New Nonparametric Tools for Complex Data and Simulations in the Era of LSST

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

Joint work with Rafael Izbicki (UCSCar) and Taylor Pospisil (CMU)

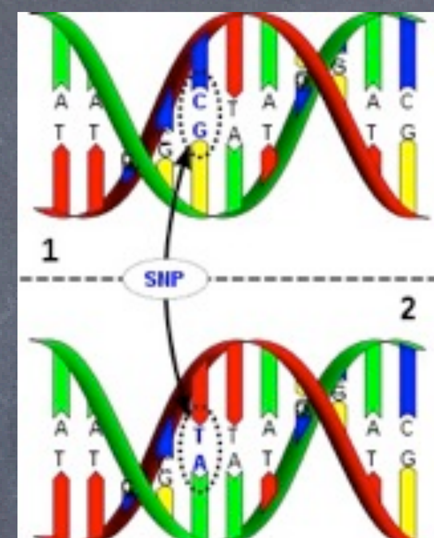
What Do Current Stats/ML Methods Do Well and Where Do They Fail?

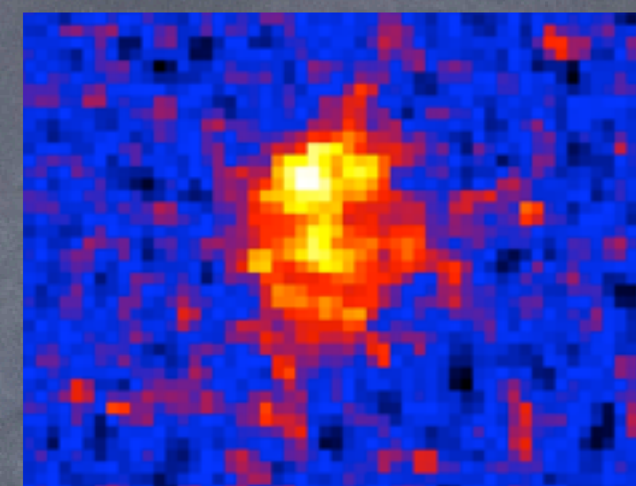
- LSST and future surveys will provide data that are wider and deeper.
- Simulation and analytical models are becoming ever sharper, reflecting more detailed understanding of physical processes.
- No doubt, statistical methods will play a key role in enabling scientific discoveries. But the question is:
  - What do current statistical learning methods do well and where do they fail?

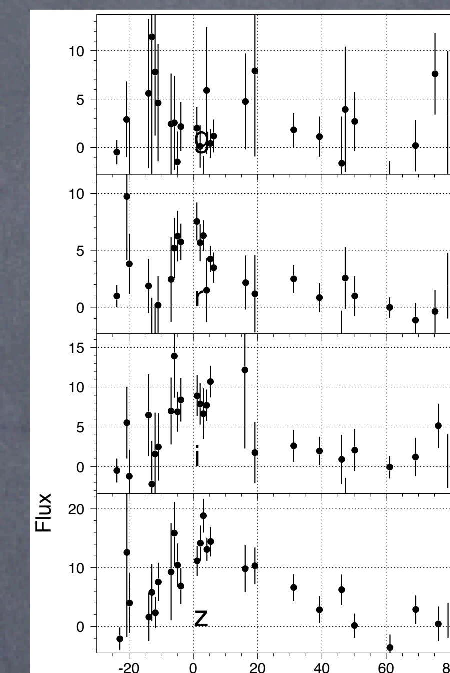
What Current Statistics and Machine Learning Methods Do well...

- Prediction (classification and regression)

x=







Given iid data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \sim f_{\mathbf{X},Y}$, there are two goals:

learning: Find an estimate $\widehat{r}(\mathbf{x})$ of the relationship between $\mathbf{X}$ and Y.

prediction: Given a new $\mathbf{X}$, predict Y; use $\widehat{Y} = \widehat{r}(\mathbf{X})$ as the prediction.

- Many ML algorithms scale well to massive data sets and can handle different types of (high-dimensional) data x.

What Current Statistics and Machine Learning Methods Don't Do Very Well…

- Modeling uncertainty beyond prediction (point estimate +/- standard error). Assessing models beyond prediction performance.

- Our objective: To develop new statistical tools that are

1. fully nonparametric

2. can handle complex data objects **x** without resorting to a few summary statistics

3. estimate and assess the quality of entire probability distributions

Next: Two Examples of Nonparametric Conditional Density Estimation ("CDE")

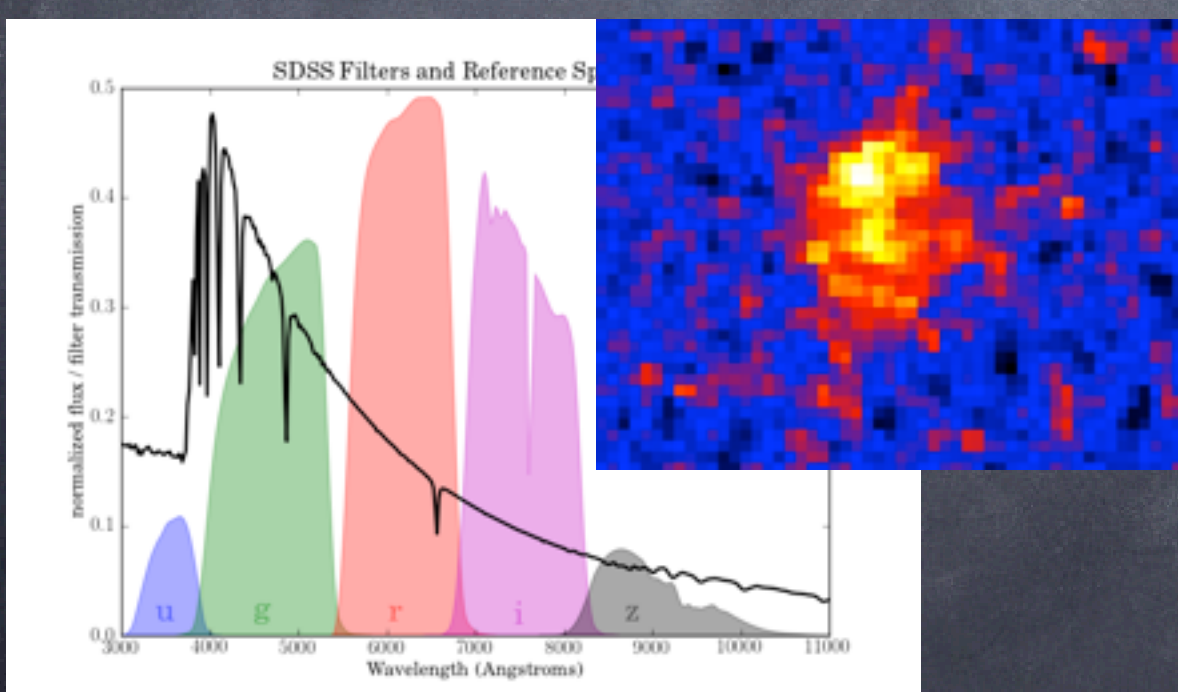
1. Photo-z estimation: Estimate $p(z|x)$ given photometric data x from individual galaxies

2. Nonparametric likelihood computation: Estimate posterior $f(\theta|x)$ using observed and simulated data, where
   θ=parameters of interest
   x=high-dim data (entire image, correlation functions, etc.)

