Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Great! Now that we know all about this super simple **Training Dataset**, let's walk through the original **Gradient Descent** algorithm step-by-step.

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$ and a differentiable **Loss Function** $L(y_i, F(x))$

In this case, a **Loss Function** is just something that evaluates how well we can predict **Weight**.

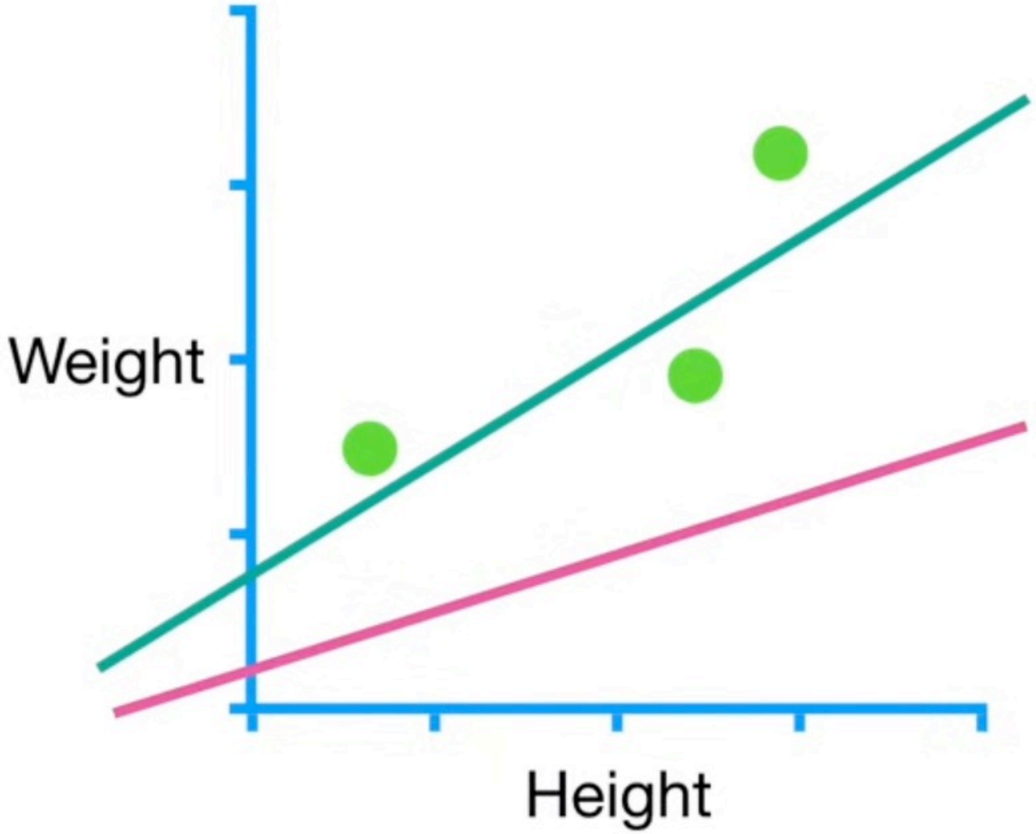
Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$ and a differentiable **Loss Function** $L(y_i, F(x))$

The **Loss Function** that is most commonly used when doing **Regression** with **Gradient Boost** is…

$$\frac{1}{2}\,(\textbf{Observed} - \textbf{Predicted})^2$$

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56



In other words, it doesn't matter if the **Loss Function** is…

$$(\textbf{Observed} - \textbf{Predicted})^2$$

…of if it's…

$$\frac{1}{2}(\textbf{Observed} - \textbf{Predicted})^2$$

…since both will tell that the **Greenish Line** has the best fit.

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

The reason why people choose this **Loss Function** for **Gradient Boost** is that when we differentiate it with respect to "**Predicted**"...

$$\frac{d}{d\,Predicted}\ \frac{1}{2}(\textbf{Observed} - \textbf{Predicted})^2$$

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

…and that leaves you with the **Observed** minus the **Predicted** multiplied by **-1**.

$$\frac{d}{d\,Predicted}\ \frac{1}{2}(\mathbf{Observed} - \mathbf{Predicted})^2$$

$$= \frac{2}{2}(\mathbf{Observed} - \mathbf{Predicted}) \times -1$$

$$= -(\mathbf{Observed} - \mathbf{Predicted})$$

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

The y_i's are the **Observed** values…

$$\frac{1}{2}(\text{Observed} - \text{Predicted})^2$$

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y, F(x))$

…and ***F(x)*** is a function that gives us the **Predicted** values.

$$\frac{1}{2}(\textbf{Observed} - \textbf{Predicted})^2$$

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

NOTE: There are other **Loss Functions** to choose from, but this is the most popular one for **Regression**.

$$\frac{1}{2}(\textbf{Observed} - \textbf{Predicted})^2$$

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

We start by initializing the model with a constant value…

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

…and that constant value is determined by this funky looking thing.

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

$\frac{1}{2}$ (**Observed** - **Predicted**)^2

…and that funky symbol, called **gamma**, refers to the **Predicted** values.

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

$$\frac{1}{2}(88 - \text{Predicted})^2 + \frac{1}{2}(76 - \text{Predicted})^2 + \frac{1}{2}(56 - \text{Predicted})^2$$

…and the "**argmin over gamma**" means we need to find a **Predicted** value that minimizes this sum.

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

$\frac{1}{2}$ (**88** - **Predicted**)^2 +

$\frac{1}{2}$ (**76** - **Predicted**)^2 +

$\frac{1}{2}$ (**56** - **Predicted**)^2

In other words, if we plot the **Observed Weights** on a number line…

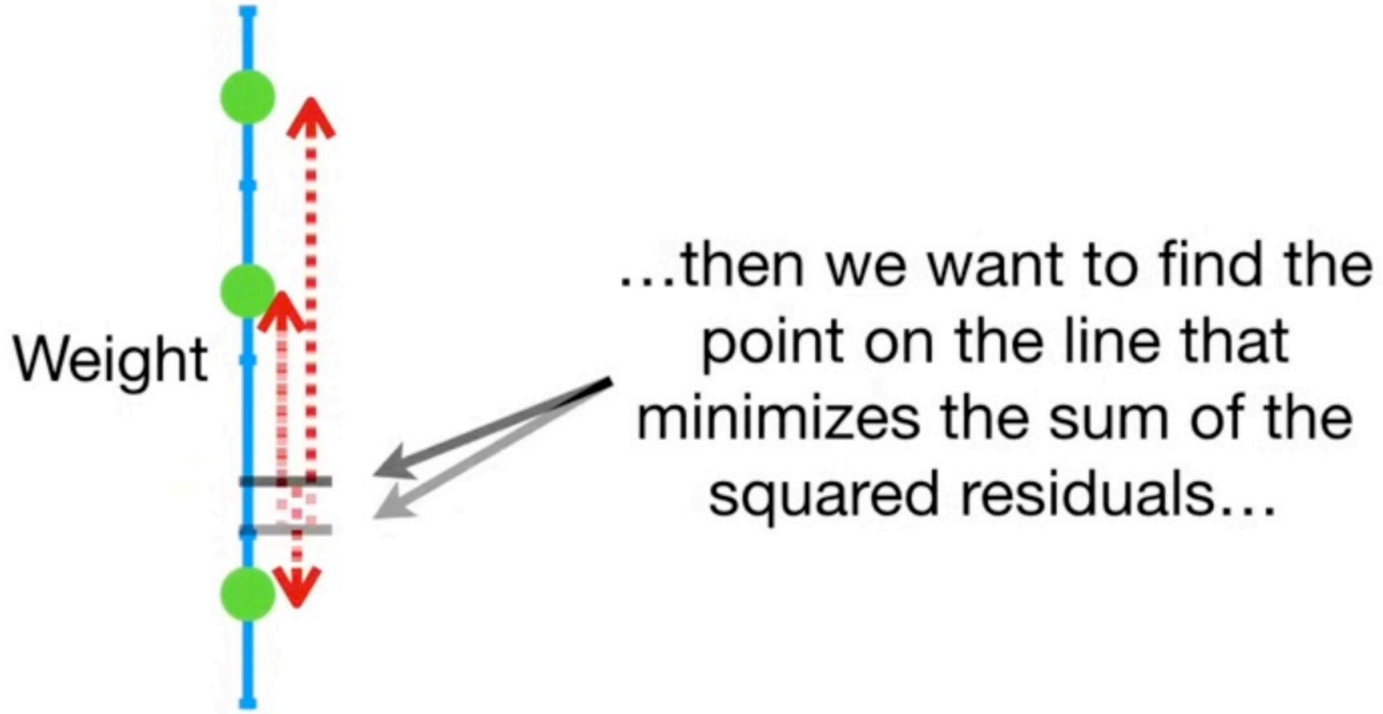
Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

