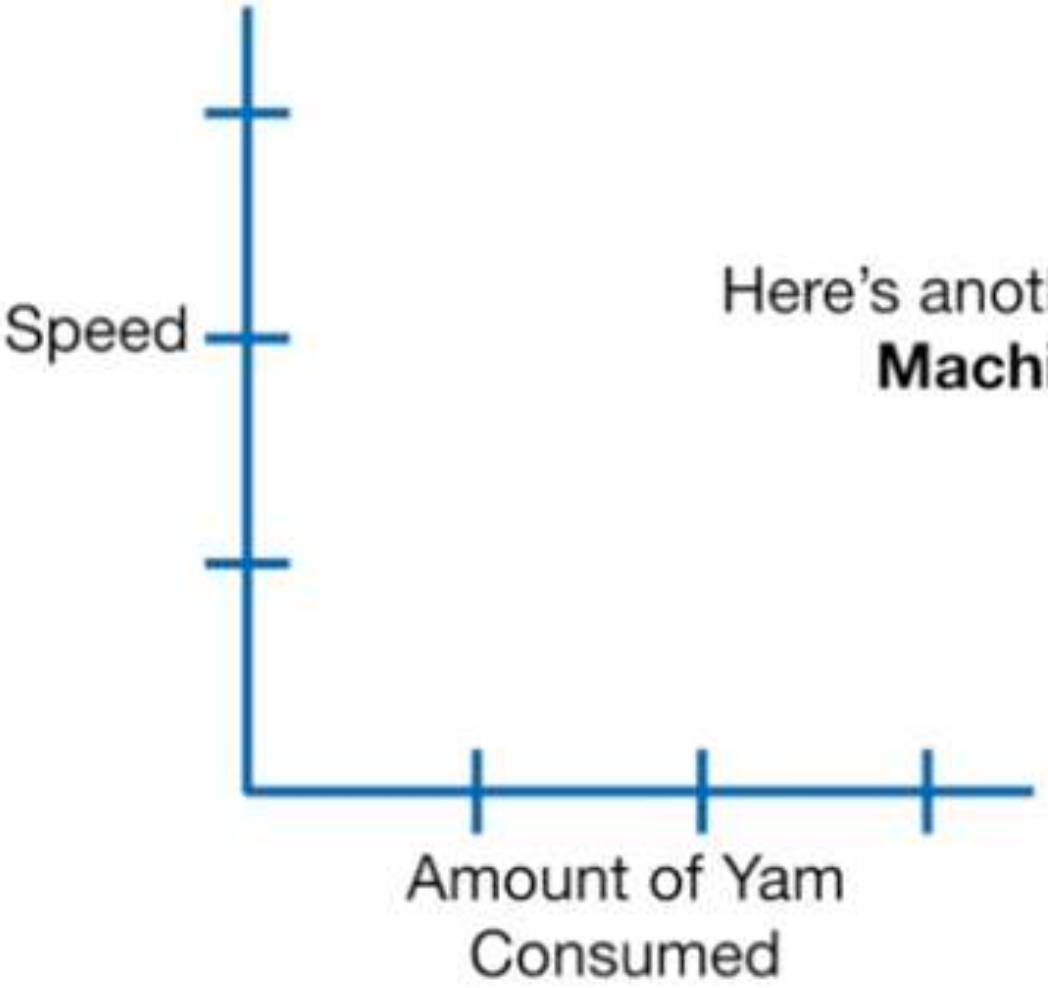
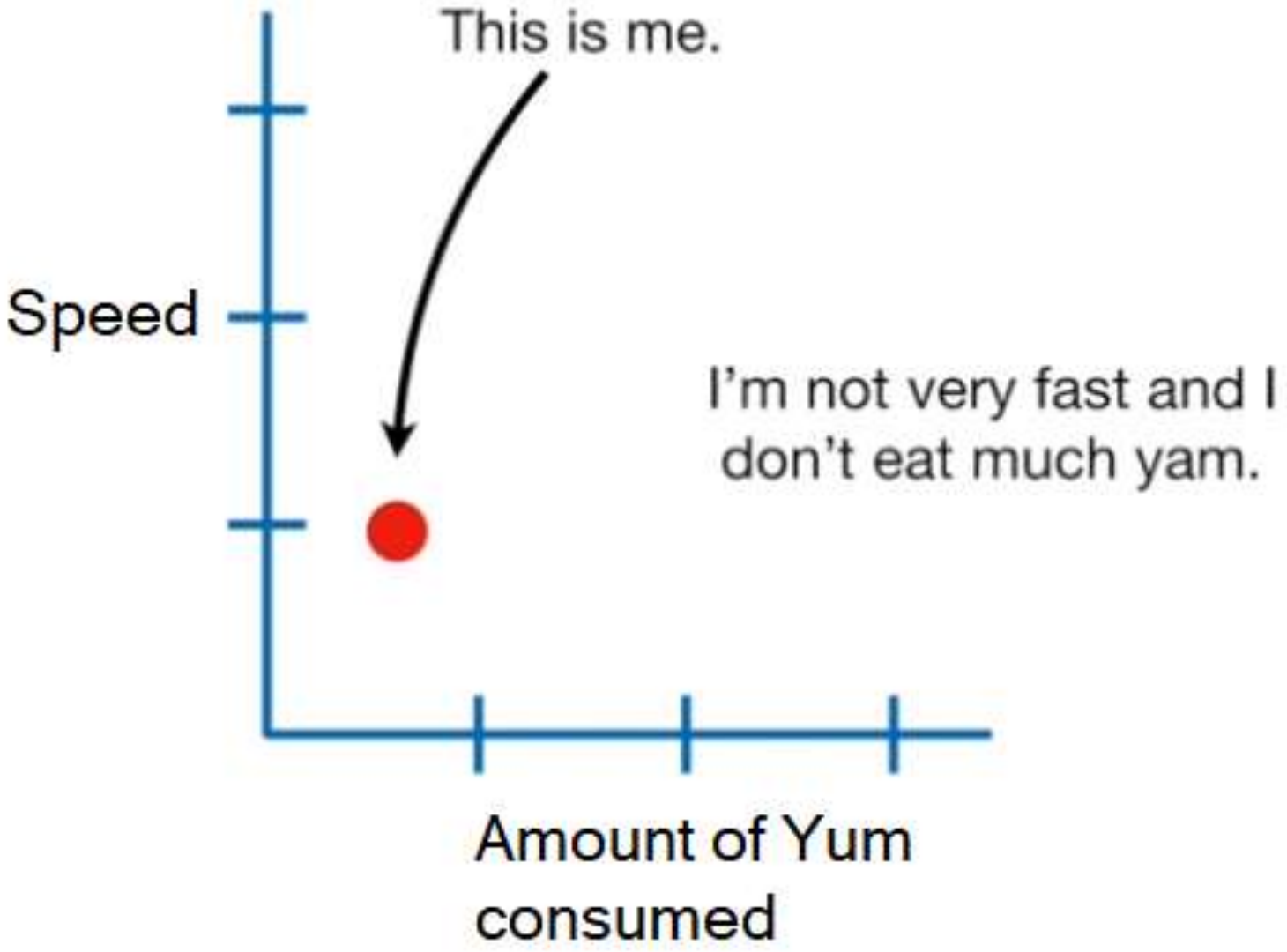
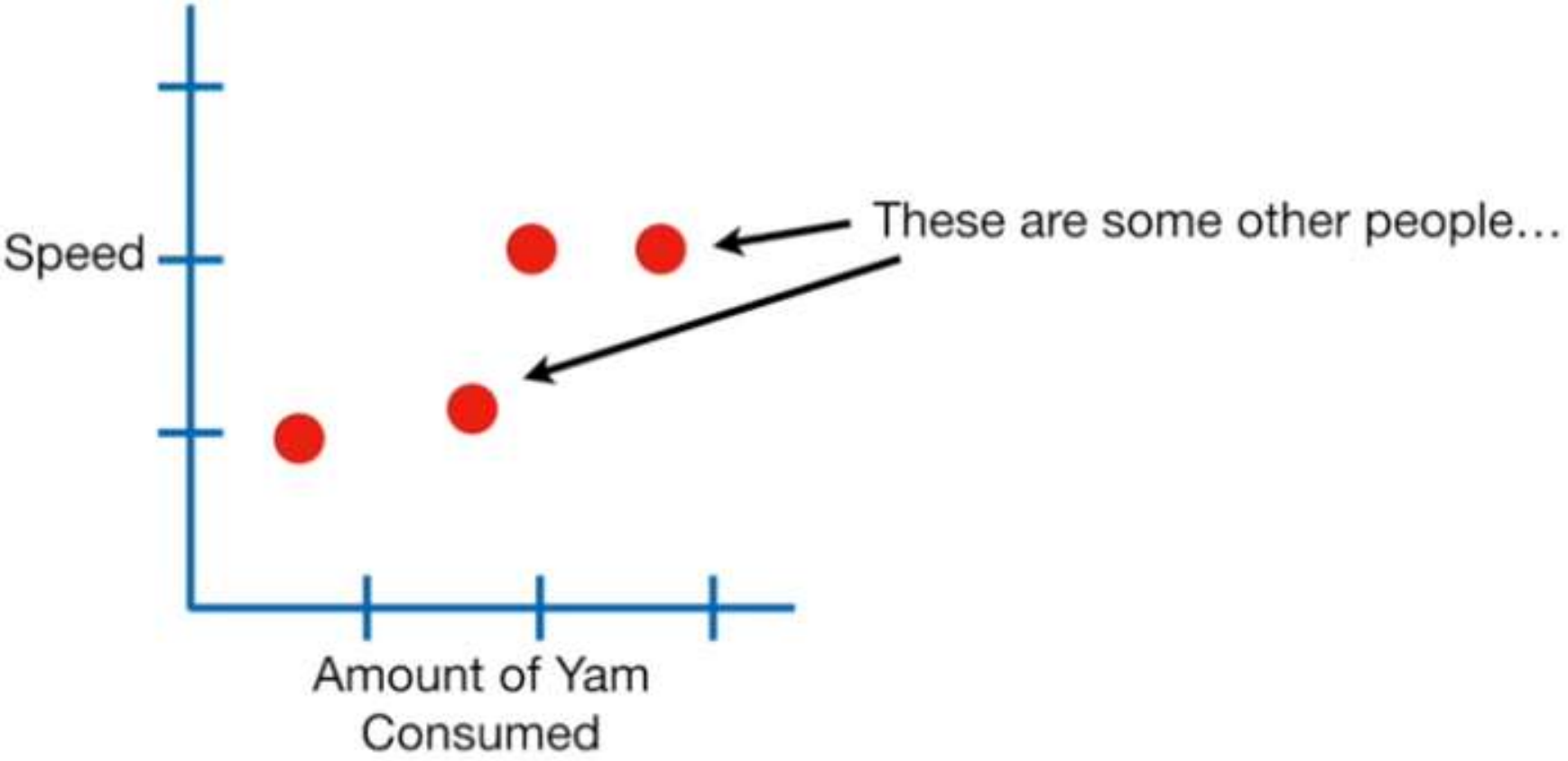
Machine Learning

