SIOP Machine Learning Competition:
Team Procrustination

Presenters: Feng Guo and Nicholas Howald

Team members: Feng Guo, Nicholas Howald,

Marie Childers, Jordan Dovel., Sami Nesnidol, Andrew Samo, & Samuel T. McAbee

April 5th, 2019

BGSU
Bowling Green State University

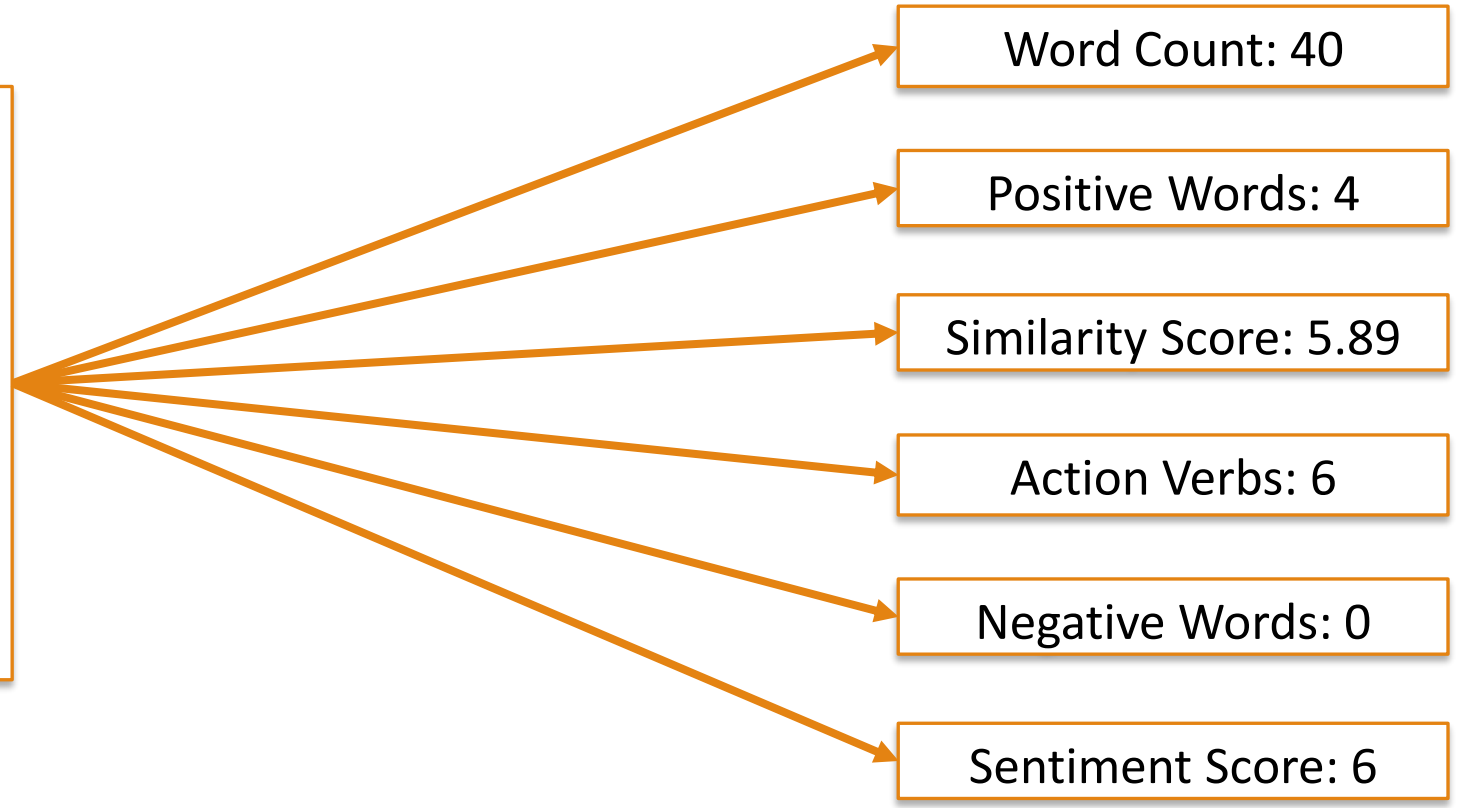
Initial Approaches

- Team members initially approached problem from multiple perspectives
- Used standard practices for cleaning text data (with minor differences)

