Naïve Bayesians

Back to Basics Series

30 Jan 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

Episode	Topic
Ep 1	Bayes Theorem
Ep 2, Ep 3	Problems with Binomial Likelihoods
Ep 4	Disease Detection
Ep 5	Naive Bayes Classification
Ep 6	Gaussian Naive Bayes Classification
Ep 7	German Tank Problem
Ep 8	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

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	S
Normal	Spam

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

8 Normal Emails		4 Spam Emails	
Dear:	8	Dear:	2
Friend	5	Friend:	1
Lunch:	3	Lunch:	0
Money:	1	Money:	4

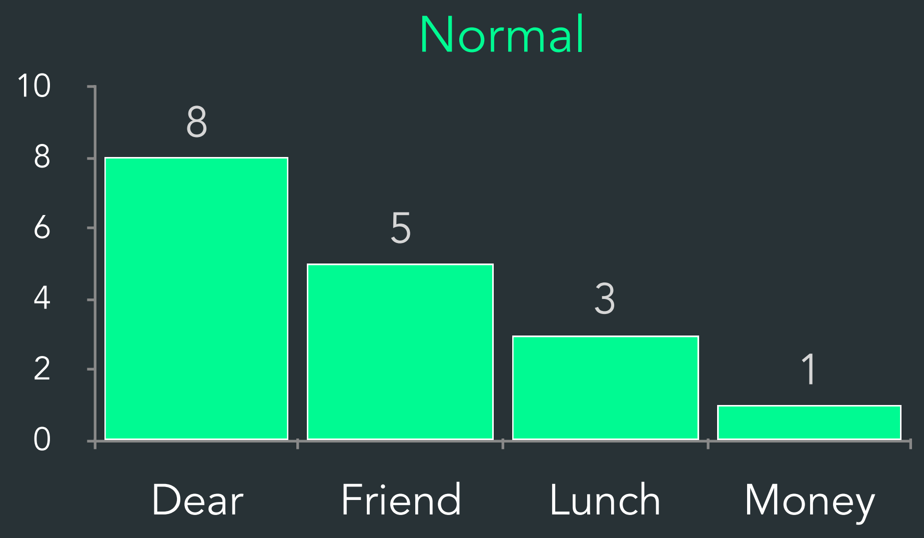
N

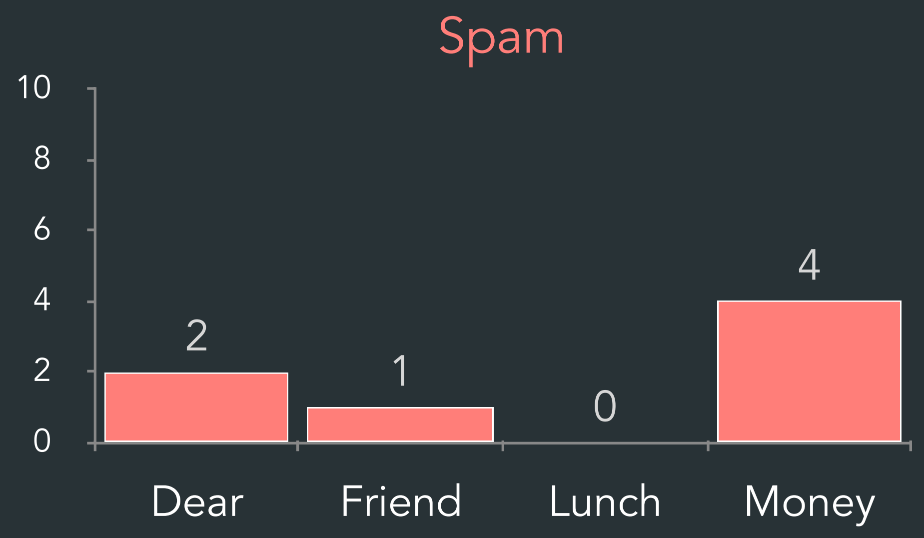
Normal

S

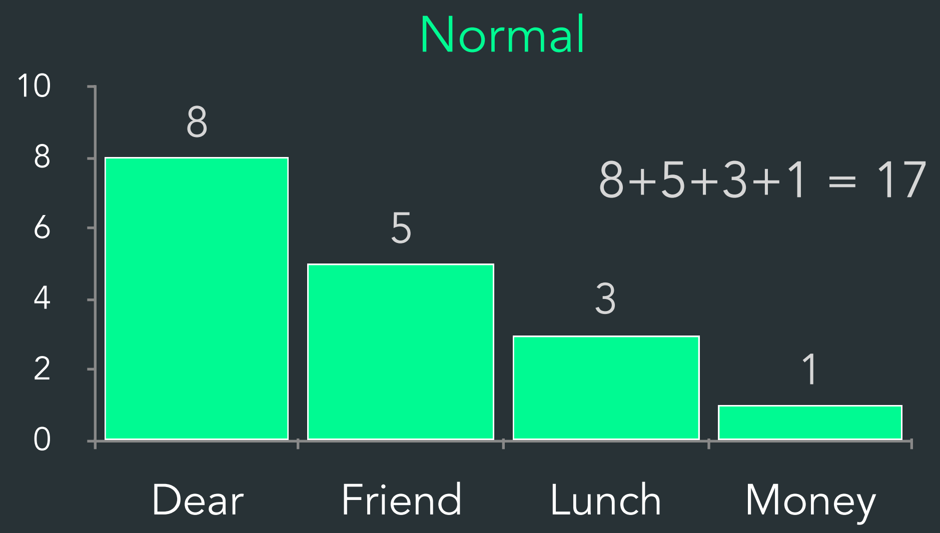
Spam

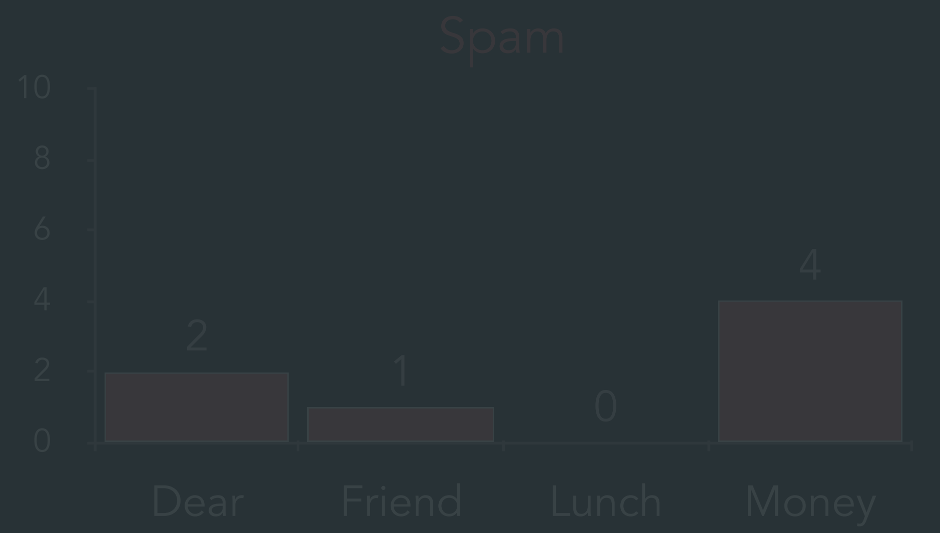
Given the words “Dear Friend” occur in this email, what's the probability it's spam?



