The Data Gap in Sports Analytics and How to Close It

Alon Harell and Ivan V. Bajić
School of Engineering Science, Simon Fraser University
Burnaby, BC, Canada, V5A 1S6
aharell@sfu.ca

Abstract

As the importance and prevalence of sports analytics grows, so does the inequality in sports data. In this paper we examine two main sources of such disparity - the perceived hierarchy of sports and privatization of data. We argue that such inequality hurts the sports analytics community in the short and long terms, and suggest ways for the deep-learning, AI, and sports analytics communities to help mitigate the issue.

Introduction

Sports analytics describes the practice of quantifying and modeling various aspects of the performance of athletes and sports teams. Tracing its origins to 1858, when Henry Chadwick began publishing statistical reports of baseball games, known as box-scores (Puerzer 2002), sports analytics has since become an intrinsic, meaningful part of sports at all levels. Analytics are used, in one way or another, by all parties in sports. Leagues use data and modeling for adjusting rules, creating policies, improving refereeing, and increasing popularity through advertising. Players and coaches can use analytics to devise training routines, manage athlete workload, design specific plays, prepare for opponents, and even make live in-game decisions such as play calls or substitutions. Teams can use data for optimizing player selection, especially under the financial limitations imposed by the team's budgets or the leagues regulations such as salary caps. This last use of analytics was made particularly famous by the 2012 Academy nominated film ”Moneyball” (Miller 2011). Sports analytics is also popular among the followers of sports, helping fans gain a deeper and more profound understanding of their beloved sport; allowing media outlets to improve the quality of their broadcast (*Second-Spectrum* 2019a); and even assisting betting agencies in determining highly substantiated betting odds.

Recently, sports analytics has experienced a proverbial boom in both complexity and popularity. This has lead to more data being collected than ever before, and more models being built upon those data. When reviewing the advances in sports analytics it is advantageous to separate them into two major components: data collection and data analysis.

Data collection involves gathering specific measurements for each individual sport. In order to be effective, data collection must be accurate, reliable, timely, and provide measurements that can be used to make meaningful insights. Starting from the relatively humble box-score, data collection has evolved significantly over the years. When collecting cumulative statistics, it is necessary to track each individual play. Eventually, a game-time measurement was added to these plays, creating what is now known as play-by-play. In recent years, due to technological advancement, data collection has expanded to include things such as automatic player and ball tracking, automatic play classification, and more. In order to achieve such advanced capabilities most sports have turned to commercially developed systems such as SportVU (*Stats-LLC.* 2019), Second Spectrum (*Second-Spectrum* 2019b), Titan Sensor (*Titan-Sensor* 2019) and others.

In *data analysis*, a variety of statistical methods are used to produce meaningful, concise and actionable metrics for teams, coaches and players. Similarly to data collection, analysis has also taken major steps forward in recent years. These steps include a more in-depth analysis of traditional box-score stats alongside models based on the newly available play-by-play or even tracking data statistics. Some advanced metrics based on traditional data include adjusted yards per attempt (AY/A) (Carroll, Palmer, and Thorn 1989) in American Football or on base percentage (OBP) (*MLB* 2019) in baseball. Some metrics based on play-by-play include the adjusted plus-minus (Ilardi 2007) in basketball, or the weighted plus-minus in soccer (Schultze and Wellbrock 2018). The introduction of player tracking data has lead to even more advanced metrics, such as expected possession value (EPV) (Cervone et al. 2014) in basketball, expected goals (xG) (Rathke 2017) in soccer, and many more.

Data Inequality

Perceived Hierarchy of Sports and Practitioners

As the popularity and complexity of sports analytics grow, so does the cost of data collection and analysis. This, in turn, leads to a growing disparity based on what we name the perceived hierarchy of sports and practitioners. Within this we include the differences between data-rich and data-poor sports, leagues, teams (Witz 2014), and genders (Mc-

Cann 2015). As an example, the Los Angeles Clippers of the NBA, list 5 employees strictly dedicated to basketball analytics in their staff directory (LA-Clippers 2019) out of a total of 64 employees in the basketball operations department. The Los Angeles Sparks of the WNBA, for comparison, list no dedicated sports analytics employees and 5 total basketball operations employees (LA-Sparks 2019).

In college sports, these differences often highlight the disparity between resource rich programs and smaller, less funded ones. For example, Duke, Louisville, and Marquette men's basketball teams have collected, and presumably taken advantage of, player tracking data using SportVU since the 2013/14 season (Witz 2014). It is safe to assume that smaller Division I schools, and most Division II and III schools, cannot afford the costs of such a system, which used to cost around $100,000 a year for an NBA franchise (though, possibly cheaper for colleges) (Witz 2014).

