Advanced Bracketology 2018

Companion Guide for the Event

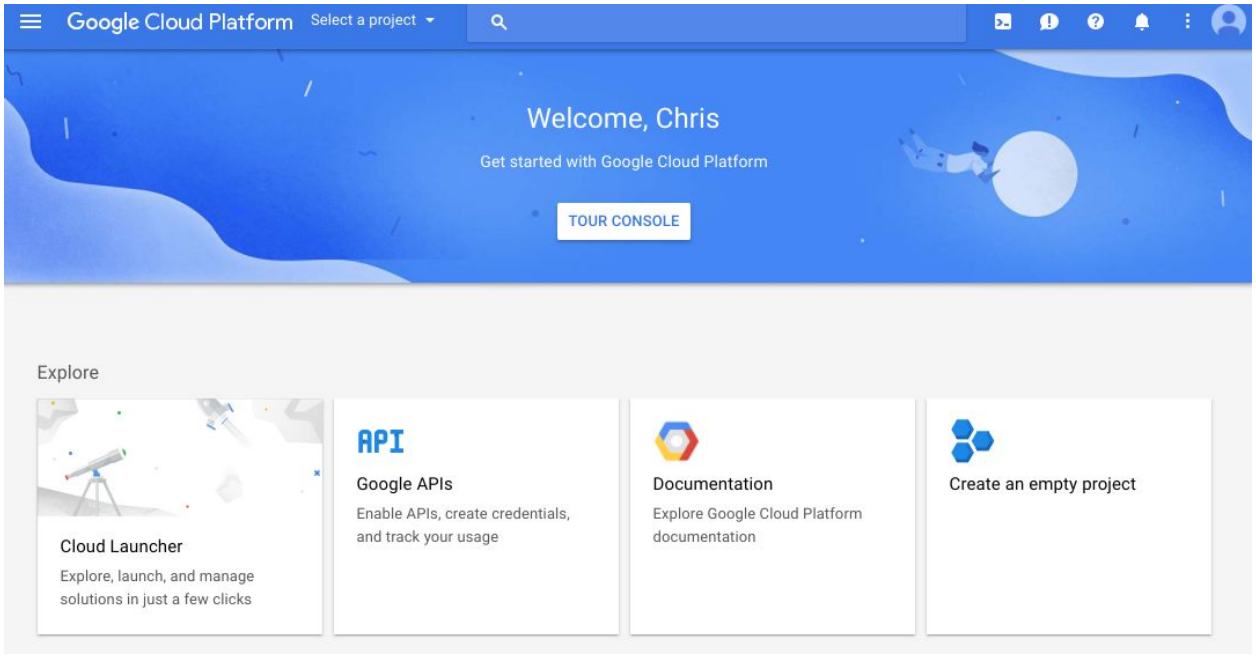
Welcome! We are so excited to learn with you today.

Contents of this Guide

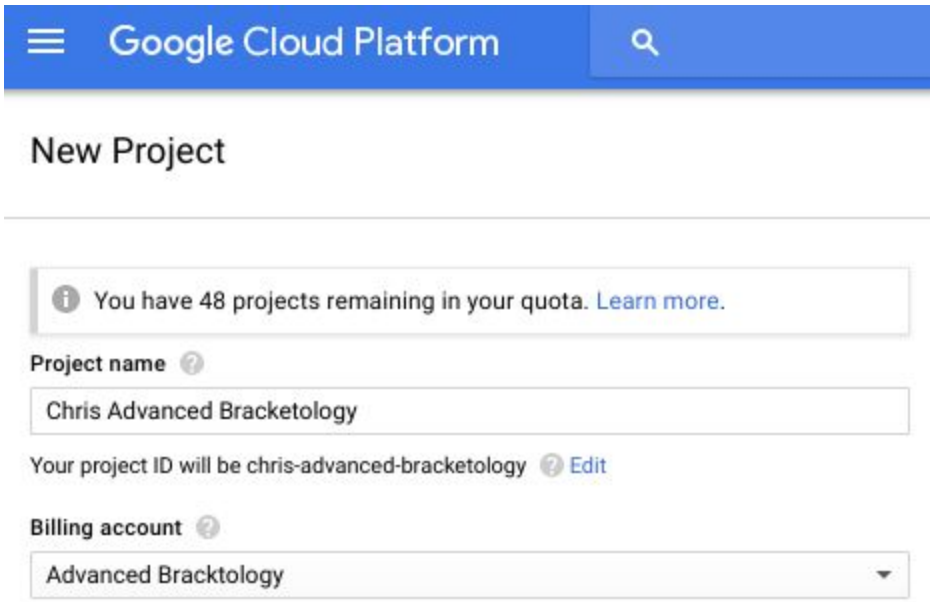
Before the event - set up

Before we get started, please follow the steps below before the event starts. Create a Google Cloud Platform (GCP) account and accept the terms of service for the Kaggle competition now so you can practice on the data during the event.

