Naïve Bayesians

Back to Basics Series

06 Feb 2021

Goal

Developing the Bayesian muscle to solve a wide range of problems

Naïve Bayesian Philosophy

Intuitive (Visual) Understanding of the Bayesian Reasoning

Ability to model real world problems in a Bayesian Setting

Starting from Simple Probabilistic modelling

Adapting it in a a Bayesian setting
And moving towards ML models

Fluency in the Calculus of Bayesian Stats & ML model

Season 2: Back to Basics

Ep 1	Ep 2	Ep 3	Ep 4	Ep 5	Ep 6	Ep 7	Ep 8
Bayes Theorem	Problems with Binomial Likelihoods		Disease Detection	Naive Bayes Classification	Gaussian Naive Bayes Classification	German Tank Problem	Waiting Times (Continuous Distributions)

Back to Basics

		Canonical Problem	Applications
Ep 1	Bayes Theorem	There are 2 boxes from which cookies can be taken from. Box A and Box B. Box A contains 10 chocolate cookies, Box B contains 5 ginger cookies. Given that you get a chocolate cookie which box was it taken from?	The Shy Librarian Problem Naive Bayes algorithm
Ep 2 Ep 3	Problems with Binomial Likelihoods	You have 2 coins C1 and C2. p(heads for C1) = .7 & P(heads for C2) = 0.6 You flip the coin 10 times. What is the probability that the given coin you picked is C1 given you have 7 heads and 3 tails?	A/B Testing
Ep 4	Disease Detection	A particular disease affects 1% of the population. There is an imperfect test for this disease: The test gives a positive result for 90% of people who have the disease, and 5% of the people who are disease-free. Given a positive test result – what is the probability of having the disease?	COVID Tests (PCR & Antibody)! Fraud Detection
Ep 5	Naive Bayes Classification	Given these words occur in this text what's the probability it's spam?	Any Classification Problem
Ep 6	Gaussian Naive Bayes Classification	Given the weights and heights of basketball players, what's the probability that person a is a basketball player given weight = w and height = h?	

Back to Basics

		Canonical Problem	Applications
Ep 7	German Tank Problem	Suppose tanks were given a serial number based on the order in which they were manufactured. Given that you've observed a tank with serial number "10", how many tanks were actually manufactured in total?	?
Ep 8	Waiting Times (Continuous Distributions)	Suppose you need to gather 10 patients for a trial. Each signup happens at time t_i (i=1, 10). How long do you have to wait after it took you 3 weeks to accrue 2 signups?	Planning Trials Estimating Queues

Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

Recap | Canonical Problem

Given the words “Dear Friend” occur in this email what's the probability it's spam?

$$P(S \mid \text{Dear Friend})$$

N

Normal

S

Spam

Canonical Problem

Given the size of an email is 1.8 MB & the time to read it is 2 seconds what's the probability it's spam?

$$P(S \mid 1.8MB, 2sec)$$

N

Normal

S

Spam

Canonical Problem Simplified

Given the size of an email is 1.8 MB & the time to read it is 2 seconds
what's the probability it's spam?

$$P(S \mid 1.8\ MB)$$

N

Normal

S

Spam

Given the size of an email is 1.8MB, what's the probability it's spam?

