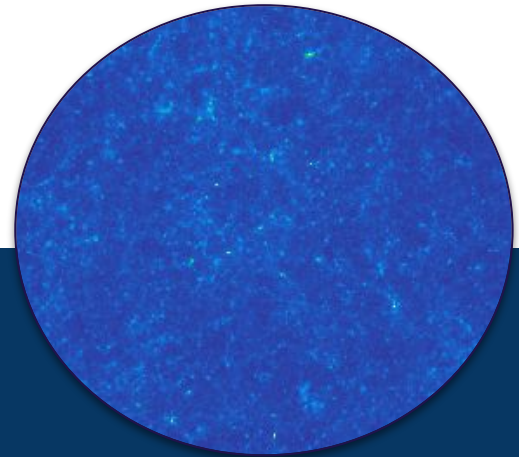
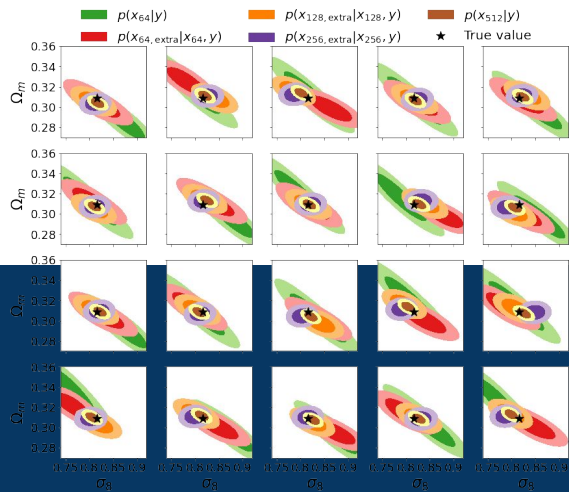


# Scientific discovery and ML in cosmology



Uroš Seljak  
UC Berkeley  
Lawrence Berkeley  
National Lab

# Outline

- ▷ **AI for Science: Artificial Intelligence (Machine Learning)** methods applied to (physical) sciences. Sometimes these are off the shelf methods, but to maximize their impact they **need to be developed specifically for physics applications** (e.g. physics symmetries, spatial structure etc.)
- ▷ Three tasks: **data generation, inference and anomaly detection**
- ▷ Each have a different training algorithm, but can be unified in a single method
- ▷ Ultimate goal: optimal data analysis (optimal inference) for **scientific discovery**
- ▷ Optimal: achieve smallest error (**generalization**), robust against unknown contaminations (**robustness**), have reliable verifications such that scientists can accept the results (**trustworthy**)



**Part 1:**

Introduction to AI/ML

# What is covered by AI/ML and by whom?

## A typical sample of topics at ICML or Neurips (leading ML conferences)

- General Machine Learning (active learning, clustering, online learning, ranking, reinforcement learning, supervised, semi- and self-supervised learning, time series analysis, etc.)
- Deep Learning (architectures, generative models, deep reinforcement learning, etc.)
- Learning Theory (bandits, game theory, statistical learning theory, etc.)
- Optimization (convex and non-convex optimization, matrix/tensor methods, stochastic, online, non-smooth, composite, etc.)
- Probabilistic Inference (Bayesian methods, graphical models, Monte Carlo methods, etc.)
- Trustworthy Machine Learning (accountability, causality, fairness, privacy, robustness, etc.)
- Applications (computational biology, crowdsourcing, healthcare, neuroscience, social good, climate science, etc.)

Who are AI practitioners? Mostly from computer science and statistics...

Growing community at the intersection of domain sciences and AI

Note: for the purpose of this talk AI=ML

# AI for Physical sciences

- 1) The goal of AI in physics is to extract scientific information using a data driven approach. Typical application is **data inference**, e.g. learn about probability distribution of some **parameters  $y$**  from **data  $x$**   $p(y|x)$
- 2) In AI this is done by using **training data  $(X,Y)$**  to learn  $p(y|x)$  of parameters  $y$  we wish to extract from some real data  $x$  which is not in the training data set
- 3) Often we do not have real data to train on. We may however have simulations. In this case simulations become training data. The corresponding concept is called **Simulation Based Inference** (also known as Likelihood Free Inference, Implicit Likelihood etc.)
- 4) While this has been very successful in some fields, in physics there are very few success stories (Experimental HEP being one). As we improve our AI methods this is likely to change (e.g. cosmology)!

# Discriminative (supervised) learning

1) Here we wish to determine parameters  $y$  from data  $x$ : a common

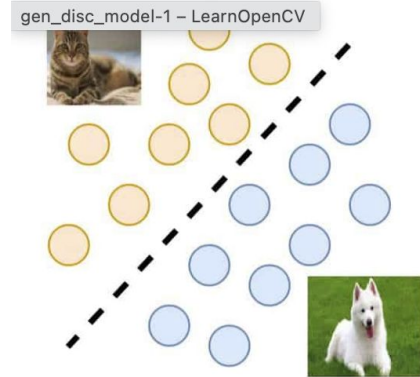
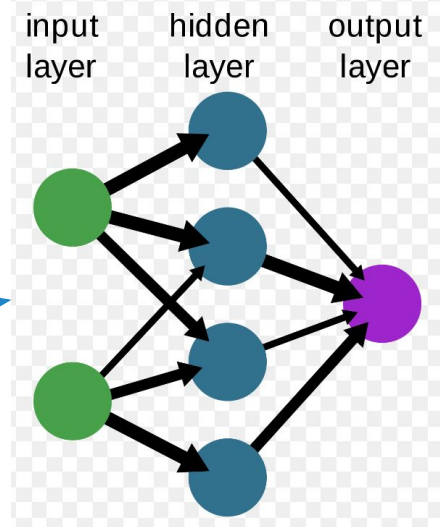
**discriminative training objective**  $\min_w \sum (y_w(x) - y_{\text{true}})^2$

over all the training data  $(X, Y_{\text{true}})$ .

2) Typically  $y_w(x)$  will be a parametrized **Neural Network** that takes in the data  $x$  and outputs  $y_w(x)$ . NN may have several layers (deep learning), but at each layer the operations are very simple: take a linear combination of previous layer inputs times weights and perform a very simple nonlinear operation  $f$  in the end, so each layer output is  $f(\mathbf{w}x)$ . We learn the weights  $w$  to minimize the loss.

3) **Posterior Estimation**: more generally, we can predict posterior  $p(y|x)$ , which may be a Gaussian defined by the mean  $y_{\text{mean}}(x)$  and the variance  $\sigma^2(x)$ , or some other distribution.

4) When  $y$  is discrete this is a **classification** problem. In this case  $p(y|x)$  represents the predicted class  $y$  probability given  $x$ .



**Discriminative**

# Generative (unsupervised) learning

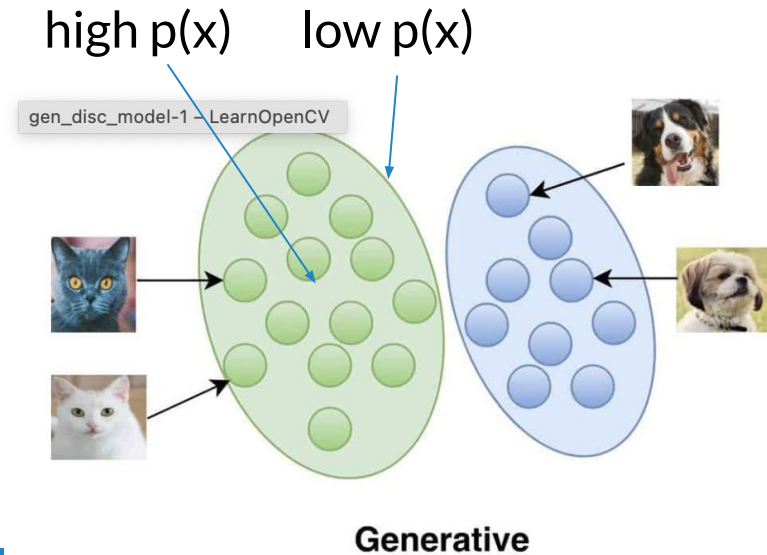
1) The goal is to generate new fake data, so it is like a simulation. This is a very active area, with many applications (e.g. ChatGPT, Dall-E2...). We can describe it as **drawing samples of data  $x$  from some probability distribution  $p_w(x)$** . Natural for unlabeled data (unsupervised learning).

