Text Classification Part 1: Machine Learning Models

CSCI 1460: Computational Linguistics

Lecture 2

Ellie Pavlick
Fall 2023

Topics

- Lecture 1 Reprise, New Quiz Makeup Policy
- Supervised Classification
- Feature Matrices and BOW Models
- Naive Bayes Text Classifier
- Logistic Regression Text Classifier
- Experimental Design in ML

Topics

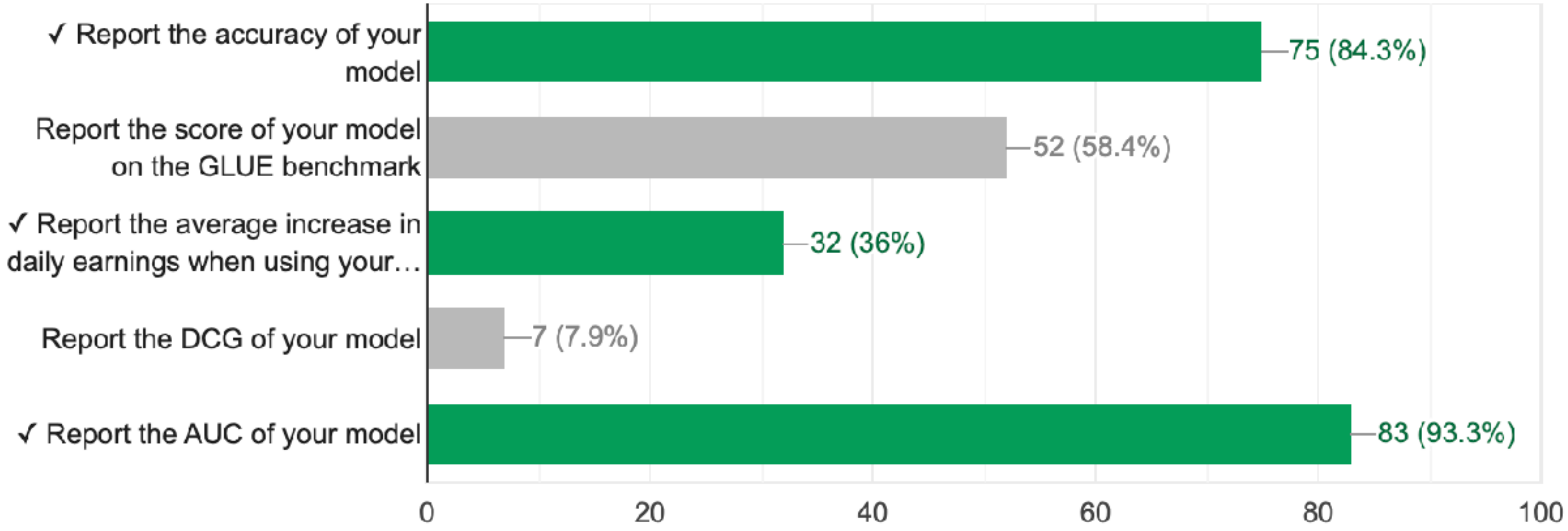
- **Lecture 1 Reprise, New Quiz Makeup Policy**
- Supervised Classification
- Feature Matrices and BOW Models
- Naive Bayes Text Classifier
- Logistic Regression Text Classifier
- Experimental Design in ML

Lecture 1 Reprise

You work for an hedge fund which is trying to use NLP to make investments based on NLP analysis of social media "buzz" about companies. You have been tasked with training the sentiment classifier. Which of the following is a reasonable way to evaluate your classifier? Check all that apply.

Copy

10 / 89 correct responses



Lecture 1 Reprise

Recall what you know about evaluating open book question answering systems. You have a very bad system. For any (question, document) input, the system returns the entire document as an answer to the question. You evaluate on the SQUAD dataset. Rank the following metrics from that which is likely to be highest to that which is likely to be lowest.

	Highest	Middle	Lowest	Points
Precision of Answer Tokens	○	◉	○	1
Exact Match Accuracy	○	○	◉	1
Recall of Answer Tokens	◉	○	○	1

Lecture 1 Reprise

- "I think the First Question is a bit ambiguous. If the model (objective) is a is a spam detection model the F1 score would be 0.72, where as if it is to classify non-spam/good emails the F1 score would be 0.61."
  - Good catch! I should have clarified, P/R/F1 are always w.r.t a specific class.
- Requests to slow down and pause more.
  - I will do this. :)
- Quiz questions are not always straightforward
  - This is intentional! They are not meant to be "tricks" but should require applying what was covered in lecture

Lecture 1 Reprise

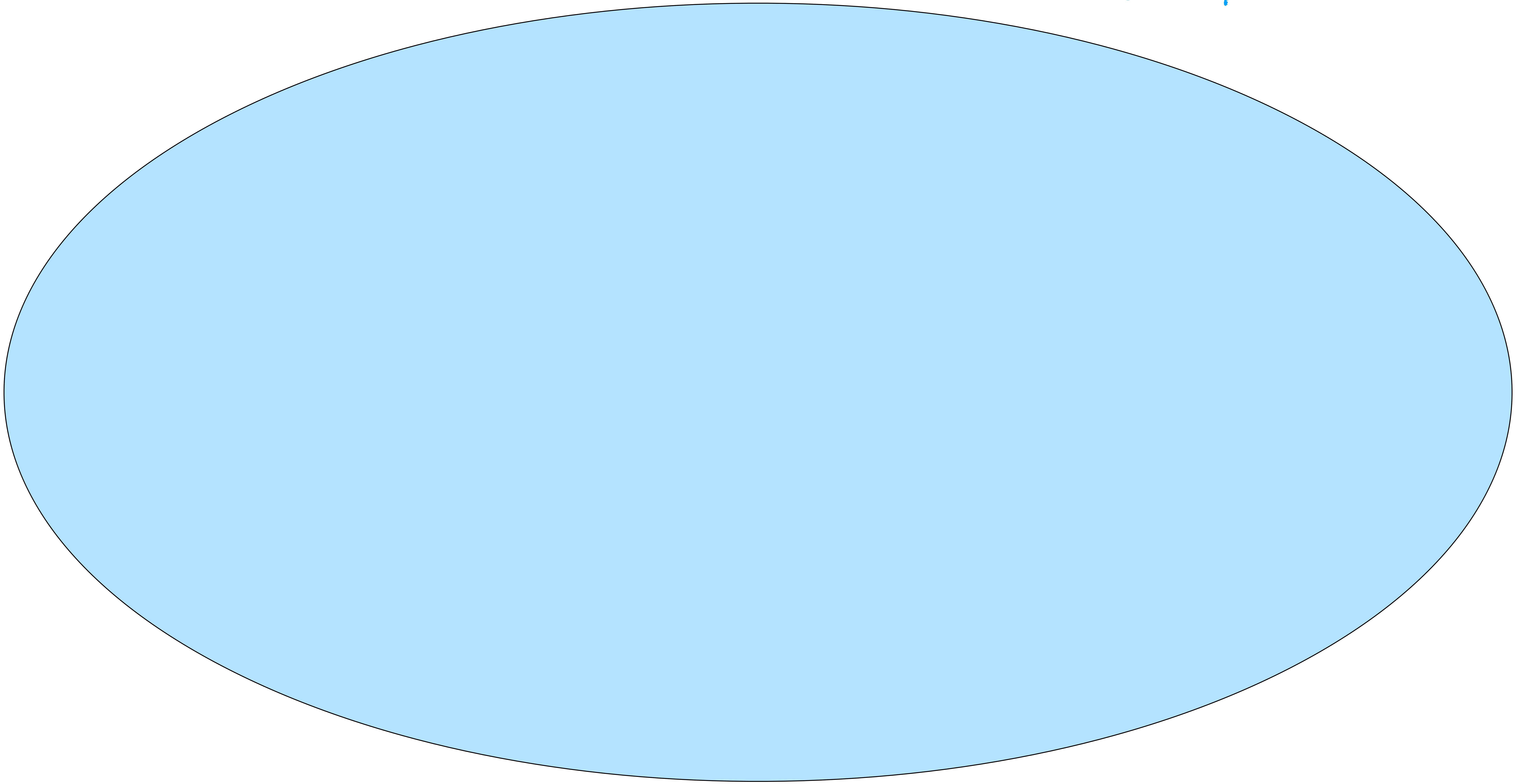
- **New Policy!**
  - The Grad TAs (Charlie, Jack) will hold conceptual hours each week
  - If you did poorly on a quiz and want to gain points back, you can go to one of their conceptual hours
  - They will cover the material and discuss in more detail. If you interact and can demonstrate understanding, they can award points back.

Topics

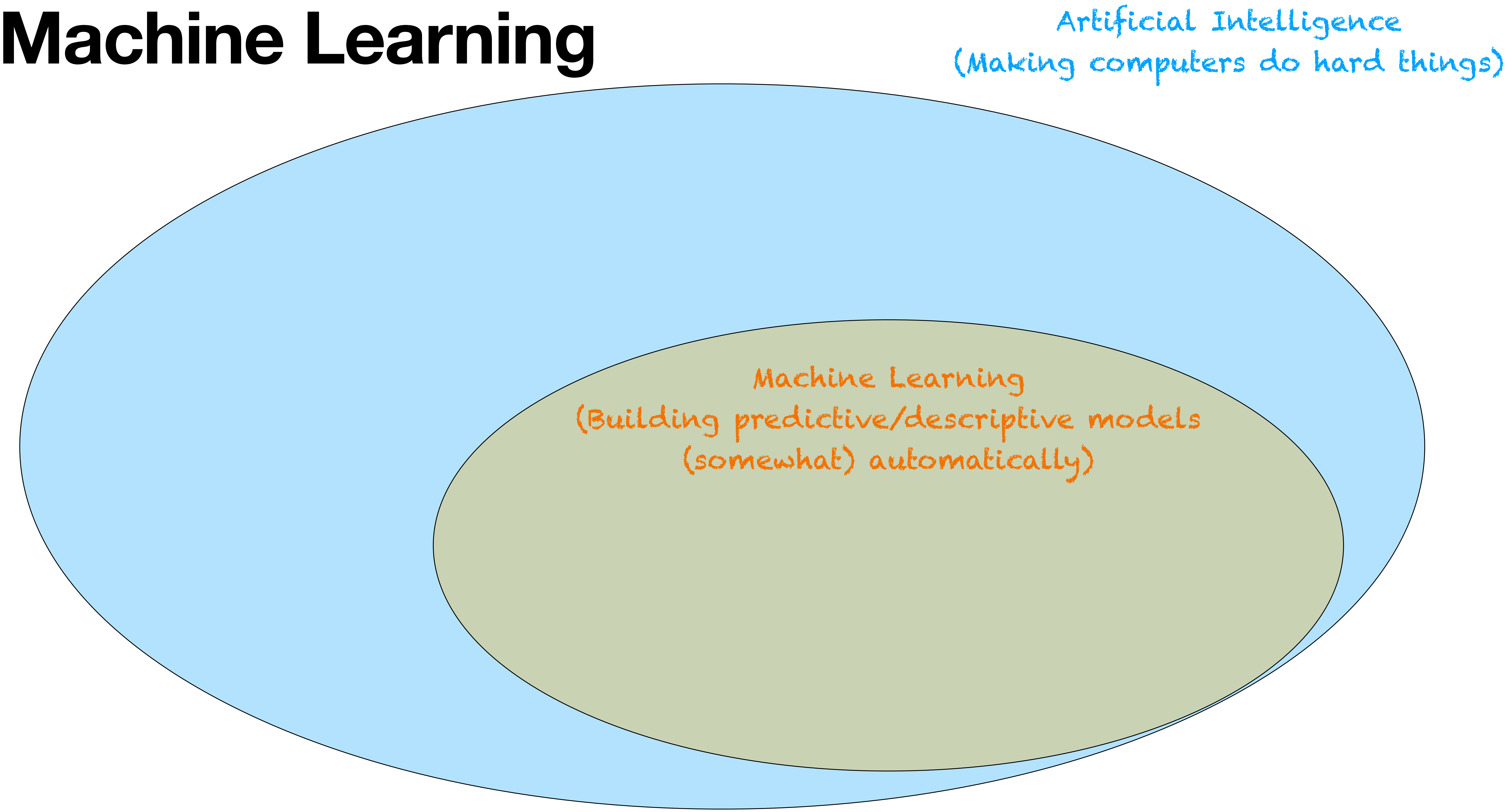
- Lecture 1 Reprise, New Quiz Makeup Policy
- **Supervised Classification**
- Feature Matrices and BOW Models
- Naive Bayes Text Classifier
- Logistic Regression Text Classifier
- Experimental Design in ML

Machine Learning

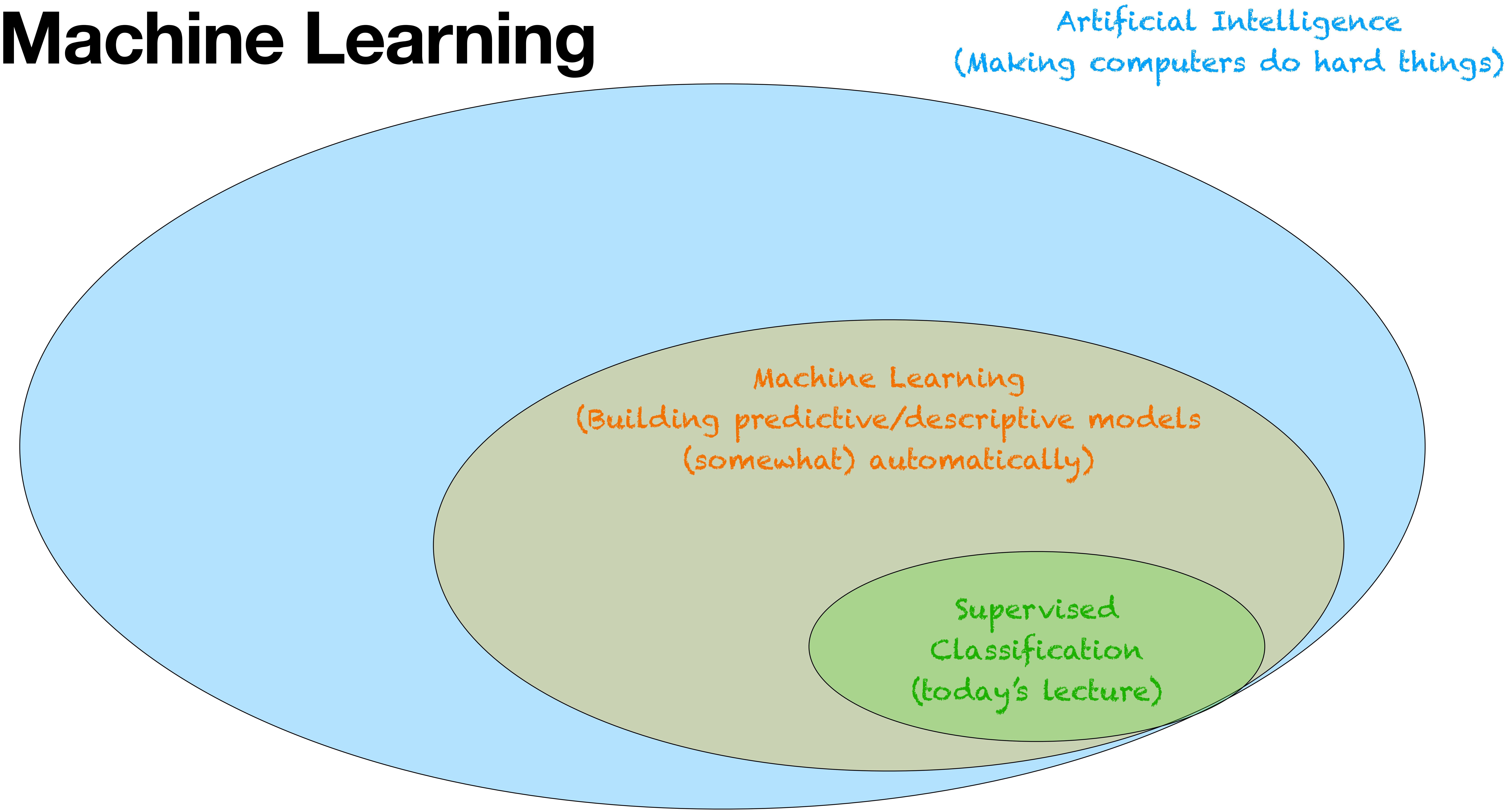
Artificial Intelligence
(Making computers do hard things)



