SPORTS ANALYTICS, GOLF, AND GAMEFORGE: INNOVATIVE ANALYTICS FOR RECOMMENDER SYSTEMS

THE GROWING DICHOTOMY BETWEEN EXPERTISE AND STATISTICS IN SPORTS ANALYTICS

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering

By
Steven Wasserman

November 1, 2021

Technical Team Members:
Rose Dennis, Rachel Kreitzer, Jerry Lu, Samuel Roberts, William Scherer, Thomas Twomey

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Prof. Catherine Baritaud, Department of Engineering and Society

Prof. William Scherer, Department of Engineering Systems and Environment

Recruitment of athletes onto collegiate sports teams has grown tremendously in recent years, with 499,217 student-athletes participating in NCAA championship sports across 19,866 collegiate teams in the United States (Durham, 2019). Two major factors have supported this significant growth: the market value of the college sports industry in America, an estimated $18.9 billion among NCAA athletic departments in 2019 (Richter, 2021), and the rise of sports analytics in recruitment of junior athletes to U.S. institutions from around the world. This latter point is particularly apparent with the recruitment of junior golfers onto American collegiate teams, whereby the composition of many U.S. teams tends towards a mixture of students geographically local to a particular university and students from across the globe (Ringler, 2015). While data analytics has provided a breath of fresh air for recruiting student athletes, and for sports recruitment in general, it readily demonstrates the growing controversy between subject-matter expertise and statistical inference, how biases in expertise influences and at times negates the insights yielded from statistical analysis.

The technical capstone project and loosely coupled STS research project proffered in this prospectus confront the problems illuminated by the addition of data analytics into the decision-making processes regarding the recruitment of junior athletes to university programs. The technical project, conducted by Fourth-Year Systems Engineering students Rose Dennis, Rachel Kreitzer, Jerry Lu, Samuel Roberts, Thomas Twomey, and Steven Wasserman and supervised by William Scherer from the Department of Engineering Systems and the Environment, strives to enable greater information sharing amongst key stakeholders by re-designing previous work products to include new user groups in the key data-driven insights created for recruiting junior golfers. These same data-driven insights hold great value, not just by providing unobtrusive suggestions on recruiting junior athletes, but because they also offer unbiased recommendations

that might otherwise be overlooked by professionals. The capacity for cognitive biases to prejudice the opinion of subject matter experts is alarming; these same subject matter experts, however, play a pivotal role in constructing contextual frameworks for data analysis to operate in, specifically enabling the translation of abstract, speculative details into refined intelligence that enhance any given decision-making process. The STS research project will traverse the emerging field of behavioral economics to elucidate the opportunities available to ameliorate the translation of information between data analytics and subject matter expertise, using athlete recruitment as a framework for thought. These projects will be executed throughout the Fall 2021 and Spring 2022 semesters, with final client deployment expected by May 2022.

SPORTS ANALYTICS, GOLF, AND GAMEFORGE: INNOVATIVE ANALYTICS FOR RECOMMENDER SYSTEMS

The size and complexity of the college athlete recruitment industry has aggrandized tremendously over the past decade, with more than 73,000 additional students drafted to 1,268 schools across the United States over the past ten years and totaling more than 500,000 NCAA student-athletes in 2021 (National Collegiate Athletic Association [NCAA], 2021). The NCAA is now one of the most powerful sports organizations in the country, whose top 25 programs are projected to grow in revenue by 116% over the next ten years, a factor more than double the NFL, NBA, NHL, or MLB (Barakat, 2017). Many programs, like football and men's basketball, have abundant financial resources to support junior recruiting for university programs - the University of Georgia, for example, spent $3.7 million in 2019 on football recruitment alone, expending the most of the $56 million spent by the 52 public Power Five conference schools on their athlete recruiting efforts that same year (Wittry, 2020). Recruiting has since been

complicated further by recent changes in student-athlete name-image-likeness (NIL) laws, entrenched in the landmark Supreme Court case *National Collegiate Athletic Association v. Alston* from June 2021 that decided NCAA regulations on "education-related benefits" violated the Sherman Antitrust Act of 1890 and carried out through a plethora of statutes passed by state legislatures to uphold athlete rights (Chappell, 2021). The tumultuous growth and inequitable distribution of recruitment resources within the college athlete recruitment process highlights the demand for data-driven insights in identifying and enrolling athletes, especially for sports with smaller teams and lesser apparent retail value such as men's and women's golf. At the University of Virginia, the ongoing research between the Department of Systems Engineering and GameForge, a Virginia-based golf analytics firm, has provided a new opportunity to ameliorate this process and give power back to the players.

