Take Me Out to (Analyze) the Ballgame

Visualization and Analysis Techniques for Big Spatial Data

Chris Comiskey

Oregon State University

August 31, 2017

Baseball, Baseball, Baseball

- Chris, rookie year

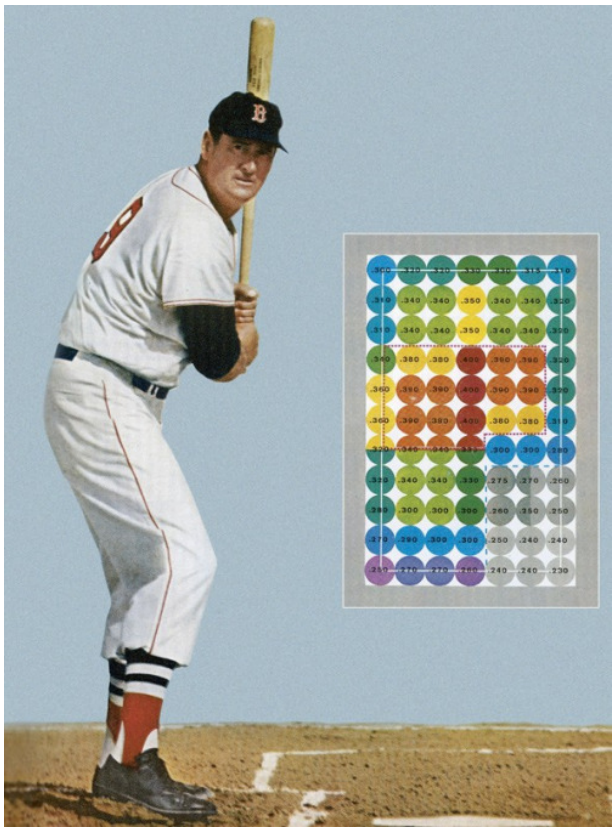
Baseball, Baseball, Baseball

- Chris, rookie year

- Chris, Boston Red Sox

Hitting Analytics, Then and Now

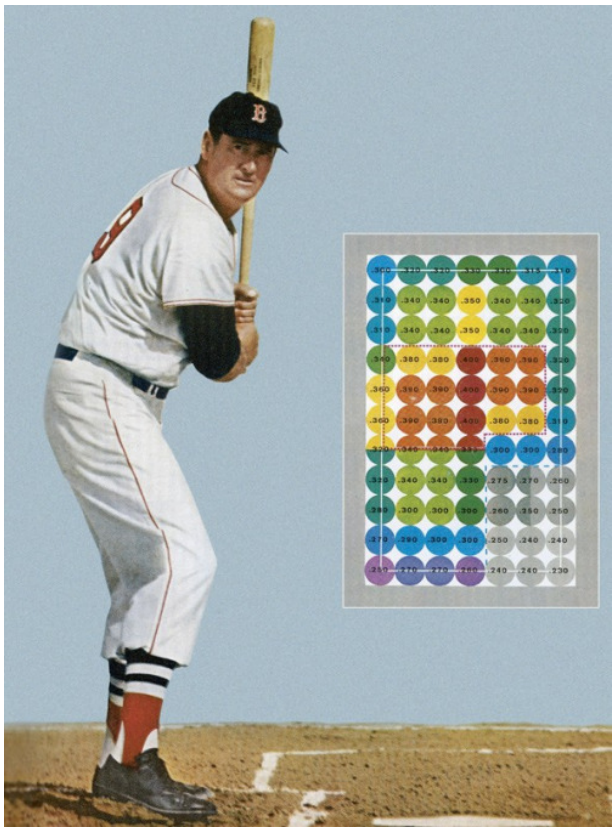
- "The Science of Hitting"$_{1970}$



- Conceptual breakthrough
- No data

Hitting Analytics, Then and Now

- "The Science of Hitting"$_{1970}$



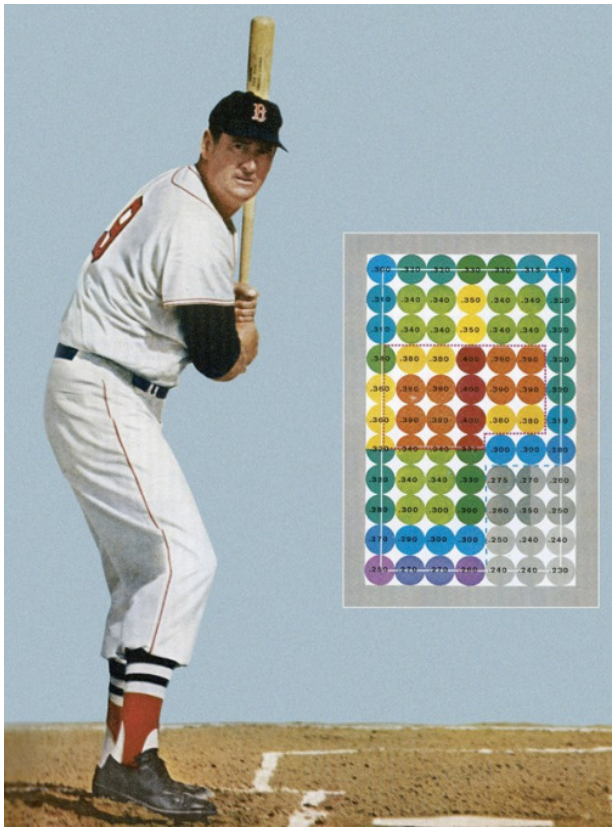
- Iconic breakthrough, hitter
- No data

- Hall of Fame exhibit

- Baseball fanatic statistician
- PITCHf/x data!

