Take Me Out to (Analyze) the Ballgame

Visualization and Analysis Techniques for Big Spatial Data

Chris Comiskey

Oregon State University

April 24, 2017

Baseball, Baseball, Baseball

- Chris, rookie year

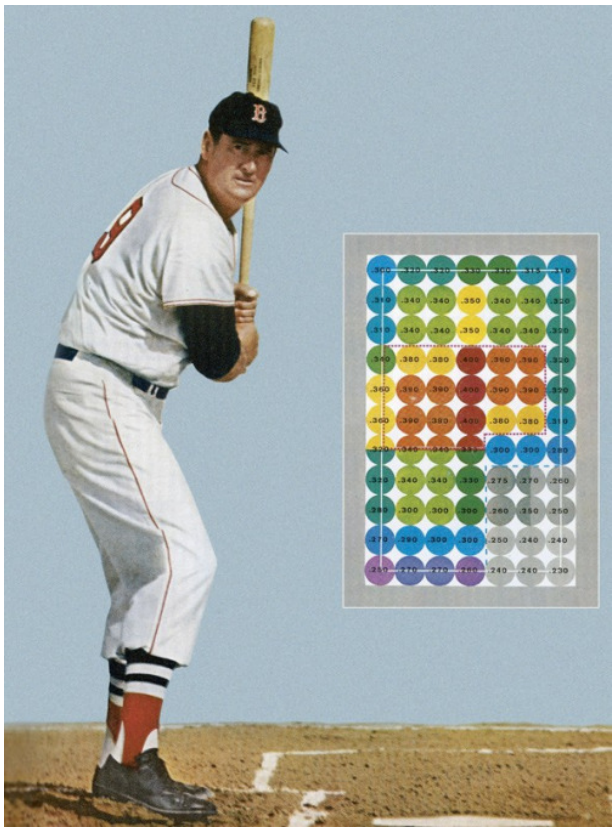
Baseball, Baseball, Baseball

- Chris, rookie year

- Chris, Boston Red Sox

Hitting Analytics, Then and Now

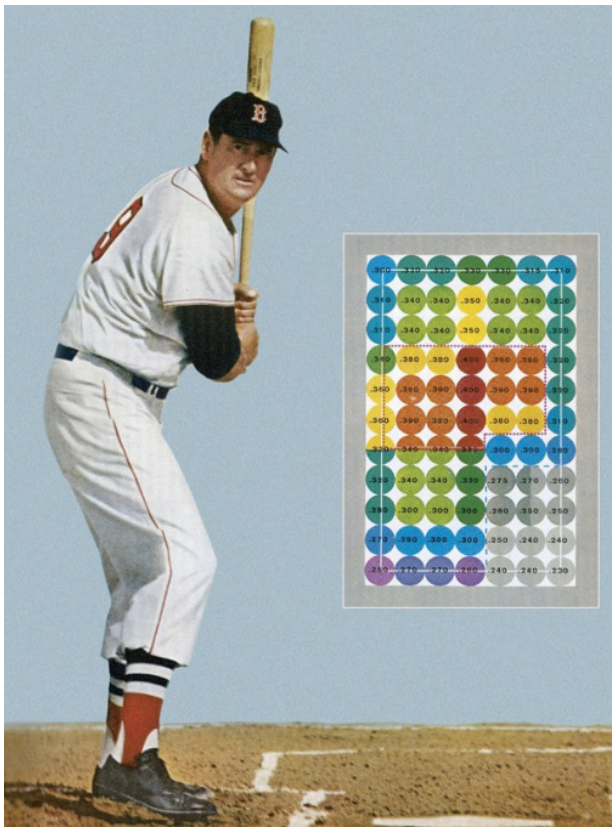
- "The Science of Hitting"$_{1970}$



- Conceptual breakthrough
- No data

Hitting Analytics, Then and Now

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



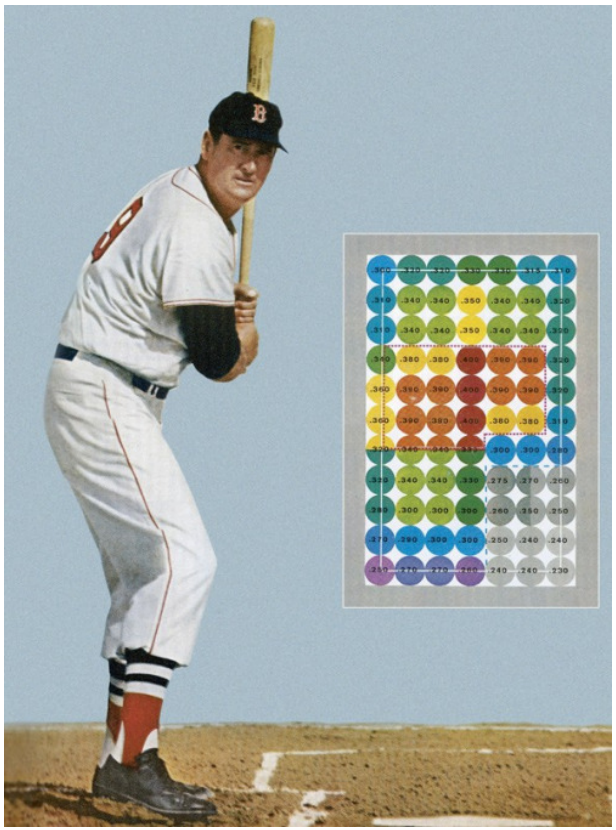
- Iconic breakthrough, hitter
- No data

- Hall of Fame exhibit

- Baseball fanatic statistician
- PITCHf/x data!

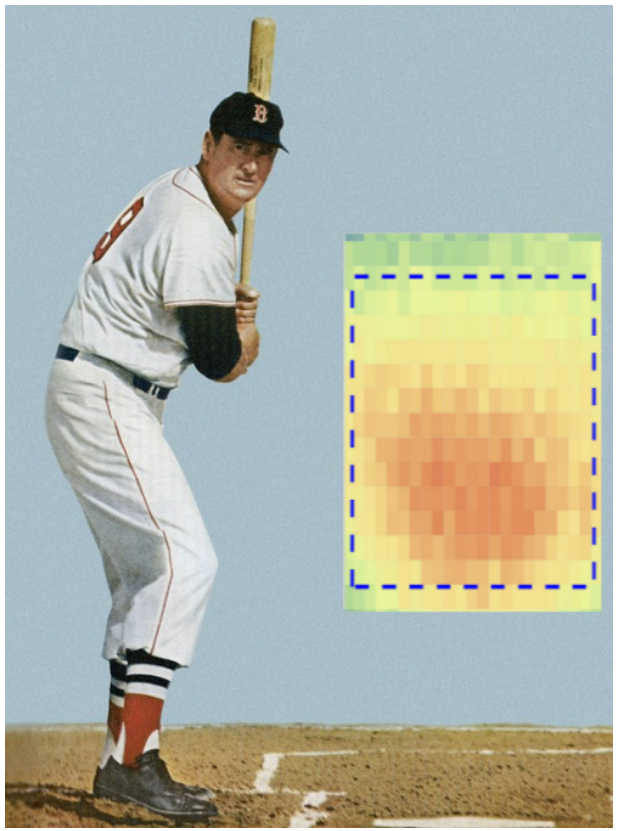
Hitting Analytics, Then and Now

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



- Iconic breakthrough, hitter
- No data

- The Statistics of Hitting



- PITCHf/x data!
- Relevant?

Relevant Research

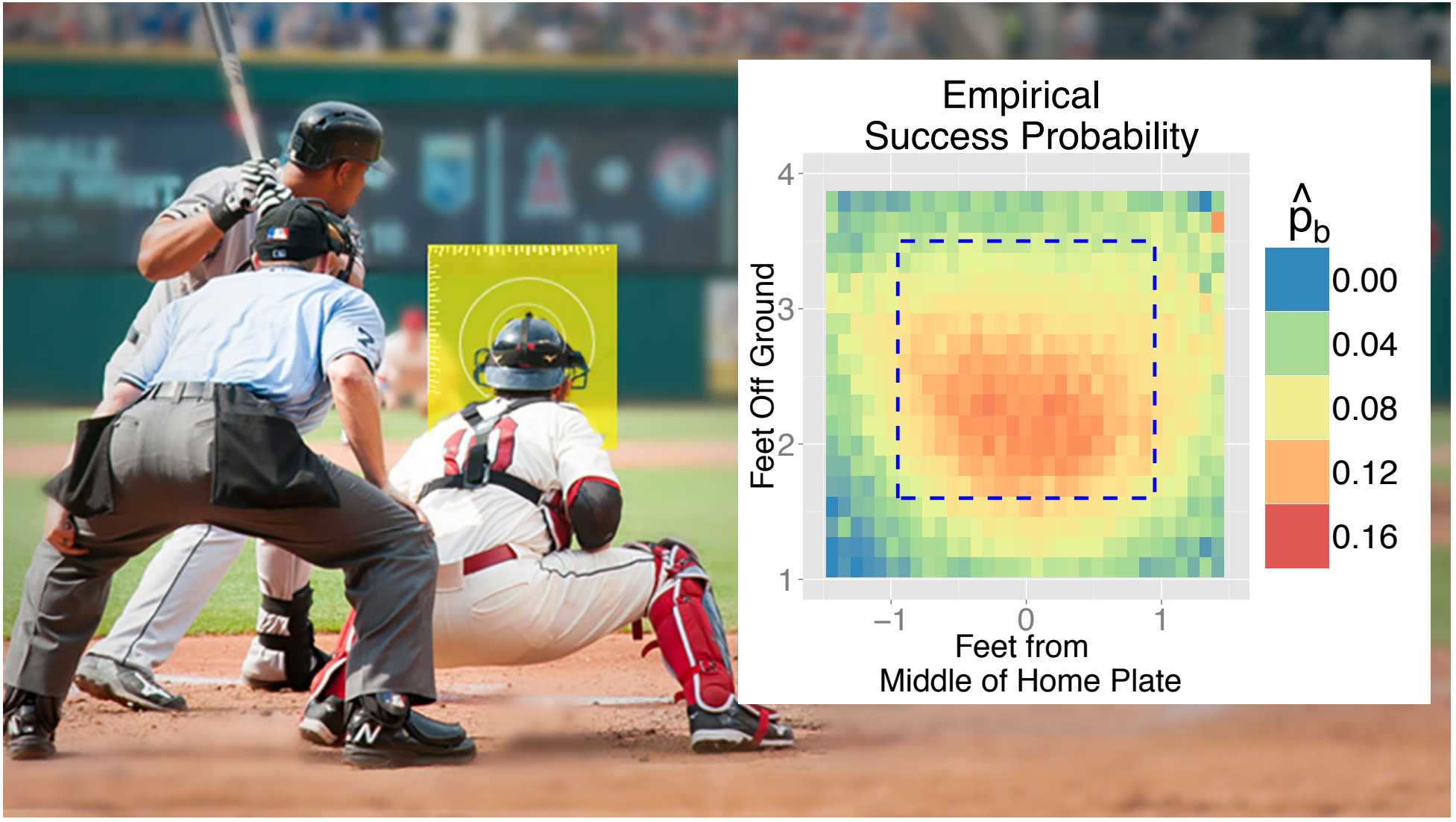
- Each MLB® team spends $\approx$ \$15 million/year
- Each MLB® team employs $\approx$ five quantitative analysts
- Heat maps popular on TV broadcasts
  - ESPN® and MLB® signed \$700 million/year contract
- Joey Votto (\$22.5 mil/year) packs dog-eared copy of "The Science of Hitting"
- "Big Data is Changing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

