3.01 Linear Regression
3.02 Regression Metrics

What is Machine Learning?

Machine learning is the process of letting your **machine** use data to **learn** the relationship between some predictors (*x*-variables) and some response(s) (*y*-variables).

This is a fancy way of saying **statistical modeling**.

Ok then, so what is statistical modeling?

Statistical modeling is the process of combining data with **statistical theory** to **model** the real-world relationship between predictors (*x*-variables) and some response(s) (*y*-variables).

This is a more down-to-earth way of saying **machine learning**.

Two* Kinds of ML

In essence, all machine learning models fall into one of two categories:

- **Supervised learning** - Given X, can we predict Y?
- **Unsupervised learning** - What does X look like, *really?* There is no Y.

*There are more. Kinda. The big third category is **reinforcement learning**, which is a field still in the process of being invented.

Two Kinds of Supervised Learning

Supervised learning models fall into two different buckets:

Regression - this is when our *y*-variable is numeric.

- *“Given the past values of the stock price of Apple, what will tomorrow’s closing price be?”*
- *“Given the annual precipitation, average temperature, and soil pH, what will this year’s harvest yield be?”*
- *“Given the square footage, number of bedrooms, number of bathrooms, and quality of school district, what will the price of this home be?”*

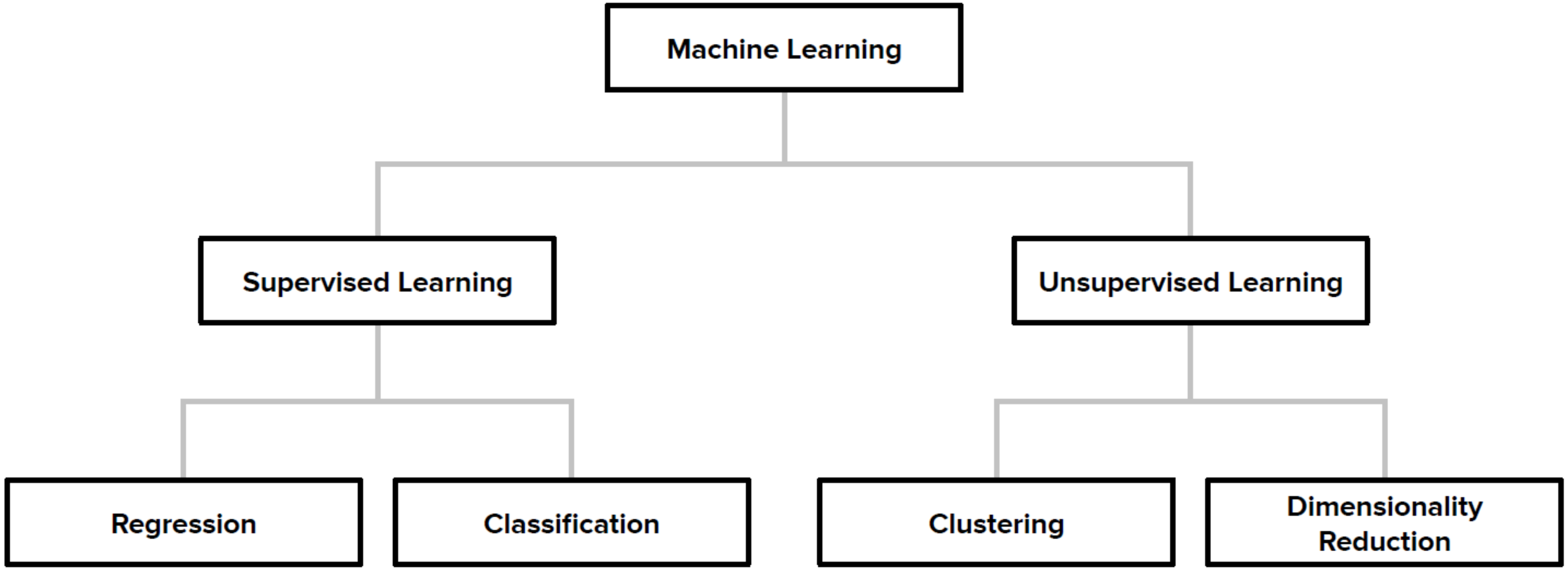
Two Kinds of Supervised Learning

Supervised learning models fall into two different buckets:

Classification - this is when our *y*-variable is a category. If it's a 0/1 yes/no kind of variable, we often call it **binary classification**. Otherwise, **multiclass classification**.