$\frac{1}{2}$(**88** - **Predicted**)^2 +

$\frac{1}{2}$(**76** - **Predicted**)^2 +

$\frac{1}{2}$(**56** - **Predicted**)^2



Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Fu**

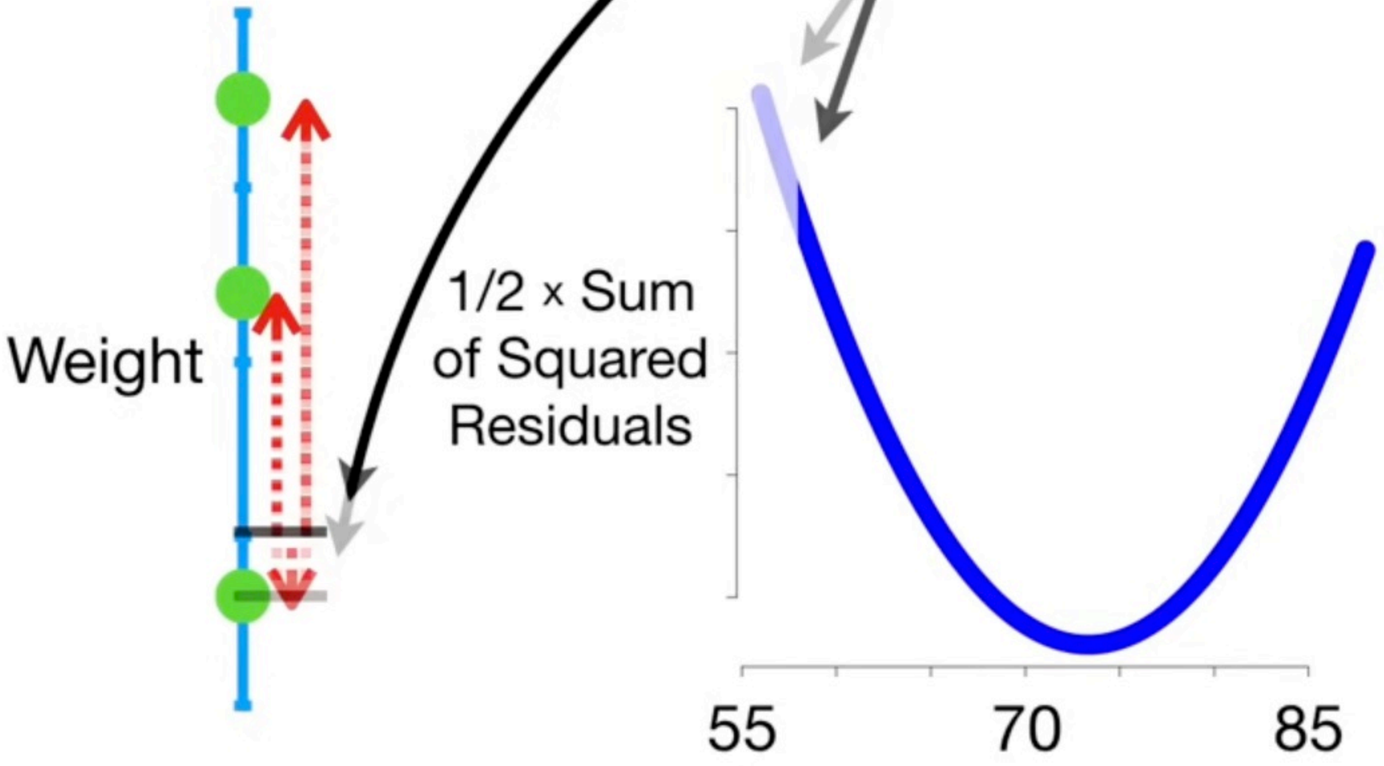
Step 1: Initialize model with a constant value: $F_0(x)$

$$\frac{1}{2}(\mathbf{88} - \mathbf{Predicted})^2 +$$

$$\frac{1}{2}(\mathbf{76} - \mathbf{Predicted})^2 +$$

$$\frac{1}{2}(\mathbf{56} - \mathbf{Predicted})^2$$

Here's a plot of one half of the sum of the squared residuals for potential **Predicted** values…



Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Fu**

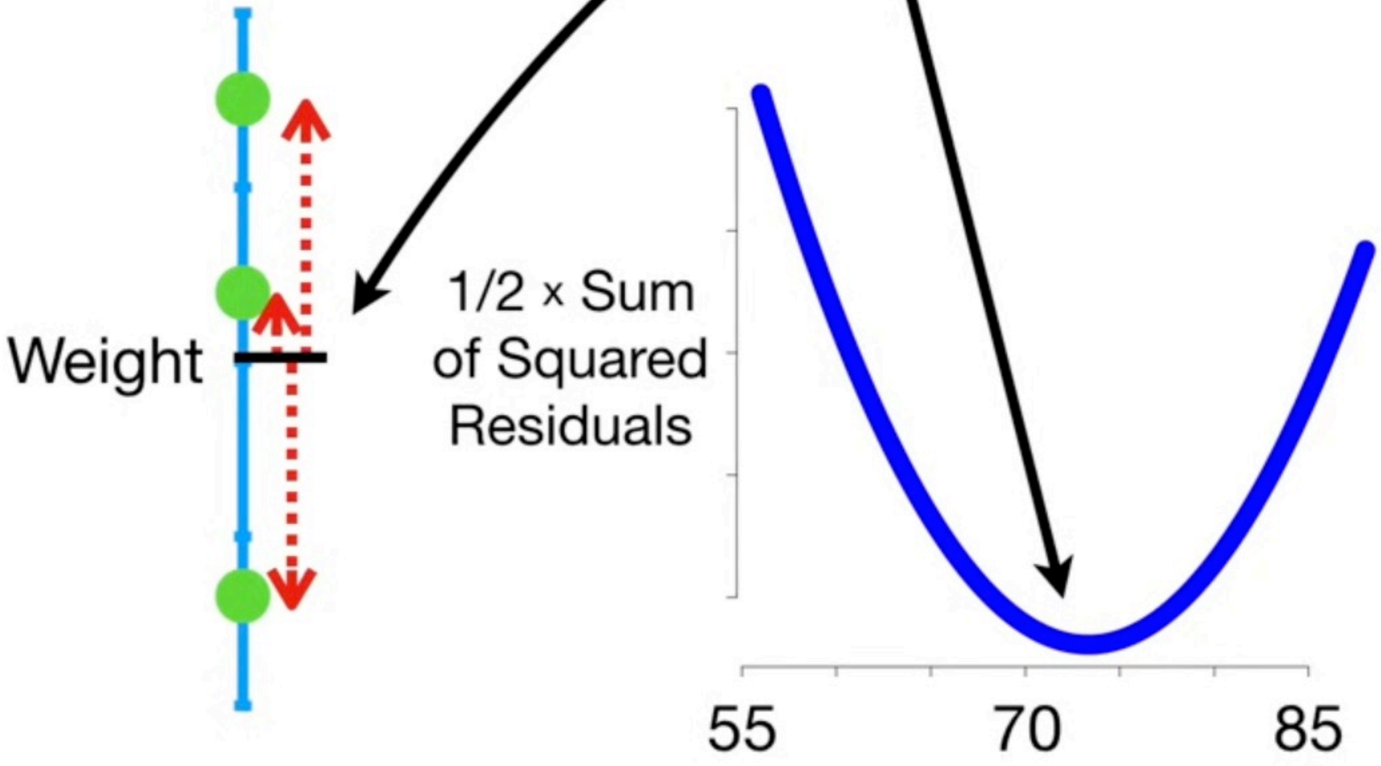
Step 1: Initialize model with a constant value: $F_0(x)$

$$\frac{1}{2}(\mathbf{88} - \mathbf{Predicted})^2 +$$

$$\frac{1}{2}(\mathbf{76} - \mathbf{Predicted})^2 +$$

$$\frac{1}{2}(\mathbf{56} - \mathbf{Predicted})^2$$

NOTE: We could use **Gradient Descent** to find the optimal value for **Predicted**…



Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Fu**

Step 1: Initialize model with a constant value: $F_0(x)$

$$\frac{1}{2}(\mathbf{88} - \mathbf{Predicted})^2 +$$

$$\frac{1}{2}(\mathbf{76} - \mathbf{Predicted})^2 +$$

$$\frac{1}{2}(\mathbf{56} - \mathbf{Predicted})^2$$

…but we can also just solve for it, because the math isn't that hard.

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

$\frac{1}{2}$(88 - Predicted)^2 + ⟶ -(88 - Predicted) +

$\frac{1}{2}$(76 - Predicted)^2 + ⟶ -(76 - Predicted) +

$\frac{1}{2}$(56 - Predicted)^2 ⟶ -(56 - Predicted)

$\frac{d}{d\,Predicted}$

= 0

Then we set sum of the derivatives equal to zero…

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

Predicted $= \frac{88 + 76 + 56}{3}$

…and we end up with the **Average** of the **Observed Weights**.

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

…the value for **gamma** that minimizes this sum…

Predicted $= \frac{88 + 76 + 56}{3}$

$\frac{1}{2}$ (**Observed** - **Predicted**)^2

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

…is the average of the **Observed Weights**.

Predicted $= \frac{88 + 76 + 56}{3}$

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

We have now created the initial predicted value, ***F₀(x)***…

$$F_0(x) = \frac{88 + 76 + 56}{3}$$

Input: Data $\{(x_i, y_i)\}_{i=1}^{n}$, and a differentiable **Loss Function** $L(y_i, F(x))$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

That means that the initial predicted value, ***F₀(x)***, is just a leaf.

$$F_0(x) = \frac{88 + 76 + 56}{3}$$

$$= 73.3$$

73.3

The leaf predicts that all samples will weigh **73.3**.

Step 2 is huge, but we'll take it one step at a time. :)

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_m I(x \in R_{jm})$

Step 2: for $m = 1$ to M:

In generic terms, we will make ***M*** trees, but in practice, most people set ***M*** = **100** and make **100** trees.

Step 2: for $m = 1$ to M:

Little ***m*** refers to an individual tree. So when little ***m*** = **1**, then we're talking about the first tree.

Step 2: for $m = 1$ to M:

When little ***m*** = big ***M***, then we're talking about the last tree…

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

Part A of **Step 2** looks nasty,
but it's not.

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = \left[- \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right] \right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

This big minus sign tells us to multiply the derivative by **-1**…

-1 x -(**Observed** - **Predicted**)

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

…and that leaves us with the **Observed** value minus the **Predicted** value.

(Observed - Predicted)

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

In other words, this nasty looking thing…

…is just a **Residual**.

(Observed - Predicted)

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

Now we plug $F_{m-1}(x)$ in for **Predicted**…

(**Observed** - $F_{m-1}(x)$)

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

…and since ***m*** = **1**, that means we plug in ***F₀(x)*** for ***F$_{m-1}$(x)***…

(**Observed** - ***F₀(x)***)

Step 2: for m = 1 to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,\ldots,n$

Now we can compute $r_{i,m}$, where ***r*** is short for **Residual**, ***i*** is the sample number and ***m*** is the tree that we're trying to build.

(**Observed** - 73.3)

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

This tells us to calculate **Residuals** for all **3** samples in the dataset.

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

Hooray! We've finished **Part A** of **Step 2** by calculating a **Residual** for each sample.

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,\ldots,n$

NOTE: Before we move on, I just want to point out that this derivative is the **Gradient** that **Gradient Boost** is named after.

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$

All this is saying is that we will build a regression tree…

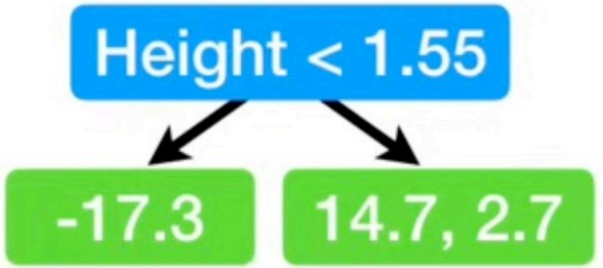
Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,\ldots,n$

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

…we to predict the **Residuals** instead of the **Weights**.

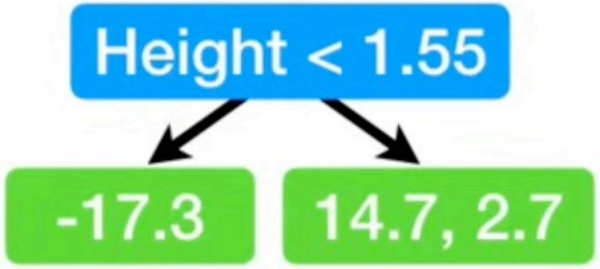


Here's the new tree.