8 Normal Emails

180	976
190	1280
256	1500
780	1798

KB

4 Spam Emails

980
1850
1950
2000

KB

N

Normal

S

Spam

Fitting a Gaussian Distribution

8 Normal Emails

180	976
190	1280
256	1500
780	1798

KB

μ = 870 KB

σ = 628 KB

4 Spam Emails

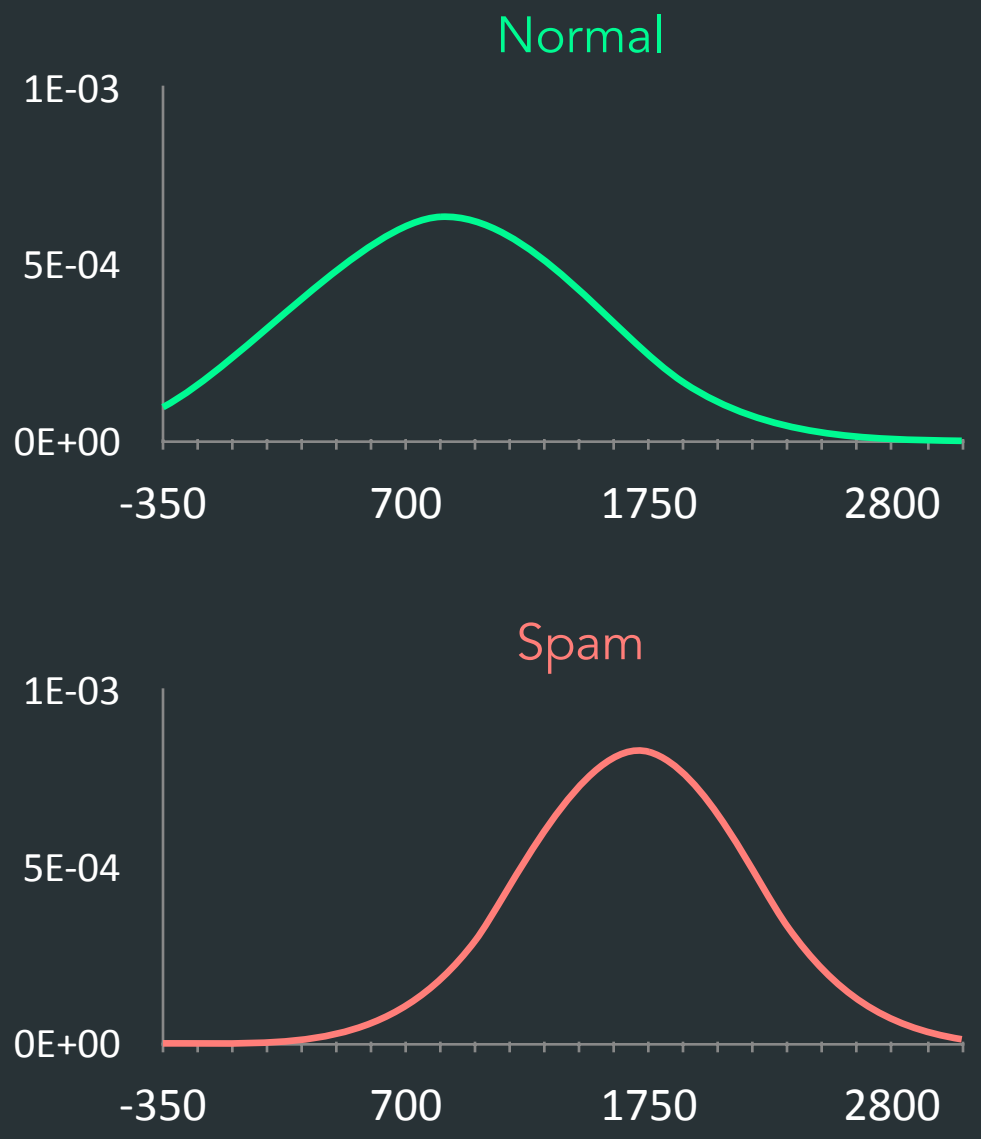
980
1850
1950
2000

KB

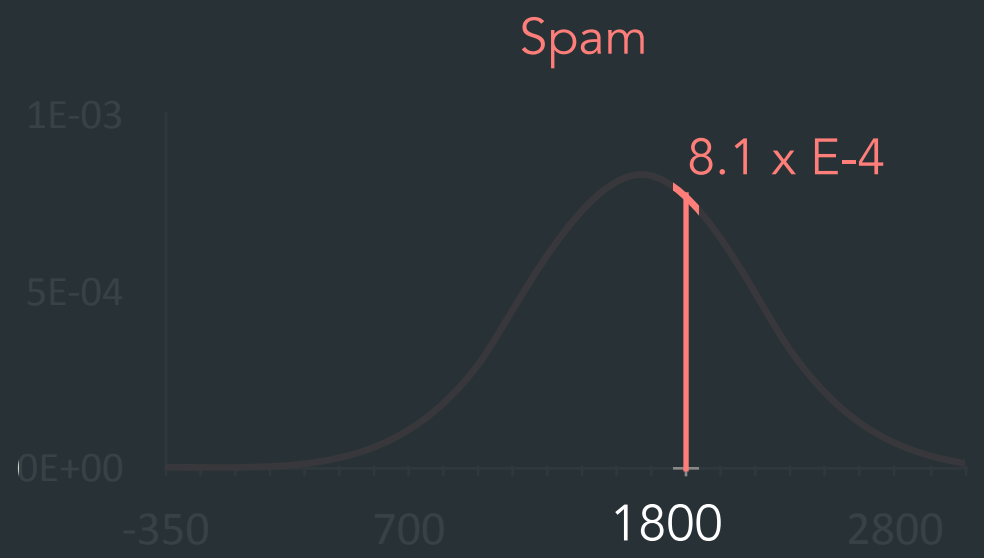
μ = 1697 KB

σ = 481 KB

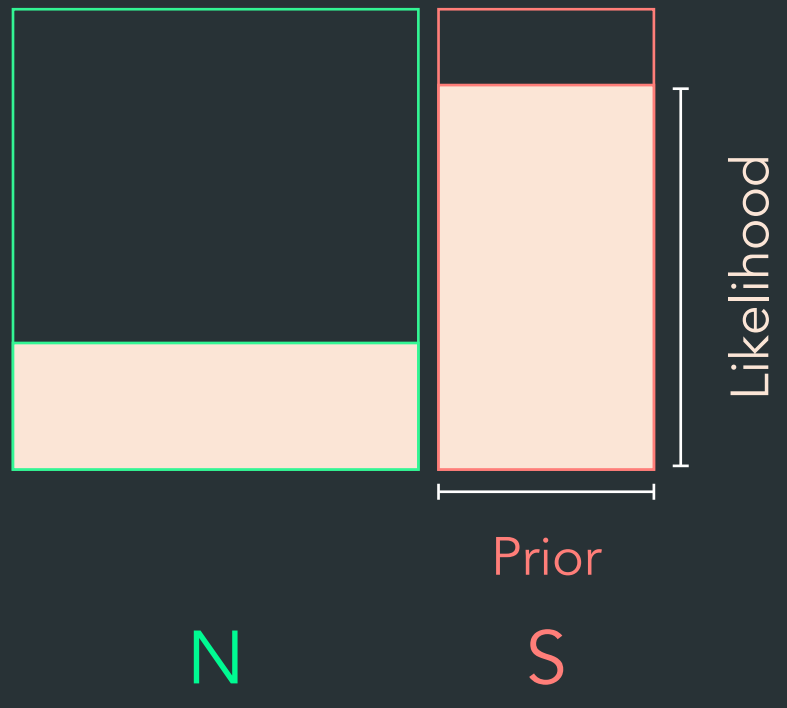
Fitting a Gaussian Distribution

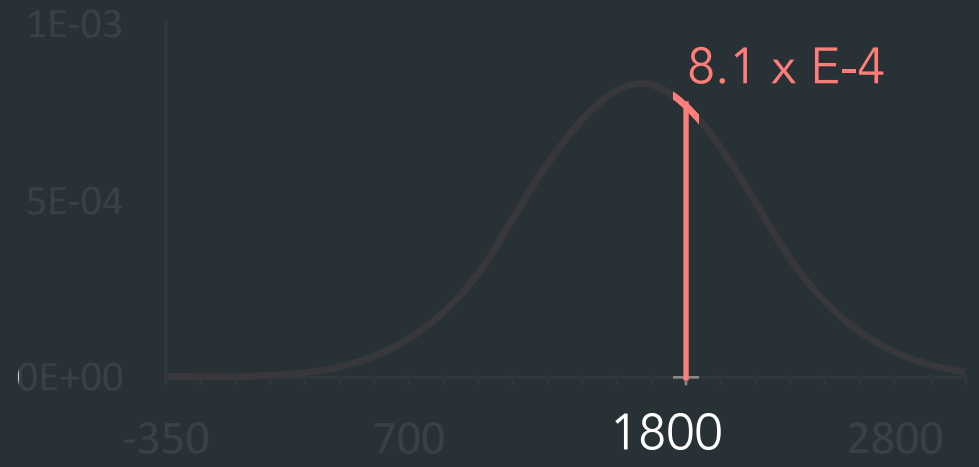


Given the size of an email is 1.8MB, what's the probability it's spam?

