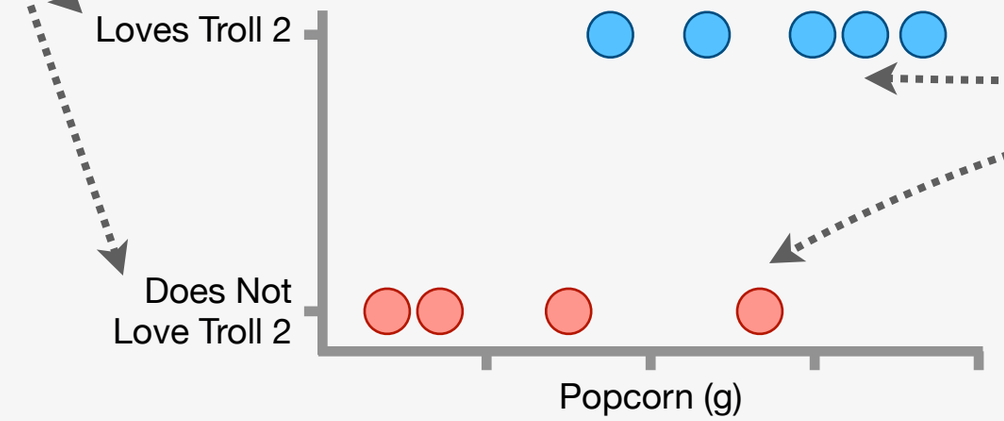
Chapter 06

Logistic Regression!!!

Logistic Regression: Main Ideas Part 1

1 **The Problem: Linear Regression** and **Linear Models** are great when we want to predict something that is ***continuous***, like **Height**, but what if we want to classify something ***discrete*** that only has two possibilities, like whether or not someone loves the movie Troll 2?



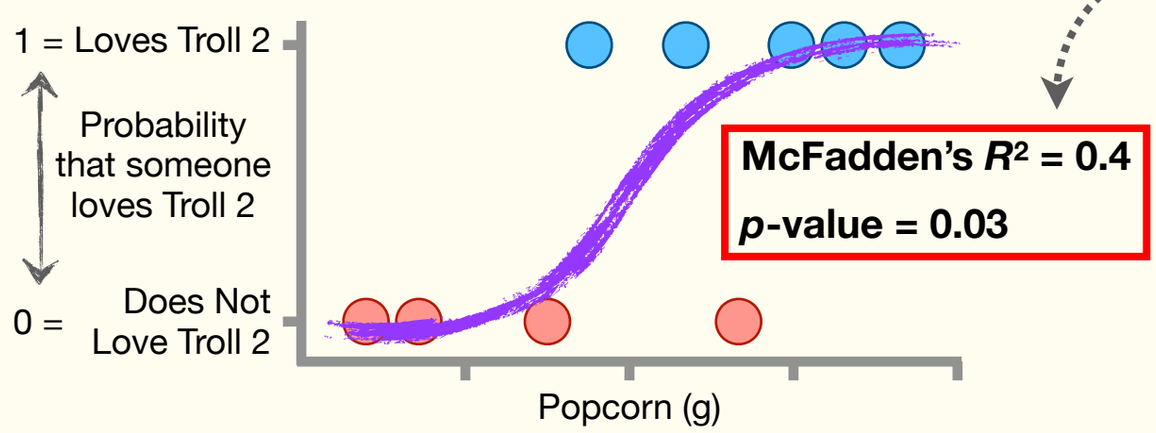
In this example, we measured the amount of Popcorn a bunch of people ate (in grams), which is ***continuous***, and whether they **Love Troll 2** or **Do Not Love Troll 2**, which is ***discrete***.

= Loves Troll 2

= Does Not Love Troll 2

The goal is to make a ***classifier*** that uses the amount of Popcorn someone eats to classify whether or not they love Troll 2.

2 **A Solution: Logistic Regression**, which probably should have been named **Logistic Classification** since it's used to classify things, fits a **squiggle** to data that tells us the predicted probability (between **0** and **1**) for ***discrete*** variables, like whether or not someone loves Troll 2.



Like **Linear Regression**, **Logistic Regression** has metrics that are similar to R^2 to give us a sense of how accurate our predictions will be, and it also calculates ***p*-values**.

Even better, all of the tricks we can do with **Linear Models** also apply to **Logistic Regression**, so we can mix and match ***discrete*** and ***continuous*** features to make ***discrete*** classifications.

BAM!!!

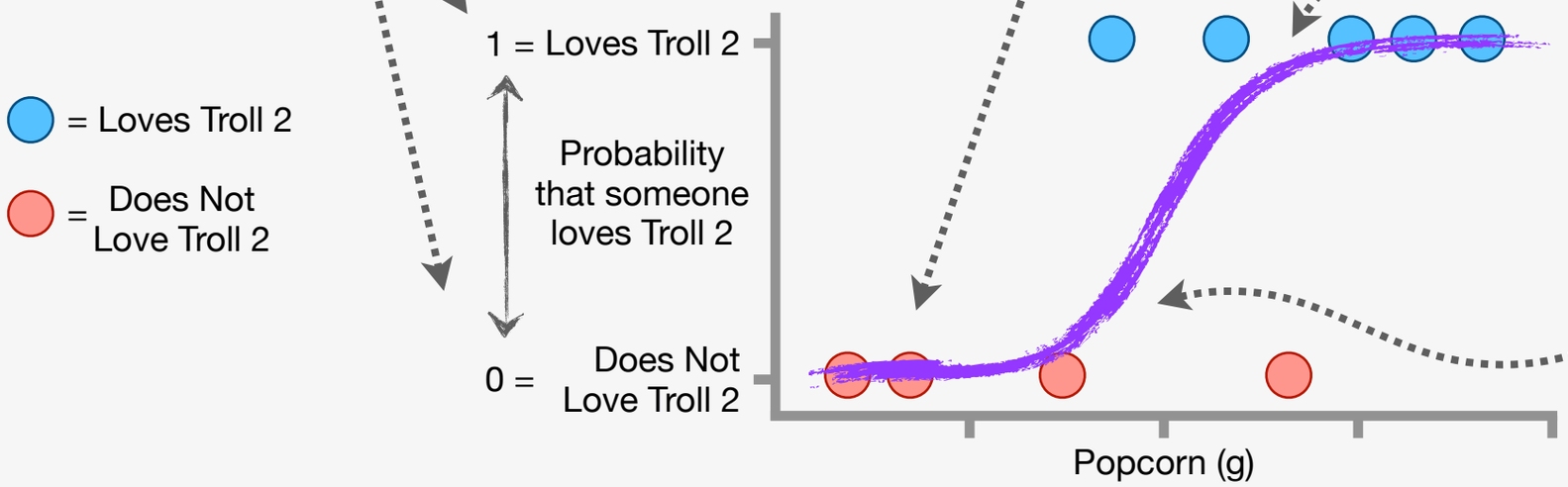
Logistic Regression: Main Ideas Part 2

3 The y-axis on a **Logistic Regression** graph represents probability and goes from **0** to **1**. In this case, it's the probability that someone loves Troll 2.

The colored dots are the **Training Data**, which we used to fit the squiggle. The data consist of **4** people who **Do Not Love Troll 2** and **5** people who **Love Troll 2**.

