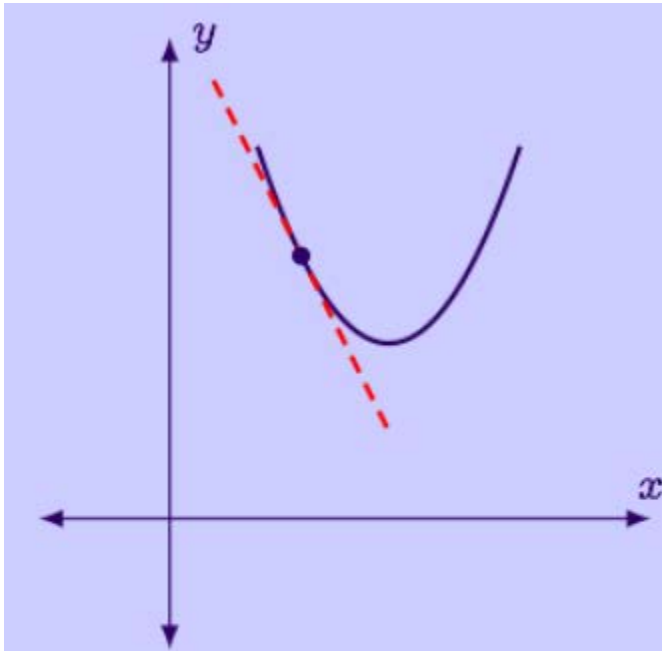
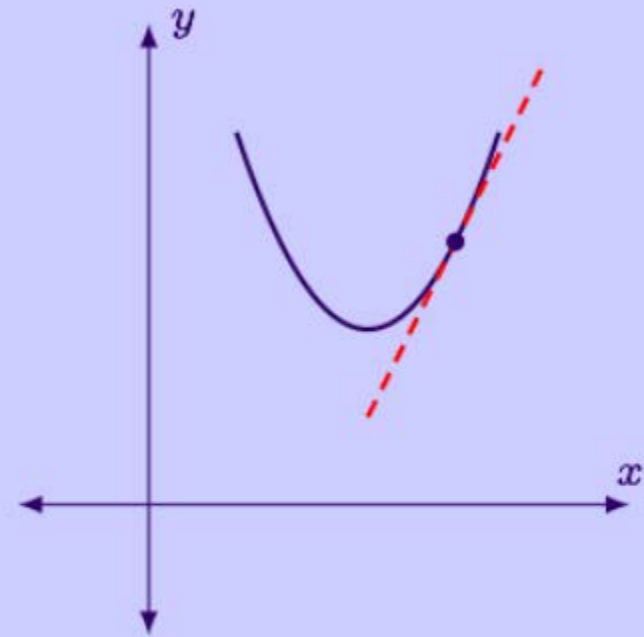


Gradient Descent

Derivative



When the tangent line has a negative slope, the function is decreasing at that point.

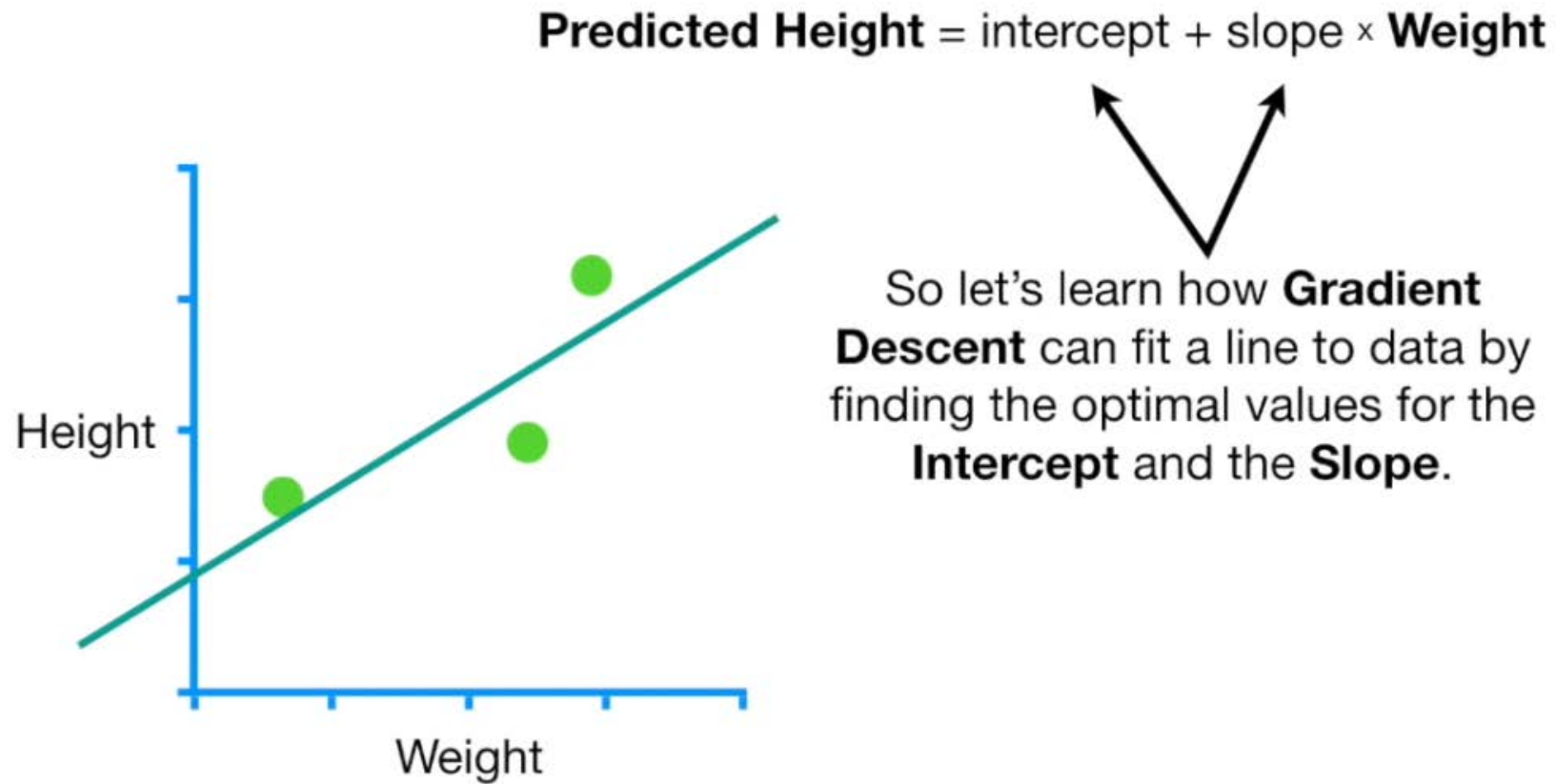


When the tangent line has a positive slope, the function is increasing at that point.

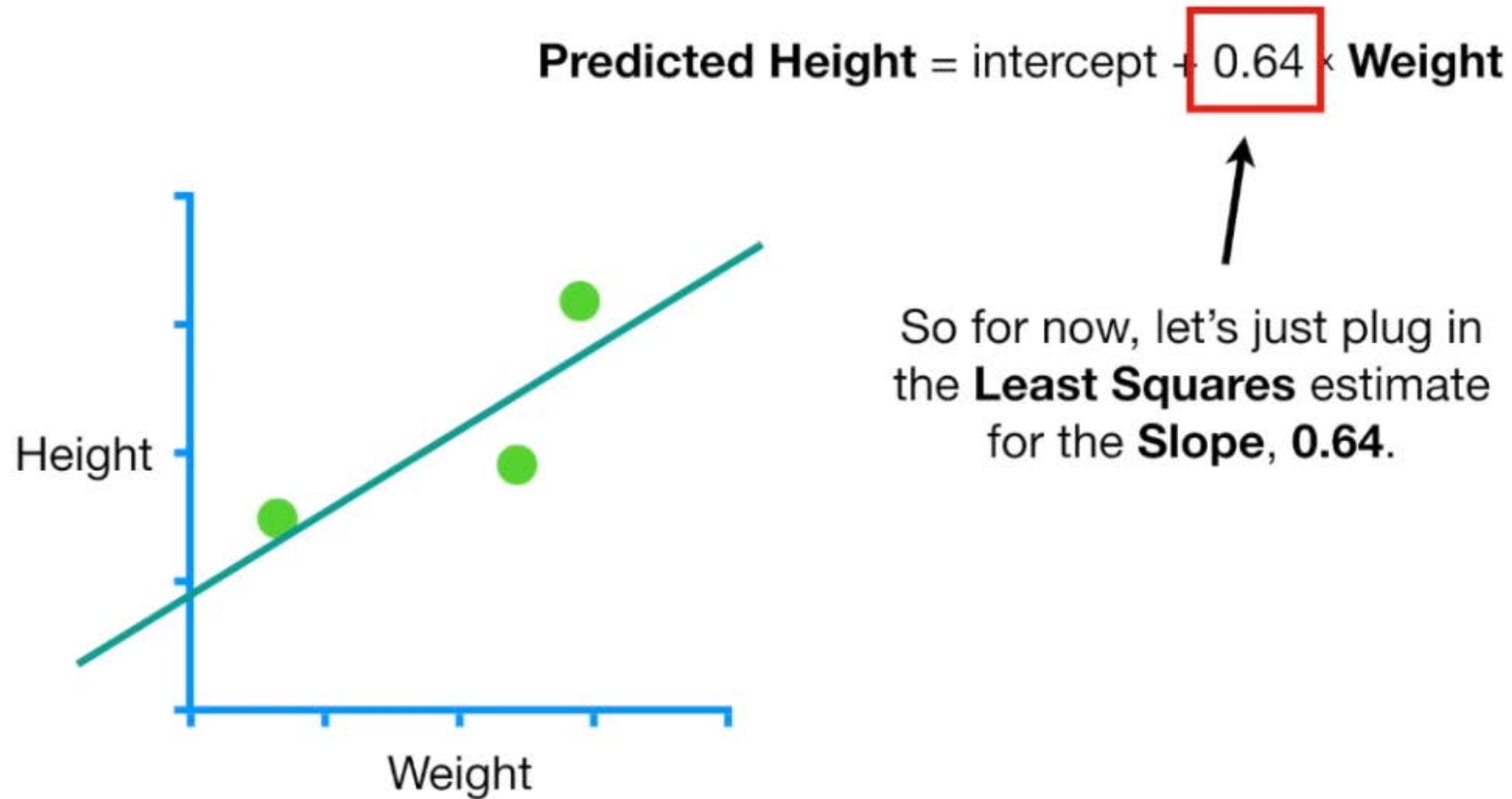
Gradient Descent

- Two or more derivatives of the same function are called Gradients.
- An algorithm which uses gradient to descent to the lowest point of a loss function.

