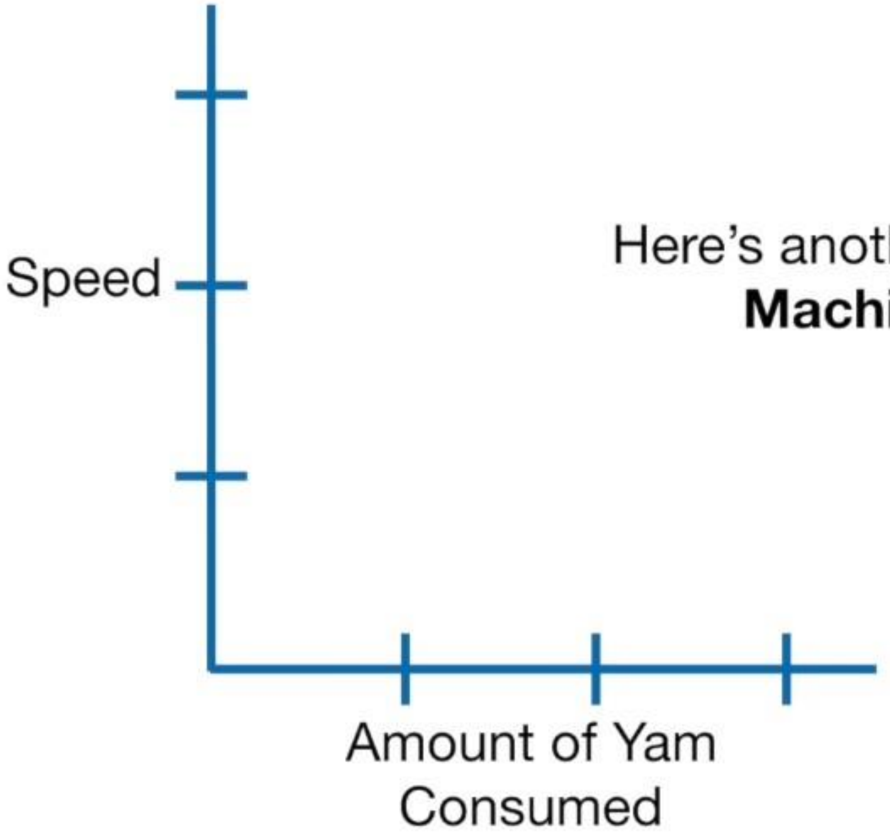
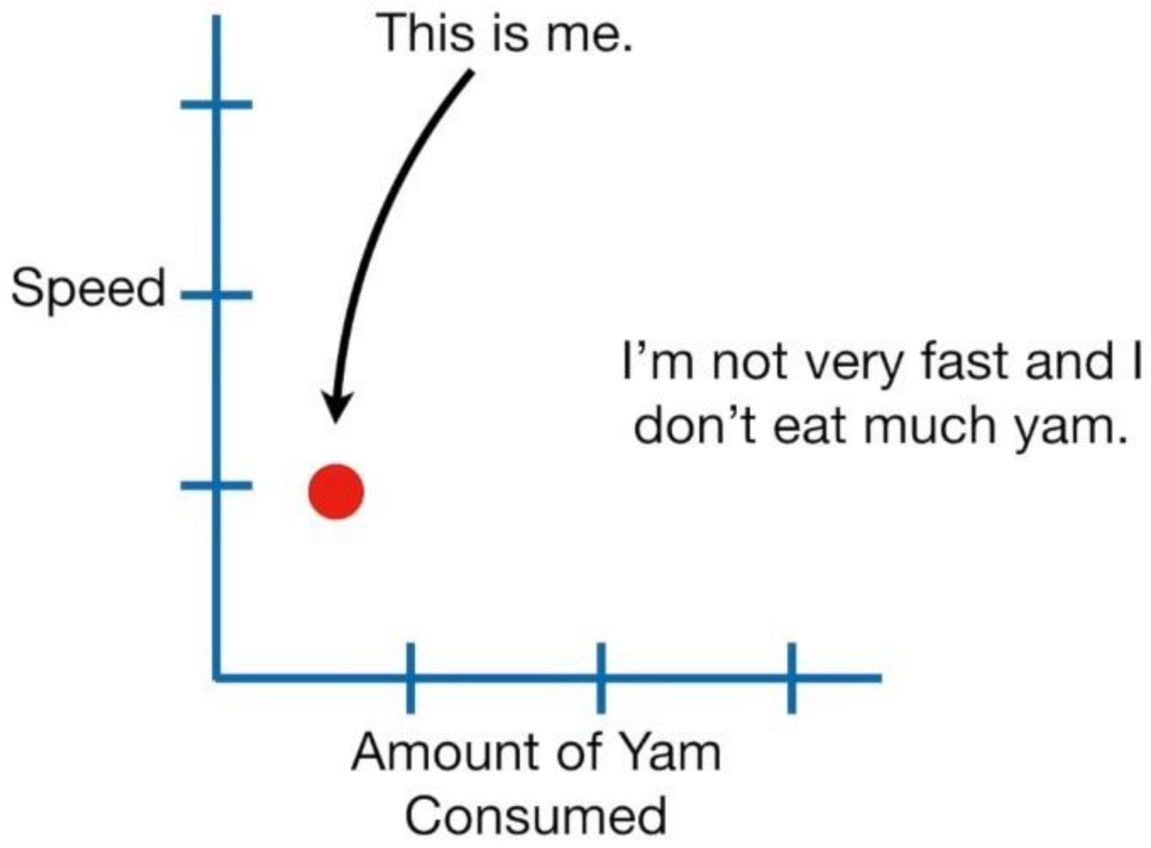
Machine Learning Methods

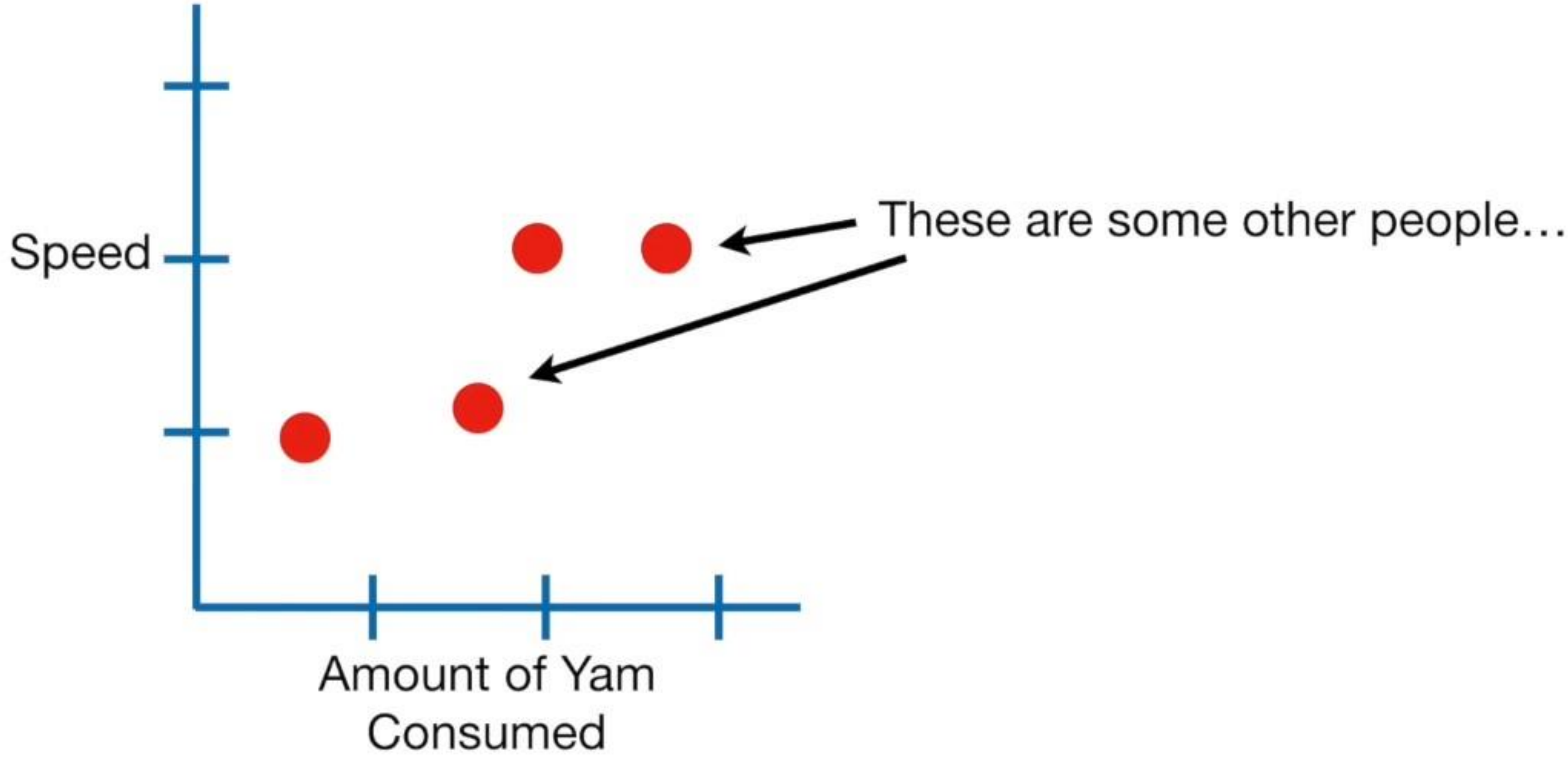
Disclaimer

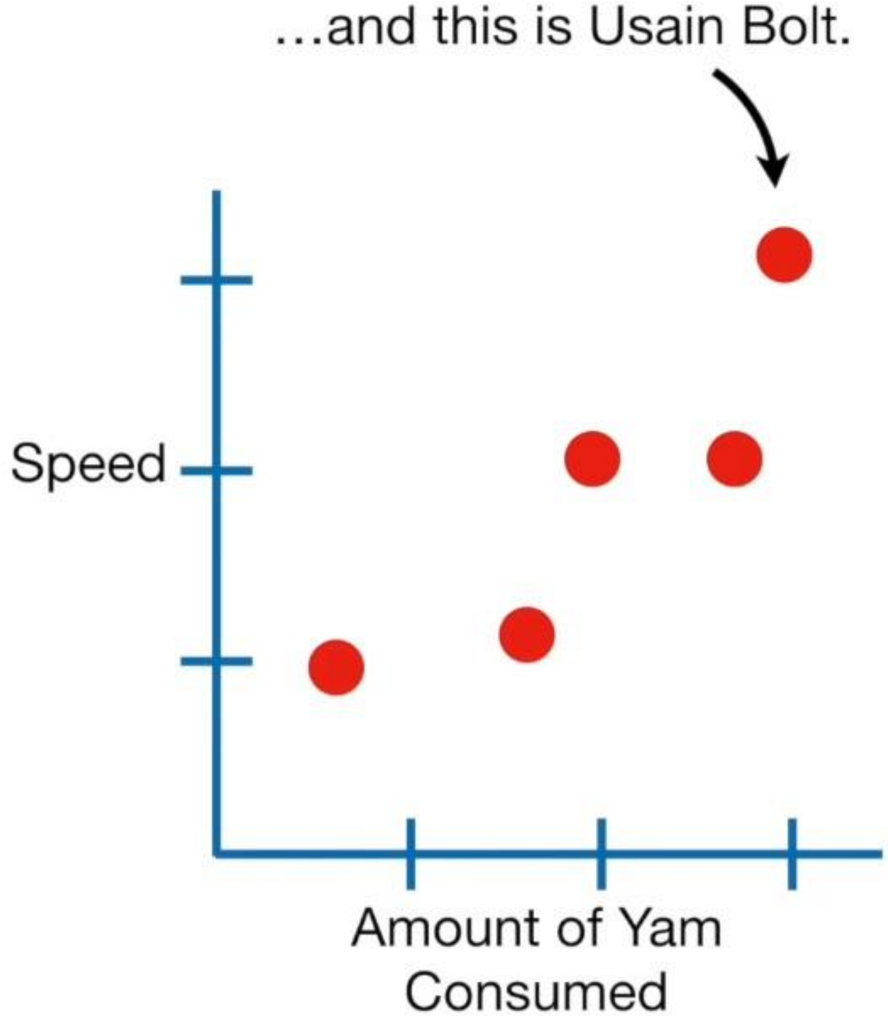
- Machine Learning is a vast domain
- Emphasize on Practical / Industry Implementation
- Take it as Experience Sharing Session

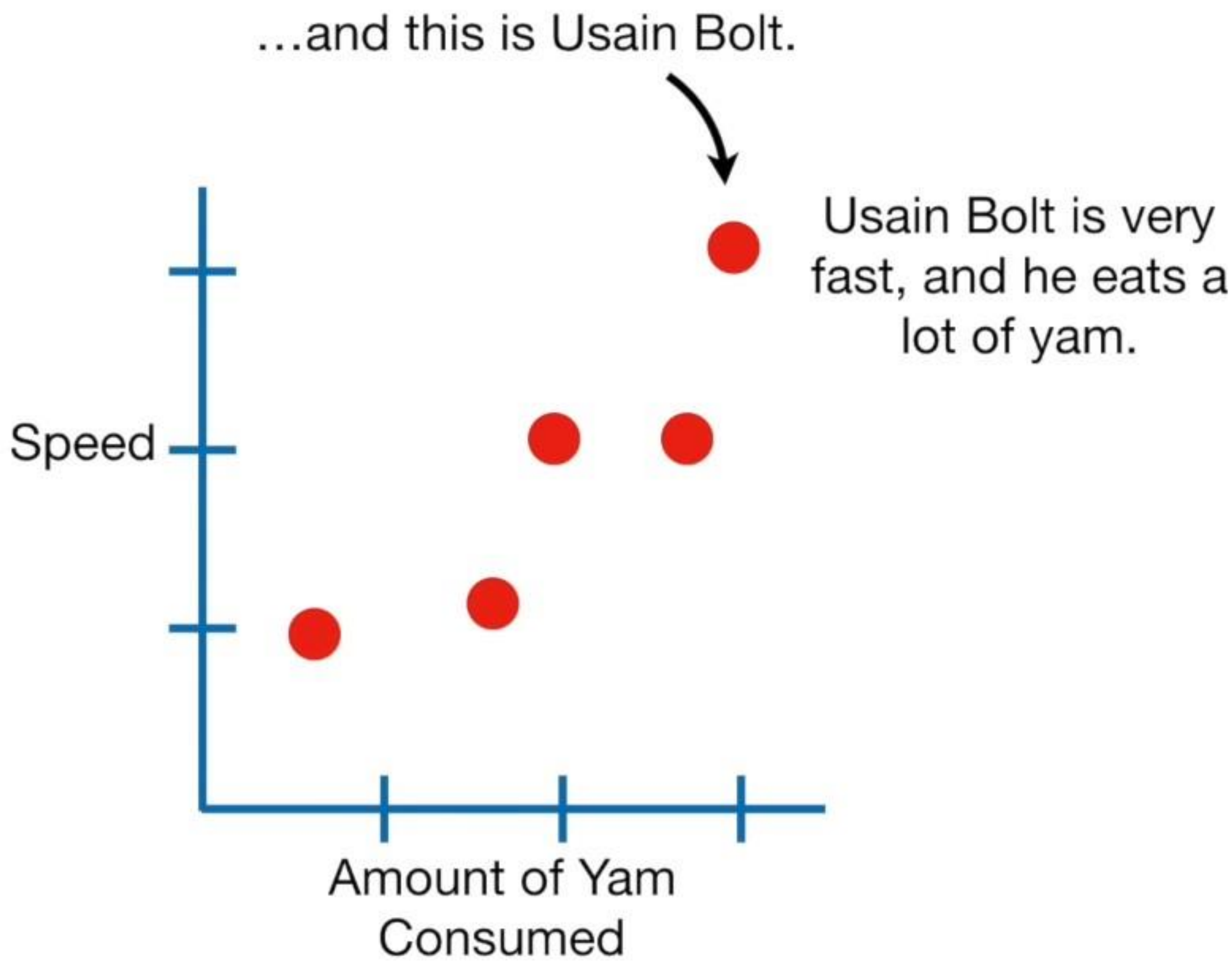


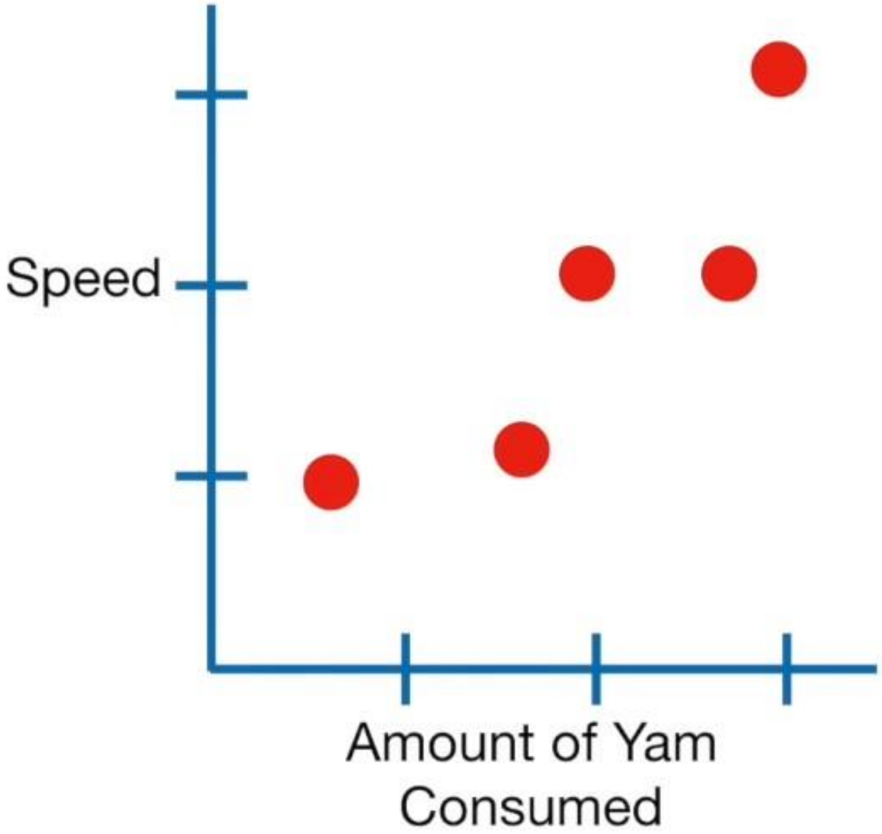
Here's another silly example of **Machine Learning**.


