Practical machine learning

NAÏVE BAYES CLASSIFIER

CLASSIFICATION

UNLABELED DATA VS LABELED DATA

Unlabeled data:

Consists of natural or human-created artefacts that can obtain easily from the world.

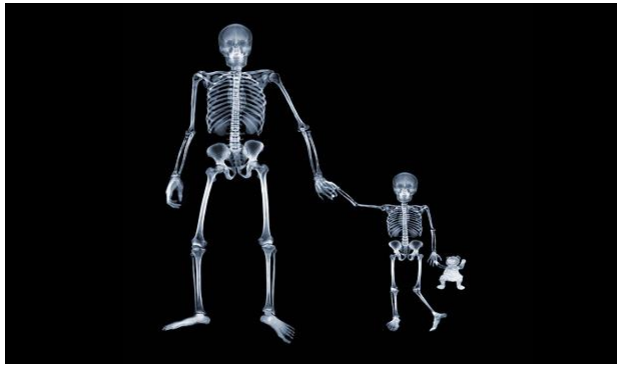
Examples:

Photos, audio recordings, videos, news articles, tweets, x-rays etc.

No "explanation" for each piece of unlabelled data, just contains the data, nothing else.

14 March 2015 Last updated at 11:43

Statue of Mahatma Gandhi unveiled in Parliament Square



UNLABELED DATA VS LABELED DATA

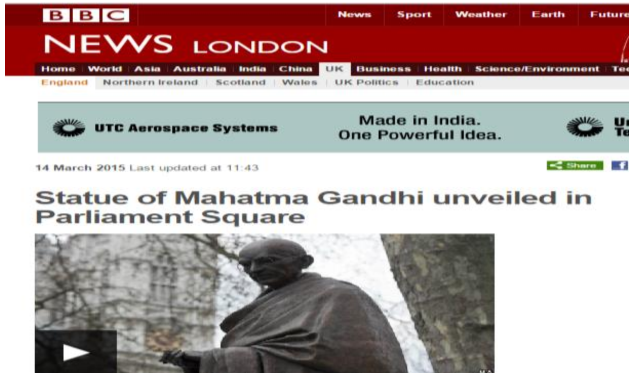
Labelled Data:

Take a set of unlabelled data and augments each piece of that unlabelled data with meaningful "tag," "label," or "class" that is somehow informative.

Example:

(a) whether this **photo contains a horse or a cow**,
(b) what the topic of this news article is,
(c) Whose X-Ray is this, etc.

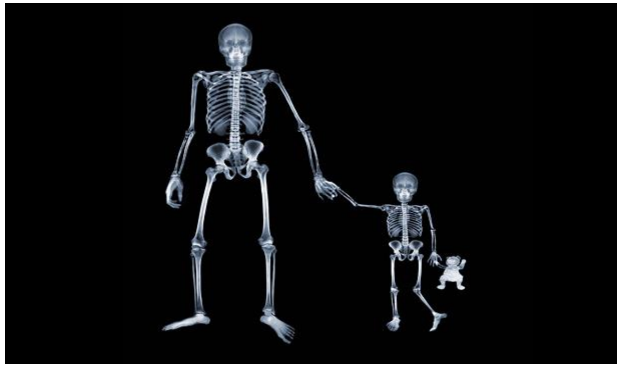
A Cow



14 March 2015 Last updated at 11:43

Statue of Mahatma Gandhi unveiled in Parliament Square

Statue of Mahatma Gandhi Unveiled in Parliament Square



X-Ray of Mr. James & His Son

WHAT IS CLASSIFICATION?

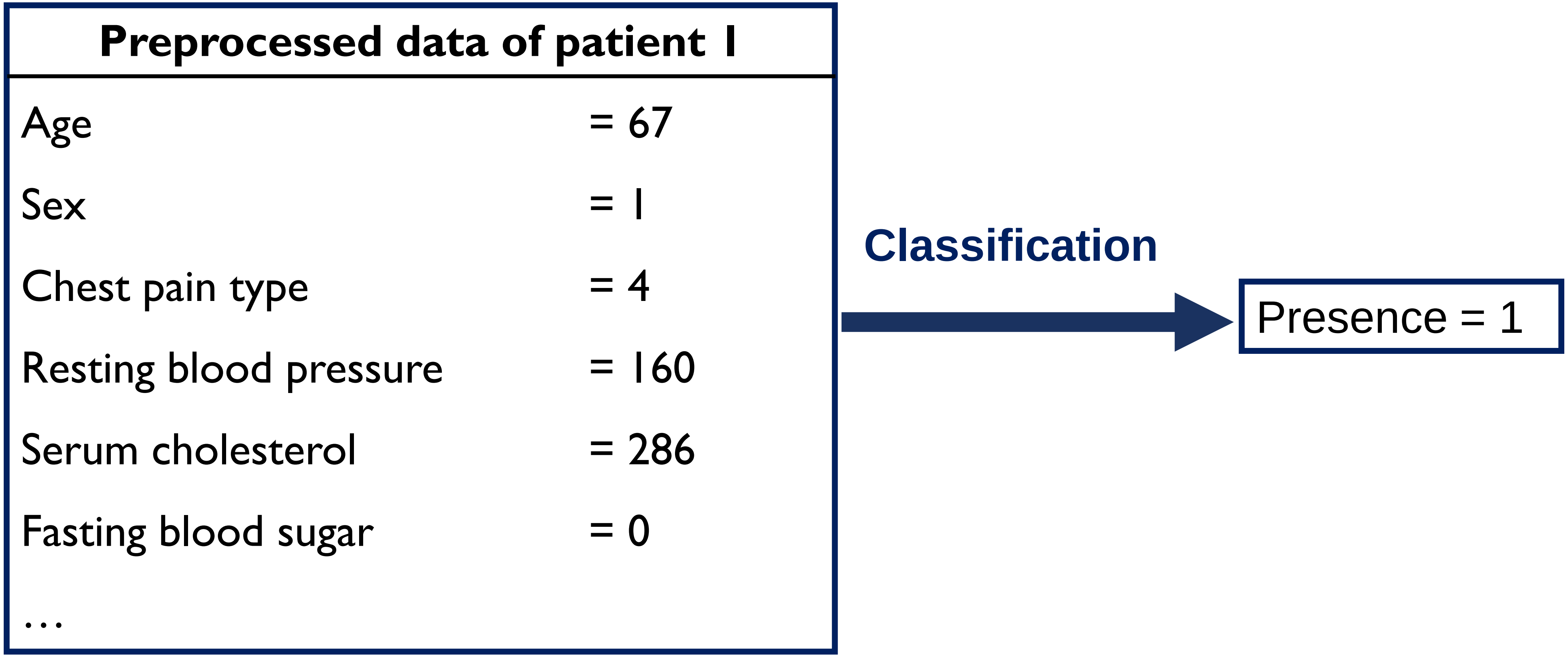
Definition: Identifying to which of a set of categories (sub-populations) a new observation belongs.

