Opportunities and Challenges in "big data" cosmology

Masahiro Takada (Kavli IPMU)

Also ask Leander Thiele …

KAVLI IPMU INSTITUTE FOR THE PHYSICS AND MATHEMATICS OF THE UNIVERSE

東京大学 THE UNIVERSITY OF TOKYO

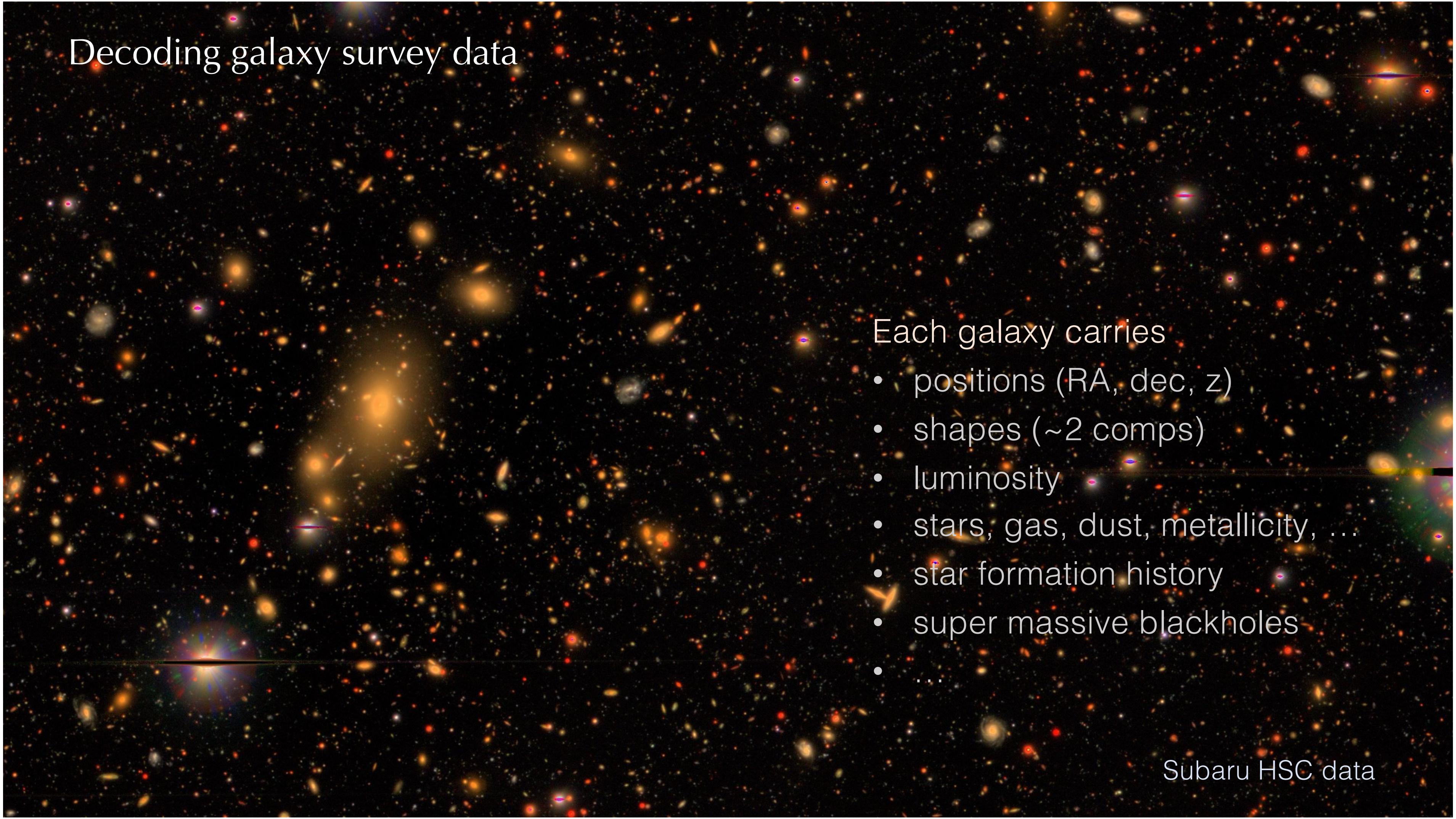
wpi CD3

@FAIRS Nagoya U., Dec 2024

Decoding galaxy survey data

Each galaxy carries

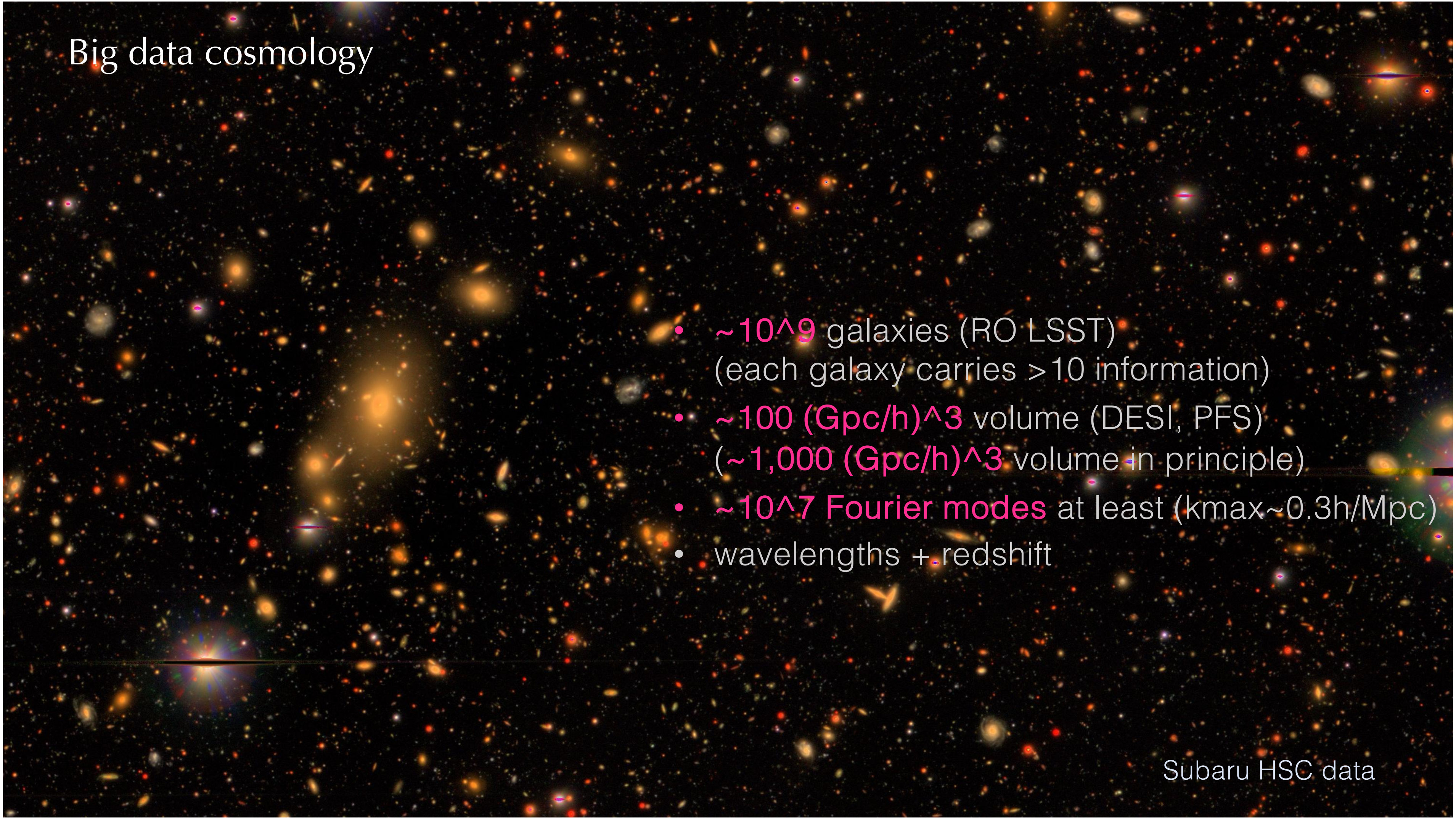
- positions (RA, dec, z)
- shapes (~2 comps)
- luminosity
- stars, gas, dust, metallicity, …
- star formation history
- super massive blackholes
- …



