Draft-Analysis of the Ancients: Predicting Draft Picks in DotA 2 using Machine Learning

Adam Summerville
UC Santa Cruz
asummerv@ucsc.edu

Michael Cook
Falmouth University
cutgarnet@gmail.com

Ben Steenhuisen
DatDota
noxville@noxville.co.za

Abstract

Analysing strategic decision-making in eSports is an increasingly important problem – for players, for teams, for commentators, for viewers and for broadcasters. Such analysis is extremely difficult, however, because of the comparatively small quantities of data, the ever-shifting state of competitive play, and the huge complexity of the game. In this paper we describe a system for predicting drafting decisions in DOTA 2, and evaluate both how the system performs compared to human experts, as well as the new kinds of analysis made possible by automation.

Introduction

Team-based competitive online games like *DOTA 2*, *League of Legends* or *CounterStrike: GO* are among the most played games today, but they are also the most *watched*, with millions of people spectating online tournaments live either in-game or via services such as *Twitch*[1]. DOTA 2, the largest eSport in terms of prize money awarded, is growing with almost uncontrollable speed – betting companies and team sponsorship deals outpace legislation and single event prize pools are exceeding $20m.

In June 2016 Will Partin published an article in *Kill Screen* titled "DOTA 2 Might Be Nearing Its Moneyball Moment", in which they explored the increasing trend of DOTA 2 teams hiring dedicated analysts to provide an edge on their opposition (**?**). Partin likens this to the trend of *sabermetrics* in baseball, where a deep focus on statistical analysis can yield valuable insights into the game itself. The term 'Moneyball' refers to the practice of making decisions in baseball based on sabermetrics analysis, and takes its name from a book about the management and performance of the Oakland Athletics team in the early 00s. Despite an increased investment in analytics and research within eSports like DOTA 2, almost all of this work is currently performed by hand.

Elsewhere, broadcasting studios and tournament organisers hire large teams of people to provide coverage of eSports events, working both on and off camera. This includes deep and complex statistical analysis of all teams in a tournament, all games leading up to that tournament, trends that emerge within a tournament, as well as live responses to games that are in-progress. All of this work is done painstakingly by hand, requiring searching through databases of information on thousands of games and hundreds of players. Analyst Brian Herren described the process as "like digging through hay to find needles" in an interview about the process (**?**).

DOTA 2 is a rich and exciting area for automated analysis. Unlike real-world sports like baseball, a replay file is automatically created for every match that has ever taken place, from the highest tier of professional play to newcomers taking their first steps. This offers an unprecedented opportunity for investigation, research and analysis, yet DOTA 2 as a problem domain has many unusual complexities, some of which we will explore in this paper, which poses challenges for existing ways of applying analytical AI techniques.

Regardless of whether DOTA 2's Moneyball moment is near, automated analysis is an important area of development for eSports in general. Teams and broadcasters are heavily reliant on manual labour even at the highest level, and could greatly benefit from assistive tools. Equally, high-quality tools could also improve eSports communities in regions with smaller or nonexistent professional scenes, by supporting the work of local broadcasters, event organisers and professional players with expert-level knowledge and insight.

In this paper we describe initial work in applying machine learning techniques to analysing the drafting phase of DOTA 2 matches. We describe the problem and the broader context of professional DOTA 2 analysis, give details on the methodology used, provide an overview of our results, evaluate the system in the context of human experts, and discuss the implications for the future of eSports analysis.

