Machine Learning: What is it?

Anastassia Kornilova

Machine Learning is Everywhere

What is Machine Learning?

"training a computer to get new *insights* from data *by itself*"
- Anastassia 2018

Insight?

- Finding similar examples
- Predicting an outcome
- Making a decision
- Giving an answer

By itself?

You did not program an answer/output for this input

Simple Example

Data:
Positive + Negative
Amazon Reviews

↓

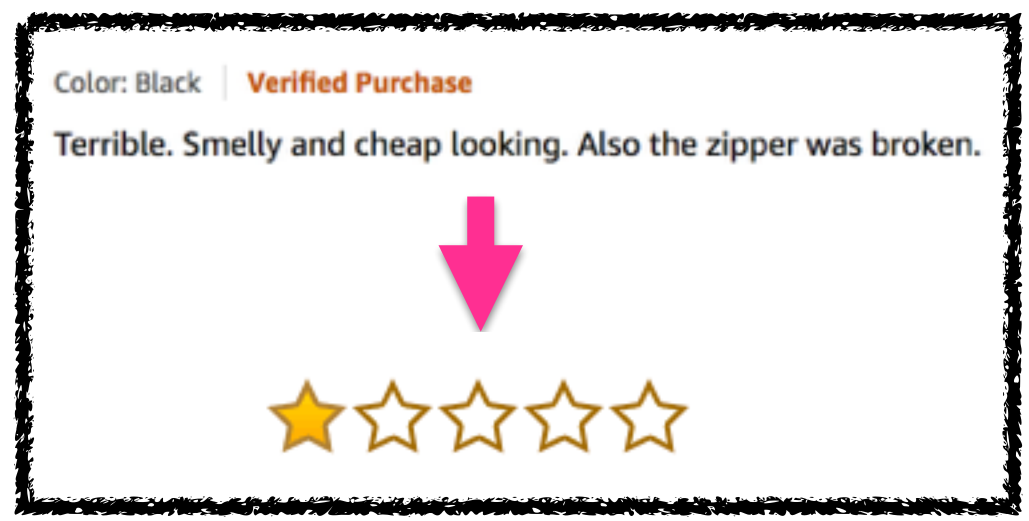
ML Algorithm

↓

Predict "stars" on a new review

Not in here

Five Stars
By Dizzle on April 3, 2018
Color: Blue | Verified Purchase
A little smaller than 24 inches but so cute and made very well for the price



Some terms

- **AI:** computers solving problems that traditionally require human intelligence; superset of ML
- **ML:** algorithms that can be applied to *NLP* and other areas
- **Natural Language Processing (NLP):** problems that *could be solved with ML*

Let’s build a model!!

SCENARIO: You are a movie director for “AI the Movie” and you want to find out if people liked your movie from Twitter. There are no or …how can you tell if people liked it ?

IDEA: Let’s look at a subset of reviews and come up with some rules for deciding if a review is good or bad

If ______ , then Good/Bad

"Amazing all around"	If **amazing** —> then **good**
"Worst movie of the year"	If **worst** —> then **bad**
"You will love the acting, I promise"	If **love** —> then **good**
"Way too slow"	If **slow** —> then **bad**
"The space battles were cool"	If **cool** —> then **good**
"Total waste of time"	If **waste of time** —> then **bad**

Let’s classify some new tweets

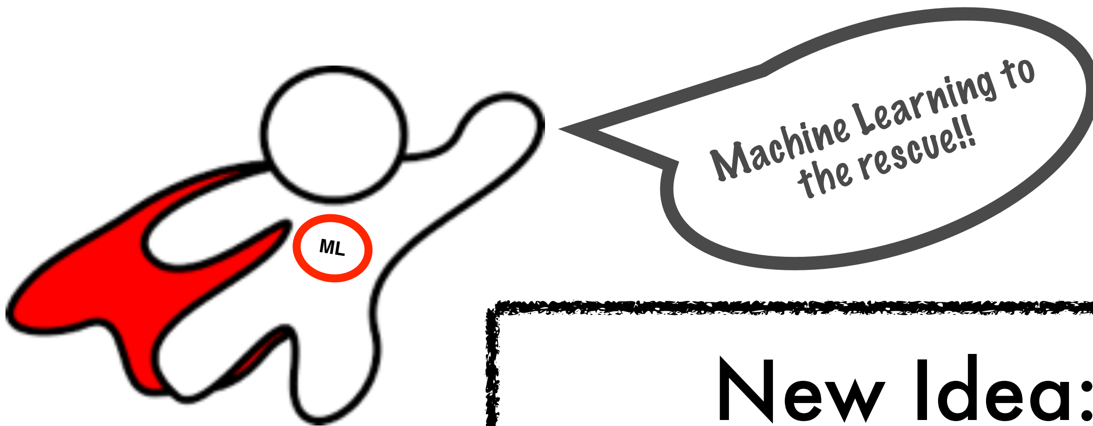
If **amazing** —> then **good**
If **worst** —> then **bad**
If **love** —> then **good**
If **cool** —> then **good**
If **waste of time** —> then **bad**
If **slow** —> then **bad**

Let's classify some new tweets

If **amazing** —> then **good**
If **worst** —> then **bad**
If **love** —> then **good**
If **cool** —> then **good**
If **waste of time** —> then **bad**
If **slow** —> then **bad**

Let’s classify some new tweets

If **amazing** —> then **good**
If **worst** —> then **bad**
If **love** —> then **good**
If **cool** —> then **good**
If **waste of time** —> then **bad**
If **slow** —> then **bad**



New Idea: If we give a computer a lot of examples, maybe it can learn the rules on its own.

Step 1: Collect the data

Rotten Tomatoes

81%	Maniac
97%	American Vandal
100%	BoJack Horseman
96%	The Sinner
58%	Manifest
53%	Marvel's Iron Fist
100%	The Deuce

This is the accidental genius of Crazy Rich Asians; it shows how wealth, in the end, makes fools of us all.

October 11, 2018 | Full Review...

The best studio romantic comedy of the year

October 10, 2018 | Rating: A- | Full Review...

Step 2: Count how many times each word appears in a good or bad review

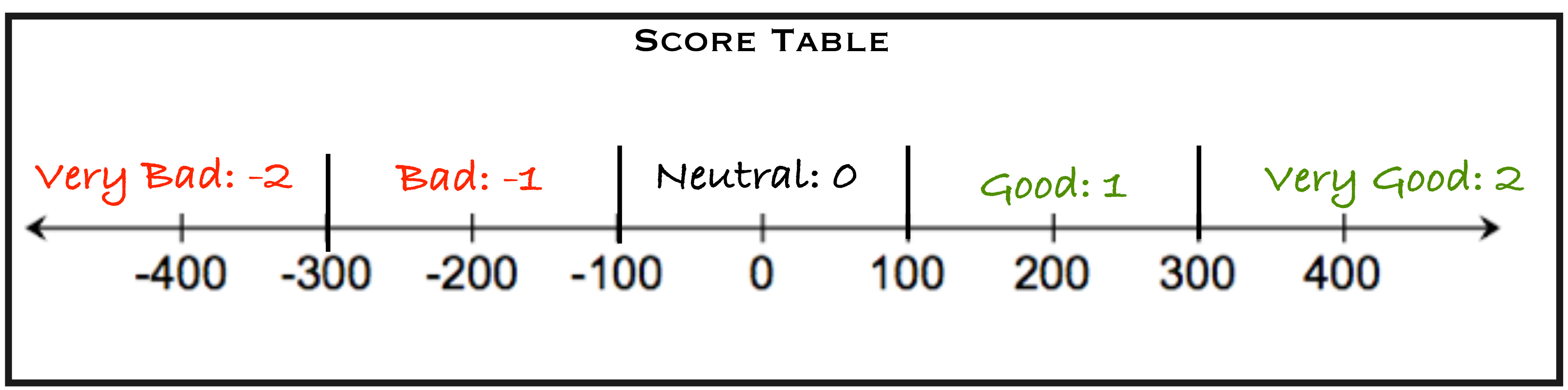
Words	Count in GOOD reviews	Count in BAD reviews		
good	300	50		
boring	20	350		
acting	800	790		
:D	540	10		
:((	20	200		
amazing	500	25		
plot	650	670		
slow	20	140		

Step 3: Count how many times each word appears in a good or bad review

Words	Count in GOOD reviews	Count in BAD reviews	Count (GOOD - BAD)	
good	300	50	300 - 50 = 250	
boring	20	350	20 - 350 = -330	
acting	800	790	800 - 790 = 20	
:D	540	10	540 - 10 = 530	
:((	20	200	20 - 200 = -140	
amazing	400	25	400 - 25 = 375	
plot	650	670	650 - 670 = -20	
slow	20	140	20 - 140 = -120	

Step 4: Assign score from table

Words	Count in GOOD reviews	Count in BAD reviews	Difference (GOOD - BAD)	Score
good	300	50	300 - 50 = 250	1
boring	20	350	20 - 350 = -330	-2
acting	800	790	800 - 790 = 20	0
:D	540	10	540 - 10 = 530	2
:((	20	200	20 - 200 = -140	-1
amazing	400	25	400 - 25 = 375	2
plot	650	670	650 - 670 = -20	0
slow	20	140	20 - 140 = -120	-1