Empirical Success Probability Heat Map

Empirical Success Probability Heat Map

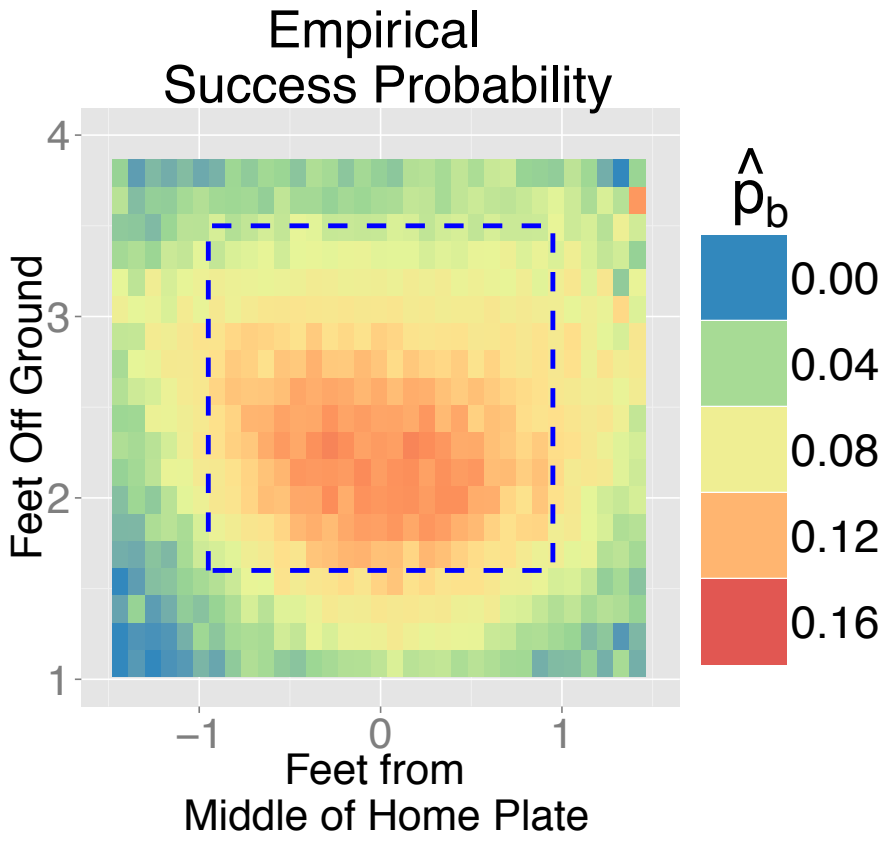


Empirical Success Probability Heat Map

- Grid vertical strike zone face
- Swing i, grid box b
- **Every swing - Bernoulli trial**
  - $Y_i = 1$, for swing success
- n_b = number of box b swings
- box b success probability:

$$\hat{p}_b = \frac{1}{n_b} \sum_{i \in b} Y_i$$

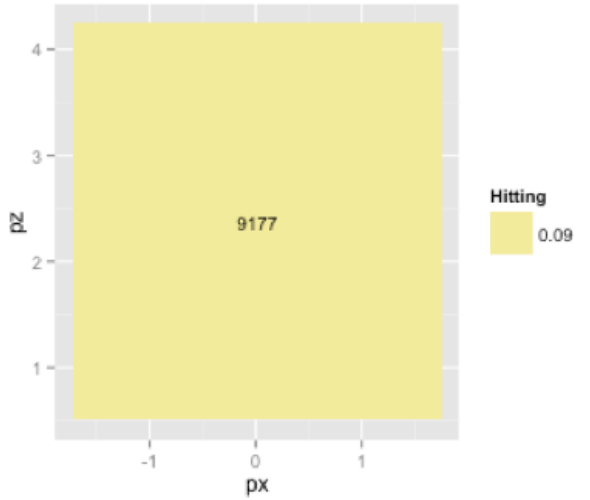
- Map $\hat{p}_b$ to box color

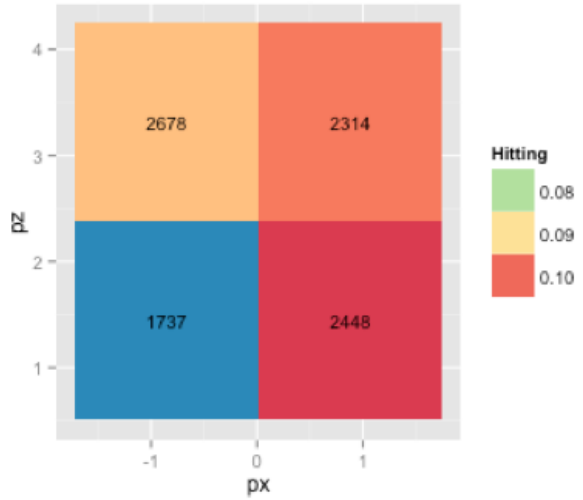


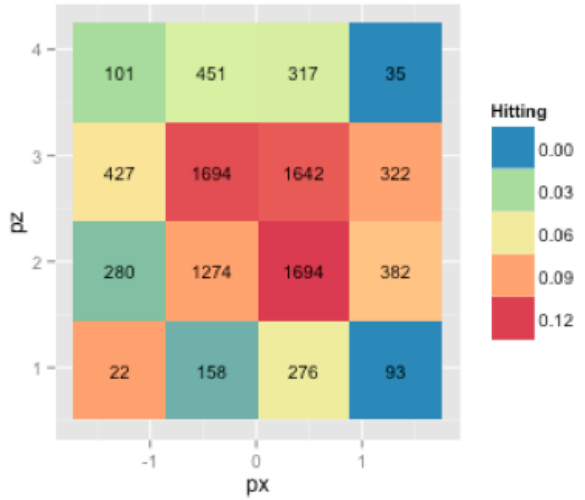
Note: $\hat{p}_b$ = **success probability**, not batting average

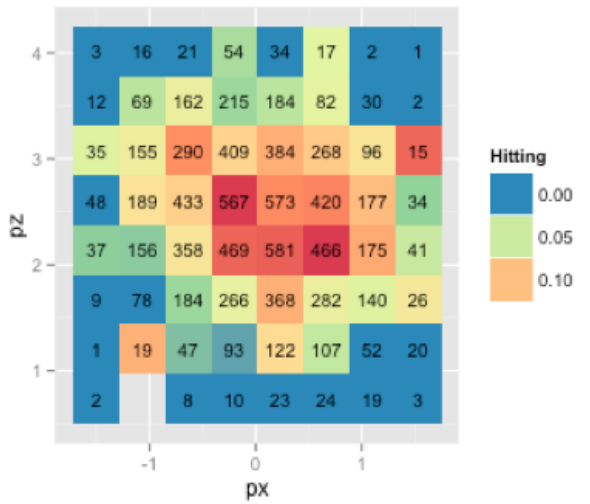
Heat Map Resolution, Jhonny Peralta

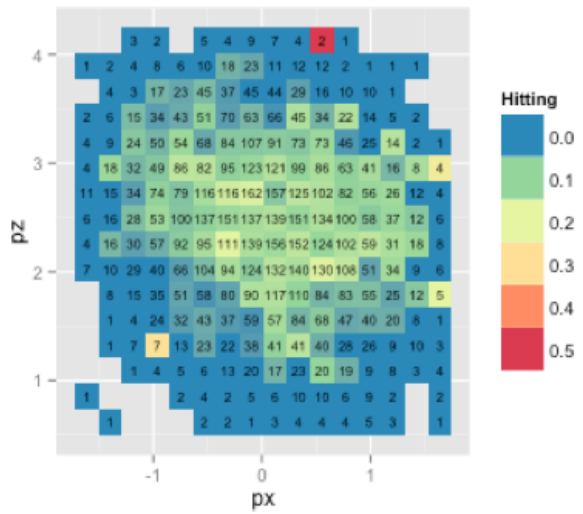
Decisions, Decisions...