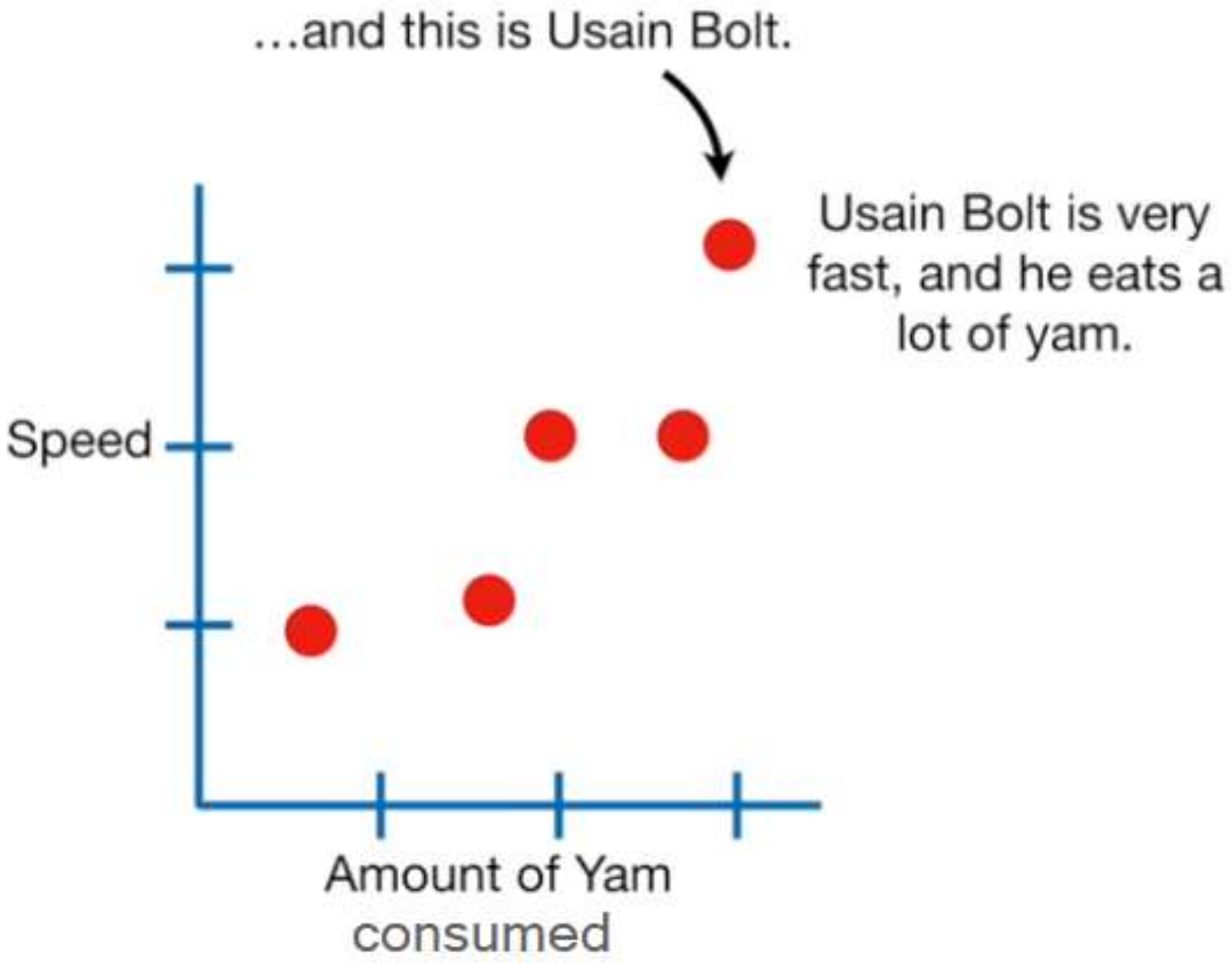
Straight Line V/S Squiggle

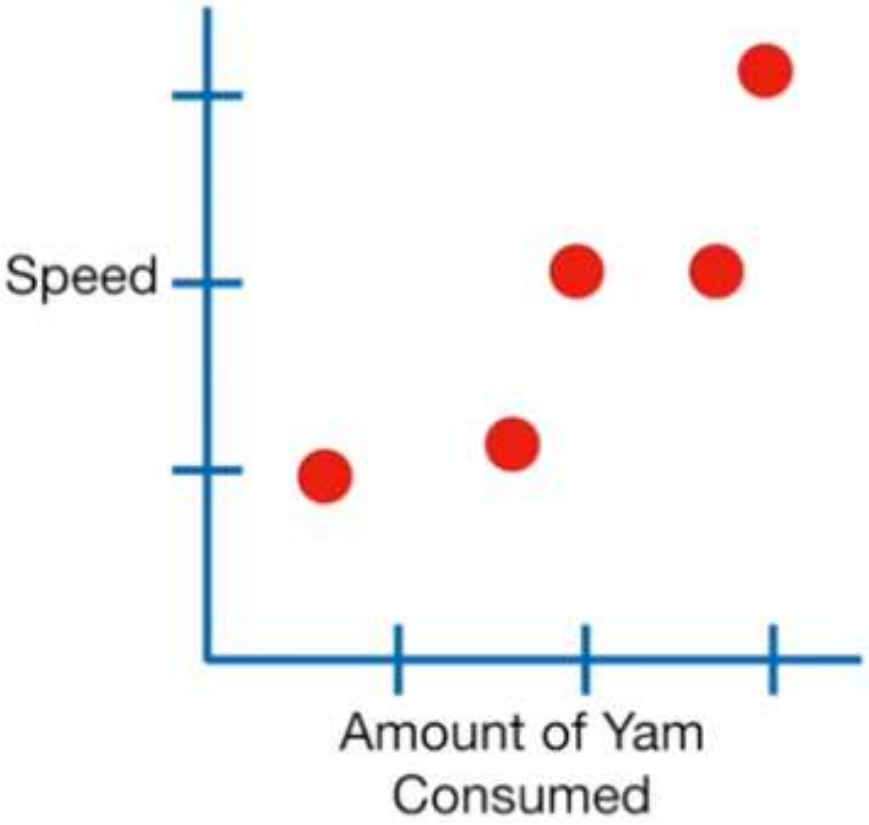


Here's another silly example of
Machine Learning.

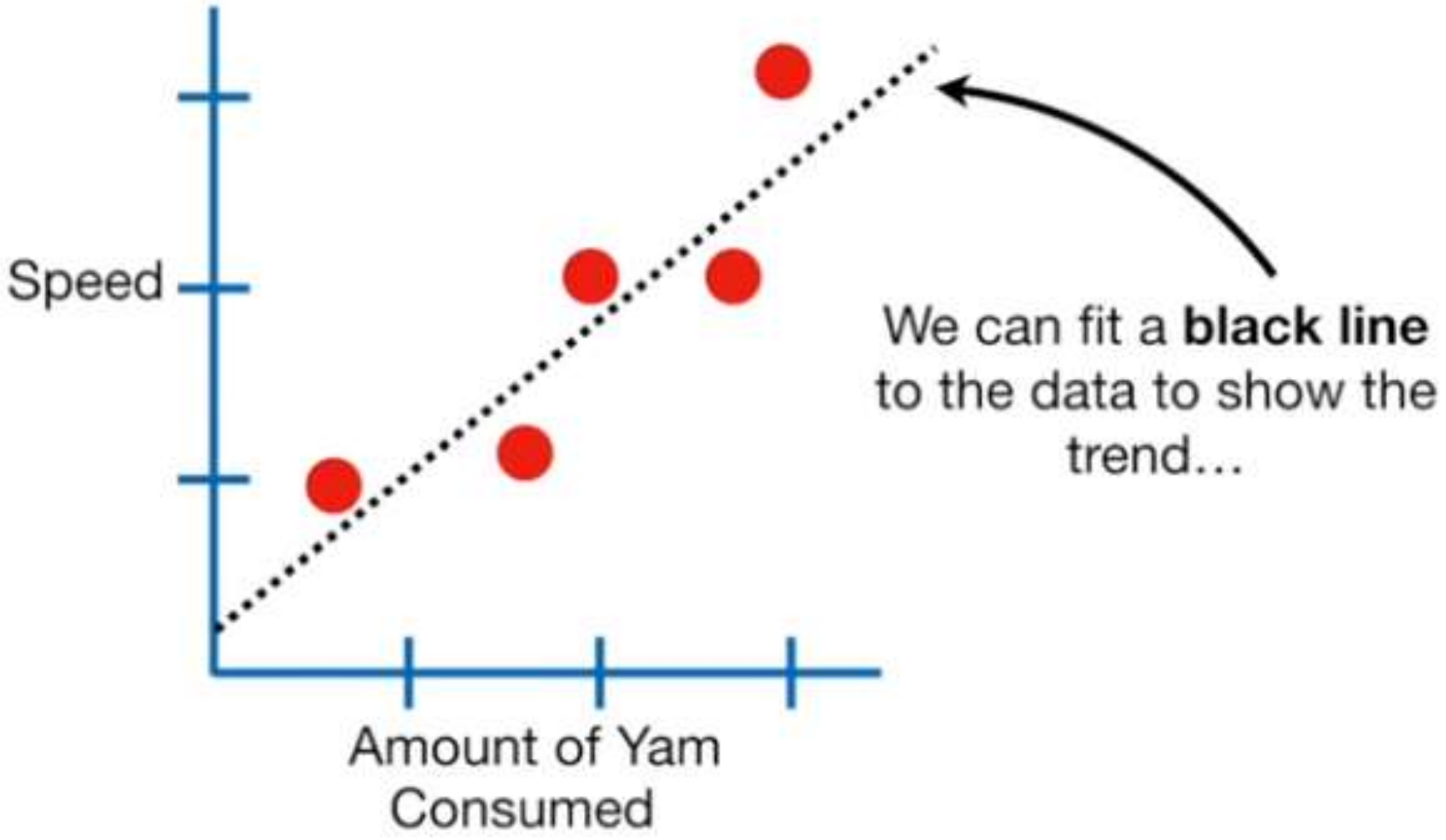

This is me.
Speed
I'm not very fast and I don't eat much yam.
Amount of Yum consumed

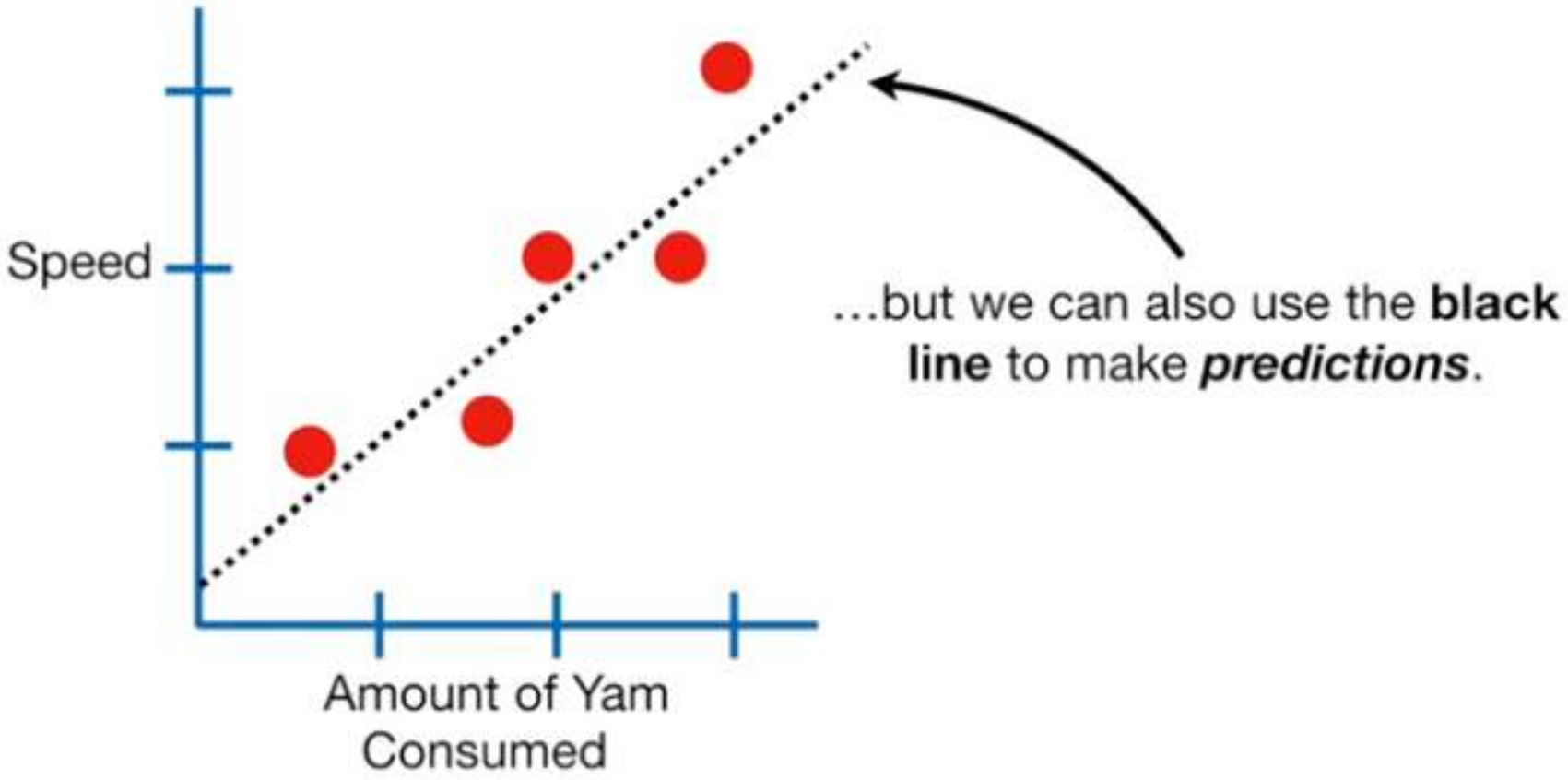

Speed
These are some other people…
Amount of Yam Consumed

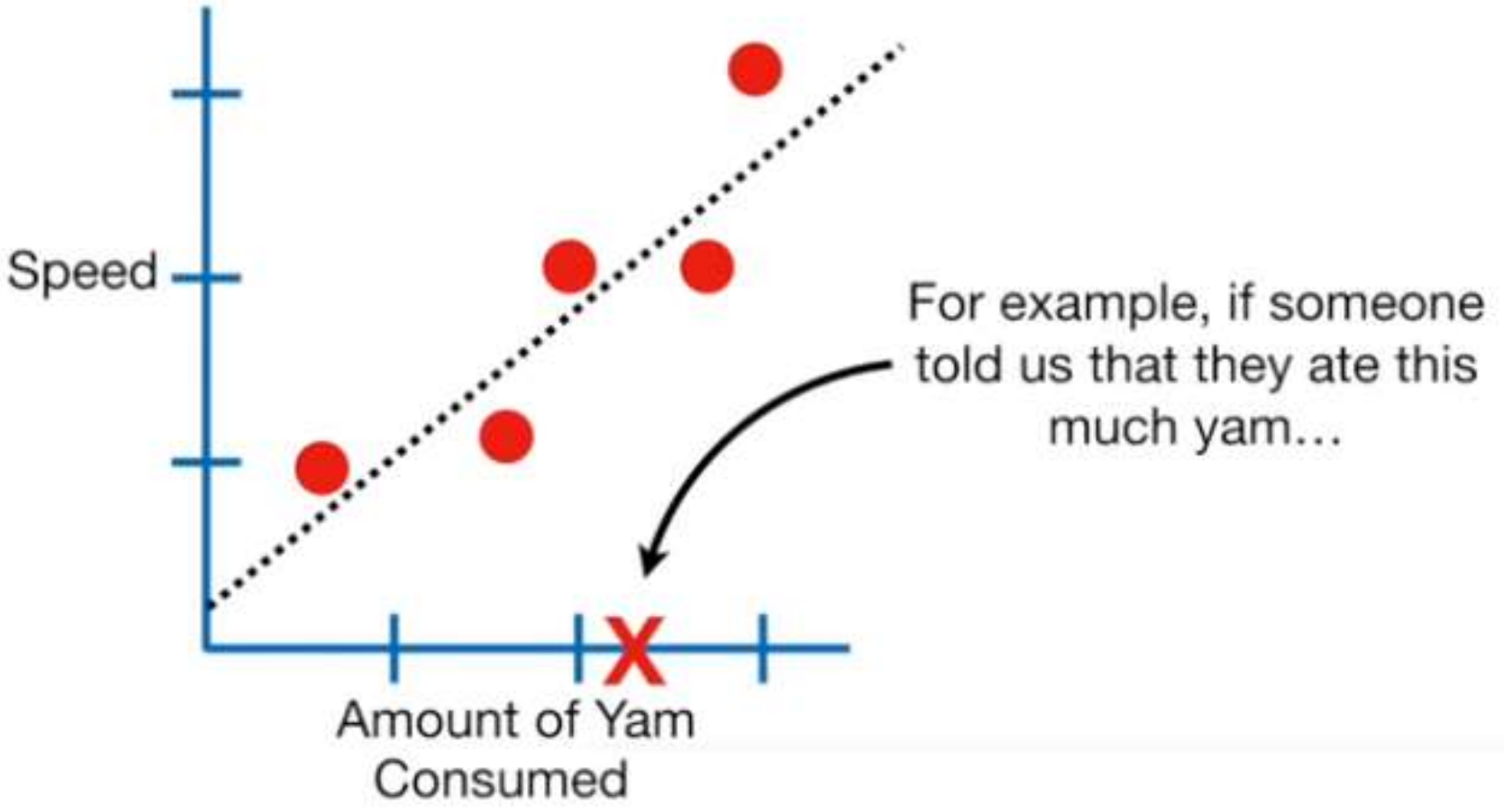

…and this is Usain Bolt.
Usain Bolt is very fast, and he eats a lot of yam.
Speed
Amount of Yam consumed

