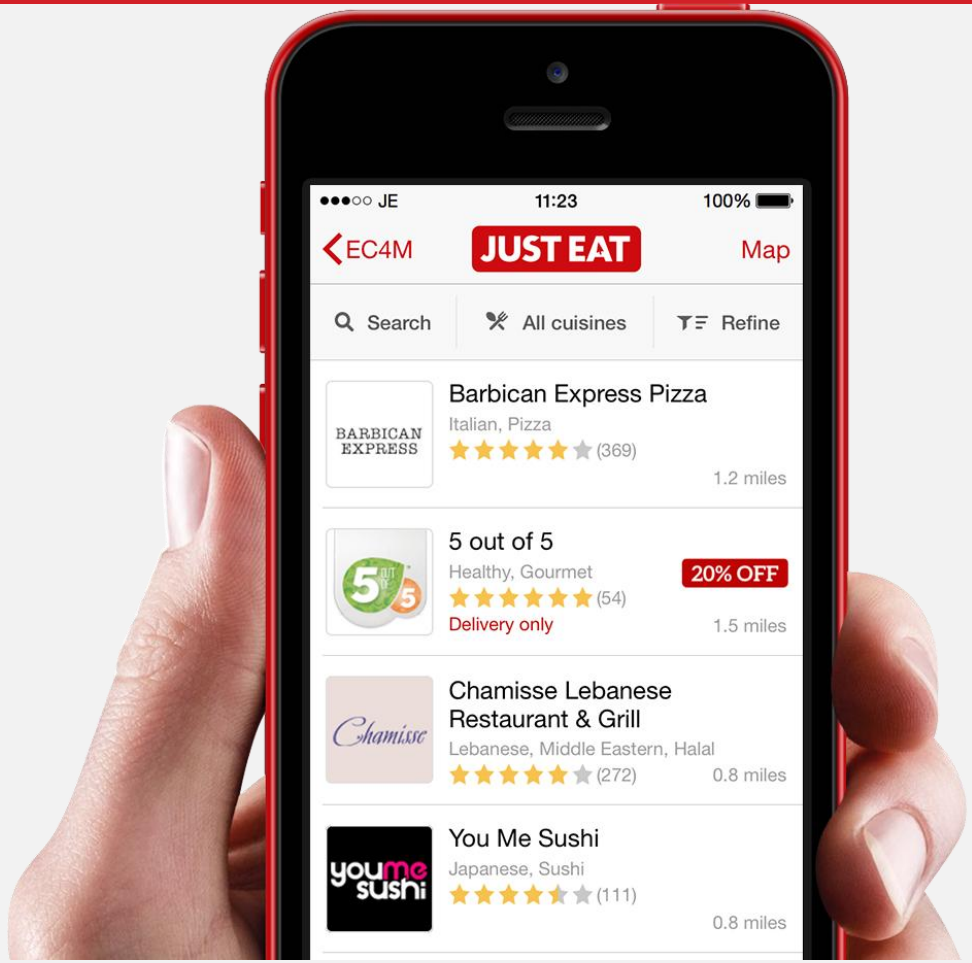
Natural language processing and Sentiment analysis

