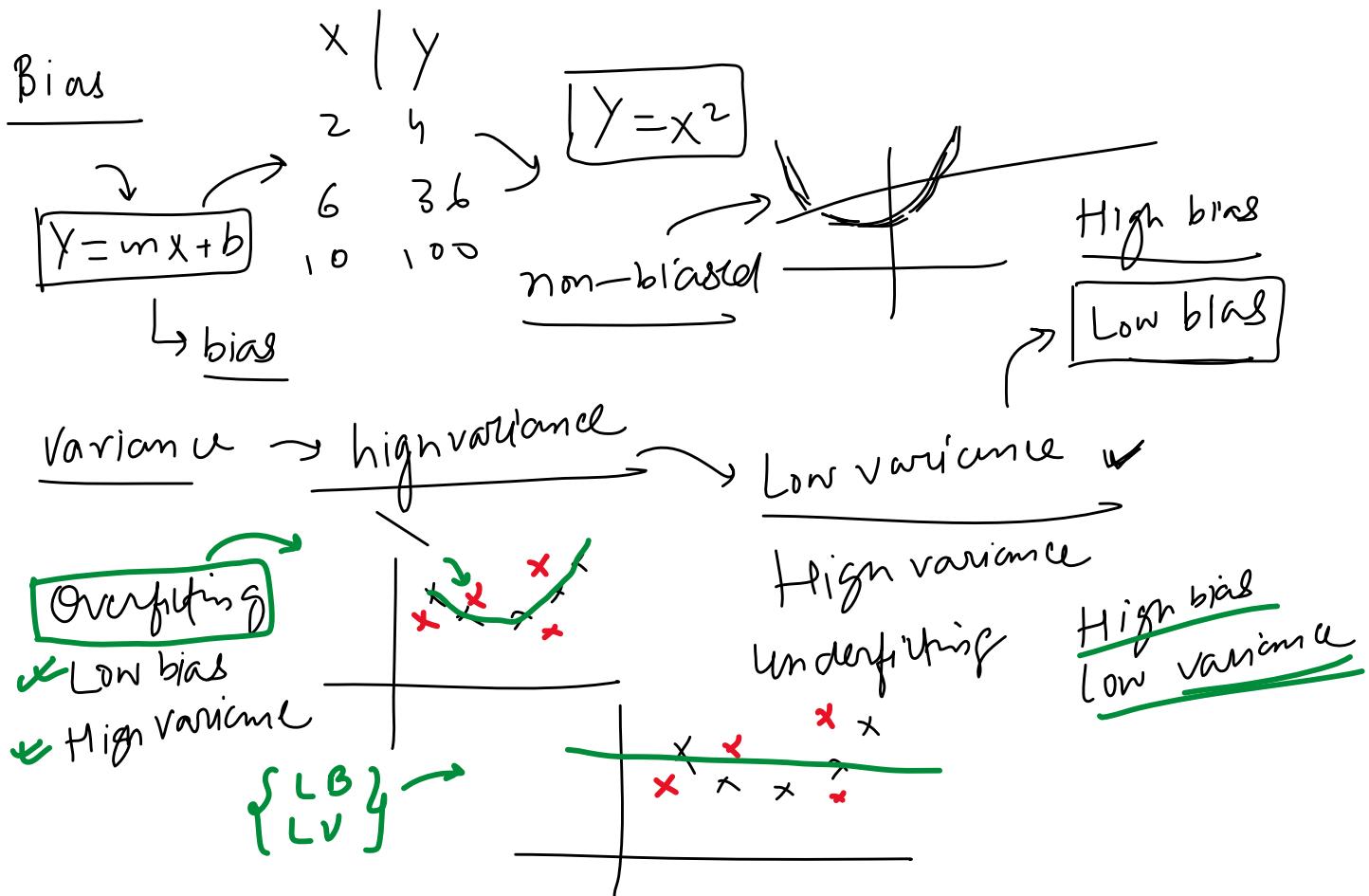
Friday, May 28, 2021 1:48 PM

# When to use Polynomial Regression

Friday, May 28, 2021 1:48 PM

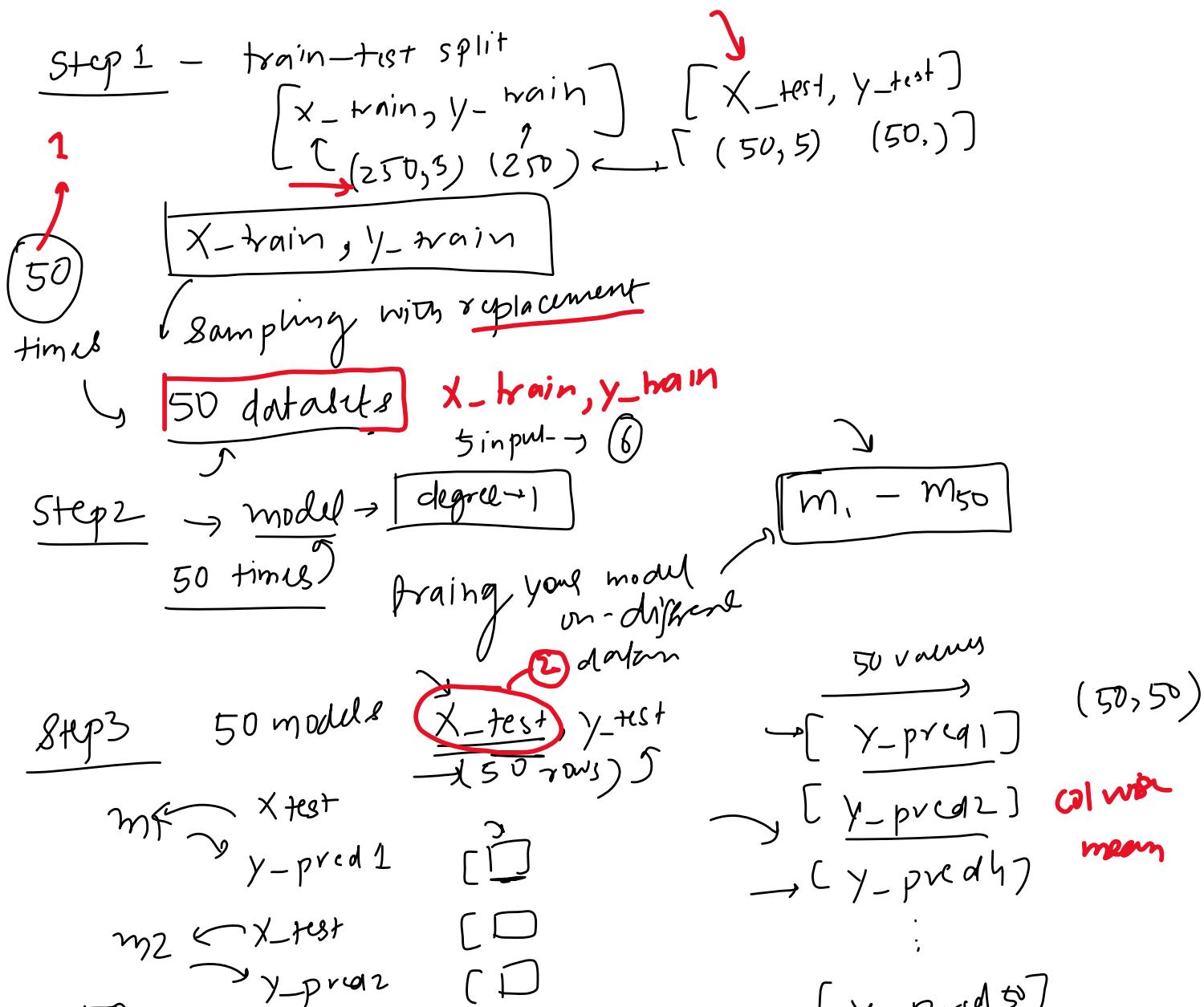
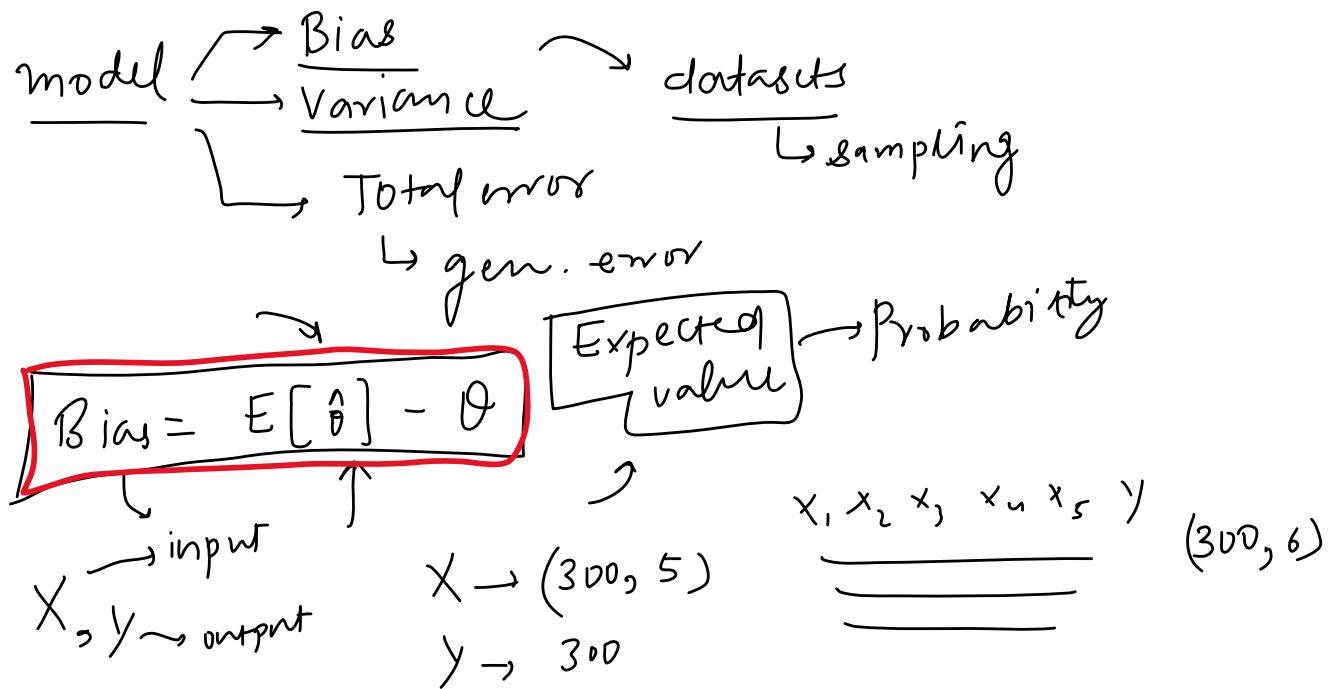
# Bias And Variance

Tuesday, June 1, 2021 12:20 PM



# How to calculate Bias and Variance

Tuesday, June 1, 2021 1:44 PM



$m_2 \leftarrow \text{`x-test'}$   
 $y_{\text{pred2}} \leftarrow$   
 $\begin{bmatrix} 50 \\ \vdots \\ 50 \end{bmatrix} \rightarrow$   
 $L \leftarrow \begin{bmatrix} \square \\ \vdots \\ \square \end{bmatrix}$

$\frac{[y_{\text{pred5}}]}{\text{mean\_prediction}}$   
 $\text{mean\_prediction} \leftarrow \begin{bmatrix} \dots \\ \text{---} \end{bmatrix}$   
 $\text{50 values}$

$\frac{(y_{\text{pred5}} - \text{mean\_prediction})^2}{50} = \boxed{\text{bias}}$

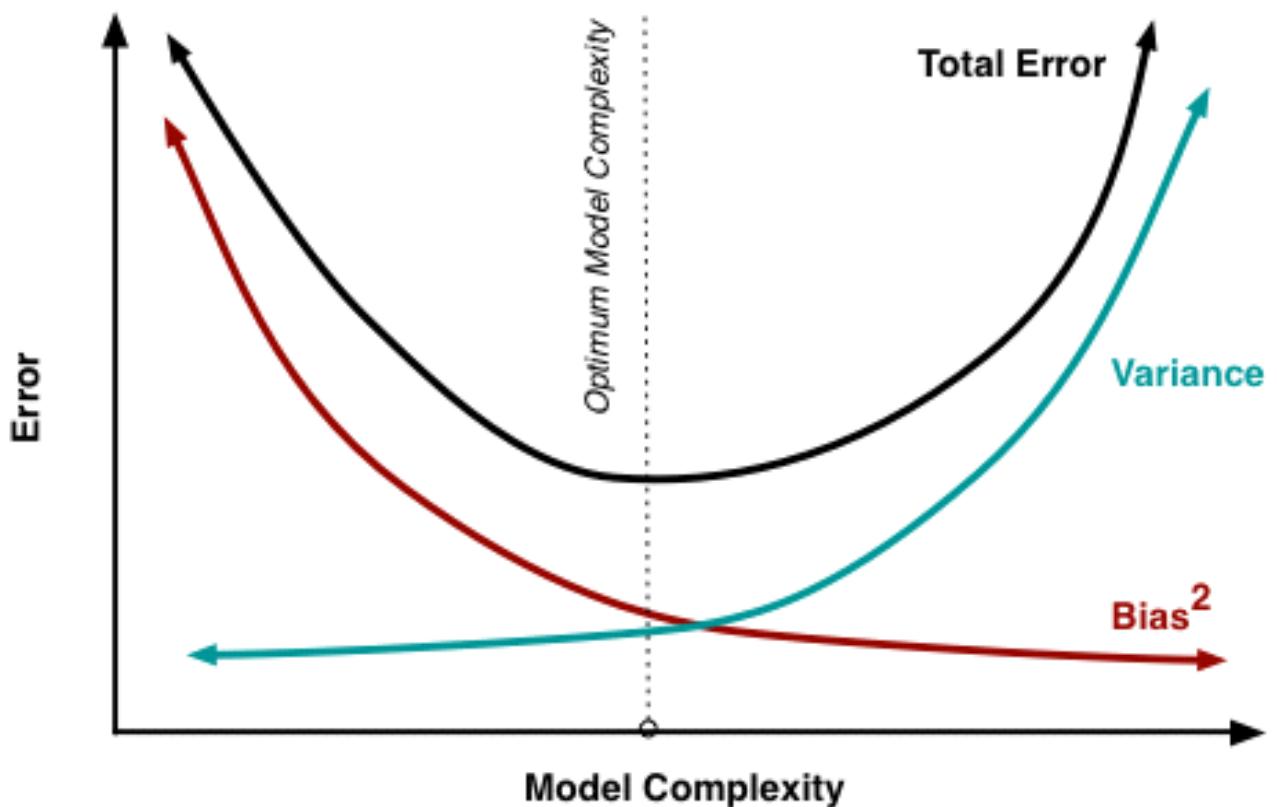
Variance

# Code

Monday, May 31, 2021 9:26 AM

# Bias Variance Decomposition for Squared Error

Monday, May 31, 2021 9:26 AM

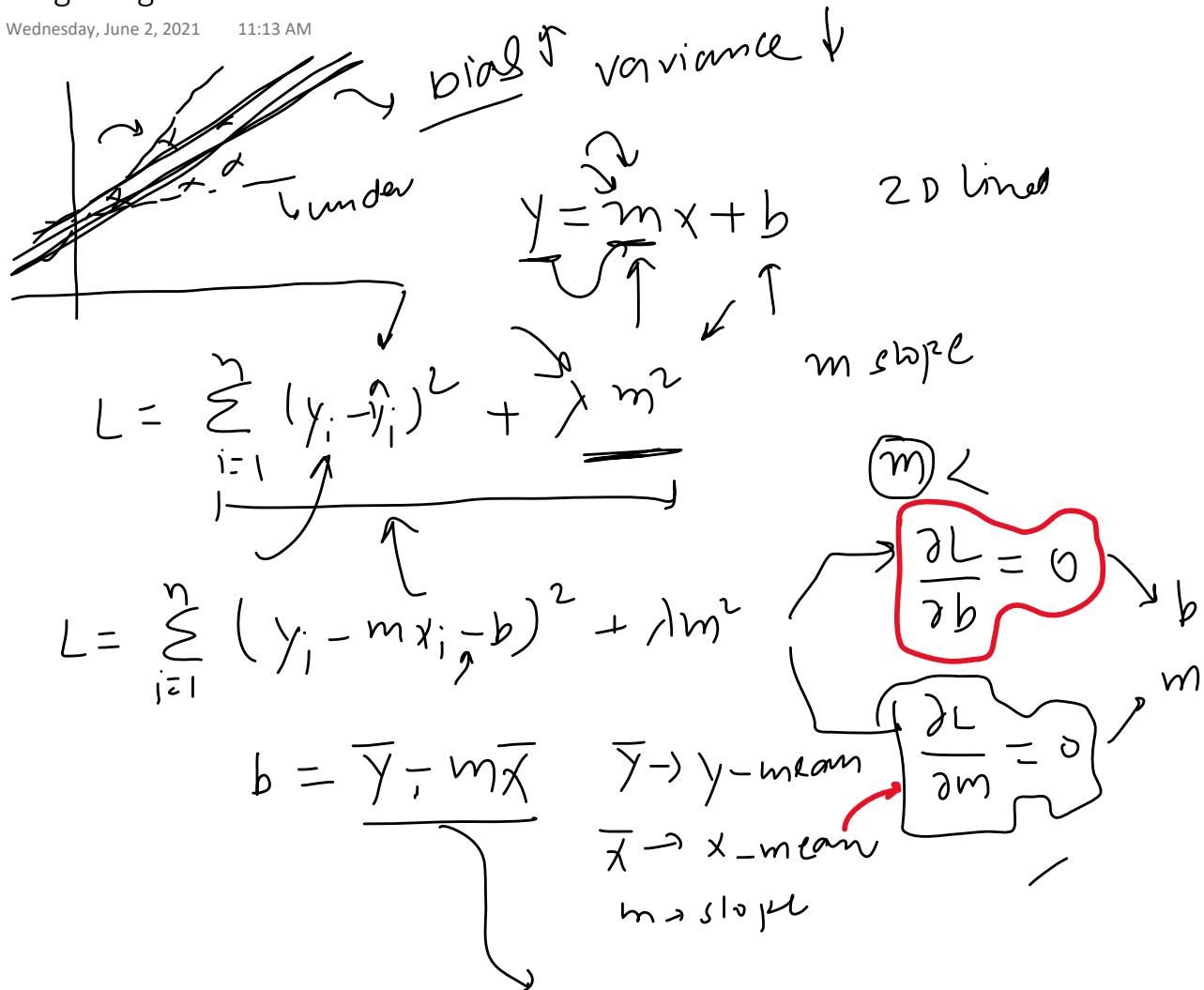


# Relationship between Bias and Variance

Monday, May 31, 2021 9:26 AM

## Ridge Regression

Wednesday, June 2, 2021 11:13 AM



$$L = \sum_{i=1}^n (y_i - mx_i - \bar{y} + \bar{mx})^2 + \lambda m^2$$

$$\frac{\partial L}{\partial m} = 2 \sum_{i=1}^n (y_i - mx_i - \bar{y} + \bar{mx}) (-x_i + \bar{x}) + 2\lambda m = 0$$

$$= -2 \sum_{i=1}^n (y_i - \bar{y} - mx_i + \bar{mx}) (x_i - \bar{x}) + 2\lambda m = 0$$

$$= \lambda n - \sum_{i=1}^n [(y_i - \bar{y}) - m(x_i - \bar{x})] (x_i - \bar{x}) = 0$$

n                    ...                    r.s. = 1 - n

$$= \lambda m - \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2 = 0$$

$$= \lambda m - \underbrace{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}_{\text{red underline}} + m \sum_{i=1}^n (x_i - \bar{x})^2 = 0$$

$$= \lambda m + m \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$$

$$= m = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2 + \lambda}$$

hyperparameter  
alpha

$$m = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\lambda = 0 = \lambda = 10,000$

$b = \bar{y} - m \bar{x}$

# Ridge Regression for 2D data

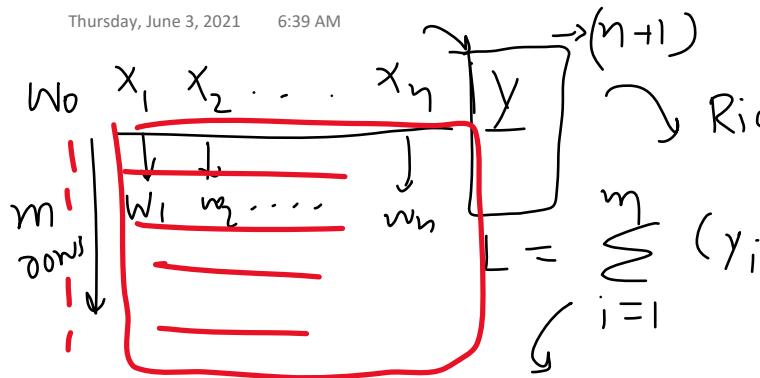
Thursday, June 3, 2021 6:38 AM

# Code

Thursday, June 3, 2021 6:39 AM

## Ridge Regression for nD data

Thursday, June 3, 2021 6:39 AM



$$= (\mathbf{x}\mathbf{w} - \mathbf{y})^\top (\mathbf{x}\mathbf{w} - \mathbf{y})$$

*m values*

$$\mathbf{y} = \begin{bmatrix} y \\ - \\ - \\ - \\ - \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} | & x_{11} & x_{12} & \dots & x_{1n} \\ | & x_{21} & x_{22} & \dots & x_{2n} \\ | & \vdots & \vdots & \ddots & \vdots \\ | & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

Normal LR  $\rightarrow$  Ridge

$$L = (\mathbf{x}\mathbf{w} - \mathbf{y})^\top (\mathbf{x}\mathbf{w} - \mathbf{y}) + \lambda \|\mathbf{w}\|^2$$

$$[\mathbf{w}_0 \mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_m] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix} \underbrace{\mathbf{w}^\top \mathbf{w}}_{\lambda(\|w_0^2 + w_1^2 + w_2^2 + \dots + w_m^2)}$$

$$(\mathbf{a} - \mathbf{b})^\top = \mathbf{a}^\top - \mathbf{b}^\top$$

$$L = (\mathbf{x}\mathbf{w} - \mathbf{y})^\top (\mathbf{x}\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^\top \mathbf{w}$$

$$L = [(\mathbf{x}\mathbf{w})^\top - (\mathbf{y})^\top] (\mathbf{x}\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^\top \mathbf{w}$$

$$= (\mathbf{w}^\top \mathbf{x}^\top - \mathbf{y}^\top) (\mathbf{x}\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^\top \mathbf{w}$$

$$= \mathbf{w}^\top \mathbf{x}^\top \mathbf{x}\mathbf{w} - \mathbf{w}^\top \mathbf{x}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{x}\mathbf{w} + \mathbf{y}^\top \mathbf{y} + \lambda \mathbf{w}^\top \mathbf{w}$$

$$L = \underbrace{w^T X^T X w}_{\text{red bracket}} - 2 \underbrace{w^T X^T y}_{\text{red bracket}} + \underbrace{y^T y}_{\text{red bracket}} + \lambda \underbrace{w^T w}_{\text{red bracket}}$$

$$\frac{dL}{dw} = \cancel{\rho} X^T X w - \cancel{\lambda} X^T y + 0 + \cancel{\lambda} \lambda w = 0$$

$$X^T X w + \cancel{\lambda} w = X^T y$$

$$(X^T X + \lambda I) w = X^T y$$

3 (4x4)  
(n x 1, n x 1)

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \boxed{w = (X^T X + \lambda I)^{-1} X^T y}$$

$$w = (X^T X)^{-1} X^T y \quad \boxed{[1]}$$

# Code

Thursday, June 3, 2021 6:39 AM

## Ridge Regression using Gradient Descent

Friday, June 4, 2021 2:17 PM

Vector form loss

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \Rightarrow \quad L = (xw - y)^T (xw - y) + \lambda w^T w$$

$\underbrace{L = (xw - y)^T (xw - y) + \lambda w^T w}_{\text{m rows}}$

$x_1 x_2 \dots x_n \circled{y}$   
 $m \text{ rows}$   
 $w_0 w_1 w_2 \dots w_n \circled{(n+1)}$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ 1 & x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$w_0, w_1, \dots, w_n$  (parameters)

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0} \quad \therefore w_1 = w_1 - \eta \frac{\partial L}{\partial w_1} \quad \dots \quad w_n = w_n - \eta \frac{\partial L}{\partial w_n}$$

$$w_{\text{new}} = w_{\text{old}} - \eta \left[ \frac{\Delta L}{\Delta w} \right] \rightarrow \text{gradient} \quad \left[ \begin{array}{c} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{array} \right]$$

$\text{L}_g \text{W}$

$$\begin{aligned} L &= \frac{1}{2} (xw - y)^T (xw - y) + \frac{1}{2} \lambda w^T w \\ &= \frac{1}{2} (w^T x^T - y^T) (xw - y) + \frac{1}{2} \lambda w^T w \\ &= \frac{1}{2} \left[ w^T x^T x w - \cancel{w^T x^T y} - \cancel{y^T w x} + y^T y \right] + \frac{1}{2} \lambda w^T w \end{aligned}$$

$$= \frac{1}{2} L^{\text{vv}} \wedge \wedge^{\text{vv}} \quad \boxed{-L' - J}$$

$$= \frac{1}{2} \left[ \underbrace{w^T x^T x w}_{2w^T x^T y} - \underbrace{2y^T w x}_{2y^T w x + y^T y} + \frac{1}{2}\lambda \underline{w^T w} \right]$$

$$\frac{dL}{dw} = \frac{1}{2} \left[ \cancel{2x^T x w} - \cancel{2y^T x} \right] + \frac{1}{2} \cancel{\lambda w}$$

$$= \boxed{x^T x w - y^T x + \lambda w} = \frac{dL}{dw} \left( \frac{\Delta L}{\Delta w} \right)$$

$w = \begin{bmatrix} w_0 & w_1 & \dots & w_n \\ 0 & 1 & \dots & 1_m \end{bmatrix}$  starting

Epochs  $w = w - n \frac{dL}{dw}$

$w \rightarrow \text{final answer}$

epoch times

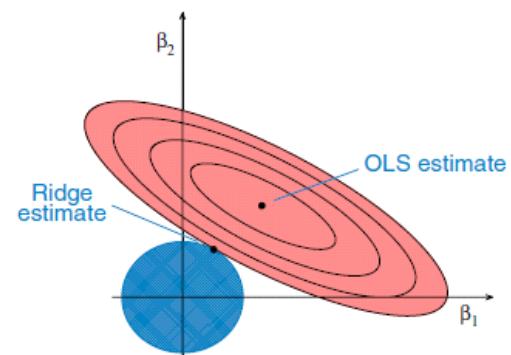
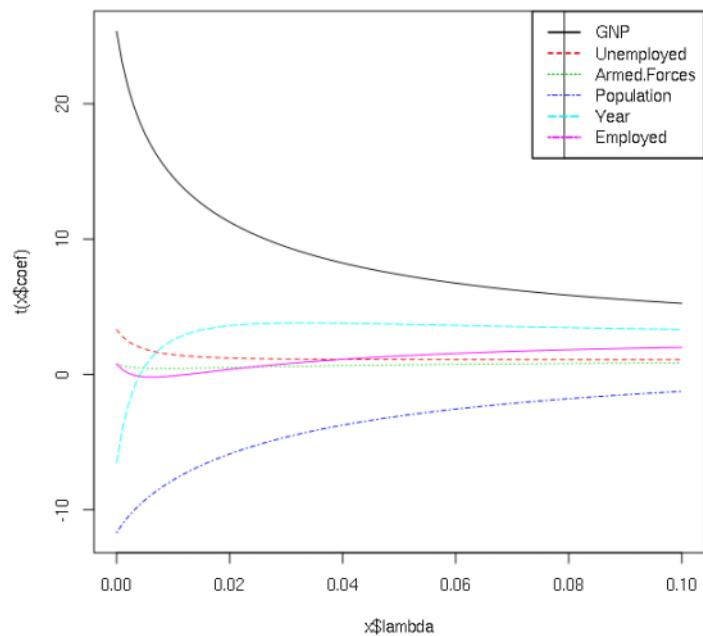
$$\hookrightarrow w = w - n \frac{dL}{dw}$$

$$\boxed{\frac{dL}{dw} = \cancel{x^T x w} - \cancel{x^T y} + \cancel{\lambda w}}$$

# Notes

Friday, June 4, 2021 4:47 PM

Why is it called ridge



## 5 Key Understandings

Saturday, June 5, 2021 4:20 PM

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \boxed{\lambda \|w\|^2}$$

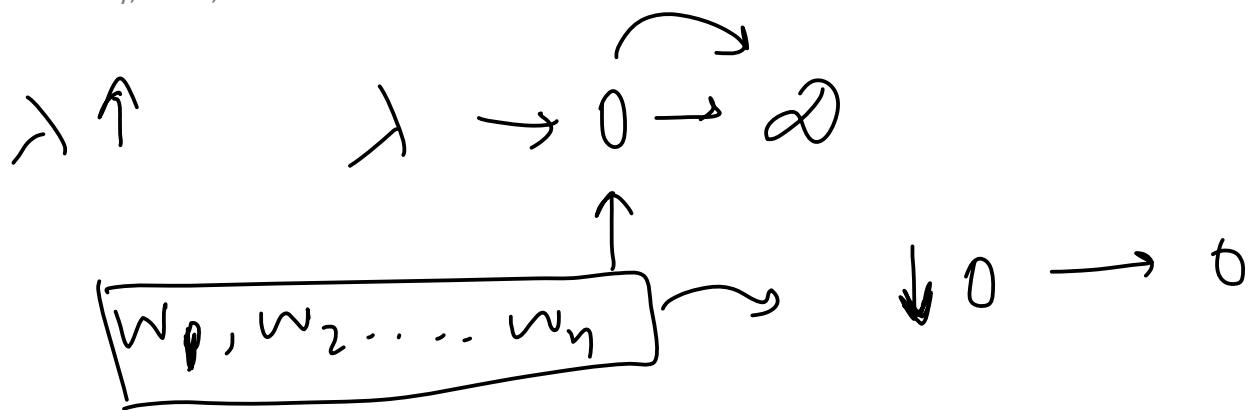
$\lambda (w_1^2 + w_2^2 + \dots + w_n^2)^2$

Shrinkage  
coef → Overfitting ↘

The diagram illustrates the Ridge Regression cost function. It starts with the sum of squared residuals term  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ , which is highlighted with a bracket labeled "Shrinkage". This term is followed by a plus sign and a regularization term  $\lambda \|w\|^2$ , which is enclosed in a box and labeled "coef". A bracket under the regularization term indicates it is equivalent to  $\lambda (w_1^2 + w_2^2 + \dots + w_n^2)^2$ . An arrow points from the word "Overfitting" towards the regularization term, suggesting that the regularization helps prevent overfitting.

