

# Enhancing Public Data Availability and Analysis of Olympic Sports: The Case of College Swimming

Anonymous

---

**Abstract.** While during the last several years popular team sports have experienced a growth in terms of data and analysis that are publicly available, this is not the case with Olympic sports. While national Olympic committees are reportedly using data to make decisions, “public analytics” have not followed suit. Part of the reason can be attributed to the lack of readily available open datasets to the public for these sports. This work aims at filling this gap by developing an open source application for downloading and analyzing data from college swimming. More specifically the application obtains data and processes them in a machine readable format from [swimcloud.com](https://swimcloud.com). Furthermore, we provide an interactive application for visualizing and analyzing the data with a focus on two specific applications: (a) swimmer progression across seasons, and (b) tapering during the season in terms of achieving optimal performance in their respective conference finals. We hope that this work will lead to more public interest and analysis in swimming and Olympic sports in general.

**1. Introduction.** Every 4 years (or more if there is a pandemic during the fourth year), athletes all over the world compete for the highest athletic achievement, a (gold) medal in the Olympic games. In between these four years, national Olympic Committees (OC) try to put assemble a team that will get on the podium the maximum possible number of times. OCs are looking for any *edge* they can find when making these decisions and hence, they turn to data to help them guide their decisions. Data can and are being used in this setting with a variety of objectives, including, identifying and investing to athletes and sports with the best chances of success, optimizing funding allocation, and supporting traditional sports science and medicine [3, 6].

For some sports - mainly running and swimming - there have been studies on the fringe of sports science and analytics, examining the relationship between pacing in long-distance *races* and performance [4, 2, 9, 8]. An interesting use of lap data from the 2016 Olympics in Rio also indicated that there was a lane bias, with athletes swimming faster in one direction as compared to the opposite one depending on the lane they were swimming [5].

However, beyond academic research and *specialized* websites/blogs Olympic sports lack a strong public analytics community. One of the potential reasons is the lack of available open data, since (the academic publications and websites we looked at, did not provide the dataset used to generate their analysis). This work aims at partially bridging this gap by collecting, curating and making available data from collegiate swimming competitions from [swimcloud.com](https://swimcloud.com). We also provide the source code of our data collection process for anyone interested in replicating the effort and collecting a different set of competitions (details are provided in Section 2). Moreover, we showcase the usage of these data through two applications: (i) swimmers’ growth curves, i.e., how do swimmers develop from season to season in terms of performance, and (ii) tapering for conference finals, that is, achieving their optimal performance during their conference finals. Finally, we have developed an interactive application that allows a user explore our data and results. We hope this work to spark more interest in the public analysis of Olympic sports.

The rest of this paper is organized as following: Section 2 describes the process of collecting data from [swimcloud.com](https://swimcloud.com), as well as, the specific data we collected. Section 3 presents the two showcase applications aforementioned, while we conclude our work in Section 4.

**2. Data Collection.** The website [swimcloud.com](http://swimcloud.com) has a very large collection of data for college swimming. However, for a non-expert, it is hard to extract data at large from the website for further analysis and identification of patterns and trends. For example, Figure 1 presents a table that one can find on the website for Pitt's swimmer Vera Blaise. This table presents Blaise's performance during the conference (ACC) finals this year. We can see a lot of information including his times during prelims, finals, timed finals, as well as, his improvement (positive or negative) as compared to his qualifying time for the finals. This is a lot of information that can be extremely helpful for analyzing the performance of an individual swimmer or a team.