Rolando P. Hong Enriquez

JUST EAT

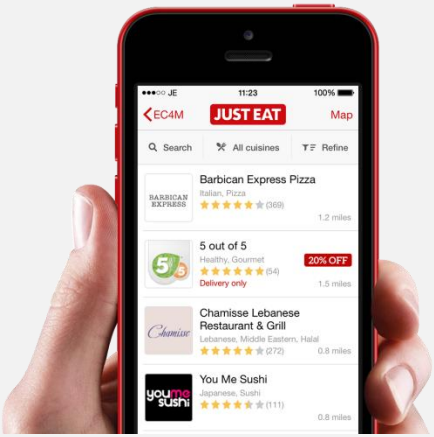
Rolando P. Hong Enriquez

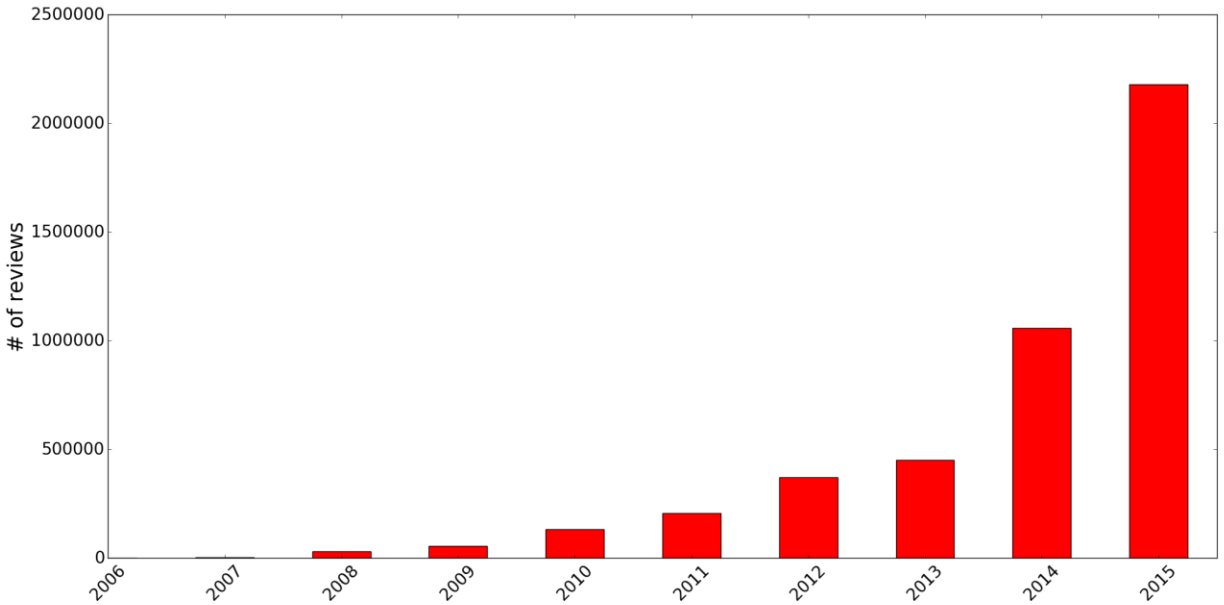
JUST EAT



Rolando P. Hong Enriquez

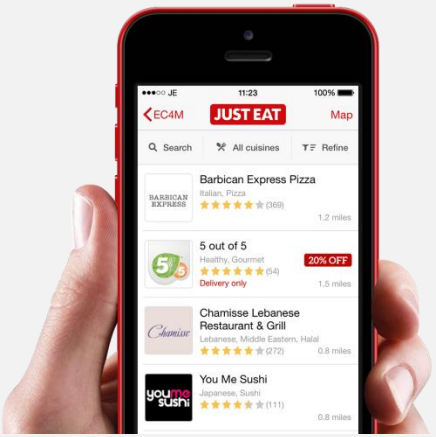
JUST EAT

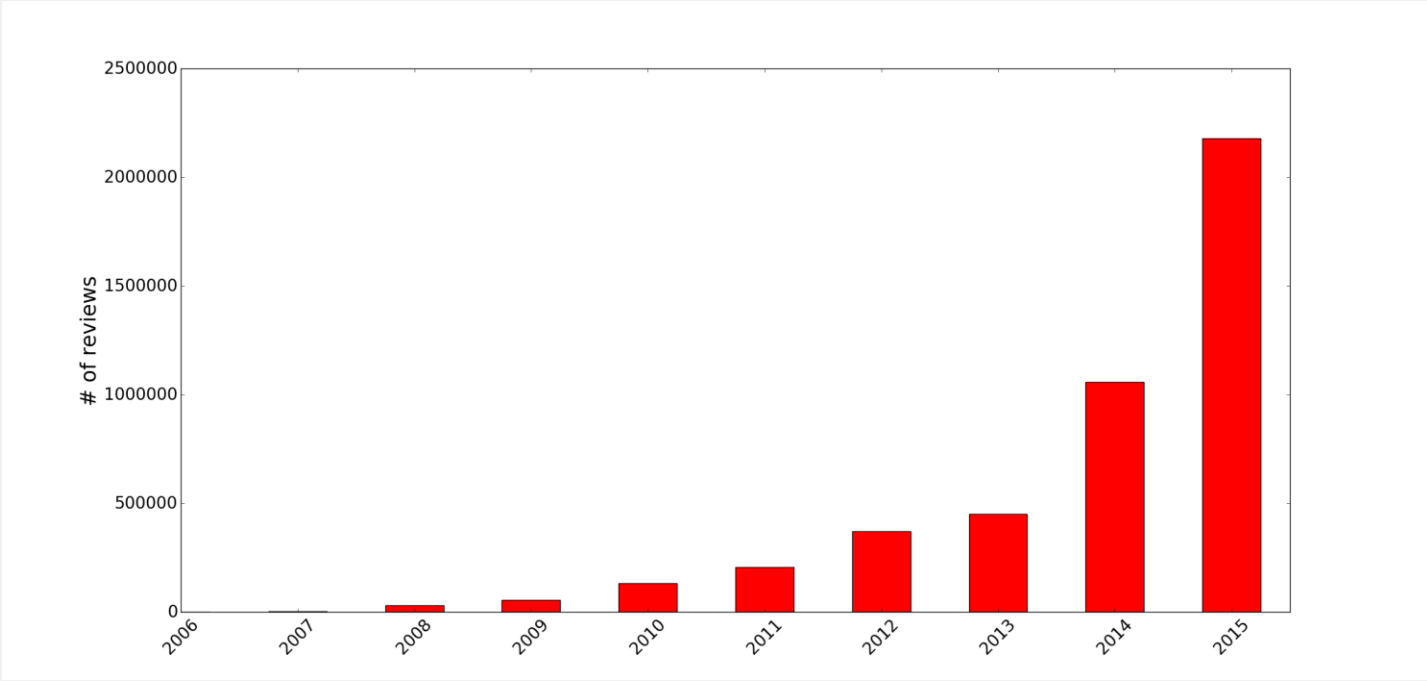




Rolando P. Hong Enriquez

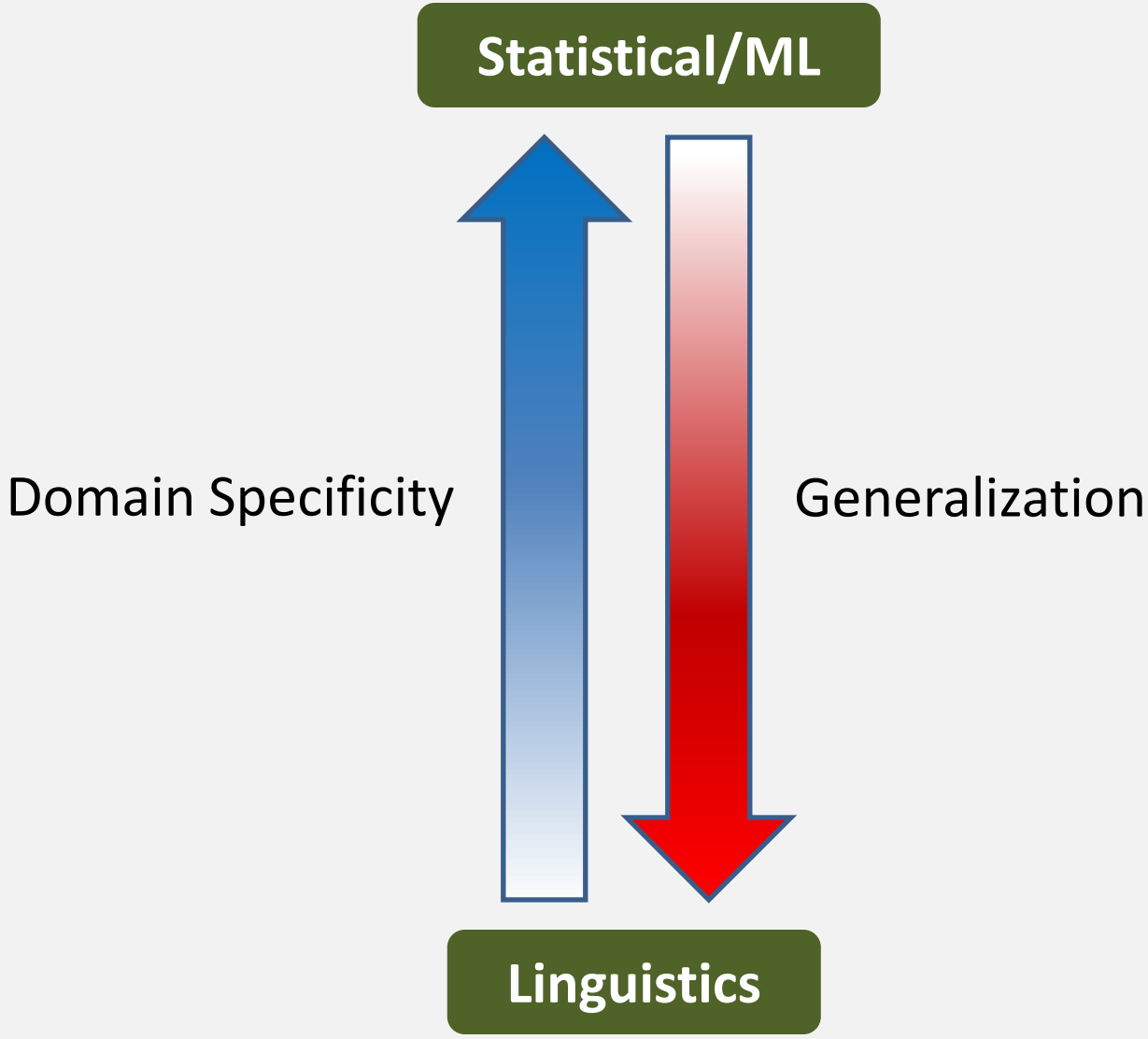
How can we efficiently analyze the sentiment in these data?





Rolando P. Hong Enriquez

Two Approaches



Two Approaches

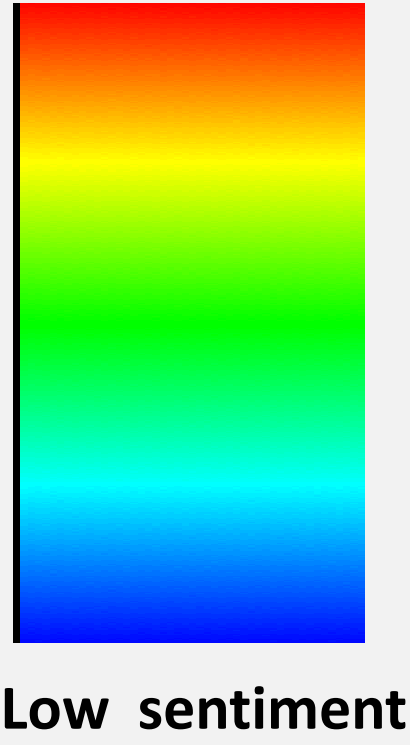
Statistical/ML

+

Linguistics

Goal

['word1' 'word2' 'word3']



THE PIPELINE

["Wheeeen I am alone, a NormallY enjoy a good pizza!! ☺"]

Rolando P. Hong Enriquez

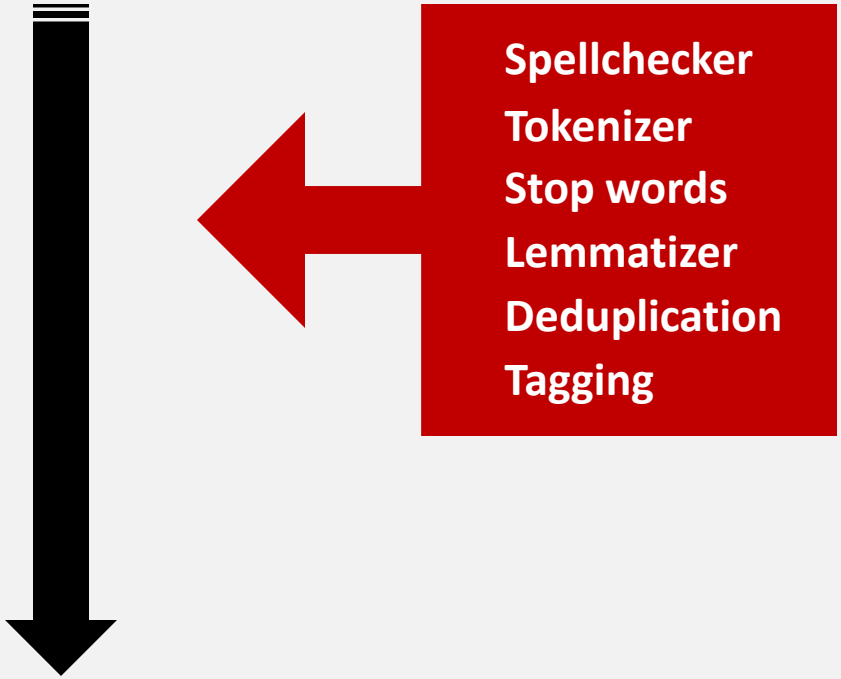
THE PIPELINE

["Wheeeen I am alone, a NormallY enjoy a good pizza!! ☺"]