OBJECTIVES

GameForge, which aims to streamline the recruitment process for junior golfers to the best U.S. colleges, provides professional consulting services to junior golfers and their coaches from around the world by maintaining a dynamic database of tournament scores and an online portal accessible to customers (Bassilios et. al., 2021). Currently, the company provides an online system with in-depth analyses of junior golfers and their rankings comparable to different professional levels within golf (Golf Academy at Heritage Point, 2021). However, they seek to expand their services to collegiate golf teams by providing services that can recommend players for coaches to recruit using a variety of metrics, both related to and straying from the sport itself. The objective of this capstone project is to provide strategic guidance and statistical inference models that can be appended and deployed within the current system to deliver such insight for NCAA golf coaches. The resulting system will yield a two-sided college golf recruiting

recommendation system for both players and coaches, building upon the in-house data that GameForge maintains and utilizing advanced analytics techniques and machine learning to provide guidance that is nascent in the world of college golf recruiting.

METHODOLOGY, RESOURCES, AND OUTCOMES

Figure 1. Timeline of previous research on sports analytics in the UVa Dept. of Systems Engineering capstone program. Note the previous engagement between UVa and GameForge, now on its third iteration (Wasserman, 2021)

Due to the prevalence and longevity of the work previously completed between UVa and GameForge, much of the baseline research and development for statistical inference has been well-established and the focus now suggests the incorporation of new data sources and minor manipulation of previously deployed models to present such information to a new set of system users. After investigating the newest set of requirements put forth by the client and analyzing the interoperability of former classification learning models, the comprehensive system augmentation was broken down into three main division-of-labor efforts to be completed for this upcoming year, including (1) creating a recommender system that can identify junior players for

schools as well as delineate best opportunities for junior players; (2) developing an independent junior player ranking system that predicts collegiate success, proprietary to GameForge; and (3) enabling team-wide evaluation of players including what-if scenario simulations for addition of new players and identification of key player archetypes.

The player-school recommender system would work both in the interest of recruiting coaches and prospective athletes; the coach can utilize the tool to seek out and evaluate the impact of junior players on their teams, both at baseline performance and over time with coaching from their staff, while junior golfers could employ the platform to forecast skills development over time, evaluate likelihood for recruitment, and assess general school fit. The system update will also include a published ranking system similar to golf ranks maintained by longstanding authorities on the sport. The primary focus of this ranking will be to forecast collegiate success in junior golfers and track current success with collegiate golfers, requiring a longitudinal model that analyzes current collegiate players against varying metrics then regresses earlier statistics on the same athletes from before they were recruited. A linear multivariate model would then be deployed to compute a score for each athlete and the rank displayed for patrons of the site. Most unique with this ranking is the ability to scale the weightings for different metrics, unlike other published rankings where such weightings are static. The final component crucial to this tasking is in a team evaluation and player archetype evaluation tool. Using *k*-means clustering and classification learning, this tool will create what-if scenarios for coaching staff when evaluating new players and "bucket" players of similar skill and performance through the development of player archetypes.

On top of the yielded outcomes described above, the capstone team will author and submit a conference paper to the Institute of Electrical and Electronics Engineers (IEEE) Systems and

Information Engineering Design Symposium (SIEDS) for peer evaluation and professional judgement, to be held in Charlottesville, VA in April 2022.

THE GROWING DICHOTOMY BETWEEN EXPERTISE AND STATISTICS IN SPORTS ANALYTICS

Gordon Moore's popular observation on the exponential growth in the number of transistors in integrated circuits back in 1965 could have never sought to generalize the explosion that technology has seen over the past several decades, let alone the last ten years. First came the expansion of mobile computing technologies, with the advent of household computers then laptops and mobile phones, and now the infiltration of data science and machine learning as cornerstones of industry. Technology has become a dominant force as new capabilities have been enabled, increased access has been secured, and more educational and training tools have become easier than ever to access though the Internet (Kozyrkov, 2018). In many ways, this poses many opportunities – the ability to discover cures to incurable diseases, to solve environmental and energy crises, to enable humanitarian aid efforts in an end-to-end lifecycle from fundraising to delivery of services. However, as we have seen with the 2016 presidential election and the COVID-19 pandemic, technology also maintains a pivotal role in guiding public opinion, spreading information both true and false, and democratizing the voices of millions on a wide variety of topics.

