CMSAC ’19: Trouble with the Curve

Jacob Danovitch

Carleton University | Microsoft Cortana

Why the name?

Source: nytimes (https://www.nytimes.com/2012/09/21/movies/trouble-with-the-curve-with-clint-eastwood-and-amy-adams.html)

The Dataset

- Scouting reports from MLB.com & FanGraphs.com circa 2013
- 20-80 grades, position, age player IDs where possible

```
Size: (9175, 26)
```

Out[2]:

	name	key_mlbam	key_fangraphs	age	year	primary_position	
5800	Luis Medina	665622	0	17.7	2017	RHP	2
895	Blake Anderson	656190	0	18.0	2014	C	2
8935	Xavier Edwards	669364	0	19.4	2019	2B	2

3 rows × 26 columns

Dataset Statistics

Out[3]:

	mean	std	min	50%	max
age	20.810431	2.307676	15.30	21.0	31.90
year	2016.715095	1.979656	2013.00	2017.0	2019.00
eta	2018.802799	2.422862	2013.00	2019.0	2025.00
Arm	53.824127	6.897962	30.00	55.0	80.00
Changeup	49.858290	5.364976	30.00	50.0	70.00
Control	49.205290	5.059892	30.00	50.0	70.00
Curveball	52.929583	5.675287	35.00	55.0	70.00
Cutter	52.475000	4.936723	40.00	50.0	70.00
Fastball	59.531873	6.665786	40.00	60.0	80.00
Field	51.633871	5.503923	30.00	50.0	80.00
Hit	49.681105	5.510597	30.00	50.0	80.00
Power	47.932818	9.420963	20.00	50.0	80.00
Run	48.837240	12.169090	20.00	50.0	80.00
Slider	52.735618	5.121491	30.00	50.0	70.00
Splitter	53.333333	7.637626	40.00	50.0	70.00
mlb_played_first	2017.065463	1.735580	2010.00	2017.0	2019.00
debut_age	23.374194	1.558851	18.89	23.4	29.27
label	0.078147	0.613681	-1.00	0.0	1.00

Positional distribution:

Out[4]:

	primary_position
RHP	0.3566
OF	0.2068
LHP	0.1196
SS	0.1171
C	0.0638
3B	0.0579
2B	0.0395
1B	0.0347
UTIL	0.0043

Label distribution:

Out[5]:

	label
0	0.6173
1	0.2304
-1	0.1523

The 20-80 Scale

How do scouts grade prospects by position?

- Lefties have better control
- Righties have better fastballs

Out[6]:

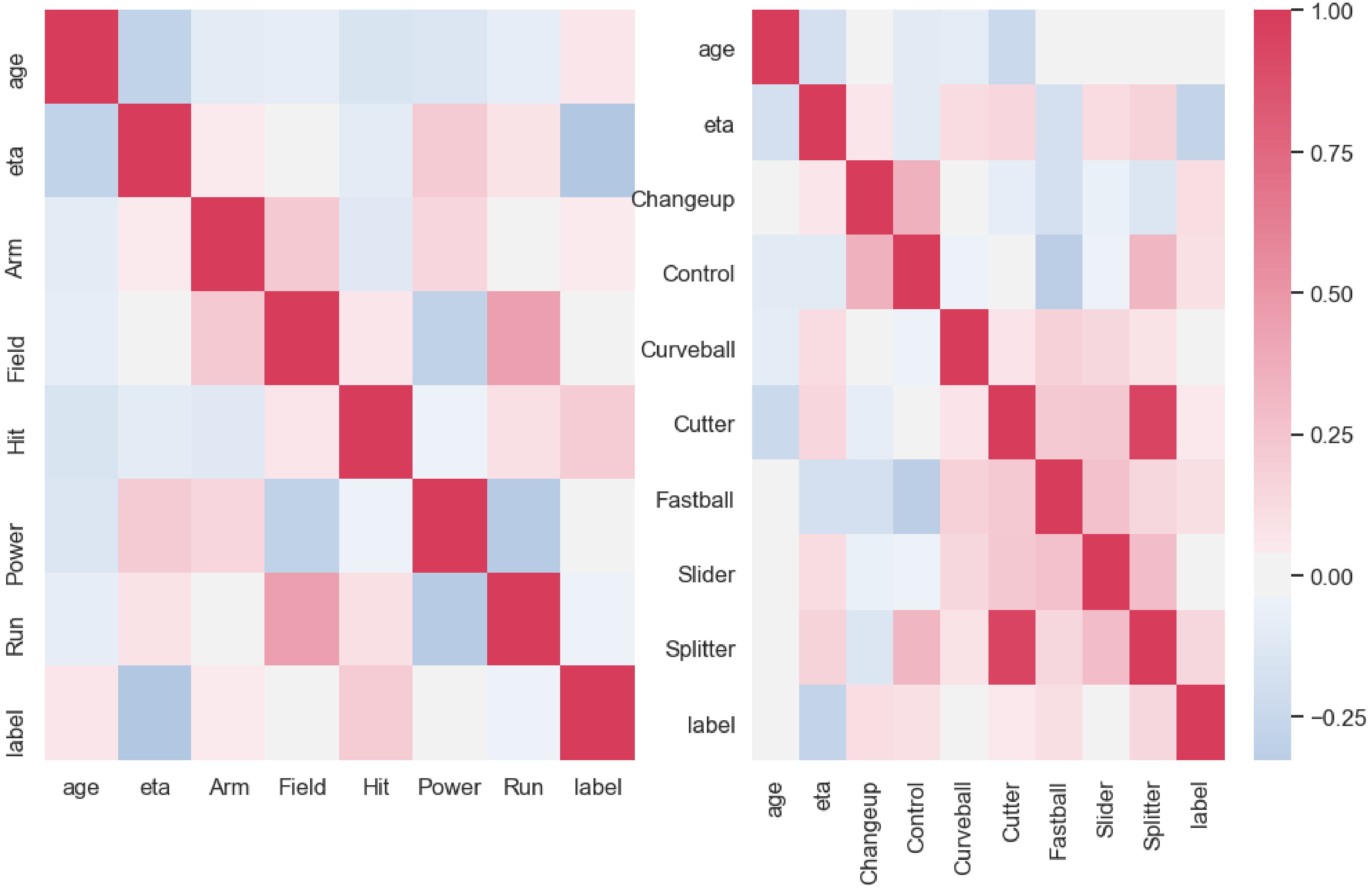
primary_position	LHP	RHP
Control	49.8371	48.9556
Fastball	56.3575	60.6458
Changeup	51.1233	49.3552
Curveball	52.6842	52.8867
Cutter	50.3409	52.9801
Slider	52.0267	52.912
Splitter	55	52.9412