This is me.
Speed
I'm not very fast and I
don't eat much yam.
Amount of Yam
Consumed

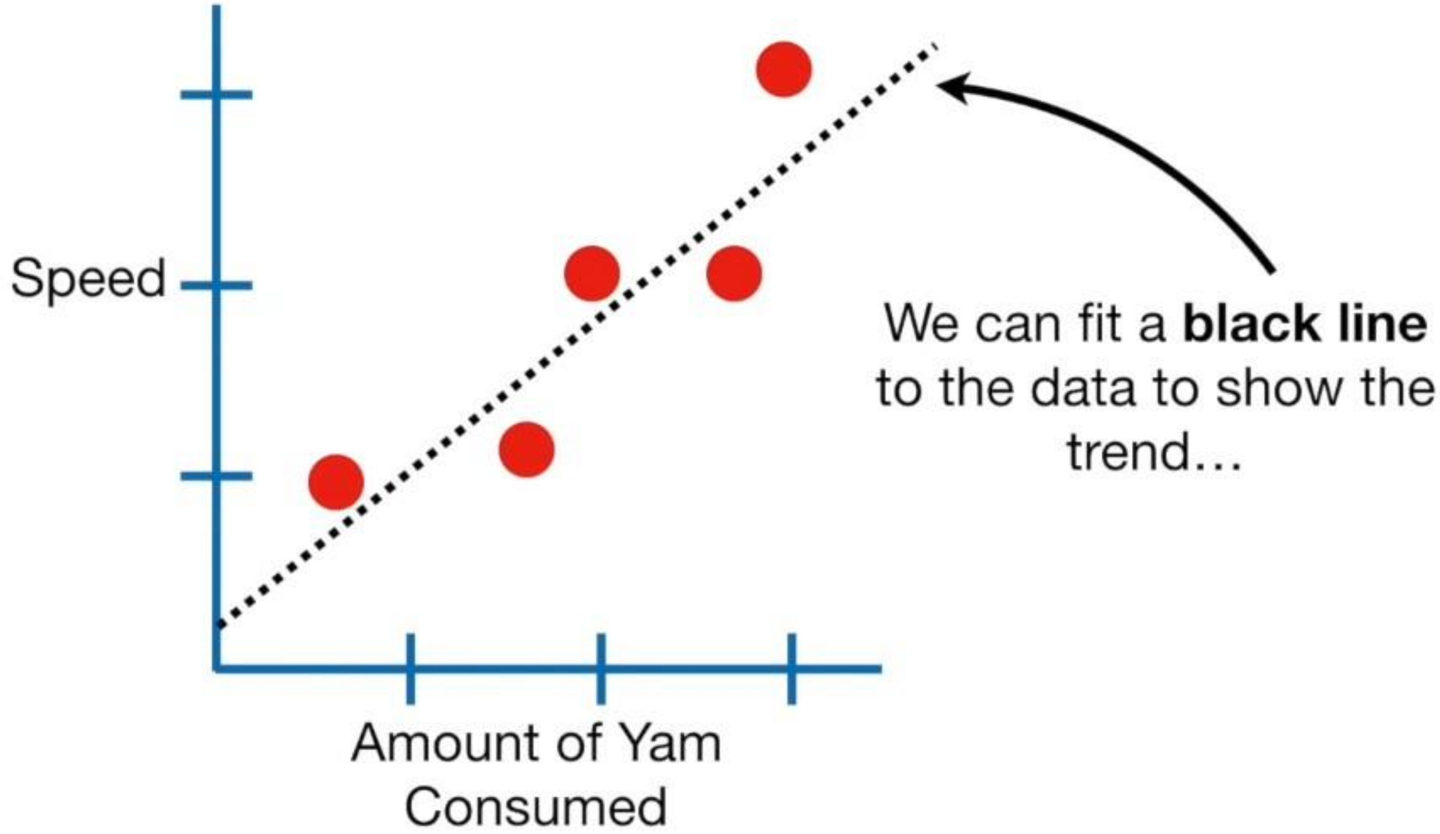

These are some other people…
Speed
Amount of Yam
Consumed

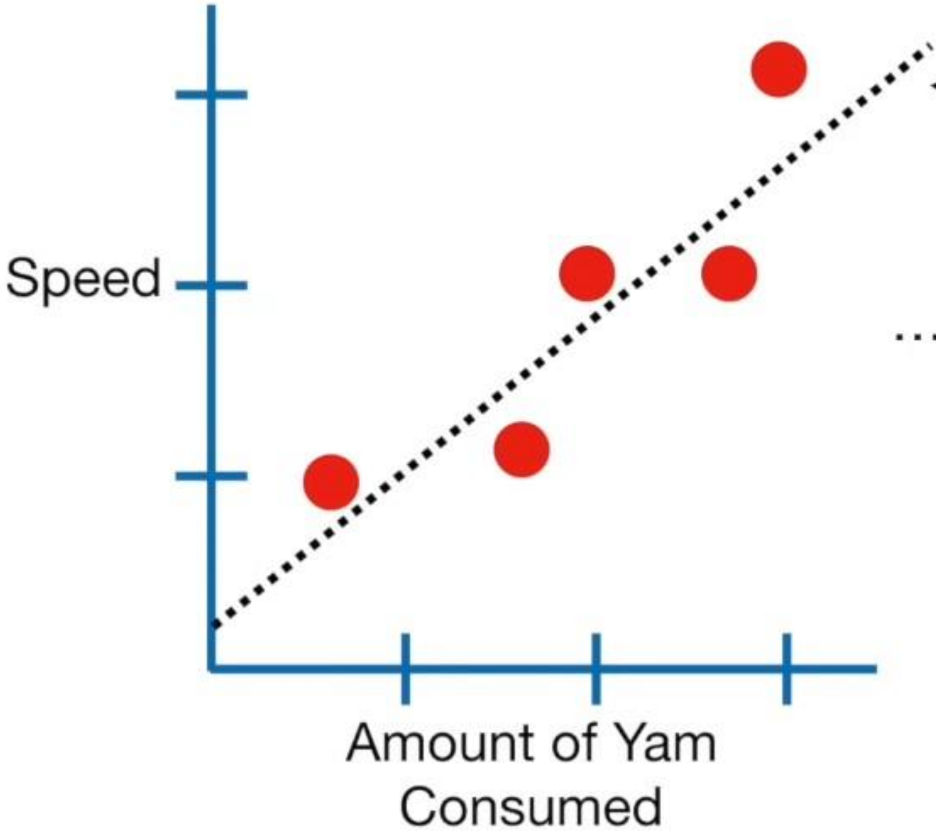

...and this is Usain Bolt.
Speed
Amount of Yam
Consumed


…and this is Usain Bolt.
Usain Bolt is very fast, and he eats a lot of yam.
Speed
Amount of Yam Consumed

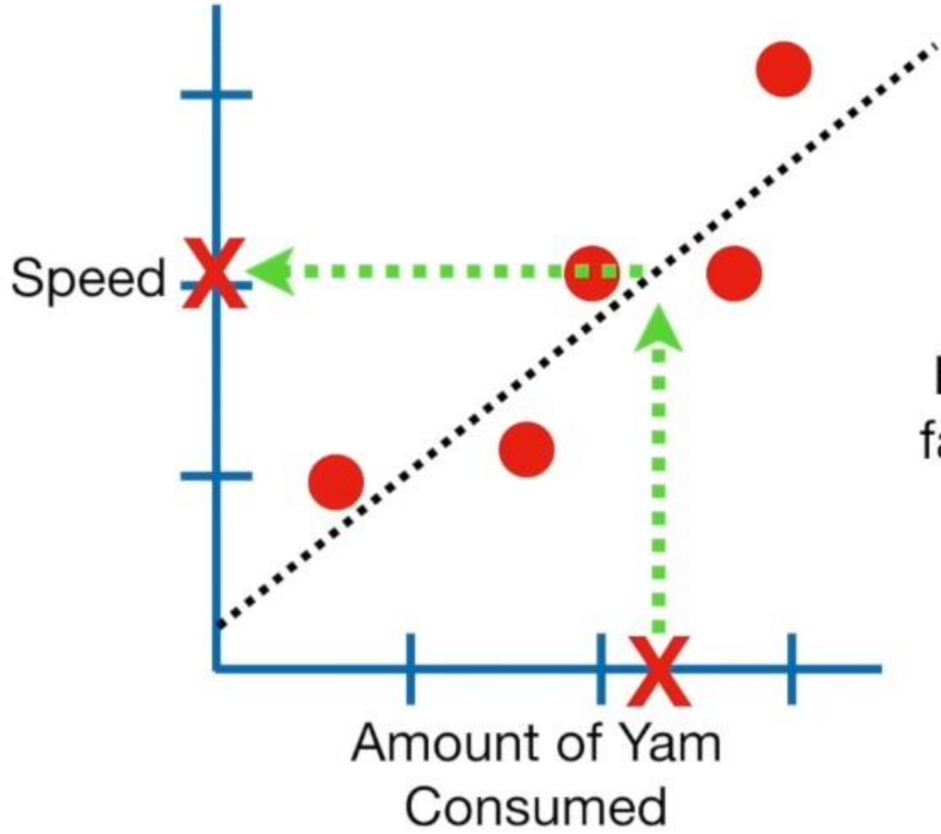


Given this pretend data, we see that the more yam someone eats, the faster they run the 100 meter dash.

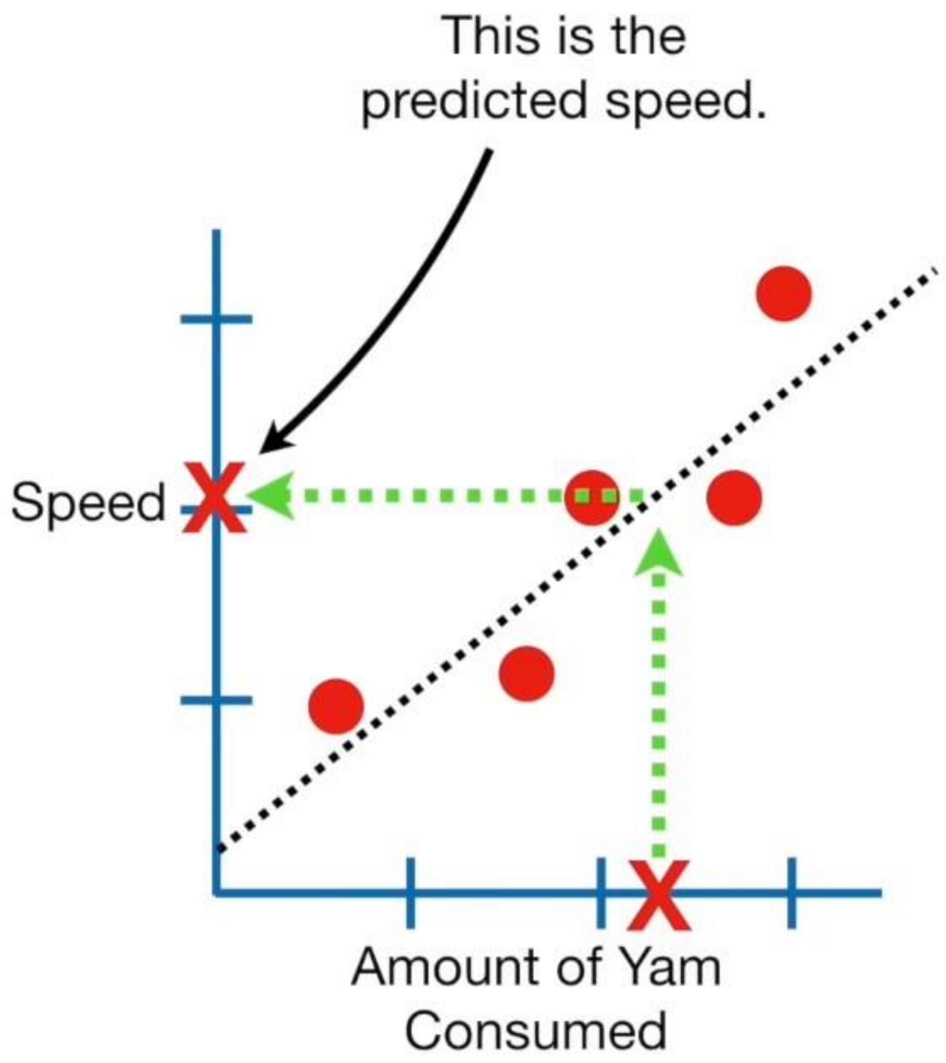

Speed
Amount of Yam Consumed
We can fit a **black line** to the data to show the trend…

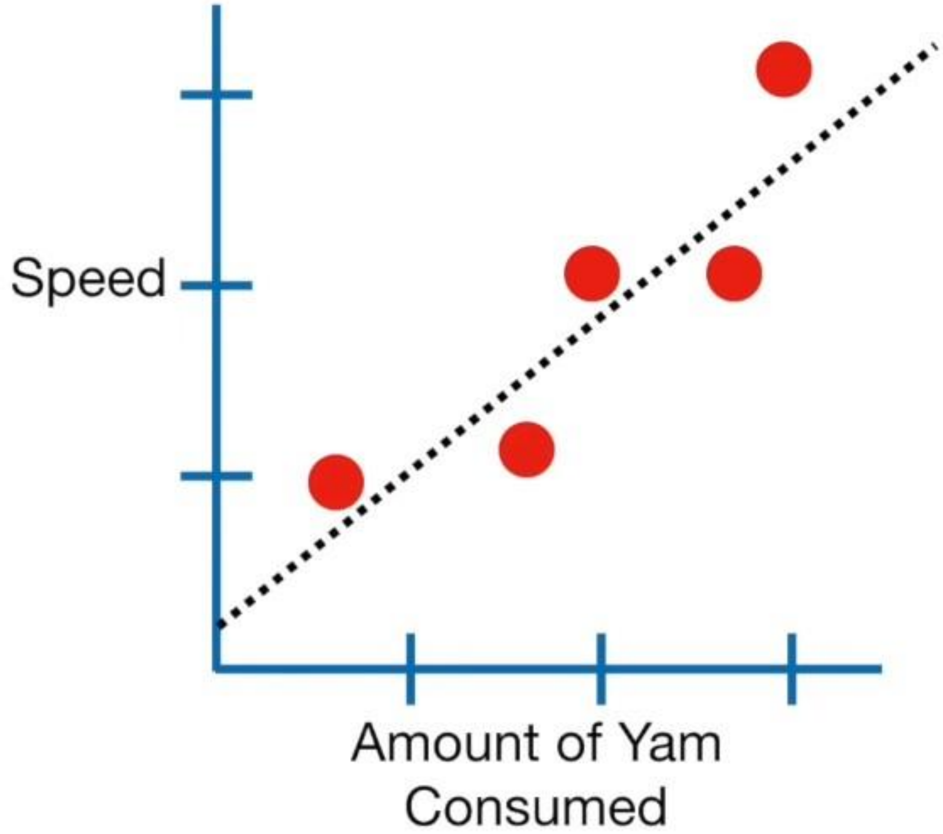


…but we can also use the **black line** to make ***predictions***.



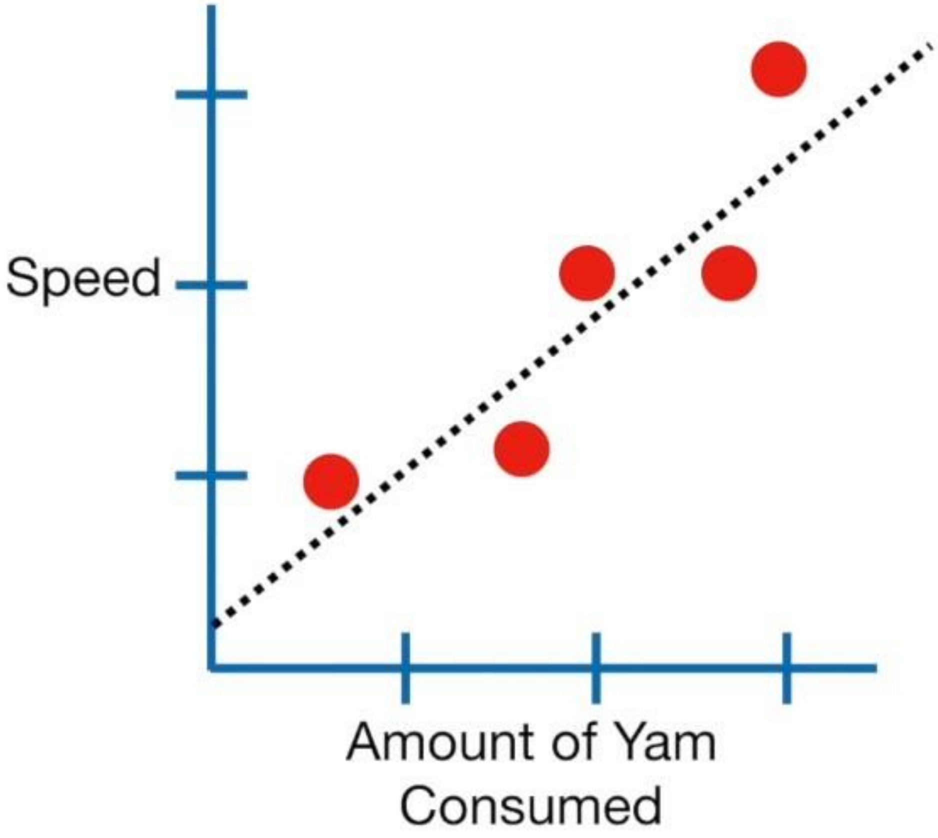
…then we could use the **black line** to predict how fast that person might run.

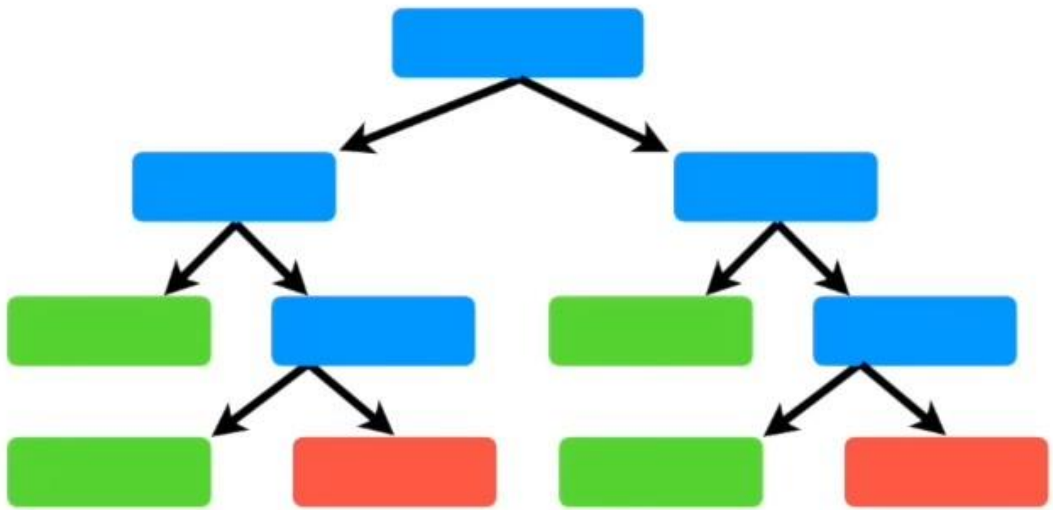

This is the
predicted speed.
Speed
Amount of Yam
Consumed



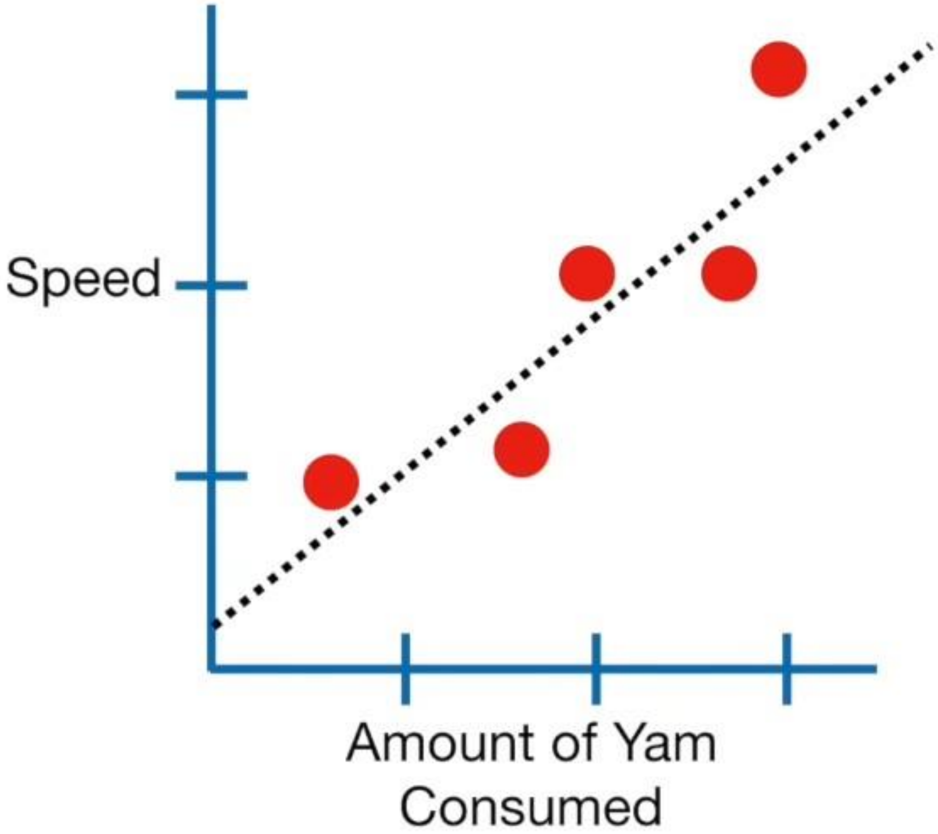
The **black line** is a type of **Machine Learning** because we can use it to make ***predictions***.

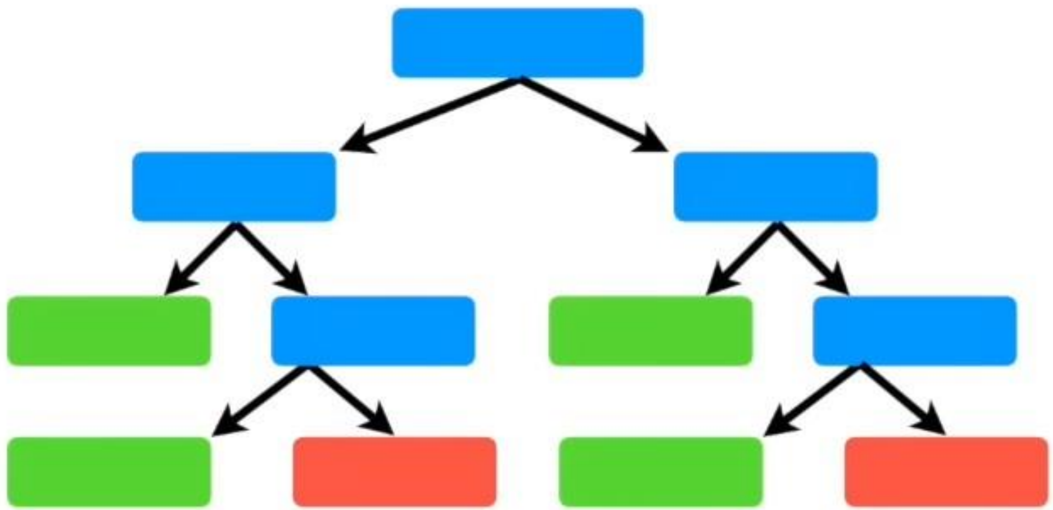
In general, **Machine Learning** is all about making ***predictions*** and ***classifications***.

