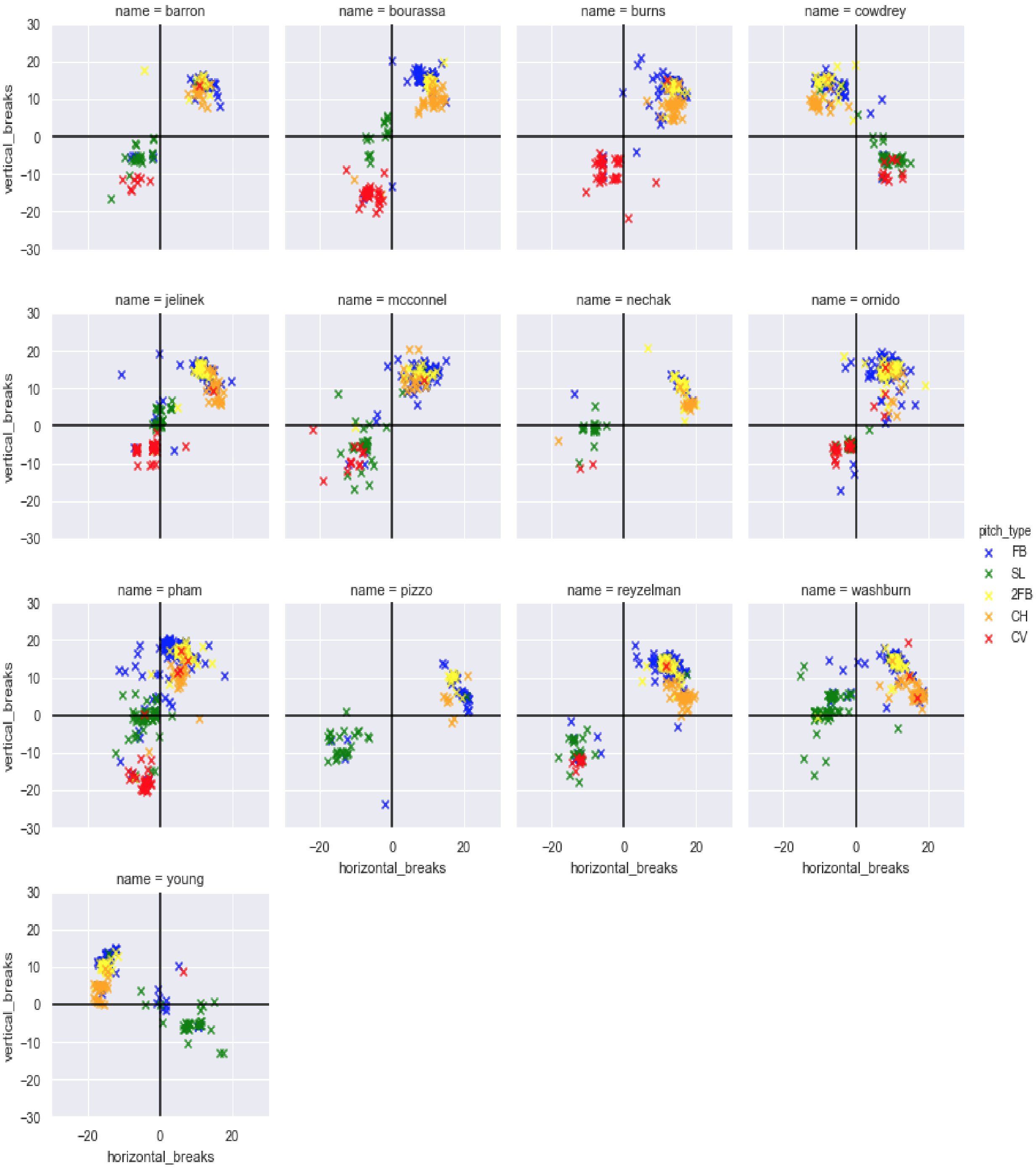
Rapsodo Dataset Analysis Back-End System Build

University of San Francisco Baseball Quantitative Studies Team

Zachariah Zanger MS, Tim Ryan, Annaliese Muth



Introduction

I think it was on a Saturday night in late January or early February when I sat in the corner of Vertigo Coffee Roasters in San Juan Bautista to do work on my computer. To be exact, I sat down to work with our growing Rapsodo dataset. Compounded by working with one arm (shoulder surgery in December), I was frustrated. Working with the Rapsodo data by downloading PDF's was inefficient and vastly time consuming in that session. It was like I kept playing whack-a-mole with new problems in the dataset trying to do things without capturing scale. At that point, I knew we needed to build a system around the dataset. A system that captured scale and could utilize a variety of frameworks for analysis.

One thing that went through my mind was our (QST) inefficient model that worked with the Pitch Execution charts from the 2019 season. I did not pump the brakes on that inefficient process/model at the time. My mind went in the direction of learning from my mistake and not letting the same thing happen with our new Rapsodo dataset. The Rapsodo dataset is large and messy (3,400 pitches from last season). There is a lot of data importing and cleaning to set up for actual analysis. Once that is complete, actual analysis is carried out. I will define an efficient system for this specific context on these two components: importing / cleaning data and conducting analysis, automating as much as responsibly possible. We have done both and laid the groundwork for future use of this dataset. I will iterate on future work in the conclusion.