- Up-the-middle spots are more defensive
- Corner guys are more power/arm oriented
- C, UTIL are jack-of-all-trades

Out[7]:

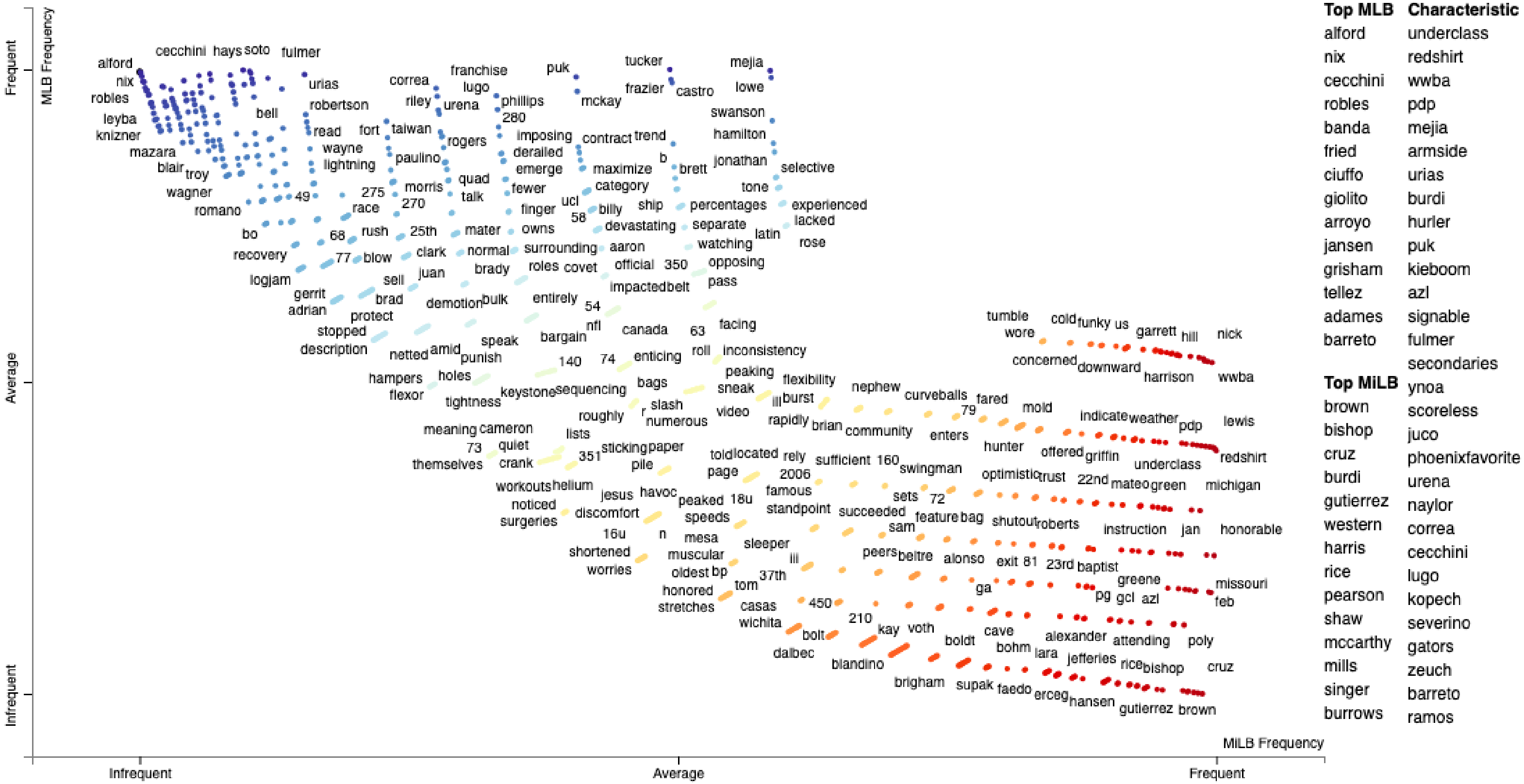
primary_position	1B	2B	3B	C	OF	SS	U
Hit	50.0324	52.2897	49.4915	47.6602	49.5577	50.0817	5
Power	55.1439	43.1034	53.0462	47.9333	49.3508	42.7895	4
Field	47.8669	49.931	48.9317	50.9398	52.5234	53.0971	5
Run	33.6151	51.5759	40.578	34.8495	54.5161	53.4155	5
Arm	50.1871	49.3069	56.1436	56	52.2856	55.9876	5