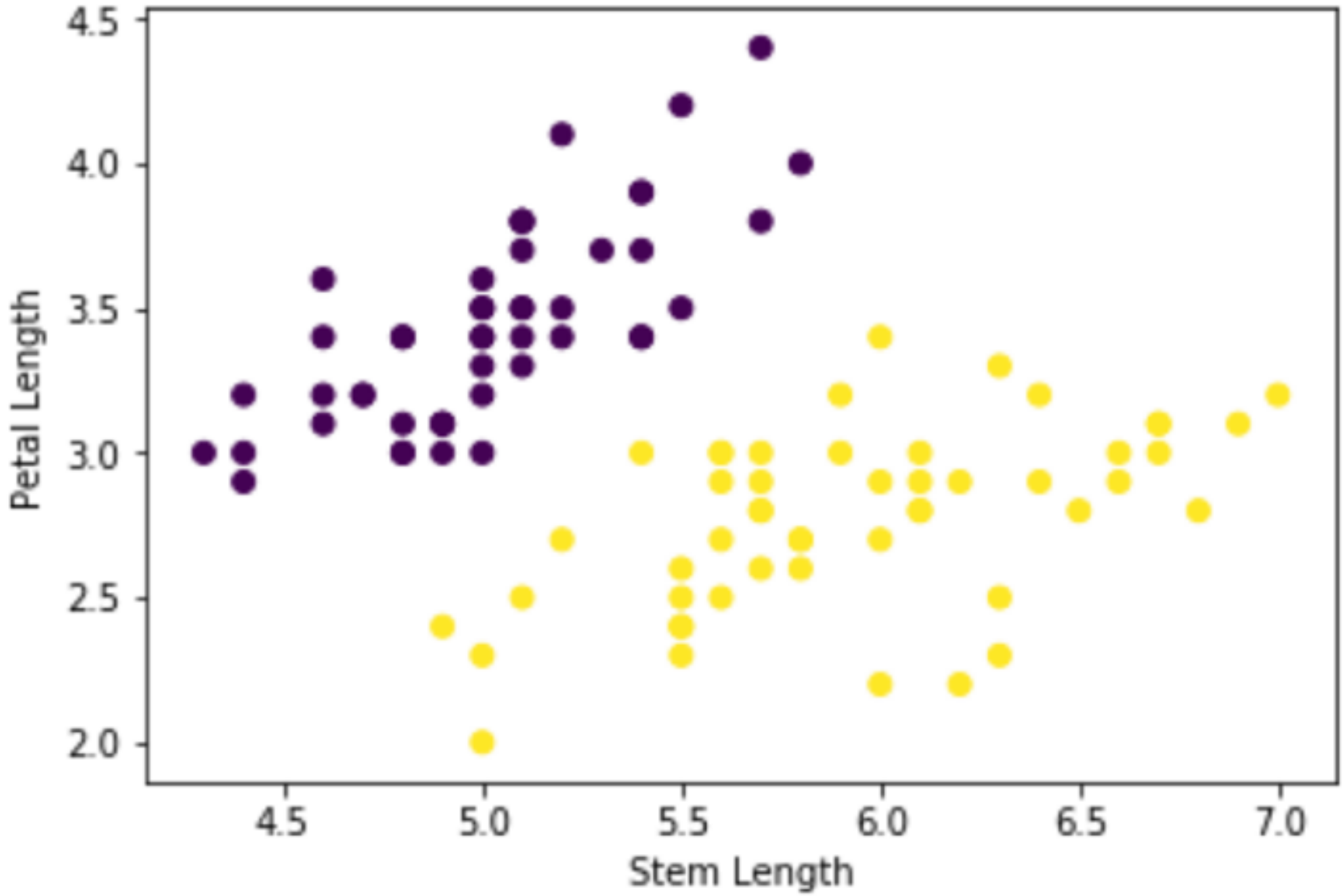
Words	Score
good	1
boring	-2
acting	0
:D	2
:((	-1
amazing	2
plot	0
slow	-1

Word	Weight
plot	0
boring	-2
Sum:	-2

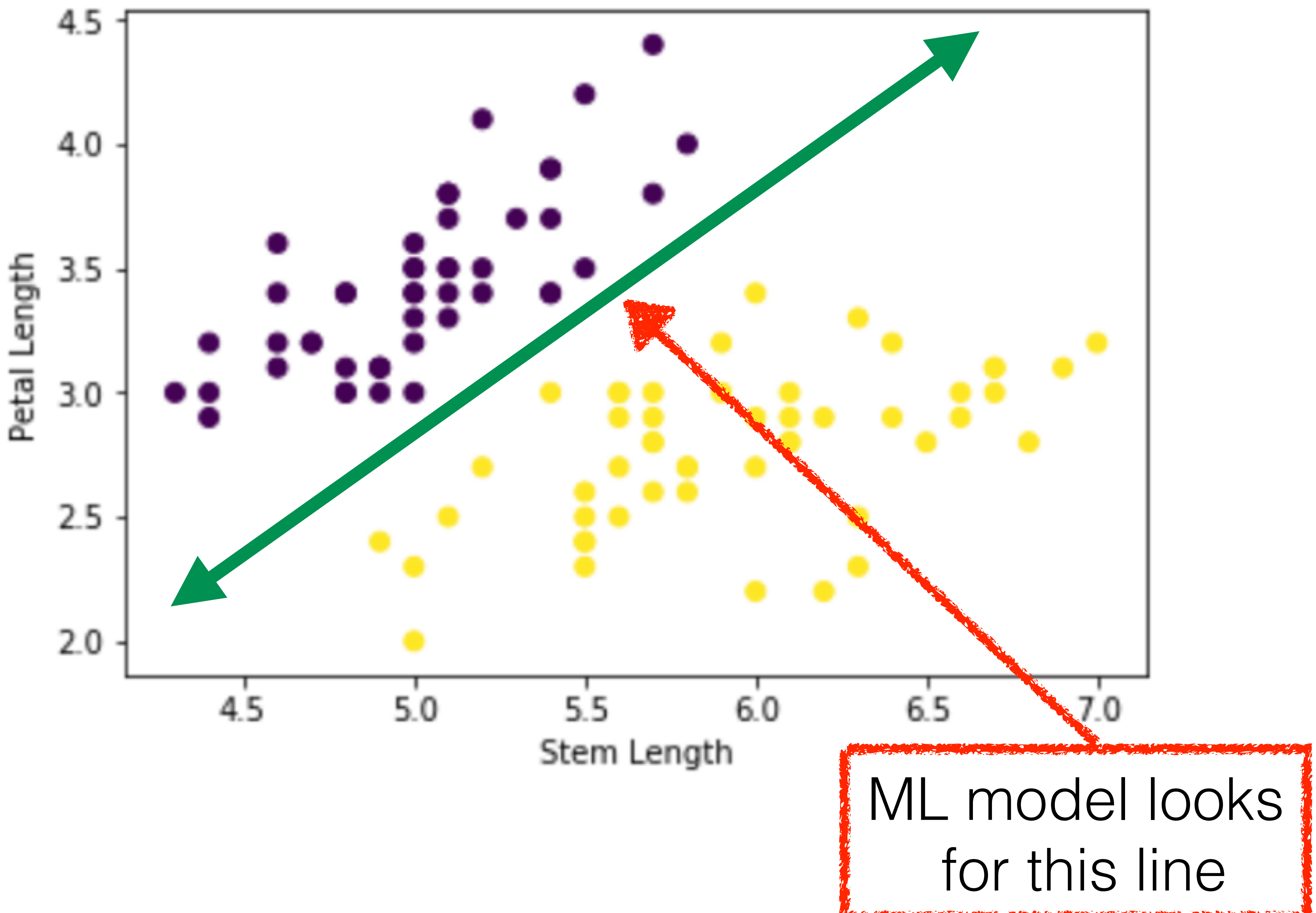
Words	Score
good	1
boring	-2
acting	0
:D	2
:((	-1
amazing	2
plot	0
slow	-1

Word	Weight
plot	0
slow	-1
amazing	2
acting	0
Sum:	1

The Flower DataSet



The Flower DataSet



A mathematical view

Model

$$y = f(x)$$

Output/Target

Input

A mathematical view

Review Sentiment = f(words)

