UVA DATA SCIENCE

# Engagement Time Machine

Griffin McCauley, Theo Thormann, Eric Tria, and Jake Weinberg

Thank you to our sponsors at hum: Will Fortin, Dylan DiGioia, Niall Little, and Dustin Smith

And our faculty advisors: Prof. Judy Fox, Ian Liu, and Prof. Jason Williamson

May 3rd, 2023

# Our Team

**Griffin McCauley**

M.S. Data Science

Sc.B. Applied Mathematics,
A.B. Economics

**Theo Thormann**

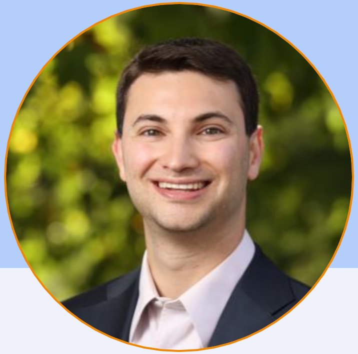
M.S. Data Science

B.S. Environmental Science
and Policy

**Eric Tria**

M.S. Data Science

B.S. Computer Science



**Jake Weinberg**

M.S. Data Science

B.S. Commerce

Agenda

01. **Project Background**

02. **Cluster Analysis**

03. **Deep Learning Implementation**

04. **Concluding Remarks**

01.

# Project Background

# Client and Project Overview

## hum

- Data analytics start-up headquartered in Charlottesville
- Operates in the academic publishing industry
- Utilizes proprietary CDP to collect first-party data across clients' online content

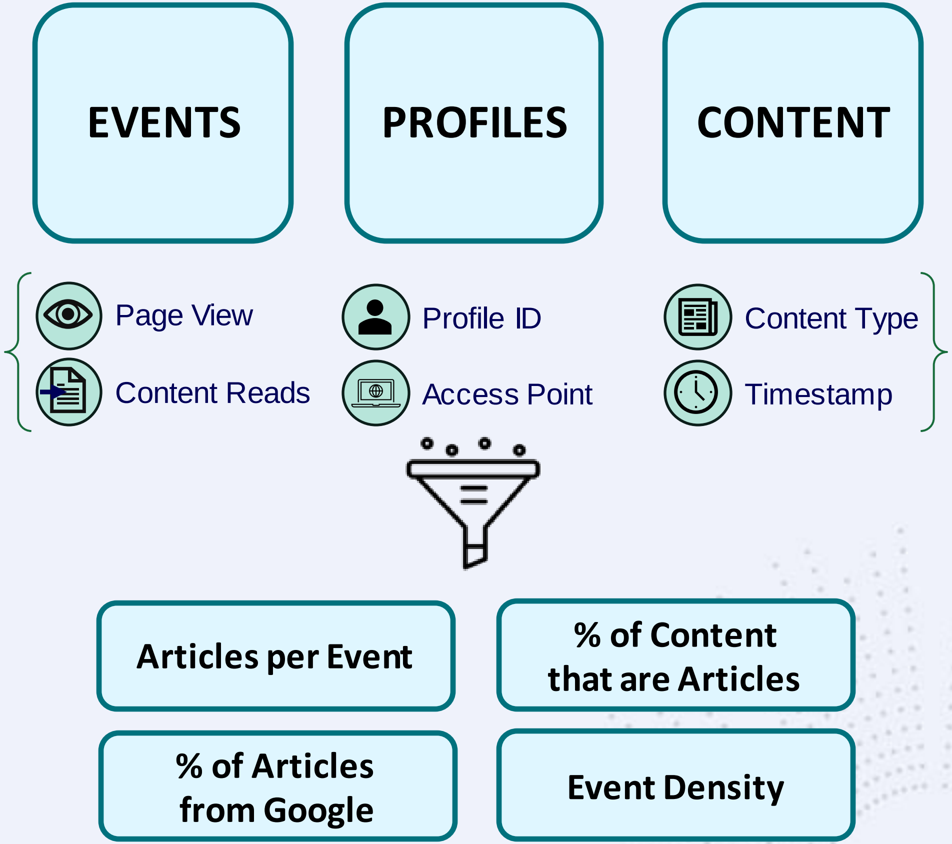
## Project Background

- Academic publishing industry is now experiencing the big data revolution
- Greater understanding of user engagement patterns has massive business implications
- **Enhance and optimize the inefficient peer reviewer selection process**