Inter-grade correlations


age
eta
Arm
Field
Hit
Power
Run
label
age
eta
Arm
Field
Hit
Power
Run
label
age
eta
Changeup
Control
Curveball
Cutter
Fastball
Slider
Splitter
label
age
eta
Changeup
Control
Curveball
Cutter
Fastball
Slider
Splitter
label
1.00
0.75
0.50
0.25
0.00
−0.25

Identifying Successful Prospects

How are successful prospects described?



Top MLB
alford
nix
cecchini
robles
banda
fried
ciuffo
giolito
arroyo
jansen
grisham
tellez
adames
barreto

Top MiLB
brown
bishop
cruz
burdi
gutierrez
western
harris
rice
pearson
shaw
mccarthy
mills
singer
burrows

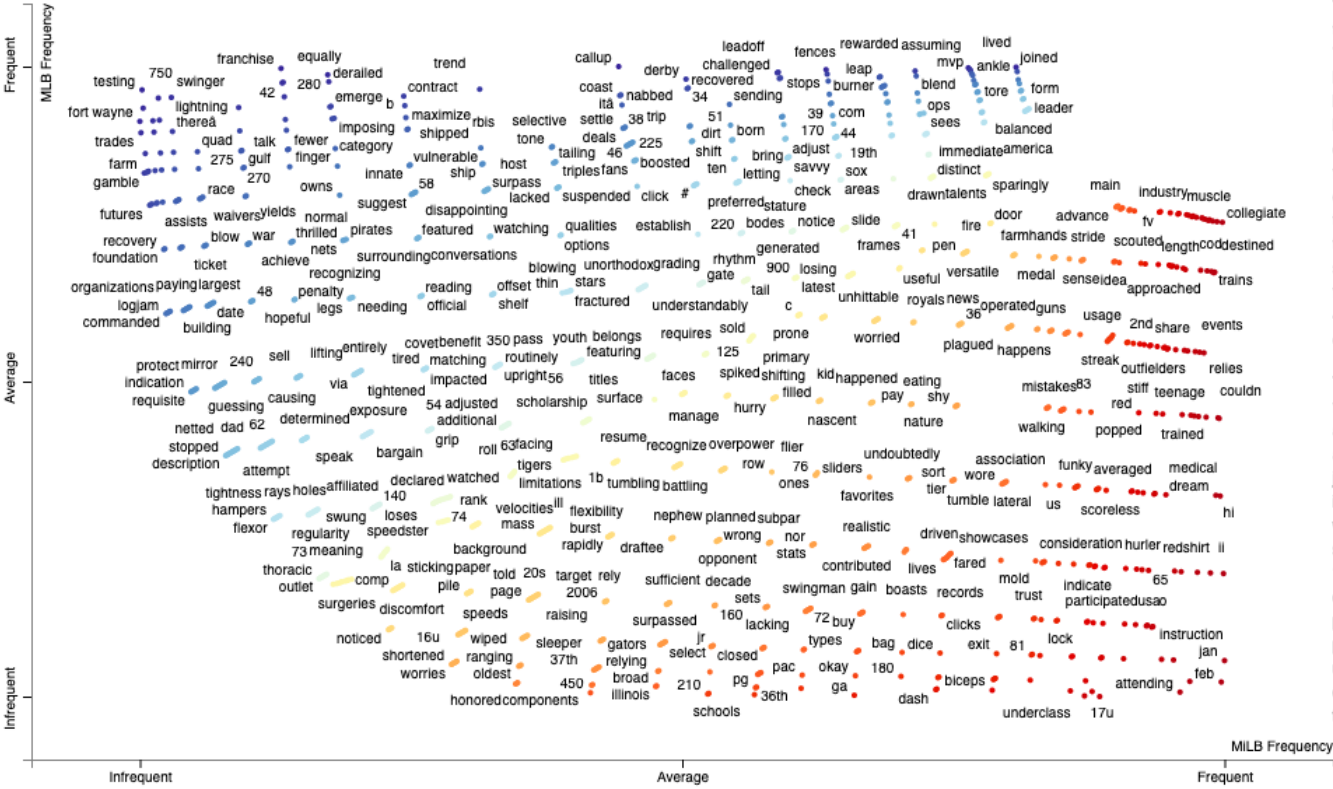
Characteristic
underclass
redshirt
wwba
pdp
mejia
armside
urias
burdi
hurler
puk
kieboom
azl
signable
fulmer
secondaries
ynoa
scoreless
juco
phoenixfavorite
urena
naylor
correa
cecchini
lugo
kopech
severino
gators
zeuch
barreto
ramos