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

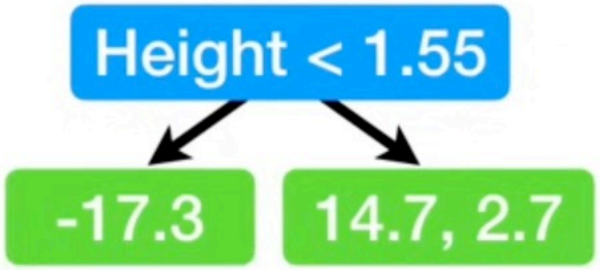
Yes, I know this is just a stump and **Gradient Boost** almost always uses larger trees.

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3



However, in order to demonstrate details of the **Gradient Boost** algorithm, we need at least one leaf with more than **1** sample in it…

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

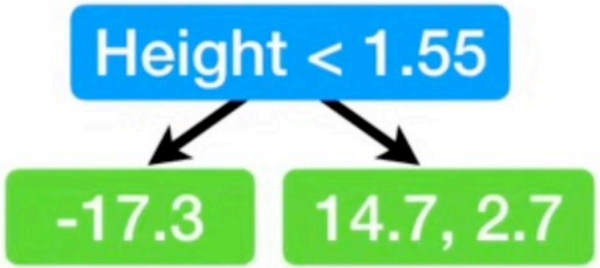


However, in order to demonstrate details of the **Gradient Boost** algorithm, we need at least one leaf with more than **1** sample in it…

…and when you only have **3** samples, then you can't have more than **2** leaves.

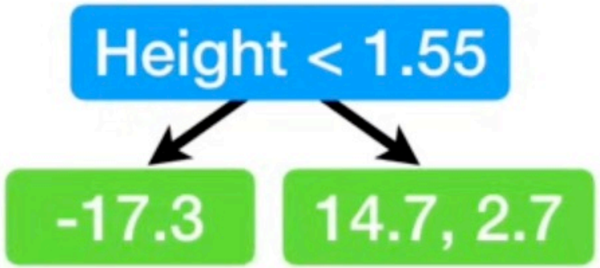
So we're stuck with using stumps, even though they are not typically used with **Gradient Boost**.

Height (m)	Favorite Color	Gender		
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3



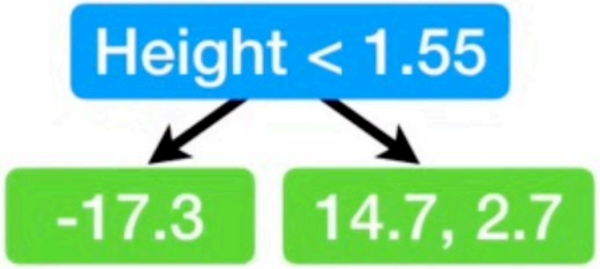
The **Residual** for the third sample, x_3, goes to the leaf on the left…

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3



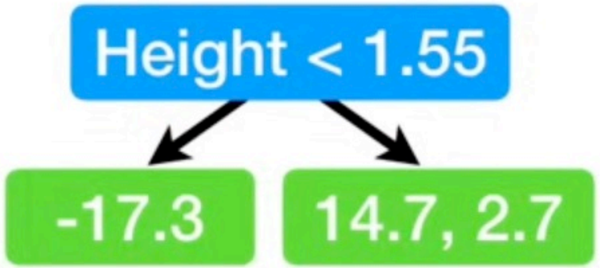
…and the **Residuals** for samples x_1 and x_2 go to the leaf on the right.

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	50	-17.3



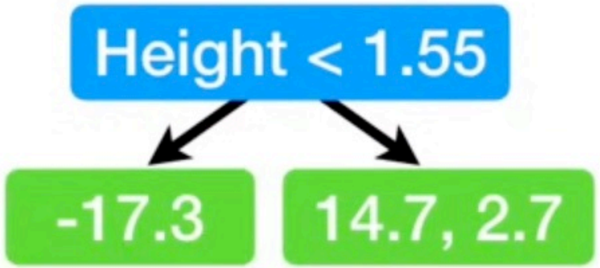
Now we need to "create terminal regions $R_{j,m}$".

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$



This part is super easy because the **Leaves** are the "terminal regions $\boldsymbol{R_{j,m}}$".

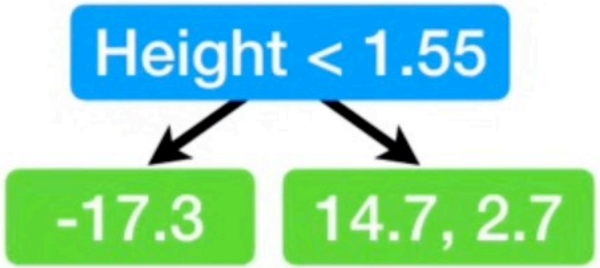
(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$



This part is super easy because the **Leaves** are the "terminal regions $R_{j,m}$".

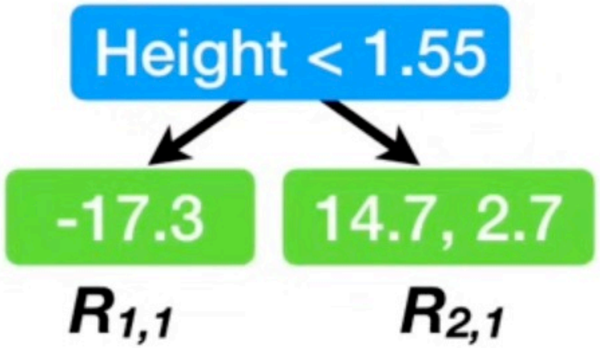
NOTE: This little ***m*** is the index for the tree we just made. Since this is the first tree, ***m*** = **1**.

…and this little ***j*** is the index for each leaf in the tree.



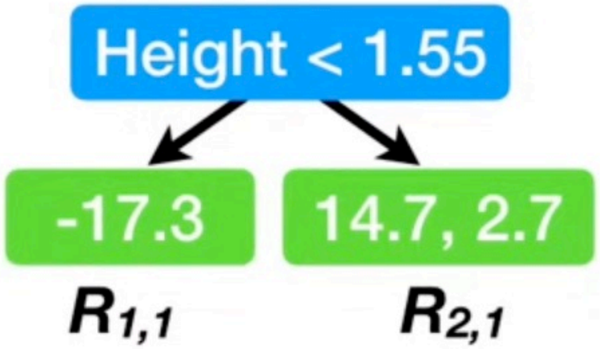
Since this tree has **2** leaves, $J_m = 2$.

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$



NOTE: It doesn't matter which leaf gets which label. However, once we give a leaf a label, we need to keep track of it.

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$



Hooray!!! We've finished **Part B** of **Step 2** by fitting a **Regression Tree** to the residuals and labeling the leaves.

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

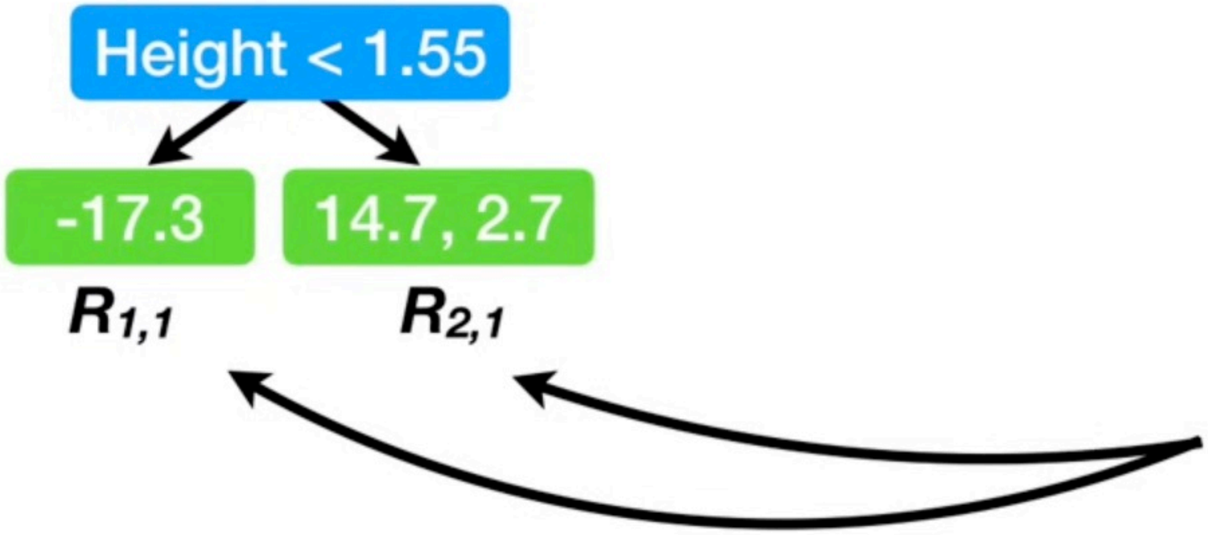
Now let's do **Part C**.

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,\ldots,n$

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1\ldots J_m$

(C) For $j = 1\ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

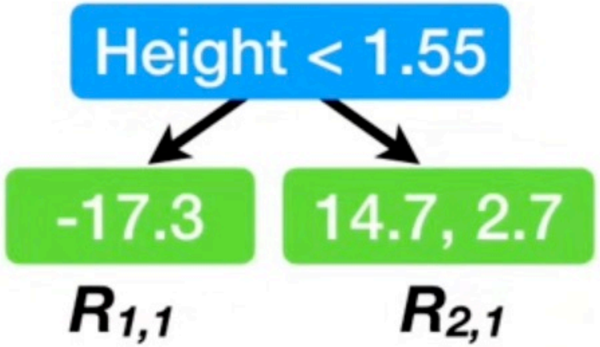


In this part, we the determine the **Output Values** for each leaf.

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

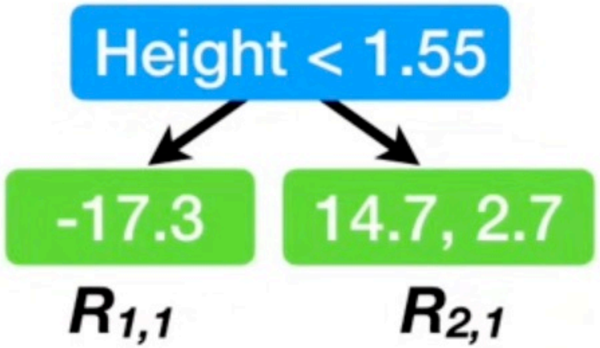
Specifically, since two residuals ended up in this leaf, it's unclear what its **Output Value** should be.

