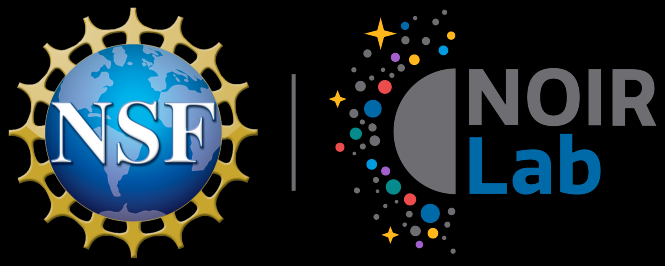
Using OSG to enable the development of a Simulation based inference Approach for Galaxy Cluster Cosmology

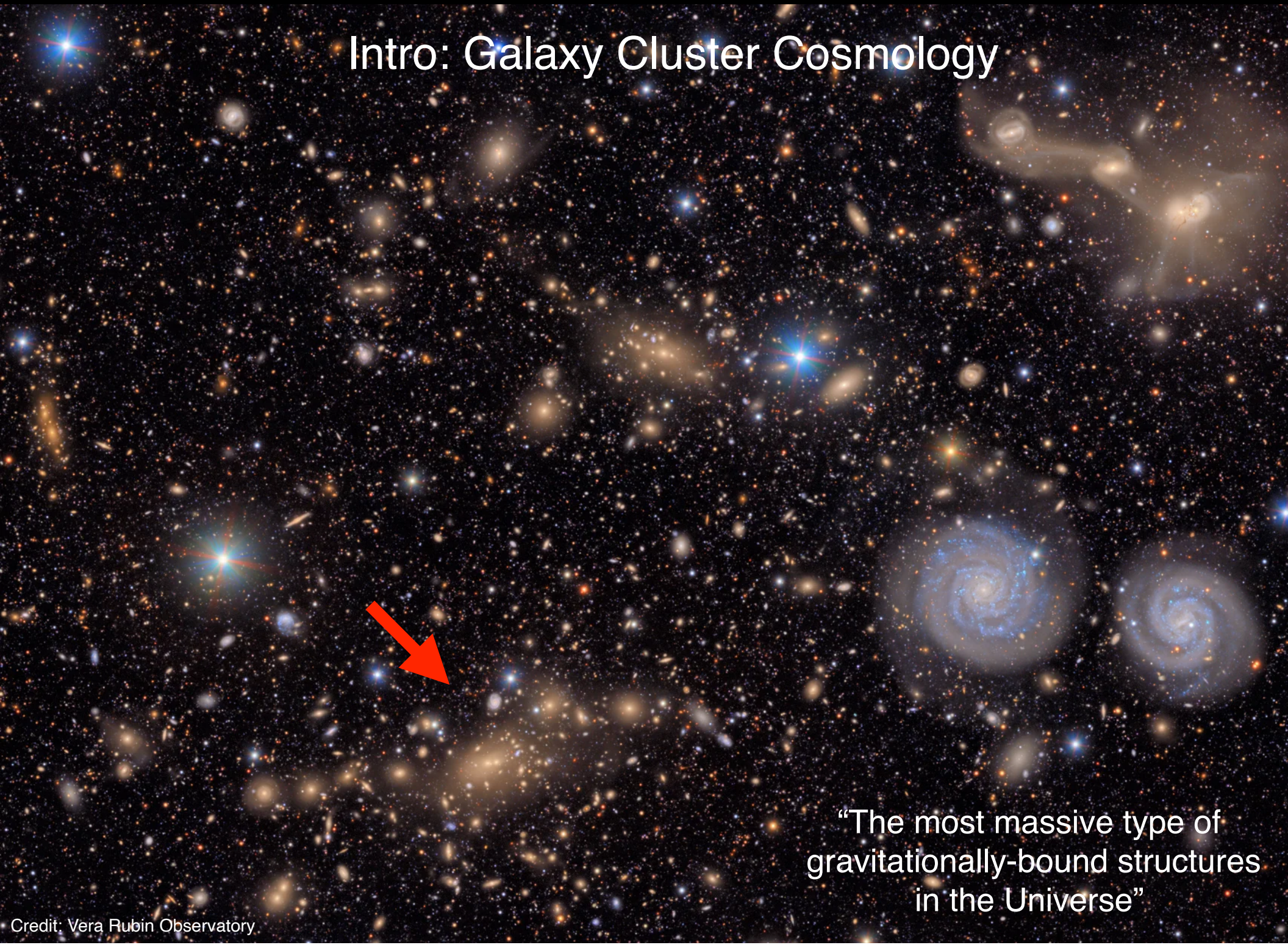


Yuanyuan Zhang
Associate Astronomer @ NSF NOIRLab

with
Moonzarin Reza, Camille Avestruz, Louis E. Strigari,
Simone Shevchuk, Francisco Villaescusa-Navarro

This talk is based on arXiv: 2409.20507 (Reza, Zhang et al.)