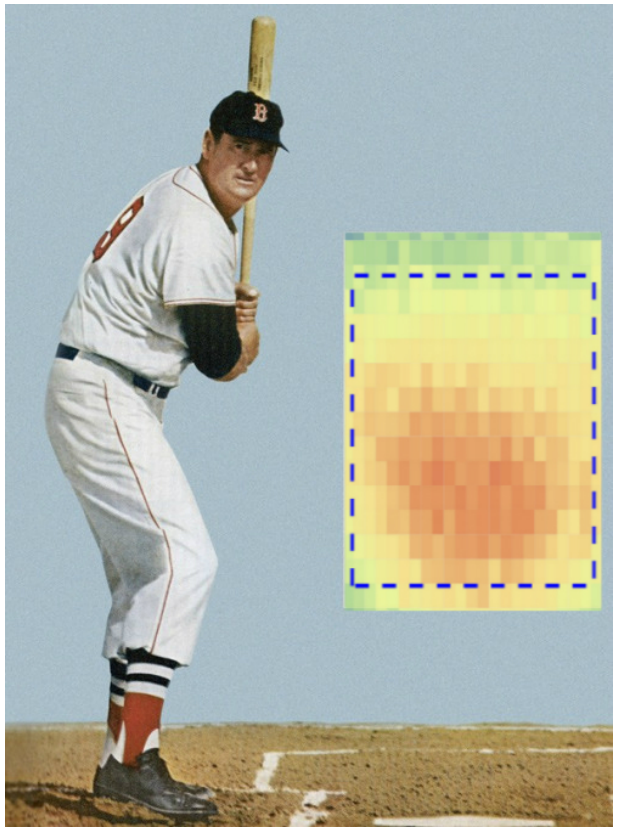
Hitting Analytics

- "The Science of Hitting"$_{1970}$



- Iconic breakthrough, hitter
- No data

- The Statistics of Hitting



- PITCHf/x data, R, heat maps
- SGLMMs, Stan, PPMs, INLA

Relevant Research

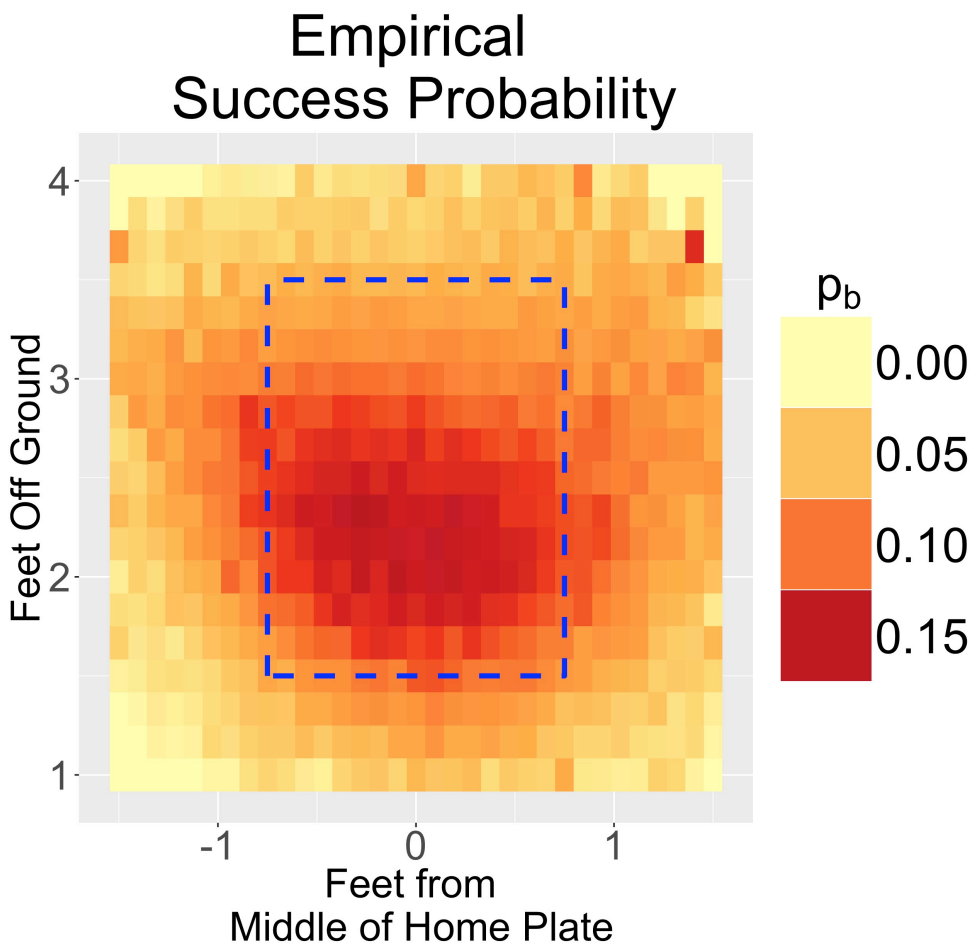
- Each MLB® team spends $\approx$ \$15 million/year
- Each MLB® team employs $\approx$ five quantitative analysts
- Heat maps popular on TV broadcasts
  - ESPN® and MLB® signed \$700 million/year contract
- Joey Votto (\$22.5 mil/year) packs dog-eared copy of "The Science of Hitting"
- "Big Data is Changing Baseball" - 1 TB/game [Delgado, 2014]

The Data

- PITCHf/x® - Sportsvision, high speed stereoscopic cameras, MLB Advanced Media, Gameday website, *open source*
  - http://www.sportvision.com/media/espn-k-zone
- MySQL - relational database management system, 2 GB, 'at bat' table 1,711,211 $\times$ 15, table joins
- Variables
  - **px** - horizontal location
  - **pz** - vertical location
  - **des** - pitch outcome
  - **ab_id** - ab bat ID number
  - **pitch_id** - pitch ID number
  - **pitch_type** - fastball, curve ball, etc.
  - **stand** - batter handedness
  - **batter** - batter ID number

Outline

Empirical Success Probability Heat Map



Empirical Success Probability Heat Map

- Swings: $i = 1, 2, \ldots, N$
- Bernoulli(π_i) trials:

$$S_i = \begin{cases} 1; & \text{swing success} \\ 0; & \text{swing failure} \end{cases}$$

- Location (x_i, y_i)
- Grid boxes $G_1, G_2, \ldots, G_B$
- Box b totals:

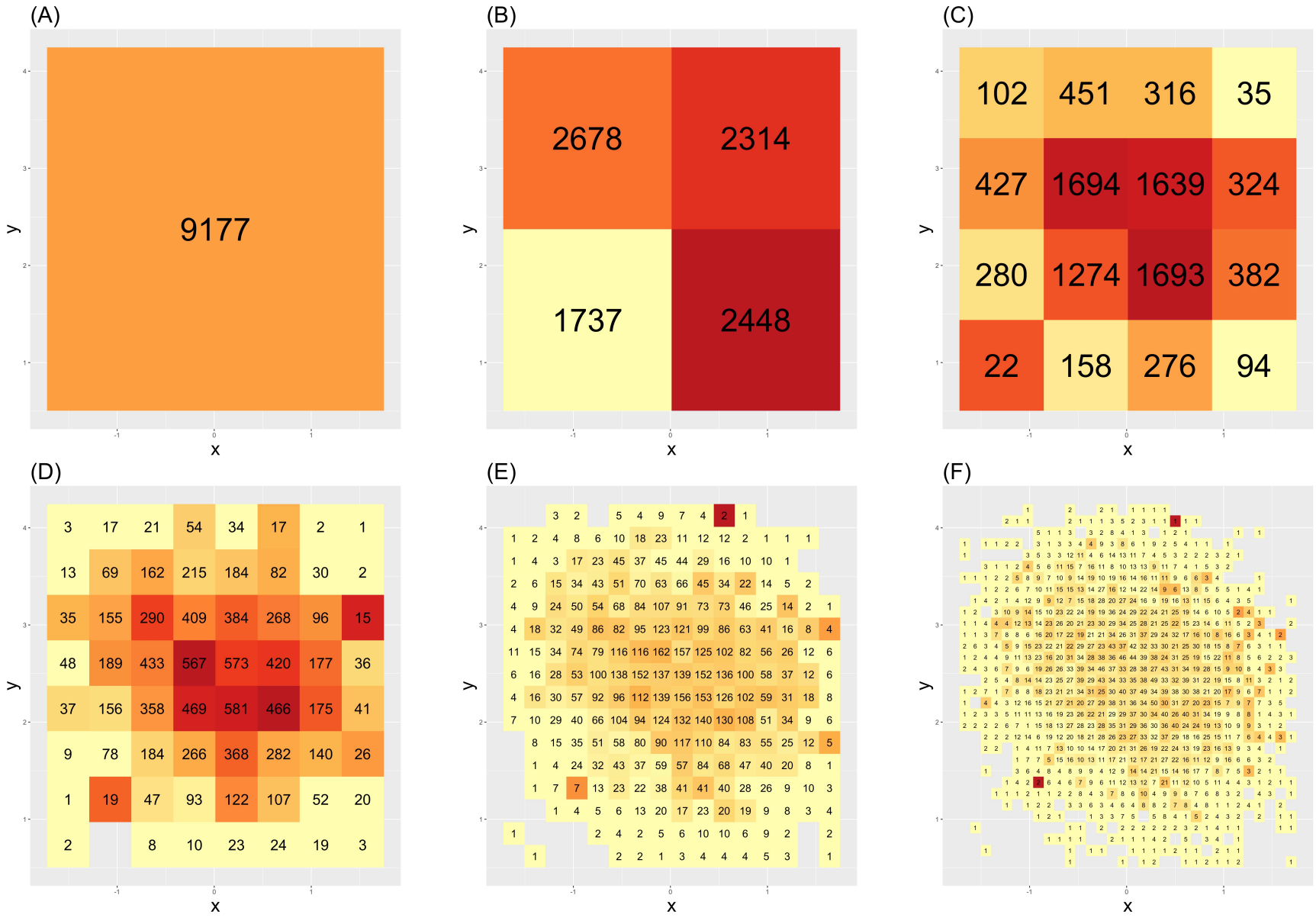
$$N_b = \sum_{i=1}^{N} I_{(x_i, y_i) \in G_b}$$