This disparity can have direct impact on team performance, allowing rich organizations to gain yet another competitive advantage. When small teams manage to overcome this gap, the benefits can be substantial. For example, the SFU men's basketball team had 2 and 4 wins in the 15-16 and 16-17 seasons, respectively. In 17-18 and 18-19, through collaboration with the school's Sports Analytics Research Group, the team improved to 9 and 14 wins, respectively.

When examining current publicly available data collection and analysis methods, we notice that most employ machine learning techniques such as non-parametric statistical modeling, deep learning, and others. These techniques are all highly data-dependent and thus are susceptible to bias caused by the data on which they are trained (Garcia 2016; Zhao et al. 2017). For this reason, the significance of the resource discrepancy goes beyond data collection and availability, impacting the analysis of data.

Data Privatization

The development and expanded monetary value organizations attribute to sports analytics have lead to a boom in sports analytics firms and products. Many companies, some of which are listed in the introduction of this paper, have been created to collect and analyse sports data. Although there is no doubt that this rapid growth has lead to incredible developments, allowing for more advanced data collection and analysis, it has also come at a price. As the significance of private business has grown, the gap between publicly available data and private data, has grown as well.

In the NBA for example, player tracking became a league-wide standard in the 2013/14 (*NBA-PR* 2013). These measurements contain 25Hz updates of each players position on the court, along with the ball. In addition to these measurements, some data are tagged and grouped to form semantically meaningful tags such as shots and shot types (pull-up, catch-and-shoot, etc.), plays (pick-and-roll, drive, etc.) and more. Unfortunately, the data released by the league only contains higher-level stats derived from these measurements, such as shot-locations, play-type stats, speeds, etc.

Similar situations occur in many other sports, such as the NFL, where similarly advanced player tracking is performed and morsels of it are released to the public under the name "Next Gen Stats". In European soccer, where regulations are less uniform than in American major sports leagues, tracking data are generally not publicly available, and advanced metrics are available directly from the leagues, through collaboration with big clubs, such as Chelsea FC (De Silva et al. 2018), or through paid services such as Prozone.

Impact on the Sports Analytics Community

We argue that aforementioned data inequality negatively impacts the sports analytics community in many ways. We divide these effects into *people-related* - referring to the members of the analytics community; and *model-related* - referring to the actual research performed. We focus on these effects, and how to mitigate them. That being said, we also believe that these issues are worth fixing in and of their own, simply to empower as many athletes, coaches, and fans as possible through advanced sports analytics.

People

Most members of the sports analytics community begin their journey to sports analytics through a passion for a sport, combined with expertise in data analysis. The benefits of data analysis are clear to us, because we have seen them in practice, in our sports of choice. However, practitioners or fans of data-poor sports, may only have very limited exposure to analytics, meaning they may not choose to pursue it. This means we are failing to attract potentially excellent researchers that are in less-privileged teams' fan bases. We could be missing out on the analytics equivalents of Damian Lillard, Becky Hammon, or Carson Wentz.

We lose even more potential researchers through the difficulty of accessing and collecting meaningful data. As in many data driven fields of science, the first step in beginning a project is to ascertain whether the relevant data exist, or can be reasonably collected. For many individuals, such as graduate students, or small teams or companies, this may become a road block that prevents them from ever entering the field. If we can make more data public, we can increase the viability of becoming a sports analytics researcher in the first place. We believe that this will greatly help the analytics community in both the short and long term.

Models

In addition to missing out on potential researchers, we claim that the data inequality also hurts the product put forth by the analytics community. In many fields, especially in data-driven ones, public datasets and competitions have been a major driving force for advancement. Perhaps the most notable example is the ImageNet Large Scale Visual Recognition Competition (Russakovsky et al. 2015). ImageNet has been a gold standard in the fields of computer vision and deep learning. In fact, before 2013, public and academic interest in neural networks was very limited. When AlexNet (Krizhevsky, Sutskever, and Hinton 2012) first won the ImageNet challenge in 2013, that immediately changed, leading to the current golden age of deep learning. Similar datasets exist in many fields such as natural language processing (Webster et al. 2018) bio-medical engineering (Wang et al.

2017) and more. In sports analytics however, there remains a dearth of publicly available data and competitions.

A notable exception are events in which professional leagues, or teams, create open competitions, to encourage public participation in the solution of a specific problem. Some examples of this are the NFL Big Data Bowl (*NFL* 2019), the NBA Hackathon (*NBA* 2019), and the Edmonton Oilers Hackathon (Dittrick 2013). These events have lead to many interesting projects, and often result in researchers being employed by the relevant leagues.