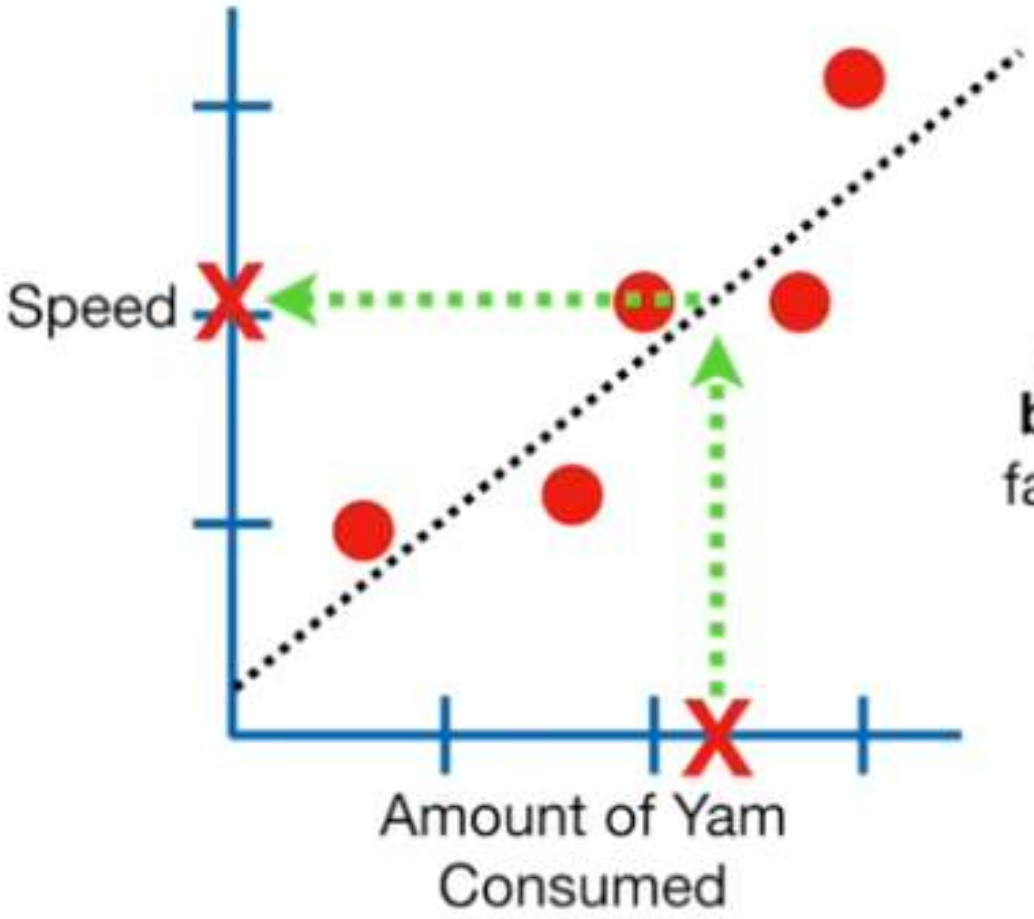


Given this pretend data, we see that the more yam someone eats, the faster they run the 100 meter dash.

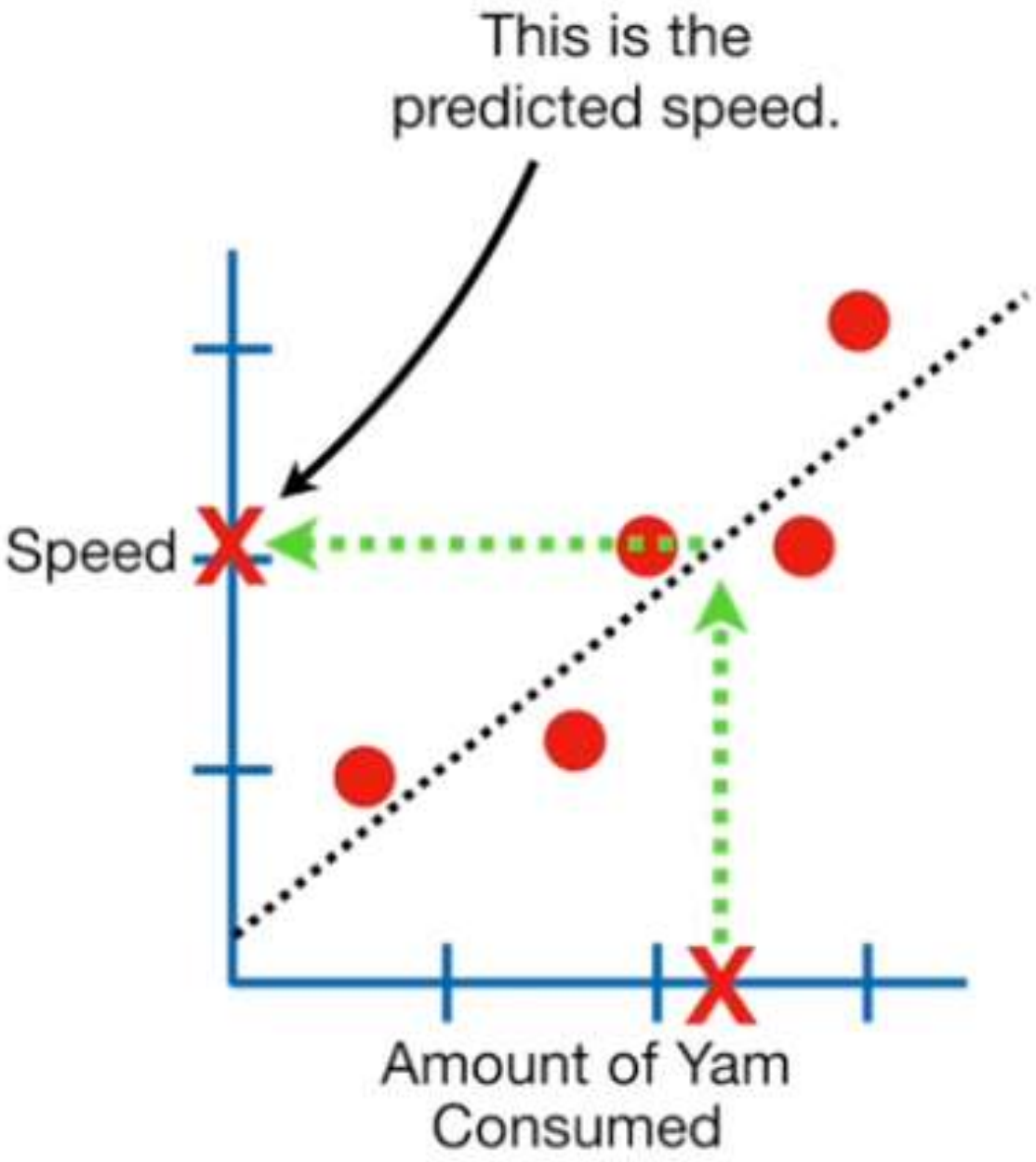

Speed
Amount of Yam Consumed
We can fit a **black line** to the data to show the trend…

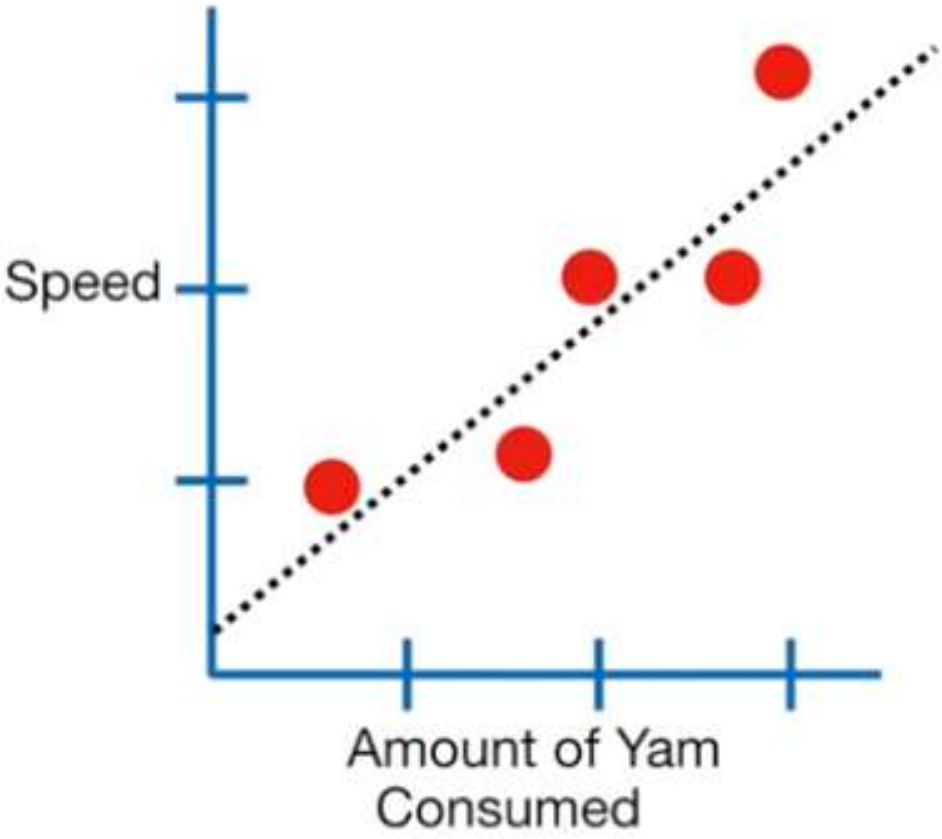

Speed
Amount of Yam Consumed
...but we can also use the **black line** to make ***predictions***.


Speed
For example, if someone told us that they ate this much yam…
X
Amount of Yam Consumed



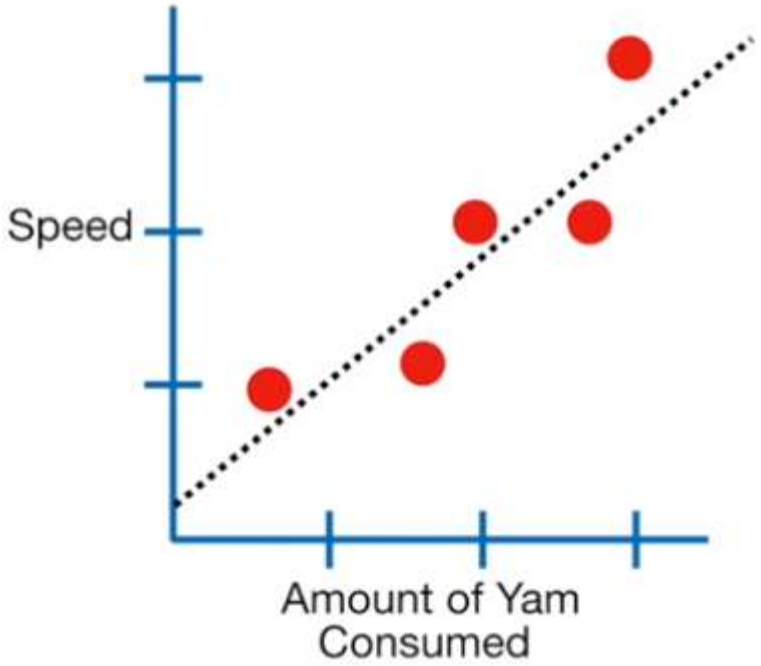
…then we could use the **black line** to predict how fast that person might run.

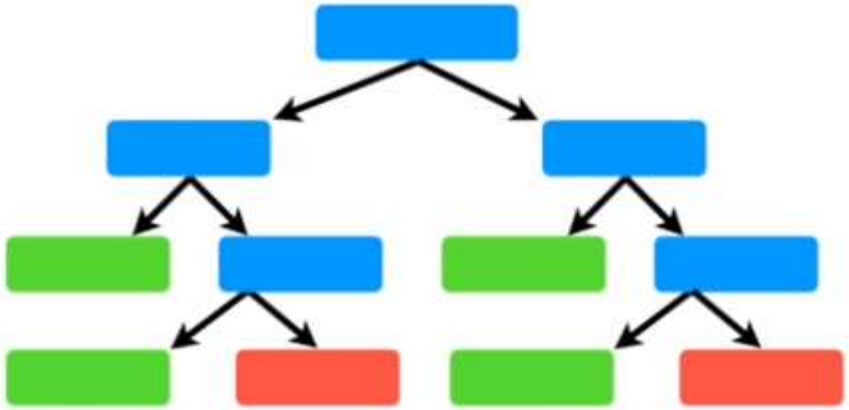

This is the
predicted speed.
Speed
Amount of Yam
Consumed

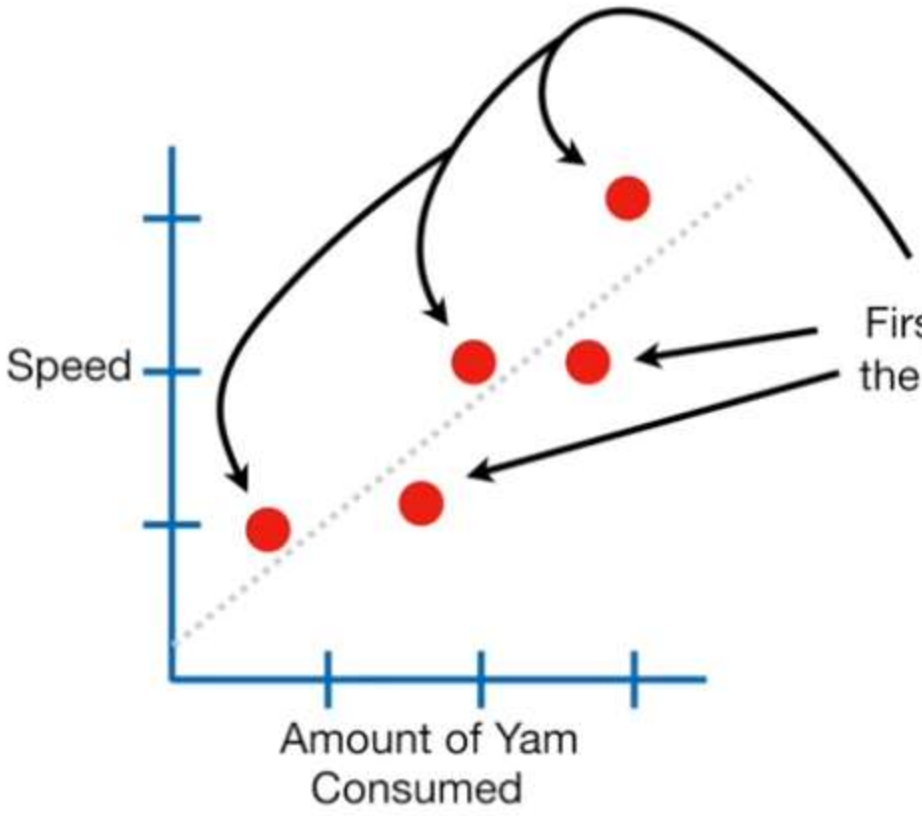


The **black line** is a type of **Machine Learning** because we can use it to make ***predictions***.