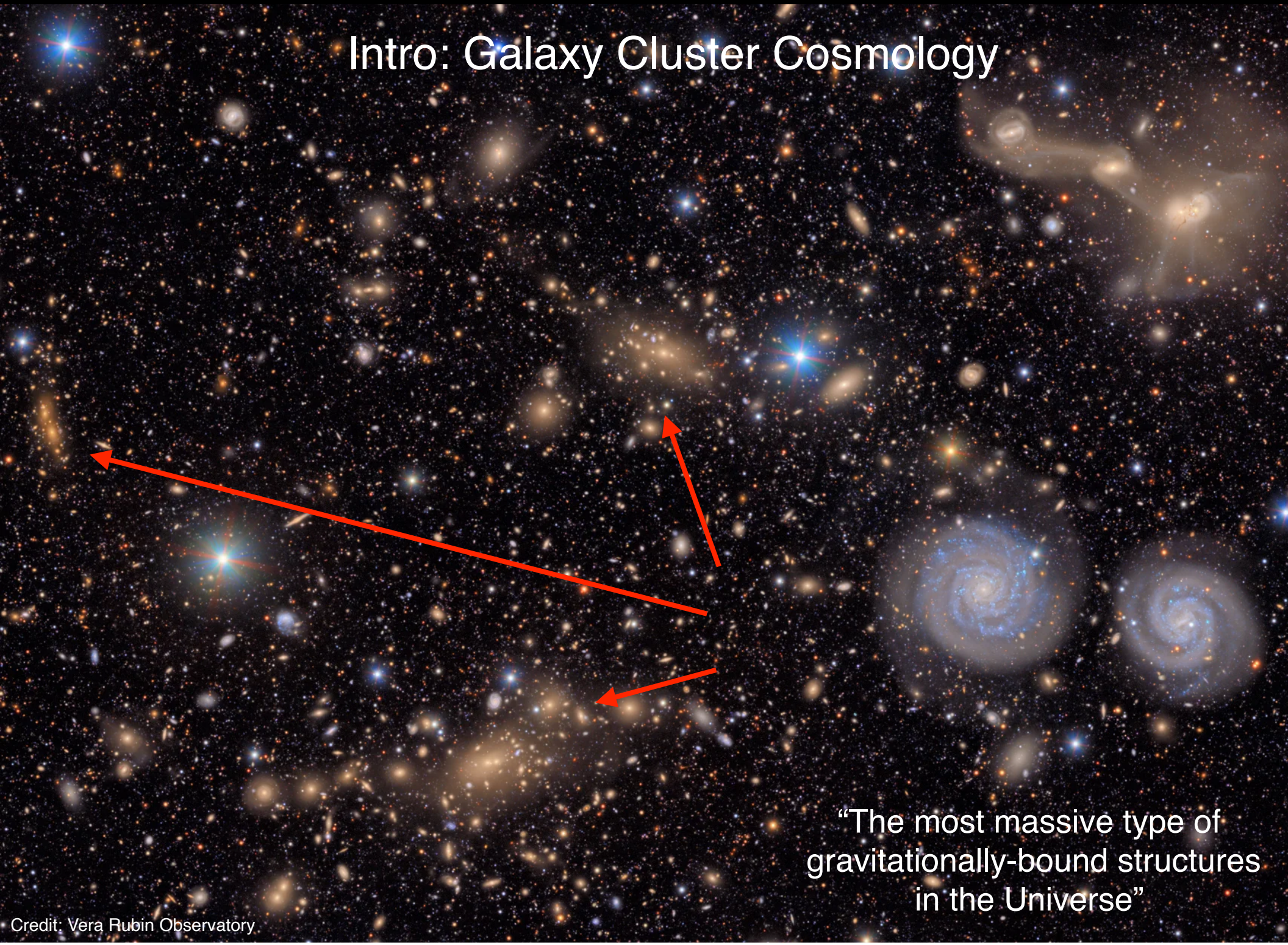
Intro: Galaxy Cluster Cosmology



"The most massive type of gravitationally-bound structures in the Universe"

Credit: Vera Rubin Observatory

Intro: Galaxy Cluster Cosmology

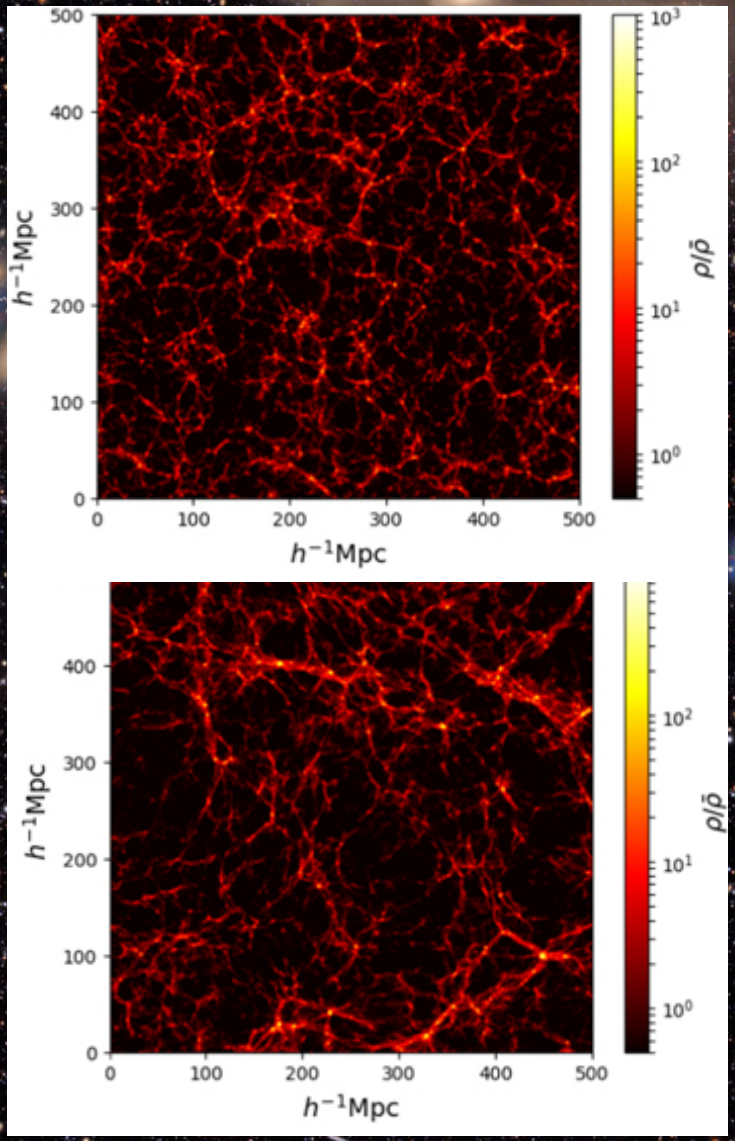


"The most massive type of gravitationally-bound structures in the Universe"

Credit: Vera Rubin Observatory

Intro: Galaxy Cluster Cosmology

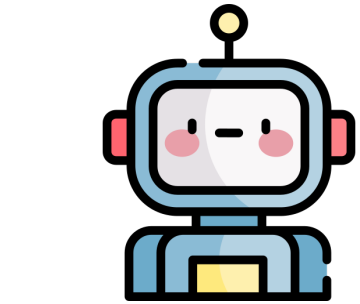
The abundance of galaxy clusters are sensitive to cosmological parameters like matter density Ω_m.