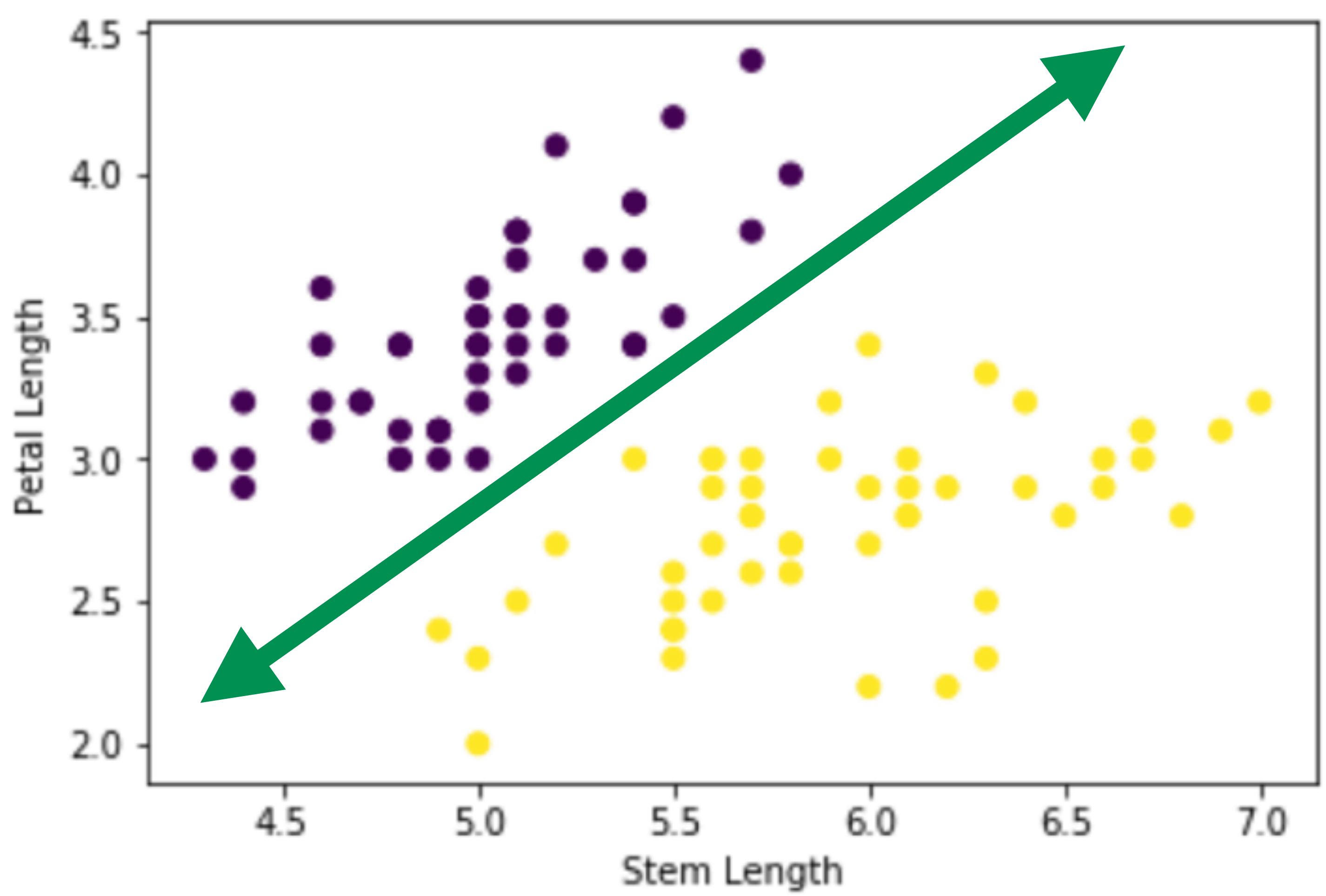
Flower Type = f(Stem Length, Petal Length)

A mathematical view

Review Sentiment = f(words)

$$\text{Review Sentiment} = \sum \text{word} \cdot w_{word}$$

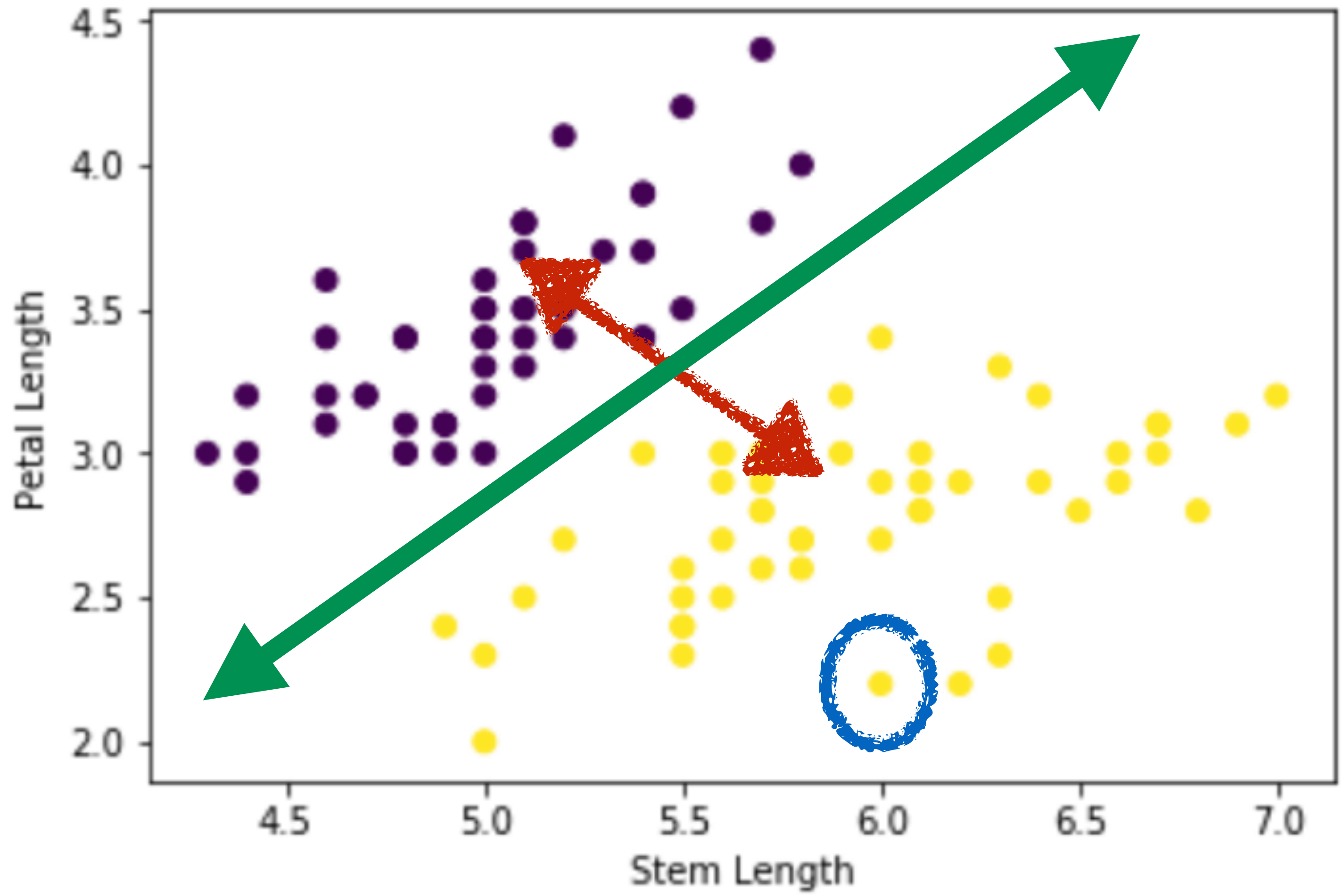
A mathematical view



Petal Length = 0.8 * Stem Length - 1.3

Target???

