# Gradient Descent

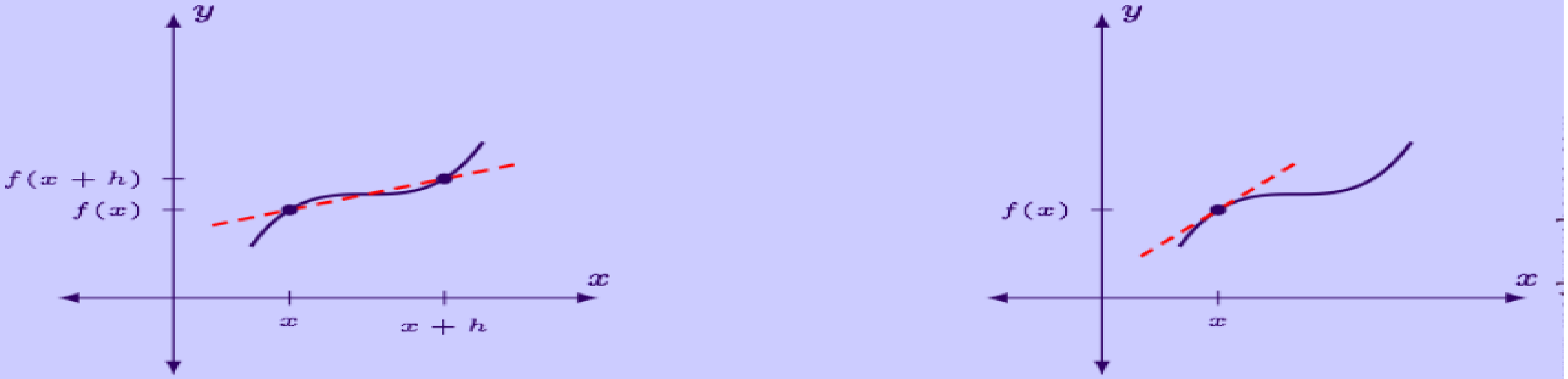
Dr. Md. Golam Rabiul Alam

# Derivative

Recall that the definition of the derivative is

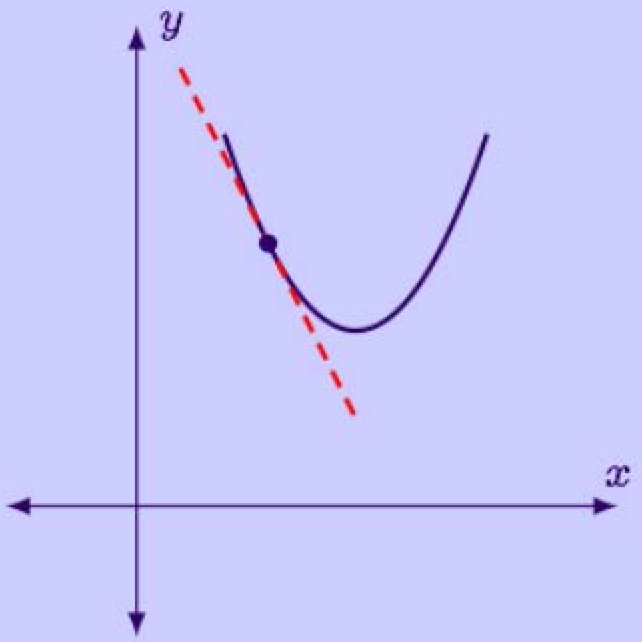
$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{(x+h) - x}.$$

Without the limit, this fraction computes the slope of the line connecting two points on the function (see the left-hand graph below).

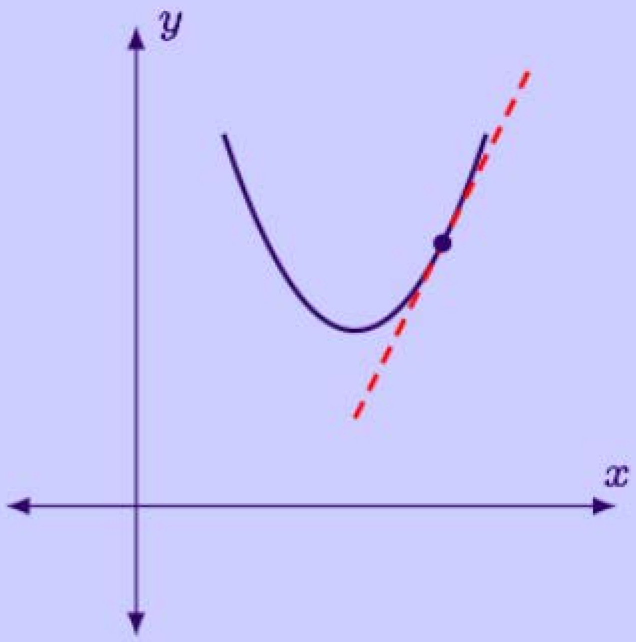



The only thing the limit does is to move the two points closer to each other until they are right on top of each other. But the fundamental calculation is still a slope. So the end result is the slope of the line that is tangent to the curve at the point $(x, f(x))$.

# Derivative




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

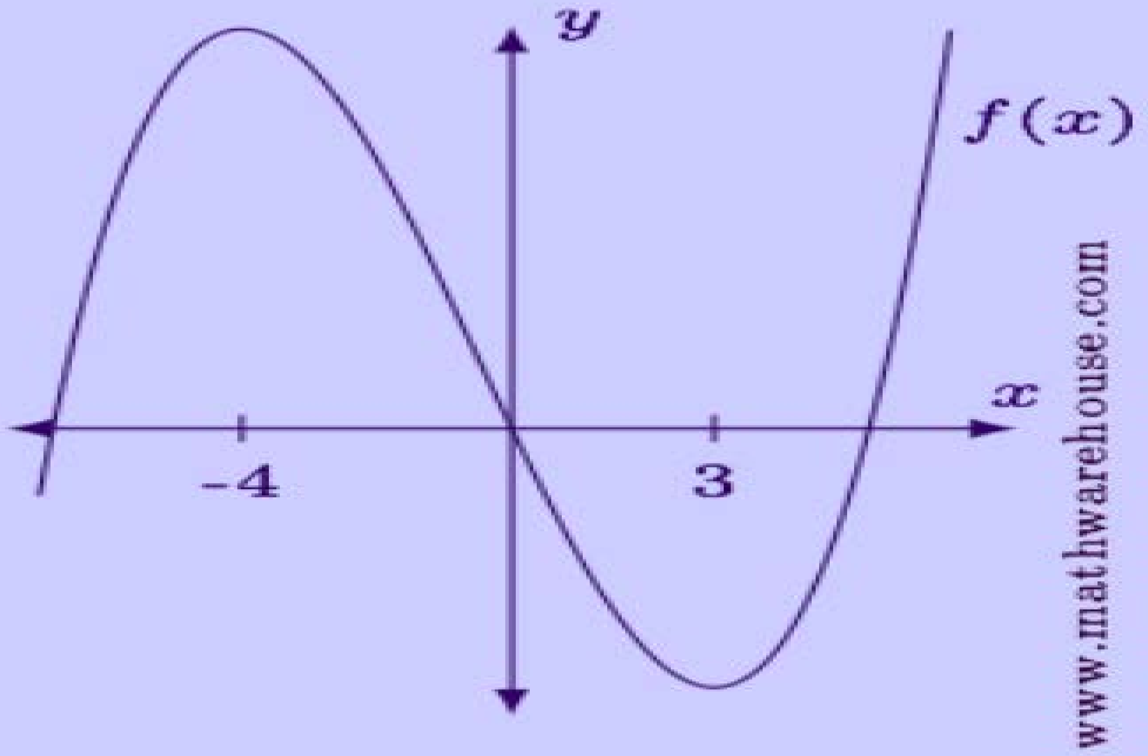



When the tangent line has a positive slope, the function is increasing at that point.

# Derivative

Suppose $f(x) = 2x^3 + 3x^2 - 72x$. Determine the intervals over which the function is increasing, and the intervals over which the function is decreasing.

The graph of the function is shown below for reference.



# Derivative

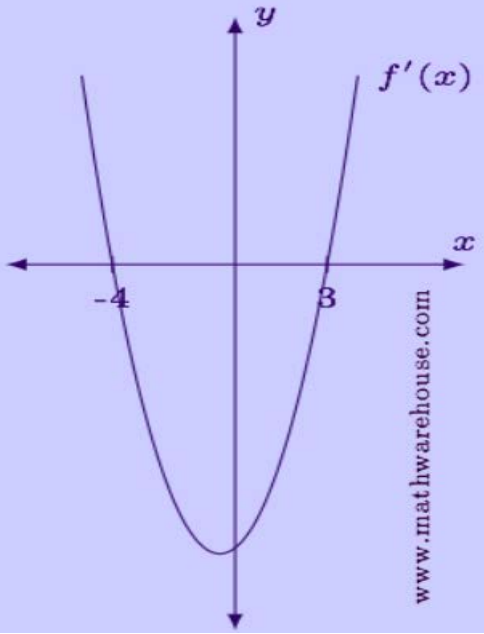
*Step 1*

Find the first derivative.

$$f'(x) = 6x^2 + 6x - 72 = 6(x^2 + x - 12) = 6(x+4)(x-3)$$

*Step 2*

Sketch a quick graph of the derivative.




*Step 3*

Interpret the graph.

We know that when the derivative is positive, the function is increasing. The graph above shows that the derivative is positive (i.e., above the $x$-axis) when $x < -4$ and when $x > 3$.

We can also see that the derivative is negative (below the $x$-axis) when $-4 < x < 3$.

*Answer*

The function is increasing on the intervals from $(-\infty, -4) \cup (3, \infty)$. Likewise, the function is decreasing over the interval $(-4, 3)$.

# Gradient Descent

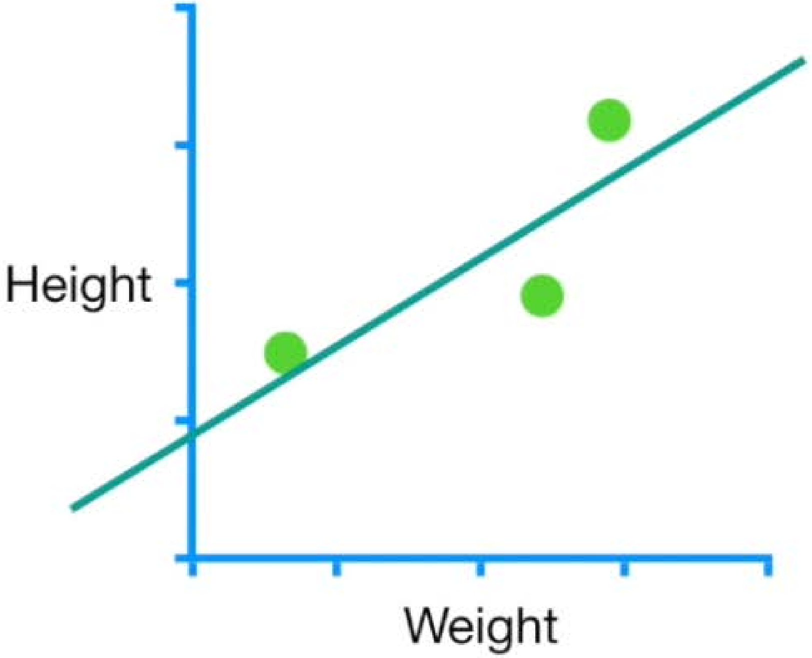
- Two or more derivatives of the same function are called Gradients.

- An algorithm which uses gradient to descent to the lowest point of a loss function.

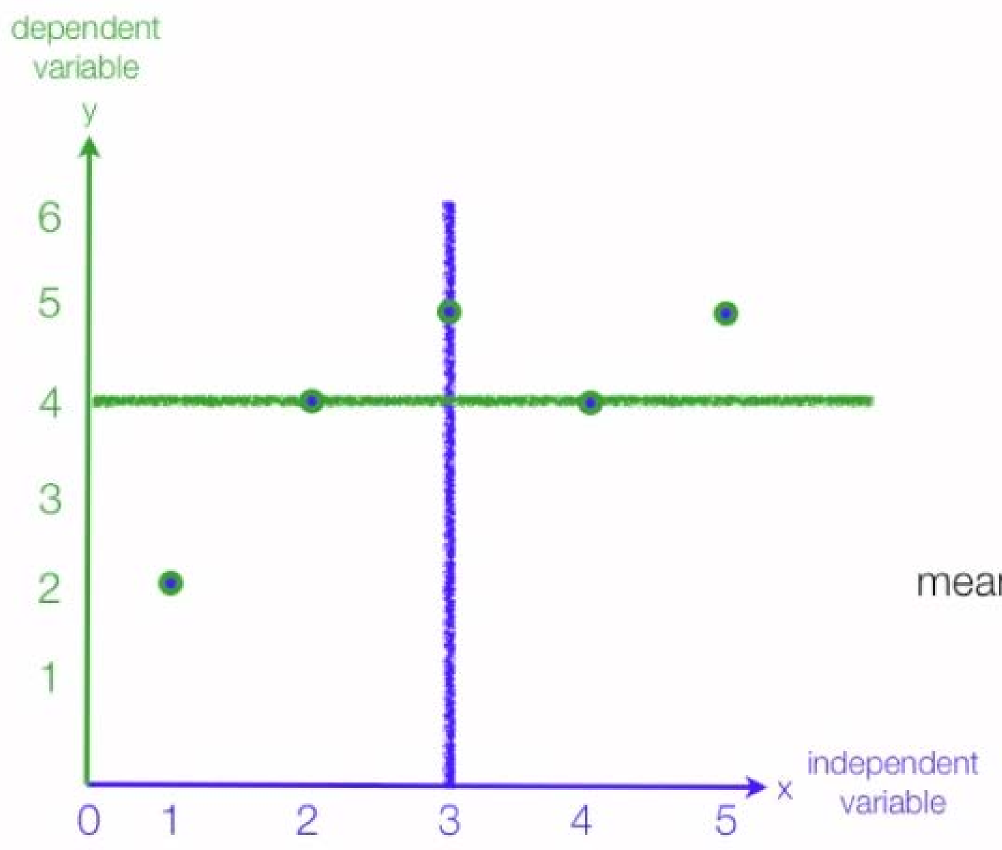
# Gradient Descent

**Predicted Height** = intercept + slope × **Weight**

So let's learn how **Gradient Descent** can fit a line to data by finding the optimal values for the **Intercept** and the **Slope**.