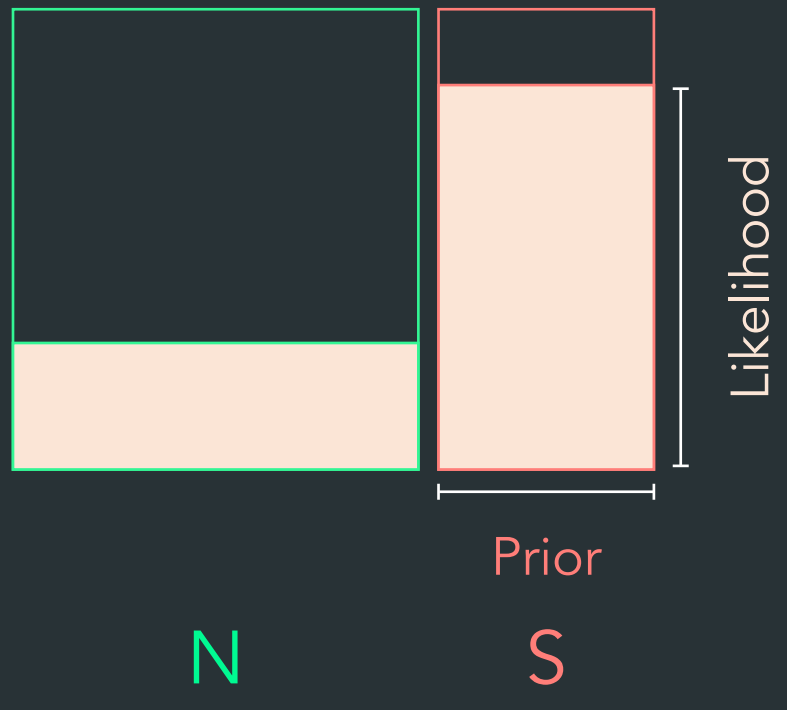


Given the size of an email is 1.8MB, what's the probability it's spam?





Given the size of an email is 1.8MB, what's the probability it's spam?



$$P(S \mid 1.8MB) \propto P(1.8MB \mid S)\, P(S)$$
$$= (8.1 \times E\text{-}4)\ \ (1/3)$$
$$= 2.7 \times E\text{-}4$$

$$P(N \mid 1.8MB) \propto P(1.8MB \mid N)\, P(N)$$
$$= (2.1 \times E\text{-}4)\ \ (2/3)$$
$$= 1.4 \times E\text{-}4$$

$$P(S \mid 1.8MB) > P(N \mid 1.8MB)$$

Canonical Problem

Given the size of an email is 1.8 MB & the time to read it is 2 seconds what's the probability it's spam?

