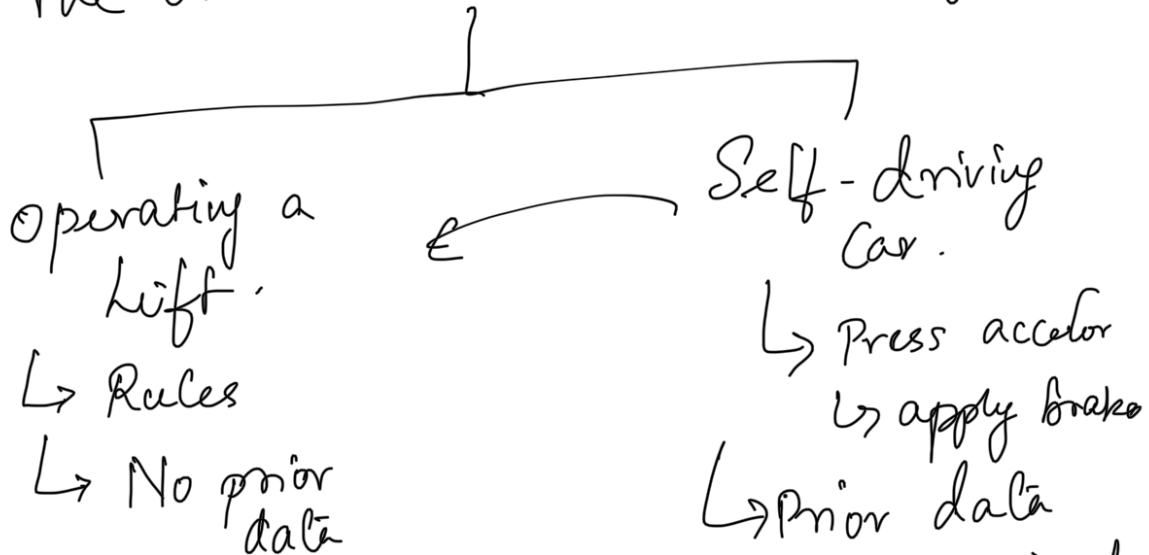


## Linear Regression Recall...

# ① Machine Learning

↳ the machine learn a task.

↓  
The overall Essence → Learning



Empirical Learning/  
Machine Learning

↳ More & different  
Scenarios.

If you don't have data → No performance.

