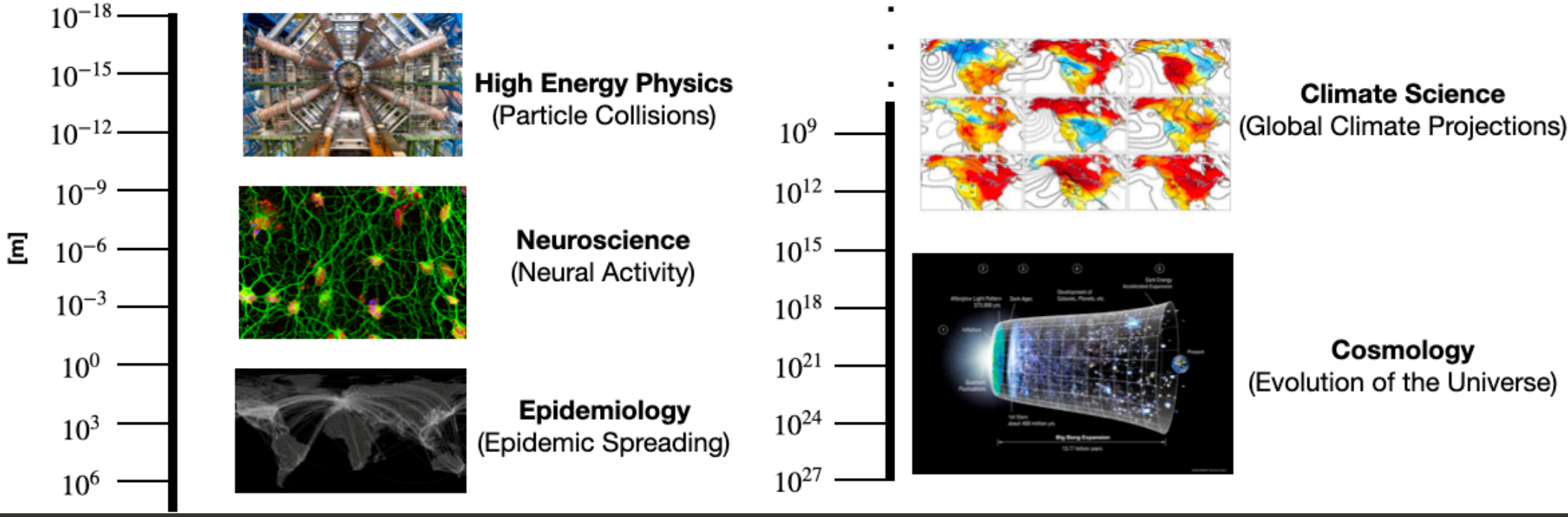
Likelihood-Free Frequentist Inference

Ann B. Lee
Department of Statistics & Data Science / Machine Learning Department
Carnegie Mellon University

Collaborators: Luca Masserano (CMU); Nic Dalmasso (JP Morgan AI); Rafael Izbicki (UFSCar); Mikael Kuusela (CMU); Tommaso Dorigo (Padova); Alex Shen (CMU)

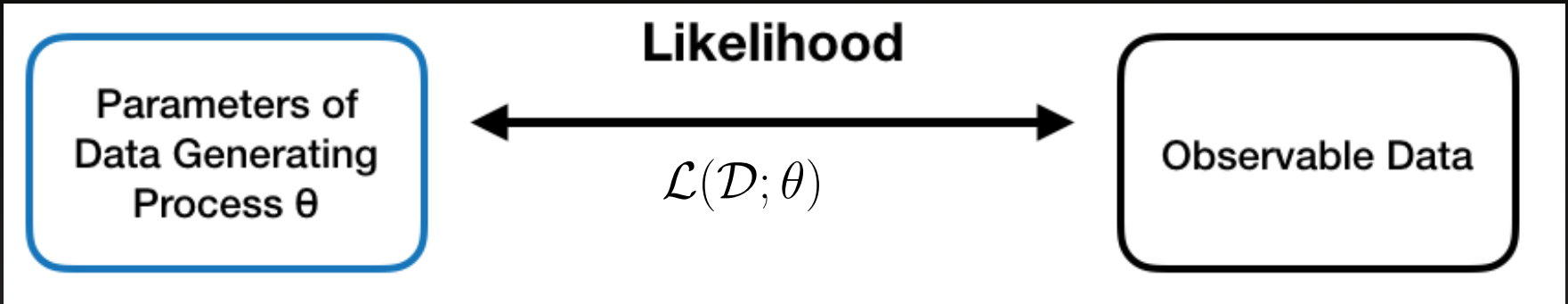
Simulators are Ubiquitous in Science



Credit: Dalmasso (adapted from Cranmer et al, 2020)

- For many complex phenomena, the only meaningful model (theory) may be in the form of simulations.

Likelihood-Based Inference



What is Likelihood-Free Inference?

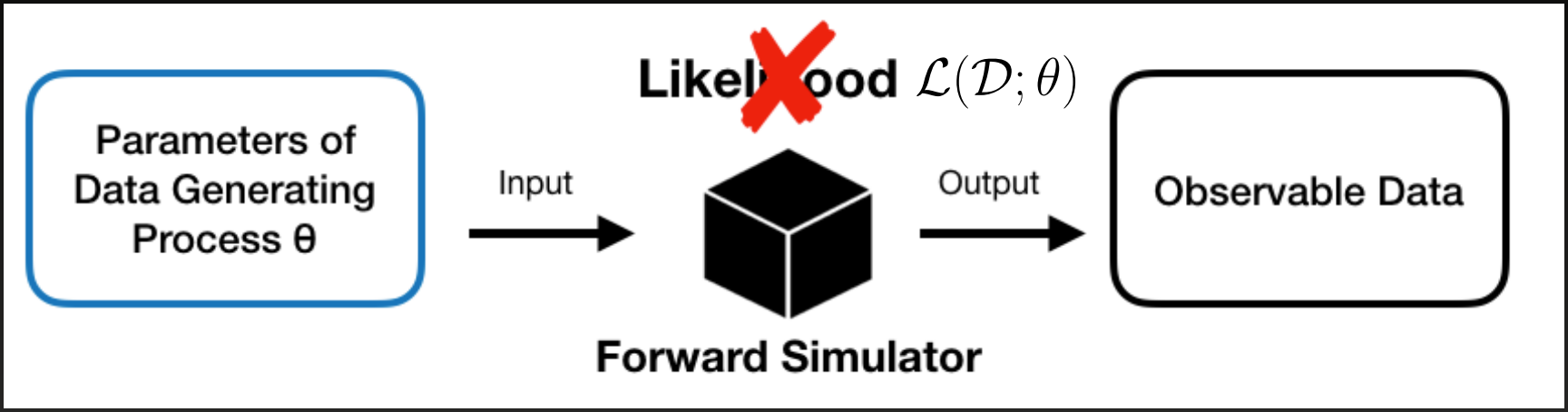


Image credit: Nic Dalmasso

- The likelihood cannot be evaluated. But it is implicitly encoded by the simulator…
- Inference on parameters in this setting is called likelihood-free inference (LFI)

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta),\ \mathbf{X} \sim F_\theta$$

Classical LFI: Approximate Bayesian Computation (ABC)

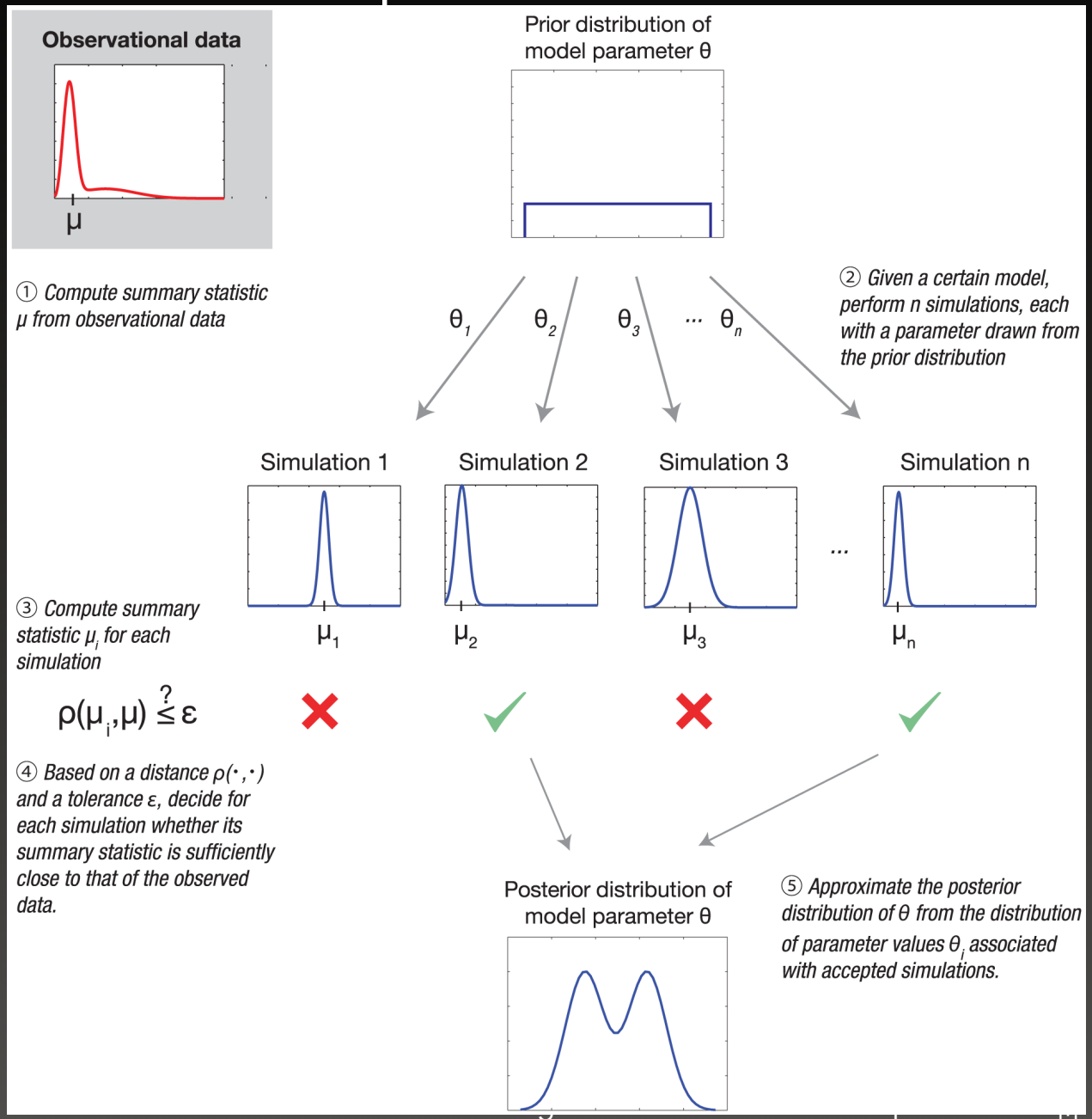


Image credit: Sunnaker et al. 2013

Changing LFI Landscape [Cranmer et al, PNAS 2019]

- More recent developments use ML algorithms to directly estimate key inferential quantities from simulated data

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta), \mathbf{X} \sim F_\theta$$

  - Posteriors, f(**θ**|x) [e.g., Papamakarios et al, 2016; Lueckmann et al, 2016; Izbicki et al, 2019; Greenberg et al, 2019]
  - Likelihoods, f(x|**θ**) or f(x|**θ**)/g(x) [e.g., Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]
  - Likelihood ratios, f(x|$\boldsymbol{\theta}_1$)/f(x|$\boldsymbol{\theta}_2$) [e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]

- These new training-based approaches can handle complex high-dimensional data without a prior dimension reduction. Provide "amortized" inference.

Changing LFI Landscape [Cranmer et al, PNAS 2019]

- More recent developments use ML algorithms to directly estimate key inferential quantities from simulated data

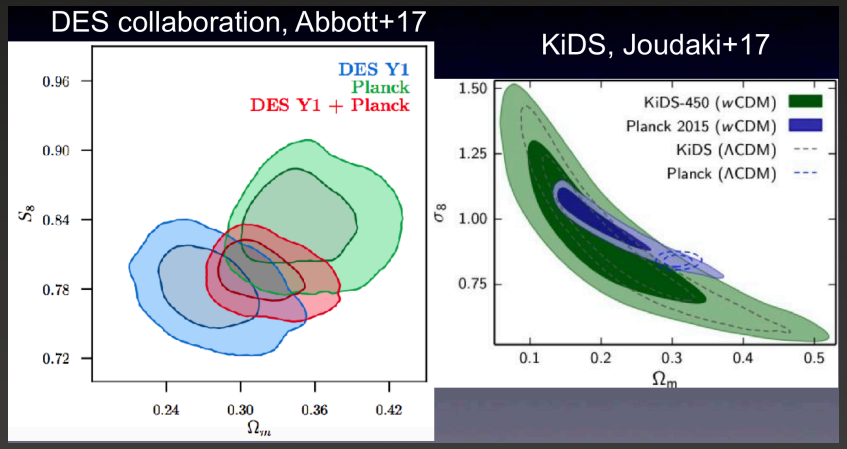
$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta), \mathbf{X} \sim F_\theta$$

  - Posteriors, f(**θ**|x) [e.g., Papamakarios et al, 2016; Lueckmann et al, 2016; Izbicki et al, 2019; Greenberg et al, 2019]
  - Likelihoods, f(x|**θ**) or f(x|**θ**)/g(x) [e.g., Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]
  - Likelihood ratios, f(x|$\boldsymbol{\theta}_1$)/f(x|$\boldsymbol{\theta}_2$) [e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]

- These new training-based approaches can handle complex high-dimensional data without a prior dimension reduction. Provide "amortized" inference.

So What's Missing in the LFI-ML Literature?

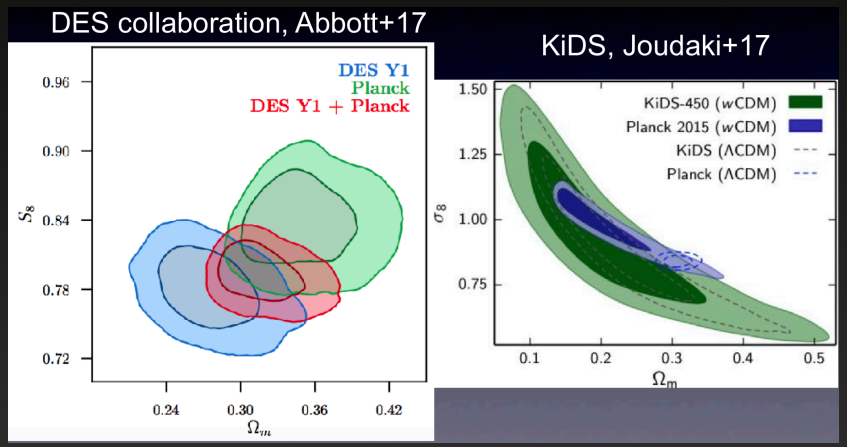
Given observed data, we would like to constrain parameters of interest using assumed theoretical/simulation model. Valid measures of uncertainty, no matter the value of the unknown parameter.



$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\middle|\theta\right) = 1-\alpha, \quad \forall\theta \in \Theta$$

- Shortage of practical inferential and diagnostic tools with finite-sample guarantees of conditional coverage.

Open Problems in LFI



Confidence sets with correct conditional coverage (for small n)?

