Evaluating plate discipline in Major League Baseball with Bayesian Additive Regression Trees

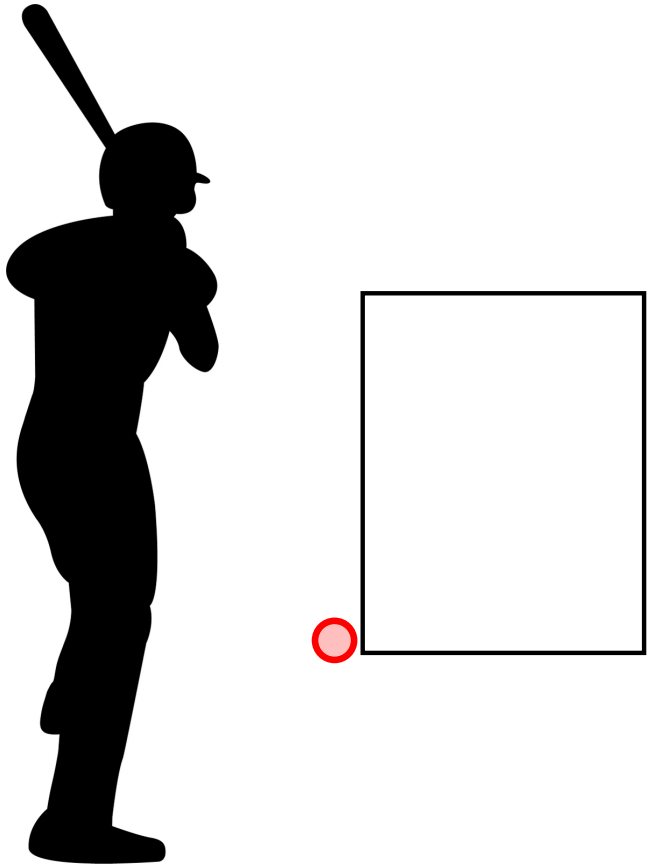
Ryan Yee

University of Wisconsin – Madison
Department of Statistics

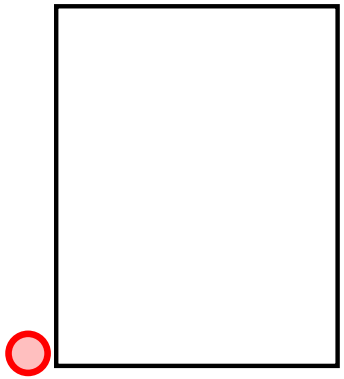
New England Symposium on Statistics in Sports

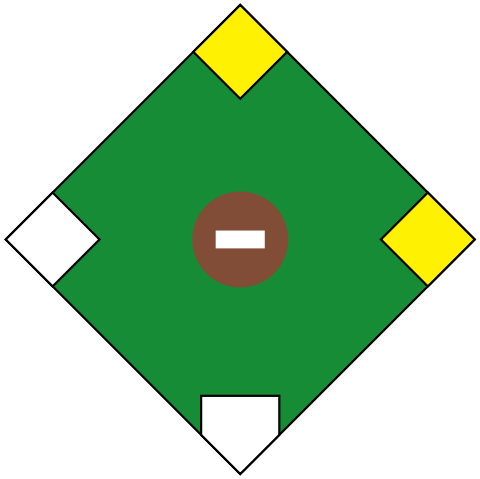
September 23, 2023

What is the optimal decision?

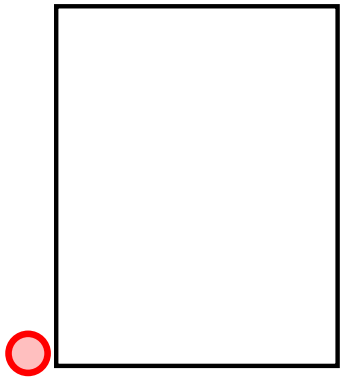


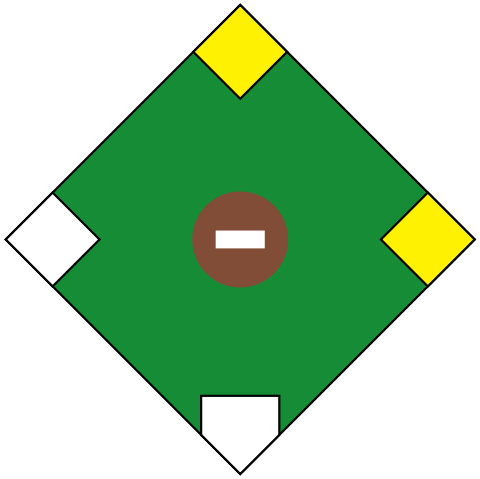
What is the optimal decision?

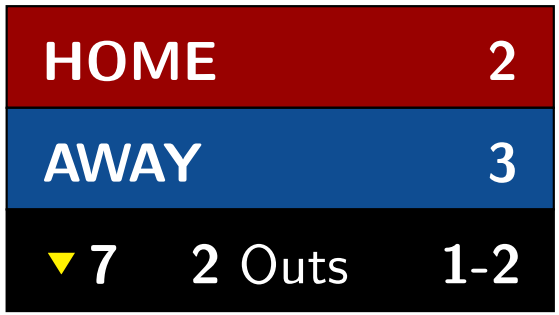




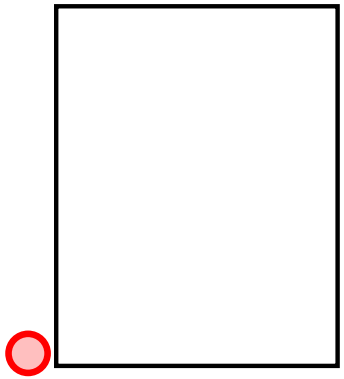
What is the optimal decision?

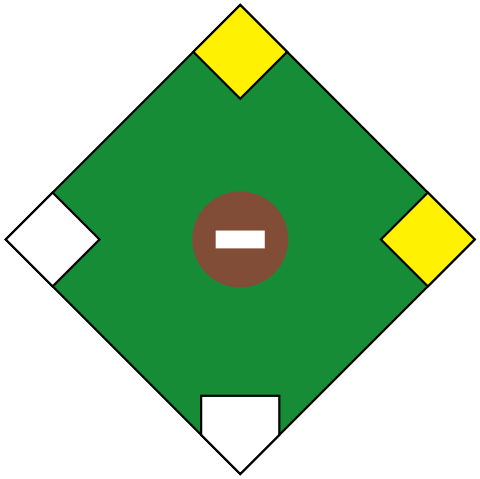


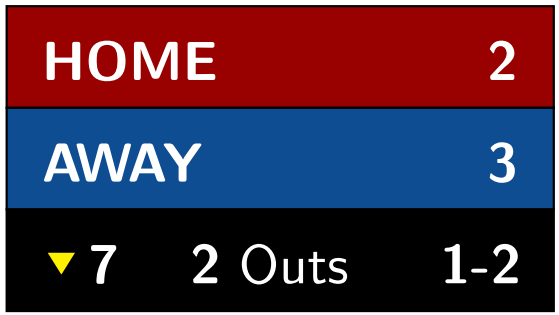




What is the optimal decision?







