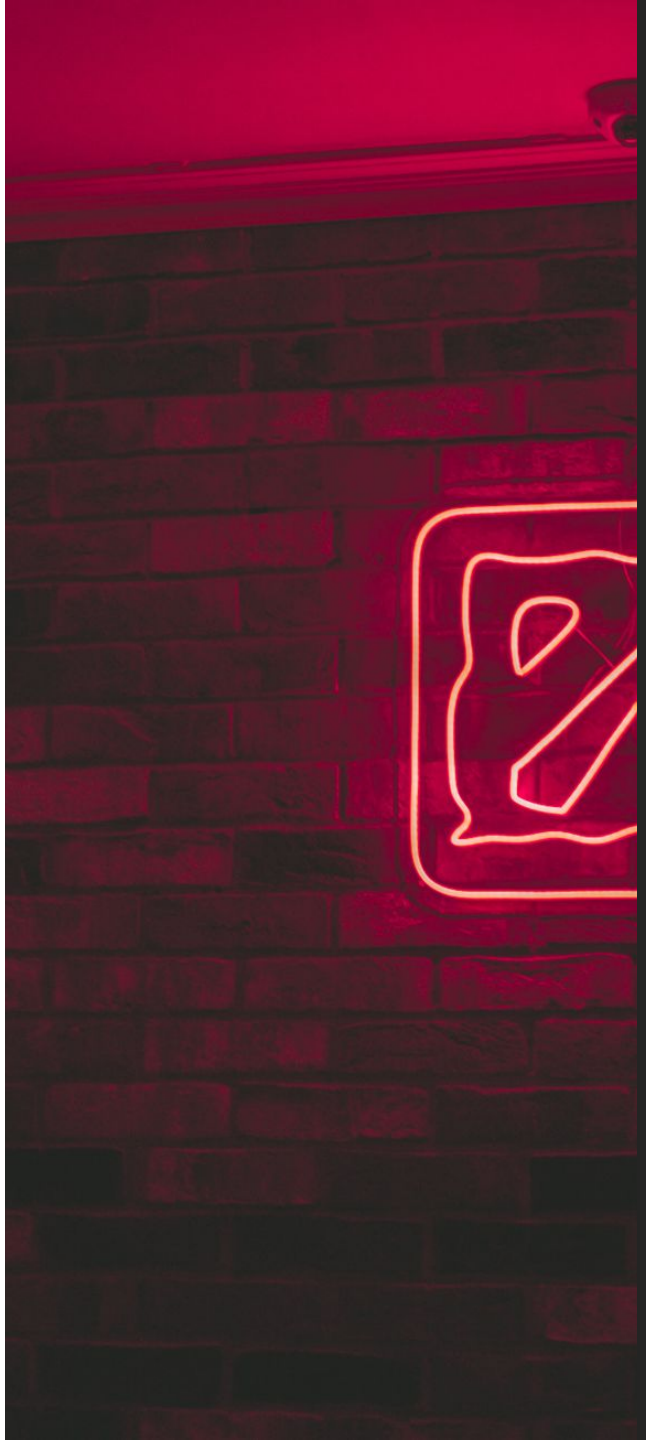


MSCA 31012DATA ENGINEERING PLATFORMS FOR ANALYTICS FINAL PROJECT

Analysis on Dota2: Democratizing OpenDota

TEAM 7: FLORA HUANG, SHUTONG LI, ZHANGCHI LIU, KAICHONG ZHANG

MEET THE TEAM

FLORA HUANG

yufeih@uchicago.edu

SHUTONG LI

shl636@uchicago.edu

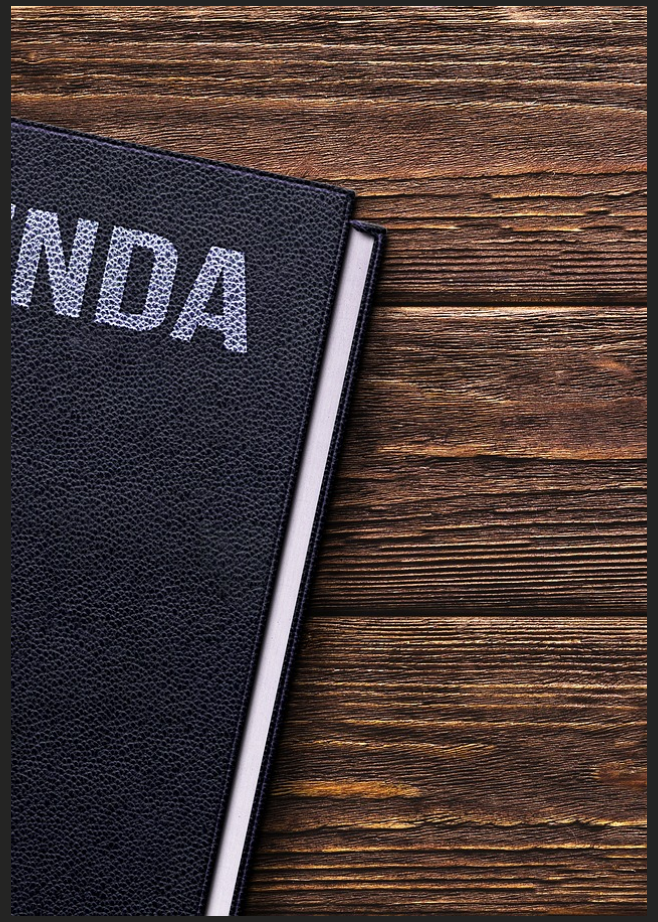
ZHANGCHI LIU

zliu4@uchicago.edu

KAICHONG ZHANG

kczhang@uchicago.edu

AGENDA



1 | EXECUTIVE SUMMARY

2 | DATA PROFILE

3 | DATA MODELS

4 | INSIGHT & VISUALIZATION

5 | CONCLUSION

EXECUTIVE SUMMARY

EXCLUSIVE SUMMARY

- DOTA 2 is a popular game worldwide, ranked the fourth most played game on Steam with 7.6 million monthly active users in 2020.
- While hundreds heroes, items, and abilities are created and updated regularly, it is often the case that the game is not balanced: some heroes are fundamentally stronger than others given their abilities and items.