$$P(S \mid 1.8MB, 2sec)$$

N

Normal

S

Spam

Given the size of an email is 1.8MB & the time to read it is 2 seconds

8 Normal Emails

Size

μ = 870 KB	
σ = 628 KB	

KB

Time to read

2	5.2
2.5	6.5
3.8	7
4.5	10

sec

4 Spam Emails

μ = 1697 KB
σ = 481 KB

KB

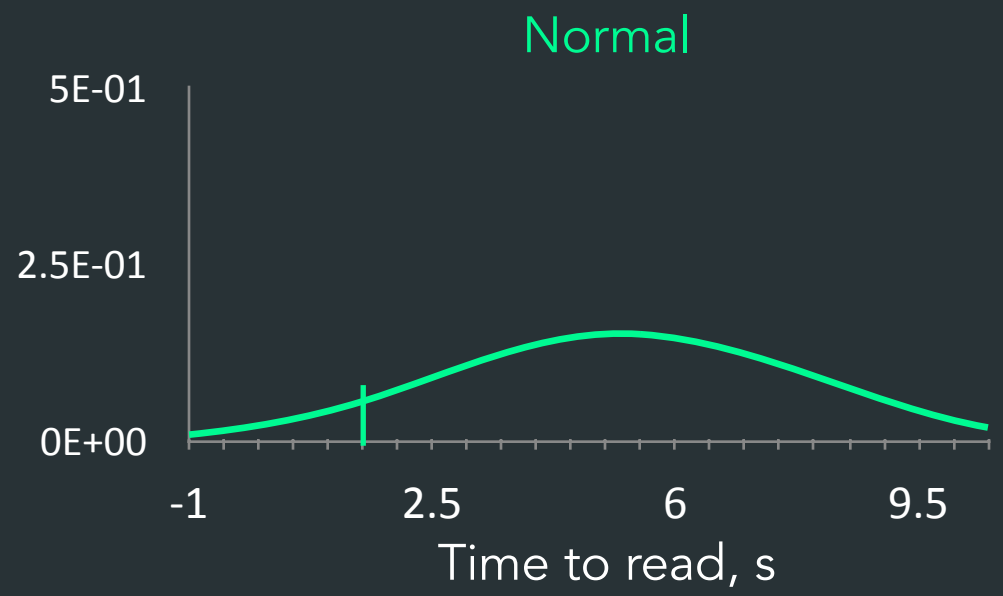
1.8
2
2.95
3.75

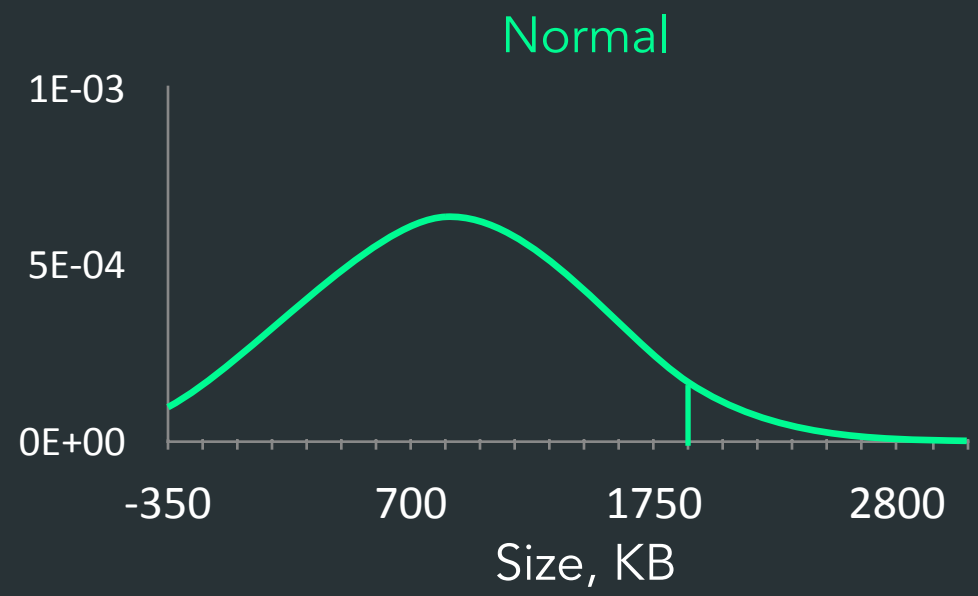
sec

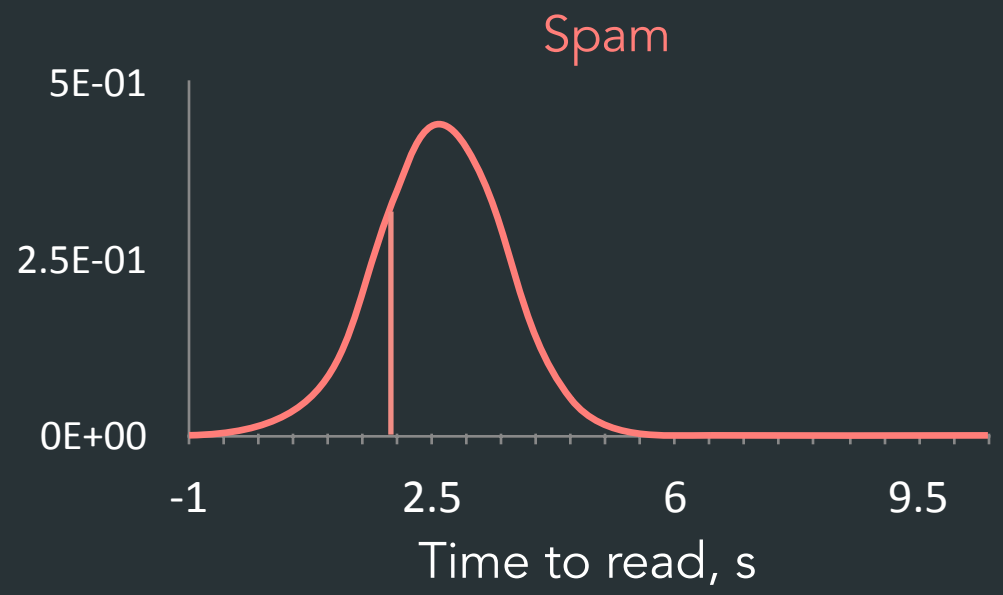
Given the size of an email is 1.8MB & the time to read it is 2 seconds

