BIG DATA AND ANALYTICS ASSOCIATION & OHI/O

PRESENT

BDAA
BIG DATA & ANALYTICS ASSOCIATION

DATA I/O

OHI/O

PROUDLY SPONSORED BY

CAS®
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Nationwide

AMEND

CROWN

covermymeds®

CapitalOne

Participant Information Packet

October 3rd, 2020

Table of Contents

The Mission

Data I/O is a collaboration between the Big Data and Analytics Association (BDAA) and the OHI/O Program. This event aims to get participants more concentrated practice than a case competition, while being shorter and more data-focused than a general hackathon. Our goal is to create a day for students to collaborate with peers and meet new friends, as well as professors and professionals.

The Challenge

Teams of 1-4 will work on projects of their choice by bringing an open source dataset or using one of the suggested datasets. Prizes will be awarded by judges to the top projects, and mentors will be available to help throughout the day.

The Schedule

9:30 AM	**Check In On Discord**
10:00 AM	**Introduction**
10:15 AM	**Analysis Begins!**
12:00PM	**AMEND Tech Talk: Predicting Sports and the So What**
4:00 PM	**Presentations**
4:45 PM	**Judging**
5:15 PM	**Awards**

Virtual Platform Information

Discord
Discord is an instant messaging platform with video chat and screen sharing capabilities. Here is a link if you would like to test Discord or look at the virtual format before the event: https://discord.gg/gwwKtfU

*You may use the same link to connect on event day.

Check in for students and mentors begins at 9:30am. Once you check in VIA google form, the appropriate role will be assigned to you on Discord, and you will have access to all needed channels. The opening ceremony begins at 10am!

Please make sure your Discord username is in the format Firstname Lastname (ex. Brutus Buckeye). If you have used Discord in the past, you can change your username in the account settings.

Additionally, both voice and text chat rooms will be available all day to connect with mentors and other students. Our sponsors want to network with you, so please take advantage of these features!

Zoom
We will be using Zoom for all large group sessions such as the Introduction, Tech Talks, and Award Ceremony. All zoom links will be sent in Discord the day of.

If you have any questions or concerns regarding this format, please email Mannix.17@osu.edu

Challenge Areas

Teams have the option to follow two paths: a **free-form data challenge** (using publicly-available data), or a **sponsored data challenge** (next page).

Below are the available challenge areas / possible award areas for the free-form data competition. While a team's project may satisfy multiple of the following areas, a team will receive an award in *only one* of them. See *Judging Information* for more on judging criteria.

Overall and Runner-up Best Project	One team will be chosen to win the overall and runner-up prize, based on cumulative scores from the areas below.
Best Insight	Conduct a meaningful analysis that illuminates a particular subject area and paves the way for action.
Best Data Visualization	Show off your best visualization! Submissions may be a static image, an interactive website, or an RShiny app.
Best Presentation	Tell a story about the data you have selected that leaves the audience with a call to action.

Note: A team participating in a sponsored data challenge **is eligible** to win the awards presented on the previous page, but will be considered for the sponsored challenge first and **cannot** win an award in both.

Sponsored Challenges

In addition to the above challenge areas, two of our sponsors are **sponsoring special challenges**! One winning team of those that participate will be chosen for each challenge.

CAS® A DIVISION OF THE AMERICAN CHEMICAL SOCIETY Challenge

The world's patent offices share the same basic mission – to provide protection for innovation within their jurisdiction. To accomplish this mission, skilled patent examiners must carefully and accurately review each detailed, technical patent application document, a long and tedious task. As a not-for-profit division of the American Chemical Society that specializes in scientific information solutions, CAS is partnering with global patent offices to address this need.

As a patent examiner reviews an inventor's patent application (the target) they are primarily searching for previously published patents (citations) that will negate it because they show that the proposed invention is not new or is an obvious application of existing inventions. When a patent is denied due to another patent, the examiner lists the citations that are in conflict in their response to the inventor.

CAS's corporate challenge, should you choose to accept it, is for you to determine the differences between two populations of target-citation pairs. Dataset provided below.