As technology continues to develop in the modern age, so does its ability to aid and influence the decisions of experts across a variety of fields and industries, be it medicine, finance, or defense. More recently, the emergence of data science and analytics has enabled better-informed decision-making through the usage of data-driven insights and calculable,

quantitative results garnered through testing and simulation. As these tools become more popular, accessible, and easy to use and learn, the role of data analysts becomes aggrandized while the power of subject-matter experts diminishes, especially as technology promotes the idea that "suddenly, everybody's an expert" (Guernsey, 2000). This new relationship between data analysts and domain experts is best demonstrated by the story of the 2002 Oakland Athletics baseball season, in which a statistician brought on staff made better recruitment selections than a team of longstanding baseball experts through the use of generalized linear modeling. (Lewis, 2003).

This example highlights a key problem in the decision-making process for any given situation, that cognitive heuristics stand in the way of effective human-made decisions. There are a variety of reasons researchers have suggested this exists – some argue that our evolutionary constructs for quick decision-action formation from our hunter-gather days thousands of years ago continue to dwell on, while others blame modern social and environmental factors that have incentivized key attributes for evaluation compared to others (Yagoda, 2018). Whatever the reason for such cognitive biases, it is important to note the exhaustive number of points from which information is derived, and how individual perception of this information is skewed and variable from person-to-person.

Considering the great number of human-centric information points alongside those obtained from data analytics products, we quickly see how expansive the amount of information is that can be utilized to make even a single decision. Figure 2 (p. 8) adopts an actor-network to illuminate this matrix of information points, generalizing broader in scope when tracing outward in from more specific examples of points of acquired information to main constructs that drive the decision-making process.