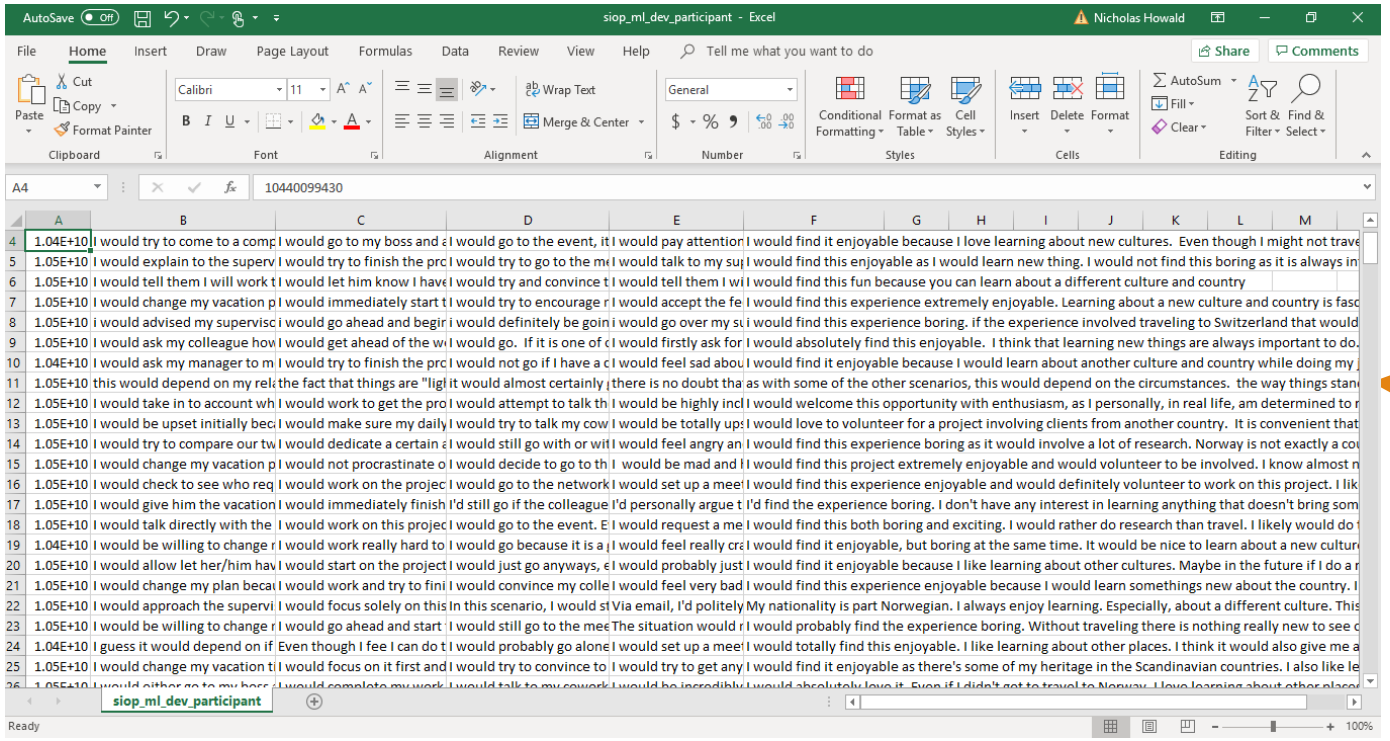
Theory-driven Approach	Data-driven Approach
Features based on structure of individual responses	Features based on patterns across entire text corpus
Features were primarily some type of word count	Features were primarily counts of combinations of words, and hidden pattern/layer from neural network model
Models typically included 20-70 features	Models typically included 1,000+ features
Primarily used ridge regression	Primarily used ridge regression

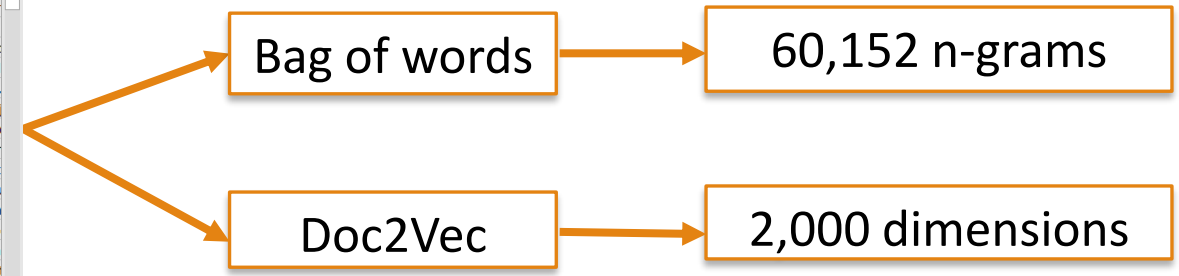
Theory-Driven Features

"I would find it enjoyable because I love learning about new cultures. Even though I might not travel I still appreciate being taught about cultures other than my own. It would also be cool to meet someone from another culture. "