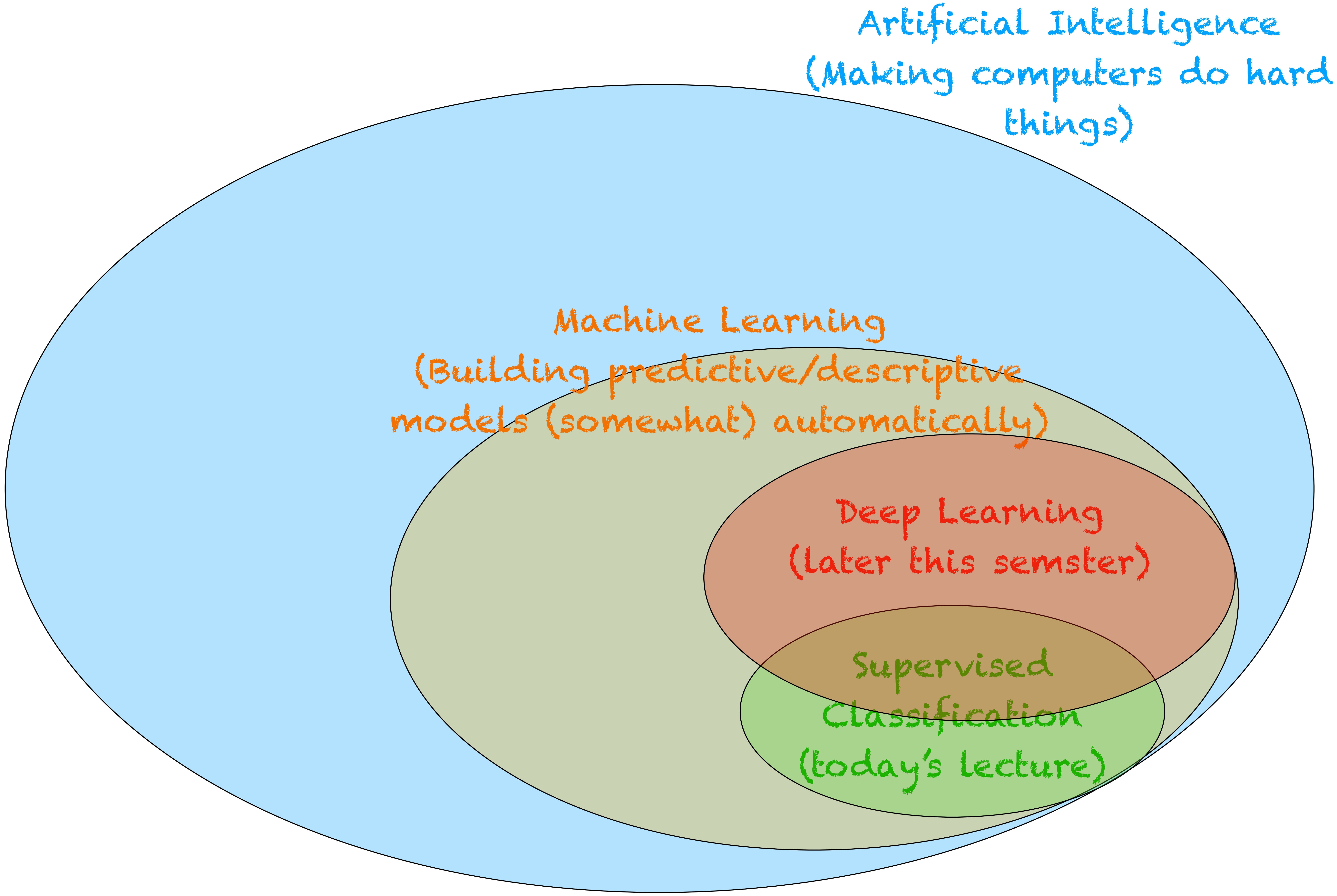
Machine Learning



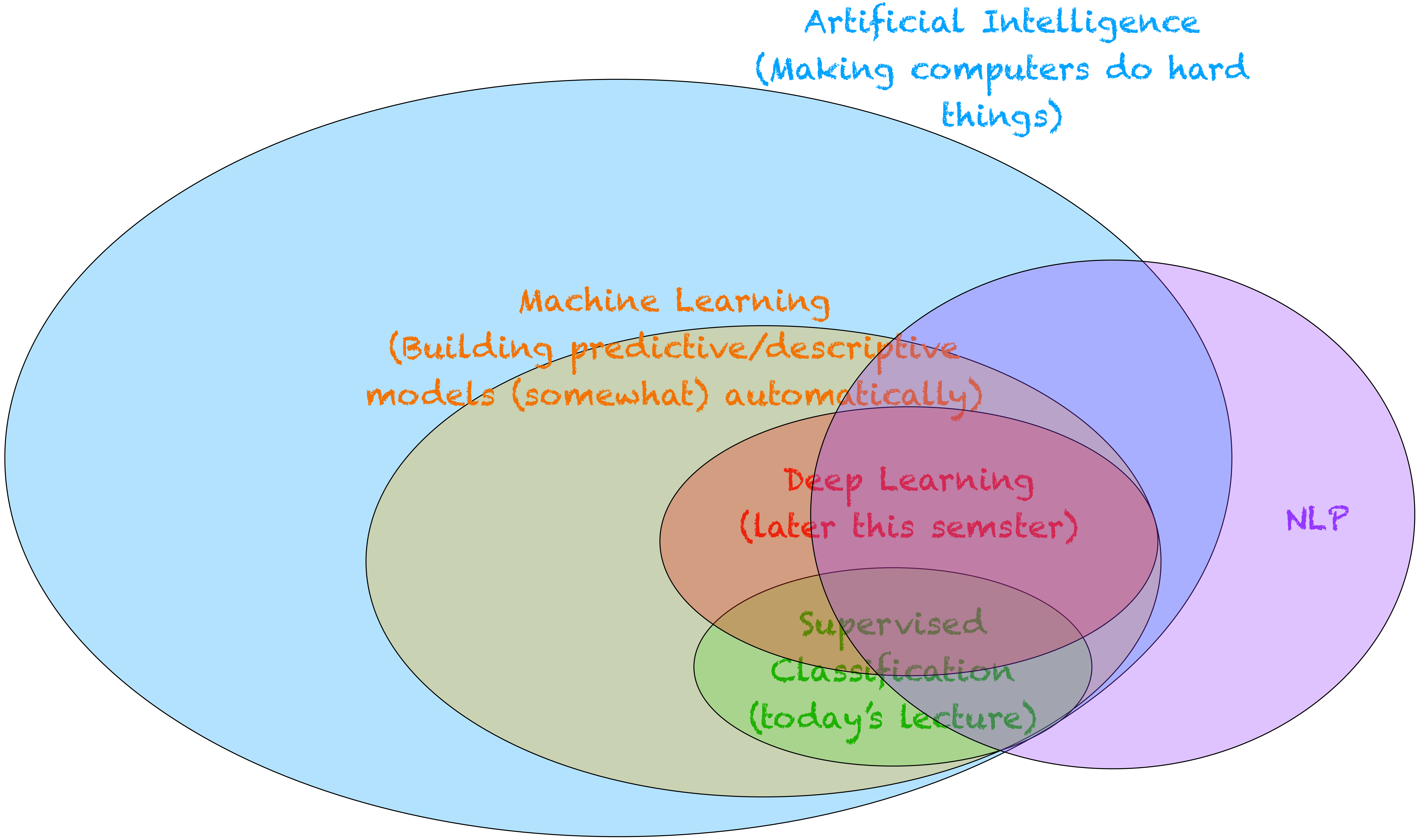
Machine Learning



Machine Learning



Machine Learning



Supervised Classification

Classification vs. Regression

Supervised Classification

Classification vs. Regression

- Classification: Your goals is to assign (discrete) labels to inputs. E.g.,
  - Is the sentiment positive or negative?
  - Which of the following topics is this article about: politics, sports, fashion?
  - Will the stock price go up or down?

Supervised Classification

Classification vs. Regression

- Classification: Your goals is to assign (discrete) labels to inputs. E.g.,
  - Is the sentiment positive or negative?
  - Which of the following topics is this article about: politics, sports, fashion?
  - Will the stock price go up or down?
- Regression: Your goal is to predict a real-valued number for a given input. E.g.,
  - How long will a person spend reading this article?
  - How many likes will this tweet get?
  - What will be the price of this company's stock tomorrow?

Supervised Classification

Supervised vs. Unsupervised

Supervised Classification

Supervised vs. Unsupervised

- Supervised: You have some examples of inputs and their true label
  - I have tons of product reviews and associated star-ratings on Amazon. Given the text of a review, can I predict the star rating?
  - I have lots of sentences. Given words 1…k in a sentence, can I predict word k+1?
  - I have lots of English documents that have been manually translated into Arabic. Given a new English document, can I translate it into Arabic?

Supervised Classification

Supervised vs. Unsupervised

- Supervised: You have some examples of inputs and their true label
  - I have tons of product reviews and associated star-ratings on Amazon. Given the text of a review, can I predict the star rating?
  - I have lots of sentences. Given words 1…k in a sentence, can I predict word k+1?
  - I have lots of English documents that have been manually translated into Arabic. Given a new English document, can I translate it into Arabic?
- Unsupervised: You only have unlabelled inputs. E.g.,
  - I have a ton of news articles. Can I cluster them into meaningful groups?
  - I have a collection of tweets from politicians. Can I detect when the discourse is changing? I.e., when new topics/priorities are emerging?

Supervised Classification

Running Example

- Predict whether or not an article will be clicked on
- Input: Article Title
- Output: 1 (clicked), 0 (not clicked)

Supervised Classification

Example

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

Supervised Classification

Example

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

“We compared 24 brands of dryer sheet so you don’t have to… 👔🩳👚”

????

Topics

- Lecture 1 Reprise, New Quiz Makeup Policy
- Supervised Classification
- **Feature Matrices and BOW Models**
- Naive Bayes Text Classifier
- Logistic Regression Text Classifier
- Experimental Design in ML

Supervised Classification

Example

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

“We compared 24 brands of dryer sheet so you don’t have to… 👔🩳👚”

????

Supervised Classification

Simple Nearest Neighbors Classifier

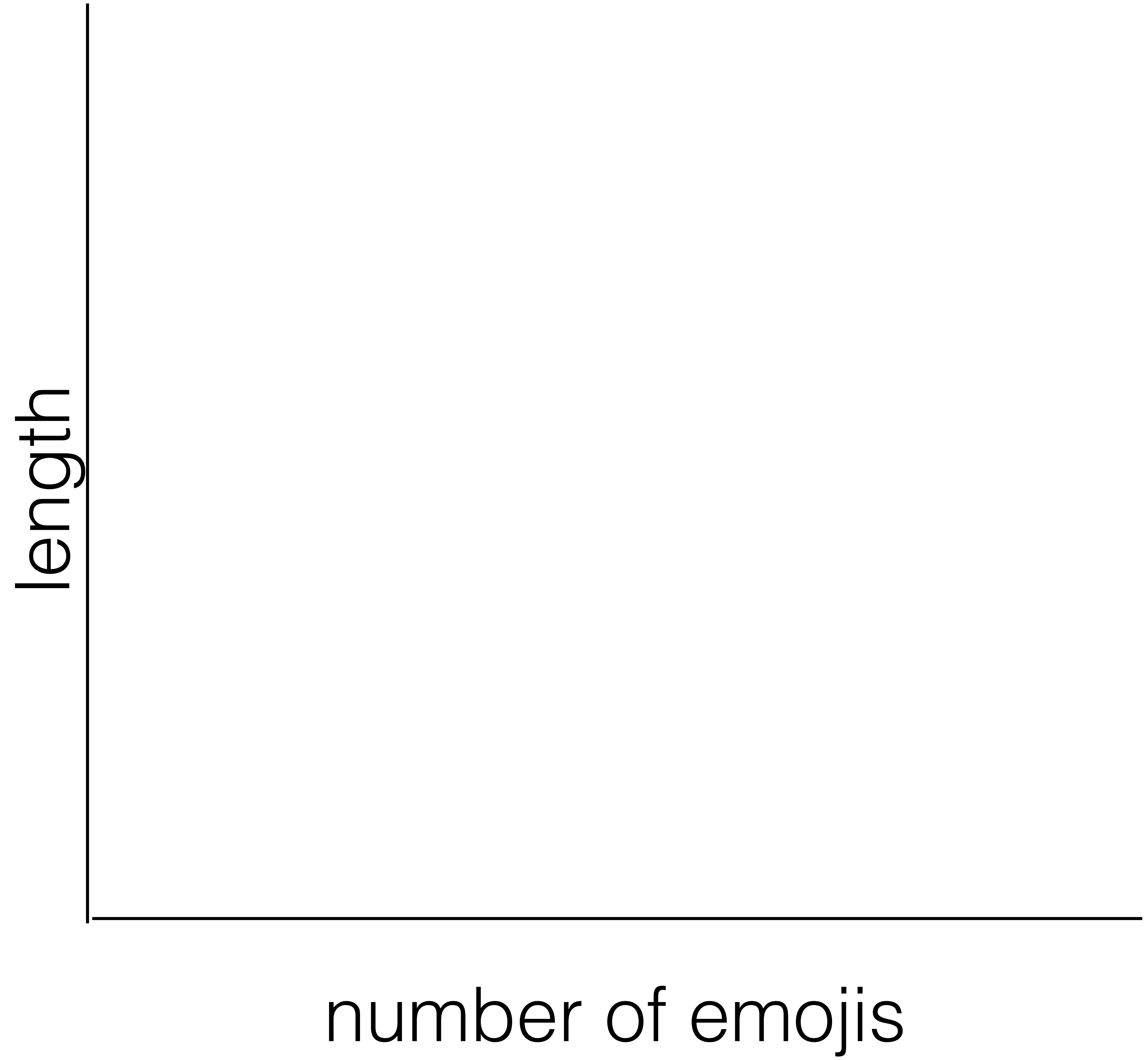
- Goal: Predict whether or not an article will be clicked on
- Basic Idea: Given a new article, find the **most similar** training article and assume it has the same label
- But you need to define “similar”. This will depend on what **features** you use!

Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

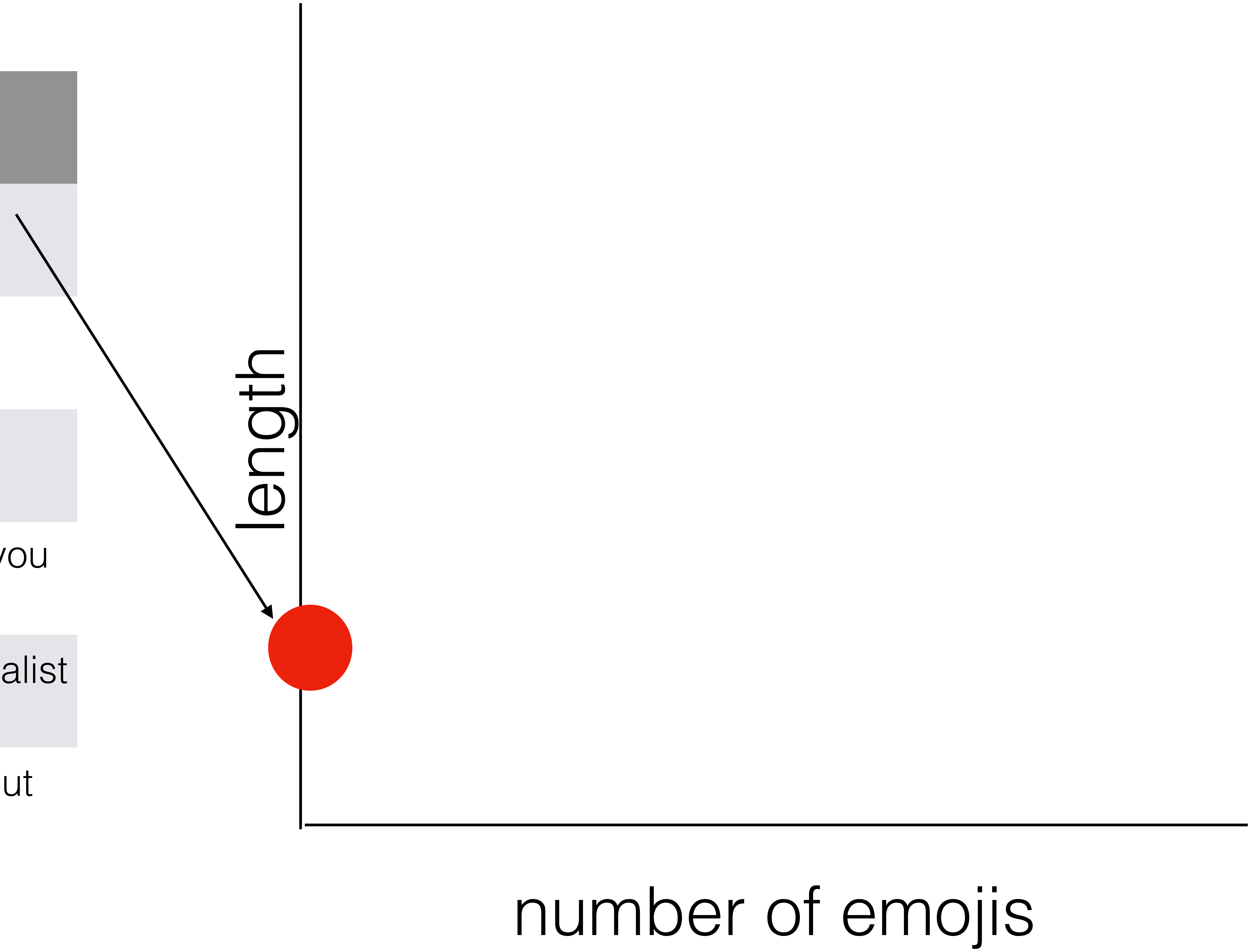


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

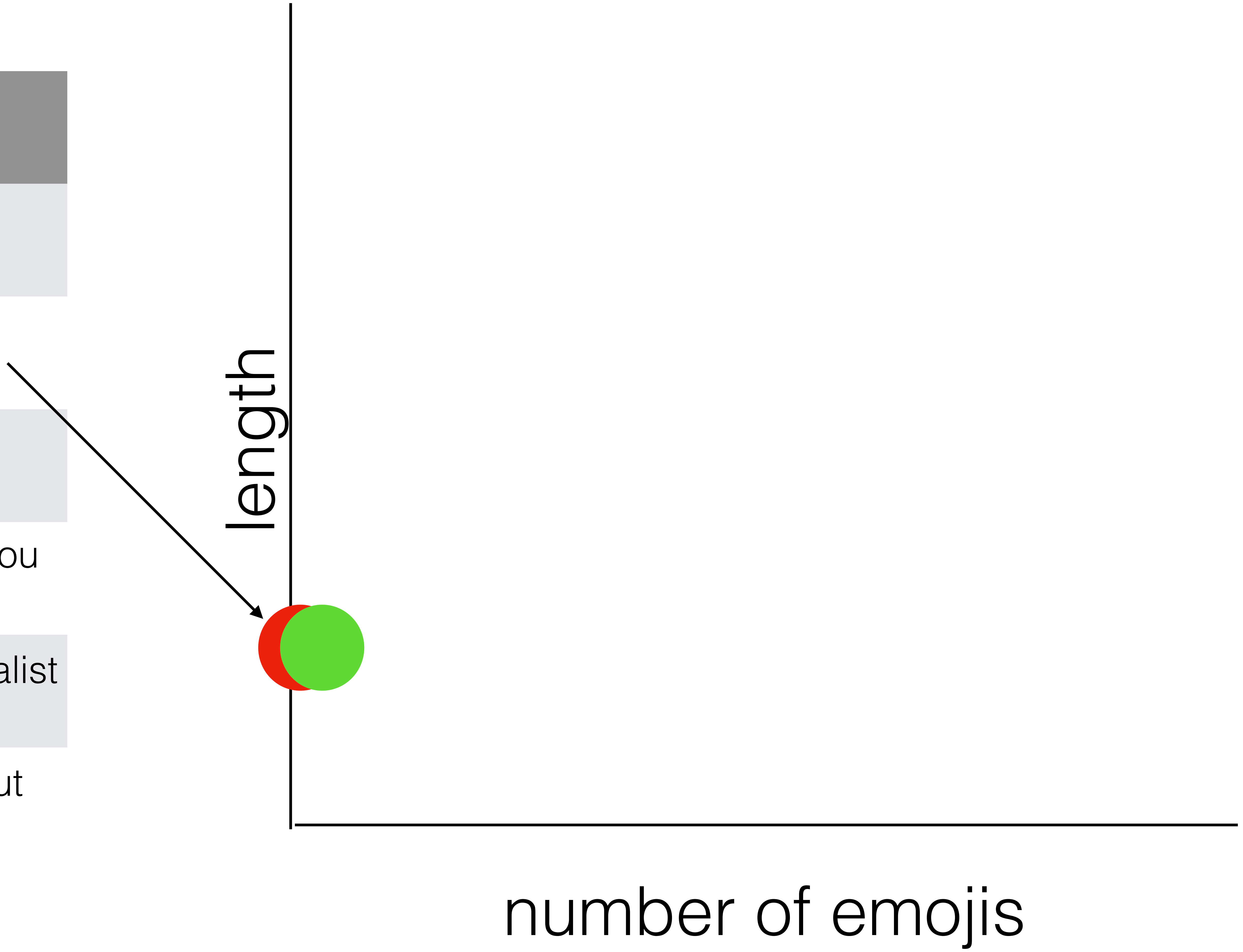


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

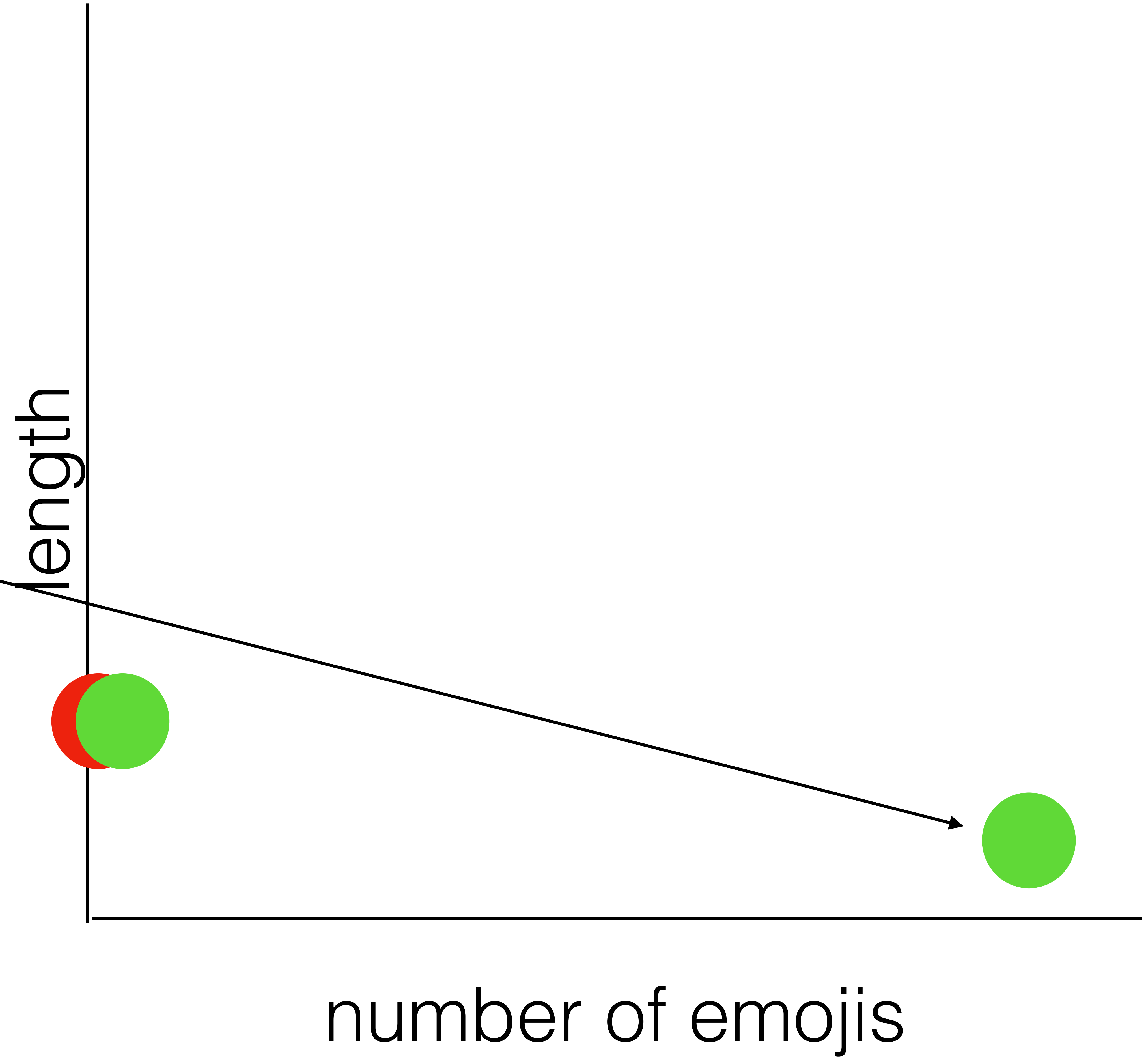


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

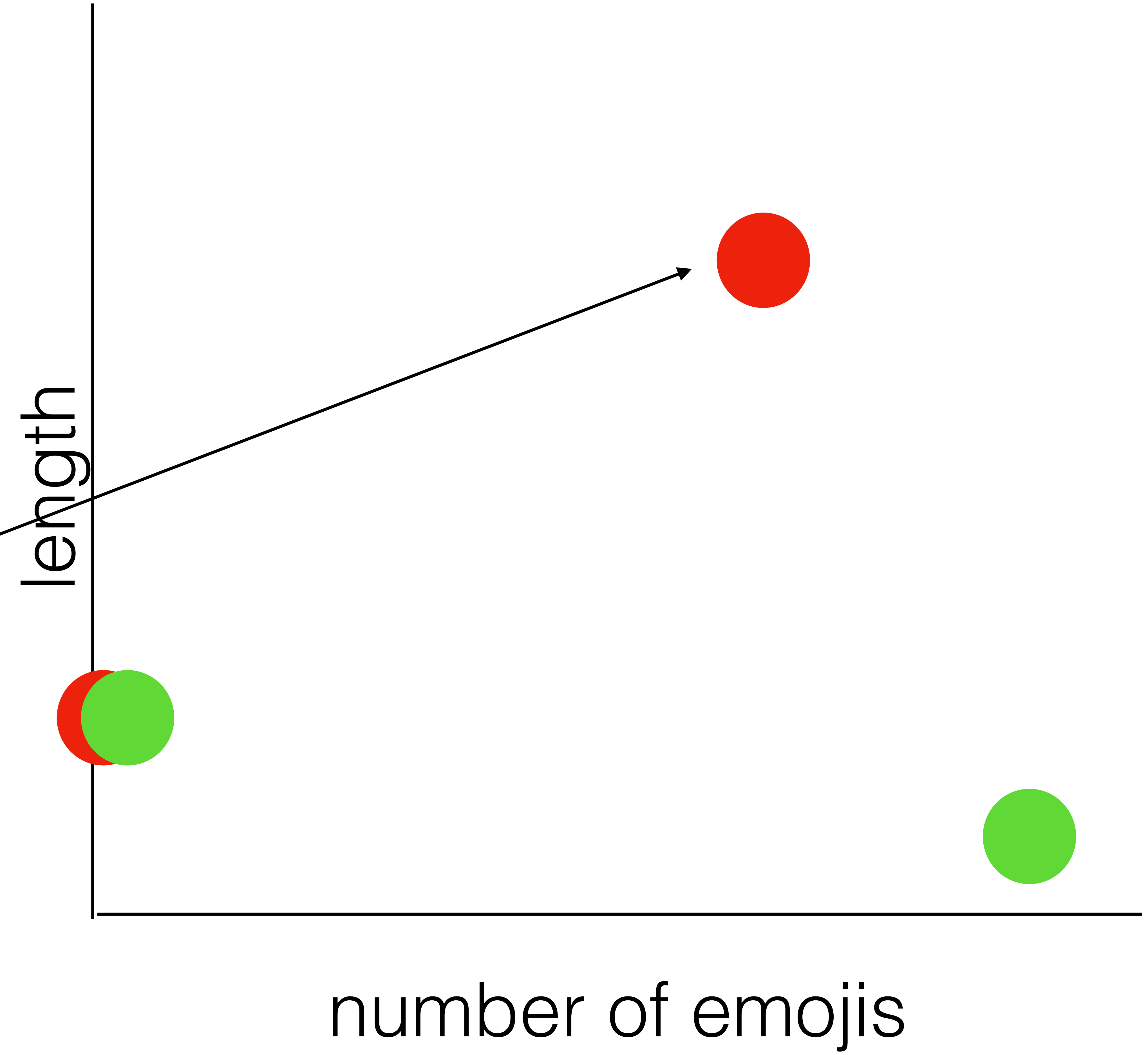


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

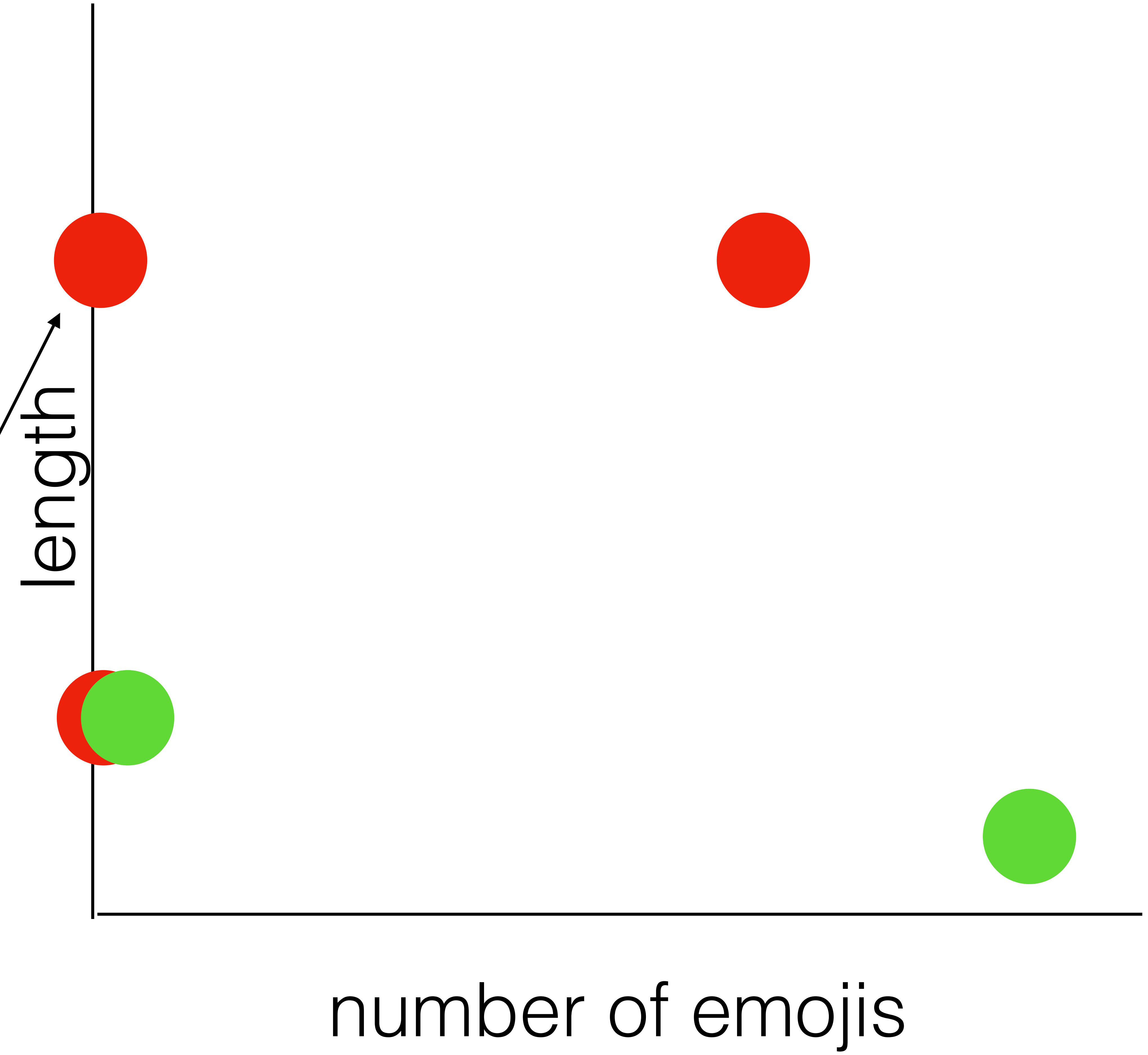


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	"New Tax Guidelines"
1	"This 600lb baby…"
1	"This. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets

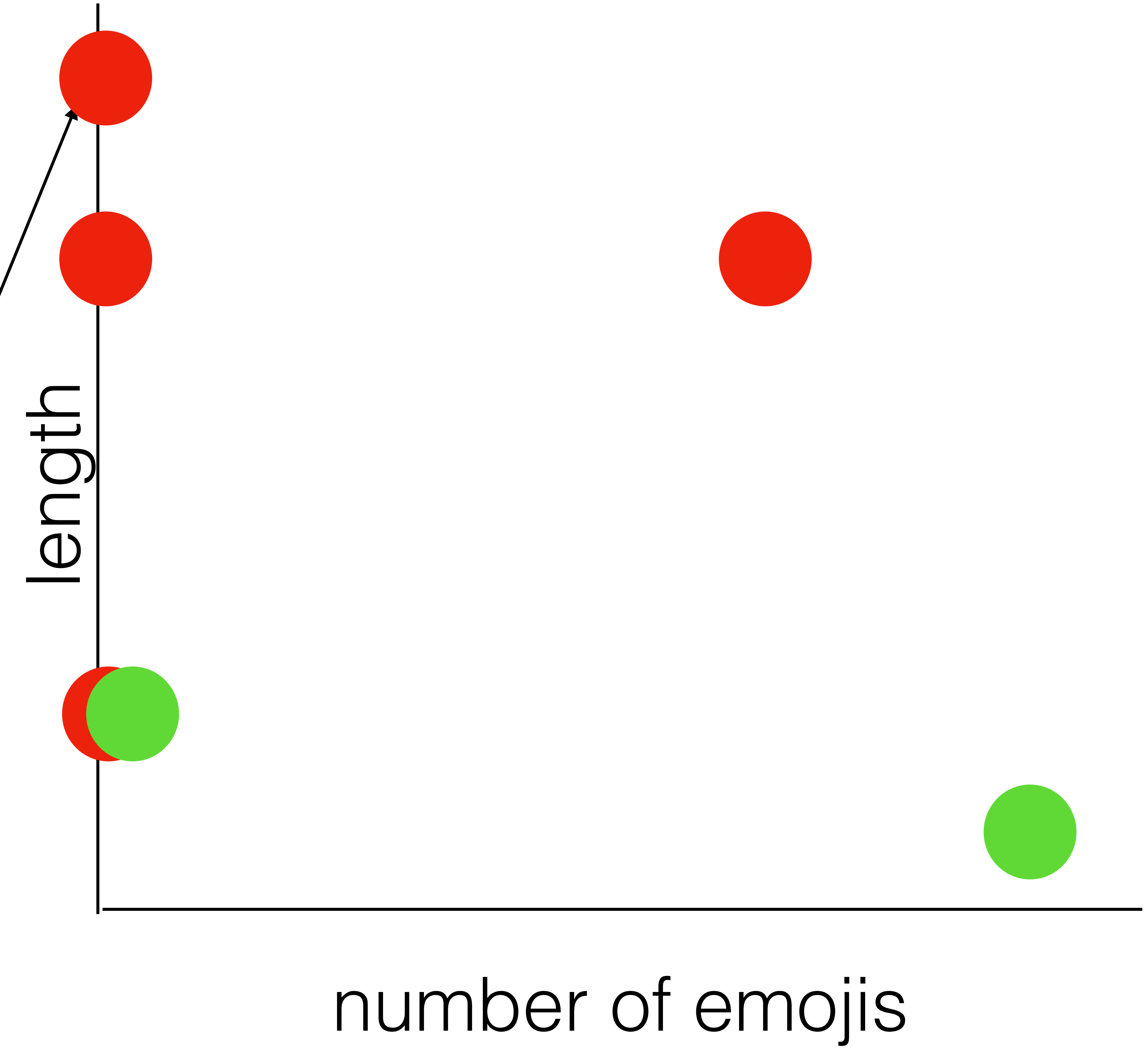


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	"New Tax Guidelines"
1	"This 600lb baby…"
1	"This. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets



