FIT1043 Introduction to Data Science

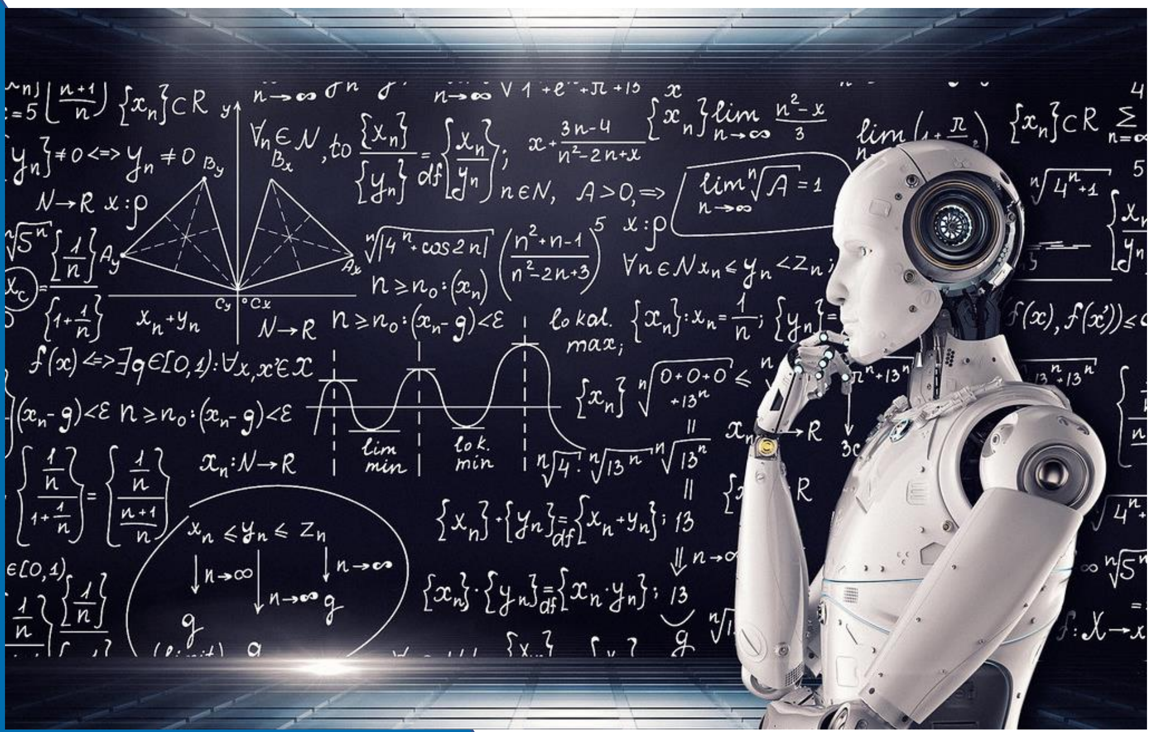
Week 5

Ian K T Tan

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Introduction to Machine Learning



MONASH University

Introduction to Machine Learning

What is Machine Learning?

- From Wikipedia: "…is the scientific study of **algorithms** and **statistical** models that computer systems use to perform a specific task without using explicit instructions, relying on **patterns** and **inference** instead."

- From Emerj: "Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in **autonomous** fashion, by feeding them **data** and information in the form of **observations** and real-world interactions.

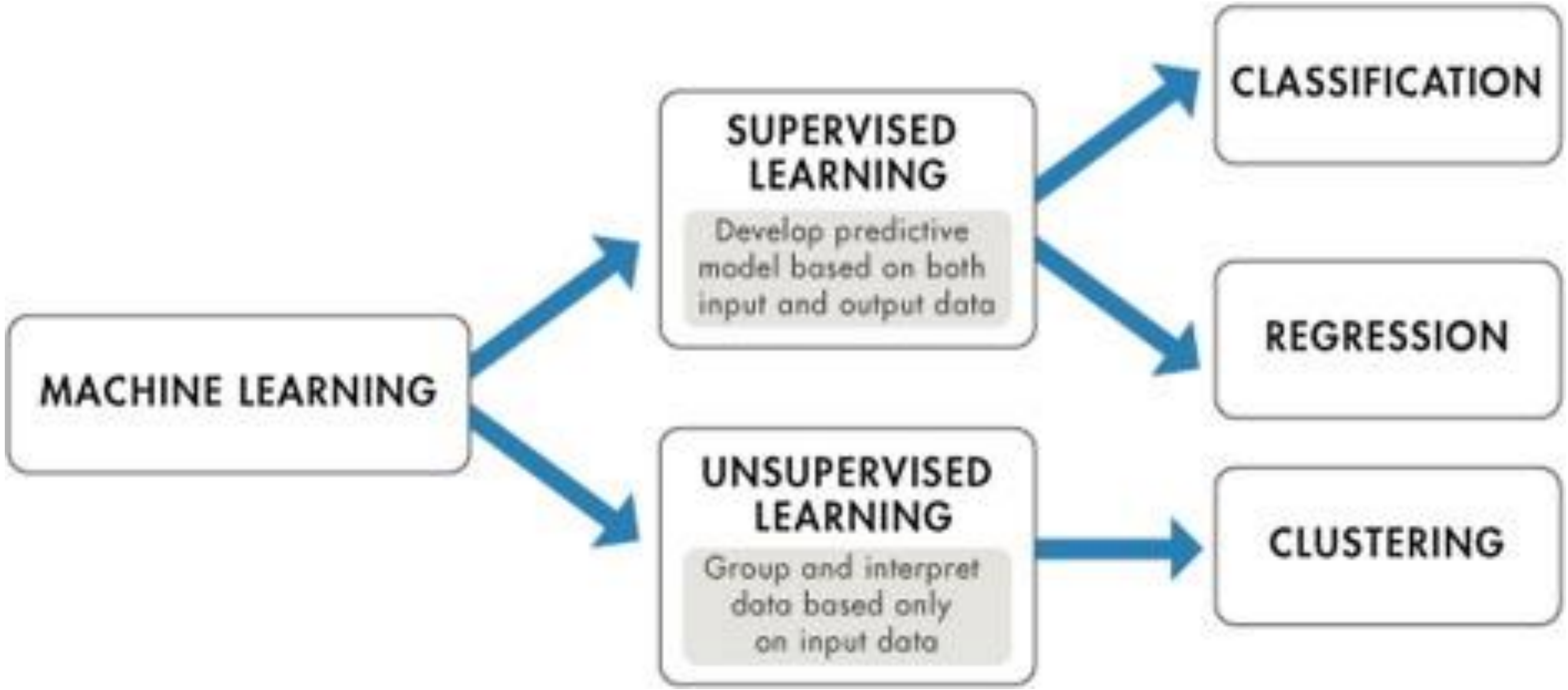
Introduction to Machine Learning

How to develop a Machine Learning model?

- Choose a measure of success
- Setting an evaluation protocol
- Developing a Benchmark Model
- Developing a Better Model and tuning its Hyperparameters

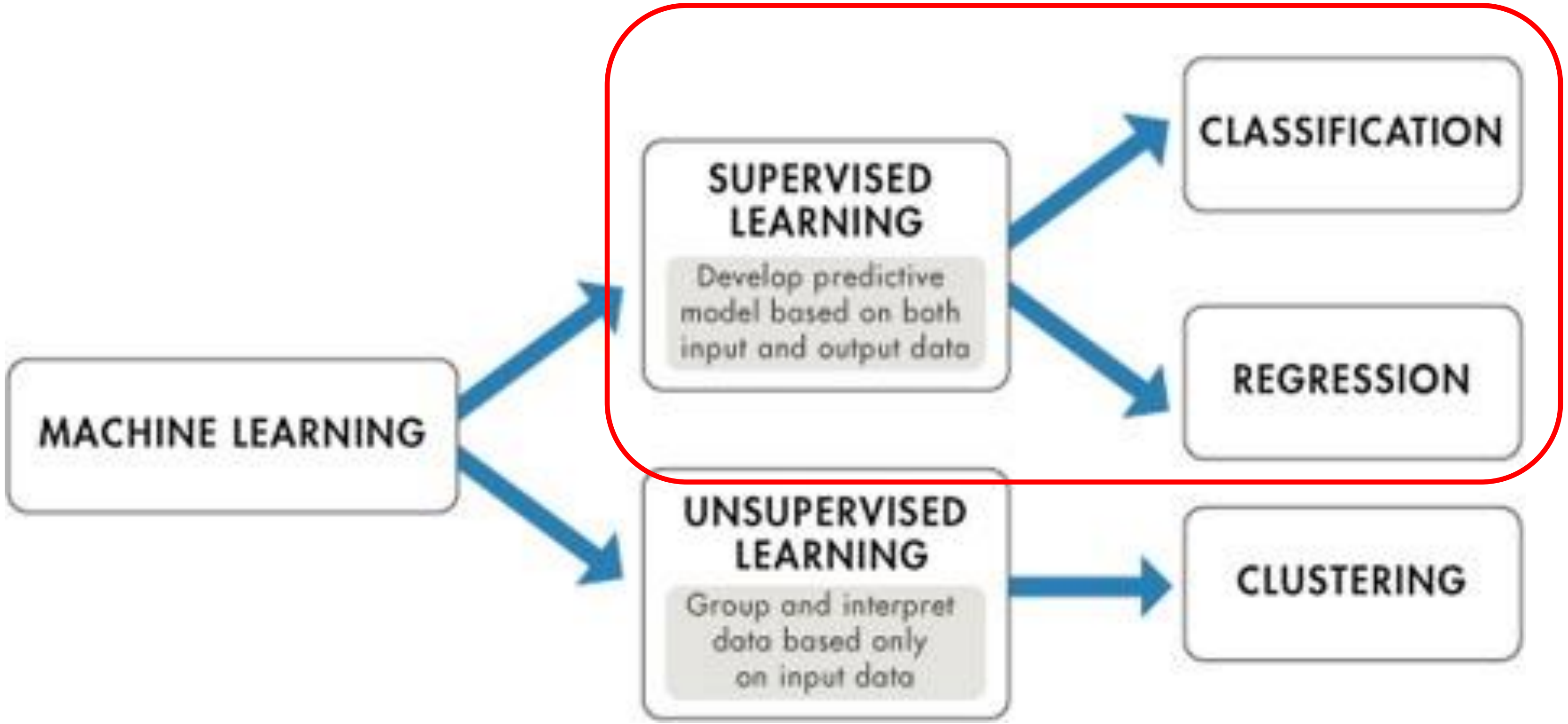
Learning Styles in ML Algorithms

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms



Learning Styles in ML Algorithms

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms



Learning Styles: *Supervised* ML

Brownlee J. (2016)