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$



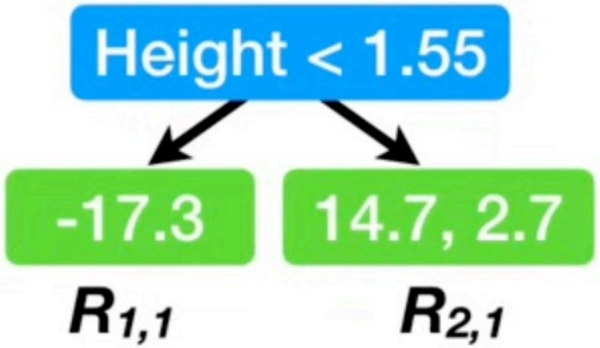
So for each leaf in the new tree…

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$



…we compute an
Output Value,
"**gamma**$_{j,m}$".

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$



The **Output Value** for each leaf is the value for **gamma** that minimizes this summation.

(C) For $j = 1 \ldots J_m$ compute

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Step 1: Initialize model with a constant value: $$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

NOTE: This minimization is like what we did in **Step 1**.

(C) For $j = 1 \ldots J_m$ compute $$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

One small difference is that now we are taking the previous **Prediction** into account…

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

…while before, since we were just starting out, there was no "previous **Prediction**".

One small difference is that now we are taking the previous **Prediction** into account…

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

The other difference is that this summation is picky about which samples it includes.

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

Step 1: Initialize model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$

…while before, the summation included all of the samples.

The other difference is that this summation is picky about which samples it includes.

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

Specifically, the x_i *in* $R_{i,j}$ means that…

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17

…since only sample,
x_3, goes to leaf $R_{1,1}$…

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17

…then only x_3 is used to calculate the **Output Value** for $R_{1,1}$…

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17

…and since only two samples, x_1 and x_2, go to leaf $R_{1,2}$…

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17

…then only x_1 and x_2 are used to calculate the **Output Value** for $R_{1,2}$.

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Let's start by calculating the **Output Value** for the leaf on the left, $R_{1,1}$.

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

That means ***j*** = **1**, since this is the first leaf, and ***m*** = **1**, since this is the first tree.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

$\frac{1}{2}$ (**Observed - Predicted**)^2

Now let's replace the generic **Loss Function**…

…with the actual **Loss Function** that we decided to use…

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{1,1}} L(y_i, F_{m-1}(x_i) + \gamma)$$

…and let's expand the summation into individual terms.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} \frac{1}{2}(y_i - (F_{m-1}(x_i) + \gamma))^2$$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

Now we plug in the value for y_3, the **Observed** value…

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(y_3 - (F_{m-1}(x_3) + \gamma))^2$$

…and the most recent **Predicted** value for x_3.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(56 - (F_{m-1}(x_3) + \gamma))^2$$

$F_0(x)$

73.3

Since ***m*** = 1, the most recent **Prediction** was $F_0(x)$, which **Predicted** that all samples weighed **73.3**…

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(56 - (F_{m-1}(x_3) + \gamma))^2$$

…and simplify what's inside the parentheses.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(-17.3 - \gamma)^2$$

Now we need to find the value for **gamma** that minimizes this equation.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(-17.3 - \gamma)^2$$

Step 1: Initialize model with a constant value: $$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

Just like **Step 1**, we can try different values for **gamma** or solve it analytically.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(-17.3 - \gamma)^2$$

$$\frac{d}{d\gamma}\frac{1}{2}(-17.3-\gamma)^2$$

First, we take the derivative of the **Loss Function** with respect to **gamma**, just like we did at the very start.

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}}\,\frac{1}{2}(-17.3-\gamma)^2$$

$$\frac{d}{d\gamma}\frac{1}{2}(-17.3-\gamma)^2 \rightarrow \boxed{17.3+\gamma=0}$$

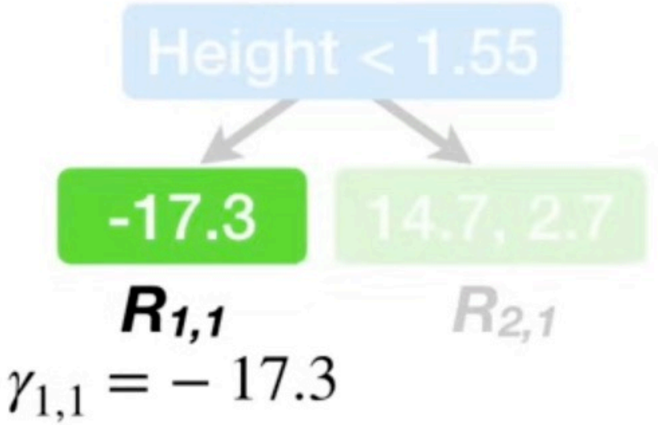
Now we set the derivative equal to **0**…

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}}\,\frac{1}{2}(-17.3-\gamma)^2$$

$$\gamma = -17.3$$

The value for **gamma** that minimizes this equation is **-17.3**.

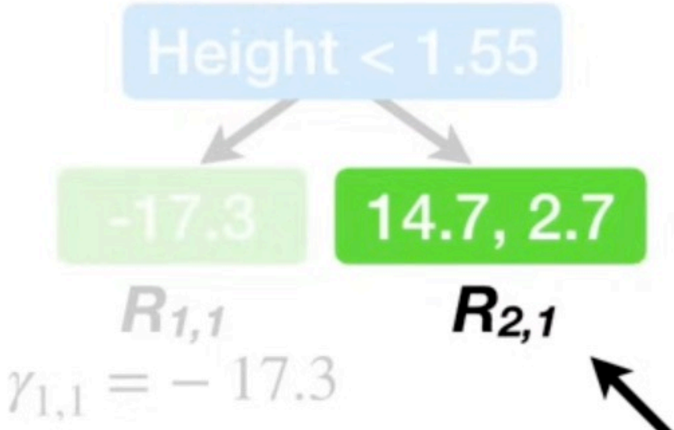
$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(-17.3 - \gamma)^2$$



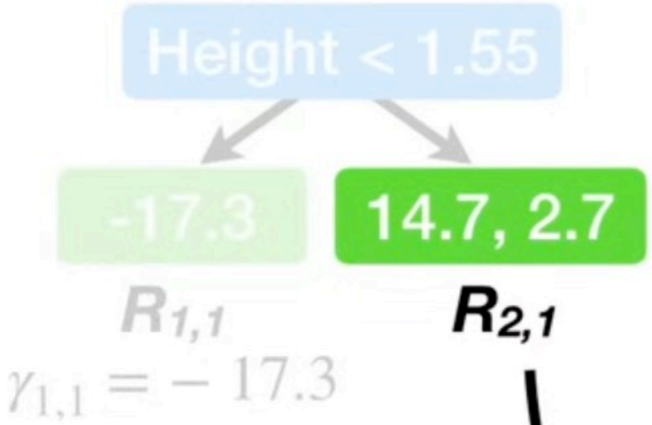
$$\gamma_{1,1} = -17.3$$

…and ultimately, the leaf, $R_{1,1}$, has an **Output Value** of **-17.3**.

$$\gamma_{1,1} = -17.3$$

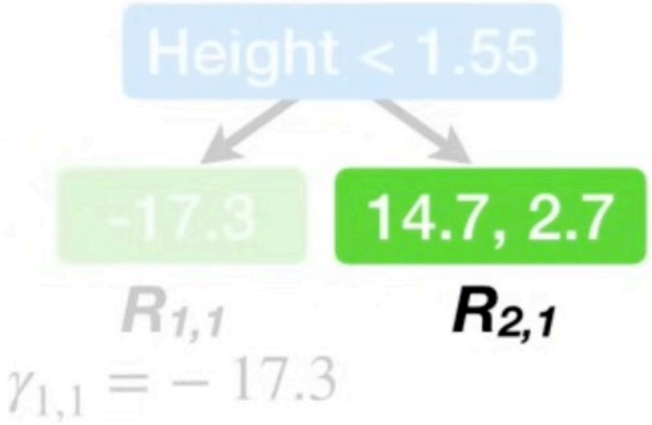


Now let's solve for the
Output Value for $R_{2,1}$



That means ***j*** = **2**, since this is the second leaf, and ***m*** = **1**, since this is still the first tree.

$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$



$$\gamma_{1,1} = -17.3$$

Height (m)	Favorite Color	Gender	Weight (kg)	$r_{i,1}$
1.6	Blue	Male	88	14.7
1.6	Green	Female	76	2.7
1.5	Blue	Female	56	-17.3

…plug in the **Observed Weights**…

$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}} \left[\frac{1}{2}(y_1 - (F_{m-1}(x_1) + \gamma))^2 + \frac{1}{2}(y_2 - (F_{m-1}(x_2) + \gamma))^2 \right]$$

…and plug in **73.3** for $F_{m-1}(x_1)$ and $F_{m-1}(x_2)$.

$F_0(x)$

73.3

$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}} \left[\frac{1}{2}(88 - (F_{m-1}(x_1) + \gamma))^2 + \frac{1}{2}(76 - (F_{m-1}(x_2) + \gamma))^2 \right]$$

Now simplify what's inside the parentheses…

$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}} \left[\frac{1}{2}(14.7 - \gamma))^2 + \frac{1}{2}(2.7 - \gamma))^2 \right]$$

$$\frac{d}{d\gamma}\left[\frac{1}{2}(14.7-\gamma))^2+\frac{1}{2}(2.7-\gamma))^2\right] \longrightarrow -14.7+\gamma+-2.7+\gamma$$

…and that gives us this derivative…

$$\gamma_{2,1}=\underset{\gamma}{\operatorname{argmin}}\left[\frac{1}{2}(14.7-\gamma))^2+\frac{1}{2}(2.7-\gamma))^2\right]$$

$$\frac{d}{d\gamma}\left[\frac{1}{2}(14.7-\gamma))^2+\frac{1}{2}(2.7-\gamma))^2\right] \longrightarrow \quad 2\gamma = 14.7 + 2.7$$

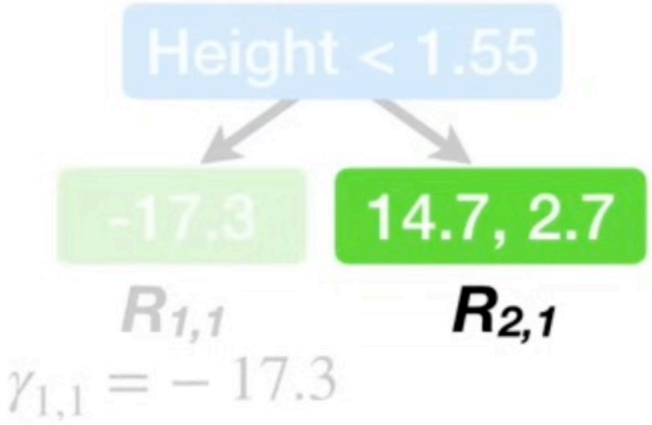
…and then we solve.