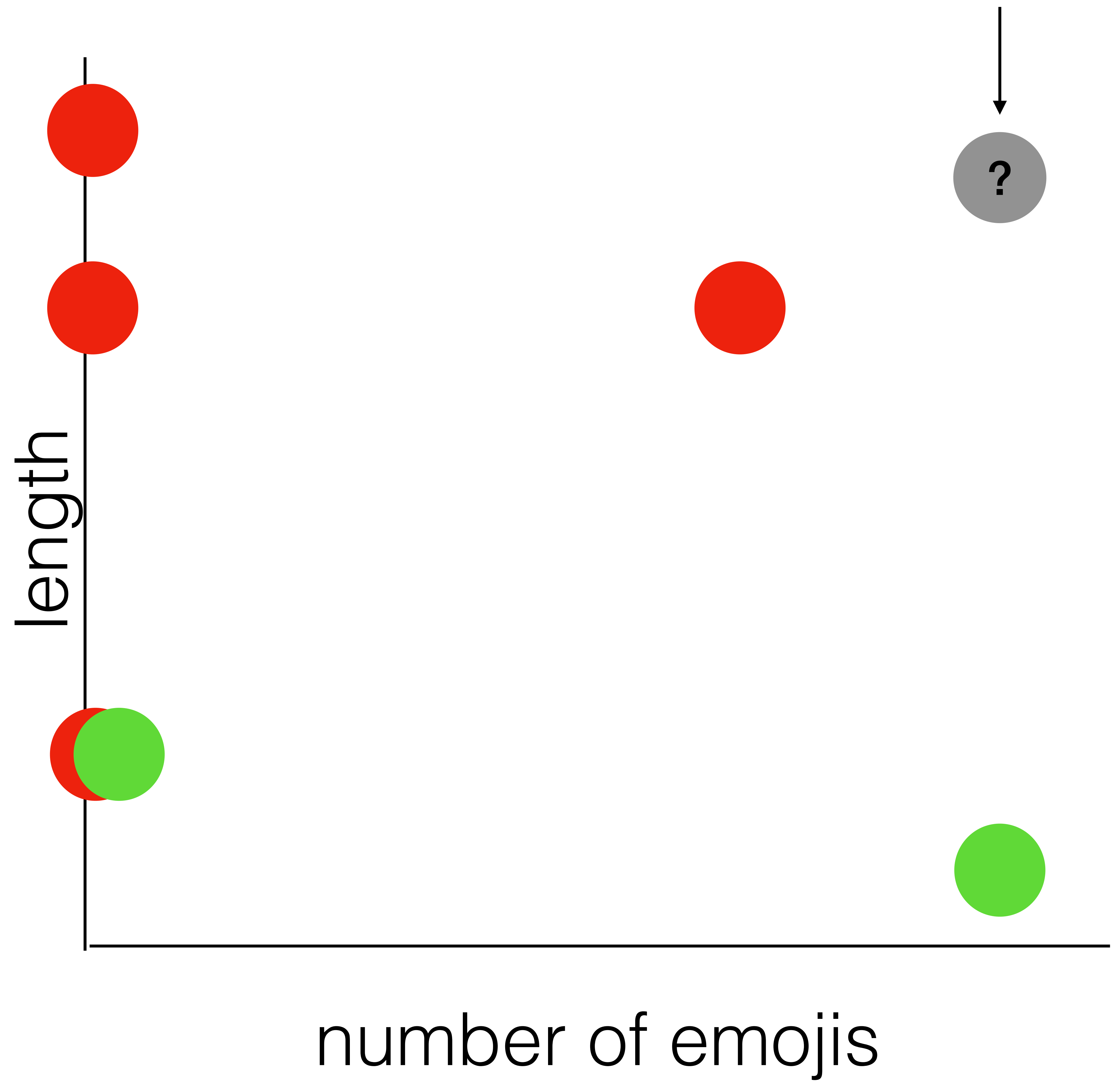
Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

“We compared 24 brands of dryer sheet so you don’t have to… 👔🩳👚”

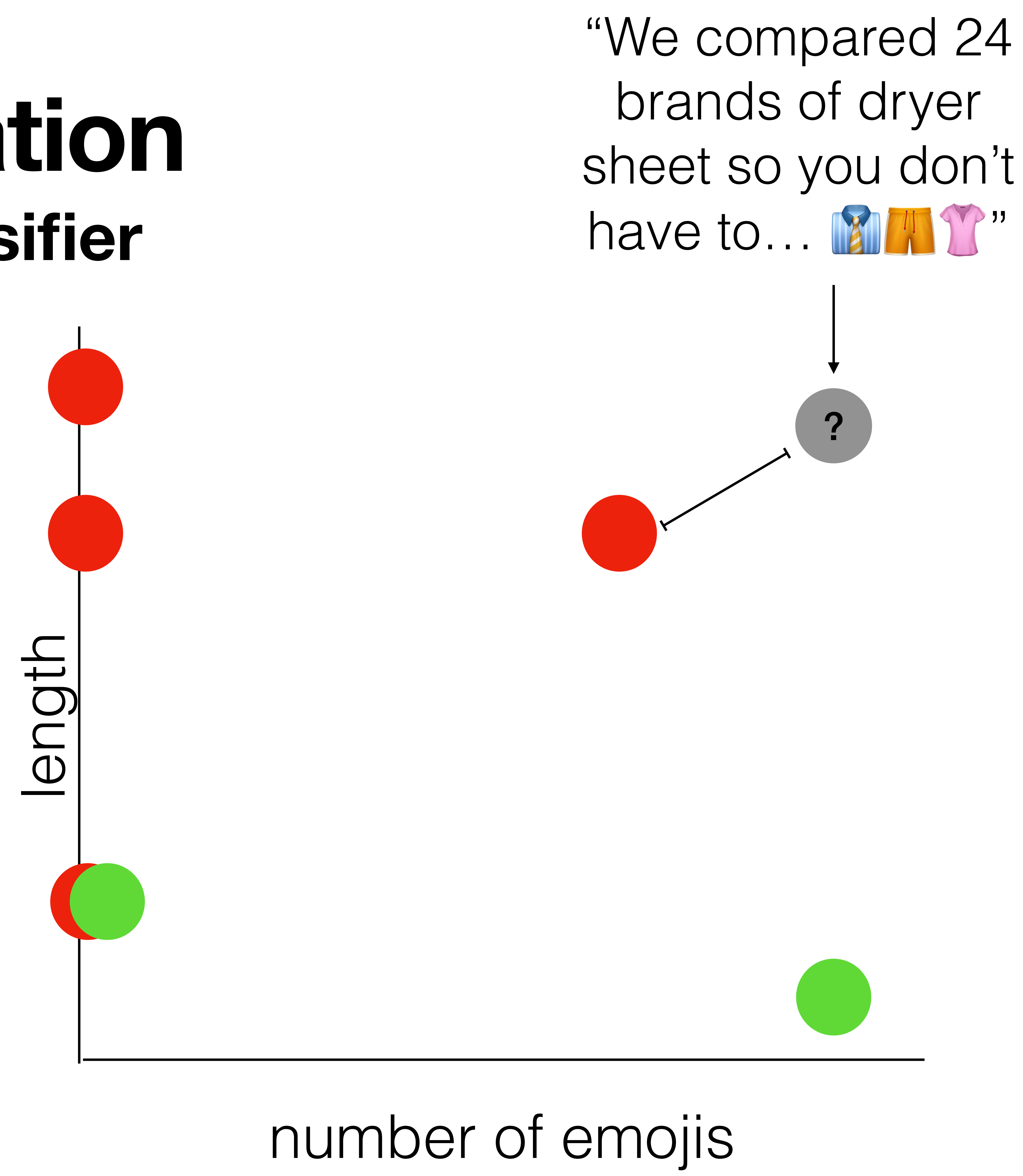


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

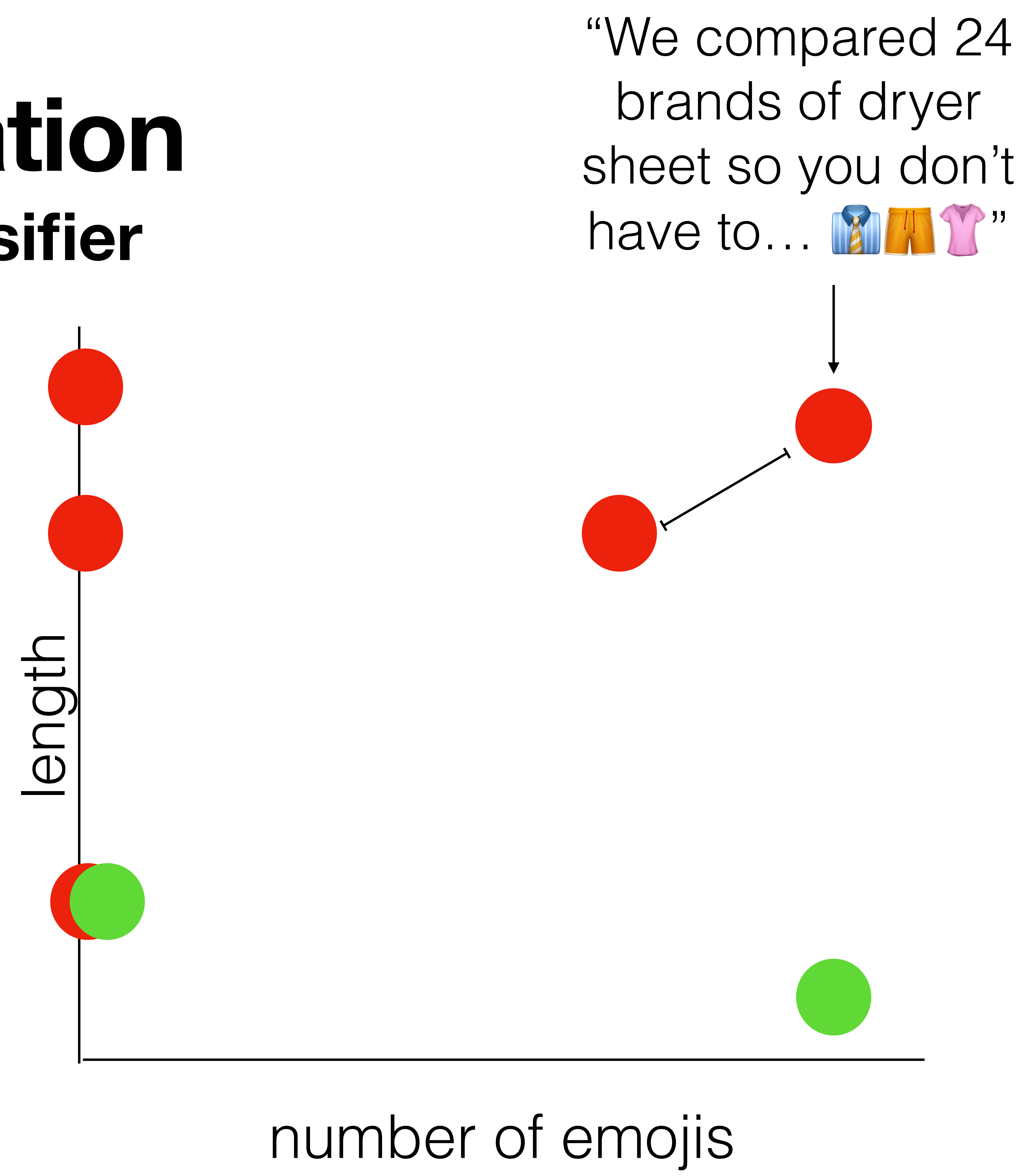


Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets



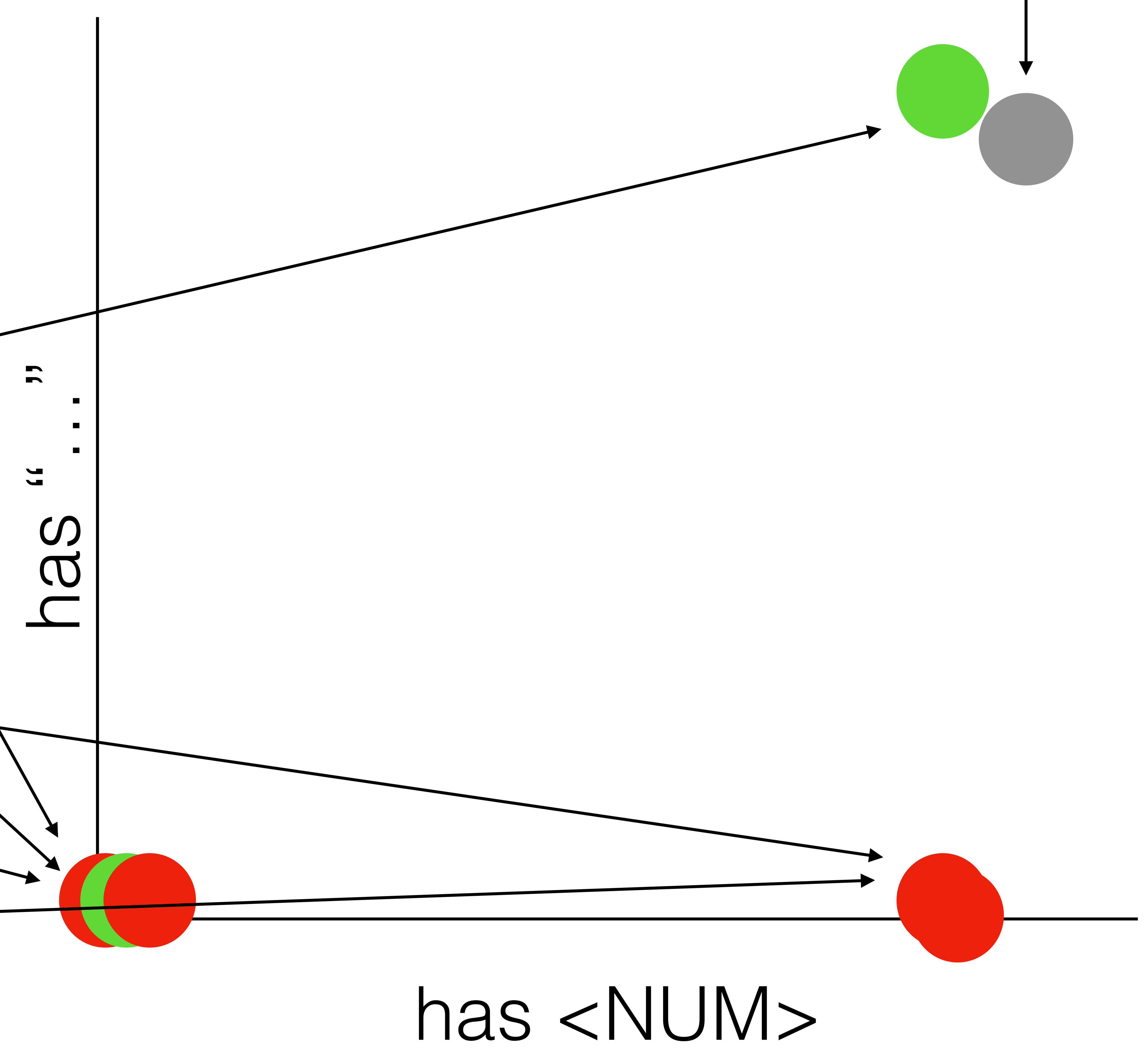
Supervised Classification

Simple Nearest Neighbors Classifier

"We compared 24 brands of dryer sheet so you don't have to… 👔🩳👚"

Training Data

Label	Article Title
0	"New Tax Guidelines"
1	"This 600lb baby…"
1	"This. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets



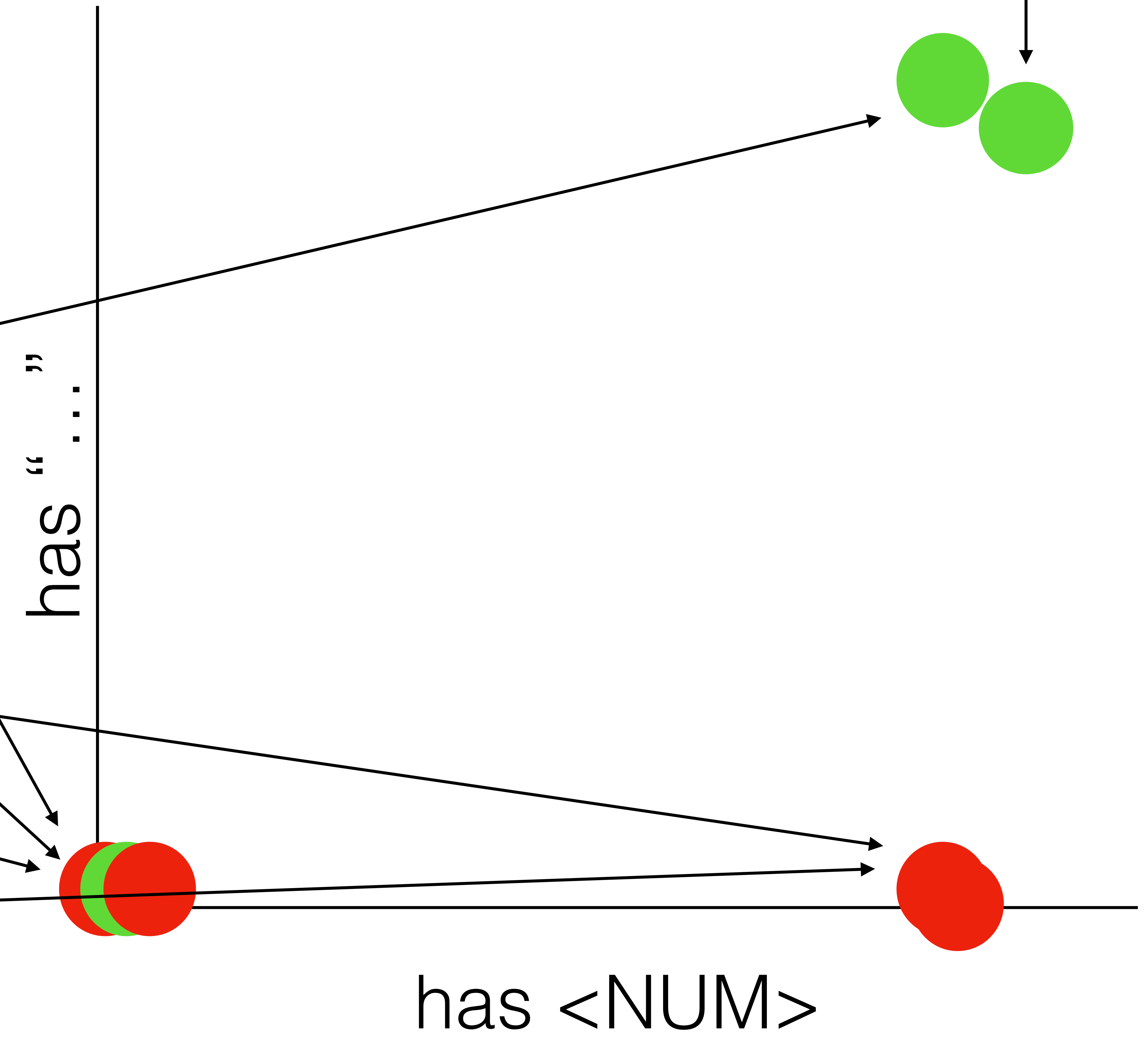
Supervised Classification

Simple Nearest Neighbors Classifier

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

“We compared 24 brands of dryer sheet so you don’t have to… 👔🩳👚”



Supervised Classification

Building Feature Matrices

- ML models require input to be represented as numeric **features**
- These can be real-valued (e.g., length of title)
- Or they can be binary (e.g., does “…” appear in the title?)
- These features are encoded in a **feature matrix**

Supervised Classification

Building Feature Matrices

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

Feature Matrix

length	number of emojis	contains NUM	contains “…”
3	0	0	0
3	0	1	1
2	3	0	0
11	2	0	0
6	0	0	0
11	0	1	0

Supervised Classification

Building Feature Matrices

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

“raw data”

Feature Matrix

length	number of emojis	contains NUM	contains “…”
3	0	0	0
3	0	1	1
2	3	0	0
11	2	0	0
6	0	0	0
11	0	1	0

Supervised Classification

Building Feature Matrices

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

Feature Matrix

length	number of emojis	contains NUM	contains “…”
3	0	0	0
3	0	1	1
2	3	0	0
11	2	0	0
6	0	0	0
11	0	1	0

X

Supervised Classification

Building Feature Matrices

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

y

Feature Matrix

length	number of emojis	contains NUM	contains “…”
3	0	0	0
3	0	1	1
2	3	0	0
11	2	0	0
6	0	0	0
11	0	1	0

Supervised Classification

Basic Bag of Words Model

- Bag of Words (BOW) Model: Model that uses the words as features
- No information about order or syntax
    - “10 reasons why cats are way better than dogs” = “10 reasons why dogs are way better than cats”
- Very strong starting point for NLP models
    - Actually, often annoyingly so. ;) Often very hard to improve over a basic BOW model for many tasks

Supervised Classification

Basic Bag of Words Model

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

Feature Matrix

New	Tax	This	600	baby	…	👇	18	healt hy	foods	that
1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Supervised Classification

Basic Bag of Words Model

Binary
(Word is either present or not. But we'll discuss some variants next lecture.)

