

The background features a light gray gradient. On the left, there is a complex network graph with dark gray nodes and lines. Scattered across the background are various thin-lined triangles of different sizes and orientations. In the top right corner, there are small, faint circles.

# **VIRAL TEXT ANALYSIS**

---

Predicting Information Propagation in  
Machine Learning Communities

## WHY DO WE CARE?

Why does information  
propagation matter?

01

## DATA

What's our data?  
Where does it come from?

02

## METHODS

How do we interpret this  
data?

03

04

## FINDINGS

What was discovered?

05

## Recommendations

What next?

06

## CONCLUSION

Closing thoughts,  
questions



# 01

## WHY DO WE CARE?

---

Why does information propagation matter?



# 01

---

## WHY DO WE CARE?

**A better model of text propagation  
means a better understanding of:**

- How companies build brands or attract customers
  - How research gets attention, citations, and funding
  - How ideas spread across cultures or teams
  - How misinformation spreads
- 





# 02 DATA

---

What's our data? Where does it come from?



## What's our data? Where does it come from?

---

- Scraped text data from Twitter users connected to ML researchers
- Accounts between 1000 and 50000 followers
- Predicting 'Retweets' as our goal



## OUR DATA IN 3 NUMBERS:

63000

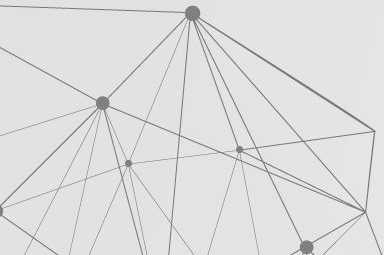
Tweets Analyzed

1524

Different Users

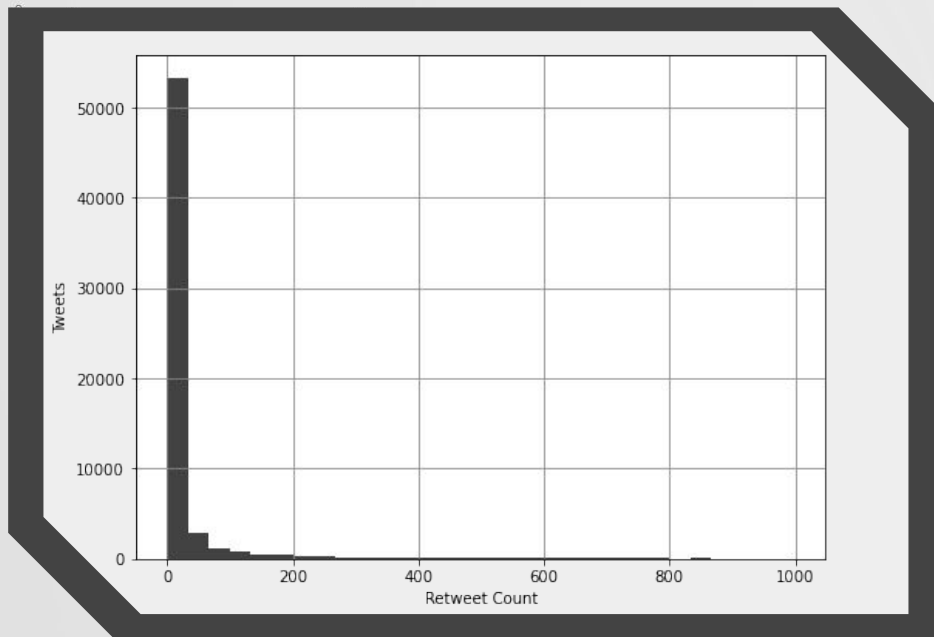
2

Machine Learning Clusters



# What is our data?

# 02



## Distribution of Retweet Count

Most content is not viral –  
Most has 0-1 retweets







# 03

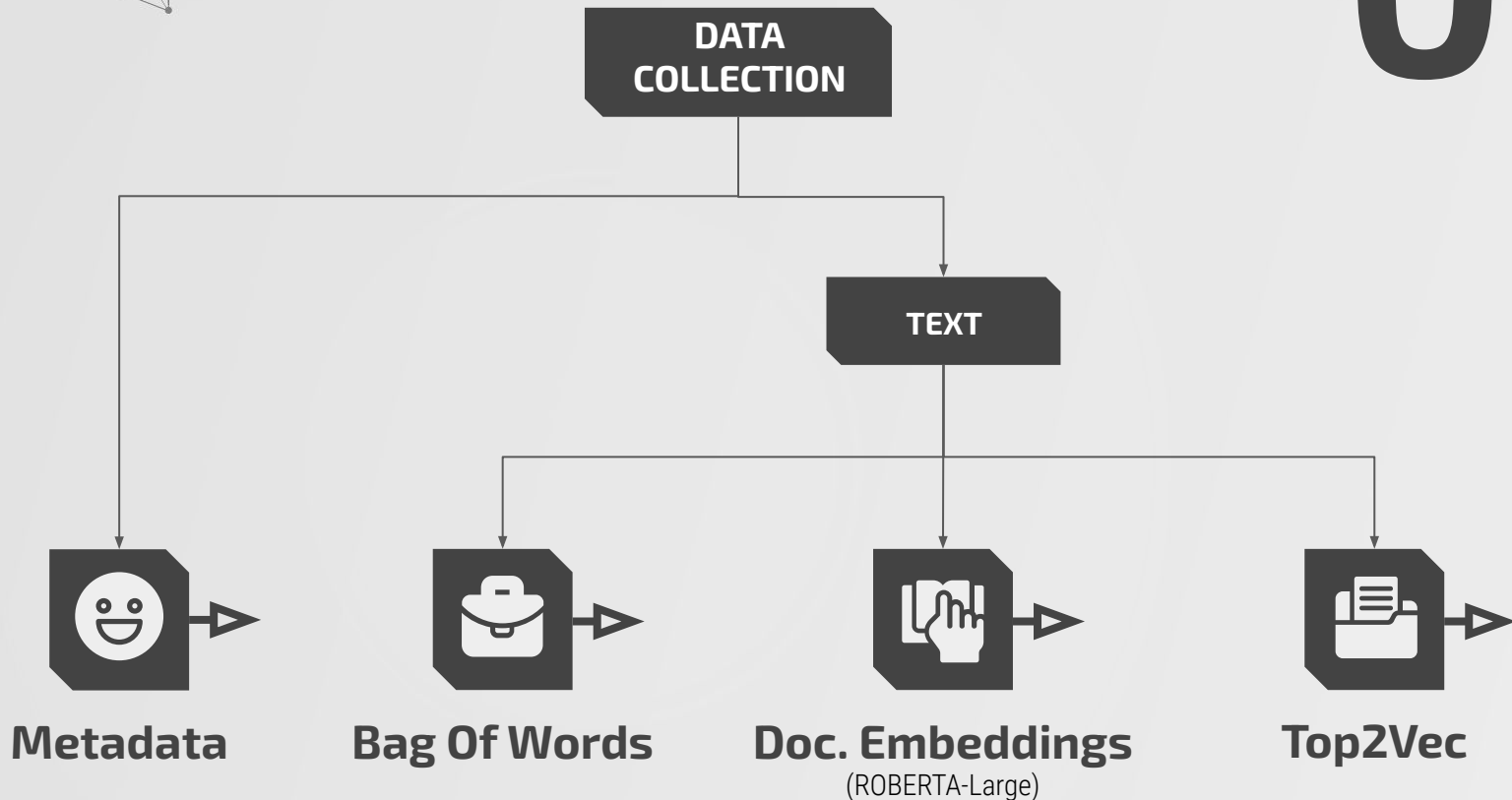
## METHODS

---

How do we interpret our data?

