NLP with H_2O

H2O.ai

Agenda

- Our Use Case
- H2O Overview
- Natural Language Processing
- Demo

Our Use Case

The Data

The Amazon Fine Food Reviews dataset consists of 568,454 food reviews Amazon users left up to October 2012

- J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews

Column	Example
Product ID	B006K2ZZ7K
User ID	A1UQRSCLF8GW1T
Helpfulness Numerator	1
Helpfulness Denominator	1
Score	5
Time	1350777600
Summary	Great taffy
Text	*“Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.”*

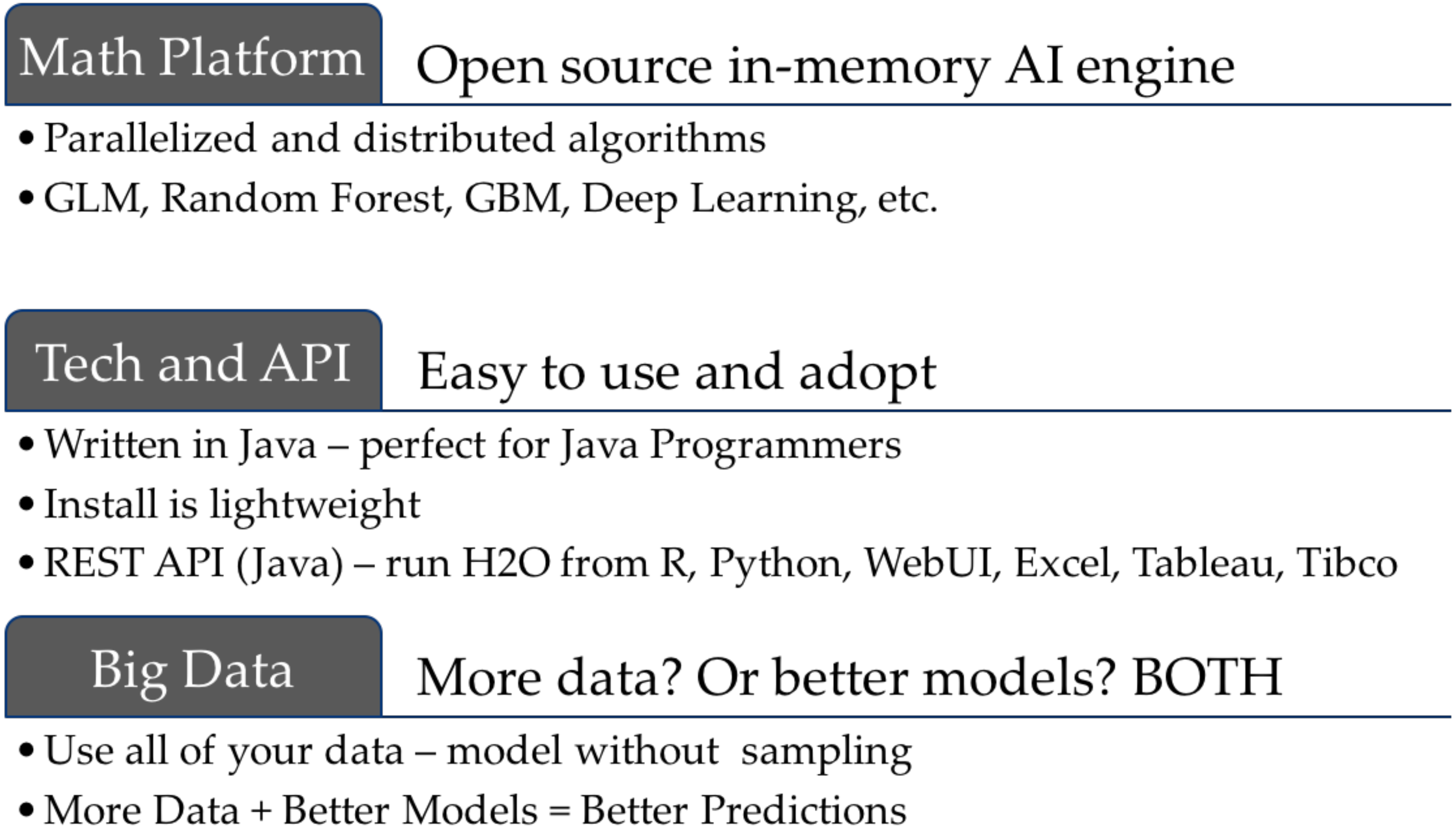
Goal

- Predict whether a food product has a good rating based on the reviews

“Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.” → ★ ★ ★ ★ ★

H2O Overview

What is H2O?



Math Platform Open source in-memory AI engine

- Parallelized and distributed algorithms
- GLM, Random Forest, GBM, Deep Learning, etc.

