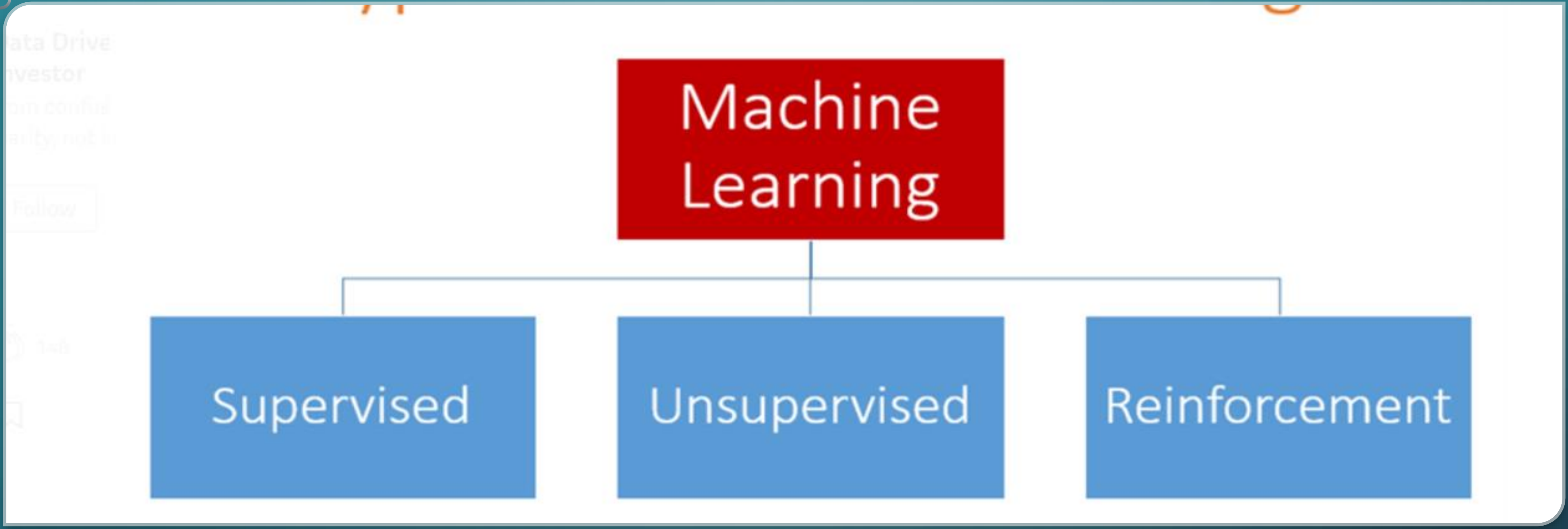
WORKSHOP ON LINEAR REGRESSION

DAY-1

WHAT IS LINEAR REGRESSION?

- Statistical approach:
  - Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.
- Machine learning approach:
  - Linear regression is used to predict a quantitative response Y from the predictor variable X with an assumption that there’s a linear relationship between X and Y.

- They are SAME!



Machine Learning

Supervised

Unsupervised

Reinforcement

LINEAR REGRESSION COMES UNDER SUPERVISED LEARNING

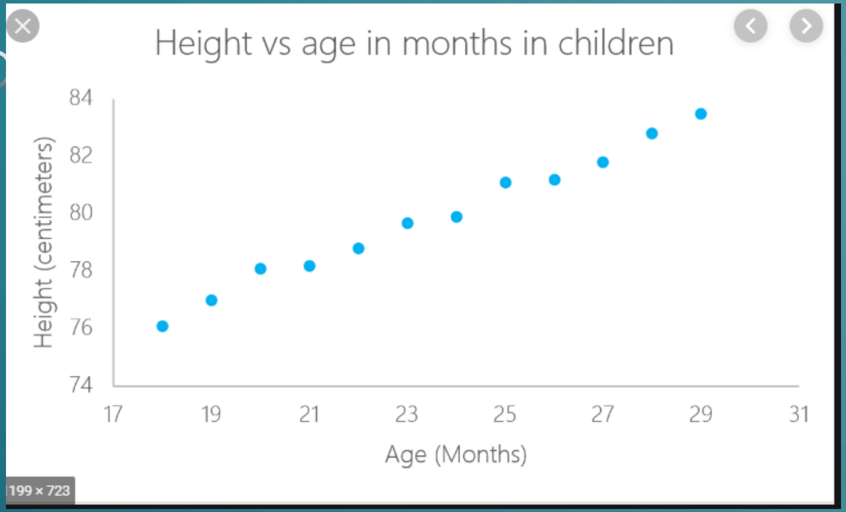
why?