$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}}\left[\frac{1}{2}(14.7-\gamma))^2 + \frac{1}{2}(2.7-\gamma))^2\right]$$

$$\gamma = \frac{14.7 + 2.7}{2}$$

We end up with the average of the **Residuals** that ended in leaf $R_{2,1}$.

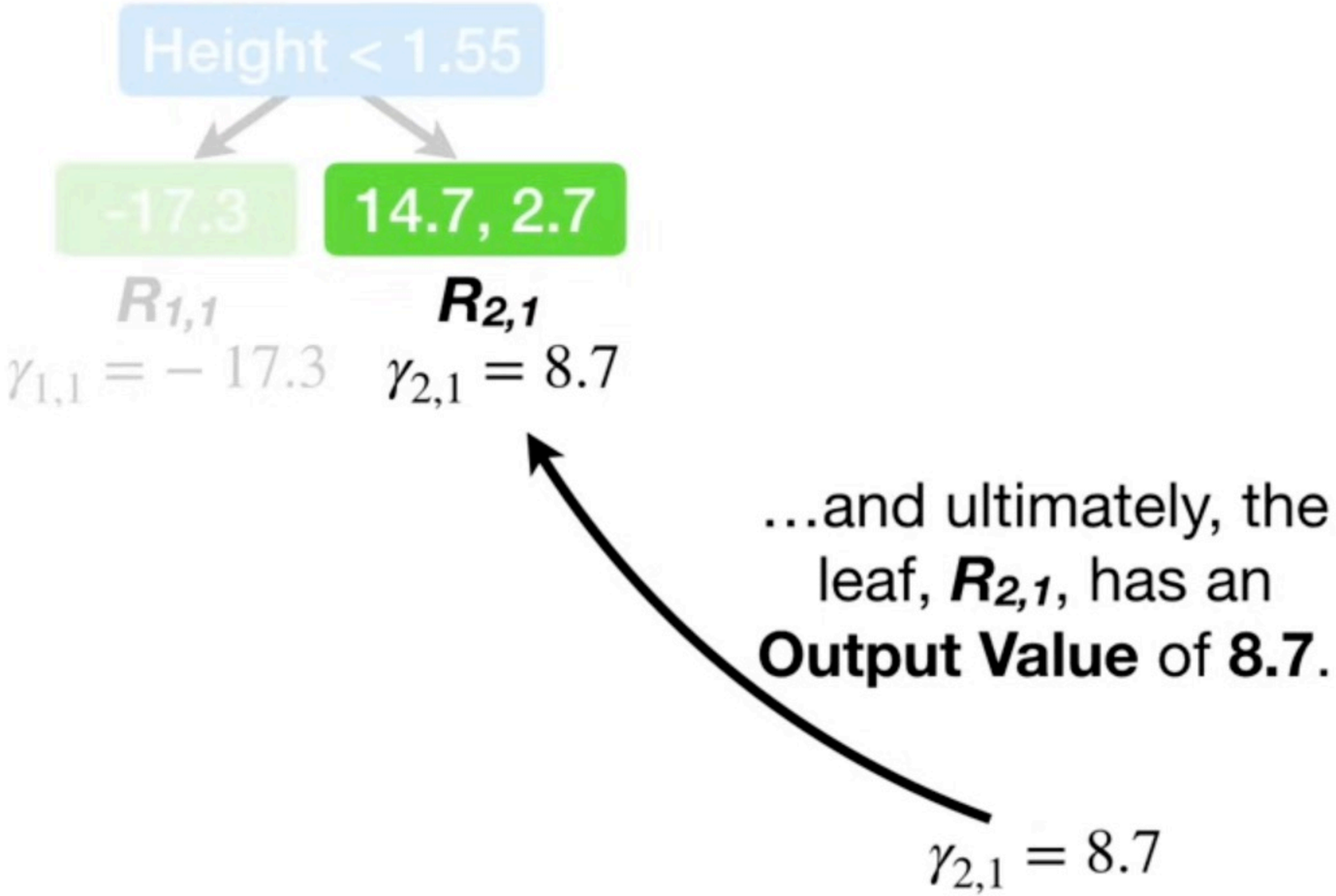
$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}} \left[\frac{1}{2}(14.7 - \gamma))^2 + \frac{1}{2}(2.7 - \gamma))^2 \right]$$

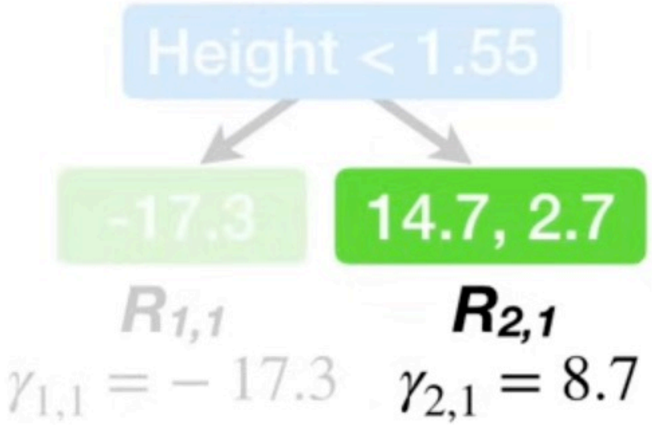


$$\gamma = \frac{14.7 + 2.7}{2} = 8.7$$

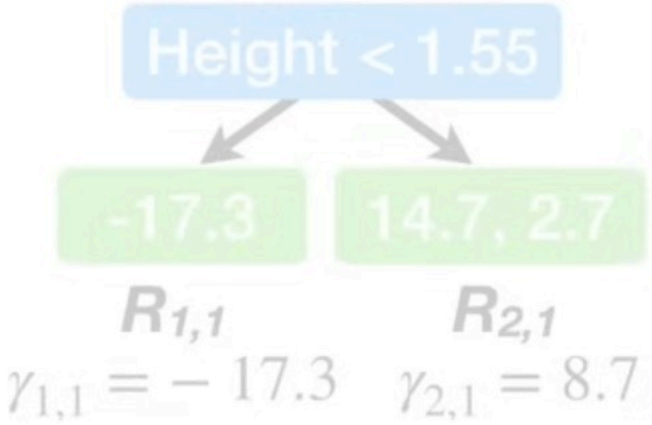
So the value for **gamma** that minimizes this equation is **8.7**…

$$\gamma_{2,1} = \underset{\gamma}{\operatorname{argmin}} \left[\frac{1}{2}(14.7 - \gamma))^2 + \frac{1}{2}(2.7 - \gamma))^2 \right]$$


Height < 1.55
-17.3
14.7, 2.7
$R_{1,1}$
$R_{2,1}$
$\gamma_{1,1} = -17.3$
$\gamma_{2,1} = 8.7$
…and ultimately, the leaf, $R_{2,1}$, has an **Output Value** of **8.7**.
$\gamma_{2,1} = 8.7$



We just saw that the **Output Value** for this leaf, $R_{2,1}$, is the average of the residuals that ended up here.



Given our choice of **Loss Function**, the **Output Values** are *always* the average of the **Residuals** that end up in the same leaf.

$$\frac{1}{2}(\textbf{Observed} - \textbf{Predicted})^2$$

Now let's do **Part D**!!!

Step 2: for $m = 1$ to M:

(A) Compute $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,...,n$

(B) Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for $j = 1 \ldots J_m$

(C) For $j = 1 \ldots J_m$ compute $\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

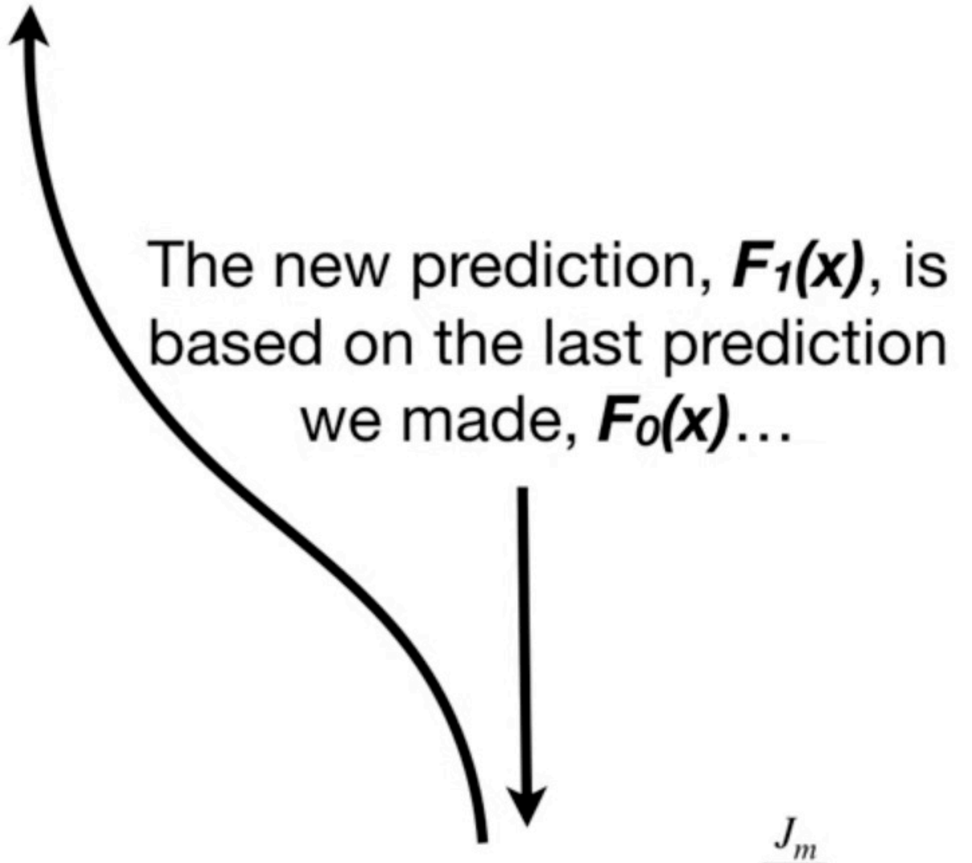
(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

Since this is our first pass through **Step 2** and ***m*** = **1**, this new prediction will be called $F_1(x)$.