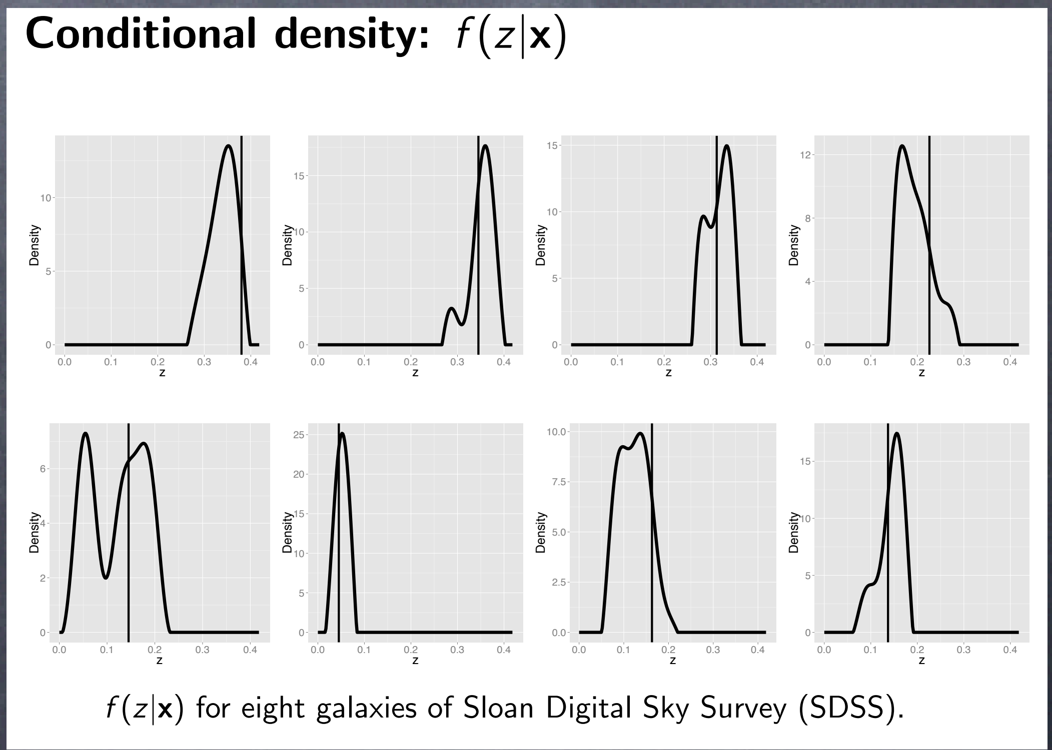
I: Photo-z Density Estimation

$$\mathcal{D} = \{(X_1, Z_1), \ldots, (X_n, Z_n), X_{n+1}, \ldots, X_{n+m}\},$$

- z = "true" redshift (spectroscopically confirmed)
- **x** = photometric colors and magnitudes of individual galaxy
- Because of degeneracies, need to estimate the full conditional density p(z|**x**) instead of just the conditional mean r(**x**)=E[Z|**x**].



Photometry



Estimates of p(z|x) from photometry

Can We Leverage the Advantages of Training-Based Regression Methods for Nonparametric CDE?

- Basic idea of "FlexCode" [Izbicki & Lee, 2017]: Expand the unknown p(z|x) in a suitable orthonormal basis $\{\varphi_i(z)\}_i$

$$p(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x})\phi_i(z)$$

- By the orthogonality property, the expansion coefficients are just conditional means (which can be estimated by regression)

$$\beta_i(\mathbf{x}) = \mathbb{E}\left[\phi_i(z)|\mathbf{x}\right] \equiv \int p(z|\mathbf{x})\phi_i(z)dz$$

1. FlexCode converts a difficult non-parametric CDE problem into a better understood regression problem.
2. We choose tuning parameters in a principled way by minimizing a "CDE loss" on a validation set.

Use Cross-Validation with a CDE Loss for Model Selection and Method Comparison

- For model selection and comparison of p(z|x) estimates, we define a conditional density estimation (CDE) loss:

$$
\begin{aligned}
L(p,\widehat{p}) &= \int\int (p(z \mid \mathbf{x}) - \widehat{p}(z \mid \mathbf{x}))^2 dz dP(\mathbf{x}) \\
&= \mathbb{E}_{\mathbf{X}}\left[\int \widehat{p}(z \mid \mathbf{X})^2 dz\right] - 2\mathbb{E}_{\mathbf{X},Z}\left[\widehat{p}(Z \mid \mathbf{X})\right] + K_f
\end{aligned}
$$

- This loss is the CDE equivalent of the MSE in regression
- Note: We can estimate the CDE loss (up to a constant) on test data without knowledge of the true densities.

An assessment of photometric redshift PDF performance in the context of LSST

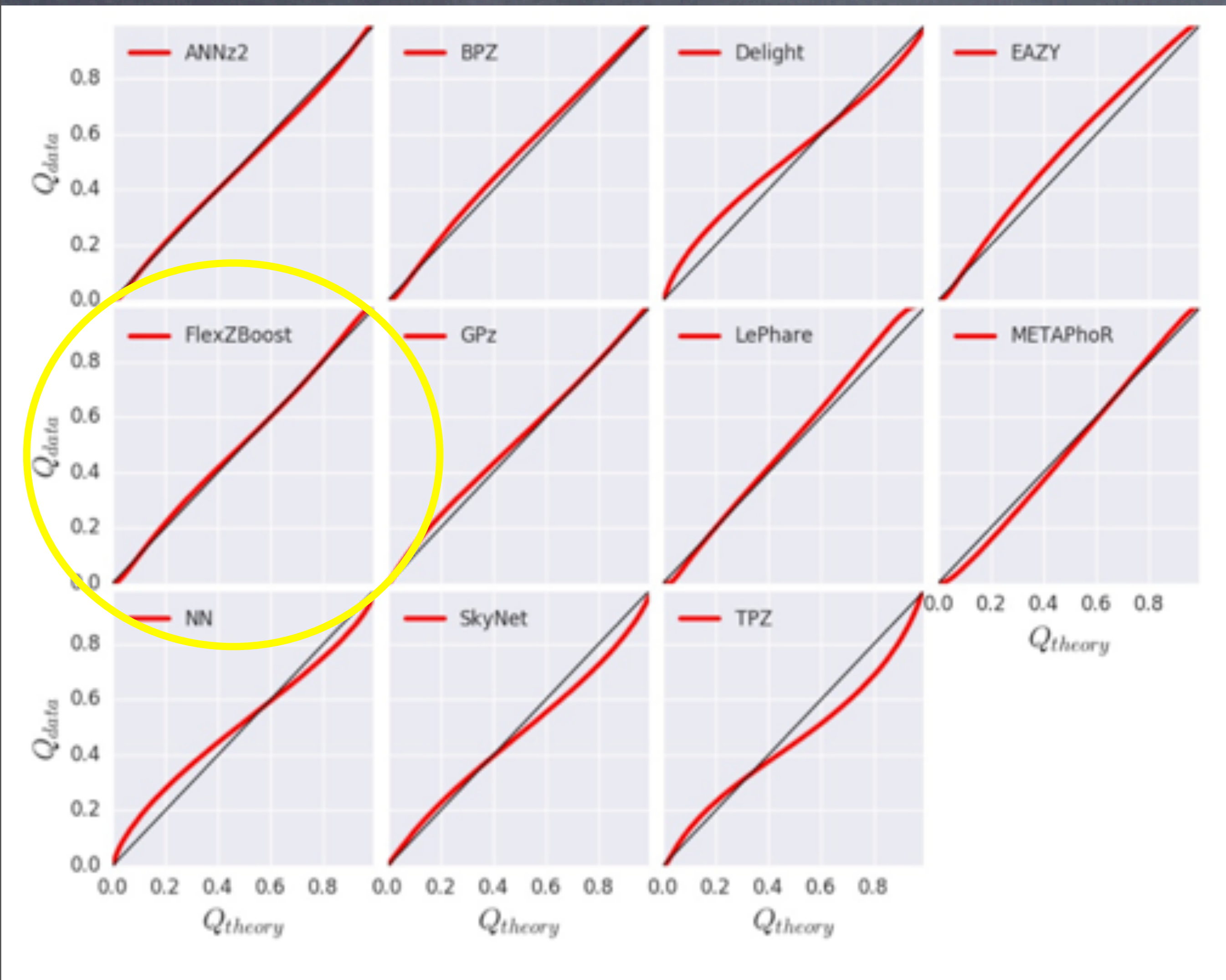
LSST-DESC Photometric Redshift Working Group

