Sports Analytics Summer Research Lab:

Spatial and Temporal Modeling in Sports

Shane T. Jensen

Department of Statistics, The Wharton School

stjensen@wharton.upenn.edu

June 14, 2023

Sports data is advancing statistical methodology

An exciting time for sports analytics: we are at the cutting edge of high resolution **spatial and temporal data situations**

I will present work by myself and others on statistical models for **variation over space and time**, while highlighting the **thought process** behind different modeling choices

Ongoing challenges in developing models that address the complexity of these high resolution data situations while still being **computationally feasible** to estimate at scale

Topics I am planning to cover

Similar thought processes and approaches can be applied to **spatio-temporal data situations** across different sports and across different time or spatial scales

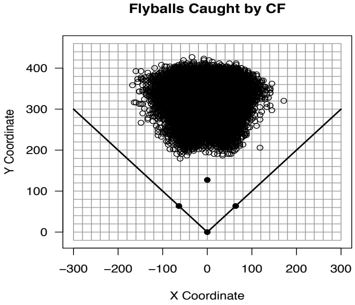
1. Hierarchical spatial models for fielding in **baseball**
2. Spatial and temporal modeling in **basketball**
3. Aging trajectories in **baseball**
4. Drafting stategy in **football**

Quantifying Fielding Performance in Baseball

- **Overall goal**: accurate evaluation of the fielding performance of each major league baseball player
- Many aspects of game (eg. hitting, pitching) are easy to quantify and tabulate
  - finite number of outcomes, baserunner configurations
- Fielding is a more **continuous** aspect of the game
  - presents a greater data and modeling challenge
- **Probit functions** used to model curves for probability of a successful fielding play

Baseball Info Solutions (BIS) Data

- **Ball-in-play data** available via Baseball Info Solutions
- 7 seasons (02-08) with 120000 balls-in-play (BIP) per year
  - Three BIP types: 42% grounders, 33% flys, 25% liners
- Each BIP is mapped to a more resolute area than zones of previous methods
- BIP **velocity** information as ordinal category



Bernoulli Successes and Failures

- The outcome of each play is either a **success or failure**:

$$S_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ BIP hit to the } i^{th} \text{ player leads to out} \\ 0 & \text{if the } j^{th} \text{ BIP hit to the } i^{th} \text{ player leads to hit} \end{cases}$$

- Observed successes and failures are modeled as **Bernoulli realizations** from an underlying probability:

$$S_{ij} \sim \mathrm{Bernoulli}(p_{ij})$$

- Each p_{ij} is a function of available data for that BIP:
  - $(x, y)_{ij}$ location, velocity V_{ij} and type of the BIP
- These probability functions will be **smooth parametric curves** that can vary between different players

Thought Process: Spatial Modeling

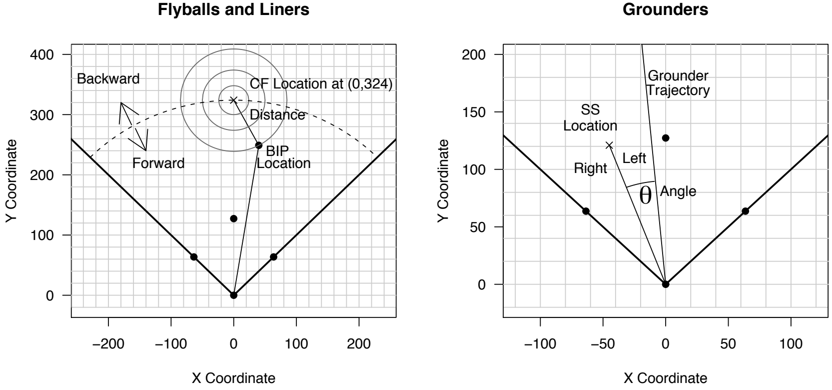
We have spatial data for every single BIP, how do we convert (x, y) coordinates into variables we can use to model **probability that BIP is successfully fielded**?