# 1. How the coefficients get affected?

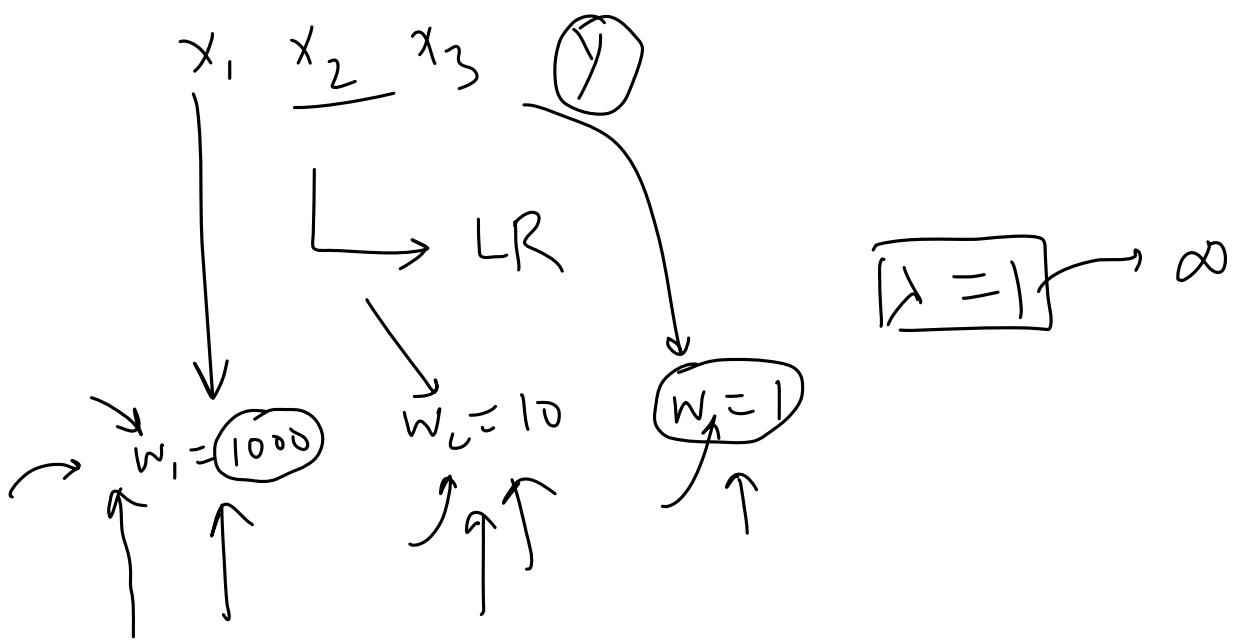
Saturday, June 5, 2021 4:20 PM



## 2. Higher Values are impacted more

Saturday, June 5, 2021 4:21 PM

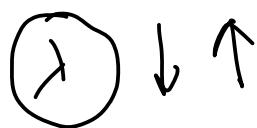
Never reaches 0



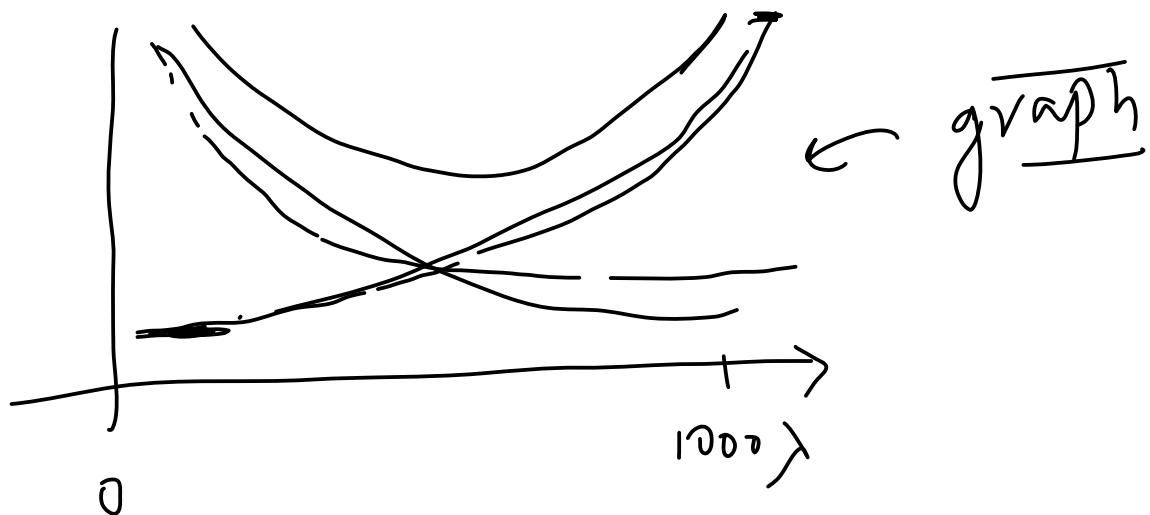
### 3. Bias Variance Tradeoff

Saturday, June 5, 2021 4:21 PM

Bias Variance



Bias ↓ overfit Variance ↑  
Bias ↑ underfitting Variance ↓



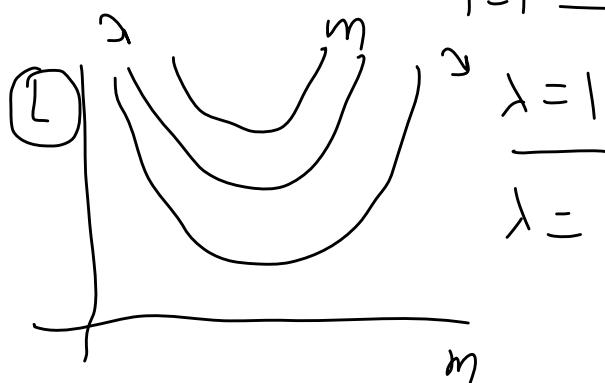
#### 4. Impact on the Loss Function

Saturday, June 5, 2021 4:21 PM

$$\lambda \rightarrow L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \underline{\lambda} \|w\|^2$$

$x, y \rightarrow [m, b] \sim b$  is constant

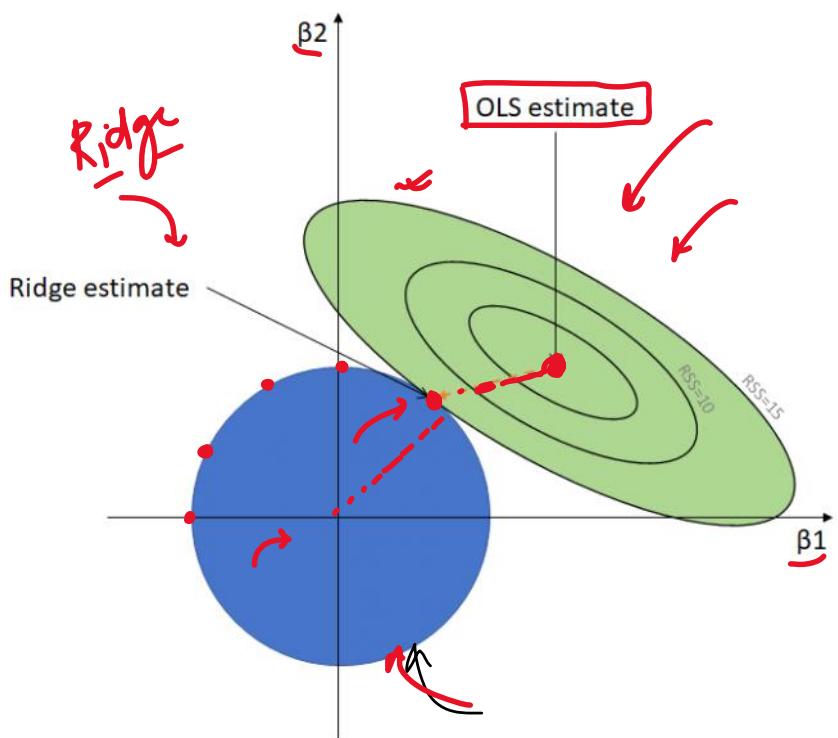
$$m \quad \underline{b=b} \quad L = \sum_{i=1}^n (\hat{y}_i - \underline{m} x_i)^2 + \underline{\lambda m^2}$$



$$b = -2.29$$

## 5. Why called Ridge

Saturday, June 5, 2021 4:22 PM



Hard constraint  
Ridge constraint

$$2 \text{ coef } \beta_1 \beta_2 \beta_0$$

$$L = \text{MSE} + \lambda \|w\|^2$$

contour

$$(y_i - \hat{y}_i)^2$$

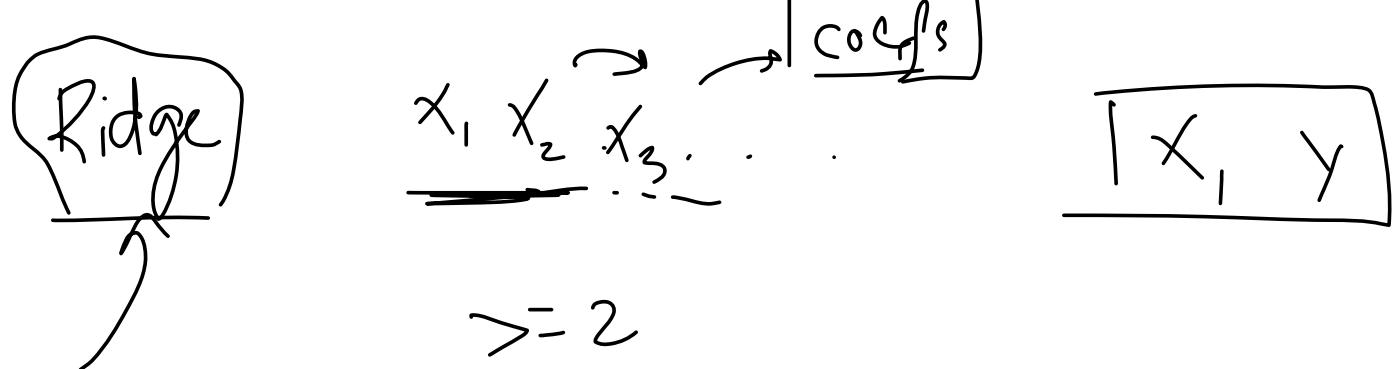
$$\sum_{i=1}^n \frac{(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2}{\text{MSE}}$$

$$\boxed{\lambda (\beta_1^2 + \beta_2^2)}$$

## Practical Tip

Monday, June 7, 2021 1:20 PM

Use ridge when there are more than 2 input cols



$\geq 2$

## Lasso Regression

Thursday, June 10, 2021 6:42 AM

L1 Regularization

overfitting

L2 reg

$$y = mx + b \quad \lambda \uparrow \quad \hat{Y} = b$$

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \|\mathbf{w}\|_1$$

$\lambda (w_1^2 + w_2^2 + \dots + w_n^2)$

$$\lambda > 0$$

$$\rightarrow 0 \quad w_1 \rightarrow w_n \rightarrow \underline{\text{coeff}}$$

overfitting | under

alpha

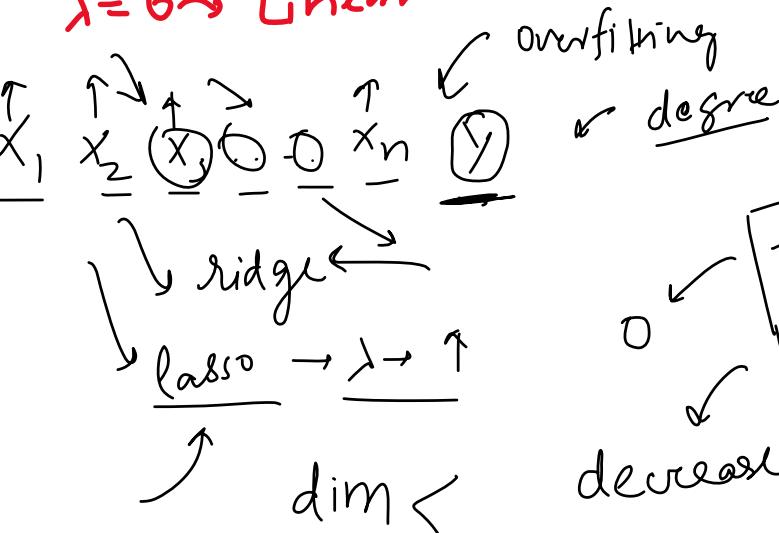
Lasso

L1 norm

$|w_1| + |w_2| + |w_3| + \dots + |w_n|$

underfitting

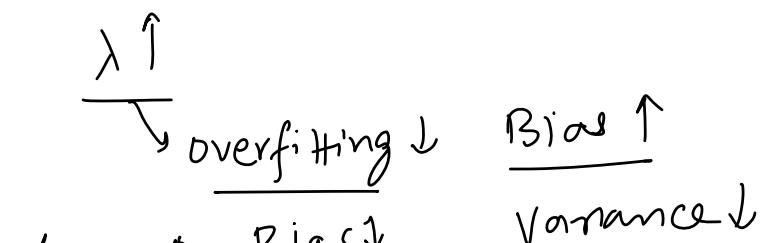
$\lambda = 0 \rightarrow \text{Linear}$



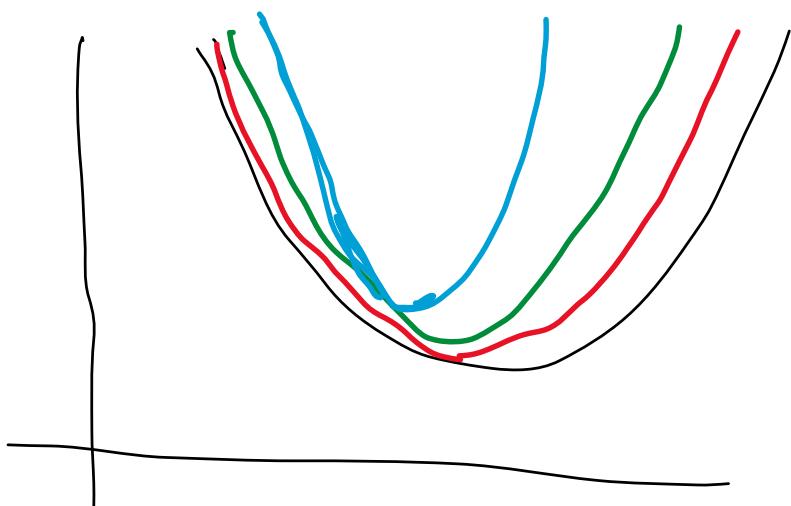
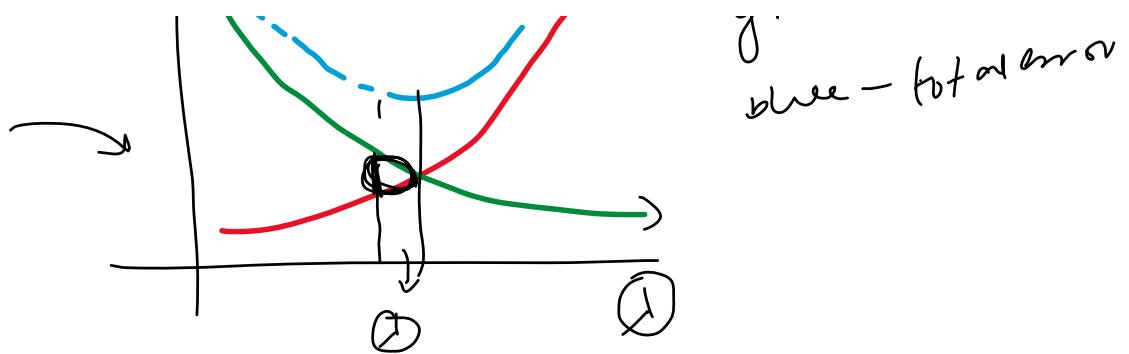
→ code implement (sklearn)

→ key point →

→ difference Ridge Vs Lasso



Red → bias  
green → variance  
blue → total error



## Understanding Sparsity

Friday, June 11, 2021 11:21 AM

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543609	561.214533	153.884786	126.734316	861.121400	52.419828
0.0001	-9.118336	-205.337133	516.880570	340.556792	-883.415291	551.553259	148.578680	125.355917	856.480254	52.467627
0.0010	-8.763583	-204.321125	518.371729	339.975385	-787.690766	475.274718	106.786540	114.632063	819.739542	52.872100
0.0100	-6.401088	-198.669767	522.048548	336.348363	-383.709187	152.663678	-66.060583	75.611090	659.869402	55.828128
0.1000	6.642753	-172.242166	485.523872	314.682122	-72.939323	-80.590053	-174.466515	83.616653	484.363285	73.584154
1.0000	42.242217	-57.305508	282.170831	198.061386	14.363544	-22.551274	-136.930053	102.023193	260.104308	98.552274
10.0000	21.174004	1.659796	63.659772	48.493240	18.421492	12.875448	-38.915435	38.842464	61.612405	35.505355
100.0000	2.858979	0.629452	7.540604	5.849997	2.710879	2.142134	-4.834047	5.108223	7.448466	4.576129
1000.0000	0.295726	0.069290	0.769004	0.597829	0.282900	0.225936	-0.495607	0.527031	0.761497	0.471029
10000.0000	0.029674	0.006995	0.077054	0.059915	0.028412	0.022715	-0.049686	0.052870	0.076321	0.047241

Ridge

Lasso  
sparsity

$\lambda \uparrow \quad w \rightarrow 0$

single  $x | y \rightarrow$

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543596	561.214523	153.884780	126.734314	861.121395	52.419828
0.0001	-9.071288	-205.337332	516.780313	340.539730	-888.652320	555.952271	150.585260	125.453044	858.639860	52.379002
0.0010	-8.264924	-204.213177	517.641106	339.751339	-826.653342	508.609613	120.899583	113.924518	836.314382	52.011583
0.0100	-1.361404	-192.944226	526.348511	332.649058	-430.205495	191.277876	-44.048113	68.990747	688.384976	47.939528
0.1000	0.000000	-113.976046	526.737112	292.635423	-82.691928	-0.000000	-152.691332	0.000000	551.077200	7.169852
1.0000	0.000000	0.000000	363.882636	27.278420	0.000000	0.000000	-0.000000	0.000000	336.135971	0.000000
10.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
100.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
1000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
10000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000

feature selection

$$m = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2 + \lambda}$$

simple

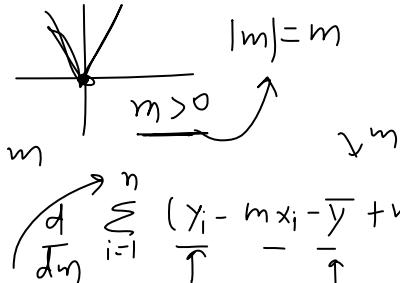
$$y = m x + b$$

$$b = \bar{y} - m \bar{x}$$

$$\bar{y} \rightarrow \text{mean}(y)$$

$$\bar{x} \rightarrow \text{mean}(x)$$

$$m = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



$$b = \bar{y} - m \bar{x}$$

$$m = ?$$

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda |m|$$

$$\frac{d}{dm} \sum_{i=1}^n \frac{(y_i - mx_i - \bar{y} + m\bar{x})^2}{T} + \frac{2\lambda m}{T}$$

$$\frac{d}{dm} L = \sum_{i=1}^n (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda m = \sum_{i=1}^n (y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda m = 0$$

$$m \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) - \lambda$$

- Lasso Regression -

$$-\sum [ (y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2 ] + \lambda = 0$$

$$-\sum (y_i - \bar{y})(x_i - \bar{x}) + m \sum (x_i - \bar{x})^2 + \lambda = 0$$

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

Lasso Coeff Sparsity

for  $m > 0$

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

$$m = \frac{(YX) - \lambda}{X^2}$$

$$\left\{ \begin{array}{l} YX = 100 \\ X^2 = 50 \end{array} \right.$$

$$m = \frac{100 - \lambda}{50}$$

$$\lambda = 0 \quad m = 2 \quad m = \frac{9}{5}$$

$$\lambda = 50 \quad m = 1 \quad \lambda > 100 \quad m = -1$$

$$m < 0 \quad m = \frac{\sum (y_i - \bar{y})(x_i - \bar{x}) + \lambda}{\sum (x_i - \bar{x})^2}$$

$$\lambda > 0$$

$$m = -\frac{100 + \lambda}{50}$$

$$\lambda = 0 \quad m = -2$$

$$\lambda = 50 \quad m = -1$$

$$\lambda = 100 \quad m = 0 \rightarrow 1$$

$$\lambda = 150 \quad m = -5$$

$$m = 1$$

for  $m < 0$

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

$$m = \frac{YX + \lambda}{X^2} = \frac{100 + \lambda}{50}$$

$$= \frac{100 + 150}{50} =$$

$$(m = \bar{m})$$

$$m = -\frac{100 - \lambda}{50} \leftarrow$$

$$= -\frac{100 - 150}{50} = -5$$

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

$$0 \quad \lambda \rightarrow \infty \quad \text{Ridge} \quad \lambda = 100000000$$

Denominator

Lasso  $\lambda \rightarrow \underline{\text{non-zero}}$

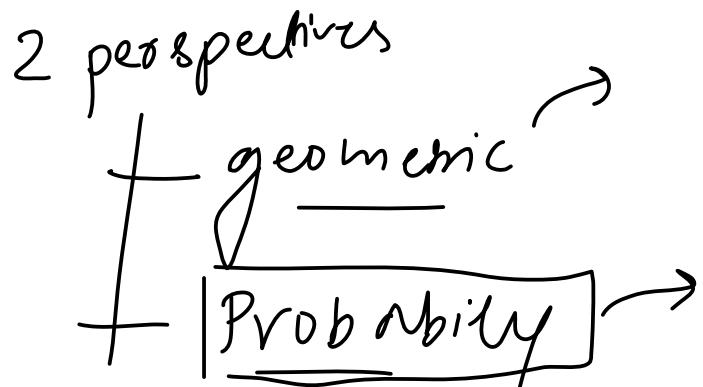
## ElasticNet Regression

Saturday, June 12, 2021 1:04 PM

$\text{Ridge} \quad \lambda(w_1^2 + w_2^2 + \dots + w_n^2)$	$\text{Lasso} \quad \lambda( w_1  +  w_2  + \dots +  w_n )$
$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \ w\ _2^2$ <span style="margin-left: 100px;"><math>\underbrace{\text{mse}}</math></span> <span style="margin-left: 100px;"><math>\underbrace{\text{overfitting}}</math></span>	$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \ w\ _1$ <span style="margin-left: 100px;"><math>\underbrace{\text{mse}}</math></span> <span style="margin-left: 100px;"><math>\lambda \uparrow \quad w \rightarrow 0</math></span> <span style="margin-left: 100px;"><math>\downarrow</math></span> <span style="margin-left: 100px;"><math>\text{feature selection}</math></span>
$\lambda \uparrow \quad w \rightarrow 0$ <span style="margin-left: 100px;"><math>\downarrow</math></span> <span style="margin-left: 100px;"><math>100 \text{ cols} \Rightarrow</math></span>	$\lambda \uparrow \quad w \rightarrow 0$ <span style="margin-left: 100px;"><math>\downarrow</math></span>
$\text{EN Reg} \quad L = \sum (y_i - \hat{y}_i)^2 + \underline{a} \ w\ ^2 + \underline{b} \ w\ $ $\lambda = 1 \quad l1\_ratio = 0.5$ $a = 0.5 \quad b = 0.5$ $\uparrow$ $l1\_ratio > 0.9$	$\left\{ \begin{array}{l} \lambda, l1\_ratio \\ \boxed{l1, \lambda} \end{array} \right.$ $q_0, q_1, \dots, q_{l1\_ratio}$ $\lambda = a + b$ $l1\_ratio = \frac{a}{a+b}$ $l1 = \frac{a}{a+b}$ $a = l1 \times \lambda$ $b = \lambda - a$
$\text{Input cols} \rightarrow \text{multicollinearity} \quad (\text{ElasticNet})$ $x_1 \quad   \quad x_2$	

