

DATA-DRIVEN BUSINESS MANAGEMENT

Working with the Data of Sports

Thomas W. Miller

Northwestern University

March 6, 2018

San Jose, California

Strata
DATA CONFERENCE

Working with the Data of Sports

There is a rich history of baseball fans and sports analysts using play-by-play information to compute traditional performance measures, such as batting average, on-base percentage, and slugging percentage for batters, and earned-run average and strikeouts per inning for pitchers.

There is no shortage of event data, with baseball records showing team and pitcher-batter matchups, outs, runners on base, and the outcome of each play, along with runs scored. But sports analytics today is more than a matter of analyzing box scores and play-by-play statistics. Faced with detailed on-field or on-court data from every game, sports teams face challenges in data management, data engineering, and analytics.

Suppose you work for the Dodgers. You seek competitive advantage through data science and deep learning. You understand traditional player performance metrics, but you are concerned that they fail to reflect performance within the context of a game.

Who should start for the Dodgers in game seven of the 2017 World Series?

2017 World Series

Setting the stage for game seven

Game 1 (Los Angeles, Tuesday, Oct. 24) Astros 1, **Dodgers** 3

Game 2 (Los Angeles, Wednesday, Oct. 25) **Astros** 7, Dodgers 6 (11 innings)

Game 3 (Houston, Friday, Oct. 27) Dodgers 3, **Astros** 5

Game 4 (Houston, Saturday, Oct. 28) **Dodgers** 6, Astros 2

Game 5 (Houston, Sunday, Oct. 29) Dodgers 12, **Astros** 13 (10 innings)

Game 6 (Los Angeles, Tuesday, Oct. 31) Astros 1, **Dodgers** 3

Desirable Measurement Attributes

Reliable	A measure should be trustworthy and repeatable.
Valid	A measure should measure the attribute it is said to measure.
Explicit	Procedures should be unambiguous and defined in detail.
Accessible	A measure should come from data that are easily obtained.
Tractable	A measure should be easy to work with.
Comprehensible	A measure should be simple and straightforward.
Transparent	The method of measurement should be fully documented.

Sports Analytics and Data Science: Winning the Game with Methods and Models
(Miller 2016).

Representing Players

Think of a player as a vector of numbers. What numbers should we use?

One-Hot Encoding

Each player vector is as long as the number of players in the study. The set of vectors represents a nominal measure, an assignment of numbers to player names.

Player-Linked Performance

Each pitcher vector is as long as the number of performance measures for pitchers. Each batter vector is as long as the number of performance measures for batting. This is feature engineering guided by an understanding of the game. Aggregate measures do not usually reflect context.

Neural Network Embeddings

Use a neural network to generate vector representations of pitchers and batters, providing more complete performance measures in context. These vector representations can then be used to evaluate teams and players, predict runs scored, and guide in-game strategy.

	Yu Darvish	Rich Hill	Clayton Kershaw	Kenta Maeda	
Yu Darvish	1	0	0	0	
Rich Hill	0	1	0	0	
Clayton Kershaw	0	0	1	0	
Kenta Maeda	0	0	0	1	
⋮	0	0	0	0	
Last Player	0	0	0	0	
⋮					
	George Springer	Jose Altuve	Carlos Correa	Yuri Gurriel	Last Player
George Springer	1	0	0	0	0
Jose Altuve	0	1	0	0	0
Carlos Correa	0	0	1	0	0
Yuri Gurriel	0	0	0	1	0
⋮					
Last Player	0	0	0	0	1

One-Hot Encoding

Each player represented by a vector of binary numbers

Scale Type: Nominal

Procedure: Simple and straightforward

Problems: Orthogonal vectors lacking context

Drawing on methods well-established in natural language processing, we can represent each player with a unique vector consisting of one 1 and $(N - 1)$ zeros, where N is the number of players in the study. One-hot vectors for any pair of players will be orthogonal (uncorrelated). One-hot vectors identify individual players, and that is all. They carry no additional information.

	Yu Darvish	Rich Hill	Clayton Kershaw	Kenta Maeda	Last Pitcher
L/R	R	L	L	R	
ERA	3.86	3.32	2.31	4.22	
WHIP	1.16	1.09	0.96	1.15	
KIP	1.21	1.22	1.15	1.04	
•					
•					
•					
Last Pitching Measure					
	George Springer	Jose Altuve	Carlos Correa	Yuri Gurriel	Last Batter
L/S/R	R	R	R	R	
BA	.283	.346	.315	.299	
OBP	.367	.410	.391	.332	
SLG	.522	.547	.550	.486	
•					
•					
•					
Last Batting Measure					

Player-Linked Performance

Each player represented by a vector of performance measures

Scale Type: Varies by vector element

Procedure: Arbitrary and often complex

Problems: Aggregate measures, usually lacking context

Each pitcher vector is as long as the number of performance measures for pitchers. Each batter vector is as long as the number of performance measures for batting. There are many potential performance measures, and there is much controversy as to which measures are best.

