– Independent Work Report Spring, 2016 –

Understanding Collegiate Track and Field Data:
Can we help high schoolers make more informed decisions?

Katie Hanss
Adviser: David Dobkin

April 29, 2016

Abstract

High school Track and Field athletes often have the privilege to choose between schools during the recruiting process; however this decision can be very difficult. We believe that the copious amounts of Track and Field data can help high school recruits evaluate the athletic performance of different teams. However, the way the data is currently organized and processed is not conducive to this. In this paper we present a new way to extract information on the athletic performances of teams from Track and Field result. We extract a number of team-metrics from the data that describe the performance of teams across 19 different events. We explore the metrics to answer interesting questions about Track and Field teams. Finally, we present a recruiting recommendation system which attempts to rank different teams according to a high school athlete's performance and personal preferences.

1. Introduction

High school Track and Field athletes often have the privilege to choose between schools during the recruiting process; however this decision can be very difficult. If you are being recruited, you generally have a number of schools which are interested in you and in the end, you have to decide where you want to commit. To get a sense of the schools – both athletically and academically – you go on recruiting visits. You meet with the coach, hear about the team's fastest runners, talk to the athletes, eat in the dinning halls and maybe attend a class or two. But the weekend-long visits are overwhelming. Everyone is trying to convince you that their school, their team, is the best. It can be hard to dig through all the information, understand what teams have to offer and decide which school is the right fit for you.

There are a number of factors a student athlete might consider as they try to choose a school. How big is the school? How big is the team? Is the coach good? Do they like the team culture? How are the academics? How are the athletics? Within the athletic performance of a school there are even more questions. How much do athletes on the team improve? How often does the team win? How often does the team compete? How fast are the athlete's best times? How fast are the team's school records?

While the more qualitative questions are very difficult to answer, we believe that Track and Field data can help high school athletes understand the athletic performance of teams. What if, instead of evaluating a school based on their best athletes or the information you absorb on your recruiting visit, you knew exactly how many seconds or meters athletes on the team improved? What if you knew how the team performed within their league and their athletes' best marks over the past 7 years? This information could help you compare schools and make a more informed decision about where you would like to go to college.

In this paper, we present a method to extract information on athletic performance from collegiate Track and Field data. We describe a way to gather Track and Field data from the Track and Field Results Reporting System (TFRRS, Direct Athletics Inc. [14]). We identify and extract team metrics which we believe will help high school athletes evaluate teams. We use the team metrics to investigate interesting questions about collegiate Track and Field teams: Do teams naturally cluster into their NCAA divisions? Do some teams improve their athletes more than others? Does competing more frequently make a team better? We propose a recruiting recommendation system which outputs a list of schools that a high school athlete might be interested in given their athletic ability and their answers to a number of questions. The recommendation system uses a high schooler's gender, best event and best mark in that event to find a subset of schools which might recruit the athlete. It asks questions about how much they care about the spread of the team, best marks on the team, improvement on the team, success of the team and the frequency of competitions.

It uses these questions to estimate which schools would be the best fit.

2. Background and Related Work

In this section, we give a brief introduction to Collegiate Track and Field. We discuss the NCAA and TFRRS, the system used to collect Track and Field results. Finally, we discuss how Track and Field data is currently used to evaluate teams and the limitations of these methods.

2.1. An Introduction to Collegiate Track and Field

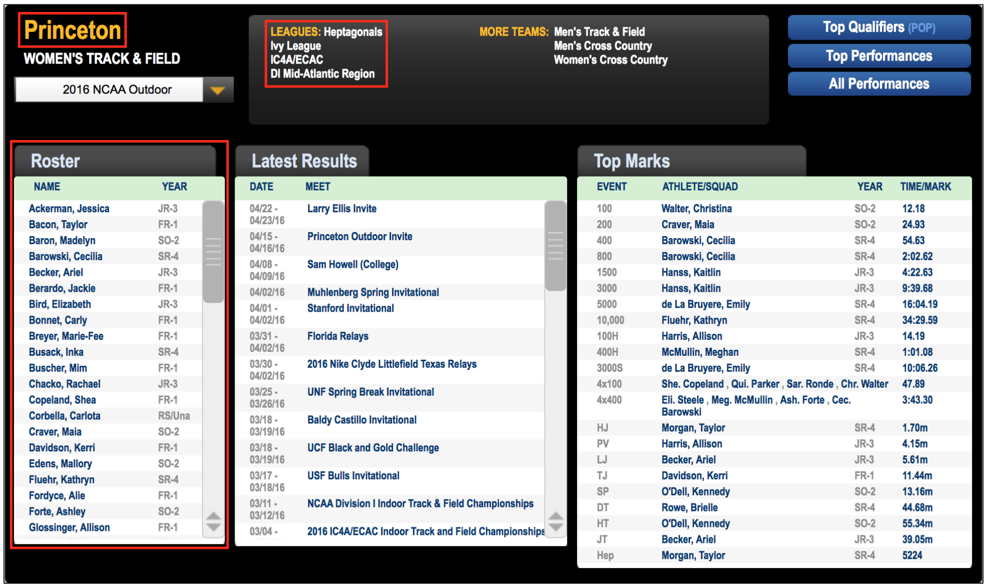
The National Collegiate Athletic Association (NCAA) is the largest governing body of collegiate athletics in the United States. It boasts 1,121 members [15] who have trained over 130,000 athletes since 2009.[1] Because it is so large, the NCAA is split up into three divisions, DI, DII and DIII based on a school's size and athletic performance. In Track and Field, DI is comprised of big schools and is the largest and highest performing devision in the NCAA. DIII is comprised of smaller schools, and is not as competitive as DI. DII is the smallest of the three divisions and is comprised of medium schools which are somewhere between the level of DI and DIII. The three divisions have separate National Championships and offer different collegiate Track and Field experiences.

Not only do the large number of team's warrant three divisions, but they also produce copious amounts of track and field data each year. In an attempt to organize this information, the NCAA partnered with a private company, Direct Athletics Inc., in 2009 to create a centralized database for track and field results called TFRRS (the Track and Field Results Recording System, Direct Athletics Inc.) [6]. Though owned and managed by Direct Athletics Inc., the database is the official source of information for NCAA Track and Field and is used to rank teams, to award "Academic All American" honors, to decide who goes to NCAA Indoor Nationals and to predict who might win NCAA Championships [6].

As a database, TFRRS is accurate and large. It not only contains data for the entire NCAA (including all divisions) and schools outside the NCAA but also, results are available from 2009 to the present and will only continue to grow. It has a page for each Track and Field team, which

[1] We produced this figure via the data we collected from TFRRS [14]

contains a team roster (Figure 1). In addition, it has a page for each athlete, which contains a list of every competition the athlete has participated in throughout their collegiate career (Figure 2). Meet directors, coaches and athletes have a vested interest in maintaining the accuracy of their results on TFRRS since it is integral in determining who competes at NCAA Regionals and Championships. Because the TFRRS accurate and complete Track and Field results, we chose it as the data sources for our project.



Princeton
WOMEN'S TRACK & FIELD
2016 NCAA Outdoor

LEAGUES: Heptagonals
Ivy League
IC4A/ECAC
DI Mid-Atlantic Region

MORE TEAMS: Men's Track & Field
Men's Cross Country
Women's Cross Country

Roster

NAME	YEAR
Ackerman, Jessica	JR-3
Bacon, Taylor	FR-1
Baron, Madelyn	SO-2
Barowski, Cecilia	SR-4
Becker, Ariel	JR-3
Berardo, Jackie	FR-1
Bird, Elizabeth	JR-3
Bonnet, Carly	FR-1
Breyer, Marie-Fee	FR-1
Busack, Inka	SR-4
Buscher, Mim	FR-1
Chacko, Rachael	JR-3
Copeland, Shea	FR-1
Corbella, Carlota	RS/Una
Craver, Maia	SO-2
Davidson, Kerri	FR-1
Edens, Mallory	SO-2
Fluehr, Kathryn	SR-4
Fordyce, Alie	FR-1
Forte, Ashley	SO-2
Glossinger, Allison	FR-1

Latest Results

DATE	MEET
04/22 - 04/23/16	Larry Ellis Invite
04/15 - 04/16/16	Princeton Outdoor Invite
04/08 - 04/09/16	Sam Howell (College)
04/02/16	Muhlenberg Spring Invitational
04/01 - 04/02/16	Stanford Invitational
03/31 - 04/02/16	Florida Relays
03/30 - 04/02/16	2016 Nike Clyde Littlefield Texas Relays
03/25 - 03/26/16	UNF Spring Break Invitational
03/18 - 03/19/16	Baldy Castillo Invitational
03/18 - 03/19/16	UCF Black and Gold Challenge
03/17 - 03/18/16	USF Bulls Invitational
03/11 - 03/12/16	NCAA Division I Indoor Track & Field Championships
03/04 -	2016 IC4A/ECAC Indoor Track and Field Championships

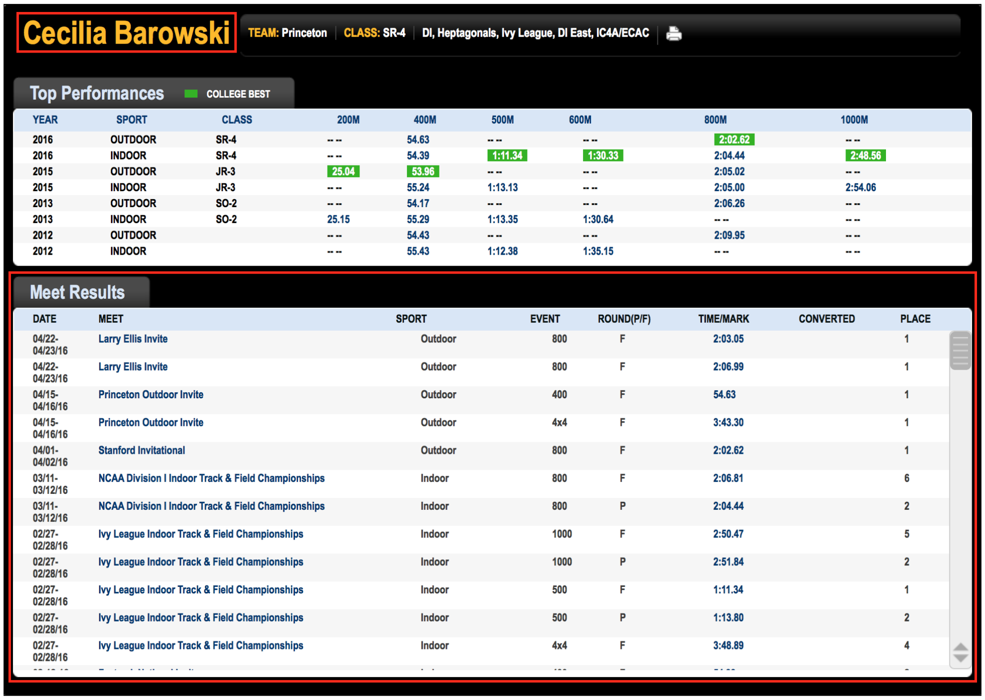
Top Marks

EVENT	ATHLETE/SQUAD	YEAR	TIME/MARK
100	Walter, Christina	SO-2	12.18
200	Craver, Maia	SO-2	24.93
400	Barowski, Cecilia	SR-4	54.63
800	Barowski, Cecilia	SR-4	2:02.62
1500	Hanss, Kaitlin	JR-3	4:22.63
3000	Hanss, Kaitlin	JR-3	9:39.68
5000	de La Bruyere, Emily	SR-4	16:04.19
10,000	Fluehr, Kathryn	SR-4	34:29.59
100H	Harris, Allison	JR-3	14.19
400H	McMullin, Meghan	SR-4	1:01.08
3000S	de La Bruyere, Emily	SR-4	10:06.26
4x100	She. Copeland , Qui. Parker , Sar. Ronde , Chr. Walter		47.89
4x400	Eli. Steele , Meg. McMullin , Ash. Forte , Cec. Barowski		3:43.30
HJ	Morgan, Taylor	SR-4	1.70m
PV	Harris, Allison	JR-3	4.15m
LJ	Becker, Ariel	JR-3	5.61m
TJ	Davidson, Kerri	FR-1	11.44m
SP	O'Dell, Kennedy	SO-2	13.16m
DT	Rowe, Brielle	SR-4	44.68m
HT	O'Dell, Kennedy	SO-2	55.34m
JT	Becker, Ariel	JR-3	39.05m
Hep	Morgan, Taylor	SR-4	5224

Figure 1: The Princeton (F) team page on the TFRRS website [12]. Fields that we scraped are boxed in red.