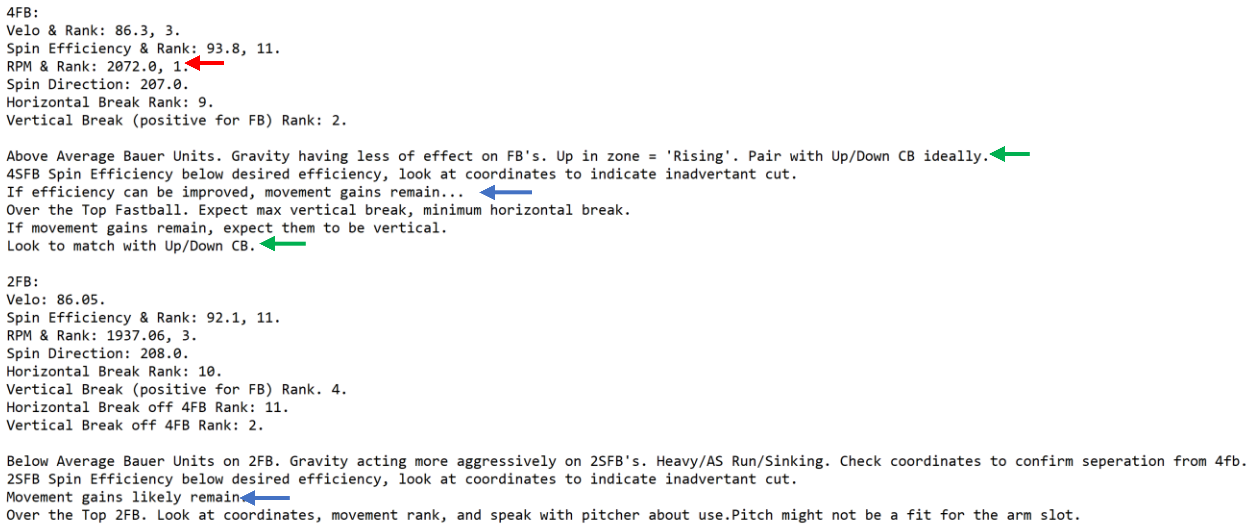
This writing serves the purpose of explaining the output of the Rapsodo system and infrastructure we have built throughout this Summer that will be applied to evaluation and development of our pitching staff going beyond the stat line, at the pitch-level. The writing does not serve the purpose of iterating on the nuts and bolts of the 2,966 lines of code we wrote over the Summer. There are three primary components of output: a text-file that generates analysis based on rules, leaderboards for any metric and different customizations, and data visualizations. *The Rapsodo camera does not measure aptitude for competition, ability to execute in games, or mental strength. These things are vital to success and are not measured with anything in the work developed below. We know that.* But we have done our best to assemble a system that handles the objectivity involved in this dataset and pull out as much as we can. We will continue to work, making this back-end system even better moving into the future.

Lastly, and most importantly, this is a team effort. Tim Ryan and Annaliese Muth have been exceptional to say the least. They did a lot of heavy lifting, answered my phone calls all Summer (some at non-working hours), always came to zoom calls ready, were creative, and were diligent in their work. This deliverable would not be possible without them.

Text File

The text file output is an automated analysis of a pitcher's population of pitches on the Rapsodo camera. You can think of it like a compass. It is not a pinpointed representation, but approximation of direction to go in and perspective of where you stand. The text file does not re-invent the wheel, but instead takes relationships known to be true among the different Rapsodo variable outputs and applies them in automated writing. This is key because with a large dataset like the Rapsodo one, there is significant risk to 'eyeball' things or perform significant amounts of mental algebra. It is feasible and advantageous to pass off the grunt work (Mental Algebra) to a computer that does not need coffee to keep working.

For a player's text file, there is a summary section for every pitch, and then rules generated writing for the end user to examine. Relationships and rules utilized are coming from Driveline's Pitch Design course as well as public resources in pitch design and shaping. I think it will be best if I break down a few parts of different text files (we can generate them for every arm on the staff from last season with one click) to explain how the text files work. Starting with Riley Ornido fastballs, both 4-Seam and 2-Seam:



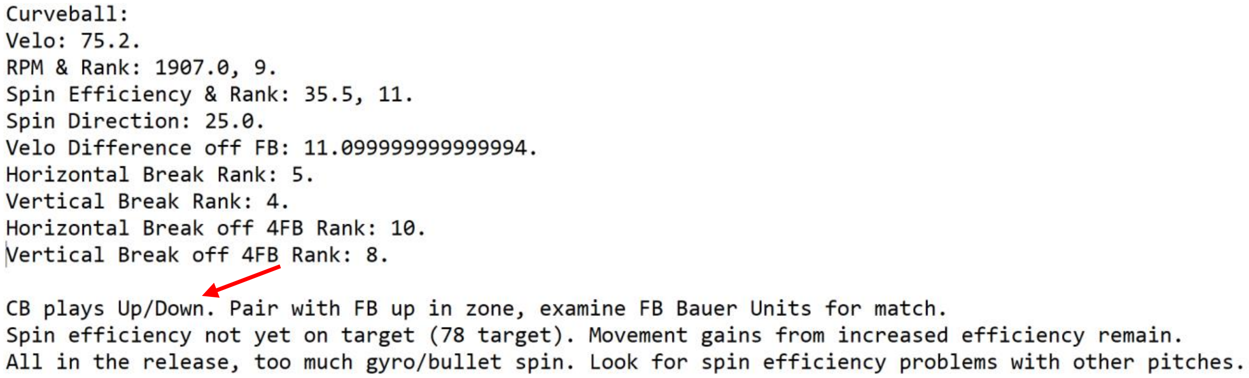
```
4FB:
Velo & Rank: 86.3, 3.
Spin Efficiency & Rank: 93.8, 11.
RPM & Rank: 2072.0, 1.
Spin Direction: 207.0.
Horizontal Break Rank: 9.
Vertical Break (positive for FB) Rank: 2.

Above Average Bauer Units. Gravity having less of effect on FB's. Up in zone = 'Rising'. Pair with Up/Down CB ideally.
4SFB Spin Efficiency below desired efficiency, look at coordinates to indicate inadvertant cut.
If efficiency can be improved, movement gains remain...
Over the Top Fastball. Expect max vertical break, minimum horizontal break.
If movement gains remain, expect them to be vertical.
Look to match with Up/Down CB.

2FB:
Velo: 86.05.
Spin Efficiency & Rank: 92.1, 11.
RPM & Rank: 1937.06, 3.
Spin Direction: 208.0.
Horizontal Break Rank: 10.
Vertical Break (positive for FB) Rank. 4.
Horizontal Break off 4FB Rank: 11.
Vertical Break off 4FB Rank: 2.

Below Average Bauer Units on 2FB. Gravity acting more aggressively on 2SFB's. Heavy/AS Run/Sinking. Check coordinates to confirm seperation from 4fb.
2SFB Spin Efficiency below desired efficiency, look at coordinates to indicate inadvertant cut.
Movement gains likely remain
Over the Top 2FB. Look at coordinates, movement rank, and speak with pitcher about use.Pitch might not be a fit for the arm slot.
```

The text is a bit small here (I apologize), but for every pitch, it starts with summary statistics. In the summary statistics, there are estimates of true velocity, spin efficiency, rpm, spin direction, and then both horizontal and vertical breaks. Some of these estimates also have rankings on the staff for the pitcher selected. In this instance, we estimate Riley Ornido true 4-Seam Fastball RPM as the highest on the staff (indicated by red arrow).