Problems with Player-Linked Performance Measures

Intractable or incomprehensible

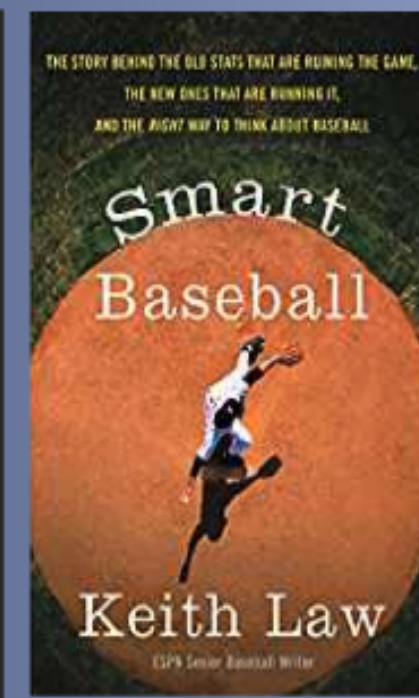
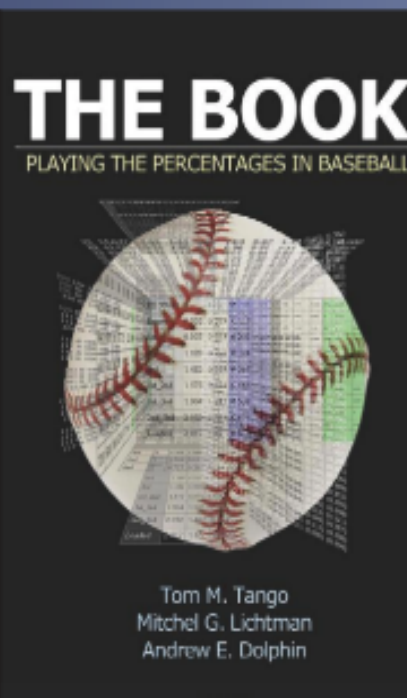
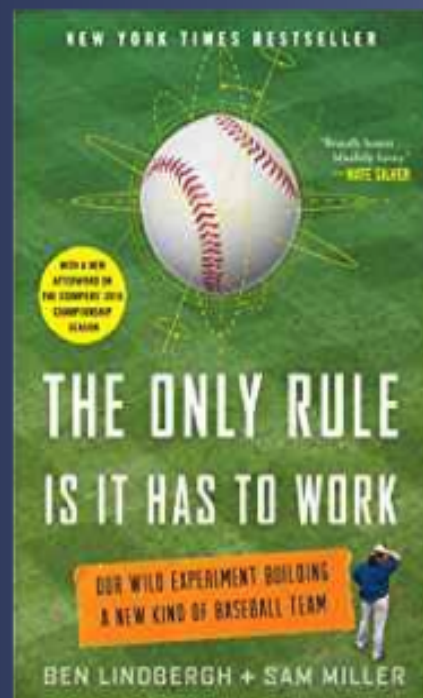
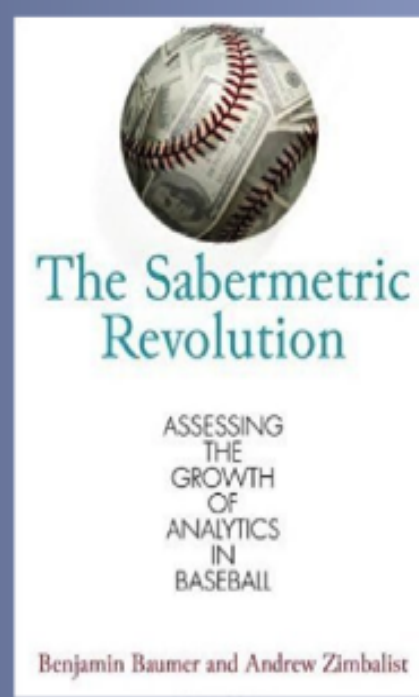
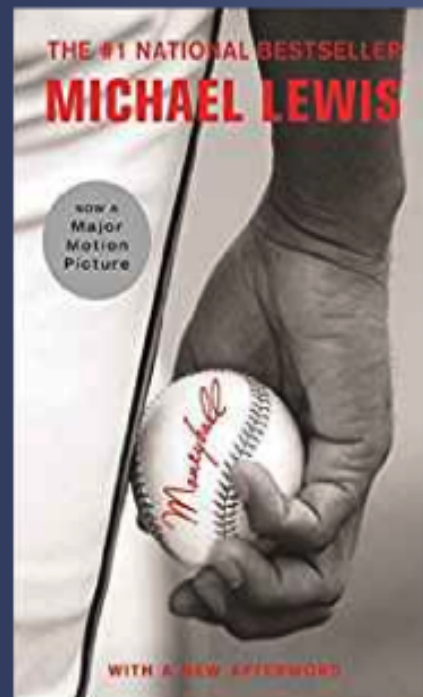
Many proposed measures are difficult to compute. Some are proprietary. Others are difficult to understand and interpret.