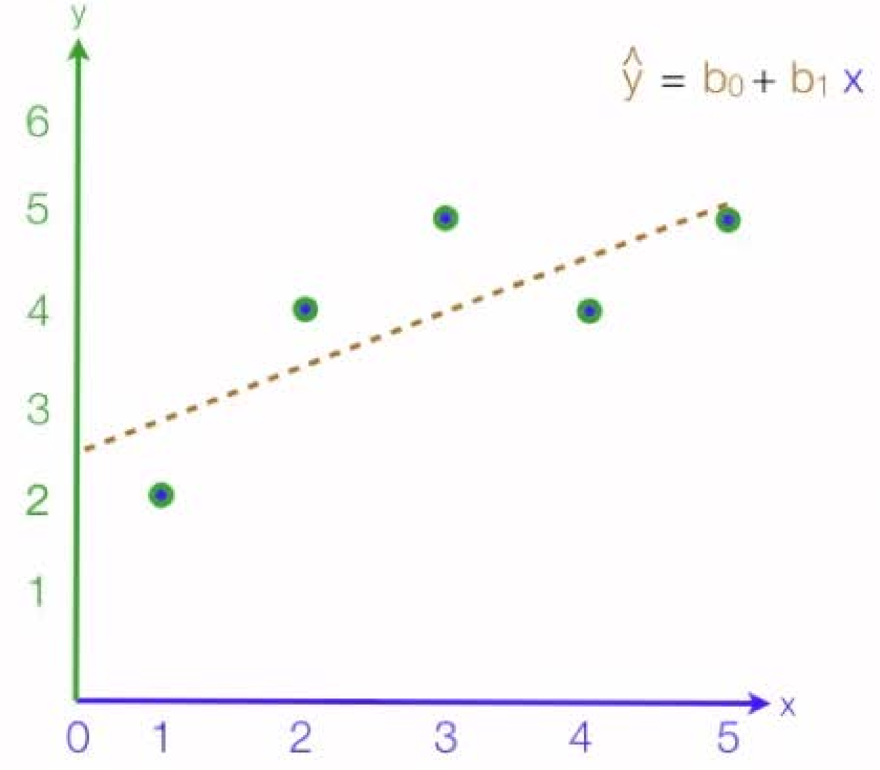


# Least Square Method




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

# Least Square Method

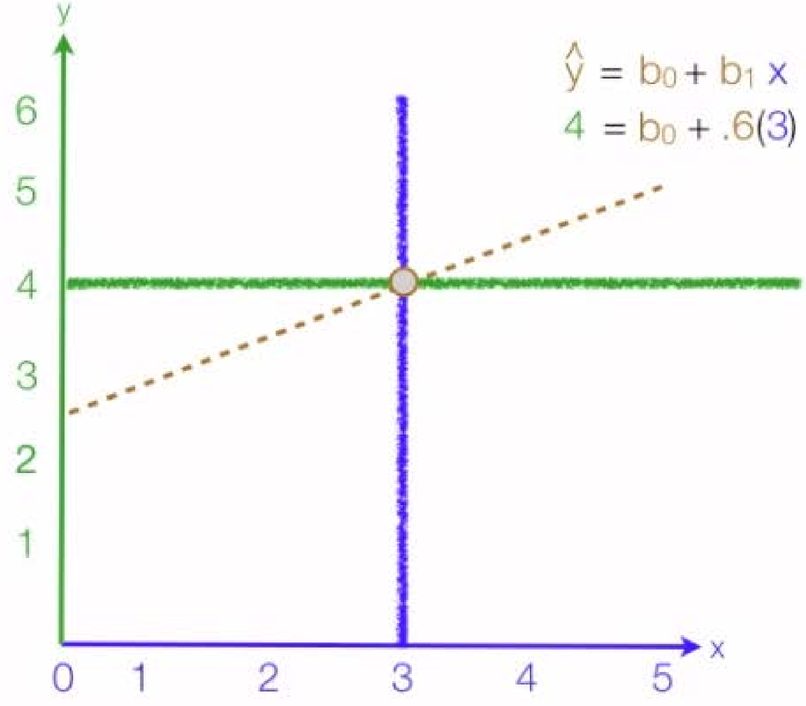



| | $x$ | $y$ | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|---|
| | 1 | 2 | -2 | -2 | 4 | 4 |
| | 2 | 4 | -1 | 0 | 1 | 0 |
| | 3 | 5 | 0 | 1 | 0 | 0 |
| | 4 | 4 | 1 | 0 | 1 | 0 |
| | 5 | 5 | 2 | 1 | 4 | 2 |
| mean | 3 | 4 | | | 10 | 6 |

sum

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

# Least Square Method




$b_0 = 2.2$

$b_1 = .6$

| | $x$ | $y$ | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|---|
| | 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 | | | | |

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

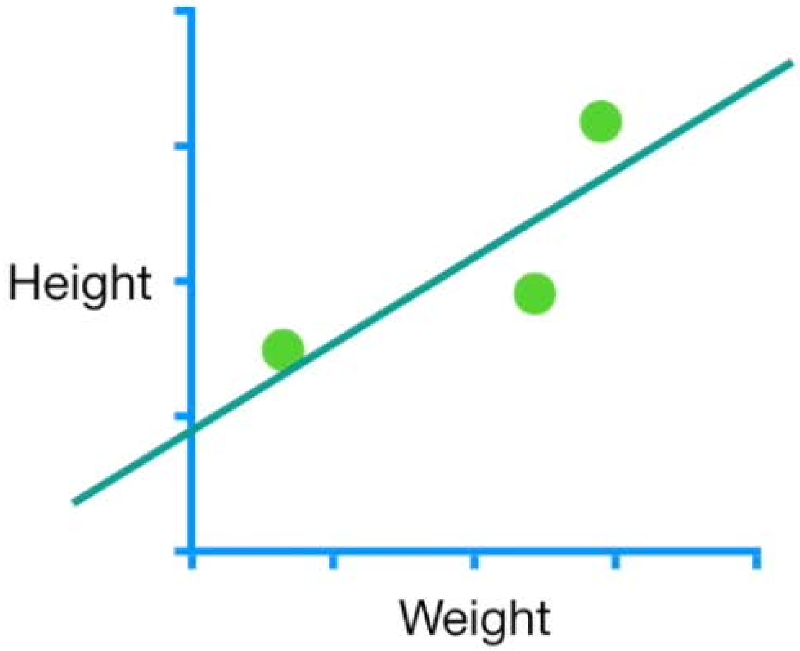
$$\begin{aligned} 4 &= b_0 + \cancel{1.8} \\ -1.8 & \quad\ \ \cancel{-1.8} \\ \hline 2.2 &= b_0 \end{aligned}$$

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

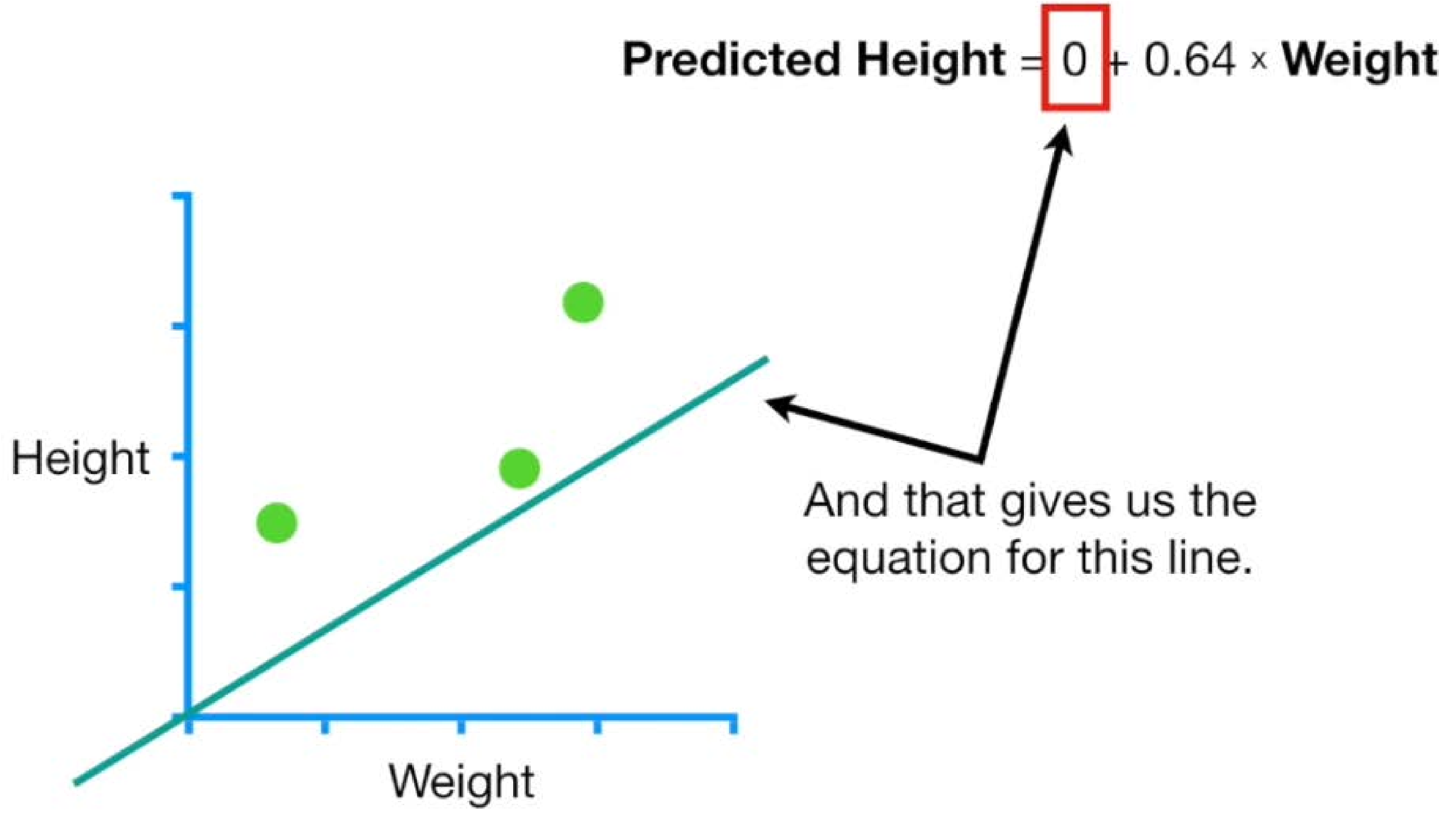
# Gradient Descent

**Predicted Height** = intercept + $\boxed{0.64}$ × **Weight**

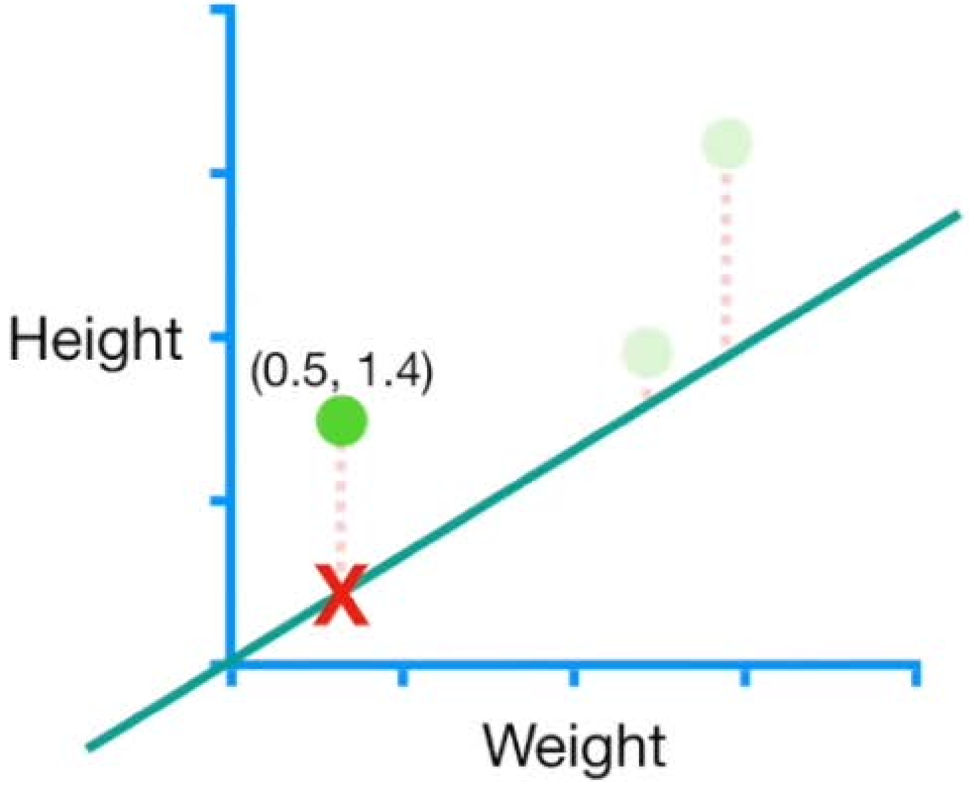
So for now, let's just plug in the **Least Squares** estimate for the **Slope**, **0.64**.



