MS in Applied Data Science | Hackathon Fall 2023
Think Beyond the Batch: Event-Driven Data Science

Sponsored by Rotational Labs & The University of Chicago Data Science Institute.

INTRODUCTION

Did you know that some of the earliest deployed machine learning apps were event-driven? Back in the 90's, email spam filters used Bayesian models to learn on the fly. Each newly flagged spam message was a new training event, an opportunity to update the model in real time.

While machine learning training programs tend to prepare us to expect all data in batches, there are a TON of natural use cases for streaming data science (maybe even more than for offline aka batchwise modeling!). In this hackathon, we'll develop skills to "think beyond the batch" by tapping into some live data feeds.

Join us to develop an online prediction model using a publicly available streaming data API, River and Ensign. River is an open-source high-performance machine learning library that is particularly well-suited for online learning tasks, where new data is continuously fed into the model as it arrives. Ensign is a data flow and database platform that empowers modern data teams to create secure, customizable, and scalable data streams for real time apps, models, and analytics, no matter where data lives. Ensign has a native integration for River. We'll host an online tutorial to get you started.

Teams of up to three people are permitted.

DATA SOURCES

Participants may use open real-time data sources available on Rotational's Data Playground or find one a data source relevant to them.

REQUIREMENTS

What to Build

Build an online prediction model for a specific use case you care about. Participants must use a publicly available real time data API; a selection will be made available. Participants will also need to create a free Ensign account.

What to Submit

Submit your project on the hackathon Canvas site: Hackathon Fall 2023.

Participants will submit a short video of no more than 10 minutes, code, and any reference materials to canvas.

SCHEDULE

	Begins	**Ends**
Kickoff Meeting	Tues Oct 10, 4pm CST	Tues Oct 10, 6pm CST
Online Tutorial	Wed Oct 11, 12pm CST	Wed Oct 11, 1pm CST
Fieldwork	Wed Oct 11, 1pm CST	Wed Oct 18, 12pm CST
Office Hours	Fri Oct 13, 12pm CST	Fri Oct 13, 1pm CST
Office Hours	Tue Oct 17, 12pm CST	Tue Oct 17, 1pm CST
Submission	Wed Oct 18, 9am CST	Wed Oct 18, 5pm CST
Judging	Thu Oct 19, 9am CST	Tue Oct 24, 5pm CST
Winners Announced	Wed Oct 25, 5pm CST	Wed Oct 25, 6pm CST

LOCATION
Online

ELIGIBILITY
Open to all MS in Applied Data Science & MS in Analytics Students and Alumni in teams of three or less.

REGISTRATION
Register by Monday, October 9 at noon CST at https://forms.gle/YQwC9nPWYDirRerU6

PRIZES
- First Place: $900 total to the team or individual + Rotational/ UChicago DSI Swag
- Second Place: $300 total to the team or individual + Rotational/ UChicago DSI Swag
- Third Place: $100 total to the team or individual + Rotational/ UChicago DSI Swag
- Honorable Mentions: Rotational/ UChicago DSI Swag

JUDGES
- Benjamin Bengfort, Ph.D. Georgetown University faculty; CEO, Rotational Labs
- Rebecca Bilbro, Ph.D., Georgetown University faculty; CTO, Rotational Labs
- Prema Roman, Georgetown University faculty; Lead Distributed Systems Engineer & MLOPs Specialist, Rotational Labs
- Arnab Bose, Ph.D, Clinical Associate Professor, Program Director of MScA Online Program University of Chicago; Chief Scientific Officer, Abzooba
- Utku Pamuksuz, Ph.D.Clinical Associate Professor, University of Chicago; Chief Scientist, Inference Analytics

JUDGING CRITERIA

Problem Statement

Identify the main topic and sub-components.

Project Purpose

- Identify the primary objectives and purpose.
- Support the purpose with relevant context.
- Describe the data set(s) including imitations, unit of analysis, time window for feature, model engineering, validation, and development samples (as needed)

Methodology

- Describe the tools used to understand the relevant structure in data
- Provide a clear description of modeling framework including:
- Model selection justification
- Model descriptions
- Feature engineering/transformations
- Model engineering /development
- Validation methodology

Results

- Address research objectives and provide:
- Presentation synthesis
- Logical Assumptions
- Descriptive analytics
- Visual storytelling

Modeling Results

- Model Performance
- Validation Results

Conclusions and Recommendations

- Actionable recommendations
- Outlined limitations
- Considerations for next steps
- Proposed future extensions

SUBMISSION GUIDELINES

Submit your project on the hackathon canvas site: Hackathon Fall 2023

Participants will submit a short video of no more than 10 minutes, code, and any reference materials to canvas. We will post more details related to the hackathon canvas site next week.

To note, recorded videos may be publicly shared in marketing initiatives. By submitting a video, you are giving permission to Rotational Labs and University of Chicago to use the video or images from the video in future marketing materials.

- The video should be a brief introduction followed by a voiceover of the model, results, and recommendations.
- The video will be recorded over Zoom.
- Upload your recorded zoom video to Panopto. Panopto is a university-provided tool for hosting and sharing video. • Submit Panopto link, code, and related materials to the hackathon canvas site.
- A panel of judges will review the materials and winners will be announced via email.
- Please review these detailed submission instructions. This section walks through the process of recording your video, uploading it to Panopto, and submitting your entire project to canvas.

SAMPLE RESEARCH & BUSINESS QUESTIONS

- **Climate Change**: Weather and climate is constantly in flux, sending important signals about the state of the world. Design and deploy a model that continuously predicts weather, temperature, or other climate activity (like hurricanes or tornadoes).
- **Earth Metrics**: The earth is a dynamic system; thousands of earthquakes occur every day. Develop a model that predicts major earthquakes near highly populated areas. records, such as patient demographics, medical history, and examination findings?
- **News Feeds**: Communities are dynamic systems. Identify a community you care about and build a model that predicts a metric based on a collection of RSS news feeds or other data sources.
- **Transportation/ Logistics**: People and things are always on the move. Develop a model to predict arrival and departure times for a transportation system such as the CTA.