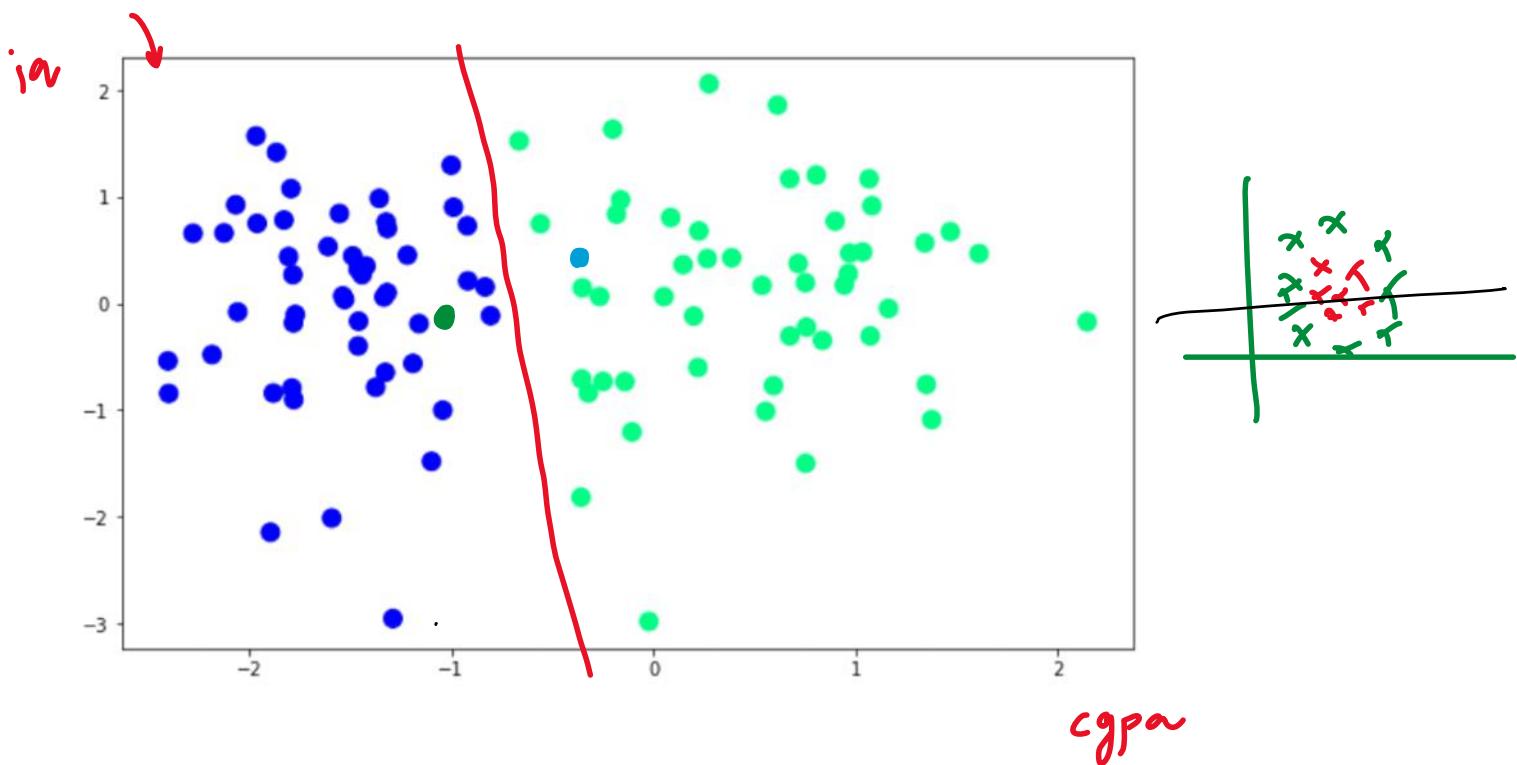
# Introduction

Tuesday, June 15, 2021 12:07 PM



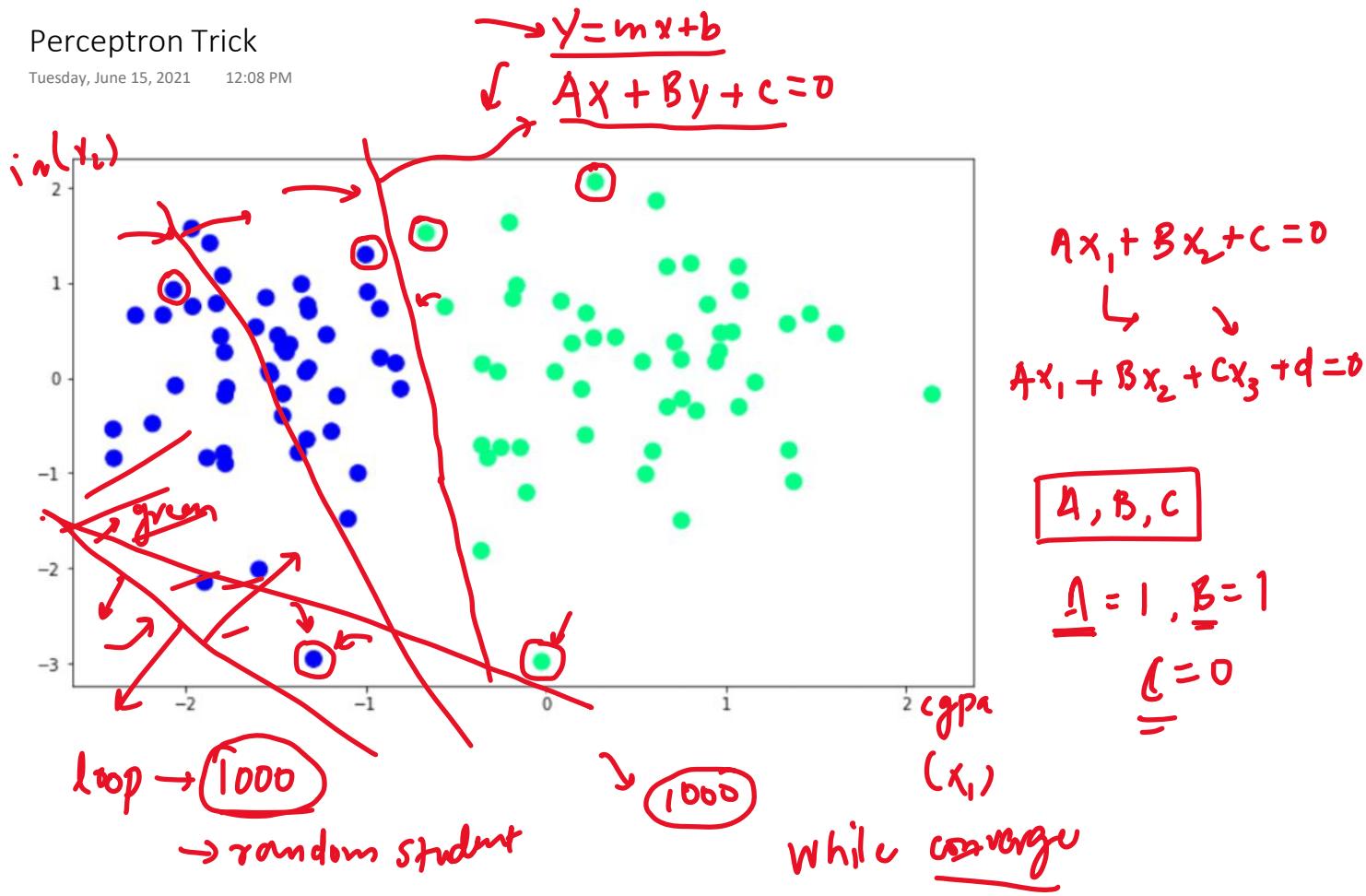
# Requirement

Tuesday, June 15, 2021 12:07 PM



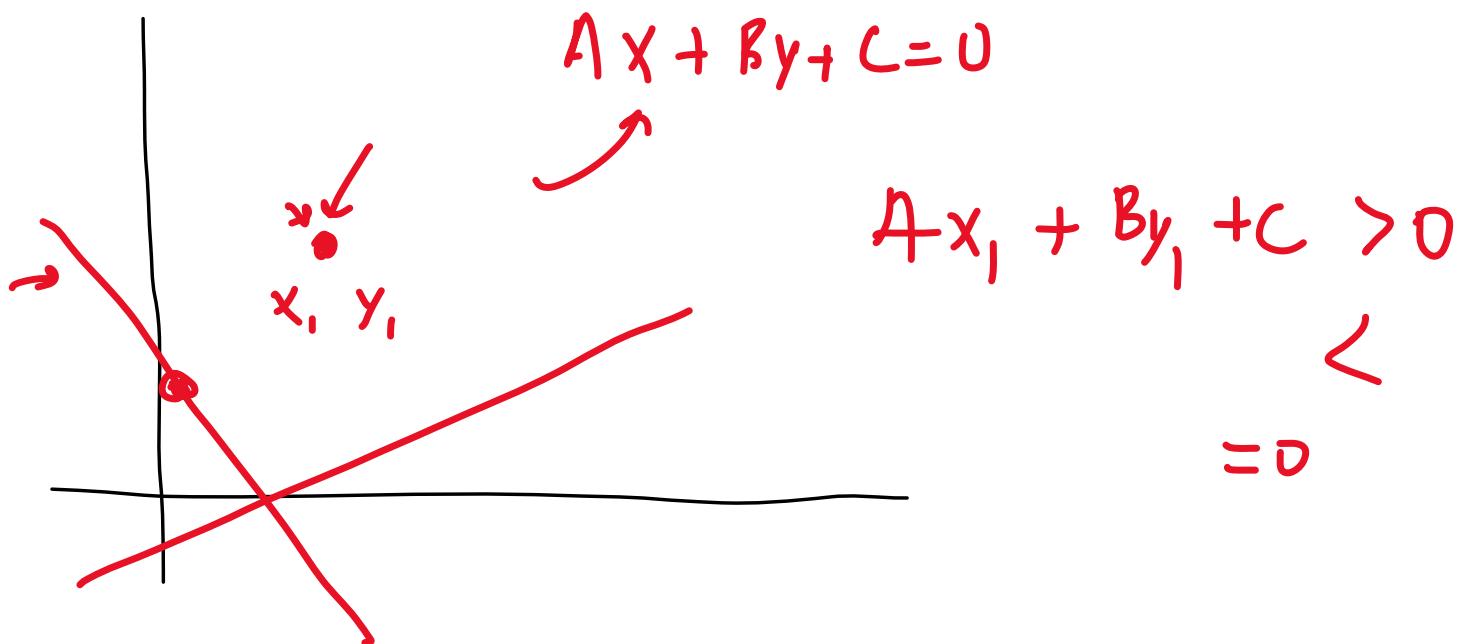
## Perceptron Trick

Tuesday, June 15, 2021 12:08 PM



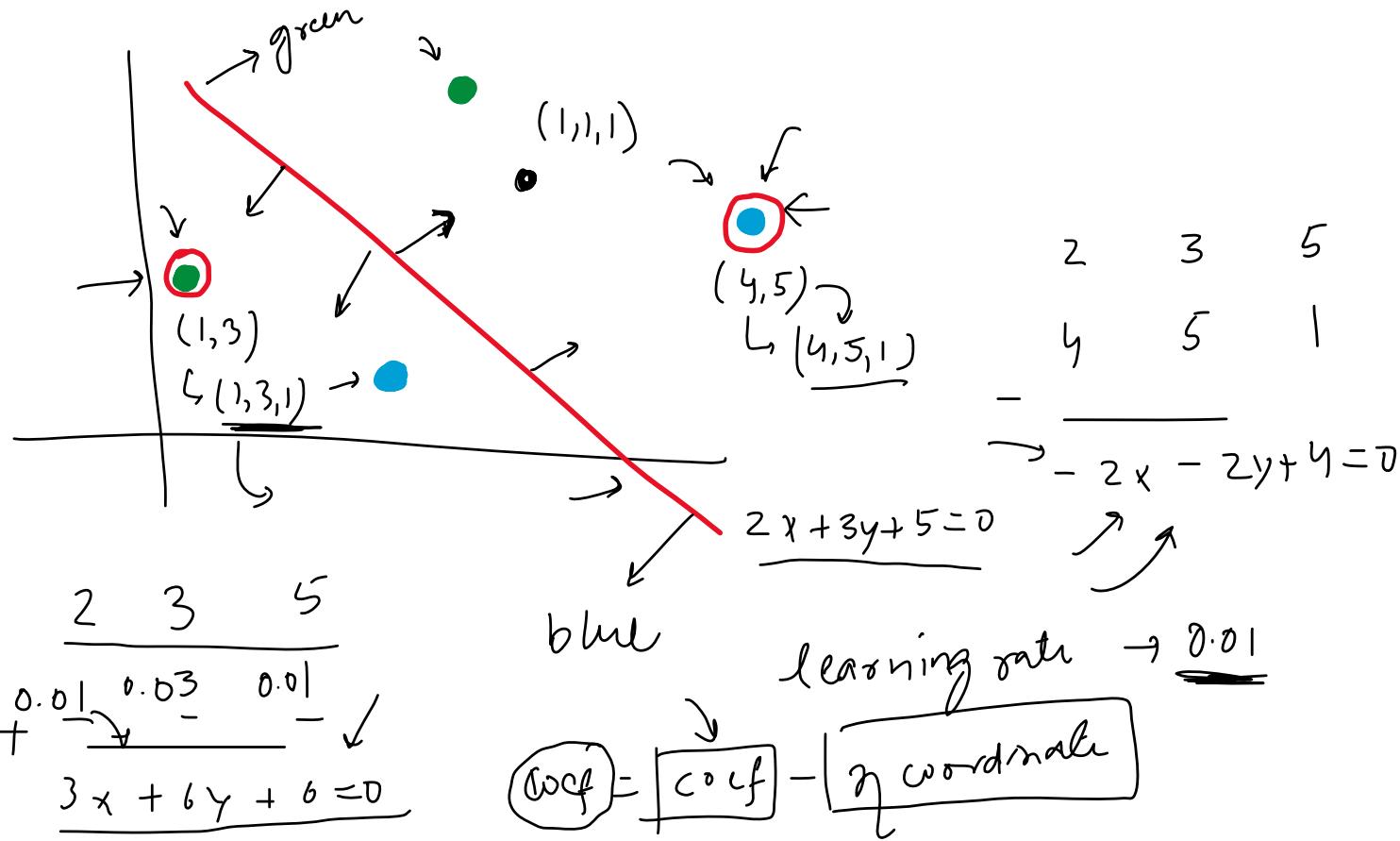
## How to label regions?

Tuesday, June 15, 2021 1:12 PM



## Transformations

Tuesday, June 15, 2021 1:31 PM



## Algorithm

	$x_0$ ( $x_1$ )	$x_2$ )	$y$ placd
	CGPA	12	
1	7.5	61	①
1	8.9	109	1
1	7.0	81	0

Tuesday, June 15, 2021

2:31 PM

$$Ax + By + C = 0$$

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

$$w_0 = C, \quad w_1 = A, \quad w_2 = B$$

$$w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$$

$$w_0 \times 1 + w_1 \times 7.5 + w_2 \times 8 \rightarrow \sum_{i=0}^2 w_i x_i = 0 \quad [w_0 \ w_1 \ w_2] \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$$

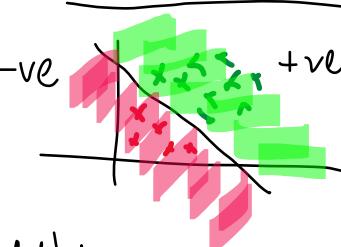
$$= \geq 0 \rightarrow ①$$

$$w_0 x_0 + w_1 x_1 + w_2 x_2$$

$$< 0 \rightarrow 0$$

$$-ve \quad +ve \quad \hookrightarrow 0$$

$$\text{Epoch} \rightarrow 1000, \eta = 0.01$$



$$x_i \in P \quad y_i \in N$$

for i in range (epochs):

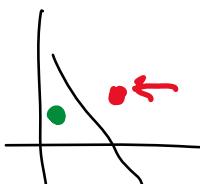
randomly select a student

if  $x_i \in N$  and  $\sum_{i=0}^2 w_i x_i \geq 0$  {

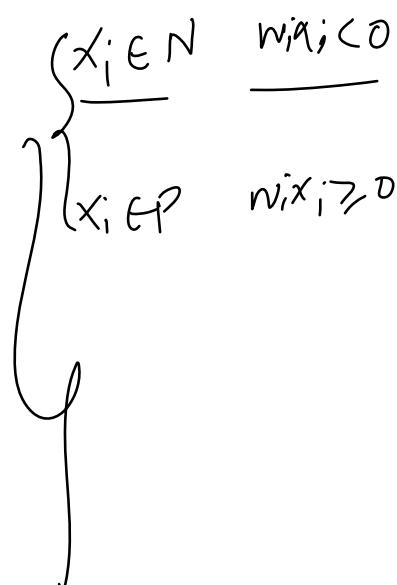
$$w_{new} = w_{old} - \eta x_i$$

if  $x_i \in P$  and  $\sum_{i=0}^2 w_i x_i < 0$

$$w_{new} = w_{old} + \eta x_i$$



$$[w_0, w_1]$$



## Simplified Algo

Tuesday, June 15, 2021 2:44 PM

```

if  $x_i \in N$  and  $\sum w_i x_i \geq 0$ 
     $w_n = w_0 - \eta x_i$ 
if  $x_i \in P$  and  $\sum w_i x_i < 0$ 
     $w_n = w_0 + \eta x_i$ 

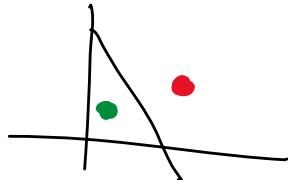
```

for  $i$  in 1000  
 random student  
 $w_n = w_0 + \eta(y_i - \hat{y}_i)x_i$

$w_n = w_0$

$w_n = w_0 + \eta x_i$

$x_i$	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
-1	0	1
0	1	-1



$$w_n = w_0 - \eta x_i$$

for  $j$  in range(epochs):  
 select a random student ( $j$ )  
 $w_n = w_0 + \eta (x_j - \hat{y}_j) x_j$

$$Ax + By + C = 0$$

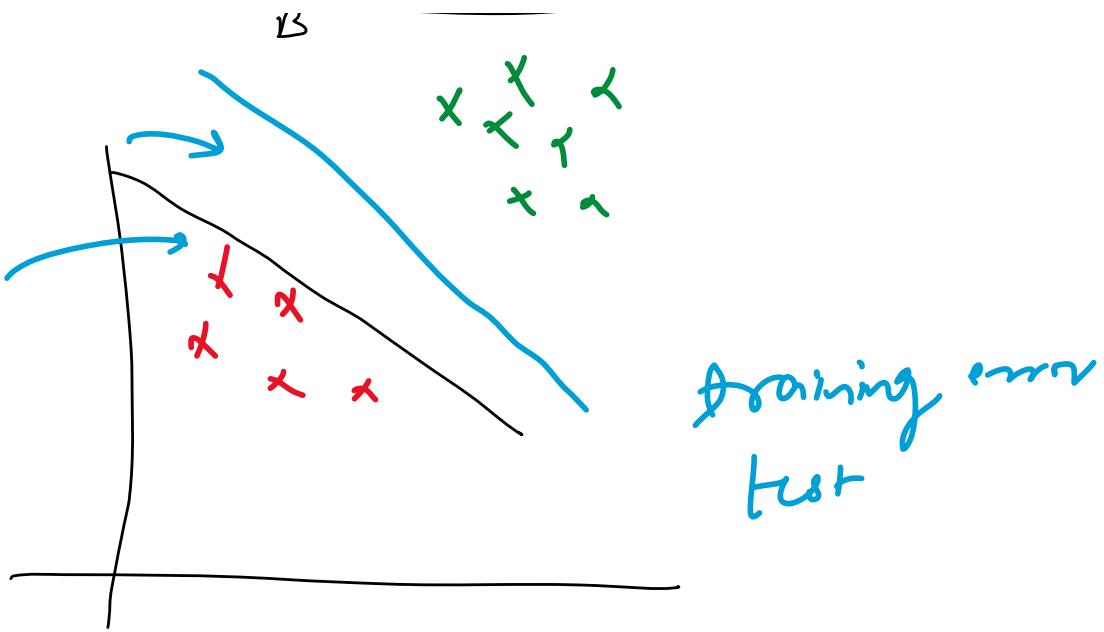
$$y = mx + b$$

$$m = -\frac{A}{B}$$

$$C = -\frac{C}{B}$$

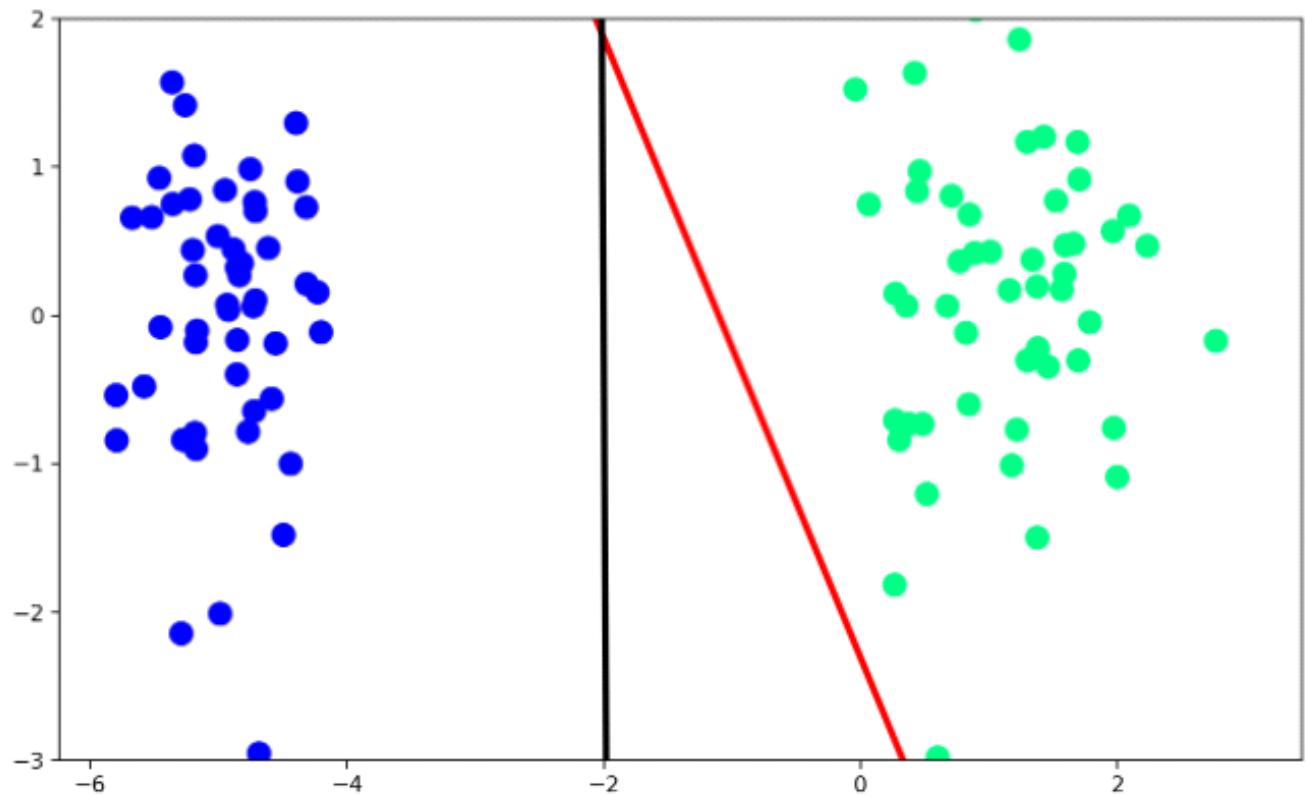
$A, B, C$

✓ ✗ ✗



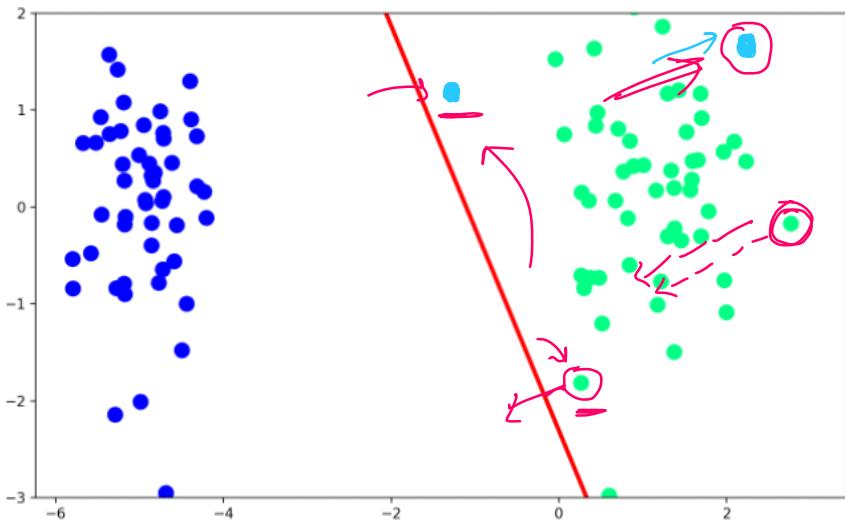
# Problem with Perceptron

Thursday, June 17, 2021 12:18 PM



## Possible Solution?

Thursday, June 17, 2021 12:18 PM



$$w_\eta = w_0 + \eta(y_i - \hat{y}_i)x_i$$

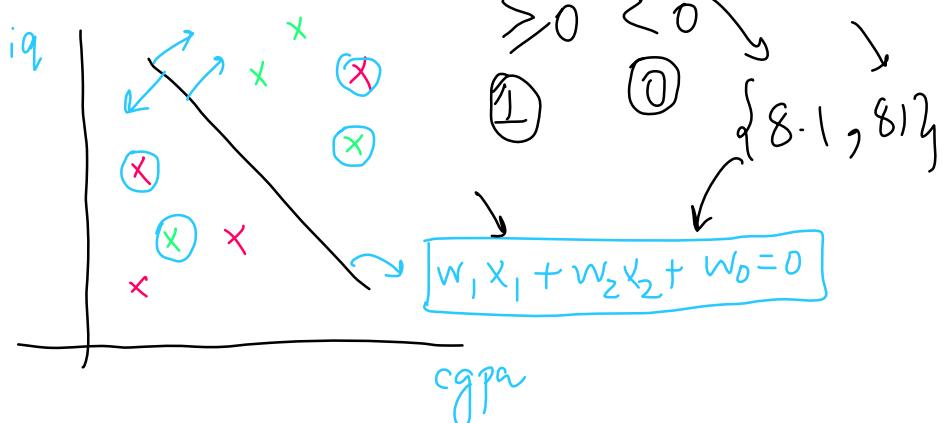
→ misclassified  
line - pull

→ correctly  
line push  
with small

$$w_\eta = w_0 + \eta \underbrace{(y_i - \hat{y}_i)}_{0} x_i$$

$$\rightarrow (y_i - \hat{y}_i) \neq 0 \text{ model predict} \\ \sum w_i x_i = [0, 1]$$

$$w_1 \times 8.1 + w_2 \times 81 + w_0 =$$



$y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
1	0	1
0	1	-1

cgpa	iq	placed
9	91	0
8.8	78	1
8.1	102	1
7.9	98	1

# The Sigmoid Function

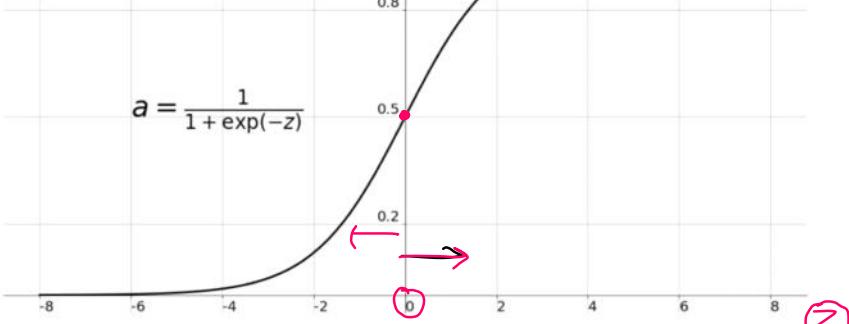
Thursday, June 17, 2021 12:19 PM

## Sigmoid Function

$$y = \sigma(z) \rightarrow 1$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$a = \frac{1}{1 + \exp(-z)}$$



$$-\infty < z < \infty$$

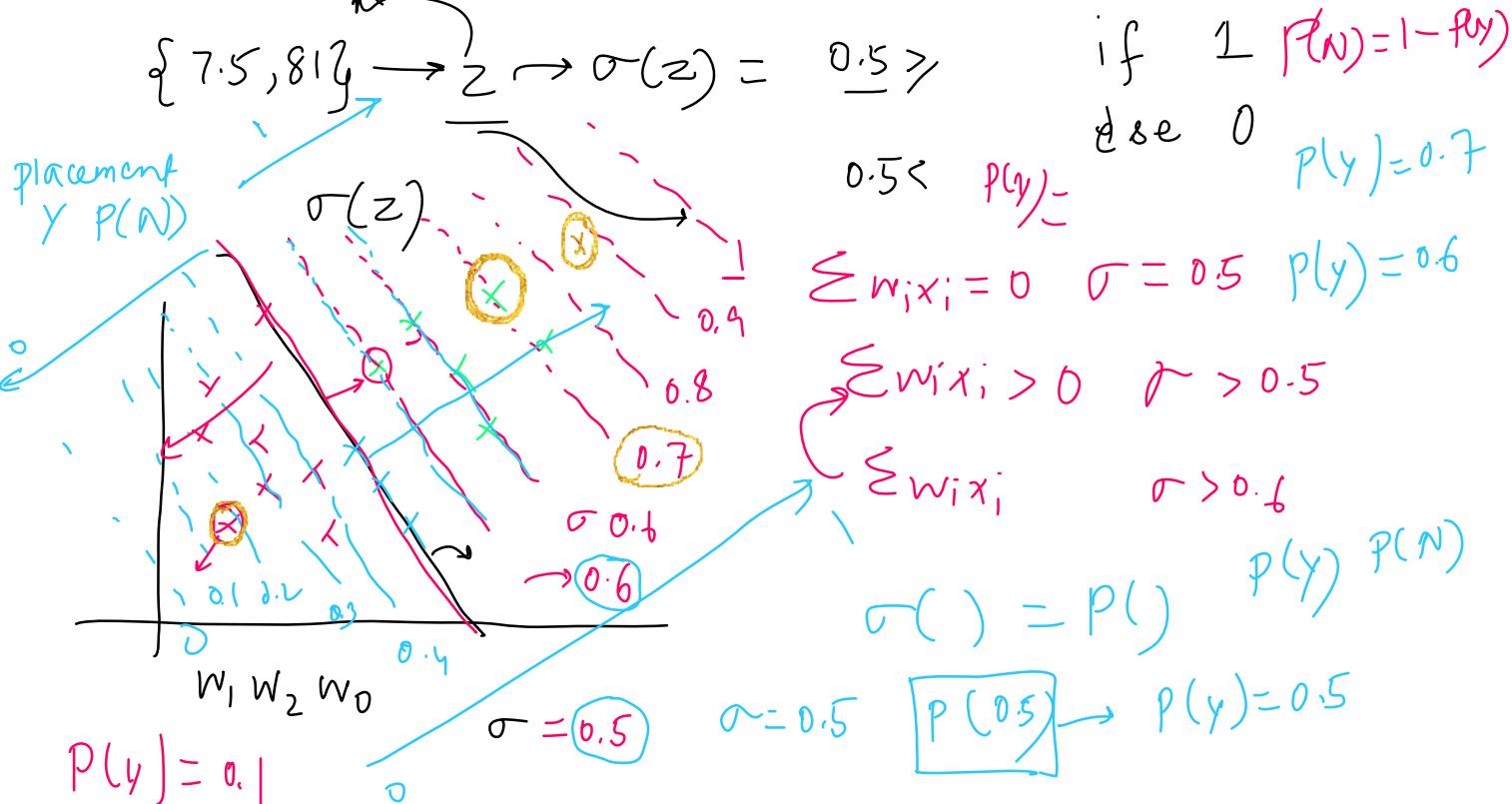
$$0 < y < 1$$

$$y_i = \frac{\{7.5, 81\} \rightarrow w_1 w_2 w_0}{w_1 \times 7.5 + w_2 \times 81 + w_0} = \sum w_i x_i = \sigma(z)$$



$$\begin{cases} z \geq 0 \rightarrow 1 \\ z < 0 \rightarrow 0 \end{cases}$$

$$\sigma(z) < 0.5$$



## Impact of Sigmoid

Thursday, June 17, 2021 3:27 PM

$$w_n = w_0 + \eta (y_i - \hat{y}_i) x_i$$

$$\hat{y}_i = \sigma(z)$$

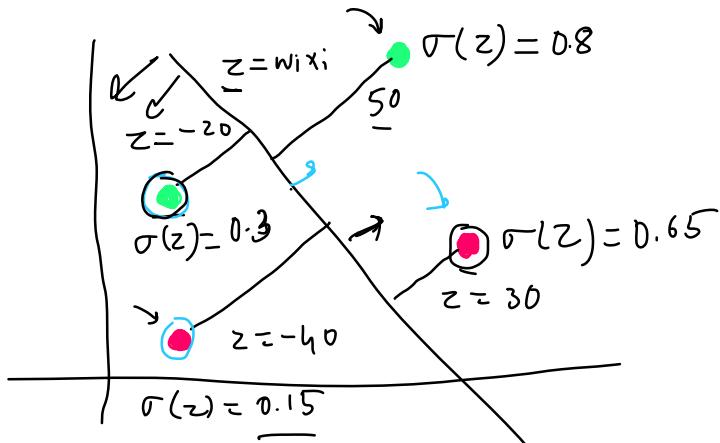
$$\text{where } z = \sum w_i x_i$$

$$w_n = w_0 + \eta \times 0.2 \times x_i$$

$$w_n = w_0 - \eta \times 0.65 \times x_i$$

$$w_n = w_0 + \eta \times 0.7 \times x_i$$

$$w_n = w_0 - \eta \times 0.15 \times x_i$$



$y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.8	0.2
0	0.65	-0.65
1	0.3	0.7
0	0.15	-0.15

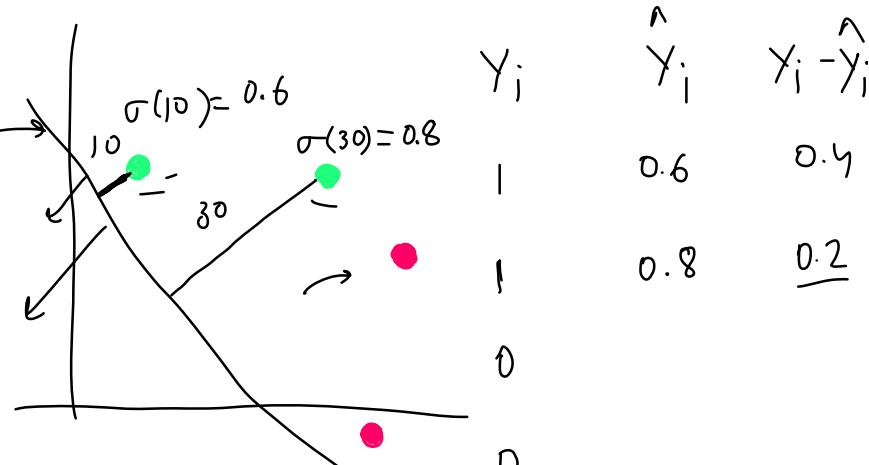
$$w_n = w_0 + \eta (y_i - \hat{y}_i) x_i$$

$$\hat{y}_i = \sigma(z)$$

$$\text{where } z = \sum w_i x_i$$

$$w_n = w_0 + [\eta \times 0.4 \times x_i] = x_1$$

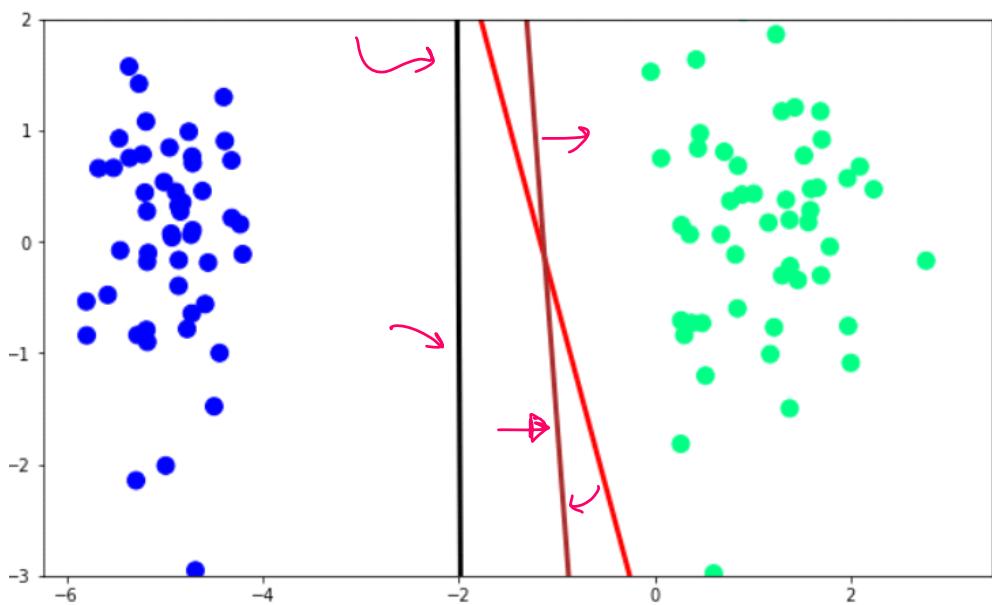
$$w_n = w_0 + [\eta \times 0.2 \times x_i] = x_2$$



$x_1 > x_2 \rightarrow \underline{\text{code implement}}$

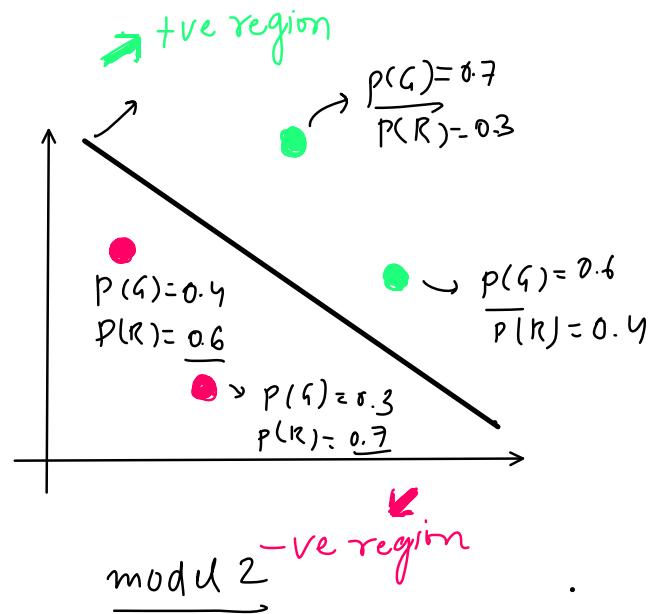
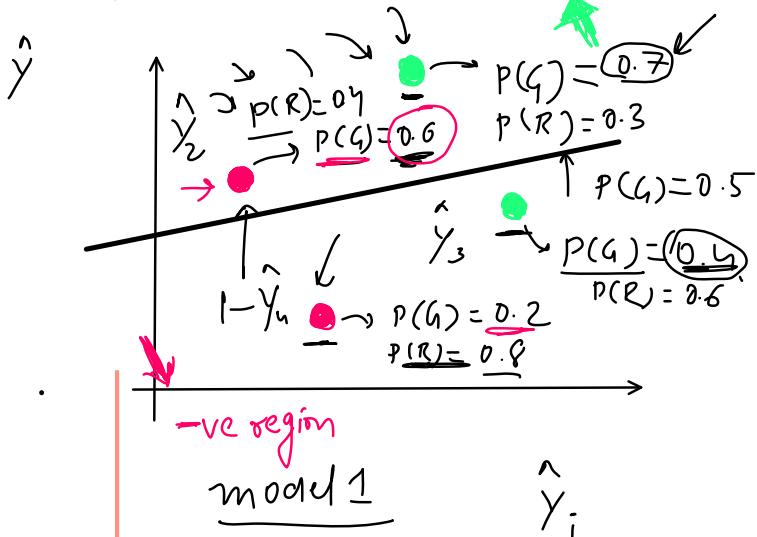
## The Problem