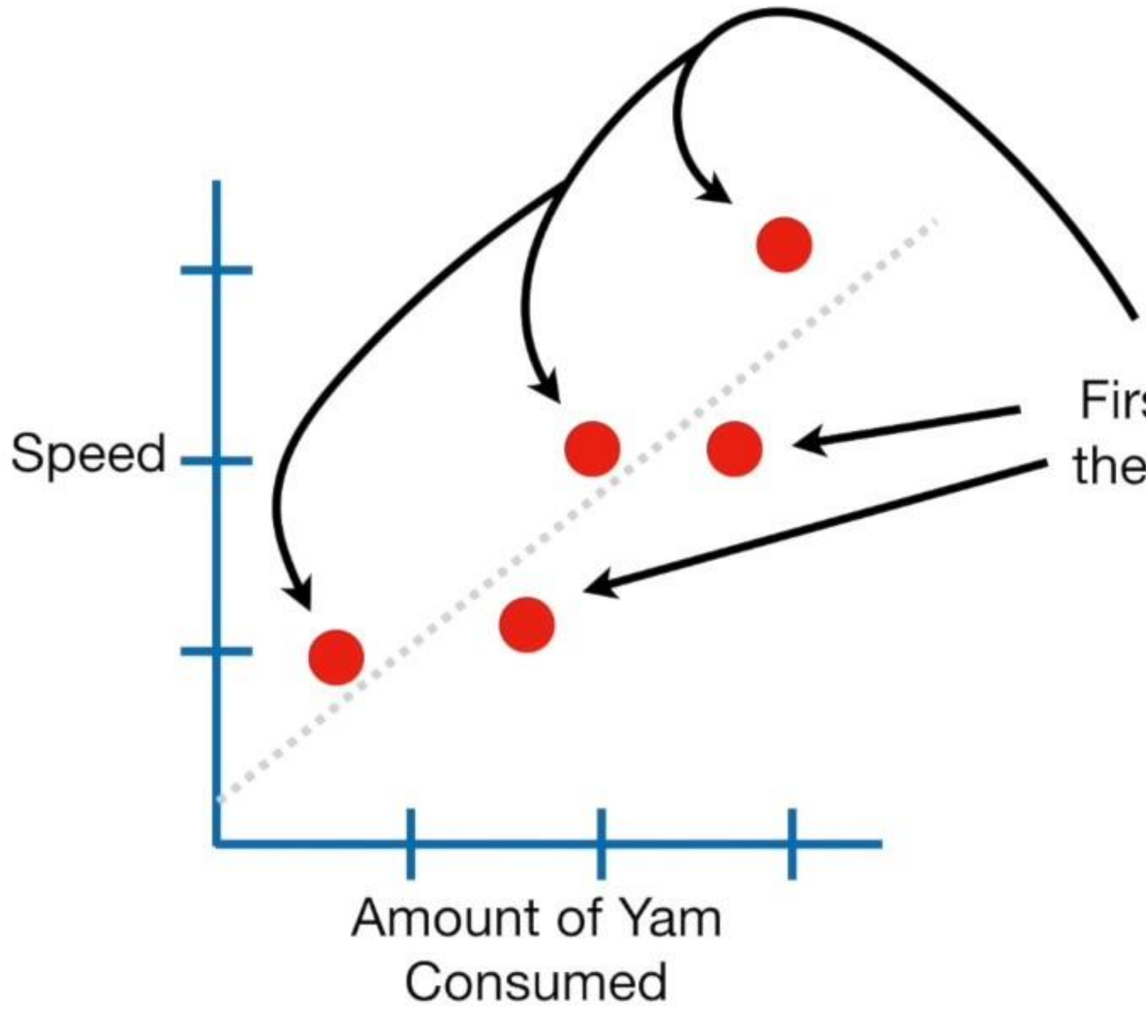




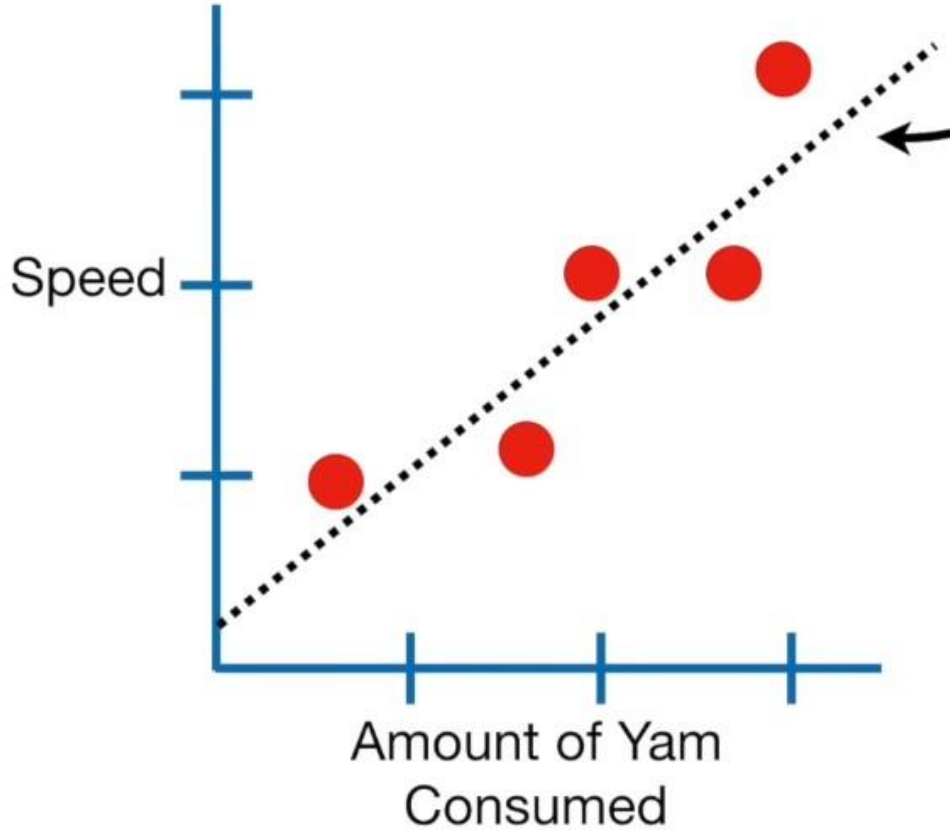
Now that we can make ***predictions*** and ***classifications***, let's talk about some of the main ideas in **Machine Learning**.



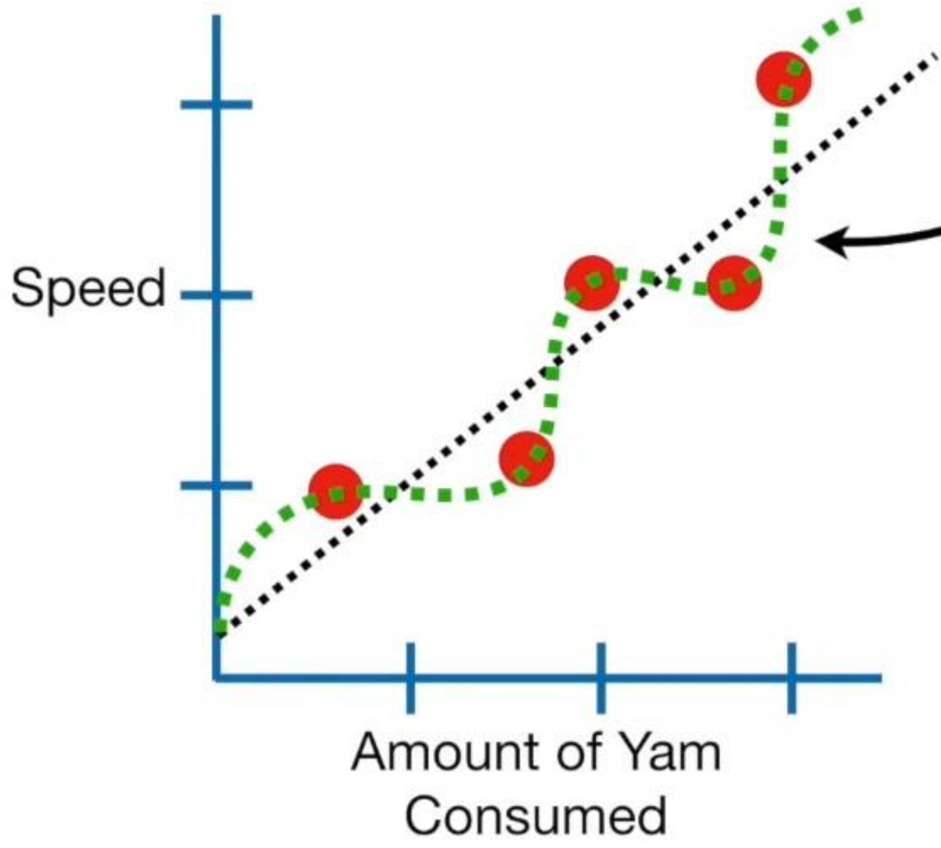




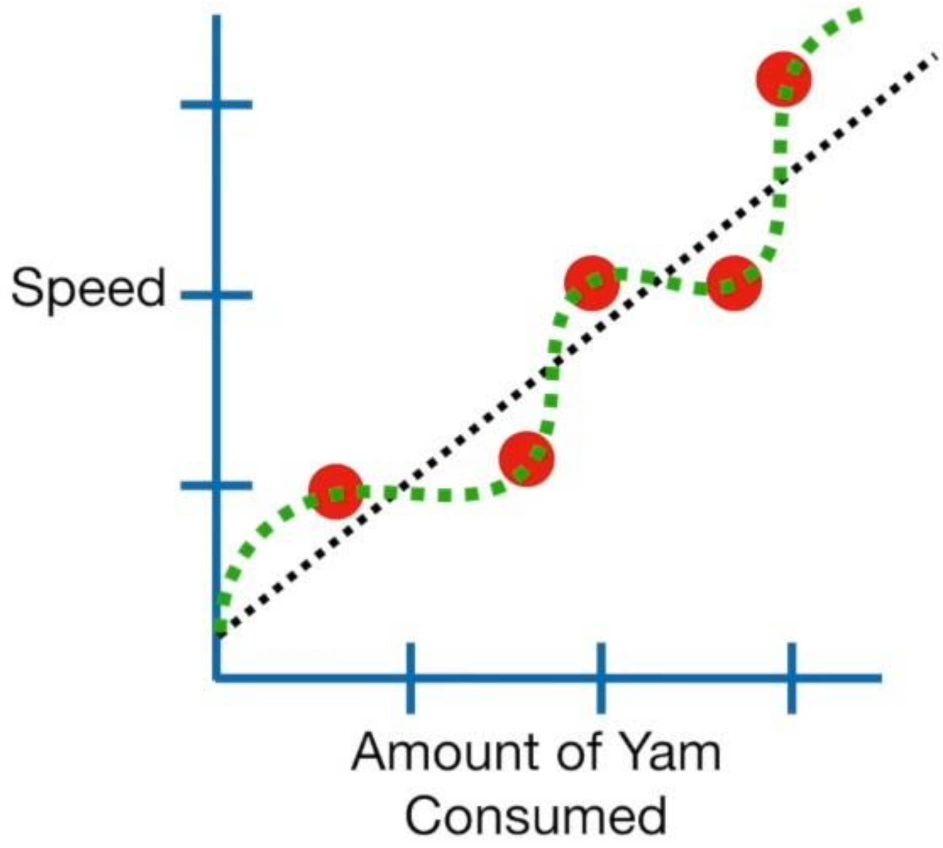
First of all, in **Machine Learning Lingo**, the original data is called **Training Data**.



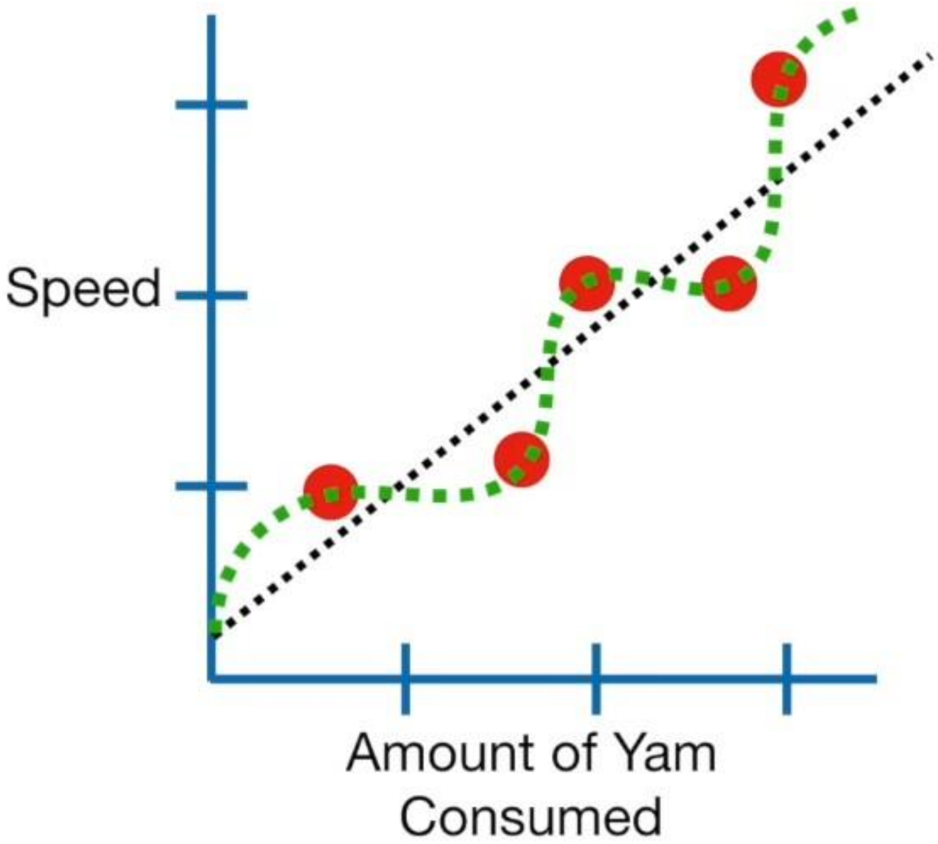
So the **black line** is fit to
Training Data.



Alternatively, we could have fit a **green squiggle** to the **Training Data**.

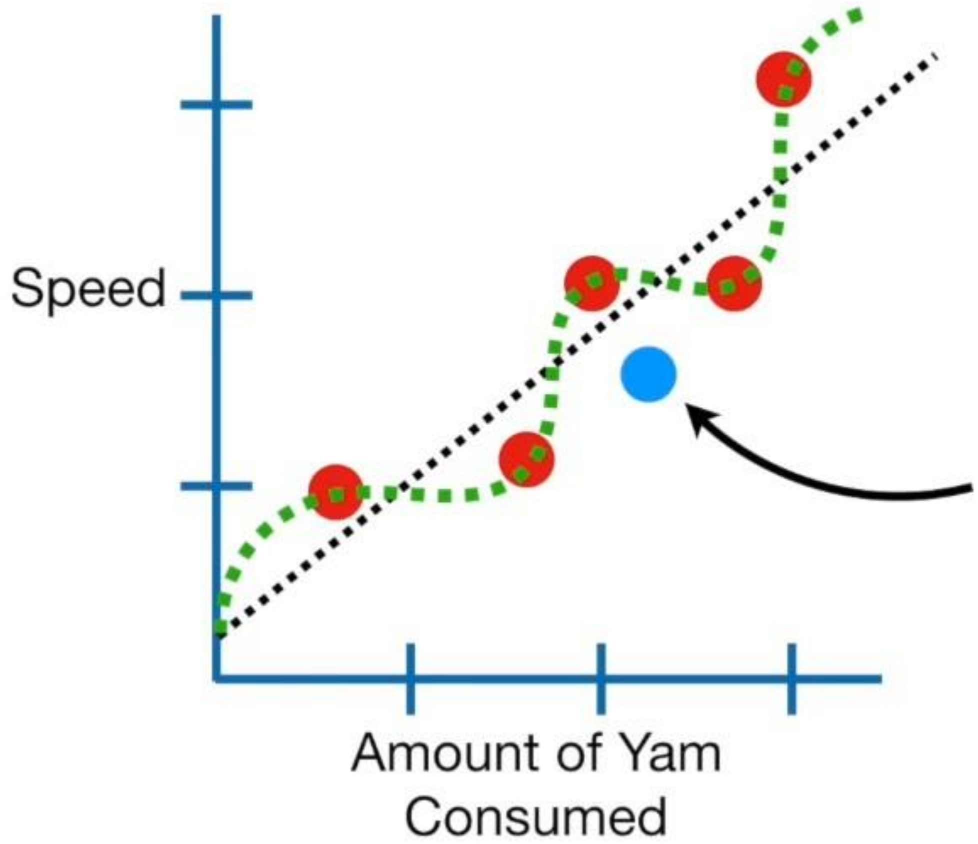


The **green squiggle** fits the **Training Data** better than the **black line**, but remember, the goal of **Machine Learning** is to make ***predictions***…

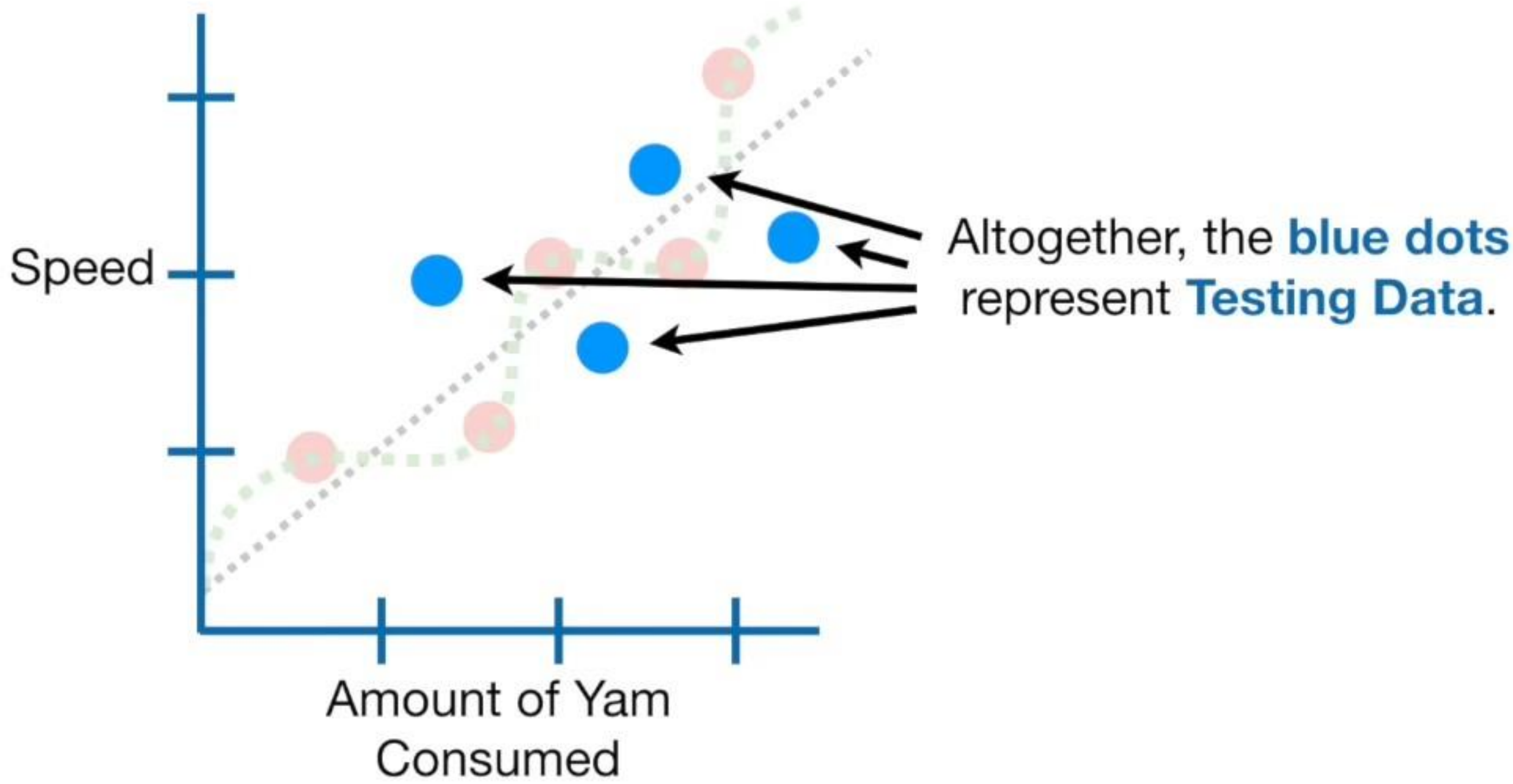


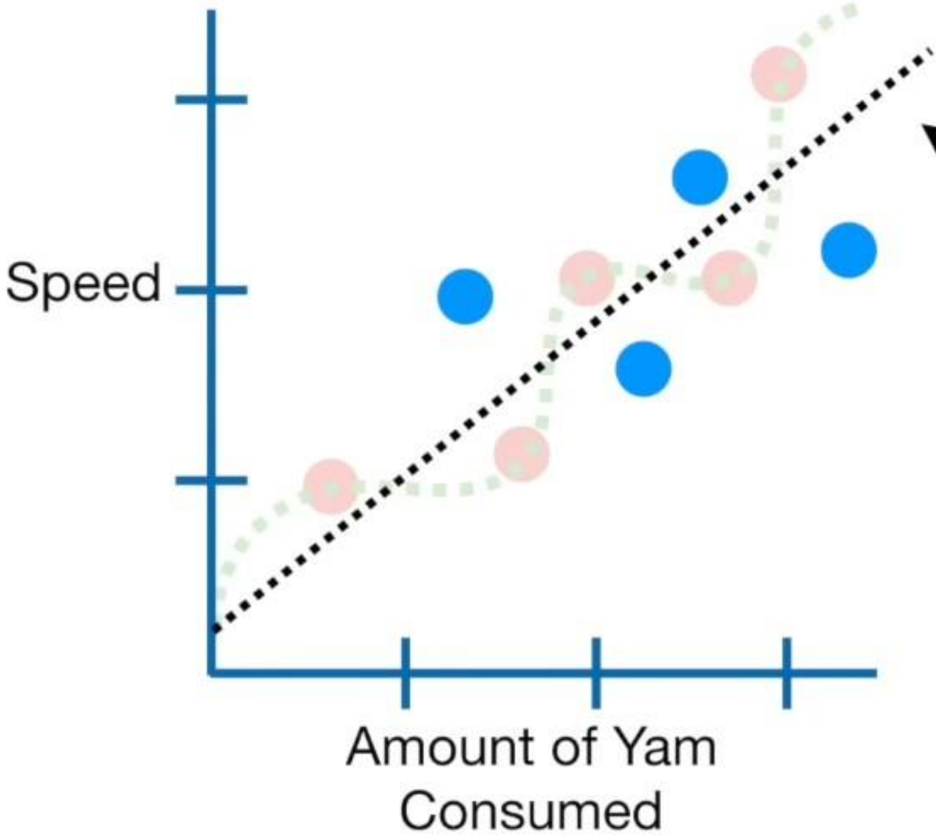
The **green squiggle** fits the **Training Data** better than the **black line**, but remember, the goal of **Machine Learning** is to make ***predictions***…

…so we need a way to decide if the **green squiggle** is better or worse than the **black line** at making ***predictions***.