- Therefore, we design a data model to store detailed statistics for every DOTA match that makes data extraction and analysis easier.
- Our goal is to promote players' performance and help designers to improve game balance via our ETL process.

How does this project help?

BUSINESS USE CASE

#1: INDIVIDUAL PLAYERS

- Sell personalized performance insight as product
- Promote community growth by increasing player competence.

#2: INTERNAL GAME DESIGNER

- Monitor balance (health) of the game
- Facilitate data-driven game design (i.e. balancing)

#3: E-SPORTS TEAM

- Provide pro-team with source of data insight
- More comprehensive team strategy
- Improved performance = better pro-scene

DATA PROFILE

DATA PROFILE

1 DATA SOURCE

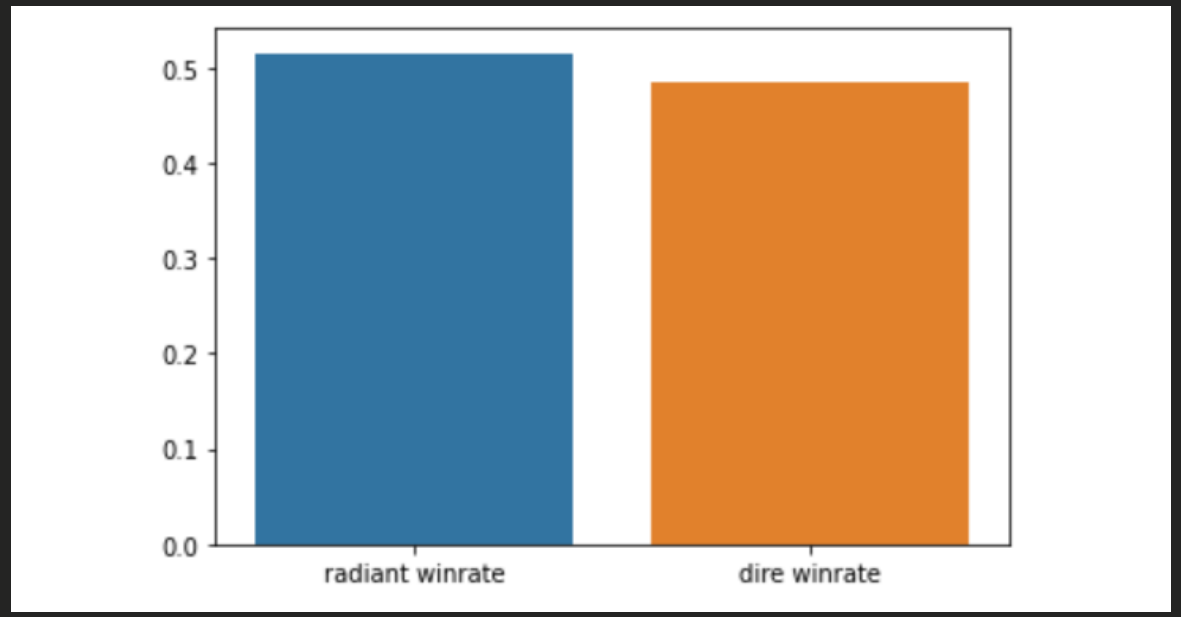
- **Single data source:** API (OpenDota) of the game through OpenAI
- **Initial data**: data dump from 2017
- **Size:**~= 14GB
- **Data Structure:** Semi Structured - JSON

