CS46N:

Big Data and Baseball

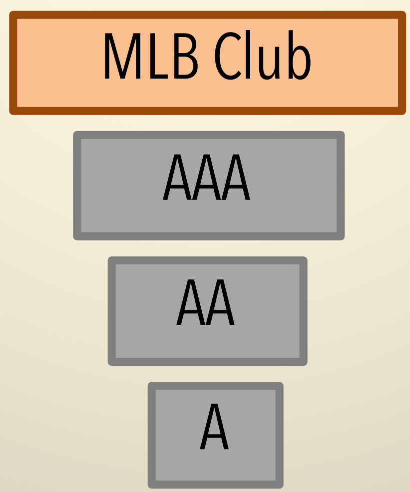
Alec Powell

Stanford CS '16

December 1, 2015

The Ecosystem of Baseball

- Major league baseball (MLB) clubs develop minor league players (younger, less experienced) in a "farm system"
- Parent club has control of a drafted player for 7 seasons once he makes the major leagues



The Case

Diamond Dollars Case Competition

presented by

CRAY
THE SUPERCOMPUTER COMPANY

What is the Perfect Cole Hamels Trade?

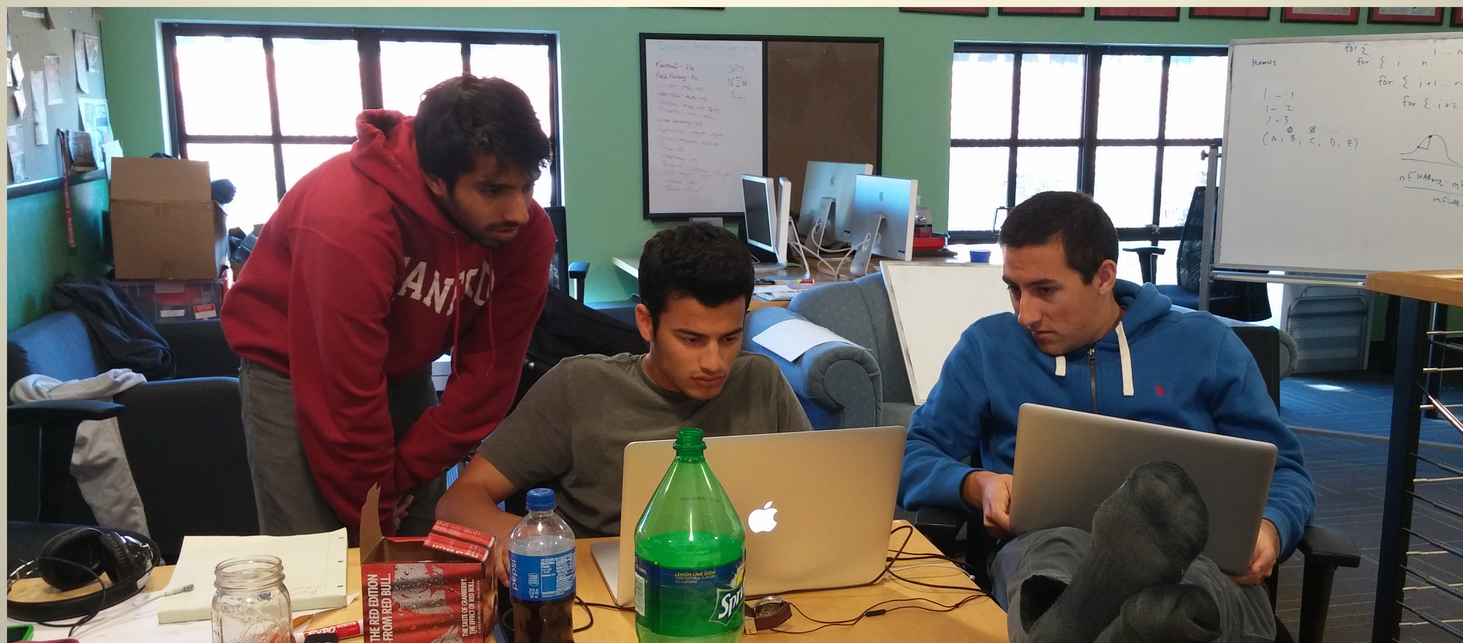
"People in all fields operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage."

— Michael Lewis, Moneyball: The Art of Winning an Unfair Game

SABR Stanford University

About Us

We're three juniors, a sophomore and a freshman from varying educational backgrounds united by our love for baseball and joined by the Stanford Sports Analytics Club.



Our Approach

Three main questions:

- How do we project the performance of Hamels and traded prospects over the coming seasons?
- How do we quantify the improvement of a team upon the addition of Hamels?
- Can we develop a quantitative metric to evaluate the value of a trade to both the Phillies and their trade partner, and how do we optimize that metric?