**Goal: Engineer a novel set of user-level features and construct a model to accurately recognize high-quality, valuable users early on in their lifecycles**

# Data

- First-Party Customer Data
- Significant events and user behavior
- From March 2022 to March 2023: roughly **2.2M users** and **13.4M user events**
- Focused on **3 tables**
- Engineered **4 main features**
- Cloud Access through **Snowflake**
- Pipeline hosted on **Amazon Web Services (AWS)**



EVENTS

PROFILES

CONTENT

Page View

Content Reads

Profile ID

Access Point

Content Type

Timestamp

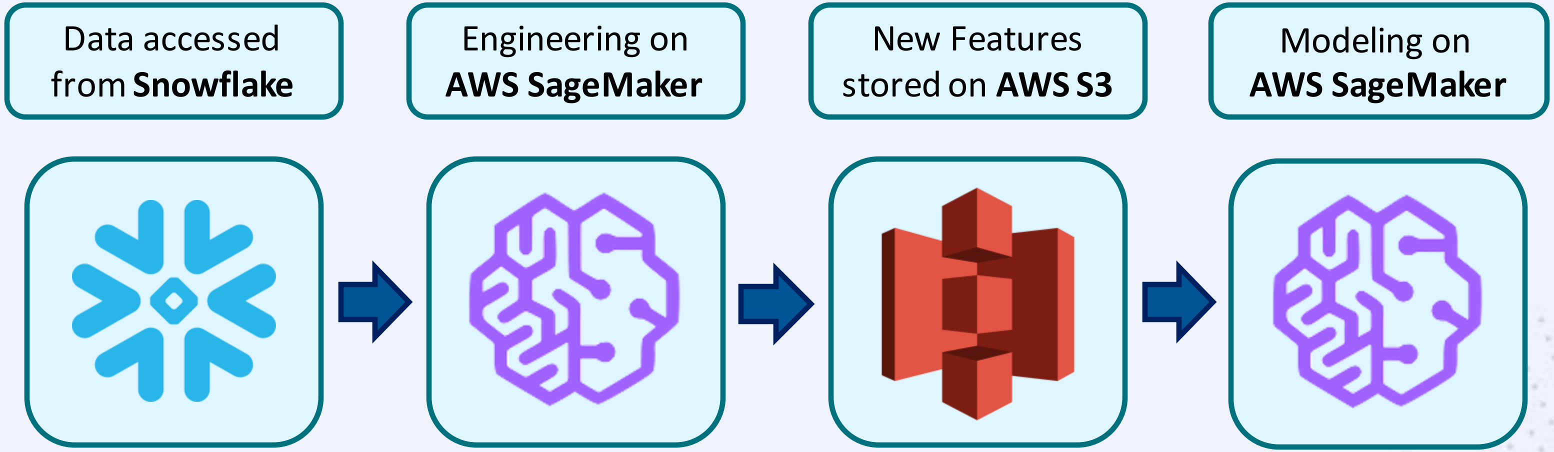
Articles per Event

% of Content that are Articles

% of Articles from Google

Event Density

# Data Pipeline



Data accessed from **Snowflake**

Engineering on **AWS SageMaker**

New Features stored on **AWS S3**

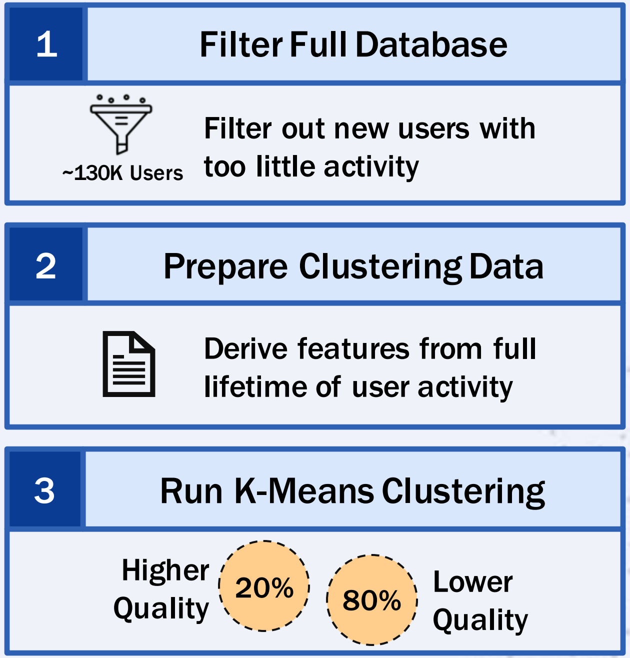
Modeling on **AWS SageMaker**

02.

