SIMULATING MLB AT-BATS

THE GOAL

Build a tool that takes in a statistical picture of a hitter and a pitcher, along with some environmental factors, and returns a distribution of possible outcomes of that interaction.

BACKGROUND: APPROACH

- Idea of an interaction
- What statistics can be used to judge players?
- Can players be described by a statistical matrix?

https://tht.fangraphs.com/can-we-predict-hitters-change-based-on-how-pitchers-approach-them/

BACKGROUND: THE DATA

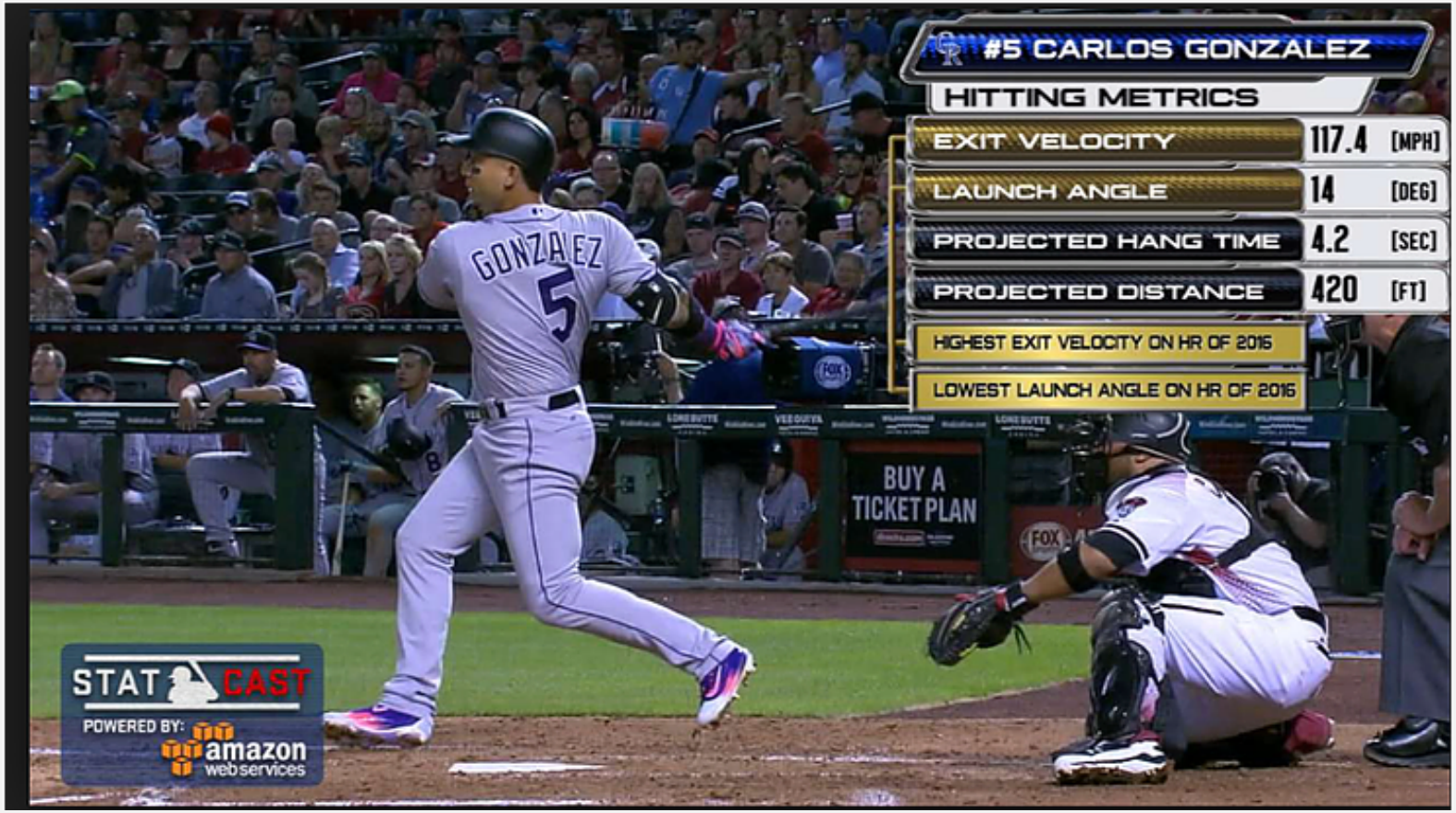
- Extensive player data
- Play by play with play outcomes over a large sample sheet
- Good news: Baseball data is everywhere!
- Retrosheet, MLB.com, Baseball Savant
- Formatting, extracting had difficult formatting, R required
- Working with text data difficult
- The community can help
- About 900k events

```
play,1,1,abrej003,32,*B*BFFBX,E6/FO/G.2-H(E6/TH)(UR)(NR);1-2
play,1,1,castw002,32,CBF*BFBB,W.2-3;1-2
com,"Mound Visit"
play,1,1,moncy001,22,FBFBFX,D8/F+.3-H(UR);2-H;1-3
play,1,1,jimee001,12,CSBS,K
play,1,1,alony001,01,CX,3/G
play,1,1,rondj002,11,FBX,9/L+
play,2,0,beckt001,01,CX,53/G
play,2,0,narvo001,11,CBX,S8/L
play,2,0,healr001,22,BCFBS,K
play,2,0,voged001,02,CCX,6/P
play,2,1,engea001,10,BX,D7/L+
play,2,1,garcl004,32,BBBCLX,S1/BG.2-3
play,2,1,andet001,11,*BFB,SB2
play,2,1,andet001,21,*BFB.X,S8/G.3-H;2-H
play,2,1,abrej003,12,CBSX,S9/G.1-3
play,2,1,castw002,00,X,64(1)3/GDP.3-H(NR)
play,2,1,moncy001,02,FFS,K
play,3,0,gordd002,30,BBBB,W
com,"Mound Visit"
play,3,0,smitm007,31,BBB11CC,SB2
play,3,0,smitm007,32,BBB11CC.B,W
com,"Mound Visit"
play,3,0,hanim001,11,BCX,8/F.2-3;1-2
play,3,0,santd002,00,X,63/G.3-H;2-3
play,3,0,brucj001,32,C*BBF*BC,K
play,3,1,jimee001,01,CX,63/G
```

