innovate achieve lead

BITS Pilani

Machine Learning (IS ZC464)

Session 2: Regression

Regression

- The environment of a system is defined by a set of variables – dependent and independent
- Regression analysis deals with statistical processes that help in building a relationship among the variables.
- Some applications include
  - ❑ Predicting the price of a product in the future
  - ❑ To predict the score of a player in a team in the coming matches
  - ❑ To predict the number of drop outs in village school

Regression

- Understand the variables (say x_1, x_2, x_3, …x_n) that are independent and have association with the output value (say y).
- The values of the variables are discrete and numerically represented.
- The output y can be a real number or an integer
- Understand the **best** possible equation for ‘f’ that fits **$y = f(x_1, x_2, x_3, \ldots x_n)$**

Consider a single variable data for regression

- Following values are observed for supervised regression

X	Y
1	1
5	5
2	2
4	4
3	3

- The dependent variable is y.
- The variable x is independent.
- Data is called as uni-variate.

Prediction

X	Y
1	1
5	5
2	2
4	4
3	3

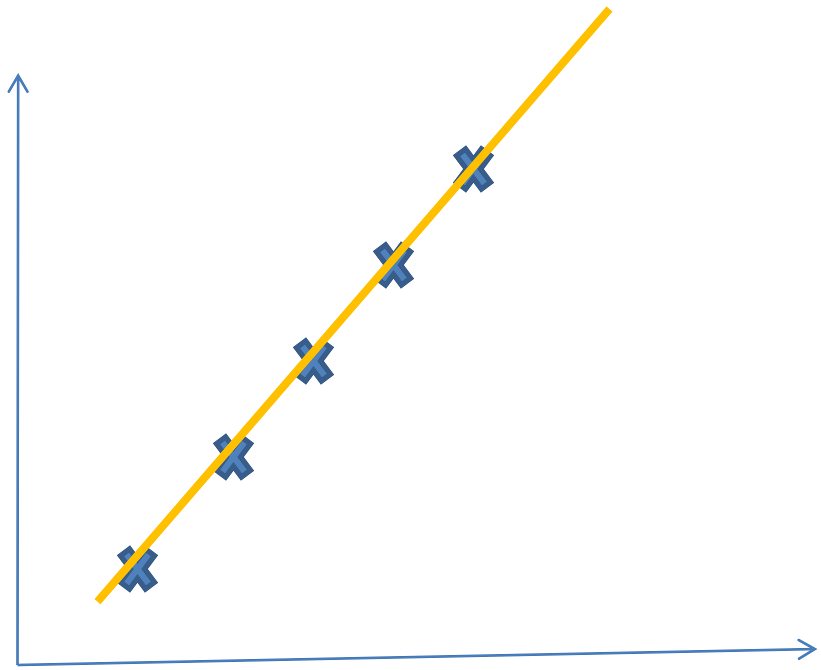
- Recall Learning: A machine with learning capability can predict about the new situation (seen or unseen) using its past experience.
- Prediction:

  Given values of x and y

  Predict value of y for x = 71
- Prediction is based on learning of the relationship between x and y
- Training data is the collection of (x,y) pairs
- Testing data is simply value of x for which value of y is required to be predicted.

Learning of a function from given sample data

Straight Line



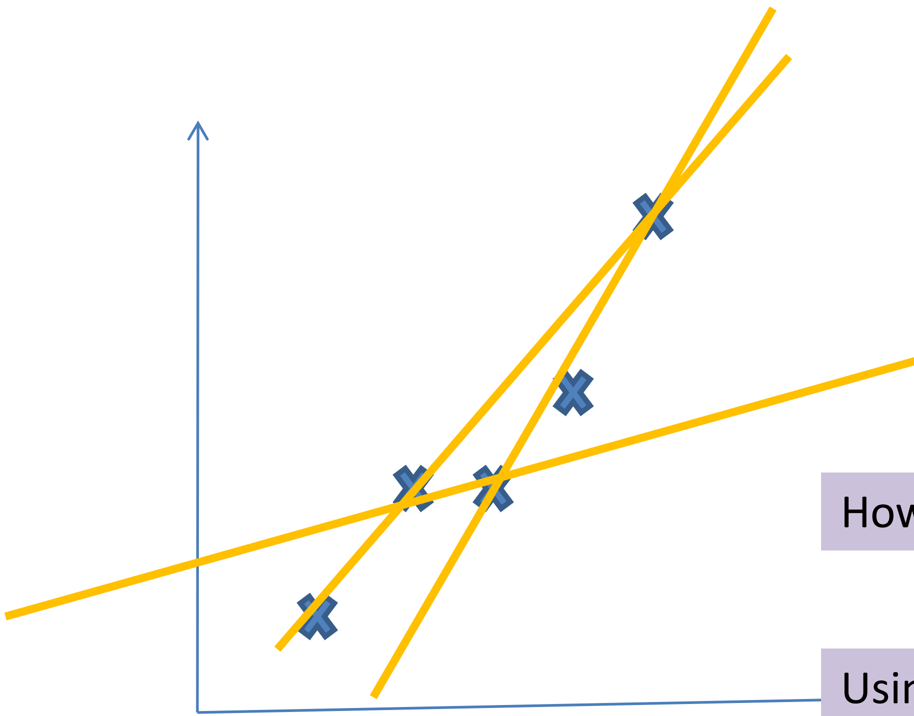
What did the system learn?

X	Y
1	1
5	5
2	2
4	3
3	2

- Y = f(x)
- Y = x
- What is its generalization ability?
- Most accurate or we can say 100%
- What if the data to train the system changes slightly? The machine can be still made to learn.

Learning of a function from given sample data-straight line learning

Straight Line



Which line fits the best?

Line is represented by parameters of slope and intercept

Machine must learn on its own-which is the best fit

How?