Rolando P. Hong Enriquez

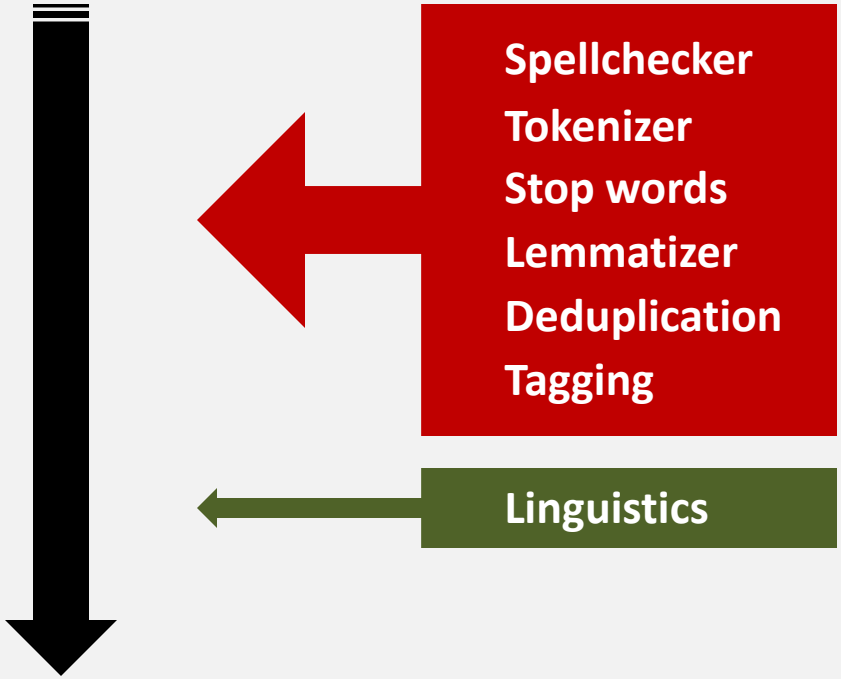
THE PIPELINE

["Wheeeen I am alone, a NormallY enjoy a good pizza!! ☺"]



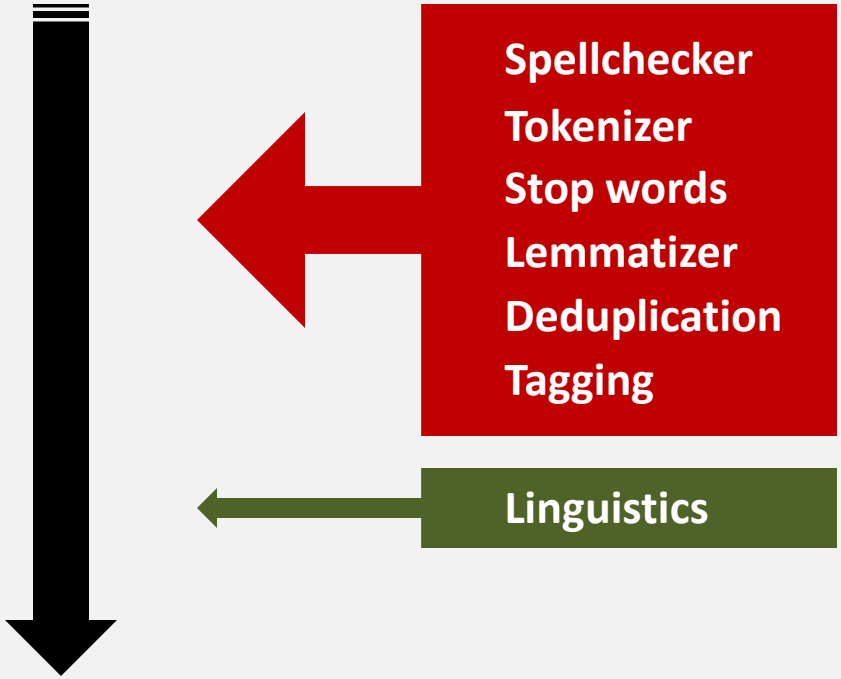
THE PIPELINE

["Wheeeen I am alone, a NormallY enjoy a good pizza!! ☺"]



THE PIPELINE

["Wheeeen I am alone, a NormallY enjoy a good pizza!! ☺"]



['alone' 'enjoy' 'good']

THE PIPELINE

["She runs a lot!!
When I run that much
I normally need to enjoy a good pizza alone in the evening."]

THE PIPELINE

["She runs a lot!!
When I run that much
I normally need to enjoy a good pizza alone in the evening."]

Tokenize

[“She runs a lot!!
When I run that much
I normally need to enjoy a good pizza alone in the evening.”]

Tokenize

[‘She’ ‘runs’ ‘a’ lot’ ‘!’ ‘!’ ‘When’ ‘I’ ‘run’ ‘that’ ‘much’
‘I’ ‘normally’ ‘need’ ‘to’ ‘enjoy’ ‘a’ ‘good’ ‘pizza’
‘alone’ ‘in’ ‘the’ ‘evening’ ’.’]

Remove stop words

['She' 'runs' 'a' lot' '!' '!' 'When' 'I' 'run' 'that' 'much'
'I' 'normally' 'need' 'to' 'enjoy' 'a' 'good' 'pizza'
'alone' 'in' 'the' 'evening' '.']

Remove stop words

['She' 'runs' 'lot' 'I' 'run' 'much' 'I' 'normally'
'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

Lemmatize

['She' 'runs' 'lot' 'I' 'run' 'much' 'I' 'normally'
'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

Rolando P. Hong Enriquez

Lemmatize

['She' 'run' 'lot' 'I' 'run' 'much' 'I' 'normally'
'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

Rolando P. Hong Enriquez

Eliminate duplicates

['She' 'run' 'lot' 'I' 'run' 'much' 'I' 'normally'
'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

Eliminate duplicates

['She' 'run' 'lot' 'much' 'I' 'normally' 'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

Tagging

['She' 'run' 'lot' 'much' 'I' 'normally' 'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

- Verbs & Adverbs
- Adjectives
- Nouns & Personal Pronouns

Rolando P. Hong Enriquez

Eliminate objective parts

['She' 'run' 'lot' 'much' 'I' 'normally' 'need' 'enjoy' 'good' 'pizza' 'alone' 'evening']

- Verbs & Adverbs
- Adjectives
- Nouns & Personal Pronouns

Rolando P. Hong Enriquez

Eliminate objective parts

['run' 'lot' 'much' 'normally' 'need' 'enjoy' 'good' 'alone']

Verbs & Adverbs

Adjectives

Rolando P. Hong Enriquez

Valence Aware Dictionary for sEntiment Reasoning [1]

['run' 'lot' 'much' 'normally' 'need' 'enjoy'
'good' 'alone']

Verbs & Adverbs

Adjectives

Rolando P. Hong Enriquez

[1] https://github.com/cjhutto/vaderSentiment

VADER

['run' 'lot' 'much' 'normally' 'need' 'enjoy' 'good' 'alone']

Verbs & Adverbs

Adjectives

VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

C.J. Hutto **Eric Gilbert**
Georgia Institute of Technology, Atlanta, GA 30032
cjhutto@gatech.edu gilbert@cc.gatech.edu

Rolando P. Hong Enriquez

VADER

['run' 'lot' 'much' 'normally' 'need' 'enjoy' 'good' 'alone']