	pitcher_name	batter_name	inning_side	inning	stand	p_throws	event	date	season	away_team	home_team	pitcher_team
0	Johnny Cueto	Mike Trout	top	1	R	R	Strikeout	2013-04-01	2013	ana	cin	cin
1	Johnny Cueto	Erick Aybar	top	1	L	R	Lineout	2013-04-01	2013	ana	cin	cin
2	Johnny Cueto	Albert Pujols	top	1	R	R	Groundout	2013-04-01	2013	ana	cin	cin
3	Jered Weaver	Shin-Soo Choo	bottom	1	L	R	Hit By Pitch	2013-04-01	2013	ana	cin	ana
4	Jered Weaver	Brandon Phillips	bottom	1	R	R	Strikeout	2013-04-01	2013	ana	cin	ana
5	Jered Weaver	Joey Votto	bottom	1	L	R	Flyout	2013-04-01	2013	ana	cin	ana

BACKGROUND: NEW AGE V OLDER DATA

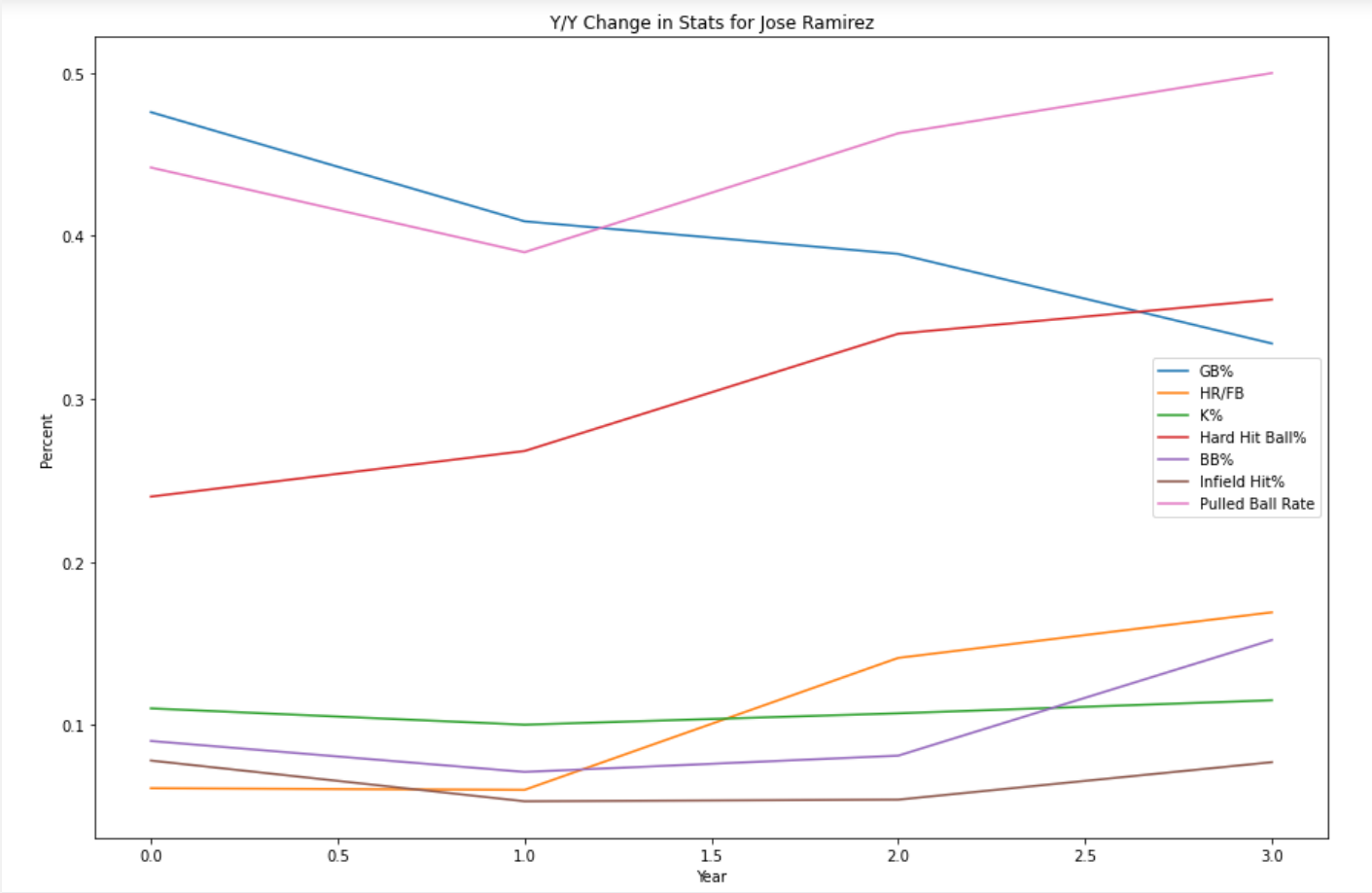
- What player data is necessary to judge players?
- Outcomes based: Fangraphs
- Physical-based: Statcast
- When does the data stabilize?
- What sample size is required



https://www.coverfoursports.com/post/working-with-data-getting-statcast-data-in-r

EDA: SIGNIFICANT DATA

- Pure bias variance trade-off
- Informative input vector vs. future use
- What can be included that is also reproducible?
- Accounting for year over year baseline change



EDA: FEATURES TO CONSIDER

- Use correlation to see what features are predictive for each outcome
- Use previous studies done by analytics websites like Fangraphs, Hardball Times

```
----------------------------------------
LD baseline: 0    0.936942
1    0.063058
Name: event, dtype: float64
Most predictive variables for LD:
hLD%                        0.030303
panglesweetspotpercent      0.027653
pLD%                        0.024726
hMed%                       0.023338
hanglesweetspotpercent      0.019761
h1B                         0.017787
hOppo%                      0.016141
pSIERA                      0.015805
hSH                         0.015049
pavg_distance               0.014597
dtype: float64
----------------------------------------
HR baseline: 0    0.967077
1    0.032923
Name: event, dtype: float64
Most predictive variables for HR:
hISO              0.084815
hHR/FB            0.080653
hbrl_pa           0.076712
pHR/9             0.075470
hbrl_percent      0.074908
hHR               0.074633
hSLG              0.071157
hbarrels          0.067352
pHR/FB            0.062387
hfbld             0.062117
dtype: float64
----------------------------------------
```