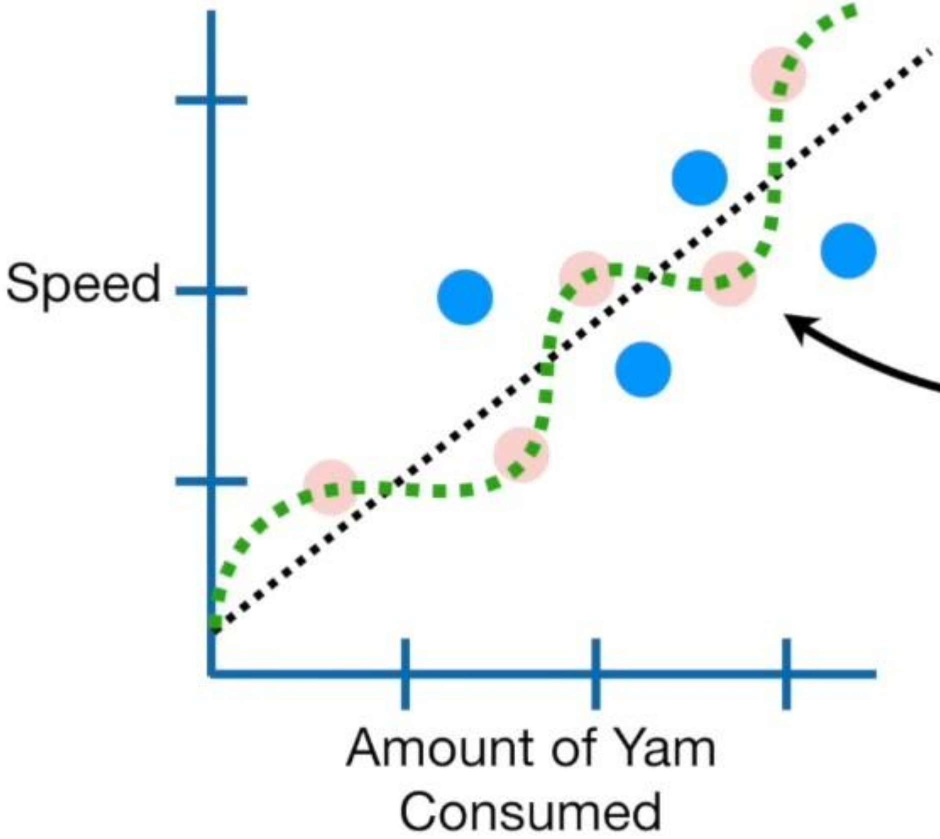


So we find a new person, and measure how fast they run and how much yam they eat…


Speed
Amount of Yam Consumed
Altogether, the **blue dots** represent **Testing Data**.

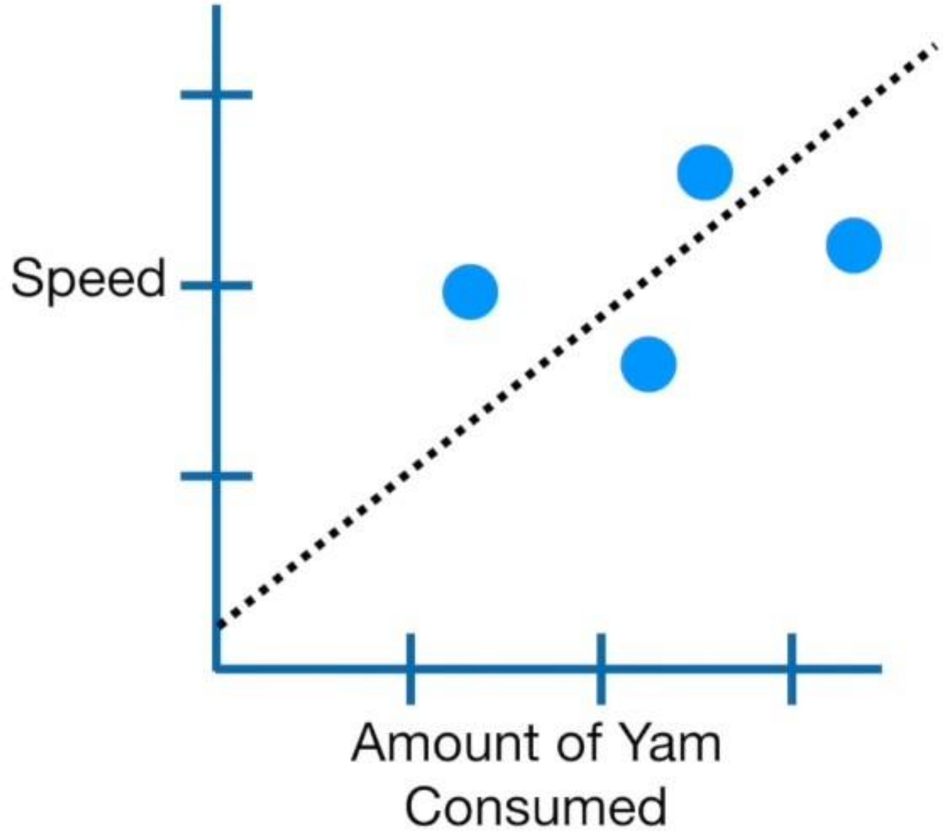


We use the **Testing Data** to compare the ***predictions*** made by the **black line**…

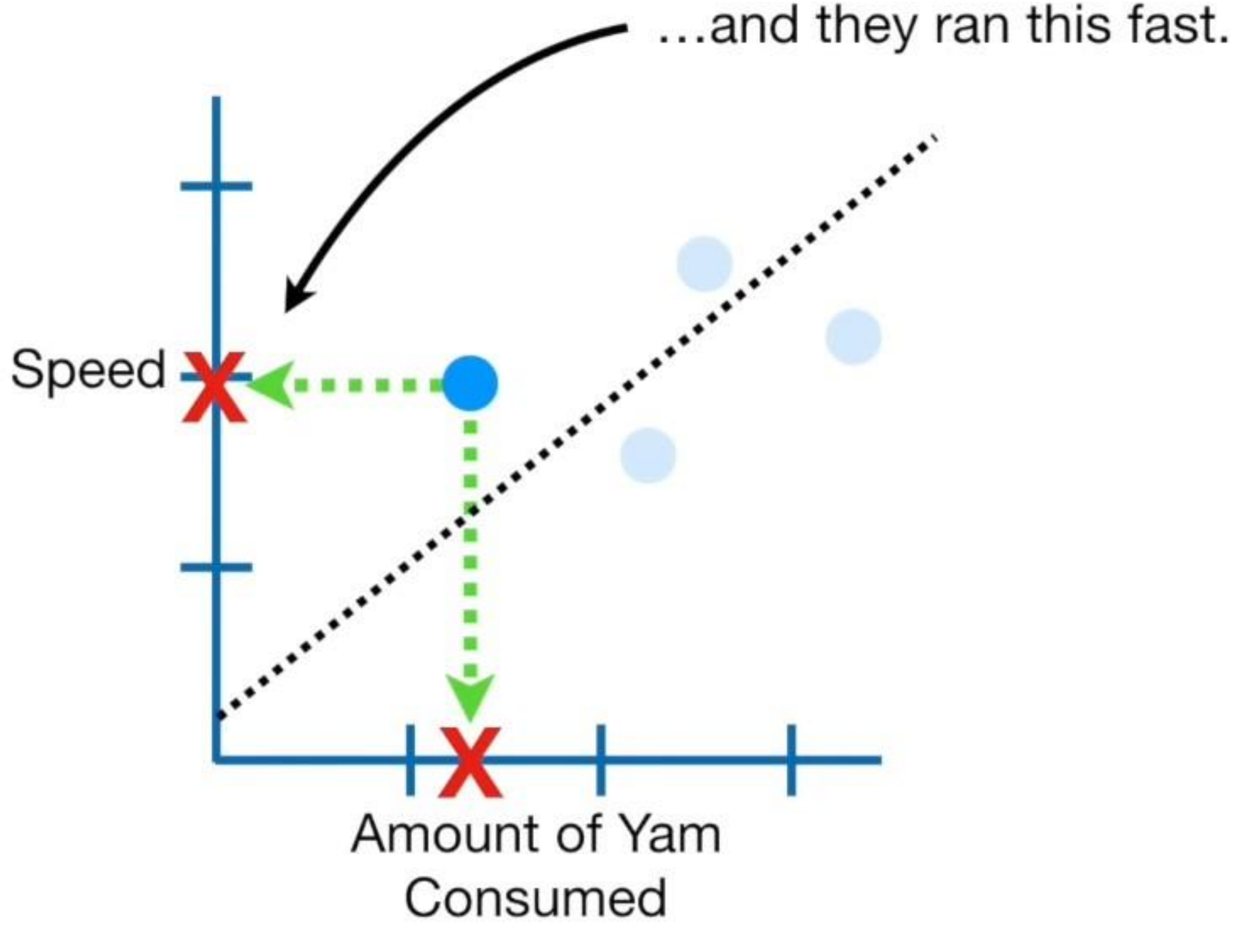


We use the **Testing Data** to compare the ***predictions*** made by the **black line**…

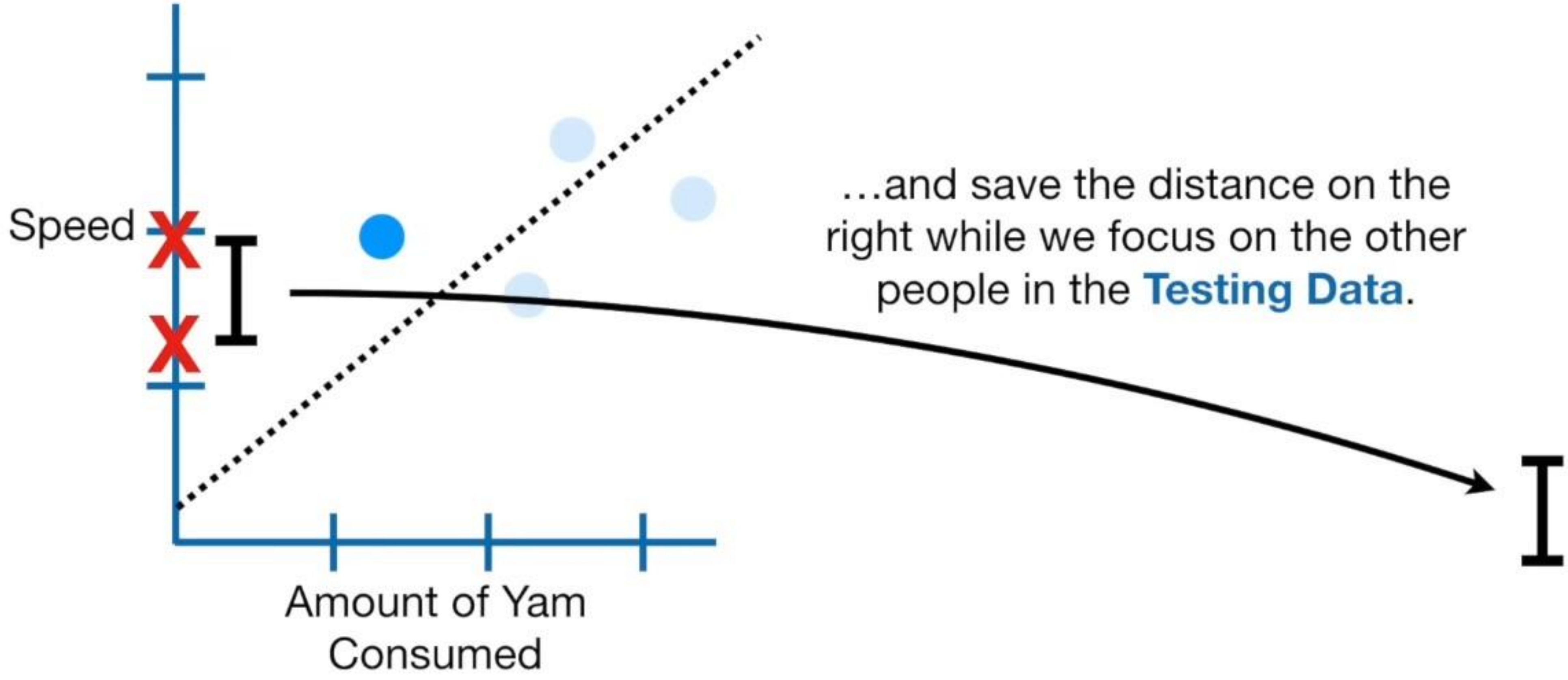
…to the ***predictions*** made by the **green squiggle**.



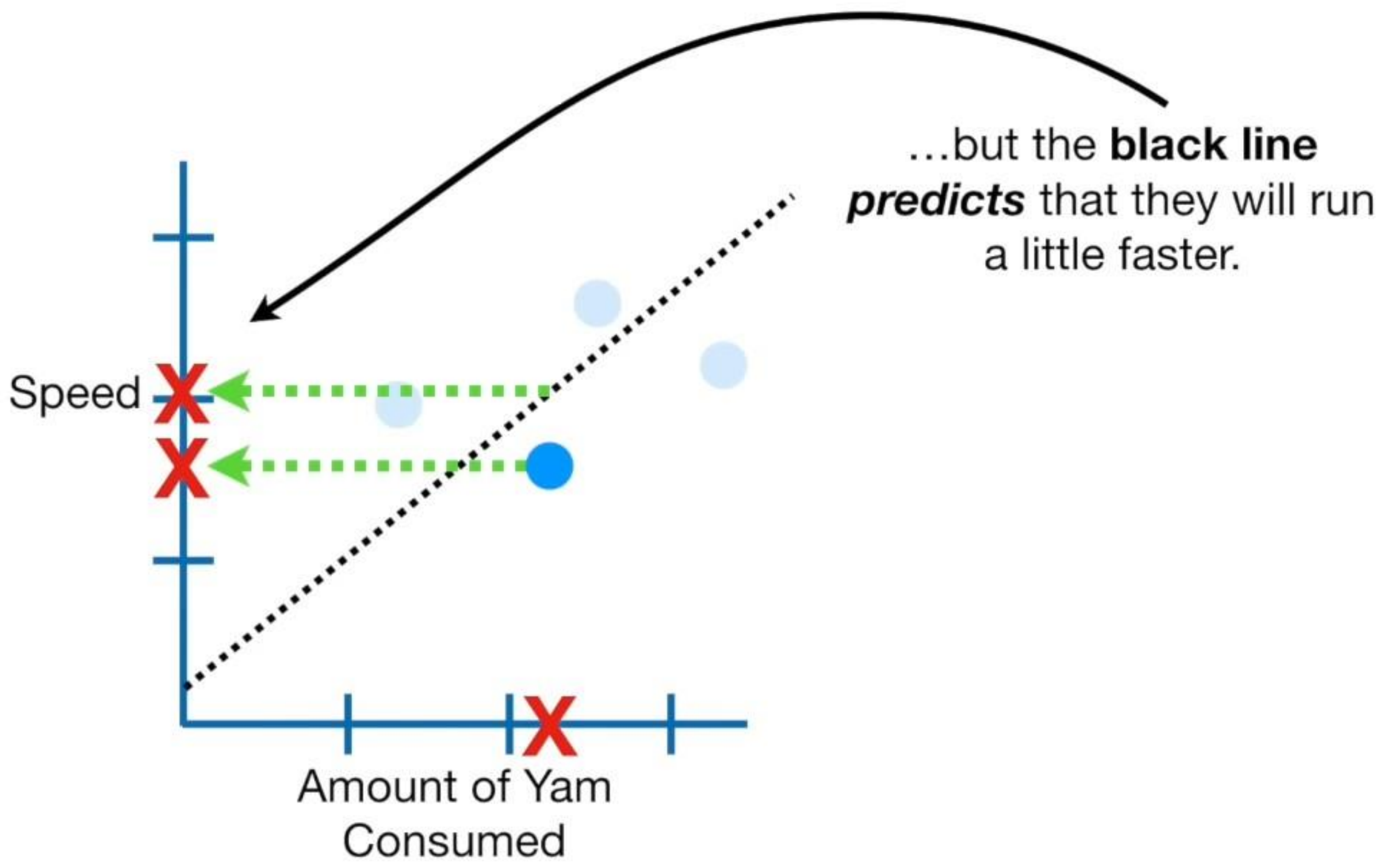
Let's start by seeing how well the **black line** predicts the speed of each person in the **Testing Data**.

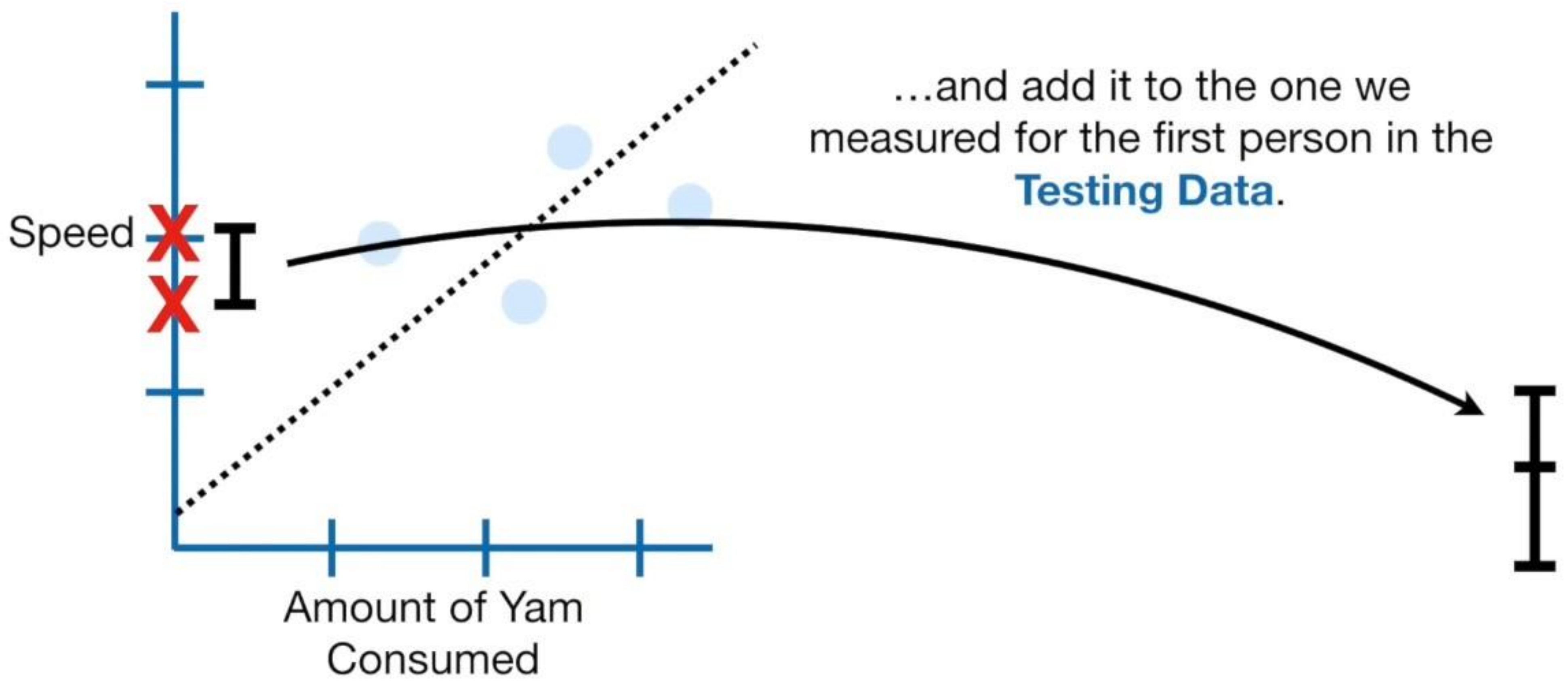


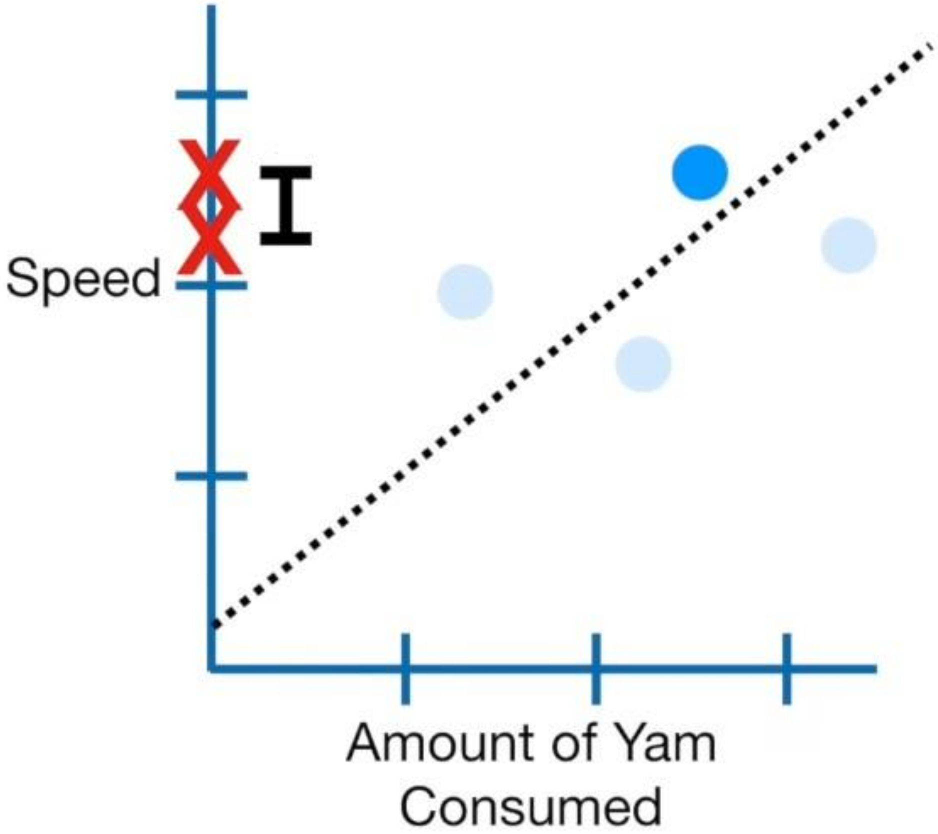
...and they ran this fast.
Speed
Amount of Yam
Consumed



