03-08-2023

CS051A

INTRO TO COMPUTER SCIENCE WITH TOPICS IN AI

14: Machine learning and Naïve Bayes

Alexandra Papoutsaki

she/her/hers

Lectures

Zilong Ye

he/him/his

Labs

Lecture 14: Machine learning and Naïve Bayes

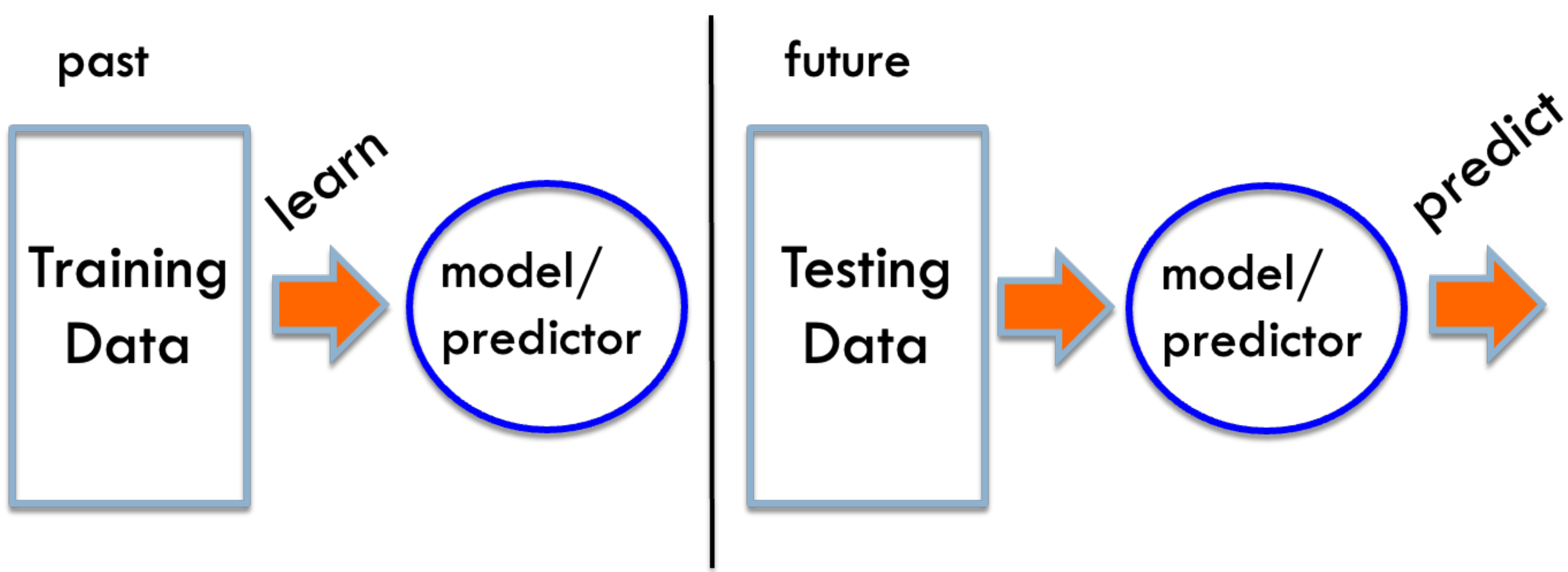
- Machine Learning
- Naïve Bayes

Machine Learning

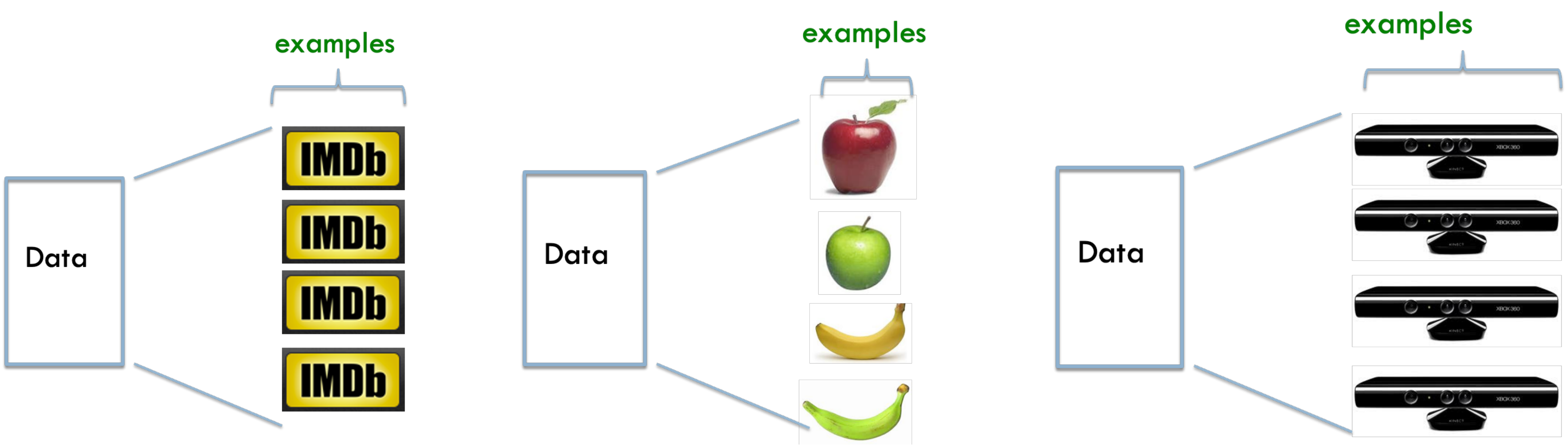
- Machine learning is about predicting the future based on the past.
  – Hal Daume III

Machine Learning

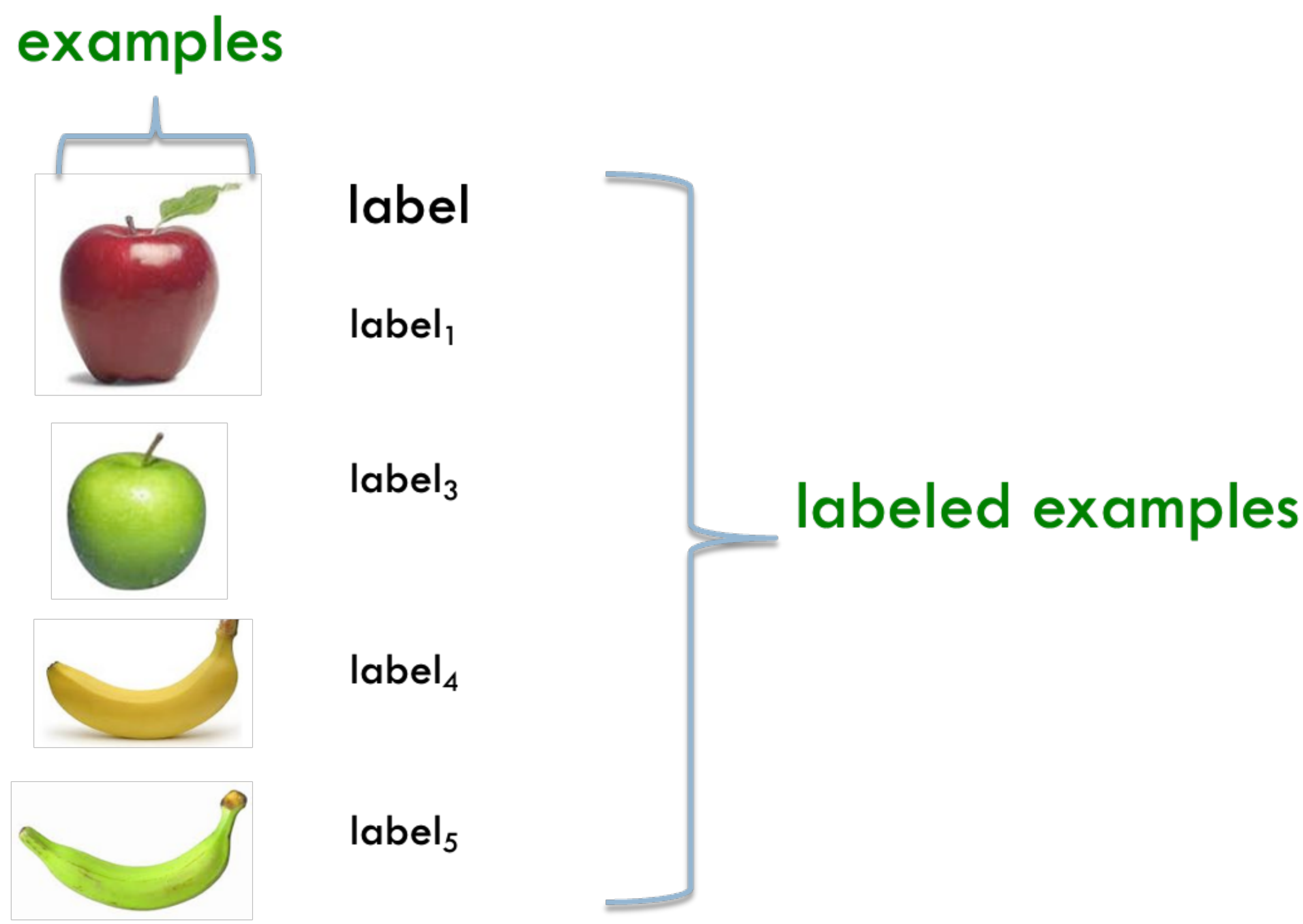
- Machine learning is about predicting the future based on the past.
  – Hal Daume III