Friday, June 18, 2021 2:20 PM



# Maximum Likelihood and Cross-Entropy

Thursday, June 17, 2021 12:19 PM



$$\hat{y} = \sigma(z)$$

$$z = \sum w_i x_i$$

$$\text{model 1} \rightarrow \frac{0.7 \times 0.4 \times 0.4 \times 0.8}{0.089}$$

$$\text{model 2} \rightarrow 0.7 \times 0.6 \times 0.6 \times 0.7 \\ = 0.176$$

$$\log(a+b) = \log a + \log b$$

$$\log(\max) = -\log(0.7) - \log(0.4) - \log(0.4) - \log(0.8)$$

$0-1 = -ve$

$$\log(0.1) > \log(0.9)$$

$\downarrow$

$\rightarrow 0.04$

Cross entropy minimize

$$-\log(\hat{y}_1) - \log(\hat{y}_2) - \log(\hat{y}_3) - \log(\hat{y}_4)$$

$(1-\hat{y})$

$$\underline{y_i = 1}$$

$$-y_i \log(\hat{y}_i) - (1-y_i) \log(1-\hat{y}_i)$$

$$-y_i \log(\hat{y}_i) = -\log(\hat{y}_i)$$

$$= -\log(0.7)$$

$$y_2 = 0$$

$$y_3 = 1$$

$$L = \sum_{i=1}^n -y_i \log(\hat{y}_i) - (1-y_i) \log(1-\hat{y}_i)$$

MSE

$$L = -\frac{1}{n} \sum_{i=1}^n y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

closed  
form  
gradient  
descent

min  
 $w_1, w_2, w_0$

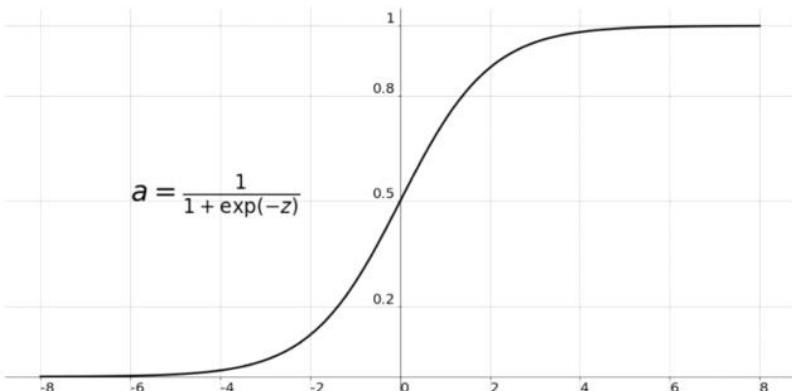
log-loss error

binary cross entropy

## Derivative of Sigmoid

Thursday, June 17, 2021 12:20 PM

## Sigmoid Function



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma'(x) = \frac{d}{dx} \left( \frac{1}{1 + e^{-x}} \right)$$

$$\frac{d}{dx} \left( \frac{1}{x} \right) = \frac{d}{dx} (x)^{-1}$$

$$= -x^{-2} = -\frac{1}{x^2}$$

$$\frac{d}{dx} \left[ \frac{1}{1 + e^{-x}} \right] = \frac{d}{dx} \left[ (1 + e^{-x})^{-1} \right] = -\frac{1}{(1 + e^{-x})^2} \frac{d}{dx} (1 + e^{-x})$$

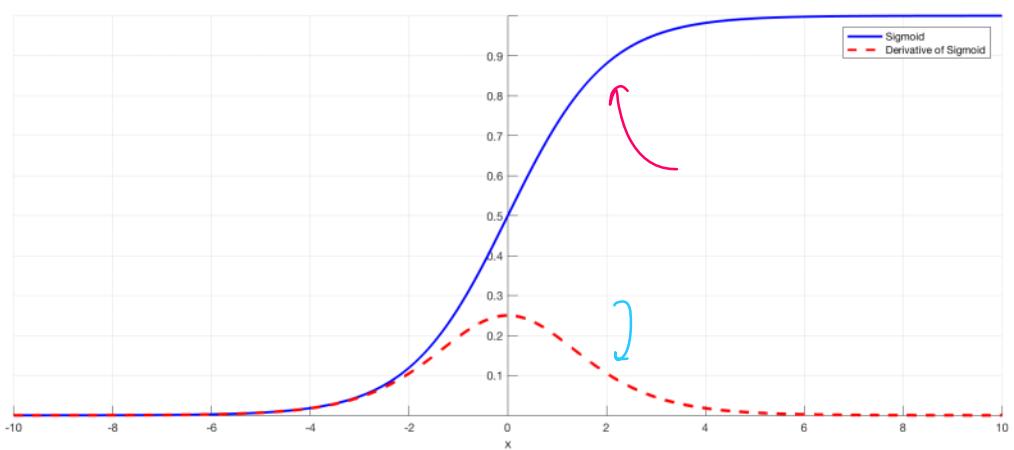
$e^{-x} = e^{-x}$        $-x = -1$

$$= -\frac{1}{(1 + e^{-x})^2} \frac{d}{dx} (e^{-x}) = -\frac{e^{-x}}{(1 + e^{-x})^2} \frac{d}{dx} (-x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$\frac{1 \cdot e^{-x}}{(1 + e^{-x})(1 + e^{-x})} = \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} = \sigma(x) \left[ \frac{e^{-x}}{1 + e^{-x}} \right]$$

$$= \sigma(x) \left[ \frac{1 + e^{-x} - 1}{1 + e^{-x}} \right] = \sigma(x) \left[ \frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}} \right]$$

$$\sigma(x) [1 - \sigma(x)] \Rightarrow \sigma'(x) = \boxed{\sigma(x) [1 - \sigma(x)]}$$



## Gradient Descent

Monday, June 21, 2021 11:50 AM

Classification  $\{x_1, x_2\} \rightarrow y$

$$w_0 w_1 x_1 + w_2 x_2 + w_0 = 0$$

$$\hat{y}_i = \sigma(z) = \sigma(\sum_{j=0}^n w_j x_j)$$

$$L = -\frac{1}{m} \sum_{i=1}^m y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

$$L(w_0, w_1, w_2) = \arg \min_{(w_0, w_1, w_2)}$$

gradient descent  
for i in epochs  
 $w_{new} = w_{old} - \eta \frac{\partial L}{\partial w_{old}}$

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}, \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1}, \quad w_2 = w_2 - \eta \frac{\partial L}{\partial w_2}$$

$w_j = w_j - \eta \frac{\partial L}{\partial w_j} \quad j=0, 1, 2, \dots, n$

$$L = -\frac{1}{m} \sum_{i=1}^m y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^m \left[ \frac{\partial L}{\partial w_j} (y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)) \right]$$

$$\hat{y}_i = \sigma(z) = \sigma(\sum_{j=0}^n w_j x_j)$$

$$\frac{\partial L}{\partial w_j} y_i \log(\hat{y}_i) = \frac{\partial L}{\partial w_j} y_i \log(\sigma(z)) - y_i$$

$$\frac{\partial L}{\partial w_j} = \frac{\partial L}{\partial z} \sigma(z) \sum_{i=1}^n w_j x_i$$

$$\frac{\partial L}{\partial w_j} = \frac{y_i - \hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} = \frac{y_i}{\hat{y}_i} \frac{\partial L}{\partial w_j} \sigma(\sum_{j=0}^n w_j x_j)$$

$$\frac{\partial L}{\partial w_j} = y_i \log(\sigma(\sum_{j=0}^n w_j x_j))$$

$$= y_i \frac{\partial L}{\partial w_j} \underbrace{\log(\sigma(\sum_{j=0}^n w_j x_j))}_{y_i(1-\hat{y}_i)}$$

$$= y_i \frac{\hat{y}_i(1-\hat{y}_i)}{\hat{y}_i} \frac{\partial L}{\partial w_j} \sum_{j=0}^n w_j x_j$$

$$y_i(1-\hat{y}_i) \sum_{j=0}^n \frac{\partial L}{\partial w_j} w_j x_j = \boxed{y_i(1-\hat{y}_i) \sum_{j=0}^n x_j}$$

$$\frac{\partial L}{\partial w_j} \frac{(1-y_i)}{\hat{y}_i} \log(1-\hat{y}_i) \Rightarrow (1-y_i) \frac{\partial L}{\partial w_j} \log(1-\hat{y}_i) \quad \hat{y}_i = \sigma(z)$$

$$\sigma(\sum_{j=0}^n w_j x_j)$$

$$\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \frac{(1-\hat{y}_i)}{\hat{y}_i} \Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \hat{y}_i \Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \sigma(z)$$

$$\Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \sigma(z) \frac{(1-\sigma(z))}{\sigma(z)} \frac{\partial L}{\partial w_j} \sigma(z) = -\frac{(1-y_i)}{(1-\hat{y}_i)} \hat{y}_i(1-\hat{y}_i) \frac{\partial L}{\partial w_j} \sigma(z)$$

$$\Rightarrow -\hat{y}_i(1-y_i) \frac{\partial L}{\partial w_j} \sum_{j=0}^n w_j x_j \Rightarrow -\hat{y}_i(1-y_i) \sum_{j=0}^n \frac{\partial L}{\partial w_j} (w_j x_j)$$

$$\Rightarrow \boxed{-\hat{y}_i(1-y_i) \sum_{j=0}^n x_j}$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^m \left[ y_i(1-\hat{y}_i) \sum_{j=0}^n x_j - \hat{y}_i(1-y_i) \sum_{j=0}^n x_j \right]$$

$$\Rightarrow -\frac{1}{m} \sum_{i=1}^m \left[ y_i(1-\hat{y}_i) - \hat{y}_i(1-y_i) \right] \sum_{j=0}^n x_j$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^m \left[ y_i (1 - \hat{y}_i) - \hat{y}_i (1 - y_i) \right] \sum_{j=0}^n x_j$$

$$= -\frac{1}{m} \sum_{j=1}^m \left[ y_i - y_i \hat{y}_i - \hat{y}_i + y_i \hat{y}_i \right] \sum_{j=0}^n x_j$$

$$\boxed{\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i) \sum_{j=0}^n x_j}$$

$$\frac{\partial L}{\partial w_0} = -\frac{1}{m} [1+0] [1+1]$$

$$= -\frac{1}{2} [1][2] = -1$$

$$\begin{array}{cccc} x_1 & x_2 & y_i & \hat{y}_i \\ 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 0 \end{array}$$

$$= -\frac{1}{2} [1] \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

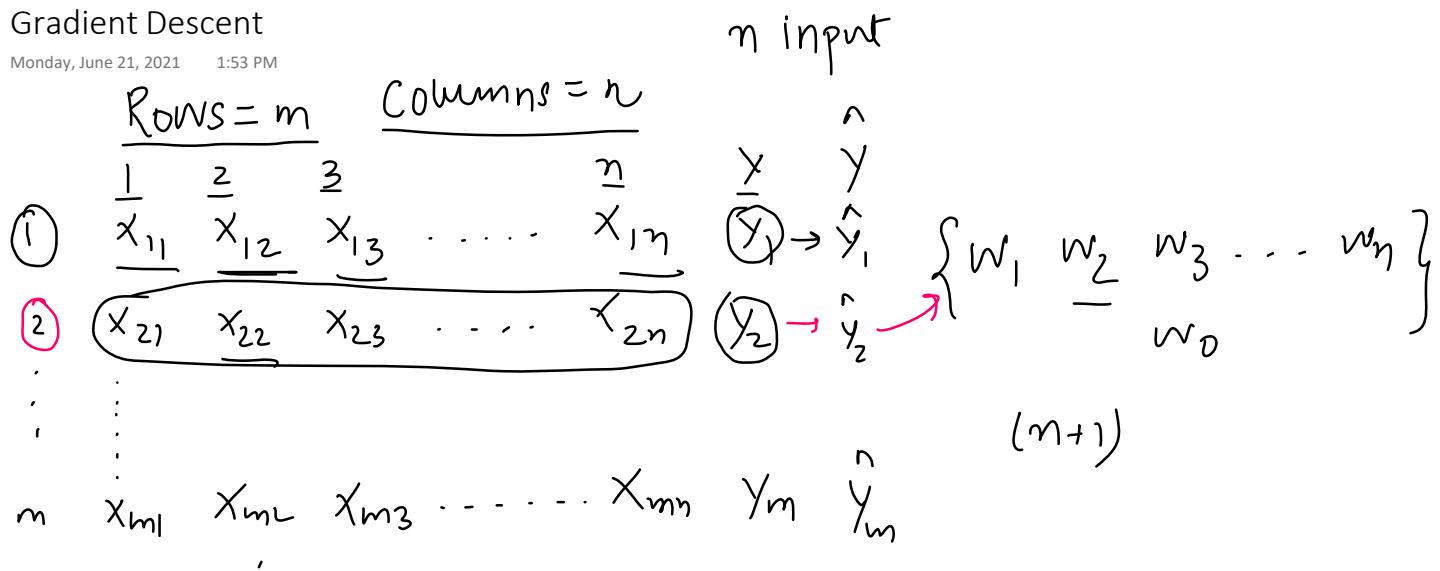
$$= -\frac{1}{2} [1] = -\frac{1}{2}$$

no. of rows = m = 2

no. of cols = n = 2

## Gradient Descent

Monday, June 21, 2021 1:53 PM



$$\sigma(w_0 + w_1 x_{11} + w_2 x_{12} + w_3 x_{13} + \dots + w_n x_{1n} + w_0) = \hat{y}_1$$

$$\sigma(w_0 + w_1 x_{21} + w_2 x_{22} + w_3 x_{23} + \dots + w_n x_{2n} + w_0) = \hat{y}_2$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \sigma(w_0 + w_1 x_{11} + w_2 x_{12} + \dots + w_n x_{1n}) \\ \vdots \\ \sigma(w_0 + w_1 x_{21} + w_2 x_{22} + \dots + w_n x_{2n}) \\ \vdots \\ \sigma(w_0 + w_1 x_{m1} + w_2 x_{m2} + \dots + w_n x_{mn}) \end{bmatrix}$$

$$\hat{Y} = \sigma \left( \begin{bmatrix} c \\ w_0 + w_1 x_{11} + w_2 x_{12} + \dots + w_n x_{1n} \\ w_0 + w_1 x_{21} + w_2 x_{22} + \dots + w_n x_{2n} \\ \vdots \\ w_0 + w_1 x_{m1} + w_2 x_{m2} + \dots + w_n x_{mn} \end{bmatrix} \right)$$

$$\hat{Y} = \sigma \left( \begin{bmatrix} w_0 \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \right)$$

$$\hat{y} = \sigma \left( \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \right)$$

$$\hat{y} = \sigma(xw)$$

$$L = -\frac{1}{m} \sum_{i=1}^m y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

$$L = -\frac{1}{m} \left[ \sum_{i=1}^m y_i \log(\hat{y}_i) + \sum_{i=1}^m (1-y_i) \log(1-\hat{y}_i) \right]$$

$(1-y) \log$   
 $(1-xw)$

$$\sum_{i=1}^m y_i \log(\hat{y}_i) = y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + y_3 \log \hat{y}_3 + \dots + y_m \log \hat{y}_m$$

$$[y_1 \ y_2 \ y_3 \ \dots \ y_m] \begin{bmatrix} \log \hat{y}_1 \\ \log \hat{y}_2 \\ \vdots \\ \log \hat{y}_m \end{bmatrix}$$

$$[y_1 \ y_2 \ y_3 \ \dots \ y_m] \log \left( \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} \right)$$

$$Y \log \hat{Y} = y \log(\sigma(xw))$$

$$L = -\frac{1}{m} [y \log \hat{y} + (1-y) \log(1-\hat{y})]$$

min

where  $\hat{y} = \sigma(xw)$   $\uparrow$  GD  $[w]$  find

where  $\hat{y} = \sigma(xw)$   $L(GD) L^W J J \dots$

Loss function in Matrix form

$$L = -\frac{1}{m} \left[ y \log(\sigma(wx)) + (1-y) \log(1 - \sigma(wx)) \right]$$

minimum

for i in epochs:

$$\rightarrow w = w - \eta \frac{\Delta L}{\Delta w}$$

gradient descent

$$w = [ \quad ]$$

$$\downarrow \downarrow \downarrow \downarrow \downarrow$$

$$w_0 \rightarrow w_n$$

$$\frac{\Delta L}{\Delta w}$$

$$\left\{ \frac{\Delta L}{\Delta w} \right\} \frac{\Delta L}{\Delta w} = \left[ \frac{\partial L}{\partial w_0}, \frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \dots, \frac{\partial L}{\partial w_n} \right]^{(n+1)}$$

$$\frac{\Delta L}{\Delta w} \quad L = -\frac{1}{m} \left[ \underbrace{y \log \hat{y}}_{\rightarrow} + (1-y) \log(1-\hat{y}) \right]$$

$$\frac{dL}{dw} =$$

$$\frac{d}{dw} y \log \hat{y} \Rightarrow y \frac{d}{dw} \log \hat{y} \Rightarrow \frac{y}{\hat{y}} \frac{d}{dL} (\hat{y})$$

$$\Rightarrow \frac{y}{\hat{y}} \frac{d}{dL} \sigma(wx) \Rightarrow \frac{y}{\hat{y}} \underline{\sigma(wx)[1-\sigma(wx)]} \frac{d}{dw} (wx)$$

$$= \frac{y}{\hat{y}} \hat{y} (1-\hat{y}) X = \boxed{Y(1-\hat{y})X} \quad \hat{y} = \sigma(wx)$$

$$= \frac{1}{\hat{y}} \quad \text{L} \quad y - \hat{y}$$

$$\frac{d}{dw} (1-y) \log(1-y) \Rightarrow (1-y) \frac{d}{dw} \log(1-\hat{y}) \Rightarrow (1-y) \frac{\frac{d}{dw} [1-\hat{y}]}{(1-\hat{y})}$$

$$= -\frac{(1-y)}{(1-\hat{y})} \frac{d}{dw} \sigma(wx) \Rightarrow -\frac{(1-y)}{(1-\hat{y})} \left[ \sigma(wx) \left[ 1 - \sigma(wx) \right] \right]$$

$$\Rightarrow -\frac{(1-y)}{(1-\hat{y})} \hat{y} (1-\hat{y}) X = \boxed{-\hat{y} (1-y) X}$$

$$\frac{dL}{dw} = -\frac{1}{m} \left[ y(1-\hat{y})X - \hat{y}(1-y)X \right]$$

$$= -\frac{1}{m} \left[ y(1-\hat{y}) - \hat{y}(1-y) \right] X$$

$$= -\frac{1}{m} \left[ y - y/\hat{y} - \hat{y} + y/\hat{y} \right] X$$

$$\boxed{\frac{\Delta L}{\Delta w} = -\frac{1}{m} (y - \hat{y}) X}$$

gd

$$\boxed{w = w - \eta \frac{1}{m} (y - \hat{y}) X}$$

$$\underline{w} = \underline{w} + \eta \frac{1}{m} (\underline{y} - \hat{\underline{y}}) \underline{x}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad (n+1, 1)$$

$$x = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} \quad (m, 1)$$

$$\underline{w} = \underline{w} + \left[ \frac{\eta}{m} \right] (\underline{y} - \hat{\underline{y}}) \underline{x}$$

$$\underline{w} = \frac{(n+1, 1)}{(n+1, 1)} \quad (1, m) \quad m, (n+1) \rightarrow (1, n+1)$$

$\downarrow$   
 $(n+1, 1)$

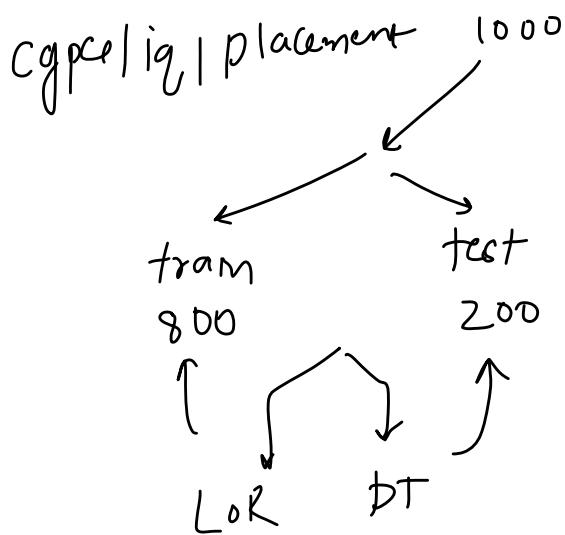
# Accuracy

Tuesday, June 22, 2021 1:12 PM

binary



Actual Label	Logistic Regression Prediction	Decision Tree Prediction
1	✓ 1	✓ 1
0	✗ 1	✗ 1
0	✓ 0	✓ 0
0	✓ 0	✓ 0
1	✓ 1	✓ 1
1	✓ 1	✓ 1
0	✗ 1	✓ 0
0	✓ 0	✓ 0
0	✓ 0	✓ 0
1	✓ 1	✓ 1



$$\text{Accuracy} = \frac{\text{no. of } \checkmark}{\text{total predictions}}$$

$$\frac{9}{16} = 90\%$$

$$\frac{8}{10} = 0.8 \rightarrow 80\%$$

## Accuracy of multi-classification problem

Wednesday, June 23, 2021 8:44 AM



Actual Label	<u>Logistic Regression Prediction</u>	<u>Decision Tree Prediction</u>
0	✓ 0	0
0	✓ 0	0
0	✓ 0	0
2	✓ 2	2
0	✓ 0	0
2	✓ 2	2
0	✓ 0	0
2	✓ 2	2
1	✓ 1	1
1	✓ 1	1

iris  
setosa, virginica / versicolor  
0 1 2

$$\text{accuracy} = \frac{\# \text{ correct predictions}}{\# \text{ total}}$$

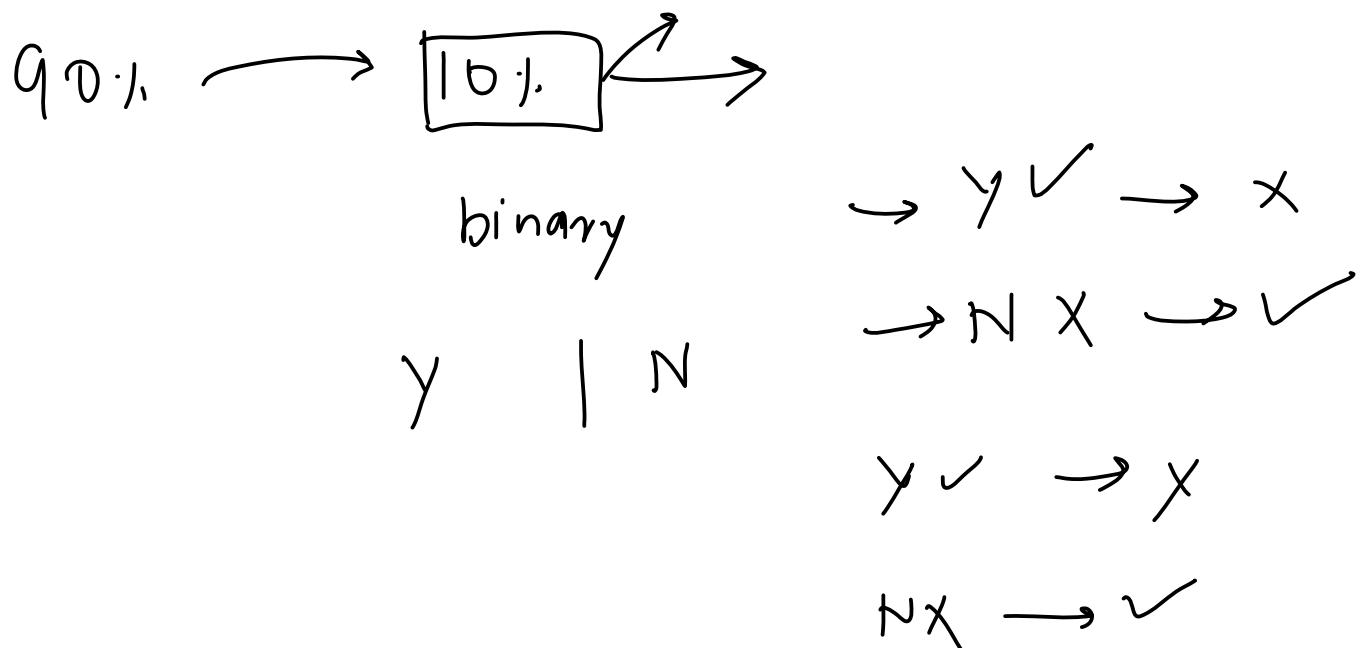
$$= \frac{10}{10} = 1 = 100\%$$

# How much accuracy is good?

Wednesday, June 23, 2021 11:14 AM

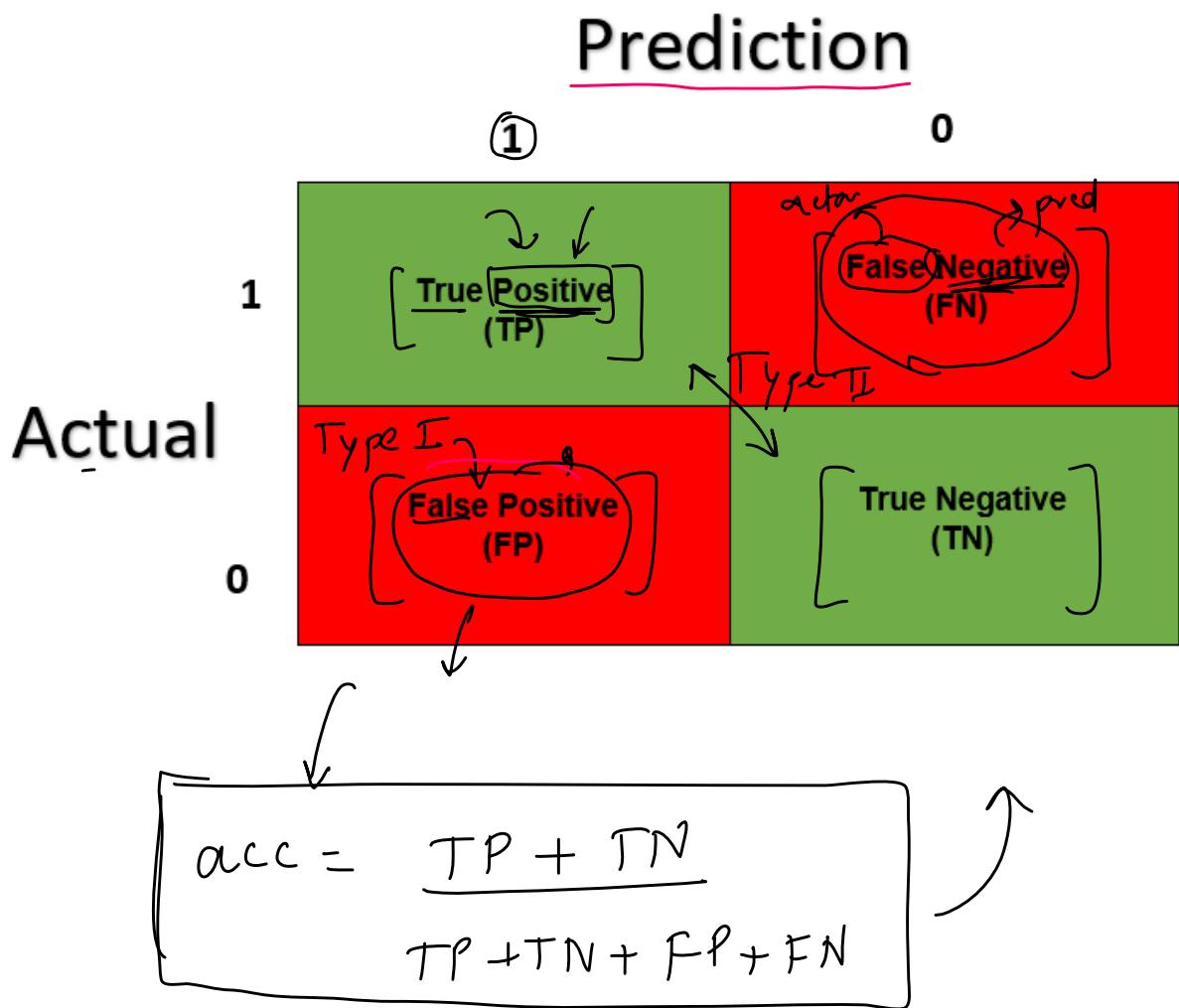
# The Problem with Accuracy

Wednesday, June 23, 2021 11:14 AM



## Confusion Matrix

Tuesday, June 22, 2021 1:12 PM



Extented  
↳ { echo }  
          { or }  
          { not }  
  
duplicate  
↳ echo

$$1+2+3=6 \rightarrow \text{outcome}$$
$$\begin{array}{r} x \\ + y \\ \hline z \end{array}$$

# Type 1 Error

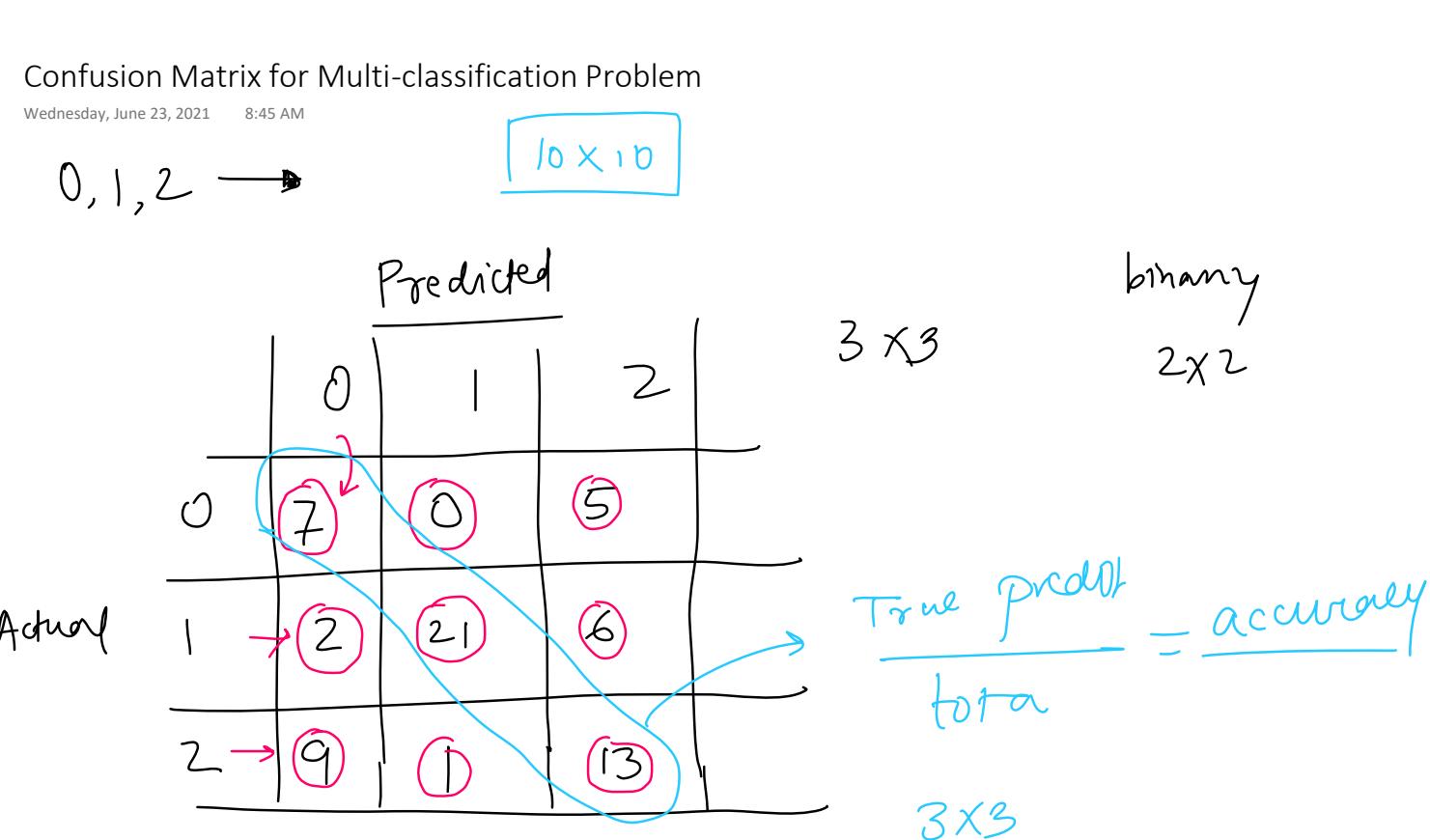
Wednesday, June 23, 2021 8:45 AM

# Type 2 Error

Wednesday, June 23, 2021 8:45 AM

## Confusion Matrix for Multi-classification Problem

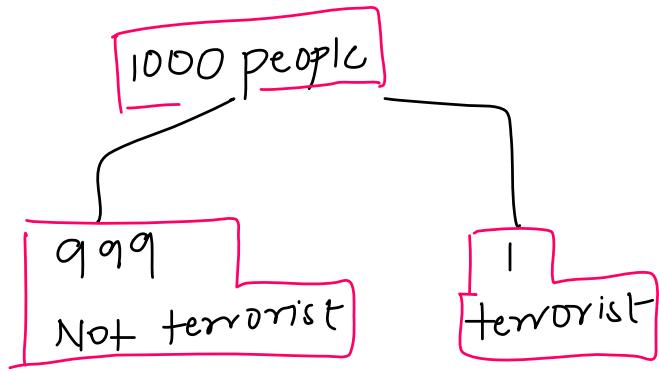
Wednesday, June 23, 2021 8:45 AM



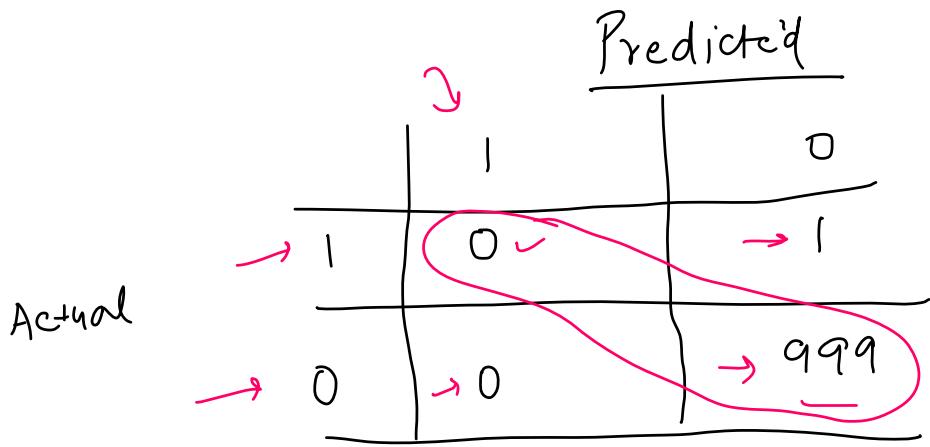
## When accuracy is misleading?