The squiggle tells us the predicted probability that someone loves Troll 2, which means that when the squiggle is close to the top of the graph, there's a high probability (a probability close to **1**) that someone will love Troll 2…

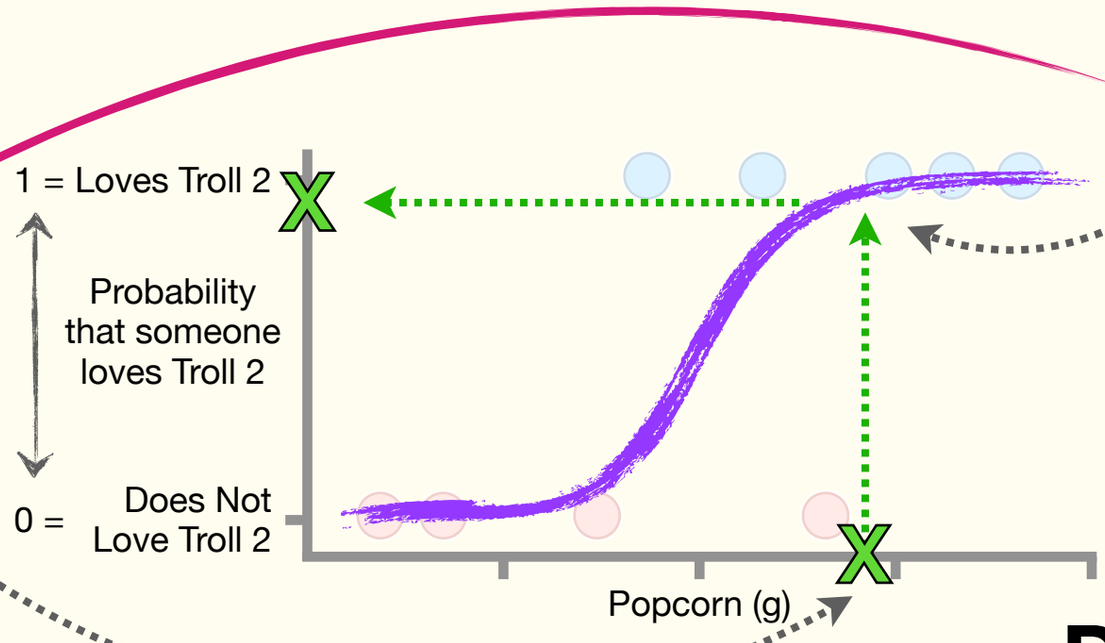
…and when the squiggle is close to the bottom of the graph, there's a low probability (a probability close to **0**) that someone will love Troll 2.



4 If someone new comes along and tells us that they ate this much Popcorn…

…then the squiggle tells us that there's a relatively high probability that that person will love Troll 2.

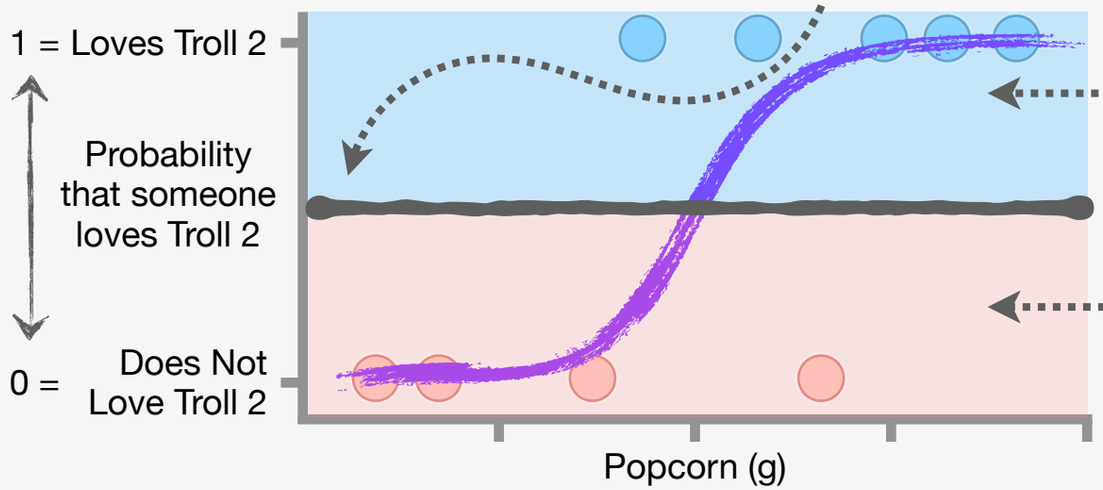
Specifically, the corresponding y-axis value on the squiggle tells us that the probability that this person loves Troll 2 is **0.96**.



DOUBLE BAM!!!

Logistic Regression: Main Ideas Part 3

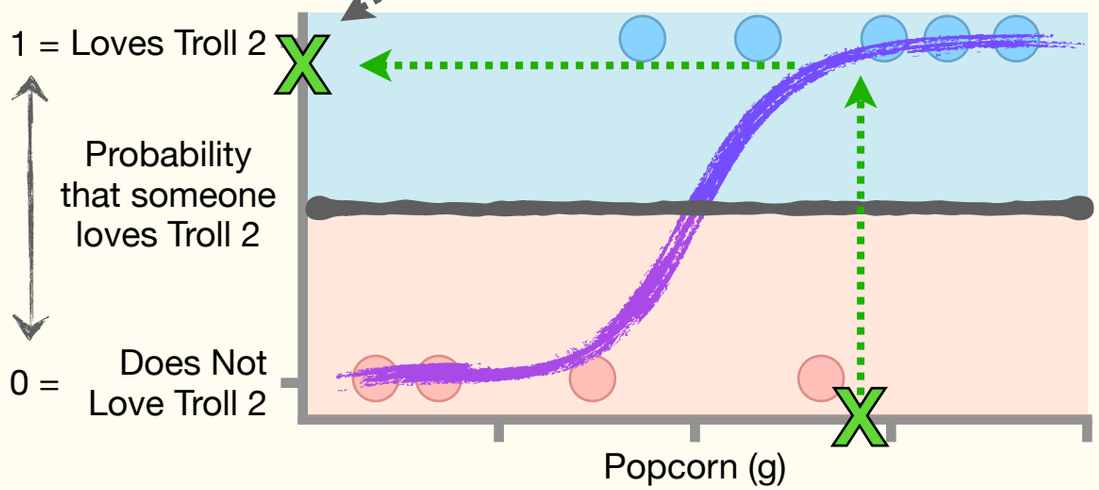
5 Now that we know the *probability* that this person will love Troll 2, we can *classify* them as someone who either **Loves Troll 2** or **Does Not Love Troll 2**. Usually the threshold for classification is **0.5**…



…and in this case, that means anyone with a probability of loving Troll 2 **> 0.5** will be classified as someone who **Loves Troll 2**…

….and anyone with a probability **≤ 0.5** will be classified as someone who **Does Not Love Troll 2**.

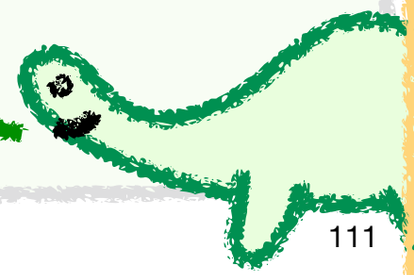
6 Thus, in this example, since **0.96 > 0.5**, we'll classify this person as someone who **Loves Troll 2**.



7 One last thing before we go: In this example, the classification threshold was **50%**. However, when we talk about **Receiver Operator Curves** (**ROCs**) in **Chapter 8**, we'll see examples that use different classification thresholds. So get excited!!!

TRIPLE BAM!!!