However, there is no way to store this information in a machine readable format for further analysis. While one can certainly manually collect this information, it should be evident that this does not scale to anything beyond anecdotal analysis. To facilitate similar efforts we have built a Python scraper to automate this process. We have also collected and curated data for all ACC teams, both men and women, since 2013. The collected data includes information for more than 2,000 male and female swimmers, including personal times for more than 17,000 events in the ACC finals. The data collected includes the following three tables:


- **Swimmers' information.** This table includes some basic information about each swimmer, including their ID (which will be used to match entries in the different tables), their name, hometown and team.
- **Swimmers' times.** This table includes the times of each swimmer at different events (currently the dataset covers all the ACC finals since 2013). The format of this table is: `Swimmer_ID, Team, Year, Event, Event_Type, Time, Improvement`. The event type should be interpreted in the context of the specific event. For example, for conference finals, the `Event_Type` can be either a prelim round, a timed final (i.e., an official "practice" time) or a final race. The `Improvement` column provides the percentage difference between the time achieved at the specific race and the qualifying time for the conference finals.
- **Swimmers' standings.** This table includes information for the points contributed to their team by each swimmer, in each of their seasons. The format of this table is: `Swimmer_ID, Starting_season, Power_Index, FR, Events-FR, Freshman_PPE, SO, Events-SO, Sophomore_ppe, JR, Events-JR, Junior_ppe, SR, Events-SR, Senior_ppe, Total_Points, Total_Events, Total_ppe`. This table includes information about the number of points each swimmer earned for their team during each season (freshmen, sophomore, junior and senior), as well as the number of events they participated in each of these seasons along with the points per event (PPE). The `Power_Index` is an indicator of their high school recruiting rank. It takes values between 1 and 100, with 1 being the top prospect. For some swimmers (mainly international) this information is missing (marked as -1).

Our scraper is also publically available<sup>1</sup> in order for people to use it for downloading their preferred data. There are several functions that are implemented that facilitate the data collection. While more details are provided with the code repository, here are the main functions currently implemented:

- `getRoster`: This function takes as an input the season and the team and returns a list with the swimmer IDs on the team's roster

---

<sup>1</sup>You can find the scraper and the data collected on the following github repository: <not provided for double blind review purposes>.

<div>  2020 Atlantic Coast Championships (M) ▾  Feb. 26–29, 2020 </div>					
Event	Time	Place	Round	Imp	
50 Y Free	19.48 <span>R</span>	–	Timed Finals	-2.0%	
50 Y Free	19.54	6th	Finals	-2.3%	
50 Y Free	19.32	3rd	Prelims	-1.2%	
100 Y Free	42.47 <span>B</span> <span>R</span>	–	Timed Finals	+0.9%	<span>SB</span>
100 Y Free	42.61	9th	Finals	+0.5%	<span>SB</span>
100 Y Free	43.14	10th	Prelims	-0.7%	
50 Y Fly	21.21 <span>X</span>	–	Finals	+0.6%	<span>SB</span>
50 Y Fly	21.10 <span>X</span>	–	Prelims	+1.1%	<span>SB</span>
100 Y Fly	46.22	6th	Finals	+1.7%	<span>SB</span>
100 Y Fly	46.21 <span>B</span>	8th	Prelims	+1.7%	<span>SB</span>