Careful thought required to ensure these **created variables** capture the relevant information in our spatial data

We do not have direct data on time, but we do have our rough measure of **velocity**

Representation for Different BIP Types

- **Two-dimensional** curves needed for **flys/liners**: success depends on velocity, direction and **distance** to BIP
- **One-dimensional** curves needed for **grounders**: success depends on velocity, direction and **angle** to BIP



Probit function for each smooth curve

- **Probit regression** used to model smooth curves for probability p_{ij} of successfully fielding BIP j by player i

  $\Phi(\cdot)$ = CDF of Normal distribution

- **Probit function for fly-balls/liners:**

$$\begin{aligned} p_{ij} &= \Phi(\beta_{i0} + \beta_{i1}\, D_{ij} + \beta_{i2}\, D_{ij}\, F_{ij} + \beta_{i3}\, D_{ij}\, V_{ij} + \beta_{i4}\, D_{ij}\, V_{ij}\, F_{ij}) \\ &= \Phi(\boldsymbol{X}_{ij} \cdot \boldsymbol{\beta}_i) \end{aligned}$$

  D_{ij} = distance to BIP, V_{ij} = vel, $F_{ij} = 1$ if forward (vs. back)

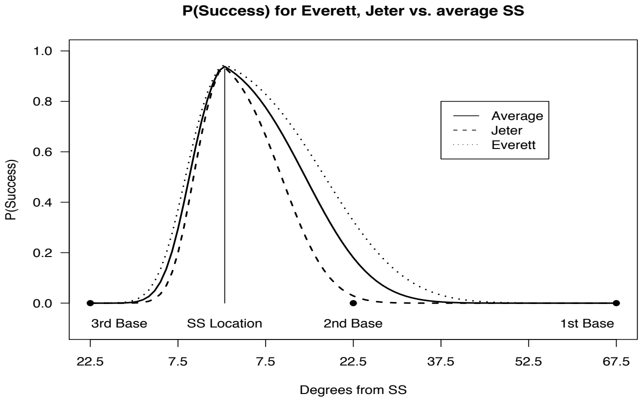
- **Probit function for grounders:**

$$\begin{aligned} p_{ij} &= \Phi(\beta_{i0} + \beta_{i1}\, \theta_{ij} + \beta_{i2}\, \theta_{ij}\, L_{ij} + \beta_{i3}\, \theta_{ij}\, V_{ij} + \beta_{i4}\, \theta_{ij}\, V_{ij}\, L_{ij}) \\ &= \Phi(\boldsymbol{X}_{ij} \cdot \boldsymbol{\beta}_i) \end{aligned}$$

  θ_{ij} = angle to BIP, V_{ij} = velocity, $L_{ij} = 1$ if left (vs. right)

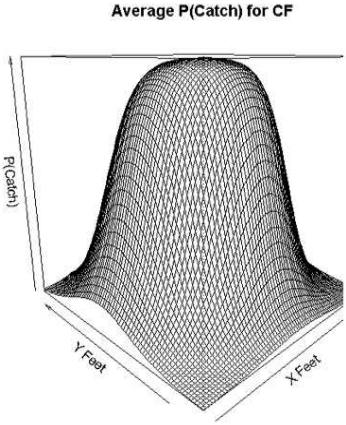
Individual Models for Grounders

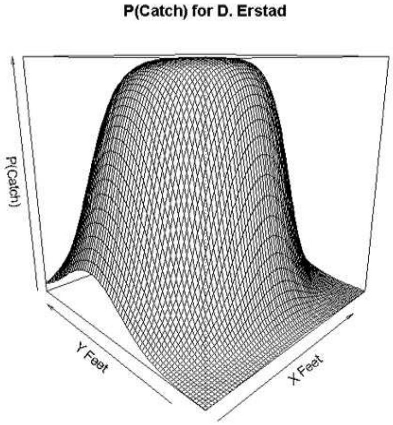
- Can fit MLE coefficients $\widehat{\beta}_i$ of **player-specific model** using only data for player i
- Compare infielders based on MLE curves for grounders



Individual Models for Fly/Liners

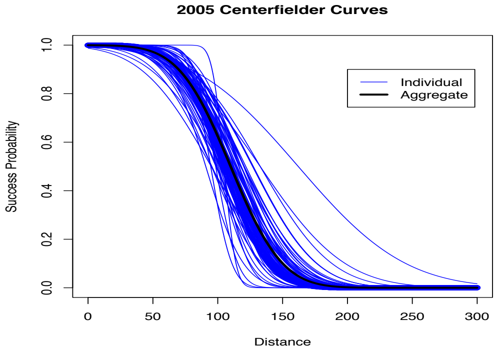
- Can again fit MLE coefficients $\widehat{\beta}_i$ of **player-specific model** using only data for player i
- Compare outfielders based on MLE curves for flys or liners





Problems with MLE Estimation

- Ideally want **curve variability**, not just MLE estimate
- Small samples for some players leads to **unstable estimates** of individual curves



Thought Process: Sharing Information

How do we capture any common information between fielders while still allowing individual fielders to stand out if they are truly exceptional?