Inconsistent

Common measures for pitching are distinct from common measures for batting. American and National League norms differ due to the designated hitter rule.

Pitcher-batter match-ups

There are small sample sizes associated with specific pitcher-batter match-ups. Traditional performance measures computed from these data are of little strategic value.



Weighted On-base Average

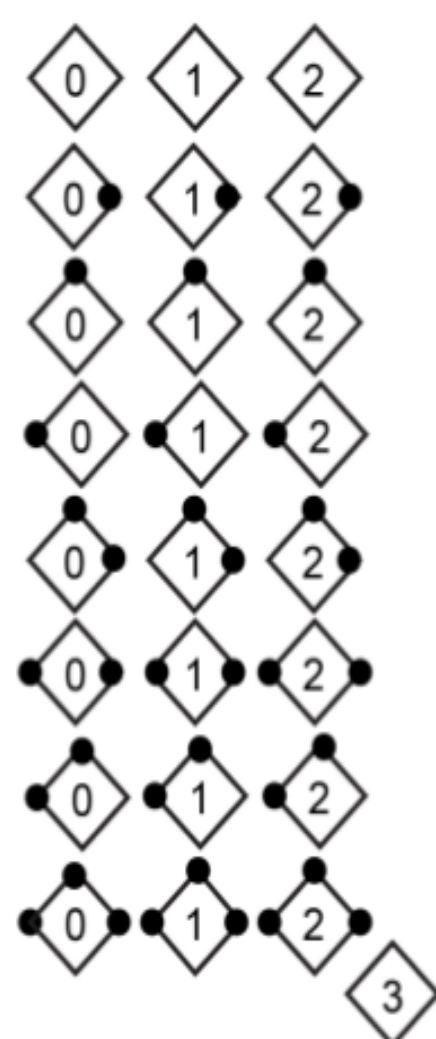
An alternative to on-base percentage plus slugging (OPS). A weighted linear combination of various hitting measures.

Scale Type: Ratio

Procedure: Complicated, weights vary by norm group

Problems: Not a simple percentage

$$wOBA = [(0.72 \times NIBB) + (0.75 \times HBP) + (0.90 \times 1B) + (0.92 \times RBOE) + (1.24 \times 2B) + (1.56 \times 3B) + (1.95 \times HR)] / PA$$
 where NIBB is the number of intentional bases on balls, HBP is the number of times hit by a pitch, 1B, 2B, 3B, and HR are the numbers singles, doubles, triples, and home runs, respectively. PA is the number of plate appearances. From *The Book* (Tango, Lichtman, and Dolphin, 2006).



State Code	Outs	Runners on Base			Expected Runs
		First	Second	Third	
[0000]	0	0	0	0	0.461
[1000]	1	0	0	0	0.234
[2000]	2	0	0	0	0.095
[0100]	0	1	0	0	0.831
[1100]	1	1	0	0	0.489
[2100]	2	1	0	0	0.214
[0010]	0	0	1	0	1.068
[1010]	1	0	1	0	0.644
[2010]	2	0	1	0	0.305
[0001]	0	0	0	1	1.426
[1001]	1	0	0	1	0.865
[2001]	2	0	0	1	0.413
[0110]	0	1	1	0	1.373
[1110]	1	1	1	0	0.908
[2110]	2	1	1	0	0.343
[0101]	0	1	0	1	1.798
[1101]	1	1	0	1	1.140
[2101]	2	1	0	1	0.471
[0011]	0	0	1	1	1.920
[1011]	1	0	1	1	1.354
[2011]	2	0	1	1	0.570
[0111]	0	1	1	1	2.282
[1111]	1	1	1	1	1.520
[2111]	2	1	1	1	0.736
[END]					0.000

From *Sports Analytics and Data Science* (Miller 2016). Expected runs estimates from FanGraphs.com (2015).

State-Transition-Based Measures

Build on well-established theory of Markov chains

Scale Type: Ratio

Procedure: Weighted averages of expected runs

Problems: Detailed calculations on play-by-play logs

Every play involves a transition from one state to another, and every state has expected runs associated with it. So every play has a positive or negative change in expected runs. Pitchers and batters are measured on the same scale, with negative values good for pitchers and positive values good for batters.

(For an example, see Brad Smith. College Baseball Analytics, Great Lakes Analytics in Sports Conference, July 2017.)

Pitcher-Batter Match-ups

Much of baseball strategy deals with pitcher-batter match-ups, but what we know about specific pitcher-batter match-ups is limited.

Aggregate measures

Common pitcher performance measures are computed across all batters, and common batter performance measures are computed across all pitchers.

These say nothing about context or situations with a specific pitcher facing a specific batter.