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

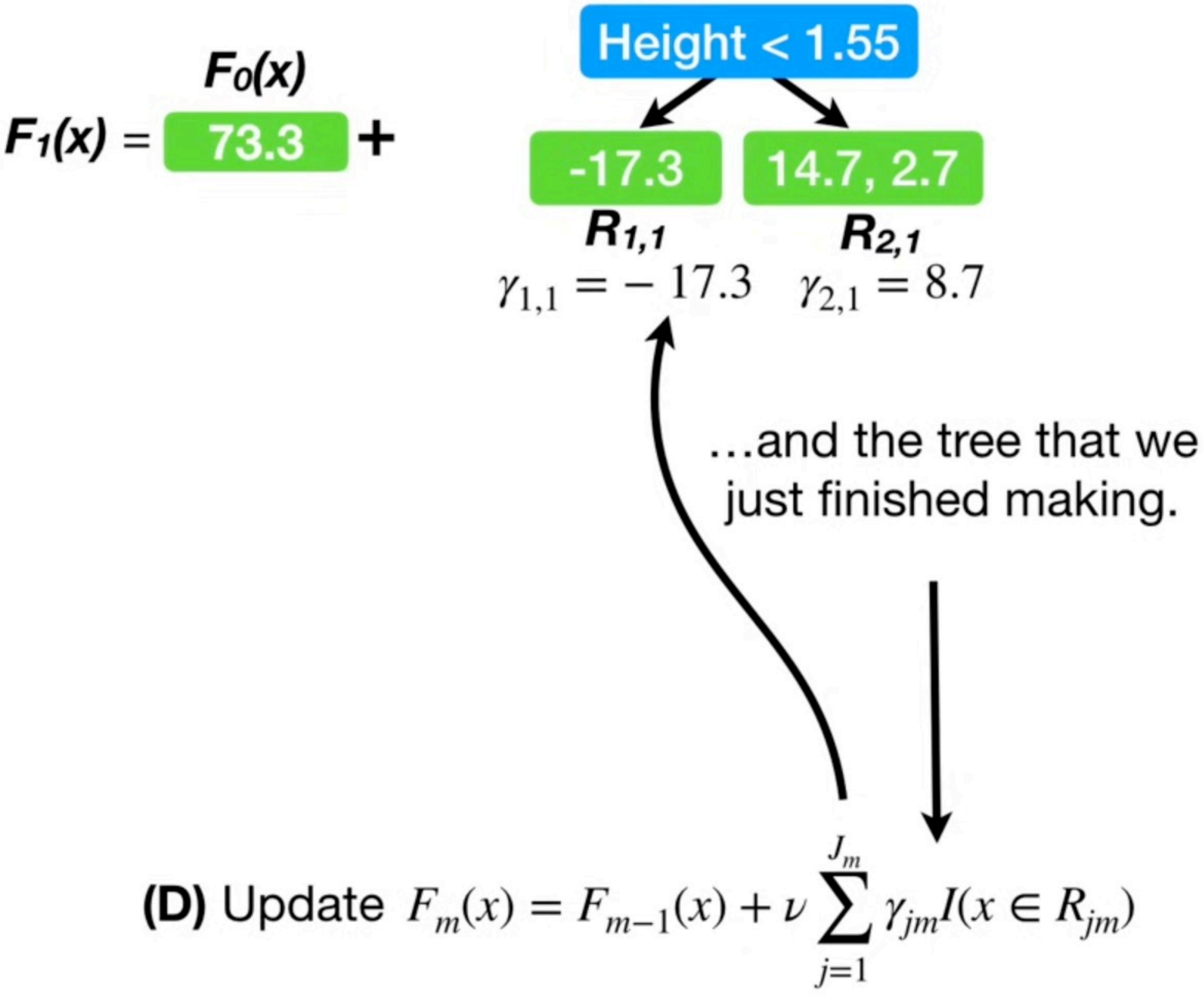
$F_0(x)$

$F_1(x) =$ 73.3

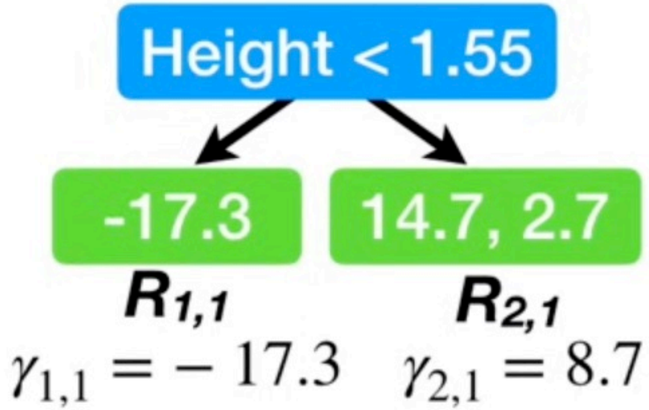


The new prediction, $F_1(x)$, is based on the last prediction we made, $F_0(x)$…

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$


$F_0(x)$
$F_1(x) =$
73.3
+
Height < 1.55
-17.3
14.7, 2.7
$R_{1,1}$
$R_{2,1}$
$\gamma_{1,1} = -17.3$
$\gamma_{2,1} = 8.7$
…and the tree that we just finished making.
(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

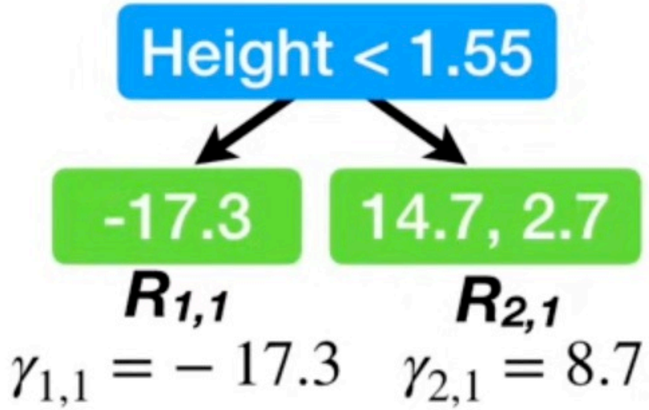
$F_1(x) = F_0(x) + 73.3$



The summation says we should add up the **Output Values**, **gamma$_{j,m}$**'s, for all the leaves, $R_{j,m}$, that a sample, x, can be found in.

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

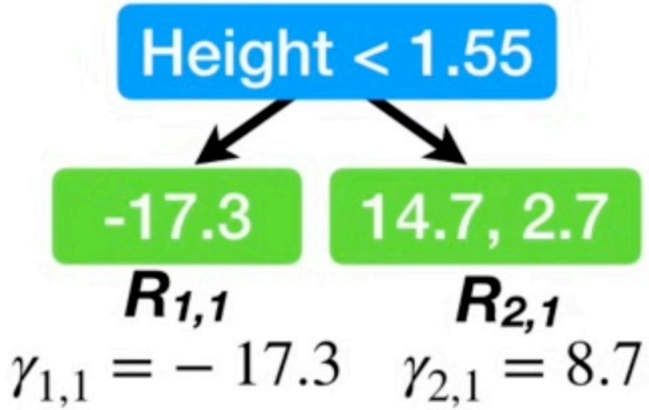
$F_1(x) = F_0(x) + \gamma_{1,1} I(x \in R_{1,1}) + \gamma_{2,1} I(x \in R_{2,1})$



The last thing in this equation is this Greek character "**nu**".

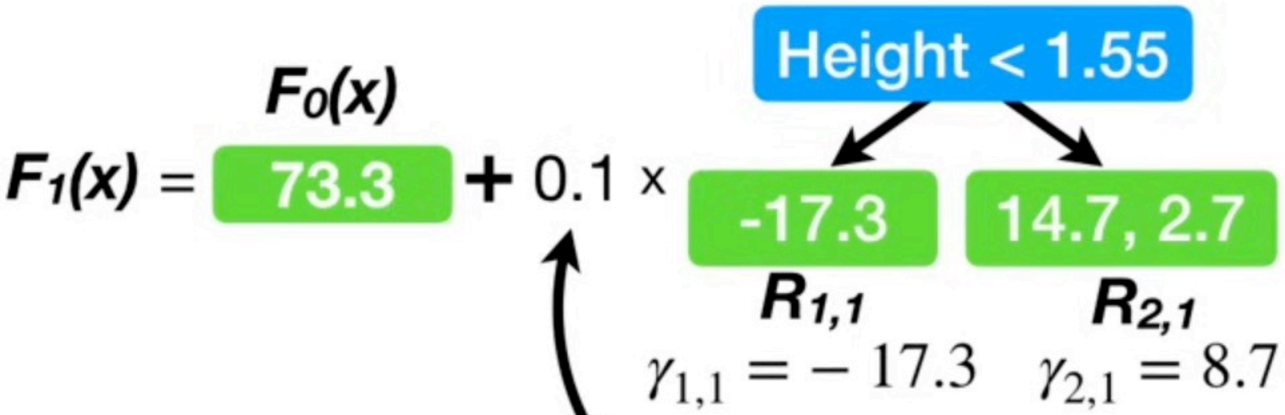
(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

$F_1(x) = F_0(x) + \gamma_{1,1} I(x \in R_{1,1}) + \gamma_{2,1} I(x \in R_{2,1})$



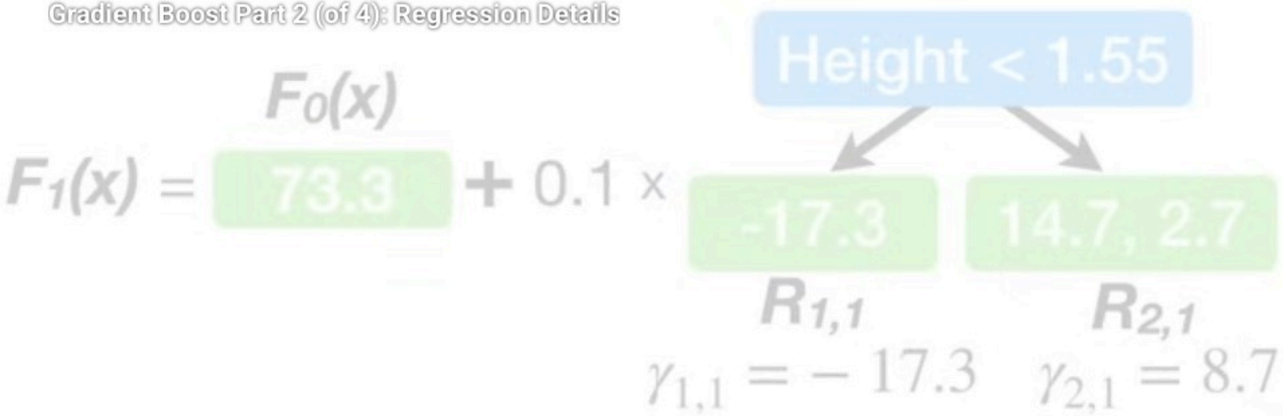
A small **Learning Rate** reduces the effect each tree has on the final prediction, and this improves accuracy in the long run.

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$



In this example, we'll set **nu** to **0.1**.

