Before We Start

We will have an opportunity for hands-on tool use!

To prepare:

1. Download RStudio: https://posit.co/downloads/
2. Download Shiny apps: https://github.com/krometis/dataworks2026
3. Install dependencies

AIRC

ACQUISITION INNOVATION
RESEARCH CENTER

Recent Methodological Advances for Integrated T&E

Mini-Tutorial, DATAWorks 2026

Dr. Justin Krometis

April 22, 2026

The views, findings, conclusions, and recommendations expressed in this material are solely those of the author(s) and do not necessarily reflect the views or positions of the United States Government (including the Department of Defense (DoD) and any government personnel)

Outline

1. Background: What and Why
2. Methods
3. Case Study 1: Bayesian Reliability
4. Case Study 2: Changing Factors
5. Case Study 3: Test Planning
6. Tool Development
7. Future Outlook

Background

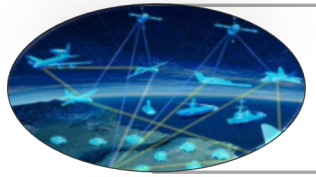
ACQUISITION INNOVATION
RESEARCH CENTER

Research Context

ARTIFICIAL INTELLIGENCE

INTEGRATED TESTING

DIGITAL ENGINEERING



KILL WEBS

Research Partners

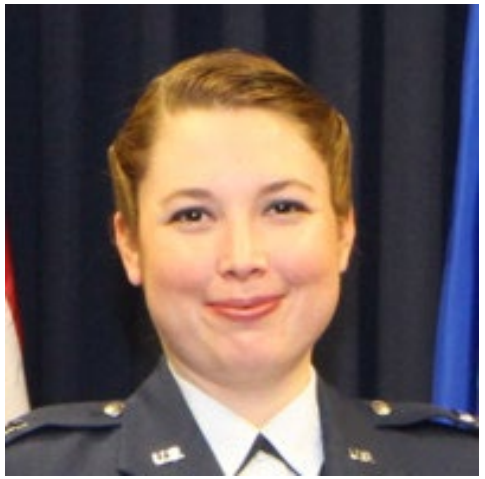
IDA

Scientific Test & Analysis Techniques Established by OSD 2012 Test & Evaluation Center of Excellence

Key Contributors

Dr. Laura Freeman
Virginia Tech

Corinne Stafford
STAT COE

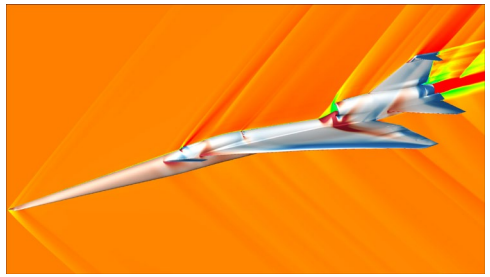


Dr. Victoria Sieck
U.S. Air Force

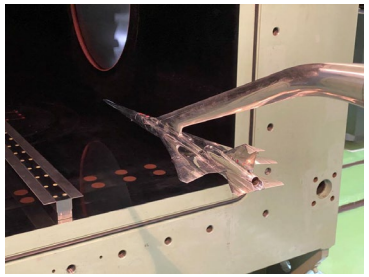
Integrated Testing

Using iterative sequential-progressive design of experiments coupled with evaluations that combine information from across the acquisition lifecycle.

Historical Data
SME Knowledge



Model Data



DT Data

OT Data

Operational
Performance

Leveraging all data enables better understanding of systems *earlier*…

…allowing for fewer or more optimal tests *later*

ALTERNATIVE T&E PATHWAY (2026 NDAA)

Alternative T&E Pathway (805(b)):

"(1)(A) design and execute a unified T&E strategy that **aligns DT and OT** to … **build system understanding throughout the test program**"

"(1)(B)(v) automated analytics tools to **assess performance trends**, reliability, and maintenance needs"

"(1)(C)(iv) **integration of supporting or complementary data** from digital twins or other model-based systems engineering tools"

"(1)(D) define general test and evaluation objectives and data needs while **allowing detailed execution plans to evolve based on test results** and emerging requirements"

Role of DOT&E (805(c)):

"(1) provide independent evaluation of test data **across all phases of the program lifecycle**"

"(1)(B) evaluating whether the program collects and analyzes sufficient raw data, learns from test results at a **pace relevant to operational needs**, and converges on military effectiveness **based on data trends**"

"(1)(D) providing **continuous oversight through ongoing analysis** of test data"

Acquisition Transformation Strategy

Maximize Acquisition Flexibility:

"It is critical that teams leverage existing authoritative **data sources, including contractor data** and automated reporting mechanisms to assess program performance."

"scaling tools, policies, and practices to **maximize the use of modeling and simulation and automated testing** will allow the Department to accelerate and continually validate software."

“It is important that the Department **invest in test and evaluation (T&E) resources for validating digital technologies, including modeling and simulation, synthetic environments**, and managing performance/risks.”

"the Department can **prioritize integrated testing to optimize test resources, avoid redundancies**, and utilize data-driven decisions to trace programs, performance, and testing"

Develop High-Performing Systems:

"The Department will integrate and scale adoption and investments in ...modeling and simulation environments for **virtual and constructive testing and training, and developmental integration and testing**, which will fuel rapid, iterative designs and technology insertion to maximize mission outcomes. These advanced techniques will **reduce the burden in test planning and execution** and the unplanned testing that results from unanticipated discovery"

"Integrate and **test new technology soonest to get key insights early** to shape weapon system development and fielding"

DoDI 5000.98

3.1(a):

"OT&E and LFT&E planning, execution, analysis, and reporting activities will use **the latest advances in science (e.g., design of experiments, statistical inference methods, or big data analytics)** to … determine, with scientific rigor, the preliminary and final operational effectiveness, suitability, survivability, and lethality (as applicable) of DoD systems."

3.1(c):

"Science and technology-based OT&E and LFT&E will **enable efficient use of data from multiple data sources** (e.g., contractor test (CT), developmental test (DT), operational test (OT), and live fire test (LFT) data or M&S results). Improved **sequential testing using Bayesian or similar inference methods** … are critical to dynamically optimize the planning, execution, analysis, and reporting of integrated T&E, OT&E, and LFT&E across the acquisition life cycle."

Considerations for Design & Evaluation of Tests

Consideration	Possible Levels (best case → worst case)
Safety	Minimal risk (not live projectile) → High risk (live fire)
Cost	$ → $$$
Resource Availability	Available → Partially available → Needs to be developed
Schedule	Easy & quick → Hard & extensive coordination
Historical operational performance data	Yes same factors → Yes but missing key factor(s) → None
Modeling and Simulation	Yes accurate and validated over time → Yes but not well understood and/or missing key factor(s) → None
Scale	Single component → Parallel systems → Series system

Considerations for Design & Evaluation of Tests

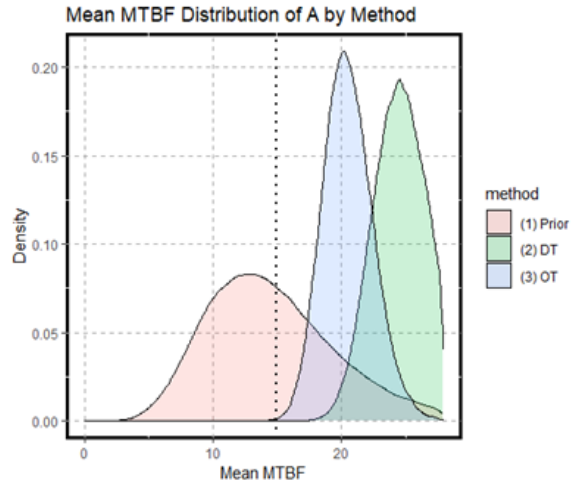
Consideration	Possible Levels (best case → worst case)
Safety	Minimal risk (not live projectile) → High risk (live fire)
Cost	$ → $$$
Resource Availability	Available → Partially available → Needs to be developed
Schedule	Easy & quick → Hard & extensive coordination
Historical operational performance data	Yes same factors →
Modeling and Simulation	Yes accurate and val and/or missing key f
Scale	Single component →

Variety of DoD programs means that a variety of analysis approaches may be appropriate!

Comprehensive Evaluations

Estimating operational performance from data…

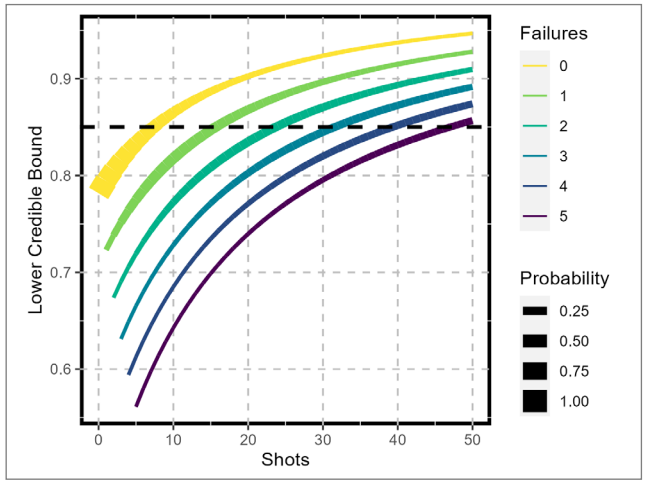
- …from multiple phases of test
- …of different fidelities/sizes
- …with different experimental factors
- …from digital representations of systems



Test Planning/Design

Given previous test results…

- …where and how many tests should we run?
- …when have we tested enough?



Methods

Frequentist Methods

Fit a statistical model to all of the data

Naïve: Assume all data is equivalent and fit to all data equally

Blocking: Try to account for differences in data sources by adding source or phase-specific factors to the model

- Example: Add a shift parameter to account for possible biases in data sources

Bayesian Inference

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta)\, P(\theta)$$

Bayesian Inference

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta)\boxed{P(\theta)}$$

Prior

Bayesian Inference

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

$$P(\theta|Y) \propto P(Y|\theta)\,P(\theta)$$

Likelihood (Data) **Prior**

Bayesian Inference

Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

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

Posterior **Likelihood (Data)** **Prior**

Bayesian Inference

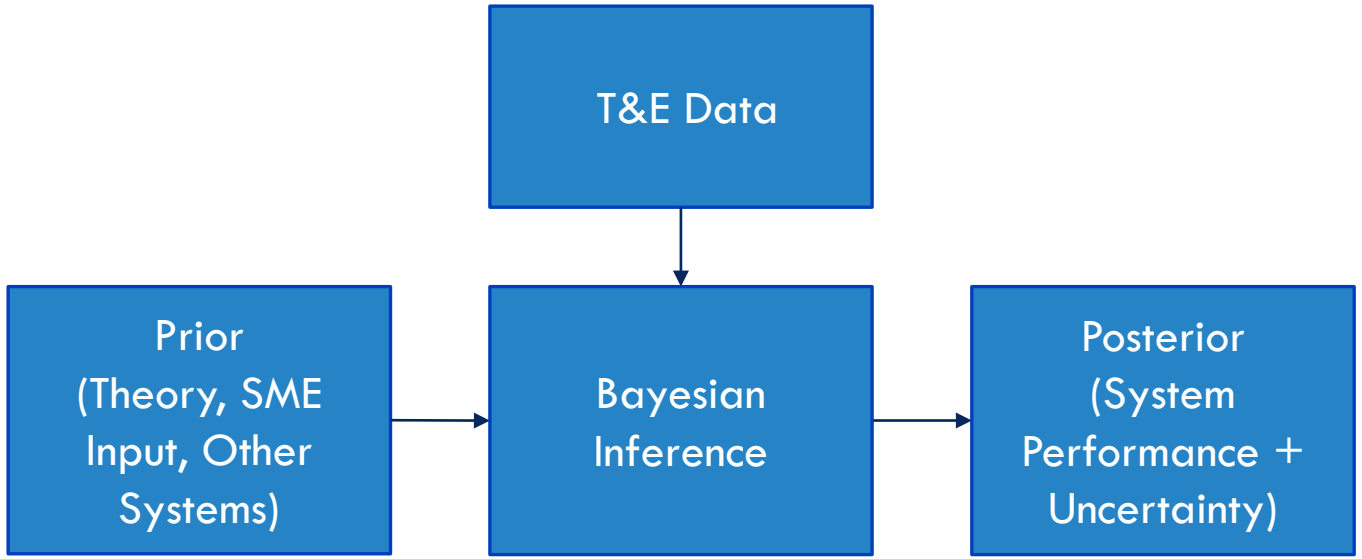
Relates the probability of a parameter value θ given data Y ($P(\theta|Y)$) to the probability of Y given θ and the probability of θ:

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

Posterior **Likelihood (Data)** **Prior**

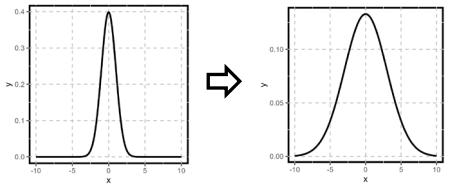
Models human learning: Understanding (prior) + Experience (likelihood) = Updated understanding (posterior)

Single Test Phase



Single Test Phase

T&E Data



Prior (Theory, SME Input, Other Systems)

Bayesian Inference

Posterior (System Performance + Uncertainty)

Downweight/Adjust as Necessary

Prior (Based on Previous Test Phase)

Informative Priors

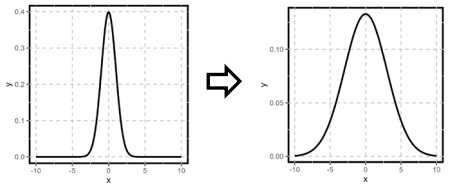
Single Test Phase

T&E Data

Prior (Theory, SME Input, Other Systems)

Bayesian Inference

Posterior (System Performance + Uncertainty)



Downweight/Adjust as Necessary

Next Test Phase

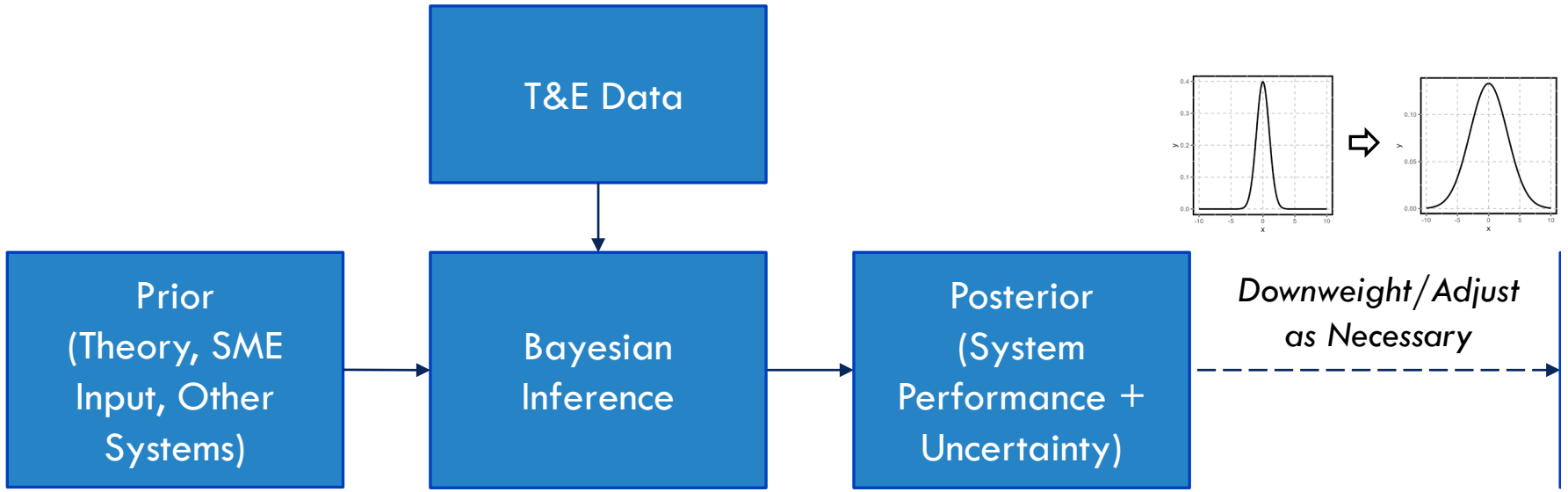
Prior (Based on Previous Test Phase)