Many measures have limited strategic value

Drilling down or filtering

We can disaggregate, drilling down or filtering the data. To advise for a particular context, we compute summary statistics on filtered data. For example, we can select all cases of a specific left-handed pitcher facing right-handed batters. We can drill down further using additional player data.

Filtered data are still aggregate data

Specific pitcher-batter match-ups

We could identify all plate appearances involving a specific pitcher and a specific batter. Traditional performance measures may be computed for these highly filtered data. For one season, the number of specific pitcher-batter match-ups will be small. With National League pitchers facing American League batters in the World Series, the sample size for many match-ups will be zero.

Small samples, large standard errors

Pitch f/x and TrackMan/Statcast

Major League Baseball Advanced Media has systems to track individual pitches, including their speed and trajectories, providing more detailed data on the tendencies of individual pitchers. In the aggregate, the locations of a pitcher's pitches can be matched up against the locations of pitches that a batter hits or misses.

Holds promise for future research

Modern Methods and Models

Data-driven ghosting

Recent developments in neural networks and deep learning have been on display at the MIT Sloan Sports Analytics Conference in the past two years. One avenue of research is data-driven ghosting. Neural networks learn movement patterns of real players. Then these learned behavior patterns are implemented in software robots. Ghosting systems are especially useful in sports with long intervals of continuous play, such as soccer and basketball. With Statcast data, we can imagine applications in baseball as well.

Benefits

Future ghosting systems are expected to assist with player training and game-day strategic planning.

Hoang M. Le, Peter Carr, Yisong Yue, and Patrick Lucey. Data-Driven Ghosting using Deep Imitation Learning, 2017 Research Papers Competition, MIT Sloan Sports Analytics Conference.

<https://www.disneyresearch.com/publication/data-driven-ghosting/>

Modern Methods and Models

Neural network embeddings

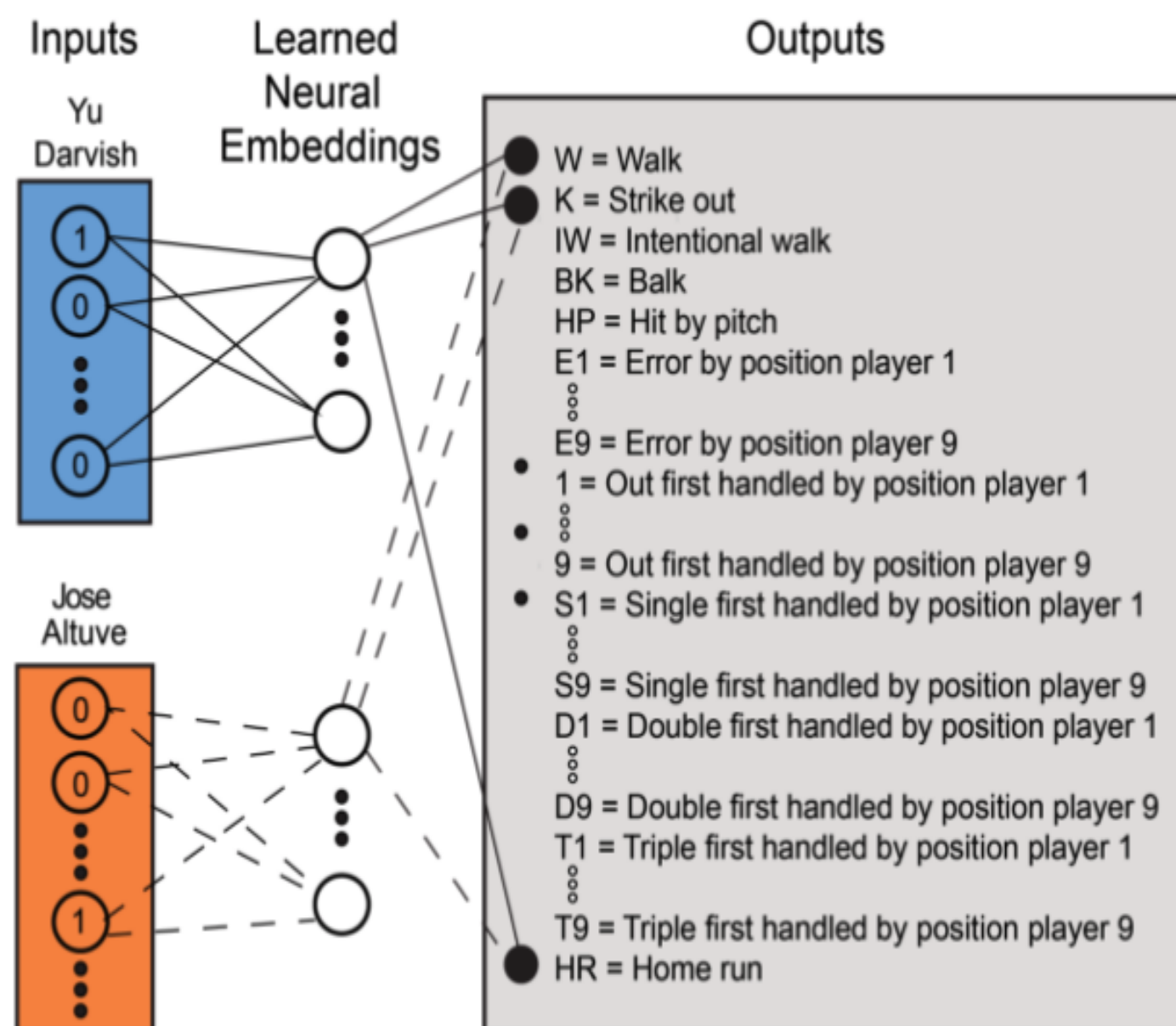
Another application of neural networks presents a radical shift in the way we think about performance measurement in baseball. Illustrated by Michael A. Alcorn (2018), this method draws on neural network applications in natural language processing and the technique known as ***word2vec***.