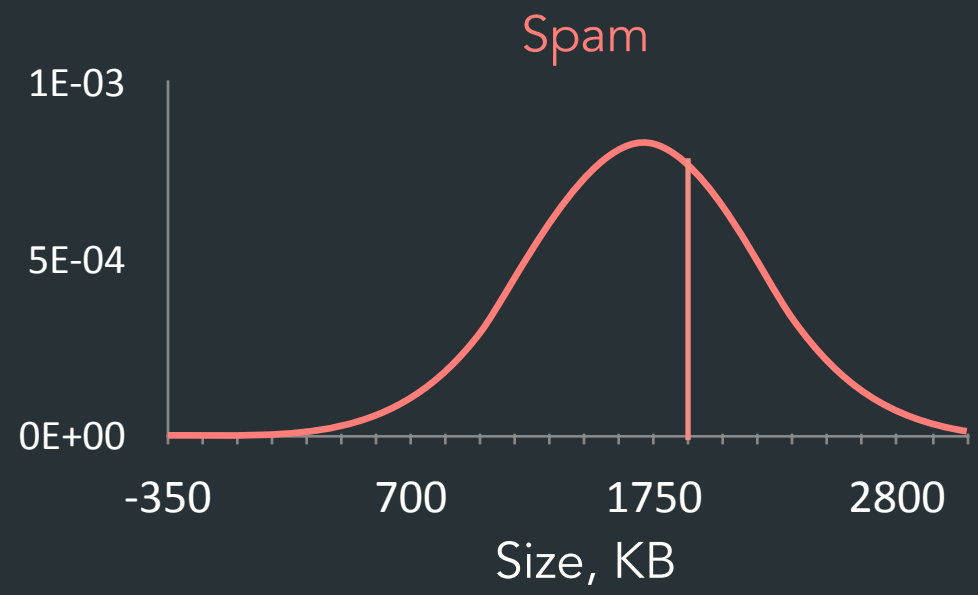
	8 Normal Emails	4 Spam Emails
Size	μ = 870 KB σ = 628 KB	μ = 1697 KB σ = 481 KB
Time to read	μ = 5.2 sec σ = 2.6 sec	μ = 2.7 sec σ = 0.9 sec