April 3, 2018

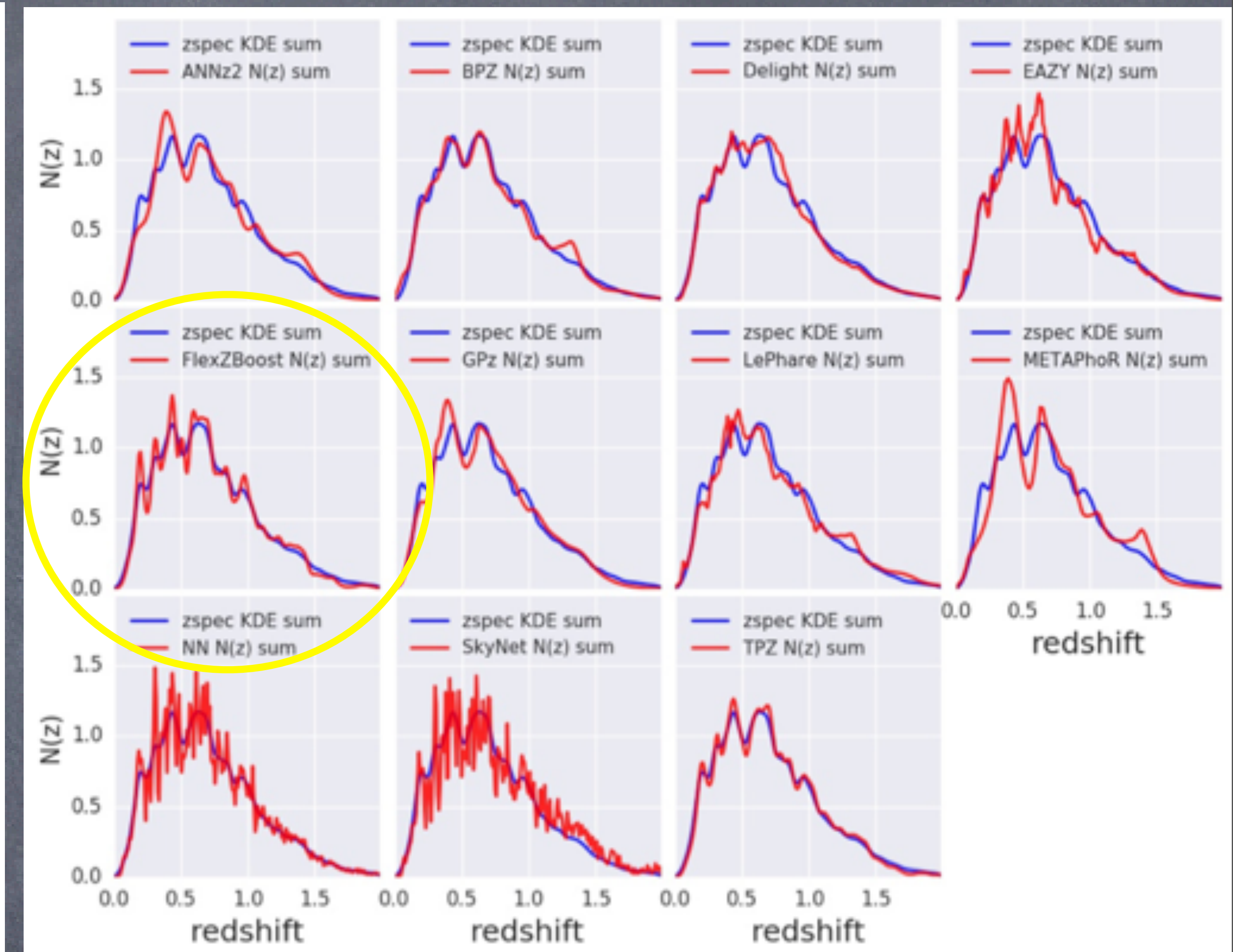
ABSTRACT
Photometric redshift (photo-z) probability distribution functions (PDFs) are a planned data product of most upcoming galaxy imaging surveys. However, the photo-z PDFs resulting from different techniques are not in general consistent with one another. We present the results of the the Large Synoptic Survey Telescope Dark Energy Science Collaboration (LSST-DESC) Data Challenge 1 (DC1), a series of tests of different photo-z PDF codes on a realistic simulation of upcoming LSST galaxy photometry catalogues. This is the first side-by-side test of photo-z PDFs produced by several popular methods in the literature, evaluated on the basis of metrics like the Kolmogorov-Smirnoff statistic, Cramer-von Mises statistic, Anderson-Darling statistic, Kullback-Leibler divergence, moments,

- We entered "FlexZBoost" into the LSST-DESC Data Challenge 1 (Buzzard v1.0 simulations with 0<z<2 and i<25, complete and representative training data and templates)
- "FlexZBoost" is a version of FlexCode that uses a Fourier basis for the basis expansion, and xgboost for regression (which scales to billions of examples)

DC 1: Side-by-Side Tests of 11 Photo-z Codes (3 Template-Based, 8 Training-Based)



QQ Plots

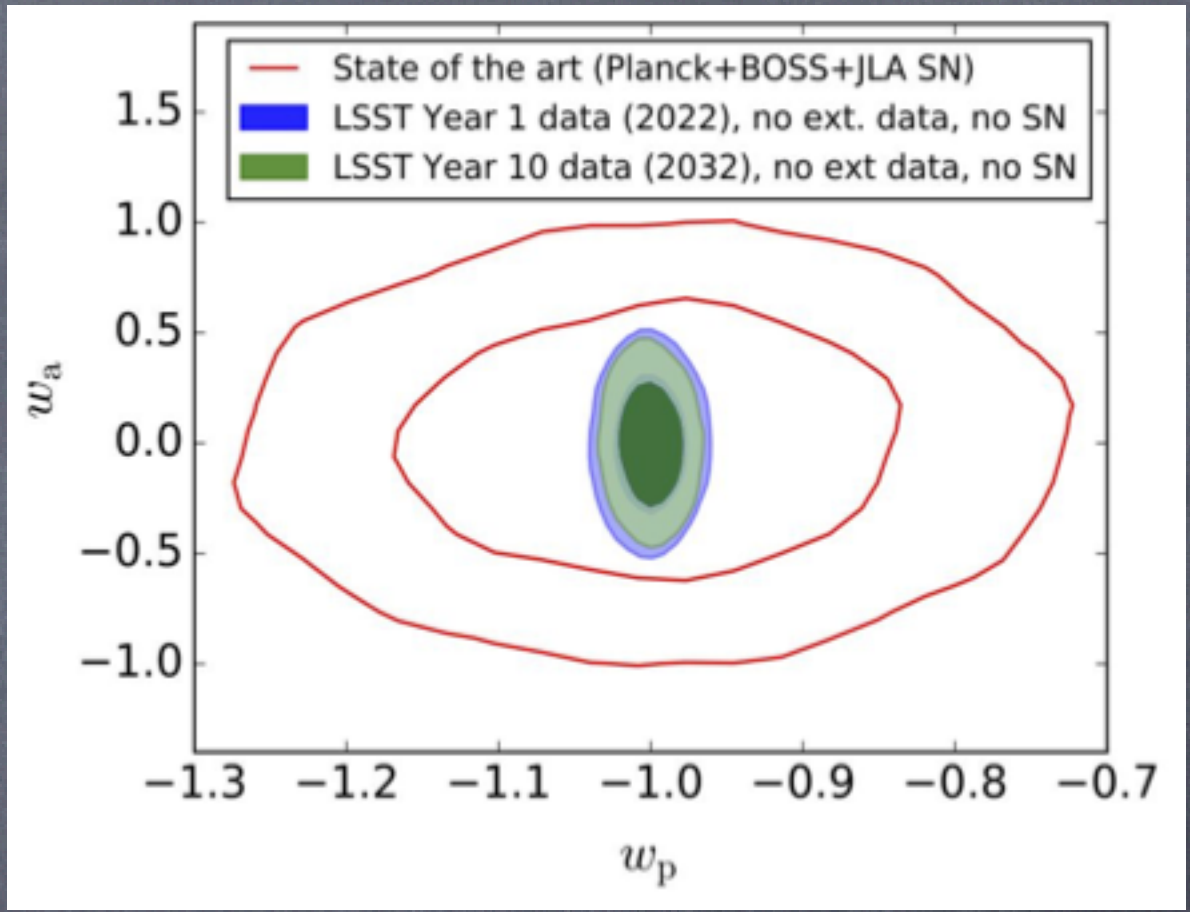


Stacked p(z) compared to true n(z)

- "FlexZBoost" shows one of the best performances in estimating both p(z) and n(z) for DC1 data with no tuning other than CV. In addition: Scales to massive data (billions of galaxies); can store p(z) estimates at any resolution losslessly with 35 Fourier coeffs/galaxy.