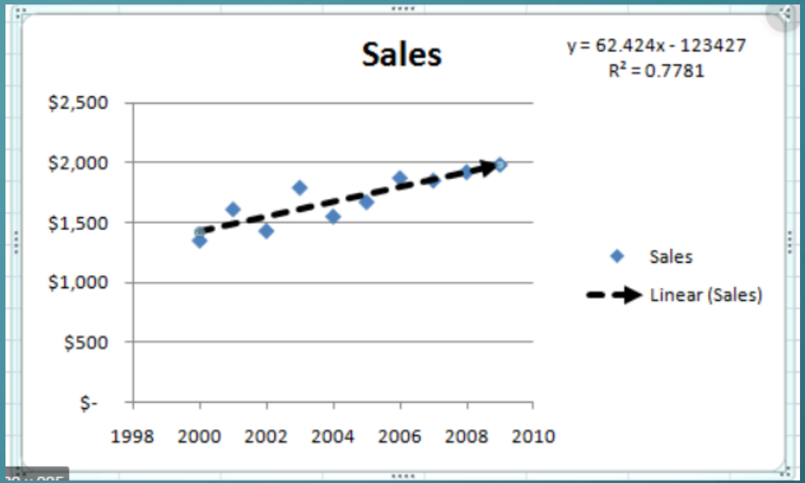
Because the training data are labelled.

