Making Major League Data Work

Carving up big data into useful applications for specific audiences

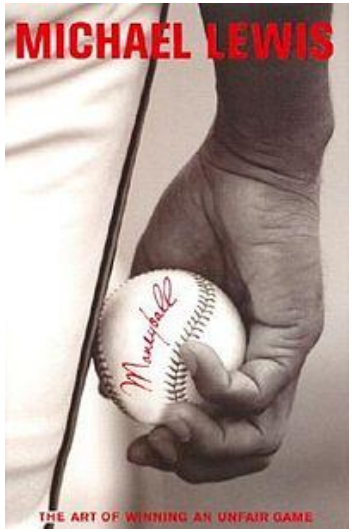
Image: http://en.wikipedia.org/wiki/Baseball

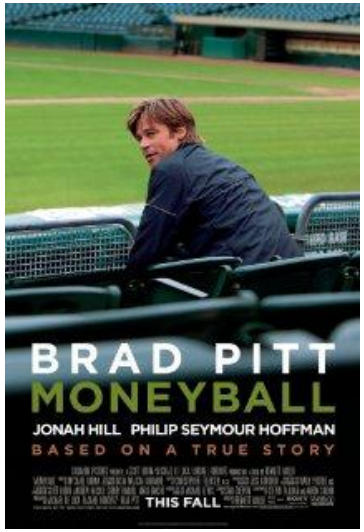
Bloomberg SPORTS oculus

Background

Bloomberg and Sports?

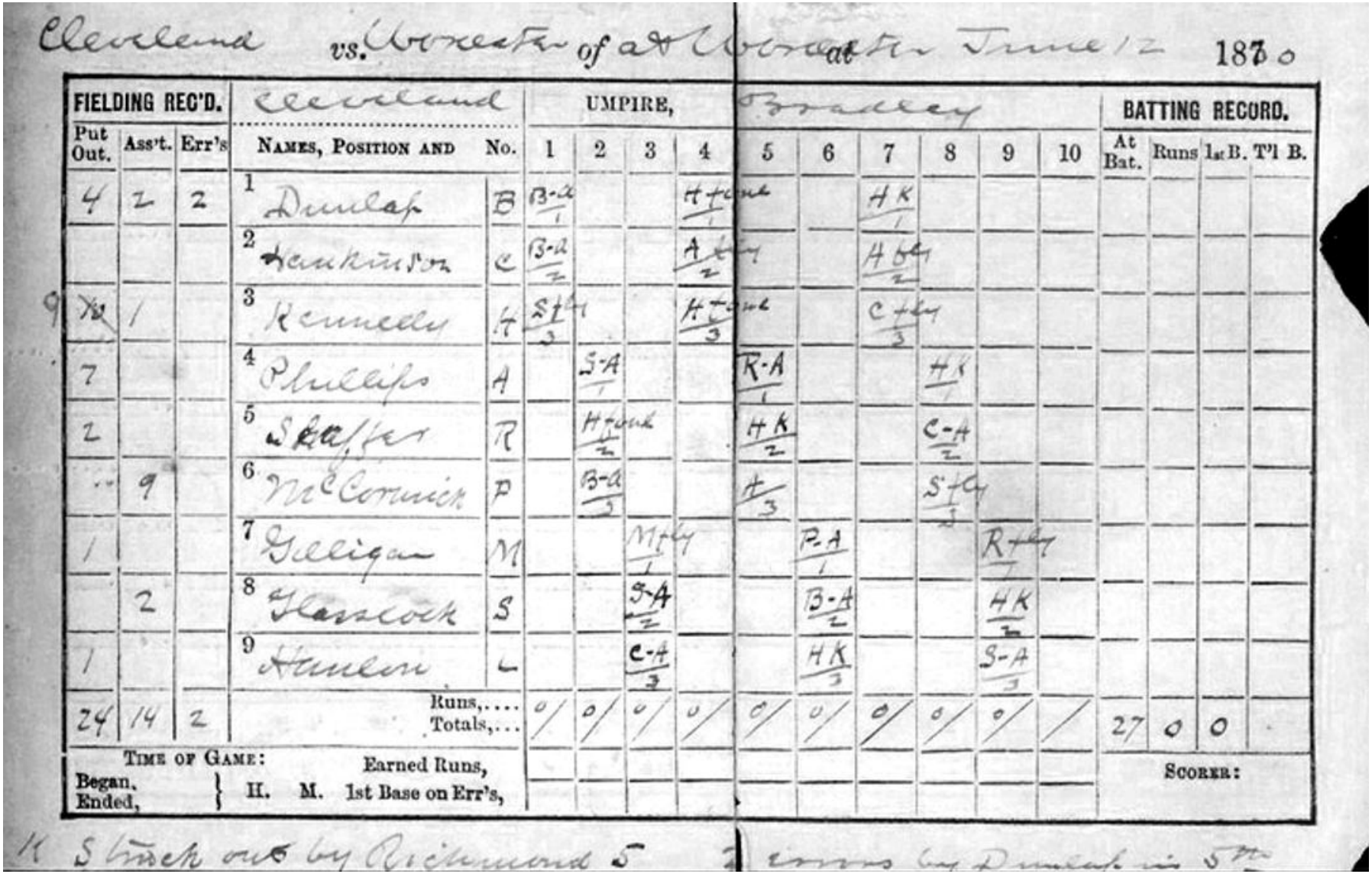
- People have used the Bloomberg Terminal to evaluate their stocks and portfolios for the past 30 years
- History steeped deeply in making data accessible
- Four years ago Bloomberg began looking at other data heavy businesses
- Plan was simple – players are stocks and teams are portfolios and start in a sport sparked by the Moneyball revolution
- Married news, biographies, and statistics
- Told a good story, not a great one
- Teamed up with Oculus to tell that great story through visualization
- Presented by Noah Schwartz, CTO Bloomberg Sports & Richard Brath, Partner at Oculus





Bloomberg SPORTS oculus

Since The 1800's Simple Statistics Have Been Tracked



Cleveland vs. Worcester of at Worcester June 12 1880

FIELDING REC'D. Put Out.	Ass't.	Err's	Cleveland — Names, Position and No.		UMPIRE, 1	2	3	4	5	6	7	8	9	10	BATTING RECORD. At Bat.	Runs	1st B.	T'l B.
4	2	2	1 Dunlap	B	B-A 1			H+ 1			H K 1							
			2 Hankinson	C	B-A 2			A+ 2			H+ 2							
9	1		3 Kennedy	H	S+ 3			H+ 3			C+ 3							
7			4 Phillips	A		S-A 1			R-A 1			H K 1						
2			5 Shaffer	R		H+ 2			H K 2			C-A 2						
	9		6 McCormick	P		B-A 3			A 3			S+ 3						
1			7 Galligan	M			M+ 1			P-A 1			R+ 1					
	2		8 Glasscock	S			S-A 2			B-A 2			H K 2					
1			9 Hanlon	L			C-A 3			H K 3			S-A 3					
24	14	2	Runs,... Totals,...		0	0	0	0	0	0	0	0	0		27	0	0	

TIME OF GAME: Began, Ended, } H. M. Earned Runs, 1st Base on Err's,

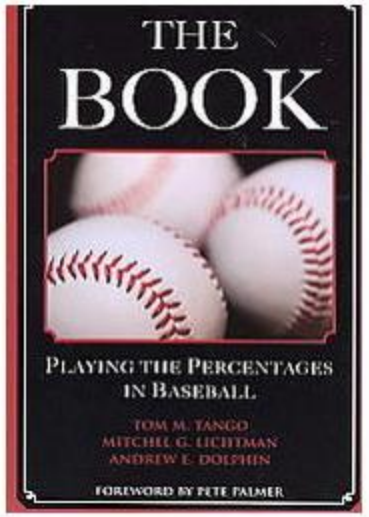
SCORER:

K Struck out by Richmond 5 ... runs by Dunlap in 5th

http://en.wikipedia.org/wiki/File:Lee-richmond-perfect-game-scorecard-2.jpeg

Statistics Evolved Over The Years

- Many changes over the next 100+ years of baseball
- For a long time people looked mainly at batting average (hits/at bats) to judge how good a hitter was
- The problem is that batters of very different types and qualities can have similar batting averages
- **On Base Percentage:** Values other important factors such as walks
- **Slugging Percentage:** Values the extra-base hits
- Triple slash: AVG/OBP/SLG
- Many people added together OBP and SLG to form OPS which is an overall measure of success
- wOBA is a further evolution which tries to differentiate between OBP (more important) and SLG (less) while still coming up with one comprehensive number. It weights all of the outcomes correctly by looking at how good they are, historically, at generating runs.



$$wOBA = \frac{(0.72 * NIBB) + (0.75 * HBP) + (0.90 * 1B) + (0.92 * RBOE) + (1.24 * 2B) + (1.56 * 3B) + (1.95 * HR)}{PA}$$

Data Becomes more Granular and Real-Time

- Yankees-Tigers ALCS Game 3 Play-by-Play Sample Feed
  - 23:21:50 Pitching Change: Phil Coke replaces Justin Verlander.
  - 23:16:23 Eduardo Nunez homers (1) on a line drive to left field.
  - **22:02:10 New York Yankees pitcher Phil Hughes left the game due to an injured back.**
  - 21:52:39 Pitching Change: Clay Rapada replaces David Phelps.
  - 21:51:21 Miguel Cabrera doubles (1) on a line drive to center fielder Curtis Granderson. Quintin Berry scores.
  - 21:22:56 Pitching Change: David Phelps replaces Phil Hughes.
  - 21:19:51 Injury Delay.
  - 21:16:51 Delmon Young homers (2) on a line drive to left field.

- Yankees-Tigers ALCS Game 3 Real-time Bloomberg News Feed
  - BSP 23:36:23 TIGERS DEFEAT YANKEES 2-1 TO LEAD ALCS THREE GAMES TO NONE
  - NYT 23:30:01 Bats: Benching Rodriguez and Swisher Was a Group Decision
  - APW 22:43:20 Tigers Lead Yankees 2-0 After 6 Innings in ALCS
  - **APW 22:11:40 Yankees RHP Hughes Leaves ALCS Game With Injury**

PITCHf/x by Sportvision

- Focal point of this presentation
- Hardware/Software solution by Sportvision that captures the physics of each pitch
- In addition to the basics, we now have the X,Y on plane of release, the X,Y crossing the plate, the break, the velocity at release, and the velocity across plate
- Debuted in 2006 MLB Playoffs
- Used in on-air graphics
- Controversial to some degree as baseball is officiated by umpires and sometimes PITCHf/x will disagree
- Used in MLB Gameday to do 3D reconstructions of pitches

Bloomberg SPORTS oculus

Pitch Analysis

- PITCHf/x allowed us to go down a level
  - from play-by-play to pitch-by-pitch
  - Instead of just knowing the output of the event we suddenly knew a lot about how the event happened
- Teams knew there was value in the data but, didn't know how to use it yet
- With 750,000 pitches per season, simple tables didn't tell the story well enough
- We needed visualizations to make the data accessible