2 FEATURES

- Highly nested, therefore hard to normalize
- Features of interest
- Match-level statistics (game duration, who won)
- Hero-level statistics
- Text corpus from chat data

EDA

Matching Statistics



radiant_win	duration
False	2480.235246
True	2374.944509

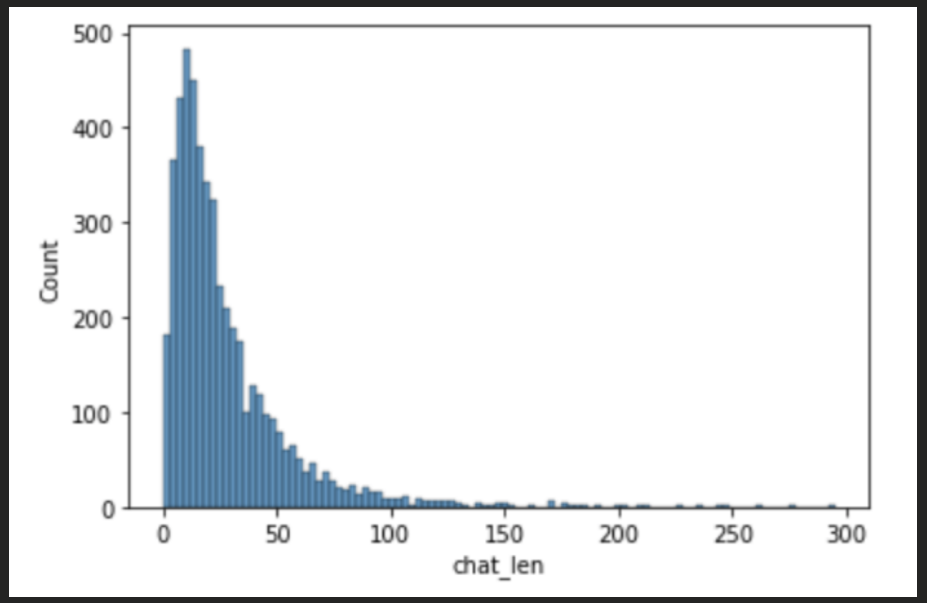
WIN RATE V.S. SIDES

Will being on one side of the map itself makes you more likely to win?

TIME DURATION V.S MATCH RESULT

Is the extra advantage offered by radiant, if any, from early game or late game?

EDA



Most matches offer a small corpus around one or two dozen chat messages

This game in particular has more than 200 messages and lasted >1hr. Seems a game won through split-pushing.

DATA MODELS

DOTA 2

DATA TOOLS

#1 COLLECT DATA

#2 EDA

#3 LOAD

#4 TRANSFORM

#5 STORAGE & QUERY

#6 VISUALIZATION

DESIGN CONSIDERATION

WHY DO WE CHOOSE NOSQL for OLTP?

- **JUSTIFICATION**
  - nested json objects
  - low redundancy in each entry
  - MUCH faster dev cycle (simple insert for data loading)
- **DATABASE**

  Singular database (named dota) with one collection (named matches)
- **COLLECTION INDEX INFORMATION**
  - base id and match_id