T&

Ba
Inf

INFORMATIVE PRIORS

SINGLE TEST PHASE

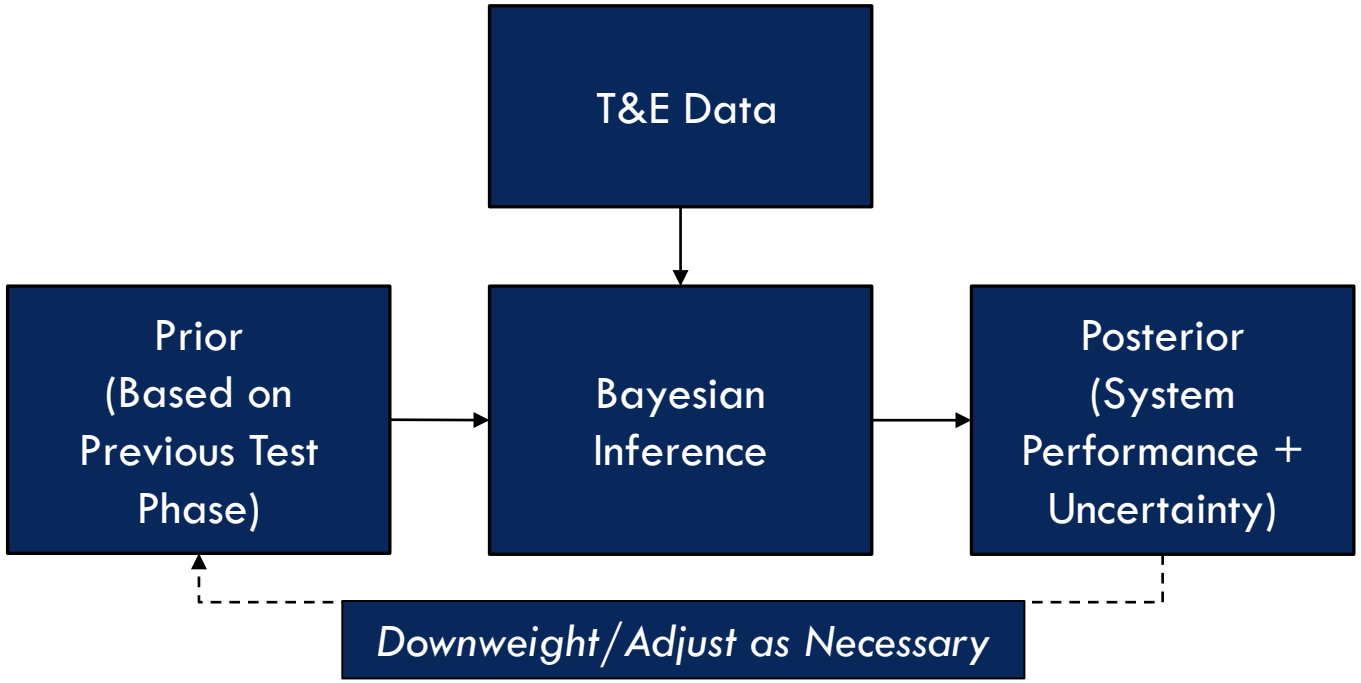


NEXT TEST PHASE

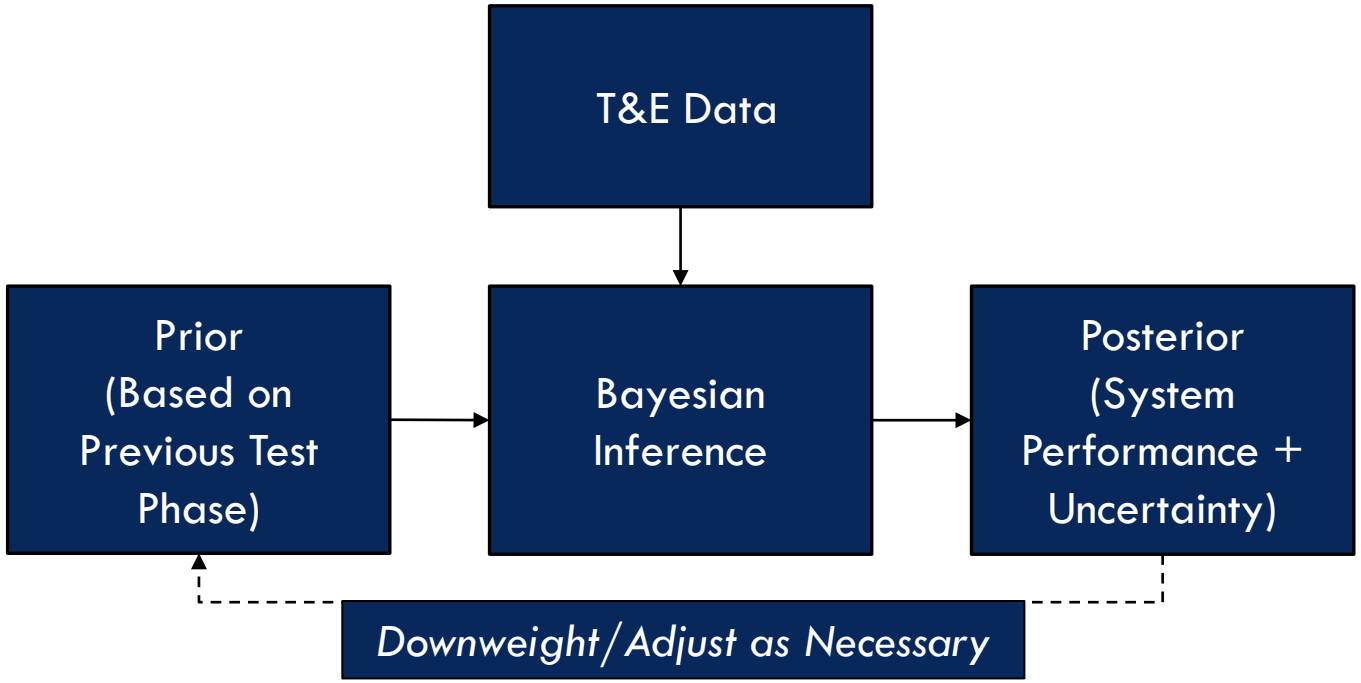
T&

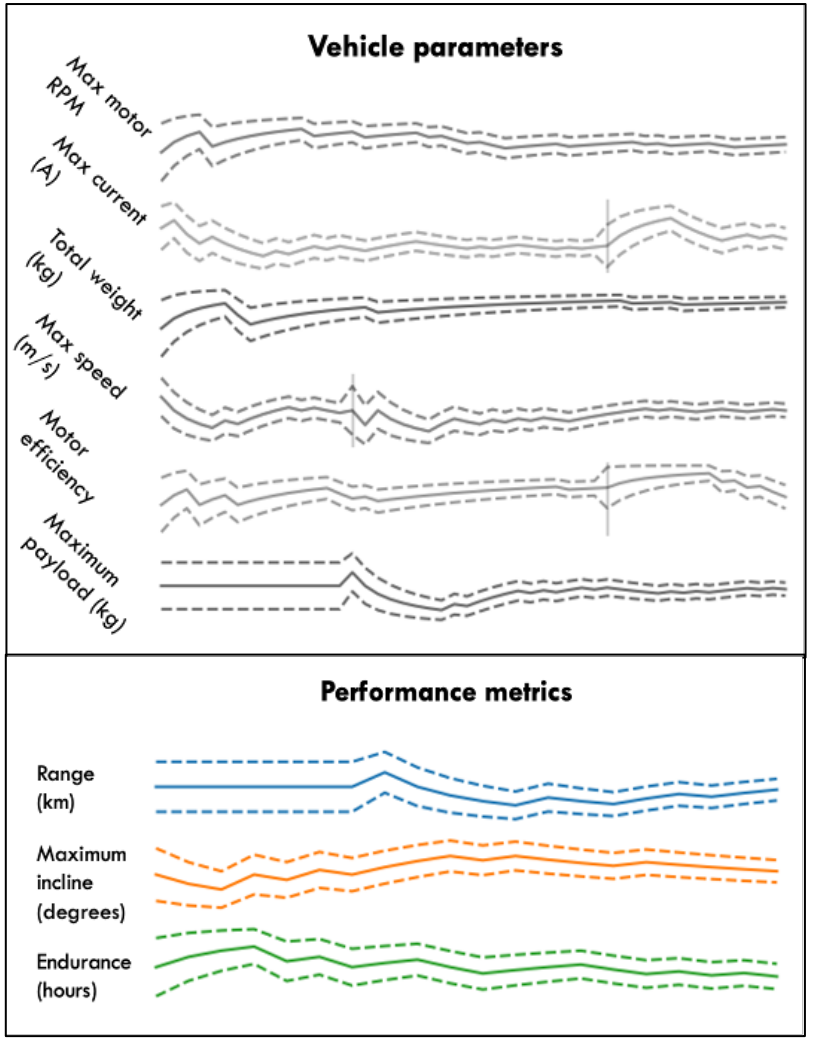
DATAWorks 2024 Tutorial on Bayesian Methods

Multiple, Integrated Test Phases



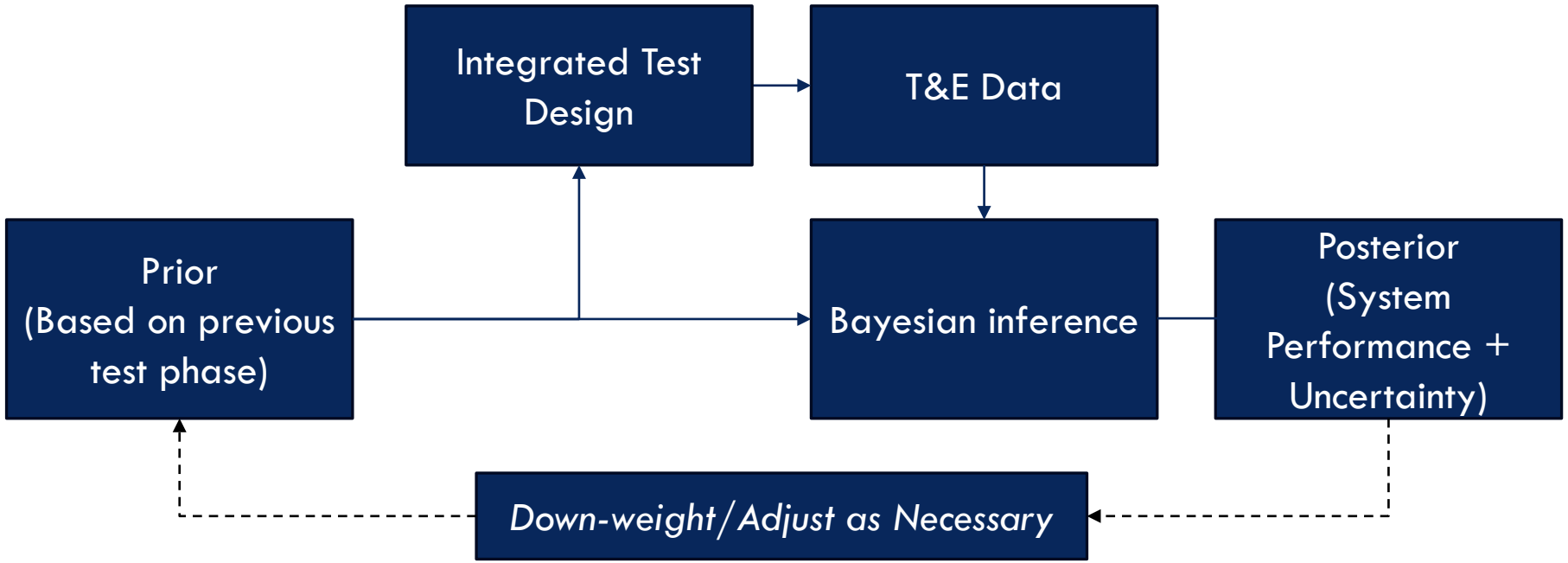
Multiple, Integrated Test Phases





(Data and metrics are notional)

Multiple, Integrated Test Phases



Case Study 1: Bayesian Reliability

ACQUISITION INNOVATION
RESEARCH CENTER

Stryker Family of Vehicles

Publicly-released reliability dataset

- Metric: Miles before system abort (MBSA)
- Both DT and OT
- Eight variants

See:

- Dickinson et al. "Statistical methods for combining information: Stryker family of vehicles reliability case study." 2015.
- Sieck et al. "A Framework for Using Priors in a Continuum of Testing." 2024.
- Krometis et al. "A Comparison of Bayesian Methods for Integrated Test and Evaluation." 2025.

Stryker Family of Vehicles

Publicly-released reliability dataset

- Metric: Miles before system abort (MBSA)
- Both DT and OT
- Eight variants

See:

- Dickinson et al. "Statistical methods for combining information: Stryker family of vehicles reliability case study." 2015.
- Sieck et al. "A Framework for Using Priors in a Continuum of Testing." 2024.
- Krometis et al. "A Comparison of Bayesian Methods for Integrated Test and Evaluation." 2025.

Informative Priors

Reliability Model

Model reliability with Weibull distribution:

$$y_i|k, \tau_i \sim \text{Weibull}(k, \tau_i)$$

where:

- k is the *shape* parameter (failure change over time)
- τ_i is the *scale* parameter (units)

We infer k and $\tau_1, \ldots, \tau_8$ from the data and then can compute the mean MBSA (MMBSA):

$$\bar{y}_i = \tau_i * \Gamma(1 + 1/k)$$

Priors

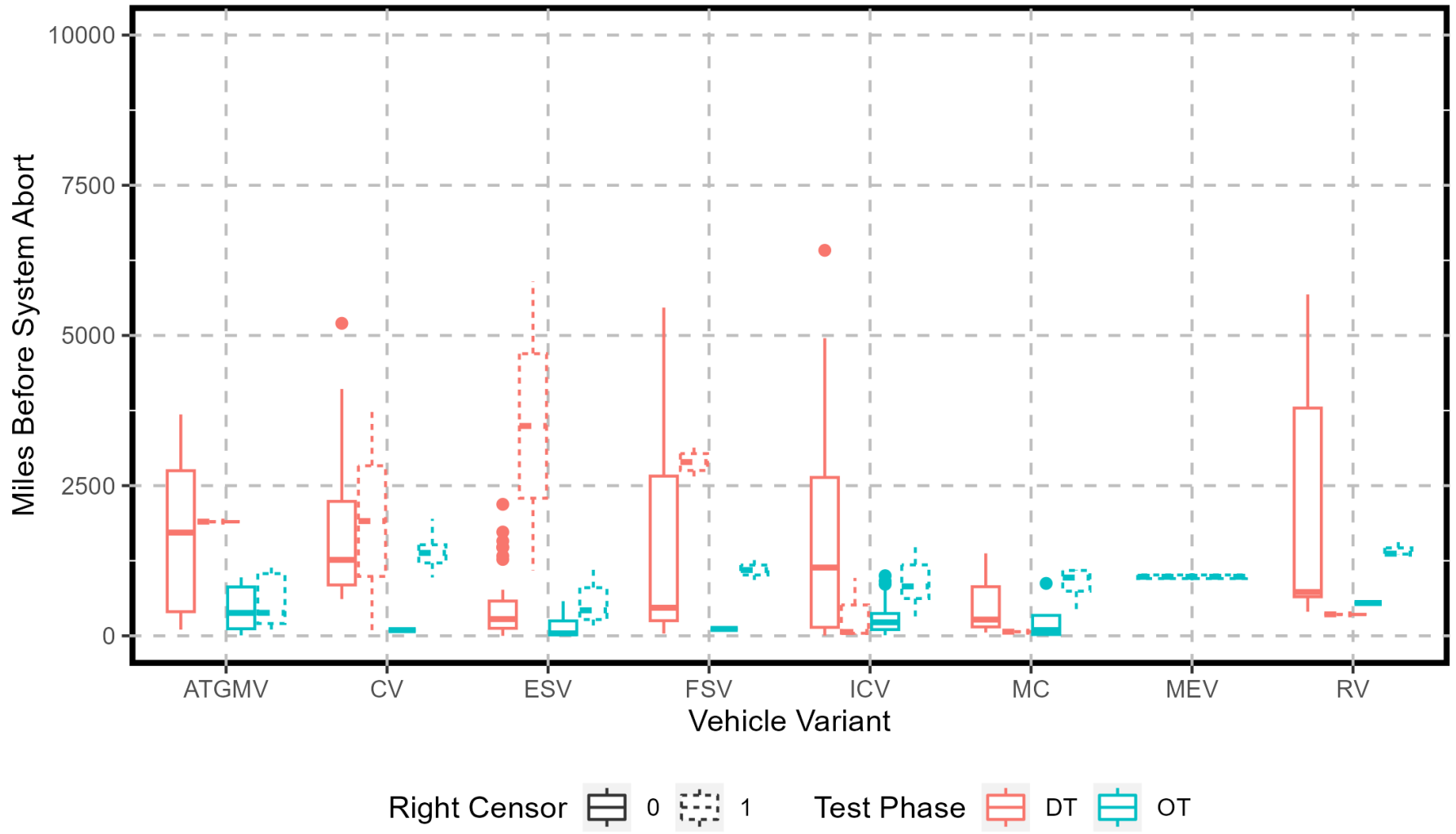
Uninformative – impart limited information via very vague priors:

- Shape parameter k: $\text{Gamma}(0.001, 0.001)$
- Scale parameter τ: $\text{Gamma}(10^{-6}, 10^{-6})$

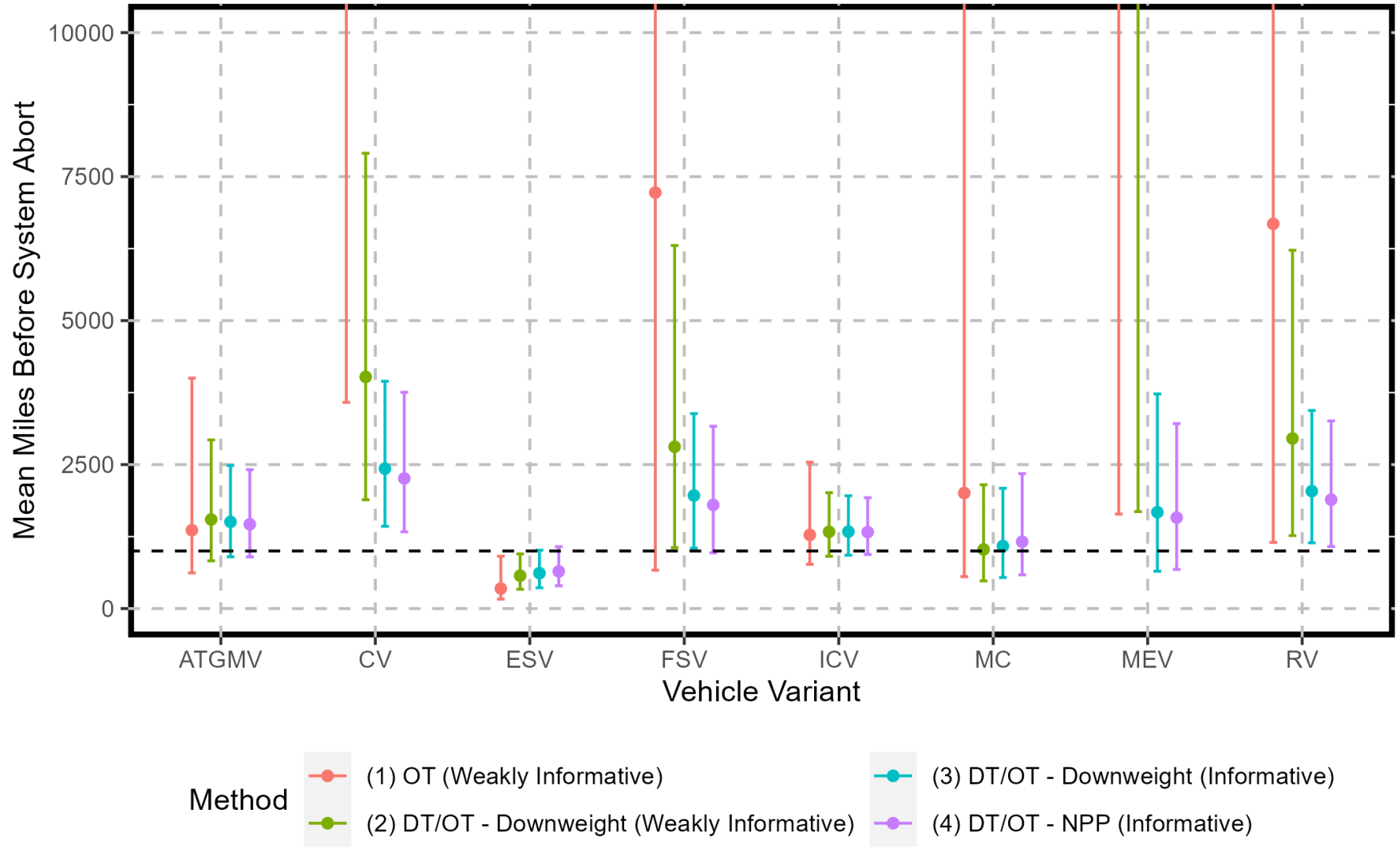
Informative – impart basic information:

- Shape parameter k: $\text{Gamma}(4,4)$ (mean 1, std. dev. 0.5)
- Scale parameter τ: $\text{Gamma}(4.6225, 0.0043)$ (median 1,000, std. dev. 500)

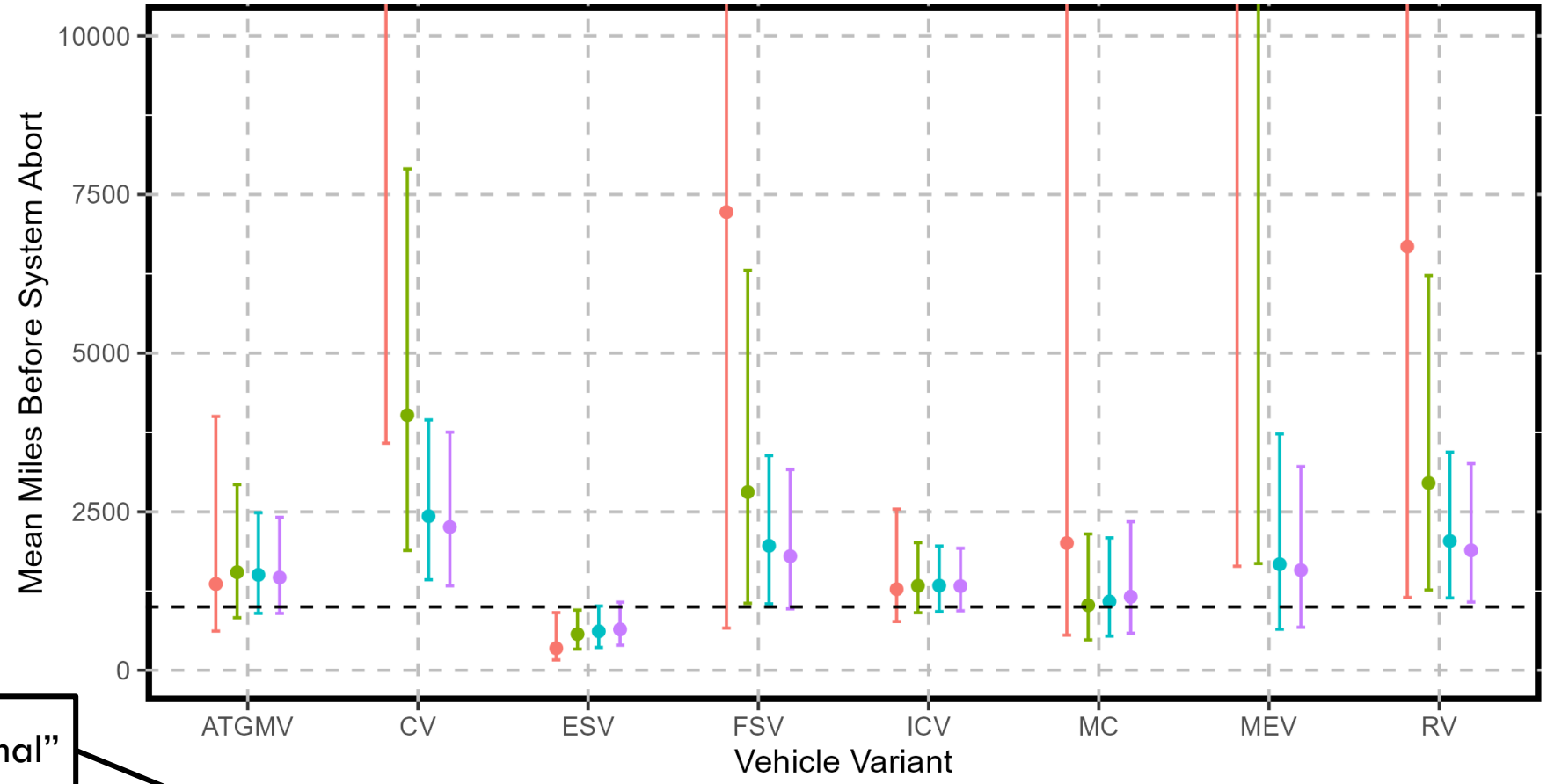
STRYKER: DATA



STRYKER: METHODS COMPARISON



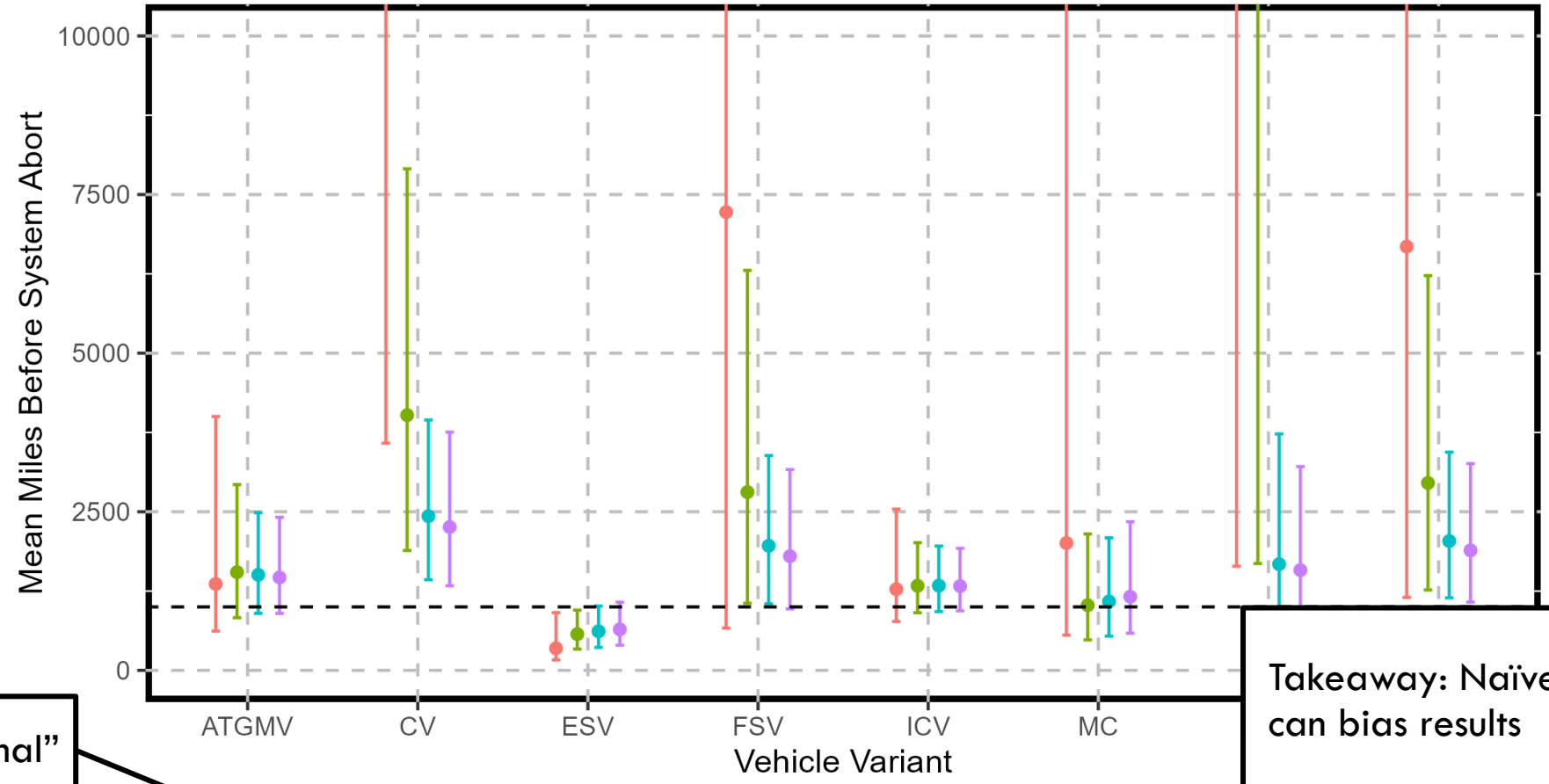
STRYKER: METHODS COMPARISON



Analog of "traditional"