Motivation & Background

DOTA 2 is an online eSport in a genre typically called *MOBA* ('multiplayer online battle arena'). The majority of gameplay takes place in a third-person action-strategy game where players control one hero in a team of five, fighting against another team of five players. The aim of the game is ultimately to destroy a particular building in the opposing team's base, but there are many approaches to achieving this, from fast-paced 15-minute sprints through to slow, 90-minute marathons. For a fuller description of DOTA 2, we direct the reader to (**?**). For the most part, an understanding of DOTA's gameplay is not necessary for this paper, since we will be focusing on the

[1] http://www.twitch.tv

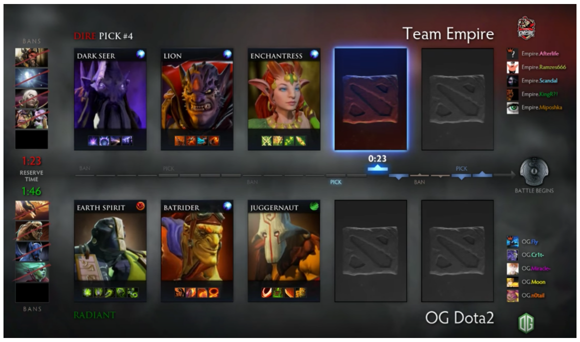

Figure 1: The drafting phase from a match between Team Empire and OG. Picked heroes appear in the large, central portraits, while bans are listed on the left in smaller images.

pre-game drafting phase, explained below.

In the most common type of DOTA 2 game, *All Pick*, each player chooses a hero from a pool of 111 possibilities, with no hero picked more than once (if one player selects a hero, it is unavailable for anyone else). Professional games do not use All Pick – they instead use *Captain's Mode*, in which a team captain makes decisions for the rest of their team. The two captains alternate in either *banning* a hero (which renders it unavailable for both teams) or *picking* a hero (which adds it to their roster) in a precise predetermined order – this process is known as *drafting*. The order of picks and bans in the current version of DOTA 2 is shown in Figure 2, while a screenshot of an in-progress draft from the spectator perspective is shown in Figure 1. Teams perform a coin-toss before a match to decide which teams picks first in the draft, and which side of the map each team starts on.

Heroes have different strengths and weaknesses depending on the skills they have, what statistics they start with, and what kind of hero they are (for example, whether they attack at range or in melee). *Very* broadly speaking, one can partition heroes into two groups: *support* heroes and *core* heroes. Core heroes need time to become useful, either because they need to purchase items with gold, or rise in levels with experience points. Support heroes tend to need less time to become useful, and may indeed be useful from the beginning of the game. However, they tend to be less useful later on as core heroes become more powerful. A common tension in a draft is the balance of core heroes to support heroes – as more heroes are revealed, the balance becomes more clear and a team's strategy may become more clear.

Decision-making in a drafting scenario is affected by many external factors. These might be simple and immediate, such as a knowledge of which heroes your team members are most skilled at playing. Others may be more variable, but still local – what heroes the opposing team has played recently, for example. Other factors may be global and always in flux – in particular the *metagame* is a term used to describe what strategies, heroes and ideas about the game are currently popular, as well as referring to the game-theoretic properties of how strategies change slowly over time. The metagame exerts a very strong influence on decision-making both at the highest level of professional play, through to the most casual public games. These factors and more are taken into account by players and analysts when thinking about game drafts.

Making predictions about what might come next in a draft is fundamental to many tasks in playing and analysing DOTA 2 matches. For players, most obviously, predicting what a team will do next confers a strategic advantage. Professional players regularly perform 'mock drafts' against knowledgeable players to practice drafting in a certain way, and compile 'bibles' of handwritten notes on drafting which they often take into tournament booths with them. For analysts, commentators, presenters and others involved in broadcasting, predicting drafting decisions is the primary driving force behind the conversations and discussion during the broadcast of a match's drafting phase. This is used in a variety of ways – to pose questions about future possibilities, to theorize about upcoming strategies, to justify decisions made based on the likelihood of future moves (for example, Team A bans a hero because they predicted Team B would pick it next) or to have broader conversations about the state of the metagame.

One of our primary motivations for selecting DOTA 2 over other MOBAs, such as League Of Legends, is that its metagame is extremely diverse. In the qualifier stage for the year's most important DOTA 2 tournament, The International, 93% of DOTA 2's heroes were either picked or banned. League of Legends' metagame is somewhat more conservative – in the 2015 World Championships, only 60% of heroes were picked or banned. A diverse metagame presents a more interesting learning problem and also signifies a more balanced game, which means that strategies and drafting are more important and more nuanced.