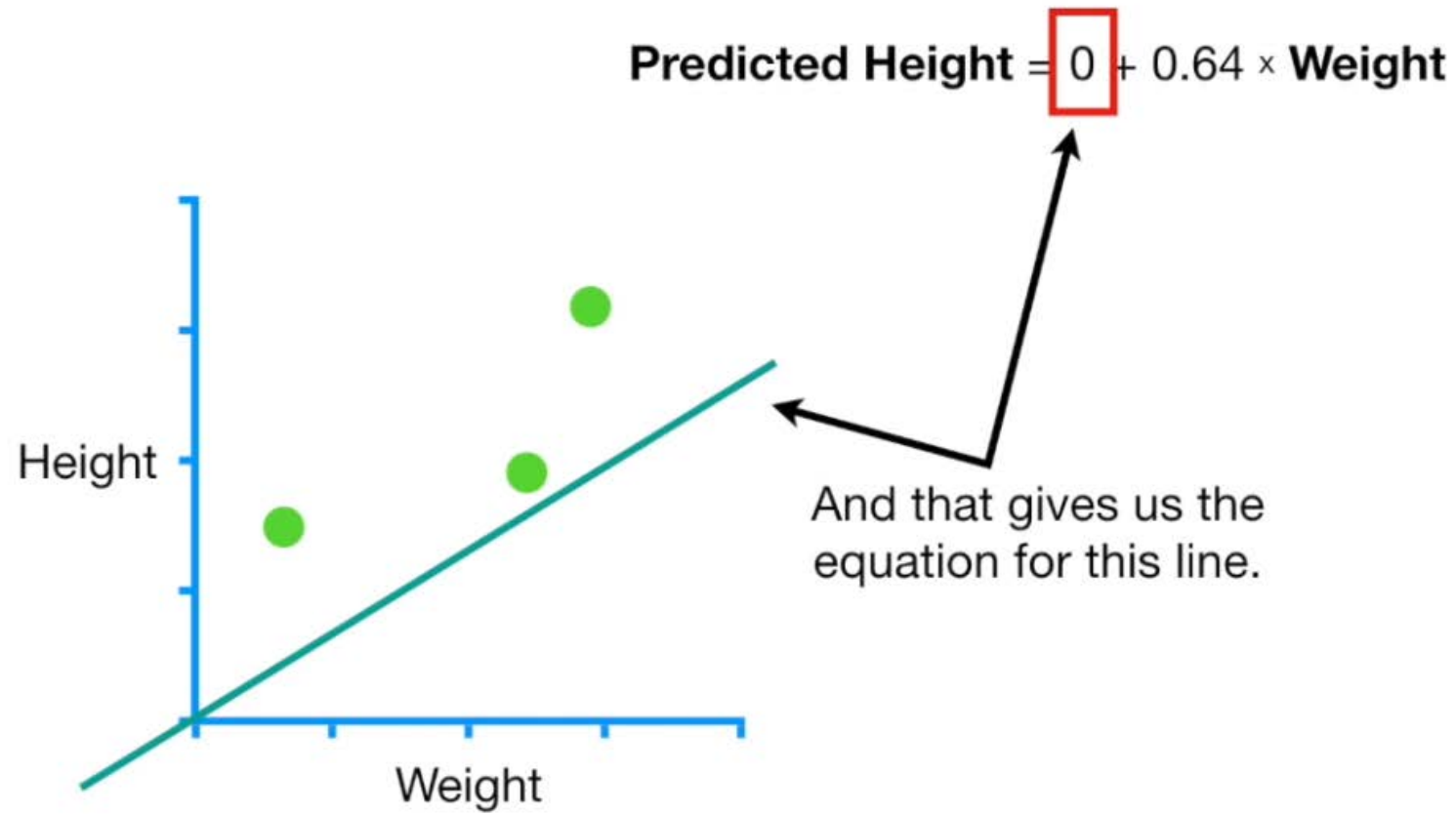
Gradient Descent



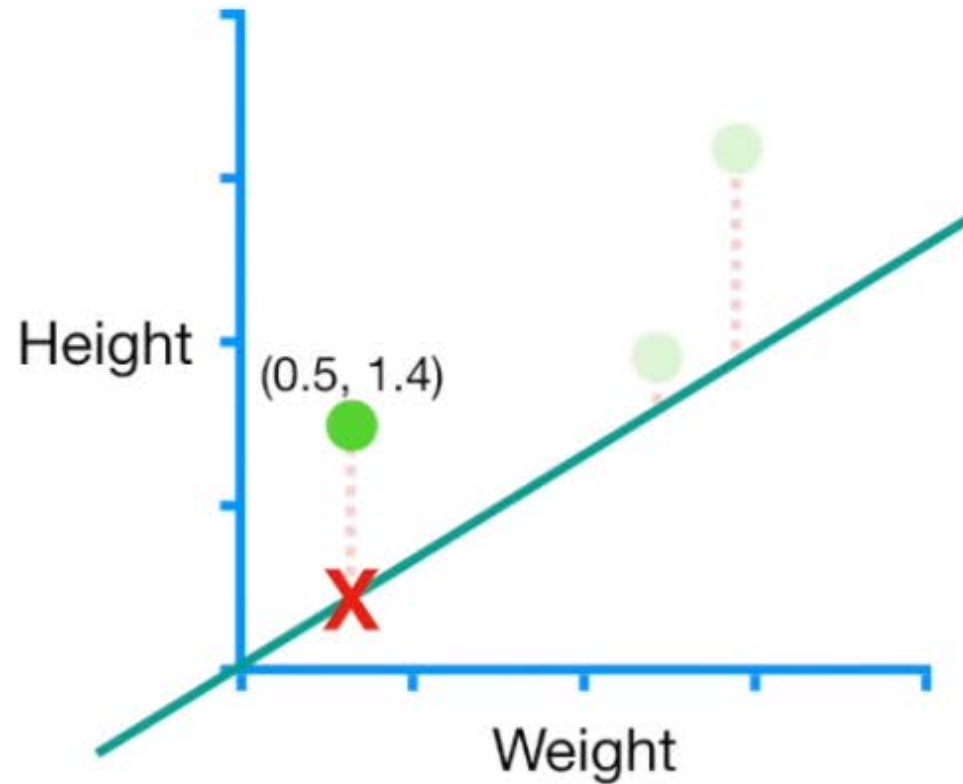
Gradient Descent



Gradient Descent



Gradient Descent

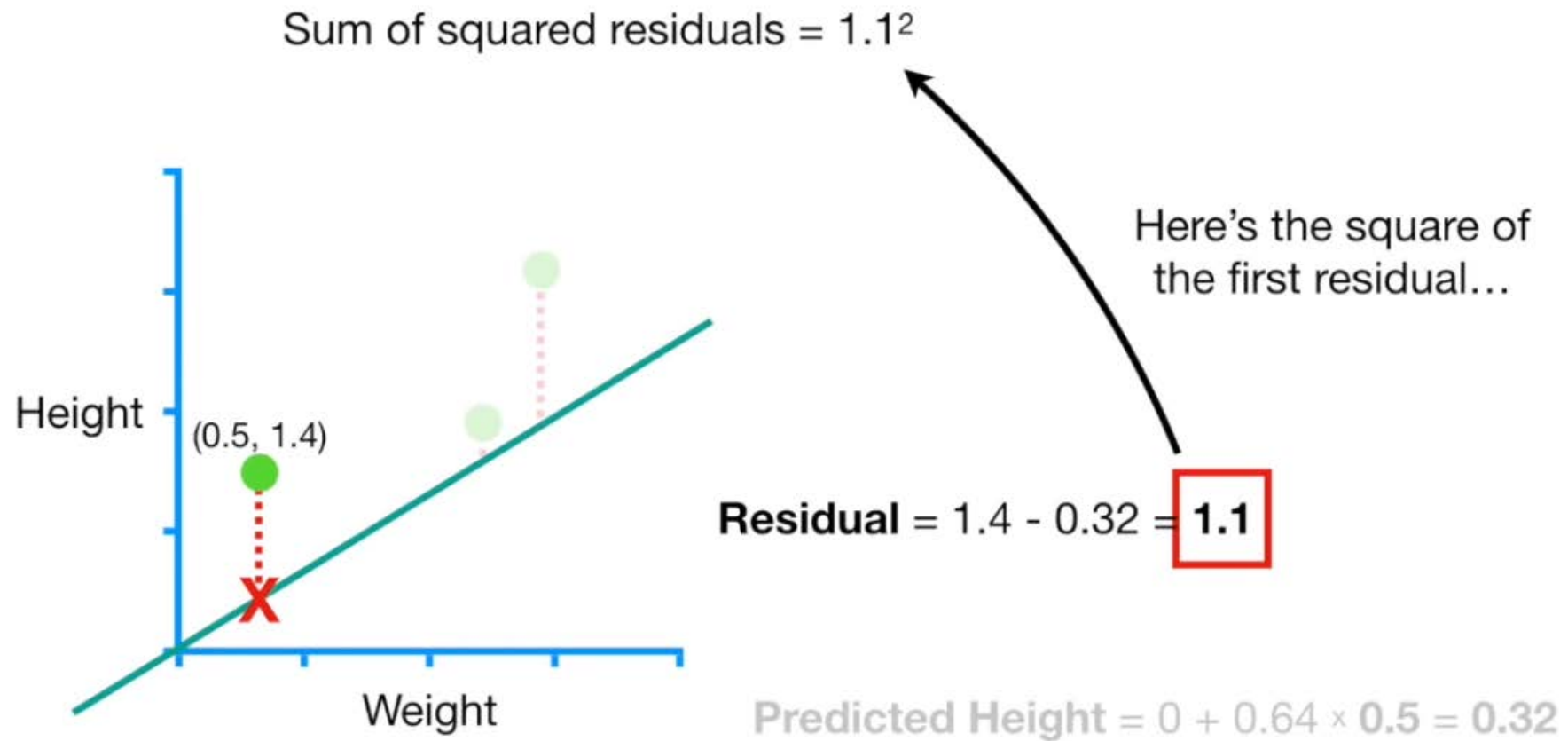


We get the **Predicted Height**, the point on the line...

...by plugging **Weight = 0.5** into the equation for the line...

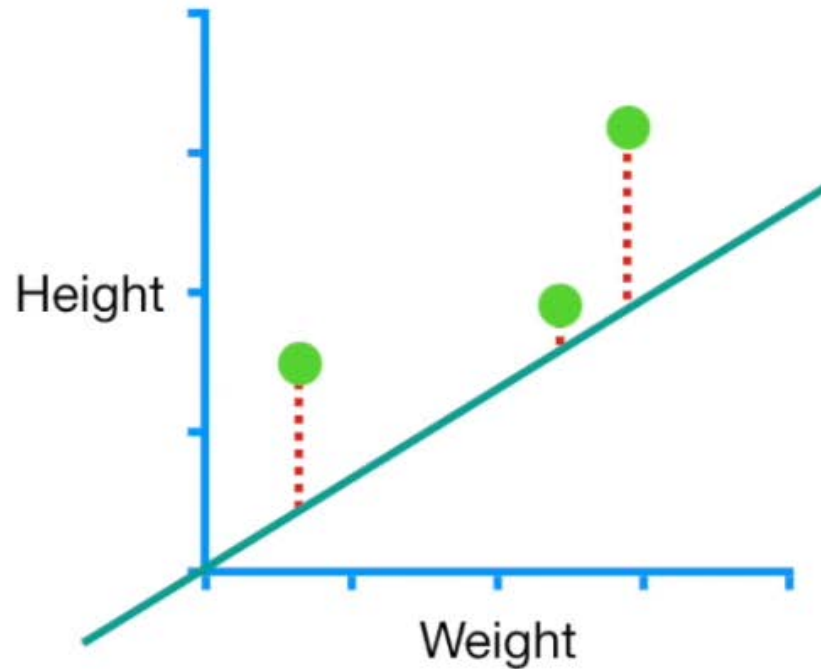
$$\text{Predicted Height} = 0 + 0.64 \times \mathbf{0.5}$$

Gradient Descent



Gradient Descent

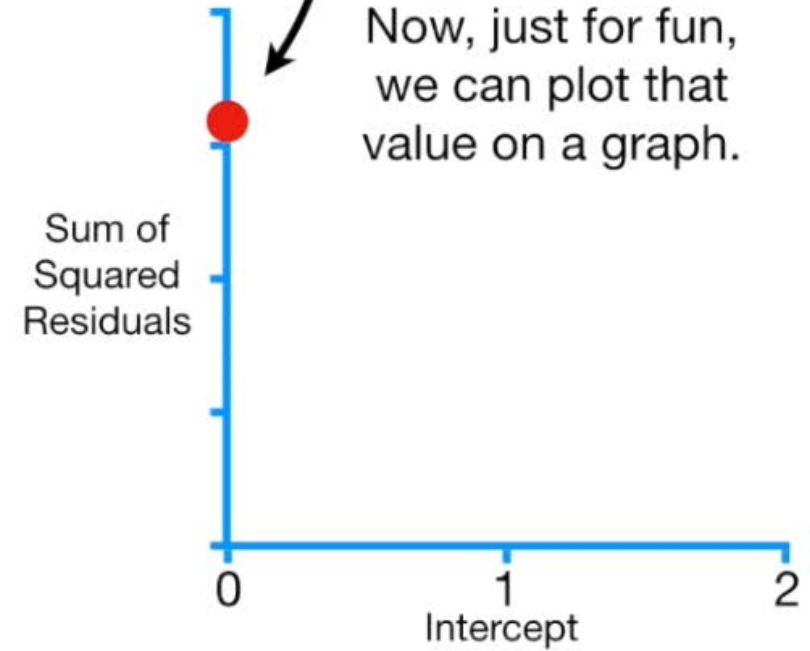
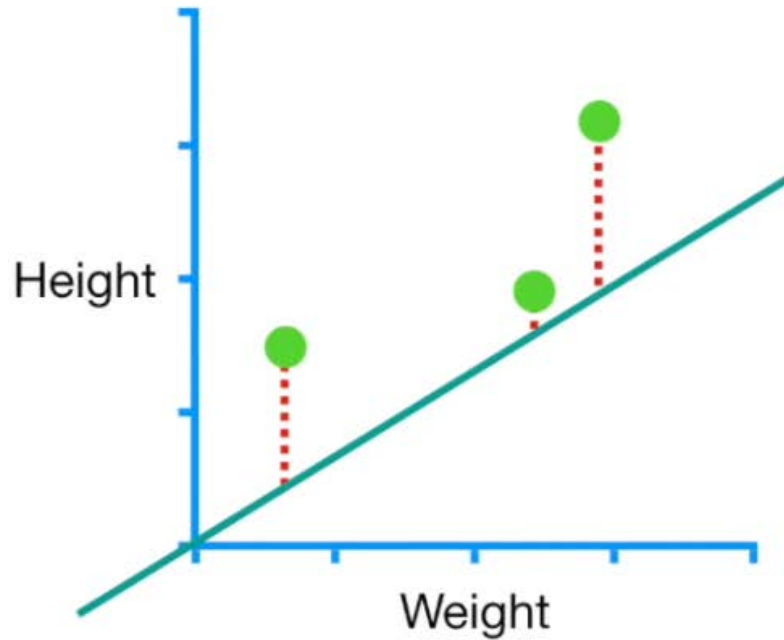
$$\text{Sum of squared residuals} = 1.1^2 + 0.4^2 + 1.3^2 = 3.1$$



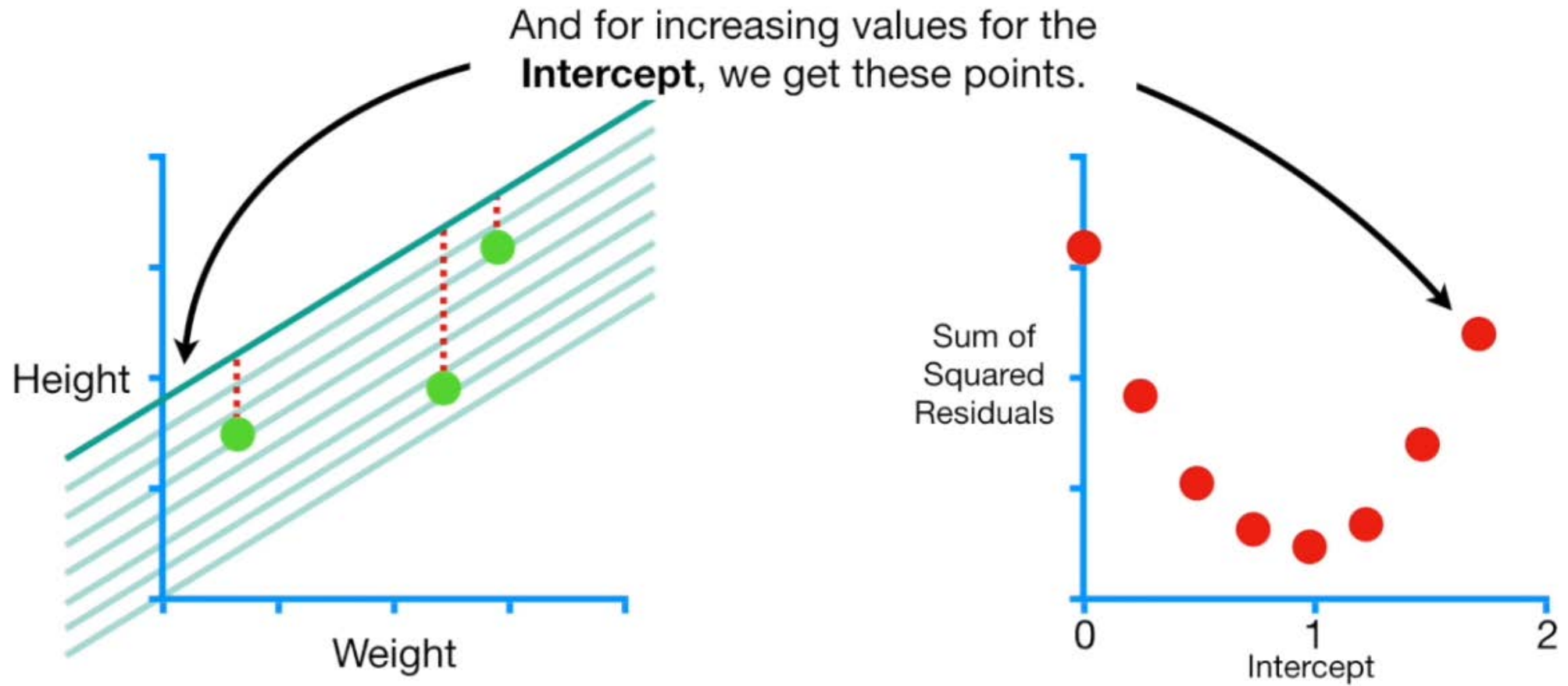
In the end, **3.1** is the Sum of the Squared Residuals.