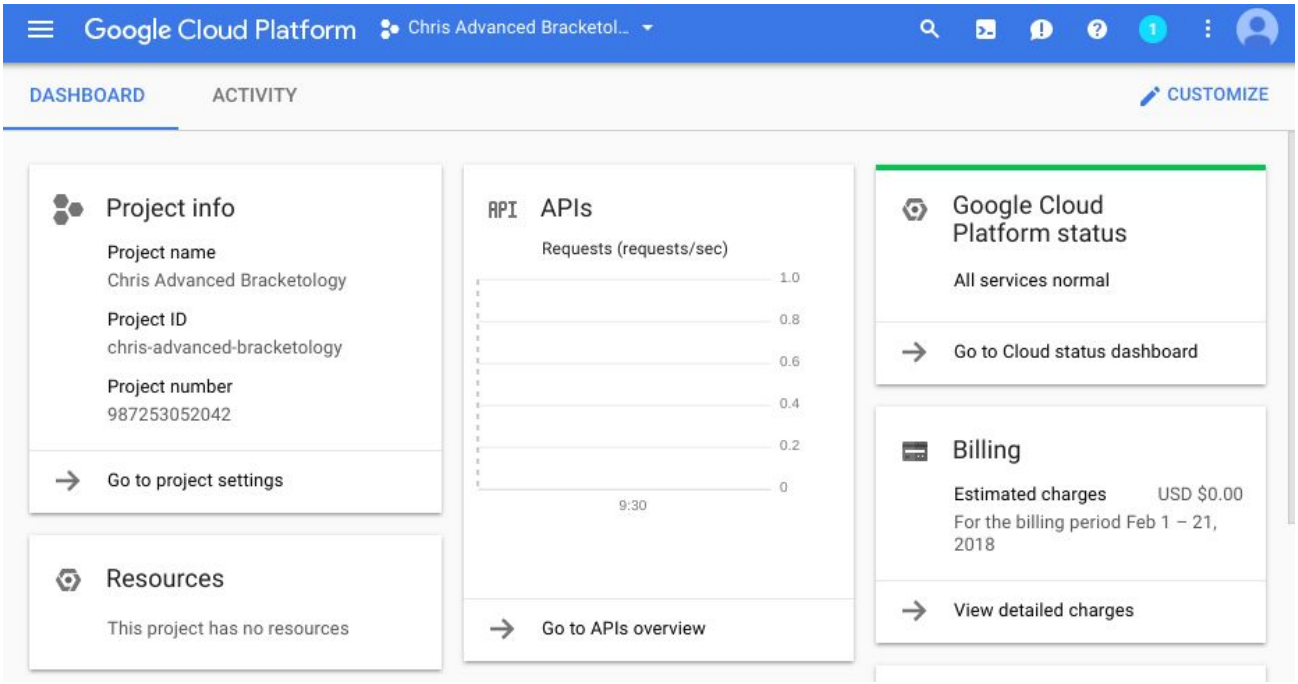
1. **Setting up your Google Cloud Platform environment**
   - **a.** Get a Google Cloud Platform (GCP) coupon for $50 at the front of the room.
   - **b.** Go to https://console.cloud.google.com/education and **sign in using your personal gmail.com account, NOT your .edu email**
   - **c.** Enter your coupon code, accept Terms and continue.
   - d. It will say “Creating project...may take a moment”
   - e. Select “Create an empty project”



i. Give it a name such as “<your first name> Advanced Bracketology”
ii. Select “Create Project”



f. You should then see a screen like below. Otherwise, use the project selector at the top blue bar to select your project.