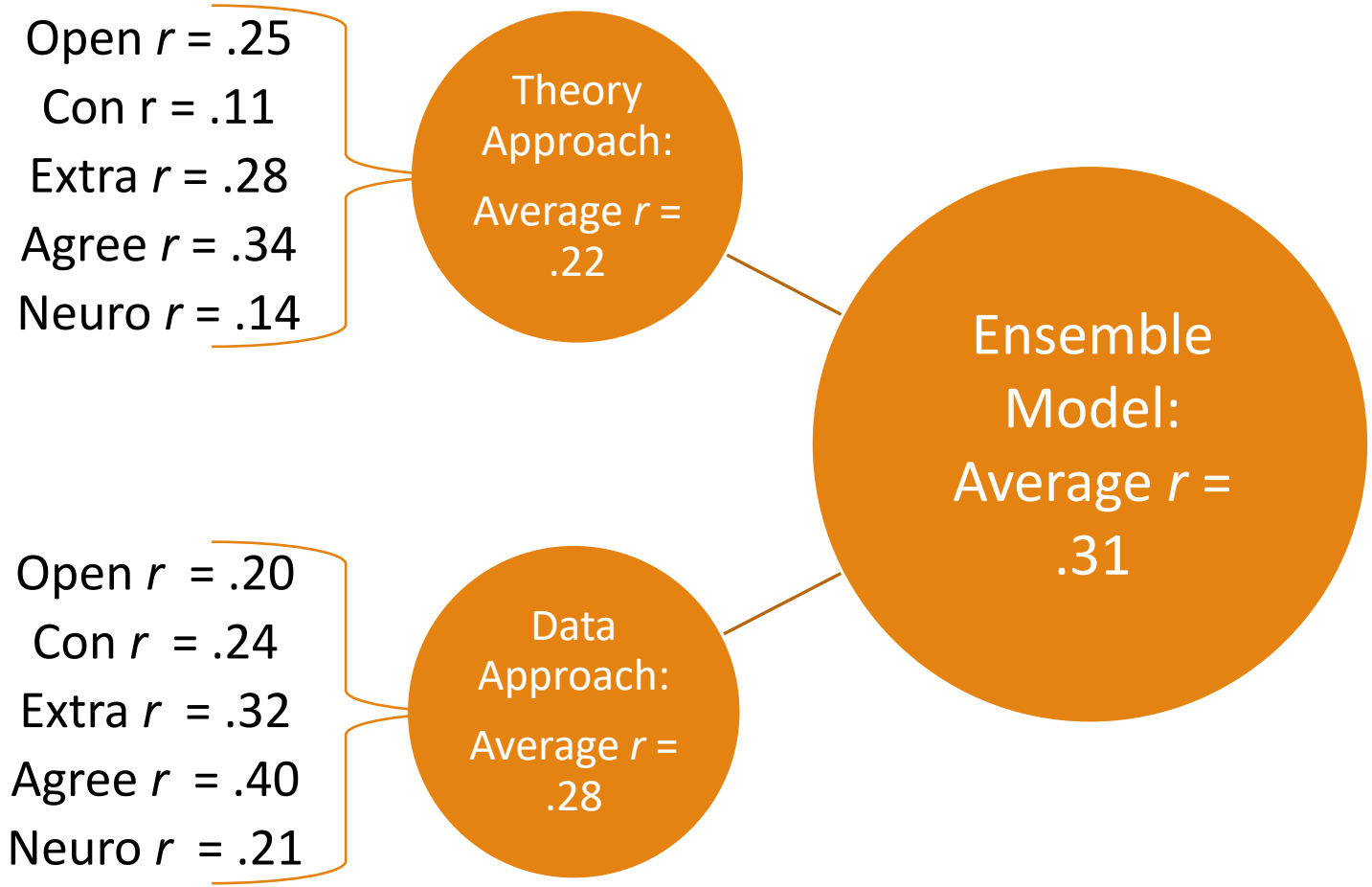


Data-driven Features





Developmental Data (Public Leaderboard)