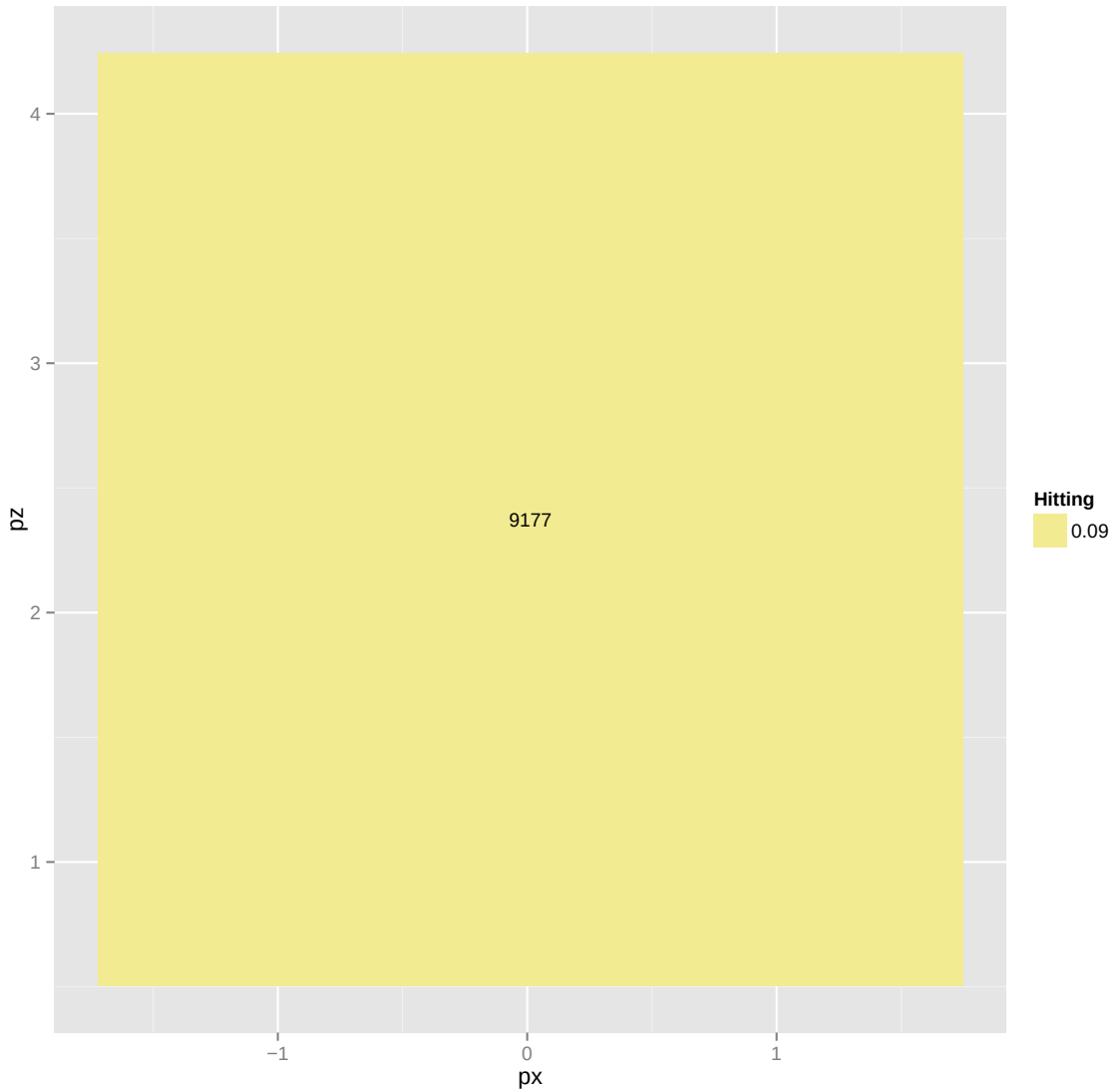






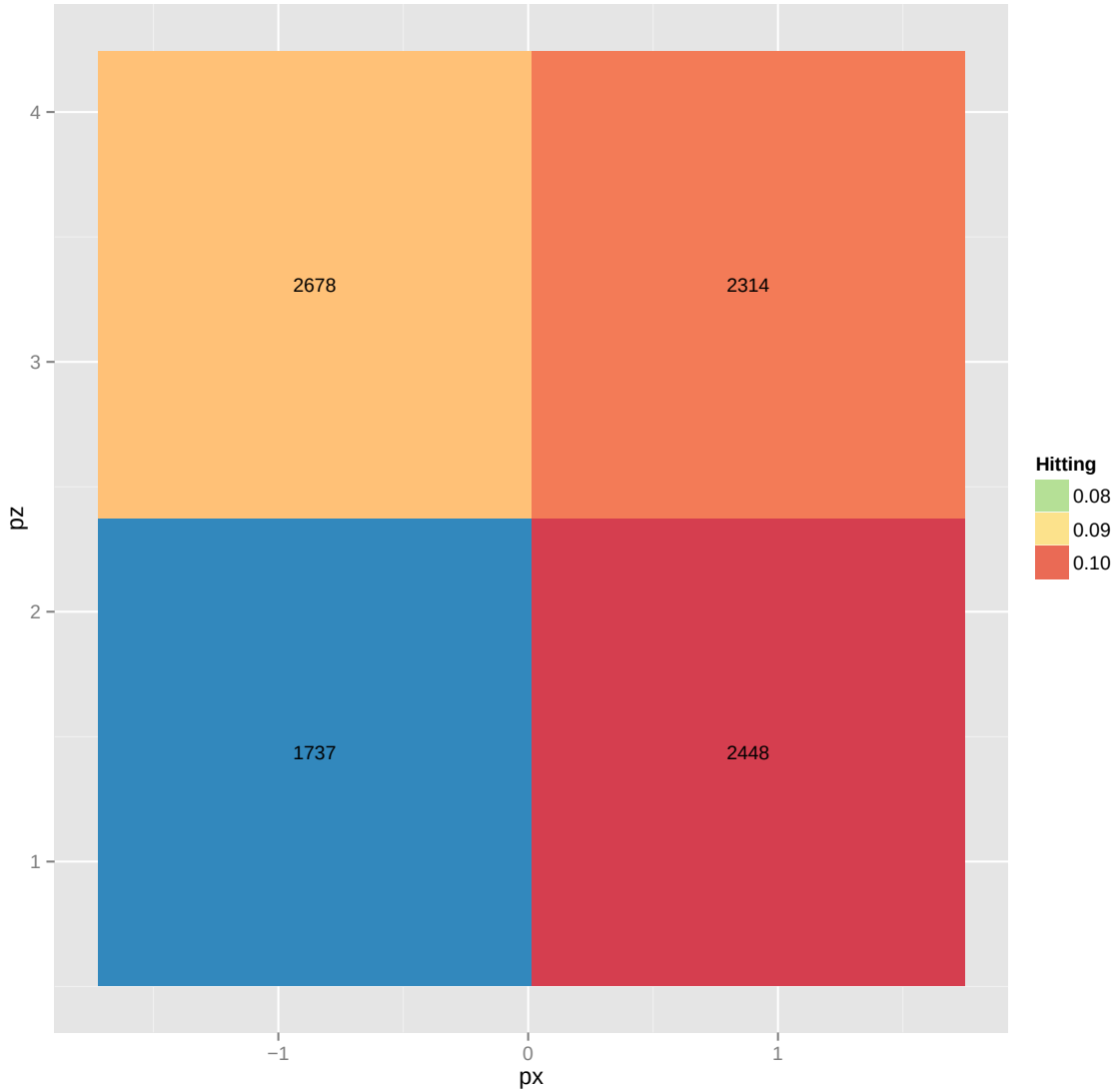
Heat Map Resolution, Jhonny Peralta

Starting from Scratch



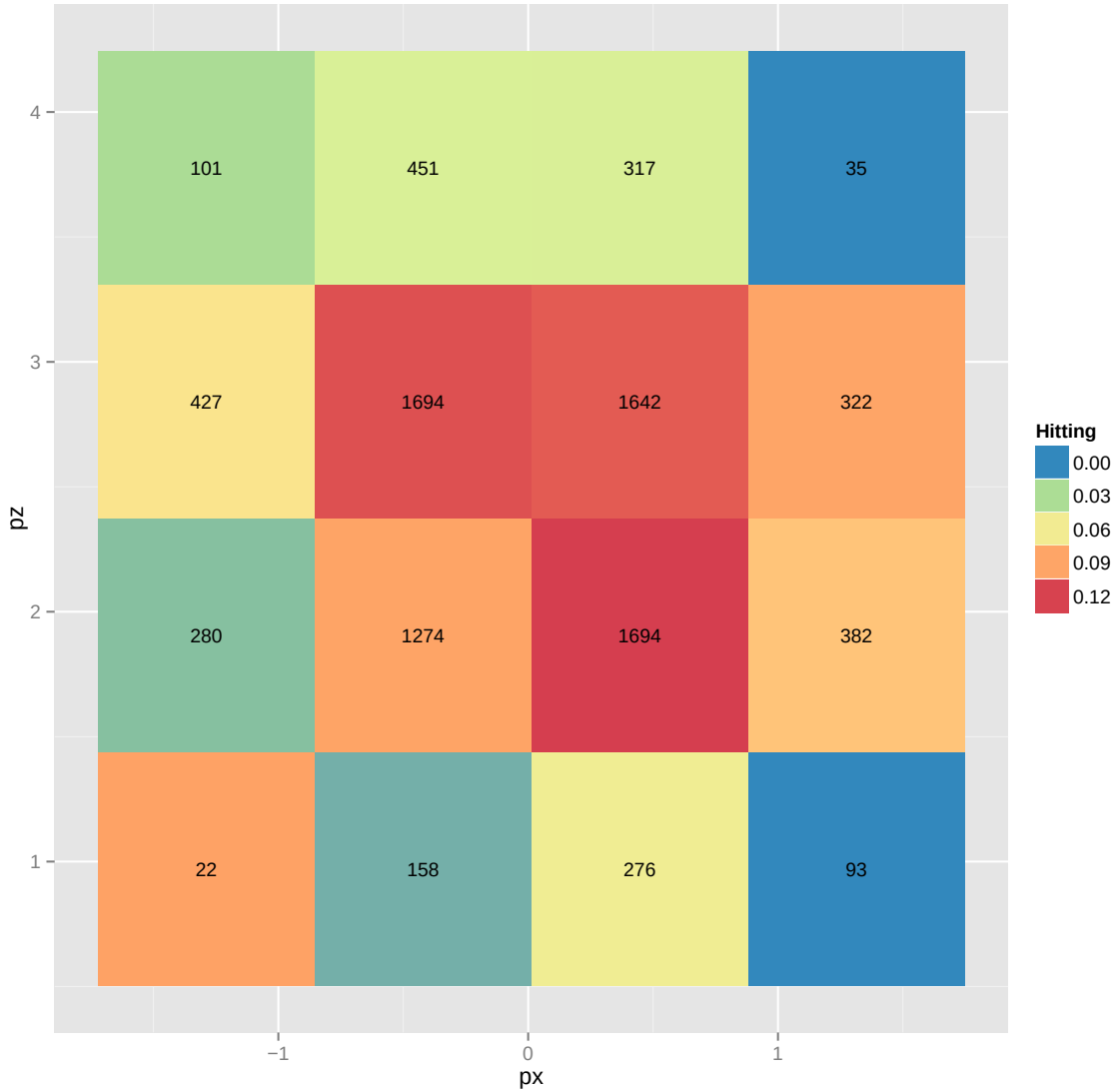
Heat Map Resolution, Jhonny Peralta

Decisions, Decsions...



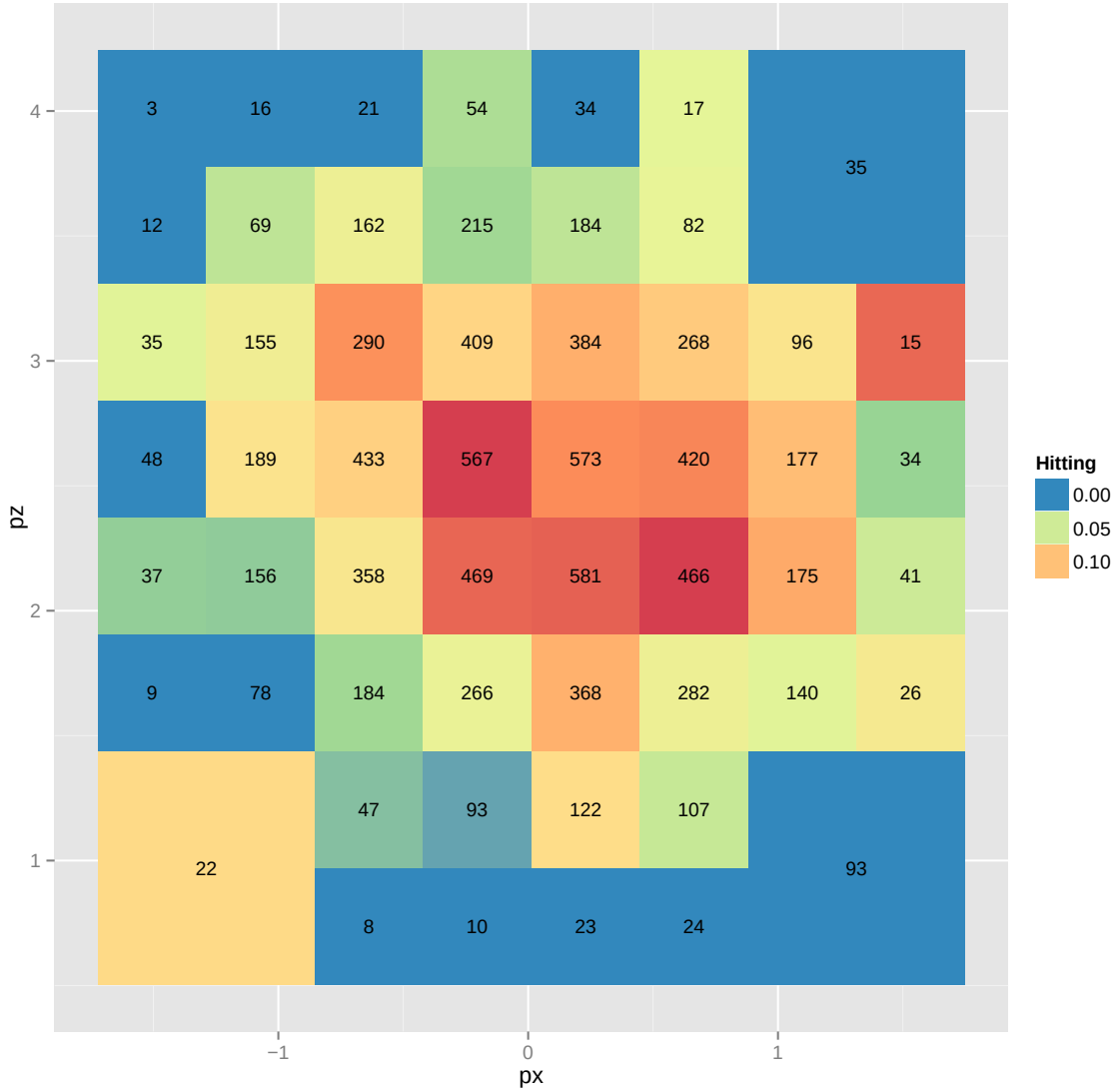
Heat Map Resolution, Jhonny Peralta

Decisions, Decsions...