# Gradient Descent



# Gradient Descent

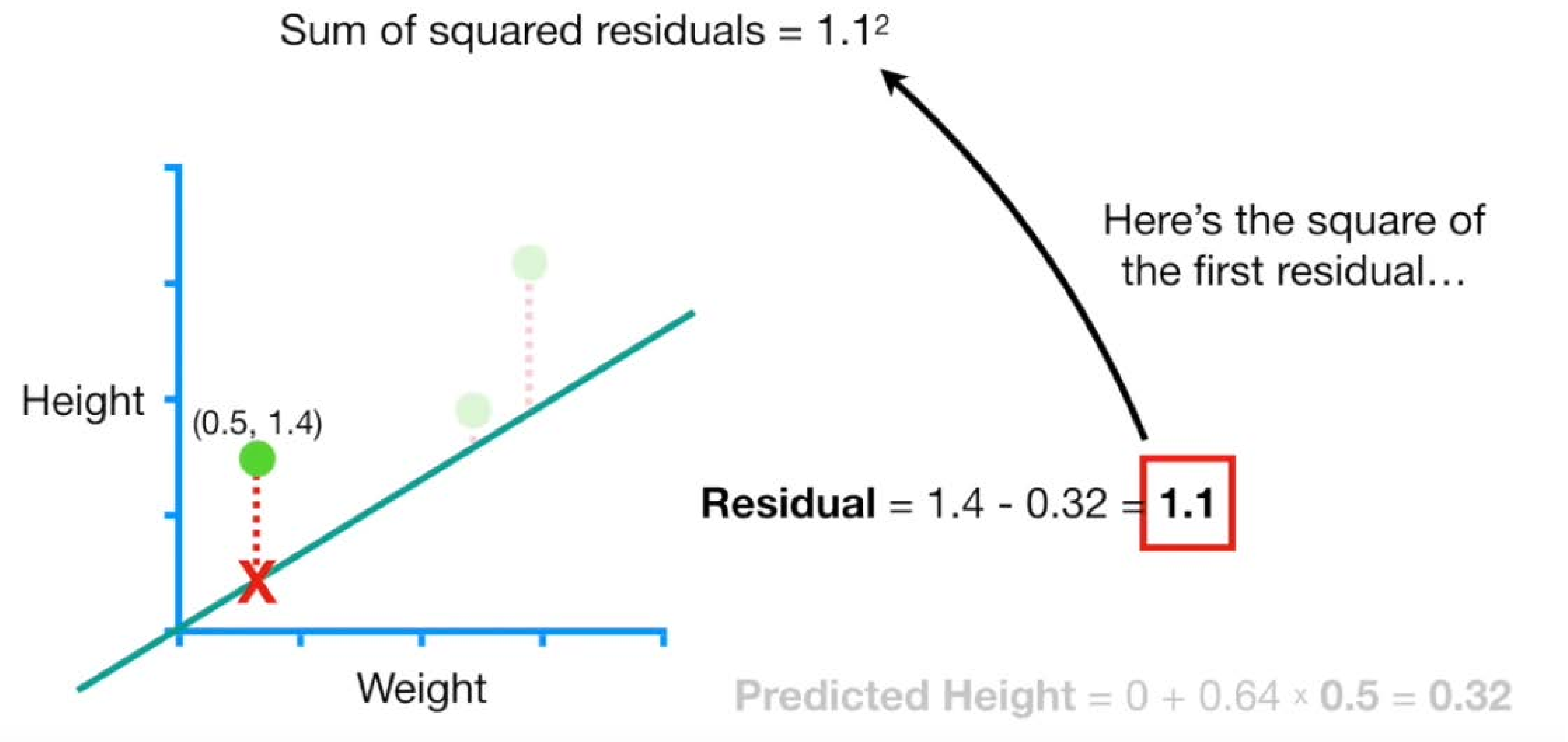



We get the **Predicted Height**, the point on the line…

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

**Predicted Height** = 0 + 0.64 × **0.5**

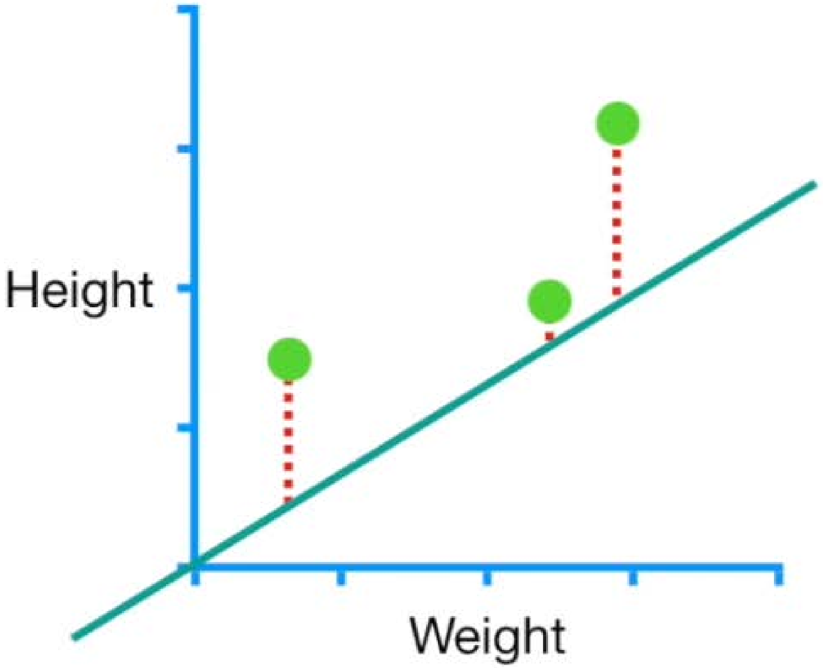
# Gradient Descent



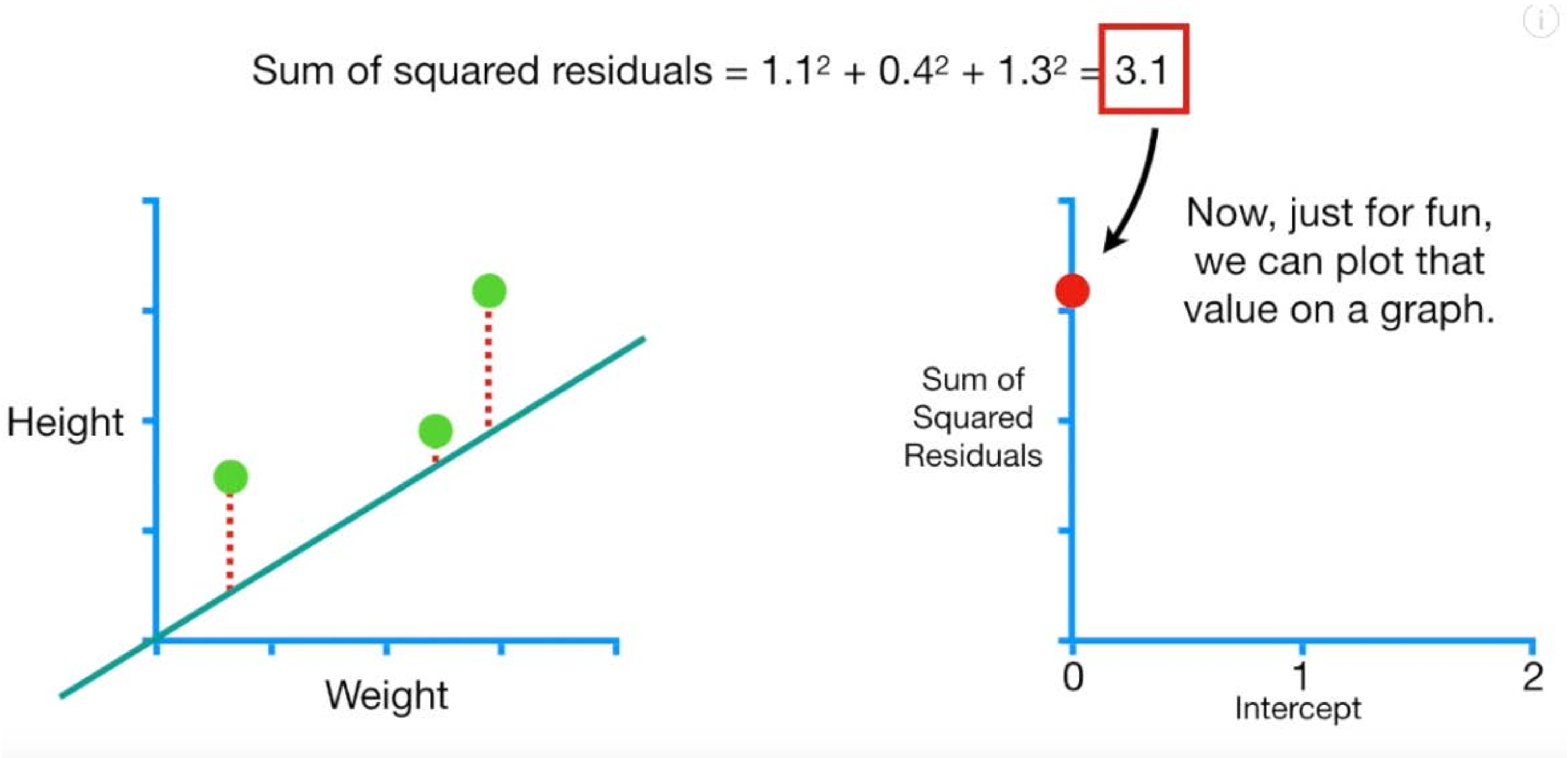
# Gradient Descent

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

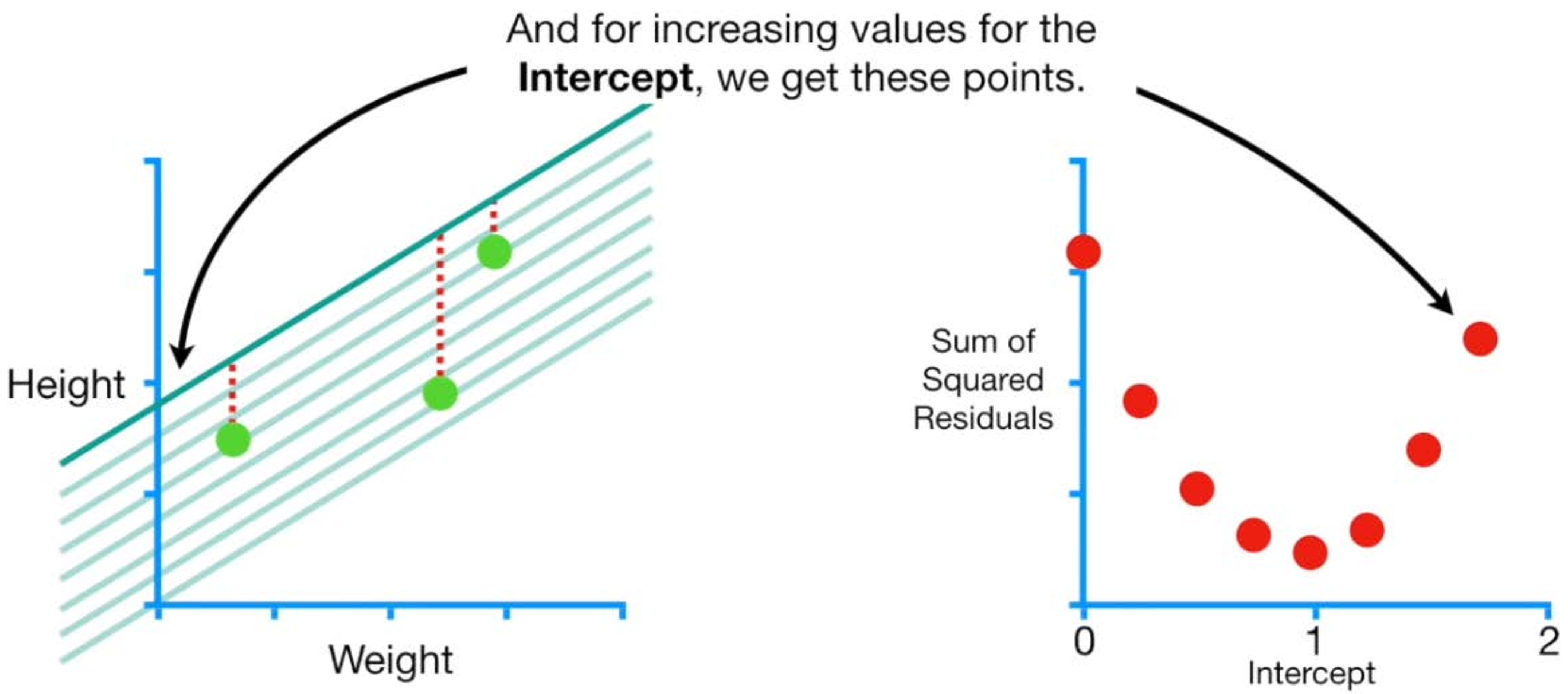
In the end, **3.1** is the Sum of the Squared Residuals.