Winning Model

Ensemble model (i.e., weighted average on ridge regressions) from

Uni & Bi-gram Bag-of-Words + Doc2Vec

Data Approach:

Average *r* = .25

- Open *r* = .19
- Con *r* = .25
- Extra *r* = .24
- Agree *r* = .32
- Neuro *r* = .24

Bag-of-Words

- All five open-ended responses were combined into one column
- Little text preprocessing involved: keep only English letters and lowercase
- No dimension reduction involved

Syntax: Pipeline(

[('**vect**', CountVectorizer(ngram_range=(1,2))),

('**tfidf**', TfidfTransformer()),

('**ridge**', Ridge(alpha=.4))])

Top Predictors from Bag-of-Words Model

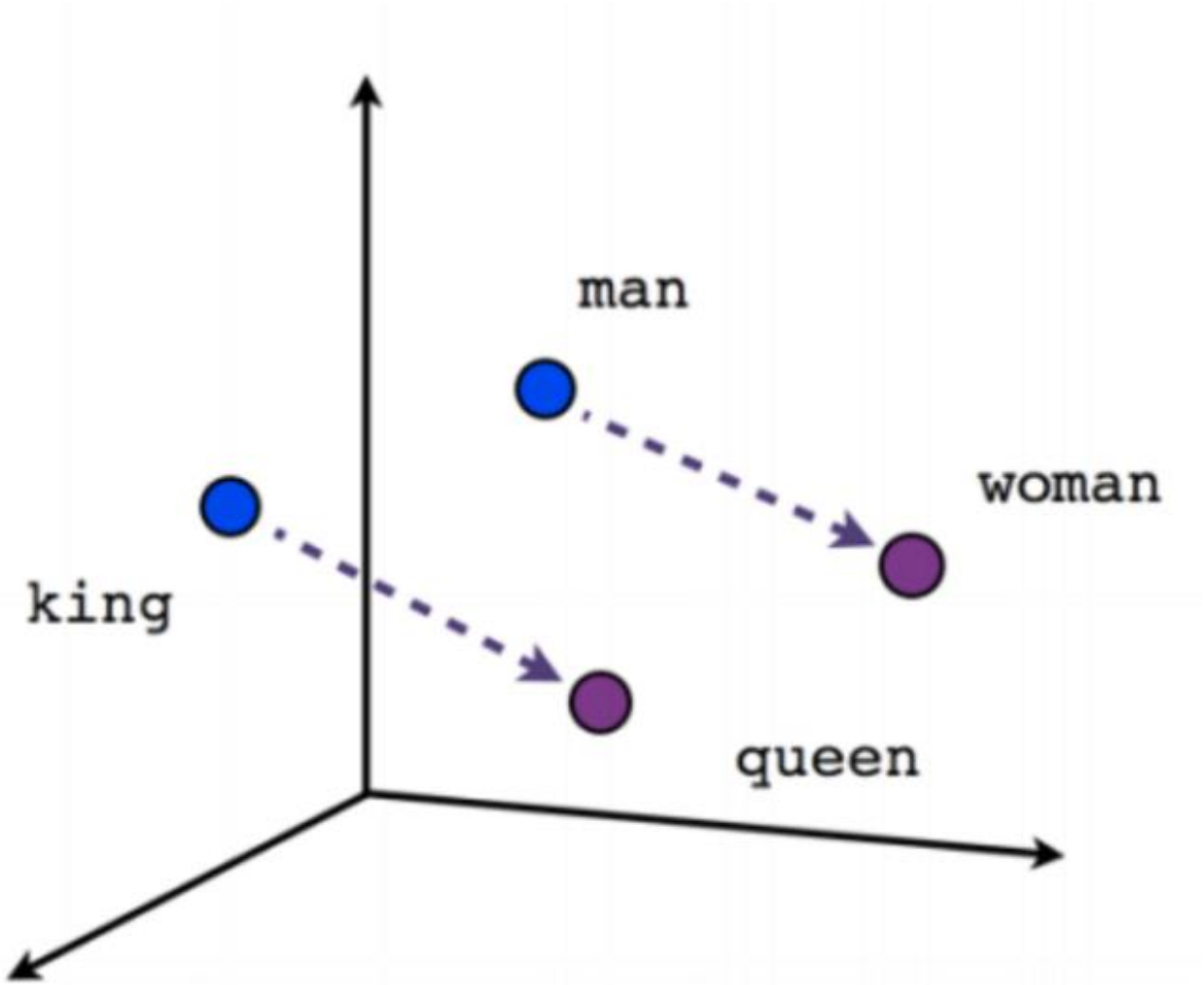
Extra:

'not go'	't'	'not'	'to the'	'try to'
'definitely'	'less'	'try'	'and explain'	'clients'
'given'	'opportunity to'	'it comes'	'can do'	'week i'
'company'	'would definitely'	'social'	'meeting i'	'work i'
'out i'	'sure i'	'if'	'love to'	'then'

Agree:

'would'	'with'	‘ boring'	'first'	'definitely'
'in the'	'it'	'know'	'to'	'this'
'free'	'i would'	'try and'	'not work'	'then'
'i think'	'love'	'would have'	'i had'	'meeting'

O: “**love**”; C: “**would**”; N: “**opportunity to**”



Doc2Vec

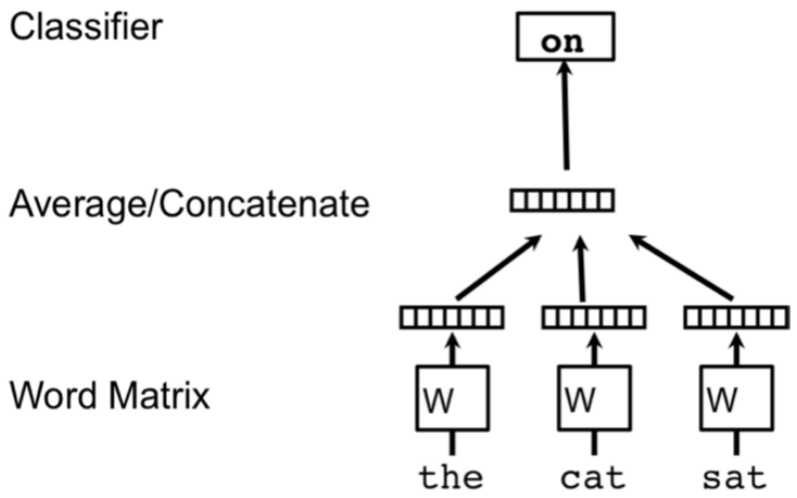
Word2Vec

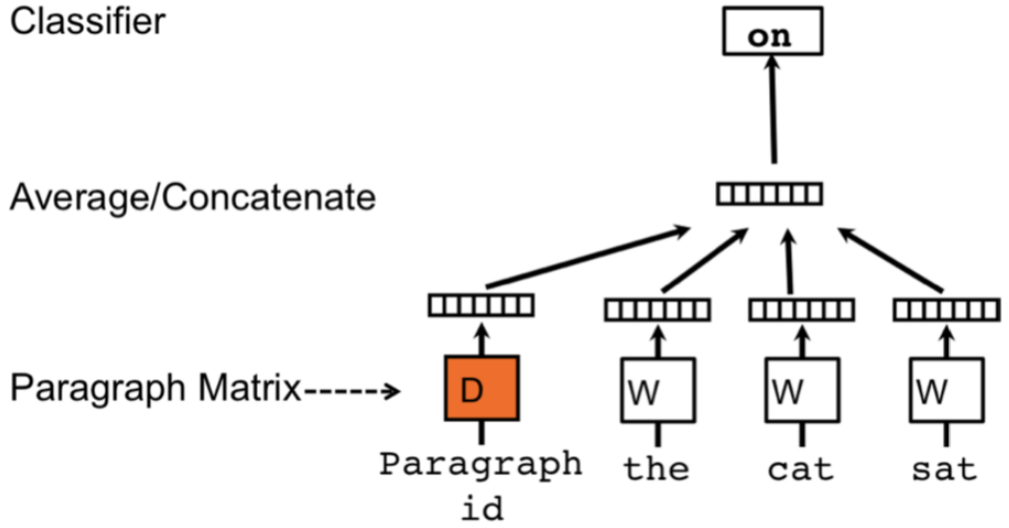
Idea: Use context (i.e., surrounding words) to predict the target word

