MAIS 202: Deliverable 1
Team 1
February 17th, 2025

1. Choice of dataset.

The goal of this project is as follows:

Within each scientific discipline, there are topics that become trending subjects of research. Some of these trends stick around for extended periods of time (e.g., large-language models in the past decade) and become major avenues of exploration for scientists. Other subfields die out when the number of publications in the field reduces over time (like Survival Analysis or Complex Analysis, both subfields of math).

This project aims to **predict the shifts in popularity** of subfields within various disciplines, namely **Mathematics, Computer Science, Physics, Biology, and the Health/Medical Sciences**. To quantify the growth or death of a field, we will primarily rely on **publication numbers** and **citation counts**.

Use cases include

- Young academics and professionals: Helps in deciding where to focus their research efforts.
- Investors and government officials: Informs decisions on where to allocate research funding.
- General public: Provides insights into trends in scientific research.

Data Source: We will be building our own dataset by querying the ArXiv, bioRxiv, medRxiv, and PubMed APIs. We will be extracting publication metrics (including publication counts, citation counts, years published, and likely more) for keywords associated with various subdisciplines of fields that upload manuscripts to our data sources. **We may supplement these by with the Semantic Scholar API, other APIs or web scraping.**

ML Models

- **Time series models**:
  - **Prophet**, **ARIMA** – To predict trends in publication numbers over time.
  - **Long Short-Term Memory (LSTM) networks** – For modeling sequential dependencies in publication trends.

2. A high-level overview of the project is given above to contextualize the data collection method.

   a) Since we'll be collecting our own data preprocessing will mostly occur during the phase of building the dataset itself. Since we will likely be working with fairly large quantity of data, we should be able to discard data with detrimental missing values. Moreover, the arXiv API and similar APIs from which we're collecting data seem to have well-populated data.

b) Since the project is a matter of time-series forecasting, we are considering using applicable models. In particular, we are considering ARIMA, Prophet, and LSTM.

c) As mentioned above, our goal is to track the number of released publications and citations over time. We can evaluate the performance of our model by comparing the predicted number of publications and citations with the actual ones. Here are some usual evaluation metrics for Arima, Prophet and LSTM networks:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

For **bigger fields** with a lot of publications, the **MAPE** evaluation metric would probably indicate the most realistic and interpretable performance of our model, as the other metrics would inflate the amplitude of the actual error, because we are dealing with potentially bigger numbers.

For **smaller field**, the other evaluation metrics mentioned above would be more suitable, as a small variation in the number of errors greatly changes the percentage error if we consider a small number of publications.

We are not aware of any previous research attempting to make the predictions we are. Therefore, it is hard to compare our results to a baseline.

3. **Application:**

Our final web app will be a dashboard-like interface where users can select a predefined field of science (math, physics, CS, biology, neuroscience, etc.) and from there select a subfield from a pre-determined list. The web app will return metrics based on the model (like predicted average publications per year) and a classification of the subfield into one of ("growing", "maintaining", or "dying") based on the prediction of the model.