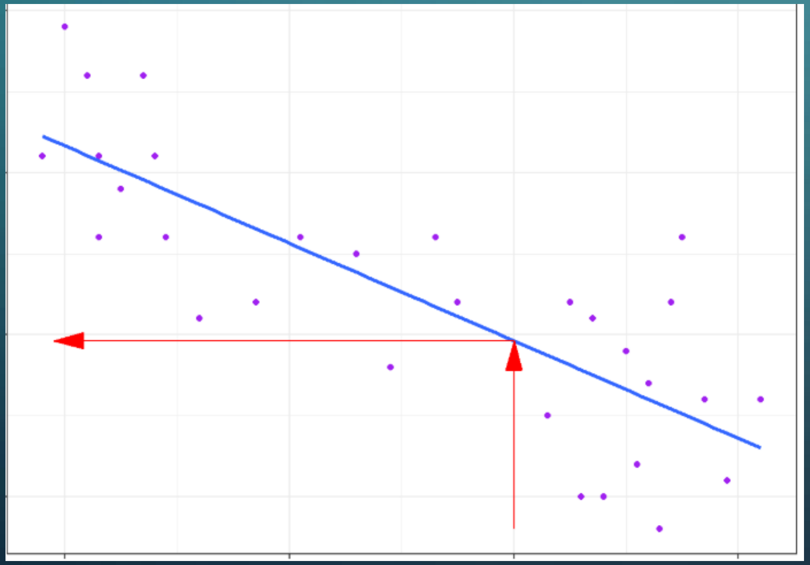
OVERVIEW

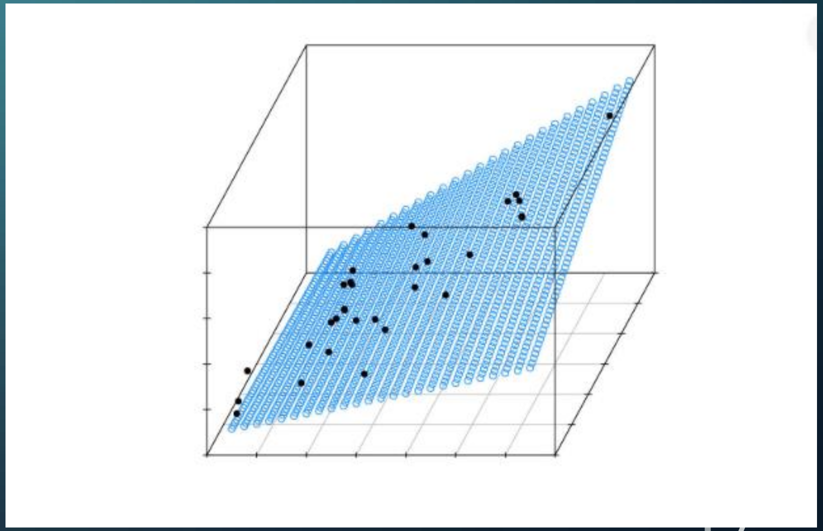
What is Linear Regression?	• A supervised algorithm that learns from a set of training samples
What is the objective?	• Find the best fitted LINE based on training data
As a data scientist what is your outcome of interest?	• Estimation, Prediction and validation of your model
How can I solve it?	• By fitting the "BEST" line between the output variable (response) input variable (explanatory).
How do I know my answer is good?	• We need to test for proximity of training data and fitted data.

Few Examples of linear regression:









IMPORTANT ASPECTS OF REGRESSION

Assumption

- What assumptions we need to make?

1. $y = \underbrace{ax+b}_{\text{Fixed part}} + \underset{\text{Random part}}{\underset{\downarrow}{\epsilon}}$
2. $\in \sim$ iid $N(0,\sigma^2)$
3. a and b are fixed and unknown
4. Other assumptions are beyond our scope

Estimation

- We need to find a and b based on some criteria
- What criteria?

--The estimated value should be very close to the predicted value **based on training data**

Validation

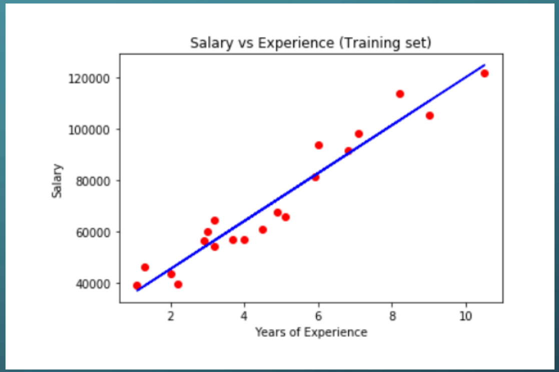
- How do I know my model is good?
- -- check how close the predicted values (**that you modelled**) to the actual value(**obtained from training data**). (Closer they are, the better your model is)
- (Measurement tool:
- p value, t statistic etc..)

CASE STUDY ON AGE AND SALARY

PROBLEM

Consider predicting the salary of an employee based on his/her age. We can easily identify that there seems to be a correlation between employee's age and salary (more the age more is the salary).

SOLUTION



Salary=age*3000+500

P value= .005

IMPORTANT ASPECTS OF REGRESSION

Assumption

- What assumptions we need to make?

1. $y = \underbrace{ax+b}_{\text{Fixed part}} \quad + \underset{\substack{\downarrow \\ \text{Random part}}}{\epsilon}$

2. Є ~ iid $N(0,\sigma^2)$

3. a and b are fixed and unknown

4. Other assumptions are beyond our scope

Estimation

- a= 3000
- b= 500

Validation

- p value=.005 is smaller than a predefined threshold (.05).
- -So the model seems to be good, based on the training data set and our initial assumptions; age seems to explain a significant amount of variation in salary

FOOD FOR THOUGHT

- If we keep increasing the age, will salary keep increasing?

BLAKE HAMENT AND KEMIL HERATH WILL ELABORATE THE CONCEPTS WITH A PRACTICAL EXAMPLE IN THE NEXT SESSION.

TO DO ITEMS FOR MEMBERS:

1. PLEASE INSTALL PYTHON AND R IN YOUR LAPTOP
2. BRING YOUR LAPTOP IN THE NEXT SESSION
3. FAMILIAR YOURSELF WITH DIFFERENT TERMS OF MACHINE LEARNING:
   a) Data sets (Training data, Validation data, Test data)
   b) Supervised learning, Unsupervised learning
   c) Overfitting, Underfitting