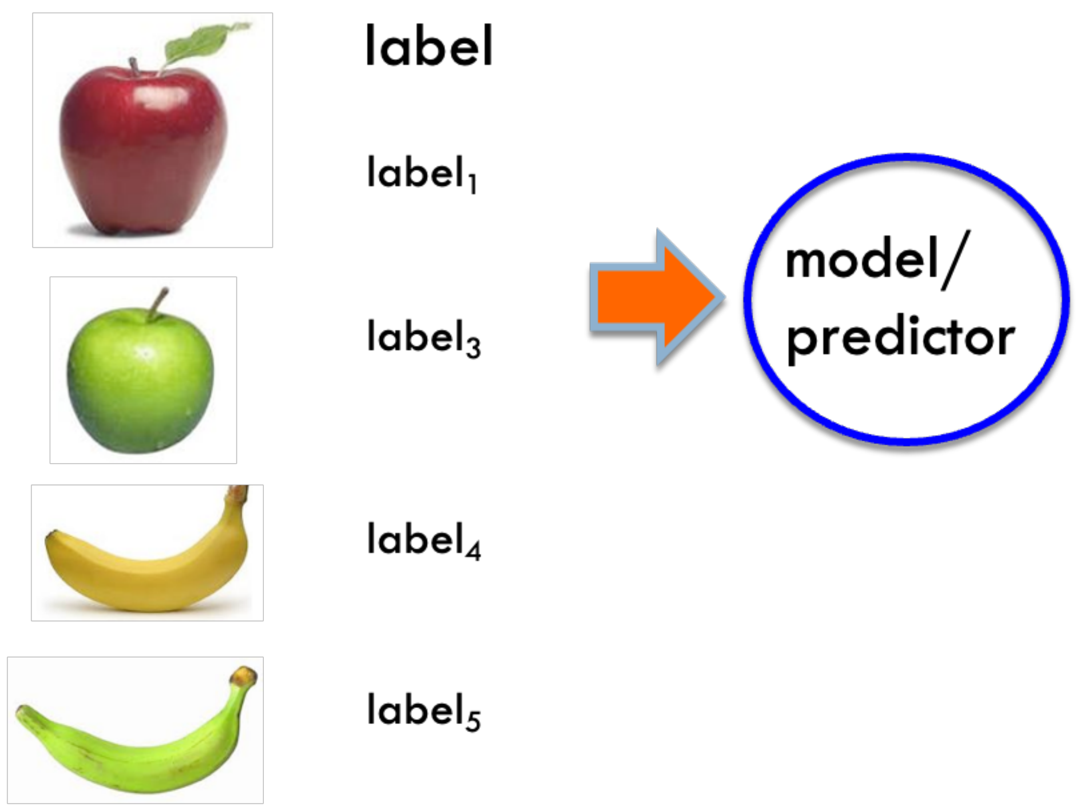
Different types of data



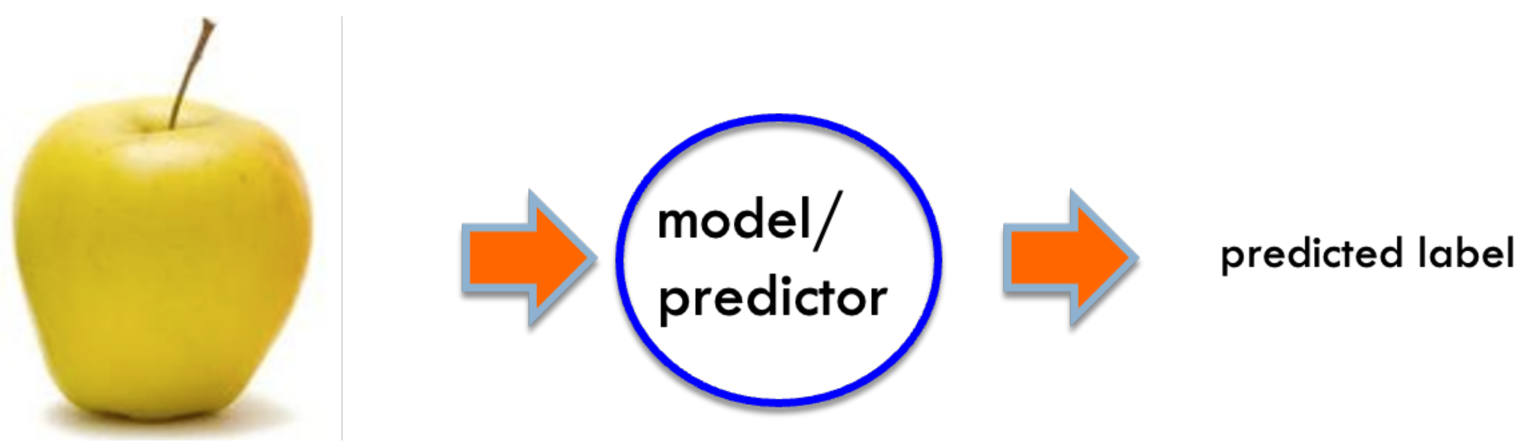
Supervised learning: given labeled examples



Supervised learning: given labeled examples build a model/predictor



Supervised learning: learn to predict new example



Supervised learning: classification

label

apple

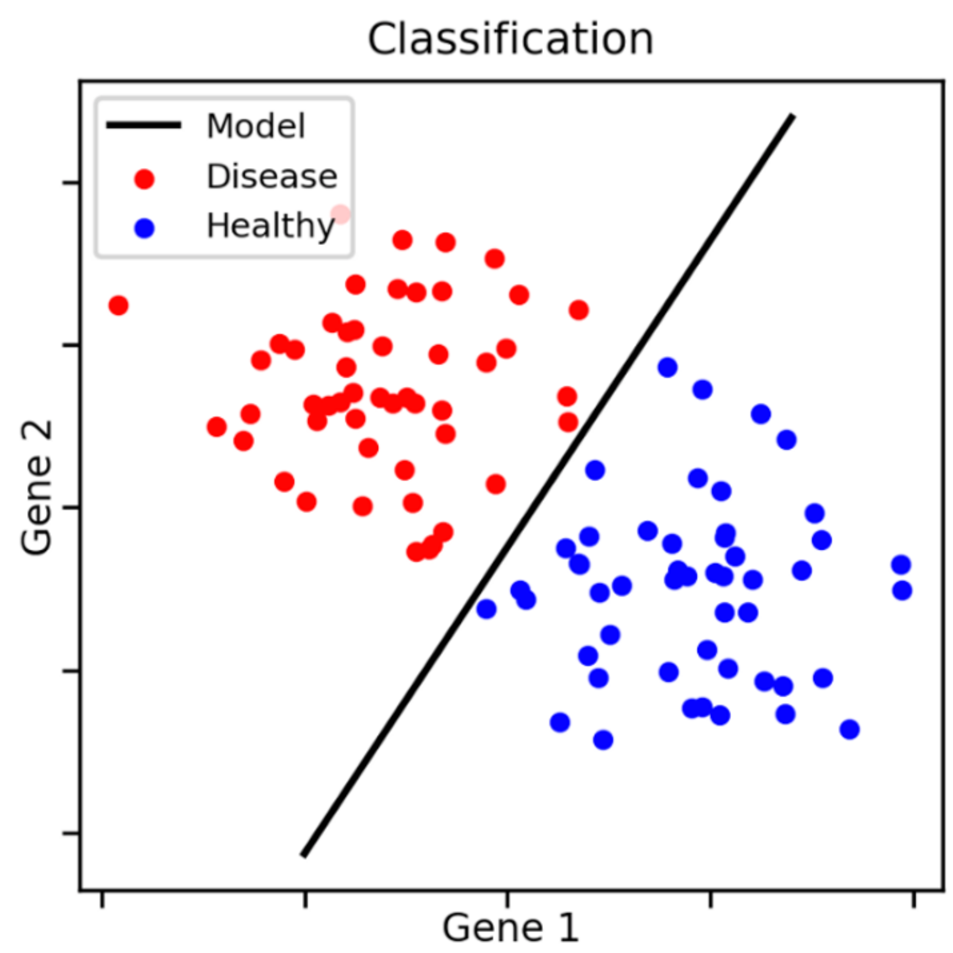
apple

banana

banana

Classification: a finite set of labels

Classification Example



Classification Applications

- Face recognition
- Character recognition
- Spam detection
- Medical diagnosis: From symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc

Supervised learning: regression

label

-4.5

10.1

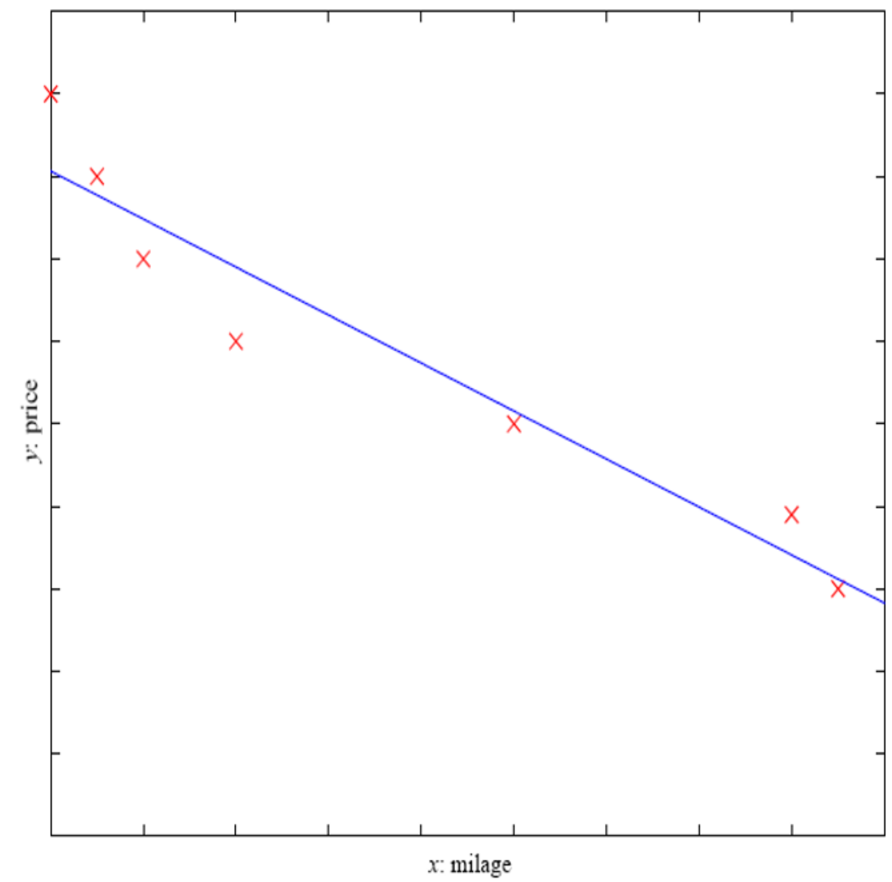
3.2

4.3

Regression: label is real-valued

Regression example

- Price of a used car
- x : car attributes
  (e.g., mileage)
- y : price



Regression Applications

- Economics/Finance: predict the value of a stock
- Epidemiology
- Car/plane navigation: angle of the steering wheel, acceleration, …
- Temporal trends: weather over time