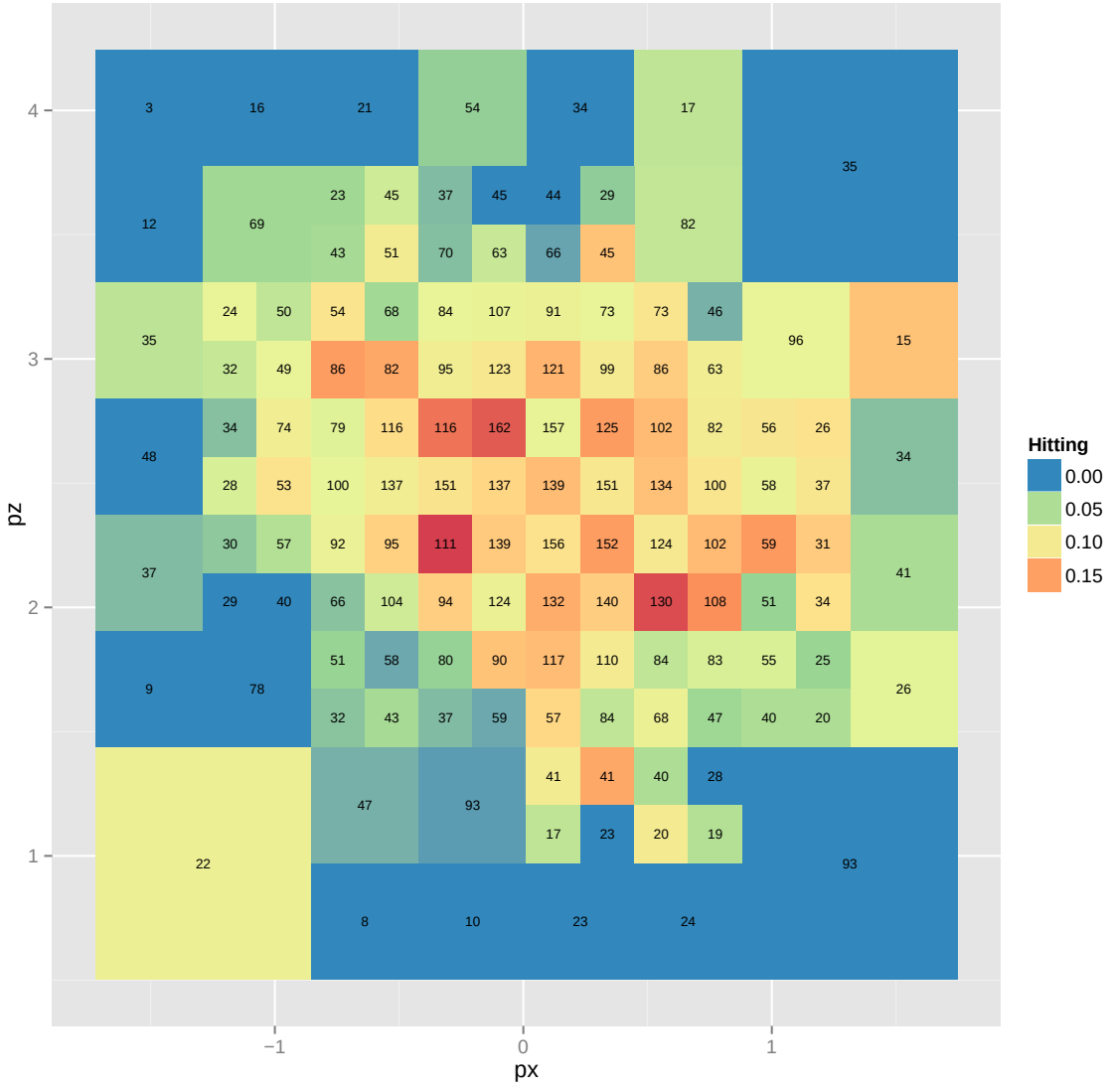
Heat Map Resolution, Jhonny Peralta

Decisions, Decsions...



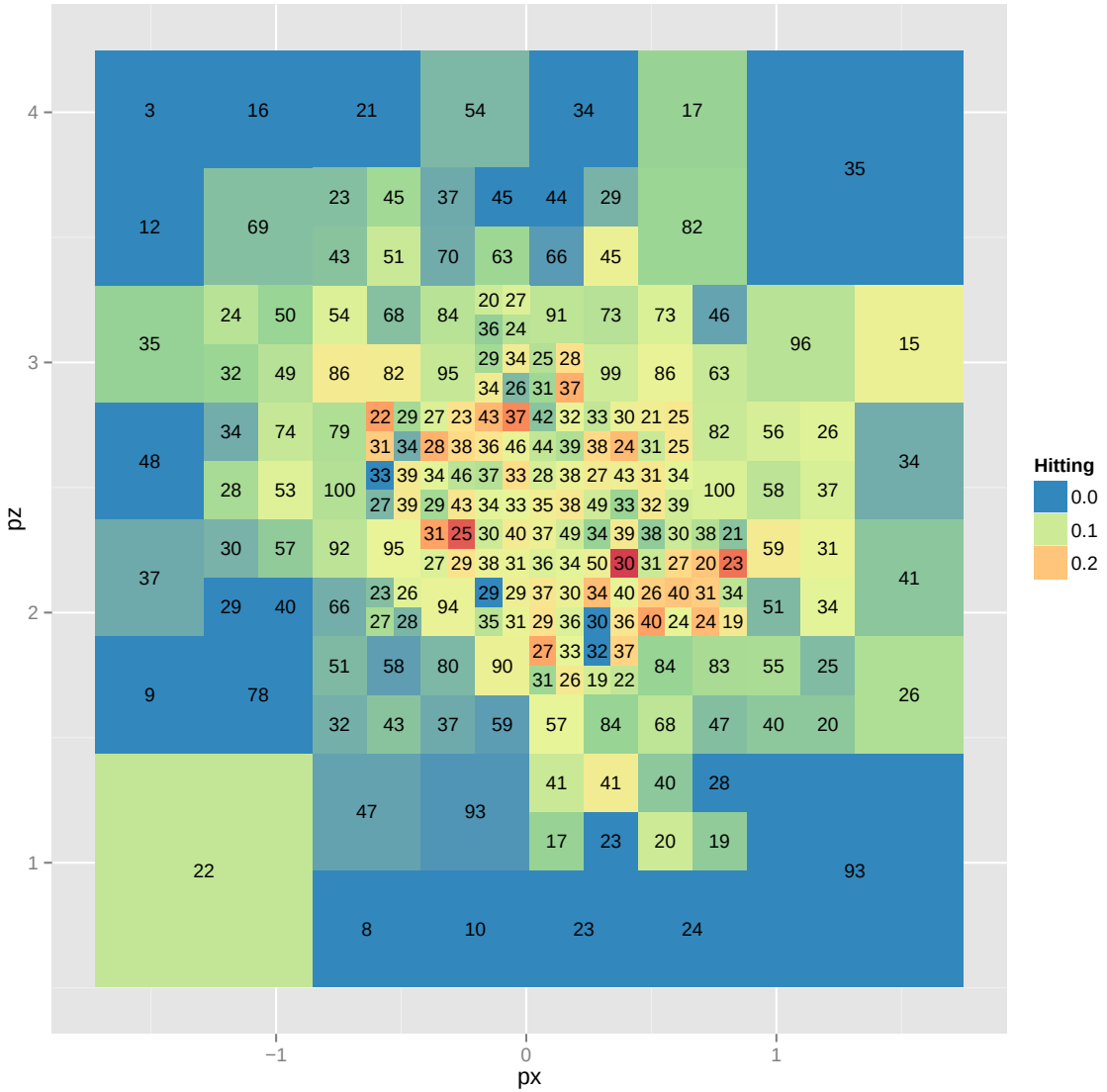
Heat Map Resolution, Jhonny Peralta

Decisions, Decsions...



Heat Map Resolution, Jhonny Peralta

Decisions, Decsions...



Best of Both (Spatial) Worlds

- Location specific resolution, rather than one global resolution
- Stopping rule $\iff$ **sample size** threshold
- Local resolution conveys data density
  - Big box $\iff$ low density
  - Bigger box, bigger variance*
- Choose subdivision algorithm

Empirical $\rightarrow$ Model

- Seek smooth(er) surface
- Models show relationships
- Models help understanding
- Modeling is fun!

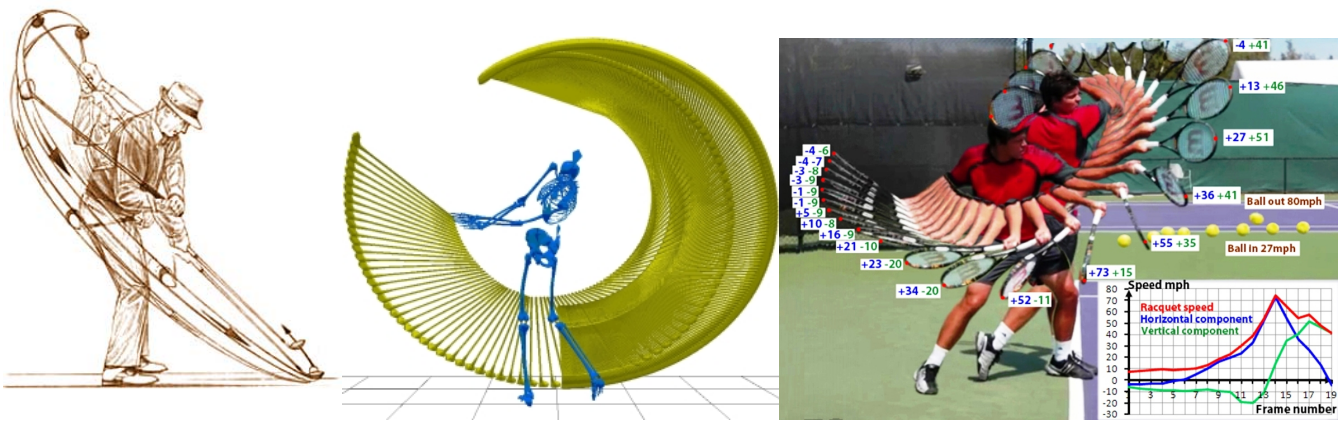
Let's Model!

- Hitter, pitcher, umpire
- Location, pitch type, previous pitch type/location
- Game state; count, baserunners
- Ballpark, day/night, weather

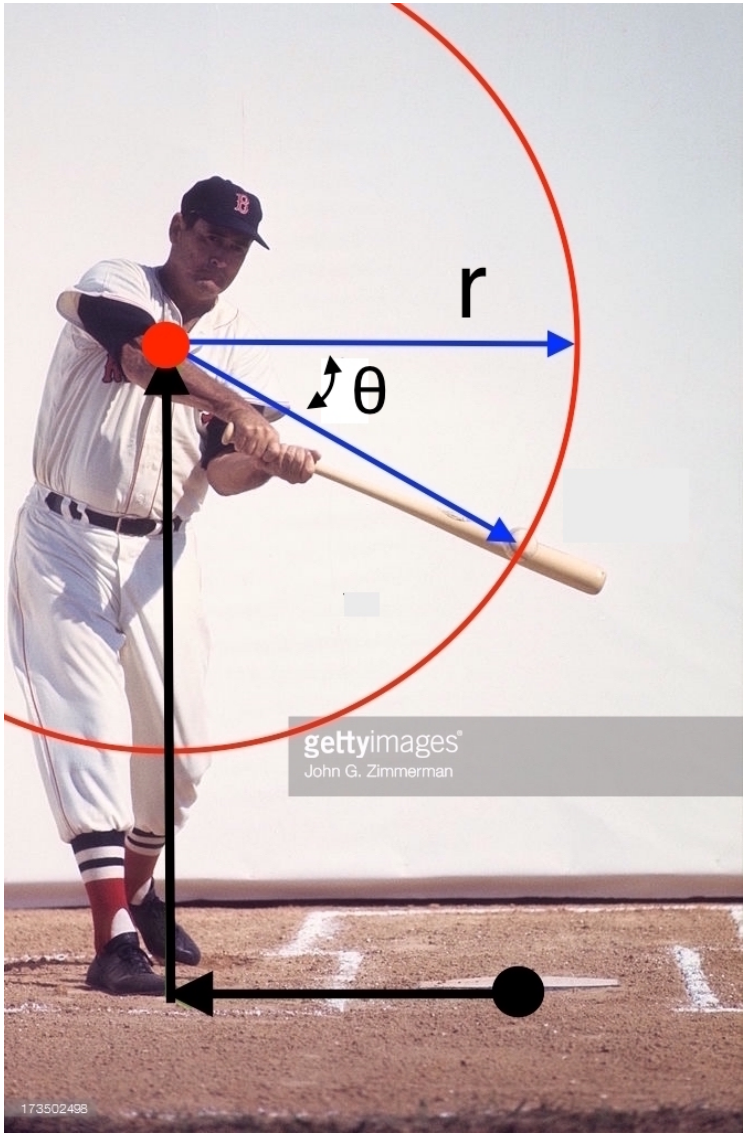
Simplifying assumption: success probabilities depend only on location (i) and hitter (j):

- $Y_{ij} \sim \text{Bernoulli}(p_{ij})$
- $\text{E}[Y_{ij}] = p_{ij}$
- $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta$
  - covariate vector $\boldsymbol{X}_{ij}$, parameter vector β
- Covariates? Biomechanics.