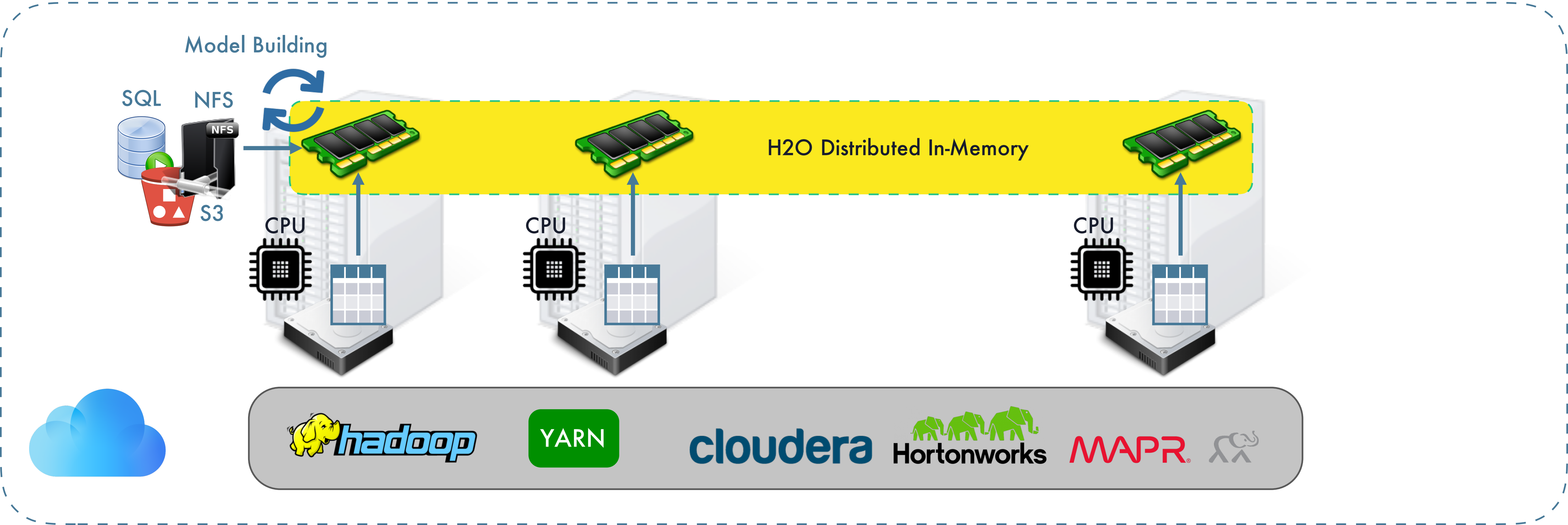
Tech and API Easy to use and adopt

- Written in Java – perfect for Java Programmers
- Install is lightweight
- REST API (Java) – run H2O from R, Python, WebUI, Excel, Tableau, Tibco

Big Data More data? Or better models? BOTH

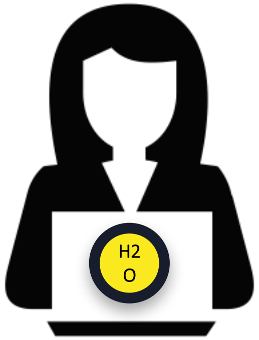
- Use all of your data – model without sampling
- More Data + Better Models = Better Predictions