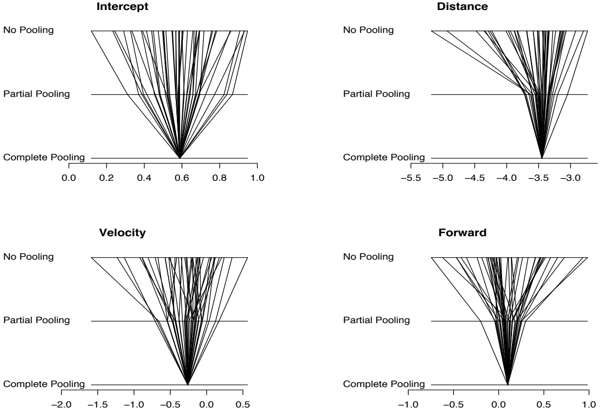
Need a compromise between two extreme modeling strategies:

1. Model each player **completely separately** (done up to now)
2. Model each player **as identical** (no individual differences)

Bayesian hierarchical models are designed to **shrink** fielders towards each other while still allowing individual differences

Examining Shrinkage of Player Coefficients

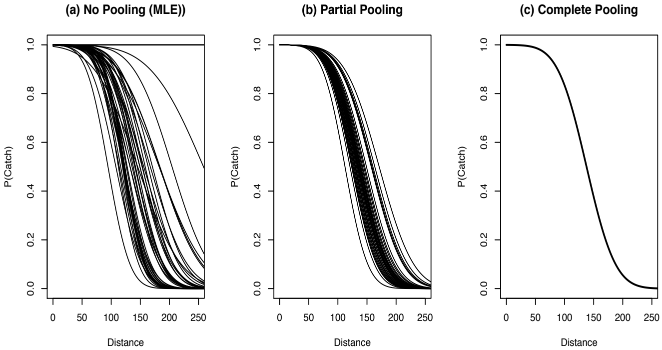
- Hierarchical model **shrinks** each β_i towards population μ



- No pooling = MLE, Partial pooling = Hierarchical model, Complete pooling = μ

Examining Shrinkage of Player Curves

- We can also examine shrinkage of player-specific curves



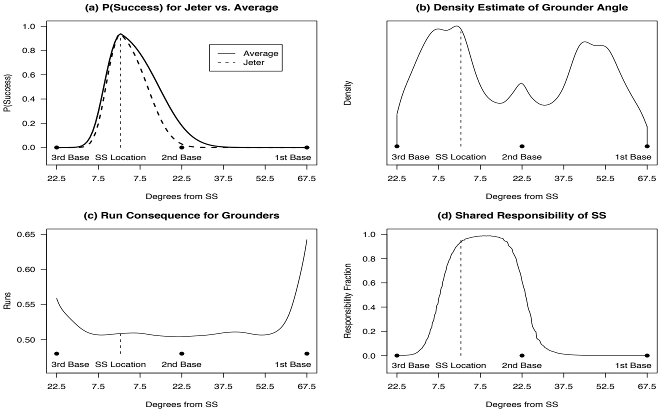
- **Amount of shrinkage** depends on sample size n_i and variation in performance through $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$
- Still **heterogeneity between players** even after shrinkage

Numerical Summary of Overall Performance

- Beyond comparing curves between players, we can derive an **overall numerical estimate** of fielder performance
  - **SAFE: Spatial Aggregate Fielding Evaluation**
- For each player, aggregate differences between individual curve (based on β_i) and overall curve (based on μ)
  - Aggregation done by **numerical integration** over fine grid of values (1D grid for grounders, 2D grid for flys/liners)
- Can calculate SAFE for each sample from posterior distribution of β_i, giving us the **posterior mean** and **95% posterior interval** of SAFE for each player

Differential Weighting in SAFE

- Our full aggregation also weights grid points by **BIP frequency**, **run value**, and **shared consequence**



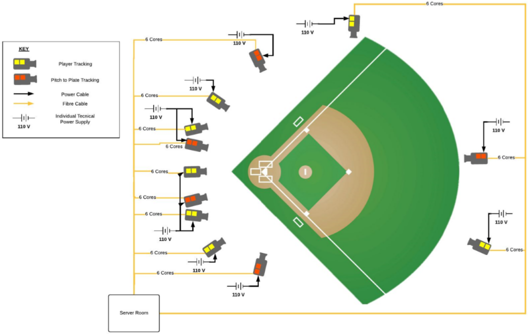
- SAFE value: **runs saved/cost** of fielder vs. average