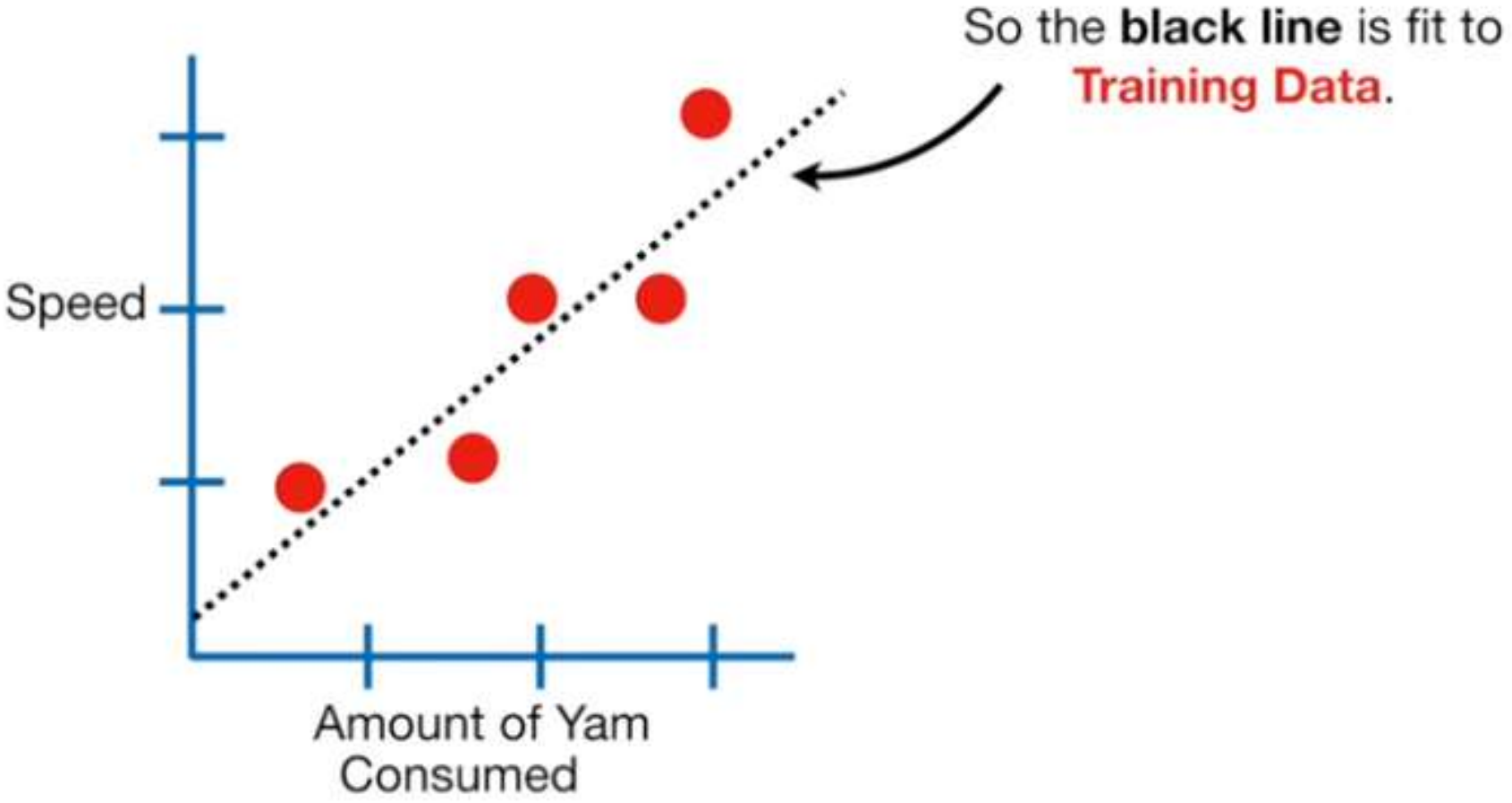
In general, **Machine Learning** is all about making ***predictions*** and ***classifications***.

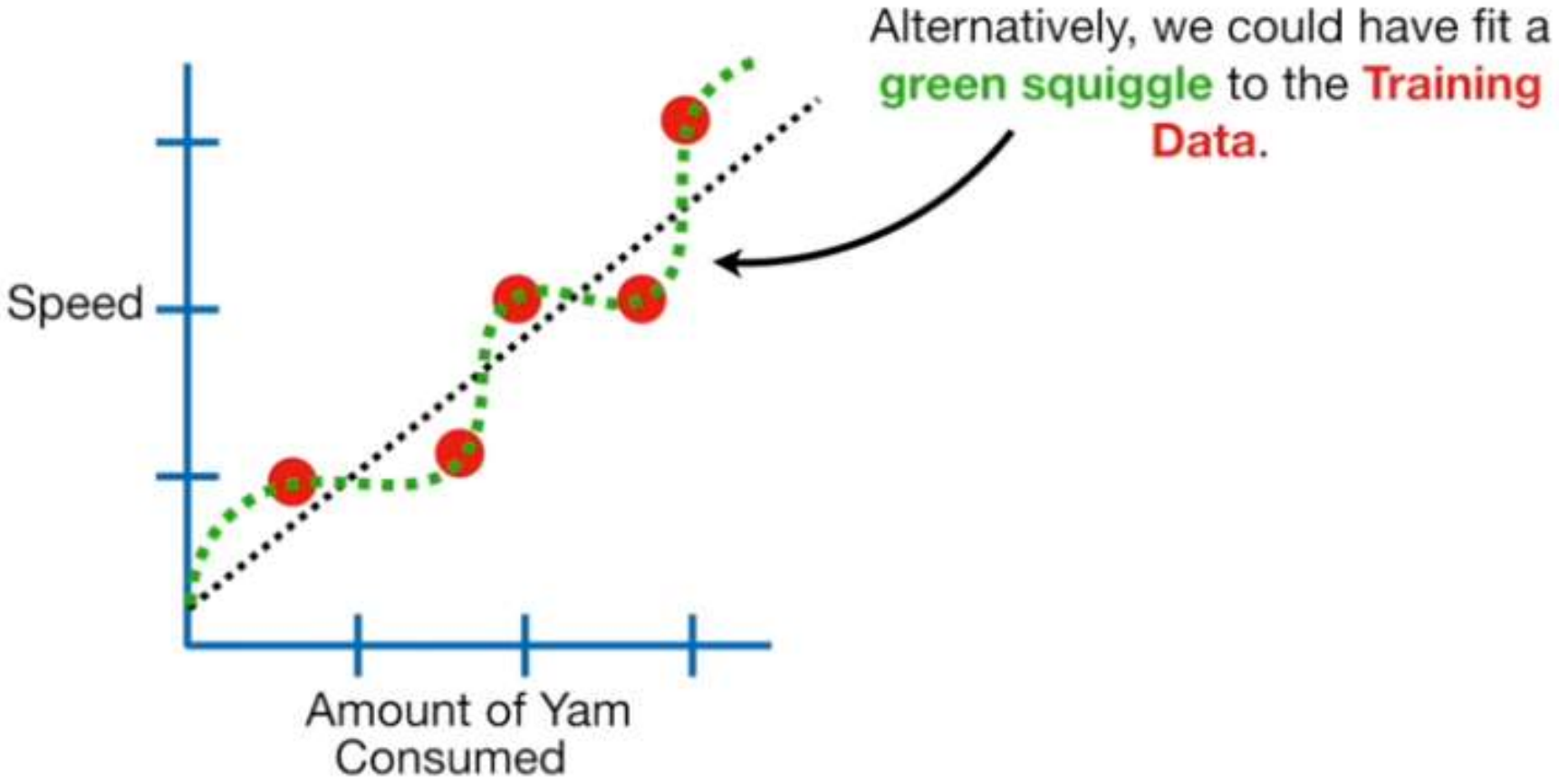


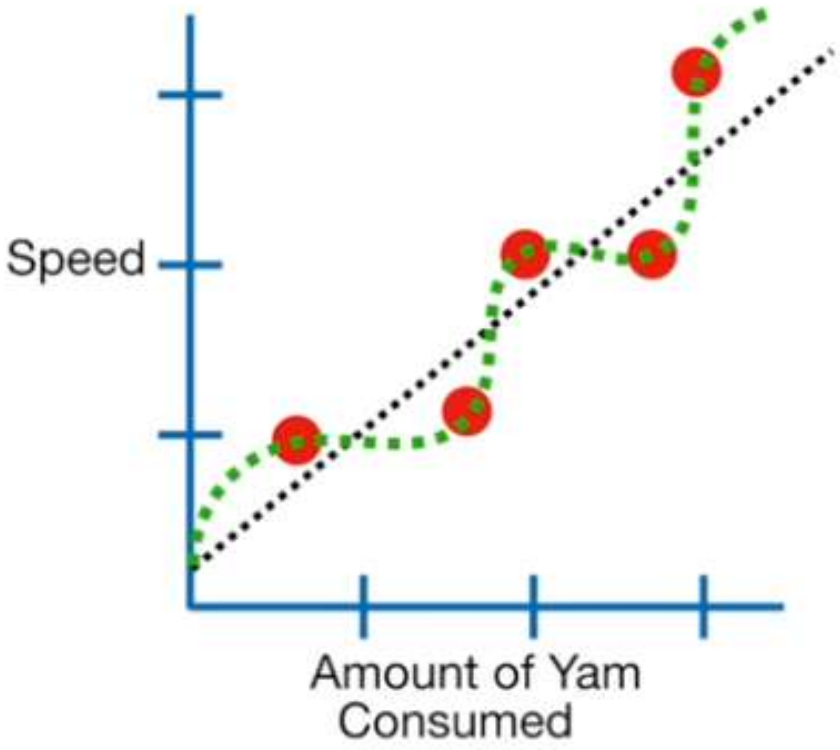




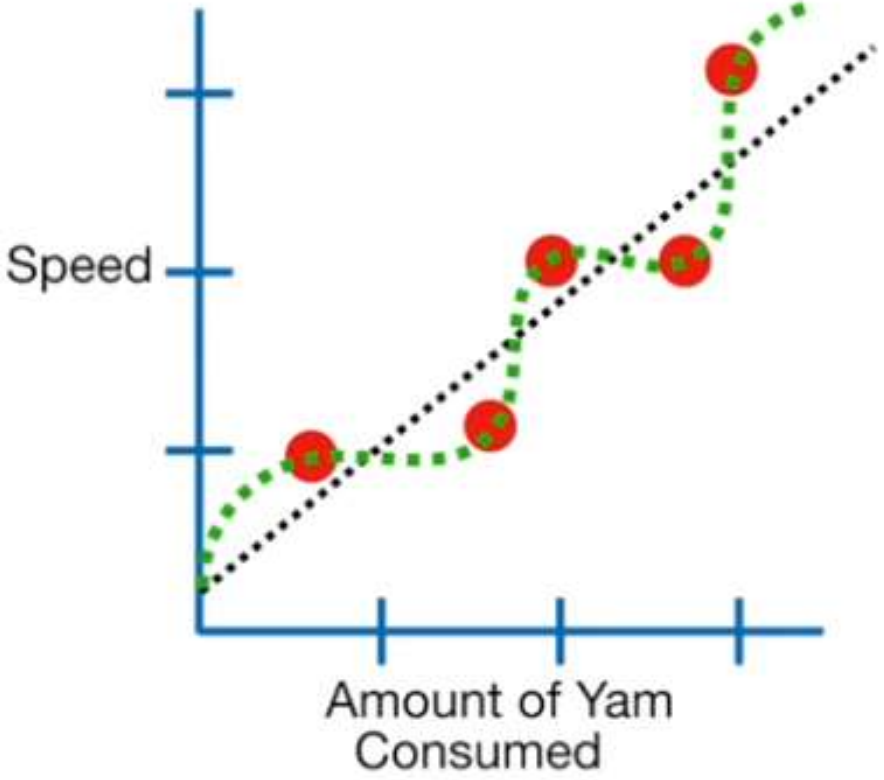
First of all, in **Machine Learning Lingo**, the original data is called **Training Data**.


