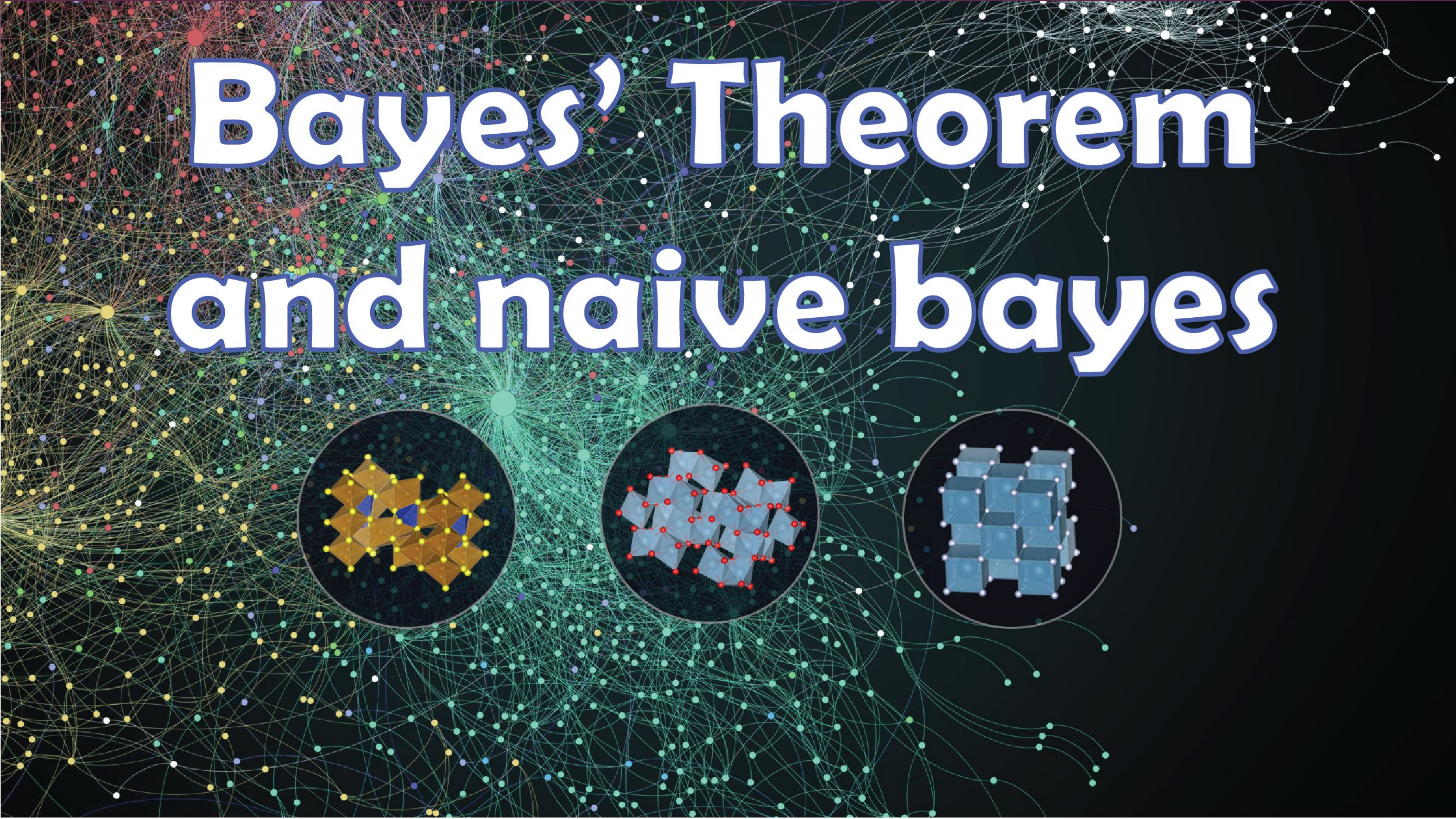
