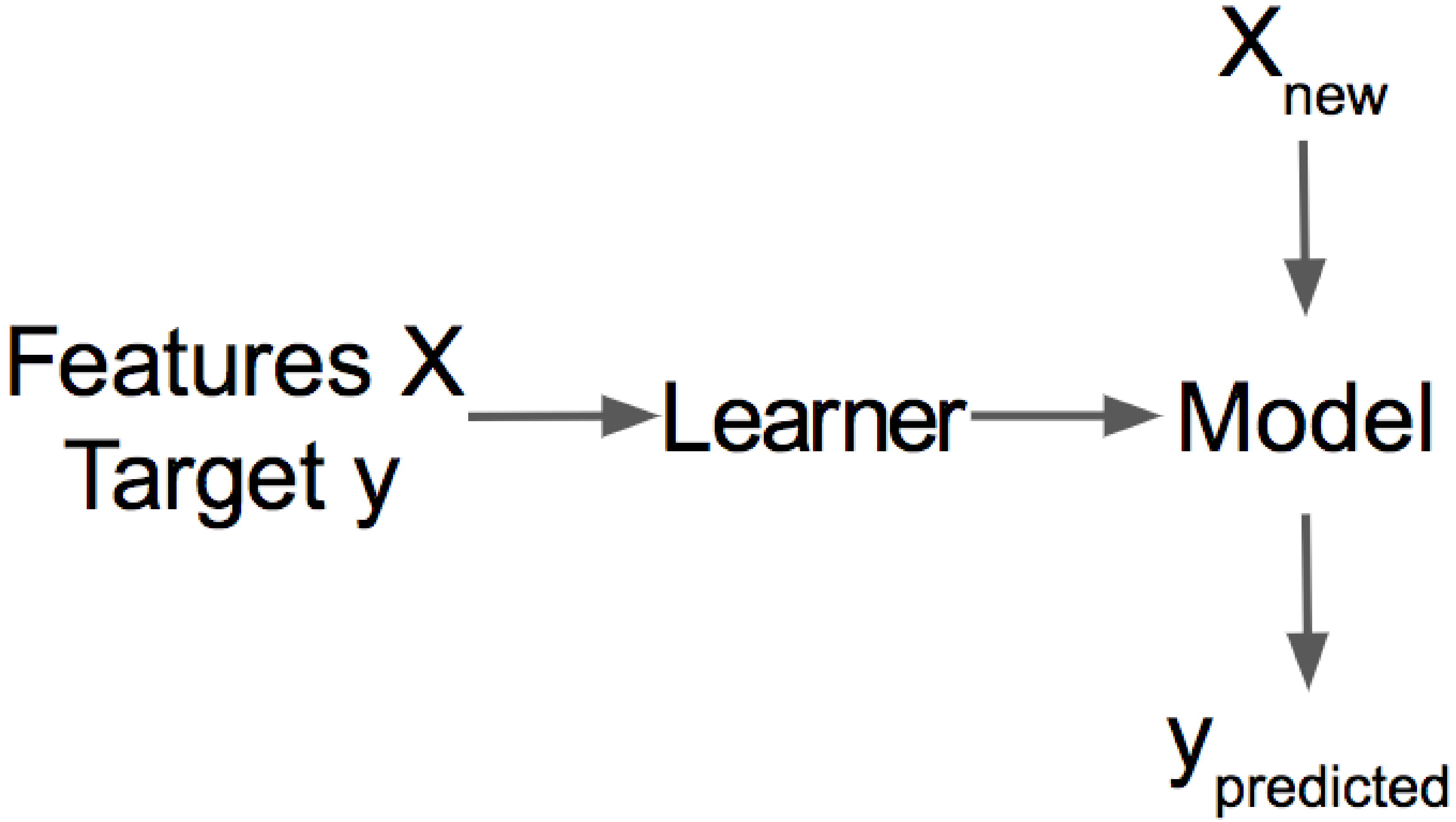
Recap – Naïve Bayes

Rina BUOY

AMERICAN UNIVERSITY
OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

Machine Learning Process



X_{new}

Features X
Target y

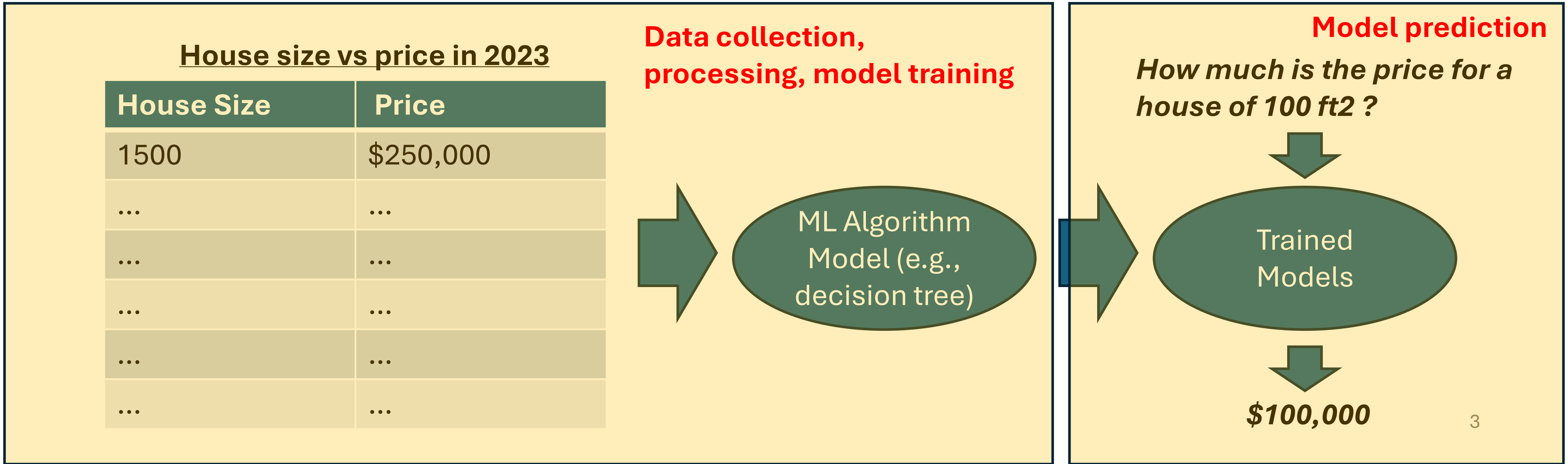
Learner

Model

$y_{predicted}$

Regression Tasks

- Imagine you have a dataset containing information about houses: their sizes (in square feet) and their prices (in dollars). Now, let's say you want to build a model that can predict the price of a house given its size.



House size vs price in 2023

House Size	Price
1500	$250,000
…	…
…	…
…	…
…	…
…	…

Data collection, processing, model training

ML Algorithm Model (e.g., decision tree)