Wednesday, June 23, 2021 8:45 AM

### Imbalanced Dataset



model → No one is terrorist



$$\text{Accuracy} = \frac{999}{999 + 1}$$
$$= 99.9\%$$

## Precision

Wednesday, June 23, 2021 8:46 AM

{ spam: 1, not spam: 0 }

Actual

①(A)

Predicted

②(B)

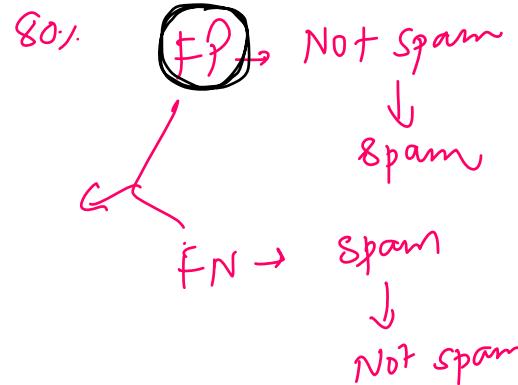
	Sent to Spam	Not sent to spam
Spam	100	170 FN
Not Spam	30 FP	700

	Sent to Spam	Not sent to spam
Spam	100	190
Not Spam	10	700

$$P_A = \frac{100}{100 + 30}$$

$$P_B = \frac{100}{100 + 10}$$

$$P_A < P_B$$



What proportion of predicted Positives is truly Positive?

①

0

Binary  
tvy

	Sent to Spam	Not sent to spam
Spam	True Positive	False Negative
Not Spam	False Positive	True Negative

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

## Recall

Wednesday, June 23, 2021 8:46 AM

has cancer: 1, no cancer: 0

Predicted

Actual

	Detected Cancer	Not Detected
Has Cancer	1000	200 (FN)
No Cancer	800 FP	8000

A 90%.

$$\text{Recall}_A = \frac{1000}{1200}$$

$R_A > R_B$

	Detected Cancer	Not Detected
Has Cancer	1000	500 ← B
No Cancer	500	8000

B 90%.

$$R_B = \frac{1000}{1500}$$

$\uparrow \downarrow \quad \text{FP} \leftarrow$

What proportion of actual Positives is correctly classified?

	Detected Cancer	Not detected Cancer
Has Cancer	<u>True Positive</u>	<u>False Negative</u>
No Cancer	False Positive	True Negative

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

## F1 Score

Wednesday, June 23, 2021 8:46 AM

$$F1 \text{ score} = \frac{2PR}{P+R}$$

$$P = 0 \quad R = 100$$

$$F1 = 0$$

$$F1 \text{ score} = 50$$

$$F1 \text{ score} = \frac{P+R}{2}$$

$$\frac{2 \times 80 \times 80}{160} = 80$$

$$\textcircled{A} \quad P = R = 80$$

$$\frac{80 + 80}{2} = 80 \quad F1 = 80$$

$$\textcircled{B} \quad P = 60 \quad R = 100$$

$$\frac{100 + 60}{2} = 80$$

$$\textcircled{C} \quad \frac{2 \times 60 \times 100}{160}$$

## Multi-class Precision and Recall

Wednesday, June 23, 2021 8:46 AM

		Binary	Multi	Dog Cat Rabbit		
Positive ① ✓ Yes No	Actual	② Class	③	$\frac{2 \times 0.86 + 0.66}{0.86 + 0.66}$		
		Predicted		$R_D, R_C, R_R$		
		Dog	Cat	Rabbit	Total	Recall
Dog		25	5	10	40	0.62
Cat		0	30	4	34	0.88
Rabbit		4	10	20	34	0.58
	Total	29	45	34		
	Precision	0.86	0.66	0.58		

↑

$$R_D = \frac{25}{40} = 0.62, R_C = \frac{30}{34} = 0.88, R_R = \frac{20}{34} = 0.58$$

macro recall ←       $F1_D, F1_C, F1_R$  ← macro f1

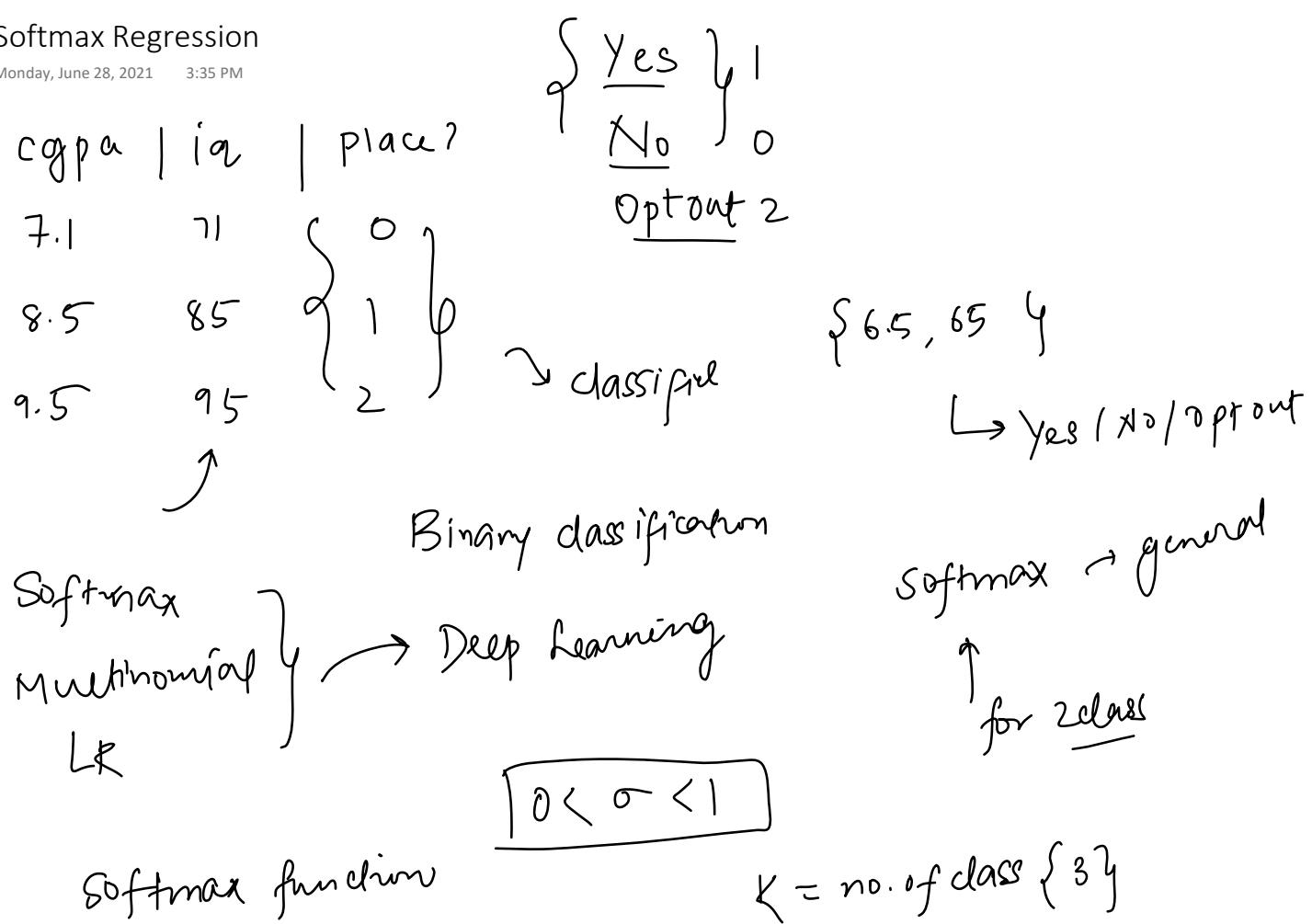
← weight recall      ← weighted f1

# Multi-class F1 Score

Thursday, June 24, 2021 11:14 AM

## Softmax Regression

Monday, June 28, 2021 3:35 PM



Yes → 1

No → 2

opt → 3

(Yes)

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

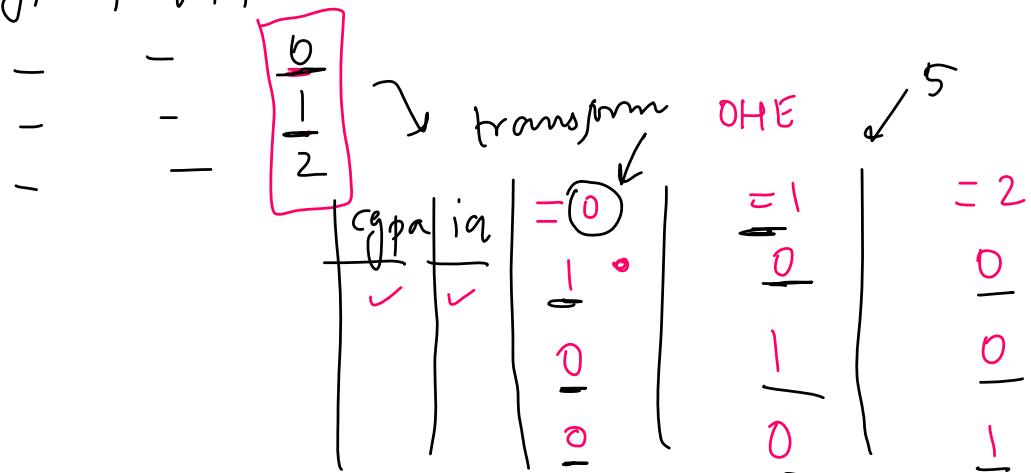
$$\sigma(z)_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\sigma(z)_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

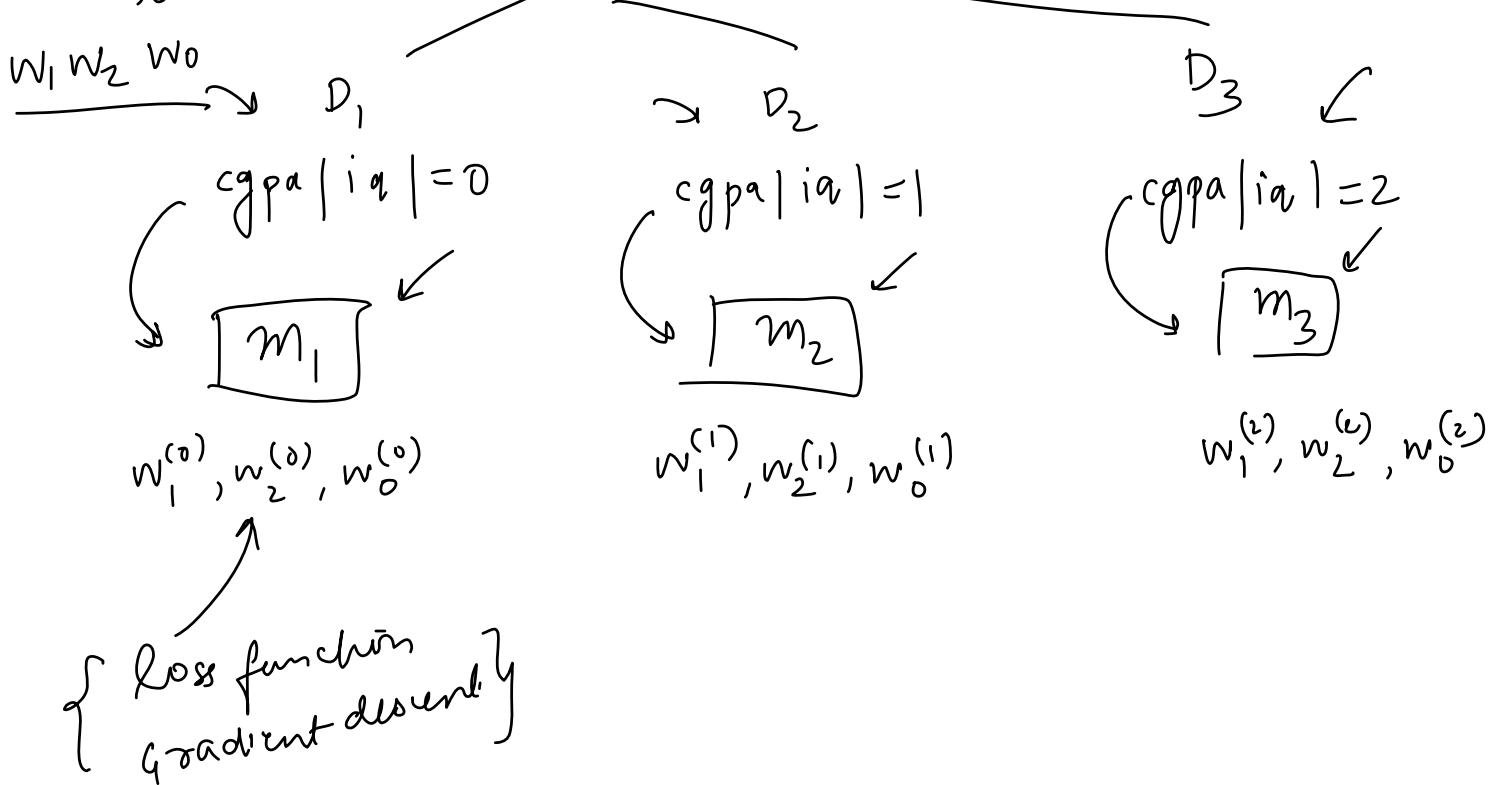
$$\sigma(z)_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\{y, N, D\}$$

cgpa | iq | place?



3 coeff



## Loss Function

Monday, June 28, 2021 3:47 PM

$$L = -\frac{1}{m} \sum_{i=1}^m y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

$\leftarrow \begin{cases} 1, 2, 3 \end{cases} \rightarrow i=1$

$$L = -\frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K y_k^{(i)} \log(\hat{y}_k^{(i)})$$

$\xrightarrow{\quad \underbrace{w_1^{(1)} w_2^{(1)} w_0^{(1)}} \quad}$

$x_1$	$x_2$	$y$	$y_{k=1}$	$y_{k=2}$	$y_{k=3}$
$x_{11}$	$x_{12}$	1	1	0	0
$x_{21}$	$x_{22}$	2	0	1	0
$x_{31}$	$x_{32}$	3	0	0	1

$y_1^{(1)} \log(\hat{y}_1^{(1)}) + y_2^{(1)} \log(\hat{y}_2^{(1)}) + y_3^{(1)} \log(\hat{y}_3^{(1)}) +$   
 $y_1^{(2)} \log(\hat{y}_1^{(2)}) + y_2^{(2)} \log(\hat{y}_2^{(2)}) + y_3^{(2)} \log(\hat{y}_3^{(2)}) +$   
 $y_1^{(3)} \log(\hat{y}_1^{(3)}) + y_2^{(3)} \log(\hat{y}_2^{(3)}) + y_3^{(3)} \log(\hat{y}_3^{(3)}) +$

$$L = \underline{y_1^{(1)} \log(\hat{y}_1^{(1)})} + \underline{y_2^{(2)} \log(\hat{y}_2^{(2)})} + \underline{y_3^{(3)} \log(\hat{y}_3^{(3)})}$$

$$\hat{y}_1^{(1)}, \hat{y}_2^{(2)}, \hat{y}_3^{(3)}$$

softmax

$$\hat{y}_1^{(1)} = \sigma(w_1^{(1)} x_{11} + w_2^{(1)} x_{12} + w_0^{(1)})$$

$$y_2^{(2)} = \sigma(w_1^{(2)} x_{21} + w_2^{(2)} x_{22} + w_0^{(2)})$$

$$y_3^{(3)} = \sigma(w_1^{(3)} x_{31} + w_2^{(3)} x_{32} + w_0^{(3)})$$

$$\begin{bmatrix} w_1^{(1)} & w_2^{(1)} & w_0^{(1)} \\ w_1^{(2)} & w_2^{(2)} & w_0^{(2)} \\ w_1^{(3)} & w_2^{(3)} & w_0^{(3)} \end{bmatrix}$$

(L)

$$\frac{\partial L}{\partial w_1^{(1)}}, \frac{\partial L}{\partial w_2^{(1)}}, \frac{\partial L}{\partial w_0^{(1)}} \dots \quad q \text{ due}$$

gradient

*q-values init=1*

$$\begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} =$$

loop  $\rightarrow$  1000 epochs

$$w_1^{(1)} = w_1^{(1)} - \eta \frac{\partial L}{\partial w_1^{(1)}}$$

$$w_2^{(1)} = w_2^{(1)} - \eta \frac{\partial L}{\partial w_2^{(1)}}$$

;

## Prediction

Monday, June 28, 2021 3:35 PM

$$\begin{aligned}
 S_x = \{7, 70\} \Rightarrow \overline{Y, N, Opt} & \\
 \downarrow & \\
 m_1 & \text{ Yes} \\
 \underline{w}_1^{(1)} \quad \underline{w}_2^{(1)} \quad w_0^{(1)} & \\
 \underline{z}_1 = 7 \times w_1^{(1)} + 70 \times w_2^{(1)} + w_0^{(1)} & \\
 \downarrow & \\
 m_2 & \text{ No} \\
 \underline{w}_1^{(2)} \quad \underline{w}_2^{(2)} \quad w_0^{(2)} & \\
 \underline{z}_2 = 7 \times w_1^{(2)} + 70 \times w_2^{(2)} + w_0^{(2)} & \\
 \downarrow & \\
 m_3 & \text{ Opt out} \\
 w_1^{(3)} \quad w_2^{(3)} \quad w_0^{(3)} & \\
 \underline{z}_3 = 7 \times w_1^{(3)} + 70 \times w_2^{(3)} + w_0^{(3)} & \\
 \downarrow & \\
 \sigma(y) = \frac{e^{\underline{z}_1}}{e^{\underline{z}_1} + e^{\underline{z}_2} + e^{\underline{z}_3}} & \sigma(N) = \frac{e^{\underline{z}_2}}{e^{\underline{z}_1} + e^{\underline{z}_2} + e^{\underline{z}_3}} \\
 & e(\delta) = \frac{e^{\underline{z}_3}}{e^{\underline{z}_1} + e^{\underline{z}_2} + e^{\underline{z}_3}} \\
 & \underline{0.35} \quad \underline{0.25} \\
 & = \underline{0.40}
 \end{aligned}$$

# Sigmoid Vs Softmax

Monday, June 28, 2021 3:47 PM

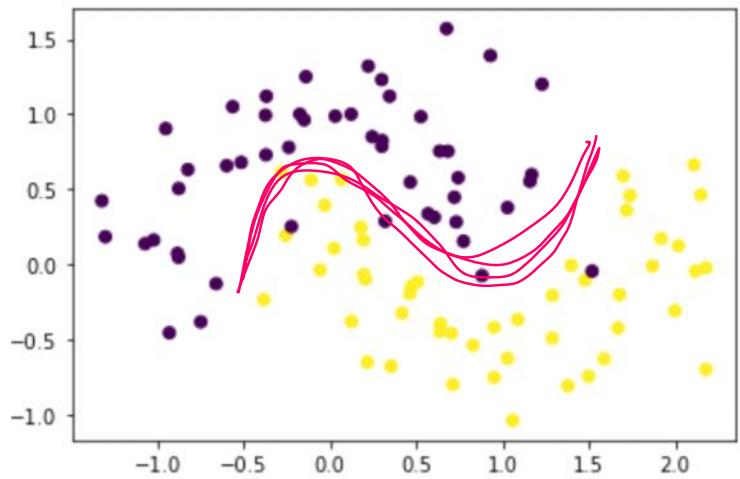
# Code Sample

Monday, June 28, 2021 3:36 PM

## Polynomial Logistic Regression

Monday, June 28, 2021 6:43 PM

$$\underline{x_1} \underline{x_2} \underline{y} \not\perp \!\!\! \perp 0, 1$$



Linear      degree

polynomial       $x^0 x^1 x^2 x^3$

degree = 2

$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \rightarrow 2 \text{ cols}$

$w_0 \rightarrow$

$\begin{array}{c} x_1^0 \\ \uparrow \\ w_1 \\ \begin{array}{c} x_1^1 \\ \downarrow \\ w_2 \\ x_1^2 \end{array} \end{array} \quad \begin{array}{c} x_2^0 \\ | \\ w_3 \\ x_2^1 \\ | \\ x_2^2 \end{array} \quad y$

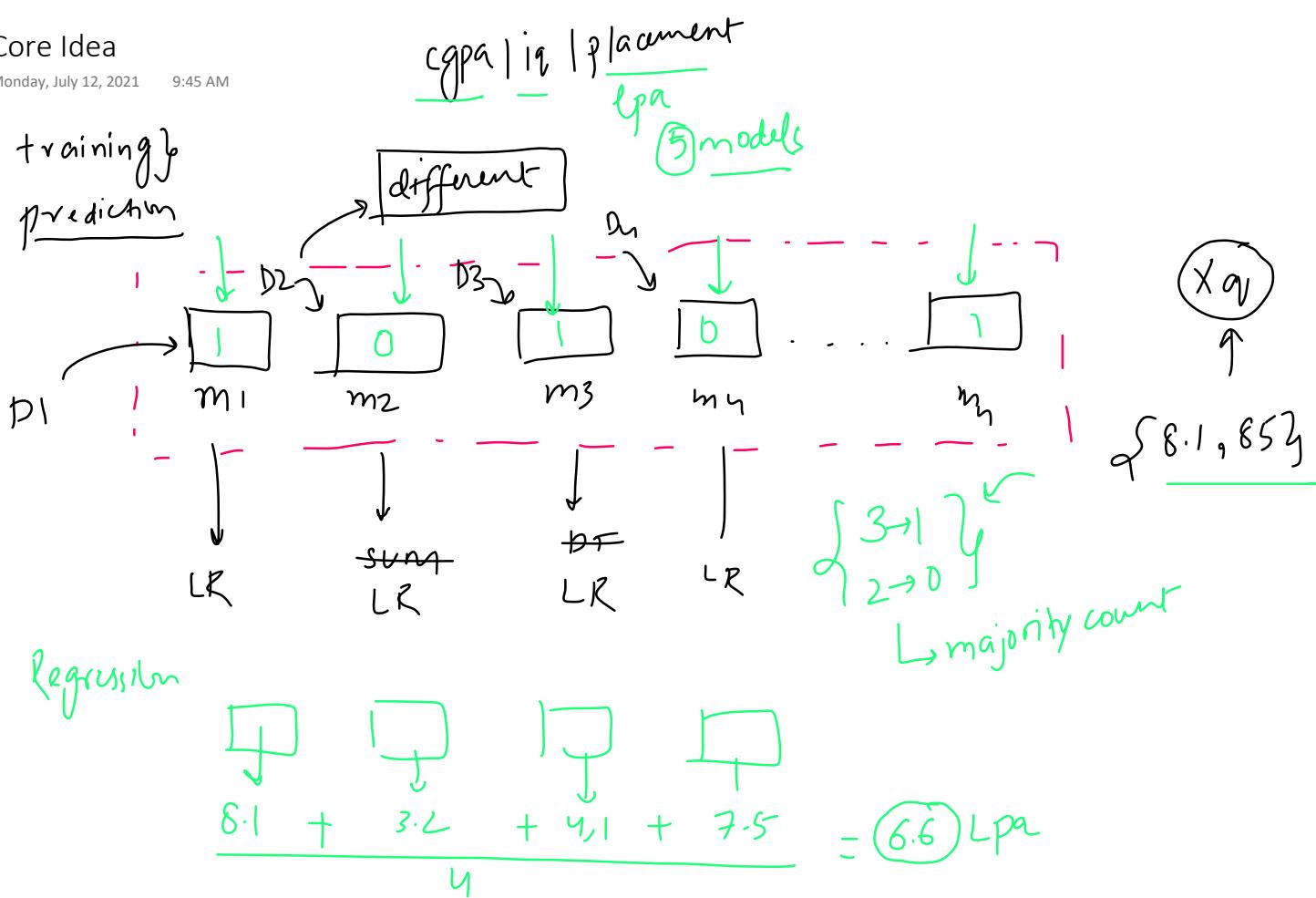
6 cols

# Wisdom of the Crowd

Monday, July 12, 2021 9:56 AM

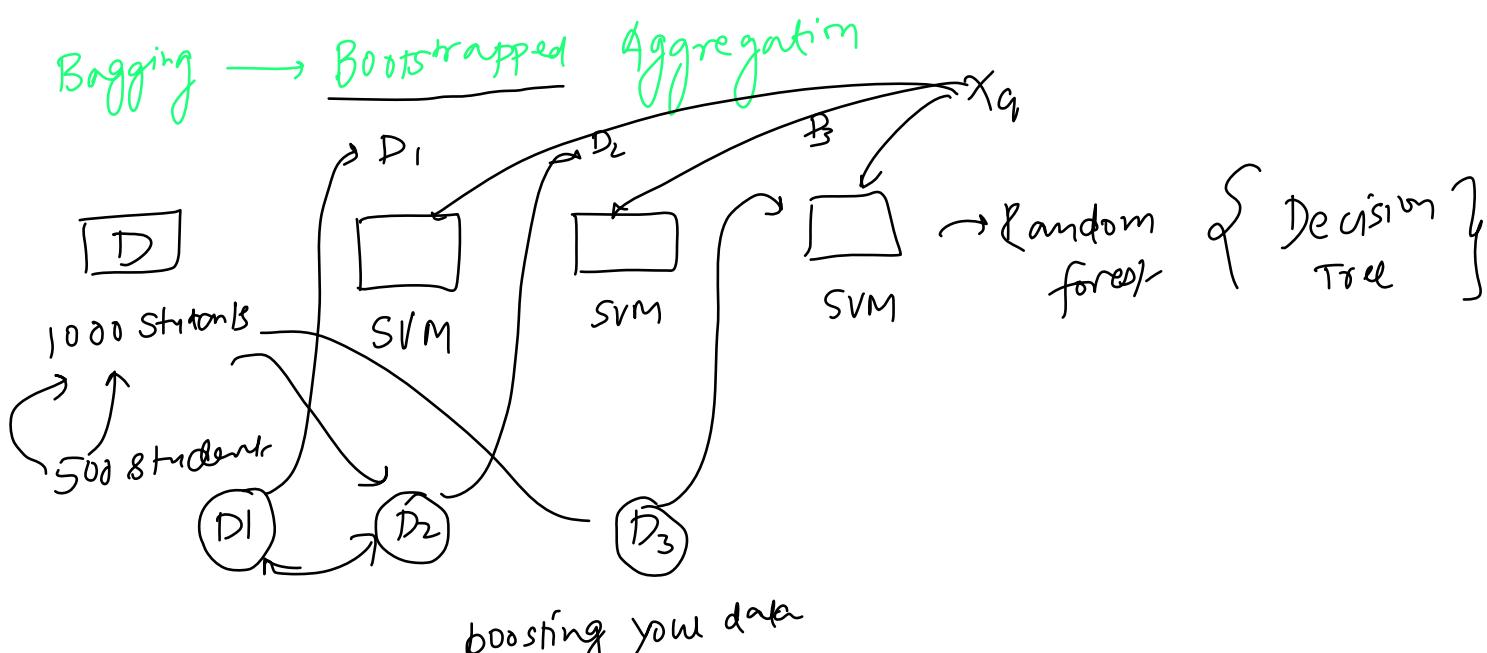
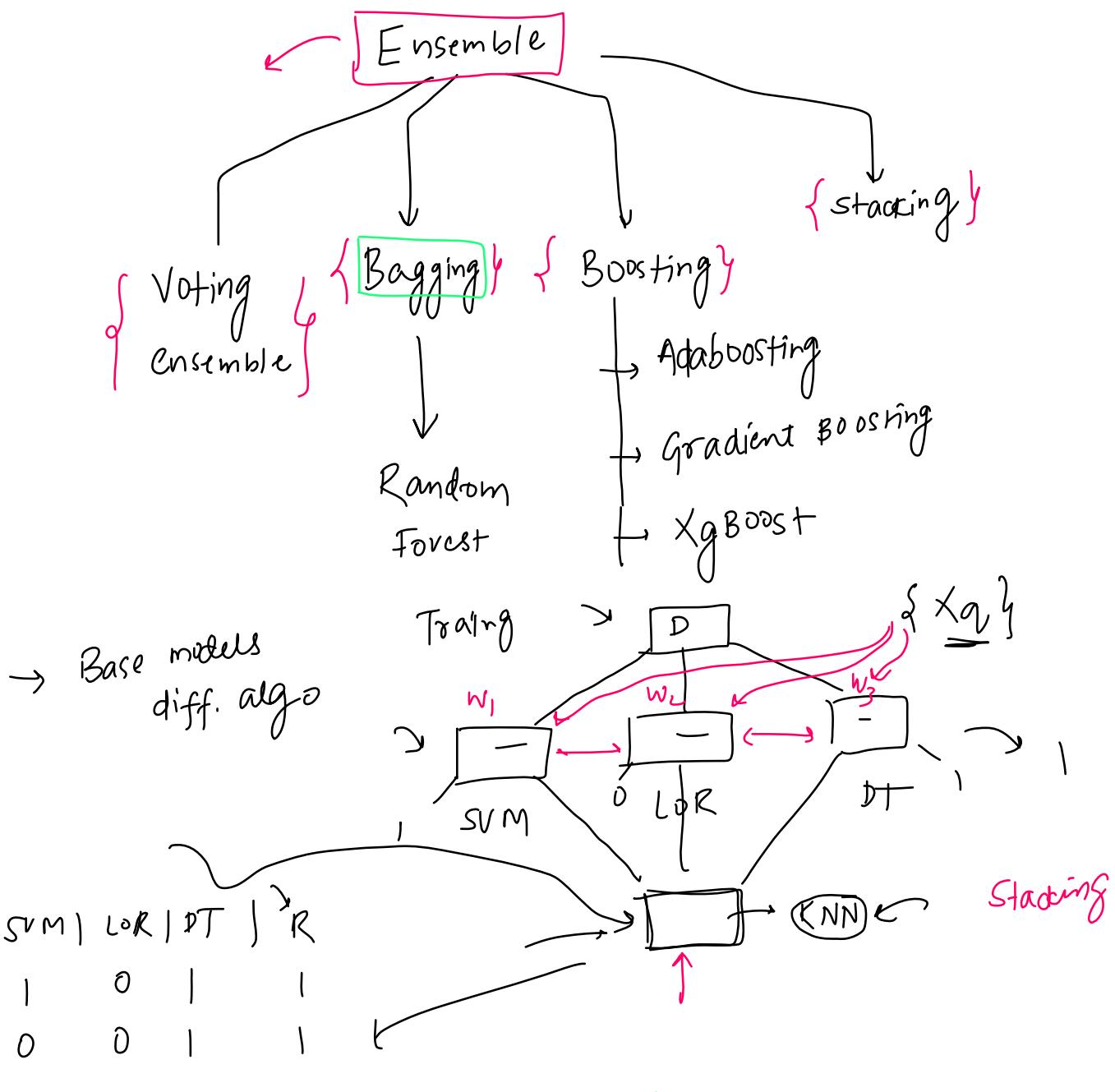
## Core Idea