Results for Middle Infielders: Best/Worst Posterior SAFE values

Ten Best 2B Player-Years			Ten Best SS Player-Years		
Name and Year	Mean	95% Interval	Name and Year	Mean	95% Interval
Julius Matos , 2002	18.1	(12.4 , 22.1)	Pokey Reese , 2004	22.6	(12.0 , 31.2)
Erick Aybar , 2007	17.6	(10.0 , 24.6)	**Adam Everett , 2007**	20.4	(10.4 , 27.4)
Junior Spivey , 2005	14.5	(4.7 , 27.1)	**Adam Everett , 2006**	17.1	(9.0 , 21.8)
Tony Graffanino , 2006	14.1	(4.6 , 27.6)	Craig Counsell , 2006	14.7	(6.9 , 21.1)
Adam Kennedy , 2008	11.3	(1.7 , 18.6)	Jorge Velandia , 2003	14.2	(3.0 , 24.0)
Willie Bloomquist , 2005	10.9	(4.3 , 17.8)	Alex Cora , 2005	14.1	(3.0 , 24.6)
Jose Valentin , 2006	10.9	(4.2 , 17.9)	Alex Rodriguez , 2003	13.5	(3.5 , 24.4)
Chase Utley , 2008	10.8	(5.7 , 17.5)	Maicer Izturis , 2004	13.2	(3.8 , 22.2)
Chase Utley , 2005	10.8	(3.1 , 17.7)	Marco Scutaro , 2008	13.0	(4.0 , 20.1)
Craig Counsell , 2005	10.8	(5.3 , 18.0)	Brent Lillibridge , 2008	11.8	(5.0 , 19.1)

Ten Worst 2B Player-Years			Ten Worst SS Player-Years		
Name and Year	Mean	95% Interval	Name and Year	Mean	95% Interval
Ronnie Belliard , 2008	-9.8	(-19.5 , 2.6)	Erick Almonte , 2003	-13.8	(-26.9 , 2.3)
Geoff Blum , 2005	-10.2	(-17.5 , -1.7)	**Derek Jeter , 2007**	-13.9	(-21.7 , -5.8)
Miguel Cairo , 2004	-10.9	(-17.9 , -3.1)	Michael Morse , 2005	-14.2	(-23.0 , -4.5)
Terry Shumpert , 2002	-11.0	(-22.2 , 0.7)	Damian Jackson , 2005	-14.5	(-30.6 , -3.5)
Roberto Alomar , 2003	-12.1	(-19.3 , -4.6)	Brandon Fahey , 2008	-15.1	(-22.4 , -8.2)
Enrique Wilson , 2004	-12.3	(-18.9 , -6.2)	Marco Scutaro , 2006	-15.1	(-22.0 , -10.0)
Alberto Callaspo , 2008	-12.4	(-20.4 , -4.5)	**Derek Jeter , 2003**	-15.6	(-24.8 , -6.4)
Dave Berg , 2002	-13.5	(-25.1 , -2.4)	Michael Young , 2004	-15.6	(-23.6 , -7.2)
Luis Rivas , 2002	-13.8	(-20.9 , -6.4)	Josh Wilson , 2007	-15.8	(-26.5 , -6.4)
Bret Boone , 2005	-15.4	(-22.4 , -8.1)	**Derek Jeter , 2005**	-18.5	(-29.1 , -9.2)

Hawk-Eye Statcast system: 12 cameras around the park for full-field optical pitch, hit, and player tracking



This new system provides **trajectories** for each ball in play as well as **starting positions** and movement of each fielder

Higher Resolution Fielding Data

With this higher resolution video-based fielding data, we could create better **spatial** (true distance traveled) and **temporal** (hang time) variables for our binary regression model

Recent rule change to limit shifting creates a **natural experiment** for studying the effects of defensive positioning

Next up, we will examine recent work to harness **high resolution spatio-temporal data** in basketball

Optical tracking data is also being used in basketball to create more detailed measures of what is happening on the court

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 514, 585–599, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1141685