Subaru HSC data

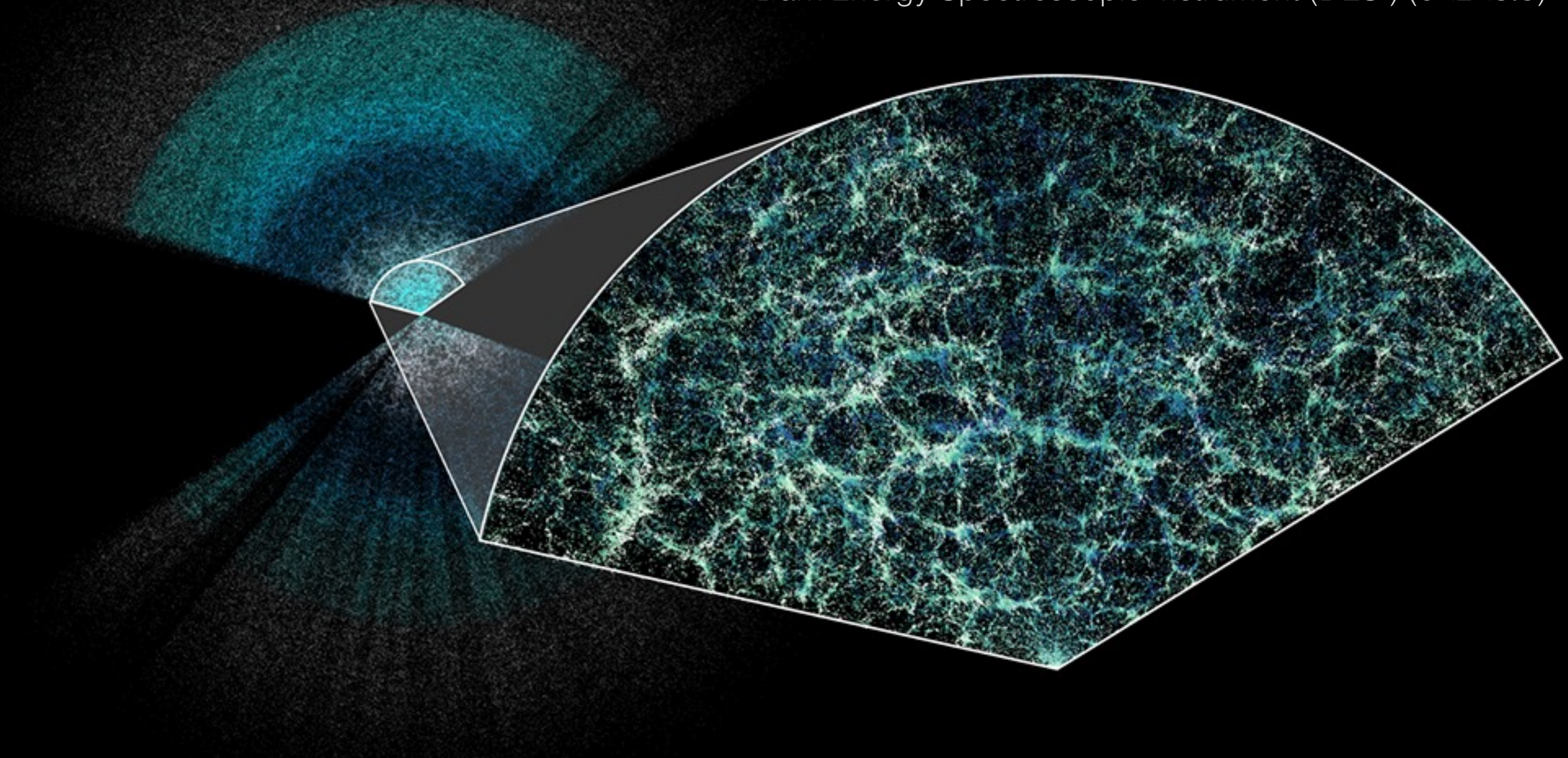
Big data cosmology

- ~10^9 galaxies (RO LSST)
  (each galaxy carries >10 information)
- ~100 (Gpc/h)^3 volume (DESI, PFS)
  (~1,000 (Gpc/h)^3 volume in principle)
- ~10^7 Fourier modes at least (kmax~0.3h/Mpc)
- wavelengths + redshift