Model prediction

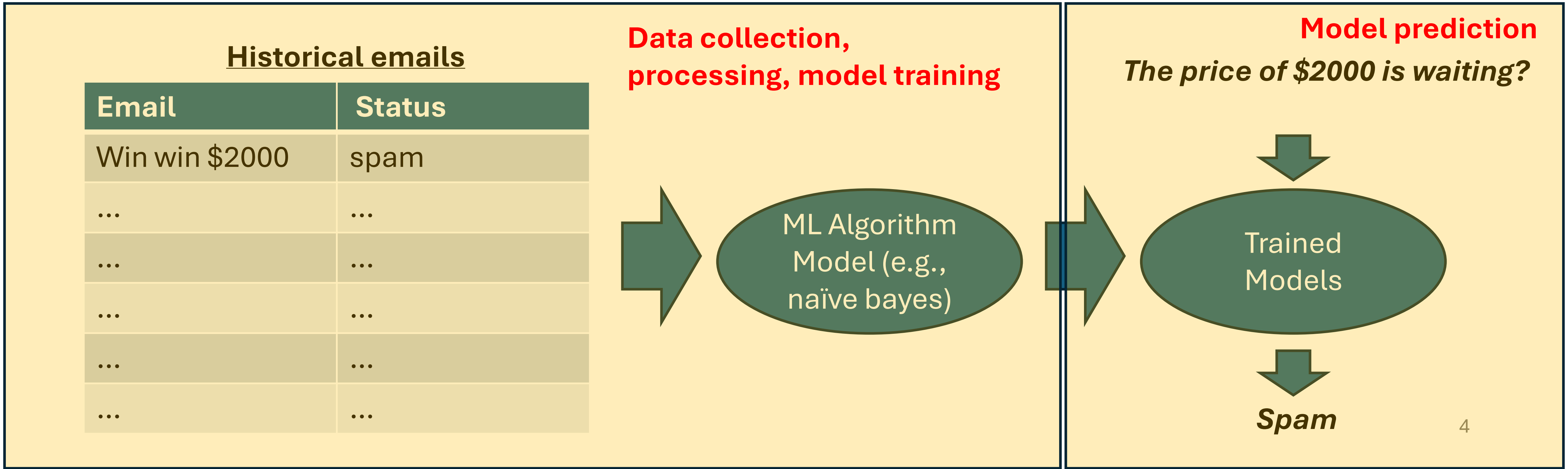
How much is the price for a house of 100 ft2 ?

Trained Models

$100,000

Classification Tasks

- Suppose you have a dataset containing information about emails: their subject lines, body text, and whether they are spam or not (classified as either "spam" or "not spam").



Historical emails

Email	Status
Win win $2000	spam
…	…
…	…
…	…
…	…
…	…

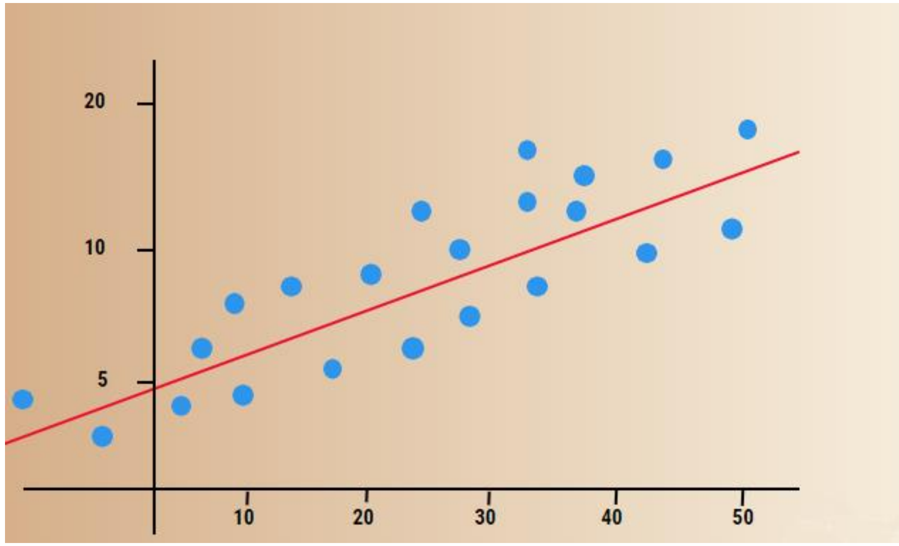
Data collection, processing, model training

Model prediction

The price of $2000 is waiting?