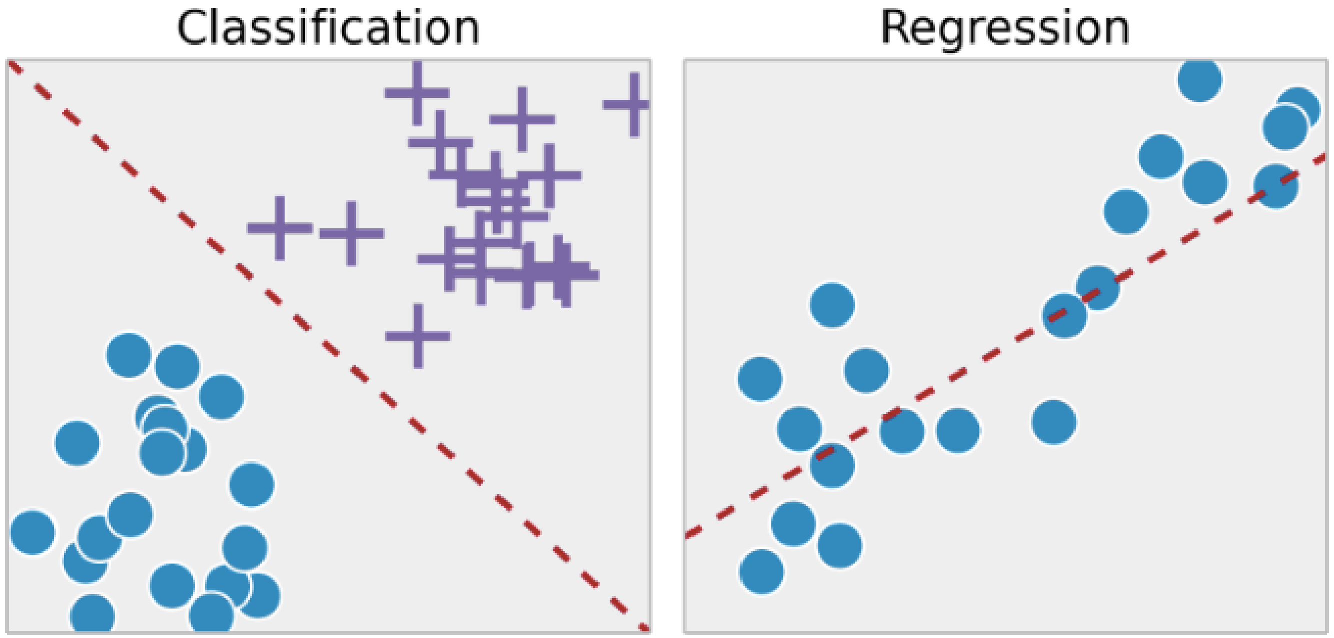
- All data is **labelled** and the algorithms learn to **predict** (infer) the output from the input data.
    - The goal is to approximate the mapping function so well that when you have new input data (x), you can predict the output variable (y) for that data.

- Example Problems:
    - Classification: The output variable is a category (e.g. “Red” or “Blue” for the Fish Classification)
    - Regression: The output variable is a real value (e.g. “dollars” or “weight”)

- Example Algorithms:
    - *Linear regression* for regression problems.
    - *Random forest (RF)* for classification and regression problems.
    - *Support vector machines (SVM)* for classification problems.

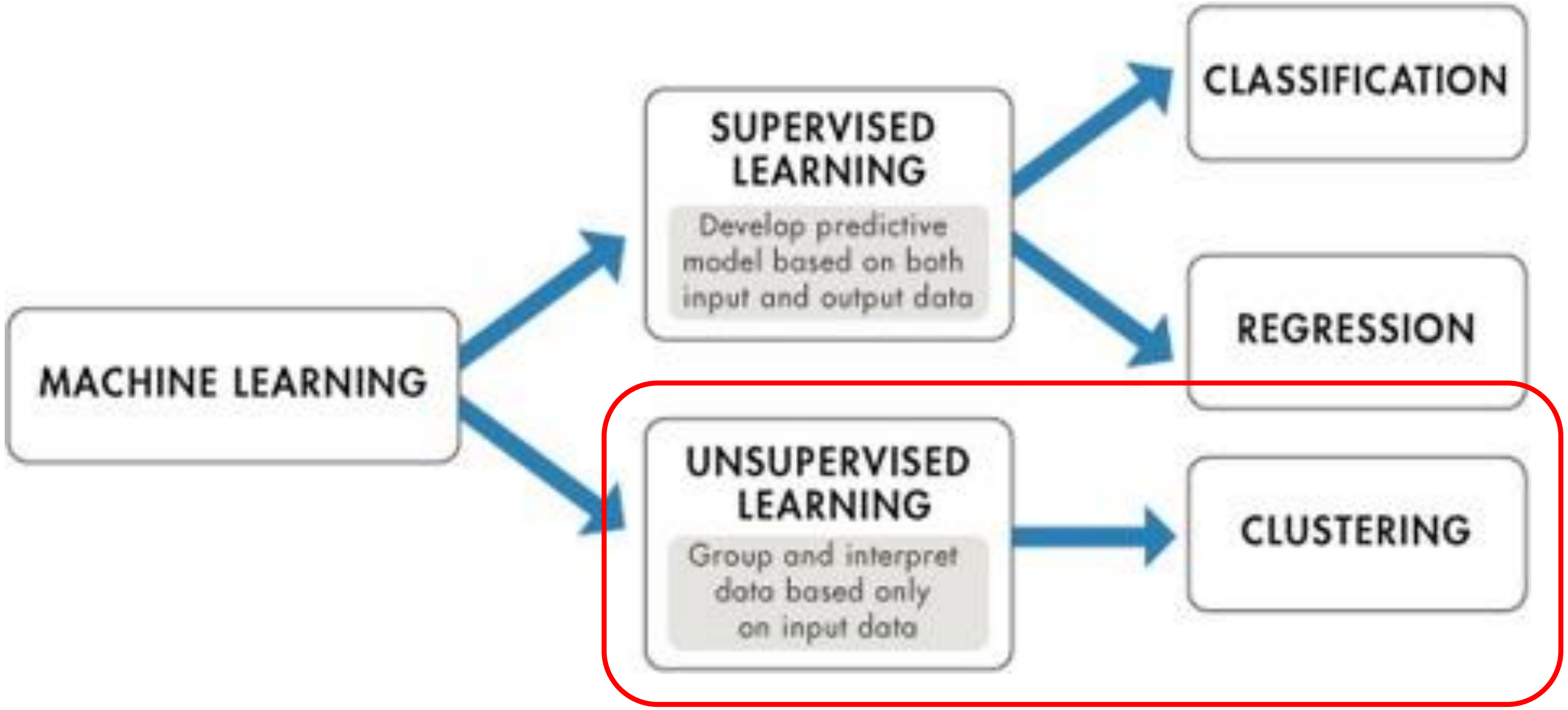
Learning Styles: Supervised ML

Brownlee J. (2016)



Learning Styles in ML Algorithms

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms



Learning Styles: *Unsupervised* ML

Brownlee J. (2016)

- All data is **unlabelled** and the algorithms learn to inherent structure from the input data.
- The goal is to model the underlying structure or distribution in the data in order to learn more about the data.

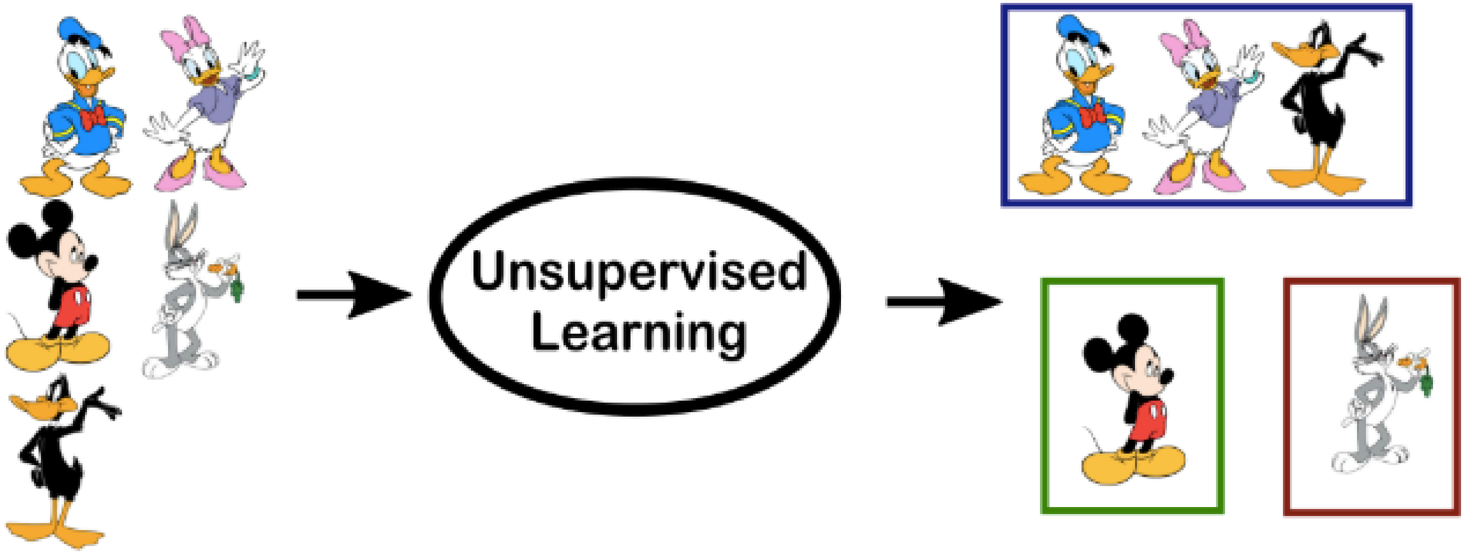
- Example Problems: face similarity detection
  - Clustering: Discover the inherent groupings in the data (e.g. grouping customers by purchasing behaviour)
  - Association: Discover rules that describe large portions of your data (e.g. people that buy X also tend to buy Y)

- Example Algorithms:
  - *k-means* for clustering problems.
  - *Apriori* algorithm for association rule learning problems.