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\middle|\theta\right) = 1-\alpha, \quad \forall\theta \in \Theta$$

- Most approaches that estimate likelihoods or likelihood ratios
  - rely on asymptotic assumptions (Wilks 1938) for downstream inference
  - do not assess validity across entire parameter space, or
  - use costly MC simulations at fixed parameter settings on a grid

Unified Inference Machinery for Frequentist LFI

- Bridges ML with classical statistics to provide:

  (i) valid inference: confidence sets and tests with finite-sample guarantees (Type I error control and power)

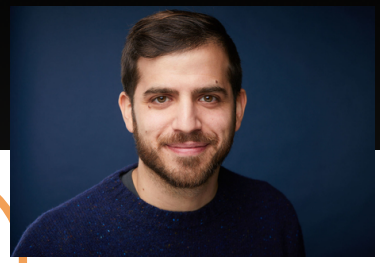
  (ii) practical diagnostics: check actual coverage across entire parameter space

- Goal: Modular and computationally efficient procedures
  - Can leverage generative, predictive and posterior algorithms
  - Compatible with any test statistic and prior

https://github.com/lee-group-cmu/lf2i

https://arxiv.org/abs/2002.10399 (ICML 2021)

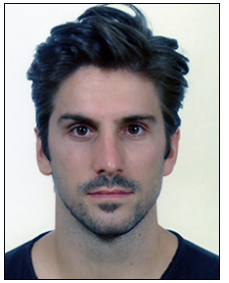
https://arxiv.org/abs/2205.15680 (AISTATS 2023)



LF2I

Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage

https://arxiv.org/abs/2107.03920



Niccolò Dalmasso[*†] NICCOLO.DALMASSO@GMAIL.COM

Luca Masserano[*‡] LMASSERA@ANDREW.CMU.EDU

David Zhao[†] DAVIDZHAO@CMU.EDU

Rafael Izbicki[§] RAFAELIZBICKI@GMAIL.COM

Ann B. Lee [†¶] ANNLEE@CMU.EDU

Abstract

Many areas of science make extensive use of computer simulators that implicitly encode likelihood functions of complex systems. Classical statistical methods are poorly suited for these so-called likelihood-free inference (LFI) settings, particularly outside asymptotic and low-dimensional regimes. Although new machine learning methods, such as normalizing flows, have revolutionized the sample efficiency and capacity of LFI methods, it remains an open question whether they produce confidence sets with correct conditional coverage for small sample sizes. This paper unifies classical statistics with modern machine learning to present (i) a practical procedure for the Neyman construction of confidence sets with finite-sample guarantees of nominal coverage, and (ii) diagnostics that estimate conditional coverage over the entire parameter space. We refer to our framework as *likelihood-free frequentist inference* (LF2I). Any method that defines a test statistic, like the likelihood ratio, can leverage the LF2I machinery to create valid confidence sets and diagnostics without costly Monte Carlo samples at fixed parameter settings. We study the power of two test statistics (`ACORE` and `BFF`), which, respectively, maximize versus integrate an odds function over the parameter space. Our paper discusses the benefits and challenges of LF2I, with a

Equivalence of Tests and Confidence Sets

- Data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\} \sim F_\theta$
- Test statistic $\lambda(\mathcal{D}; \theta)$
- Critical values

$$\text{Reject } H_0 : \theta = \theta_0 \iff \lambda(\mathcal{D}; \theta_0) < C_{\theta_0, \alpha}$$

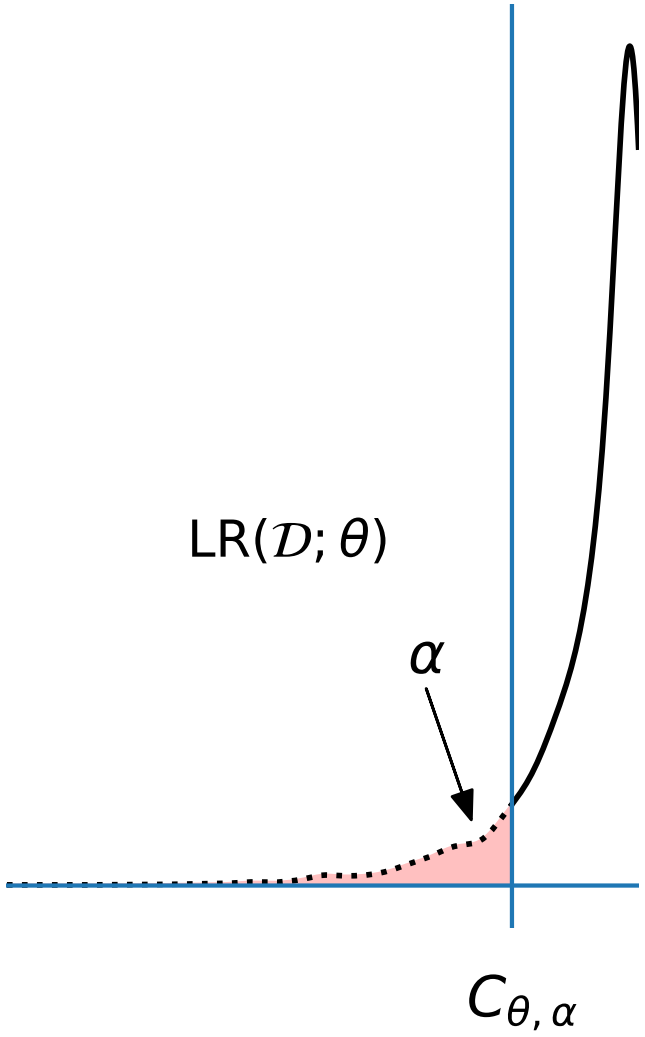
Theorem (Neyman 1937)

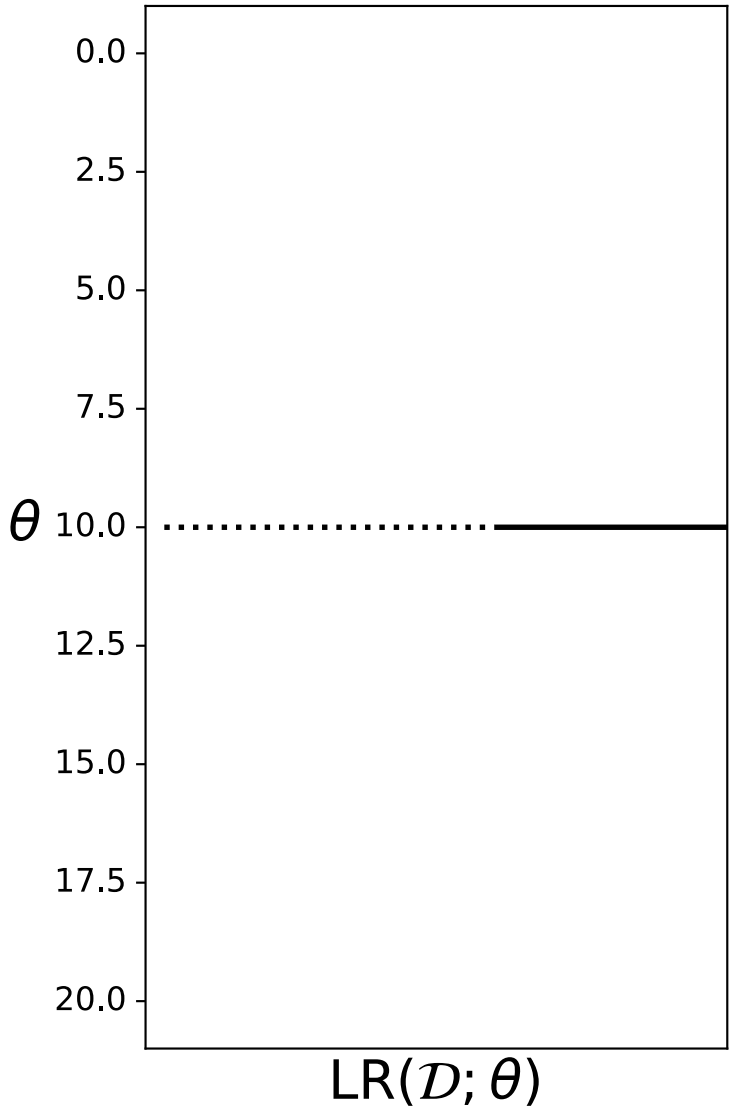
Constructing a $1 - \alpha$ *confidence set for* θ *is equivalent to testing*

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

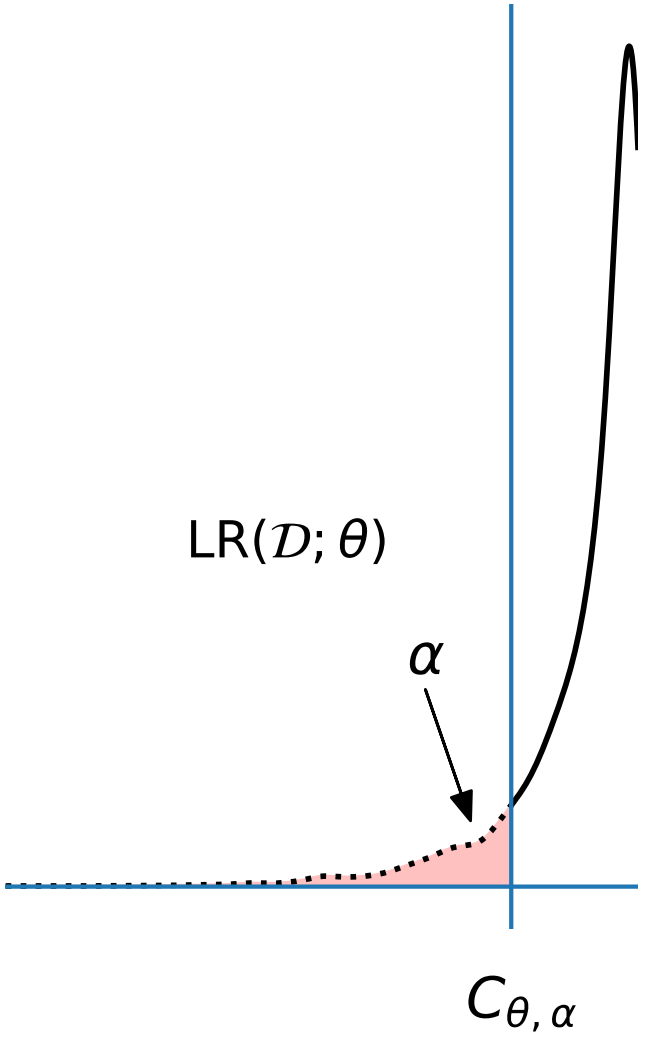
for every $\theta_0 \in \Theta$.

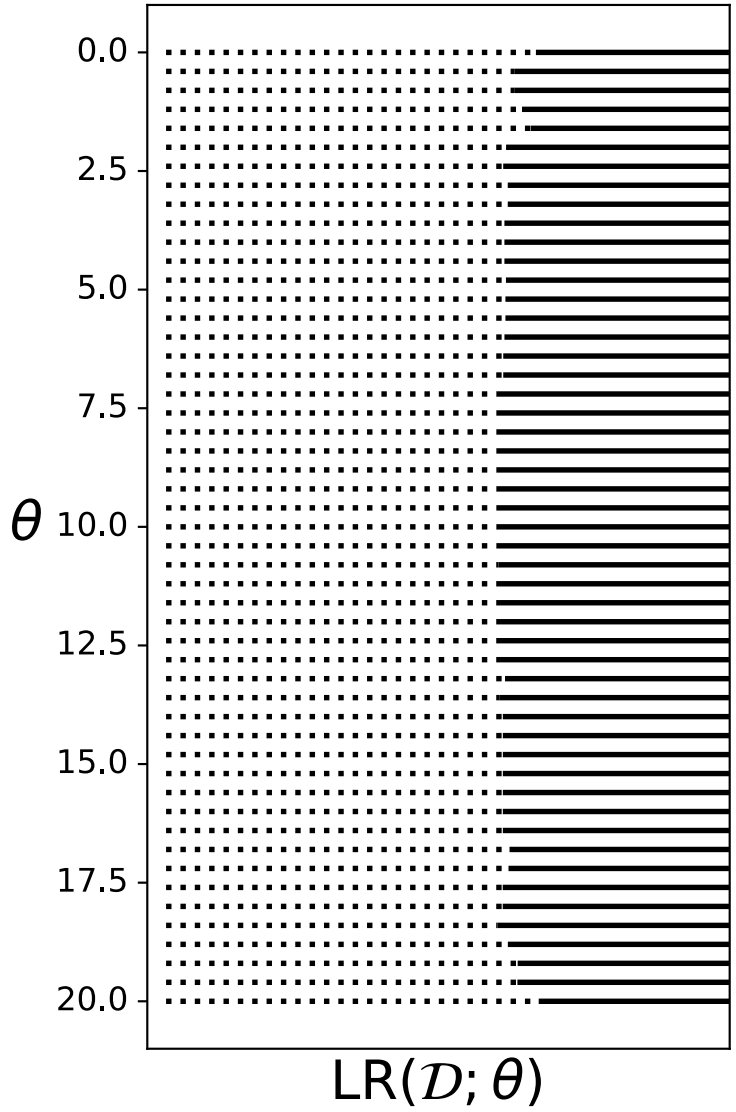
1. Fixed θ. Find the rejection region for test statistic λ.



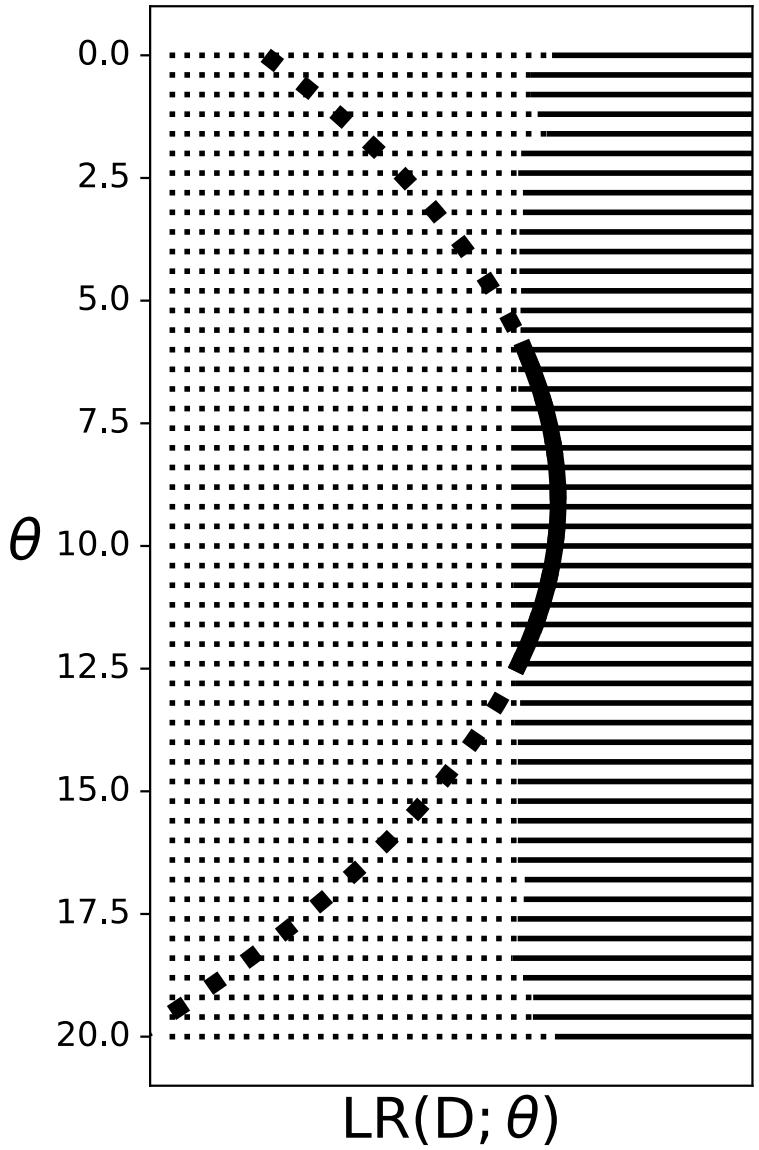


2. Repeat for every θ in parameter space.

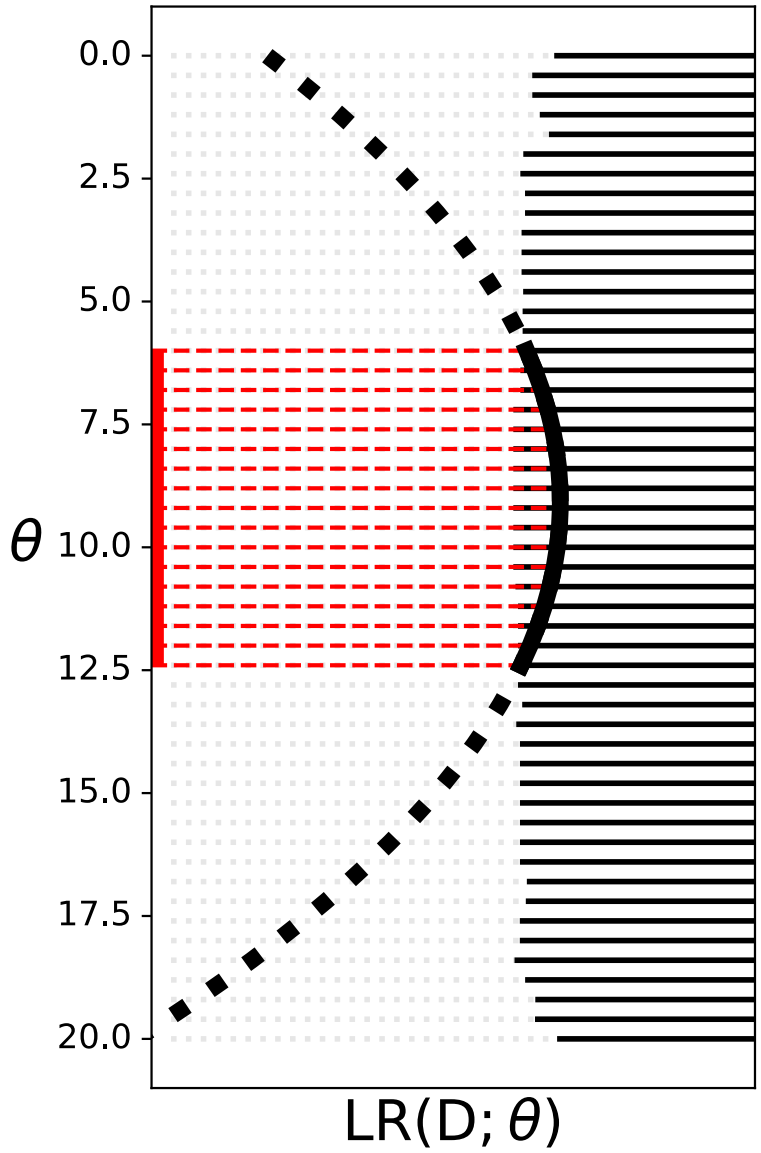




3. Observe data $\mathcal{D} = \mathrm{D}$. Evaluate $\lambda(\mathrm{D}; \theta)$.



4. Construct $(1 - \alpha)$ confidence set for θ.



Challenges

- Neyman construction itself. L. Lyons, "*Open Statistical Issues in Particle Physics*", AOAS 2008:

However, in practice, it is very hard to use the Neyman frequentist construction when more than two or three parameters are involved: software to perform a Neyman construction efficiently in several dimensions would be most welcome. The

- Validation of frequentist coverage. R. Cousins: "*Lectures on Statistics in Theory: Prelude to Statistics in Practice*", arXiv:1807.05996, 2018:

A complete, rigorous check of coverage considers a fine multi-D grid of *all* parameters, and *for each multi-D point in the grid*, generates an ensemble of toy MC pseudo-experiments, runs the full analysis procedure, and finds the fraction of intervals covering the μ_t of interest that was used for that ensemble. I.e., one calculates $P(\mu_t \in [\mu_1, \mu_2])$, and compares to C.L.

