

# INTRODUCTION TO DATA SCIENCE

This lecture is  
based on course by E. Fox and C. Guestrin, Univ of Washington

15/12/2021

WFAiS UJ, Informatyka Stosowana  
I stopień studiów

# Classification

2

## ❑ An intelligent restaurant review system

It's a big day & I want to book a table at a nice Japanese restaurant



# What is a sentiment of the review

3



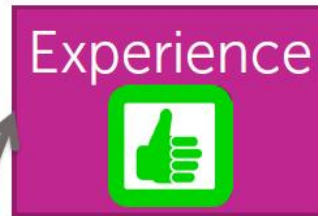
## Positive reviews not positive about everything

Sample review:

Watching the chefs create incredible edible art made the experience very unique.

My wife tried their ramen and it was pretty forgettable.

All the sushi was delicious!  
Easily best sushi in Seattle.



# Topic sentiments

4

## From reviews to topic sentiments

All reviews  
for restaurant

★★★★★ 7/21/2015

This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered right of scallop, Japanese snapper (seasonal), and the agodashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The toki, as recommended by other Yelpers was amazing. It's more chewy and the sauce/garnish is the perfect amount of flavor for the delicate toki.

★★★★★ 9/1/2015

Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have reservations, banged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

★★★★★ 6/9/2015

I came here having high expectations due to the reviews of the place, but I was bit disappointed. The restaurant is small so do make reservations when you come here. Dishes cost from \$4-26 each and dishes are small.

Novel intelligent  
restaurant review app

Experience

★★★★★

Ramen

★★★

Sushi

★★★★★

Easily best sushi  
in Seattle.

15/12/2021

# Intelligent restaurant review system

5

## All reviews for restaurant

★★★★★ 7/21/2015  
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

★★★★★ 6/11/2015  
Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have resos, banged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

★★★★★ 6/9/2015  
I came here having high expectations due to the reviews of this place, but I was bit disappointed. The restaurant is small so do make reservations when you come here. Dishes cost from \$4-26 each and dishes are small.

## Break all reviews into sentences

The seaweed salad was just OK, vegetable salad was just ordinary.

I like the interior decoration and the blackboard menu on the wall.

All the sushi was delicious.

My wife tried their ramen and it was pretty forgettable.

The sushi was amazing, and the rice is just outstanding.

The service is somewhat hectic.

Easily best sushi in Seattle.

# Core building block

6

Easily best sushi in Seattle.



Sentence Sentiment  
Classifier



Easily best sushi in Seattle.



# Intelligent restaurant review system

7

## All reviews for restaurant

★★★★☆ 7/21/2015  
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

★★★★☆ 6/11/2015  
Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have reservations, barged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

★★★☆☆ 6/9/2015  
I came here having high expectations due to the reviews of this place, but I was bit disappointed. The restaurant is small so do make reservations when you come here. Dishes cost from \$4-26 each and dishes are small.

## Break all reviews into sentences

The seaweed salad was just OK, vegetable salad was just ordinary.

I like the interior decoration and the blackboard menu on the wall.

All the sushi was delicious.

My wife tried their ramen and it was pretty forgettable.

The sushi was amazing, and the rice is just outstanding.

The service is somewhat hectic.

Easily best sushi in Seattle.

Sentence  
Sentiment  
Classifier

## Average predictions

Sushi  
★★★★★

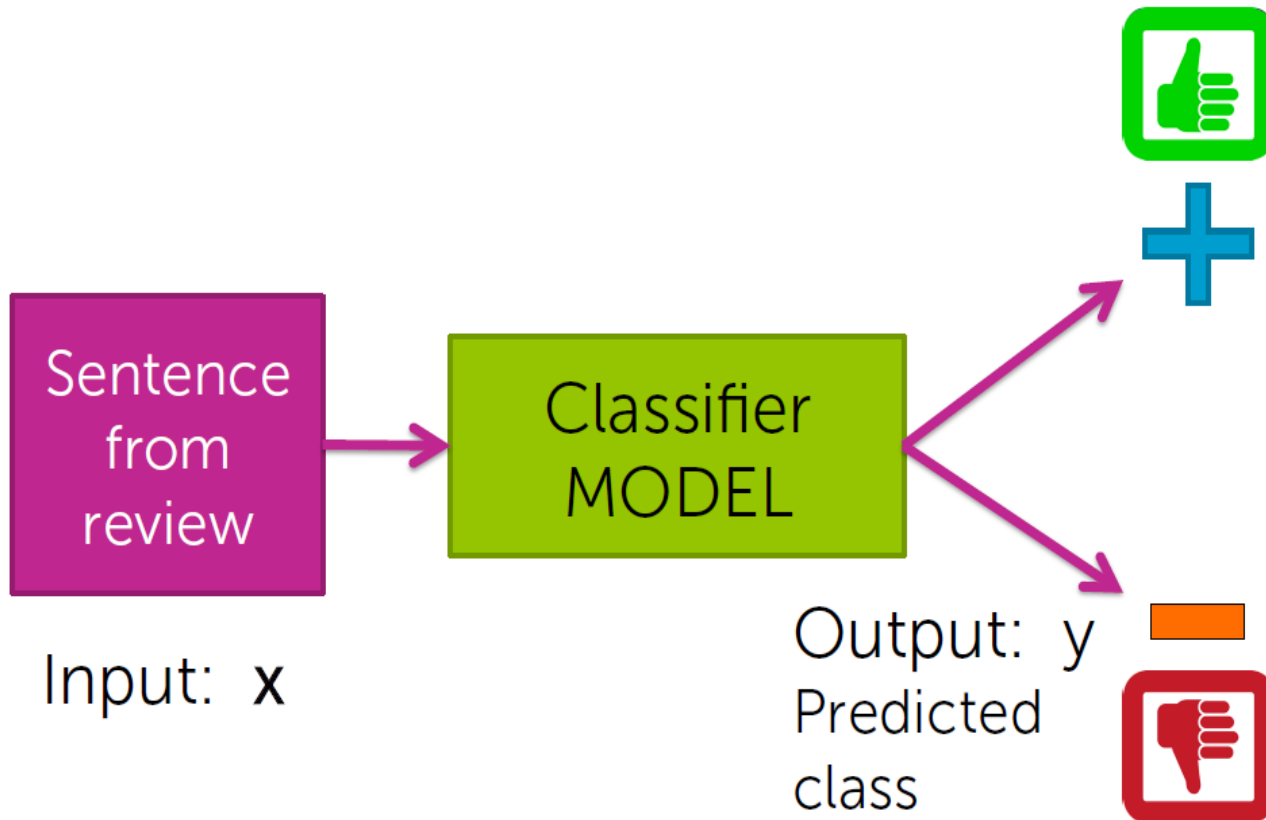
Most  
👍 & 👎

Easily best  
sushi  
in Seattle.