## ① Statistics → Understand the data

↳ phenomenon that have generated  
the data

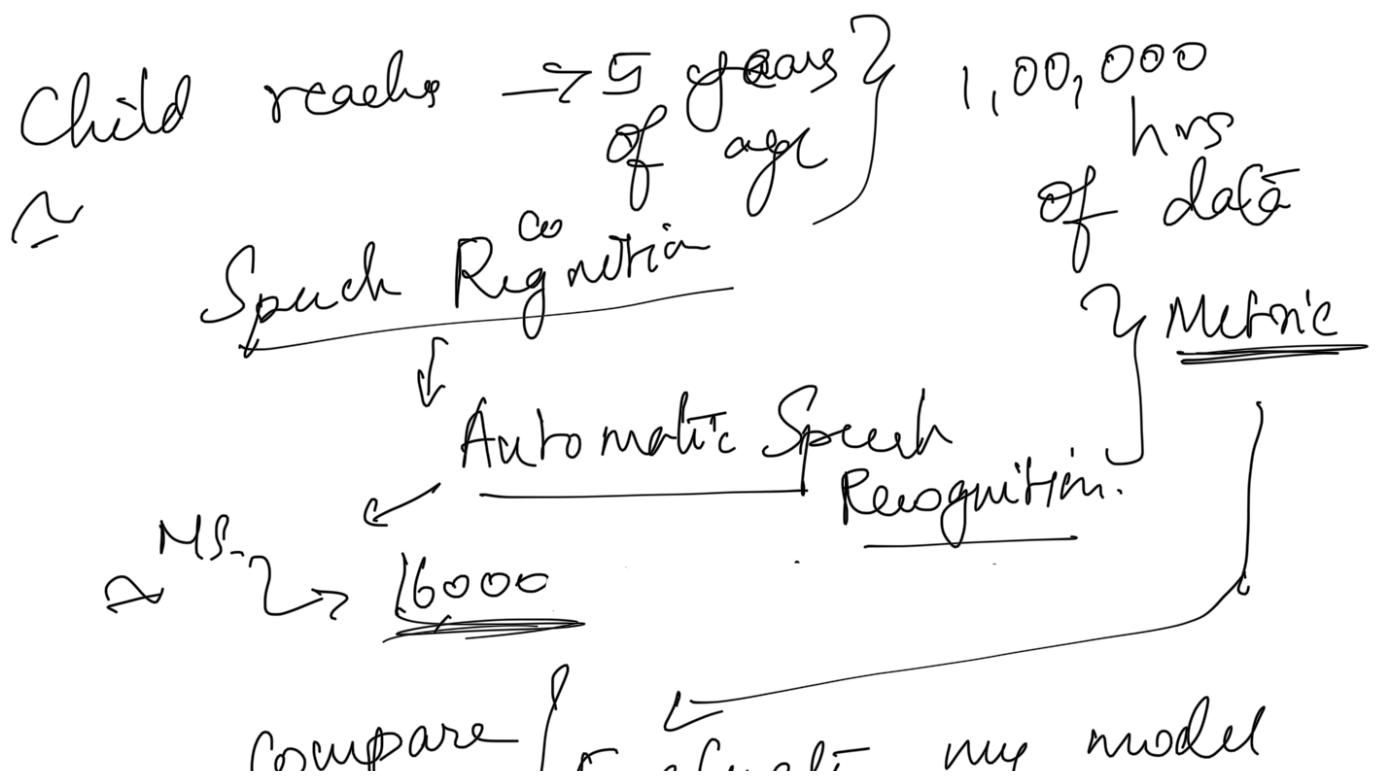
