python

Class: Machine Learning

★ Topic ★

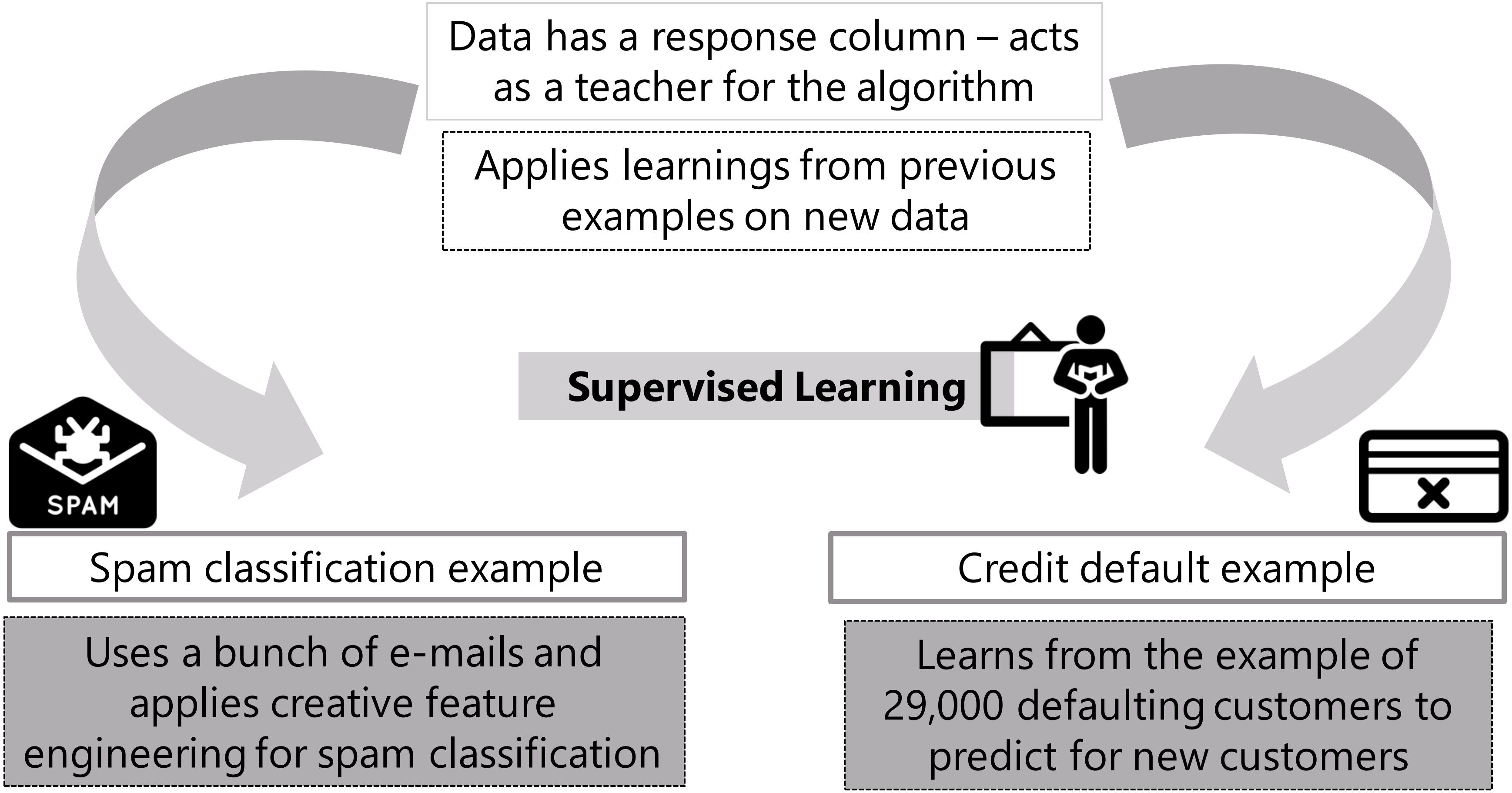
Types of Tasks, Machine Learning Algorithms and Linear Regression

Machine Learning

What Kinds of Problems Does it Solve?

Machine Learning algorithm tasks can be broadly categorized

Types of Tasks



Data has a response column – acts as a teacher for the algorithm

Applies learnings from previous examples on new data

Supervised Learning

Spam classification example

Uses a bunch of e-mails and applies creative feature engineering for spam classification

Credit default example

Learns from the example of 29,000 defaulting customers to predict for new customers

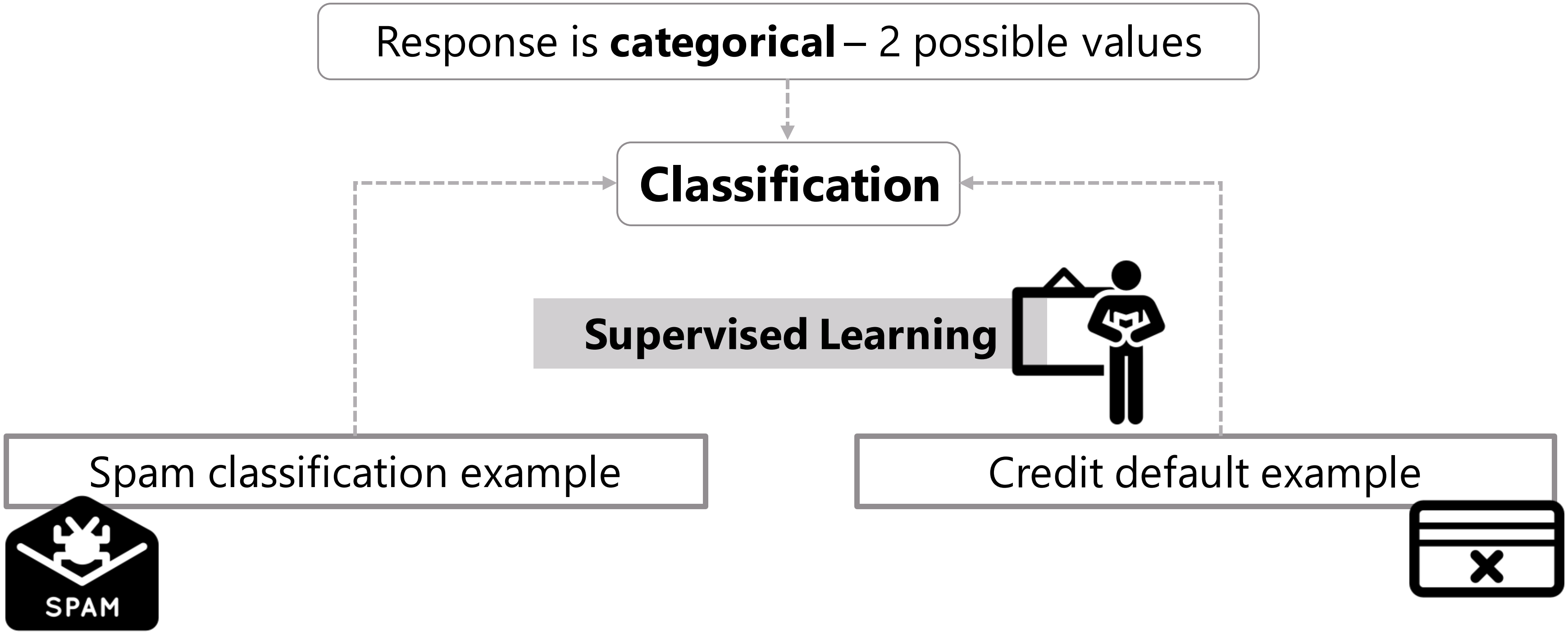
Types of Tasks

2 kinds of problems

Supervised Learning

Distinction depends on the type of response

Types of Tasks



Response is **categorical** – 2 possible values

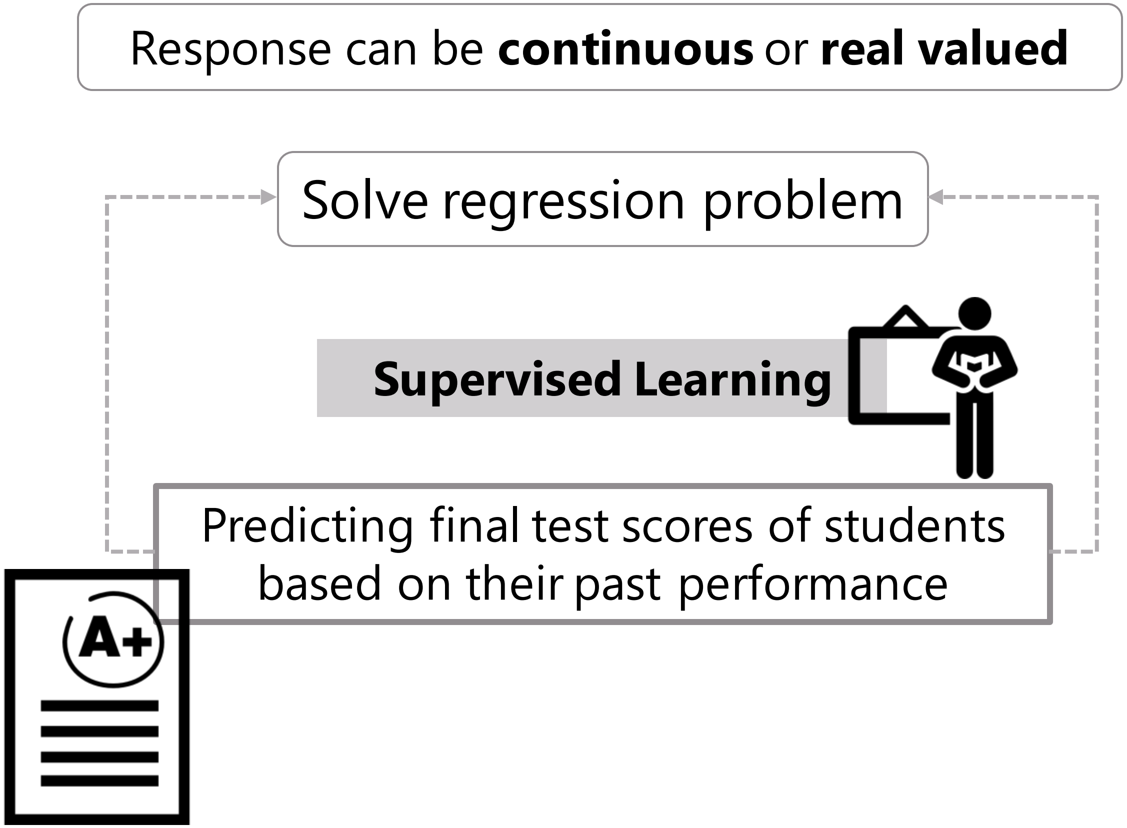
Classification

Supervised Learning

Spam classification example

Credit default example

Types of Tasks



Response can be **continuous** or **real valued**

Solve regression problem

Supervised Learning

Predicting final test scores of students based on their past performance

Types of Tasks

The machine gets no teacher – the data does not have a definitive response column

Unsupervised Learning

The computer finds meaningful patterns only from features

Most useful in finding groups in data, based on the features

Exercise

From the credit data, can you think of a case where you might need to group customers based only on the information regarding bill amounts and payment history?

Types of Tasks

Roots in behavioural psychology

The computer is shown each data point sequentially

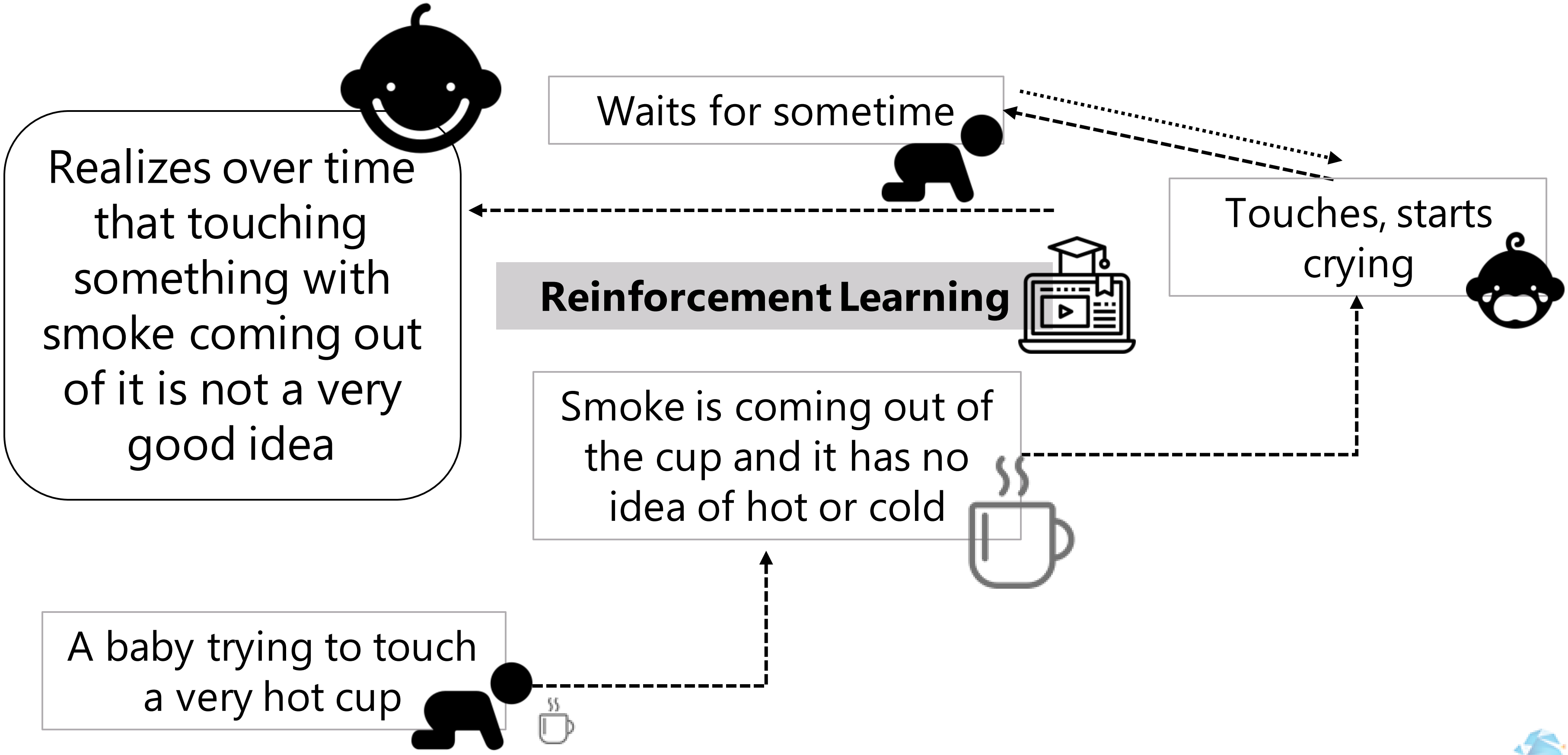
Reinforcement Learning

If it gets it right, it gets a reward and is penalized otherwise

While starting, it typically makes many mistakes but learns gradually

Types of Tasks

Example



A baby trying to touch a very hot cup

Smoke is coming out of the cup and it has no idea of hot or cold

Touches, starts crying

Waits for sometime

Reinforcement Learning

Realizes over time that touching something with smoke coming out of it is not a very good idea

