



Probability Review

Agenda - Schedule

1. Inferential Statistics
2. Probability Review
3. Probability in Data Analysis
4. Break
5. TLAB

$$\text{Odds} = \frac{\text{4 blue circles}}{\text{6 yellow circles}}$$

$$\text{Probability} = \frac{\text{4 blue circles}}{\text{10 total circles (4 blue + 6 yellow)}}$$

[Odds vs probabilities](#)



Agenda - Announcements

- Week 5 Pre-Class Quiz due 4/8 (TO BE POSTED)
- TLAB #2 Due 4/21
- Cohort A
 - Still have office hours available!



Agenda - Goals

- Review basic probability
- Understand how sample calculations differ from population
- Understand & apply Bayes Theorem

Inferring from a Sample



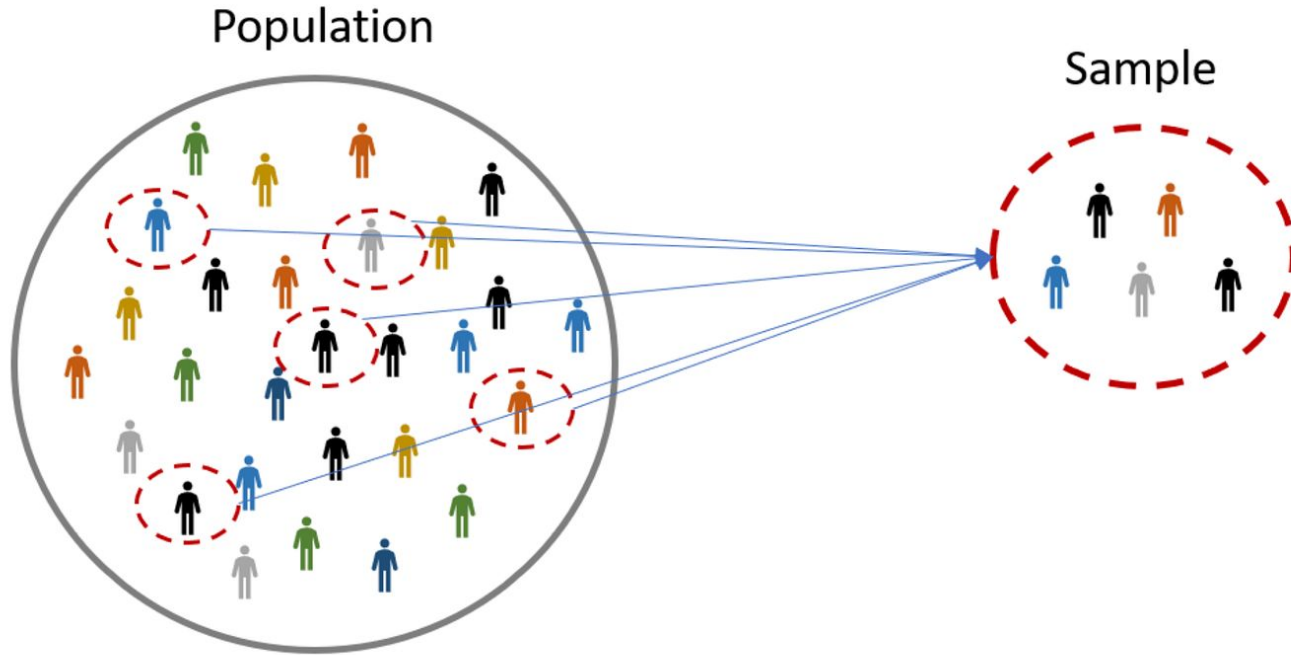
Inferential Statistics

Now that we have a good understanding of basic descriptive statistics, we need to turn our attention to **inferential statistics**.

This entails the mathematics of **inferring** outcomes/relationships/patterns on a **population** using a **sample**.

This is also the heart of **machine learning**:

- **Solving self-driving** by training ml algorithms on a subset of driving videos
- **Recognizing dogs** by training ml algorithms on a subset of dog photos



We often use humans when discussing population vs sample, but this applies to **any dataset**. Ex: All AAPL stock prices since Dec 14, 1984 (**population**) & AAPL stock price on Dec 12, 2024 (**sample**).

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$ <p> σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size </p>	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ <p> s^2 = sample variance x_i = value of i^{th} element \bar{x} = sample mean n = sample size </p>

Our equations for measures of dispersion change based on which subset we are calculating our metrics on. This is **because a sample is an estimate**, and a **population is ground truth**.

Population vs Sample

The reasoning is as follows:

Sample variance is a biased estimator (i.e. *sample variance will skew this measure based on what the sample mean is*)

In order to remove this bias we need to divide our calculation not necessarily by **N** (*the size of sample*), but rather by the **samples' degrees of freedom**.

A **degree of freedom** is defined as the number of independent values that affect the measure. For our sample variance this is **N-1**. **But why?**

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

**Estimated population
variance**

Population vs Sample

When we use sample mean in our variance, the last value of our mean calculation is entirely dependent on the previous values.

For example, let's say that we are calculating variance on the dataset to the right.

Considering that we know the sample mean of our dataset (80), the last value that we add to our sample mean is entirely dependent on the other values.

It is impossible for the last value to be anything other than 100.

73	73	76	77	81	100
----	----	----	----	----	-----

$$\begin{aligned}\text{Mean (average)} &= \frac{\text{Sum}}{\text{Count}} \\ &= \frac{73 + 73 + 76 + 77 + 81 + 100}{6} \\ &= \frac{480}{6} \\ &= 80\end{aligned}$$

How many truly independent values are in this calculation?



Population vs Sample

Therefore, when we include the degrees of freedom in our variance, we must subtract one value off of our **N**.

This is because all values except the last value in our sample are independent degrees of freedom (**N-1**).

This also has the added benefit of giving us a more accurate estimate of the population variance.

This is the heart of this discussion: sample calculations are estimates.

Population calculations are ground truth values.