2.2. Current Ranking System

Before describing our methods for analyzing TFRRS data, we would like to explain how it is currently used to rank teams. The United States Track and Field and Cross Country Coaching Association (USTFCCCA) – the main body which produces team Track and Field rankings – tries to predict which teams might win NCAA championships [9]. They do this by considering one factor: how likely is each athlete on the team to score at the NCAA Championship in a given event? If an athlete is very likely to score at the NCAA Championship in an event, they are awarded a large number of points; if they may score, they are awarded fewer points; and if they have a very low



Cecilia Barowski

TEAM: Princeton | CLASS: SR-4 | DI, Heptagonals, Ivy League, DI East, IC4A/ECAC

Top Performances 🟩 COLLEGE BEST

YEAR	SPORT	CLASS	200M	400M	500M	600M	800M	1000M
2016	OUTDOOR	SR-4	-- --	54.63	-- --	-- --	2:02.62	-- --
2016	INDOOR	SR-4	-- --	54.39	1:11.34	1:30.33	2:04.44	2:48.56
2015	OUTDOOR	JR-3	25.04	53.96	-- --	-- --	2:05.02	-- --
2015	INDOOR	JR-3	-- --	55.24	1:13.13	-- --	2:05.00	2:54.06
2013	OUTDOOR	SO-2	-- --	54.17	-- --	-- --	2:06.26	-- --
2013	INDOOR	SO-2	25.15	55.29	1:13.35	1:30.64	-- --	-- --
2012	OUTDOOR		-- --	54.43	-- --	-- --	2:09.95	-- --
2012	INDOOR		-- --	55.43	1:12.38	1:35.15	-- --	-- --

Meet Results

DATE	MEET	SPORT	EVENT	ROUND(P/F)	TIME/MARK	CONVERTED	PLACE
04/22-04/23/16	Larry Ellis Invite	Outdoor	800	F	2:03.05		1
04/22-04/23/16	Larry Ellis Invite	Outdoor	800	F	2:06.99		1
04/15-04/16/16	Princeton Outdoor Invite	Outdoor	400	F	54.63		1
04/15-04/16/16	Princeton Outdoor Invite	Outdoor	4x4	F	3:43.30		1
04/01-04/02/16	Stanford Invitational	Outdoor	800	F	2:02.62		1
03/11-03/12/16	NCAA Division I Indoor Track & Field Championships	Indoor	800	F	2:06.81		6
03/11-03/12/16	NCAA Division I Indoor Track & Field Championships	Indoor	800	P	2:04.44		2
02/27-02/28/16	Ivy League Indoor Track & Field Championships	Indoor	1000	F	2:50.47		5
02/27-02/28/16	Ivy League Indoor Track & Field Championships	Indoor	1000	P	2:51.84		2
02/27-02/28/16	Ivy League Indoor Track & Field Championships	Indoor	500	F	1:11.34		1
02/27-02/28/16	Ivy League Indoor Track & Field Championships	Indoor	500	P	1:13.80		2
02/27-02/28/16	Ivy League Indoor Track & Field Championships	Indoor	4x4	F	3:48.89		4

Figure 2: A Princeton (F) athlete's page on the TFRRS website [2]. Fields that we scraped are boxed in red.

chance of scoring, they are awarded no points. If $score(a_{i,j,k})$ is the score of the i^{th} athlete in event j on team k, then team k's overall score is $\sum_i \sum_j score(a_{i,j,k})$.

The question then becomes, what is $score(a_{i,j,k})$? There are two factors that USTFCCCA believes are important in determining this: (1) "place points" and (2) "bonus points" [9]. Place points are awarded to athletes based on their national ranking in an event. Because only the top 24 participate in NCAA championships, USTFCCCA only awards place points to the top 50 athletes in each event. Further, because it gets exponentially harder to improve the better you are, the place points are awarded "at a regressive exponential rate" [9] with 1st place receiving 30 points and 50th place receiving 0.01 points. Bonus points, on the other hand, describe the idea that you are likely to score more than your national rank suggests if you are very close to the top performers in the field. You are also more assured to score the number of points your national rank suggests if there is a large gap in the field behind you [9]. Place points and bonus points together determine an athlete's points in an event. The cumulative score of athlete points on a team determines a team's ranking on the USTFCCCA list [9].

While the USTFCCCA rankings are good at predicting a team's performance at an NCAA

Championship, they offer limited information on a team's athletic performance. USTFCCCA ranking do not indicate how much a team's athletes have improve. A school with a high ranking clearly has very fast runners; however, it is not clear whether they were already excellent when they arrived or if they have improved while training there. In addition, while it is clear that the team with a high ranking has some excellent runners, it is impossible to know how deep the team is. A school with a few outstanding athletes would receive a high ranking regardless of the performance of everyone else. Finally, the USTFCCCA ranking provide no information on how many athletes are on the team, how often the team competes, or how the team places during week-to-week competitions, which are less competitive than the NCAA Championship. These limitation mean that the USTFCCCA rankings – and the way TFRRS data is currently processed – is not as helpful to high school athletes as it could be.

3. Approach

As it stands, there is a lot of data available on Track and Field results, but the way it is currently presented does not allow high schoolers to evaluate the athletic performance of teams. Using the current ranking system, high school athletes can understand which teams have (at least some) good athletes, but it is difficult to tell how much the athletes have improved over 4 years, how deep the team is or how successful they are within their league. Furthermore, because only the top teams in the nation are ranked, the ranking system is not at all helpful to an athlete looking at lower caliber teams.

The goal of our project, therefore, is to structure Collegiate Track and Field results in a way that is meaningful to high school recruits. The key idea behind achieving this is that there are team metrics we can extract from the TFRRS data which are more relevant and useful to high school Track and Field athletes than the ones USTFCCCA currently measures. While there is no objective truth behind which metrics are the most useful to high school athlete, in this section, we will explain the metrics we extracted and why we think they are useful to high school recruits.

3.1. Athlete Metrics

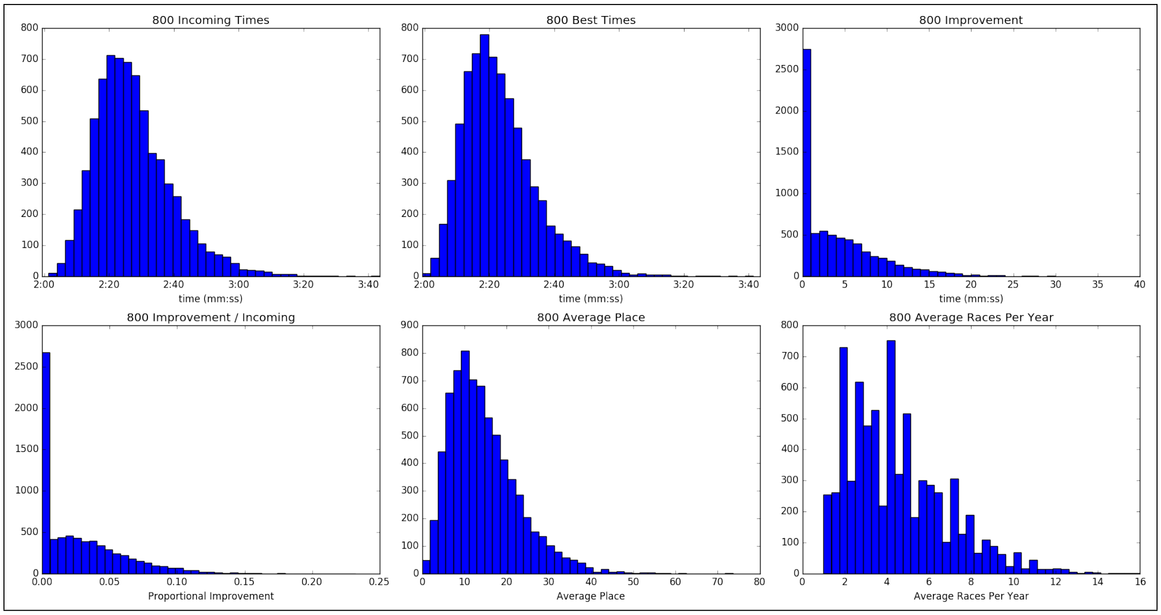


Figure 3: A histogram of athlete metrics for all female 800m runners in the TFRRS data.

In order to give high schoolers a sense of the athletic performance and experience on a team, we thought it was important to first understand the athletic experience of individuals who have competed in college. Therefore, our goal in extracting athlete metrics was to capture the athletic experience of an individual.

While there is no objective way to do this, we reflected on what variables a person might think about when they evaluate his or her college running, throwing or jumping experience. As seen in Figure 3, we settle on 6 athlete metrics which describe an individuals athletic experience through college.

Incoming mark represents an athlete's best mark in high school in an event. This metric is important because it represents how good the athlete was in high school. If we assume that the athlete was recruited for the particular event, it also suggests what calibre of mark got them recruited.

Best Mark represents an athlete's fastest mark in college in an event. In a sport where your mark defines exactly how fast you've run or far you've jumped or thrown, an athletes' best mark indicates how an athlete measure up against the competition.

Improvement ($Incoming\ Mark - Best\ Mark$) measures how much an athlete has improved in college. Since the goal of Track and Field is to run faster or throw farther, improvement represents to what extent an athlete has achieved this. In addition, we believe that if an athlete improves a lot throughout college they will reflect more positively on their athletic experience than if they remained stagnant or got worse.

That said, we also realized that it is harder to improve if you are already very good. Therefore it might be more meaningful to think about **improvement ratio** ($\frac{Improvement}{Incoming\ Mark}$) since this metric speaks more to percent improvement since high school.

Average place describes what an athlete's average place was in an event. Since winning is always fun and often boosts confidence, we felt that the average place of an athlete was an important metric in understanding his or her college running experience.

Finally, **average number of races per year** measures how many times an athlete competed in an event per year on average. It describes how often a school allowed athletes to compete and how much the individual athlete chose (or was able to) compete. The meaning of this metric is a bit hard to decipher – if an athlete didn't compete a lot is it because they were injured? Were they not good enough? Does the coach believe in racing infrequently? Still if competing is the goal, this measures how much an athlete was able to do that.

3.2. Team Metrics

We believe that team metrics can allow high school students to understand the athletic performance of different schools. Therefore, with an idea of how to measure an individual's athletic experience, we moved on to measure a team's athletic performance.

Since a team is a collection of many athletes, we noticed that every team has a distribution of athlete metrics. For example, Princeton Female (F) athletes have a distribution of incoming marks in an event, best marks in an event etc. To capture the team distributions of the 6 athlete-event metrics, we recorded their respective means and standard deviations. Therefore, while athletes only have 6 metrics per event, teams have 12 metrics per event (Figure 4). In addition we realized

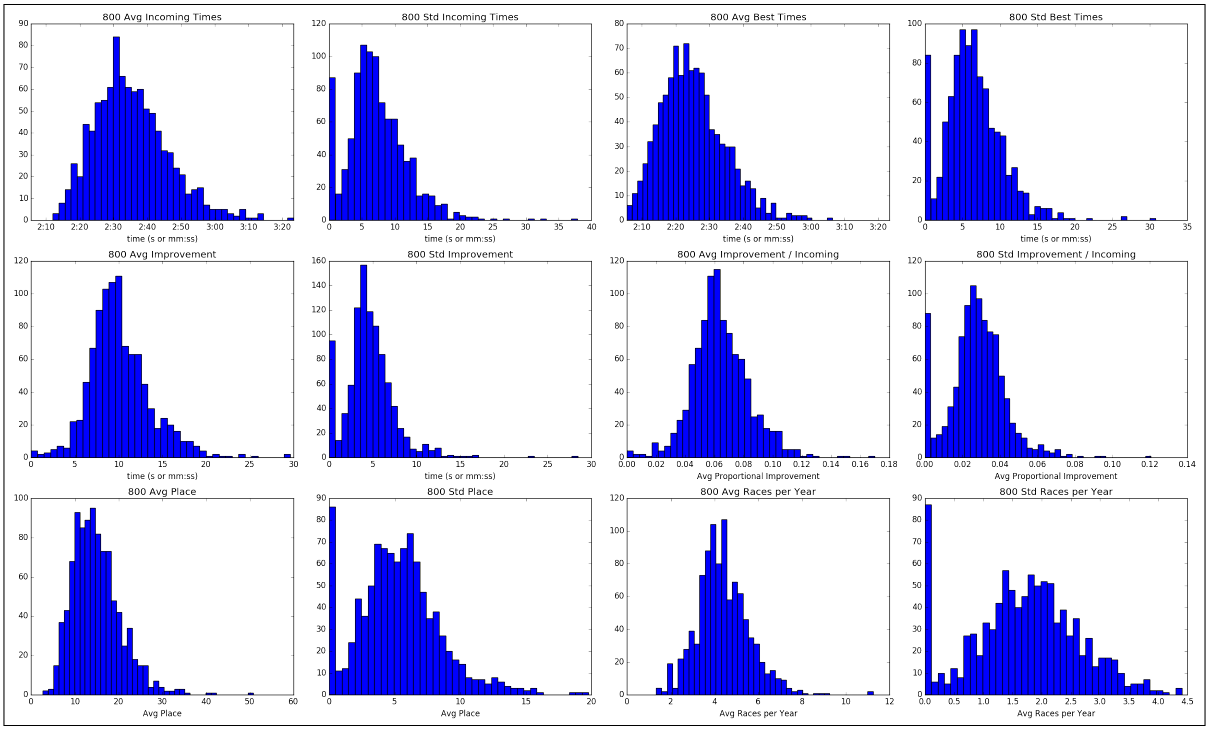


Figure 4: A histogram of team metrics for the 800 for all female teams in the TFRRS data. The record metric is not shown.

that teams have an overall best mark in every event. Therefore, we added **record** metrics to the 12 event-specific team metrics.