- on the basis of a training set of data containing observations (or instances) whose category membership (Class label) is known.

- Data: A set of data records (also called examples, instances or cases) described by
  - *k* attributes: $A_1, A_2, \ldots A_k$.
  - a class: Each example is labelled with a pre-defined class.

- Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

EXAMPLE1 – CLASSIFY A DISEASE



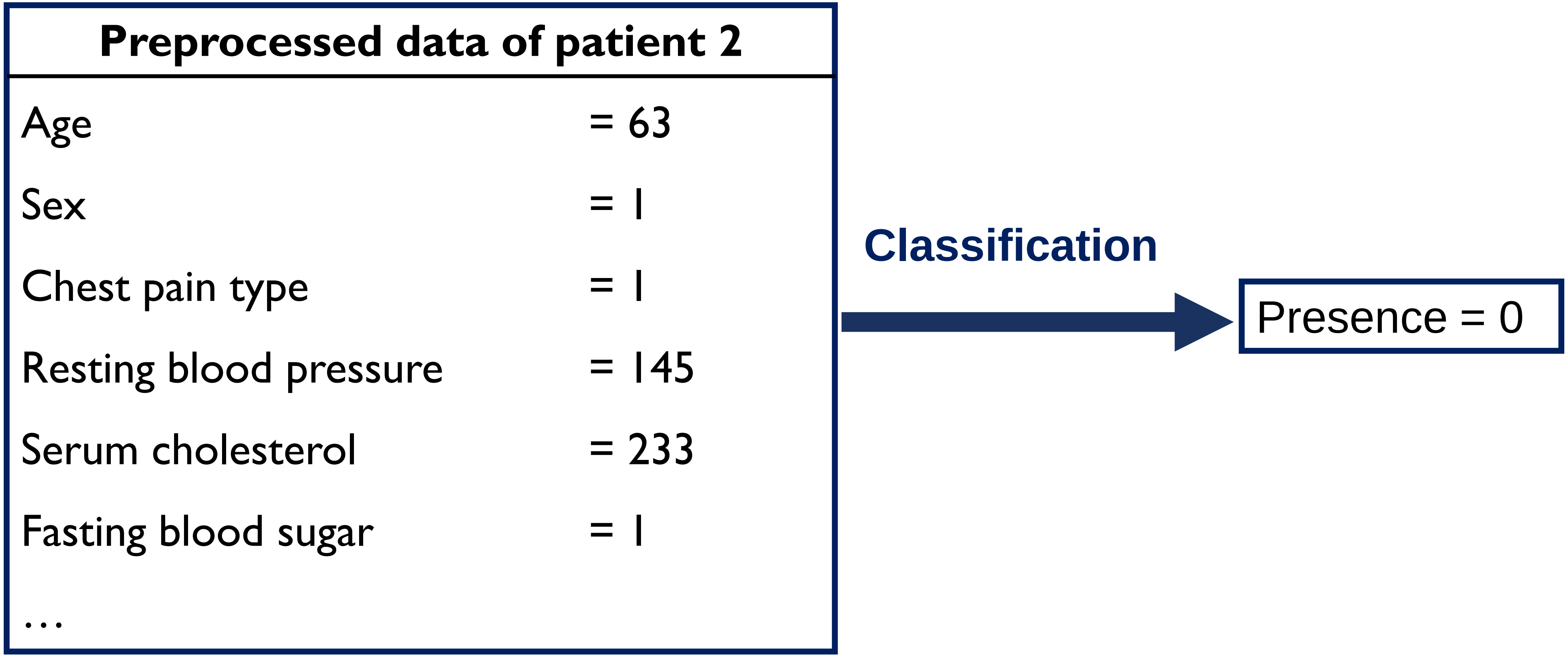
Preprocessed data of patient 1

Age	= 67
Sex	= 1
Chest pain type	= 4
Resting blood pressure	= 160
Serum cholesterol	= 286
Fasting blood sugar	= 0
…	