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

Estimated Population
Variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

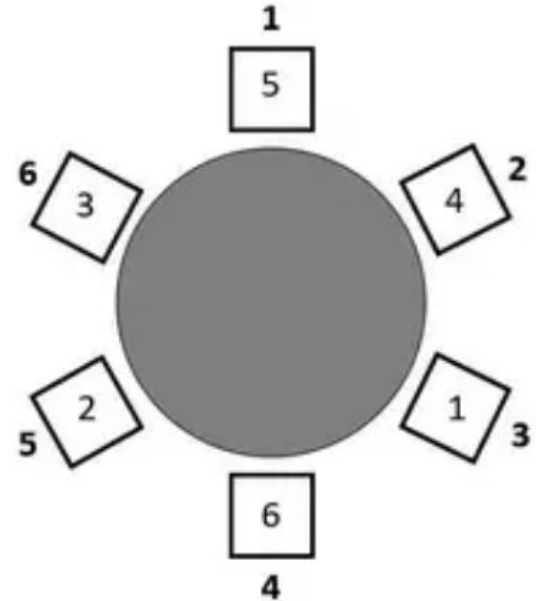
Population Variance

Chair Metaphor

Think of it another way: assume we have 30 chairs (pieces of data), after we allow 29 people to randomly sit (random selection) then the last person (last piece of data) **MUST** be in the 30th chair.

This person is completely dependent on where everyone else sat. **They are no longer independent!**

Therefore given a calculated mean, that means 29 of our numbers could have been anything but the last one must be fixed (based on the other 29).



Population vs Sample

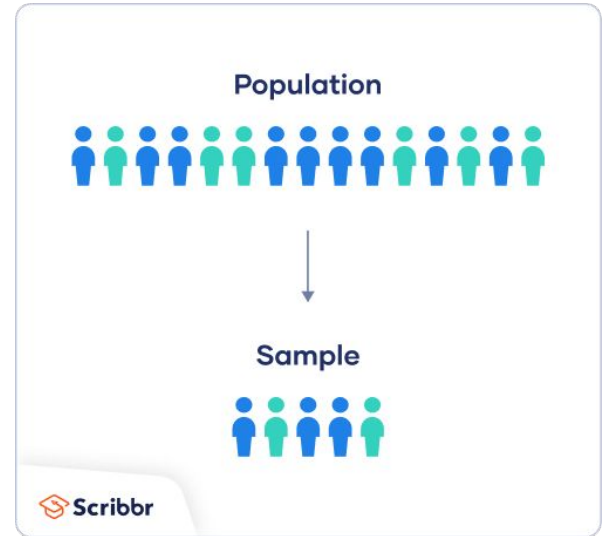
Now that we understand that our sample is just an **estimate**, the next question you should be asking yourself is:

*How **good** of an estimate is our sample of the population?*

In order to answer this question, we must explore another field of maths which allow us to make predictions such as:

- Computer Vision
- Probabilistic Language Models
- Self-driving cars

This entail **probability**



Probability Review



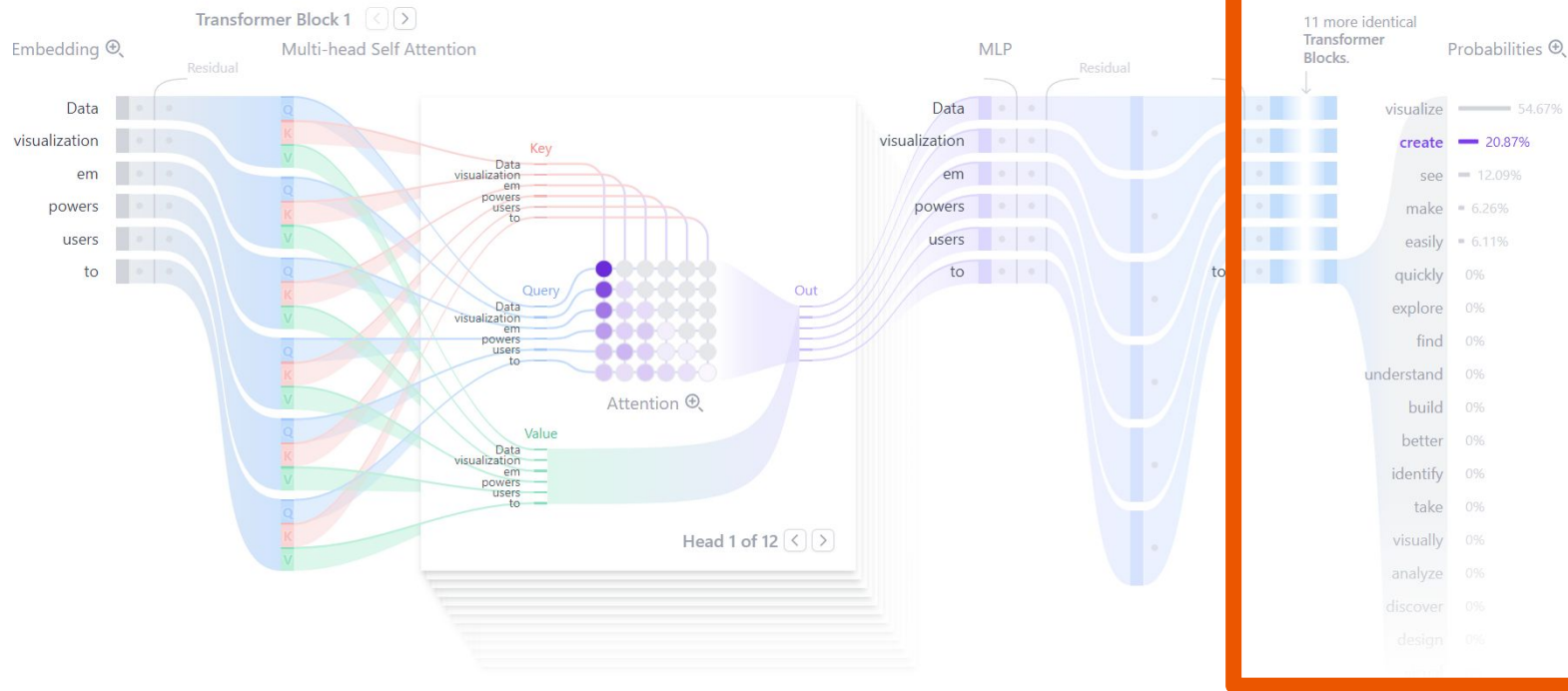
Probability

Yesterday, we discussed the usage of the OpenAI API to predict responses for text data.