SABR Stanford University

Our Approach

Three main questions:

- How do we project the performance of Hamels and traded prospects over the coming seasons?
- How do we quantify the improvement of a team upon the addition of Hamels?
- Can we develop a quantitative metric to evaluate the value of a trade to both the Phillies and their trade partner, and how do we optimize that metric?

What Data?

Baseball stats from Fangraphs, baseball-reference.com:

- Wins Above Replacement (WAR)
- Weighted On-base Average (wOBA)
- Fielding-Independent Pitching (FIP)

Demographic information:

- Age
- Minor/Major League Level
- Service Time

Our Database

We have:

- Minor league player-seasons data (one long Excel file…)
- List of Baseball America Top-10 rated prospects for each trade partner team
- Age, career WAR, and service time for each player on each MLB team

Projecting Minor League Prospects

How do we project the future performance of minor league prospects?

In a trade for Cole Hamels, it is likely that the Phillies will want one or multiple "top" prospects in return.

For the sake of simplicity, we only considered the top 10 prospects on each team (according to Baseball America preseason rankings).

SABR Stanford University

Projecting Minor League Prospects

1. Accumulate a database of every minor league player's seasons from 2010-2014 (all levels, all players)
2. For every 2015 top-10 prospect… find similar historical minor league seasons (age, level, wOBA/FIP)
3. Narrow down by only considering similar seasons from former top-10 minor league prospects
4. For every player with a "similar" former season, find career MLB seasons played, career MLB fWAR
5. For each of those old players, average MLB fWAR per season to predict future of our 2015 prospect
6. Return projected career fWARs for every 2015 prospect

Stanford University

Projecting Minor League Prospects

As input, the program takes the list of top-10 prospects for a trade partner team

The program finds comparable statistical seasons to each prospect's 2014 season adjusted for **age** (i.e., 22) and **level** (i.e., AA)

- "Comparable" season defined to be +/- 5% of comparison statistic
- Batters compared using wOBA
- Pitchers compared using FIP

SABR Stanford University

An Aside: Similarity Algorithms

- Jaccard similarity
- Cosine similarity
- K-nearest neighbors
- K-means clustering

Applications: Recommender systems, classfication, & more

Our Approach

Three main questions:

- How do we project the performance of Hamels and traded prospects over the coming seasons?
- How do we quantify the improvement of a team upon the addition of Hamels?
- Can we develop a quantitative metric to evaluate the value of a trade to both the Phillies and their trade partners, and how do we optimize that metric?

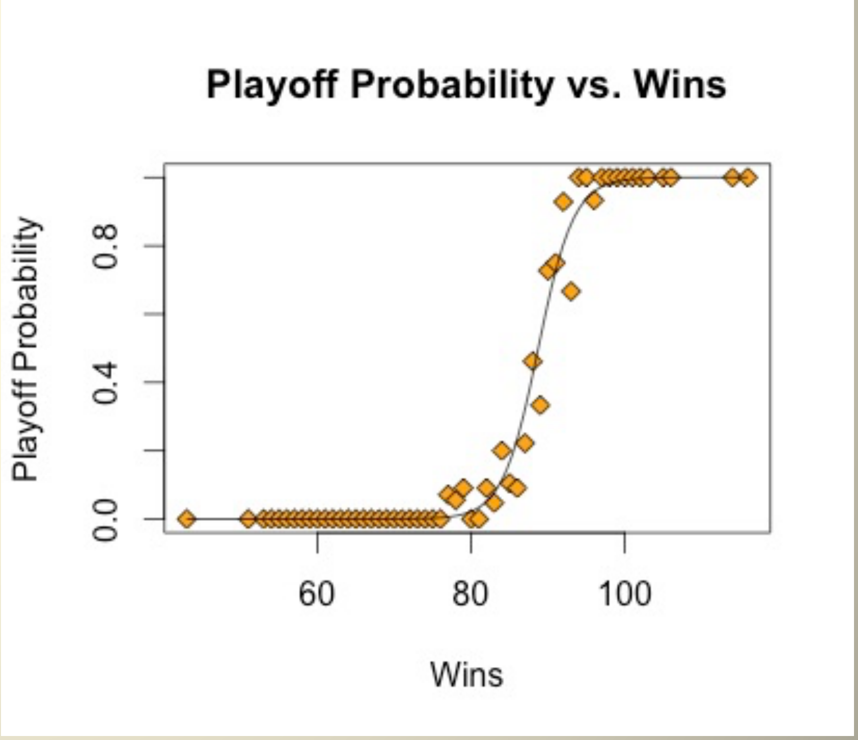
Correlating Wins to Playoffs

Compiled every season from 1995-2014 and noted which teams made the playoffs, aggregated into logistic regression

$$\text{Playoff probability} = \frac{1}{1 + e^{-(A + (B \times \text{wins}))}}$$

A = -60.25773
B = 0.69183

Observation: very short "sweet spot" where probability changes dramatically with wins