(one possible) Generative training objective:  $\text{maximize}_w \log p_w(x)$

2) For data inference more interesting is **likelihood estimation**, density learned conditionally on  $y$ ,  $p(x|y)$ .

Use Bayes Theorem  $p(y|x) = p(x|y)p(y)/p(x)$

3) enables anomaly detection: an outlier has low  $\max_y p(x|y)$ , generalization of chi squared test



# Generative versus discriminative learning

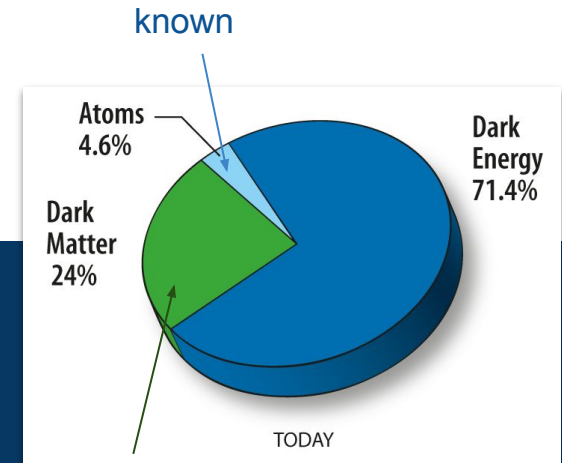
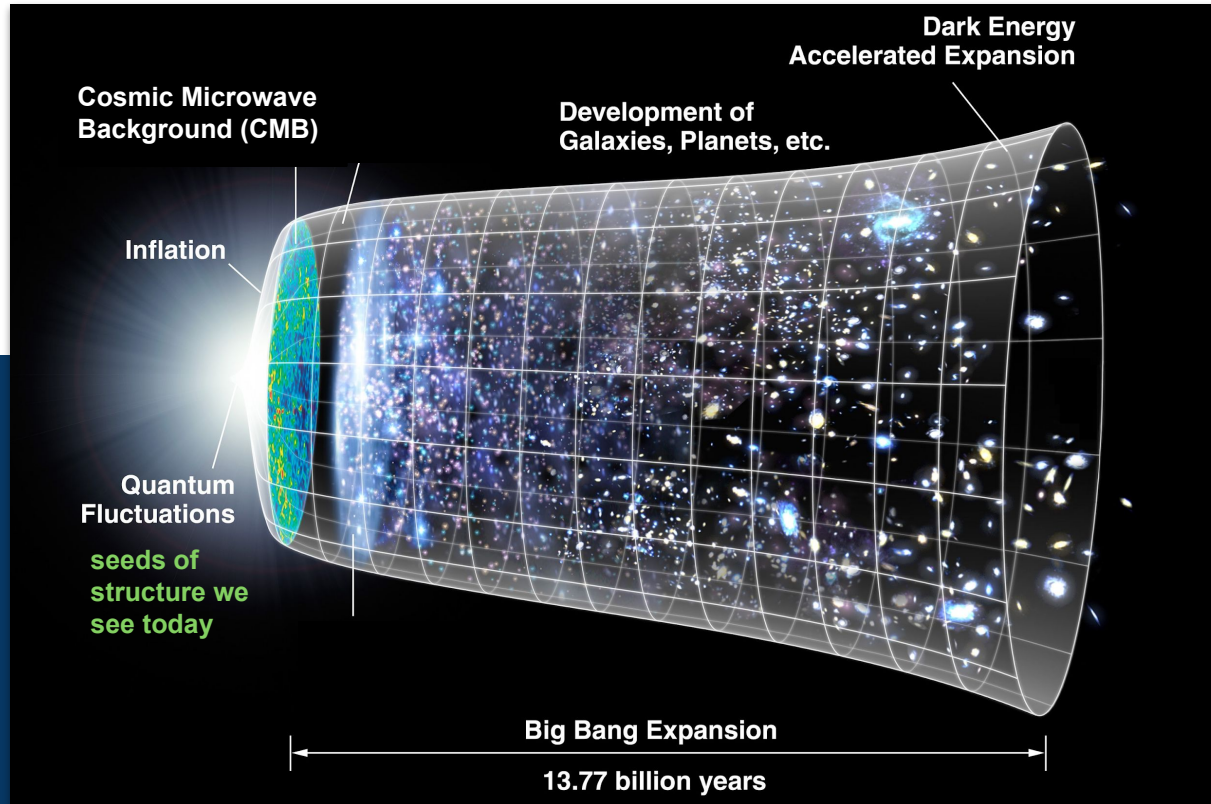
- 1) Both methods can give  $p(y|x)$ , but which is better? For regression or classification the traditional answer (originated by Vapnik) is that discriminative is always better
- 2) Later work (e.g. Ng & Jordan 2003) gives a more nuanced answer: discriminative is always at least as good or better in the limit of large data, but generative may reach its own asymptotic limit faster (using less training data).
- 3) Modern view (this talk): hybrid generative+discriminative training gives the best of both worlds



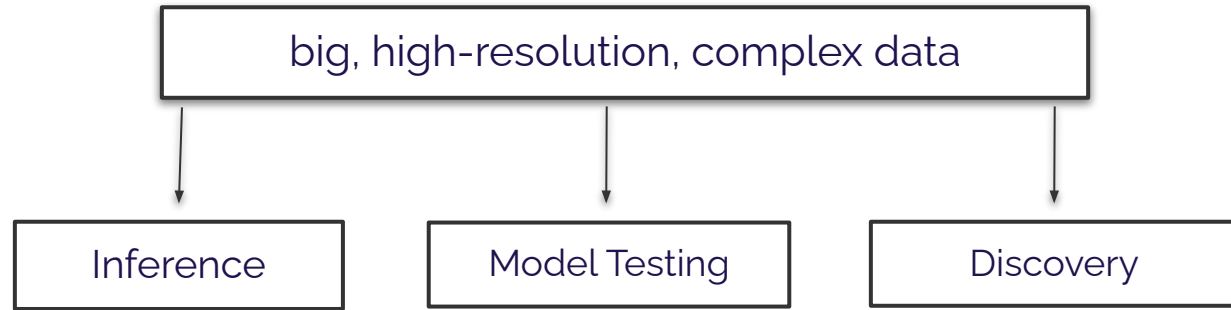


Introduction to cosmology

# Current Cosmological Standard Model



# Open questions and how to answer them



Nature of dark energy? Time-dependence?

unknown signals

Mass of neutrinos? ( $M_\nu$ )

new physics

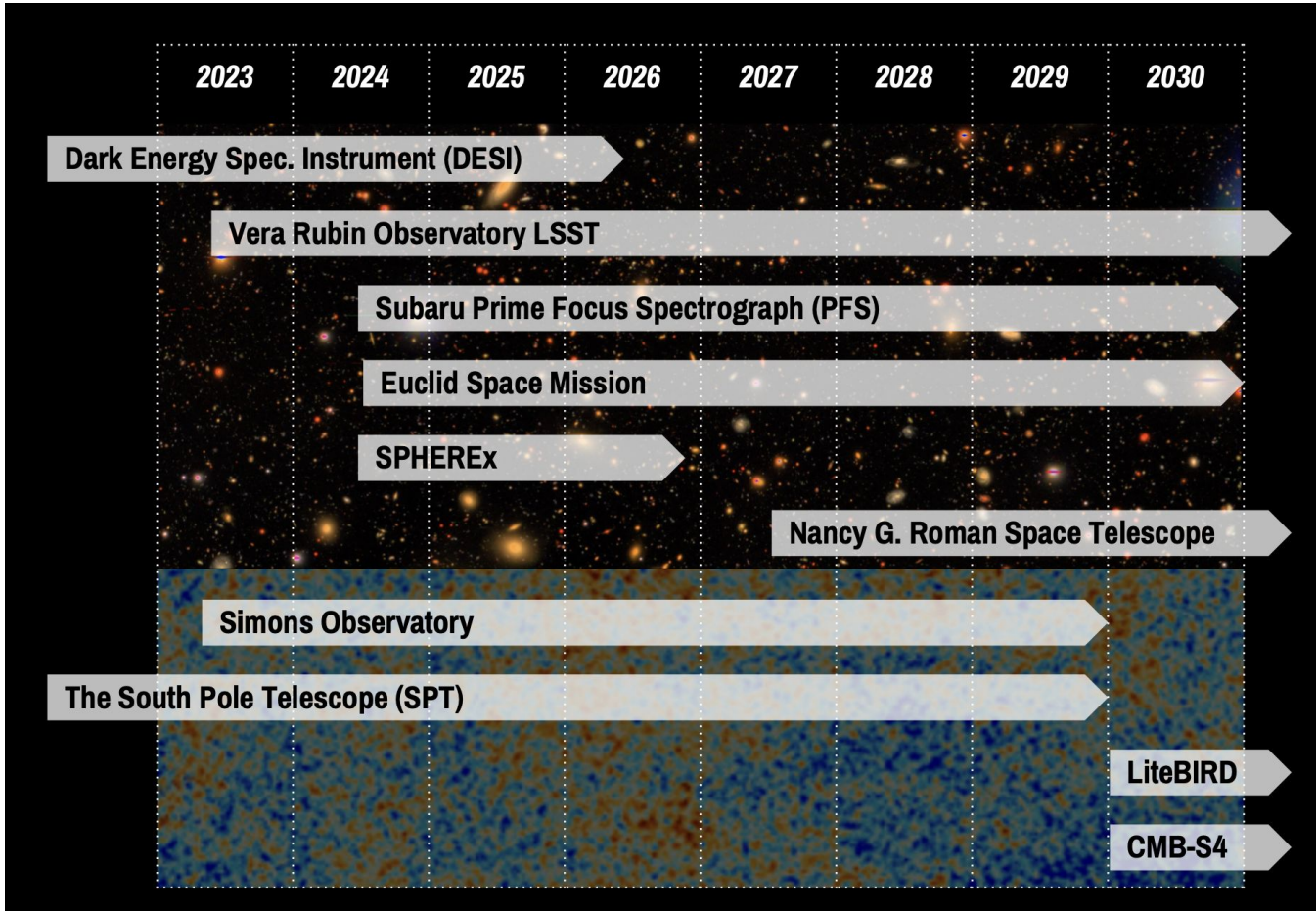
Tensions between parameters inferred from different observables?