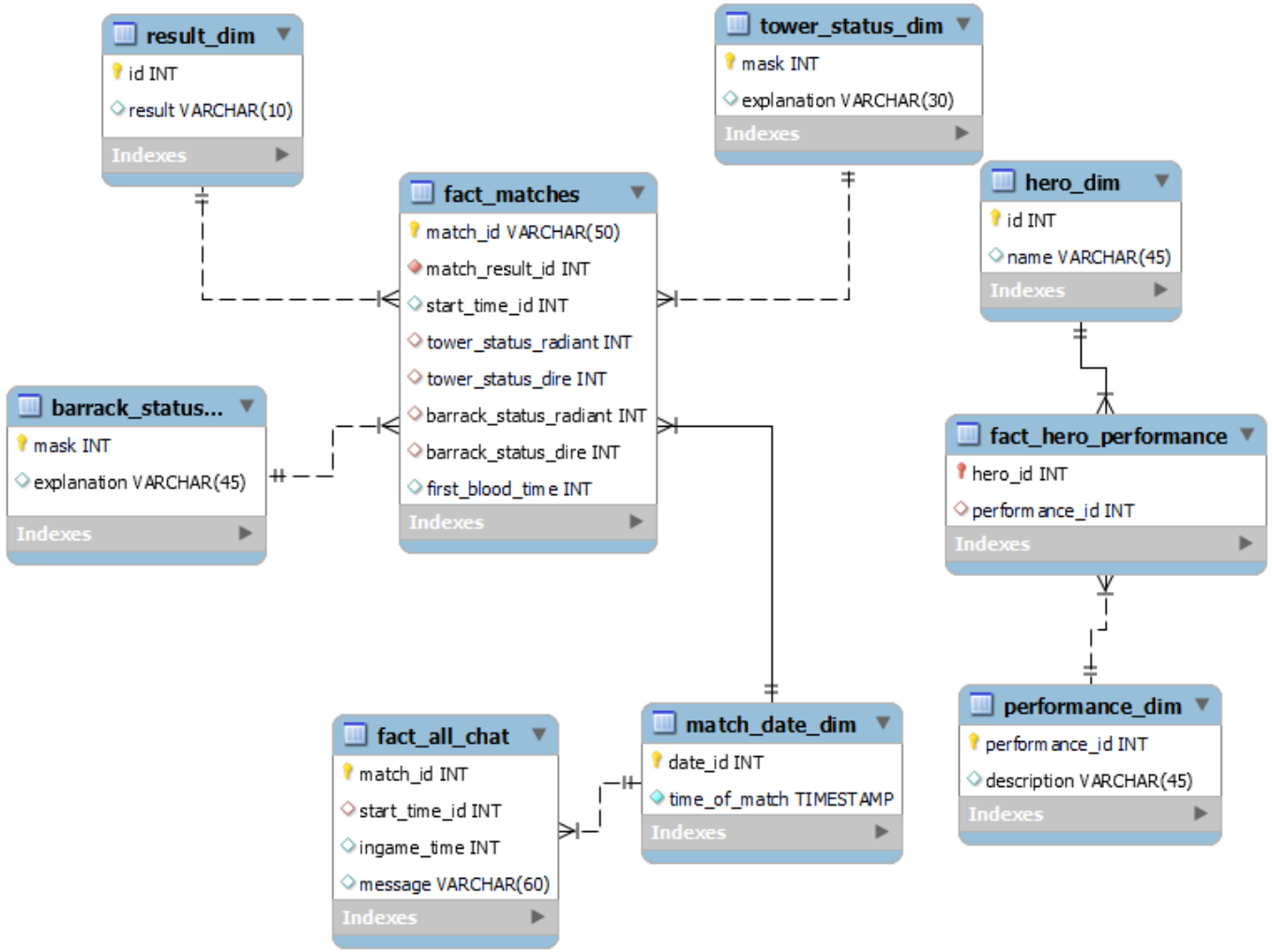
DESIGN CONSIDERATION Cont'd

What would the design look like on other platforms?

- **RELATIONAL DB**
  - Unpack/flatten nested values
  - 0-1 NF hardest to normalize
  - A lot of weak entity
- **NEO4J**
  - Node: Heroes, players, chat and matches
  - Players play matches, matches contains heroes and chat

OLAP MODEL

- GCP CLOUD SQL instance
- Connected through MySQL Workbench
- Figure on the right is the star schema for the analytical facts that we are interested in
- Foward Engineered into the Database