H2O Cluster

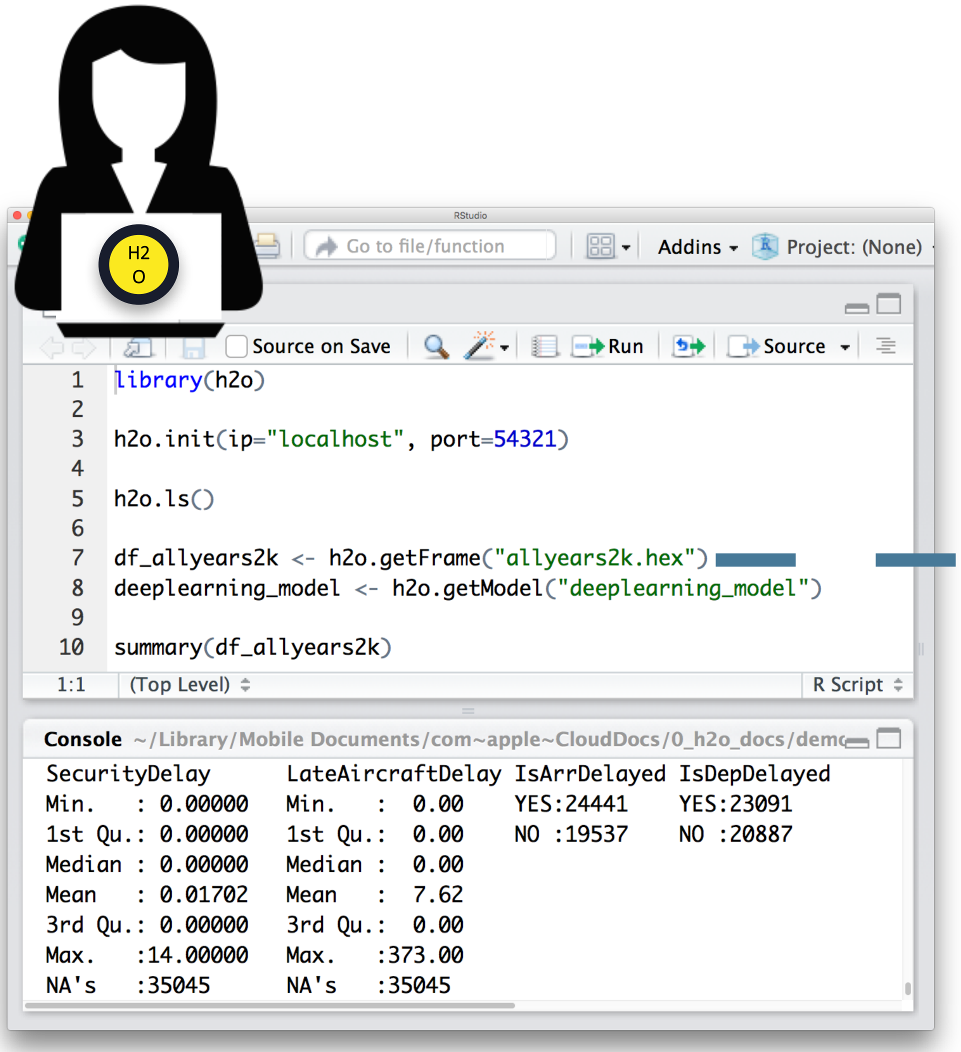

Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

H2O Cluster

H2O Clients

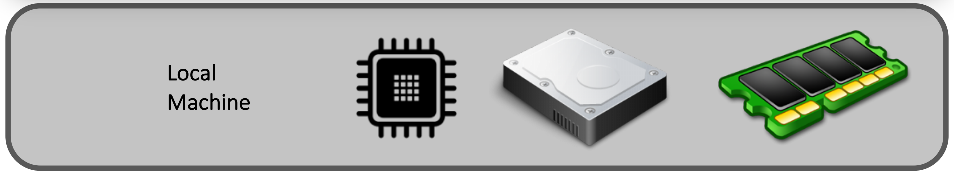


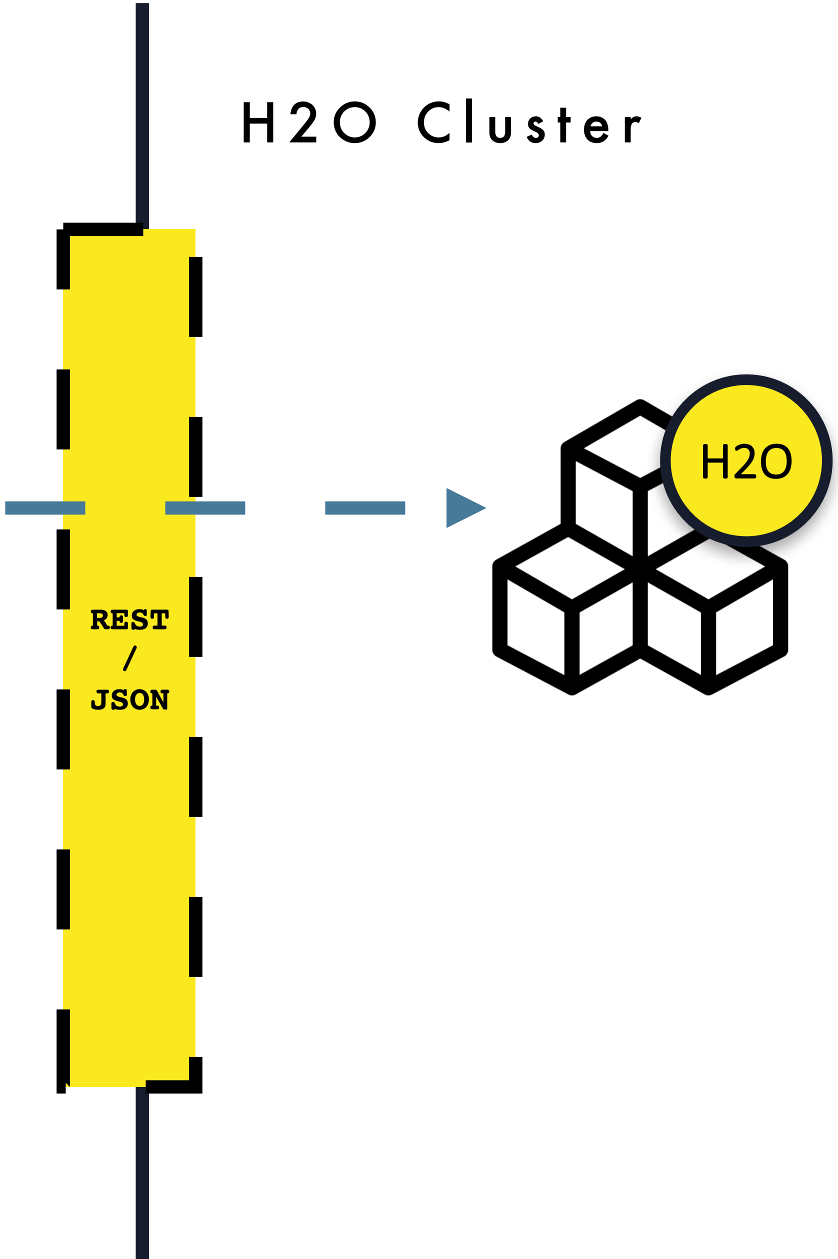
H2O Cluster



```r
library(h2o)

h2o.init(ip="localhost", port=54321)

h2o.ls()

df_allyears2k <- h2o.getFrame("allyears2k.hex")
deeplearning_model <- h2o.getModel("deeplearning_model")

summary(df_allyears2k)
```

```
SecurityDelay      LateAircraftDelay IsArrDelayed IsDepDelayed
Min.   : 0.00000   Min.   :  0.00    YES:24441    YES:23091
1st Qu.: 0.00000   1st Qu.:  0.00    NO :19537    NO :20887
Median : 0.00000   Median :  0.00
Mean   : 0.01702   Mean   :  7.62
3rd Qu.: 0.00000   3rd Qu.:  0.00
Max.   :14.00000   Max.   :373.00
NA's   :35045      NA's   :35045
```





Natural Language Processing

What is NLP?