Batter: Shohei Ohtani

Umpire: Ángel Hernández

Evaluating Plate Discipline

- It is not enough to just look at location...
  - Analysis depends on location, game factors, and personnel
  - Some pitches are harder to evaluate that others

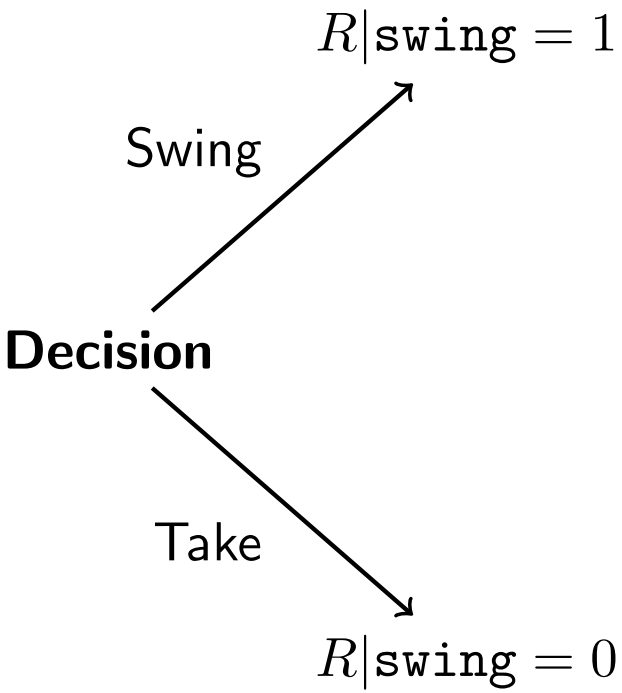
Evaluating Plate Discipline

- It is not enough to just look at location...
  - Analysis depends on location, game factors, and personnel
  - Some pitches are harder to evaluate that others
- Traditional plate discipline metrics: `O-Swing%`, `Z-Swing%`
  - Analysis relies solely on the location of the pitch
  - Treats every pitch inside (outside) the strike zone the same

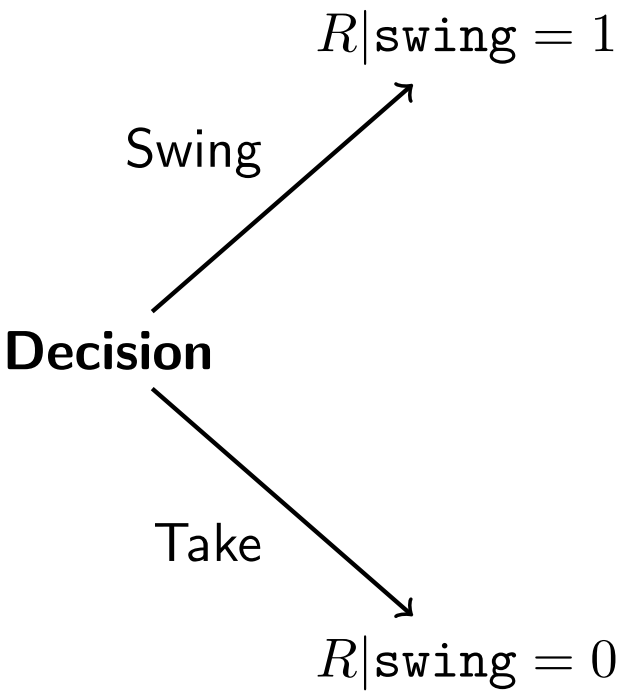
Evaluating Plate Discipline

- It is not enough to just look at location...
  - Analysis depends on location, game factors, and personnel
  - Some pitches are harder to evaluate that others
- Traditional plate discipline metrics: `O-Swing%`, `Z-Swing%`
  - Analysis relies solely on the location of the pitch
  - Treats every pitch inside (outside) the strike zone the same
- Our Contribution
  - Introduce framework for evaluating plate discipline which accounts for contextual factors in addition to location
  - Framework facilitates uncertainty quantification in evaluations