Training Data

Label	Article Title
0	"New Tax Guidelines"
1	"This 600lb baby…"
1	"This. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets

Feature Matrix

New	Tax	This	600	baby	…	👇	18	healt hy	foods	that
1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Supervised Classification

Basic Bag of Words Model

Very High Dimensional
(usually 10s or 100s of 1000s of
features)

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

Feature Matrix

New	Tax	This	600	baby	…	👇	18	healt hy	foods	that
1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Supervised Classification

Basic Bag of Words Model

"Sparse"
The vast majority of values are 0

Training Data

Label	Article Title
0	"New Tax Guidelines"
1	"This 600lb baby…"
1	"This. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets

Feature Matrix

New	Tax	This	600	baby	…	👇	18	healt hy	foods	that
1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Supervised Classification

Basic Bag of Words Model

Training Data

Label	Article Title
0	“New Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they **don’t** want **you to** know about **dryer sheets**

“We compared 24 brands of **dryer sheet** so **you don’t** have **to**… 👔🩳👚”

????

Topics

- Lecture 1 Reprise, New Quiz Makeup Policy
- Supervised Classification
- Feature Matrices and BOW Models
- **Naive Bayes Text Classifier**
- Logistic Regression Text Classifier
- Experimental Design in ML

Naive Bayes Classifiers

- Nearest Neighbors Classifier gives a simple way of choosing the best label Y given some features X
- But what if we want to be more principled? Actually, estimate/maximize P(Y|X)
- Naive Bayes is one model for doing this
- It relies on Bayes Rule in order to do the computations

Naive Bayes Classifiers

- Hard to estimate P(Y|X) directly
  - (X is a complex distribution, and we aren't going to see lots of examples of the same X)
- Use Bayes Rule to flip the computation around, and then make it more tractable

	X										
Y	New	Tax	This	600	baby	…	👇	18	healthy	foods	that
0	1	1	0	0	0	0	0	0	0	0	0
1	0	0	1	1	1	1	0	0	0	0	0
1	0	0	1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Naive Bayes Classifiers

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Naive Bayes Classifiers

Bayes Rule

Likelihood Prior

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Posterior

Marginal

Naive Bayes Classifiers

Bayes Rule

	X				
Y	New	Tax	This	600	baby
0	1	1	0	0	0

Naive Bayes Classifiers

Bayes Rule

X

Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(Y=0 \mid New=1, Tax=1, This=0, \ldots)$$

Naive Bayes Classifiers

Bayes Rule

X

Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(Y=0|New=1, Tax=1, This=0, \ldots)$$

$$P(New=1, Tax=1, This=0, \ldots|Y=0)P(Y=0)$$ Bayes Rule

Naive Bayes Classifiers

Bayes Rule

X					
Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(Y=0|New=1, Tax=1, This=0, \ldots)$$

$$P(New=1, Tax=1, This=0, \ldots|Y=0)P(Y=0)$$

$$P(New=1, Tax=1, This=0, \ldots, Y=0)$$

Equivalent to joint distribution of label and features

Naive Bayes Classifiers

Bayes Rule

X

Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(Y=0|New=1, Tax=1, This=0, \ldots)$$

$$P(New=1, Tax=1, This=0, \ldots|Y=0)P(Y=0)$$

$$P(New=1, Tax=1, This=0, \ldots, Y=0)$$

$$P(New=1 \mid Tax=1, This=0, \ldots, Y=0)\; P(Tax=1, This=0, \ldots, Y=0)$$

"Chain Rule"

Naive Bayes Classifiers

Bayes Rule

X

Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(C|x_1, x_2, \ldots, x_k)$$
$$=P(x_1|x_2, \ldots, x_k, C)P(x_2|x_3, \ldots, x_k, C)\ldots P(x_k|C)P(C)$$

Can keep applying chain rule
to produce an equation with
one term per feature (x_i)

Naive Bayes Classifiers

Bayes Rule

	X				
Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(C|x_1, x_2, \ldots, x_k)$$
$$=P(x_1|x_2, \ldots, x_k, C)P(x_2|x_3, \ldots, x_k, C)\ldots P(x_k|C)P(C)$$

Wait—what was the point of this? We are still stuck with hard to estimate quantities (since most feature combos are only seen once)

Naive Bayes Classifiers

Bayes Rule

	X				
Y	New	Tax	This	600	baby
0	1	1	0	0	0

$$P(C|x_1, x_2, \ldots, x_k)$$
$$=P(x_1|x_2, \ldots, x_k, C)P(x_2|x_3, \ldots, x_k, C)\ldots P(x_k|C)P(C)$$
$$=P(x_1|C)P(x_2|C)\ldots P(x_k|C)P(C)$$

Naive Assumption: Assume features are independent

Naive Bayes Classifiers

Worked Example

	X			
Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

x	P(x|Y=1)	P(x|Y=0)
Tax	??	??
This	??	??
600	??	??
👇	??	??

Naive Bayes Classifiers

Worked Example

	X			
Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

x	P(x|Y=1)	P(x|Y=0)
Tax	??	??
This	??	??
600	??	??
👇	??	??

Naive Bayes Classifiers

Worked Example

Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	??
This	??	??
600	??	??
👇	??	??

Naive Bayes Classifiers

Worked Example

	X			
Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	??	??
600	??	??
👇	??	??

Naive Bayes Classifiers

Worked Example

	X			
Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	??
600	??	??
👇	??	??

Naive Bayes Classifiers

Worked Example

X

Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	??	??
👇	??	??

Naive Bayes Classifiers

Worked Example

	X			
Y	Tax	This	600	👇
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1
1	0	0	1	1

Model Parameters

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5

Naive Bayes Classifiers

Worked Example

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

600 Awesome Tax Policies

???

Naive Bayes Classifiers

Worked Example

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

600 Awesome Tax Policies

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

Naive Bayes Classifiers

Worked Example

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

600 Awesome Tax Policies

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

Naive Bayes Classifiers

Worked Example

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

600 Awesome Tax Policies

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

Domain knowledge or estimate from data

Naive Bayes Classifiers

Worked Example

P(Y=1)	P(Y=0)
0.3	0.7

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

600 Awesome Tax Policies

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

Domain knowledge or estimate from data

Naive Bayes Classifiers

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

Worked Example

P(Y=1)	P(Y=0)
0.1	0.9

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

P(Y=0 l"600 Awesome Tax Policies") = 0.075 x 0.9 = 0.0675

Naive Bayes Classifiers

Worked Example

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

P(Y=1)	P(Y=0)
0.1	0.9

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

P(Y=0 |“600 Awesome Tax Policies”) = 0.075 x 0.9 = 0.0675

P(Y=1 |“600 Awesome Tax Policies”) = 1.0 x 0.6 x 0.5 x 0.2 x 0.1

Naive Bayes Classifiers

Worked Example

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

P(Y=1)	P(Y=0)
0.1	0.9

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

P(Y=0 |“600 Awesome Tax Policies”) = 0.075 x 0.9 = 0.0675

P(Y=1 |“600 Awesome Tax Policies”) = 0.06 x 0.1 = 0.006

Naive Bayes Classifiers

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

Worked Example

P(Y=1)	P(Y=0)
0.1	0.9

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	0.2	0.5

P(Y=0 l“600 Awesome Tax Policies”) = 0.075 x 0.9 = 0.0675

P(Y=1 l“600 Awesome Tax Policies”) = 0.06 x 0.1 = 0.006

Prediction:
y = 0

Naive Bayes Classifiers

Worked Example

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

P(Y=1)	P(Y=0)
0.1	0.9

x	P(x|Y=1)	P(x|Y=0)
Tax	0.5	1.0
This	0.5	0.5
600	1.0	0.5
👇	1.0	0.5
Awesome	0.6	0.3
Policies	**1.0**	0.5

Note: It is possible for the prior to outweigh the evidence!

P(Y=0 |“600 Awesome Tax Policies”) = 0.5 x 0.3 x 1.0 x 0.5 x 0.9 = **0.075** x 0.9 = **0.07**

P(Y=1 |“600 Awesome Tax Policies”) = 1.0 x 0.6 x 0.5 x 1.0 x 0.1 = **0.3** x 0.1 = **0.03**

Topics

- Lecture 1 Reprise, New Quiz Makeup Policy
- Supervised Classification
- Feature Matrices and BOW Models
- Naive Bayes Text Classifier
- **Logistic Regression Text Classifier**
- Experimental Design in ML

Logistic Regression

Logistic Regression vs. Naive Bayes

- Similar to NB, we want to estimate P(Y|X)
- NB does this by modeling the **joint distribution** P(X,Y)
  - This is called a **generative** model
- LR will do this by modeling the **conditional distribution** P(Y|X) directly
  - This is called a **discriminative** model

Logistic Regression

Logistic Regression vs. Naive Bayes

- In short: Same basic goal, just different models for doing it
- Some Trends:
  - NB often thought to be better with smaller data
  - LR maybe better in general
- But good to know both! Usually best to try both and see which is better for your data/problem

Logistic Regression

Logistic Regression Overview

- Logistic Regression is based on Linear Regression
- Linear Regression is a regression model (predicts a real value)
- Logistic Regression is a classification model (predicts 0 or 1)

Linear Regression

- $y = wx + b + e$
  - y = label
  - x = feature
  - w = weight/slope
  - b = intercept
  - e = error

Linear Regression

- $y = wx + b + e$
  - y = label
  - x = feature
  - w = weight/slope
  - b = intercept
  - e = error
- num_clicks $\cong w_1x_1 + w_2x_2 + \ldots + w_nx_n$

Linear Regression

- $y = wx + b + e$
  - y = label
  - x = feature
  - w = weight/slope
  - b = intercept
  - e = error

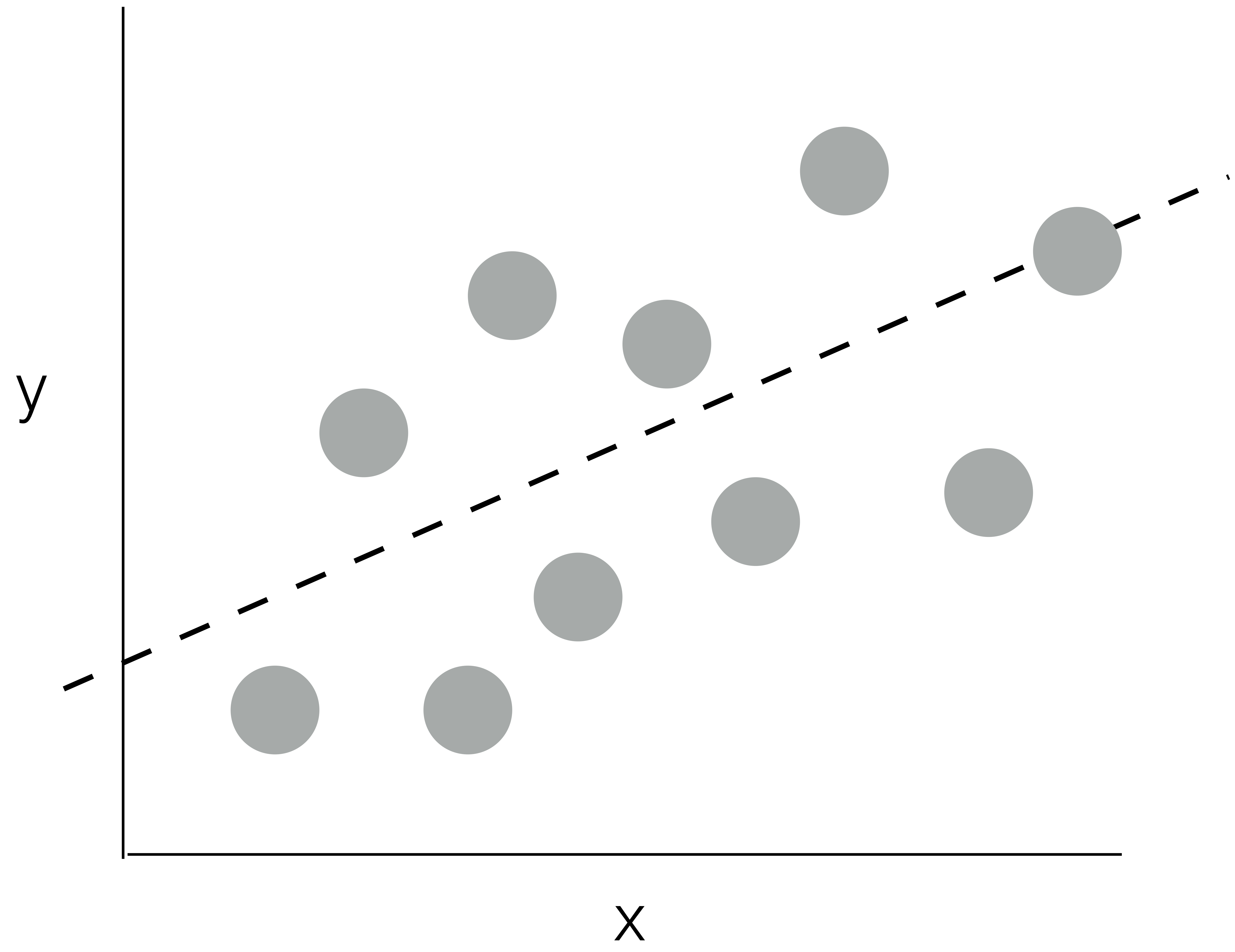
$$\begin{bmatrix} w_1 & w_2 & w_3 & \dots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix} = w_1x_1 + w_2x_2 + \dots + w_nx_n$$

dot product

- num_clicks $\cong w_1x_1 + w_2x_2 + \dots + w_nx_n = w \cdot x$

Logistic Regression Classifiers

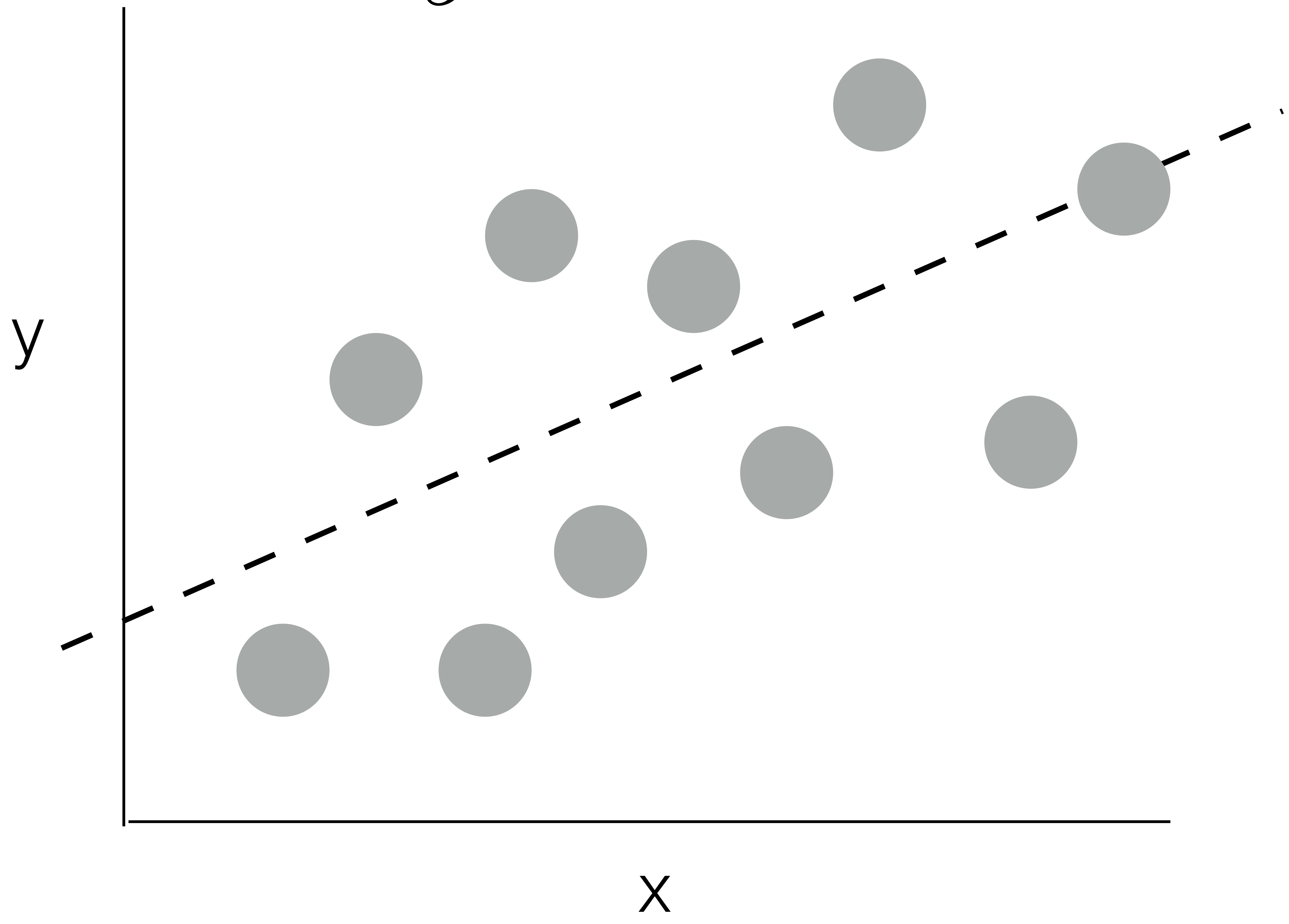
Relation between Linear Regression and Logistic Regression



