Before you begin...

- You'll be meeting virtually with your assigned student team on or around **Thursday, August 14th.**

- This template deck is what **we ask all Challenge Advisors to fill in and be ready to present during this first meeting.**

- Your assigned student team will have specific questions about the project related to their **Project Brief and Workplan assignment** (due to Break Through Tech on September 7, 2025). This Challenge Project Overview Template provides dedicated space to help you answer many of these questions (e.g., project milestones, preferred project workspaces).

Advice from former students: How to scope out and present your project

- Give us **structure, clear goals, and a timeline** to work toward at the outset of the project. Open-ended prompts can be hard for us since we're used to college course assignments.

- Provide **guidance on the initial steps** for us to take. Don't assume we know what to do, even if it's obvious to you! Some of us are new to Python and data science.

- Provide us with a **useable dataset** that (a) includes proper documentation and (b) doesn't require so much cleaning or pre-processing that it prevents us from getting started.

- **Help us anticipate potential challenges** during later phases of the project (and, when the time comes, help us tackle them by pointing us in the right direction).

- Suggest **resources** for us to better understand the problem space and possible approaches. What would **you** have found helpful as an undergrad?

YouTube Viral Video Longevity

Google

BREAK
THROUGH
TECH

August 2024

We’re excited to be your Challenge Advisors!

Woon Ket Wong (He/Him)
Google
Software Engineer
woonketw@google.com

Company overview

Nestled in sunny Mountain View, California, Google is a global tech titan where innovation and fun collide. Born in 1998 from the bright minds of Larry Page and Sergey Brin, this powerhouse now boasts over 190,000 employees, all fueled by a shared passion for pushing boundaries. From its iconic search engine that connects us to the world's knowledge to cloud computing and cutting-edge software, Google's playful spirit infuses everything they do. Their sprawling campus is a testament to their creativity, with slides, bikes, and even nap pods to keep the inspiration flowing. While they may be a tech giant, Google's heart remains young, constantly exploring new ways to make the world a more connected and exciting place.

A problem isn't truly solved until it's solved for all. Googlers build products that help create opportunities for everyone, whether down the street or across the globe. Bring your insight, imagination and a healthy disregard for the impossible.

AI Studio Challenge Project Overview

CHALLENGE SUMMARY

Build a machine learning model to predict which YouTube videos are likely to become viral or trending. The model should consider early engagement metrics, video metadata, and potentially external factors like news events or social media trends.

YOUR TEAM'S OBJECTIVE

- Clean and preprocess the data that will be used to train your model
- Extract relevant features like engagement features (views, likes-to-dislike ratio, share count) and content features (video length, keywords, topic categories)
- Train and Validate your model on your prepared dataset

DESIRED OUTCOMES

Present a machine learning model that can successfully predict a YouTube videos potential to become trending.

Business context

YouTube's recommendation algorithm plays a crucial role in driving video discovery and popularity. By understanding the factors that contribute to a video's viral potential, YouTube can improve its recommendations and provide valuable insights to creators on how to optimize their content strategy.

Suggested ML approach

Unleashing the Power of Data to Forecast Video Success

Problem Type

- This YouTube video forecasting project is primarily a supervised learning problem, where we aim to predict a categorical outcome (viral or not) based on various features. The machine learning techniques involve either regression (predicting a continuous probability of virality) or classification (predicting whether a video will be viral or not).
- The problem may also utilizes Natural Language Processing (NLP) to analyze text data and potentially deep learning for more complex pattern recognition.

Recommended Algorithms/Models

- **Start Simple**
  - Logistic Regression: Easy to interpret, good baseline model.
  - Decision Trees/Random Forests: Handles non-linear relationships, feature importance insights.
- **Explore Advanced Technique**
  - Gradient Boosting (XGBoost, LightGBM): Often top performers in competitions.
  - Neural Networks (Multi-layer Perceptron, CNNs, RNNs): Can learn intricate patterns, especially with large datasets.
  - Transformer Models (BERT, GPT): Powerful for NLP tasks, may improve analysis of video metadata.

Rationale for Recommendations

- **Start with interpretable models** to understand feature relationships and gain initial insights.
- **Progress to more complex models** if performance is unsatisfactory, leveraging their ability to capture intricate patterns.
- **Consider deep learning** if you have a large dataset and computational resources, as they can excel at handling high-dimensional data and complex relationships.
- **NLP techniques are essential** for extracting meaningful information from text-based features.

Model Evaluation Metrics

- **Accuracy:** Overall proportion of correct predictions.
- **Precision:** How many of the predicted viral videos were actually viral.
- **Recall:** How many of the actual viral videos were correctly predicted.
- **F1-score:** Harmonic mean of precision and recall, balances both metrics.
- **AUC-ROC Curve:** Visualizes the model's ability to distinguish between classes across different thresholds.

Tip: Choose the best model based on your specific dataset and requirements. Experimentation and iteration are key to success!

Data overview

- The following link contains a dataset of daily trending YouTube videos collected over several years.
  - https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/data
- If you want to gather the data yourself, Google requires a credit card on file to obtain your own API key. We don't expect you to do this. If there's data that you would like to access via an API call, feel free to reach out to me, and I will try to make the API call for you.