15/12/2021

# Classifier

8

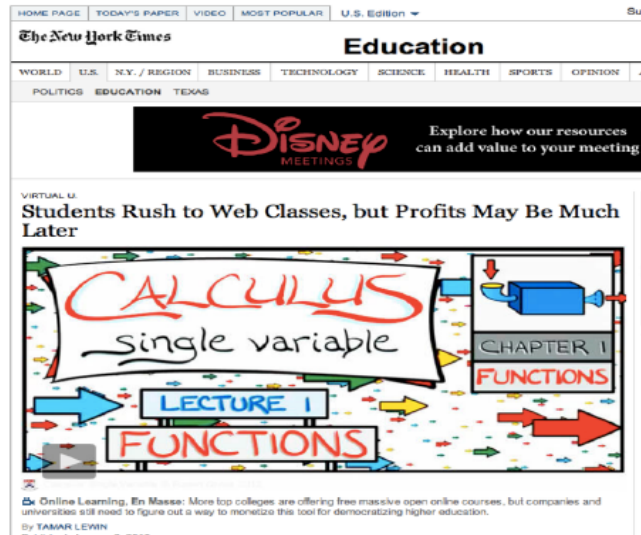




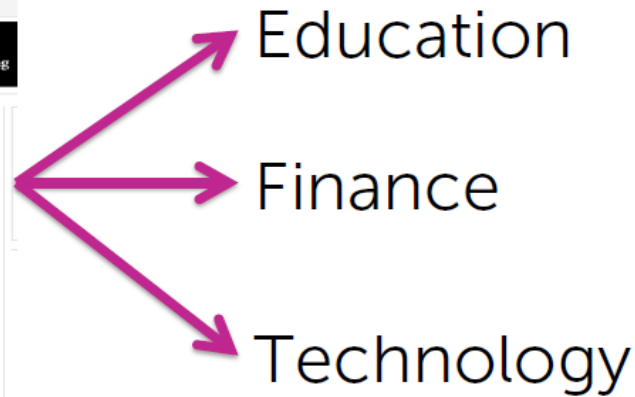
# Multiclass classifier

9

*Output  $y$  has more than 2 categories*



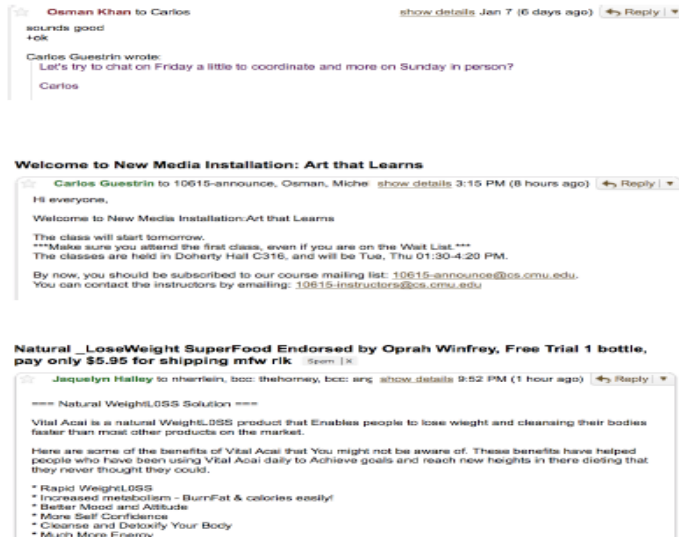
Input:  $x$   
Webpage



Output:  $y$

# Spam filtering

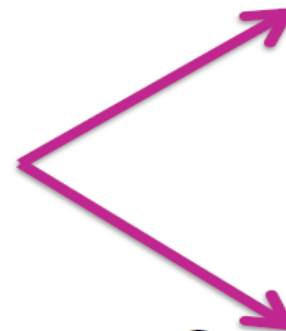
10



Input:  $x$

Text of email,  
sender, IP, ...