"Naïve" Bayes

STRYKER: METHODS COMPARISON

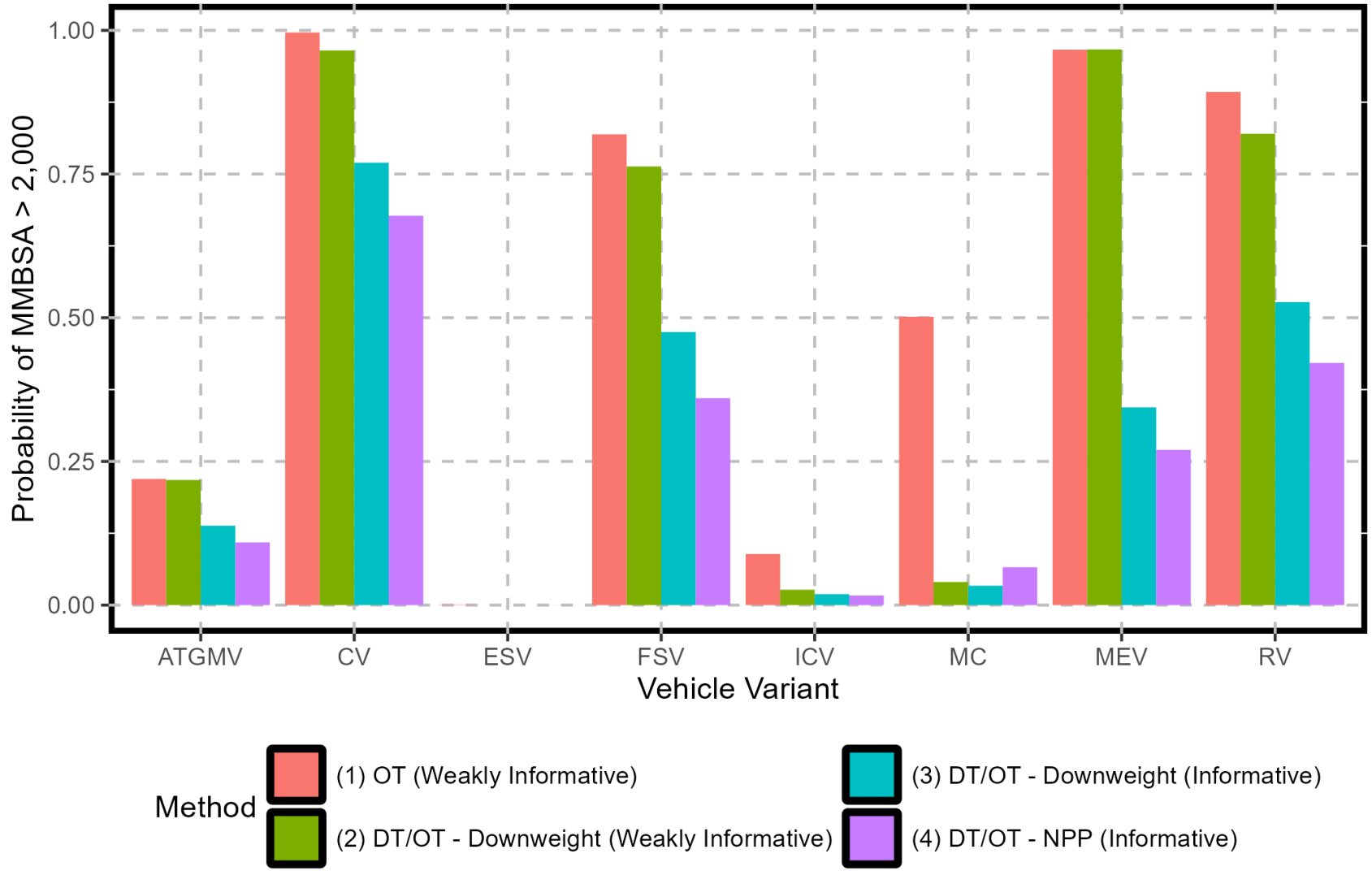


Analog of "traditional"

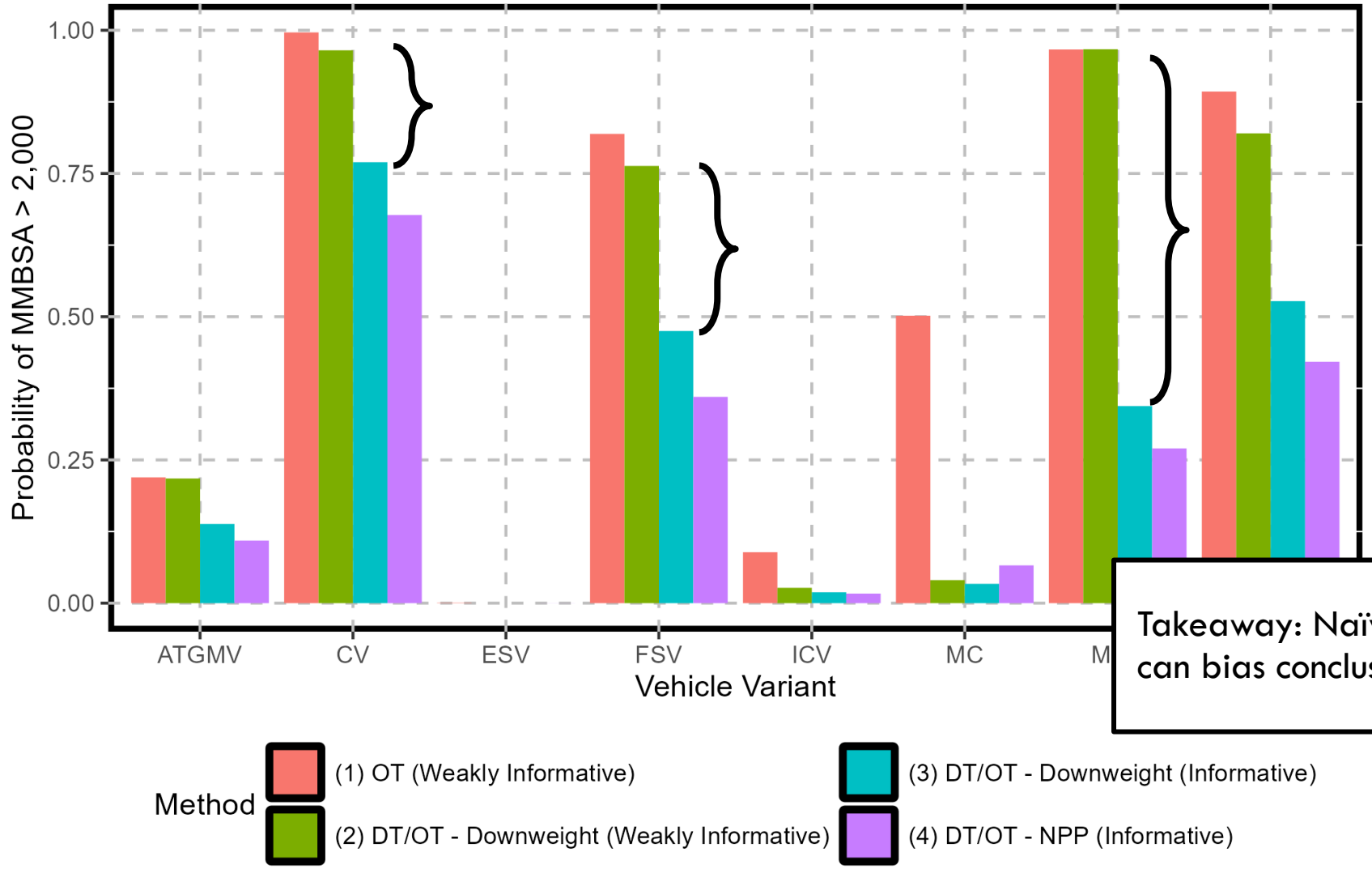
"Naïve" Bayes

Takeaway: Naïve approaches can bias results

STRYKER: METHODS COMPARISON



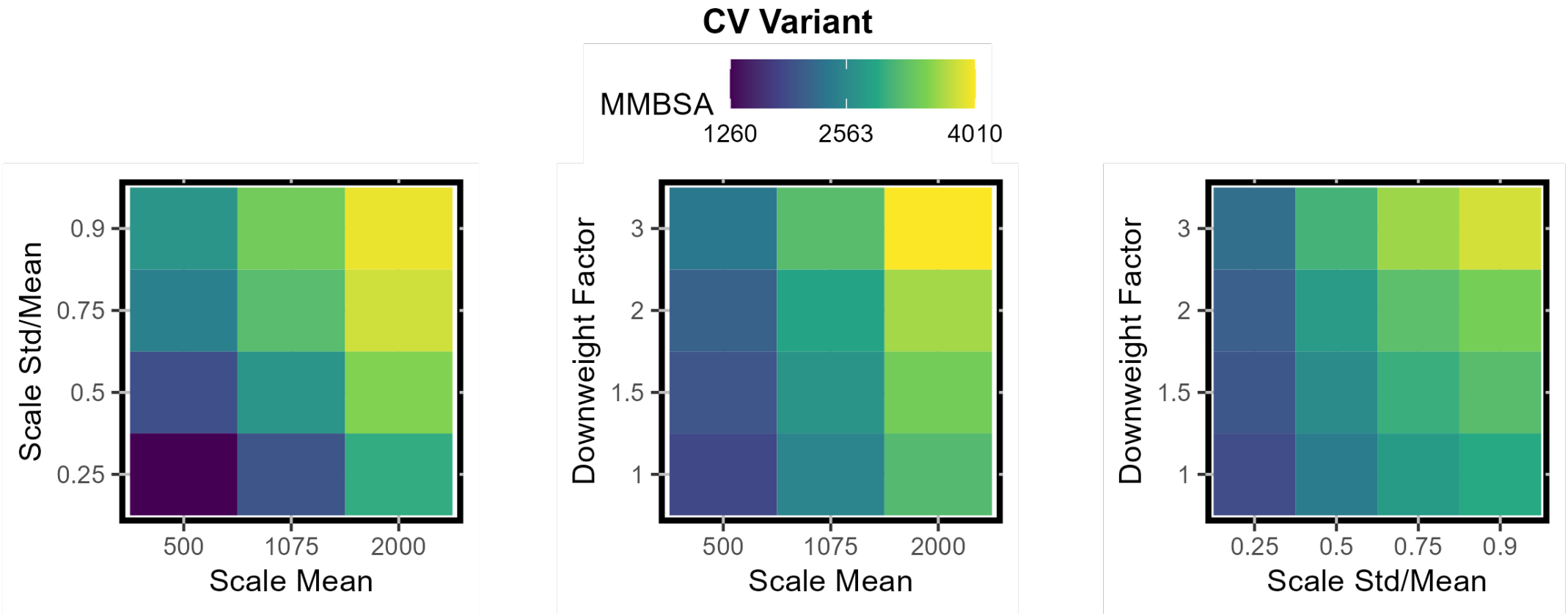
STRYKER: METHODS COMPARISON



Takeaway: Naïve approaches can bias conclusions

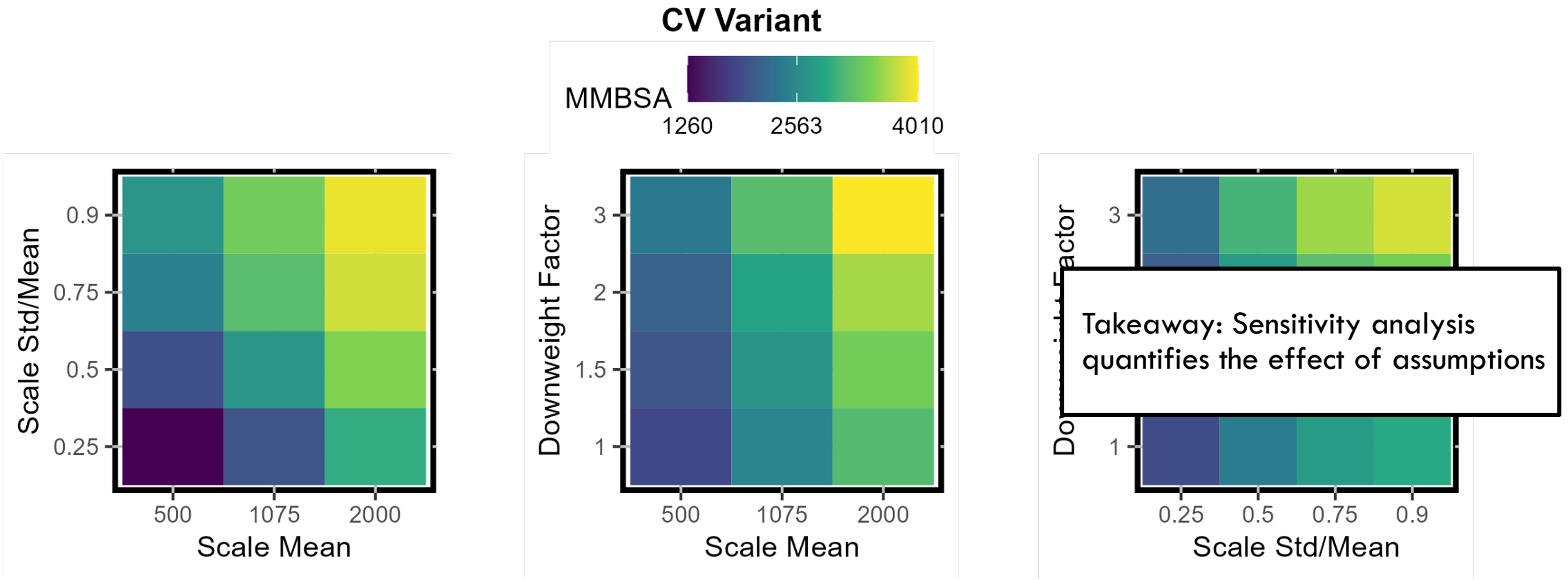
STRYKER: SENSITIVITY ANALYSIS

Key concern among Bayesian skeptics: How do modeling assumptions (including the prior) affect estimates?



Stryker: Sensitivity Analysis

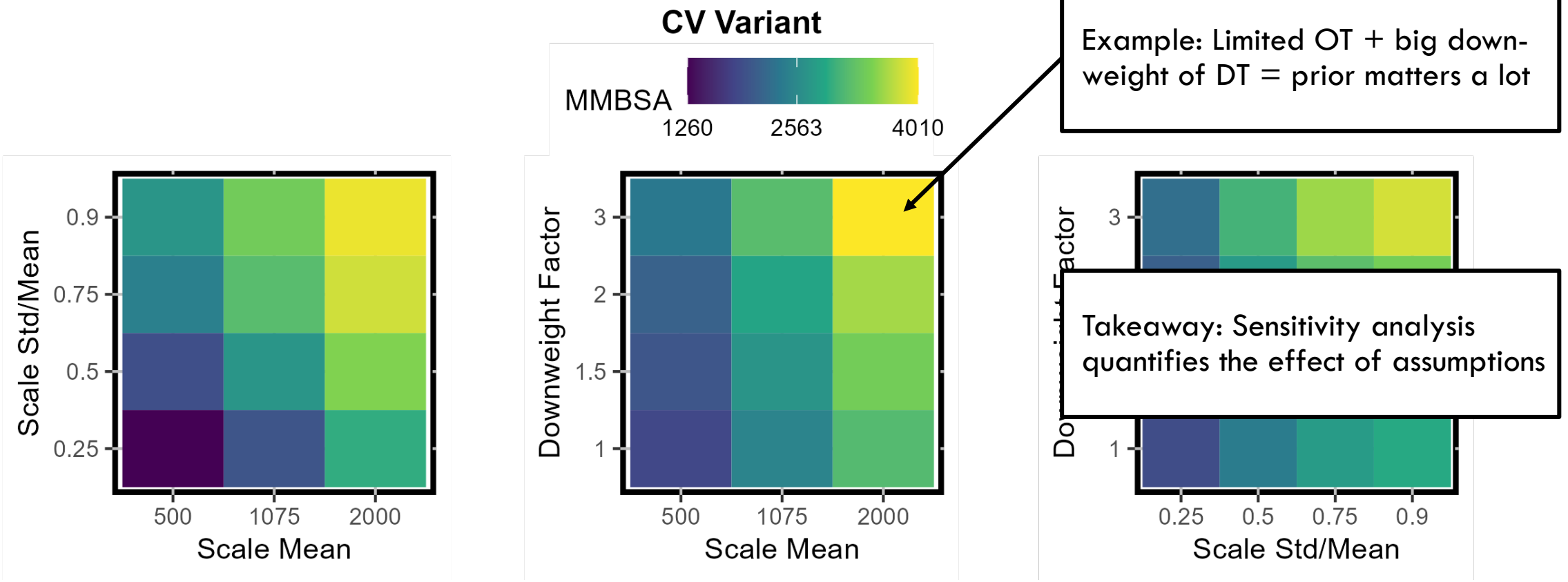
Key concern among Bayesian skeptics: How do modeling assumptions (including the prior) affect estimates?



Takeaway: Sensitivity analysis quantifies the effect of assumptions

Stryker: Sensitivity Analysis

Key concern among Bayesian skeptics: How do modeling assumptions (including the prior) affect estimates?

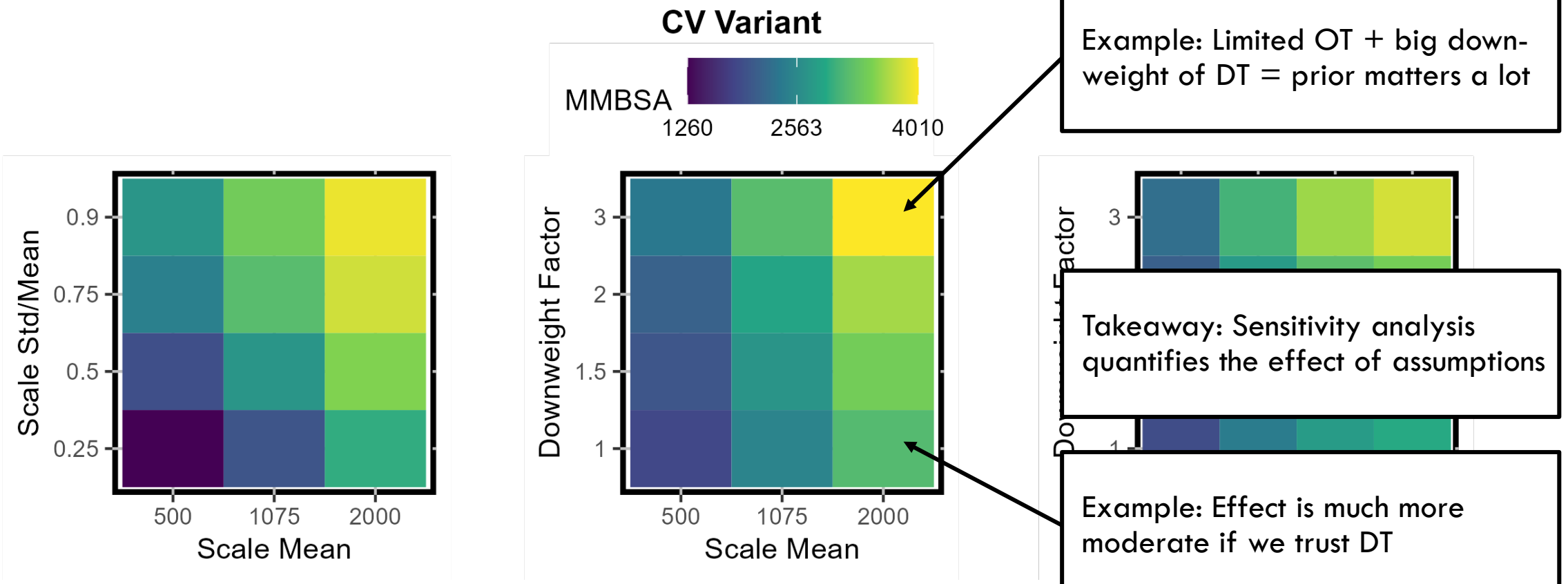


Example: Limited OT + big down-weight of DT = prior matters a lot

Takeaway: Sensitivity analysis quantifies the effect of assumptions

Stryker: Sensitivity Analysis

Key concern among Bayesian skeptics: How do modeling assumptions (including the prior) affect estimates?