Classification

Presence = 1

EXAMPLE1 – CLASSIFY A DISEASE

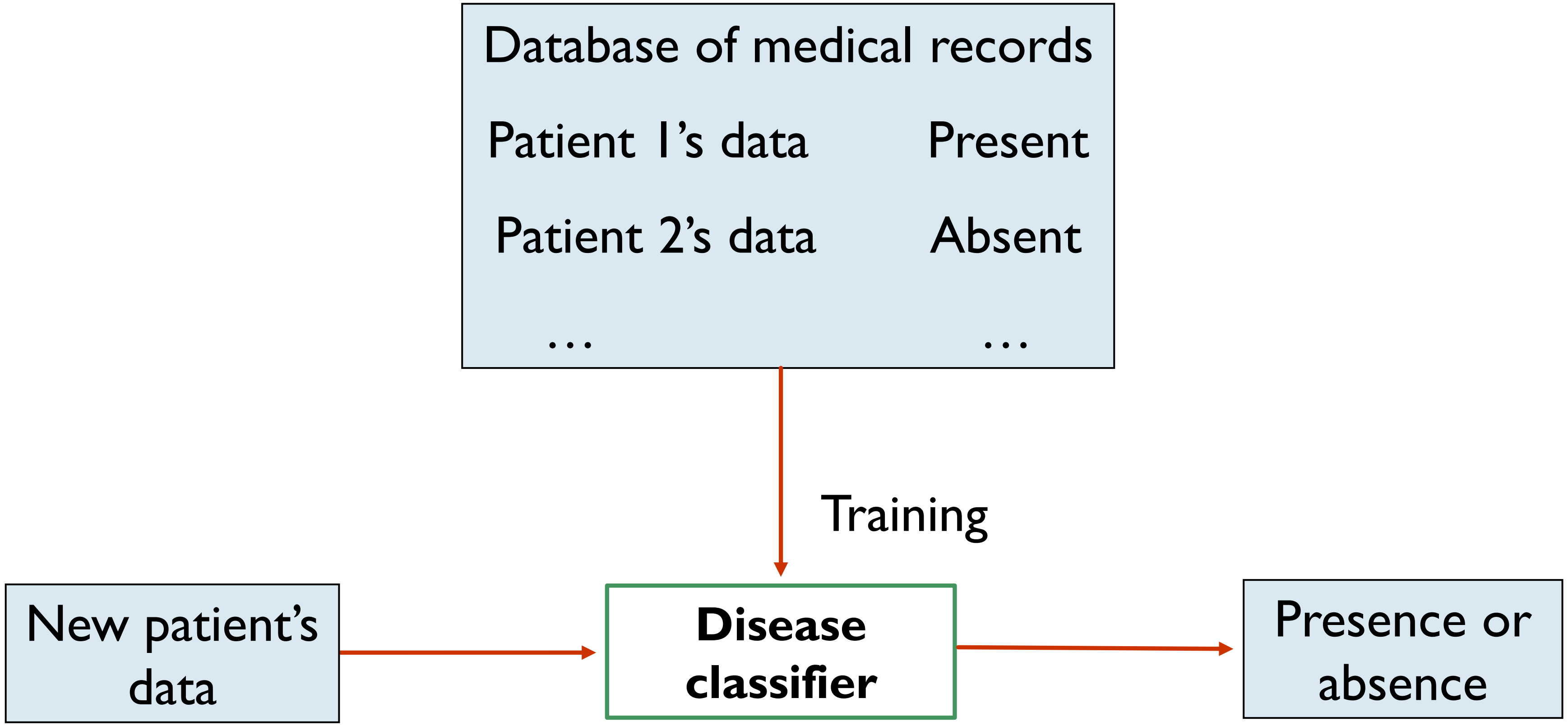


Preprocessed data of patient 2	
Age	= 63
Sex	= 1
Chest pain type	= 1
Resting blood pressure	= 145
Serum cholesterol	= 233
Fasting blood sugar	= 1
…	

Classification

Presence = 0

EXAMPLE1 – CLASSIFY A DISEASE – MACHINE LEARNING



Database of medical records

Patient 1's data Present

Patient 2's data Absent

… …

Training

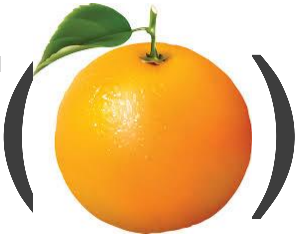
New patient's data

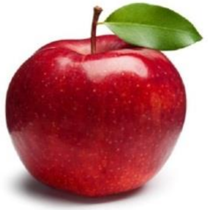
Disease classifier

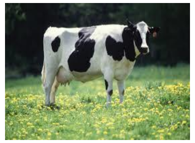
Presence or absence

EXAMPLE2 – IMAGE CLASSIFICATION

Apply a **prediction function** to a **feature representation** of the image to get the desired output:

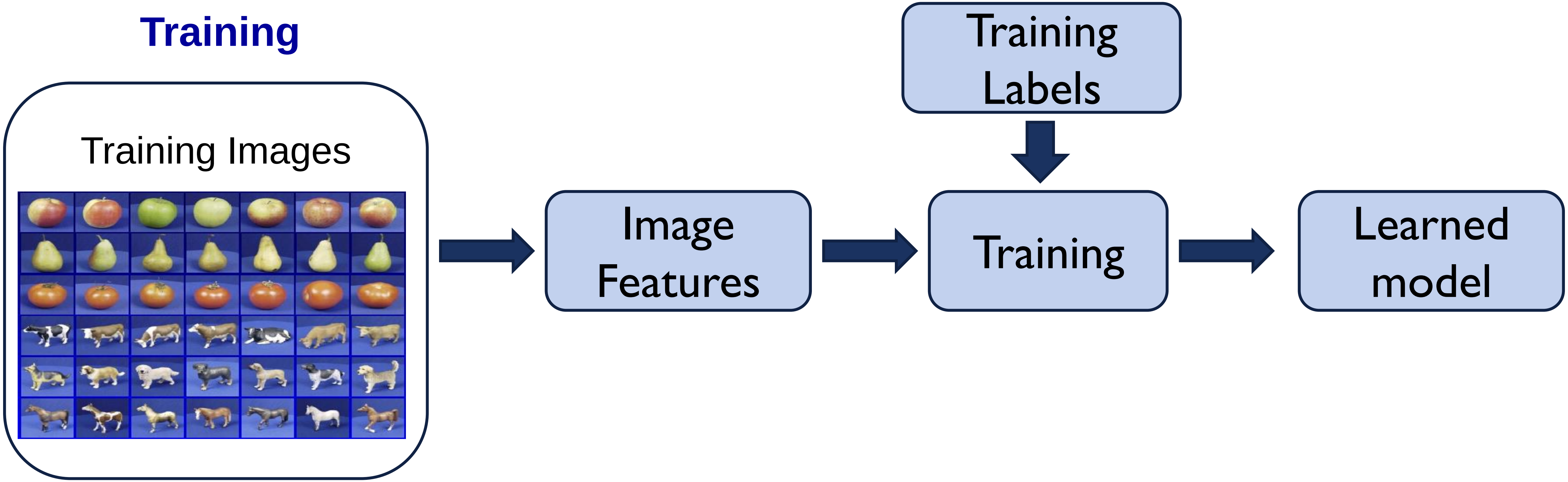
f() = “Orange”

f() = “Apple”

f() = “Cow”