Logistic Regression Classifiers

Relation between Linear Regression and Logistic Regression

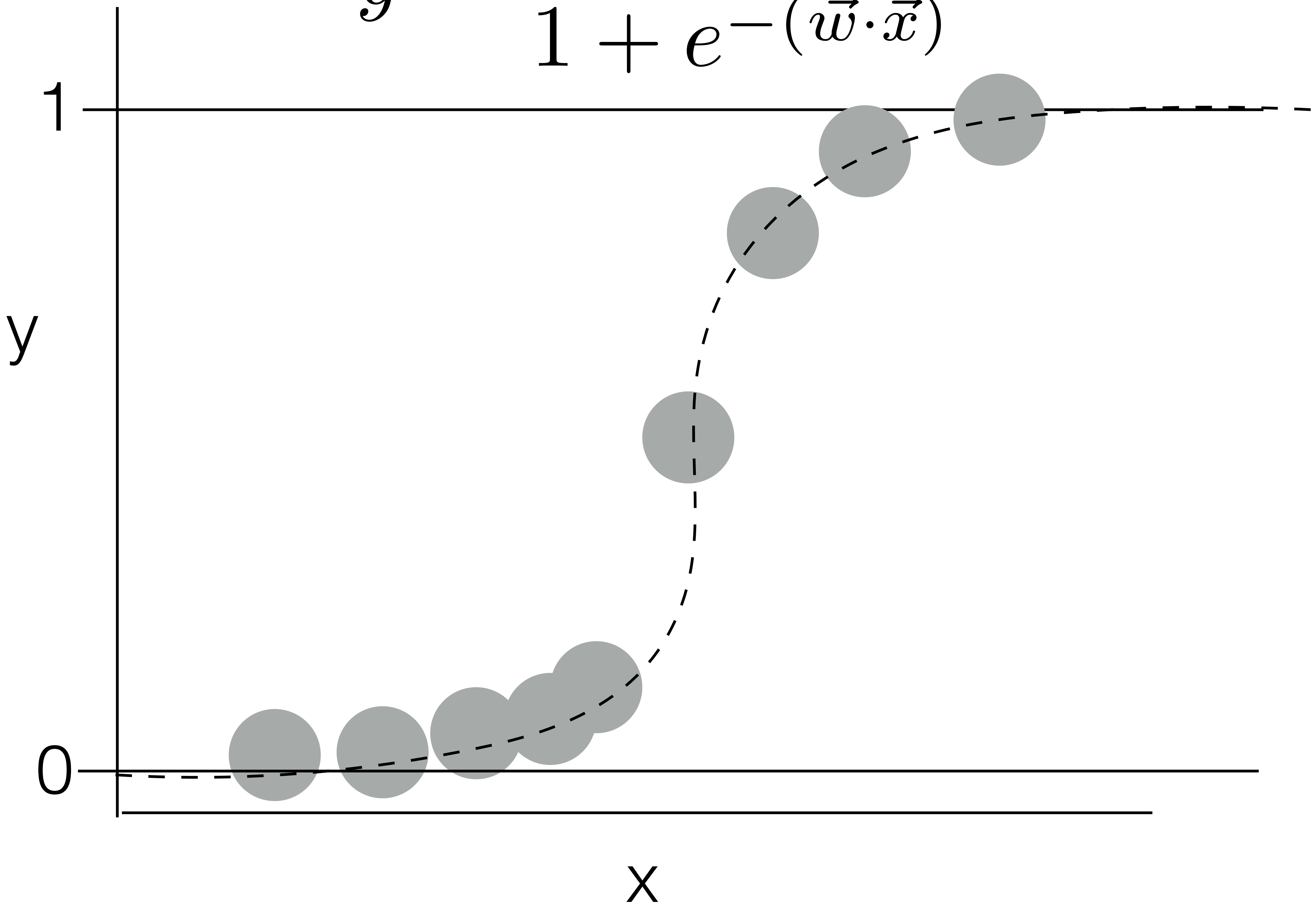
$$y = \vec{w} \cdot \vec{x}$$



Logistic Regression Classifiers

Relation between Linear Regression and Logistic Regression

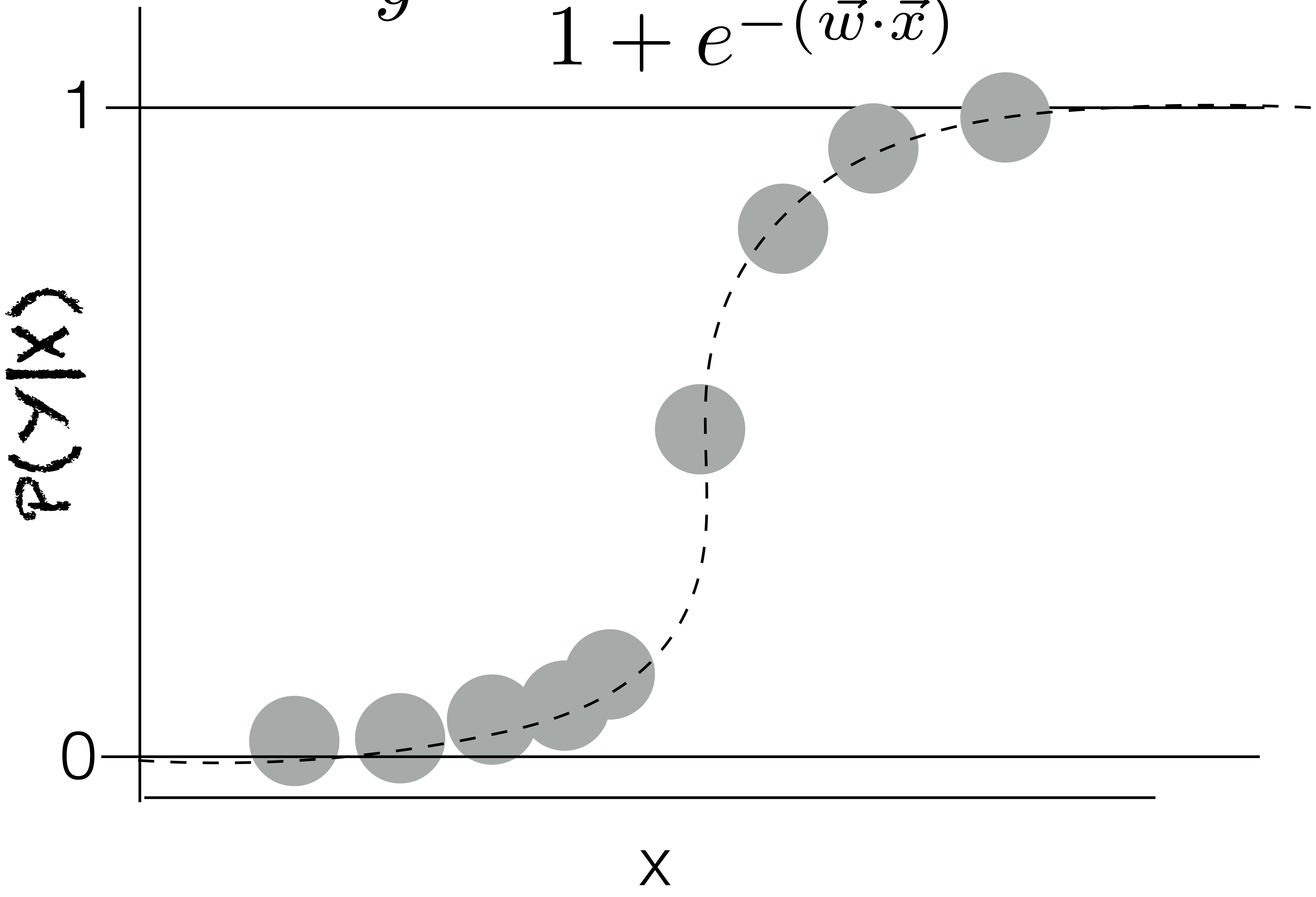
$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



Logistic Regression Classifiers

Relation between Linear Regression and Logistic Regression

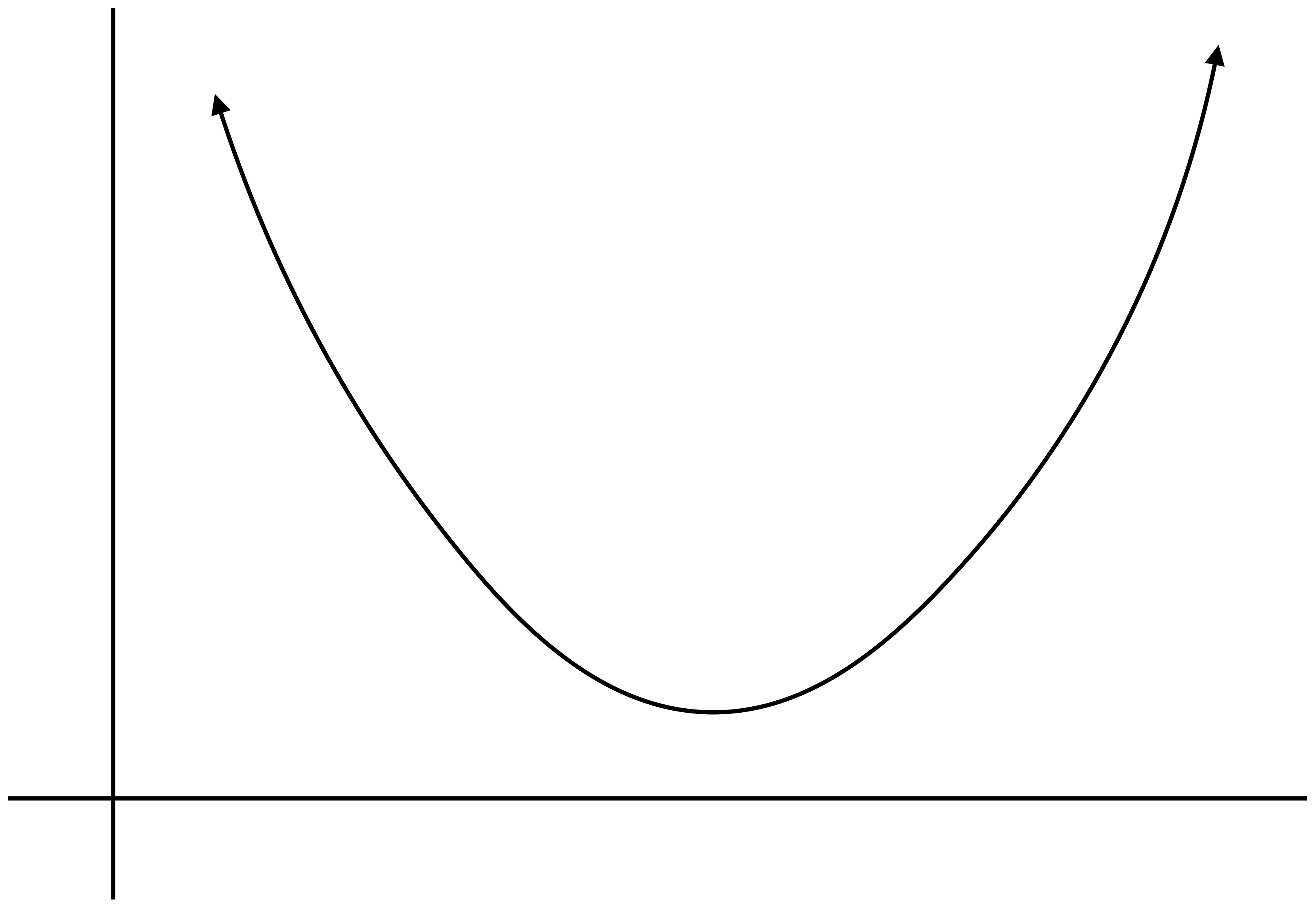
$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



Logistic Regression Classifiers

Training with Gradient Descent

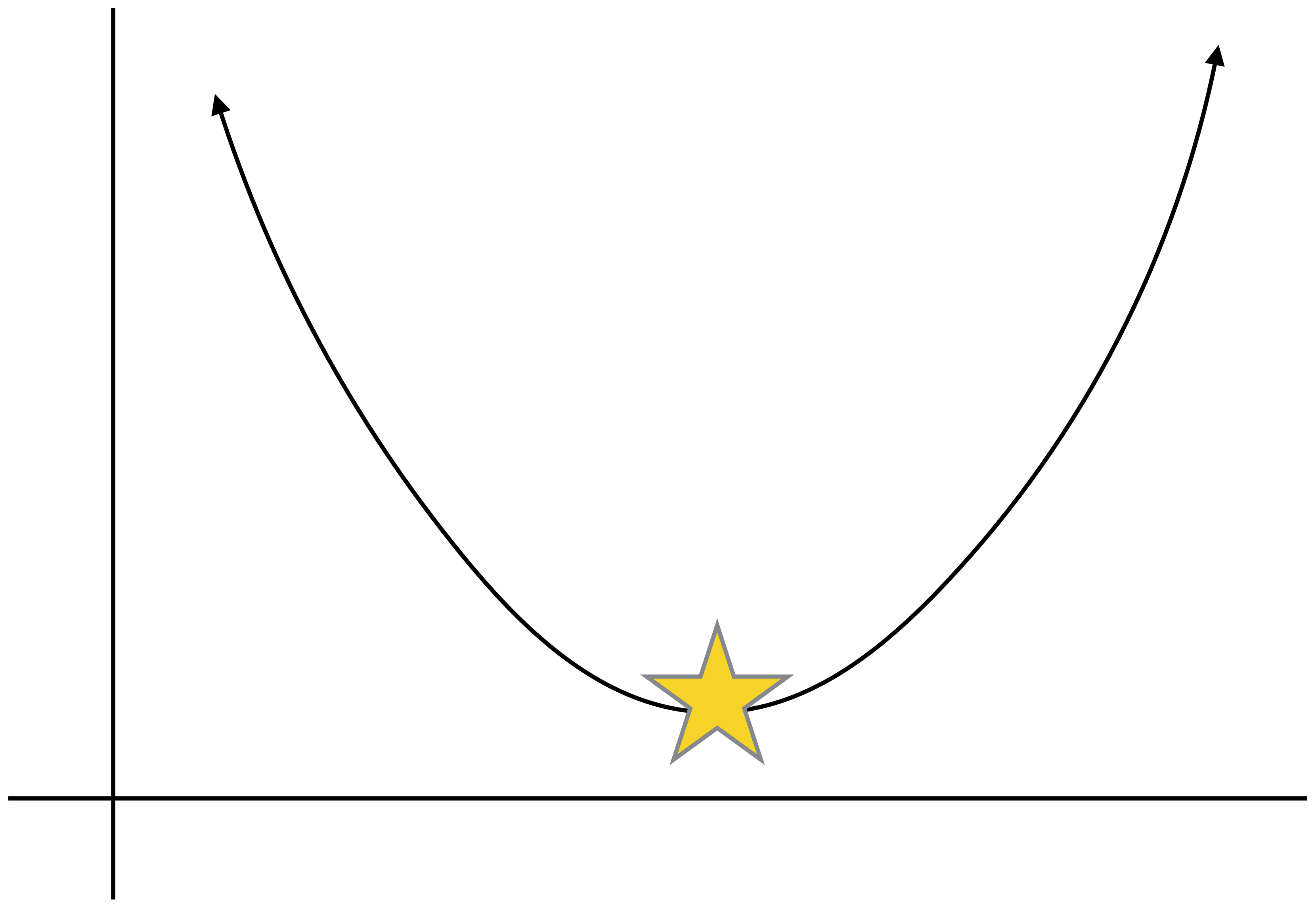
minimize log loss: $-Y log \hat{Y} + (1-Y) log(1-\hat{Y})$