Example: Limited OT + big down-weight of DT = prior matters a lot

Takeaway: Sensitivity analysis quantifies the effect of assumptions

Example: Effect is much more moderate if we trust DT

STRYKER: LESSONS LEARNED

Demonstrated the promise of Bayesian methods for integrating test data to improve estimates of operational performance

Highlighted that care should be taken in the development of assumptions, as attempts at objectivity can inadvertently bias results

Case Study 2: Changing Factors

Background: Synthetic Radar Models

IDA created[1] the following model of a synthetic counterfire radar:

$$Y = 79 - 6B + 4D - 7.5F + 5AF - 5.5BD + 4.5DF + 4D^2 - 9F^2$$

with the following factors:

Design Factor	Label	Type	Levels
Quadrant Elevation	A	Continuous	Low, High
Aspect Angle	B	Continuous	Incoming, Crossing
Munition Type	C	Categorical	Mortar, Rockets, Artillery
Shot Range	D	Continuous	Low, High
Operating Mode	E	Categorical	90, 360
Radar to Weapon Range	F	Continuous	Low, High

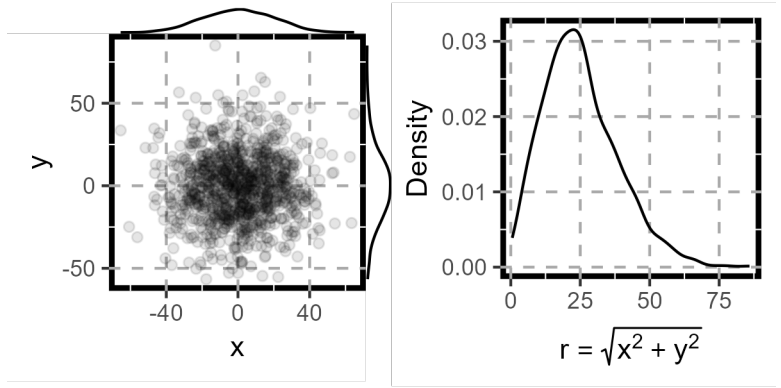
[1]Ahrens, Monica, Rebecca Medlin, Keyla Pagán-Rivera, and John W. Dennis. "Case Study on Applying Sequential Analyses in Operational Testing." *Quality Engineering* 35, no. 3 (December 12, 2022): 534–45. https://doi.org/10.1080/08982112.2022.2146510.

USING THE MODELS

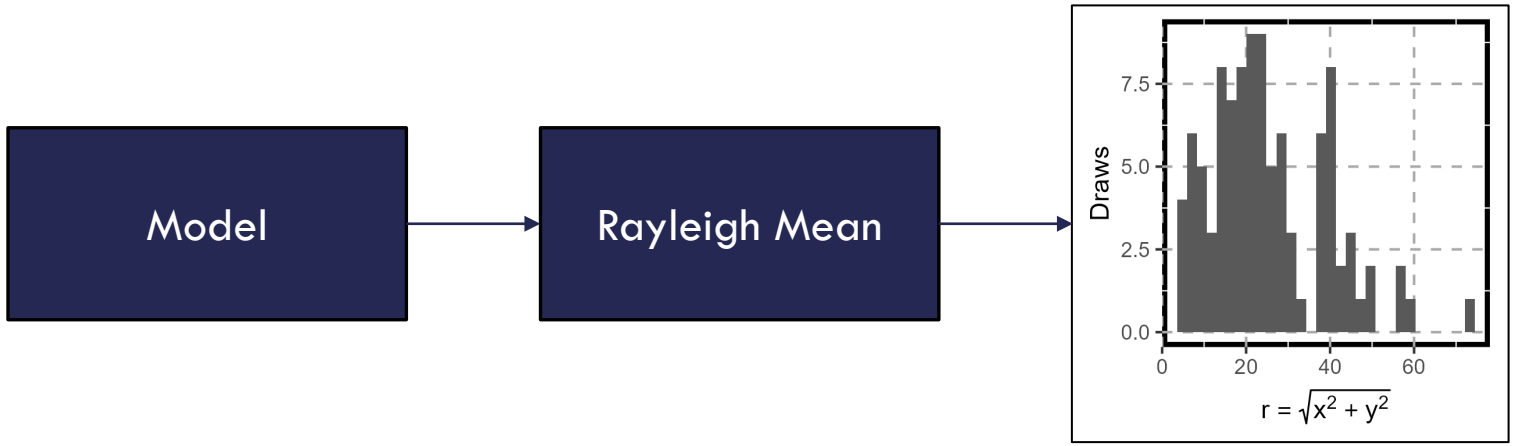
Scenario	Rationale	Formula
Operations ("Real Life")	Most complicated – full model	$79 - 6B + 4D - 7.5F - 5.5B * D + 4.5D * F + 5A * F + 4D^2 - 9F^2$
Operational Testing (OT)	Less fidelity than operations – drop quadratic terms	$79 - 6B + 4D - 7.5F - 5.5B * D + 4.5D * F + 5A * F$
Developmental Testing (DT)	Drop Quadrant Elevation (A)	$79 - 6B + 4D - 7.5F - 5.5B * D + 4.5D * F$
Modeling & Simulation (M&S)	Drop Radar to Weapon Range (F)	$79 - 6B + 4D - 5.5B * D$

Generating Data

Assume location error is normally distributed in two dimensions: Rayleigh distribution



Models give distribution mean, which can then be used to generate data



Synthetic Data Generation

Can compare:

- Analysis methods
- Design of experiments techniques (test designs via skpr package)

For problems with:

- Different numbers of test phases/data sources
- Varying data sizes, e.g., trials and reps by phase
- Evolving test factors
- Shifts/biases in test data (e.g., in M&S data)
- Different error/noise in measurements

Comparing Methods

Five methods considered:

- Frequentist:
  - Using OT data only
  - All data, Blocking: With shift factors added for M&S and DT
  - All data, Without blocking: No shift factors for M&S and DT
- Bayesian informative priors w/ downweighting:
  - Resetting intercept uncertainty to prior value
  - Doubling intercept uncertainty

Key metric: RMSE between Rayleigh means

- Fitted model vs. Operational model
- Computed on full factorial dataset generated using the Operational model

Comparing Methods

Five methods considered:

- Frequentist:
  - Using OT data only
  - All data, Blocking: With shift factors added for M&S and DT
  - All data, Without blocking: No shift factors for M&S and DT
- Bayesian informative priors w/ downweighting:
  - Resetting intercept uncertainty to prior value
  - Doubling intercept uncertainty

Key metric: RMSE between Rayleigh means

- Fitted model vs. Operational model
- Computed on full factorial dataset generated using the Operational model

Benefit of working with synthetic data: We know the "truth"!

Example 1: Limited OT

Consider scenarios where OT is limited and M&S is quite a bit larger than DT

Intuition:

- Integrated testing should provide a benefit
- Challenge of managing different data sizes and changing test factors

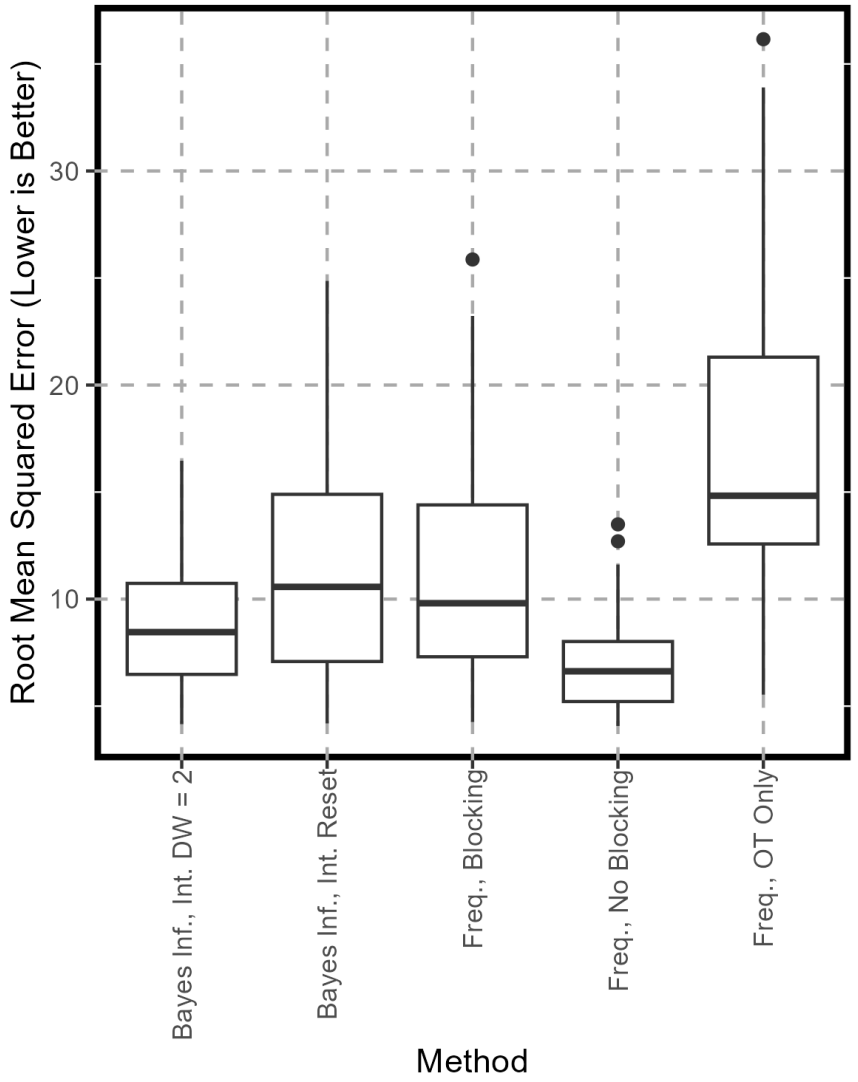
Parameter	Value
M&S Trials	Full factorial (9 trials)
M&S Reps	100
DT Trials	10, 20, 40
DT Reps	5, 10
OT Trials	10, 20
OT Reps	1, 2
DT Optimality	D
OT Optimality	D

(Additional assumption: No more than 200 DT datapoints.)

Example 1: Summary

Takeaways:

- Integrated evaluation helps provide better models
- Without blocking seems to do a little better

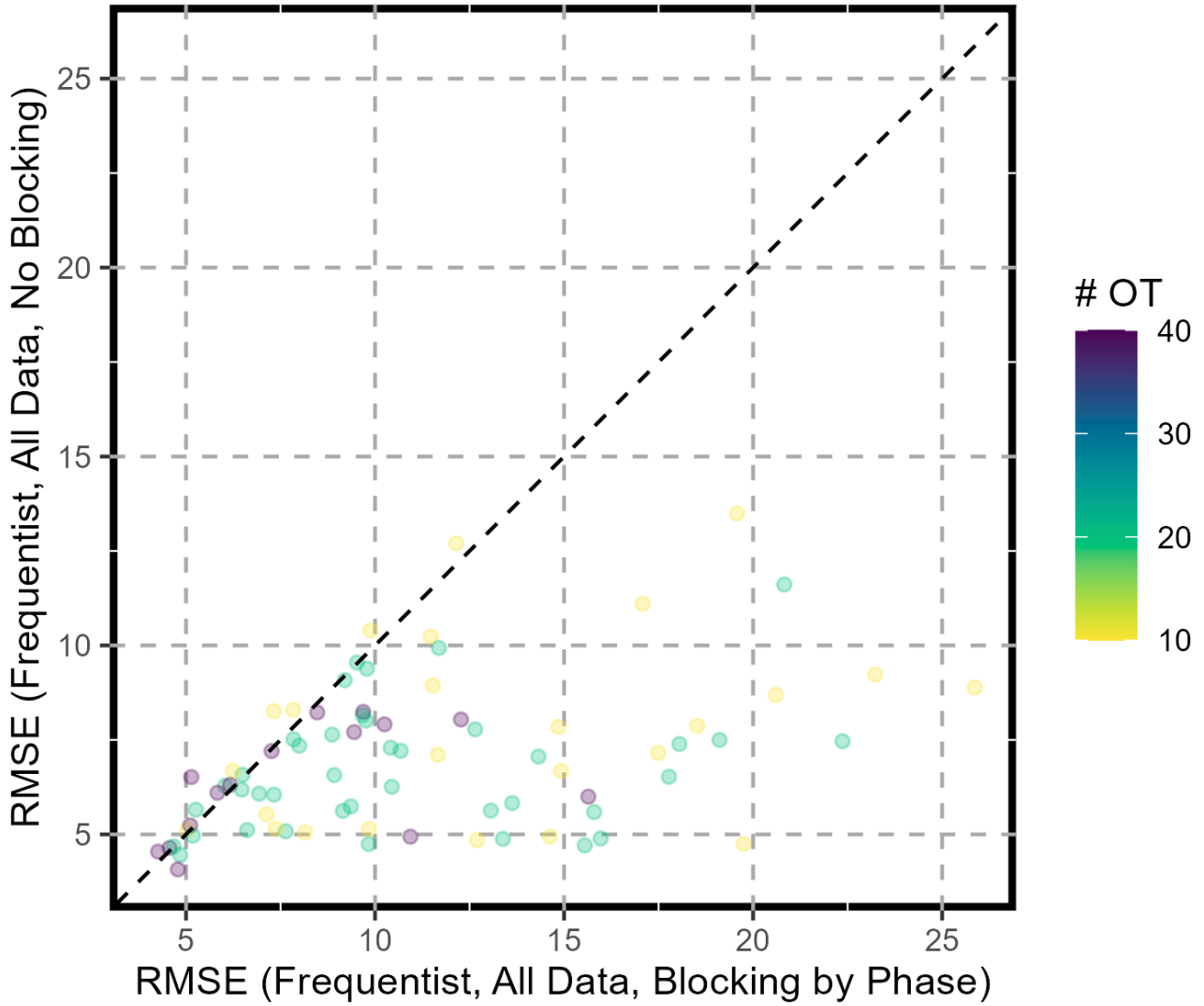


Example 1: Effect of Blocking

Error is lower for integrated model *without* blocking

Takeaways:

- Blocking shifts just fits bias due to noise in small datasets
- Blocking is bad?

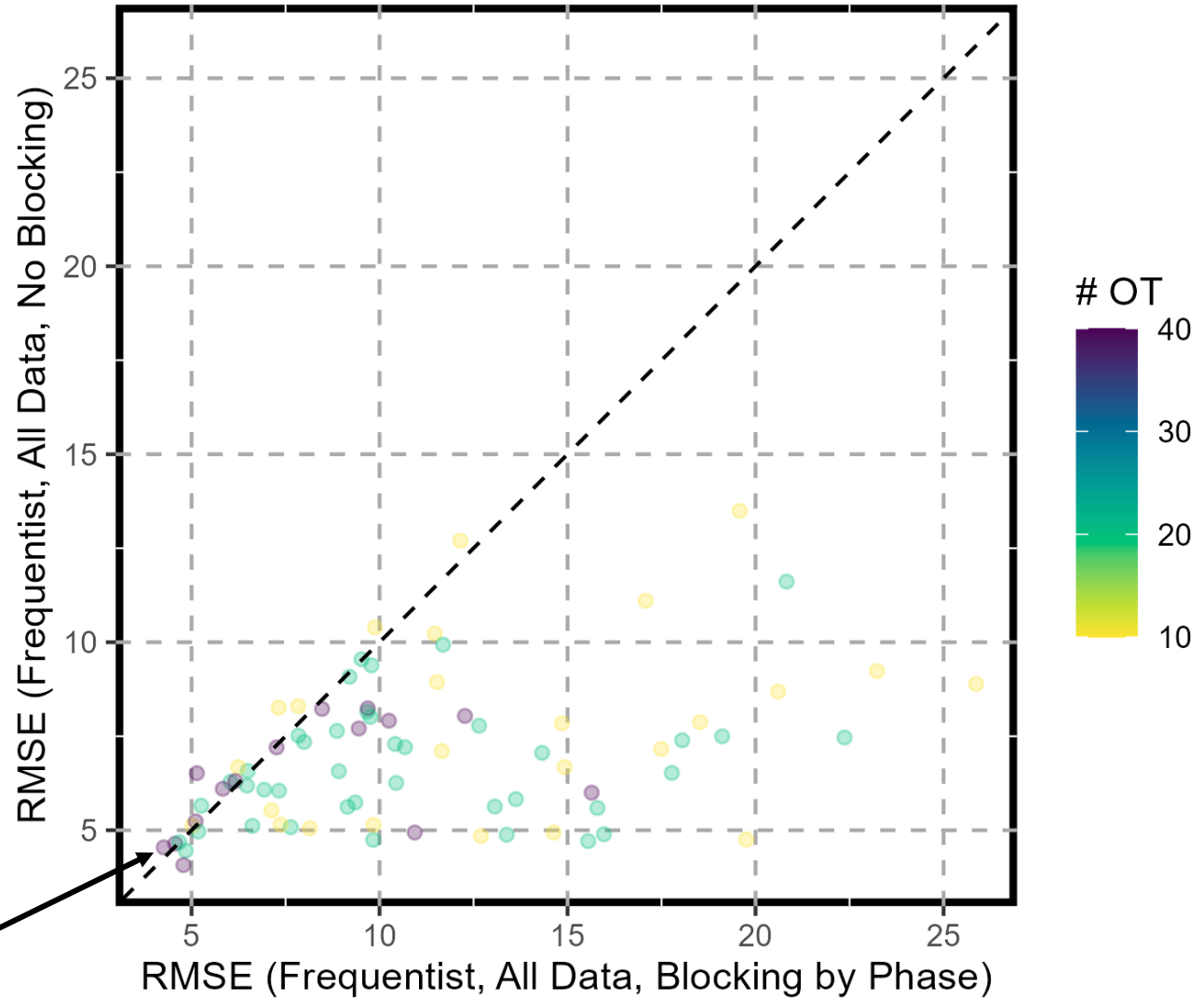


Example 1: Effect of Blocking

Error is lower for integrated model *without* blocking

Takeaways:

- Blocking shifts just fits bias due to noise in small datasets
- Blocking is bad?

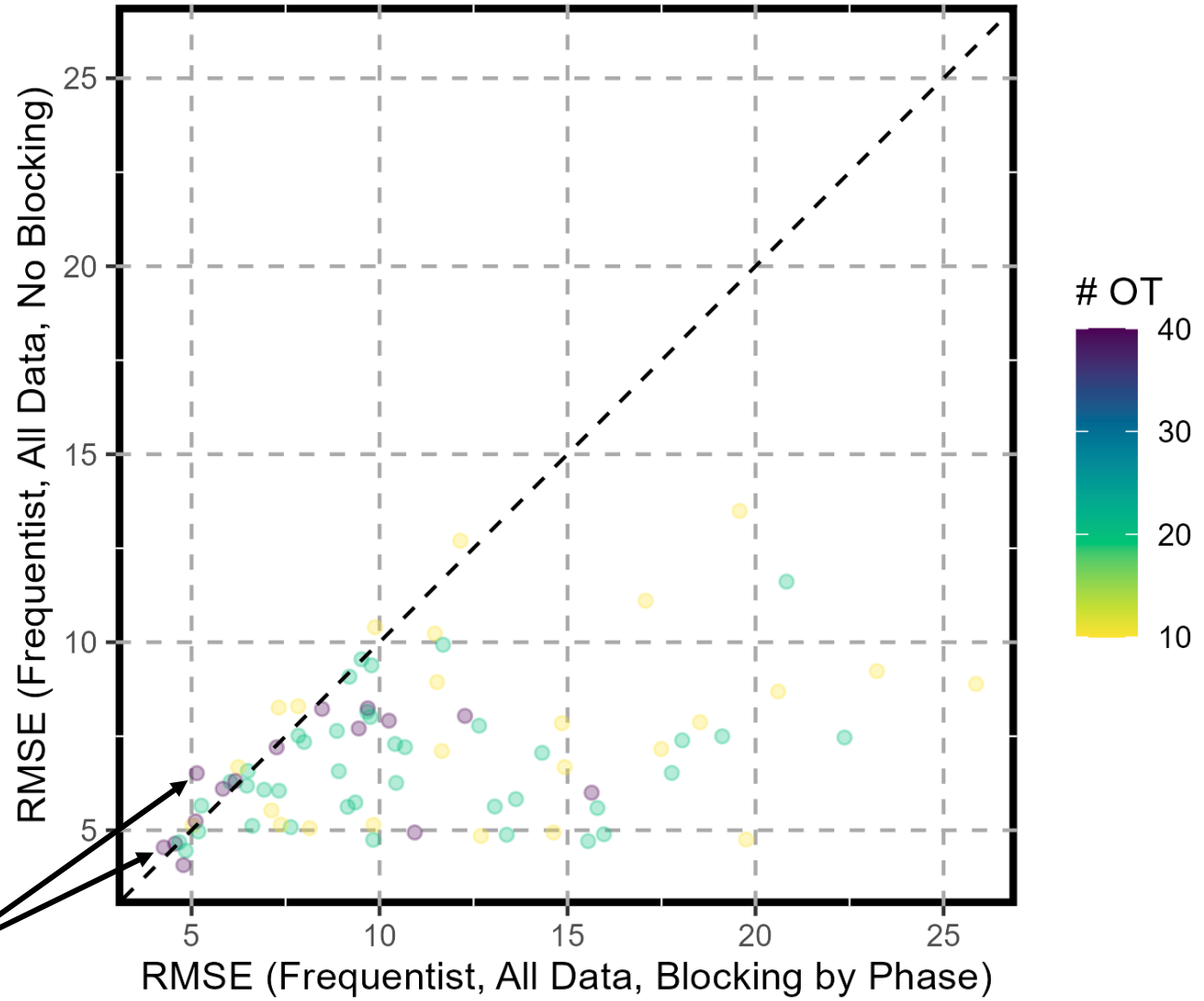


Example 1: Effect of Blocking

Error is lower for integrated model *without* blocking

Takeaways:

- Blocking shifts just fits bias due to noise in small datasets
- Blocking is bad?



