Calibration and Validation of Approximate Likelihood Models

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

Joint work with Nic Dalmasso, Rafael Izbicki, Ilmun Kim, and David Zhao

Simulators are Ubiquitous in Science

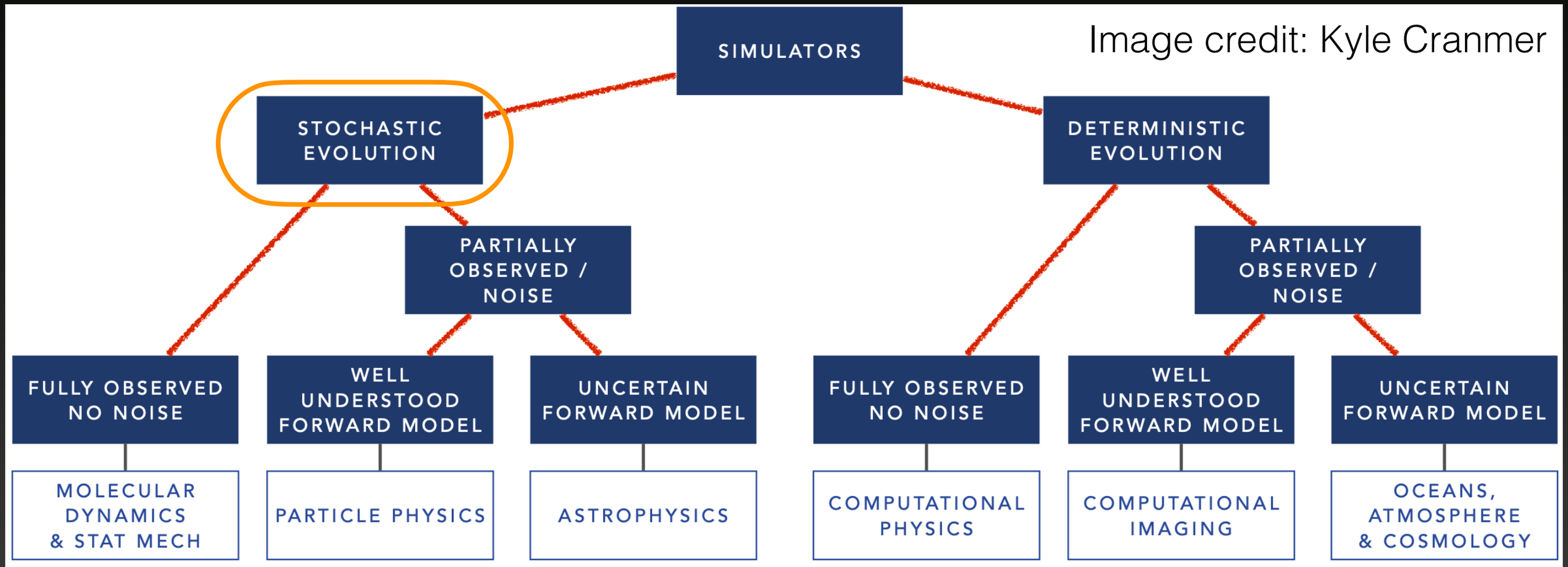
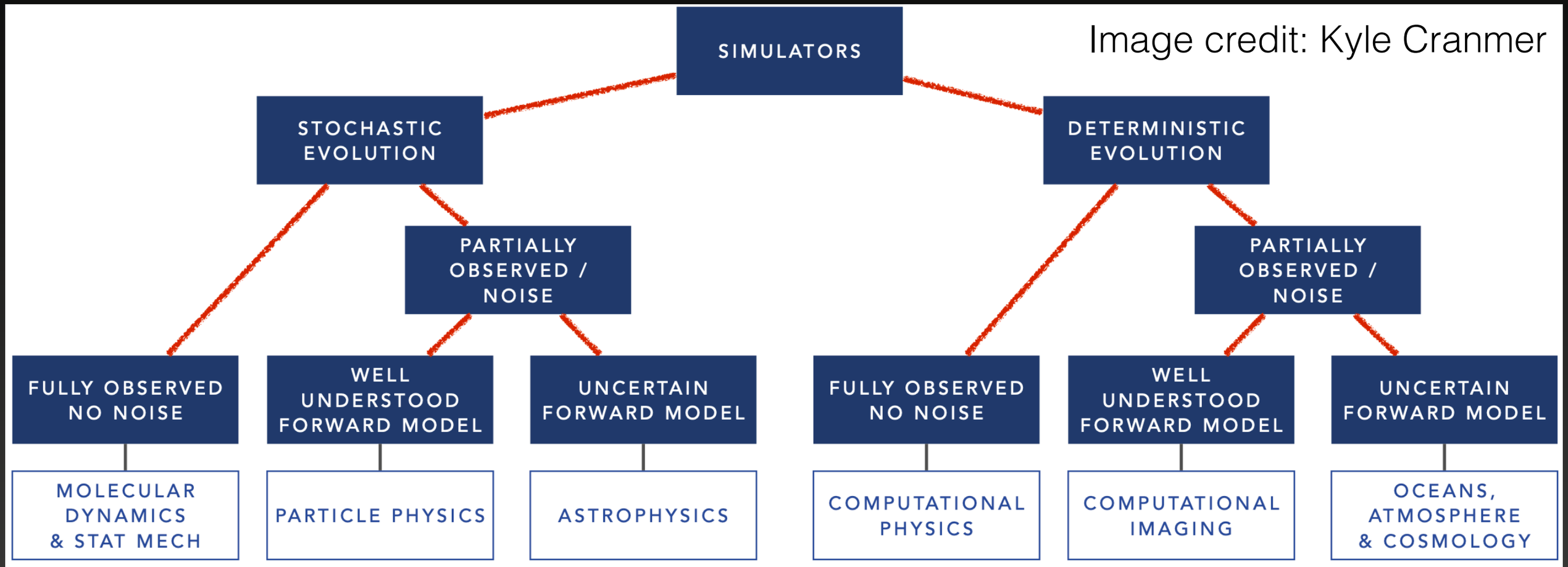


Image credit: Kyle Cranmer

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

Simulators are Ubiquitous in Science



- Simulators may be good at simulating realistic data — but often poorly suited for the inverse problem of inferring the underlying scientific mechanisms. High-fidelity simulations can also be slow => fit approximate model (emulator)

Statistical Challenges for Complex Models

- Forward problem: Does data from the approximate model have the same distribution as high-fidelity (simulated or observed) data?
  - Ask if two distributions are different, and if so, how they differ in high dimensions (capture dependencies between all variables)?

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

- Inverse problem: Suppose we have a forward model F_θ that implicitly encodes the relationship between parameter θ of interest (input) and high-dimensional observable data $\mathbf{X}$ (output).
  - Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, can we infer the true parameters θ with valid measures of uncertainty (confidence sets)?

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$

Statistical Challenges for Complex Models

- Forward problem: Does data from the approximate model have the same distribution as high-fidelity (simulated or observed) data?
  - Ask if two distributions are different, and if so, how they differ in high dimensions (capture dependencies between all variables)?

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

- Inverse problem: Suppose we have a forward model F_θ that implicitly encodes the relationship between parameter θ of interest (input) and high-dimensional observable data $\mathbf{X}$ (output).
  - Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, can we infer the true parameters θ with valid measures of uncertainty (confidence sets)?

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$

Ex 1: Comparing Distributions in High Dimensions.

How are the Morphologies of the Galaxy Populations Different?

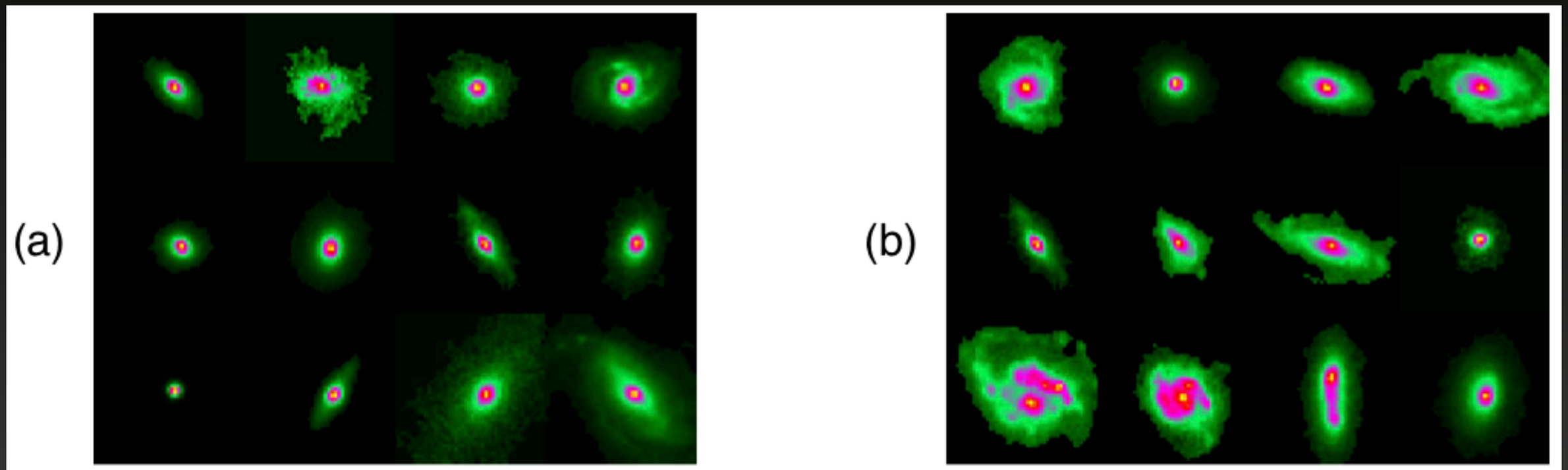
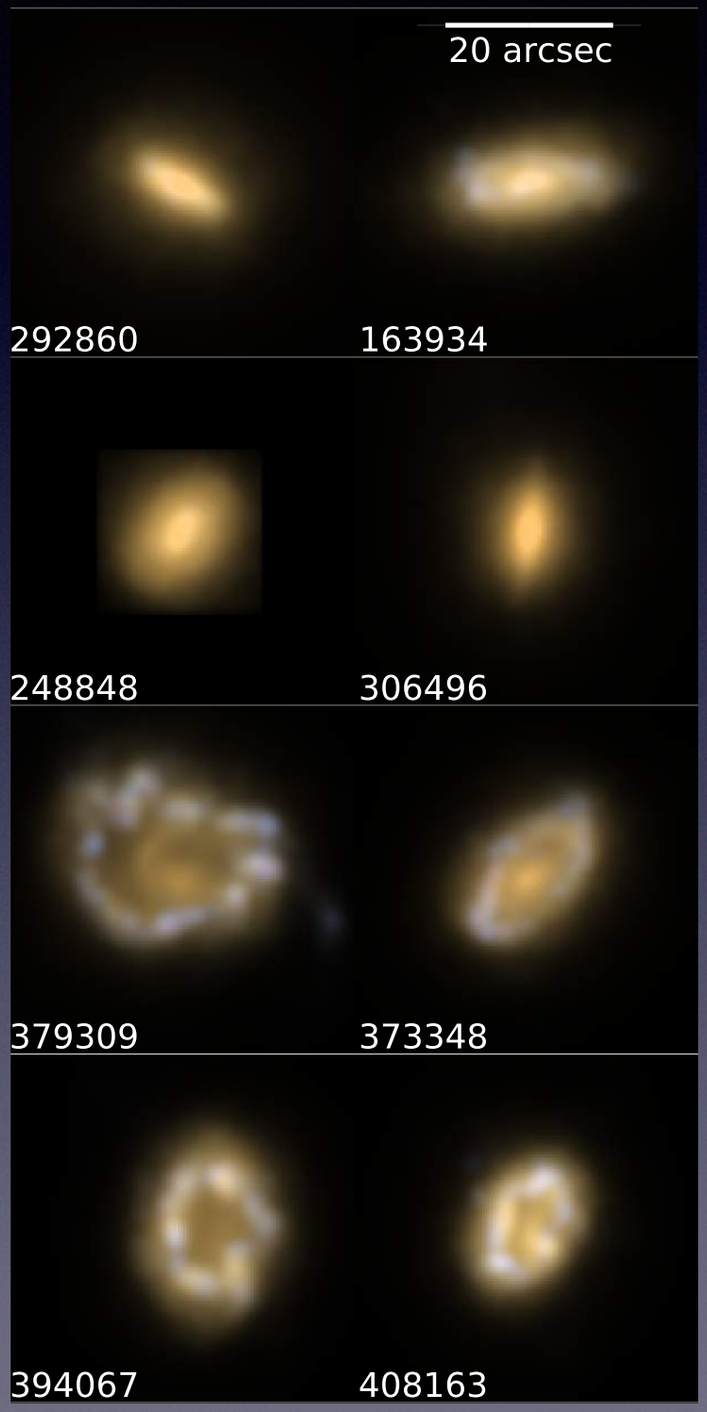


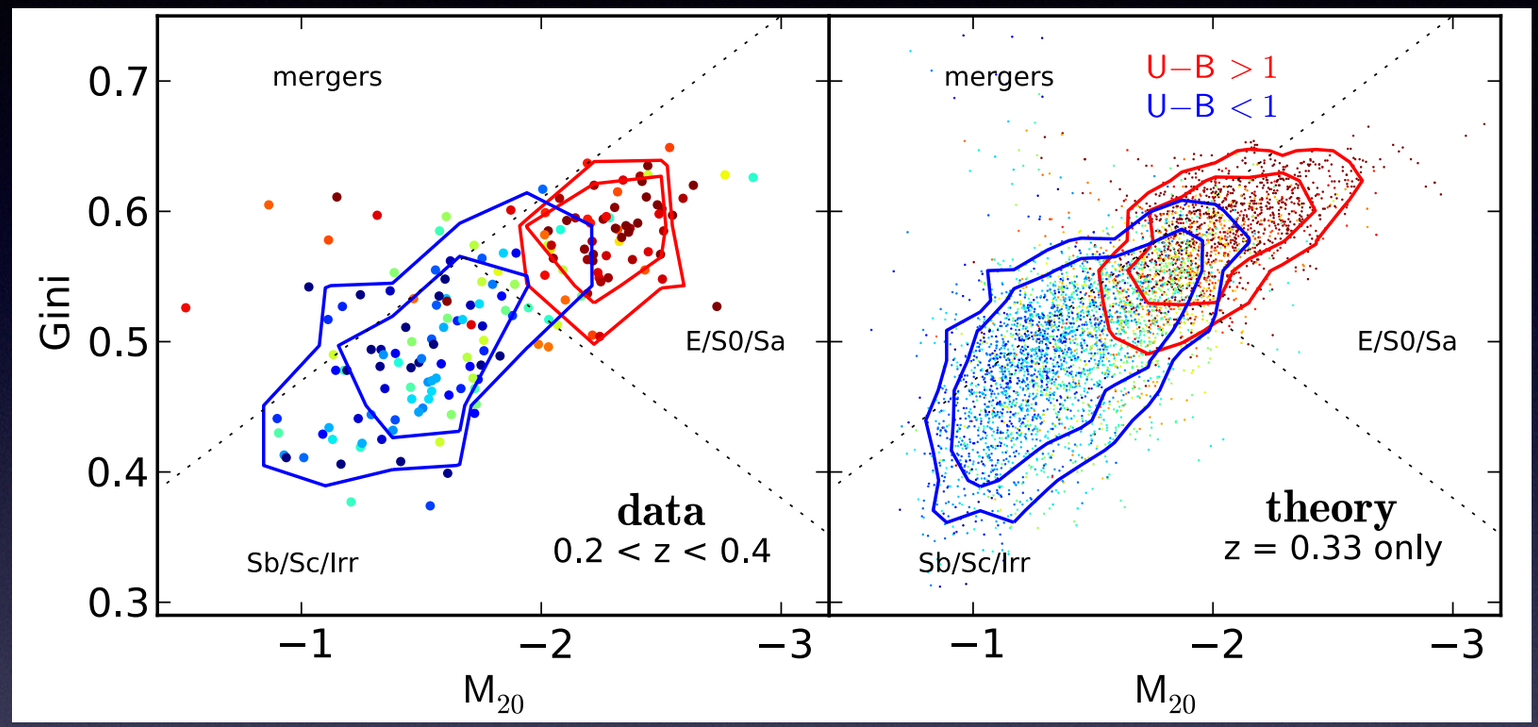
Figure 7: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

- Can we answer the question if, and if so, how two populations are different without just looking at histogram of a few individual features?

[Figure credit: Dalmasso et al, 2019]



Snyder et al. (2015)

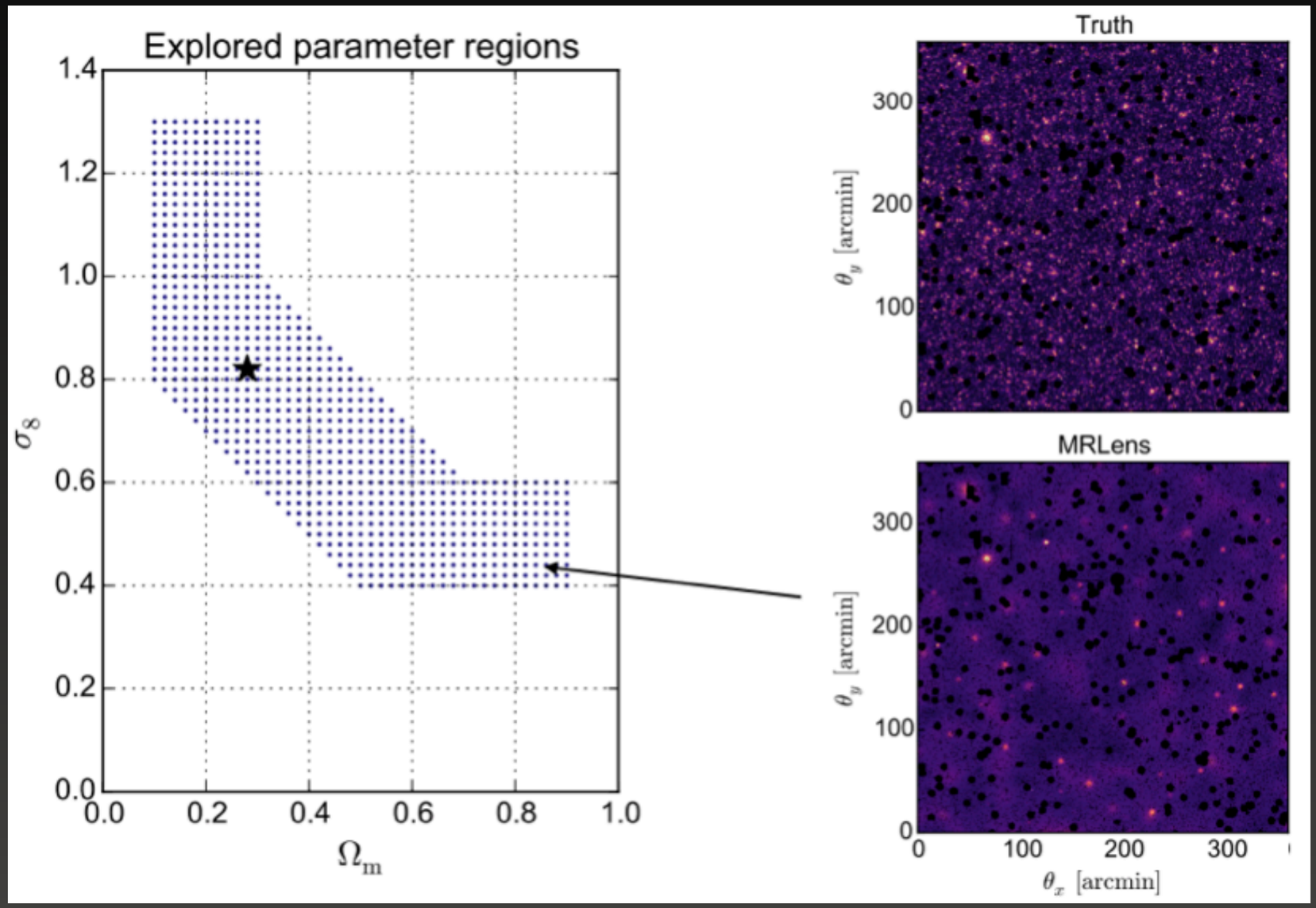


Snyder et al. (2015)

With regard to our first statistical aim, we wish to identify regions in the sample space where the distributions F and G are significantly different and to use this information, e.g., to infer redshift evolution (given two observed samples) or to inform improvements in simulation codes (by comparing simulation output at one wavelength to *HST* data at that same wavelength), etc.