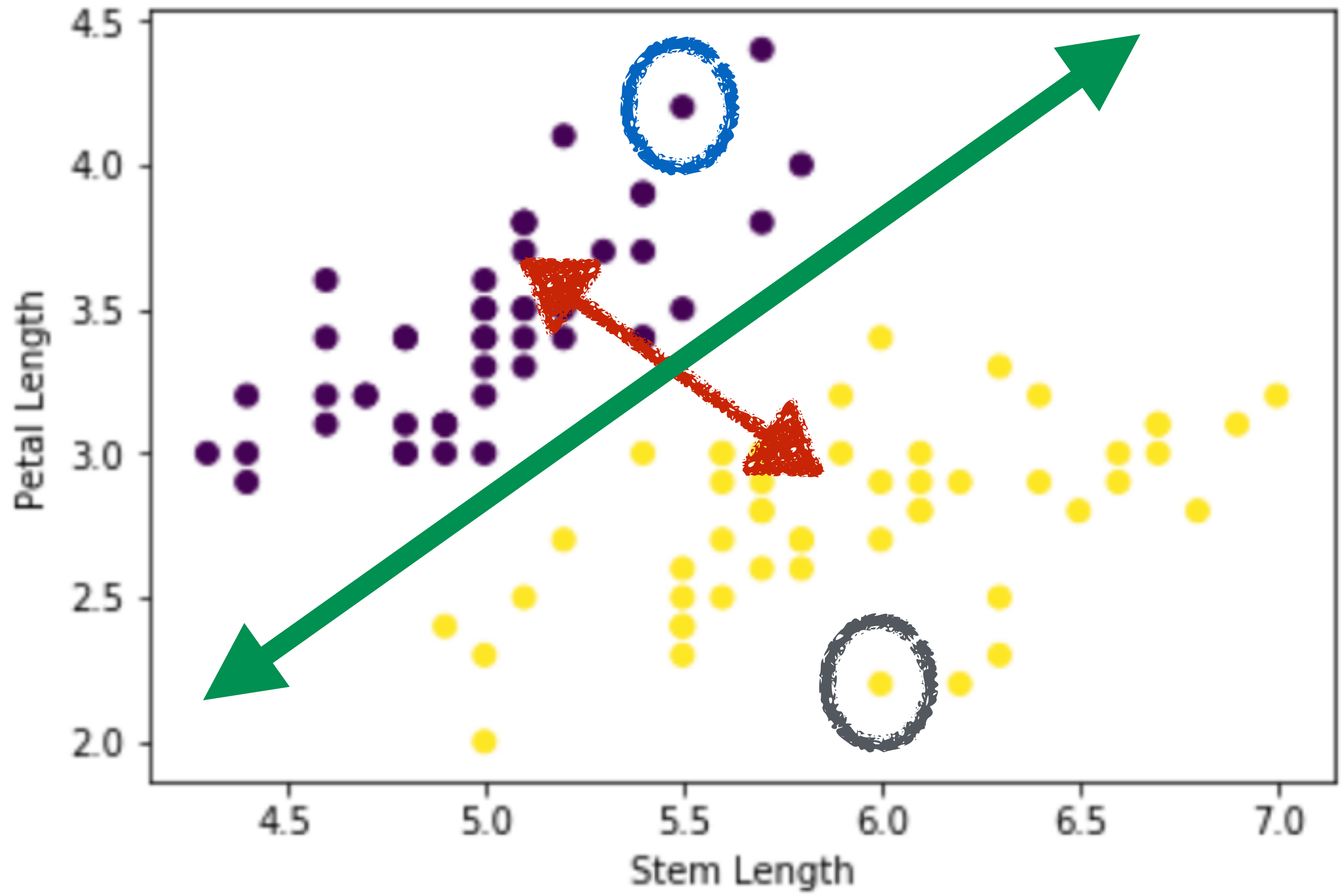
A mathematical view



0.8 * SL - 1.3 - PL

$\boxed{0.8 * 6 - 1.3 - 2.1 = 1.4}$

A mathematical view

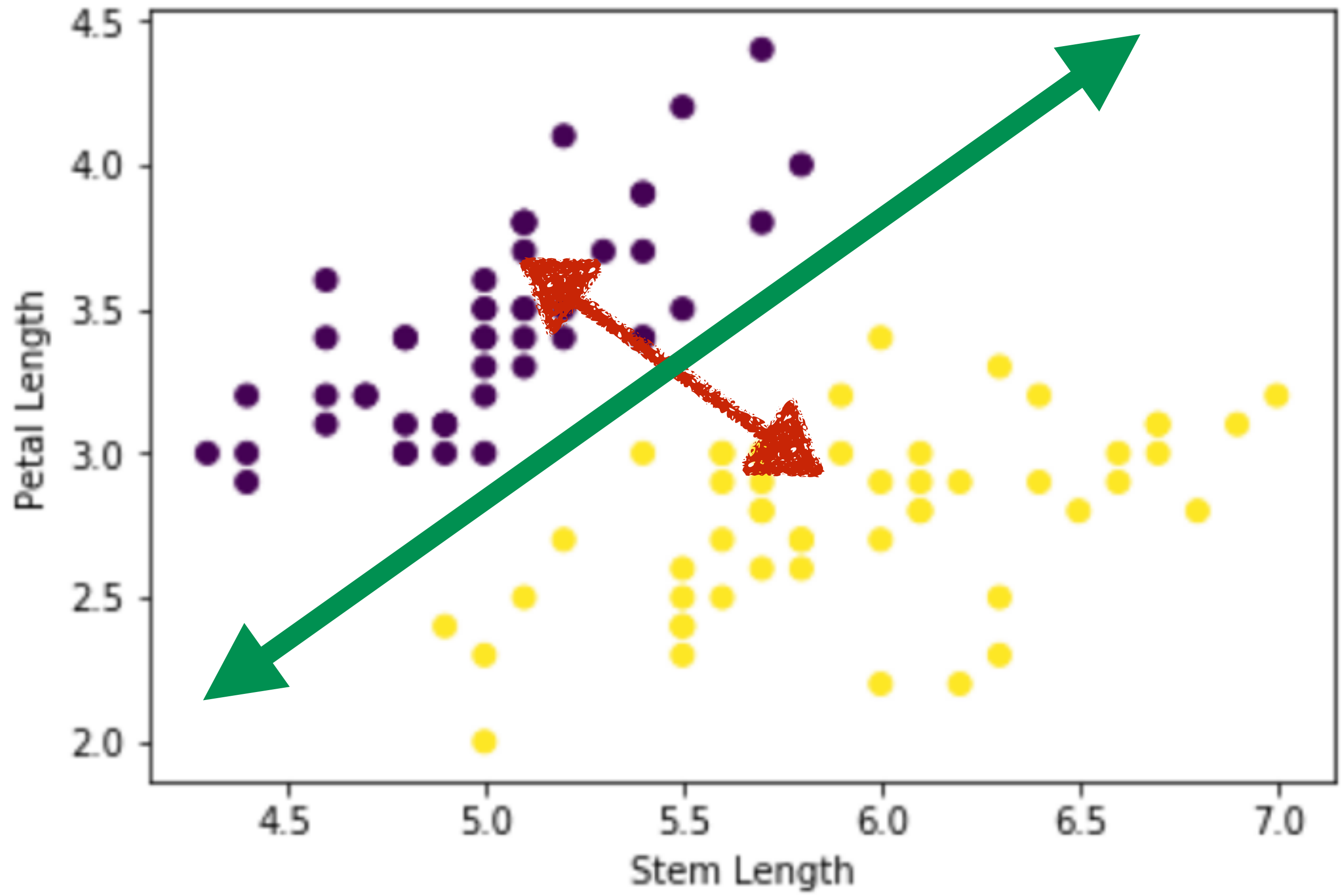


0.8 * SL - 1.3 - PL

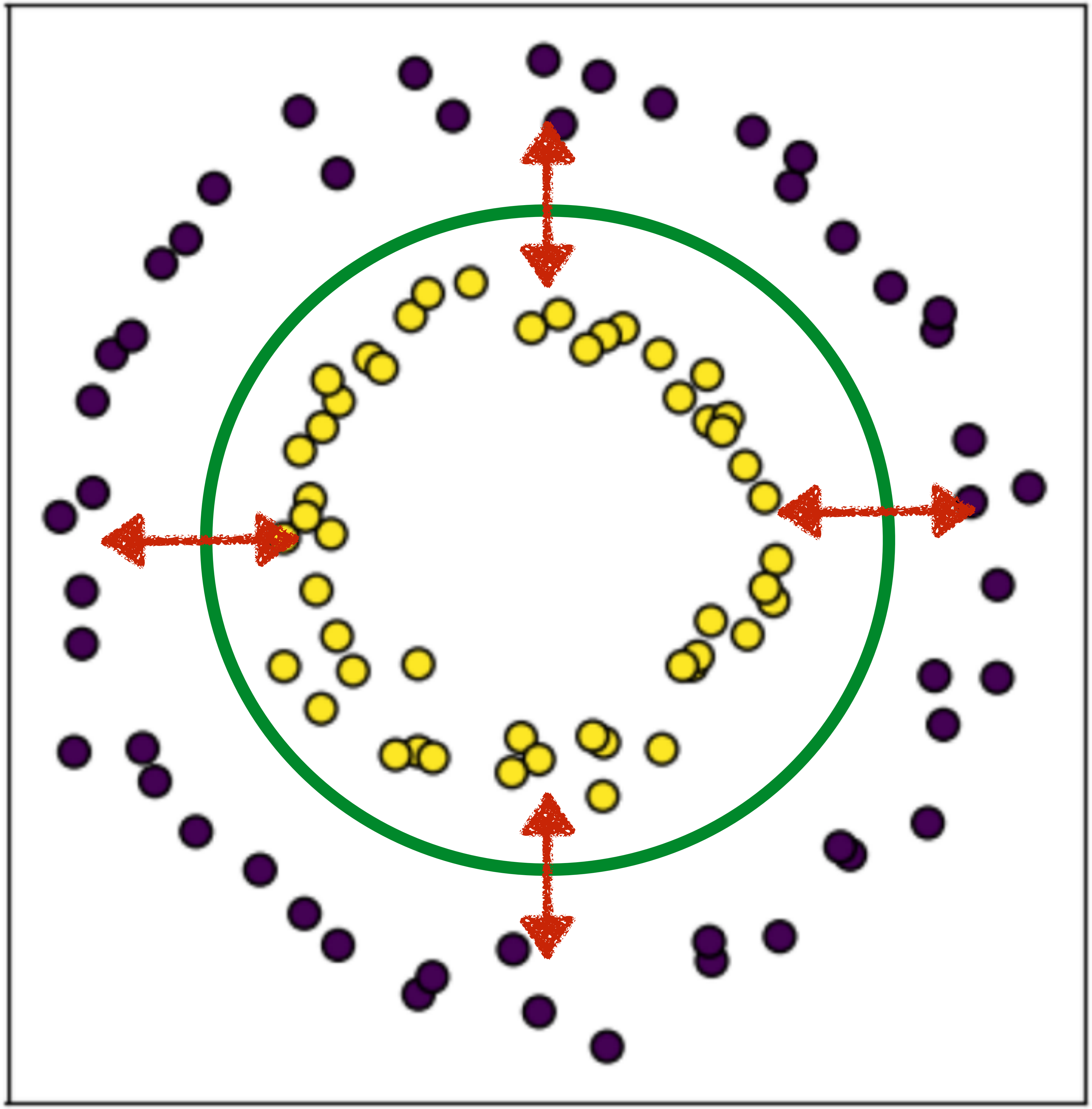
0.8 * 6 - 1.3 - 2.1 = 1.4