VADER score ≠ 0

Rolando P. Hong Enriquez

VADER

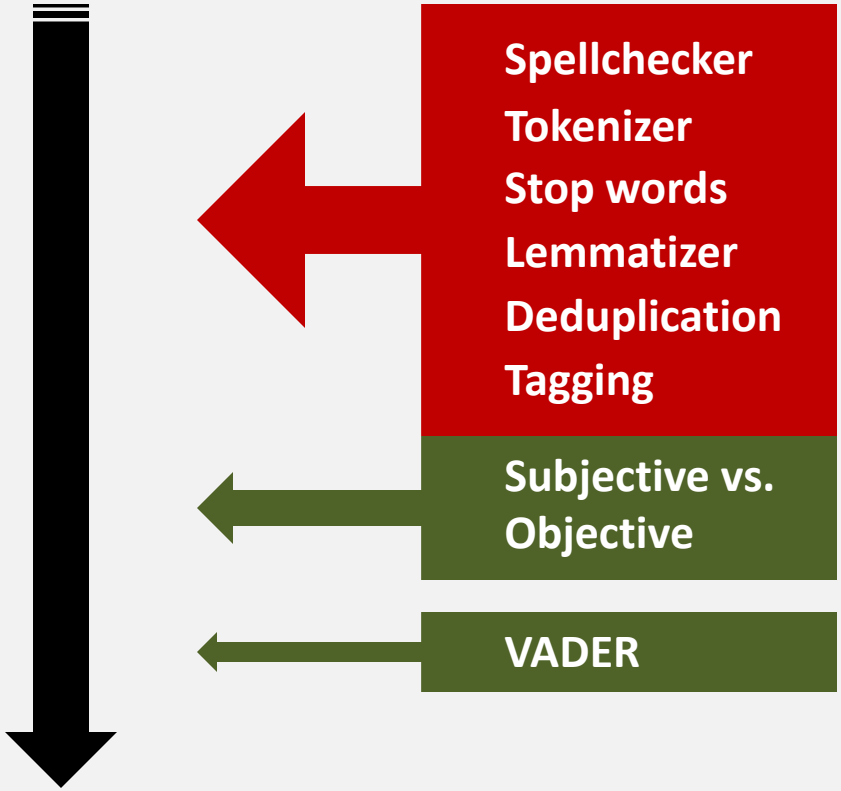
['enjoy' 'good' 'alone']

VADER score ≠ 0

Rolando P. Hong Enriquez

THE PIPELINE

["Wheeeen I am alone, a NormallY enjoy a good pizza!! ☺"]



['alone' 'enjoy' 'good']