Another area with currently untapped potential, is cross-sport research. This is perhaps most important in aspects that are common across a variety of sports, such as general athletic conditioning and evaluation, training load planning, injury prevention, etc. As a result of the significant disparity in data, researchers in sports that are lower on the perceived hierarchy have a diminished ability to produce meaningful insights, which could later benefit the entire community.

Similarly to cross-sport research, cross-organization data in the same sports is also currently difficult to utilize. This greatly impacts one of the major problems today in sports analytics - quantifying the potential of young athletes to succeed as professionals. For example, consider the data disparity between NCAA sports and their equivalent professional leagues, and within the NCAA teams themselves. When attempting to evaluate a college athlete, an analyst is unable to utilize many of the tools that are available to them when analyzing professionals, leading to inferior results. The disparity within the NCAA leads to the fact that many smaller school athletes are even harder to evaluate, potentially hurting both the teams and the athletes.

Bridging the Data Gap

There are many reasons why the data gap persists, and why it might never be entirely closed. However, having established the negative impact this gap has on the sports analytics community, it is imperative to consider ways to mitigate it.

Create Public Data

The first, most important, but most achievable step we suggest, is to continue building publicly accessible dataset and models. Recently, there has been a growing amount of academic publications on sports analytics from fields such as computer vision (Thomas et al. 2017), statistical modelling (Santos-Fernandez, Mengersen, and Wu 2019), operations research (Vaziri et al. 2018), etc. However, with the exception of a few notable commonly used datasets such as (Karpathy et al. 2014; Yu et al. 2018; Yurko, Ventura, and Horowitz 2019), the majority of papers published have relied on data mining from various online resources to obtain data. Furthermore, unlike many disciplines, code sharing in sports analytics is rather limited. The lack of knowledge sharing makes comparing, evaluating, and most importantly building upon previous work challenging, and can be avoided through direct publication of datasets and projects.

The publications of datasets along with code can also help analysts in lower-resource scenarios close the data gap. For example, a player tracking algorithm from video, such as (Takahashi et al. 2018) was developed using men's soccer video, but could be used as a basis for creating a similar solution for women's soccer, or perhaps for other field sports such as lacrosse. Once such solutions are made public, they can be used, with some modification, to annotate data and create public datasets for low-resource scenarios. We believe that through such efforts, data can be made available in scenarios that have hitherto not had access to data.

Reverse the Privatization Trend

Private data owners, such as sports teams, leagues, and commercial analytics firms, have a variety of reasons for keeping their data private. These include a competitive edge, privacy concerns (especially in college or high-school leagues), and monetary considerations. However, the sports community can use its strength to encourage the publication of such data. This may be achieved through collaborations with former colleagues who are now members of private organizations. More importantly, it can be achieved through providing substantial value to data owners from publishing their data, while addressing their concerns.

As mentioned earlier, hackathons and data bowls are a great example of scenarios where data is made public, in a manner that benefits private organizations. We believe that such win-win events can be used to advocate for more data to be made public. To address the competitive edge concerns, the data released can be from previous seasons, or anonymized, addressing privacy concerns as well. Furthermore the creation of periodic competitions, similar to ImageNet, can create considerable value to competition sponsors while also creating an industry standard for comparing the quality of various sports analytics models or methods.

Summary

In this paper, we wish to bring to light the significant gap in data availability that exists in the sports analytics community. Such data disparity notably exists between sports, leagues (especially between pros and amateurs), and unfortunately, genders. We have demonstrated the negative effects caused by this data gap, and by the gap between private and public data. We then presented several approaches for bridging this gap while benefiting all sides. We hope that through these steps, we can grow the sports analytics community, inspire new researchers, and improve the products put forth.

References

Carroll, B.; Palmer, P.; and Thorn, J. 1989. *The Hidden Game of Football*. Warner books.

Cervone, D.; D'Amour, A.; Bornn, L.; and Goldsberry, K. 2014. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA*, volume 28, 3.

De Silva, V.; Caine, M.; Skinner, J.; Dogan, S.; Kondoz, A.; Peter, T.; Axtell, E.; Birnie, M.; and Smith, B. 2018. Player tracking data analytics as a tool for physical performance management in football: A case study from chelsea football club academy. *Sports* 6(4):130.