Strength of Relationship, Across Years

Statcast Metric	y1 → y2
Avg EV: FB & LD	0.82
Max EV	0.81
Brl/BBE %	0.79
Avg EV: all	0.78
Avg dist	0.75
Brl/PA %	0.74
95+ mph %	0.73
Avg FB dist	0.64
Avg EV: GB	0.61
Max dist	0.56
Avg HR dist	0.47

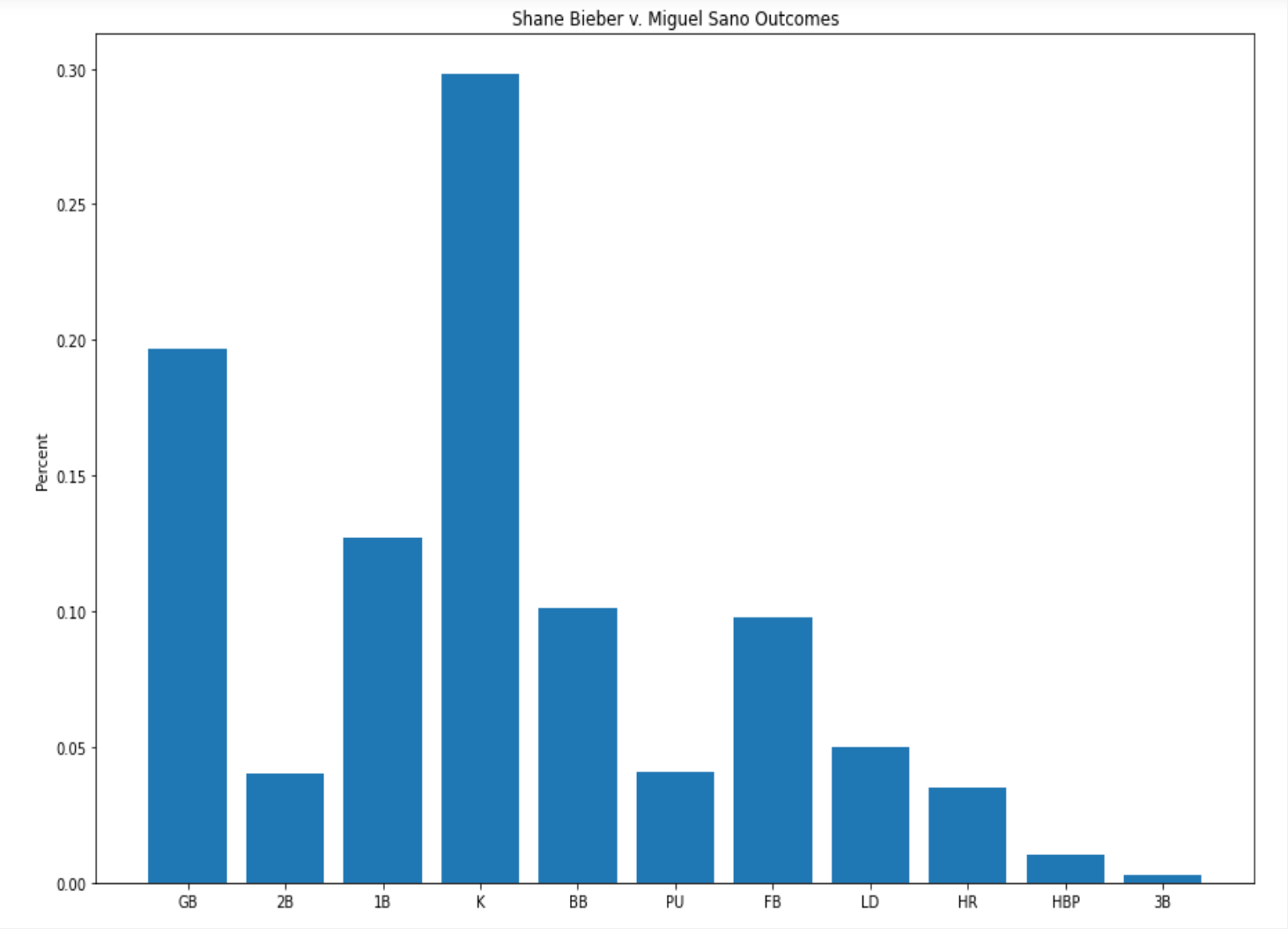
SOURCE: Statcast

FITTING THE MODEL: NEURAL NET

- ‘Softmax’ best for multiclass classification
- Data leakage is OK here!
- Due to time constraints, different configurations were done by hand
- Room for improvement
- Never going to be able to improve all that much above percentage of highest class
- .24 K v .28 accuracy