- Box b empirical success:

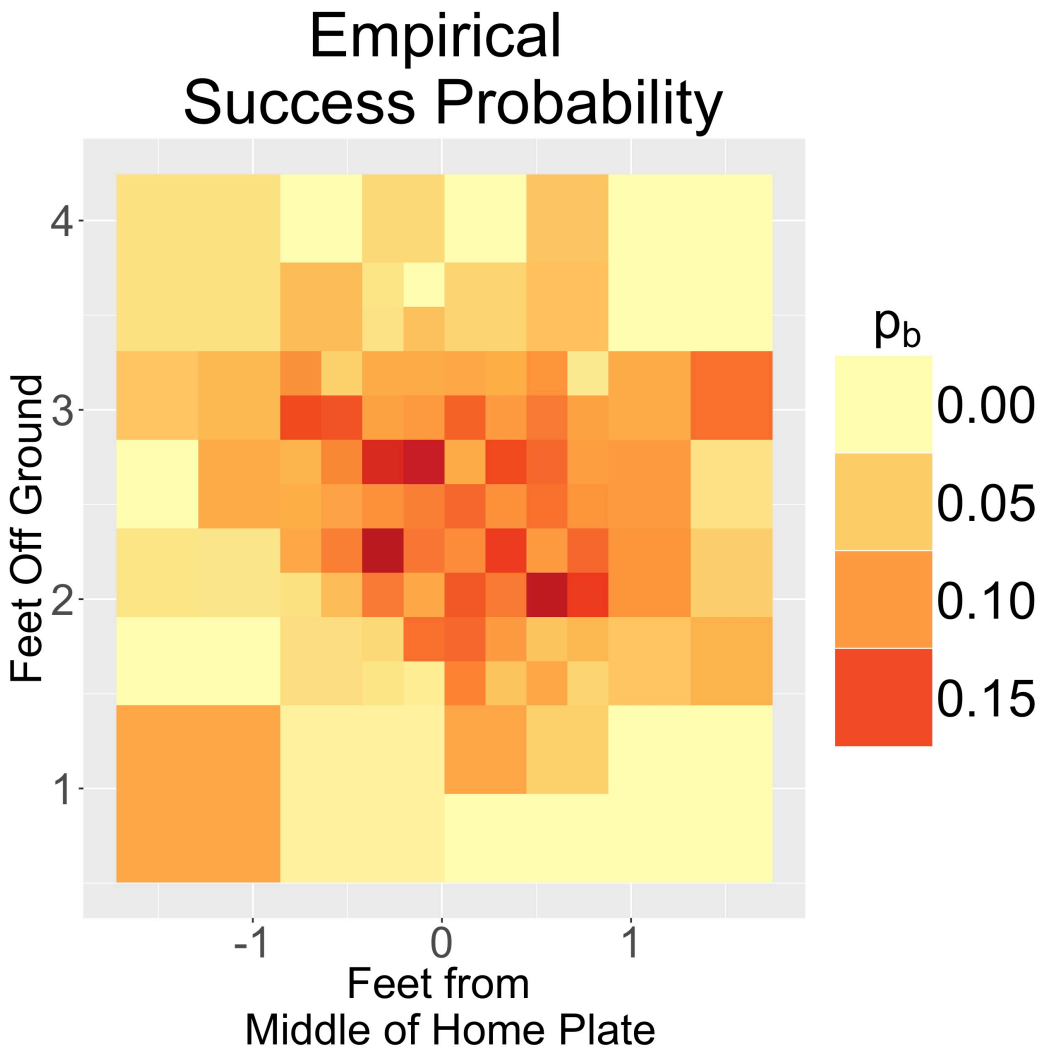
$$p_b = \frac{1}{N_b} \sum_{i=1}^{N} S_i I_{(x_i, y_i) \in G_b}$$