But... the ideal of a fine grid is usually impractical.

How Do we Turn the Neyman Construction and Validation into Practical Procedures?

The Neyman construction requires one to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

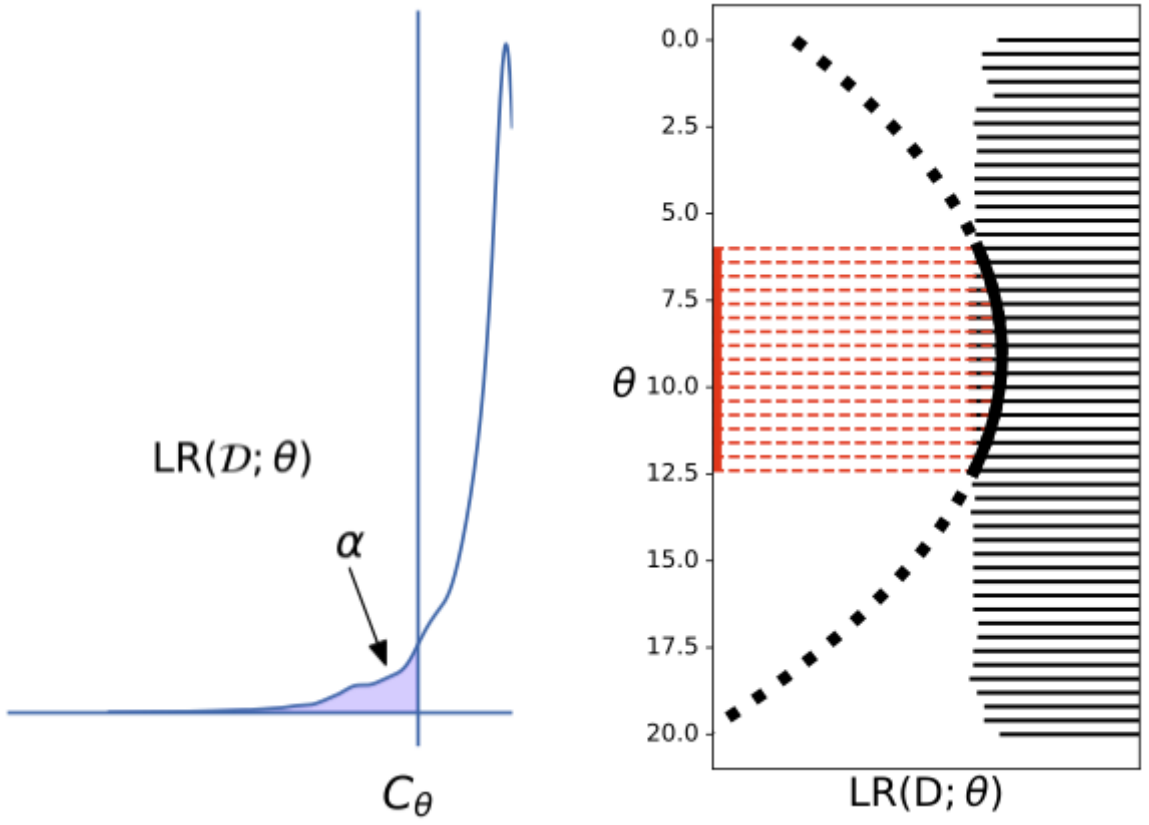
for **every** $\theta_0 \in \Theta$.

Key insight:

1. Test statistic $\lambda(\mathcal{D}; \theta)$
2. Critical values $C_{\theta_0,\alpha}$ or p-values $p(D; \theta_0)$ of the test
3. Coverage $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\right)$ of the constructed confidence set

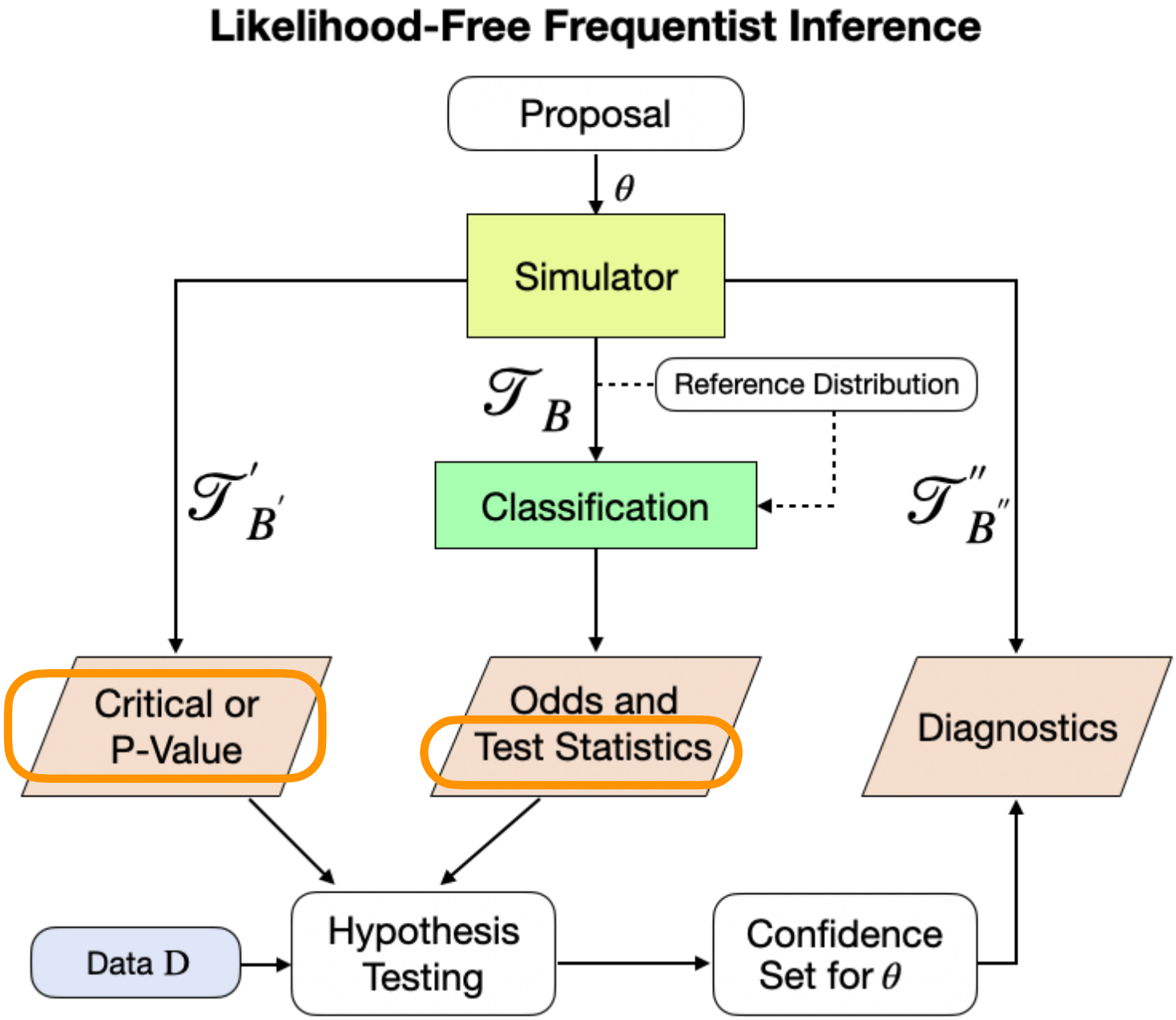
are **conditional distribution functions** of the (unknown) parameters, and often vary smoothly across the parameter space Θ.

Efficient Construction of Finite-Sample Confidence Sets



Rather than running a batch of Monte Carlo simulations for every null hypothesis $\theta = \theta_0$ on, e.g., a fine enough grid in Θ, we can interpolate across the parameter space using training-based ML algorithms.

Our Inference Machinery



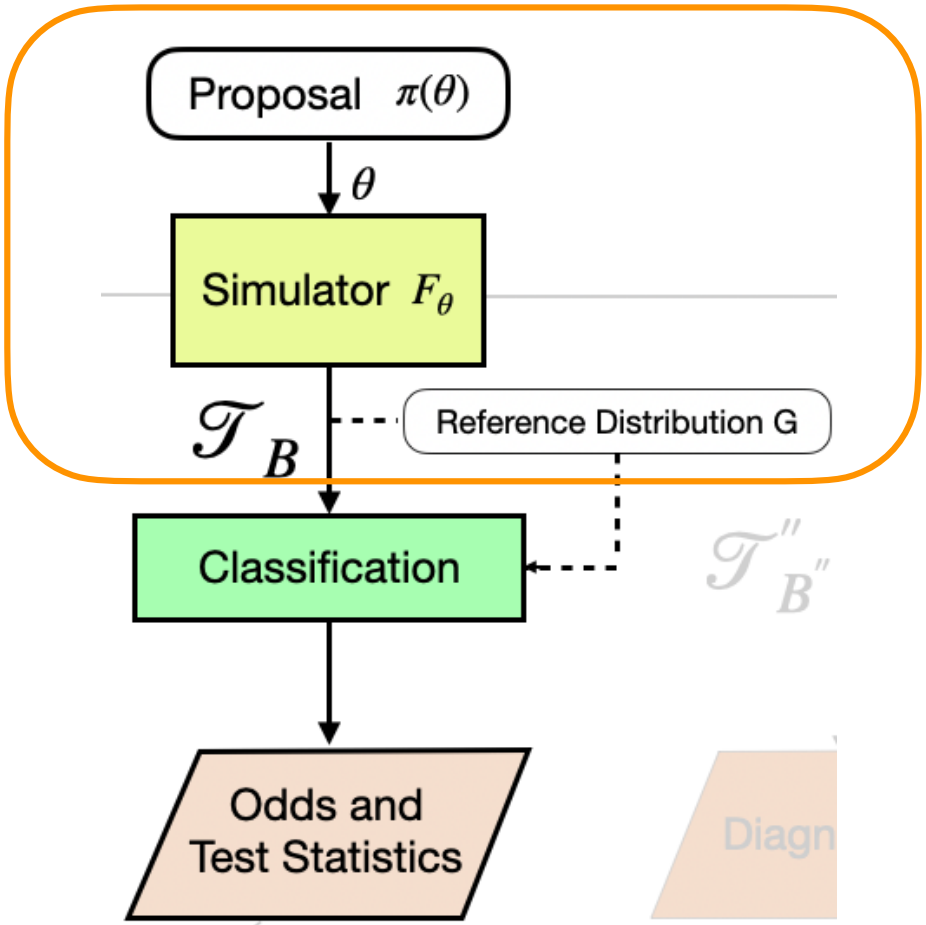
Test Statistics: Leverage ML Classification/ Prediction Algorithms

- Examples of LF2I test statistics:
  - classification/odds → **ACORE** (approximate LRT) [Dalmasso et al 2020; arXiv:2002.10399]
  - classification/odds → **BFF** (approximate Bayes Factor) [Dalmasso et al 2021; arXiv:2107.03920]
  - prediction or posterior estimation → **WALDO** (modified Wald test statistic) [Masserano et al 2022; arXiv:2205.15680]

Center Branch: Estimating Odds and Test Statistic

Parameter: $\theta \in \Theta$

Simulated data: $\mathbf{X},\ \mathbf{x} \in \mathcal{X}$. *Observed data*: $\mathbf{X}^{\text{obs}}, \mathbf{x}^{\text{obs}} \in \mathcal{X}$.



1. Proposal distribution $\pi(\theta)$ over the parameter space Θ
2. Forward simulator F_θ
   - $F_{\theta_1} \neq F_{\theta_2}$ for $\theta_1 \neq \theta_2 \in \Theta$
3. Reference distribution G over the feature space $\mathcal{X}$
   - $F_\theta \ll G$ for all $\theta \in \Theta$
4. A simulated sample of size B to estimate odds and test statistic

Estimate Odds via Probabilistic Classification

Simulate two samples:

- $\{(\theta_k, \mathbf{X}_k, Y_k = 1)\}_{k=1}^{B/2}$, where $\theta \sim \pi(\theta)$, $\mathbf{X} \sim F_\theta$
- $\{(\theta_l, \mathbf{X}_l, Y_l = 0)\}_{l=1}^{B/2}$ where $\theta \sim \pi(\theta)$, $\mathbf{X} \sim G$

Probabilistic classifier r:

$$r : (\theta, \mathbf{X}) \longrightarrow \mathbb{P}(Y = 1 | \mathbf{X}, \theta)$$

Define the **odds** at $\theta \in \Theta$ and fixed $\mathbf{x} \in \mathcal{X}$ as

$$\mathbb{O}(\mathbf{x}; \theta) := \frac{\mathbb{P}(Y = 1 | \mathbf{x}, \theta)}{\mathbb{P}(Y = 0 | \mathbf{x}, \theta)} = \frac{f_\theta(\mathbf{x})}{g(\mathbf{x})}$$

Interpretation: Chance that $\mathbf{x}$ was generated from F_θ rather than G.

Test Statistics Based on Odds: ACORE and BFF

Suppose we want to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1, \text{ where } \Theta_1 = \Theta_0^c$$

For observed data $\mathrm{D} = \{\mathbf{X}_1^{\text{obs}}, ..., \mathbf{X}_n^{\text{obs}}\}$, we define:

- ACORE (Approximate Computation via Odds Ratio Estimation):

$$\widehat{\Lambda}(\mathcal{D}; \Theta_0) := \log \frac{\sup_{\theta \in \Theta_0} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta)}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta)}$$

- BFF (Bayesian Frequentist Factor):

$$\widehat{\tau}(\mathcal{D}; \Theta_0) := \frac{\int_{\Theta_0} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta) \, d\pi_0(\theta)}{\int_{\Theta_0^c} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta) \, d\pi_1(\theta)}.$$

where π_0 and π_1 are the restrictions of a proposal distribution π_τ over Θ to Θ_0 and Θ_0^c, respectively.

ACORE and BFF are Approximations of the LR Statistic and the Bayes Factor respectively!