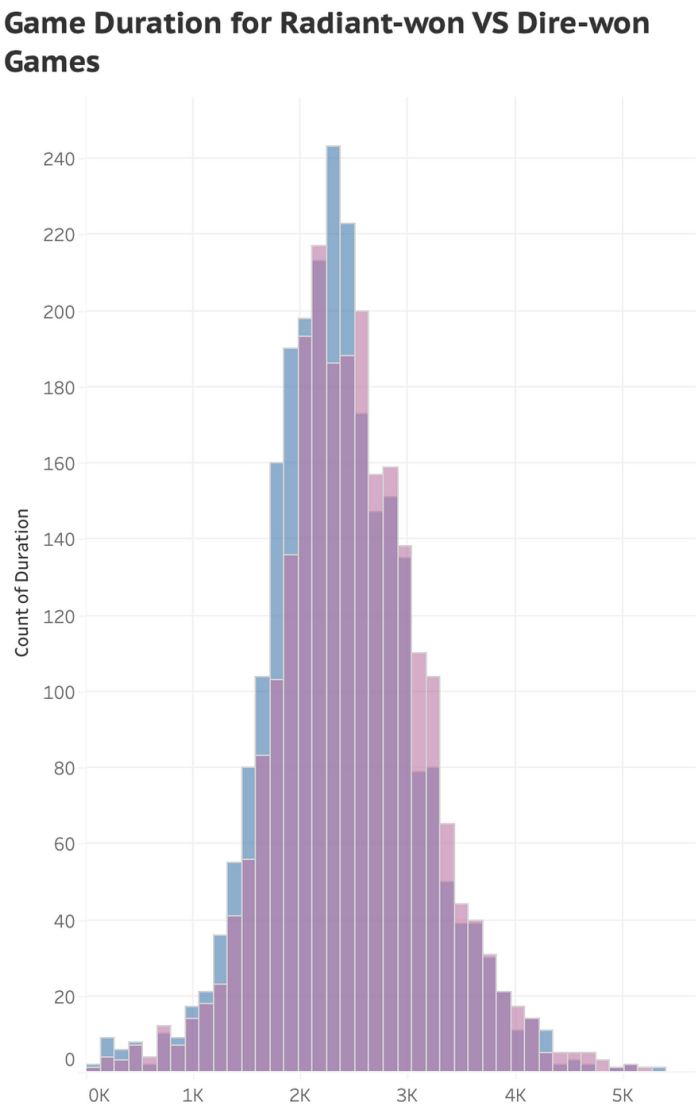


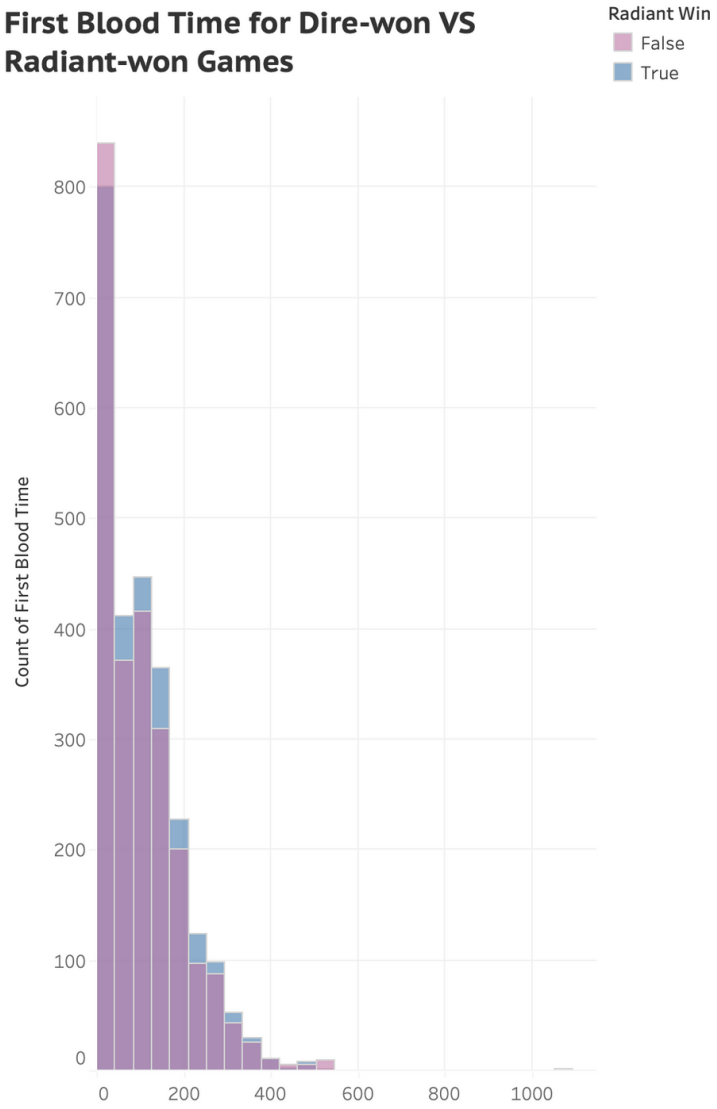
INSIGHT & VISUALIZATIONS

DOTA 2

Map-Side Advantage

- FROM EDA we know being on radiant side of the map has higher winrate
- The BI extracted from OLAP tells us:
  - Games won by radiant side ends earlier
  - First blood happens earlier for radiant-won games
- We can postulate that the advantage offered by radiant seem to come from *early-game*.