Logistic Regression Classifiers

Training with Gradient Descent

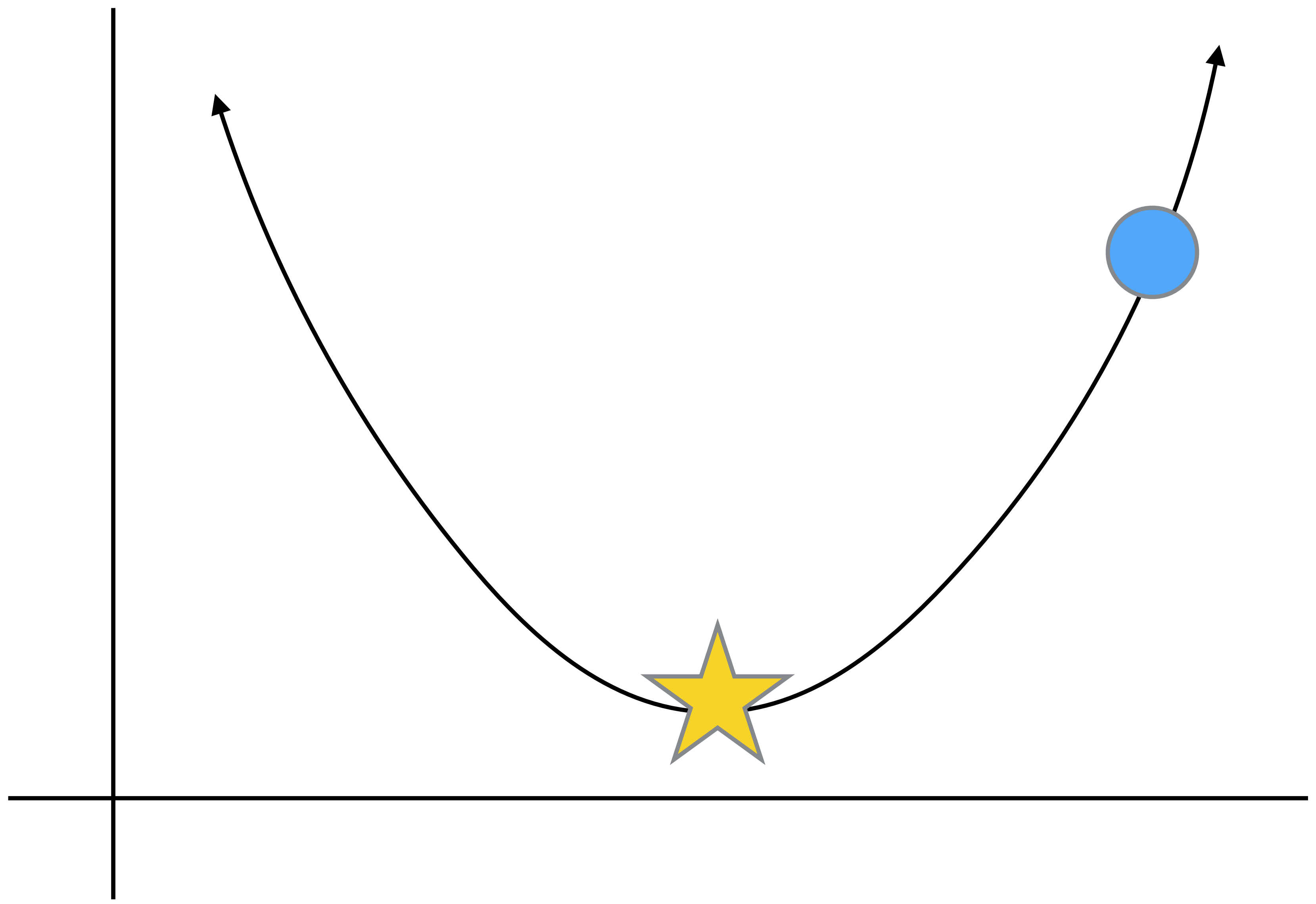
minimize log loss: $-Y log\hat{Y} + (1 - Y)log(1 - \hat{Y})$



Logistic Regression Classifiers

Training with Gradient Descent

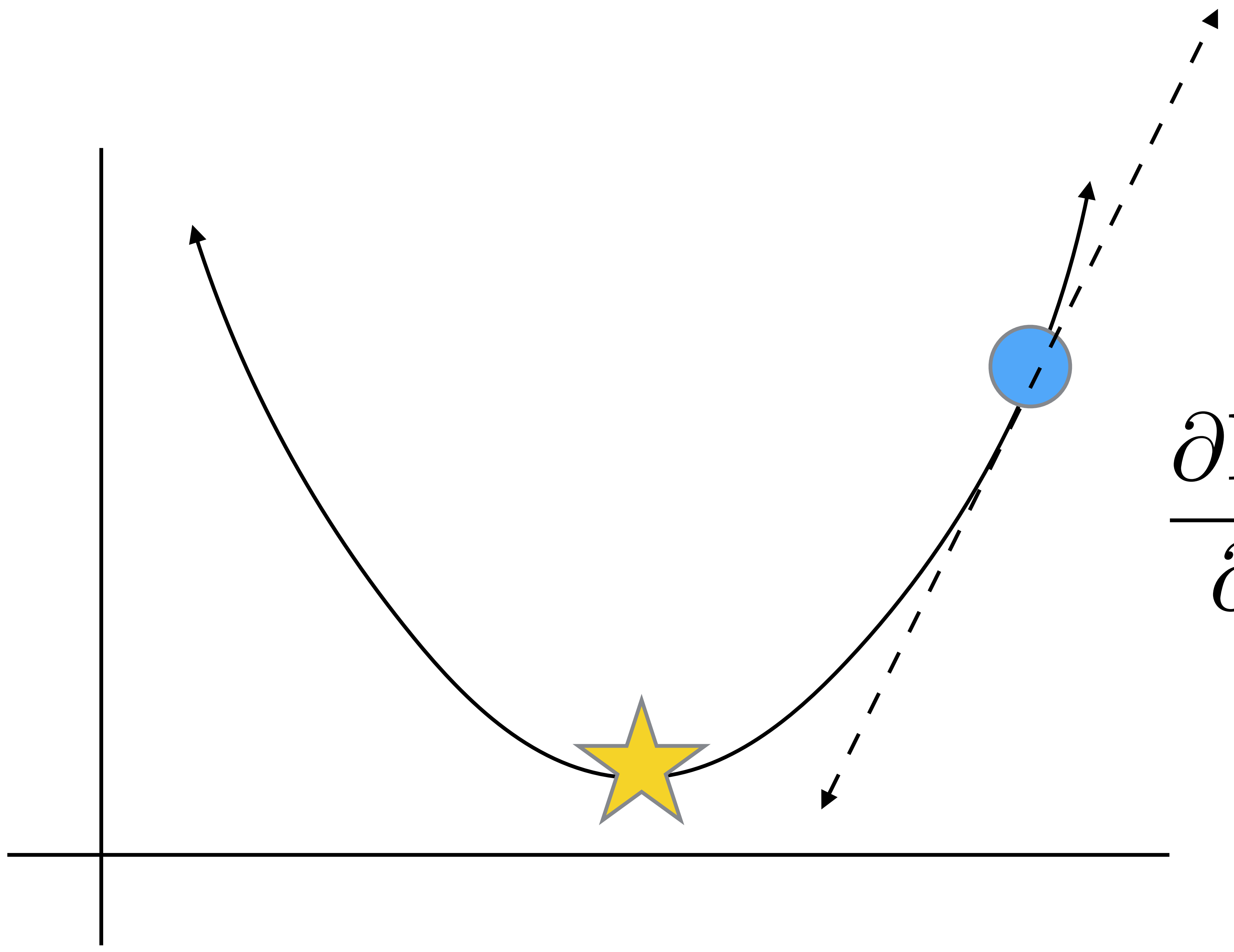
minimize log loss: $-Y log \hat{Y} + (1-Y) log(1-\hat{Y})$



Logistic Regression Classifiers

Training with Gradient Descent

minimize log loss: $-Y log \hat{Y} + (1 - Y) log(1 - \hat{Y})$

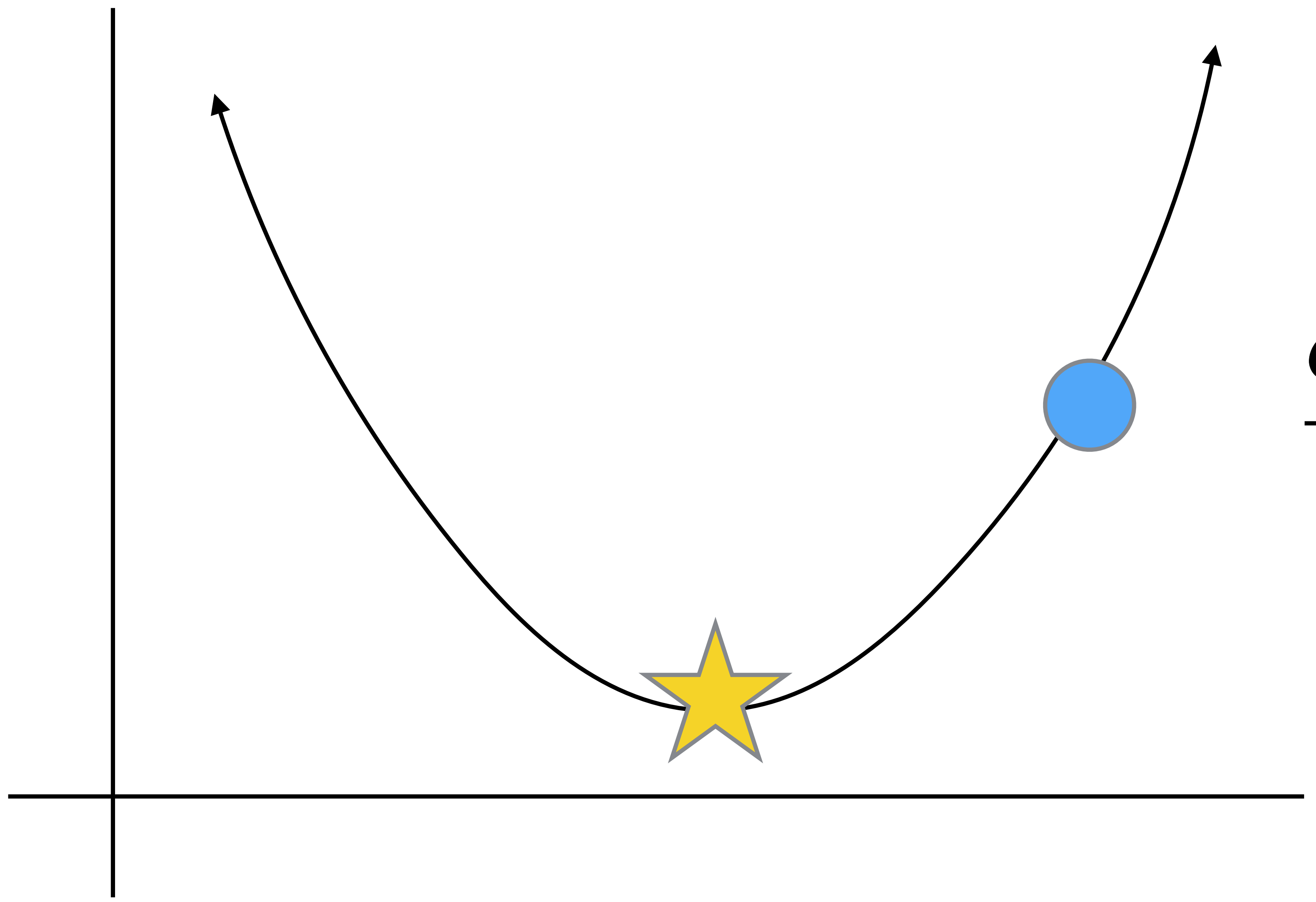


$$\frac{\partial \text{loss}}{\partial w}$$

Logistic Regression Classifiers

Training with Gradient Descent

minimize log loss: $-Y log\hat{Y} + (1 - Y) log(1 - \hat{Y})$



$$\frac{\partial \text{loss}}{\partial w}$$

Logistic Regression Classifiers

Interpreting Weights

Naive Bayes

x	P(x|Y=1)
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

Logistic Regression Classifiers

Interpreting Weights

Logistic Regression

x	???
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

Logistic Regression Classifiers

Interpreting Weights

Logistic Regression

x	???
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

WTF does this weight mean?

(a) There is a 1.0 probability of observing "600" given Y = 1

(b) There is a 1.0 probability that Y = 1 given we observe "600"

(c) 1 is the co-efficient on the "600" variable in the best fit linear regression.

(d) 1 is the co-efficient on the "600" variable in linear regression that minimizes the log loss.

Logistic Regression Classifiers

Interpreting Weights

Logistic Regression

x	???
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

WTF does this weight mean?

(a) There is a 1.0 probability of observing "dramatic" given Y = 1

(b) There is a 1.0 probability that Y = 1 given we observe "dramatic"

(c) 1 is the co-efficient on the "dramatic" variable in the best fit linear regression.

(d) 1 is the co-efficient on the "dramatic" variable in linear regression that minimizes the log loss.

Logistic Regression Classifiers

Inference

Logistic Regression

x	w
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

600 Awesome Tax Policies

???