MONASH University

Learning Styles: *Unsupervised* ML

Brownlee J. (2016)



Theory of Data Analysis

Introduction to Learning Theory

MONASH University

What is Learning Theory?

From Wikipedia:

(Computational) learning theory is a subfield of Artificial Intelligence devoted to ***studying the design and analysis of machine learning***.

Truth

Heart Disease Diagnosis

- For a single patient the “truth” can be measured directly
- How can you measure the “true” model?
    - Collect infinite data
    - But even if you can, it is a dynamic problem

Quality

To evaluate the quality of results derived from learning, we need notions of **value,** we will review quality and value using William Tell's Apple Shot (a Swiss folklore).

- William Tell forced to shoot the apple on his son's head

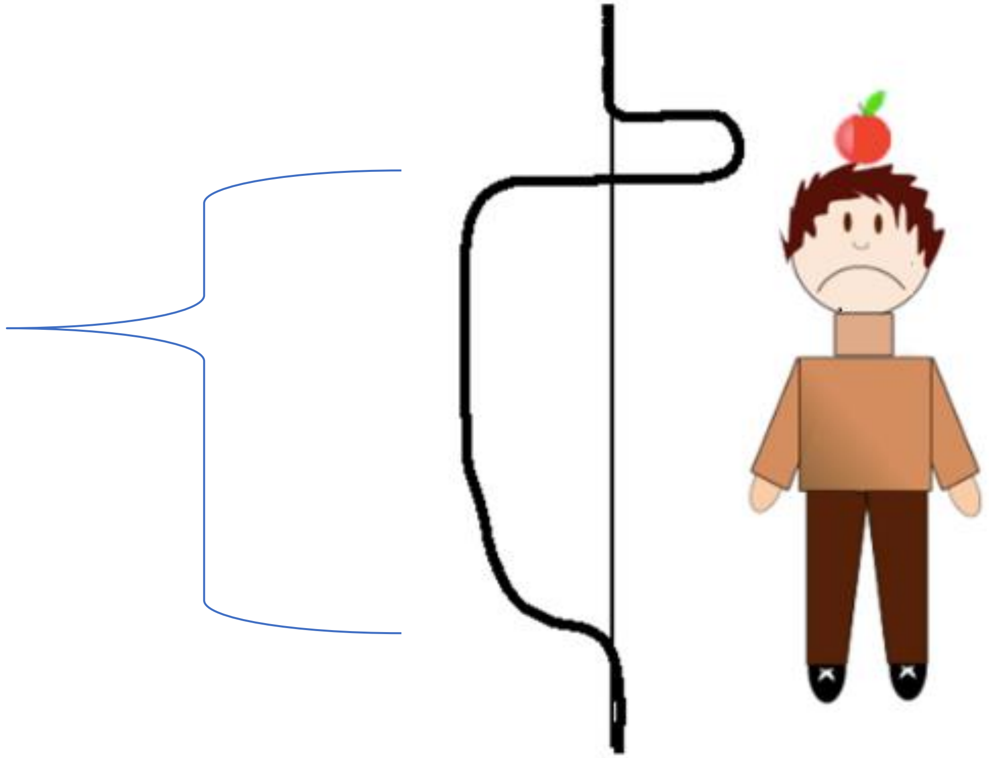
- If he strikes it, he gets both their freedom

William Tell's Apple Shot

This shows "value" as a function of height.

- Loss varies depending on where it strikes

- How do you compare loss of life versus gain of freedom?



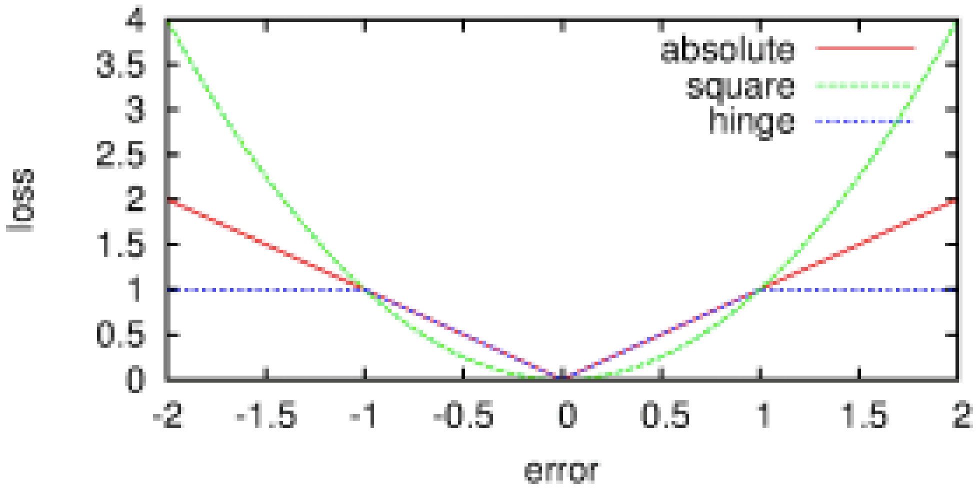
Quality

May be the consequence of your actions (making a prediction is a kind of action)

Can be measured on a positive or negative scale

Loss: positive when things are bad, negative (or zero) when they're good
Gain: positive when things are good, negative when they're not
Error: measure of "miss", sometimes a distance, but not a measure of quality

Quality is a Function of Error



Error measures the distance between the prediction and the actual value, where “0” means no error, prediction was exactly right.

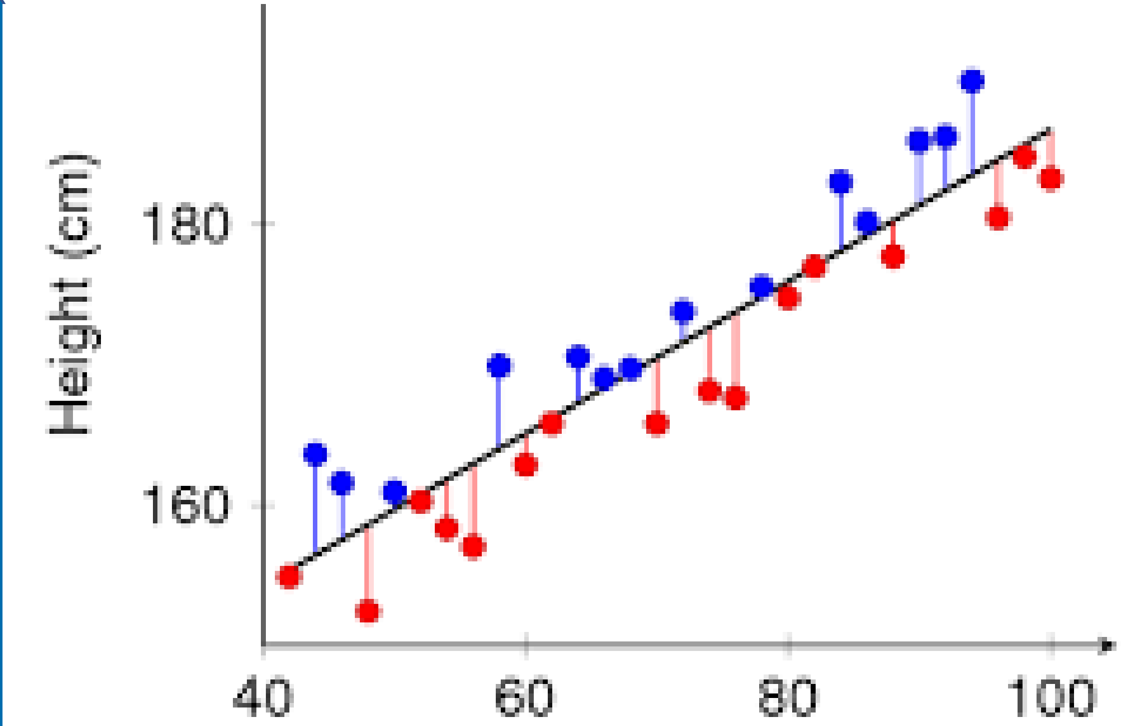
We can convert error to a measure of quality using a loss function, e.g.:

absolute-error(x) = $| x |$

square-error(x) = $x * x$

hinge-error(x) = $| x |$, if $| x | \leq 1$; 1 otherwise

Regression



Regression

$ $$ $$$ $$$$

?