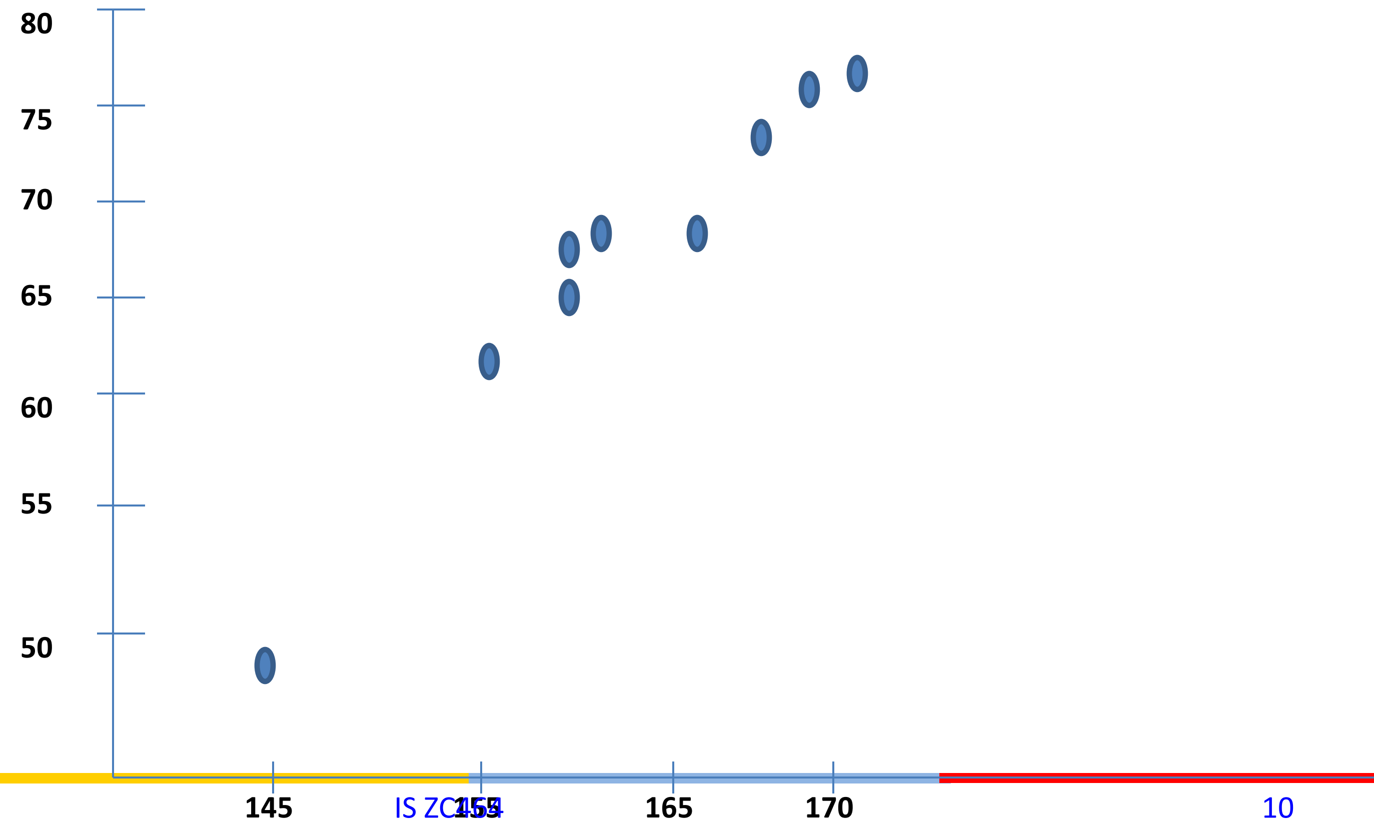
Using the data – known as training data i.e. (x,y) pair

Understanding ERROR

- Consider an example of using height and weight

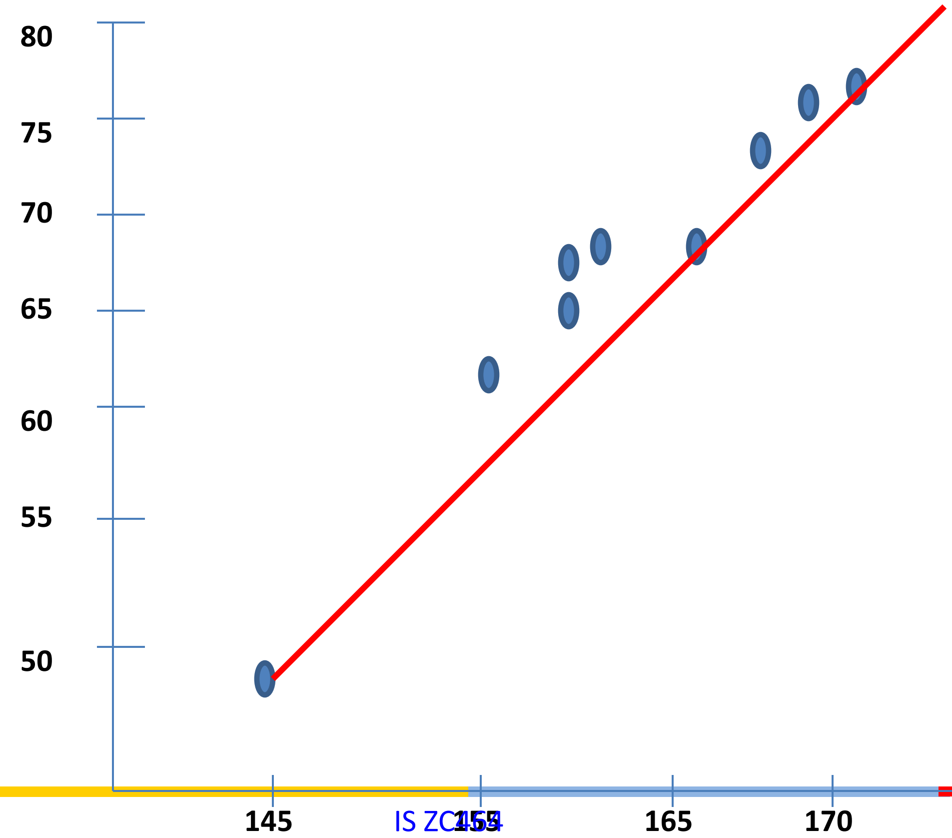
Height (in cm)	Weight (in Kg)
145	48
165	68
155	62
160	65
170	75
163	67
171	76
167	72
159	65

Understanding ERROR



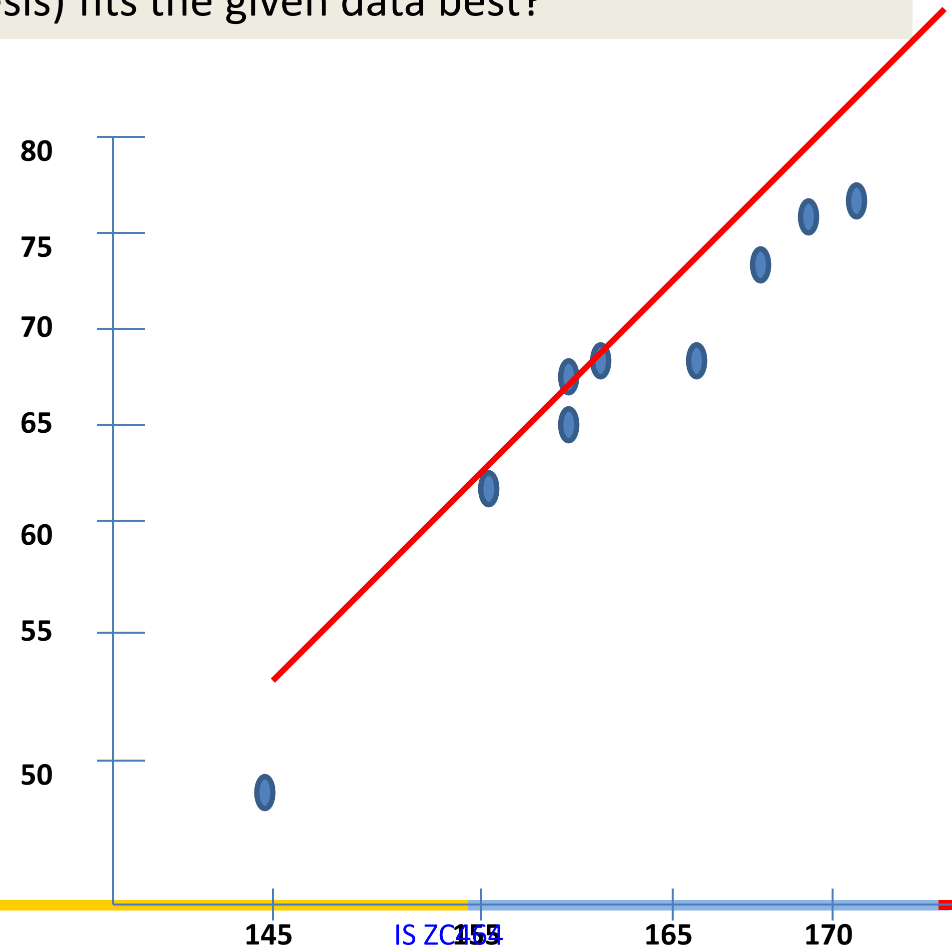
Understanding ERROR

Which line (hypothesis) fits the given data best?