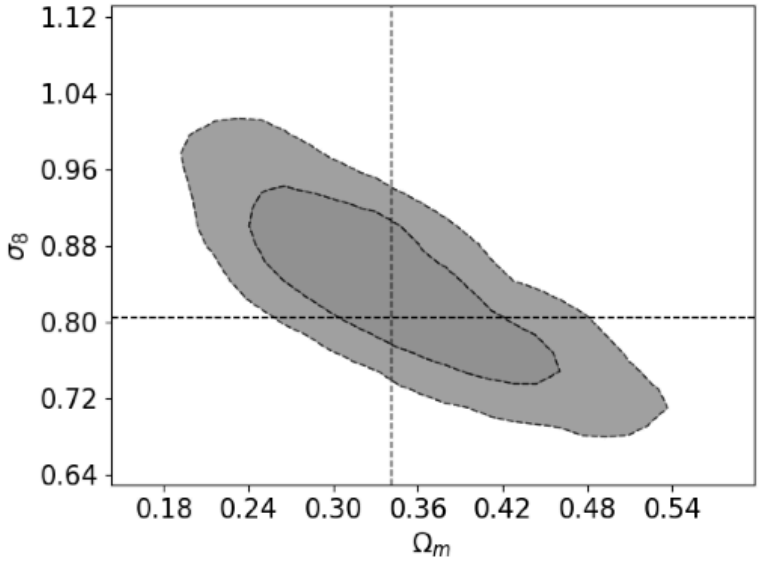


Villaescusa-Navarro et al. 2020

Credit: Vera Rubin Observatory

Galaxy Cluster Cosmology + ML





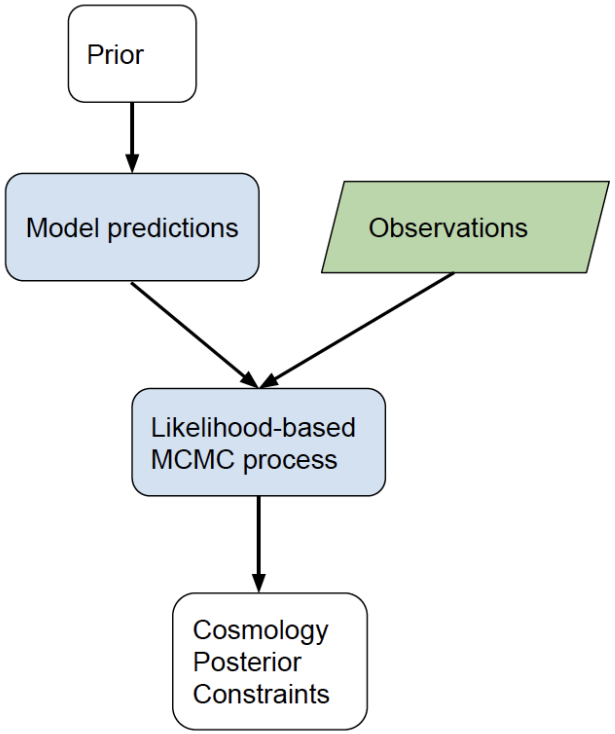
Credit: arXiv:2309.06593 Mock test

Simulation Based Inference (SBI) with Neural Posterior Estimation:

- Decide on the observable/data vectors and parameters.
- Generate simulations for training.
- Train a ML model (e.g., Mixture Density Networks).
- Derive posterior parameter distributions, for the real observables/data vectors.

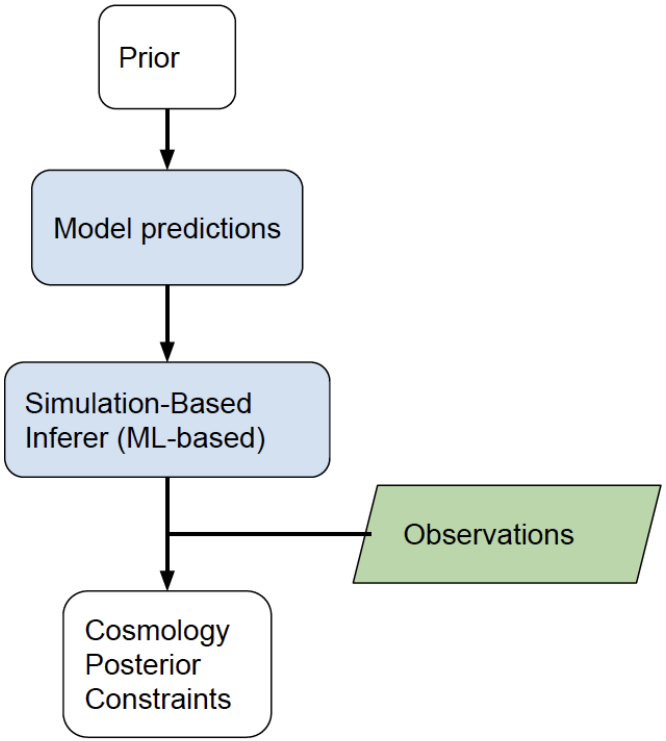
SBI VS the more traditional MCMC approach

The more traditional “likelihood” -based Markov Chain Monte Carlo (MCMC)



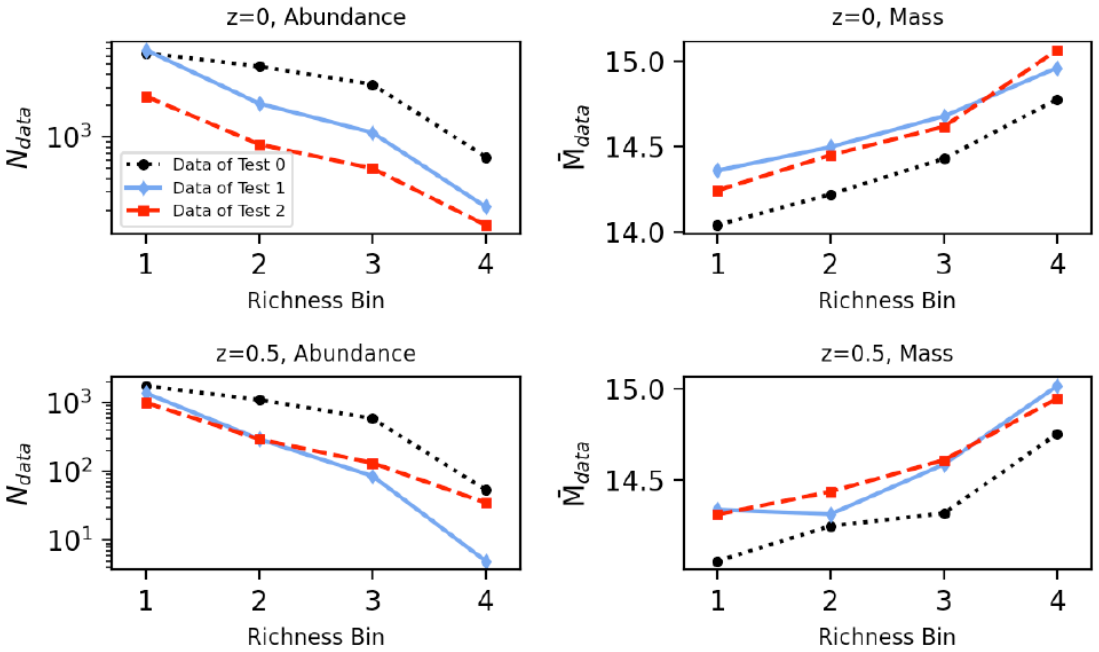
- Likelihood-based sampling needs HPC to derive posterior constraints.
- There is a “scaling” ceiling in MCMC.