Heat Map Resolution Selection

Jhonny Peralta



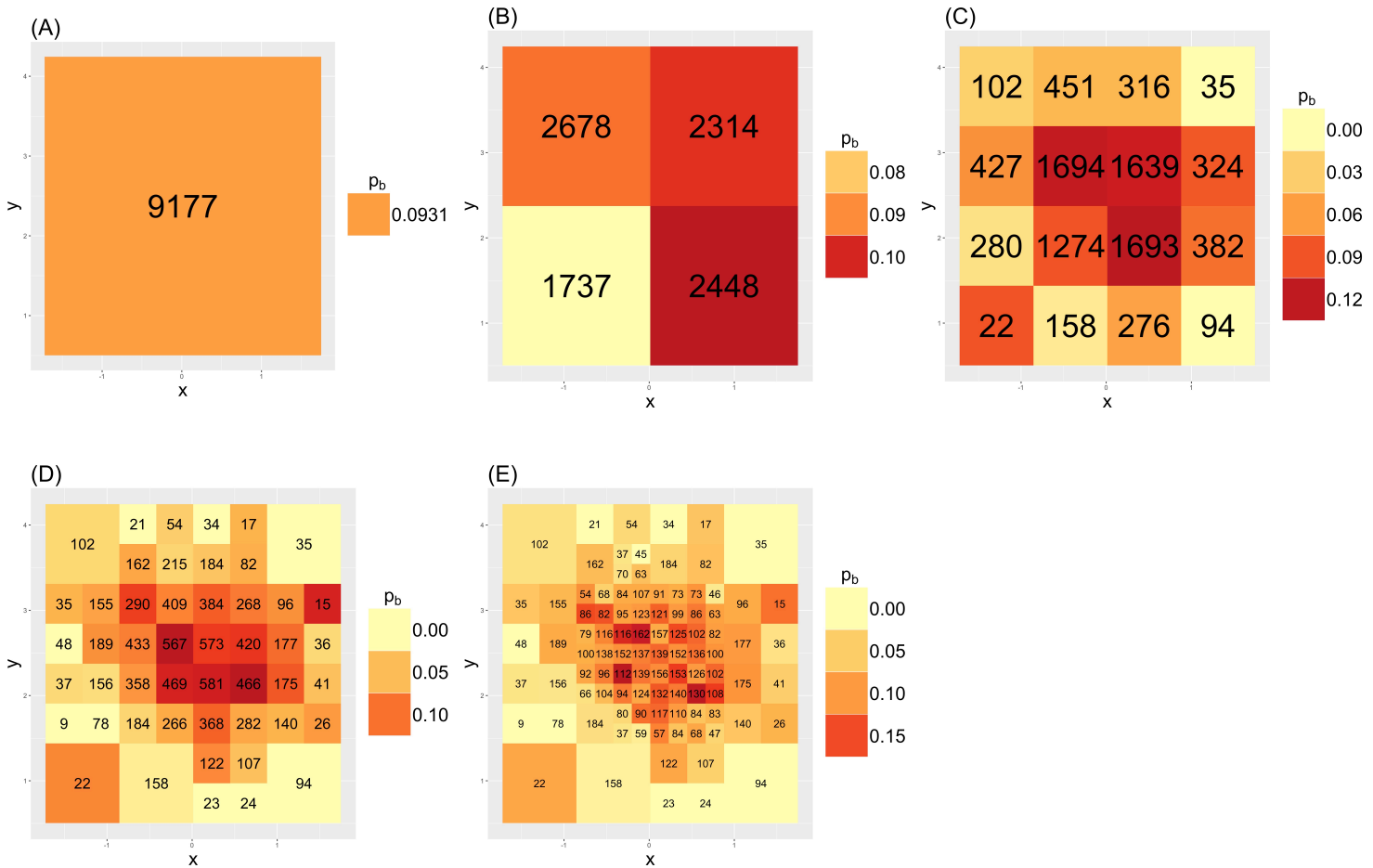
A Thing of Beauty



Variable-Resolution Heat Maps

Jhonny Peralta

- VR algorithm
  - Stopping rule
  - Subdivision method



Variable-Resolution Heat Maps

Combine Resolutions

- Sample size stopping rule
- Combine resolutions
- Resolution conveys varying data abundance
- Improvements:
  - **.gif** option
  - Alternate stopping rule types
  - Alternate subdivision method

varyres(...) in **varyres**

varyres {varyres} R Documentation

A variable-resolution heat map generator

Description

This function creates variable resolution heat maps accroding to a stopping rule

Usage

```
varyres(dataset, cutoff, fun = mean, max = 6)
```

Arguments

`dataset`	data frame with spatial data: x-coordinates (x), y coordinates (y), and Bernoulli responses at those locations (res)
`cutoff`	Box subdivisions cease when a box sample size drops below the cutoff
`fun`	Function to apply to responses in each box
`max`	The maximum number of subdivision iterations the algorithm will perform

Value

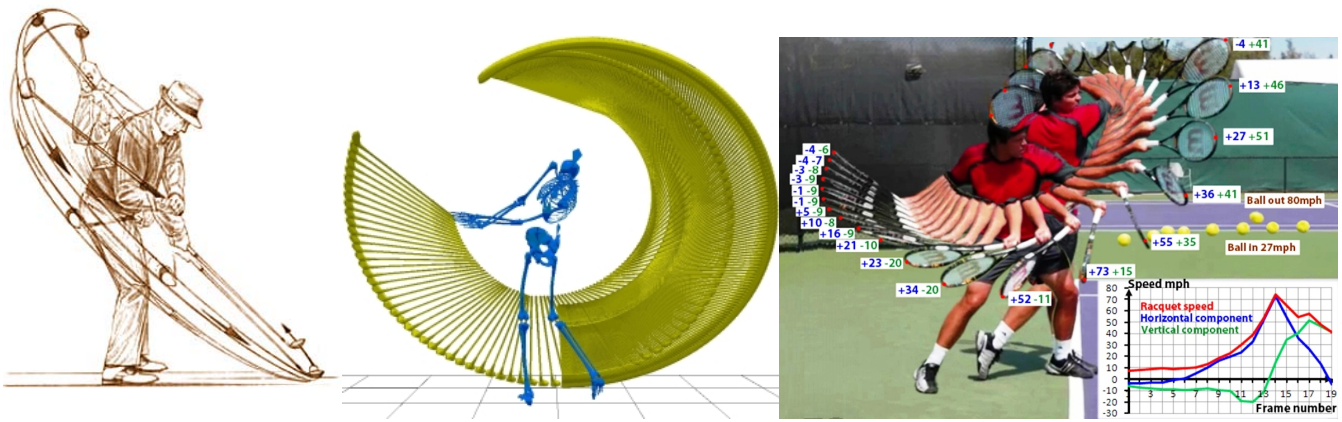
A list containing a data frame for each iteration of the subdivision algorithm; and a vector of the number of boxes eligible for subdivision at each iteration.

Examples

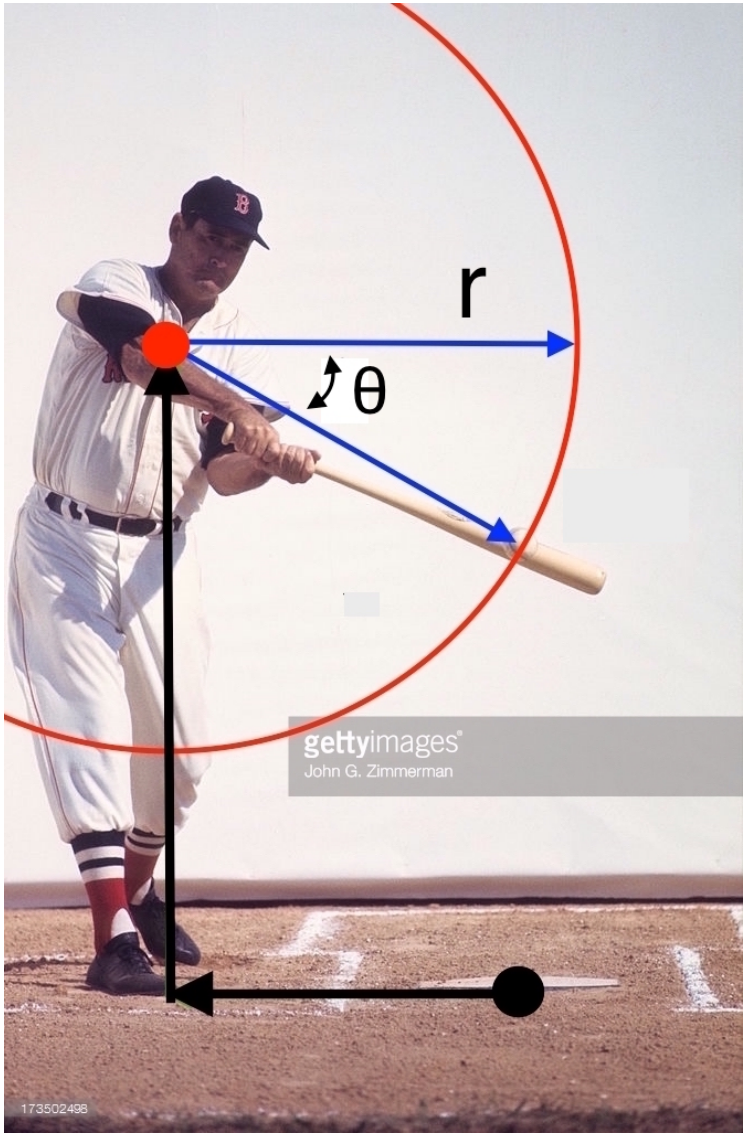
```
data(hitter)
data <- varyres(hitter, mean, cutoff = 200, max = 6)
mapit(data[[4]])
```