**Figure 1.** An example of a data table that can be found on [swimcloud.com](https://swimcloud.com).

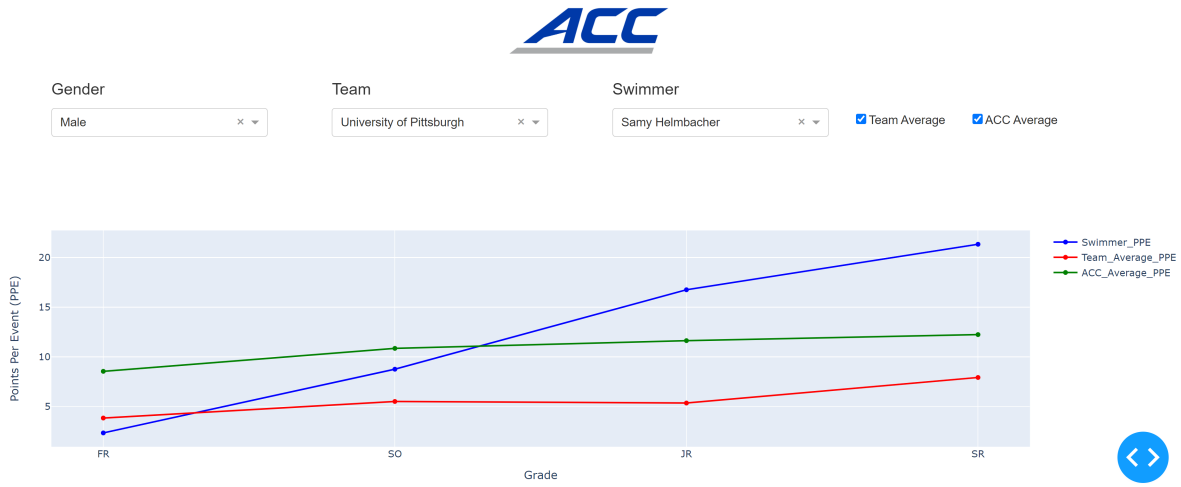
- `findEntries`: This function takes as an input a swimmer ID and a competition/meeting and returns the events that the swimmer participated during the competition (in our data the competition is the ACC championship)
- `getData`: This function (calls the function `getPoints`) takes as input a team and a competition and returns the points gained from each swimmer on the roster in that meeting
- `getPowerIndex`: This function takes as input a swimmer and returns is `Power_Index`, i.e., his ranking as a high-school prospect

These functions are currently implemented as a (fully functioning) stand alone script, but we plan on transforming them to a appropriate python library that will serve as an API to the underlying functionality.

**3. Modeling Applications.** In the rest of this paper we focus on showcasing the use of the data collected by focusing on two specific applications, namely, projecting the growth of a swimmer in terms of conference finals performance, as well as, the ability of a swimmer to perform their best during the conference finals.

**3.1. Swimmer Growth Curve.** Every swimmer participates at a maximum of 7 events during conference finals and based on their performance they earn points for their team. Therefore, one of the possible ways to evaluate the contribution of a swimmer  $s$  to their team is by calculating the average points per event,  $\pi_s$ , that they earned. Using  $\pi_s$  we can then explore whether swimmer  $s$  follows the expected trajectory of their peers - both overall, as well as, within the same team. We can think of this as the swimmers' growth curves. For example Figure 2, visualizes the development (in terms of points per event) of all swimmers over their college career. We also have highlighted a specific swimmer and we can visually compare their growth with the rest of the pool (no pun intended).

Using these data we can start estimating the relationship between the college a swimmer attempted and their overall performance. For example, if we build a model for  $\pi_s$  given their recruiting power



**Figure 2.** An example of a swimmer's growth curves compared to the conference and team averages.

index and the college they attended can we rank schools based on the effect they have on the swimmers' performance? Essentially, if we adjust for the “inherent” talent of a swimmer does the school they attend correlates with their overall performance? Of course, it goes without saying that this is a purely observational analysis and we cannot assign causality to these claims.

For this we build a Bayesian linear regression [1], where we model  $\pi_s$  as a normal distribution, whose mean is a linear combination of a set of independent features, including the college  $c$  they attended and their recruiting power index  $r$ :

$$(3.1) \quad \pi_s \sim \mathcal{N}(a_0 + a_c \cdot c + a_r \cdot r, \sigma^2)$$

The goal is to estimate the posterior distribution of the coefficients  $a_0$ ,  $a_c$  and  $a_r$ , as well as,  $\sigma$ . Given that the college attended is a categorical variable, we will essentially get a set of coefficients, one for each team. We also assign the following prior distributions for these parameters:

- $\sigma$ : Half Cauchy distribution with  $\beta = 10$
- $a_0, a_c, a_r$ : normal distribution with 0 mean and standard deviation of 5