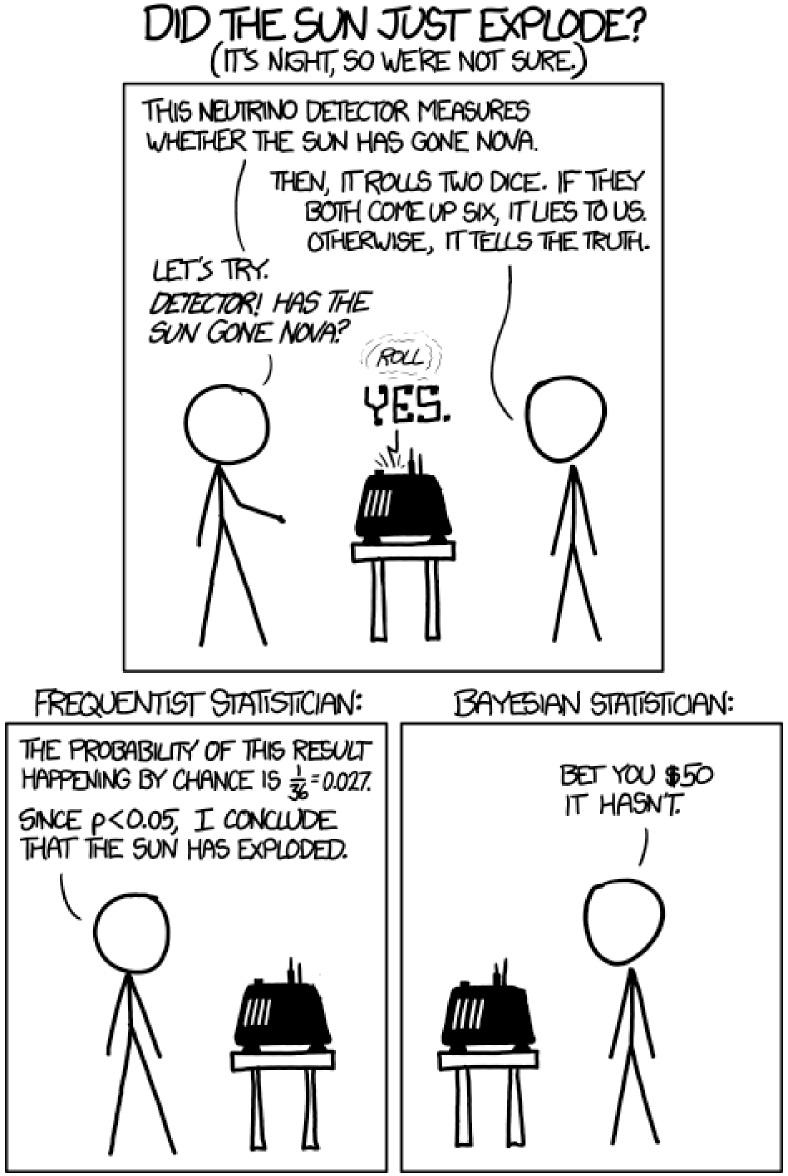
