Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 01

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic
- Topic
- Topic
- Topic
- Topic

clone

Topics to be Covered

Topic • Basic of ML

Topic • Steps in ML

Topic • data, optimisation, Model

Topic →

Topic

Syllabus

Machine Learning: → Syllabus

Basic
↓
advance

(i) Supervised Learning: regression and classification problems, simple linear regression, multiple linear regression, ridge regression, logistic regression, k-nearest neighbour, naive Bayes classifier, linear discriminant analysis, support vector machine, decision trees, bias-variance trade-off, cross-validaθon methods such as leave-one-out (LOO) cross-validation, k-folds cross- validation, multi-layer perceptron, feed-forward neural network;

(ii) Unsupervised Learning: clustering algorithms, k-means/k-medoid, hierarchical clustering, top-down, bottom-up: single-linkage, multiple- linkage, dimensionality reduction, principal component analysis.

Prerequisites

Prerequisites

- Mathematics (Linear Algebra) → Matrix
- Mathematics (probability) →
- Mathematics (basic Calculus)

$\frac{d}{dx}$ | int.

Basics of Machine Learning

- What is Machine Learning
- What is optimisation
- What is a model

⇒ Gate 2025 ⇒ "Podenge" Boom → Mtech

"Your positive action combined with positive thinking results in success."

← Prepare

Basic → high level

Matrix

Probab

focus → "Gate" → INPTEL

→ Question → Gate

College Exam ✓

Class H.W / app

More than enough Gate

- ML ⇒ 8-10 marks

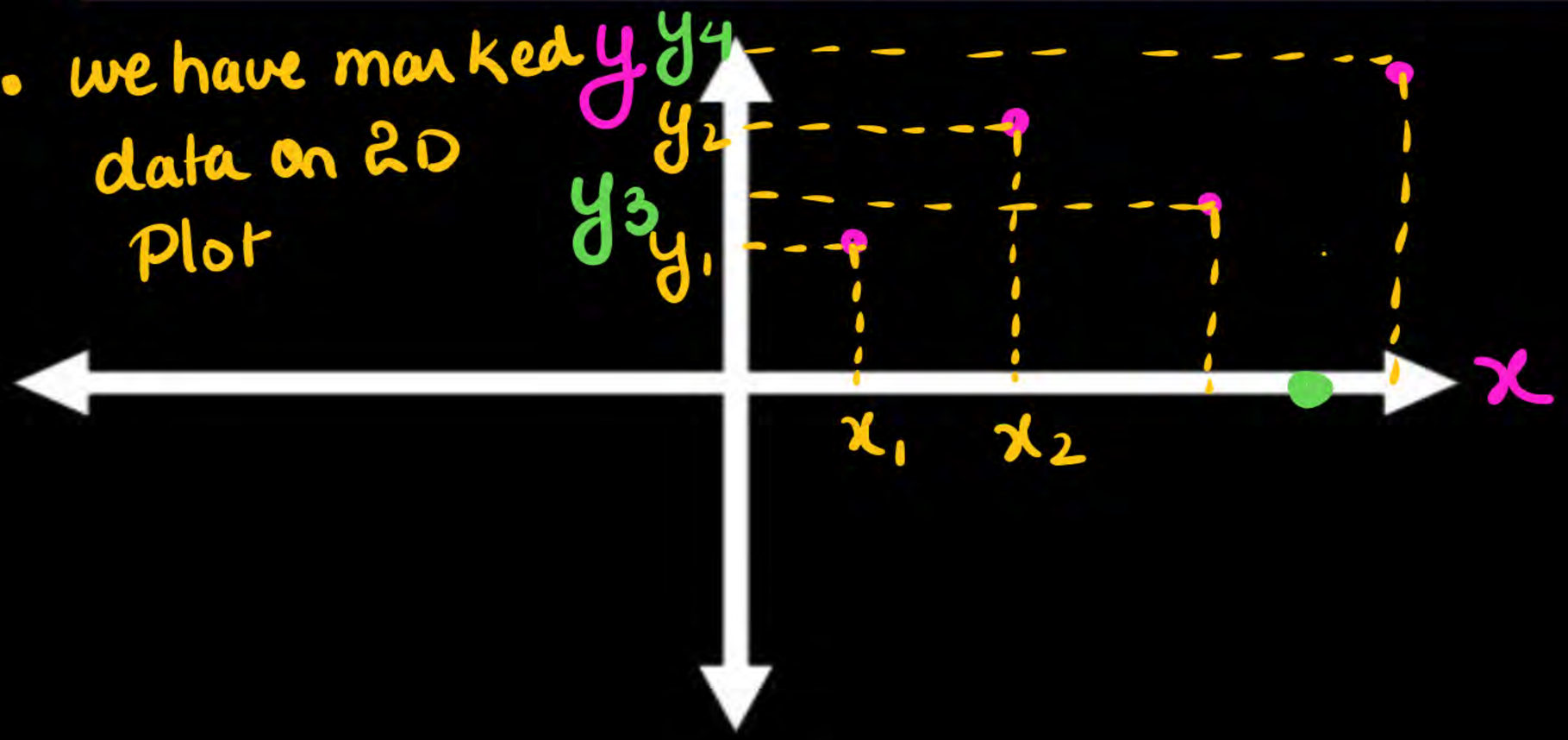
- (2 Hour + 50 lectures ⇒ 100-105 Hours)

  Mon- friday + weekends

What is the General Problem in Machine Learning

Lets start with a basics

- we have marked data on 2D Plot



we have 4 data Points

(x_1, y_1)

(x_2, y_2)