In addition to these event-specific metrics, we thought a number of size related metrics might be relevant to high school athletes as well. Is the team big or small? Do they have a lot of sprinters or just a few? These questions are relevant in helping high school athletes understand what their experience will be like in college. If you are a sprinter, you might want to join a team that has a high proportion of sprinters so that you have a big group to train with. To represent these characteristics, we decided to also include the following size-related team metrics:

1. **Number of Athletes** – how many athletes have competed on the team?
2. **Proportion of Sprinters** – what proportion of athletes compete in sprint events?
3. **Proportion of Middle Distance** – what proportion of athletes compete in middle distance events?
4. **Proportion of Distance** – what proportion of athletes compete in distance events?
5. **Proportion of Hurdlers** – what proportion of athletes compete in hurdle events?

6. **Proportion of Throwers** – what proportion of athletes compete in throwing events?
7. **Proportion of Jumpers** – what proportion of athletes compete in jumping events?
8. **Proportion of Multis** – what proportion of athletes compete in multi events (the heptathlon or pentathlon)?

In summary, we extracted 13 metrics per event across the 19 events in our dataset. In addition, we extracted 8 size-related metrics. In total, we extracted 255 metrics per team.

4. Implementation

In this section, we will explain the steps we took to complete our project. We will start with an explanation of our scraping method. We will then move on to explain how we cleaned the data and extracted athlete and team metrics. We will describe the methods we used to perform a general exploration of Track and Field teams. Finally, we will discuss the implementation of our recommendation system.

4.1. TFRRS Scraping

We will first explain the TFRRS Scraping process at a high level, and then discuss specific problems we ran into and how we solved them.

4.1.1. A high level view of the scraping process. On TFRRS, each Track and Field team has a unique team page that is archived back to 2009; however, in order to scrape these pages, we first had to identify their URLs. Therefore, we built a "team links" scraper, which used TFRRS built in search feature to get a list of each team with their corresponding team page link on TFRRS.

With (team, link) tuples, we were able to scrape the TFRRS website. Our scraper travels to each team page and extracts the team's name, it's division and a list of athlete links from the team's roster from 2009 - present 1. It then goes to each athlete page and gathers a list the athlete's meet results (Figure 2). The extracted information is stored in a JSON tree (Figure 5), where each team is the parent node of their athletes and each athlete is the parent node of a number of lists which contain the meet results from each event they have competed in. In addition to the "teams" branch of the tree, we also constructed "events" and "meets" branches. The "events" branch consists of lists of

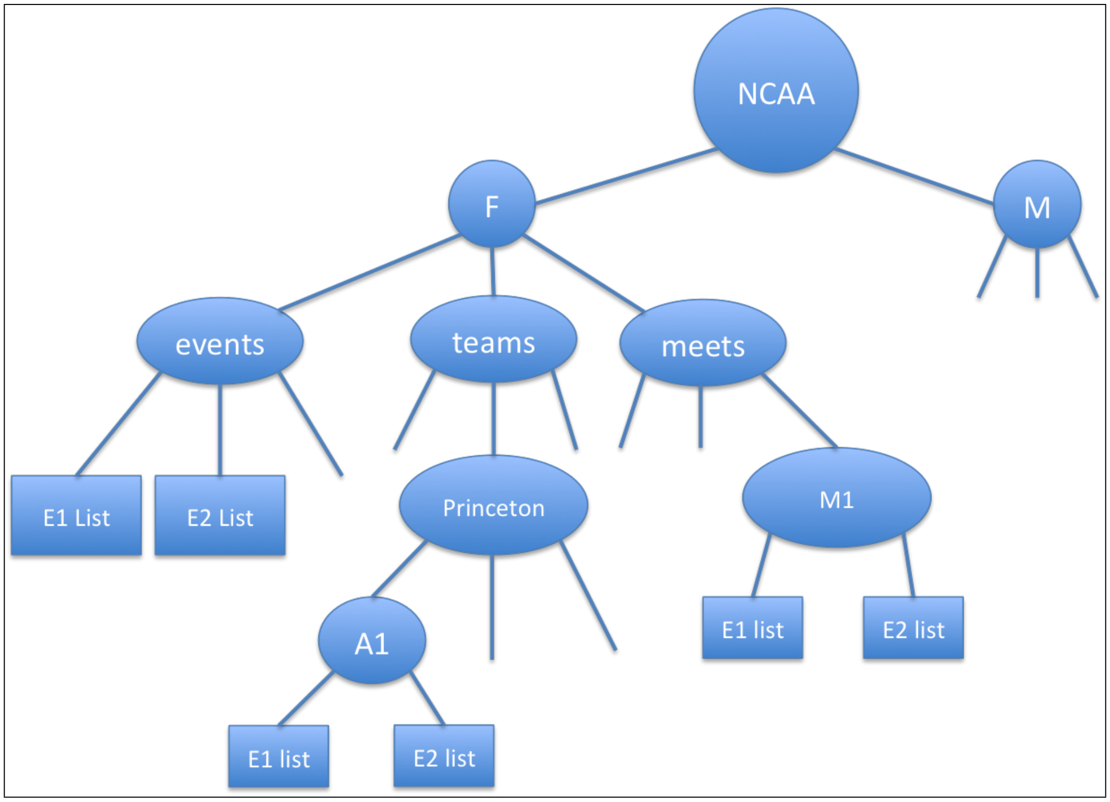


Figure 5: A diagram of the tree structure we build while scraping. E1 and E2 List represent a list of marks in a particular event. A1 represents an athlete.

athletes’ best marks in different events (an TFRRS leaderboard of sorts). The "meets" branch has a node for every meet recorded on TFRRS with the corresponding results. However, for the purposes of our project, we did not use these two additional branches.

4.1.2. Specific scraping problems and their solutions. While scraping, we ran into two problems: (1) Which athletes do we take from the team’s roster? (2) How do we standardize units for meet result marks? We will discuss each problem separately.

(1) Which athletes do we take from the team’s roster? This might seem like a straight forward question – shouldn’t we scrape all of them? But then we began to wonder: If an athlete joins the team and then quits, should they be representative of a team’s performance? Should current freshman, sophomores and juniors, who are still competing and improving, be representative of a team’s ability to develop athletes? We decided that, for the purposes of our project, we should not include these athletes on a team’s roster. Therefore we only scraped athletes who competed in their senior year of college.

This decision came with a few draw backs. To begin with, many 2 year colleges have Track and Field programs. However, because their athletes graduate from the school at the end of their sophomore year, we did not scrape any athletes for these teams[2]. Additionally, we recognize that there might be valuable team information hidden in athletes who did not compete their senior year. Did they quit? Were they injured? In this way, the dropout rate of athletes on a team might be interesting to measure. However, because we only scraped athletes who competed in their senior year of college we were unable to investigate it.

(2) How do we standardize units for meet result marks? Be it the time of a race or the distance of a jump, the mark of an athlete is a numerical representation of how they performed. Because of this, we need a uniform way to represent marks in each event. However, throughout TFRRS, there are many different ways to represent a mark. For instance high jump might be recorded in meters (1.7m) or feet (5'7")[3] and a time might be formatted in second (25.04) or in minutes and seconds (2:02.62)[4]. In order to extract numeric values from the various representations, we created a parser, which recognizes the representations described above. It converts all times to seconds and all distances to meters. However, if it is unable to parse a mark, the scraper does not record that performance.

Again there are drawbacks to this method. Occasionally, there are times which contain hours. For instance, TFRRS claims an athlete ran 10:54:12.6 in a 5k[5]. While we believe that these times are unreasonable and probably errors in the data, we completely ignored all of these race results. In addition, there is often a mark of "NP" or "DNF" meaning that the athlete competed but did not receive a mark in the event (they may have dropped out of a race or fouled all of their jumps). It might be interesting to investigate the occurrence of such values; however, our parser was unable to process them and these results were not scraped.

[2] For an example, see `https://www.tfrrs.org/teams/KS_jcollege_f_Allen_County_CC.html`

[3] For an example, see `https://www.tfrrs.org/results/36998_2551308.html?athlete_hnd=4181640`

[4] For an example, see `https://www.tfrrs.org/athletes/3733029/Princeton/Cecilia_Barowski.html`

[5] For an example, see `https://www.tfrrs.org/results/xc/3930.html?athlete_hnd=3829964`

4.2. Cleaning the Data

We cleaned the data in three phases: (1) subset the data to a limited number of events; (2) remove all marks that are unreasonably good or bad; (3) remove all athletes with no marks and all teams with no athletes. We will discuss each phases briefly.

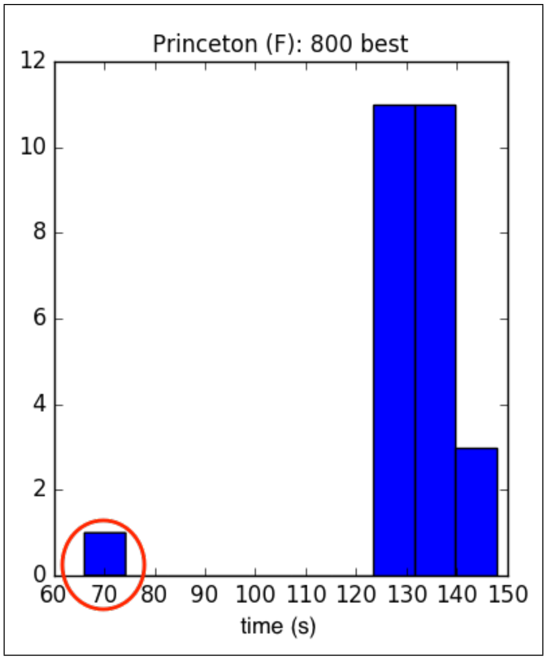


Figure 6: A histogram of Princeton (F) 800m best times before the data was cleaned. One athlete, circled in red, boasts an 800 time of 66 seconds, well below the collegiate national record.

With 229 unique events[6] represented in the TFRRS data, we realized we would need to limit the number we investigated. We wanted to include a handful of the most common events in Track and Field and settled on the 19 female and 19 male events that are part of the Olympic Games (Table 1) [1]. Therefore, we limited the data to only these events.

We then did some preliminary data analysis and realized that some of the marks were blatantly wrong. For instance, one Princeton Track and Field athlete boasts an 800m time of 1:06.07[7] (Figure 6) – well below the collegiate record of 1:59.11 [3]. (We suspect this time should be for the 400m hurdle event.) In order to avoid such obviously incorrect data, we removed all marks which were

[6] We produced this figure via the data we collected from TFRRS [14]

[7] See the 800m result from the 2008 Sam Howell Invitational at, https://www.tfrrs.org/athletes/1214169/Princeton/Christine_Brozynski.html

Male	Female
100m	100m
200m	200m
400m	400m
800m	800m
1500m	1500m
5000m	5000m
10,000m	10,000m
110m Hurdles	100m Hurdles
400m Hurdles	400m Hurdles
3000m Steeple Chase	3000m Steeple Chase
High Jump	High Jump
Long Jump	Long Jump
Triple Jump	Triple Jump
Pole Vault	Pole Vault
Shot Put	Shot Put
Discus Throw	Discus Throw
Hammer Throw	Hammer Throw
Javelin Throw	Javelin Throw
Decathlon	Heptathlon

Table 1: Track and Field events in the 2016 Olympic games. We limited the data we analyzed to these events.

better than the collegiate record in their respective events. Additionally, we created upper limits for each event (normally double or triple the collegiate record) so that we could remove marks that were unreasonably bad.

After we removed marks that were unreasonably bad or good we did some further investigation of the data and noticed that there were some athletes who had no marks and some teams that had no athletes. In the former case, we noticed that some athletes had empty pages on TFRRS[8]. We also suspect that removing marks that were unreasonably good or bad rendered some athletes markless. In the latter case, we realized that all 2 year colleges lack athletes because we only scraped those who competed in their senior year (and 2 year college have no seniors). However in many cases, teams had no athletes on their TFRRS rosters at all[9]. Regardless of the reason, we decided that we should remove all athletes with no marks and all teams with no athletes.

After performing both of these steps, our data seemed sufficiently reasonable.

[8] For example, see https://www.tfrrs.org/athletes/4102979/Smith/Kelly_Beauvais.html

[9] For an example, see https://www.tfrrs.org/teams/70740.html

4.3. Extracting Metrics

With cleaned TFRRS data, we then extracted the metrics outlined in Section 3. We first extracted athlete metrics outlined in Section 3.1 for every athlete in the TFRRS data. We then used the athlete metrics to extract the team metrics outlined in Section 3.2 for every team in the TFRRS data. While most of the athlete metrics were very easy to find using the TFRRS data, "incoming mark" was more difficult. Therefore, we will focus our discussion on extracting this metric.

The "incoming mark" of an athlete is his or her best mark in high school. However, because the TFRRS database only contains college results, this is impossible to measure directly. Therefore, we had to estimate an athlete's best mark in high school using his or her college data. Despite researching the topic, we found no previous work on how to approach this problem, so we crafted what we believe to be a reasonable solution: we define an athlete's "incoming mark" as his or her best mark in an event during their first year of competition in that event. While this might seem like a big and somewhat arbitrary estimate, we believe that because there are so many transitions freshman year, athletes do not normally improve very much over this time. We also found that in many cases, our "incoming mark" estimates were close to team's recruiting standards (see Section 5.4.1).

By extracting team metrics for all events and size metrics for all teams, we generated a feature vector representation of each team. Each team had 13 metrics (features) per event and our data contained 19 events (see Section 4.2). In addition, each team had 8 size-related features Therefore, each team was modeled as a 1 *x* 255 feature vector.