# HOW DO WE INTERPRET OUR DATA?

# 03



# MACHINE LEARNING MODELS

# 03



**Linear Regression**



**XGBoost**



**TabNet**



- Many algorithms tested (ask for details)
- Both Regression and classification



**Random Forest**



**Deep Learning**



**1D CNN**

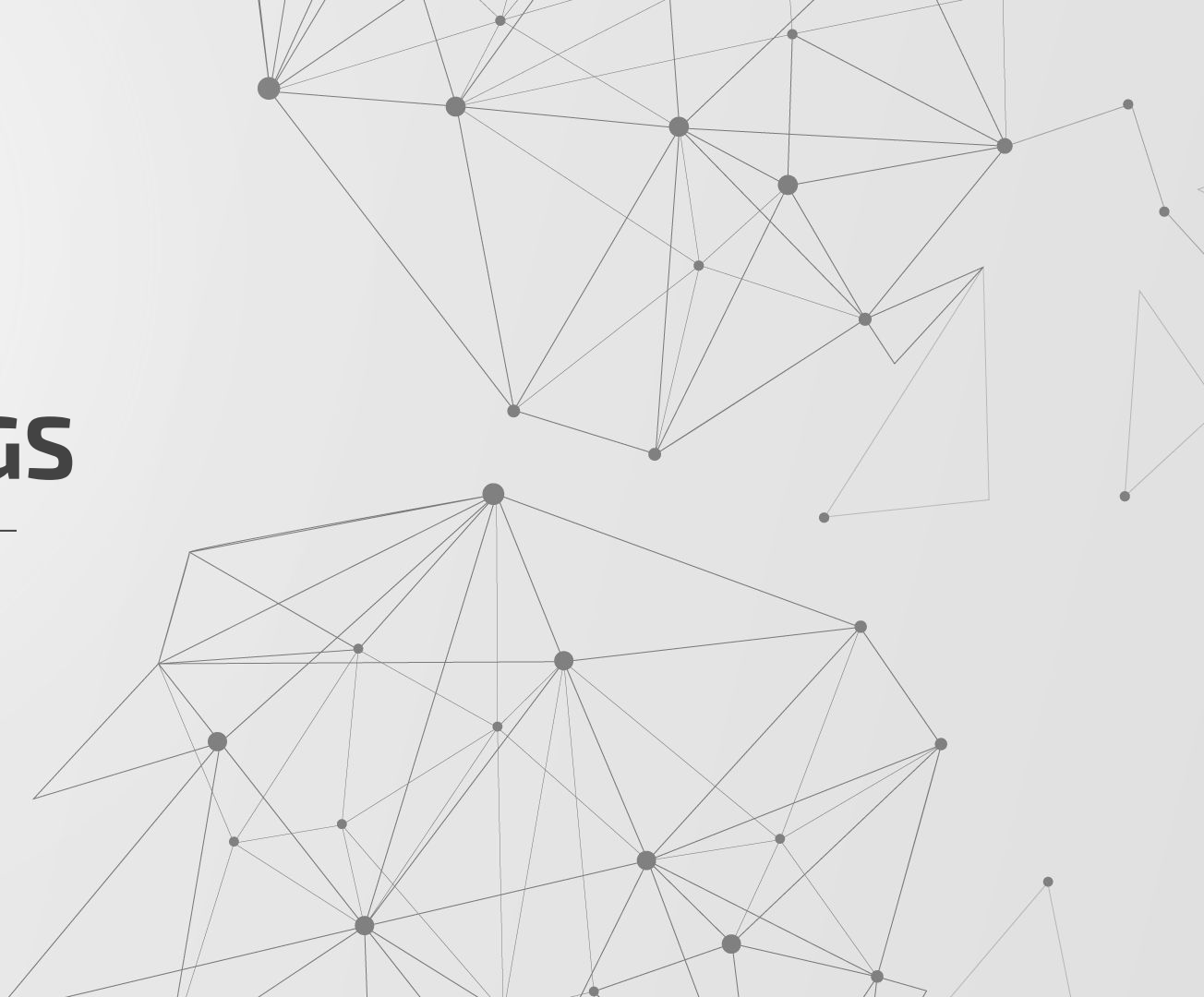


# 04

## FINDINGS

---

What was discovered?



# What was discovered?

# 04



## Viral Text is Different

There is a measurable difference between viral text and nonviral text - we could explain about 11% of variance ( $R^2$ ) using NLP



## Viral Topics

Some topics are clearly more viral, e.g. talking about OpenAI, or hiring phd candidates