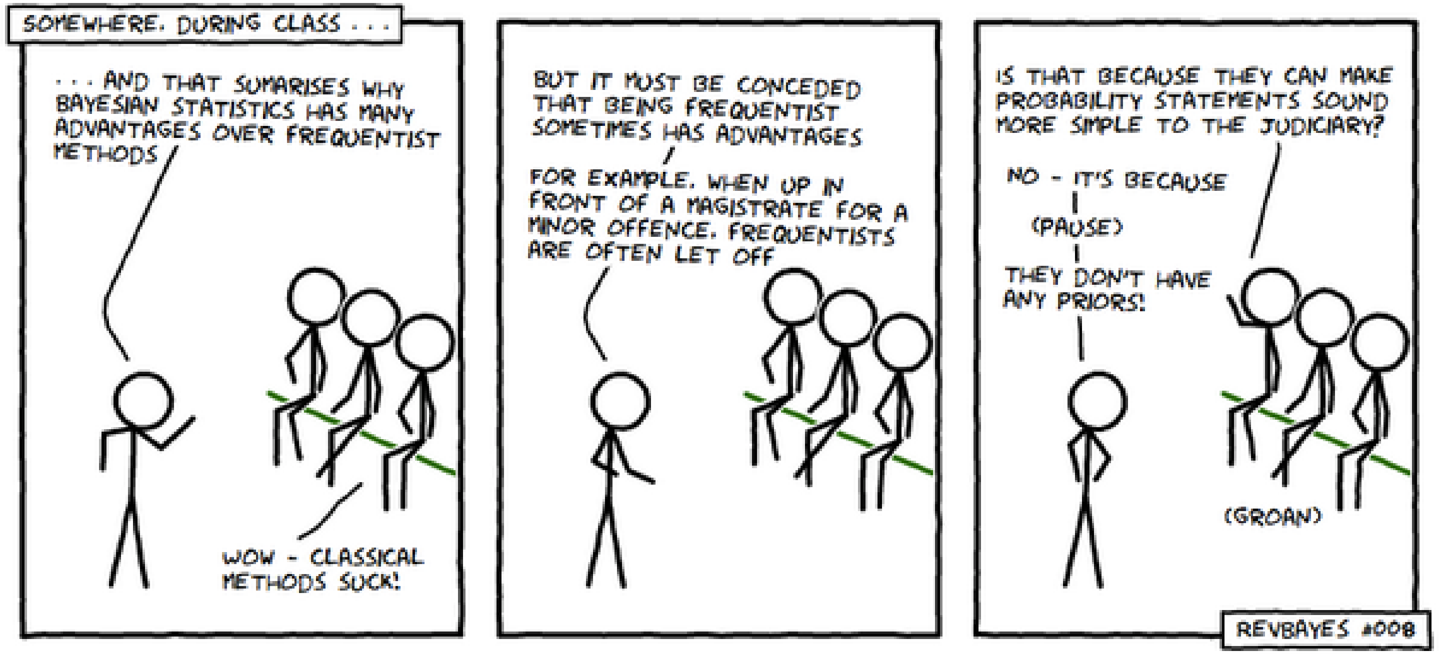