Example 2: Limited OT with M&S, DT Shifts

Consider scenarios where OT is limited and M&S is quite a bit larger than DT

- Add random bias to M&S, DT model intercepts

	M&S	DT	OT
Intercept	99.6	86.0	79

Intuition:

- Integrated testing should provide a benefit
- Biases in some of the data might change results?

Parameter	Value
M&S Trials	Full factorial (9 trials)
M&S Reps	100
DT Trials	10, 20, 40
DT Reps	5, 10
OT Trials	10, 20
OT Reps	1, 2
DT Optimality	D
OT Optimality	D

(Additional assumption: No more than 200 DT datapoints.)

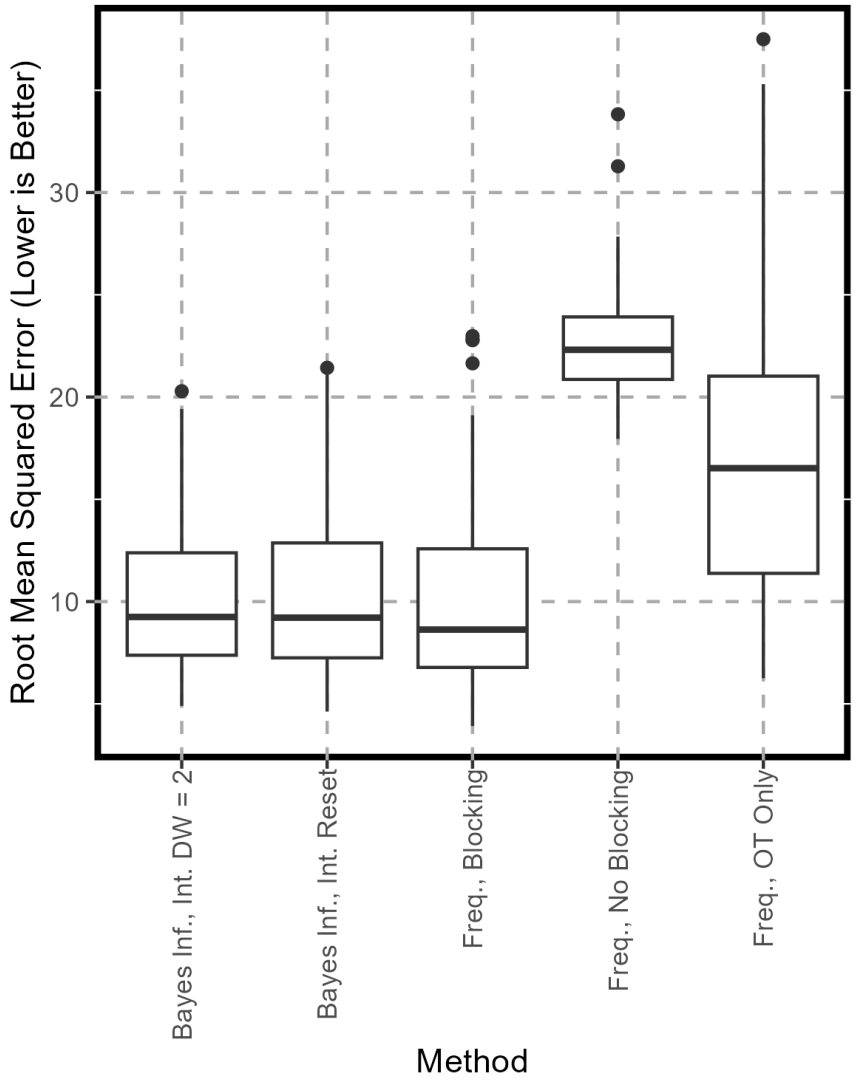
Example 2: Summary

Integration without blocking makes estimates

- Worse than using OT only
- Much worse than other integration methods

Takeaways:

- Integrated testing *mostly* helps provide better models
- **But** some care must be taken in how the data is integrated

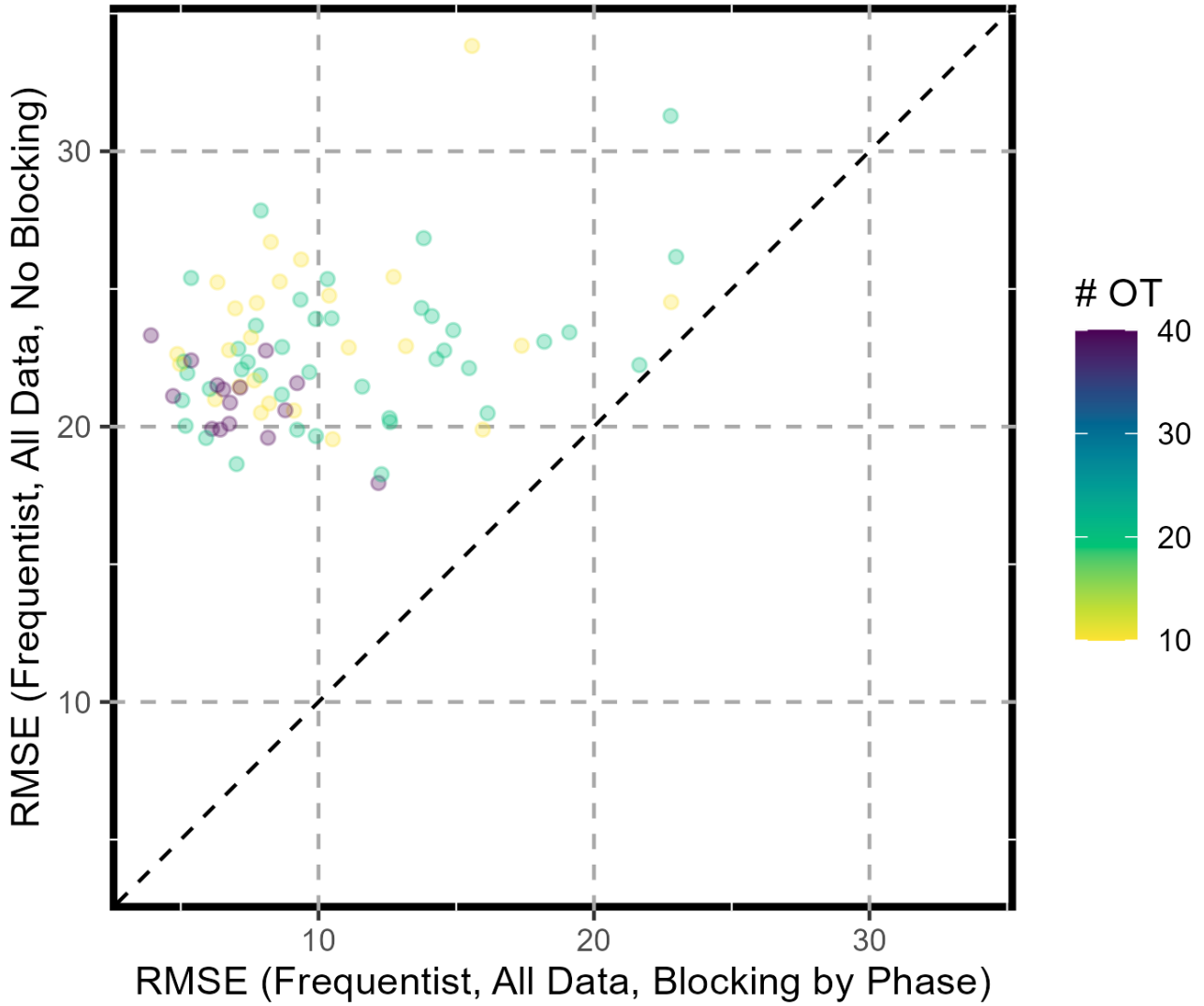


Example 2: Blocking

Error is dramatically lower for integrated model *with* blocking

Takeaways:

- Blocking allows accounting for differences in data sources
- Blocking is good?

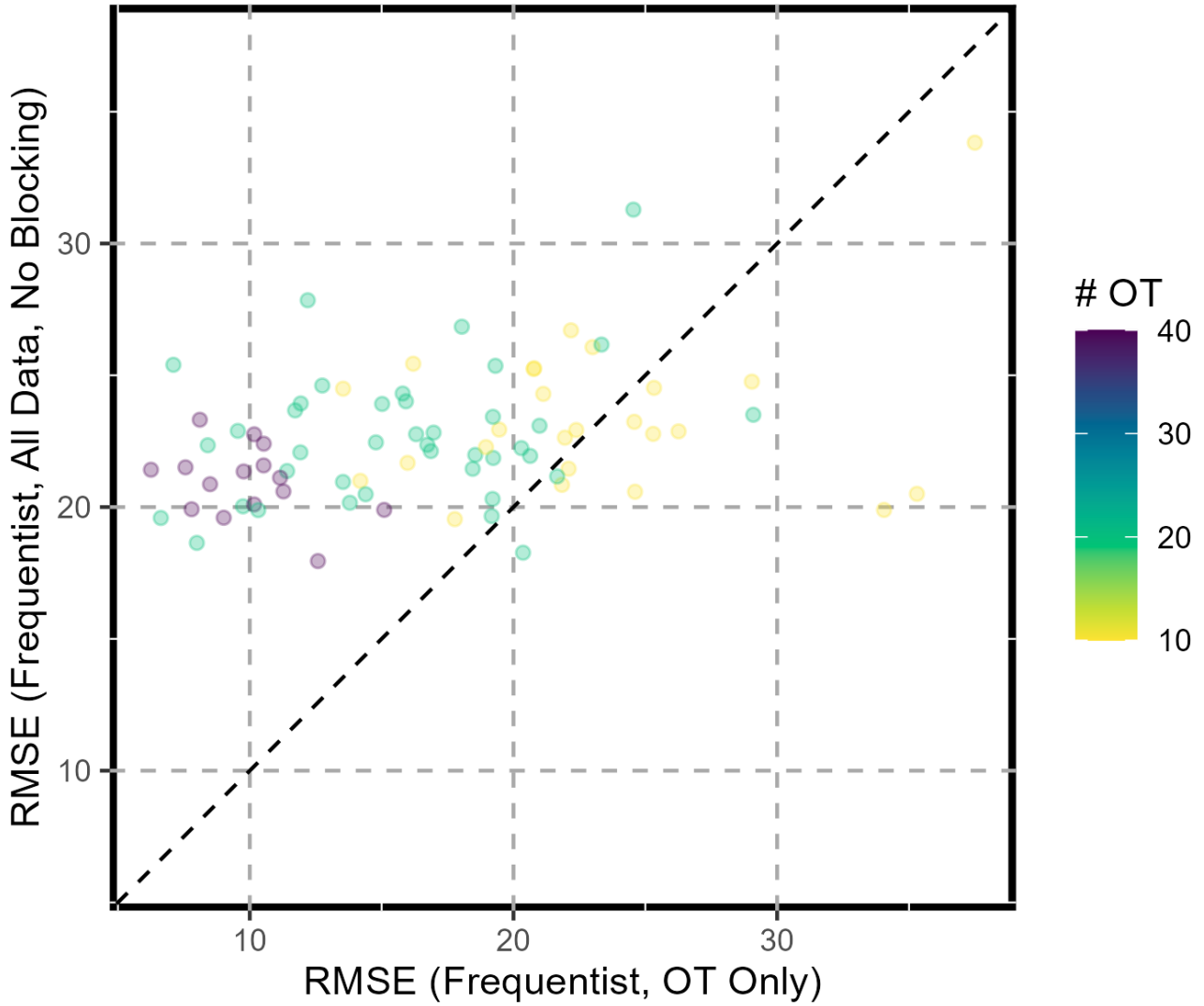


Example 2: Non-Blocking vs. Single-Phase

Non-blocking model actually makes estimates *worse* than single phase model

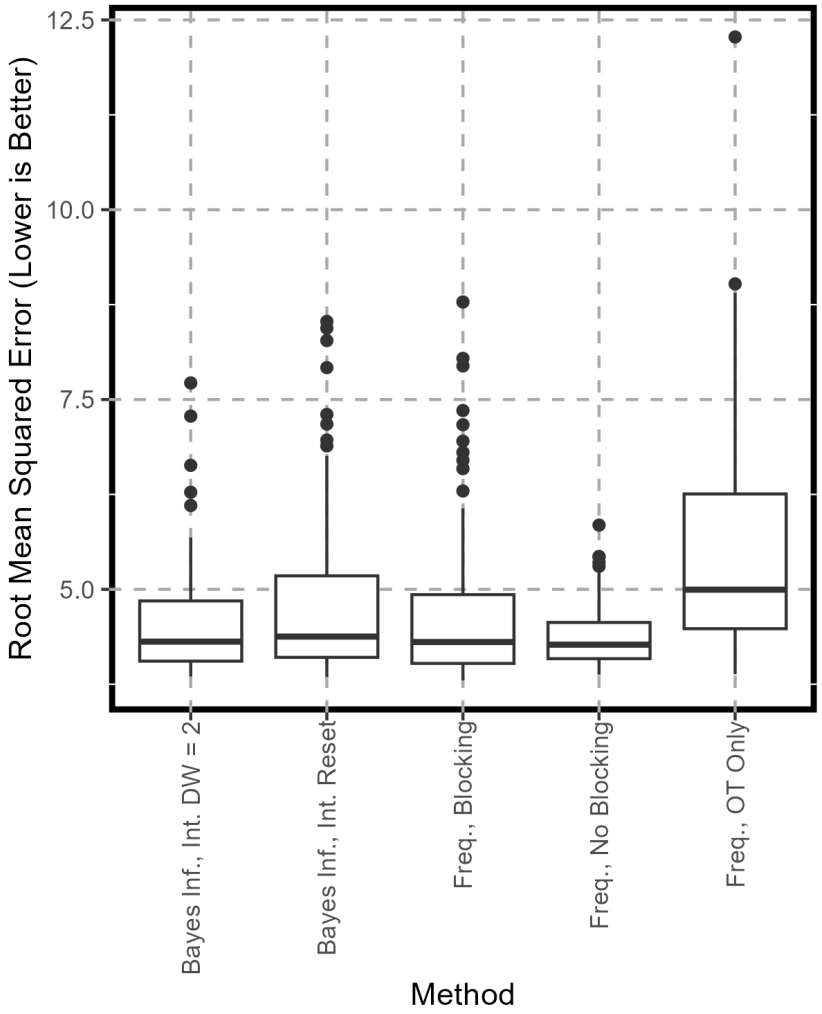
Takeaways:

- Integration of information can make analysis worse if not done carefully

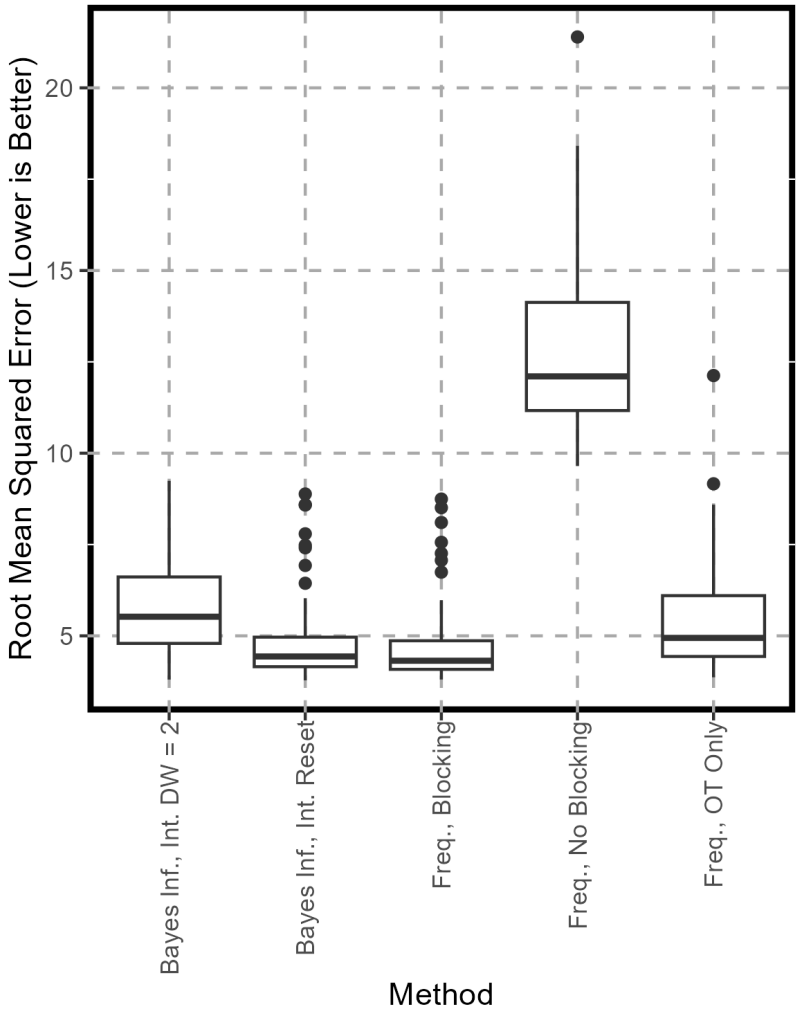


EXAMPLE 3: LARGE OT

Unbiased M&S/DT



Biased M&S/DT



Case Study 3: Binary Data

ACQUISITION INNOVATION
RESEARCH CENTER

Bayesian Binary Test Planning (STAT COE Work)

Given

- prior $Beta(a_0, b_0)$ on probability of failure p (e.g., assembled from downweighted previous test results)
- binary test results a successes and b failures,

the Bayesian posterior distribution is $Beta(a_0 + a, b_0 + b)$.

Also, given a number of tests n, the probability of getting a successes (and $b = n - a$ failures) is given by the binomial distribution

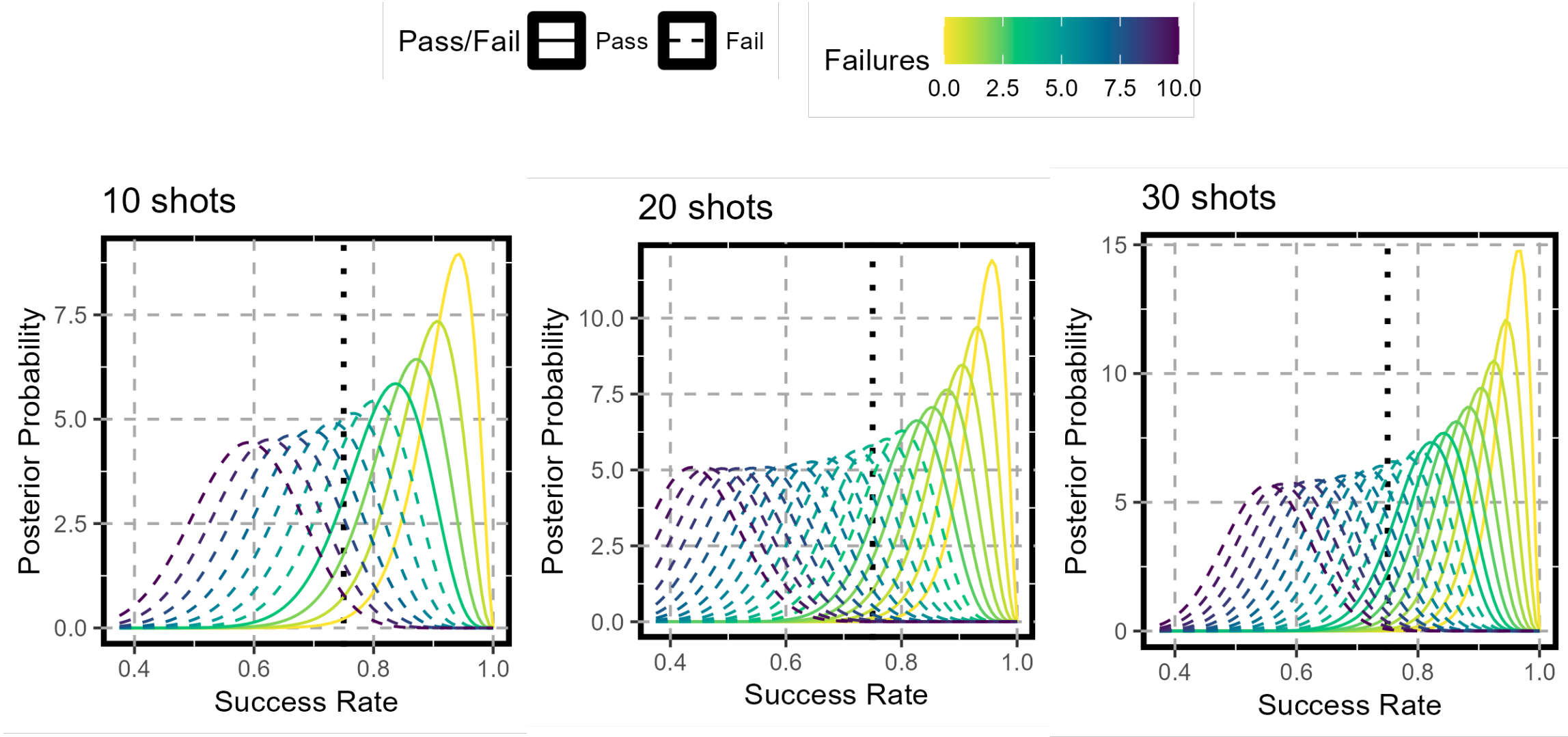
$$\binom{n}{a} p^a (1-p)^{n-a}$$

where p is the true reliability.