So the **black line** is fit to
Training Data.
Speed
Amount of Yam
Consumed


Alternatively, we could have fit a green squiggle to the Training Data.
Speed
Amount of Yam Consumed

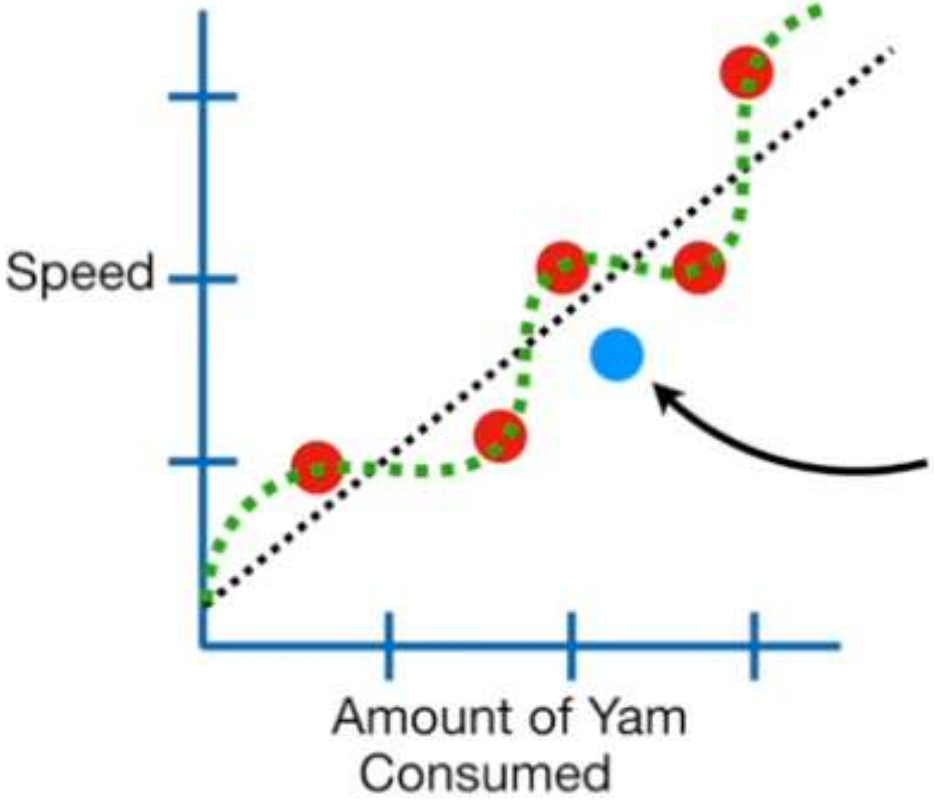


The **green squiggle** fits the **Training Data** better than the **black line**, but remember, the goal of **Machine Learning** is to make ***predictions***...

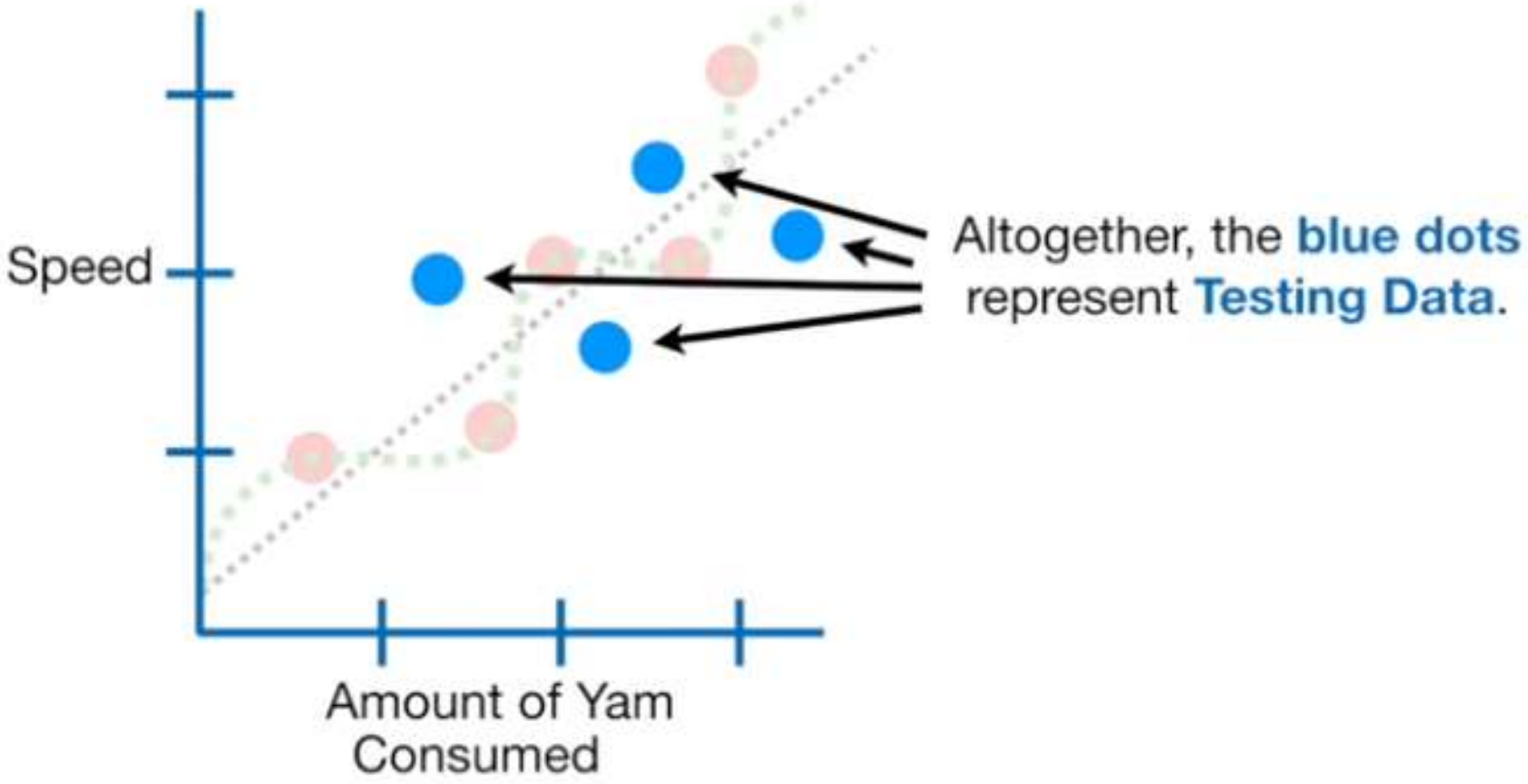


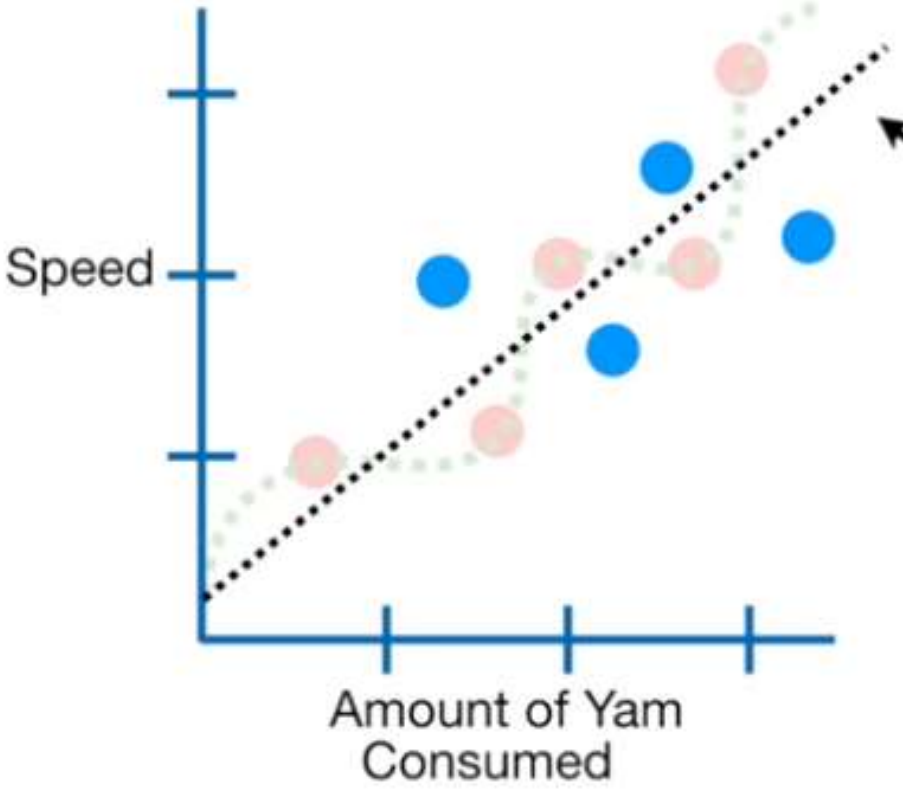
The **green squiggle** fits the **Training Data** better than the **black line**, but remember, the goal of **Machine Learning** is to make ***predictions***…

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

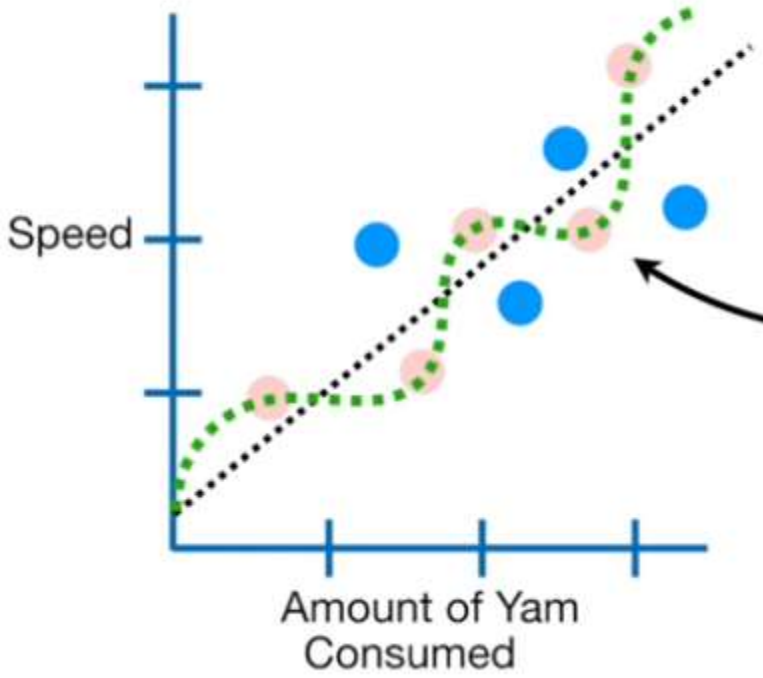


So we find a new person, and measure how fast they run and how much yam they eat…