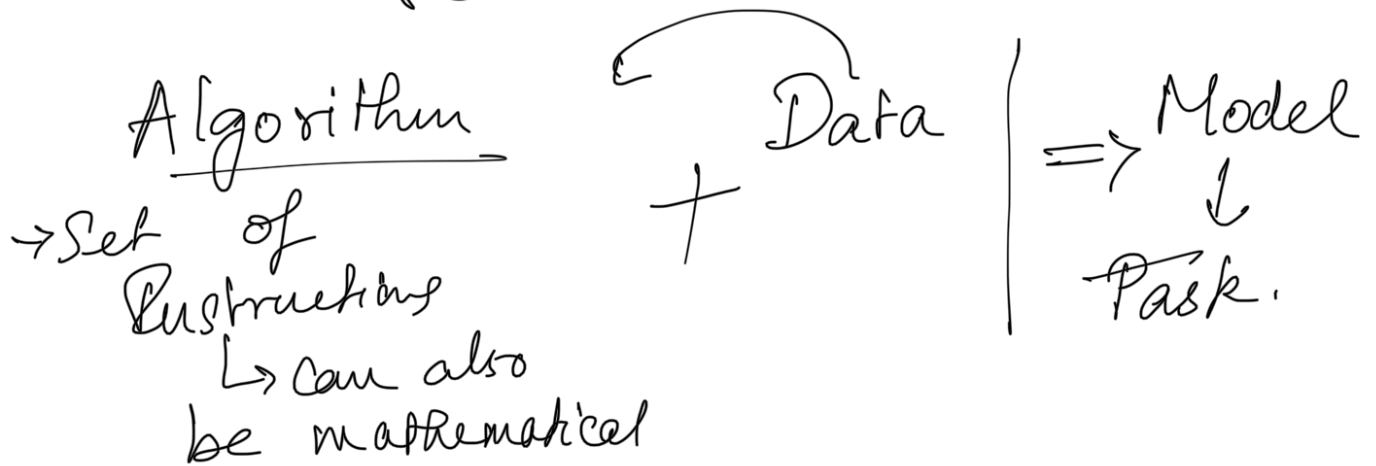
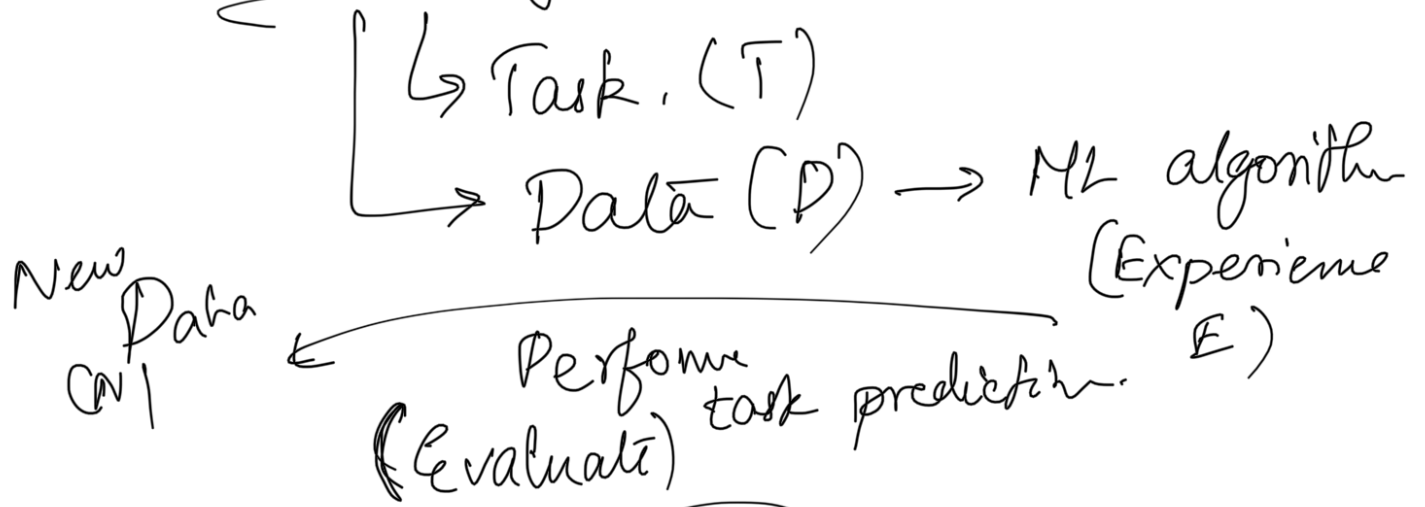
↳ often have understanding to  
perform task → different  
hypothesis