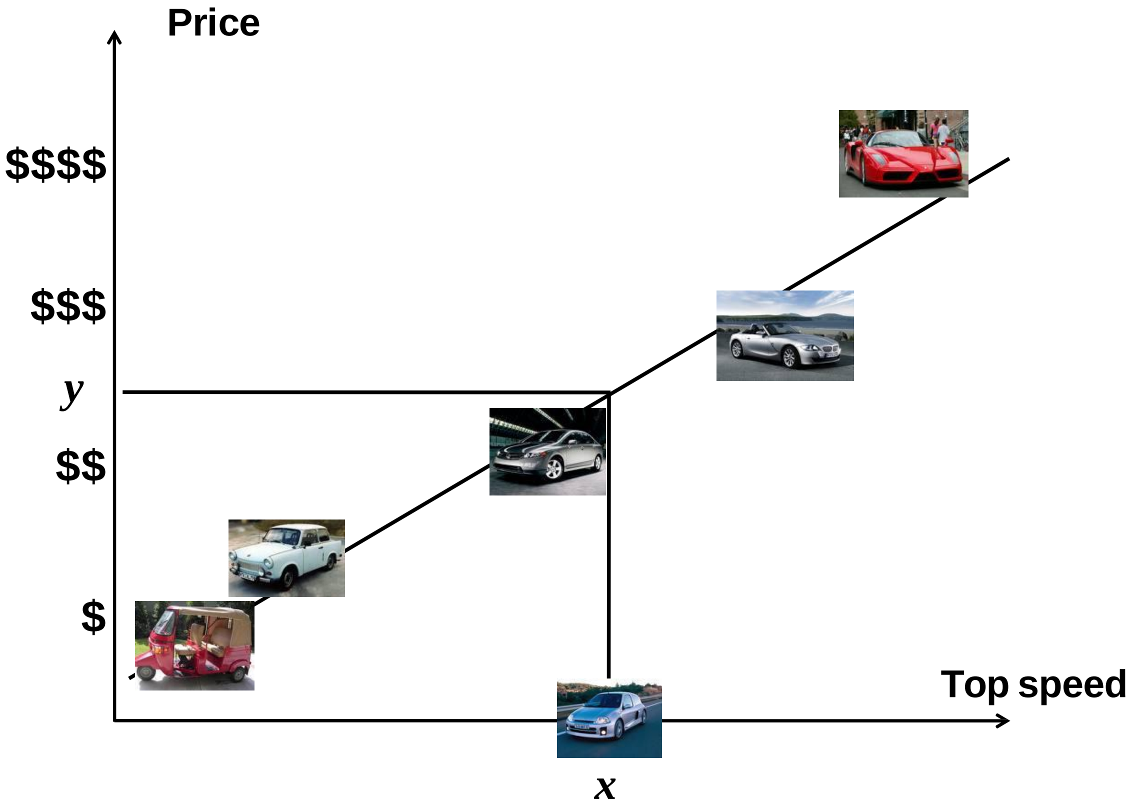
Regression

$ $$ $$$ $$$$

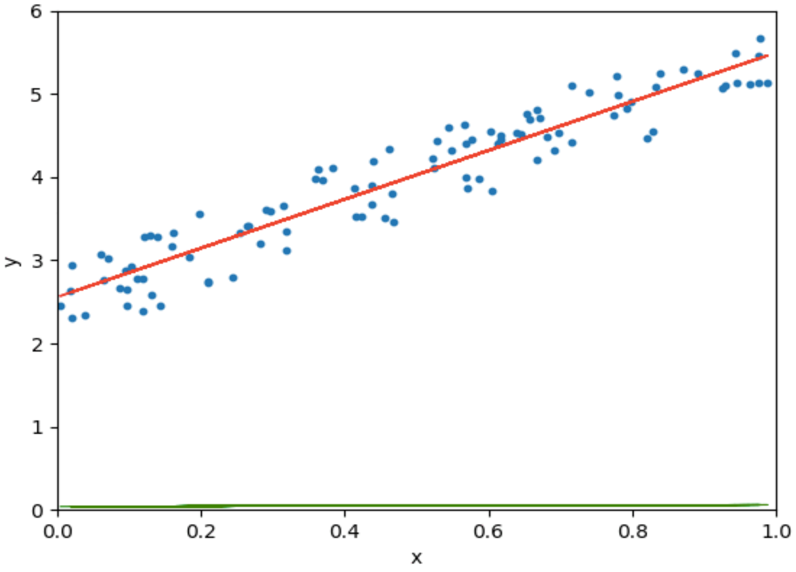
?

How much is this car worth?

Regression



Linear Regression



Regression fits a very simple equation to the data:

$$\hat{y}(x; \vec{a}) = a_0 + a_1 x$$

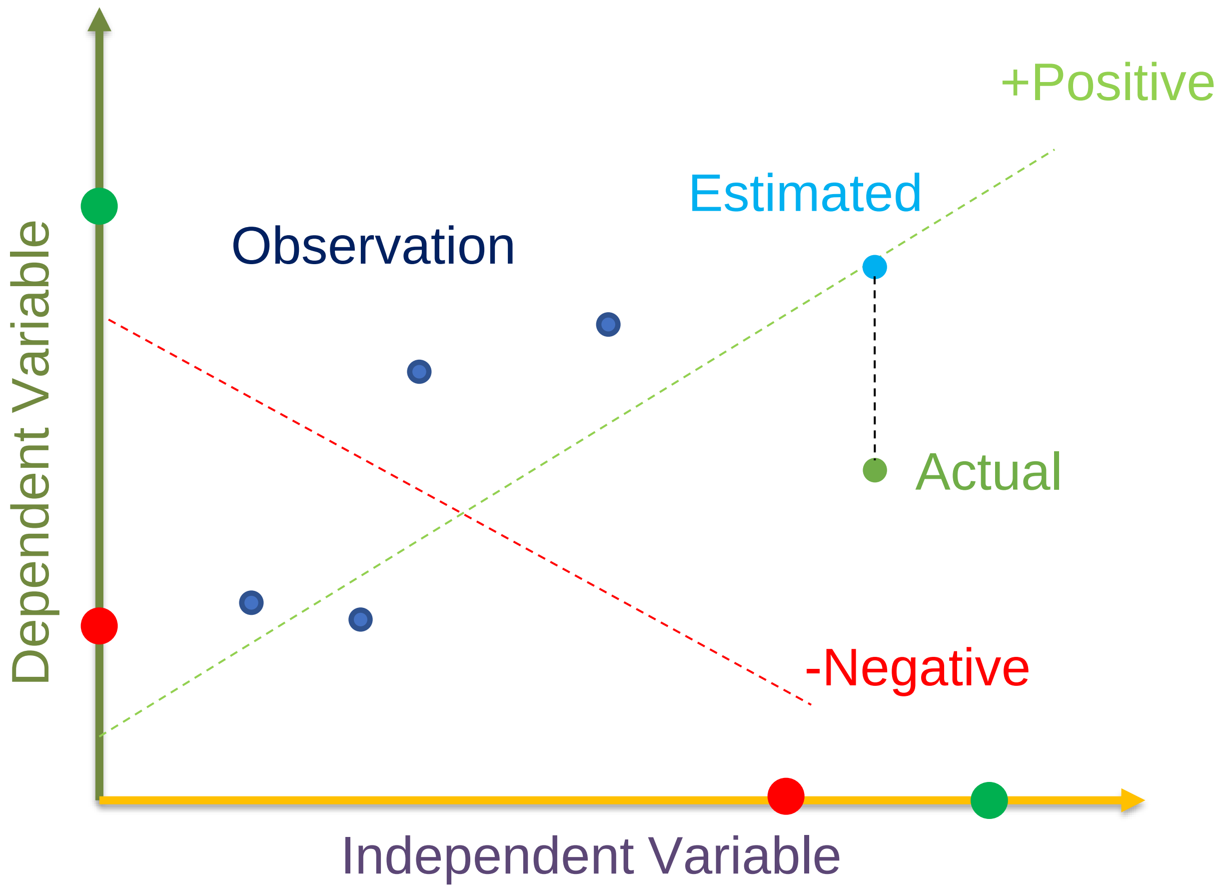
Data is shown with blue dots, red line is the "linear fitted model"

Here $\hat{y}(x; \vec{a})$ is the for prediction for y at the point x using the model parameters (a_0, a_1), i.e. the intercept and slope terms.

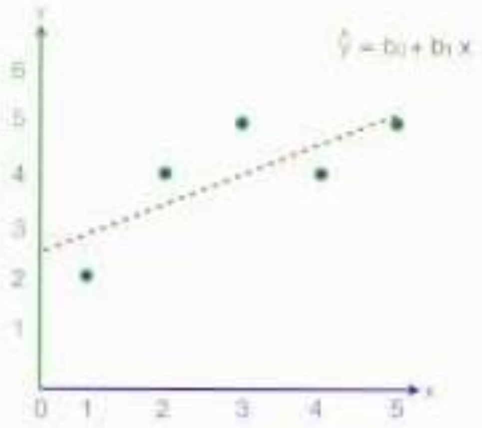
Given some data pairs (x_1, y_1), ...,(x_N , y_N), we fit a model by finding the vector $\vec{a}$ that minimises the loss function:

$$\text{mean square error} = MSE_{train} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}(x_i; \vec{a}) - y_i\right)^2$$