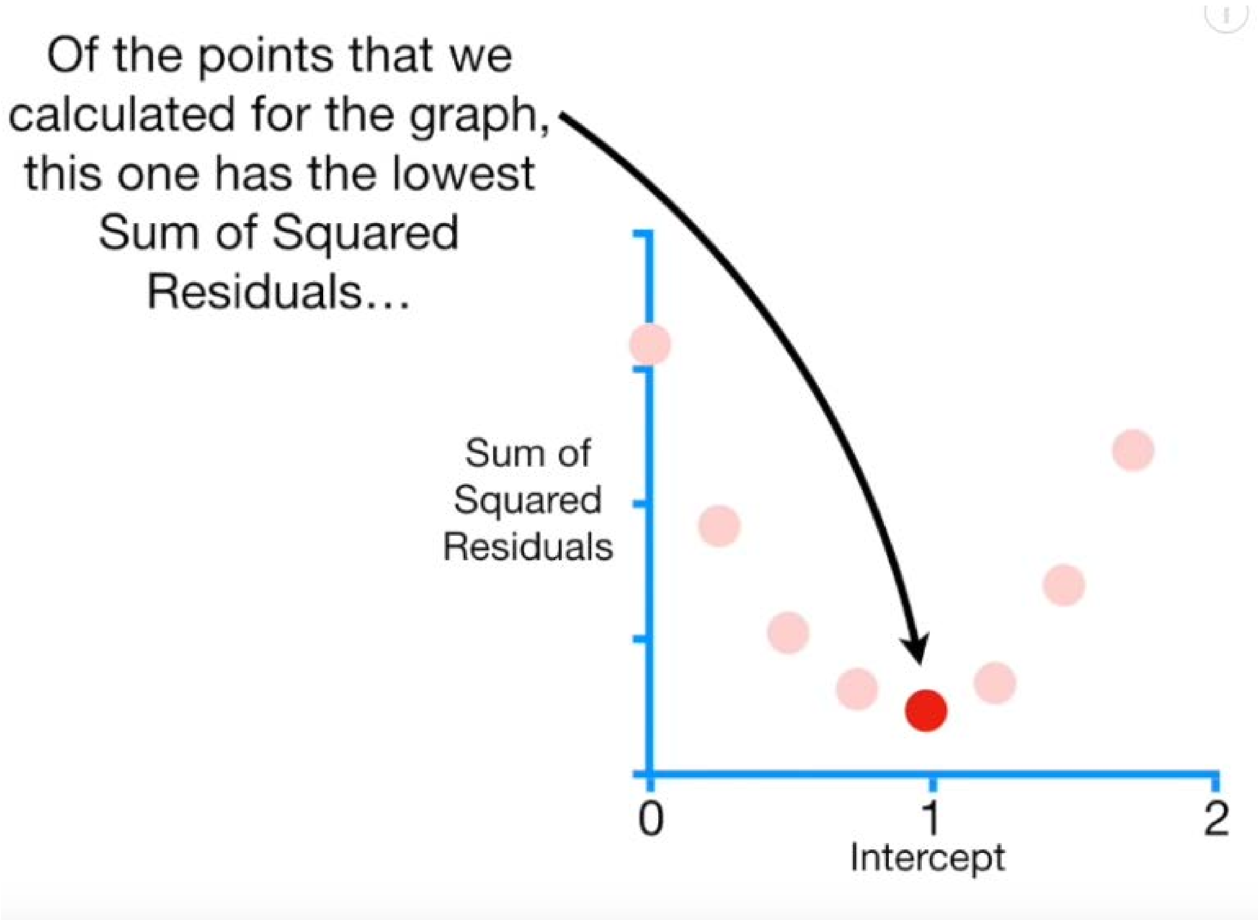
# Gradient Descent



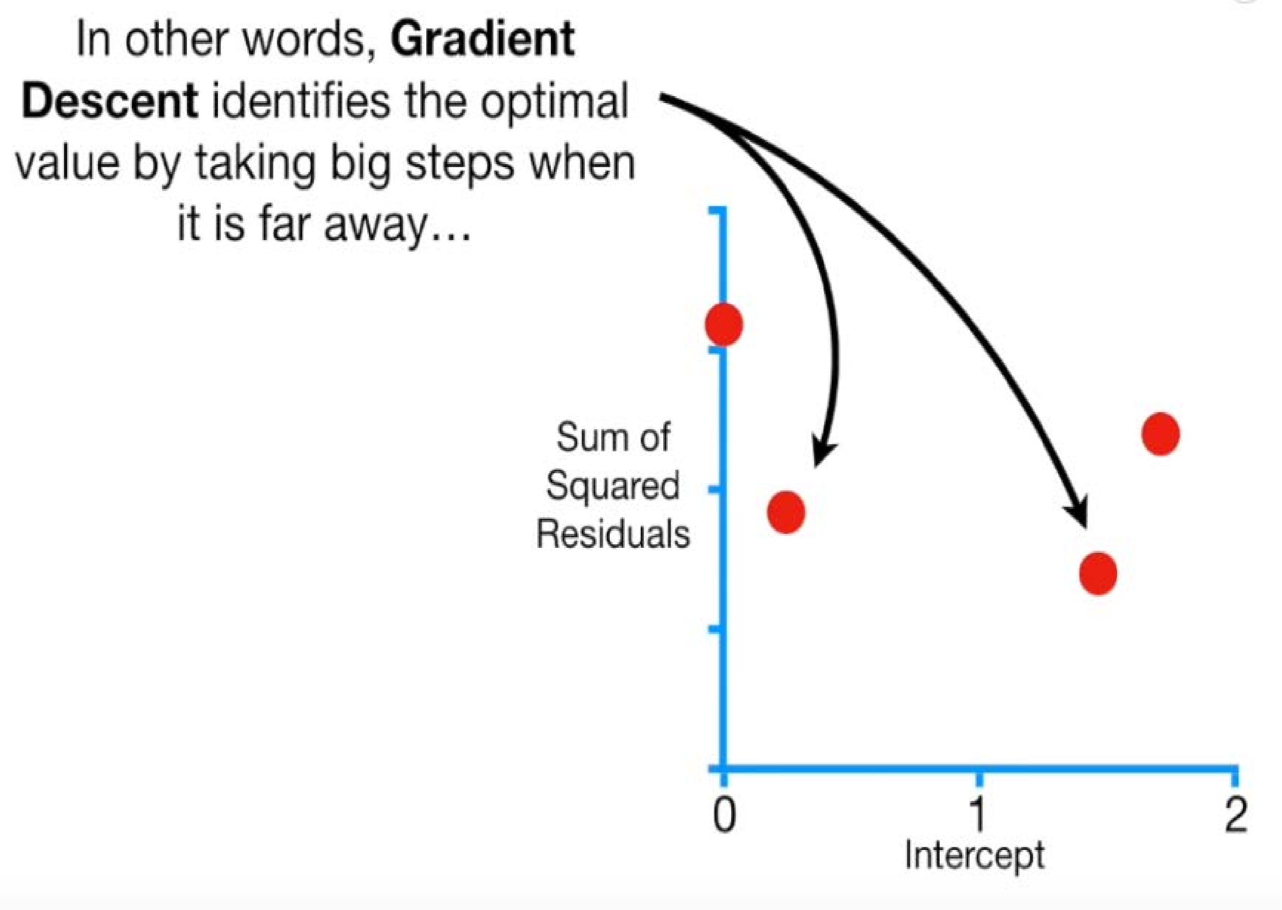
# Gradient Descent

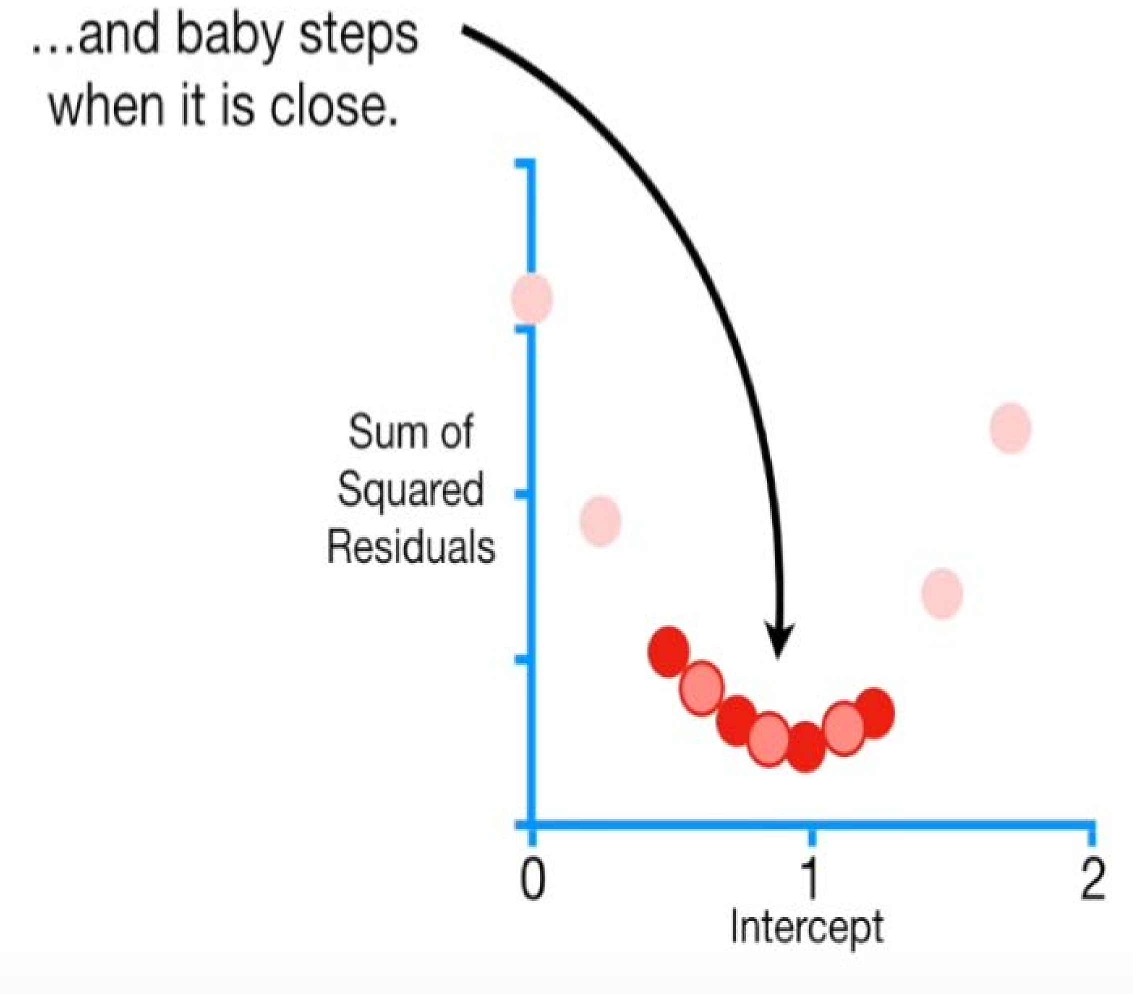


# Gradient Descent

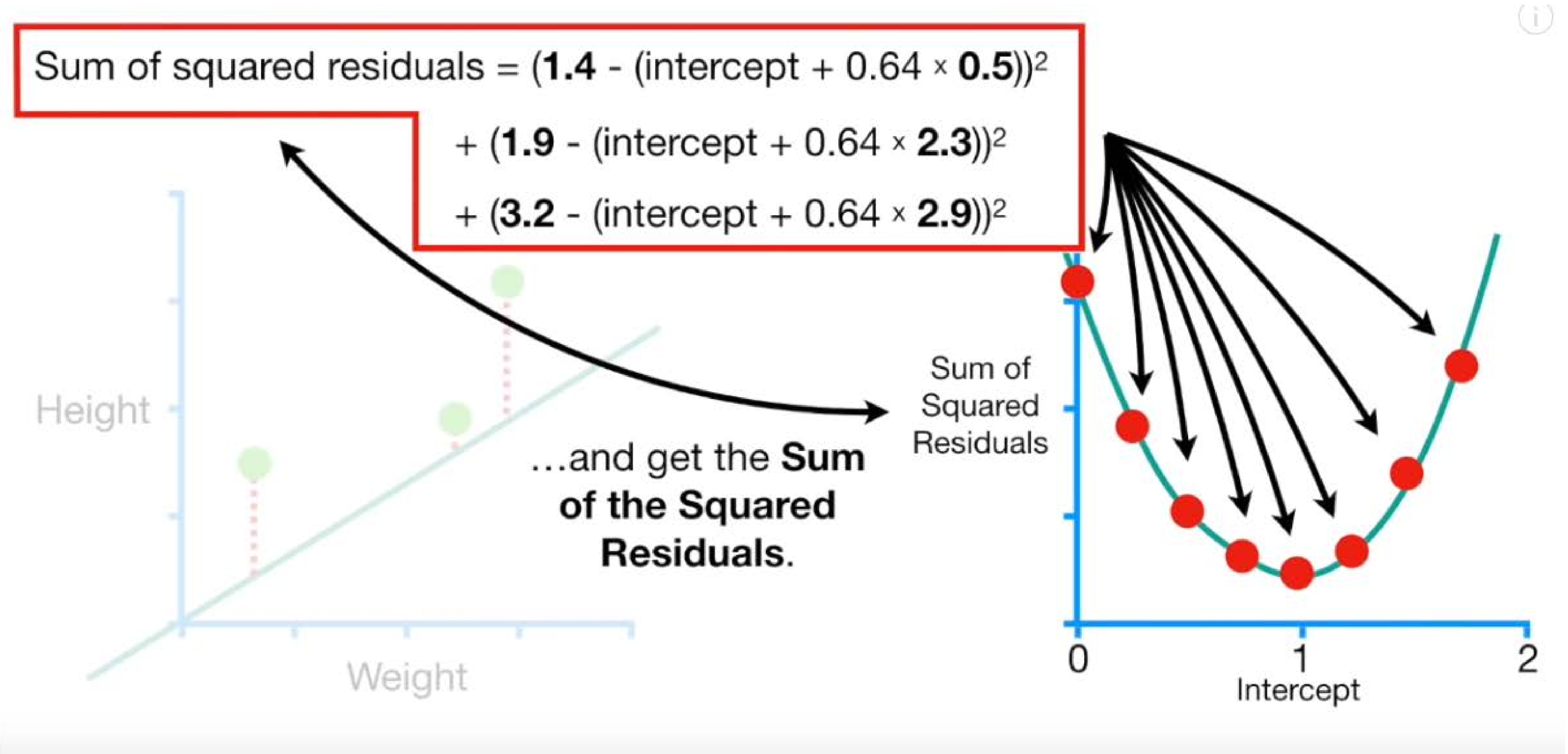


# Gradient Descent



# Gradient Descent



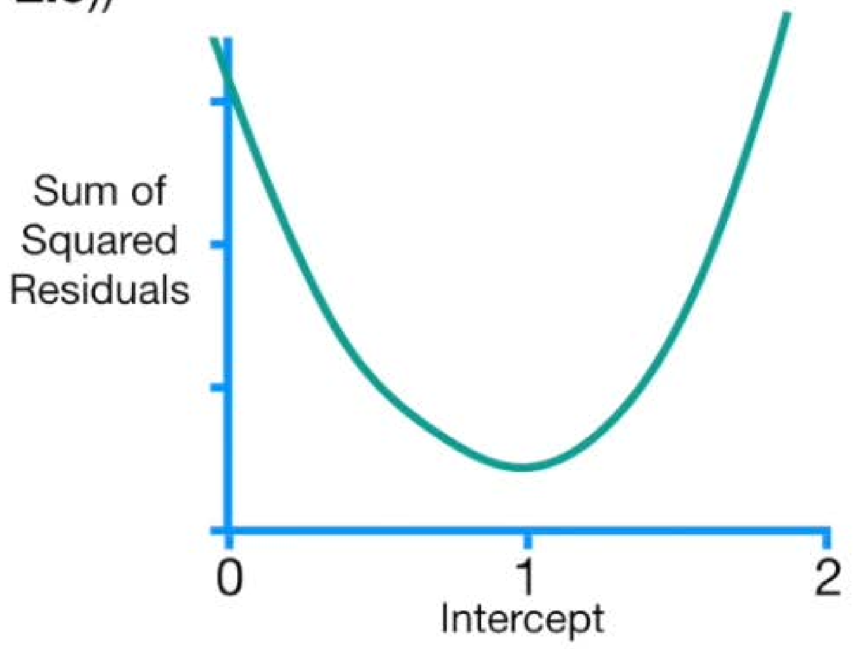
# Gradient Descent

Sum of squared residuals = $(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5}))^2$

$+ (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$

$+ (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$

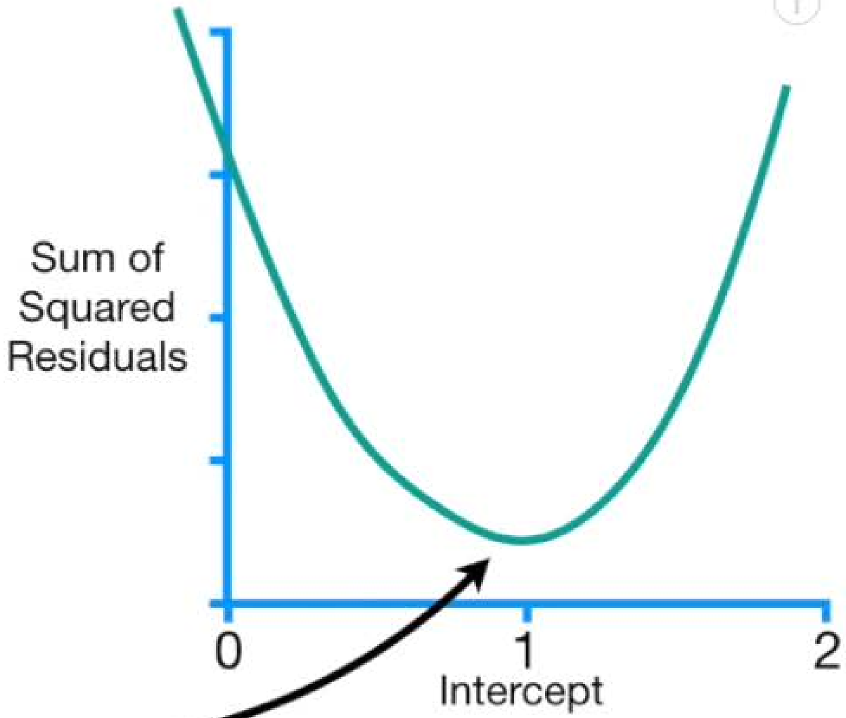
So let's take the derivative of the Sum of the Squared Residuals with respect to the **Intercept**.



# Gradient Descent

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

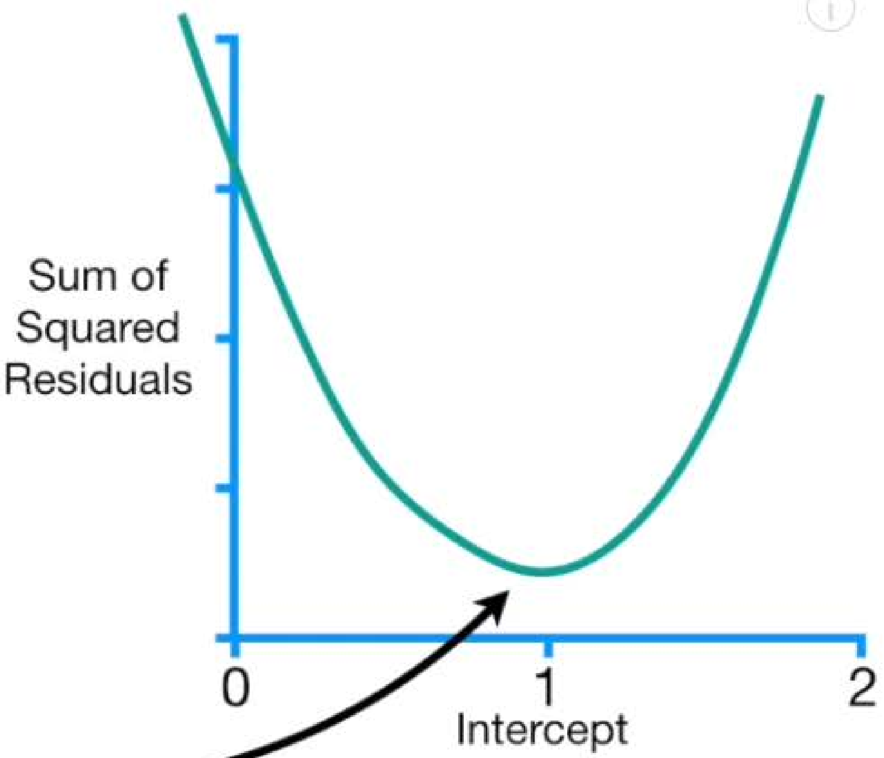
Now that we have the derivative, **Gradient Descent** will use it to find where the Sum of Squared Residuals is lowest.