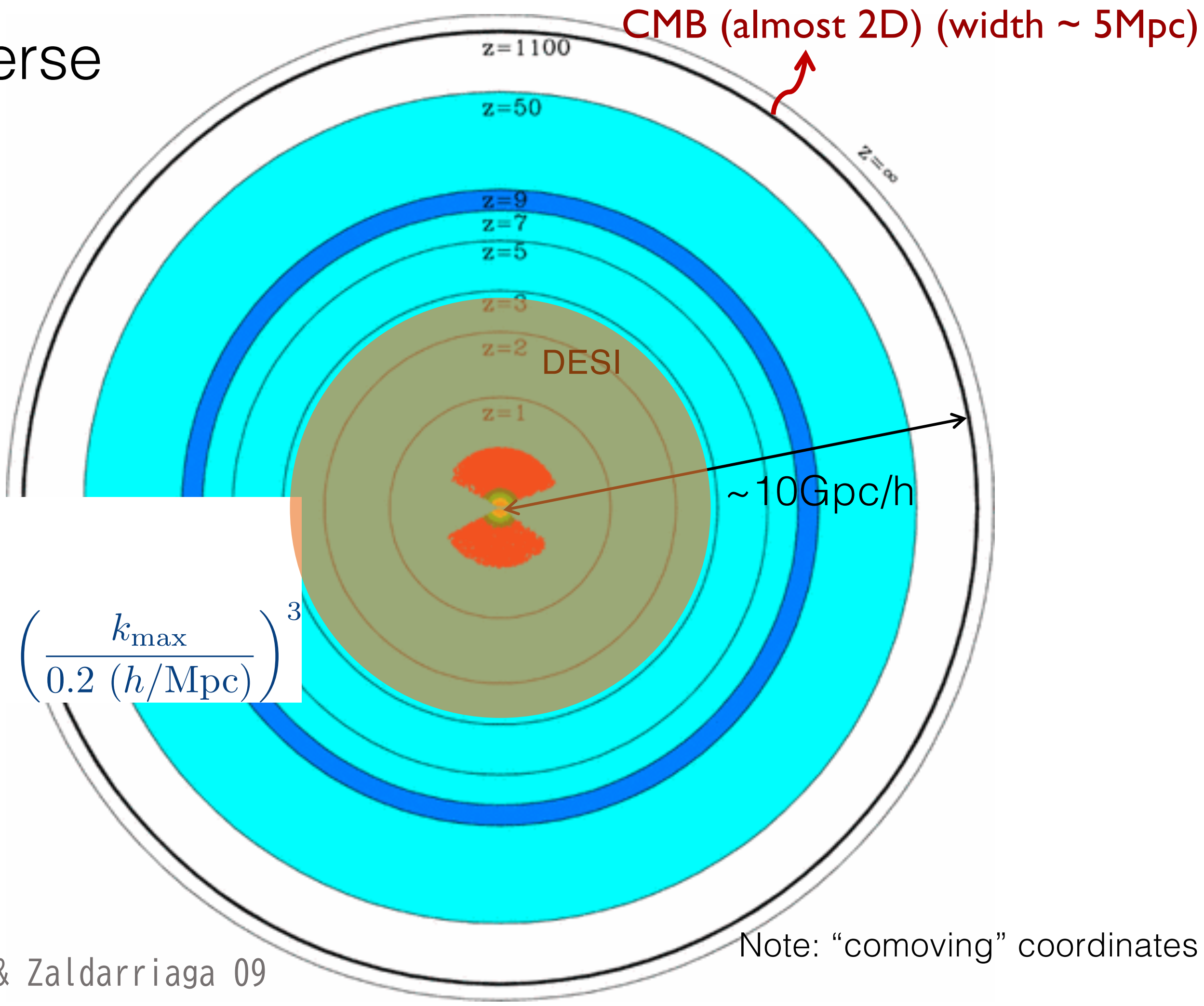
Subaru HSC data

A 3D map of the universe with
Dark Energy Spectroscopic Instrument (DESI) ($0<z<3.5$)



Credit: Claire Lamman/DESI

“Observable” universe



The number of Fourier modes

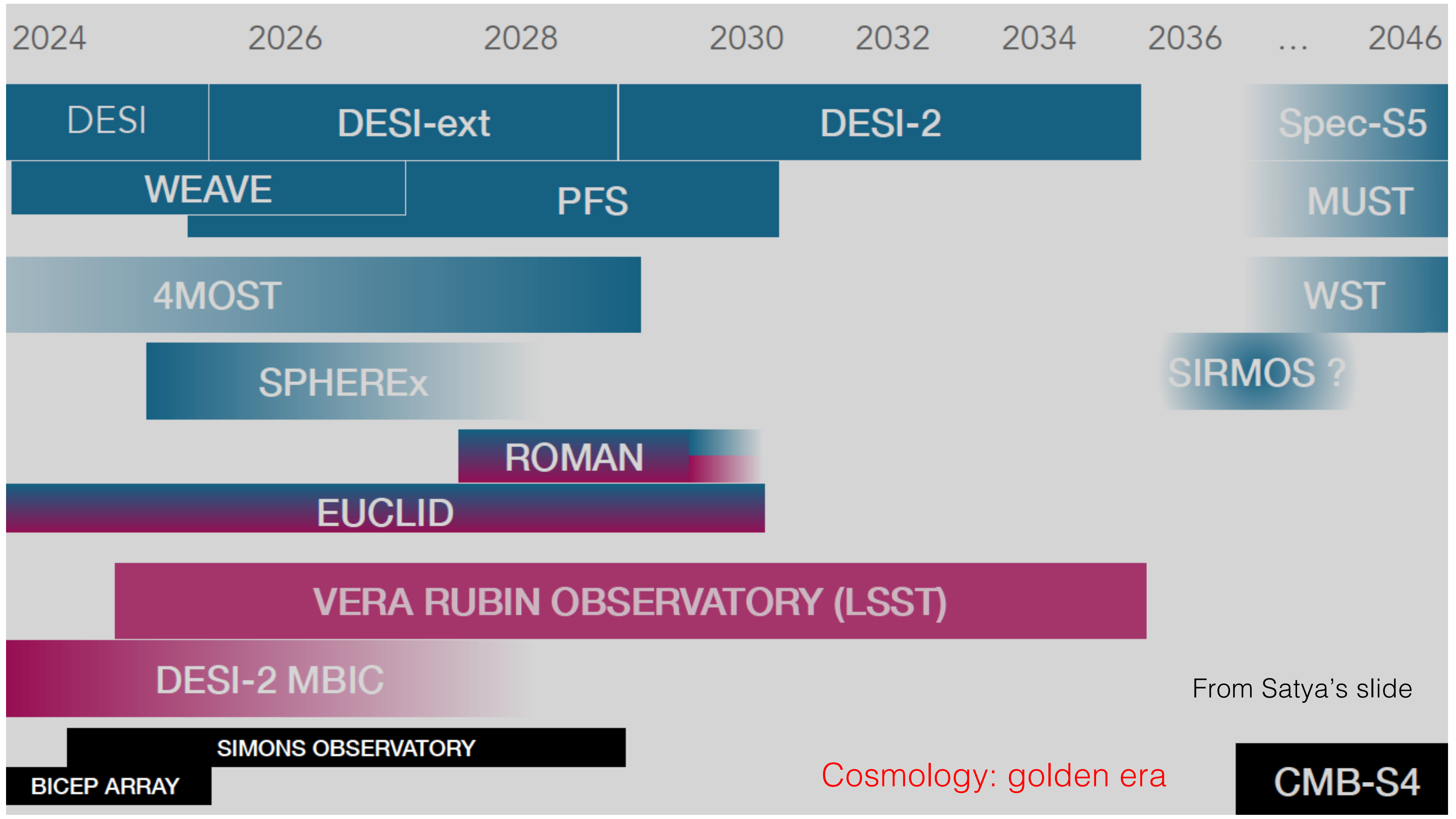
$$N_{\rm mode} = \frac{V_F}{k_F^3} \sim V_{\rm survey} k_{\rm max}^3$$

$$\sim 10^7 \left(\frac{V_{\rm s}}{100\ ({\rm Gpc}/h)^3}\right)\left(\frac{k_{\rm max}}{0.2\ (h/{\rm Mpc})}\right)^3$$