Fitting multiple Gaussian Distributions

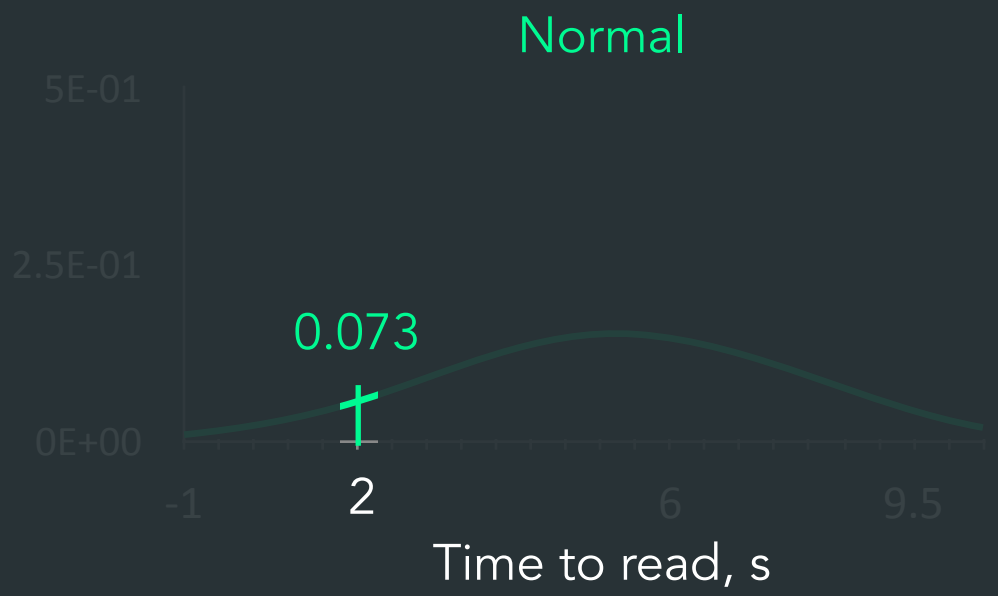


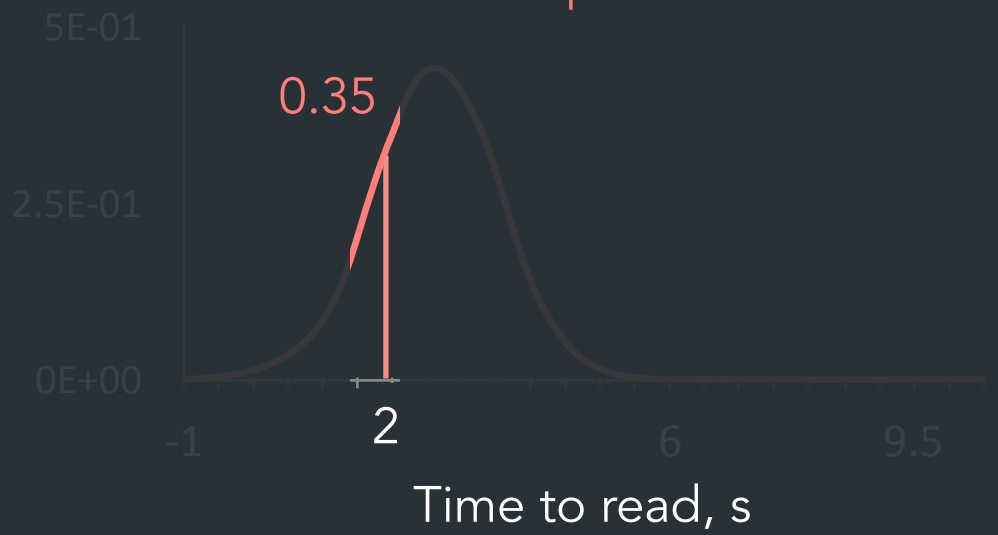






Gaussian Naive Bayes Solution





Gaussian Naive Bayes Solution

$$P(S \mid 1.8MB, 2s) \propto P(1.8MB, 2s \mid S) \times P(S)$$

$$\approx P(1.8MB \mid S) \times P(2s \mid S) \times P(S)$$