Pitch Prediction

upload.wikimedia.org/wikipedia/commons/3/34/Phil_Hughes_2008.jpg

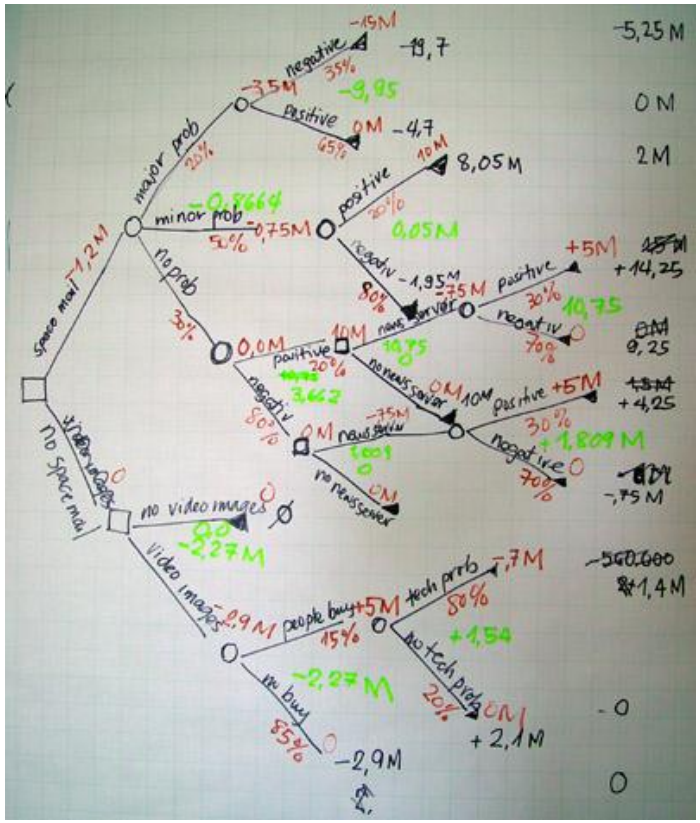
Predicting Pitches

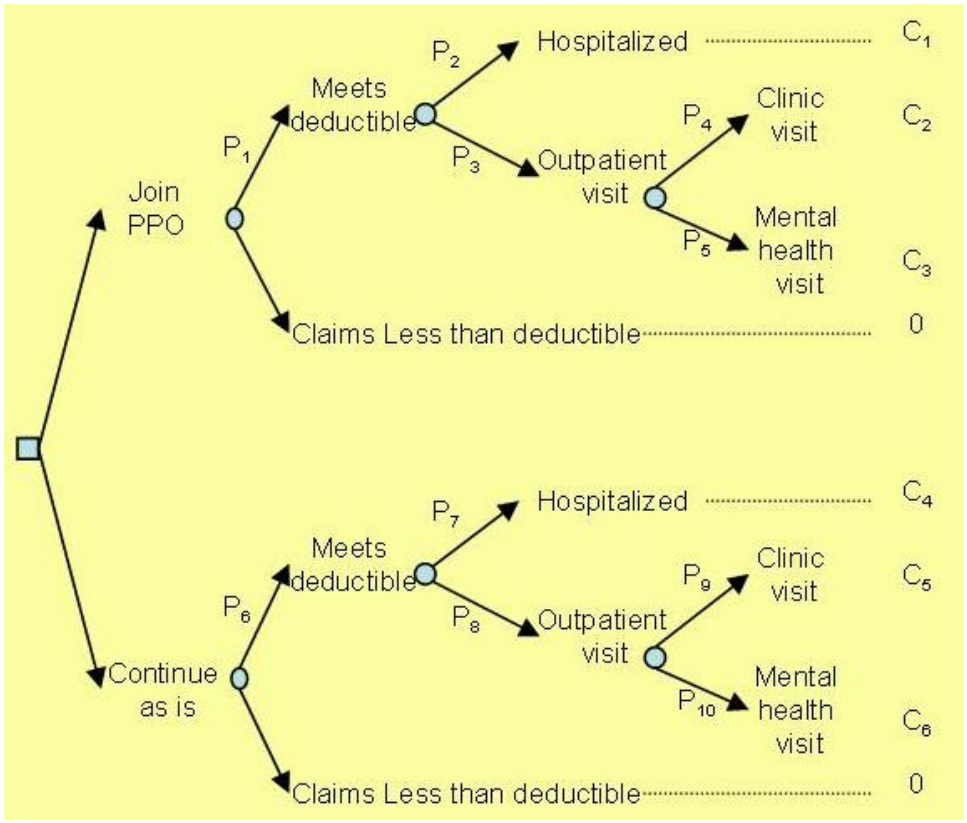
If the batter knows what kind of pitch is coming they have a big advantage. Pitchers are very aware of this and try to be less predictable. One use of PITCHf/x data is to try to predict what pitch a pitcher is going to thow:

- C.J. Wilson studies himself as much as his opponents… He says, "**Pitchers fall into traps. They get predictable with pitch sequences."** – USA Today, June 10, 2011
- "James Shields said, it was just time. Time to **stop being so predictable in his pitch selection**." – Tampa Bay Times, August 23, 2012
- "Johnny Cueto throws a lot of things at you and **he's not really predictable at all**," Brooks Raley says. – Star Tribune, Aug 12, 2012

At Bats Are Decision Trees

- We treated each at bat as a decision tree
- When pitch type A happens, 30% of the time we see another A, 20% of the time we see B, and 30% of the time we see C
- The problem is that it is too difficult to see patterns in a tree
- If we could distill the data, players could take specific details and use them in-game





http://en.wikipedia.org/wiki/Decision_tree

http://gunston.gmu.edu/healthscience/730/DecisionTrees.asp?E=0

Bloomberg SPORTS oculus

A Case Study: Aníbal Sánchez

Sánchez throws five pitches, giving him a variety of weapons to use :

- a four-seam fastball
- a sinker
- a slider
- a changeup
- a curveball

- From Wikipedia

Anibal Sanchez Delivers Ace Performance as Detroit Tigers Extend Lead in the ALCS – Fox News

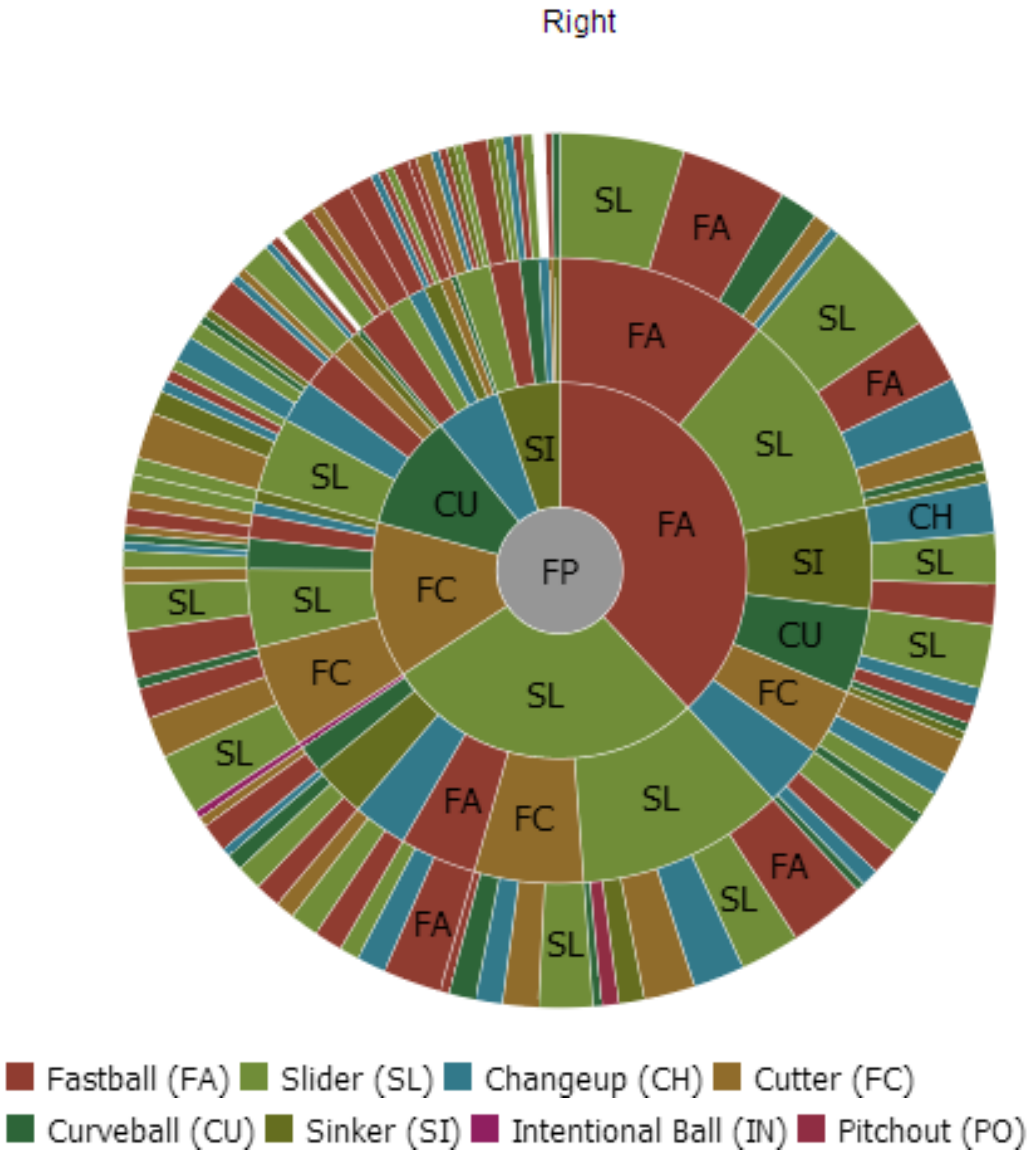
Anibal Sanchez silences Yanks – ESPN

"I try to think backwards," Sanchez said, "... I try to mix my speeds."

Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

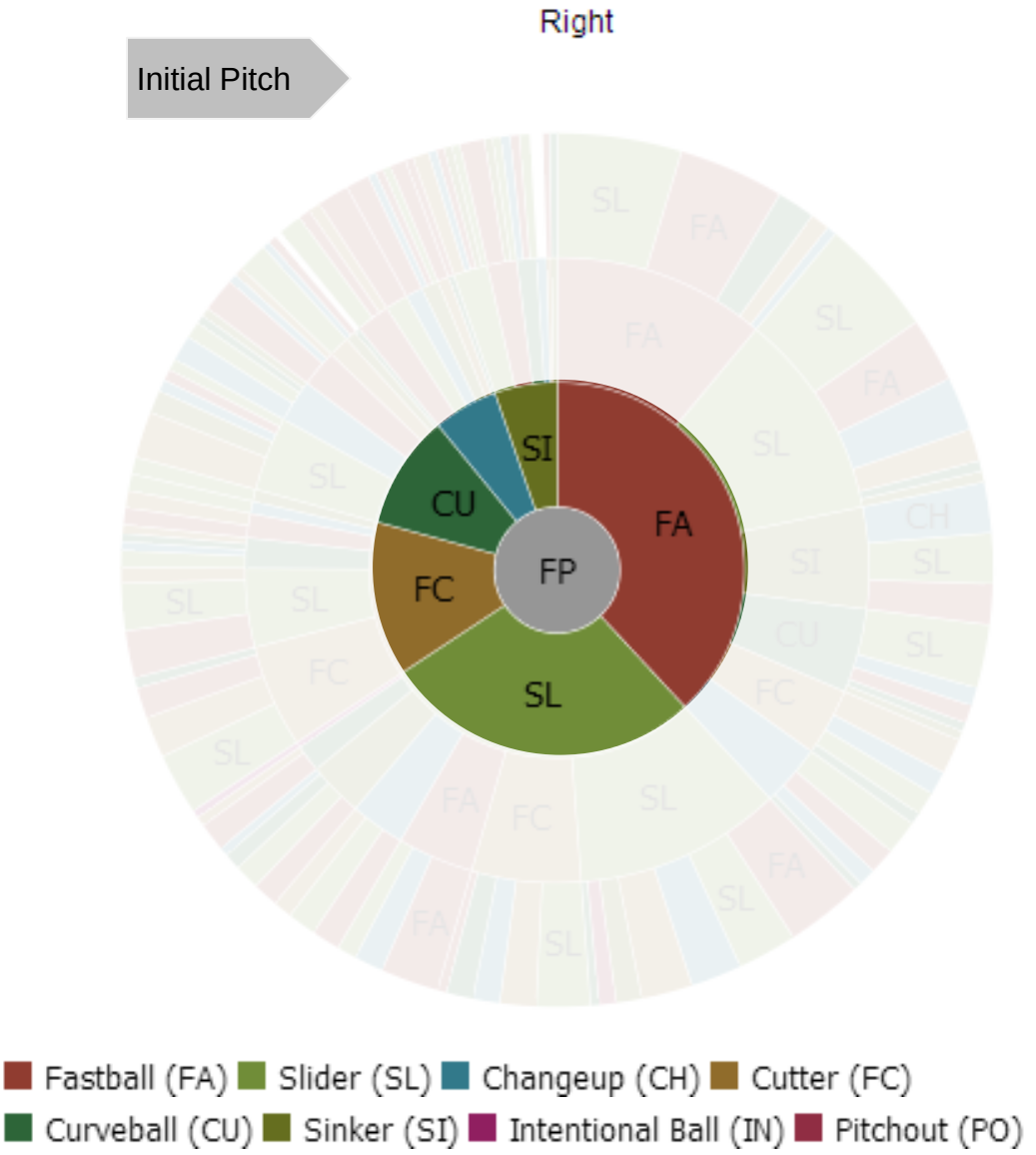
Pitches are based on the handedness of the opponent. These are all the first 3 pitches from Aníbal Sánchez to right handed hitters.



Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

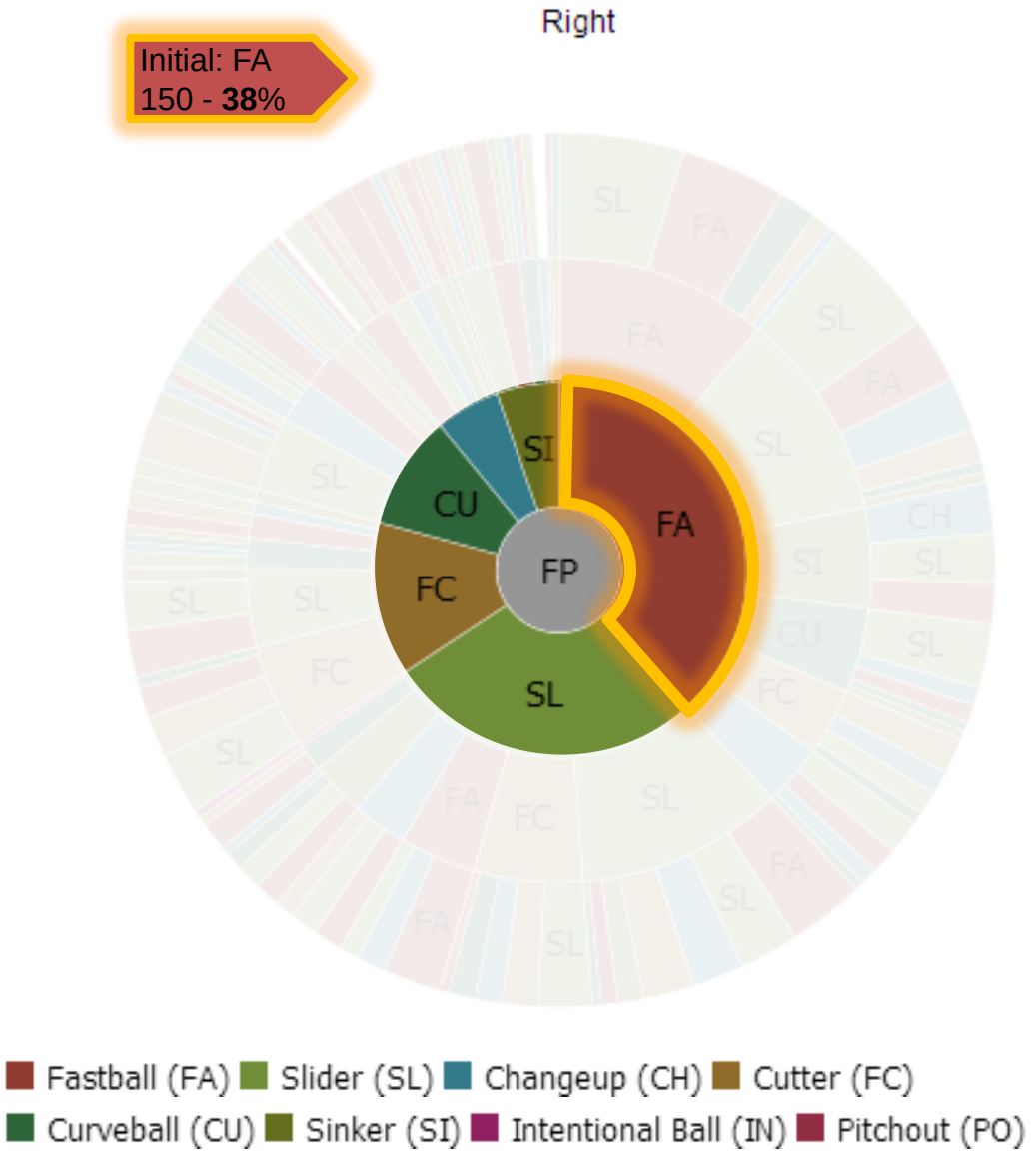
Each ring represents a successive pitch with the first ring being the first pitch of the at bat.



Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

Each ring represents a successive pitch with the first ring being the first pitch of the at bat.

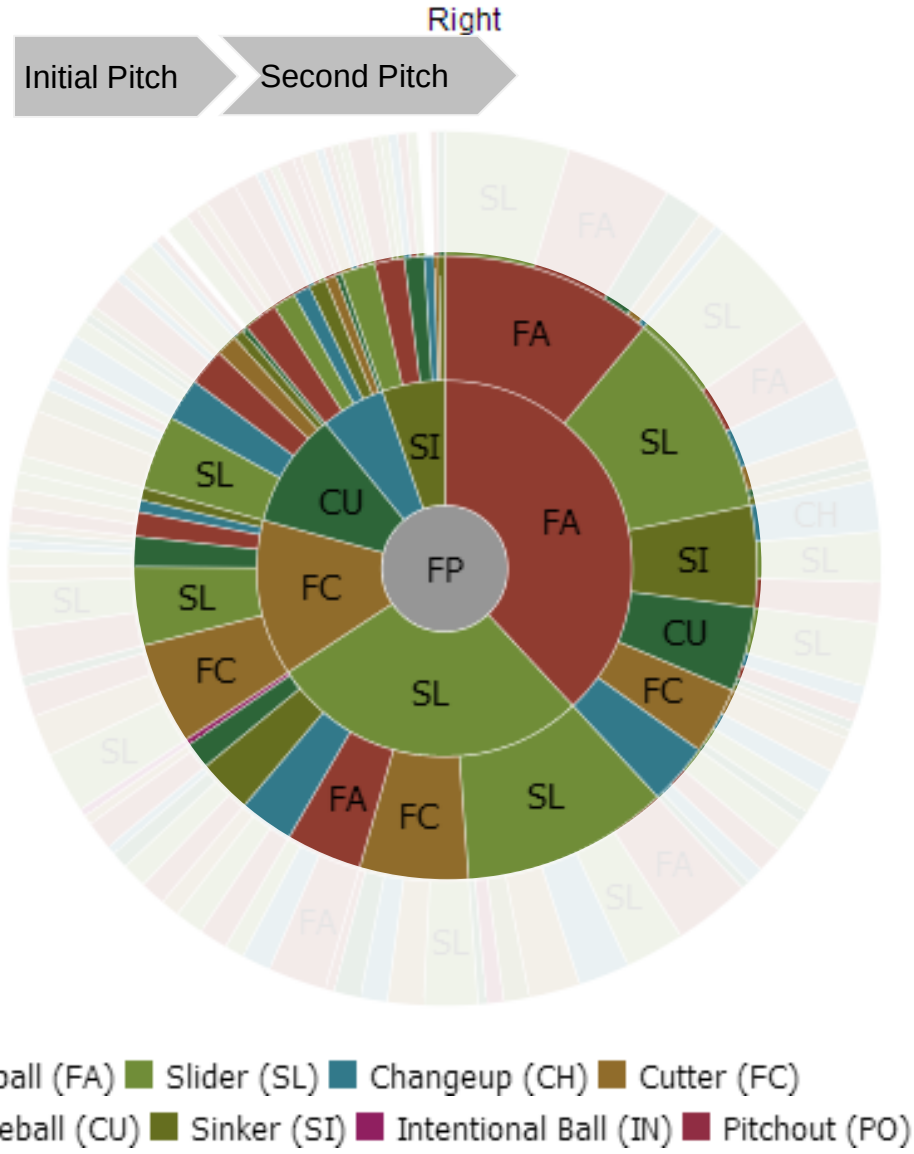


Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

Each ring represents a successive pitch with the first ring being the first pitch of the at bat.

The second ring is the second pitch, aligned to the first pitch.

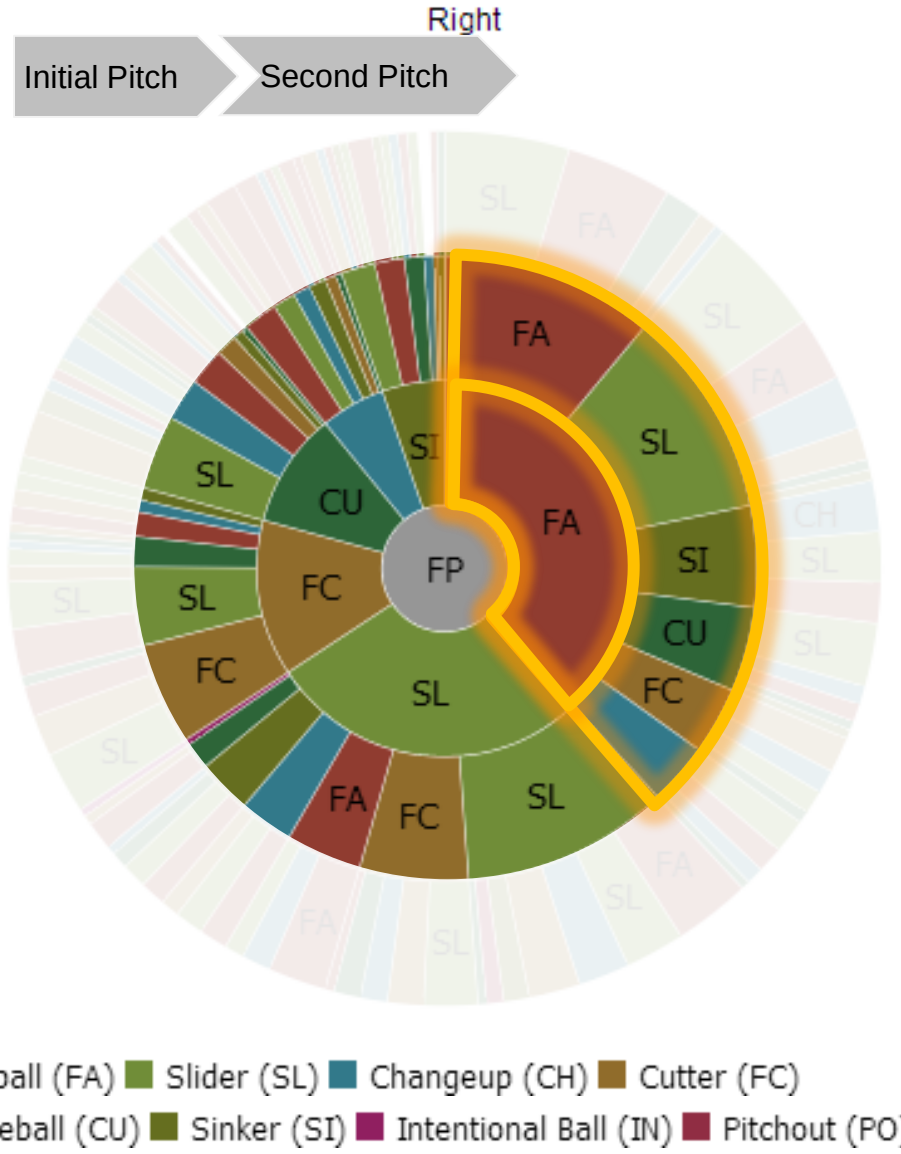


Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

Each ring represents a successive pitch with the first ring being the first pitch of the at bat.

The second ring is the second pitch, aligned to the first pitch.



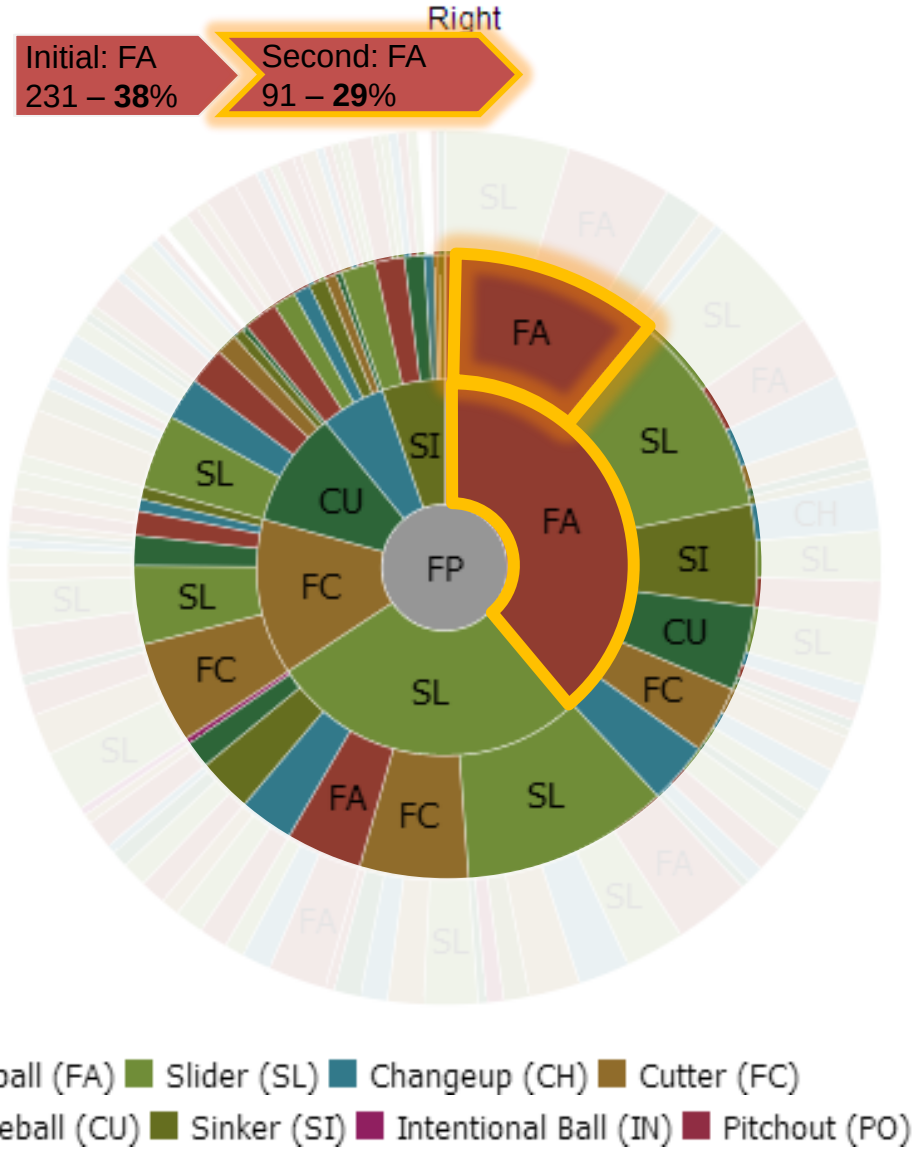
Bloomberg SPORTS oculus

Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

Each ring represents a successive pitch with the first ring being the first pitch of the at bat.