Outline

- Rotational movement
- Change impact point $\Longrightarrow$ biomechanical adjustments
- Adjustments affect impact conditions, outcome probabilities



Baseball Biomechanics



- Polar coordinates
- Translate origin
- Radius and angle biomechanically meaningful
- Big challenge: new origin location
  - Fleisig, ASMI
  - Dowling, Motus
- Let's model Jhonny Peralta, Mr. 9172

Generalized Linear Model

$$Y_i|\mathbf{X}_i(\mathbf{s}_i) \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \mathbf{X}_i(\mathbf{s}_i)\beta$$

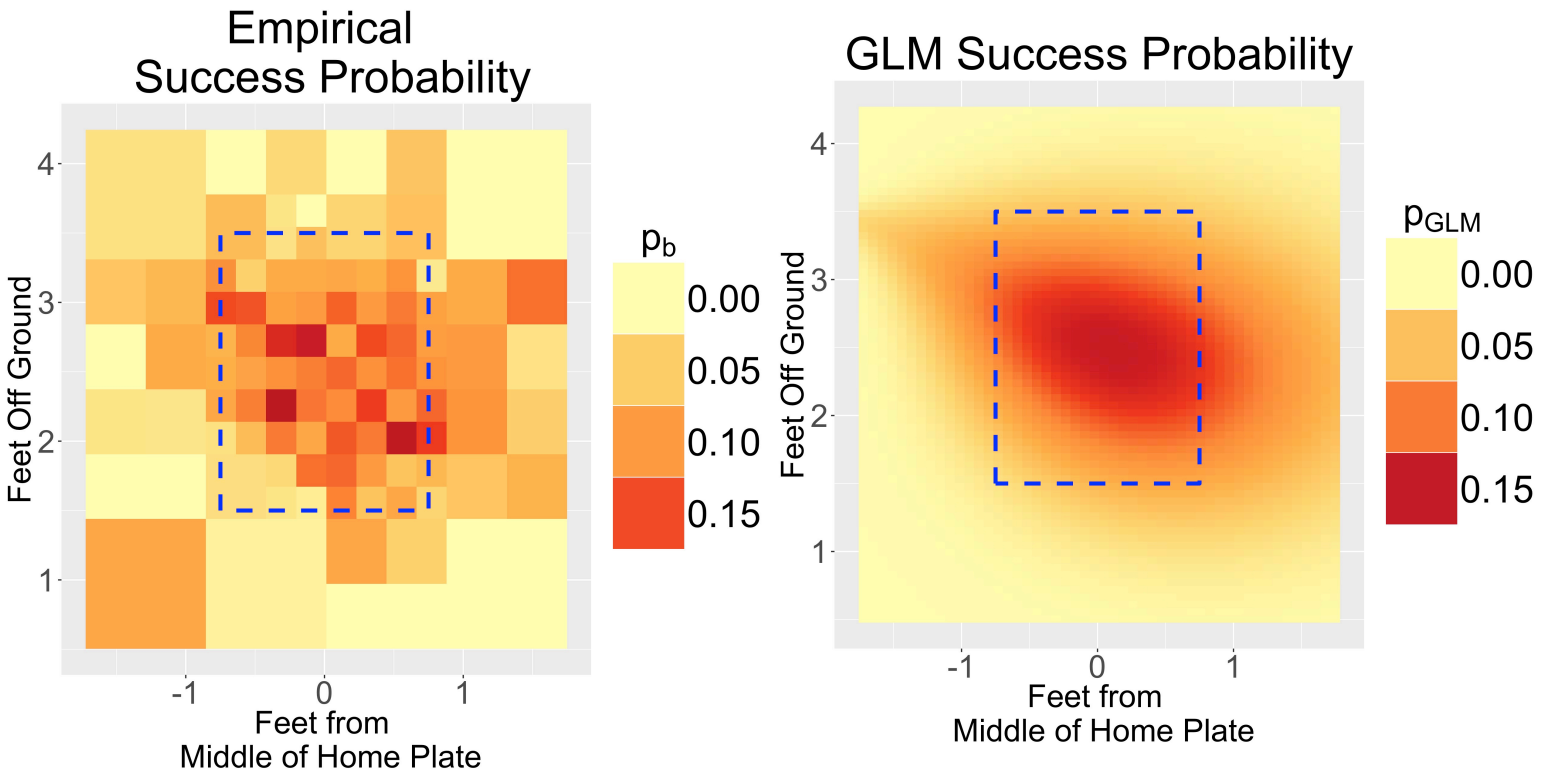
- Swings: $i = 1, 2, \ldots, N$
- Pitch location: $\mathbf{s}_i = (x_i, y_i)$
- Biomechanical covariates: $\mathbf{X}_i(\mathbf{s}_i)$
- Fit to Jhonny Peralta data

Logistic Regression Model

Jhonny Peralta

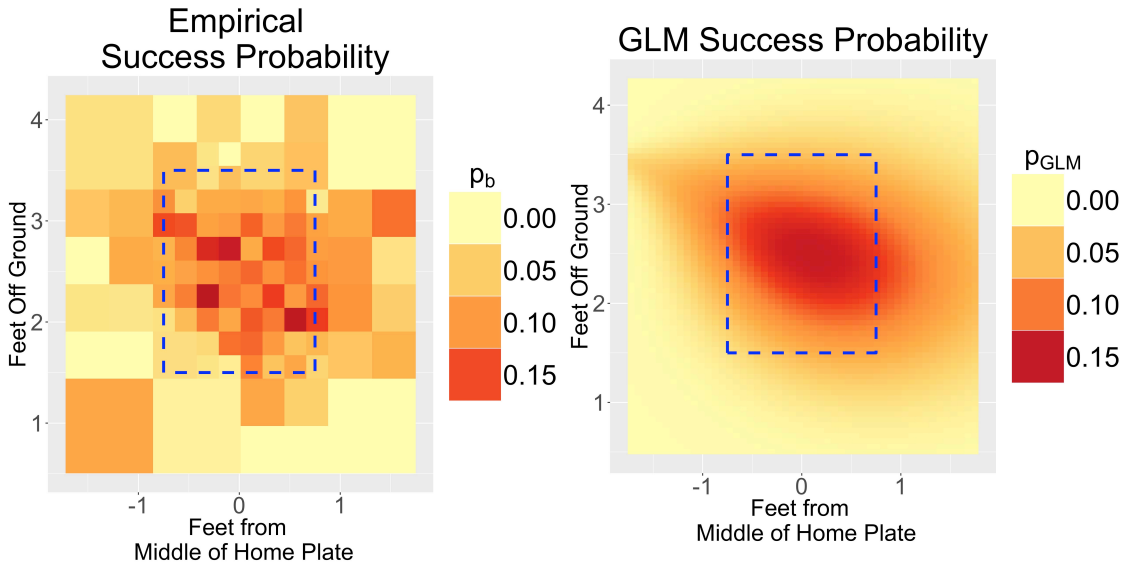
$$Y_i|\mathbf{X}_i(\mathbf{s}_i) \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \mathbf{X}_i(\mathbf{s}_i)\beta,$$