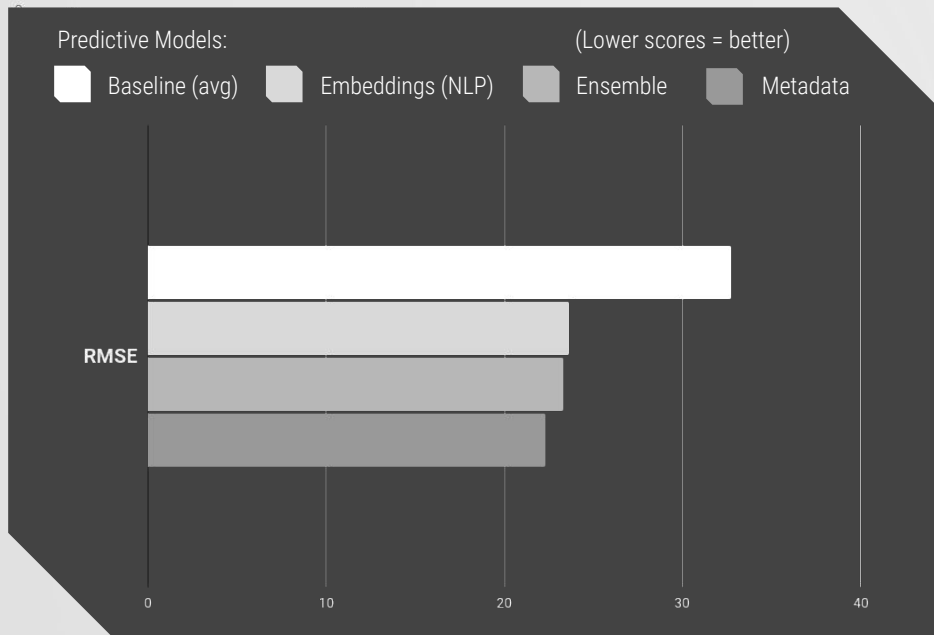


## Quantifiability

Specific words/topics can be measured, e.g. '100daysofcode' had 16% correlation ( $R$ ) with retweets

# What was discovered?

# 04



## REGRESSION PERFORMANCE

- XGB on metadata remains more predictive than NLP
- Ensemble methods yet to improve performance

## Classification Performance on Holdout Data

# 04

	Precision	Recall	F1-Score	Support
Not Viral	0.98	0.98	0.98	2520
Viral	.27	0.26	0.26	70
Accuracy	>50 retweets		97%	2590

The background features a complex network of thin grey lines connecting various-sized dark grey circular nodes. These nodes are scattered across the slide, with a higher concentration on the right side. Some nodes are isolated, while others are part of larger, interconnected clusters. The overall effect is a modern, technical, and interconnected aesthetic.

# 05

## RECOMMENDATIONS

What next?



## What Next?



### Process Integration

How might this fit in a marketing dashboard? Could the service be used to market itself?



### Improve the Model

Collecting more data and improving on embeddings from newer large language models



### Reuse the Pipeline

Can we predict citation counts of research papers based on their titles and abstracts? Etc.



# 06

## CONCLUSIONS

In a nutshell...

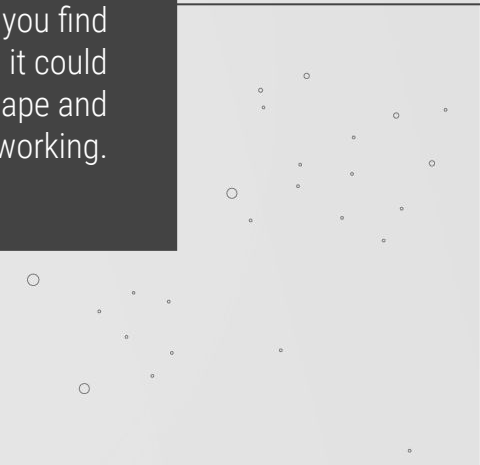


## In a nutshell...

# 06

We were able to use machine learning to predict 26% of viral tweets. That means we could start to preemptively score marketing and communications for virality.

Virality may always be hard to predict – like predicting the stock market, if you find something that works, it could change the landscape and stop working.





# THANKS!

Any questions?

blakemcme@gmail.com  
[github.com/thegrandblooms](https://github.com/thegrandblooms)  
[linkedin.com/in/blakemcme](https://linkedin.com/in/blakemcme)

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

**Please keep this slide for attribution.**

# Technical Details

## CLASSIFICATION PERFORMANCE

- Random Forest had the best F1 score at .31 on Val.
- Also the best Area Under the Curve at 82%