Total matter content in the Universe,  $\Omega_m$ ?

transient signals

Signatures of inflation?

# We are in the golden era of cosmology: huge investments in new experiments



# New surveys

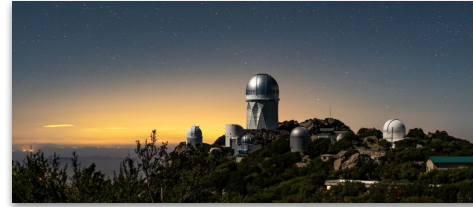
## Vera Rubin Observatory (LSST)



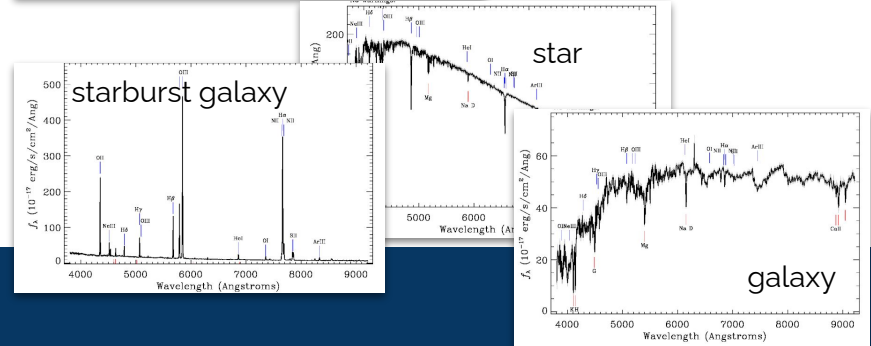
**20 billion** galaxies  
**17 billion** stars  
**20 terabyte** data/day



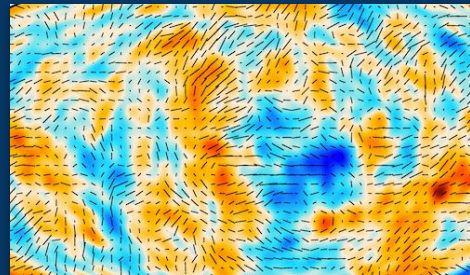
## Dark Energy Spectroscopic Instrument (DESI)



**40 million** galaxies  
**10 million** stars



## CMB-S4



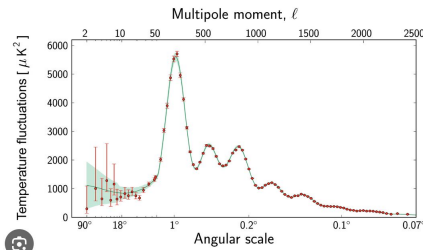
Cosmic Microwave Background

**50 PB** total database

# Current cosmological data analysis: 2 point correlations

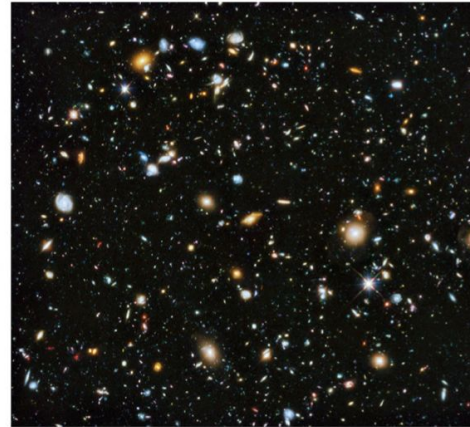
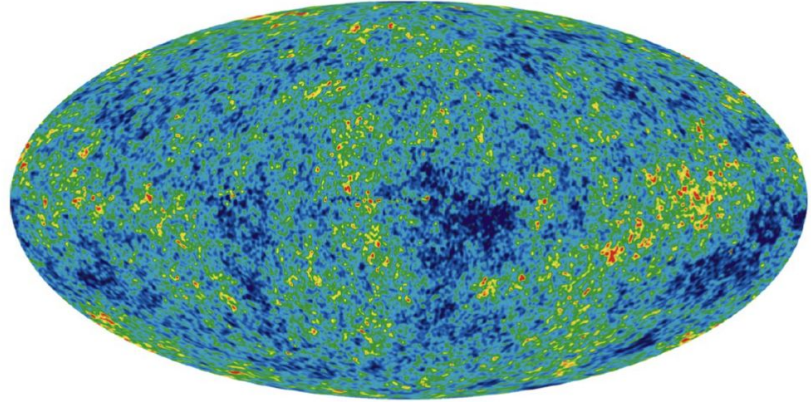
## Gaussian density field

Fully described by the power spectrum



## Non-Gaussian density field

$P(k)$ ,  $B(k)$ , peaks, voids, ...



# Many proposals for higher order statistics

- Lensing bispectrum (Coulton+2018)
- Lensing Minkowski functionals (Marques+2018)
- Lensing peak counts (Li+2018, Ajani+2020)
- Lensing probability density function (Liu+2020)
- Matter probability density function (Uhlemann+2020)
- Redshift-space bispectrum (Hahn+2020)
- Marked power spectrum (Massara+2021)
- Wavelets (Cheng+2021, Valogiannis+2021)
- Voids (Bayer+2021, Kreisch+2021)
- ...

Constraining neutrino mass with tomographic weak lensing one-point probability distribution function and power spectrum  
Jia Liu\* and Mathew S. Madhavacheril

Constraining neutrino mass with weak lensing Minkowski Functionals  
Gabriela A. Marques,<sup>a,1</sup> Jia Liu,<sup>b</sup> José Manuel Zorrilla Matilla,<sup>c</sup> Zoltán Haiman,<sup>c</sup> Armando Bernui<sup>a</sup> and Camila P. Novaes<sup>a</sup>

**Fisher for complements: Extracting cosmology and neutrino mass from the counts-in-cells PDF**








Cora Uhlemann<sup>1,2</sup>, Oliver Friedrich<sup>3,4</sup>, Francisco Villaescusa-Navarro<sup>5,6</sup>, Arka Banerjee<sup>7,8,9</sup>, Sandrine Codis<sup>10</sup>

**Constraining  $M_L$  with the Bispectrum I: Breaking Parameter Degeneracies**

ChangHoon Hahn <sup>a,b</sup> Francisco Villaescusa-Navarro<sup>c,d</sup> Emanuele Castorina<sup>a,b</sup> Roman Scoccimarro<sup>c</sup>