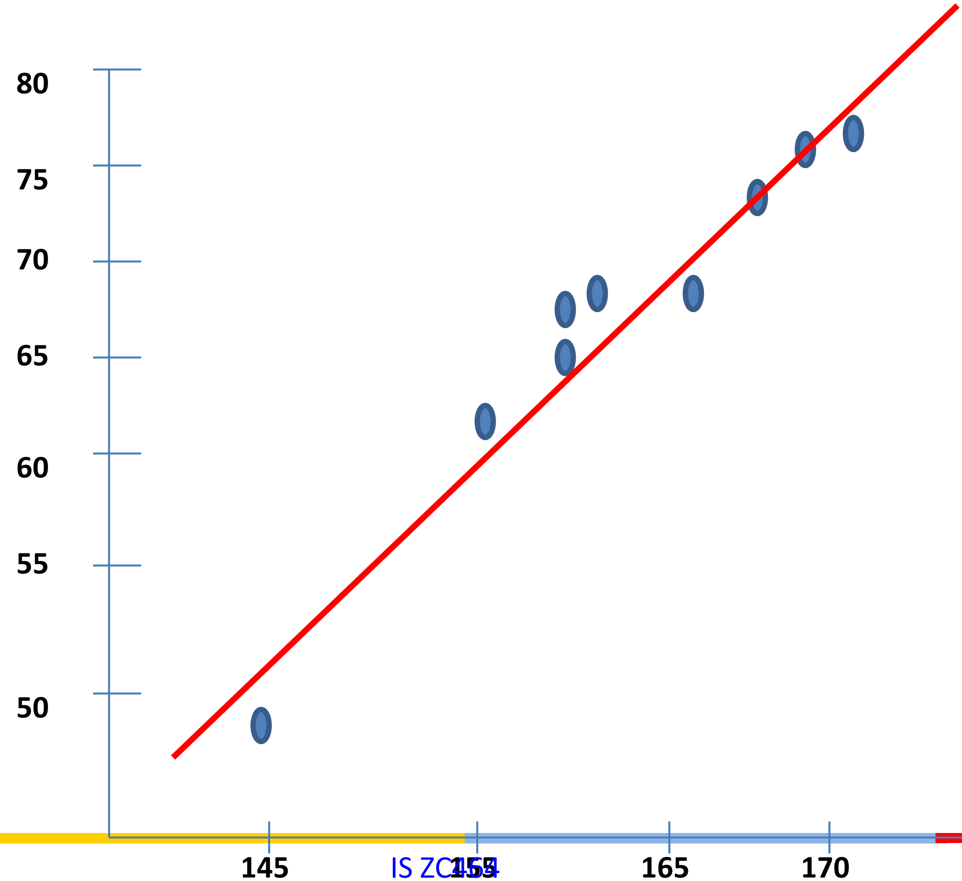
Understanding ERROR

Which line(hypothesis) fits the given data best?



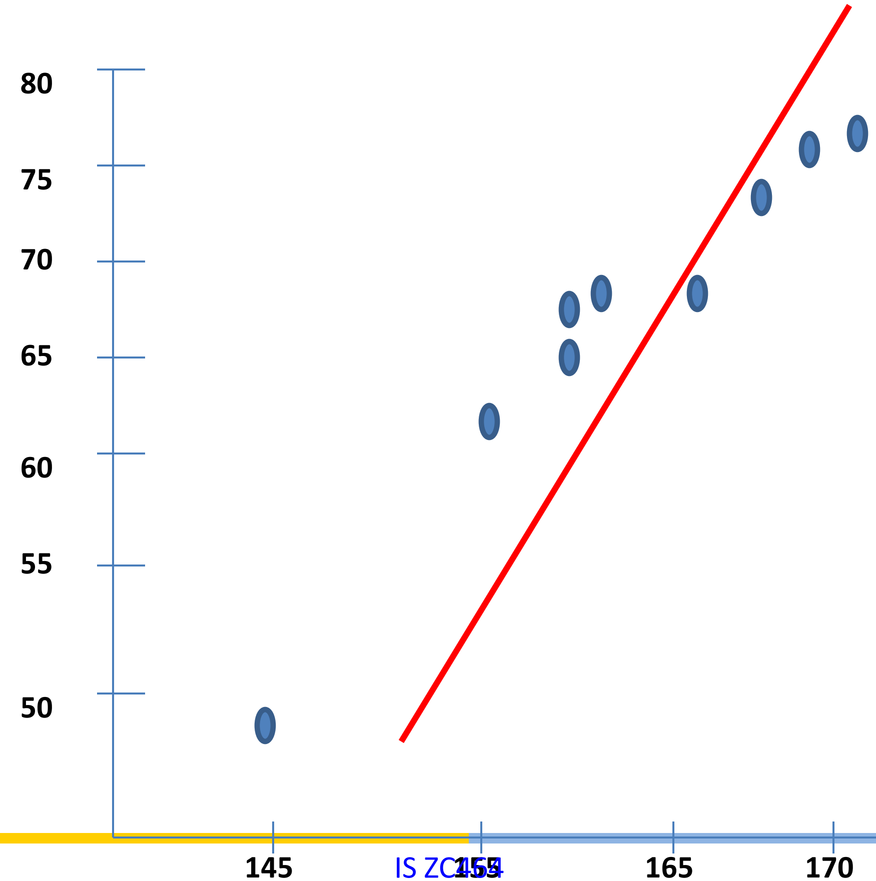
Understanding ERROR

Which line(hypothesis) fits the given data best?



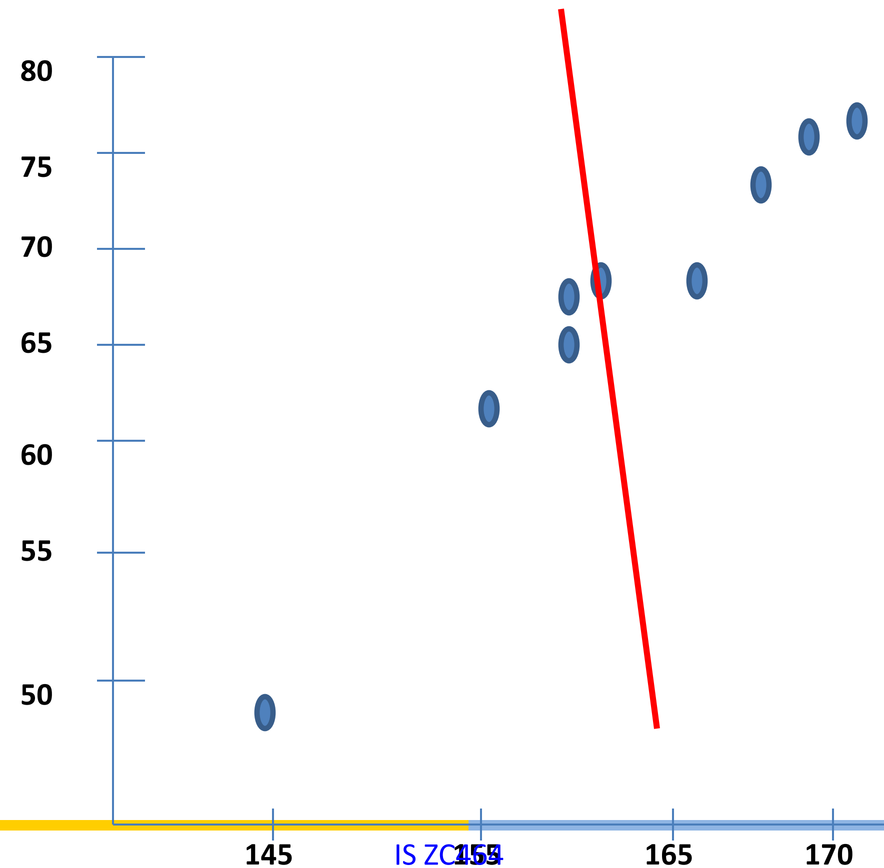
Understanding ERROR

Which line(hypothesis) fits the given data best?