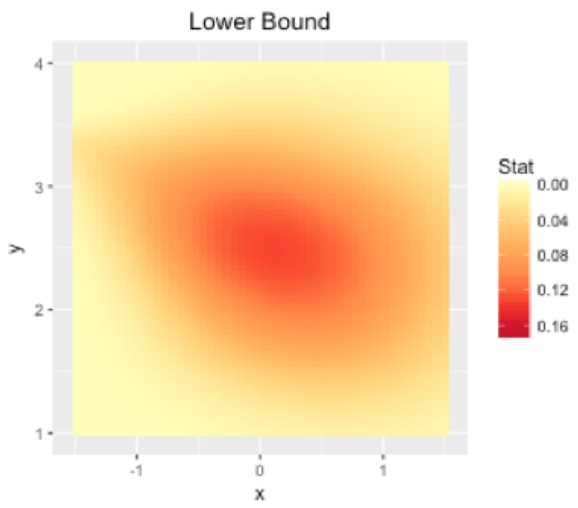
Logistic Regression Model

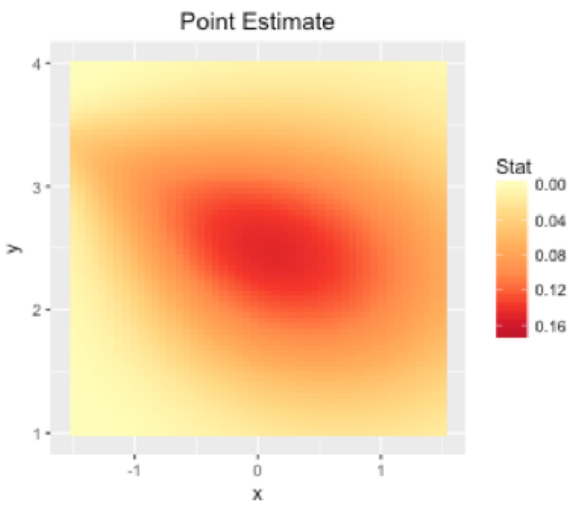
Jhonny Peralta

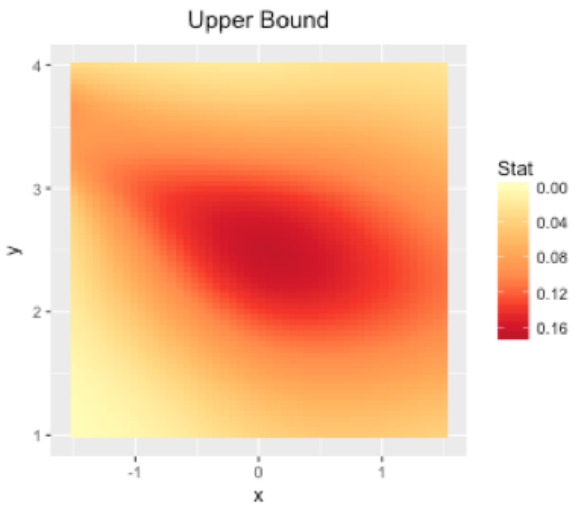


- Hosmer-Lemeshow (logistic regression) GOF test
  - H_0: Well fit
    H_A: Lack of fit
  - p-value = 0.8217
- Confidence intervals

Interactive Confidence Intervals

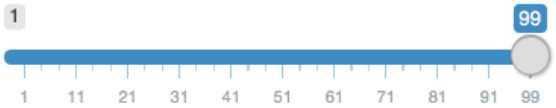






This is the 99 % confidence interval layer.

Confidence Interval



mapapp: An R Package

mapapp: An R Package

- Challenge: Where does user stop and package begin?
- Future improvement

```
all_in_one <- get_CI(model, x, y, levels)
shinyHMCI(all_in_one)
```

1. `get_CI(...)` — creates proper data structure
2. `shinyHMCI(all_in_one)` — creates Shiny app

...and **.gif** option.

Outline

Spatial Generalized Linear Mixed Model (SGLMM)

$$Y_i|\mathbf{X}_i(\mathbf{s}_i) \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \mathbf{X}_i(\mathbf{s}_i)\beta + w(\boldsymbol{s}_i)$$

- $w(\boldsymbol{s}_i)$ — spatial random effect, location $\boldsymbol{s}_i$.
- $\boldsymbol{w}(\boldsymbol{s}) = (w(\boldsymbol{s}_1), w(\boldsymbol{s}_2), \ldots, w(\boldsymbol{s}_N))$ — vector
- $\boldsymbol{w}(\boldsymbol{s})$ - Gaussian Random Field (GRF)

Gaussian Random Field: $\mathbf{w}(\mathbf{s})$

$$\boldsymbol{w}(\boldsymbol{s})|\boldsymbol{\theta} \sim \text{MVN}(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$$

$$\Sigma(\phi, \sigma^2)_{i,k} = \sigma^2 \exp(-||\boldsymbol{s}_i - \boldsymbol{s}_k||/\phi)$$

- Spatial exponential covariance
  - $||\boldsymbol{s}_i - \boldsymbol{s}_k||$ - Euclidean distance
  - σ^2 - scale parameter
  - ϕ - range parameter.
- Notice: $\Sigma(\boldsymbol{\theta})$ — $n \times n$

Computational Cost: “Big N” Problem

- Peralta: $n = 9177$
- $\boldsymbol{w}(\boldsymbol{s})$: 9177×9177 cov. matrix
- MCMC iterations require: Σ^{-1}, determinant of Σ
- $\mathcal{O}(n^3)$ rate of increase:

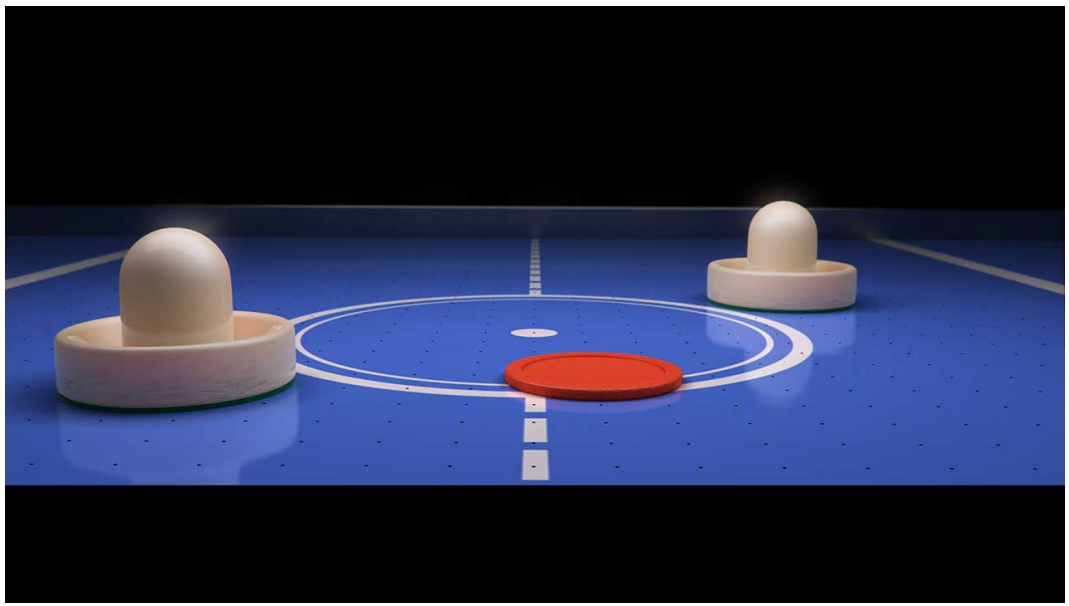
$$t(n) \leq M \cdot n^3$$

$$\text{as } n \to \infty$$

Outline

Hamiltonian Monte Carlo (HMC)

- Stan uses Hamiltonian proposal mechanism
  1. Disk on surface, with position (β, θ) and momentum
  2. Randomly sample momentum (auxiliary)
  3. Calculate new position (parameters)
  4. That's your Metropolis proposal.



Stan Computational Optimization

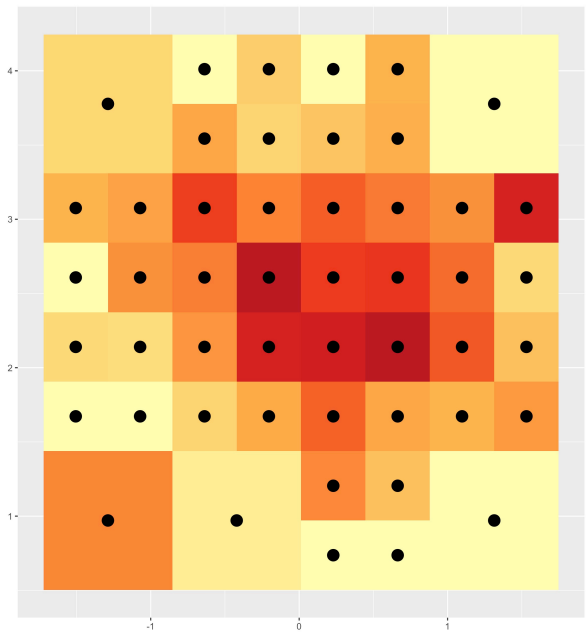
- ϕ, β: informative/proper priors $\rightarrow$ identifiability/cost/convergence
- QR factorization of X, Cholesky decomposition of $\Sigma(\theta)$
- Matrices & vectors faster than loops & scalars
- n = 25: 40 seconds $\rightarrow$ 3 seconds
- n = 2000 — overnight, 350 draws

Outline

Predictive Process Models (PPMs)

[Banerjee et al., 2008]

- Knots: $\boldsymbol{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_m^*\}$
  - $m << n$



$$\text{logit}(\pi_i|\boldsymbol{s}_i) = \boldsymbol{X}_i(\boldsymbol{s}_i)\beta + \widetilde{w}(\boldsymbol{s}_i)$$

$$\widetilde{\boldsymbol{w}}(\boldsymbol{s}) \sim \text{MVN}\{0, \widetilde{\Sigma}(\boldsymbol{\theta})\}$$

$$\widetilde{\Sigma}(\boldsymbol{\theta})_{i,j} = \boldsymbol{\sigma}^{*T}(\boldsymbol{s}_i;\boldsymbol{\theta}) \cdot \boldsymbol{\Sigma}^{*-1}(\boldsymbol{\theta}) \cdot \boldsymbol{\sigma}^*(\boldsymbol{s}_j;\boldsymbol{\theta})$$

- $\boldsymbol{\sigma}^*(\boldsymbol{s}_i;\boldsymbol{\theta}) = \text{Cov}(\boldsymbol{s}_i, \boldsymbol{S}^*)$
- $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}) = \text{Var}(\boldsymbol{S}^*)$

PPM Results

- Implement in **spBayes** [Finley et al., 2013].
- MCMC chains did not converge — trace-plots:

  - n = 1000, knots = 97, 10K samples, $\approx$ 6.7 mins
  - n = 1000, knots = 49, 30K samples, $\approx$ 7 mins
  - n = 3000, knots = 49, 80K samples, $\approx$ 54 mins
- Speed issue for extending MCMC chains

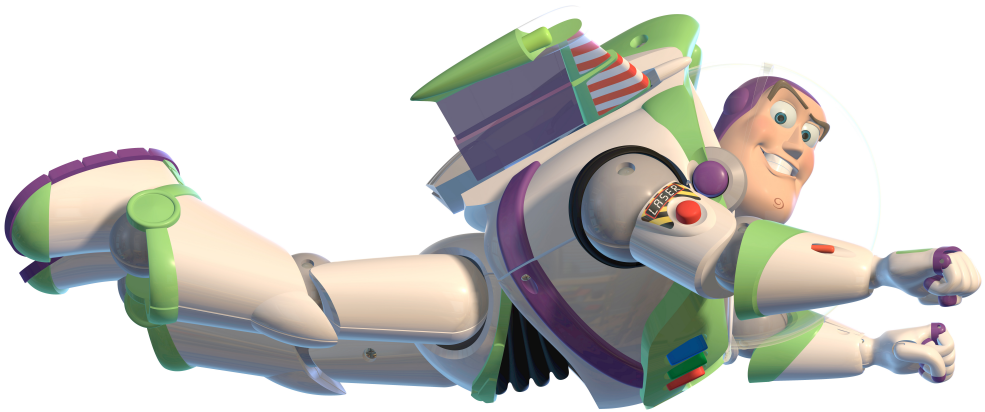
Outline

Integrated Nested Laplace Approximation (INLA)

[Rue et al., 2009]

$$\text{logit}(\pi_i) = \boldsymbol{X}_i(\boldsymbol{s}_i)\beta + \boldsymbol{w}(\boldsymbol{s}_i)$$
$$\boldsymbol{w}(\boldsymbol{s})|\boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{0}, \Sigma(\boldsymbol{\theta}))$$

- Assume Matérn covariance
- Two parts (continuous domain)
  - Part 1: Stochastic Partial Differential Equation (SPDE)
  - Part 2: Integrated Nested Laplace Approximation (INLA)