II. A New CDE Approach to Fast Nonparametric Likelihood Computation

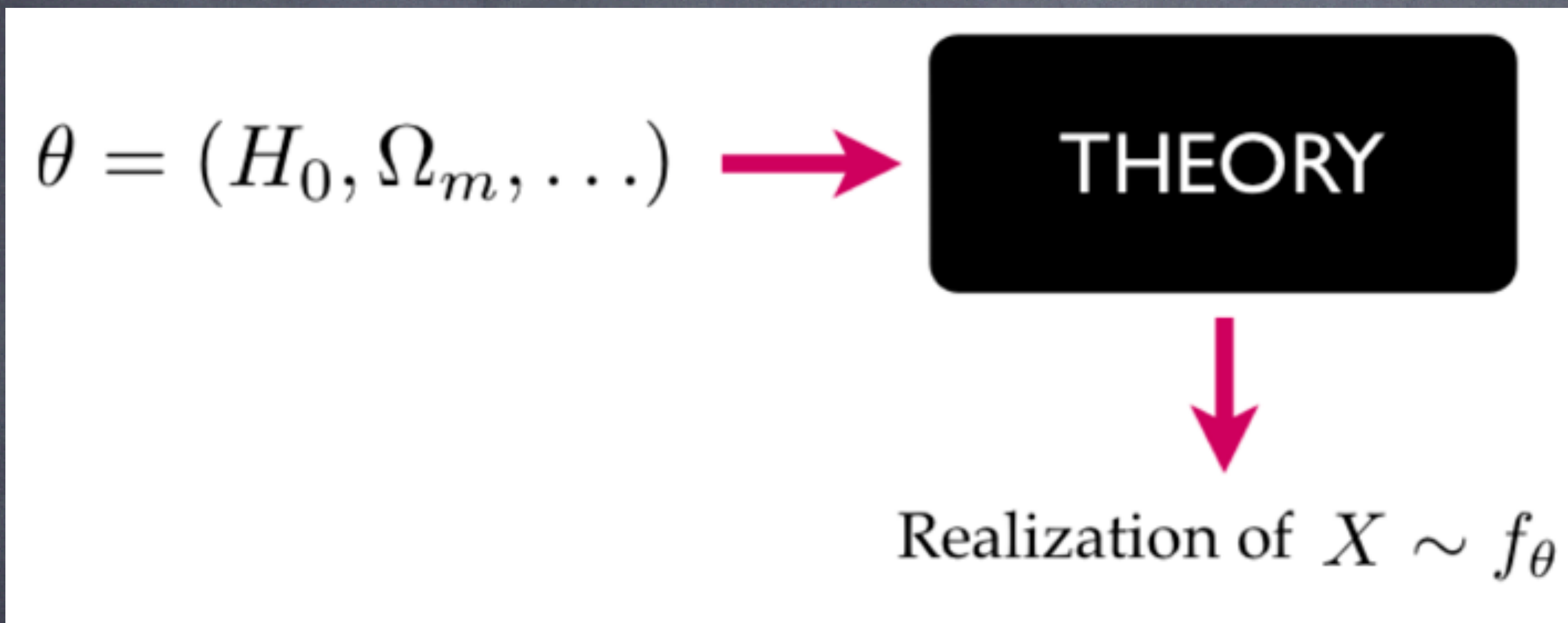
- Fig: LSST will greatly increase the cosmological constraining power compared to current state of the art



- Standard Gaussian likelihood models may become questionable at LSST precision. (Several works explore non-Gaussian alternatives and “varying covariance” models, e.g. Eifler et al)

- How about fully nonparametric methods? Could e.g ABC and likelihood-free methods be made practical for LSST science?

Approximate Bayesian Computation (ABC) Driven By Repeated Simulations From a Forward Model



1. Draw θ from prior $\pi(\theta)$
2. Simulate x from forward model f_θ
3. Accept θ if $\mathrm{dist}(S(\mathbf{x}), S(\mathbf{x}_{\mathrm{obs}})) < \epsilon$
4. Return to Step 1

Creates a sample from approximation of posterior:

$$\pi_\epsilon(\theta \mid \mathrm{dist}(S(\mathbf{x}), S(\mathbf{x}_{\mathrm{obs}})) < \epsilon) \;\approx\; \pi(\theta \mid \mathbf{x}_{\mathrm{obs}})$$

Several Outstanding Issues with ABC

1. ABC requires repeated forward simulations (which may not be computationally feasible)

2. need to choose approximately sufficient summary statistics of the data

ABC creates a sample from approximation of posterior:

$$\pi_\epsilon(\theta \mid \mathrm{dist}(S(\mathbf{x}), S(\mathbf{x}_{\mathrm{obs}})) < \epsilon) \;\approx\; \pi(\theta \mid \mathbf{x}_{\mathrm{obs}})$$

Equality if and only if $\epsilon = 0$ and $S(\cdot)$ is sufficient for θ.

3. not clear how to assess the performance of ABC methods without knowing the true posterior

We propose ABC-CDE [Izbicki, Lee and Taylor 2018]: Combines ABC with CDE Training-Based Method

- Idea: Take the output from ABC (at a high acceptance rate)

$$(\theta_1, \mathbf{x}_1), (\theta_2, \mathbf{x}_2), \ldots, (\theta_B, \mathbf{x}_B) \sim f_\epsilon(\theta, \mathbf{x}),$$

where no ABC corresponds to $\epsilon \to \infty$; that is, an "acceptance rate" of 1.