[Slide credit: Peter Freeman]

Ex 2: Comparing Distributions in High Dimensions. Calibrate Forward Model with Internal Parameters



- Goal: Simulate weak lensing data to constrain parameters of the Lambda CDM model in “Big Bang” cosmology

[Image credit: Lin & Kilbinger, 2018]

Statistical Methods for Comparing Distributions of High-Dimensional Data

Electronic Journal of Statistics
Vol. 13 (2019) 5253–5305
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1648

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **471,** 3273–3282 (2017) doi:10.1093/mnras/stx1807
Advance Access publication 2017 July 18

Local two-sample testing: a new tool for analysing high-dimensional astronomical data

P. E. Freeman,★ I. Kim and A. B. Lee
Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

Niccolò Dalmasso,[1] Ann B. Lee,[1] Rafael Izbicki,[2] Taylor Pospisil,[3] Ilmun Kim,[1] Chieh-An Lin[4]

https://arxiv.org/abs/1905.11505 (AISTATS 2020)

HECT: High-Dimensional Ensemble Consistency Testing for Climate Models

https://arxiv.org/abs/2010.04051 (NeurIPS Workshop 2020)

Niccolò Dalmasso[*,1] **Galen Vincent**[*,1] **Dorit Hammerling**[2] **Ann B. Lee**[1]

[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]Department of Applied Mathematics and Statistics, Colorado School of Mines
ndalmass@stat.cmu.edu

Statistical Setting: Two-Sample Test

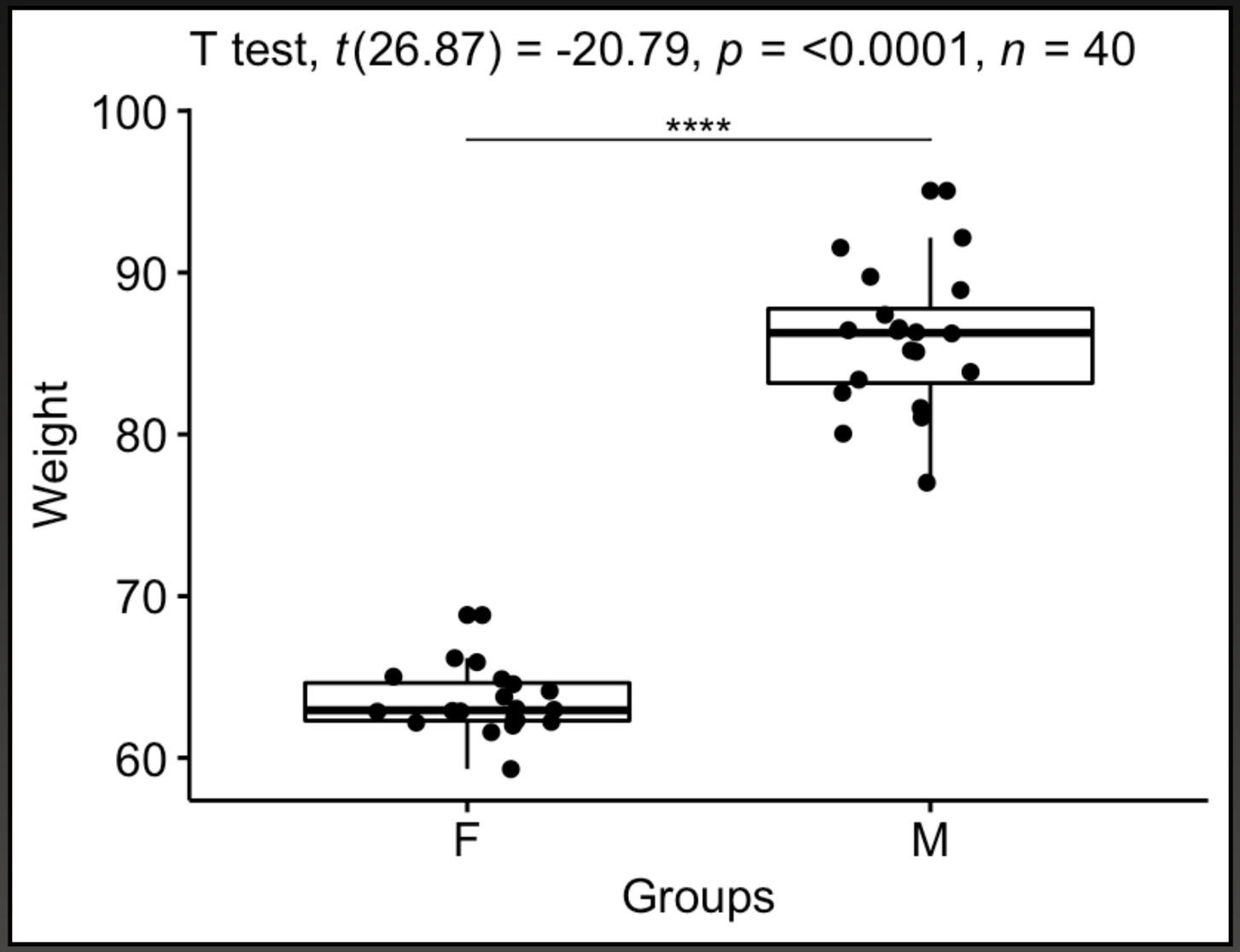
Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_m^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_n^1 \sim P_1$$

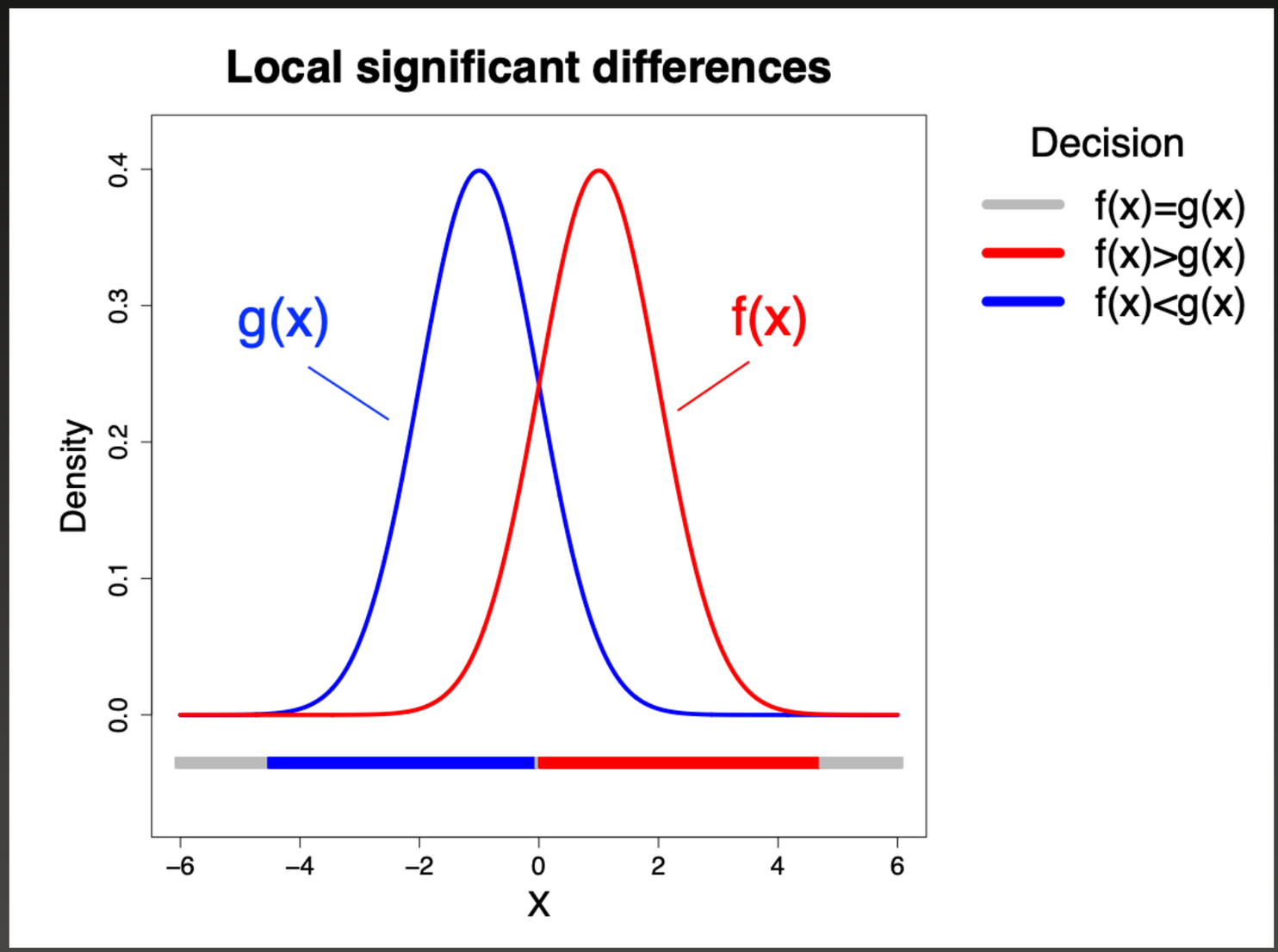
A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : f(\mathbf{x}|Y = 0) = f(\mathbf{x}|Y = 1) \text{ for all } \mathbf{x} \in \mathcal{X}$$

Comparing Distributions:
Traditional Approaches Focus on Univariate Tests

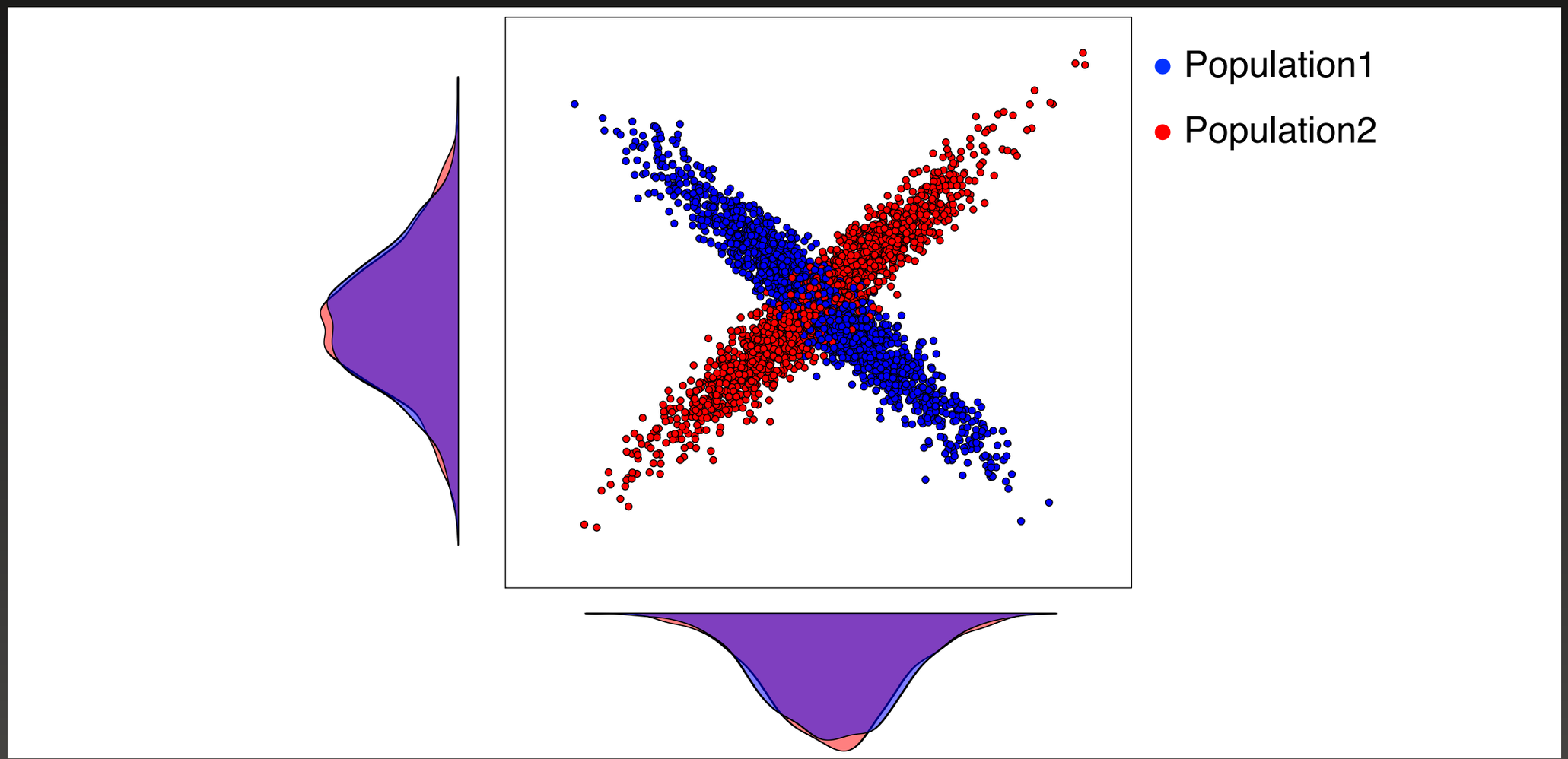


1. We are looking for regions in the sample space where the two populations have significantly different densities



[Figure credit: Ilmun Kim 2016]

2. We are searching for differences in high-dimensional space (e.g., each data point could represent an image or a sequence of images)



[Figure credit: Ilmun Kim 2016]

Two-Sample Test via Regression

[Kim, Lee and Lei 2019]

Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_m^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_n^1 \sim P_1$$

A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : f(\mathbf{x}|Y = 0) = f(\mathbf{x}|Y = 1) \text{ for all } \mathbf{x} \in \mathcal{X}$$

Two-Sample Test via Regression

[Kim, Lee and Lei 2019]

Suppose we have two samples:

$$\mathbf{X}_1^0, \ldots, \mathbf{X}_m^0 \sim P_0 \quad \text{and} \quad \mathbf{X}_1^1, \ldots, \mathbf{X}_n^1 \sim P_1$$

A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : f(\mathbf{x}|Y=0) = f(\mathbf{x}|Y=1) \text{ for all } \mathbf{x} \in \mathcal{X}$$

By Bayes rule, this is equivalent to testing

$$H_0 : \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1) \text{ for all } \mathbf{x} \in \mathcal{X}$$

Why Two-Sample Test via Regression?

$$H_0: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1) \ \text{ for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \neq \mathbb{P}(Y=1) \ \text{ for some } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}} = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{m}(\mathbf{X}_i) - \widehat{\pi}_1\right)^2 .$$

- Can adapt to any structure in X for which there is a suitable regression technique.
- The power of the test is directly related to the the MISE of the chosen regression estimator [Kim et al, 2019]
- The regression test tells you not only if but also how the two samples are different in space of observables

If the chosen regression estimator has a small MISE, the power of the test is large over a wide region of the alternative hypothesis

Theorem 1. *Suppose that the regression estimator $\widehat{m}(\mathbf{x})$ is a linear smoother satisfying*

$$\sup_{m \in \mathcal{M}} \mathbb{E} \int_{\mathcal{X}} \left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 dP_X(\mathbf{x}) \leq C_0 \delta_n, \quad (2)$$

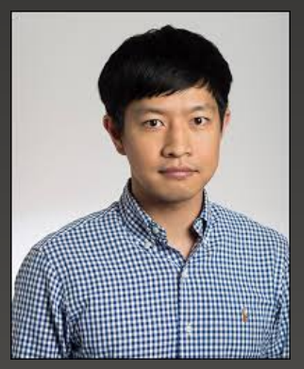
where C_0 is a positive constant, $\delta_n = o(1)$, $\delta_n \geq n^{-1}$, and $\mathcal{M}$ is a class of regressions $m(\mathbf{x})$ containing constant functions. Let t^*_α be the upper α quantile of the permutation distribution of the test statistic $\widehat{T}'$ on validation data.[1] *Then for any $\alpha, \beta \in (0, 1/2)$, there exists a universal constant C_1 such that*

- *Type I error:* $\mathbb{P}_0\left(\widehat{T}' \geq t^*_\alpha\right) \leq \alpha$, *and*
- *Type II error:* $\sup_{m \in \mathcal{M}(C_1\delta_n)} \mathbb{P}_1\left(\widehat{T}' < t^*_\alpha\right) \leq \beta$

against the class of alternatives $\mathcal{M}(C_1\delta_n)$ defined by

$$\left\{m \in \mathcal{M} : \int_{\mathcal{X}} \left(m(\mathbf{x}) - \pi_1\right)^2 dP_X(\mathbf{x}) \geq C_1\delta_n\right\},$$

for n sufficiently large.



[Ref: Kim, Lee, & Lei; EJS 2019]

Why Two-Sample Test via Regression?

$$H_0: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1) \ \text{ for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \neq \mathbb{P}(Y=1) \ \text{ for some } \mathbf{x} \in \mathcal{X}$$

$$\widehat{\mathcal{T}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}(\mathbf{X}_i) - \widehat{\pi}_1\right)^2 .$$

- Can adapt to any structure in X for which there is a suitable regression technique
- The power of the test is directly related to the the MISE of the chosen regression estimator [Kim et al, 2019]
- The regression test tells you not only if, but also how, the two samples are different in space of observables

Let's return to the galaxy morphology example…

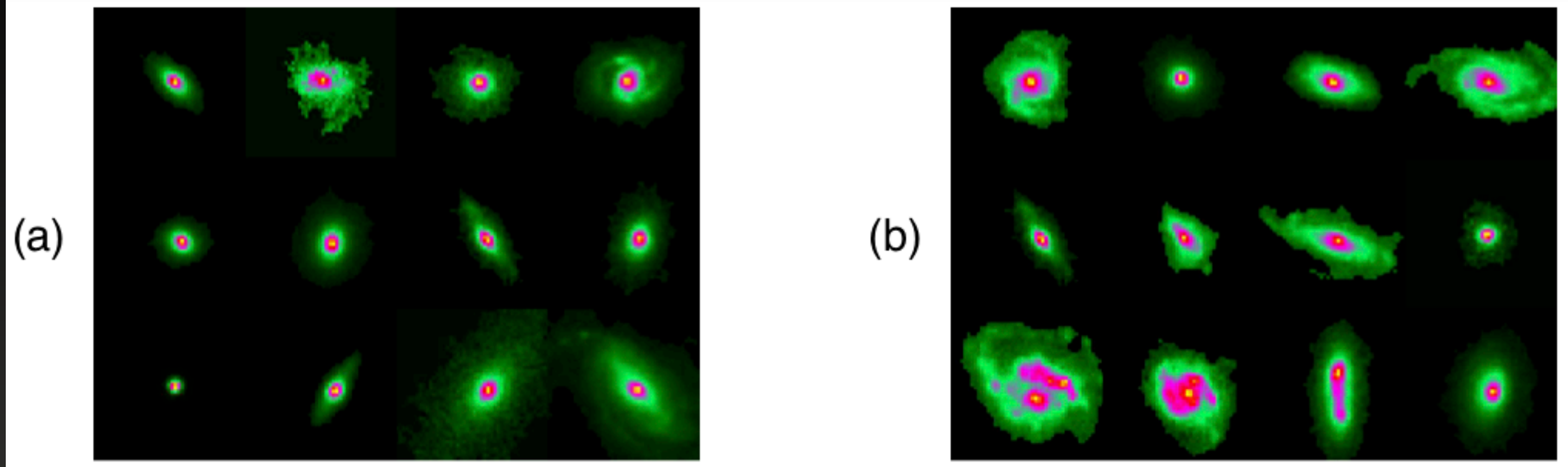


Figure 7: Examples of galaxies from (a) the low-SFR sample $\mathcal{S}_0$ versus (b) the high-SFR sample $\mathcal{S}_1$.

- Divide 2736 galaxies from the CANDELS program into two populations based on SFR: "Low SFR" vs "High SFR" sample
- Consider seven morphology summary statistics jointly
- Are the morphologies the same or not (compared to chance)?

Regression Test to Identify If and How Two Distributions Differ in 7-Dim Feature Space

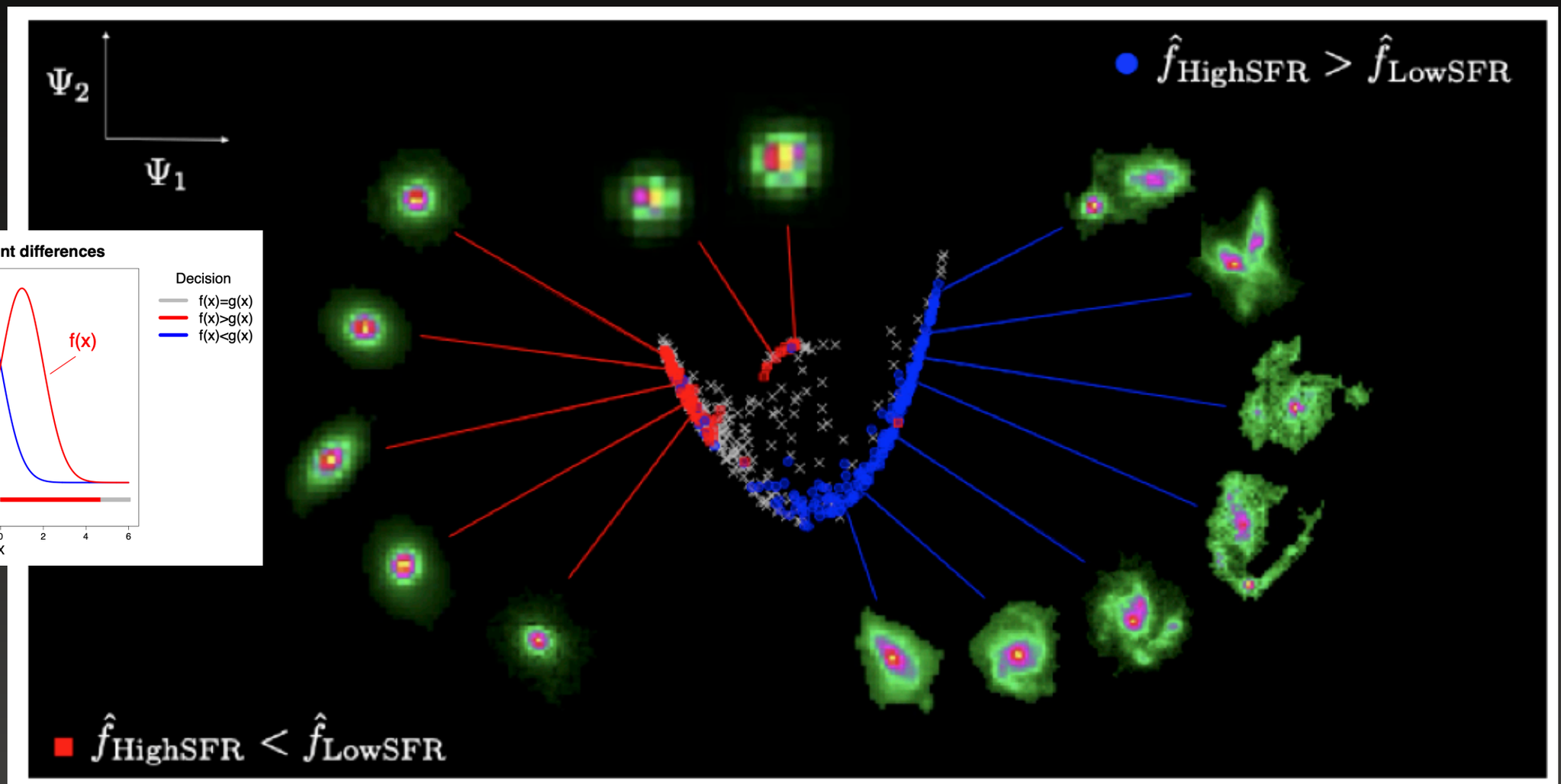


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].

We can use a similar approach to compare forward models with internal parameters

$$\begin{aligned} \text{Test} \quad & H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta \\ \text{versus} \quad & H_1 : \widehat{\mathcal{L}}(\mathbf{x};\theta) \neq \mathcal{L}(\mathbf{x};\theta) \text{ for some } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta \end{aligned}$$

- Our framework can help answer:
  - IF one needs to improve the emulator model
  - WHERE in parameter space Θ the fit might be poor
  - HOW the distributions of emulated and high-fidelity simulated data may differ in observable space χ

Two-Step Procedure: Local Test at Each Parameter. Global Test of Uniformity of Local P-Values.