Bayes' Theorem
and naive bayes

Bayesian vs Frequentists is one of the great debates in machine learning





https://xkcd.com/1132/

What are the key differences?

Frequentist	***Bayesian***
Data is a repeatable random sample - there is a frequency	Data is fixed, probably a few samples.
No expression of belief (formally not present) => Objective view on probability.	Approach deals with belief (Formally present) => Subjective view on probability. It helps to update their beliefs in the evidence of new data (thus creating posterior distribution).
Provides us with a point estimate using MLE and Least Squares Estimate.	Provides us with a posterior distribution with high density interval with mean, mode and median stats.
Parameters are fixed and unknown.	Parameters are unknown, random and described probabilistically.
Confidence Interval: Over an infinite sample size taken from population, 95% of these contain the true population value.	**Credible Interval:** A 95% probability that the population value is within the limits of the interval.
Statistical Hypothesis Testing with p-value and significance level is employed to deduce a solution in the decision-making process.	Bayes Factor considered a direct test of null and alternate hypothesis, yielding a measure of strength of evidence.
Less Computationally intensive	Computationally intensive

Before we get started, we need to talk about conditional probability

$LaYS_3$

$LiMn_2O_4$

$SrTiO_3$

SiO_2

$LaAlO_3$

MgO

$YScS_3$

$CaTiO_3$

$La_2Zr_2O_7$

CaF_2

$NaCl$

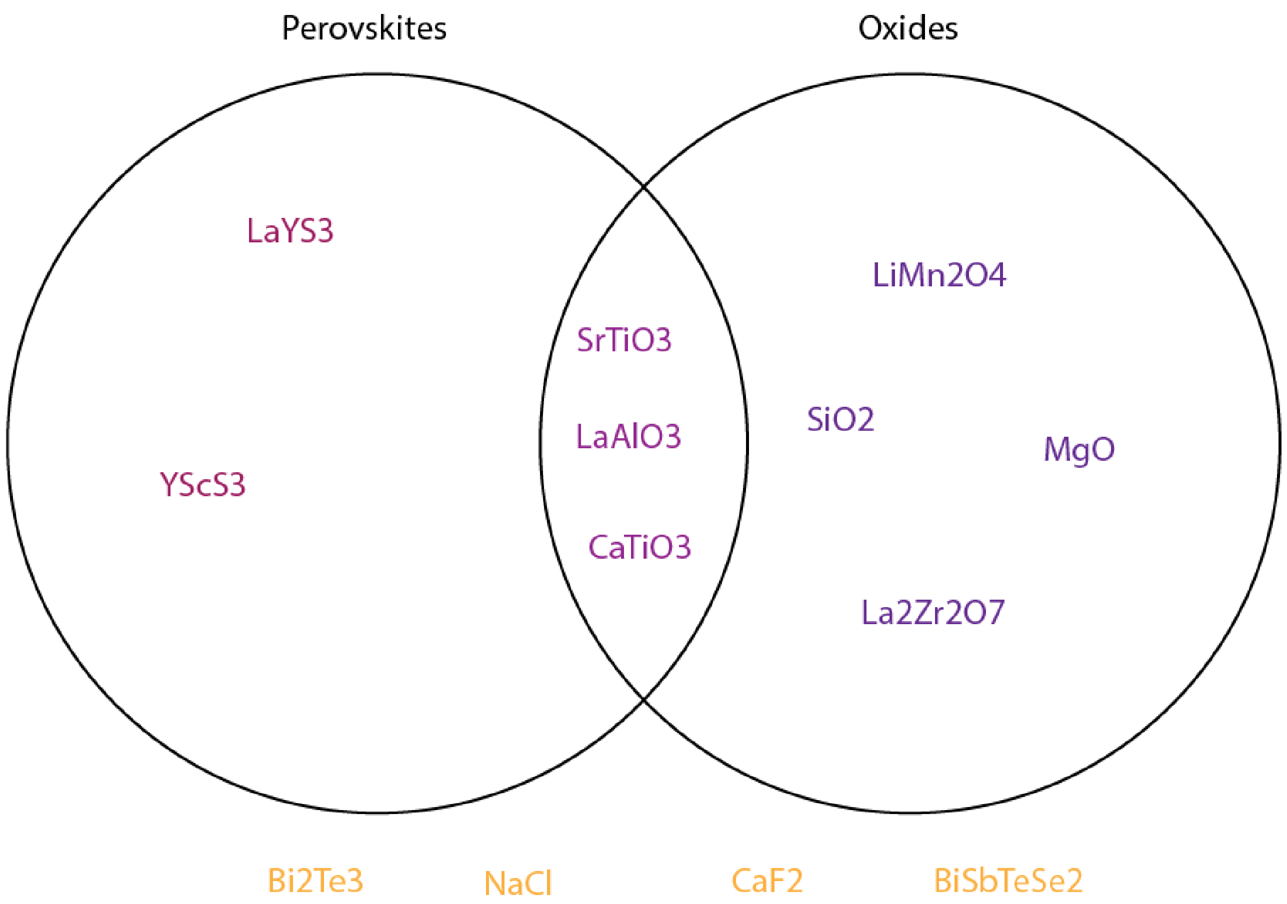
Bi_2Te_3

$BiSbTeSe_2$

Conditional Probabilities, Clearly Explained!!!, StatQuest, Josh Starmer, YouTube https://youtu.be/_IgyaD7vOOA