# Gradient Descent

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

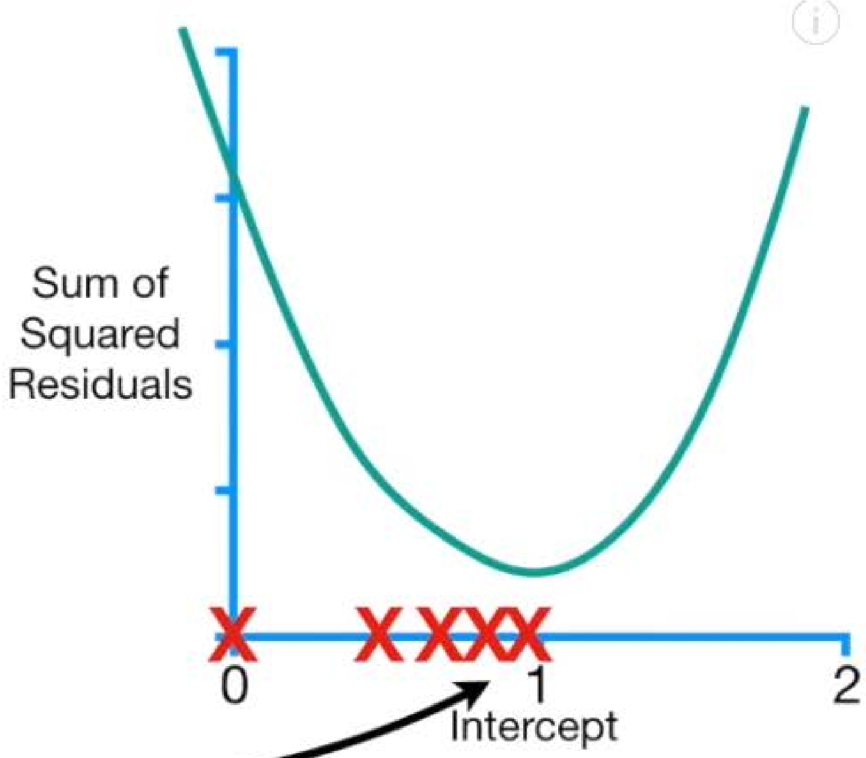
**NOTE:** If we were using **Least Squares** to solve for the optimal value for the **Intercept**, we would simply find where the the slope of the curve = **0**.



# Gradient Descent

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

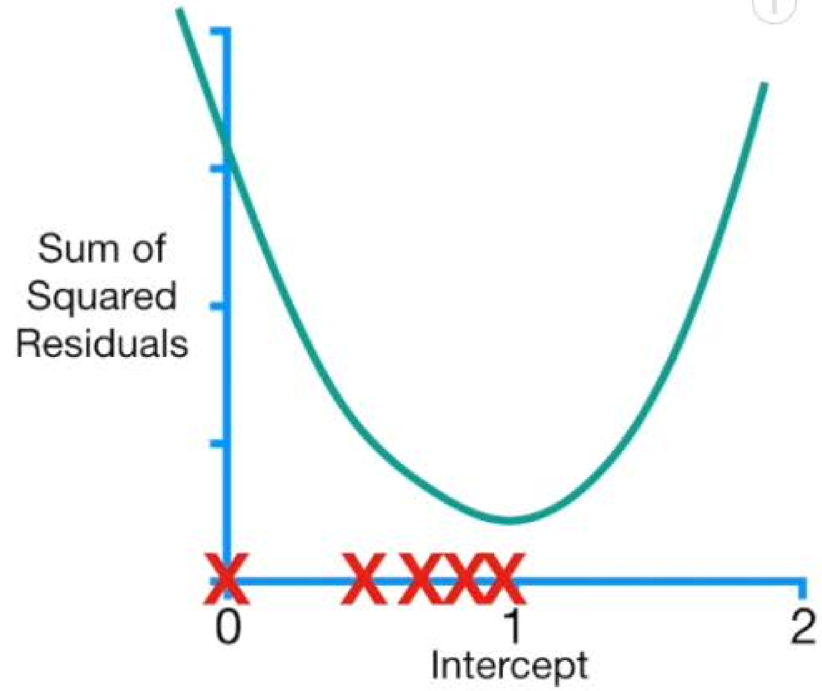
In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.



# Gradient Descent

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

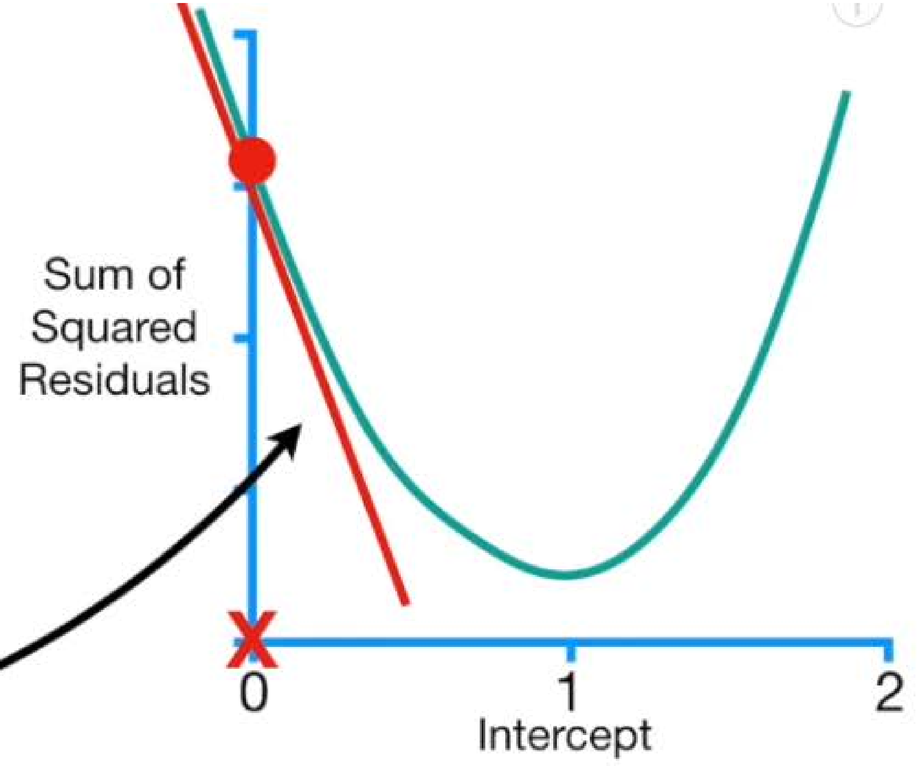
This makes **Gradient Descent** very useful when it is not possible to solve for where the derivative = **0**, and this is why **Gradient Descent** can be used in so many different situations.



# Gradient Descent

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

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



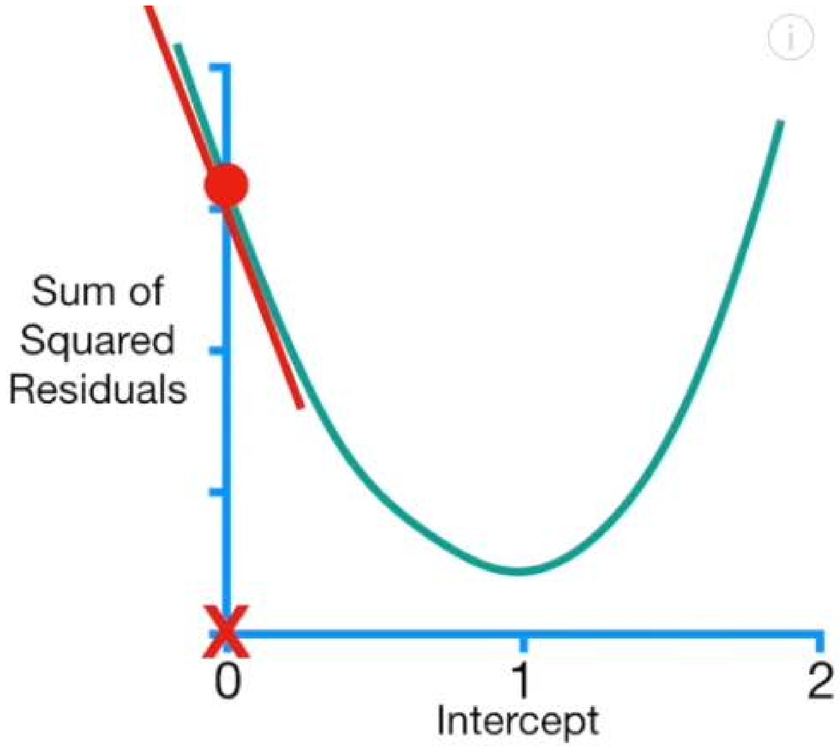
# Gradient Descent

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

$$= -5.7$$

**Step Size** = -5.7 × 0.1

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



# Gradient Descent

$$\frac{d}{d\ intercept} \text{ Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

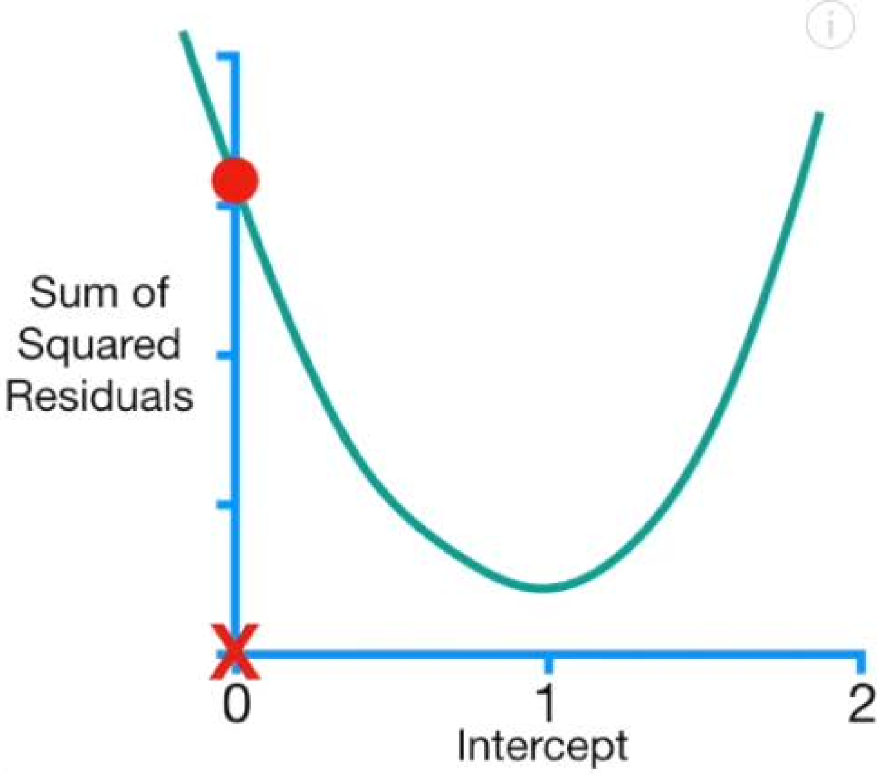
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