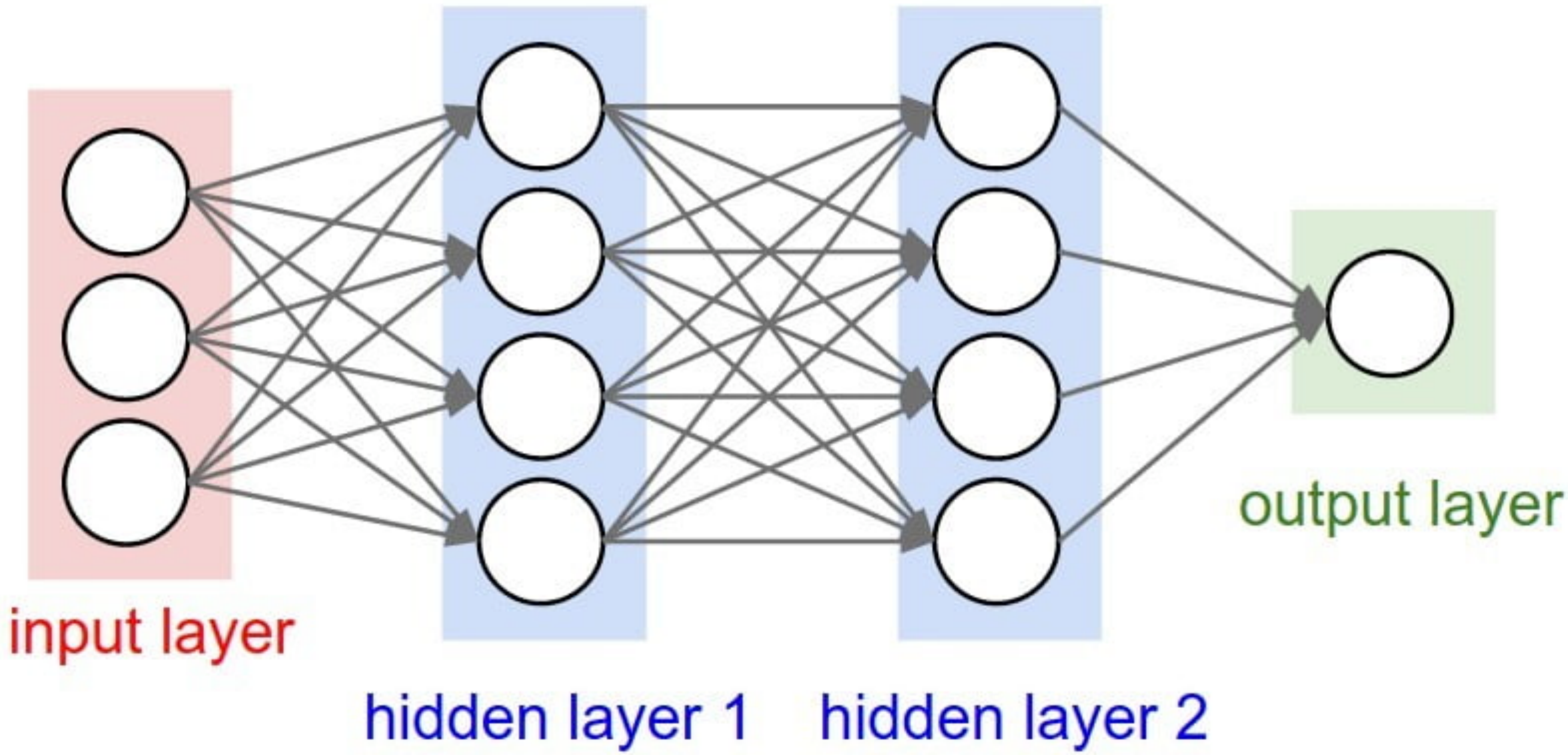
0.8 * 5.5 - 1.3 - 4.3 = -1.2

A mathematical view



$$\boxed{\text{Flower Type} = \text{Sign}(0.8 * SL - 1.3 - PL\ )}$$



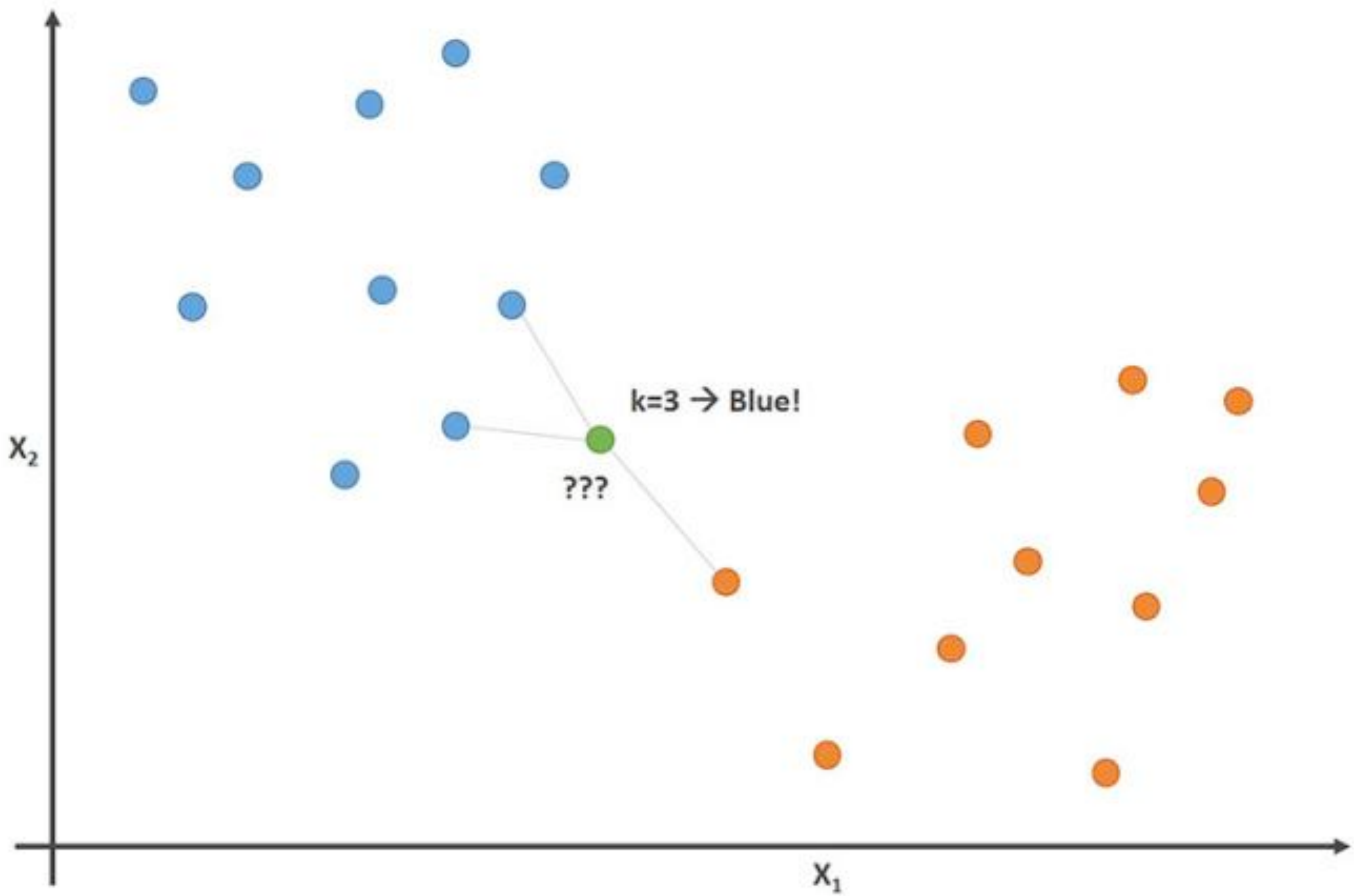


$$y = f(x)$$

Method Names

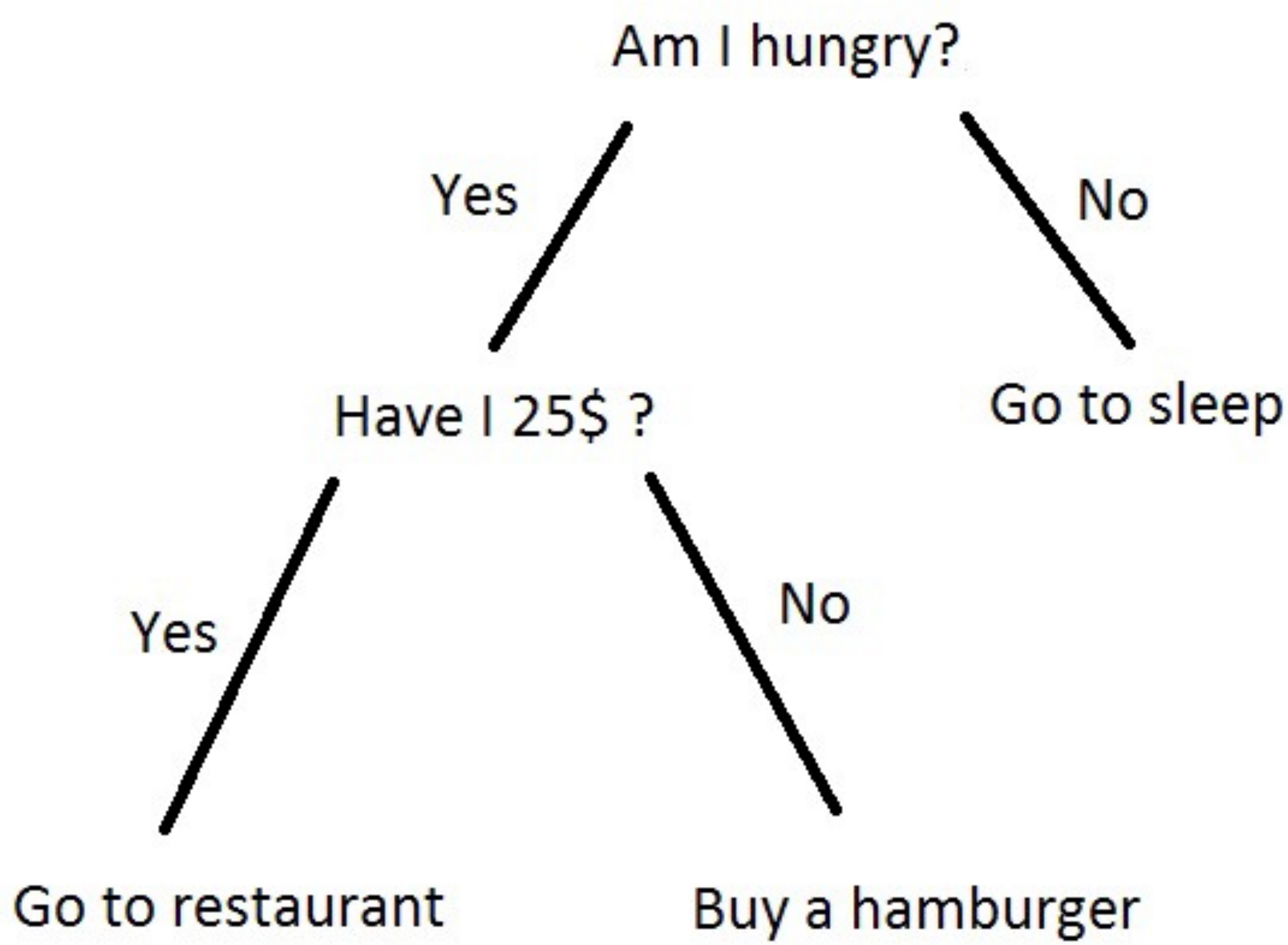
- Perceptron