Before we get started we need to talk about conditional probability

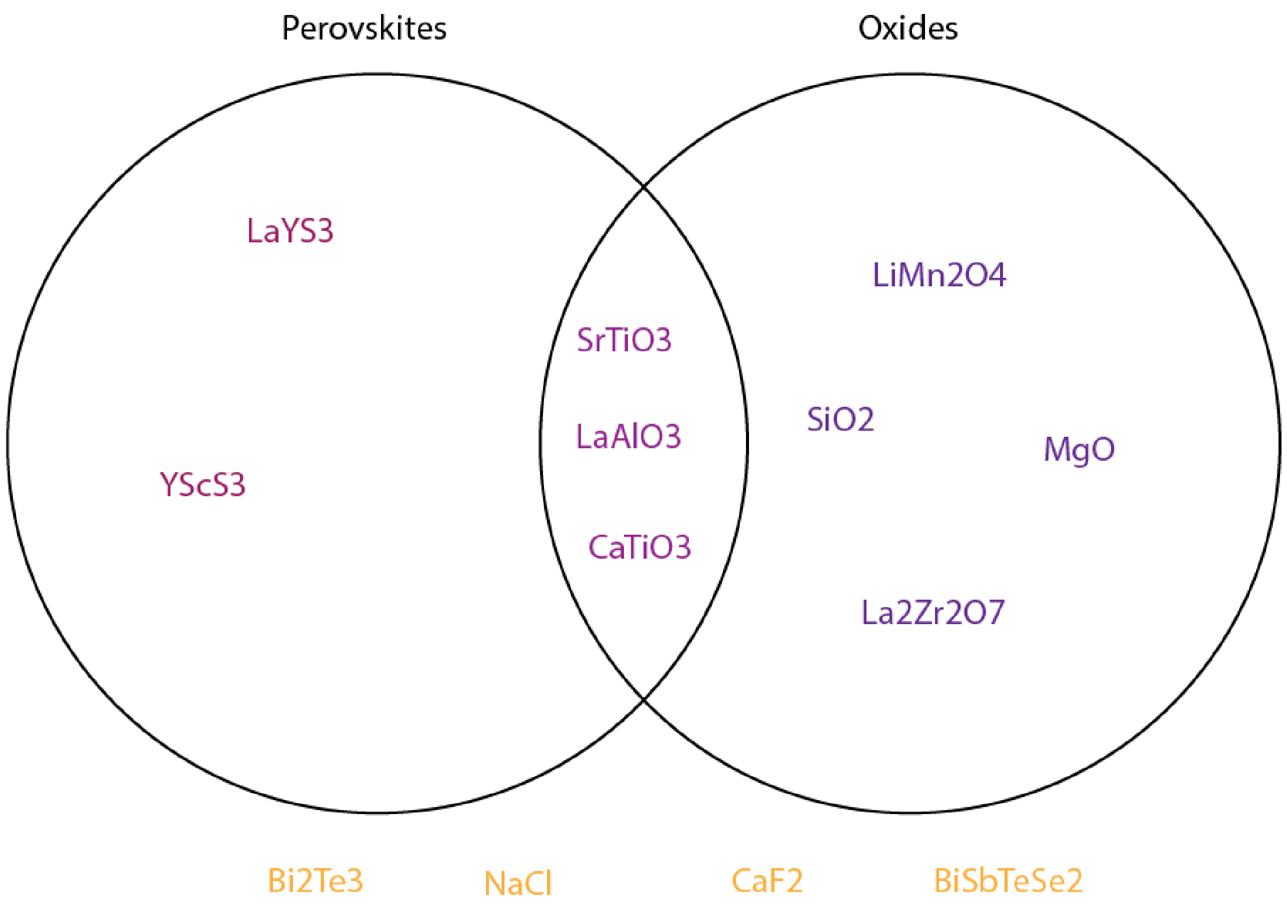
$LaYS_3$
$LiMn_2O_4$
$SrTiO_3$
SiO_2
$LaAlO_3$
MgO
$YScS_3$
$CaTiO_3$
$La_2Zr_2O_7$
CaF_2
$NaCl$
Bi_2Te_3
$BiSbTeSe_2$