What do you notice about the most frequent words used to describe successful prospects?

All the most discriminative terms are player names!

```
Out[10]:
```

	term	MiLB freq	MLB freq	MLB Score
35864	alford	0	47	1.000000
132332	nix	0	43	0.999147
29913	cecchini	0	38	0.997598
109491	robles	0	35	0.996513
36508	banda	0	34	0.996077
130563	fried	0	31	0.994667
79964	arroyo	0	30	0.994142
189461	ciuffo	0	30	0.994142
102142	tellez	0	29	0.993580
219935	grisham	0	29	0.993580



Post-entity masking

Top MLB
testing
750
lightning
blockbuster
hagerstown
exceptionally
fort wayne
designated
trades
thereâ
swinger
franchise
42
equally

Top MiLB
feb
armside
attending
jan
instruction
weekend
specialist
17u
effectiveness
participated
price
underclass
o
teen

Characteristic
redshirt
callup
outfielders
projectability
armside
scouted
labrum
hurler
signable
barreling
midseason
leadoff
farmhands
secondaries
scoreless
uppercut
stateside
rebounded
unhittable
backstops
changeups
unspectacular
tweener
bloodlines
sneaky
parlayed
stocky
nabbed
groundout
baserunning

Not perfect, but better!

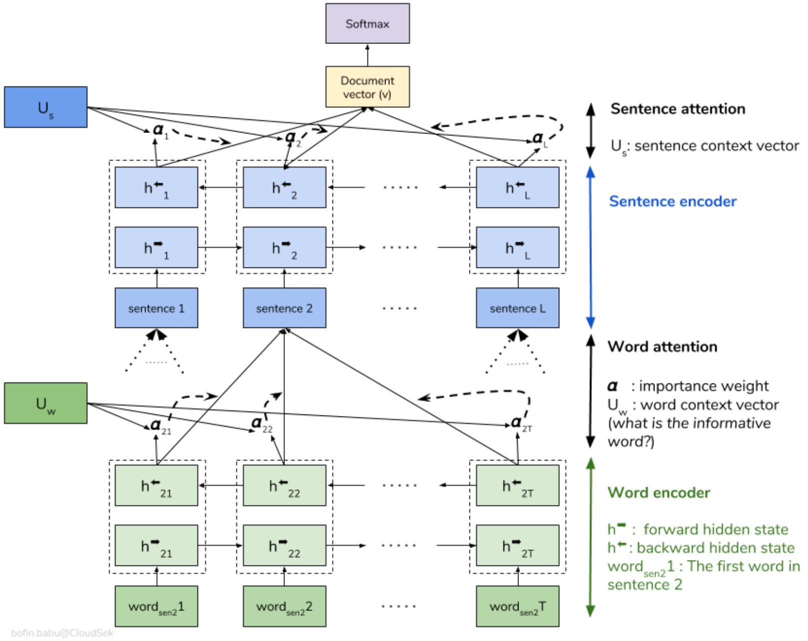
Out[12]:

	term	MiLB freq	MLB freq	MLB Score
716	trade	210	218	1.000000
582	the package	20	42	0.999735
1151	at age	144	167	0.999512
236	as part	87	105	0.997560
2214	traded	116	118	0.997343
9034	youngest	54	78	0.997062
744	three team	17	35	0.997009
8289	that sent	39	61	0.996948
576	organization deal	34	56	0.996236
14306	age 20	30	53	0.996151

Classifying successful prospects

Task Definition

- **Task**: Sequence of tokens ⟶ binary label
- **Solution**: Hierarchical Attention Network (among others)



Source: [medium.com (https://medium.com/analytics-vidhya/hierarchical-attention-networks-d220318cf87e)](https://medium.com/analytics-vidhya/hierarchical-attention-networks-d220318cf87e)