EXAMPLE2 - MACHINE LEARNING FRAMEWORK

Training



Training Images

Image Features

Training Labels

Training

Learned model

Testing

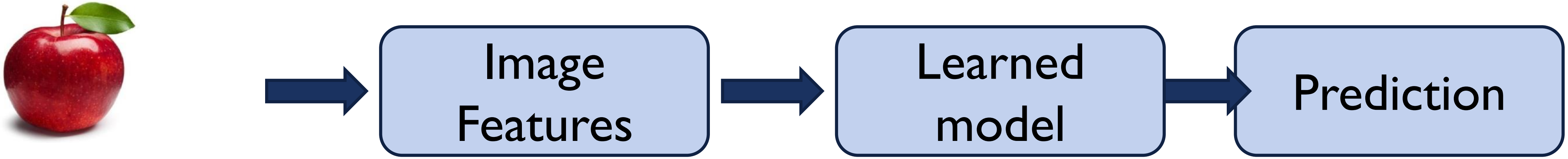


Image Features

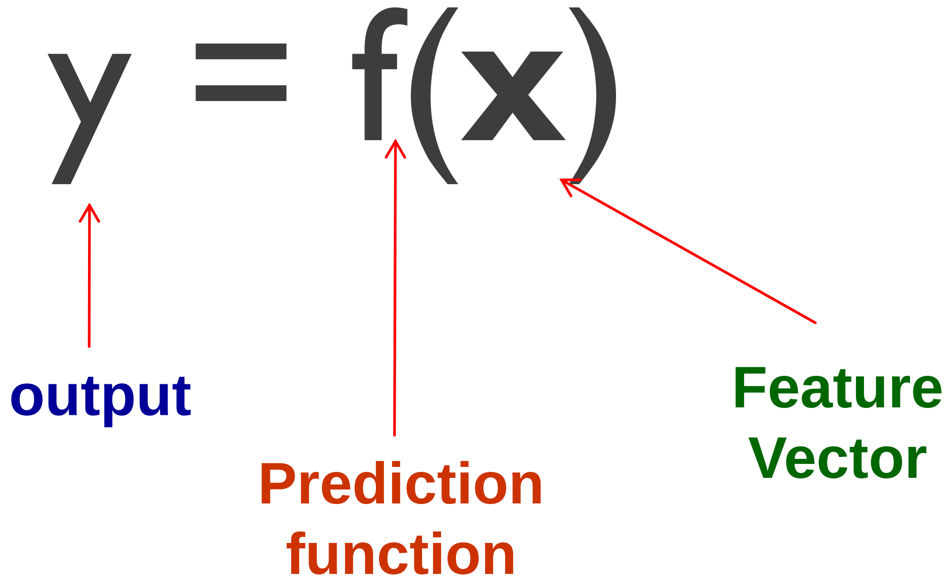
Learned model

Prediction

Test Image

Slide credit: D. Hoiem and L. Lazebnik

THE MACHINE LEARNING FRAMEWORK



$$y = f(\mathbf{x})$$

output

Prediction function

Feature Vector

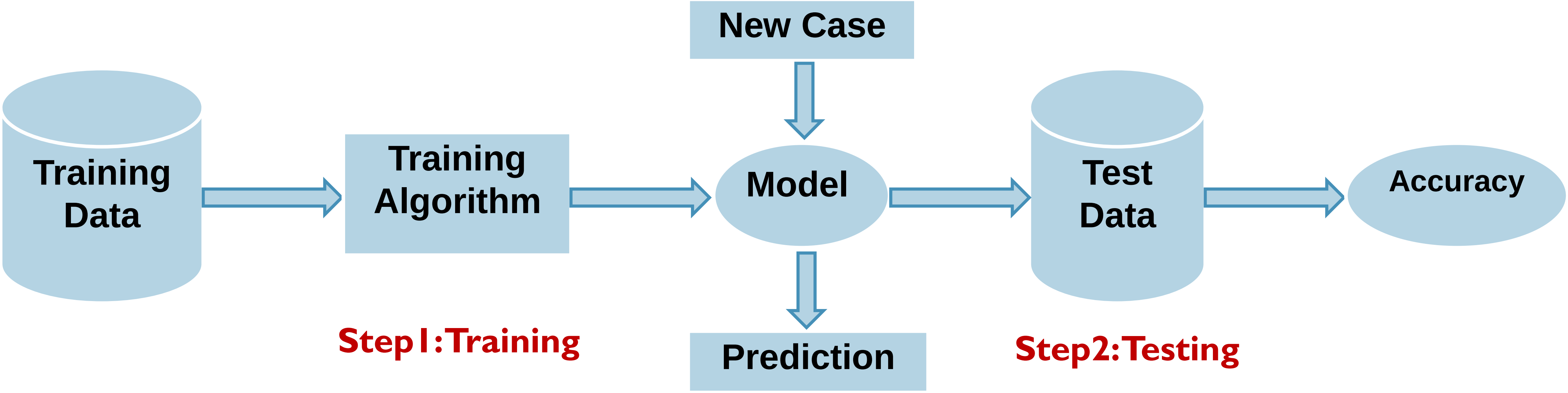
- **Training:** Given a training set of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function "f" by minimizing the prediction error on the training set.
- **Testing:** apply "f" to a never before seen test example "x" and output the predicted value $y = f(x)$.

Slide credit: L. Lazebnik

SUPERVISED LEARNING – TWO STEP PROCESS

- Learning (Training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$



ALGORITHMS

- Naïve Bayesian classification
- Decision tree induction
- Random Forest
- Gradient Boosting
- Support vector machines (SVM)
- K-nearest neighbor (KNN)

NAÏVE BAYES CLASSIFIER

WHAT WE LEARN?

- Basic principles of probability
- Understanding Naïve Bayes
- Basic concepts of Bayesian methods
- Conditional probability with Bayesian theorem
- Naïve Bayes Algorithm
- Case Study

70 % probability of rain today evening

Use past data to predict future events

A 70 percent chance of rain implies that in 7 out of the 10 past cases with similar conditions, precipitation occurred somewhere in the area.

UNDERSTANDING NAÏVE BAYES

- The technique descended from the work of the 18th century mathematician **Thomas Bayes**.
- Developed foundational principles to describe the probability of events, and how probabilities should be revised in the light of additional information.
- These principles formed the foundation for what are now known as **Bayesian methods**.