Lemma (Fisher's Consistency)

If $\quad \widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{X}) = \mathbb{P}(Y = 1|\theta, \mathbf{x}) \, \forall \theta, \mathbf{X}$

1. $\implies \widehat{\Lambda}(\mathcal{D}; \Theta_0) = \mathrm{LR}(\mathcal{D}; \Theta_0) \equiv \log \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathcal{D};\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathcal{D};\theta)},$

2. $\implies \widehat{\tau}(\mathcal{D}; \Theta_0) = \mathrm{BF}(\mathcal{D}; \Theta_0) \equiv \frac{\mathbb{P}(\mathcal{D}|H_0)}{\mathbb{P}(\mathcal{D}|H_1)} = \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D};\theta) d\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D};\theta) d\pi_1(\theta)}.$

Note: The Bayes factor is often used as a Bayesian alternative to significance testing but here we are treating it as a frequentist test statistic.

Test Statistics Based on Odds: ACORE and BFF

Suppose we want to test:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

For observed data $\mathrm{D} = \{\mathbf{X}_1^{\text{obs}}, ..., \mathbf{X}_n^{\text{obs}}\}$, we define

- ACORE (Approximate Computation via Odds Ratio Estimation):

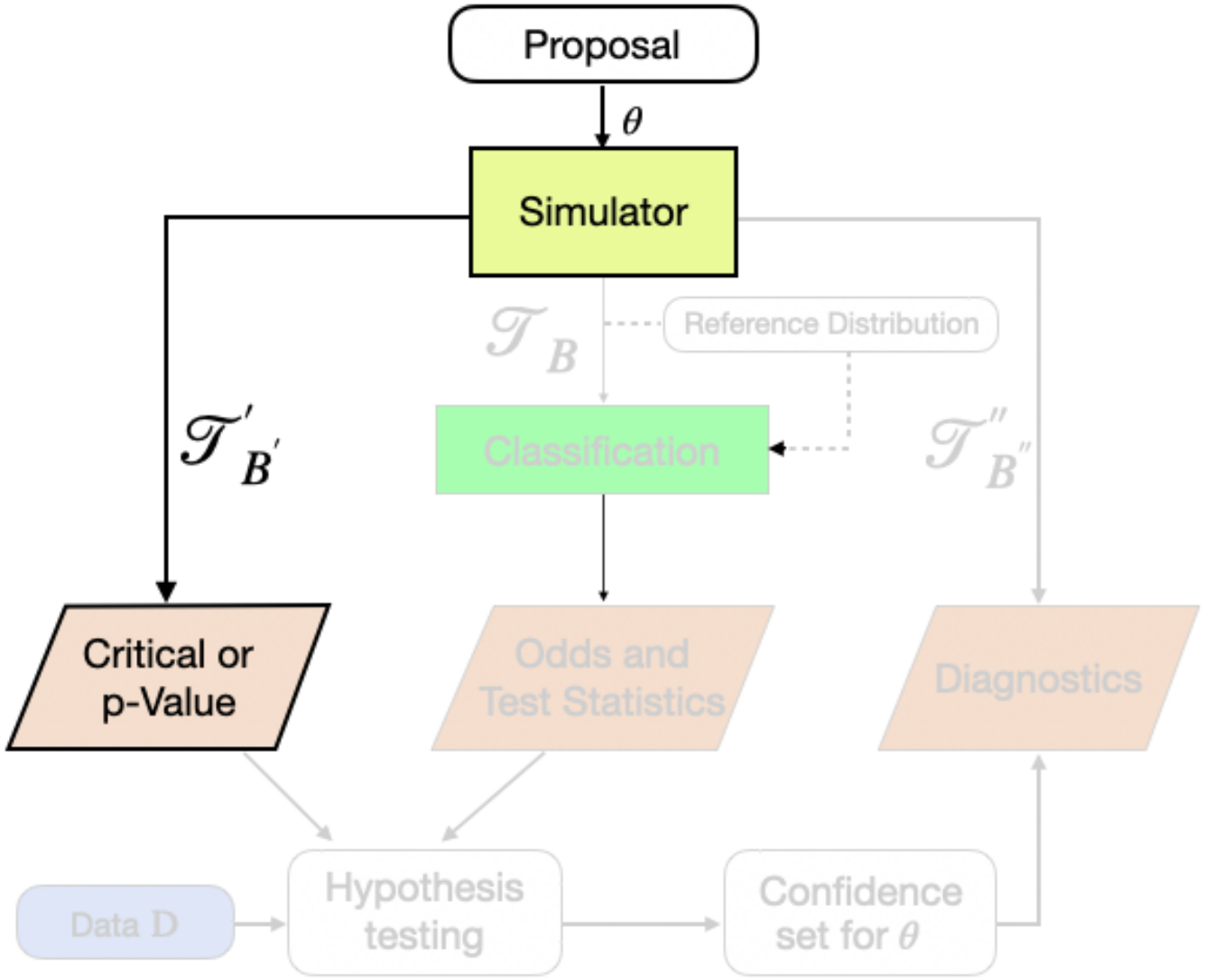
$$\widehat{\Lambda}(\mathcal{D}; \theta_0) := \log \frac{\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}$$

- BFF (Bayesian Frequentist Factor):

$$\widehat{\tau}(\mathcal{D}; \theta_0) := \frac{\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\int_\Theta \left(\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta) \right) d\pi_\tau(\theta)}.$$

where $\pi_\tau(\theta)$ is a probability distribution over the parameter space.

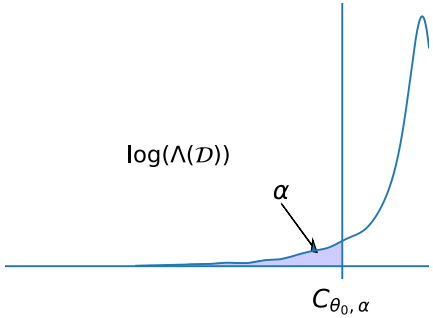
Left Branch: Estimate Critical Values or P-Values



We use B' simulations to estimate critical values.

Estimating Critical Values $C_{\theta_0,\alpha}$

To control Type I error at level α:
Reject $H_0 : \theta = \theta_0$ when $\lambda(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$, where

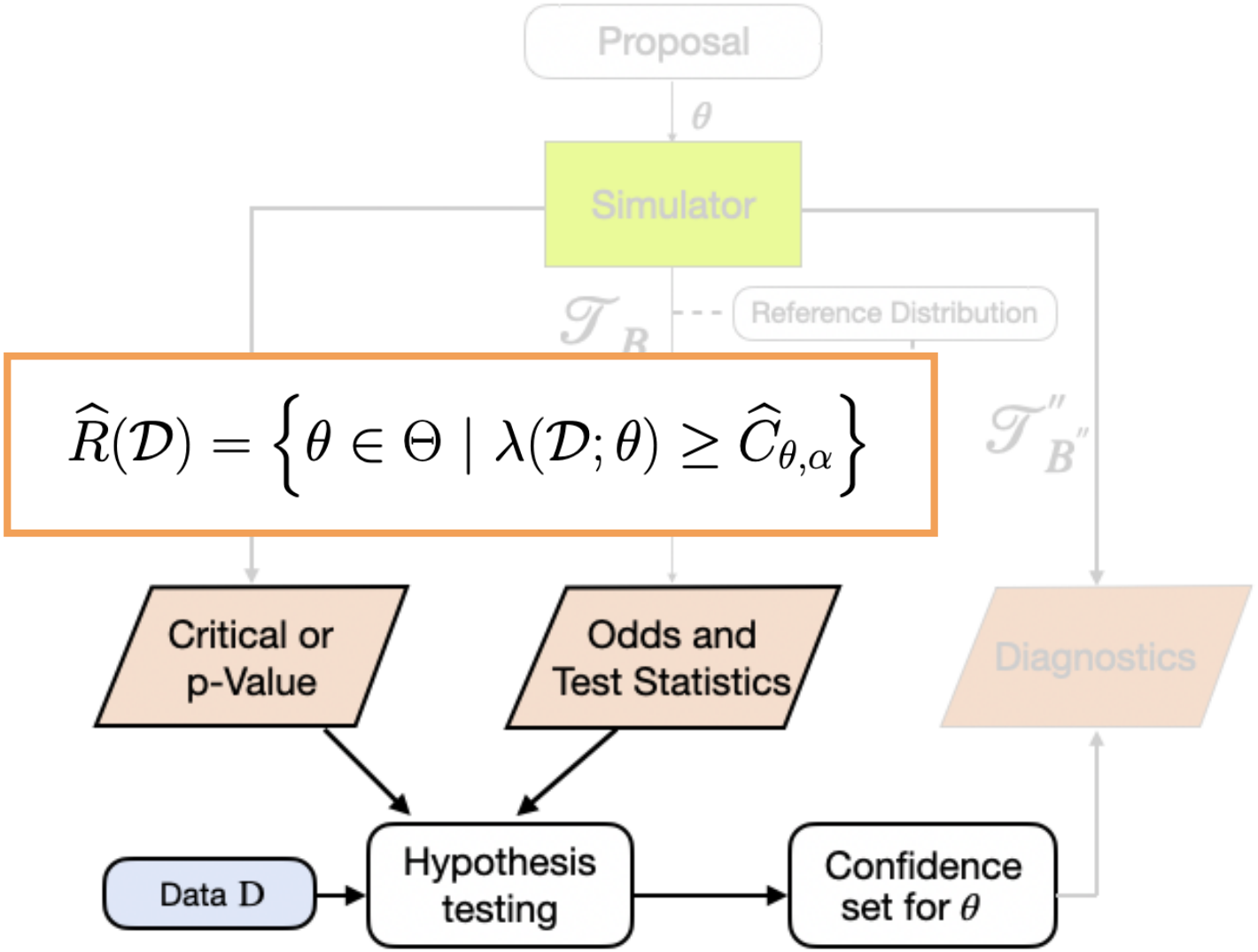


$$C_{\theta_0,\alpha} = \arg\sup_{C \in \mathbb{R}} \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0} \left(\lambda(\mathcal{D}; \theta_0) < C \right) \leq \alpha \right\}.$$

Problem: Need to compute $\mathbb{P}_{\mathcal{D}|\theta} \left(\lambda(\mathcal{D}; \theta) < C \right)$ for every $\theta \in \Theta$.

Solution: $F_{\lambda|\theta}(C \mid \theta) \equiv \mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \theta) < C \mid \theta)$ is a conditional CDF, so we can estimate its α-quantile via quantile regression $F^{-1}_{\lambda|\theta}(\alpha|\theta)$.

Construct Confidence Set via Neyman Inversion



Are the Constructed Confidence Sets Valid?

Theorem (Validity for any test statistic)

Let $C_{B'}$ be the critical value of a level-α test based on the statistic $\lambda(\mathcal{D}; \theta_0)$. Then, if the quantile regression estimator is consistent,

$$C_{B'} \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} C^*,$$

where C^* is such that

$$\mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \theta_0)) \leq C^*) = \alpha.$$

If B' is large enough, we can construct a confidence set with guaranteed nominal coverage regardless of the observed sample size n.

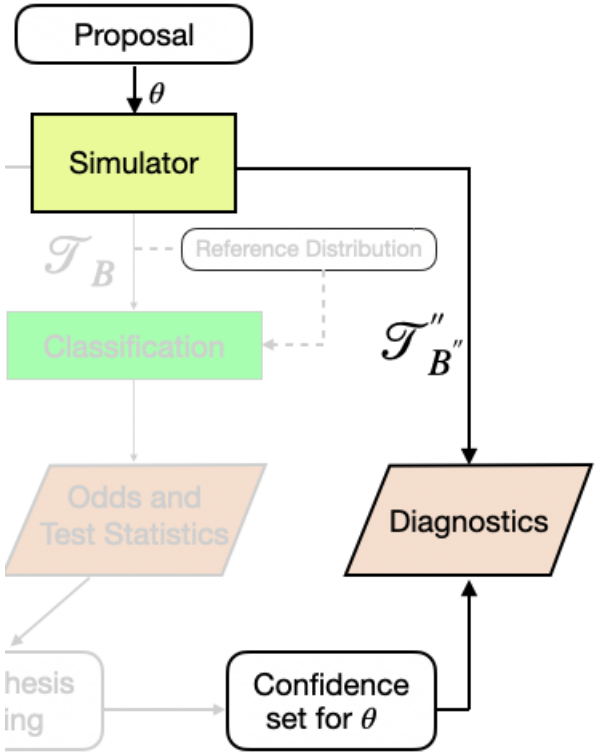
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

How do we check coverage of constructed confidence sets across Θ?

Note:

$$\widehat{R}(\mathcal{D}) = \left\{\theta \in \Theta \mid \lambda(\mathcal{D};\theta) \geq \widehat{C}_{\theta,\alpha}\right\}$$

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D}) \mid \theta\right) = \mathbb{E}_{\mathcal{D}|\theta}\left[\mathbb{I}\left(\theta \in \widehat{R}(\mathcal{D})\right) \mid \theta\right]$$

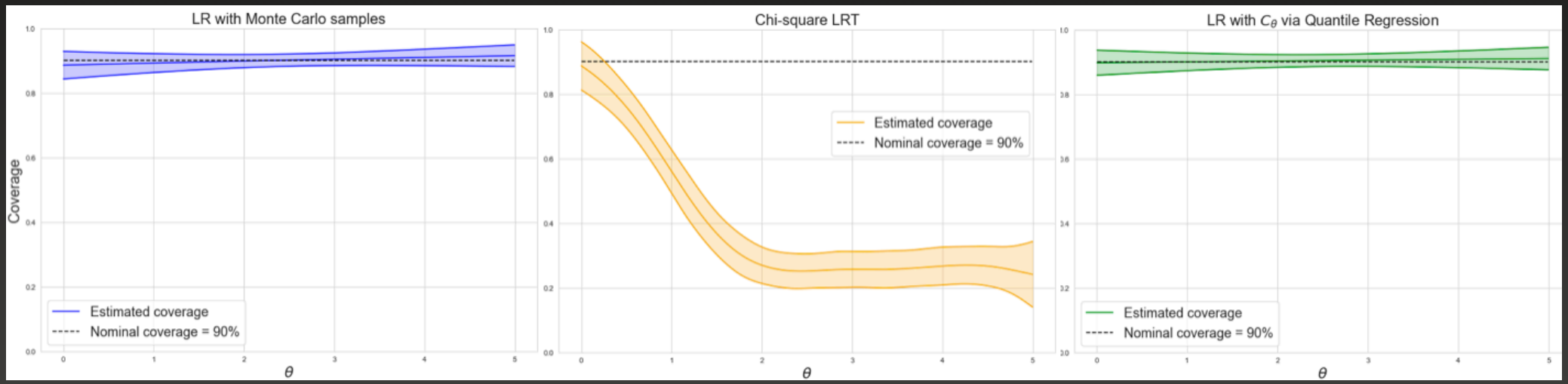


1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $\widehat{R}(\mathcal{D}_i)$
3. For $\{\theta_i, \widehat{R}(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i := \mathbb{I}(\theta_i \in \widehat{R}(\mathcal{D}_i))$ on θ_i.

How close is the actual coverage to the nominal confidence level $1-\alpha$?