Michael A. Alcorn, (batter|pitcher)2vec: Statistic-Free Talent Modeling With Neural Player Embeddings, MIT Sloan Sports Analytics Conference, February 23, 2018. <http://www.sloansportsconference.com/wp-content/uploads/2018/02/1008.pdf>
Python Jupyter Notebook at <https://github.com/airalcorn2/batter-pitcher-2vec>

Benefits

Provides a method for predicting pitcher-batter outcomes for players who have never faced one another.

So what about Yu Darvish as the starting pitcher in game seven of the 2017 World Series? Let's see what neural network embeddings say.



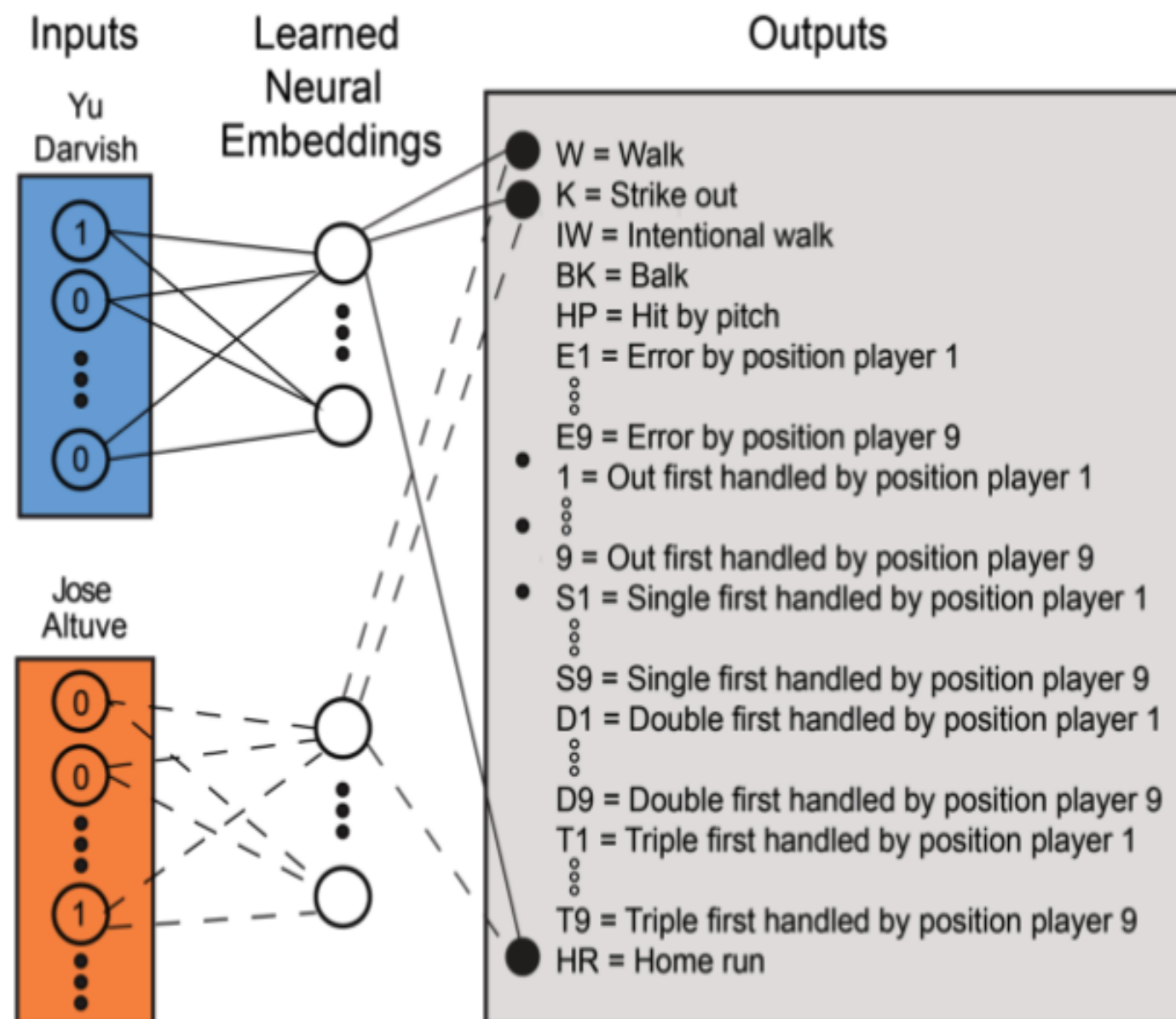
Neural Network Embeddings

Each player represented as a vector of real numbers

An example of what is possible when new methods are applied to old data, Alcorn's research draws on play-by-play data from 2013 to 2016, Retrosheet event files. (We add one more year's data to make predictions for the 2017 World Series.)

This is a simple neural network with one hidden layer.

The job of the network is to learn pitcher and batter vectors (embeddings) that predict the outcome of a plate appearance. "Context" in this model refers to the pitcher-batter combination.



Neural Network Embeddings

Each player represented as a vector of real numbers

Scale Type: Undefined scale, real numbers

Procedure: Neural network learns from data

Problems: Hard to describe in words

Real number vectors are generated from raw event data. No feature engineering required. No need for arbitrary measures of pitching or hitting prowess. We can use these real number vectors to evaluate teams and players, predict runs scored, and guide in-game strategy.

Neural Network Embeddings

How do they stack up as measurements?

Reliable	Reliably computed for any given network structure.
Valid	(Problem) No words to describe what is being measured.
Explicit	Estimation follows from standard optimization methods.
Accessible	Vectors estimated from data that are easily obtained.
Tractable	Vectors are easy to work with.
Comprehensible	(Problem) Neural network "black box" is hard to put into words.
Transparent	The method is fully documented with public-domain code.