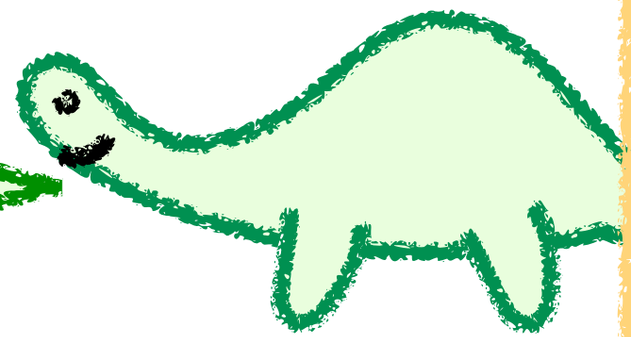
In a few pages, we'll talk about how we fit a **squiggle** to **Training Data**. However, before we do that, we need to learn some fancy terminology.



Terminology Alert!!! Probability vs. Likelihood: Part 1

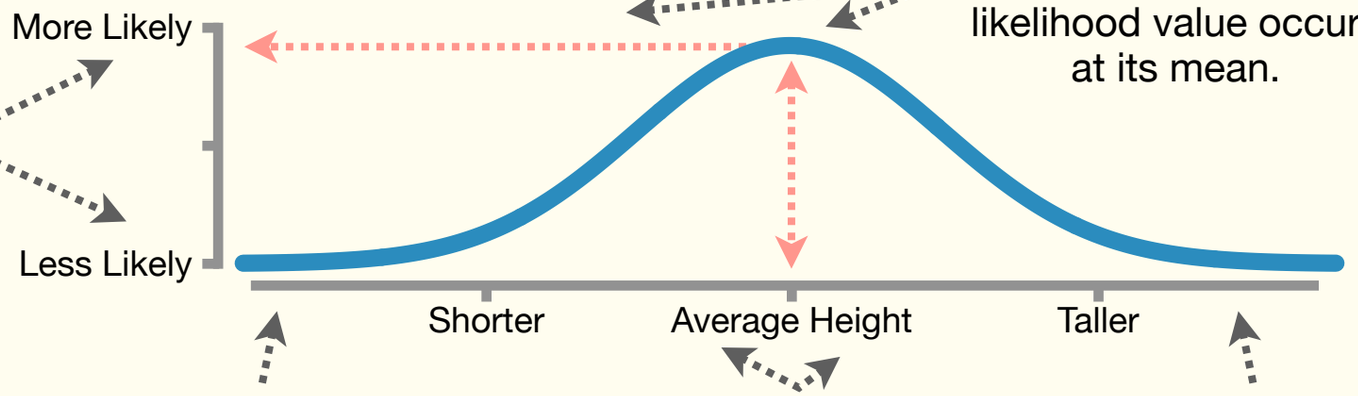
Hey **Norm**, don't the words **Probability** and **Likelihood** mean the same thing?

In casual conversations, yes, we can use **Probability** and **Likelihood** interchangeably. But unfortunately, in the context of statistics, these terms have different uses and, in some cases, completely different meanings.



1 Way back in **Chapter 3**, when we described the **Normal Distribution**, we saw that the y-axis represented **Likelihood**.

In this specific example, the y-axis represents the **Likelihood** of observing any specific Height.



The **Normal Distribution**'s maximum likelihood value occurs at its mean.

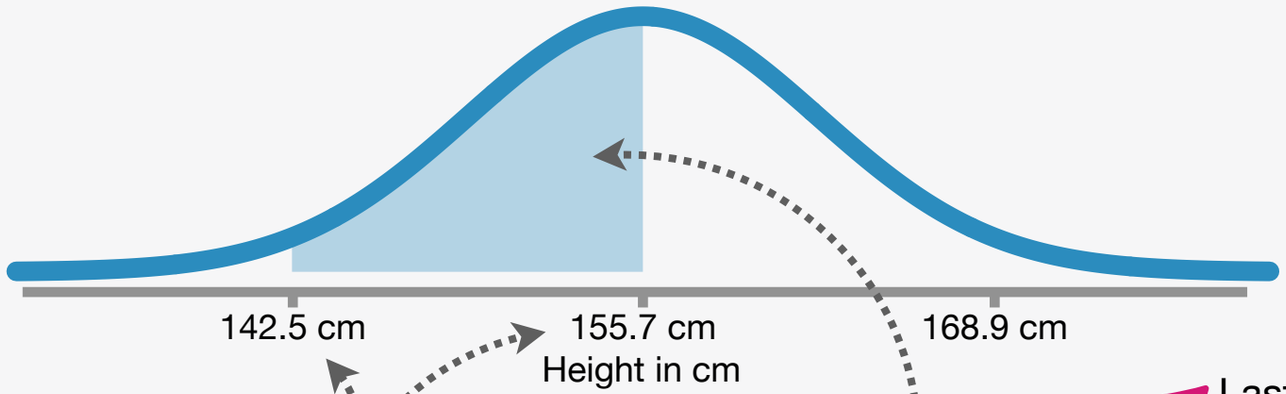
For example, it's relatively rare to see someone who is super short…

…relatively common to see someone who is close to the average height…

…and relatively rare to see someone who is super tall.

Terminology Alert!!! Probability vs. Likelihood: Part 2

2 In contrast, later in **Chapter 3**, we saw that **Probabilities** are derived from a **Normal Distribution** by calculating the **area under the curve** between two points.



For example, given this **Normal Distribution** with **mean = 155.7** and **standard deviation = 6.6**, the probability of getting a measurement between **142.5** and **155.7** cm…

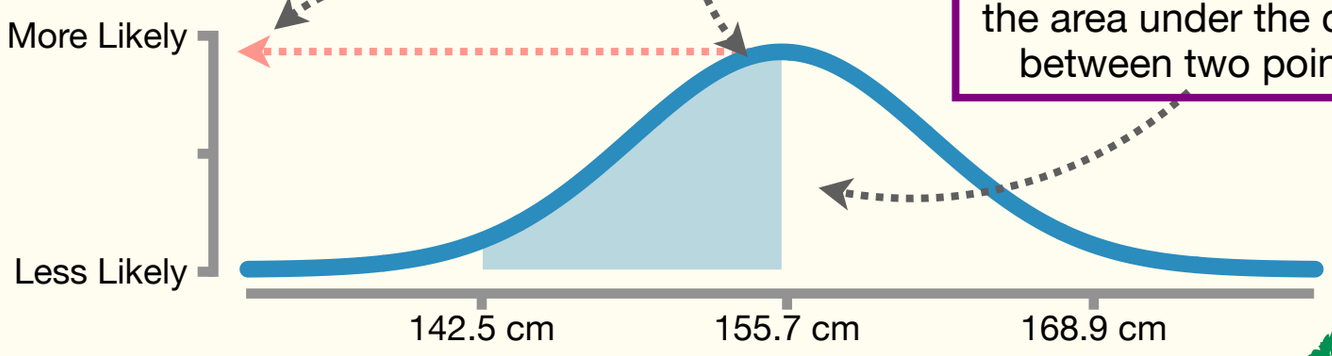
…is equal to this area under the curve, which, in this example, is **0.48**. So, the probability is **0.48** that we will measure someone in this range.