Word2Vec

Doc2Vec

Continuous Bag of Words **vs.** **Distributed Memory Model (DM)**



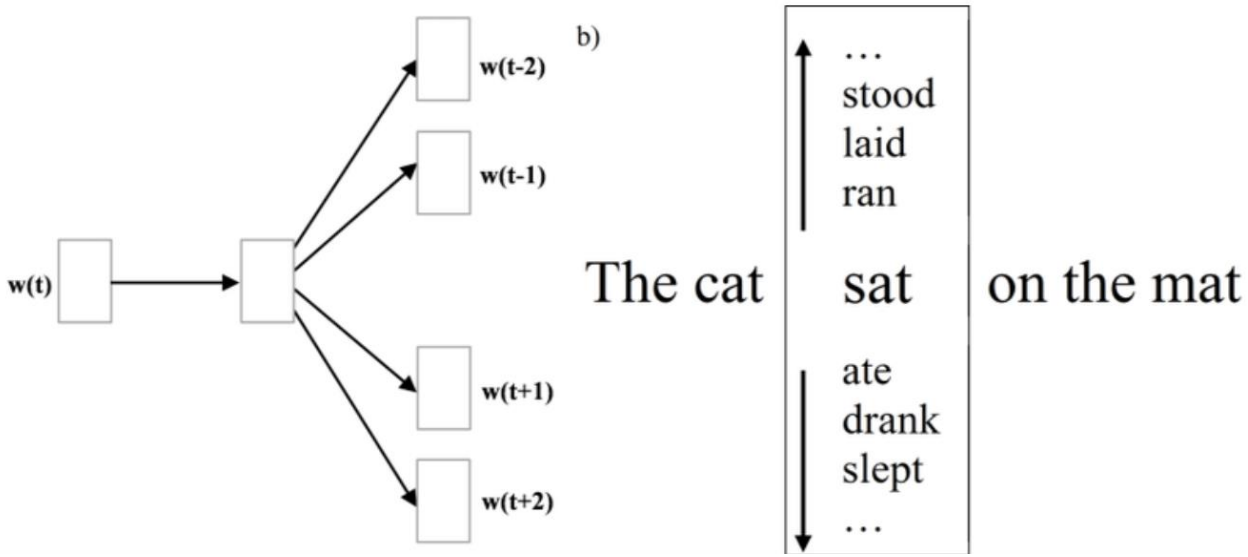


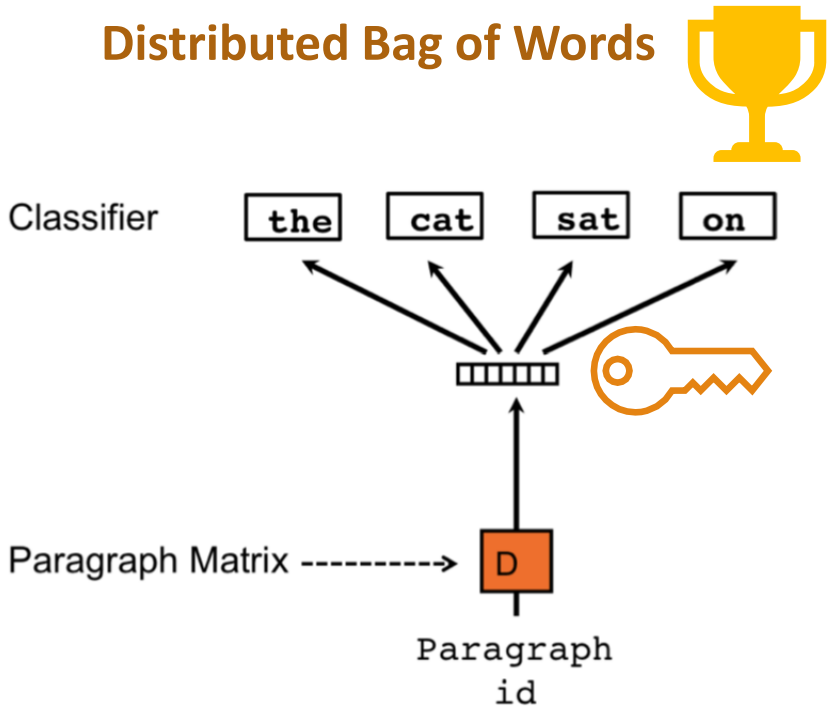
Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

Word2Vec

Doc2Vec

Skip-Gram Model **vs.** **Distributed Bag of Words**





Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

Doc2Vec model using Gensim

Training data + Test data both trained (unsupervised model)

Parameters:

- 2000 features/dimensions/neurons from the hidden layer
- 10 epochs

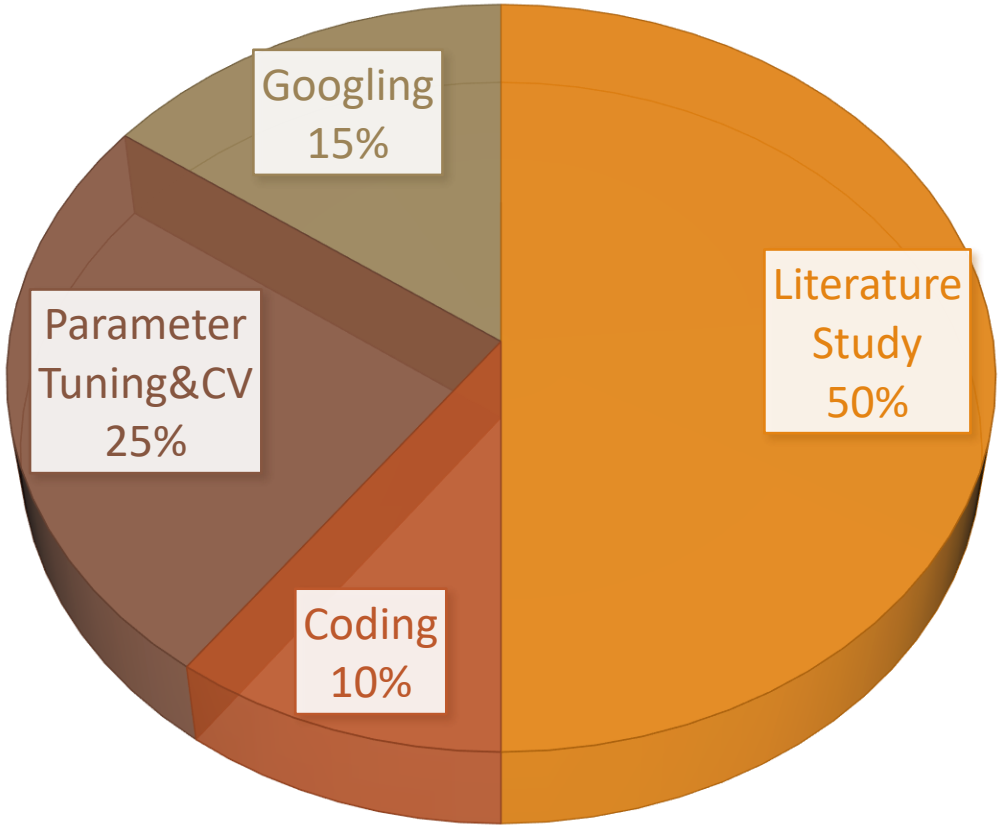
Ridge regression conducted based on features extracted from Doc2Vec model

Doc2Vec Model

Findings: Check Response Similarity on Neuroticism

A colleague of yours has requested vacation for the same week as you. According to your supervisor one of you has to take a different week of vacation because it would be too busy at work if both of you are absent. Your colleague is not willing to change their vacation plans. What would you do and why?		
ID_457	ID_778 (Most Similar)	ID_221 (Most dissimilar)
'i would ask my coworker why they would not change their plans after that i would try to figure out some sort of compromise perhaps one of us could have this vacation time and next time the other would get the vacation time if we could not reach an agreement i would ask my boss to step in'	'i would see if there was any way i could switch vacation times with him if he was unwilling i would ask my boss if there was a way i could take my vacation at a later time when no one else was taking their’	'i would not be willing to change plans it is the supervisor s job to advise on what weeks are available and scheduling it depends on who submitted the vacation request first depending on what i may actually have planned for my vacation i may or may not be willing to take a different week’

TIME ALLOCATION



"BACKGROUND RESEARCH IS IMPORTANT (PAPERS, BLOGS, PACKAGE TUTORIALS, KAGGLE DISCUSSIONS, ETC.)"

"PYTHON > R???"

"LEARNED ABOUT DATA SCIENCE WORKFLOW AND PROJECT MANAGEMENT/WORKING ON DATA ANALYSIS WITH A GROUP."