- Lasso
- Ridge Classifier
- Stochastic Gradient Descent (SGD)
- Support Vector Machine (SVM)

K-Nearest Neighbors / K-Means

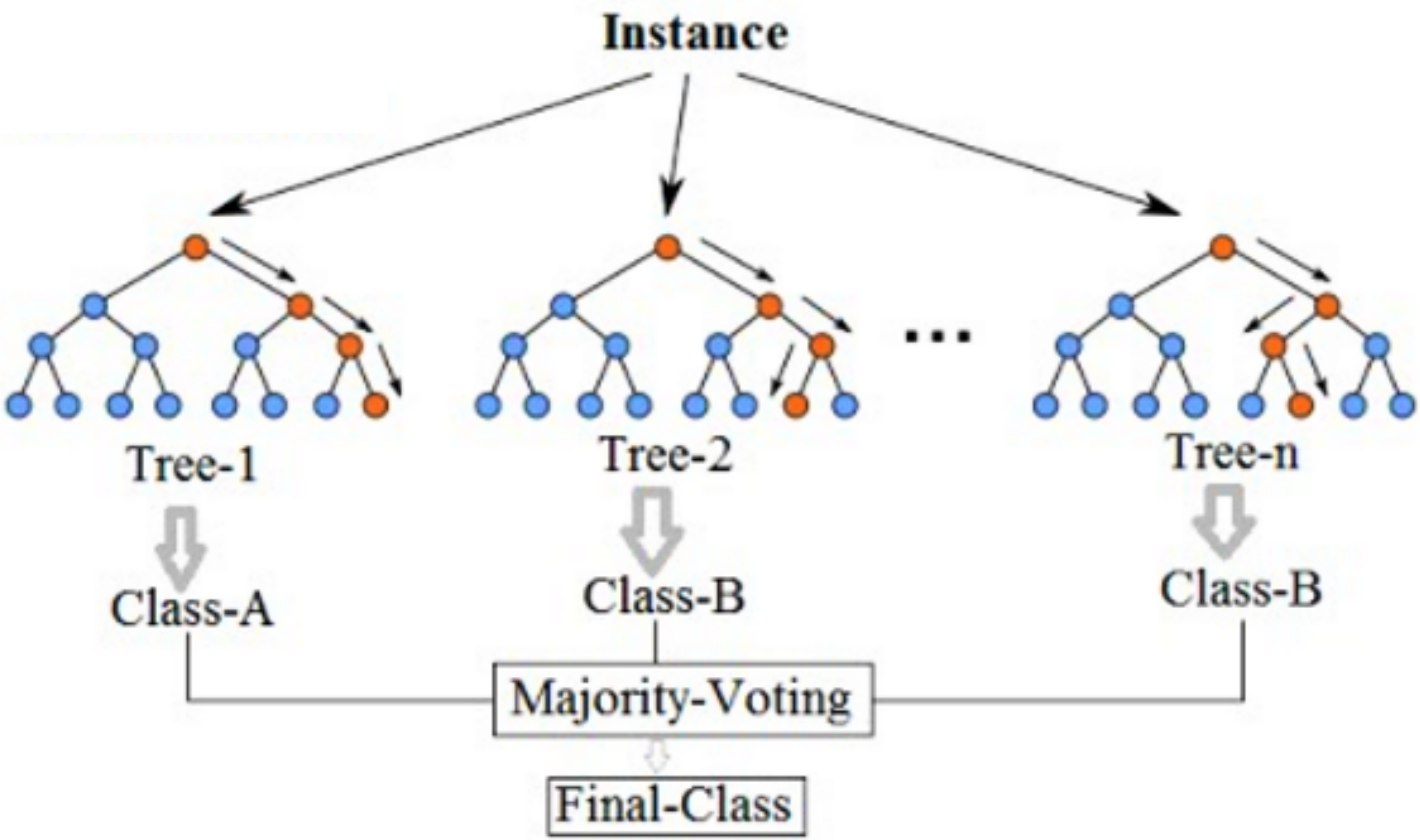


https://www.kdnuggets.com/2017/09/rapidminer-k-nearest-neighbors-laziest-machine-learning-technique.html