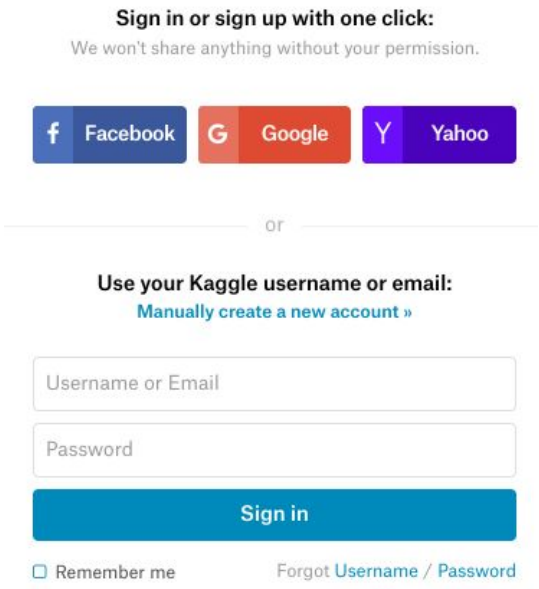


2. Setting up on Kaggle

a. Go to kaggle.com - select “Sign In” in the top right corner blue button

b. Sign up using a Facebook, Google, or Yahoo account or create a custom username/email.

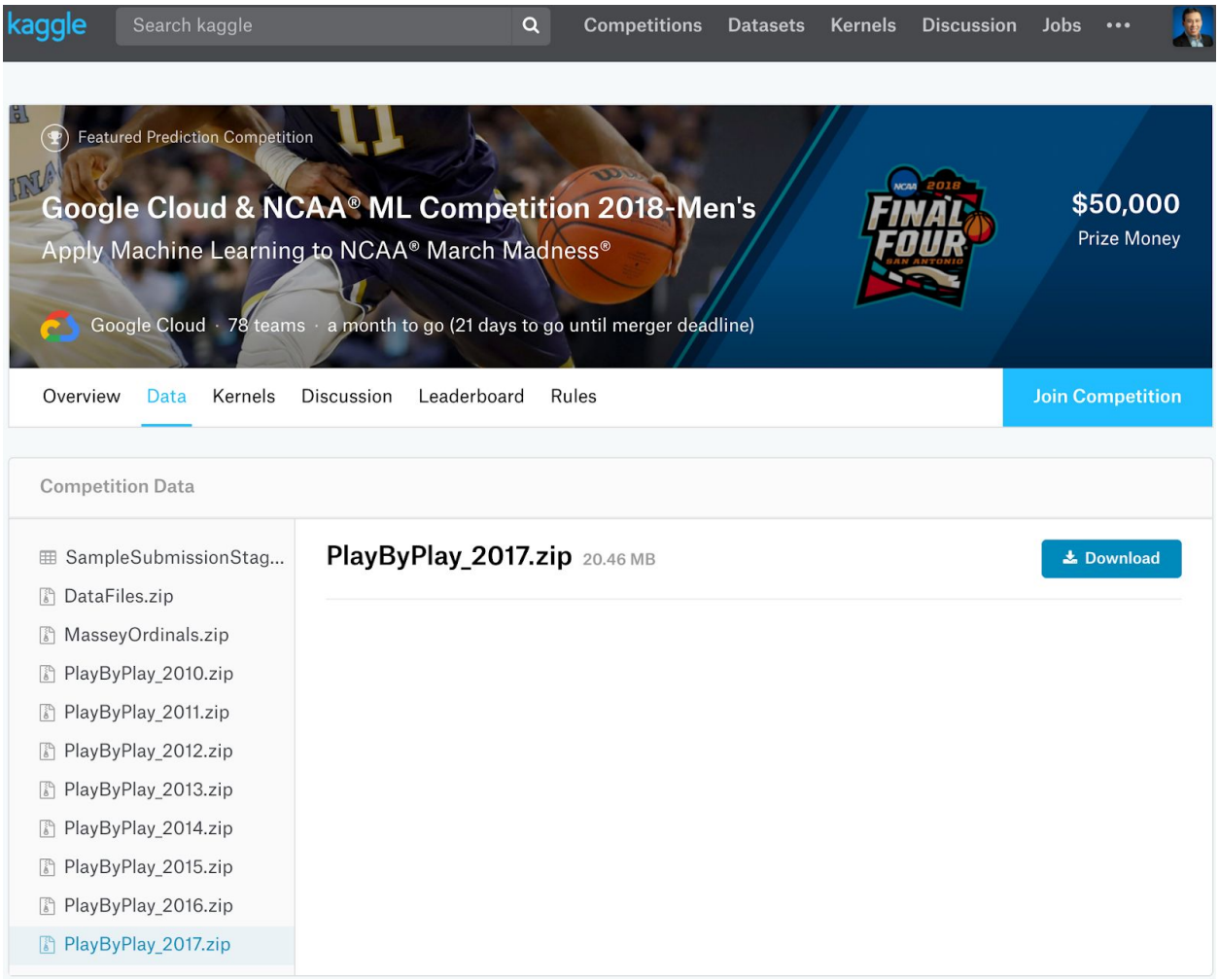
   i. To sign up manually, fill out your username, display name, email address, and password. Accept terms and privacy policy.