How to Calculate Linear Regression



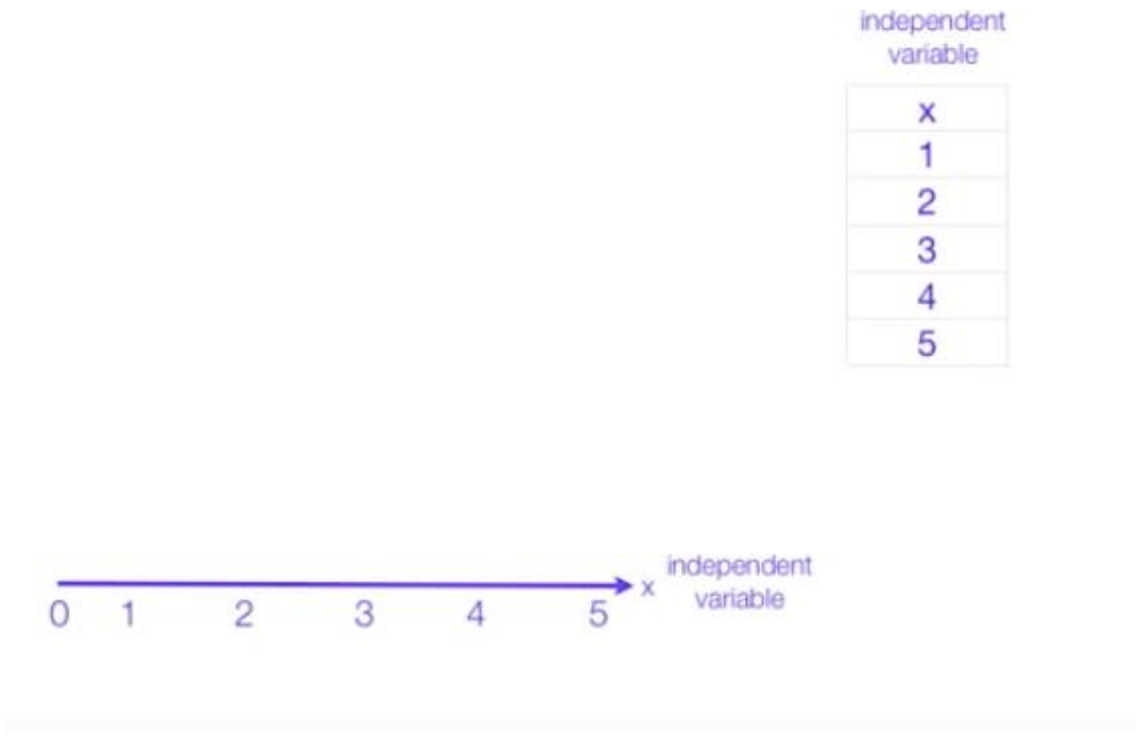
How to Calculate Linear Regression



	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	-1
	3	5	0	1	0	
	4	4	1	0	1	
	5	5	2	1	4	
mean	3	4				

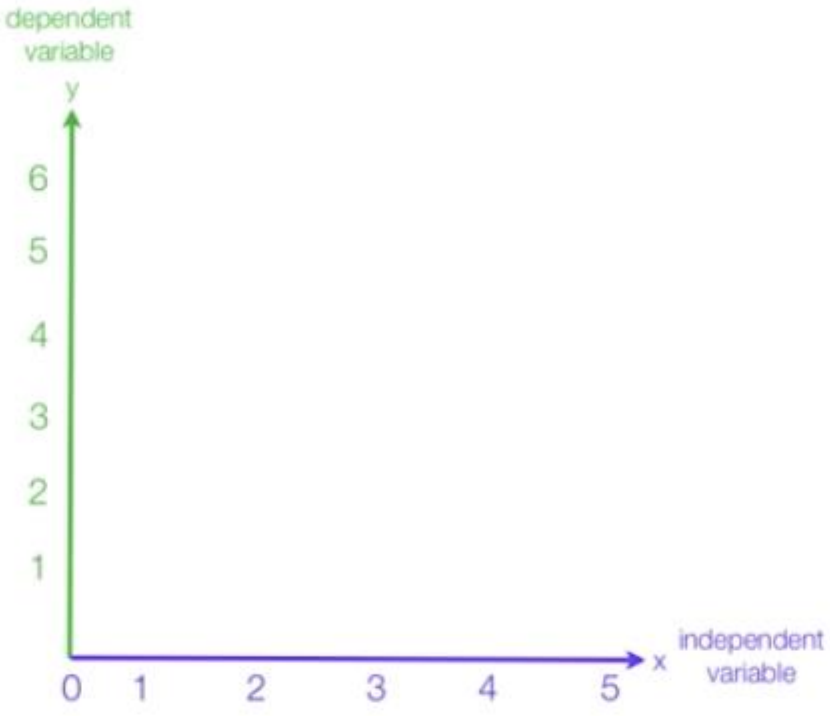
https://www.youtube.com/watch?v=JvS2triCgOY

How to Calculate Linear Regression



https://www.youtube.com/watch?v=JvS2triCgOY

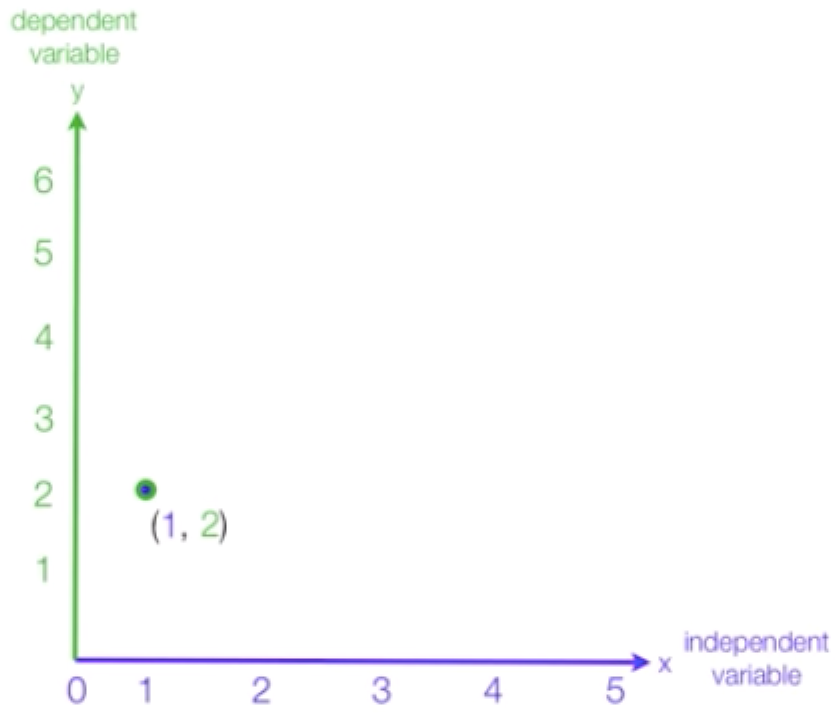
How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5

https://www.youtube.com/watch?v=JvS2triCgOY

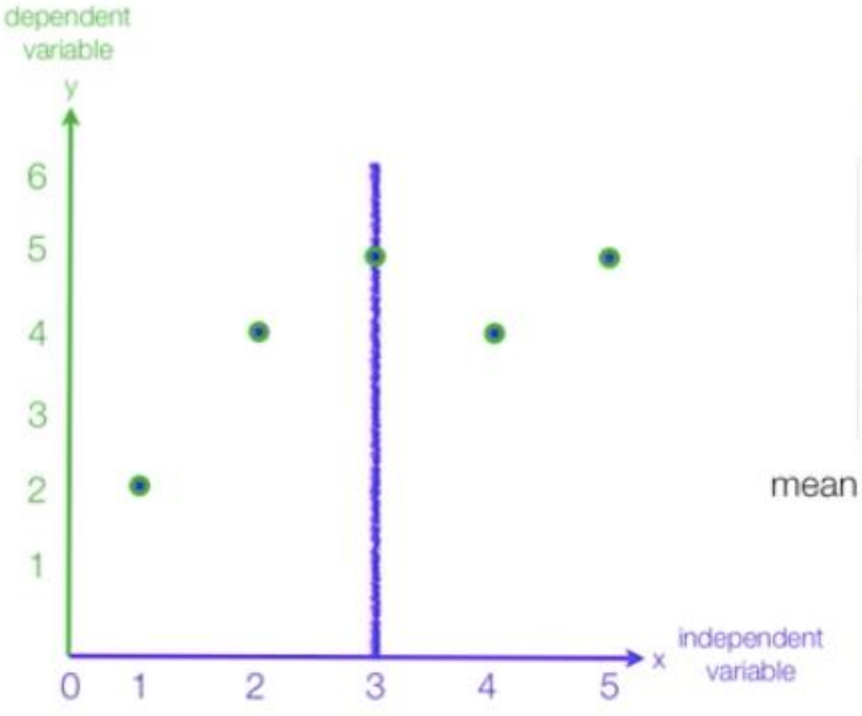
How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5

https://www.youtube.com/watch?v=JvS2triCgOY

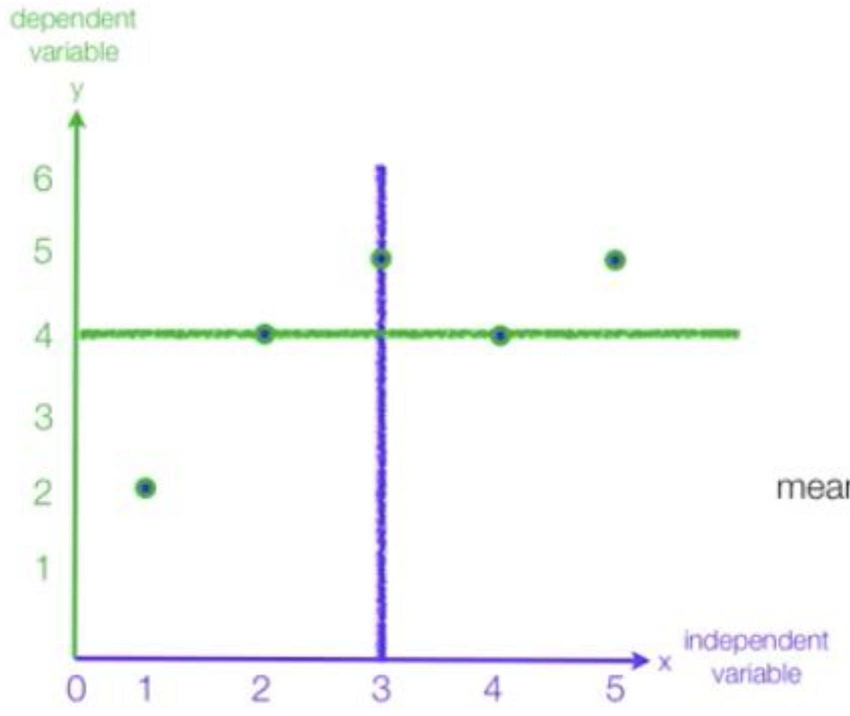
How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5
mean 3	