We run the model for both male and female teams using Markov Chain Monte Carlo (MCMC) with 5 chains and 2000 samples each. The results are presented in Table ???. This table provides the expected value of the posterior for each variable, as well as, the 97% credible interval of this distribution. The way to interpret these results are as follows. Let's take Boston College men's team as an example. A swimmer that gets recruited and attends Boston College, is projected to perform on average almost 3 points per event lower than expected based on his recruiting power index. The benefit of the Bayesian approach is that we get the full posterior distribution. With this in the case of Boston College we can see that there is still a small probability ( 6%) that this effect is less than half a point. Of particular interest for the men's teams appears is the University of Miami, where it has the largest expected value out of all the schools, but it also exhibits a much higher uncertainty compared to other schools. The main reason for this is that there are only 10 data points for swimmers from the University

Variable	Coefficient	CI (3%) (Men)	CI (97%) (Men)	Coefficient	CI (3%) (Women)	CI (97%) (Women)
Intercept	6.5	3.77	9.16	6.0	3.47	8.69
Power Index	-0.14	-0.19	-0.1	-0.12	-0.154	-0.09
$\sigma$	6.49	6.15	6.72	6.43	6.14	6.7
UNC	-0.72	-3.6	2.23	3.24	0.25	6.17
Miami (FL)	11.0	4.36	17.3	-1.92	-4.99	1.04
Duke	-2.65	-5.53	0.44	-0.47	-3.38	2.47
BC	-2.93	-5.88	0.1	-2.09	-5.02	0.733
NCSU	4.42	1.57	7.34	4.13	1.31	7.02
Virginia	2.24	-0.75	5.14	4.87	1.94	7.71
FSU	-1.17	-4.19	1.58	0.19	-2.73	2.94
Pitt	-3.1	-5.97	-0.22	-1.78	-4.57	1.12
Gergia Tech	-2.03	-4.98	0.91	-2.52	-5.44	0.42
Louisville	0.96	-1.92	4.0	2.3	-0.71	5.08
Notre Dame	0.07	-2.81	3.01	0.25	-2.63	3.15
Virginia Tech	0.06	-2.78	3.1	-0.73	-3.48	2.17

Table 1

The coefficients and credible intervals for the Bayesian regression for  $\pi_s$ .

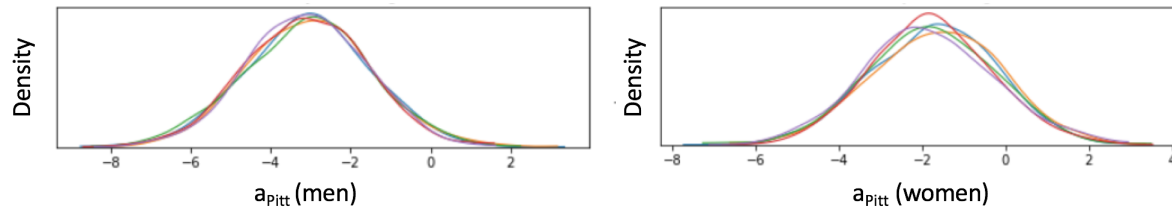


Figure 3. The posterior distribution for the coefficient associated with the University of Pittsburgh's men's and women's teams.

of Miami (the next fewest number of data points is 81 for Duke). For women's team the University of Virginia and North Carolina State have the highest expected points per event improvement for their swimmers as compared to what one might have expected based on the recruiting power index of these swimmers. Figure 3 further presents the posterior distribution obtained from each MCMC chain for the University of Pittsburgh.

A probabilistic *ranking* of teams obtained from these coefficients can serve as a way to assess the *coaching* effect on swimmers. It is clear that different colleges are able to recruit different quality of swimmers. However, do schools help swimmers succeed beyond what is expected from their prior skill/quality? For example, University of Virginia is considered one of the best swimming schools in the ACC, but they are able to recruit male swimmers that are highly ranked (average power index is 18.3; the lower the better). On the contrary, the University of Miami recruits students with an average

power index of 79.8; much lower compared to Virginia. However, they seem to be able to equip these swimmers with the appropriate tools for succeeding, performing on average 10 points per event higher than what expected. Now we have to provide here the caveats we mentioned above that we have fairly few data for the University of Miami, and that these relationships identified are not causal.