## ② psychology — Understand behaviours exhibited by people.

## ③ Data Mining — Seeks to find the

pattern inside the data that is understandable by people.

⇒ The Learning



metric

1 Evaluation

← WER

3/10

① Task

② Data →

Adding more data

③ feed the Data with an

algorithm

④ Evaluate the model (predictions)

⑤ perform prediction in real time  
↓  
taking it live

Meal

1

2

3

4

5

6

Tip amount \$

5.00

17.00

11.00

8.00

14.00

5.00

One variable

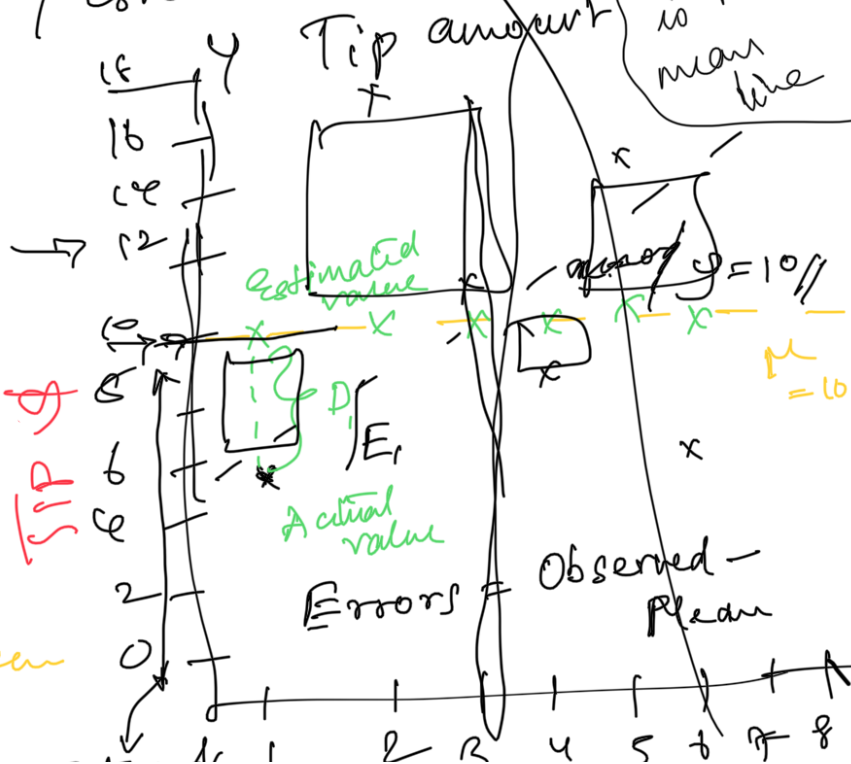
① Task → predict

Estimate future

② Data | y-

③ C.T

④ Evaluate the model



Can I make use of Mean at C.T??

Intercept

Mean #

Variance

$$\frac{\text{Sum of Squares}}{n} \Rightarrow \frac{\text{Sum}(\text{Observed} - \text{Mean})^2}{n}$$

Deviations

Mean Sum of Squares

Sum of Squared Deviations

Error

Mean Sum of Squared Deviations

Error

Before we proceed can we do better??

Benchmark. Baseline

SSE

SSE

SST

120

Algorithm → reduces my SSE

Competing agent

$$y = 10$$

rate of change =  $\frac{\text{Change in } y}{\text{change in } x} = \frac{0}{\infty}$

vertical line

$x = 3$

rate of change =  $\frac{\infty}{0} \rightarrow \infty$

Function

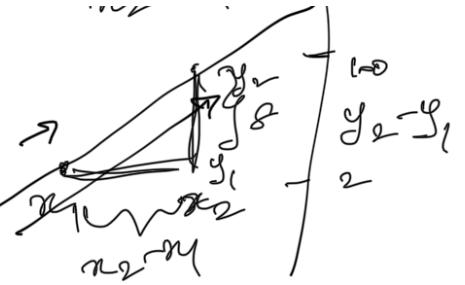
$\Rightarrow \Delta y \rightarrow \text{delta } y$

$\frac{y_2 - y_1}{x_2 - x_1}$

Slope =  $\frac{\text{rise}}{\text{run}}$

$\Delta x$

$y = \text{Slope} \cdot x + \text{Intercept}$

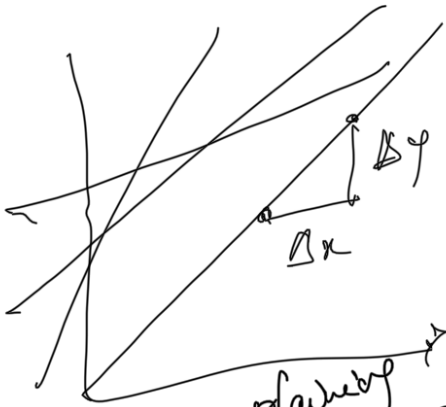


define your target as a function of  $x$

$y = mx + c$

$m$  ← slope

$c$  ← Intercept



Total bill (\$)