The section below the summary statistics is where the computer performs analysis based on rules and relationships programmed in. At this point, it should be stressed that this is a high-level analysis and that the goal of this component is to eliminate mental algebra. The end user still must pick up where the text file ends and continue the analysis. In this section, multiple rules are triggered that suggest an Up/Down Curveball pairs best with Riley Ornido's high spin / positive vertical break 4-Seam Fastball (indicated by green arrows).

From the text file, rules for both 4-Seam and 2-Seam Fastballs are triggered that spin efficiency is an issue. Riley Ornido is not getting as much true spin on his Fastballs and this is capping movement (both horizontal and vertical). On the same note, a rule is triggered (blue arrows) that if spin efficiency can be fixed, movement gains from spin efficiency remain.

At this point, one would look down to the section of Riley Ornido's text file to see if his Curveball matches as an Up/Down Curveball to pair with the text suggested by the green arrows. There is a match (red arrow below)

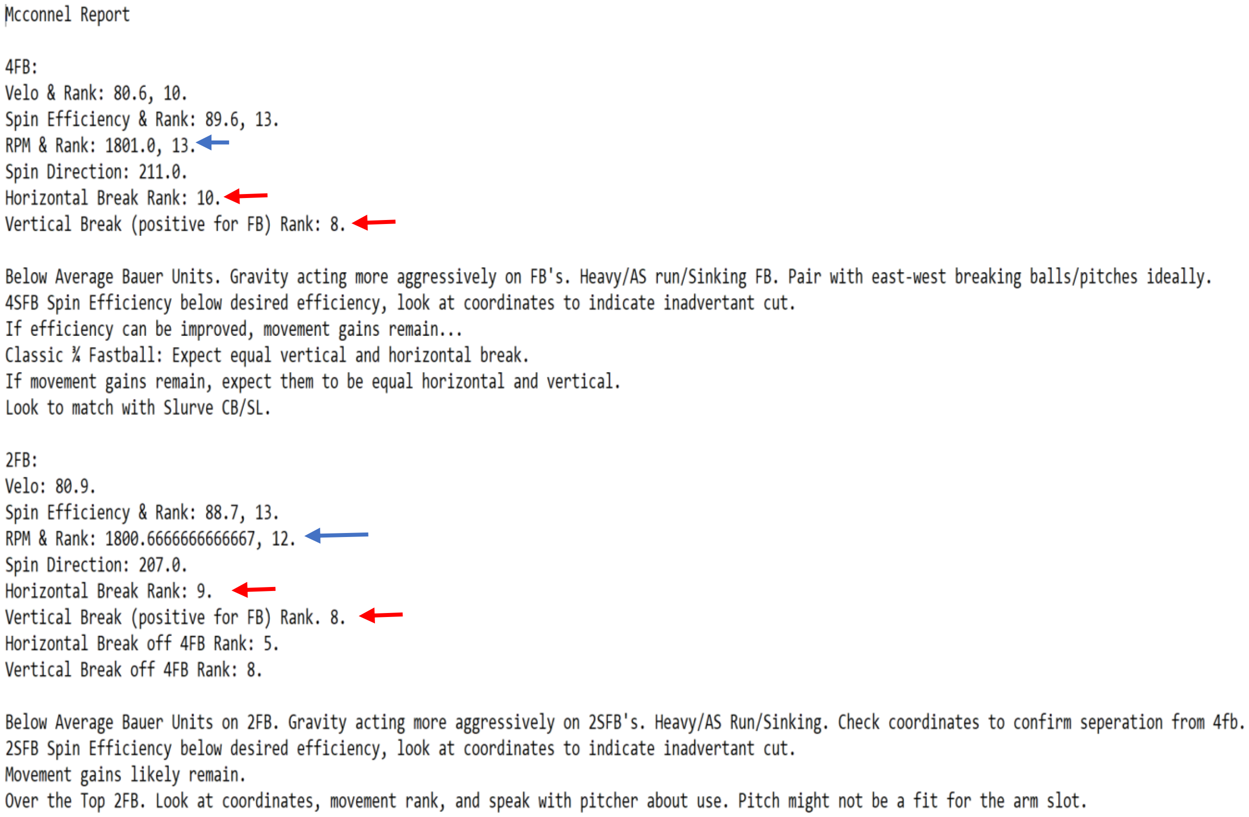


```
Curveball:
Velo: 75.2.
RPM & Rank: 1907.0, 9.
Spin Efficiency & Rank: 35.5, 11.
Spin Direction: 25.0.
Velo Difference off FB: 11.099999999999994.
Horizontal Break Rank: 5.
Vertical Break Rank: 4.
Horizontal Break off 4FB Rank: 10.
Vertical Break off 4FB Rank: 8.

CB plays Up/Down. Pair with FB up in zone, examine FB Bauer Units for match.
Spin efficiency not yet on target (78 target). Movement gains from increased efficiency remain.
All in the release, too much gyro/bullet spin. Look for spin efficiency problems with other pitches.
```