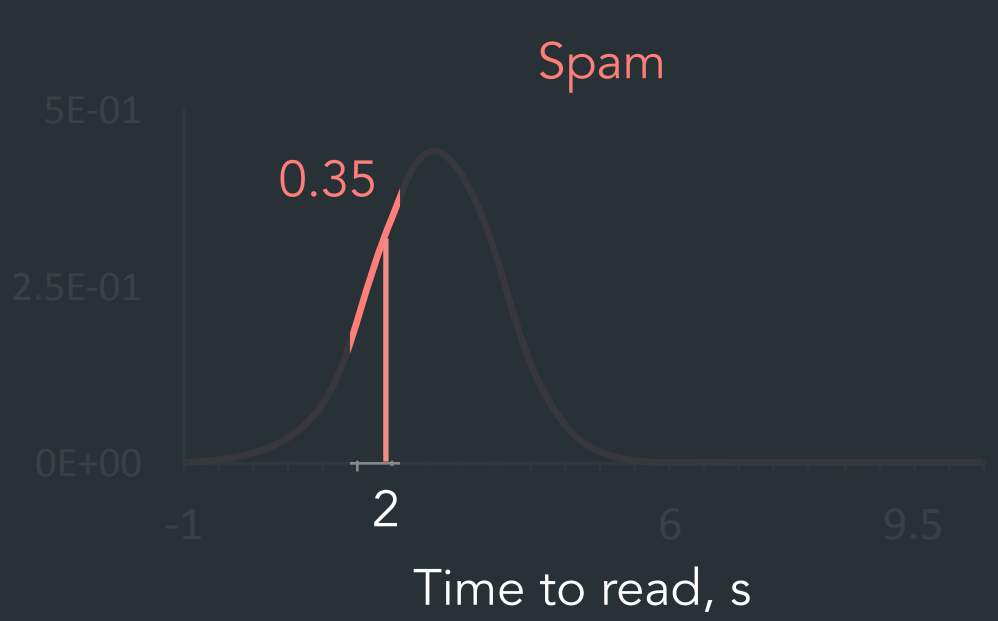
Naïve

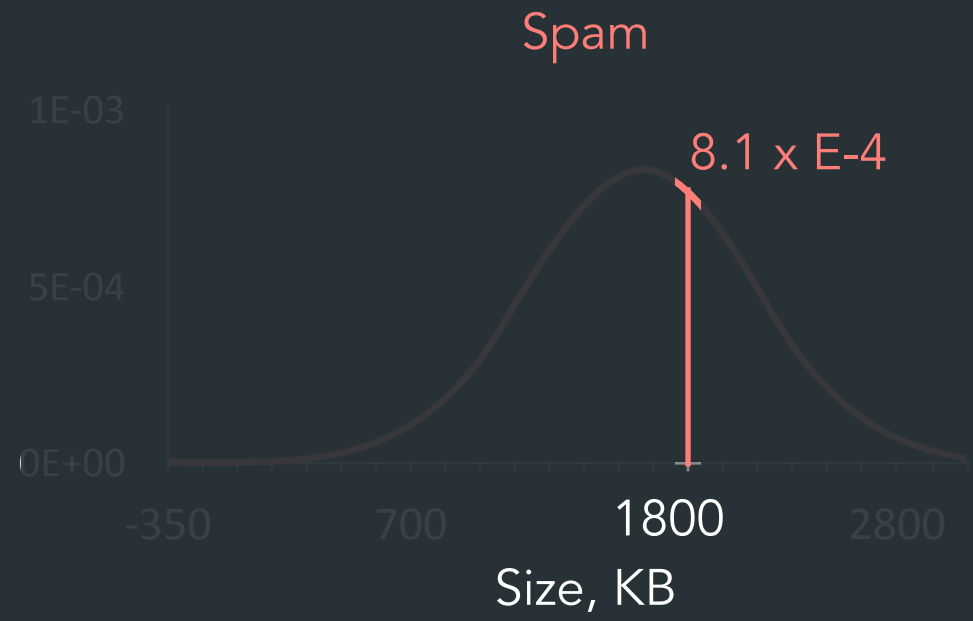
Gaussian Naive Bayes Solution

$$P(S \mid 1.8MB, 2s) \propto P(1.8MB, 2s \mid S) \times P(S)$$

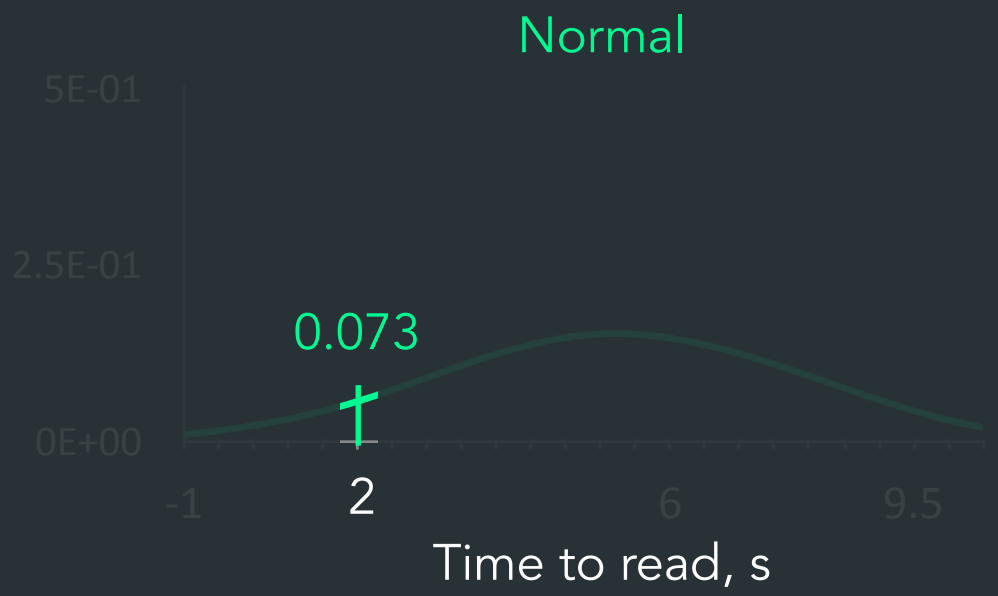
$$\approx P(1.8MB \mid S) \times P(2s \mid S) \times P(S)$$