Gradient Descent

$$\text{Sum of squared residuals} = 1.1^2 + 0.4^2 + 1.3^2 = 3.1$$

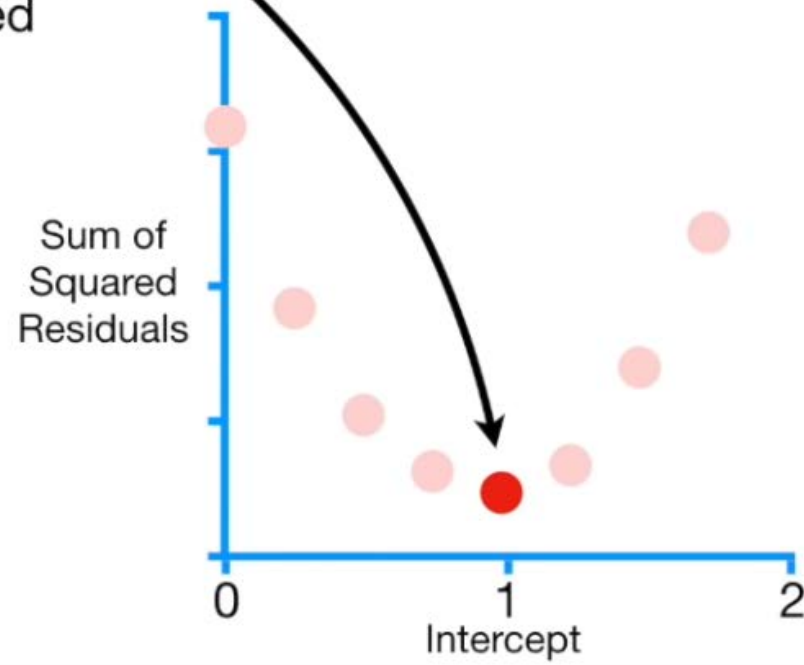


Gradient Descent



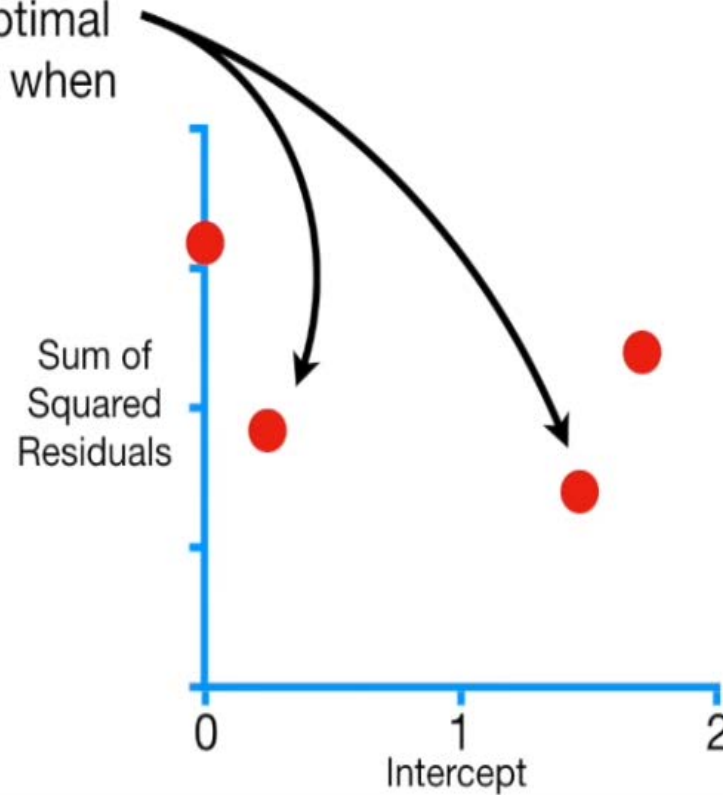
Gradient Descent

Of the points that we
calculated for the graph,
this one has the lowest
Sum of Squared
Residuals...

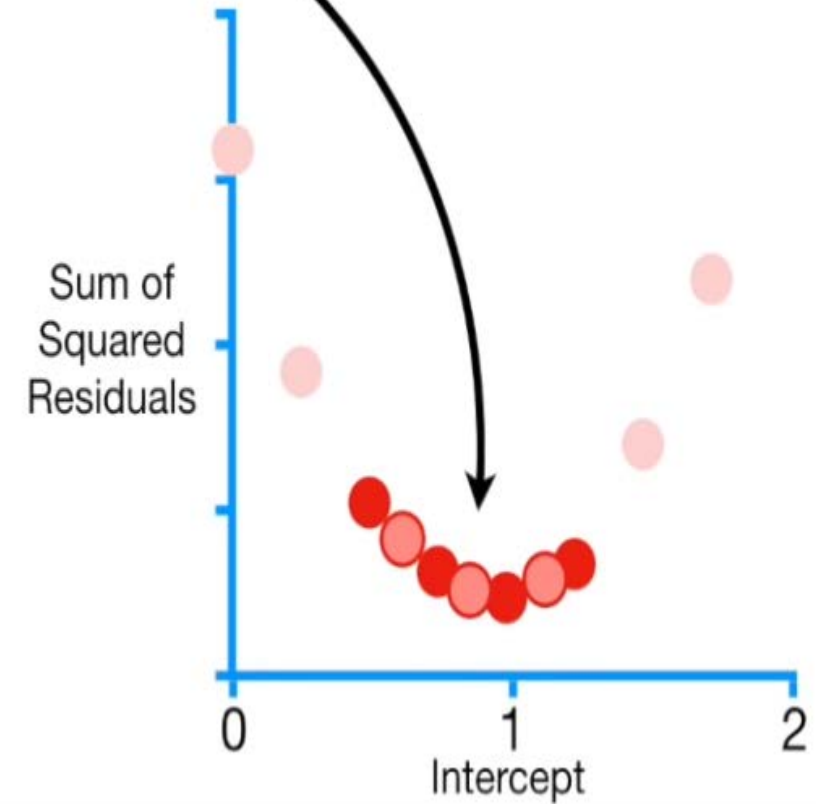


Gradient Descent

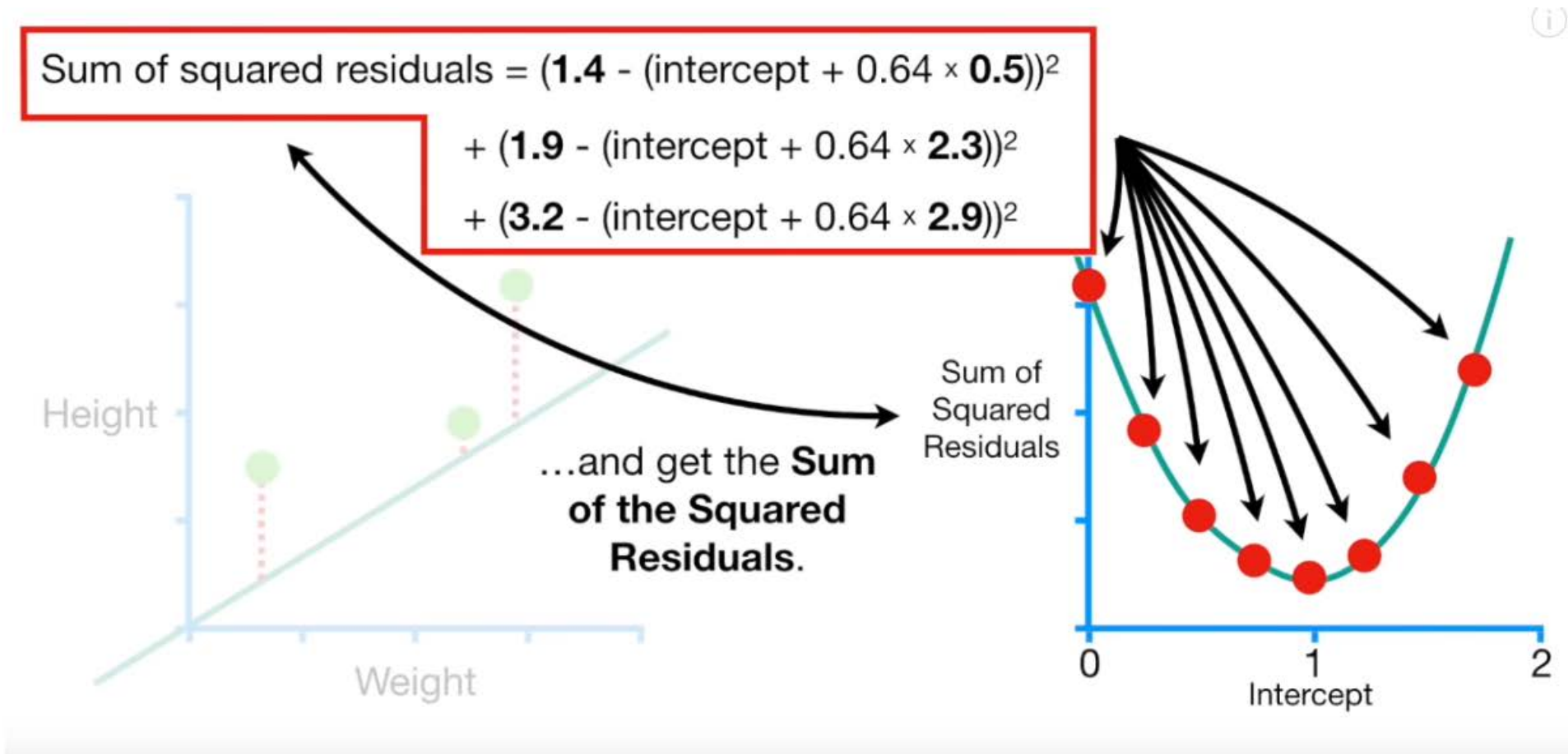
In other words, **Gradient Descent** identifies the optimal value by taking big steps when it is far away...



...and baby steps when it is close.