Algorithm 1 Local Test

Input: parameter value θ_0, two-sample testing procedure, number of draws from the true model, $n_{\text{sim},0}$ and from the estimated model, $n_{\text{sim},1}$

Output: p-value p_{θ_0} for testing if $L(\mathbf{x};\theta_0)=\widehat{L}(\mathbf{x};\theta_0)$ for every $\mathbf{x}$

1: Sample $\mathcal{S}_0 = \{\mathbf{X}_1^{\theta_0}, \ldots, \mathbf{X}_{n_{\text{sim},0}}^{\theta_0}\}$ from $\mathcal{L}(\mathbf{x};\theta_0)$.
2: Sample $\mathcal{S}_1 = \{\mathbf{X}_1^{*}, \ldots, \mathbf{X}_{n_{\text{sim},1}}^{*}\}$ from $\widehat{\mathcal{L}}(\mathbf{x};\theta_0)$.
3: Compute p-value p_{θ_0} for the comparison between $\mathcal{S}_0$ and $\mathcal{S}_1$.
4: **return** p_{θ_0}

- For the local test, our regression test allows us to accommodate any data type with interpretable diagnostics.
- Global test is consistent against all alternatives if the local test is consistent.

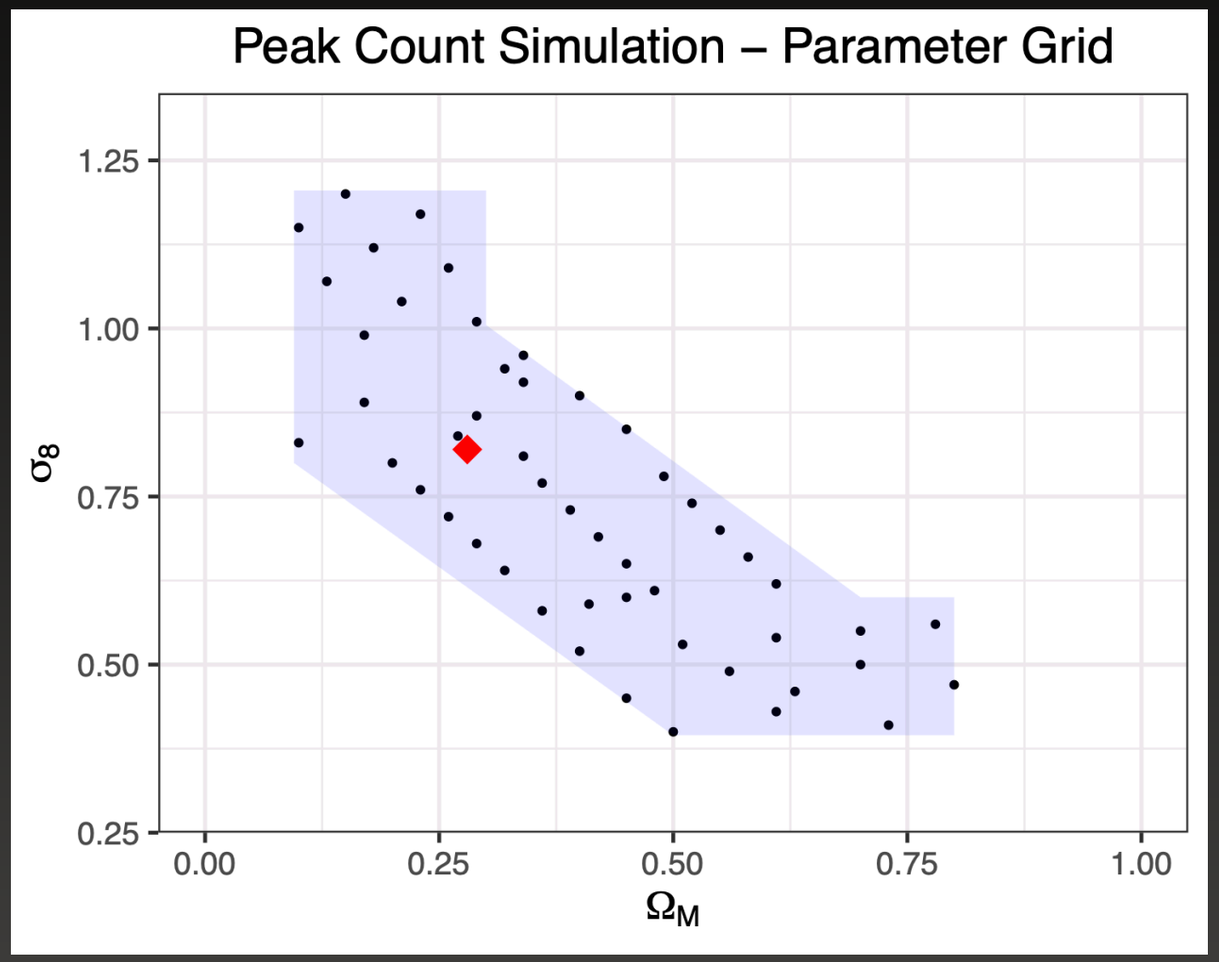
Algorithm 3 Global Test

Input: reference distribution $r(\theta)$, B, uniform testing procedure

Output: p-value p for testing if $L(\mathbf{x};\theta) = \widehat{L}(\mathbf{x};\theta)$ for every $\mathbf{x}$ and θ

1: **for** $i \in \{1, \ldots, B\}$ **do**
2: sample $\theta_i \sim r(\theta)$
3: compute p_{θ_i} using Algorithm 1
4: **end for**
5: Compute p-value p for testing if $(p_{\theta_i})_{i=1}^{B}$ has a uniform distribution.
6: **return** p

Back to Example: Simulate Weak Lensing Data to Constrain Cosmological Parameters

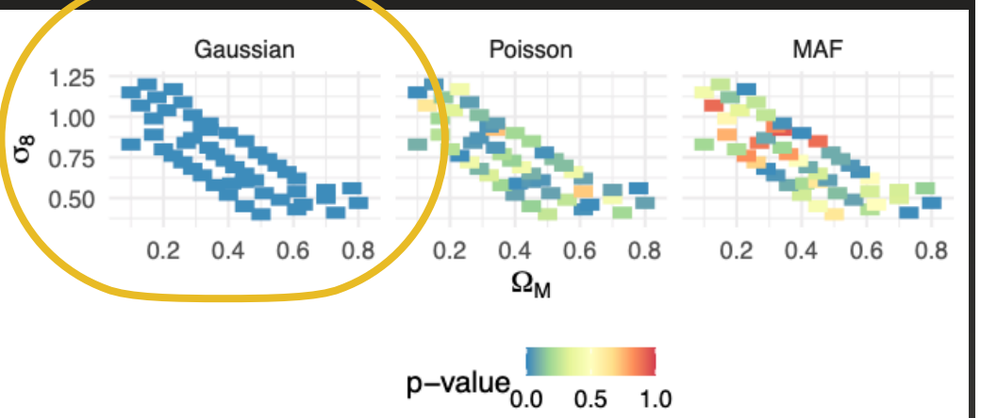


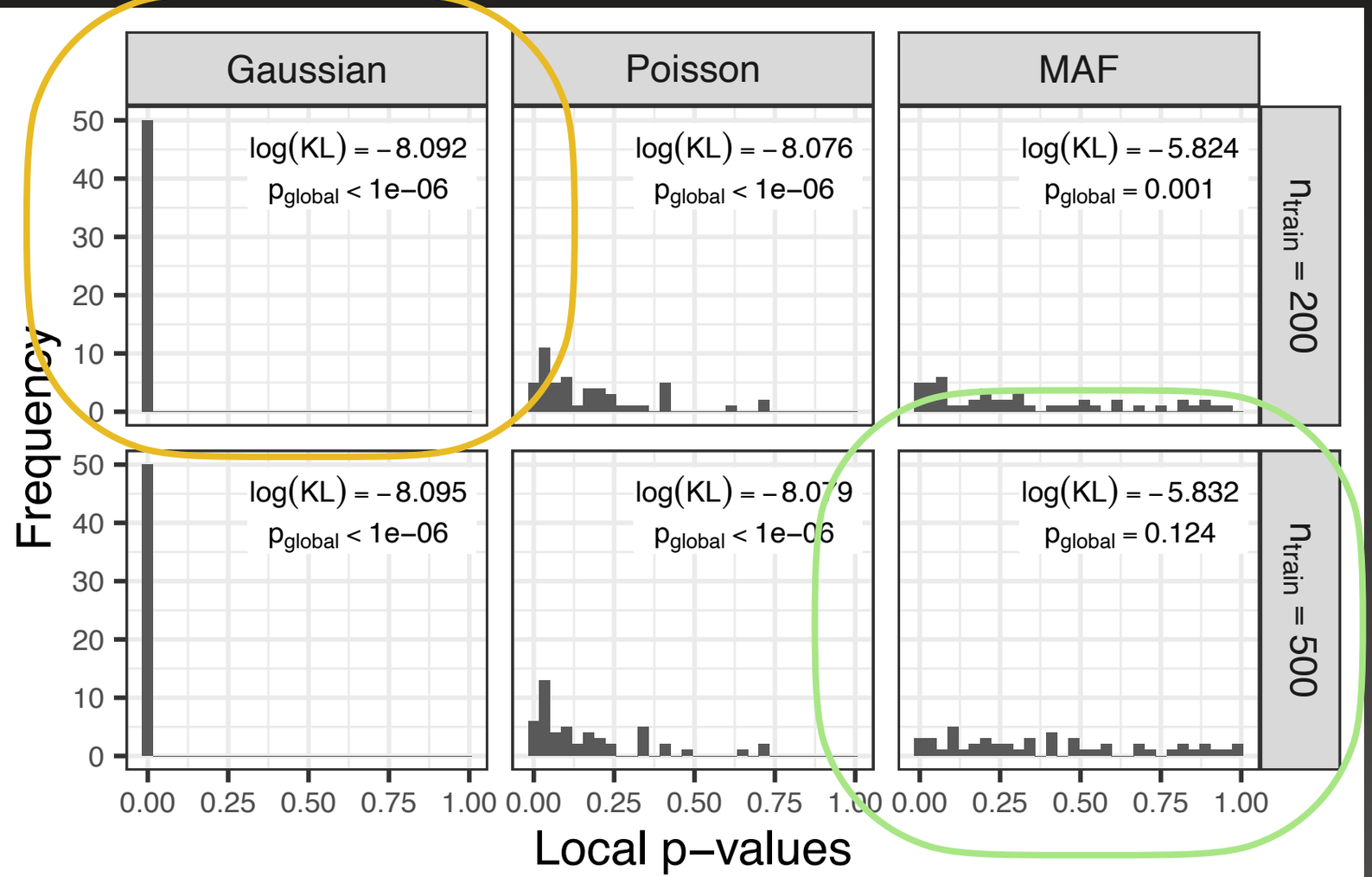
- Use CAMELUS [Lin & Kilbinger 2015] to simulate weak lensing convergence maps
  ⇒ binned peak counts $x \in \mathbb{N}^7$
- Batch of 200 train + 200 test simulations at 50 different cosmologies/parameter settings.
- Fit 3 different likelihood models: Gaussian, Poisson, MAF

$$\begin{aligned} \text{Test} \quad & H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta \\ \text{versus} \quad & H_1 : \widehat{\mathcal{L}}(\mathbf{x};\theta) \neq \mathcal{L}(\mathbf{x};\theta) \text{ for some } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta \end{aligned}$$

Do we need more simulations to fit the data well or are the current fits good enough?

- Based on the KL loss we would choose the Gaussian likelihood model — but our local test p-values reveal that the Gaussian model is rejected at all parameter values.





Even if it's not feasible to simulate more data, our regression test provides valuable diagnostics…

$$H_0: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y=1) \ \text{ for all } \mathbf{x} \in \mathcal{X}$$
$$H_1: \ \mathbb{P}(Y=1|\mathbf{X}=\mathbf{x}) \neq \mathbb{P}(Y=1) \ \text{ for some } \mathbf{x} \in \mathcal{X}$$

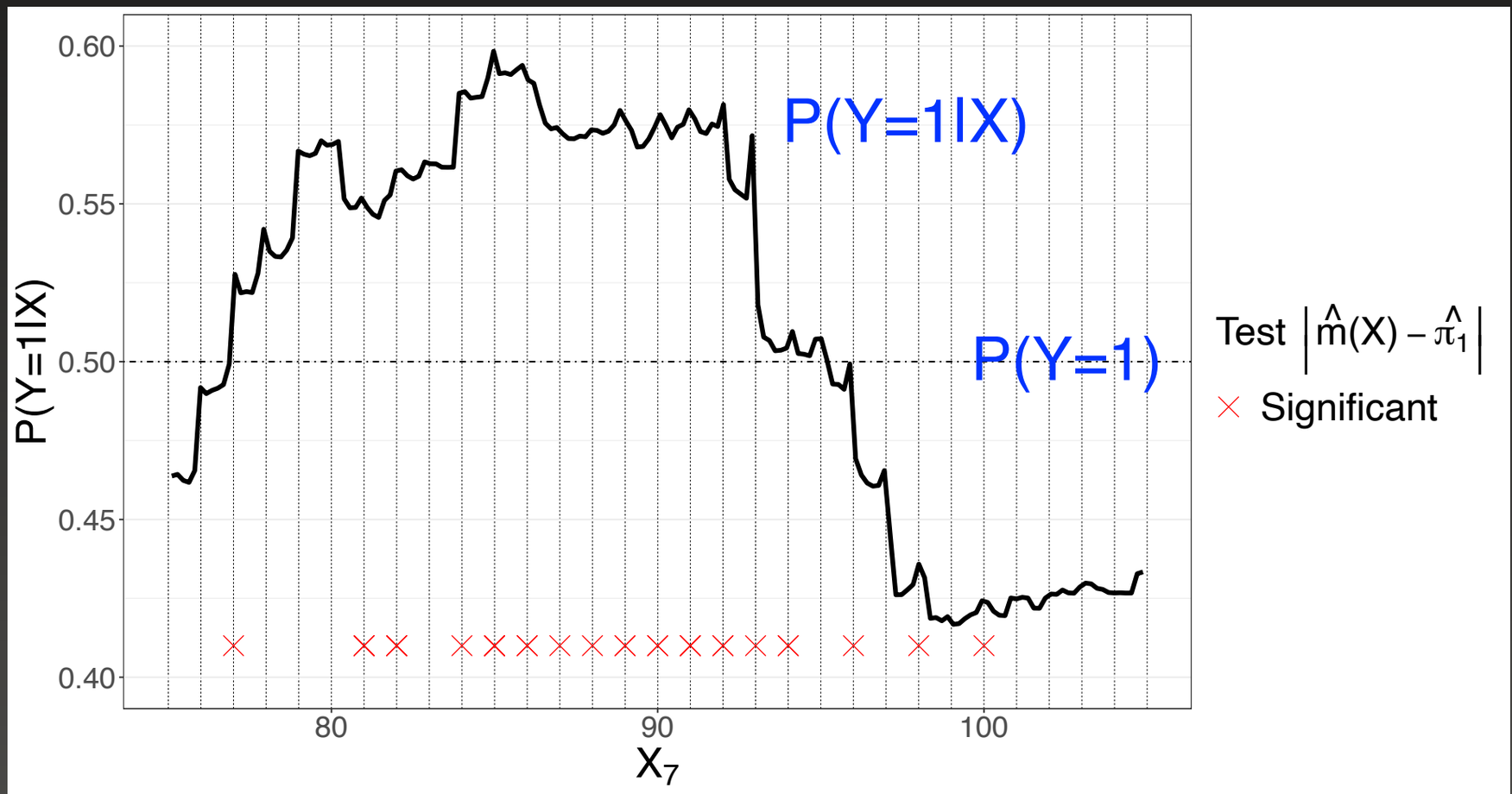
$$\widehat{\mathcal{T}} = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{m}(\mathbf{X}_i) - \widehat{\pi}_1\right)^2.$$

- the difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ provides information on how well the emulator fits the simulator in feature space: we can test whether $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ is statistically significantly higher!

Emulator diagnostics: Our regression test tells us how the two samples are different in $\mathbb{N}^7$

- According to our random forest regression, bins with low counts (e.g. bin X_7) contribute the most to the rejection of the Gaussian model.

Partial dependence plot for variable X_7. The regression test is distinguishing between the discrete true distribution and the approximate Gaussian continuous distribution.



Statistical Challenges for Complex Models

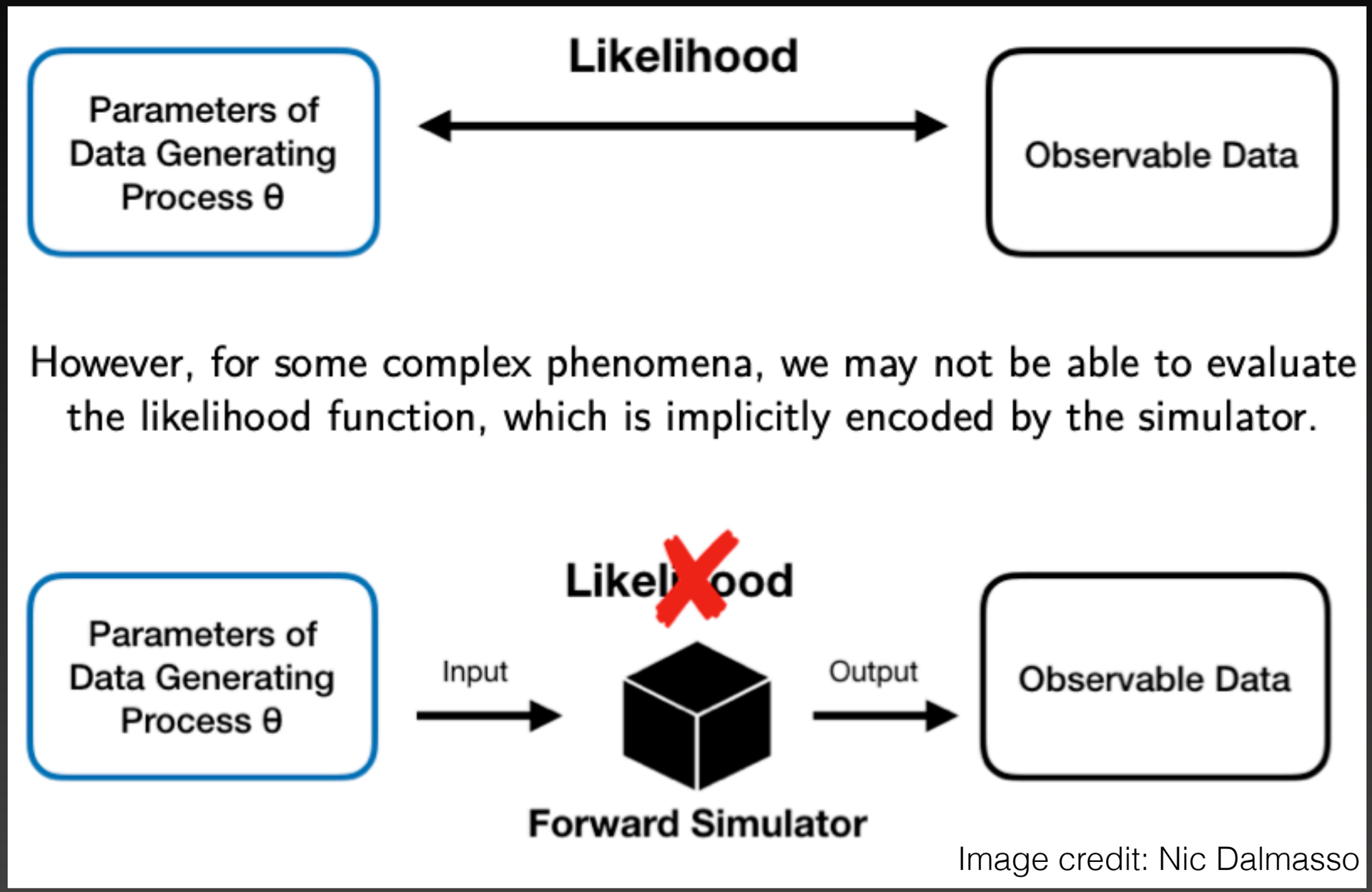
- Forward problem: Does data from the approximate model have the same distribution as high-fidelity (simulated or observed) data?
  - Ask if two distributions are different, and if so, how they differ in high dimensions (capture dependencies between all variables)?

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

- Inverse problem: Suppose we have a forward model F_θ that implicitly encodes the relationship between parameter θ of interest (input) and high-dimensional observable data $\mathbf{X}$ (output).
  - Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, can we infer the true parameters θ with valid measures of uncertainty (confidence sets)?

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$

What is Likelihood-Free Inference?



- Inference on parameters in the second setting is called likelihood-free inference (LFI).