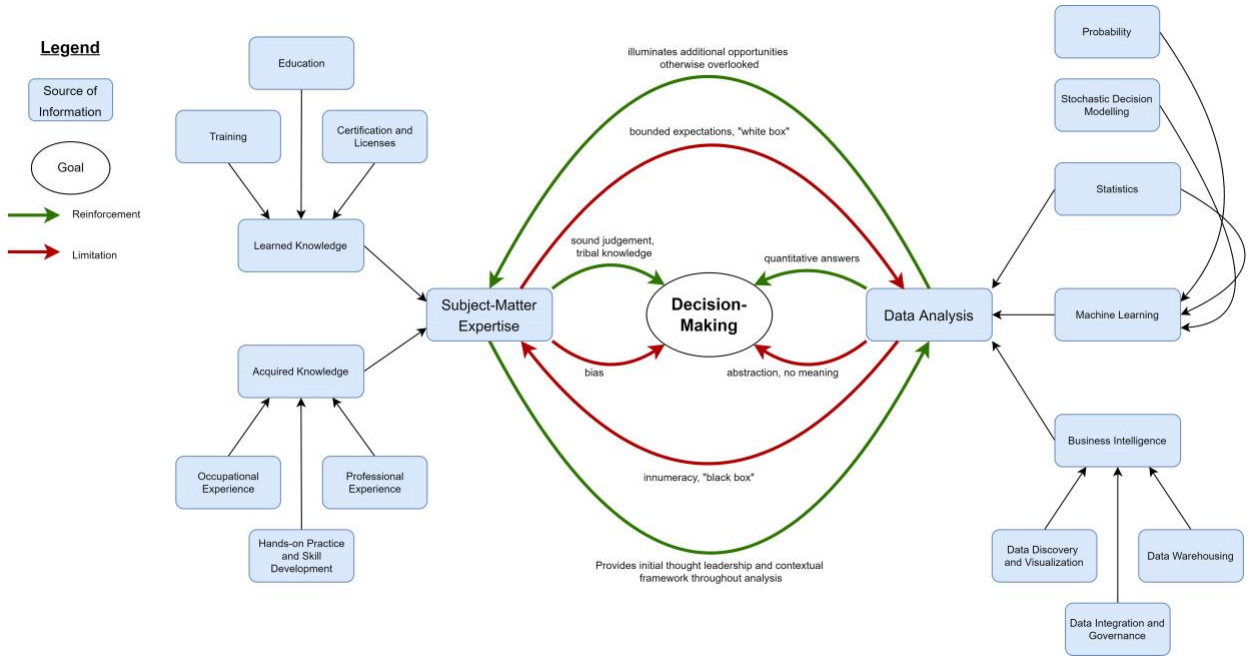

Figure 2. Information sources relevant to generic decision-making. Modern-day decision-making derives its reasoning from subject-matter expertise and data analysis, which can be further specified and compartmentalized into more basic constructs of intelligence. (Adapted by Wasserman (2021) from Jolviet & Heiskanen, 2010)

FURTHERED COMPLICATIONS

The idea that cognitive biases and heuristics, the mental shortcuts and systematic yet flawed patterns of thinking that arise in judgement and decision problems, cause dilemma in perception and analysis of information dates to the 1970s (Yagoda, 2018). Israeli social scientists Amos Tversky and Daniel Kahneman coined the phrase “cognitive bias,” and have since influenced the major critiques in the field of behavioral economics, including that of the Oakland A’s rise to success. In his novel *Moneyball*, author Michael Lewis suggests that statistics was able to demonstrate market inefficiencies in the recruitment of baseball players that A’s were then able to exploit in enhancing their team’s performance over the next season (Lewis, 2003). This book prompted criticism from a variety of academics, including economics professors Richard Thaler of Chicago Booth and Cass Sunstein of Harvard Law, who suggested instead that

the real highlight of the story told was not the suggestion of market inefficiencies, but rather was the discovery of underlying behavioral economics that caused leading experts in baseball to misjudge players when recruiting for professional teams (Thaler & Sunstein, 2003). This notion highlights the real issue that lies at the crux of our aforementioned problem – that foundational concepts rooted in behavioral economics cause complications in translating information between the subject matter experts (SMEs) and the data analytics tools that are expected to arrive at conclusions together.

ANALYZING THE PROBLEM

Cognitive biases in information retention and analysis amongst subject matter experts (SMEs) disable these authorities from trusting the bias-ephemeral answers that technology delivers. SMEs and technology must work in tangent to overcome the problem cognitive biases and deliver better-informed decisions that arise from a bottom-up evaluation of all available information. The objective of this research is to suggest a methodology of remediation using the foundational methodologies and models from the field of behavioral economics to aid in disabling the plethora of cognitive biases at-play when making data-informed decisions. This will be achieved by adapting a thought framework known as the Social Construction of Technology (SCOT), established by technological sociologists Wiebe Bijker and Trevor Pinch throughout the 1980s (Bijker, Bönig, & Oost, 1984; Bijker, Hughes, & Pinch, 1987). This SCOT analysis will provide a systemic look into how biases fit into the decision-making process, exhibiting points in the process that could be then ameliorated through adoption of the ideas from behavioral economics.

Figure 3 (p. 10) presents an early iteration on this SCOT model, showing a current state of decision-making without behavioral economics and ideal future state with the inclusion.

While this model is not complete, it demonstrates that the adoption of behavioral economics can have beneficial effects on the insights from both human-centric subject matter experts and technology-centric data analytics tools. The negotiation space between domains in the current state present weaknesses that the engineer must either relegate in significance, or better yet, regulate with the inclusion of behavioral economics in the suggested future state to prevent such weaknesses from sustaining within any generic decision-making system.