The second ring is the second pitch, aligned to the first pitch.



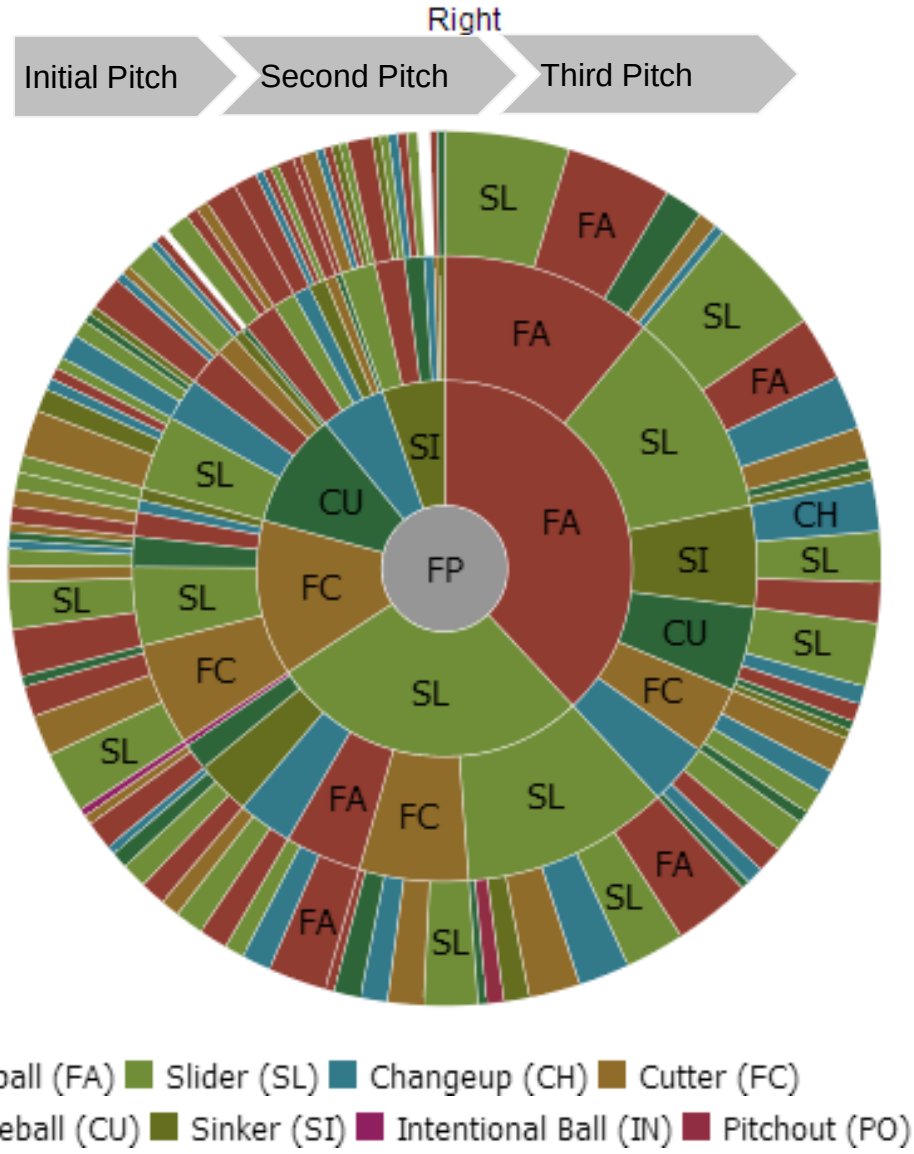
Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

Each ring represents a successive pitch with the first ring being the first pitch of the at bat.

The second ring is the second pitch, aligned to the first pitch.

And a third ring for the third pitch…

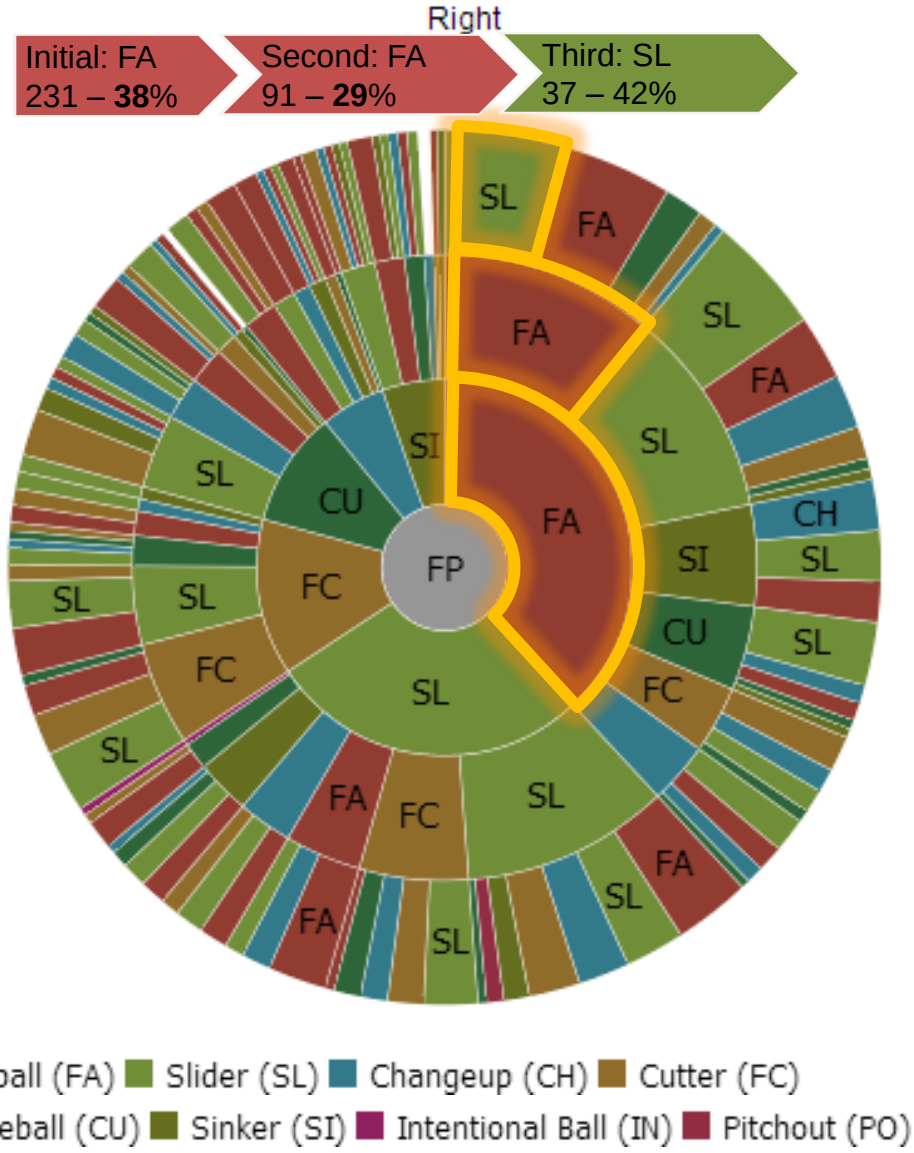


Bloomberg SPORTS oculus

Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

Against left-handed hitters, the most common pitch sequence is fastball-fastball-slider.

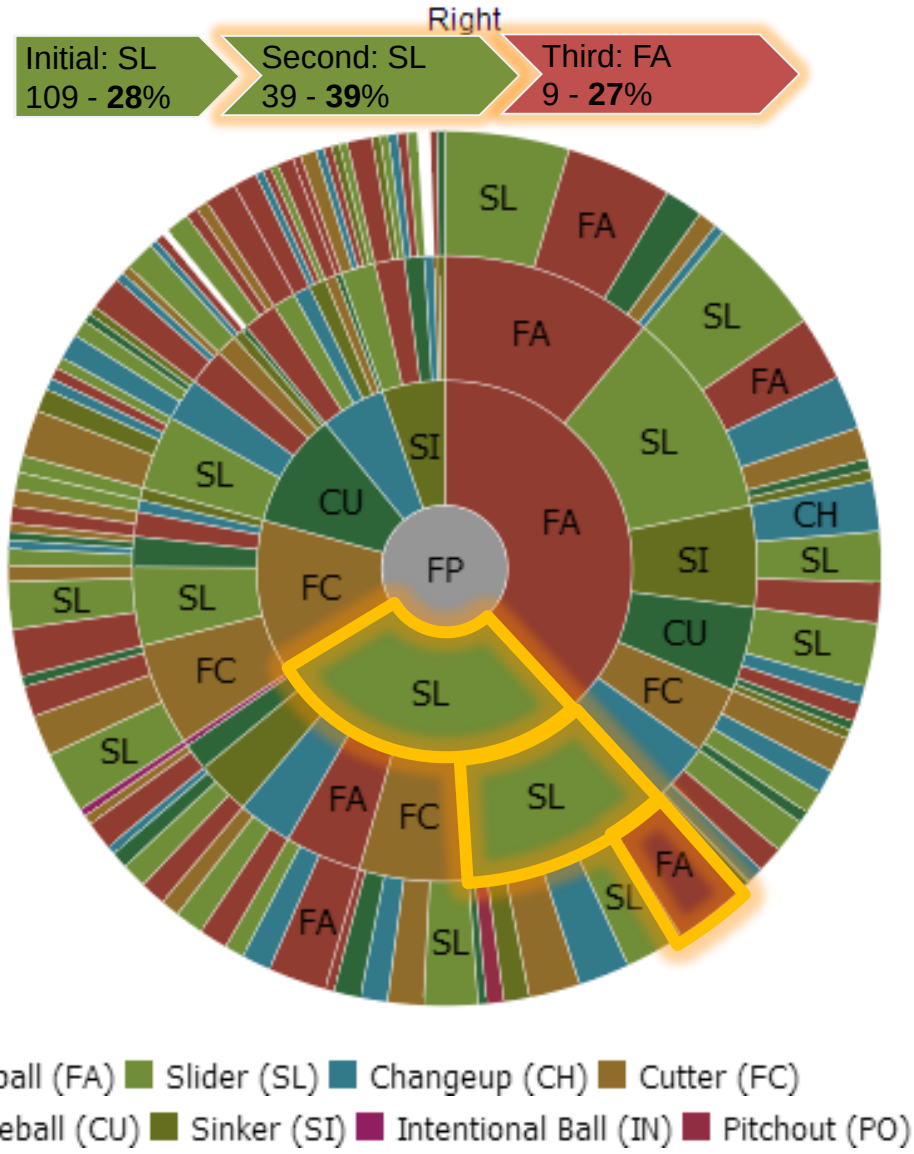


Bloomberg SPORTS oculus

Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

But, if the starting pitch isn't a fastball, he is most likely to throw a slider and most likely to continue with another slider.