(CMB: ~10^6)

Tegmark & Zaldarriaga 09

Note: “comoving” coordinates

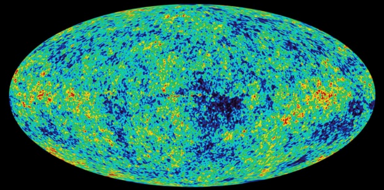

2024
2026
2028
2030
2032
2034
2036
...
2046
DESI
DESI-ext
DESI-2
Spec-S5
WEAVE
PFS
MUST
4MOST
WST
SPHEREx
SIRMOS ?
ROMAN
EUCLID
VERA RUBIN OBSERVATORY (LSST)
DESI-2 MBIC
From Satya's slide
SIMONS OBSERVATORY
BICEP ARRAY
Cosmology: golden era
CMB-S4

We need to be ready …

Rubin Observatory LSST (Chile): 2025-

Cosmic structure formation: time-evolution of primordial perturbations

The primordial information of the fluctuations in the linear regime is preserved



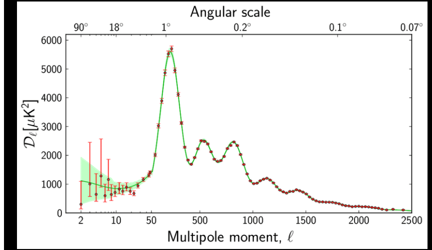
CMB at z~1000

standard model of cosmology: ΛCDM

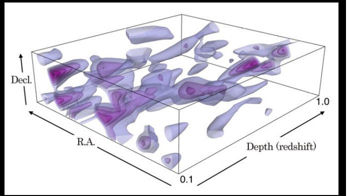
time

gravity (dark matter)

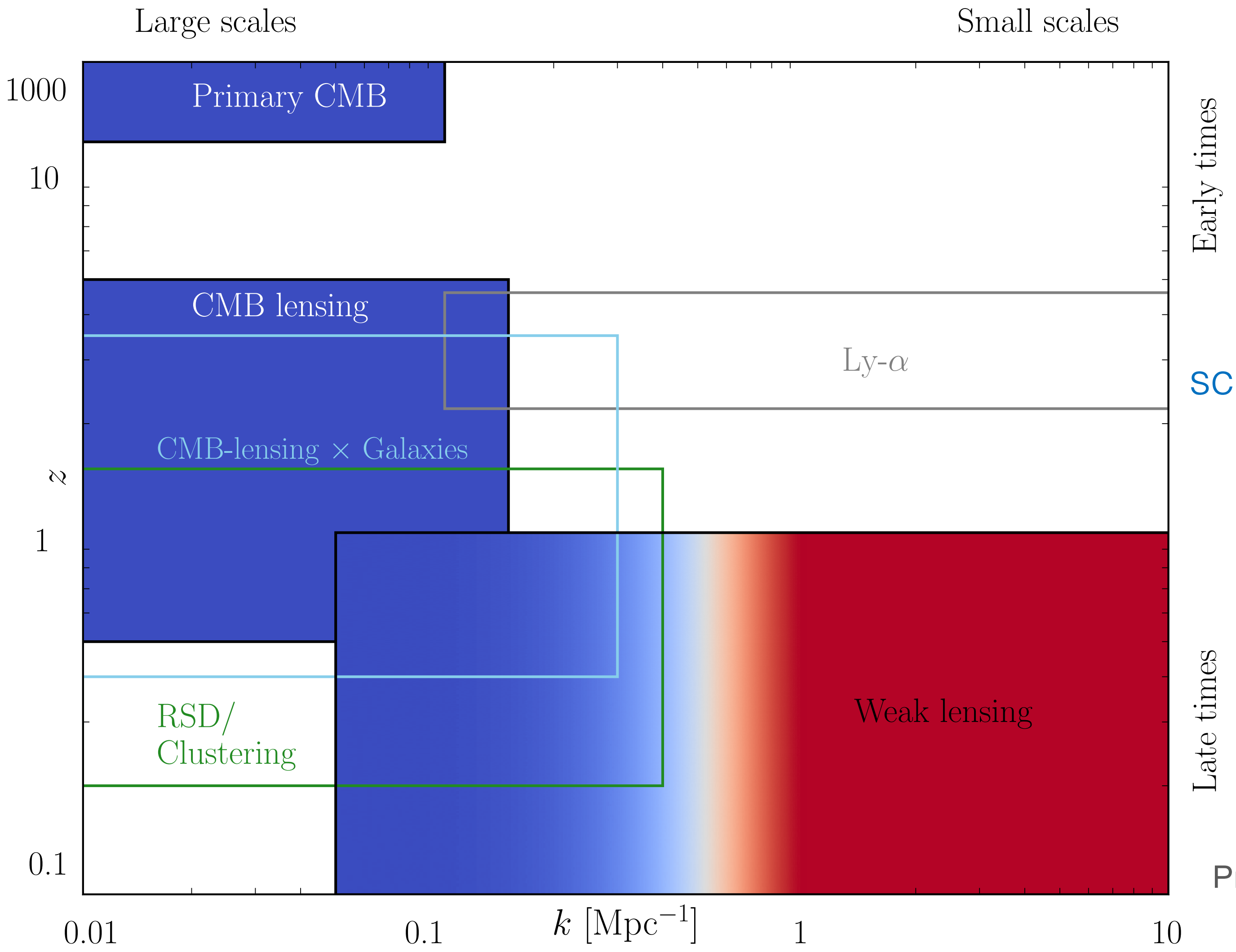
cosmic expansion (dark energy)