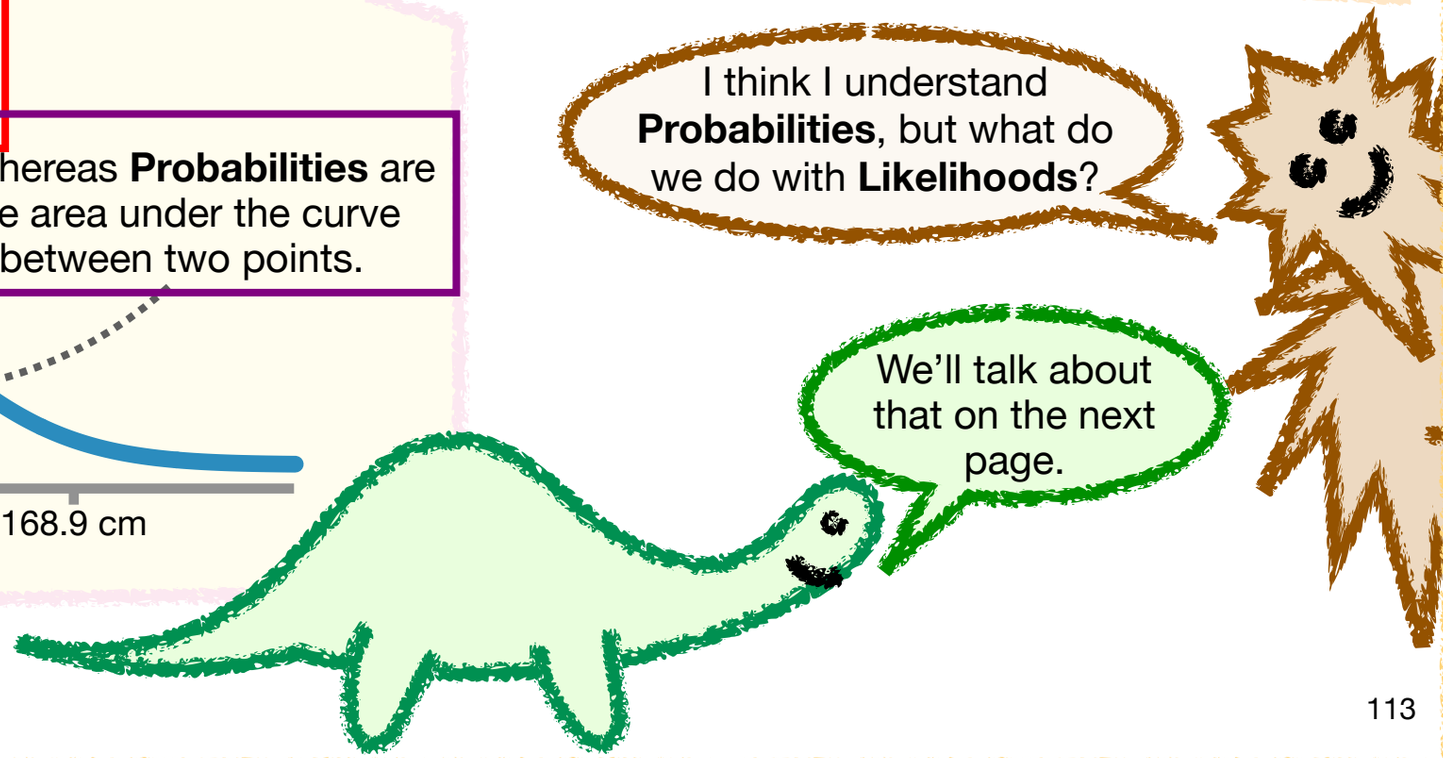
Lastly, in **Chapter 3** we mentioned that when we use a **Continuous Distribution**, like the **Normal Distribution**, the probability of getting any specific measurement is always **0** because the area of something with no width is **0**.

3 So, in the case of the **Normal Distribution**…

…**Likelihoods** are the y-axis coordinates for specific points on the curve…

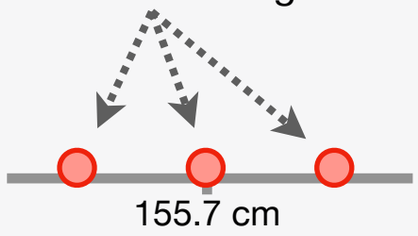
…whereas **Probabilities** are the area under the curve between two points.



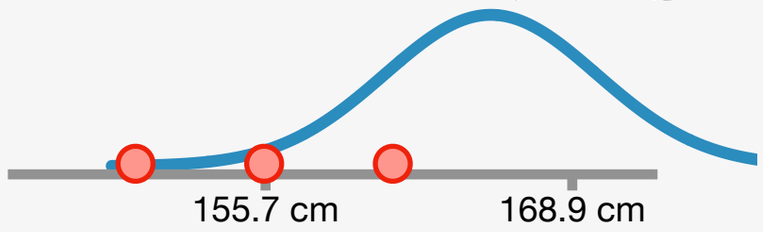


Terminology Alert!!! Probability vs. Likelihood: Part 3

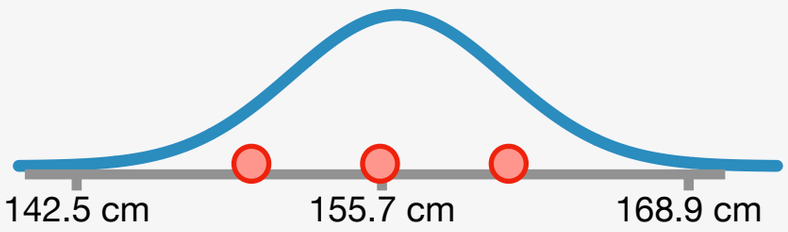
4 **Likelihoods** are often used to evaluate how well a statistical distribution fits a dataset. For example, imagine we collected these three Height measurements…



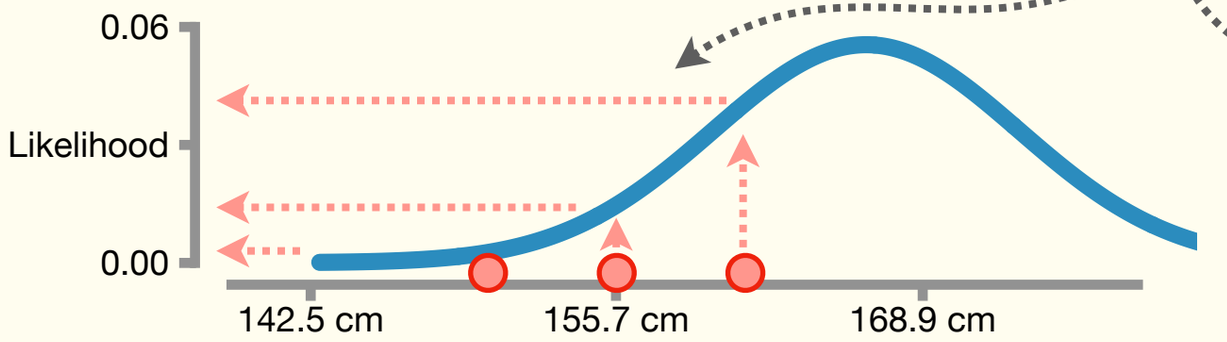
…and we wanted to compare the fit of a **Normal Curve** that has its peak to the right of the data…



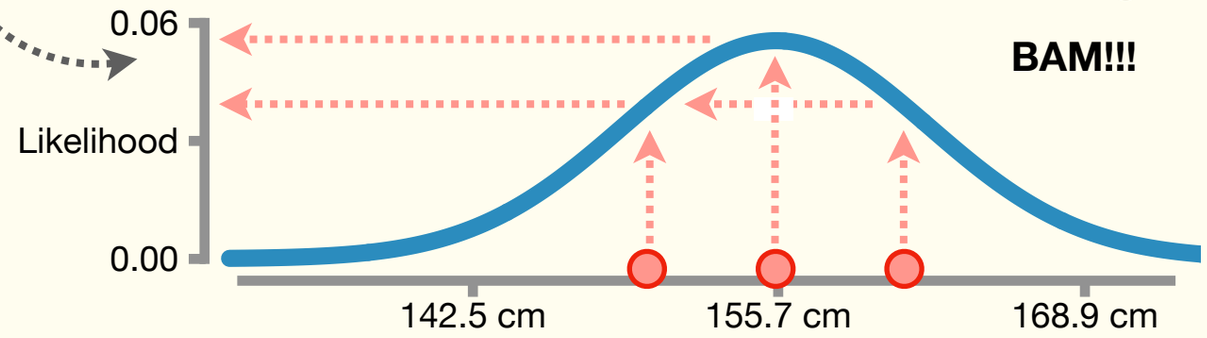
…to the fit of a **Normal Curve** that is centered over the data.