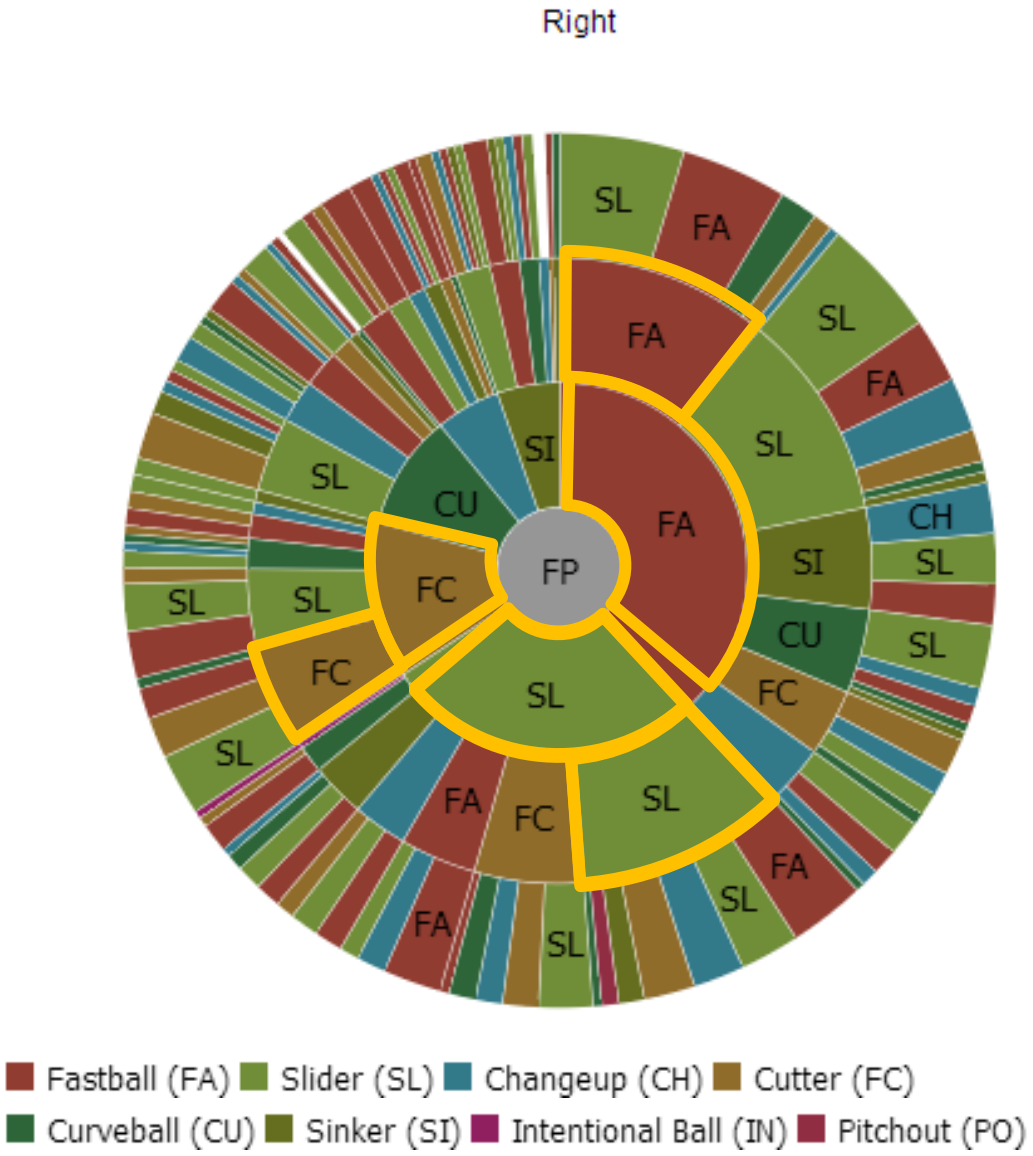
Bloomberg SPORTS oculus

Aníbal Sánchez: Detroit Pitcher

Visual Analysis of 350 pitches in 2012

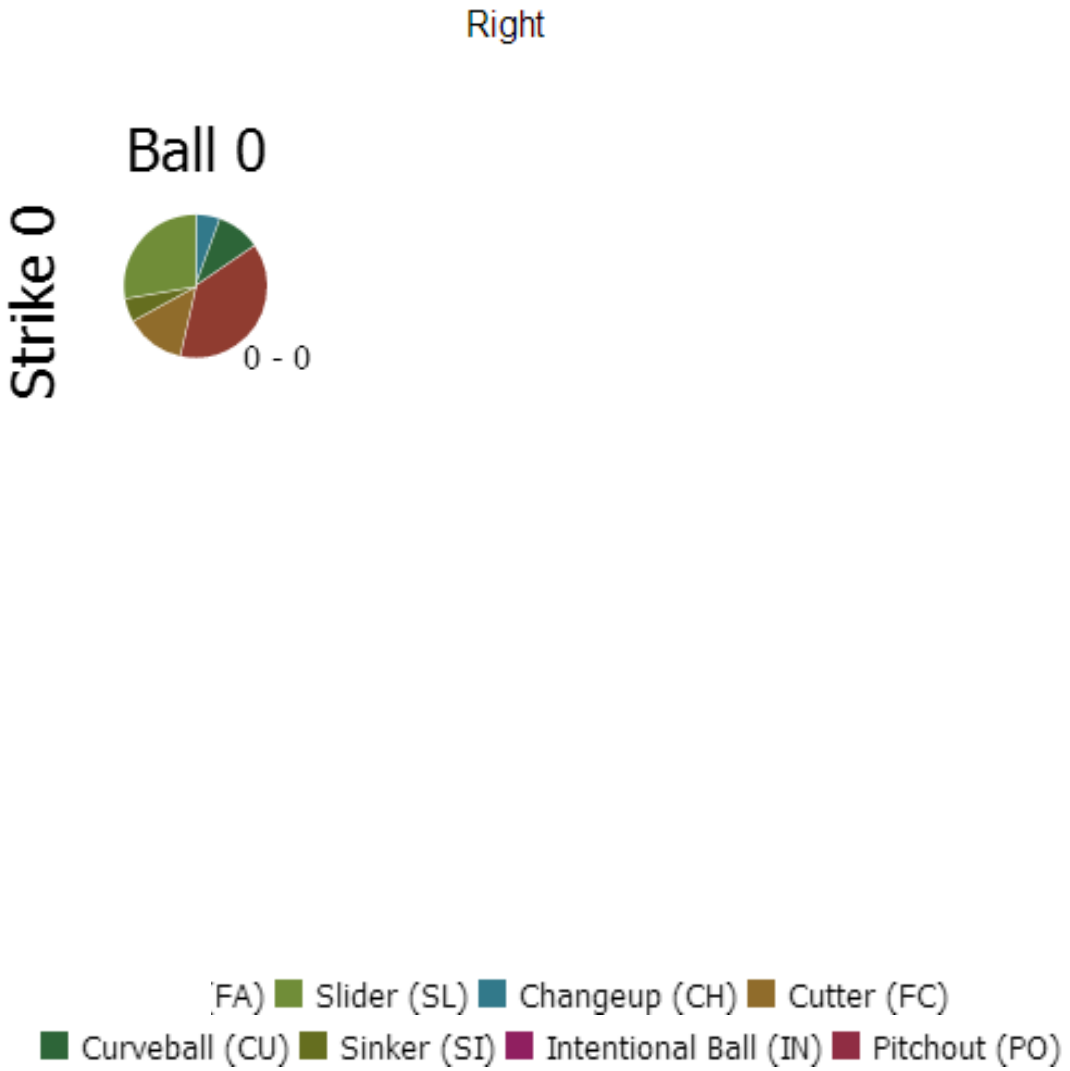
Interesting insight: The second pitch is likely to be the same as the first pitch…

Takeaway: Batters should remember that Aníbal Sánchez is likely to follow his first pitch with the same pitch.



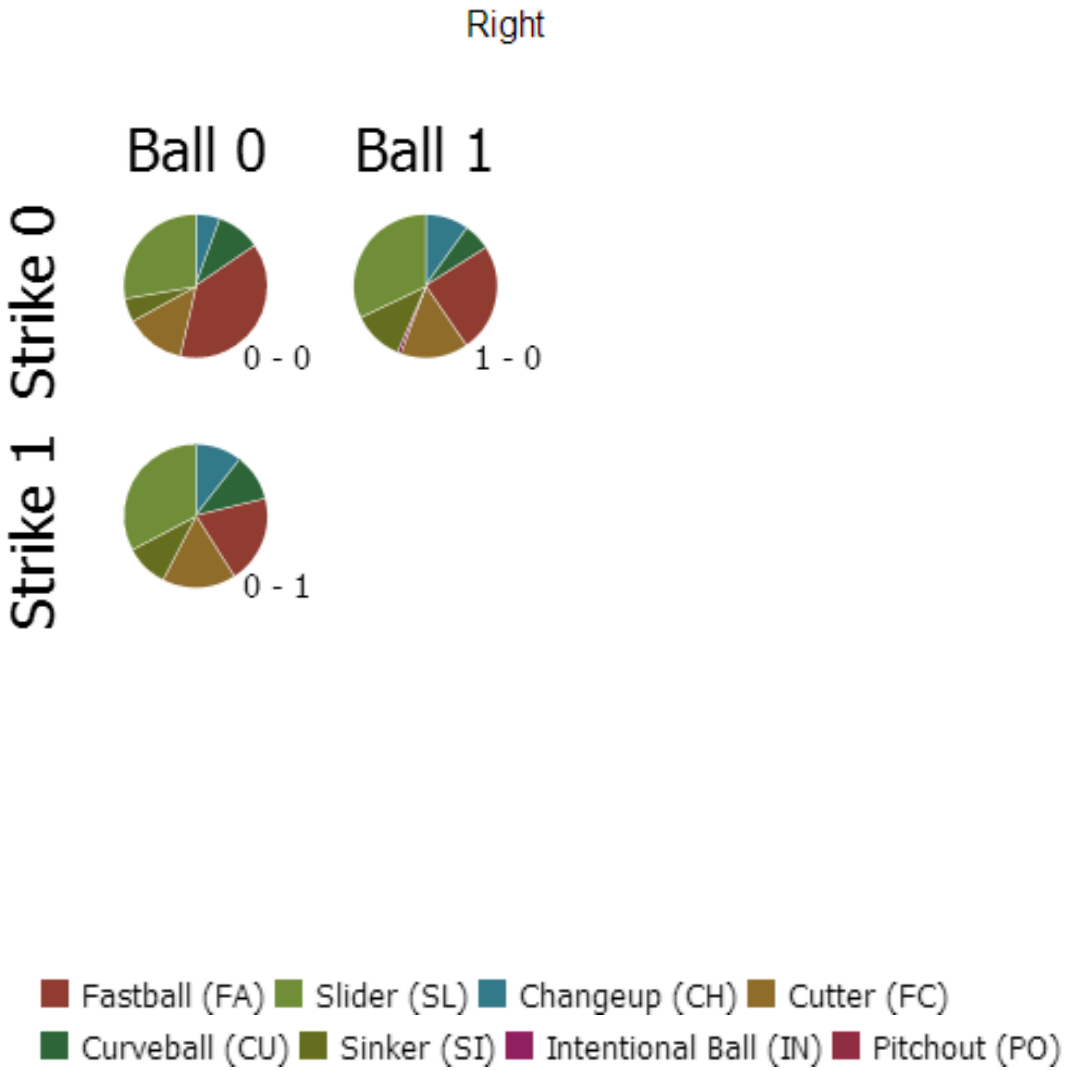
Aníbal Sánchez Pitches When Ahead / Behind in Count

- Count matters too
- Against right-handed hitters, Aníbal favors his Fastball initially