- A way for computers to understand text and language

- Used For:
  - Sentiment Analysis
  - Topic Identification
  - **Improving Supervised Learning Models**

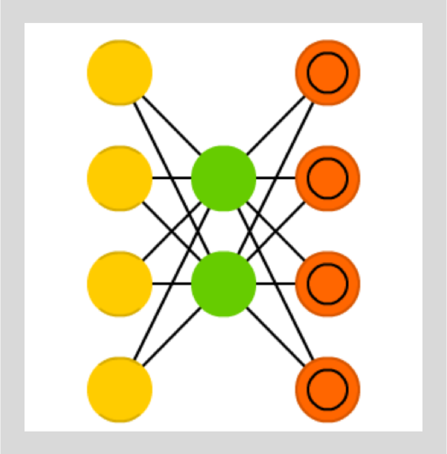
Methods for Natural Language Processing

Bag of Words

- does the word exist in the document?

Count Based

- how often do words occur in each document?
- TF-IDF



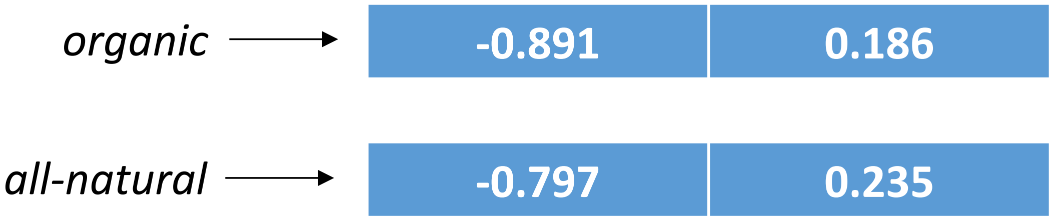
Predictive

- train models predicting a word or sentence from its neighbors
- Word2Vec, RNN

Word Embeddings

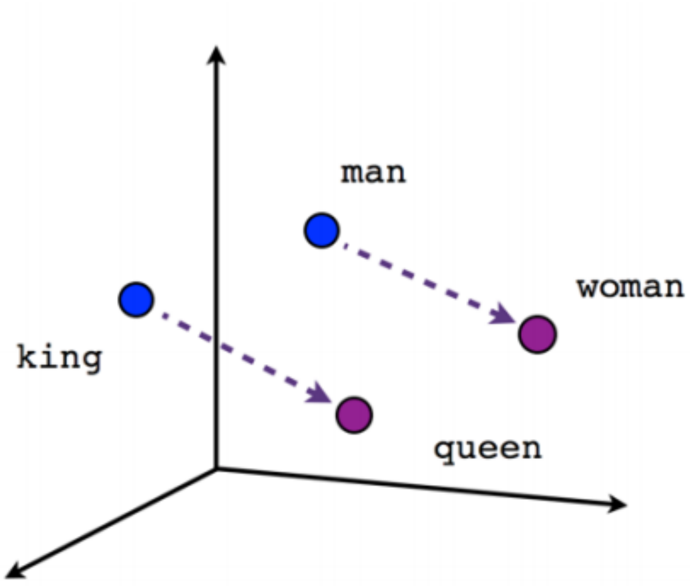
- What?
  - Mapping of words to vectors from a high dimensional space (100 – 1000)

- Why?
  - Embeddings capture the meaning of the word
  - Semantically similar words are close to each other

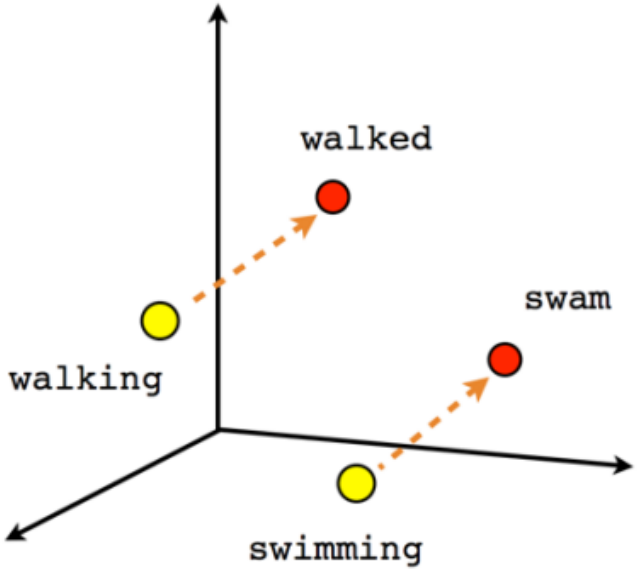


organic →	-0.891	0.186
all-natural →	-0.797	0.235

Word Embeddings



Male-Female



Verb tense

https://www.tensorflow.org/tutorials/word2vec

Word2Vec Algorithm

How do we use a neural network to capture the semantic meaning of words?