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



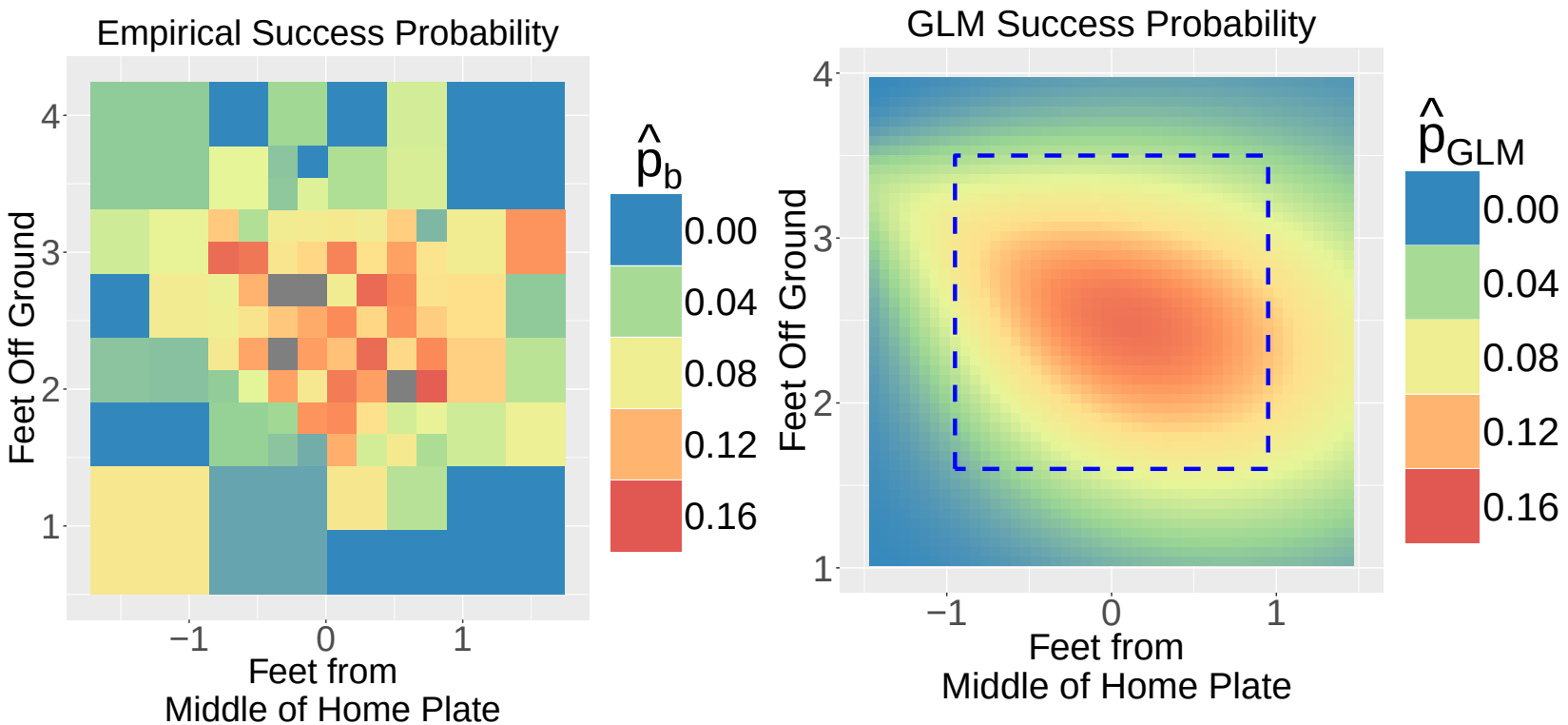
Baseball Biomechanics



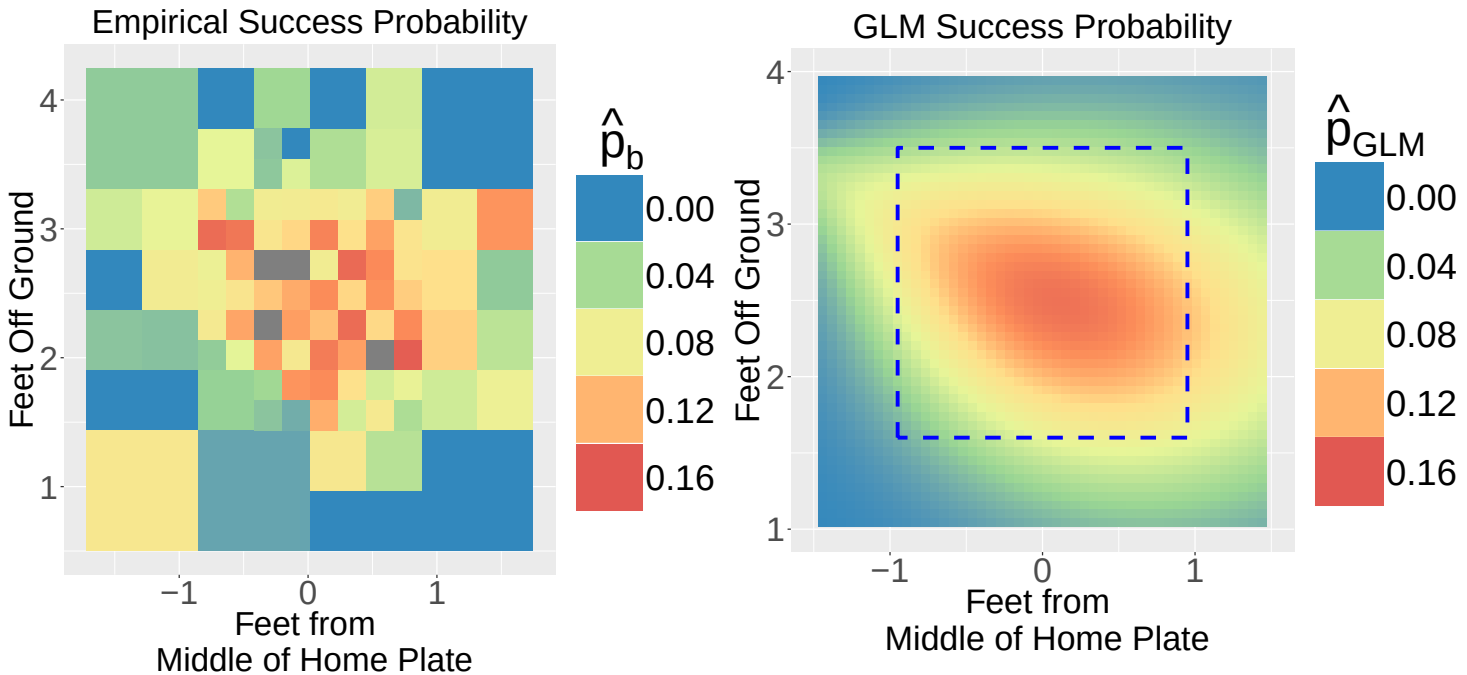
- Polar coordinates
- Translate origin
- Radius and angle biomechanically meaningful
- Big challenge: new origin location
  - Fleisig, ASMI
  - Dowling, Motus
- Let's model Jhonny Peralta, Mr. 9172

Logistic Regression Model, Jhonny Peralta

- $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta$
  - swing i, hitter j = Jhonny Peralta



Logistic Regression Model, Jhonny Peralta

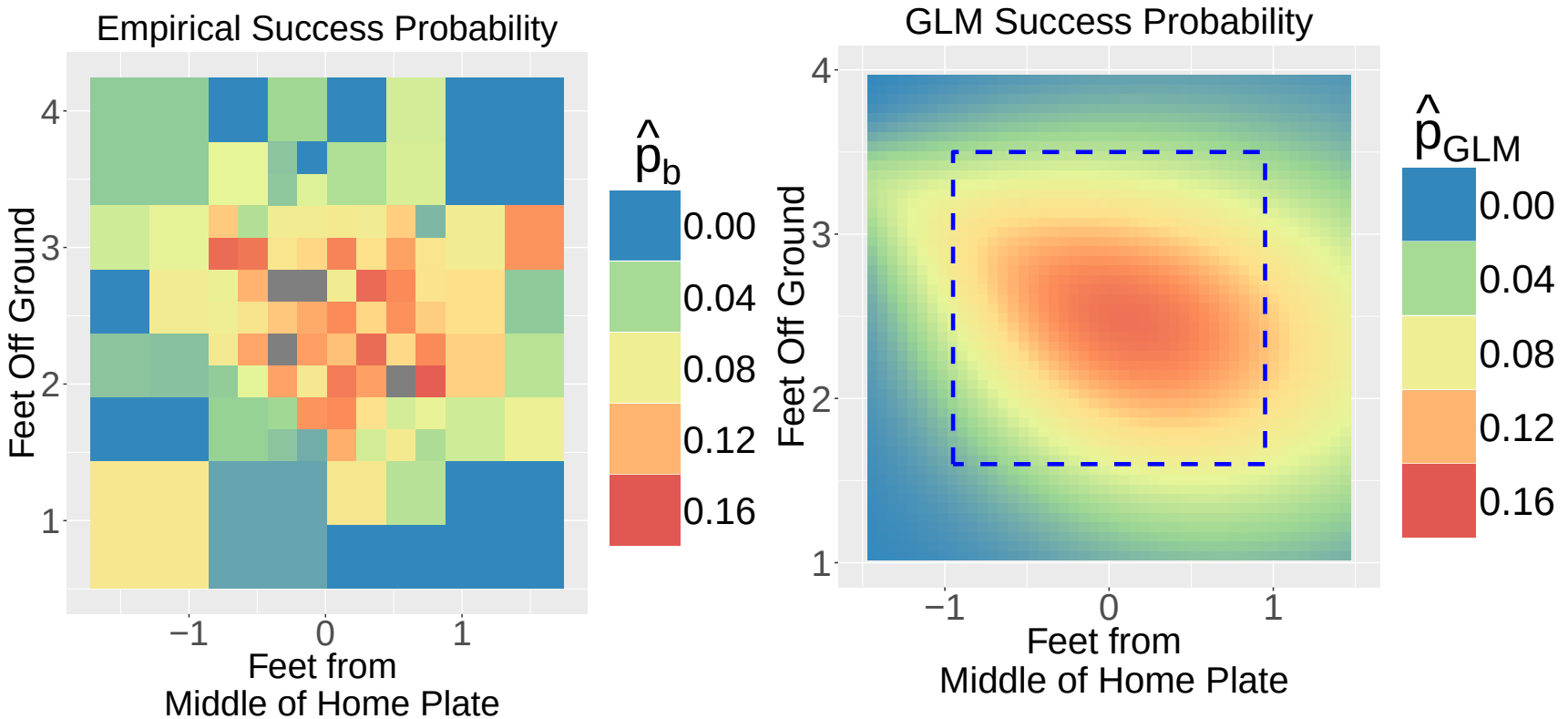


- Hosmer-Lemeshow (logistic regression) GOF test
  - Like Pearson χ^2 test, but cont. covariates (no categories)
  - H_0: Well fit
    H_A: Lack of fit
  - p-value = 0.8217

Logistic Regression Model

What about a confidence interval?