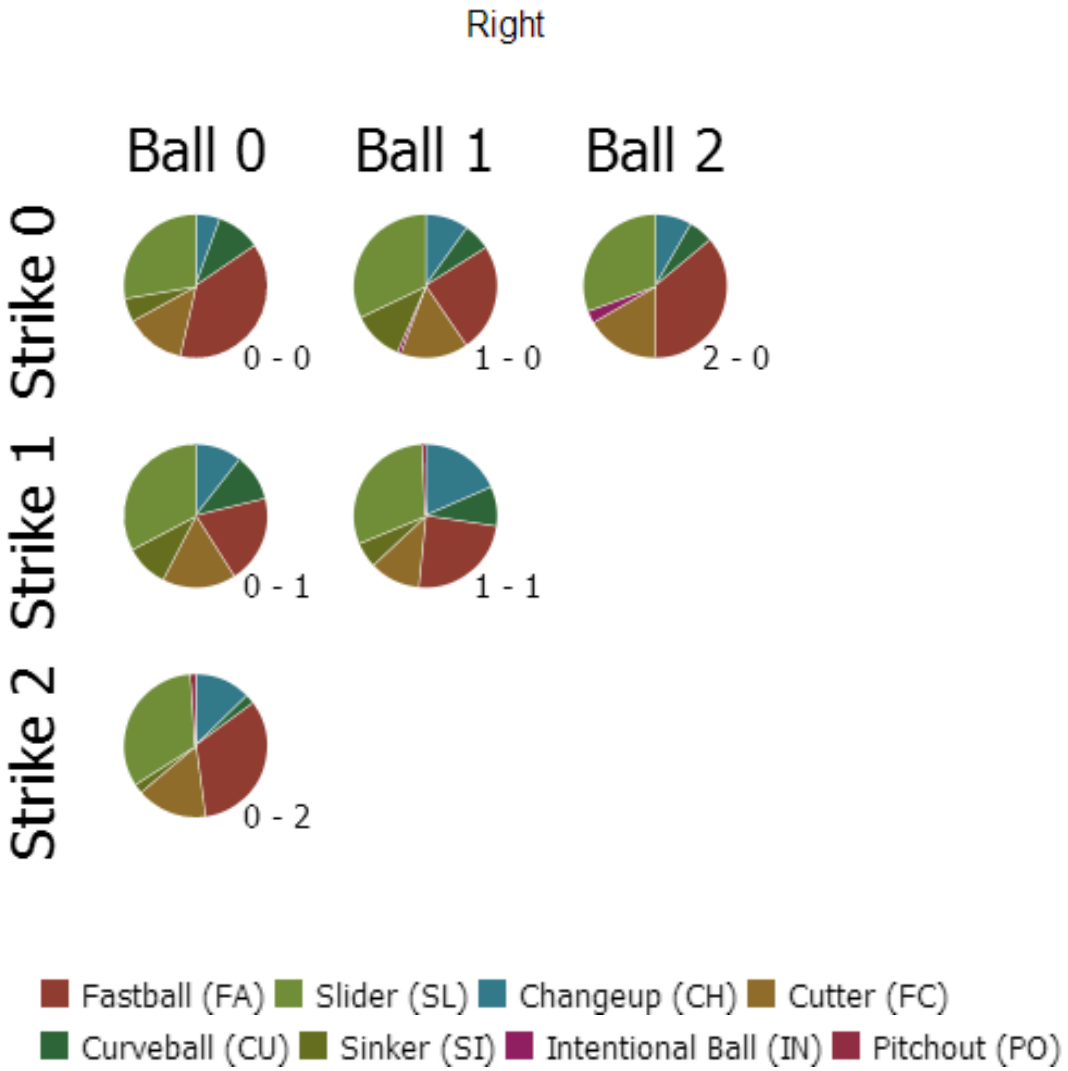
Aníbal Sánchez Pitches When Ahead / Behind in Count

- Count matters too
- Against left-handed hitters, Aníbal favors his Fastball initially
- Then Slider for the second pitch



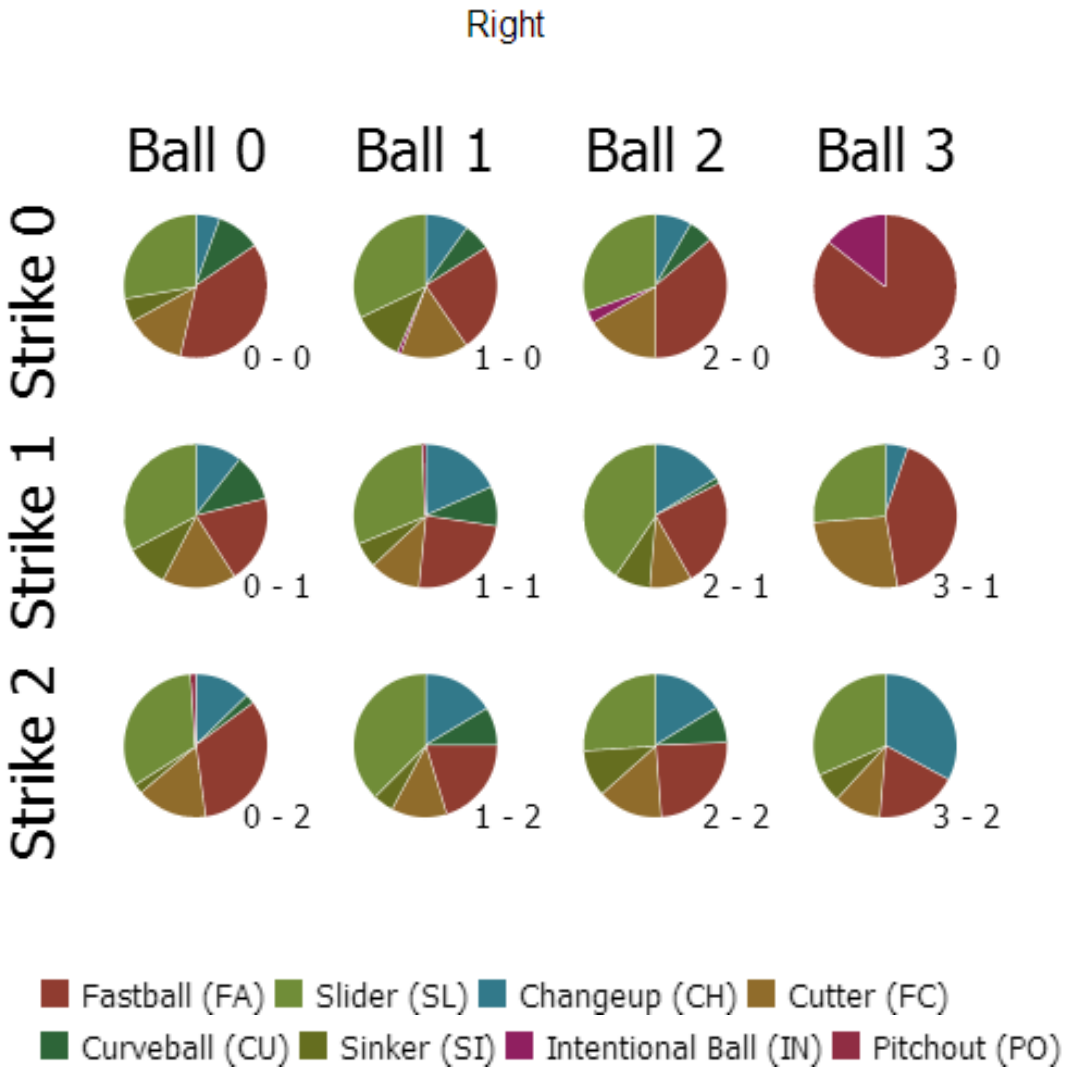
Aníbal Sánchez Pitches When Ahead / Behind in Count

- Count matters too
- Against left-handed hitters, Aníbal favors his Fastball initially
- Then Slider for the second pitch



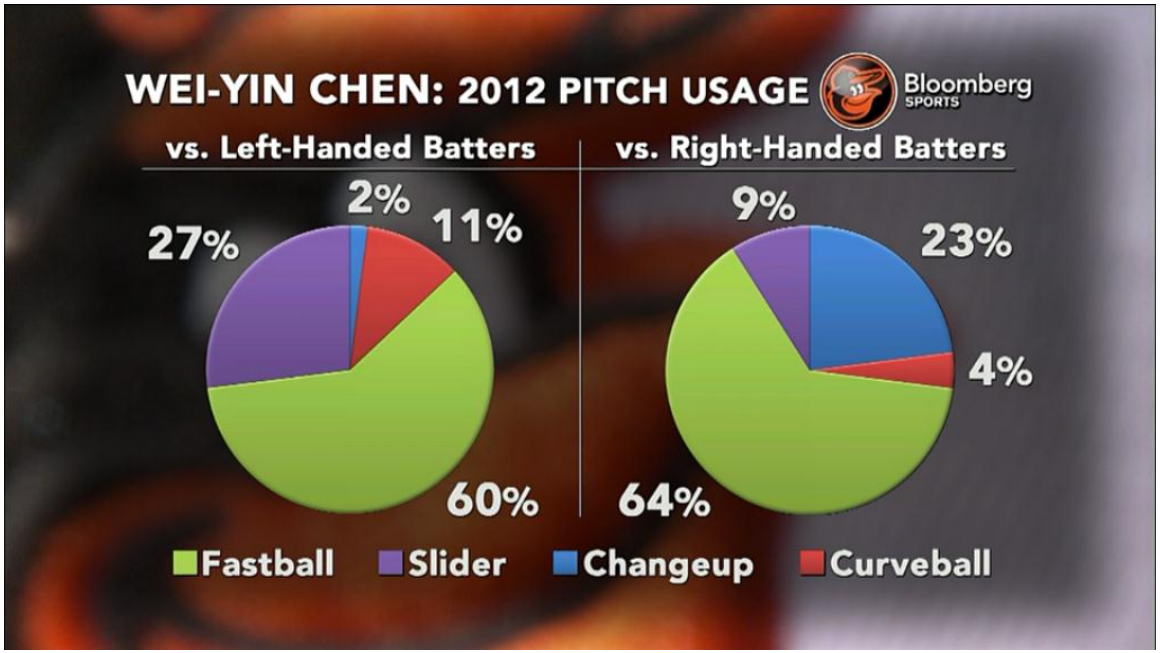
Aníbal Sánchez Pitches When Ahead / Behind in Count

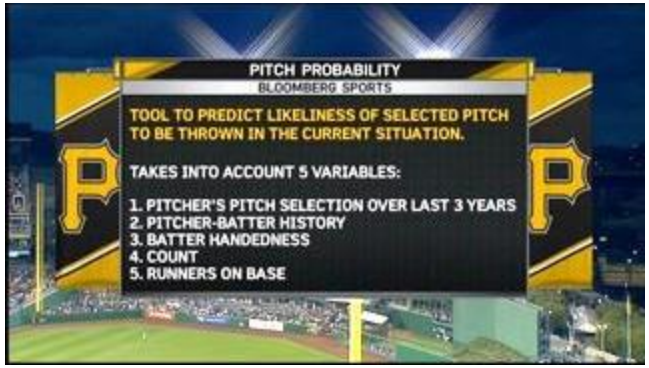
- Count matters too
- Against left-handed hitters, Aníbal favors his Fastball initially
- Then slider for the second pitch
- Shifts to Fastballs when behind in the count
- Likely to use the Changeup at full count.



Predicting Pitches – for a TV Audience

- A chart on TV only lasts for 10-30 seconds
- Must be immediately understandable – no explanation about the graphic, must be able to be able to immediately discuss content
- Therefore, for TV audience, really simplified data:





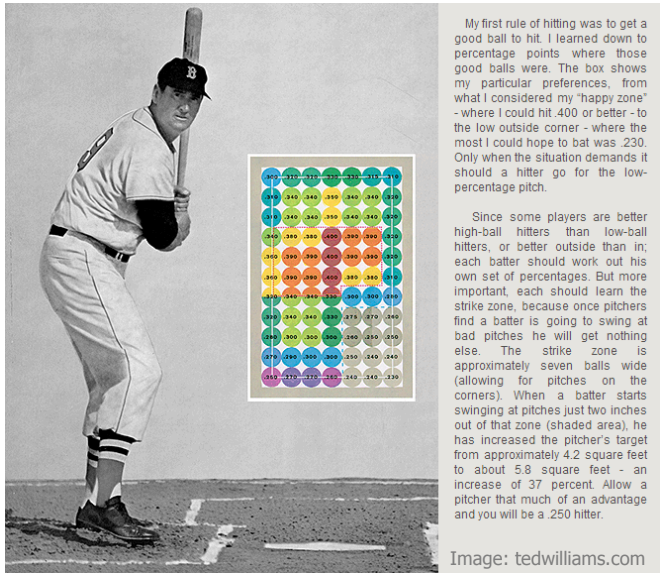
Bloomberg SPORTS oculus

Batting Analysis

Batter Zones

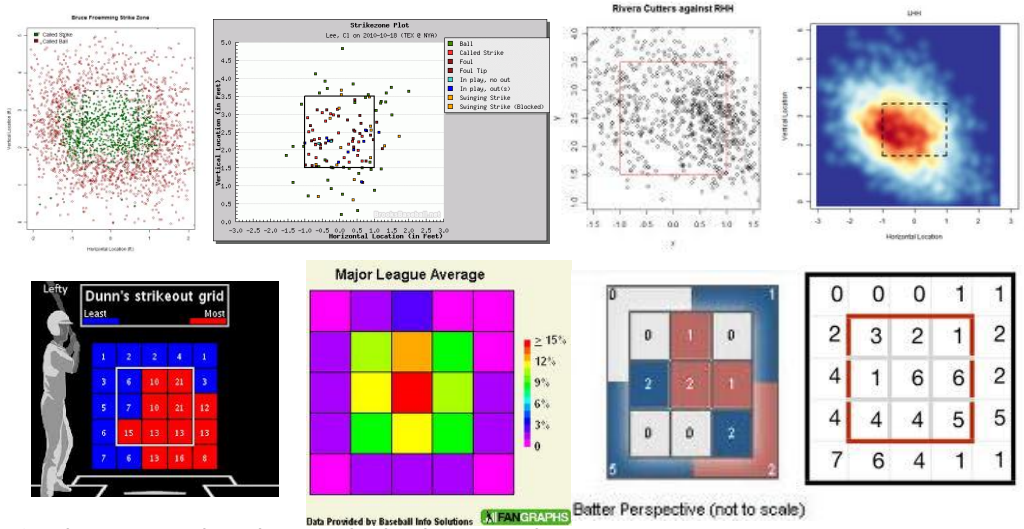
1. Long known

- Where the pitch goes through the strike zone makes a difference to batter's performance.
- E.g. Ted Williams Strike Zone (1968) ➔