Additional considerations

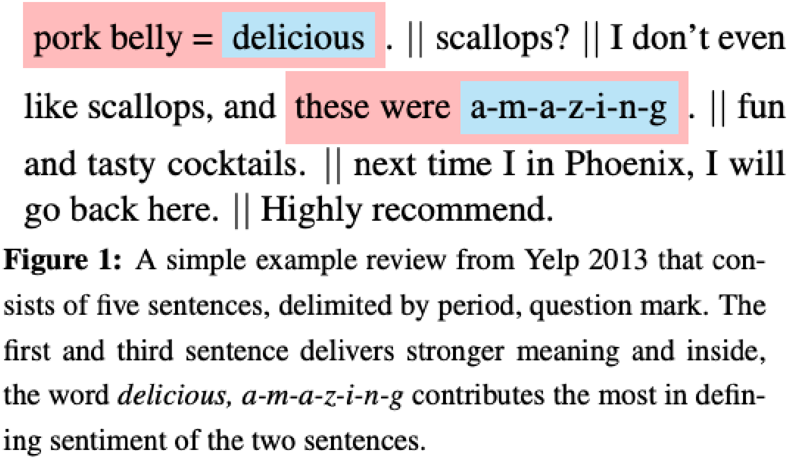
Problem	Solution
Heavy class imbalance	Resampling + loss reweighting
Data sparsity	Data augmentation
(Relatively) small corpus	Pre-trained GloVe embeddings

Results

Model	Accuracy	F1
Bag-Of-Embeddings	64.65%	53.78%
TextCNN	69.02%	56.42%
LSTM+SelfAttn	68.64%	54.65%
BCN	73.52%	43.33%
HAN	66.00%	54.07%

Hyperparameters: link (https://github.com/jacobdanovitch/jdnlp/blob/master/experiments/twtc.json)

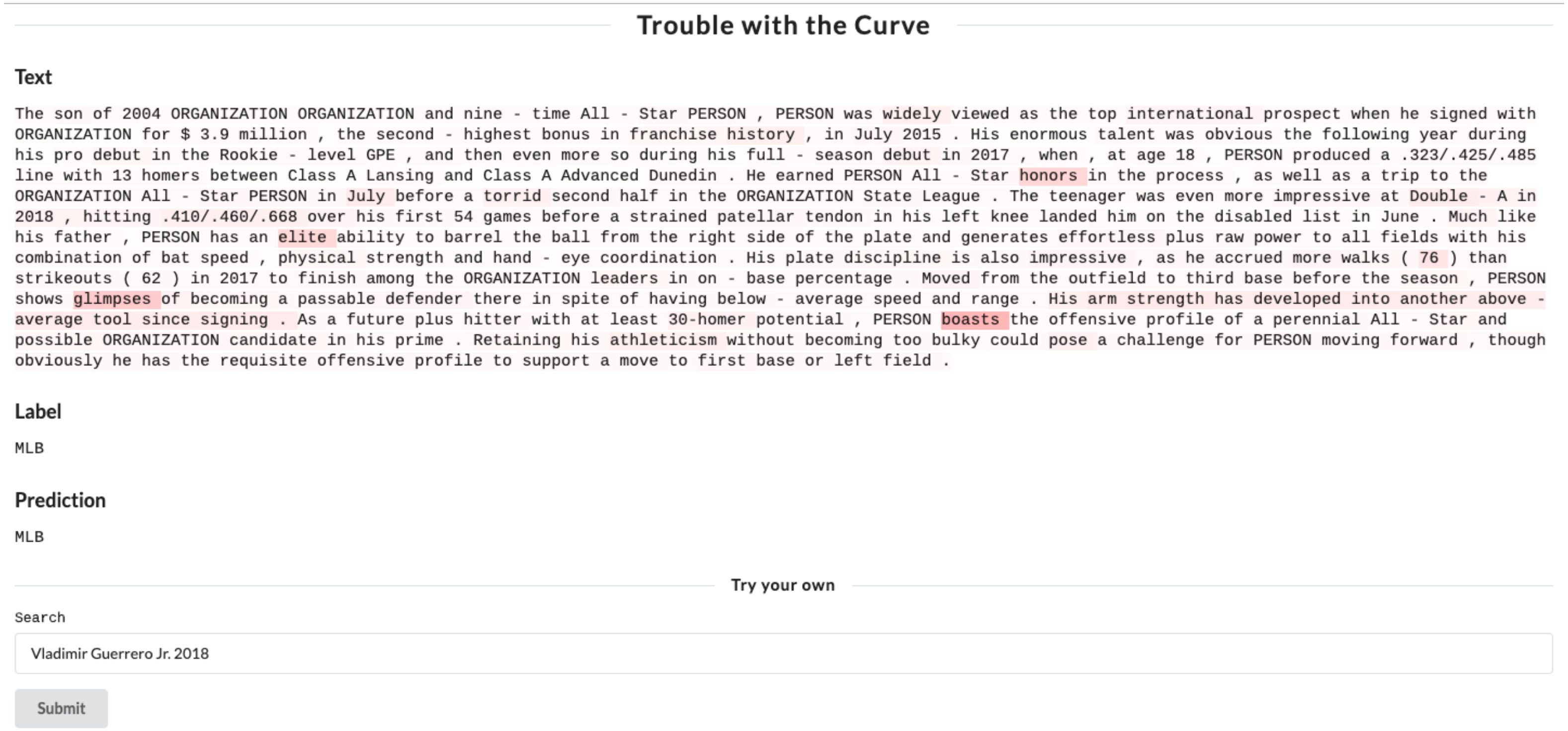
- Why did you use a HAN if it wasn't even the best one?



pork belly = delicious . || scallops? || I don’t even like scallops, and these were a-m-a-z-i-n-g . || fun and tasty cocktails. || next time I in Phoenix, I will go back here. || Highly recommend.