$$= (8.1 \times E\text{-}4) \times (0.35) \times (1/3) = 9.39\ E\text{-}5$$





Gaussian Naive Bayes Solution

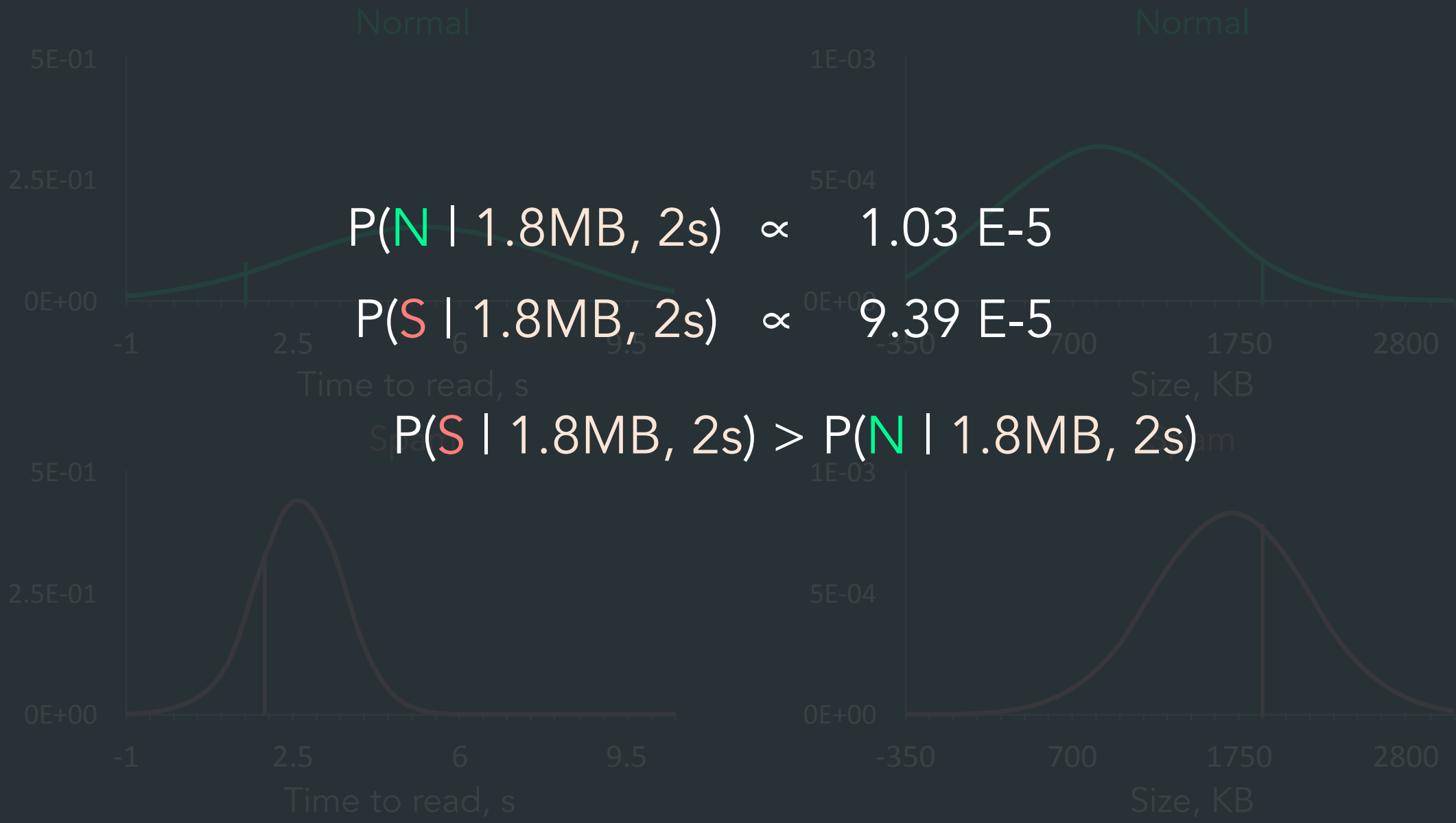


$$P(N \mid 1.8MB, 2s) \propto P(1.8MB \mid N) \times P(2s \mid N) \times P(N)$$

$$= (2.1 \times E\text{-}4) \times (0.073) \times (2/3)$$

$$= 1.0\ E\text{-}5$$

Given the size of an email is 1.8MB & the time to read it is 2 seconds...

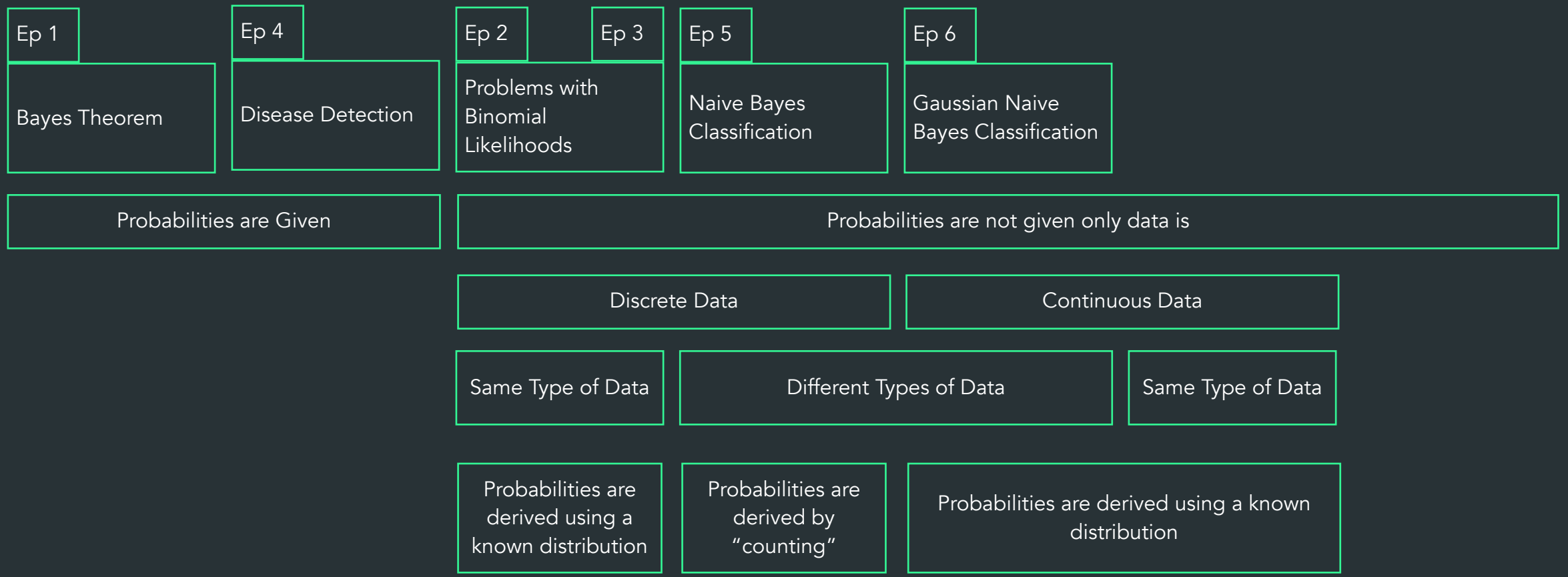


$$P(N \mid 1.8MB, 2s) \propto 1.03\ E\text{-}5$$

$$P(S \mid 1.8MB, 2s) \propto 9.39\ E\text{-}5$$

$$P(S \mid 1.8MB, 2s) > P(N \mid 1.8MB, 2s)$$

Takeaways


Ep 1
Bayes Theorem
Ep 4
Disease Detection
Ep 2
Ep 3
Problems with Binomial Likelihoods
Ep 5
Naive Bayes Classification
Ep 6
Gaussian Naive Bayes Classification
Probabilities are Given
Probabilities are not given only data is
Discrete Data
Continuous Data
Same Type of Data
Different Types of Data
Same Type of Data
Probabilities are derived using a known distribution
Probabilities are derived by "counting"
Probabilities are derived using a known distribution

References

StatQuest: Naive Bayes, Clearly Explained
https://www.youtube.com/watch?v=O2L2Uv9pdDA

StatQuest: Gaussian Naive Bayes, Clearly Explained!!!
https://www.youtube.com/watch?v=H3EjCKtlVog