2. Now Pitchf/x data

- One-offs analyses require effort ➔



Google image search: strike zone plot / strike zone grid

3. BUT!!!

- Coaches and front office need ease-of-use
- Players have limited time

Pitches

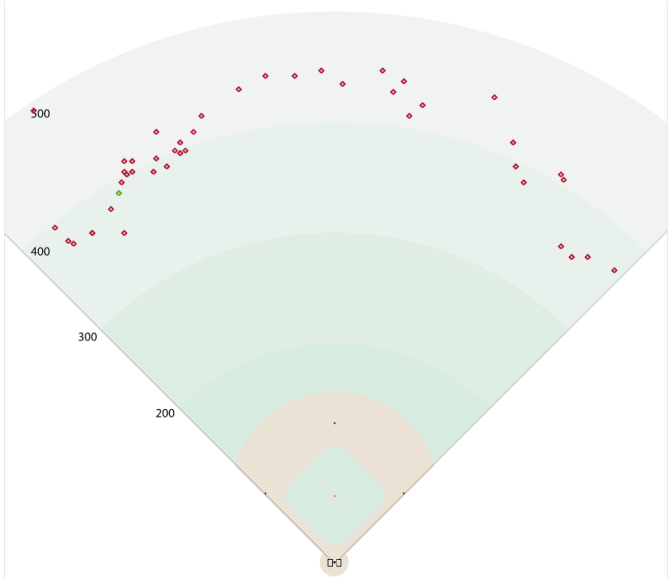
- Allows users search through the full history of PitchFX data
- Quick definitions
- **AVG:** Batting Average. It measures how often you successfully get onto base with a hit. Higher is better. The 2012 league average was .255 and 1.000 is the highest possible value.
- **OBP:** On-base percentage. It measures how often you get on base and similar to AVG except that you also get credit for walks and being hit by a pitch. Higher is better. The 2012 league average was .319 and 1.000 is the highest possible value.
- **SLG:** Slugging percentage. It measures the overall power of your hitting and similar to AVG except that the hits are weighted by how many bases they earned. Singles get one base, doubles get two bases and are worth twice as much, etc. Higher is better. The 2012 league average was .405 and 4.000 is the highest possible value and would require that every at bat be a Homerun.
- **OPS:** OBP + SLG. Its an overall measure of the hitter. Higher is better. The 2012 league average was .724 and 5.000 is the highest possible value and would require that every at bat be a homerun.
- **Pull Hitter:** A hitter who usually hits the balls to the side of the plate from which he bats. For example, if you are right-handed then you bat standing the left side of the plate. If you usually hit the ball into left field you're a pull hitter as you're pulling it left.

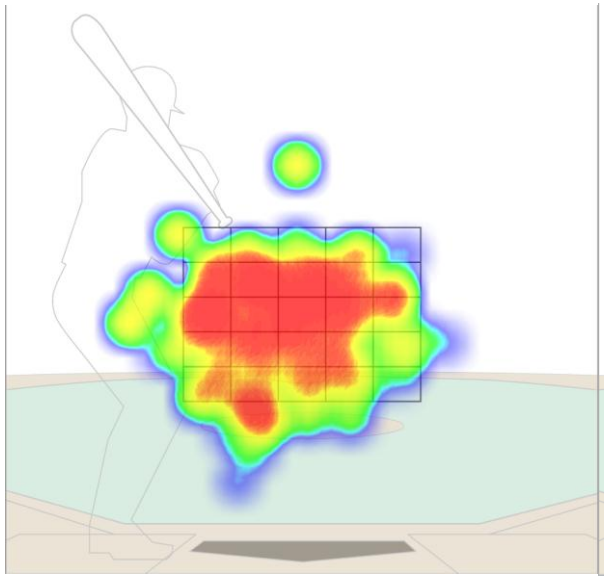
PitchFX Analysis

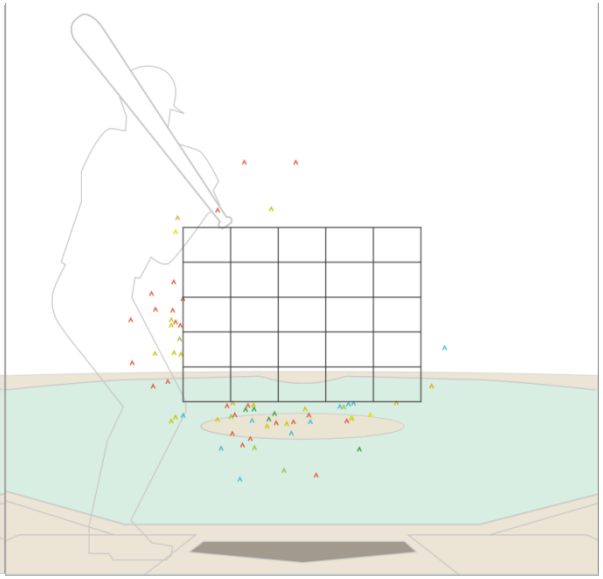
Name	Bats	AVG	OBP	SLG	OPS
Miguel Cabrera	Right	.330	.393	.606	.999
Prince Fielder	Left	.313	.412	.528	.940
Pablo Sandoval	Switch	.283	.342	.447	.789
Mike Trout	Right	.326	.399	.564	.963

- All good or great hitters and all above league average. Sandoval is not in the same league as the others but, you'll see why we included him
- The problem is that the numbers we have right now only tell us that they're good and that as a pitcher or a fielder you need a plan when facing them. How do we get that plan?

Miguel Cabrera (RH, Bats on Left Side)





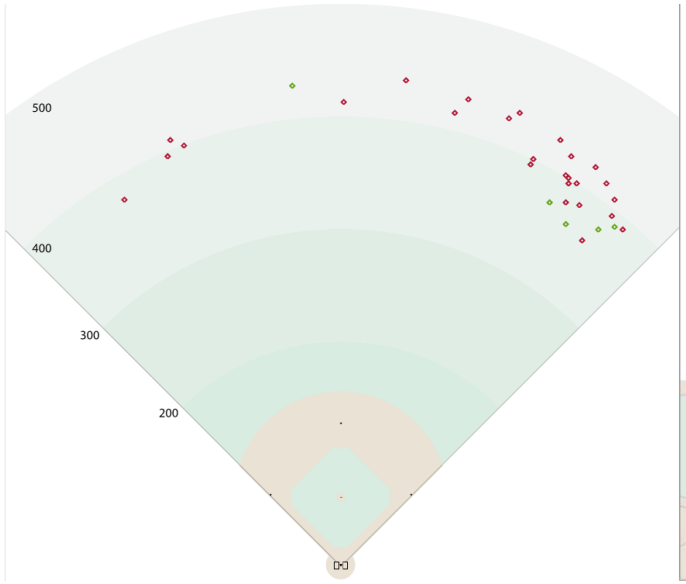


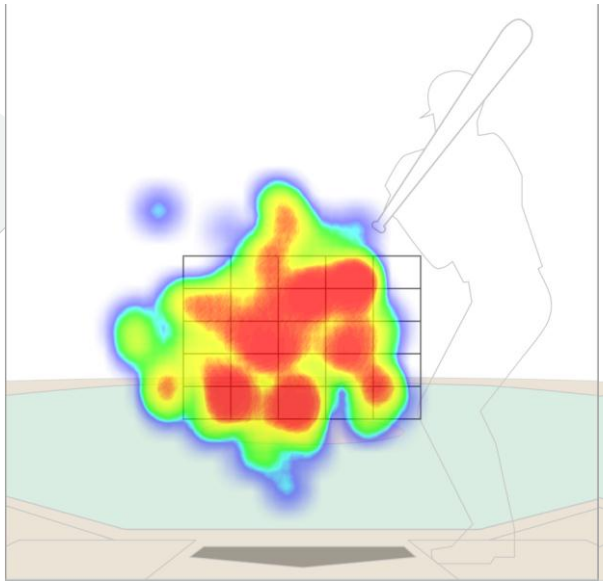
2012 Regular Season
Miguel Cabrera spreads his homeruns so you need to worry about him going deep in all directions meaning fielders should play deep.

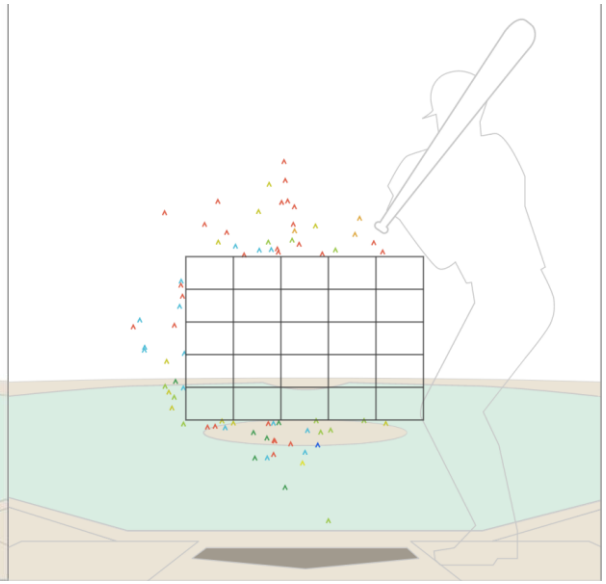
2012 Regular Season
Miguel Cabrera's heatmap shows that he is hot all over the strike zone leaving no real safe spot for the pitcher. If he is taking less hits down and away, its probably because he is walking on those pitches.

2011 – 2012 Regular Season
66 hits outside of the strike zone over the last 2 years with few of those hits away from him.

Prince Fielder (LH, Bats on Right Side)





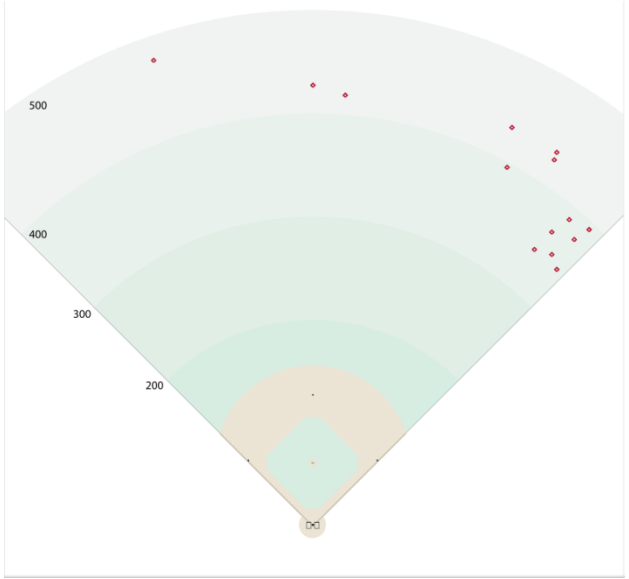


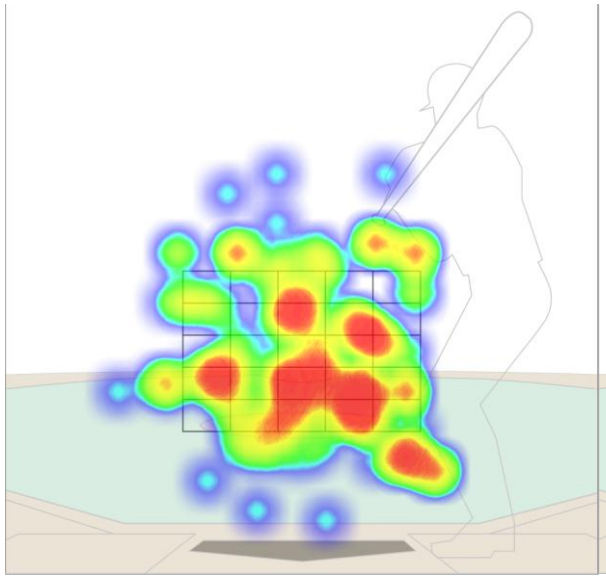
2012 Regular Season
Prince Fielder pulls his homeruns. Batting on the right-side of the plate, most of his homeruns go to right-field. Fielders should play deep in right.