Pitcher-Batter Match-ups

Who should start for the Dodgers in game seven of the 2017 World Series?

Game 1. Los Angeles, Tuesday, October 24 (Astros 1, **Dodgers** 3)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	Series
											ERA
Clayton Kershaw	7.0	83	57	0.69	3	1	1	0	11	1	1.29
Brandon Morrow	1.0	10	7	0.70	0	0	0	0	0	0	0.00
Kenley Jansen	1.0	14	9	0.64	0	0	0	0	1	0	0.00

Game 2. Los Angeles, Wednesday, October 25 (**Astros** 7, Dodgers 6, 11 innings)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	Series
											ERA
Rich Hill	4.0	60	42	0.70	3	1	1	3	7	0	2.25
Kenta Maeda	1.1	25	17	0.68	1	0	0	0	0	0	0.00
Tony Watson	0.2	1	1	1.00	0	0	0	0	0	0	0.00
Ross Stripling	0.0	4	0	0.00	0	0	0	1	0	0	0.00
Brandon Morrow	1.0	14	10	0.71	2	1	1	0	0	0	4.50
Kenley Jansen	2.0	29	20	0.69	3	1	1	0	1	1	3.00
Josh Fields	0.0	6	4	0.67	3	2	2	0	0	2	INF
Tony Cingrani	1.0	5	5	1.00	0	0	0	1	0	0	0.00
Brandon McCarthy	1.0	21	12	0.57	2	2	2	0	0	1	18.00

Game 3. Houston, Friday, October 27 (Dodgers 3, **Astros** 5)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	Series
											ERA
Yu Darvish	1.2	49	31	0.63	6	4	4	1	0	1	21.60
Kenta Maeda	2.2	42	28	0.67	1	0	0	1	2	0	0.00
Tony Watson	1.0	18	11	0.61	2	1	0	0	1	0	0.00
Brandon Morrow	0.2	13	8	0.62	1	0	0	1	2	0	3.38
Tony Cingrani	0.2	8	6	0.75	1	0	0	1	0	0	0.00
Ross Stripling	1.1	15	11	0.73	1	0	0	0	1	0	0.00

Game 4. Houston, Saturday, October 28 (**Dodgers** 6, Astros 2)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	ERA
Alex Wood	5.2	84	49	0.58	1	1	1	2	3	1	3.48
Brandon Morrow	1.1	14	9	0.64	0	0	0	0	0	0	1.46
Tony Watson	1.0	9	6	0.67	0	0	0	0	0	0	3.00
Kenley Jansen	1.0	14	9	0.64	1	1	1	0	1	1	4.50

Game 5. Houston, Sunday, October 29 (Dodgers 12, **Astros** 13, 10 innings)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	Series
											ERA
Clayton Kershaw	4.2	94	57	0.61	4	6	6	3	2	1	5.40
Kenta Maeda	0.2	25	15	0.60	2	1	1	1	1	1	1.93
Tony Watson	0.2	9	8	0.89	0	0	0	0	0	0	0.00
Brandon Morrow	0.0	6	5	0.83	4	4	4	0	0	2	11.25
Tony Cingrani	1.1	20	13	0.65	1	1	1	0	2	1	3.00
Ross Stripling	0.2	8	5	0.63	1	0	0	0	0	0	0.00
Kenley Jansen	1.2	33	19	0.58	2	1	1	1	1	0	4.76