Framework

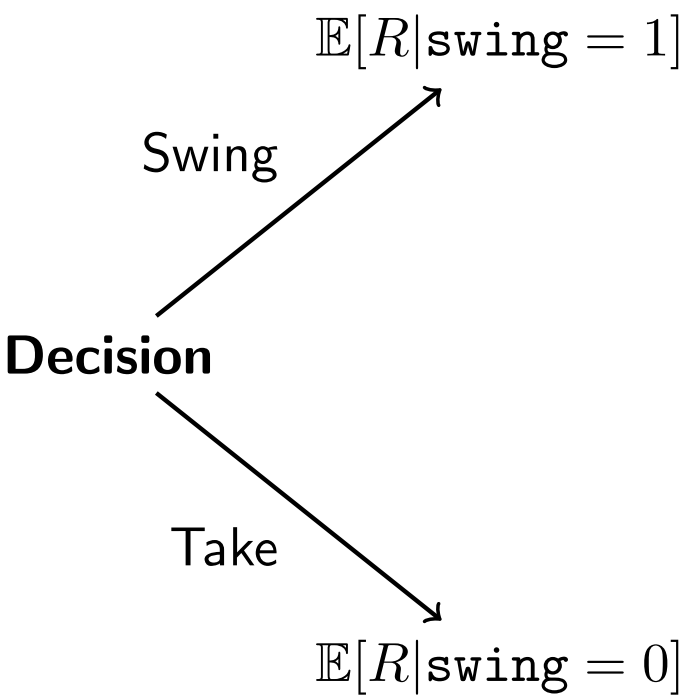


Framework



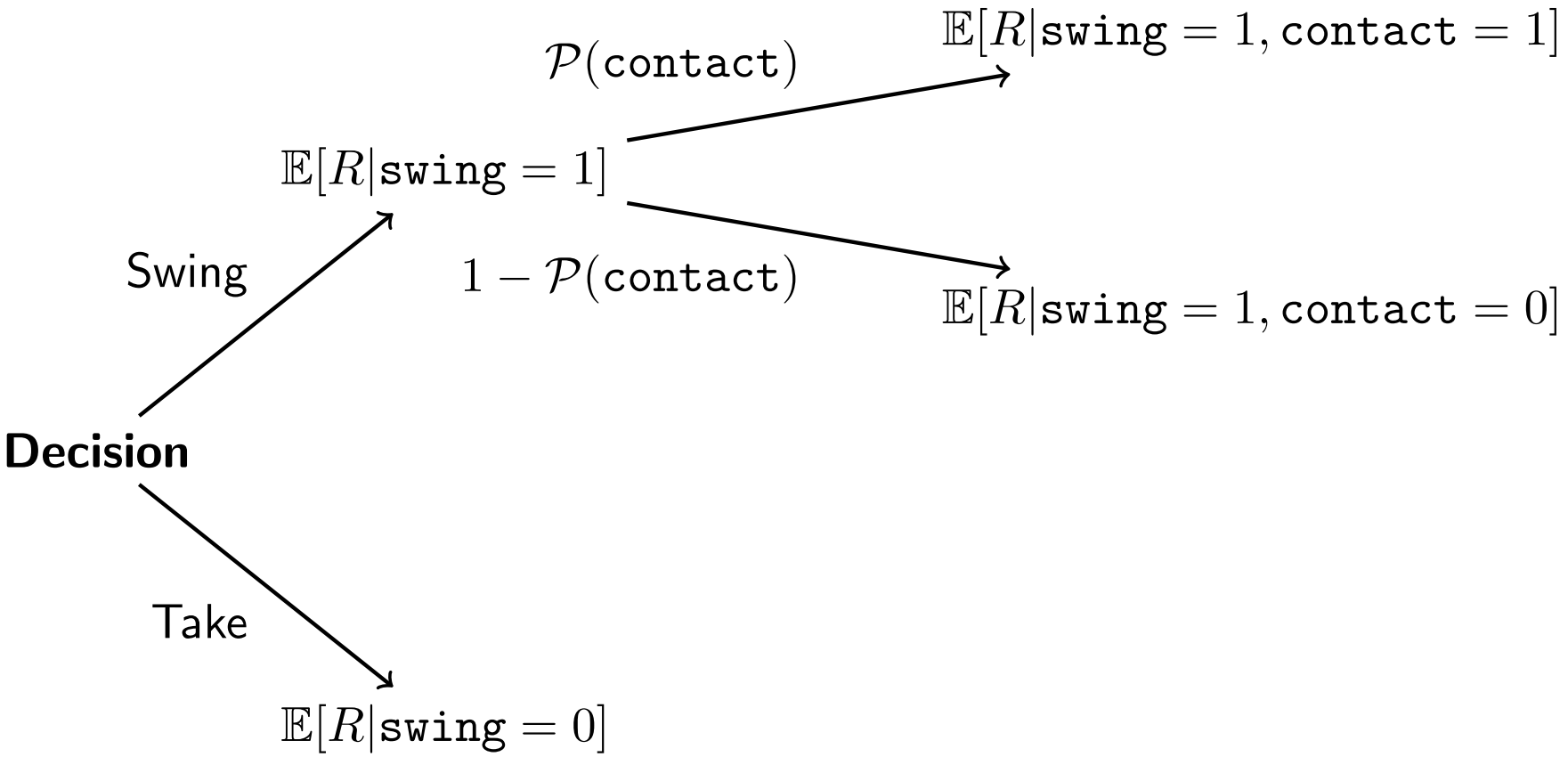
The optimal decision is to swing if $R|\texttt{swing} = 1 > R|\texttt{swing} = 0$

Framework



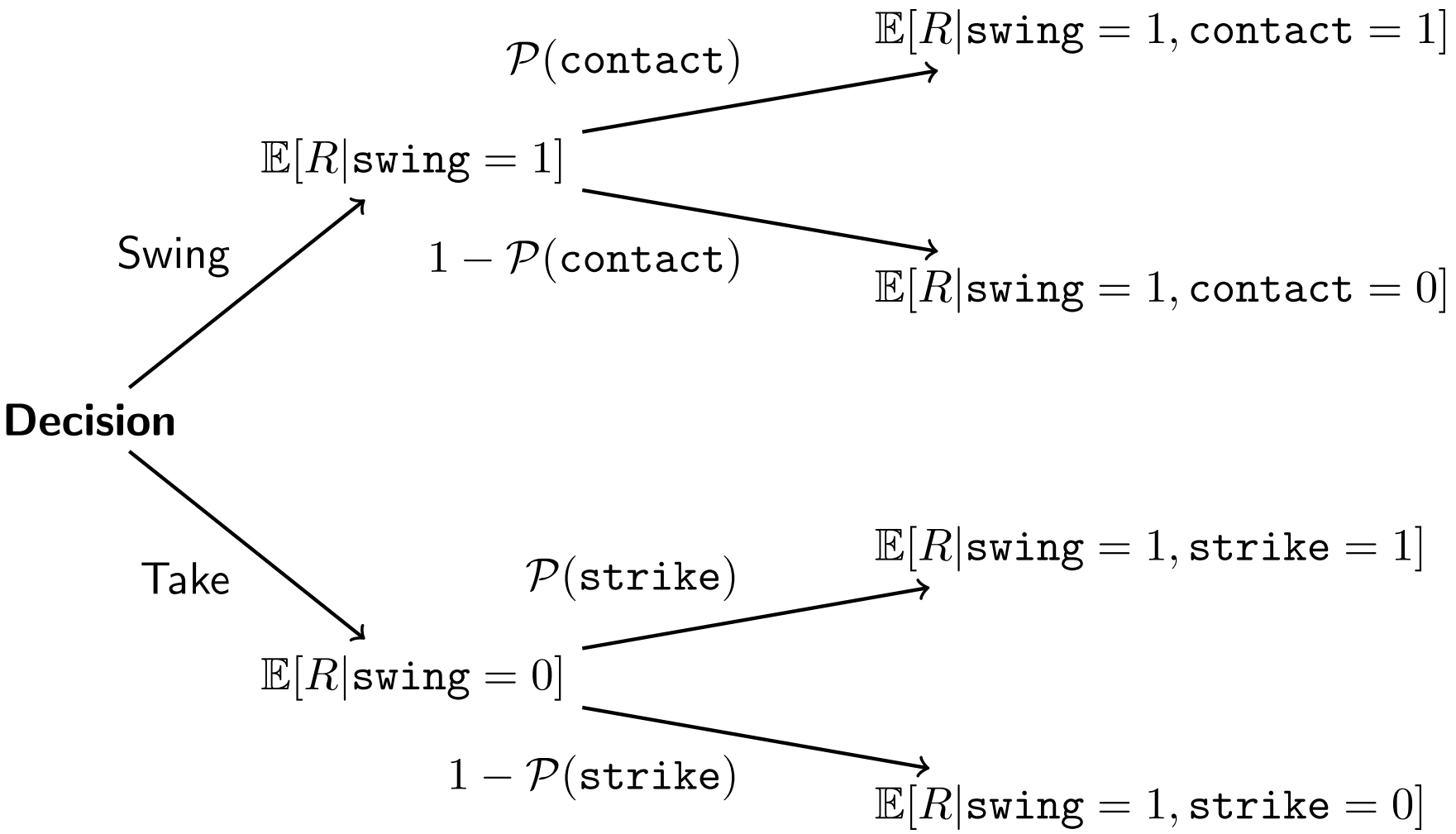
The optimal decision is to swing if $\mathbb{E}[R|\texttt{swing} = 1] > \mathbb{E}[R|\texttt{swing} = 0]$