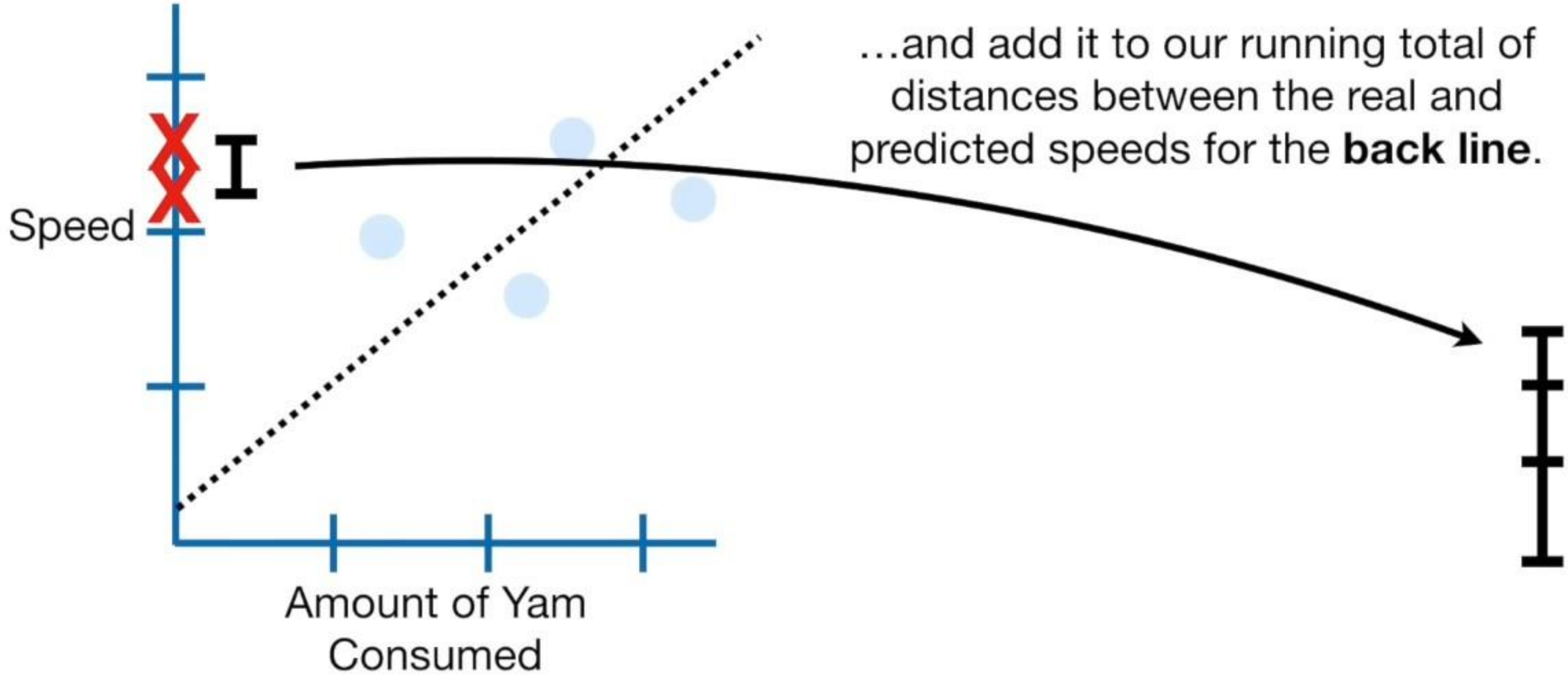
...and save the distance on the right while we focus on the other people in the **Testing Data**.


...but the **black line** ***predicts*** that they will run a little faster.
Speed
Amount of Yam Consumed

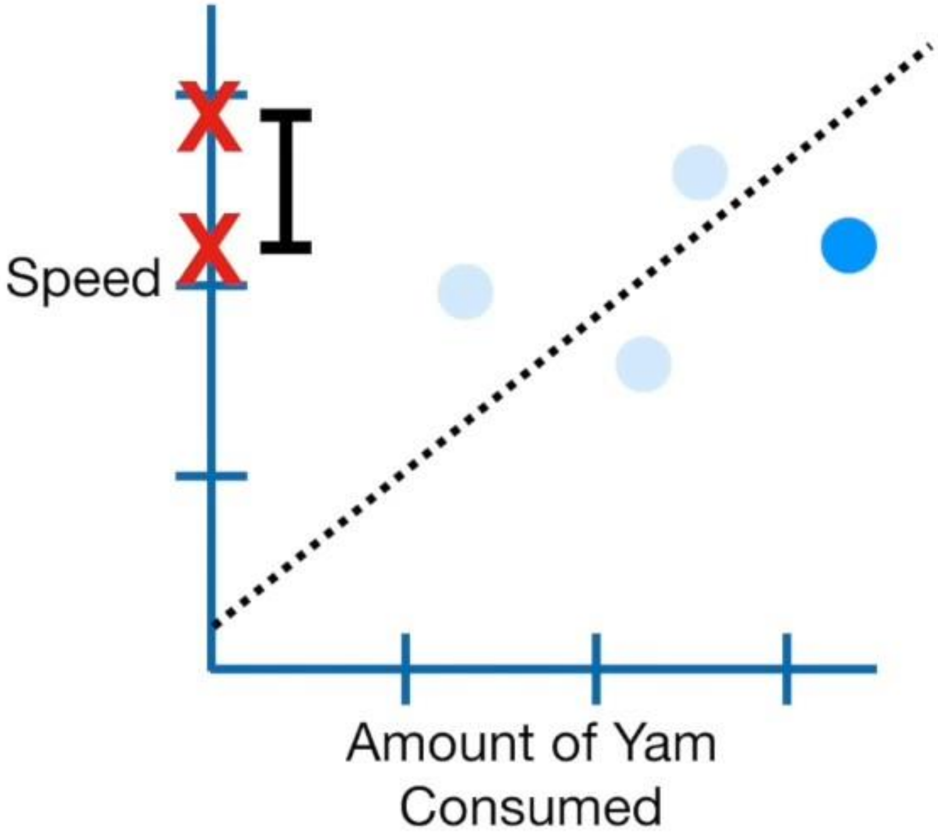

...and add it to the one we measured for the first person in the **Testing Data**.
Speed
Amount of Yam Consumed



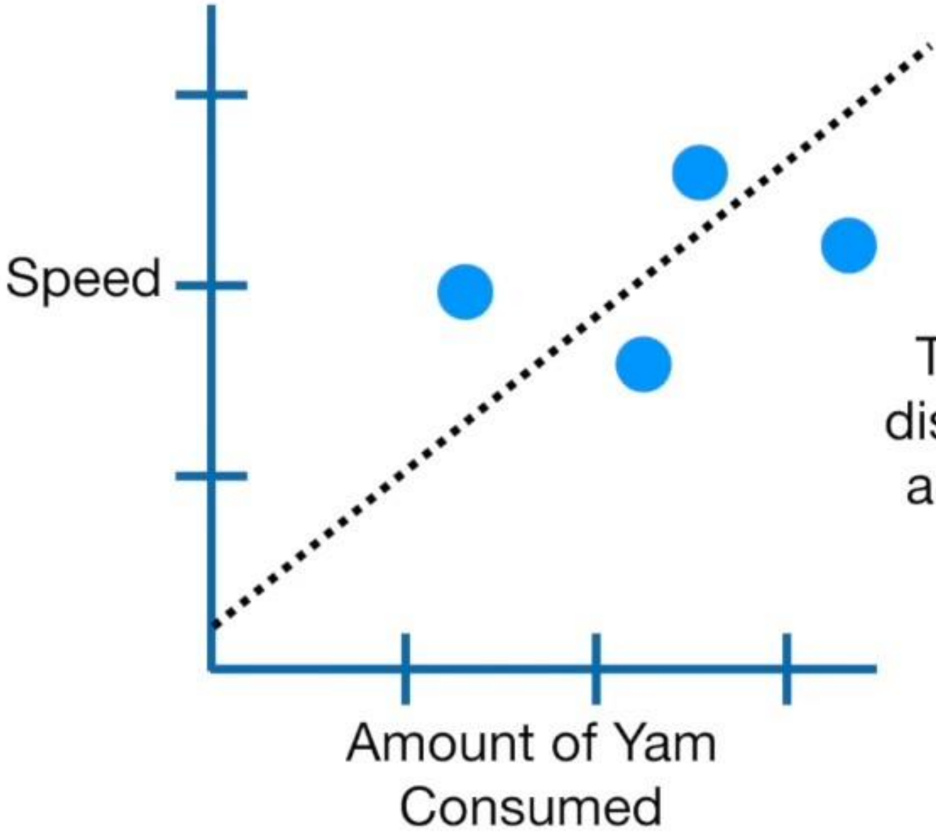
Then we measure the distance between the real and the predicted speed for the third person in the **Testing Data**…



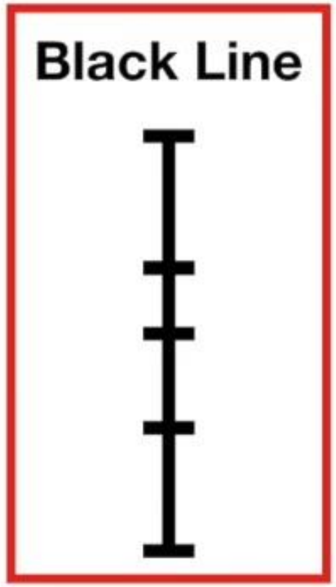
...and add it to our running total of distances between the real and predicted speeds for the **back line**.
Speed
Amount of Yam Consumed



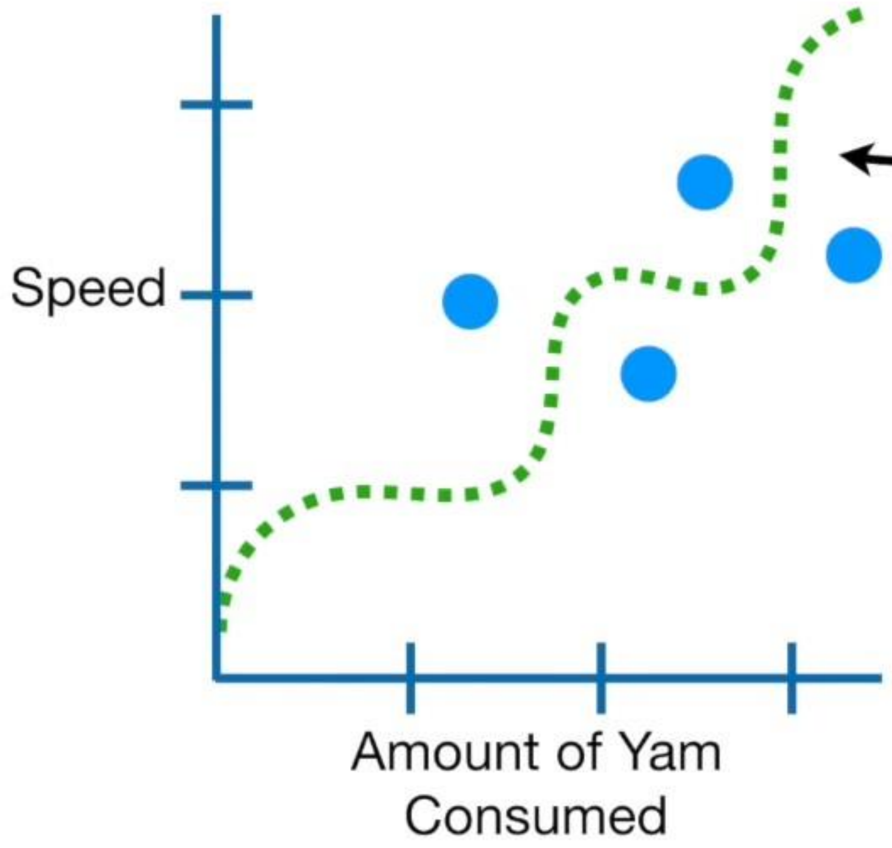
Then we do the same thing for the fourth person in the **Testing Data**.

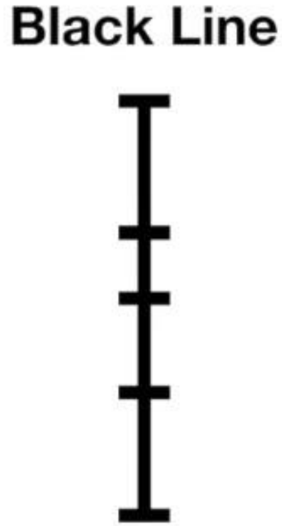


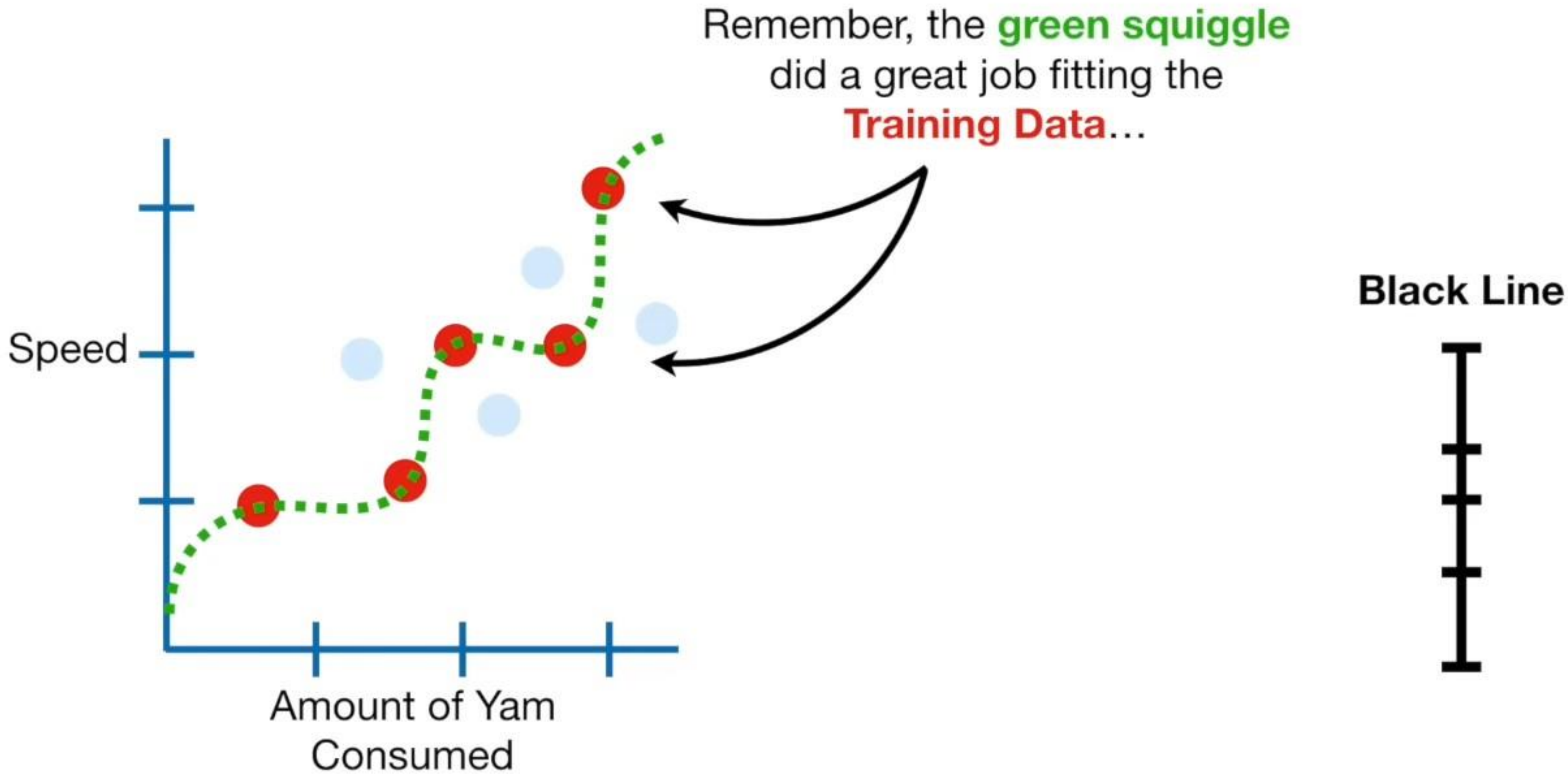
This is the sum of all the distances between the real and predicted speeds for the **black line**.



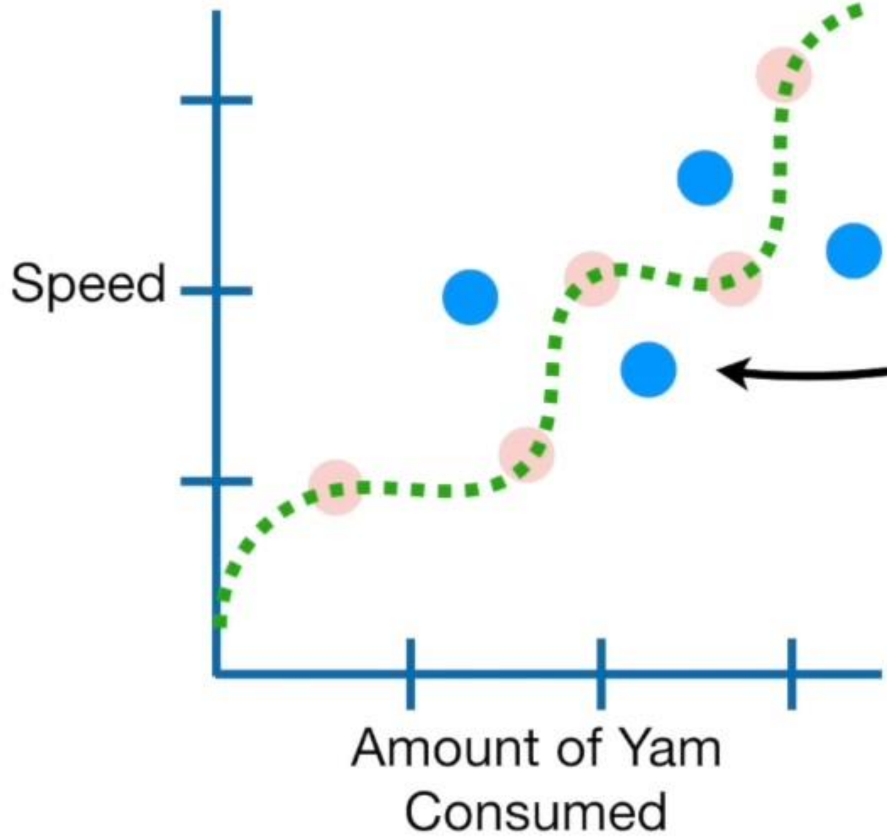
Now let's calculate the distances between the real and predicted speeds using the **green squiggle**.