Types of Tasks

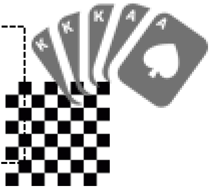
Example

Relatively new way of solving some problems in Machine Learning

Reinforcement Learning

Logic games are traditionally defined as a sequence of decisions

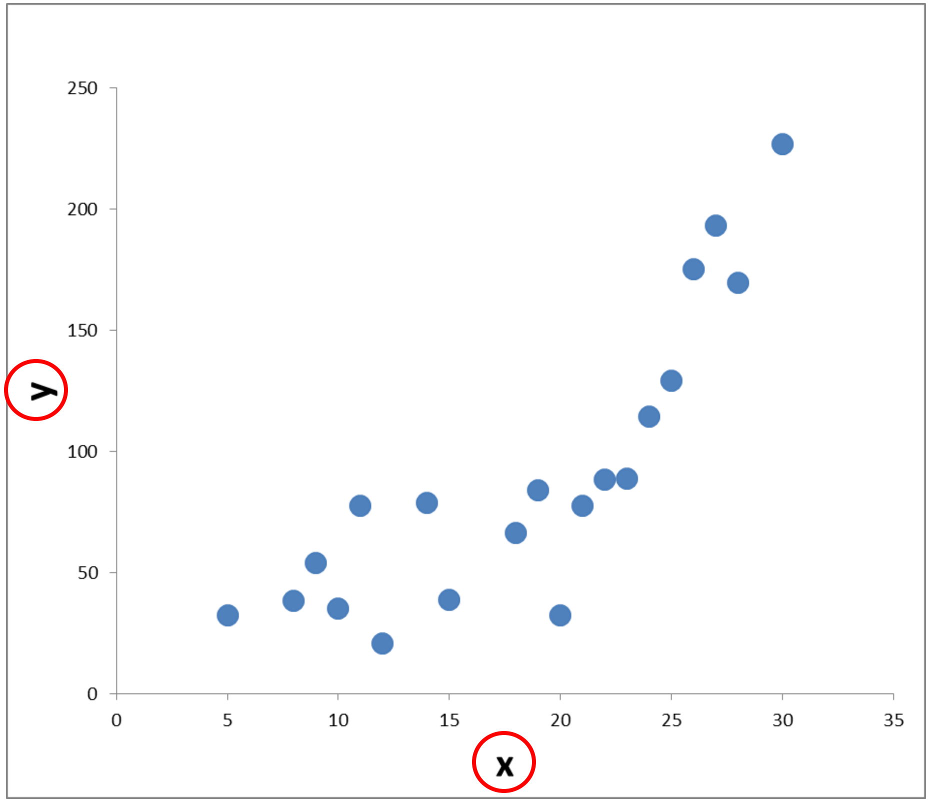
Example: Poker, Backgammon, Othello, Chess



Supervised Machine Learning Algorithm

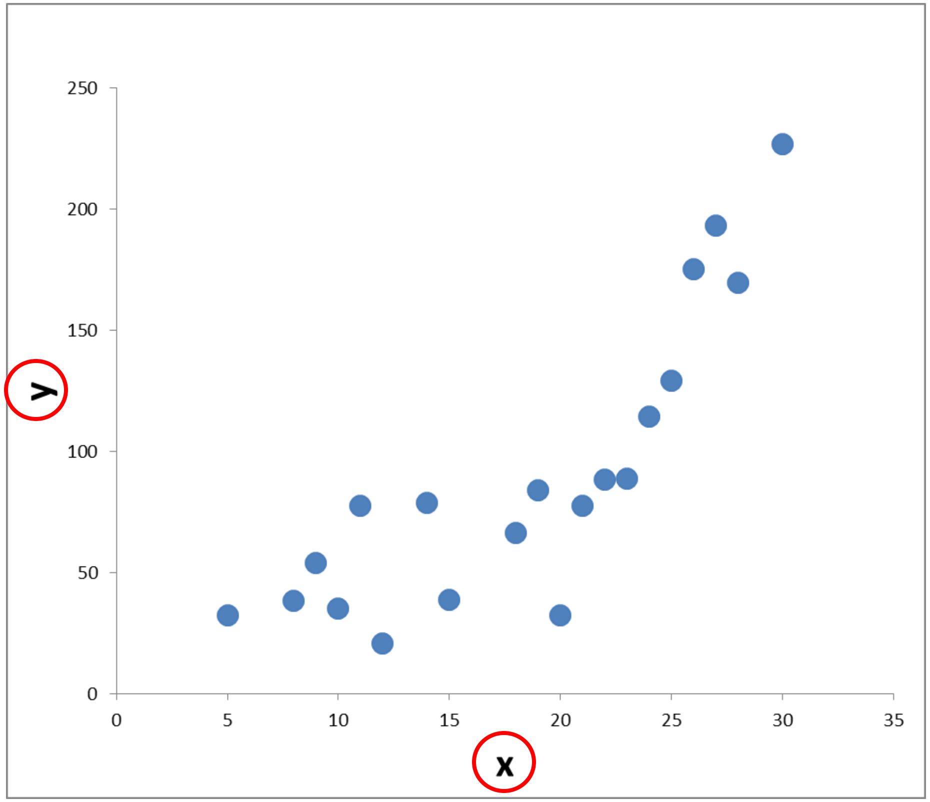
Requires a response, popularly denoted by **y** and at least one feature, denoted by **x**

How Does it Work?



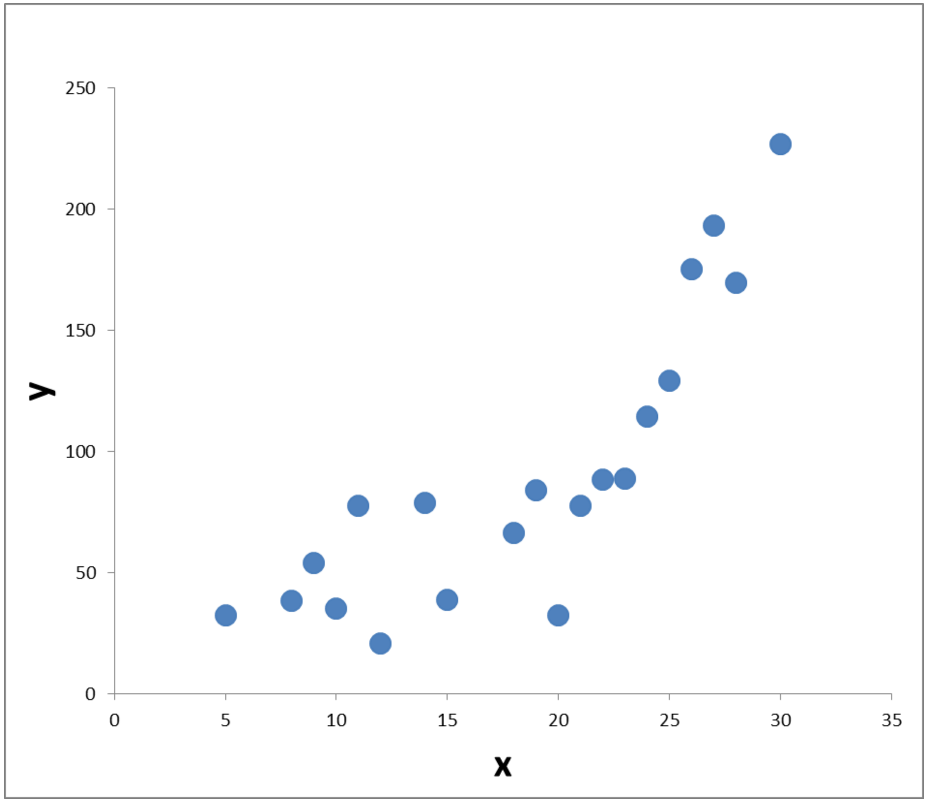
As **x** increases, the value of **y** tends to increase

How Does it Work?



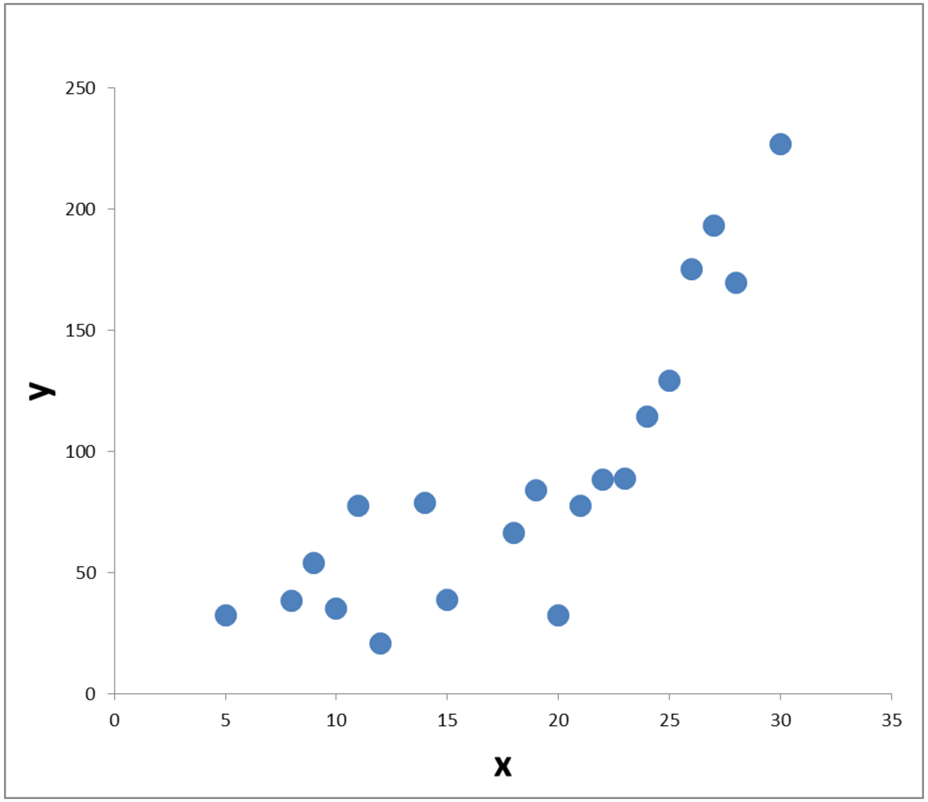
The exact relationship between the 2 variables is not known

How Does it Work?



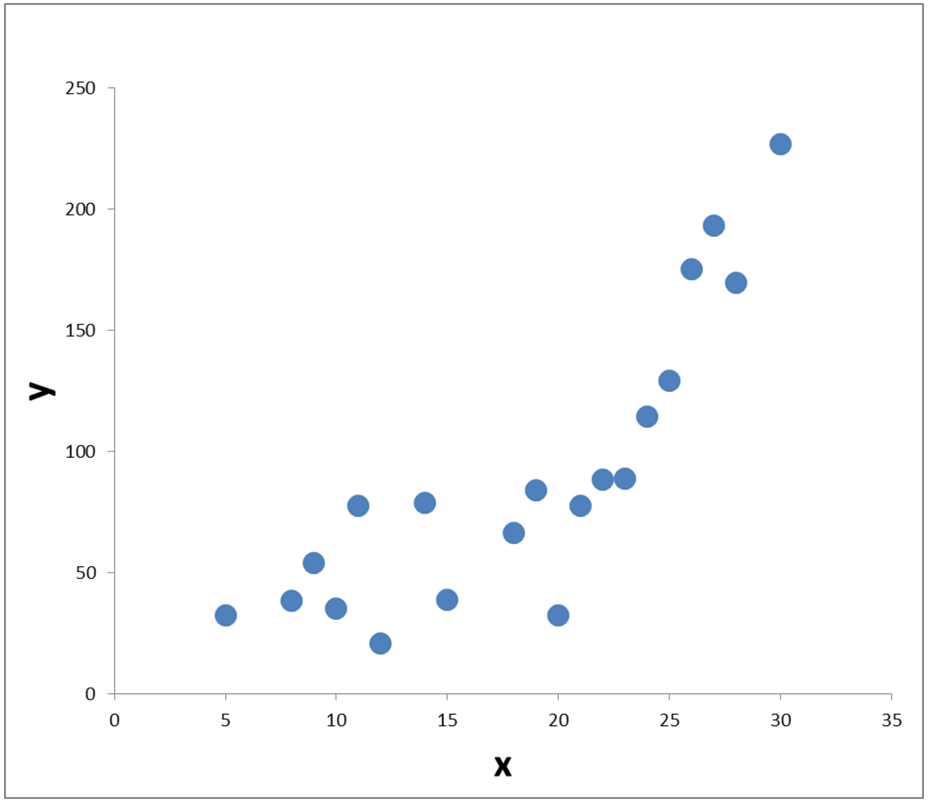
The task for the computer is to find something that approximates the true relationship

How Does it Work?



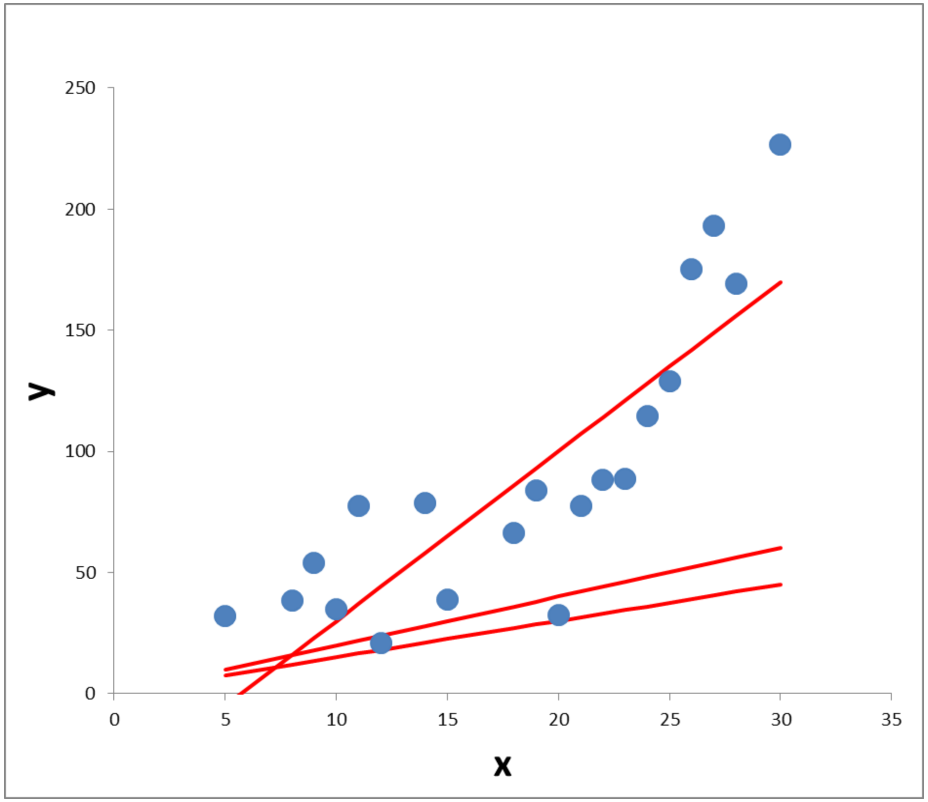
Finding a function capital **F** that best describes the data

How Does it Work?