Rolando P. Hong Enriquez

Binarize

1 2 3 4 5

[‘play’ ‘funny’ ‘sad’ ‘good’ ‘bad’] **Minimal Dictionary**

Rolando P. Hong Enriquez

Binarize

1 2 3 4 5

[‘play’ ‘funny’ ‘sad’ ‘good’ ‘bad’] **Minimal Dictionary**

‘This food is very good’ **Review**

Rolando P. Hong Enriquez

Binarize

1 2 3 4 5

[‘play’ ‘funny’ ‘sad’ ‘good’ ‘bad’] **Minimal Dictionary**

‘This food is very good’ **Review**

[0 0 0 1 0] **Binarized Review**

Rolando P. Hong Enriquez

Small Intro to ML models

Rolando P. Hong Enriquez

ML models

Discriminative

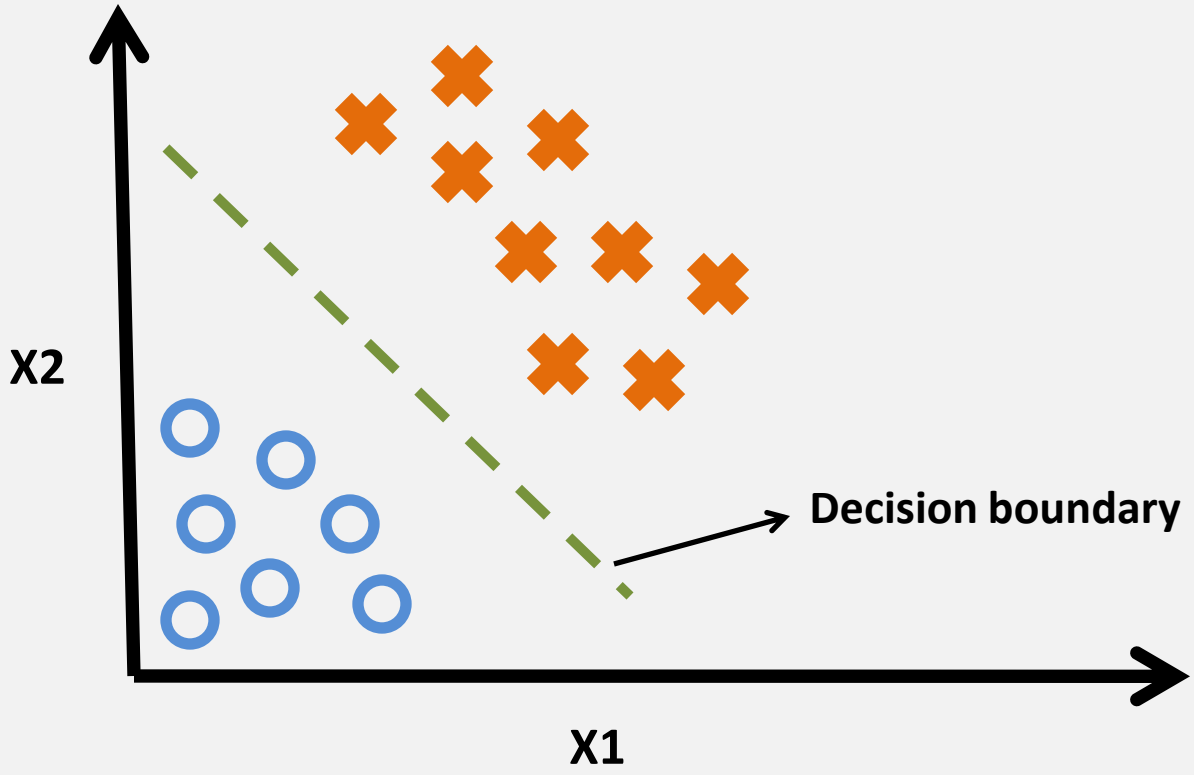
e.g., Linear/Logistic Regression

Generative

e.g., Naïve Bayes