ΛCDM=~6 parameters
+Gaussian fluctuations



Galaxy surveys measure "**clumpiness**" of the late universe

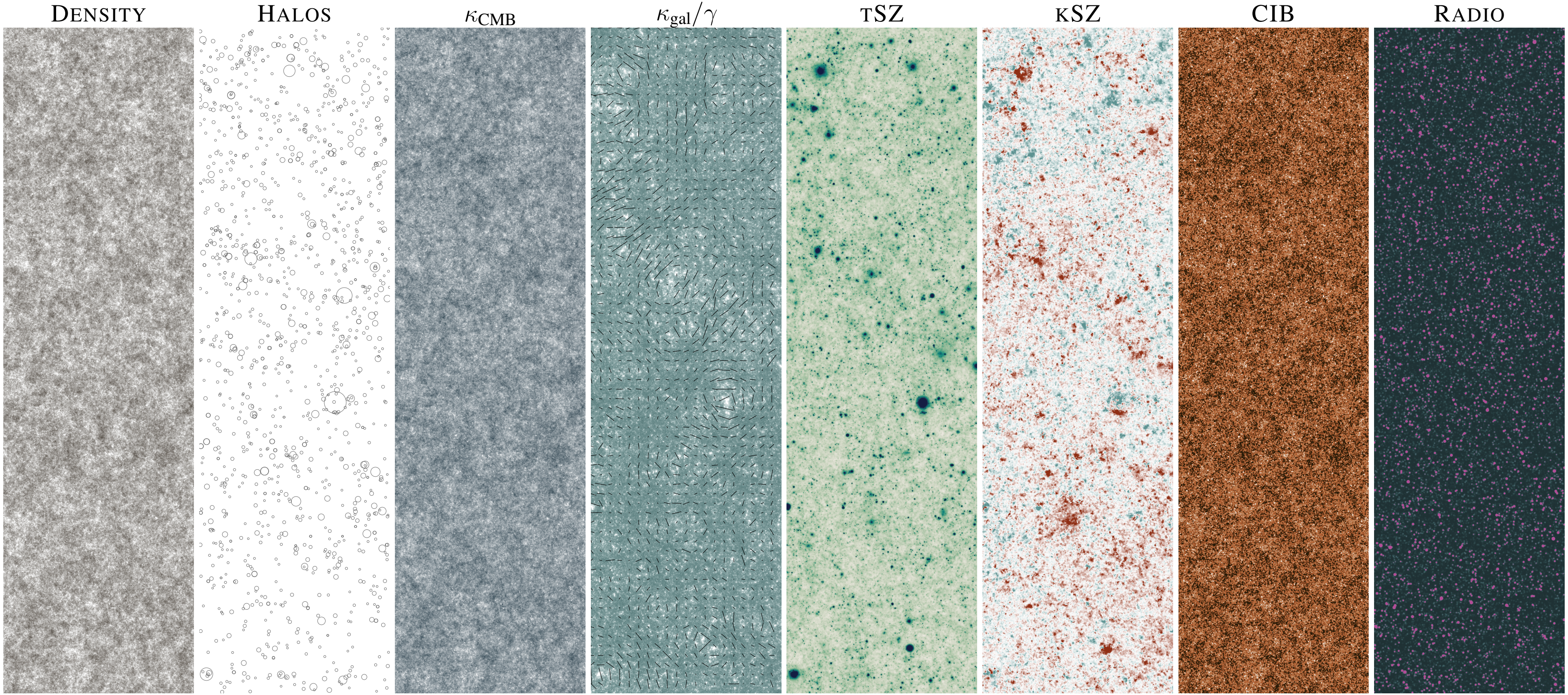


$$\delta(\mathbf{k}, z)$$

scale and time (redshift)

Preston, Amon & Efstathiou 23

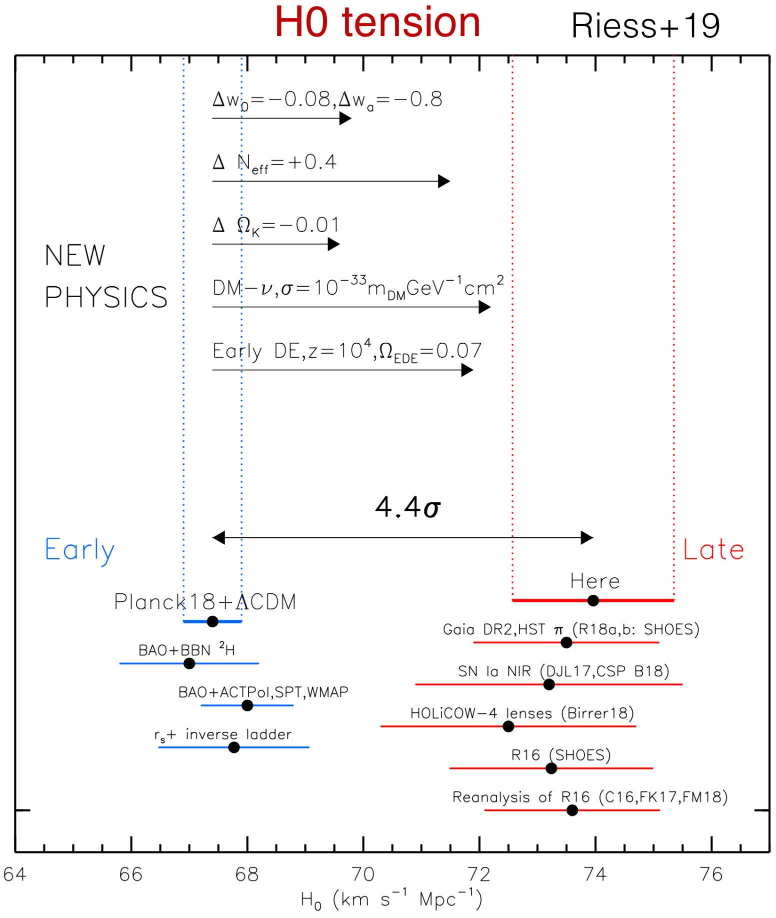
Multi-components maps (many observables in the same large-scale structure)