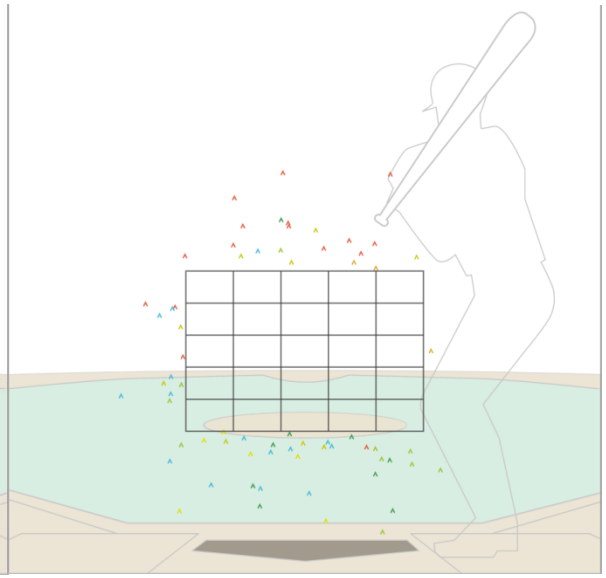
2012 Regular Season
A heatmap of Prince Fielder's SLG. As you can see he is hot all over the strike zone and is especially dangerous up and in.

2011 – 2012 Regular Season
75 hits outside of the strike zone over the last two years. The hits are all around the zone except for on the inside which is especially interesting when you consider how hot he is up and in.

Pablo Sandoval (LH, Bats on Right Side)





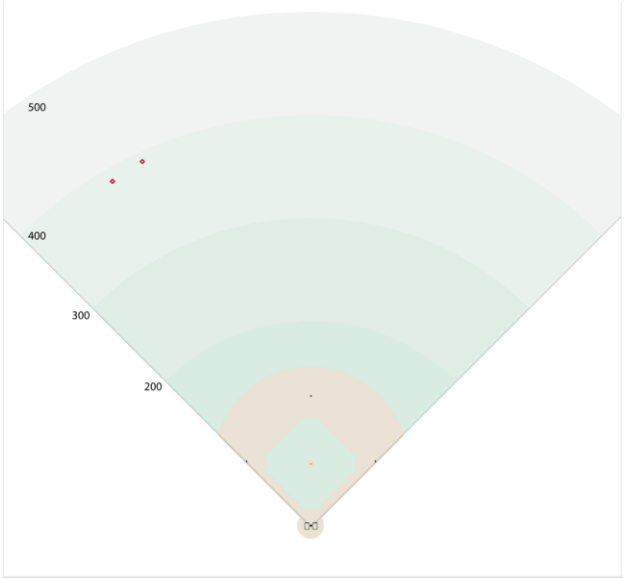


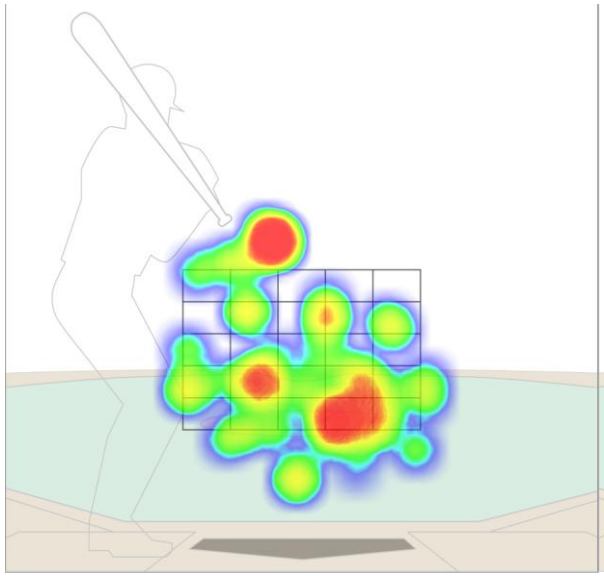
2012 Regular Season
Batting left-handed, Pablo pulls his homeruns. He isn't much of a homerun threat in general but, you might want to play deeper against him in right-field.

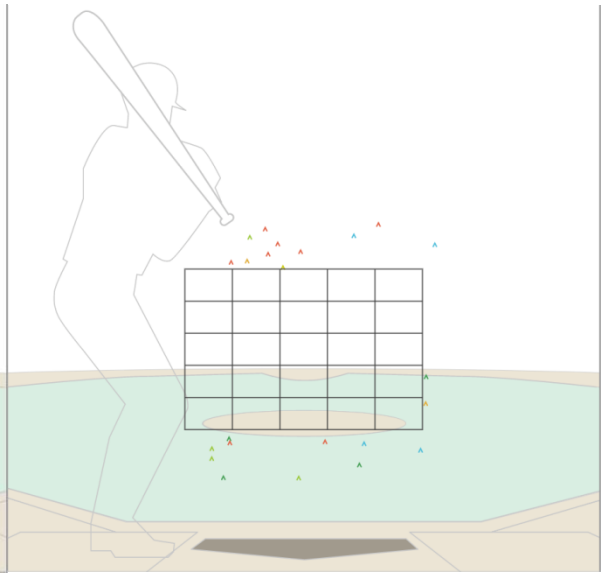
2012 Regular Season
Batting left-handed, Pablo isn't very hot in any one area. It's all over the place and not very predictable.

2011 – 2012 Regular Season
68 hits outside of the strike zone over the last two years. The hits are all over and in some cases very far from the zone. Most hitters would not swing at those pitches.

Pablo Sandoval (RH, Bats on Left Side)





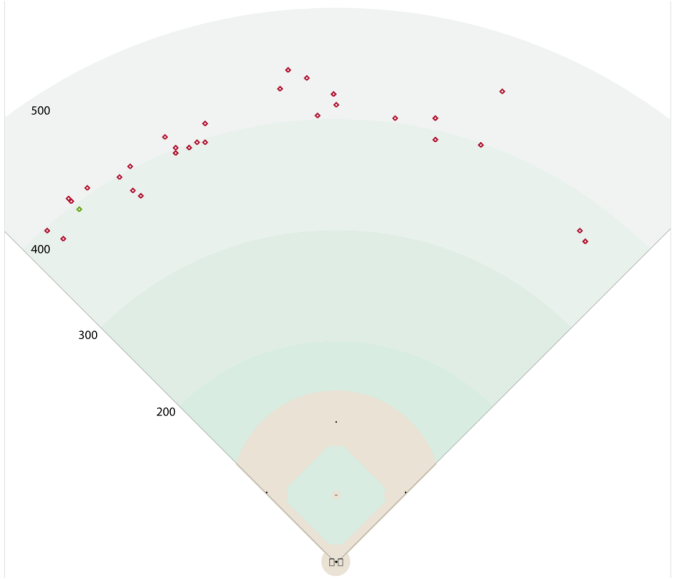


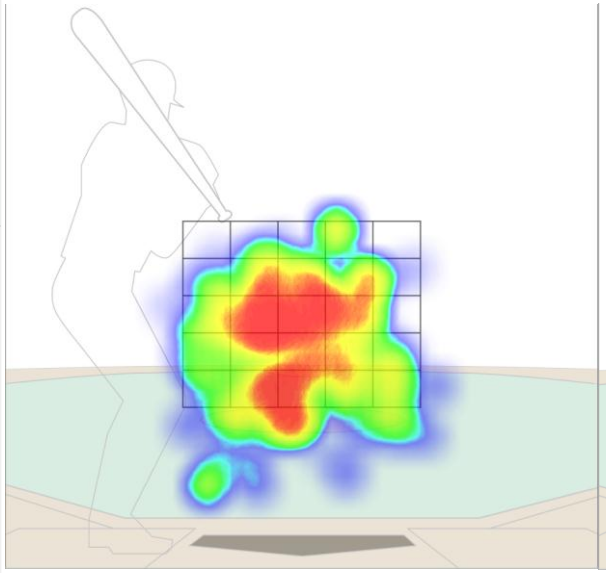
2012 Regular Season
Almost no homeruns to speak of batting right-handed. Both did pull to the left.

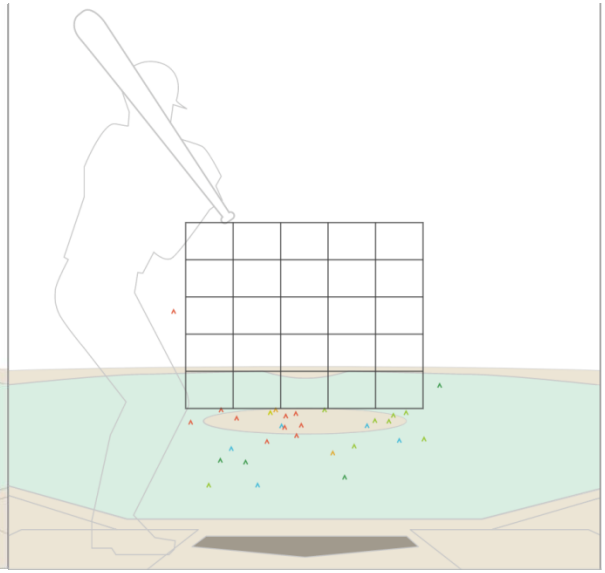
2012 Regular Season
Batting right-handed, Pablo isn't very hot in any one area. It's all over the place and not very predictable. Surprisingly consistent with his left-handed heatmap in that he is hot to the bottom-right.

2011 – 2012 Regular Season
23 hits outside the strike zone over the last two years. The hits are either above or below the strike zone when Pablo is batting right-handed.

Mike Trout (RH, Bats on Left Side)







2012 Regular Season
Mike Trout has a slight pull to his homeruns but, can be dangerous anywhere. Definitely play deep in left-field.

2012 Regular Season
Mike Trout is very hot down the middle. This suggests an extreme amount of patience and a batter with a keen eye. He probably waits for fastballs or breaking balls that don't break.

2011 – 2012 Regular Season
40 hits outside of the strike zone over the last two years. The hits are mostly down in the zone.

PitchFX Analysis Complete

Name	Pull?	Heat?	Outside Hit %
Miguel Cabrera	No	Nowhere is safe	16%
Prince Fielder	Yes	Most dangerous inside	21%
Pablo Sandoval	Yes	Inconsistent, overlap batting left and right	37%
Mike Trout	Slight	Patient, dangerous down the middle	19%

- All different hitters and you would struggle to find anything consistent about them other than them all being good
- Cabrera, Fielder, and Trout are hitting balls to different parts of the park and taking them in different places in the zone. The only consistency is that most hits are inside the zone.
- Sandoval slugs inconsistently across the zone, pulls using either hand, and has more than a third of hits off of balls outside of the zone
- AVG/OBP/SLG/OPS alone only tells us if someone is good or great – it cant tell us how or why
- Thanks to PitchFX we can see that these players are unique, that they're getting the job done differently, they require different strategies, and that you can be good by being unusual

Live Demo

Lots of interactive data

- 4,035,089 pitches
- 2,057,089 videos
- 17 filters
- 3 data sets
- In-memory cache with indices across multiple fields

Lots of visual techniques

- Multi-variate high-performance glyphs
  e.g. color = pitch type
  shape = pitch result
- Heatmaps (based on heatmap.js)
- Linked interaction, workflow to videos
- Significant iterations over four years working with Oculus

Does anyone have a favorite player?

Front Office Baseball

- **The goal is to make fantasy baseball easy**
- Lots of workflows that needed to be streamlined
- Fantasy baseball depends more on the overall numbers and less on the specific pitch-by-pitch details
- Ironically, fantasy baseball required us to build a projections engine
- The staff responsible for projections are the same kind of staff that would have that responsibility at an MLB team
- **We had 8 of the top 10 teams by record correct in 2011** – the Rangers and Cardinals were ranked exactly correctly

Bloomberg SPORTS oculus

Today and Beyond

- MLB now measures the ball off the bat (HITf/x) and is working on a system which will measure all movement on the field (FIELDf/x)
- PITCHf/x helped start a revolution in sabermetrics and HITf/x and FIELDf/x could lead to more interesting discoveries about the game. Defense is one area which is still very difficult to measure statistically and FIELDf/x data is expected to be something of a holy grail for baseball defensive statistics
- In other words, the data is still growing… fast

In Conclusion

- Baseball has a lot of data
- Managing this data is difficult
- Statistical concepts can be used in conjunction with rich visuals to make the data consumable