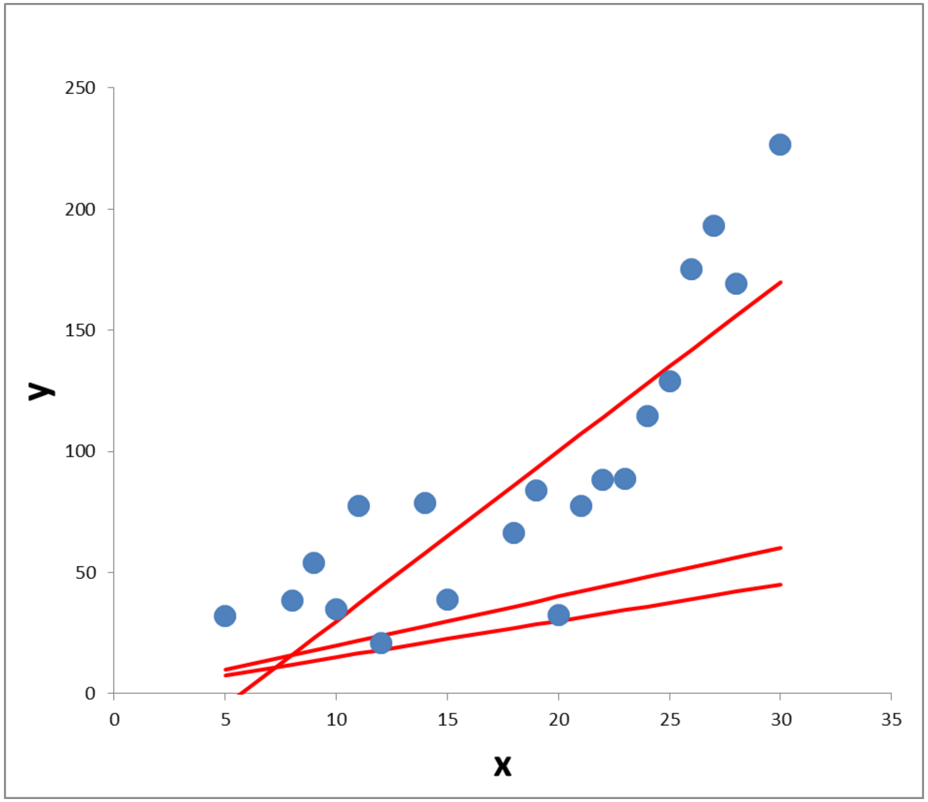
F can have many forms that are not necessarily algebraic

How Does it Work?



Assumption: F is linear

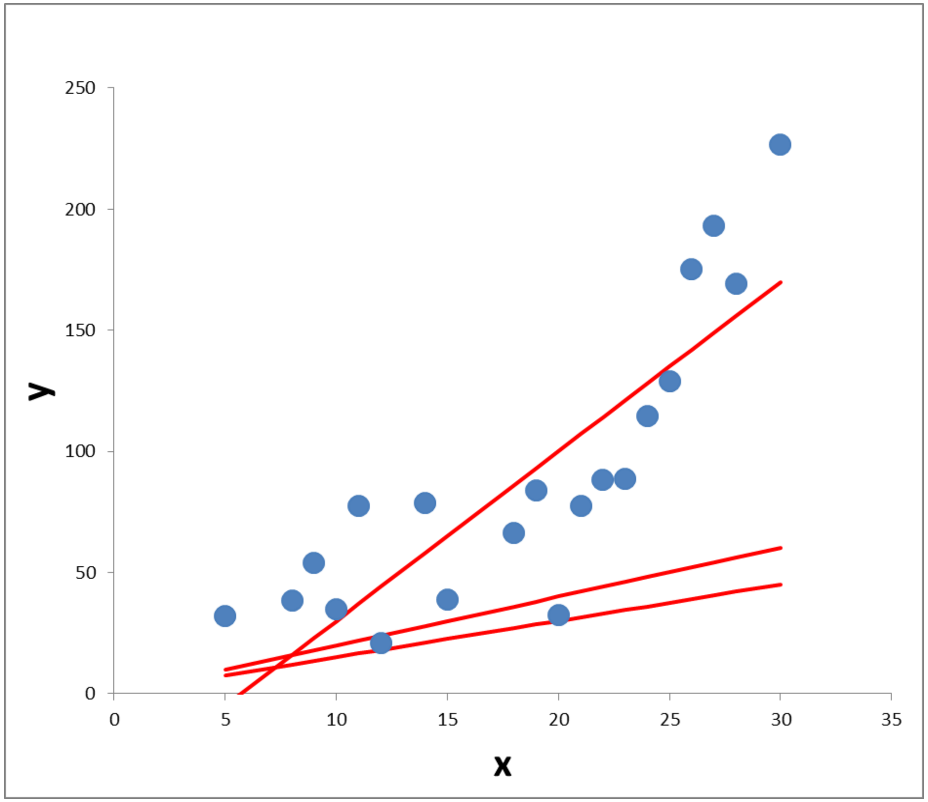
Machine Learning Algorithm



Equation: y = a + b*x*

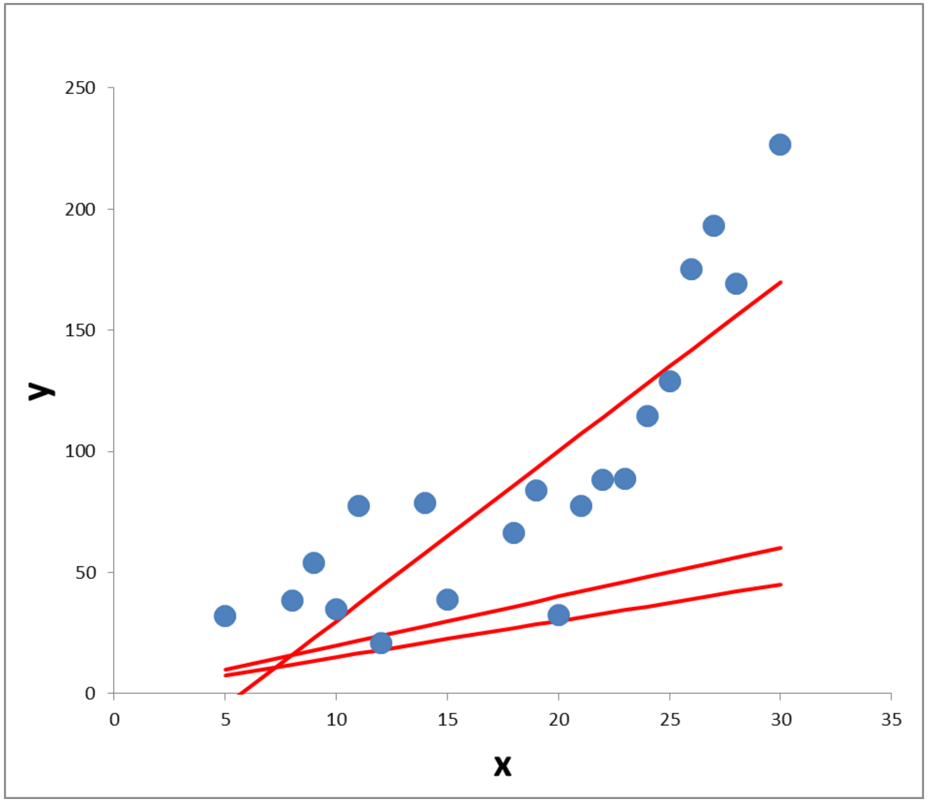
Therefore, F(x) = a + b x

Machine Learning Algorithm



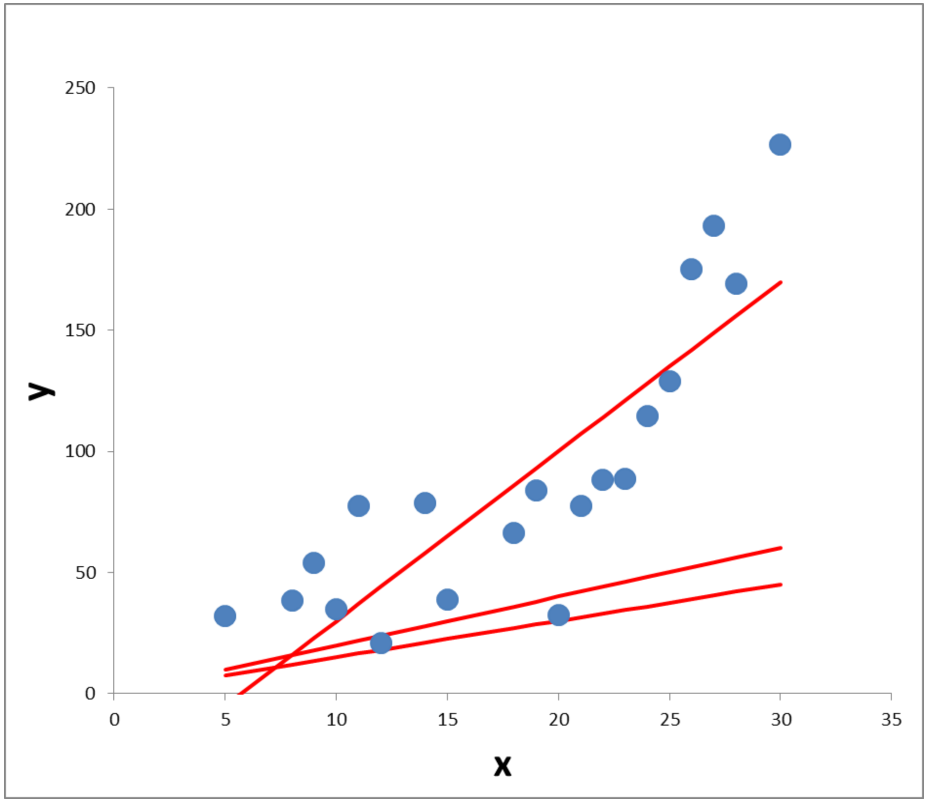
Finding the straight line that best describes the given data

Machine Learning Algorithm



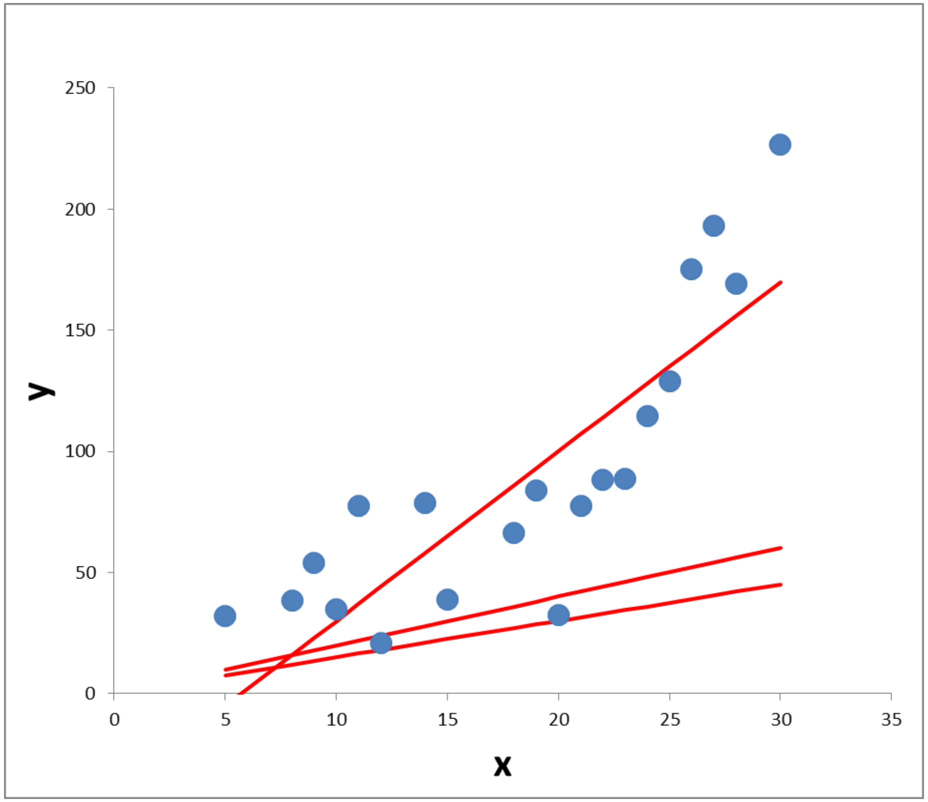
There are many red lines possible that seem to pass through the data

Machine Learning Algorithm



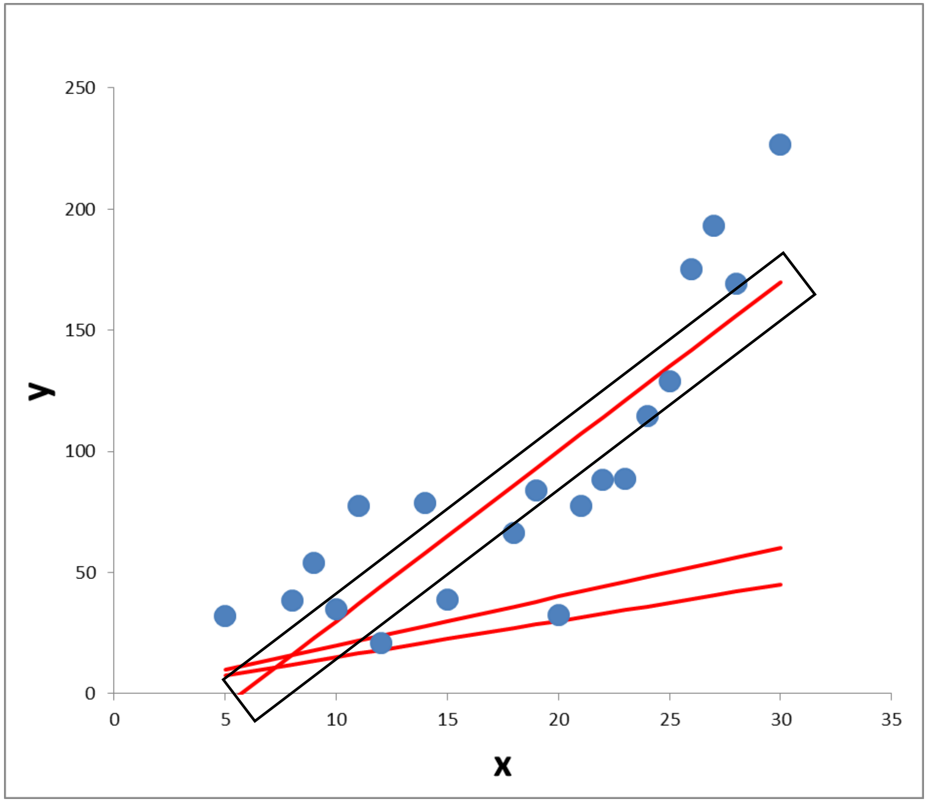
All are not equally good!