Related Work

Research work analysing DOTA 2 and other games in the MOBA genre are becoming more common, although most of this analysis is restricted to the gameplay phase rather than drafting. For example, in (**?**) the authors consider the problem of detecting engagements between groups of players before they happen by analysing their movements in the game. Other work, such as in (**?**), focuses on the heroes that make up a team, but their interest is in predicting the success of a team based on the selected heroes. In (**?**) the authors explore how changes to the internal economy of a MOBA influences the heroes chosen by players in casual play, which is closer to the area we investigate here, but focused on everyday gameplay motivations rather than professional play.

A related task to draft analysis is the problem of hero suggestions – proposing a good hero for a player given a situation and that player's needs. Work in (**?**) follows this concept with machine learning techniques for League of Legends. The problem of hero suggestion differs from the work here in two important ways: first, we consider the entire draft, including bans, for both teams; but most importantly, we are interested in what is *likely* to be picked, not what is necessarily best. This means we are interested in recognising common trends, personal preferences and emerging themes, rather than acting as an expert coach to suggest the best next move. This is important for coaches and commentators interested in a professional game, as they are most interested

	B	B	B	B	P	P	P	P	B	B	B	B	P	P	P	P	B	B	P	P
Team 1 [First Pick]	●		●		●			●		●		●		●		●		●	●	
Team 2 [Second Pick]		●		●		●	●		●		●		●		●		●			●

Figure 2: The order of picks and bans in Captain's Mode as of version 6.88. The draft progresses from left to right. A column marked 'B' indicates a Ban, while 'P' indicates a Pick. Note that the team with second pick picks two heroes in a row, and also gets to pick last.

in what the team drafting is going to do – which may not be the same as optimal.

Drafting is also the name given to a similar notion in real-world sports, although their function as part of the game is something different. Drafting refers to the selection of players prior to a season starting, based on the assumption of how they will perform during a season. Some research has gone into this process, such as an investigation into the order of player selection in (**?**) or an attempt to predict quarterback performance in the NFL as in (**?**). Real-world sports drafting differs from our chosen problem domain in many important ways. Most importantly, it is assessing the value of human beings at accomplishing a task, rather than the strategic worth of a decision within a system of known, static rules. It is also worth noting that drafting in sports is a much broader and vaguer process, that takes place across all teams in a sport simultaneously, and is performed once per year. Drafting in eSports is a competitive head-to-head process that happens every single game, and is an integral part of the game's design. In general, we feel it shares little in common with its real-world namesake (although many MOBAs now have transfer windows and collegiate teams, perhaps indicative of a future where both forms of drafting take place in eSports).

Methodology

In the following sections we will (1) detail the dataset and (2) discuss the different machine learning techniques used for this analysis, Bayes Nets (BNs) and Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs).

Dataset

When a game of DOTA 2 is played, a replay file is generated containing enough information to recreate the game so it can be watched again in-client. DatDOTA[2] is a site which downloads replay files from officially-recognised tournaments, and then parses and extracts detailed information from each match. This is then loaded into a searchable database which is used to find things from the winrate of a particular hero on a patch, to the average time it takes a particular player to build a particular item, on a particular hero.

The dataset used in this paper is derived from all professional matches parsed by DatDOTA between October 5th 2015, and December 16th 2015, covering 1518 matches in total. This includes high-profile tournaments like the $3m Frankfurt Major, as well as smaller, local tournaments held at a national level. This period spans almost the entirety of DOTA 2 patch 6.85, which was released on September 24th 2015. The first few weeks of patch data were unavailable for our study due to technical complications with the database, meaning that we are missing 126 matches in total. Fortunately this only includes a single non-minor tournament, a two-day ESL invitational. In Future Work we discuss using larger datasets in the future, and we are working closely with DatDOTA to build more comprehensive datasets for use in the future.