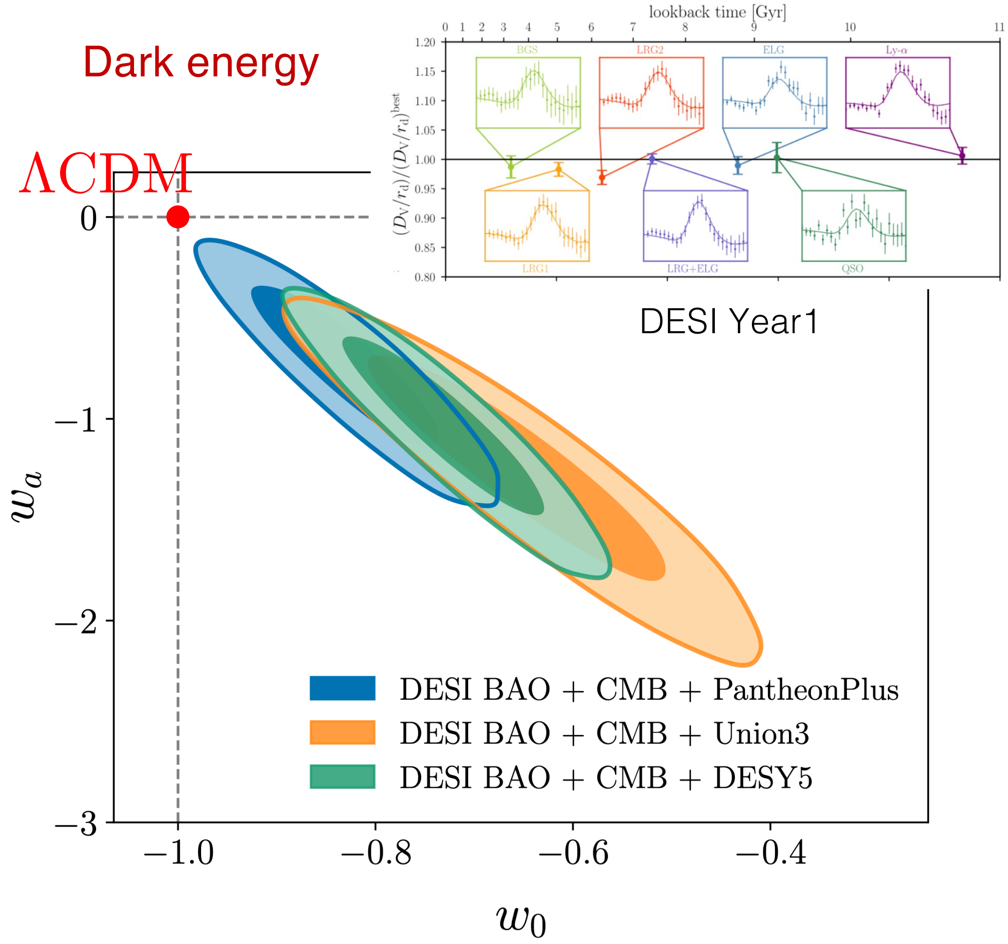
Omori 2022

A hint of new physics beyond ΛCDM – discovery potential

H0 tension



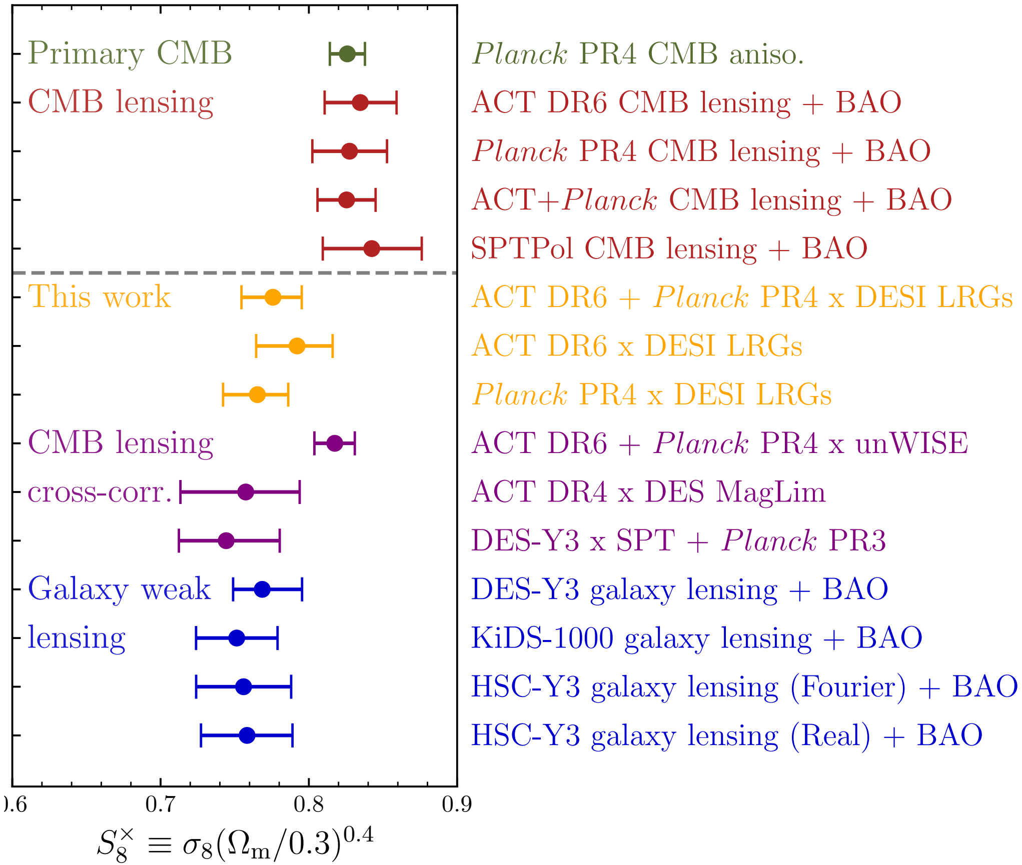
Dark energy



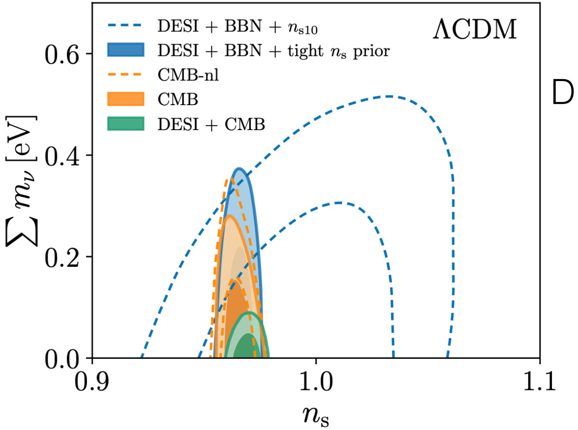
A hint of new physics beyond ΛCDM – discovery potential

S8 tension

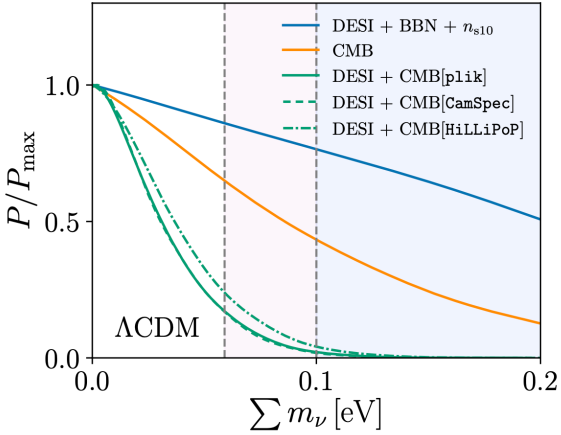
Kim+24 (DESI+ACT)