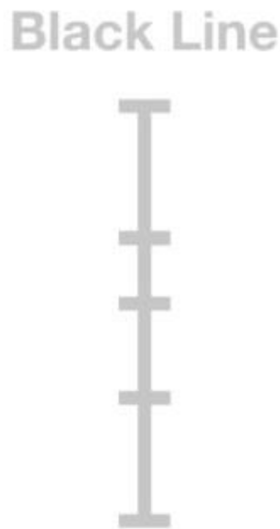


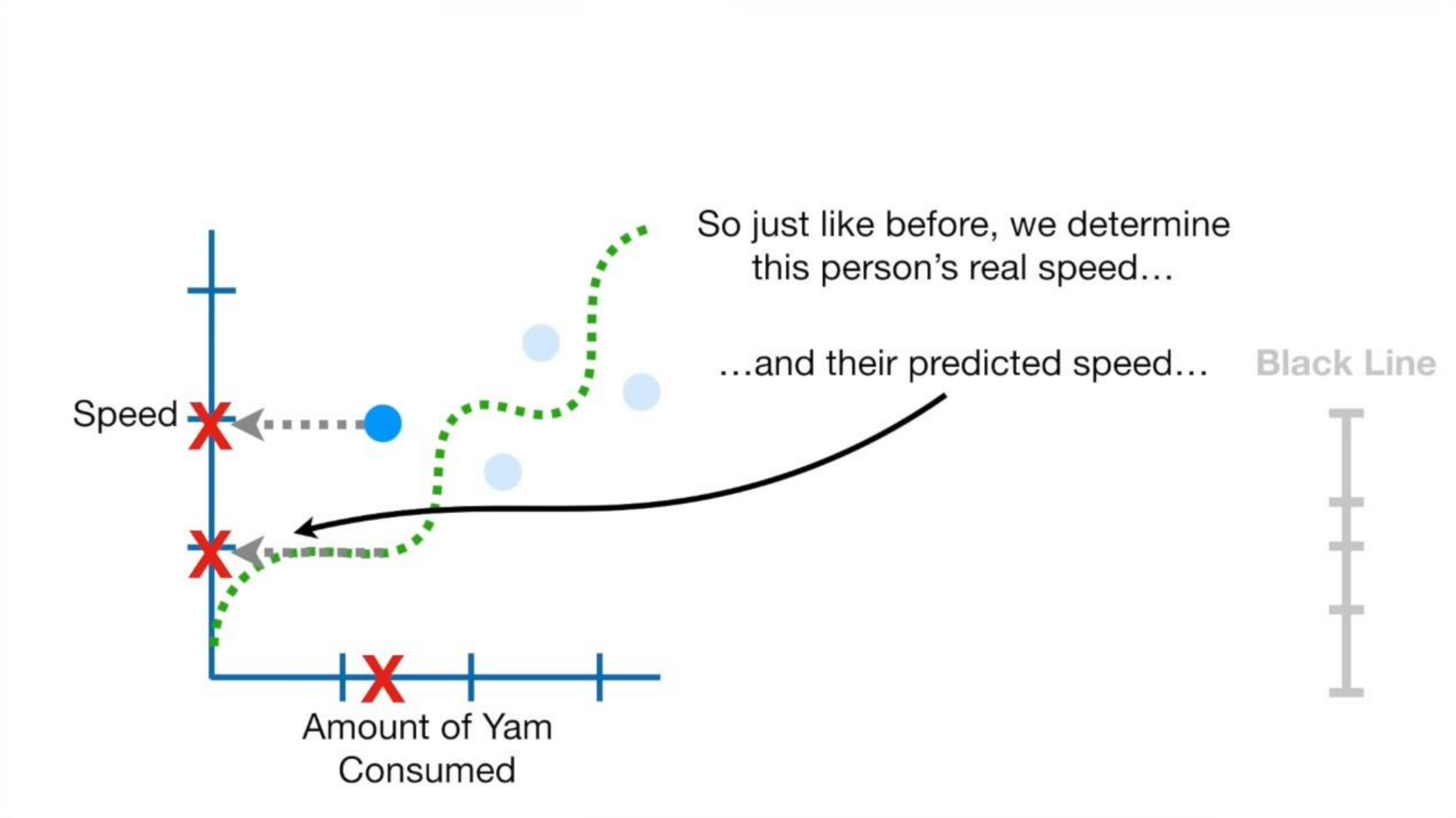


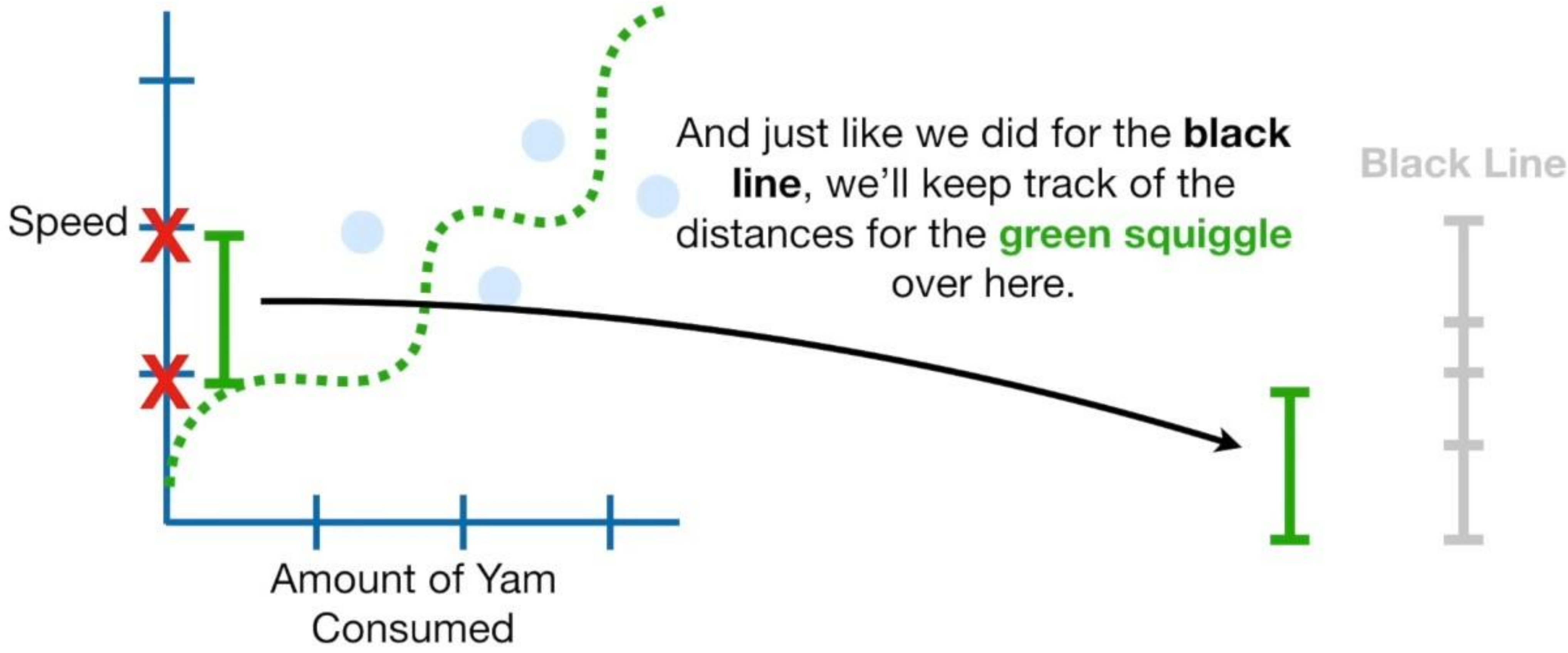

Remember, the green squiggle did a great job fitting the Training Data…
Speed
Amount of Yam Consumed
Black Line

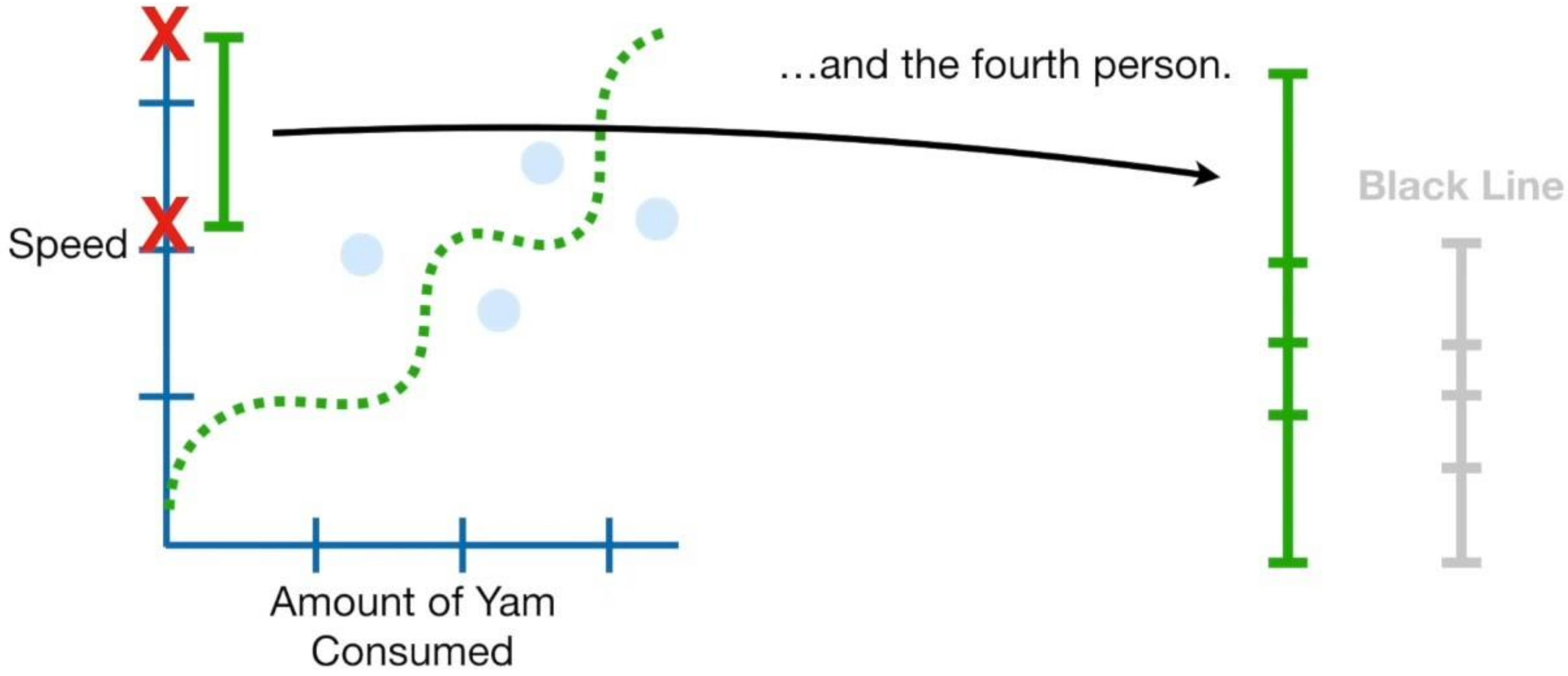
...but when we are doing **Machine Learning**, we're more interested in how well the **green squiggle** can make ***predictions*** with new data.



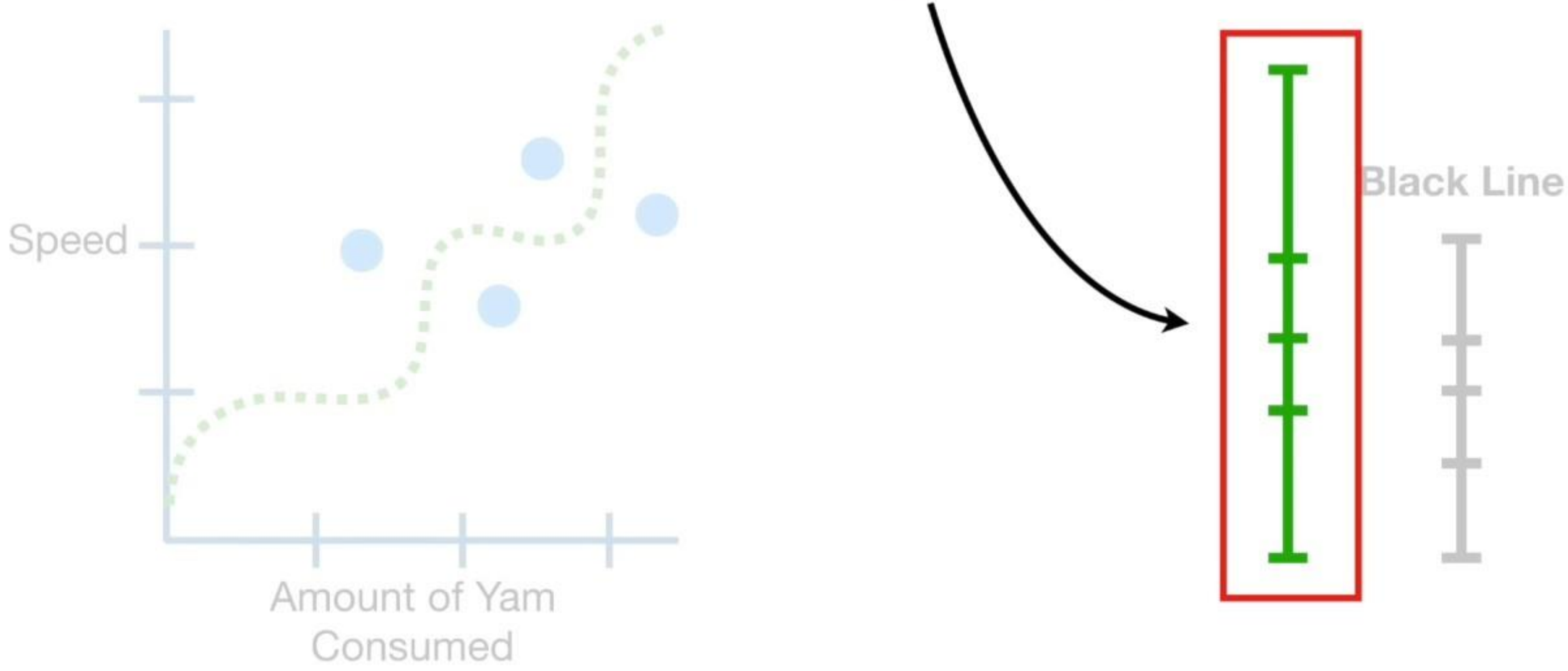



So just like before, we determine
this person's real speed…
…and their predicted speed…
Black Line
Speed
Amount of Yam
Consumed

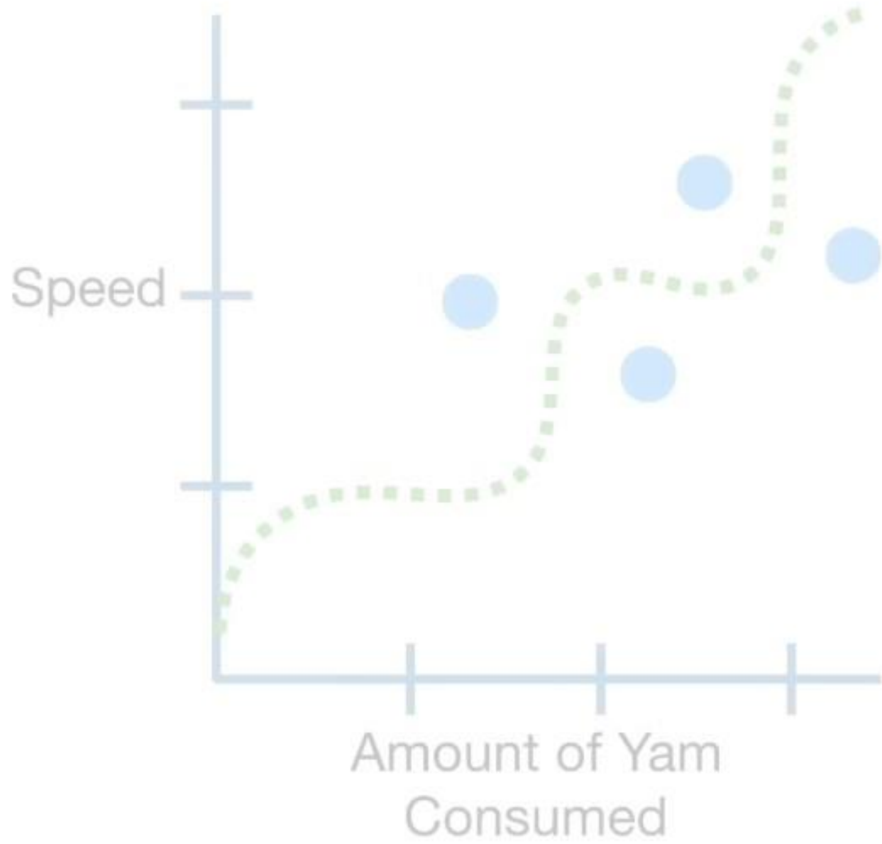

Speed
Amount of Yam Consumed
And just like we did for the **black line**, we'll keep track of the distances for the **green squiggle** over here.
Black Line

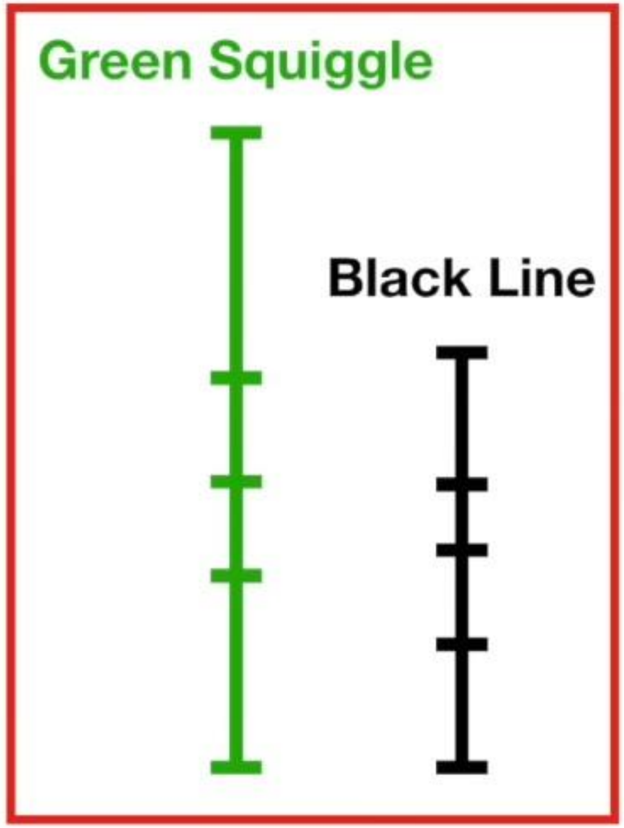

...and the fourth person.
Black Line
Speed
Amount of Yam Consumed

This is the sum of the distances between the real and predicted speeds for the **green squiggle**.

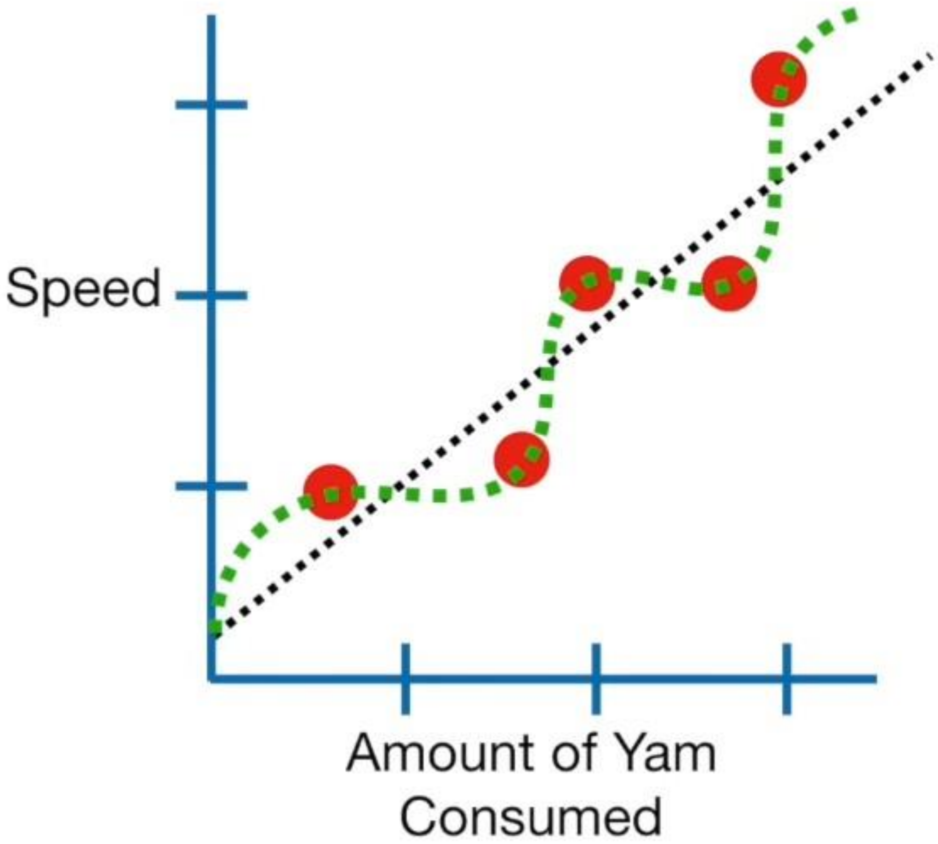


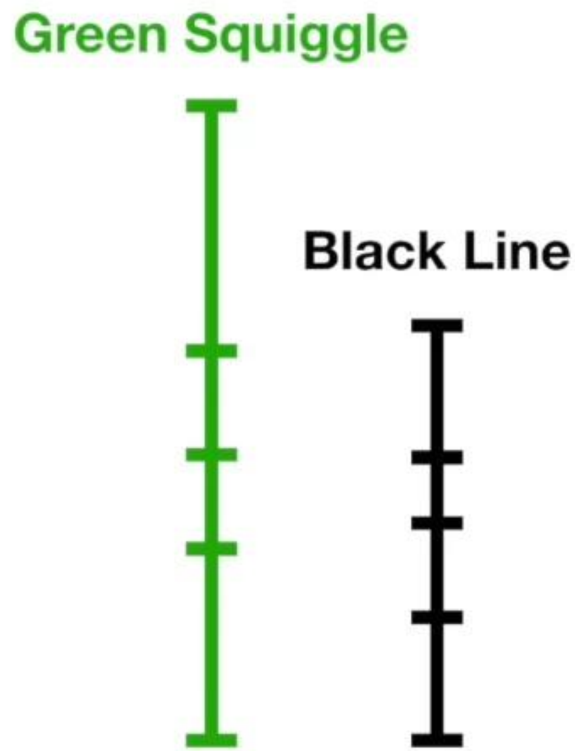
The sum of the distances is larger for the **green squiggle** then the **black line**…