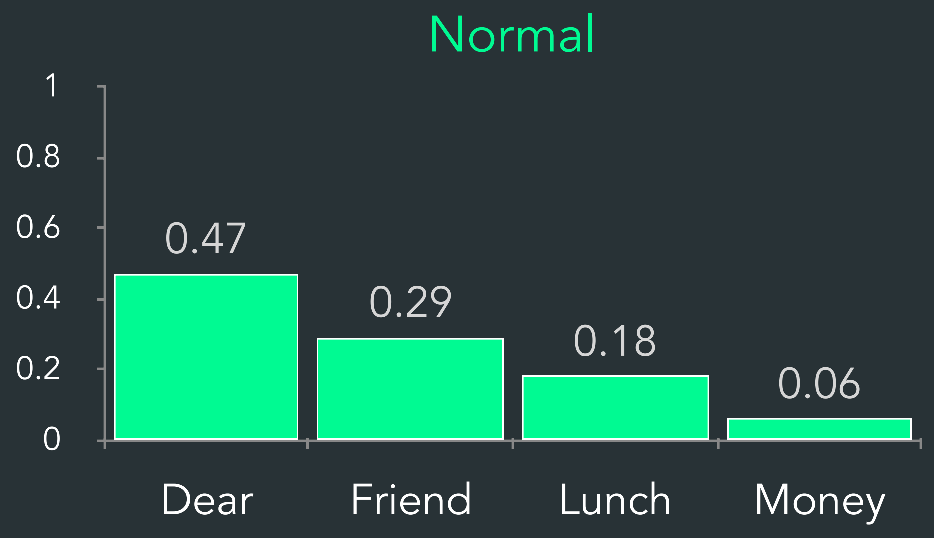


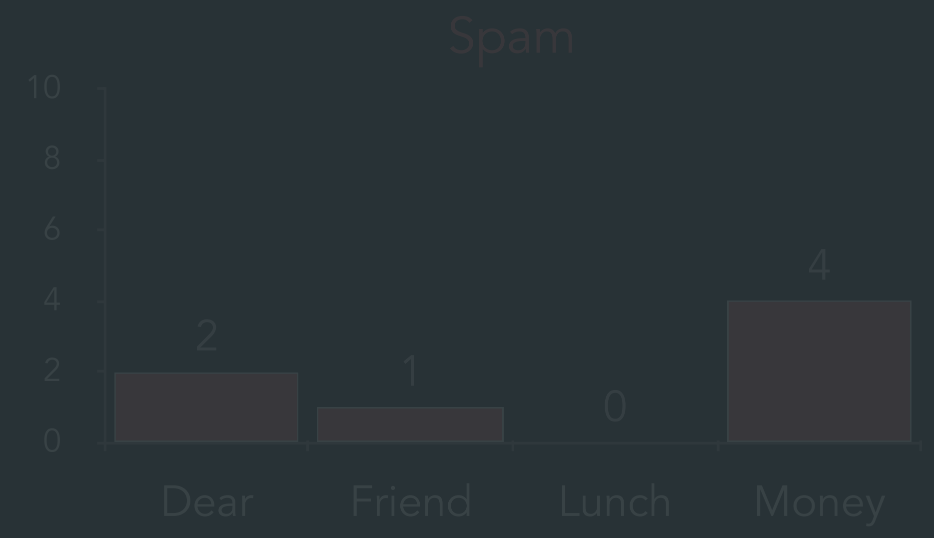
Given the words "Dear Friend" occur in this email, what's the probability it's spam?



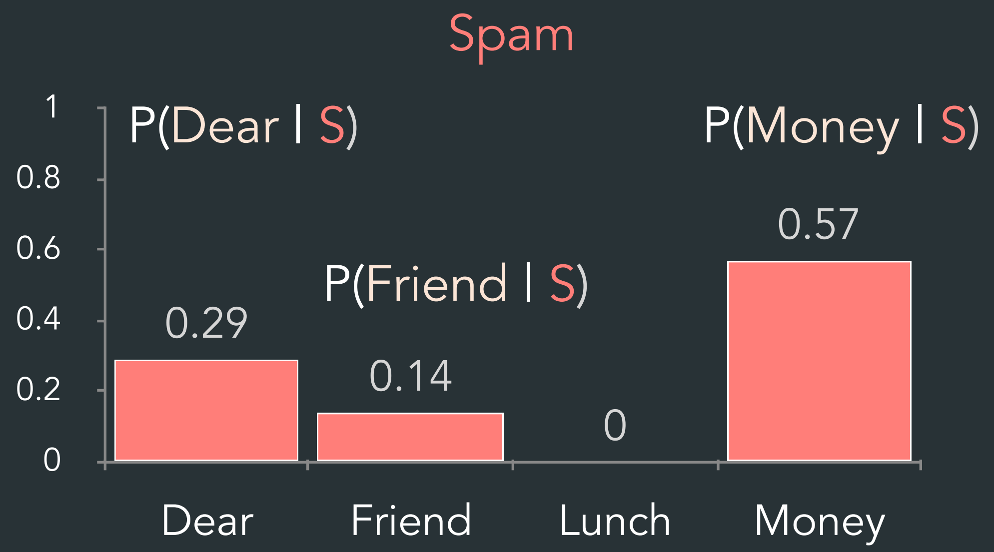


Given the words "Dear Friend" occur in this email, what's the probability it's spam?