Supervised learning: ranking

label

1

4

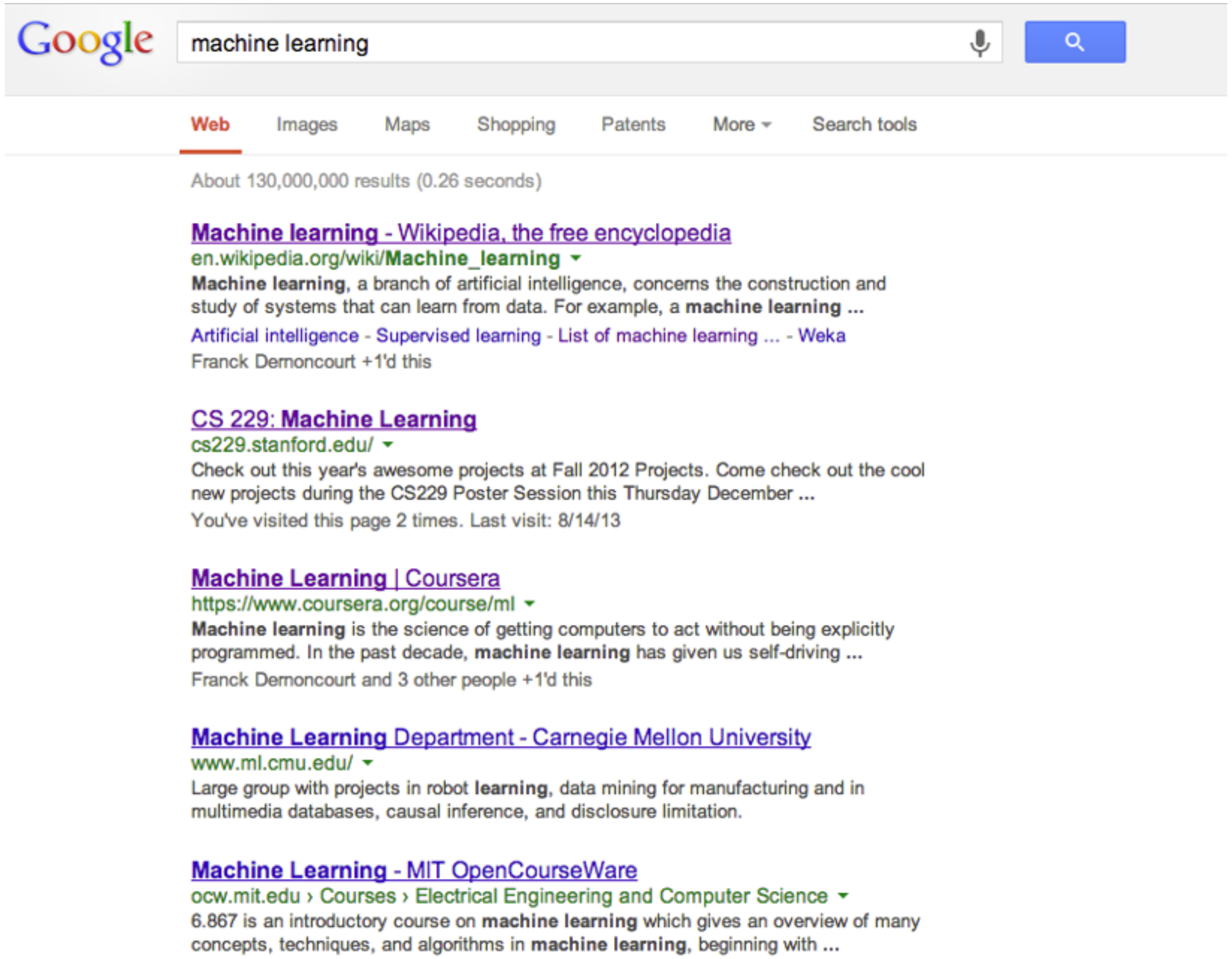
2

3

Ranking: label is a ranking

Ranking example

- Given a query and a set of web pages, rank them according to relevance



Ranking Applications

- User preference, e.g., Netflix “My List” -- movie queue ranking
- Spotify
- flight search (search in general)

Unsupervised learning: given data/examples but no labels





Unsupervised learning applications

- learn clusters/groups without any label
- customer segmentation (i.e. grouping)
- image compression
- bioinformatics: learn motifs

Reinforcement learning

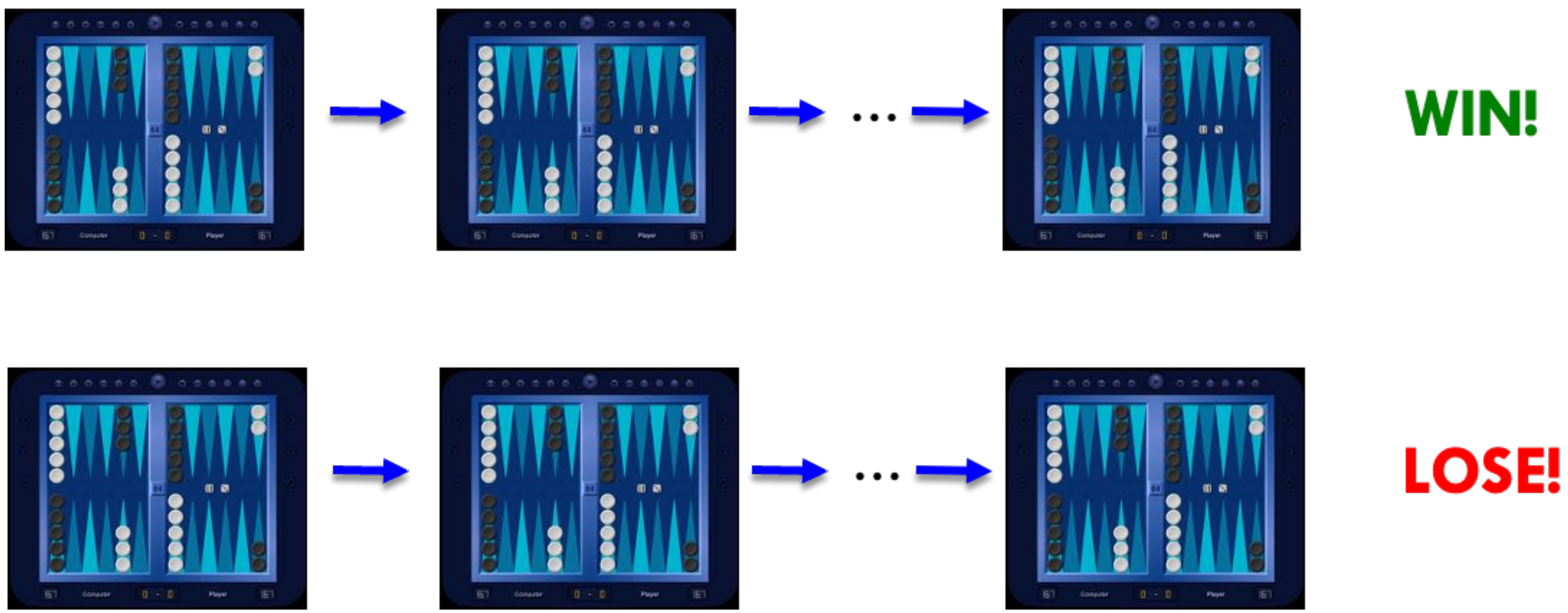
- Given a sequence of examples/states and a reward after completing that sequence, learn to predict the action to take in for an individual example/state

left, right, straight, left, left, left, straight	GOOD
left, straight, straight, left, right, straight, straight	BAD

left, right, straight, left, left, left, straight	18.5
left, straight, straight, left, right, straight, straight	-3

Reinforcement learning example

- Backgammon

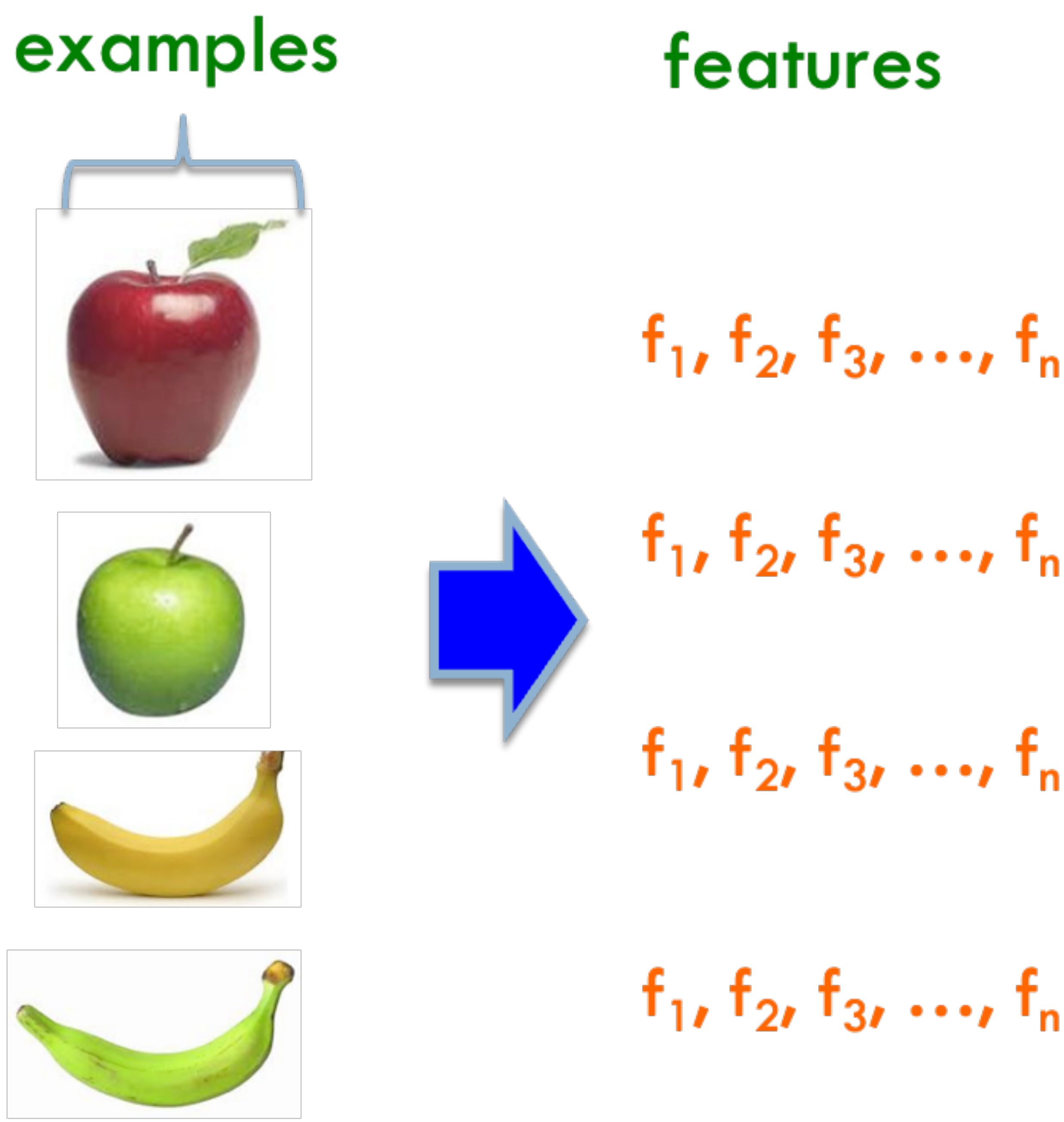


- Given sequences of moves and whether or not the player won at the end, learn to make good moves.