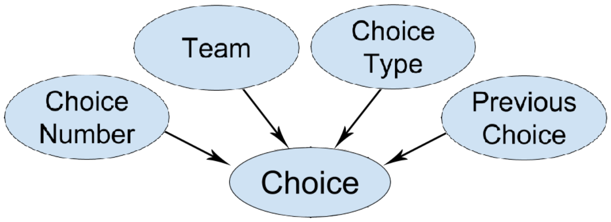

Figure 3: Topology of the *Full* Bayes Net. The *Choice Number* network does not have the Team, Choice Type, or Previous Choice nodes. The *Nulligram* is simply the Choice node.

Bayes Nets

BNs are a graphical structure that represent probabilistic dependencies between Random Variables. A BN is a Directed Acyclic Graph (DAG) where each node can either be an observed or latent variable. For this work we considered 3 different BNs. Figure 3 shows the topology of the *Full* network. In mathematical terms this network represents:

$Pr(C|N, T, C_t, C_{i-1})$

where C is the current choice, N is the number of the choice, T is which team is making the choice, C_t is they type of choice (Ban or Pick), and C_{i-1} is the previous choice made. Two simpler models were considered the *Choice Number* network which is simply

$Pr(C|N)$

and the *Nulligram* model which is the baseline model that considers no context, i.e.

$Pr(C)$

Ideally, a longer history would be used to enable learning more complex patterns, but this is intractable for two reasons. First, given the 111 heroes each step into the past increases the size of the network by a factor of 111 which quickly explodes the size of the network (2 steps in the past is beyond the limits of the 32-bit memory space used by GeNIe &

[2] http://www.datdota.com

SMILE). While using sparse matrices can make the memory and performance constraints feasible, ultimately it cannot help the larger problem. Without considering any previous choices, the *Full* network already consists of 8,880 entries nearly 6 times as many entries as the number of drafts considered. Increasing the size of the network will only increase the sparsity.

Long Short-Term Memory

LSTMs represent the current state of the art for sequence prediction. LSTMs are a neural network topology that allow for the efficient training of RNNs capable of learning across many time steps. Standard RNNs are only capable of learning a few time steps into the past given what is commonly known as the "vanishing gradient problem". During back-propagation the weight on the recurrent edge will be updated multiplicatively meaning that any weight < 1 will rapidly approach zero as the number of time steps considered increases (and any weight > 1 will rapidly explode). LSTMs instead have an additive recurrent edge which will only increase or decrease at a linear rate with the number of time steps. LSTMs also have gates on the input, output, and remembering of the recurrent value, allowing the LSTM cell to selectively choose when to remember, when to forget, and when to produce output.

For this work the LSTMs are trained on a One-Hot encoding which is predicted via a Soft-Max categorical output. The models are trained using Categorical Cross-Entropy loss. The encoding consists of a vocabulary of 137 words which is composed of:

- 111 Heroes
- 2 Ban/Pick choices
- 2 Teams
- 20 digits
- 1 special *Start* word
- 1 special *Null* word

A snippet of sample input looks like:

```
  Null Null Start Ban 1 TeamA Doom Ban 2
TeamB Tusk
```

Each choice in the draft consists of 4 words:

- Choice Type (Ban/Pick)
- Choice Number (1-20)
- Team (Team A/Team B)
- Hero

This means that each draft consists of 80 words, hence a sequence length of 80 words was used during training. Between each draft there are 80 filler *Null* words followed by the *Start* word. Ideally, this would not be necessary as the forgetting capabilities of the LSTM would be capable of recognizing the *Start* word to forget all previous knowledge; However, given the lack of data the system was learning the orderings of the drafts and was learning dependencies between drafts (e.g. if it knew the last 2 picks of the previous drafts it could have increased knowledge about what the first ban would be). The 80 *Null* words prevent this from happening given the length of training sequences.

For this work we used and LSTM network consisting of 4 hidden layers each consisting of 512 LSTM cells, and a dropout rate of 20% to reduce overfitting. The networks were trained using Torch (**?**).