Spam

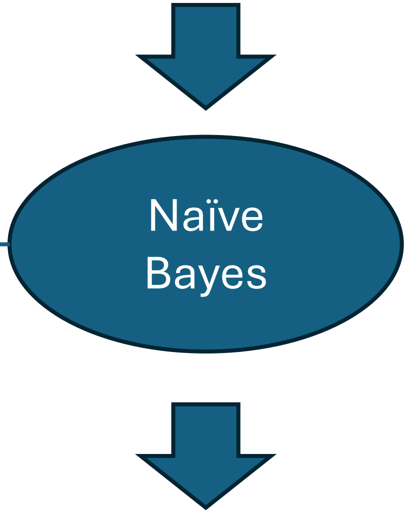
Machine Learning as a mapping function



$$y = f(x)$$

$f(x)$ = a ML algorithm

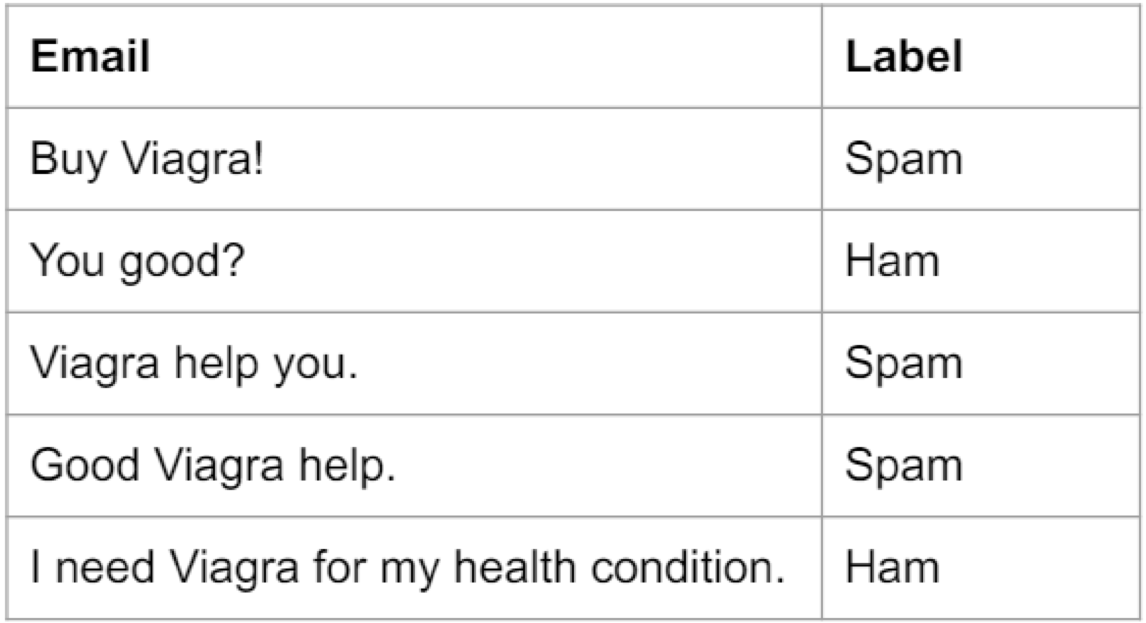
The prize of $2000 is waiting for you !!!