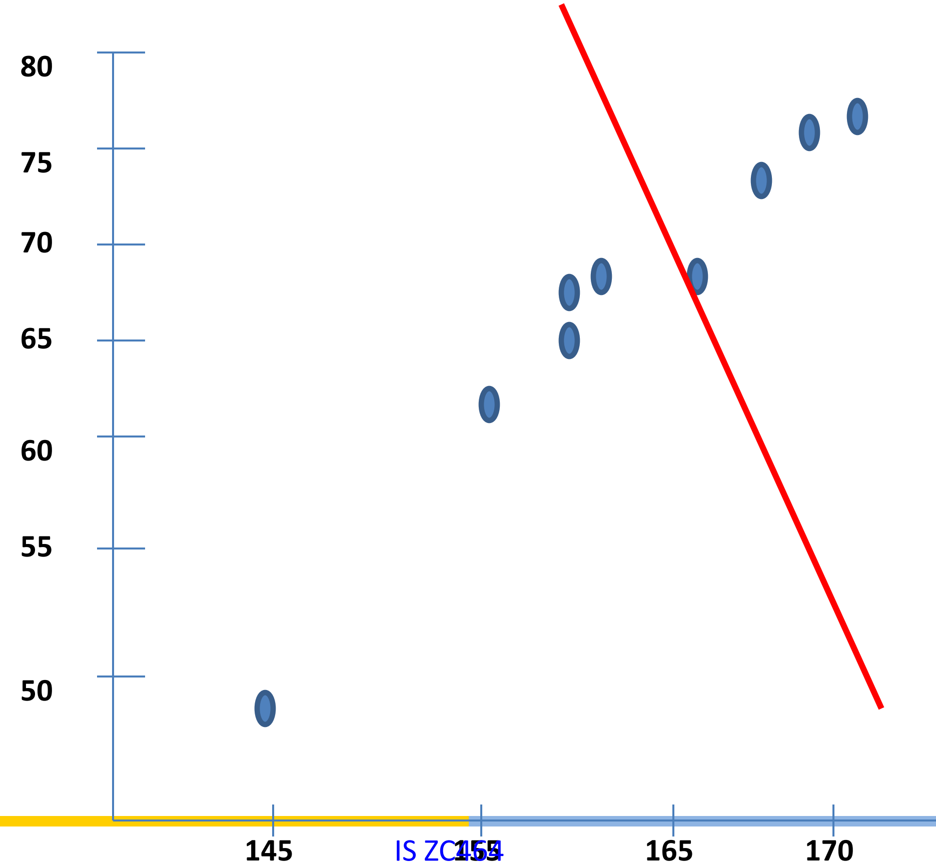
Understanding ERROR

Which line(hypothesis) fits the given data best?



Understanding ERROR

Which line(hypothesis) fits the given data best?

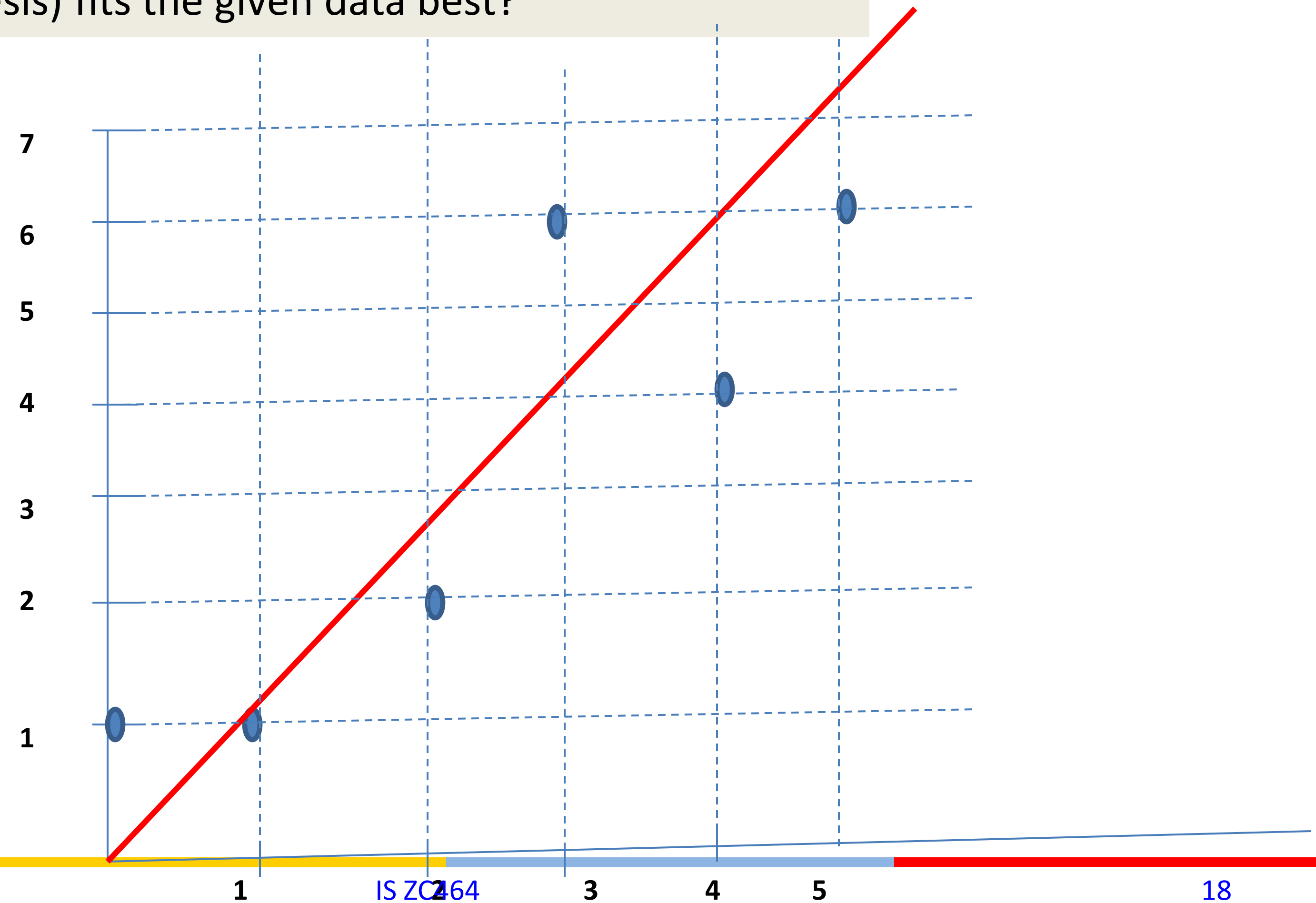


In 2D space the line parameters are two

- Slope and intercept
- Can be called as w_1 and w_2
- In order to find a line that best fits the given data, we must find w1 and w2 in such a way that the sum of the squared error is minimum