Results

We split the data into 11 folds, training the models using 10 fold cross-validation and keeping the 11th fold as a test set. Accuracy can be seen in figure 4. Of the computational models, the LSTMs had the highest accuracy with a success rate of 11.94% on the held out test data. Reported accuracy for the LSTM is only for the heroes, not any of the additional meta-tokens in the sequence (e.g. `TeamA` or `Ban`).

We wanted to better show how well the systems do as function of depth into the draft which can be seen in figure 5. The *Choice Number* model was used as it was not statistically different from the more complex BN model and therefore represents a better balance of simplicity and accuracy. The first 3 picks do not differ much between the LSTM and BN which makes sense for multiple reasons. (1) The first 4 picks of the draft were quite common (Shadowfiend, Doom, Tusk, and Queen of Pain) and (2) the LSTM does not have much sequence data to work off of meaning there is little room for it to distinguish itself. However, as the draft goes along the LSTM by and large outperforms the BN handily (being up to 2.4× as accurate on the final pick).

Comparison With Humans

To assess the quality of our predictions, we sought to compare the performance of our system with real-world analysis of professional-level game drafts. We transcribed the commentary for the draft phase of several DOTA 2 matches and assessed the kinds of predictions made by people during discussion, and how accurate they ultimately were. One difficulty here is that predictions are rarely confidently made by analysts – far more common were weaker assertions, for example proposing that one of a number of heroes *might* be picked or banned, or making a prediction but then couching it in a possible explanation for why it may not happen. We tried where possible to focus on firm predictions only, where analysts made statements to the effect that a specific hero *would* be picked or banned by a specific team. In total, we analyzed 23 professional games from the same period our dataset is pulled from, and observed 101 predictions in total. The full transcripts and annotated predictions can be found online[3].

In Figure 4 we break down the human predictions in three different summaries. The *Lenient* summary considers predictions as single units - if a hero is correctly predicted, we consider the prediction correct. Under *Strict*, different elements of a prediction are considered separately, e.g. if a human predicts that a team will pick a specific hero in a phase, that actually is 2 predictions, one for each possible time the team could make that choice, or if a human predicts

[3] https://github.com/gamesbyangelina/dota2transcripts

Method	Avg. Eval Success %	Avg. Test Success %
Bayes Net - Nulligram	4.93	4.95
Bayes Net - Choice Number	9.91	10.05
Bayes Net - Full	10.55	11.48
LSTM	13.42*	11.94*
Human – Lenient	31.48	
Human – Strict	13.11	
Human – All Possible	6.66	

Figure 4: Average accuracy of the Top 3 picks for evaluation and test sets across the 10-folds. Each fold consists of 139 drafts consisting of 20 choices each, for a total of 2780 data points. Of the non-human methods the LSTM performs the best with $p < 1e - 4$ using Fisher's exact test. The human experts perform better, but given the limited sample size it is not statistically significant.

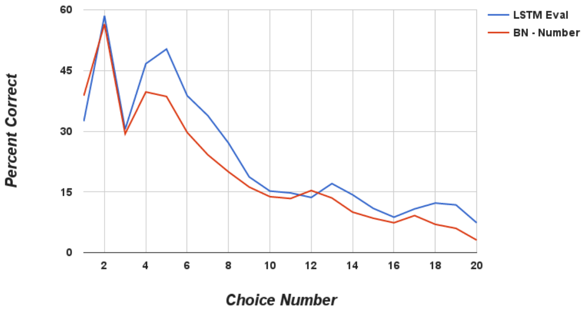

Figure 5: Percentage correct for top 3 by draft phase.

that a team will at some point pick a given hero that is actually 5 predictions. Finally, *All Possible* considers every single pick and ban in the draft even if a human never makes a prediction. Essentially, this says that if a human does not make a prediction they are making a null prediction of "I don't know". The AI is forced to make all possible predictions, no matter its certainty, so it is closest to the *All Possible*, but this unfairly punishes the humans who did not attempt to make all possible predictions so the true human value is likely in between these two ranges.

