

## ➤ Background

- Project intros
- Current state of the project

## ➤ Goal

- Problem articulation
- Has your idea been done by others?

## ➤ Dataset(s)

- Basic EDA's
- Database set up

## ➤ Methods

- What methods are you planning to use?
- Any related literature to support your proposed methods?

## ➤ References

# Capstone Project Proposal

Team Members: Griffin McCauley, Theo Thormann, Eric Tria, Jake Weinberg

Supervisor: Prof. Judy Fox

# Checklist of Goals

1. Understand trends in the data given to us by the client and figure out what data is useful in creating a model and what information is noise
2. Create a user retention model, which finds patterns in user events to identify user churn risk and predict resubscription behavior
3. Build an RNN to analyze the user event sequences on the platform
4. Produce a 6-page publishable paper (in IEEE format), along with an oral presentation, on our work summarizing what the model the team has created and what we have found using our model
5. Package and share our model, GitHub, and research findings with our client

# Background



- Customer data platform (CDP)
- Use that data to help publishers understand their customers and content
- Focus on educational publishers

## Project Background

- Publishers want to maximize user retention
- To do this we will utilize user "events" to create models
- User events include pageviews, citations, scrolling activity, and more

### Goals

### Critical Activities

#### Fall

- Complete kickoff and onboarding
  - Gain subject matter expertise
  - Complete EDA ✓
- 
- Gain access to client data
  - Understand necessary analysis tools ✓

Goal: Create models using user events to maximize user retention and potentially other use cases for HUM

# Tasks and Timeline

**January** – Finish all set up and background research to prepare for model execution

- Choose how to embed our data and input it to an RNN
- Determine what criteria to use for isolating training data and classifying users
- Coerce the data into the correct format and perform a trial run on a simple RNN model

---

**March** – Refine the model and perform hyperparameter tuning

- Determine a subset of hyperparameters that we want to tune and select a reasonable range of values for these
- Retrain and evaluate the model for a variety of hyperparameter configurations

**February** – Successfully complete initial training of user retention model

- Build code that comprises the backbone of our model

---

**April** – Prepare the final model for deployment and ensure customizability for different clients

- Extract insights from the model to market to clients
- Package the model for easy interpretability and implementation by our client

# Datasets

- Use first-party customer data, which includes significant user events and actions collected by the client
  - Events include things like page views, page scrolls, and citations
- While our data is mostly cleaned for us by our client, our group will engineer the data to effectively prepare it for use in the model
- Data is hosted on Snowflake
  - We will use Python integration through the Snowpark API
- Machine Learning models will run on AWS SageMaker

# Datasets

## EVENT

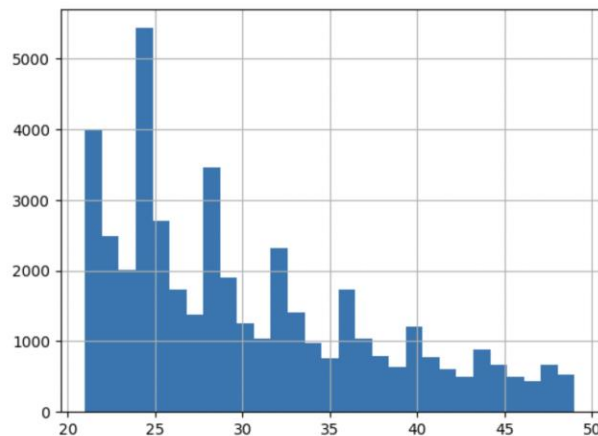
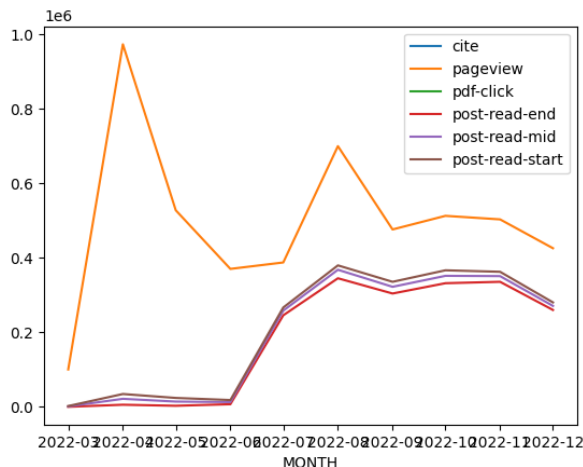
```
1  create or replace view CORE.CLIENT.EVENT(  
2      CLIENT,  
3      ID,  
4      TAGS,  
5      META,  
6      DAY,  
7      KEYWORDS,  
8      REFERER,  
9      UTM_CAMPAIGN,  
10     UTM_CONTENT,  
11     UTM_MEDIUM,  
12     UTM_SOURCE,  
13     UTM_TERM,  
14     SET_PROFILE,  
15     SET_USER,  
16     IP,  
17     USER_AGENT,  
18     SOURCE,  
19     URL,  
20     VISITOR_ID,  
21     DATE,  
22     EVENT,  
23     CONTENT_ID,  
24     CREATED,  
25     UPDATED  
26 ) ROW ACCESS POLICY #unknown_policy  
27 as  
28 select * from public.EVENT;
```

## PROFILE

```
1  create or replace view CORE.CLIENT.PROFILE(  
2      CLIENT,  
3      ID,  
4      USER_ID,  
5      EMAILS,  
6      CAMPAIGNS,  
7      CREATED,  
8      UPDATED,  
9      DOMAINS,  
10     FIRST_VISIT,  
11     IDENTIFIED_ON,  
12     IDENTIFYING_REFERER,  
13     IDENTIFYING_UTM,  
14     LAST_ACTIVE,  
15     ORGANIZATION_IDS,  
16     SEGMENTS,  
17     PROPERTIES,  
18     METRICS,  
19     PERCENTILES,  
20     USER_SIDES  
21 ) ROW ACCESS POLICY #unknown_policy  
22 as  
23 select * from public.profile;
```

## Datasets (EDA)

- Recurring events to use: cite, pageview, pdf-click, post-read-start, post-read-mid, post-read-end
- A good portion of RUP users have between 20 and 50 events for 2022.



# Methods

- We currently plan on implementing an RNN for the user retention model
  - Our sequence data will contain information regarding the event type and time of execution, and we will likely use an event-time joint embedding (Li et al., 2017) in order to incorporate both components into the model input
- Details of the model architecture are still under consideration
  - Potential use of an LSTM hidden layer for capturing longer-term sequence dependencies
  - Probable use of a softmax activation layer for multiclass classification of the output
- Depending on client needs, we may also add another RNN for sequence classification in addition to next-step event prediction



## Conclusions & Future Work (TBD)

- What solutions you have accomplished and how do they compare with related work?
- Any insights or observations from this work?
- What can be done in future work?

# References

- Li, Y., Du, N., Bengio, S. (2017, July 31). Time-Dependent Representation for Neural Event Sequence Prediction. Arxiv. Retrieved January 12, 2023, from <https://arxiv.org/abs/1708.00065>
- Savsunenko, O. (2020, January 4). How-to encode time property in recurrent neural networks. Towards Data Science. Retrieved January 12, 2023, from <https://towardsdatascience.com/how-to-encode-time-property-in-recurrent-neutral-networks-friday-experiment-c14c39ba9755>