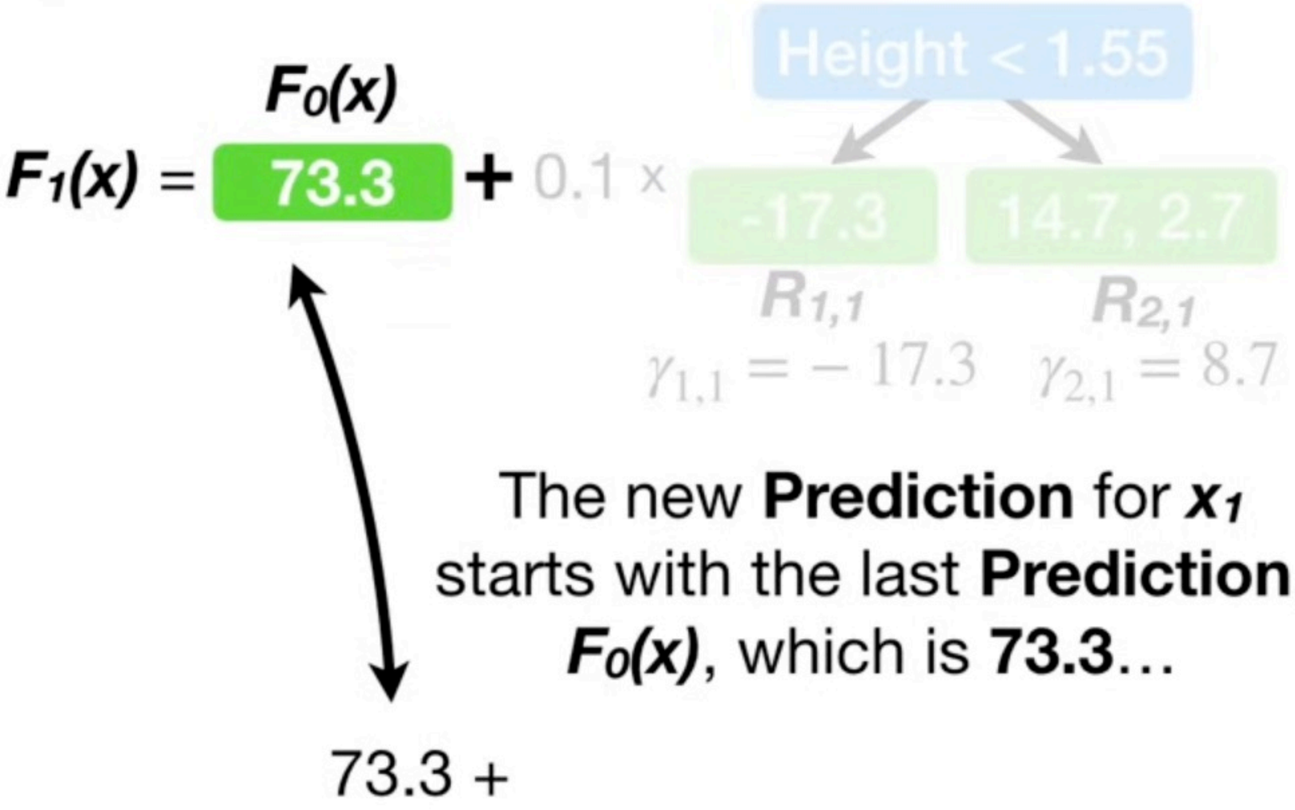
(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$



Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

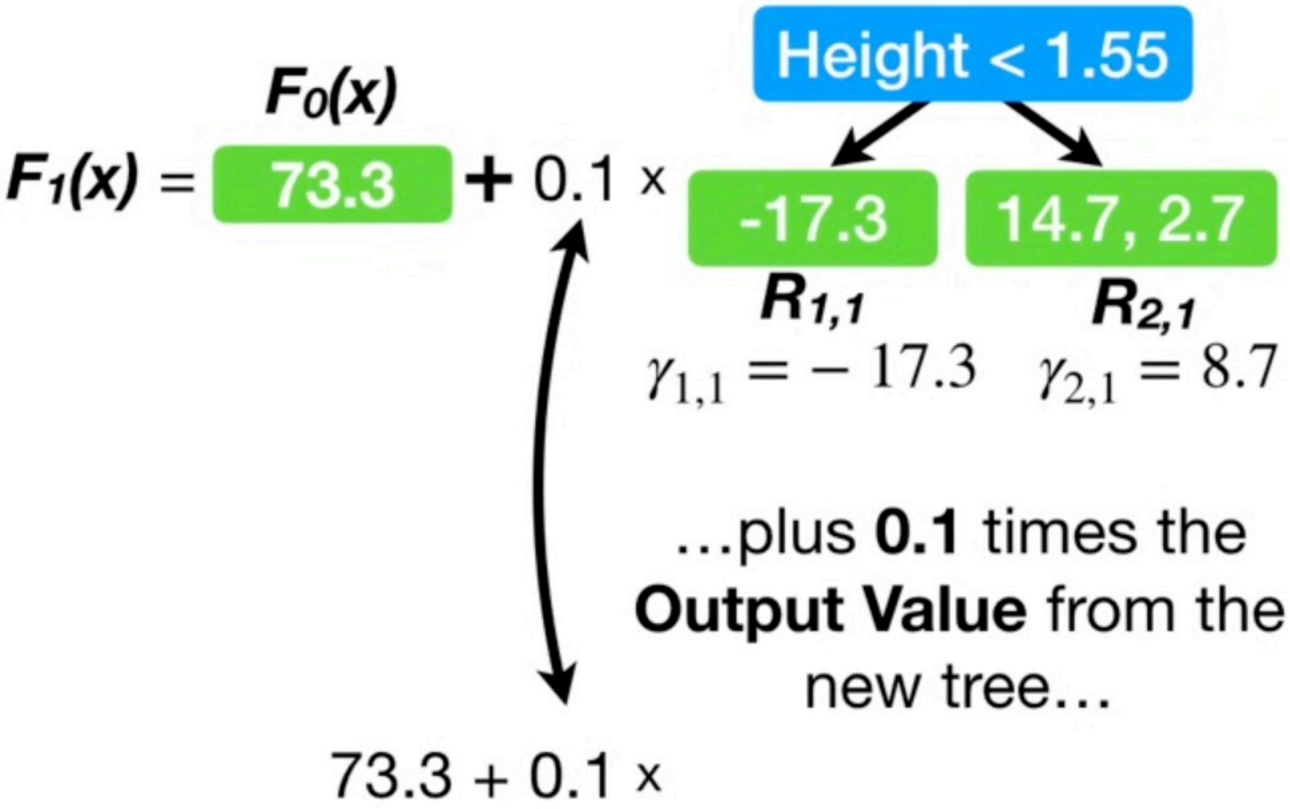
Now we will use $F_1(x)$ to make new **Predictions** for each sample.

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$



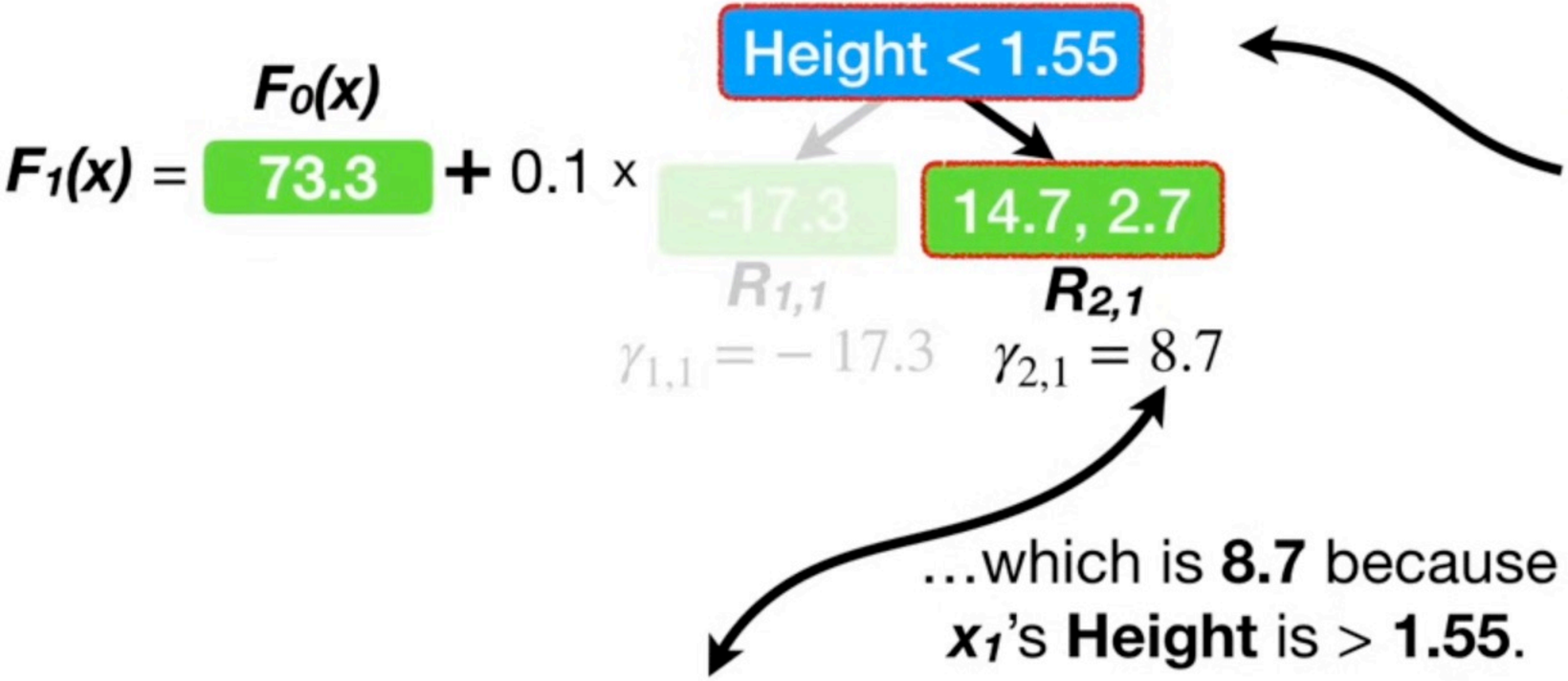
Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$



Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$



Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

$F_1(x) = F_0(x) + 0.1 \times$

Height (m)	Favorite Color	Gender	Weight (kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56

Height < 1.55

-17.3 ($R_{1,1}$)

14.7, 2.7 ($R_{2,1}$)

$\gamma_{1,1} = -17.3$

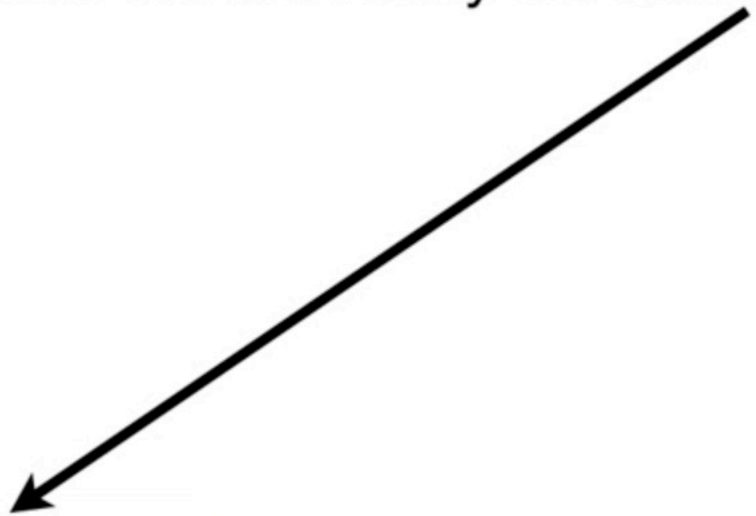
$\gamma_{2,1} = 8.7$

$73.3 + 0.1 \times 8.7 = 74.2$

The new **Prediction** for the first sample is **74.2**, which is slightly closer to the **Observed Weight**, **88**, than the first **Prediction**, **73.3**.