DETECTING NEUTRINO MASS BY COMBINING MATTER CLUSTERING, HALOS, AND VOIDS  
ADRIAN E. BAYER<sup>1,2,\*</sup>, FRANCISCO VILLAESCUSA-NAVARRO<sup>3,4,†</sup>, ELENA MASSARA<sup>5,4</sup>, JIA LIU<sup>1,2,6</sup>, DAVID N. SPERGEL<sup>3,4</sup>, LICIA VERDE<sup>7,8</sup>, BENJAMIN D. WANDEL<sup>9,10,4</sup>, MATTEO VIEL<sup>11,12,13,14</sup>, SHIRLEY HO<sup>4,3,15</sup>

**The GIGANTES dataset: precision cosmology from voids in the machine learning era**

CHRISTINA D. KREISCH <sup>1</sup>, ALICE PISANI <sup>1</sup>, FRANCISCO VILLAESCUSA-NAVARRO <sup>1</sup>, DAVID N. SPERGEL <sup>1,2</sup>, BENJAMIN D. WANDEL <sup>2,3,4</sup>, NICO HAMAUS <sup>5</sup> AND ADRIAN E. BAYER <sup>6,7</sup>

Constraining neutrino masses with weak-lensing multiscale peak counts

Virginia Ajani,<sup>1,\*</sup> Austin Peel,<sup>2</sup> Valeria Pettorino,<sup>1</sup> Jean-Luc Starck,<sup>1</sup> Zack Li,<sup>3</sup> and Jia Liu<sup>4,3</sup>

Constraining Neutrino Mass with the Tomographic Weak Lensing Bispectrum  
William R. Coulton

**Using the Marked Power Spectrum to Detect the Signature of Neutrinos in Large-Scale Structure**

Elena Massara,<sup>1,2,\*</sup> Francisco Villaescusa-Navarro,<sup>3,2</sup> Shirley Ho,<sup>2,3,4</sup> Neal Dalal,<sup>5</sup> and David N. Spergel<sup>2,3</sup>

Constraining neutrino mass with tomographic weak lensing peak counts

# New Data require new Data Analyses

that

- 1) can handle such large datasets
- 2) Is optimal: can extract their full information content

Can we capitalize on the recent success of AI/ML and Statistical Inference? **YES!**

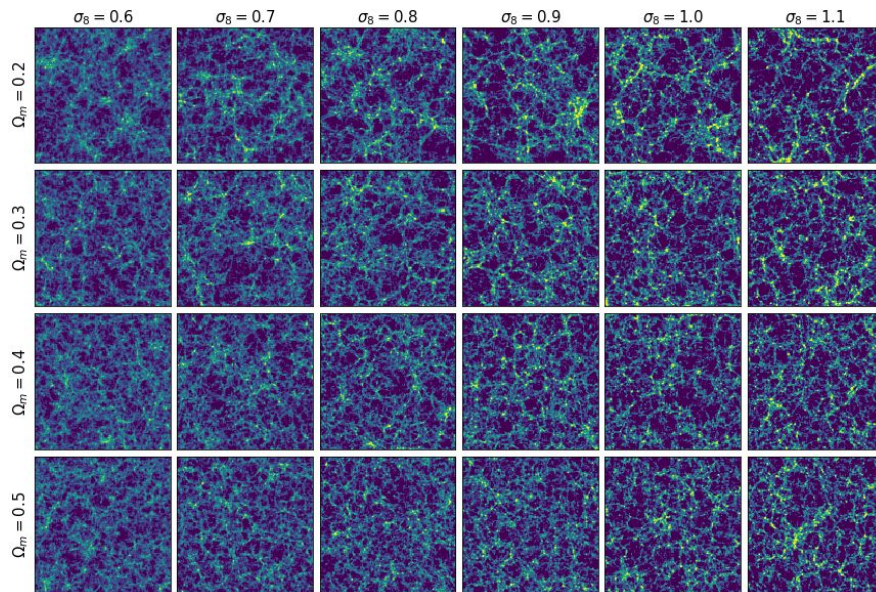
Can we take off the shelf AI/ML codes to do so? **NO!**

We need to account for symmetries, spatial correlations, high dimensionality and high stochasticity



# Stochastic and high dimensional nature of cosmology data

- 1) Each realization is different, information is in the correlations on all scales
- 2) Data is pixelized or voxelized, with  $O(10^7)$  or more dimensions
- 3) Each training dataset is an N-body simulation, very expensive
- 4) Very difficult to “interpolate” given a finite number of simulations
- 5) Our approach: learn the data structure first using generative learning, followed by generative+discriminative learning to reach optimality





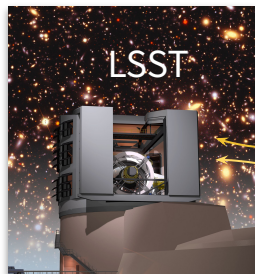
**Part 3:**

AI for Physics: Robust and Optimal Analysis of  
Weak Gravitational Lensing with Normalizing  
Flows

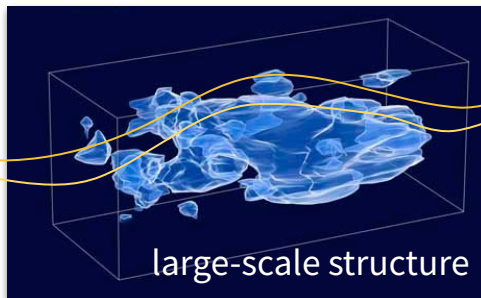
Work led by Biwei Dai



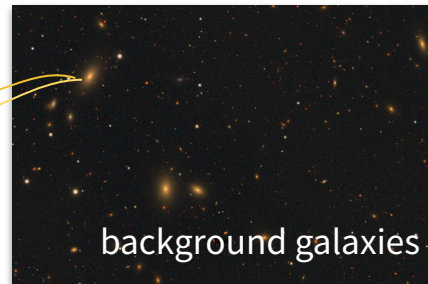
# Weak Lensing of Galaxies



LSST



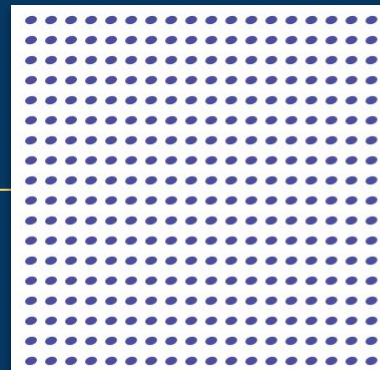
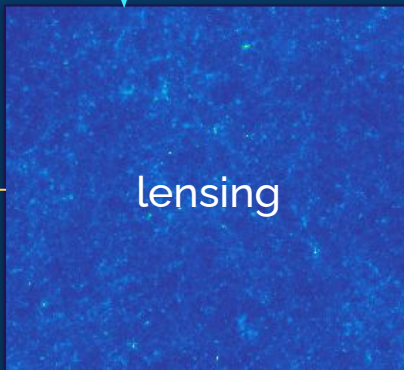
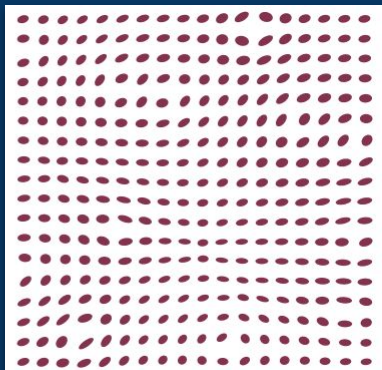
large-scale structure



background galaxies

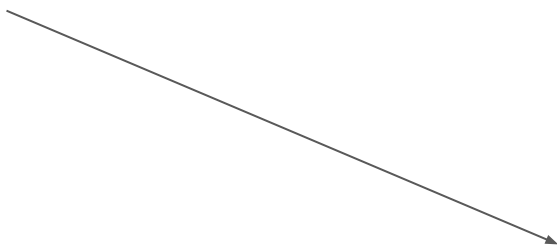
Rubin data size:  
 $10^9$  galaxies

*projection*



# Optimal cosmological data analysis

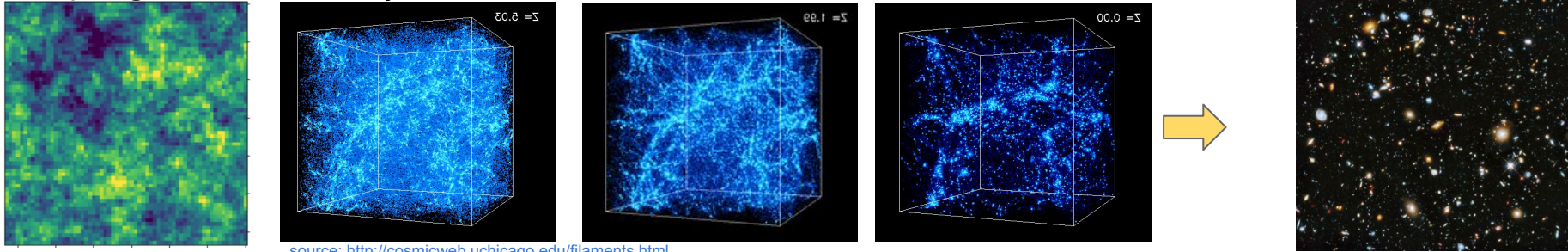
- **Optimal:** Extract **maximum** amount of information from the data
- **Bayesian Inference:** MCMC sampling of initial conditions and parameters
- **Supervised learning with AI/ML** to predict posterior  $p(y|x)$ : discriminative
- **Learning the likelihood  $p(x|y)$  directly with AI/ML: generative**
  - Con: High dimensionality of  $x$ .
  - Model: Normalizing Flows



