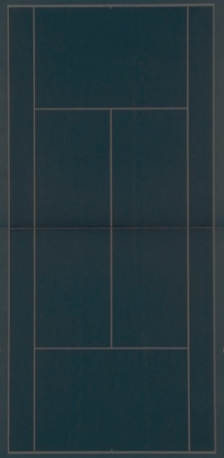
OLYMPIC PREDICTIONS

PRESENTATION OUTLINE



TEAM 12

JOHN TEMPLE

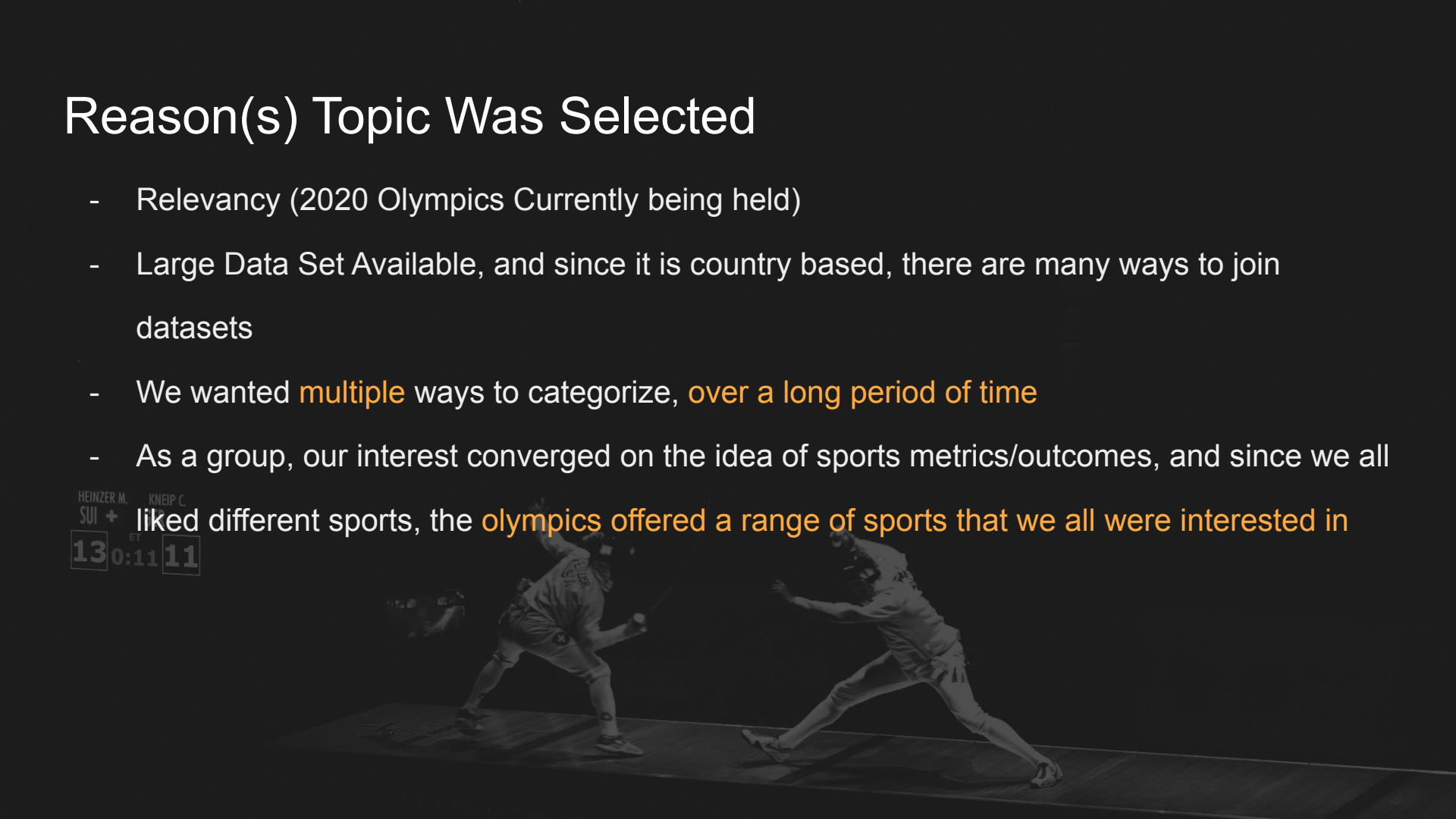
TREY GRAHAM

TUCKER FISCHER

HUNTER HUANG

Reason(s) Topic Was Selected

- Relevancy (2020 Olympics Currently being held)
- Large Data Set Available, and since it is country based, there are many ways to join datasets
- We wanted multiple ways to categorize, over a long period of time
- As a group, our interest converged on the idea of sports metrics/outcomes, and since we all liked different sports, the olympics offered a range of sports that we all were interested in



Description of the Source of the Data

Data Scraping

- Olympic Medal Counts by Year/Country
  - Olympedia – Overview
- Olympic Competition Results by Year/Discipline/Competition/Athlete
  - Olympedia – Results by Games

Olympic Data will be scraped using BeautifulSoup, stored in Pandas/DataFrames for cleaning, and exported out as a CSV

Curated Data

- GDP Data per Country per Year
  - TRADING ECONOMICS | 20 million INDICATORS FROM 196 COUNTRIES
- Historical Population Model
  - Historical population data and projections (1950-2050)
- Geographical Data of Countries
  - LM Nixon - Latitude & Longitude Of World Capital Cities

Curated Data will be stored in Pandas/DataFrames for cleaning, and joined with other CSV data using SQL

Questions the team hopes to answer with the data

What country-based metric has the greatest effect on medal count?

- GDP, Population, Average Income, Proximity to Games
- Average Height/Weight for each Gender

Which countries are most efficient in:

- Competition Based Metrics:
  - Getting medals per athlete
  - Getting medals per discipline/competition/gender
- Country Based Metrics:
  - Getting medals per GDP
  - Getting medals per Population
  - Getting medals per Proximity/Region

How will countries categorically fit?

- Based on both Competition and Country Metrics:
  - Countries with High/Low Entries and High/Low Medal Count
  - Countries that statistically specialize in certain disciplines
  - Countries that statistically get “spoil” (random) results most frequently

What is the prediction for Tokyo 2020?