Framework



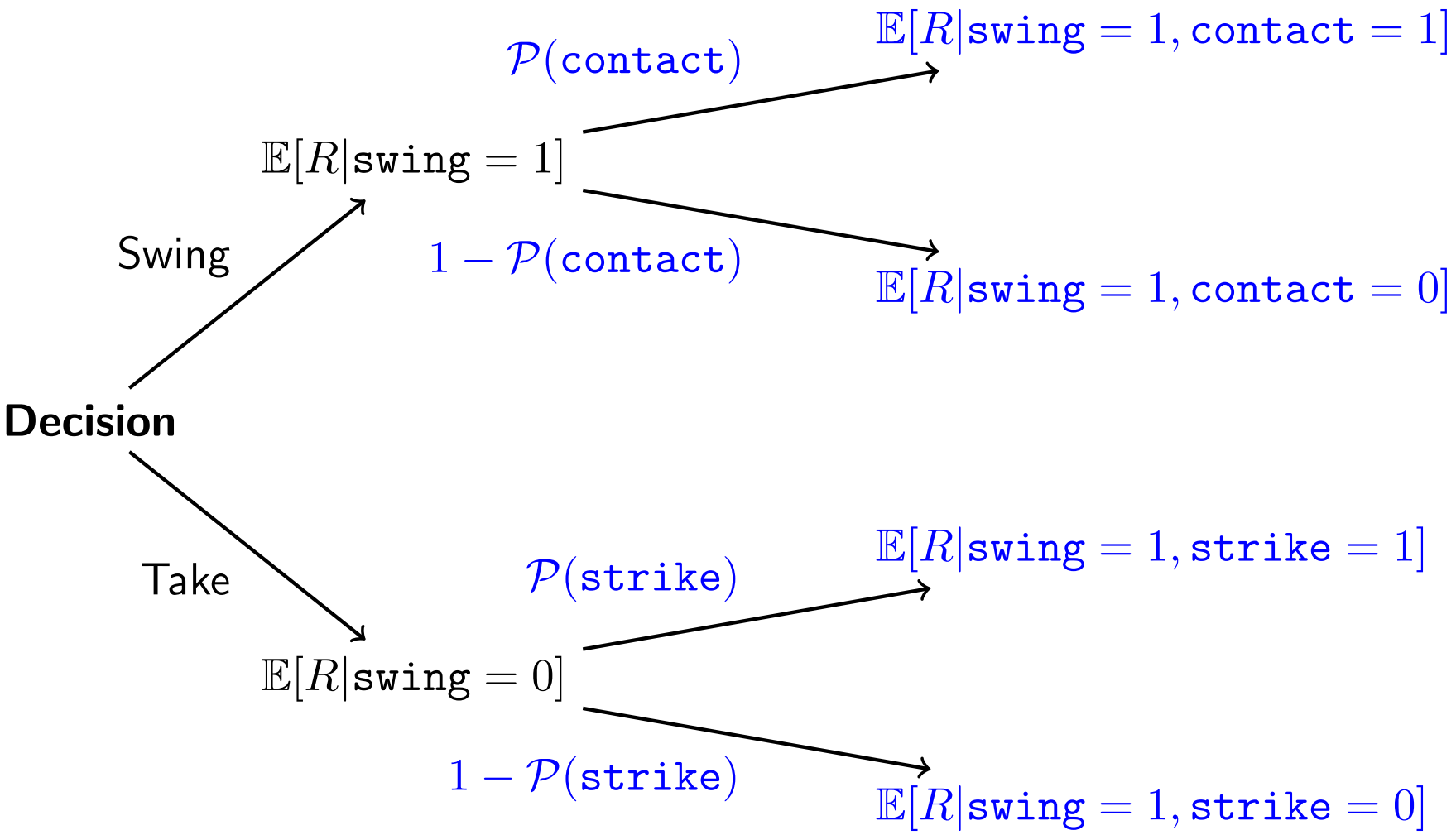
The optimal decision is to swing if $\mathbb{E}[R|\texttt{swing} = 1] > \mathbb{E}[R|\texttt{swing} = 0]$

Framework



The optimal decision is to swing if $\mathbb{E}[R|\texttt{swing} = 1] > \mathbb{E}[R|\texttt{swing} = 0]$

Framework



The optimal decision is to swing if $\mathbb{E}[R|\texttt{swing} = 1] > \mathbb{E}[R|\texttt{swing} = 0]$

Estimation

- Desiderata: estimate each intermediate quantity as functions of their conditioning variables

$$\mathcal{P}(\texttt{strike}) = f(\texttt{game state, personnel, location, ...})$$

  - Nonlinear
  - Complex interactions
  - Functional form difficult to specify
  - Propagate uncertainties from constituent models

Estimation

- Desiderata: estimate each intermediate quantity as functions of their conditioning variables

$$\mathcal{P}(\texttt{strike}) = f(\texttt{game state, personnel, location, ...})$$

  - Nonlinear
  - Complex interactions
  - Functional form difficult to specify
  - Propagate uncertainties from constituent models

- Solution: Bayesian Additive Regression Trees (BART; see Chipman, et. al.)
  - Nonparametric Bayesian model that approximates the target function through an ensemble of piecewise constant step functions
  - Ideal for modeling complex interactions due to tree-based structure
  - Do not need to specify the functional form of the function
  - Generates samples from the posterior distribution of target function

Estimation

- Desiderata: estimate each intermediate quantity as functions of their conditioning variables

$$\mathcal{P}(\texttt{strike}) = f(\texttt{game state, personnel, location, ...})$$