Use symmetries and multiscale architecture to reduce the curse of dimensionality

# Generating training data with N-body simulations

cosmological structure formation: an N-body simulation and a galaxy formation model  
A lot of progress in recent years with differentiable fast simulations (FastPM, FlowPM, PWMD...)



**random initial fluctuations**

random realization is independent from cosmological parameters

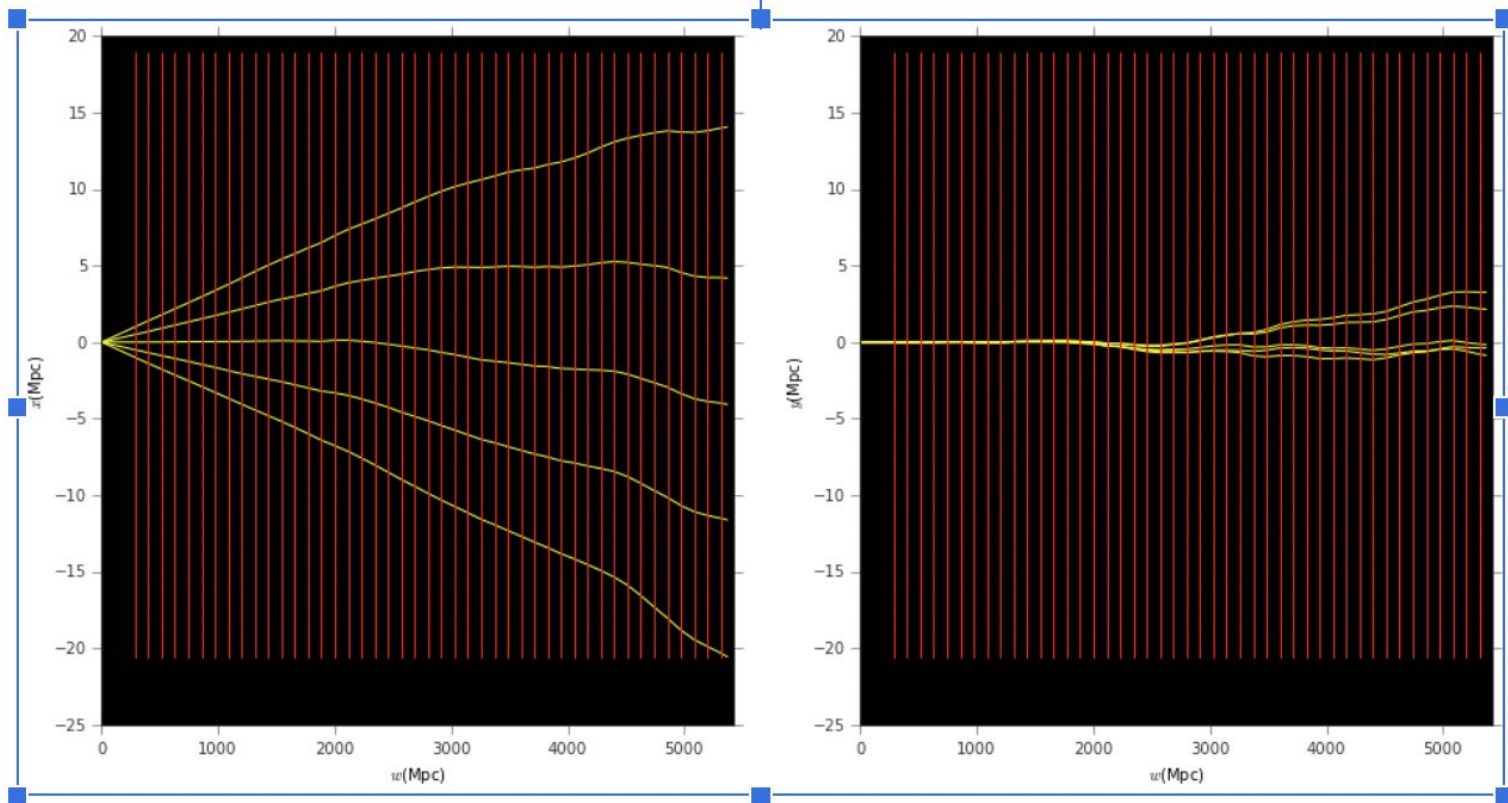
**structure formation**

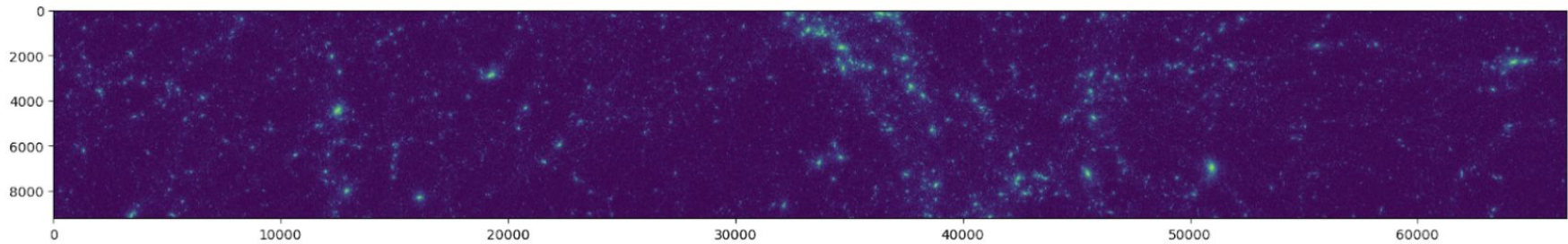
depend on cosmological parameters

**galaxies**

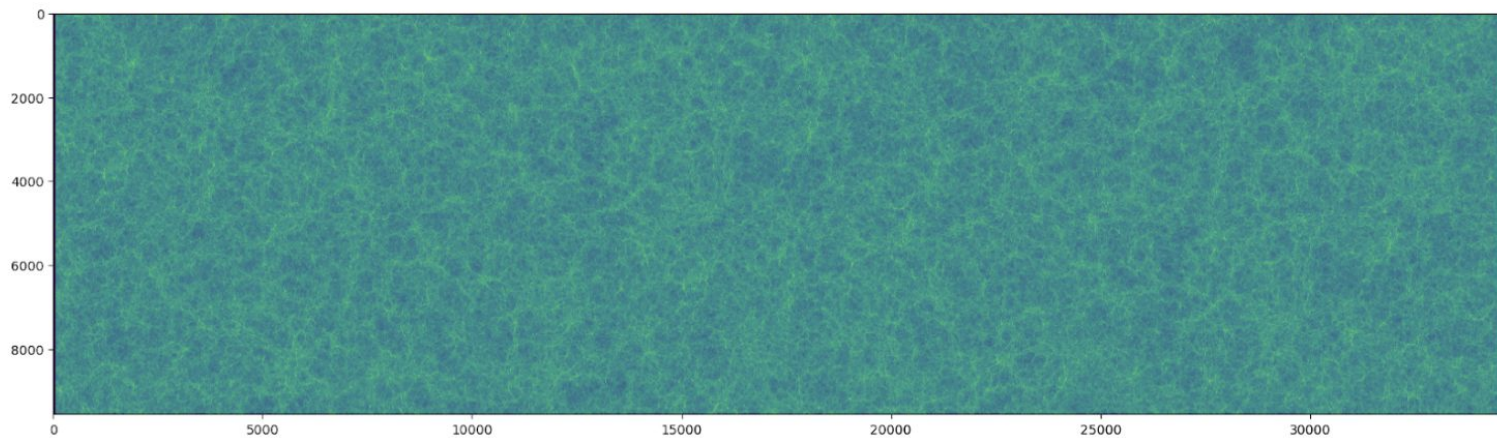
# Ray-tracing

- multi-lens-plane algorithm from LensTools package





Another example lensing plane at higher redshift:

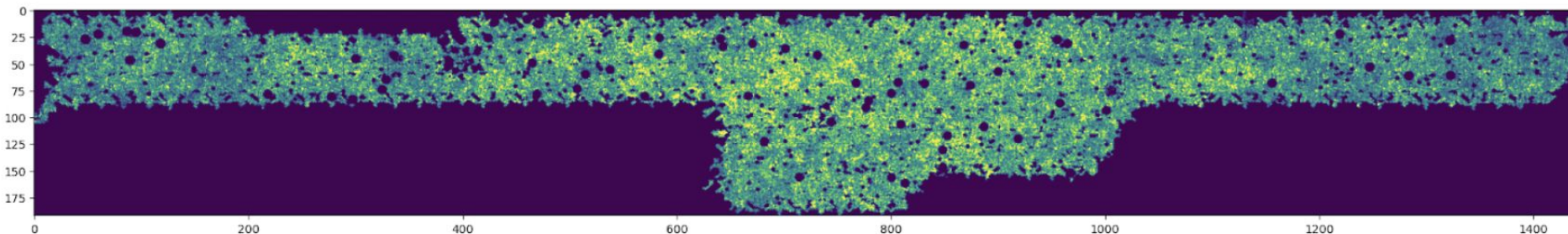


# Stack all the lens planes, add survey mask

HypersupremeCam (HSC) data on Subaru

Later Rubin (LSST) etc.

Real world complications easy to add: survey mask, noise etc.



We now have realizations of data  $x$

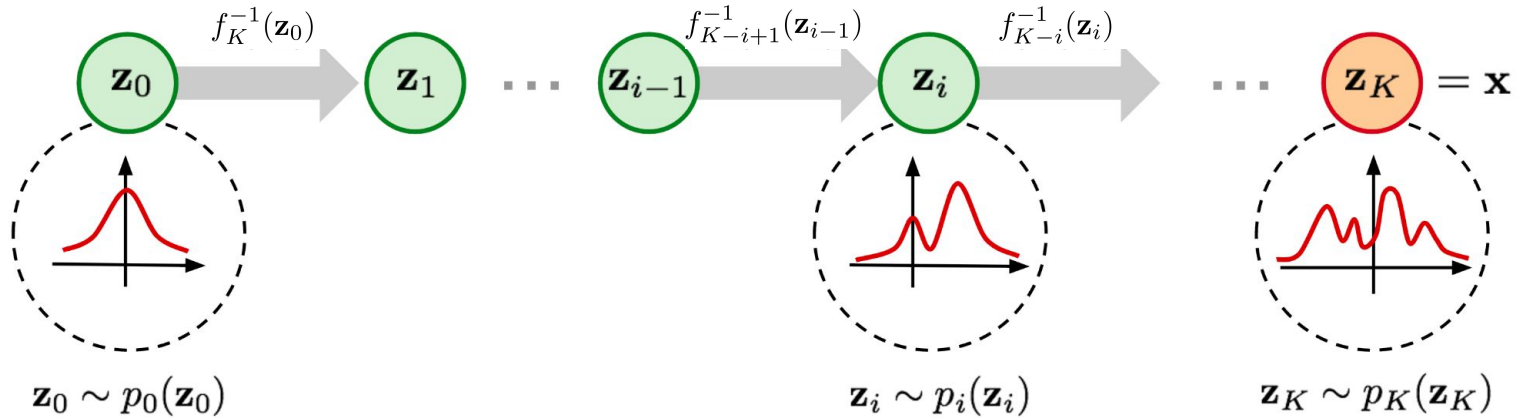
For each simulation we can make many maps by making different projections

Repeat the process for different values of cosmological parameters  $y$

$y$ : we can vary initial density amplitude, matter density, dark energy, neutrino mass etc.



# Normalizing Flows

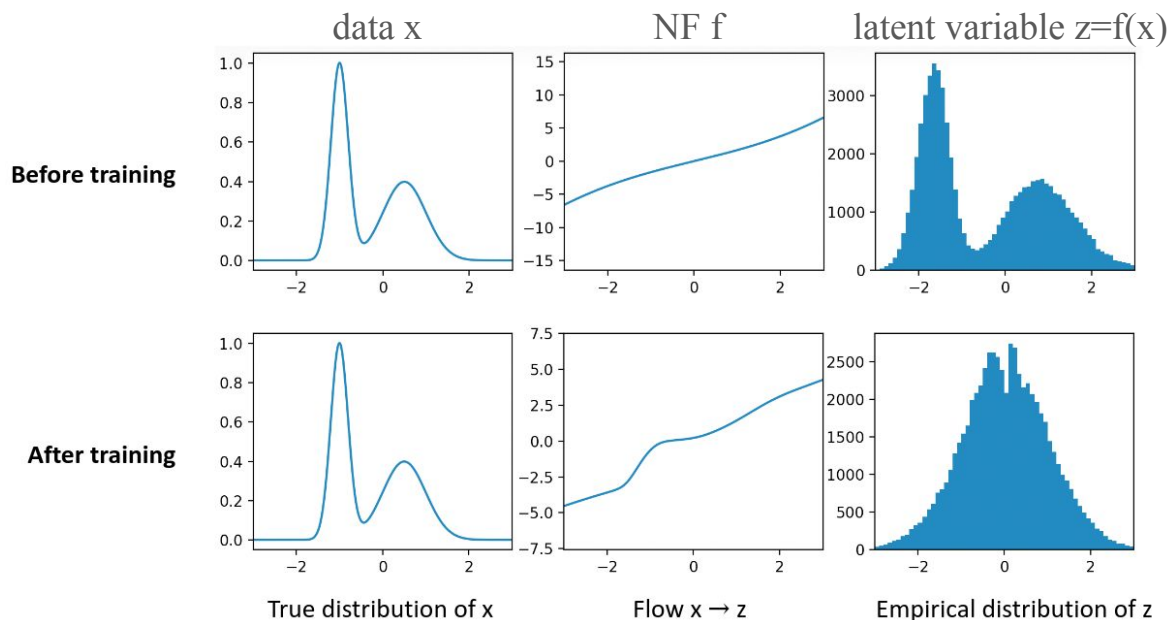


- Bijective mapping  $f$  between data  $\mathbf{x}$  and latent variable  $\mathbf{z}$  ( $\mathbf{z} = f(\mathbf{x})$ ,  $\mathbf{z} \sim \pi(\mathbf{z})$ )
  - Evaluate density:  $p(\mathbf{x}) = \pi(f(\mathbf{x})) |\det(df/dx)|$
  - Sample:  $\mathbf{x} = f^{-1}(\mathbf{z})$  ( $\mathbf{z} \sim \pi(\mathbf{z})$ )

Credit:  
<https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html>

# Normalizing Flow training

- 1D example



- Training objective: maximize  $\log p(x)$  over training data

$$p(x) = \pi(f(x)) |\det(df/dx)|$$

- Evaluate density:  $p(x) = \pi(f(x)) |\det(df/dx)|$

- Sample:  $x = f^{-1}(z)$  ( $z \sim \pi(z)$ )