4.4. Data Exploration

With our team feature extracted, we decided to answer some interesting questions about the data. In this section, we will explain the questions we investigated and the methods we used to do so. However we will leave all results to the respective Results sections.

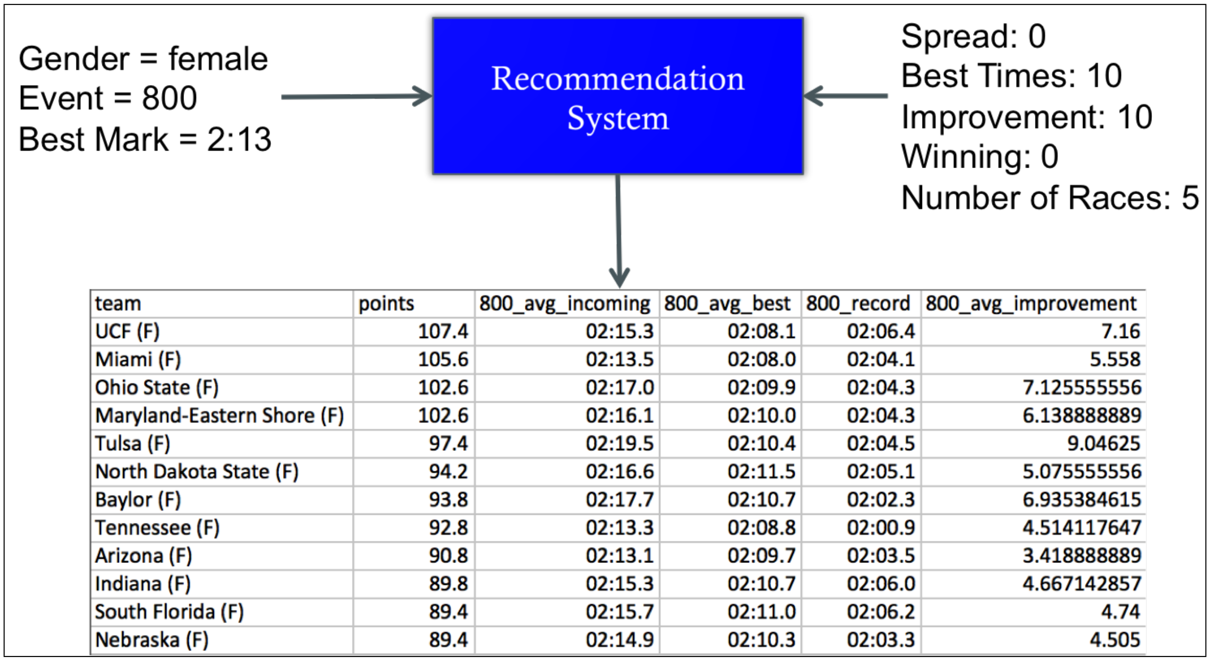
4.4.1. Clustering. Our first question was: Do the schools naturally cluster into DI, DII, DIII and non-NCAA schools? As outlined in Section 2.1, it is generally thought that DI school are better than DII are better than DIII and that good high school athletes should compete at DI schools. We were curious to see the team features we extracted displayed enough differences to naturally cluster into DI, DII and DIII.

In order to perform this clustering, we first used sci-kit learn's normalize method to normalized our data [16]. We then applied sci-kit learn's k-means clustering algorithm [16], which used euclidean distance and an iterative algorithm to place samples into K clusters in some p-dimensional feature space. We clustered teams using all 255 features. We also clustered teams using 13 event specific features – the 12 event features described in Section 3 with the corresponding size-related feature (proportion of sprinters, proportion of middle distance etc.) also described in Section 3. We explored many different values of K and visualized all results via pie charts of the proportion of DI, DII, DIII and non-NCAA teams in each cluster. (For clustering results, see Section 5.1)

4.4.2. Metric Relationships. The next question we asked was: What is the relationship between the different metrics we extracted? A high school athlete might be interested in knowing what the relationship is between a team's average incoming mark in an event and their average best mark in that event. If a high schooler is trying to decide between a team that competes a lot and one that does not, it might be interesting to understand the correlation between average races per year and best times. To answer some of these questions we generated scatter plots where each point was a team, the x axis corresponded to one feature of interest and the y axis corresponded to the other feature of interest. (For clustering results, see Section 5.2)

4.4.3. Principal Component Analysis. The final question we asked was: Which features cause the most variance in teams? Features which cause a lot a variance among teams are potentially features that high schooler might want to pay more attention to. To examine this question we implemented sci-kit learn's principal component analysis (PCA) [16]. PCA searches the p-dimensional feature space for the orthogonal dimensions of highest variance. The first principal component accounts for the most variance in the data and while the last principal component accounts for the the least. We visualized PCA results via heat maps of the principal component vectors. (For PCA results, see Section 5.3)

4.5. Recommendation System



team	points	800_avg_incoming	800_avg_best	800_record	800_avg_improvement
UCF (F)	107.4	02:15.3	02:08.1	02:06.4	7.16
Miami (F)	105.6	02:13.5	02:08.0	02:04.1	5.558
Ohio State (F)	102.6	02:17.0	02:09.9	02:04.3	7.125555556
Maryland-Eastern Shore (F)	102.6	02:16.1	02:10.0	02:04.3	6.138888889
Tulsa (F)	97.4	02:19.5	02:10.4	02:04.5	9.04625
North Dakota State (F)	94.2	02:16.6	02:11.5	02:05.1	5.075555556
Baylor (F)	93.8	02:17.7	02:10.7	02:02.3	6.935384615
Tennessee (F)	92.8	02:13.3	02:08.8	02:00.9	4.514117647
Arizona (F)	90.8	02:13.1	02:09.7	02:03.5	3.418888889
Indiana (F)	89.8	02:15.3	02:10.7	02:06.0	4.667142857
South Florida (F)	89.4	02:15.7	02:11.0	02:06.2	4.74
Nebraska (F)	89.4	02:14.9	02:10.3	02:03.3	4.505

Figure 7: A high level diagram of our recruiting recommendation system.

Having performed a thorough investigation of the data, we created a recommendation system that outputs a list of teams which might be "good fits" given a high schooler's best event and best mark in that event. We divided our recommendation system into two steps: (1) Given a student athlete's best marks in an event, which teams might recruit them? (2) Given a student athlete's preferences to a handful of questions, can we assign a "fit" score to each team? Our recommendation system uses step (1) to create a "recruitment" list consisting of all teams the high school athlete might get recruited to. It then uses step (2) to sort this list based on the "fit" score. The recommendation system produces a cvs file which contains the list of schools the high school athlete might get recruited to sorted by their "fit" score (Figure 7). We will spend the remainder of the section discussing how we implemented each of the two steps.

4.5.1. Determining teams that will recruit an athlete. Though recruitment standards exist for some schools, the standards are neither centralized nor ubiquitous. Therefore, we had to estimate which schools might recruit an athlete. To do this, we we made several assumptions. We assume that teams are always trying to get better and therefore will not recruit you unless you are as good as or better than their average incoming mark in an event. For example, if you run a 2:12.8 in the 800m and Princeton's average incoming 800m mark is 2:13.9[10], then you can get recruited there. However if you run a 2:14.8, we assume you cannot be recruited to Princeton.

We also decided that high school athlete will not want to be "too good" for a team. For example, if you run a 2:12.8 in the 800m and Paine College's average incoming 800m mark is 2:52.2[11], you can certainly get recruited there, but the team is probably too slow for you. In the same way that Google co-founder Larry Page is overqualified to work Princeton's OIT, an athlete can be overqualified to run at a school. While it is difficult to measure how slow a team must be in order for an athlete to be "too good" to run there, we decided to use a reasonable approximation: if you are more than two standard deviations better than a team's average incoming mark, you are too good for the team. In other words, you are "too good" for a team in the 800m if:

$$(your\, 800\, time) < (team's\, average\, 800\, incoming\, time) - 2 * (team's\, stdev\, of\, incoming\, 800\, times)$$

Using both of these ideas together, we claim that you can get recruited to a school if your mark in an event is better than their average incoming mark in that event and your mark is worse than two standard deviations away from their average incoming mark in that event.

4.5.2. Assigning a "fit" score to each team. If we assume that "fit" is entirely based on how much a team improves its athletes, we can calculate "fit" score as follows:

1. Sort the teams by average improvement proportions
2. $Fit\_Score(team) = N - index\_of(team)$, where N is the number of teams being considered.

However, after some contemplation, we decided that the idea of "fit" is very individual and it limits our recommendation system to assume that a team's improvement proportion is the only important metric. Therefore, to try and personalize our "fit" algorithm, we incorporated a number of

[10] We produced this figure via the data we collected from TFRRS [14]. For the Princeton TFRRS page see https://www.tfrrs.org/teams/NJ_college_f_Princeton.html

[11] We produced this figure via the data we collected from TFRRS [14]. For the Paine TFRRS page see https://www.tfrrs.org/teams/GA_college_f_Paine.html

questions to gauge which variables were most important to each athlete. We then used these values to weight the "fit" score accordingly.

The questions our system asks of high school users are:

1. On a scale of -10 to 10, how much do you care about the **spread** of the team? (0 = don't care; 10 = REALLY want spread to be small; -10 = REALLY want spread to be big)
2. On a scale of -10 to 10, how much do you care about the **best marks** on the team? (0 = don't care; 10 = REALLY want best marks to be good; -10 = REALLY want best marks to be bad)
3. On a scale of -10 to 10, how much do you care about how much the team **improves** its athletes? (0 = don't care; 10 = REALLY want to improve a lot; -10 = REALLY don't want to improve)
4. On a scale of -10 to 10, how much do you care about **winning**? (0 = don't care; 10 = want to be on a team that does really well compared to their competition; -10 = want to be on a team that does not do well compared to their competition)
5. On a scale of -10 to 10, how much do you care about how much you **compete per year**? (0 = don't care; 10 = want to compete a lot; -10 don't want to compete a lot)

Each of these questions speak to a specific team metric we extracted. Question 1 asks how much you care about a team's standard deviation of incoming times and best times. Question 2 asks how much you care about a team's average best time. Question 3 asks how much you care about a team's average improvement proportion. Question 4 asks how much you care about a team's average place. Question 5 asks how much you care about a team's average number of races per year. We used this information to create a weighted fit score as follows:

```
total_weight = $\sum_q value_q$;
fit_score of all teams = 0;
for all questions whose value is not 0 do
    sort the teams by the appropriate metric ;
    fit_score$_{team} += \frac{value_q}{total\_weight}(N - index\_of(team))$;
end
```

Using this method, our recommendation system attempts to personalize "fit" scores so that many of the team metrics we gathered are incorporated into our analysis of the teams. By combining our estimated teams that will recruit an athlete and the fit scores for those teams, our system attempts to provide high school athletes with a personalized list of teams they should consider looking into.

5. Results

In this section, we will first present cluster results. We will then move on to talk about the relationships among some of the team metrics we extracted. Finally we will give an interpretation of our PCA analysis and discuss our recommendation system.

5.1. Clustering

We found that if we represent the schools as 255 dimensional feature vectors and cluster with $K = 3$ clusters, we can identify a DI and DIII cluster for both men's and women's team. As seen in Figure 8, female cluster 0 is majority DI and female cluster 2 is majority DIII. Additionally, the remaining cluster in both the male and female data appears to have more representation from DII schools and schools which are not part of the NCAA (non-NCAA). (We clustered with $K = 4$ but it did not separate out a "DII" cluster or "non-NCAA" cluster.)