**3.2. Season-Ending Taper.** Next we are interested in examining whether we can identify team effects on the improvement of swimmers' performance within the season. Tapering is a training practise based on which for a period of time ranging from a few days, up to 3 weeks athletes fine-tune their swimming, while the volume of the work in the pool is cut drastically [10, 7]. The goal is to *optimize* the performance during the most important competitions. Being able to be on top-form during the conference finals is a result of good tapering and this could/should relate with coaching, as well as, strength and conditioning training. While of course we cannot know if all teams include in their training season-ending taper (even though it is highly unlikely that any of them does not), we are focusing on the times recorded by the swimmers during their conference finals. In particular, we will examine the percentile improvement in each event as compared to their qualifying time for the same event.

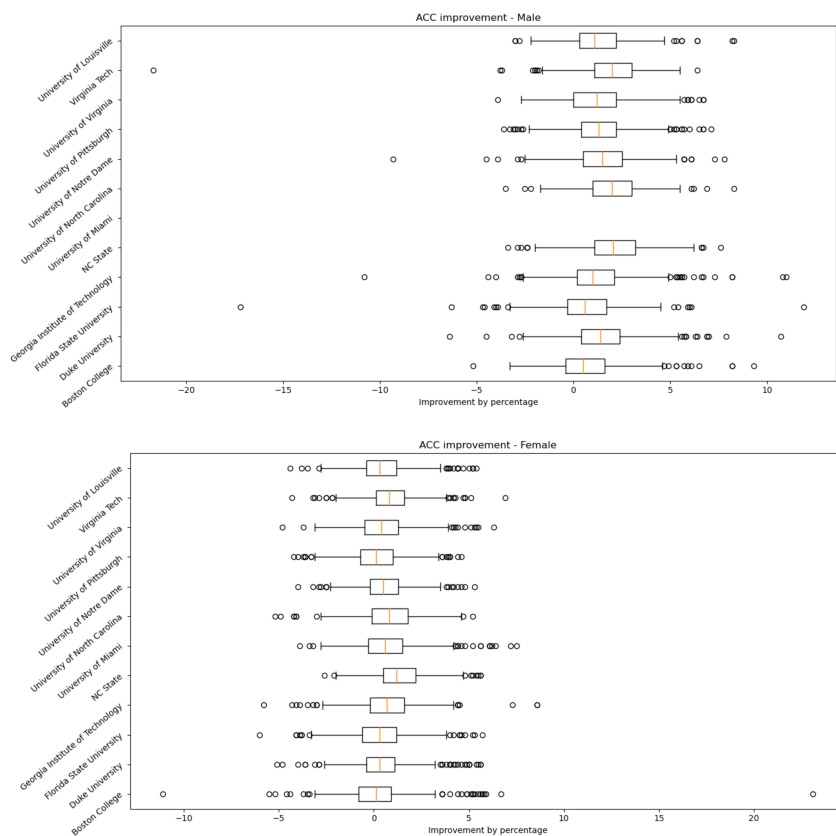
Figure 4 presents the boxplot for all the swimmers of each school (across all the seasons covered by the dataset) and their percentage improvement for every event. As we can see, overall there is a trend for an aggregate (but not statistically robust) improvement during the conference finals at the end of the season. In particular, for male swimmers there is an average improvement overall of 1.41% (with standard deviation of 1.75), while for female swimmers there is an average improvement of 0.6% (with standard deviation of 1.5). Nevertheless, overall there do not seem to be specific schools that exhibit significant improvement compared to their peers in terms of improvement at the conference finals.