Classical LFI: Approximate Bayesian Computation (ABC)

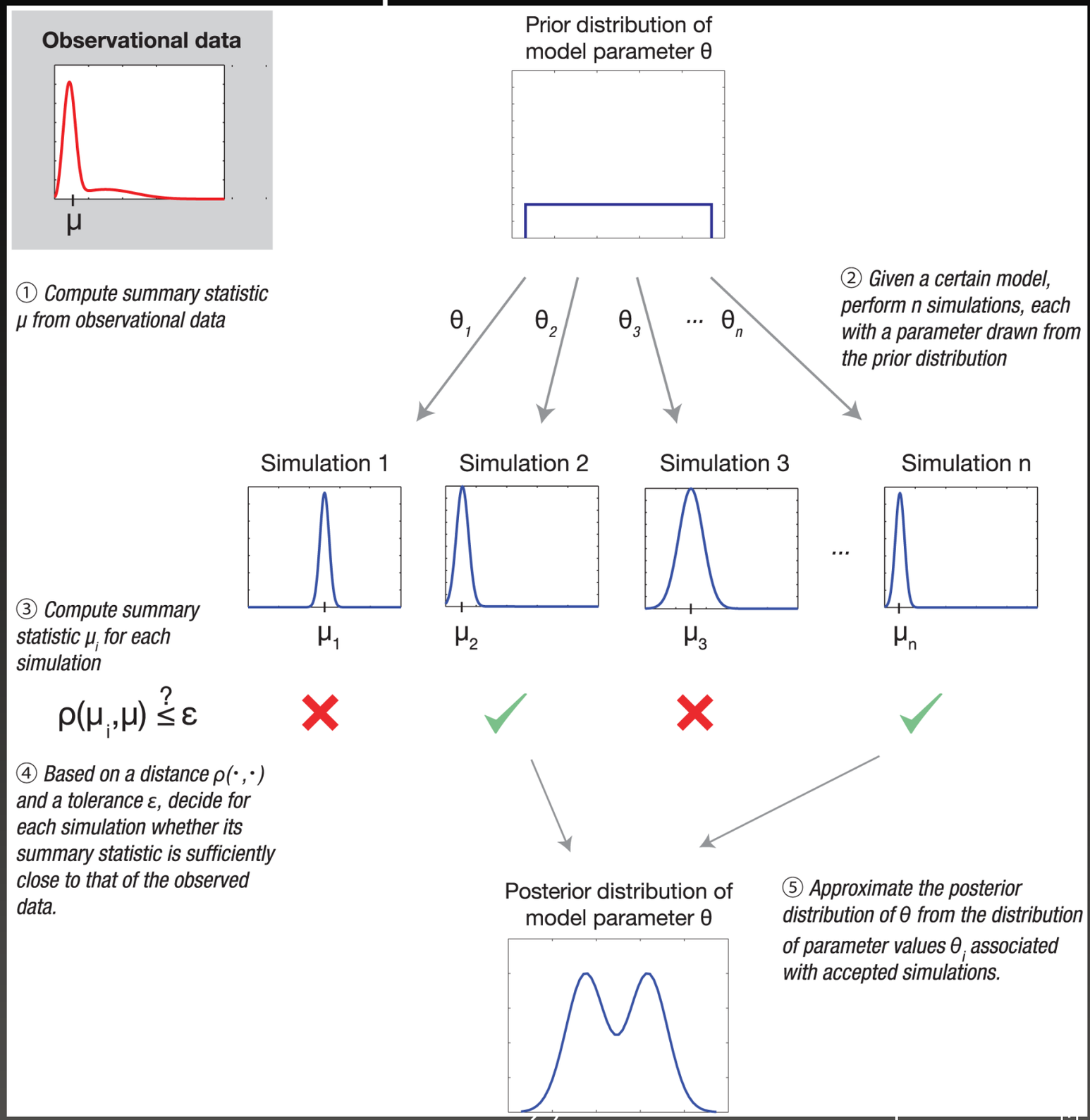


Image credit: Sunnaker et al. 2013

Changing LFI Landscape

- 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 [e.g., Papamakarios et al, 2016; Lueckmann et al, 2016; Izbicki et al, 2019; Greenberg et al, 2019]
  - Likelihoods [e.g., Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]
  - Likelihood ratios [e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]

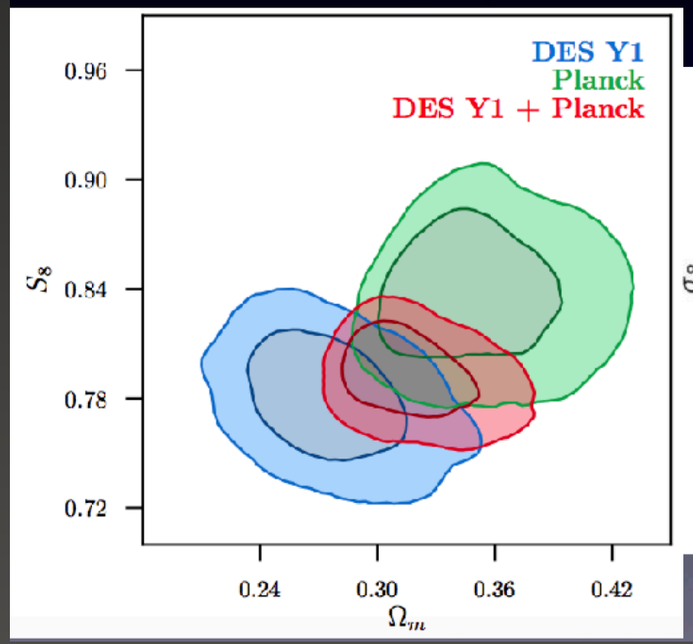
- These new training-based approaches provide amortized inference. Can handle complex high-dimensional data without relying on summary statistics.

So What's Missing?

- Statistical tests and confidence sets are the hallmarks of scientific inference but have not received much attention in LFI.
- Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, can we infer the true parameters θ with valid measures of uncertainty for small n?

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$

DES collaboration, Abbott+17



KiDS, Joudaki+17

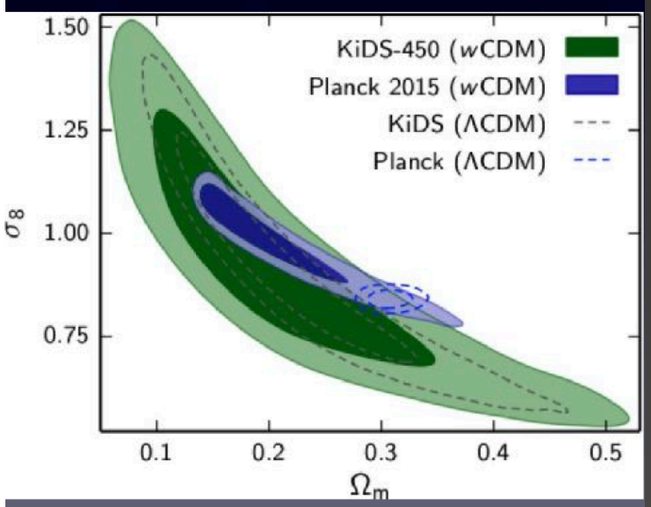


Image credit: Rachel Mandelbaum (Cosmo21)

A New Inference Machinery for Frequentist LFI

- Bridges ML with classical statistics to provide:

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

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

  (iii) modular procedures: compatible with any test statistic and different types of data

Preliminary work presented at ICML 2021 by Nic Dalmasso

https://arxiv.org/abs/2002.10399

Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting

Niccolò Dalmasso [1] Rafael Izbicki [2] Ann B. Lee [1]

Abstract

Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that allow scientists to make inferences about the underlying process that generated the observed data. A key question is whether one can still construct hypothesis tests and confidence sets with proper coverage and high power in a so-called likelihood-free inference (LFI) setting; that is, a setting where the likelihood is not explicitly known but one can forward-simulate observable data according to a stochastic model. In this paper, we present ACORE (Approximate Computation via Odds Ratio Estimation), a frequentist approach to LFI that first formulates the classical likelihood ratio test (LRT) as a parametrized classification problem, and then uses the equivalence

1. Introduction

Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that relate observed data to properties of the underlying statistical model. Most frequentist procedure with good statistical performance (e.g., high power) require explicit knowledge of a likelihood function. However, in many science and engineering applications, complex phenomena are modeled by forward simulators that *implicitly* define a likelihood function: For example, given input parameters θ, a statistical model of our environment, climate or universe may combine deterministic dynamics with random fluctuations to produce synthetic data $\mathbf{X}$. Simulation-based inference without an explicit likelihood is called *likelihood-free inference* (LFI).

The literature on LFI is vast. Traditional LFI methods, such as Approximate Bayesian Computation (ABC; Beaumont et al. 2002; Marin et al. 2012; Sisson et al. 2018), estimate posteriors by using simulations sufficiently close to

Equivalence of Tests and Confidence Sets

Key ingredients:

- data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$
- a test statistic, such as the likelihood ratio statistic $\mathrm{LR}(\mathcal{D}; \theta_0)$
- an α-level critical value $C_{\theta_0,\alpha}$

$$\text{Reject the null hypothesis } H_0 \text{ if } \mathrm{LR}(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$$

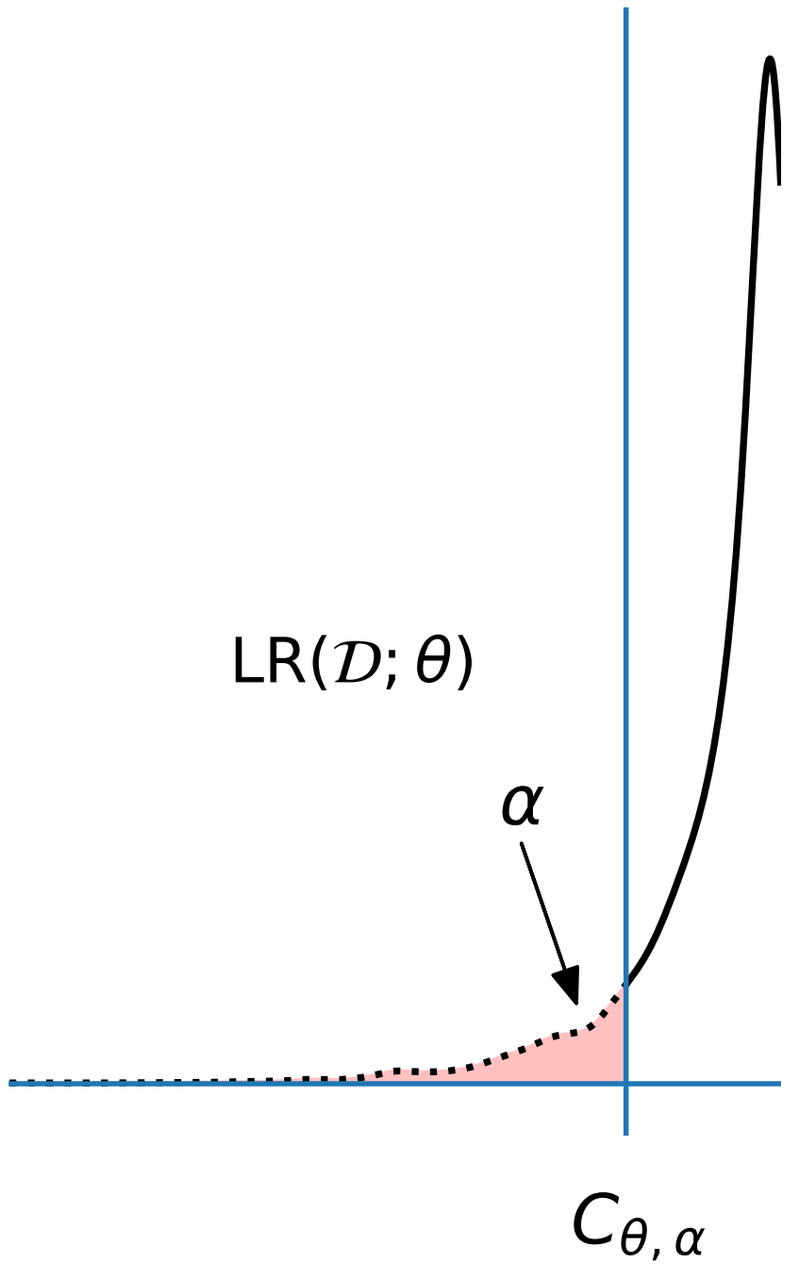
Theorem (Neyman 1937)

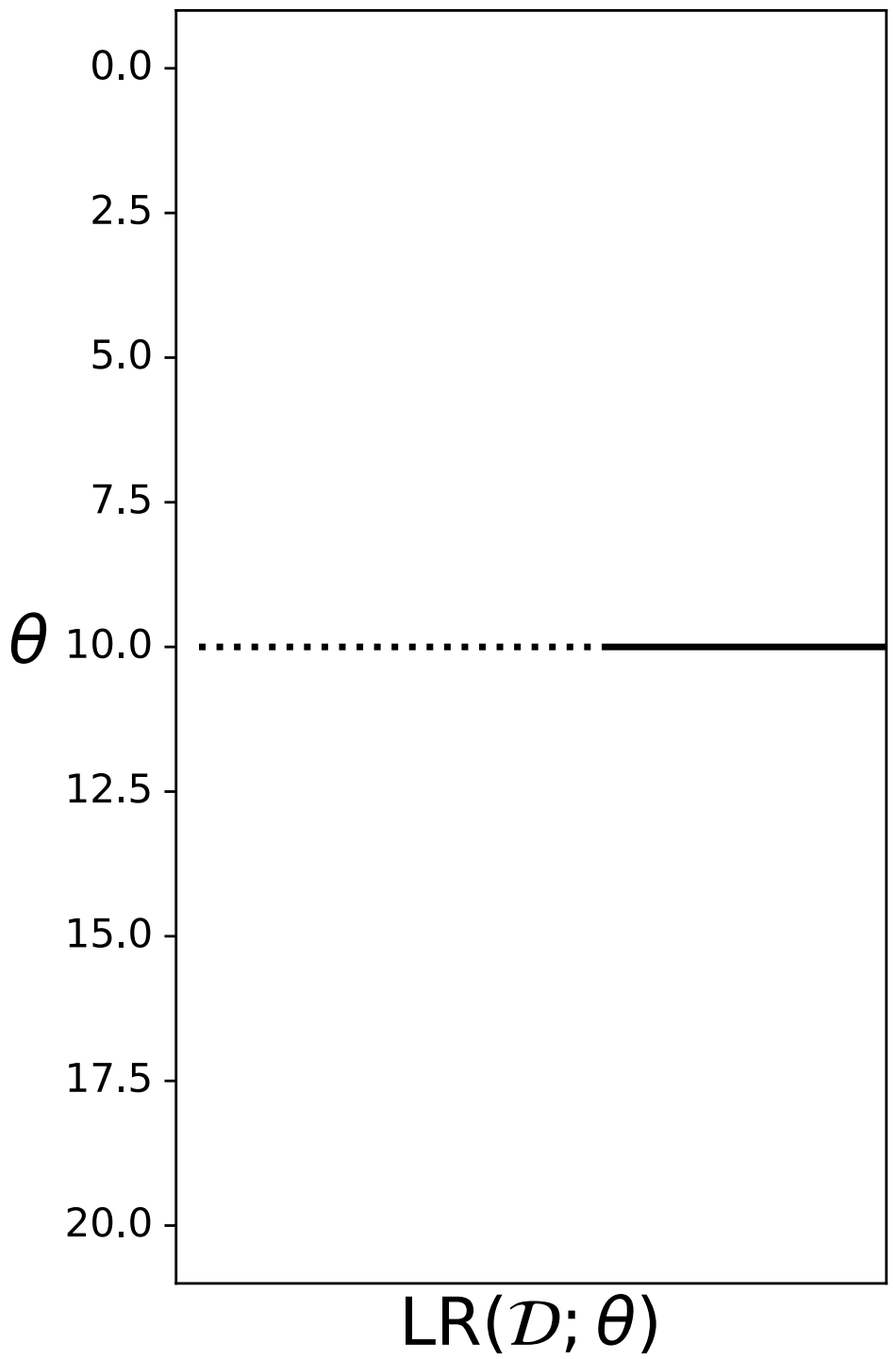
Constructing a $1 - \alpha$ confidence set for θ is equivalent to testing

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

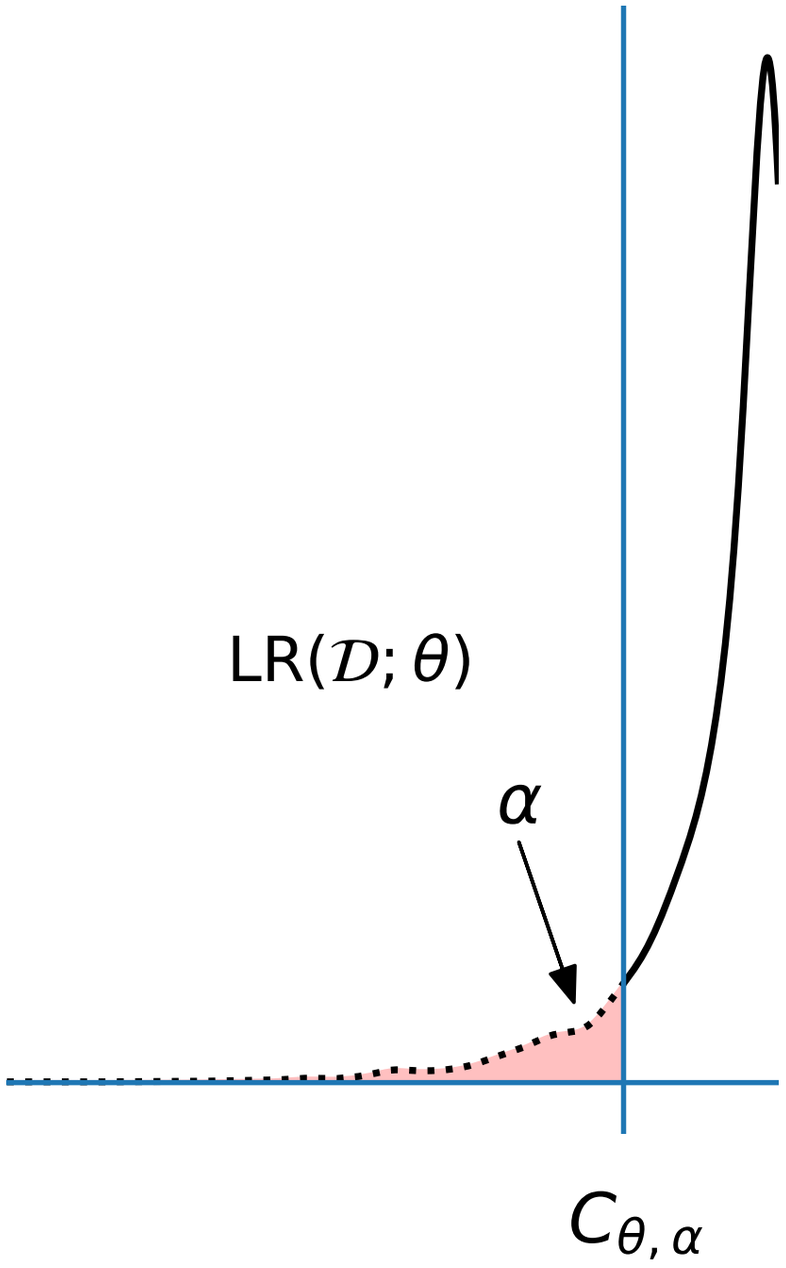
for every θ_0 in the parameter space.

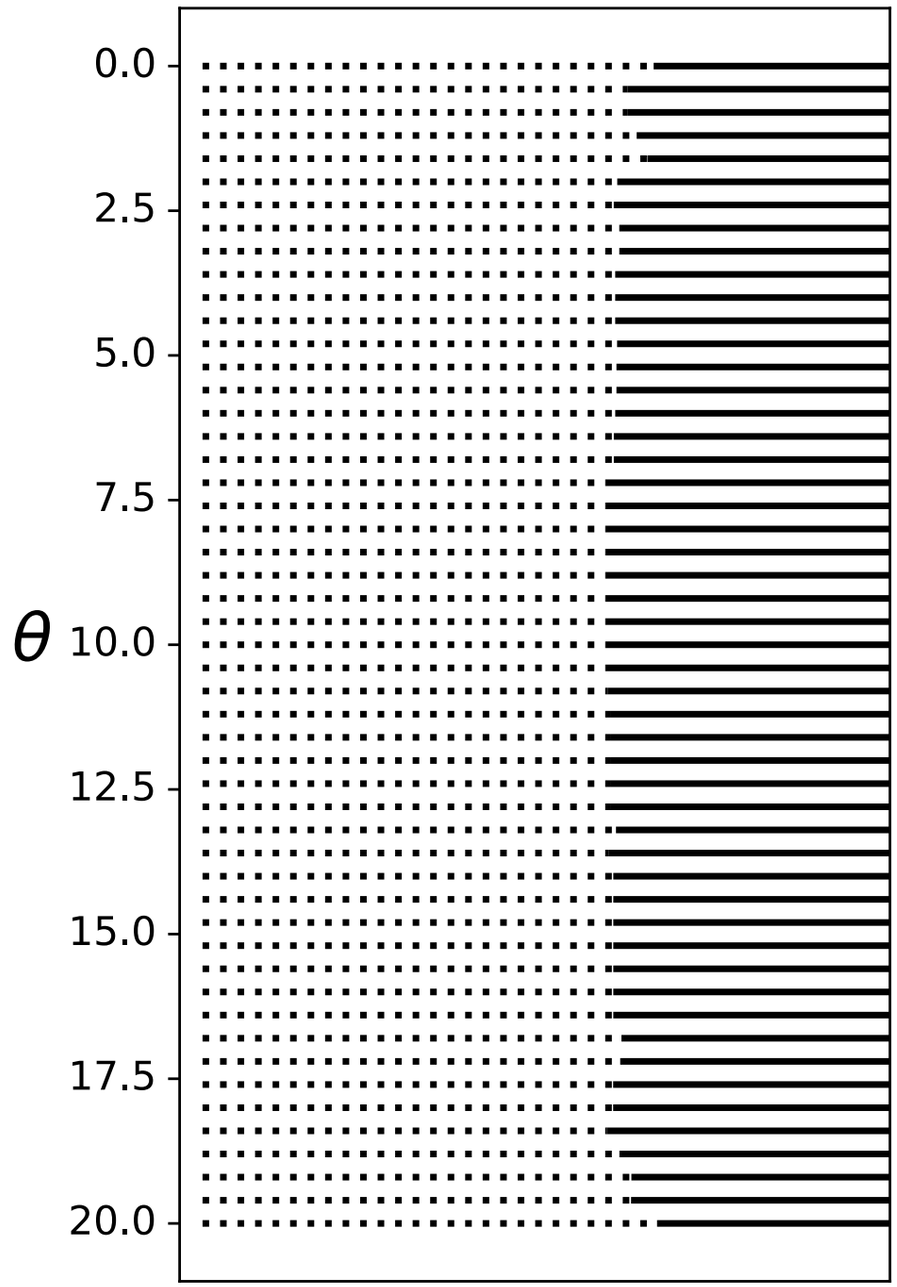
1. Fixed θ. Find the rejection region for test statistic Λ.