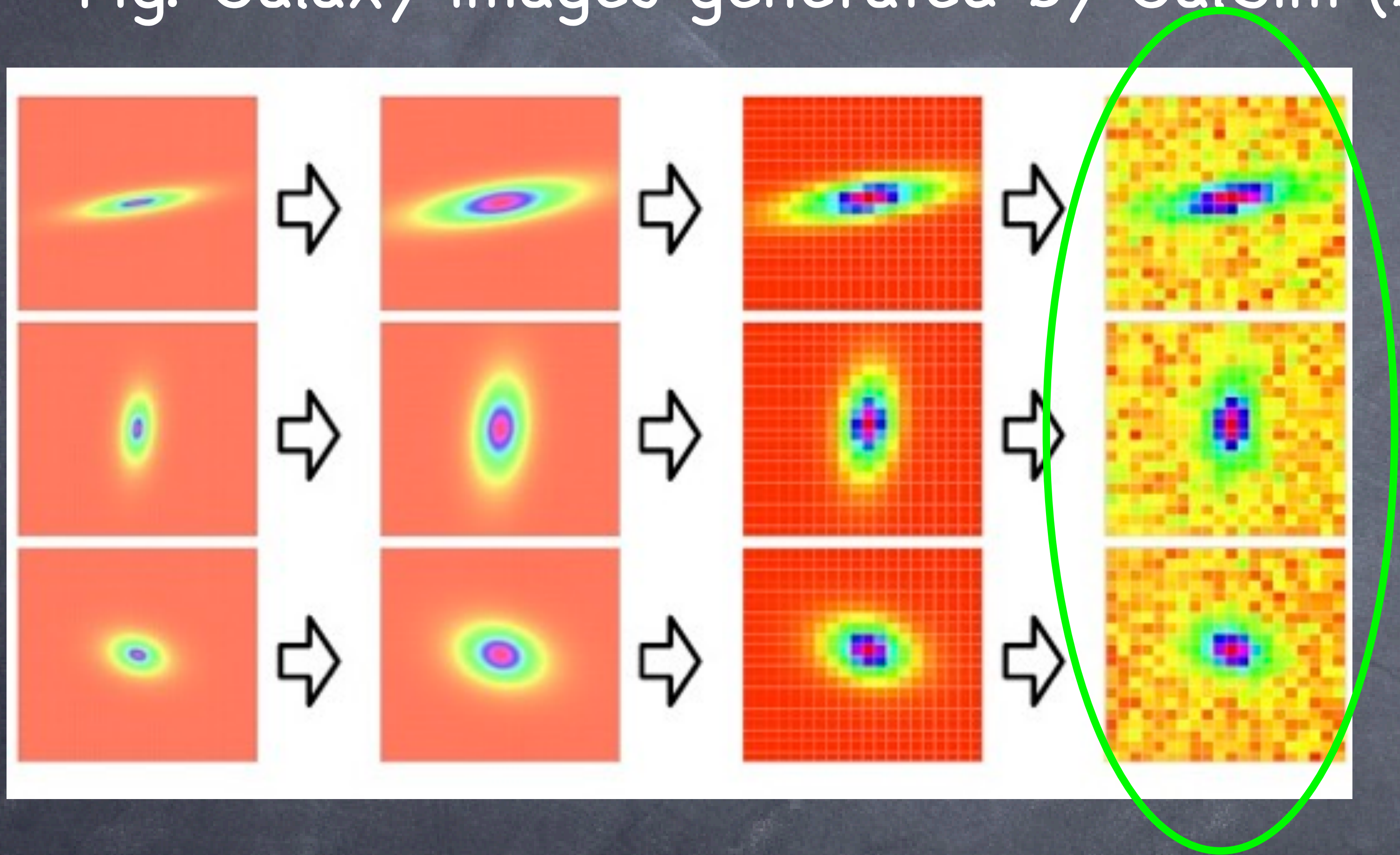
and then directly estimate the posterior $\pi(\theta|x_0)$ at observed data x_0 using a CDE training-based method

1. Can adapt CDE method to different types of high-dimensional data (entire images, correlation functions, etc.). Dimension reduction is implicit in the choice of CDE method.

2. Can use our "CDE loss" to choose which model is closest to the truth --- even without knowing the true posterior.

Example: Nonparametric Likelihood Computation with Entire Images (No Summary Statistics; No ABC)

Fig: Galaxy images generated by GalSim (blurring, pixelation, noise)



θ=(rotation angle, axis ratio)

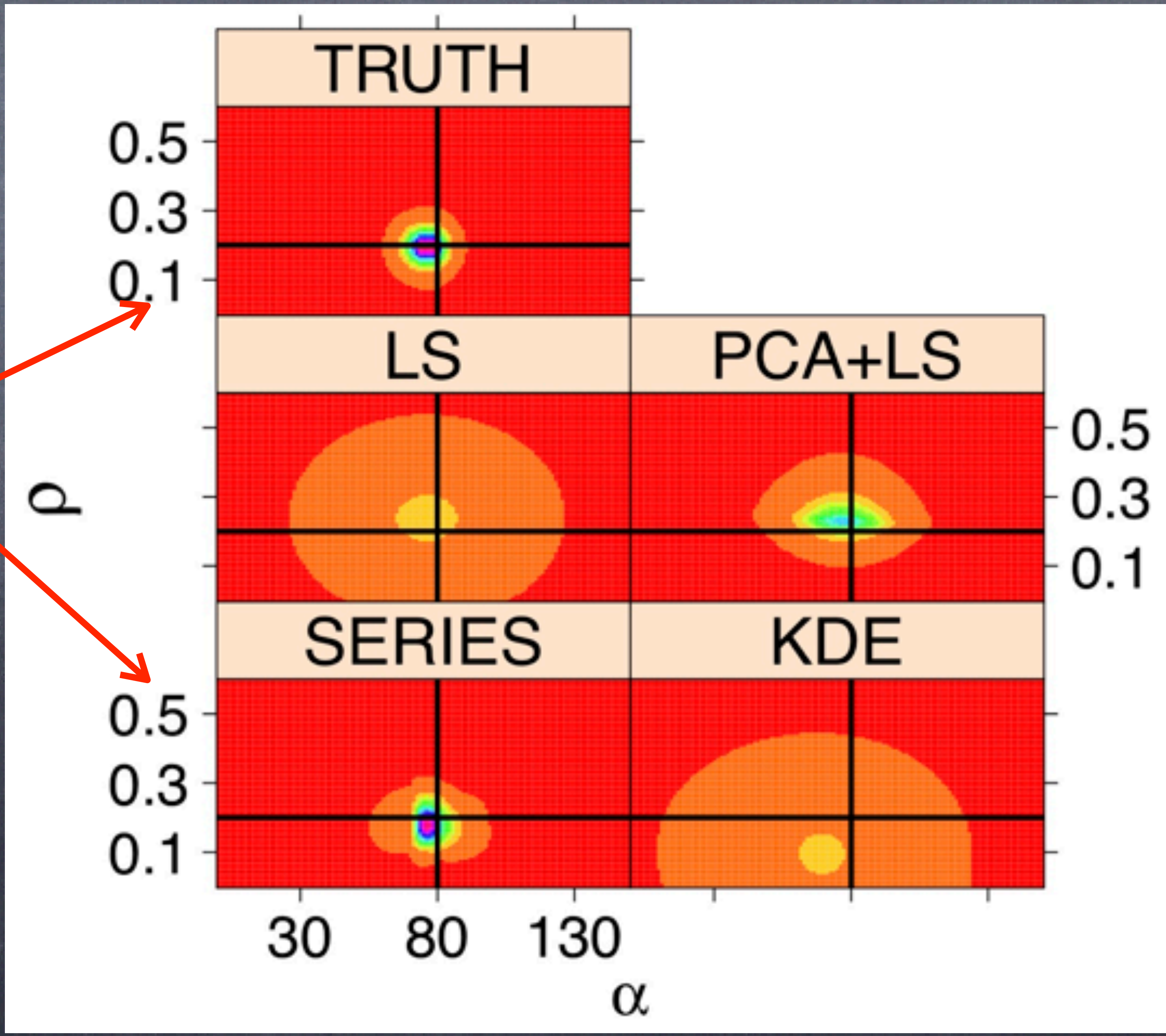
x: entire image

- Use a uniform prior and forward model, to simulate a sample $(\theta_1, x_1), \ldots, (\theta_B, x_B)$
- Estimate the likelihood $L(\theta) \propto f(\mathbf{x}|\theta)$ directly via CDE. No summary statistics (entire images); no MCMC or ABC iterations

Even Decent Performance With Uniform Prior and Without ABC Iterations and Summary Statistics

- Unknown parameters: rotation angle α, axis ratio ρ
- Contours of the estimated likelihood for different CDE methods

The spectral series estimator (bottom left) comes close to the true distribution (top)