Figure 1: A simple example review from Yelp 2013 that consists of five sentences, delimited by period, question mark. The first and third sentence delivers stronger meaning and inside, the word *delicious, a-m-a-z-i-n-g* contributes the most in defining sentiment of the two sentences.

Source: Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., & Hovy, E.H. (2016). Hierarchical Attention Networks for Document Classification. HLT-NAACL.



Trouble with the Curve

Text

The son of 2004 ORGANIZATION ORGANIZATION and nine - time All - Star PERSON , PERSON was widely viewed as the top international prospect when he signed with ORGANIZATION for $ 3.9 million , the second - highest bonus in franchise history , in July 2015 . His enormous talent was obvious the following year during his pro debut in the Rookie - level GPE , and then even more so during his full - season debut in 2017 , when , at age 18 , PERSON produced a .323/.425/.485 line with 13 homers between Class A Lansing and Class A Advanced Dunedin . He earned PERSON All - Star honors in the process , as well as a trip to the ORGANIZATION All - Star PERSON in July before a torrid second half in the ORGANIZATION State League . The teenager was even more impressive at Double - A in 2018 , hitting .410/.460/.668 over his first 54 games before a strained patellar tendon in his left knee landed him on the disabled list in June . Much like his father , PERSON has an elite ability to barrel the ball from the right side of the plate and generates effortless plus raw power to all fields with his combination of bat speed , physical strength and hand - eye coordination . His plate discipline is also impressive , as he accrued more walks (76) than strikeouts (62) in 2017 to finish among the ORGANIZATION leaders in on - base percentage . Moved from the outfield to third base before the season , PERSON shows glimpses of becoming a passable defender there in spite of having below - average speed and range . His arm strength has developed into another above - average tool since signing . As a future plus hitter with at least 30-homer potential , PERSON boasts the offensive profile of a perennial All - Star and possible ORGANIZATION candidate in his prime . Retaining his athleticism without becoming too bulky could pose a challenge for PERSON moving forward , though obviously he has the requisite offensive profile to support a move to first base or left field .