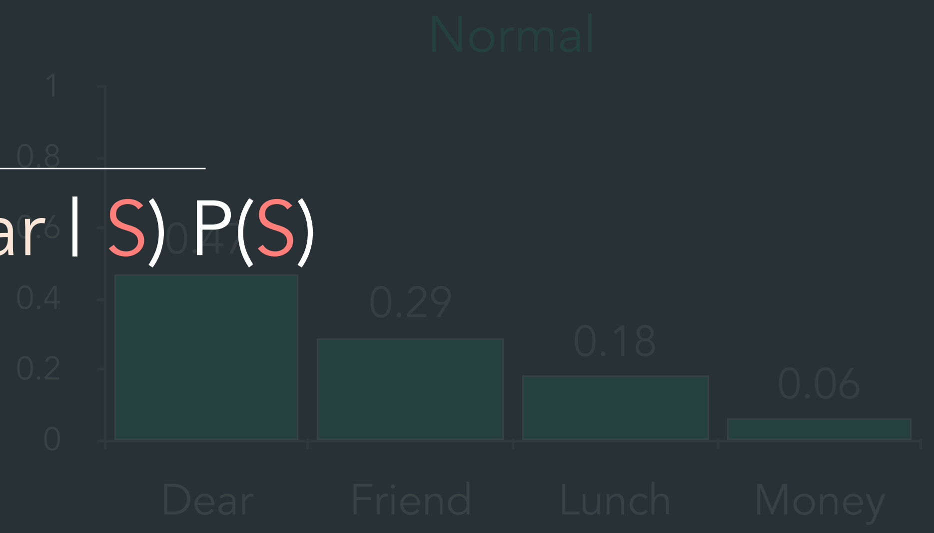


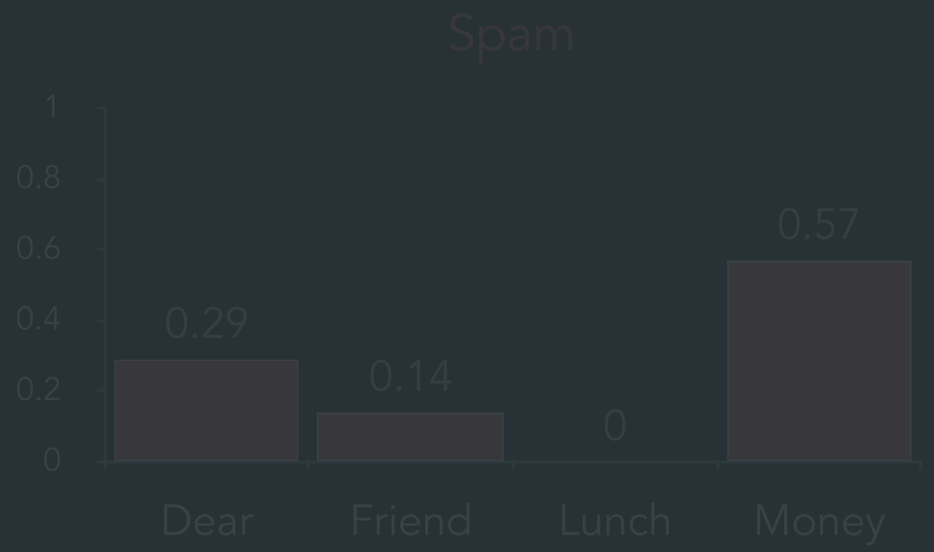
Given the words “Dear Friend” occur in this email, what's the probability it's spam?



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

$$P(S \mid \text{Dear}) = \frac{P(\text{Dear} \mid S)\, P(S)}{P(\text{Dear} \mid N)\, P(N) + P(\text{Dear} \mid S)\, P(S)}$$



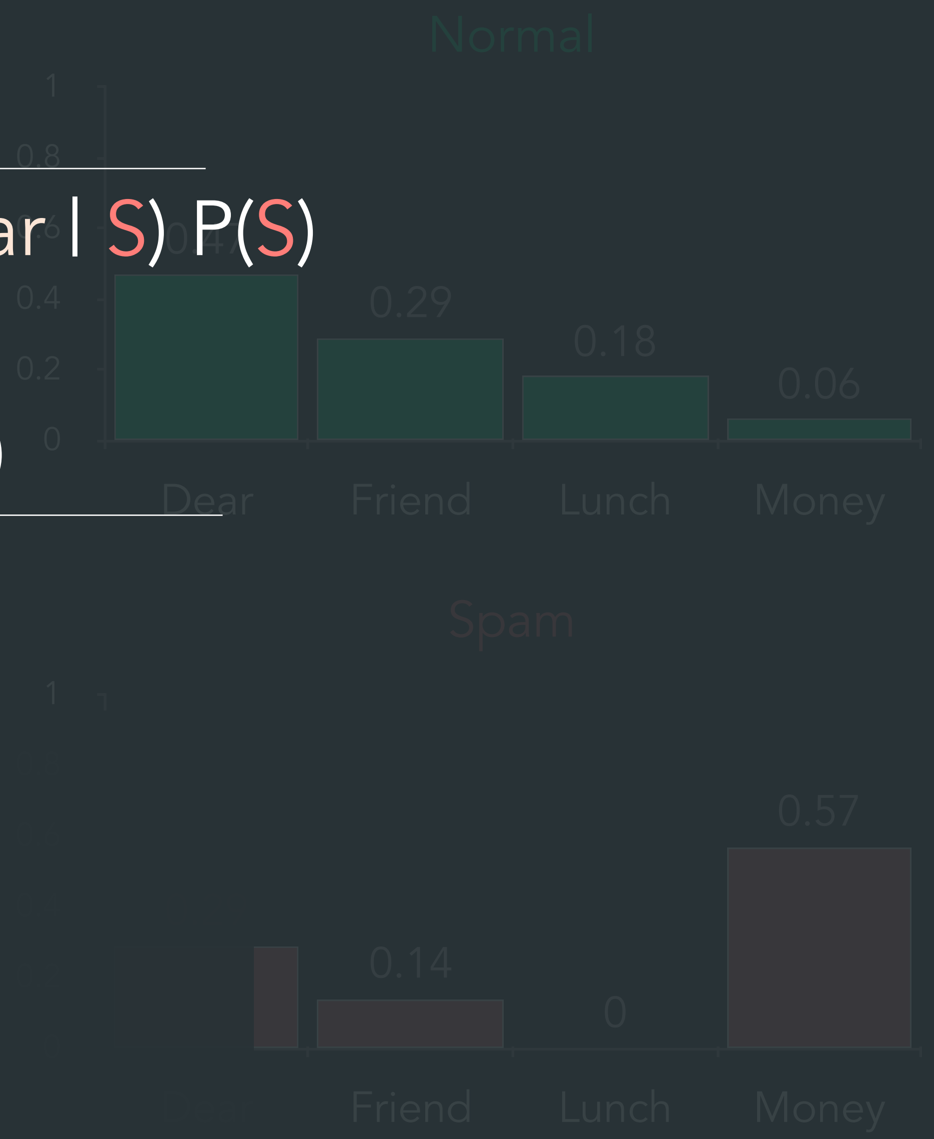


Given the words "Dear Friend" occur in this email, what's the probability it's spam?

