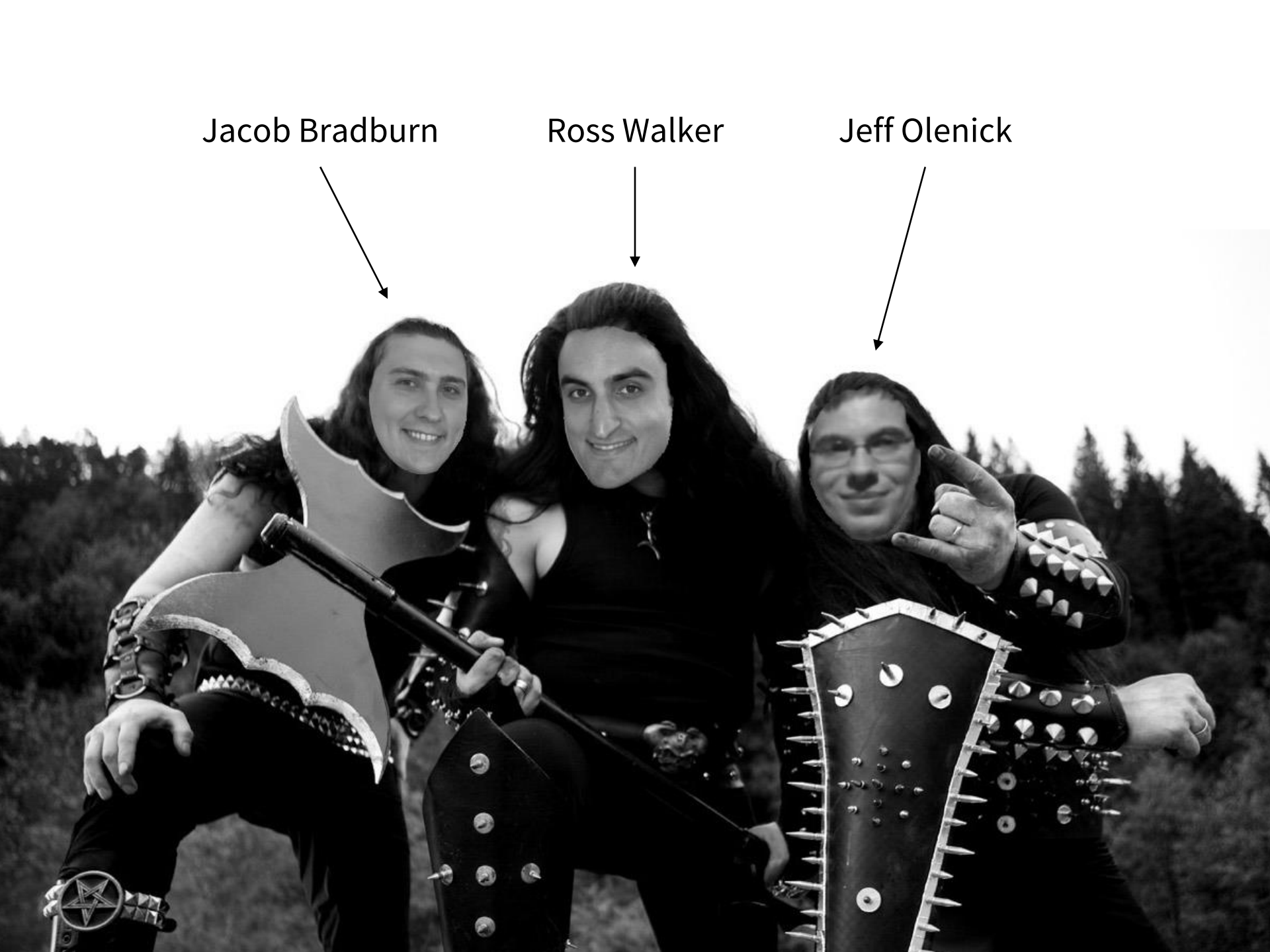
2019 SIOP Machine Learning Competition

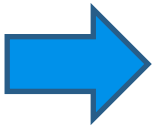
Team Name:

Logistic Aggression


Jacob Bradburn
Ross Walker
Jeff Olenick

Feature Extraction

Unstructured Text



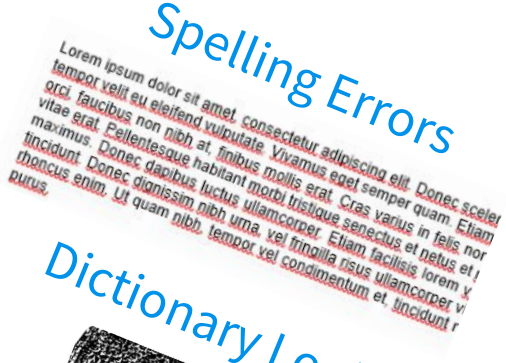
Numeric Variables

Document Term Matrix

	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0		0		0	1	1	1

Profanity

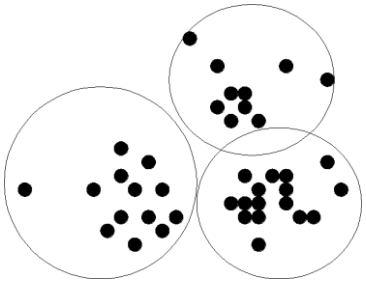
Spelling Errors



Dictionary Lookups

Lexical Diversity

Topic Models



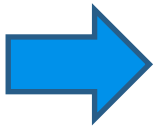
Sentiment

Readability

Parts of Speech

Model Development

Numeric Variables



Predictions of Big 5 Scores

Model Development

Linear Models

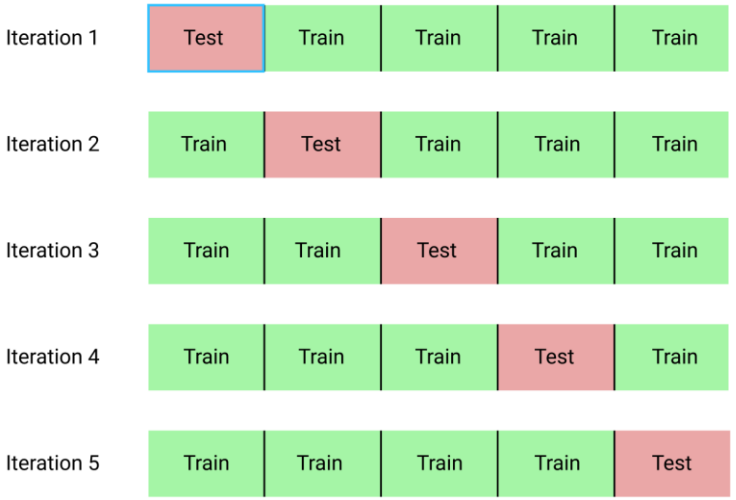
- Principal Components
- Partial Least Squares

Non-Linear Models

- Support Vector Machines
- Random Forest
- Boosted trees
  - GBM
  - XGBoost

Model Validation

1. Local Validation
   - k-fold CV
   - Repeated CV
     1. Randomly select a hold-out set (n = 300)
     2. Calculate mean *r*
     3. Repeat
2. Development Set
3. Test Set