Not spam



Spam

Output:  $y$

12

©2015 Emily Fox & Carlos Guestrin

15/12/2021

# Image classification

11

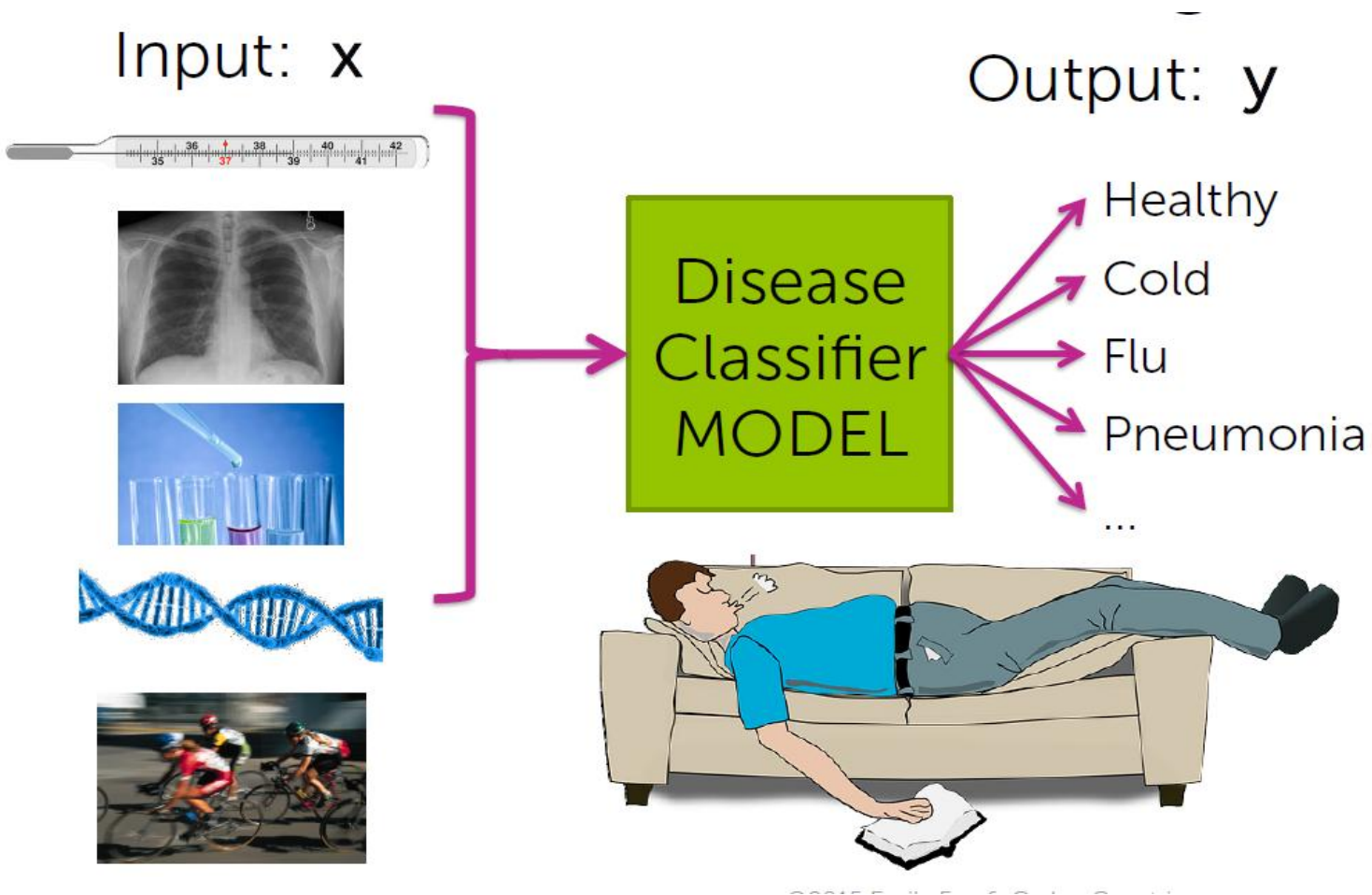


Input:  $x$   
Image pixels

Output:  $y$   
Predicted object

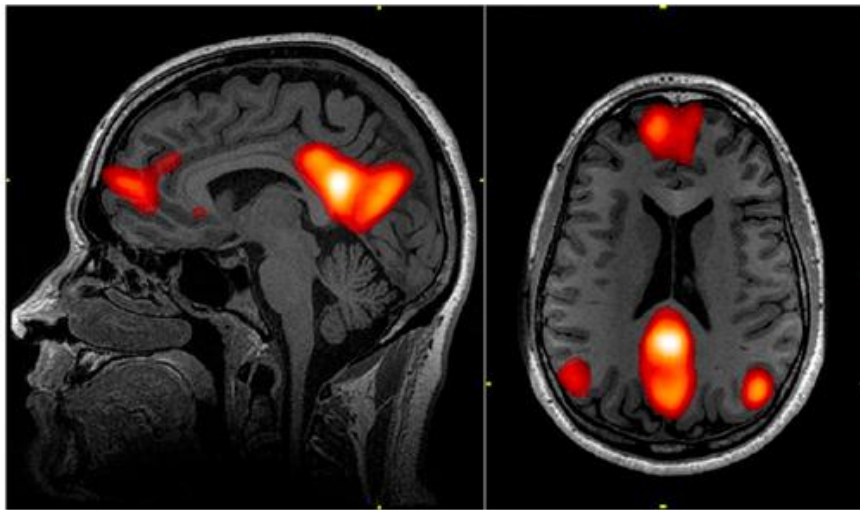
# Personalized medical diagnosis

12

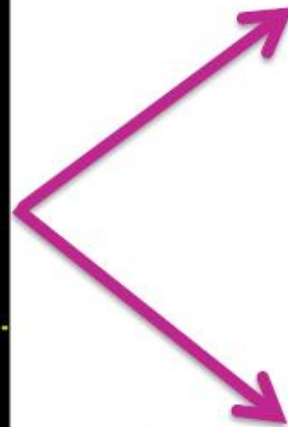


# Reading your mind

13



"Hammer"

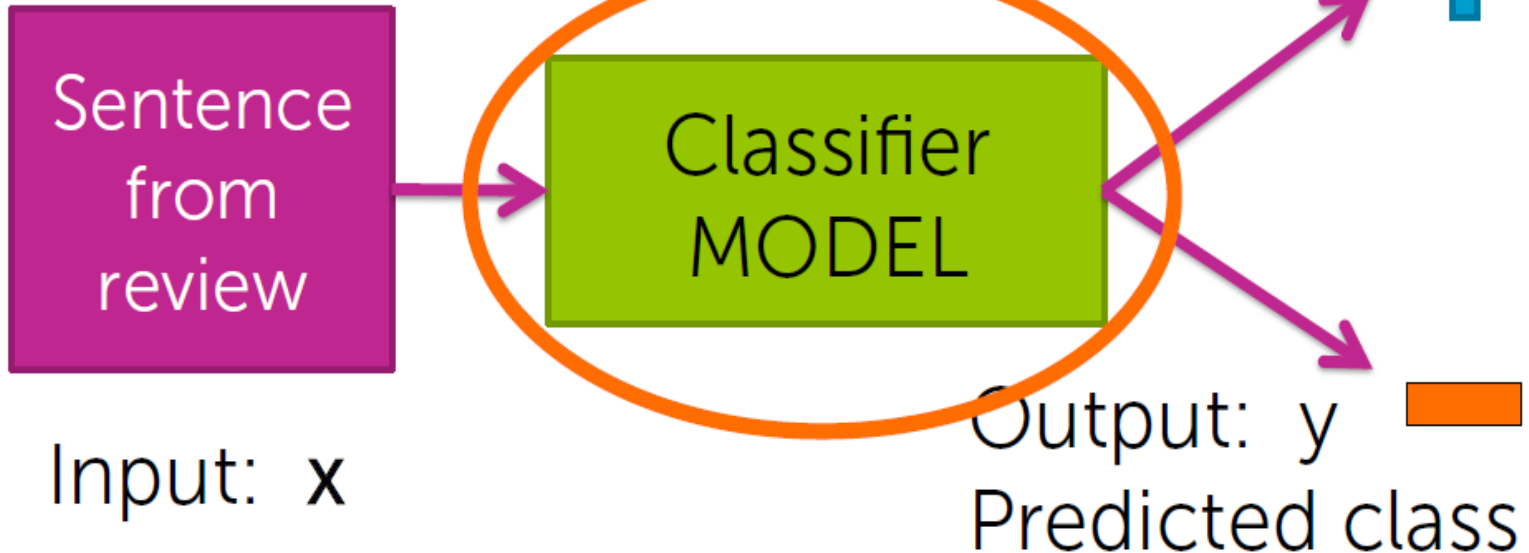


"House"

# Representing classifiers

14

How does it work???