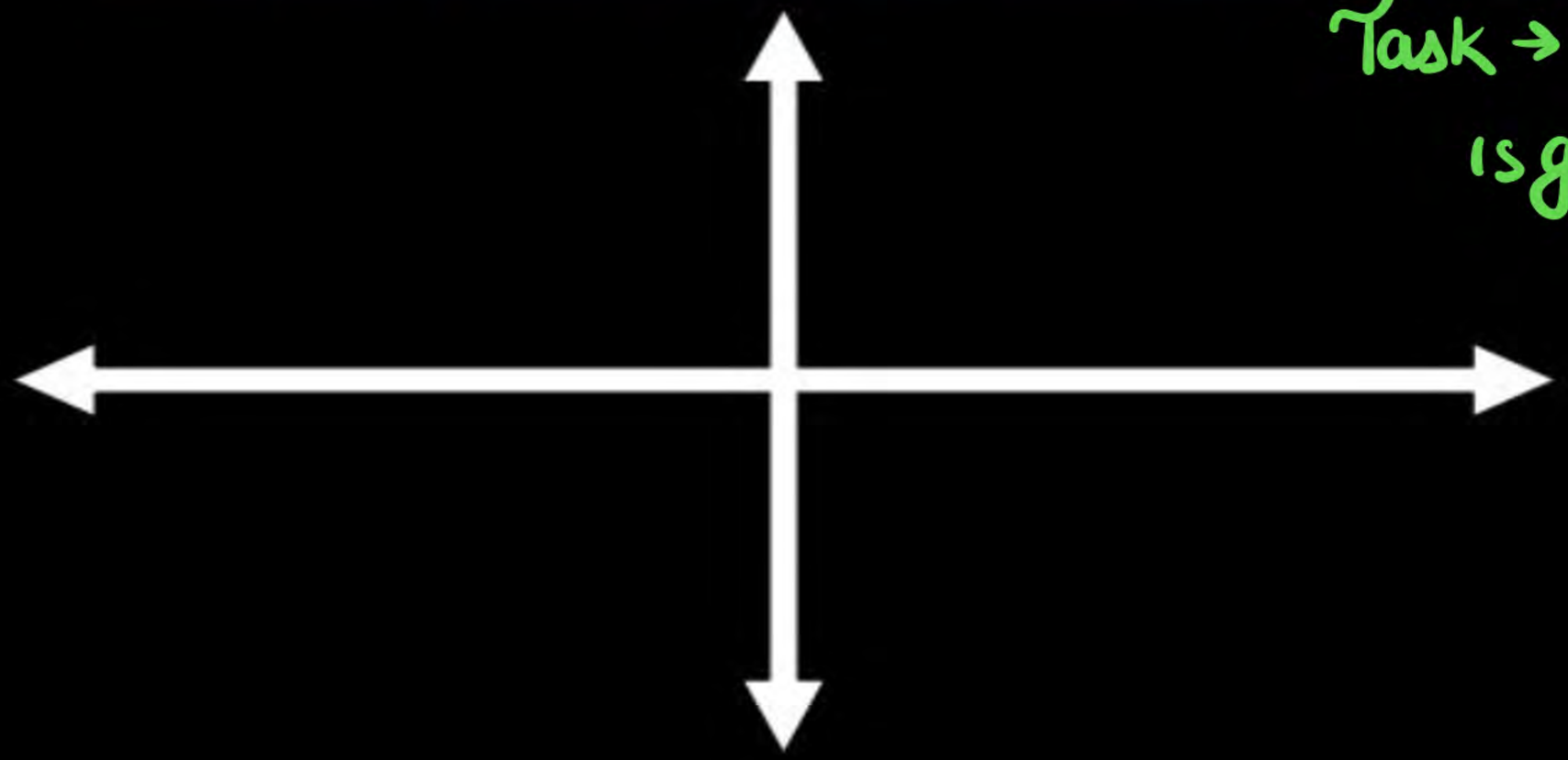
(x_3, y_3)

(x_4, y_4)

What is the General Problem in Machine Learning

How you will predict/ find the value of y for new x

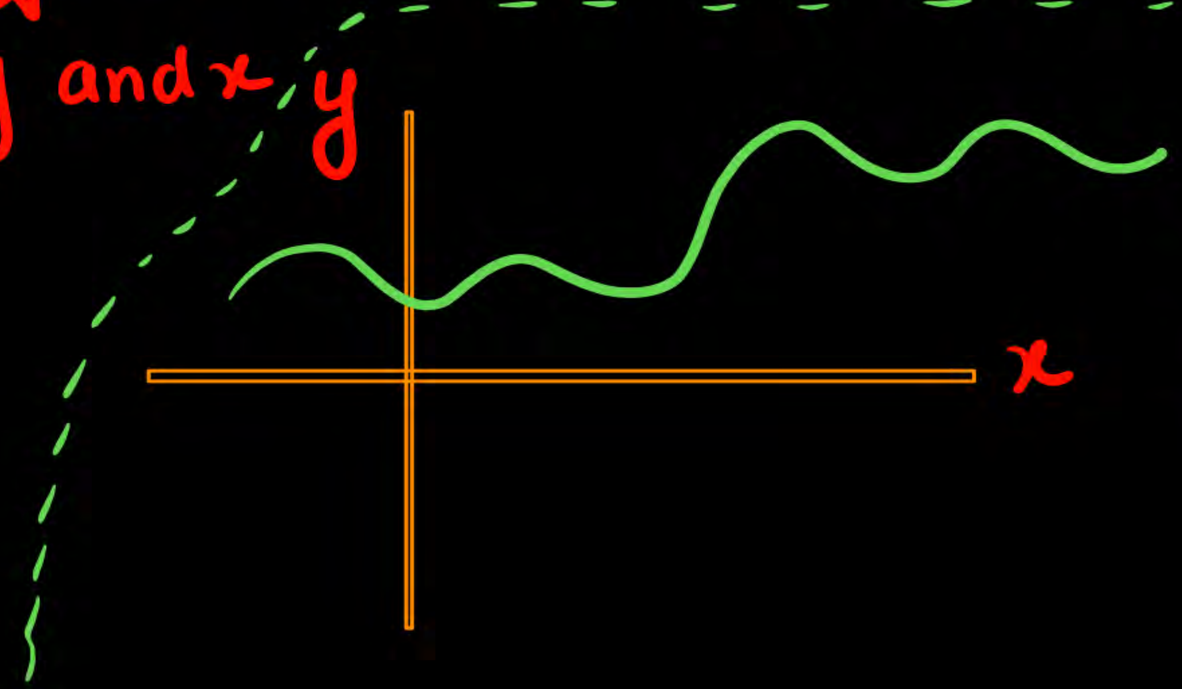
Task → if any new point is given find its 'y'