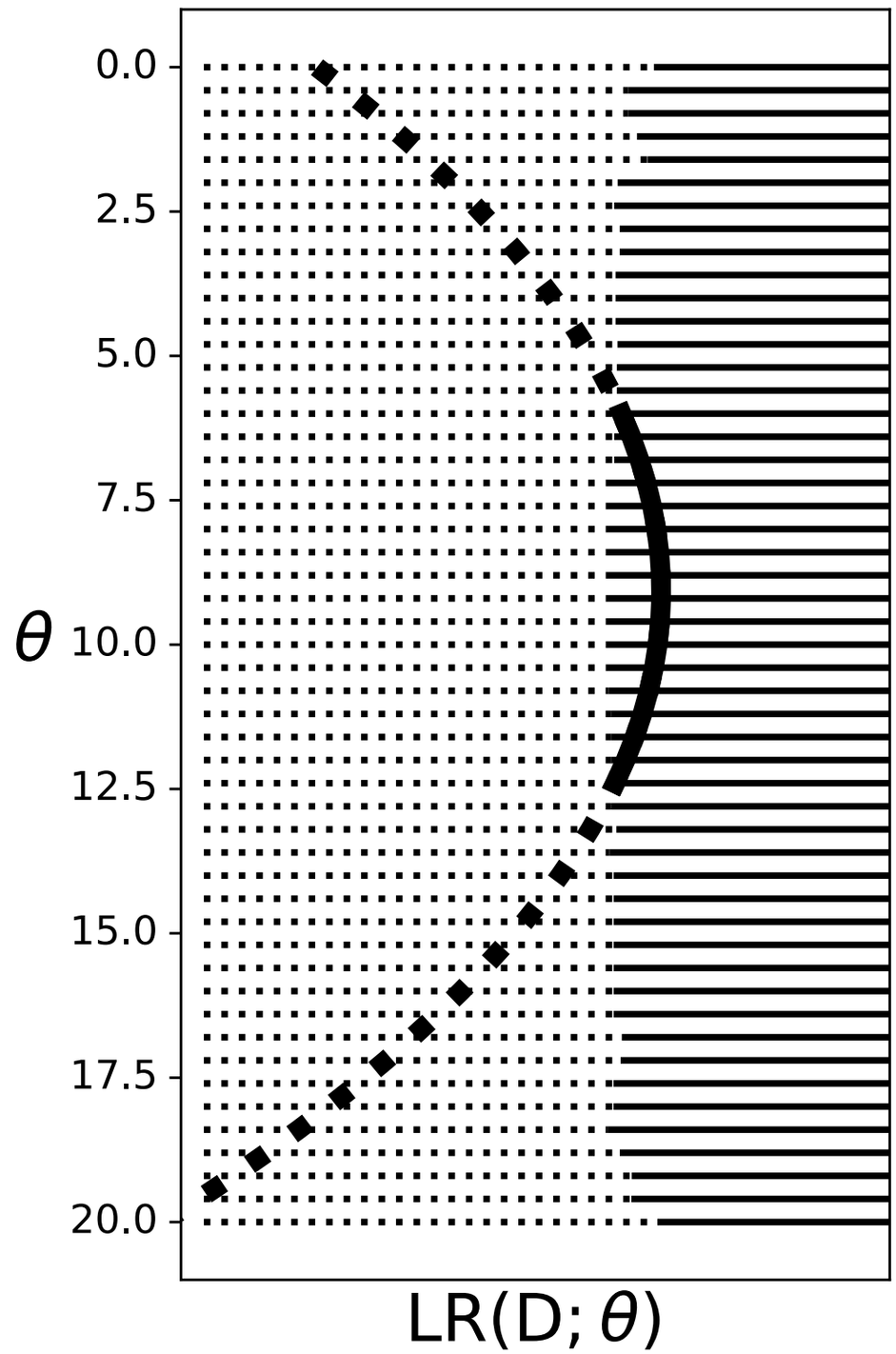


2. Repeat for every θ in parameter space.

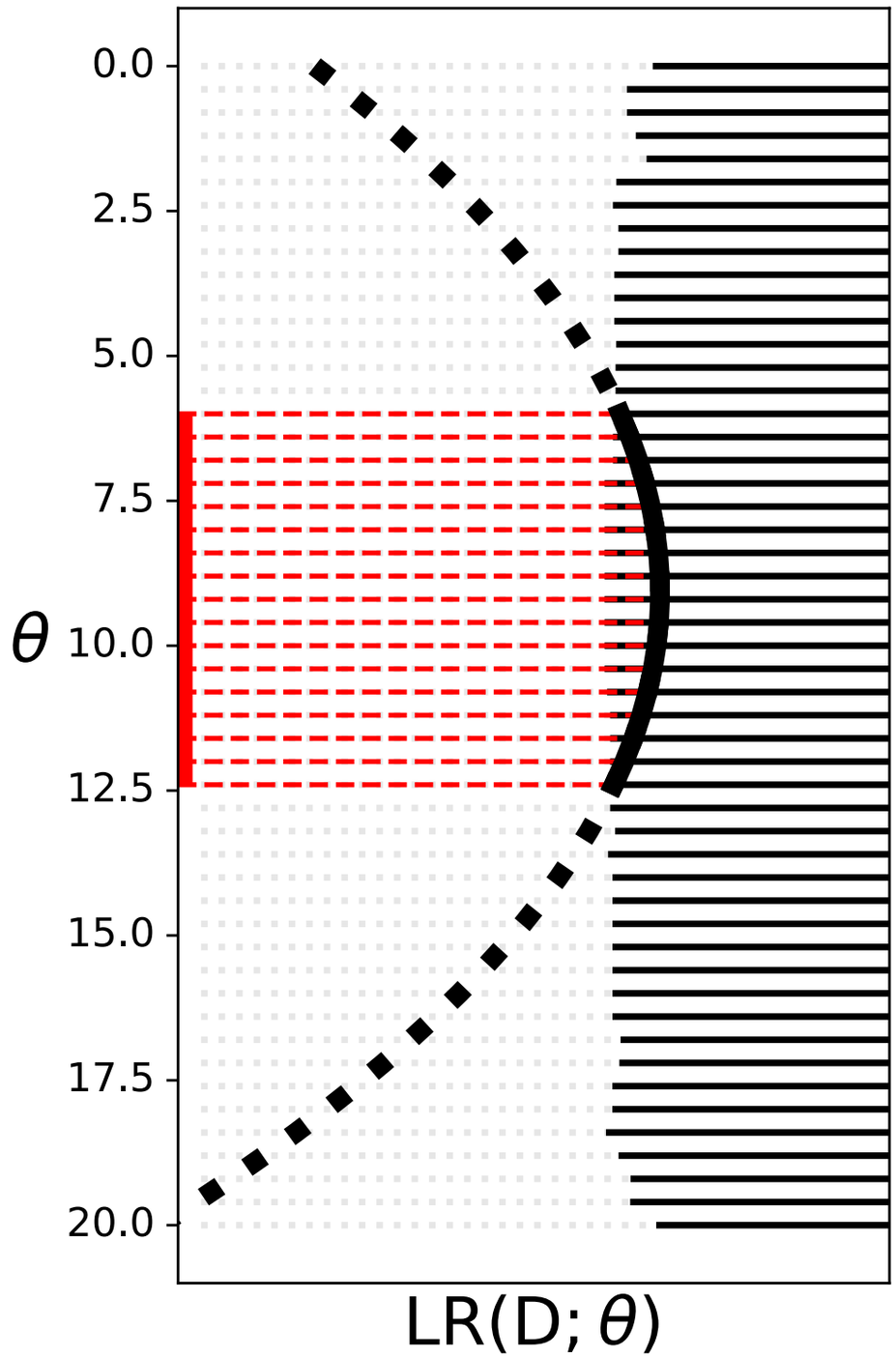




3. Observe data $\mathcal{D} = \mathrm{D}$. Calculate $\Lambda(\mathrm{D}; \theta)$.



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



How do we turn the construction into practical procedures?

"Wrinkle": The Neyman construction requires one to test

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

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

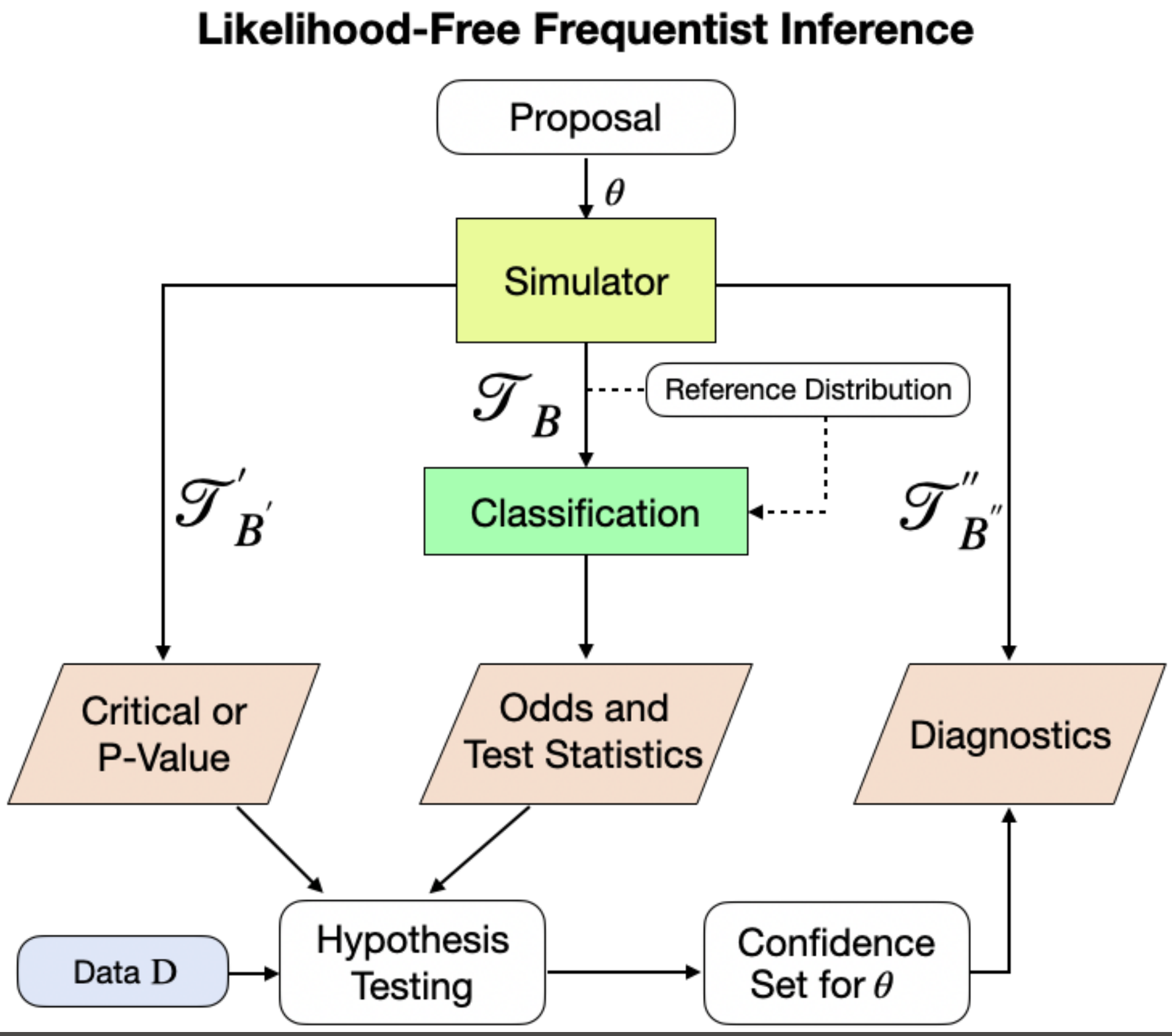
Key Realization: The main inferential quantities like

1. the **test statistic** $\tau(\mathcal{D}; \theta_0)$,
2. the **critical value** $C_{\theta_0,\alpha}$ or the p-value $p(D; \theta_0)$ of the test
3. the **coverage** $\mathbb{P}[\theta_0 \in R(\mathcal{D})]$ of the confidence set

are conditional distribution functions which often vary smoothly as a function of the (unknown) parameters.

Rather than relying solely on samples at fixed parameter settings (standard Monte Carlo approach), we can interpolate across the parameter space with ML algorithms.

Our Inference Machinery



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 Statistic Based on Odds Ratios (ACORE)

Odds Ratio: $\mathbb{OR}(\mathbf{x}; \theta_0, \theta_1) = \frac{\mathbb{O}(\mathbf{x};\theta_0)}{\mathbb{O}(\mathbf{x};\theta_1)}$

Interpretation: Chance that $\mathbf{x}$ was generated from θ_0 rather than θ_1.

Suppose we want to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \notin \Theta_0$$

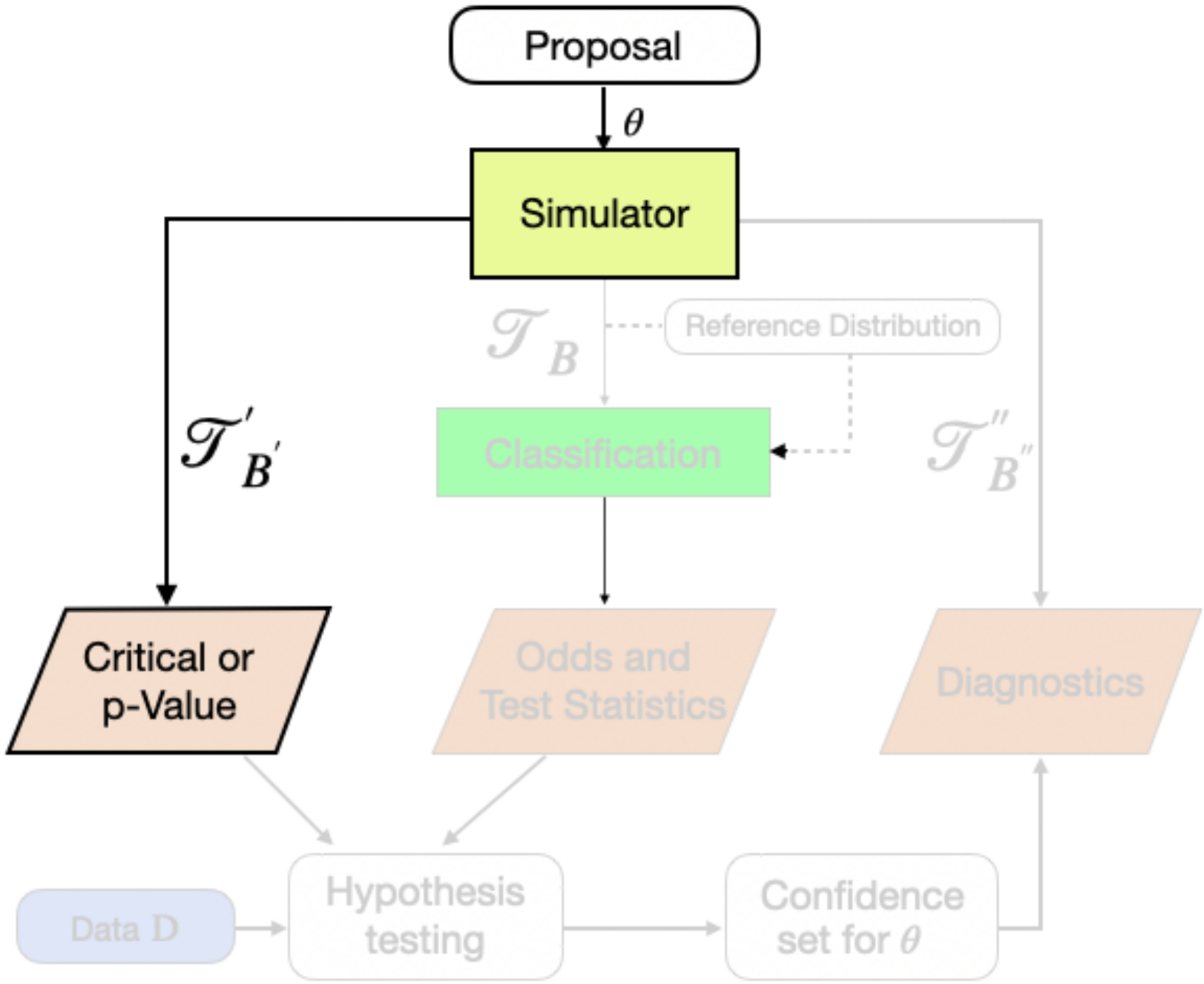
We define ACORE test statistic:

$$\widehat{\tau}(\mathcal{D}; \Theta_0) := \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^{n} \log \left(\widehat{\mathbb{OR}}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1) \right)$$

Theorem (Fisher's Consistency)

$$\text{If} \quad \widehat{\mathbb{P}}(Y = 1 | \theta, \mathbf{x}) = \mathbb{P}(Y = 1 | \theta, \mathbf{x}) \, \forall \theta, \mathbf{x} \implies \widehat{\tau}(\mathcal{D}; \Theta_0) = \text{LR}(\mathcal{D}; \theta_0)$$

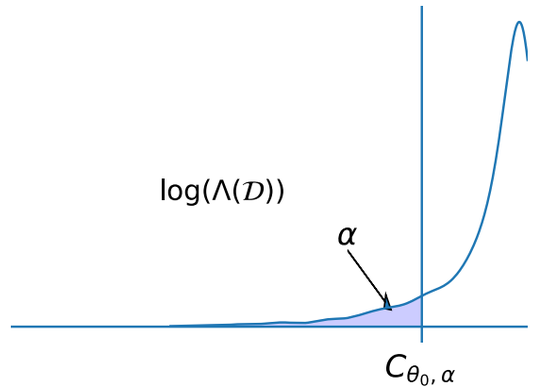
Left Branch: Estimate Critical Values or P-Values



We use B' simulations to estimate critical values.

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

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

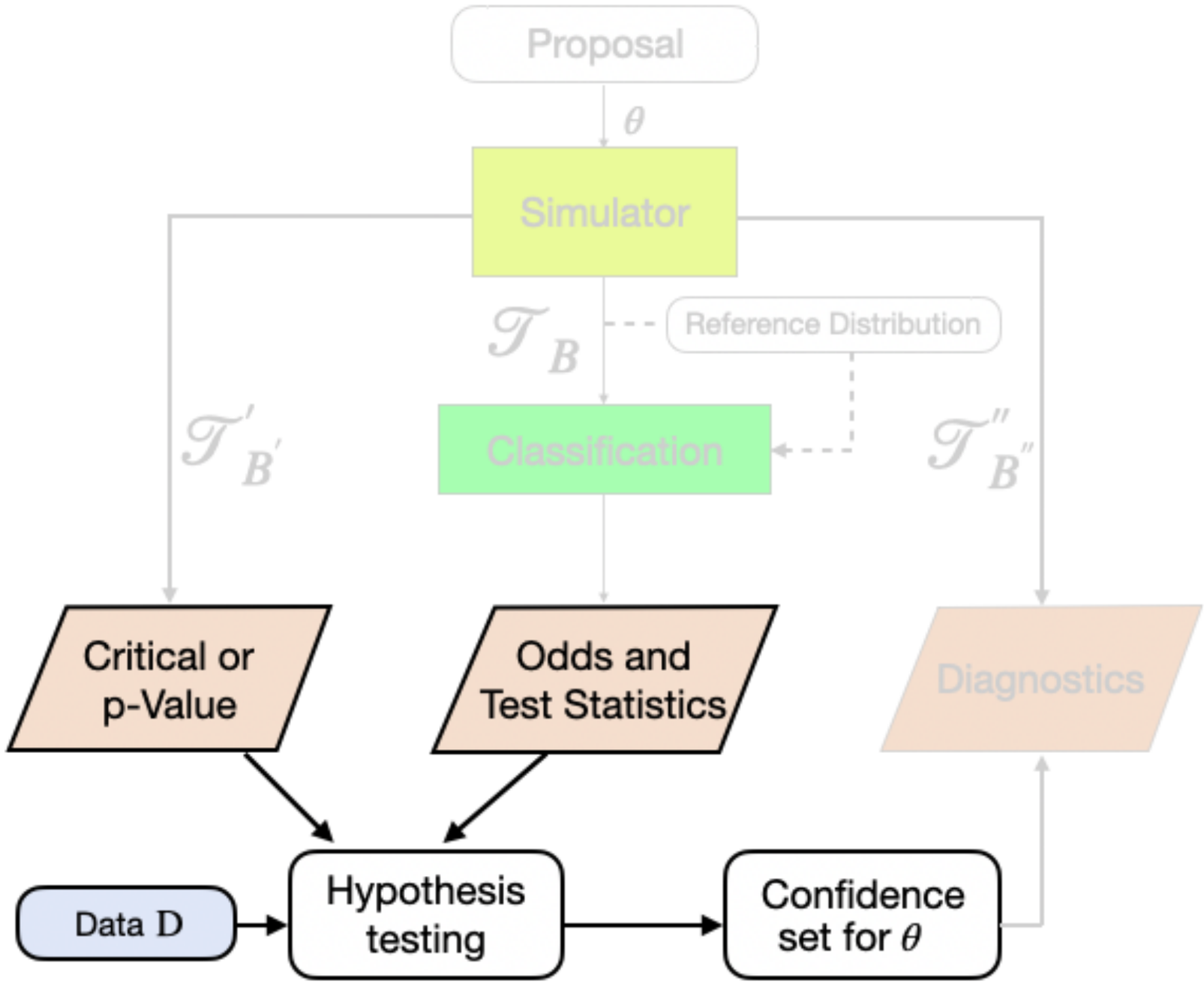


$$C_{\theta_0,\alpha} = \arg\sup_{C \in \mathbb{R}} \{C : \mathbb{P}\left(\tau(\mathcal{D}; \theta_0) < C \mid \theta = \theta_0\right) \leq \alpha\}.$$

Problem: Need to estimate $\mathbb{P}\left(\tau(\mathcal{D}; \theta) < C \mid \theta\right)$ for every $\theta \in \Theta$.