# Simple threshold classifier

15

List of positive words	List of negative words
great, awesome, good, amazing,...	bad, terrible, disgusting, sucks,...




Sentence from review




**Simple threshold classifier**  
Count positive & negative words in sentence

If *number of positive words* > *number of negative words*:

$\hat{y} =$  

Else:

$\hat{y} =$  

Input:  $x$

# Simple threshold classifier

16

List of positive words	List of negative words
great, awesome, good, amazing,...	bad, terrible, disgusting, sucks,...

Sushi was great, the food was awesome, but the service was terrible.

## Simple threshold classifier

Count positive & negative words in sentence

2

If *number of positive words* > *number of negative words*:

$\hat{y} =$  +

Else:

$\hat{y} =$  -


1



# Problems with threshold classifier

17

- How do we get list of positive/negative words?
- Words have different degrees of sentiment:
  - Great > good
  - How do we weigh different words?
- Single words are not enough:
  - *Good* → Positive
  - *Not good* → Negative



Addressed  
by learning  
a classifier

Addressed  
by more  
elaborate  
features

# A (linear) classifier

18

Will use training data to learn a weight for each word

Word	Weight
good	1.0
great	1.5
awesome	2.7
bad	-1.0
terrible	-2.1
awful	-3.3
restaurant, the, we, where, ...	0.0
...	...

# Scoring a sentence

19

Word	Weight
good	1.0
great	<u>1.2</u>
awesome	<u>1.7</u>
bad	-1.0
terrible	<u>-2.1</u>
awful	-3.3
restaurant, the, we, where, ...	0.0
...	...

Input x:

Sushi was great,  
the food was awesome,  
but the service was terrible.

$$\begin{aligned} \text{Score}(x) &= 1.2 + 1.7 - 2.1 \\ &= 0.8 \end{aligned}$$

$$\text{Score}(x) > 0 \Rightarrow +$$

if

$$\text{Score}(x) < 0 \Rightarrow -$$

Called a linear classifier, because output is weighted sum of input.

# Simple linear classifier

20

Word	Weight
...	...

Sentence  
from  
review

Input:  $\mathbf{x}$

## Simple linear classifier

$Score(x)$  = weighted count of words in sentence

If  $Score(x) > 0$ :

$\hat{y} =$  

Else:

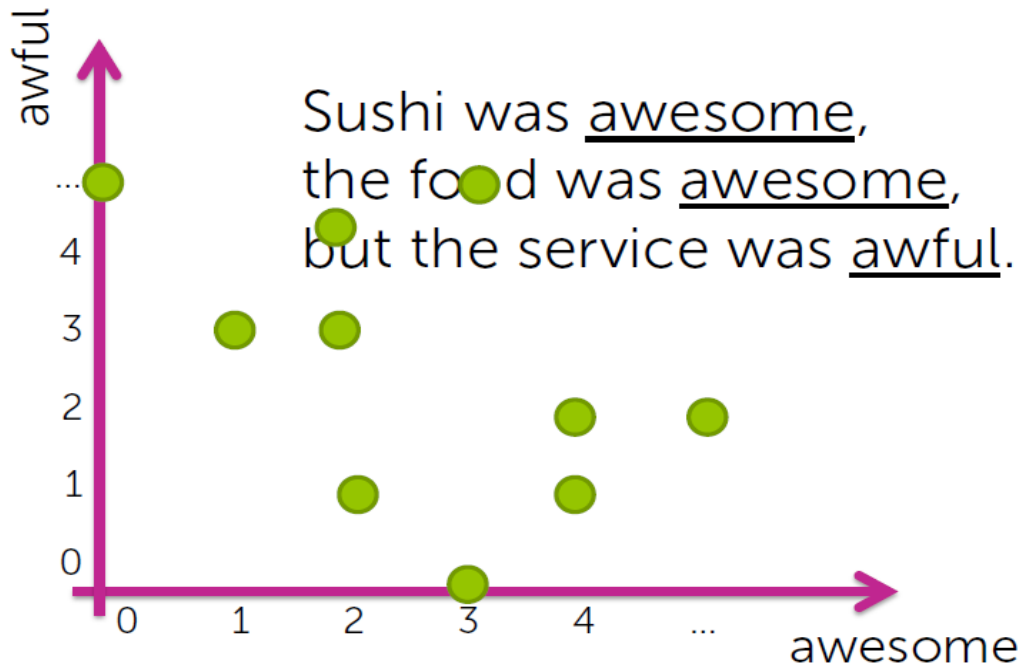
$\hat{y} =$  

# Suppose only two words had non-zero weight

21

Word	Weight
awesome	1.0
awful	-1.5

→  $\text{Score}(x) = 1.0 \# \text{awesome} - 1.5 \# \text{awful}$

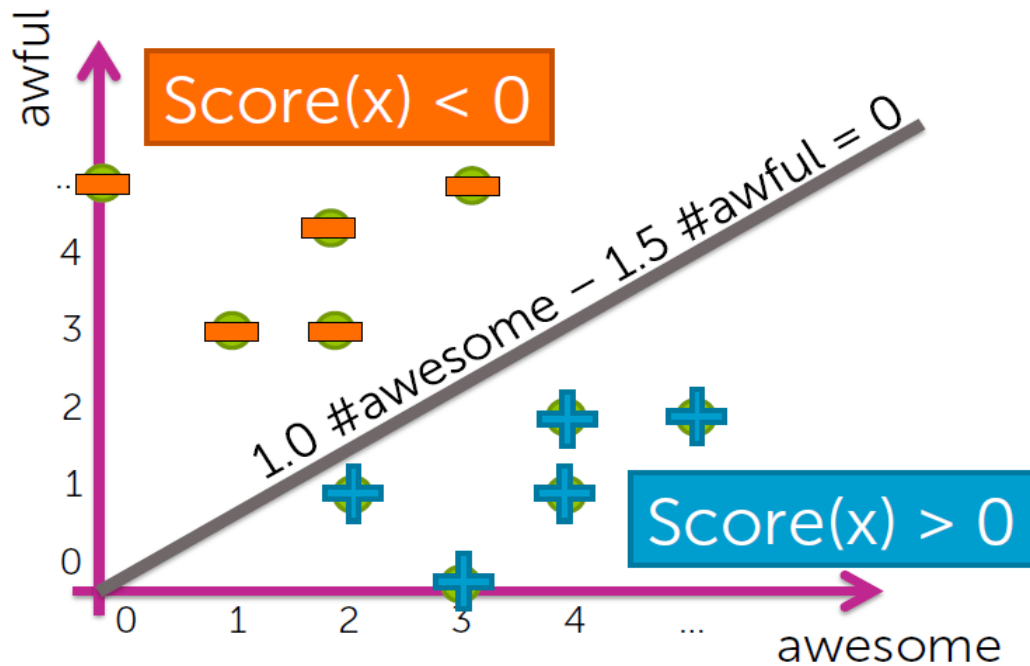


# Decision boundary example

22

Word	Weight
awesome	1.0
awful	-1.5

→  $\text{Score}(x) = 1.0 \# \text{awesome} - 1.5 \# \text{awful}$

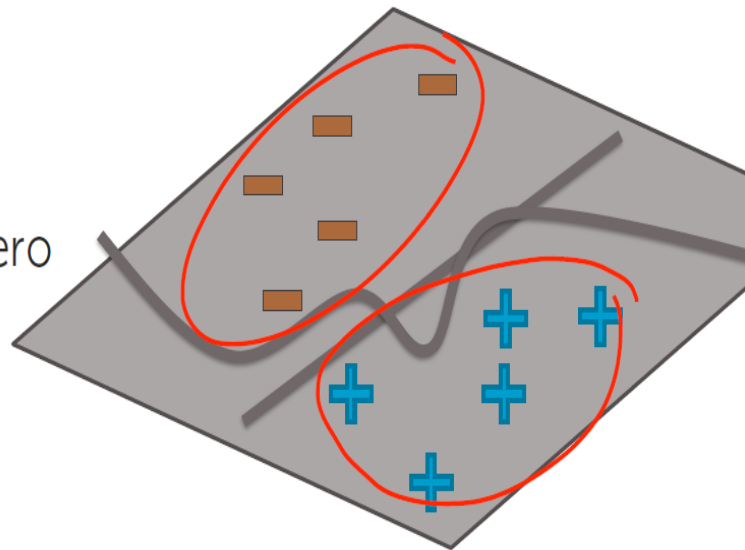


# Decision boundary

23

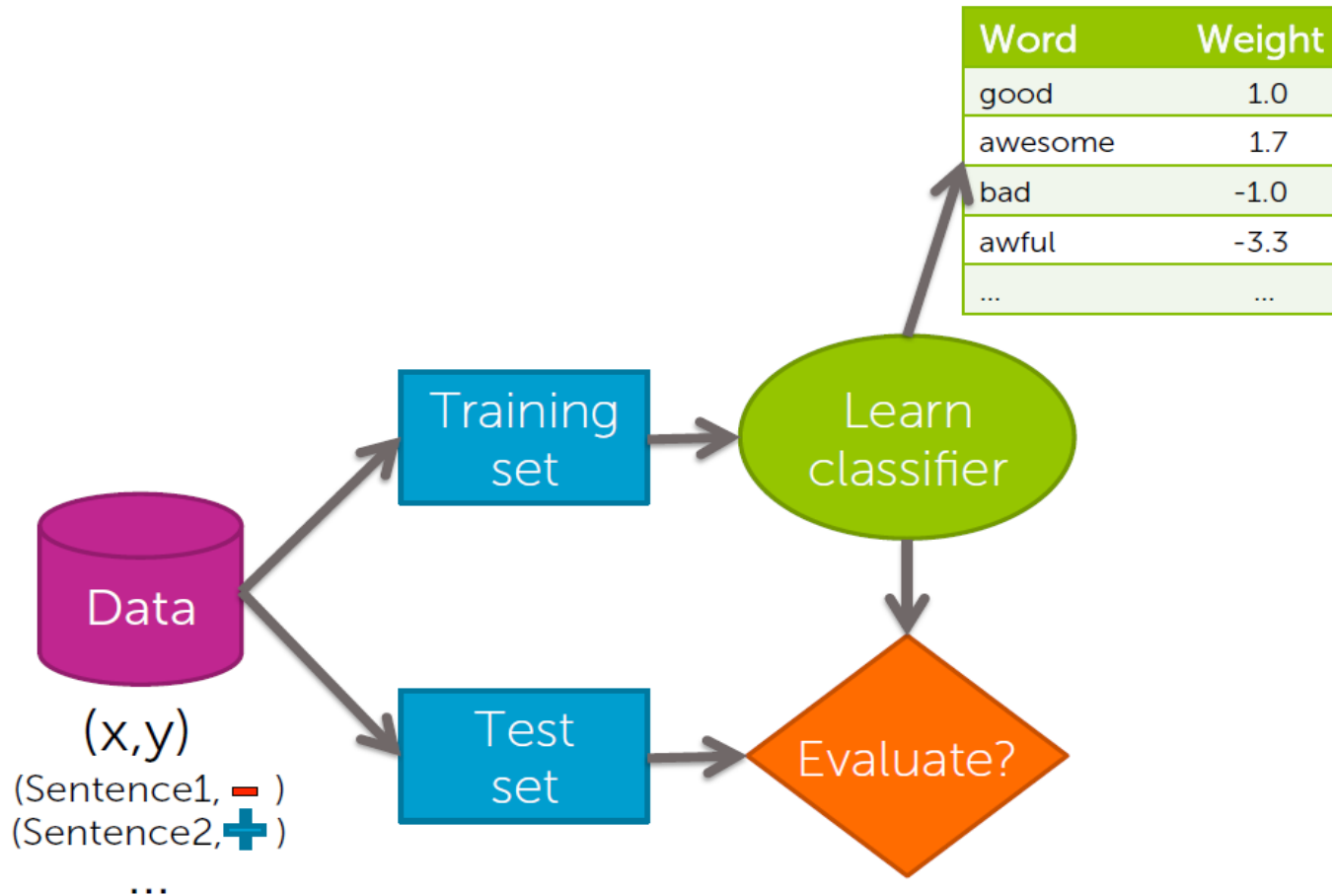
## Separates positive & negative predictions