What is the General Problem in Machine Learning

This is called a function of _____ variable

⇒ So using the available data we will find relation b/w y and x



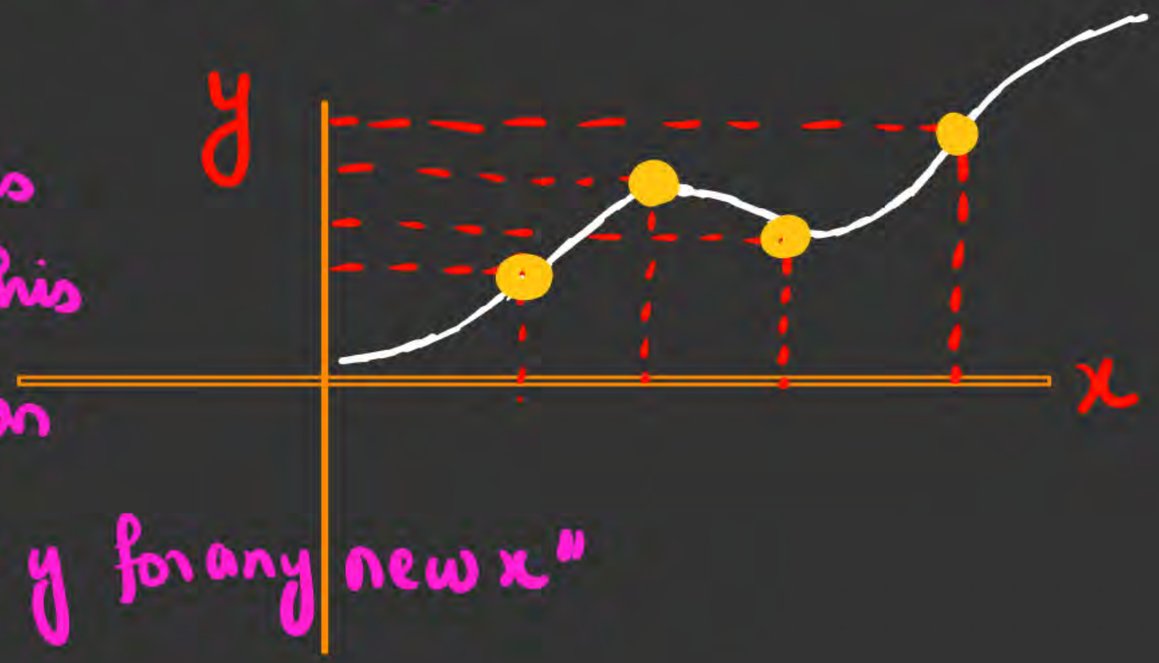
⇒

← this plot

y = function of x

$y = f(x)$

So from the available data we try to find the relation b/w y and x i.e $y = f(x)$



"Now using this plot, using this y, x Relation we can find y for any new x"

⇒ How to find plot b/w y and x

⇒ we are open to Choose any plot

⇒ So We will draw a Random plot.

What is the General Problem in Machine Learning

These already known points relating Y and X is called data

So, * So data was available in (x, y) format

x → data value

y → this is the label or the output

* Target was ⇒ To predict value of y for new 'x'.

* How to do this ⇒ So using the data we find relation b/w y and x. $y = f(x)$

What is the General Problem in Machine Learning

So using the data we learn a relation between Y and X and this process is called Training process.

⇒ So the Training Process ⇒ To find the Relation b/w y and x.

⇒ $y = f(x)$ ⇒ dependent Variable → 'y'

Independent Variable → 'x'

⇓

y is the function of Single Variable

What is the General Problem in Machine Learning

What is Data

It refers to the set of observations or measurements that can be used to train a machine-learning model.

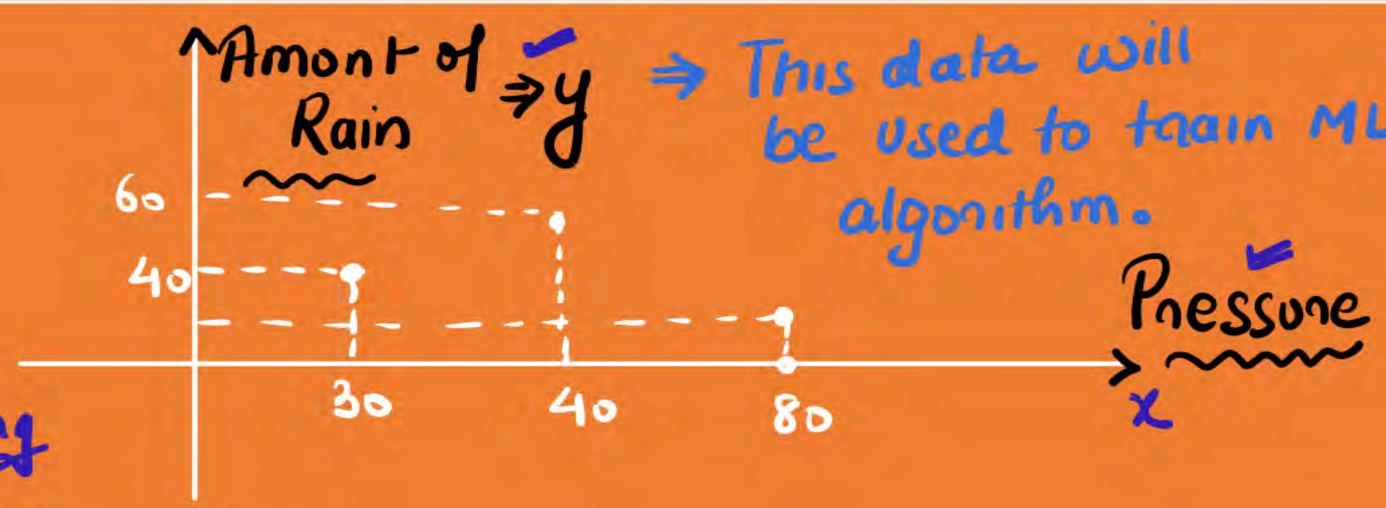
⇒ example

⇓

How data is Collected

⇓

- Survey ✓
- buy ✓
- expert
- web sourcing



⇒ This data will be used to train ML algorithm.