- Frame the problem as a supervised learning problem
  - Given an input word predict the neighboring words

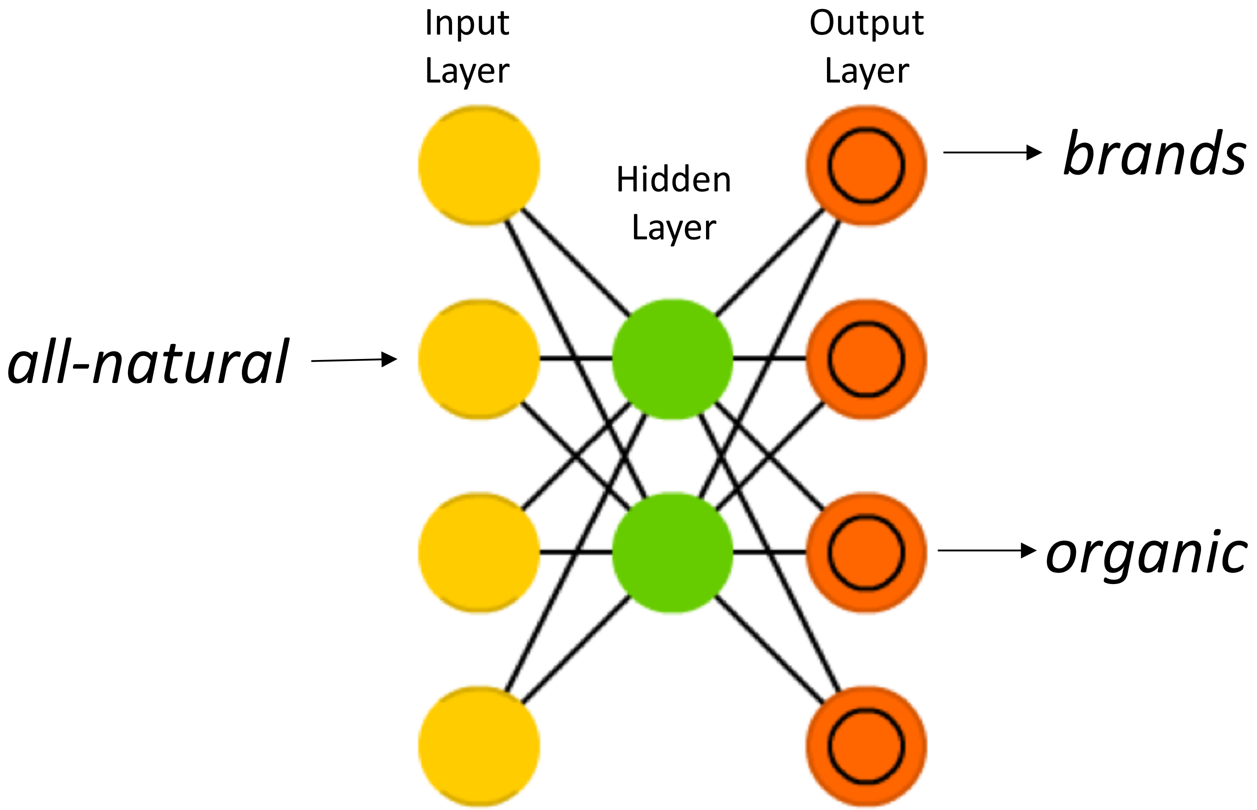
“It's even better than the ***organic, all-natural brands*** *I have tried.”*

Given: *all-natural*

Predict: *organic, brands*

Word2Vec Algorithm

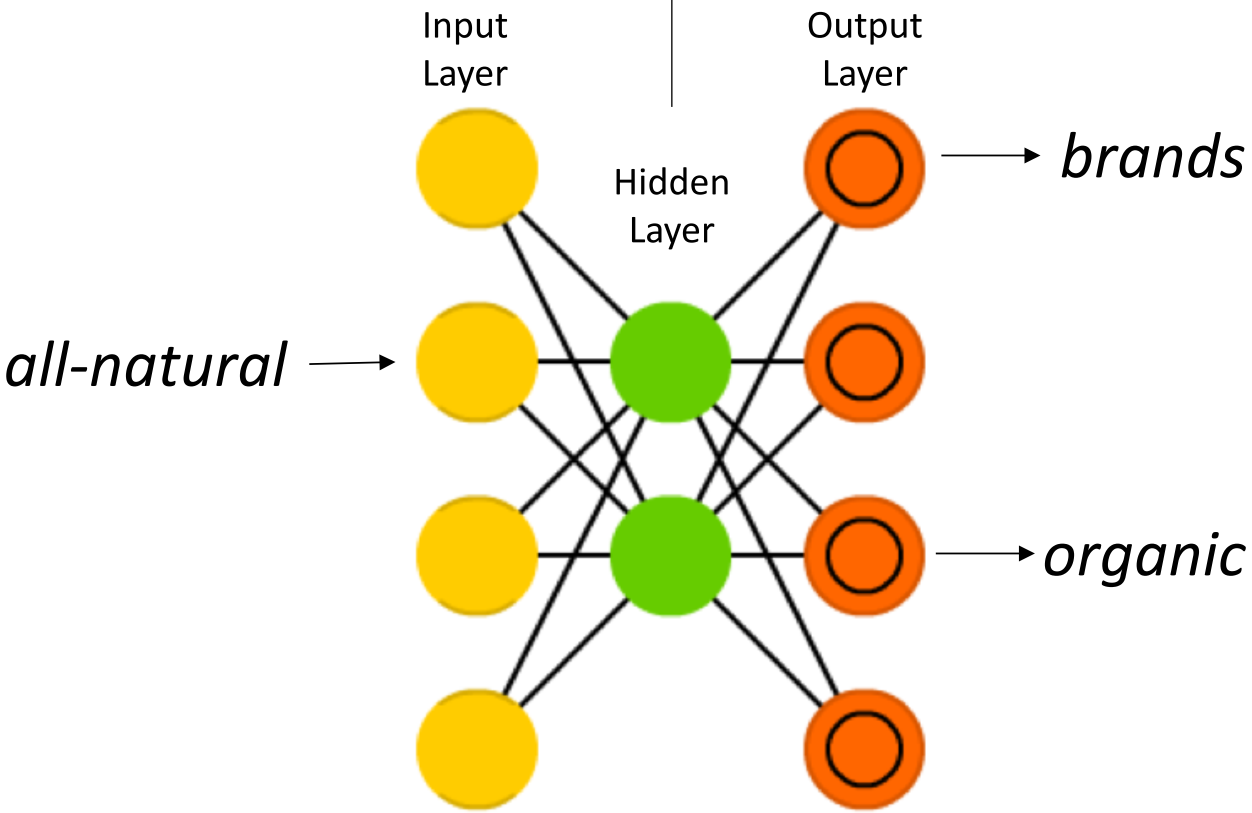
*“It's even better than the **organic, all-natural brands** I have tried.”*