https://www.youtube.com/watch?v=JvS2triCgOY

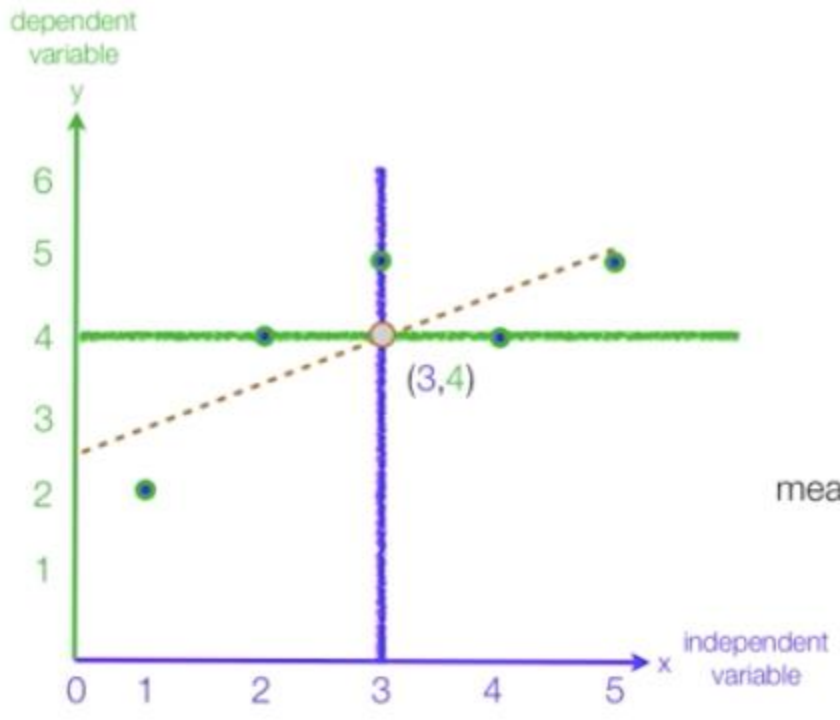
How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5
mean 3	4

https://www.youtube.com/watch?v=JvS2triCgOY

How to Calculate Linear Regression

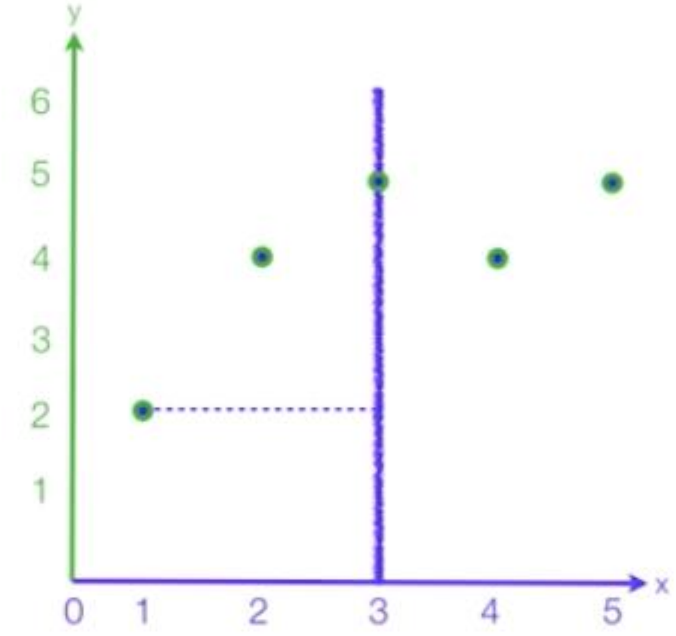


independent variable x	dependent variable y
1	2
2	4
3	5
4	4
5	5

mean

https://www.youtube.com/watch?v=JvS2triCgOY

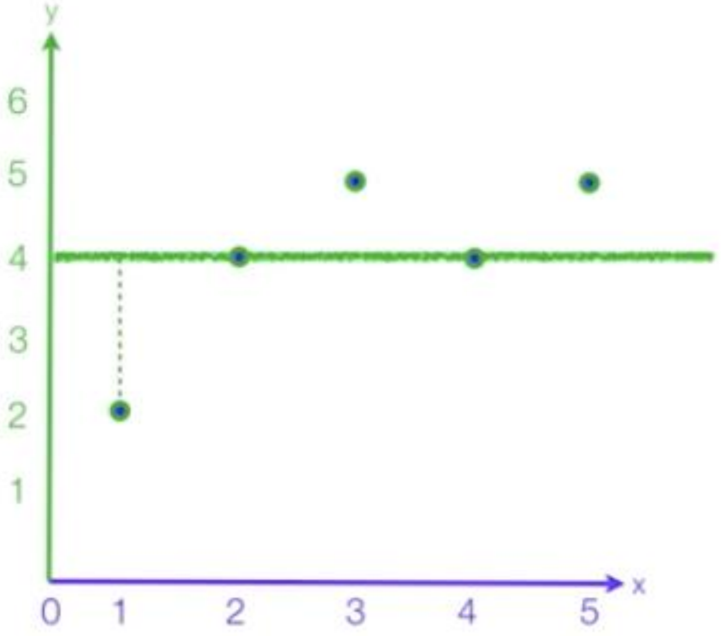
How to Calculate Linear Regression



x	y	$x - \bar{x}$	
1	2	1 - 3	
2	4		
3	5		
4	4		
5	5		
mean 3	4		

https://www.youtube.com/watch?v=JvS2triCgOY

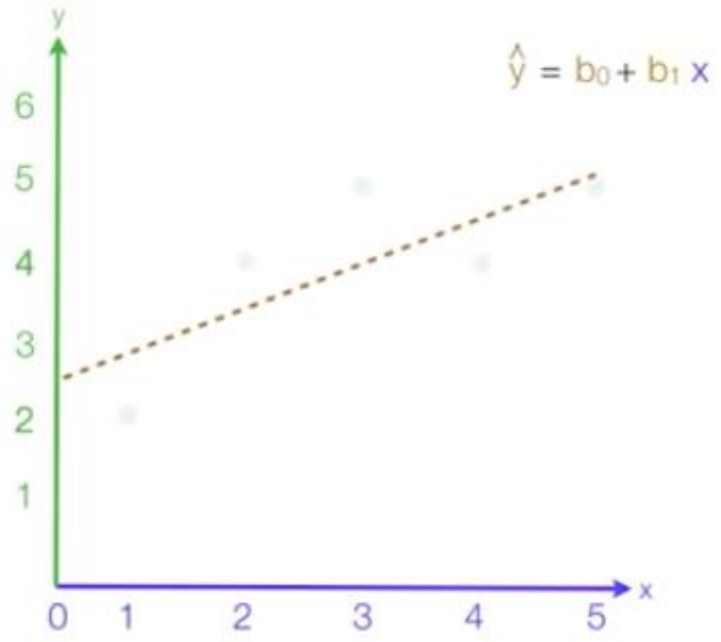
How to Calculate Linear Regression



	x	y	$x - \bar{x}$	$y - \bar{y}$
	1	2	-2	2 - 4
	2	4	-1	
	3	5	0	
	4	4	1	
	5	5	2	
mean	3	4		

https://www.youtube.com/watch?v=JvS2triCgOY

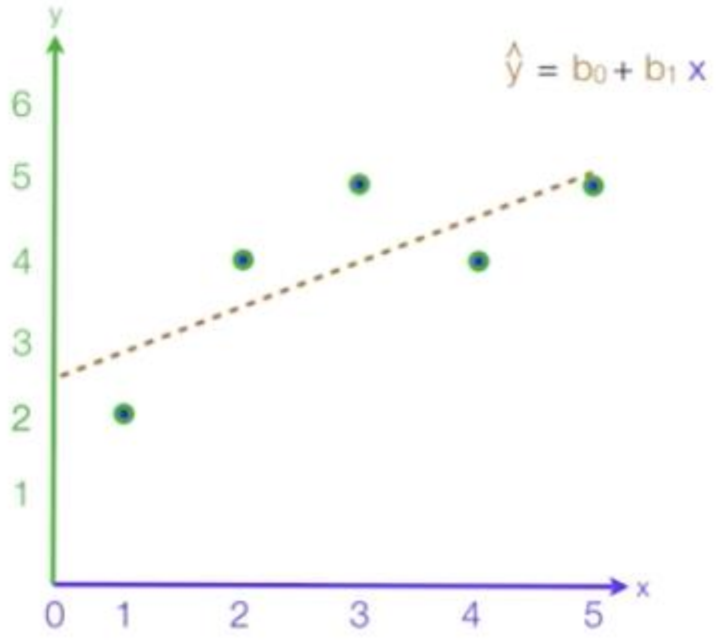
How to Calculate Linear Regression



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2

https://www.youtube.com/watch?v=JvS2triCgOY

How to Calculate Linear Regression



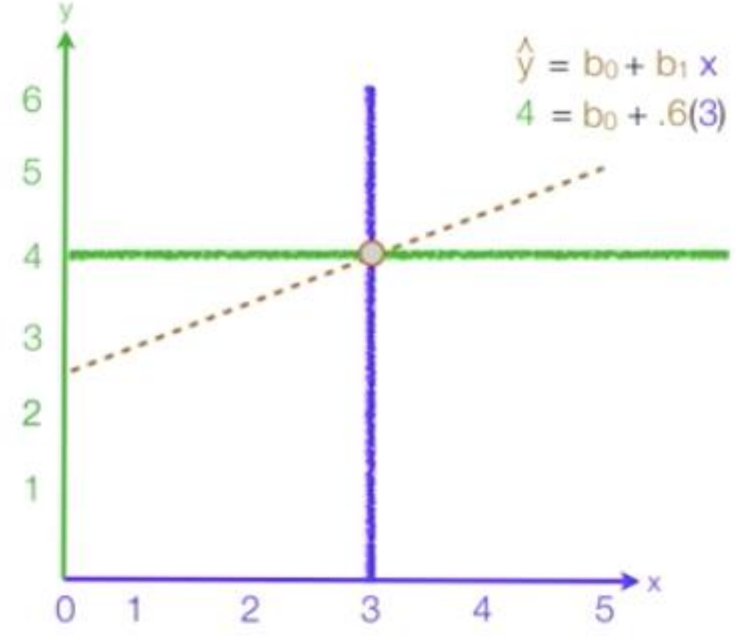
	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
					10	6
mean	3	4				