- *"Given this person's demographic information, how many tabs they have open, and where they live, will they make a purchase on my site?"*
- *"Given radar readouts, past weather, and almanac data, will it rain tomorrow?"*
- *"Given how many hours you study, how many hours you sleep, and your course load, will you pass the final exam?"*

Roadmap of ML



Machine Learning

Supervised Learning

Unsupervised Learning

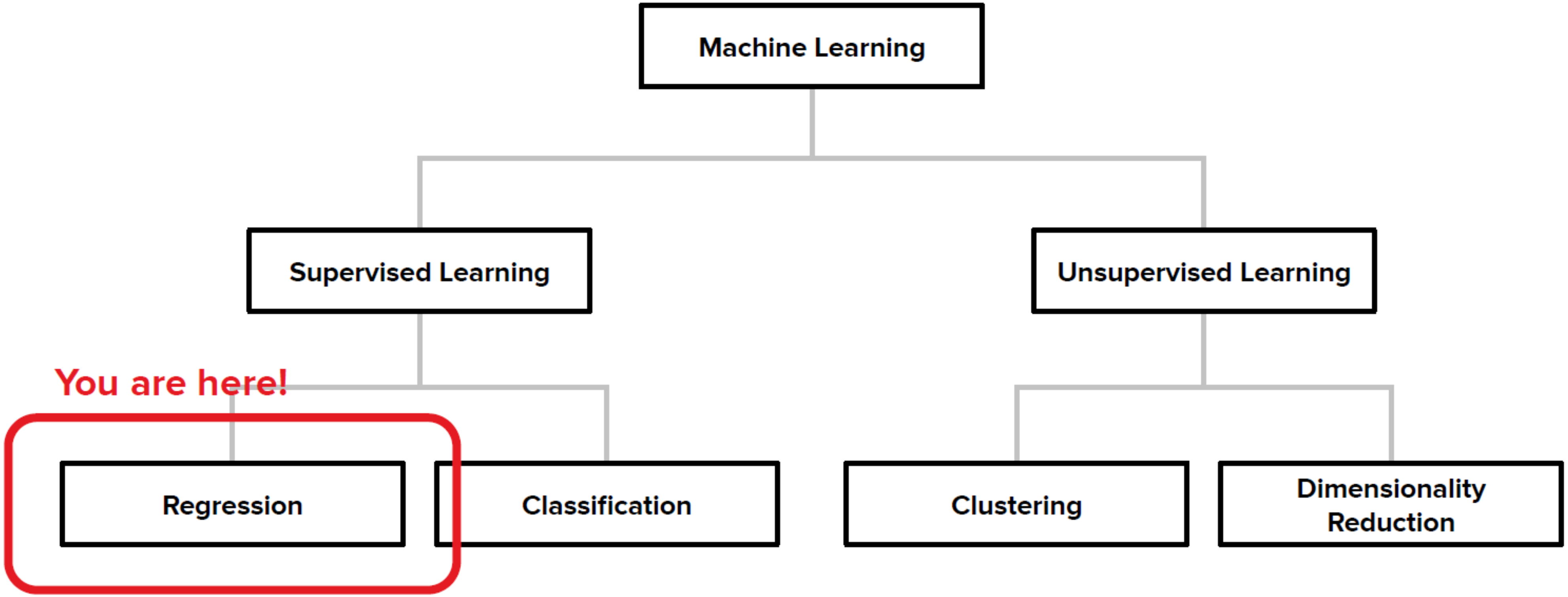
Regression

Classification

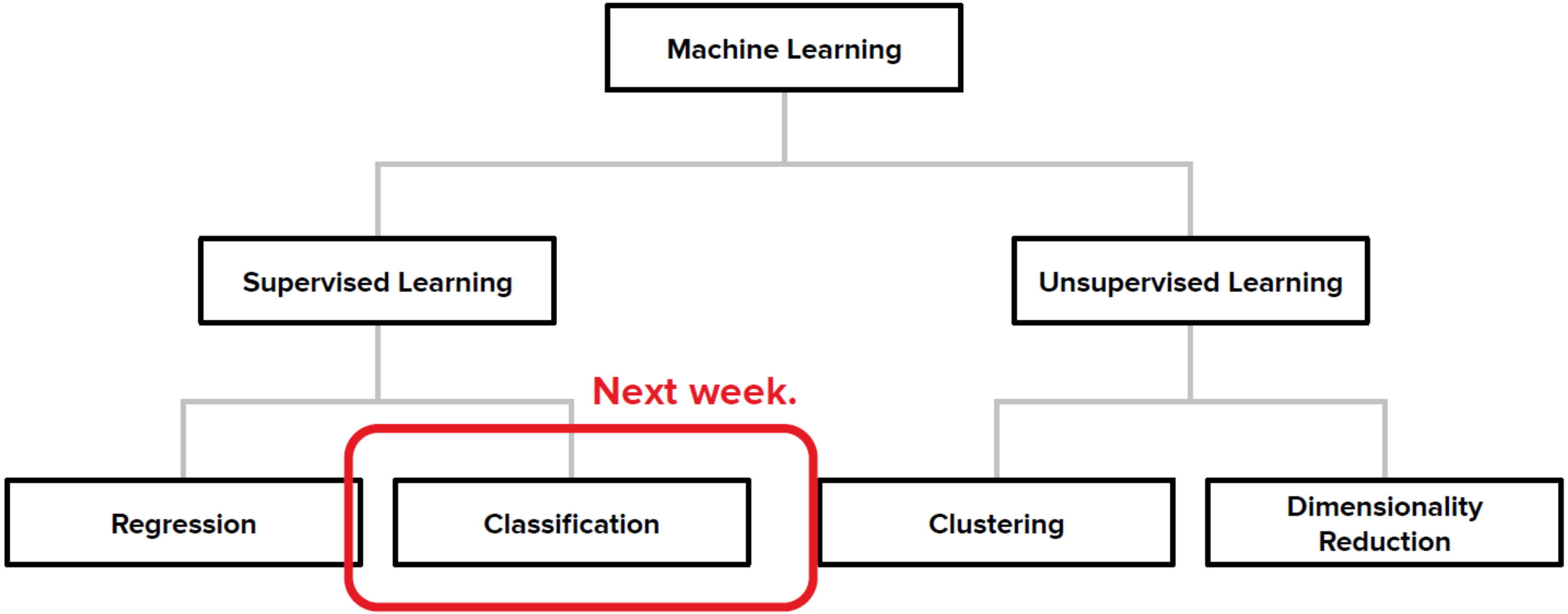
Clustering

Dimensionality Reduction

Roadmap of ML



Roadmap of ML


Machine Learning

Supervised Learning

Unsupervised Learning

Next week.