In an attempt to understand what distinguished one cluster from another, we also examined the centroid of each cluster. It is clear from Figure 8 that the centroids of clusters 0 and 2 are opposite of each other. Cluster 0's centroid has lower (better) marks in running events towards the left of the centroid graph and higher (better) marks in field events towards the right. Cluster 2's centroid has

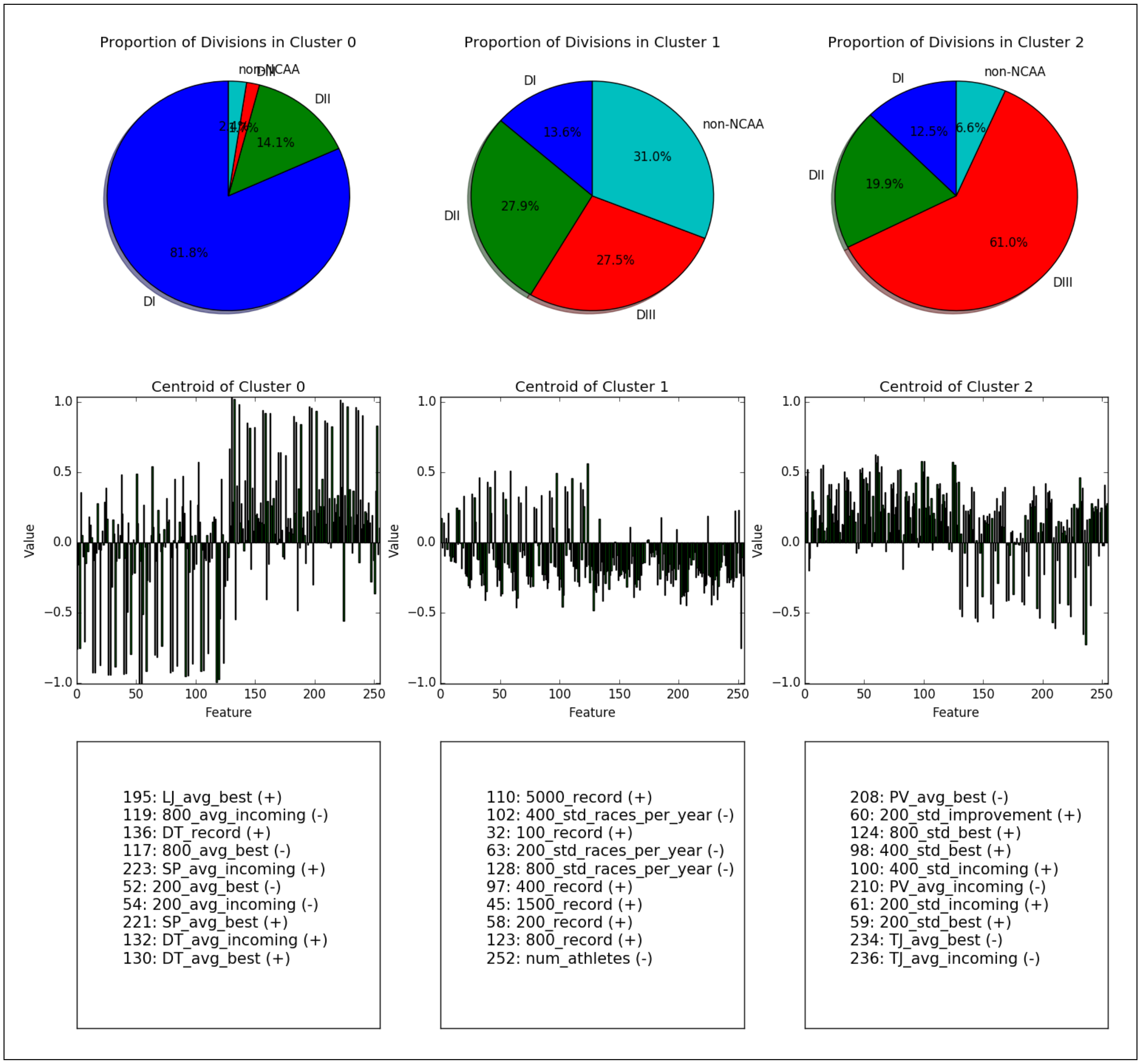


Figure 8: The clustering of female teams when teams are described as 255-dimensional vectors and $K = 3$. The top row of graphs displays a pie chart of the proportion of divisions in each cluster. The next row displays the centroids of the respective clusters. The final row displays the 10 features of each centroid that changed the most. As a point of reference, 32.1% of female teams in our data are DI, 21.9% are DII, 29.5% are DIII and 16.5% are non-NCAA.

higher (worse) marks in running events towards the left and lower (worse) marks in field events towards the right. In an attempt to be more specific about this observation, we also printed the 10 features of the centroid which deviated most from the unclustered data. For both men and women, the "DIII" centroid is most different in that it has much worse sprinting and field performances and the "DI" centroid is most different in that it has much better sprinting and field performances. For example, in Figure 8, cluster 0 has better long jump, shot put, 200m and 800m marks while cluster

2 has worse pole vault, triple jump, 200m, 400m and 800m marks. These data suggest that field and sprint performance are major factors in separating DI and DIII schools in both women's and men's teams.

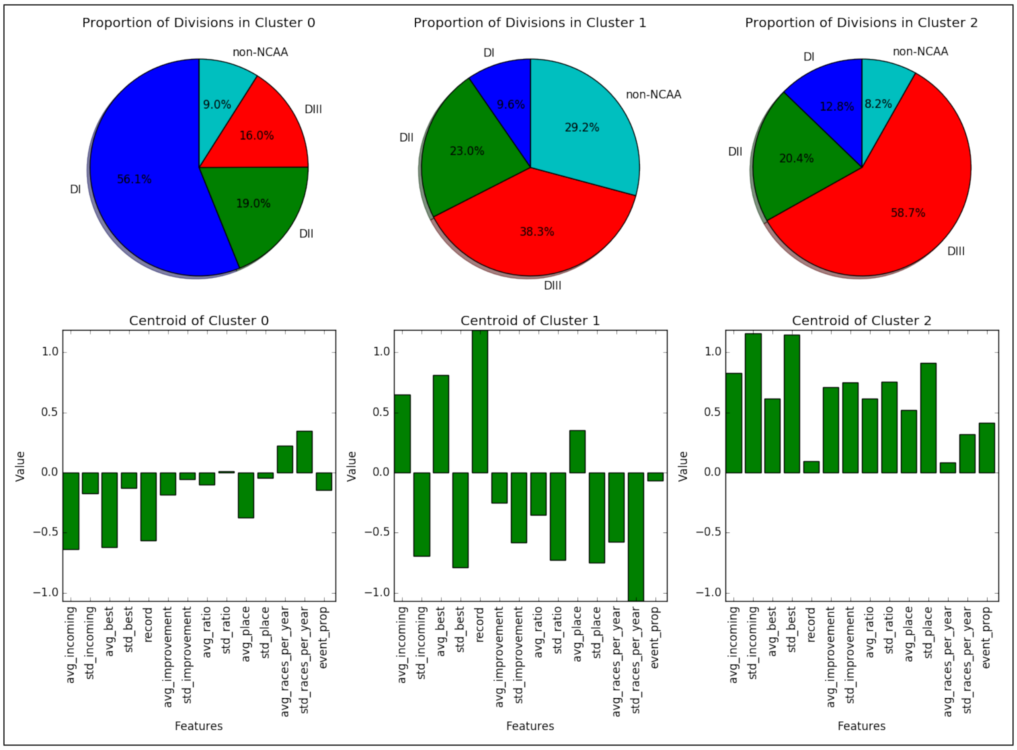


Figure 9: The clustering of female teams in the 100m when teams are described as 13-dimensional vectors and $K = 3$. The top row of graphs displays a pie chart of the proportion of divisions in each cluster. The next row displays the centroids of the respective clusters.As a point of reference, 32.1% of female teams in our data are DI, 21.9% are DII, 29.5% are DIII and 16.5% are non-NCAA.

In addition to clustering schools represented as 255 dimensional feature vectors, we looked specifically at each event and clustered schools via 13 event-specific features (the 12 event-specific features along with the proportion of athletes in that event type) (Figure 9). we were able to identify clusters that were majority DI and DIII for all 19 events we examined for both the men's and women's teams. However, we noticed that instead of being able to identify a "DI / non-NCAA" cluster, we were often left with two "DI" or "DIII" clusters (Figure 9). These data suggest that the differences between DI and DIII schools is perhaps stronger than the differences between DII and non-NCAA schools. In addition, we noticed that while average incoming, best and record metrics seemed more prominent in clustering 255-dimensional teams, standard deviation metrics seemed more prominent in clustering 13-dimensional teams.

We think it's important to recognize that with $K = 3$ clusters, there was never a cluster that was completely DI nor completely DIII. This suggests that while the teams do naturally cluster into DI and DIII schools, there are a significant number of DIII schools which fit more with the DI cluster and visa versa. We believe this suggests that while performance is linked to the division of a school, it is possible to find high performing DIII teams and low performing DI teams. Because of this, a high school student looking to run in college might want to consider certain DI or DIII schools regardless of how accomplished they are in their event.

5.2. Metric Relationships

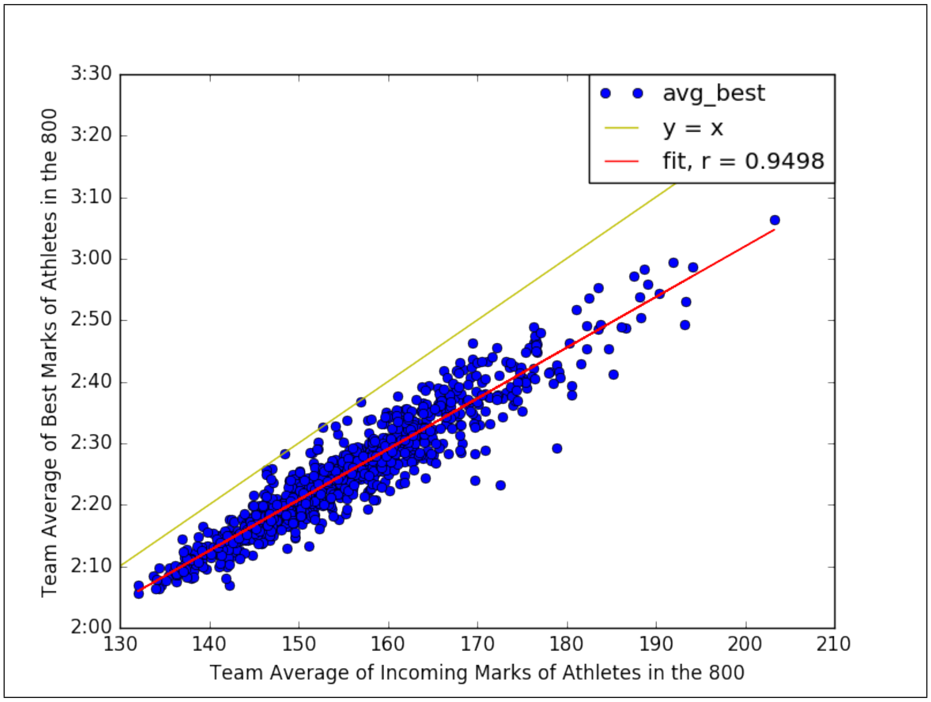


Figure 10: A scatter plot of female teams' average best time versus their average incoming time in the 800. Each point in the scatter represents a female team. Teams that lie below the line $y = x$ improve their athletes in the 800m on average. The The lower the team is from $y = x$ the more the team has improved it's athletes on average.

In order to better understand the metrics we extracted from the TFRRS data, we examined the correlation between metrics via scatter plots. In every event for both male and female athletes, we found a strong correlation between a team's average incoming time in an event and their average best time in the event (Figure 10). What is more, we found that none of the teams lie above the line $y = x$ in running events nor below the line $y = x$ in field events. This suggests that on average, teams improve their athletes in all events. That being said, by looking at how far away each point is from the line $y = x$, it is clear that even with the same incoming marks, some teams see larger improvement than others. These data suggest that high schoolers who decide to run in college will, on average, improve. However it also suggests that they might be more successful at certain schools.

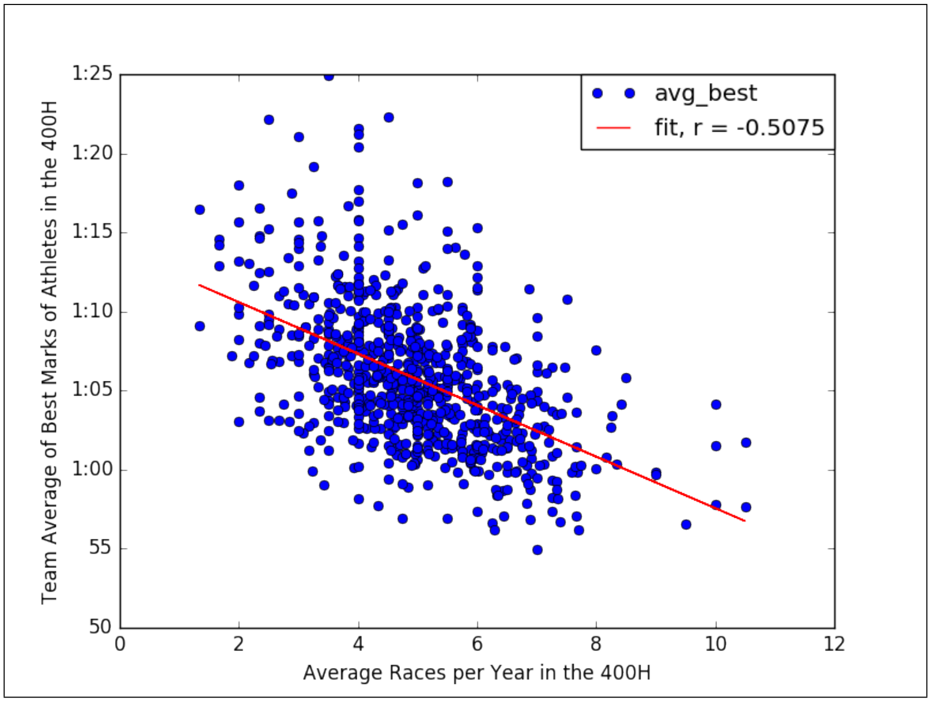


Figure 11: A scatter plot of female teams' average best time versus their average number of races per year in the 400 Hurdles. Each point in the scatter represents a female team.

We also found the relationship between a team's average number of competitions per year and their best marks particularly interesting. While field events were fairly correlated with a team's average best time in the event ($r \approx 0.54$), there was much less correlation in running events ($r \approx 0.21$). However, this general trend was not true of the 100m, 110m and 400m hurdle events, where the correlation between number of competitions per year and a team's average best time was stronger ($r \approx 0.50$, Figure 11). While we are not entirely sure why this is the case, we hypothesize that field and hurdles are niche events where constant practice and competition makes a bigger difference. This information might be interesting to high school athletes who compete in these areas.