sum

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

https://www.youtube.com/watch?v=JvS2triCgOY

How to Calculate Linear Regression



	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
mean	3	4			10	6

$$4 = b_0 + .6(3)$$

$$4 = b_0 + 1.8$$

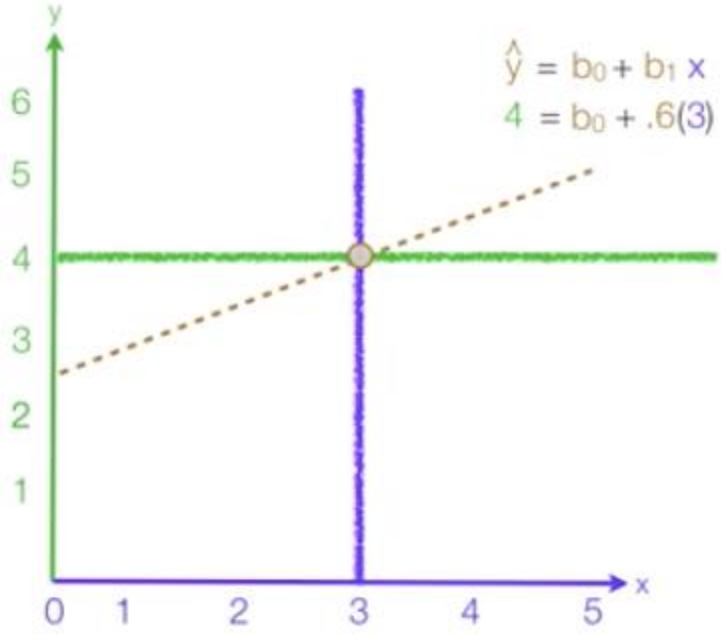
$$-1.8 \qquad -1.8$$

$$2.2 = b_0$$

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

https://www.youtube.com/watch?v=JvS2triCgOY

How to Calculate Linear Regression



$b_0 = 2.2$

$b_1 = .6$

$\hat{y} = 2.2 + .6x$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
mean 3	4			10	6

$$4 = b_0 + .6(3)$$

$$4 = b_0 + \cancel{1.8}$$

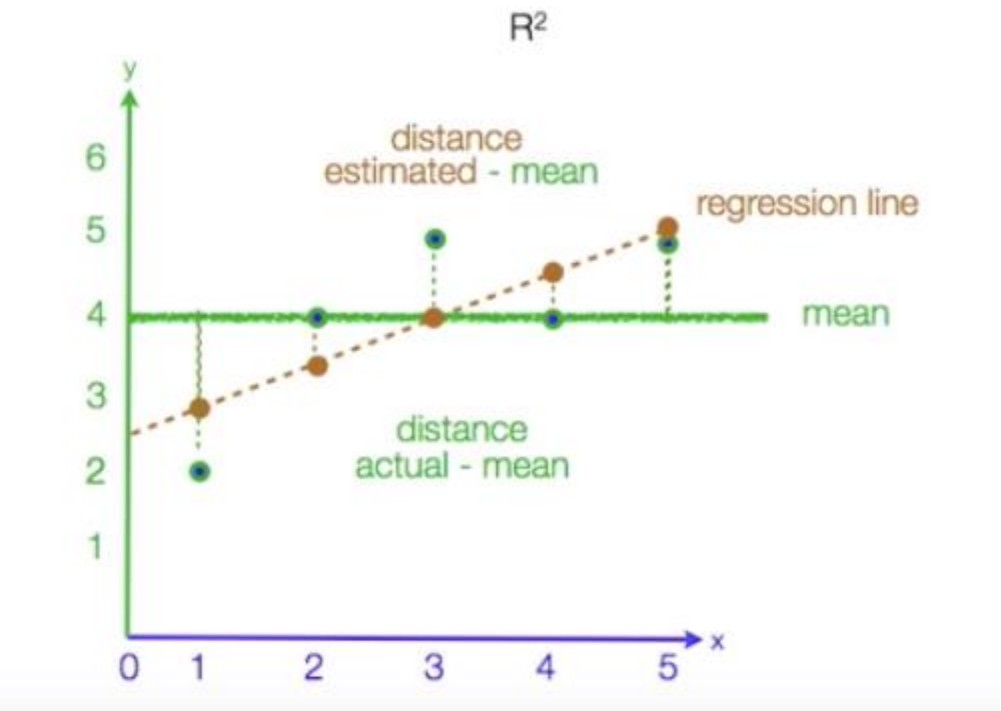
$$\underline{-1.8 \quad\quad -\cancel{1.8}}$$

$$2.2 = b_0$$

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

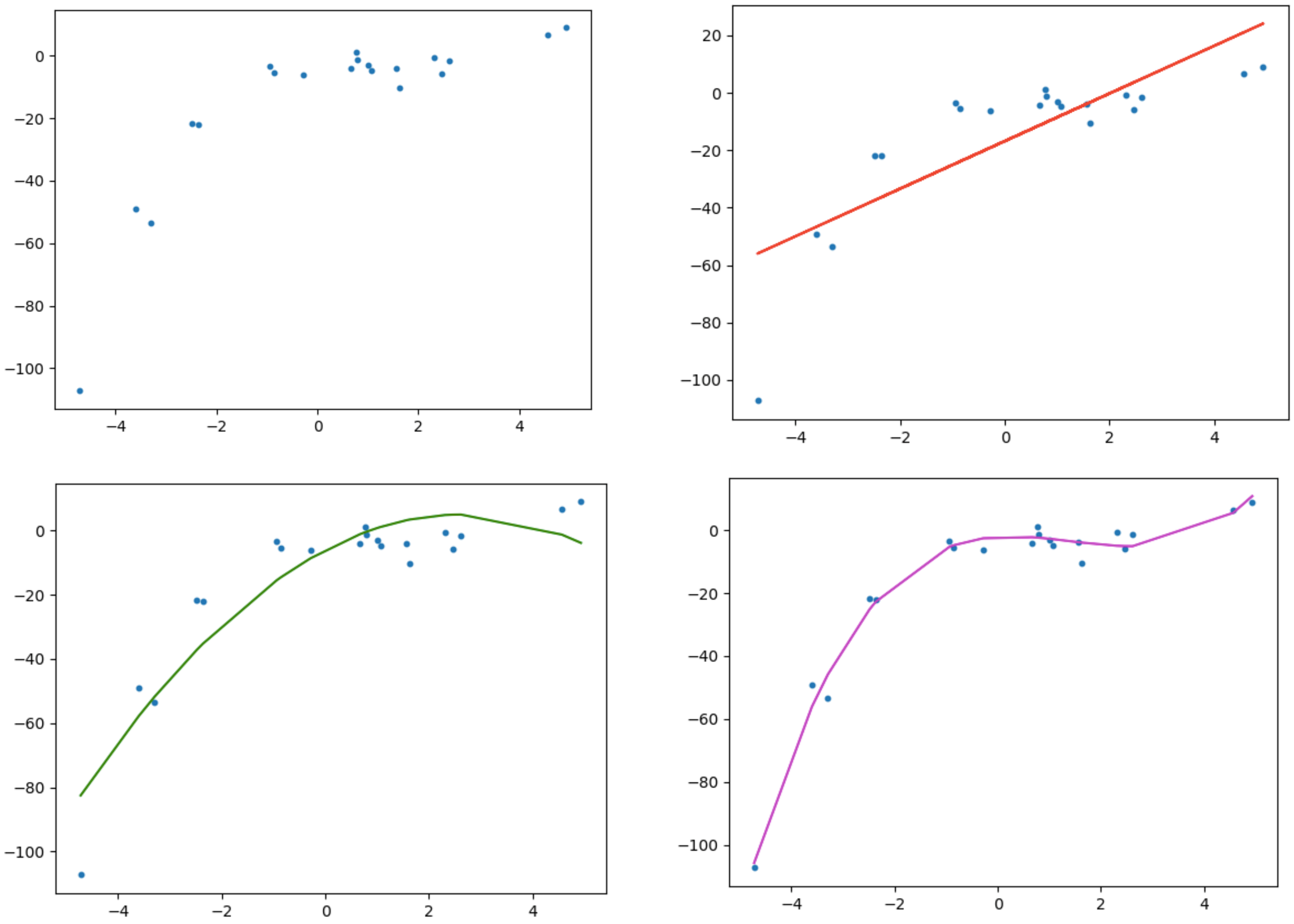
https://www.youtube.com/watch?v=JvS2triCgOY

How to Calculate Linear Regression



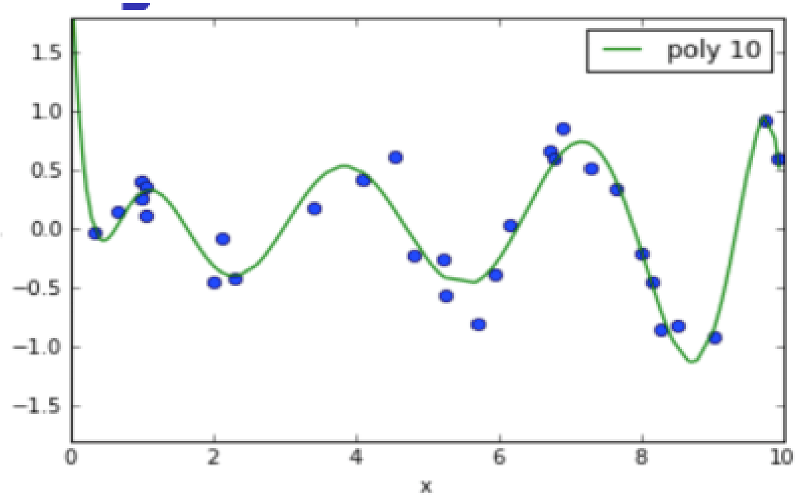
https://www.youtube.com/watch?v=JvS2triCgOY

Polynomial Regression



Polynomial Regression

- Data is shown with blue dots, green line is the "polynomial fitted model"