For now, assume the following parameters:

Parameter	Value
Prior successes (a_0)	17.84
Prior failures (b_0)	2.656
Prior mean ($a_0/(a_0 + b_0)$)	0.87
Prior sample size ($a_0 + b_0$)	20.496
True reliability (p)	0.85
Threshold	0.75
Lower credible interval	0.2

Results by Shots & Failures

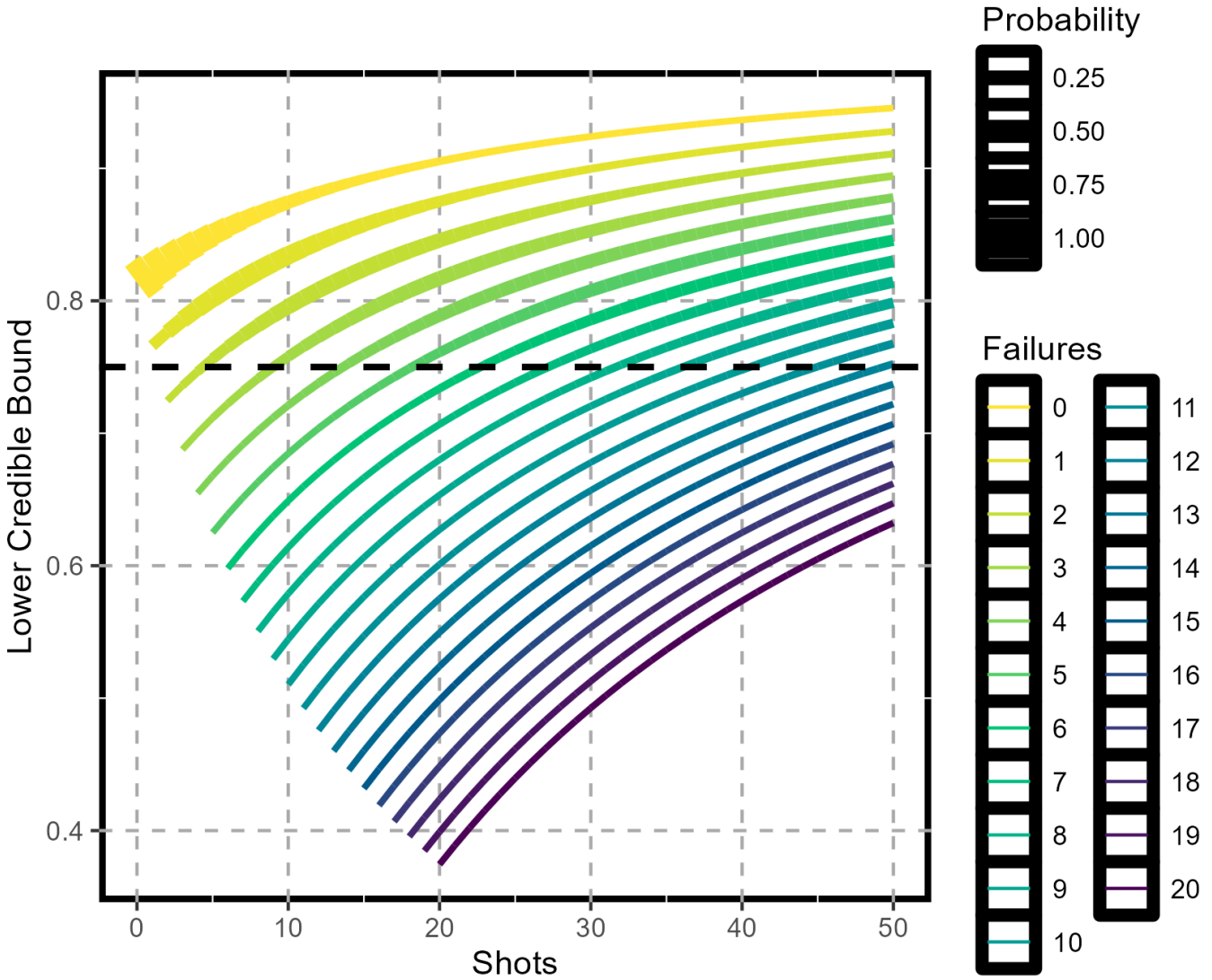


Lower Credible Bound

Compare lower credible bounds vs. threshold

Above threshold = pass

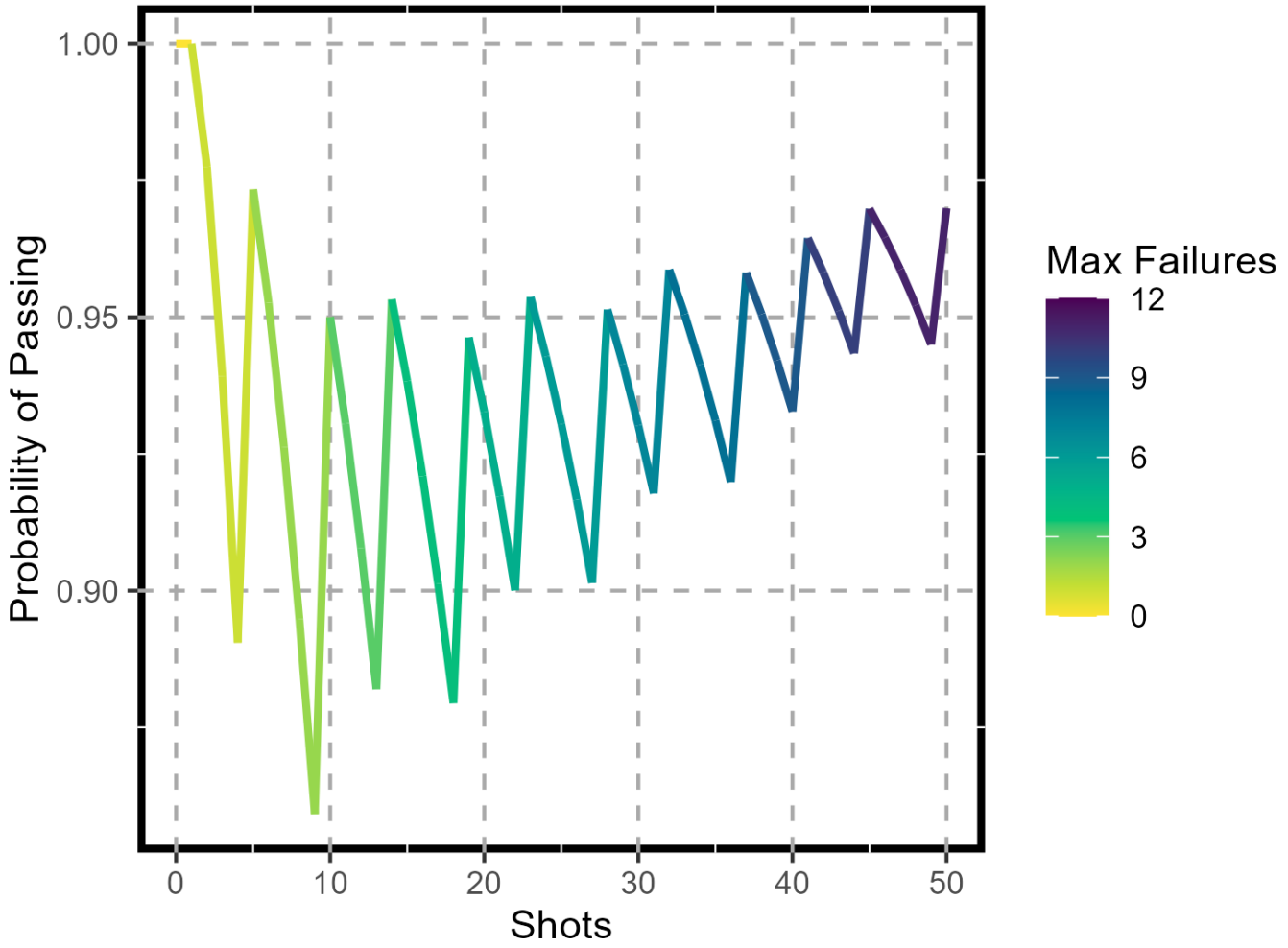
Thicker lines = more probable outcome (according to binomial distribution)



Probability of Passing by Shot

Sum probability (line thickness) above threshold from previous plot to get probability of passing

Increases occur when a new number of failures becomes "acceptable"

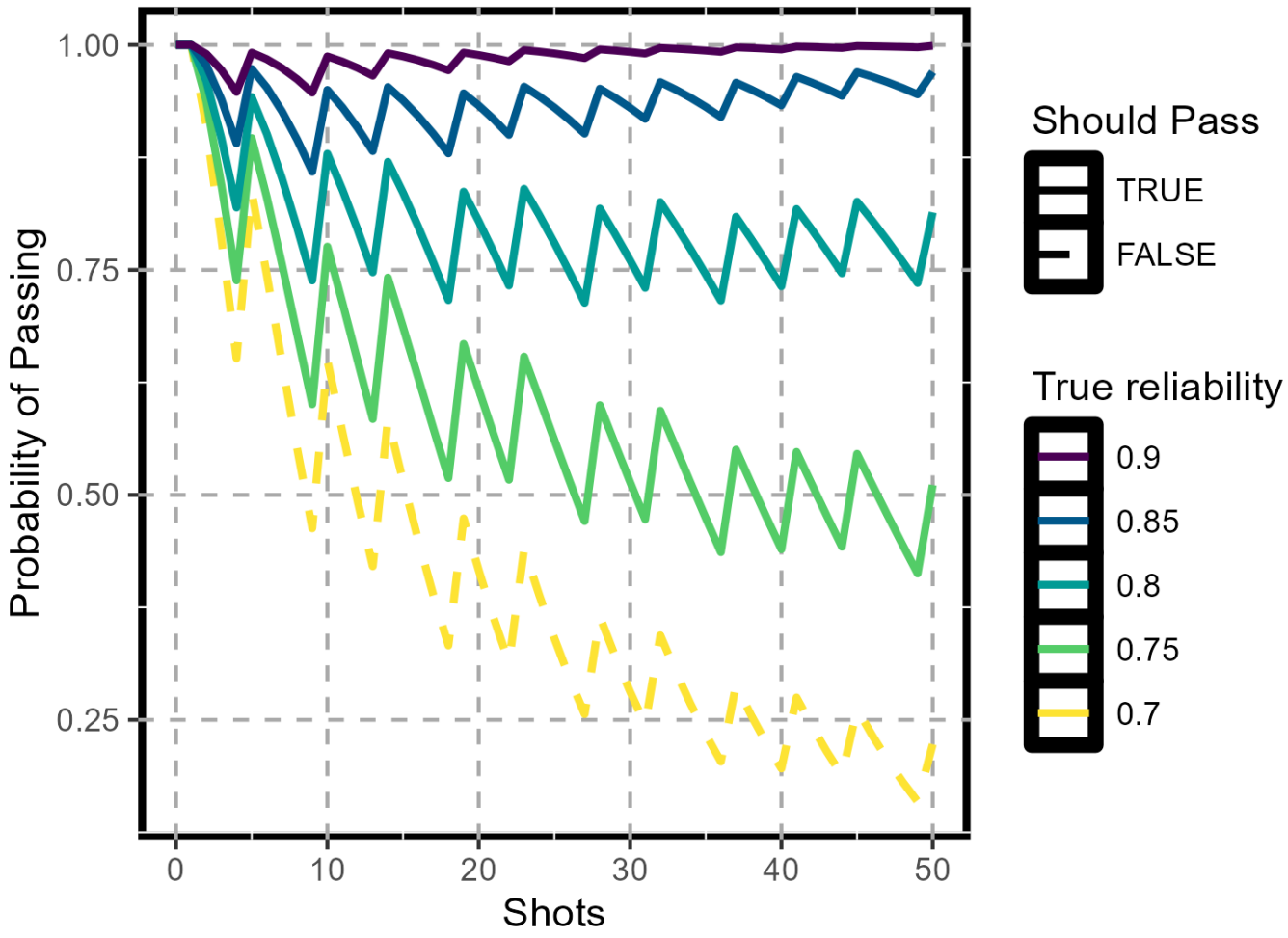


Results by True Reliability

Previous results assumed a true reliability to compute binomial probability

Can compare probabilities of passing across true reliability values

These can be used to size tests to discriminate between candidate reliabilities



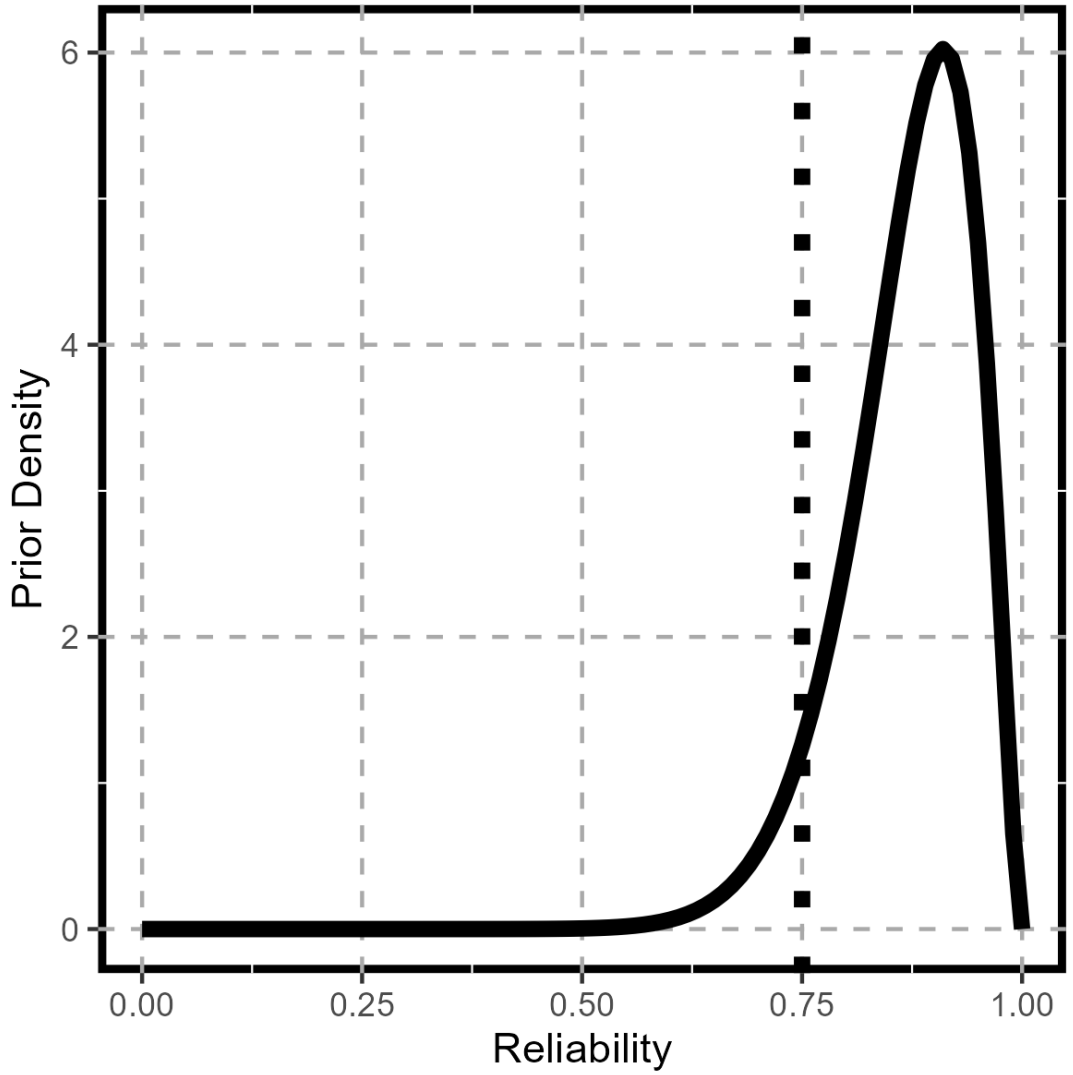
Fully Bayesian Treatment (VT Work)

In the Bayesian framework, we have beliefs about the value of the true reliability p. Can we use this information?

Prior Predictive Distribution

Integrate potential test results across true reliability p, weighted by the prior (right)

Gives the probability of test outcomes given our *current* knowledge



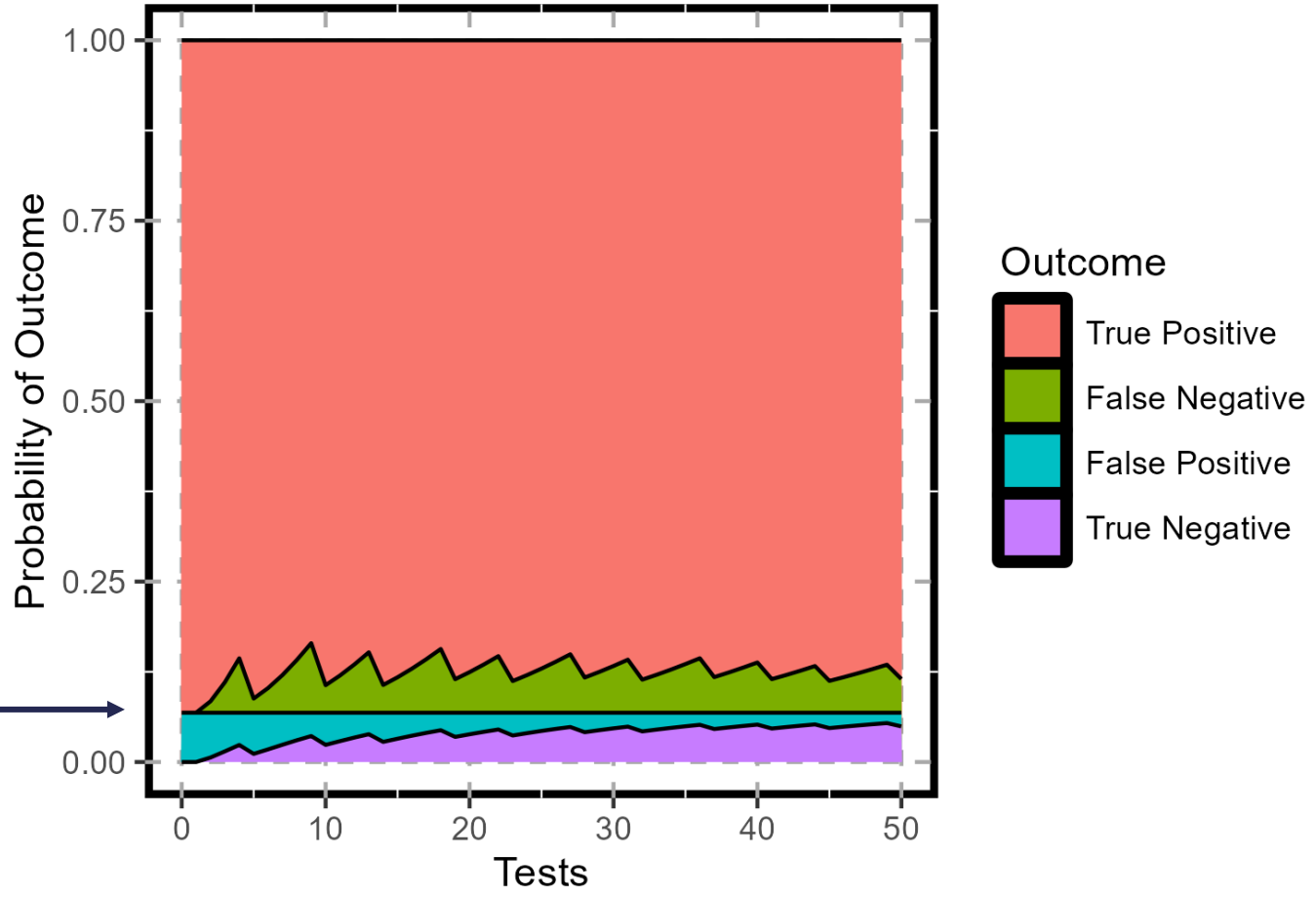
Computational Notes

1. Make a fine grid of p values and evaluate the prior probability at each
2. Similarly, make a grid of number of tests (n) and number of successes (a) values
3. Evaluate:
    1. The probability of getting a successes for each values of n and p
    2. Whether the posterior associated with n tests and a successes meets the acceptance threshold
4. Use quadrature to approximately integrate the acceptance/rejections across p, weighted by the prior, considering the cases where p meets or fails to meet the threshold

Results by Decision (Same Example)

Confusion matrix: Bin results by system (above/below threshold) and outcome (pass/fail)

Horizontal line: Probability of whether system is good or bad is constant (given by prior) because we do not yet have additional information

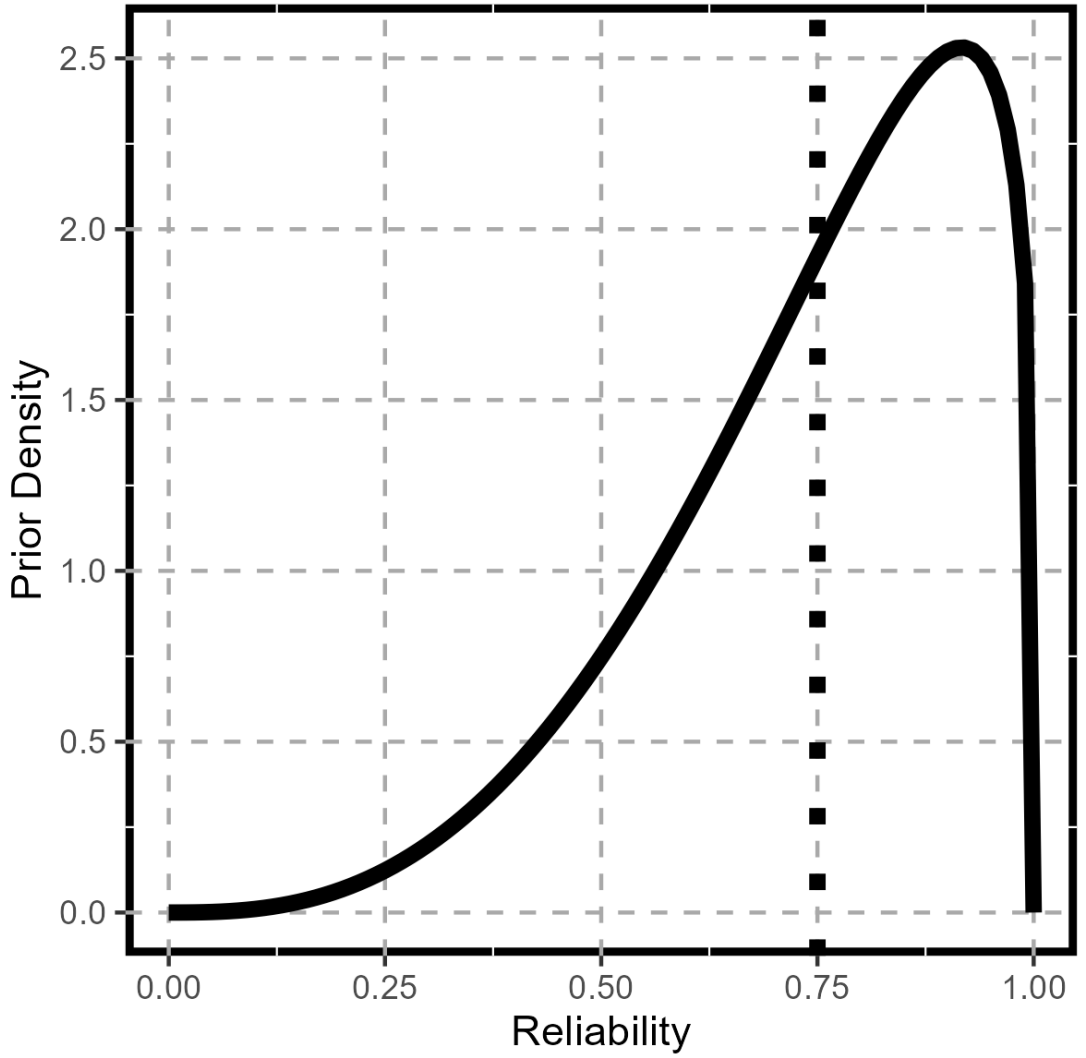


(Note: The prior for this example already meets the acceptance threshold, so negative values start at 0.)