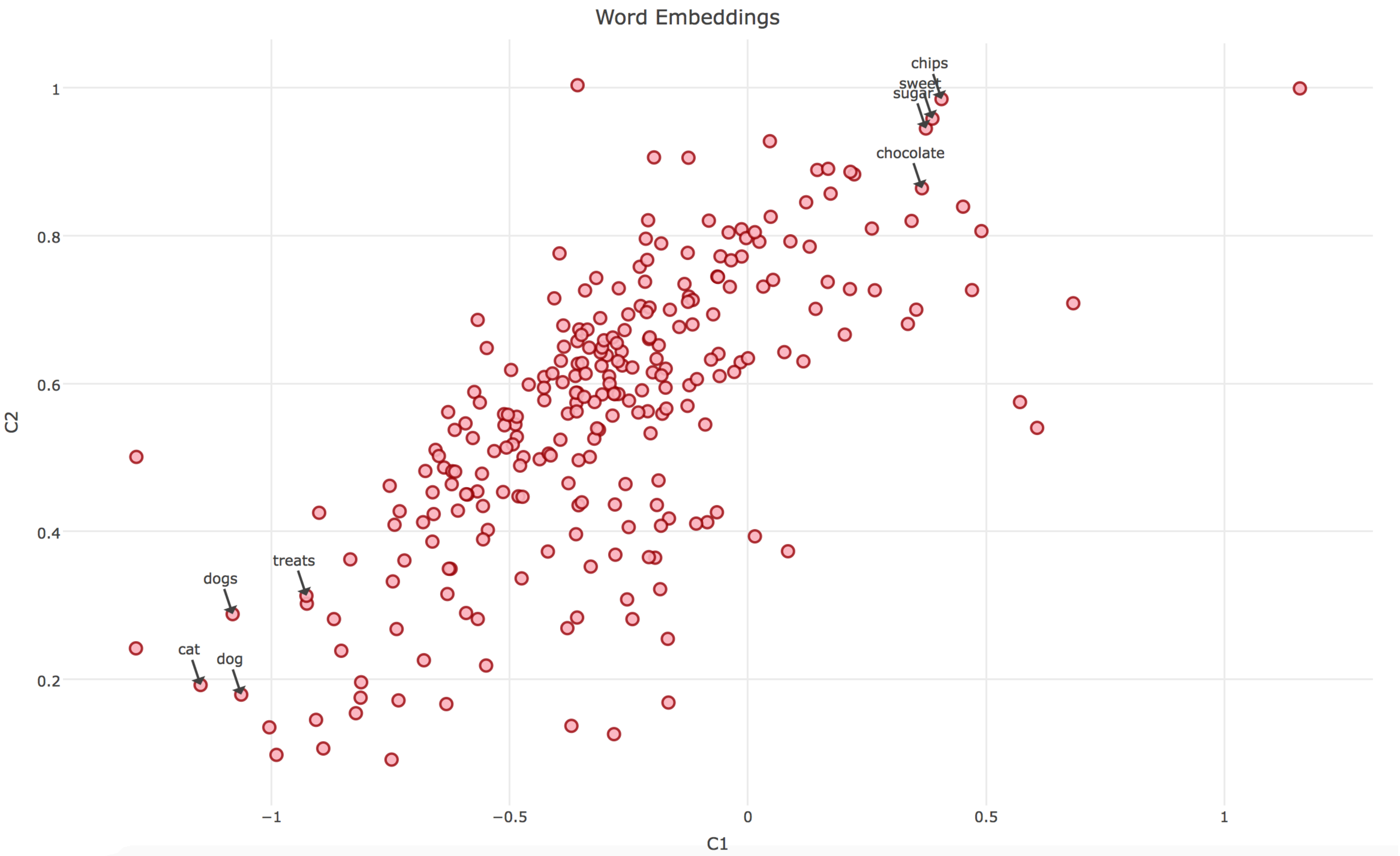
Word2Vec Algorithm

Matrix from Hidden Layer =
Word Embeddings

Word	C1	C2
brands	0.647	0.235
all-natural	-0.797	0.235
organic	-0.891	0.186
tried	-0.751	0.409



Word Embeddings



Word2Vec Usage

- Word2Vec Embeddings are typically used as a pre-processing step to a supervised learning algorithm