One key component that we mentioned is that ChatGPT is a **probabilistic** model which *predicts* the next word to use in a response using historical data.

A good chunk of the ML models that we will be using also estimate probabilities.

Therefore, let's get reacquainted with the laws of probability.



While this architecture is fairly complex, note the output of this model is just a **probability distribution** which ChatGPT will use to select the most likely word.



Probability

In probability, we always start with a known model of the world, but we do not have the data.

This is different from statistics, where we have the data, but we do not know the model of the world.

The most basic “model” is expressed using the following notation.

$$P(\text{heads}) = 0.5$$

The probability of heads is 50%

We always express probability as a number between 0 and 1.

I know what my initial assumptions are, but what are the chances I will get 3 heads in a row out of 5 flips?



Probability Formulas

Let's go over a **few definitions** and useful **frameworks** which we will eventually use when we get to the mathematics of machine learning.

First we have an **elementary event**. These are one of the outcomes we are guaranteed to have in our **sample space**. The sample space is the set of all possible events:

- Flipping heads on a quarter is an **elementary event**, heads and tails are **the sample space**.
- Rolling 1 on a die is an **elementary event**, rolling 1,2,3,4,5, or 6 is the **sample space**.



Probability Formulas

The probability of the elementary event is quite **simply the ratio** of a singular **event occurring in the sample space**.

$$P(\text{heads}) = \frac{\text{Number of Events where we flip Heads}}{\text{Total Events}} = \frac{\text{Sample Space}}{H \ T}$$



Probability Formulas

The probability of the elementary event is quite **simply the ratio** of a singular **event occurring in the sample space**.

$$P(\text{heads}) = \frac{1}{2} \quad \frac{\text{Sample Space}}{H T}$$



Probability Formulas

The probability of the elementary event is quite **simply the ratio** of a singular **event occurring in the sample space**.

$P(\text{heads}) =$

0.5

Sample Space

H T



Probability Formulas

While this is a simple discussion, we can extract a few more definitions out of this example:

$$P(X) = 1$$

The event will always occur.

$$P(X) = 0$$

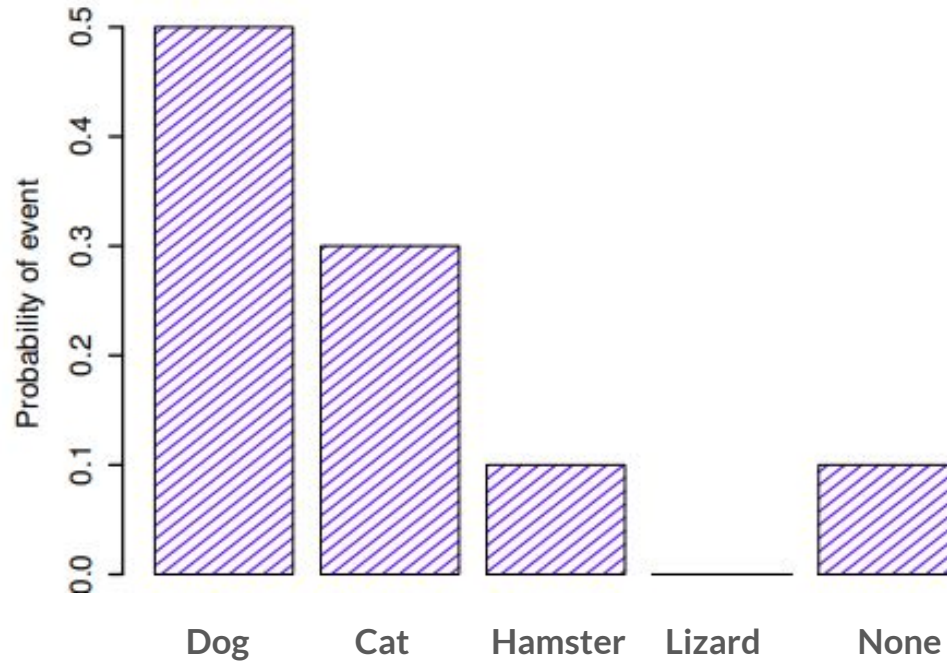
The event will never occur.

Law of total probability

All probabilities of events must add up to 1.

Event	Probability
Dog	0.5
Cat	0.3
Hamster	0.1
Lizard	0.0
No pet	0.1

Let's increase our sample space in order to discuss more complex terms. Here we describe the probabilities that an american household has some specific pet.



Graphing these values gives us the **probability distribution**.



Probability Formulas

Using this sample space we can go on to also discuss **non-elementary events**.

These entail **subsets of events**. For example, if we ask ourselves “*What is the probability that someone owns a Dog or Cat*” we can express this event as **E**, which entails