Simulation Based Inference (SBI) (Cranmmer 2020)



- Generate training “simulations” is the most computationally intensive component.
- Each simulation takes minutes to generate, but completely independent from other simulations. Well suited for HTC.

Setting up the analysis: Build training data and testing data sets on OSPool



Visualization of some of the training data vectors.
** We assumed no astrophysical systematic effects. Will build more complicated models for future applications!

The cluster cosmology observables:

- A 16 element data vector — Galaxy cluster counts in richness bins, and average masses in richness bins.

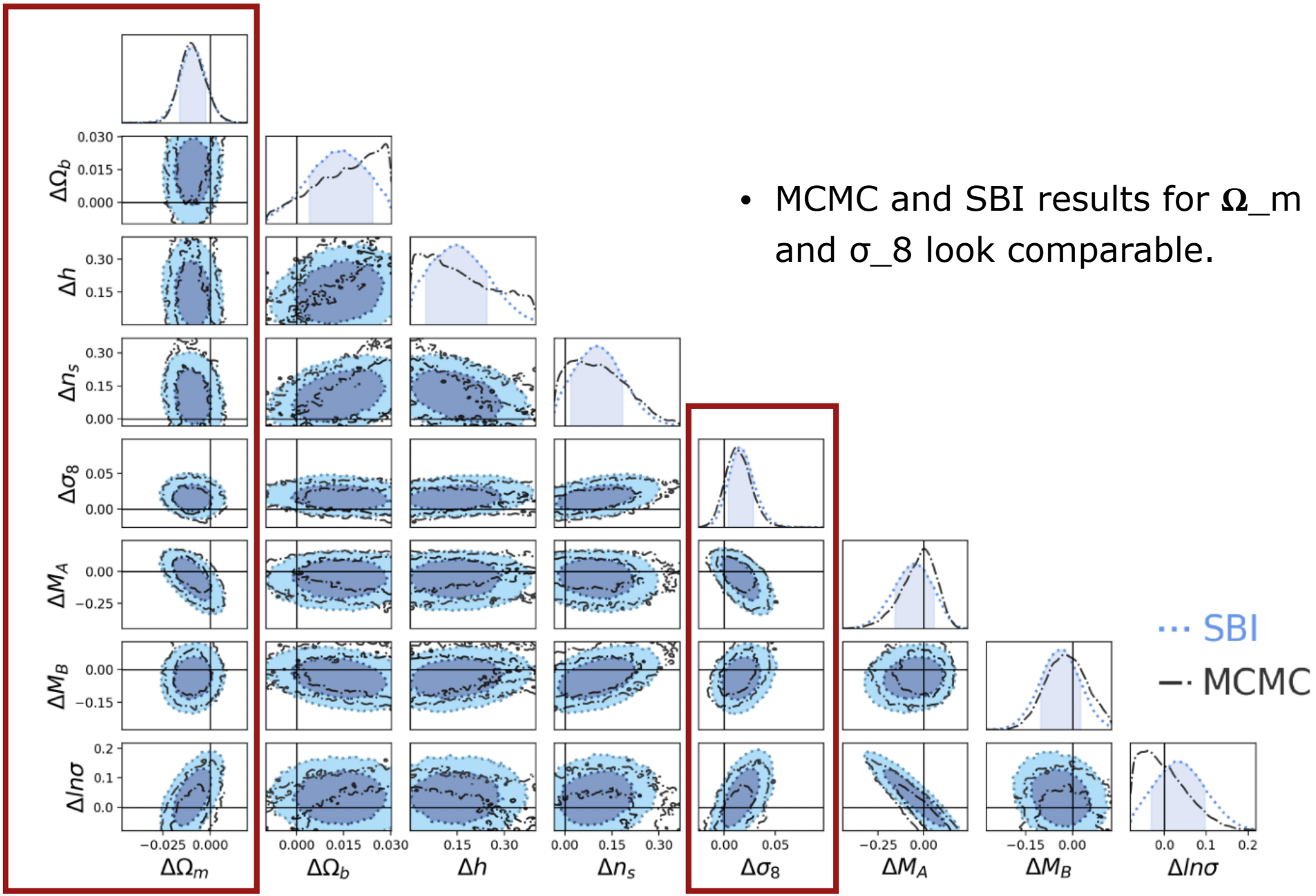
The training/testing samples:

- Sample cosmological and astrophysical parameters from flat distributions.
- Run the analytical "simulation" process to generate training/testing samples.
- Each simulation generation is an independent process. Ran on OSPool for O(100,000) samples, ~O(1,000) CPU hours.

Inference:

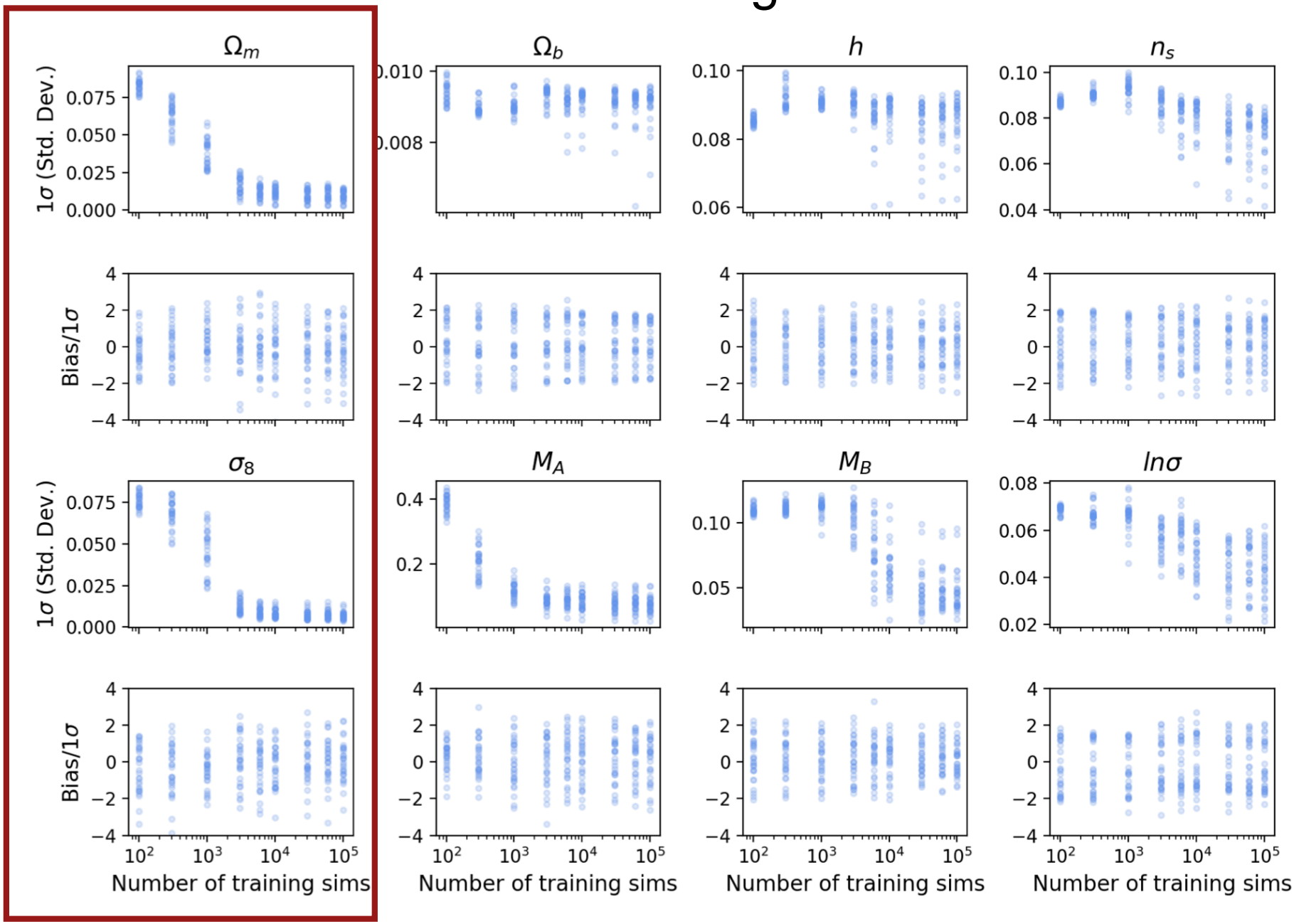
- Training, testing and further analyses all done on a laptop.