- For linear classifiers:
  - When 2 weights are non-zero  
→ line
  - When 3 weights are non-zero  
→ plane
  - When many weights are non-zero  
→ hyperplane
- For more general classifiers  
→ more complicated shapes



# Training a classifier = Learning the weights

24





# Classification error & accuracy

25

- Error measures fraction of mistakes

$$\text{error} = \frac{\# \text{ of mistakes}}{\text{Total \# of sentences}}$$

- Best possible value is 0.0

- Often, measure **accuracy**

- Fraction of correct predictions

$$\text{accuracy} = \frac{\# \text{ of correct}}{\text{Total \# of sentences}}$$

- Best possible value is 1.0

# What if you ignore the sentence and just guess?

26

- For binary classification:
  - Half the time, you'll get it right! (on average)
    - ➔ accuracy = 0.5
- For k classes, accuracy =  $1/k$ 
  - 0.333 for 3 classes, 0.25 for 4 classes,...

At the very, very, very least,  
you should healthily beat random...  
Otherwise, it's (usually) pointless...

# Is a classifier with 90% accuracy good?

## Depends...

27

2010 data shows:  
*"90% emails sent are spam!"*

Predicting every email is spam  
gets you 90% accuracy!!!

Majority class prediction

Amazing performance when  
there is class imbalance

(but silly approach)

- One class is more common than others
- Beats random (if you know the majority class)

# What is a good accuracy?

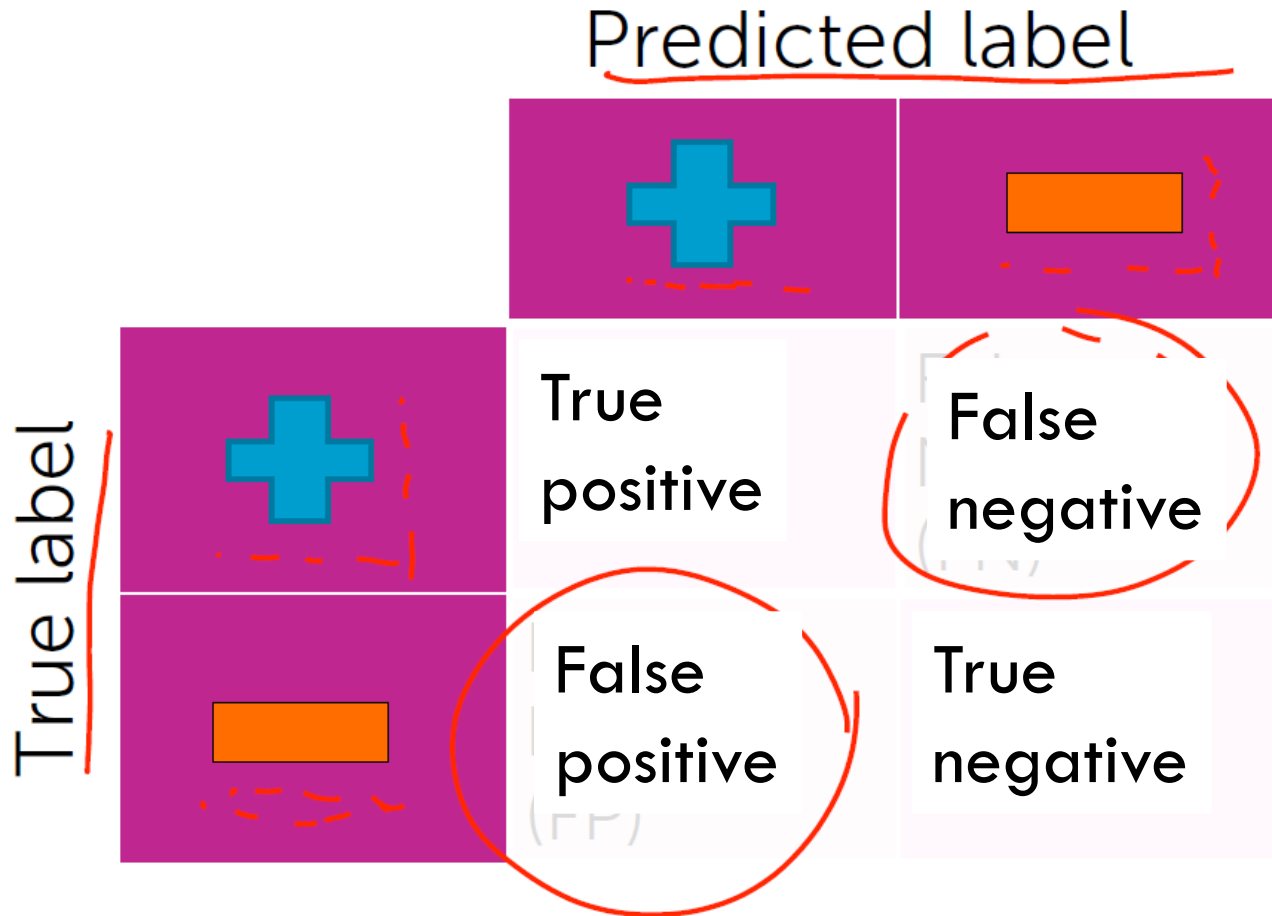
28

So, always be digging in and asking the hard questions about reported accuracies

- Is there class imbalance?
- How does it compare to a simple, baseline approach?
  - Random guessing
  - Majority class
  - ...
- Most importantly:  
*what accuracy does my application need?*
  - What is good enough for my user's experience?
  - What is the impact of the mistakes we make?