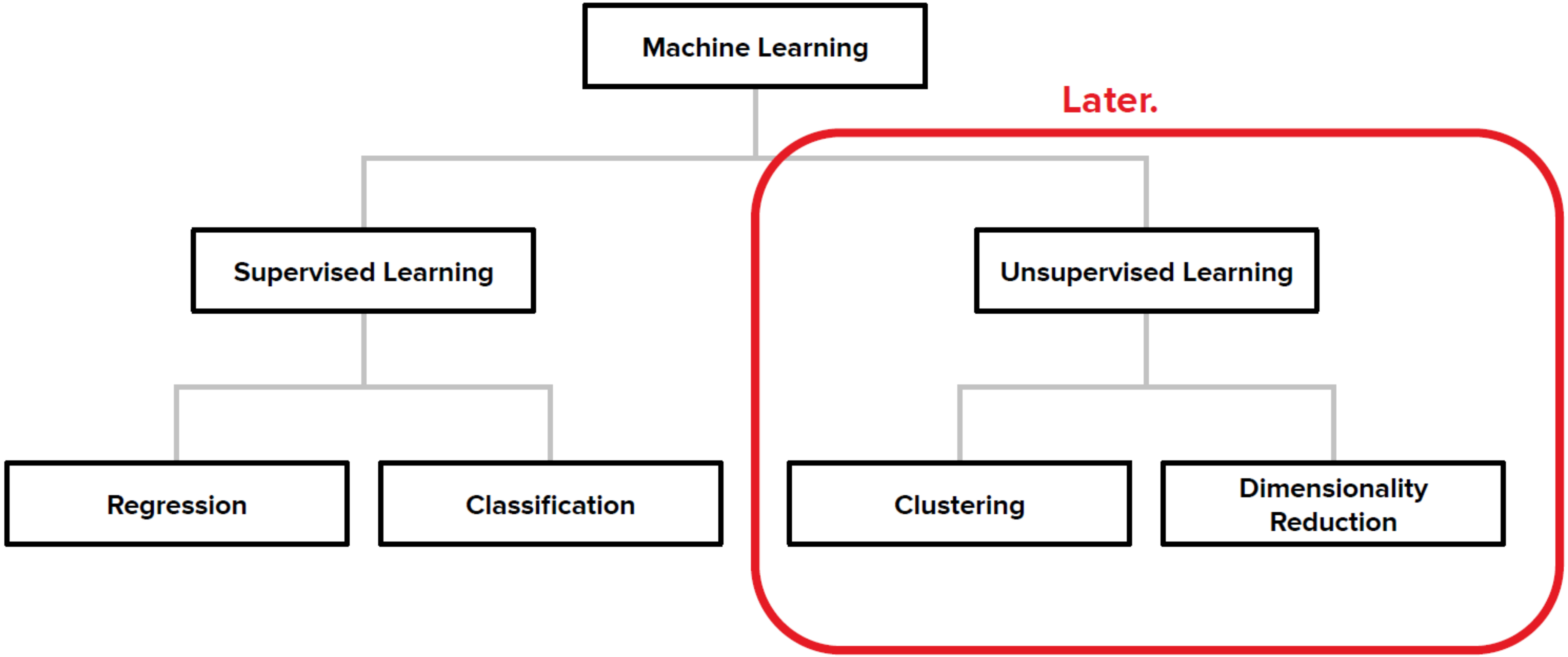
Regression

Classification

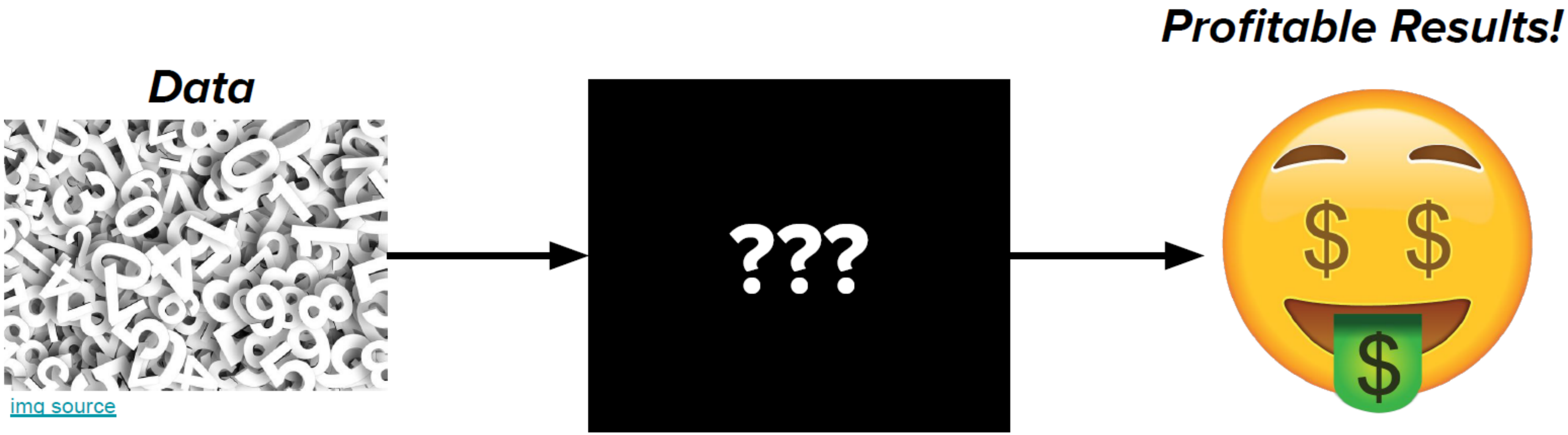
Clustering

Dimensionality Reduction

Roadmap of ML



Supervised Learning Transparency



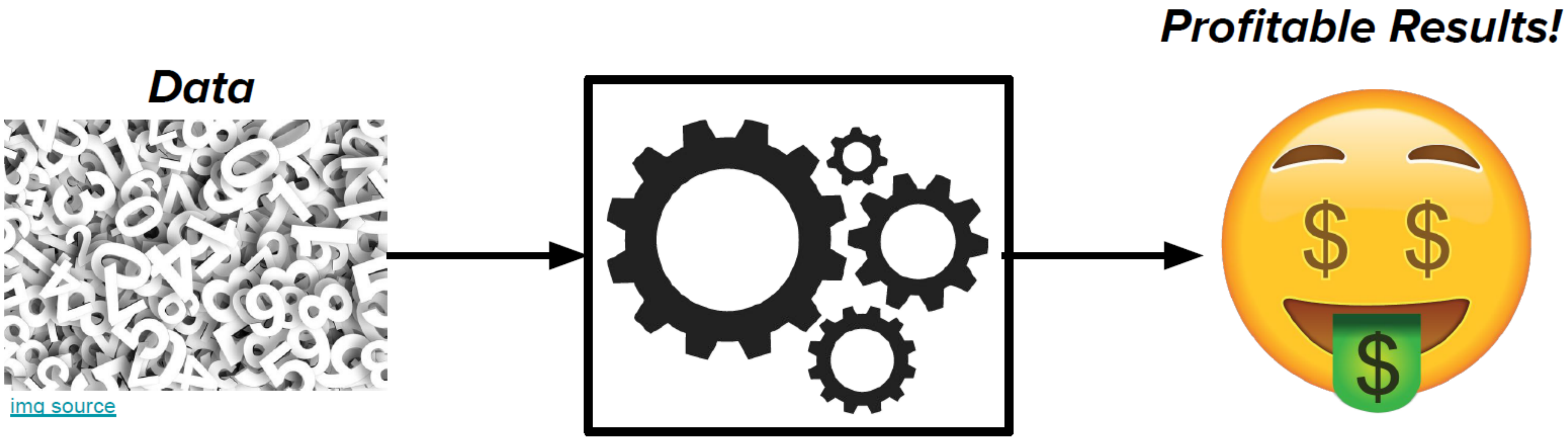
Data

???

Profitable Results!

img source

Supervised Learning Transparency



Data

Profitable Results!

img source

What is linear regression?

Remember this?

$$y = mx + b$$

What is linear regression?

$$y = mx + b$$

In data science it gets changed to:

$$y = \beta_0 + \beta_1 x_1$$

What is linear regression?

$$y = \beta_0 + \beta_1 x_1$$

Our model isn't going to be perfect. The things our model doesn't capture are errors and denoted by ε (epsilon).