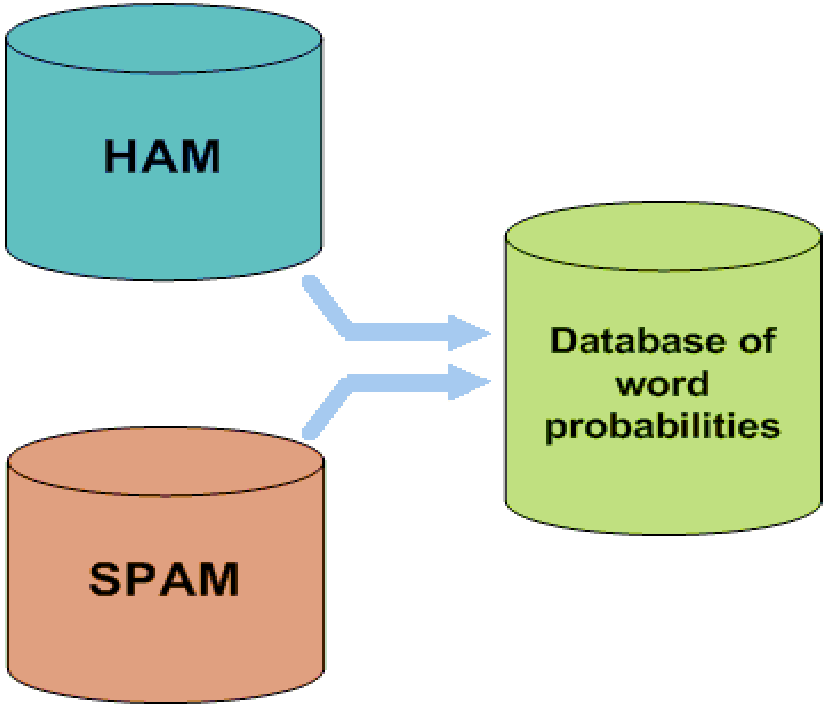
Thomas Bayes
1701 - 1761

UNDERSTANDING NAÏVE BAYES…

- probability is a number between 0 and 1 (that is, between 0 percent and 100 percent), which captures the chance that an event will occur in the light of the available evidence.
- The lower the probability, the less likely the event is to occur.
- A probability of 0 indicates that the event will definitely not occur, while a probability of 1 indicates that the event will occur with 100 percent certainty.
- Classifiers based on Bayesian methods utilize training data to calculate an observed probability of each outcome based on the evidence provided by feature values.
- When the classifier is later applied to unlabeled data, it uses the observed probabilities to predict the most likely class for the new features.

UNDERSTANDING NAÏVE BAYES…

- Bayesian classifiers have been used for:
  - Text classification, such as junk e-mail (spam) filtering
  - Intrusion or anomaly detection in computer networks
  - Diagnosing medical conditions given a set of observed symptoms



WHAT IS PROBABILITY ?

- if it rained 3 out of 10 days with similar conditions as today, The probability of rain today can be

  3 / 10 = 0.30 or 30%

- if 10 out of 50 prior email messages were spam, then The probability of any incoming message being spam can be

  10 / 50 = 0.20 or 20%.

- Notation:

  - P(Rain) = 0.3, P(NoRain) = 1 - 0.3 = 0.7
  - P(Spam) = 0.2 P(Ham) = 1 – 0.2 = 0.8

Spam and Ham are mutually exclusive and exhaustive events. They cannot occur at the same time. Complement A is denoted with A^1

WHAT IS CONDITIONAL PROBABILITY ?

We need to use a more careful formulation of the relationship between two events.

Use Bayesian methods.

CONDITIONAL PROBABILITY USING BAYES' THEOREM

- How to represent relationship between dependent events using **Bayes' theorem ?**
- How to revise an estimate of the probability of one event in light of the evidence provided by another event ?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (1)$$

- P(A|B) is read as the probability of event A, given that event B occurred.
- This is known as **conditional probability**, since the probability of A is dependent (that is, conditional) on what happened with event B.

CONDITIONAL PROBABILITY USING BAYES' THEOREM…

- *P(A ∩ B)* → a measure of how often *A* and *B* are observed to occur together
- *P(B)* → a measure of how often *B* is observed to occur in general.
- We can write eq(1) in another form:

$$P(A \cap B) = P(A|B) * P(B) \qquad \textbf{(2)}$$

And, $$P(A \cap B) = P(B \cap A)$$

Hence, $$P(A \cap B) = P(B|A) * P(A) \qquad (3)$$

From (1) and (3), we obtain,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

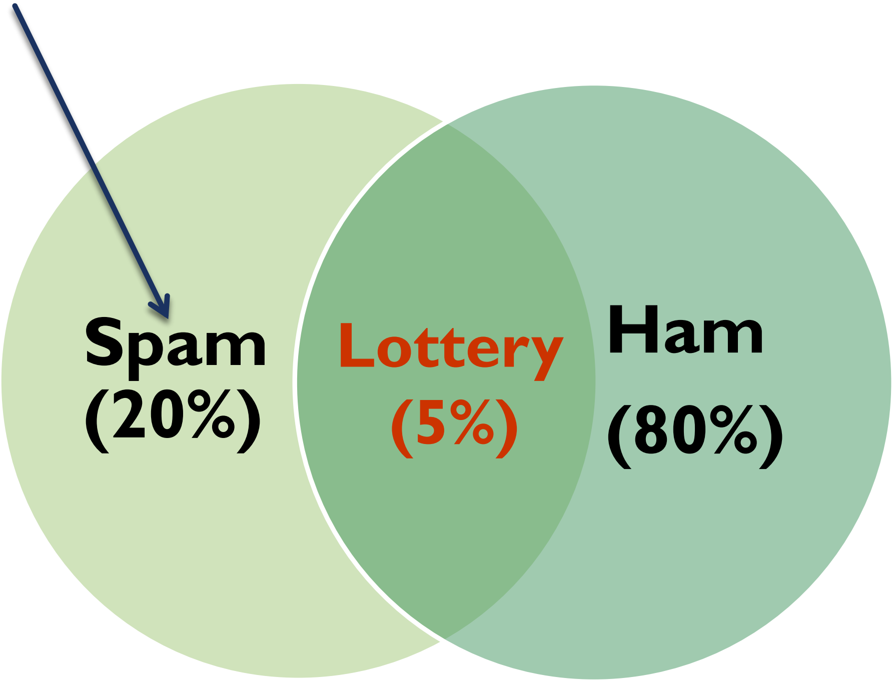
This is the traditional way in which Bayes' theorem has been specified

HOW TO APPLY BAYES THEOREM ?

$$\underbrace{P(Spam \mid Lottery)}_{\text{Posterior Probability}} = \frac{\overbrace{P(Lottery \mid Spam)}^{\text{Likelihood}} * \overbrace{P(Spam)}^{\text{Prior Probability}}}{\underbrace{P(Lottery)}_{\text{Marginal Likelihood}}}$$