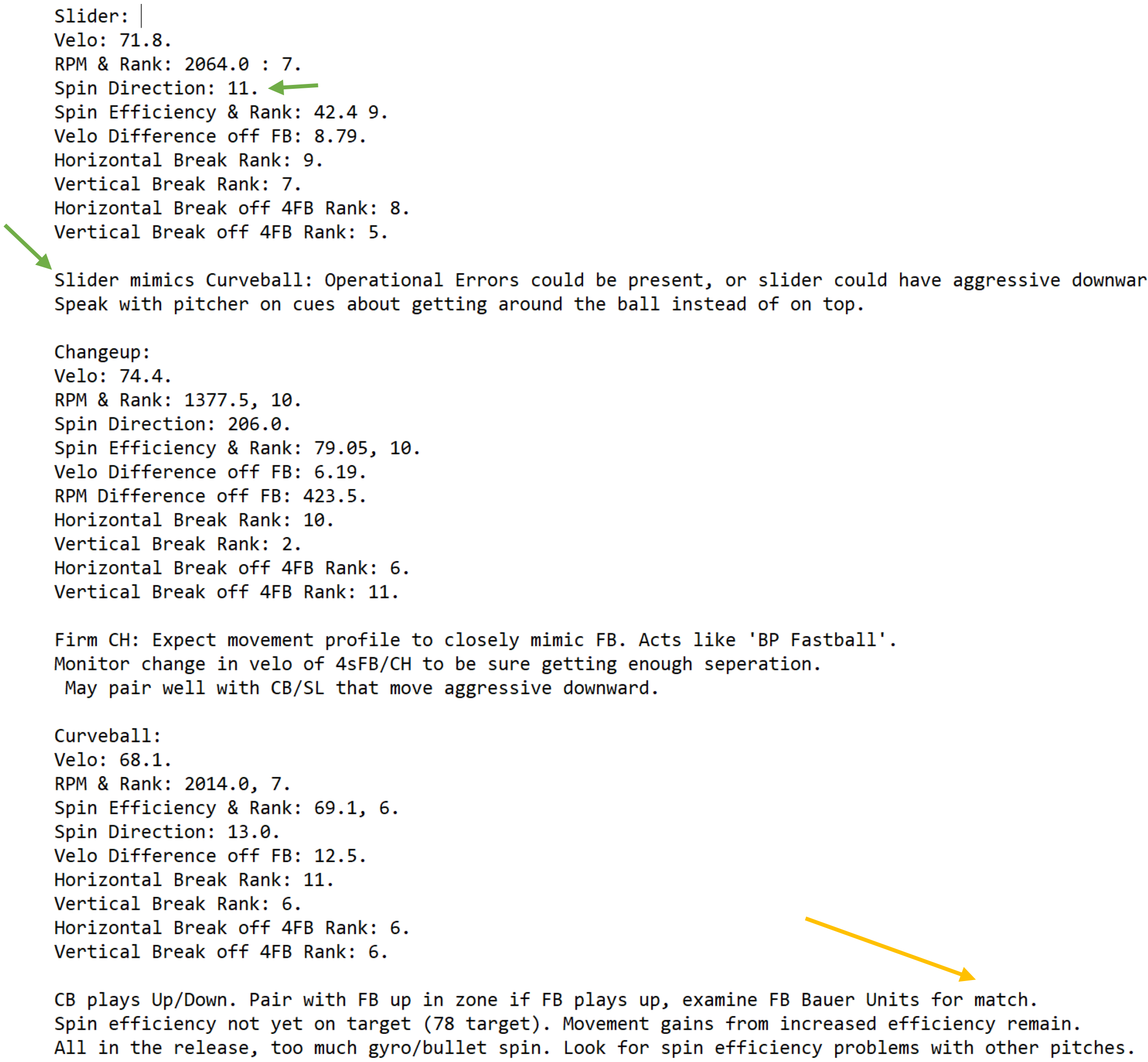
We also see in the text that Riley Ornido has the same spin efficiency problems with his Curveballs as with his Fastballs. This example is to show that from the automated text generation, we were quickly able to conclude that: *Riley Ornido throws a high spin 4-Seam Fastball that is optimally paired with an Up/Down Curveball. Riley Ornido's Curveball type is a match with his 4-Seam Fastball. Both pitches suffer from too much inefficient gyro/bullet spin. If by pitch release cue'ing, both pitches will increase movement gains and separation.*

Again, this is not re-inventing the wheel. This is being efficient and handing off the grunt work to the computer. In a few clicks, we were able to perform objective analysis and gain high level insight into where gains remain. We could pick up the analysis from here and conduct a more thorough deep dive.

This is what the entire text file looks like for Mat McConnel:



```
Mcconnel Report

4FB:
Velo & Rank: 80.6, 10.
Spin Efficiency & Rank: 89.6, 13.
RPM & Rank: 1801.0, 13.
Spin Direction: 211.0.
Horizontal Break Rank: 10.
Vertical Break (positive for FB) Rank: 8.

Below Average Bauer Units. Gravity acting more aggressively on FB's. Heavy/AS run/Sinking FB. Pair with east-west breaking balls/pitches ideally.
4SFB Spin Efficiency below desired efficiency, look at coordinates to indicate inadvertant cut.
If efficiency can be improved, movement gains remain...
Classic ¾ Fastball: Expect equal vertical and horizontal break.
If movement gains remain, expect them to be equal horizontal and vertical.
Look to match with Slurve CB/SL.

2FB:
Velo: 80.9.
Spin Efficiency & Rank: 88.7, 13.
RPM & Rank: 1800.6666666666667, 12.
Spin Direction: 207.0.
Horizontal Break Rank: 9.
Vertical Break (positive for FB) Rank. 8.
Horizontal Break off 4FB Rank: 5.
Vertical Break off 4FB Rank: 8.

Below Average Bauer Units on 2FB. Gravity acting more aggressively on 2SFB's. Heavy/AS Run/Sinking. Check coordinates to confirm seperation from 4fb.
2SFB Spin Efficiency below desired efficiency, look at coordinates to indicate inadvertant cut.
Movement gains likely remain.
Over the Top 2FB. Look at coordinates, movement rank, and speak with pitcher about use. Pitch might not be a fit for the arm slot.
```



```
Slider:
Velo: 71.8.
RPM & Rank: 2064.0 : 7.
Spin Direction: 11.
Spin Efficiency & Rank: 42.4 9.
Velo Difference off FB: 8.79.
Horizontal Break Rank: 9.
Vertical Break Rank: 7.
Horizontal Break off 4FB Rank: 8.
Vertical Break off 4FB Rank: 5.

Slider mimics Curveball: Operational Errors could be present, or slider could have aggressive downward action.
Speak with pitcher on cues about getting around the ball instead of on top.

Changeup:
Velo: 74.4.
RPM & Rank: 1377.5, 10.
Spin Direction: 206.0.
Spin Efficiency & Rank: 79.05, 10.
Velo Difference off FB: 6.19.
RPM Difference off FB: 423.5.
Horizontal Break Rank: 10.
Vertical Break Rank: 2.
Horizontal Break off 4FB Rank: 6.
Vertical Break off 4FB Rank: 11.

Firm CH: Expect movement profile to closely mimic FB. Acts like 'BP Fastball'.
Monitor change in velo of 4sFB/CH to be sure getting enough seperation.
 May pair well with CB/SL that move aggressive downward.

Curveball:
Velo: 68.1.
RPM & Rank: 2014.0, 7.
Spin Efficiency & Rank: 69.1, 6.
Spin Direction: 13.0.
Velo Difference off FB: 12.5.
Horizontal Break Rank: 11.
Vertical Break Rank: 6.
Horizontal Break off 4FB Rank: 6.
Vertical Break off 4FB Rank: 6.

CB plays Up/Down. Pair with FB up in zone if FB plays up, examine FB Bauer Units for match.
Spin efficiency not yet on target (78 target). Movement gains from increased efficiency remain.
All in the release, too much gyro/bullet spin. Look for spin efficiency problems with other pitches.
```