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

Construct Confidence Set via Neyman Inversion

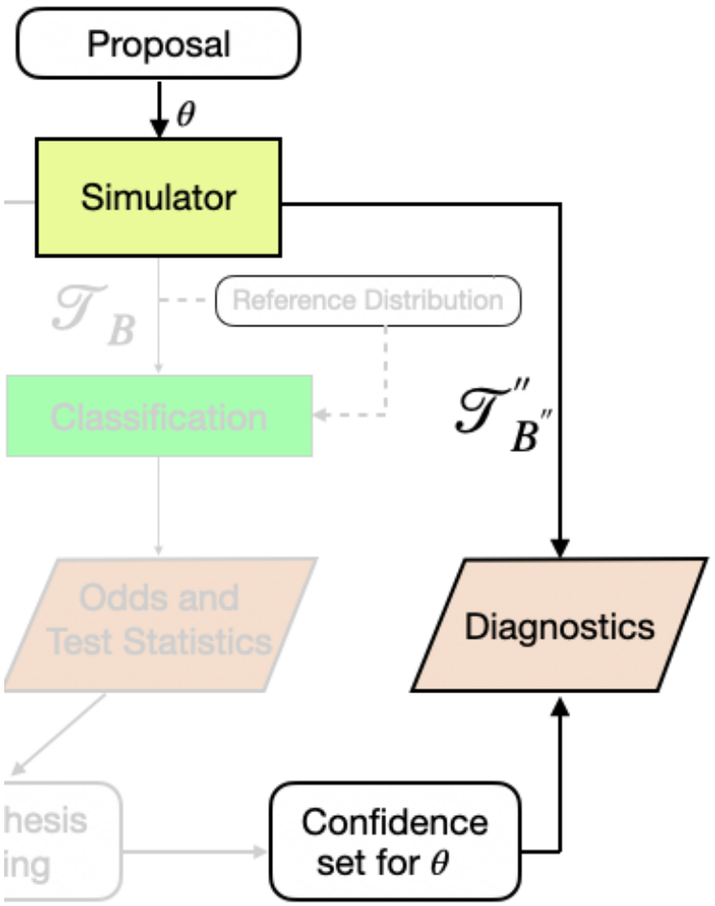


Right Branch: Checking Actual Coverage Across Θ

How do we check coverage $\mathbb{P}(\theta \in R(\mathcal{D}))$ as a function of $\theta \in \Theta$?

Note: $\mathbb{P}(\theta \in R(\mathcal{D})|\theta) = \mathbb{E}[\mathbb{I}(\theta \in R(\mathcal{D}))|\theta]$

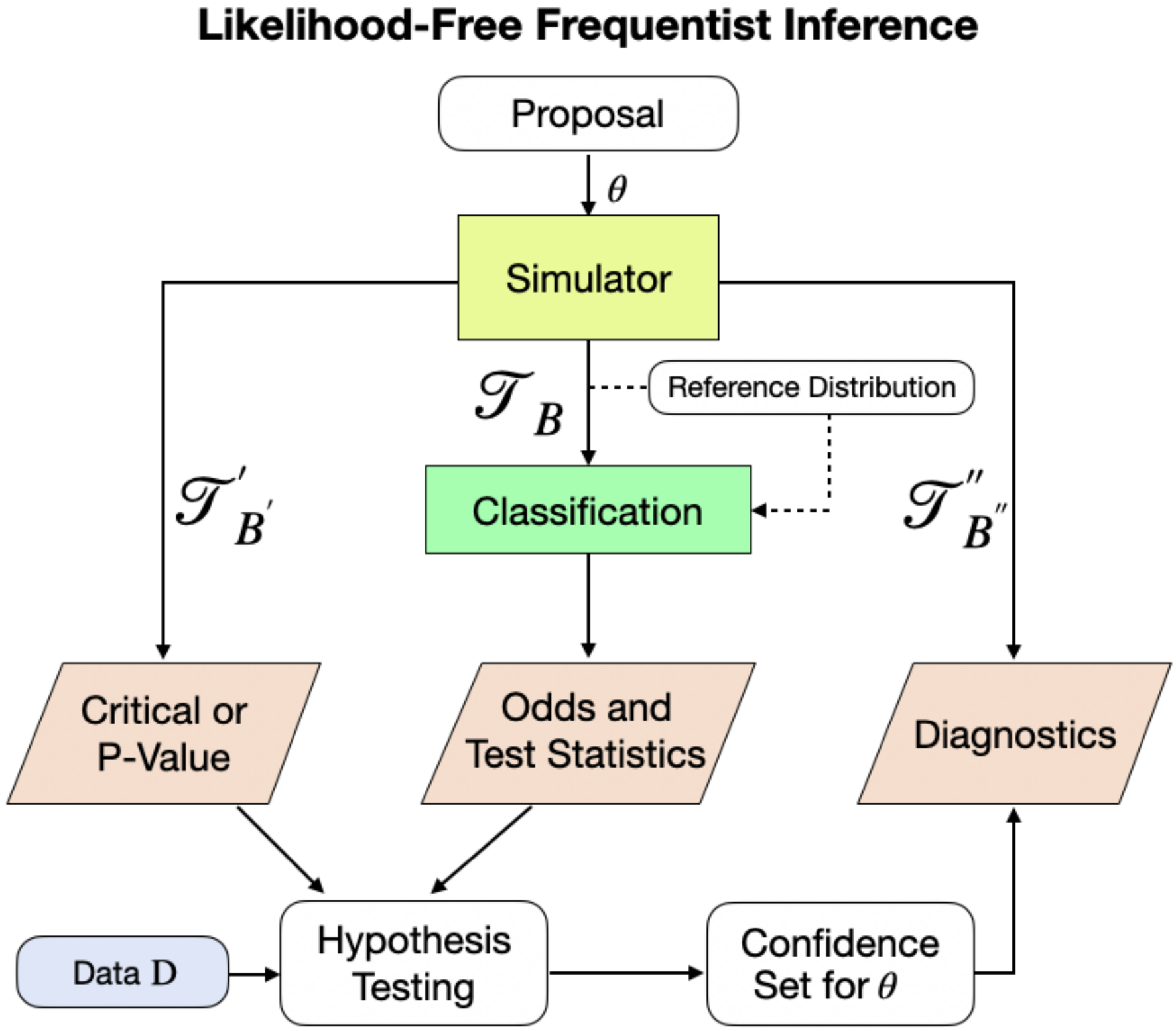
That is, we can estimate empirical coverage across the entire parameter space by regression (probabilistic classification):



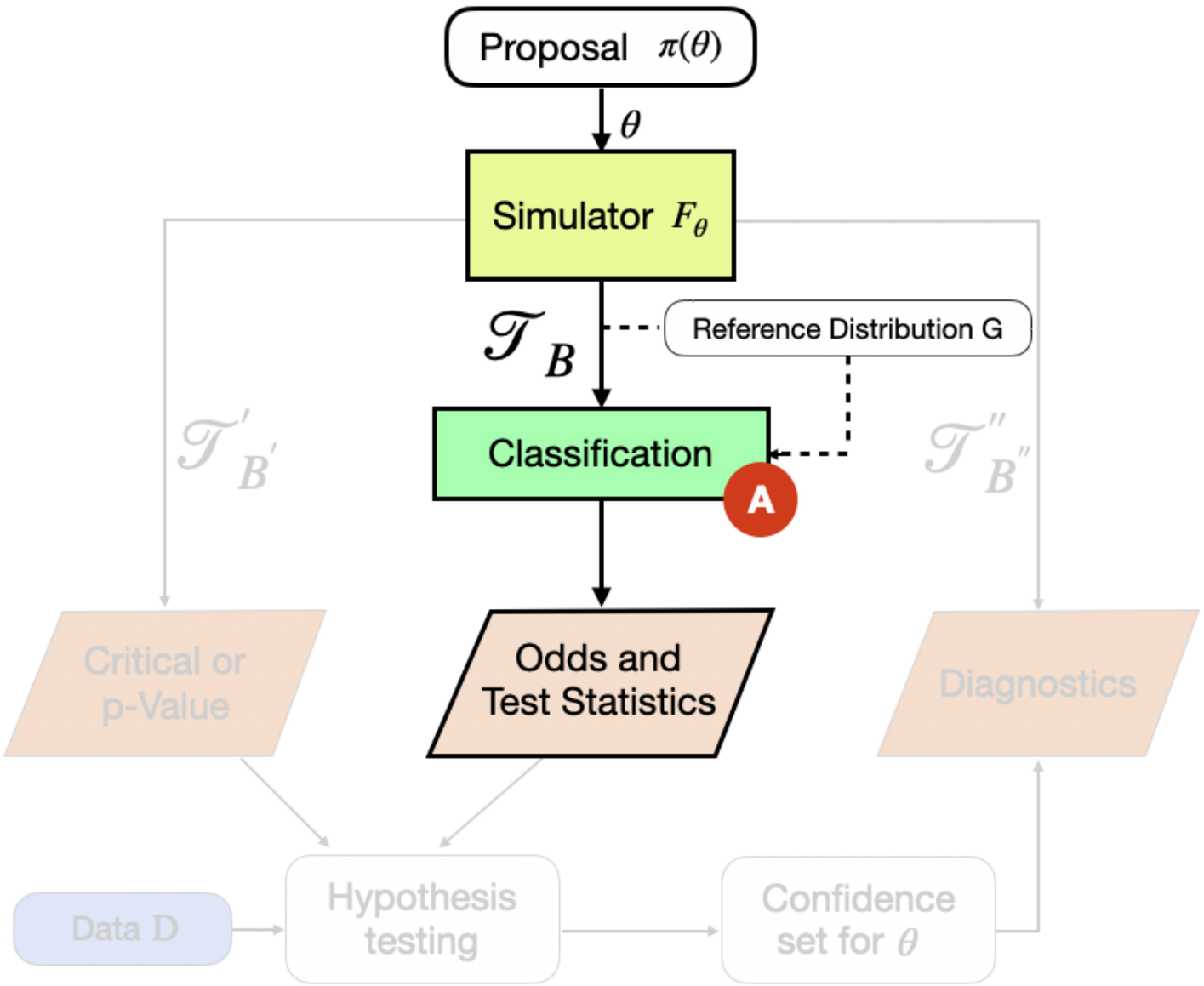
1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $R(\mathcal{D}_i)$
3. For $\{\theta_i, R(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i = \mathbb{I}(\theta_i \in R(\mathcal{D}_i))$ on θ_i

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

How do we implement our LFI machinery in practice?

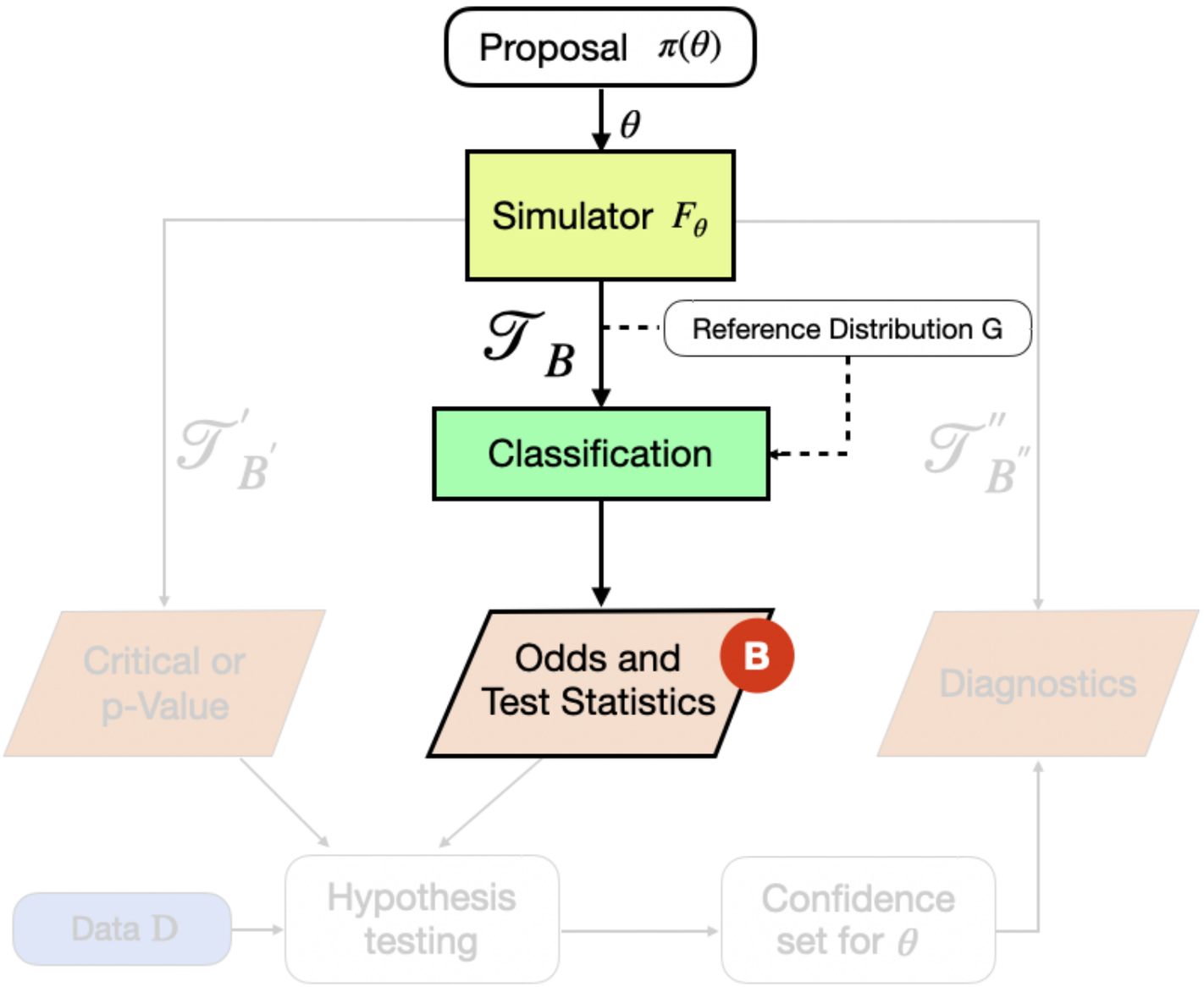


A Practical Strategy



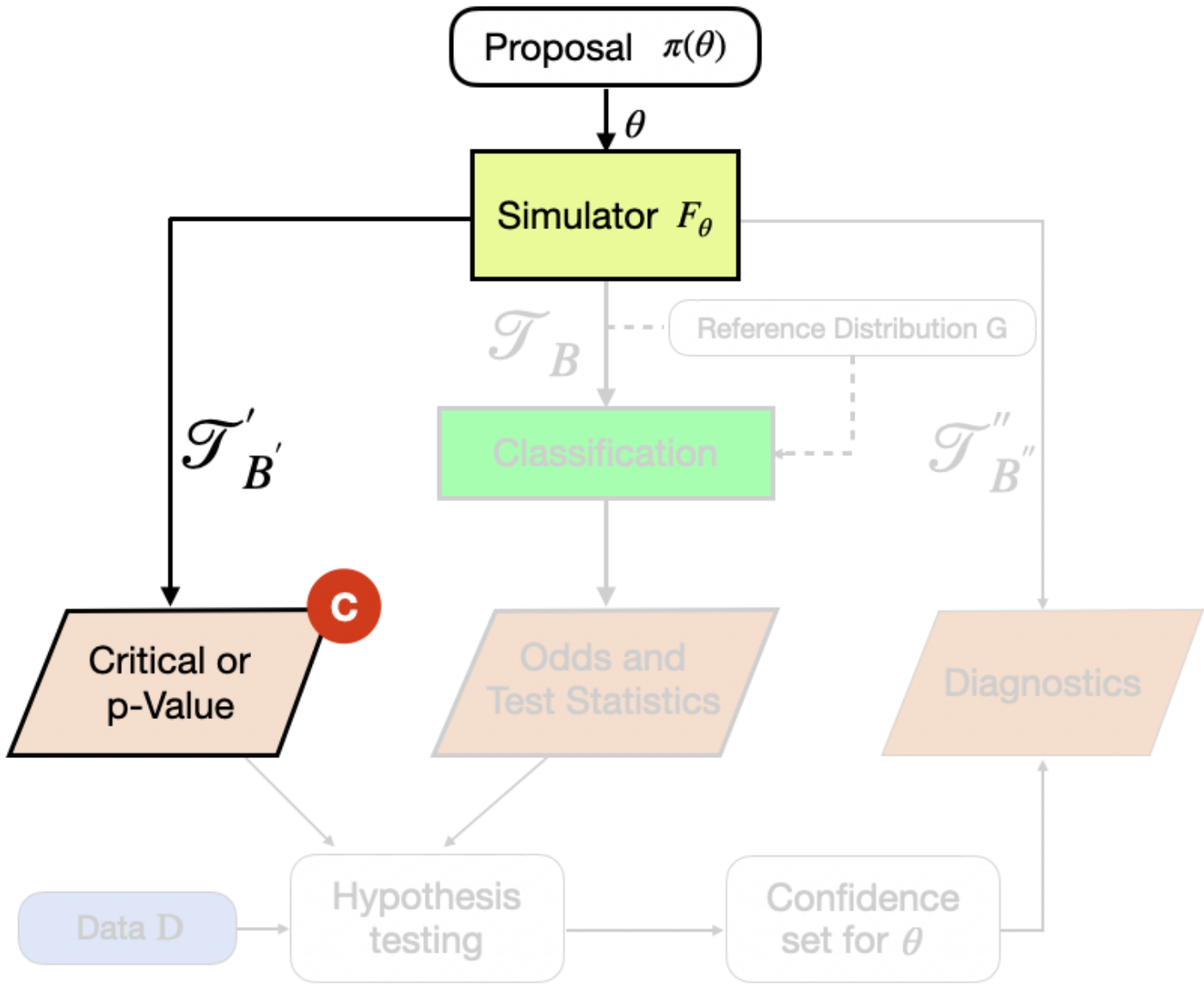
(A) Use the cross-entropy on a held-out set to select probabilistic classifier and sample size B for learning the odds

A Practical Strategy



(B) Compute the optimization step in `ACORE` with available computational budget.

A Practical Strategy



(C) Use our diagnostic tool to determine the quantile regression algorithm and sample size B' to achieve nominal coverage across Θ.

Toy Example: Signal Detection in High Energy Physics

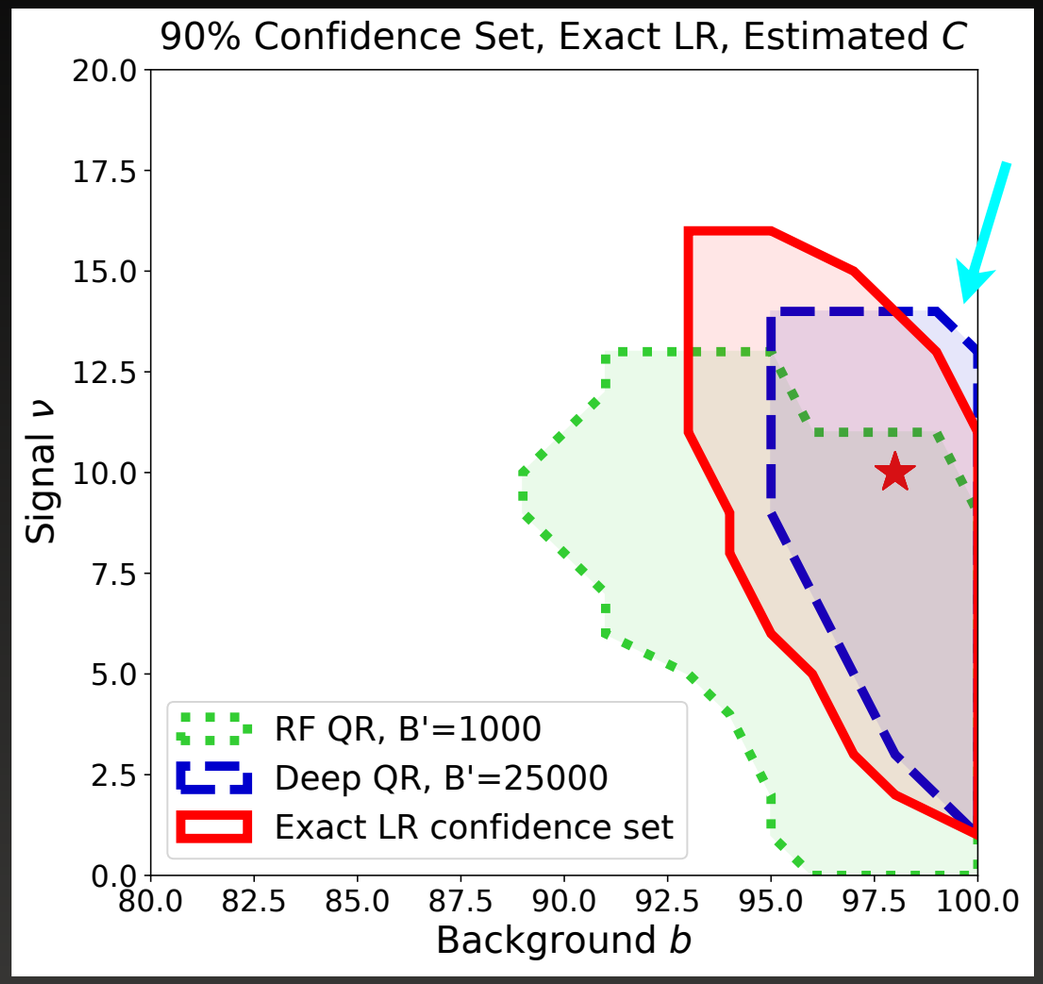
- Particle collision events counted under the presence of a background process.

$$\text{Observed data } D = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{10})$$
$$\mathbf{X} = (N, M), \text{ where } N \sim \text{Poisson}(b + \nu),\ M \sim \text{Poisson}(b)$$

- The observed data D consist of n=10 realizations of X=(N,M), where
  - N is the number of events in the signal region,
  - M is the number of events in the background/control region
- Unknown parameters:
  - intensity of signal (ν); intensity of background (b)