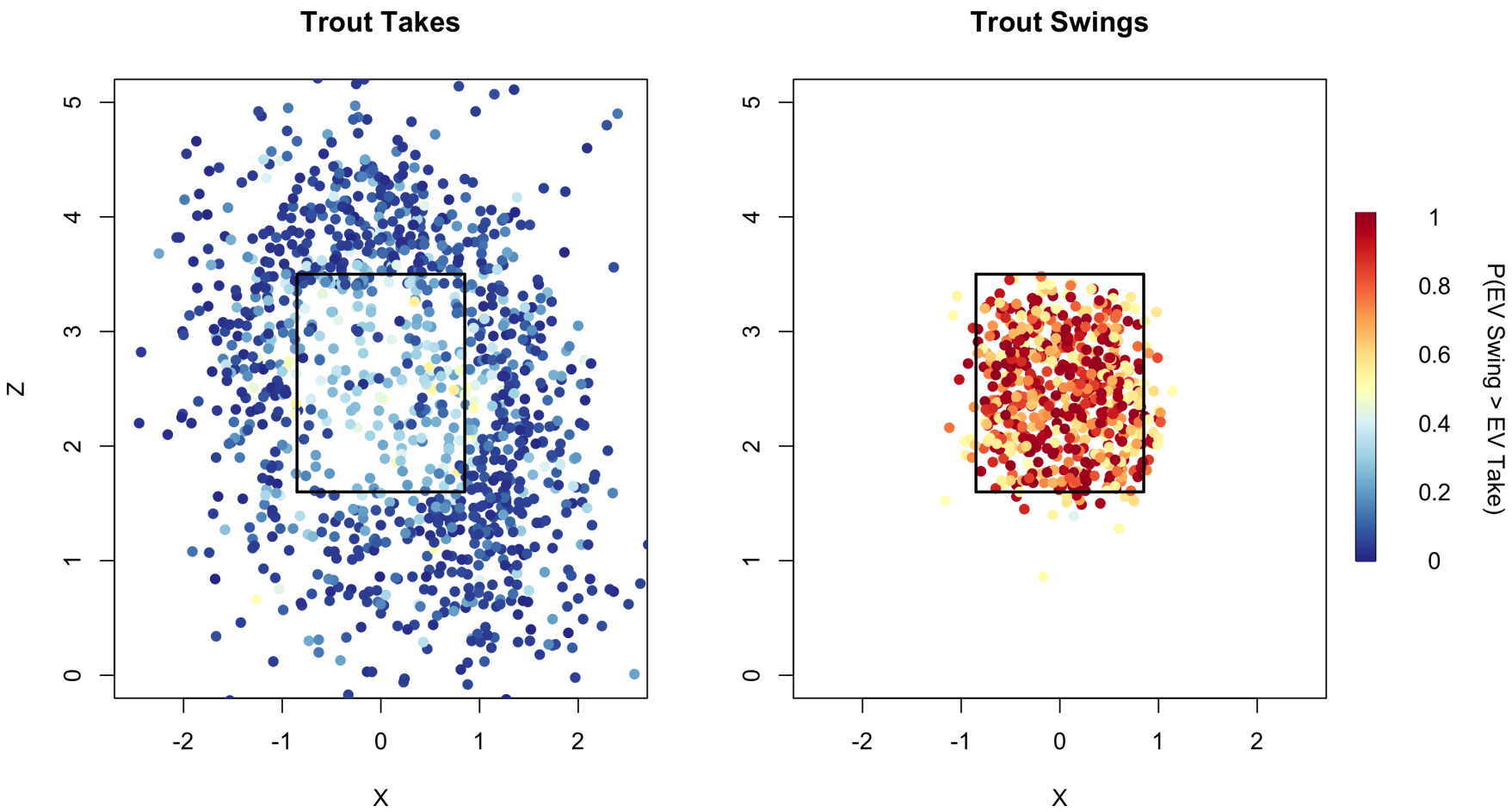
  - Nonlinear
  - Complex interactions
  - Functional form difficult to specify
  - Propagate uncertainties from constituent models

- Solution: Bayesian Additive Regression Trees (BART; see Chipman, et. al.)
  - Nonparametric Bayesian model that approximates the target function through an ensemble of piecewise constant step functions
  - Ideal for modeling complex interactions due to tree-based structure
  - Do not need to specify the functional form of the function
  - Generates samples from the posterior distribution of target function

- Use same approach for modeling $\mathcal{P}(\texttt{contact})$ and $\mathbb{E}[R|\texttt{swing}, \texttt{outcome}]$

Results

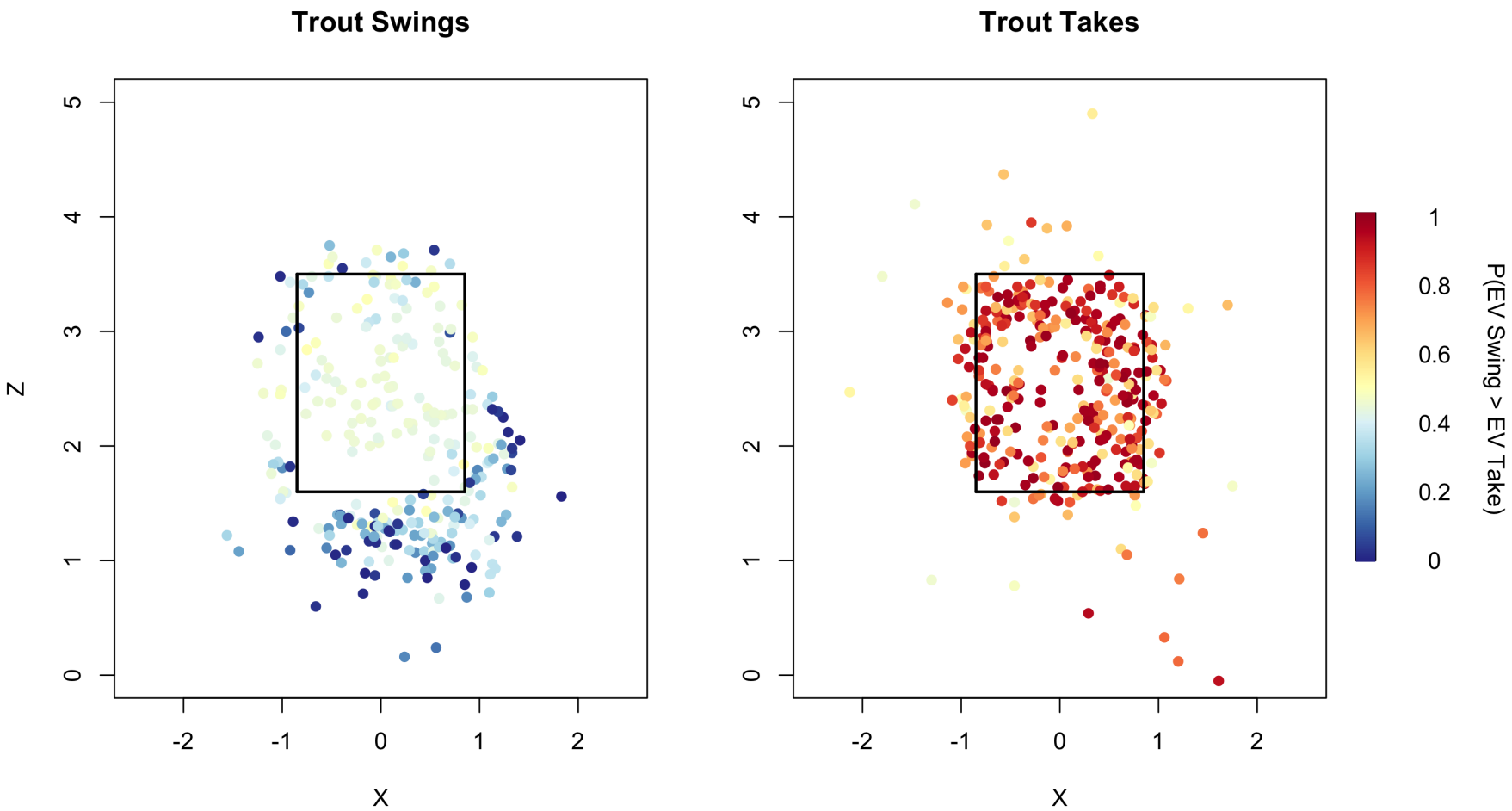
Pitches where Trout made the optimal decision