What is the General Problem in Machine Learning

How we collect data

done → 4 method

What is the General Problem in Machine Learning

Lets take an example of predicting rain

- Rain depend on may factors → we have to choose only 1 or 2 factors
- How we decide 'y' and 'x' in data → This depend on the goal

y	x
Size of House	Price of House
⋮	⋮

Now Q find out a function to predict value of Size of house if price of house is given.

Since we have to predict $\underbrace{\text{Size of house}}_{y}$ for new $\underbrace{\text{Price of house}}_{x}$

Here we will find Relation b/w Size of house $= f($Price of house$)$

What is the General Problem in Machine Learning

Lets take an example of predicting rain

done

What is the General Problem in Machine Learning

What is the problem Statement

Task | Goal → That is to find y for new x.

To do this →

What is the General Problem in Machine Learning

What is a Model

we find Relation b/w y and x

i.e. $y = f(x)$

This function is obtained after training

This is called the Model

What is a Model → It is Simply the fxn or Relation b/w y and x obtained after training

data

data is
Process/ analyze

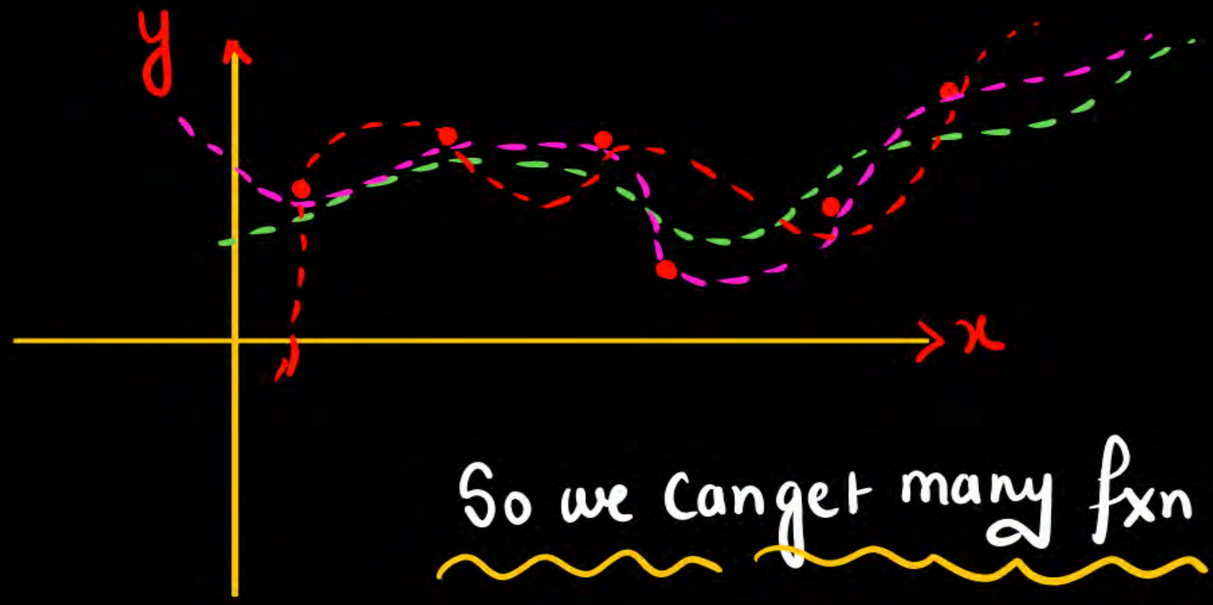
and then we find
Relation b/w $y - x$

Model
$y = f(x)$

• The process involved is called the training Process } ⇒ we will study.

What is the General Problem in Machine Learning

What is the Optimization



← data is marked on 2D plane

Harish Model

Parthiv model

Sid model

So we can get many fxn (models) after training → So the optimization step is done to find the best model.

Basic Understanding – Predict Rain in Your city

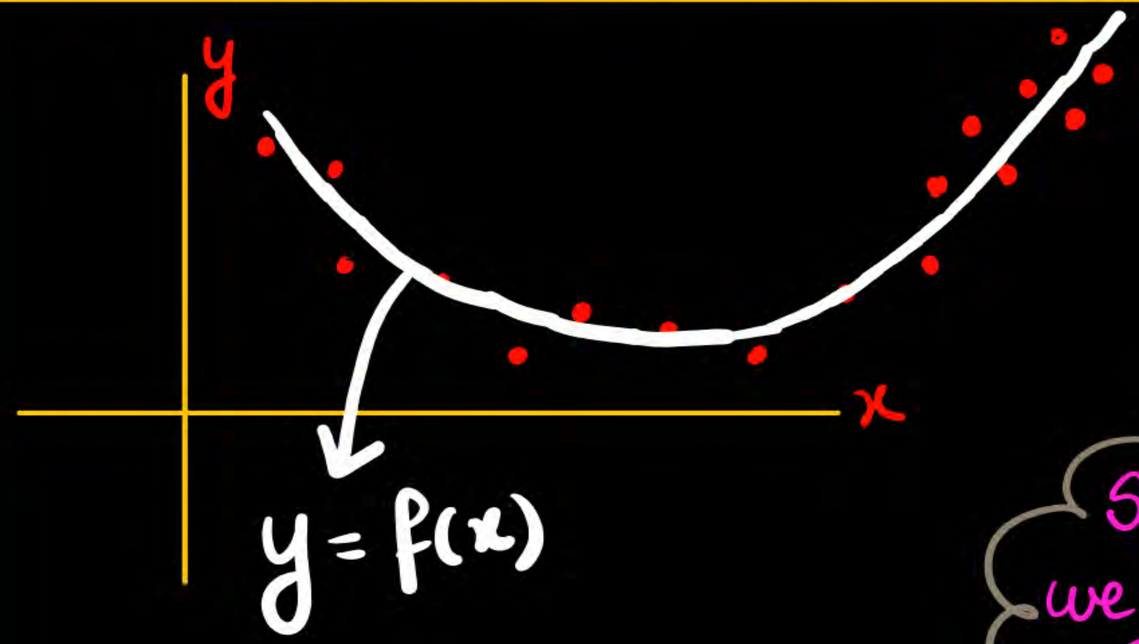
What should be the format of the data ?

Skip

We have two type of Data : -

Basic Understanding – Predict Rain in Your city

What should be the format of the data ?