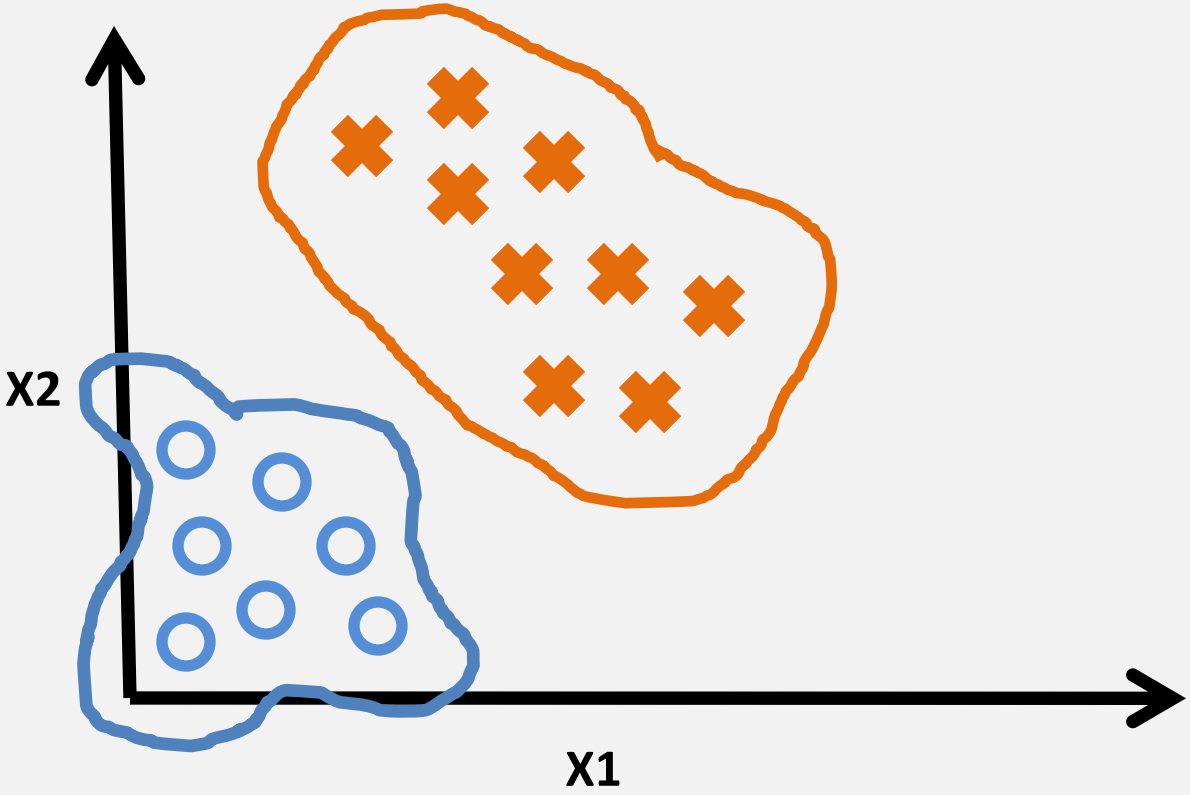
ML models

Discriminative



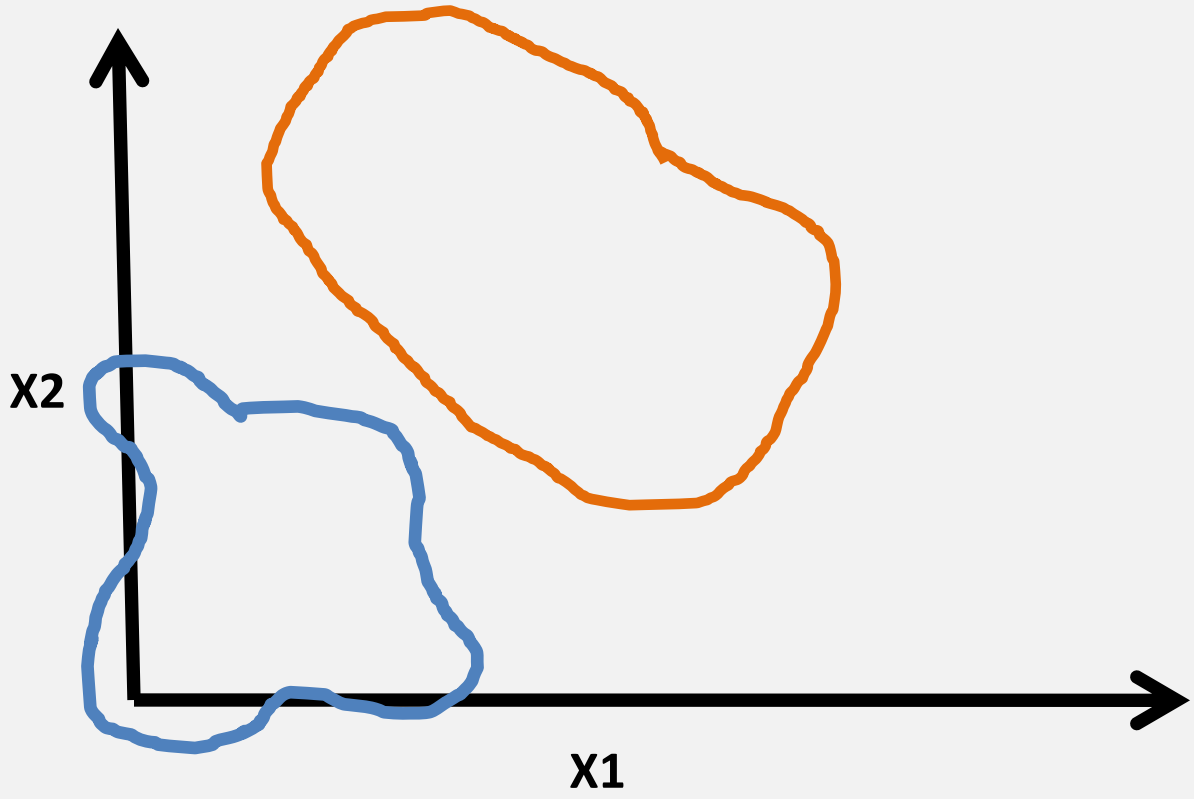
ML models

Generative



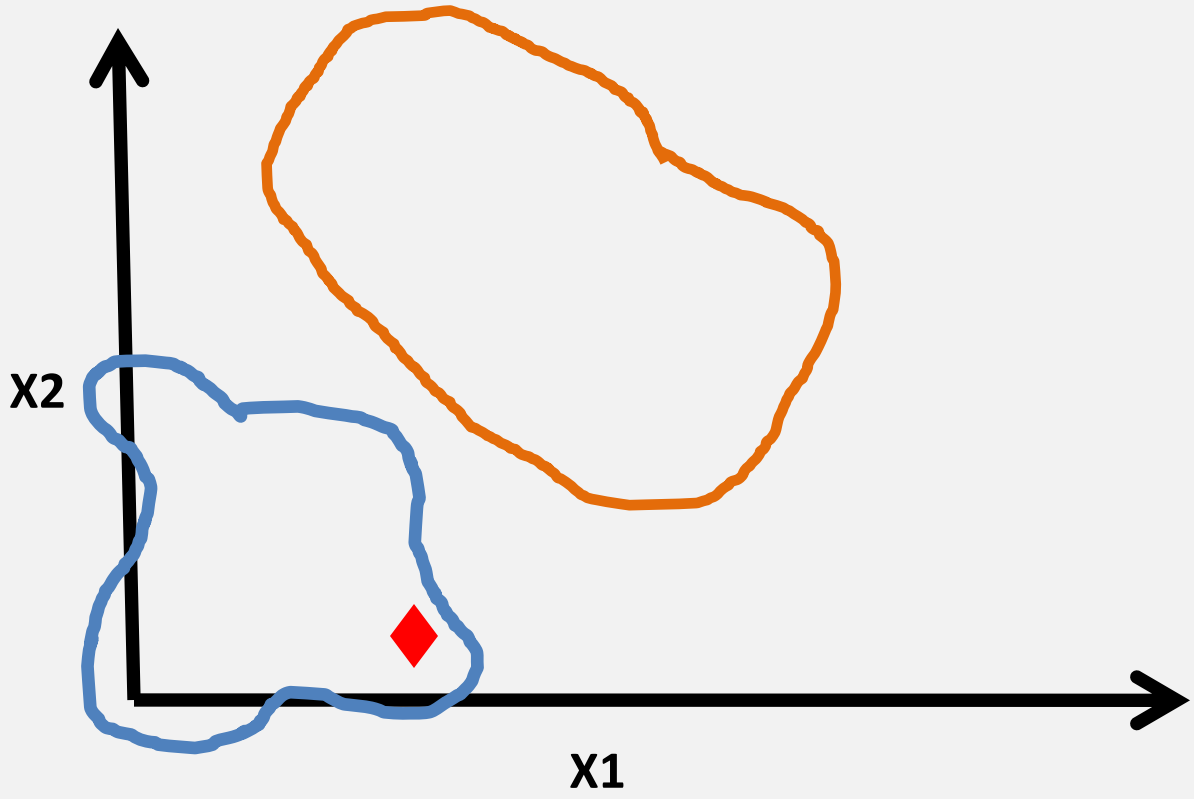
ML models

Generative

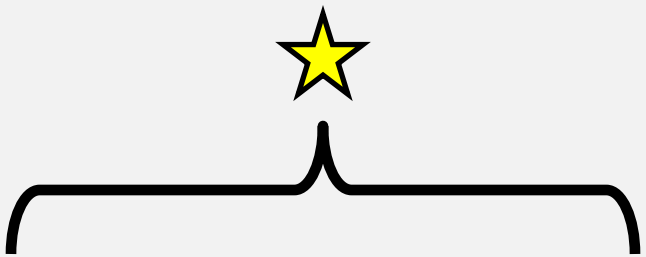


ML models

Generative



Training data



[1 1 0 1 0] **Review 1**

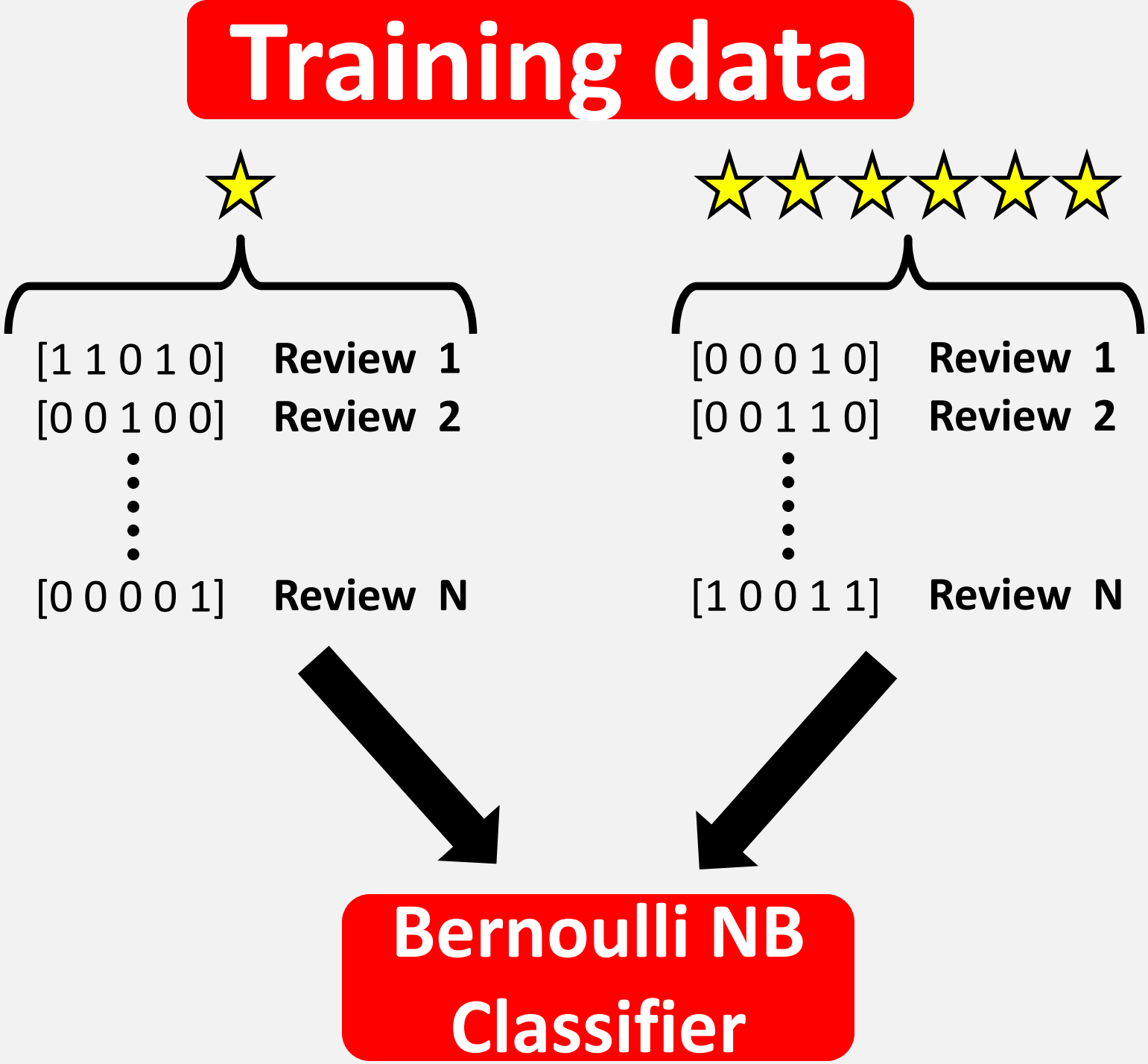
[0 0 1 0 0] **Review 2**

[0 0 0 0 1] **Review N**

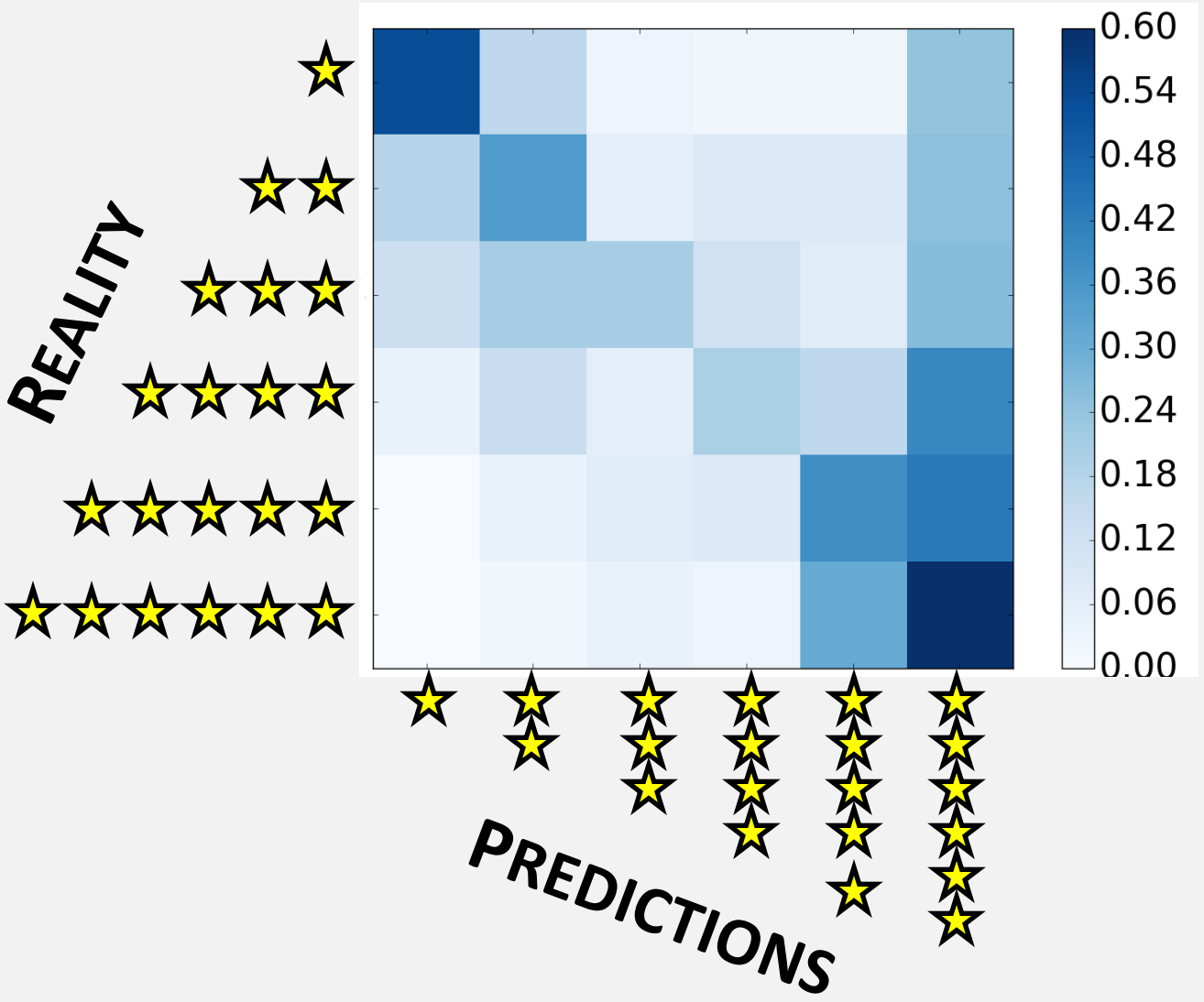
[0 0 0 1 0] **Review 1**

[0 0 1 1 0] **Review 2**

[1 0 0 1 1] **Review N**

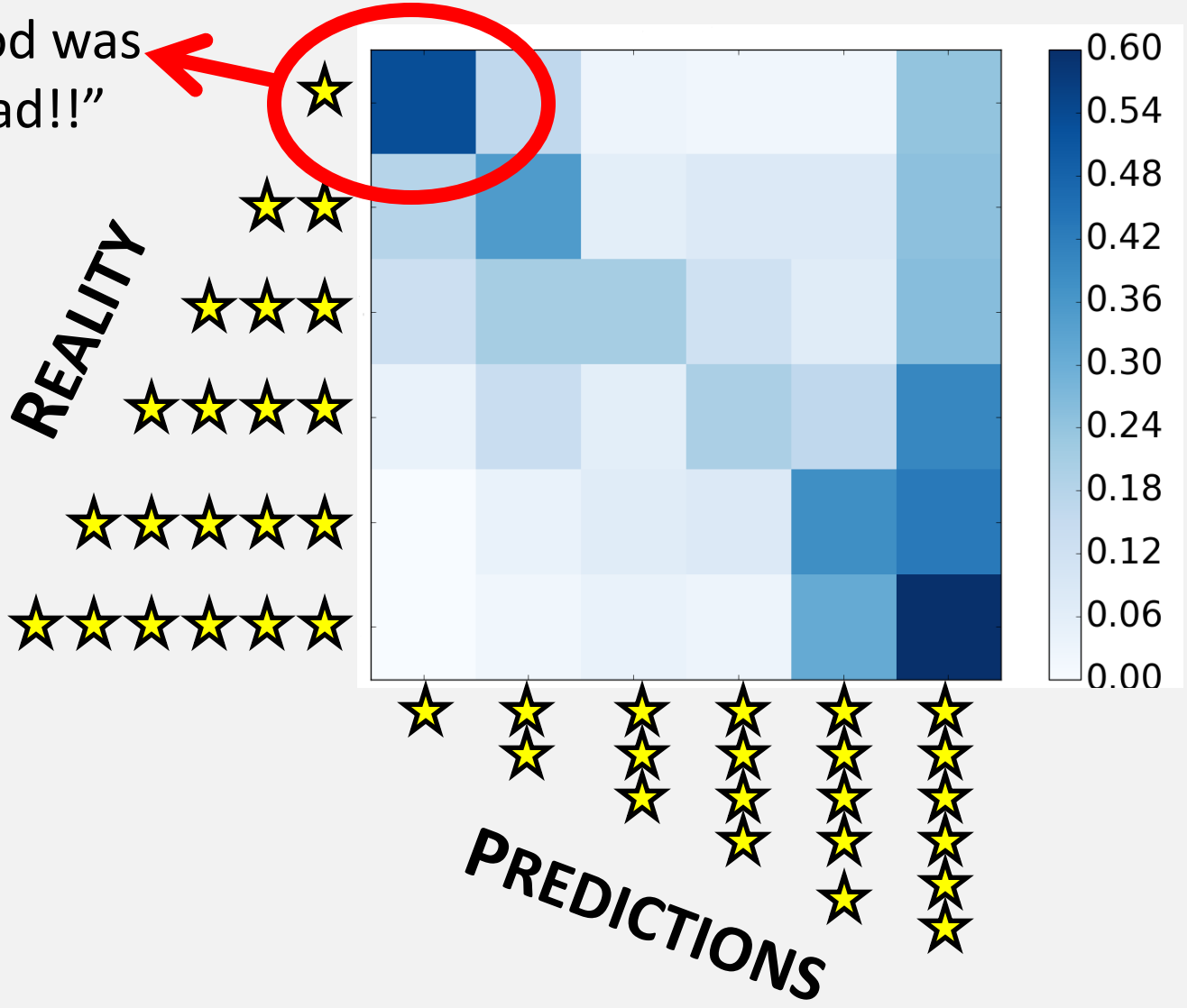

Training data
[1 1 0 1 0] Review 1
[0 0 1 0 0] Review 2
[0 0 0 0 1] Review N