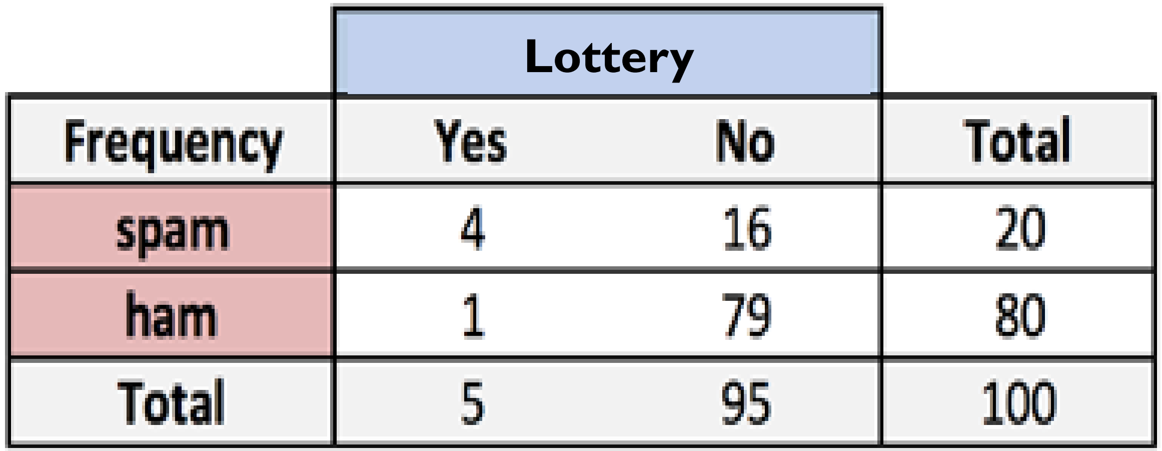
PRIOR PROBABILITY - EXAMPLE

- Without knowledge of an incoming message's content, the best estimate of its spam status would be *P(spam)*, the probability that any prior message was spam, which we calculated previously to be 20 percent. This estimate is known as the **prior probability**.



PRIOR PROBABILITY – EXAMPLE …

- Without knowledge of an incoming message's content, the best estimate of its spam status would be *P(spam)*, the probability that any prior message was spam, which we calculated previously to be 20 percent. This estimate is known as the **prior probability**.

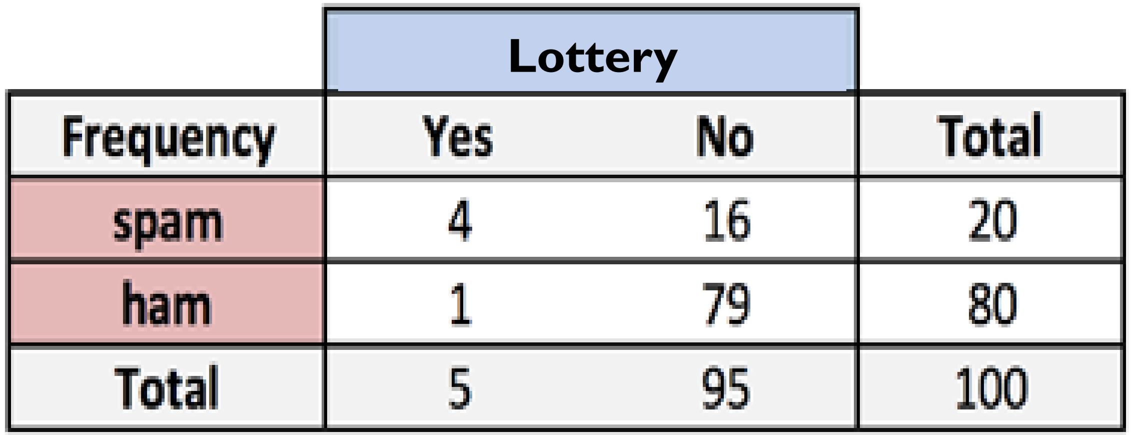


	Lottery		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

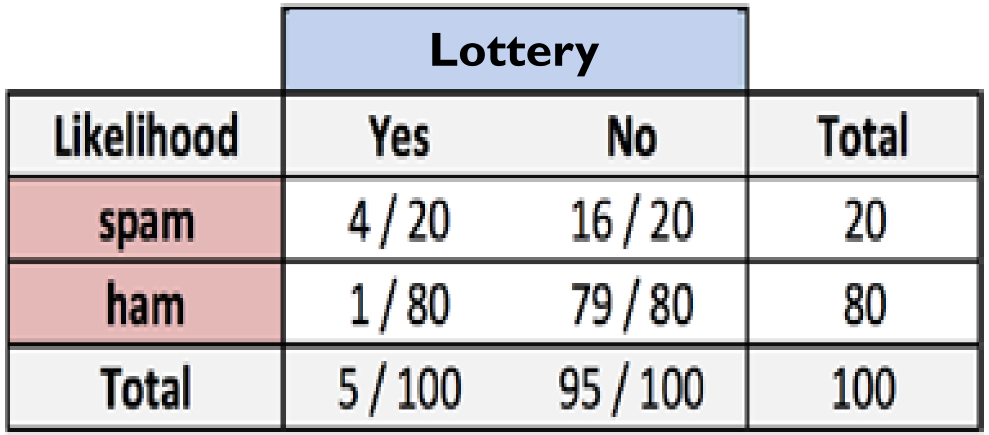
Prior Probability = 20/100 = 0.2 (OR) 20%

LIKELIHOOD

- Suppose that you obtained additional evidence by looking more carefully at the set of previously received messages to examine the frequency that the term Lottery appeared.
- The probability that the word Lottery was used in previous spam messages, or *P(Lottery|spam)*, is called the **likelihood**.



	Lottery		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

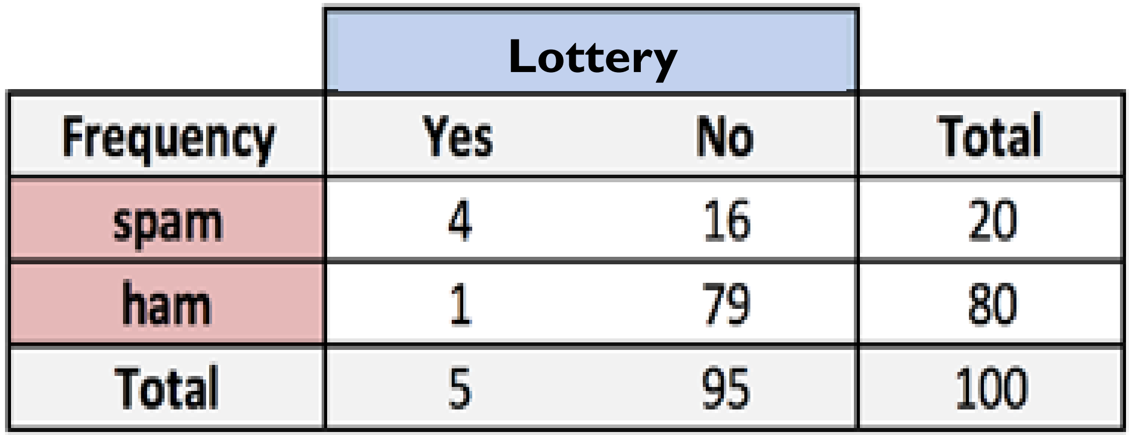


	Lottery		
Likelihood	Yes	No	Total
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