- So the model i.e $y = f(x) \Rightarrow$
- CID $\Rightarrow$

So in ML we try to find Pattern of data

- we donot have to follow Step by step

• this is the available data.

CID ⇒

• Abhar ⇒

• Preetam ⇒

• Which model is Better ⇒ Preetam model ⇒

• in Both the model we have some error bcoz the predicted value of y by the model and actual value of data point are different.

which is a better model

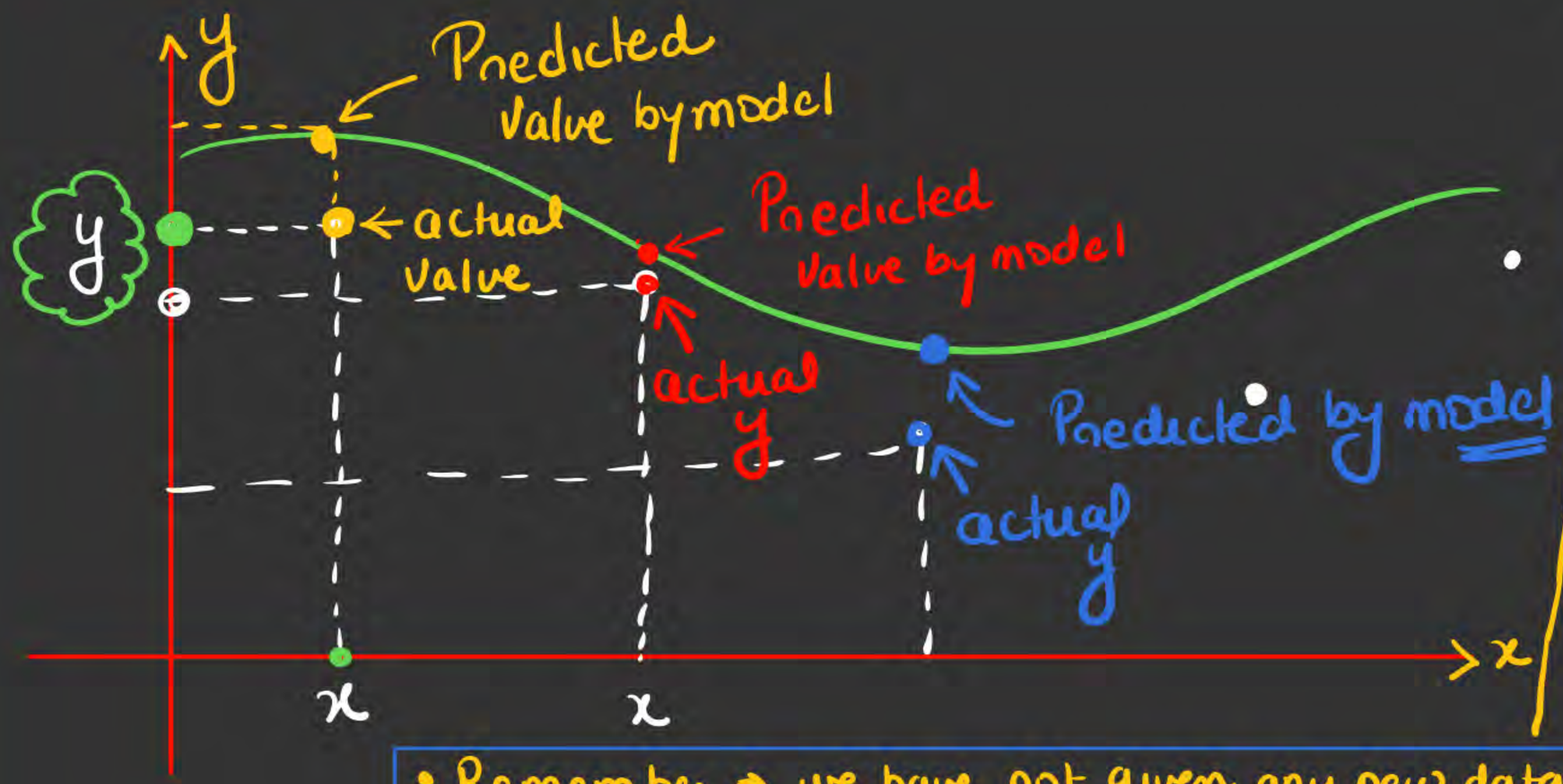
⇓

So we have more than 1 model

⇓

The best model is that which give min error b/w the actual y value and y value given by model

• → data available



Here we are finding the goodness of model by looking at y_{actual} and $y_{predicted}$

• Remember → we have not given any new data point

So after training we can get many model ⇒ • optimization.

Best model ⇒ • The model which give min "error"

i.e min gap b/w y_{actual} and

$y_{predicted}$ by model.

• Remember here we are talking about the already available data ⇒ Also called as training data.

Basic Understanding – Predict Rain in Your city

- Now you must create a mathematical model to predict rain in your city?

So basically model will try to learn the pattern of the data

Model is Simply mathematical function that try to minimize the error on the Training data.