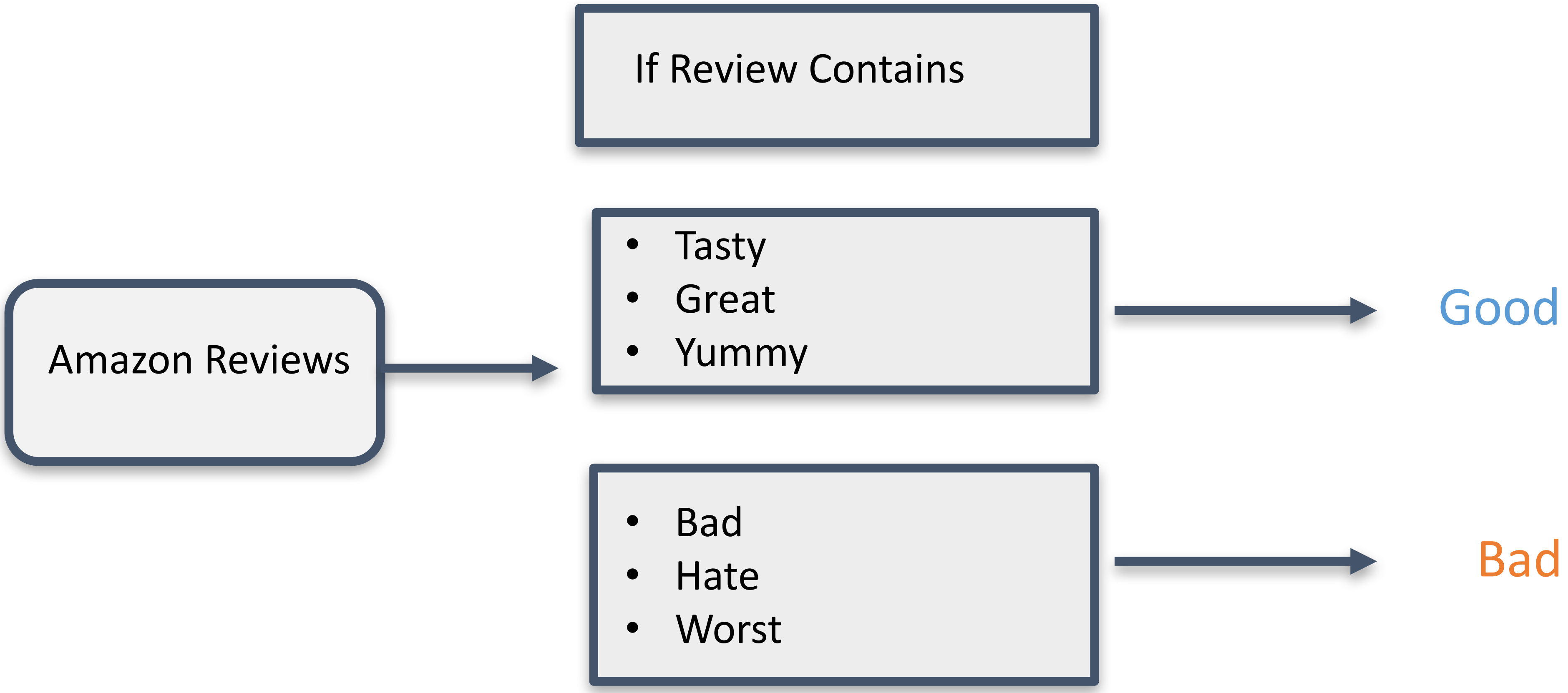
“It's even better than the ***organic, all-natural brands*** *I have tried.”*

Word	C1	C2
even	0.602	-0.499
better	0.765	-0.437
brands	0.647	0.235
all-natural	-0.797	0.235
organic	-0.891	0.186
tried	-0.751	0.409