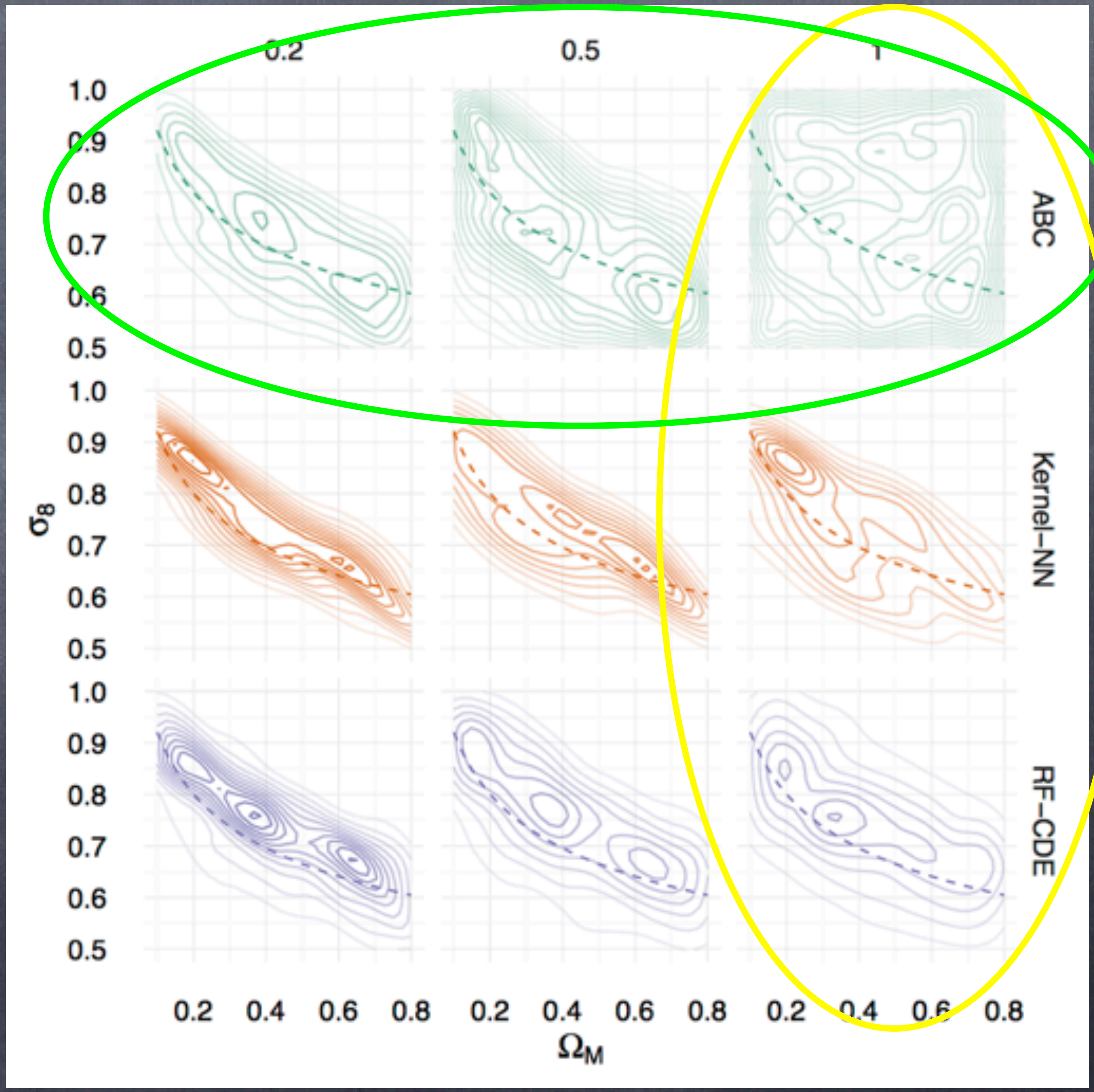


Toy Example of Cosmological Parameter Inference for Weak Lensing Mock Data via ABC-CDE.

- Use GalSim to generate a cosmic shear grid realization with shape noise. Input two-point correlation functions to ABC.

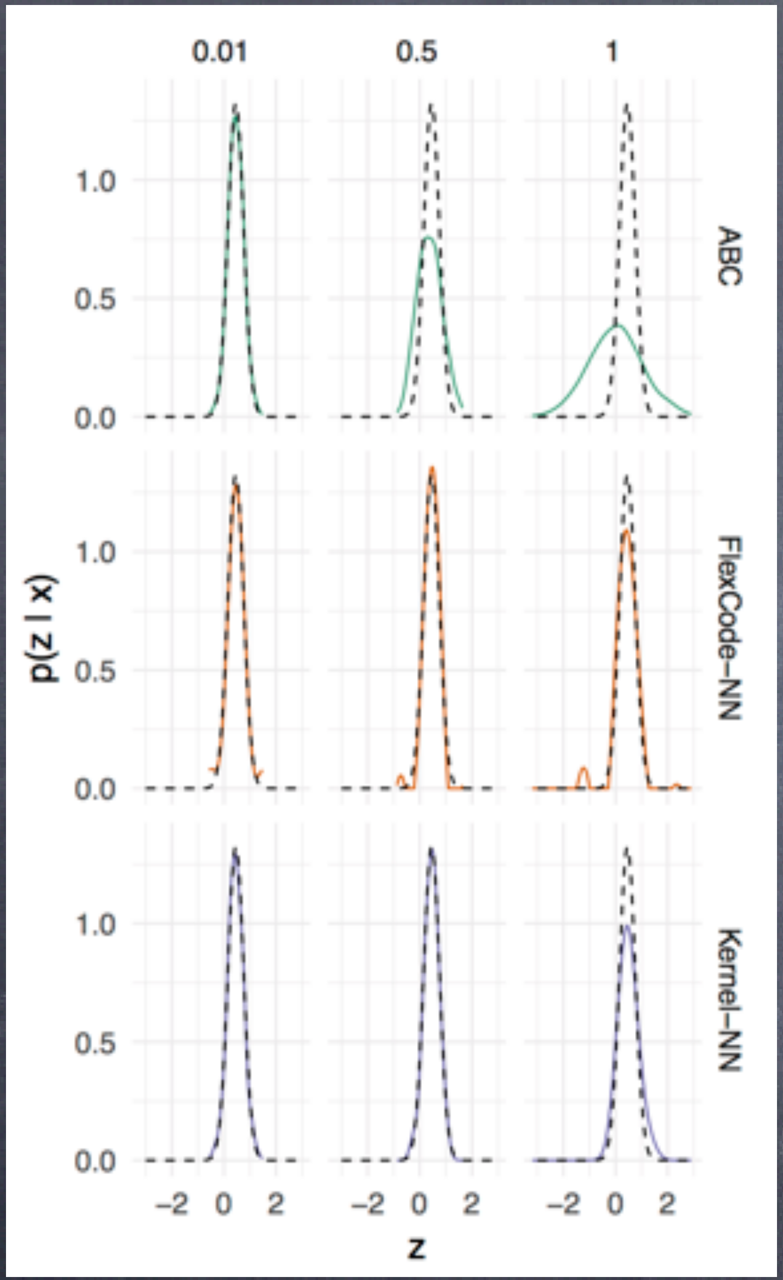
Fig: Estimated posteriors of Ω_M and σ_8 for ABC (top row) and two ABC-CDE methods (middle and bottom rows).