Taylor & Francis
Taylor & Francis Group

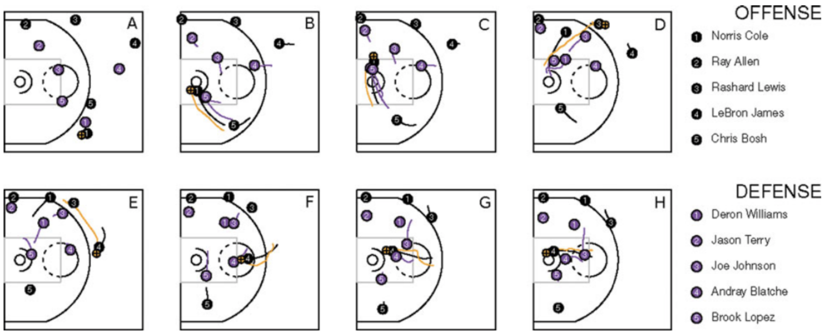
A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes

Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry

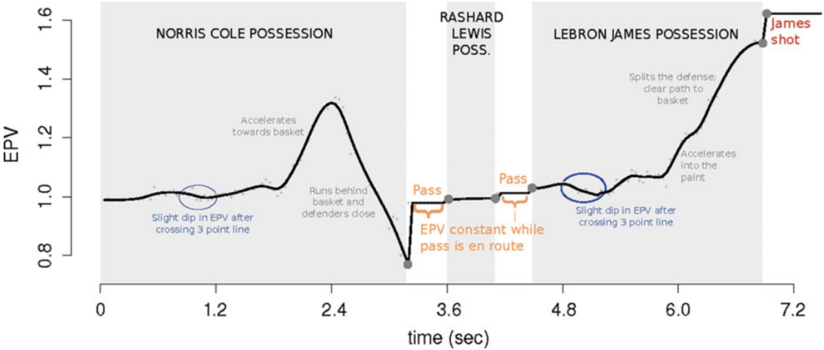
Can evaluate players on their **real time decisions and outcomes** at a very high level of **temporal resolution**

Optical player tracking data

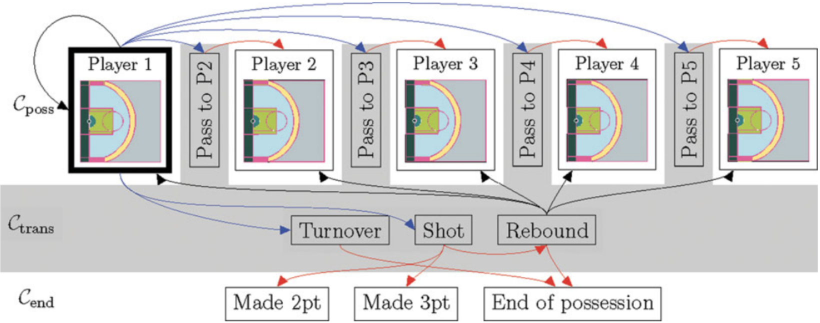
Framework for using **optical player tracking data** to estimate the **expected number of points** obtained by the end of a possession



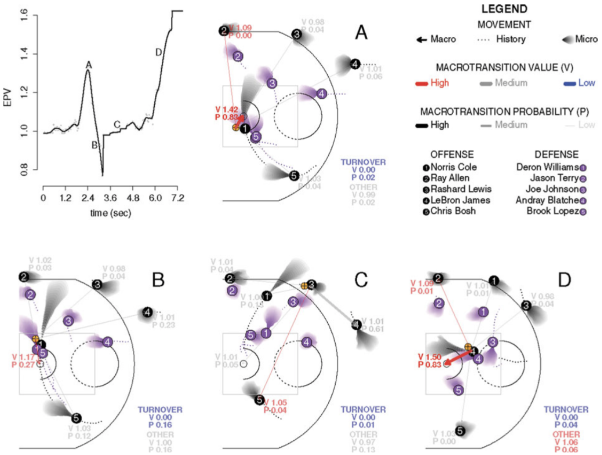
Expected possession value (EPV) derives from a stochastic process model for the **evolution of a basketball possession**



Multiple levels of modeling used to differentiate between **continuous movements** of players and **discrete events** such as shot attempts and turnovers



Challenge of High Resolution Modeling



Great approach to evaluating real time decision making and outcomes but implementation is **computationally challenging**