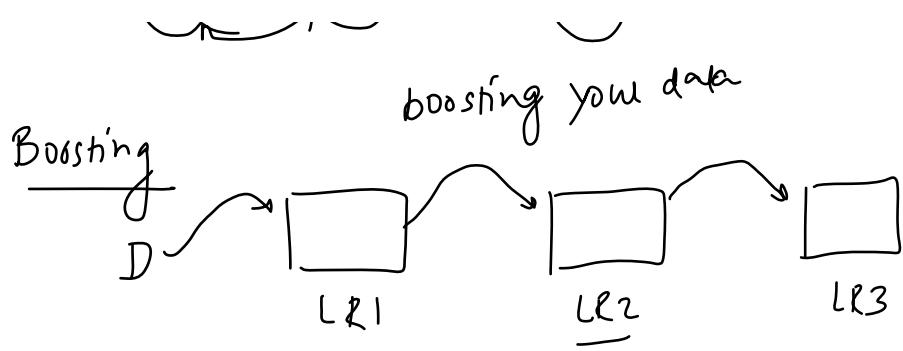
Monday, July 12, 2021 9:45 AM



# Type of Ensemble Learning

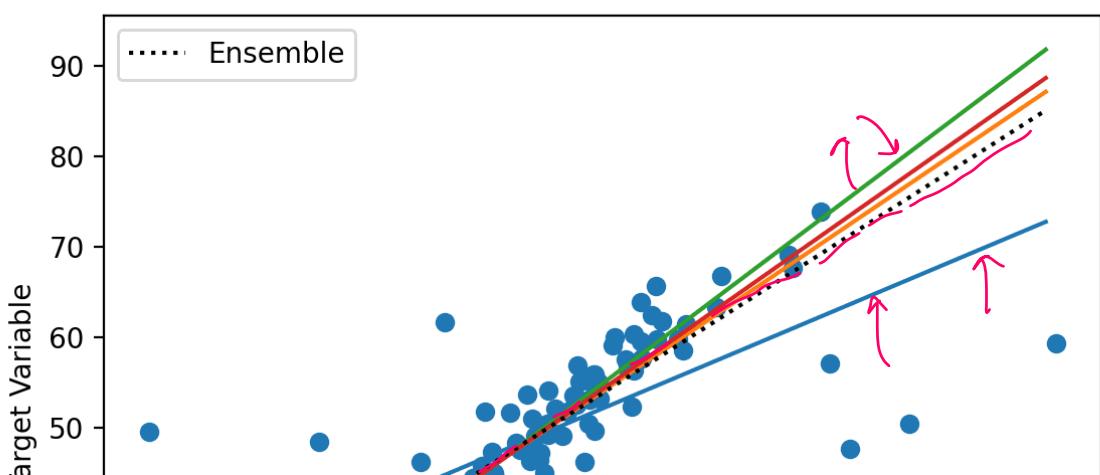
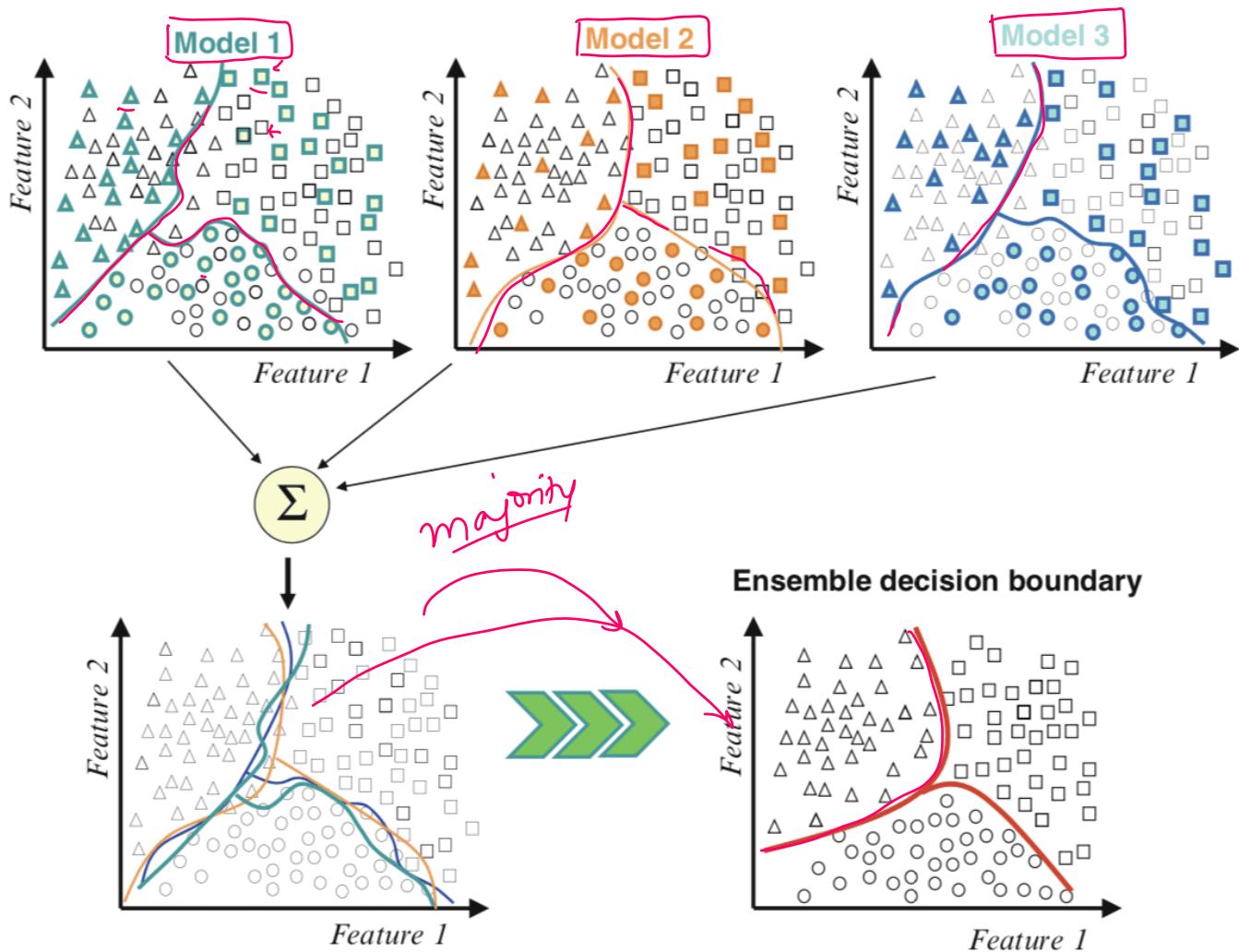
Monday, July 12, 2021 9:45 AM

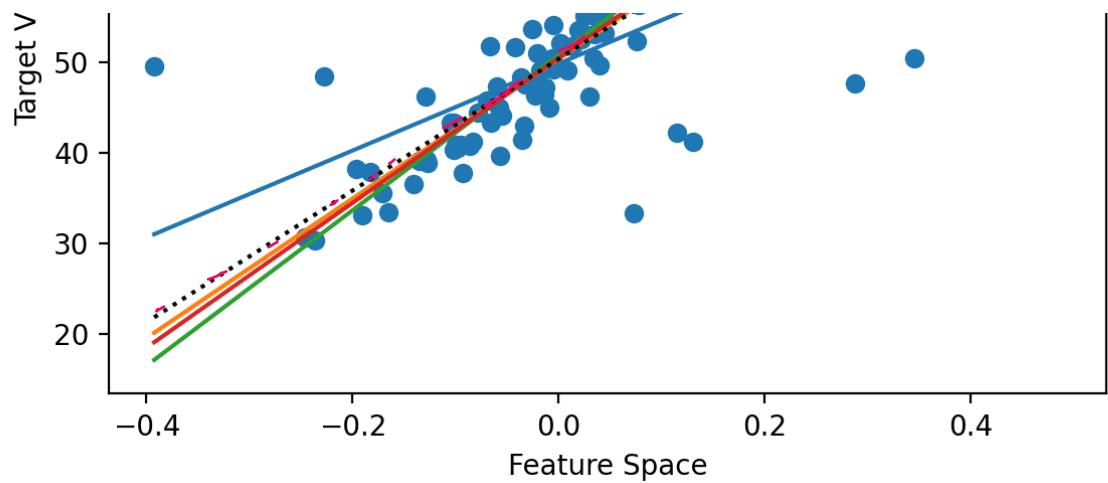




## Why it works?

Monday, July 12, 2021 9:45 AM





## Benefits

Monday, July 12, 2021 9:45 AM

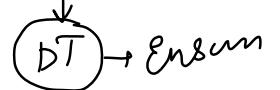
1) Improvement in performance

2) Bias Variance



Low Bias + Low Variance

Low Bias High Variance



Low Bias → LV

3) Robustness

High bias Low Variance

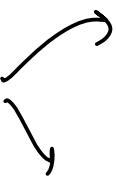
Low bias



# When to use?

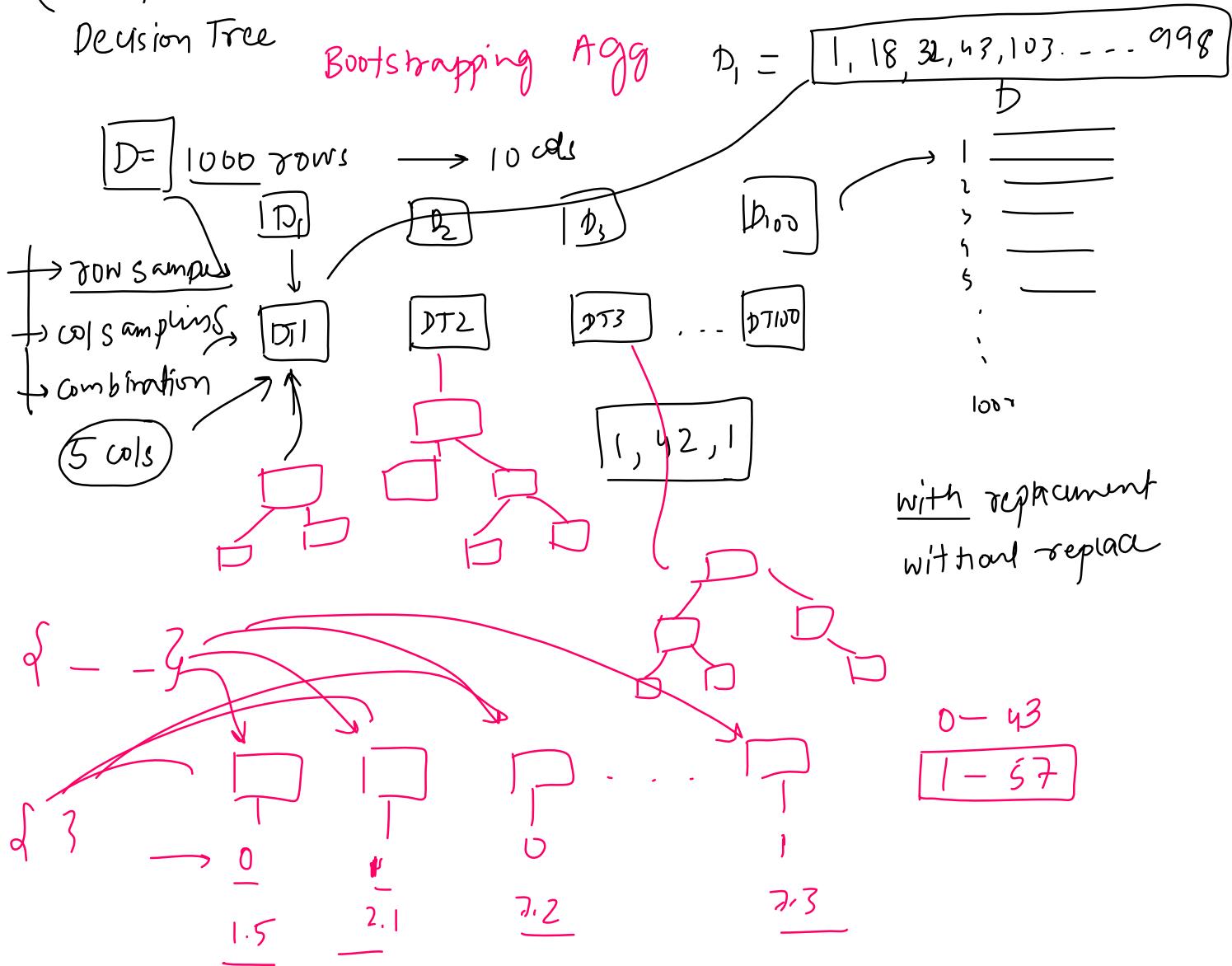
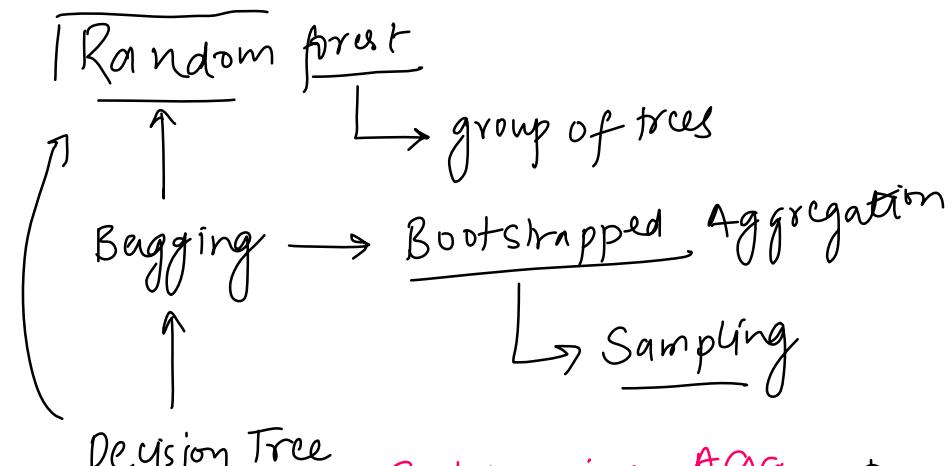
Monday, July 12, 2021 9:46 AM

Always



# Intuition

Monday, July 19, 2021 4:53 PM

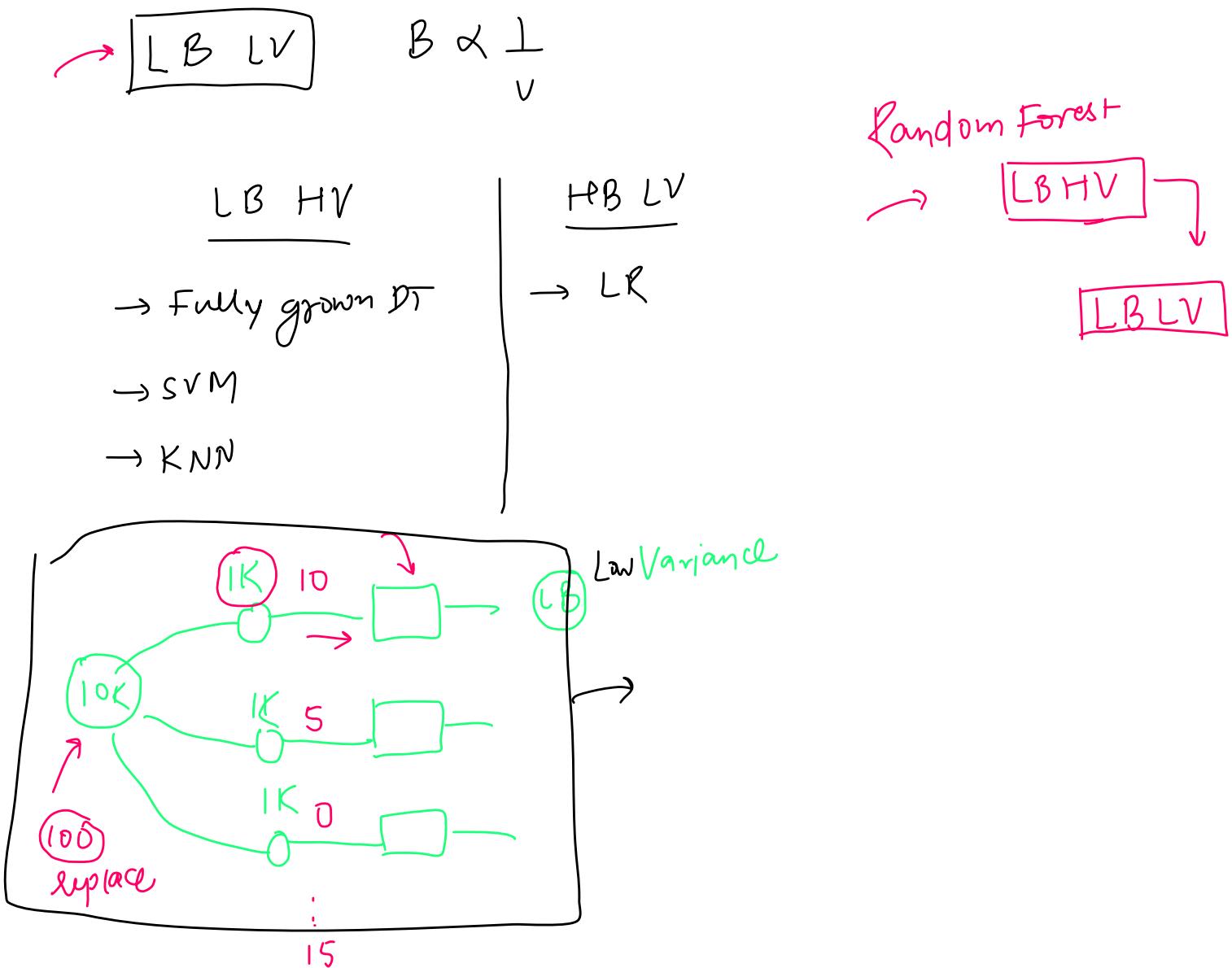


# Demo

Monday, July 19, 2021 5:05 PM

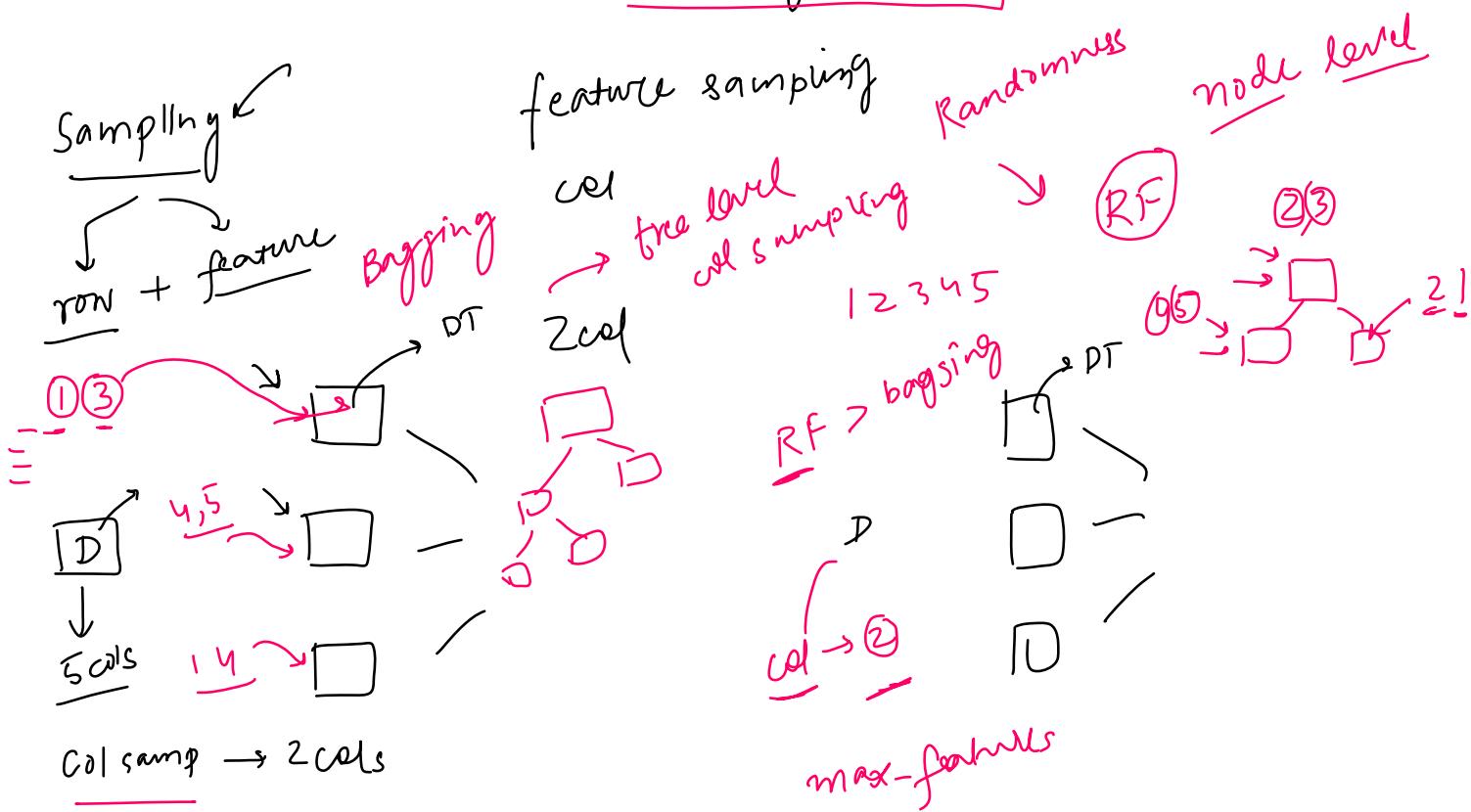
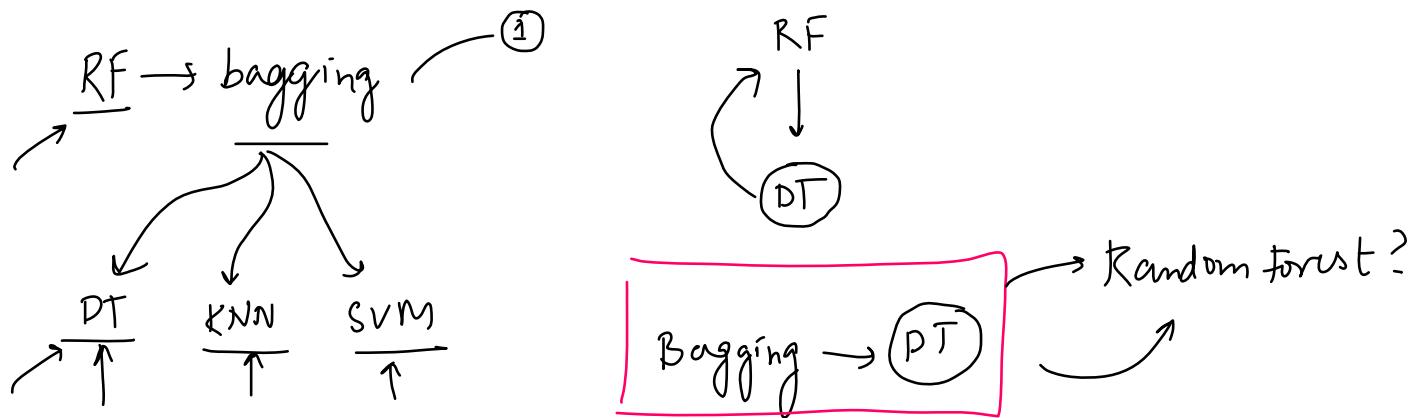
# Bias Variance and Random Forest

Tuesday, July 20, 2021 10:52 AM



# Bagging Vs Random Forest

Monday, July 19, 2021 4:53 PM



# Hyperparameters

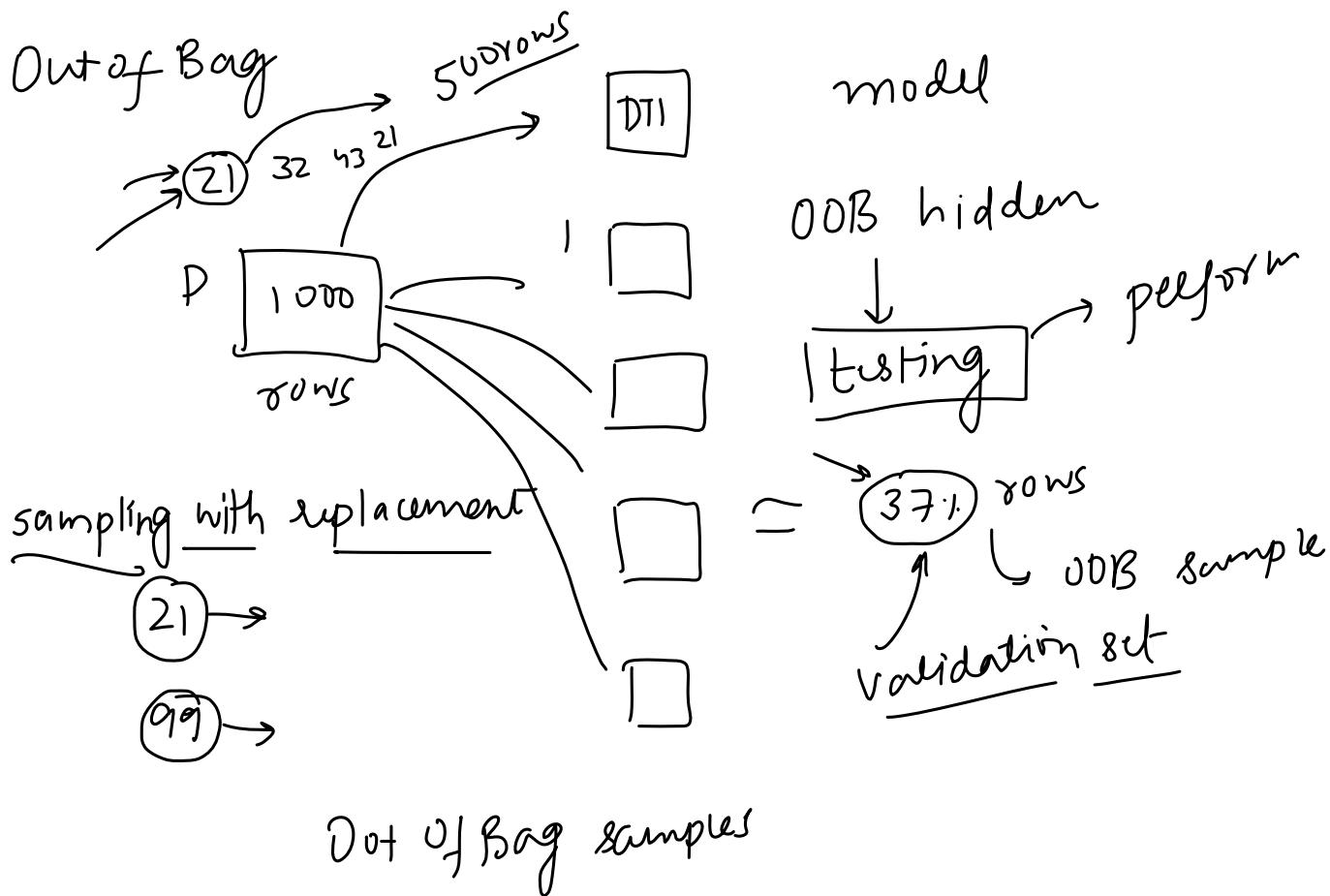
Monday, July 19, 2021 4:56 PM

# Code Demo

Monday, July 19, 2021 5:05 PM

## OOB Evaluation

Monday, July 19, 2021 5:14 PM

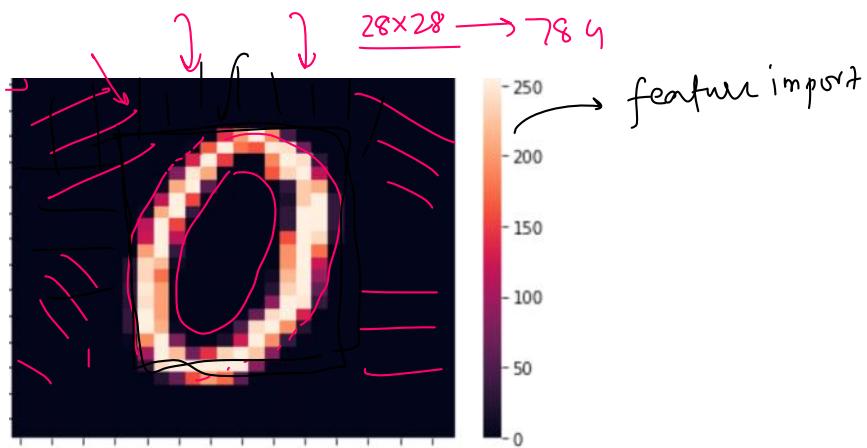


## Feature Importance

Monday, July 19, 2021 4:56 PM

MNIST

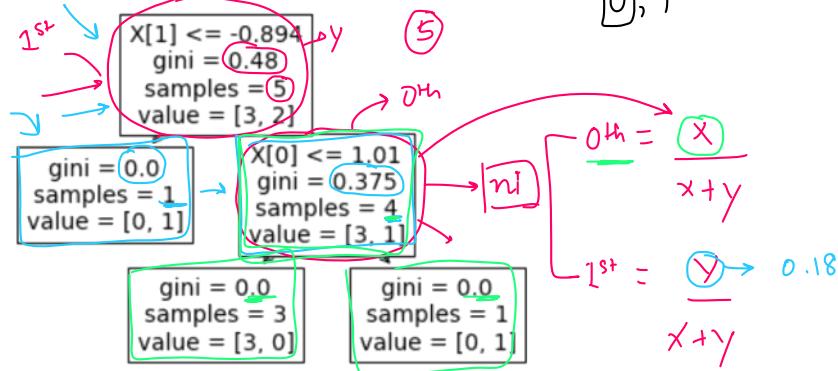
label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel774	pixel775	pixel776	pixel777	pixel778	42000 images
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	



Formula

$$ni = \frac{N-t}{N} \left[ \text{impurity} - \left( \frac{N-t-\gamma}{N-t} \times \text{right\_impurity} \right) - \left( \frac{N-t-\gamma}{N-t} \times \text{left\_impurity} \right) \right]$$

$$f_{ik} = \frac{\sum_{j \in \text{node split on feature } k} ni}{\sum_{j \in \text{all nodes}} ni}$$



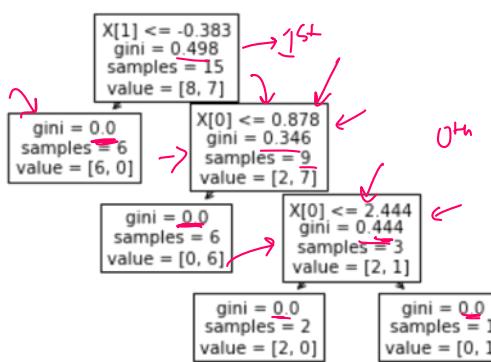
$$= \frac{5}{5} \left[ 0.48 - \frac{4}{5} \times 0.375 - \frac{1}{5} \times 0.0 \right] = 0.18$$

$$\frac{4}{5} \left[ 0.375 \right] = X = 0.8 \times 0.375 = 0.30 = X$$

$$0^{th} = \frac{0.3}{0.3 + 0.18} = 0.625$$

$$0^{th} = \frac{0.3}{0.3 + 0.18} = 0.625$$

$$1^{st} = \frac{0.18}{0.3 + 0.18} = 0.375$$



0<sup>th</sup> Node

$$\frac{15}{15} \left[ 0.49 - \frac{9}{15} \times 0.34 \right]$$

$$= 0.290$$

$$f_i[0] = \frac{0.11 + 0.08}{0.29 + 0.11 + 0.08} = 0.48$$

$$= 0.39$$

0<sup>th</sup> Node

$$\frac{9}{15} \left[ 0.34 - \frac{3}{15} \times 0.44 \right]$$

$$= 0.118$$

$$\frac{3}{15} \left[ 0.44 \right]$$

$$= 0.088$$

$$f_i[1] = \frac{0.29}{0.48} = 0.60$$

# Extra Trees

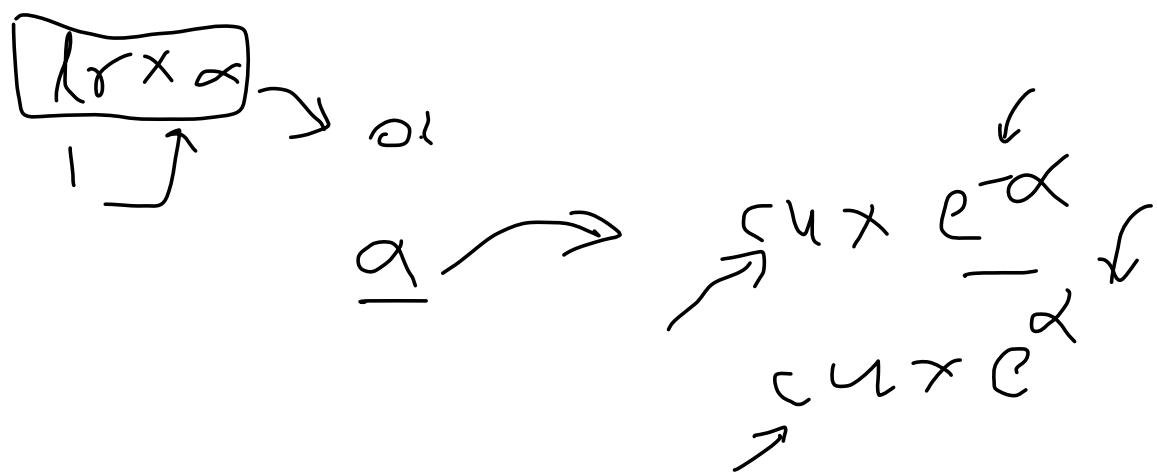
Monday, July 19, 2021 4:57 PM

# Before Starting

Monday, August 9, 2021 10:42 AM

i) weak classifiers ( $> 50\%$ )

$$\alpha = \frac{1}{2} \log \left( \frac{1-p}{p} \right)$$



# The Big Idea

Monday, August 9, 2021 10:43 AM

# Step by Step Breakdown

Monday, August 9, 2021 10:43 AM

# Points to remember

Monday, August 9, 2021 10:43 AM

# Code Walkthrough

Monday, August 9, 2021 10:43 AM

# Example and Hyperparameters

Monday, August 9, 2021 10:43 AM

## Algorithm

Monday, September 20, 2021 8:23 AM

$$\rightarrow \{(x_i, y_i)\}_{i=1}^n \quad \eta=3 \quad x_i \ y_i$$

Input: training set  $\{(x_i, y_i)\}_{i=1}^n$  a differentiable loss function  $L(y, F(x))$ , number of iterations  $M$ .