   ii. Complete registration through a verification email. Click “Activate”.

   iii. You are now a Kaggle member!

c. Download the data
    i. Navigate to the competition pages: Women's | Men's
    ii. Go to "Rules" tab, accept the rules at the bottom. You may need to verify the account using an SMS message to your phone.
    iii. Click "Data" tab and download relevant .zip files containing the .csv's
        1. We suggest downloading the `Men's PlayByPlay_2017.zip` for the purposes of our workshop



During the event

Vocabulary

- **Big Data**
    - Extremely large data sets
- **Data science**
    - The science of drawing knowledge and insights from data sets, both structured and unstructured
- **Machine Learning**
    - Field of computer science that gives computers the ability to learn without being explicitly programmed
- **Bracketology**

- The activity of predicting the participants in and outcomes of the games in a sports tournament, especially the NCAA college basketball tournament

The Game

- Which factors help you predict the probability of winning a basketball game?
  - A list of recommended predictors
- What further reading can you do? Below are links to research articles about "bracketology:"
  - Can the NCAA basketball tournament seeding be used to predict margin of victory?
  - Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament
  - March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis
  - A method to March madness? Institutional logics and the 2006 National Collegiate Athletic Association Division I men's basketball tournament
  - A simple and flexible rating method for predicting success in the NCAA basketball tournament
  - Improved NCAA basketball tournament modeling via point spread and team strength information
  - Calculating Strength of Schedule, and Choosing Teams for March Madness

The Data

- Watch this clip to learn about the NCAA Data set - video here, written description here
  - **BQ/SQL Script for Denormalized tables**
    i. Supplementing video, here is a series of SQL scripts that can convert the competition files into a denormalized view.
    ii. These are the actual fully denormalized tables, if you'd like to save time. Download them here. Or query them directly off of the public BigQuery datasets (Women's | Men's)
  - (optional) Explanation of Massey Ordinals file - video here

BigQuery

- Open your BigQuery console (https://bigquery.cloud.google.com/) Press "Compose Query"

Query 1

1. Let's try a simple query to find out total rows in Wikipedia public BigQuery dataset!

Copy and paste this into "New Query" box

```
SELECT COUNT(*) FROM publicdata:samples.wikipedia;
```

- Select "Run Query"
- Delete query from box when finished

Query 2

2. Let's try a bigger one to find how many Wikipedia articles have your team's mascot in the title?
Copy and paste into "New Query". Run Query.

```
#standardSQL

# Demo processing 10 Billion Wikipedia records
SELECT
  language,
```

```
  title,
  SUM(views) AS views
FROM
  `bigquery-samples.wikipedia_benchmark.Wiki10B`
WHERE
  title LIKE '%SCHOOLMASCOT%'
GROUP BY
  language,
  title
ORDER BY
            views DESC;
```

Query 3

3. Clear the box. Let's try another and find out which Wikipedia article has been most vandalized?
Copy and paste into "New Query". Run Query.

```
SELECT title, COUNT(*) as revert_count
FROM publicdata:samples.wikipedia
WHERE wp_namespace = 0 and comment
CONTAINS 'vandalism'
GROUP BY title
ORDER by revert_count DESC LIMIT 10;
```

Query 4

4. Clear the box. Let's find out what were the most common names in the US based off of another BigQuery public data set of names from the US Census?
Copy and paste into "New Query". Run Query.

```
SELECT
  names_step_1.name AS names_step_1_name,
  names_step_1.gender AS names_step_1_gender,
  COALESCE(CAST(SUM(names_step_1.number) AS FLOAT),0) AS namesstep1totalpopulat_1
FROM [bigquery-public-data:usa_names.usa_1910_2013] AS names_step_1
GROUP EACH BY 1,2
ORDER BY 3 DESC
LIMIT 500
```

Machine Learning (ML) APIs

Google Cloud Platform offers many great APIs to pre-trained ML models. Complete individual tasks like text sentiment analysis & translation, speech transcription and image & video recognition using our pre-trained models. You can get a feel for how to use and apply Machine Learning technology without having to train your own model and can build several interesting projects by leveraging the existing APIs for Video Intelligence, Vision, Translate, Speech, and Natural Language.