# Types of mistakes

29



# Cost of mistakes

30

Cost of different types of mistakes can be different (& high) in some applications





	Spam filtering	Medical diagnosis
False negative	Annoying	Disease not treated
False positive	Email lost <i>Higher cost</i>	Wasteful treatment

# Confusion matrix: binary classification

31

100 test examples

Predicted label

		
 60	50	10
 40	5	35

$$\text{accuracy} = \frac{85}{100} = 0.85$$

# Confusion matrix: multiclass classification

32

100 test examples

		Predicted label		
		Healthy	Cold	Flu
True label	Healthy 70	60	8	2
	Cold 20	4	12	4
	Flu 10	0	2	8

$$\text{accuracy} = \frac{80}{100} = 0.8$$



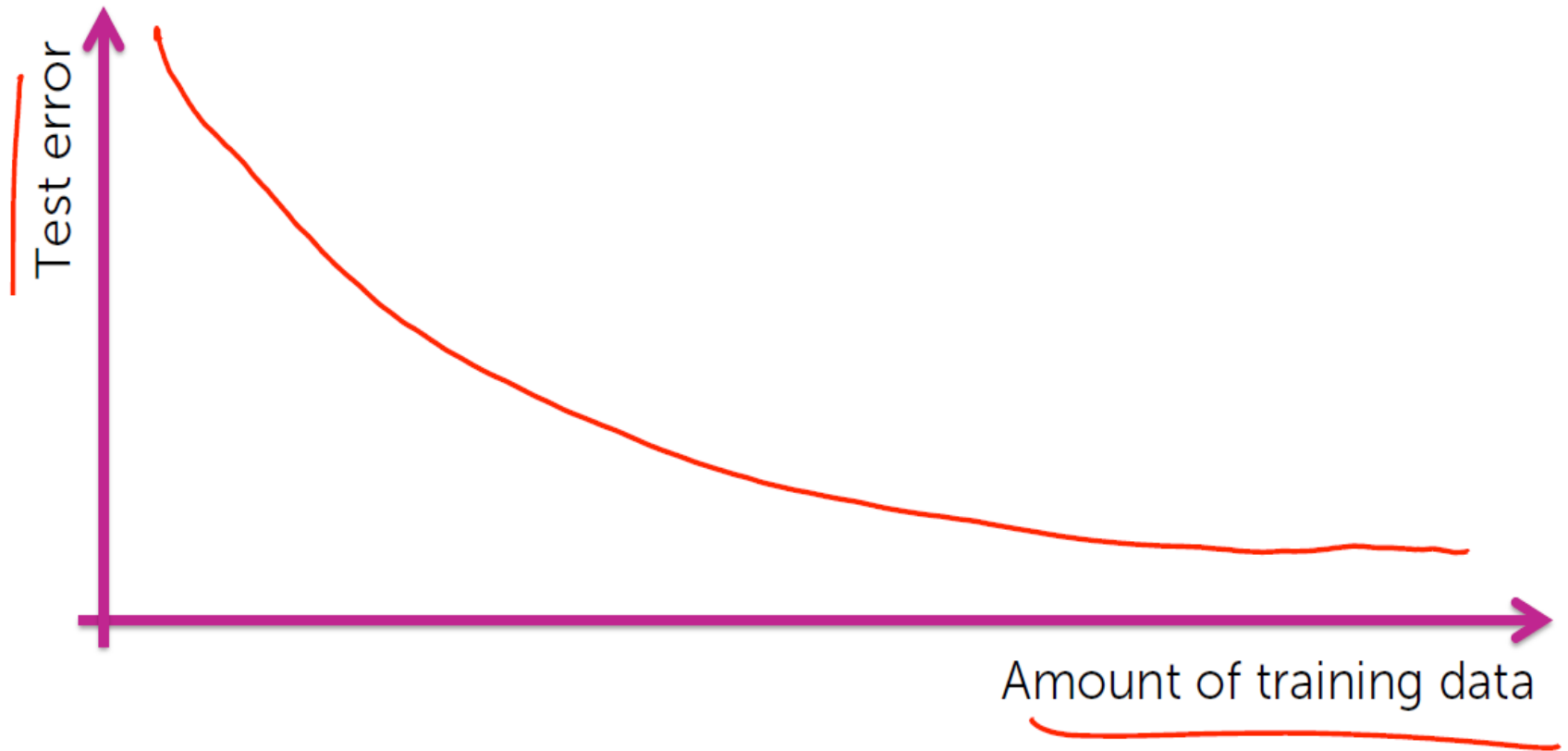
# How much data does a model need to learn?

33

- The more the merrier 😊
  - But data quality is most important factor
- Theoretical techniques sometimes can bound how much data is needed
  - Typically too loose for practical application
  - But provide guidance
- In practice:
  - More complex models require more data
  - Empirical analysis can provide guidance