5 First, we determine the **Likelihoods**, the y-axis coordinates on the curves, for each data point…



…and, by eye, we can see that, overall, the **Likelihoods** are larger when we center the curve over the data.



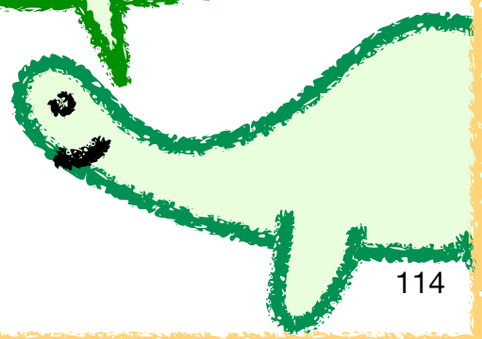
And larger likelihoods suggest that the centered curve fits the data better than the one shifted to the right.

BAM!!!

6 **NOTE:** When we try to fit a **Normal Curve** to data, we can't use **Probabilities** because, as we saw in **Chapter 3**, the probability for a specific point under a **Normal Curve** is always **0**.

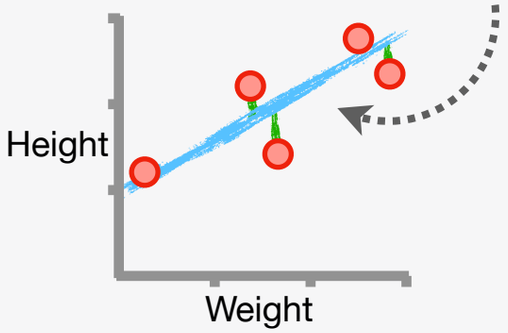
7 Lastly, as we just saw with the **Normal Distribution**, **Probabilities** and **Likelihoods** can be different. However, as we'll soon see, this is not always the case. In other words, sometimes **Probabilities** and **Likelihoods** are the same. When **Probabilities** and **Likelihoods** are the same, we could use either **Probabilities** or **Likelihoods** to fit a curve or a squiggle to data. However, to make the terminology consistent, when we're fitting a curve or squiggle to data in a statistical context, we almost always use **Likelihoods**.

Now that we know how to use **Likelihoods** to fit curves, let's talk about the main ideas of fitting a squiggle to data for **Logistic Regression**.

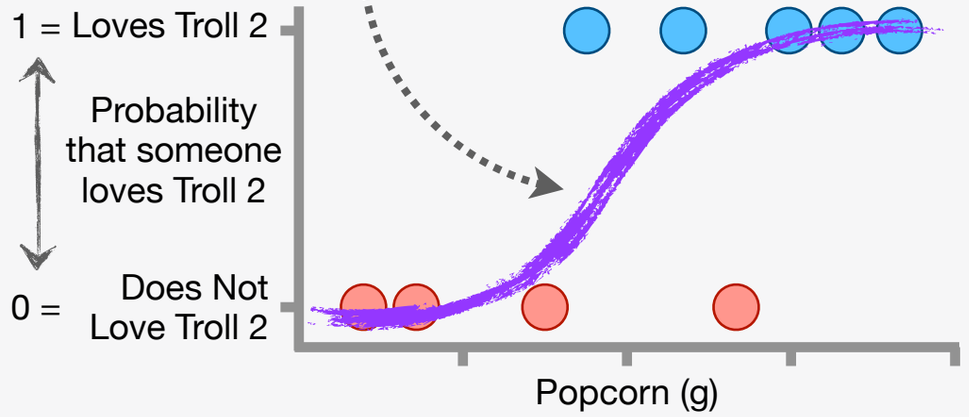


Fitting A Squiggle To Data: Main Ideas Part 1

1 When we use **Linear Regression**, we fit a **line** to the data by minimizing the **Sum of the Squared Residuals** (**SSR**).



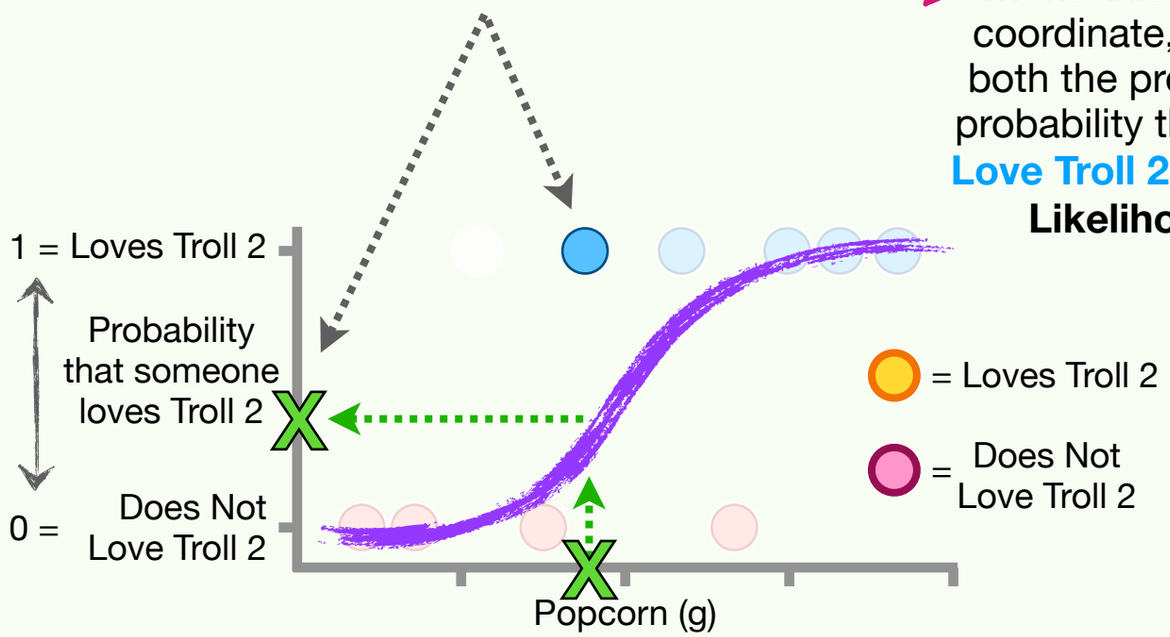
In contrast, **Logistic Regression** swaps out the **Residuals** for **Likelihoods** (y-axis coordinates) and fits a **squiggle** that represents the **Maximum Likelihood**.