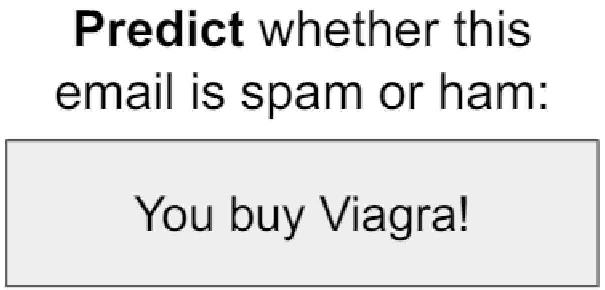
Spam

Spam Filter Task

Training Set



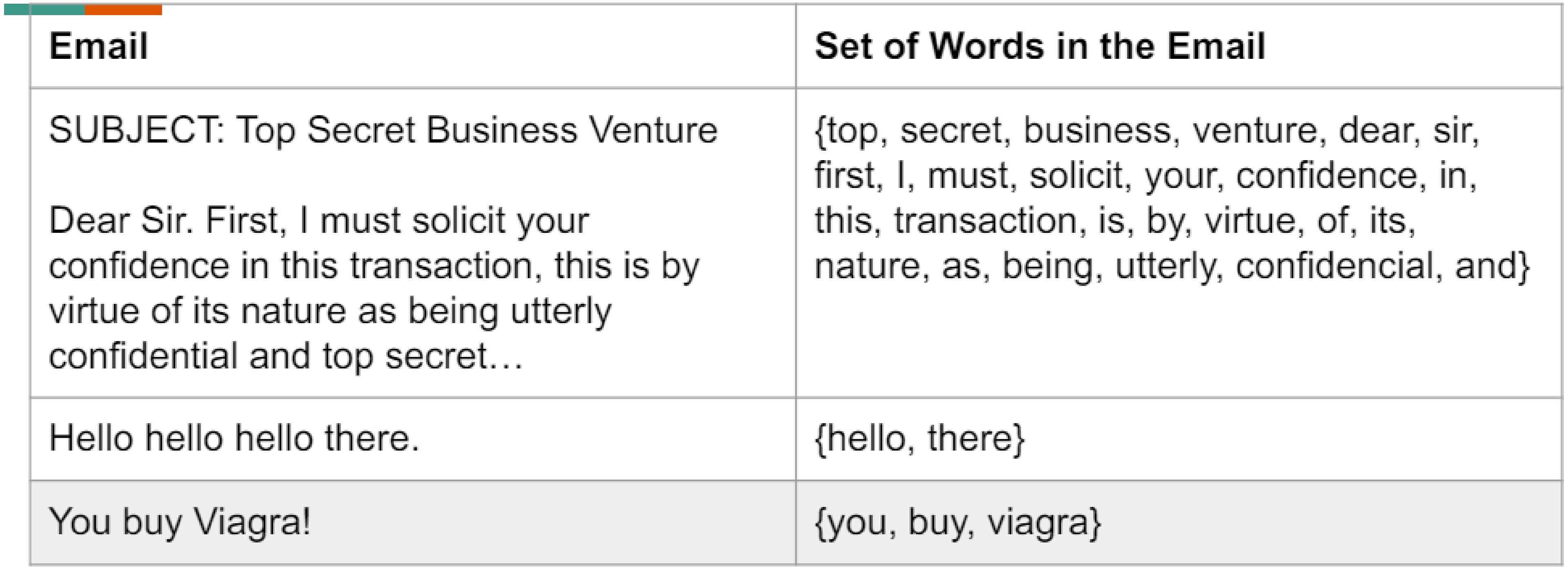
Email	**Label**
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham



Predict whether this email is spam or ham:

You buy Viagra!

Emails as word collections



Email	Set of Words in the Email
SUBJECT: Top Secret Business Venture Dear Sir. First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret…	{top, secret, business, venture, dear, sir, first, I, must, solicit, your, confidence, in, this, transaction, is, by, virtue, of, its, nature, as, being, utterly, confidencial, and}
Hello hello hello there.	{hello, there}
You buy Viagra!	{you, buy, viagra}

For simplicity, we will

- Ignore Duplicate Words
- Ignore Punctuation
- Ignore Casing

Idea

Compute and Compare:

$$\mathbb{P}(\text{spam} \mid \text{"You buy Viagra!"}) \qquad \mathbb{P}(\text{ham} \mid \text{"You buy Viagra!"})$$

Then predict whichever is larger! Can we get away with just computing one of them?

Equivalently, note that these add to 1, so we can just compute $\mathbb{P}(\text{spam} \mid \text{"You buy Viagra!"})$

and if it is greater than 0.5, then we predict **spam**.

Otherwise, we predict **ham**.

Note: We resolve the tie in favor of **ham**.

Naive Bayes Classifier - The naive part

It is somewhat unlikely that we have the email ”You buy Viagra!” in our training data. (In this case we don’t!)