The Hamels Effect

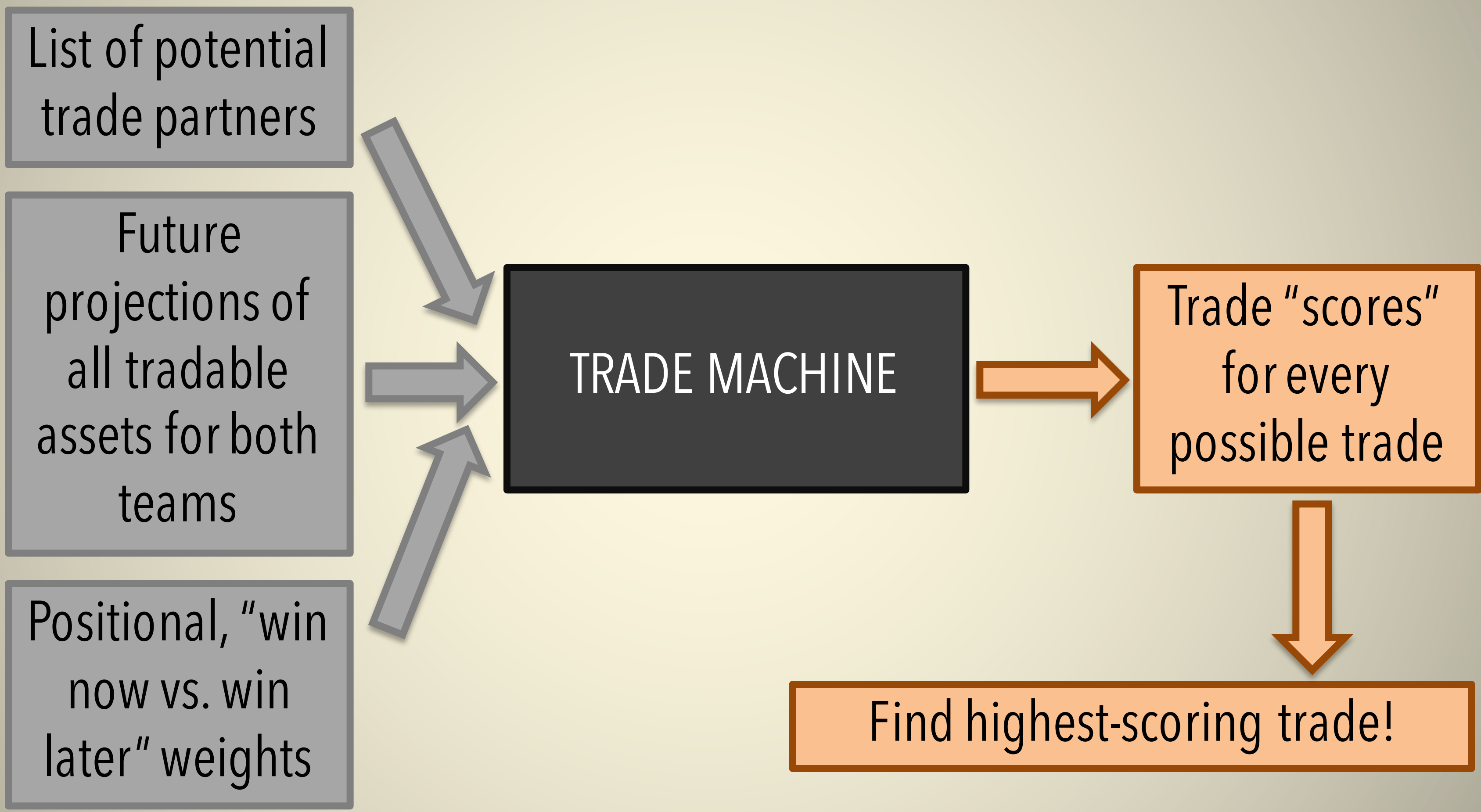
1. **Cleveland Indians** — Change: +46.6%
   Before Hamels: 85.5 W / 24.6 % playoffs
   With Hamels: 88.4 W / 71.2 % playoffs
2. **Toronto Blue Jays** — Change: +46.3%
   Before Hamels: 86.2 W / 34.6 % playoffs
   With Hamels: 89.2 W / 80.9 % playoffs
3. **Detroit Tigers** — Change: +44.4%
   Before Hamels: 86.3 W / 36.2 % playoffs
   With Hamels: 89.2 W / 80.6 % playoffs
4. **New York Yankees** — Change: +38.9%
   Before Hamels: 85.2 W / 21.1 % playoffs
   With Hamels: 87.7 W / 60.0 % playoffs
5. **Seattle Mariners** — Change: +37.8%
   Before Hamels: 84.9 W / 15.9 % playoffs
   With Hamels: 87.3 W / 53.7 % playoffs
6. **Los Angeles Angels** — Change: +31.9%
   Before Hamels: 83.7 W / 8.6 % playoffs
   With Hamels: 86.5 W / 40.5 % playoffs