Conditional Probabilities, Clearly Explained!!!, StatQuest, Josh Starmer, YouTube https://youtu.be/_IgyaD7vOOA

Before we get started, we need to talk about conditional probability

LaYS3
LiMn2O4
SrTiO3
SiO2
LaAlO3
MgO
YScS3
CaTiO3
La2Zr2O7
CaF2
NaCl
Bi2Te3
BiSbTeSe2



	Is an oxide	Not an oxide
Is a perovskite	3	2
Not a perovskite	4	4

Conditional Probabilities, Clearly Explained!!!, StatQuest, Josh Starmer, YouTube https://youtu.be/_IgyaD7vOOA

You're already familiar with probability

What's the probability of the next material being an oxide perovskite?

$$p(oxide\ perovskite) =$$

$$p(oxide) =$$

$$p(perovskite) =$$

You're already familiar with probability

What's the probability of the next material being an oxide perovskite?

$$p(oxide\ perovskite) = \frac{3}{3+4+4+2} = \frac{3}{13} = 0.231$$

$$p(oxide) = \frac{7}{13} = 0.538$$

$$p(perovskite) = \frac{5}{13} = 0.385$$

We can use our contingency table to introduce conditional probability!

If the next material is an oxide, what's the probability it's a perovskite?

$$p(oxide\ perovskite\ |oxide) =$$

What's the probability it's a perovskite if you know it's not an oxide?

$$p(perovskite, not\ an\ oxide\ |\ not\ an\ oxide) =$$

Conditional probability is **scaled** by knowledge we already have

If the next material is an oxide, what's the probability it's a perovskite?

$$p(oxide\ perovskite\ |oxide) = \frac{3}{7} = 0.429$$

What's the probability it's a perovskite if you know it's not an oxide?

$$p(perovskite, not\ an\ oxide\ |\ not\ an\ oxide) = \frac{2}{6} = 0.333$$

Bayes theorem extends from this conditional probability

What's the probability the next material is non-oxide perovskite given that…

$$p(not\ oxide\ \&\ perovskite\ |perovskite) = \frac{p(not\ oxide\ \&\ perovskite)}{p(perovskite)}$$

$$p(not\ oxide\ \&\ perovskite\ |not\ an\ oxide) = \frac{p(not\ oxide\ \&\ perovskite)}{p(not\ an\ oxide)}$$

Can you predict these conditional probabilities if you don't know the probability of being perovskite but not oxide???