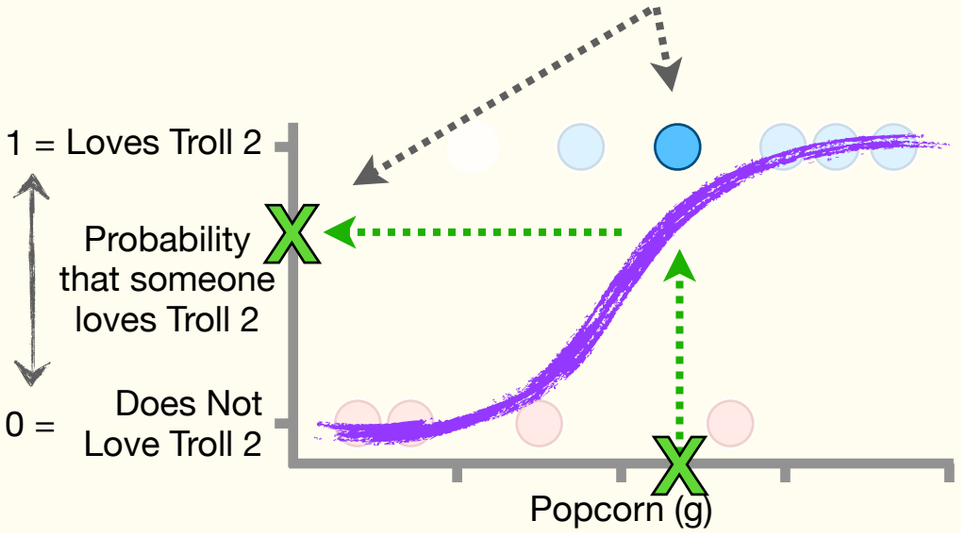
2 However, because we have two classes of people, one that **Loves Troll 2** and one that **Does Not Love Troll 2**, there are two ways to calculate **Likelihoods**, one for each class.

3 For example, to calculate the **Likelihood** for this person, who **Loves Troll 2**, we use the **squiggle** to find the y-axis coordinate that corresponds to the amount of Popcorn they ate…

…and this y-axis coordinate, **0.4**, is both the predicted probability that they **Love Troll 2** and the **Likelihood**.

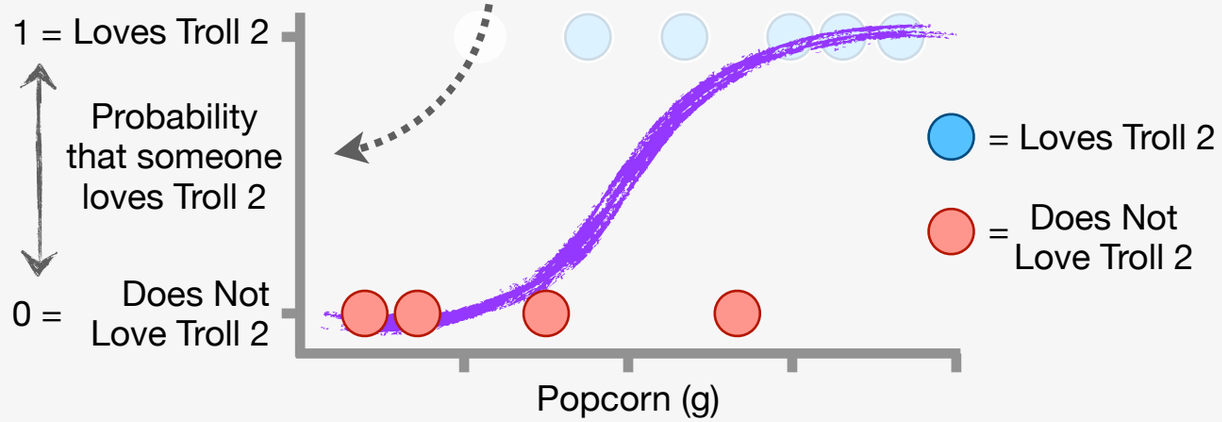


4 Likewise, the **Likelihood** for this person, who also **Loves Troll 2**, is the y-axis coordinate for the **squiggle**, **0.6**, that corresponds to the amount of Popcorn they ate.



Fitting A Squiggle To Data: Main Ideas Part 2

5 In contrast, calculating the **Likelihoods** is different for the people who **Do Not Love Troll 2** because the y-axis is the probability that they **Love Troll 2**.



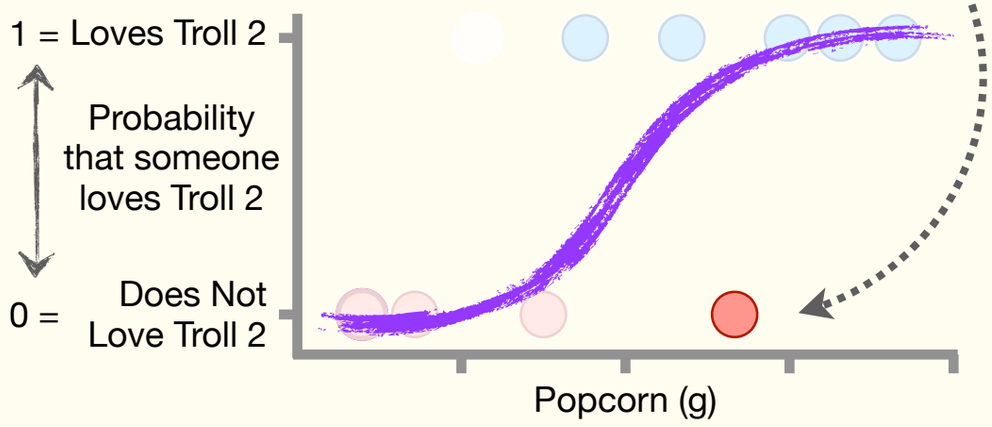
The good news is, because someone either loves Troll 2 or they don't, the probability that someone does not love Troll 2 is just **1** minus the probability that they love Troll 2…

p(**Does Not Love Troll 2**) = 1 - p(**Loves Troll 2**)

…and since the y-axis is both probability and likelihood, we can calculate the **Likelihoods** with this equation.

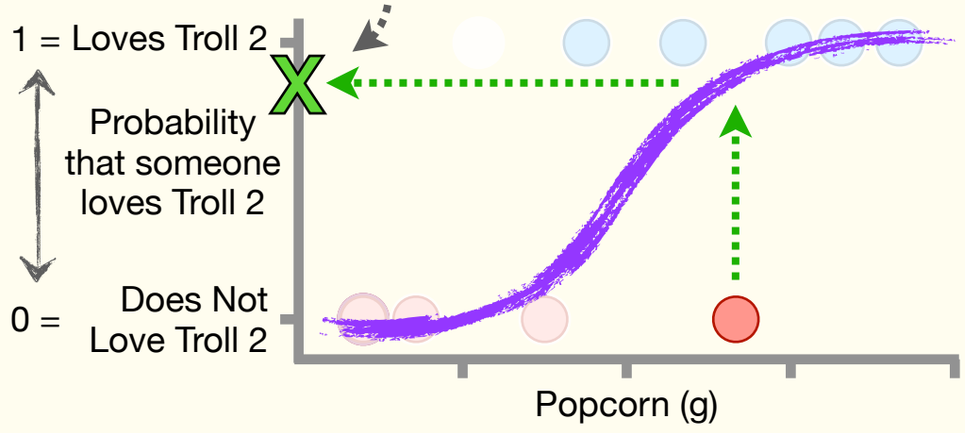
L(**Does Not Love Troll 2**) = 1 - L(**Loves Troll 2**)

6 For example, to calculate the **Likelihood** for this person, who **Does Not Love Troll 2**…



…we first calculate the **Likelihood** that they **Love Troll 2**, **0.8**…

…and then use that value to calculate the **Likelihood** that they **Do Not Love Troll 2** = **1 - 0.8 = 0.2**.