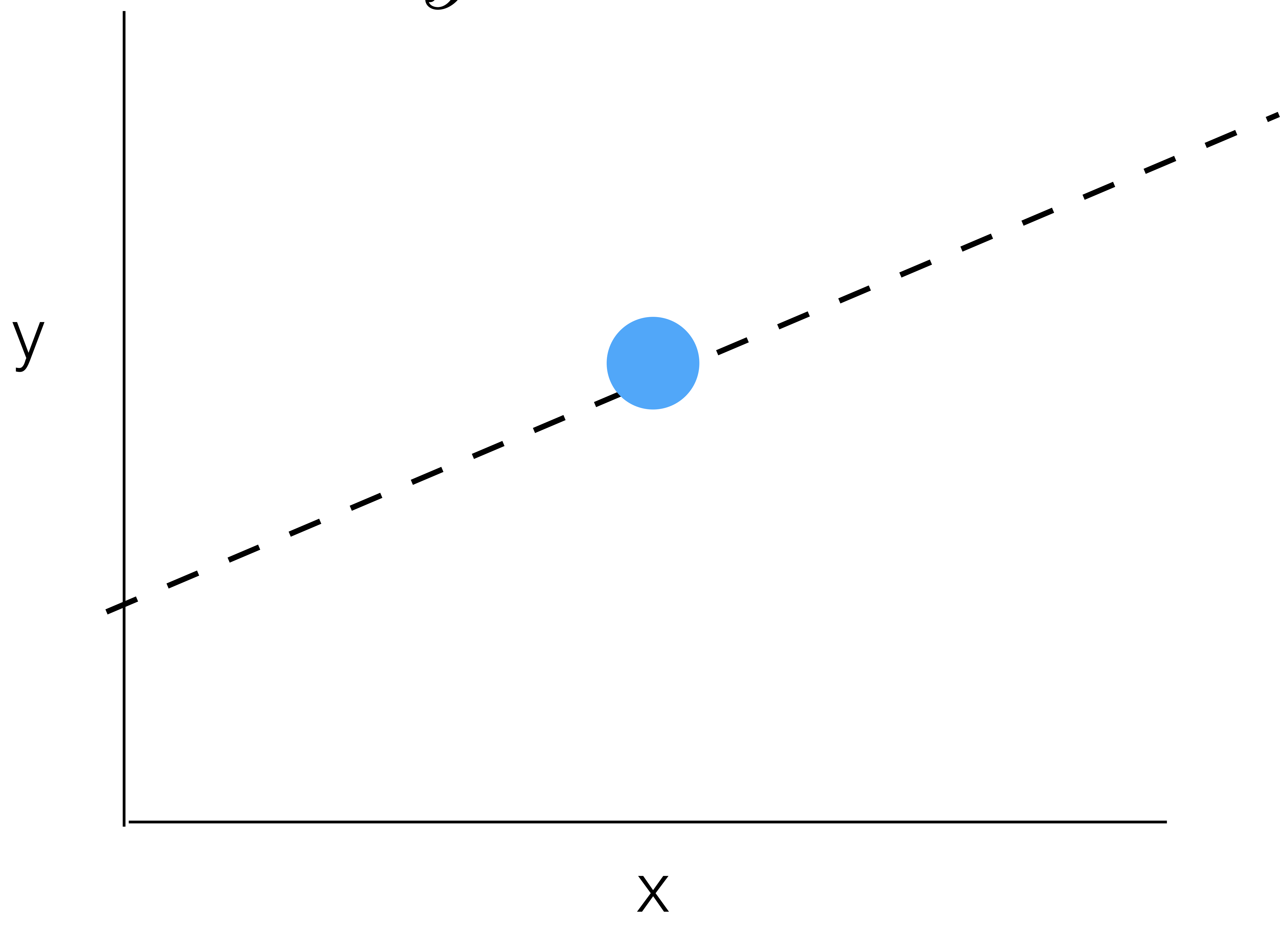
Logistic Regression Classifiers

Inference

x	w
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

600 Awesome Tax Policies

$$y = \vec{w} \cdot \vec{x}$$

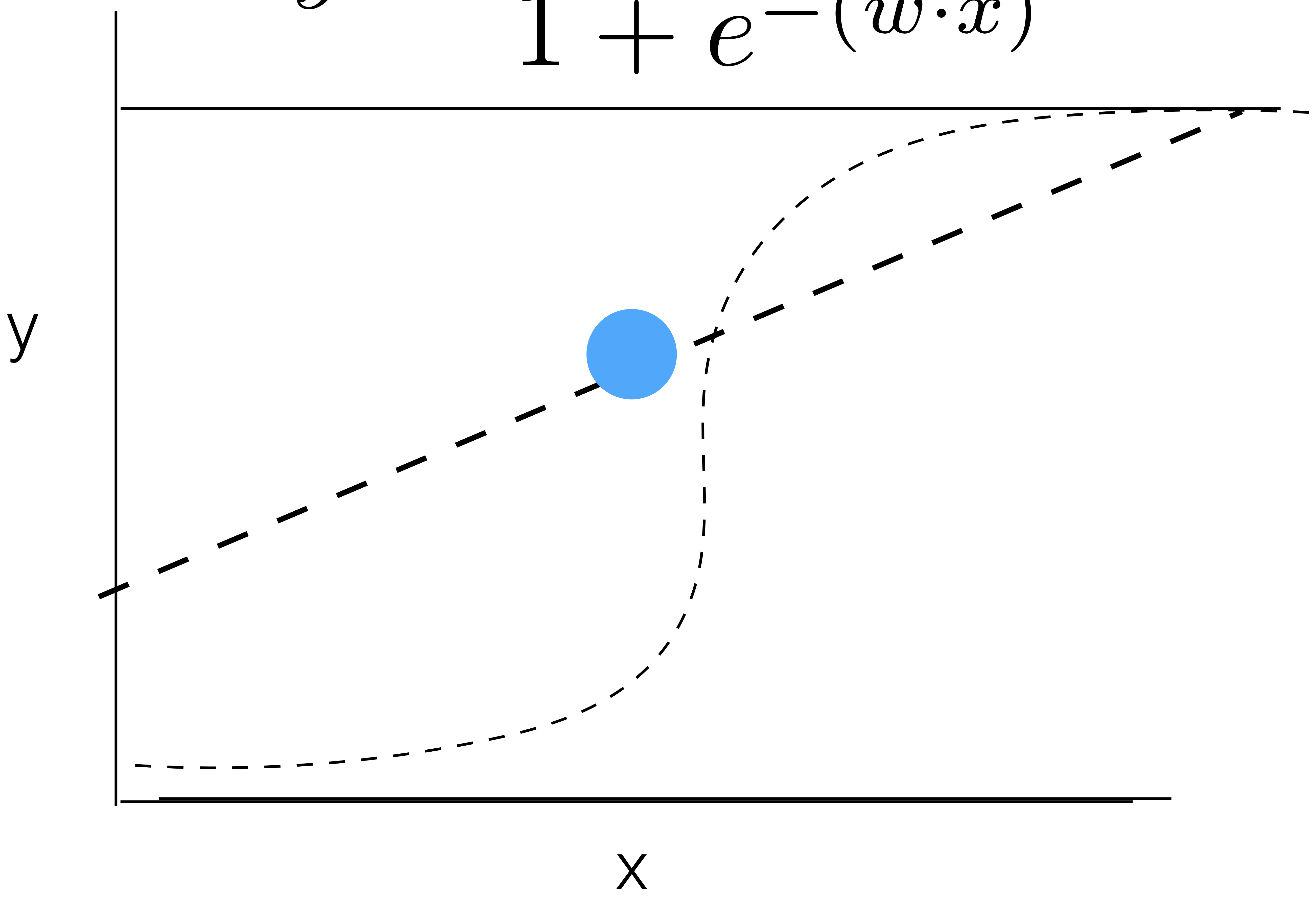


Logistic Regression Classifiers

Inference

x	w
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



600 Awesome Tax Policies

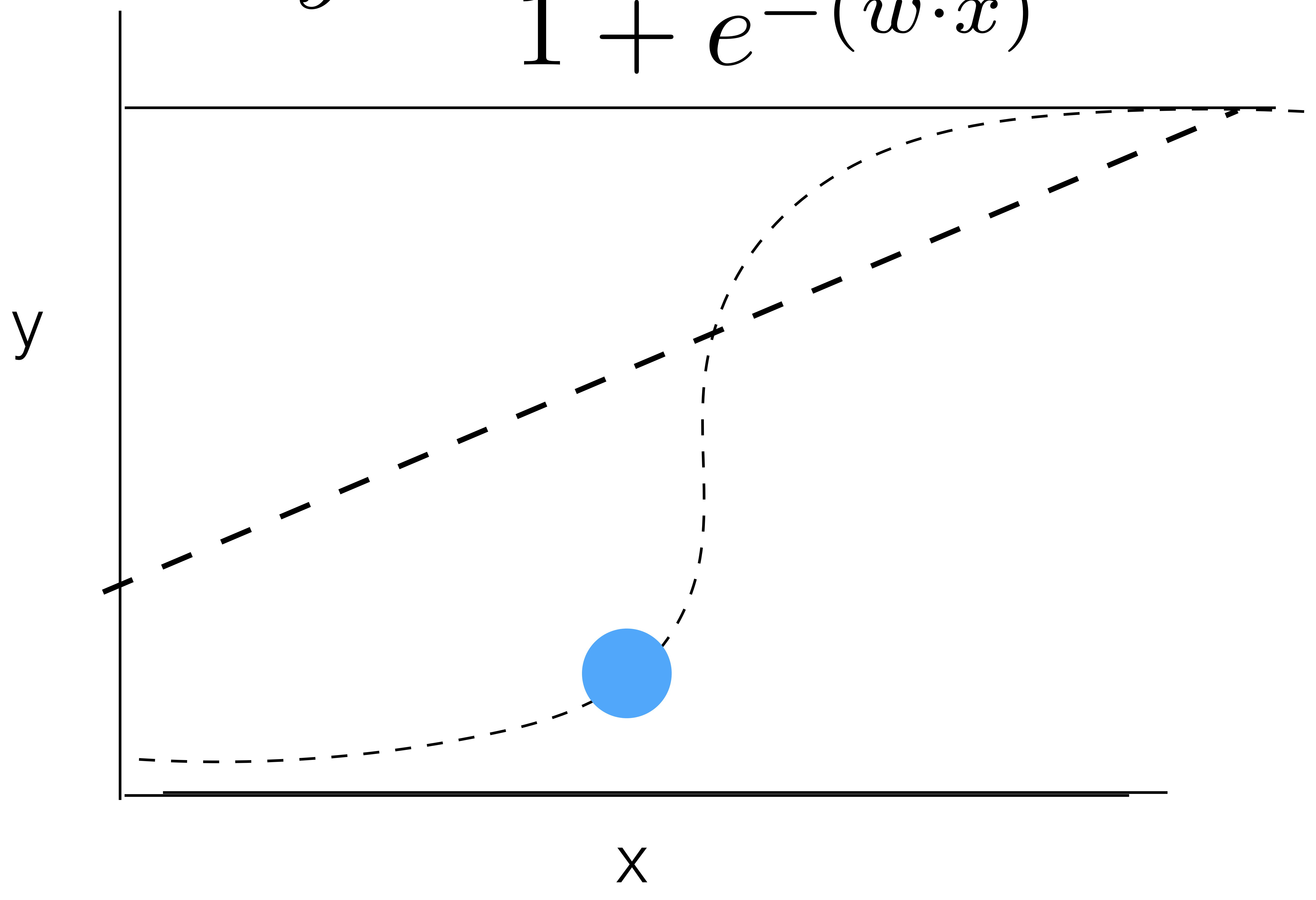
Logistic Regression Classifiers

Inference

x	w
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

600 Awesome Tax Policies

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



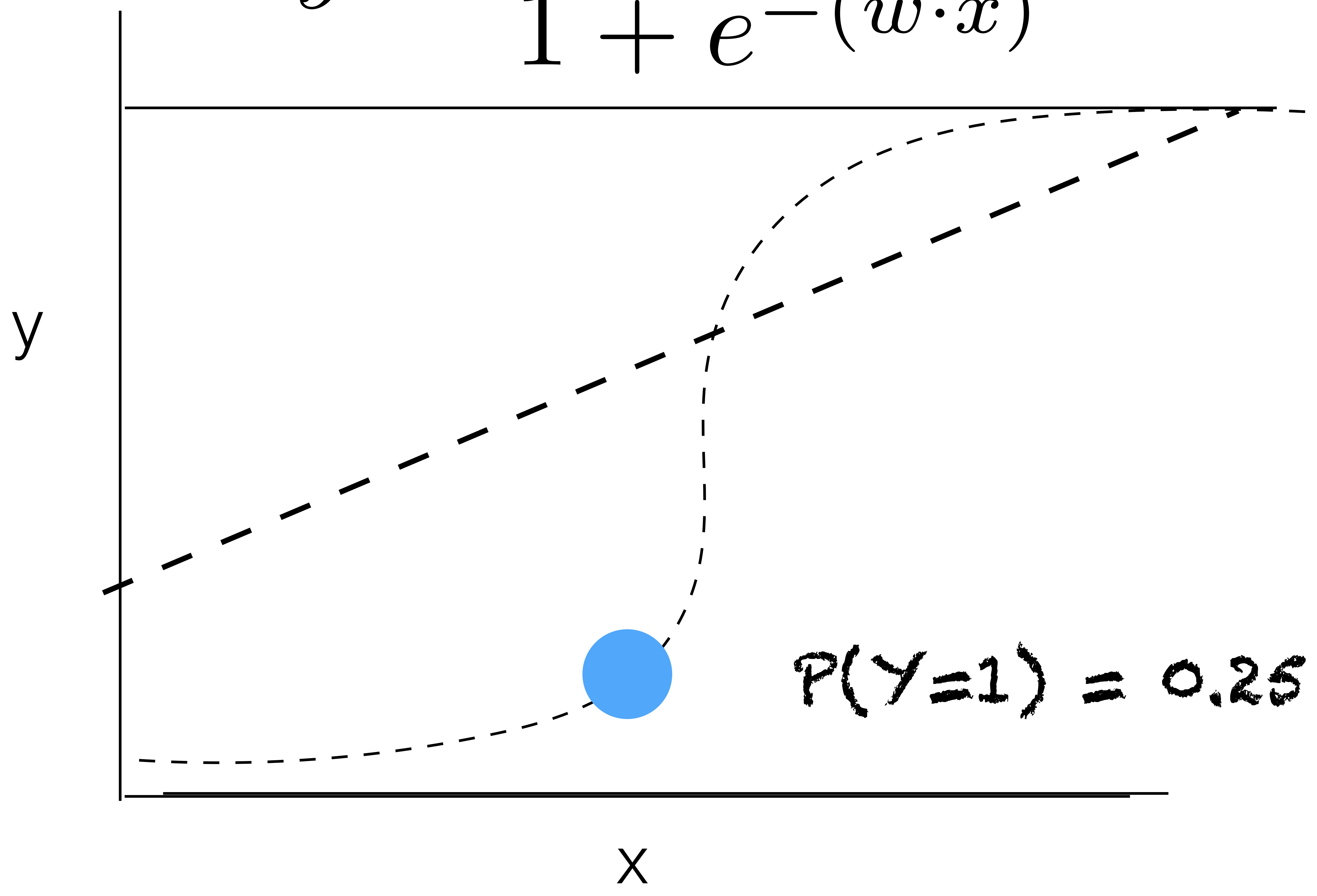
Logistic Regression Classifiers

Inference

x	w
Tax	0.5
This	0.5
600	1.0
👇	1.0
Awesome	0.6
Policies	0.2

600 Awesome Tax Policies

$$y = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$



Topics

- Lecture 1 Reprise, New Quiz Makeup Policy
- Machine Learning 101: Nearest Neighbors Classifier
- Naive Bayes Text Classifier
- Logistic Regression Text Classifier
- **Experimental Design in ML**

Overfitting

- Models **overfit** when then model idiosyncrasies of the training data that don't necessarily generalize

Overfitting

- Models **overfit** when then model idiosyncrasies of the training data that don't necessarily generalize
- E.g., assuming that the presence of the word "This" *definitely* means an article will be clicked

Training Data

Label	Article Title
0	"New Tax Guidelines"
1	"**This** 600lb baby…"
1	"**This**. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets

Train-Test Splits

- What we really care about is performance on new data we haven’t seen before
  - I.e., data the model hasn’t trained on
- We need to simulate this scenario
- We “hold out” some data from training, so model can’t use it to set parameters
- Then we evaluate on the held out data

Train-Test Splits

Training Data

Label	Article Title
0	“Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

x	P(x|Y=1)	P(x|Y=0)
Tax	0.00	0.33
This	0.67	0.00
600	0.33	0.00
👇	0.33	0.00
Guidelines	0.0	0.33

Train-Test Splits

Training Data

Label	Article Title
0	“Tax Guidelines”
1	“This 600lb baby…”
1	“This. 👇👇👇”
1	“18 healthy foods that are actually killing you from within. 🥗☠️”
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

x	P(x|Y=1)	P(x|Y=0)
Tax	0.00	0.33
This	0.67	0.00
600	0.33	0.00
👇	0.33	0.00
Guidelines	0.0	0.33

Tax Guidelines ???

Train-Test Splits

Training Data

Label	Article Title
Hidden!	
1	“This 600lb baby…”
1	“This. 👇👇👇”
Hidden!	
0	“The Brothers Karamazov: a neo-post-globalist perspective”
0	4 things they don’t want you to know about dryer sheets

x	P(x|Y=1)	P(x|Y=0)
~~Tax~~	~~0.00~~	~~0.33~~
This	1.00	0.00
600	0.50	0.00
👇	0.50	0.00
~~Guidelines~~	~~0.0~~	~~0.33~~

Train-Test Splits

Training Data

Label	Article Title
0	"Tax Guidelines"
1	"**This** 600lb baby…"
1	"**This**. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets

x	P(x|Y=1)	P(x|Y=0)
~~Tax~~	~~0.00~~	~~0.33~~
This	1.00	0.00
600	0.50	0.00
👇	0.50	0.00
~~Guidelines~~	~~0.0~~	~~0.33~~

Tax Guidelines ???

Train-Test Splits

Training Data

Label	Article Title
0	"Tax Guidelines"
1	"**This** 600lb baby…"
1	"**This**. 👇👇👇"
1	"18 healthy foods that are actually killing you from within. 🥗☠️"
0	"The Brothers Karamazov: a neo-post-globalist perspective"
0	4 things they don't want you to know about dryer sheets

(more on dealing with low counts next lecture!)

x	P(x|Y=1)	P(x|Y=0)
~~Tax~~	~~0.00~~	~~0.33~~
This	1.00	0.00
600	0.50	0.00
👇	0.50	0.00
~~Guidelines~~	~~0.0~~	~~0.33~~

Tax Guidelines ???

Train-Test Splits
Train/Dev/Test Splits

- In real ML settings, we typically split into three sets
  - Train: Used to train the model, i.e., set the parameters
  - Dev: Used during development, to inspect and choose "hyperparameters"
    - E.g., comparing whether to use NB or LR
  - Test: To test the model. In good experimental design, only allowed to evaluate on test one time, to avoid "cheating"

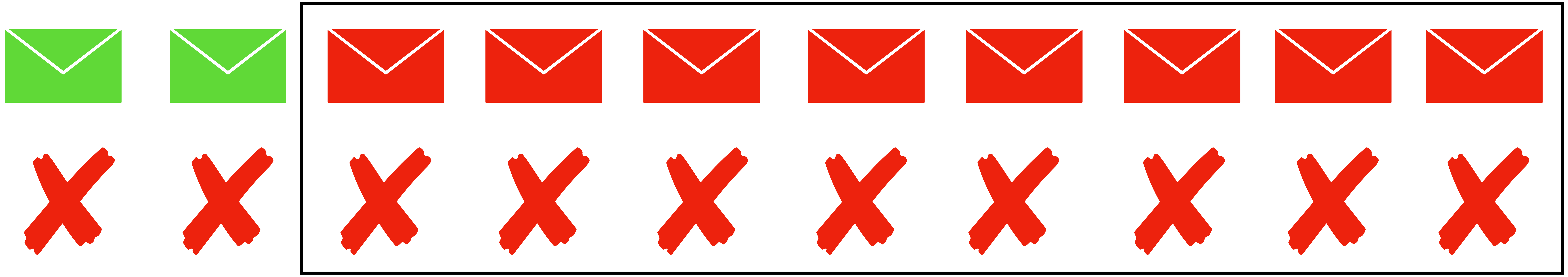
Train-Test Splits

K-Fold Cross Validation

```
for i in range(0, K):
  train_data, test_data = split(data)
  model = model.train(train_data)
  score = model.evaluate(test_data)
```

Baselines

- Baseline: A simpler model/way of solving the problem
  - Used to put results in context
- E.g., 80% sounds pretty good, but if “always predict spam” gets 80% accuracy, then an ML model which gets 80% is not very impressive…



Baselines

- Common baselines to report:
  - Random: Guess at random
  - “Most frequent class”: For classification tasks, always predict whichever label is most common in the training set
  - Prior state-of-the-art (SOTA): i.e., the “defending champ”
  - Various task-specific heuristics, e.g.,
    - For QA: pick the first name in the passage
    - For IR: sort documents according to length
    - Usually requires some creativity

Baselines

- “Skylines”: upper bounds on performance
- Common skylines:
  - Human performance on the task
  - Performance under ideal conditions (e.g., how good would my QA system be if we tell it which sentence to look at…)

Train Test Splits

Beyond IID

- i.i.d.: train and test data are drawn from the same distribution
    - I.e., take a dataset, randomly shuffle it, and split it into 80% train/20% test
    - This is the most standard setting, a “traditional ML” setting
- In real applications, test isn’t always i.i.d., i.e.,
    - You want to build a model of customers that generalizes to new markets (train in China, test in US)
    - You want to forecast disease spread that generalizes to the future (train in 2010—2019, test in 2020—2021)
    - You want to screen applicants for an internship, based on data on success of past interns (train data is from 1980—2000 when company was primarily white upper middle class, new applicant pool is more racially and socio-economically diverse)

Train Test Splits

Practice Question!

- Context: Social media director for a PR company
- Data: Instagram posts for 5000 new musicians, plus subsequent likes/reposts/comments and sales records.
- Goal: Predict the popularity of a post so we can optimize visibility of new clients.

How should I define my train/test splits here?

a) i.i.d., i.e., randomly split posts into train/test
b) hold out posts from the most recent year
c) hold out posts from 10% of artists
d) hold out least popular 10% of posts

Train Test Splits

Practice Question!

- Context: Social media director for a PR company
- Data: Instagram posts for 5000 new musicians, plus subsequent likes/reposts/comments and sales records.
- Goal: Predict the popularity of a post so we can **optimize visibility of new clients**.

How should I define my train/test splits here?

a) i.i.d., i.e., randomly split posts into train/test
b) hold out posts from the most recent year
c) hold out posts from 10% of artists
d) hold out least popular 10% of posts

Train Test Splits

Practice Question!

- Context: Hedge fund manager
- Data: Social media chatter about companies plus daily stock prices for those companies over past .
- Goal: Detect when there is going to be another “GameStop situation”…i.e., sudden spike in a stoke’s price

How should I define my train/test splits here?

a) i.i.d.: randomly split daily returns into train/test
b) hold out data from the most recent N years
c) hold out data from 10% of companies
d) hold out data from companies that experienced spikes

Train Test Splits

Practice Question!

- Context: Hedge fund manager
- Data: Social media chatter about companies plus daily stock prices for those companies over past .
- Goal: **Detect when** there is going to be another “GameStop situation”…i.e., sudden spike in a stoke’s price

How should I define my train/test splits here?

a) i.i.d.: randomly split daily returns into train/test
b) hold out data from the most recent N years
c) hold out data from 10% of companies
d) hold out data from companies that experienced spikes