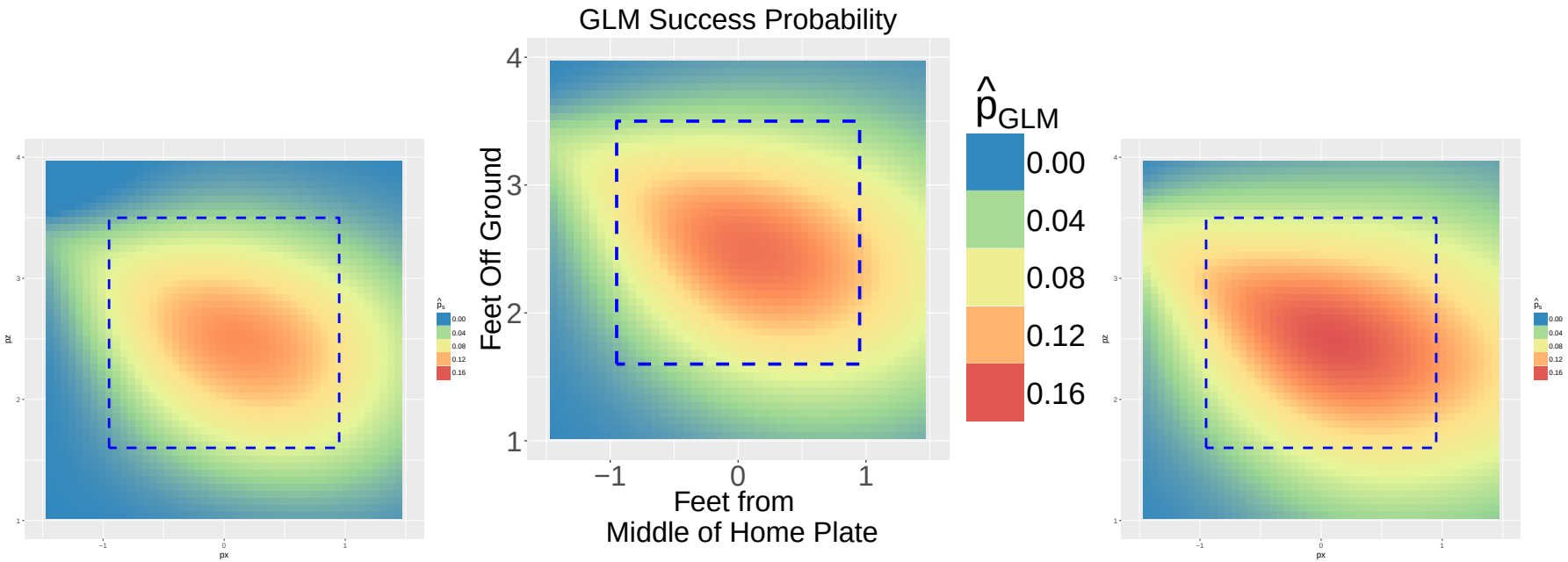
- $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta$
  - swing i, hitter j



Logistic Regression Model

The State of the Art

- $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta$



- Think about 1D and 2D CIs

Meet SHINY!!!

Spatial Logistic Regression Model

What about a spatial random effect?

- $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta + w_{ij}$
- Missing covariates
- Tobler's First Law of Geography
- Unexplained **spatial** variation in the mean

$\rightarrow$ Spatial Generalized Linear Mixed Model (SGLMM)

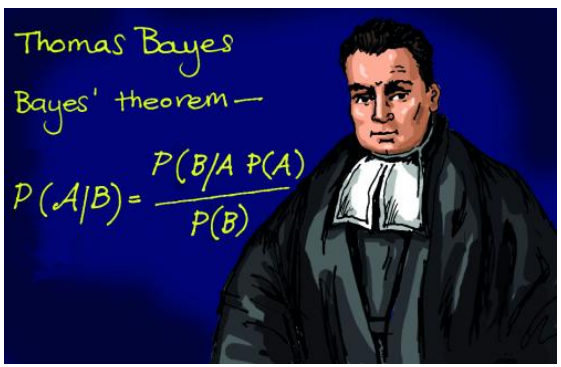
Spatial Generalized Linear Mixed Models

- SGLMM: $\text{logit}(p_j) = \boldsymbol{X}_{ij}\beta + w_{ij}$
- Gaussian Random Field (GRF)
  - $\boldsymbol{w}|\boldsymbol{\theta} \sim MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$
- Covariance function
  - $\Sigma(\phi, \sigma^2)_{i,j} = \sigma^2 exp(-||\boldsymbol{s}_i - \boldsymbol{s}_j||/\phi)$
  - $||s_i - s_j||$ - Euclidean distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$
  - σ^2 - scale parameter
  - ϕ - range parameter.

Challenges

$$\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta + w_{ij}$$

- We do not observe p_{ij}, instead Bernoulli(p_{ij}) trial outcome
- We do not observe n latent w_{ij}
- w_{ij} has a correlation structure, unknown correlation parameters $\rightarrow$ p_{ij} has complicated correlation structure
- Classical methods infeasible
  - Imagine writing, maximizing the likelihood, partial derivatives



Bayesian Hierarchical Model

$$f(\theta|Y) \propto f(Y|\theta)f(\theta)$$

- $Y_{ij}|p_{ij} \sim \text{Bernoulli}(p_{ij})$
- $\text{logit}(p_{ij}) = \boldsymbol{X}_{ij}\beta + w_{ij}$
  - $\beta \sim f_\beta(\beta)$
  - $\boldsymbol{w}_{ij}|(\sigma^2, \phi) \sim \text{MVN}(\mathbf{0}, \Sigma_{\sigma^2,\phi})$
  - $\sigma \sim f_\sigma$, $\phi \sim f_\phi$
- $f(\beta_j, \sigma, \phi|Y_{ij}) \propto f(Y_{ij}|\beta_j, w) \cdot f_\beta(\beta_j) \cdot f(w|\sigma, \phi) \cdot f_\sigma(\sigma) \cdot f_\phi(\phi)$

Great structure, now implementation.

Markov Chain Magic

Markov Chain Monte Carlo (MCMC)

- Most popular Bayesian approach, by far
- Target and update parameters (or groups of) in sequence, procedurally reduced dimension
- Produce **samples from posterior**; no closed form
- It works!
- Can take time, converges asymptotically, correlated draws

There's just one more thing.

One More Big (N) Problem

- GRF $\boldsymbol{w}_j$, recall $n \times n$ correlation matrix
- Recall multivariate normal pdf
  - $n \times n$ correlation matrix $\rightarrow$ invert $n \times n$ matrix
  - Jhony Peralta: 9172×9172!
- Iterative MCMC approach, invert lots of high dimension matrices
- Known as “Big N” problem

How bad can it be? Plus, we have physics, and Stan!!

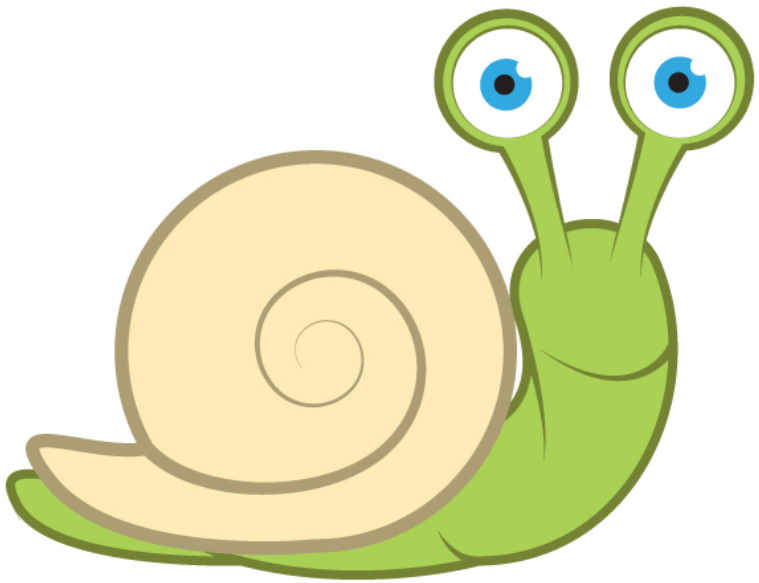
Hamiltonian Markov Chain (HMC) in Stan

- Proposal distribution - key component of MCMC algorithm
- Stan proposal machinery: HMC, from physics (molecular motion)
  - $H(q(t), p(t)) = U(q(t)) + K(p(t))$
  - Total energy of a system
  - t = time
  - **q = position (variables of interest)**, U = potential energy
  - p = momentum (auxiliary), K = kinetic energy
- $H(q, p) = -\log f_q(q) + p^T \boldsymbol{M}^{-1} p/2$
- Tidy partial derivatives
- **Short version**: randomly sample momentum p (auxiliary), calculate new position q — that's your Metropolis proposal.

Stan will be unstoppable!

Slow, Slow, Slow

Meet Stan the snail.



Slow, Slow, Slow

Meet Stan the snail.

How slow? Prohibitively slow. Think geologic time scales; earth is 4.54 billion years old.

Optimization

Give Stan a turbo-boost!

Stan Optimization

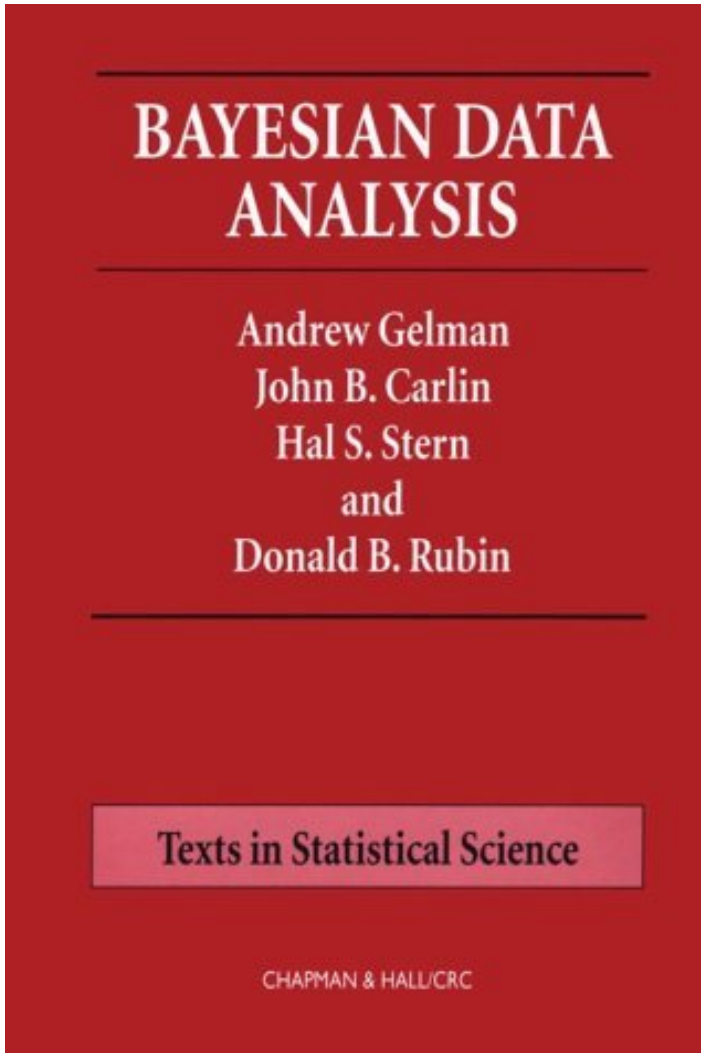
Stan Modeling Language

User's Guide and Reference Manual

Stan Development Team

Stan Version 2.12.0

Tuesday 6th September, 2016



- Stan Development Team Members
  - Andrew Gelman
  - Bob Carpenter
  - Rob Trangucci

Stan Optimization

- Informative (sharp tailed) prior on exponential covariance length-scale parameter ϕ for model identifiability (normal or log-normal)
- Computing time/convergence - “Complex models” such as spatial hierarchical models require proper priors for all β coefficients.
- QR factorization on covariate matrix X; reparameterize, adjust priors

$$
\begin{aligned}
\boldsymbol{X} &= \boldsymbol{QR} \\
\boldsymbol{X}\beta &= \boldsymbol{QR}\beta \rightarrow \text{Let } \boldsymbol{\theta} = \boldsymbol{R}\beta \\
\boldsymbol{X}\beta &= \boldsymbol{Q}\boldsymbol{\theta} \\
\beta &= \boldsymbol{R}^{-1}\boldsymbol{\theta}
\end{aligned}
$$

Stan Optimization

- Add noise to the covariance matrix diagonal, ensure numerical positive-definiteness
- Cholesky decomposition and tactical reparameterization

```
L = cholesky_decompose(Sigma);
Z ~ normal(0, 1);
Z_mod = L * Z;
Y ~ bernoulli_logit(Q*theta + Z_mod);
```