5.3. Principal Component Analysis

If we represent each team as a 1 *x* 255 feature vector, our PCA results indicated the average incoming, average best and record mark in each event cause the most variance among teams. As seen in Figure 12, it takes around 125 principal components to explain around 95% of the variance in the female team data. The same is true of the male team data. This suggests that there is a lot of variance in the data – which makes sense since the feature space is relatively large. If we graph the first 12 principal components (PCs), we find that a number of features contribute significantly to PC1, suggesting that there are certain features which cause the most variance among teams (Figure 13). If we zoom in on these features in order to identify them, we find that the features with the high contributions to the first PC are the average incoming, average best and record mark in each event (Figure 14). These results are congruent with the clustering results, which also suggested that, in the 255 dimensional feature space, clusters are different by their performance (average incoming and average best marks).

If we instead subset by event and represent each team as a 1 *x* 13 feature vector, our PCA results indicate that features which incorporate standard deviation (for example standard deviation of incoming marks) often cause the most variance among both male and female teams, across all events. If we graph the first 4 PCs for female teams in the 800, we find that features which incorporate standard deviation contribute heavily to PC1. Additionally, the average incoming, average best and record marks in the 800 contribute heavily to PC2 (Figure 15). The same trends were seen across most events for both male and female teams. These results are congruent with the clustering results, which suggested that, in the 13 dimensional feature space, standard deviation features become more important in differentiating clusters.

Finally, since our recommendation system ultimately assigns "fit" scores to a subset of schools that a given high school athlete might get recruited to, we decided to perform PCA analysis on this

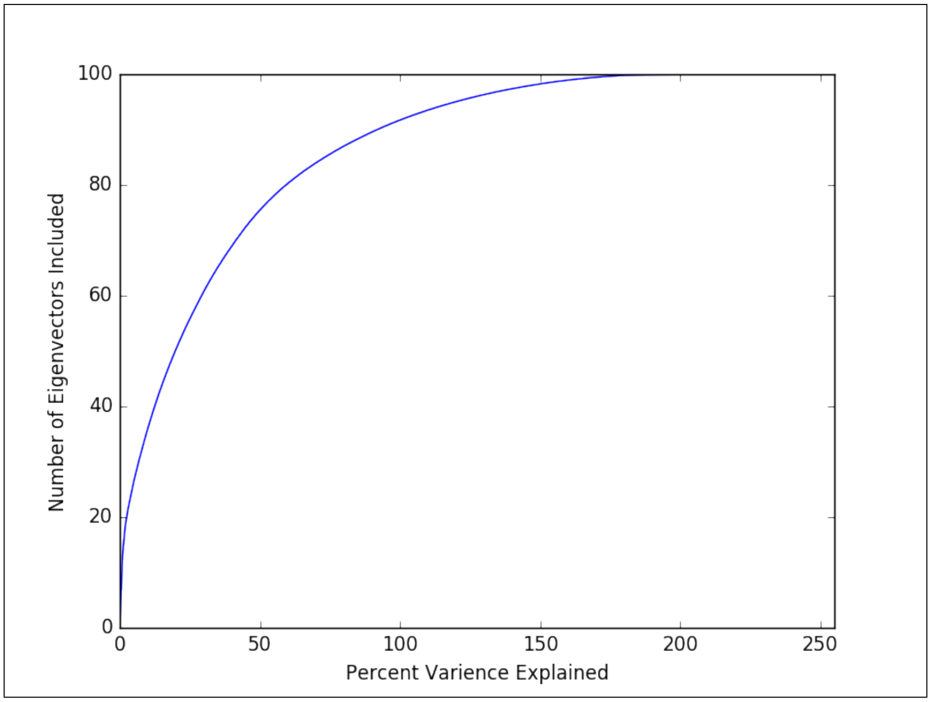


Figure 12: The explained variance of the principal components in 255-dimensional feature space for female teams. It takes approximately 125 PCs to explain around 95% of team variance.

smaller subset. Compared with PCs for female teams in the 800 shown above (Figure 15), PCs for female teams that (according to our system) would recruit a 2:10 in the 800 are more differentiated on their average and standard deviations of improvement and improvement proportion (Figure 16). While we did not examine every possible mark / event combination, other events generally displayed a similar trend. These data suggest that if we look at a subset of schools that might recruit a high school athlete in his / her best event, the features which cause the most variance among this subset are the improvement metrics in the event. Given this, it might be informative for the high schooler to know which teams yield better improvement metrics.

5.4. Recommendation System

As described in Section 4.5, our recommendation system was implemented in 2 phases: (1) which teams might recruit an athlete? (2) assign a "fit" score to each of those teams. In this section, we will evaluate each phase separately.

5.4.1. Evaluating Recruitment Estimates. As described in Section 4.5.1, our recommendation system first finds a subset of schools that a high school athlete might get recruited to. In order to do

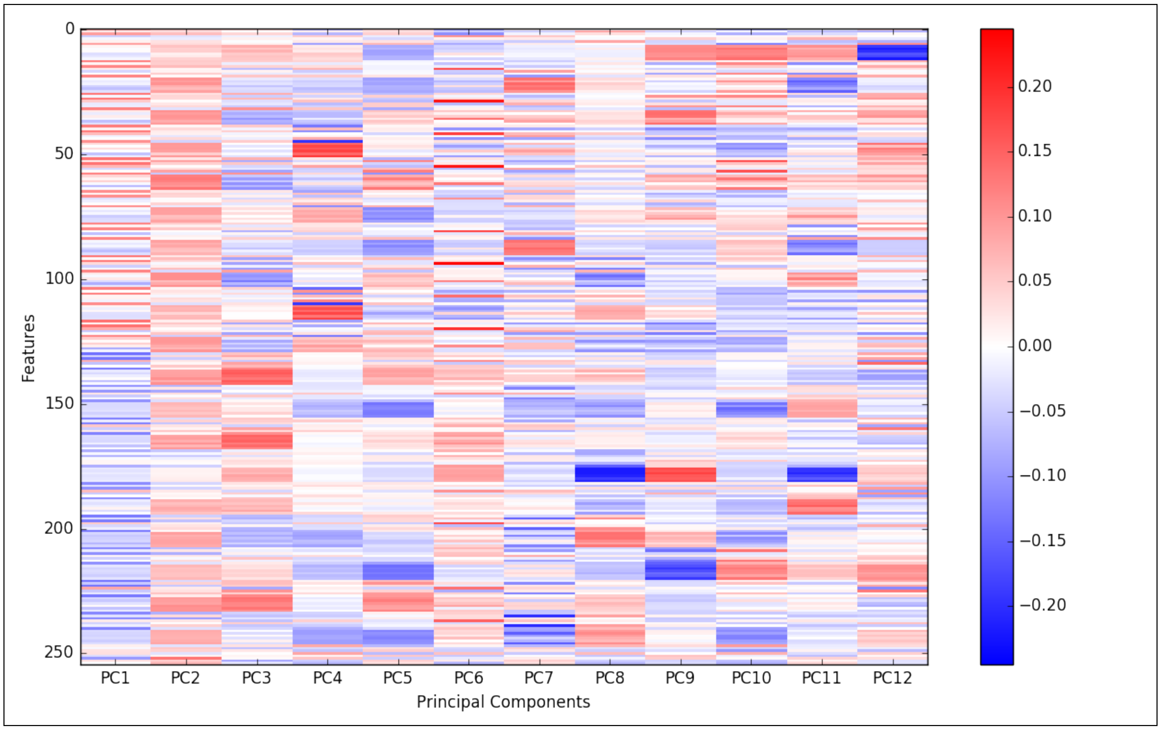


Figure 13: Principal components for female teams in the 255-dimensional feature space. Each color band represents the importance of that feature in the PC. There are a number of features (the darker bands) which contribute more to PC1. These bands are the incoming, best and record marks for all events (Figure 14).

this it assumes two things: (1) in order for a team to recruit an athlete in an event the athlete must be faster than the team's average incoming mark in the event. (2) a team is too slow for an athlete if the athlete is better than one standard deviation better than the team's incoming marks in the event. While the second assumption is very difficult to validate, we were able evaluate the validity of our first assumption using known recruiting standards within the ivy league. As seen in Table 2, our method is very accurate in some events while there is room for improvement in others. For example, our predictions for the 200m, High Jump, Triple Jump and Shot Put are strongly correlated with the recruiting standards ($r > 0.86$); however our predictions for the 1500m are not very accurate ($r = 0.04$). While we would certainly benefit from including more schools in our validation (our validation was done with $n = 9$), this suggests our method for predicting recruiting standards should be improved.

That said, we noticed that there were a few outliers and decided to investigate. As seen in Figure 17, there are a few teams in each event who have significantly more or less percent error than the

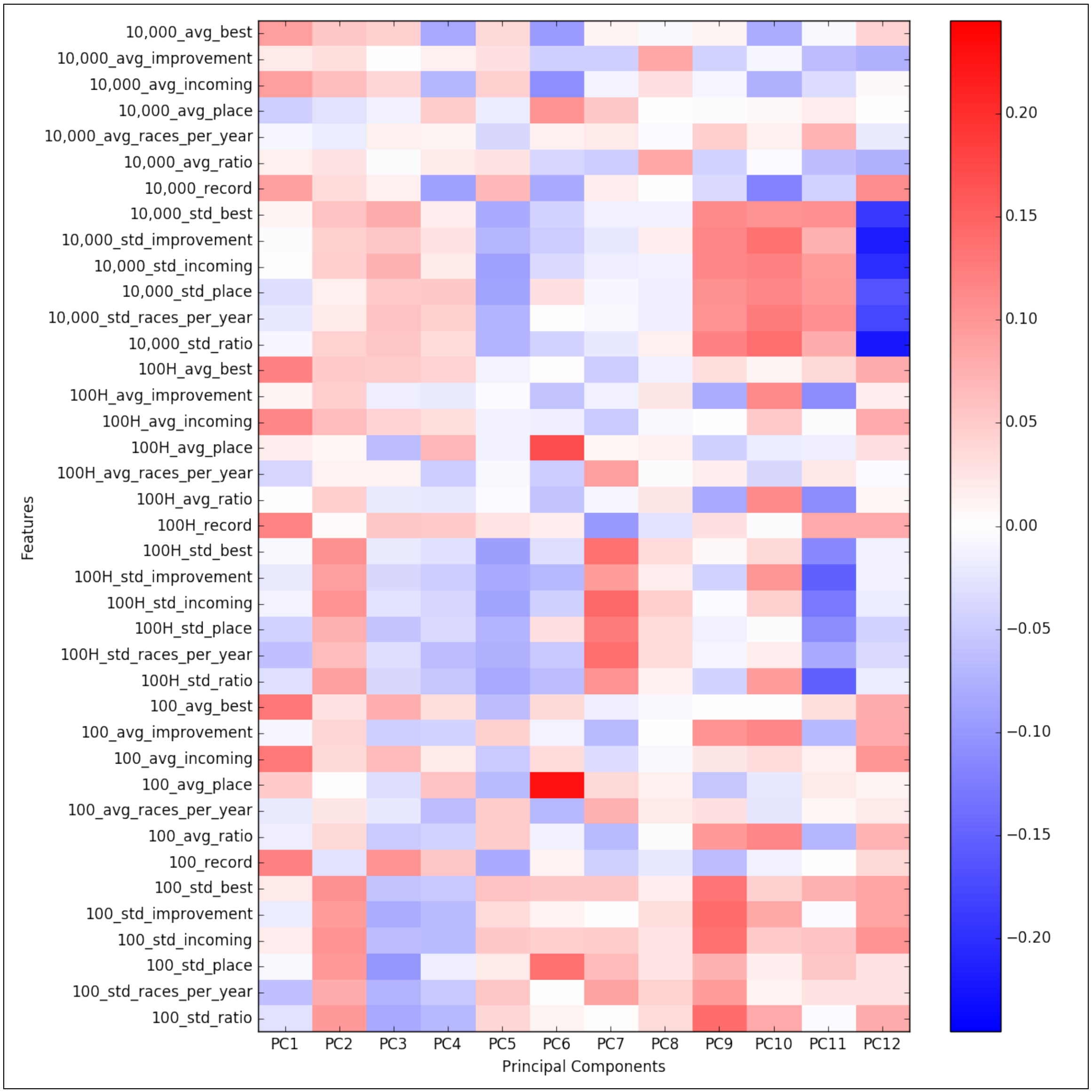


Figure 14: Zoomed in vie of principal components for female teams in the 255-dimensional feature space. From this picture, we can see that the dark bands in Figure 13 are the incoming, best and record marks for every event.

rest. We found that the outlier in the 100m is Penn (F) who's recruiting standard is 12.34 while their average incoming time is 13.12. The outlier in the 400m is Harvard (F) who's recruiting standard is 56 seconds while their average incoming time is 57 seconds. The outlier in the Long Jump is Harvard (F) who's recruiting standard is 5.5m while their average incoming mark is 4.6m. The upper outlier in the Pole Vault is Harvard (F) who's recruiting standard is 3.66m while their average incoming mark is 3.19m; the lower outlier is Columbia (F) who's recruiting standard is 3.51m while their average incoming mark is 3.57m. First, it is interesting that all of our outliers are female teams.