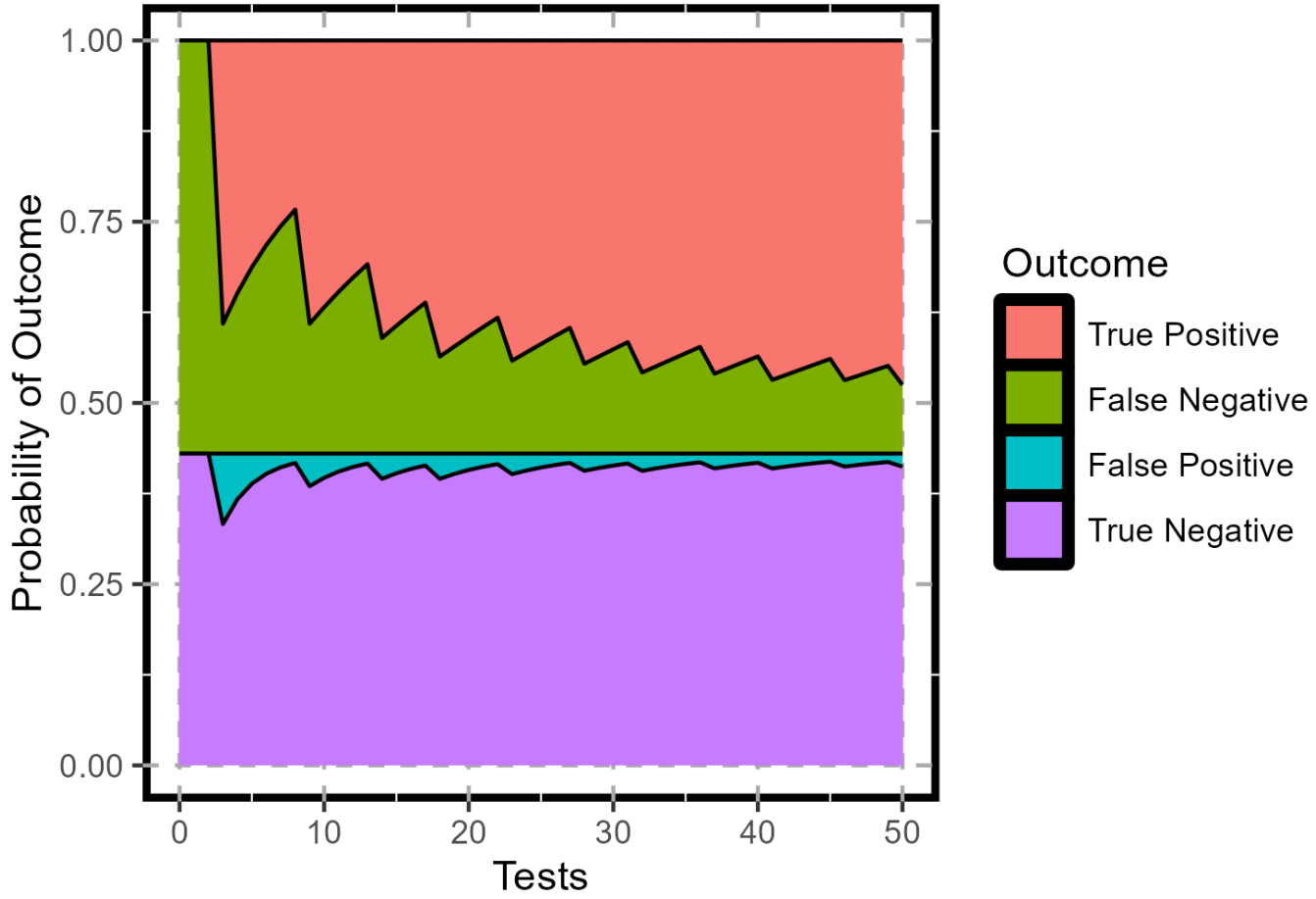
Example: A Less Informative Prior

Assume the following parameters, representing a more ambivalent prior:

Parameter	Value
Prior successes (a_0)	3.75
Prior failures (b_0)	1.25
Prior mean ($a_0/(a_0+b_0)$)	**0.75**
Prior sample size (a_0+b_0)	5
Threshold	**0.75**
Lower credible interval	0.2



Results for Less Informative Prior

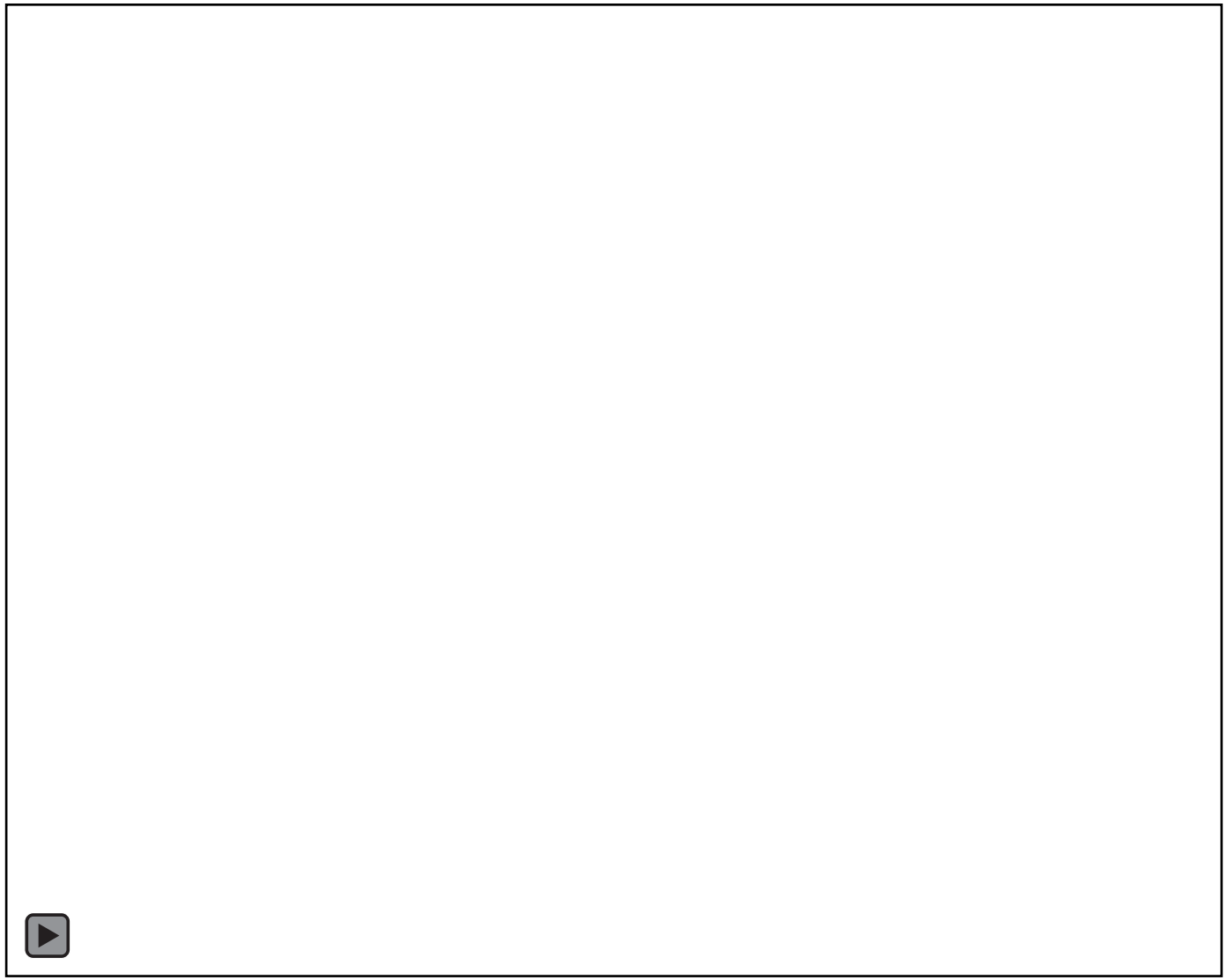


Iterate as Test Results Come In

Iterate as Test Results Come In

Iterate as Test Results Come In (Worse Luck)

Iterate as Test Results Come In (Worse Luck)



Binary Tests: Takeaways

The Bayesian approach provides a way to incorporate prior knowledge and test results

The prior predictive distribution provides full characterization of probable test outcomes given current knowledge

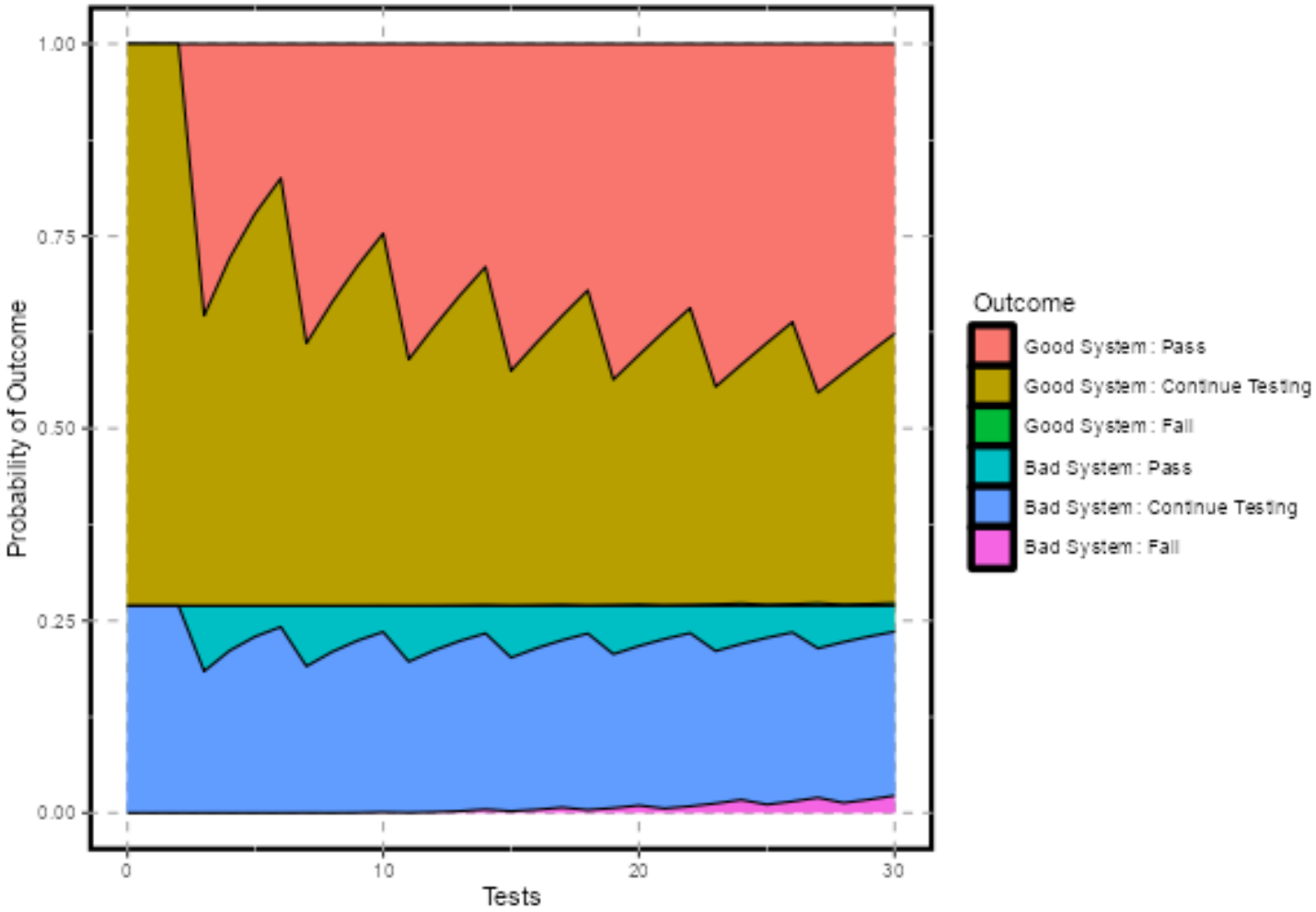
Future work: Add a "keep testing" category for when the posterior is not confident in either direction

Binary Tests: Takeaways

The Bayesian approach provides a way to incorporate prior knowledge and te

The prior predictive dis characterization of pro knowledge

Future work: Add a "ke posterior is not confide



Tool Development

Advanced statistical methods can be hard to implement

Tools allow users to:

- Conduct analysis using pre-written and validated code
  - Externally hosted: Without installation of special software or libraries
- Allow experimentation & exploration of assumptions
- Illustrate and disseminate methods
- Provide examples for training materials and tutorials

Key technology: Shiny

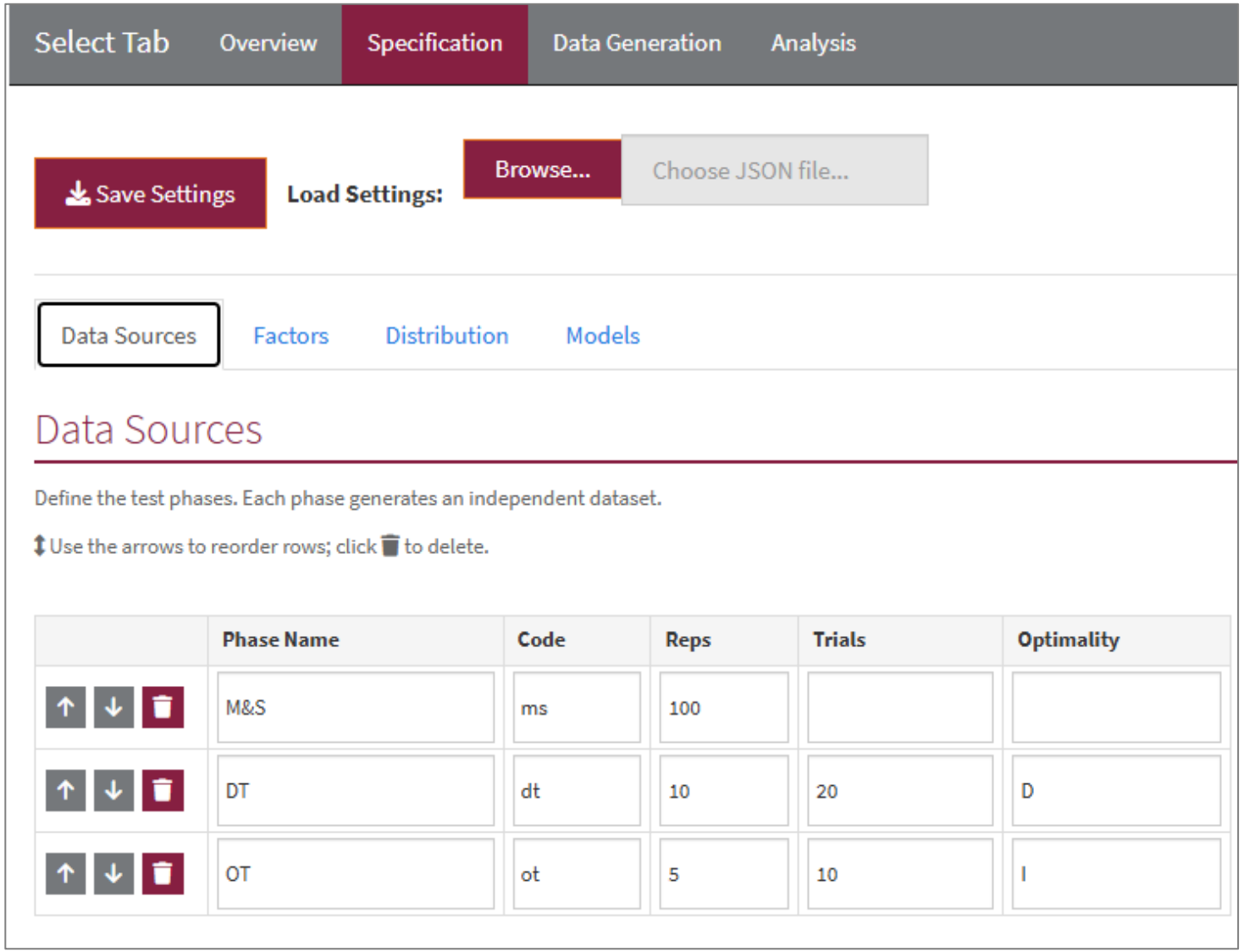
- Facilitates creation of graphical interfaces from R or Python
- Server version allows browser-based access for multiple users

Data Generation

Mimic DoW/D challenges

- Different numbers of test phases/data sources
- Varying data sizes, e.g., trials and reps by phase
- Evolving test factors
- Shifts/biases in test data (e.g., in M&S data)
- Different error/noise in measurements

Basic analysis (right now)

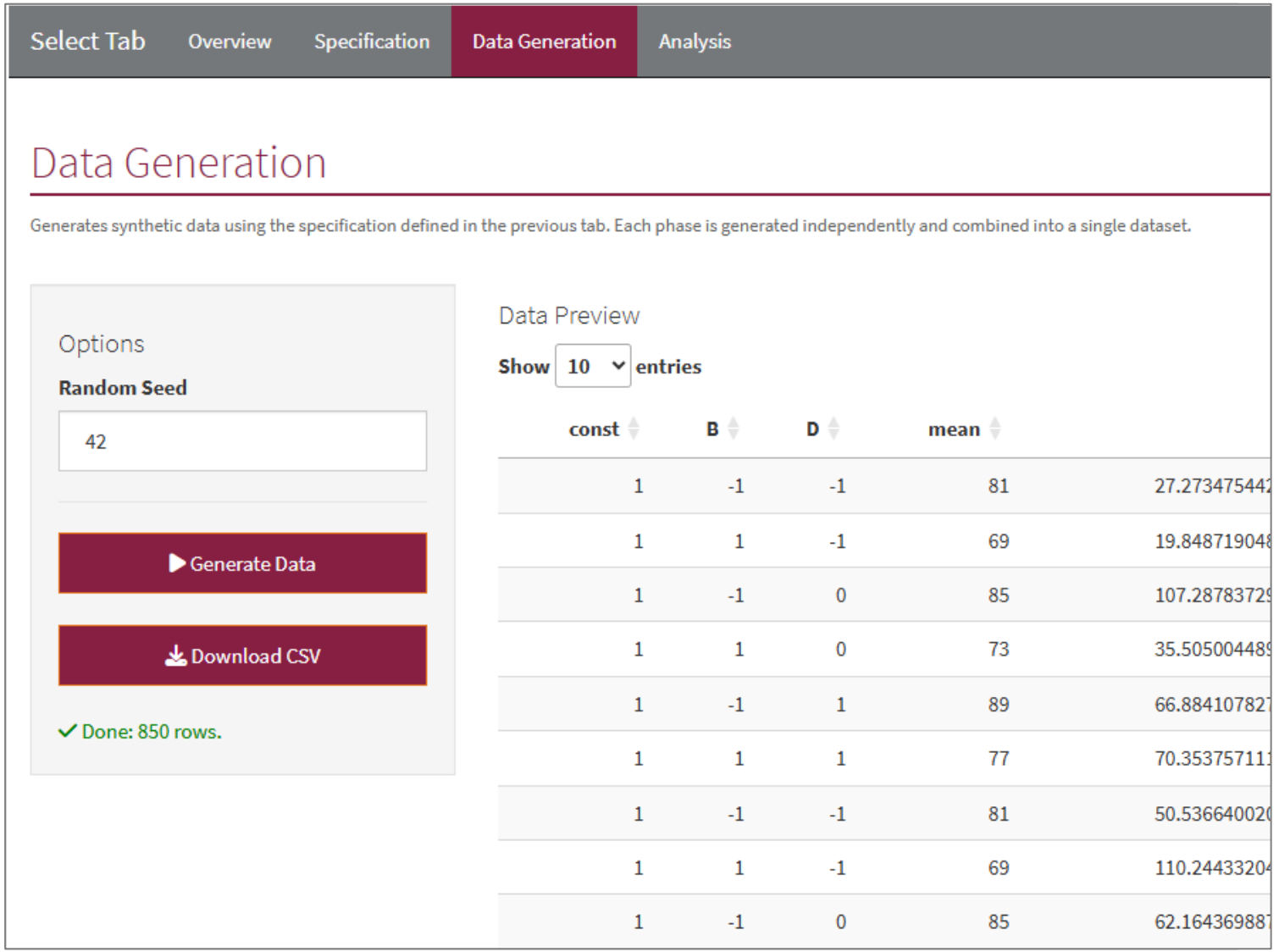


Data Generation

Mimic DoW/D challenges

- Different numbers of test phases/data sources
- Varying data sizes, e.g., trials and reps by phase
- Evolving test factors
- Shifts/biases in test data (e.g., in M&S data)
- Different error/noise in measurements