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

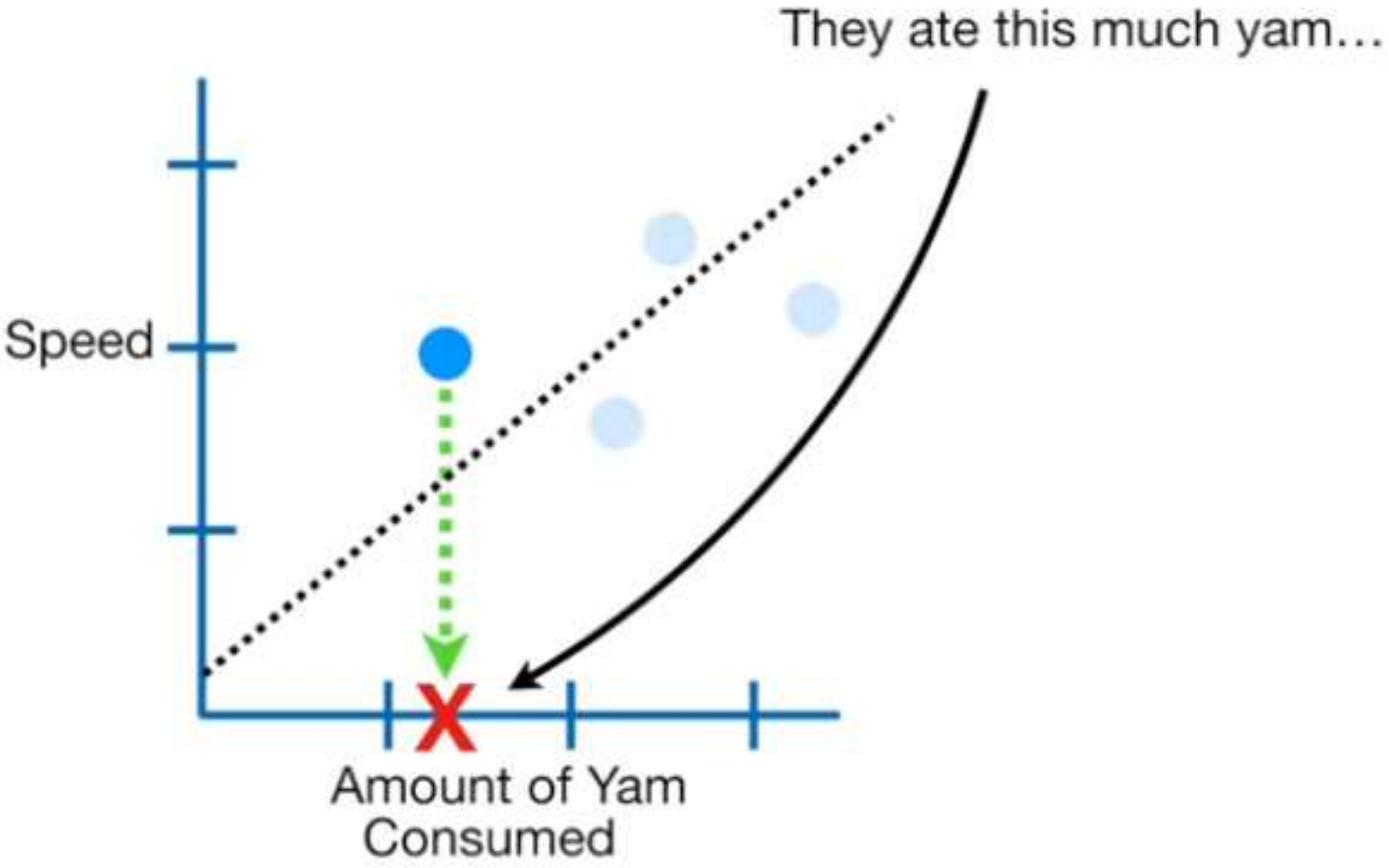


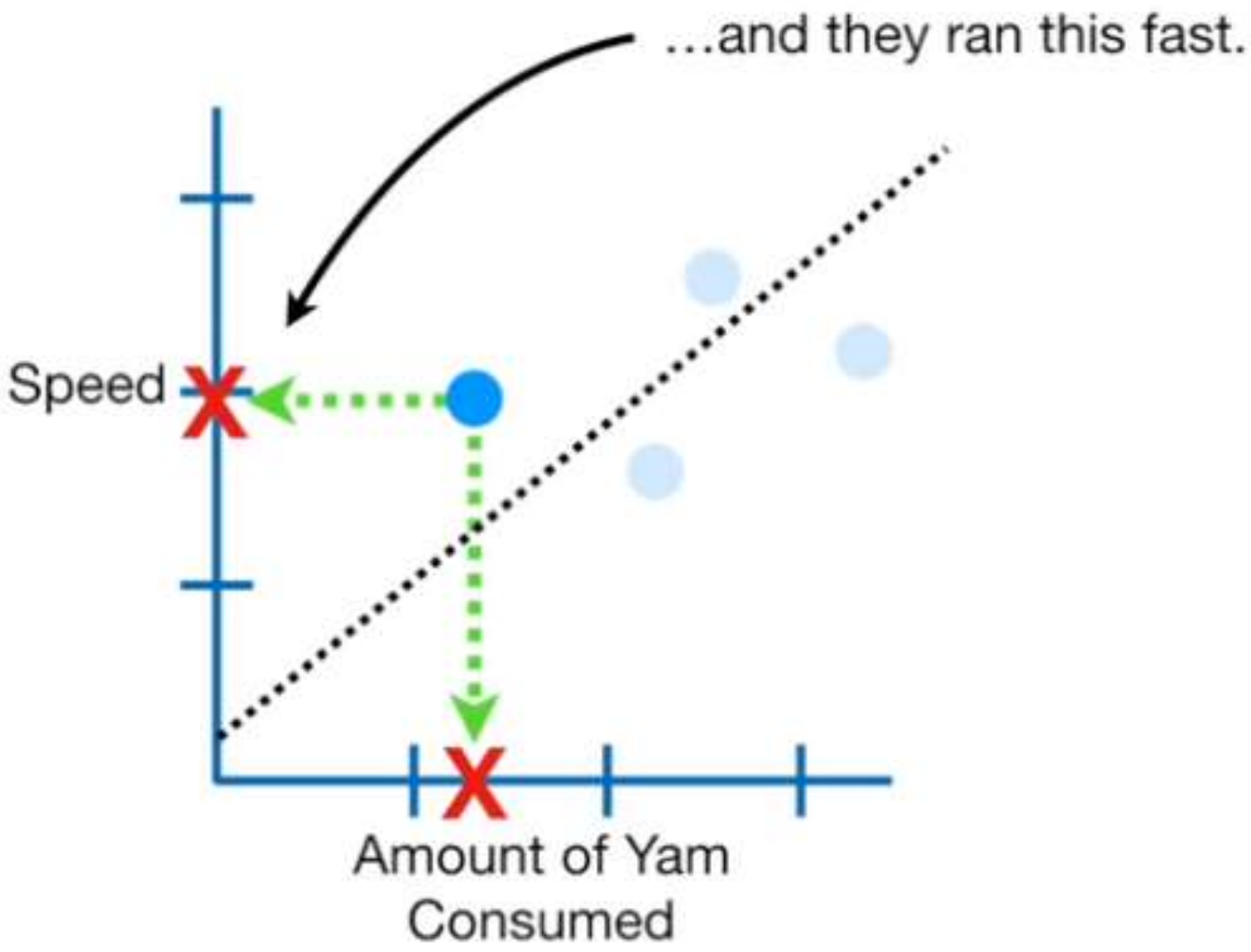
We use the **Testing Data** to compare the ***predictions*** made by the **black line**…

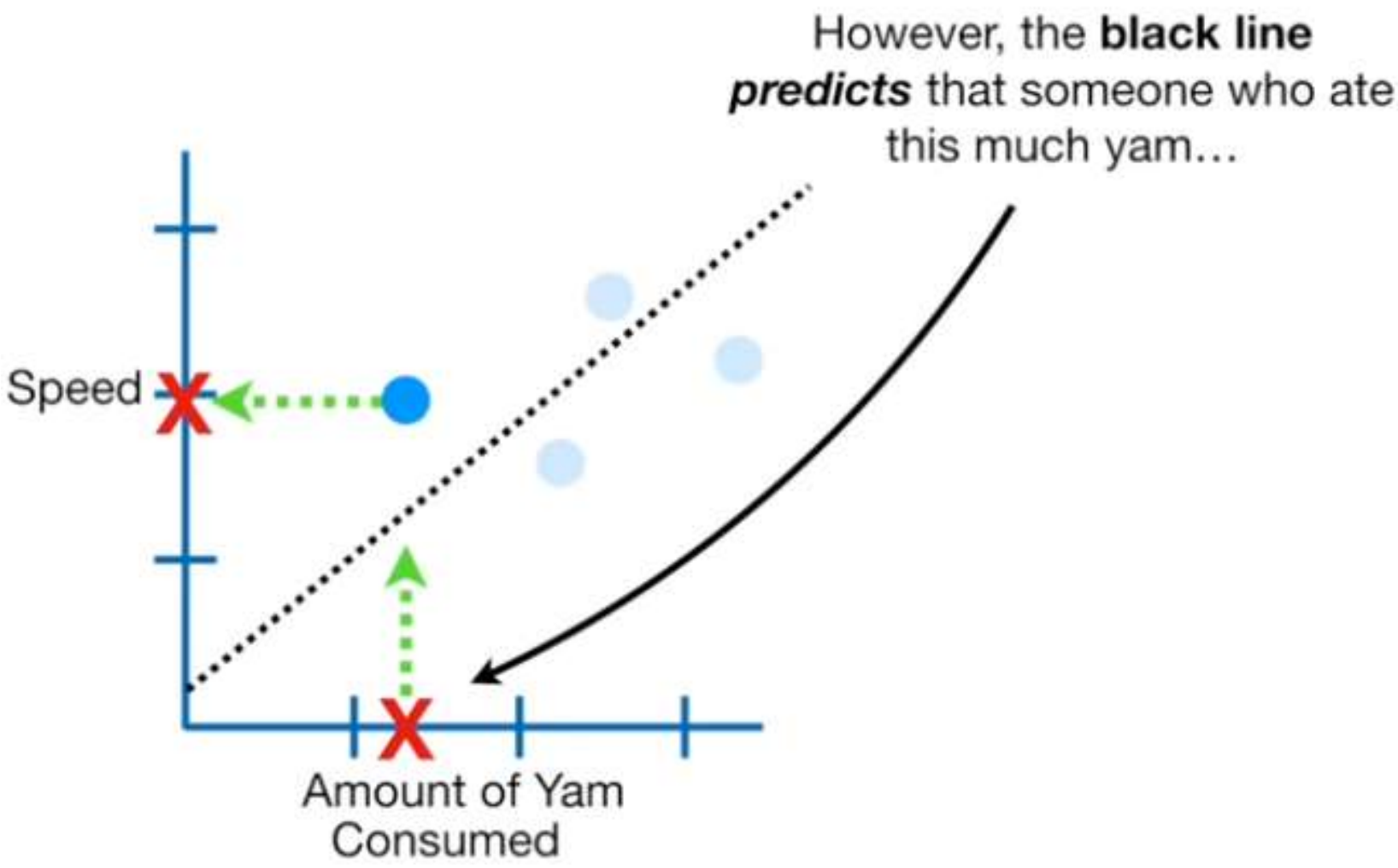


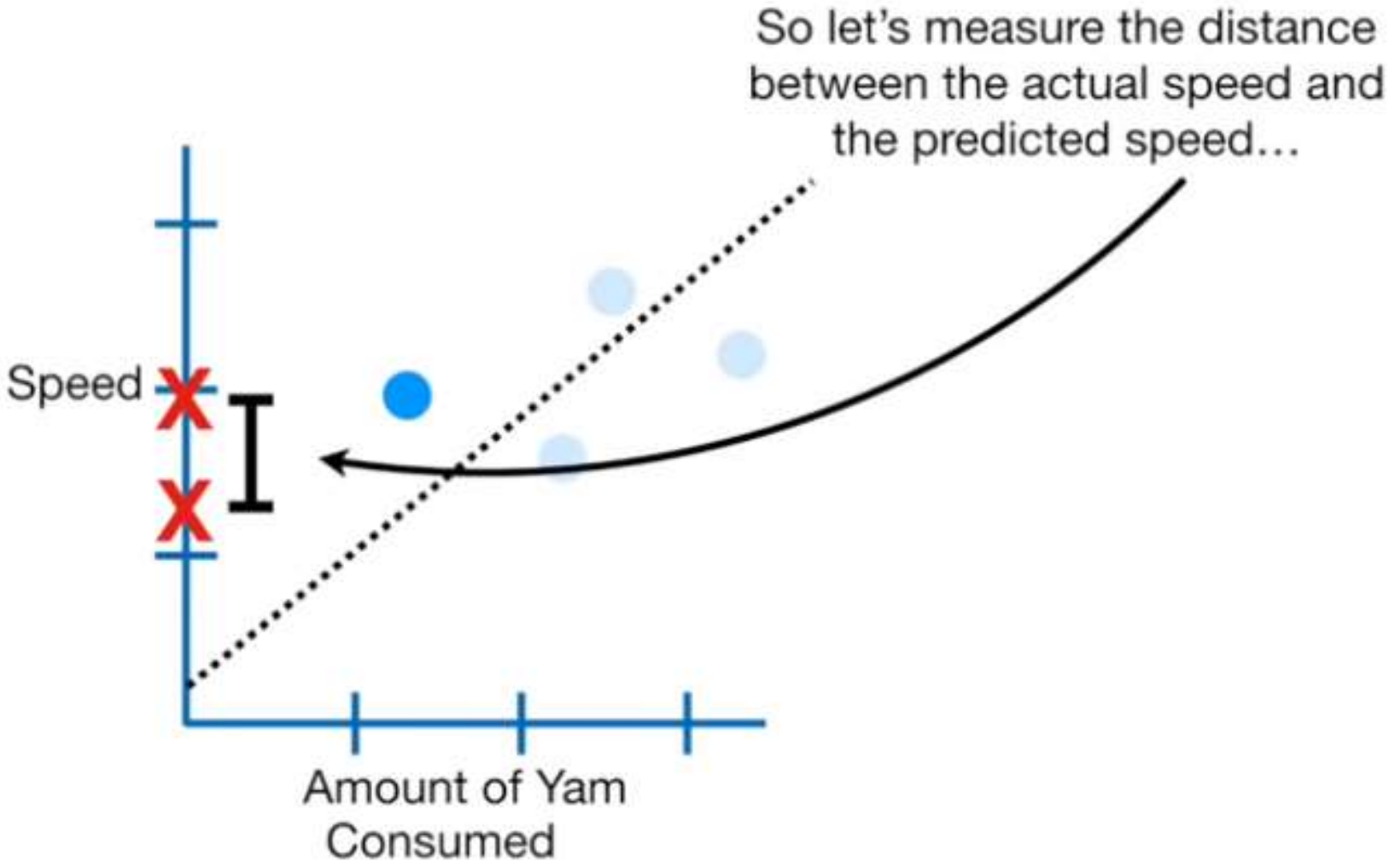
We use the **Testing Data** to compare the ***predictions*** made by the **black line**…

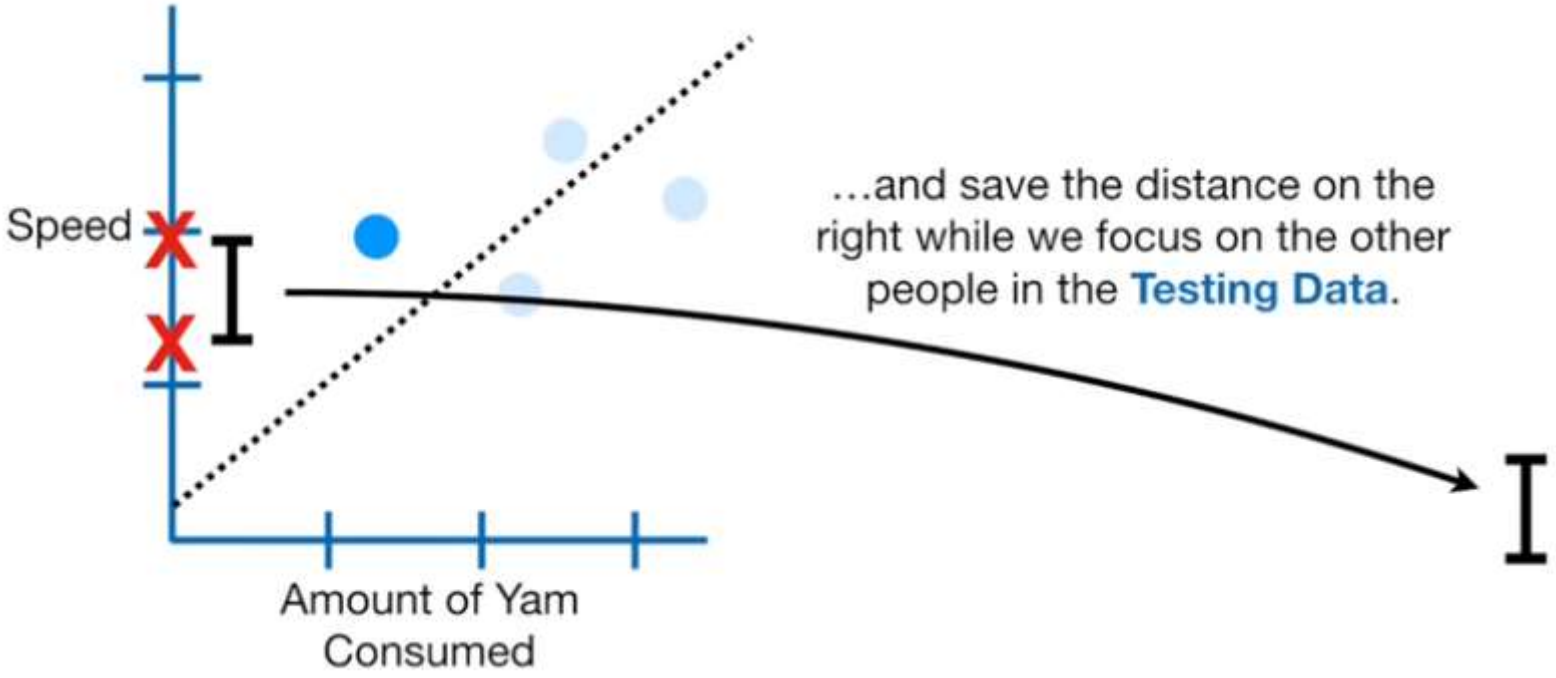
…to the ***predictions*** made by the **green squiggle**.