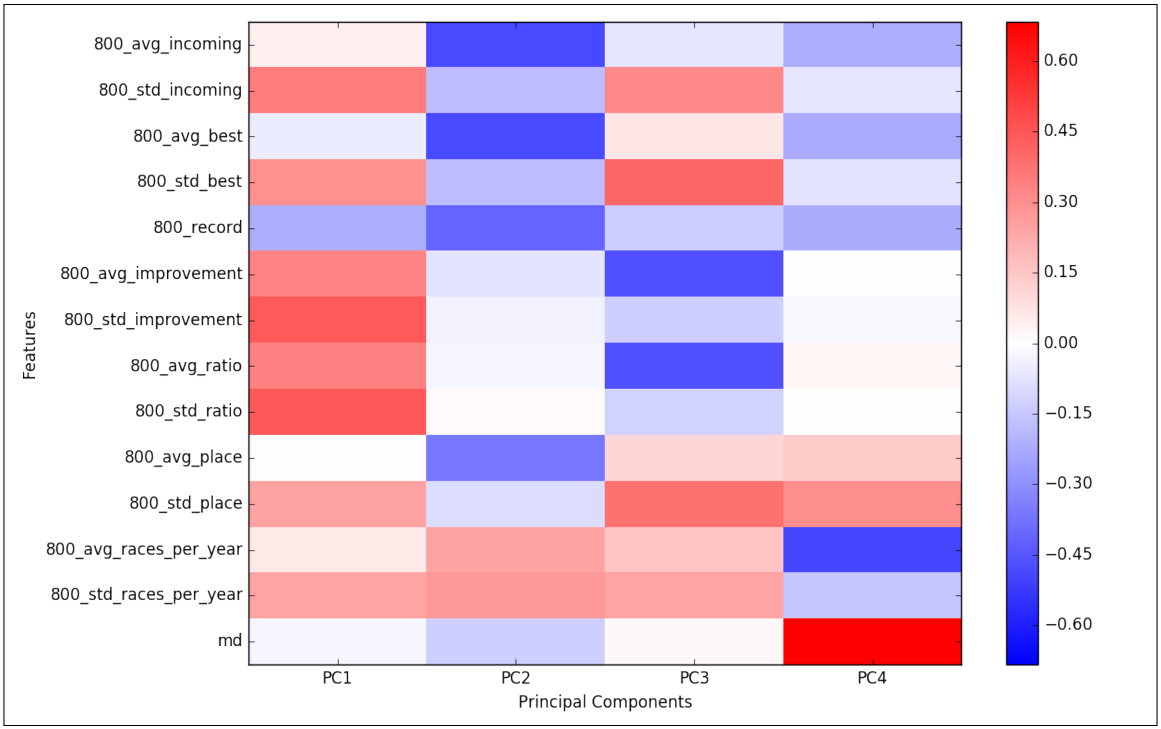


Figure 15: Principal components for female teams in the 800m 13-dimensional feature space. Each color band represents the importance of that feature in the PC. Standard deviation features appear to contribute more to PC1 while incoming, best and record marks appear to contribute more to PC2.

Second, it is interesting that Harvard's team consistently comes up. We investigated further and realized that while most recruiting standards across the ivy league are similar, the Harvard Women's Team expectations are significantly higher. For instance, they have the fastest 400m standards by a second (all other teams require 57 seconds); they have the fastest 800m standards by 4 seconds (they require 2:09 while the next fastest recruiting time is 2:13); they have the fastest 1500m standards by 10 seconds (they require 4:30 while the next fastest recruiting time is 4:40); they have the farthest hammer throw standards by 5m (they require 50m while the next best recruiting mark is 45m). Recruiting decisions are made by the coach – so we have no real authority to say they are exaggerating their standards. However, given that I was recruited to Harvard with a 2:13 in the 800m and a 4:41 in the 1500m, we suspect that these times are not necessarily accurate.

We don't want to belabor the point too much or appear to be bias against Harvard's team, but we believe that this suggests our method for evaluating recruitment estimates is not particularly accurate. If we do not believe the published standards, should we use them as our ground truth? We believe that a better approach would involve searching for athletes' best times in high school and

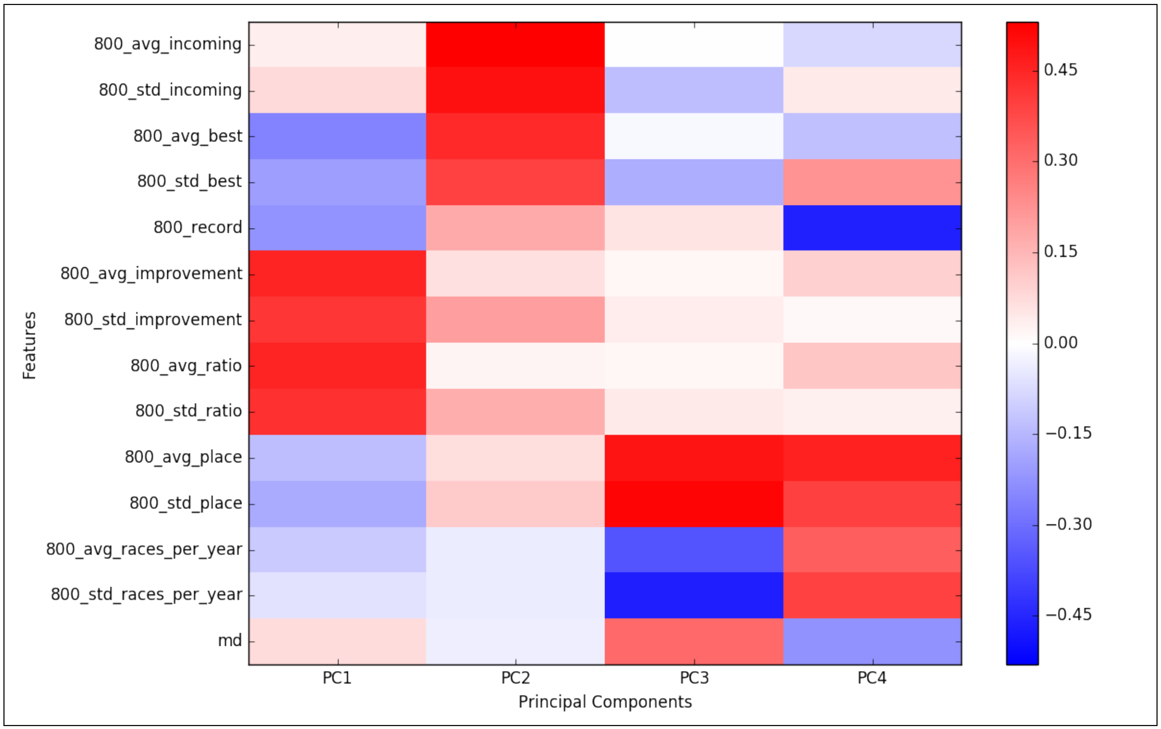


Figure 16: Principal components for female teams who would recruit a 2:10 in the 800m 13-dimensional feature space. Incoming times seem to be much less important to PC1 while the importance of improvement and the improvement proportion have increased.

estimating the "ground truth" recruiting standard as the average of those time.

5.4.2. Evaluating "Fit" Scores. With a subset of teams an athlete might get recruited to, our recommendation then assigns a "fit" score to each school based on 5 questions outlined in Section 4.5.2. As seen in Figures 18 and 19, if we keep the gender, best event and best mark inputs constant but change the answers to the questions, the "fit" score of schools change enough to reorder the teams. If we say that (spread preference, best mark preference, improvement preference, winning preference, competitions per year preference) = (5, 5, 10, 10, 5), Georgetown (F) is clearly the highest ranking school (Figure 18). However if we instead set our preferences to (-10, 5, 10, 10, 5), LSU (F) has the most "fit" points (Figure 19).

There is no ground truth for the accuracy of "fit" points. (In fact, this is why we don't eliminate any schools based on "fit" points.) Because of this, we were unable to rigorously evaluate our system. However we can provide quantitative results based on this example. An athlete might choose the preferences which yielded Figure 18's results if they are looking for a very stacked team that had good improvements and performs well within their league. In this case, Georgetown

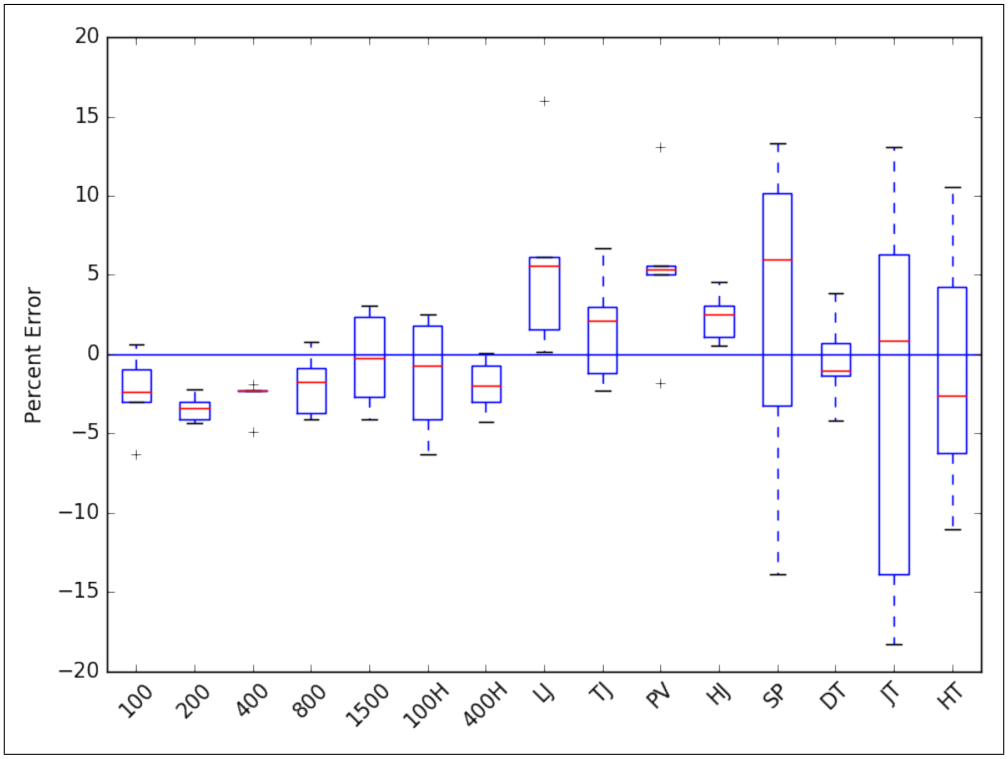


Figure 17: Box plots of the percent error between a team's recruiting standard and our estimated standard in different events. $error = \frac{recruiting\ standard - estimated\ standard}{recruiting\ standard} * 100$**. The outlier in the 100m is Penn (F), The outlier in the 400m is Harvard (F), The outlier in the Long Jump is Harvard (F), The upper outlier in the Pole Vault is Harvard (F) and the lower outlier is Columbia (F)**

(F) is ranked first because it has low standard deviation among incoming and best 800m times. Additionally it displays good proportional improvement in the 800m (3.3%) and has a low average place of around 7th. While Georgetown (F) is not the best in many of these metrics (for instance Miami has a 4.1% proportional improvement) the combination of qualities that Georgetown (F) displays gives it a very high "fit" score.

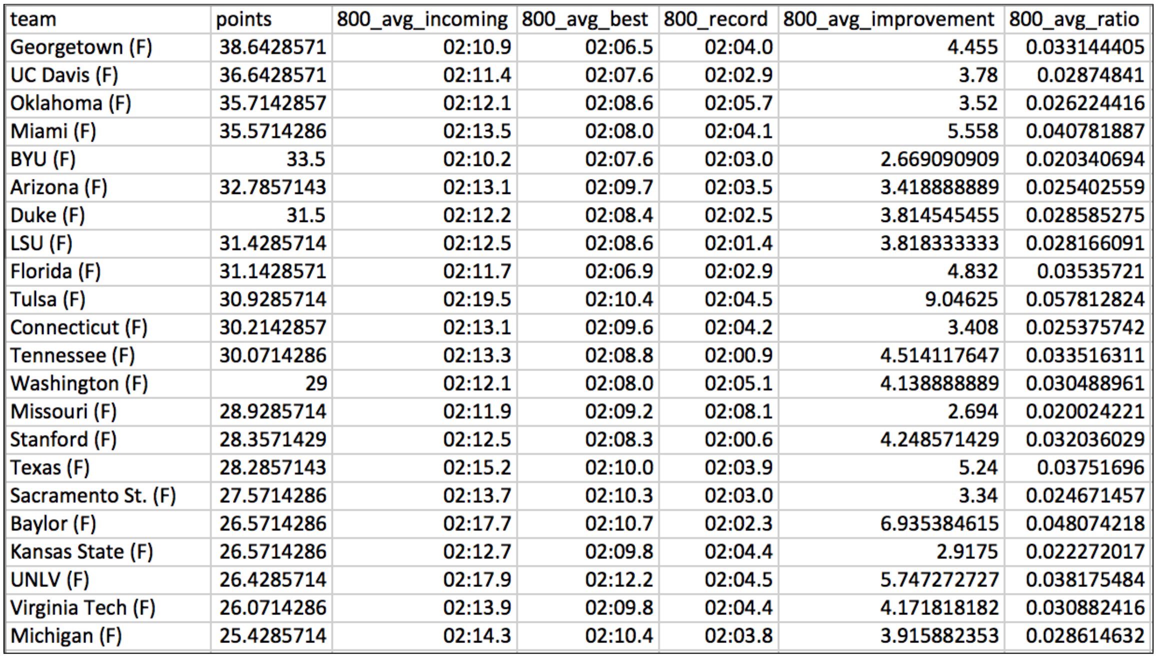
An athlete might choose the preferences which yielded Figure 19's results if they are looking for a team that has a lot of variation in talent but still manages to have good improvements and perform well within their league. In this case, LSU (F) is ranked first because despite high standard deviations among incoming and best 800m times, it still has good proportional improvement in the 800m (2.8%) and has a low average place around 7th. Again it is not one metric which gives LSU (F) this high ranking (Tulsa has a much higher standard deviation of incoming times) but rather the combination of qualities which yield this score.