Thought Process: Different Time Scales

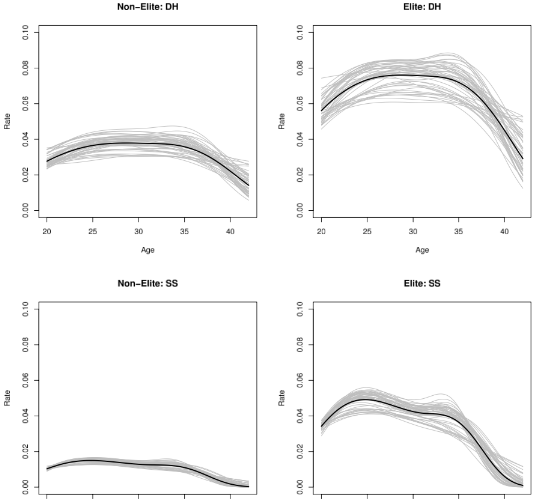
It is very common to **transfer models or techniques** between different sports that have similar spatial data situations

We can also adapt similar regression and hidden Markov modeling techniques to very **different time scales**:

1. Late Career Aging Trajectories in **baseball**
2. Early Career Draft Stategy in **football**

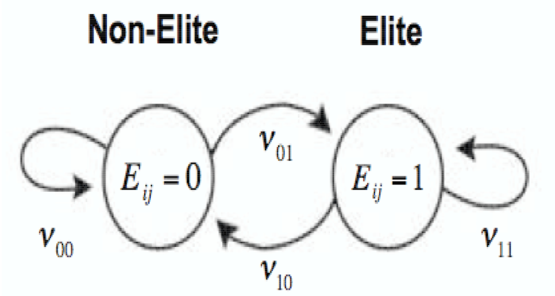
Modeling Career Trajectories in Baseball

Another area of current research is prediction over much longer time scales, e.g. **career trajectories** of baseball players



Cubic B-splines provide a flexible model for **career trajectory shapes** that can differ by position

A **hidden Markov model** is also used to allow for players to transition between elite and non-elite performance



In addition to late career performance, it is also important to predict **early career** performance, especially when making **drafting decisions**

DE GRUYTER JQAS 2014; 10(4): 381–396

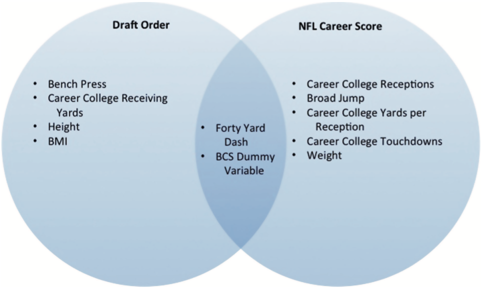
Jason Mulholland and Shane T. Jensen*

Predicting the draft and career success of tight ends in the National Football League

In **football**, how should we balance college performance vs. other data like the NFL combine?

Drafting of Tight Ends in Football

Regression models for **NFL Draft** versus **early NFL career** of tight ends based on college, combine, and physical measures



Current drafting decisions are NOT optimally calibrated in terms of using the **best predictors** of early career NFL performance

Common Thought Processes

Careful thought and **subject domain knowledge** required to create variables from spatial data that can be used for modeling

Hierarchical models capture common performance across players but still allow exceptional players to distinguish themselves

Modeling strategies can be transferred between **different sports** as well as adapted for **different time scales** of interest

Looking Forward

Statistical models provide a principled way to address the complexity in high resolution **spatial and temporal** data

An ongoing challenge is the **computational feasibility** of model estimation and interpretation. We need methods that can handle the increasingly complex data that will become available.

An interesting future challenge will be adapting to these methods to **other sports situations**: larger playing surfaces (soccer) and faster play action (hockey). Or both (e-sports)!

Wharton Moneyball Show and Podcast

New show every week on Sirius XM 132 or as Podcast!

Papers discussed in this talk

Jensen, S.T., Shirley, K., and Wyner, A.J. (2009). **Bayesball: a Bayesian hierarchical model for evaluating fielding in major league baseball.** Annals of Applied Statistics 3:491-520

Cervone, D., DAmour, A., Bornn, L, and Goldsberry, K. (2016). **A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes.** Journal of the American Statistical Association 111: 585-599

Jensen, S.T., McShane, B. and Wyner, A.J. (2009). **Hierarchical Bayesian modeling of hitting performance in baseball.** Bayesian Analysis 4:631-674

Mulholland, J. and Jensen, S.T. (2014). **Predicting the Draft and Career Success of Tight Ends in the National Football League.** Journal of Quantitative Analysis of Sports 10:381-396