- Matrix algebra and vectors, over 'for loops' and scalars

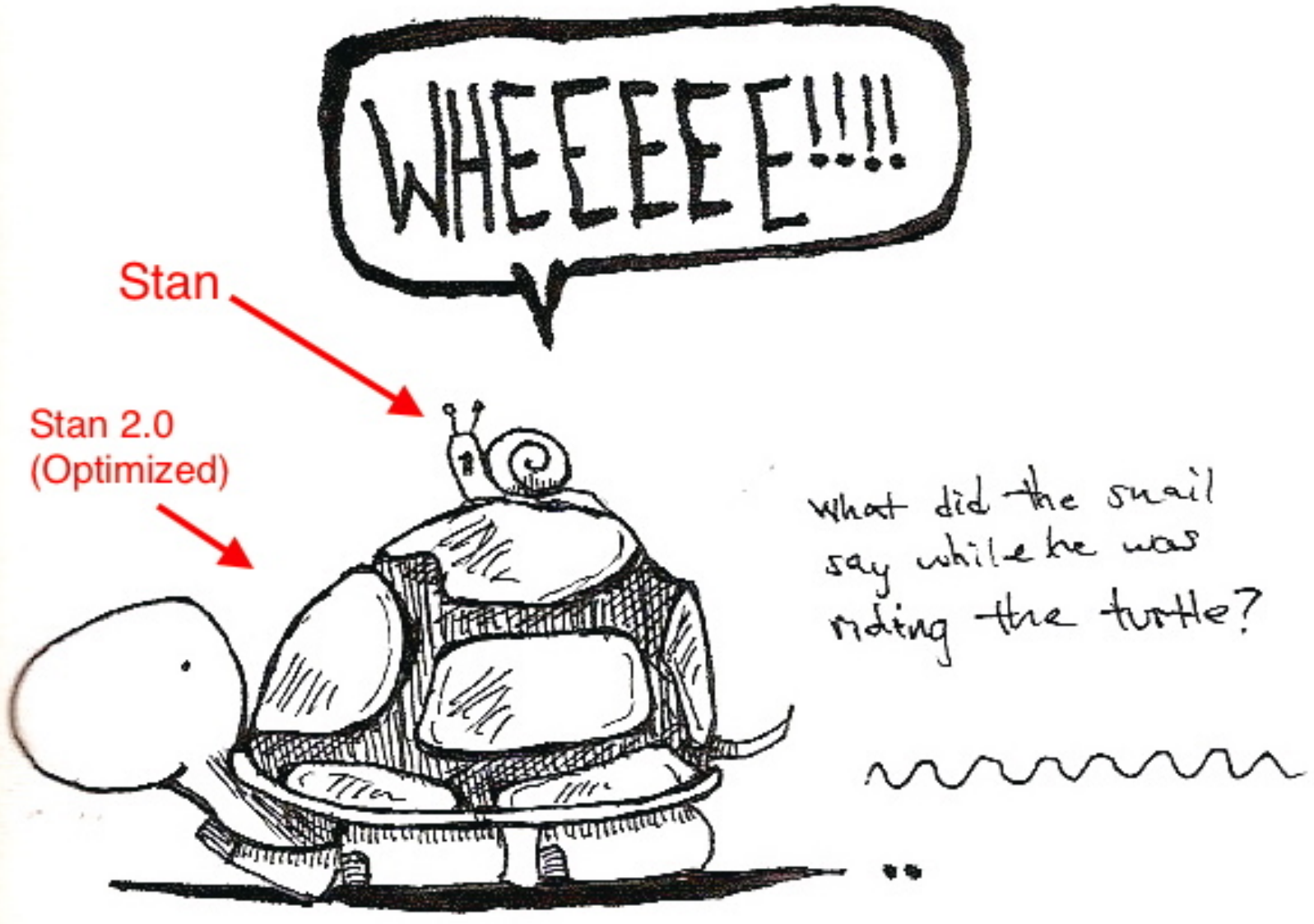
```
hit ~ bernoulli_logit(X*beta + Z)
```

is faster than

```
for (n in 1:N)
    hit[n] = bernoulli_logit(X[n]*beta[n] + Z[n]);
```

On your mark. Get set...

Go Stan 2.0!


WHEEEEE!!!!
Stan
Stan 2.0
(Optimized)
what did the snail
say while he was
riding the turtle?

How Slow is Stan 2.0?

- N = 1000: 7 hours 45 mins for 1500 draws
  - Note: w/o spatial effect: 6 seconds.
- N = 2000 overnight, 350 draws

The “Big N” Problem Revisited

- SGLMM computational costs
  - Increase at a rate of $\mathcal{O}(n^3)$:

$$t(n) \leq M \cdot n^3 \text{ as } n \to \infty$$

  - Spatial random effect covariance matrix Σ is $n \times n$
  - Every iteration requires Σ^{-1}
  - Determinant of Σ
- Computational bottleneck

Persistence

Secret 7. You can do it. A PhD is 10% intelligence and 90% persistence.

Predictive Process Models (PPMs)

Dimension Reduction

NCAR workshop, Andrew Finley

- MSU Deptartments of Forresty and Geography

“Gaussian predictive process models for large spatial data sets” [Banerjee et al., 2008]

- Competitive approach with computational advantages
- Project original process onto lower dimensioned subspace; generate parent process realizations at knots
- “Spatial sampling design for parameter estimation of the covariance function” [Zhu and Stein, 2005]

PPM Procedure

Nuts and Bolts

$$\text{logit}(p_{ij}|\boldsymbol{s}_{ij}) = \boldsymbol{X}_{ij}(\boldsymbol{s}_{ij})\beta_j + w(\boldsymbol{s}_{ij})$$
$$\boldsymbol{w}(\boldsymbol{s})|\boldsymbol{\theta} \sim \text{GRF}(\boldsymbol{0}, \boldsymbol{C}(\boldsymbol{\theta}))$$

- $\boldsymbol{w}(\boldsymbol{s})$ - $n \times 1$ vector of random effects, at locations $\boldsymbol{s}$
- $\boldsymbol{0}$ - $n \times 1$ zero vector
- $\boldsymbol{C}(\boldsymbol{\theta})$ - $n \times n$, symmetric, positive-definite covariance matrix
  - $\boldsymbol{\theta}$ - covariance parameters
- $C(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta})$ - covariance of random effects at locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$
  - $C(\boldsymbol{\theta}) = [C(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta})]_{i,j=1}^{n}$

PPM Knots

- $\boldsymbol{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_m^*\}$ — m knot locations
  - $m << n$
- $\boldsymbol{w}^* = \left[w(\boldsymbol{s}_i^*)\right]_{i=1}^m$ – knot location random effects
- $\boldsymbol{C}^*(\boldsymbol{\theta}) = \left[C(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*)\right]_{i,j=1}^m$ – knot covariance matrix
  - $m \times m$ — much smaller dimension
- $\boldsymbol{w}^* | \boldsymbol{\theta} \sim \text{GRF}\{\boldsymbol{0}, \boldsymbol{C}^*(\boldsymbol{\theta})\}$ – m-dimensional GRF

PPM, Krig $w(\boldsymbol{s}_0)$

- Interpolate $w(\boldsymbol{s}_0)$ using: (i) m knots, (ii) parent process covariance structure, (iii) kriging procedure
- $\tilde{w}(\boldsymbol{s}_0)$ - interpolated random effect

$$\begin{aligned}\tilde{w}(\boldsymbol{s}_0) &= E[w(\boldsymbol{s}_0)|\boldsymbol{w}^*] \\ &= \boldsymbol{c}^T(\boldsymbol{s}_0;\boldsymbol{\theta}) \cdot \boldsymbol{C}^{*-1}(\boldsymbol{\theta}) \cdot \boldsymbol{w}^*\end{aligned}$$

- $\boldsymbol{c}(\boldsymbol{s}_0;\boldsymbol{\theta}) = \left[C(\boldsymbol{s}_0, \boldsymbol{s}_j^*;\boldsymbol{\theta})\right]_{j=1}^{m}$ — $m \times 1$ vector, Cov$(w(\boldsymbol{s}_0),$ knots$)$
- Spatially varying linear combination
- Minimize squared error loss function (linear predictors, GRFs)

Now replace $w(\boldsymbol{s})$ with $\tilde{w}(\boldsymbol{s})$.

The Predictive Process! $\tilde{w}(\boldsymbol{s})$

Another Gaussian Random Field

$$\tilde{\boldsymbol{w}}(\boldsymbol{s}) \sim \text{GRF}\{0, \tilde{C}(\cdot)\}$$
$$\tilde{C}(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta}) = \boldsymbol{c}^T(\boldsymbol{s}; \boldsymbol{\theta}) \cdot \boldsymbol{C}^{*-1}(\boldsymbol{\theta}) \cdot \boldsymbol{c}(\boldsymbol{s}'; \boldsymbol{\theta})$$

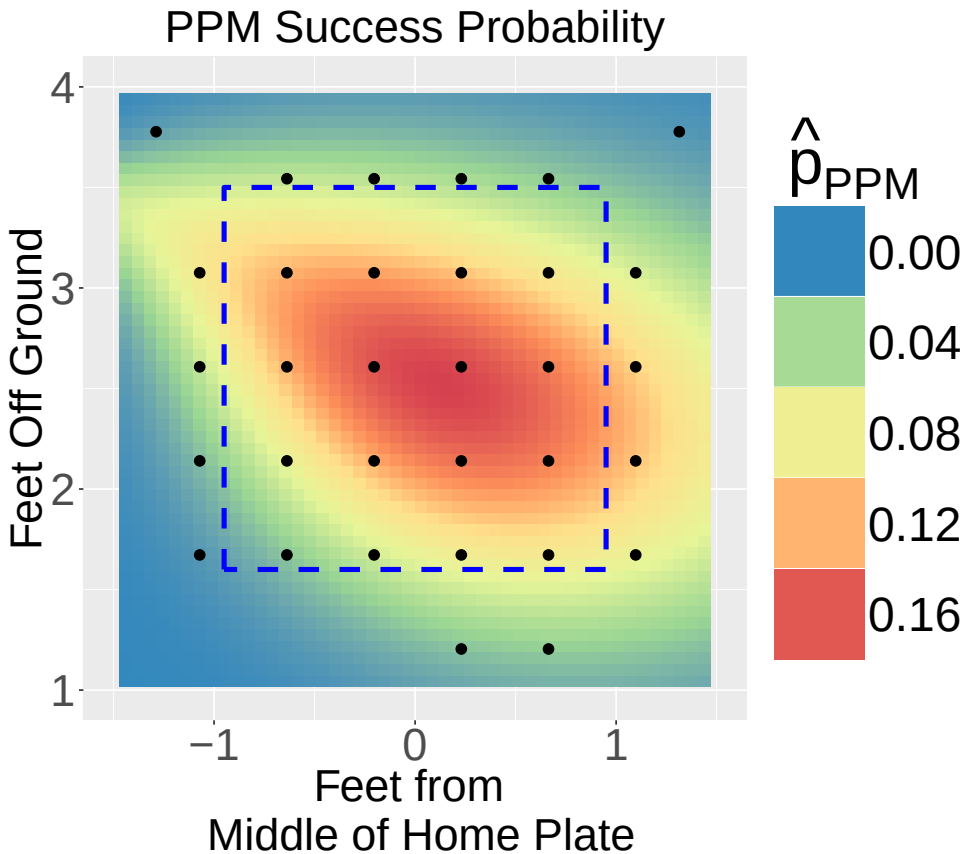
And a predictive process model!

$$\text{logit}(p_{ij}|\boldsymbol{s}_{ij}) = \boldsymbol{X}_{ij}(\boldsymbol{s}_{ij})\beta_j + \tilde{w}(\boldsymbol{s}) \tag{1}$$