# Cluster Analysis

# Clustering

- **Problem:** No industry standard for what constitutes a high-quality user
- Needed to define our own training labels
- **Solution:** Labeled users via K-means clustering analysis
- Found that the two clusters represented higher- and lower-quality users
- Clusters can be used to identify peer reviewer targets



1 Filter Full Database

~130K Users

Filter out new users with too little activity

2 Prepare Clustering Data

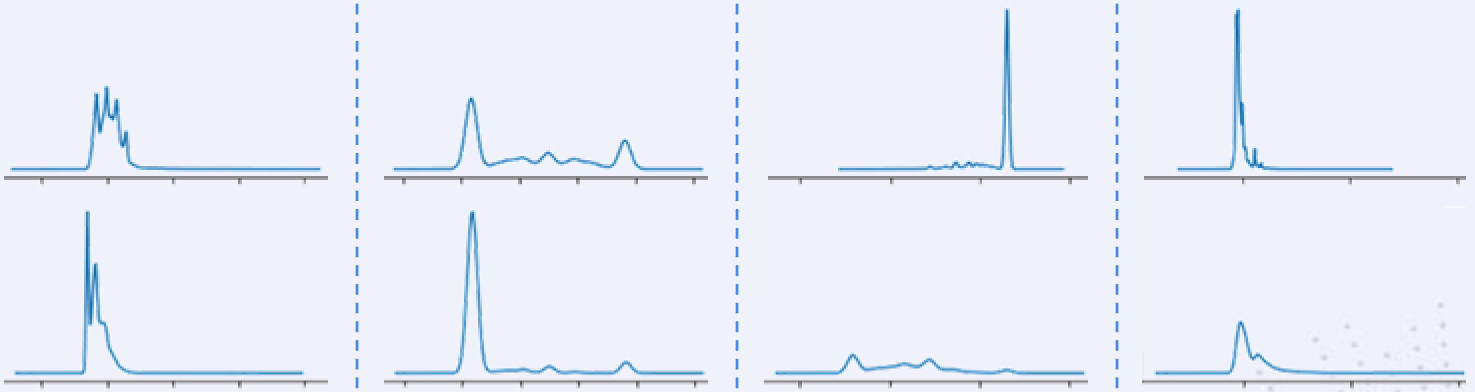
Derive features from full lifetime of user activity

3 Run K-Means Clustering

Higher Quality 20%

80% Lower Quality

# User Profiles



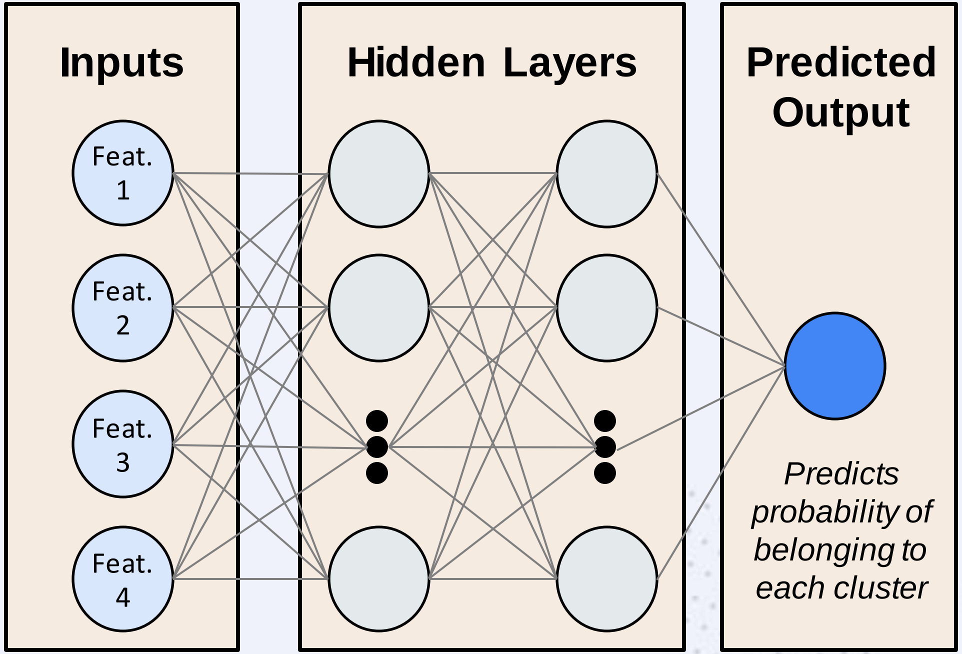
| | Articles per Event | % of Articles Found Through Google | % of Content Read That Are Articles | Events per Day |
|---|---|---|---|---|
| High Quality Users Have: | Low number of articles per event | Low % of articles found through Google | Low % of content that are articles | High number of events per day |