Machine Learning Algorithm



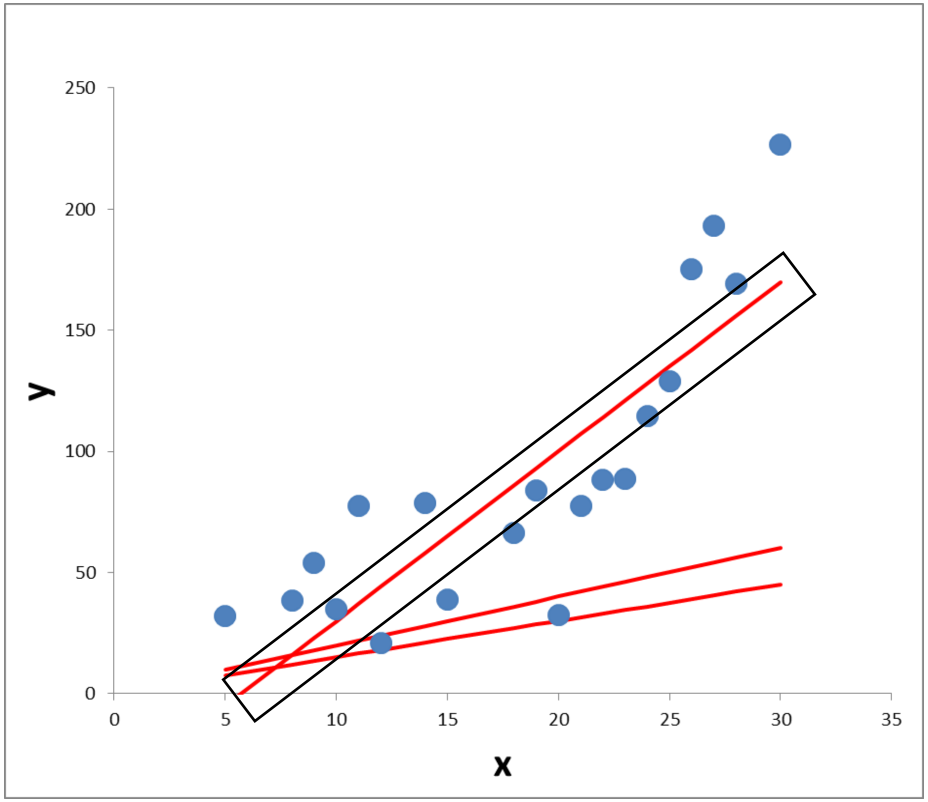
2 lines pass through the outside of the data boundary, 1 line passes through the middle of the data

Machine Learning Algorithm



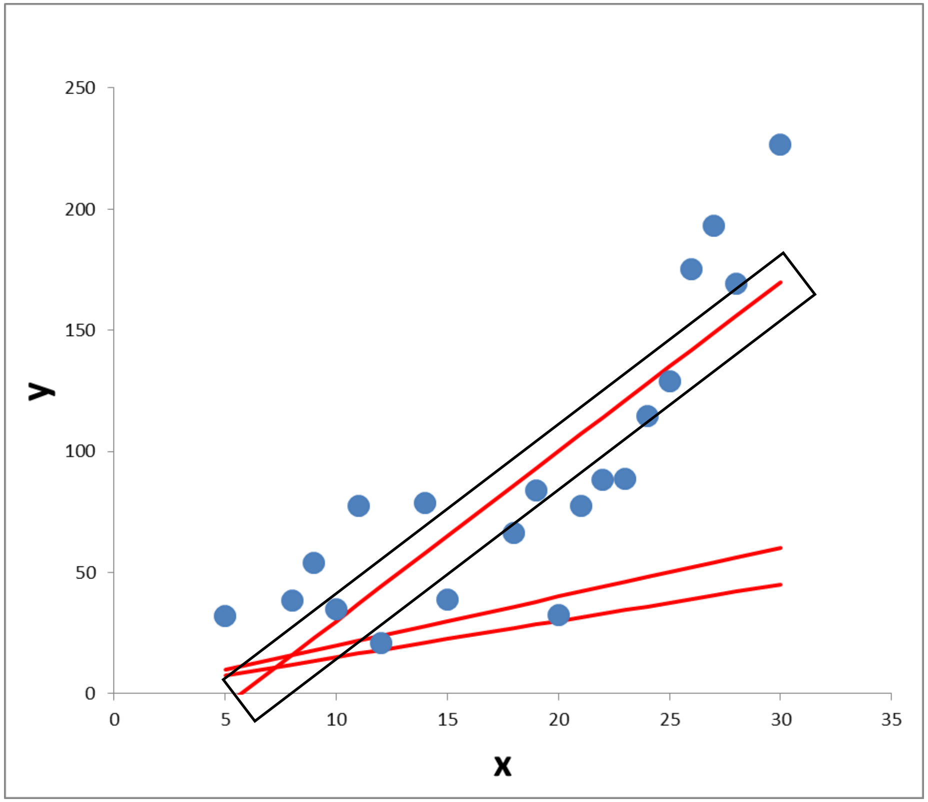
2 lines pass through the outside of the data boundary, 1 line passes through the middle of the data

Machine Learning Algorithm



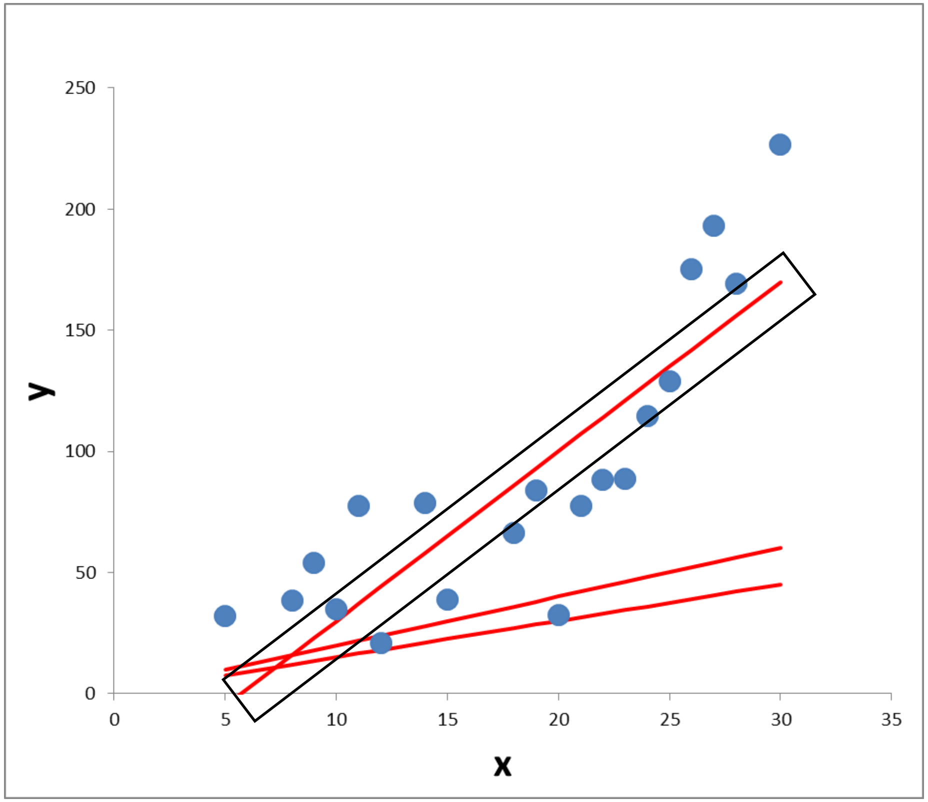
Why do you think that this line better summarizes the data than the other 2 lines?

Machine Learning Algorithm



Given these lines, how does the computer find the best line?

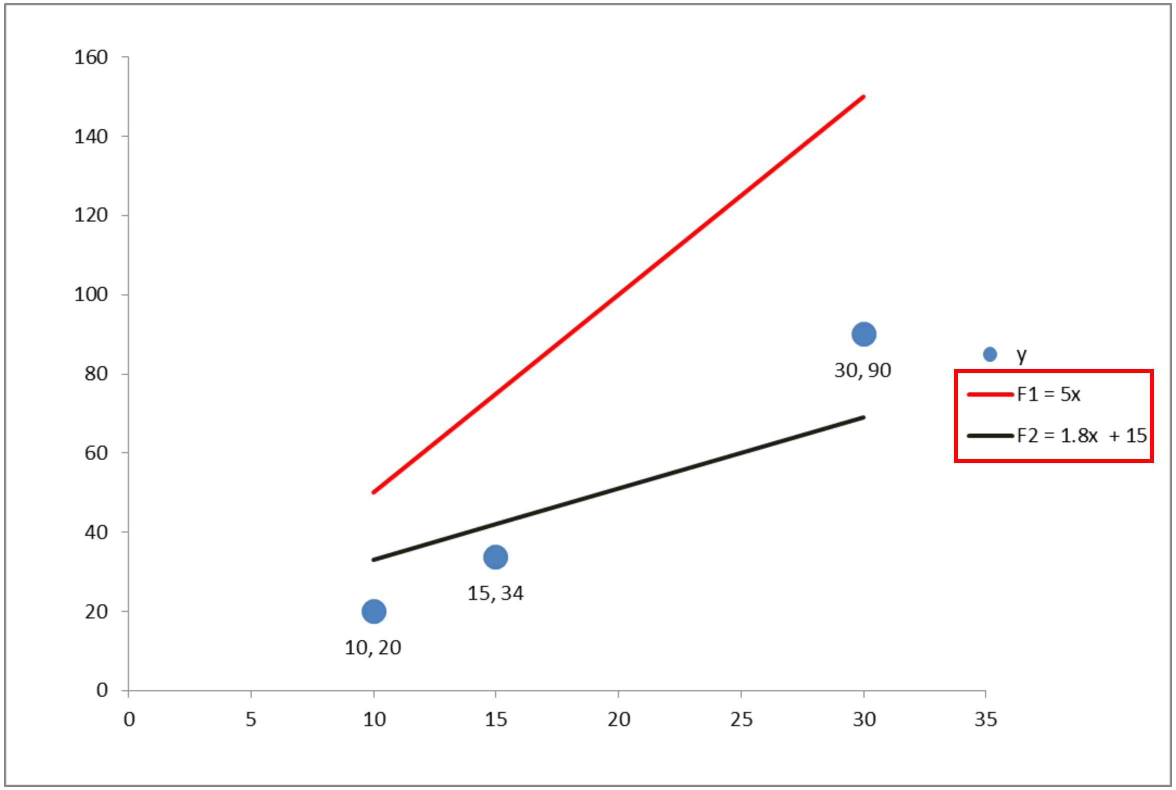
Machine Learning Algorithm



Define an error for each line; compute how much error each line makes

Linear Regression

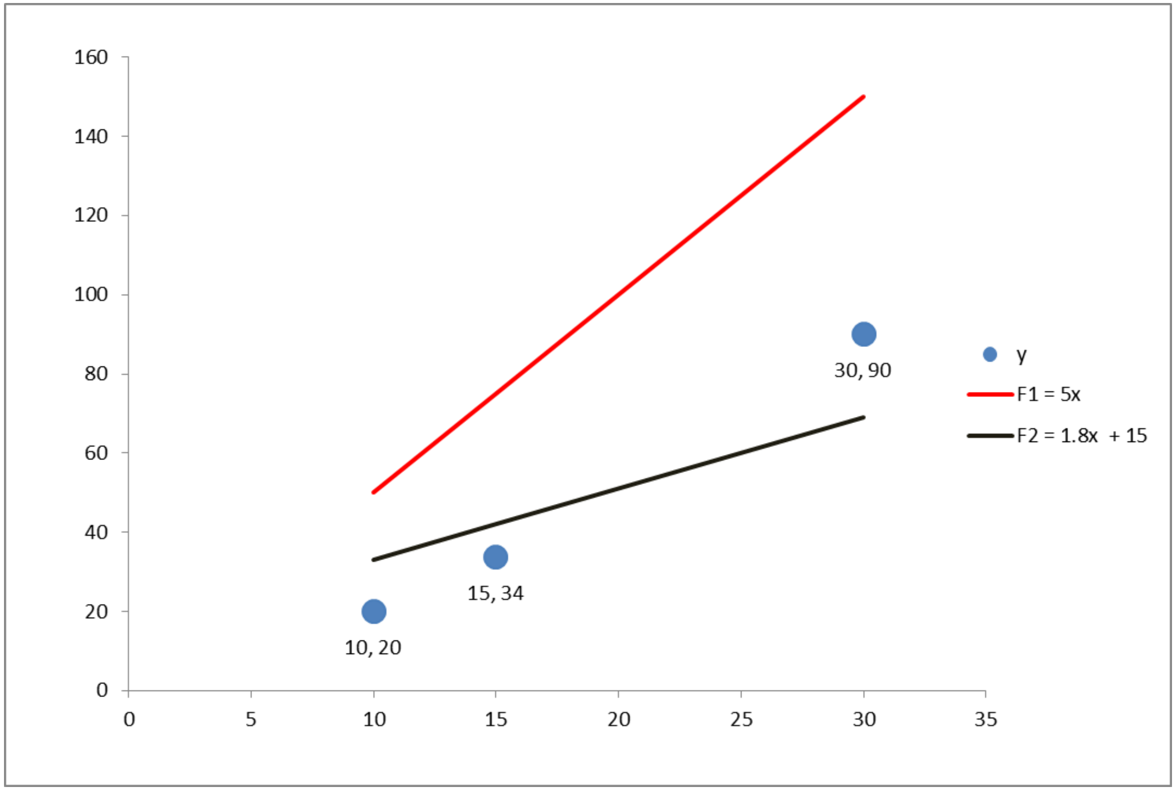
x	y
10	20
15	34
30	90



There are 2 functions **F1** and **F2** that act as candidate lines for summarizing the data