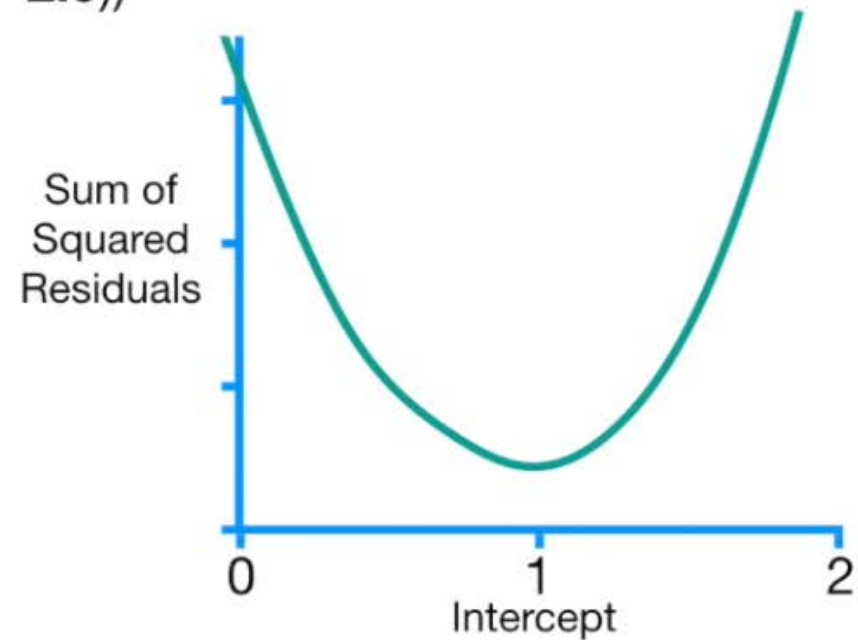
Gradient Descent



Gradient Descent

$$\begin{aligned}\text{Sum of squared residuals} = & (\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2 \\ & + (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2 \\ & + (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2\end{aligned}$$

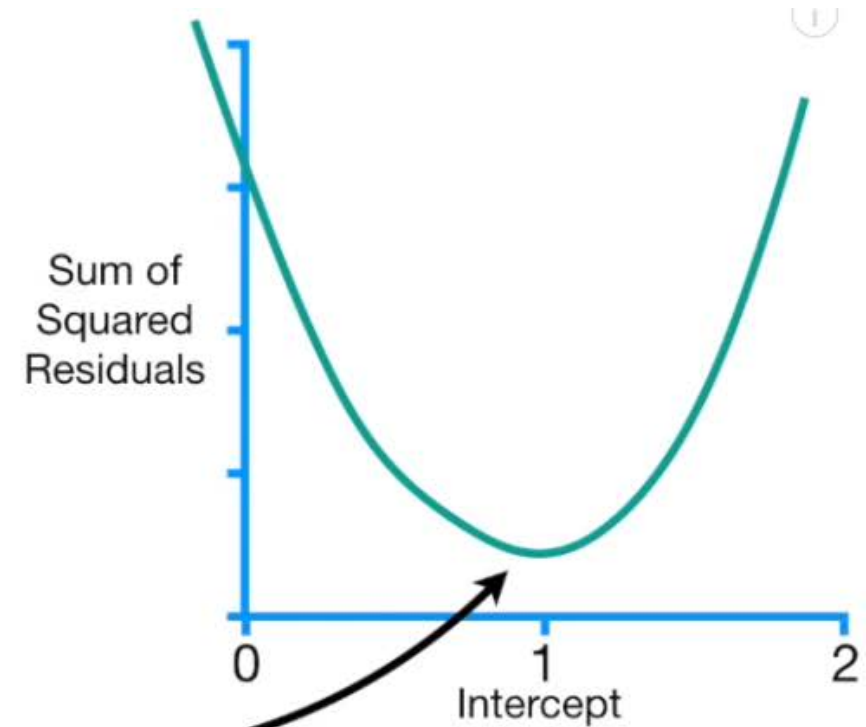
So let's take the derivative
of the Sum of the
Squared Residuals with
respect to the **Intercept**.



Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (\text{intercept} + 0.64 \times 0.5)) \\ & + -2(1.9 - (\text{intercept} + 0.64 \times 2.3)) \\ & + -2(3.2 - (\text{intercept} + 0.64 \times 2.9)) \end{aligned}$$

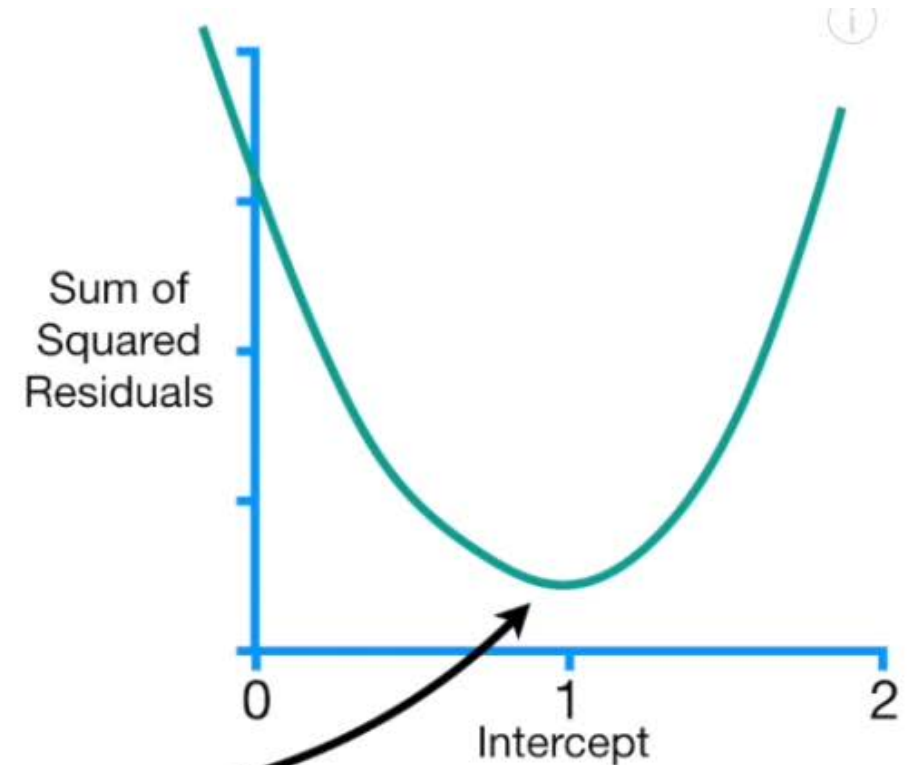
Now that we have the derivative,
Gradient Descent will use it to find
where the Sum of Squared
Residuals is lowest.



Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (\text{intercept} + 0.64 \times 0.5)) \\ & + -2(1.9 - (\text{intercept} + 0.64 \times 2.3)) \\ & + -2(3.2 - (\text{intercept} + 0.64 \times 2.9)) \end{aligned}$$

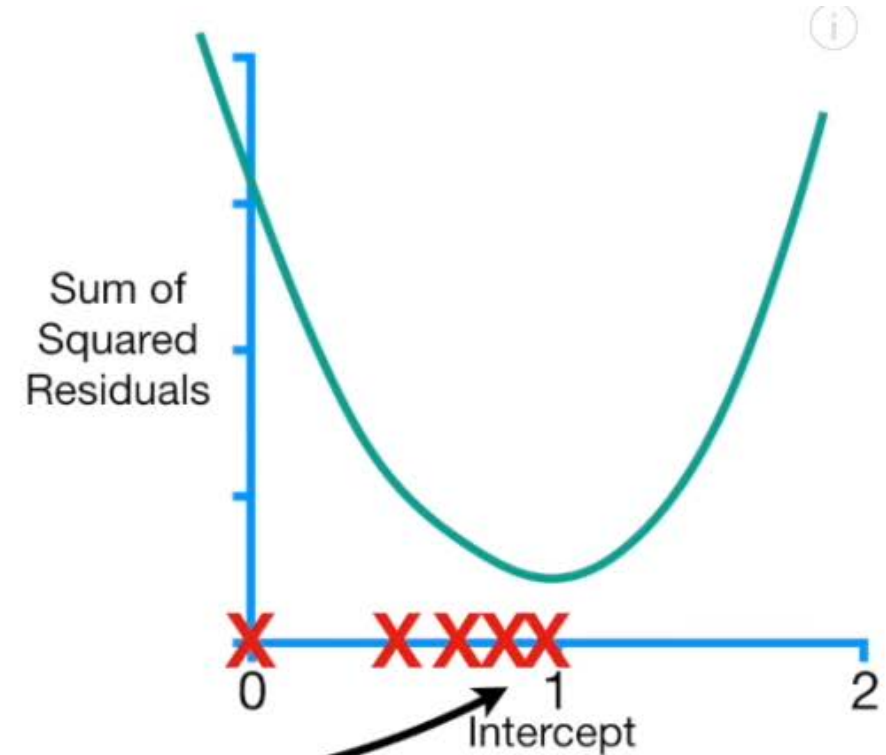
NOTE: If we were using **Least Squares** to solve for the optimal value for the **Intercept**, we would simply find where the the slope of the curve = **0**.



Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (\text{intercept} + 0.64 \times 0.5)) \\ & + -2(1.9 - (\text{intercept} + 0.64 \times 2.3)) \\ & + -2(3.2 - (\text{intercept} + 0.64 \times 2.9)) \end{aligned}$$

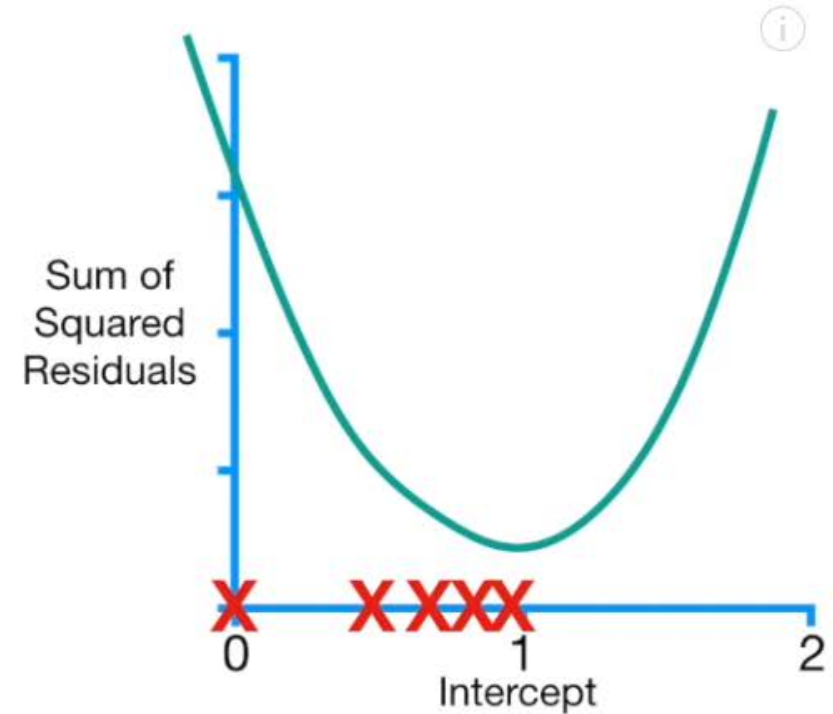
In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.



Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (\text{intercept} + 0.64 \times 0.5)) \\ & + -2(1.9 - (\text{intercept} + 0.64 \times 2.3)) \\ & + -2(3.2 - (\text{intercept} + 0.64 \times 2.9)) \end{aligned}$$

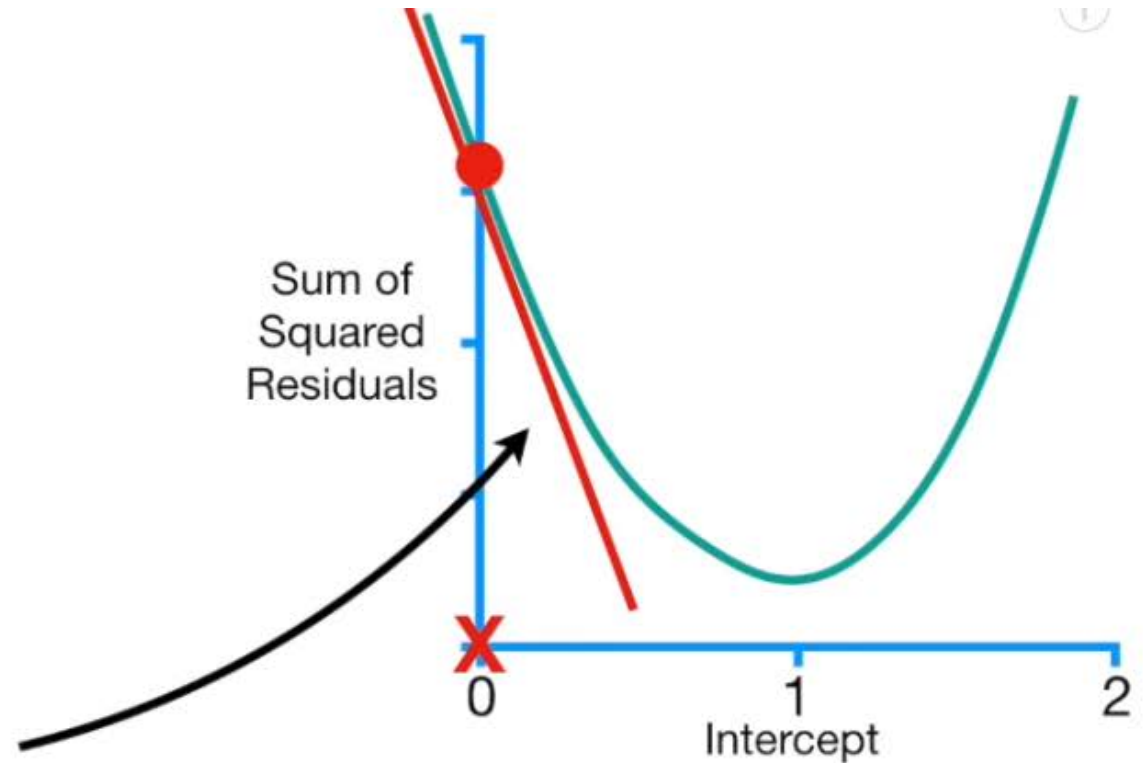
This makes **Gradient Descent** very useful when it is not possible to solve for where the derivative = 0, and this is why **Gradient Descent** can be used in so many different situations.



Gradient Descent

$$\begin{aligned} \frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} = & -2(1.4 - (0 + 0.64 \times 0.5)) \\ & + -2(1.9 - (0 + 0.64 \times 2.3)) \\ & + -2(3.2 - (0 + 0.64 \times 2.9)) \\ & = -5.7 \end{aligned}$$

So when the **Intercept** = 0,
the slope of the curve = **-5.7**.