$\rightarrow$  1. Initialize  $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$ .

$\rightarrow$  2. For  $m = 1$  to  $M$ :

$\rightarrow$  (a) For  $i = 1, 2, \dots, N$  compute

$\nearrow$  row  $\nearrow$   
 $i/m \rightarrow$

$$r_{im} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

residual / pseudo-residual

(b) Fit a regression tree to the targets  $r_{im}$  giving terminal regions  $R_{jm}$ ,  $j = 1, 2, \dots, J_m$ .

(c) For  $j = 1, 2, \dots, J_m$  compute

$$\boxed{\gamma_{jm}} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update  $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$ .

3. Output  $\hat{f}(x) = \boxed{f_M(x)}$ .

$$\oplus \quad f_1(x) = \overbrace{f_0(x)} + \textcircled{dT}$$

$$f_2(x) = f_1(x) + dT_2$$

$$f_2(x) = f_1(x) + dT_2$$

$$f_0(x) + dT$$

$$f_1(x) + dT_1$$

$$f_4(x) = f_0(x) + \dots$$

$\rightarrow$  recursion

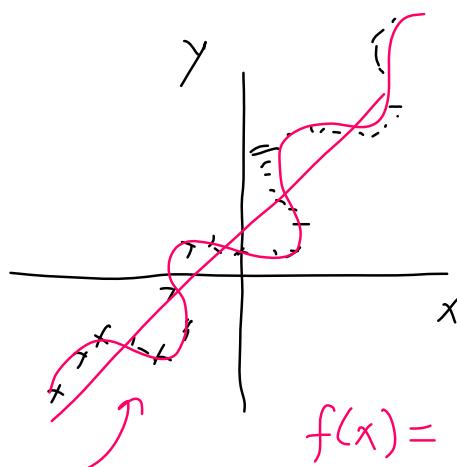
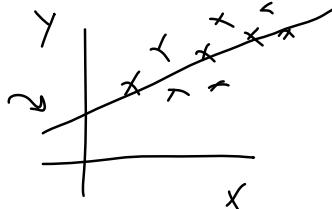
$$f_1(x) \ f_2(x)$$

## Additive Modelling

Monday, September 20, 2021 8:28 AM

$$\begin{array}{c} \text{---} \\ \text{x} \end{array} \left| \begin{array}{c} \text{y} \rightarrow f() \\ \text{x} \rightarrow \end{array} \right. \Rightarrow y = f(x)$$

$$x_1, x_2, x_3 | y \quad y = f(x_1, x_2, x_3)$$



$$f(x) = x + \sin x$$

additive

$$F(x) = f_0(x) + f_1(x) + f_2(x) + \dots$$

↓      ↓  
DI      PT

$$y = x \quad y = \sin(x)$$

## Explanation

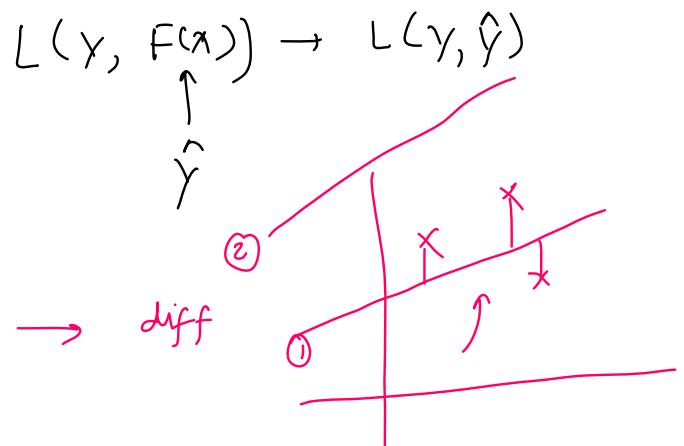
Monday, September 20, 2021 8:25 AM

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

↓  
actual      ↓  
pred

$L = \frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

$\frac{1}{2}$



$$\textcircled{1} = \underline{10} = 5 \text{ ✓}$$

$$\textcircled{2} = \underline{20} = 10$$

$$y = f(x) \rightsquigarrow$$

$$f(x) = f_0(x) + \overbrace{f_1(x) + f_2(x) + \dots + f_n(x)}$$

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, \underline{\gamma})$$

$$L = \frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^n (y_i - \underline{\gamma})^2$$

$$\frac{d f_0(x)}{d \gamma} = \frac{d}{d \gamma} \frac{1}{2} \sum_{i=1}^n (y_i - \underline{\gamma})^2 = \frac{1}{2} \sum_{i=1}^n \frac{d}{d \gamma} (y_i - \underline{\gamma})^2$$

$$\sum_{i=1}^n (y_i - \underline{\gamma}) \frac{d}{d \gamma} (y_i - \underline{\gamma}) = - \sum_{i=1}^n (y_i - \underline{\gamma}) = 0$$

$$\sum_{i=1}^n (\underline{\gamma} - y_i) = 0$$

$$\sum_{i=1}^3 (\underline{\gamma} - y_i) = 0 \Rightarrow (\underline{\gamma} - 192) + (\underline{\gamma} - 144) + (\underline{\gamma} - 91) = 0$$

$$\sum_{i=1}^3 (\gamma - \bar{y}_i) = 0 \Rightarrow (\gamma - 192) + (\gamma - 144) + (\gamma - 9) = 0$$

$$3\gamma = 192 + 144 + 9$$

mean  
 $F_m(x) = f_0(x)$   
mean of output

$$\boxed{\gamma = \frac{192 + 144 + 9}{3}}$$

$$F(x) = \underbrace{f_0(x)}_{\text{mean}} + \underbrace{f_1(x)}_{\uparrow} + \underbrace{f_2(x)}_{\uparrow} + \dots + \underbrace{f_m(x)}_{\uparrow}$$

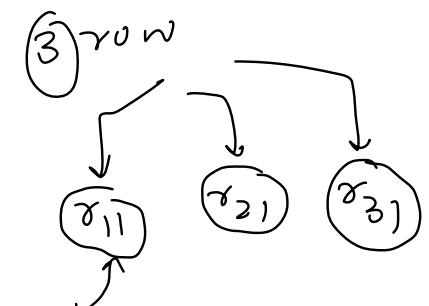
mean  
(leaf)

$$\boxed{m=1}$$

$$\sigma_{im} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$$

$$\sigma_{ii} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_0}$$

$$\hat{y}_i = f(x_i)$$



$$\sigma_{ii} = - \left[ \frac{\partial}{\partial \hat{y}_i} L(y_i, \hat{y}_i) \right]_{f=f_0}$$

$$L = \frac{1}{2} \sum_{j=1}^n (y_j - \hat{y}_j)^2$$

$$\sigma_{ii} = - \left[ \frac{\partial}{\partial \hat{y}_i} \frac{1}{2} (\underline{y_i - \hat{y}_i})^2 \right]_{f=f_0}$$

$$= \left[ (y_i - \hat{y}_i) \right]_{f=f_0} = \left[ (y_i - f(x_i)) \right]_{f=f_0}$$

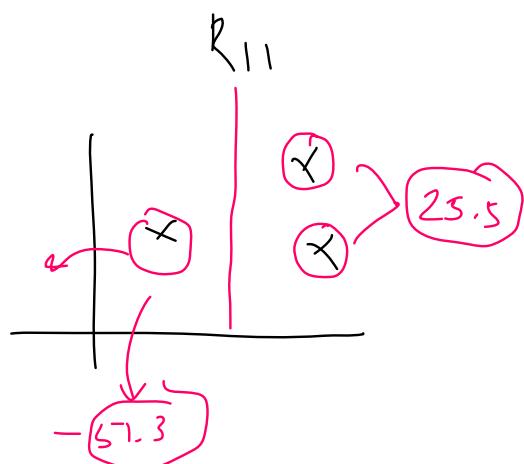
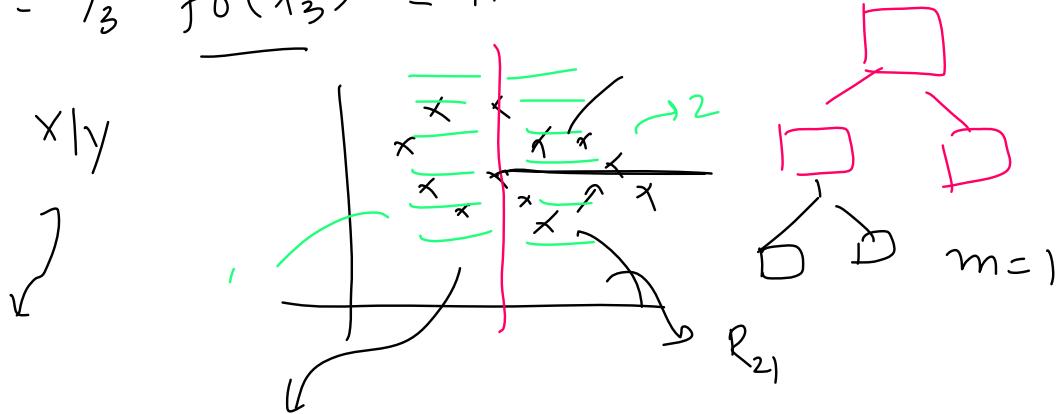
$$= \lfloor (y_i - \hat{y}_i) \rfloor_{f=f_0} = \lfloor (y_i - f(x_i)) \rfloor_{f=f_0}$$

$$\varepsilon_{j1} = \underline{(y_i - f_0(x_i))}$$

$$\varepsilon_{11} = y_1 - \underline{f_0(x_1)} = 192 - 142 =$$

$$\varepsilon_{21} = y_2 - \underline{f_0(x_2)} = 144 - 142 =$$

$$\varepsilon_{31} = y_3 - \underline{f_0(x_3)} = 91 - 142 =$$



$$\underline{\gamma_{jm}} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

$$\underline{\gamma_{j1}} \rightarrow \underline{\gamma_{11}} = \arg \min_{\gamma} \sum_{x_i \in R_{11}} L(y_i, f_{m-1}(x_i) + \gamma)$$

$$\begin{array}{c} \downarrow \quad \downarrow \\ \gamma_{11} \quad \gamma_{21} \end{array} \quad \textcircled{\text{1}} \quad \nearrow \quad \overline{\uparrow}$$

$$\gamma_{11} = \arg \min_{\gamma} \sum_i (y_i - (f_0(x_i) + \gamma))^2$$

$$\frac{\partial L}{\partial \gamma} = \frac{1}{n} \times \sum_i (y_i - (f_0(x_i) + \gamma)) \frac{\partial}{\partial \gamma} (\underbrace{y_i - f_0(x_i)}_{=0} - \underbrace{\gamma}_{=0}) = 0$$

$$-(y_i - f_0(x_i) - \gamma) = 0$$

$$y_i - f_0(x_i) - \gamma = 0$$

$$\gamma_{11} = q_1 - 142 - \gamma = 0$$

$$y = 91 - 142 = -51$$

$$\gamma_{21} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{21}} L(y_i, f_0(x_i) + \gamma)$$

$$= \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n (y_i - (f_0(x_i) + \gamma))^2$$

$$= - \sum_{i=1}^n (y_i - f_0(x_i) - \gamma) = 0$$

$$= \sum_{i=1}^n (y_i - f_0(x_i) - \gamma) = 0$$

$$= y_1 - f_0(x_1) - \gamma + y_2 - f_0(x_2) - \gamma = 0$$

336  
284

$$= 192 - 142 - \gamma + 144 - 142 - \gamma = 0$$

$$336 - 284$$

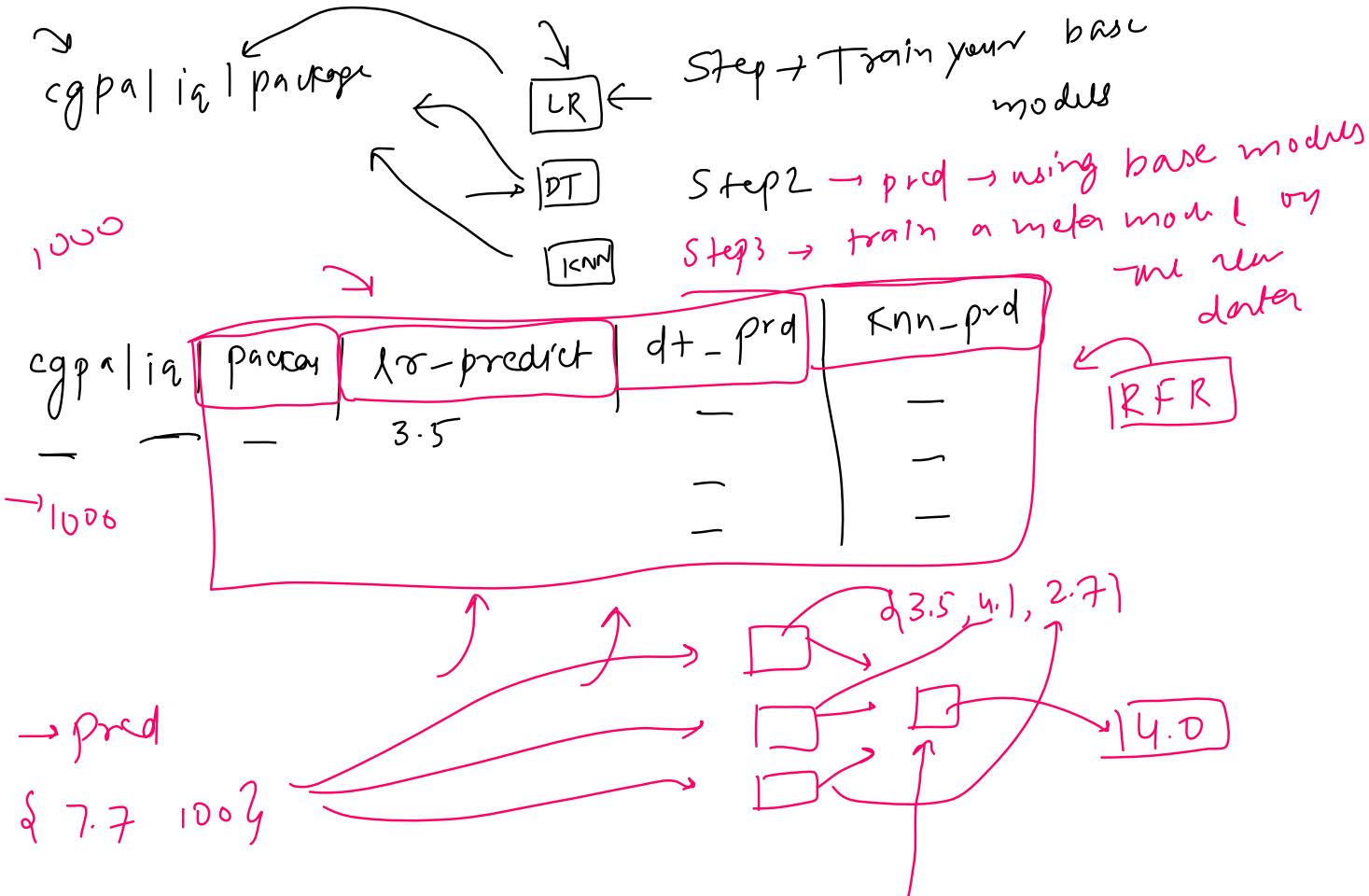
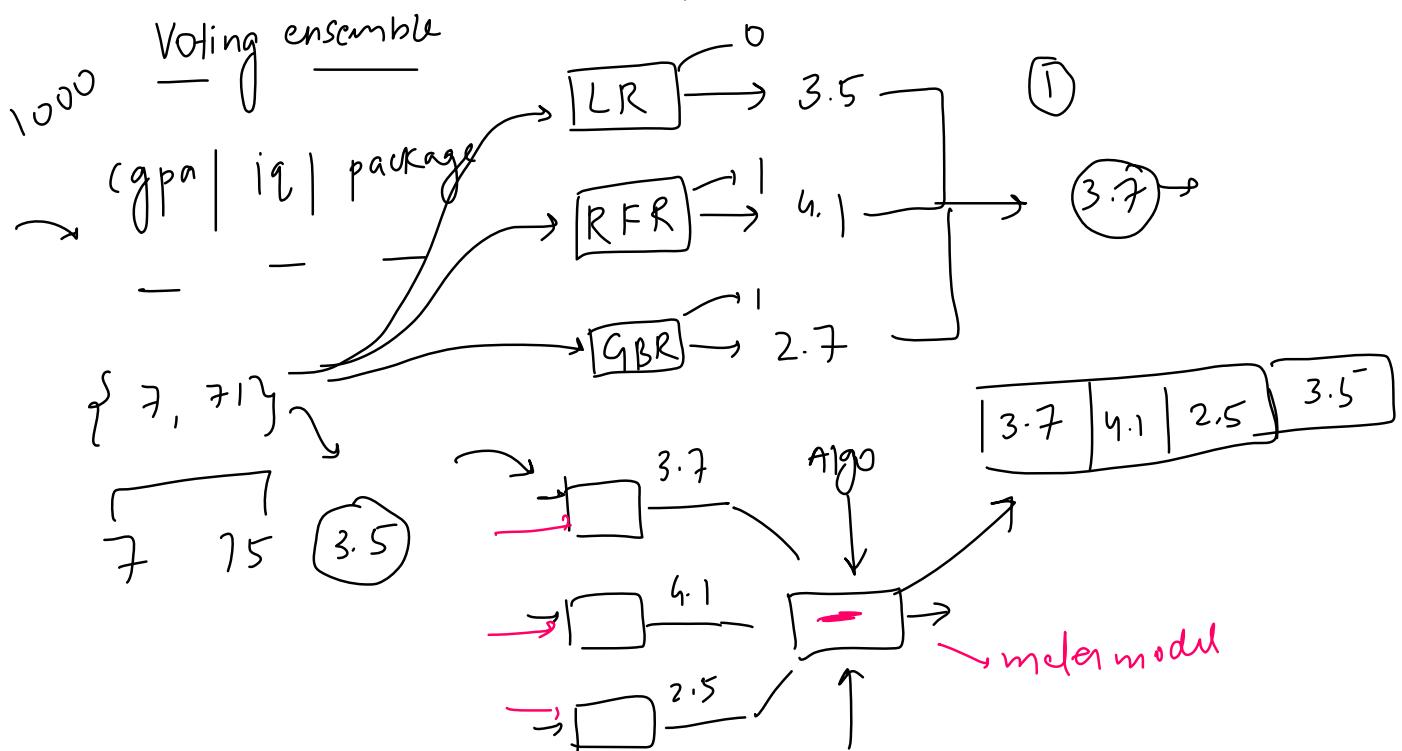
$$52 - 2\gamma = 0$$

$$\gamma = \frac{52}{2} - 26$$

## Introduction

Thursday, September 30, 2021 3:47 PM

## Stacking



## Boosting / Bagging

→ base model

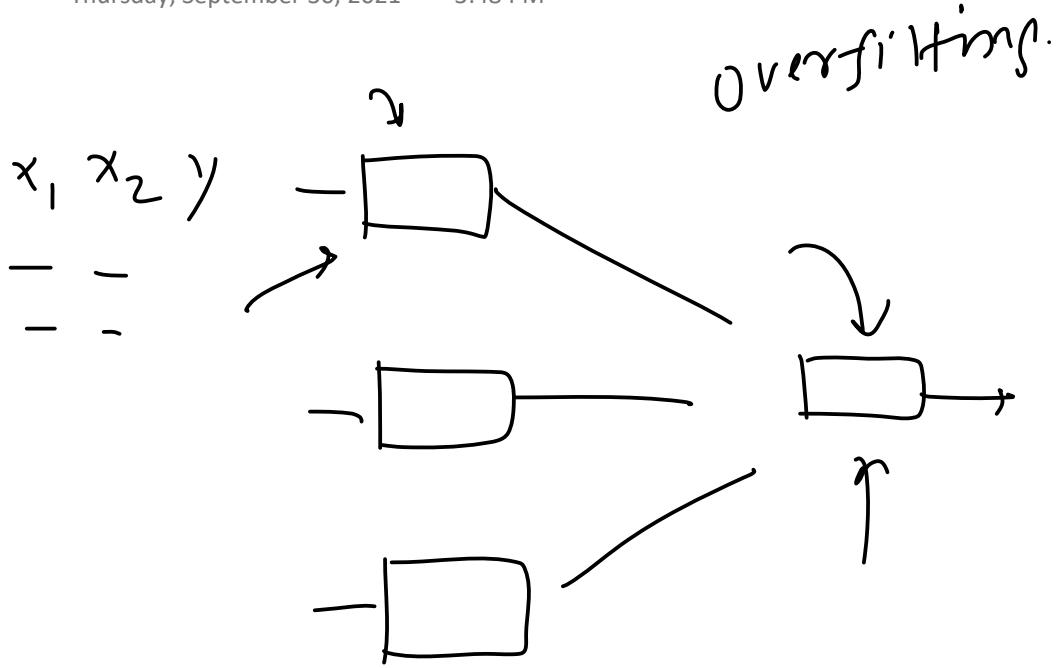
→ majority

→ training

# Problem

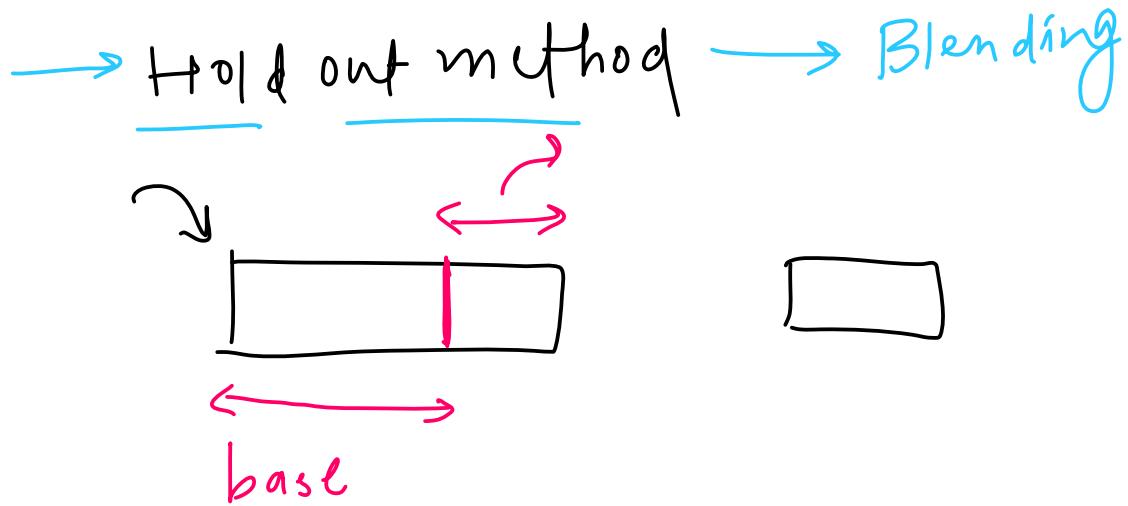
Thursday, September 30, 2021

3:48 PM



# Solutions

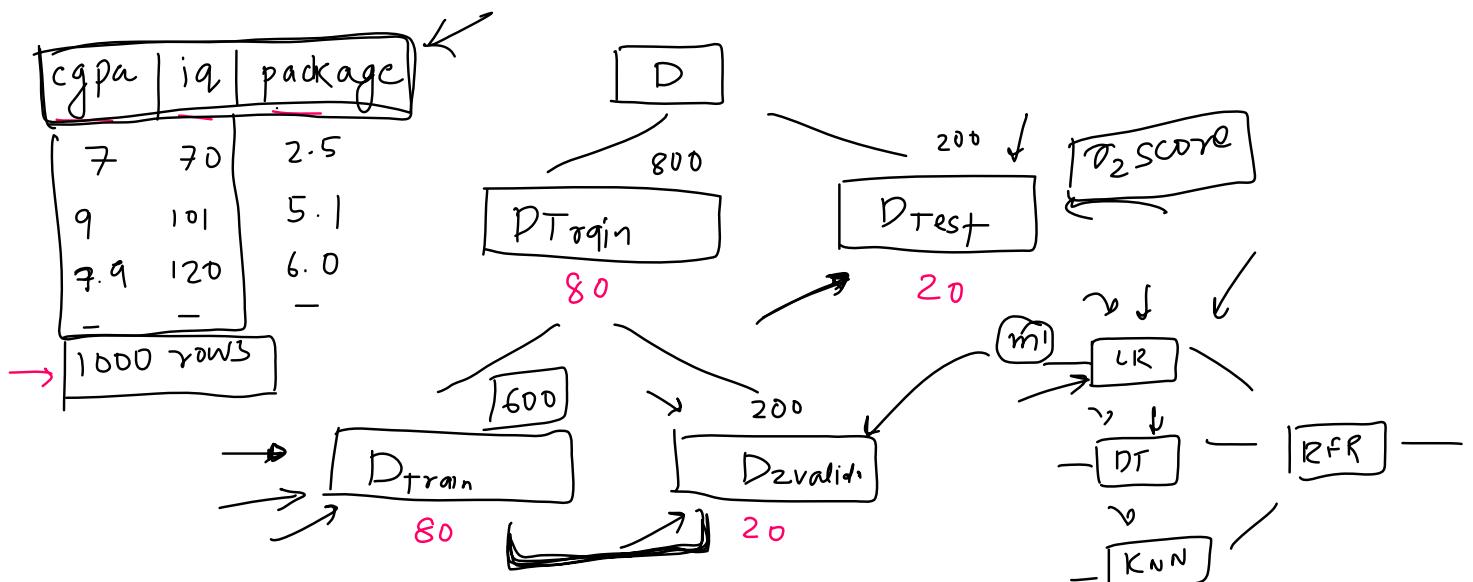
Thursday, September 30, 2021 3:48 PM



→ K-fold method → Stacking

## Hold Out Approach - Blending

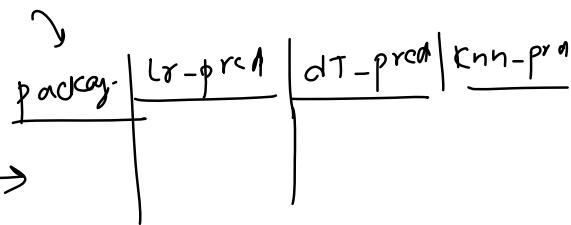
Thursday, September 30, 2021 3:48 PM



Step 1 → train 3 base modules (3 trained on  $D_{train}$ )