Linear Regression

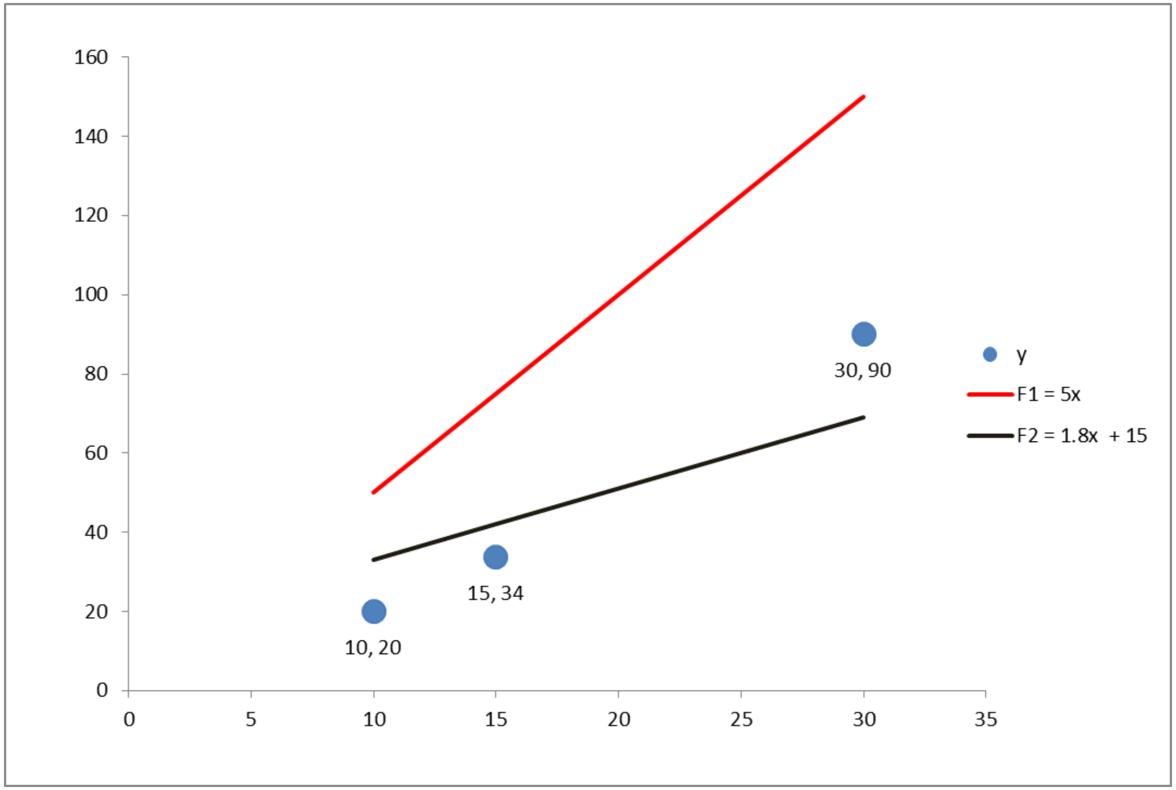
x	y
10	20
15	34
30	90



Plotting the data points along straight lines

Linear Regression

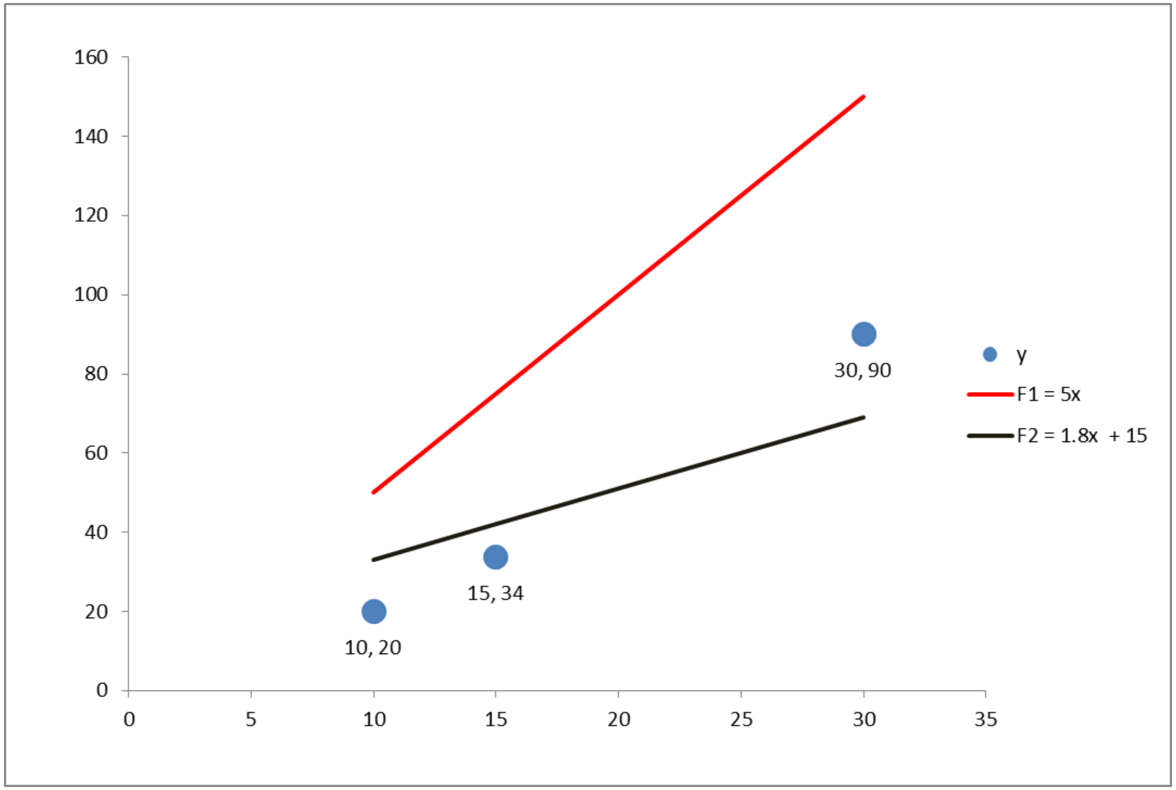
x	y
10	20
15	34
30	90



F2 is a better representation of data than **F1**

Linear Regression

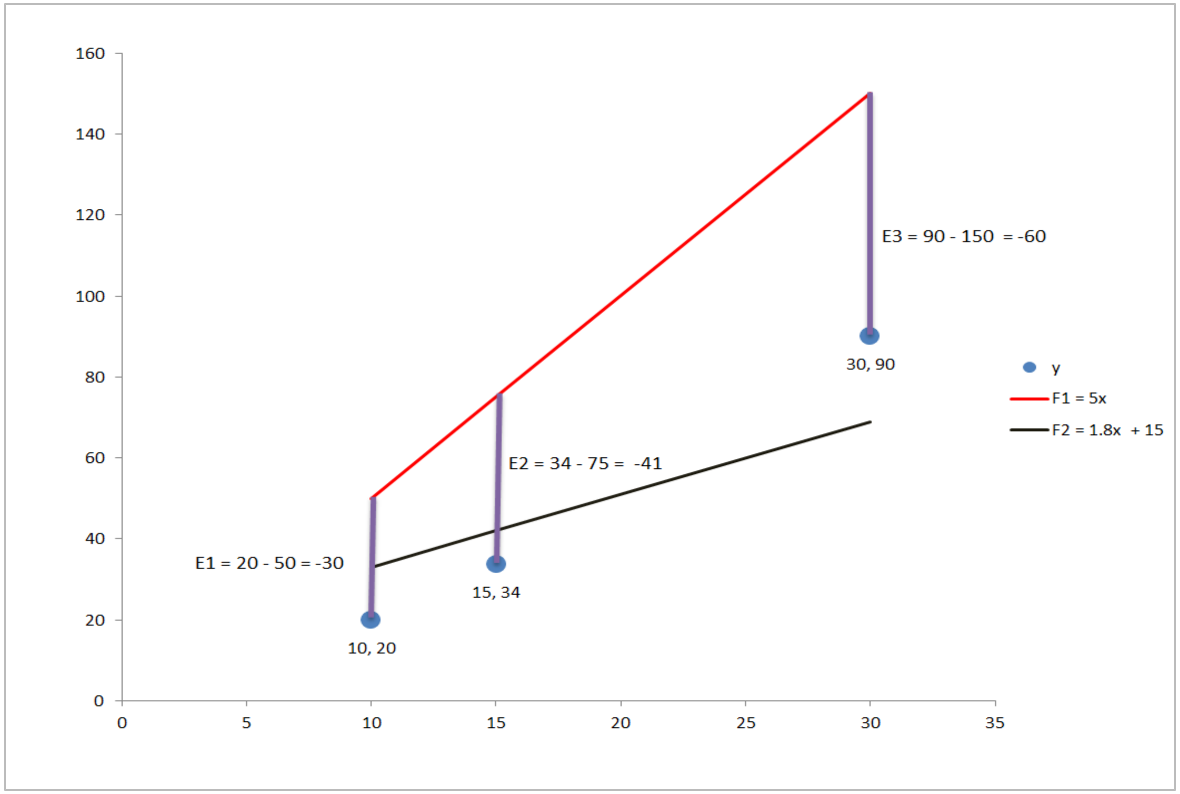
x	y
10	20
15	34
30	90



The computer needs to calculate how much error each line makes while summarizing the data

Calculating Errors

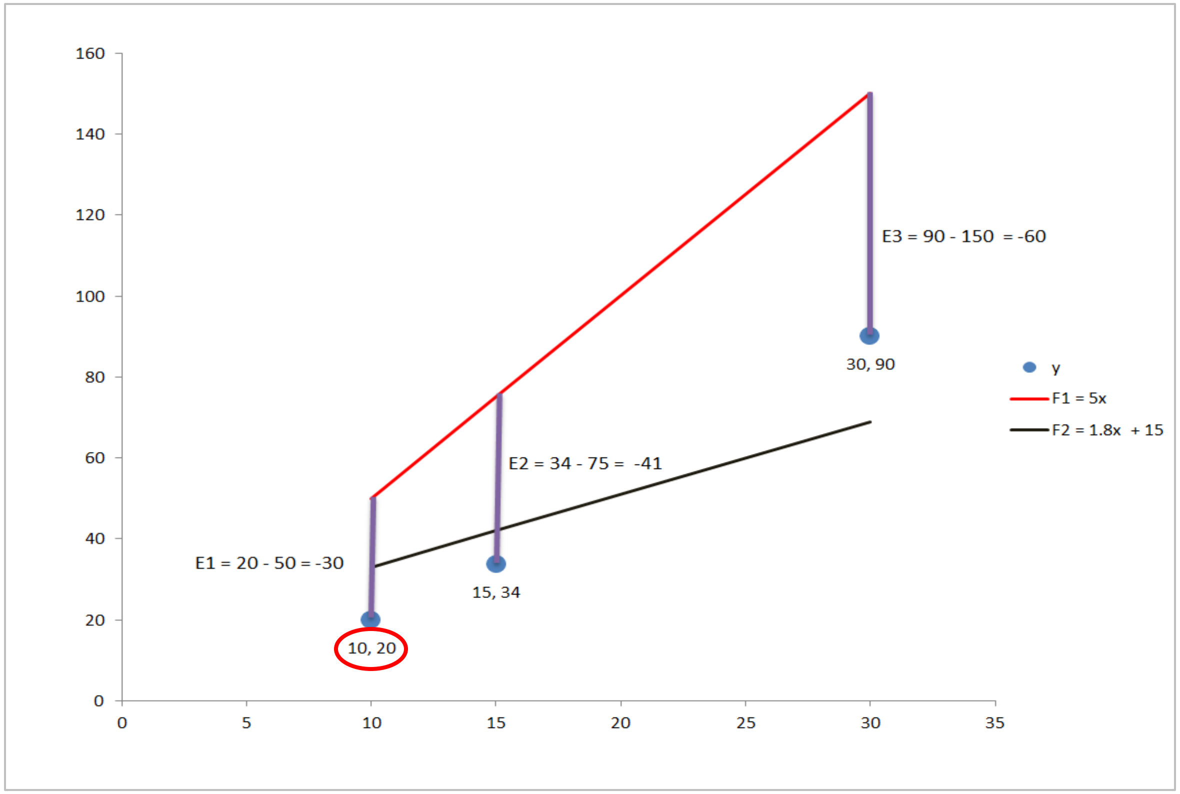
x	y
10	20
15	34
30	90



The error measure calculates vertical distance from points in the line to the data points

Calculating Errors

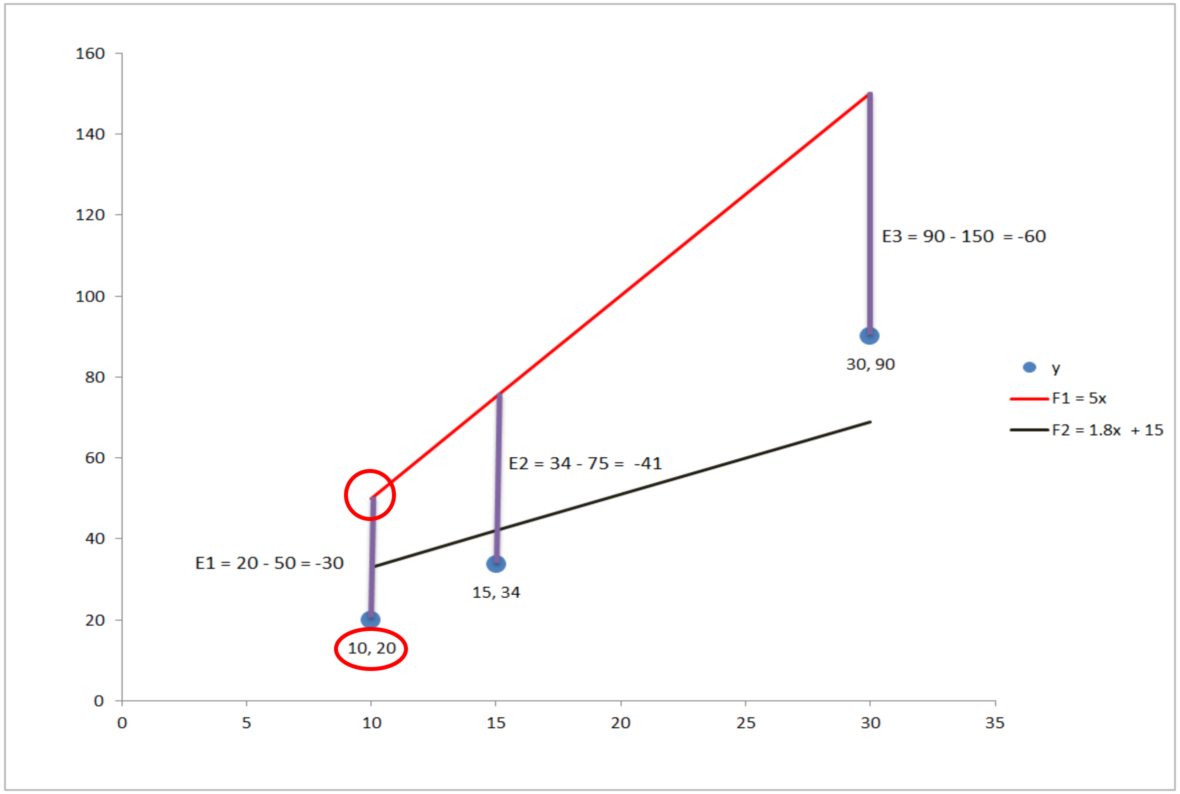
x	y
10	20
15	34
30	90



The error measure calculates vertical distance from points in the line to the data points