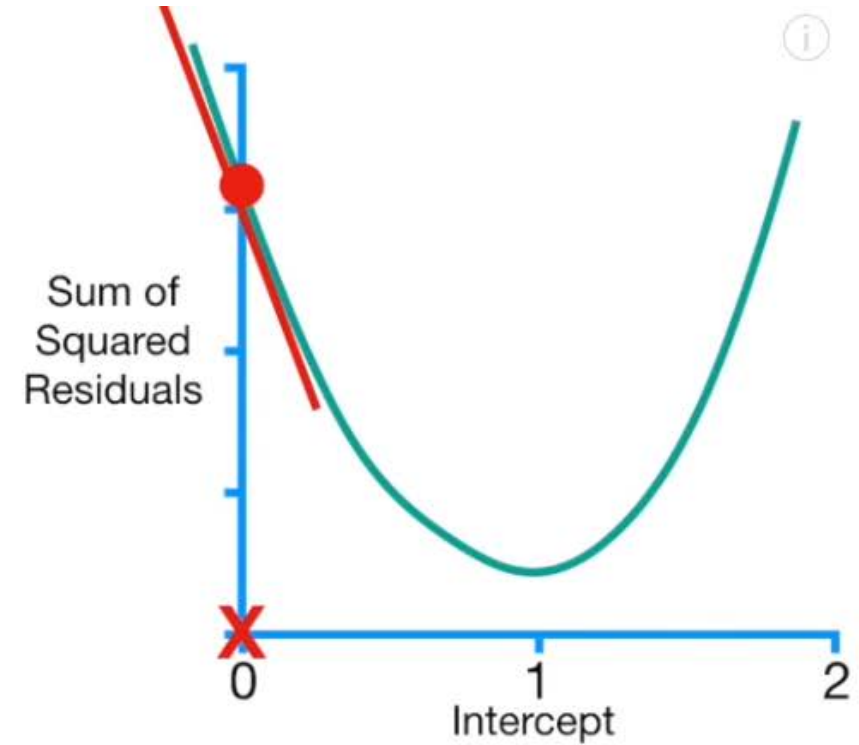


Gradient Descent

$$\begin{aligned} \frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} &= \\ &= -2(1.4 - (0 + 0.64 \times 0.5)) \\ &\quad + -2(1.9 - (0 + 0.64 \times 2.3)) \\ &\quad + -2(3.2 - (0 + 0.64 \times 2.9)) \\ &= -5.7 \end{aligned}$$

$$\text{Step Size} = -5.7 \times 0.1$$

...by a small number called
The Learning Rate.

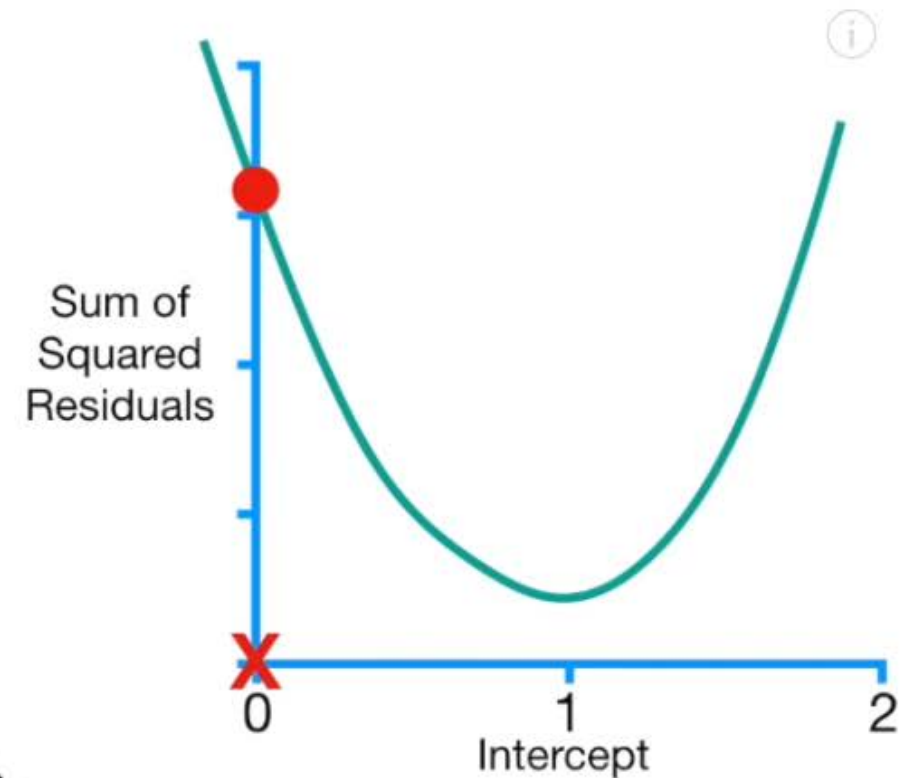


Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (0 + 0.64 \times 0.5)) \\ & + -2(1.9 - (0 + 0.64 \times 2.3)) \\ & + -2(3.2 - (0 + 0.64 \times 2.9)) \\ & = -5.7 \end{aligned}$$

$$\text{Step Size} = -5.7 \times 0.1 = -0.57$$

$$\text{New Intercept} = \text{Old Intercept} - \text{Step Size}$$



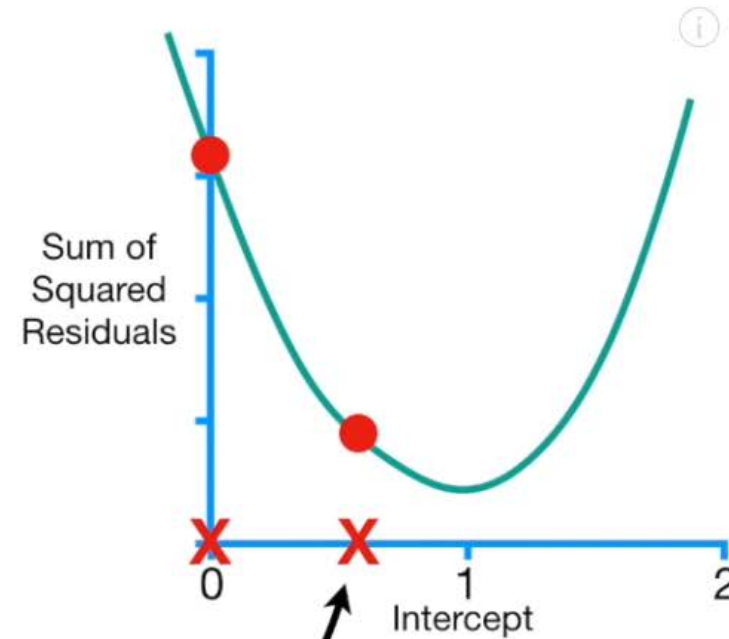
Gradient Descent

$$\begin{aligned} \frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} &= \\ &= -2(1.4 - (0 + 0.64 \times 0.5)) \\ &\quad + -2(1.9 - (0 + 0.64 \times 2.3)) \\ &\quad + -2(3.2 - (0 + 0.64 \times 2.9)) \\ &= -5.7 \end{aligned}$$

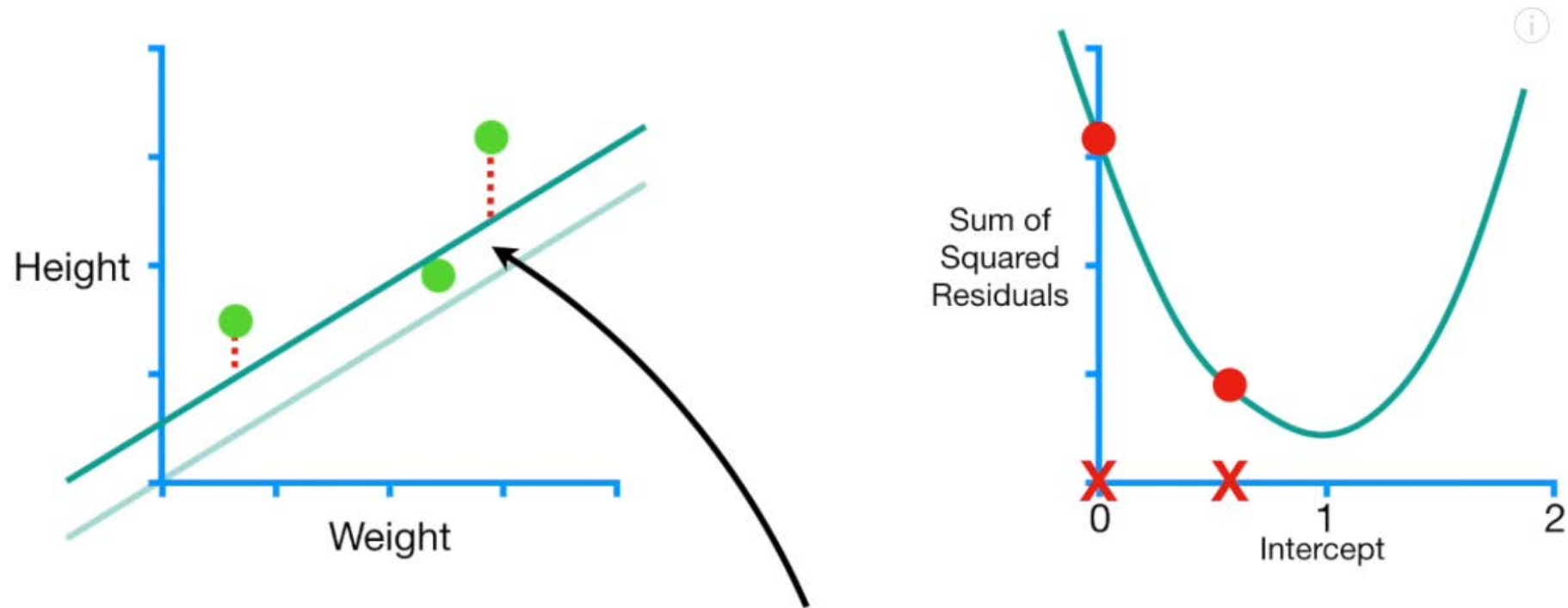
$$\text{Step Size} = -5.7 \times 0.1 = -0.57$$

$$\text{New Intercept} = 0 - (-0.57) = \boxed{0.57}$$

...and the the **New Intercept = 0.57.**



Gradient Descent

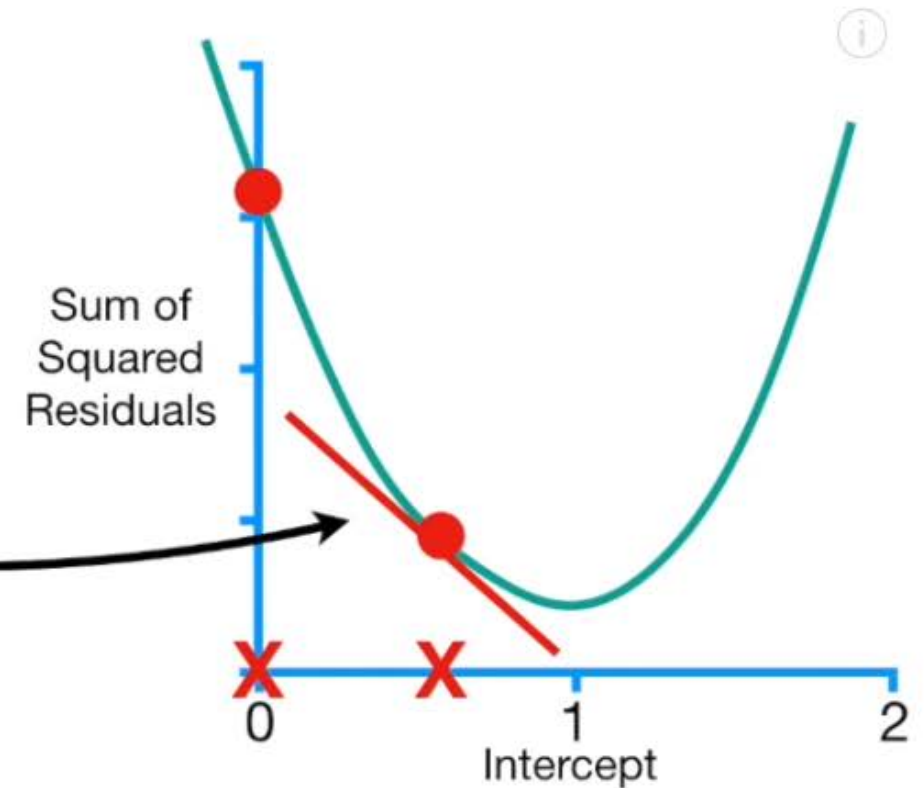


...we can see how much the residuals shrink when the
Intercept = 0.57.

Gradient Descent

$$\begin{aligned} \frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} = & -2(1.4 - (0.57 + 0.64 \times 0.5)) \\ & + -2(1.9 - (0.57 + 0.64 \times 2.3)) \\ & + -2(3.2 - (0.57 + 0.64 \times 2.9)) \\ = & \boxed{-2.3} \end{aligned}$$

...and that tells us the
slope of the curve = **-2.3**.