[0 0 0 1 0] Review 1
[0 0 1 1 0] Review 2
[1 0 0 1 1] Review N
Bernoulli NB Classifier

Testing the Classifier

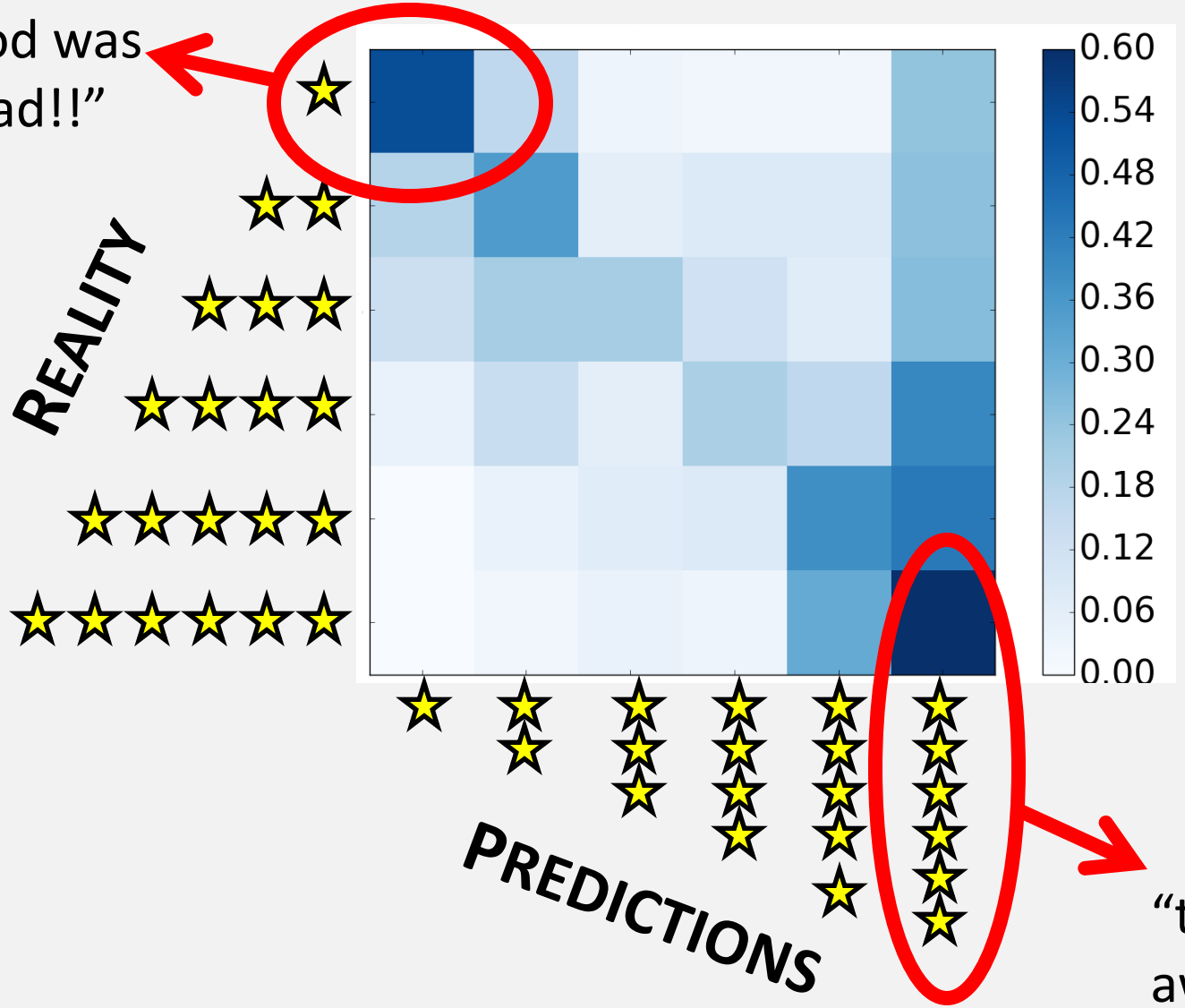


Testing the Classifier

“the food was really bad!!”



Testing the Classifier



Some comparisons

Fang and Zhan *Journal of Big Data* (2015) 2:5
DOI 10.1186/s40537-015-0015-2

Journal of Big Data
a SpringerOpen Journal

METHODOLOGY Open Access

Sentiment analysis using product review data

Xing Fang* and Justin Zhan

Some comparisons

Fang and Zhan *Journal of Big Data* (2015) 2:5
DOI 10.1186/s40537-015-0015-2

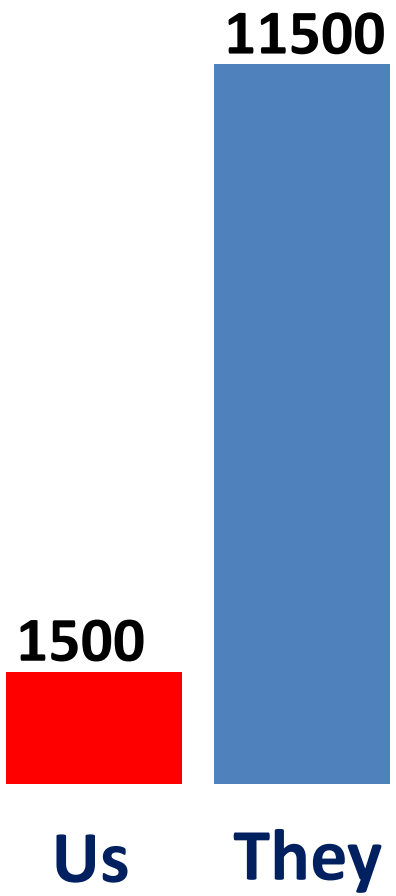
Journal of Big Data
a SpringerOpen Journal