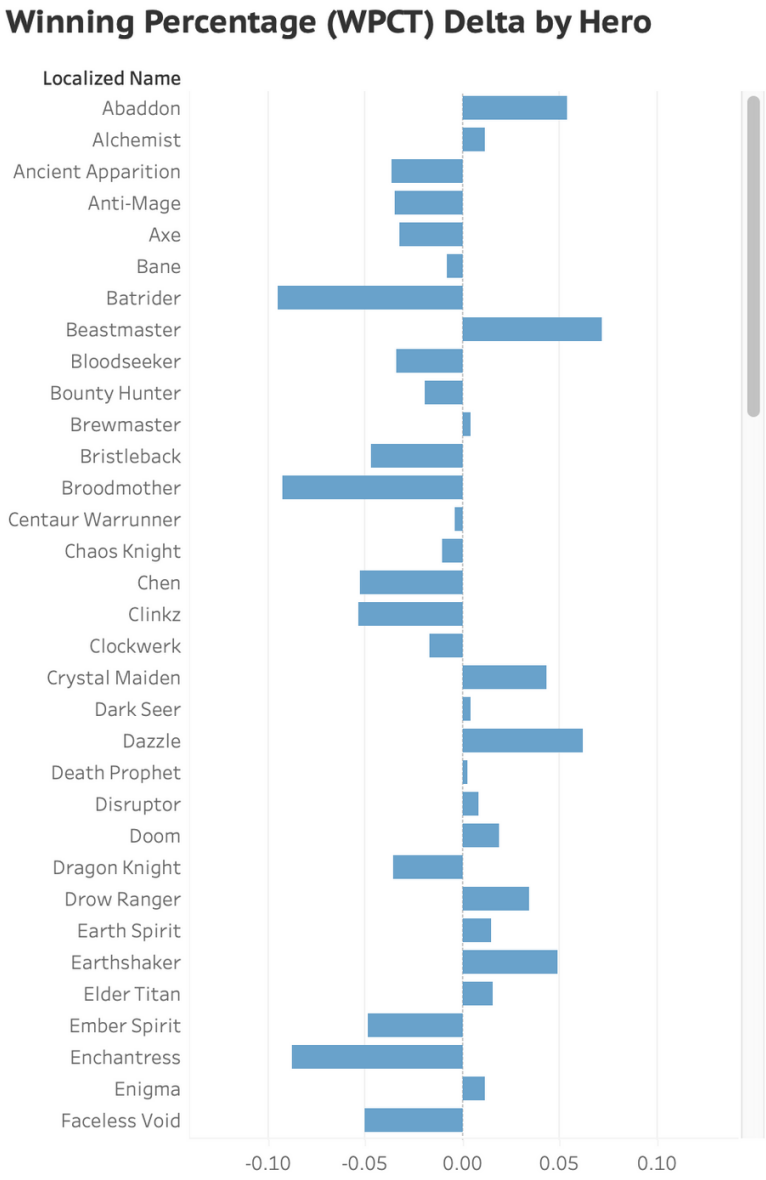


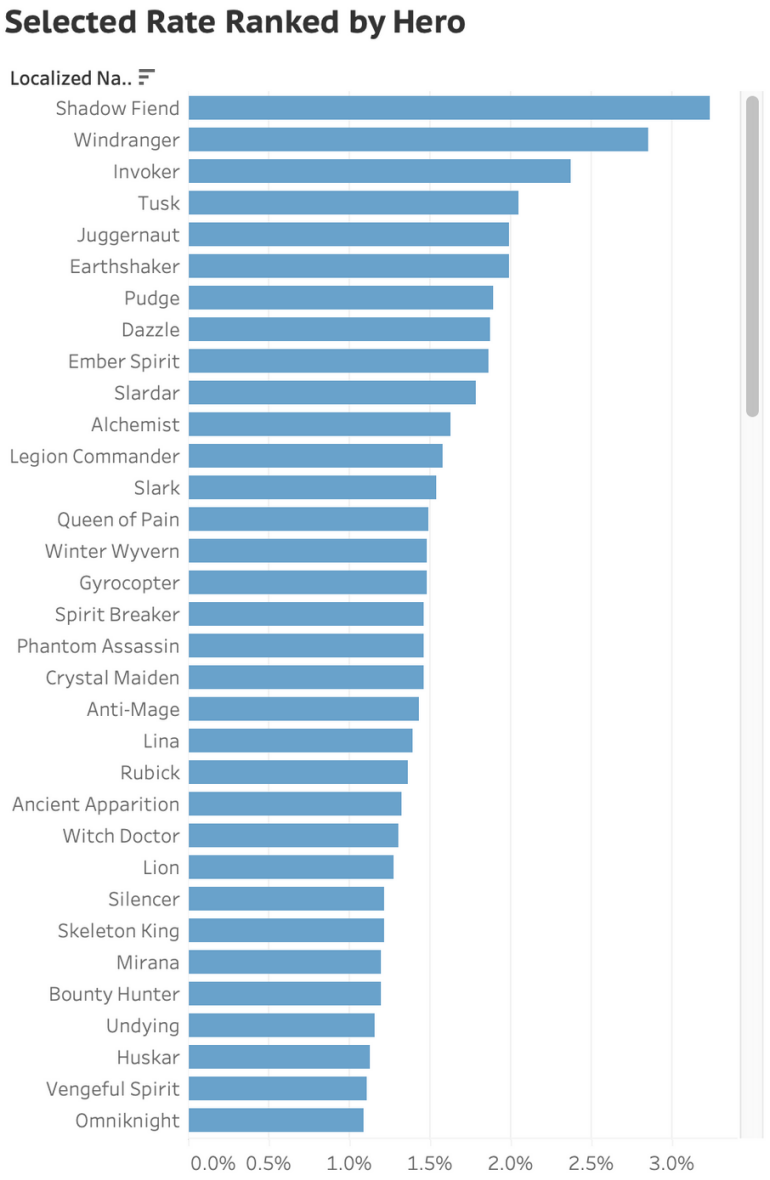


DOTA 2

Tracking Hero Performance

- Balance among heroes is essential to the game
- Our OLAP enables us to extract hero pick rate and winrate from player level data and hence track their perforamence
- Winrate and pickrate are the two often monitored feature