A simple example to understand ERROR

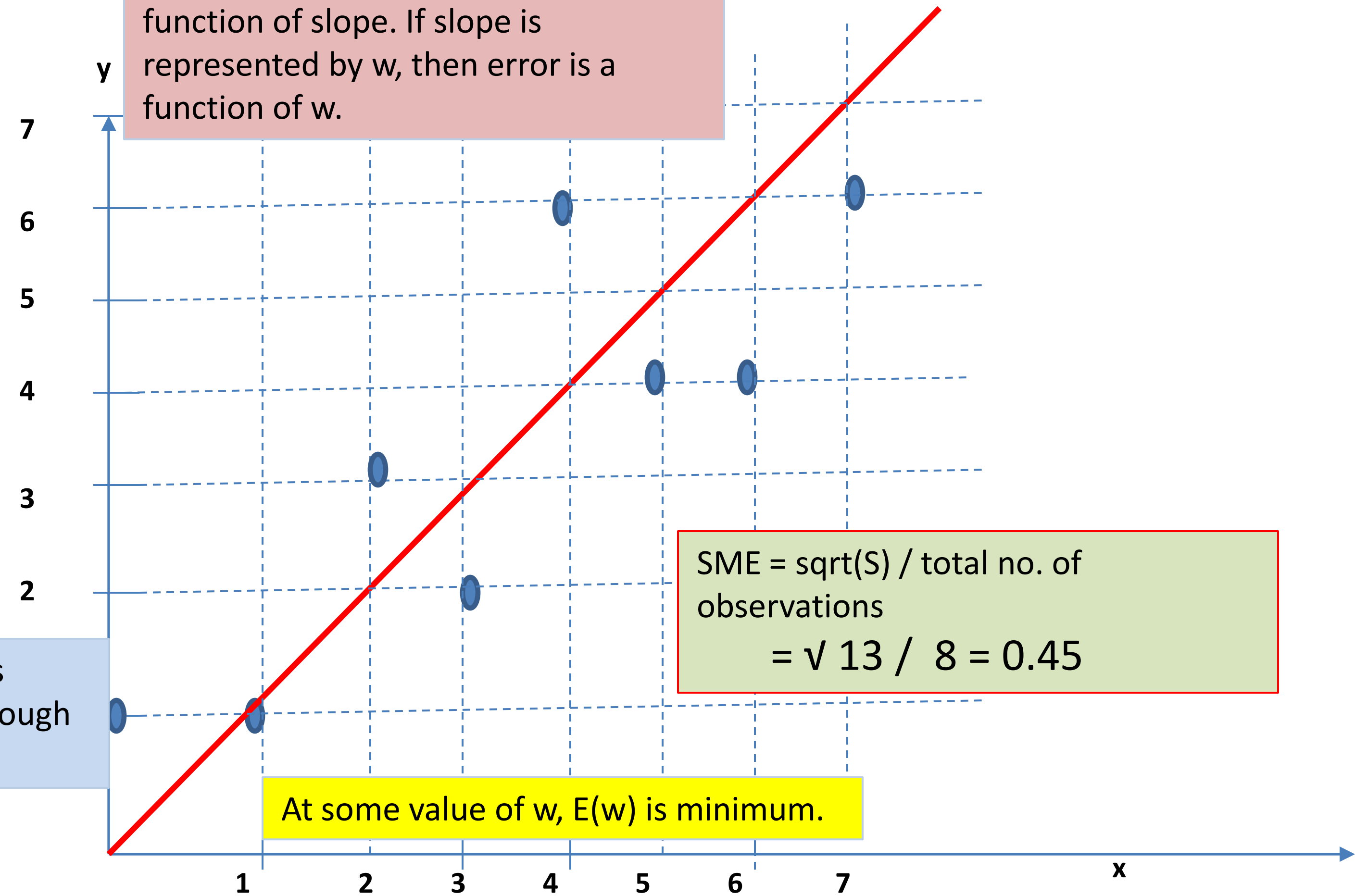
Which line(hypothesis) fits the given data best?



Compute the Squared Mean Error (line is y=x)

Sum of squares
(S) = 1*1
+ 0
+ (-1)*(-1)
+1*1
+(-2)*(-2)
+1*1
+2*2
+1*1
=13

Therefore we can say that error is the function of slope. If slope is represented by w, then error is a function of w.



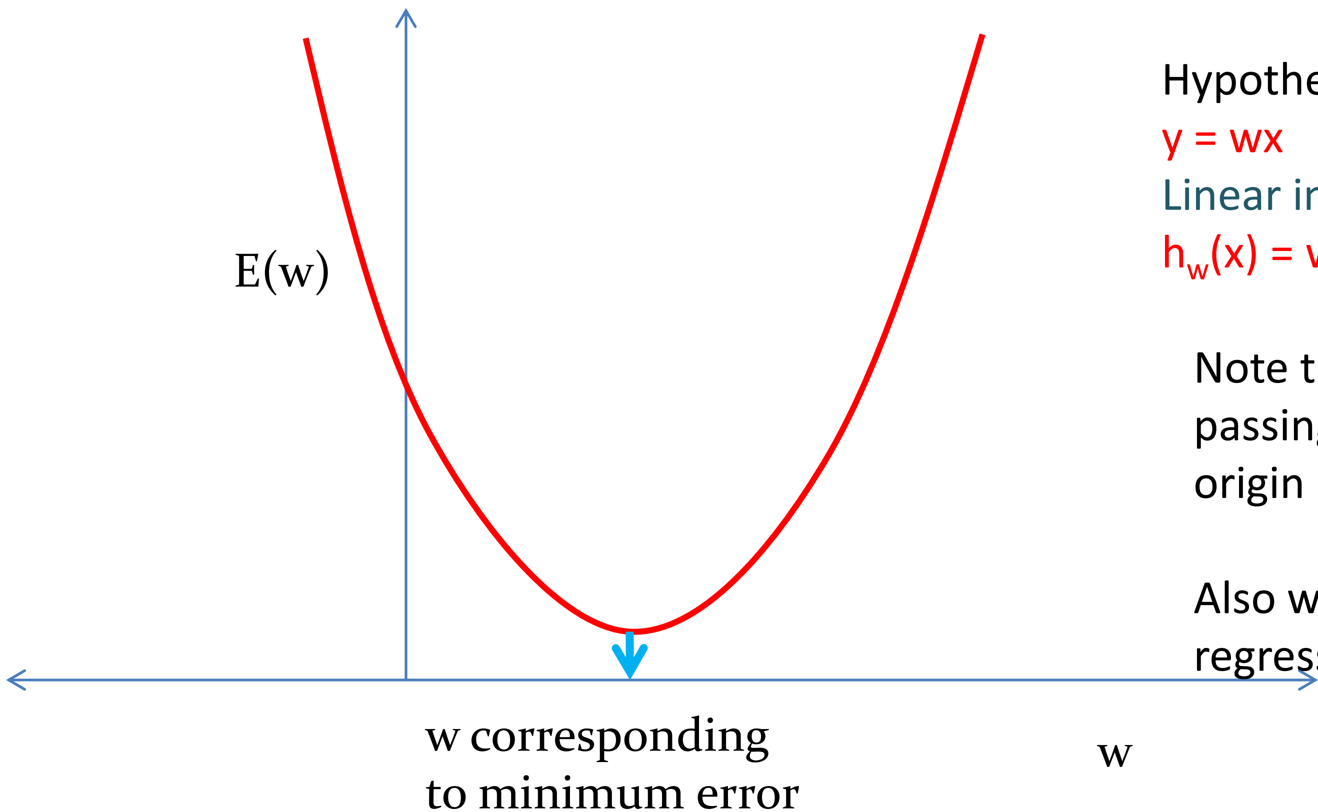
Error will be different if the line's slope is different (line passes through origin)

SME = sqrt(S) / total no. of observations

= √ 13 / 8 = 0.45

At some value of w, E(w) is minimum.