Dittrick, R. 2013. Hockey fans, analytic gurus make compelling cases in oilers hackathon. *NBA.com*. [Online], Available: https://www.nhl.com/oilers/news/hockey-fans-analytic-gurus-make-compelling-cases-in-oilers-hackathon/c-665367 [Accessed: Nov. 16 2019].

Garcia, M. 2016. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal* 33(4):111–117.

Ilardi, S. 2007. Adjusted plus-minus: An idea whose time has come. *82 Games*. [Online], Available: http://www.82games.com/ilardi1.htm [Accessed: Nov. 11 2019].

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60:84–90.

LA-Clippers. 2019. Los Angeles Clippers staff and directory. *NBA.com*. [Online], Available: https://www.nba.com/clippers/club-directory [Accessed: Sep 16 2019].

LA-Sparks. 2019. Front office - Los Angeles Sparks. *WNBA.com*. [Online], Available: https://sparks.wnba.com/front-office-update [Accessed: Sep 16 2019].

McCann, A. 2015. Hey, nate: There is no ‘rich data’ in women’s sports. *FiveThirtyEight*. [Online], Available: https://fivethirtyeight.com/features/hey-nate-there-is-no-rich-data-in-womens-sports/ [Accessed: Sep 19 2019].

Miller, B. 2011. *Moneyball*. Columbia Pictures.

Puerzer, R. J. 2002. From scientific baseball to sabermetrics: Professional baseball as a reflection of engineering and management in society. *NINE: A Journal of Baseball History and Culture* 11(1):34–48.

Rathke, A. 2017. An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise* 12(2):514–529.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.

Santos-Fernandez, E.; Mengersen, K. L.; and Wu, P. 2019. Bayesian methods in sport statistics. *Wiley StatsRef: Statistics Reference Online* 1–8.

Schultze, S. R., and Wellbrock, C.-M. 2018. A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics* 4(2):121–131.

Takahashi, M.; Yokozawa, S.; Mitsumine, H.; and Mishina, T. 2018. Real-time ball-position measurement using multi-view cameras for live football broadcast. *Multimedia Tools and Applications* 77(18):23729–23750.

MLB. 2019. Standard stats glossary - on base percentage. [Online], Available: http://m.mlb.com/glossary/standard-stats/on-base-percentage [Accessed: Sep. 20 2019].

NBA-PR. 2013. NBA expands partnership with STATS LLC. to unveil player tracking technology for all 30 nba teams. [Online], Available: https://pr.nba.com/nba-stats-llc-partnership/ [Accessed: Sep. 20 2019].

NBA. 2019. 2018 hackathon recap. *NBA.com*. [Online], Available: https://hackathon.nba.com/2018-hackathon-recap [Accessed: Nov. 15 2019].

NFL. 2019. The NFL’s inaugural big data bowl. [Online], Available: https://operations.nfl.com/the-game/big-data-bowl/2019-big-data-bowl [Accessed: Nov 17 2019].

Second-Spectrum. 2019a. 2019 CourtVision highlights. *Vimeo.com*. [Online], . Available: https://vimeo.com/314544307 [Accessed: Oct 19 2019].

Second-Spectrum. 2019b. Win more games. [Online], Available: https://corp.synergysportstech.com/products/basketball-team-products/ [Accessed: Sep. 21 2019].

Stats-LLC. 2019. Stats sportvu basketball player tracking. [Online], Available: https://www.stats.com/sportvu-basketball/ [Accessed: Sep 20. 2019].

Titan-Sensor. 2019. Real-time GPS tracking. [Online], Available: https://www.titansensor.com/titan-realtime-gps/ [Accessed: Sep 22 2019].

Thomas, G.; Gade, R.; Moeslund, T. B.; Carr, P.; and Hilton, A. 2017. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding* 159:3–18.

Vaziri, B.; Dabadghao, S.; Yih, Y.; and Morin, T. L. 2018. Properties of sports ranking methods. *Journal of the Operational Research Society* 69(5):776–787.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106.

Webster, K.; Recasens, M.; Axelrod, V.; and Baldridge, J. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics* 6:605–617.

Witz, B. 2014. College basketball data aplenty for those who can afford it. *New York Times*. [Online], Available: https://www.nytimes.com/2014/03/25/sports/ncaabasketball/sportvu-offers-college-basketball-data-for-those-who-can-afford-it.html [Accessed: Sep 19 2019].

Yu, J.; Lei, A.; Song, Z.; Wang, T.; Cai, H.; and Feng, N. 2018. Comprehensive dataset of broadcast soccer videos. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 418–423. IEEE.

Yurko, R.; Ventura, S.; and Horowitz, M. 2019. nflwar: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports* 15(3):163–183.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.