Game 6. Los Angeles, Tuesday, October 31 (Astros 1, **Dodgers** 3)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	Series ERA
Rich Hill	4.2	58	45	0.78	4	1	1	1	5	1	2.08
Brandon Morrow	1.0	14	12	0.86	1	0	0	0	1	0	9.00
Tony Watson	0.1	12	4	0.33	0	0	0	1	0	0	0.00
Kenta Maeda	1.0	14	9	0.64	1	0	0	0	0	0	1.59
Kenley Jansen	2.0	19	18	0.95	0	0	0	0	3	0	3.52

Game Seven?

Yu Darvish [0.518 0.532 0.451 0.459 0.563 0.460 0.518 0.509 0.570]
Rich Hill [0.491 0.550 0.526 0.492 0.538 0.533 0.639 0.453 0.528]
Clayton Kershaw [0.504 0.566 0.444 0.628 0.534 0.522 0.521 0.433 0.615]
Kenta Maeda [0.480 0.518 0.510 0.479 0.541 0.457 0.452 0.463 0.510]
Alex Wood [0.511 0.475 0.508 0.575 0.532 0.530 0.589 0.507 0.494]
Jose Altuve [0.527 0.365 0.534 0.458 0.471 0.312 0.487 0.408 0.373]
Alex Bregman [0.602 0.428 0.499 0.508 0.604 0.443 0.443 0.527 0.462]
Carlos Correa [0.454 0.430 0.463 0.518 0.569 0.438 0.522 0.345 0.530]
Marwin Gonzalez [0.429 0.464 0.528 0.499 0.451 0.524 0.495 0.493 0.480]
Yulieski Gurriel [0.516 0.425 0.614 0.445 0.580 0.374 0.500 0.538 0.433]
Brian McCann [0.467 0.441 0.564 0.513 0.494 0.551 0.417 0.512 0.644]
Josh Reddick [0.508 0.485 0.605 0.490 0.478 0.571 0.412 0.501 0.544]
George Springer [0.531 0.398 0.391 0.486 0.557 0.408 0.554 0.391 0.506]

No words, just numbers

These are the neural network embeddings for the 2017 World Series. We go directly from events on the field to numbers describing pitching and hitting, numbers that may be used to predict future events on the field. Each player is represented by a vector of nine real numbers. Vectors for pitchers are distinct from vectors for batters. We consider five potential starting pitchers for the **Dodgers** and eight position players expected to be in the starting lineup for the **Astros**.

Pitcher	Batter	Out Fielder 1	Out Fielder 2	● ● ●	Strikeout	Walk
Yu Darvish	Jose Altuve	0.0189	0.0070		0.1646	0.0690
Yu Darvish	Alex Bregman	0.0122	0.0060		0.2634	0.0989
Yu Darvish	Carlos Correa	0.0158	0.0055		0.2820	0.1204
Yu Darvish	Marwin Gonzalez	0.0242	0.0081		0.2925	0.0689
Yu Darvish	Yulieski Gurriel	0.0184	0.0072		0.1673	0.0532
Yu Darvish	George Springer	0.0119	0.0050		0.2528	0.0963
Yu Darvish	Brian McCann	0.0165	0.0057		0.2124	0.0984
Yu Darvish	Josh Reddick	0.0136	0.0056		0.3471	0.1154
Rich Hill	Jose Altuve	0.0232	0.0062		0.1499	0.0738
Rich Hill	Alex Bregman	0.0153	0.0054		0.2447	0.1080
Rich Hill	Carlos Correa	0.0198	0.0049		0.2636	0.1323
Rich Hill	Marwin Gonzalez	0.0313	0.0075		0.2808	0.0778
Rich Hill	Yulieski Gurriel	0.0229	0.0064		0.1547	0.0579
Rich Hill	George Springer	0.0156	0.0047		0.2460	0.1102
Rich Hill	Brian McCann	0.0215	0.0054		0.2058	0.1121
Rich Hill	Josh Reddick	0.0171	0.0051		0.3237	0.1265
Clayton Kershaw	Jose Altuve	0.0246	0.0062		0.1548	0.0395
Clayton Kershaw	Alex Bregman	0.0165	0.0055		0.2559	0.0585
Clayton Kershaw	Carlos Correa	0.0215	0.0051		0.2783	0.0724
Clayton Kershaw	Marwin Gonzalez	0.0330	0.0075		0.2881	0.0414
Clayton Kershaw	Yulieski Gurriel	0.0237	0.0063		0.1556	0.0302
Clayton Kershaw	George Springer	0.0169	0.0048		0.2599	0.0603
Clayton Kershaw	Brian McCann	0.0232	0.0055		0.2158	0.0609
Clayton Kershaw	Josh Reddick	0.0186	0.0052		0.3432	0.0695
Kenta Maeda	Jose Altuve	0.0225	0.0069		0.1162	0.0517
Kenta Maeda	Alex Bregman	0.0154	0.0062		0.1959	0.0781
Kenta Maeda	Carlos Correa	0.0201	0.0057		0.2131	0.0966
Kenta Maeda	Marwin Gonzalez	0.0304	0.0083		0.2176	0.0545
Kenta Maeda	Yulieski Gurriel	0.0218	0.0069		0.1174	0.0397
Kenta Maeda	George Springer	0.0149	0.0052		0.1882	0.0761
Kenta Maeda	Brian McCann	0.0203	0.0058		0.1550	0.0762
Kenta Maeda	Josh Reddick	0.0178	0.0060		0.2684	0.0947
Alex Wood	Jose Altuve	0.0272	0.0058		0.0997	0.0536
Alex Wood	Alex Bregman	0.0192	0.0054		0.1734	0.0835
Alex Wood	Carlos Correa	0.0249	0.0050		0.1873	0.1025
Alex Wood	Marwin Gonzalez	0.0387	0.0074		0.1969	0.0595
Alex Wood	Yulieski Gurriel	0.0266	0.0060		0.1019	0.0416
Alex Wood	George Springer	0.0194	0.0047		0.1737	0.0849
Alex Wood	Brian McCann	0.0263	0.0053		0.1425	0.0847
Alex Wood	Josh Reddick	0.0221	0.0052		0.2372	0.1011