SBI and MCMC results are comparable.



- MCMC and SBI results for Ω_m and σ_8 look comparable.

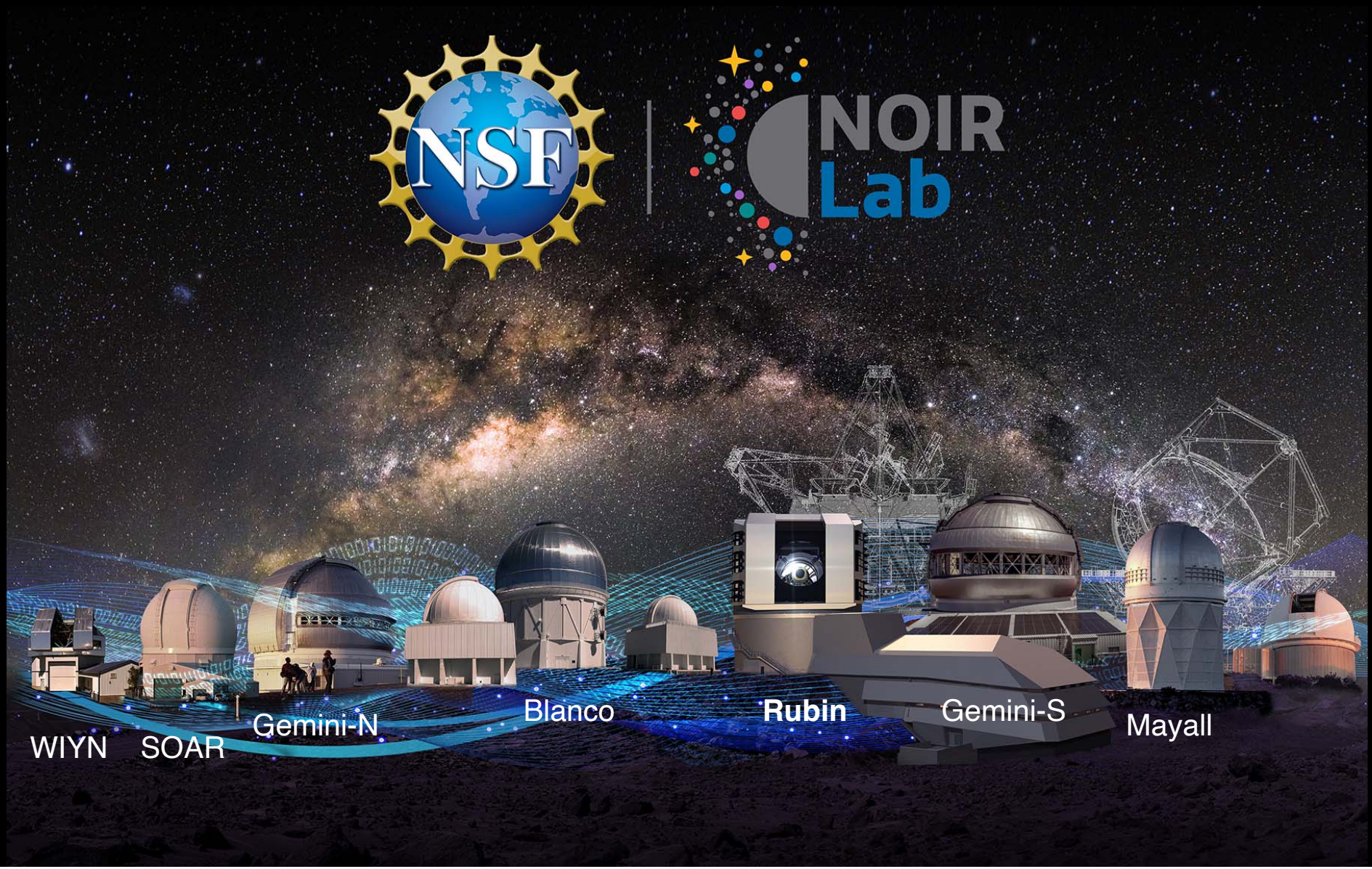
Posterior accuracies change when increasing the number of training simulations.