We **naively assume that words are conditionally independent from each other, given the label (In reality, they aren’t):**

$$\mathbb{P}(\{\text{“you”},\ \text{“buy”},\ \text{“viagra”}\} \mid \text{spam})$$

$$\approx \mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})$$

Then we estimate for example that

$$\mathbb{P}(\text{“you”} \mid \text{spam}) = \frac{\text{number of spam emails containing “you” (in training set)}}{\text{number of spam emails (in training set)}}$$

Example

$$\mathbb{P}\left(\text{spam} \mid \text{“You buy Viagra”}\right)$$

$$= \frac{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{ham})\ \mathbb{P}(\text{ham})}$$

$$= \frac{\mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})\mathbb{P}(\text{spam})}{\mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})\mathbb{P}(\text{spam}) + \mathbb{P}(\text{“you”} \mid \text{ham})\mathbb{P}(\text{“buy”} \mid \text{ham})\mathbb{P}(\text{“viagra”} \mid \text{ham})\mathbb{P}(\text{ham})}$$

$$= \frac{\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}}{\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{2}{5}} \approx 0.7544$$

Email	Label
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

$$\mathbb{P}(\text{spam}) = \frac{3}{5} \qquad \mathbb{P}(\text{ham}) = \frac{2}{5}$$

$$\mathbb{P}(\text{“you”} \mid \text{spam}) = \frac{1+1}{3+2} = \frac{2}{5} \qquad \mathbb{P}(\text{“you”} \mid \text{ham}) = \frac{1+1}{2+2} = \frac{1}{2}$$

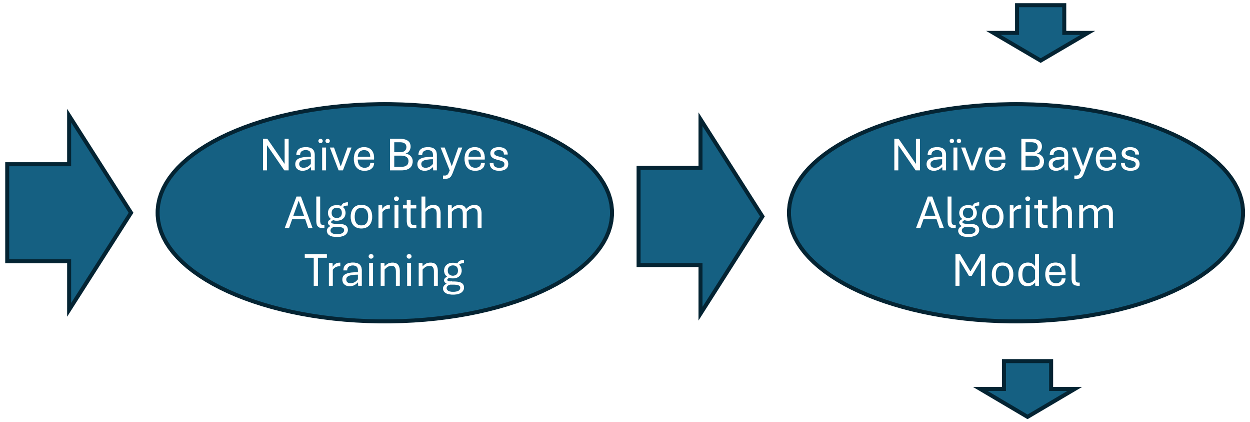
$$\mathbb{P}(\text{“buy”} \mid \text{spam}) = \frac{1+1}{3+2} = \frac{2}{5} \qquad \mathbb{P}(\text{“buy”} \mid \text{ham}) = \frac{0+1}{2+2} = \frac{1}{4}$$

$$\mathbb{P}(\text{“viagra”} \mid \text{spam}) = \frac{3+1}{3+2} = \frac{4}{5} \qquad \mathbb{P}(\text{“viagra”} \mid \text{ham}) = \frac{1+1}{2+2} = \frac{1}{2}$$

Think about Khmer fake facebook post detection

Post	Status
5ឆ្នាំទៀតខ្ញុំចង់!និយាយមិនអោយខុស់	OKAY
... #តារាចម្រៀង #សិល្បៈ #កម្សាន្ត #ព័ត៌មានថ្មីៗ	NOT OKAY
...	...
...	...
...	...

#លោកតាស្លូង



P(OKAY| #លោកតាស្លូង)

vs.

P(NOT OKAY| #លោកតាស្លូង)

Other cases

- Think about news article classification ?
- Think about sentiment analysis ?
- Think about product reviews ?
- Etc.