"USING THE CARET PACKAGE IN R CAN SIMPLIFY MODEL-BUILDING"

" BETTER GRASP OF MACHINE LEARNING CONCEPTS IN GENERAL"

LESSONS LEARNED

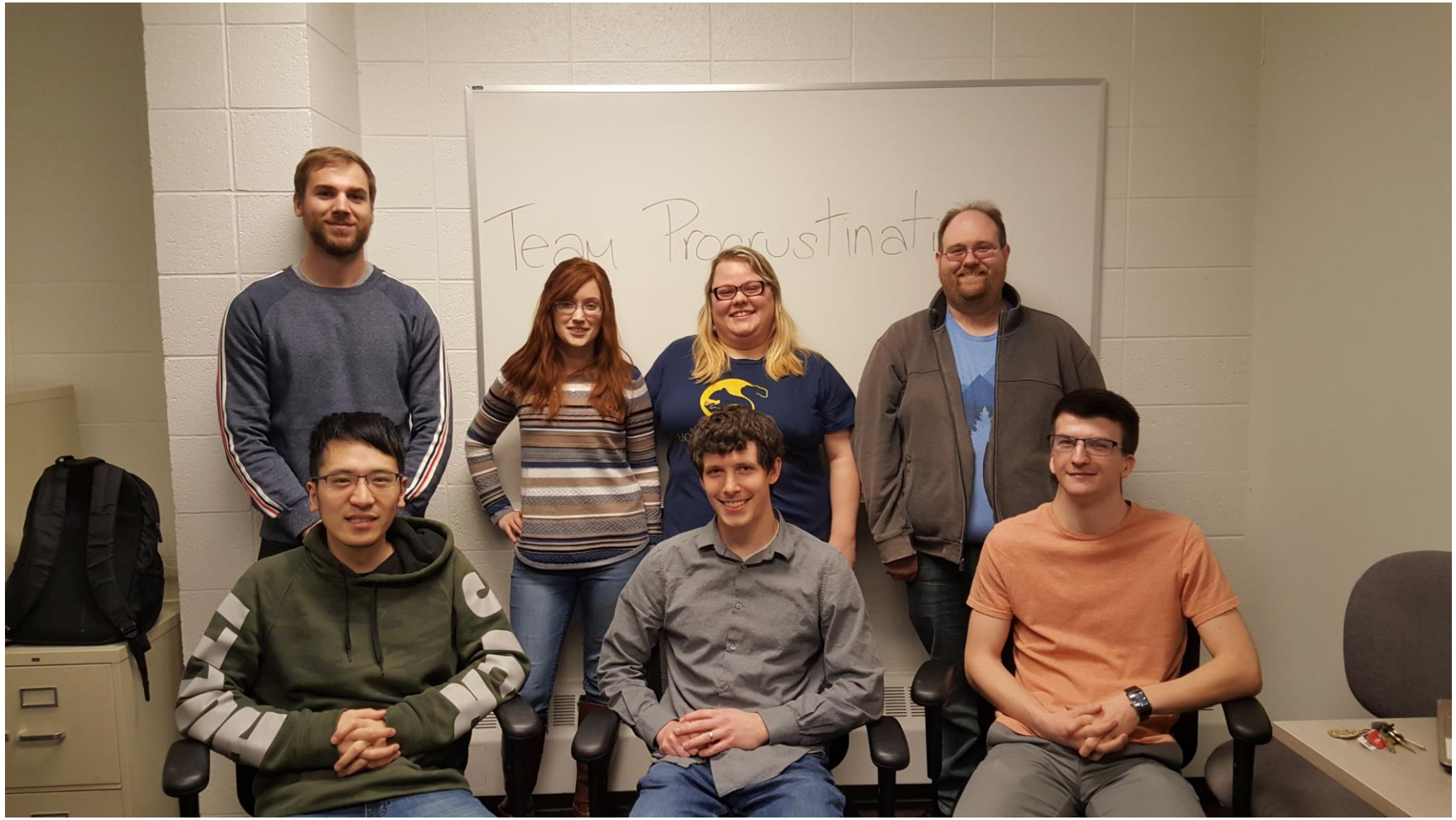
"... A COMBINATION OF INDUCTIVE AND DEDUCTIVE APPROACHES HELPS TO CAPTURE THE VARIANCE IN DIFFERENT

" MACHINE LEARNING APPROACHES INVOLVE A GREAT NUMBER OF STATISTICAL AND TECHNICAL DECISIONS"

"A GOOD COMPUTER IS CRUCIAL !"

"CROSS VALIDATION AND TUNING PARAMETERS ARE VITAL AND TIME-CONSUMING"

Thank You!



For more information on this and related projects, please contact Dr. Sam McAbee (smcabee@bgsu.edu)