From the text and the summary statistics, starting with Fastballs, Mat McConnel generates some of the lowest spin on his Fastballs on the staff (blue arrows). We should expect a lot of either vertical or horizontal break on both pitches. This part could be confusing, but since the higher the vertical break on Fastballs the higher the ranking (Pham is #1), Mat McConnel generates more sinking vertical movement on his Fastballs in comparison to the staff (hence the lower ranking). Of the two (Horizontal and Vertical Movement), Mat McConnel Fastballs are likely to generate more vertical break downward (red arrows).

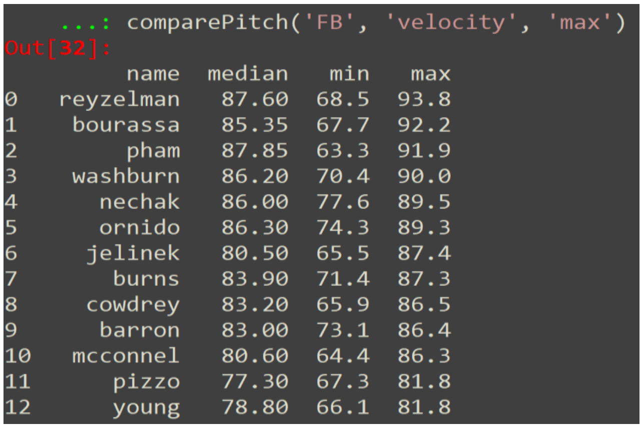
Moving on to breaking pitches, Mat McConnel gets a rule triggered for his Slider spin direction. The spin direction is rotating at essentially 7:00 and looking a lot like a Curveball as it rotates. The text communicates this (green arrows). This could start a conversation with Mat McConnel about how he feels about both pitches and which he feels better about. He has options. We do get a rule stating Mat McConnel's Curveball is an Up/Down Curveball and we should look at if there is a match with his Fastballs (orange arrow). But there is not.

From the text file, we see no clear matches in pitch combinations (maybe his firm Changeup and Up/Down Curveball match). Is this a bad thing? No, Mat McConnel is on his way to developing as a Division 1 pitcher. His Fastballs, both 4-Seam and 2-Seam, already generate sink, and this could be a starting point building out his repertoire. Just because there are no answers now does not mean there will be none in the future. We must continue to monitor and most importantly listen to how Mat feels about his repertoire moving forward.

Leaderboards

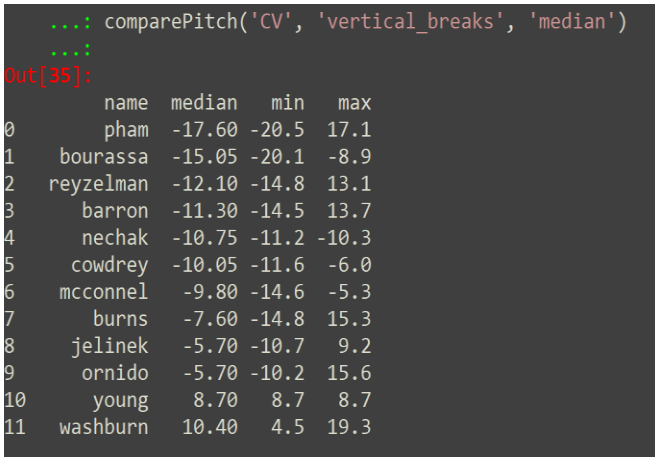
We have automated the process of creating leaderboards on the staff. I am going to list off a few examples. We are able to automatically generate leaderboards for the minimum, maximum, and median (most likely approximation) of any Rapsodo data field.