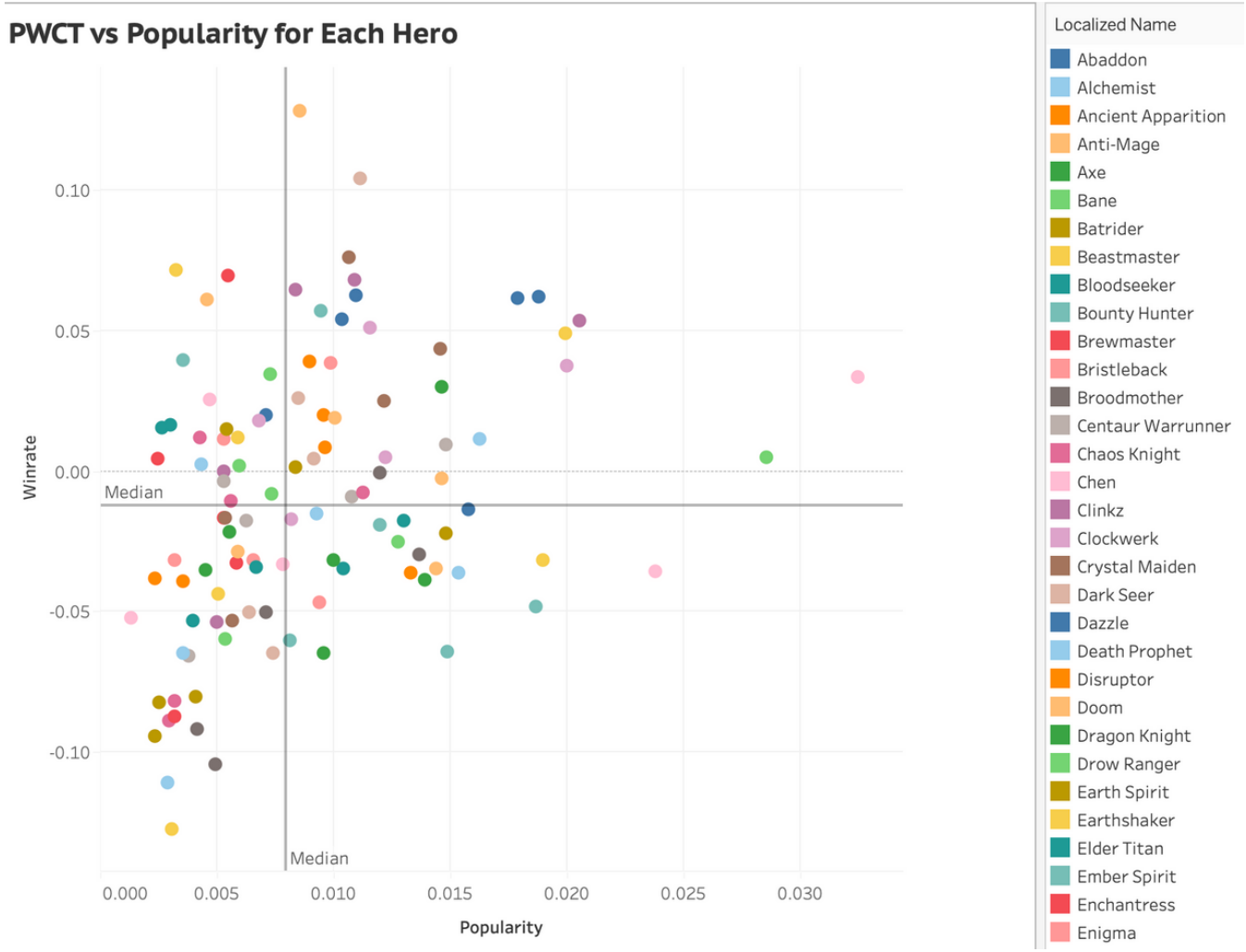




DOTA 2

Leveraging Balance with Popularity

- Winrate and pickrate often affects each other
- Combining performance with popularity. We can define heroes into four quadrant
  - popular but bad (norm for popular heroes)
  - popular but good (needs nerfing)
  - unpopular but good (situational picks)
  - unpopular and bad (needs buffing/rework)