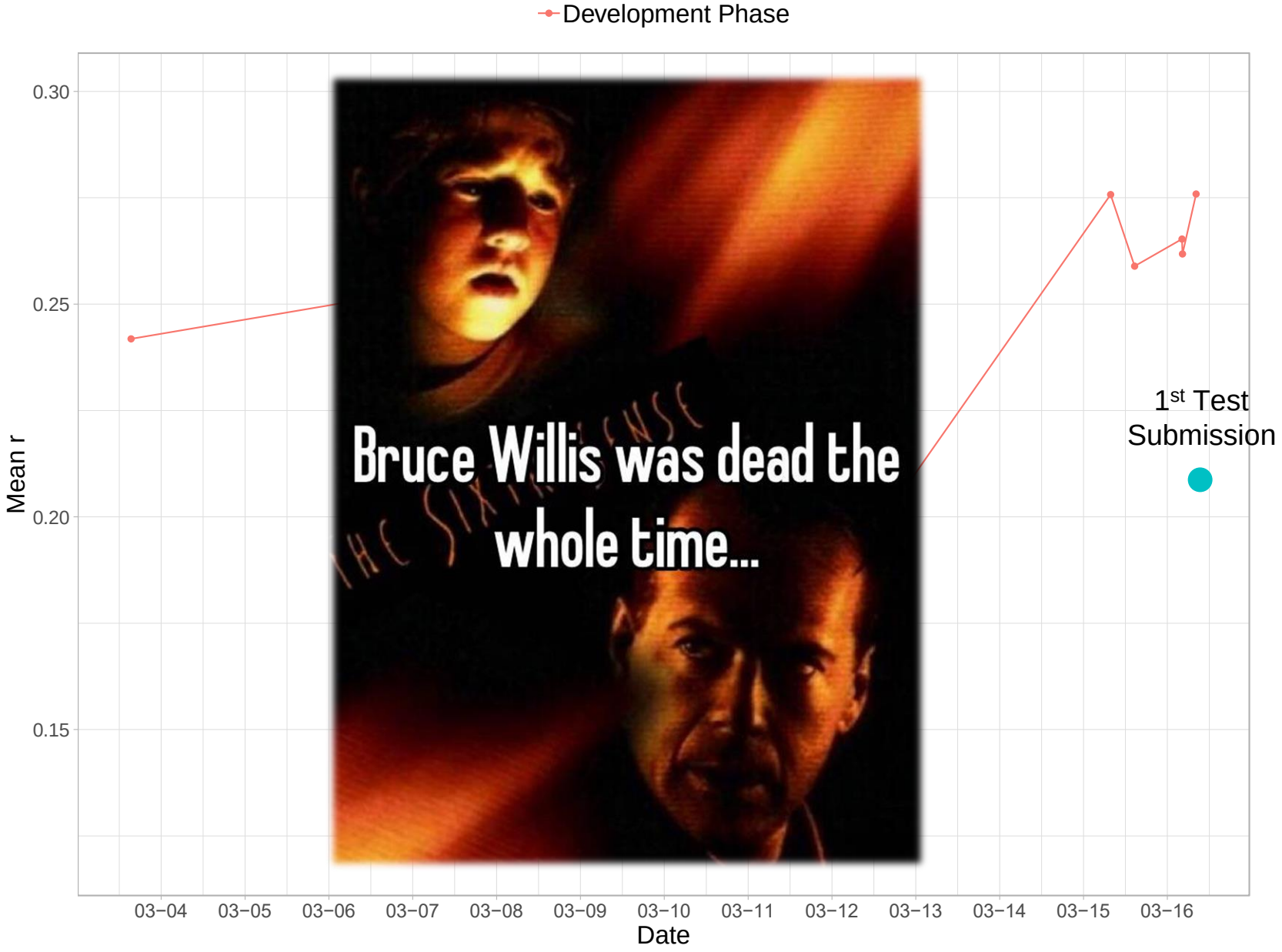
Model Development

Linear Models

- Principal Components
- Partial Least Squares

Non-Linear Models

- Support Vector Machines
- Random Forest
- Boosted trees
  - GBM
  - XGBoost


Development Phase
0.30
0.25
0.20
0.15
Mean r
Bruce Willis was dead the whole time...
1st Test Submission
03−04
03−05
03−06
03−07
03−08
03−09
03−10
03−11
03−12
03−13
03−14
03−15
03−16
Date


Development Phase
Test Phase
0.30
0.25
0.20
0.15
Mean r
03−04
03−05
03−06
03−07
03−08
03−09
03−10
03−11
03−12
03−13
03−14
03−15
03−16
03−17
03−18
Date

How do we prevent overfitting?

How do we prevent overfitting?

1. Variable Selection

- Only keep variables with zero-order correlation p-values below a threshold

2. Elastic Net Regression

- *Ridge* = coefficients asymptote at zero
- *Lasso* = coefficients can actually reach zero
- Parameters
  - α = blend of ridge & lasso
  - λ = penalty weight

Top 5 Heaviest Beta Weights

Extraversion

"I would not (go)" (-.11)

"I would go anyway" (.08)

"I wouldn't want to…" (-.07)

"Love" (.07)

Sentiment on social Q (.07)

Agreeableness

Lexical Diversity R (.09)

"Probably" "depend on" (-.08)

"Getting paid" (-.08)

"If X, I would Y" (-.07)

"(First) request" (-.07)

Conscientiousness

Depth (-.08)

"One / Both of us" (-.06)

"If X, I would Y" (-.06)

Liberal Values (.06)

Sentiment on travel Q (.06)

Neuroticism

"Try" (.07)

"Need to find out" (-.07)

"I would go anyway" (-.06)

Anxiety (.06)

Commas (.05)

Openness

Sentiment on travel Q (.08)

"I would go anyway" (.08)

"I would not (go)" (-.07)

"Discuss" (-.06)

Negated Positive Words (-.06)

Lessons Learned

1. **Make 1 test submission ASAP!**

2. **Pay attention to bias/variance tradeoff during local validation.**

3. **Things we could have done better:**
   - Linear transformations
   - Multivariate analyses

Thanks!

Ross Walker: riwalker@msu.edu
Jacob Bradburn: bradbu17@msu.edu
Jeff Olenick: olenickj@msu.edu