Decision Tree



Random Forest



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

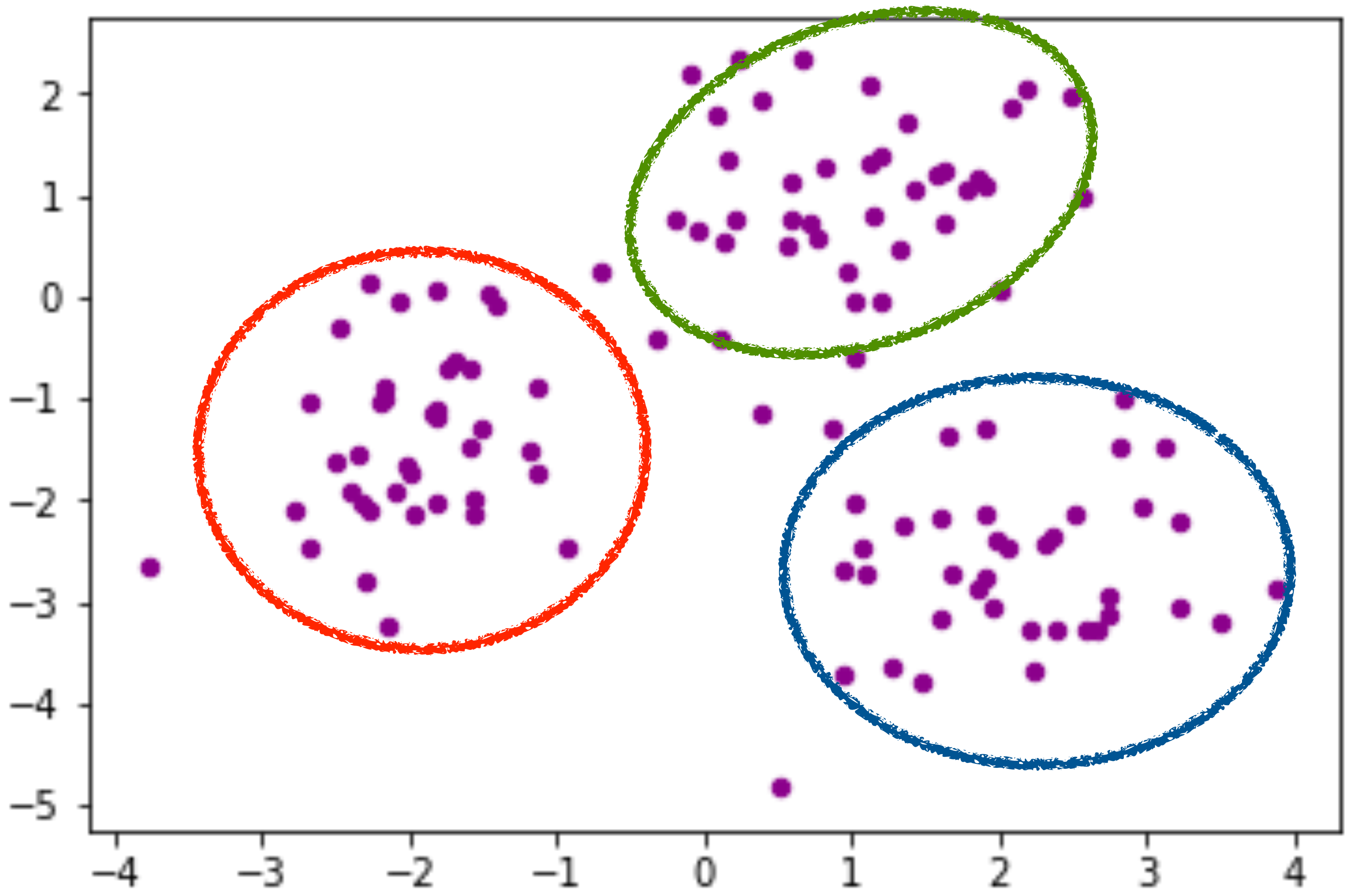
Evaluation: how good is my model?

- Compare the *true* label to the predicted one:
    - Accuracy: # Correct / # Total
- Model must be evaluated on a a different set of data than the one used to train the model.
    - Called *train* and *test* data

Supervised vs unsupervised learning

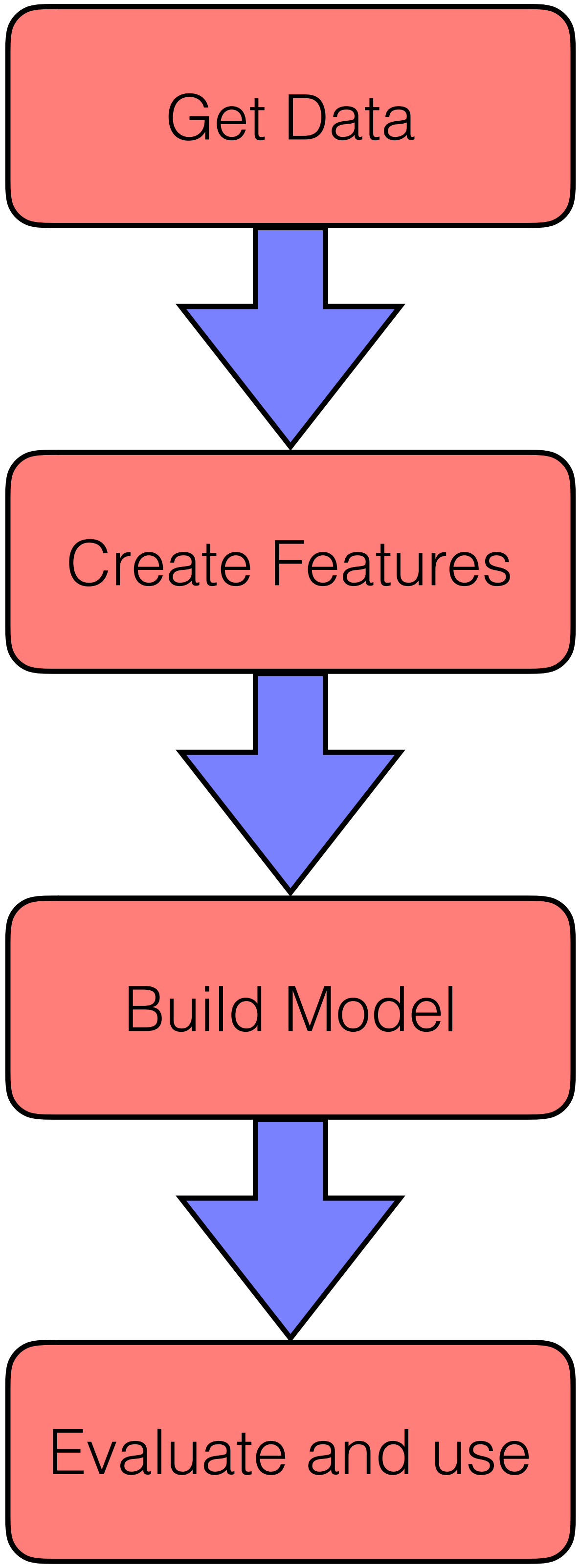
- Supervised (our example model)
  - Data contains both **inputs** and **outputs**
  - Learn to predict **output** from the **input**
- Unsupervised
  - Data contains **input** but **no output**
  - Learn patterns/clusters in the data

Unsupervised Example



E.g: What are the common themes in these movie reviews

Big Picture Process



ML in the Wild

FN Outlook

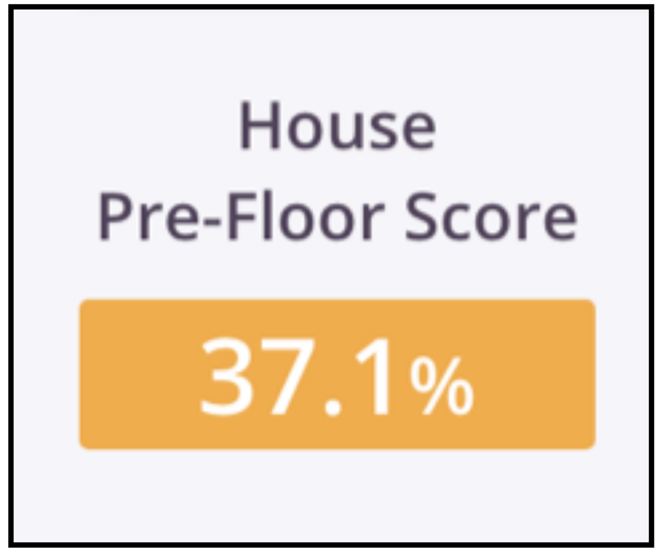
Compared to other bills in Pennsylvania, this bill is more likely to pass.

We calculated this based on **the strength of the bill sponsor, the language in the bill,** and **the network of the cosponsors.**

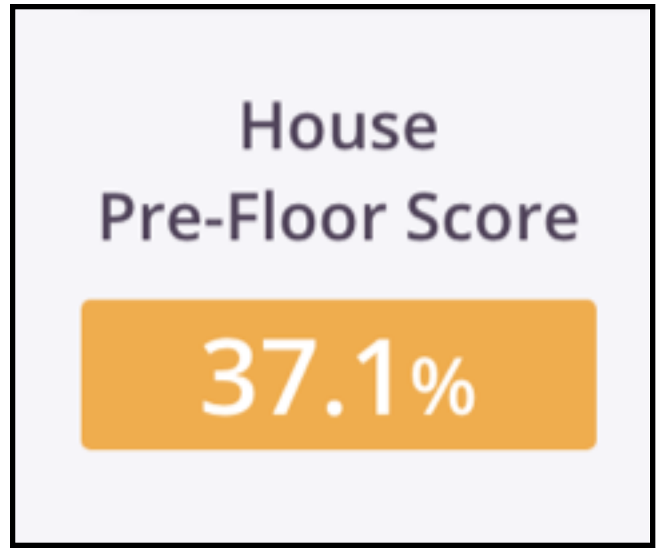
- ↑ Rosita C. Youngblood is the House Minority Caucus Secretary.
- ⇆ Last session, approximately 32.8% of bills introduced in the House were enacted.
- ↑ At least one bill with similar language passed in a previous session in this legislature.
- ↓ Historically, bills with a similar number of sponsors passed 7.5% of the time in this chamber.