Event probability distributions

Fifty-one distinct events were observed in the training data, beginning with an out first fielded by the player at position 1 (the pitcher) and ending with a walk. So there are fifty-one event probabilities for each pitcher-batter combination. These estimated probability distributions flow from the neural network embeddings.

Expected event probabilities can then be used to drive game-day simulations.

Who should pitch for the Dodgers in game seven?

DODGER PITCHER	DAYS RESTED	SERIES ERA	EXPECTED ERA
Clayton Kershaw	2	5.40	3.10
Yu Darvish	4	21.6	4.02
Rich Hill	0	2.08	4.65
Alex Wood	3	3.48	4.73
Kenta Maeda	0	1.59	4.74

Simulation results

To compute the expected earned run average (ERA) for a Dodgers pitcher, we run game-day simulations on events selected at random from estimated event probability distributions. Estimates come from each pitcher-batter match-up using neural network embeddings. In running simulated games for each pitcher, we make a few simplifying assumptions, such as Astros pitchers make outs when batting, there are no stolen bases, no sacrifices, and all errors are one-base errors. A single advances all runners one base, a double two bases, and so on. Summary results are for 2,000 simulated games for each pitcher.

Game Seven Results

Yu Darvish was removed from the game in the second inning. Kershaw and Wood performed well in relief roles, as did Morrow and Jensen. The Houston Astros win the series.

Game 7. Los Angeles, Wednesday, November 1 (**Astros** 5, Dodgers 1)

Pitcher	IP	Pitches	Strikes	S/P	H	R	ER	BB	K	HR	Series ERA
Yu Darvish	1.2	47	30	0.64	3	5	4	1	0	1	21.60
Brandon Morrow	0.1	3	3	1.00	0	0	0	0	1	0	8.44
Clayton Kershaw	4.0	43	34	0.79	2	0	0	2	4	0	4.02
Kenley Jansen	1.0	20	12	0.60	0	0	0	1	1	0	3.12
Alex Wood	2.0	25	18	0.72	0	0	0	0	3	0	1.17

Winning the Game with Methods and Models

Not to impugn the gods of Sabermetrics, but maybe we have learned as much as we can from arbitrary measures and spreadsheet formulas. Suppose we use neural networks to go directly from events on the field to numbers that predict future events on the field.

Better algorithms

Machine learning begets neural networks, neural networks beget deep learning, and deep learning fosters artificial intelligence. If we can learn so much using a simple neural network, imagine what may be possible with deeper networks and better algorithms.

More data

If we can do so much with public-domain event data, imagine the competitive value afforded by the full range of Statcast data.

The right people

Having the right players on the team is a primary source of competitive advantage. This applies for teams on and off the field. The good news for team owners:

There are no salary caps for analysts and data scientists.

About Me

`thomas-miller-0@northwestern.edu`

I am the faculty director of the data science program at Northwestern University. I have developed and taught many courses in the program, including *Practical Machine Learning*, *Web and Network Data Science*, and *Information Retrieval and Real-time Analytics*. I also consult with businesses about performance and value measurement, data science methods, information technology, and best practices for building teams of data scientists and data engineers. Among my books about data science, *Sports Analytics and Data Science: Winning the Game with Methods and Models* reviews many concepts from this presentation.

Data science at Northwestern University

`http://sps.northwestern.edu/program-areas/graduate/data-science/index.php`