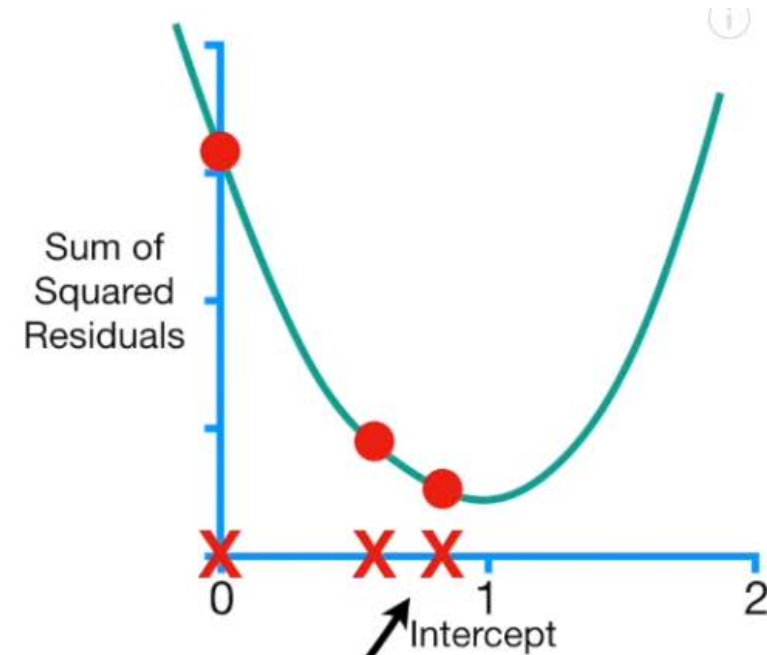
Gradient Descent

$$\begin{aligned} \frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} = & -2(1.4 - (0.57 + 0.64 \times 0.5)) \\ & + -2(1.9 - (0.57 + 0.64 \times 2.3)) \\ & + -2(3.2 - (0.57 + 0.64 \times 2.9)) \\ = & -2.3 \end{aligned}$$

$$\text{Step Size} = -2.3 \times 0.1 = -0.23$$

$$\text{New Intercept} = 0.57 - (-0.23) = \boxed{0.8}$$

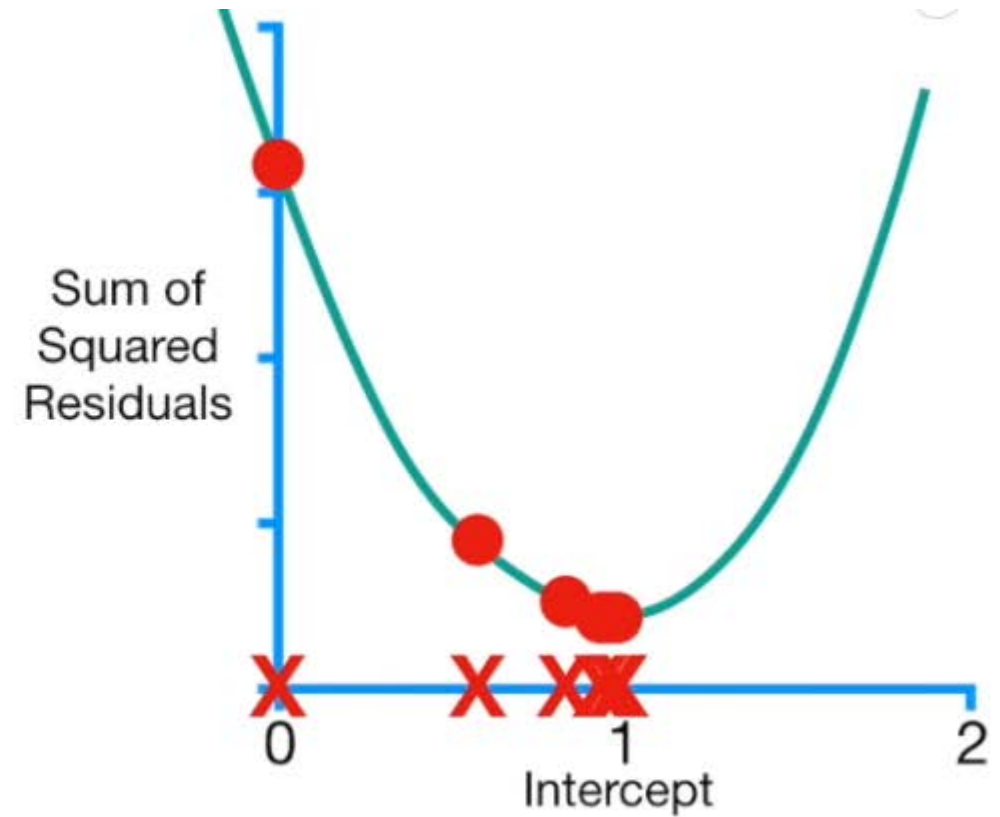
...and the **New Intercept = 0.8**



Gradient Descent

Gradient Descent stops
when the **Step Size** is **Very
Close To 0**.

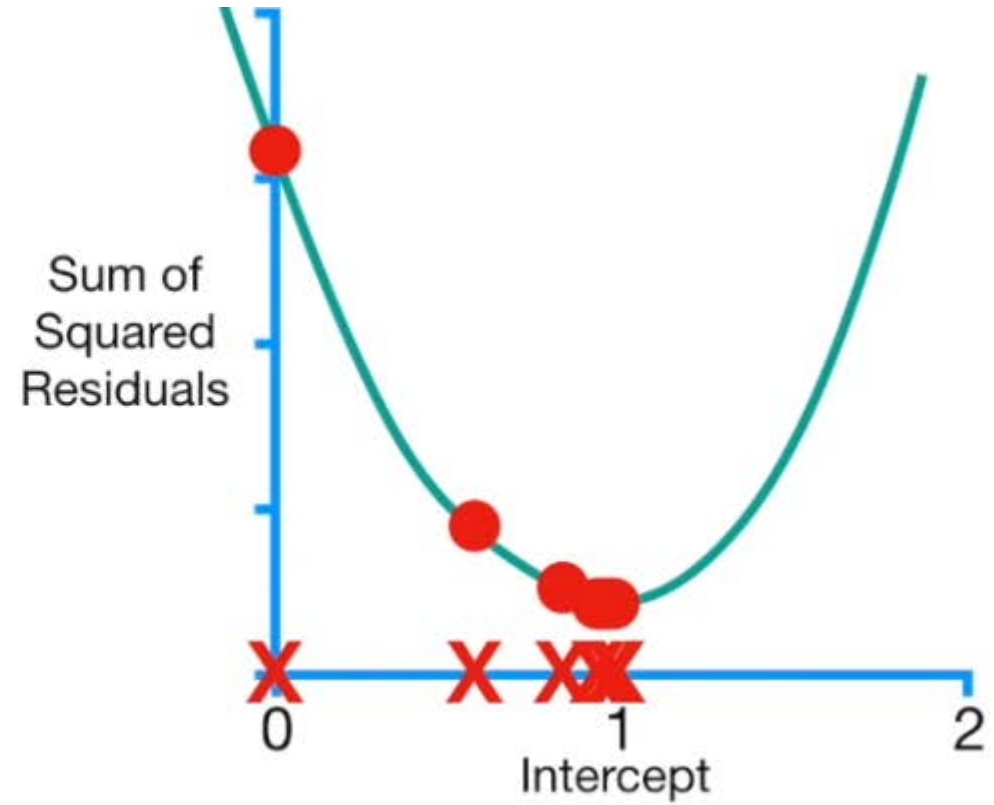
$$\text{Step Size} = \text{Slope} \times \text{Learning Rate}$$



Gradient Descent

In practice, the
Minimum Step Size = 0.001
or smaller.

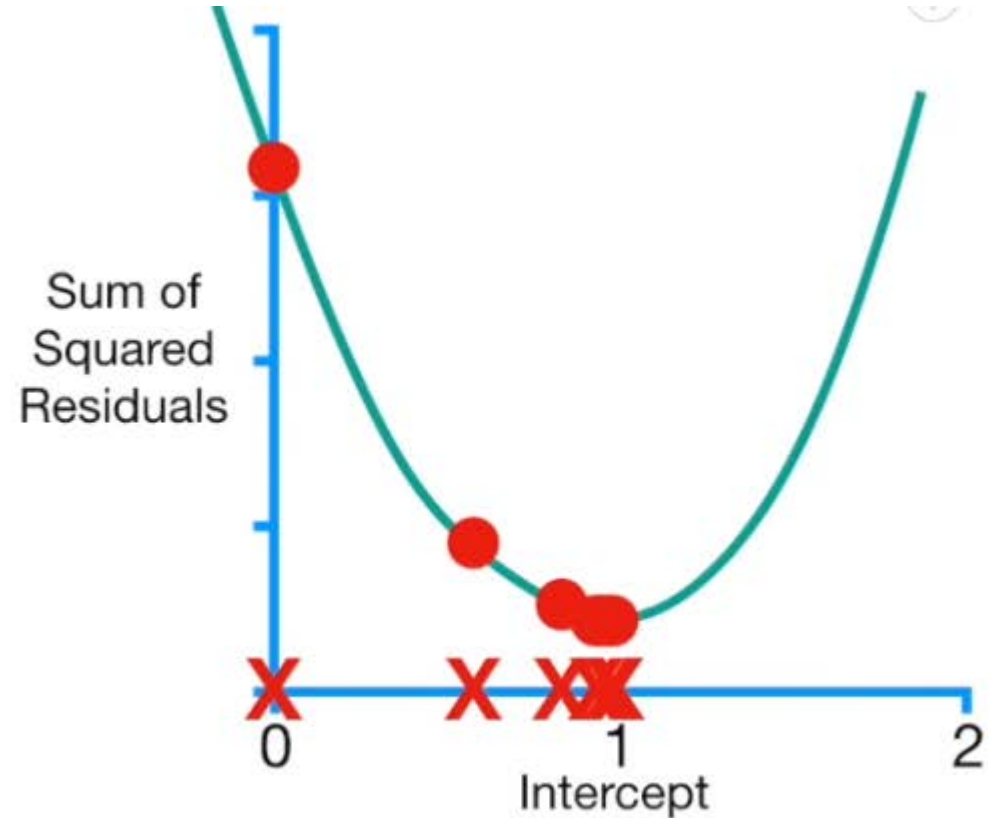
$$\text{Step Size} = \text{Slope} \times \text{Learning Rate}$$



Gradient Descent

That said, **Gradient Descent** also includes a limit on the number of steps it will take before giving up.

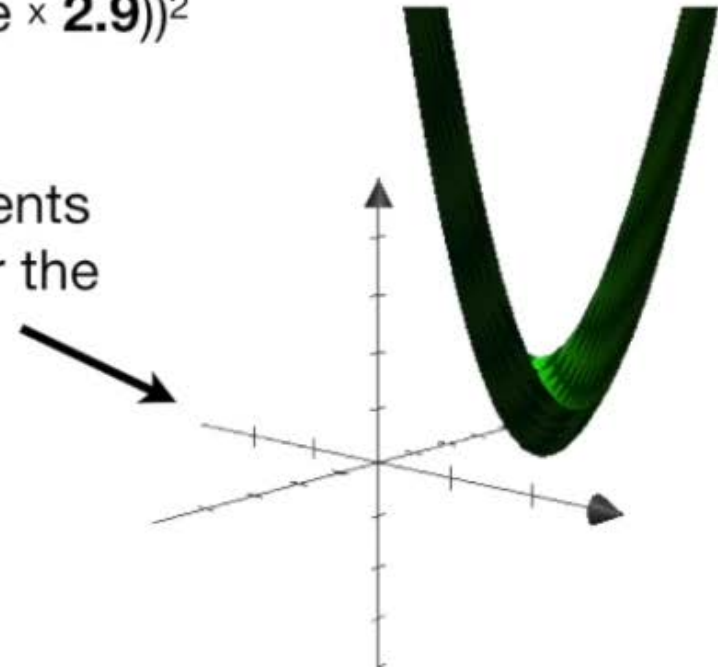
In practice, the **Maximum Number of Steps = 1,000** or greater.



Gradient Descent

$$\begin{aligned}\text{Sum of squared residuals} = & (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2 \\ & + (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2 \\ & + (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2\end{aligned}$$

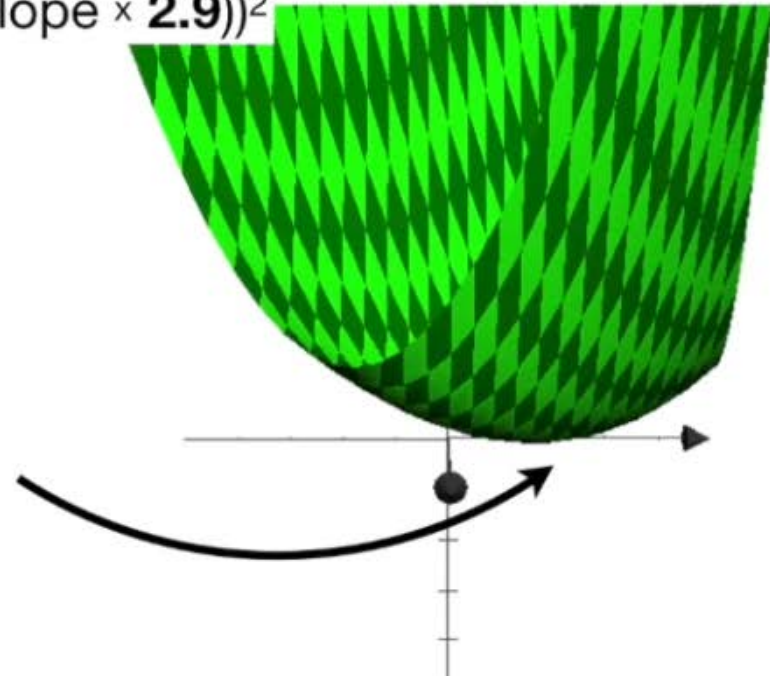
...this axis represents
different values for the
Slope...



Gradient Descent

$$\begin{aligned}\text{Sum of squared residuals} = & (\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2 \\ & + (\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2 \\ & + (\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2\end{aligned}$$


We want to find the values for the **Intercept** and **Slope** that give us the minimum Sum of the Squared Residuals.



Gradient Descent


$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})) \\ & + -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3})) \\ & + -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9})) \end{aligned}$$

Here's the derivative of the
Sum of the Squared
Residuals with respect to
the **Intercept**...



$$\frac{d}{d \text{ slope}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})) \\ & + -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2 \\ & + -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2 \end{aligned}$$

...and here's the derivative
with respect to the **Slope**.



Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (\text{intercept} + \text{slope} \times 0.5)) \\ & + -2(1.9 - (\text{intercept} + \text{slope} \times 2.3)) \\ & + -2(3.2 - (\text{intercept} + \text{slope} \times 2.9)) \end{aligned}$$


Now let's plug in **0** for the **Intercept** and **1** for the **Slope**...

$$\frac{d}{d \text{ slope}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2 \times 0.5(1.4 - (\text{intercept} + \text{slope} \times 0.5)) \\ & + -2 \times 2.9(3.2 - (\text{intercept} + \text{slope} \times 2.9))^2 \\ & + -2 \times 2.3(1.9 - (\text{intercept} + \text{slope} \times 2.3))^2 \end{aligned}$$

Gradient Descent

$$\frac{d}{d \text{ intercept}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2(1.4 - (0 + 1 \times 0.5)) \\ & + -2(1.9 - (0 + 1 \times 2.3)) \\ & + -2(3.2 - (0 + 1 \times 2.9)) \end{aligned} = -1.6$$


Step Size_{Intercept} = -1.6 × **Learning Rate**



...now we plug the
Slopes into the **Step
Size** formulas...

$$\frac{d}{d \text{ slope}} \text{ Sum of squared residuals} =$$
$$\begin{aligned} & -2 \times 0.5(1.4 - (0 + 1 \times 0.5)) \\ & + -2 \times 2.9(3.2 - (0 + 1 \times 2.9))^2 \\ & + -2 \times 2.3(1.9 - (0 + 1 \times 2.3))^2 \end{aligned} = -0.8$$

Step Size_{Slope} = -0.8 × **Learning Rate**



Gradient Descent

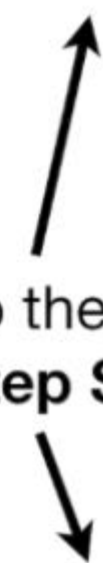
Step Size_{Intercept} = -1.6×0.01

NOTE: The larger **Learning Rate** that we used in the first example doesn't work this time. Even after a bunch of steps, **Gradient Descent** doesn't arrive at the correct answer.

Step Size_{Slope} = -0.8×0.01

Step Size_{Intercept} = $-1.6 \times 0.01 = -0.016$

Anyway, we do the math and get two **Step Sizes**.



Step Size_{Slope} = $-0.8 \times 0.01 = -0.008$

This means that **Gradient Descent** can be very sensitive to the **Learning Rate**.

Gradient Descent

$$\text{Step Size}_{\text{Intercept}} = -1.6 \times 0.01 = \boxed{-0.016}$$

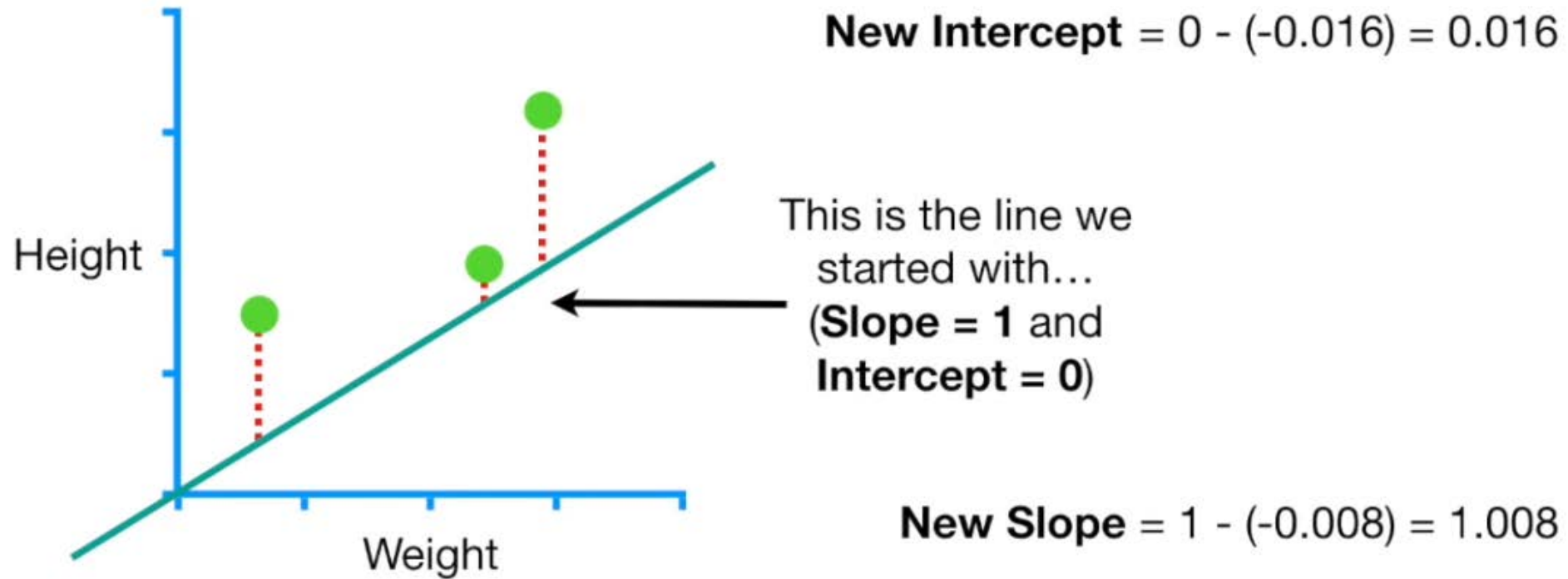
$$\text{New Intercept} = 0 - (-0.016) \leftarrow$$

...and the
Step Sizes...

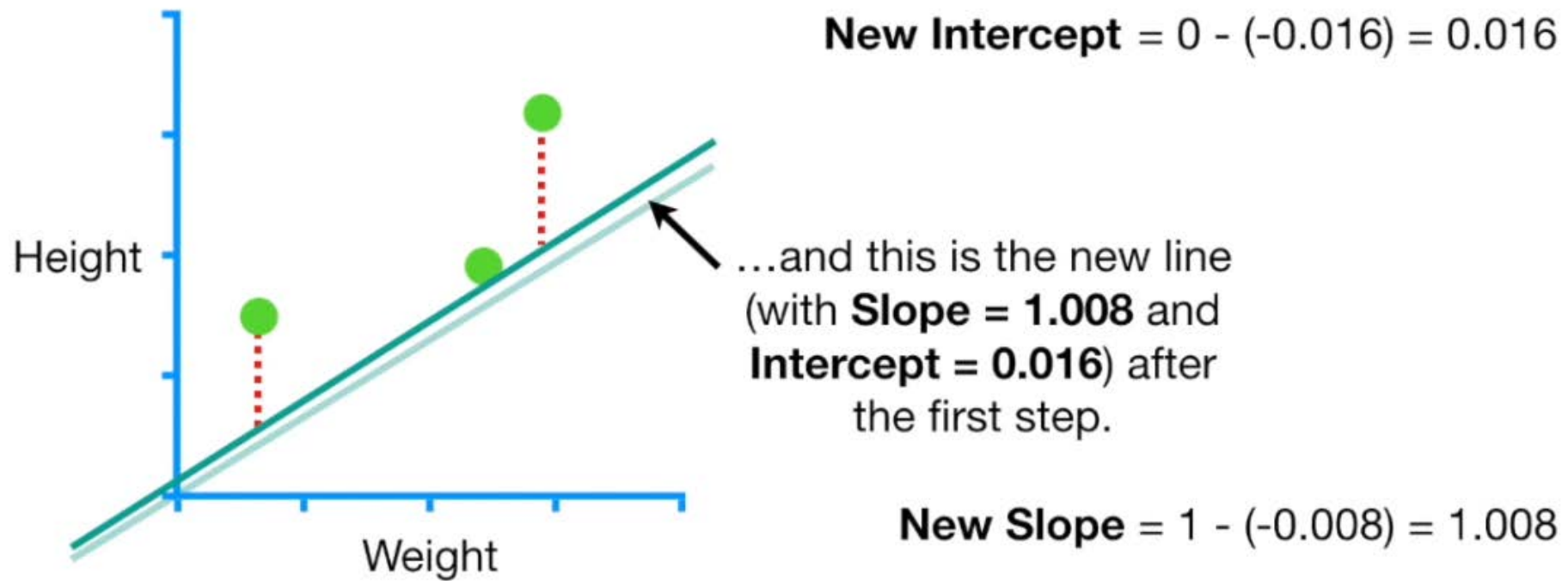
$$\text{Step Size}_{\text{Slope}} = -0.8 \times 0.01 = \boxed{-0.008}$$