Max Fastball Velocity



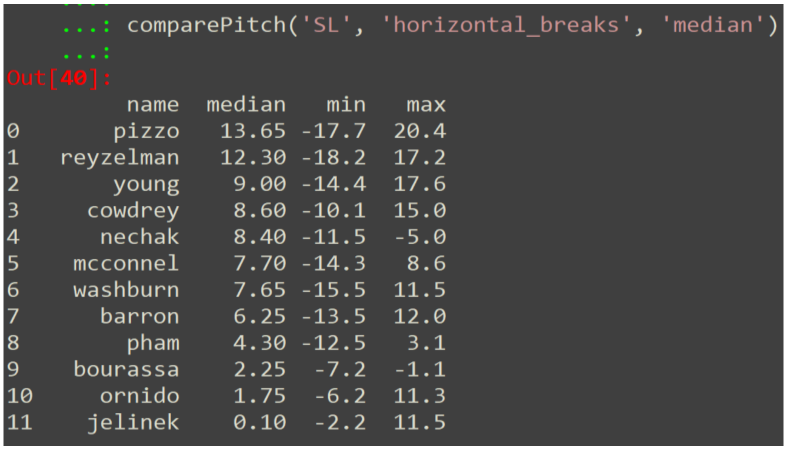
```
    ...: comparePitch('FB', 'velocity', 'max')
Out[32]:
         name  median   min   max
0   reyzelman   87.60  68.5  93.8
1    bourassa   85.35  67.7  92.2
2        pham   87.85  63.3  91.9
3    washburn   86.20  70.4  90.0
4      nechak   86.00  77.6  89.5
5      ornido   86.30  74.3  89.3
6     jelinek   80.50  65.5  87.4
7       burns   83.90  71.4  87.3
8     cowdrey   83.20  65.9  86.5
9      barron   83.00  73.1  86.4
10   mcconnel   80.60  64.4  86.3
11      pizzo   77.30  67.3  81.8
12      young   78.80  66.1  81.8
```

Median (Middle) Curveball Vertical Break



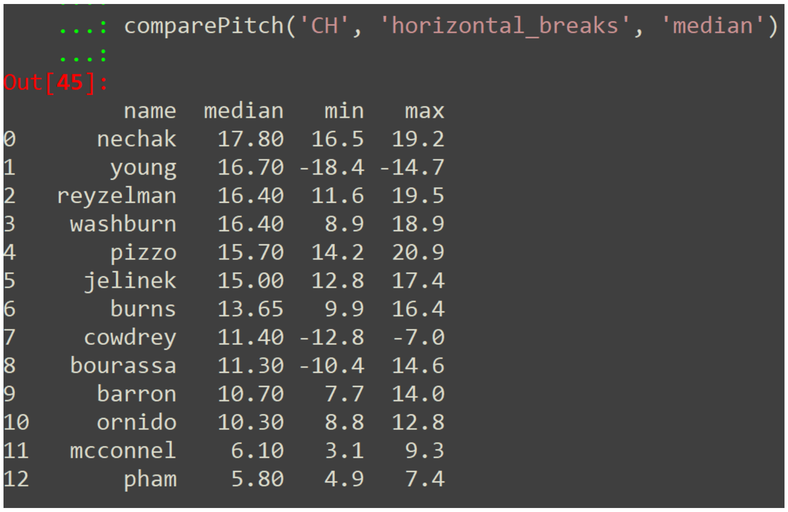
```
    ...: comparePitch('CV', 'vertical_breaks', 'median')
    ...:
Out[35]:
         name  median   min   max
0        pham  -17.60 -20.5  17.1
1    bourassa  -15.05 -20.1  -8.9
2   reyzelman  -12.10 -14.8  13.1
3      barron  -11.30 -14.5  13.7
4      nechak  -10.75 -11.2 -10.3
5     cowdrey  -10.05 -11.6  -6.0
6    mcconnel   -9.80 -14.6  -5.3
7       burns   -7.60 -14.8  15.3
8     jelinek   -5.70 -10.7   9.2
9      ornido   -5.70 -10.2  15.6
10      young    8.70   8.7   8.7
11   washburn   10.40   4.5  19.3
```

Median (Middle) Slider Horizontal Break



```
    ...: comparePitch('SL', 'horizontal_breaks', 'median')
    ...:
Out[40]:
         name  median   min   max
0       pizzo   13.65 -17.7  20.4
1   reyzelman   12.30 -18.2  17.2
2       young    9.00 -14.4  17.6
3     cowdrey    8.60 -10.1  15.0
4      nechak    8.40 -11.5  -5.0
5    mcconnel    7.70 -14.3   8.6
6    washburn    7.65 -15.5  11.5
7      barron    6.25 -13.5  12.0
8        pham    4.30 -12.5   3.1
9    bourassa    2.25  -7.2  -1.1
10     ornido    1.75  -6.2  11.3
11    jelinek    0.10  -2.2  11.5
```