Constructed Confidence Set for a Particular X^{obs}

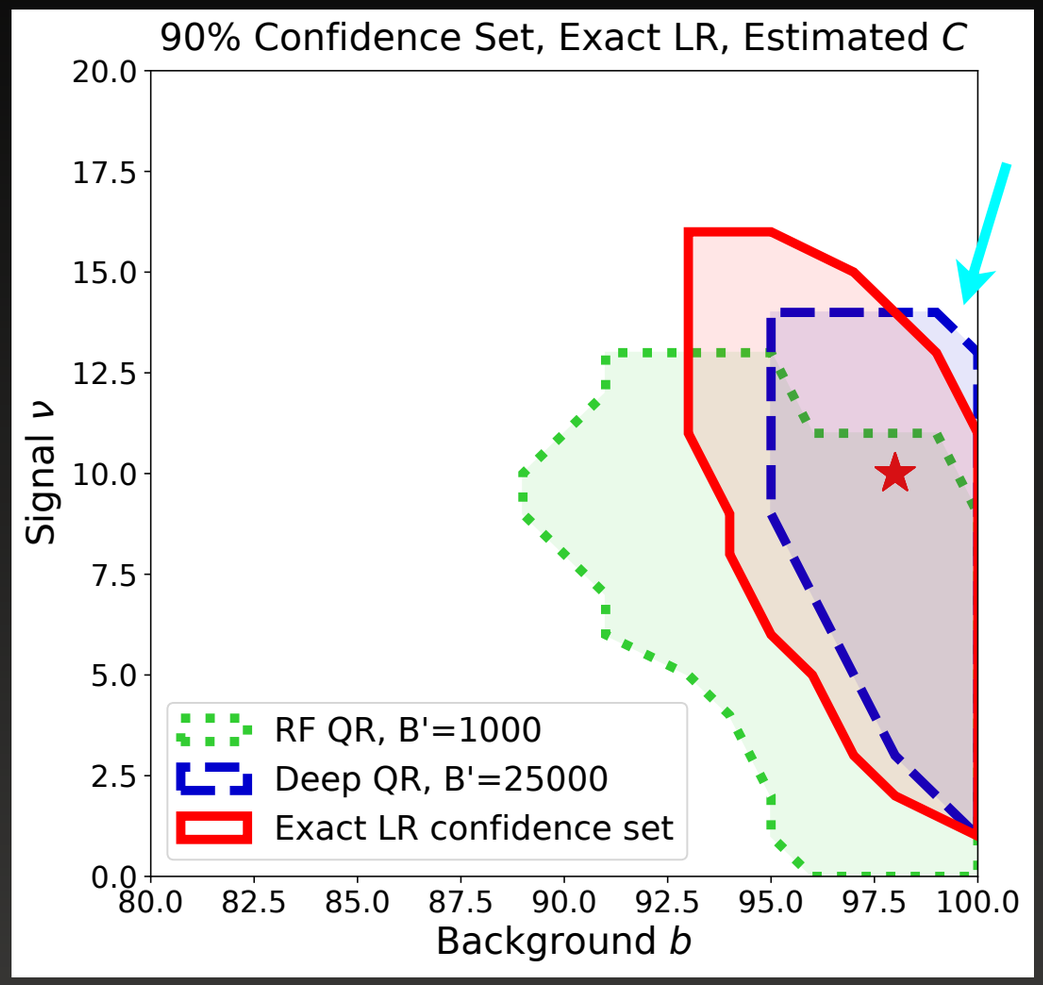
Our proposed strategy selects the BLUE confidence region

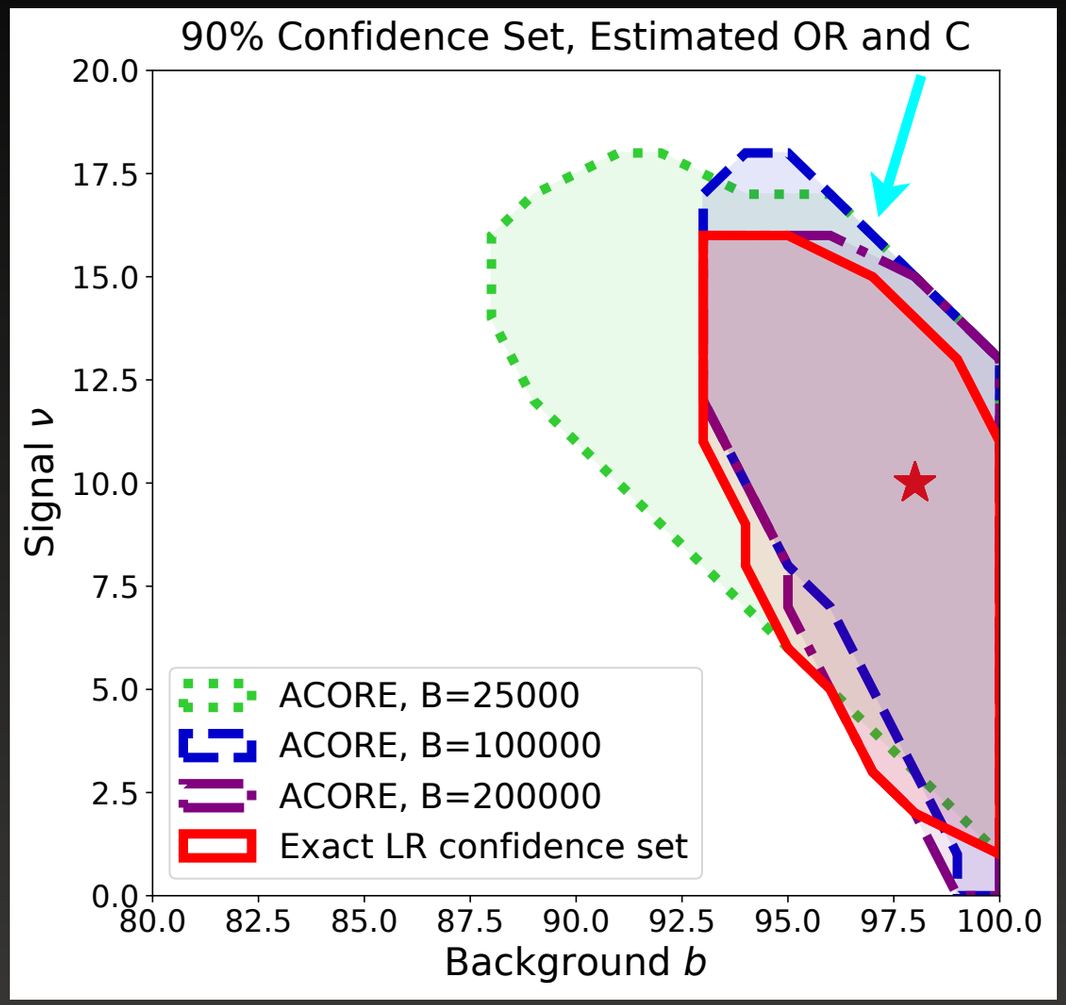


- Left: 90% confidence set computed with the exact LR statistic but estimated critical value. Estimating C can be challenging.
- Right: 90% confidence set with both estimated LR statistic and critical value. This is the true LFI setting.

Constructed Confidence Set for a Particular X^{obs}

Our proposed strategy selects the BLUE confidence region

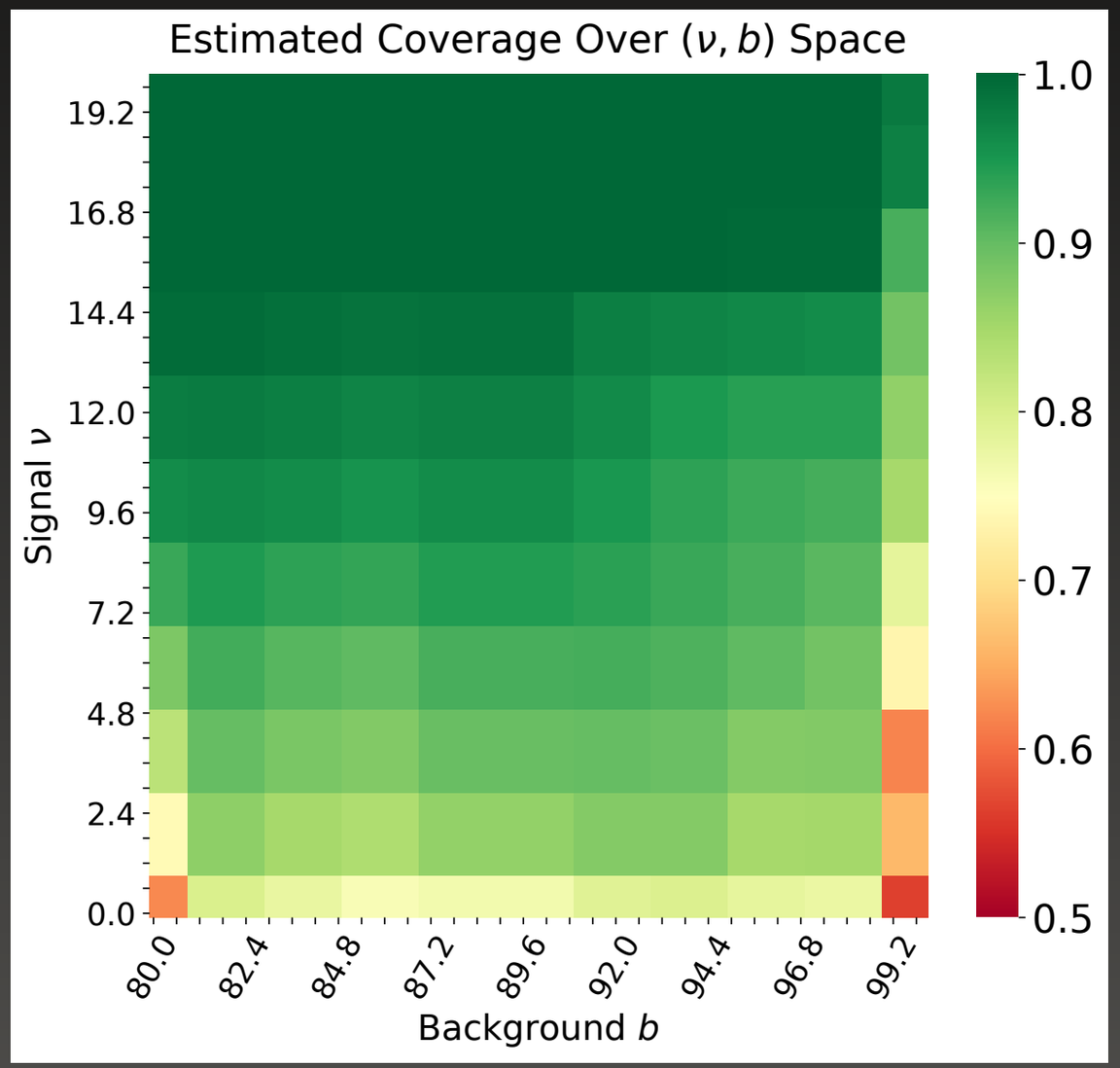




- Left: 90% confidence set computed with the exact LR statistic but estimated critical value. Estimating C can be challenging.
- Right: 90% confidence set with both estimated LR statistic and critical value. This is the true LFI setting.

Diagnostics: Do We Achieve Nominal Coverage (Type I Error Control) Across the Parameter Space?

$$\mathbb{P}[\theta_0 \in R(\mathcal{D}) \mid \theta = \theta_0] \geq 1 - \alpha \quad \text{for all } \theta_0 \in \Theta$$



Heat map of estimated coverage for a confidence set that did not pass our goodness-of-fit diagnostic

- Overall coverage of confidence set is correct (92% vs the 90% nominal coverage)
- However, the set undercovers in low-signal and high-background regions.

Take Away: Forward Problems - Validation

- We can leverage regression methods (probabilistic classifiers) to identify IF and HOW two samples differ.

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

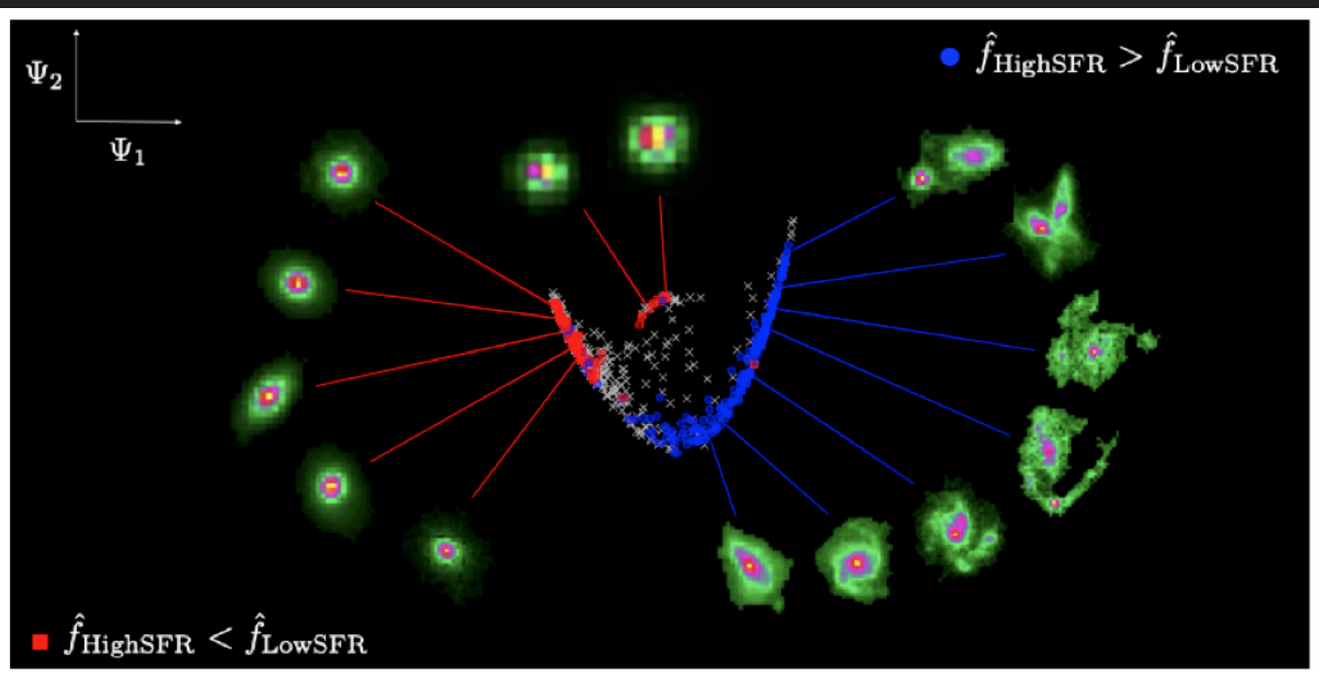
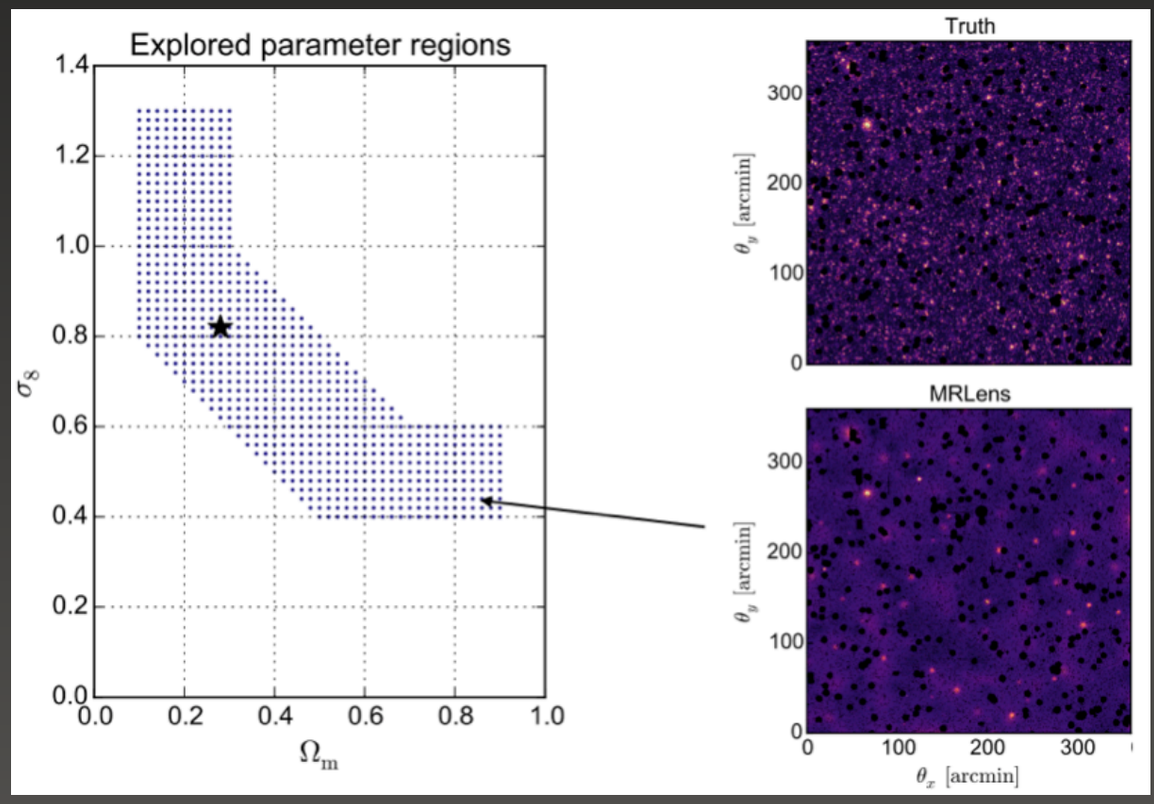
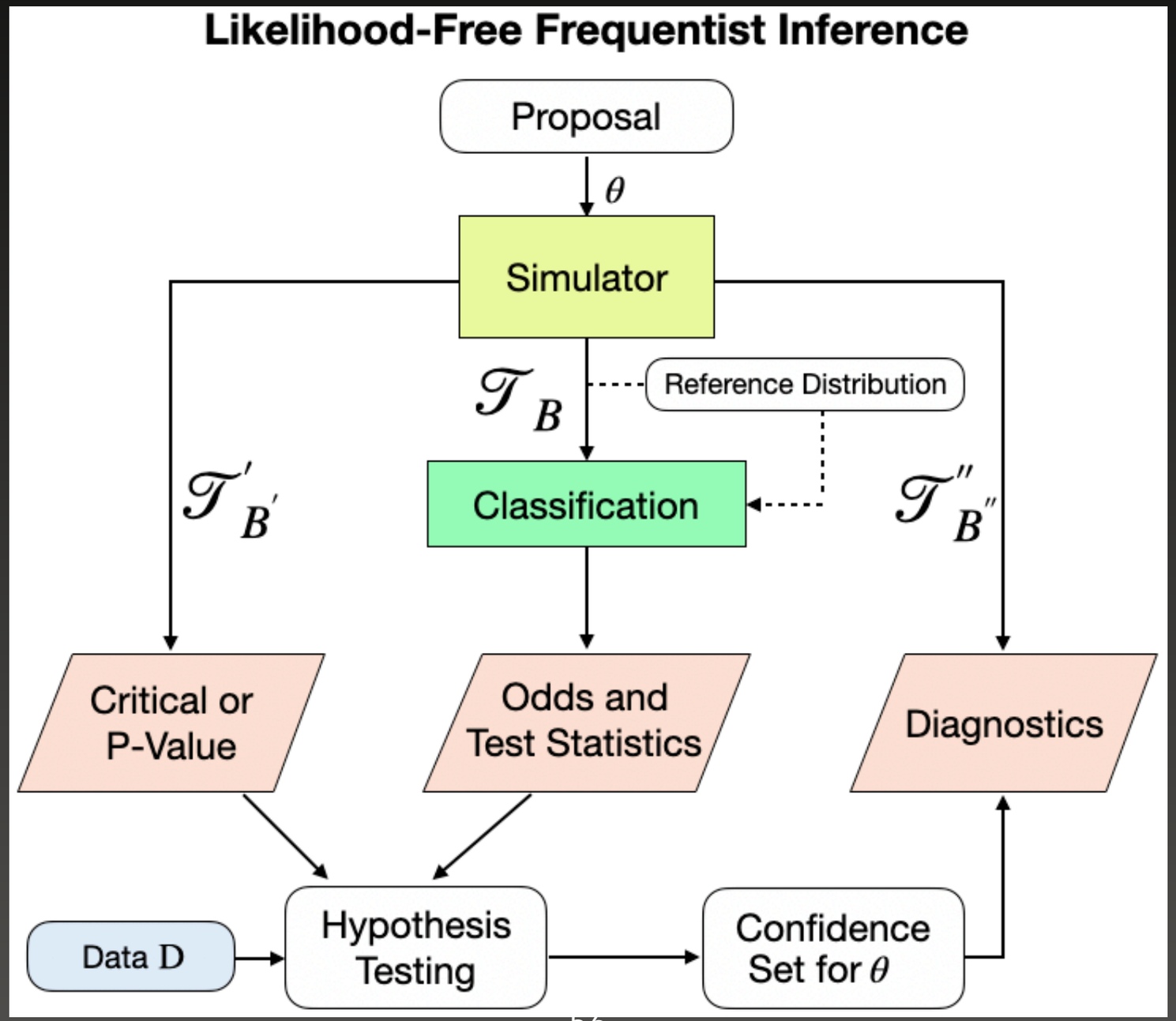


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].



Take-Away: Inverse Problem - Calibration

- We can construct confidence sets with nominal coverage, and provide diagnostics, even without a tractable likelihood.