Ex: Estimate Critical Values (GMM; n=1000) & Run Diagnostics Across the Parameter Space

$$X_1, \ldots, X_n \sim 0.5N(\theta, 1) + 0.5N(-\theta, 1)$$



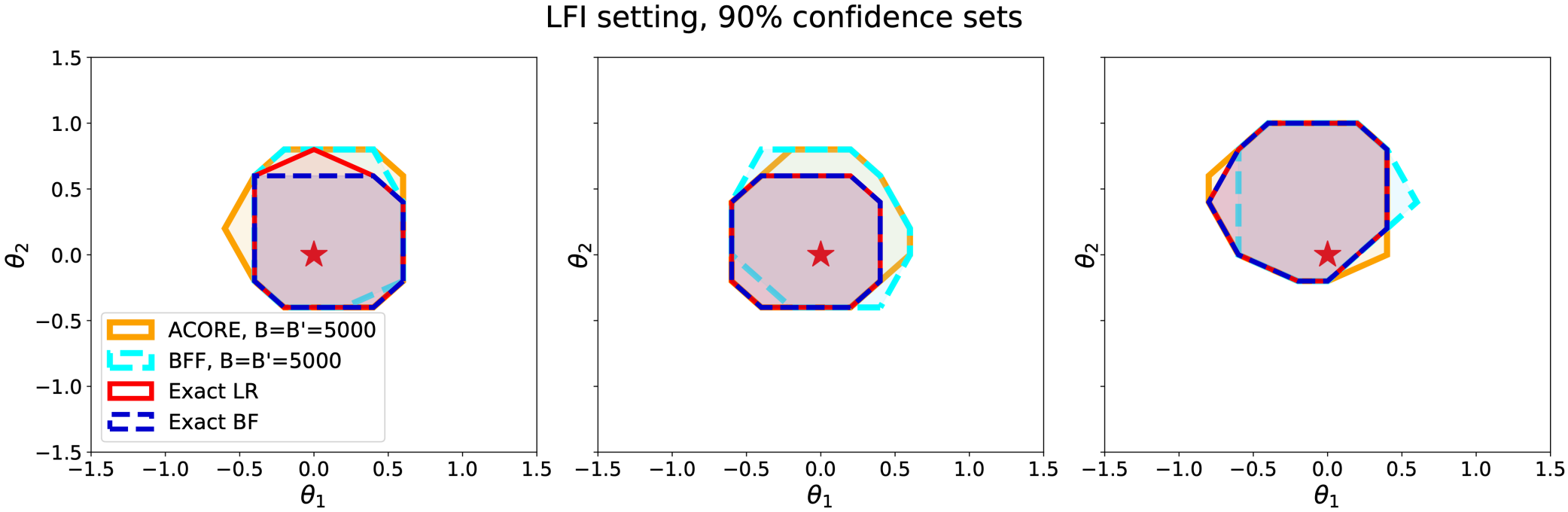
(Left) LR with 1000 MC simulations at each θ on a fine grid

(Center) Assume chi-squared distribution of LR statistic

(Right) LR with quantile regression with B'=1000 simulations total

Ex: Construct Confidence Sets (MVG data)

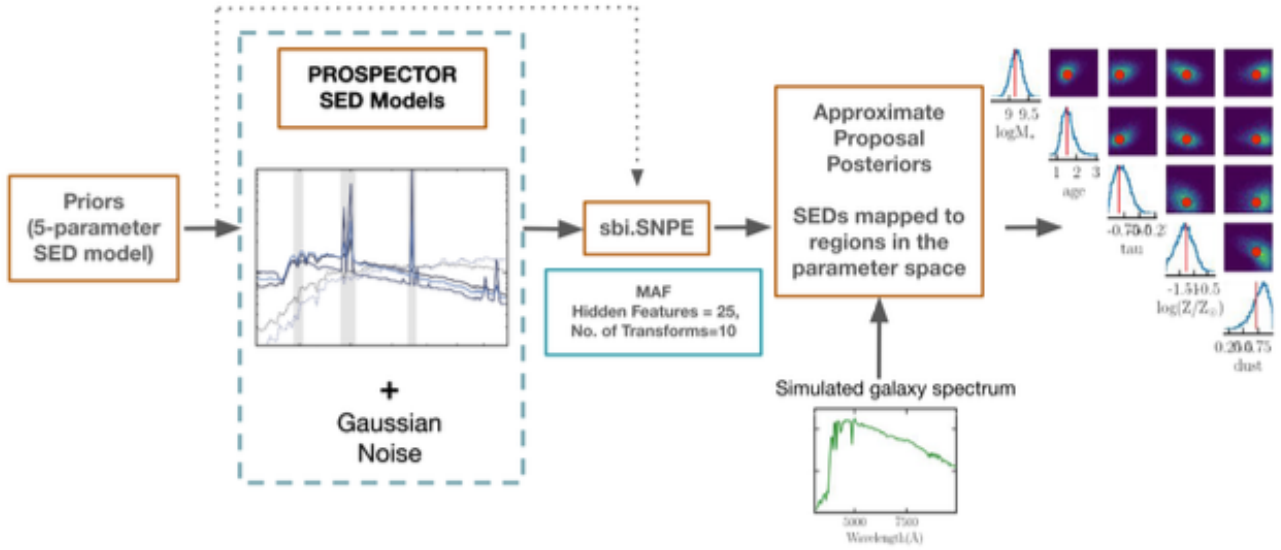
$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N(\boldsymbol{\theta}, \mathbf{I}_d), \text{ where } n = 10, \ \boldsymbol{\theta} = \mathbf{0}$$



When d=2, ACORE and BFF confidence sets (for B=B'=5000) are similar in size to the Exact LR confidence sets.

LF2I scales well for <10 parameters

Astronomy: Infer galaxy parameters from SEDs via NPE



Why? Advent of billion-galaxy surveys with complex data needs efficient modeling of spectral energy distributions (SEDs) with robust uncertainty quantification

How? Combine SBI and NPE to infer galaxy parameters (5-parameter model)

Goal: use Waldo to obtain reliable constraints and check their validity against those obtained via NPE

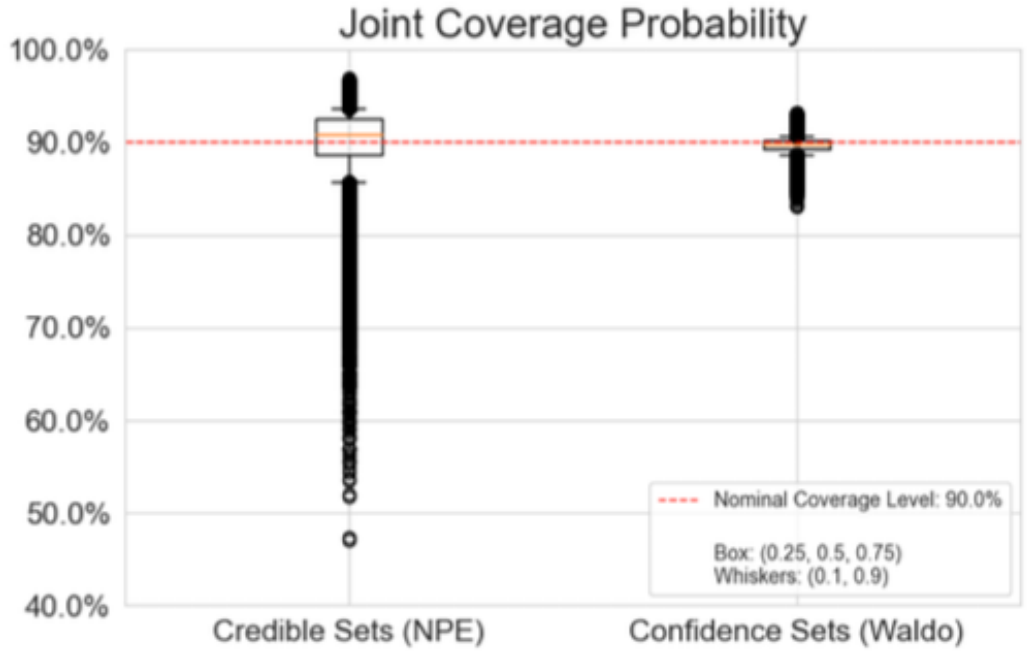
Image taken from Khullar et al. (2022)

13

$$\tau^{\text{WALDO}}(\mathcal{D};\boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

Coverage across the entire parameter space

$$r(\theta) := \mathbb{P}(\theta \in \mathscr{R}(\mathscr{D}) \mid \theta), \quad \theta \in \mathbb{R}^5$$

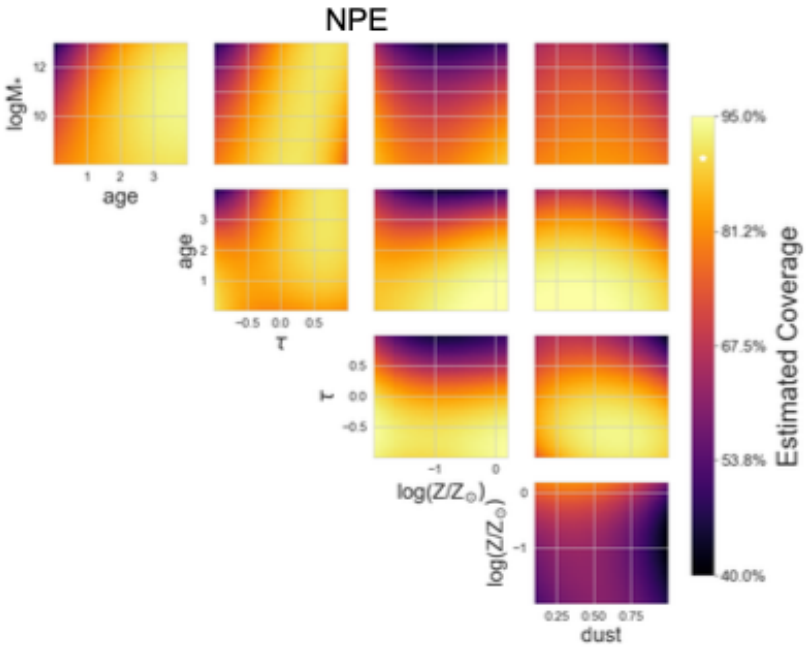


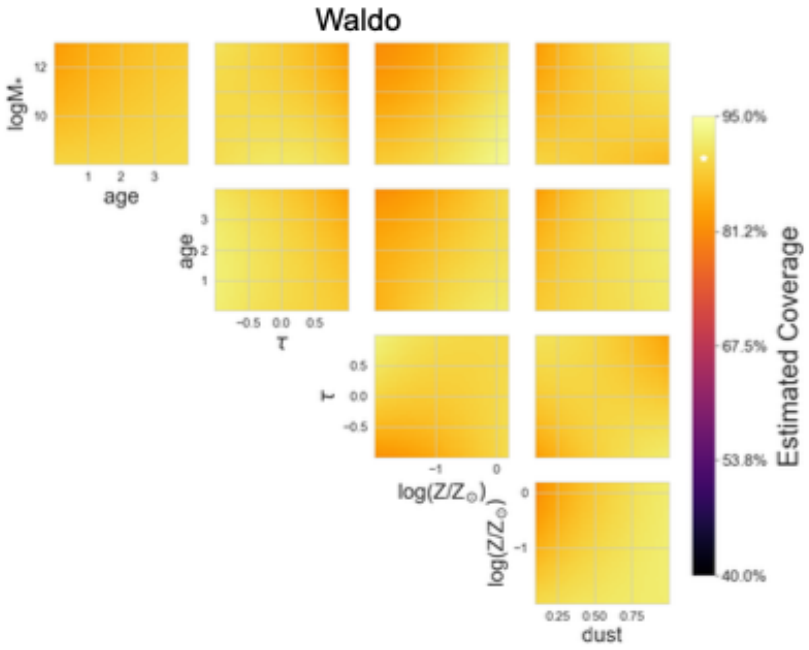
15

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

Profiled dependence of coverage probability vs parameters

$$\min_{\theta_i \theta_j \theta_k} r(\theta), \quad \theta \in \mathbb{R}^5$$





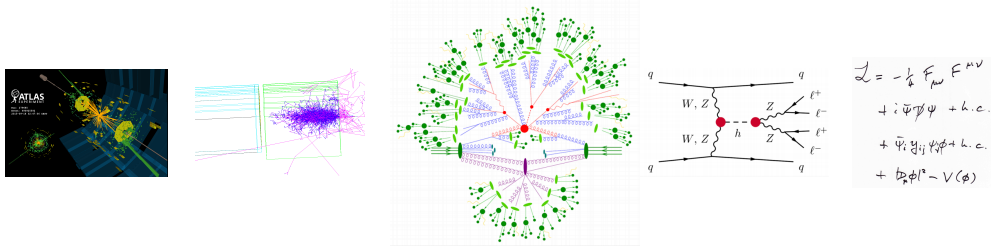
17

LF2I scales well for <10 parameters. However…

The parameters θ

One more issue: the "theory" space is not the only thing effecting the data

- **every step of the forward process comes with its own parameters**
  (we understand the process generally but need additional knobs to model the data)



$$p(x|\theta) = \int \mathrm{d}z_d \mathrm{d}z_h \mathrm{d}z_p \quad p(x|z_d, \theta_x)\; p(z_d|z_h, \theta_d)\; p(z_h|z_p, \theta_h)\; p(z_p|\theta_p, \theta_{\mathrm{th}})$$

$p(x|z_d)$ $p(z_d|z_h)$ $p(z_h|z_p)$ $p(z_p|\theta)$

nuisance parameters

core "theory" parameters of inferest (e.g. "Higgs Mass"

12

Credit: Lukas Heinrich

How do we Handle Nuisance Parameters?

In many applications, the parameter space can be decomposed as $\Theta = \mathcal{M} \times \mathcal{N}$, where $\mathcal{M}$ contains the *main parameters* μ of interest, and $\mathcal{N}$ contains the *nuisance parameters* ν not of immediate interest.

Suppose we want to test

$$H_{0,\mu_0} : \mu = \mu_0 \quad \text{versus} \quad H_{1,\mu_0} : \mu \neq \mu_0 \quad \text{for } \mu_0 \in \mathcal{M}$$

How does one solve this problem within our inference machinery?

Nuisance-Parameterized LF2I

Test composite vs composite hypotheses:

$$H_{0,\mu_0} : \theta \in \Theta_0 \quad \text{vs} \quad H_{1,\mu_0} : \theta \in \Theta_1,$$

where $\Theta_0 = \{(\mu_0, \nu) \mid \nu \in \mathcal{N}\}$, and $\Theta_1 = \Theta_0^c$.

- `ACORE` test statistic (by maximizing estimated odds)

$$\widehat{\Lambda}(\mathcal{D}; \mu_0) := \log \frac{\sup_{\nu \in \mathcal{N}} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; (\mu_0, \nu))}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta)}$$

- `BFF` test statistic (by integrating estimated odds)

$$\widehat{\tau}(\mathcal{D}; \mu_0) := \frac{\int_{\mathcal{N}} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; (\mu_0, \nu)) \; d\pi_0(\nu)}{\int_{\Theta_1} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta) \; d\pi_1(\theta)}.$$

where $\pi_0(\nu)$ is a distribution over $\mathcal{N}$, the nuisance parameter space.

But Critical Value Estimation is Difficult with Many NPs

Remember: To guarantee frequentist coverage by Neyman's inversion technique, we need to test null hypotheses

$$H_{0,\mu_0} : \mu = \mu_0 \quad \text{versus} \quad H_{1,\mu_0} : \mu \neq \mu_0 \quad \text{for } \mu_0 \in \mathcal{M}$$

by comparing test statistics to the cutoffs $\widehat{C}_{\mu_0} := \inf_{\nu \in \mathcal{N}} \widehat{C}_{(\mu_0,\nu)}$.

That is, one needs to control the type I error at each μ_0 for *all* possible values of the nuisance parameters.

Can lead to numerically unwieldy and costly computations if the number of nuisance parameters is large (>10 NPs).

Hybrid Approaches to Critical Value Estimation

- h-ACORE: Hybrid Resampling or Profiling[1] of Nuisance Parameters
  - Compare ACORE test statistic with the *hybrid cut-off*

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\Lambda(\mathcal{D};\mu_0)\big|(\mu_0,\widehat{\nu}_{\mu_0})}\left(\alpha \,\middle|\, \mu_0, \widehat{\nu}_{\mu_0}\right)$$

  where the quantile regression is based on a train sample $\mathcal{T}'$ generated at *fixed* $\widehat{\nu}_{\mu_0}$.
- h-BFF: Integration of Nuisance Parameters
  - Compare BFF test statistic with the *approximate cut-off*

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\tau(\mathcal{D};\mu_0)\big|\mu_0}\left(\alpha \,\middle|\, \mu_0\right)$$

  where we draw the train sample $\mathcal{T}'$ from the entire parameter space $\Theta = \mathcal{M} \times \mathcal{N}$, but apply quantile regression *using* μ *only*

[1] Van der Vaart, 2000; Chuang & Lai, 2000; Feldman, 2000; Sen et al. 2009

Hybrid Methods and Confidence Sets

- Hybrid methods (which maximize or average over nuisance parameters) do not always control the type I error of statistical tests.
- "*For small sample sizes, there is no theorem as to whether profiling or marginalization will give better frequentist coverage for the parameter of interest*" (Cousins 2018)
- Can our diagnostic tools provide guidance as to which method to choose for the problem at hand?