Median (Middle) Changeup Horizontal Break



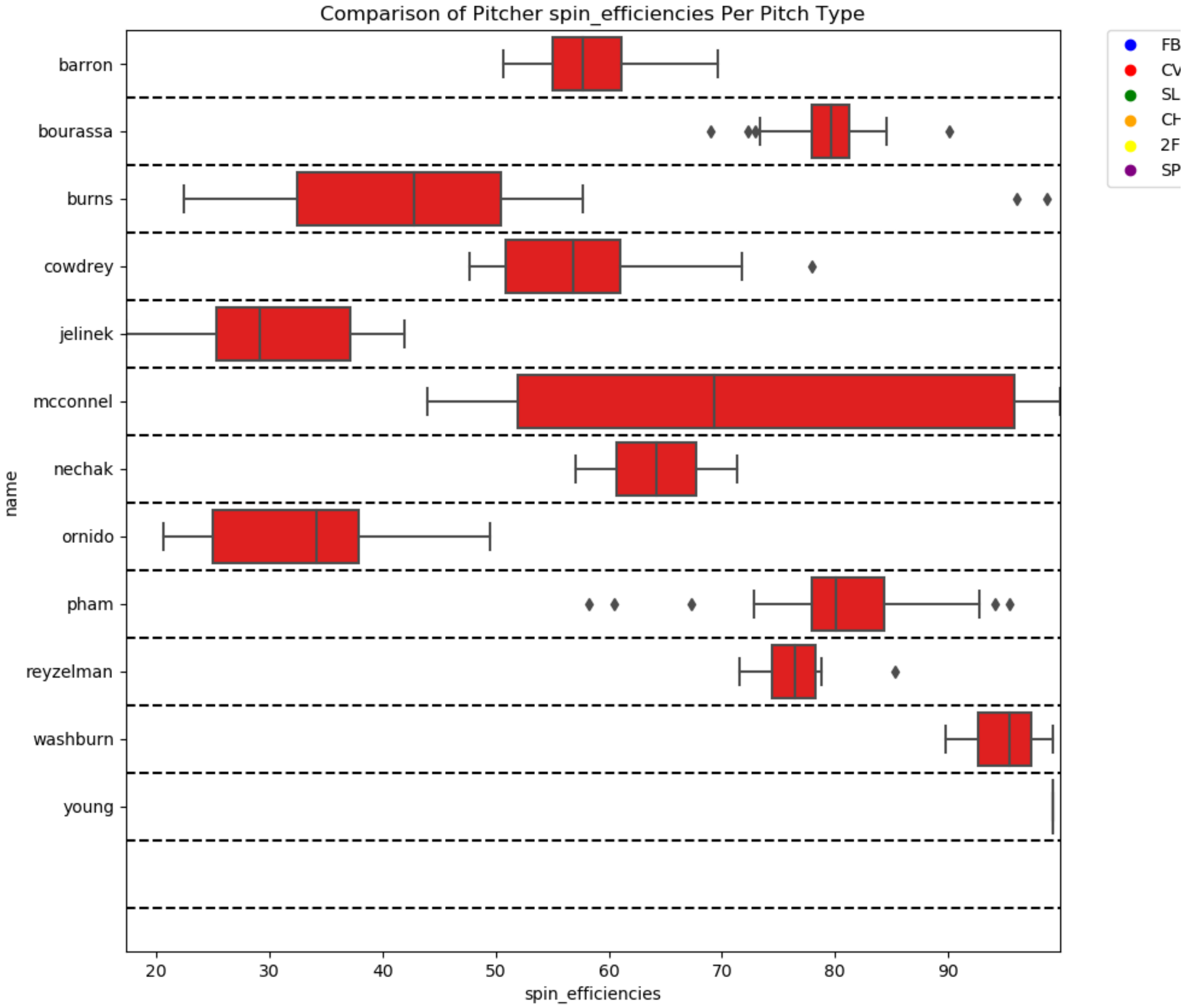
```
    ...: comparePitch('CH', 'horizontal_breaks', 'median')
    ...:
Out[45]:
         name  median   min   max
0      nechak   17.80  16.5  19.2
1       young   16.70 -18.4 -14.7
2   reyzelman   16.40  11.6  19.5
3    washburn   16.40   8.9  18.9
4       pizzo   15.70  14.2  20.9
5     jelinek   15.00  12.8  17.4
6       burns   13.65   9.9  16.4
7     cowdrey   11.40 -12.8  -7.0
8    bourassa   11.30 -10.4  14.6
9      barron   10.70   7.7  14.0
10     ornido   10.30   8.8  12.8
11   mcconnel    6.10   3.1   9.3
12       pham    5.80   4.9   7.4
```

The leaderboards for metrics at the pitch level will move throughout the year and will be updated automatically.

<u>Data Visualizations</u>

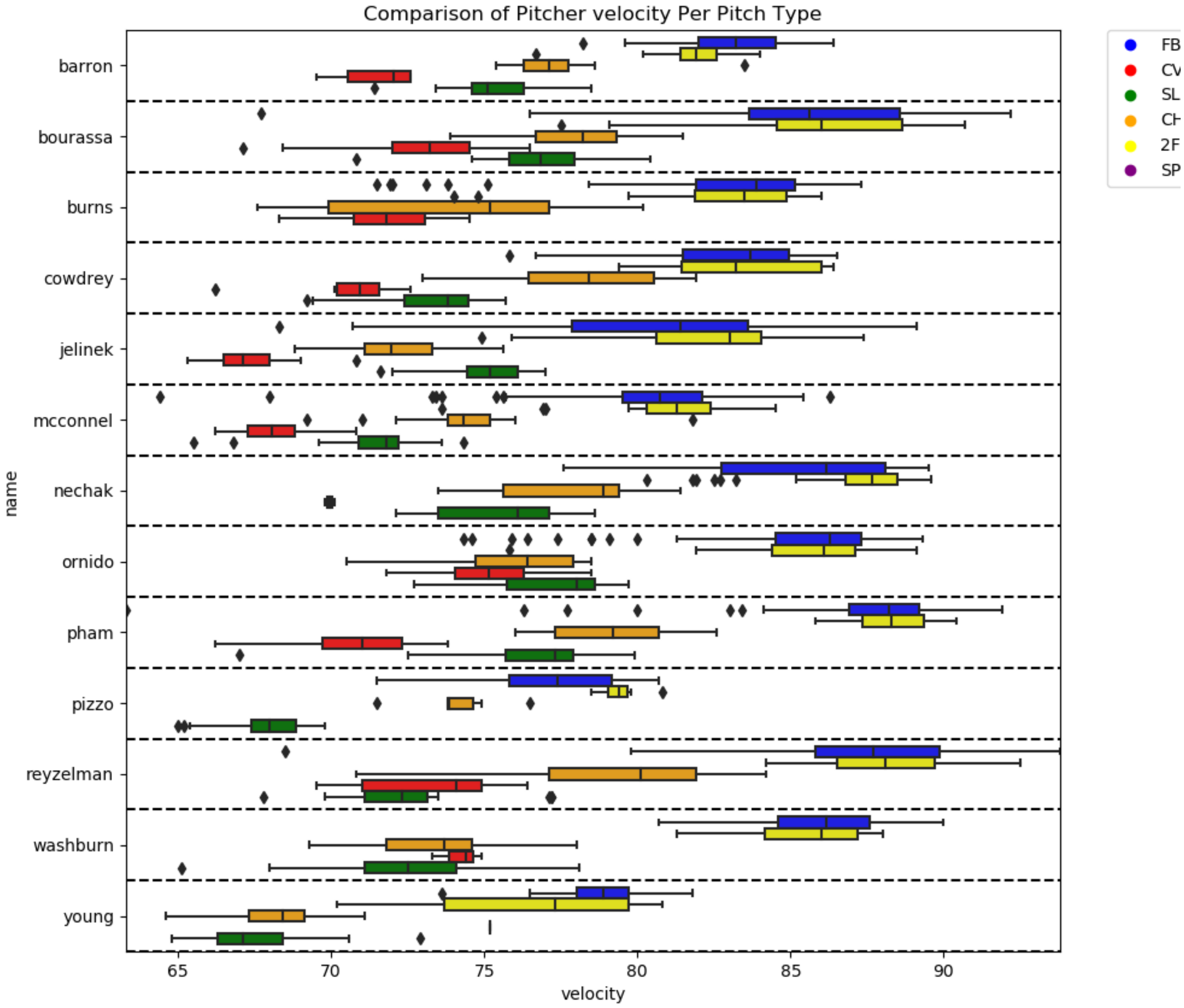
We have automated box-and-whisker plots for visualizing what data looks like at a high level. I am not going to get too into the weeds of boxplots, but in the plots: the left and right corner of the colored box represents the 25th and 75th percentile of the data. The line down inside the box represents the middle value (median). The whiskers, or lined ends of the plot, represent reasonable outer limits. And individual points outside the whiskers represent extreme values.