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

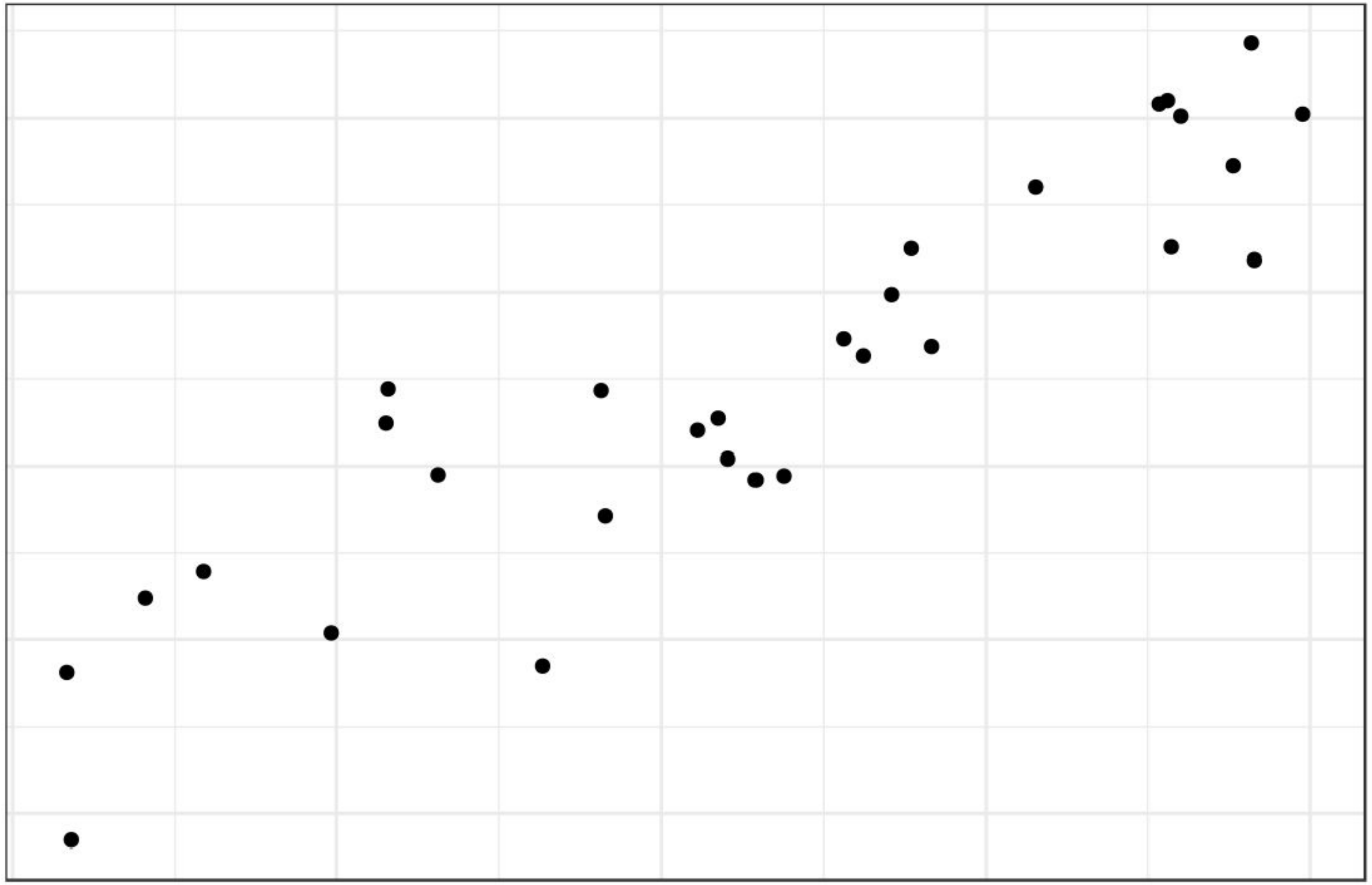
What is linear regression?

In ordinary least squares linear regression (often just referred to as **OLS**), we try to predict some response variable (*y*) from at least one independent variable (*x*). We believe there is a **linear** relationship between the two:

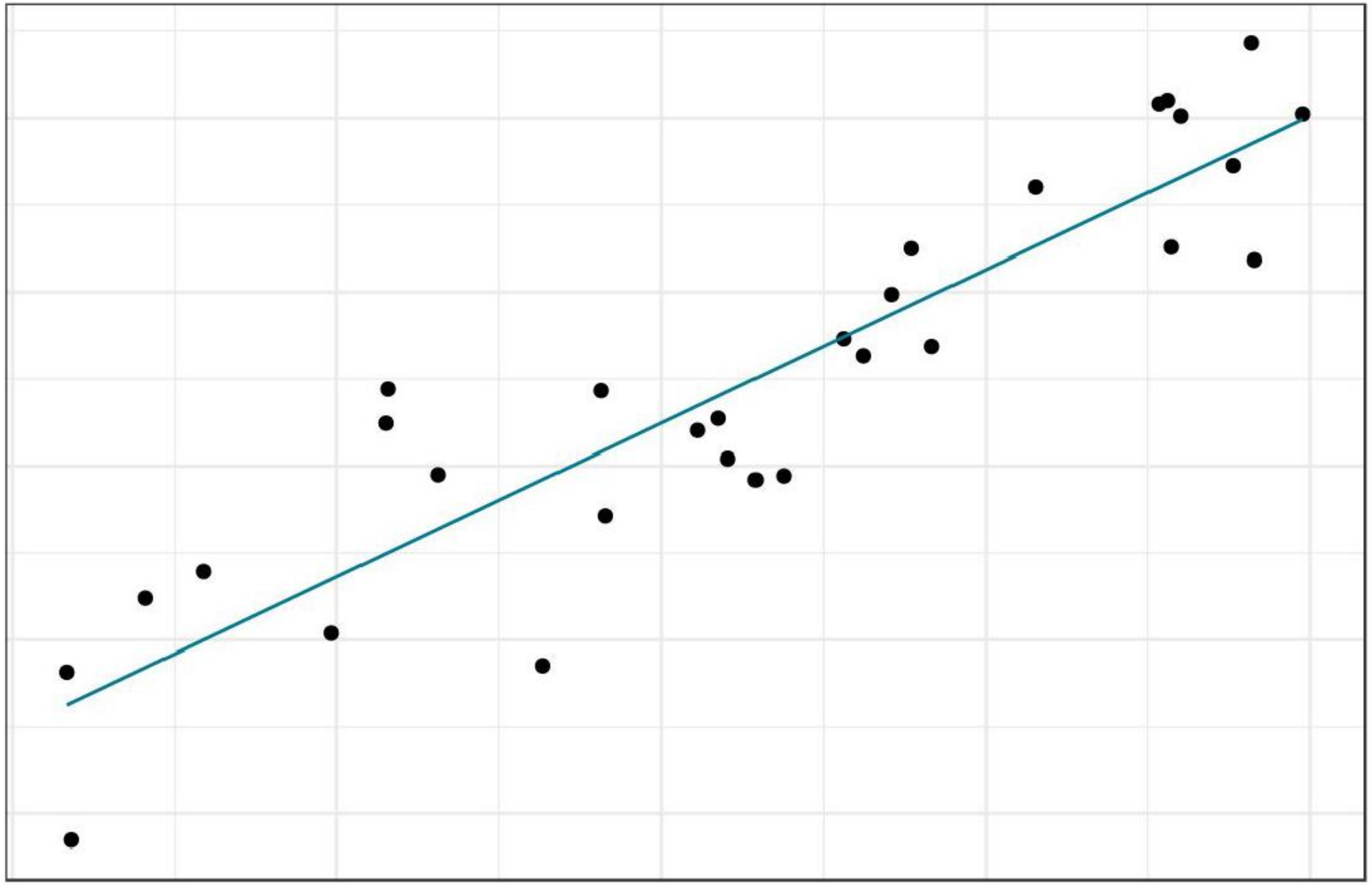
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Again, that funny looking “e” stands for “error” - it’s random noise inherent in our prediction because nothing will be perfect.