Concept

- function complexity inc

$\Rightarrow y = mx + c \approx$ Plot :- Straight line

$\Rightarrow y = ax^2 + bx + c \approx$ Parabola

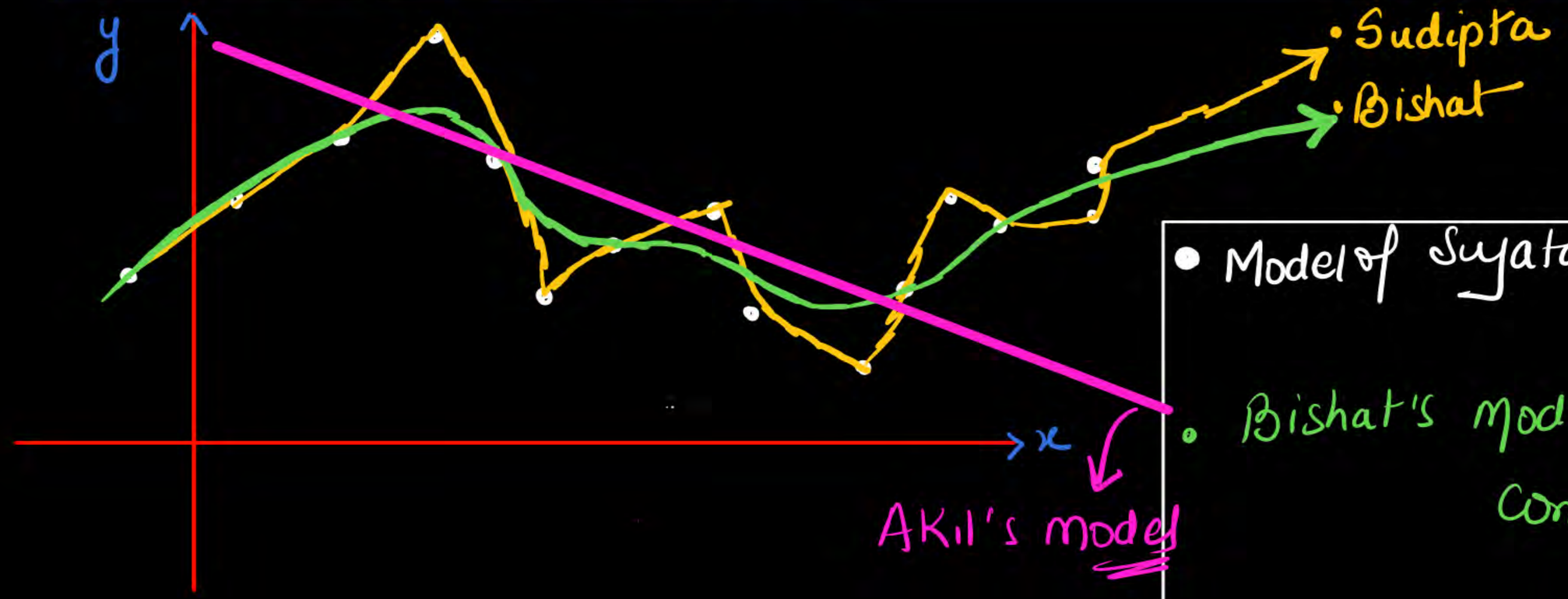
$\Rightarrow y = ax^3 + bx^2 + cx + d \approx$

$\Rightarrow y = ax^4 + bx^3 + cx^2 + dx + e \Rightarrow$

- as the complexity of the fxn inc the Plot became more and more "ajeeb", Complicated.

Basic Understanding – Predict Rain in Your city

- **Example of a model**



- Model of Sujata → more complex
- Bishal's model is less complex

- Sujata's Model → good model → becoz $y_{actual} = y_{predicted\ by\ model}$
  - → Highly Complex
  - → Rote learning / Remember data [Not under Stand Pattern]

- Bishal Model → Has some error but less Complex and Understant pattern of data easily

- Akil model → Very Simple, But is not following Pattern of data

Basic Understanding – Predict Rain in Your city

- **We can have a simple model**
  - → Problem of Cannot learn pattern of data.
  - → large error on training data

- **We can have a very complicated problem**
  - → Rote learning of data, will not learn pattern of data.

Basic Understanding – Predict Rain in Your city

We can have a simple model

⇒ So in general we donot need very Complex / very simple model.

Basic Understanding – Predict Rain in Your city

Problem in Simple Models ?

done

Basic Understanding – Predict Rain in Your city

We can have a very complicated problem

done

Basic Understanding – Predict Rain in Your city

Problem in Complex Models ??

ex Suyata Mode

Rote learning

$y_{actual} = y_{predicted}$

Bacche Wala Example

Why Rote learning is Bad ⇒

data

- we may add some error or some noise

- we can say that the data has actually parabolic pattern (R) at there are some noisy points.
- if we create complex model ⇒ Rote learning then our model will include the noisy points

- Thus this model will not be good.