DOTA 2

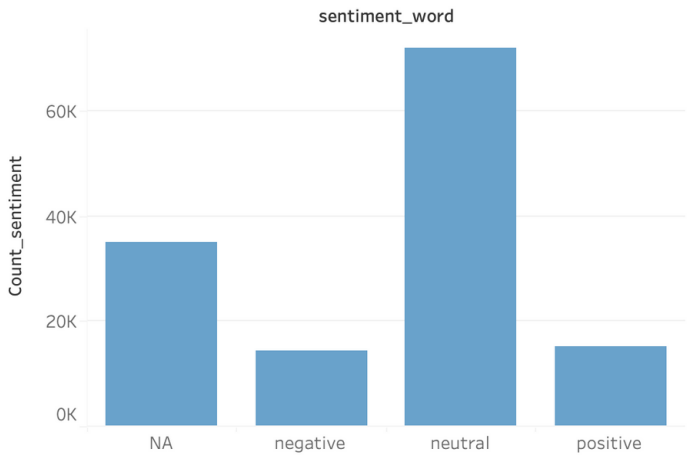
Sentiment Visualization

- We used lexicon-based sentiment classification

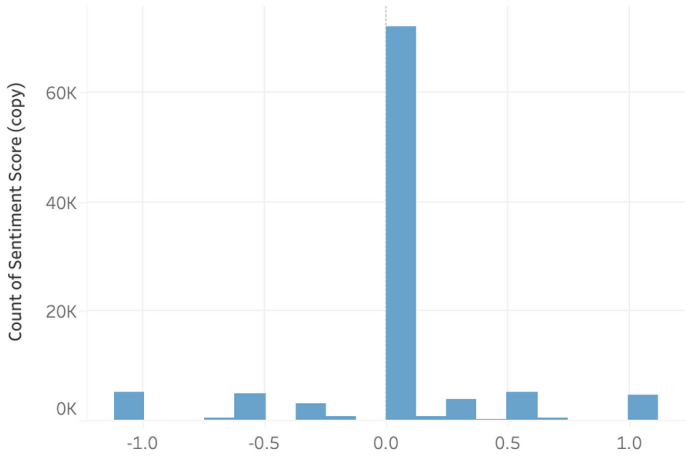
- Many slangs do not have a sentiment score

- However, judging even by naive occurence of words we can see that the overall environment of DOTA2 is very toxic

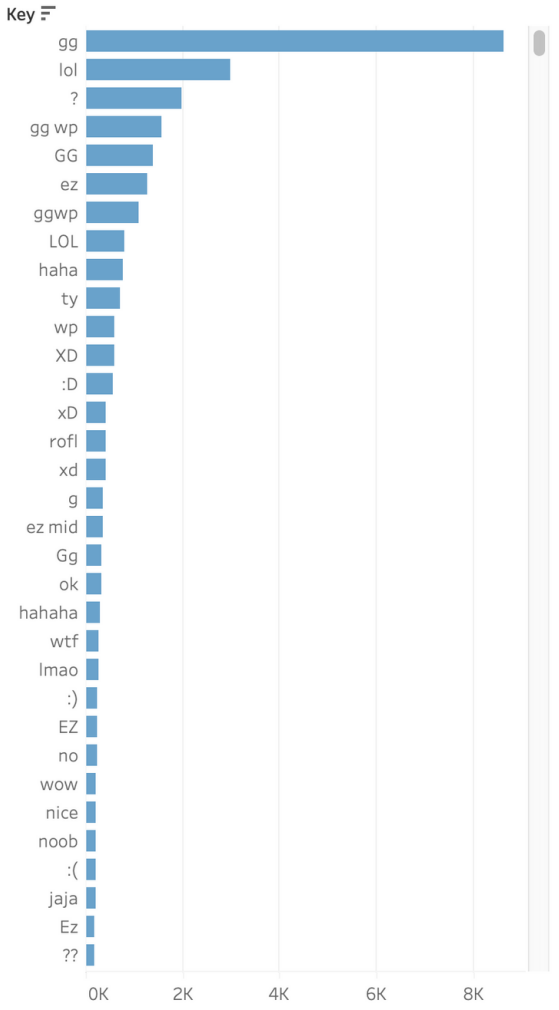
Number of Positive and Negative Chat in Selected Game



Distribution of Sentiment Score



Chats that Occur by Frequency



CONCLUSION

DOTA 2

Recommendation and Future Scope

OLTP STAGE

OLTP shall be expanded to extract personal performance data and be connected with apps/websites that provides a platform to track user-level data.

MODELING STAGE

Current sentiment analysis for match chat is done using lexicon lookup. A better methodology may be researched by either expanding the sentiment dictionary or training a model using labeled data.

DOTA 2

Lessons learned

DATA COLLECTION

- We should take into account of the time and difficulties of data modeling when collecting the data so that can we be more prepared for the upcoming workload.
- We should consider more dimensions of the data that can potentially enrich our business insights.

DATA MODELING

- As we are using NoSQL and Mongodb for data modeling, we should start familiarizing the related concepts and techniques earlier to avoid the procrastination due to inexperience.
- We should have better time management and division of project to facilitate our progress.

BUSINESS INSIGHTS

- We should design the connection between data modeling and business insights in advance to avoid going back and forth to create and transform variables.
- We should be more aware of the timeliness of our business insights and what do the business insights truly deliver in practice.