$$P(S \mid \text{Dear}) = \frac{P(\text{Dear} \mid S)\,P(S)}{P(\text{Dear} \mid N)\,P(N) + P(\text{Dear} \mid S)\,P(S)}$$

$$P(S \mid \text{Dear}) = \frac{P(\text{Dear} \mid S)\,P(S)}{P(\text{Dear})}$$

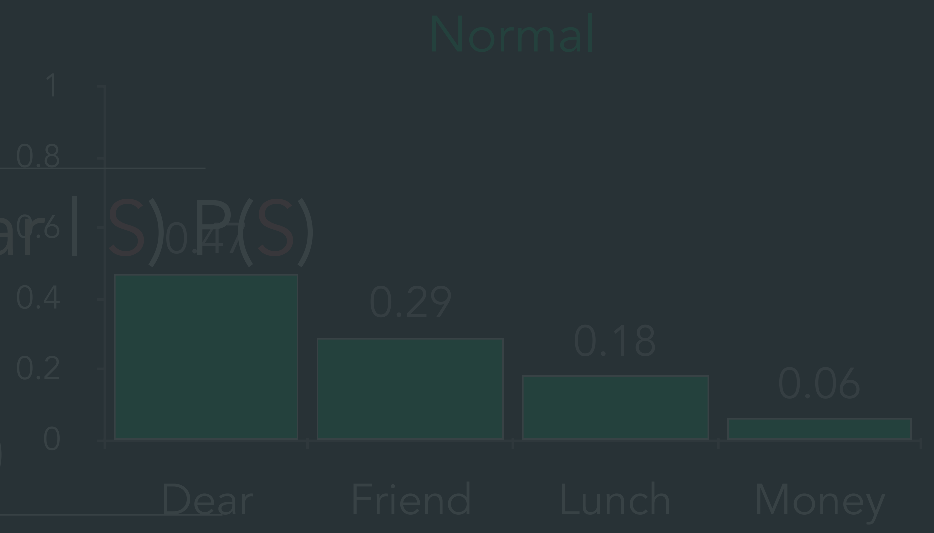
$$P(S \mid \text{Dear}) \propto P(\text{Dear} \mid S)\,P(S)$$

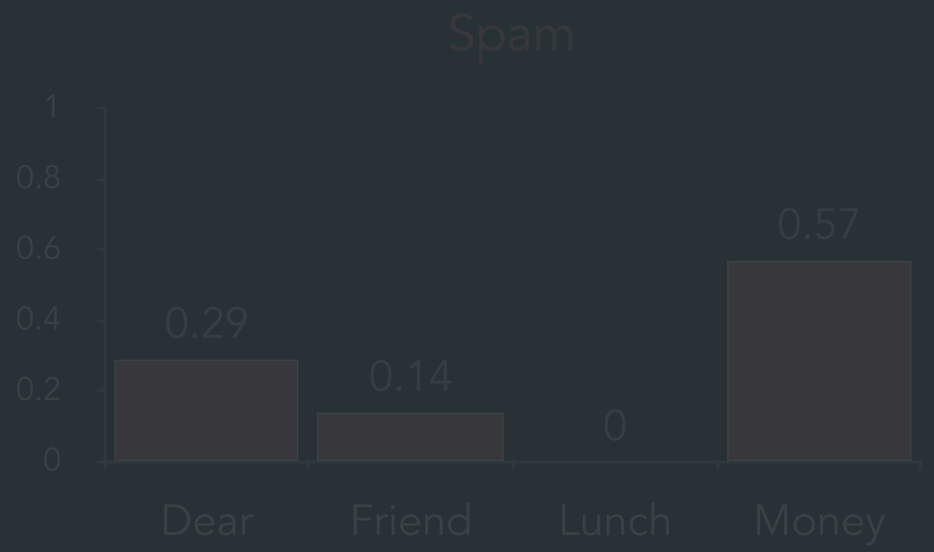


Given the words “Dear Friend” occur in this email, what's the probability it's spam?

$$P(S \mid \text{Dear}) = \frac{P(\text{Dear} \mid S)\,P(S)}{P(\text{Dear} \mid N)\,P(N) + P(\text{Dear} \mid S)\,P(S)}$$

$$P(S \mid \text{Dear}) = \frac{P(\text{Dear} \mid S)\,P(S)}{P(\text{Dear})}$$





$$P(S \mid \text{Dear}) \propto P(\text{Dear} \mid S)\,P(S)$$

Proportional to

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

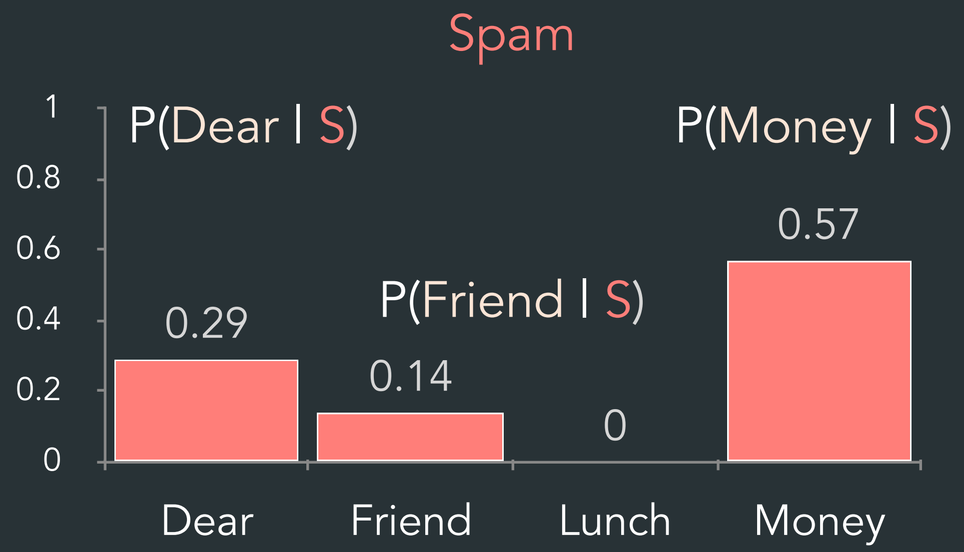
$$P(S \mid \text{Dear Friend}) \stackrel{?}{=}$$