Plotting error when y=f(x)



Hypothesis function

$y = wx$

Linear in one variable

$h_w(x) = wx$

Note that the line is passing through the origin as c = 0.

Also w is the slope of the regression line.

Multi-variate Regression

- Includes many variables as independent variables X = <$\mathbf{x_1, x_2, x_3, \ldots x_n}$ >
- There is one dependent variable (say y).
- The regression model builds a relationship between y and X such that $\mathbf{y = f(x_1, x_2, x_3, \ldots x_n)}$
- The regression line is a n-dimensional line.
- The equation of the regression line is

$$y = w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n$$

Understanding of the error surface for linear regression in one variable

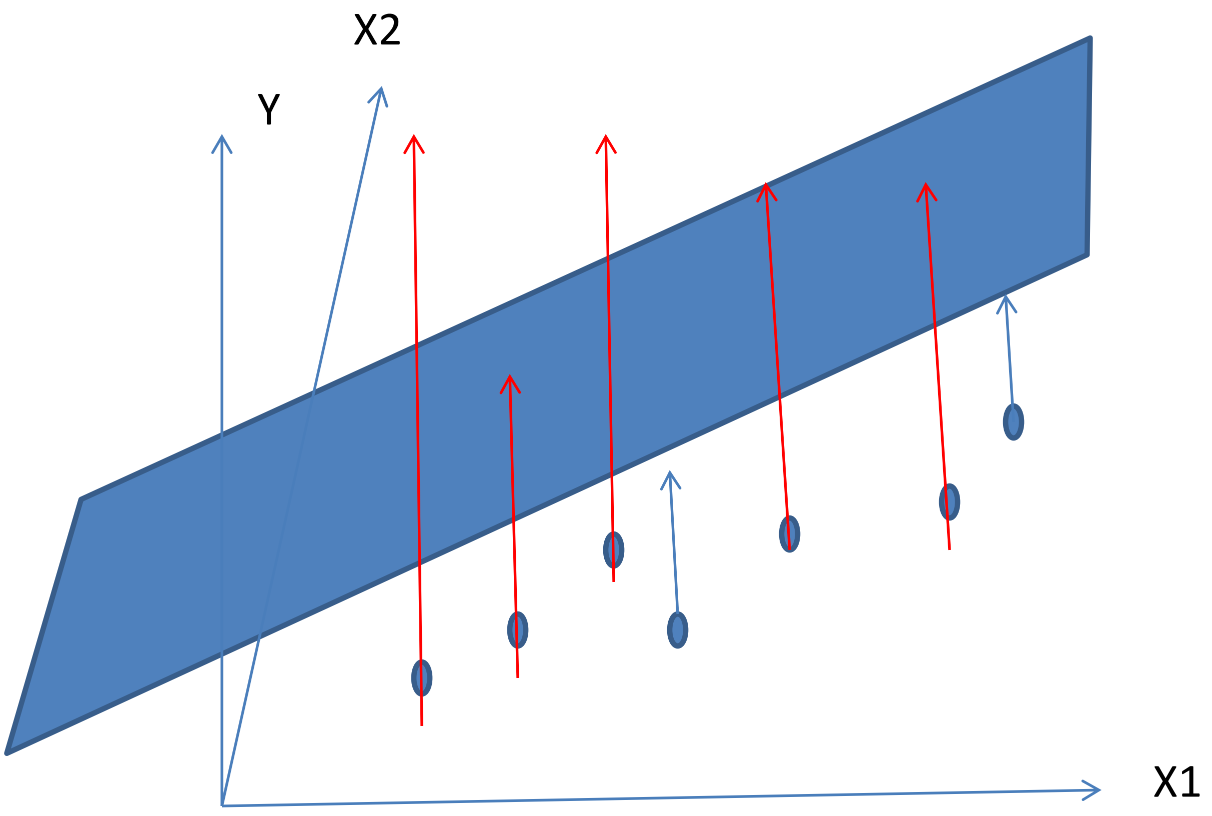
- Consider m observations $<x^1,y^1>$, $<x^2,y^2>$, ...$<x^m,y^m>$.
- An hypothesis $h_w(x)$ that approximates the function that fits best to the given values of y
- There is likely to be some error corresponding to each observation (say i).
- The magnitude of such error is $y^i - h_w(x^i)$
- Objective is to find such w that minimizes the sum of squares of errors