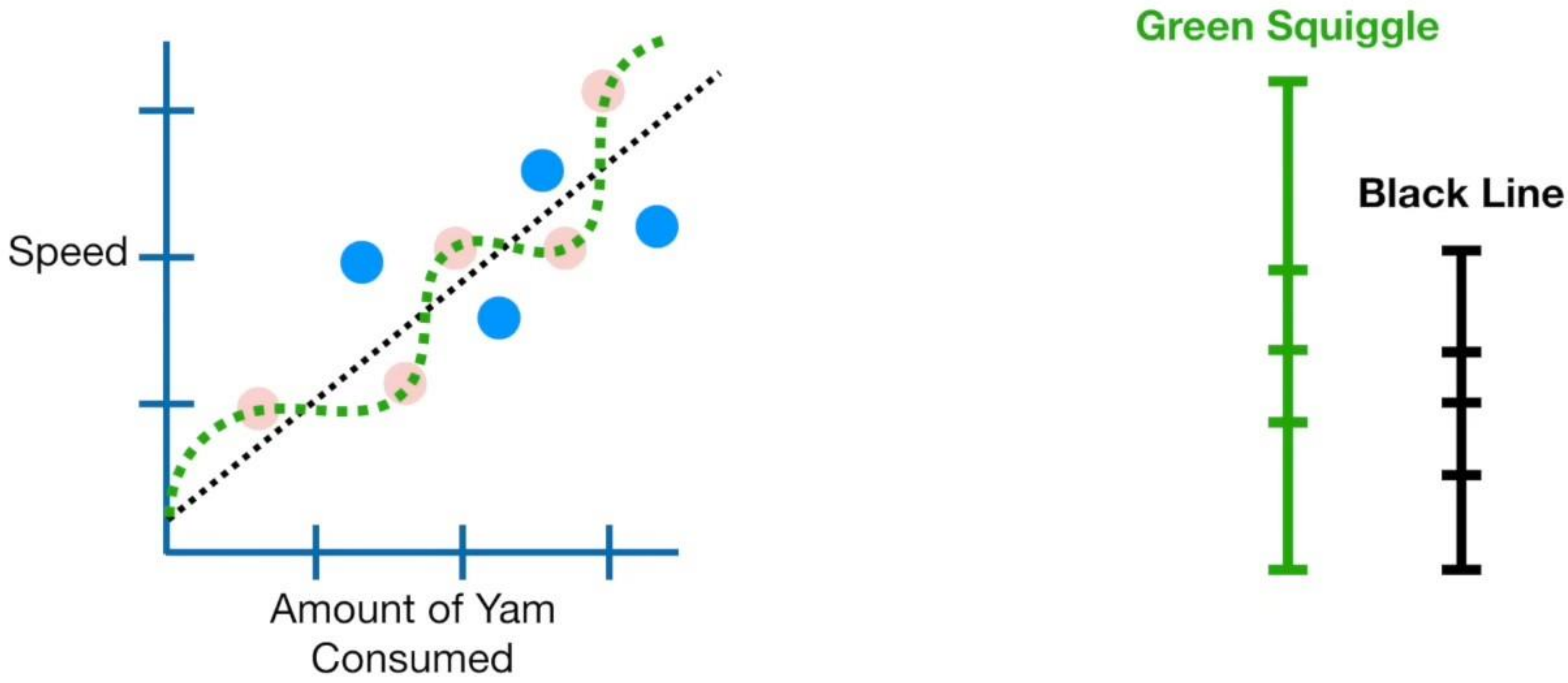


In other words, even though the
green squiggle fit the **Training Data**
way better than the **black line**…

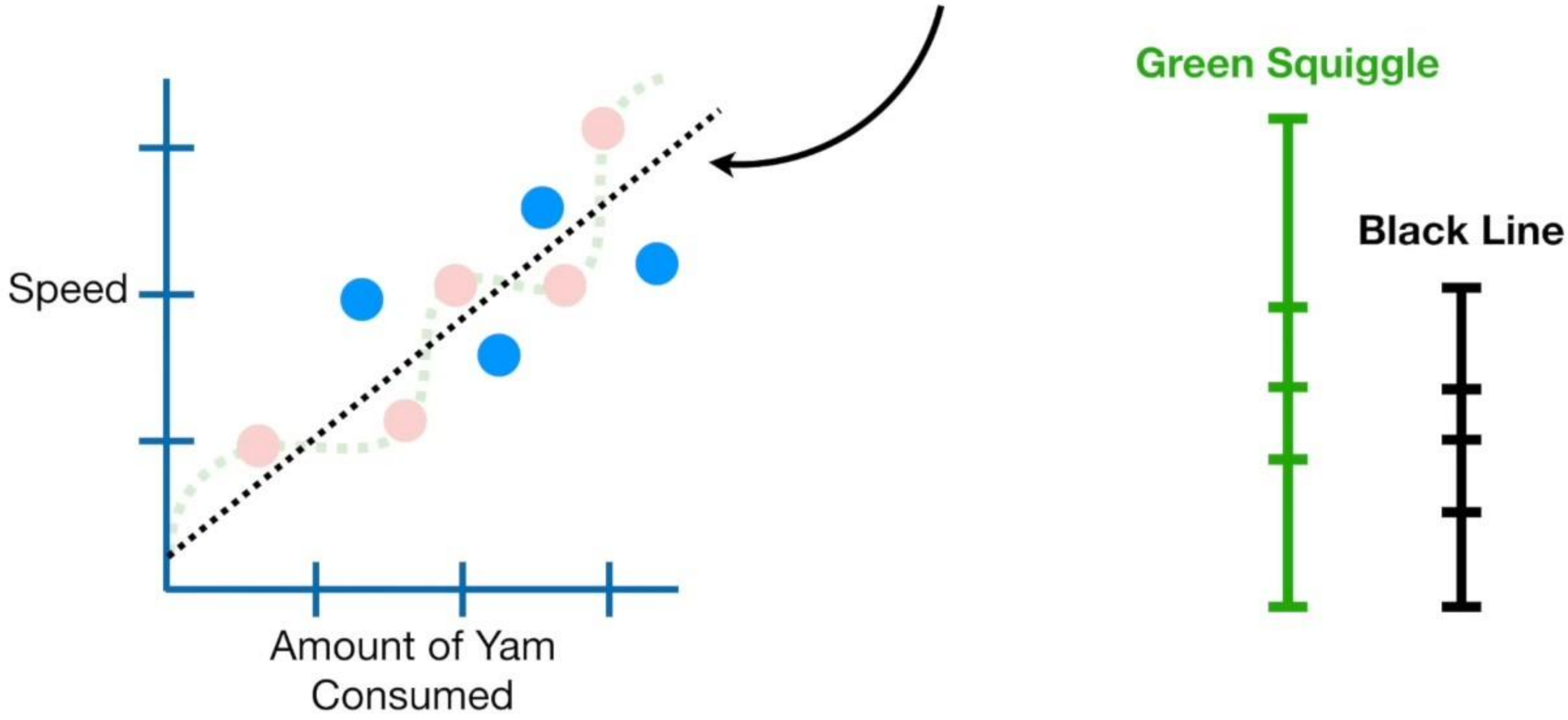




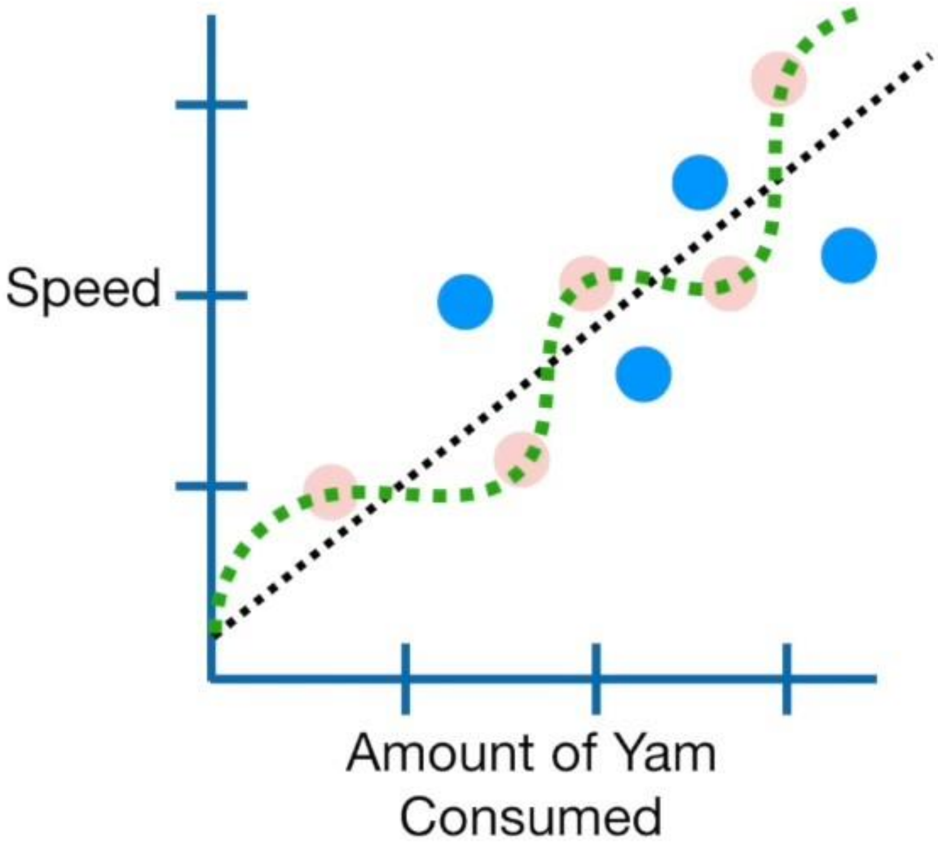
...the **black line** did a better job predicting speeds with the **Testing Data**.

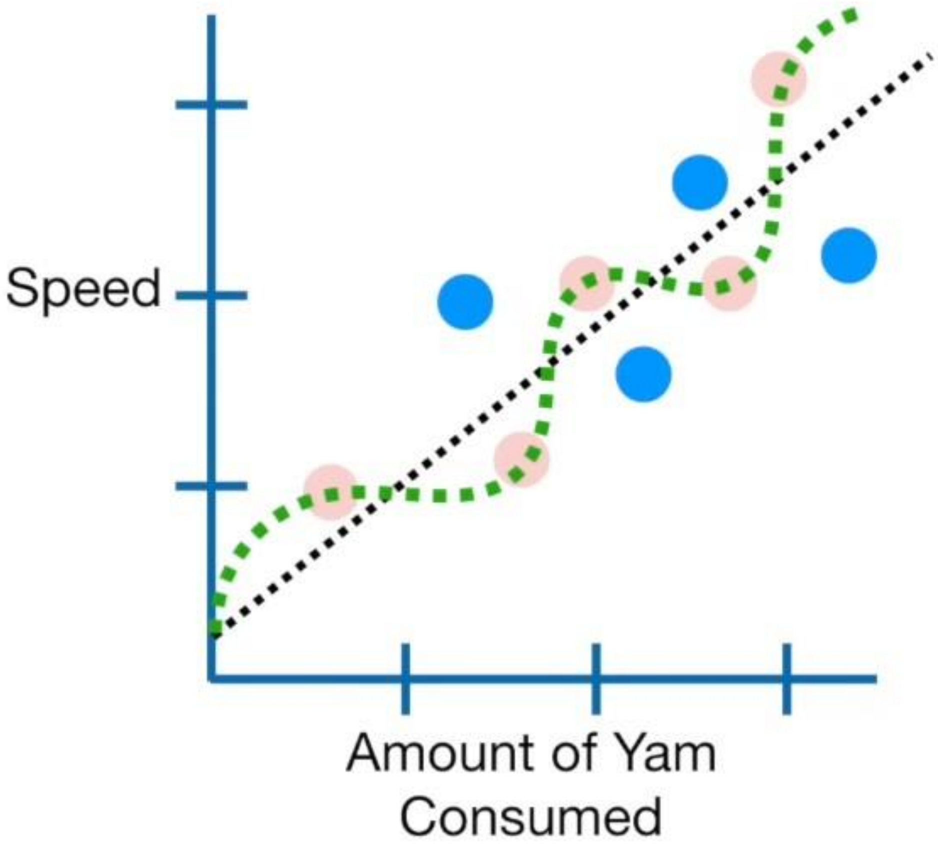


…we would choose the **back line**.



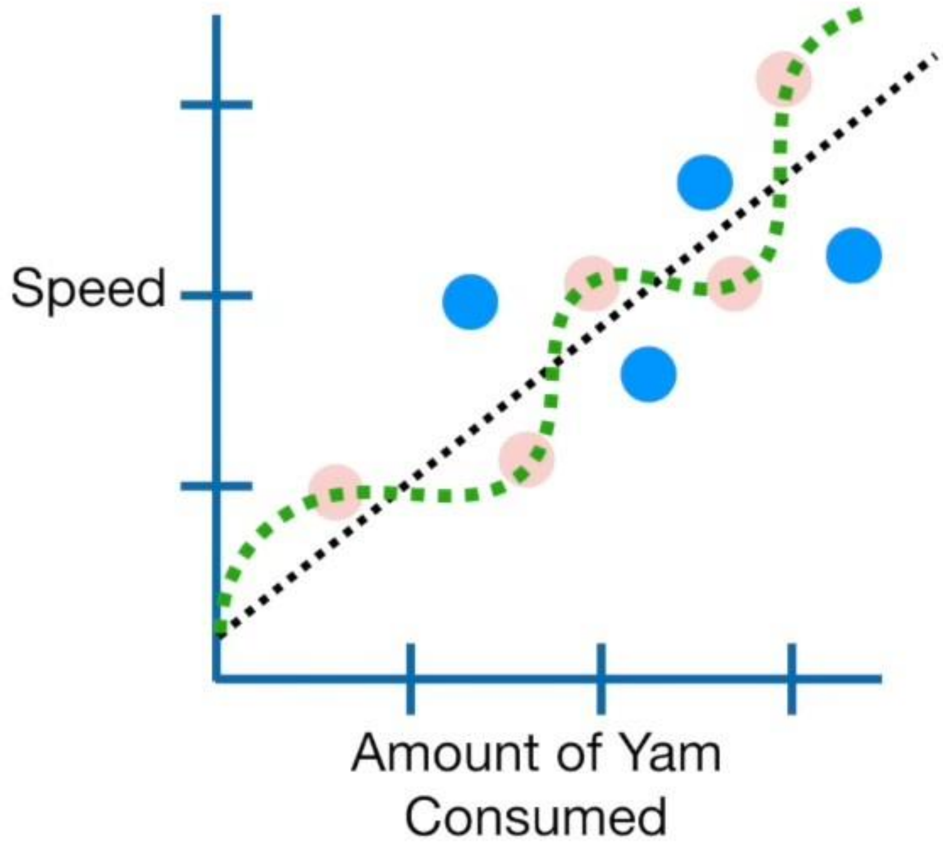
This example teaches two main ideas about **Machine Learning**.





This example teaches two main ideas about **Machine Learning**.

First, we use **Testing Data** to evaluate **Machine Learning** methods.

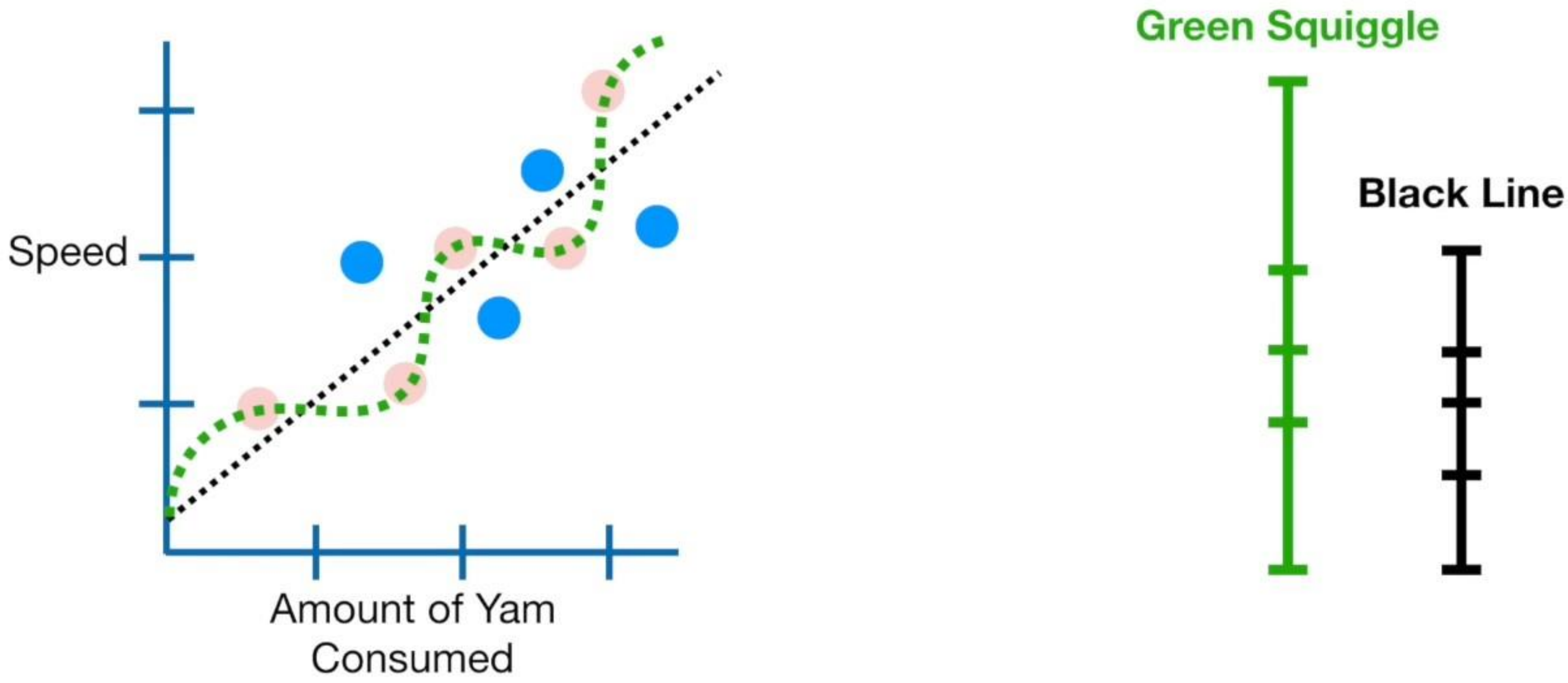


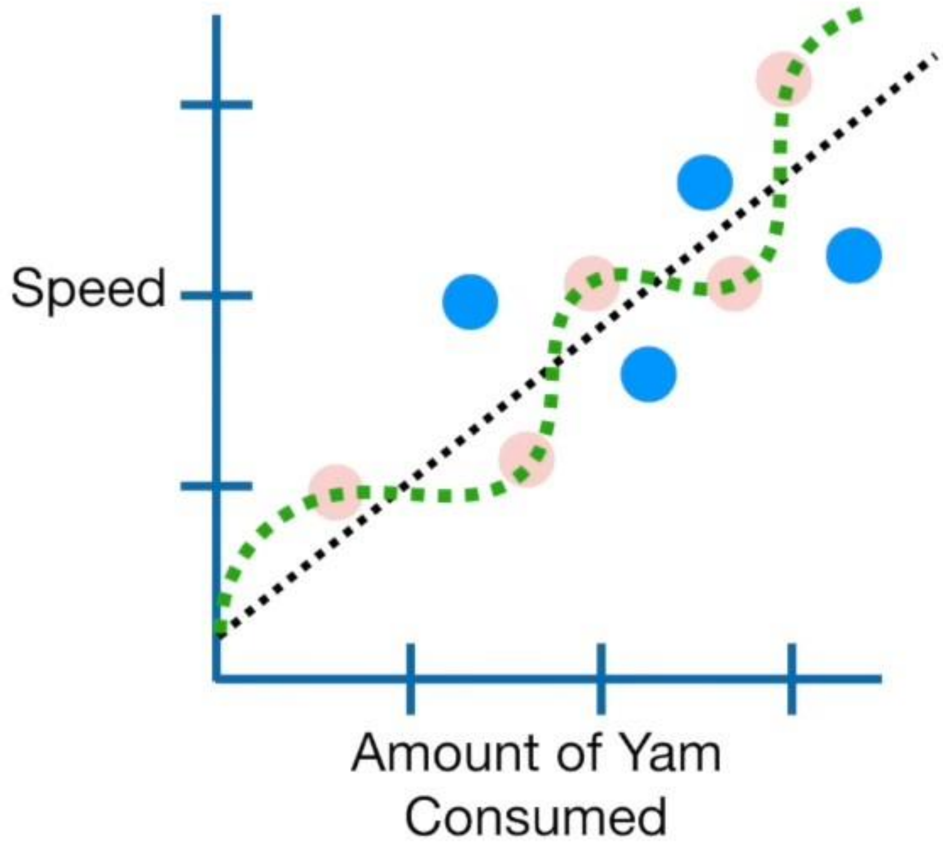
This example teaches two main ideas about **Machine Learning**.