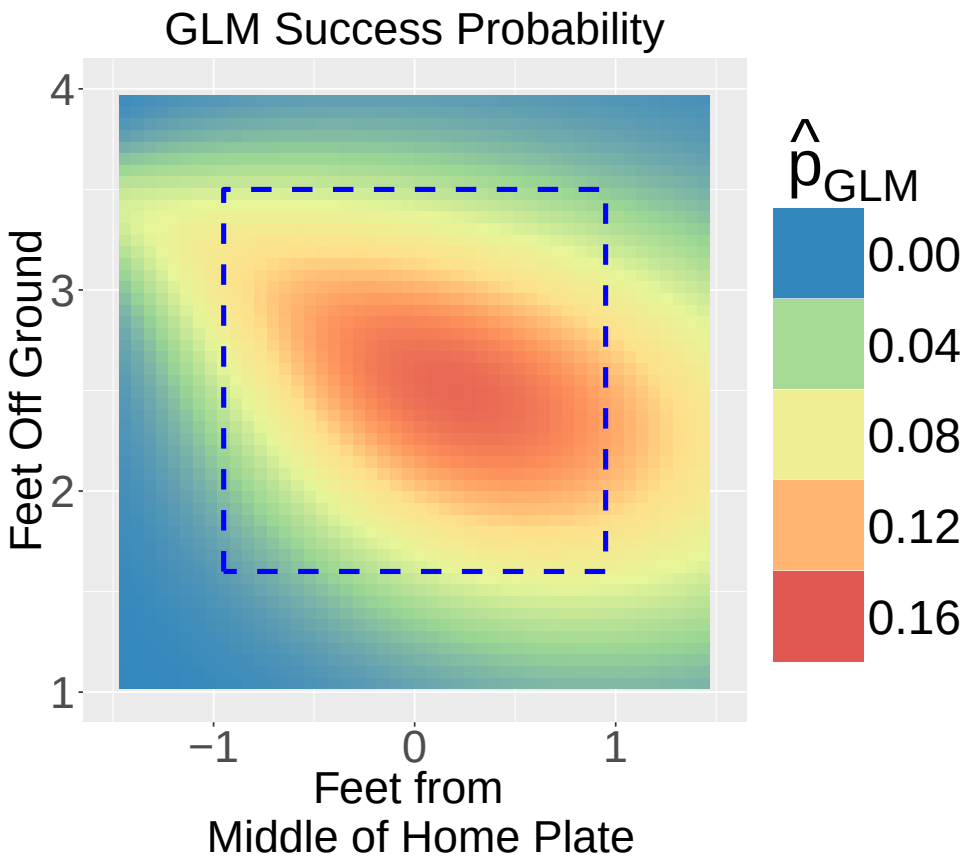
- Note dimension of covariance matrix.
- Implement in `spBayes` [Finley et al., 2007].
- Faster?

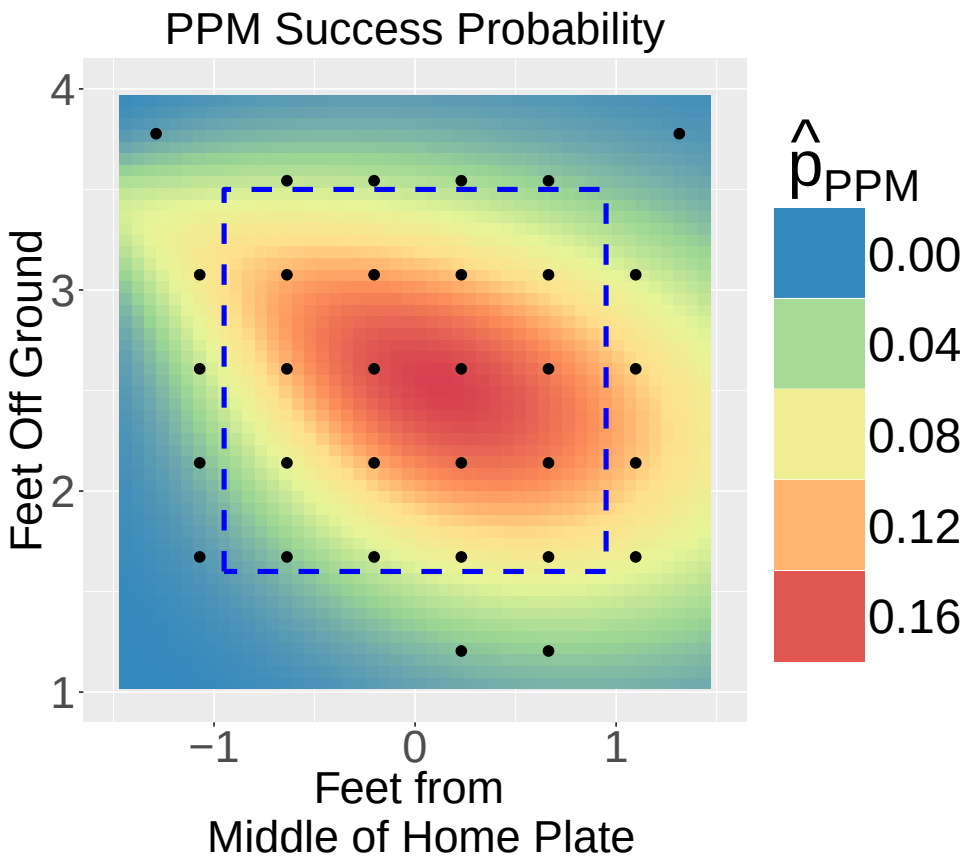
PPMs, Plot and Results

- n = 500, knots = 97, 10K samples, $\approx$ 4 mins
- 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



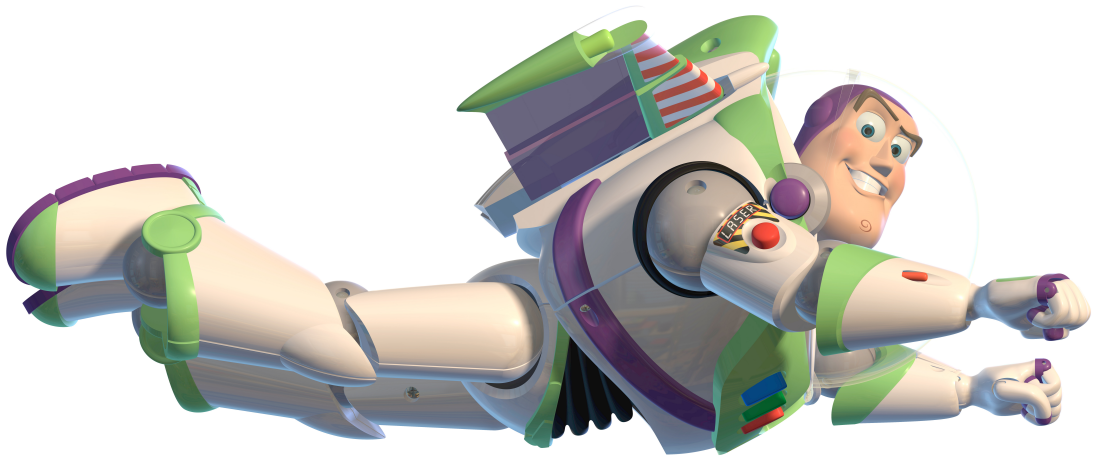
PPM, Plot and Results





Current Work, SPDE-INLA

- What about n = 9172?
- Approximation method, works well for Bayesian hierarchical models with latent GRF
- Essentially a two part process for continuous domain
  - Part 1: SPDE
  - Part 2: INLA
- To n = 9172... and beyond!!



Approximation Part 1, SPDE

SPDE - Stochastic partial differential equation [Lindgren et al., 2011]

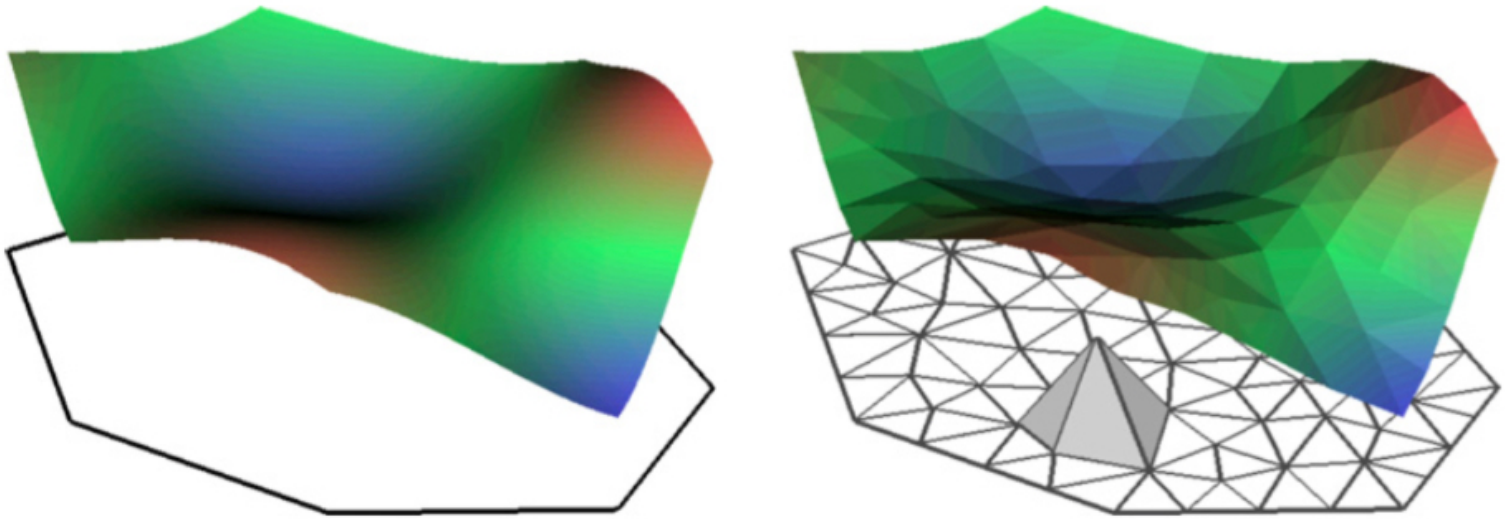
$$(\kappa^2 - \Delta)^{\alpha/2} x(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s})$$

- Convert GRF to Gaussian **Markov** Random Field (GMRF)
  - Sparse precision matrix
- Finite Element Method: project SPDE onto basis representation
- Triangular mesh, defines piecewise linear basis representation

$$\tilde{x}(\boldsymbol{s}) = \sum_k \psi_k(\boldsymbol{s}) x_k$$

  - $\psi_k(\boldsymbol{s})$ — deterministic basis functions
  - x_k — weights, explicit and sparse precision matrix
  - $\tilde{x}(\boldsymbol{s})$ approximates SPDE solution

SPDE, A Picture Worth 1000 Equations



(a) A continuous function.

(b) A piecewise linear approximation.

[Simpson et al., 2012]

Approximation Part 2, INLA

INLA - Integrated Nested Laplace Approximation [Rue et al., 2009]

1. Gaussian approximation: $\tilde{p}(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{y})$
2. Laplace approximation:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{w})}{p(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{w}=\boldsymbol{w}^*(\boldsymbol{\theta})} \approx \frac{p(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{w})}{\tilde{p}(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{w}=\boldsymbol{w}^*(\boldsymbol{\theta})}$$

   - $\boldsymbol{w}^*(\boldsymbol{\theta}) = \text{argmax}_w \tilde{p}(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{y})$

3. Numerical integration: $p(\theta_k|\boldsymbol{y}) \approx \int \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-k}$
4. Numerical integration: $p(w_j|\boldsymbol{y}) \approx \int \tilde{p}(w_j|\boldsymbol{\theta}, \boldsymbol{y}) \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$

Current Work, SPDE-INLA

- Implement in R-INLA [Lindgren and Rue, 2015]
- Computational costs - increase at a rate of $\mathcal{O}(n^{3/2})$
- Trade-offs - speed for bias (Binomial data)

Next Steps

- Cool SPDE-INLA pictures
- Model evaluation, scoring rules [Bickel, 2007]
- Cross validation study
- Variable-resolution heat map R package?
- Dynamic Heat Map CIs – R package?

Thanks for listening. Questions?

Rick Delgado. Beyond moneyball: How big data is changing baseball, 2014. URL http://www.sporttechie.com/2014/11/11/beyond-moneyball-how-big-data-is-changing-baseball/.

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.

Zhengyuan Zhu and Michael L Stein. Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, 134(2):583–603, 2005.

Andrew O Finley, Sudipto Banerjee, and Bradley P Carlin. spbayes: an r package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19 (4):1, 2007.

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.

Daniel Simpson, Finn Lindgren, and Håvard Rue. Think continuous: Markovian gaussian models in spatial statistics. *Spatial Statistics*, 1: 16–29, 2012.

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 and Håvard Rue. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19), 2015.

J Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49–65, 2007.