$$P(E) = P(\text{Dog}) + P(\text{Cat})$$

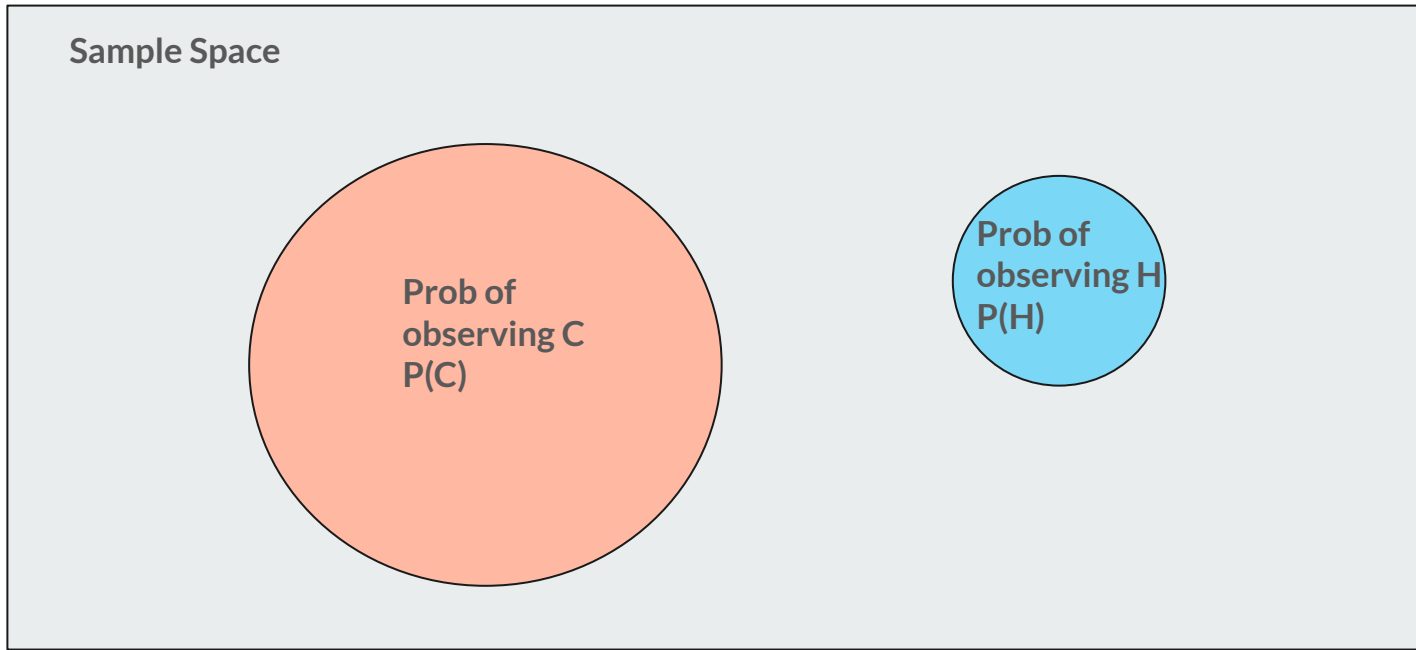
Event	Probability
Dog	0.5
Cat	0.3
Hamster	0.1
Lizard	0.0
No pet	0.1

English	Notation	Formula
not A	$P(\neg A)$	$= 1 - P(A)$
A or B	$P(A \cup B)$	$= P(A) + P(B) - P(A \cap B)$
A and B	$P(A \cap B)$	$= P(A B)P(B)$

Lastly, we also discuss a few rules that our probabilistic events must abide by. This entails **not**, **or**, and **and**.

One key factor to note is that these events that we currently discussing are **independent events**. (The outcomes of prior events does not impact future events)

In later lectures, we will discuss **dependent events** which entail more complex formulation and reasoning.



A very useful visual to understand how to calculate probability entails the **venn diagram**. As we are discussing independent events, we can rest assured that our events will never intersect.

Probability in Data Analysis - Bayes Theorem



Probability in Data Analysis

We can also use probability to make inferences about our data.

A couple of examples include:

- If someone made a purchase **yesterday**, what is the likelihood that they make a purchase **today**?
- If someone listened to **song X** and **song Y** what is the likelihood they will listen to **song Z**?

In order to do this, we need to discuss an alternate **branch of statistics**.



Bayes Theorem - Frequentist World-View

So far, in our discussion of statistics and probability, we have assumed a **frequentist world-view**.

That is, we define probability as simply a long-run frequency. For example, do you want to figure out the probability of flipping heads on a quarter? **Simply run an experiment!**

number of flips	11	12	13	14	15	16	17	18	19	20
number of heads	8	8	9	10	10	10	10	10	10	11
proportion	.73	.67	.69	.71	.67	.63	.59	.56	.53	.55



Bayes Theorem - Frequentist Pros

We like this world-view because it is:

Objective:

we run experiments to figure out the world

Unambiguous:

you and I will calculate the same probability from one experiment



Bayes Theorem - Frequentist Cons

But, there are also negatives to this world-view:

No use of prior information:

when running our experiment, we close ourselves off from any prior information.

Inflexible:

Frequentists rely on large amounts of data to make statements on probability.

Bayes Theorem - Frequentist Cons

To further explain the limitations of the frequentist perspective, let's say you take a covid test. I tell you that there is a 80% chance that this test is reliable

From our previous perspective, how do we use this information? Well we basically chuck it in the trash.



Bayes Theorem - Bayesian Statistics

Enter Thomas Bayes.

He introduced probability as a **degree of belief** to which a “rational agent” assigns a truth to an event...

...as opposed to a ratio of experiments.

This **degree of belief** is updated with **new information**. Let's see an example...




Full-time minister, part-time statistician

$$P(H|E) = \frac{P(H) P(E|H)}{P(E)}$$

Here's the formula, let's **break down its components** before going through our covid example.



Remember, the “|” is the symbol for conditional probability statements.

“Hypothesis given Event”

$$P(H|E) = \frac{P(H) P(E|H)}{P(E)}$$


The probability the **hypothesis** is true given the **event** is equal to

Probability of the hypothesis
occurring (also known as the
prior belief!)


$$P(H|E) = \frac{P(H) P(E|H)}{P(E)}$$


The probability the **hypothesis** is true given the **event** is equal to


Probability of the hypothesis occurring (also known as the **prior belief!**)

Probability of seeing **event** given that the **hypothesis** is true (also known as the **likelihood***)