CROWN Challenge

Forklift impacts (i.e. collisions) are a serious safety concern within warehouse operations. The telemetry data provided contains operational forklift information including operator authentications, safety checklists, forklift utilization data, and finally impact event data. We challenge you to look through the data and find new insights to what may contribute to impact events (e.g. impacts over time or failed inspection checklists to impacts).

Datasets

This event is structured as a **free-form data hack**, meaning that a team can use any publicly-available data, even in conjunction with data provided in sponsored challenges. Below are some datasets that can get a team started.

CAS Patent Data

Source: CAS

Are you up for the challenge? See the previous page for background on this data. CAS mentors will be available all day to assist you in your analysis.

Link to Data: Will be available 10/3/2020

Crown Forklift Impact Data

Source: Crown

What causes forklift impacts, a major safety concern in warehouse operations? Intrigued? Dig through this data and see what you can find! Crown employees will be available all day for any questions you may have.

Link to Data: Will be available 10/3/2020

NFL Play by Play 2009 - 2018

Source: Kaggle

The dataset made available on Kaggle contains all the regular season plays from the 2009-2018 NFL seasons. The dataset has 356,768 rows and 100 columns. Each play is broken down into great detail containing information on: game situation, players involved, results, and advanced metrics such as expected point and win probability values.

Interested in sports analytics? Check out AMEND's tech talk "Predicting Sports and the So What" at 12:00pm.

Link to Data: Will be available 10/3/2020

Judging Information

Each team will have **five minutes** to present their project to the judges, allowing for approximately **one minute** of questions. Teams' projects will be judged on the following three criteria, each with a **maximum score of 10 points**:

Insight (/10)

- How technically and statistically sound are the insights that you are providing?
- Were the assumptions and limitations of your models/algorithms discussed appropriately?
- Does your analysis offer relevant and well-thought-out recommendations for action?

Data Visualization (/10)

- Are you visualizing your data in a creative, engaging way?
- Is your data story easy to follow and accurately represented?
- Are your visualizations well-balanced (thought put into figure-ground relationship, spacing, colors, etc)?

Presentation (/10)

- How well are you presenting your insight and content to the judges?
- Are you able to cover all the content you had prepared in your presentation?
- Do you appear informed and enthusiastic about your topic of focus, based on your analysis?
- Is your team cohesive and cooperative?

After judging concludes, the following winning teams will be chosen: one **overall** winner, one **runner-up** winner, **one winner for each criteria,** and one winner for the **CAS Sponsored Challenge**.

Suggested Data Platforms and Tools

This event is structured as a **free-form data hack**, meaning that a team can use any publicly-available data, even in conjunction with data provided in sponsored challenges. Below are some data platforms that can get a team started, but note that they are not required to use these data platforms or tools.

Tools

Data Visualization

Want to make some super cool infographics with your data? Check out these industry leading tools if you would like to create beautiful visualizations without the programming hassle!

- Power BI
- Tableau
- Rawgraphs.io

Web scraping

Can only find the data you want to analyze/visualize on your favorite website? Use web scraping to pull the data from the website for further analysis.

- Python tutorial
- R tutorial

Additional Data Platforms

Kaggle

Kaggle is an online community of data scientists and machine learners, run by Google. Kaggle allows users to find and publish datasets, in addition to participating in competitions to solve data science challenges.

data.world (search data)

data.world is home to the world's largest collaborative data community, which is free and open to the public.

Data.gov

The home of the U.S. Government's open data. Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

Presentation Tips

- The presentations will take place in the team voice channels in Discord. Test your screen sharing and microphones before the judges enter your room.
- Any format is acceptable, but we recommend a PowerPoint or some other presentation format.
- Include visuals! Pictures, graphs, interactive visualizations, and everything in between make your presentation more appealing and memorable.
- Clearly explain your process. Why did you choose the dataset you worked with? Why did you choose the software you worked in? How did you come across any interesting insights?
- Be prepared to answer questions from the judges.

While it is not required, we encourage all teams of all skill sets to present their findings. No matter how “complete” your project is, it is worth sharing. Plus, you should never pass up the chance to win such awesome prizes!