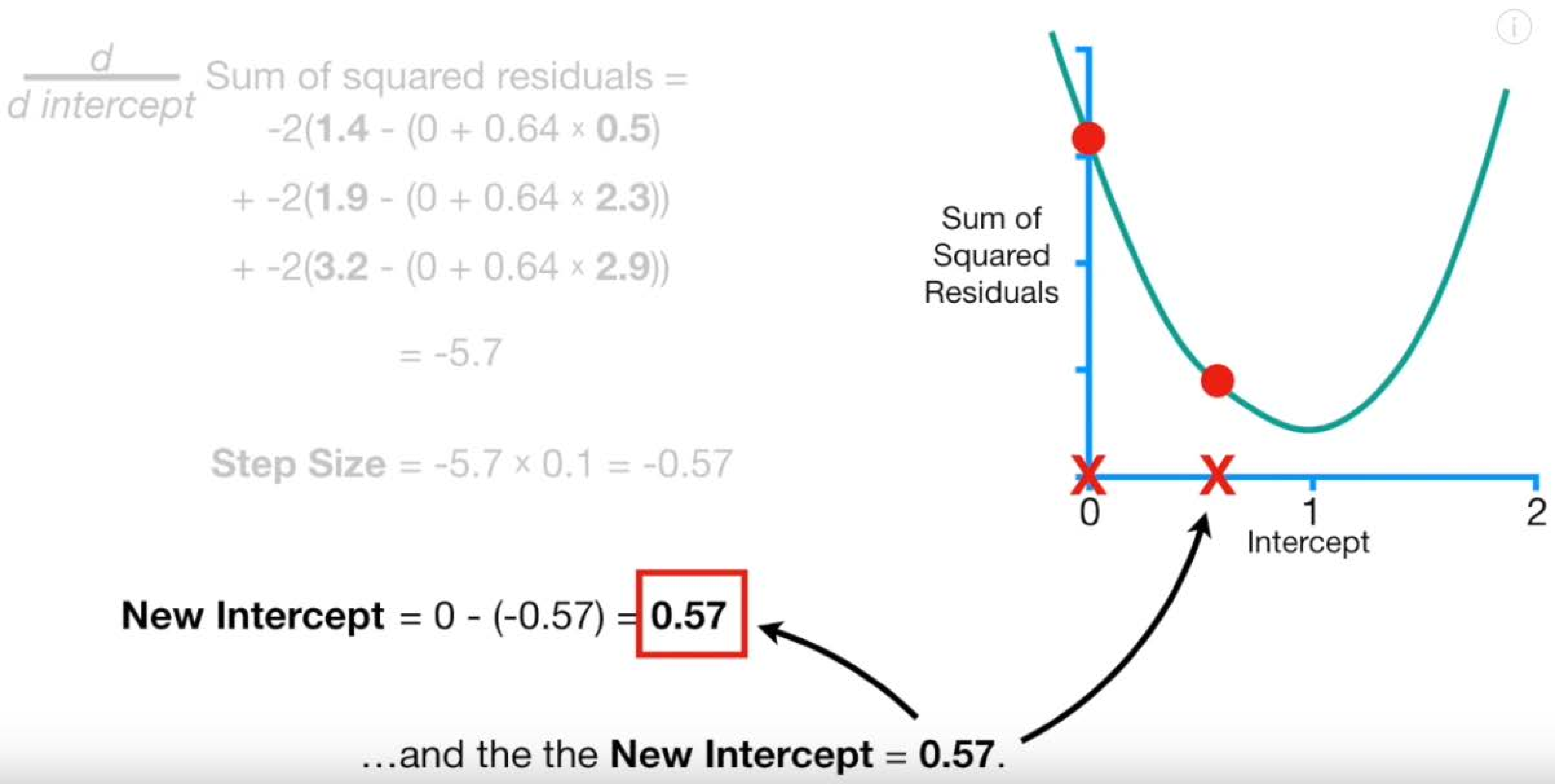
$$= -5.7$$

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

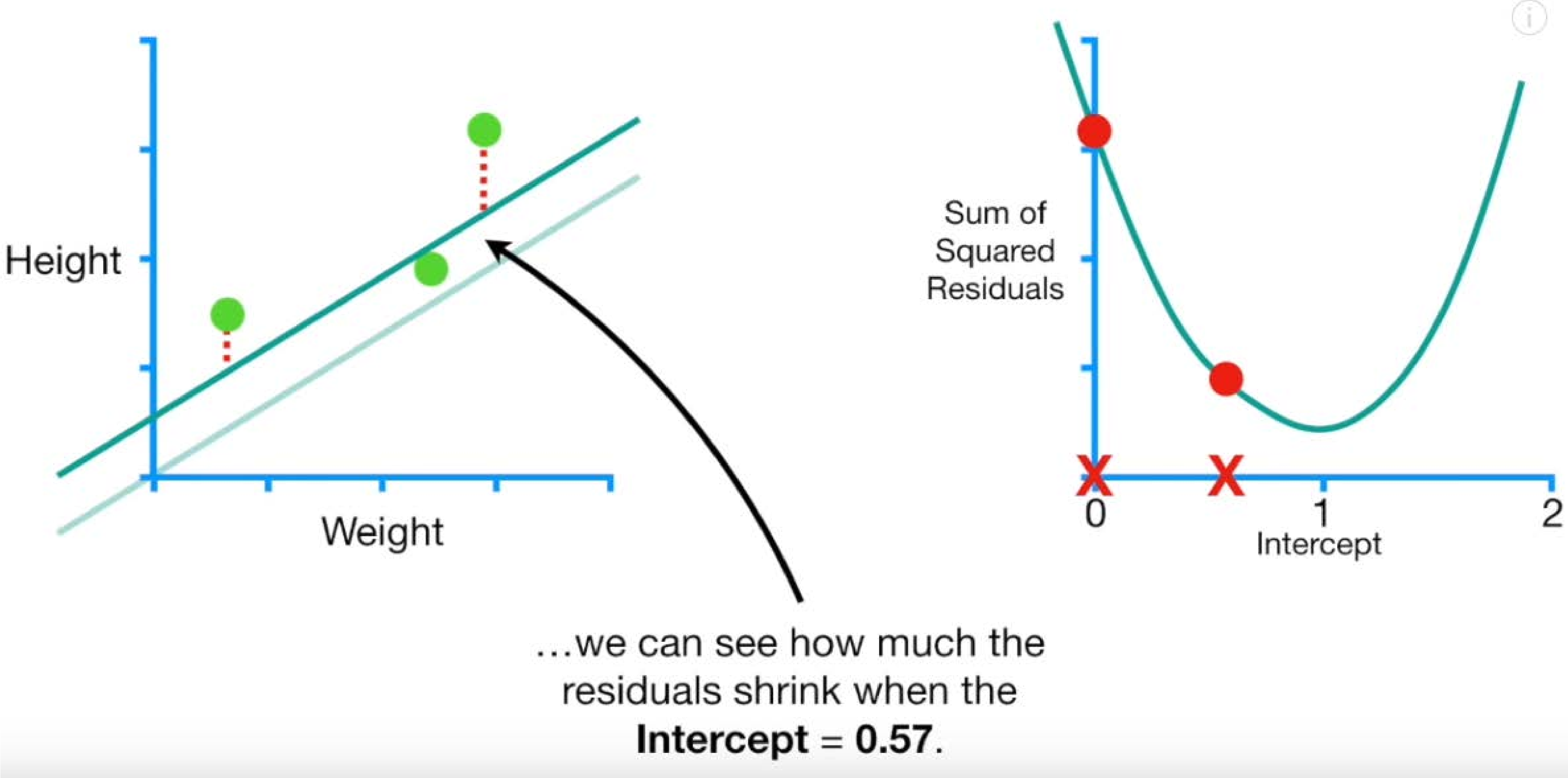
**New Intercept** = **Old Intercept** - **Step Size**



# Gradient Descent



# Gradient Descent



# Gradient Descent

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

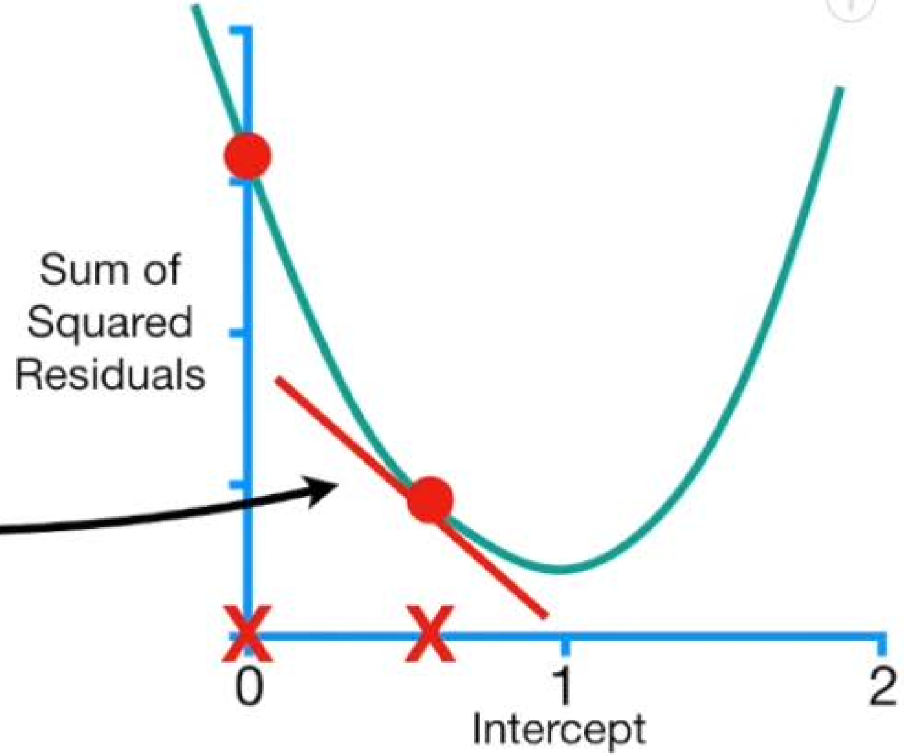
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3}))$$

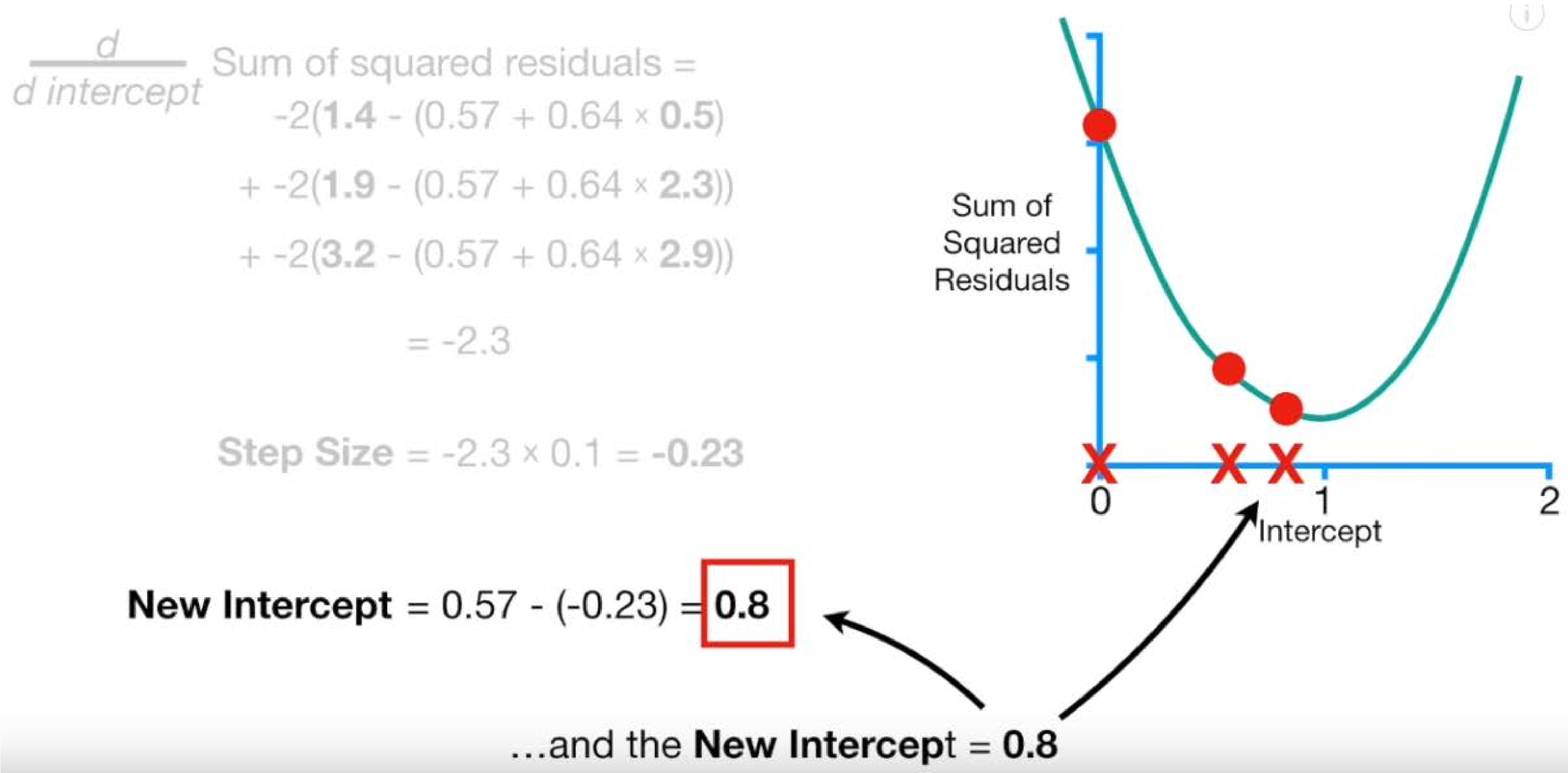
$$+ -2(\mathbf{3.2} - (0.57 + 0.64 \times \mathbf{2.9}))$$

$$= \boxed{-2.3}$$

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



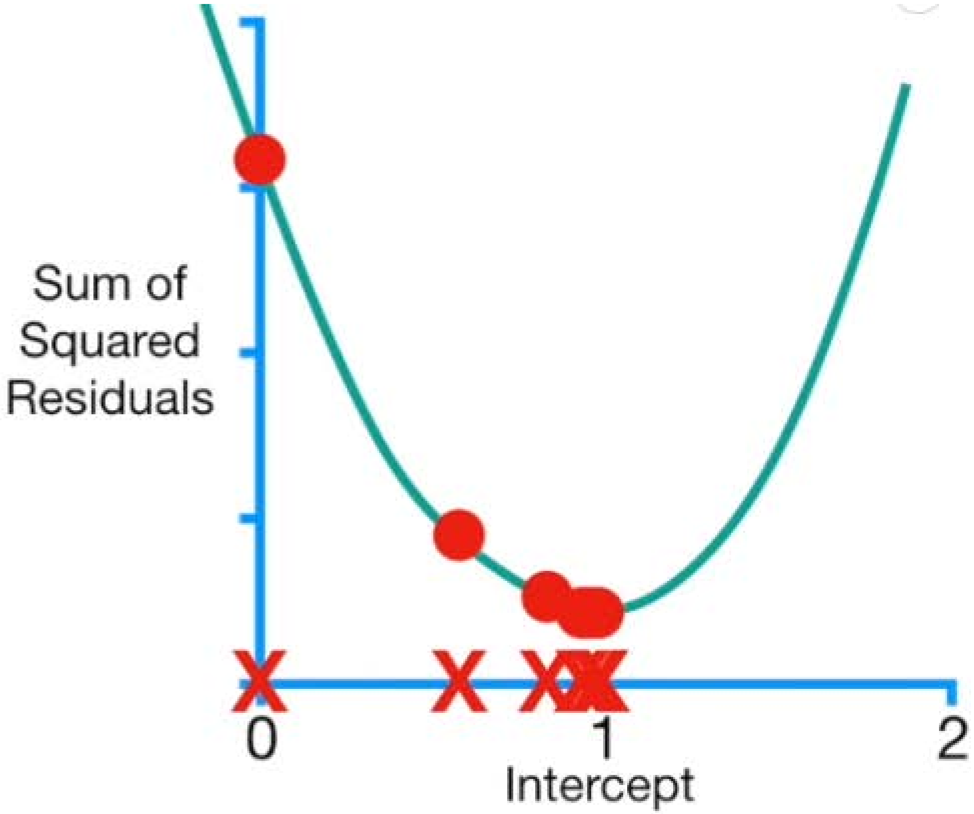
# Gradient Descent



# Gradient Descent

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

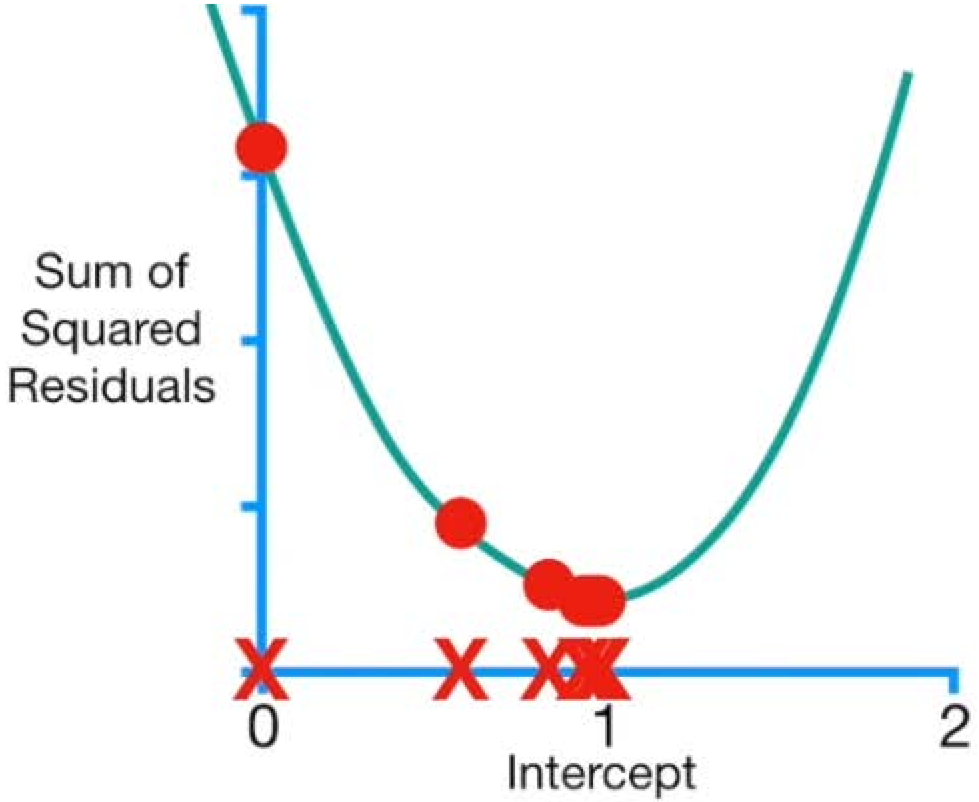
**Step Size** = **Slope** × **Learning Rate**