# Learning curves

34

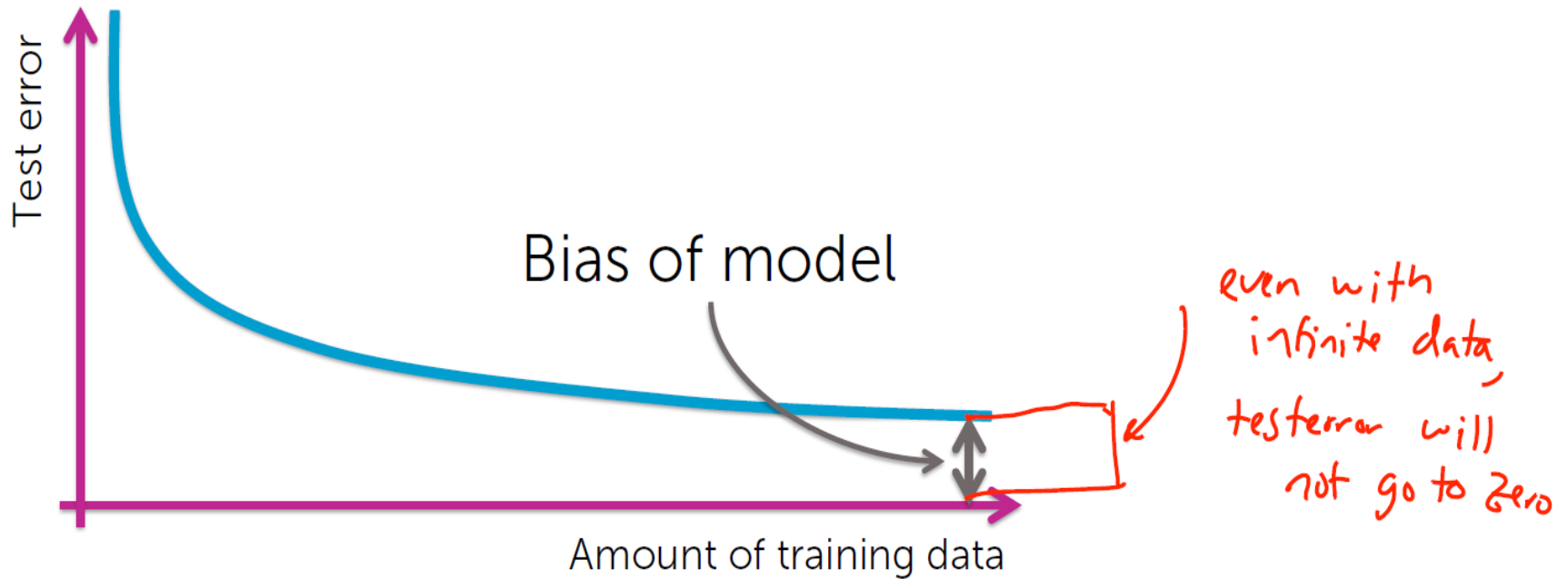


# Learning curves

35

Is there a limit?

Yes, for most models...



# More complex models tend to have less bias...

36

Sentiment classifier using single words can do OK, but...

Never classifies correctly:  
"The sushi was not good."

More complex model:  
consider pairs of words (bigrams)

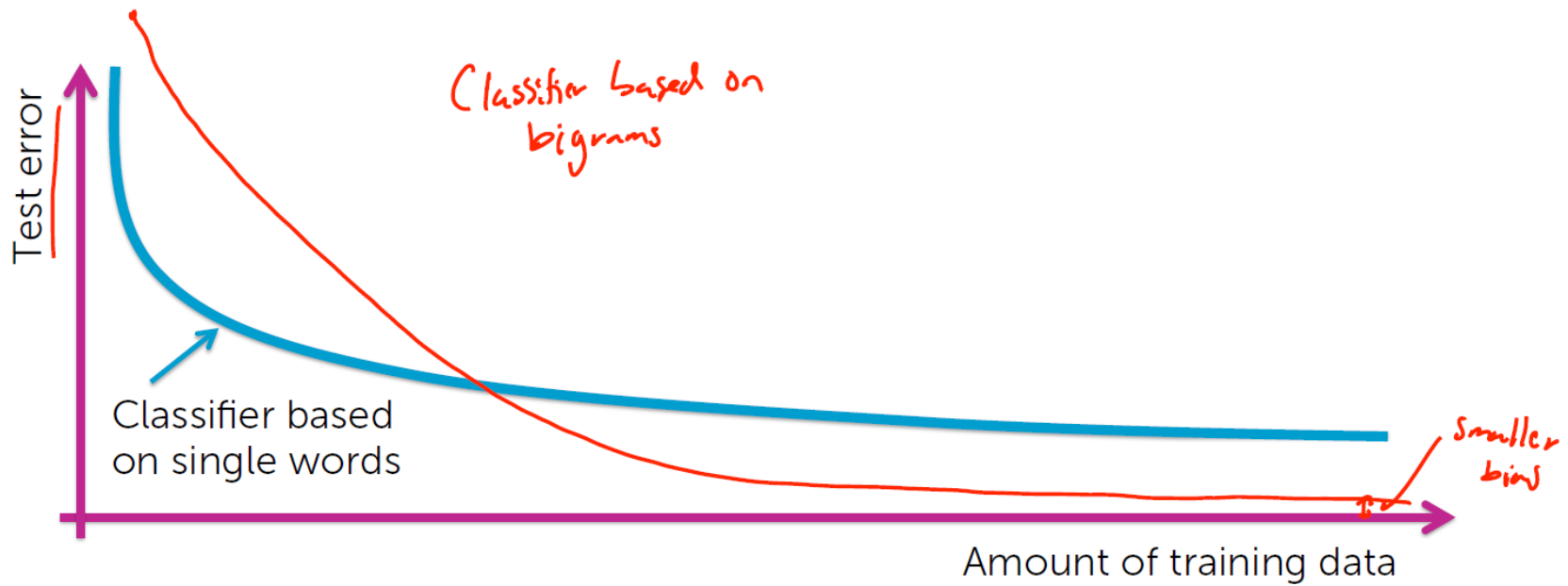
Word	Weight
good	+1.5
not good	-2.1

Less bias →  
potentially more accurate,  
needs more data to learn

# Classification based on bigrams

37


Models with less bias tend to need more data to learn well, but do better with sufficient data



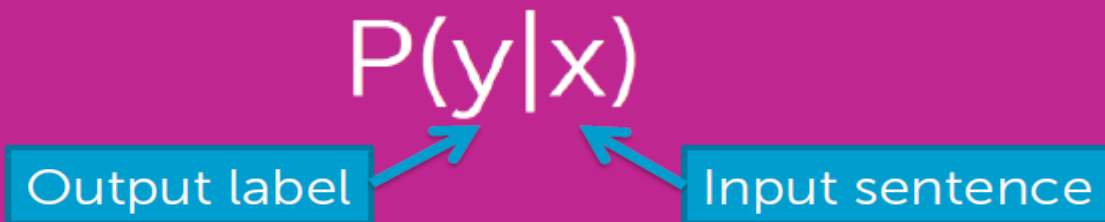
46

# How confident is your prediction?

38

- Thus far, we've outputted a prediction 
- But, how sure are you about the prediction?
  - *"The sushi & everything else were awesome!"* ←  $P(y=+|x) = 0.99$
  - *"The sushi was good, the service was OK."* ←  $P(y=+|x) = 0.55$

Many classifiers provide a confidence level:



Extremely useful in practice

# We have discussed how to

39

- Identify a classification problem and some common applications
- Describe decision boundaries and linear classifiers
- Train a classifier
- Measure its error
  - Some rules of thumb for good accuracy
- Interpret the types of error associated with classification
- Describe the tradeoffs between model bias and data set size
- Use class probability to express degree of confidence in prediction