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

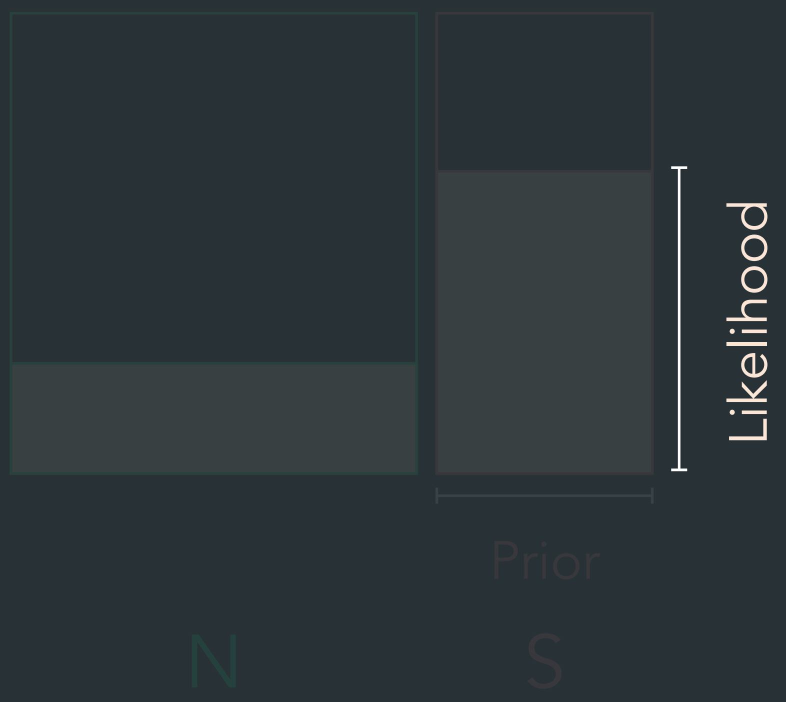
or

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

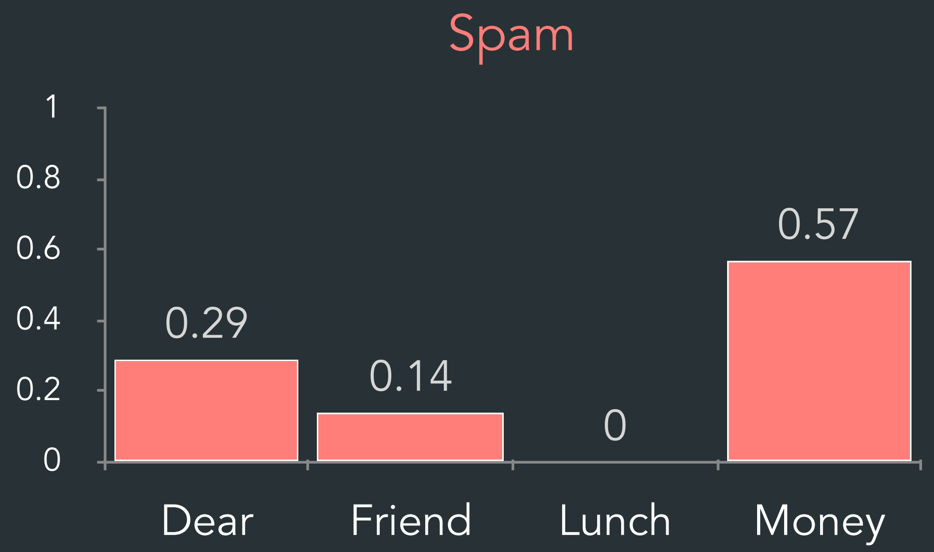
or ??



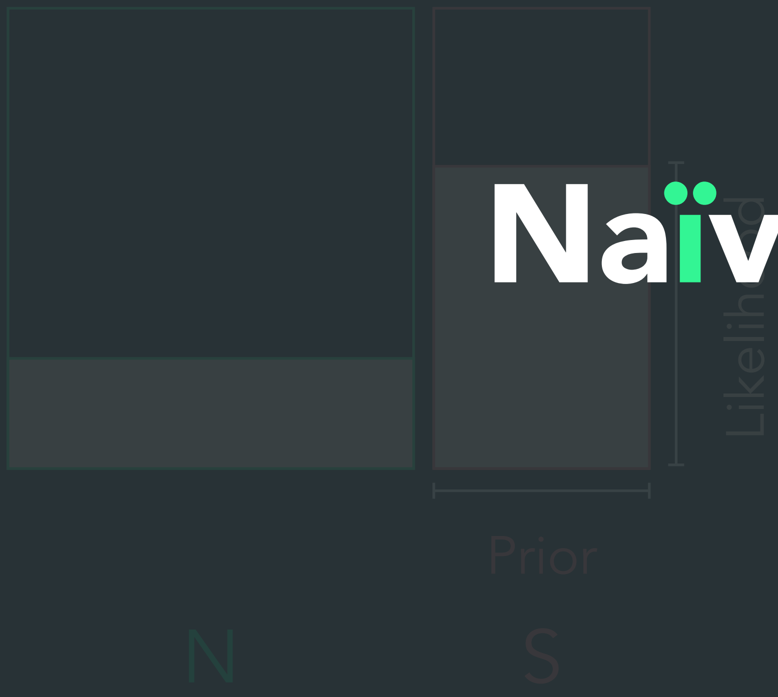
Given the words "Dear Friend" occur in this email, what's the probability it's spam?



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



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



Naïve

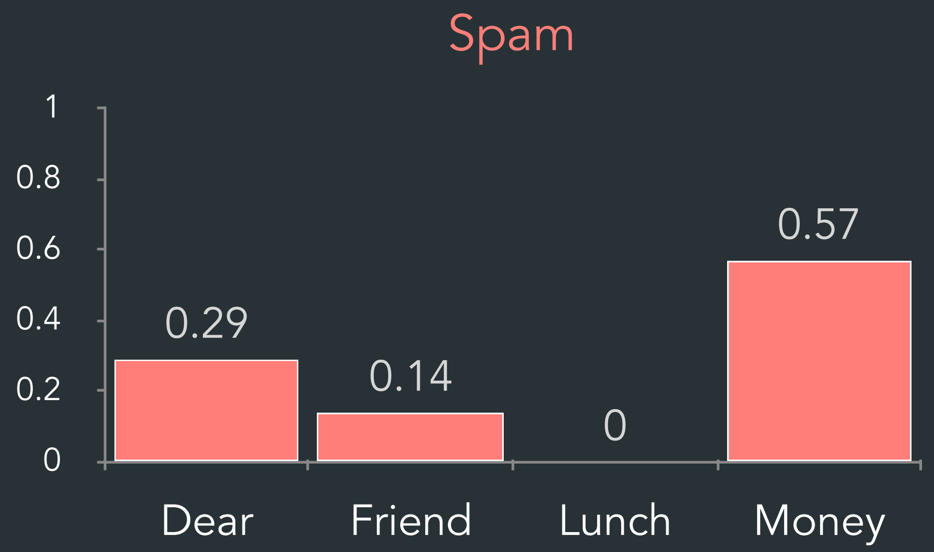
$$P(\text{ Dear Friend} \mid S) \approx P(\text{ Dear} \mid S) \times P(\text{ Friend} \mid S)$$