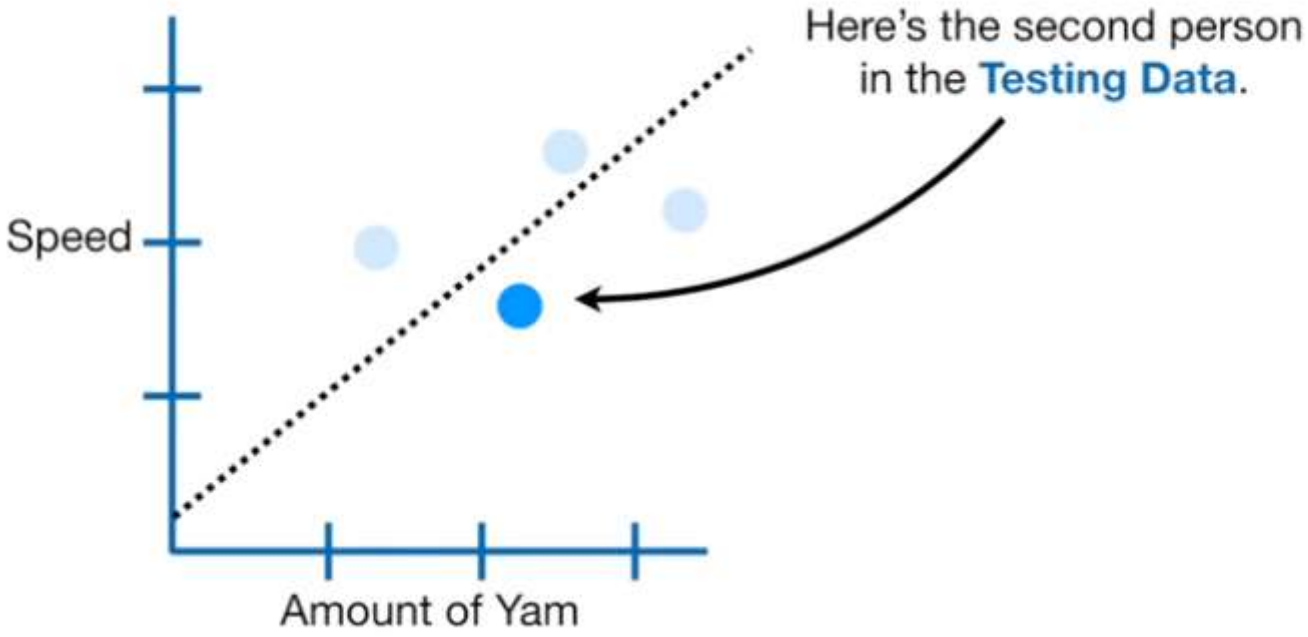

They ate this much yam…
Speed
Amount of Yam
Consumed

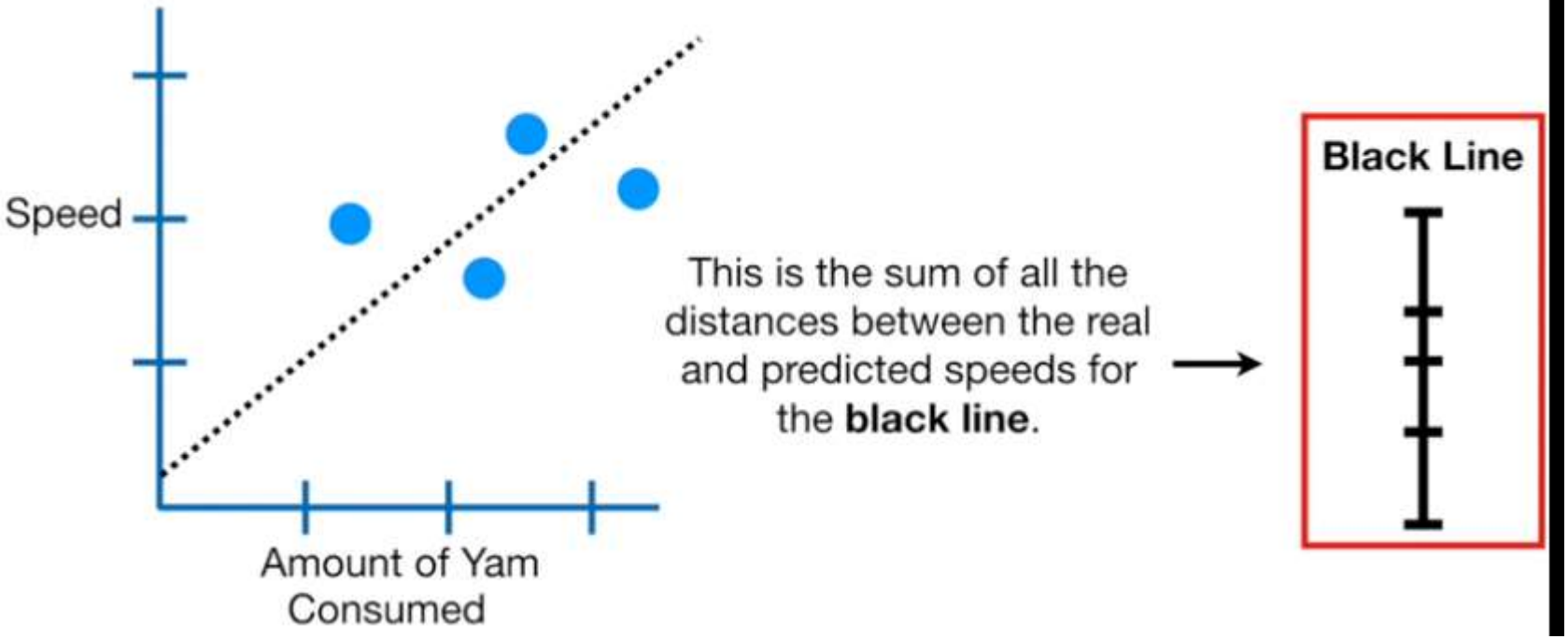

…and they ran this fast.
Speed
Amount of Yam Consumed

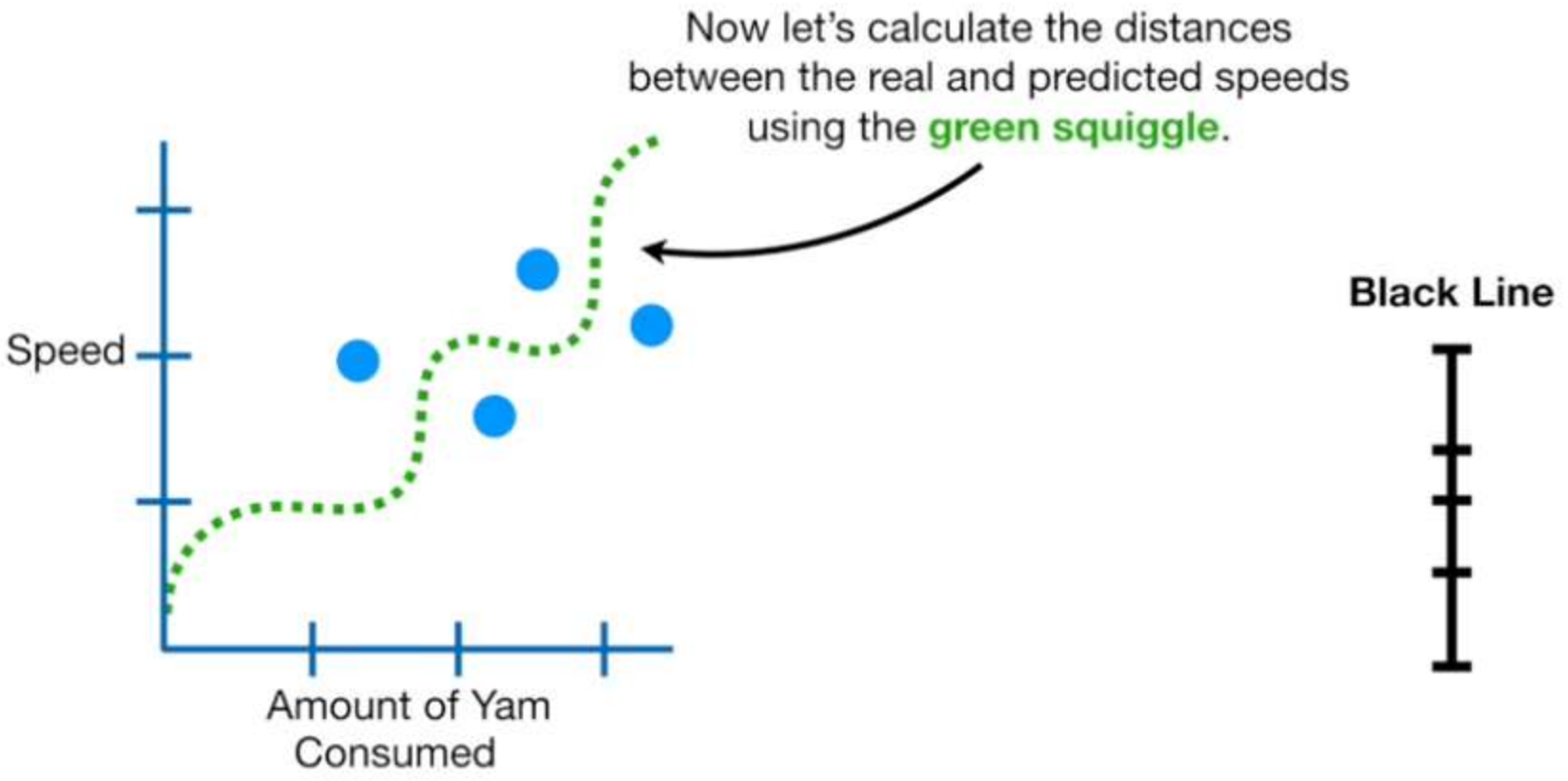

However, the **black line** ***predicts*** that someone who ate this much yam…
Speed
Amount of Yam Consumed

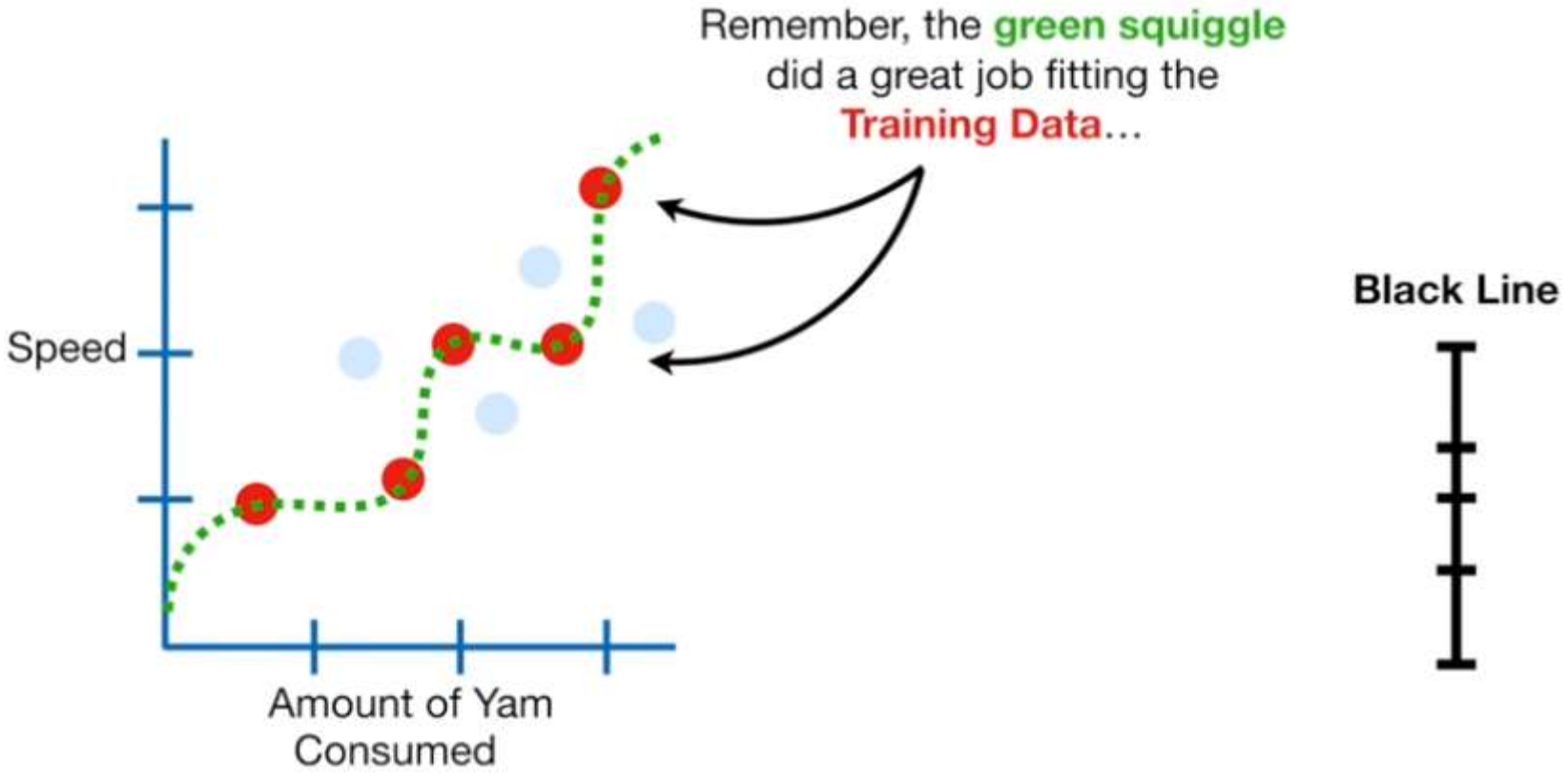

So let's measure the distance between the actual speed and the predicted speed…
Speed
Amount of Yam Consumed