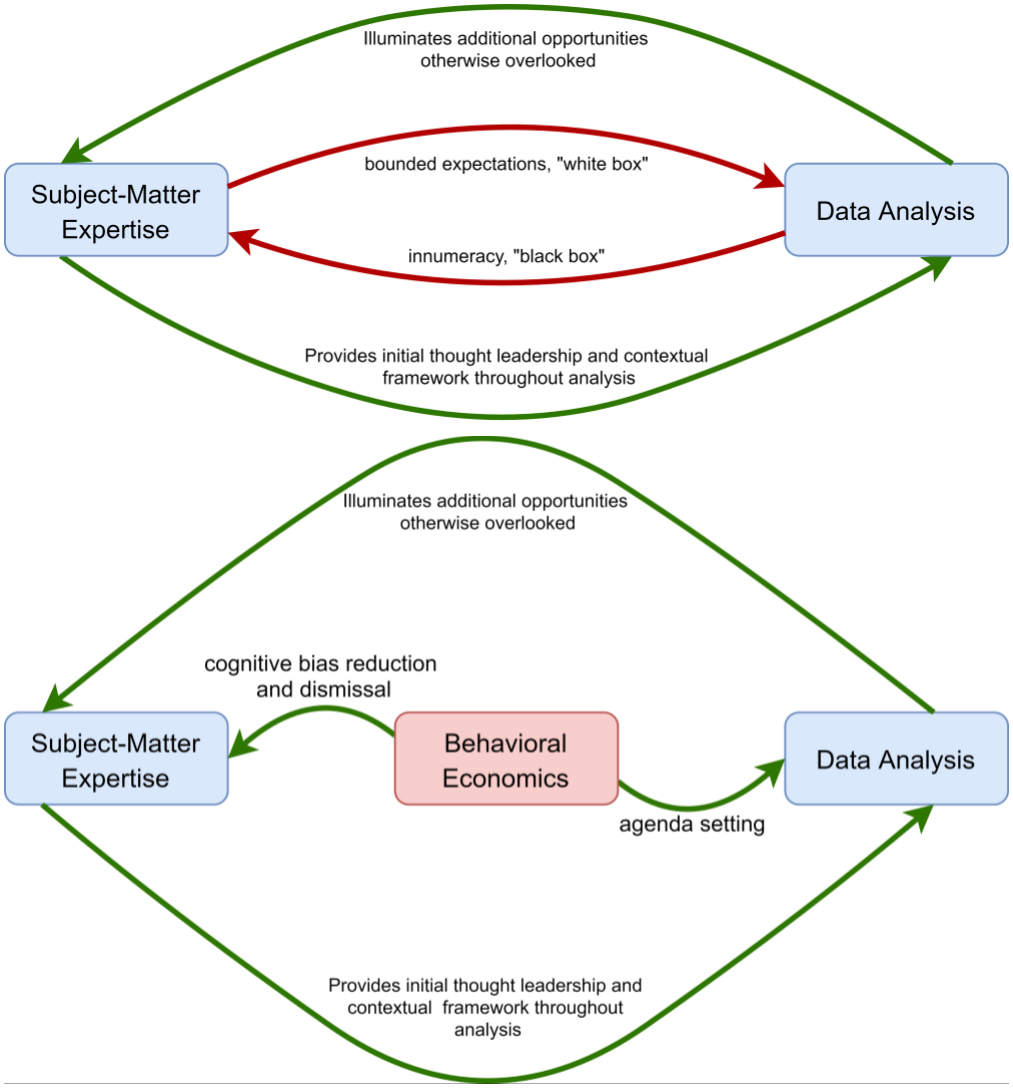

Figure 3. Current state versus suggested future state for generic decision-making system. By introducing behavioral economics principles, the negotiation space within the SCOT model is diminished or resolved altogether. (Adapted by Wasserman (2021) from Bijker, Bönig, & Oost, 1984)

BRINGING ABOUT CHANGE

As technology continues to evolve and mutate, so should our capacity evolve and mutate the necessary ideas and objectives even before the deployment and utilization of such technologies. Data analytics has provided a whole new realm of opportunity, as demonstrated here with the recruitment of junior athletes to some of the most powerful universities in the world. An exciting project, filled with a variety of tasks and responsibilities including the development of machine learning models, database query operations, and a user-facing interactive portal, also readily demonstrates an unprecedented cause for concern in the growing controversy between subject-matter expertise and statistical inference, and edifies how biases in expertise influences and at times negates the insights yielded from statistical analysis. If we can understand the cognitive biases that inhibit good domain expertise and better educate the fallacies of innumeracy that plague mainstream society, decision-makers can rest easy knowing they have information necessary at their disposal to guide the world forward with great positive impact.