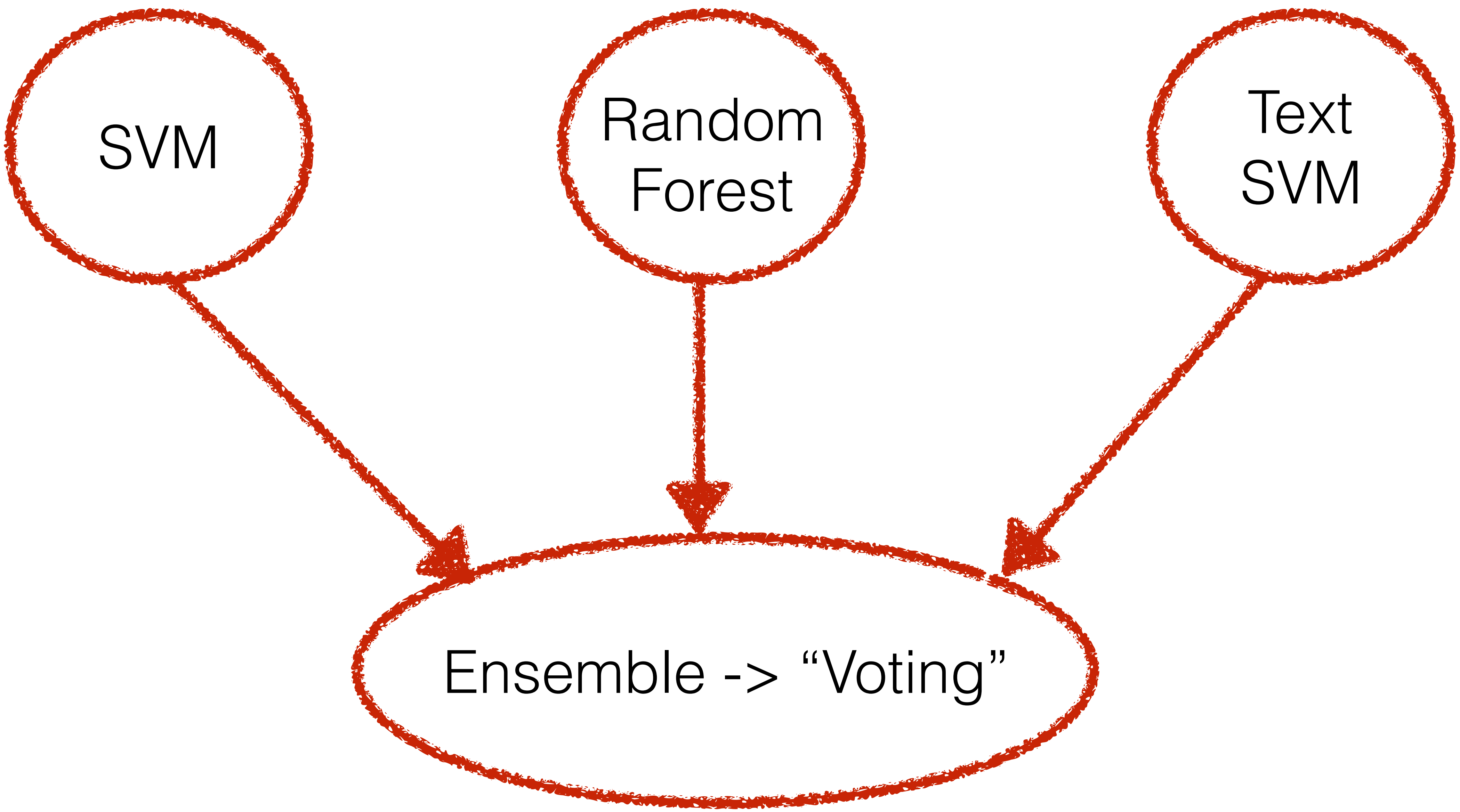
Click to Hide Analytics

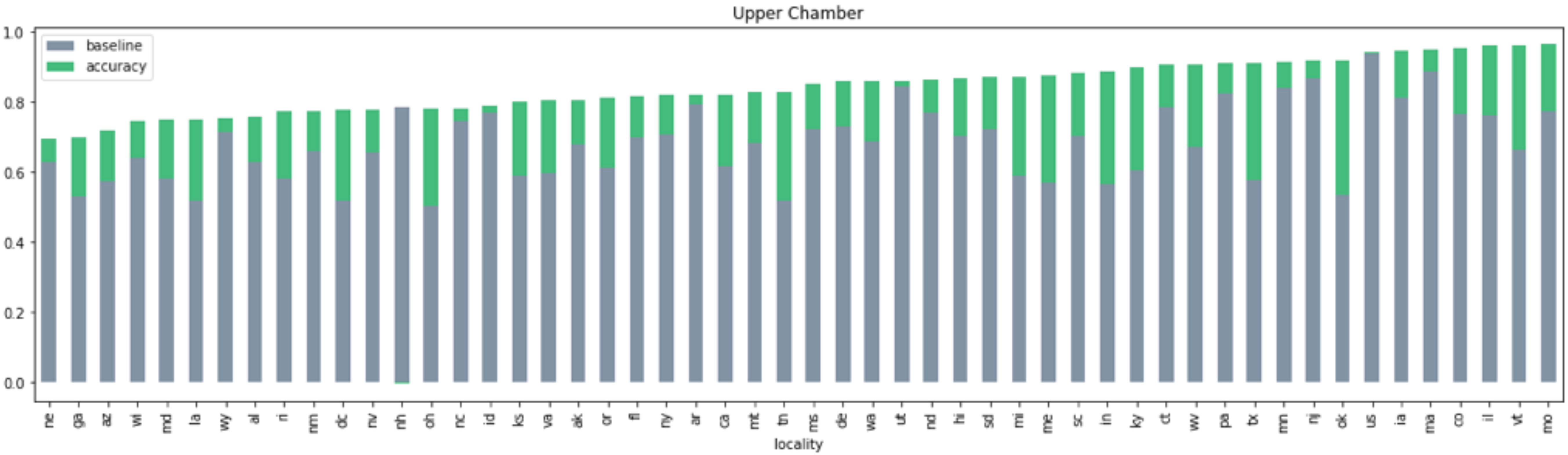
House Pre-Floor Score	House Floor Score	Senate Pre-Floor Score	Senate Floor Score
37.1%	84.5%	97.5%	71.4%

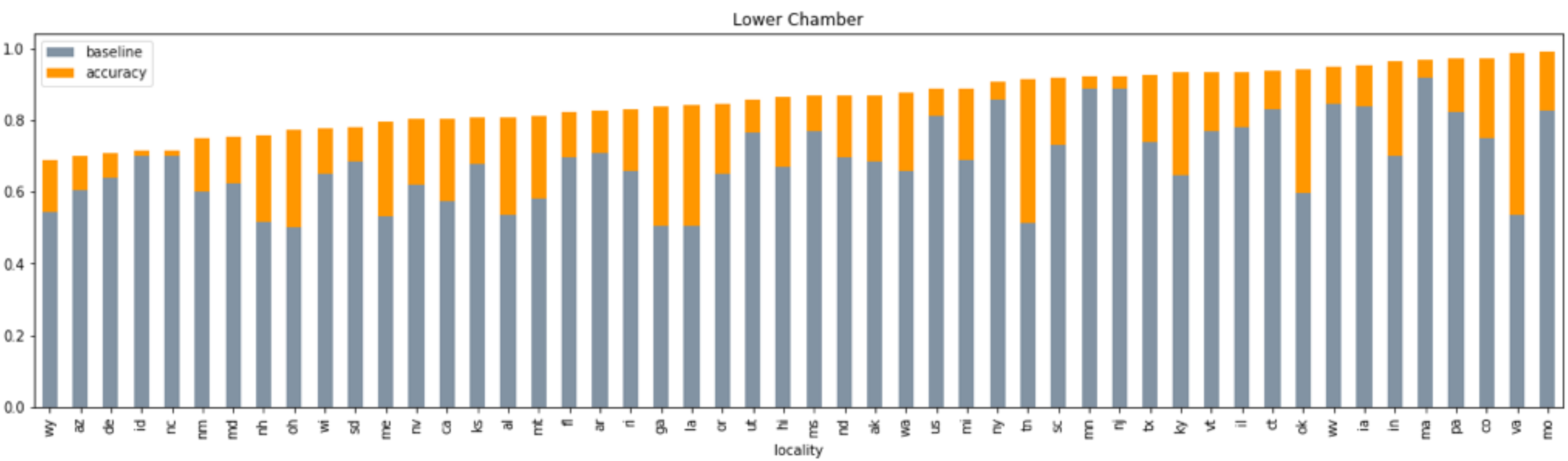


- Features:
  - # Sponsors
  - Sponsors' ideology and effectiveness
  - Leadership positions
  - Committee
  - Text


House
Pre-Floor Score
37.1%


SVM
Random Forest
Text SVM
Ensemble -> “Voting”


Upper Chamber
baseline
accuracy
1.0
0.8
0.6
0.4
0.2
0.0
ne ga az wi md la wy al ri nm dc nv nh oh nc id ks va ak or fl ny ar ca mt tn ms de wa ut nd hi sd mi me sc in ky ct wv pa tx mn nj ok us ia ma co il vt mo
locality


Lower Chamber
baseline
accuracy
1.0
0.8
0.6
0.4
0.2
0.0
wy az de id nc nm md nh oh wi sd me nv ca ks al mt fl ar ri ga la or ut hi ms nd ak wa us mi ny tn sc mn nj tx ky vt il ct ok wv ia in ma pa co va mo
locality

Questions?

Link to page with more terms defined

- Feature engineering
- Feature Selection
- Precision/Recall
- SVM/SVC

A few more useful terms

- Features:
- Classification
- Regression
- Accuracy
- Supervised vs unsupervised learning