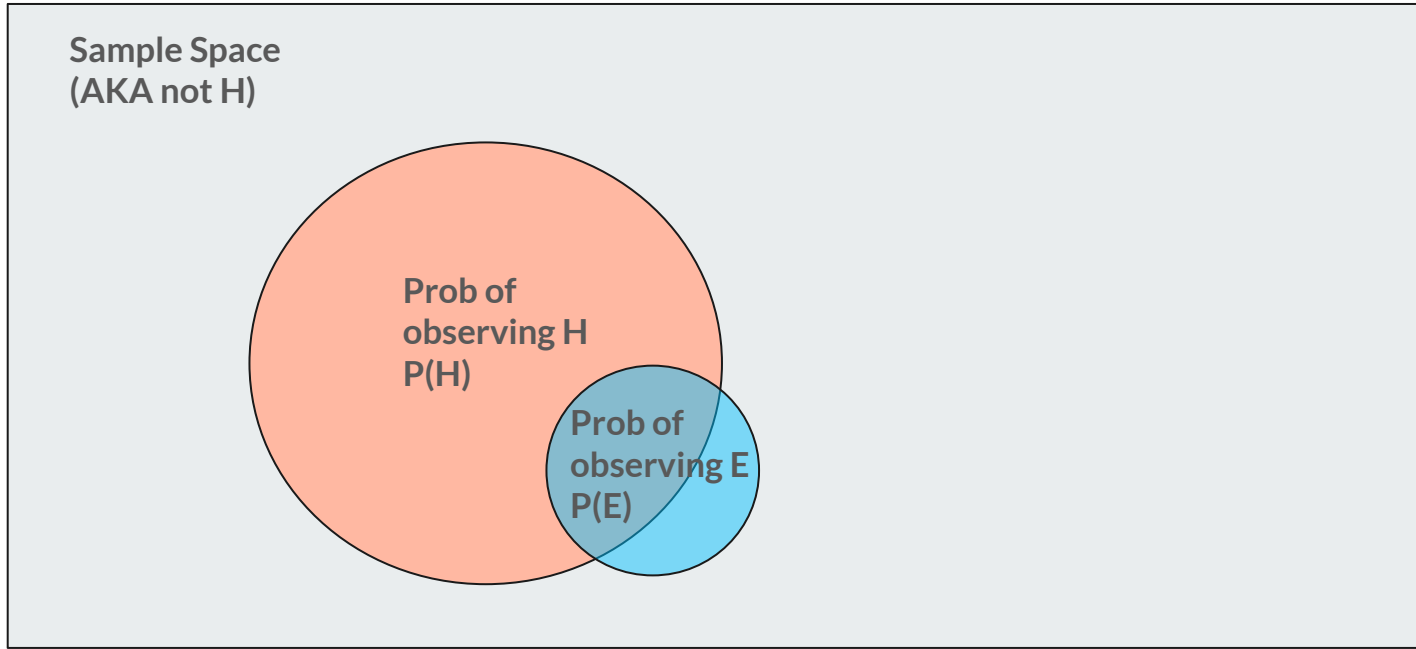
$$P(H|E) = \frac{P(H) P(E|H)}{P(E)}$$

The diagram shows the formula for Bayes' Theorem. The numerator consists of two terms, P(H) and P(E|H), separated by a space. A handwritten arrow points from the text 'Probability of the hypothesis occurring (also known as the prior belief!)' to P(H). Another handwritten arrow points from the text 'Probability of seeing event given that the hypothesis is true (also known as the likelihood*)' to P(E|H). The denominator is P(E). A handwritten arrow points from the text 'The probability the hypothesis is true given the event is equal to' to P(H|E).

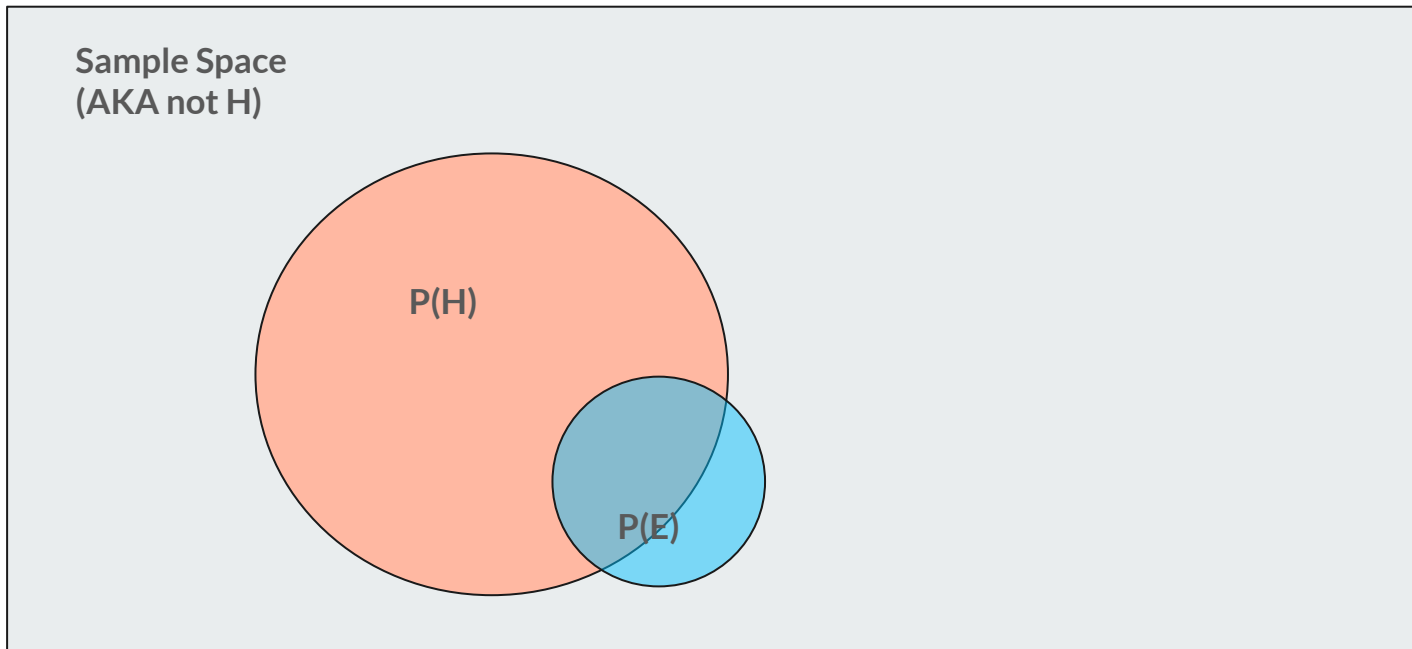
The probability the **hypothesis** is true given the **event** is equal to