Note: Pitches shown from umpire's perspective (Trout stands to the left)

Results

Pitches where Trout made the sub-optimal decision



Note: Pitches shown from umpire's perspective (Trout stands to the left)

Takeaways

- Introduced framework for determining the optimal swing decision which utilizes contextual information about the game when the pitch was thrown
- Bayesian approach facilities uncertainty propagation from constituent models to our determination of the optimal swing decision
- Intermediate quantities estimated using Bayesian additive regression trees which outperform traditional approaches

Takeaways

- Introduced framework for determining the optimal swing decision which utilizes contextual information about the game when the pitch was thrown
- Bayesian approach facilities uncertainty propagation from constituent models to our determination of the optimal swing decision
- Intermediate quantities estimated using Bayesian additive regression trees which outperform traditional approaches
- Addition results such as summary metrics can be found in the pre-print

✉ Email: ryan.yee@wisc.edu

☞ Website: https://ryanyee3.github.io

Pre-print

Rshiny app

Takeaways

- Introduced framework for determining the optimal swing decision which utilizes contextual information about the game when the pitch was thrown
- Bayesian approach facilities uncertainty propagation from constituent models to our determination of the optimal swing decision
- Intermediate quantities estimated using Bayesian additive regression trees which outperform traditional approaches
- Addition results such as summary metrics can be found in the pre-print