$$E_{min}(w) = \text{Minimize}_w \sum_i (y^i - h_w(x^i))^2$$

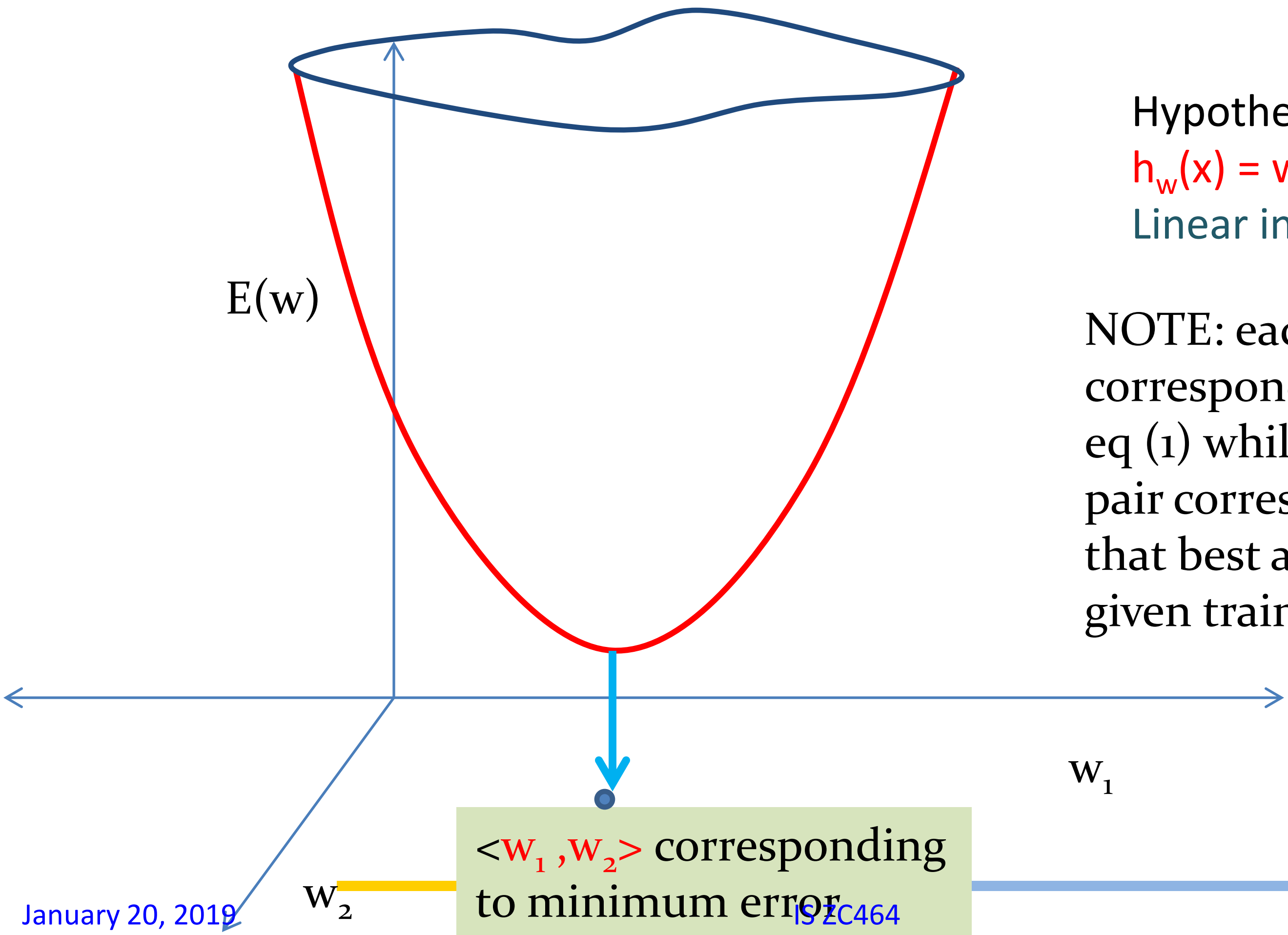
Linear Regression in two variables

Number of hours of work (X1)	Number of items produced (X2)	Average Wages paid to each employee (Y)
89	4	300
66	1	220
78	3	290
111	6	340
44	1	230
77	3	290
80	3	280

Interpreting regression as a plane in two dimensional space



Plotting error when $y=f(x_1,x_2)$



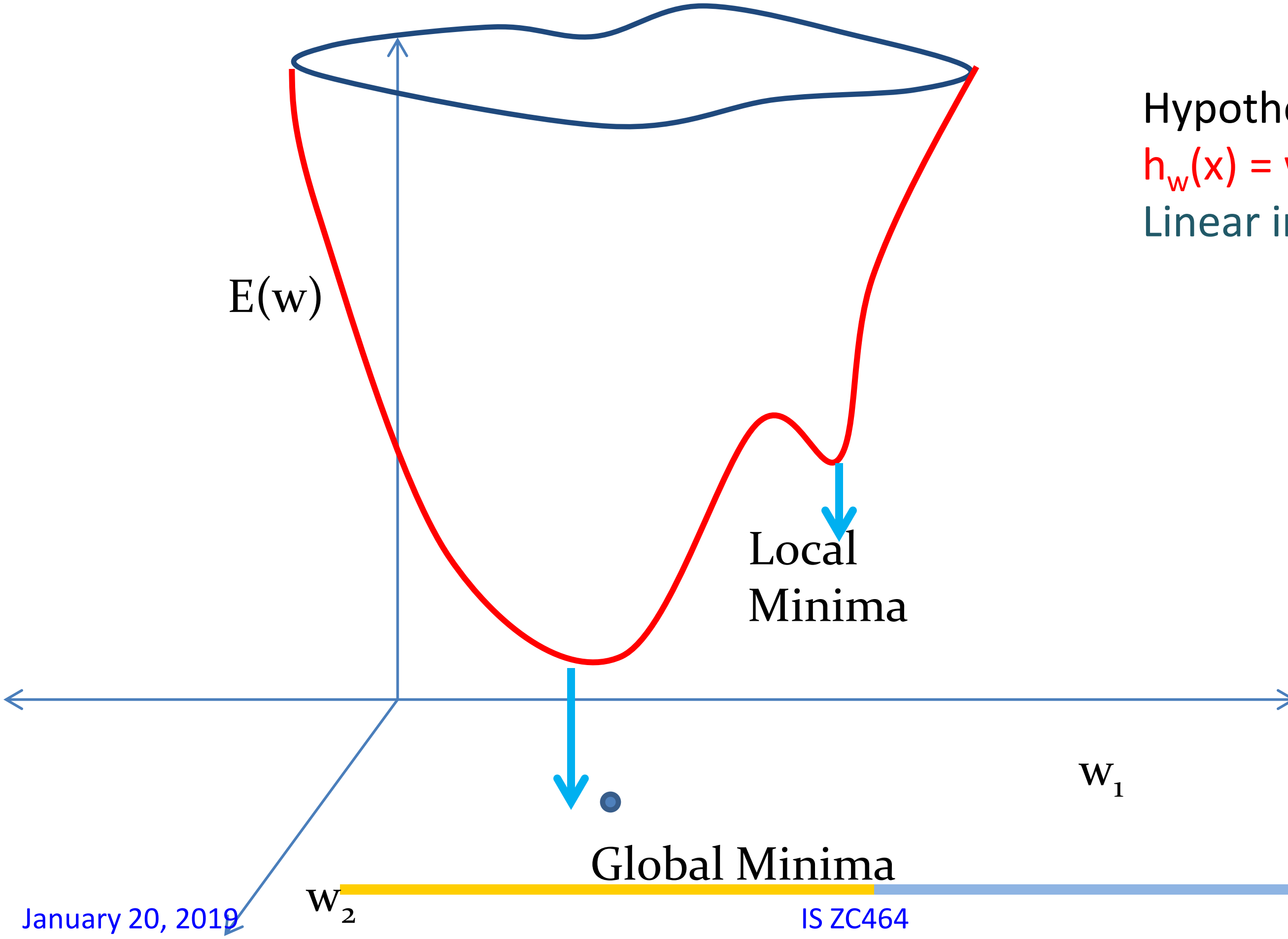
Hypothesis function

$h_w(x) = w_1x_1 + w_2x_2$ …….(1)

Linear in two variables

NOTE: each pair < w_1 ,w_2> corresponds to a line given by eq (1) while only one such pair corresponds to the line that best approximates the given training data