Poisson Counting Experiment

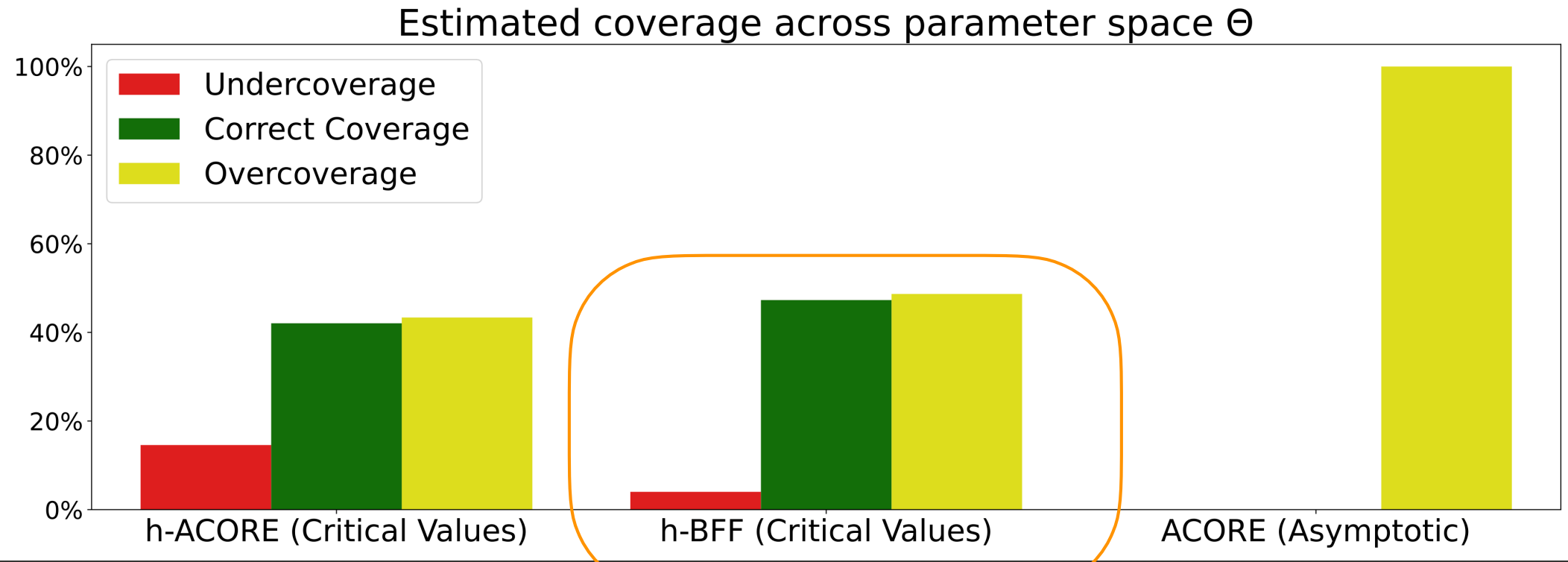
[cf., Lyons, 2008; Cowan et al, 2011; Cowan, 2012]

- Particle collision events counted under the presence of a background process.

$$\text{Observed data } \mathcal{D} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{10})$$
$$\mathbf{X} = (N_b, N_s), \text{ where } N_b \sim \text{Pois}(\gamma b),\ N_s \sim \text{Pois}(b + \varepsilon s)$$

- The observed data D consist of n=10 observations of X=(N_B, N_S), where
  - N_B is the # of events in the background region (assume γ=1)
  - N_S is the # of events in the signal region
- Unknown parameters:
  - signal strength (s); two nuisance parameters (b and ϵ)

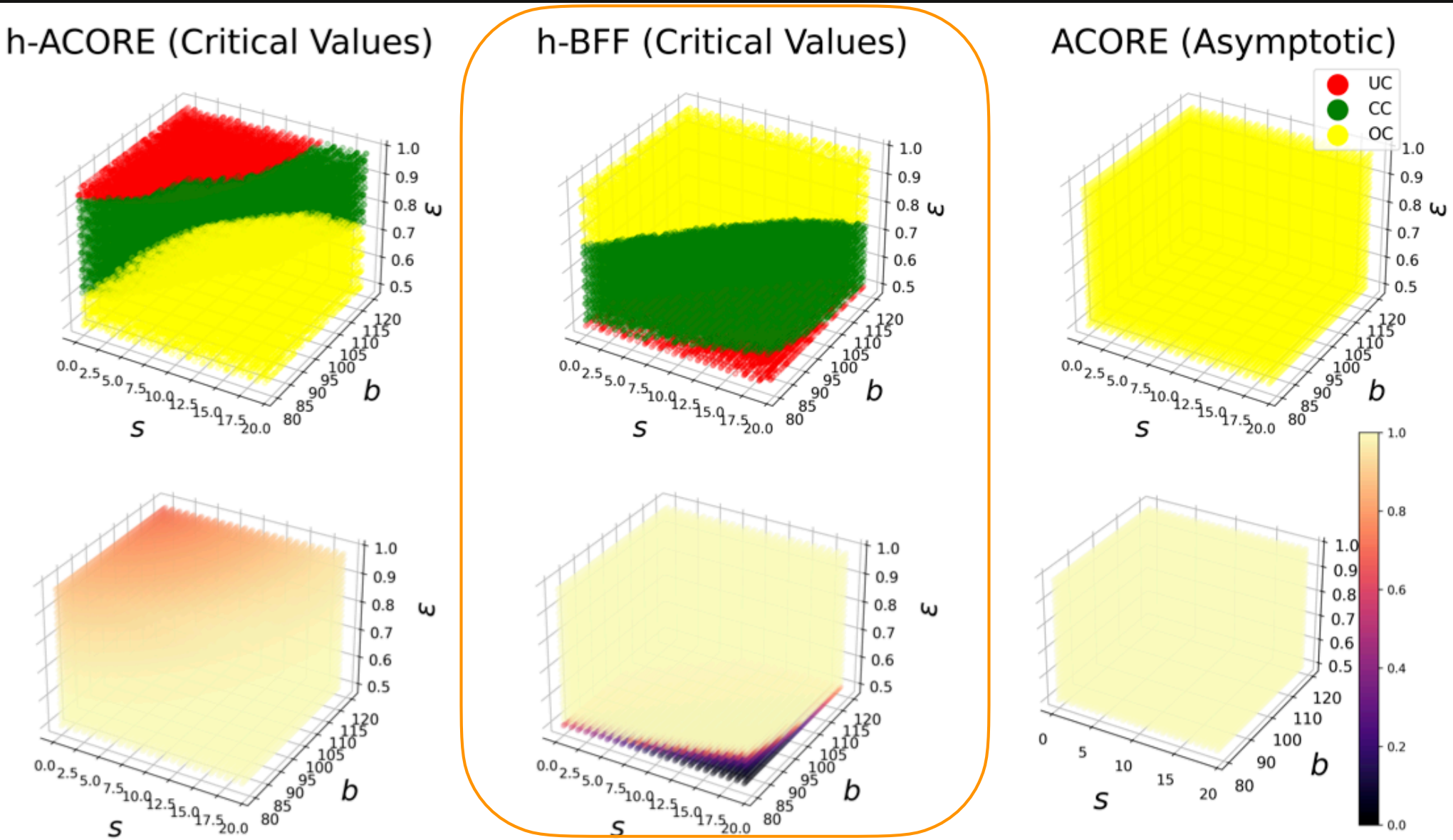
Diagnostics to Check Coverage Across the Entire Parameter Space



- h-BFF (averages over nuisance parameters) performs the best in terms of having the largest proportion of the parameter space with CC and only a small fraction of the parameter space with UC

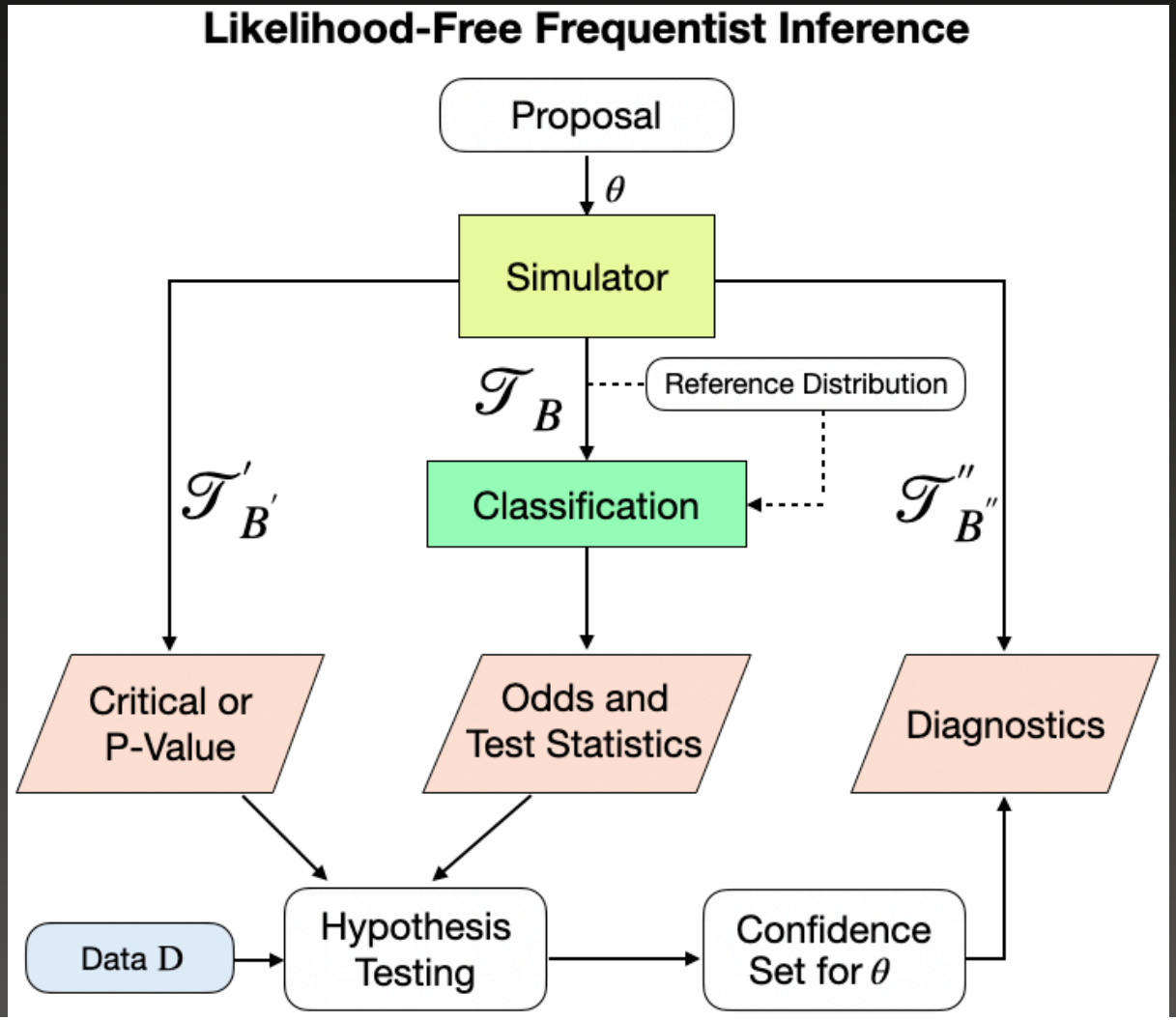
Our diagnostic tool can identify regions in parameter space with UC, CC and OC

(Bottom: heat maps of upper limit of 2**σ** prediction band)



Take-Away: LF2I

- Can construct finite-sample confidence sets with nominal coverage, and provide diagnostics, even without a tractable likelihood. (Do not rely on large n, or costly MC samples)



Take-Away: LF2I

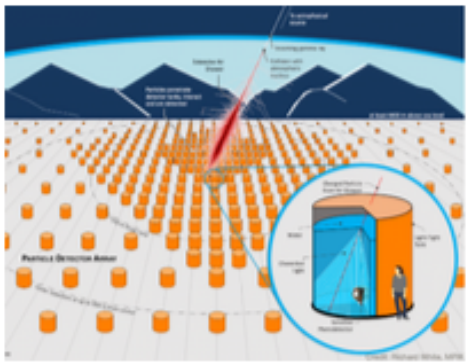
- Validity: Any existing or new test statistic — that is, not only estimates of the LR statistic — can be used in our framework to create frequentist confidence sets. (~10 parameters)
- Power: Hardest to achieve in practice. Area where most statistical and computational advances will take place.
- Nuisance parameters and diagnostics: No guarantee that hybrid methods are valid. However, we have a practical tool for assessing coverage across the entire parameter space.

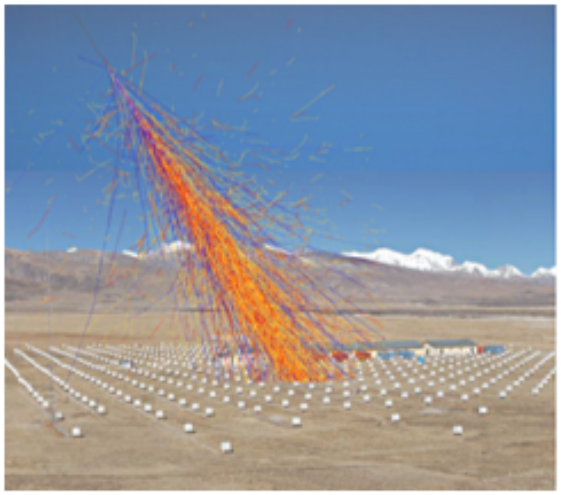
https://github.com/lee-group-cmu/lf2i

Current Projects (2023-)

- Constructing test statistics that are invariant to nuisance parameters (with Luca Masserano and Rafael Izbicki) → next time?
- Nuisance-parametrized LF2I of atmospheric cosmic-ray showers (with Alex Shen, Tommaso Dorigo, Michele Doro, Luca Masserano) → next talk by Alex!

Right: a simulated photon-induced air shower over the LHAASO array in China. Below: a representation of the SWGO array.





Acknowledgments

- Nic Dalmasso (JP Morgan AI)
- Rafael Izbicki (UFSCar)
- Luca Masserano (CMU)
- Mikael Kuusela (CMU)
- Tommaso Dorigo (INFN/Padova)
- David Zhao (CMU)

original LF2I framework

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

This work is funded in part by NSF DMS-2053804 and NSF PHY-2020295.

EXTRA SLIDES START HERE

Likelihood-Free Inference (LFI)

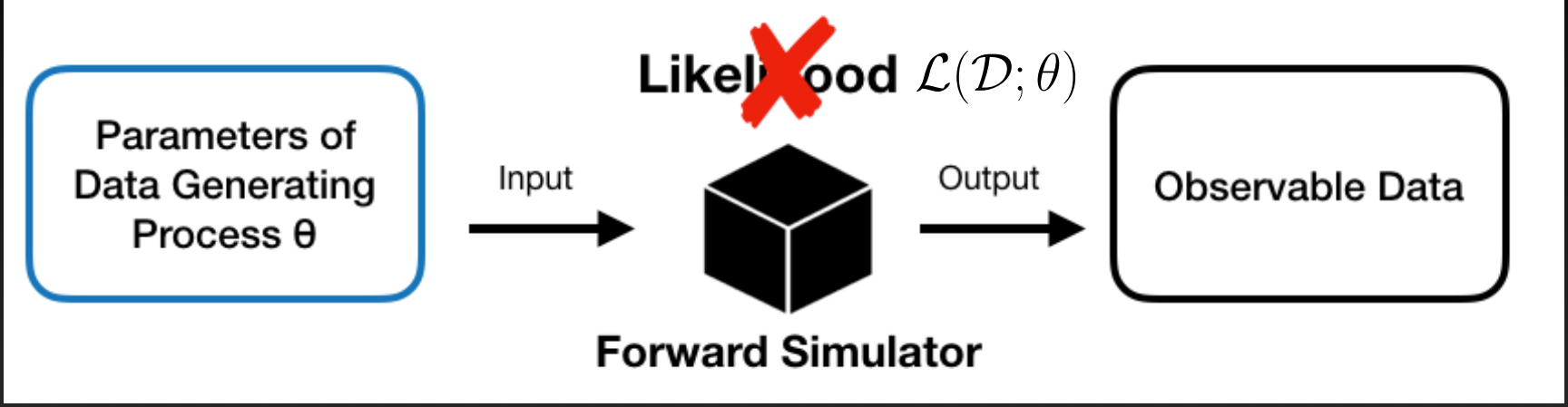


Image credit: Nic Dalmasso

- The likelihood cannot be evaluated. But it is implicitly encoded by the simulator…
- Inference on parameters in this setting is called likelihood-free inference (LFI)

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta), \mathbf{X} \sim F_\theta$$

Predictive AI Approach Can Be Very Powerful, But One Needs to Correct for Bias

[with Luca Masserano, Tommaso Dorigo, Rafael Izbicki and Mikael Kuusela]

Data coming from Dorigo et al. (2020): ~ 400'000 **simulated muons** with true incoming energy sampled uniformly between 100 and 2000 GeV.