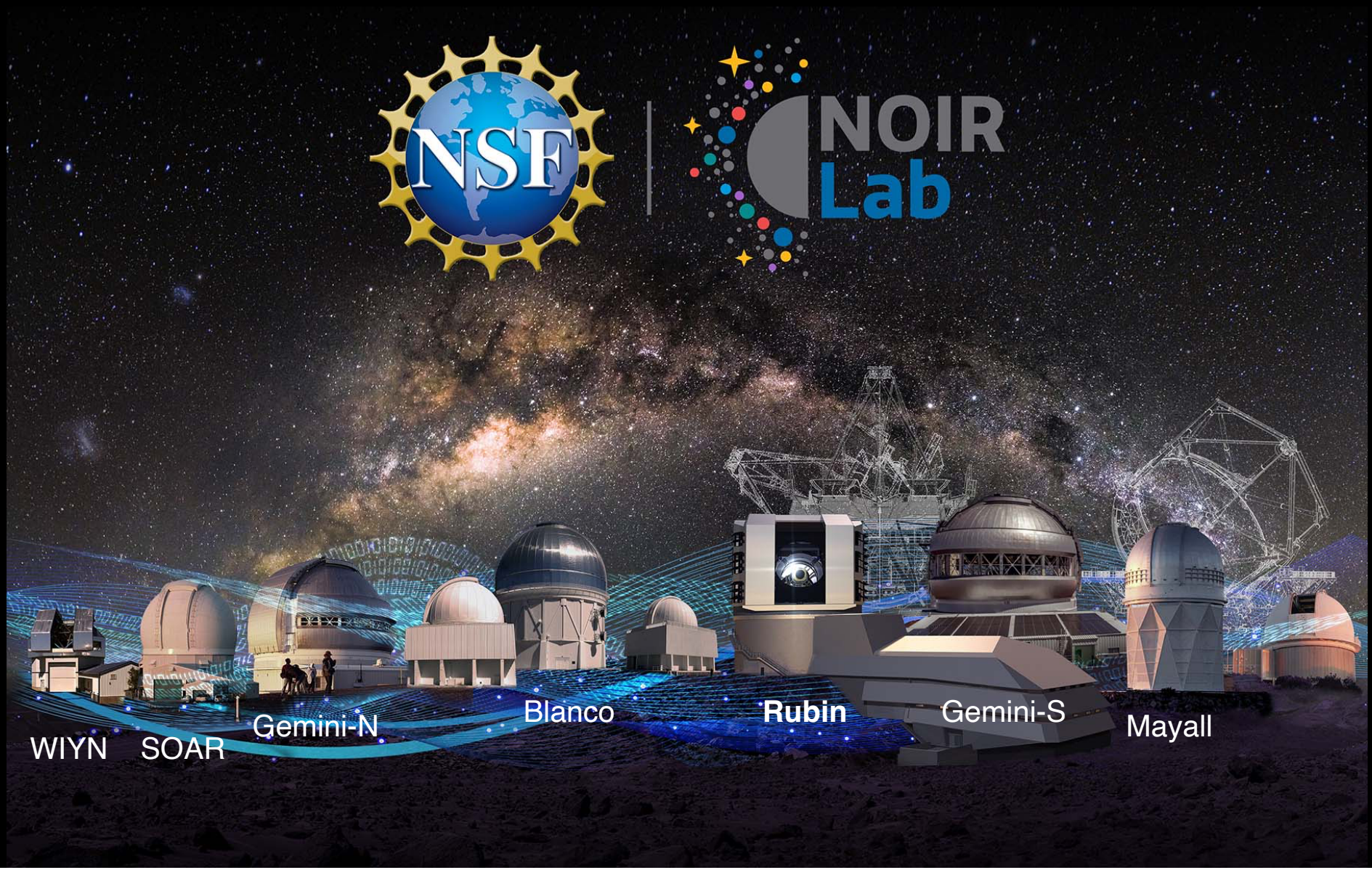
Thank you OSPool!

- SBI method can be accurate enough for a cluster cosmology application. The results are generally comparable to MCMC.
- The computing intensive part is in generating the training data sets. For a galaxy cluster cosmology application, it's well suited for high-throughput computing.

Why am I really here?
Learn to put data/users close to compute.



Peta-Bytes of (non-Rubin) Data holdings (from the Gemini, Blanco, Mayall telescopes and others) at NOIRLab available through **Astro Data Lab** and **Astro Data Archive.**



ASTRO DATALAB

NOIRLab Surveys

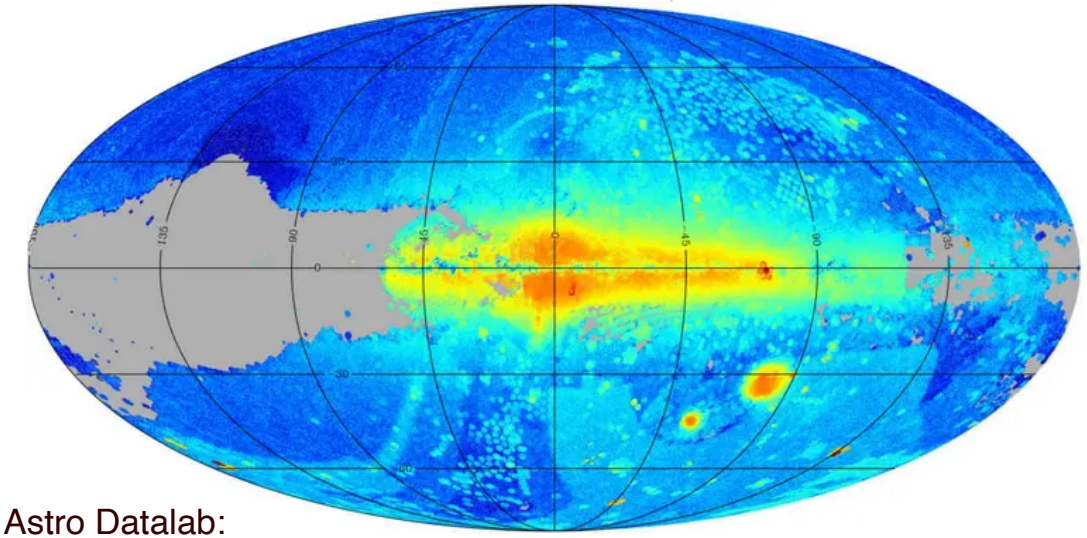
- Buzzard ›
- DECaPS ›
- DELVE ›
- DES ›
- DESI ›
- LS ›
- NSC ›
- SMASH ›