Representing examples

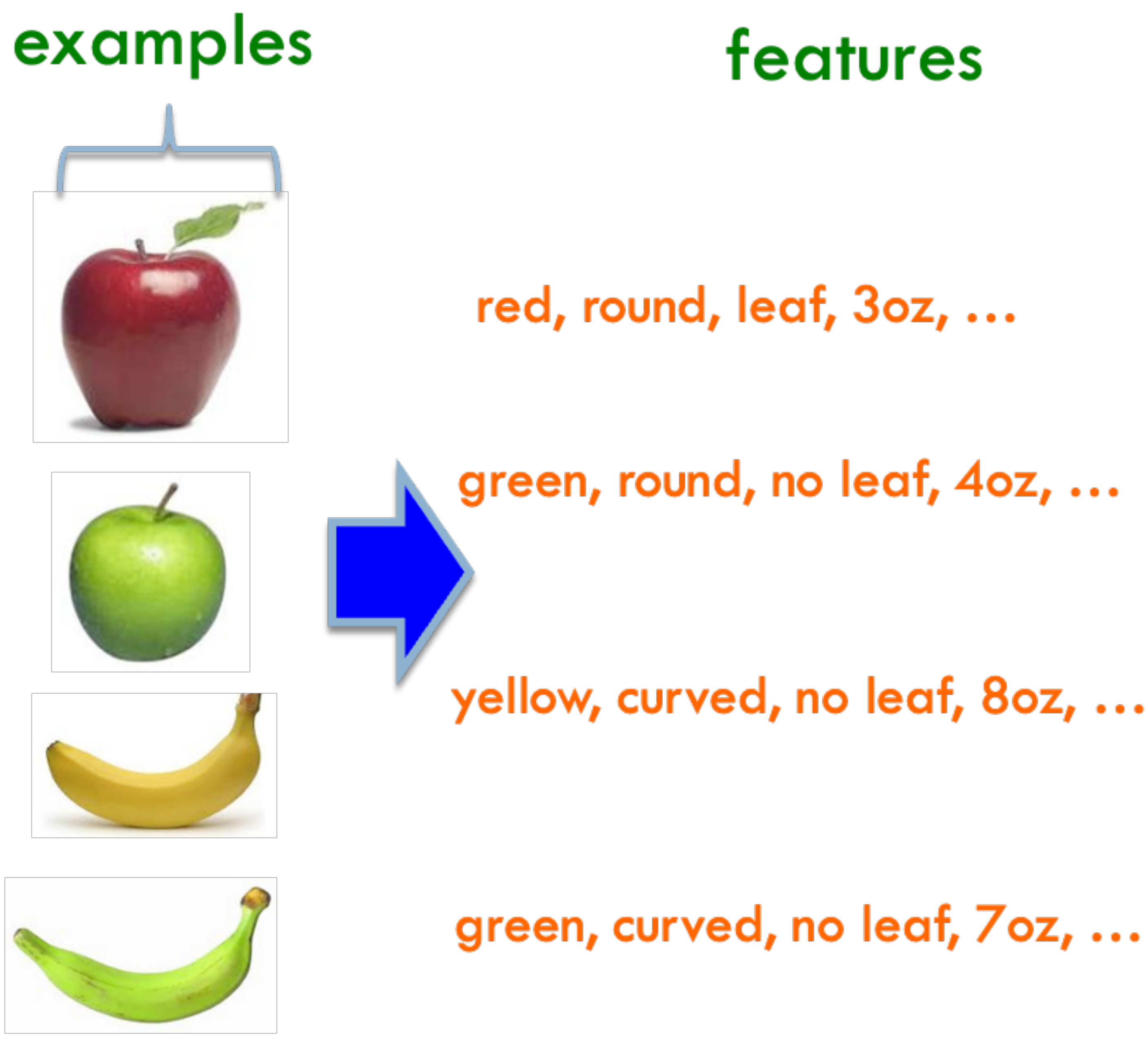
- What is an example?
- How is it represented?

Features



- How our algorithms actually "view" the data
- Features are the questions we can ask about the examples

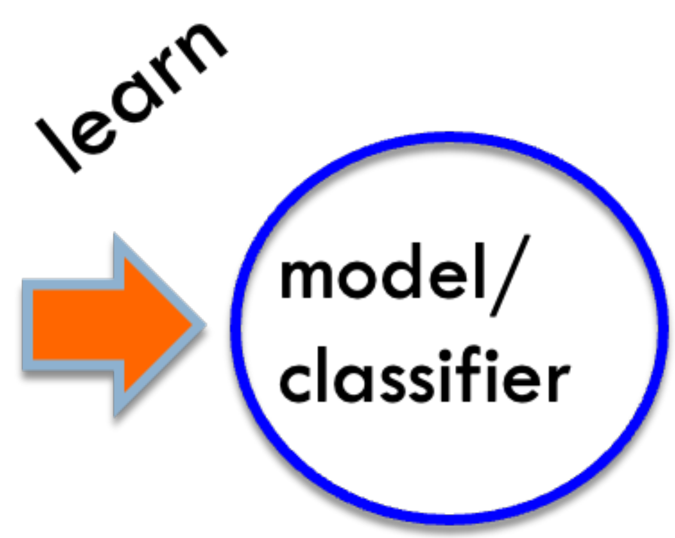
Features



- How our algorithms actually "view" the data
- Features are the questions we can ask about the examples

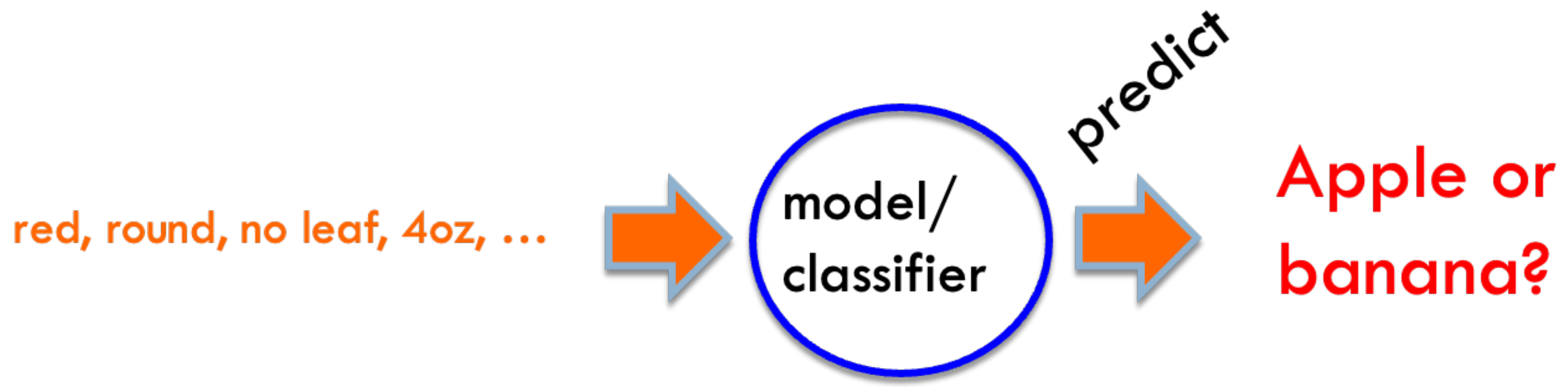
Classification revisited

examples	label
red, round, leaf, 3oz, …	apple
green, round, no leaf, 4oz, …	apple
yellow, curved, no leaf, 8oz, …	banana
green, curved, no leaf, 7oz, …	banana



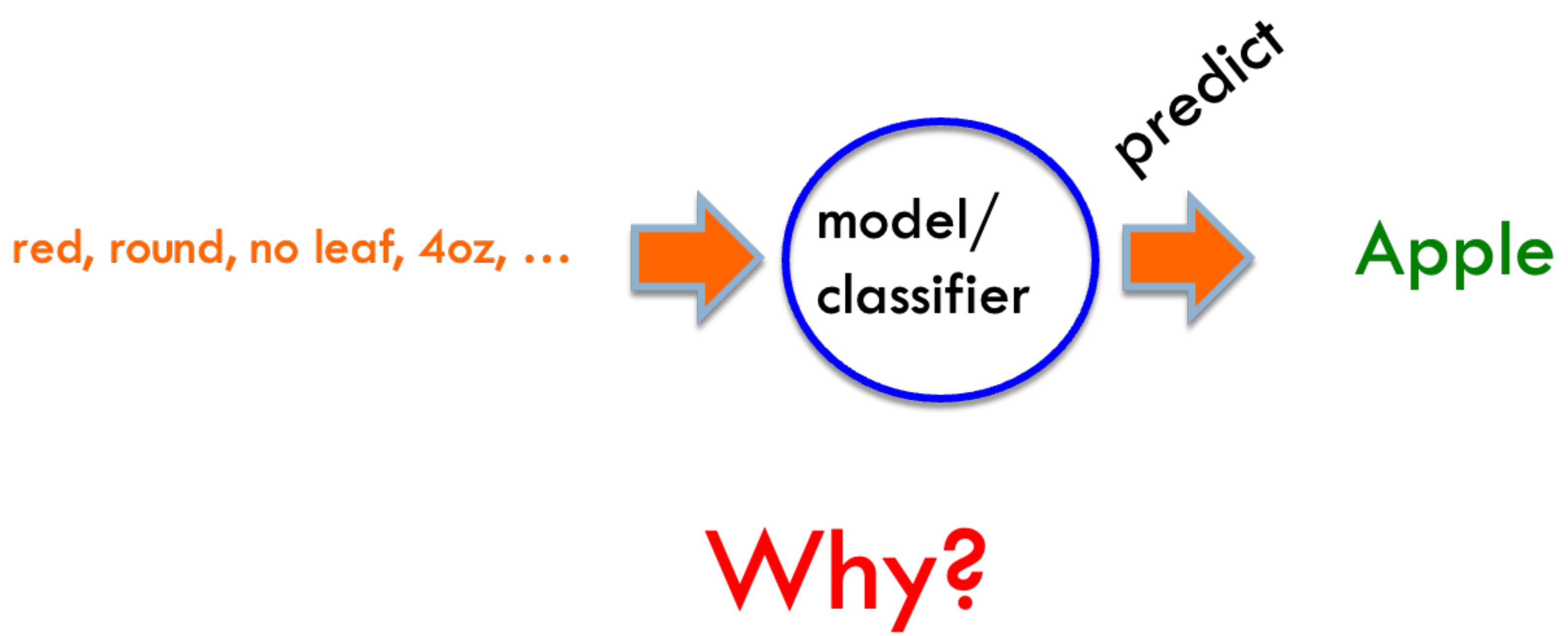
- During learning/training/induction, learn a model of what distinguishes apples and bananas based on the features.