EXTRA SLIDES START HERE

Results from Joint Analysis in 7 Dimensions:
Galaxies with significantly higher representation in low-SFR sample (top) vs in high-SFR sample (bottom)

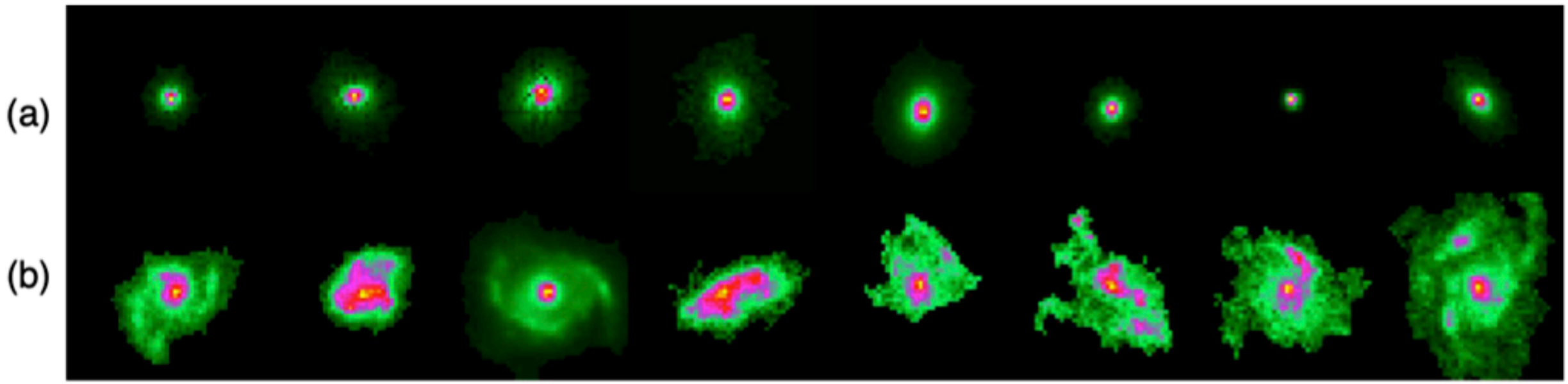
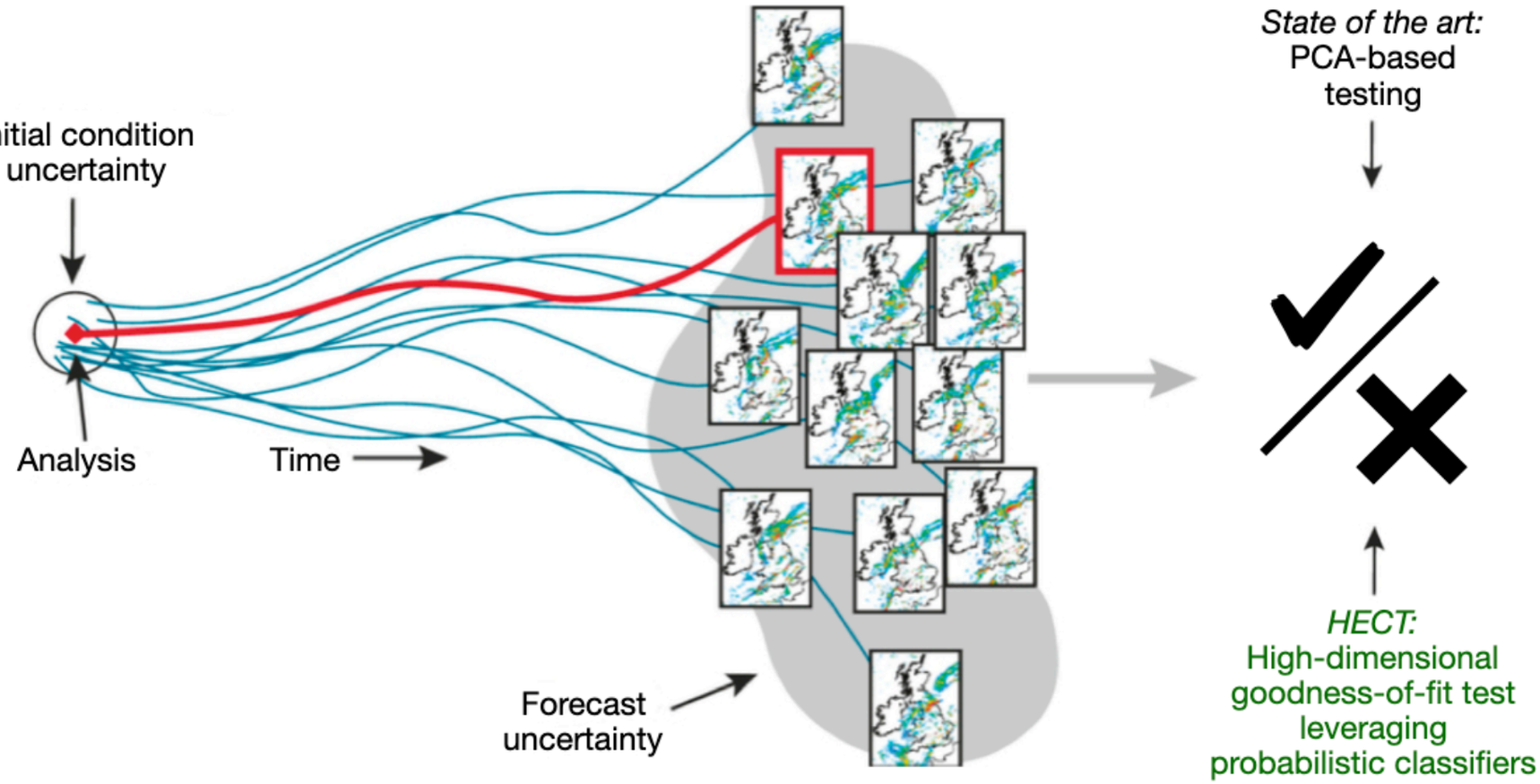


Figure 6: Galaxies in the test set with the highest significant difference $|\widehat{m}(\mathbf{x}) - \widehat{\pi}_1|$ according to our local test in feature space, Algorithm 4. (a) Galaxies that are more representative of the low-SFR sample $\mathcal{S}_0$, and (b) galaxies more representative of the high-SFR sample $\mathcal{S}_1$. The first group of galaxies presents undisturbed and concentrated morphologies, while the latter galaxies appear more extended and/or disturbed. This is in line with what is expected by astronomers when comparing actual low-SFR and high-SFR galaxies.

Quality Assurance of Simulation Models

by High-Dimensional Ensemble Consistency Testing (HECT)



[Dalmasso, Vincent, Hammerling and Lee 2020]

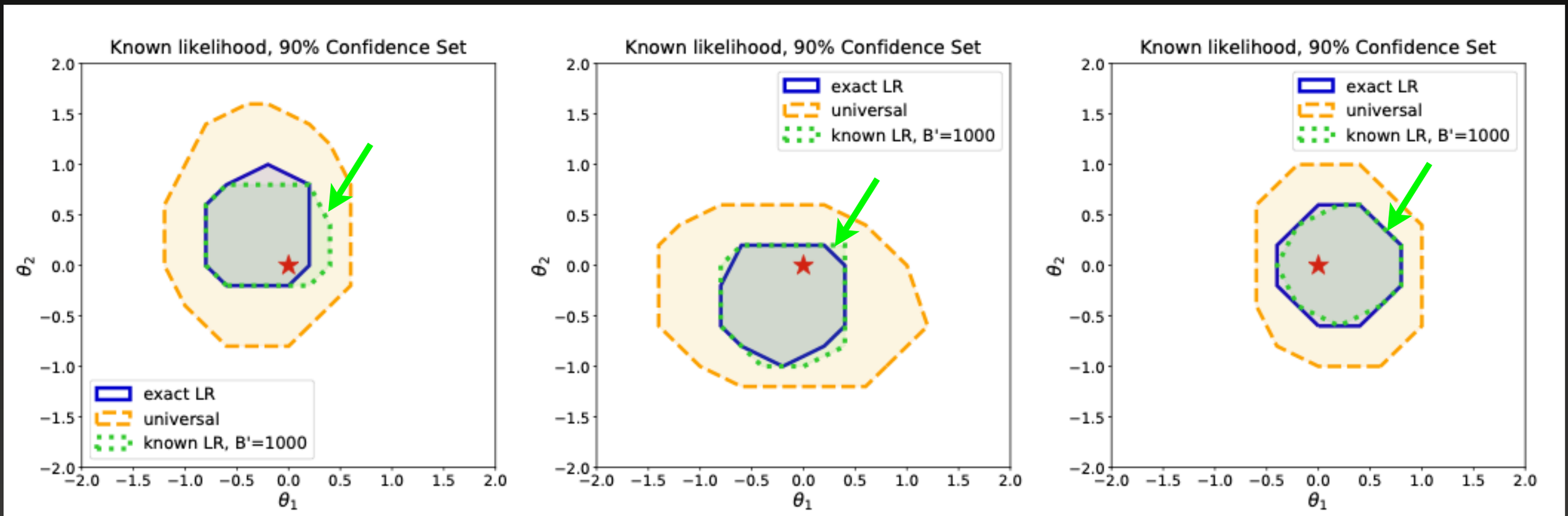
How Does Our Approach Scale?

Consider an example where the forward model is just a MVG distribution $N(\boldsymbol{\theta}, \mathbf{I}_d)$. Construct a confidence set for the unknown mean $\boldsymbol{\theta} \in \mathbb{R}^d$.

- Suppose the observed data are $\mathbf{X}_1, \ldots, \mathbf{X}_{10} \sim N(\mathbf{0}, \mathbf{I}_d)$, so $n = 10$ and $\boldsymbol{\theta} = \mathbf{0}$ (unknown parameter).
- We first assume that the likelihood can be evaluated, but that we do not know the distribution of the test statistic under the null.
- Compare our results to exact LRT and exact BF (baseline).
- Compare our results to Crossfit LRS — a "universal inference"[1] method for constructing valid finite-sample confidence sets in such settings without regularity conditions by averaging the LR over two data splits.

[1] Wasserman, Ramdas, and Balakrishnan; PNAS 2020

Known Likelihood Setting in 2D: Examples of Constructed Confidence Sets (All Valid)

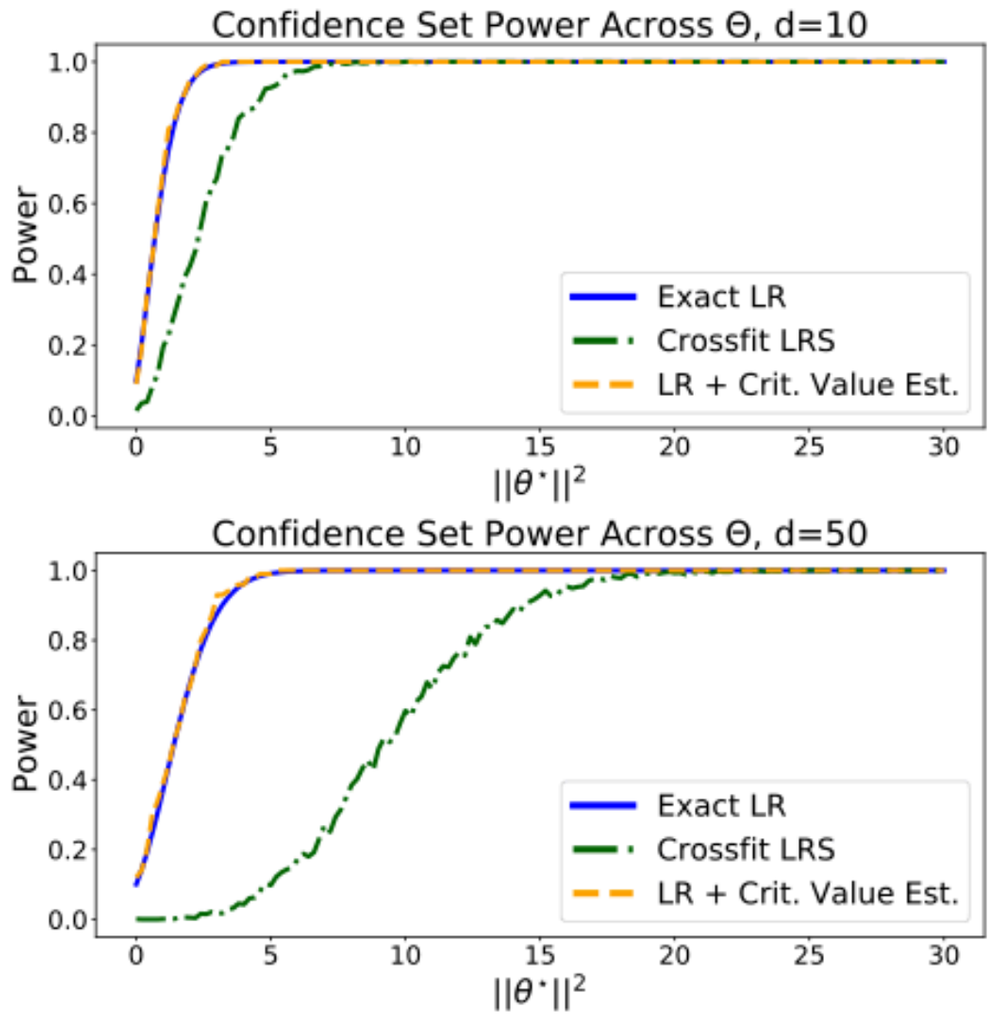


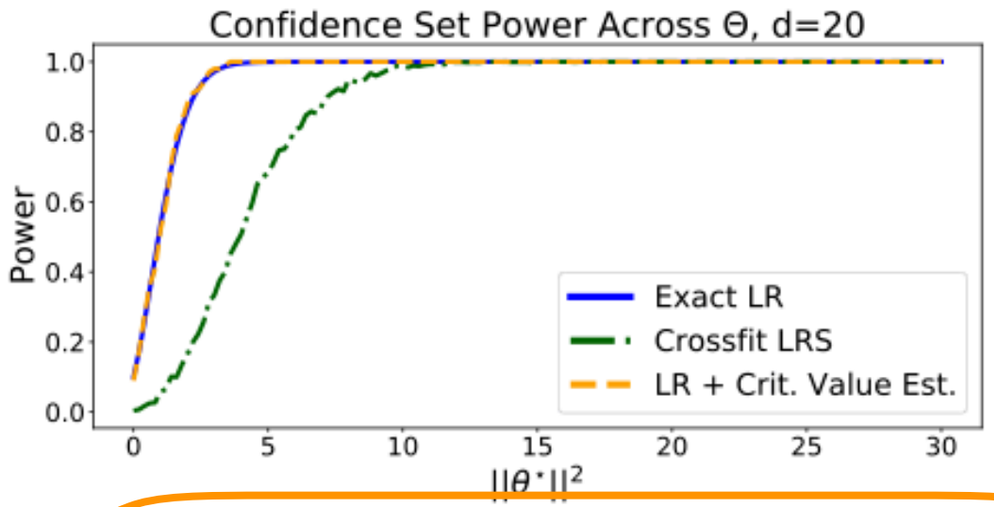
When d=2, our method "known LR" (GREEN) returns confidence sets that are similar to "exact LR" (BLUE), but smaller than the more conservative universal inference approach with "crossfit LR"

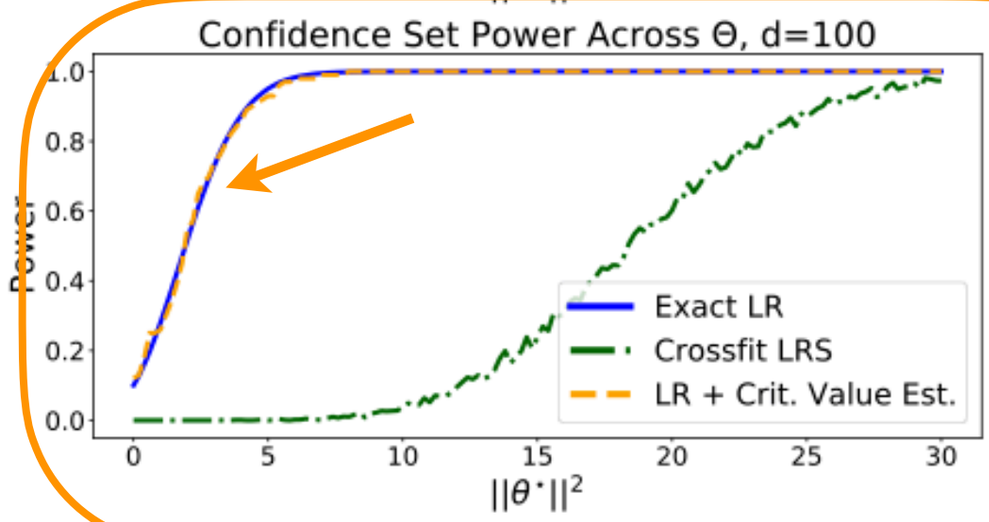
Coverage and Power for Known Likelihood

Gaussian Model, Known Likelihood Setting

	d=10	d=20	d=50	d=100
Coverage of LR + Crit. Value Est.	0.91 ± 0.03	0.91 ± 0.03	0.88 ± 0.03	0.88 ± 0.03
Coverage of Crossfit LRS	0.993 ± 0.008	0.997 ± 0.005	1.000 ± 0.000	1.000 ± 0.000

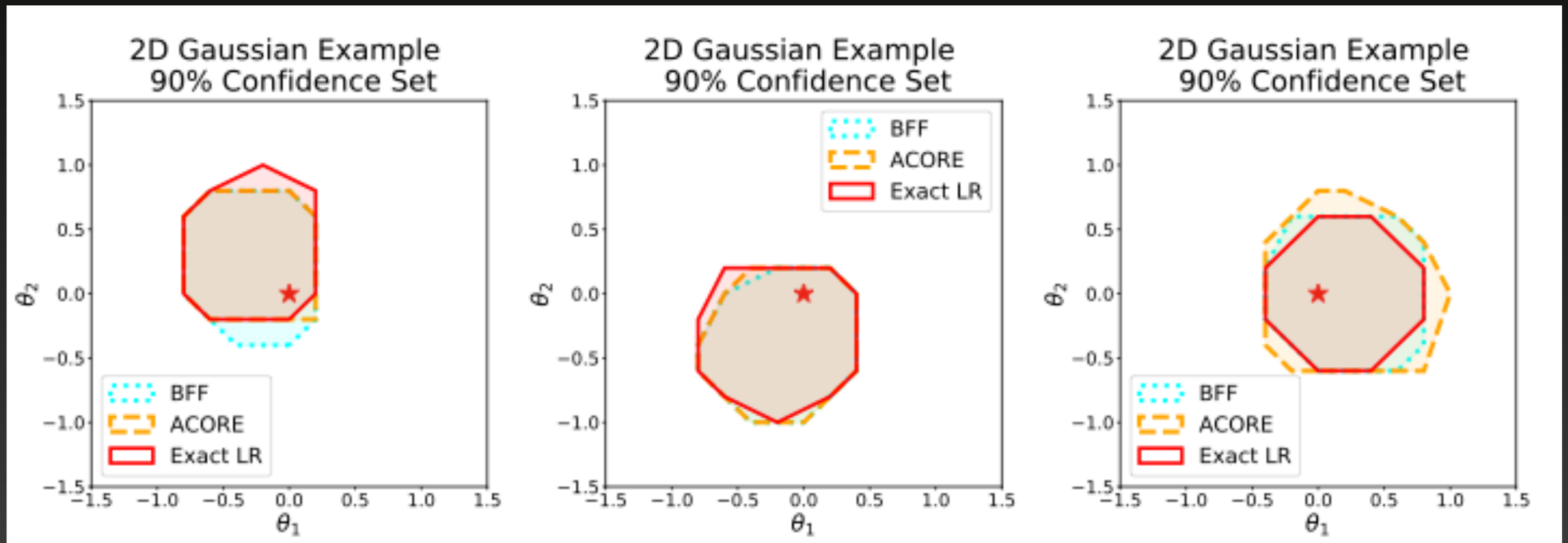






Our approach for estimating critical values yields the same power as the exact tests even in high dimensions, with a modest sample size of B'=5000.

LFI Setting in 2D: Examples of Constructed Confidence Sets (All Valid)

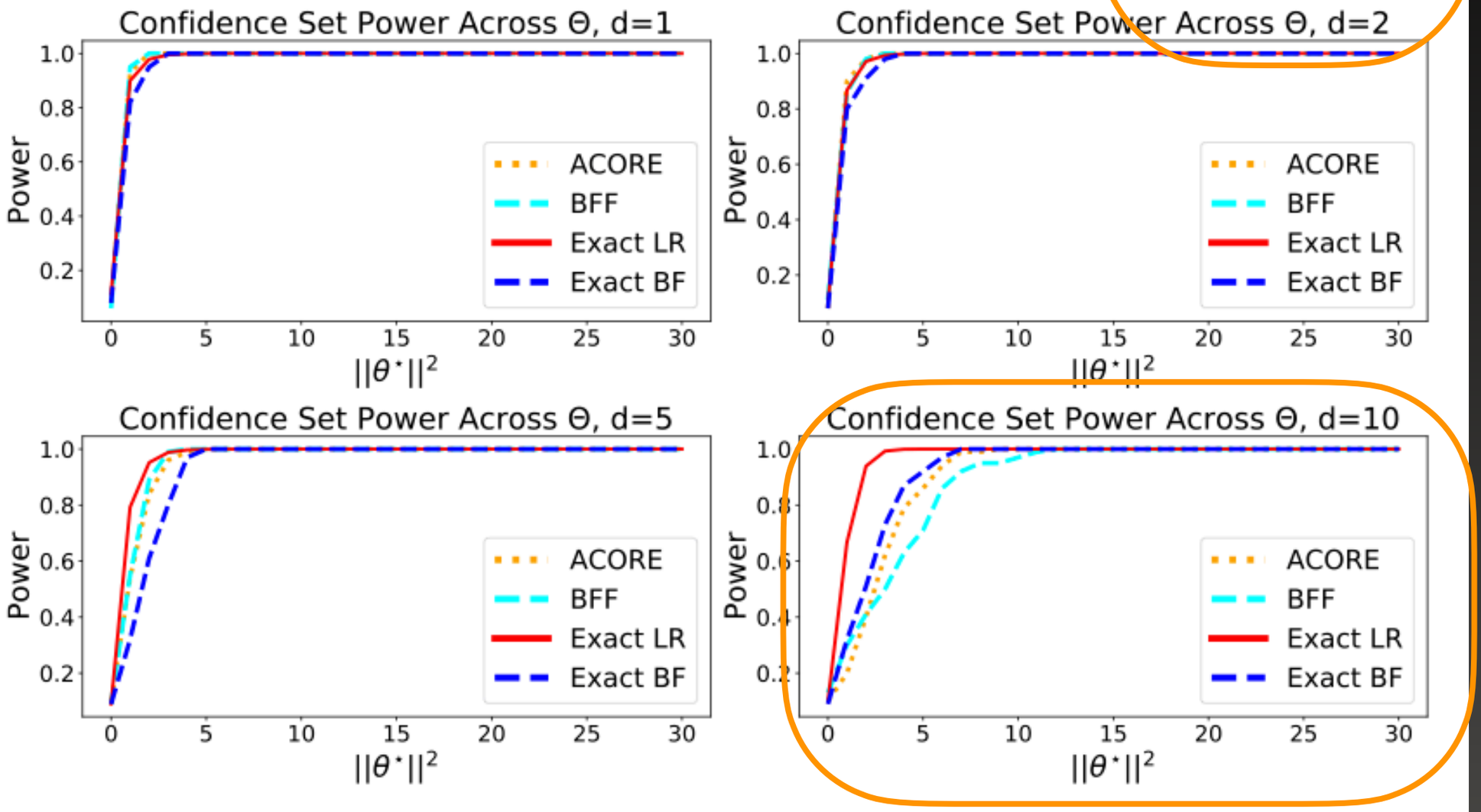


When d=2, ACORE and BFF confidence sets (for B=B′=5000) are comparable with the exact LR confidence sets.

Coverage and Power in an LFI Setting

Gaussian Model, Likelihood-Free Inference Setting

	d=1	d=2	d=5	d=10
Coverage of ACORE	0.92 ± 0.07	0.92 ± 0.07	0.90 ± 0.03	0.90 ± 0.03
Coverag e of BFF	0.94 ± 0.06	0.89 ± 0.10	0.91 ± 0.03	0.91 ± 0.03



In higher dimensions, ACORE and BFF confidence sets are still valid but lose some power with respect to their exact counterparts.