To further examine this potential school effects, we use an approach similar to that of the previous section. In particular, we model the percentage improvement during conference finals as compared to the corresponding qualifying time through a Bayesian regression. The dependent variable is this percentage change between the qualifying time and the time during the finals  $\delta$ . Again we model this through a normal distribution whose expected value is a linear combination of the school  $c$  the swimmer attended and the event type  $e$  (e.g., 50 meter free etc.):

$$(3.2) \quad \delta_s \sim \mathcal{N}(b_0 + b_c \cdot c + b_e \cdot e, \sigma^2)$$

We again use a normal distribution with mean 0 and standard deviation 5 for all the independent variables and a half Cauchy for the model standard deviation. Table 2 presents the results for both men and women. Here a smaller coefficient  $b_c$  is better (ideally negative), since this corresponds to smaller percentage increase in the swimming time compared to the qualifiers. As we can see, the posterior distributions for the school effects are more spread as compared to the regression on the points per event, and they all cover both positives and negative values. Essentially, it appears that in terms of conditioning and achieving their best performance during the conference finals, the school attended is not associated with any quantifiable benefit (or setback) on a swimmer's time.

**3.3. Interactive Application.** We have also developed an interactive web application with which a user can explore the data and metrics discussed above. The application is available at: <https://accswimming.herokuapp.com>. The dashboard is implemented using the Python library `dash`, while



**Figure 4.** Boxplot for the aggregate percentage improvement of all swimmers per school (top: male, bottom: female)

we also provide the source code. There are two tabs that can choose between the two type of results. The one tab allows a user can select a swimmer of choice and obtain their progression through their college career in terms of points per event. There are also control boxes for plotting the conference and team average curves for comparison. Furthermore, a user can choose the second tab and identify explore the time improvement for a chosen player and event during the conference finals.

**4. Conclusions and Discussion.** The main objective of our study is to provide a footprint for those interested in performing analysis of Olympic sports. We develop this footprint using college swimming as our testbed, and we begin by developing an open source crawler for collecting detailed data from [swimcloud.com](https://swimcloud.com). Using this crawler we collected data from the ACC since the 2012 season. We further showcased the usefulness of these data by modeling through Bayesian linear regression the school effects on the performance of swimmers in terms of points earned for the school, as well as, times recorded during the conference finals. Our results indicate that in terms of points earned per finals event, after adjusting for the *skill/talent* of a freshman swimmer, there are programs that *stretch* the output of an athlete beyond what is expected from a similar swimmer.

While our analysis at this point is more descriptive - e.g., ranking schools based on how they have improved their incoming talent - we believe that our efforts in collecting these data and making it easier



Variable	Coefficient	CI (3%) (Men)	CI (97%)	Coefficient	CI (3%) (Women)	CI (97%)
intercept	1.48	-2.02	4.89	0.75	-2.97	4.11
sigma	1.67	1.64	1.70	1.49	1.46	1.51
BC	-0.54	-3.16	2.24	-0.49	-3.29	2.18
Duke	0.26	-2.36	3.04	-0.24	-2.94	2.52
FSU	-0.58	-3.14	2.25	-0.32	-3.15	2.31
Georgia Tech	-0.02	-2.69	2.69	0.06	-2.67	2.80
NCSU	0.93	-1.64	3.75	0.74	-1.96	3.51
Louisville	0.04	-2.53	2.84	-0.14	-2.94	2.53
Miami	—*	—*	—*	0.09	-2.61	2.87
UNC	0.80	-1.80	3.58	0.26	-2.50	2.97
Notre Dame	0.25	-2.32	3.07	-0.06	-2.84	2.63
Pitt	0.07	-2.54	2.86	-0.46	-3.19	2.28
Virginia	-0.04	-2.61	2.78	-0.16	-2.90	2.57
Virginia Tech	0.79	-1.69	3.69	0.25	-2.51	2.96
100YBack	-0.11	-2.63	2.05	-0.15	-2.34	2.28
100YBreast	-0.25	-2.63	2.05	-0.21	-2.44	2.20
100YFly	0.07	-2.42	2.26	0.02	-2.22	2.42
100YFree	-0.42	-2.89	1.79	-0.02	-2.32	2.31
1650YFree	-0.50	-2.94	1.76	-0.16	-2.37	2.27
200YBack	-0.17	-2.62	2.05	-0.14	-2.39	2.25
200YBreast	-0.10	-2.53	2.13	-0.26	-2.49	2.15
200YFly	-0.30	-2.72	1.95	-0.46	-2.70	1.92
200YFree	-0.63	-3.07	1.61	-0.13	-2.47	2.17
200YIM	0.10	-2.32	2.35	-0.17	-2.39	2.24
400YIM	-0.17	-2.58	2.10	-0.19	-2.44	2.21
500YFree	-0.65	-3.07	1.61	-0.44	-2.73	1.91
50YBack	0.55	-1.94	2.75	0.92	-1.25	3.42
50YBreast	1.69	-1.59	4.68	—*	—*	—*
50YFly	1.85	-1.63	5.65	—*	—*	—*
50YFree	-0.39	-2.89	1.80	0.12	-2.13	2.49