Basic analysis (right now)

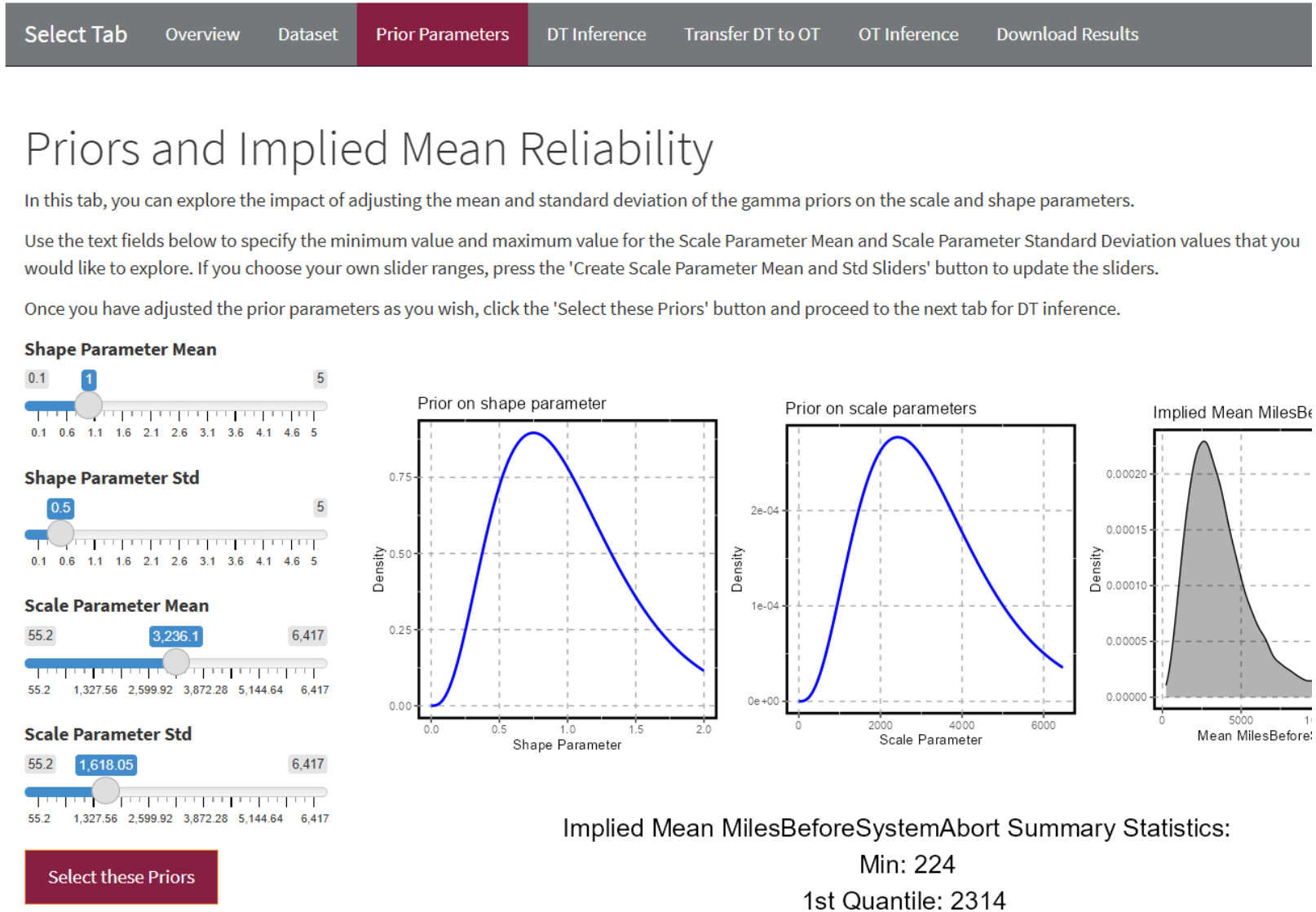


Capabilities:

- Upload data
- Select priors
- Conduct DT inference
- Downweight
- Conduct OT Inference

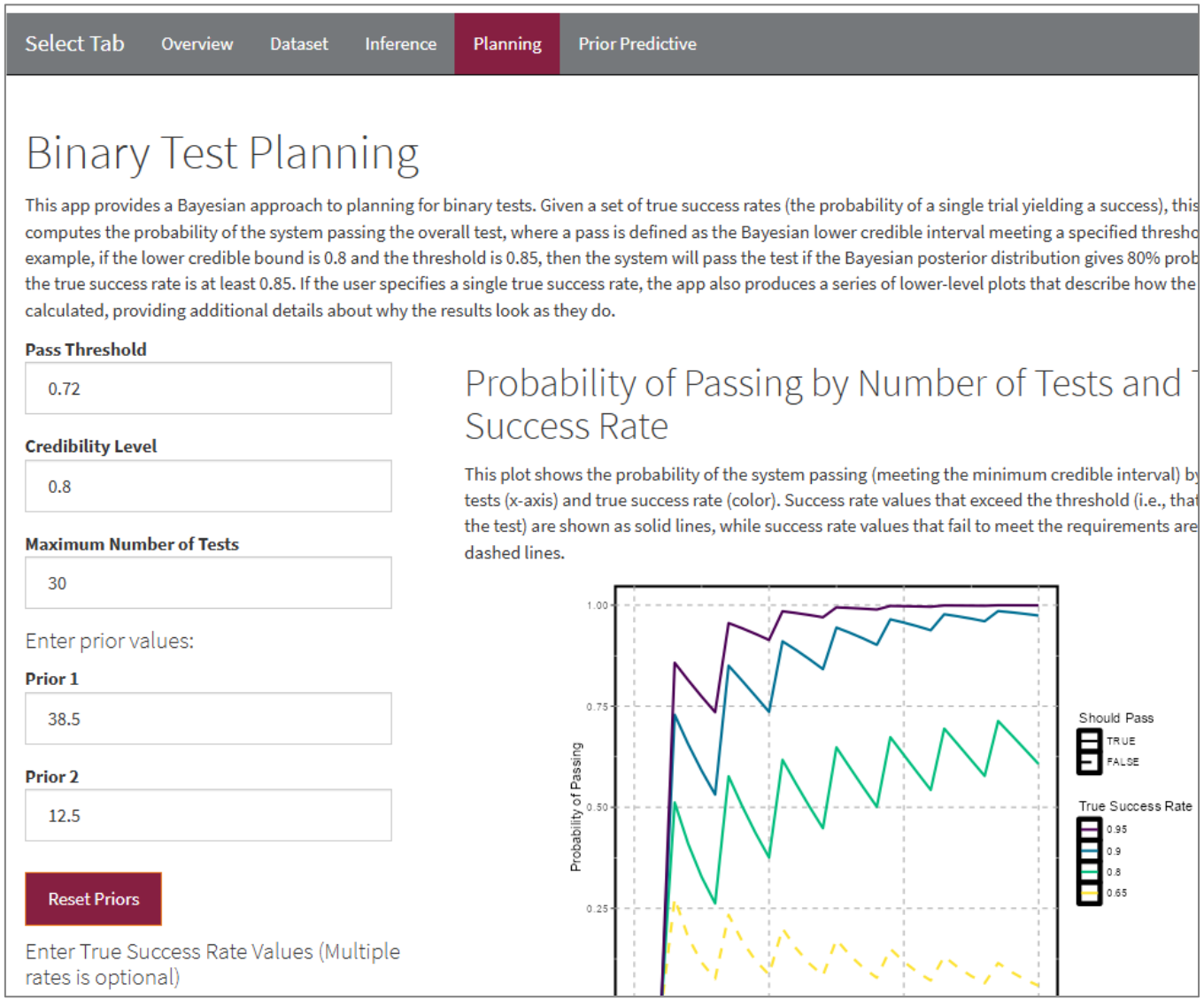
Challenge:

- Inference (MCMC) breaks on free versions of Shiny servers



Functionality

- Upload data
  - Arbitrary number of test phases
- Compute posteriors
- Evaluate test plans
- Compute prior predictive



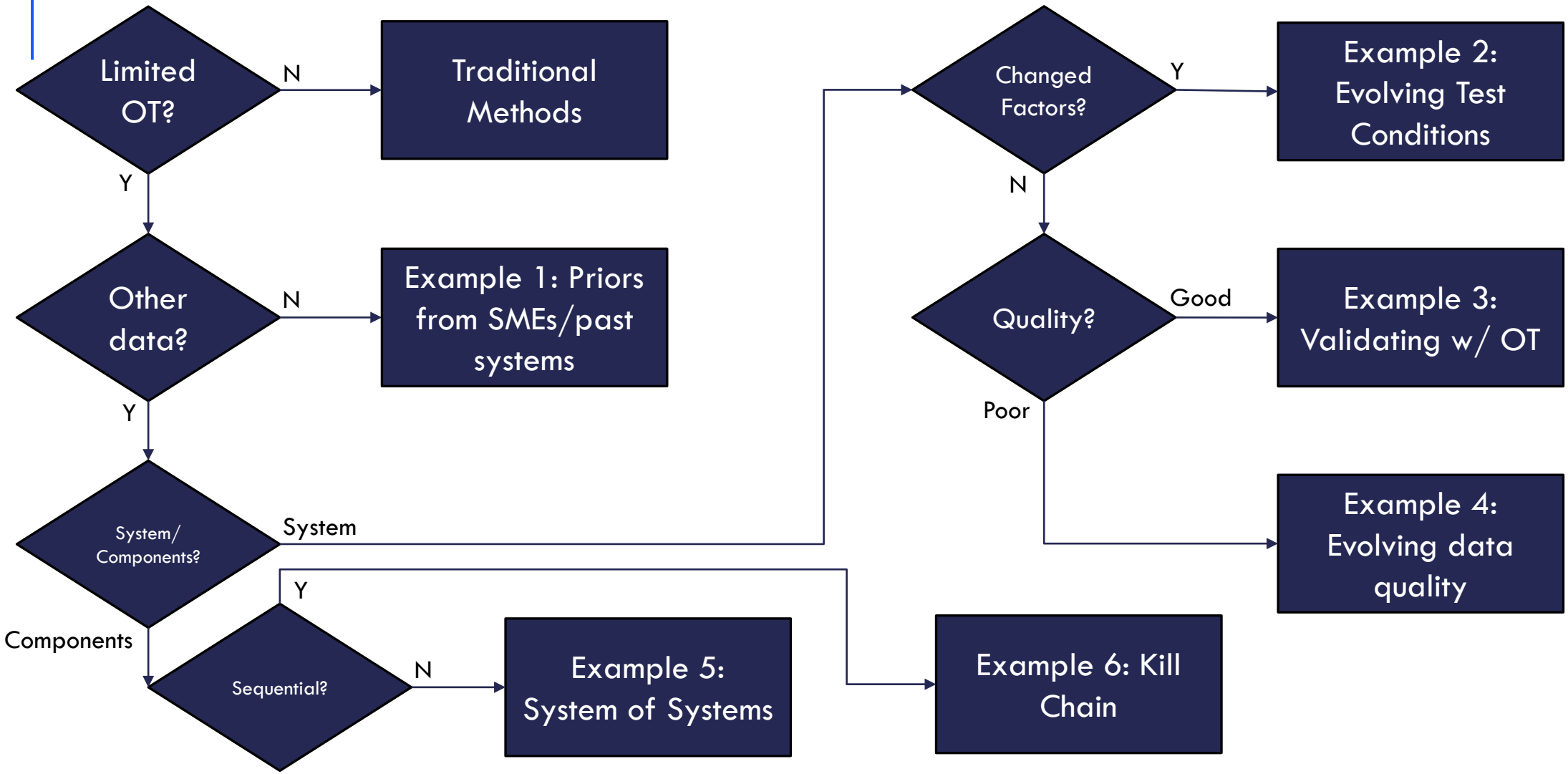
Download: https://github.com/krometis/dataworks2026

Hosted:

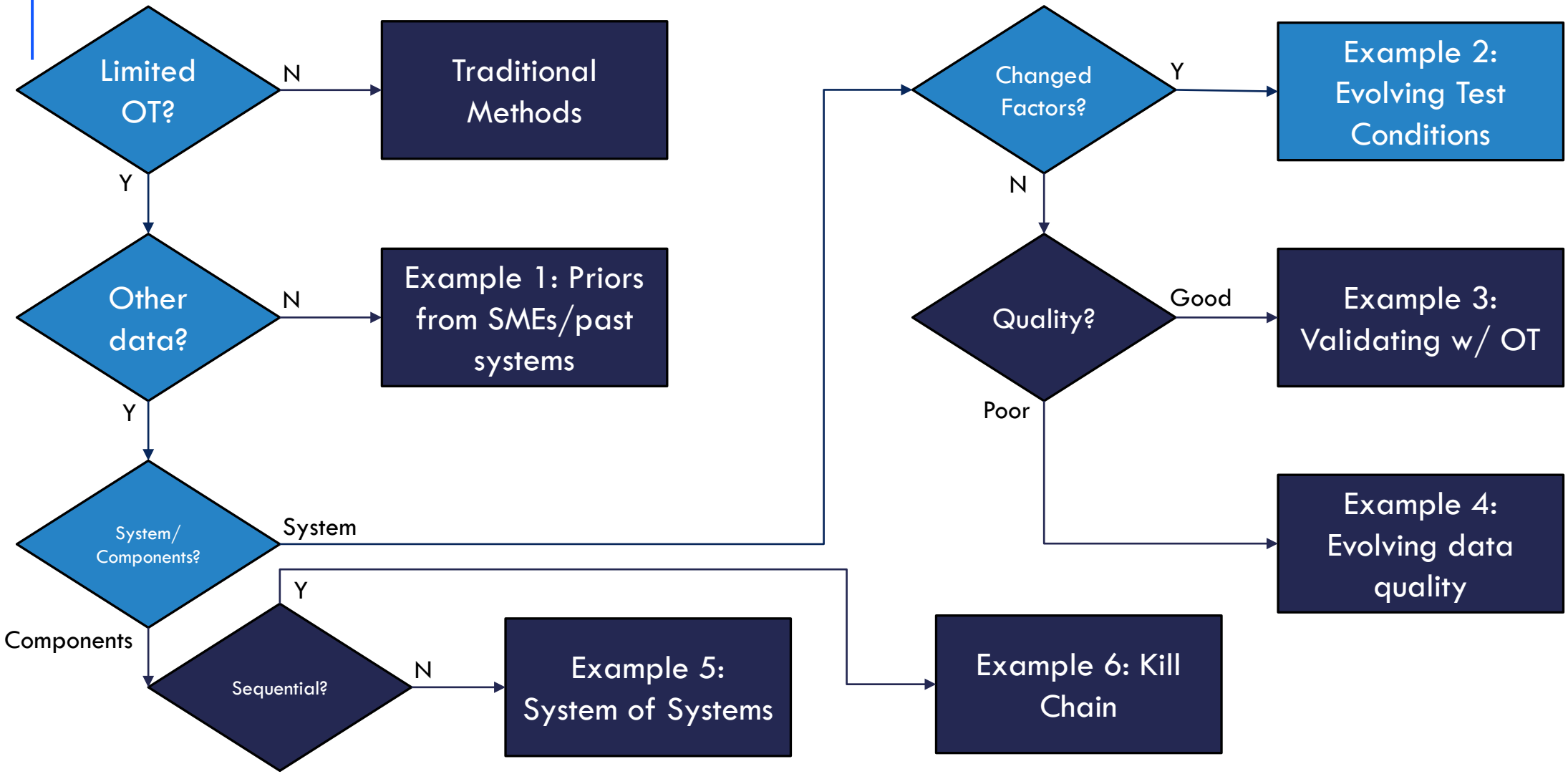
- Data Generation: https://krometis.shinyapps.io/datagen/
- Reliability: https://krometis.shinyapps.io/reliability/
  - Warning: Analysis will crash!
- Binary: https://krometis.shinyapps.io/binary_multi/

Next Steps

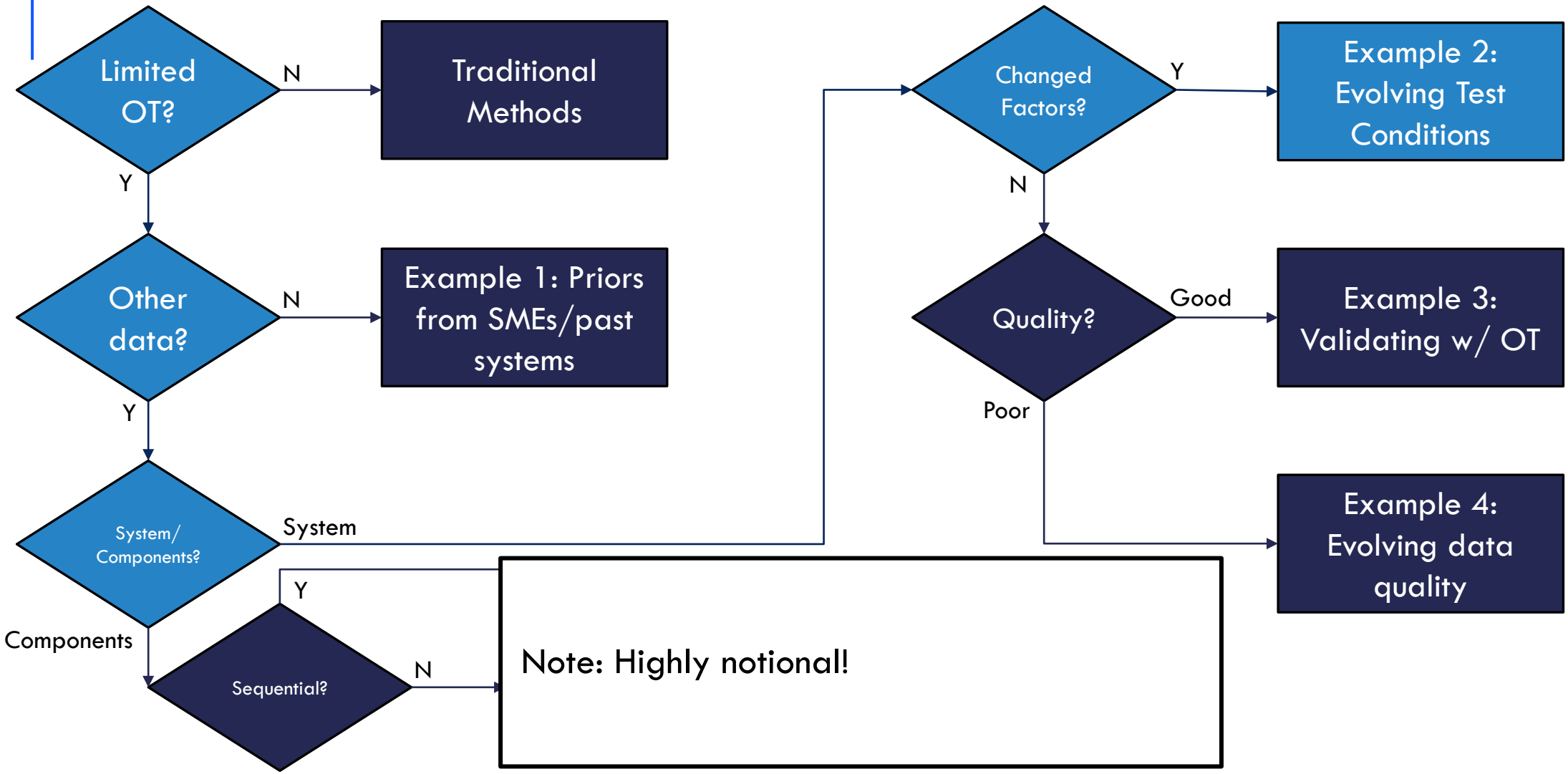
Long-Term Goal: Decision Tree for Best Practices



Long-Term Goal: Decision Tree for Best Practices



Long-Term Goal: Decision Tree for Best Practices



QUESTIONS/THOUGHTS/FEEDBACK?

Concerns/questions about methods

Barriers to implementation & adoption