✉ Email: ryan.yee@wisc.edu

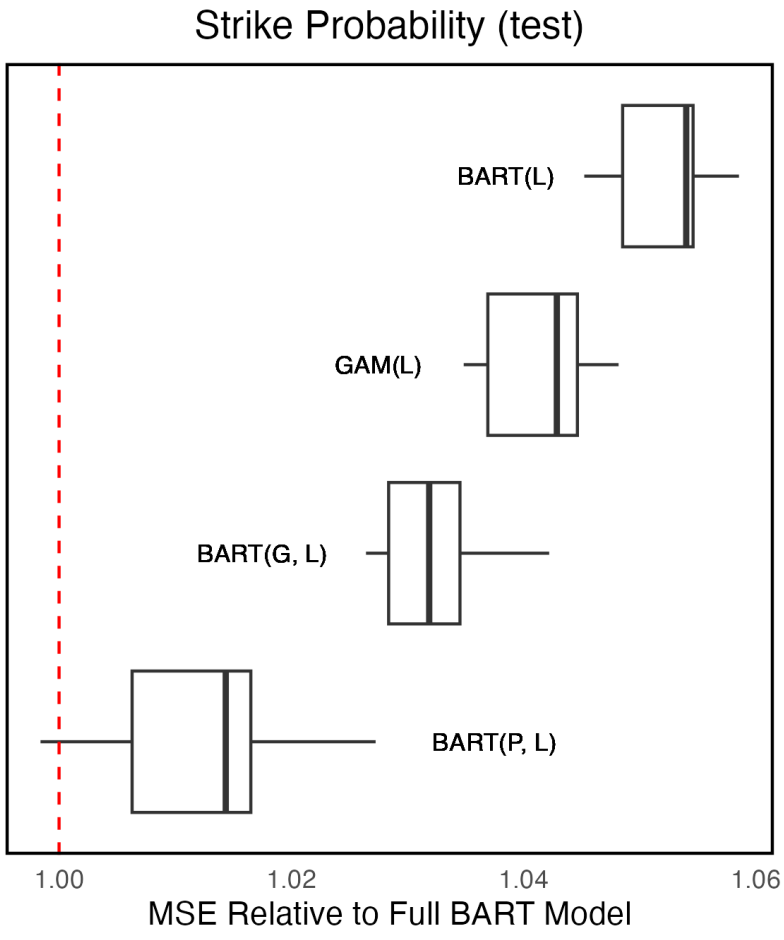
☞ Website: https://ryanyee3.github.io

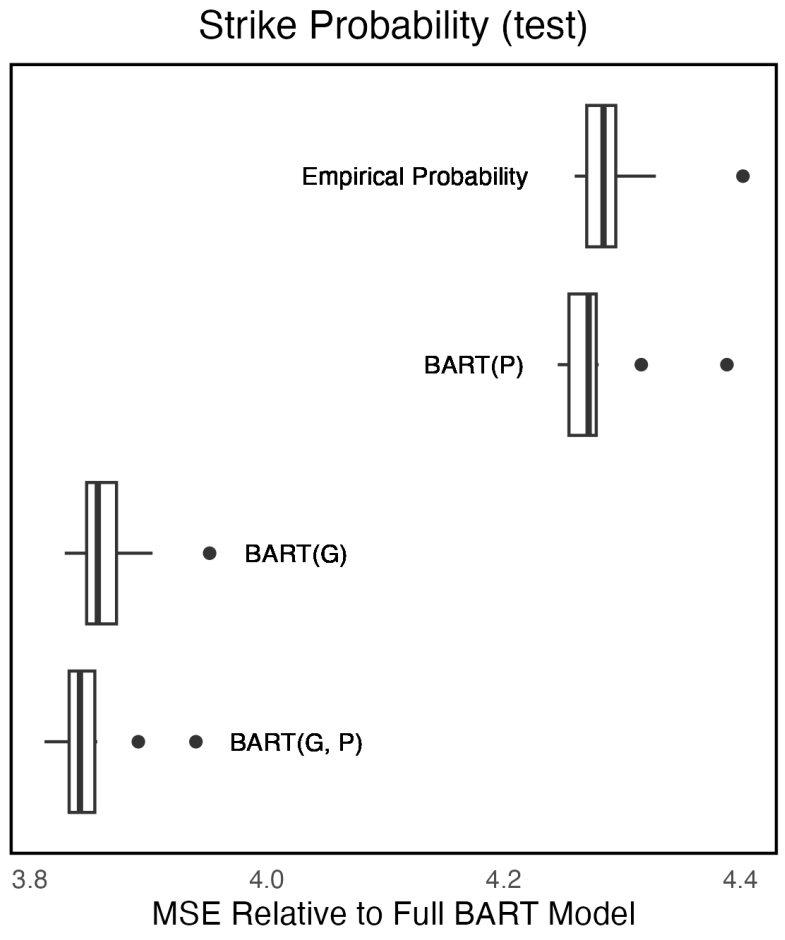
Pre-print

Rshiny app

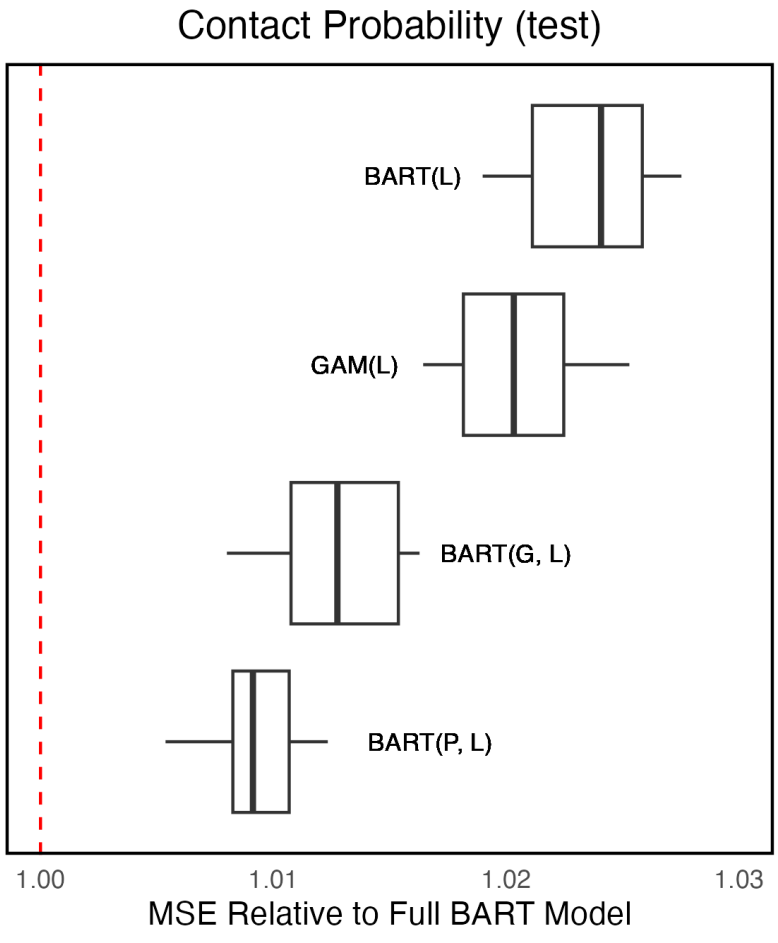
- Thank you!!