Our Approach

Three main questions:

- How do we project the performance of Hamels and traded prospects over the coming seasons?
- How do we quantify the improvement of a team upon the addition of Hamels?
- Can we develop a quantitative metric to evaluate the value of a trade to both the Phillies and their trade partners, and how do we optimize that metric?

The Trade Machine



The Trade Machine

Computes scores for Phillies and other team by simulating player exchanges between teams

For each player traded from team A to B

- Subtract that player's value to team A from Score A
- Add that player's value to team B from Score B

Value to each team differs based on weights

In the end, we have two scores: S_A, S_B

The Trade Machine

Our goal:

Maximize $S_A + S_B$

(ensures maximum utility for both teams)

subject to $| S_A - S_B | < t$, where t is a threshold

(ensure that the trade is fair to both teams)

SABR Stanford University

The Trade Machine: Example Input

	A	B	C	D	E	F	G
1	Player	Position	Level	fWAR15	fWAR16	fWAR17	fWAR18
2	Dellin Betances	P	MLB	1.5	1.4	1.3	1.2
3	David Carpenter	P	MLB	0.4	0.3	0.3	0.2
4	Nathan Eovaldi	P	MLB	1.7	1.8	1.9	1.9
5	Didi Gregorius	SS	MLB	1.8	2.0	2.2	2.2
6	Bryan Mitchell	P	MLB	-0.9	-0.8	-0.7	-0.7
7	Michael Pineda	P	MLB	1.5	1.5	1.5	1.4
8	Austin Romine	C	MLB	0.3	0.3	0.3	0.3
9	Chasen Shreve	P	MLB	0.2	0.2	0.3	0.3
10	Masahiro Tanaka	P	MLB	3.2	3.2	3.1	2.9
11	Adam Warren	P	MLB	0.5	0.5	0.4	0.4
12	Chase Whitley	P	MLB	0.2	0.2	0.2	0.2
13	Justin Wilson	P	MLB	-0.2	-0.2	-0.2	-0.3
14	Luis Severino	P	AA	2.5	2.7	2.9	4.1
15	Greg Bird	1B	AA	1.2	1.3	1.9	2.6
16	Gary Sanchez	C	AA	-2.7	-2.5	-1.3	-0.8
17	Ian Clarkin	P	A+	1.2	1.4	1.5	1.7
18	Rob Refsnyder	2B	AAA	0.8	1.1	1.3	1.3
19	Jacob Lindgren	P	AA	1.5	1.7	2.3	3.2
20	Miguel Andujar	3B	A	-0.1	-0.1	-0.1	-0.1

List of tradable MLB assets and prospects

Projected WARs, 2015-18

Trade 1

Phillies get:

Ketel Marte
(SS, AAA)

D.J. Peterson
(3B, AA)

Taijuan Walker
(RHP, MLB)

Patrick Kivlehan
(3B, AA)

Edwin Diaz
(RHP, A)

Mariners get:

Cole Hamels
(LHP, MLB)

Darin Ruf
(1B, MLB)

Phillies Score: +2.010

Mariners Score: +2.104

Total Trade Score: +4.114

Stanford University

Trade 2

Phillies get:

Gary Sanchez
(C, AA)

Greg Bird
(1B, AA)

Didi Gregorius
(SS, MLB)

Rob Refsnyder
(2B, AAA)

Jacob Lindgren
(LHP, AA)

Yankees get:

Cole Hamels
(LHP, MLB)

Ben Revere
(OF, MLB)

Phillies Score: +2.692

Yankees Score: +2.784

Total Trade Score: +5.476

SABR

Stanford University

Data + Other Sports

- Football
  - Impact of weather
  - Bayesian draft analysis
  - Fantasy Football Machine Learning (yours truly)
- Basketball
  - Player efficiency rating
  - Advanced defensive metrics
- Soccer
  - Predictive shot-taking
- Hockey
  - Elo ratings
  - Offensive line shift productivity
- Join Sports Analytics!!

Stanford University

Big Data/CS Classes To Take:

- CS145 – Databases
- CS124 – Natural Language Processing
- CS246 – Mining Massive Datasets
- CS221 – Artificial Intelligence
- CS229 – Machine Learning

Thank you!