$$p(not\ oxide\ \&\ perovskite) = unkown$$

The standard way of writing Bayes' Theorem is similar

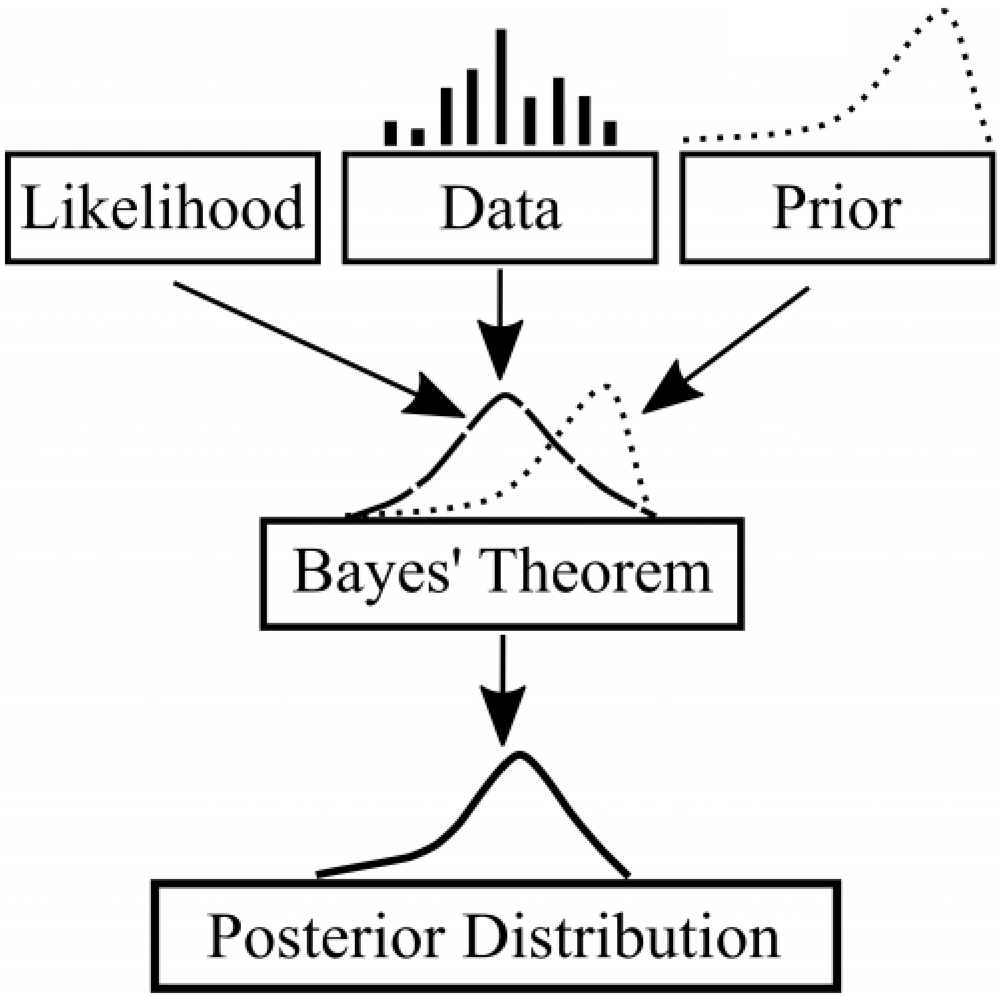
$$p(A\&B|B) = \frac{p(A\&B)}{p(B)}$$
$$p(A\&B) = p(A\&B|A)p(A) = p(A\&B|B)p(B)$$

A=not an oxide
B=perovskite

$$p(A\&B|B) = \frac{p(A\&B|A)p(A)}{p(B)}$$
$$p(A\&B|A) = \frac{p(A\&B|B)p(B)}{p(A)}$$

If you didn't have all the info… but you had part and then made a guess, you could still calculate conditional probabilities

How do you use Bayes’ Theorem in machine learning?



Let's start with a Naïve Bayes (Multinomial Naïve Bayes classifier)

Consider a classifier for insulator vs metal

If you have labeled data, you could go element by element in your feature vector and figure out the probability of each element being present for your insulator vs metal classes

$$p(Insulator|Coordination\ Number = 6) = 0.5$$
$$p(Insulator|oxide) = 0.8$$
$$p(Insulator|Density < 5) = 0.6$$

$$p(Metal|Coordination\ Number = 6) = 0.3$$
$$p(Metal|oxide) = 0.3$$
$$p(Metal|Density < 5) = 0.4$$

We need to start with an initial (prior) guess and then we update it

For a new material, when predicting insulator vs metal, we can start with a guess based on our prior experience. If 2/3 of our labeled data was insulators, then we can guess that p=0.66 the next one will be insulator.