Python libraries

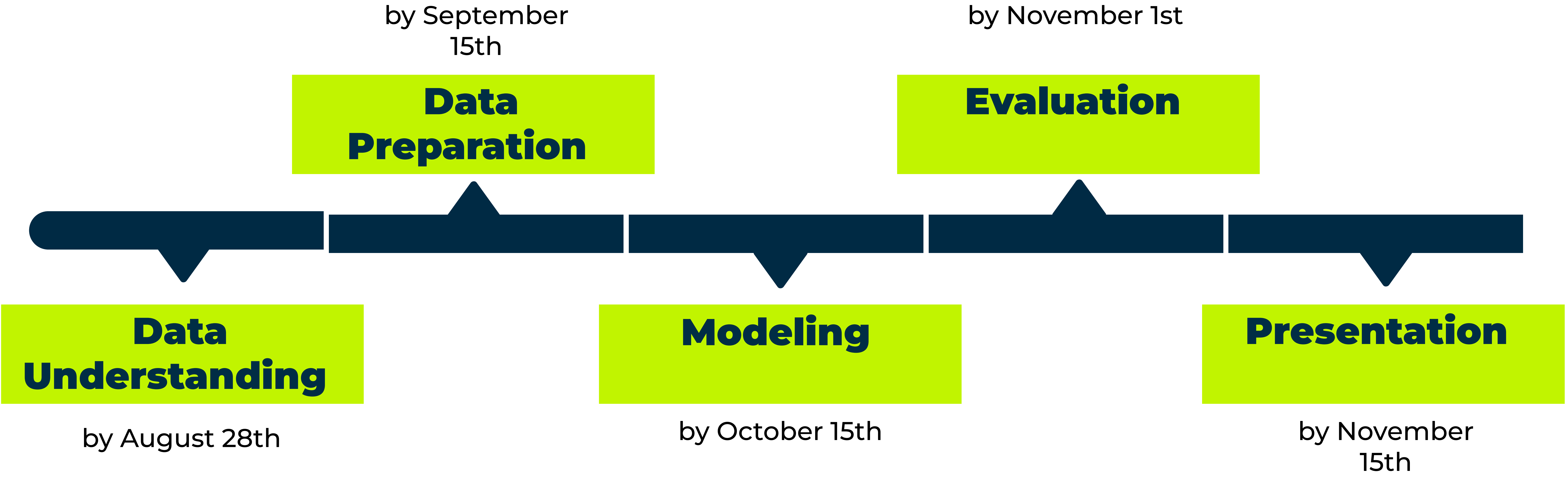
- NumPy: The fundamental package for numerical computation in Python. It provides powerful tools for working with arrays and matrices.
- Pandas: A library for data manipulation and analysis. It offers data structures like DataFrames, making it easy to load, clean, transform, and explore your data.
- Scikit-learn: The go-to library for classical machine learning algorithms. It provides a simple and consistent interface for a wide variety of tasks like classification, regression, clustering, and dimensionality reduction.
- PyTorch: Another popular deep learning library, known for its dynamic computation graphs and flexibility. It's favored by researchers and for projects that require more customization.

Suggest tools and workspaces

- Google Colab: A free cloud-based Jupyter Notebook environment that provides access to powerful GPUs and TPUs, making it easy to train large models without needing expensive hardware.
- We suggest using GitHub for code collaboration.
- Notion is a great tool for project management. There are many premade templates that you can use to plan your project.

Project milestones and timeline

These are the milestones for your Challenge Project. They are roughly aligned to the CRISP-DM process you learned about in your ML Foundations course.



Data Understanding

by August 28th

Data Preparation

by September 15th

Modeling

by October 15th

Evaluation

by November 1st

Presentation

by November 15th

MILESTONE #1 TASKS

Data preprocessing

- Collect initial data: Acquire the necessary data and (if necessary) load it into your analysis tool.
- Describe data: Examine the data and document its surface properties like data format, number of records, or field identities.
- Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
- Verify data quality: How clean/dirty is the data? Document any quality issues.

How we’ll work together this semester

Check-in meetings	● Provide a brief overview of your progress, highlight any key accomplishments, challenges or roadblocks encountered, and discuss questions if there are any. ● We’ll discuss the milestones and priorities for the next meeting.
Reporting	● Weekly in offline communication, and biweekly in meeting.
Communication	● The best way to reach Woon Ket are by email. You can expect a response within 48 hours.
Tools and platforms	● GitHub ● Google Colab ● Notion
Other project norms	● Weekly planning session where the team breaks down executable tasks that can be accomplished within 3 days.

How to get started

Here's what I suggest for your immediate next steps. I'll follow up on your progress and help address any challenges in our next check-in meeting:

Review these slides and note down questions

I'll email you a copy of this deck. Review it as a team and note down any questions you'd like to discuss in our next meeting.

Complete your "Project Brief and Workplan"

Continue working on your Project Brief and Workplan assignment, which is due next month. We'll review it again in our next meeting.

Research dataset and set up tool

Attain and research data set. Also set up your dev environment.

Questions?

What questions do you have?

Anything I can help clarify?

What are you most excited about?

Anything you’re unsure about?