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

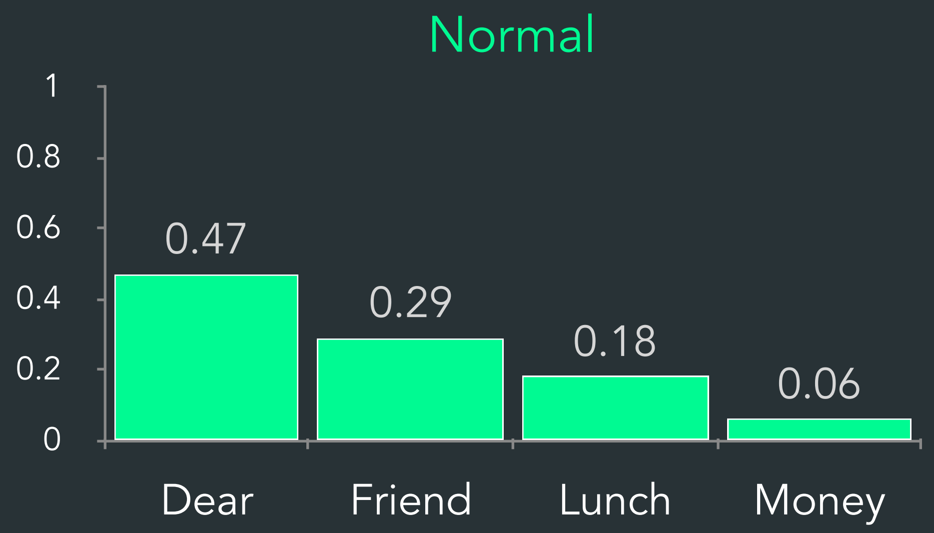
$$P(S \mid \text{Dear Friend}) \propto P(\text{Dear} \mid S) \times P(\text{Friend} \mid S) \times P(S)$$

$$= (0.29) \times (0.14) \times (0.33)$$

$$= 0.0135$$



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

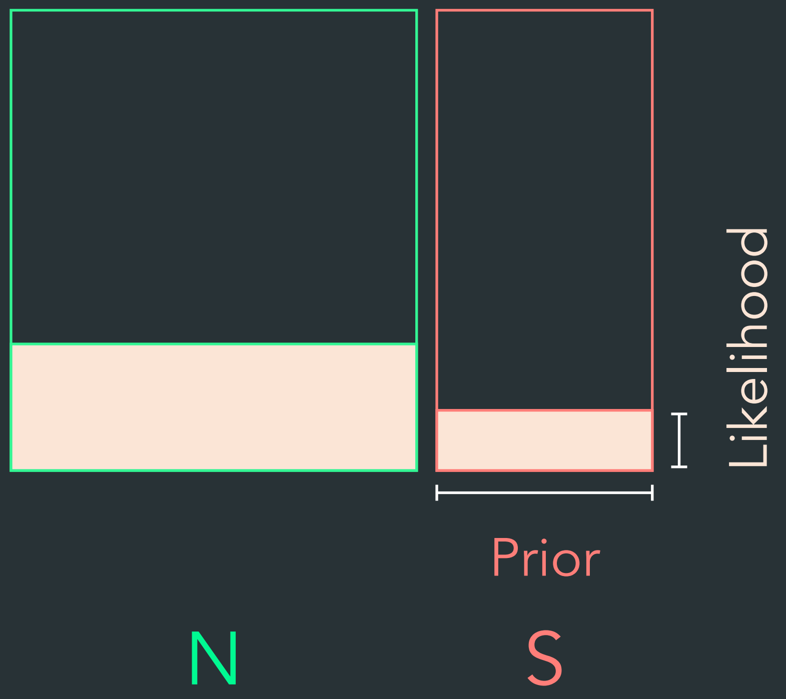


$$P(N \mid \text{Dear Friend}) \propto P(\text{Dear Friend} \mid N) \times P(N)$$

$$\approx P(\text{Dear} \mid N) \times P(\text{Friend} \mid N) \times P(N)$$

$$= (0.47) \times (0.29) \times (0.67) = 0.0909$$

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



$$P(S \mid \text{Dear Friend}) \propto 0.0135$$

$$P(N \mid \text{Dear Friend}) \propto 0.0909$$

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

Classify as N

Alternative: Bayes Factor

Odds $a : b$

Probabilities: $a/(a + b)$

$S : N$

$$B = \frac{P(\text{Dear Friend} \mid S)}{P(\text{Dear Friend} \mid N)}$$

$B \times S : N$

Alternative: Bayes Factor

Odds $a : b$

Probabilities: $a/(a + b)$

$S : N$

$$B = \frac{P(\text{Dear Friend} \mid S)}{P(\text{Dear Friend} \mid N)}$$

$B \times S : N$

Alternative: Bayes Factor

Odds $a : b$ Probabilities: $a/(a + b)$

$S : N$

$$B = \frac{P(\text{Dear Friend} \mid S)}{P(\text{Dear Friend} \mid N)}$$

$B \times S : N$

Alternative: Bayes Factor

Odds $a : b$

Probabilities: $a/(a + b)$

$S : N$

$$B = \frac{P(\text{Dear Friend} \mid S)}{P(\text{Dear Friend} \mid N)} \approx \frac{P(\text{Dear} \mid S) \times P(\text{Friend} \mid S)}{P(\text{Dear} \mid N) \times P(\text{Friend} \mid N)}$$