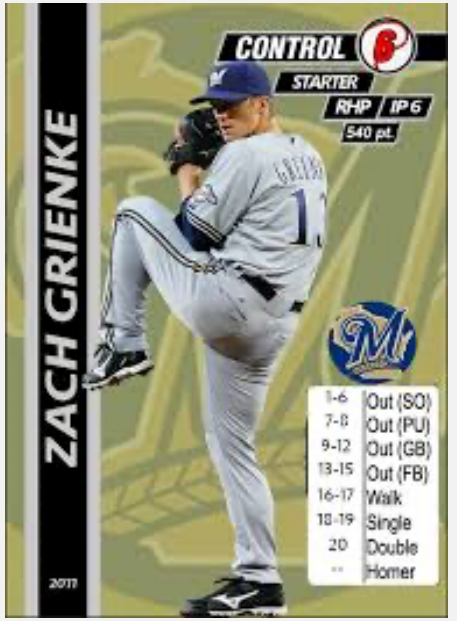
CLASSIFYING VS. DISTRIBUTION

- Want the probabilities of softmax, not predictions
- Dealing with issues of unbalanced classes difficult
- Using Categorical Accuracy, TopKCategorical Accuracy
- Struggles to assess situations with lower chances of major outcomes



FUNCTIONALIZE MODEL

- Define the simAB function
- In take pitcher/hitter names and seasons to pull stats from
- Options to change batter/pitcher handedness
- Construct input array into shape required for the neural network
- Return output into easily distilled format



https://shlabotnikreport.wordpress.com/tag/2001-mlb-showdown/

DISTRIBUTION

```
In [963]: simAB('Shane Bieber','Miguel Sano',output=('probs',0))
```

Out[963]:

	GB	2B	1B	K	BB	PU	FB	LD	HR	HBP	3B
0	0.197	0.04	0.127	0.298	0.101	0.041	0.098	0.05	0.035	0.01	0.003

INDIVIDUAL RESULTS

```
In [974]: simAB('Shane Bieber','Miguel Sano',output=('pa',10))

Out[974]: array([['1B'],
                 ['GB'],
                 ['BB'],
                 ['BB'],
                 ['GB'],
                 ['PU'],
                 ['BB'],
                 ['FB'],
                 ['GB'],
                 ['K']], dtype='<U3')
```

STATS OVER A NUMBER OF PA

```
In [980]: simAB('Shane Bieber','Miguel Sano',output=('statline',600))
```

Out[980]:

	PA	BA	HR	2B	SLG	OBP	oPS	BABIP	K	BB	pIP	pWHIP
0	600	0.214	33	14	0.426	0.288	0.714	0.257	185	59	140.2	1.23

FURTHER AREAS OF SEARCH

- Baseball Wise:
- Customize ballpark
- Customize weather
- Add pitcher fatigue, runners on base, other game contexts
- Include MAP estimate builder within function
- Batted Ball Dilemma, Statcast Data

- Computation Wise:
- Fit more neural networks
- Attempt to further correct unbalanced classes issues
- Compress outcomes

NEXT STEPS: GAME SIMULATION

- Roughly 300 conditional statements/loops
- Feed in lineup of hitters and pitchers
- Situations very rough
- No sac bunting, stealing, base running complexity
- Extra conditionals required for relief pitcher specification

inning	half	outs	home_runs	away_runs	1stB	2ndB	3rdB
3	1	1	3	1	0	1	1

SimAB = 'GB'

inning	half	outs	home_runs	away_runs	1stB	2ndB	3rdB
3	1	2	4	1	0	0	1

SimAB(next hitter in lineup array, pitcher)

A BIG THANK YOU