Box and Whisker Plots of Curveball Spin Efficiencies



The more efficient the spin on a Curveball, the better. Spin efficiency measures how consistently the ball rotates end over end in flight. In the visual above, a smaller red box means there is less variance in the values. Landen Bourassa throws a consistent and high spin efficient Curveball. His box is very narrow and concentrated toward the middle value of approximately 81% spin efficiency. Luckily this visual came up, because the text file does not tell the whole story for Mat McConnel's Curveball. Mat McConnel, at times, throws a very efficient Curveball. The spin efficiency, combined with the spin direction and total spin, are the three primary contributors to break on the Curveball. From this visual, we could say Mat McConnel has it in him and has showed it in the past to throw Curveballs with high amounts of spin efficiency.

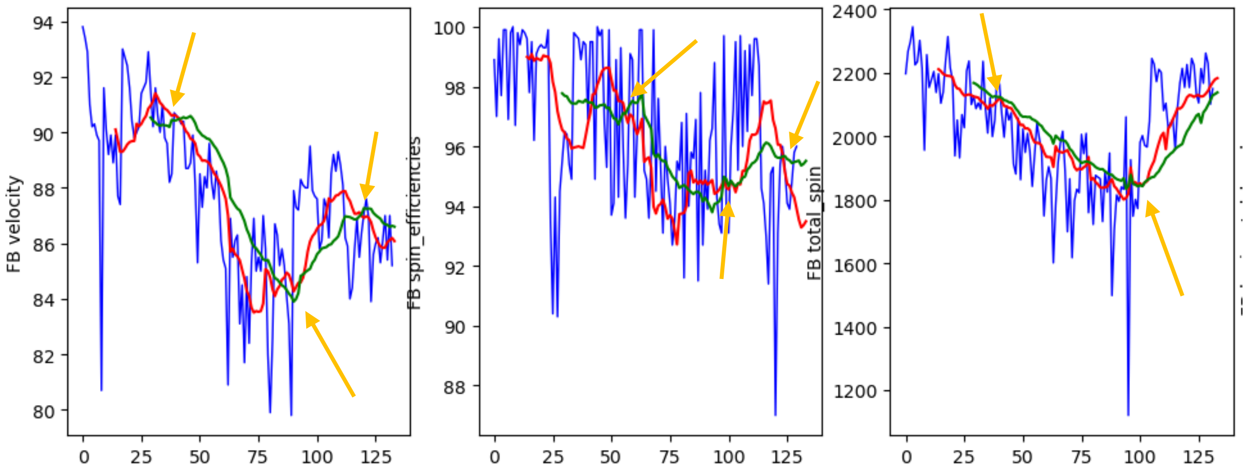
Box and Whisker Plots of Velocities for Different Pitches



This is just an example of velocities for different pitch types of every pitcher on the staff. Alex Pham and Eric Reyzelman back off velocity very well moving from 4-Seam Fastball to Curveball. Grant Nechak has a higher max velocity on his 2-Seam Fastball and the velocity is more consistent (smaller box) than his 4-Seam Fastball (bigger box). Alex Pham’s text file says he throws a “Firm” Changeup. The text file says to monitor how well he can back off MPH’s, since he is not picking up much break on his Firm Changeup. From the visual, Pham will most likely back off ~10 mph between the two pitches.

These are two examples of boxplots we can dial up easily with one click. The infrastructure is already in place to generate these plots very quickly for any customization.

Moving Average Visuals (And Prediction)



The blue lines represent, over time, Eric Reyzelman 4-Seam Fastball Velocities, Spin Efficiencies, and Spin Rate. The point of these visuals is not to just explain trends, but predict movement, either upward or downward. The red lines represent the 15-pitch moving average. The green lines represent the 30-pitch moving average.

In short, *when the 15-pitch moving average crosses BELOW the 30-pitch moving average, we expect a sustained negative downward movement in output. When the 15-pitch moving average crosses ABOVE the 30-pitch moving average, we expect a sustained upward movement in output.*

The arrows indicate different points when a cross occurred and a sustained upward or downward movement followed. We have not done it yet, but the goal is to incorporate in the text file a rule trigger for when a cross occurs for specific fields. We can customize for different pitches very easily. We also need to do more testing on this, because this is very much original thinking and not proven to be applicable to something like the Rapsodo dataset.

Conclusion

We will continue to work. I remain steadfast in my belief that as bigger datasets become more prevalent in the college game, programs that are unable to capture scale and build the back-end systems will fall behind. No model is perfect. And while this is more of a system, it also is not perfect. *The biggest obstacle, a lot like all Baseball modeling, is capturing the context that things happen* in. The text file could say one thing, but if we fail to pick up on something like a player making mechanical adjustments or experimenting with different cue's, we could lose explanatory power in the file. We will do our best to stay on top of context. And of course, we will continue to build better rules, try to make things clearer, and unlock new insights from this rich dataset.

I have ideas with respect to applying the Rapsodo dataset to things like recruiting and probabilistic pitch risk; but we (QST) must shift our focus to rollout and implementation of the things the system does for this Fall. Given the uncertainty around the Fall season, we will do our part to uncover as much objectivity as possible in the dataset in preparation for the 2021 campaign.