Today we are focusing on Vision, Speech, and Natural Language. Try it yourself!

- Vision API demo
    - Open link
    - Select photo to analyze

    - Drag into the “Try the API” dotted box
    - See results!
- Speech API demo
    - Open link
    - Select language and click “Start Now” to record your speech
    - See your speech converted to text in realtime
- Natural Language API demo
    - Open link
    - Copy this tweet into the demo box and click “Analyze” to extract information, understand sentiment or parse intent.
        - Tweet: *Auburn on a run! No. 10 Tigers have made 7 of last 9 shots and hold 66-61 lead over Kentucky with 3:57 to play on ESPN2.*
        - Tweet: *PROVIDENCE PULLS THE UPSET! Five Friars reach double figures in 76-71 win over No. 3 Villanova.*

The Kaggle Competition

- **Getting Started with Kaggle.** What Kaggle is, how to sign up, how to navigate site, how to navigate the competitions page specifically - video here
- **Submissions & Scoring.** How the competition is structured, what makes it different from other bracket challenges, how the evaluation metric works, and how to make a simple submission from BigQuery (Google Cloud Platform) or Kernels (Kaggle) - video here, written version here
    - **BigQuery/SQL Submission**. Supplementing video, here is a SQL script that can be run in BigQuery to generate a simple submission file.
    - **Kernel Submission**. Supplementing the video, Starter Kernel on Kaggle that walks through a basic model with submission file output written in Python
        - Women’s Kernal | Men’s Kernal

Analyzing the data sets with GCP

Follow this step by step guide of how to analyze the NCAA data on GCP, specifically using BigQuery and Datalab.

- **Follow along here: Guide to analyzing NCAA data on Google Cloud (goo.gl/QZnBGA)**

After the Event

Additional GCP Resources

- Resources from our Hackathon Team!

Resources for applying to Google

- How to apply
    - Visit google.com/students to find the best role(s) for you.
    - Upload your resumé and transcript (unofficial is fine).
    - Click submit!
- Role information
    - Software Engineer, University Graduate
        - Deadline: N/A
        - g.co/swegrad
    - Software Engineer, YouTube

        - Deadline: N/A
        - google.com/students (search for YouTube)
    - Software Engineer, Tools & Infrastructure
        - Deadline: N/A
        - google.com/students
- Guides
    - Google's Guide to Technical Development: g.co/techdevguide
- Suggested texts
    - *Cracking the Coding Interview,* 6th Edition
    - *The Algorithm Design Manual*, Steven S Skiena
- Practice questions
    - Visit Project Euler for tutorials and questions
- Serious practice questions
    - Sign up for an account on leetcode.com and have your practice answers reviewed by others
    -
- Opportunities for higher ed learners interested in using GCP and in computer science more broadly
    - Google offers many programs, resources and opportunities for those interested in cloud computing, computer science, and GCP specifically. Follow our Keyword blog for the latest. Some resources to highlight are:
    - GCP for your classes - Your instructors can apply for GCP Education Grants to get free GCP credits to use in classes and with student groups. More info at cloud.google.com/edu.
    - GCP for start-ups - Our team offers mentorship, training and free credits for startups that want to use GCP. Learn more at cloud.google.com/developers/startups/
    - For CS & GCP for HIgher Ed - edu.google.com/highered and visit the pages on GCP and CS

Kaggle Competition Info

Q: What do I submit in the competition? Is it a bracket?
A: You don't submit a bracket. To enter into the competition you will submit a .csv of predictions for every possible match-up of the 64 teams selected into each tournament (Men's & Women's) E.g. Duke vs UNC, Duke vs UVA, etc. Your prediction file will contain the probabilities that the first team will beat the second team in every matchup. You'll include predictions for every matchup, even if they don't happen. You are scored using the logloss metric, calculating the sum of the negative natural logarithm of each prediction for the 64 games that do happen (the other predictions are ignored). More details on scoring are contained here in the Men's Division I Competition & the Women's Division I Competition. The data page for each competition contains the dataset & a sample submission file, as well.

Q: Is the Kaggle competition only open to university students?
A: No, the competition is open to anyone as per the terms at kaggle.com

Q: Can I apply as part of a team?
A: Yes! No limit on # of team members.