# Gradient Descent

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

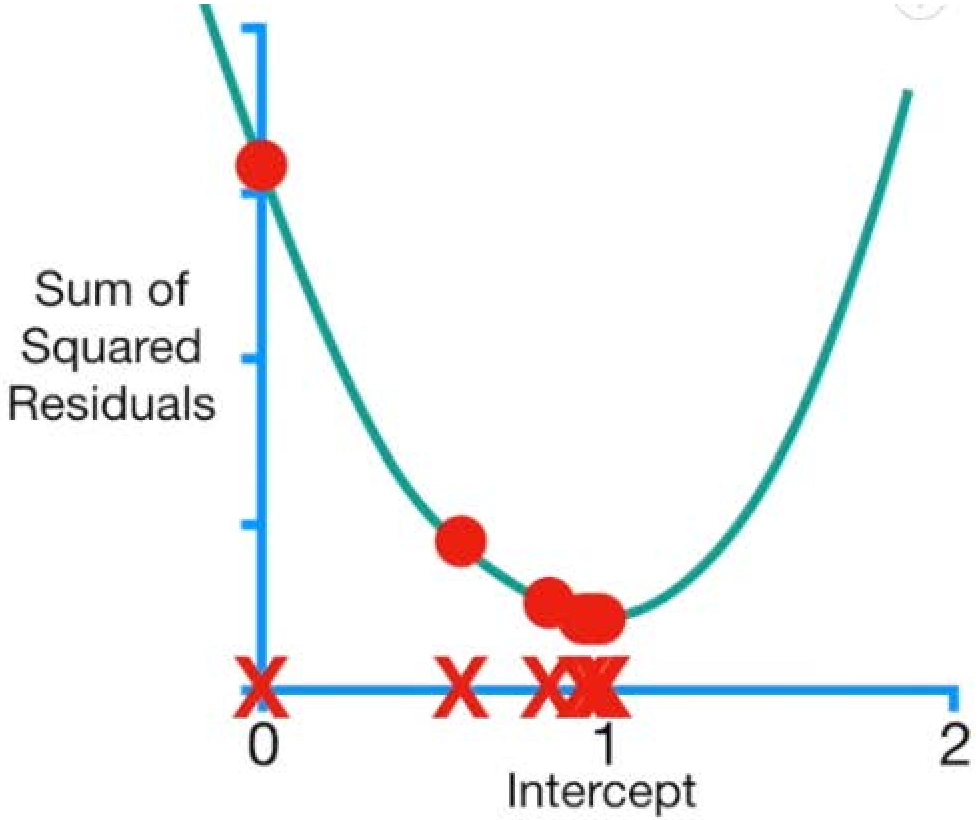
**Step Size = Slope × Learning Rate**



# Gradient Descent

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

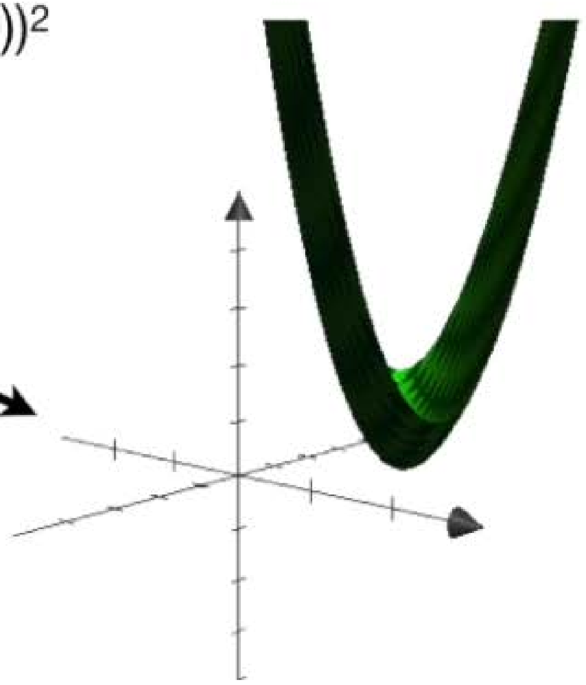
In practice, the **Maximum Number of Steps = 1,000** or greater.



# Gradient Descent

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

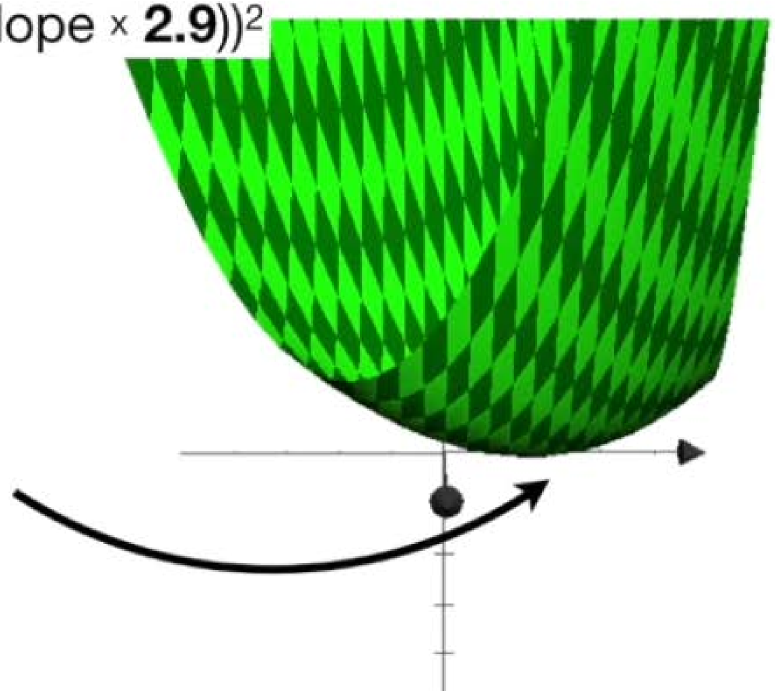
…this axis represents different values for the **Slope**…



# Gradient Descent

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

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



# Gradient Descent

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

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

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

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

# Gradient Descent

$$\frac{d}{d\ intercept} \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

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

$$\frac{d}{d\ slope} \text{Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

# Gradient Descent

$$\frac{d}{d\ intercept} \text{ Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

$$+ -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) \boxed{= \mathbf{-1.6}}$$

$$\textbf{Step Size}_{\text{Intercept}} = -1.6 \times \textbf{Learning Rate}$$

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

$$\frac{d}{d\ slope} \text{ Sum of squared residuals} =$$

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 \boxed{= \mathbf{-0.8}}$$

$$\textbf{Step Size}_{\text{Slope}} = -0.8 \times \textbf{Learning Rate}$$

# Gradient Descent

**Step Size**$_{\text{Intercept}}$ = -1.6 x 0.01

**Step Size**$_{\text{Intercept}}$ = -1.6 x 0.01 = **-0.016**

**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.

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

**Step Size**$_{\text{Slope}}$ = -0.8 x 0.01 = **-0.008**

**Step Size**$_{\text{Slope}}$ = -0.8 x 0.01

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

# Gradient Descent

**Step Size**$_{\text{Intercept}}$ = -1.6 x 0.01 = $\boxed{\textbf{-0.016}}$

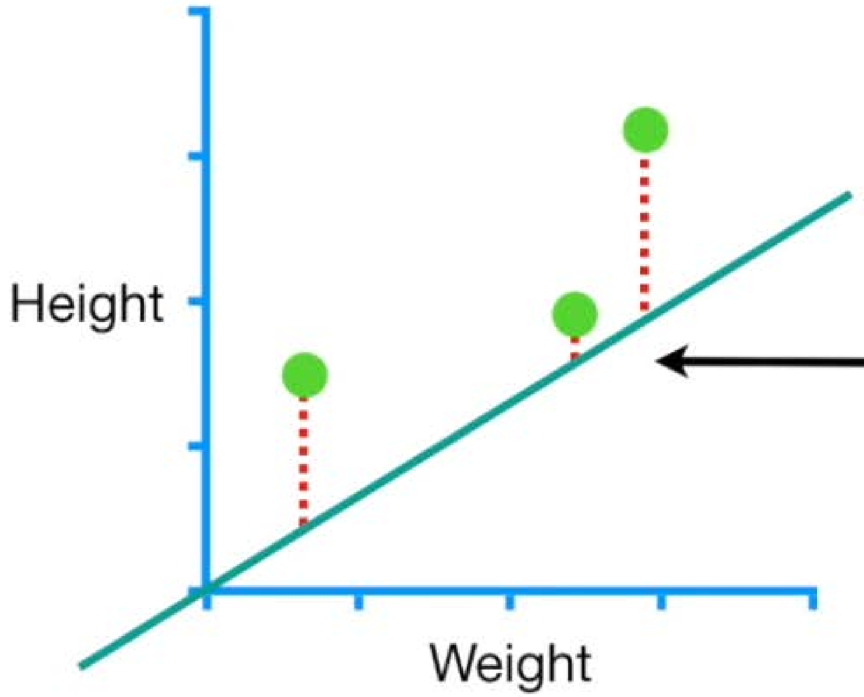
**New Intercept** = 0 - (-0.016)

...and the
**Step Sizes**...

**Step Size**$_{\text{Slope}}$ = -0.8 x 0.01 = $\boxed{\textbf{-0.008}}$

**New Slope** = 1 - (-0.008)

# Gradient Descent

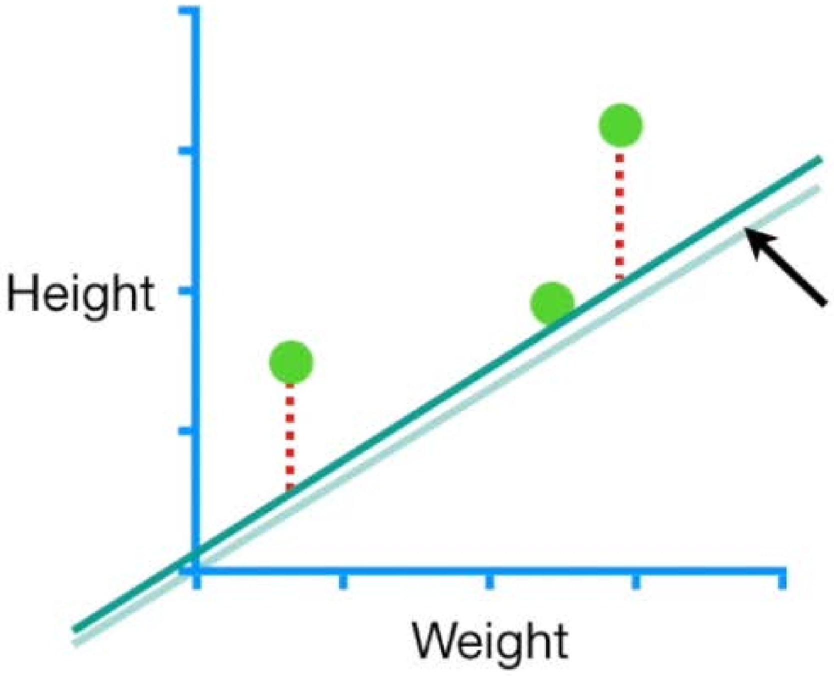



**New Intercept** = 0 - (-0.016) = 0.016

This is the line we started with… (**Slope = 1** and **Intercept = 0**)

**New Slope** = 1 - (-0.008) = 1.008

# Gradient Descent

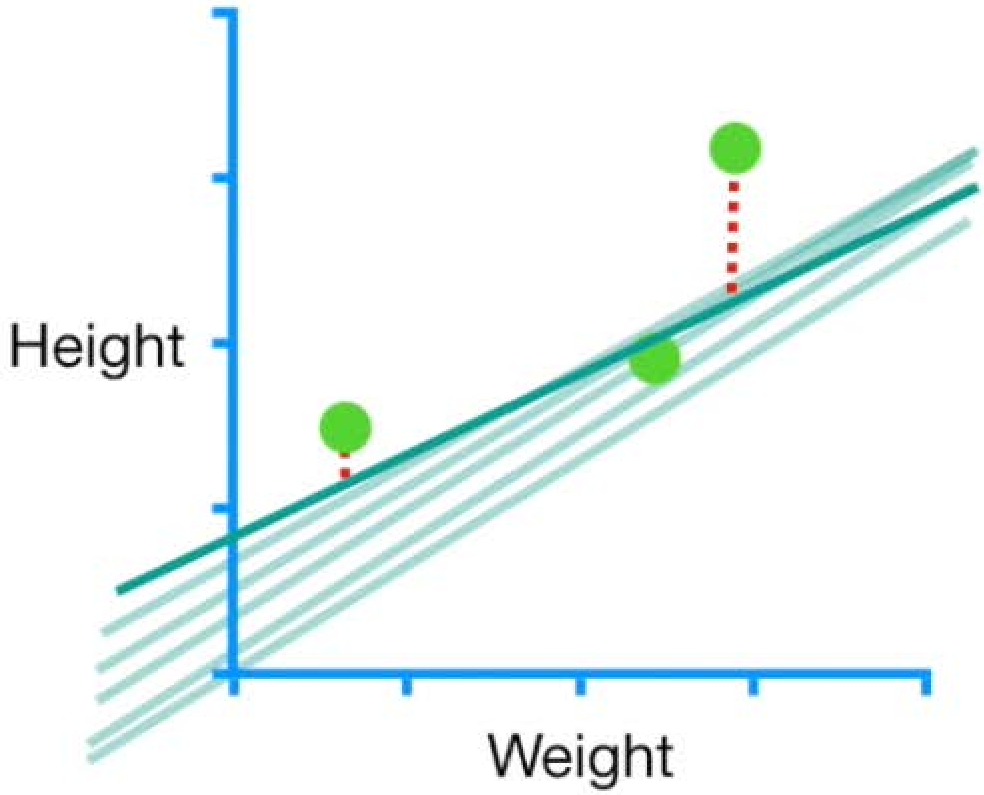



**New Intercept** = 0 - (-0.016) = 0.016

…and this is the new line (with **Slope = 1.008** and **Intercept = 0.016**) after the first step.