Event	Correlation Coefficient (r)
100m	0.75
200m	0.94
400m	0.67
800m	0.42
1500m	0.04
110m Hurdles	0.29
400m Hurdles	0.52
High Jump	0.91
Long Jump	0.59
Triple Jump	0.86
Pole Vault	0.26
Shot Put	0.87
Discus Throw	0.62
Hammer Throw	0.57
Javelin Throw	0.62

Table 2: The correlation coefficient between the average incoming mark for teams and those team's recruiting standards published online [7][13][5][4][10][11] for various events. 9 teams in the Ivy League (5 men's teams and 4 women's teams) were evaluated.

6. Conclusion and Future Work

In summary, we have described a way to gather Track and Field data from TFRRS, extract team metrics which are useful to high school athletes, and used these metrics to create a high school recruiting recommendation system. Along the way, we also discovered interesting things about Track and Field teams. Teams naturally cluster into DI and DIII schools; however DII and non-NCAA seem harder to distinguish. On average all teams improve their athletes across all events – so you decide to compete in college, you will likely improve regardless of where you go. However even among schools with similar average incoming times in an event, we found that average best times (and therefore average improvement) still varied. We noticed a strong correlation between the average number of competitions per year and field and hurdle events. It is possible that competing often in these events is important. Finally, we found that the features which cause the most variability among schools are the average incoming, best and record marks in each event. However, if we look at schools in a particular event, the standard deviation of these metrics becomes much more important while the average plays a secondary role.

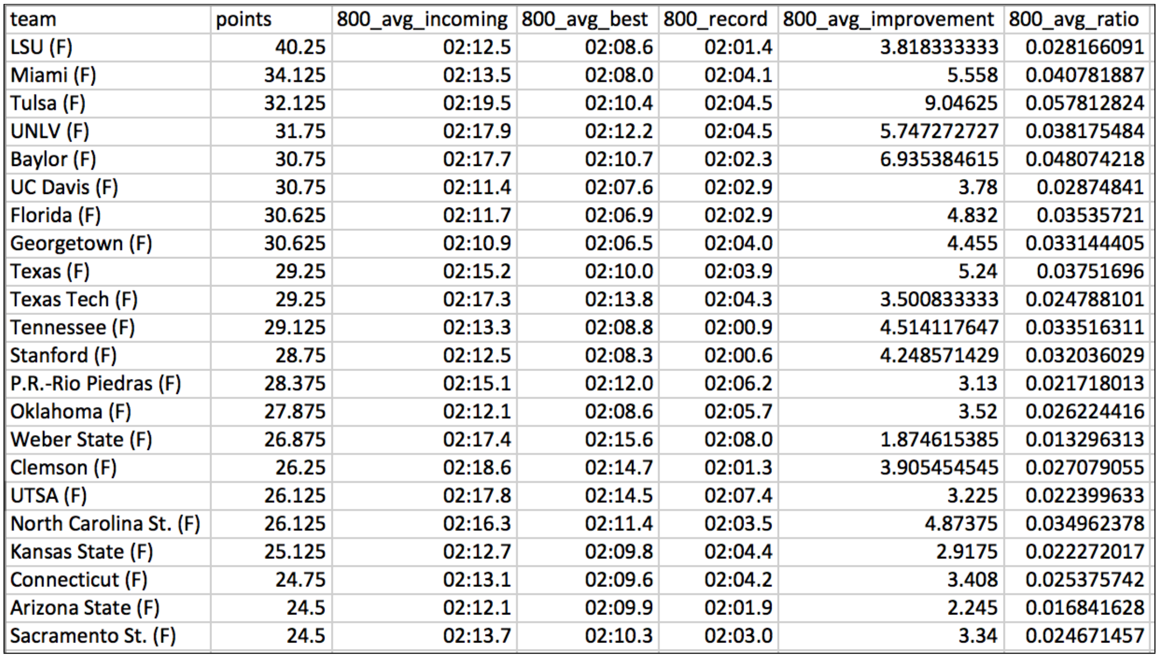


team	points	800_avg_incoming	800_avg_best	800_record	800_avg_improvement	800_avg_ratio
Georgetown (F)	38.6428571	02:10.9	02:06.5	02:04.0	4.455	0.033144405
UC Davis (F)	36.6428571	02:11.4	02:07.6	02:02.9	3.78	0.02874841
Oklahoma (F)	35.7142857	02:12.1	02:08.6	02:05.7	3.52	0.026224416
Miami (F)	35.5714286	02:13.5	02:08.0	02:04.1	5.558	0.040781887
BYU (F)	33.5	02:10.2	02:07.6	02:03.0	2.669090909	0.020340694
Arizona (F)	32.7857143	02:13.1	02:09.7	02:03.5	3.418888889	0.025402559
Duke (F)	31.5	02:12.2	02:08.4	02:02.5	3.814545455	0.028585275
LSU (F)	31.4285714	02:12.5	02:08.6	02:01.4	3.818333333	0.028166091
Florida (F)	31.1428571	02:11.7	02:06.9	02:02.9	4.832	0.03535721
Tulsa (F)	30.9285714	02:19.5	02:10.4	02:04.5	9.04625	0.057812824
Connecticut (F)	30.2142857	02:13.1	02:09.6	02:04.2	3.408	0.025375742
Tennessee (F)	30.0714286	02:13.3	02:08.8	02:00.9	4.514117647	0.033516311
Washington (F)	29	02:12.1	02:08.0	02:05.1	4.138888889	0.030488961
Missouri (F)	28.9285714	02:11.9	02:09.2	02:08.1	2.694	0.020024221
Stanford (F)	28.3571429	02:12.5	02:08.3	02:00.6	4.248571429	0.032036029
Texas (F)	28.2857143	02:15.2	02:10.0	02:03.9	5.24	0.03751696
Sacramento St. (F)	27.5714286	02:13.7	02:10.3	02:03.0	3.34	0.024671457
Baylor (F)	26.5714286	02:17.7	02:10.7	02:02.3	6.935384615	0.048074218
Kansas State (F)	26.5714286	02:12.7	02:09.8	02:04.4	2.9175	0.022272017
UNLV (F)	26.4285714	02:17.9	02:12.2	02:04.5	5.747272727	0.038175484
Virginia Tech (F)	26.0714286	02:13.9	02:09.8	02:04.4	4.171818182	0.030882416
Michigan (F)	25.4285714	02:14.3	02:10.4	02:03.8	3.915882353	0.028614632

Figure 18: The output of our recruiting recommendation system when (spread preference, best mark preference, improvement preference, winning preference, competitions per year preference) = (5, 5, 10, 10, 5)

The trends we noticed during out data exploration make us more confident that a team's success is defined by many variables – not just by their best athlete in each event. As described in Section 2.2, the USTFCCCA ranks schools based on their top athletes in every event. Under the assumption that a team's top performing athletes are key in its ranking, we would expect the "record" variable we extracted to be very important throughout our analysis. However, this was not necessarily the case. When we clustered teams and examined the 10 most different features of the centroids, a team's record in each event only came up a handful of times. While PCA analysis of teams revealed that records in each event were an important contributor to variance among schools, PCA analysis of teams in particular events showed that records were less important. Because of these observations, we are confident that our metrics describe schools in a way that is impossible using the USTFCCCA ranking system. This is not to say that our system is better, but rather is detecting numerical trends that USTFCCCA does not.

There are a number of limitations in our data and problems in our approach that should be improved upon in the future. To begin with, it is very unfortunate that we do not have high school performances in our dataset. While there is a high school Track and Field result database called



team	points	800_avg_incoming	800_avg_best	800_record	800_avg_improvement	800_avg_ratio
LSU (F)	40.25	02:12.5	02:08.6	02:01.4	3.818333333	0.028166091
Miami (F)	34.125	02:13.5	02:08.0	02:04.1	5.558	0.040781887
Tulsa (F)	32.125	02:19.5	02:10.4	02:04.5	9.04625	0.057812824
UNLV (F)	31.75	02:17.9	02:12.2	02:04.5	5.747272727	0.038175484
Baylor (F)	30.75	02:17.7	02:10.7	02:02.3	6.935384615	0.048074218
UC Davis (F)	30.75	02:11.4	02:07.6	02:02.9	3.78	0.02874841
Florida (F)	30.625	02:11.7	02:06.9	02:02.9	4.832	0.03535721
Georgetown (F)	30.625	02:10.9	02:06.5	02:04.0	4.455	0.033144405
Texas (F)	29.25	02:15.2	02:10.0	02:03.9	5.24	0.03751696
Texas Tech (F)	29.25	02:17.3	02:13.8	02:04.3	3.500833333	0.024788101
Tennessee (F)	29.125	02:13.3	02:08.8	02:00.9	4.514117647	0.033516311
Stanford (F)	28.75	02:12.5	02:08.3	02:00.6	4.248571429	0.032036029
P.R.-Rio Piedras (F)	28.375	02:15.1	02:12.0	02:06.2	3.13	0.021718013
Oklahoma (F)	27.875	02:12.1	02:08.6	02:05.7	3.52	0.026224416
Weber State (F)	26.875	02:17.4	02:15.6	02:08.0	1.874615385	0.013296313
Clemson (F)	26.25	02:18.6	02:14.7	02:01.3	3.905454545	0.027079055
UTSA (F)	26.125	02:17.8	02:14.5	02:07.4	3.225	0.022399633
North Carolina St. (F)	26.125	02:16.3	02:11.4	02:03.5	4.87375	0.034962378
Kansas State (F)	25.125	02:12.7	02:09.8	02:04.4	2.9175	0.022272017
Connecticut (F)	24.75	02:13.1	02:09.6	02:04.2	3.408	0.025375742
Arizona State (F)	24.5	02:12.1	02:09.9	02:01.9	2.245	0.016841628
Sacramento St. (F)	24.5	02:13.7	02:10.3	02:03.0	3.34	0.024671457

Figure 19: The output of our recruiting recommendation system when (spread preference, best mark preference, improvement preference, winning preference, competitions per year preference) = (-10, 5, 10, 10, 5)

MileSplit [8], the data is much more protected and difficult to extract. However, this information is crucial to any successful recommendation system and ought to be added in the future. Secondly, our calculation of "fit" score could be more statistically rigorous. Currently, we assign "fit" points by sorting teams by a particular metric (for example, average best time in an event). However we do not take confidence intervals into account in this sorting process. A future system would perform the sorting process using a more rigorous method.

7. Acknowledgements

Huge thanks to Emily for helping me come up with the idea and to all of my teammates for pretending to be interested in my graphs. Thanks to Kathryn, Lizzie, Emily and Sarah for reading over the report! And huge thanks to Professor Dobkin and Ted for helping figure out to do with all the data.

References

[1] “2016 usot - qualifying standards,” http://www.usatf.org/Events---Calendar/2016/U-S--Olympic-Team-Trials---Track---Field/Athlete-Information/Qualifying-Standards.aspx.

[2] "Cecilia barowski," https://www.tfrrs.org/athletes/3733029/Princeton/Cecilia_Barowski.html.

[3] "Collegiate outdoor records," http://www.ustfccca.org/assets/record-book/collegiate-outdoor-all-time-top-ten-with-AC.pdf.

[4] "Columbia recruiting standards," http://www.gocolumbialions.com/ViewArticle.dbml?DB_OEM_ID=9600&ATCLID=210268000.

[5] "Dartmouth recruiting standards," http://www.dartmouthsports.com/ViewArticle.dbml?DB_OEM_ID=11600&ATCLID=204999331.

[6] "Direct athletics products," https://www.directathletics.com/products.html?tfrrs=1.

[7] "Harvard recruiting standards," http://www.gocrimson.com/sports/track/2014-15/Recruiting_Standards.pdf.

[8] "Mile split," http://www.milesplit.com.

[9] "Ncaa division i outdoor track and field rankings," http://www.ustfccca.org/team-rankings-polls-central/division-i-rankings.

[10] "Penn men recruiting standards," http://www.pennathletics.com/pdf9/3429309.pdf.

[11] "Penn women recruiting standards," http://www.pennathletics.com/pdf9/3429407.pdf?ATCLID=209448925&SPSID=8693&SPID=542&DB_LANG=C&DB_OEM_ID=1700.

[12] "Princeton," https://www.tfrrs.org/teams/NJ_college_f_Princeton.html.

[13] "Princeton women's recruiting standards," http://www.goprincetontigers.com/ViewArticle.dbml?ATCLID=208663264.

[14] "The track and field results reporting system, direct athletics inc." https://www.tfrrs.org/index.html.

[15] "What is the ncaa?" http://www.ncaa.org/about/resources/media-center/ncaa-101/what-ncaa.

[16] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.