03.

# Deep Learning Implementation

# Deep Learning Model Structure

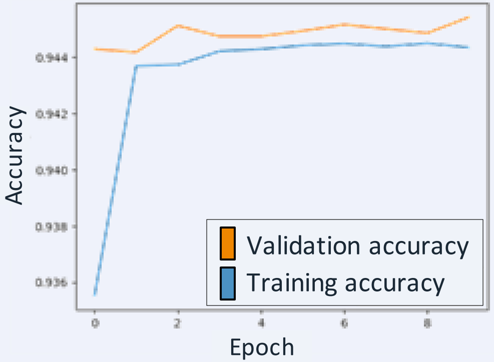
- Built a deep learning MLP model to assign each user to a cluster
- Used same features as with clustering, but only derived from early user activity
- Structure enables Hum to customize model for other clients and new applications



# Results

## Training Curve

*Shows effectiveness of model training and ability for model to classify users*

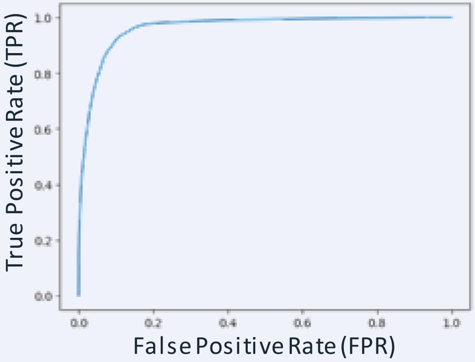


**95%**

**Test Accuracy**

## ROC Curve

*Shows diagnostic capability of classifier (scale: [0,1]; higher is better)*

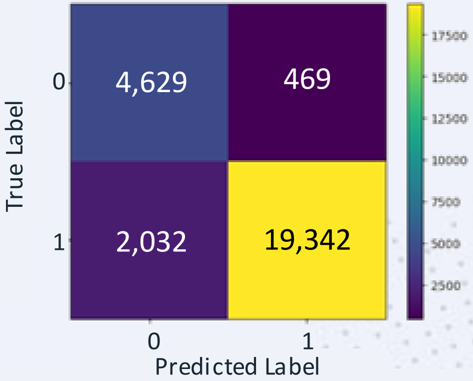


**0.96**

**AUC**

## Confusion Matrix

*Shows classification accuracy and misclassification types*



**91%** **TPR & TNR**

**98%** **PPV**

**70%** **NPV**

04.

# Concluding Remarks

# Current State & Next Steps

## Project Impacts

- **Found that user lifetime behavior can be predicted very early on**
- Constructed a robust model framework that can be easily extended to other academic publishers
- Classified user engagement with high accuracy based on novel features

## Future Applications

- **Identify potential peer reviewers based solely on reading behaviors**
- Tailor recommended content and ads based on user activity
- Incorporate information for other clients and more granular user data

# Acknowledgements

Thank you to our sponsors at hum:

**Will Fortin, Dylan DiGioia, Niall Little, and Dustin Smith**

And our faculty advisors:

**Prof. Judy Fox, Ian Liu, and Prof. Jason Williamson**

# Thank you for your time!
# We hope you enjoyed.