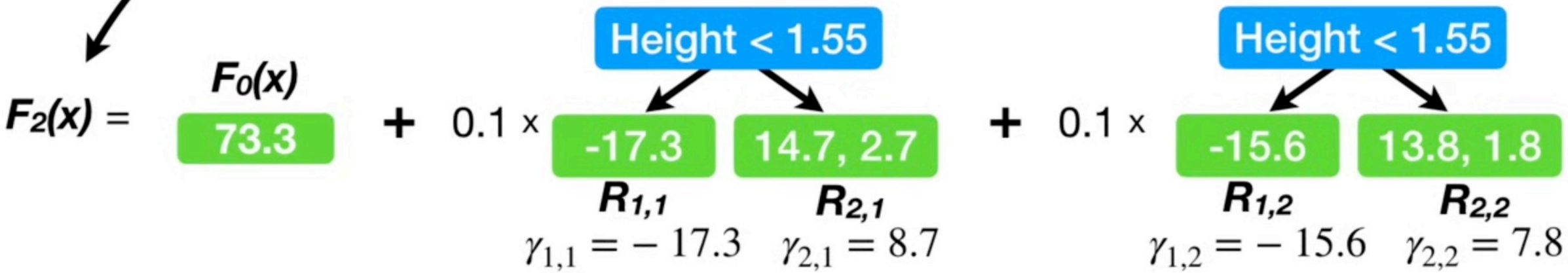
(D) Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

Now we are ready for **Gradient Boost's** 3rd and final step!



Step 3: Output $F_M(x)$

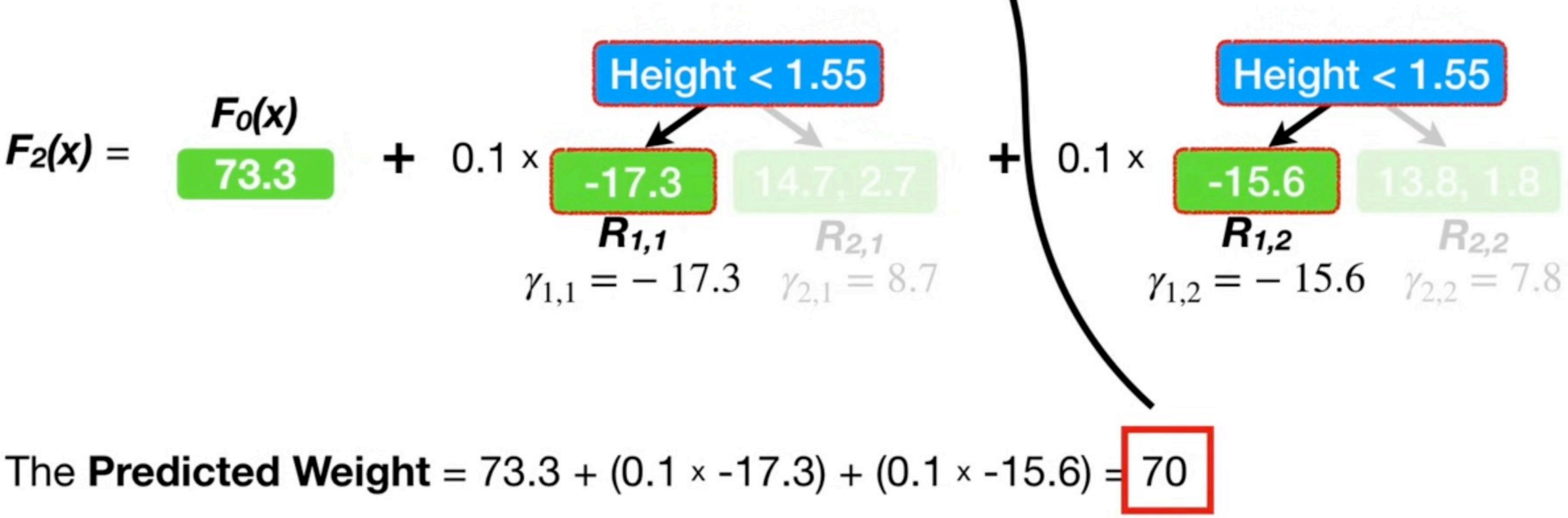
If $M = 2$, then $F_2(x)$ is the output from the **Gradient Boost** algorithm.



Step 3: Output $F_M(x)$

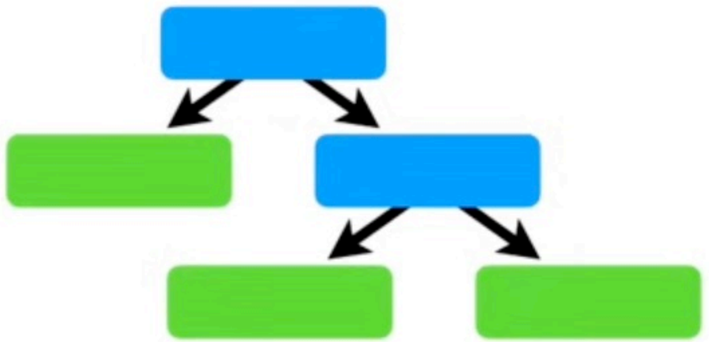
Height (m)	Favorite Color	Gender	Weight (kg)
1.4	Green	Male	70

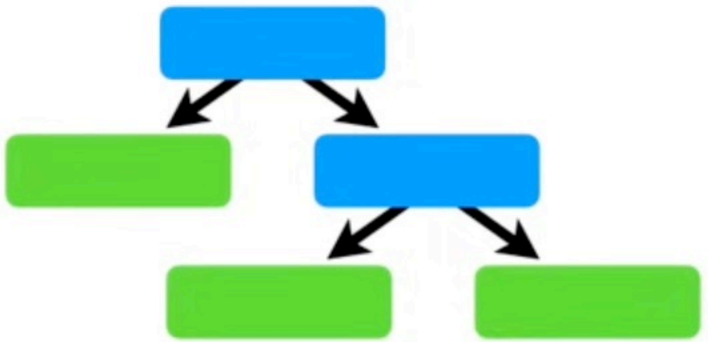
Gradient Boost
Predicts that this person **Weighs 70** kg.

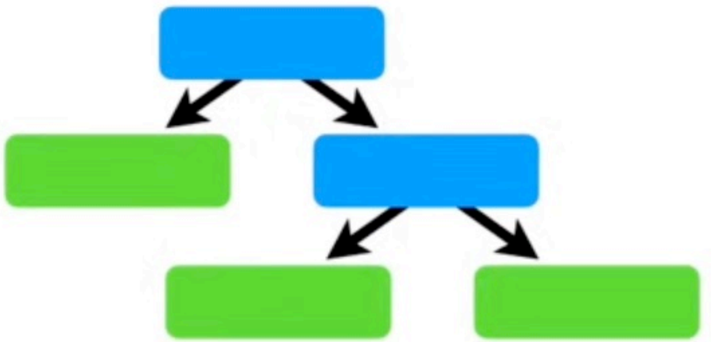


The **Predicted Weight** = 73.3 + (0.1 × -17.3) + (0.1 × -15.6) = 70

NOTE: Before we go, I want to remind you that **Gradient Boost** usually uses trees larger than stumps.

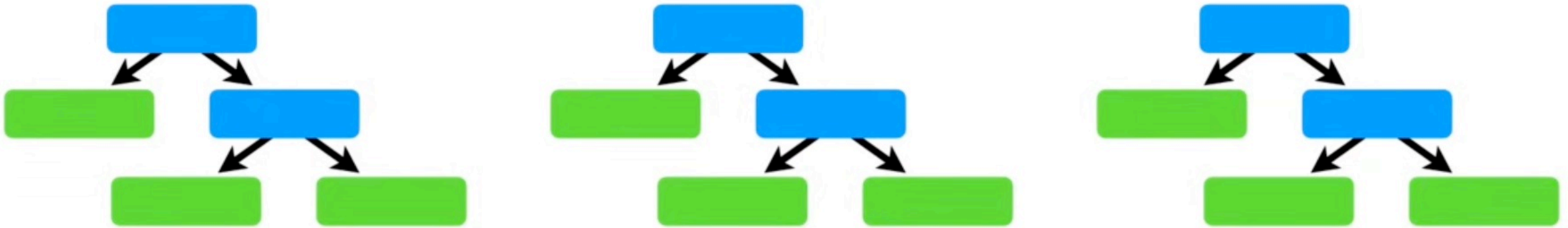






NOTE: Before we go, I want to remind you that **Gradient Boost** usually uses trees larger than stumps.

I only used stumps in this tutorial because our **Training Dataset** was so darn small.



⭐ SUPER INTUITIVE ANALOGY (you'll love this)

Think of Gradient Boosting as a shooter trying to hit a target:

- First shot misses → that's your initial model
- Second shot is made by looking at **where you missed → direction of correction**
- Third shot corrects remaining error
- Each new shot corrects remaining error (gradient)

The "direction of correction" is the **gradient**.

Each correction is **learned by a decision tree**.

⭐ FINAL ANSWER (interview-quality)

> "We compute the gradient because it tells us the direction in which predictions must change to reduce the loss.
> A decision tree is then trained to approximate that gradient — meaning the tree learns how much correction is needed in different regions of feature space.
> Adding this tree to the model moves predictions in the direction that reduces loss, exactly like gradient descent but in function space."