$B \times S : N$

Alternative: Bayes Factor

Odds $a : b$

Probabilities: $a/(a + b)$

4 Spam Emails

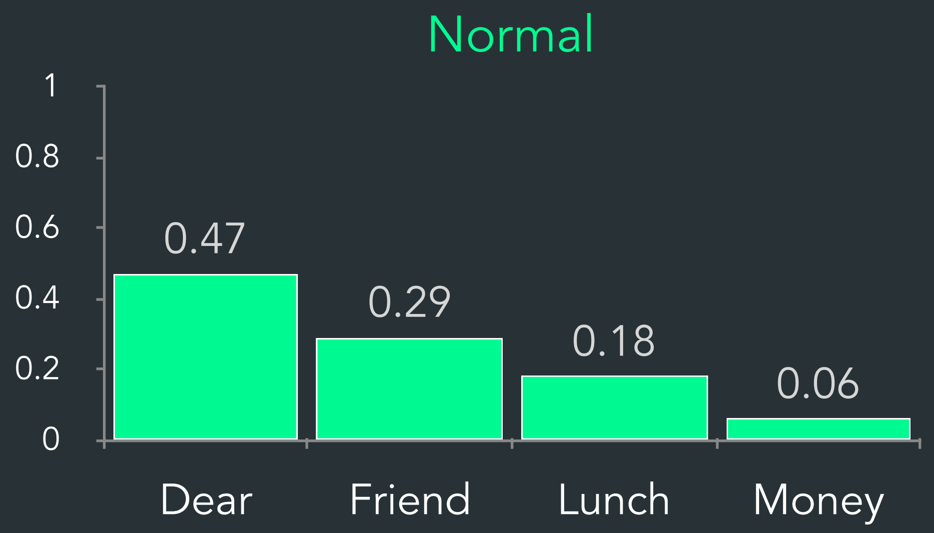
8 Normal Emails

1 : 2

B

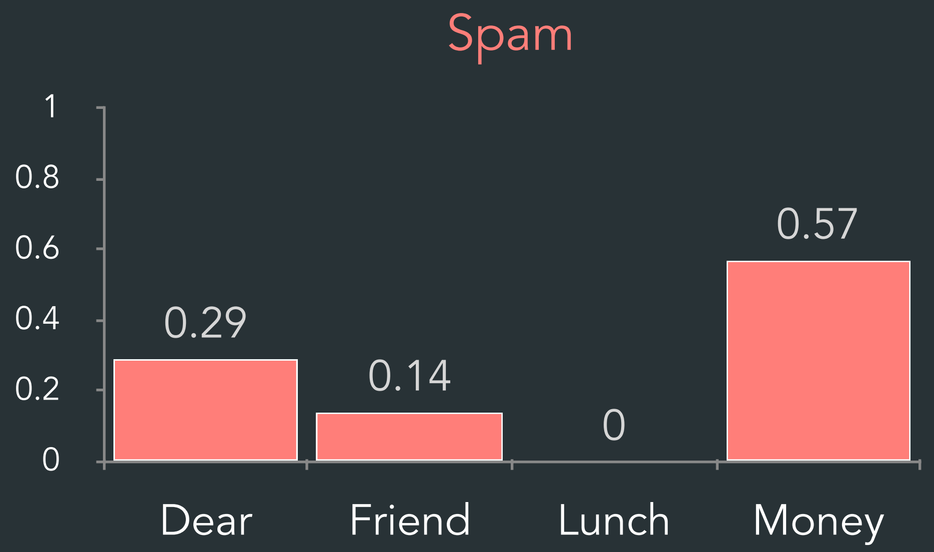
B 1 : 2

Alternative: Bayes Factor



1 : 2

$$B \approx \frac{P(\text{Dear} \mid S) \times P(\text{Friend} \mid S)}{P(\text{Dear} \mid N) \times P(\text{Friend} \mid N)}$$



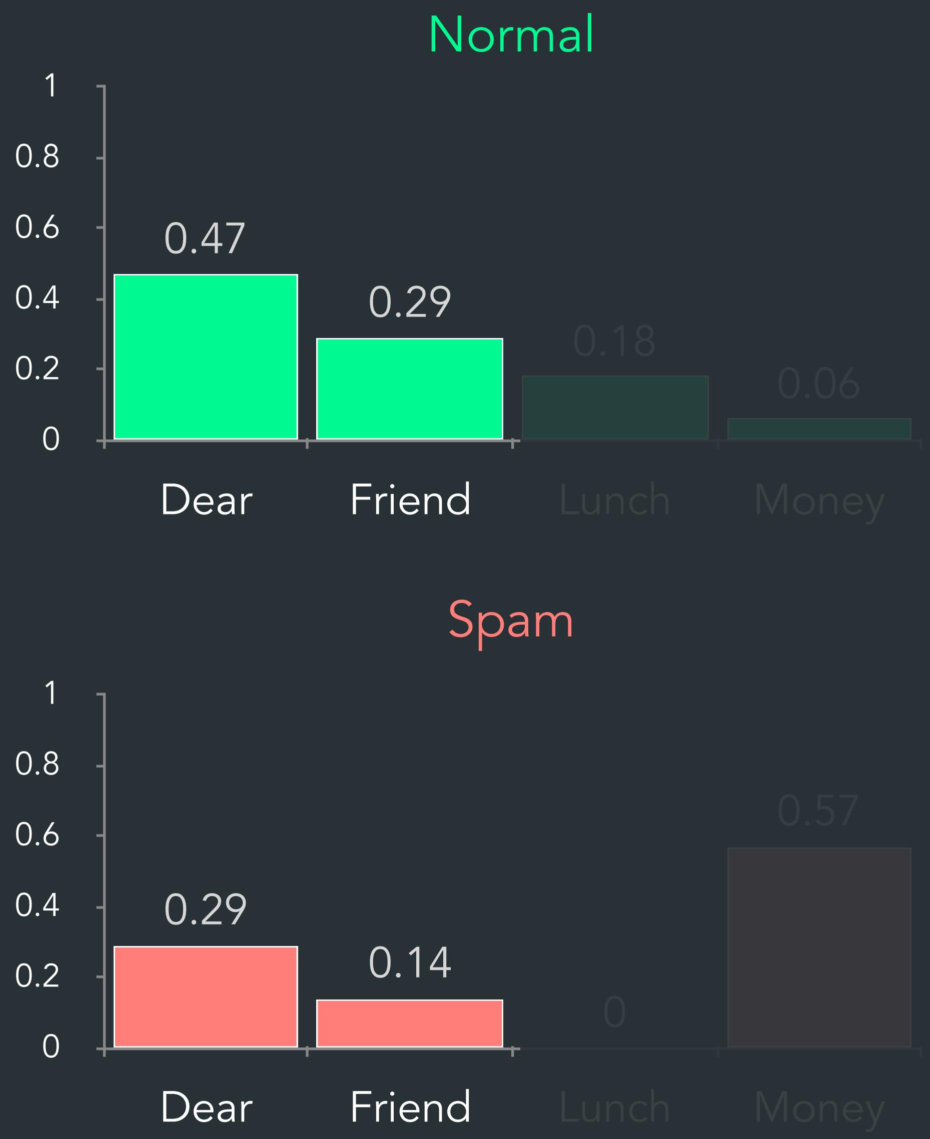
B 1 : 2

Alternative: Bayes Factor

1 : 2

$$B = \frac{(0.29) \times (0.14)}{(0.47) \times (0.29)} = 0.298$$

(0.298)1 : 2



Back to Basics

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

Ep 8

Waiting Times (Continuous Distributions)

Probabilities are Given

Probabilities are not given only data is

Discrete Data

Probabilities are derived using a known distribution

Probabilities are derived by "counting"

Recap | What is the probability that the given coin you picked is C_1?