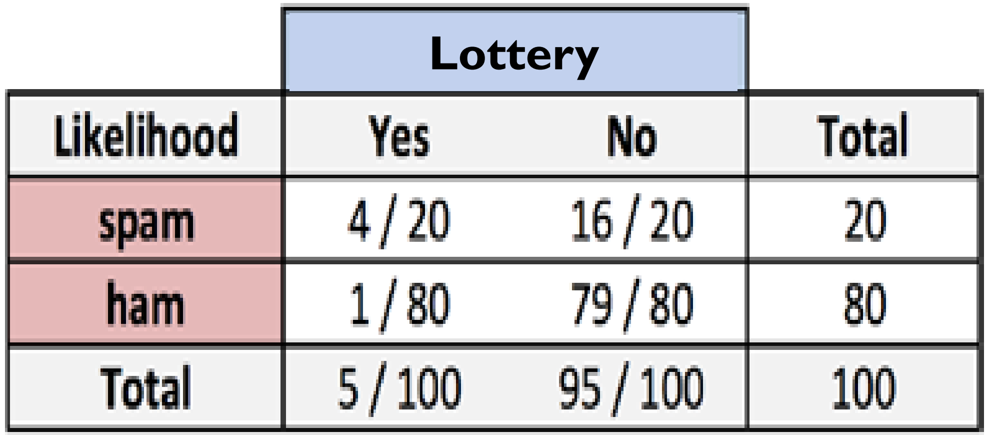
Likelihood = *P(Lottery =Yes | spam) = 4/20 = 0.20*

MARGINAL LIKELIHOOD

- The probability that Lottery appeared in any message at all, or *P(Lottery)*, is known as the **marginal likelihood**..



	Lottery		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100



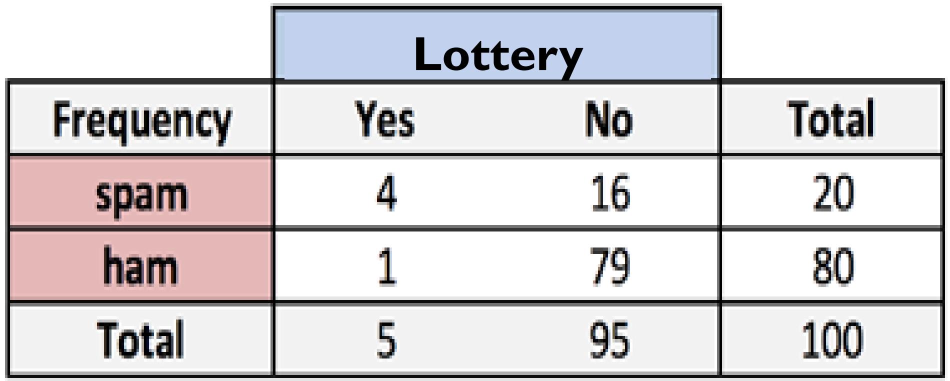
	Lottery		
Likelihood	Yes	No	Total
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

Marginal Likelihood = *P(Lottery =Yes) = (4 + 1)/100 = 5/100 = 0.005*

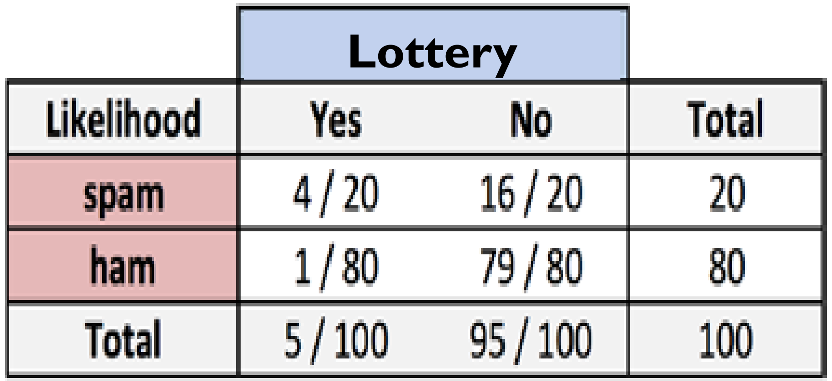
POSTERIOR PROBABILITY

- By applying Bayes' theorem to the evidence, we can compute **a posterior probability** that measures how likely the message is to be spam.

- If the posterior probability is greater than 50 percent, the message is more likely to be spam than ham and it should perhaps be filtered.



	Lottery		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100



	Lottery		
Likelihood	Yes	No	Total
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