Aggregate

C1	C2
-0.071	0.0215

Rule Based Model



If Review Contains

Amazon Reviews

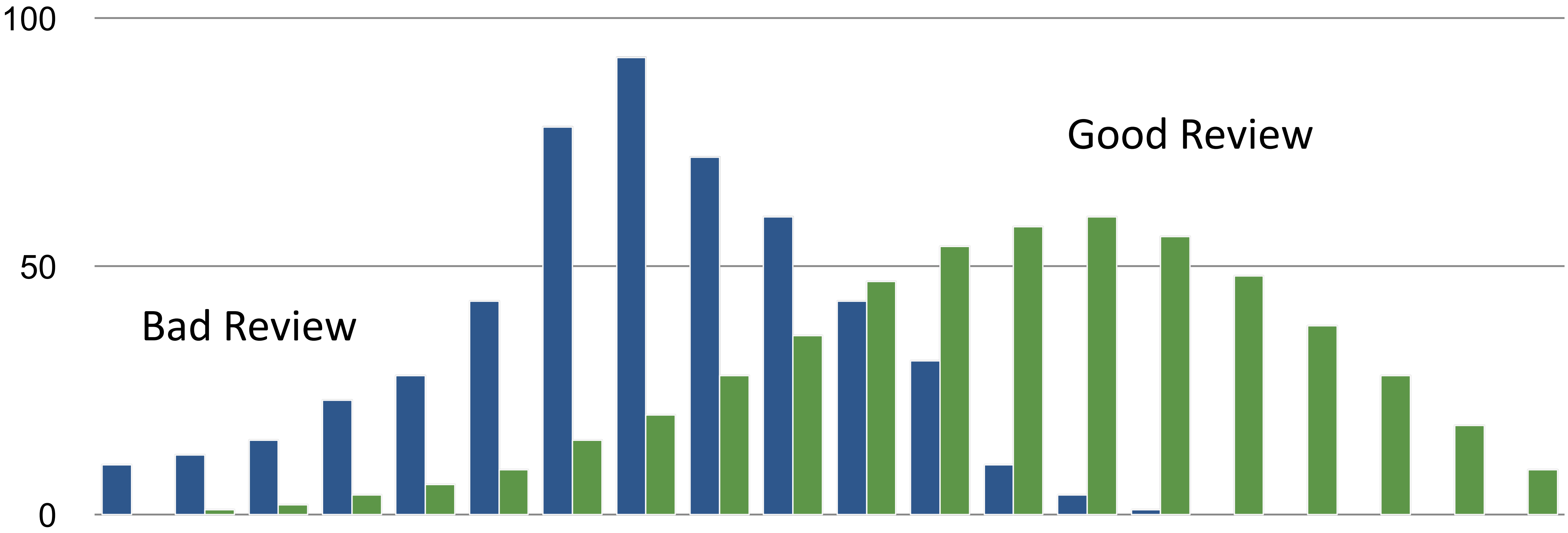
- Tasty
- Great
- Yummy

Good

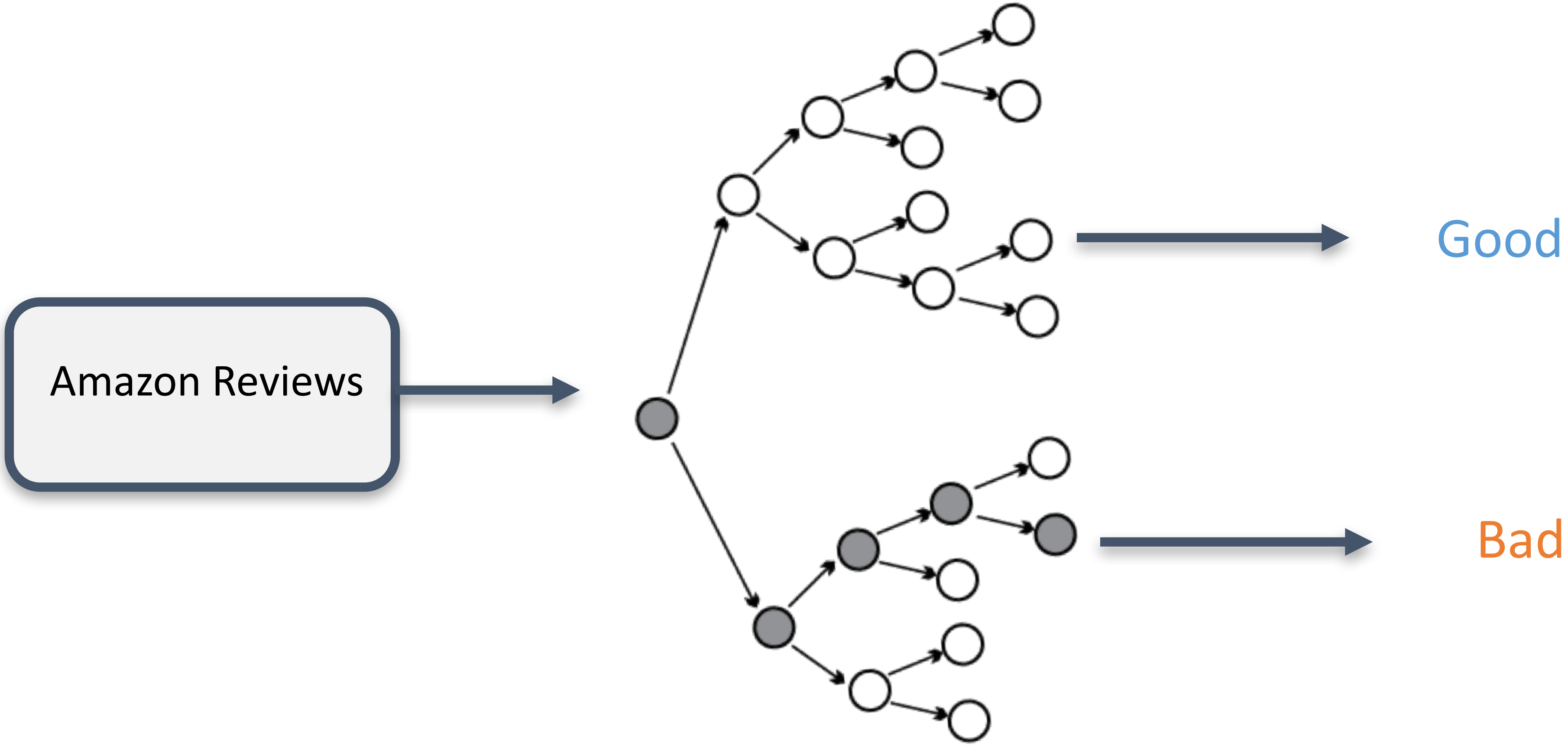
- Bad
- Hate
- Worst

Bad

Machine Learning Model



Machine Learning Model



Amazon Reviews

Good

Bad

Demo

Workflow

Use Case: Predict whether a food product has a good rating based on the review.

1. Tokenize Reviews
   - Break up reviews into separate words
   - Filter words: remove stop words like “the” and “if”
2. Train a Word2Vec Model
3. Use model to transform reviews to vectors
4. Train a supervised learning model to predict good rating

End