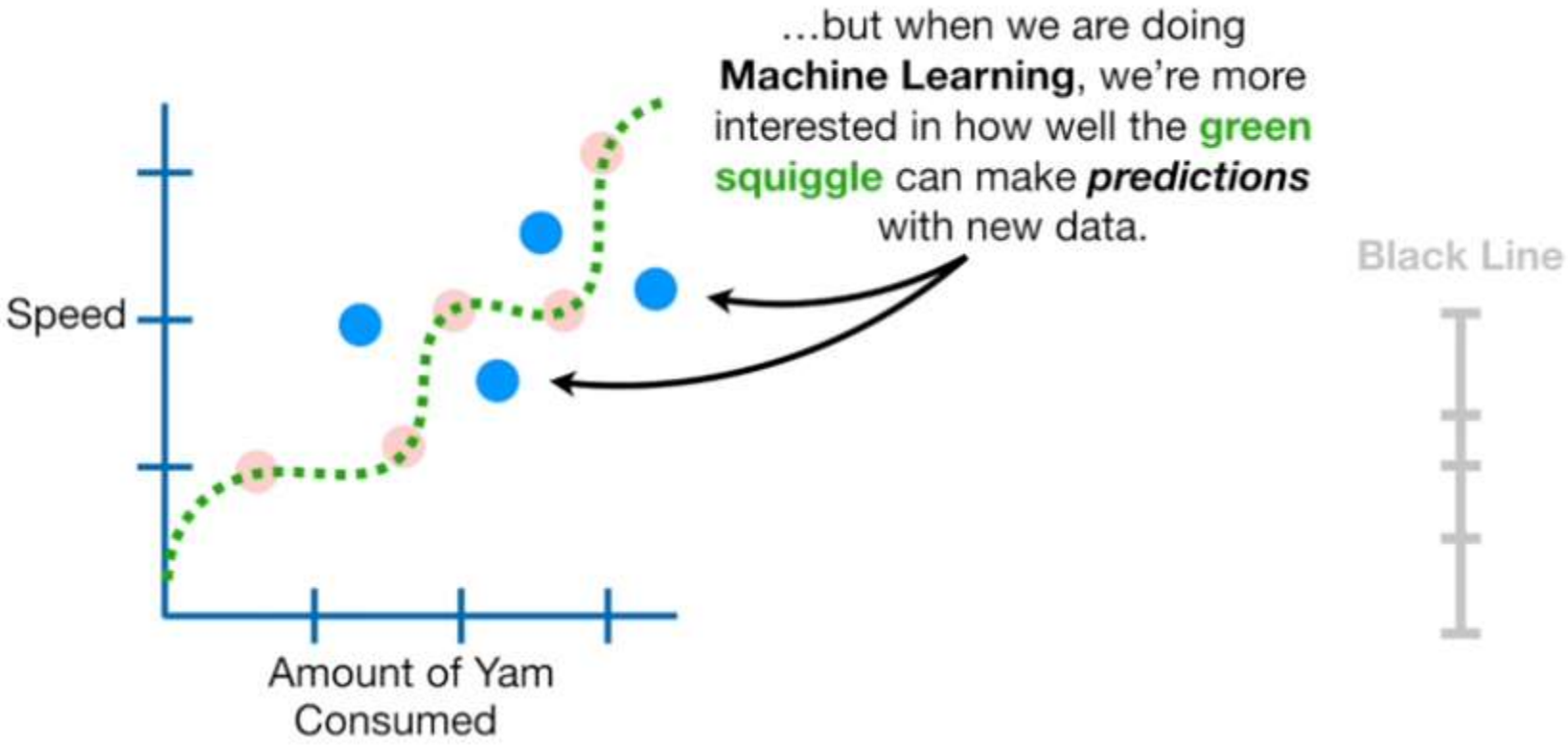

Speed
Amount of Yam Consumed
…and save the distance on the right while we focus on the other people in the Testing Data.

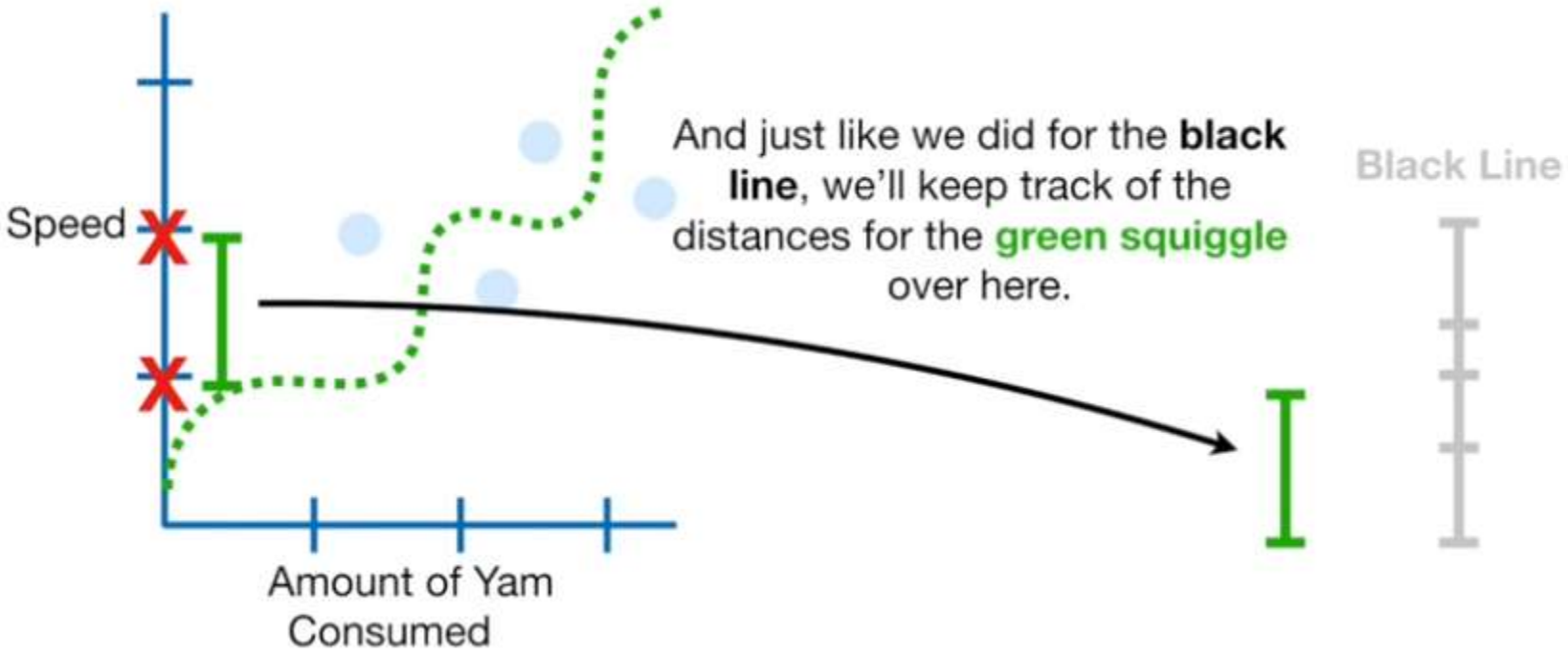

Here's the second person
in the Testing Data.
Speed
Amount of Yam

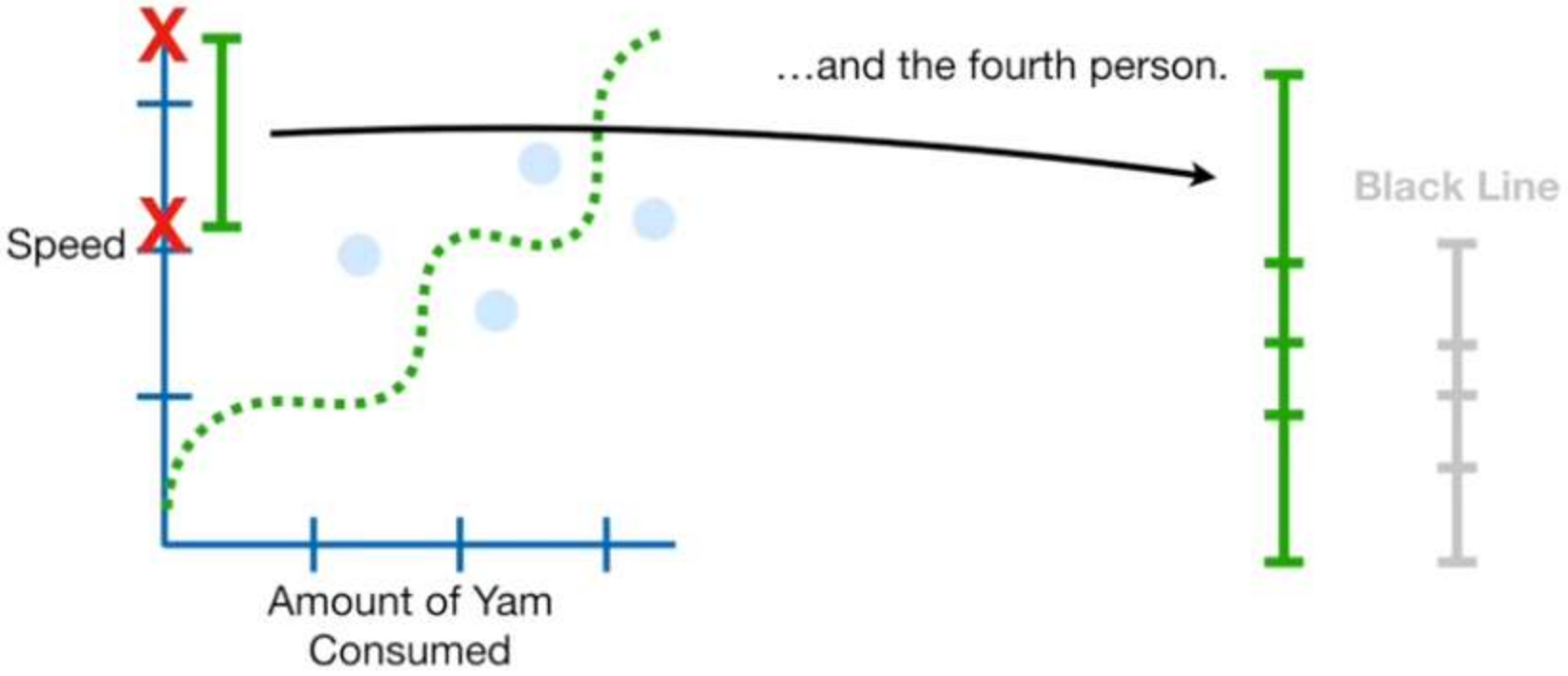
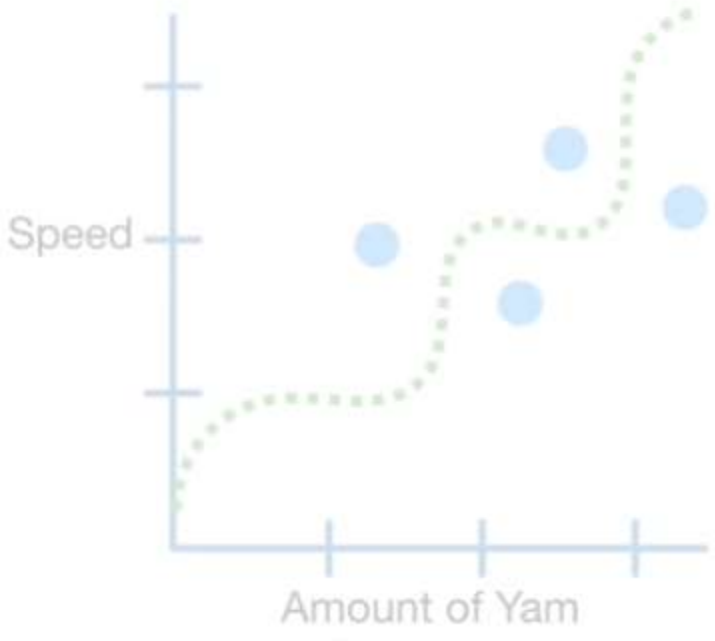
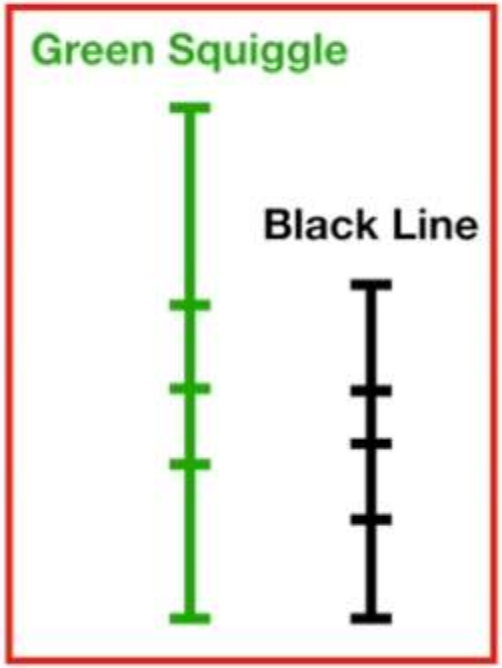

Speed
Amount of Yam Consumed
This is the sum of all the distances between the real and predicted speeds for the **black line**.
Black Line


Now let's calculate the distances between the real and predicted speeds using the green squiggle.
Speed
Amount of Yam Consumed
Black Line


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.
Speed
Amount of Yam
Consumed
Black Line


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


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

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