$$P(Spam \mid Lottery\,) = \frac{P(Lottery \mid Spam)\ * P(Spam)}{P(Lottery)}$$

Posterior Probability = *(4/20) * (20/100) / (5/100) = 0.80*

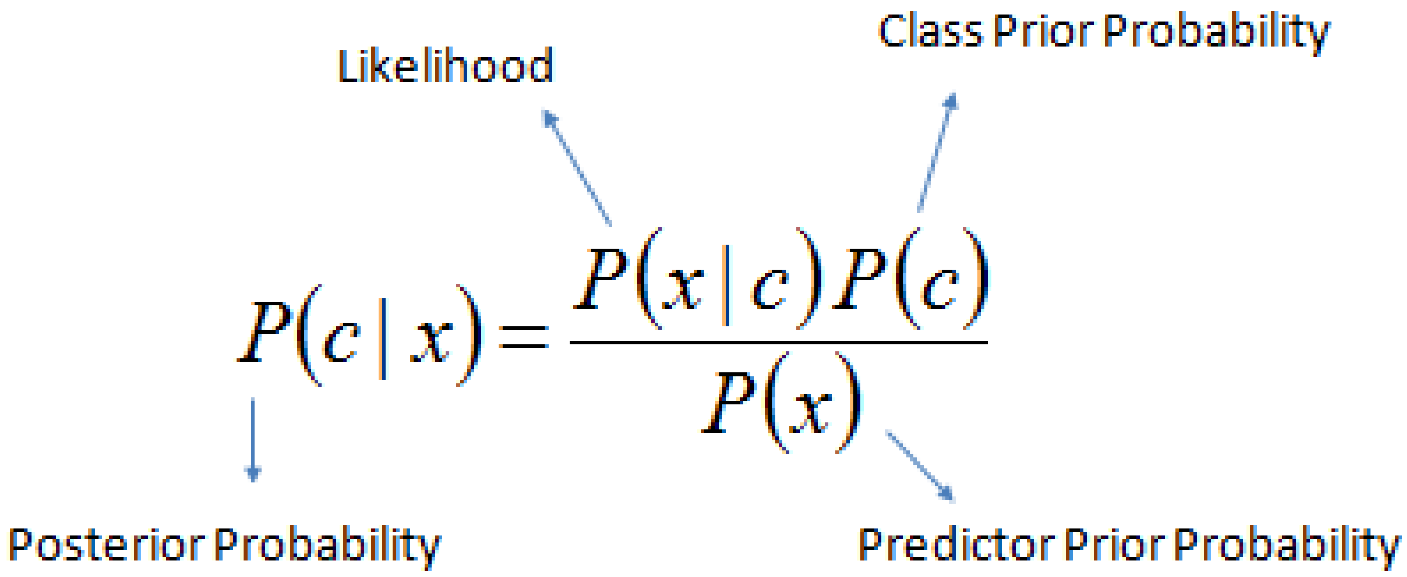
PREDICTION

The probability is 80 percent that a message is spam, given that it contains the word "Lottery".

In light of this result, any message containing this term should probably be filtered.

THE NAIVE BAYES ALGORITHM

- The **Naive Bayes** algorithm uses Bayes' theorem.
- It is named as such because it makes some "naive“ assumptions about the data. It assumes that all of the features in the dataset are equally important and independent.
- Naive Bayes classifier assume that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors.



$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Likelihood

Class Prior Probability

Posterior Probability

Predictor Prior Probability

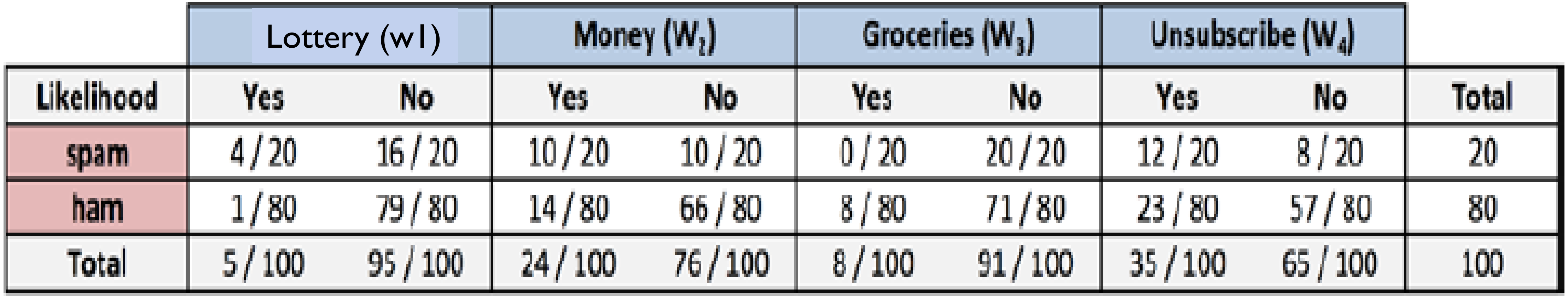
$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

THE NAIVE BAYES ALGORITHM…

- The Naive Bayes algorithm is named as such because it makes some "naive“ assumptions about the data.
- In particular, Naive Bayes assumes that all of the features in the dataset are equally important and independent.

CLASSIFICATION WITH NAIVE BAYES ALGORITHM…

- Let's extend our spam filter by adding a few additional terms to be monitored in addition to the term Lottery: Money, Groceries, and Unsubscribe.
- The Naive Bayes learner is trained by constructing a likelihood table for the appearance of these four words (labeled *W1*, *W2*, *W3*, and *W4*), as shown in the following diagram for 100 e-mails:



	Lottery (w1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		
Likelihood	Yes	No	Yes	No	Yes	No	Yes	No	Total
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

CLASSIFICATION WITH NAIVE BAYES ALGORITHM…

- As new messages are received, we need to calculate the posterior probability to determine whether they are more likely to be spam or ham, given the likelihood of the words found in the message text.
- For example, suppose that a message contains the terms Lottery and Unsubscribe, but does not contain either Money or Groceries.
- Using Bayes' theorem, we can define the problem as shown in the following formula.
- It captures the probability that a message is spam, given that *Lottery* = *Yes*, *Money* = *No*, *Groceries* = *No*, and *Unsubscribe* = *Yes*:

$$P(\text{spam}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4|\text{spam})P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

P(SPAM) = 0 **P(HAM) = 1**

IS THERE ANY PROBLEM WITH NB?

- Suppose that we received another message, this time containing all four terms:

  Lottery, Groceries, Money, and Unsubscribe.

Using the Naive Bayes algorithm as before, we can compute the likelihood of spam as:

$$P(\text{spam}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto P(W_1|\text{spam})P(\neg W_2|\text{spam})P(\neg W_3|\text{spam})P(W_4|\text{spam})P(\text{spam})$$

$$P(\text{ham}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto P(W_1|\text{ham})P(\neg W_2|\text{ham})P(\neg W_3|\text{ham})P(W_4|\text{ham})P(\text{ham})$$

- Likelihood(SPAM) = (4/20) * (10/20) * (0/20) * (12/20) * (20/100) = 0
- Likelihood(HAM) = (1/80) * (14/80) * (8/80) * (23/80) * (80/100) = 0.00005

P(SPAM) = 0/0+0.857 = 0 **P(HAM) = 1**

Does this prediction make sense? Probably not.

IS THERE ANY PROBLEM WITH NB?

- Consequently, *P(spam|groceries) = 0%.*
- This problem might arise if an event never occurs for one or more levels of the class. For instance, the term Groceries had never previously appeared in a spam message. Consequently, *P(spam|groceries) = 0%*
  because probabilities in the Naive Bayes formula are multiplied in a chain, this 0 percent value causes the posterior probability of spam to be zero.
- What is the solution??

Laplace Estimator

LAPLACE ESTIMATOR

- The Laplace estimator essentially adds a small number to each of the counts in the frequency table, which ensures that each feature has a nonzero probability of occurring with each class.
- Typically, the Laplace estimator is set to 1, which ensures that each class-feature combination is found in the data at least once.

- Likelihood(SPAM) = (5/24) * (11/24) * (1/24) * (13/24) * (20/100) = 0.0004
- Likelihood(HAM) = (2/84) * (15/84) * (9/84) * (24/84) * (80/100) = 0.0001

P(SPAM) = 80% = 0.8

P(HAM) = 20% = 0.2

CASE STUDY: FILTERING SPAM MAILS

Thank You