$$\text{New Slope} = 1 - (-0.008) \leftarrow$$

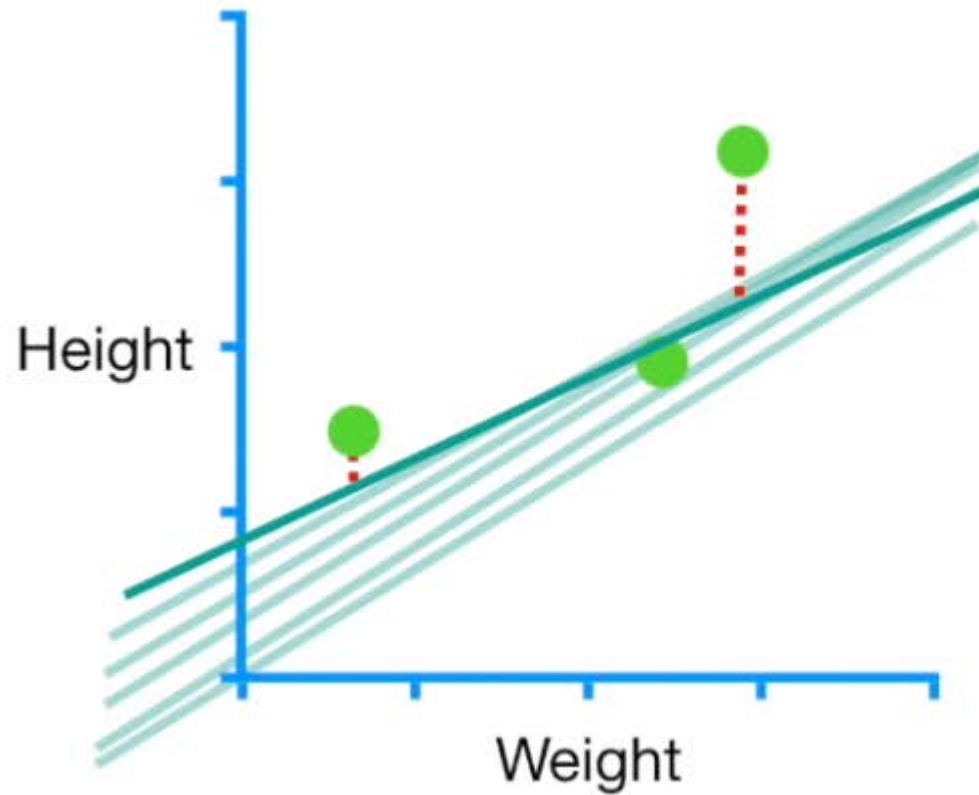
Gradient Descent



Gradient Descent

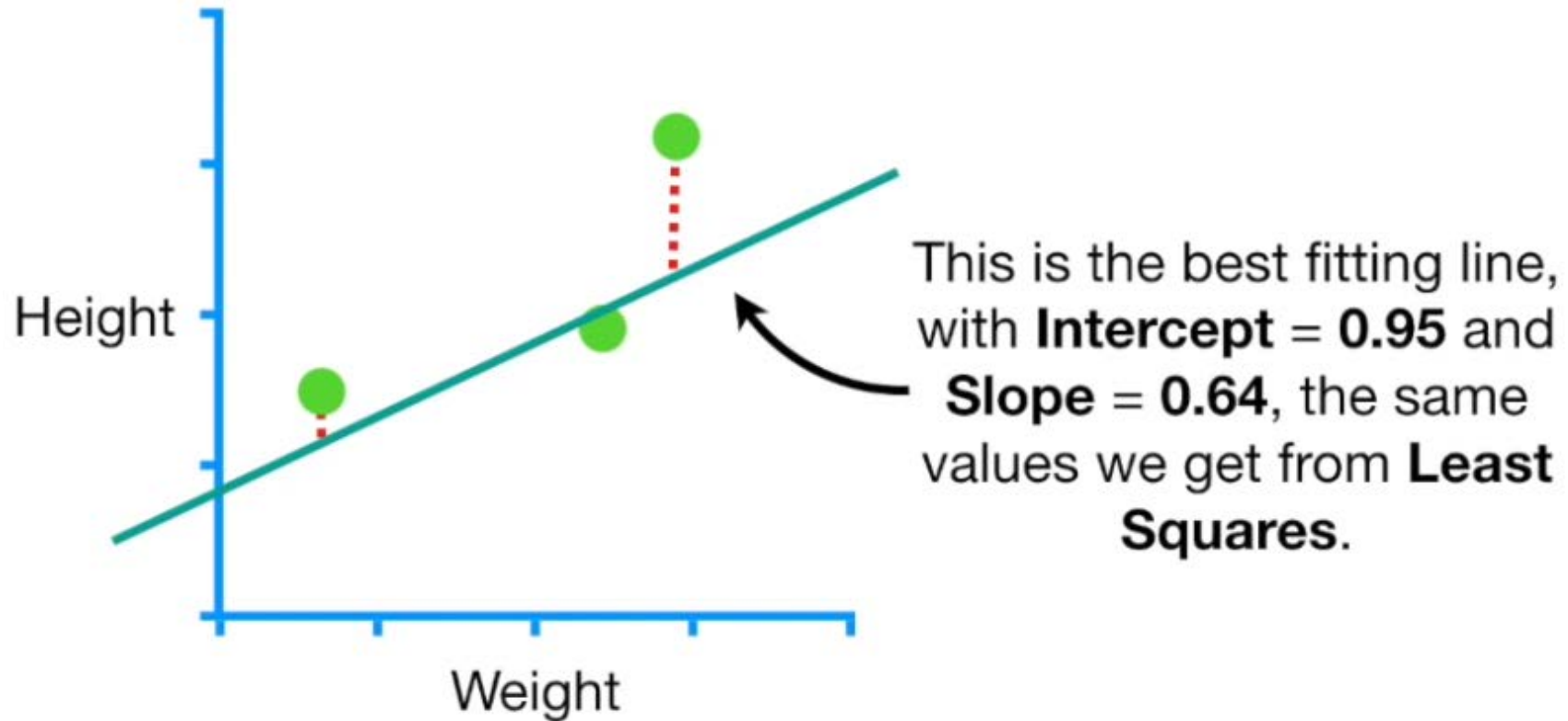


Gradient Descent

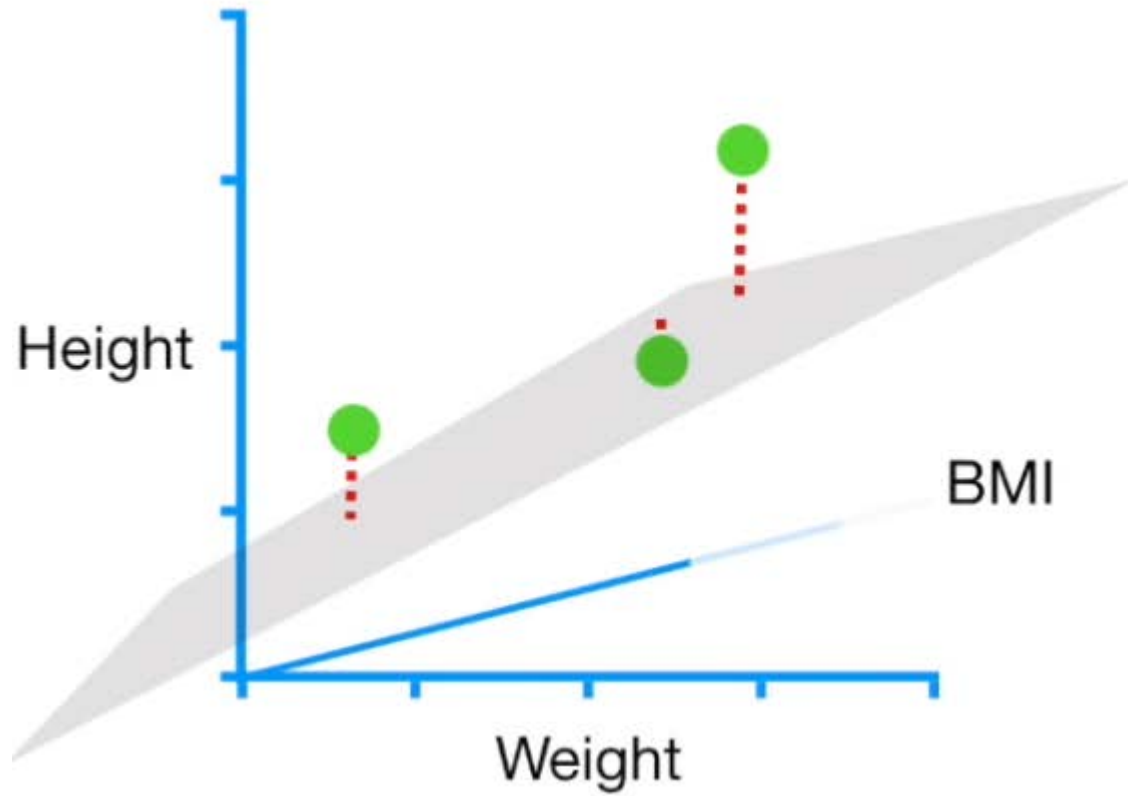


Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

Gradient Descent



Gradient Descent



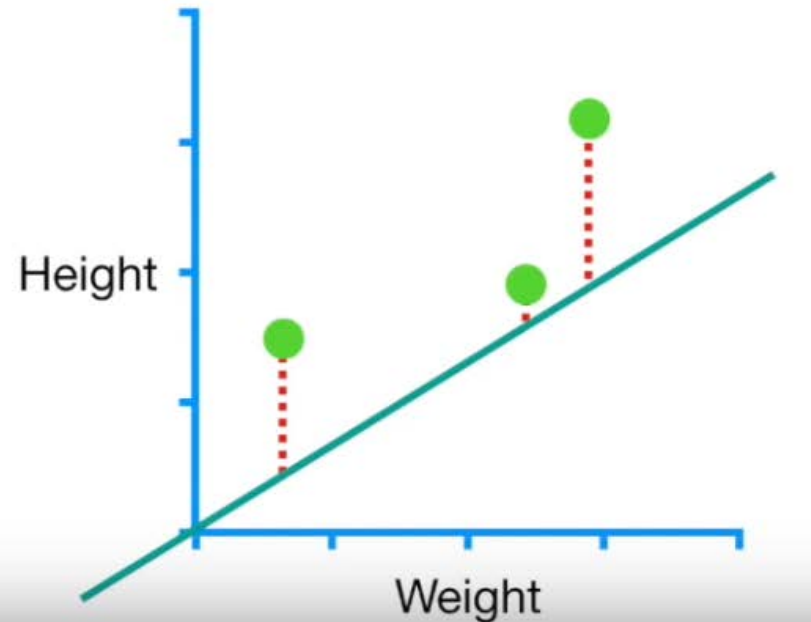
If we had more parameters, then we'd just take more derivatives and everything else stays the same.

Gradient Descent

$$\begin{aligned}\text{Sum of squared residuals} &= (\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2 \\ &\quad + (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2 \\ &\quad + (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2\end{aligned}$$

However, there are tons of other **Loss Functions** that work with other types of data.

Regardless of which **Loss Function** you use, **Gradient Descent** works the same way.



Gradient Descent

Regression Losses

Mean Square Error/Quadratic Loss/L2 Loss

Mathematical formulation :-

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

Mean Squared Error

Mean Absolute Error/L1 Loss

Mathematical formulation :-

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

Mean absolute error

Mathematical formulation :-

$$MBE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n}$$

Mean bias error

Mathematical formulation :-

$$SVM Loss = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

SVM Loss or Hinge Loss

Mathematical formulation :-

$$CrossEntropyLoss = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$

Cross entropy loss

Steps in Gradient Descent

Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Step 2: Pick random values for the parameters.

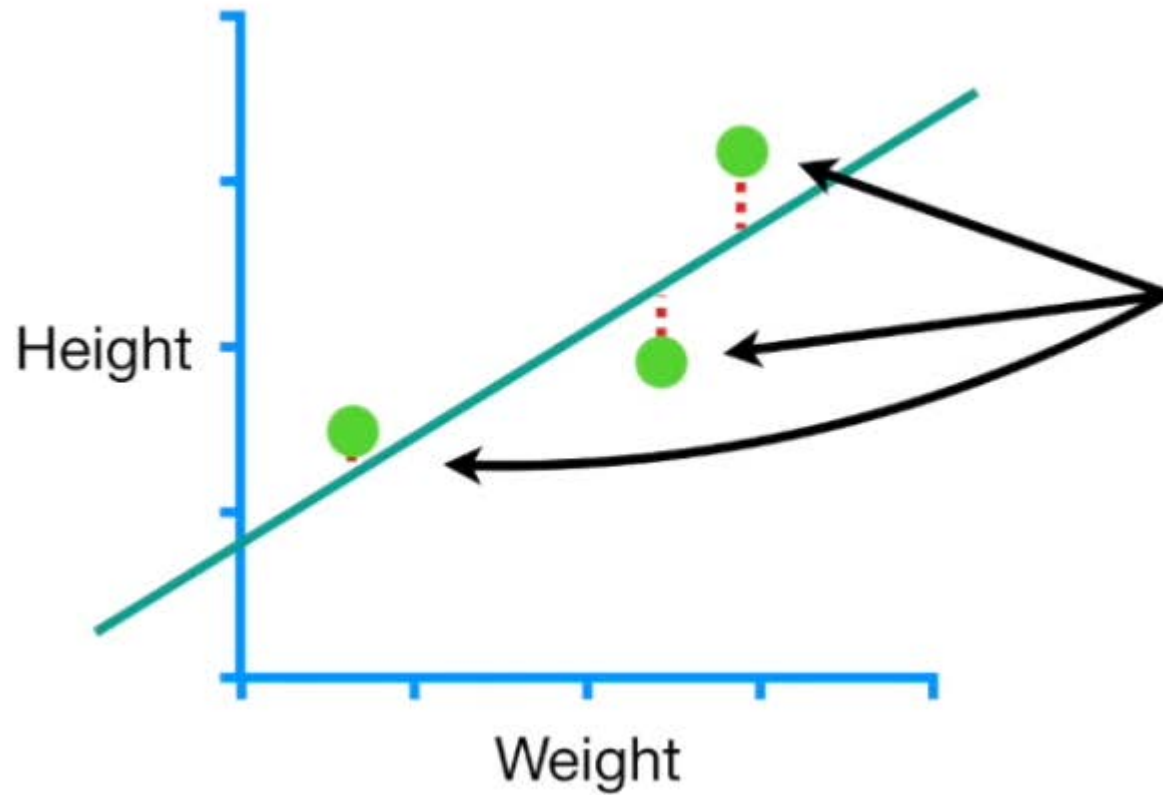
Step 3: Plug the parameter values into the derivatives (ahem, the **Gradient**).

Step 4: Calculate the Step Sizes: **Step Size** = **Slope** × **Learning Rate**

Step 5: Calculate the New Parameters:

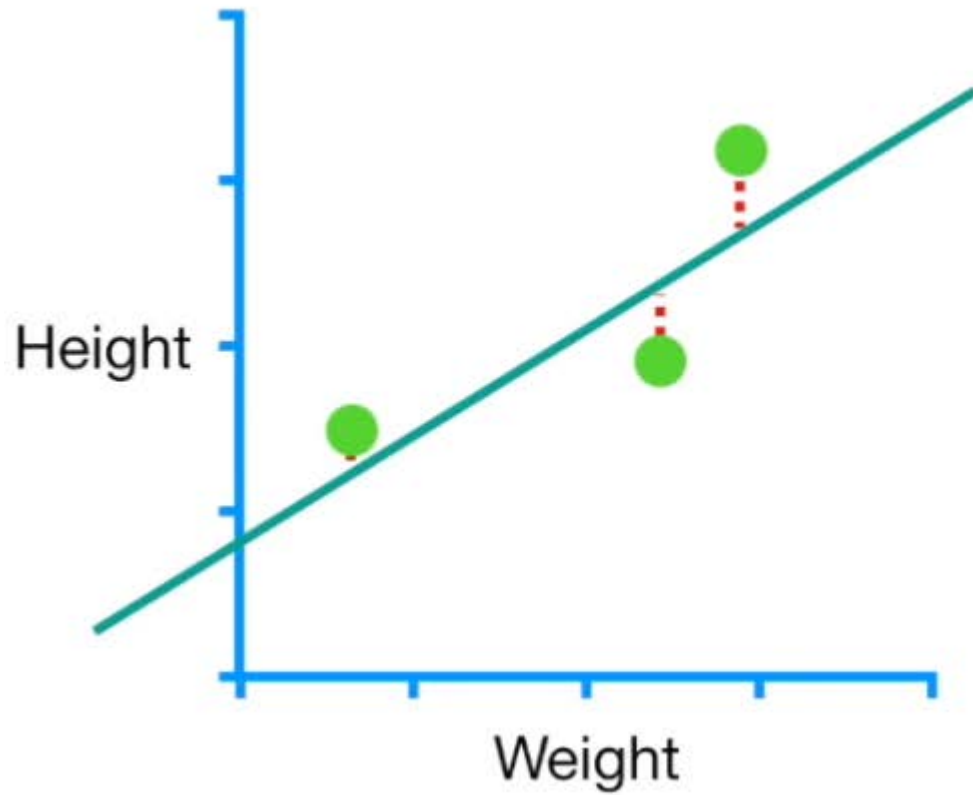
$$\text{New Parameter} = \text{Old Parameter} - \text{Step Size}$$

Gradient Descent



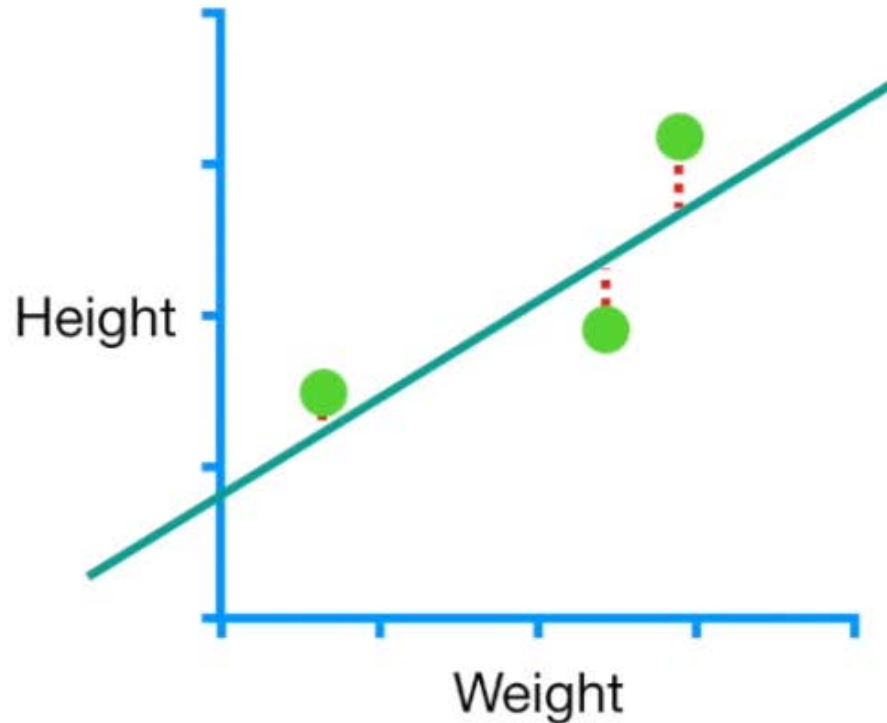
In our example, we only had three data points, so the math didn't take very long...

Gradient Descent



...but when you have millions of data points, it can take a long time.

Stochastic Gradient Descent (SGD)



So there is a thing called **Stochastic Gradient Descent** that uses a randomly selected subset of the data at every step rather than the full dataset.