34.00  
108.00  
64.00  
88.00  
99.00  
51.00

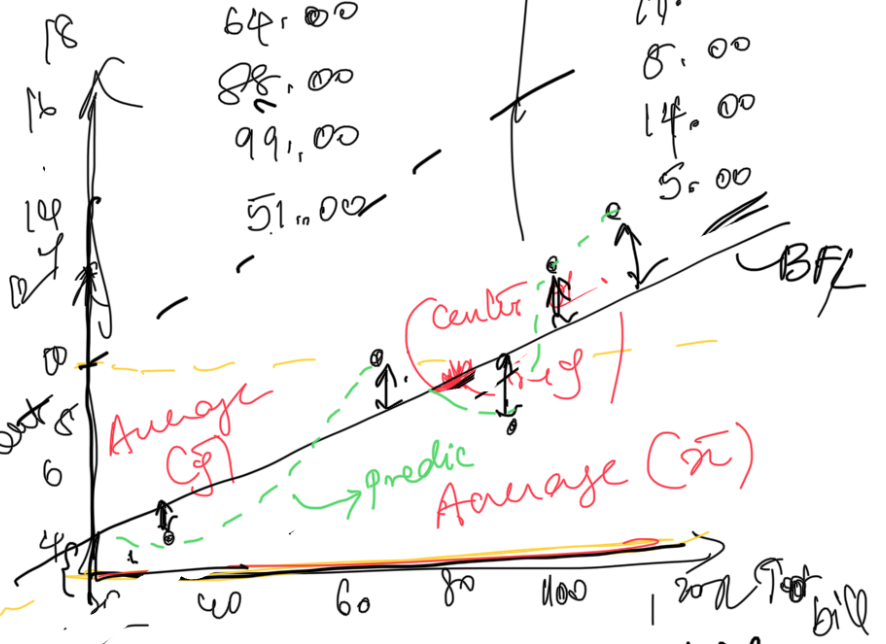
Tip amount (\$)

5.00  
17.00  
11.00  
8.00  
14.00  
5.00

Regression explaining Relationship correlation

Brute force method

Between iterations



Best fit line → line which has min Error.

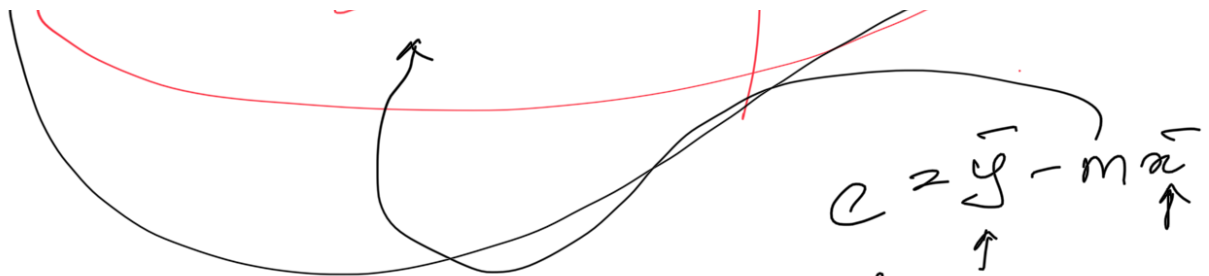
$$m = \frac{\Delta y}{\Delta x}$$

$$y = mx + c$$

$$c = y - mx$$

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

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



$$SSE = 30.075$$

$$SSE = SST = 120$$

$$120 \rightarrow 30.075$$

$$120 - 30.075$$

$$89.925$$

new  
Regression  
line  
  
old  
Mean  
line

$$\begin{aligned} SST &= 120 \\ SSE &= 30.075 \\ SSR &= 89.925 \end{aligned}$$

Total Error is Explained by my regression line  
of the Error is Explained by my regression line  
how good is my

Goodness of fit

Coefficient of Determination

$$r^2 = \frac{SSR}{SST}$$

Root-  
how good is my

$$\frac{no}{no} = 1$$

$$r^2 = 1$$

Best Model  
New/Reg. line

89.9

Assumptions for Linear Regression.

- (1) Linear Relationship bet  $X$  &  $Y$ .
- (2) Data should have little to no  
Multicollinearity - Write up on why/how  
Multicollinearity is bad!

Several Independent Variables (X) are correlated among themselves

2 variables are perfectly co-linear  $+1/-1$  affect the statistics inferred bet.

$$X \rightarrow Y$$

Computes Coefficient

read on how/why write an article

Sample !!!

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + \dots + \epsilon$$

Multicollinearity with influence on when computing this coefficient.

(3) Residuals should be normally distributed



(4) Residuals should have homoscedasticity

Homoscedasticity  $\rightarrow$  Same variance the errors should have similar variances.

(4) ~~infinite - old on Assumptions of linear regression.~~



VV VV W  
|  
L Zwischenapp :  $\rightarrow$  gloss

V MF  $\rightarrow$  Variation <sup>dann</sup> Seiflatten  $\hookrightarrow$   $\infty$   
Berechnung