INLA: Step 1

- Parameter vector: $\boldsymbol{\rho} = (\beta^{\mathsf{T}}, \widetilde{\boldsymbol{w}}^{\mathsf{T}})^{\mathsf{T}}$
  - $\boldsymbol{Q}$: precision matrix of $\boldsymbol{\rho}$
- Hyperparameter vector: $\boldsymbol{\theta} = (\kappa, \sigma)$

1. Gaussian approximation:

$$p(\boldsymbol{\rho}|\boldsymbol{\theta}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta}) p(\boldsymbol{\rho}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

$$p(\boldsymbol{\rho}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\rho}^T \boldsymbol{Q} \boldsymbol{\rho} + \sum_i \log p(y_i|\boldsymbol{\rho}, \boldsymbol{\theta})\right)$$

$$p_G(\boldsymbol{\rho}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\rho} - \boldsymbol{\mu})^T (\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))(\boldsymbol{\rho} - \boldsymbol{\mu})\right)$$

   - Gaussian kernel: $\boldsymbol{c}$ and $\boldsymbol{\mu}$ depend on second order Taylor expansions of $f(\boldsymbol{\rho}) = \sum_i \log p(y_i|\boldsymbol{\rho}, \boldsymbol{\theta})$

INLA: Step 2

- Fact: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta})p(\boldsymbol{\rho}|\boldsymbol{\theta})}{p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})}$$

- Bayes proportionality:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta})p(\boldsymbol{\rho}|\boldsymbol{\theta})}{p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})} \cdot p(\boldsymbol{\theta}) \end{aligned}$$

2. For a given $\boldsymbol{\theta}$, let $\boldsymbol{\rho}_0 = \text{argmax}_{\rho} p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})$. Then,

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}|\boldsymbol{\rho}_0, \boldsymbol{\theta})p(\boldsymbol{\rho}_0|\boldsymbol{\theta})}{p_G(\boldsymbol{\rho}_0|\boldsymbol{y}, \boldsymbol{\theta})} \cdot p(\boldsymbol{\theta})$$

INLA: Step 2

- Fact: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta})p(\boldsymbol{\rho}|\boldsymbol{\theta})}{p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})}$$

- Bayes proportionality:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{\rho}, \boldsymbol{\theta})p(\boldsymbol{\rho}|\boldsymbol{\theta})}{p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})} \cdot p(\boldsymbol{\theta}) \end{aligned}$$

2. For a given $\boldsymbol{\theta}$, let $\boldsymbol{\rho}_0 = \text{argmax}_{\rho} p(\boldsymbol{\rho}|\boldsymbol{y}, \boldsymbol{\theta})$. Then,

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}|\boldsymbol{\rho}_0, \boldsymbol{\theta})p(\boldsymbol{\rho}_0|\boldsymbol{\theta})}{p_G(\boldsymbol{\rho}_0|\boldsymbol{y}, \boldsymbol{\theta})} \cdot p(\boldsymbol{\theta})$$

INLA: Step 3 & Step 4

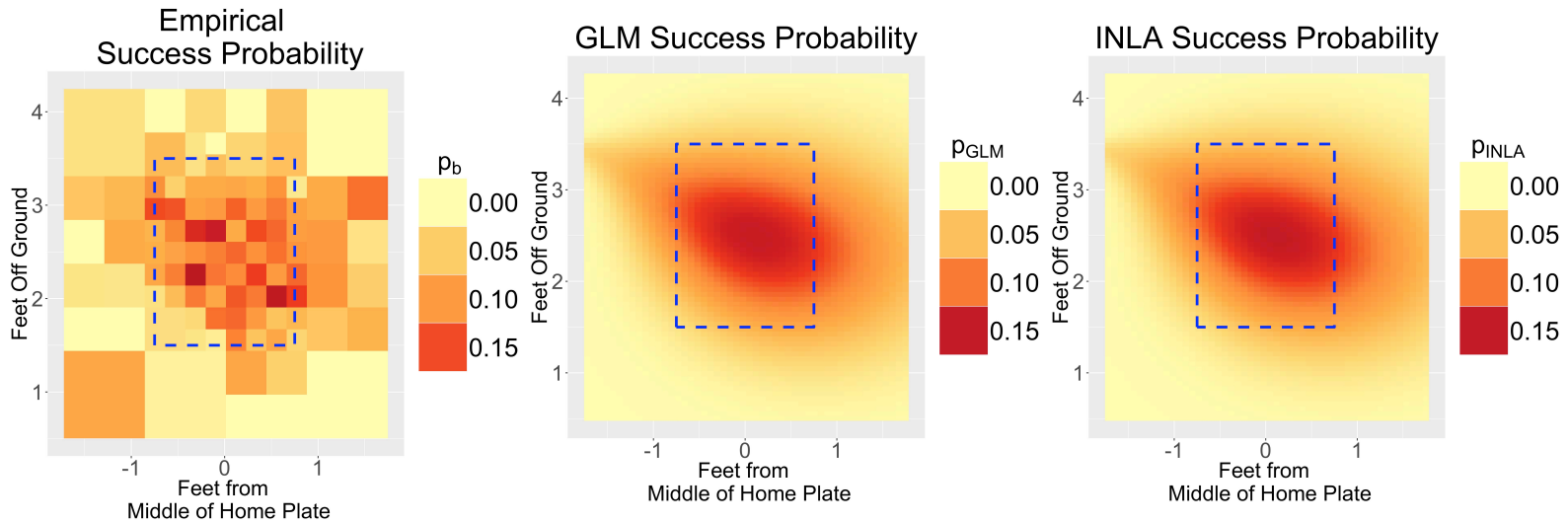
3. Numerical integration:

$$p(\rho_j|\boldsymbol{y}) \approx \int p_{\mathsf{G}}(\rho_j|\boldsymbol{\theta}, \boldsymbol{y})\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

4. Numerical integration:

$$p(\theta_k|\boldsymbol{y}) \approx \int \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-k}$$

INLA Model Fit: 34 seconds



INLA		
ρ_i	$\hat{\rho}_i$	SE
κ	3.23	$\pm$ 1 SE: (1.35, 7.54)
σ	0.11	$\pm$ 1 SE: (0.05, 0.26)

- $\hat{\kappa}$ — long range correlation
- $\text{SE}(w(\boldsymbol{s})) = 0.11$.
- $p_i = 0.15 \rightarrow p_i \pm 1 \cdot \text{SE} = (0.137, 0.165)$
- $p_i = 0.15 \rightarrow p_i \pm 2 \cdot \text{SE} = (0.125, 0.181)$

Summary

- Resolution selection? Variable-resolution heat maps and **varyres**
- Heat map confidence intervals? Interactive HMCIs and **mapapp**
- Fitting big data SGLMMs to baseball data?
  - Stan - inadequate
  - PPM - Slow and did not converge
  - INLA - Fast, successful(?); to be continued...

Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

Andrew O Finley, Sudipto Banerjee, and Alan E Gelfand. spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *arXiv preprint arXiv:1310.8192*, 2013.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498, 2011.