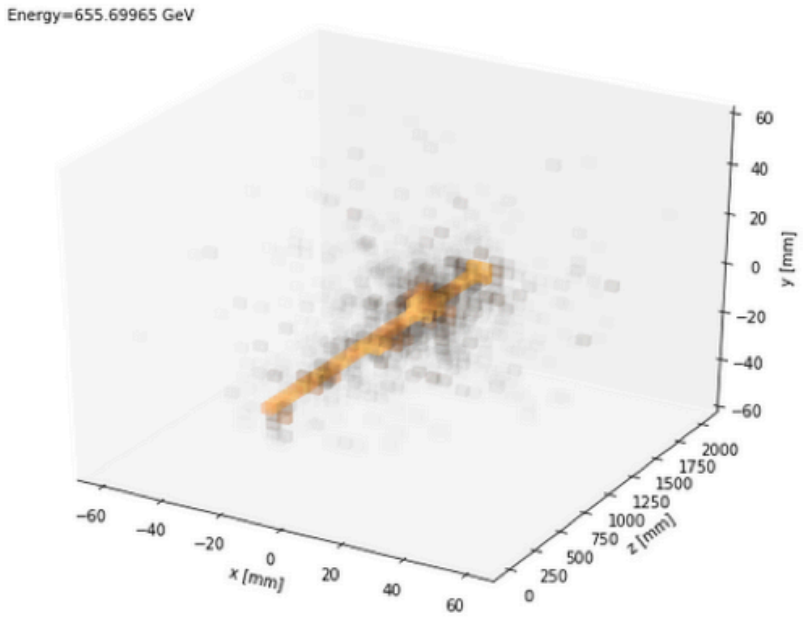


Figure 4: Muon entering the calorimeter in z direction.

[Kieseler et al., July 2021 arXiv:2107.02119]

1. Bias

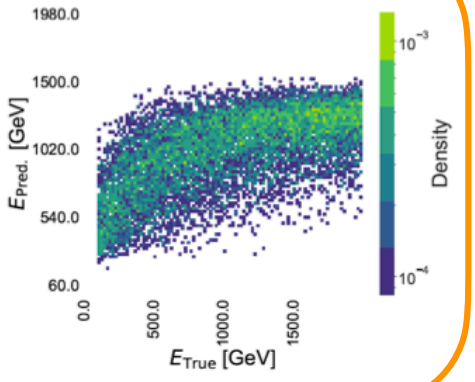


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

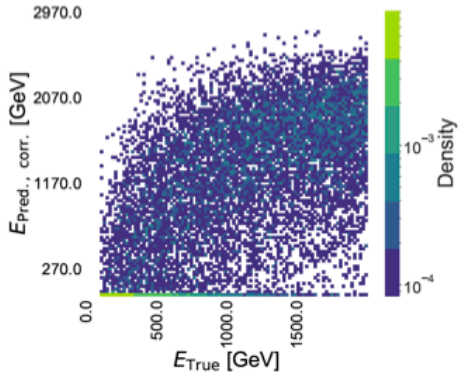


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

$$\mathbb{E}[\theta|X] \neq \theta^{\star}$$

Source: Dorigo et al 2020.
Slide credit: Luca Masserano

https://arxiv.org/abs/2205.15680 (AISTATS 2023)

Simulation-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms or Posterior Estimators for Inverse Problems

Luca Masserano[1] **Tommaso Dorigo**[2] **Rafael Izbicki**[3] **Mikael Kuusela**[1] **Ann B. Lee**[1]

[1]Department of Statistics & Data Science, Carnegie Mellon University

[2]INFN, Sezione di Padova [3]Department of Statistics, Federal University of São Carlos

Abstract

Predictive algorithms, such as deep neural networks (DNNs), are used in many domain sciences to directly estimate internal parameters of interest in simulator-based models, especially in settings where the observations include images or other complex high-dimensional data. In parallel, modern neural density estimators, such as normalizing flows, are becoming increasingly popular for uncertainty quantification, especially when both parameters and observations are high-dimensional. However, parameter inference is an inverse problem and not a prediction task; thus, an open challenge is to construct *conditionally valid* and *precise* confidence regions, with a guaranteed probability of covering the true parameters of the data-generating process, no matter what the (unknown) parameter values are, and without relying on large-sample theory. Many simulator-based inference (SBI) methods are indeed known to produce bi-

1 INTRODUCTION

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{\left(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0\right)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

many science applications, however, one is often interested in the "inverse" problem of estimating the internal parameters of a data-generating process with reliable measures of uncertainty. The parameters of interest, which we denote by $\boldsymbol{\theta}$, are then not directly observed but are the "causes" of the observed data $\mathbf{x}$.

In order to make inference on internal parameters, one needs a statistical model that relates the (unknown) parameters to the observed data. In science and engineering, simulations are often used to model the behavior of complex systems in lieu of an analytical likelihood, when the latter is too complicated to be evaluated explicitly. Let $\mathcal{D} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ denote observable data, where the "sample size" n refers

Back to muon energy calorimeter problem: LF2I/Waldo Confidence Sets Derived from CNN Predictions: Correct Coverage Across the Parameter Space

Data coming from Dorigo et al. (2020): ~ 400'000 **simulated muons** with true incoming energy sampled uniformly between 100 and 2000 GeV.

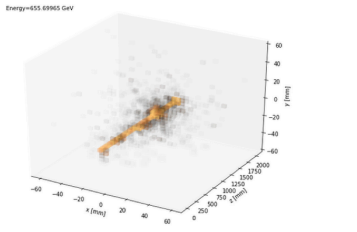
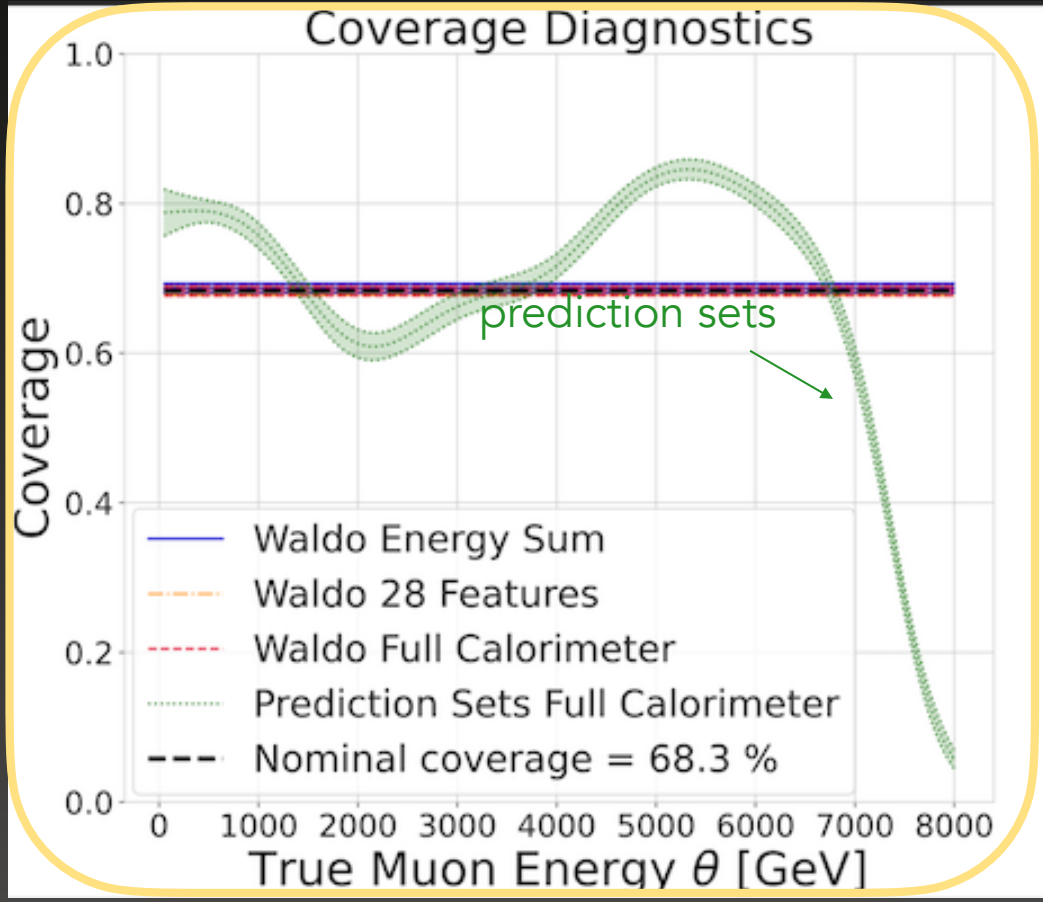


Figure 4: Muon entering the calorimeter in z direction.



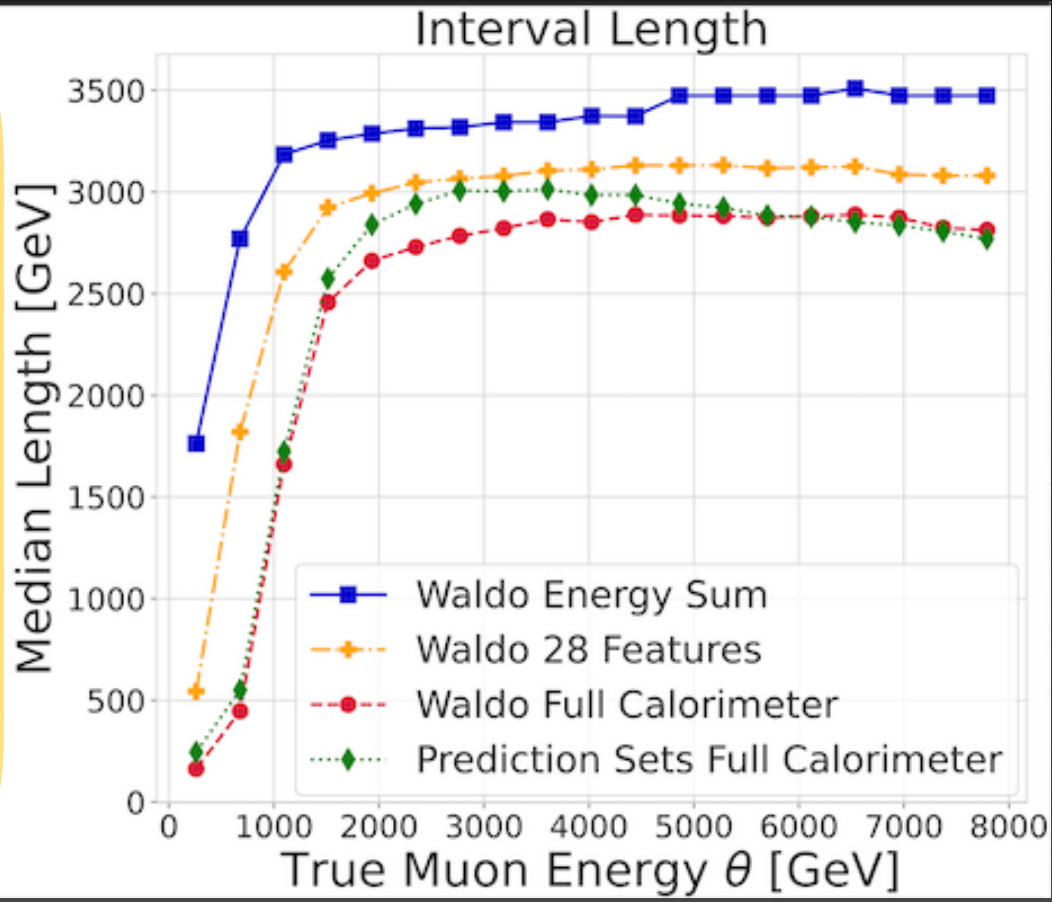
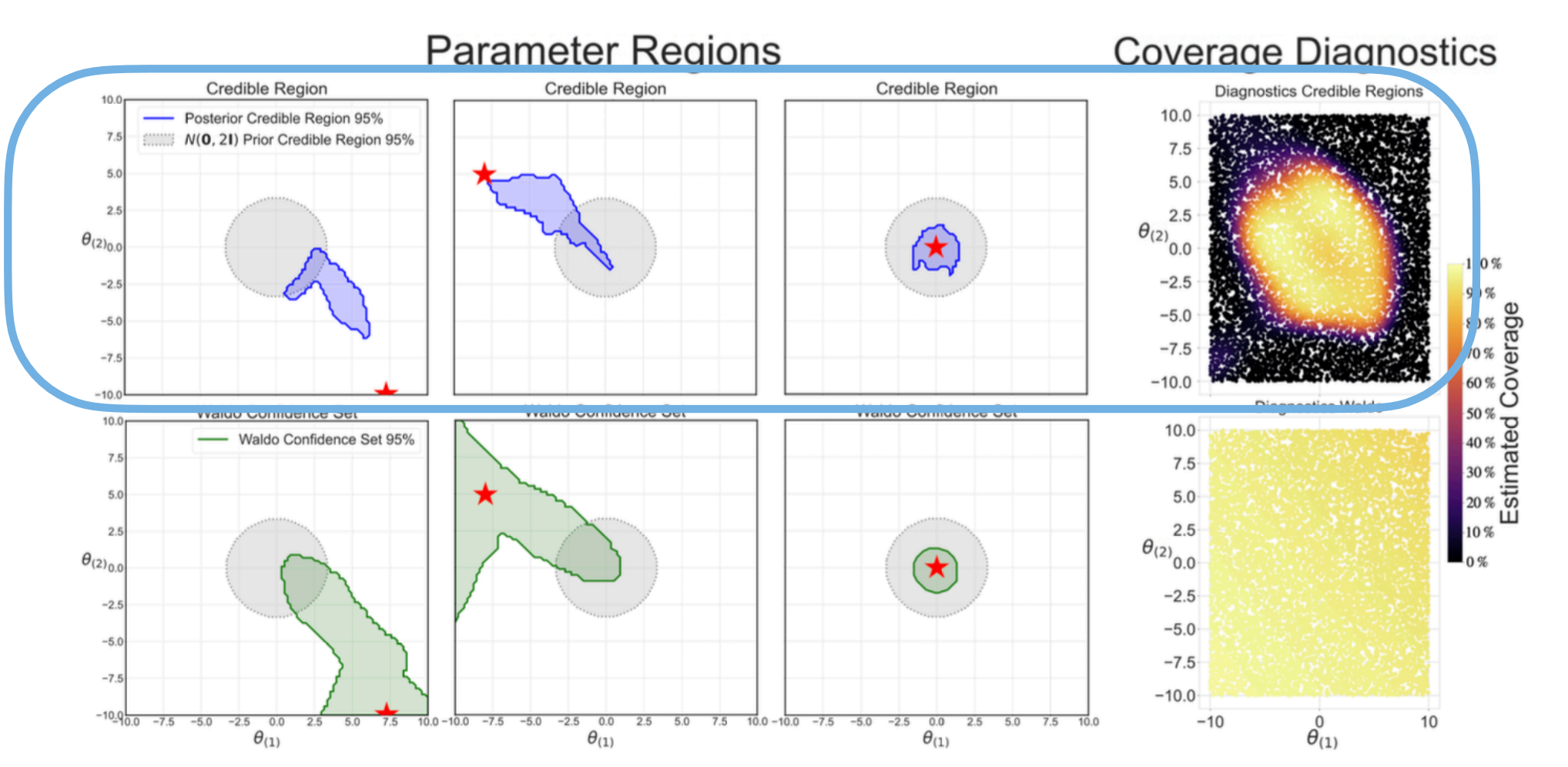