Classification revisited



- The model can then classify a new example based on the features.

Classification revisited



- The model can then classify a new example based on the features.

Classification revisited

Training data		Test set	
examples	label		
red, round, leaf, 3oz, …	apple		
green, round, no leaf, 4oz, …	apple	red, round, no leaf, 4oz, …	?
yellow, curved, no leaf, 4oz, …	banana		
green, curved, no leaf, 5oz, …	banana		

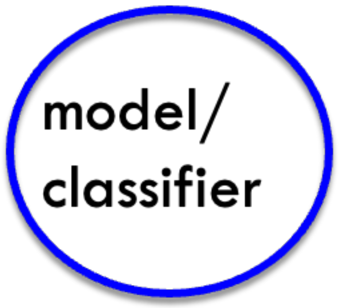
Classification revisited

Training data: examples	label	Test set
red, round, leaf, 3oz, …	apple	
green, round, no leaf, 4oz, …	apple	red, round, no leaf, 4oz, … ?
yellow, curved, no leaf, 4oz, …	banana	
green, curved, no leaf, 5oz, …	banana	

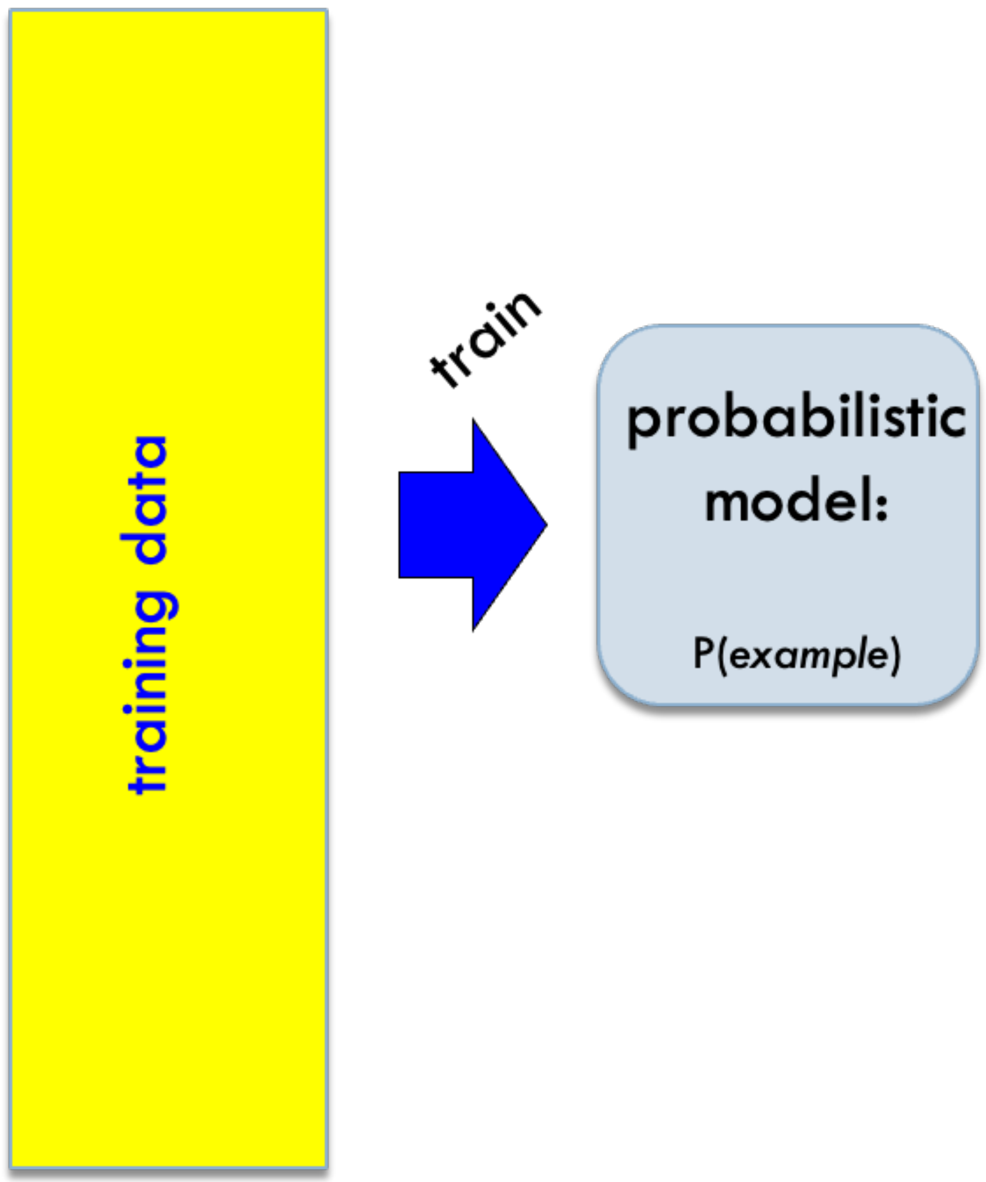
Learning is about ***generalizing*** from the training data

Models

- We have many, many different options for the model
- They have different characteristics and perform differently (accuracy, speed, etc.)

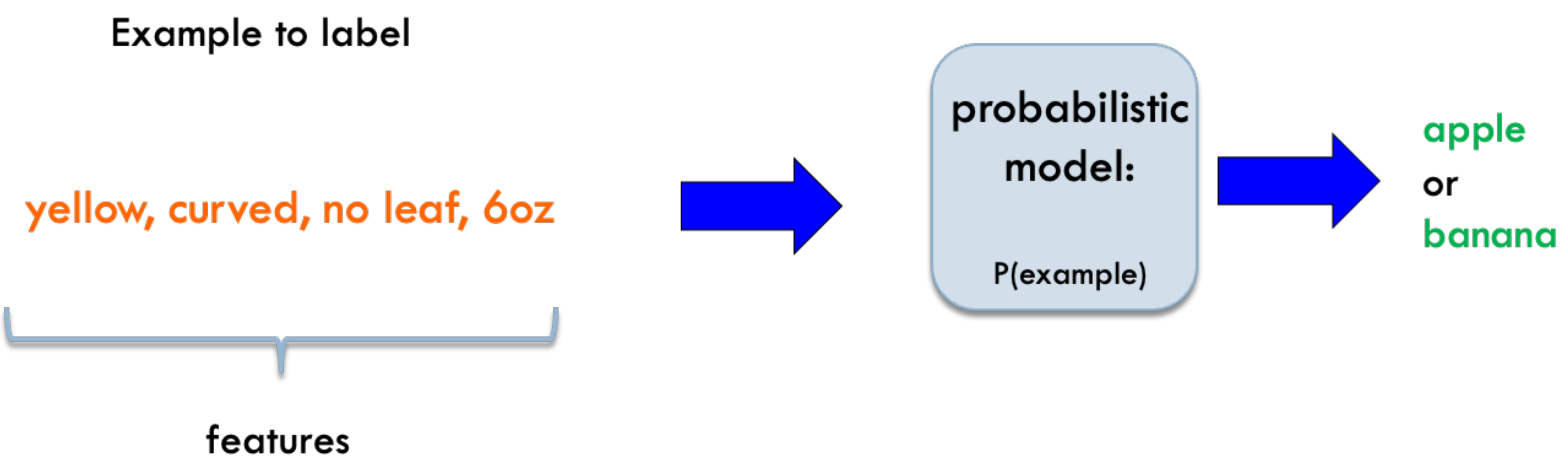


Probabilistic modeling



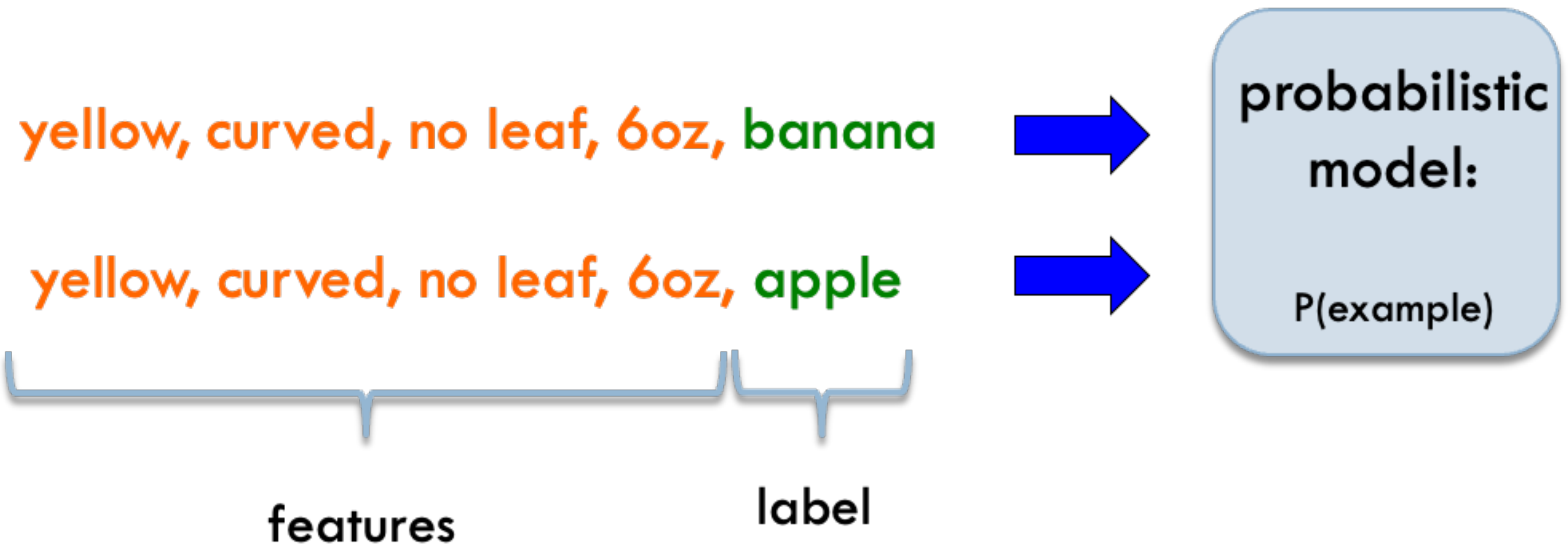
Model the data with a probabilistic model which tells us how likely a given data example is