Better evaluation against human experts is needed in future studies. We are currently working with a group of expert analysts and commentators and hope to conduct a more formal study in the near future.

Using Confusion Matrices For Analysis

One additional hypothesis we wished to explore with this work was the idea that confusion matrices produced by these models would reveal information about hero similarity. The confusion matrix indicates where a model confuses two or more heroes (i.e. they chose hero A when instead it was B). We theorized that commonly-confused heroes might relate to conventional wisdom about heroes which are interchangeable or similar. In the case of the models we have produced so far, we have no found any such relationship.

In Figure 6 we show two sets of heroes that might commonly be thought to be interchangeable or related, and how often they were chosen at different phases of a draft. As you can see, even though these heroes have similar in-game utility, they have very different patterns of how they are selected in drafts. There are many reasons why this might be – some might be stronger when picked later in a draft, some might be dependent on other picks or bans. Regardless, what this shows is that while a hero's functionality in game is important, it has little impact on how players draft. This warrants further investigation in future work, with larger datasets, but we believe this is good evidence for why draft analysis, independent of hero mechanics or performance, is important.

Future Work

The system we've presented in this paper represents only the beginning of work on draft predictions, much less the broader problem of general techniques for eSports analytics. In this section we explore some of the immediate points of future work, as well as the longer-term goals for this research area.

Patch & Metagame Shift

As the metagame is explored and changed by players it inevitably shifts towards an equilibrium where increasingly optimal strategies are discovered and reused. To mitigate this, games such as DOTA 2 are often patched to rebalance the game, improving underused aspects of the game and weakening overused aspects. Patch notes are released a few days before it is applied to the game, but the impact of a patch is quite significant. DOTA 2 releases a major patch every three months on average, with smaller 'tweak' patches applied in between. On average, the last six DOTA 2 patches changed 84 heroes (out of a pool of approximately 111) and 39 items. Strategies and common drafts disappear overnight as the game is transformed and new ideas must now be explored.

The days after a new patch are stressful for players and broadcasters, but they pose a particular problem for machine learning systems. Because our models are trained on data

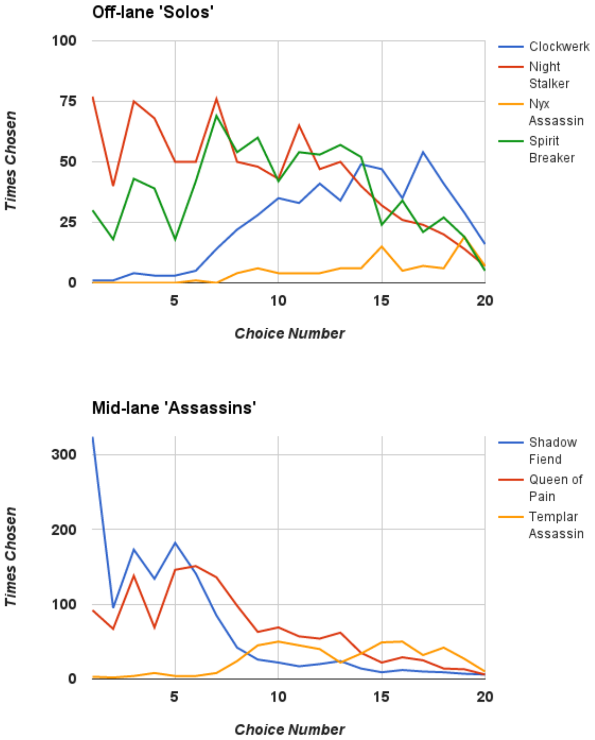

Figure 6: Likelihood of heroes being confused with one another. Top: Off-lane ’solos’. Bottom: Mid-lane ’assassins’.

from particular patches, the new patch means that our data is no longer as relevant. Further work will be needed to investigate how draft systems can adapt to patches and still provide analysis. One common thing done by experts in the early weeks of a patch is hypothesising which heroes will be highly valued. This is effectively a static analysis of the patch rather than of the metagame. A separate system designed to analyse patch notes, perhaps armed with a model of gameplay, could possibly replace or support a machine learning system in the early phases of a new patch.