Gemini LLPs

- GNIRS-DQS ›
- GOGREEN and GCLASS ›

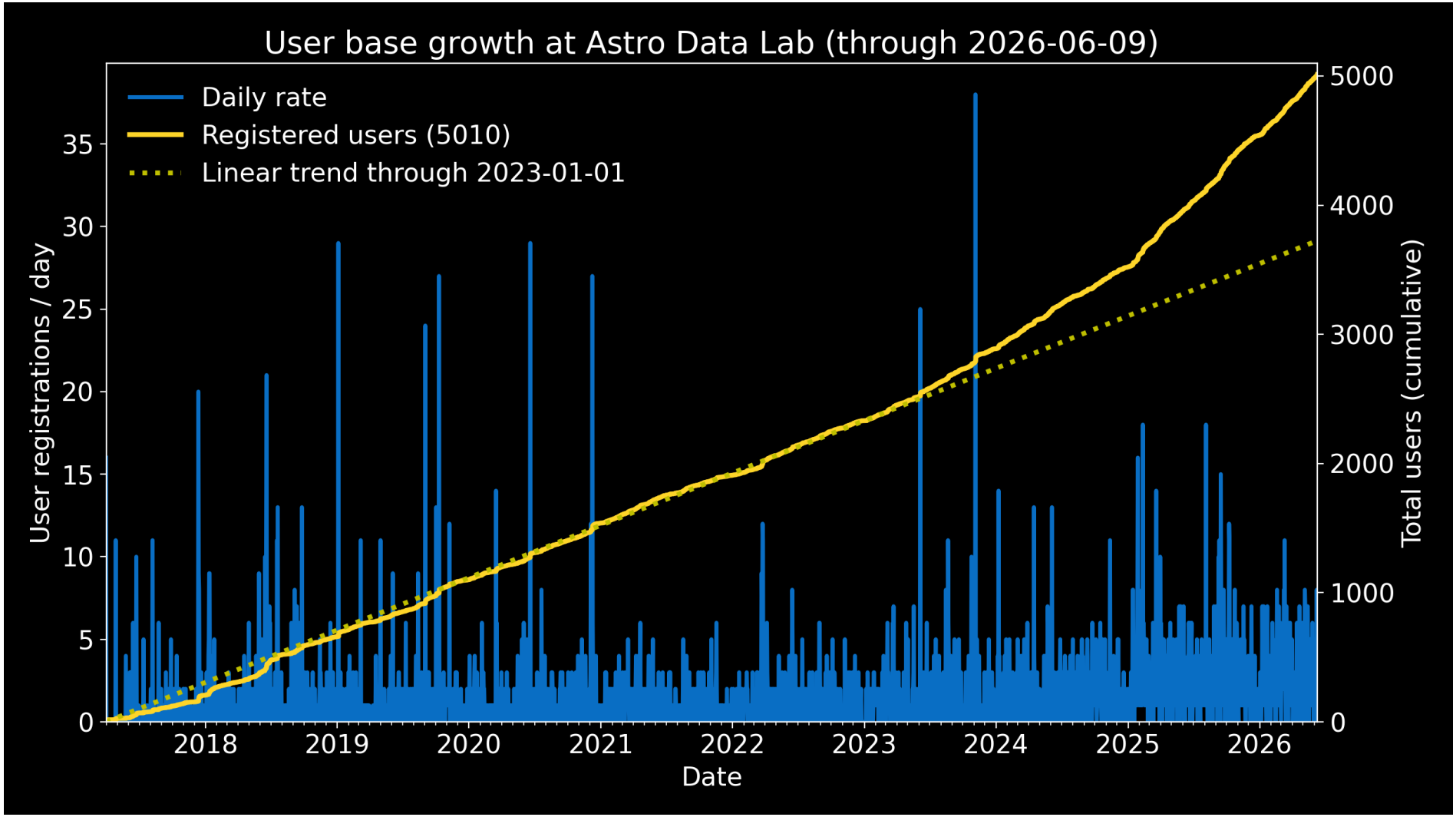
External Surveys

- AllWISE ›
- CatWISE ›
- DAD ›
- DeMCELS ›
- Euclid ›
- Gaia ›
- Hipparcos ›
- KS4 ›
- LSST SIM ›
- PGIR ›
- PHAT ›
- PHATTER ›
- SDSS ›
- SGA ›
- SkyMapper ›
- S-PLUS ›
- 2MASS ›
- Tycho-2 ›
- UKIDSS ›
- unWISE ›
- USNO ›
- VHS ›
- VMC ›

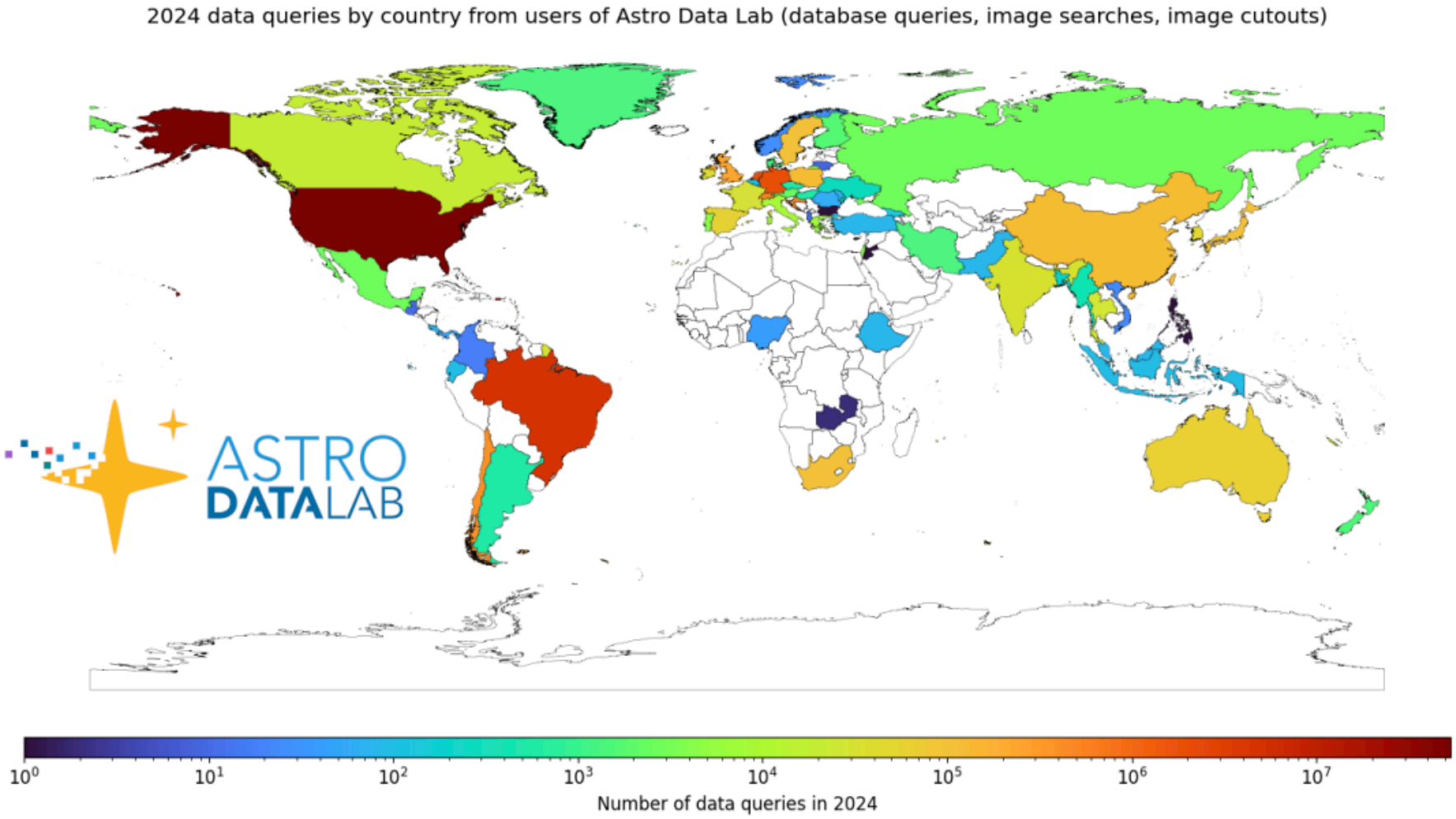


PI of Astro Datalab:
Robert Nikutta

Peta-Bytes of (non-Rubin) Data holdings at NOIRLab
Available through **Astro Data Lab** and Astro Data Archive