Other Questions

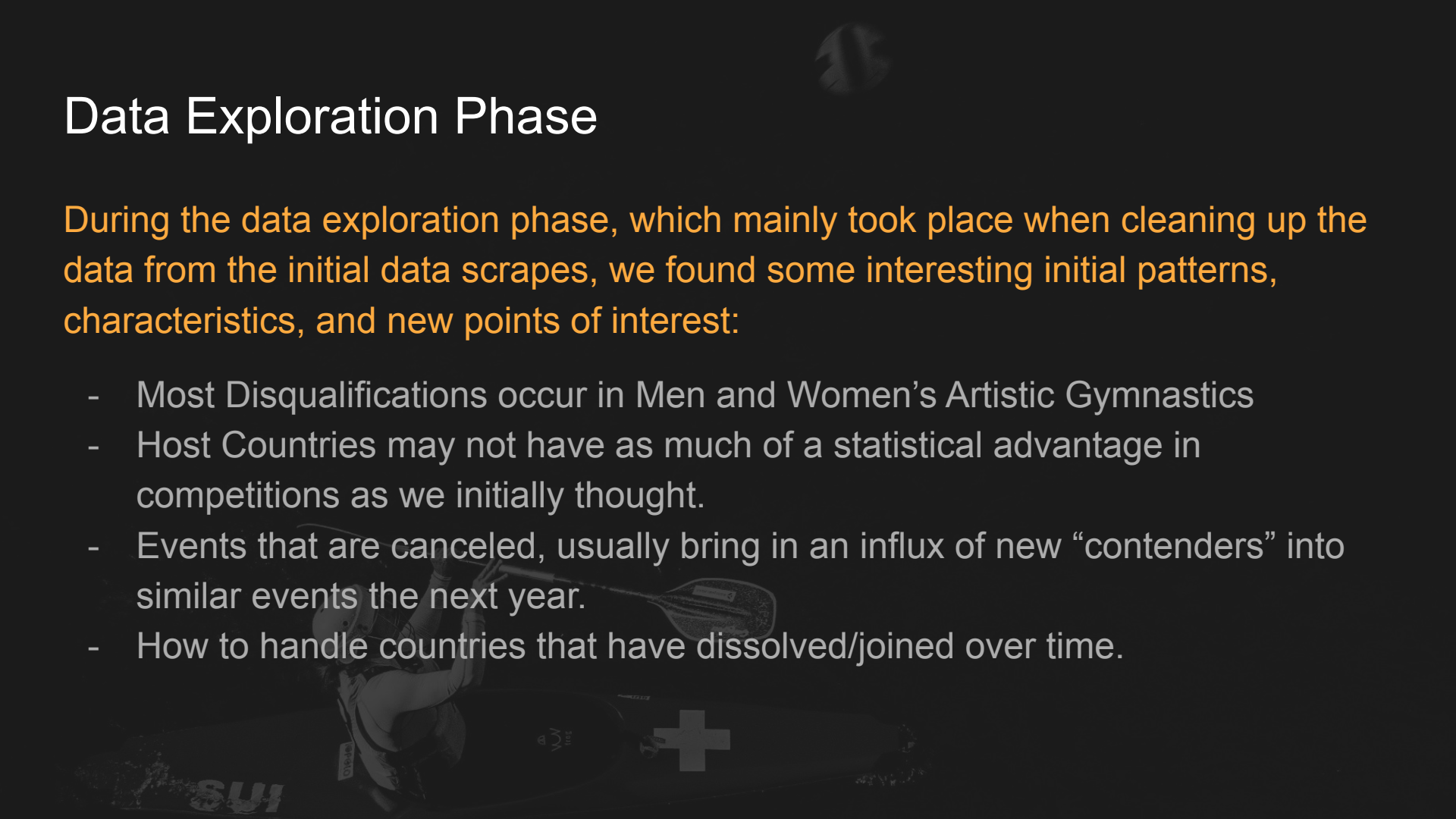
As we explore the data, our team has come up with more “fun” questions:

- How would a country consisting only of all disqualified olympic athletes look/perform?
- Which competitions offers the most random results? Least?
- What is the longest streak of medals for a competition?
- Who finishes second the most?

Data Exploration Phase

During the data exploration phase, which mainly took place when cleaning up the data from the initial data scrapes, we found some interesting initial patterns, characteristics, and new points of interest:

- Most Disqualifications occur in Men and Women’s Artistic Gymnastics
- Host Countries may not have as much of a statistical advantage in competitions as we initially thought.
- Events that are canceled, usually bring in an influx of new “contenders” into similar events the next year.
- How to handle countries that have dissolved/joined over time.



Analysis Phase

- For the analysis phase, our team is going to prioritizing analyzing the relationship of medal counts per year by country with the country data to see which is the most significant indicator of how well a particular country does in the olympics.

- We have discussed using both a weighted and non-weighted point system for medal counts when creating our model.

- We will use machine learning in two ways:
  - Categorize Countries based on their results in each competition discipline by year, and country demographic
  - Regression model to forecast the medal counts for the Tokyo 2020 Olympic games.

- For visualizations we intend on creating:
  - Heatmaps, to compare location of medal winners
  - Scatter plots, to compare country population of medal winners
  - Spider charts, to compare a country's performance in each of the disciplines