Label

MLB

Prediction

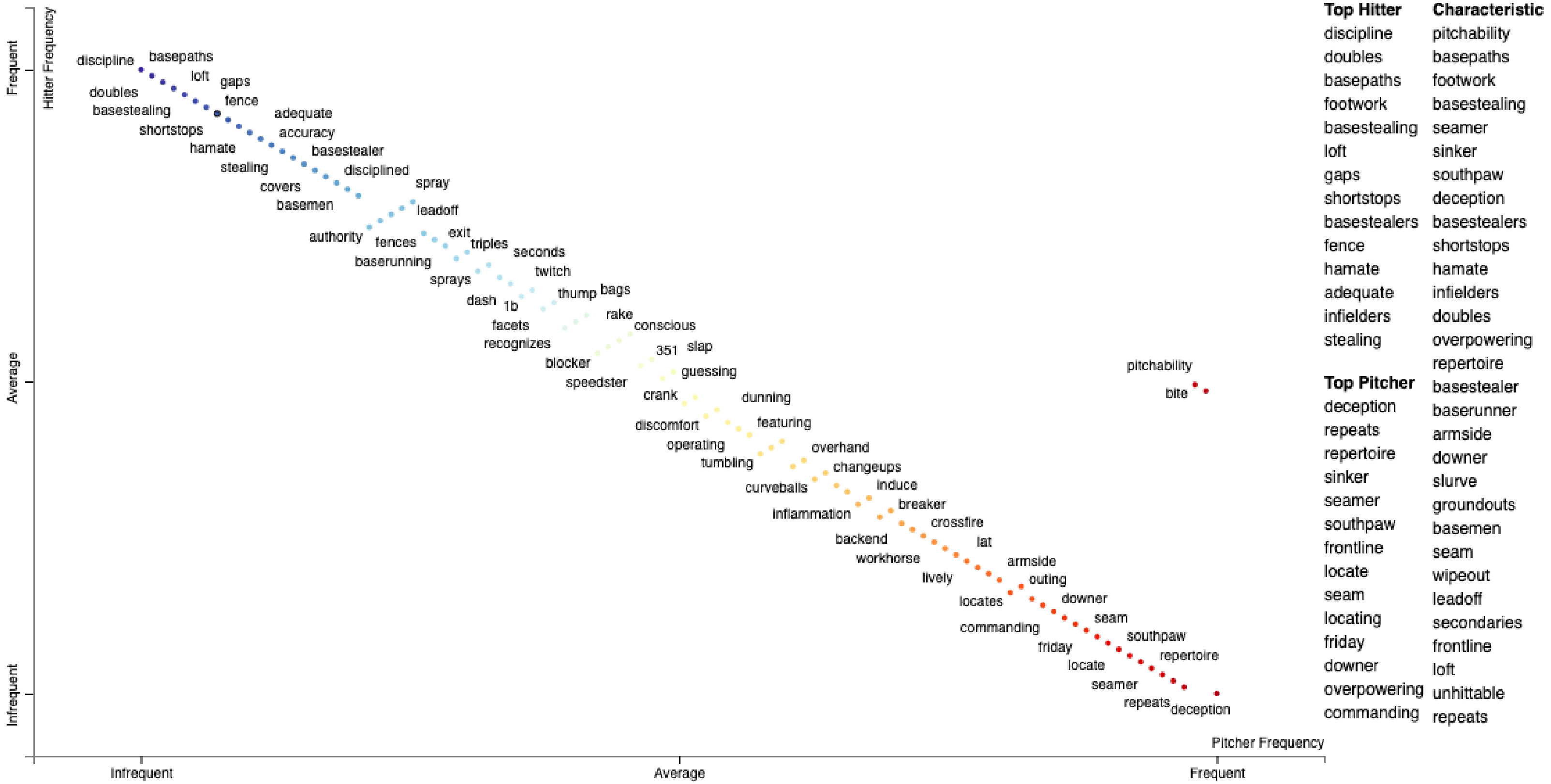
MLB

Search

Vladimir Guerrero Jr. 2018

Scouting the scouting reports

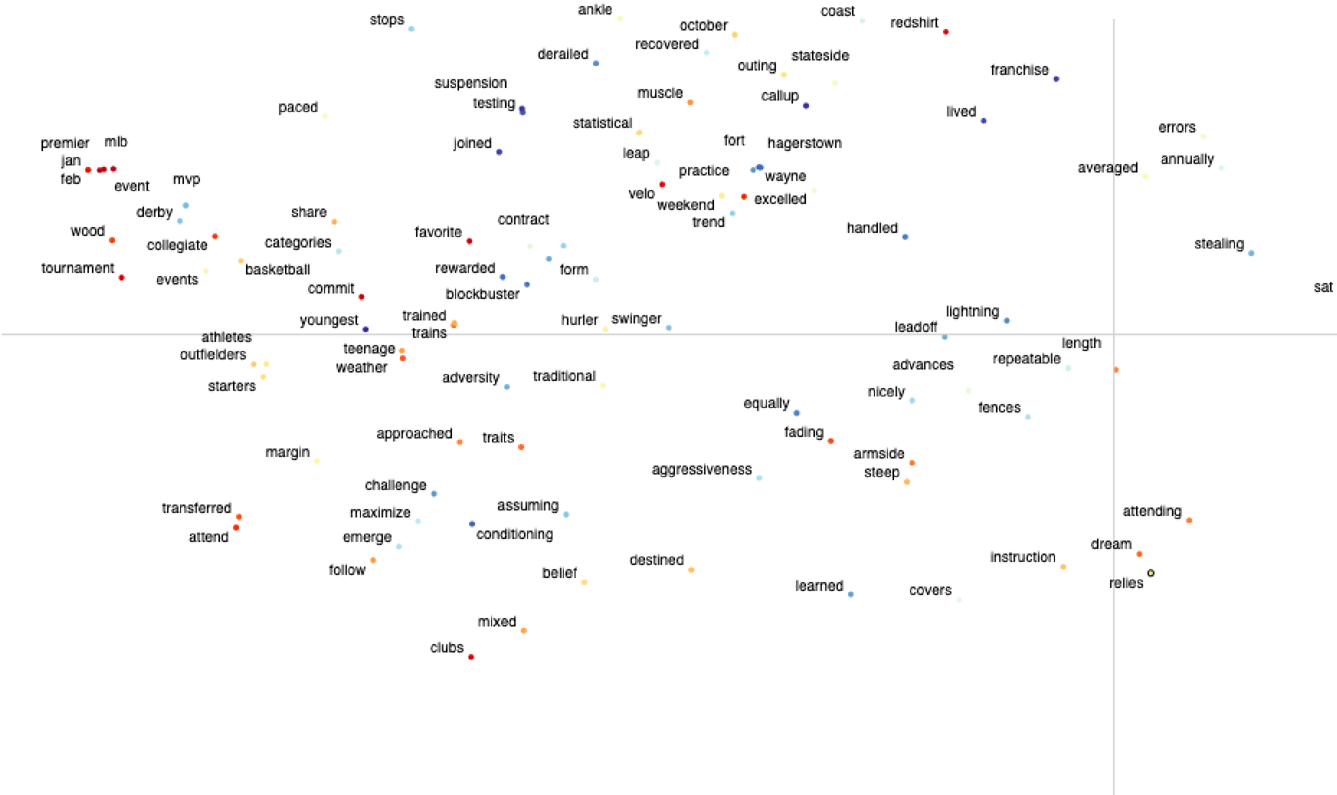
Language variation



Variation in the reports of hitters and pitchers

Semantic similarity

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

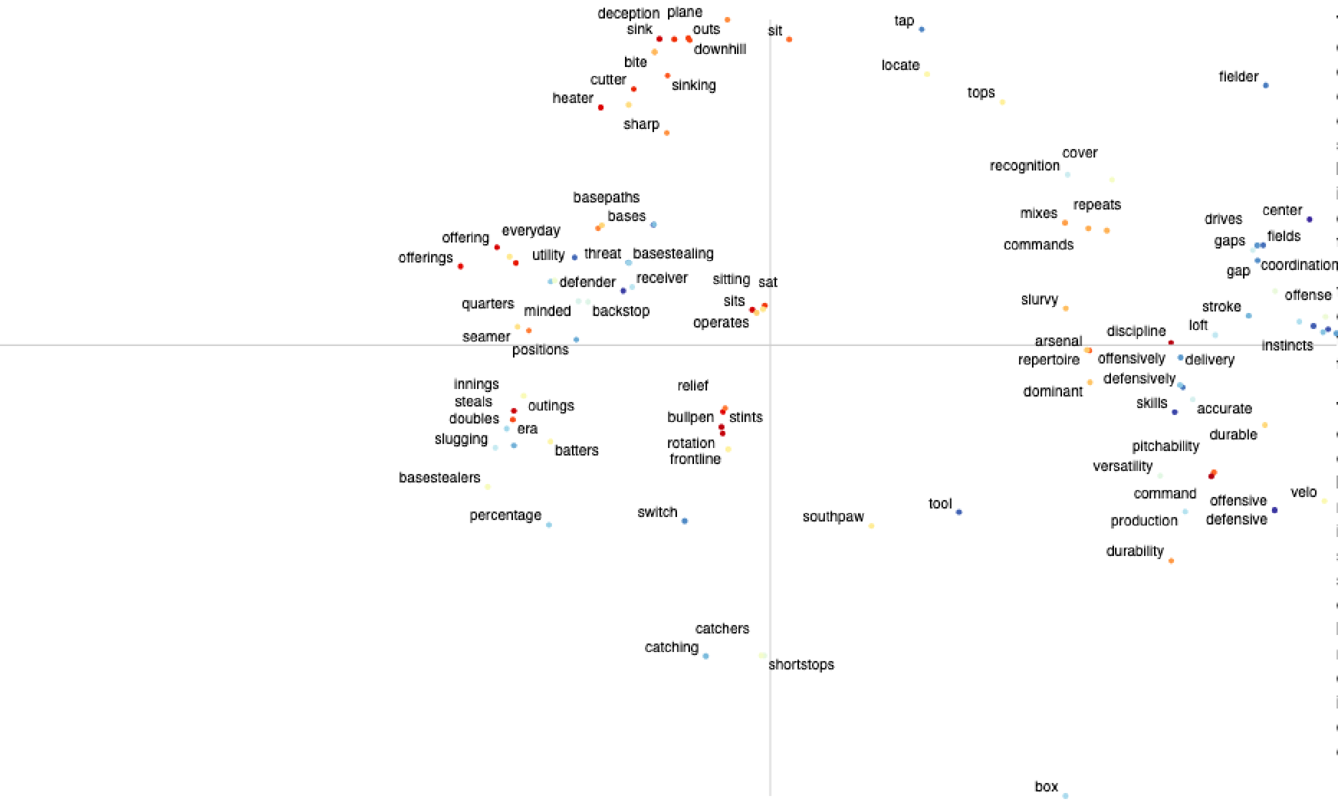


Top MLB
youngest
callup
suspension
franchise
joined
lived
fort
testing
conditioning
wayne
rewarded
blockbuster
handled
equally

Top MiLB
mlb
event
premier
commit
tournament
favorite
sat
clubs
jan
redshirt
velo
feb
practice
attend

Characteristic
redshirt
velo
callup
outfielders
armside
hurler
leadoff
excelled
stateside
repeatable
aggressiveness
youngest
fading
mlb
derailed
outing
collegiate
adversity
destined
hagerstown
traits
rewarded
stealing
starters
steep
fences
paced
relies
approached
averaged

Word-level similarity by success



Top Hitter

center
offensive
defensive
defender
skills
bases
instincts
defense
tool
everyday
fields
defensively
fielder
tap

Top Pitcher

command
delivery
bullpen
rotation
innings
sink
sits
offering
heater
relief
offerings
inning
cutter
deception

Characteristic

bullpen
defensively
innings
pitchability
basepaths
stints
quickness
instincts
fielder
slurvy
basestealing
offensively
footwork
seamer
defender
slugging
sinker
backstop
southpaw
basestealers
shortstops
inning
rotation
defensive
velo
sits
catchers
batters
deception
sinking

Word-level similarity by position

Conclusion

- Lessons learned
- Future directions

Thank you!

Questions?

🔗 jacobdanovitch.me (http://jacobdanovitch.me) || jacobdanovitch (https://github.com/jacobdanovitch/)