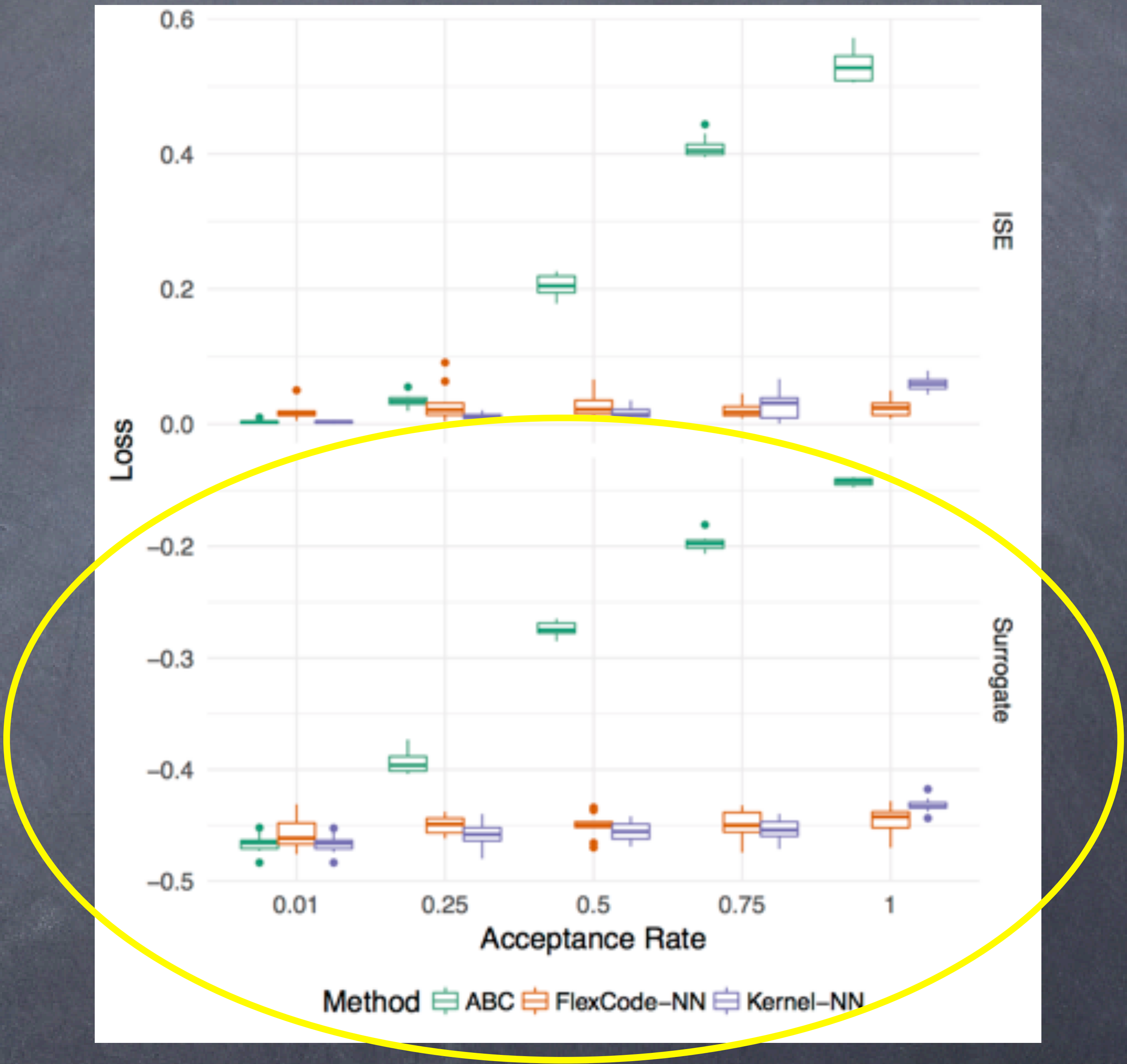
ABC-CDE posteriors concentrate around the degeneracy line at higher acceptance rates; that is, with fewer simulations.



Toy Example with 1D Normal Posterior:

Estimated CDE Loss Tells Us Which Method is Best.





- Bottom right: CDE loss estimated from data for three different methods (at varying acceptance rates). By comparing these values we can tell which estimate is closest to the true posterior.

Summary: Nonparametric CDE Approach to Inference

- We are developing fast nonparametric CDE tools that go beyond prediction and estimate entire posteriors and likelihoods from observed and simulated data

  1. potentially explore different types of high-dimensional data

  2. principled method of comparing estimates without knowing the true posterior

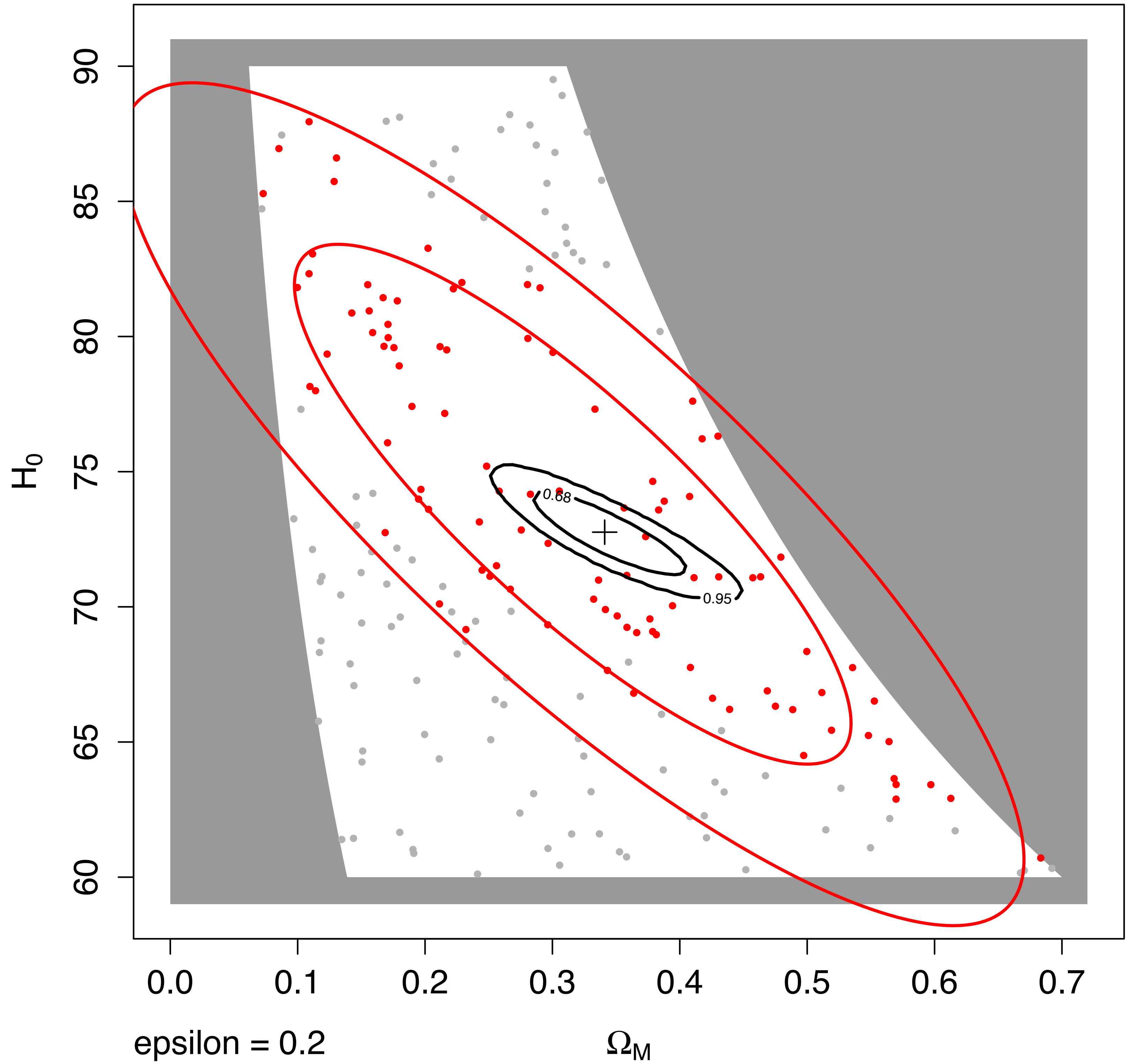
- Please contact me for questions: annlee@cmu.edu

Acknowledgements

- Rafael Izbicki (Stats at UFSCar, Brazil)
- Taylor Pospisil (Stats & Data Science at CMU)
- CMU AstroStats: Peter Freeman, Chad Schafer, Nic Dalmasso, Michael Vespe
- U. Pitt. Astro.: Jeff Newman, Rongpu Zhu
- LSST-DESC: Sam Schmidt, Alex Malz & pz wg, Tim Eifler, Rachel Mandelbaum, Chien-Hao Lin