# How to design Normalizing Flows for physics?

## NF by itself requires:

- Easy evaluation of the inverse
- Easy evaluation of the Jacobian determinant

## Physics applications require:

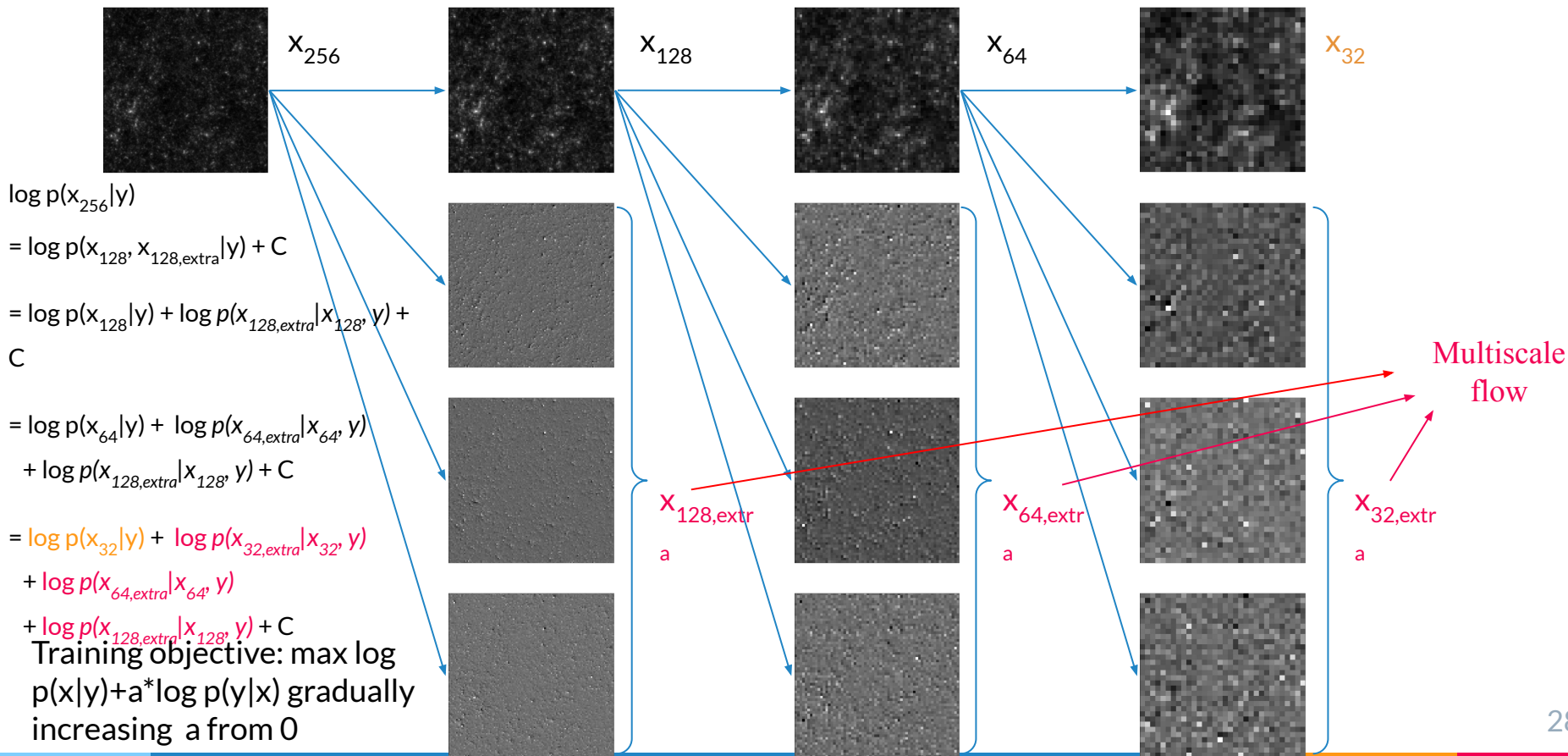
- Spatial Structure modeling, high dimensionality (pixels, voxels): coarse graining, multi-scale correlations
- Symmetries (translation, rotation etc.)
- Ability to learn from very stochastic data

How do we parametrize our NF, given these considerations?

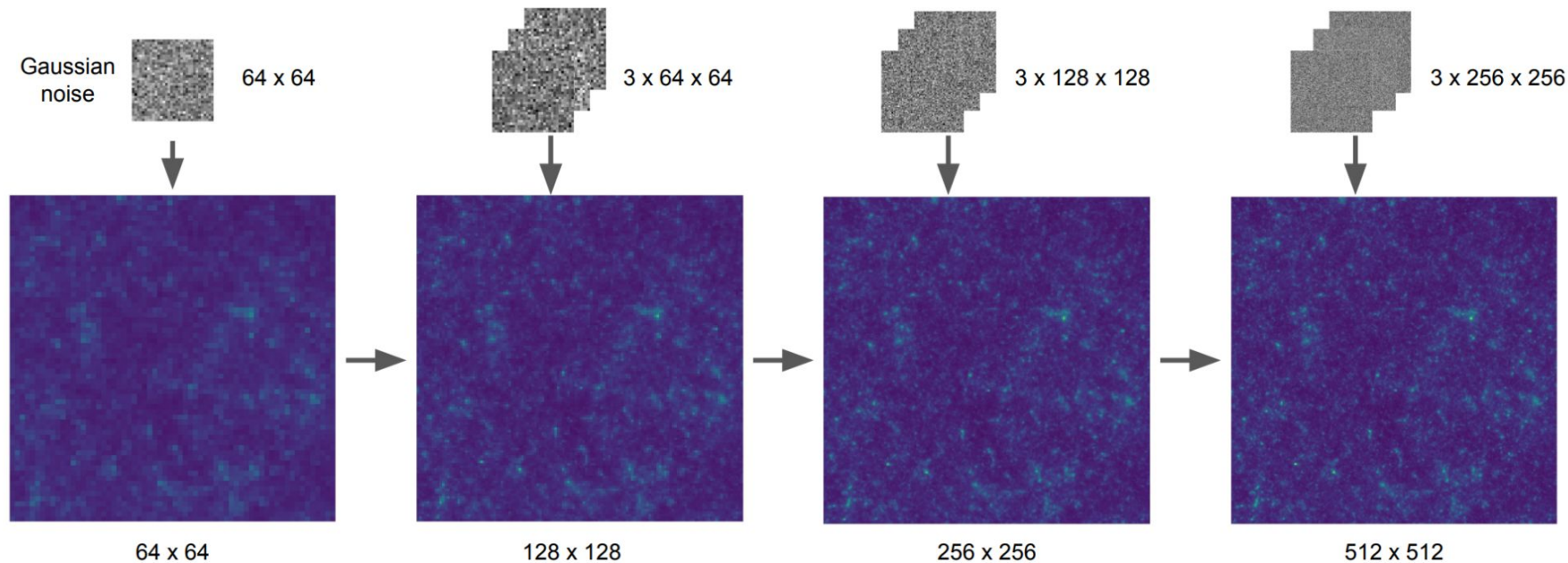
- 1) MultiScale Flow (Dai & Seljak 2023)
- 2) Translational and Rotational Normalizing Flow (Dai & Seljak 2021)

# Multiscale flow (MSF): a wavelet based flow (Dai & Seljak, 2023)

- Consider a cosmological field with  $256^2$  resolution:

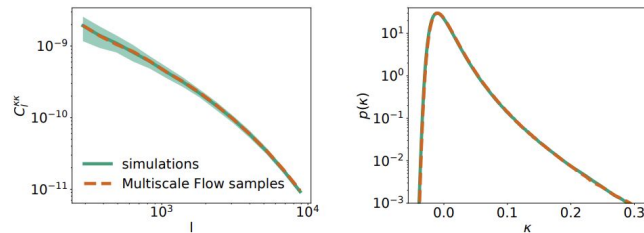


# MSF: A fast and accurate simulator of WL maps



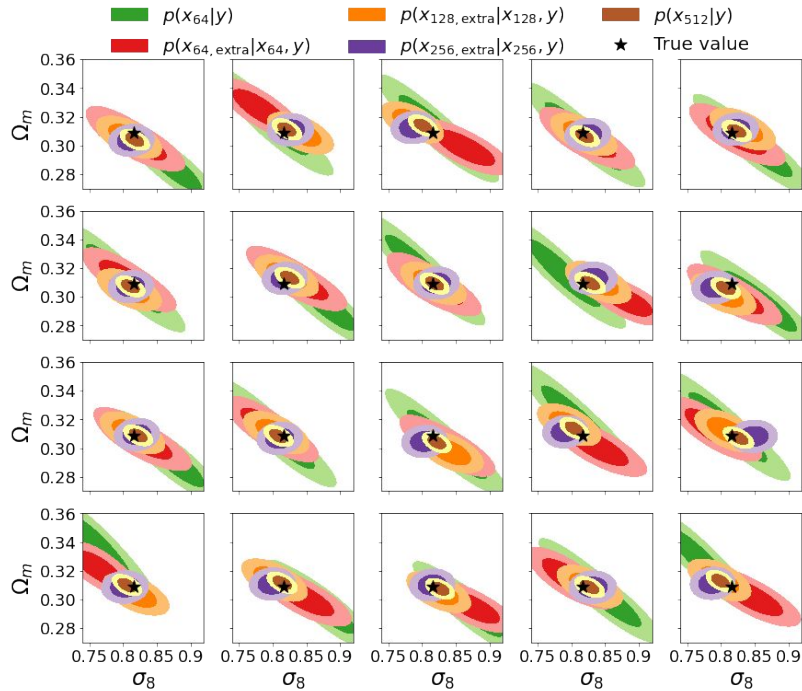
Perfect agreement with real simulations on the power spectrum and one point distribution