Calculating Errors

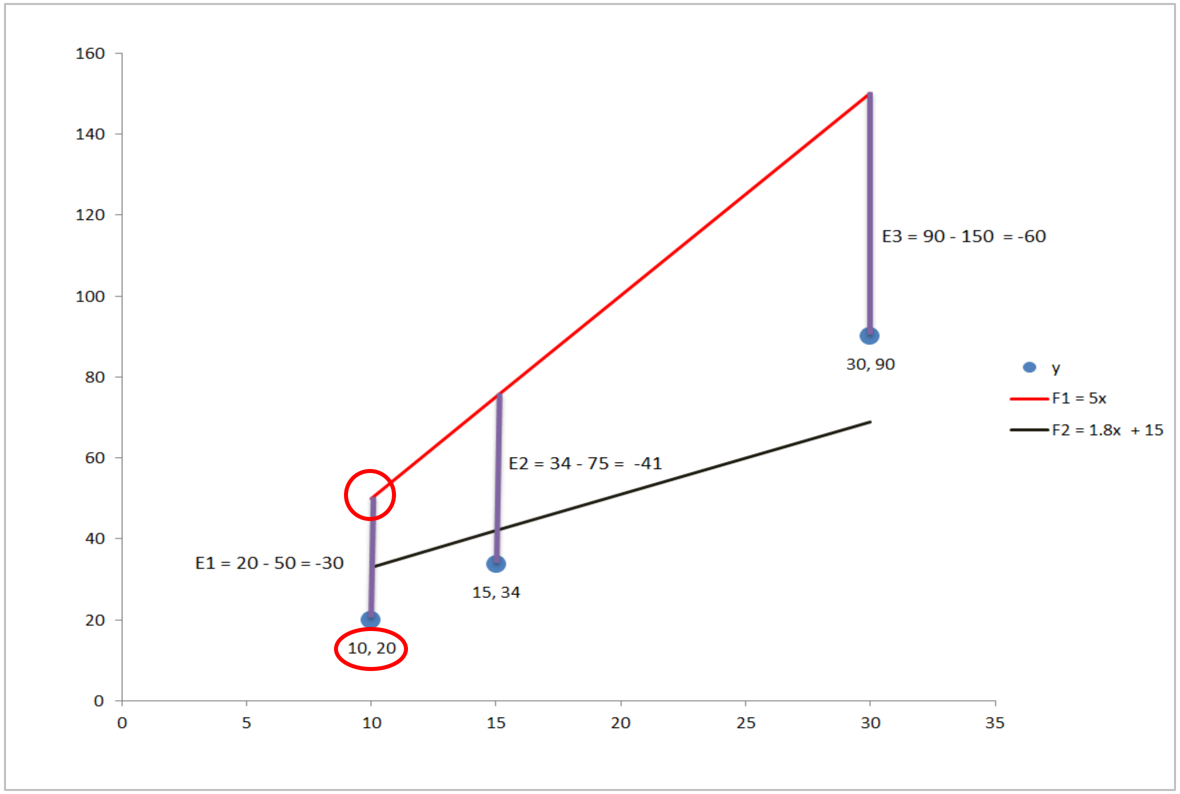
x	y
10	20
15	34
30	90



The function **F1** which is 5 times of **x**, puts the line at **50** on the **y** axis

Calculating Errors

x	y
10	20
15	34
30	90

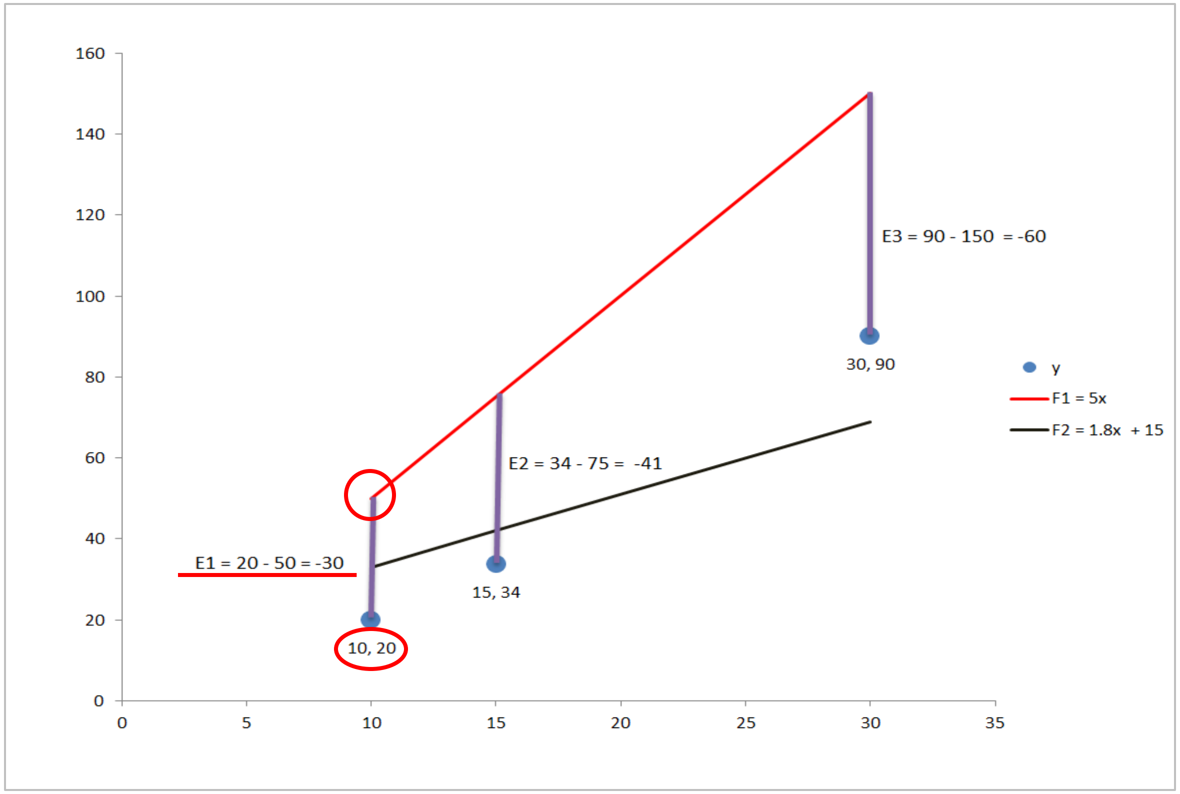


This according to **F1** is the expected value of **y** when **x** is equal to **10**

Calculating Errors

x	y
10	20
15	34
30	90

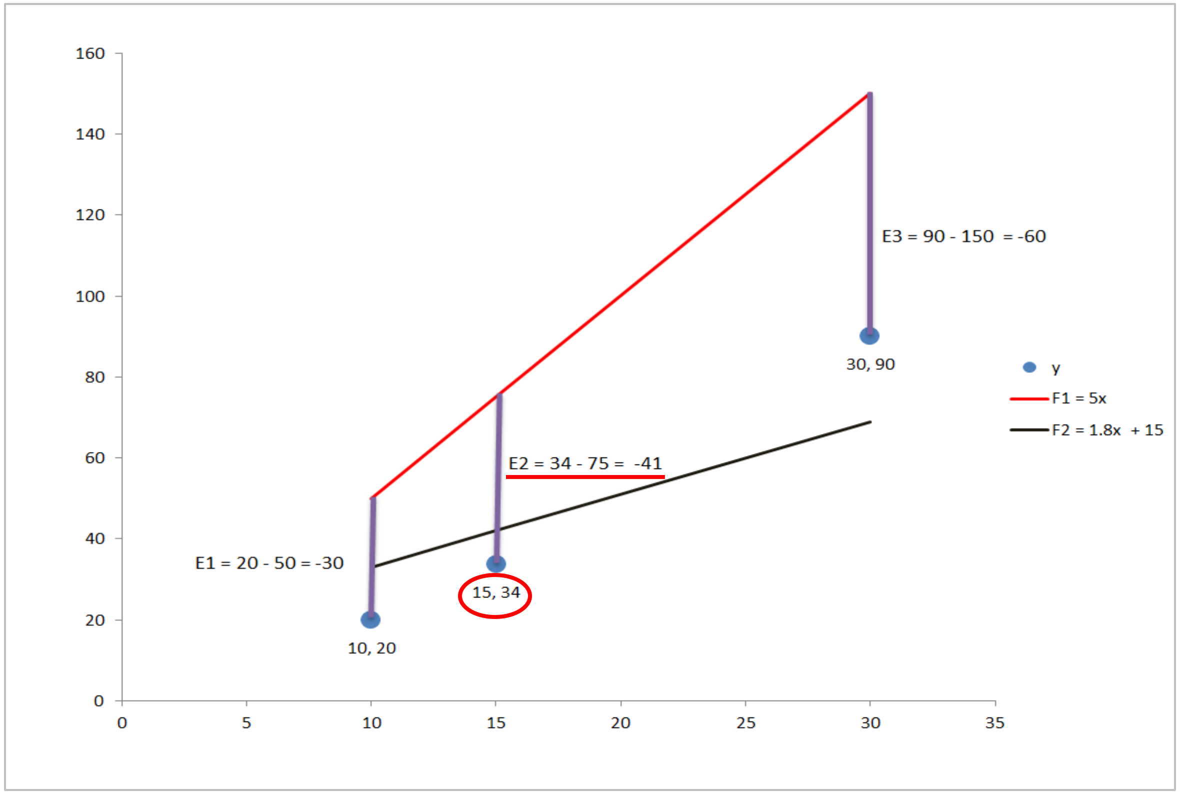
The error the line makes for the first point is the difference between the observed **y** and the expected **y**



Calculating Errors

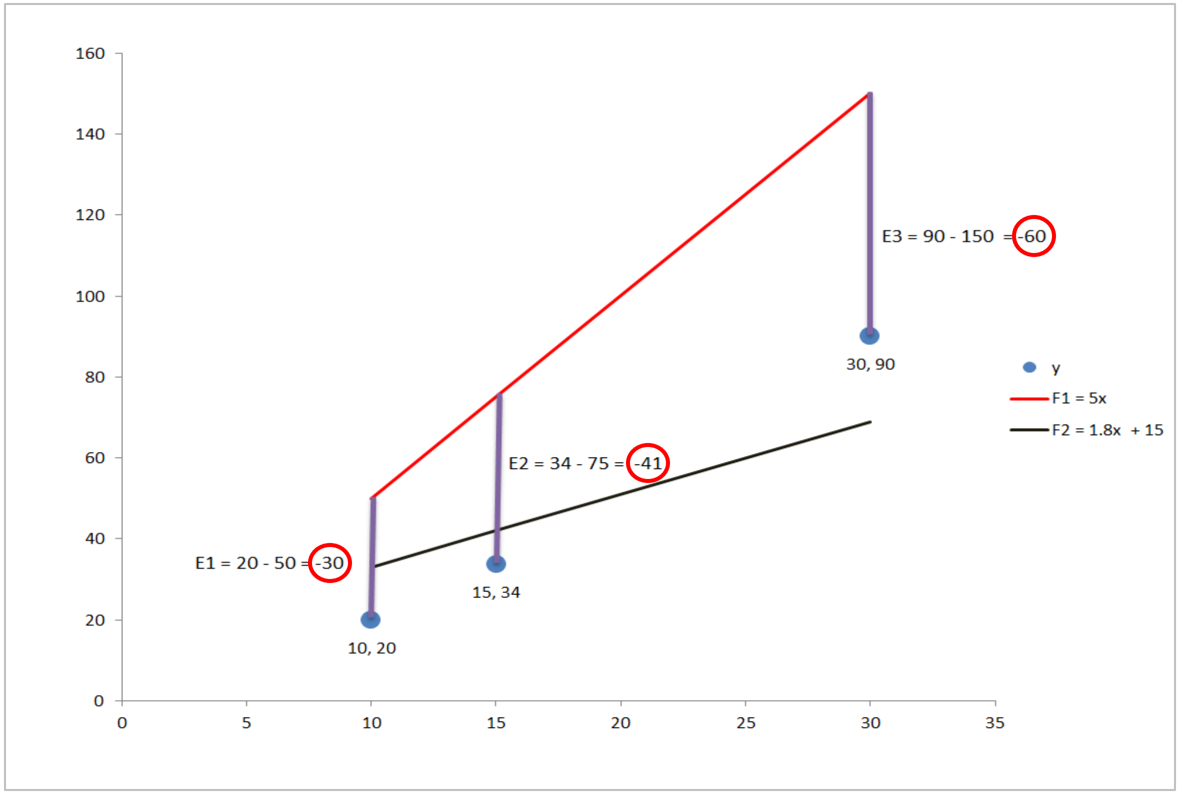
x	y
10	20
15	34
30	90

The error for the second point **15, 34** is **-41**



Calculating Errors

x	y
10	20
15	34
30	90



Sum up errors for all points to get the total error that line **F1** makes over entire data