Probabilistic models



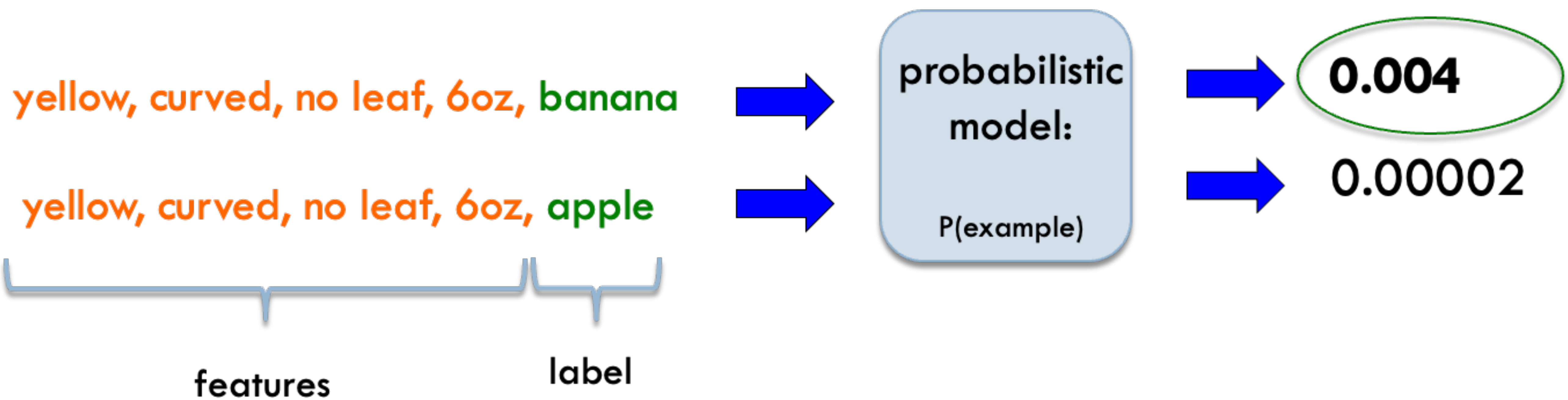
Probabilistic models

For each label, ask for the probability



Probabilistic models

Pick the label with the highest probability



Probability basics

- A probability distribution gives the probabilities of all possible values of an event
- For example, say we flip a coin three times. We can define the probability of the number of times the coin came up heads.

P(num heads)
P(3) = ?
P(2) = ?
P(1) = ?
P(0) = ?

Probability distributions

- What are the possible outcomes of three flips (hint, there are eight of them)?

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

Probability distributions

- Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = ?
P(2) = ?
P(1) = ?
P(0) = ?

Probability distributions

- Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = ?
P(2) = ?
P(1) = ?
P(0) = ?

Probability distributions

- Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = ?
P(1) = ?
P(0) = ?

Probability distributions

- Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = ?
P(1) = ?
P(0) = ?

Probability distributions

- Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = 3/8
P(1) = ?
P(0) = ?

Probability distributions

- Assuming the coin is fair, what are our probabilities?

$$\text{probability} = \frac{\text{number of times it happens}}{\text{total number of cases}}$$

T T T
T T H
T H T
T H H
H T T
H T H
H H T
H H H

P(num heads)
P(3) = 1/8
P(2) = 3/8
P(1) = 3/8
P(0) = 1/8

Probability distributions

- A probability distribution assigns probability values to all possible values.
- Probabilities are between 0 and 1, inclusive.
- The sum of all probabilities in a distribution must be 1.

P(num heads)
P(3) = 1/8
P(2) = 3/8
P(1) = 3/8
P(0) = 1/8

Probability distributions

- A probability distribution assigns probability values to all possible values.
- Probabilities are between 0 and 1, inclusive.
- The sum of all probabilities in a distribution must be 1.

P
P(3) = 1/2
P(2) = 1/2
P(1) = 1/2
P(0) = 1/2

P
P(3) = -1
P(2) = 2
P(1) = 0
P(0) = 0

Examples of probability distributions

- probability of heads
  - (distribution options: heads, tails)
- probability of passing class
  - (distribution options: pass, fail)
- probability of rain today
  - (distribution options: rain or no rain)
- probability of getting an 'A'
  - (distribution options: A, B, C, D, F)

Conditional probability distributions

- Sometimes we may know extra information about the world that may change our probability distribution.

- P(X|Y) captures this (read “probability of X given Y”)
  - Given some information (Y) what does our probability distribution look like?
  - Note that this is still just a typical probability distribution.

Conditional probability example

P(pass 51a)
P(pass) = 0.9
P(not pass) = 0.1

Unconditional probability distribution

Conditional probability example

P(pass 51a)
P(pass) = 0.9
P(not pass) = 0.1

Unconditional probability distribution

P(pass 51a | don't study)
P(pass) = 0.5
P(not pass) = 0.5

Still probability distributions over passing CS51A

P(pass 51a | do study)
P(pass) = 0.95
P(not pass) = 0.05

Conditional probability distribution

Conditional probability example

P(rain in LA)
P(rain) = 0.05
P(no rain) = 0.95

Unconditional probability distribution

Conditional probability example

P(rain in LA)
P(rain) = 0.05
P(no rain) = 0.95

Unconditional probability distribution

P(rain in LA | January)
P(rain) = 0.2
P(no rain) = 0.8

Still probability distributions over raining in LA

P(rain in LA | not January)
P(rain) = 0.03
P(no rain) = 0.97

Conditional probability distribution

Joint distribution

- Probability over two events: P(X,Y)
- Has probabilities for all possible combinations over the two events.

51Pass, EngPass	P(51Pass, EngPass)
true, true	0.88
true, false	0.01
false, true	0.04
false, false	0.07

Joint distribution

- Still a probability distribution
- All questions/probabilities that we might want to ask about these two events can be calculated from the joint distribution.

51Pass, EngPass	**P(51Pass, EngPass)**
true, true	0.88
true, false	0.01
false, true	0.04
false, false	0.07

What is P(51Pass = true)?

Joint distribution

- There are two ways that a person can pass 51: they can do it while passing or not passing English
- P(51Pass=true) = P(true, true) + P(true, false) = 0.89

51Pass, EngPass	P(51Pass, EngPass)
true, true	0.88
true, false	0.01
false, true	0.04
false, false	0.07

Relationship between distributions

$$P(X,Y) = P(Y) * P(X|Y)$$

joint distribution — unconditional distribution — conditional distribution

- Can think of it as describing the two events happening in two steps:
- The likelihood of X and Y happening:
  - How likely it is that Y happened?
  - Given that Y happened, how likely is it that X happened?

Relationship between distributions

$$P(51Pass, EngPass) = P(EngPass) * P(51Pass|EngPass)$$

- The probability of passing CS51 and English is:
  - Probability of passing English *
  - Probability of passing CS51 **given** that you passed English.

Can also view it with the other event happening first:

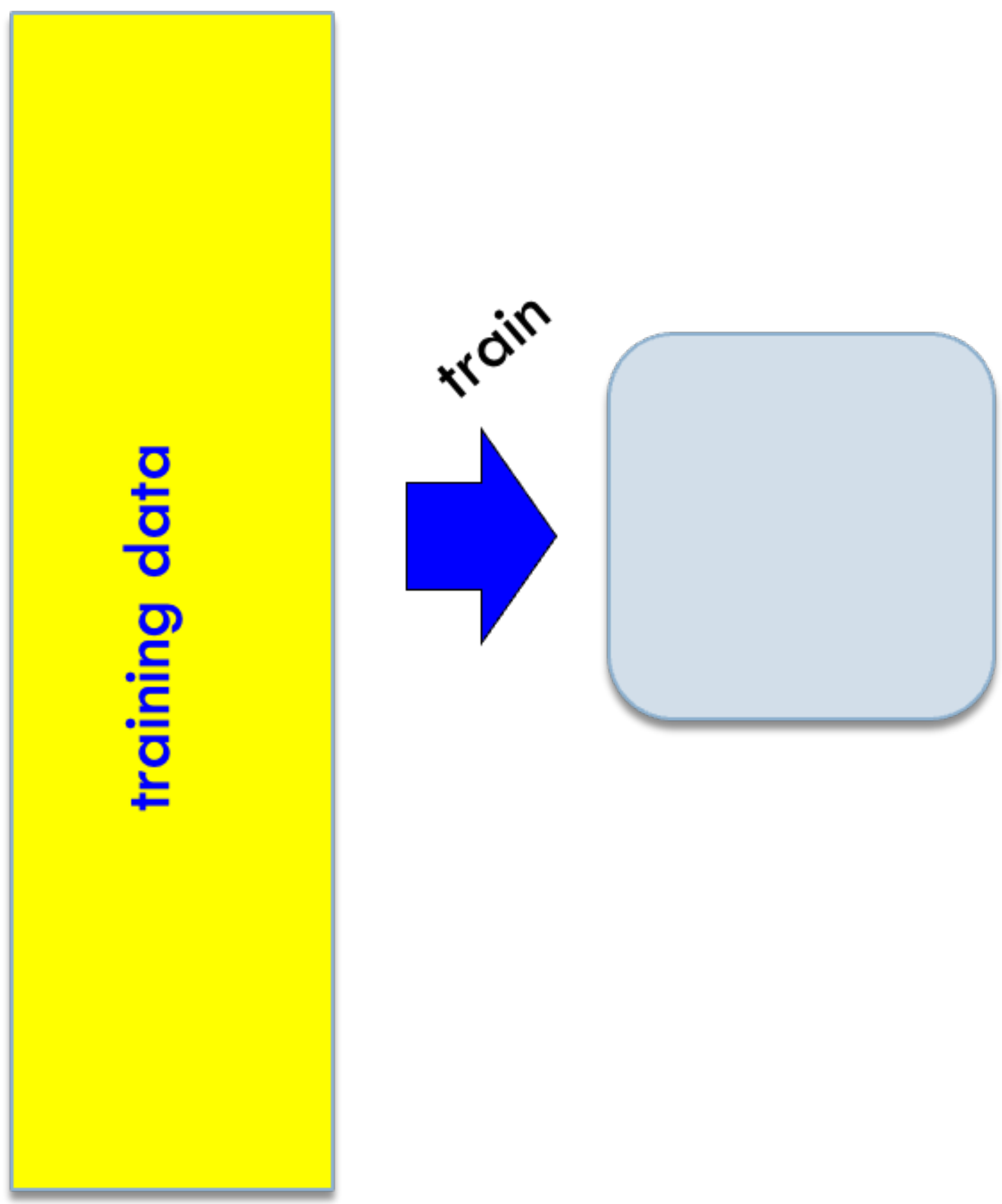
$$P(51Pass, EngPass) = P(51Pass) * P(EngPass|51Pass)$$

- The probability of passing CS51 and English is:
  - Probability of passing CS51 *
  - Probability of passing English **given** that you passed CS51.