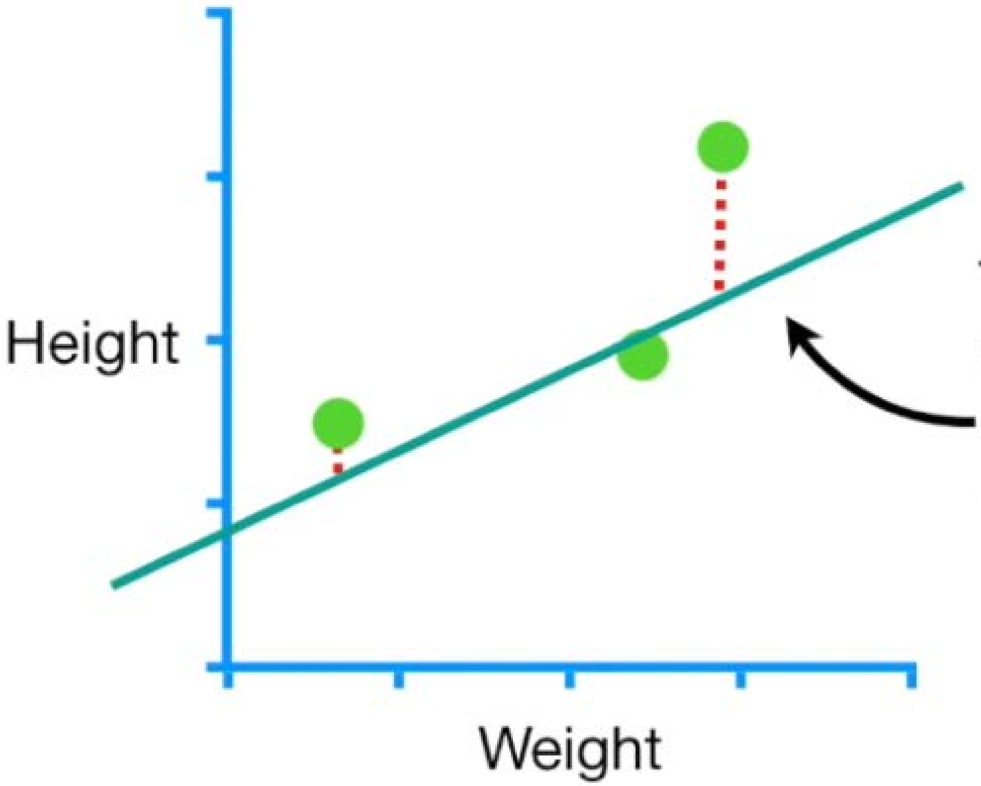
**New Slope** = 1 - (-0.008) = 1.008

# Gradient Descent



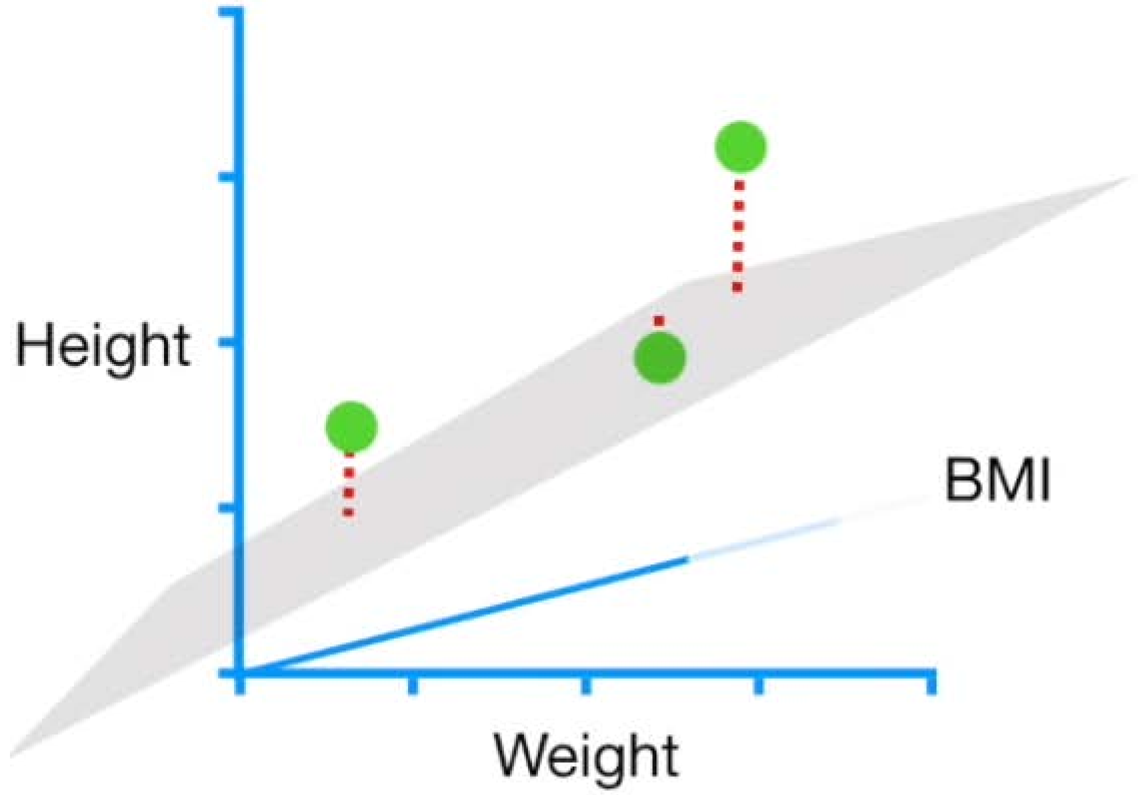

Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.

# Gradient Descent




This is the best fitting line, with **Intercept** = **0.95** and **Slope** = **0.64**, the same values we get from **Least Squares**.

# Gradient Descent



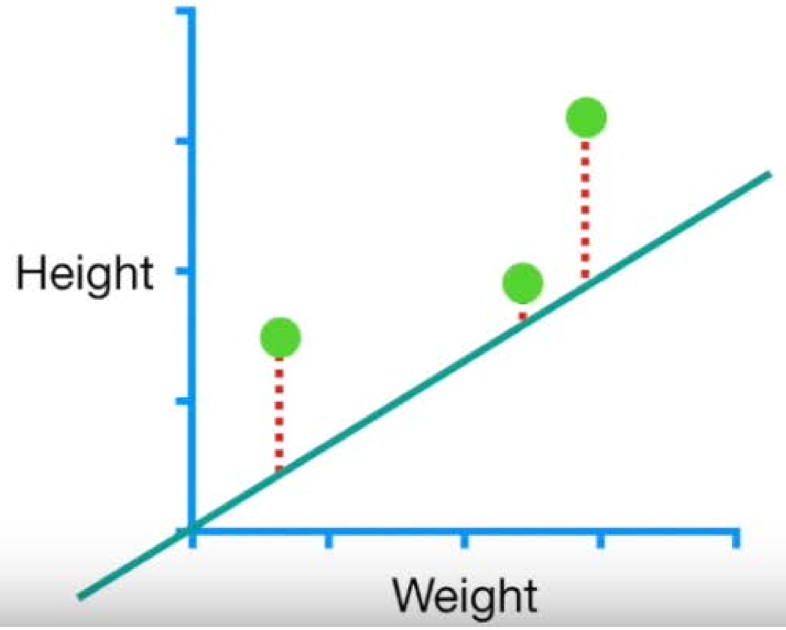

If we had more parameters, then we'd just take more derivatives and everything else stays the same.

# Gradient Descent

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

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 :-*

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

Mean absolute error

*Mathematical formulation :-*

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

Mean bias error

*Mathematical formulation :-*

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

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.

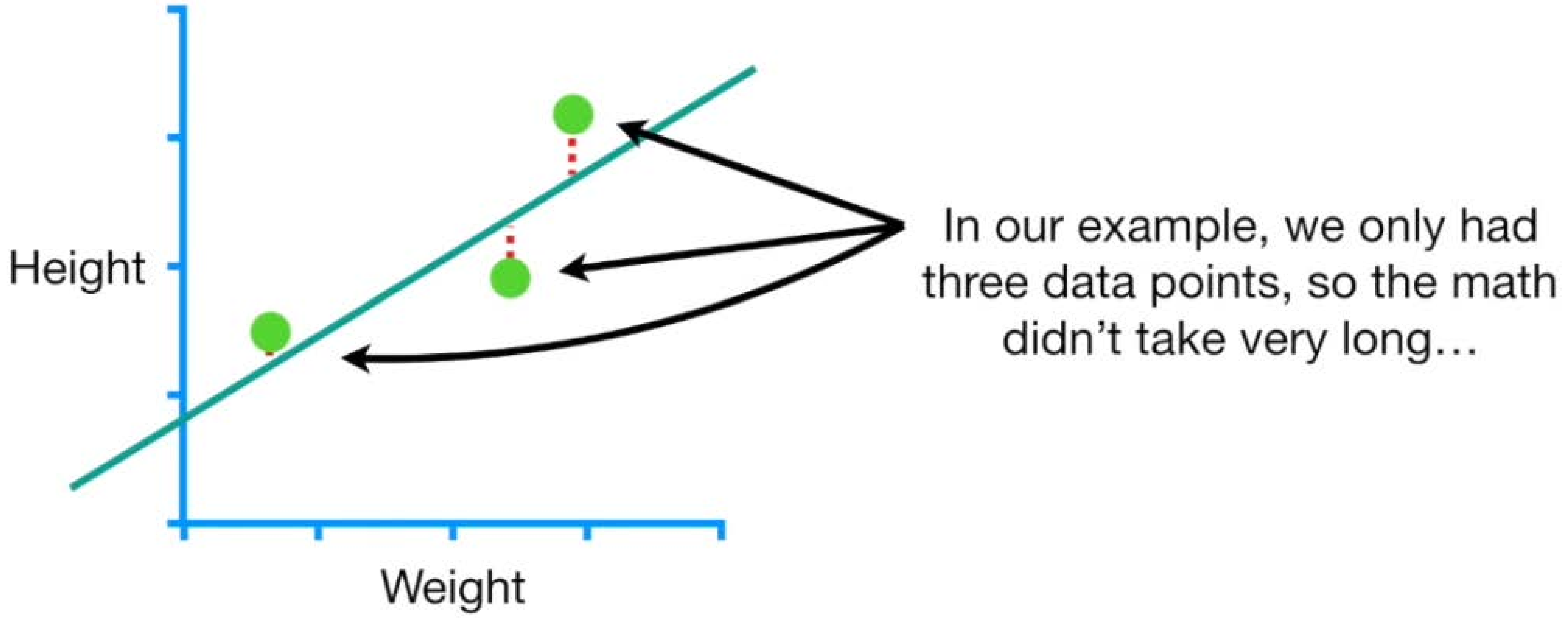
**Step 3:** Plug the parameter values into the derivatives (ahem, the **Gradient**).

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

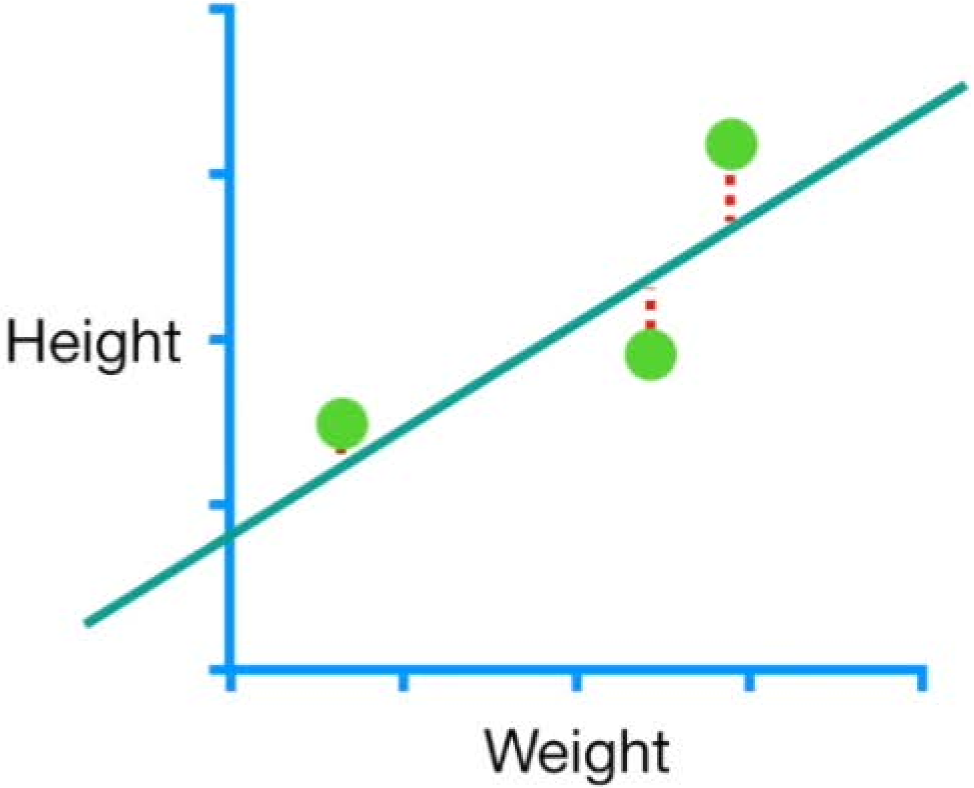
**Step 5:** Calculate the New Parameters:

**New Parameter** = **Old Parameter** - **Step Size**

# Gradient Descent

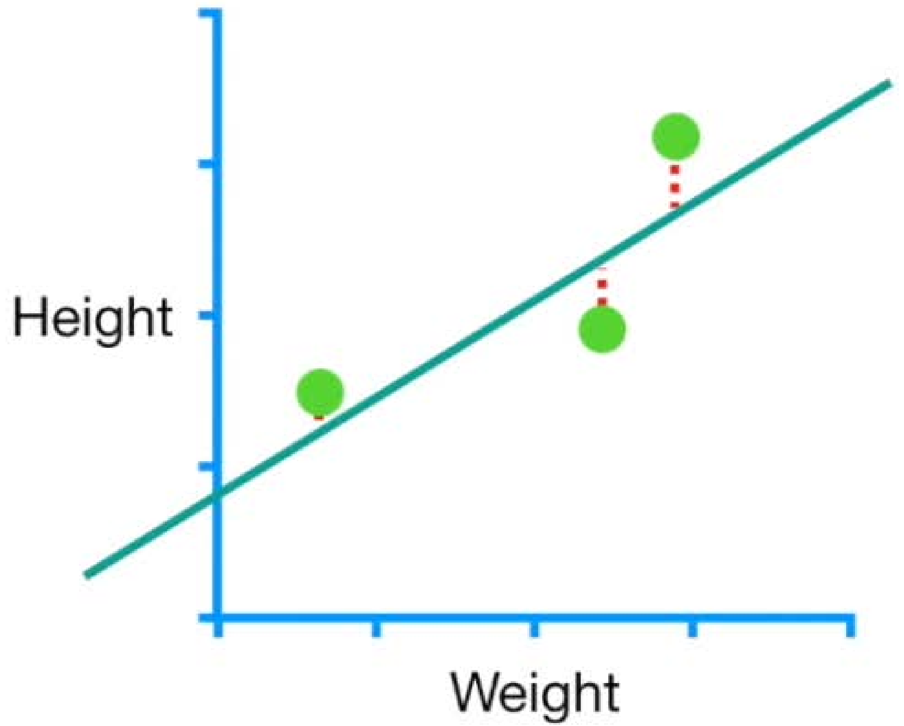


# Gradient Descent




...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.