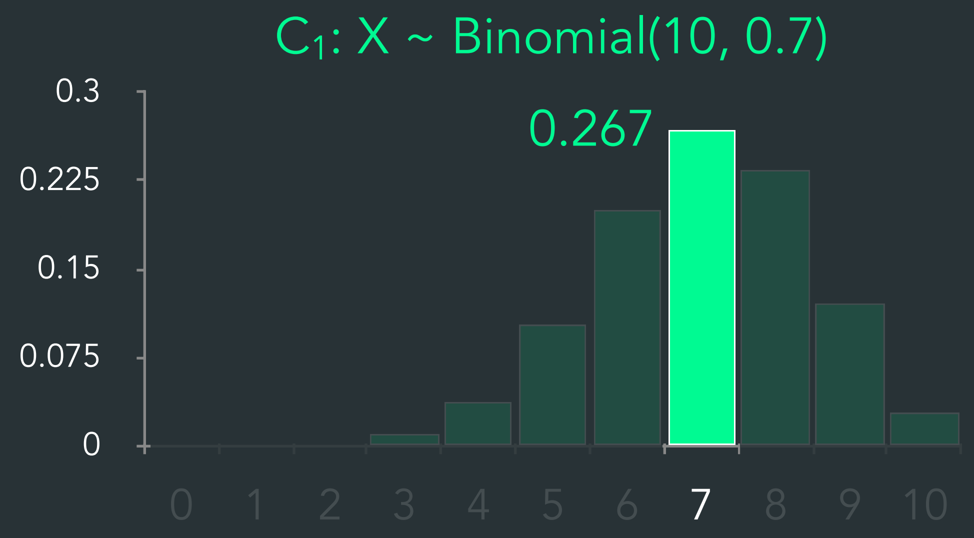
You have 2 coins C_1 and C_2.

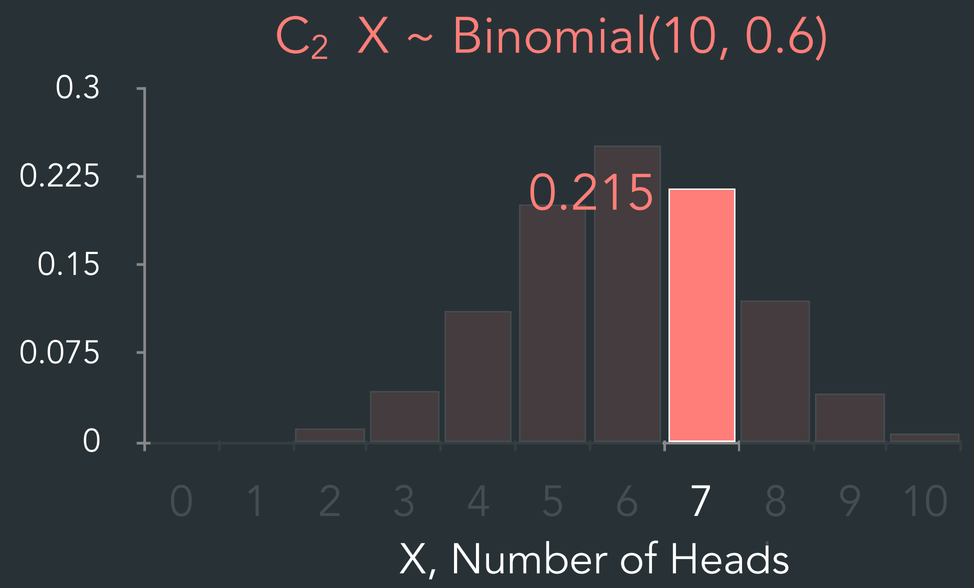
p(heads for C_1) = 0.7

p(heads for C_2) = 0.6

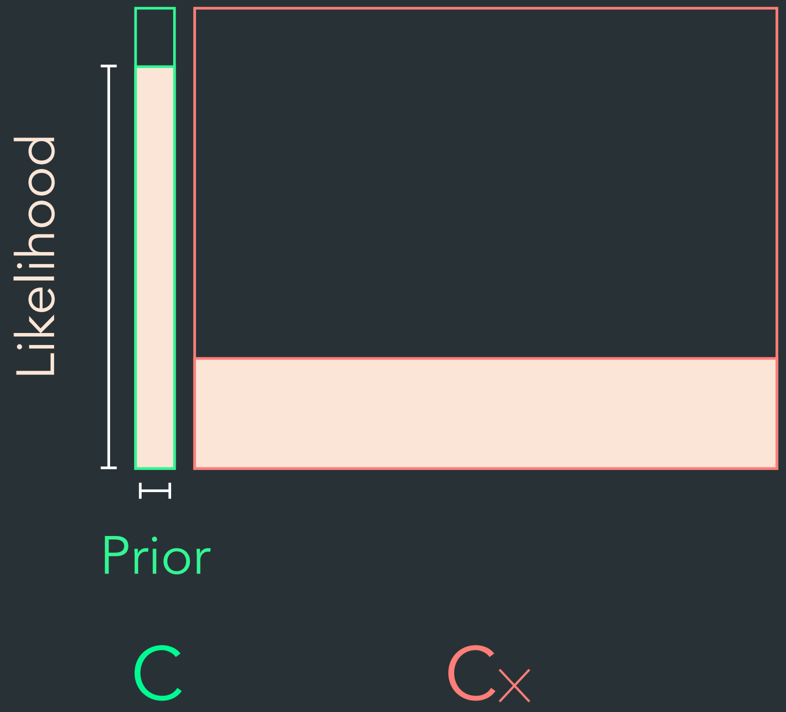
You flip one of the coins 10 times and get **7 heads** and **3 tails**

What is the probability that the given coin you picked is C_1?





Recap | Given a positive test result, what is the probability of having the disease?



$$P(C \mid +) = \frac{P(+|C)\,P(C)}{P(+|C)\,P(C) + P(+|C_x)\,P(C_x)}$$

$\boxed{P(C) = 0.01}$

$P(C_x) = 0.99$

$\boxed{P(+|C) = 0.9}$

$\boxed{P(+|C_x) = 0.05}$

Takeaways

Naive Bayes Algorithm

Bayes Factor for an Alternative Formulation

References

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