Problem 1 – Predict Population of bacteria in a lab

We must create a model with following data

Time	Population
0	50
10	200

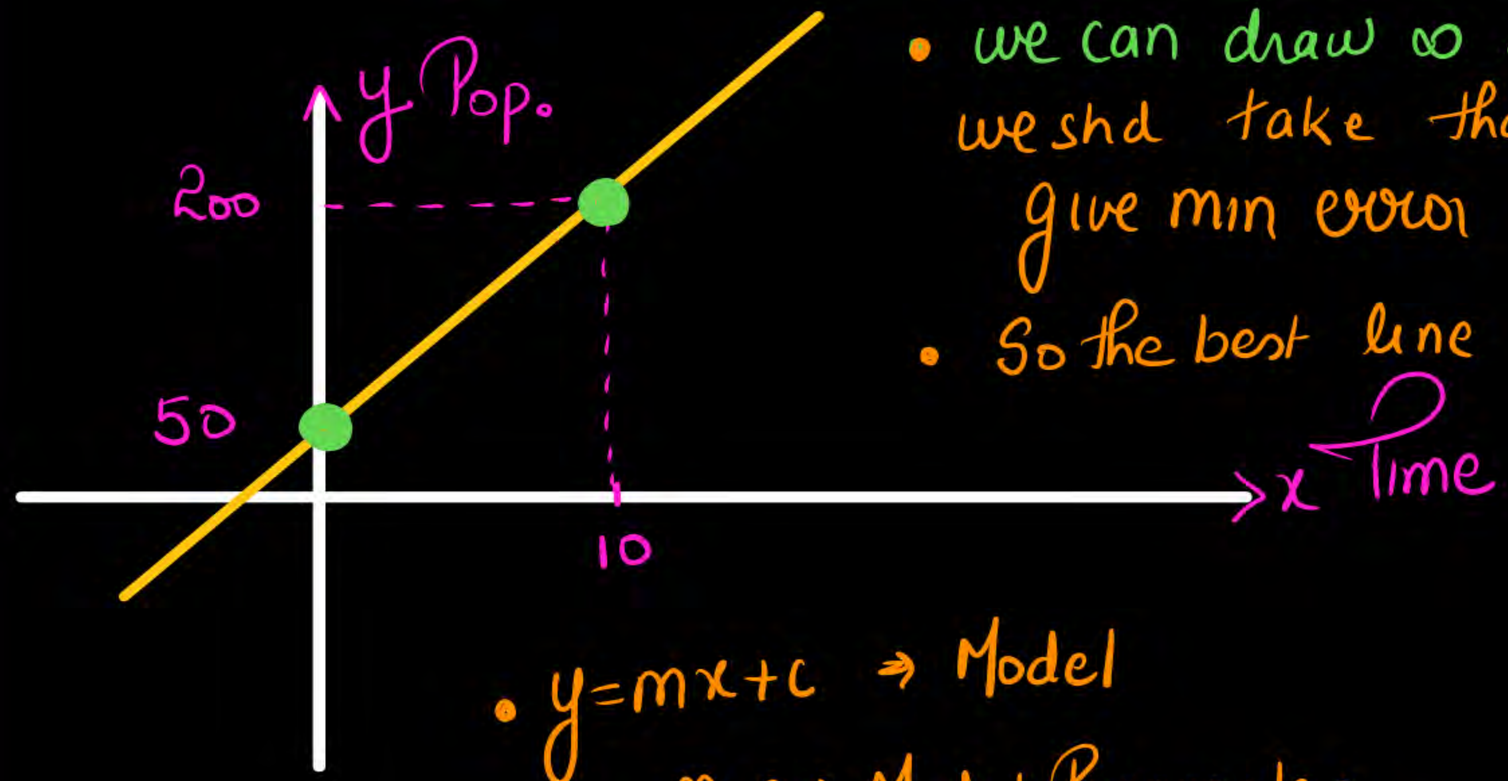
⇒ What shd be y and x

Since we have to predict Population @ t = 20

y x

Now predict the population at t = 20

Problem 1 – Predict Population of bacteria in a lab



- we can draw ∞ straight line, but we shd take that line which give min error b/w y_{actual} and y_{pred}.
- So the best line ⇒ which pass through Both points
  ⇓
- $y_{actual} = y_{pred}$

- $y = mx + c$ ⇒ Model
  m, c ⇒ Model Parameters

Problem 1 – Predict Population of bacteria in a lab

So we created a mathematical model

So 'c' = Intercept

- $y = mx + c$ @ $x = 0$, $y = 50$

  $\Rightarrow 50 = m(0) + c \Rightarrow c = 50$

- $y = mx + 50$ @ $x = 10$ $y = 200$

  $\Rightarrow 200 = 10m + 50$, $m = 15$

What are Parameters of a Model

Here 'm' and 'c' are Parameters of model.

So $\Rightarrow$ $y = 15m + 50$ → Best line

Problem 1 – Predict Population of bacteria in a lab

So we created a mathematical model

What is optimisation

Problem 2 – Predict Sale of I-phone based on Age of customer

We must create a model with following data

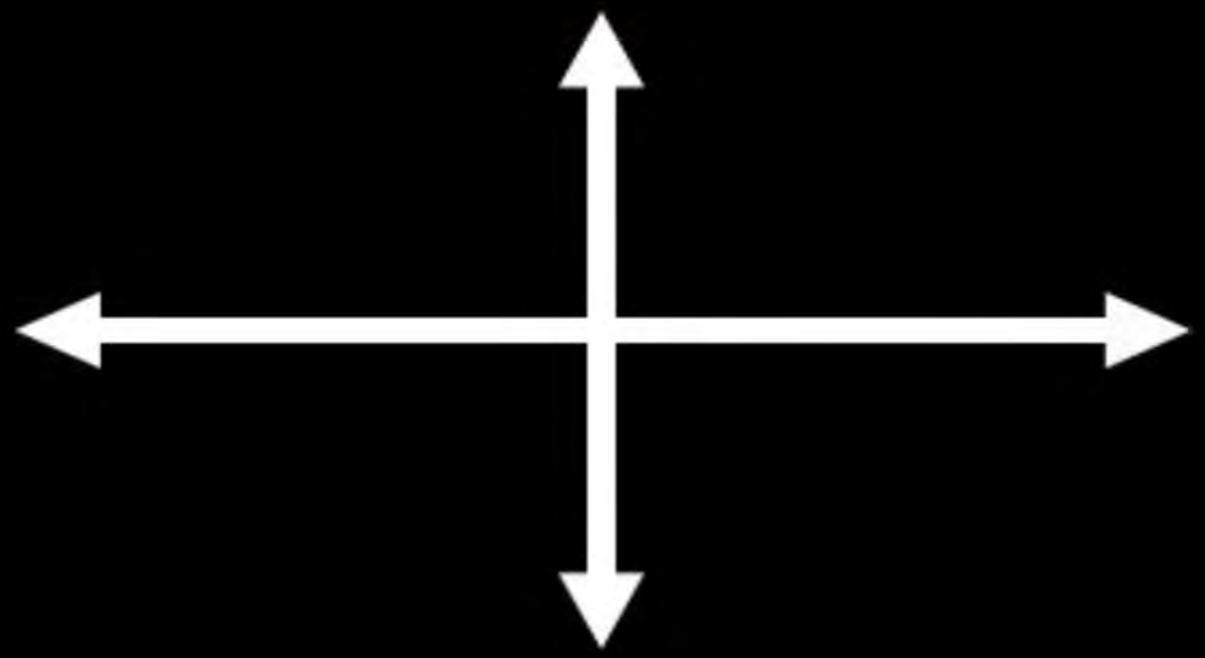
Age	Sale of I-Phone (in a month)
30	300
40	400

Now predict the Sale of I-Phone at Age = 20

Problem 2 – Predict Sale of I-phone based on Age of customer

We don't have any expert now, and data has only two Points.
So _____

What is the best model now ?



Problem 2 – Predict Sale of I-phone based on Age of customer

Now we have to find the best parameters..