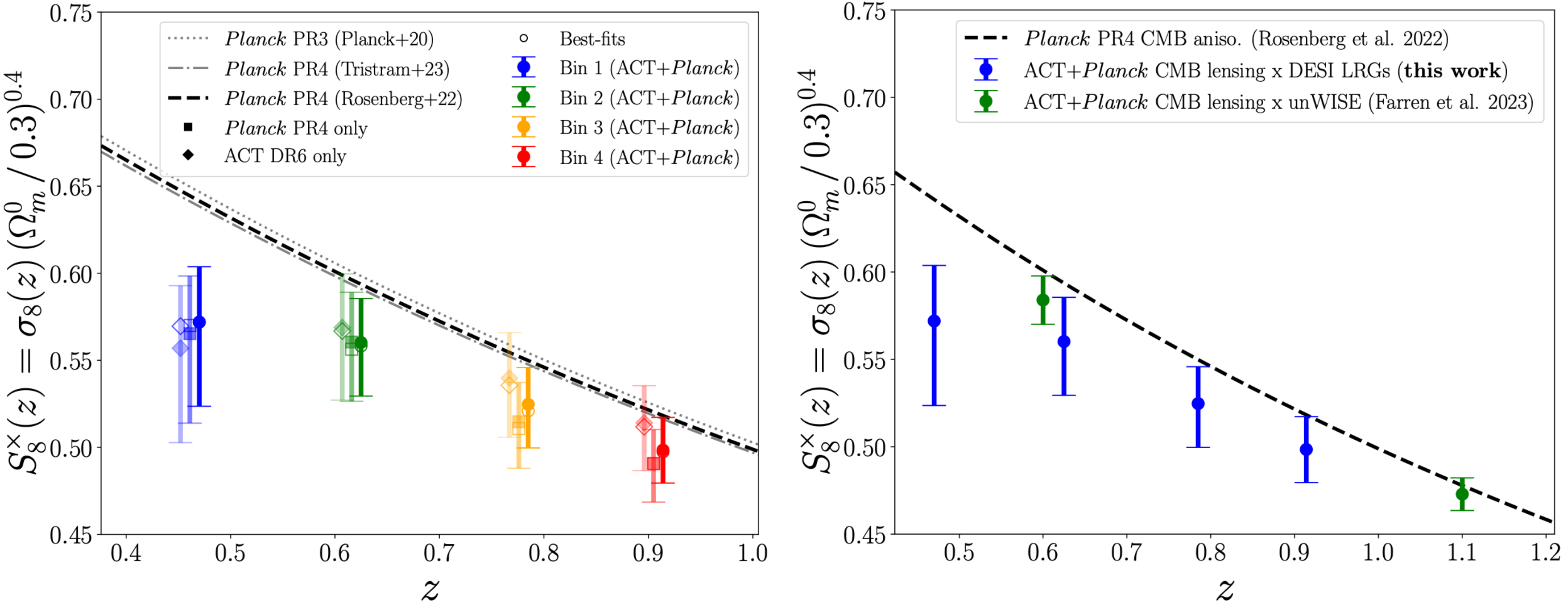
neutrino mass



DESI Year1



Kim+24 (DESI+ACT)



Discovery potential and Opportunities for AI/ML in fundamental cosmology

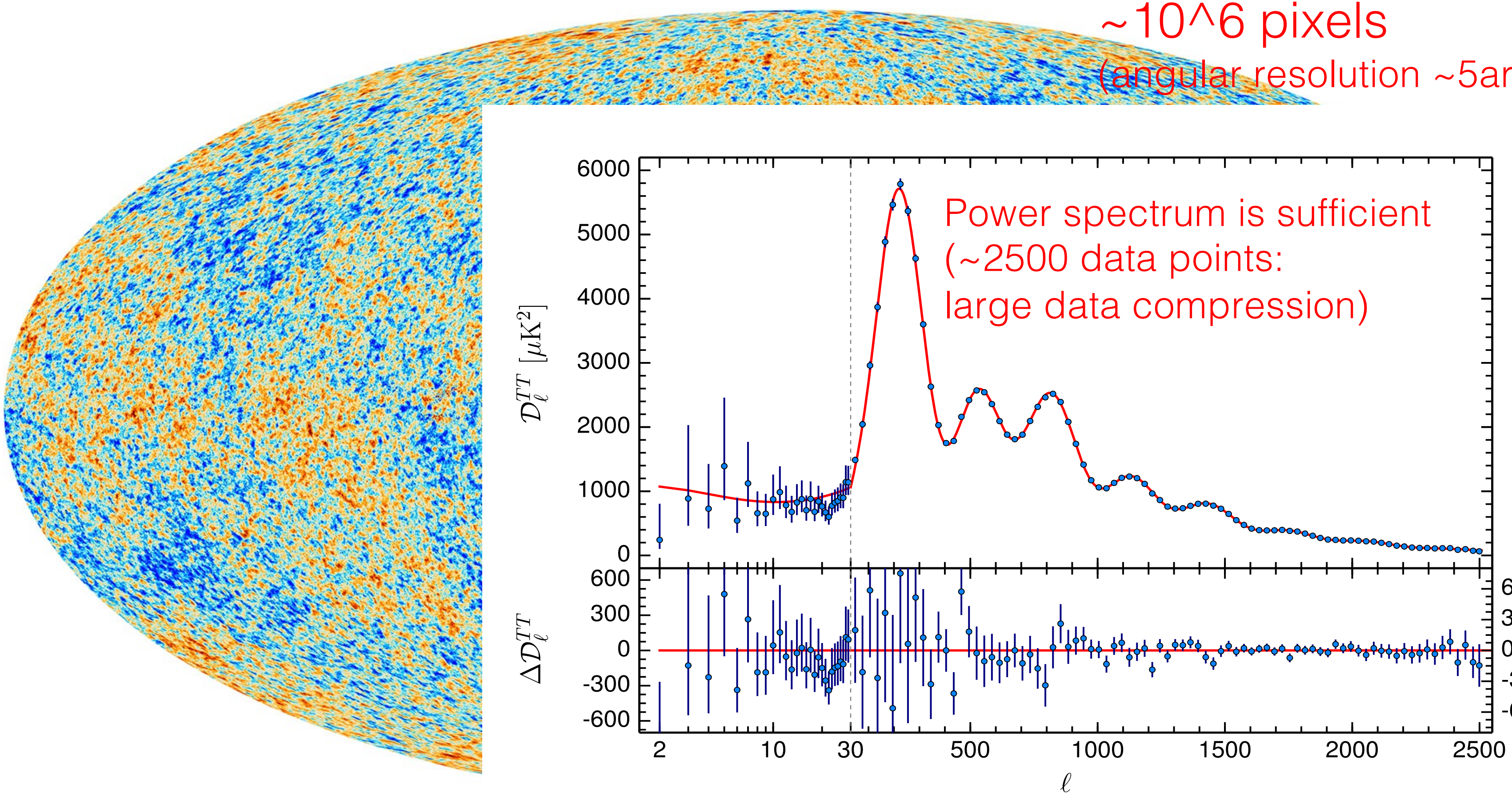
- Cosmic expansion (dark energy, H0 tension, curvature)
- Growth of cosmic structures (S8 tension, modified gravity)
- Primordial non-Gaussianity (inflation physics) (+parity violation)
- Neutrino mass (note: synergy with particle physics)
- Nature of dark matter