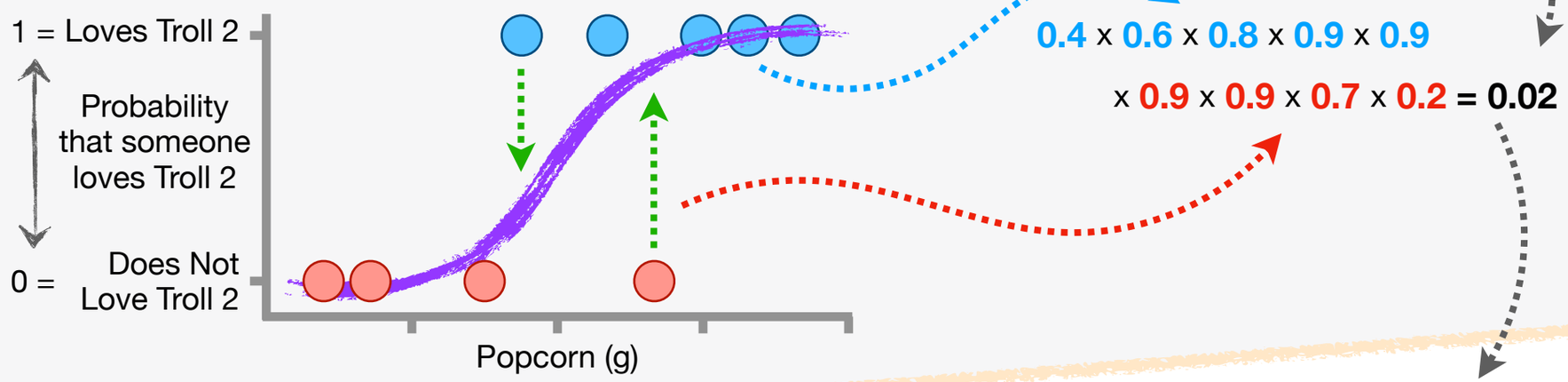


Bam!

Fitting A Squiggle To Data: Main Ideas Part 3

7 Now that we know how to calculate the **Likelihoods** for people who **Love Troll 2** and people who **Do Not Love Troll 2**, we can calculate the **Likelihood** for the entire **squiggle** by multiplying the individual **Likelihoods** together…

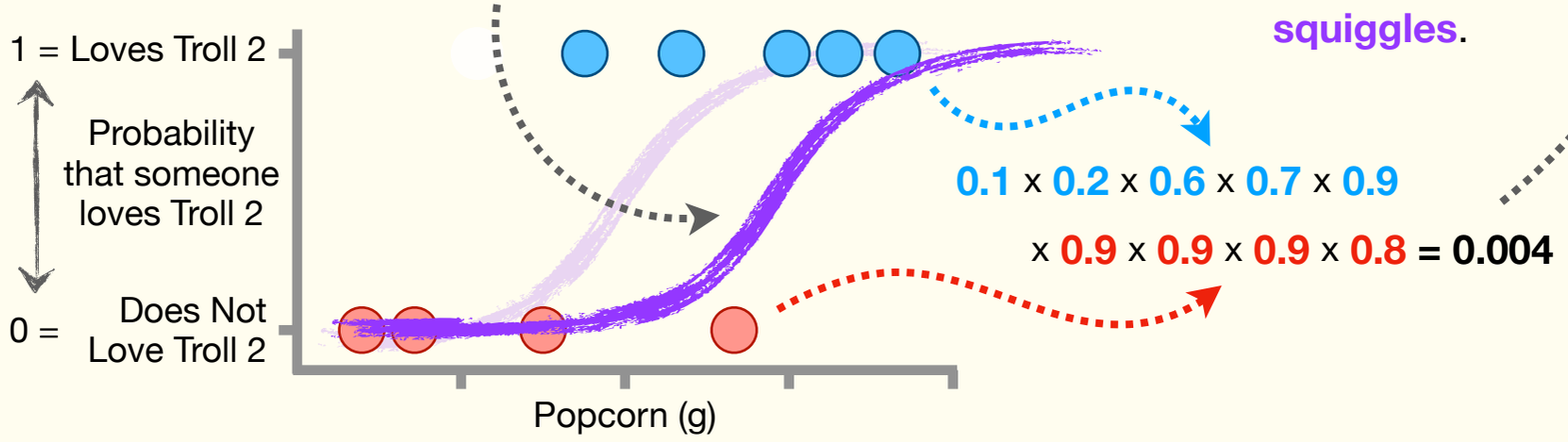
…and when we do the math, we get **0.02**.



VS.

8 Now we calculate the **Likelihood** for a different **squiggle**…

…and compare the total **Likelihoods** for both **squiggles**.



9 The goal is to find the **squiggle** with the **Maximum Likelihood**.

In practice, we usually find the optimal **squiggle** using **Gradient Descent**.

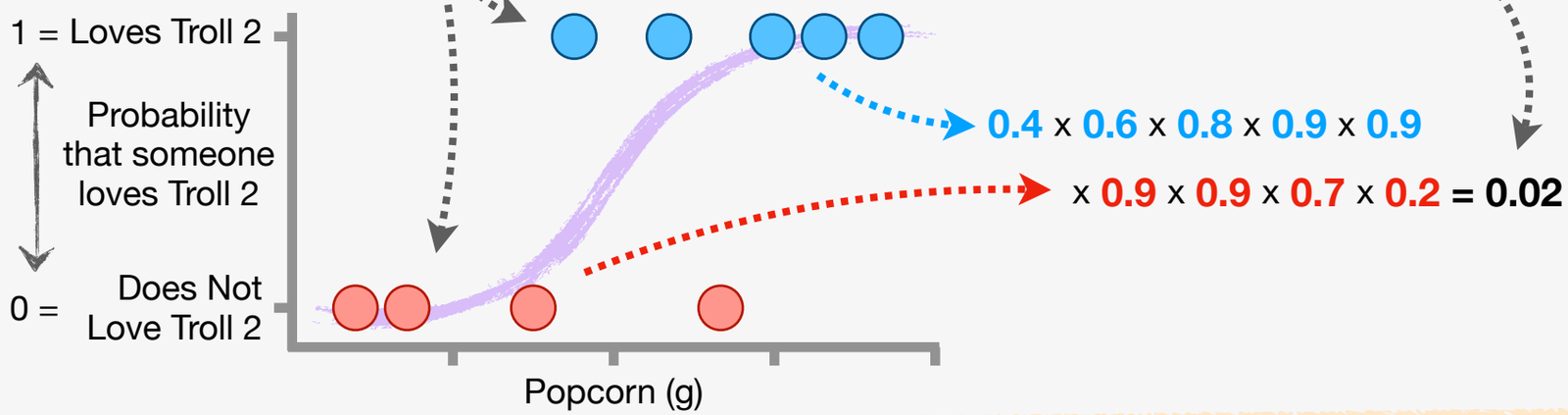
TRIPLE BAM!!!

Fitting A Squiggle To Data: Details

If you'd like to learn more details about **Logistic Regression**, scan, click, or tap this QR code to check out the **'Quests** on YouTube!!!