METHODOLOGY Open Access

Sentiment analysis using product review data

Xing Fang* and Justin Zhan

Number of features



Some comparisons

Fang and Zhan *Journal of Big Data* (2015) 2:5
DOI 10.1186/s40537-015-0015-2

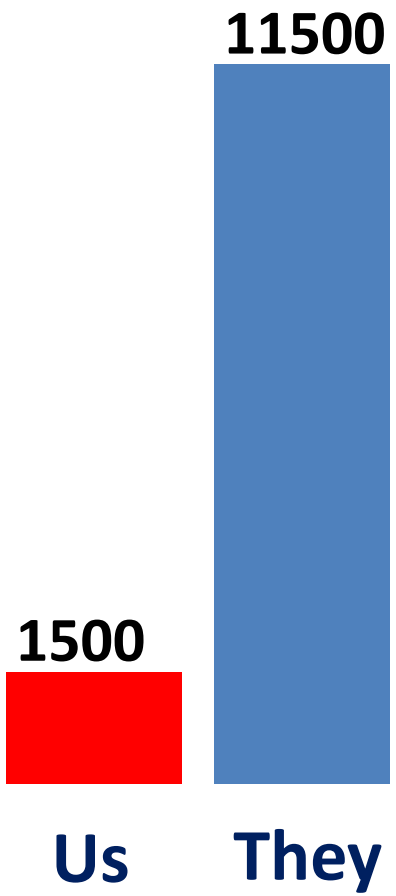
Journal of Big Data
a SpringerOpen Journal

METHODOLOGY Open Access

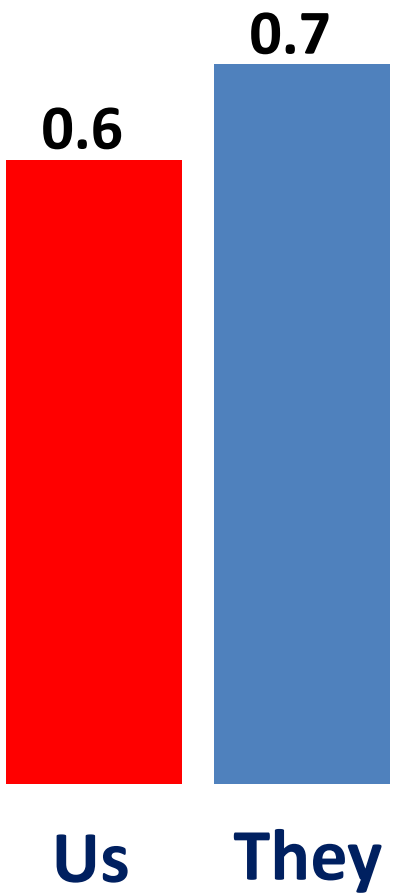
Sentiment analysis using product review data

Xing Fang* and Justin Zhan

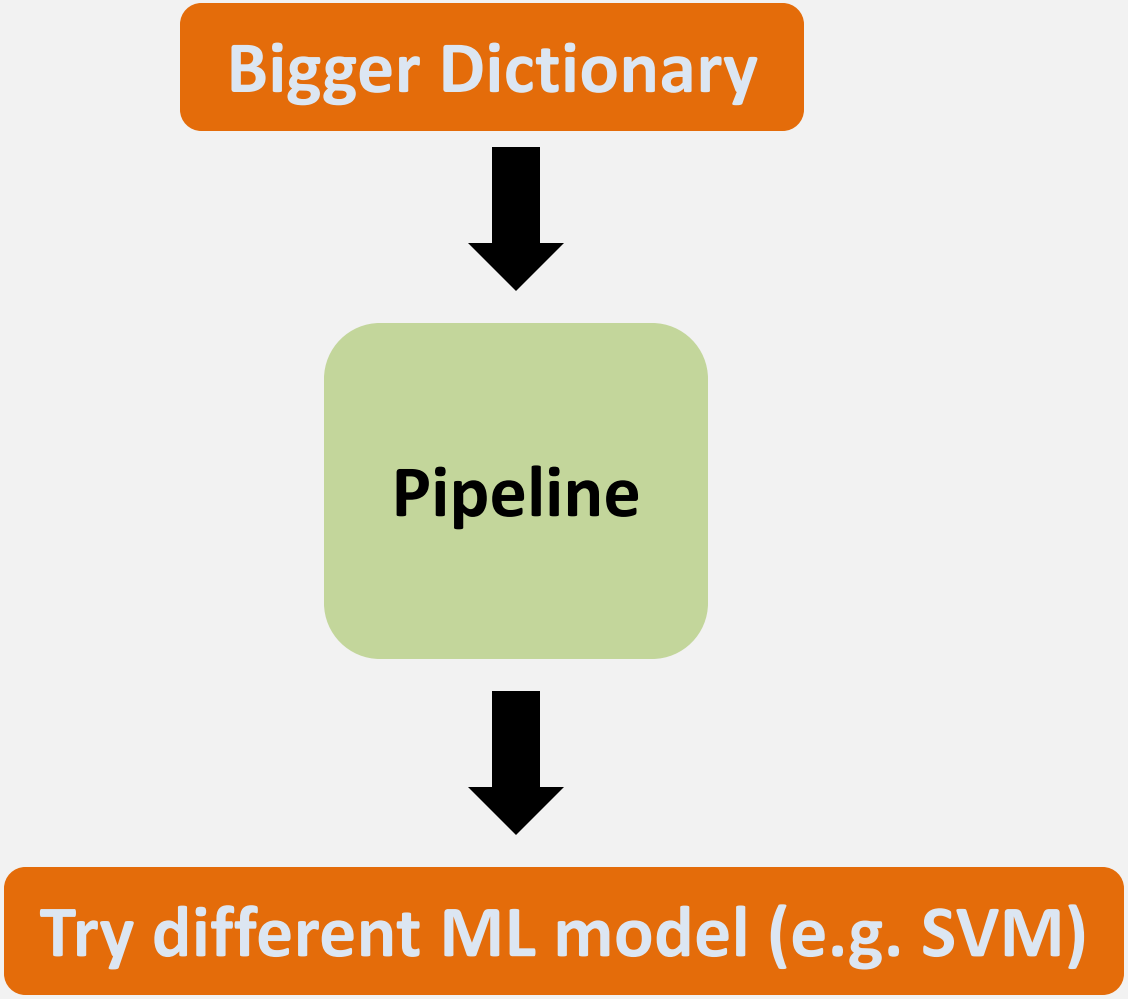
Number of features



F1 scores



Improving



Thank you

Rolando P. Hong Enriquez

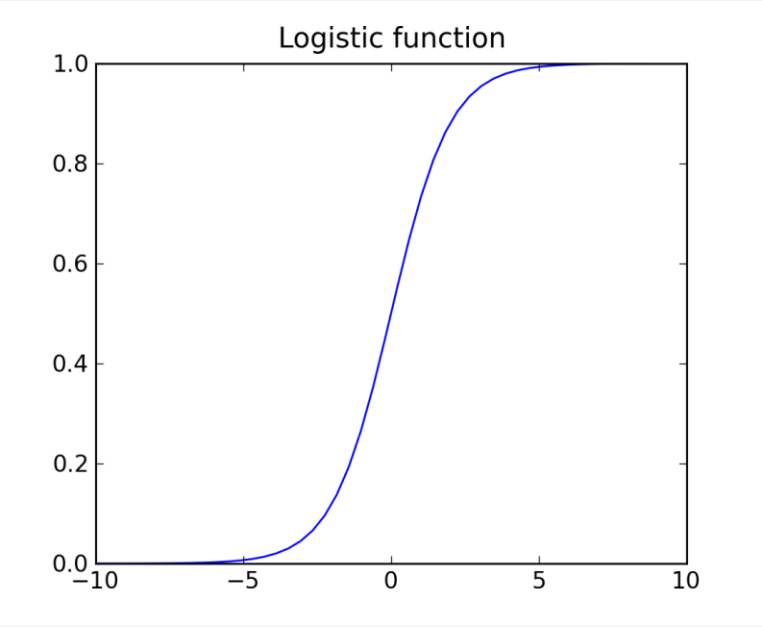
ML models

What does all that means from the perspective of probabilities?

ML models

Discriminative

p(y|x) **Estimates "directly" the probability of y given x**



$$y = \frac{1}{1+e^{x}}$$

$$y = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$$ **Classifying in classes '0' and '1'**

e.g., y = 1 if x > 0

ML models

Generative

p(x|y) **Estimates “directly” the probability of y given x**