Notice how stochastic the data is



# Reliable Uncertainty Quantification with NF posteriors (generalization: accurate prediction on out of sample data)

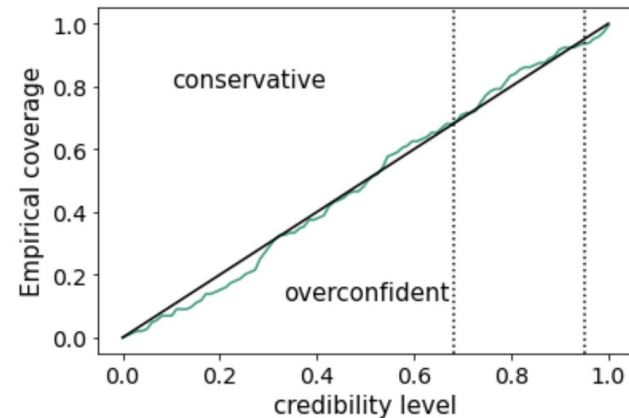
- Consistent posteriors from different scales



Posterior  $p(y|x)=p(x|y)p(y)/p(x)$   
prior  $p(y)$  is assumed to be flat here

On simulated data the errors are properly calibrated and in agreement with frequentist notions of error quantification (68/95% of true simulation values are within 68/95% posterior contours)

Empirical coverage of Multiscale Flow posterior:



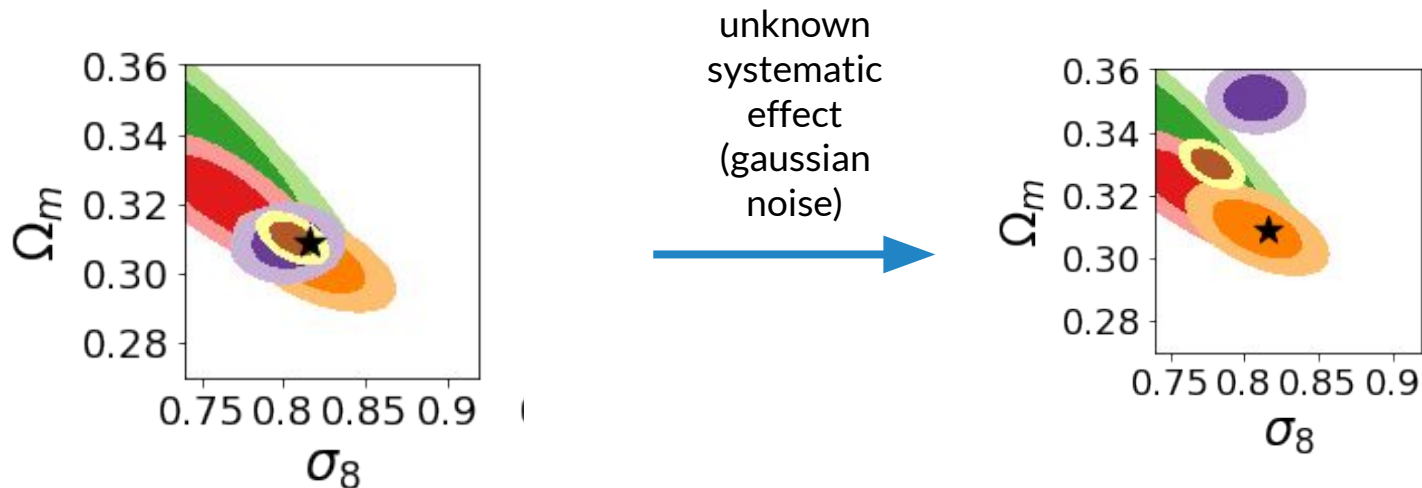
# Trustworthy AI

Robustness – how to identify unknown unknowns (anomalies)?

1) scale dependence of unknown systematic effects on WL maps

- Consistent posteriors from different scales

- Inconsistent small scale posterior



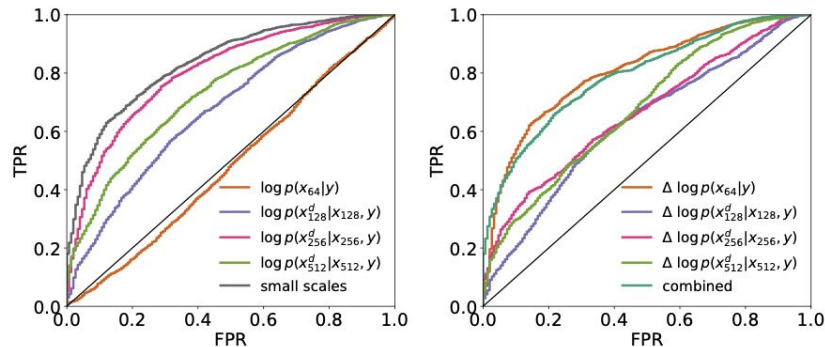
One of the few ways to identify unknown unknowns!

# Trustworthy AI

## 2) anomaly detection with density estimation

Max density estimation  $\max_y p(x|y)$  can be used for unsupervised anomaly detection: if the data are an outlier then the density will be low even when maximized over  $y$

This can be done as a function of scale

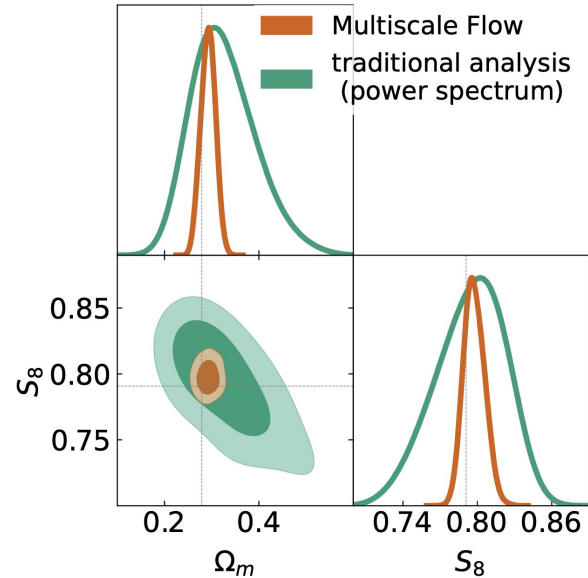


**Fig. 5.** Top panel: scale-dependent posterior analysis of a baryon-corrected convergence map using Multiscale Flow trained on dark-matter-only maps (left), and Multiscale Flow trained on BCM maps (right). Bottom panel: ROC curve of identifying distribution shift with  $\log p$  (left) and  $\Delta \log p$  (right). The "small scales" in the lower left panel represent combining the three small scale terms. In these experiments, we consider  $30 \text{ arcmin}^{-2}$  galaxy shape noise.

**Two independent ways to identify unknown unknowns!**



# How much better is MSF versus standard power spectrum?



Example: simulated HSC or Rubin (LSST) WL data

**MSF up to 10 x better than power spectrum**

Two key cosmological parameters are amplitude of fluctuations  $S_8$  and dark matter density  $\Omega_m$

**Equivalent to 10x larger survey! It is rare to achieve such improvements solely from a better analysis**

**Better than other recent methods (CNN, scattering transform)**

# Future plans

Add complications: intrinsic alignments, baryonic feedback, redshift uncertainty etc.

Finish the analysis of HSC data: could significantly improve cosmological parameters

Develop the analysis pipeline for Rubin and Euclid: much larger area needs much large simulation volumes

# Summary

- **AI for Cosmology:** AI can be used for synthetic data generation (simulations), data analysis, corrupted data restoration, anomaly detection...
- For many physics applications combining generative and discriminative learning is better than discriminative only, and offers more information (trustworthy AI)
- AI for cosmological weak lensing analysis: potential improvements of 10x over power spectrum in weak lensing
- **AI for science:** every field is different, success is not guaranteed, but there are many fields where AI has the potential to significantly improve current baselines