$$P(H|E) = \frac{P(H) P(E|H)}{P(E)}$$


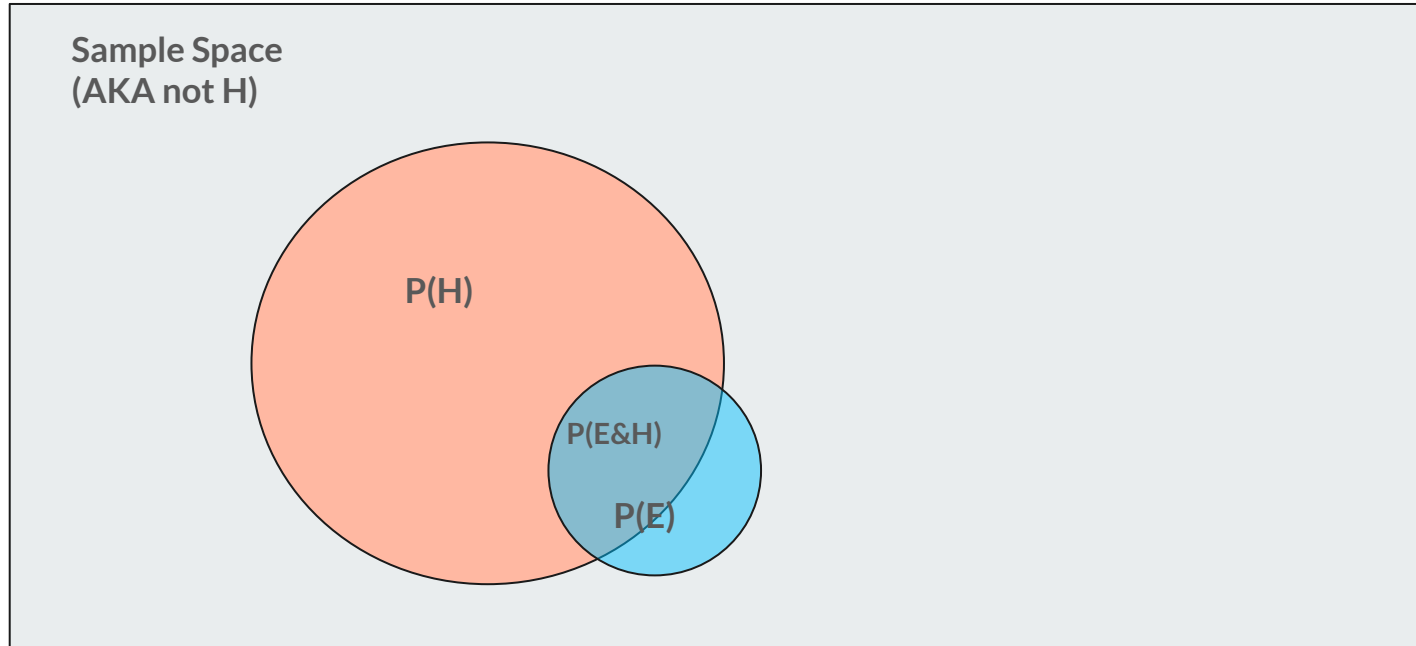
Divided by the probability of **event** occurring. This part is sometimes deceptively simple. Keep in mind that we want to consider the **event occurs given the hypothesis is true and given the hypothesis is false!**



Let's understand how to calculate this. I find diagrams to be the most helpful in visualizing this calculation

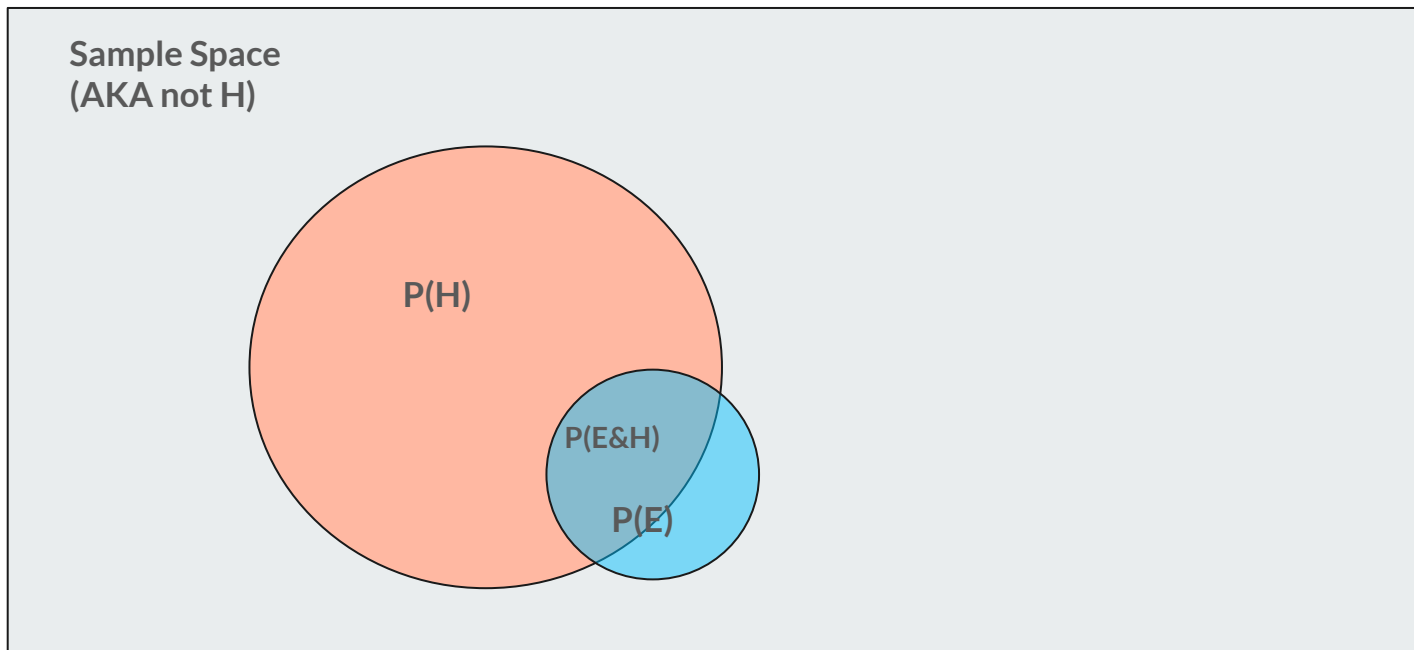


To calculate $P(E)$, we have to consider the intersection of $P(E)$ and $P(H)$, as well as the space where $P(E)$ exists outside of $P(H)$