Problem 3 – Predict Sale of I-phone based on Age of customer

We must create a model with following data

Age	Sale of I-Phone (in a month)
30	300
40	400
50	300

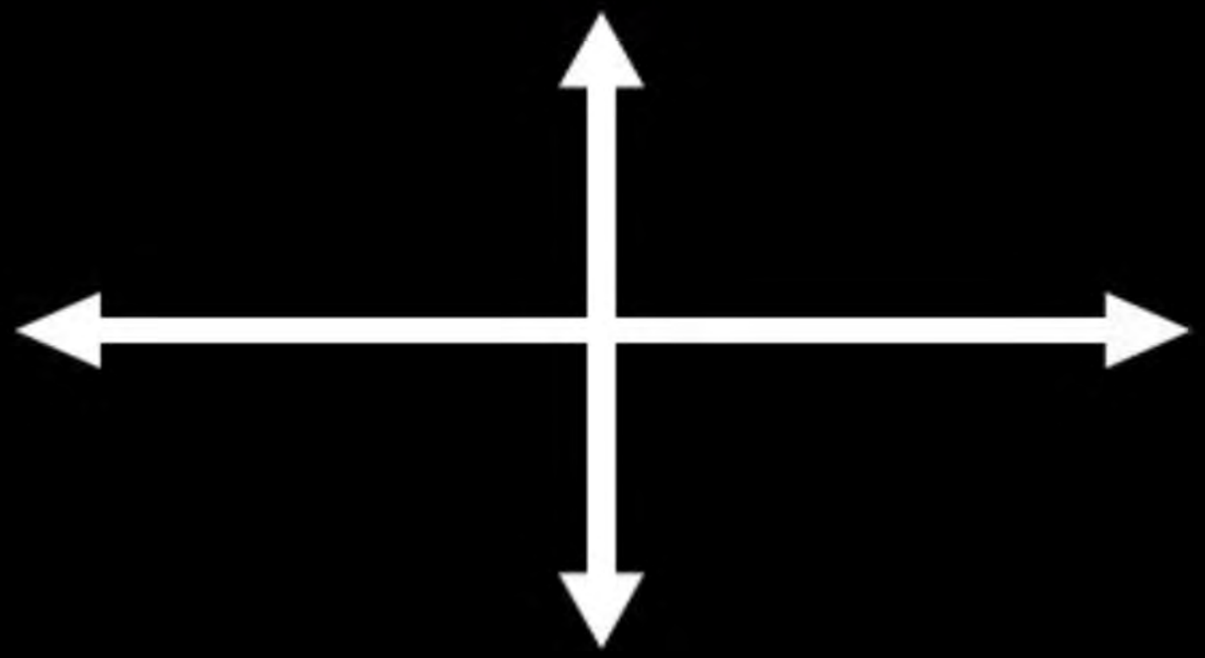
Now predict the Sale of I-Phone at Age = 20

Problem 3 – Predict Sale of I-phone based on Age of customer

We don't have any expert now, and data has only two Points. So

_____

What is the best model now ?



Problem 3 – Predict Sale of I-phone based on Age of customer

We don't have any expert now, and data has only two Points. So

_____

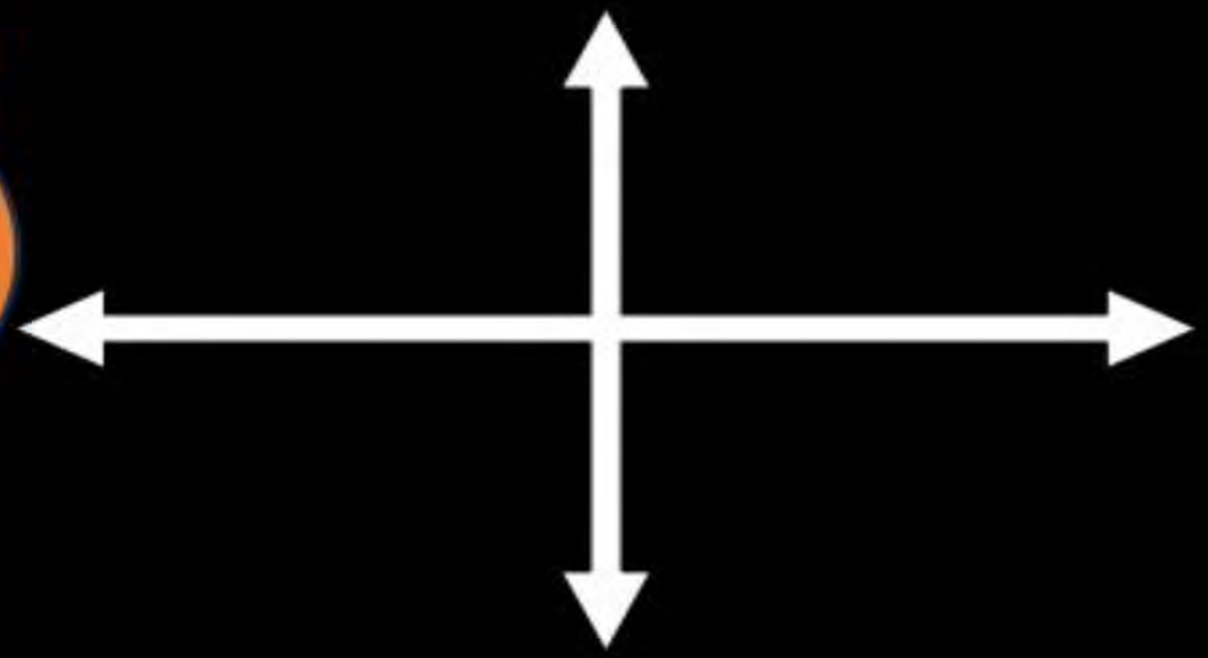
But we will try to find the linear model only.

So, we must find the model that try to

Problem 3 – Predict Sale of I-phone based on Age of customer

Creating the best model

Now we have that is called the predicted value of input.



Problem 3 – Predict Sale of I-phone based on Age of customer

Creating the best model

Loss Functions ?? (RSS-Residual Sum of Squares)

The residual sum of squares (RSS), also known as the sum of squared residuals (SSR) or the sum of squared estimate of errors (SSE), is the sum of the squares of residuals

Problem 3 – Predict Sale of I-phone based on Age of customer

Now how to find the best parameters ??

Formulae to find direct value of m and c

Problem 3 – Predict Sale of I-phone based on Age of customer

Now how to find the best parameters ??

Example

Obtain a linear regression for the data in below table assuming that y is the independent variable.

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

Problem 3 – Predict Sale of I-phone based on Age of customer

Now how to find the best parameters ??

Basics of Machine Learning

2 mins Summary

- Topic — optimization
- Topic — model
- Topic — Data
- Topic — Training
- Topic

- Complex model = Simple model
- Simple Subject.

Bye Bye

THANK - YOU