REFERENCES

Barakat, C. (2017, June 21). The business of athletic recruiting. *Sparks Rowing*. https://sparksrowing.com/blog/the-business-of-athletic-recruiting

Bassilios, M., Jundanian, A., Barnard, J., Donnelly, V., Kreitzer, R., Adams, S., & Scherer, W. (2021, April 29-30). *Developing a recommendation system for collegiate golf recruiting* [Paper presentation]. IEEE Symposium on Systems and Information Engineering Design 2021, Charlottesville, VA, United States. https://doi.org/10.1109/SIEDS52267.2021.9483777

Bijker, V., Bönig, J., and Oost, E. (1984). Current state versus suggested future state for generic decision-making system [Figure 3]. The social construction of technological artefacts. *Zeitschrift für Wissenschaftsforschung*, 2, 39-52.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. The MIT Press.

Chappell, W. T. (2021, July 14). U.S. Supreme Court rules that the NCAA's limits on education-related benefits violet federal antitrust law. *JD Supra*. https://www.jdsupra.com/legalnews/the-u-s-supreme-court-s-ruling-attacks-5506149/

Durham, M. (2019, November 19). More college students that ever before are student-athletes. *National Collegiate Athletic Association*. https://www.ncaa.org/about/resources/media-center/news/more-college-students-ever-are-student-athletes

Golf Academy at Heritage Point (2021). Future junior champions. *Paul Horton Golf*. https://www.paulhortongolf.com/future-junior-champions/

Guernsey, L. (2000, February 3). Suddenly, everybody's an expert. *The New York Times*. https://www.nytimes.com/2000/02/03/technology/suddenly-everybody-s-an-expert.html

Jolviet, E. & Heiskanen, E. (2010). Information sources relevant to generic decision-making [Figure 2]. Blowing against the wind – An exploratory application of actor network theory to the analysis of local controversies and participation processes in wind energy. *Energy Policy*, 38, 6746-6754.

Kozyrkov, C. (2018, December 4). What great data analysts do — and why every organization needs them. *Harvard Business Review.* https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them

Lewis, M. (2003). *Moneyball: The art of winning an unfair game.* W. W. Norton & Company.

Lewis, M. (2016). *The undoing project: A friendship that changed our minds.* W. W. Norton & Company.

Metwalli, S. (2021, March 2). 5 tools to detect and eliminate bias in your machine learning Models. *Towards Data Science*. https://towardsdatascience.com/5-tools-to-detect-and-eliminate-bias-in-your-machine-learning-models-fb6c7b28b4f1

National Collegiate Athletic Association. (2021). *Spots sponsorship and participation research.* National Collegiate Athletic Association. https://www.ncaa.org/about/resources/research/sports-sponsorship-and-participation-research

Richter, F. (2021, July 2). U.S. college sports are a billion-dollar game. *Statista.* https://www.statista.com/chart/25236/ncaa-athletic-department-revenue/

Ringler, L. (2015, March 12). Anatomy of NCAA college golf: Where do players come from? *Golfweek.* https://golfweek.usatoday.com/2015/03/12/ncaa-college-golf-playersgeography-statistics/

Schrager, A. (2021, August 27). Behavioral economics doesn't have to be a total loss. *The*

Washington Post. https://www.washingtonpost.com/business/behavioral-economics-doesnt-have-to-be-a-total-loss/2021/08/27/ee38eeac-072a-11ec-b3c4-c462b1edcfc8_story.html

Thaler, R. & Sunstein, C. (2003, September 1). Who's on first. *The New Republic.* https://newrepublic.com/article/61123/whos-first

Wasserman, S. (2021). Current state versus suggested future state for generic decision-making system [Figure 3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Sciences, University of Virginia. Charlottesville, VA

Wasserman, S. (2021). Information sources relevant to generic decision-making [Figure 2]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Sciences, University of Virginia. Charlottesville, VA

Wasserman, S. (2021) Timeline of previous research on sports analytics in the UVa Dept. of Systems Engineering capstone program [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Sciences, University of Virginia. Charlottesville, VA

Wittry, A. (2020, February 23). An analysis of college football recruiting costs. *AthleticDirectorU*. https://www.athleticdirectoru.com/articles/an-analysis-of-football-recruiting-costs/

Yagoda, B. (2018, August 4). The cognitive biases tricking your brain. *The Atlantic.* https://www.theatlantic.com/magazine/archive/2018/09/cognitive-bias/565775/