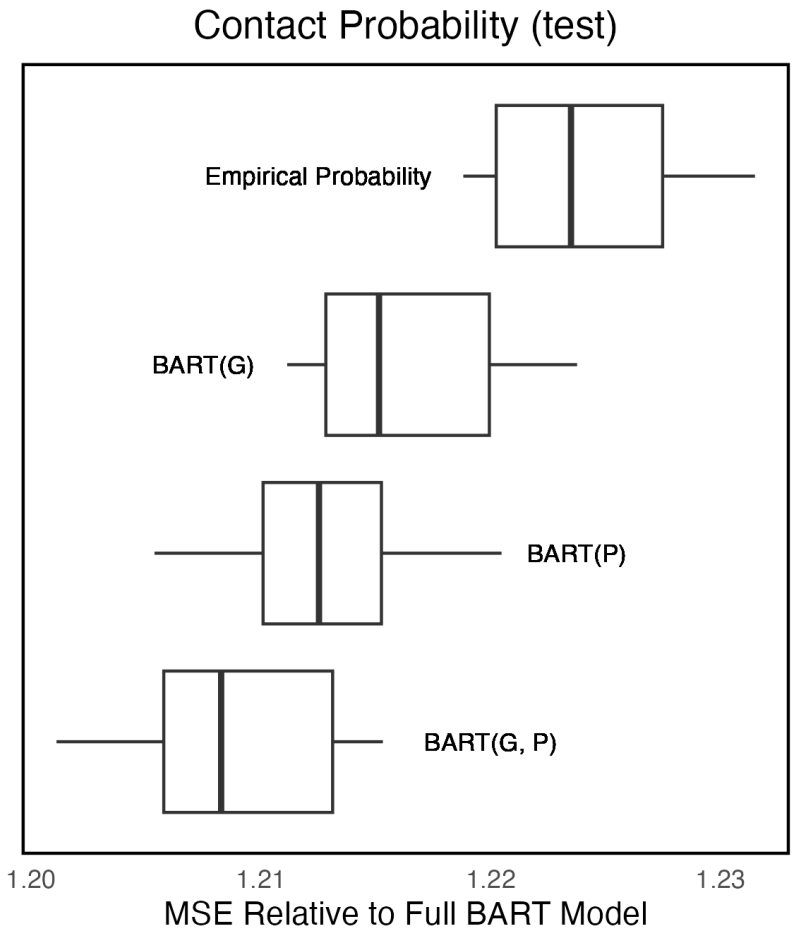
Appendix – Strike Model Validation



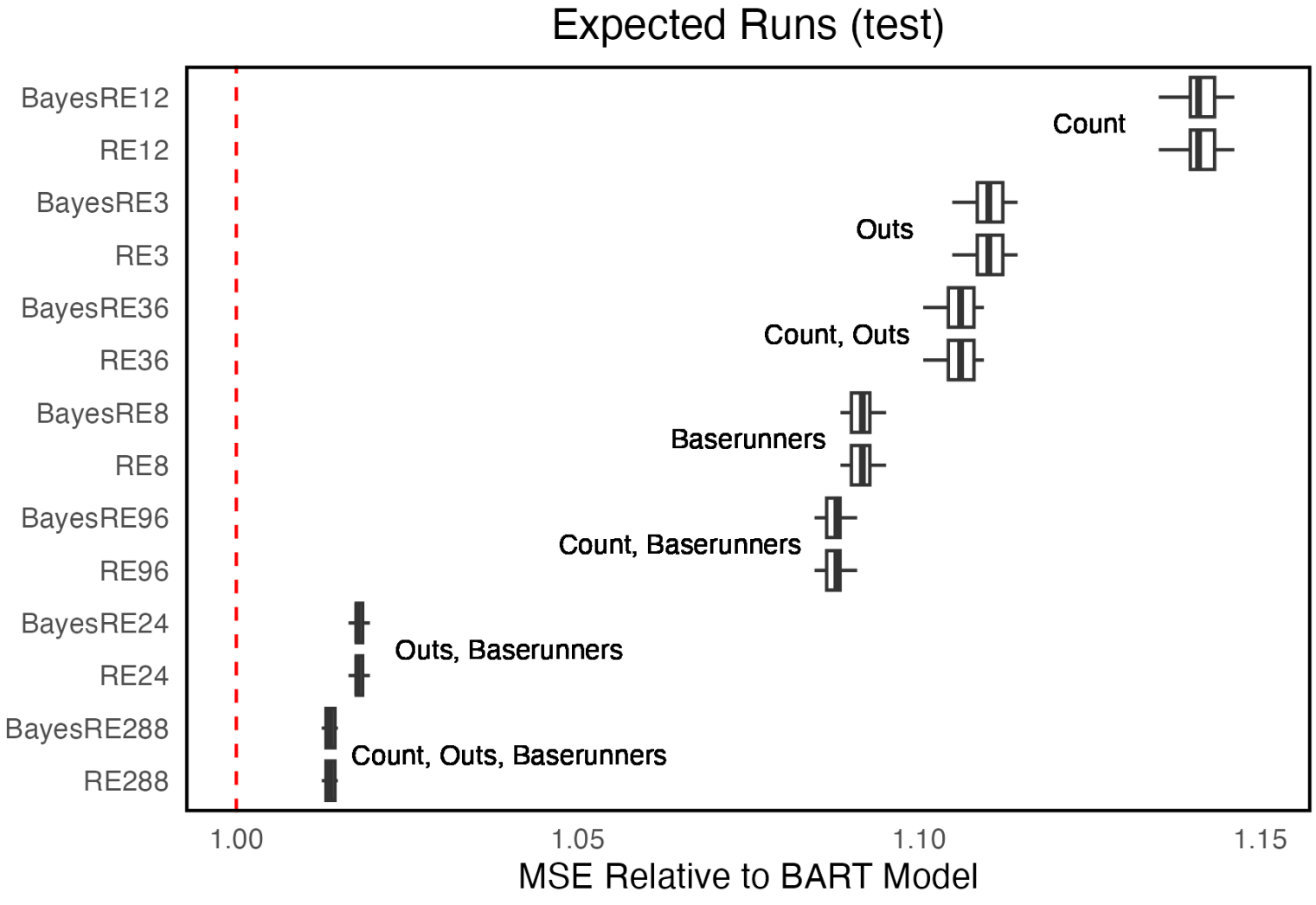
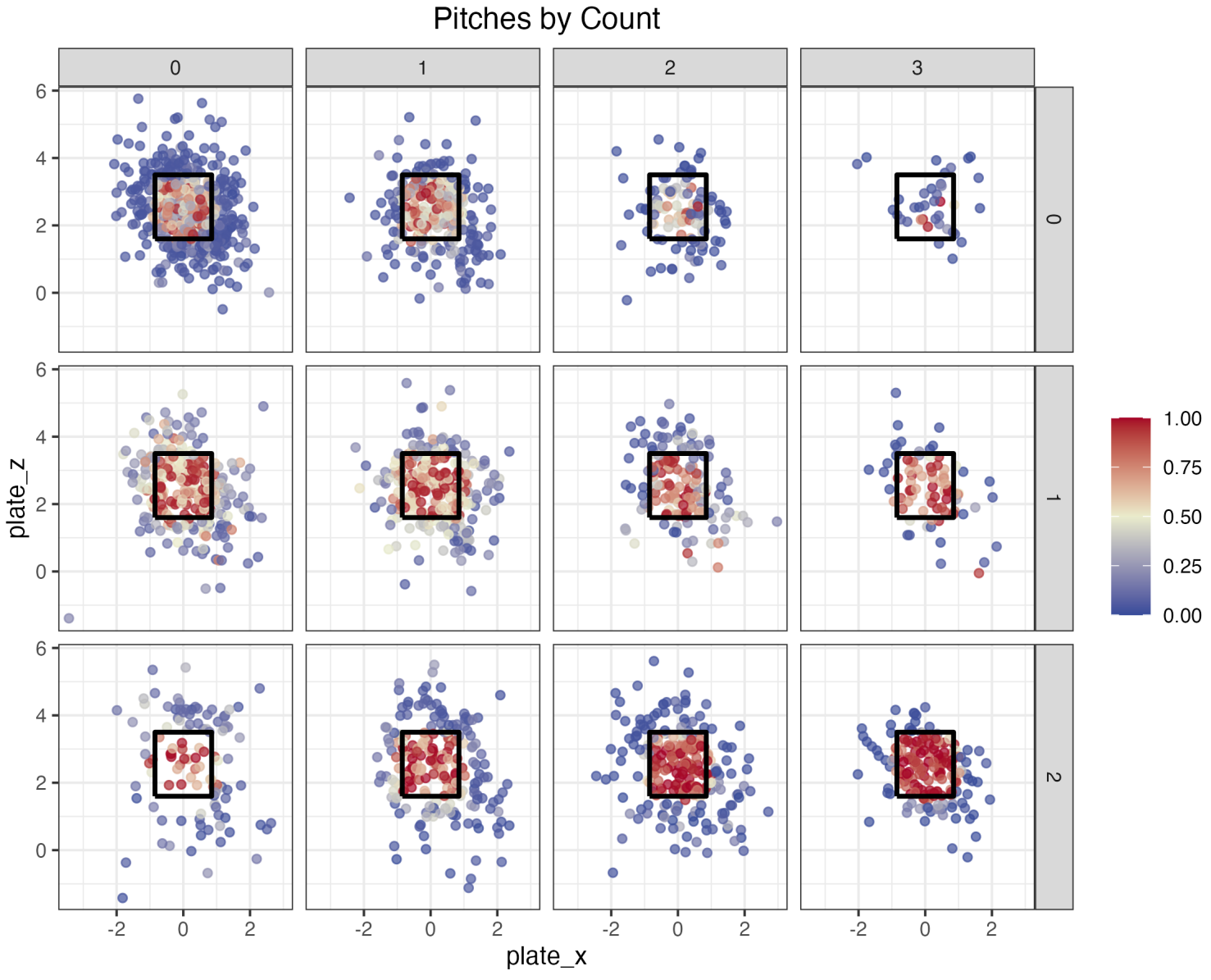


Appendix – Contact Model Validation





Appendix – Expected Runs Model Validation



Appendix – Additional Figures