Lecture 14: Machine learning and Naïve Bayes

- Machine Learning
- Naïve Bayes

Back to probabilistic modeling



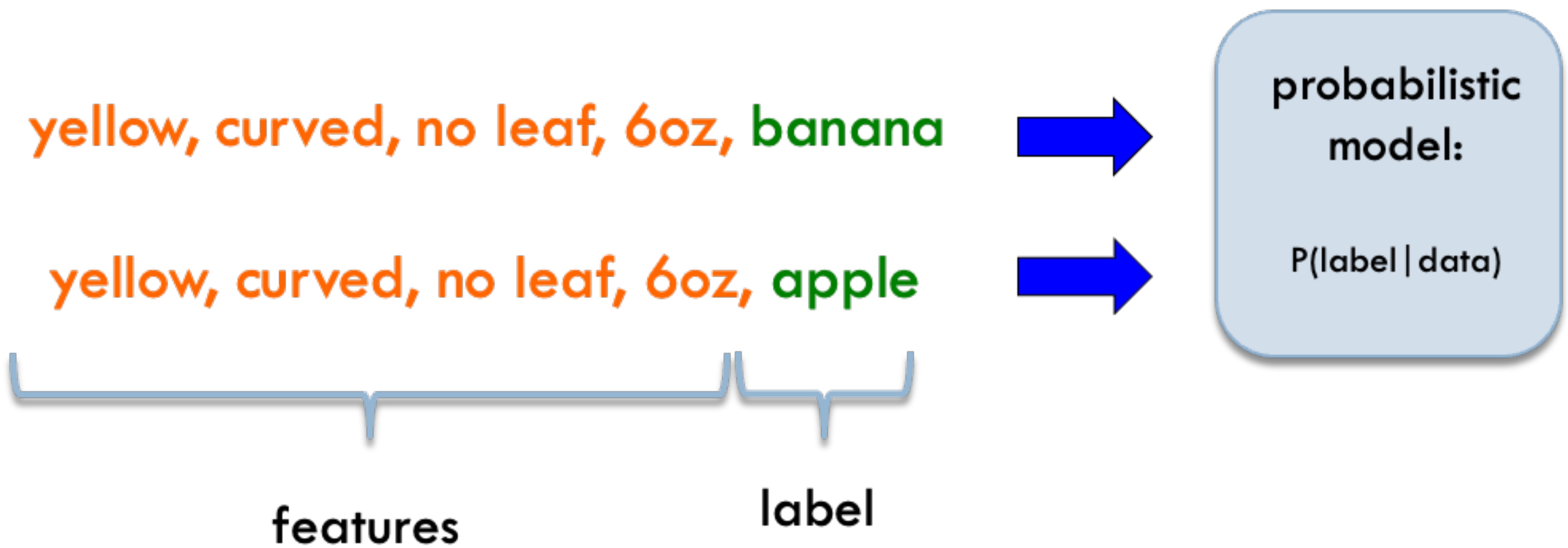
Build a model of the conditional distribution:

P(label | data)

How likely is a label given the data

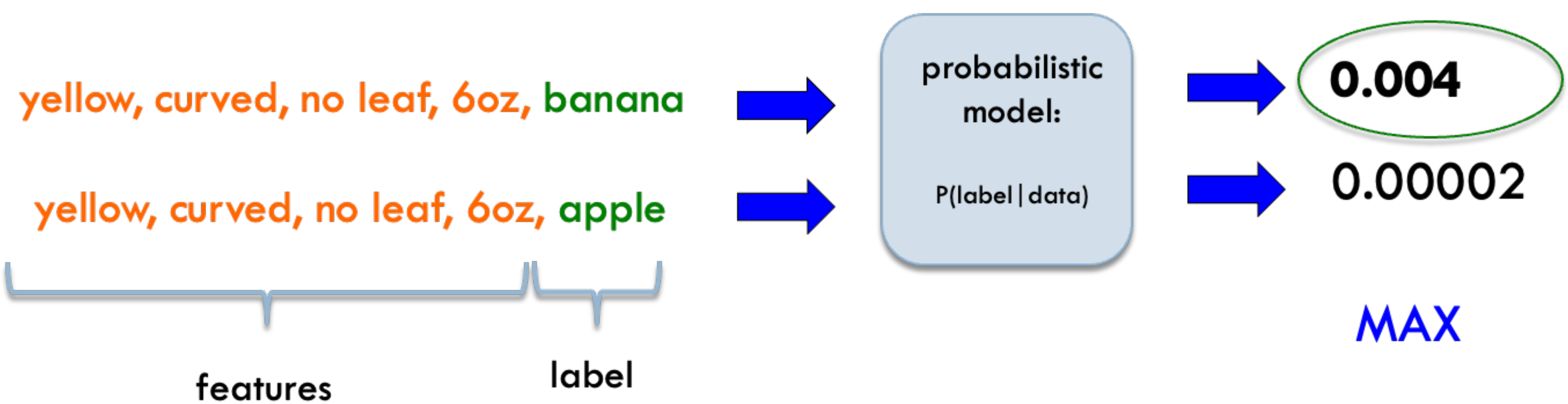
Back to probabilistic modeling

- For each label, calculate the probability of the label given the data.



Back to probabilistic modeling

- Pick the label with the highest probability.



Naïve Bayes model

- Two parallel ways of breaking down the joint distribution.

$$P(data, label) = P(label) * P(data|label)$$

$$P(data, label) = P(data) * P(label|data)$$

$$P(label) * P(data|label) = P(data) * P(label|data)$$

- What is P(label|data)?

Naïve Bayes model

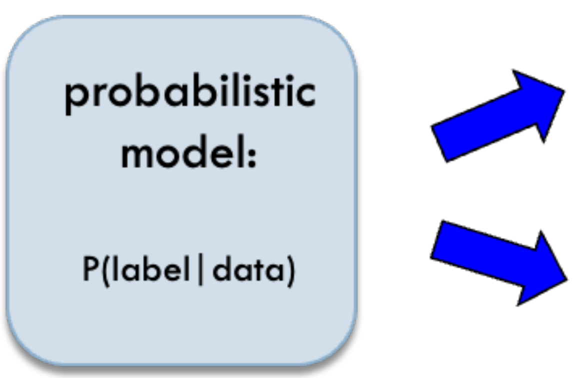
- Bayes' rule

$$P(label) * P(data|label) = P(data) * P(label|data)$$

$$P(label|data) = \frac{P(label) * P(data|label)}{P(data)}$$

Naïve Bayes

$$P(label|data) = \frac{P(label) * P(data|label)}{P(data)}$$



$$\frac{P(positive) * P(data|positive)}{P(data)}$$

MAX

$$\frac{P(negative) * P(data|negative)}{P(data)}$$

One observation

- For picking the largest, P(data) doesn't matter.

$$\frac{P(positive) * P(data|positive)}{P(data)}$$

MAX

$$\frac{P(negative) * P(data|negative)}{P(data)}$$

One observation

- For picking the largest, P(data) doesn't matter.

$$P(positive) * P(data|positive)$$

MAX

$$P(negative) * P(data|negative)$$

One observation

- If we assume that P(positive) == P(negative), then

$$P(positive) * P(data|positive)$$

MAX

$$P(negative) * P(data|negative)$$

- becomes

$$P(data|positive)$$

MAX

$$P(data|negative)$$

P(data|label)

$$P(data|label) = P(f_1, f_2, \ldots, f_n|label)$$
$$\approx P(f_1|label) * P(f_2|label) * \ldots P(f_n|label)$$

- This is generally not true!
- However…, it makes our life easier.
- This is why the model is called **Naïve** Bayes.