Plotting error when $y=f(x_1,x_2)$

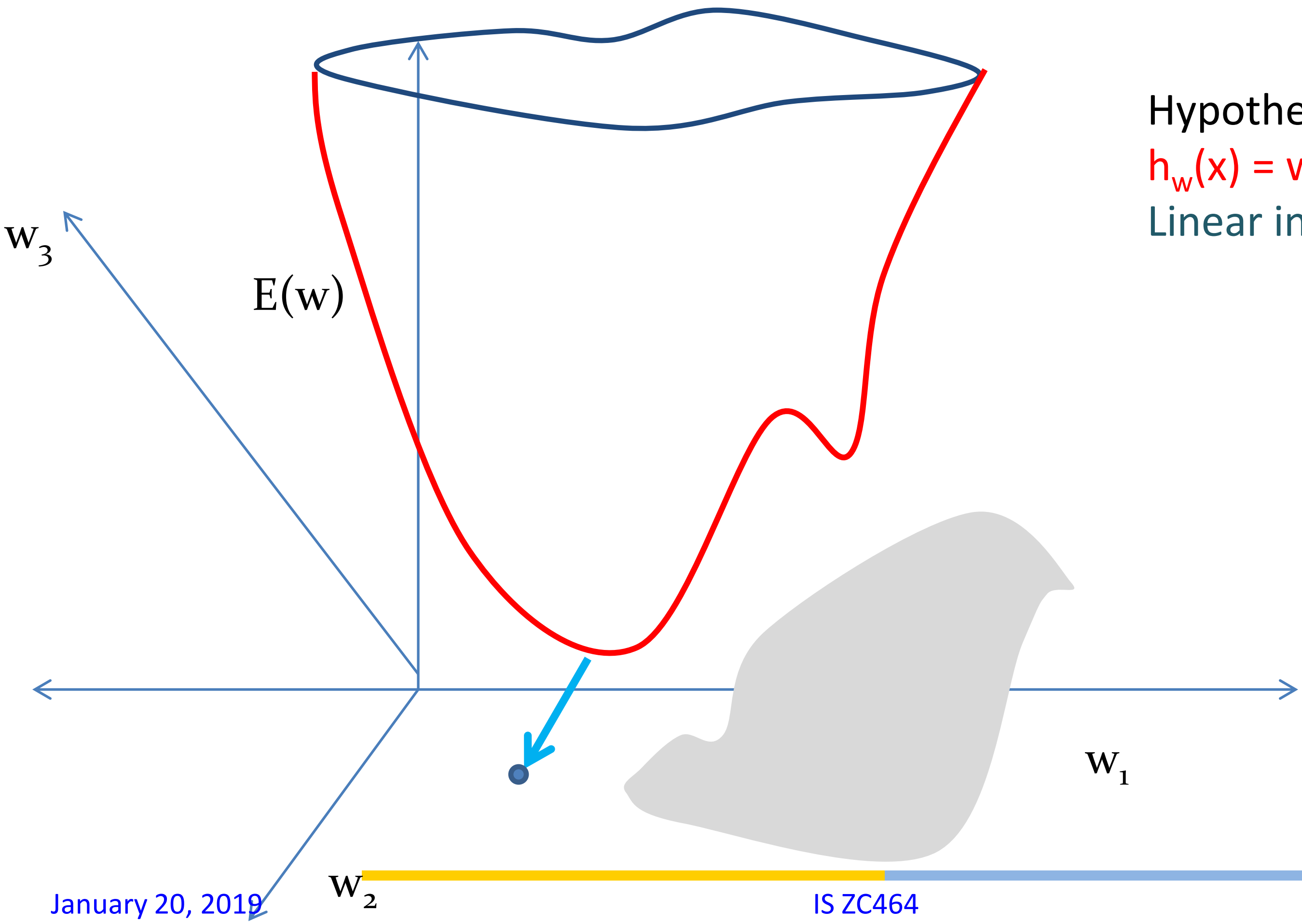


Hypothesis function

$h_w(x) = w_1x_1 + w_2x_2$ ....(1)

Linear in two variables

Difficult to visualize when $y=f(x_1,x_2, x_3)$



Hypothesis function

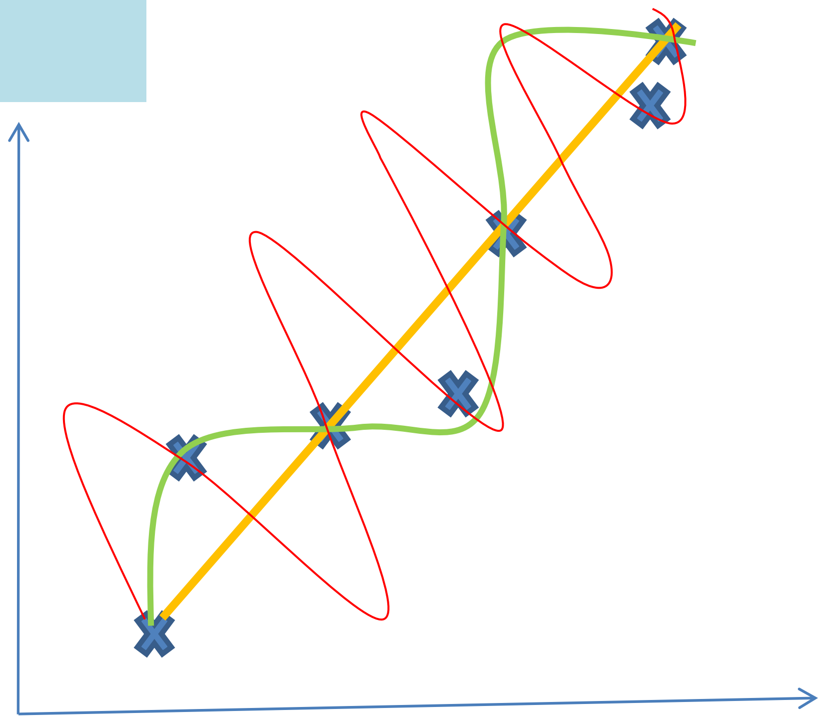
$h_w(x) = w_1x_1 + w_2x_2 + w_3x_3$

Linear in three variables

Learning of a function from given sample data-polynomial curve Learning

T: prediction of y-value for given x-value
P: least error
E: experience by training

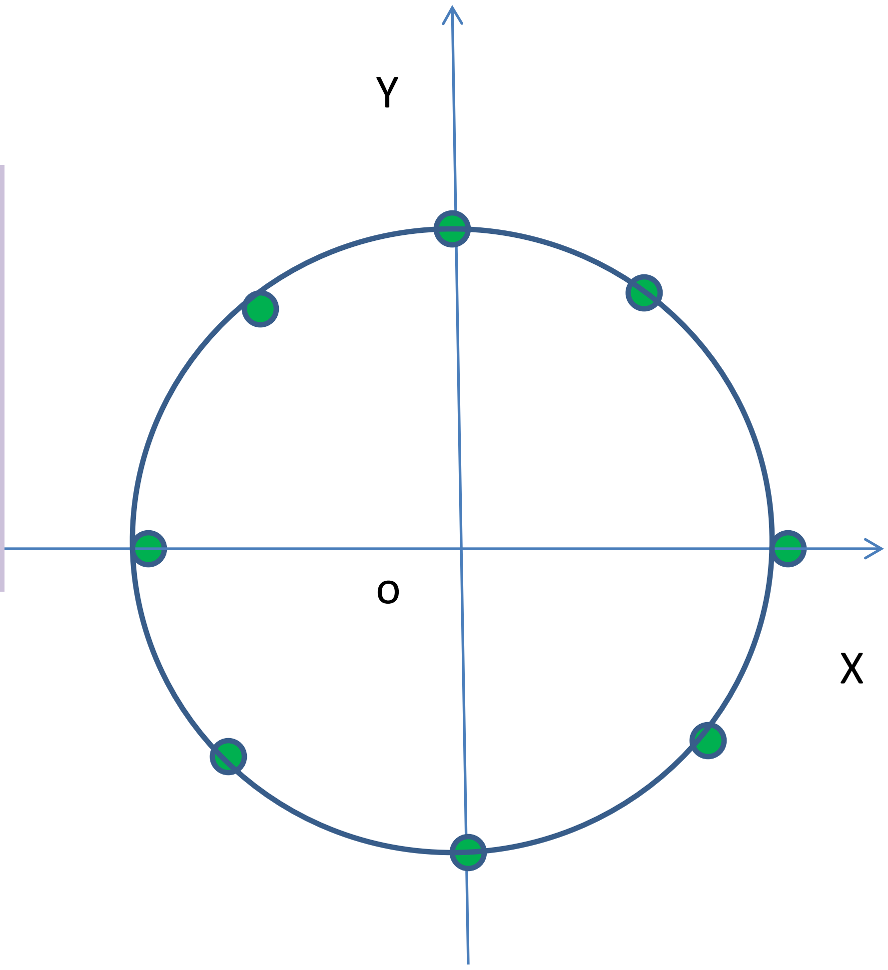
1. Straight Line
2. Sinusoidal Curve
3. Other higher order polynomial



Generalization in Function Approximation

X	Y
1	0
0	1
0	-1
0.6	0.8
0.6	-0.8
-0.6	0.8
-0.6	-0.8
-1	0

Generalization
If the NN answers --
What is f(-0.25)?
Or
f(0.001)
correctly



$Y = \pm \sqrt{(1-X^2)}$