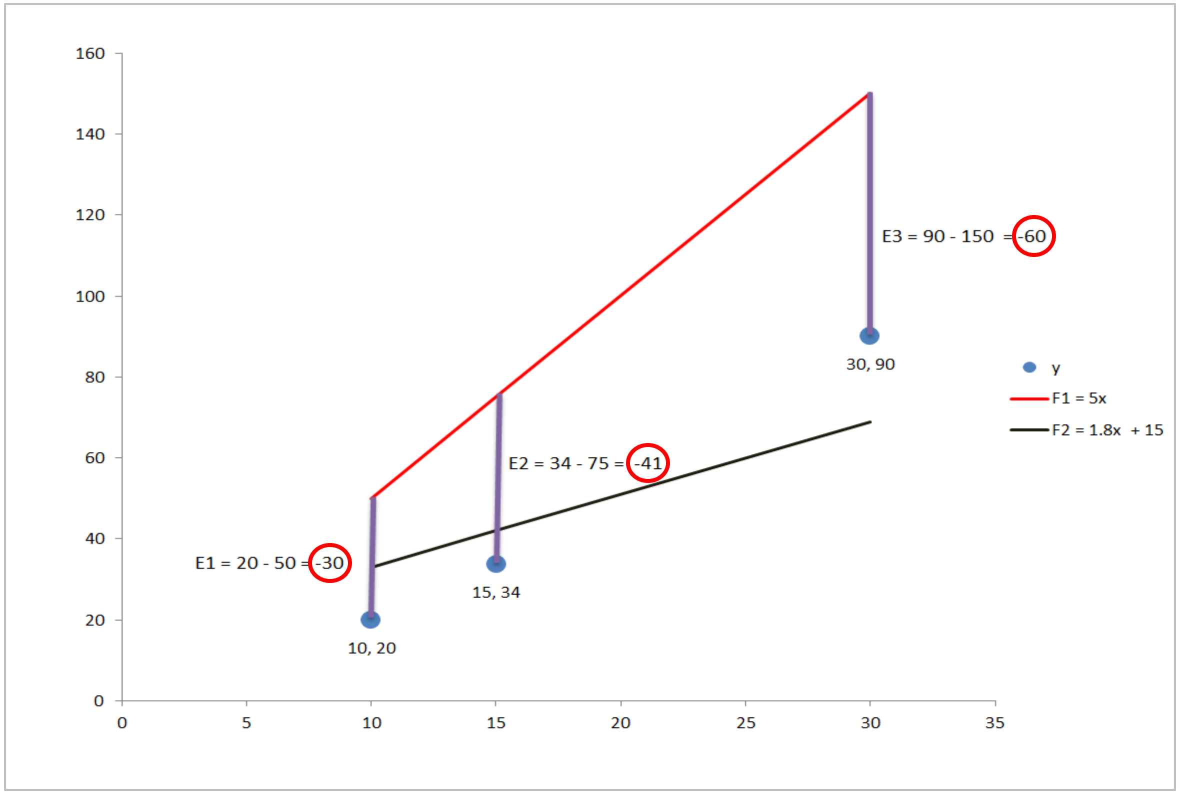
Calculating Errors

x	y
10	20
15	34
30	90

Total error for **F1**

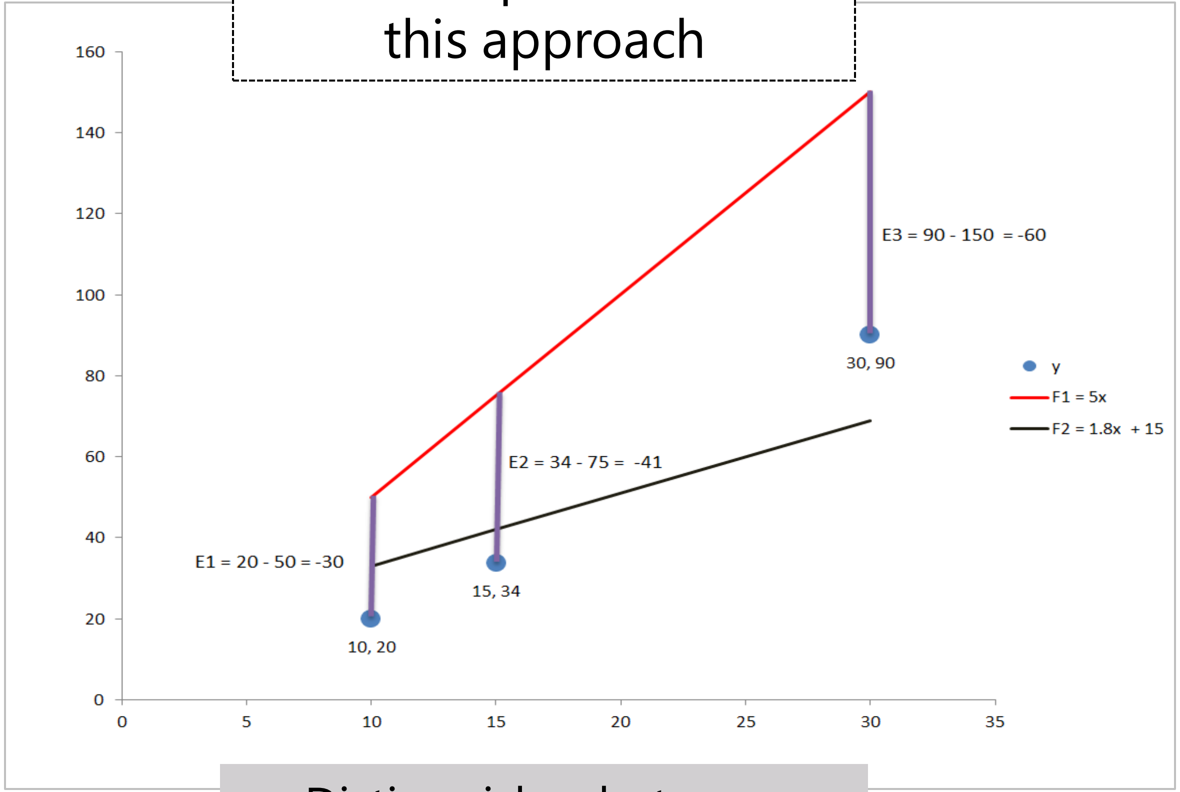
= E1 + E2 + E3

= -30 – 41 – 60 = -131



Calculating Errors

x	y
10	20
15	34
30	90



Potential problem with this approach

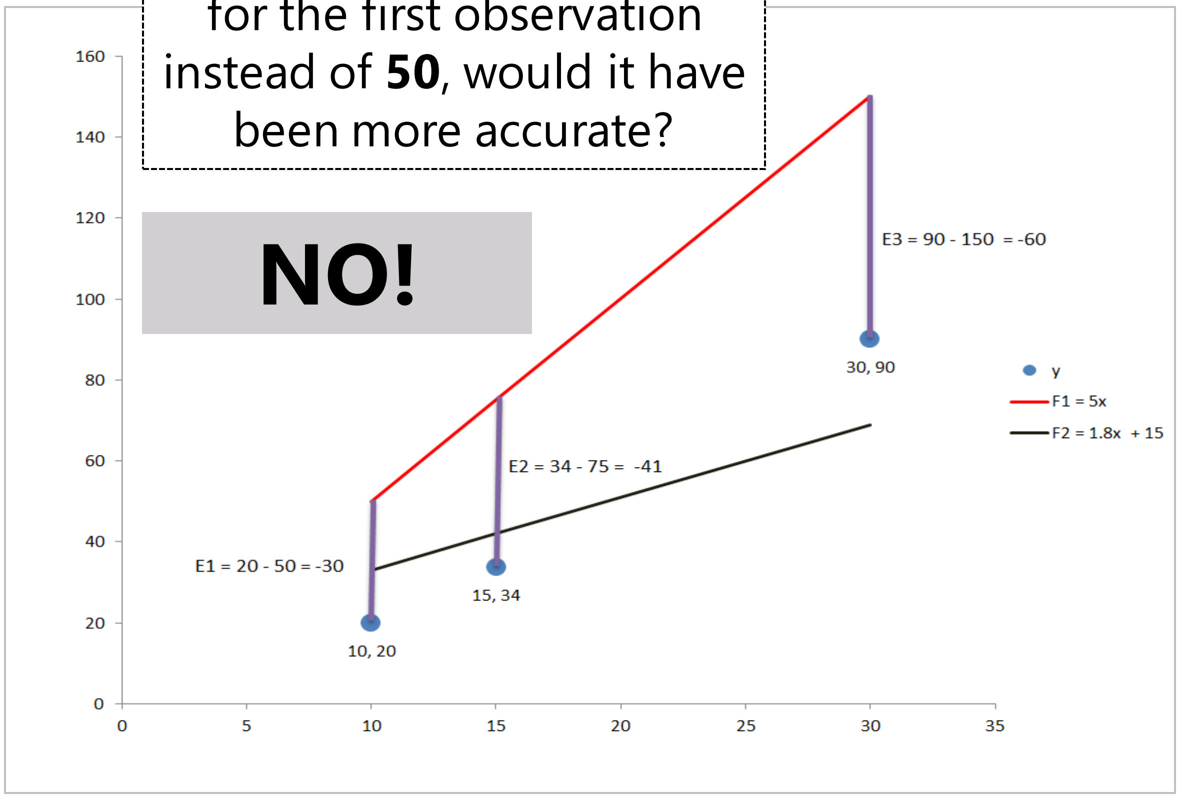
Distinguishes between negative and positive error

Calculating Errors

x	y
10	20
15	34
30	90

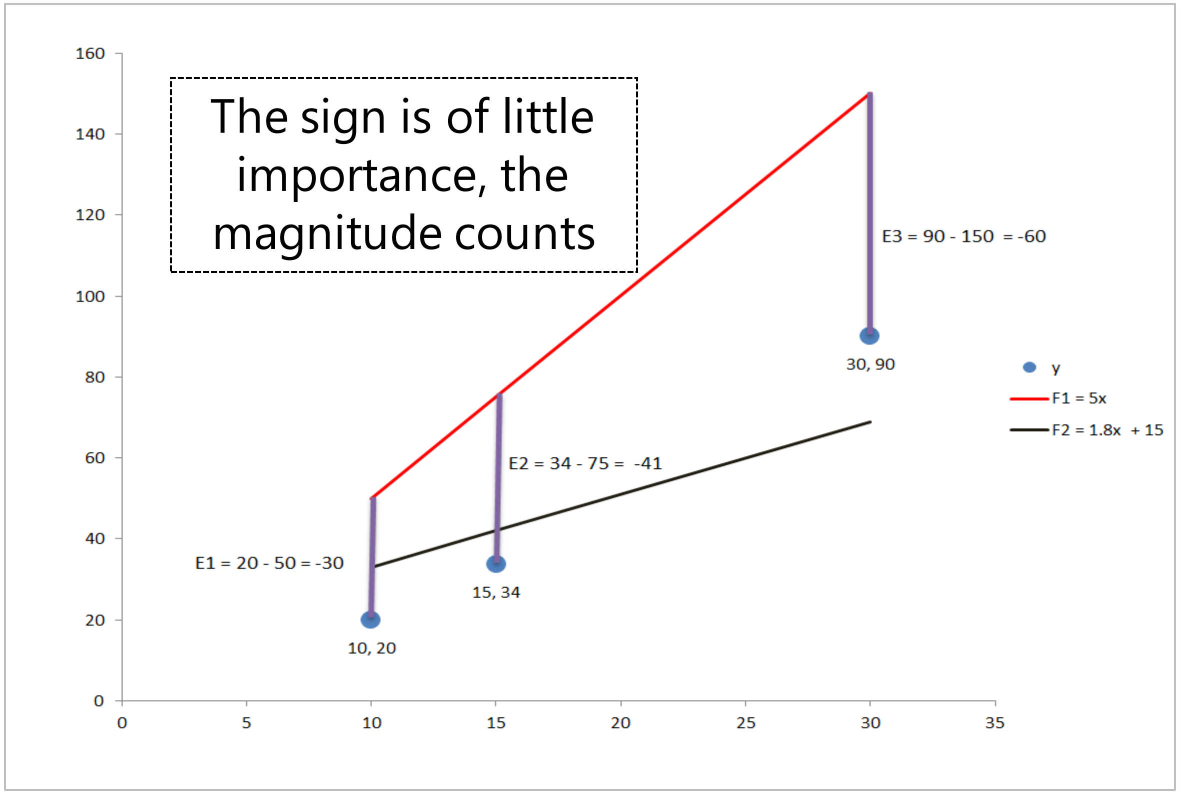
Had the line predicted **-10** for the first observation instead of **50**, would it have been more accurate?

NO!