- Over **5,000** registered users from over 90 countries.
- Several ways to access and use the data: a Jupyter server, a web interface and a command line interface tool etc. **Tens of millions** of data queries each year.
- We are interested to put the users closer to HPC/HTC style compute (currently only modest local Hardware).



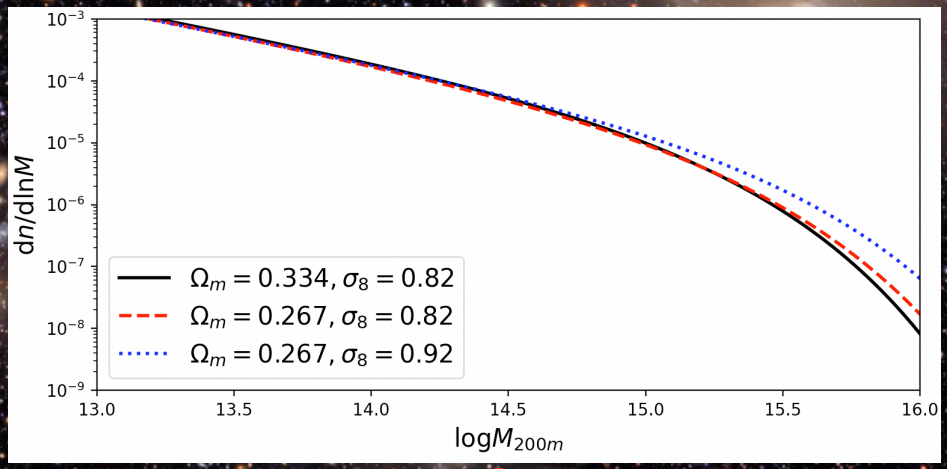
In 2024: 64.2M data queries from at least 72 countries, 82% from US

- Over **5,000** registered users from over 90 countries.
- Several ways to access and use the data: a Jupyter server, a web interface and a command line interface tool etc. **Tens of millions** of data queries each year.
- We are interested to put the users closer to **HPC/HTC style compute** (currently only modest local Hardware).

Back-up slides

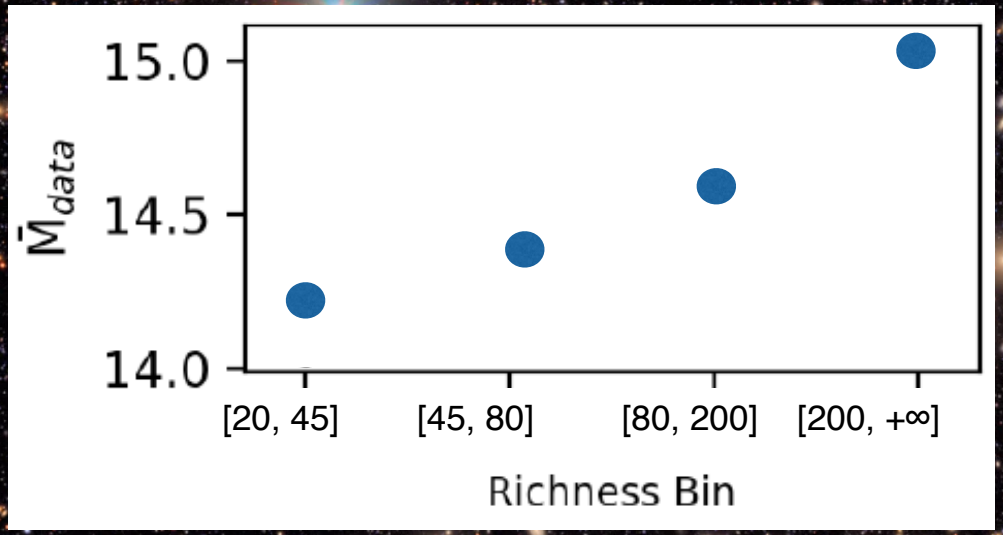
Intro: Galaxy Cluster Cosmology

"Halo mass function"

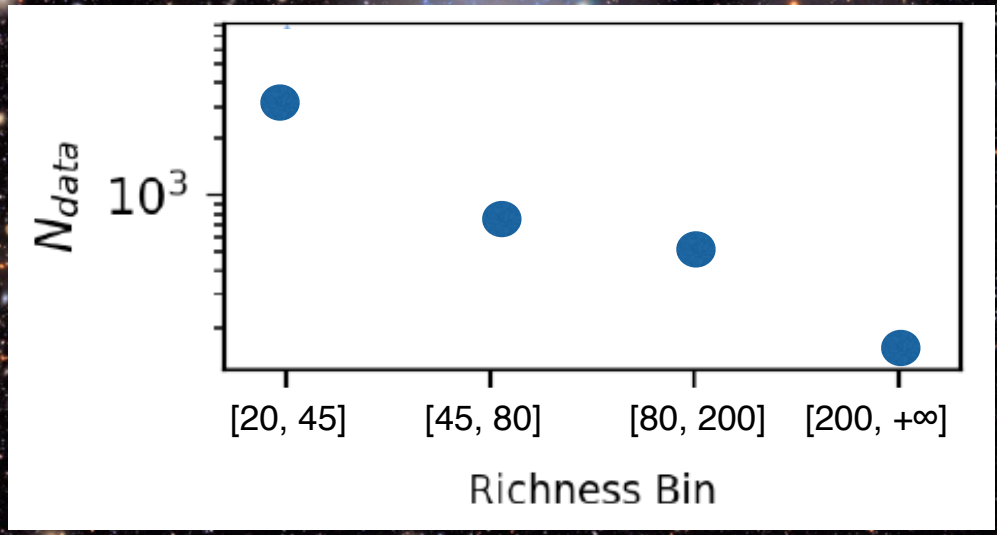


Potential Observables:

- Galaxy cluster counts,
- Average masses,
- Other quantities, e.g., different kinds of clustering signals.



+



Credit: Vera Rubin Observatory