First, we use **Testing Data** to evaluate **Machine Learning** methods.

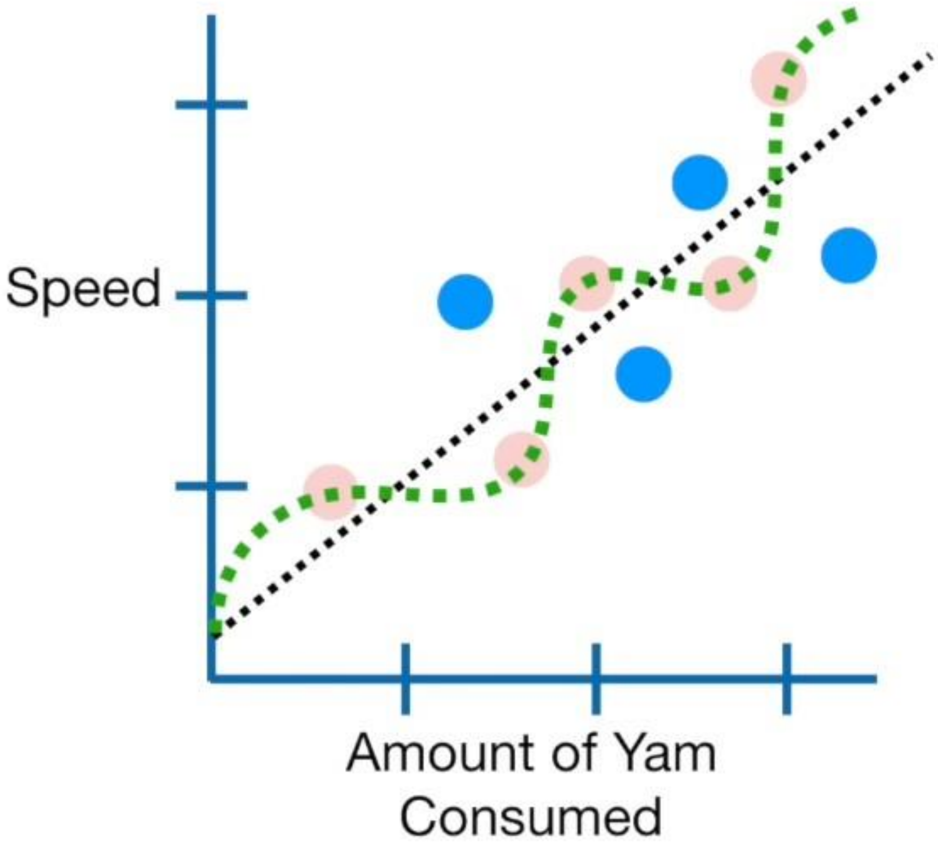
Second, don't be fooled by how well a **Machine Learning** method fits the **Training Data**.

NOTE: Fitting the **Training Data** well but making poor *predictions*, is called the **Bias-Variance Tradeoff**.





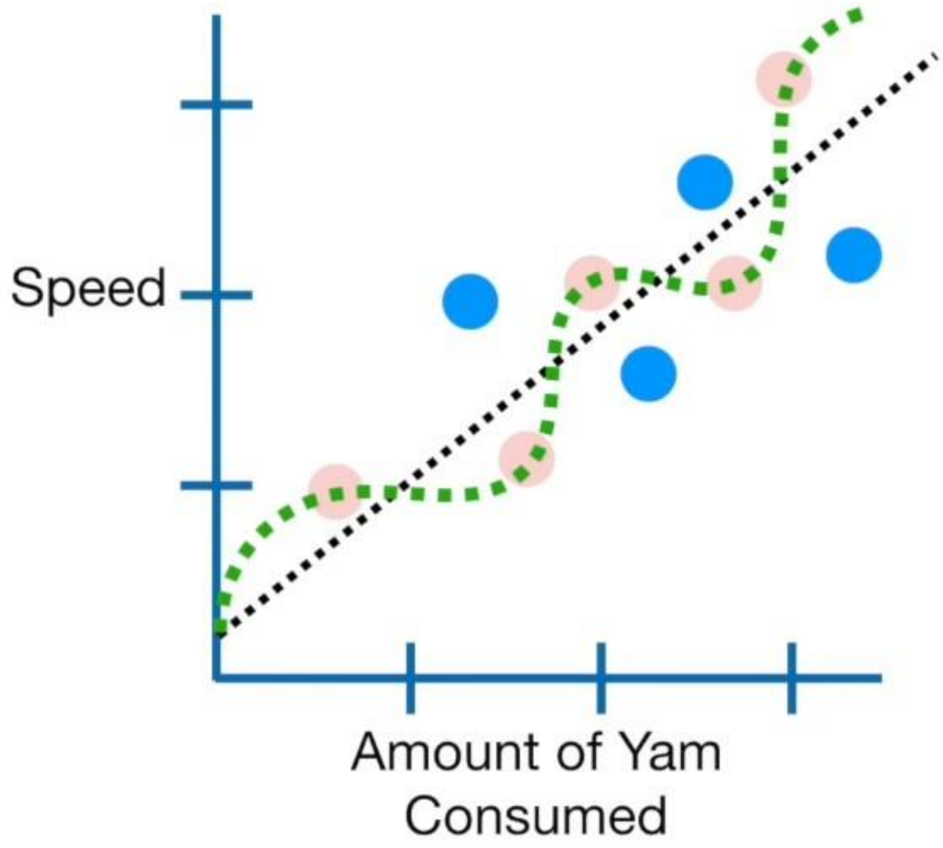
Before we move on, you may be wondering why we used a simple **black line** and a silly **green squiggle** instead of a…



Before we move on, you may be wondering why we used a simple **black line** and a silly **green squiggle** instead of a…

Deep Learning Convolutional Neural Network

or…

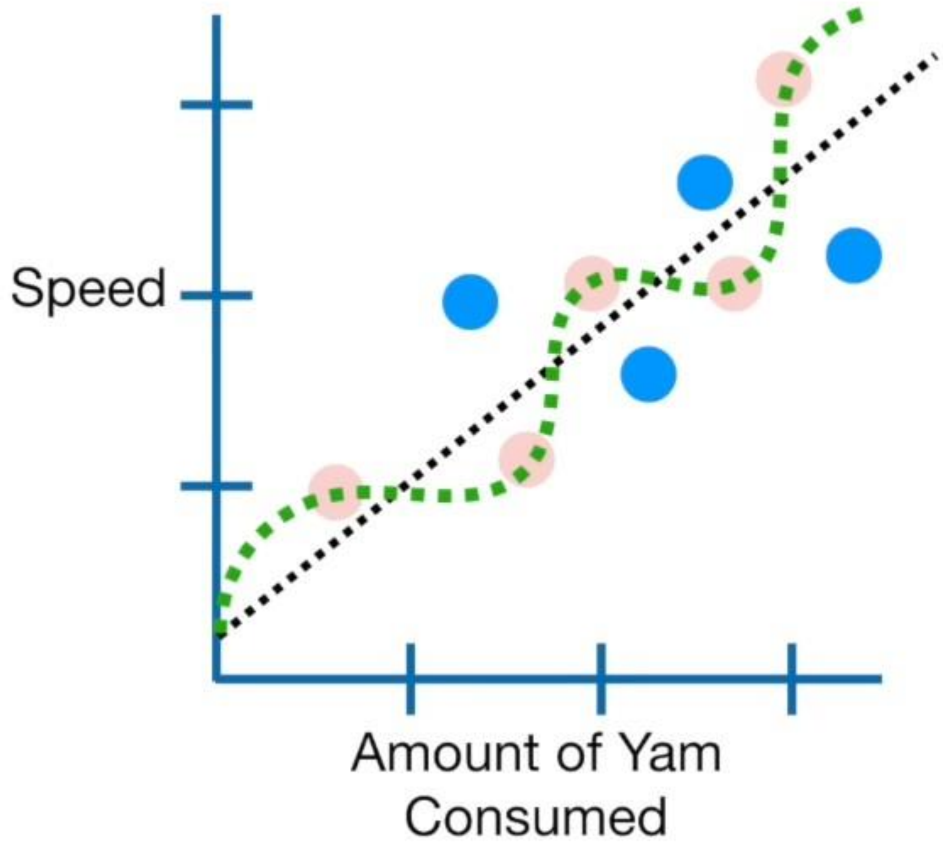


Before we move on, you may be wondering why we used a simple **black line** and a silly **green squiggle** instead of a…

Deep Learning Convolutional Neural Network

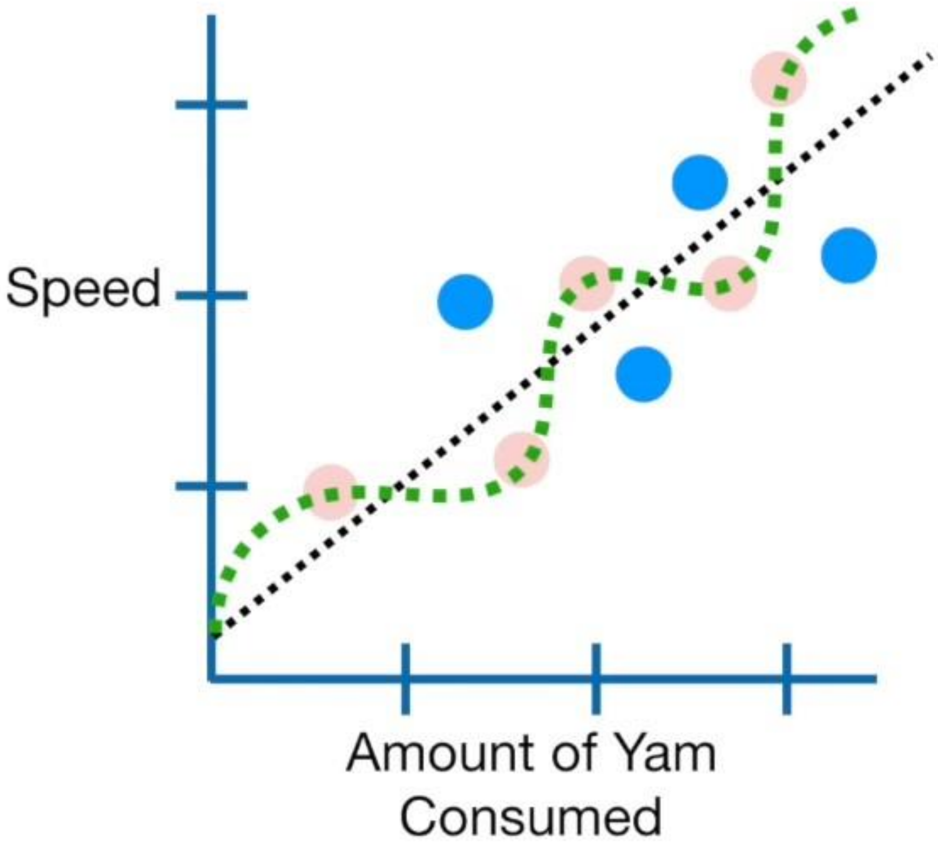
or…

[insert newest, bestest, most fancy machine learning method here].



There are tons of fancy sounding **Machine Learning** methods…

…and each year something new and exciting comes on the scene…

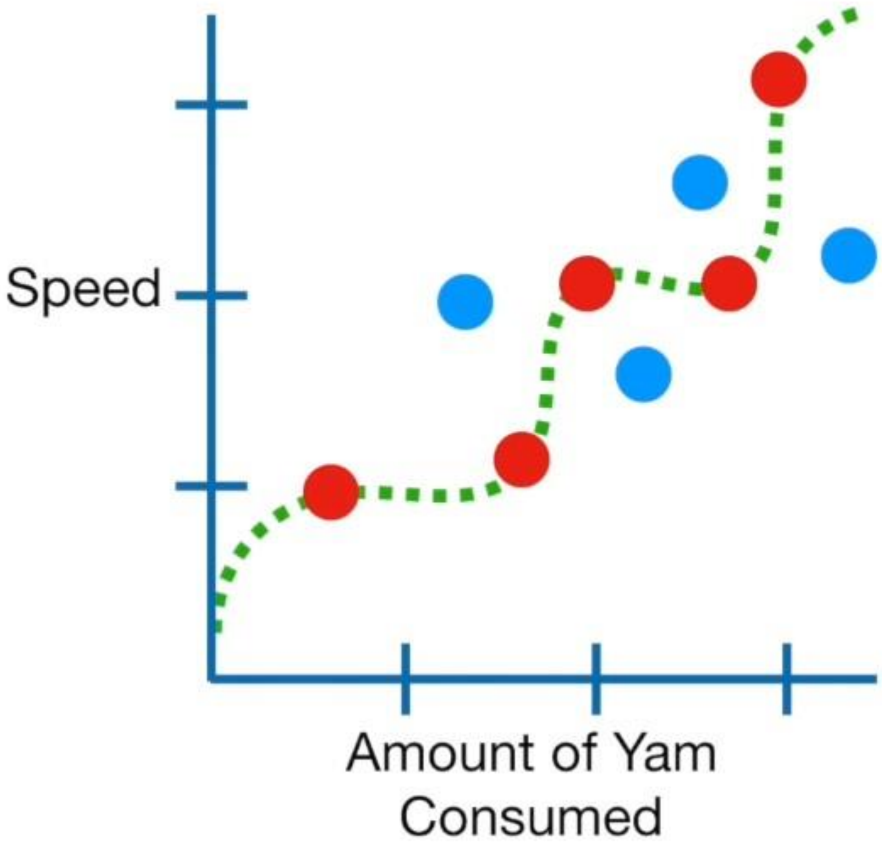


There are tons of fancy sounding **Machine Learning** methods…

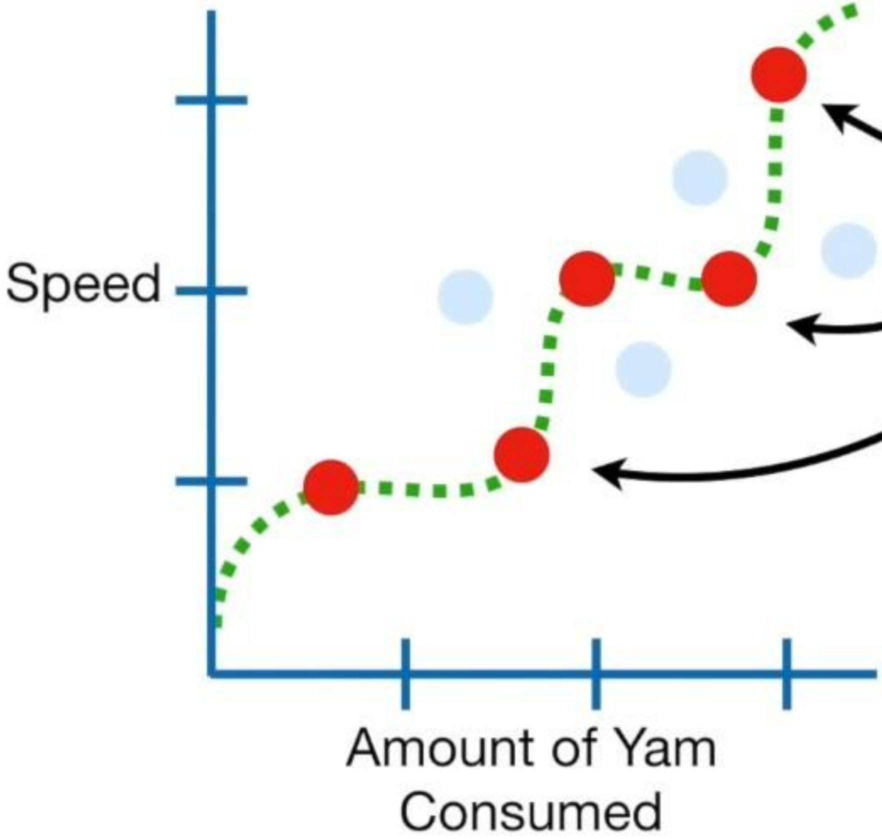
…and each year something new and exciting comes on the scene…

…but regardless of what you use, the most important thing isn't how fancy it is, but how it performs with **Testing Data**.

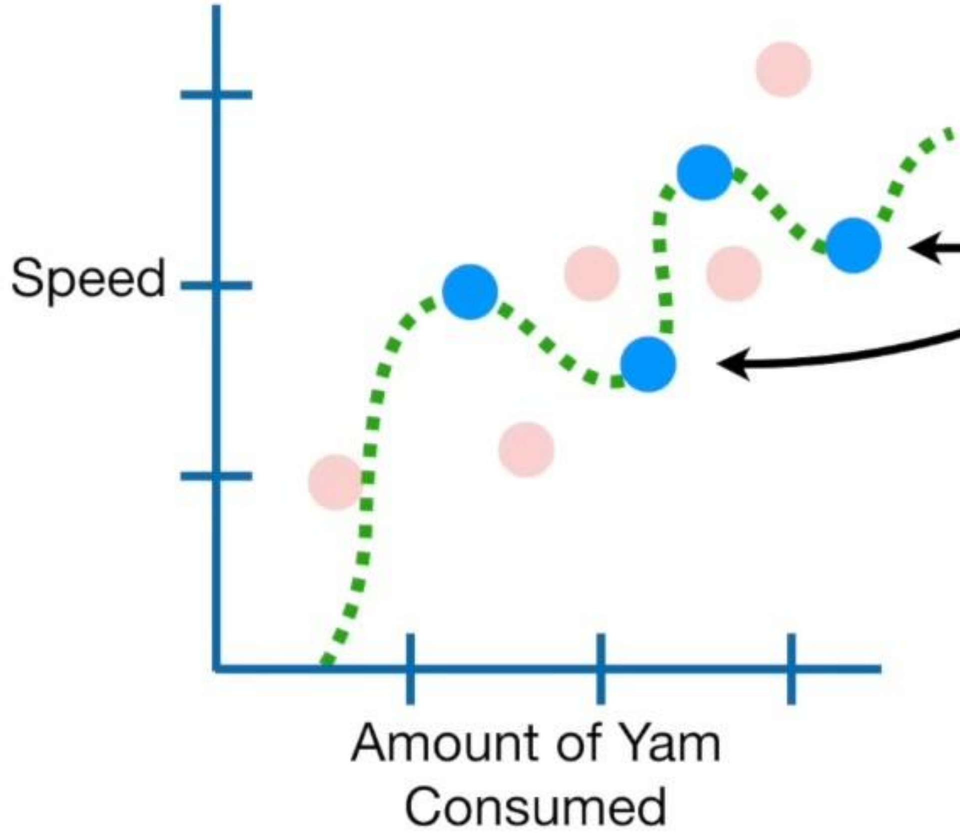
One last thing before we go.



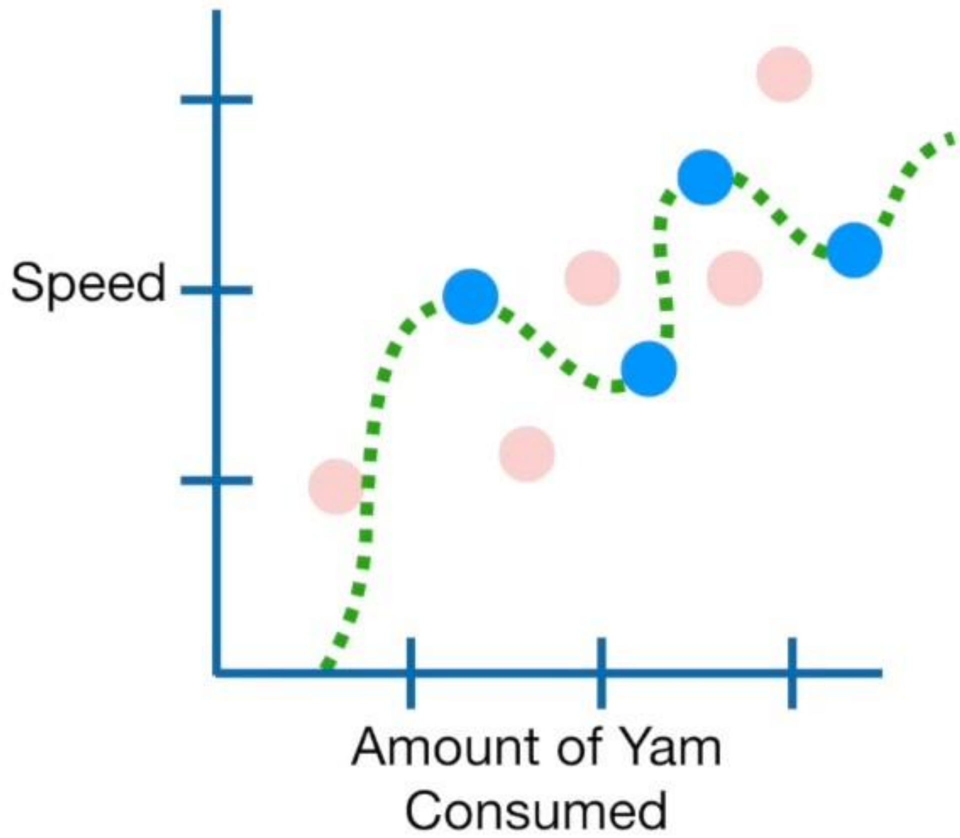
You may be wondering how we decide which data go into the **Training Set** and which data go into the **Testing Set**.



Earlier, we just arbitrarily decided that these **red dots** were the **Training Data**…



…but the **blue dots** could have, just as easily, been the **Training Data**.



The good news is that there are ways to determine which samples should be used for **Training Data** and which samples should be used for **Testing Data**…