Figure credit: Luca Masserano

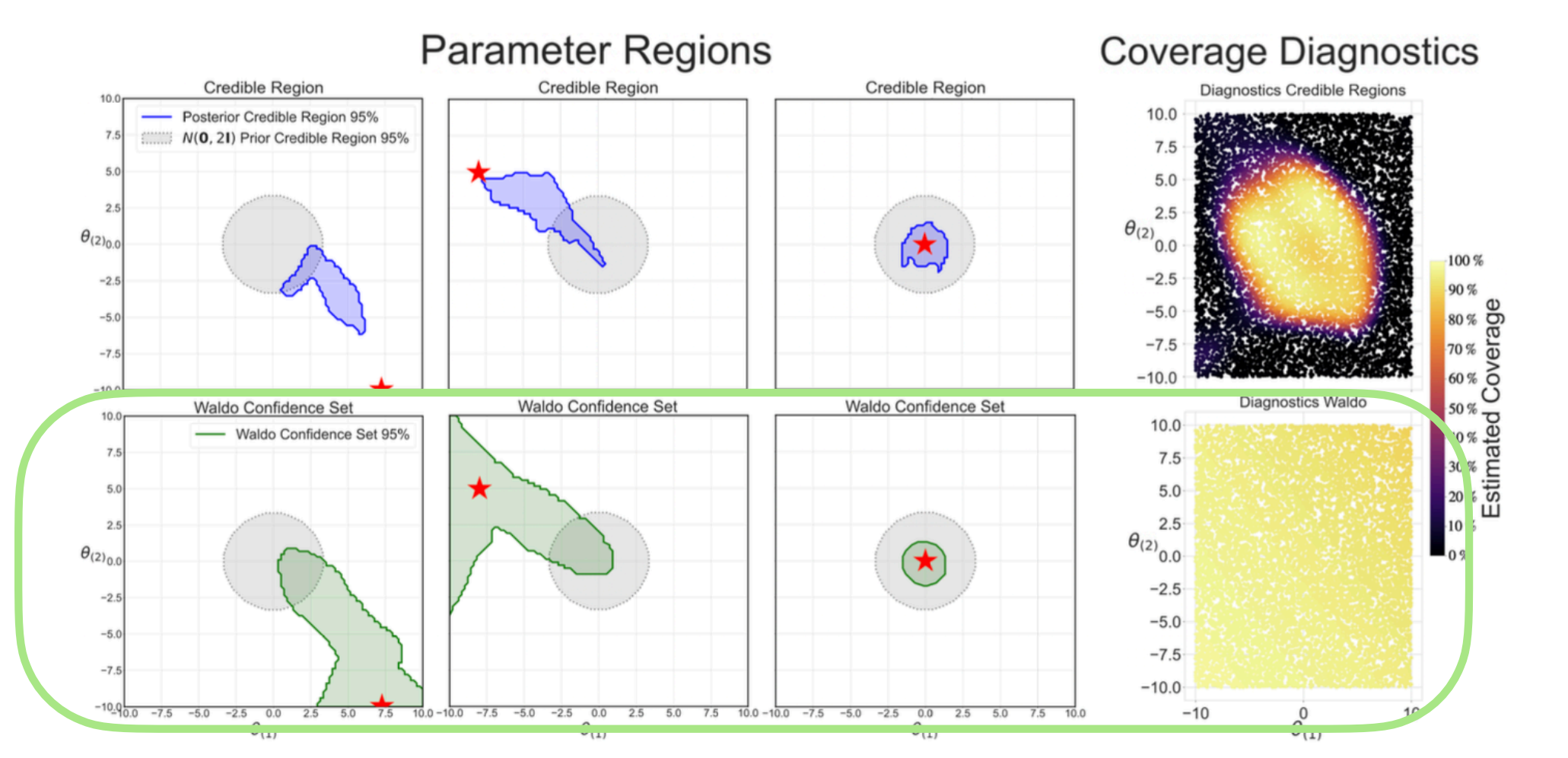
Ex: Credible Regions from Neural (NF) Posteriors

$$\mathcal{D}|\boldsymbol{\theta} \sim \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \odot \mathbf{I}), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^2 \text{ and } n = 1$$



Blue contours: 95% credible regions from Normalizing Flows (overly confident when prior is poorly specified)

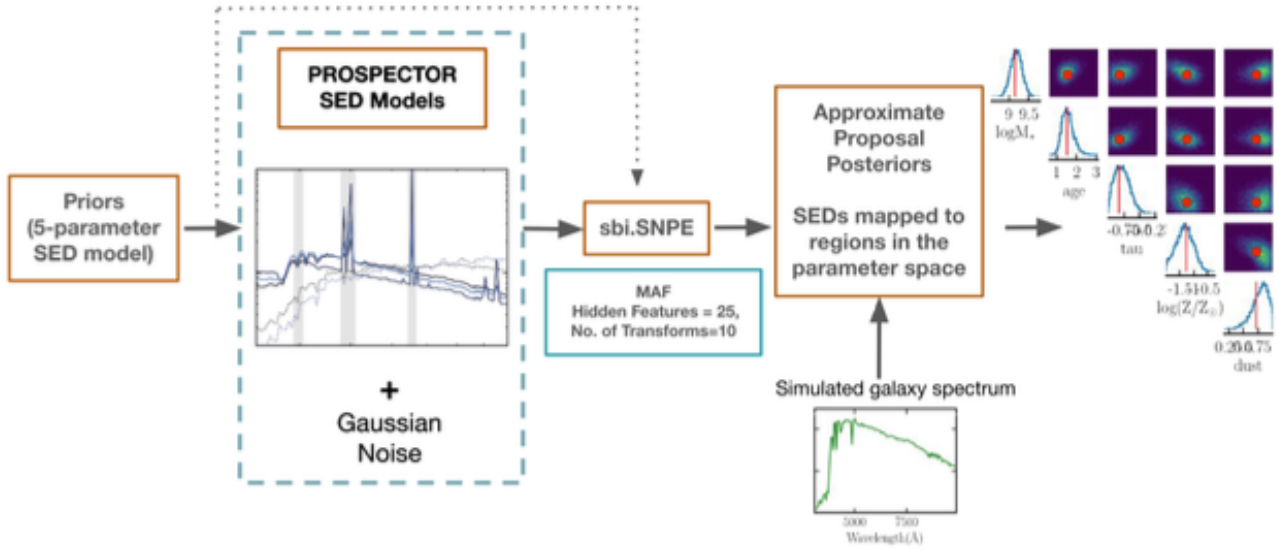
Ex: LF2I/Waldo Confidence Sets Derived from the Same Neural Posteriors ⇨ Correct Coverage



Waldo guarantees coverage everywhere, even if the prior poorly specified. Well-specified prior ⇨ power (tighter constraints)

$$\tau^{\text{WALDO}}(\mathcal{D};\boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

Astronomy: Infer galaxy parameters from SEDs via NPE



Why? Advent of billion-galaxy surveys with complex data needs efficient modeling of spectral energy distributions (SEDs) with robust uncertainty quantification

How? Combine SBI and NPE to infer galaxy parameters (5-parameter model)

Goal: use Waldo to obtain reliable constraints and check their validity against those obtained via NPE

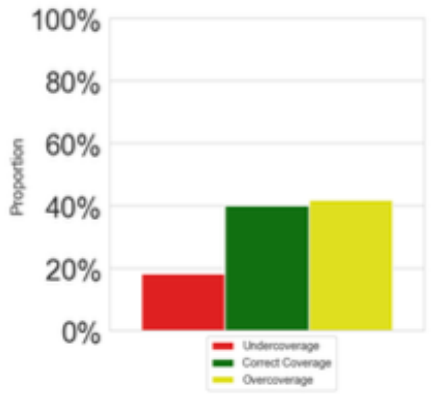
Image taken from Khullar et al. (2022)

13

Coverage across the entire parameter space

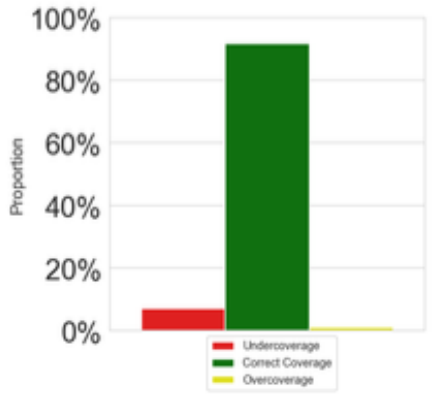
$$r(\theta) := \mathbb{P}(\theta \in \mathscr{R}(\mathscr{D}) \mid \theta), \quad \theta \in \mathbb{R}^5$$

NPE



Proportion of diagnostics simulations whose coverage probability $r(\theta)$ is over/at/under the nominal coverage level

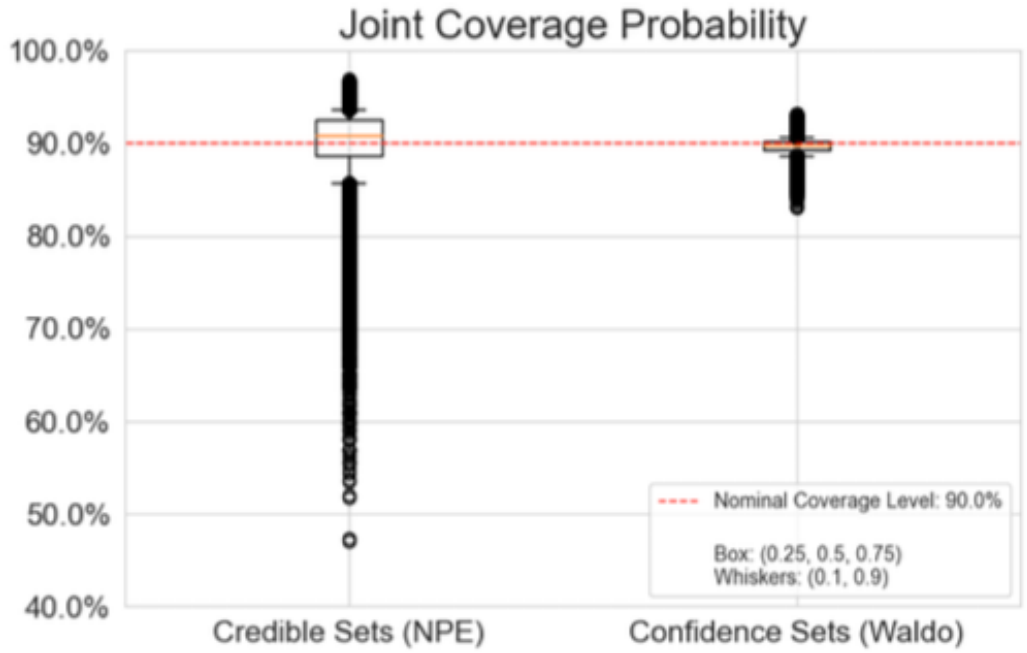
Waldo



- **Waldo** significantly improves the reliability of the constraints on the galaxy parameters, relative to **NPE**
- This is only a partial view on the results. Regions marked as under-coverage or over-coverage might largely differ in the actual coverage probabilities

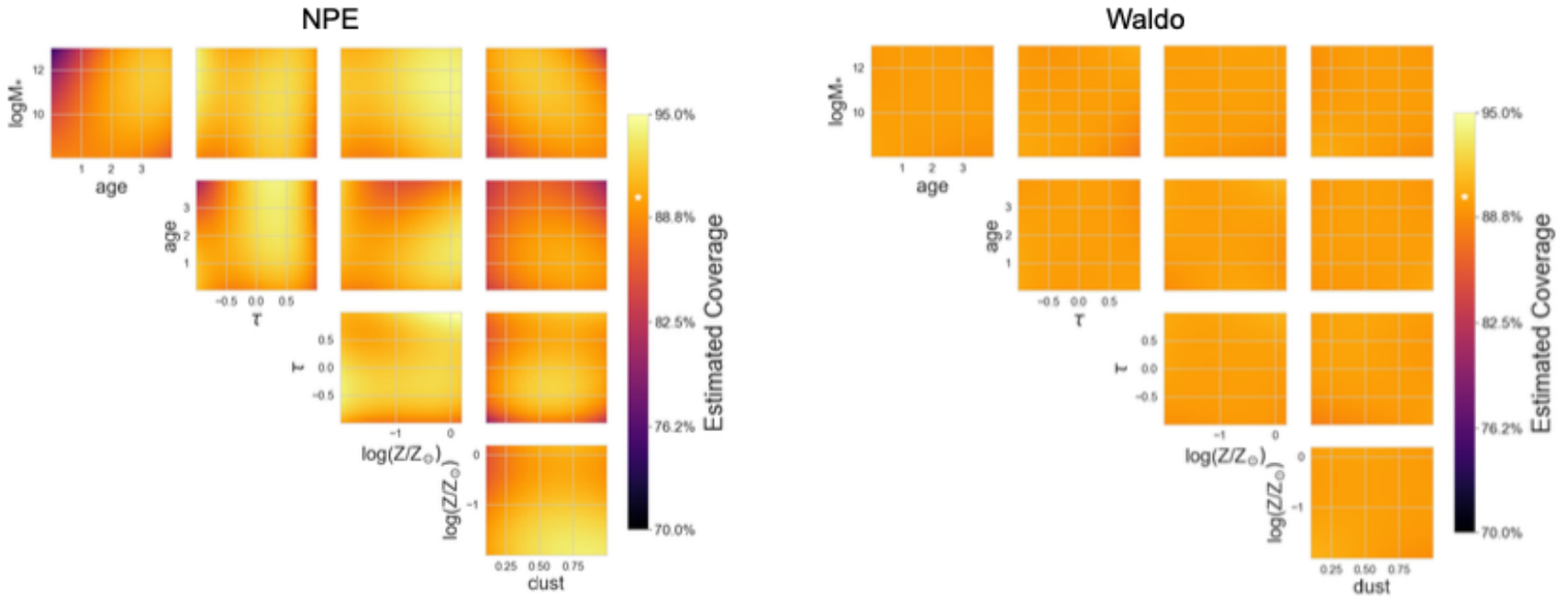
Coverage across the entire parameter space

$$r(\theta) := \mathbb{P}(\theta \in \mathscr{R}(\mathscr{D}) \mid \theta), \quad \theta \in \mathbb{R}^5$$



Partial dependence of coverage probability vs parameters

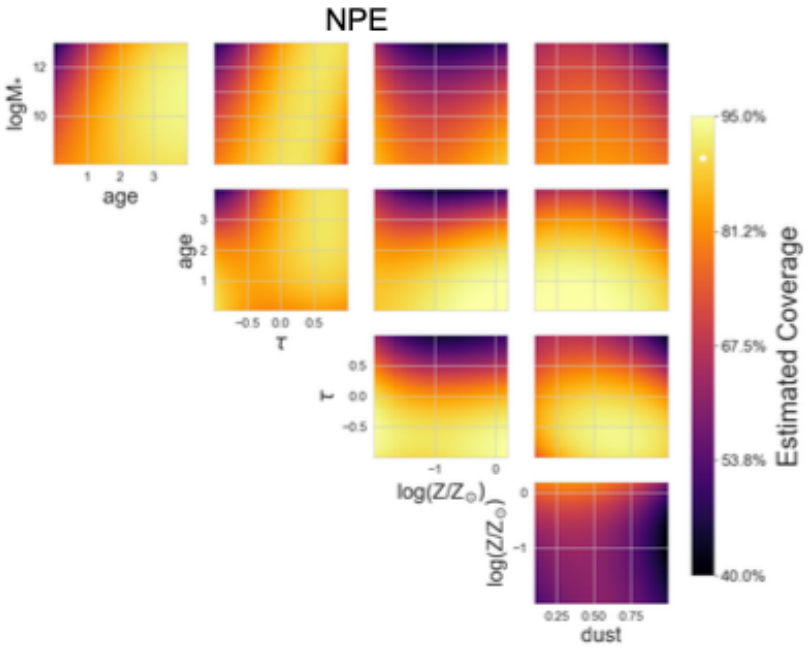
$$\int_{\theta_i\theta_j\theta_k} r(\theta)\mathrm{d}\theta_i\mathrm{d}\theta_j\mathrm{d}\theta_k, \quad \theta \in \mathbb{R}^5$$

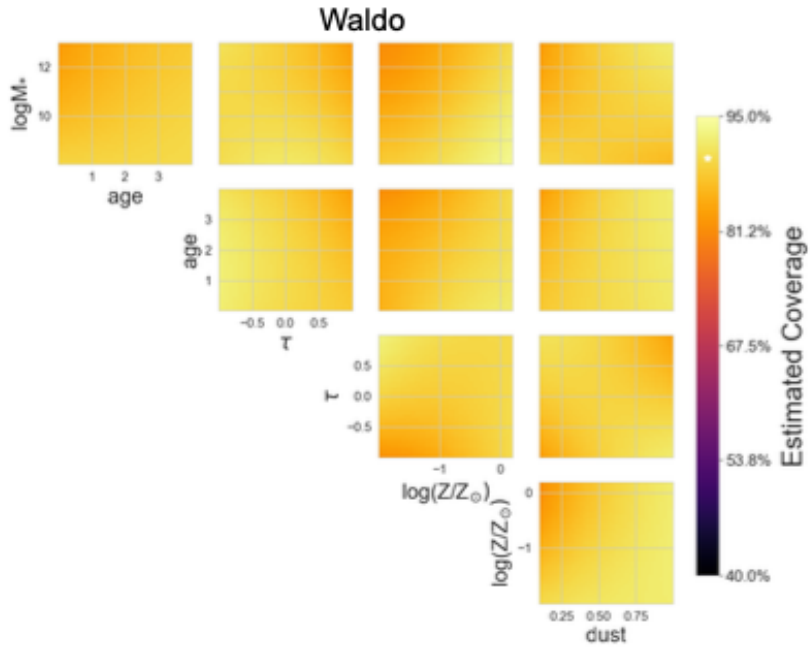


16

Profiled dependence of coverage probability vs parameters

$$\min_{\theta_i \theta_j \theta_k} r(\theta), \quad \theta \in \mathbb{R}^5$$





17

Example of parameter regions when NPE undercovers

Confidence regions (Waldo, green) and credible regions (NPE, blue) obtained from three observations sampled from the same true parameter