Naïve Bayes

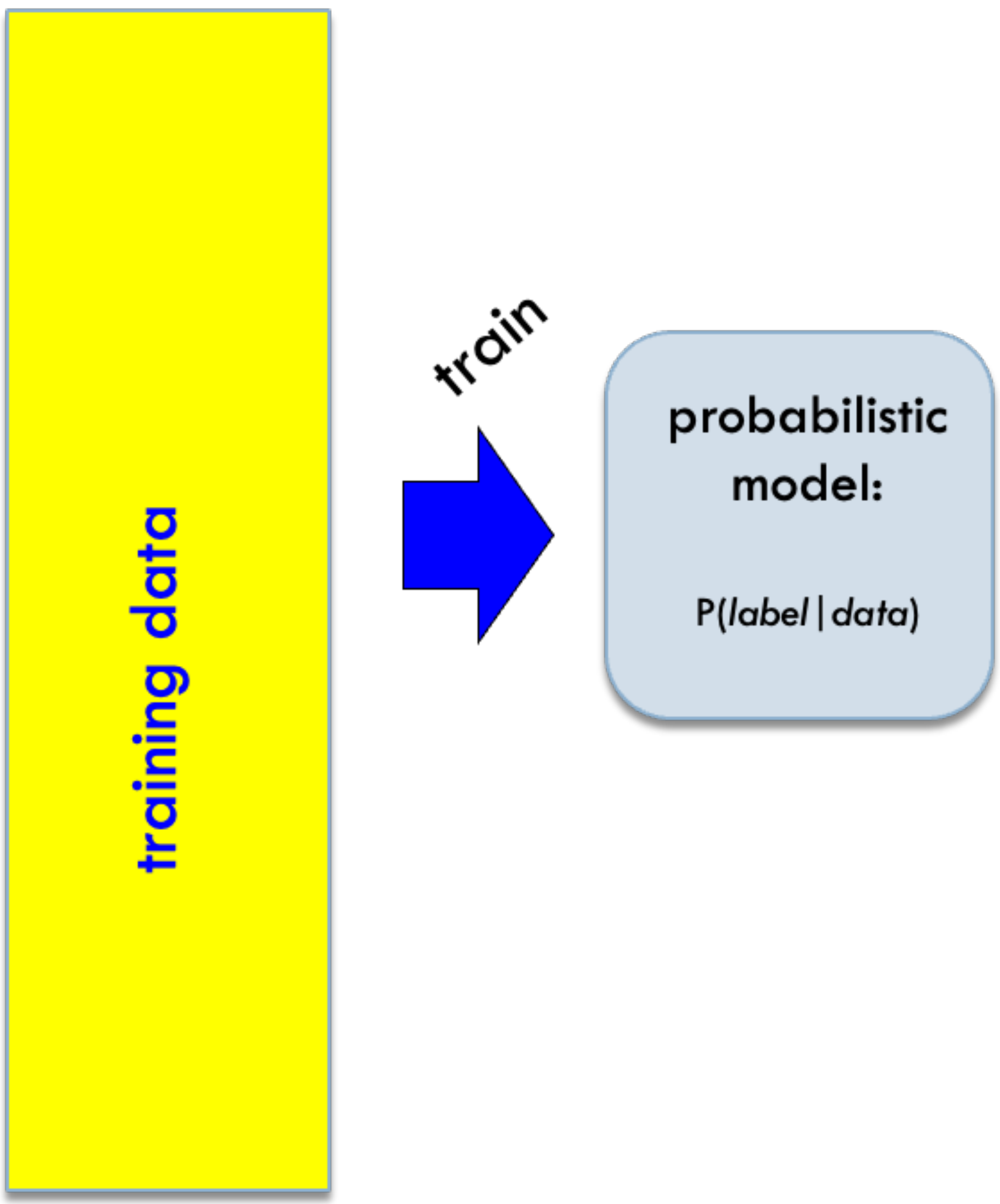
$$P(f_1|positive) * P(f_2|positive) *...* P(f_n|positive)$$

MAX

$$P(f_1|negative) * P(f_2|negative) *...* P(f_n|negative)$$

Where do these come from?

Training Naïve Bayes



An aside: P(heads)

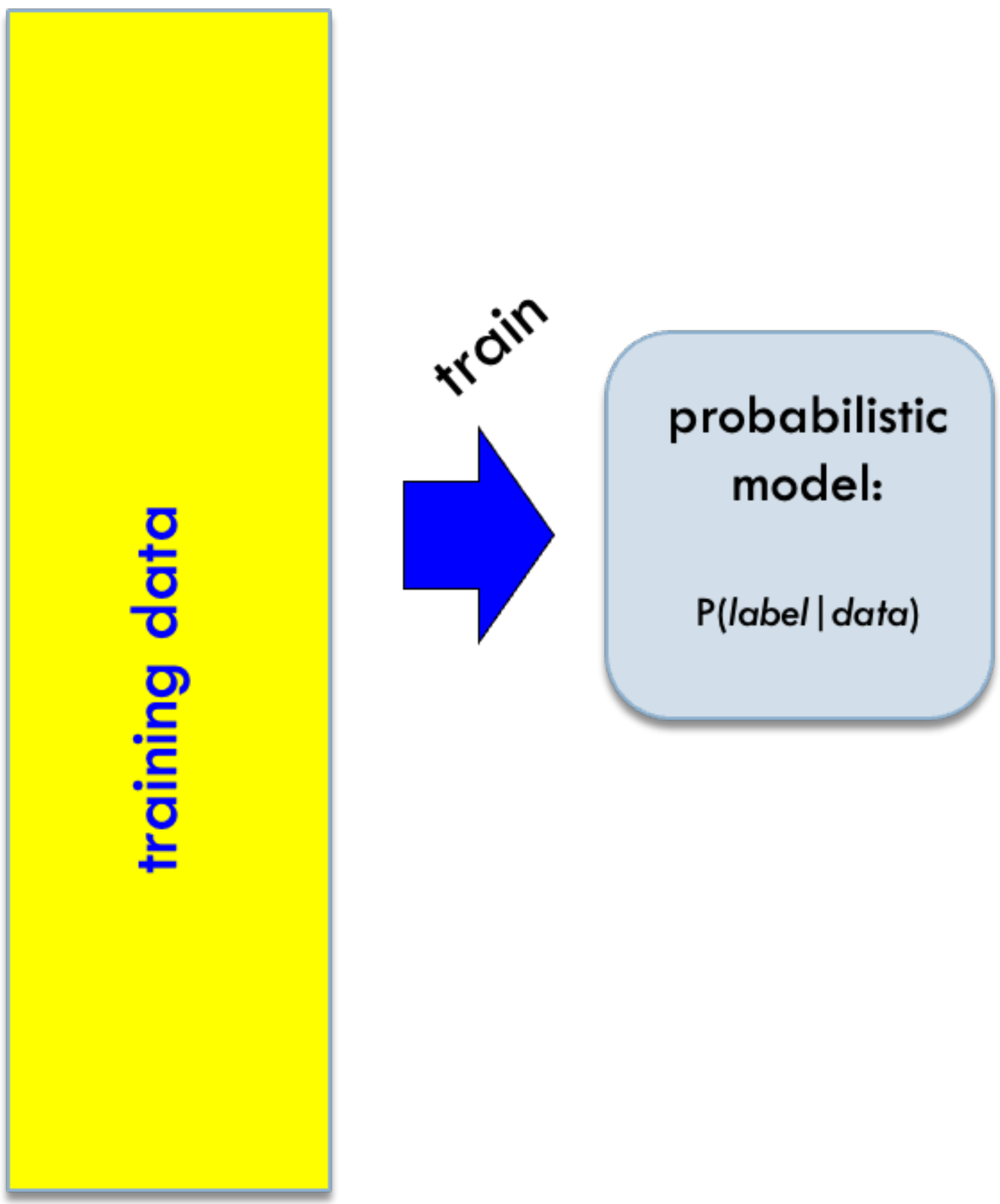
- What is the P(heads) on a fair coin?
  - 0.5
- What if you didn't know that, but had a coin to experiment with?
  - $P(heads) = \frac{number\ of\ times\ heads\ came\ up}{total\ number\ of\ coin\ tosses}$

P(feature|label)

$$P(heads) = \frac{number\ of\ times\ heads\ came\ up}{total\ number\ of\ coin\ tosses}$$

- Can we do the same thing here? What is the probability of a feature given positive, i.e. the probability of a feature occurring in in the positive label?
- $$P(feature|positive) = \frac{number\ of\ positive\ examples\ with\ that\ feature}{total\ number\ of\ positive\ examples}$$

Training Naïve Bayes



- Count how many examples have each label.
- For all examples with a particular label, count how many times each feature occurs.
- Calculate the conditional probabilities of each feature for all labels:

$$P(feature|label) = \frac{number\ of\ ``label"\ examples\ with\ that\ feature}{total\ number\ of\ examples\ with\ that\ label}$$

Classifying with Naïve Bayes

- For each label, calculate the product of P(feature|label) for each label.

Naïve Bayes Text Classification

- Given examples of text in different categories, learn to predict the category of new examples

- Sentiment classification: given positive/negative examples of text (sentences), learn to predict whether new text is positive/negative

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that I loved it	I loved that I hated it

Text Classification Training

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that I loved it	I loved that I hated it

- We'll assume words just occur once in any given sentence

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

- For each word and each label, learn: P(word | label)

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 3/3 = 1.0

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0

P(loved | positive) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 1.0
P(hated | positive) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0

P(loved | positive) = 1.0

P(hated | positive) = 1/3

…

P(I | negative) = 1.0

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 1.0
P(hated | positive) = 1/3
…

P(I | negative) = 1.0
P(movie | negative) = ?

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Training the model

Positive	Negative
I loved it	I hated it
I loved that movie	I hated that movie
I hated that loved it	I loved that hated it

P(I | positive) = 1.0
P(loved | positive) = 1.0
P(hated | positive) = 1/3
…

P(I | negative) = 1.0
P(movie | negative) = 1/3
…

$$P(word|label) = \frac{number\ of\ times\ word\ occured\ in\ \text{“}label\text{”}\ examples}{total\ number\ of\ examples\ with\ that\ label}$$

Classifying

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- Notice that each of these is its own probability distribution

P(it | positive) = 2/3
P(no it | positive) = 1/3

$$P(f_1|positive) * P(f_2|positive) * \ldots * P(f_n|positive)$$

MAX

Classifying

$$P(f_1|negative) * P(f_2|negative) * \ldots * P(f_n|negative)$$

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- How would we classify: "I hated movie"?

P(I | positive) * P(hated | positive) * P(movie | positive) = 1.0 * 1/3 * 1/3 = 1/9

P(I | negative) * P(hated | negative) * P(movie | negative) = 1.0 * 1.0 * 1/3 = 1/3

Classifying

$$P(f_1|positive) * P(f_2|positive) *...* P(f_n|positive)$$

MAX

$$P(f_1|negative) * P(f_2|negative) *...* P(f_n|negative)$$

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- How would we classify: "I hated the movie"?

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Classifying

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- What are these?

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Classifying

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- 0! Is this is a problem?

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Classifying

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- Yes, they make the entire product go to 0!

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Classifying

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- Our solution: assume any unseen word has a small, fixed probability, e.g., in this example 1/10.

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) =

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) =

Classifying

P(I | positive)	= 1.0	P(I | negative)	= 1.0
P(loved | positive)	= 1.0	P(hated | negative)	= 1.0
P(it | positive)	= 2/3	P(that | negative)	= 2/3
P(that | positive)	= 2/3	P(movie | negative)	= 1/3
P(movie | positive)	= 1/3	P(it | negative)	= 2/3
P(hated | positive)	= 1/3	P(loved | negative)	= 1/3

- Our solution: assume any unseen word has a small, fixed probability, e.g., in this example 1/10.

P(I | positive) * P(hated | positive) * P(the | positive) * P(movie | positive) = 1/90

P(I | negative) * P(hated | negative) * P(the | negative) * P(movie | negative) = 1/30

Full disclaimer

- I've fudged a few things on the Naïve Bayes model for simplicity.
- Our approach is very close, but it takes a few liberties that aren't technically correct, but it will work just fine.

Homework

- No homework for the week
- Sign up for group presentations