Step 2 → You form new dataset

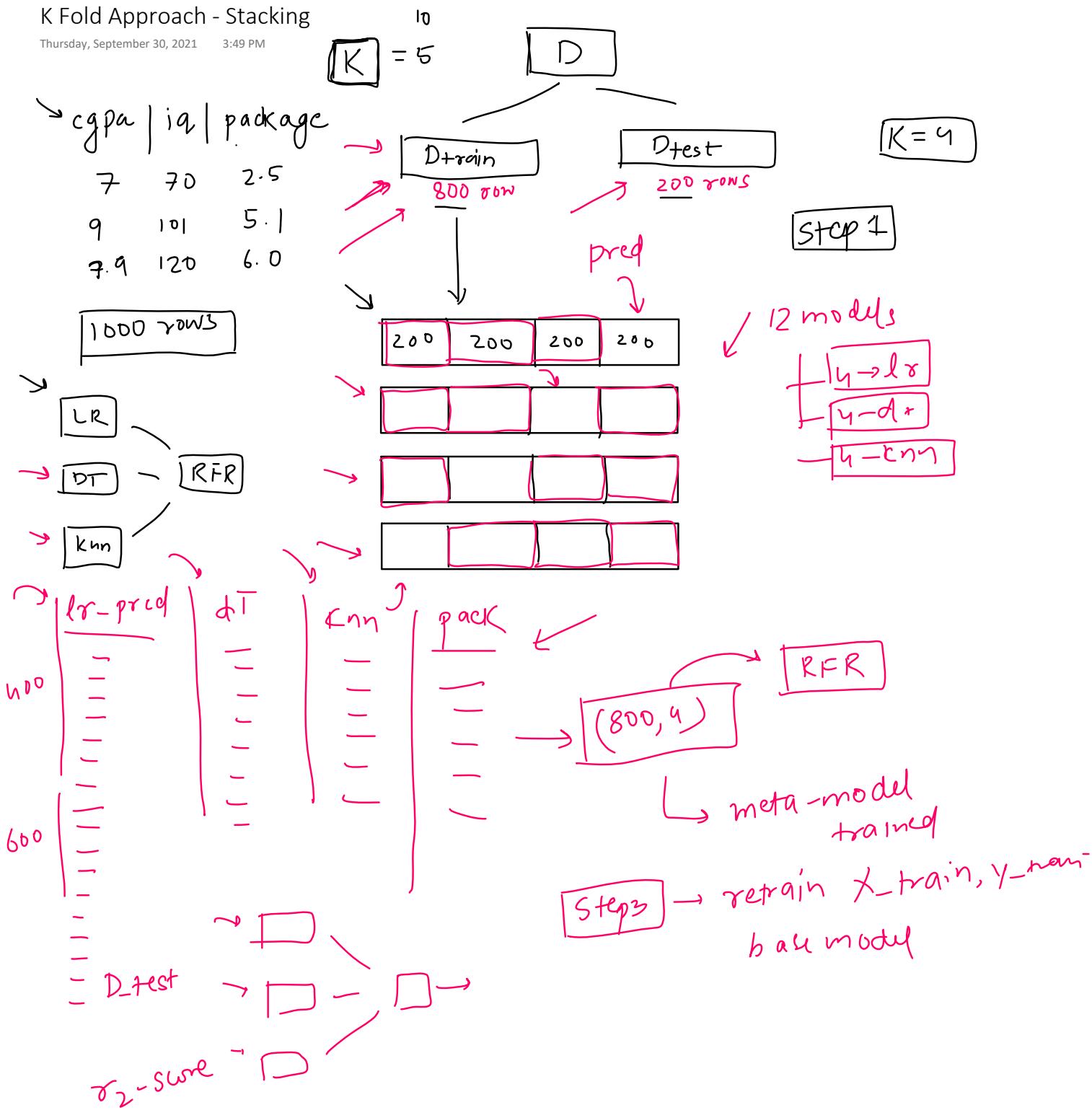
Step 3 → Train the meta-module on  
└ meta-modules



Blending

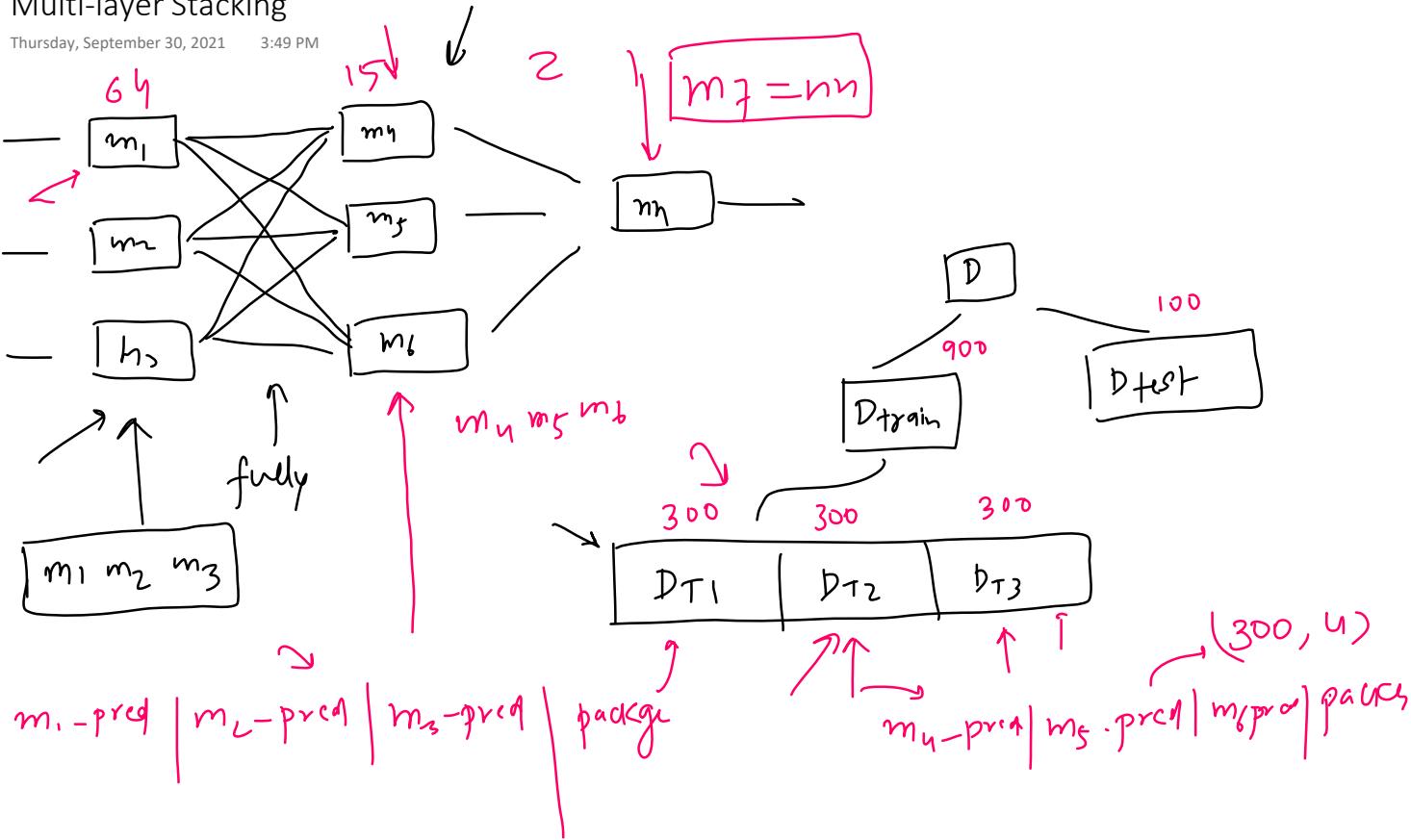
## K Fold Approach - Stacking

Thursday, September 30, 2021 3:49 PM



## Multi-layer Stacking

Thursday, September 30, 2021 3:49 PM

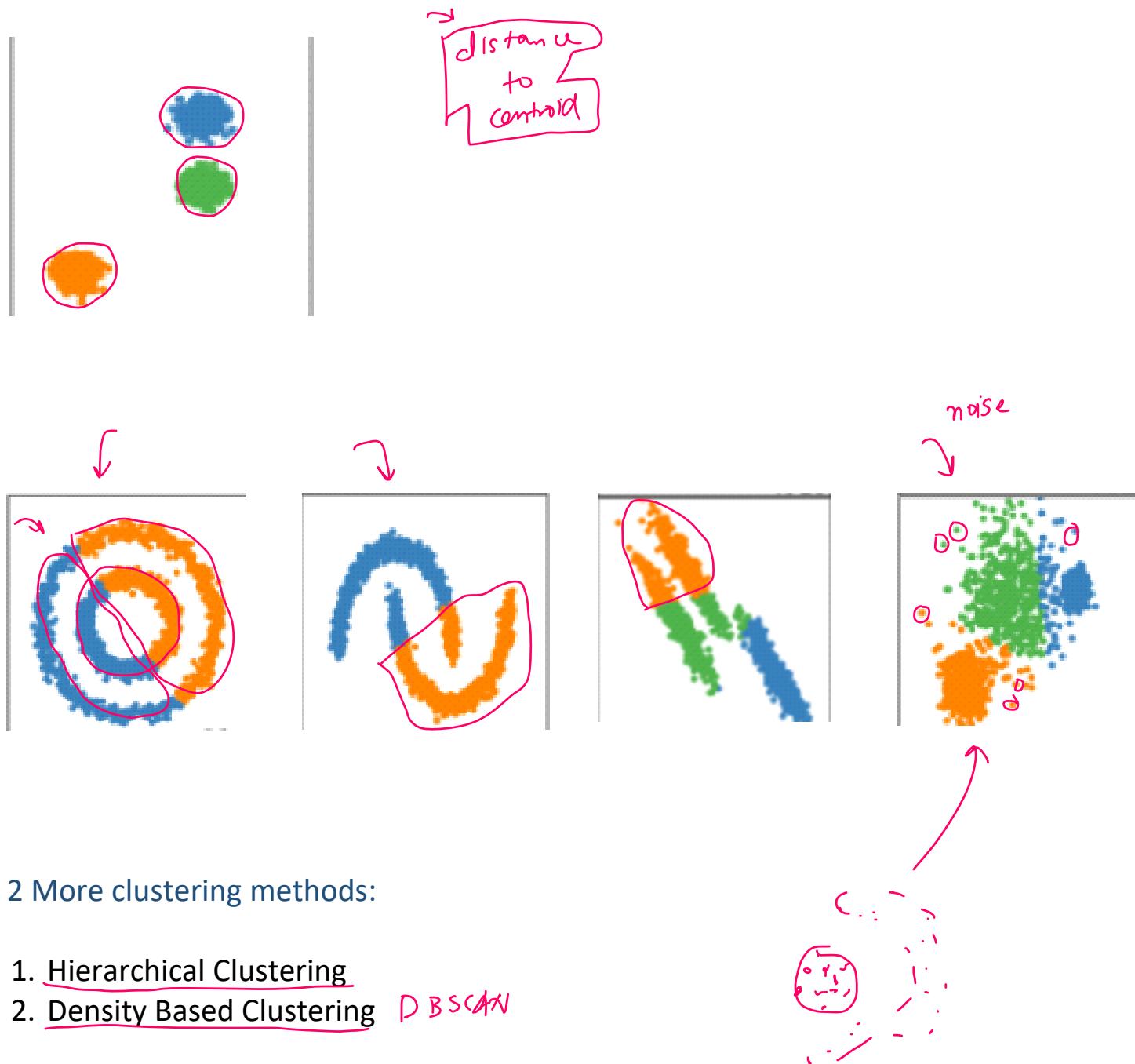


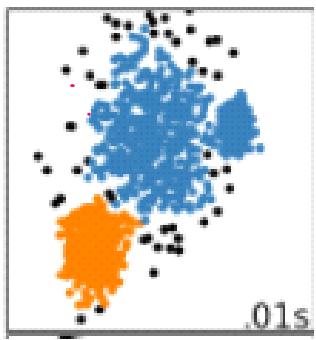
# Demo

Thursday, September 30, 2021 3:49 PM

## Need of Other Clustering Methods

Saturday, November 6, 2021 7:38 AM



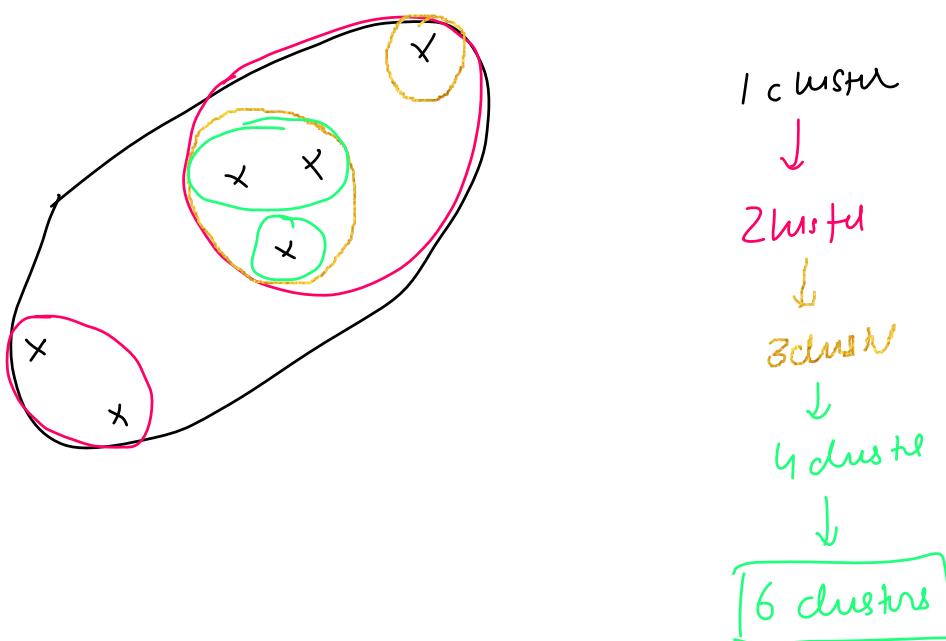
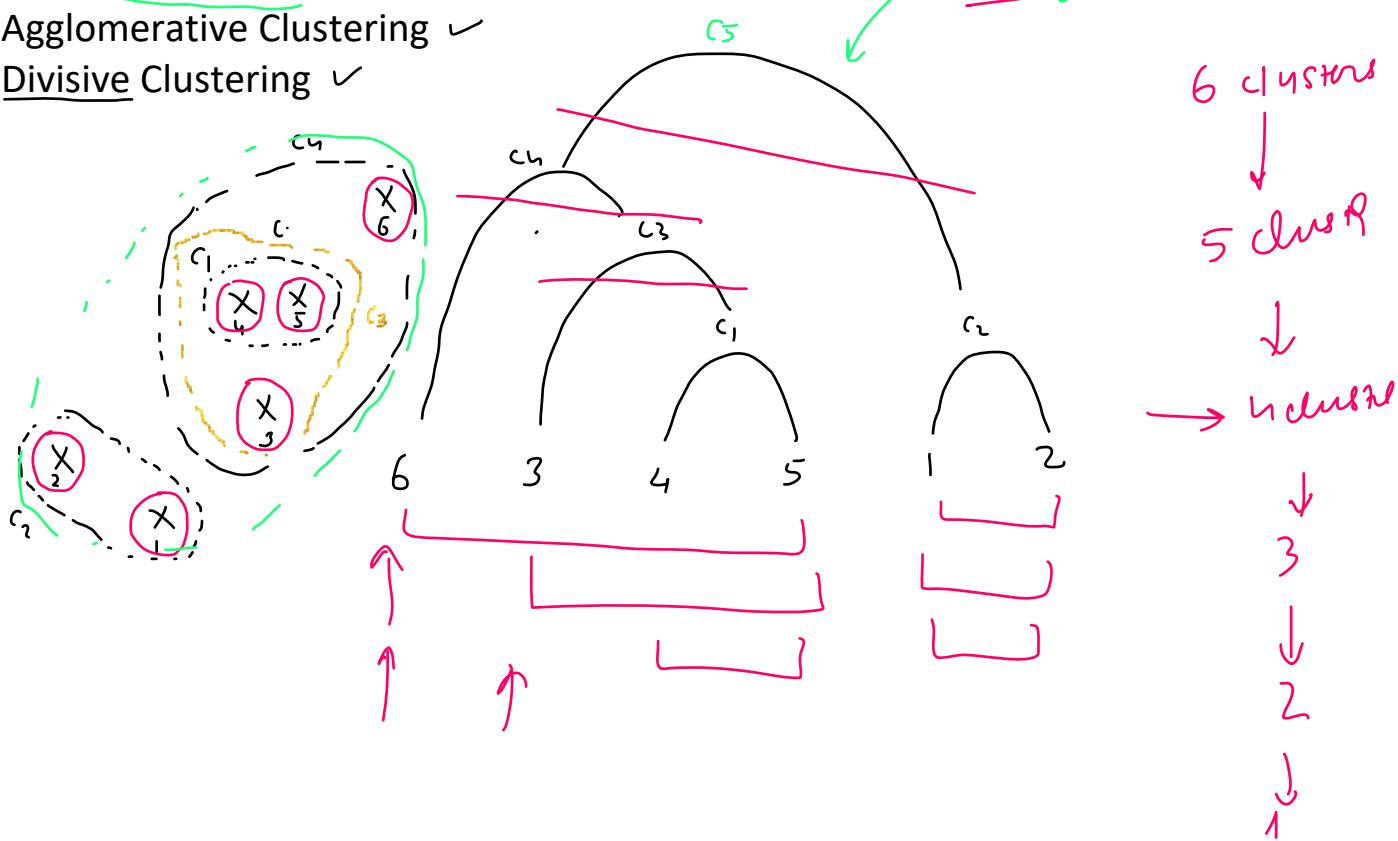


# Introduction

Saturday, November 6, 2021 7:39 AM

## Types of Hierarchical Clustering:

1. Agglomerative Clustering ✓
2. Divisive Clustering ✓



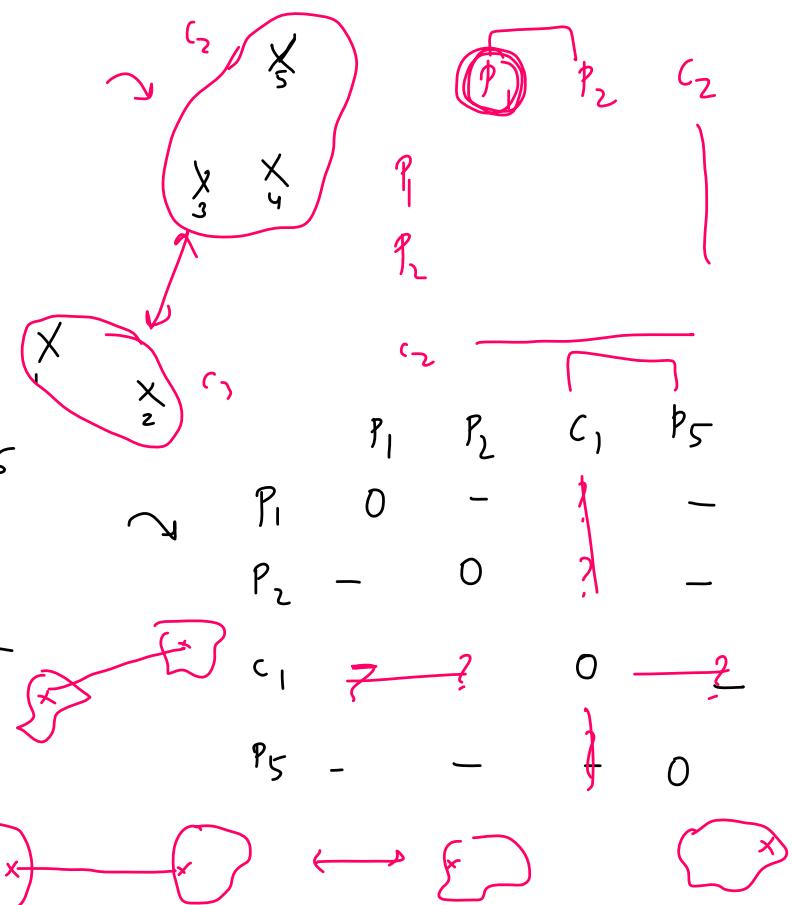
## Algorithm

Saturday, November 6, 2021 7:39 AM

1. Initialize the Proximity Matrix
2. Make each point a cluster
3. Inside a loop
  - a. Merge the 2 closest clusters
  - b. Update the Proximity Matrix
4. Until only one cluster is left

$n$  points  $n \times n$

	$P_1$	$P_L$	$\boxed{P_3 \quad P_4}$	$P_5$
$P_1$	0	-	-	-
$P_2$	-	0	-	-
$\boxed{P_3 \quad P_4}$			$c_2 \quad c_3$	



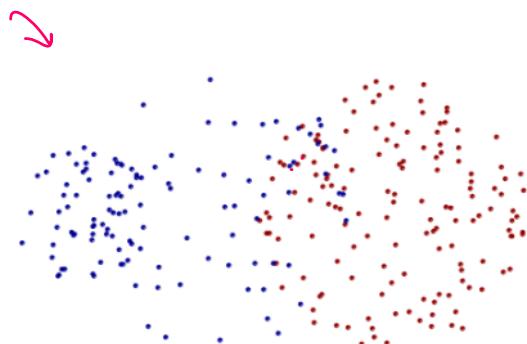
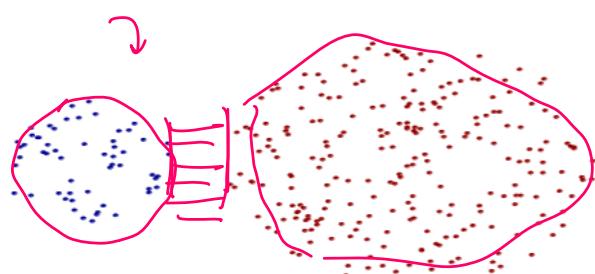
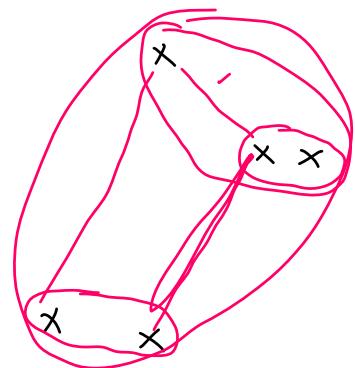
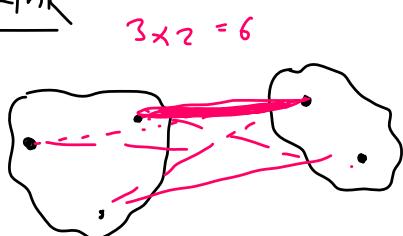
## Types

Saturday, November 6, 2021 7:39 AM

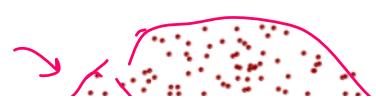
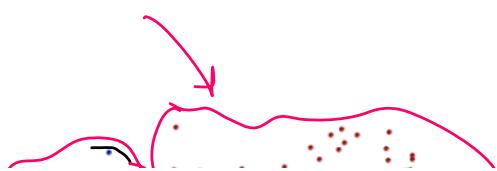
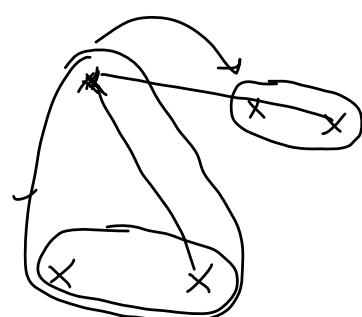
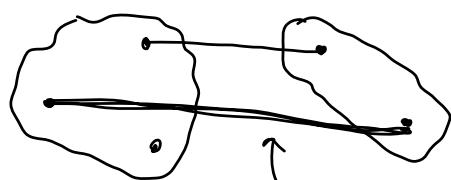
### Types of Agglomerative Clustering

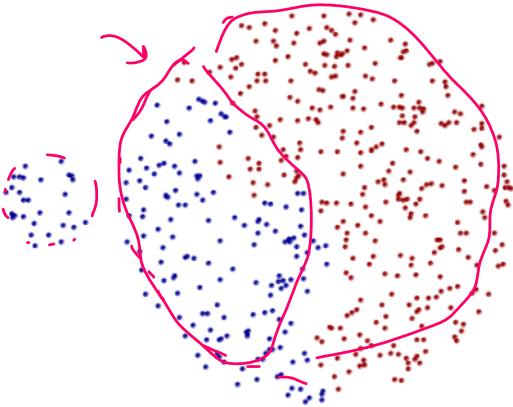
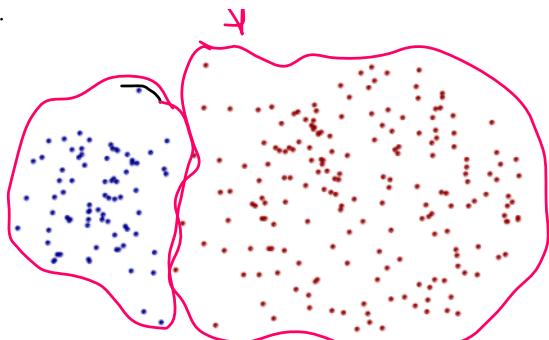
1. Min (Single-link)
2. Max (Complete Link)
3. Average
4. Ward

#### 1. Single link

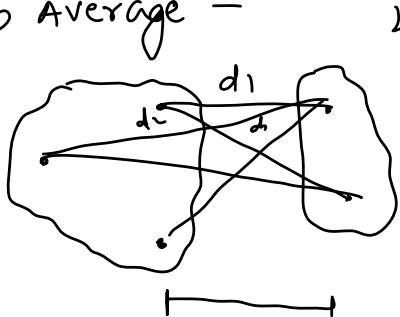


#### 2. Complete link (Max)



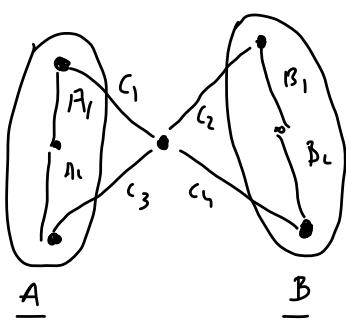


3. Group Average -



$$\frac{d_1 + d_2 + d_3 + \dots + d_6}{12 \times 3} =$$

4. ward  $\rightarrow$  SK Ward



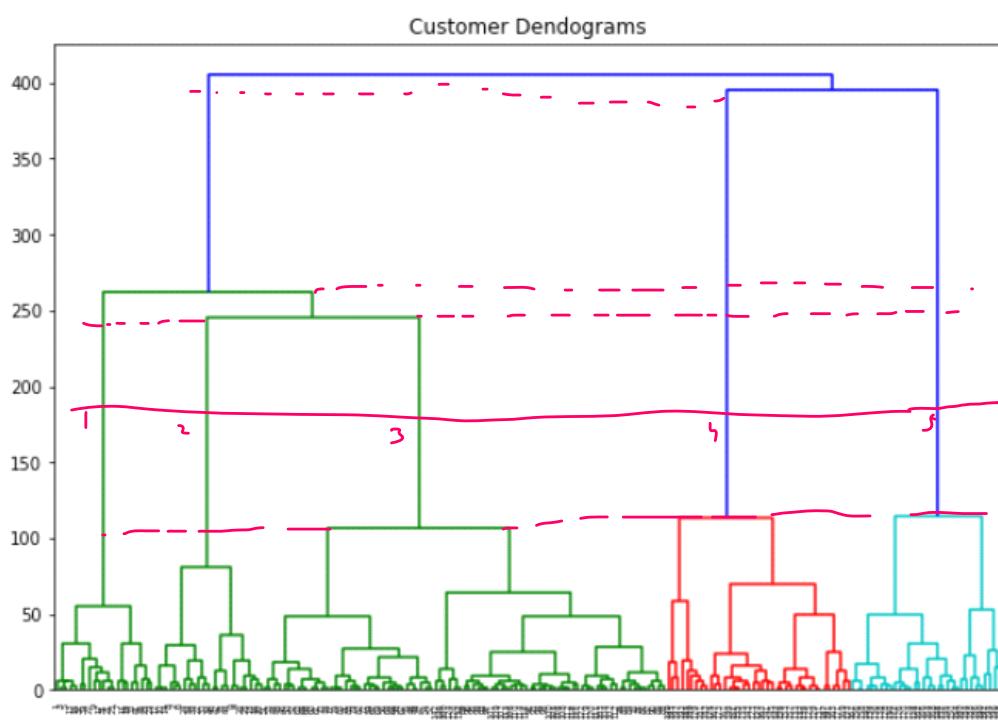
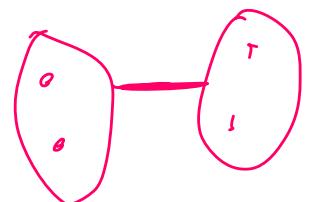
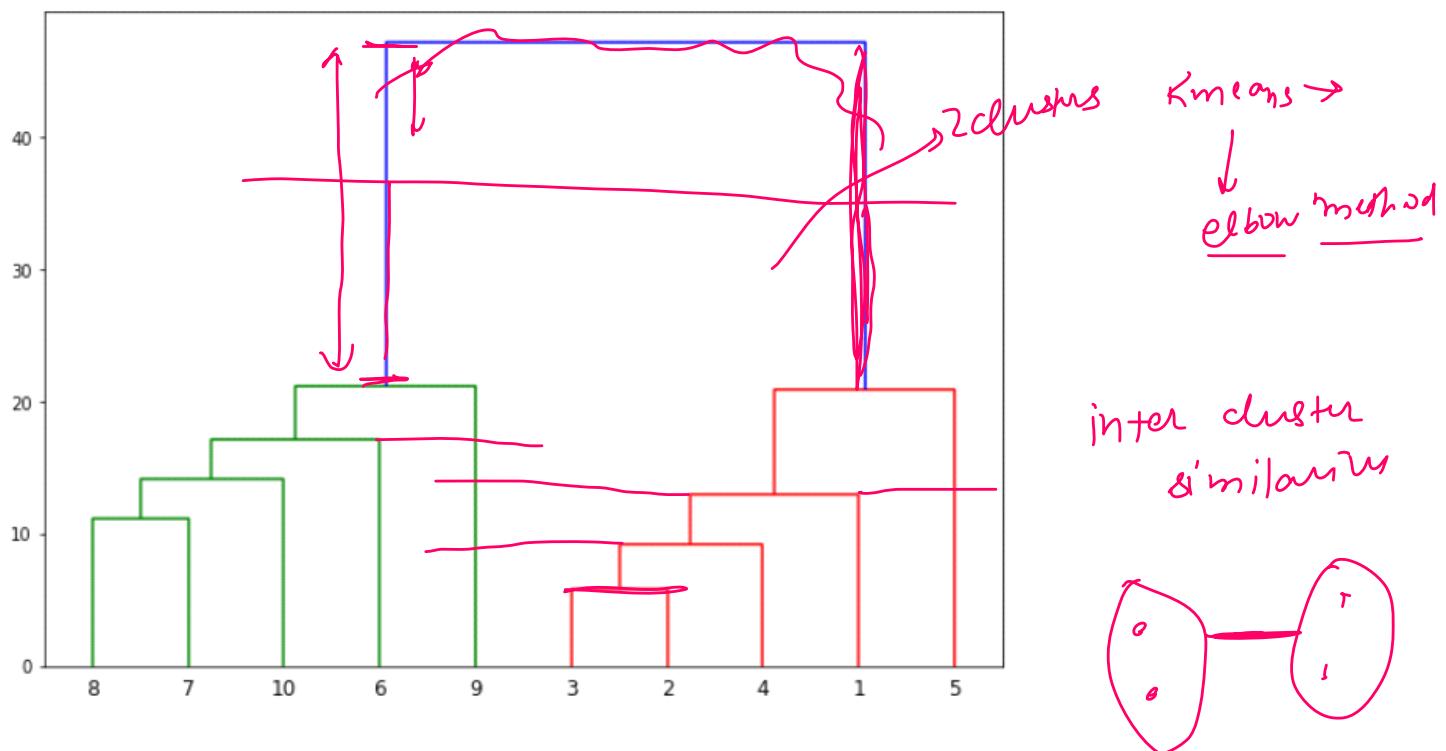
$$dist = \frac{c_1^2 + c_2^2 + c_3^2 + c_4^2 - A_1^2 - A_2^2 - B_1^2 - B_2^2}{A \rightarrow B}$$

$\leftarrow$

Variance minimize

# How to find the ideal number of clusters

Saturday, November 6, 2021 1:50 PM



# Hyperparameter

Saturday, November 6, 2021 7:40 AM

# Code Example

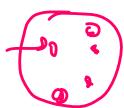
Saturday, November 6, 2021 7:40 AM

## Benefits/Limitations

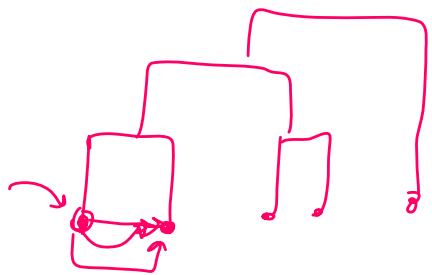
Saturday, November 6, 2021 7:39 AM

### Benefits

1) Widely applicable



2) Dendogram



### Limitations

...  
 $n$

$n$

$n \times n$

$10^6$

$10^{12}$  bytes

$10^{43}$

RAM