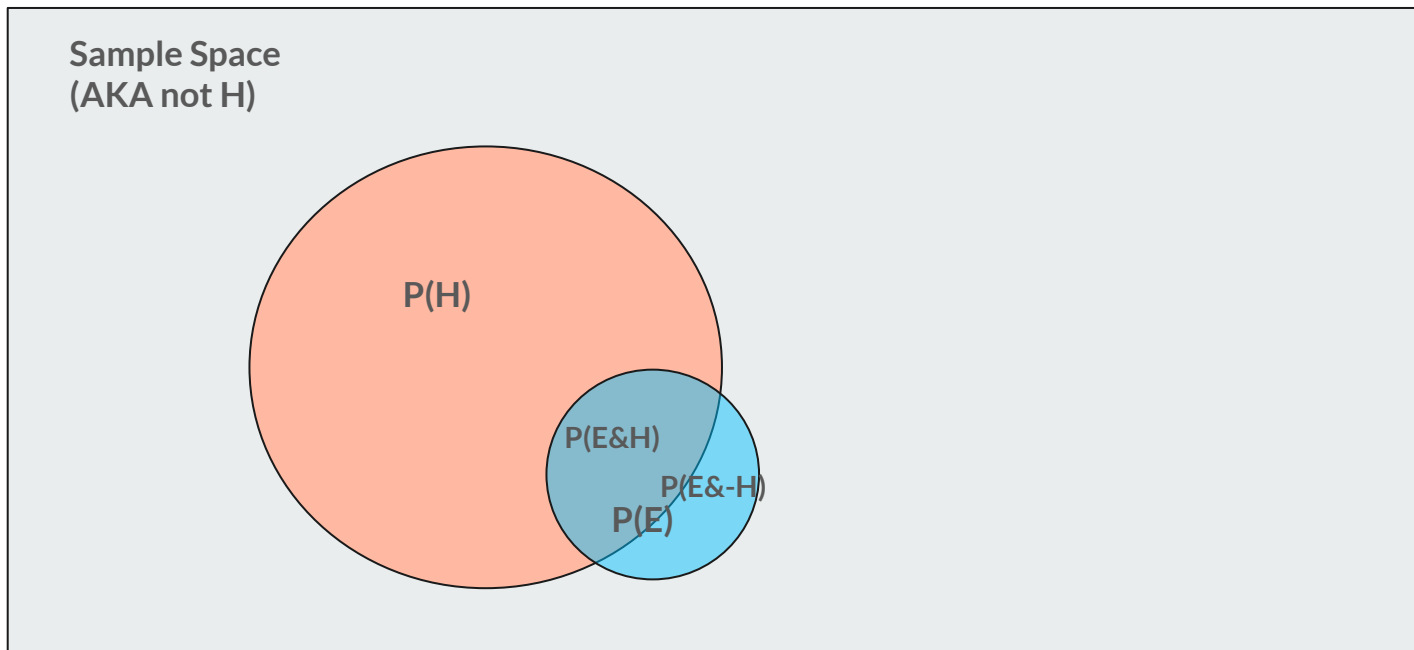
This intersect can be labeled **$P(E \& H)$**

$$P(H)P(E|H) +$$



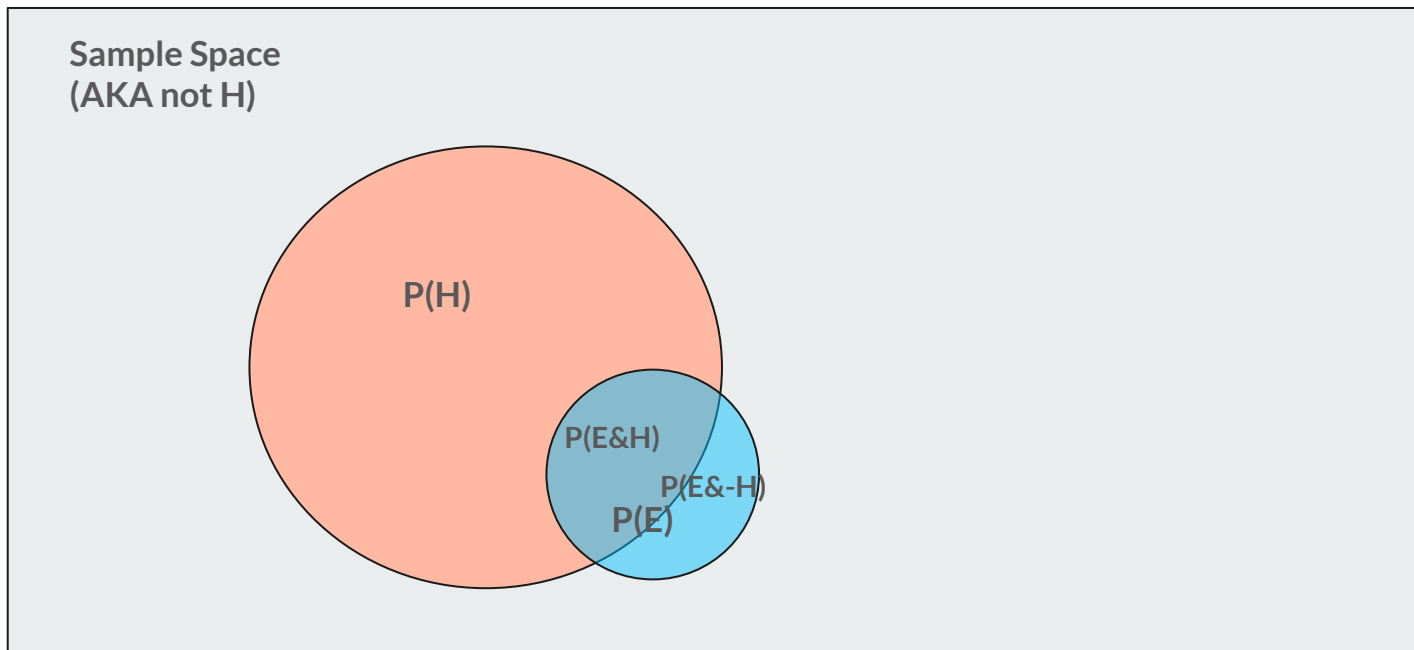
In terms of conditional probability, that is $P(H)P(E|H)$

$$P(H)P(E|H) +$$



And lastly we have $P(E \& -H)$. Can anyone express this in terms of conditional probability as well?

$$P(H)P(E|H) + P(-H)P(E|-H)$$



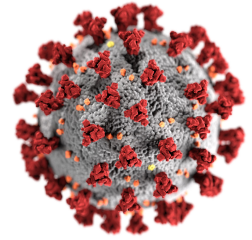
$P(-H)P(E|-H)$. This allows us to express $P(E)$ as $P(H)P(E|H) + P(-H)P(E|-H)$

$$P(H|E) = \frac{P(H) P(E|H)}{P(H)P(E|H) + P(-H)P(E|-H)}$$



This is expressed as the sum of the probability that the **hypothesis is true** and the **event occurs given the hypothesis is true**, and the **probability the hypothesis is false** and the **event occurs given the hypothesis is false**.