Calculating Errors

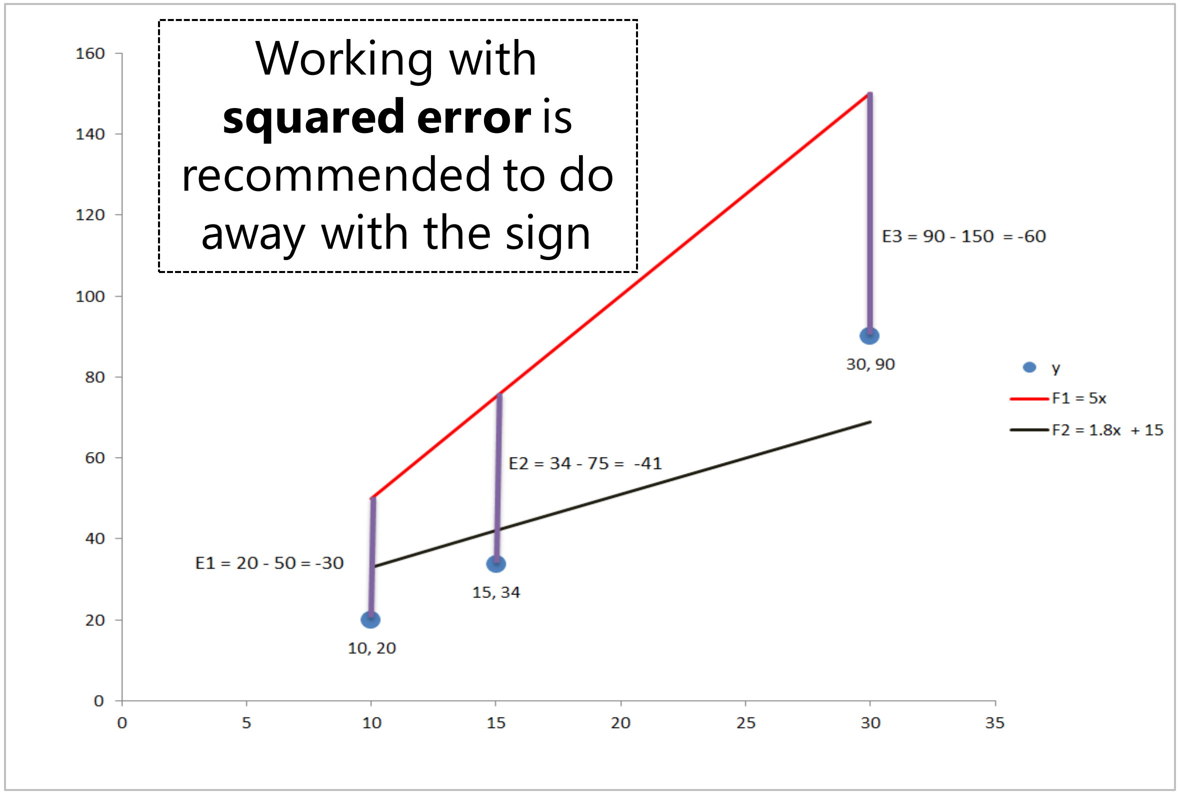
x	y
10	20
15	34
30	90



The sign is of little importance, the magnitude counts

Calculating Errors

x	y
10	20
15	34
30	90

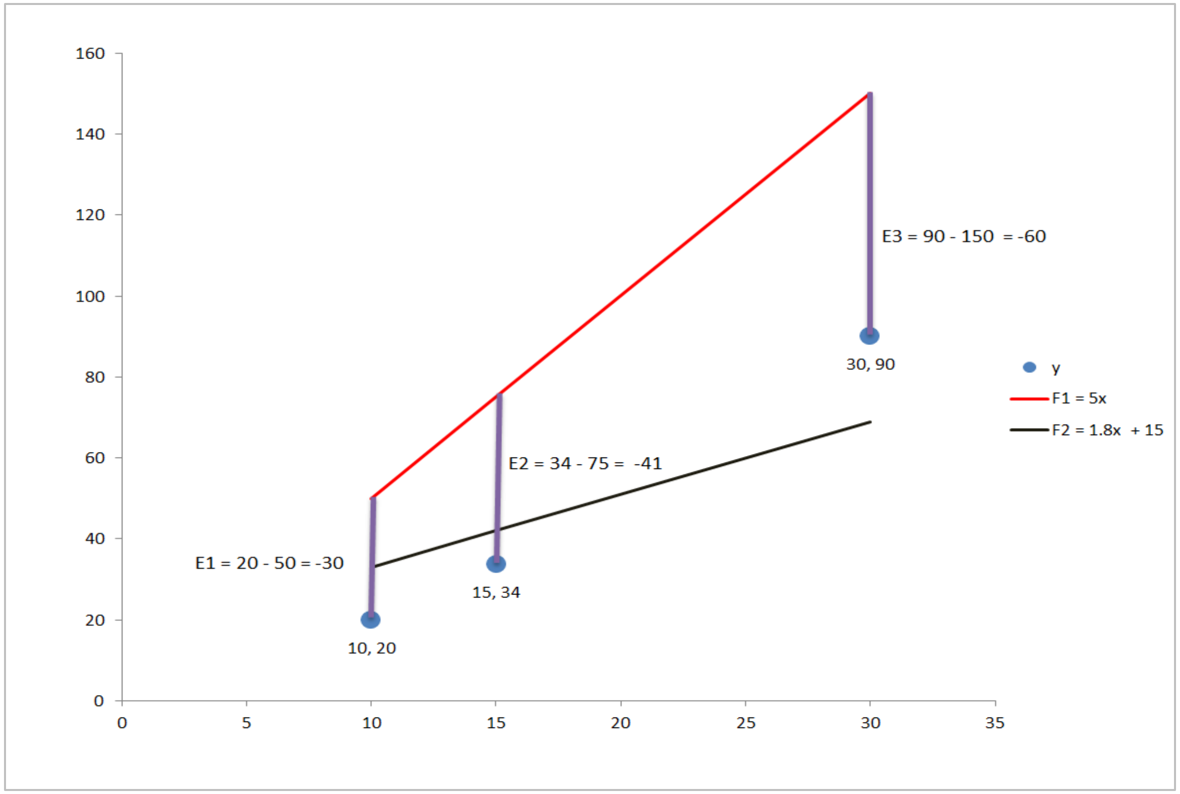


Working with **squared error** is recommended to do away with the sign

Calculating Errors

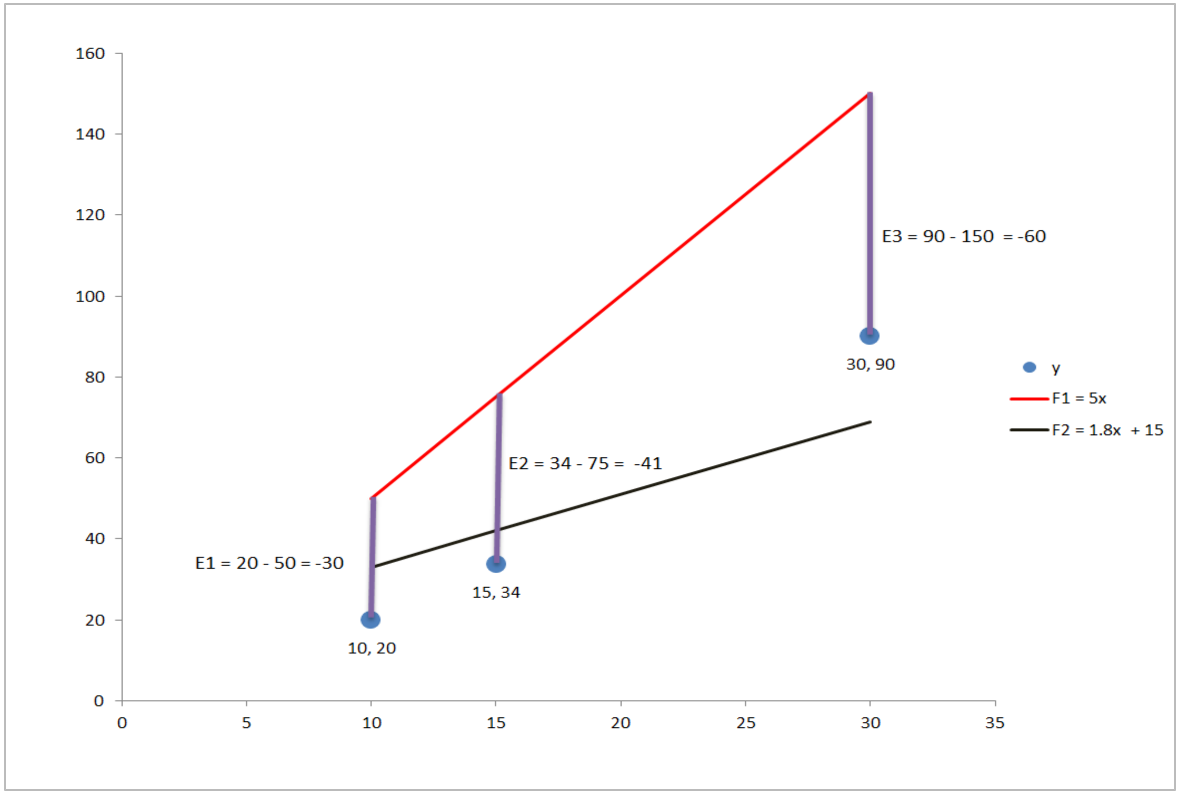
x	y
10	20
15	34
30	90

Calculate the error for each point made by the line



Calculating Errors

x	y
10	20
15	34
30	90

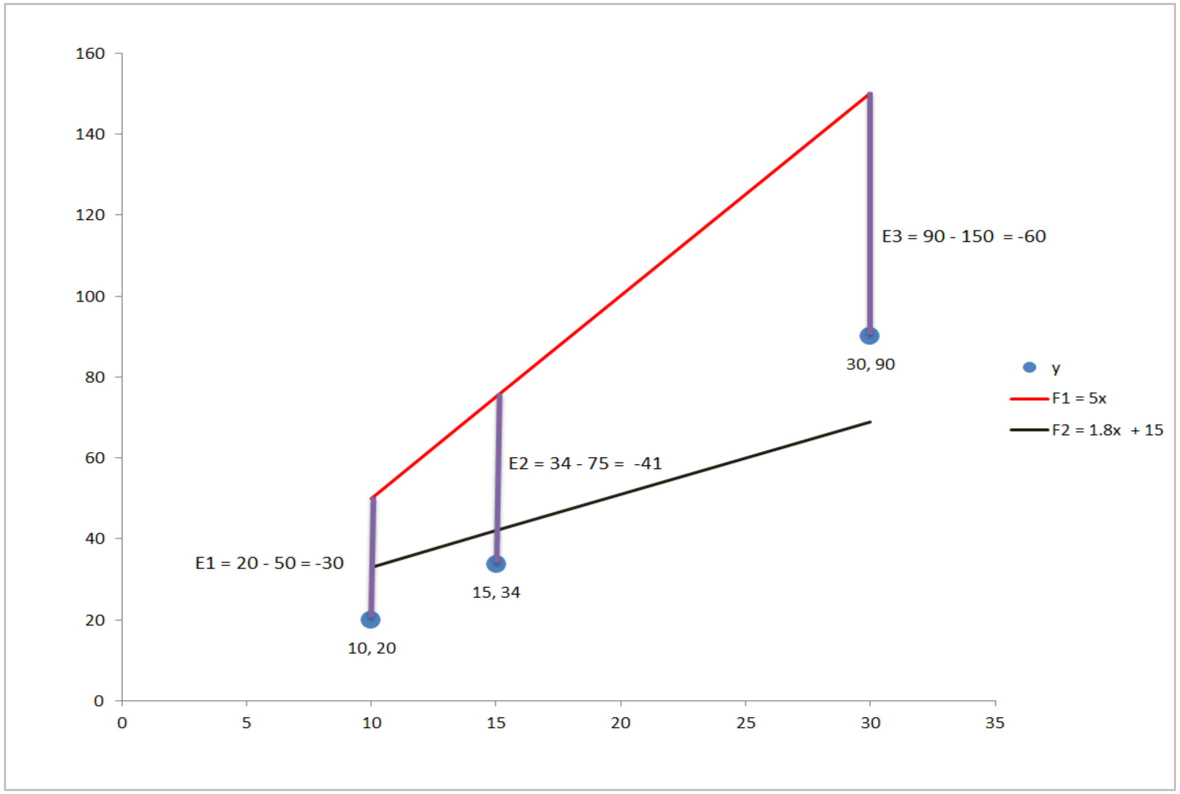


Take the squares and sum them up to get the total squared error made by the line **F1** equals **5** times **x** in summarizing the given data

Calculating Errors

x	y
10	20
15	34
30	90

A prediction of **-10** and **50** for the first point with **x = 10**, both get a squared error of **900**, instead of **-30** and **+30** respectively



Calculating Errors

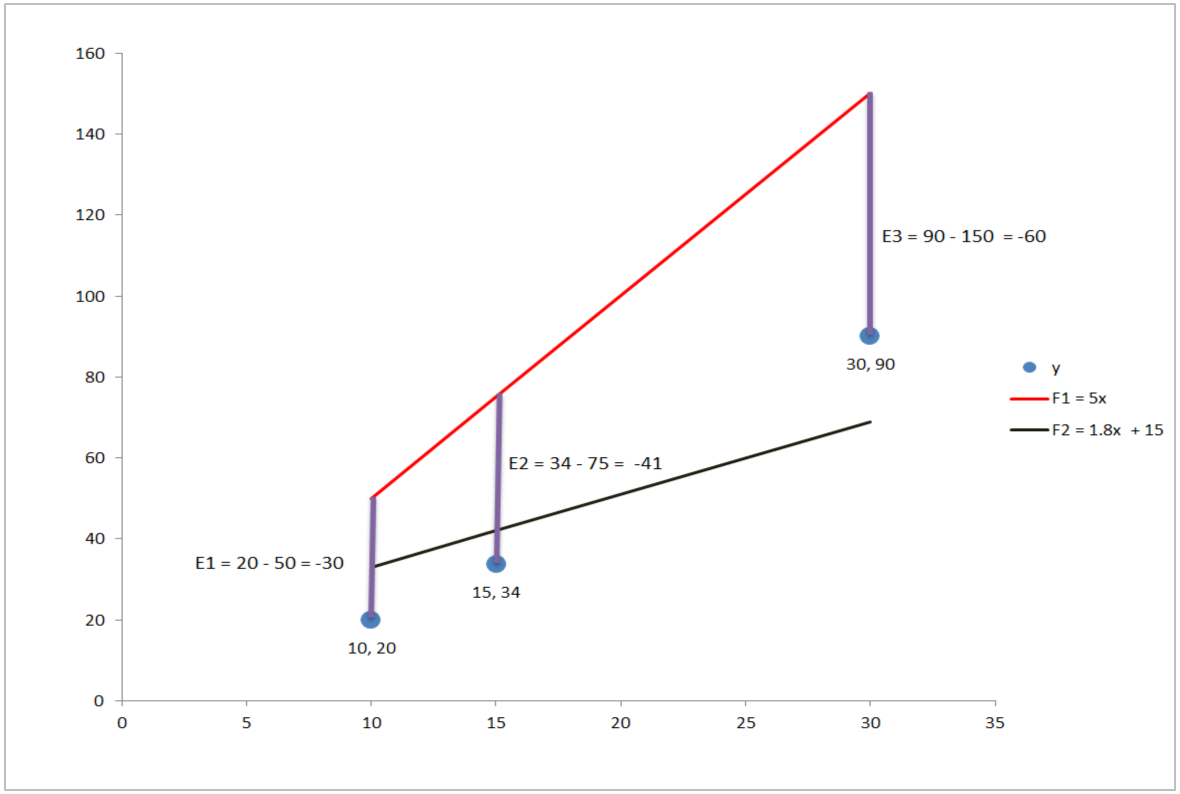
x	y
10	20
15	34
30	90

Total squared error for **F1**

$$= E_1^{\ 2} + E_2^{\ 2} + E_3^{\ 2}$$

= 900 + 1681 + 3600

~ **6,181**



What We Have Learnt

Introduction to Machine Learning – Definitions

Examples of Machine Learning in Real Life

Feature Engineering – Converting Raw Data to Meaningful Information

Types of Tasks Performed in Machine Learning

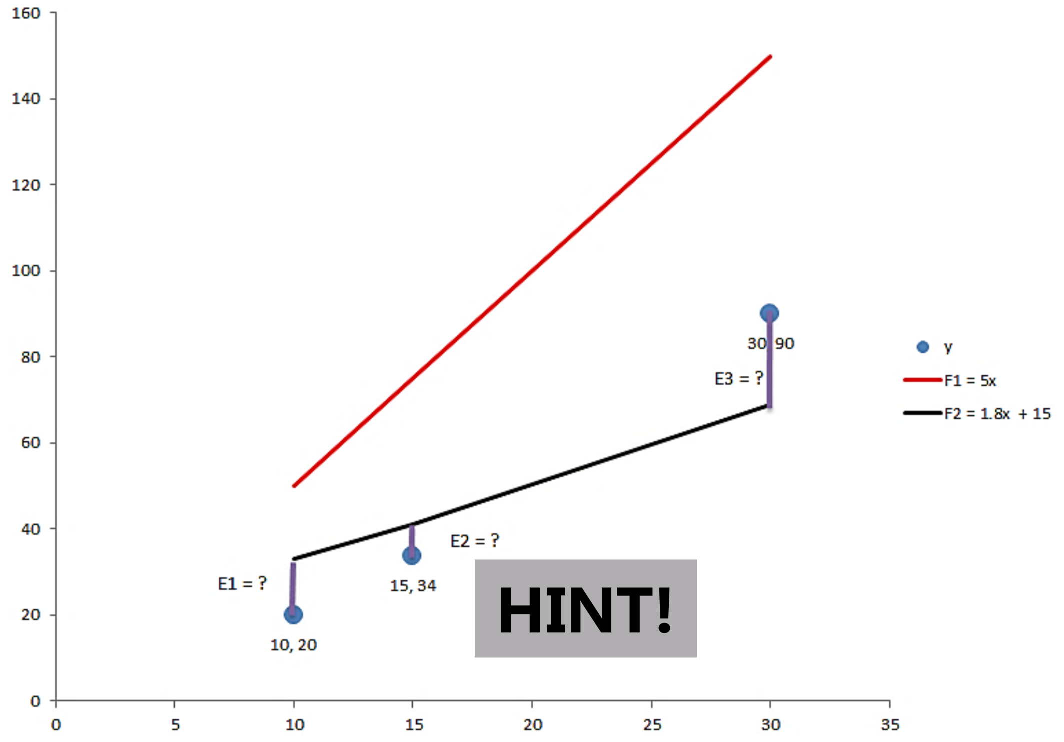
Quantify Error Made by an Algorithm in a Supervised Learning Problem with a Continuous Response

Example – Error for an Algorithm Can be Manually Calculated

Exercise

Calculate the total squared error made by the line F2

x	y	F2 = 1.8x + 15
10	20	33
15	34	41
30	90	69



Calculating Errors

How is it Done?

There can be infinitely many possible straight lines that can summarize a given data set but all of them are not equally good

Calculating Errors

How is it Done?

So how do we find out the best possible line?

Mathematics

Calculating Errors

How is it Done?

Ordinary Least Squares

Based on some simple linear algebra and directly gives the best line

Mathematics

Calculating Errors

How is it Done?

More of an iterative procedure, based out of traditional numerical analysis

Gradient Descent

Mathematics

Calculating Errors

How is it Done?

For those getting started, it is not required to go through the mathematics

Most software, designed for Data Science can directly give the results for linear regression

Calculating Errors

How is it Done?

You can search the internet to understand the computational aspects of Ordinary Least Squares or Gradient Descent

Recap

Types of Tasks, Machine Learning Algorithms and Linear Regression

Types of Tasks

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Supervised Machine Learning Algorithm

Exercise

Calculating Errors

Next

Using Scikit Learn for Machine Learning