1 The example we've used so far has a relatively small **Training Dataset** of only **9** points total…

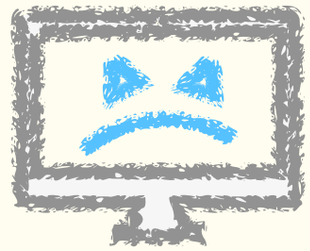
…so when we multiplied the **9** **Likelihoods** together, it was easy, and we got **0.02**.



0.4 x **0.6** x **0.8** x **0.9** x **0.9**

x **0.9** x **0.9** x **0.7** x **0.2** = **0.02**

2 However, if the **Training Dataset** was much larger, then we might run into a computational problem, called **Underflow**, that happens when you try to multiply a lot of small numbers between **0** and **1**.



Technically, **Underflow** happens when a mathematical operation, like multiplication, results in a number that's smaller than the computer is capable of storing.

Underflow can result in errors, which are bad, or it can result in weird, unpredictable results, which are worse.

3 A very common way to avoid **Underflow** errors is to just take the **log** (usually the **natural log**, or **log base *e***), which turns the multiplication into addition…

log(0.4 x 0.6 x 0.8 x 0.9 x 0.9 x 0.9 x 0.9 x 0.7 x 0.2)

= log(0.4) + log(0.6) + log(0.8) + log(0.9) + log(0.9)

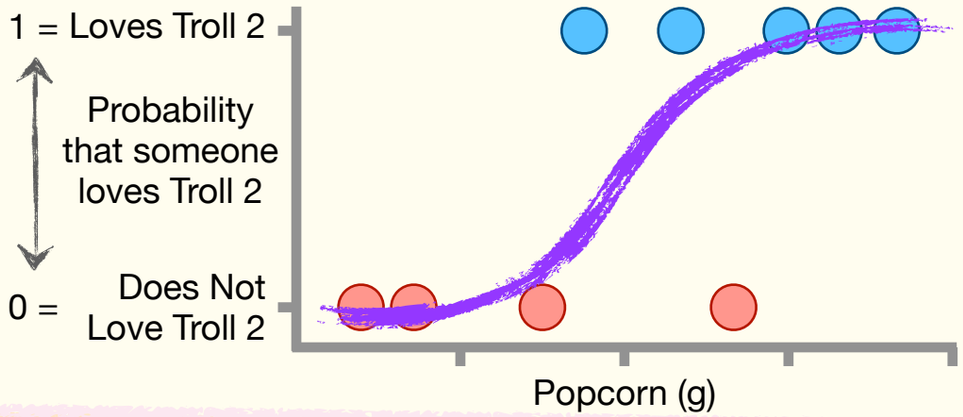
+ log(0.9) + log(0.9) + log(0.7) + log(0.2)

= -4.0

…and, ultimately, turns a number that was relatively close to **0**, like **0.02**, into a number relatively far from **0**, **-4.0**.

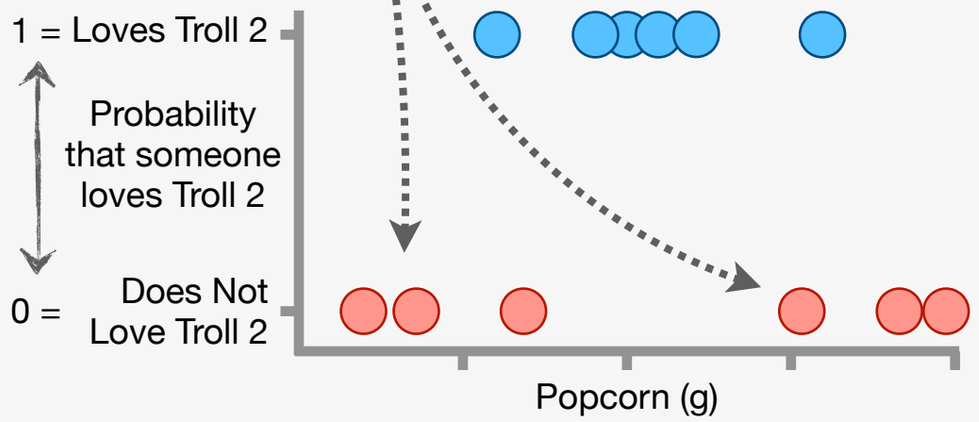
Logistic Regression: Weaknesses

1 When we use **Logistic Regression**, we assume that an **s-shaped squiggle** (or, if we have more than one independent variable, an s-shaped surface) will fit the data well. In other words, we assume that there's a relatively straightforward relationship between the Popcorn and Loves Troll 2 variables: if someone eats very little Popcorn, then there's a relatively low probability that they love Troll 2, and if someone eats a lot of Popcorn, then there's a relatively high probability that they love Troll 2.

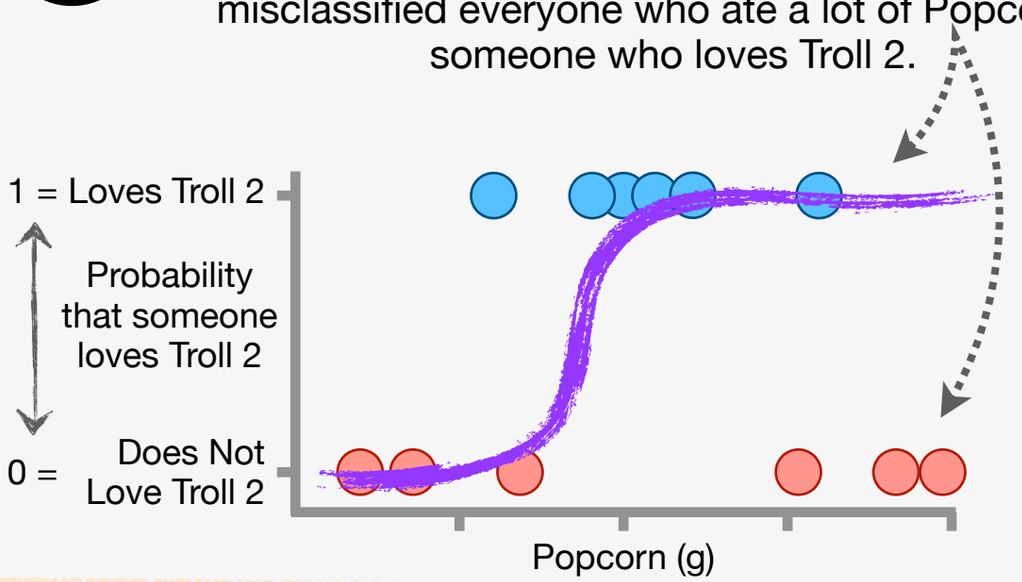


2 In contrast, if our data had people who ate a little or a lot of Popcorn and do not love Troll 2…

…and people who ate an intermediate amount did not, then we would violate the assumption that **Logistic Regression** makes about the data.



3 And if we used **Logistic Regression** to fit an **s-shaped squiggle** to the data, we would get a horrible model that misclassified everyone who ate a lot of Popcorn as someone who loves Troll 2.



4 Thus, one of the limitations of **Logistic Regression** is that it assumes that we can fit an **s-shaped squiggle** to the data. If that's not a valid assumption, then we need a **Decision Tree** (**Chapter 10**), or a **Support Vector Machine** (**Chapter 11**), or a **Neural Network** (**Chapter 12**), or some other method that can handle more complicated relationships among the data.