Disclaimer: These metrics are **fabricated** so don't use this for health decisions.



Covid Test Experiment

Let's say you take a covid test back in March 2021 with **90% reliability** (*i.e. it catches 90% of positive cases*) and it results in **positive**.

You might think your likelihood of having covid is 90%, but it is not so simple.

Using Bayes Theorem, we can determine the actual likelihood that we have Covid using the previously mentioned formula.

H = "We have Covid"

E = "Test evaluates to positive"

Let's apply this framework to our covid test

$P(H|E)$ = Likelihood we have covid **given the test shows true** (what we are trying to solve for)

$P(H)$ = Likelihood we have covid (this estimate is tough to calculate, we can just use the prevalence of Covid in the population)

$P(E|H)$ = Likelihood the test is positive **given we have covid** (reliability)

$P(E|-H)$ = Likelihood the test is positive **given we don't have covid** ($1 - P(E|H)$)

$P(-H)$ = Likelihood we don't have Covid ($1 - P(H)$)



$$P(H|E) = \frac{P(H) P(E|H)}{P(H)P(E|H) + P(-H)P(E|-H)}$$



$P(H|E)$ = (what we are trying to solve for)

$P(H)$ = 0.012 (percent of NY that had Covid in March '21)

$P(E|H)$ = 0.9

$P(E|-H)$ = 0.1

$P(-H)$ = 0.988

We then simply plug and chug these numbers into the calculation below.



$$P(H|E) = \frac{P(H) P(E|H)}{P(H)P(E|H) + P(-H)P(E|-H)}$$



$P(H|E)$ = (what we are trying to solve for)

$P(H)$ = 0.012 (percent of NY that had Covid in March '21)

$P(E|H)$ = 0.9

$P(E|-H)$ = 0.1

$P(-H)$ = 0.988

We then simply plug and chug these numbers into the calculation below.


$$P(H|E) = \frac{0.012 * 0.9}{0.012 * 0.9 + 0.1 * 0.988}$$

Simplifying this calculation we get a probability of ~0.10

This states that we have a ~9% probability that you have COVID given this test catches 90% of positive cases.

This is quite different from our initial assumption of 90%!

However, let's say we take **another** Covid test and this too shows up as positive? How can we measure the new likelihood that we have Covid?



$$P(H|E) = 0.089$$



Well we simply re-use bayes theorem again.

$P(H|E)$ = Likelihood we have covid **given the test shows true** (*what we are trying to solve for*)

$P(H)$ = Likelihood we have covid

$P(E|H)$ = Likelihood the test is positive **given we have covid** (*reliability*)

$P(E|-H)$ = Likelihood the test is positive **given we don't have covid** ($1 - P(E|H)$)

$P(-H)$ = Likelihood we don't have Covid ($1 - P(H)$)

However instead of using $P(H) = 0.012$, **what is the new prior probability that we have Covid?** (Think back to the previous test)

$$P(H|E) = \frac{P(H) P(E|H)}{P(H)P(E|H) + P(-H)P(E|-H)}$$



Well we simply re-use bayes theorem again.

$$P(H|E) = ???$$

$$P(H) = 0.09$$

$$P(E|H) = 0.9$$

$$P(E|-H) = 0.1$$

$$P(-H) = 0.9$$

The beautiful think about Bayes Theorem is that we can use our previous knowledge for this new calculation. Frequentist statistics offers no room for this kind of calculation. Our previous likelihood is our new prior.

$$P(H|E) = \frac{P(H) P(E|H)}{P(H)P(E|H) + P(-H)P(E|-H)}$$



Again, we plug these values in

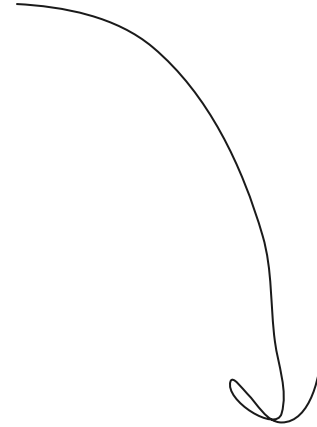
$$P(H|E) = ???$$

$$P(H) = 0.09$$

$$P(E|H) = 0.9$$

$$P(E|-H) = 0.1$$

$$P(-H) = 0.91$$



$$P(H|E) = \frac{0.09 * 0.9}{0.09 * 0.9 + 0.91 * 0.1}$$



Simplifying this calculation we get a probability of 0.5

This states that we now have a 47% probability that you have COVID given this test catches 80% of positive cases.

Notice that this likelihood has now increased given the fact that this is the second positive test.

What do you think will occur to the likelihood of us having COVID with a 3rd positive test??

Again, this is the beauty of Bayes Theorem. We are able to update our beliefs with evidence.

This is going to be relevant during Phase 2 when we start discussing ML algorithms.

$$P(H|E) = 0.47$$

Wrap-Up

Lab (Due 04/21)



Vancouver, Canada

You are a growth analyst at a Vancouver-based consulting firm called Monica Group. Your manager is spearheading the completion of a new analytical tool which will automatically label if a review is positive, neutral, negative, or irrelevant.

You will be kicking off completion of this milestone by independently implementing a minimal-viable-product. **This will be a Python pipeline that ingests a text-file of review data and interfaces with the Open AI API in order to automatically label each review.**

We will release API keys on 4/1



Thursday Review Session

On Thursday we will continue our review of TLAB #2.



Jupyter: scratchpad of the data scientist

If you understand what you're doing, you're not learning anything. - Anonymous