Table 2

The coefficients and credible intervals for the Bayesian regression for  $\delta_s$ . (\*: no data for swimmers/event.)

for people to access them, can have positive effects in growing the interest of the analytics community. In the near future we plan on transforming the stand alone crawling scripts to an API through a python library that will significantly improve the user friendliness aspect of the software. Finally, we plan to focus on *predictive* applications of these data, and in particular identifying the *optimal* strategy for allocating events to swimmers during finals. For instance, is it better to have a swimmer participate in 4 events and perform slightly above average in all of them, or participate in 2 events and excel in both?



## REFERENCES

- [1] G. E. BOX AND G. C. TIAO, *Bayesian inference in statistical analysis*, vol. 40, John Wiley & Sons, 2011.
- [2] A.-W. DE LEEUW, L. A. MEERHOFF, AND A. KNOBBE, *Effects of pacing properties on performance in long-distance running*, *Big Data*, 6 (2018), pp. 248–261.
- [3] R. B. FERGUSON, *Team gb: Using analytics (and intuition) to improve performance*, *MIT Sloan Management Review*, 54 (2013), p. 1.
- [4] C. FOSTER, M. SCHRAGER, A. C. SNYDER, AND N. N. THOMPSON, *Pacing strategy and athletic performance*, *Sports Medicine*, 17 (1994), pp. 77–85.
- [5] J. GUO, *These charts clearly show how some olympic swimmers may have gotten an unfair advantage.* <https://www.washingtonpost.com/news/wonk/wp/2016/09/01/these-charts-clearly-show-how-some-olympic-swimmers-may-have-gotten-an-unfair-advantage/>, September 2016.
- [6] F. KIRWAN, *Data driven resource allocation: A us olympic committee perspective*, *Sports Analytics Innovation Summit*, (2015).
- [7] Y. LE MEUR, C. HAUSSWIRTH, AND I. MUJKA, *Tapering for competition: A review*, *Science & Sports*, 27 (2012), pp. 77–87.
- [8] K. E. MCGIBBON, D. PYNE, M. SHEPHARD, AND K. THOMPSON, *Pacing in swimming: a systematic review*, *Sports Medicine*, 48 (2018), pp. 1621–1633.
- [9] S. G. P. MENTING, M. T. ELFERINK-GEMSER, B. C. HUIJGEN, AND F. J. HETTINGA, *Pacing in lane-based head-to-head competitions: A systematic review on swimming*, *Journal of sports sciences*, 37 (2019), pp. 2287–2299.
- [10] I. MUJKA, *The influence of training characteristics and tapering on the adaptation in highly trained individuals: a review*, *International journal of sports medicine*, 19 (1998), pp. 439–446.