"prior probability"

Then we can consider the features available.... oxide, density=4g/cc

$$p(new\ material\ is\ Metal) = p(Metal) * p(Metal|oxide) * p(Metal|Density < 5)$$
$$= \frac{1}{3} * 0.3 * 0.4 = 0.04$$

$$p(new\ material\ is\ Insulator)$$
$$= p(Insulator) * p(Insulator|oxide) * p(Insulator|Density < 5) = \frac{2}{3} * 0.8 * 0.6 = 0.32$$

This approach breaks if we come across a feature that was never in training data

Since probability is now just prior probability times the probability of each feature, if any are zero then the whole probability becomes zero

$$p(new\ material\ is\ Metal) = p(Metal) * p(Metal|perovskite) * p(Metal|Density < 5)$$
$$= \frac{1}{3} * 0 * 0.4 = 0$$

To fix this, we should add a small probability of each feature element.

We can also use Gaussian naïve Bayes to do our classification

Consider a classifier for insulator vs metal
In your labeled data collect average and standard deviation for our features

Insulator:
Average density = 4g/cc, stdev = 2g/cc
Average atomic size = 1.3Angstroms, stdev = 0.2Angstroms
Average electronegativity difference = 2.1, stdev = 0.4

Metal:
Average density = 4g/cc, stdev = 2g/cc
Average atomic size = 0.6Angstroms, stdev = 0.2Angstroms
Average electronegativity difference = 0.4, stdev = 0.3

We can represent these data as Gaussians

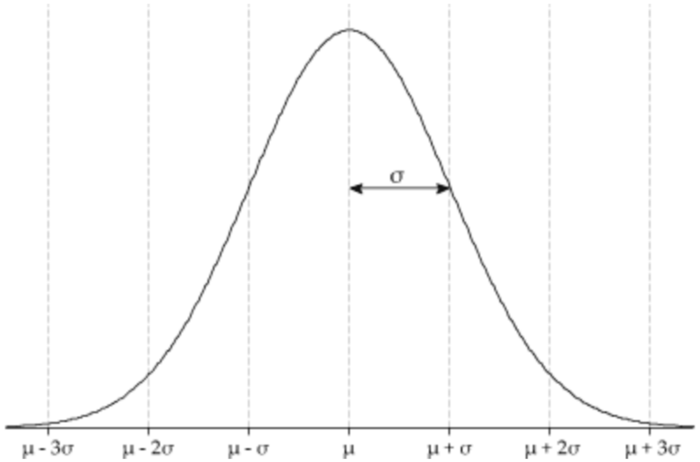
Normal distribution

The Normal distribution is arguably the most important continuous distribution. It is used throughout the sciences, because of a remarkable result known as the *central limit theorem*, which is covered in the module *Inference for means*. Due to the phenomenon behind the central limit theorem, many variables tend to show an empirical distribution that is close to the Normal distribution.

If X has a **Normal distribution** with mean μ and standard deviation σ, then we write that $X \stackrel{d}{=} N(\mu, \sigma^2)$; the probability density function of X is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \qquad \text{for } x \in \mathbb{R}.$$

This distribution is so important that it is well known in general culture, where it is often referred to as the **bell curve** — for example, in the controversial 1994 book by R. J. Herrnstein entitled *The Bell Curve: Intelligence and Class Structure in American Life*.



Detailed description

Figure 2: The pdf of a Normal random variable with mean μ and standard deviation σ.

https://amsi.org.au/ESA_Senior_Years/SeniorTopic4/4f/4f_2content_3.html

Just like with Naïve Bayes, we start with our initial (prior) probabilities

We calculate our probability of being metal or insulator for a new material using Gaussians

New material density 3, avg size 0.9, avg electronegativity difference 1.5

$$p(new\ material\ Metal)$$
$$= p(Metal) * L(Metal|Density\ 3) * L(Metal|size\ 0.9) * L(Metal|\chi\ 1.5) =$$

$$p(new\ material\ Insulator)$$
$$= p(Insulator) * L(Insulator|Density\ 3) * L(Insulator|size\ 0.9) * L(Insulator|\chi\ 1.5) =$$

We often add log of these scores to prevent “underflow” when probabilities are really small or to help even out features that are not equally important.


Gaussian processes