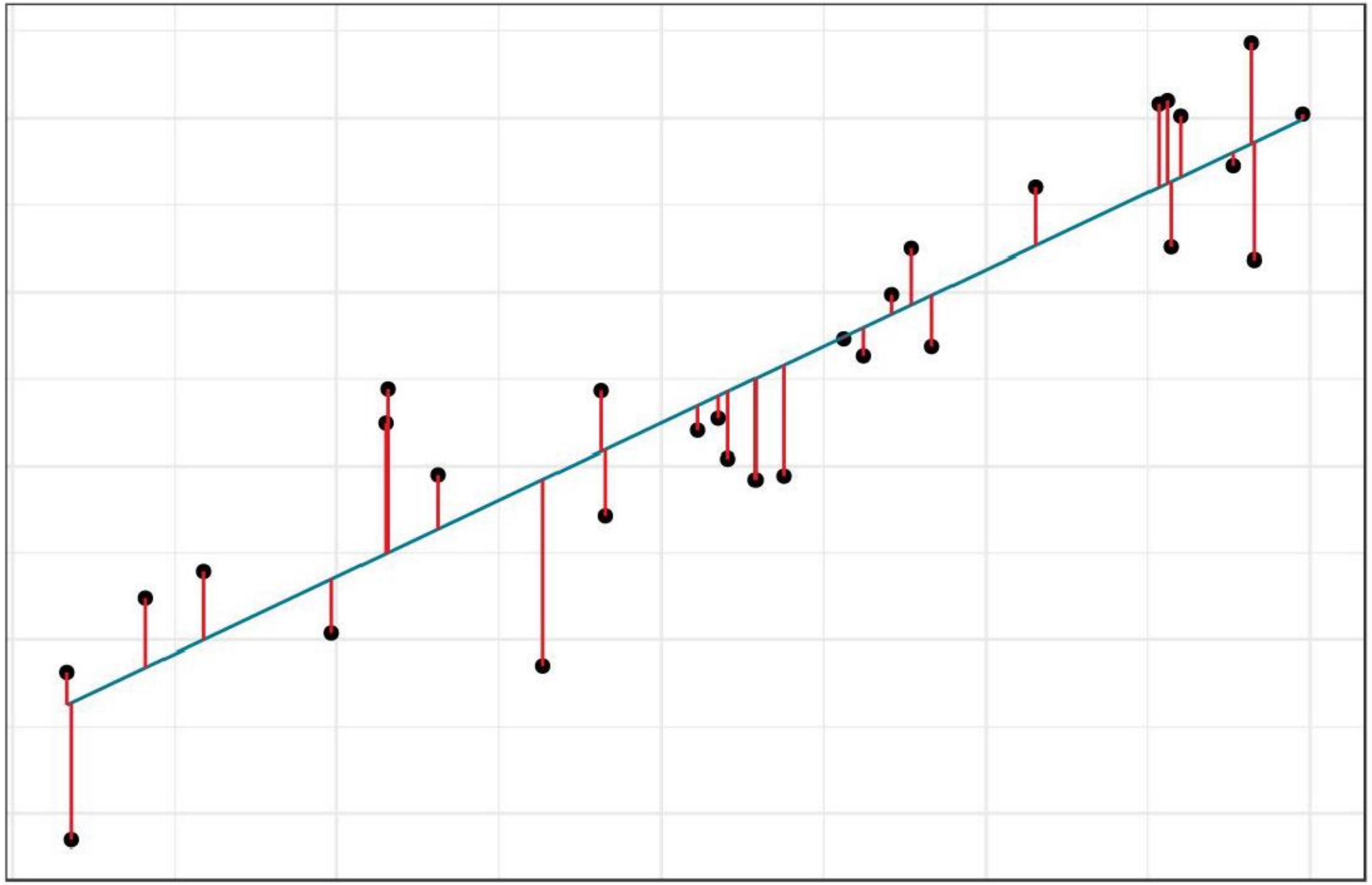
Graphically, this is:



Graphically, this is:

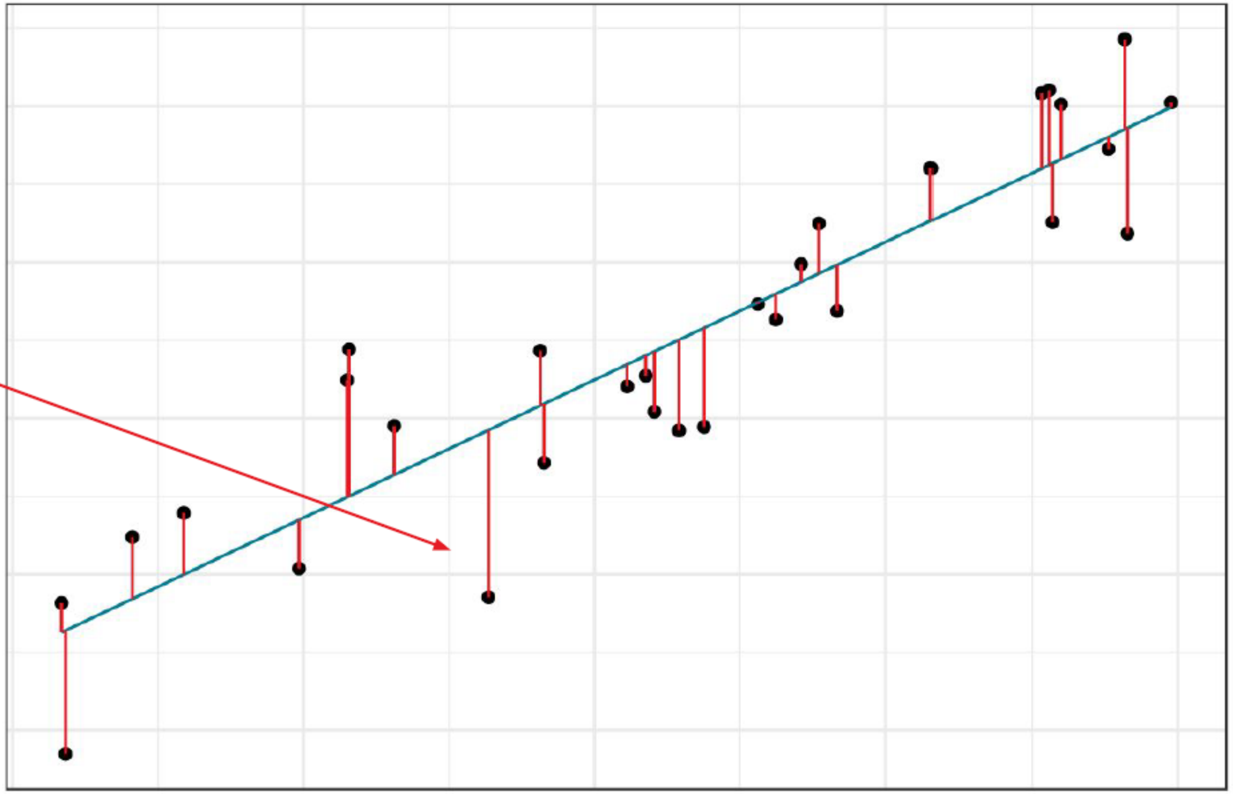


Graphically, this is:



Graphically, this is:

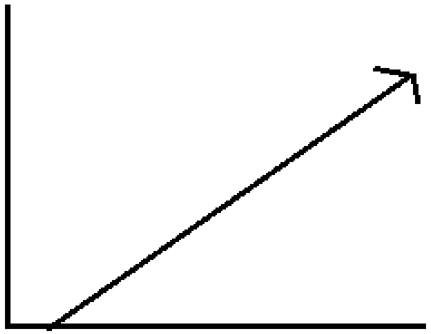
The difference between the actual and the predicted is called a **residual**, and the line of "best fit" minimizes all of these residuals.



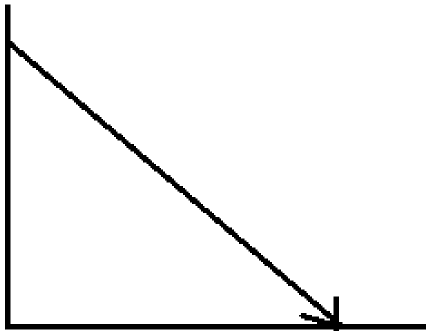
Specifically, we minimize the **sum of the squared residuals**, hence the term "least squares".

OLS

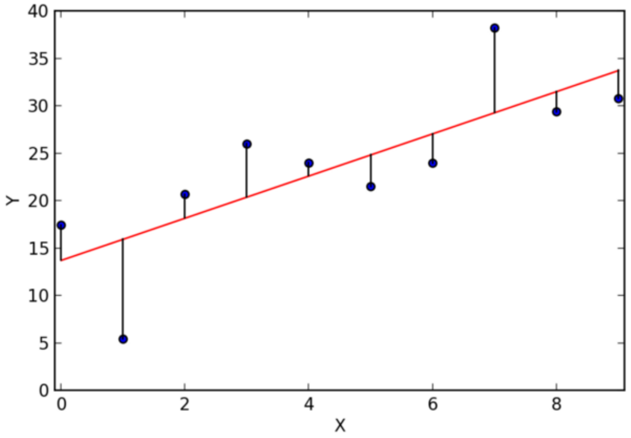
- **OLS** stands for Ordinary Least Squares and the method “Least Squares” means that we’re trying to fit a regression line that would minimize the square of distance from the regression line
- R-squared = % of variance our model explains or how well the model can predict the real values



Positive Linear Relationship



Negative Linear Relationship



OLS Assumptions

Conducting OLS comes with some pretty steep assumptions that should be satisfied before believing the results. Luckily, there's a nice acronym to remember them:

- **L** - Linearity. Relationship between *x* and *y* should be approximately linear.
- **I** - Independence. Your observations should not affect one another.
- **N** - Normality. Our residuals should be approximately normally distributed.
- **E** - Equal variances, aka "**homoscedasticity**". Residuals should have approximately equal variances for each *x*.

Categorical Features

How do we work with categorical variables in our model? We can simply use **binary categorical features** as 0/1 variables.

But what if our variable has more than two **levels**?

First some more notation!

Categorical Features for Linear Regression

- Create dummy variables to convert categorical data into numeric data in binary form so that they can be used in a regression model

- A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study.

- Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups.

What to do if our LINE assumptions are violated?!

A common scenario in linear model is when you have:

- A slightly **curvilinear** relationship between *x* and *y*
- Very right-skew residuals
- Residuals that tend to spread out from right to left (a “fan shape”)

One quick fix that should improve all of these issues is doing **log regression**. That is, simply take the natural log of *y* before modeling!

Regression Metrics

- **R square or coeff. of determination**
  - Shows percentage variation in y which is explained by all the x variables together.
  - Higher the better. It is always between 0 and 1.
  - It can never be negative – since it is a squared value.

- **Mean Squared Error (MSE)**
  - measures the average of the squares of the errors
  - that is, the average squared difference between the estimated values and the actual value

What is R-Squared?

- R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by independent variable(s) in a regression model.

- Whereas correlation explains the strength of the relationship between an independent and dependent variable, R2 explains to what extent the variance of one variable explains the variance of the second variable.
  - So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.