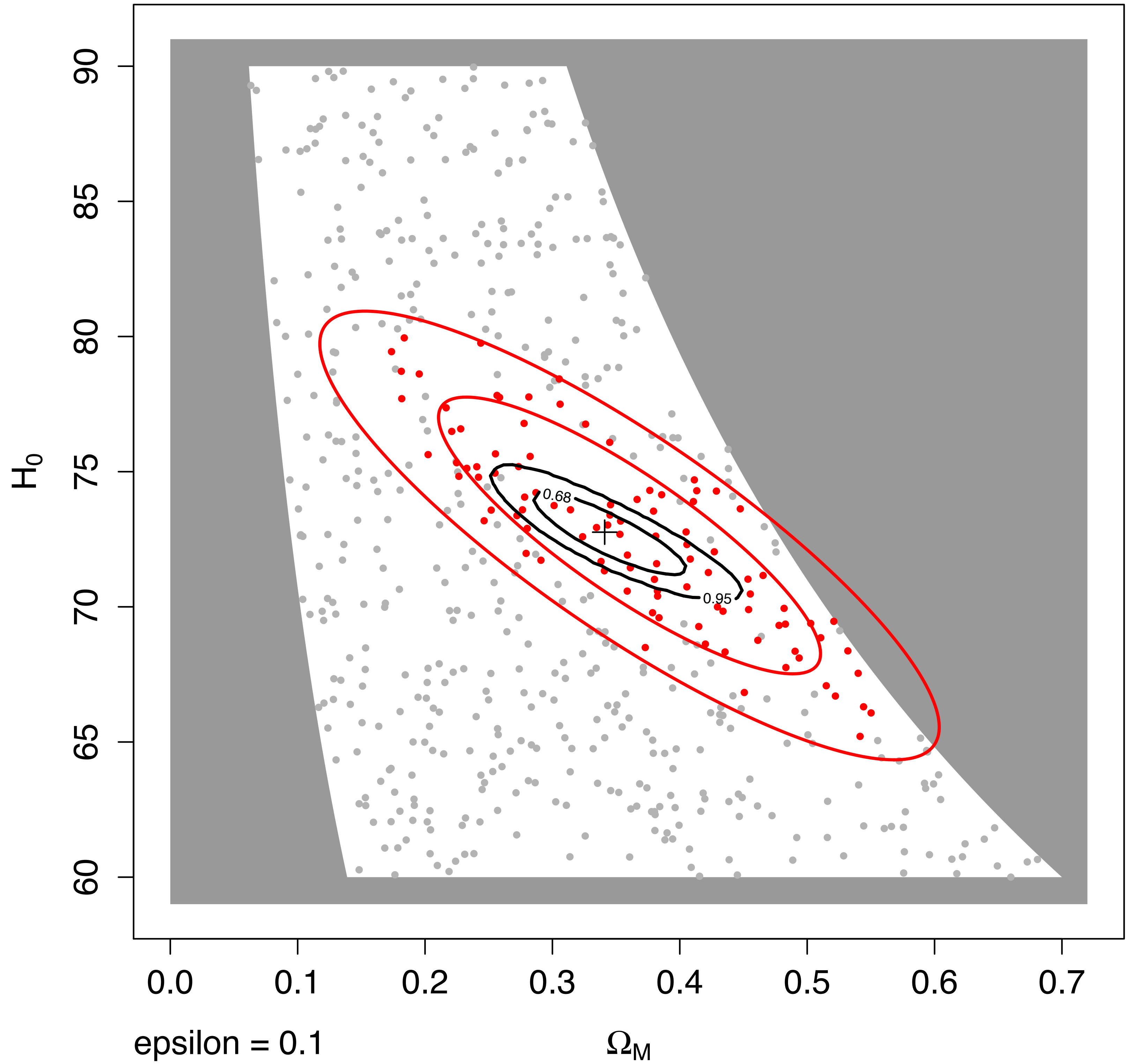
Contact: annlee@cmu.edu

EXTRA SLIDES START HERE



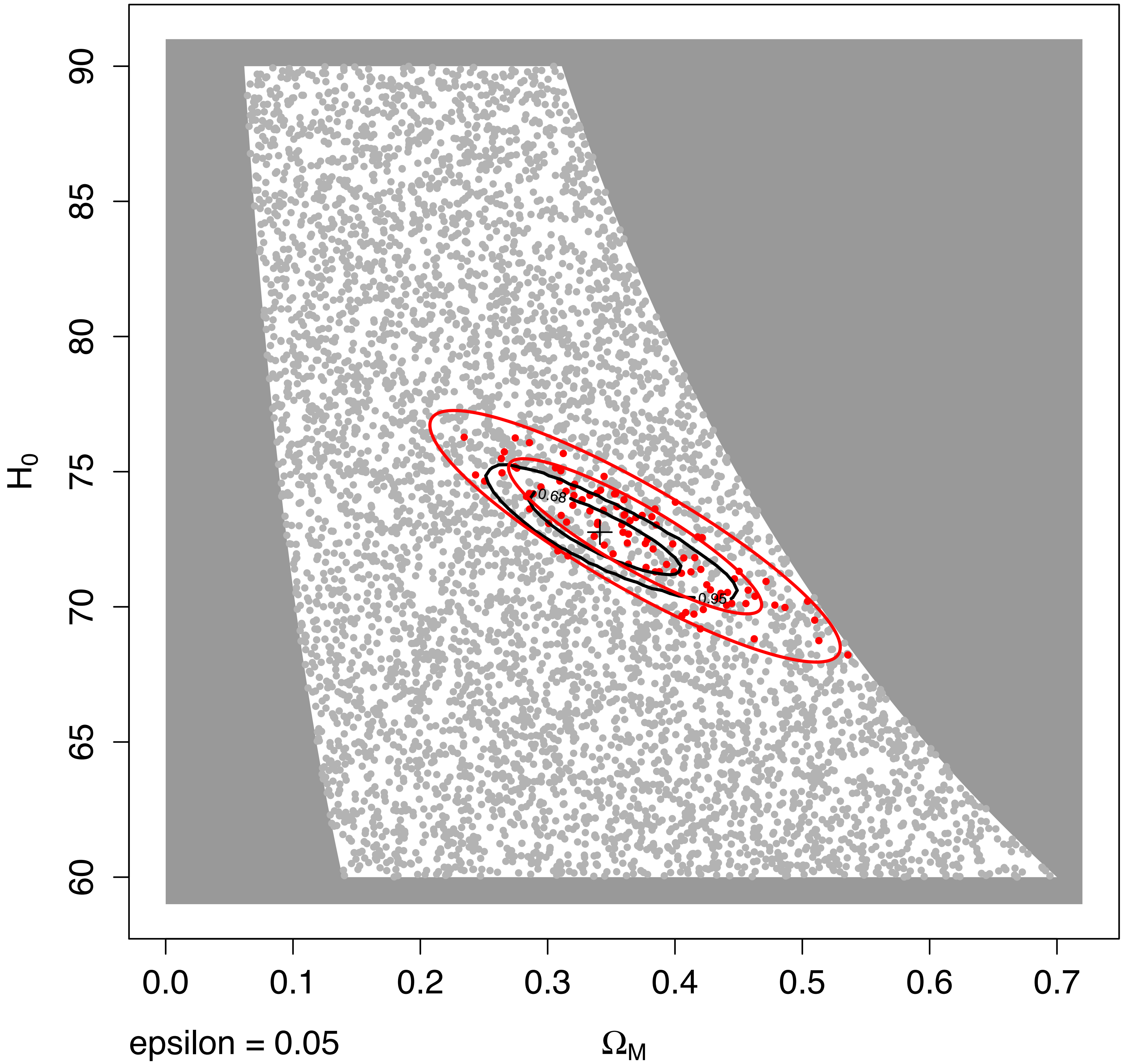
Basic rejection approach applied to SNe data

ABC applied to SNe data; see Weyant/Schafer/Wood-Vasey (ApJ 2013)



Basic rejection approach applied to SNe data

[Courtesy of Chad Schafer]



Basic rejection approach applied to SNe data

[Courtesy of Chad Schafer]