Variable Epoch Analysis

For the purposes of this paper we trained a system on games from a single patch, which provides a model of that patch. We did this because we believe a system trained on a single patch’s metagame is likely to be more accurate. However, some knowledge transcends a particular patch, and instead reflects general truths about DOTA 2 as a game. For example, certain heroes are designed around *armour reduction*, which makes a target more vulnerable to physical attacks. Heroes which reduce armour thus tend to pair well with heroes who do physical damage, and this remains true regardless of patch (although its relevance may rise and fall with the metagame).

We believe that by training models on different segments of professional DOTA 2 history, from individual patches through to models trained on years of data, that we might be able to build a more sophisticated analysis that takes into account both local metagame shifts as well as long-term trends. Different models can compare predictions and weight suggestions differently – allowing the system to potentially make distinctions between good heroes generally, and heroes which are particularly good in the current metagame.

Incorporating Amateur Match Data

In contrast to other work that exists studying DOTA 2 matches, our models do not take into account public matches, only professional ones. Although the pool of professional DOTA 2 matches is small (a matter of thousands of matches, compared to millions of public games) we made this decision because the professional metagame is slightly different compare to casual play, and we wanted to analyse that dataset in isolation. However, public trends do affect professional DOTA 2, and are also reflective of longer-term trends in the game (as we discussed in the previous subsection). As such, we would like to investigate how this data can be incorporated into our system in the future.

Conclusions

In this paper we described a system for predicting drafting decisions in professional games of DOTA 2. We motivated the work by pointing towards the depth and richness of available data, as well as the complexity and relevance of professional eSports analysis. We described our approach to the problem, explained how we tested the system using data collected from professional matches, and then provided an overview of results with a comparison to real-world examples of draft analysis. Finally, we looked at the challenges posed by this area in the future.

eSports analysis is an area rich with interesting problems for research to contribute to, backed up by a wealth of available data that far exceeds the detail and scale possible with real-world sports. We’re excited by the potential in this area, and the way this work might be able to impact the emerging and rapidly changing worlds of broadcasting, organisation, competition and spectatorship around this genre of games.

References

Aitchinson, K., and Herren, B. 2013. Interview - the international 3. http://tinyurl.com/herreninterview.

Collobert, R.; Bengio, S.; and Marithoz, J. 2002. Torch: A modular machine learning software library.

Godec, K. 2015. Welcome To DOTA, You Suck. https://purgegamers.true.io/g/dota-2-guide/.

Lee, C.-S., and Ramler, I. 2015. Investigating the impact of game features and content on champion usage in league of legends. In *Proceedings of the Foundations of Digital Games Conference*.

Partin, W. 2016. DOTA 2 might be nearing its Moneyball moment. https://killscreen.com/articles/dota-2-moneyball-moment/.

Sabik, A., and Bhattacharya, R. 2015. Data-driven Recommendation Systems for Multiplayer Online Battle Arenas. Master’s thesis, Johns Hopkins University.

Schubert, M.; Drachen, A.; and Mahlmann, T. 2016. Esports analytics through encounter detection. In Sloan, M., ed., *Proceedings of the MIT Sloan Sports Analytics Conference*.

Staw, B. M., and Hoang, H. 1995. Sunk costs in the NBA: Why draft order affects playing time and survival in professional basketball. *Administrative Science Quarterly* 40(3):474–494.

Wolfson, J.; Addona, V.; and Schmicker, R. H. 2011. The quarterback prediction problem: Forecasting the performance of college quarterbacks selected in the NFL draft. *Journal of Quantitative Analysis in Sports* 7(3).

Yang, P., and Roberts, D. L. 2013. Knowledge discovery for characterizing team success or failure in ARTS games. In *Computational Intelligence in Games (CIG), IEEE Conference on*.