- **Polynomial regression** uses the same linear regression infrastructure to fit a higher order polynomial. In this case we fit a 10-th order polynomial:

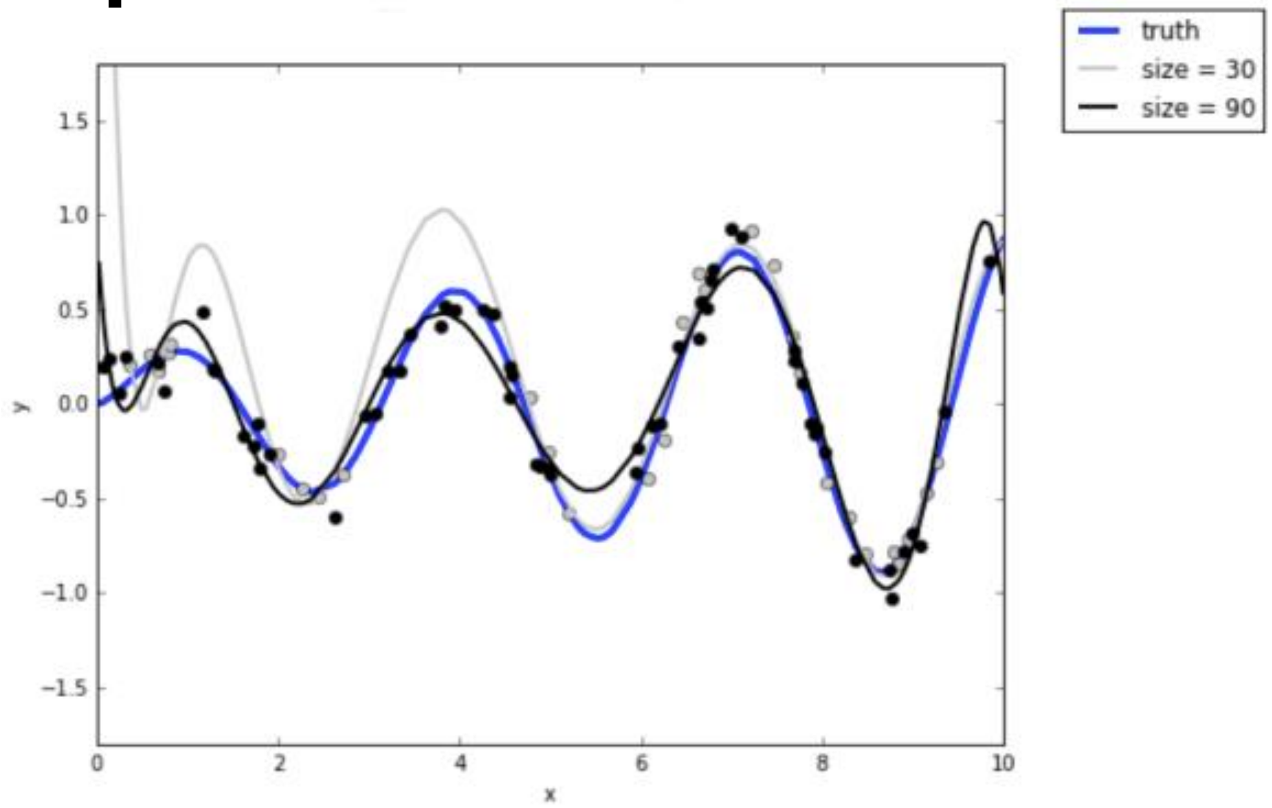
$$\hat{y}(x;\vec{a}) = a_0 + a_1 x + a_2 x^2 + \ldots a_9 x^9 + a_{10} x^{10} = \sum_{i=0}^{10} a_i x^i$$

- By finding the vector $\vec{a}$ that for a given set of data pairs (x_1, y_1), ...,(x_N, y_N) minimises the loss function:

$$\text{mean square error} = MSE_{train} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}(x_i;\vec{a}) - y_i\right)^2$$

MONASH University

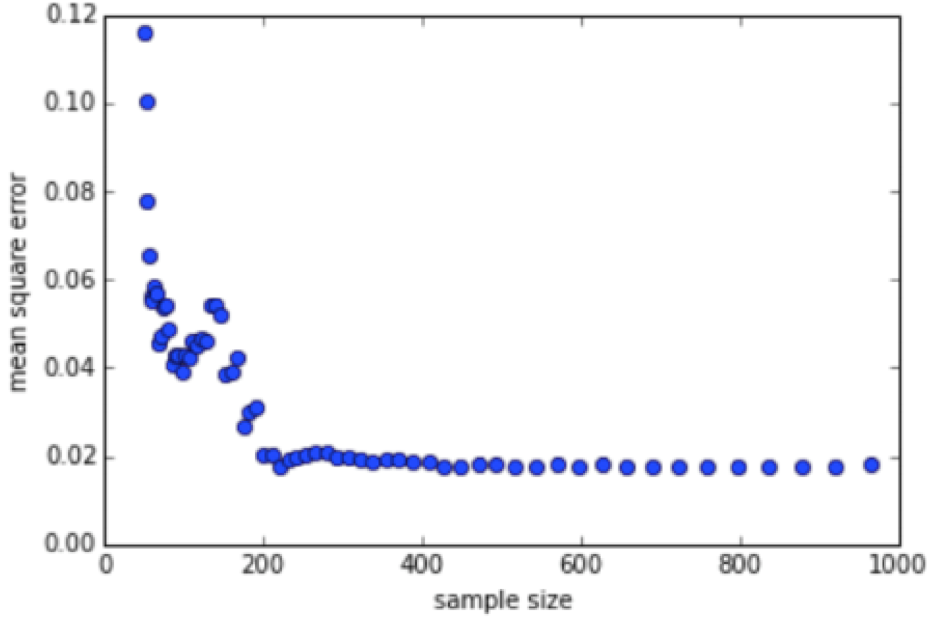
More Data Improves the Fit

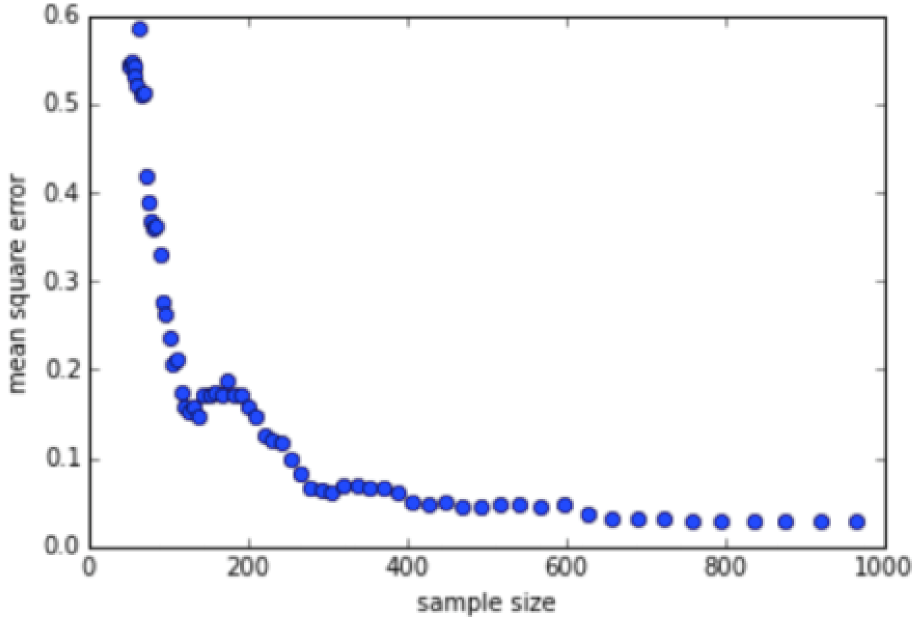


- Blue line is true model that generated the data (before noise was added).
- Grey curve is model fit to 30 data points
- Black curve is model fit to 90 data points

In general, more data means better fit

More Data Improves the Fit





MSE decreases as the amount of training data grows

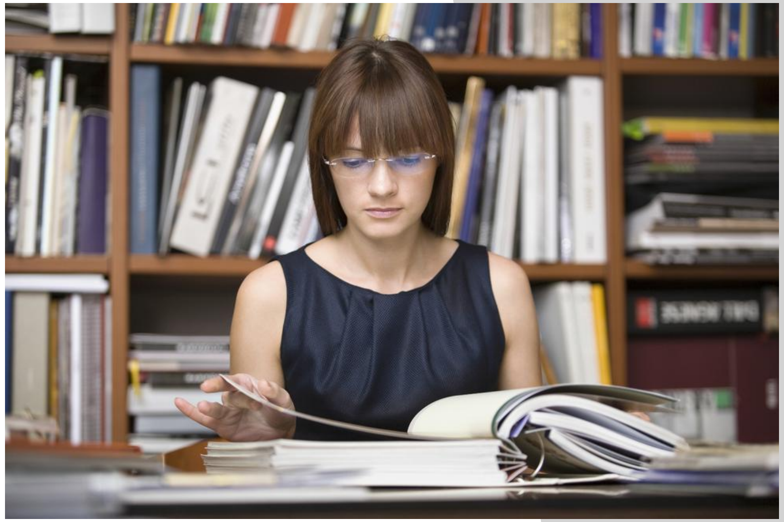
- These plots are called **learning curves**
- Different learning algorithms exhibit different behaviour (rate of decay)

Home Activities

Suggested Activities for the week

Videos

Watch David Longstreet's easy to follow video on regression and on Calculate R^2 using regression analysis.



Reading

Read pages 16 – 18 of Jason Brownlee, "Master Machine Learning Algorithms: discover how they work and implement them from scratch", 2016

MONASH University

Recap: Learning Outcomes

Week

By the end of this week you should be able to:

- Explain what are models and predictive models
- Analyse predictive models in different examples
- Understand how to evaluate predictive models
- **Analyse how to estimate linear regression model**
- Apply linear regression and polynomial regression on different data sets using Python