Discovery potential and Opportunities for AI/ML

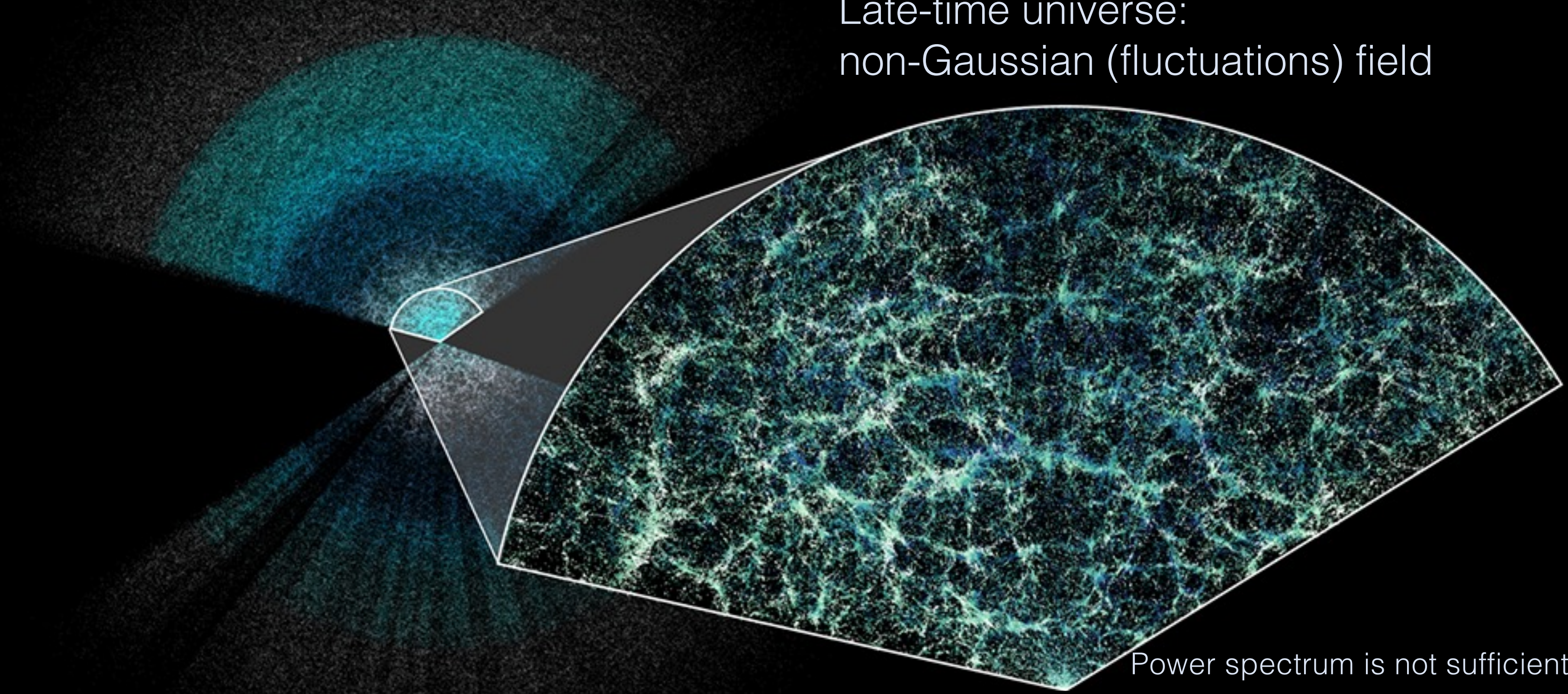
- Cosmic expansion (dark energy, H0 tension, curvature)
  - Galaxy surveys: BAO, supernovae, GW sirens, [weak lensing (WL)]
- Growth of cosmic structures (S8 tension, modified gravity)
  - Galaxy surveys: WL, galaxy clustering, redshift space distortion (RSD), galaxy clusters
- Primordial non-Gaussianity (inflation) (+parity violation)
  - Galaxy surveys: power spectrum, bispectrum, intrinsic alignments, WL
- Neutrino mass (note: synergy with particle physics)
  - Galaxy surveys: galaxy clustering, WL, galaxy clusters
- Nature of dark matter
  - small-scale probes – Lyman-alpha, dwarf galaxies, microlensing, …

CMB = (nearly) Gaussian (fluctuations) field

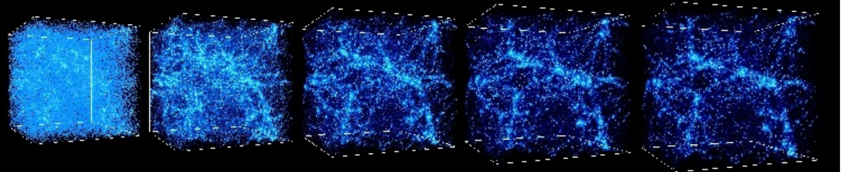
~10^6 pixels
(angular resolution ~5arcmin)



Late-time universe:
non-Gaussian (fluctuations) field



Power spectrum is not sufficient



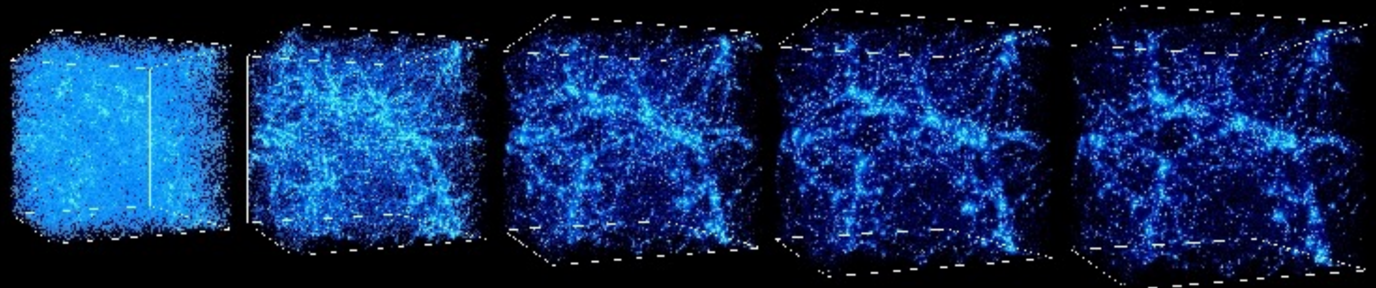
Credit: Claire Lamman/DESI

Challenges in fundamental cosmology – opportunities for ML

- How can we extract the maximum amount of information from galaxy survey data?
  - Can we recover the information of the initial Gaussian field? (angular 2D + redshift in the light-cone volume)
  - Up to which k_max can we use for cosmology?
  - Systematic effects: baryonic effects, observational systematics (Galactic dust) …
  - Properties of galaxies (which are impossible to accurately model from first principles) or very small-scale data can be used for cosmology
- How can we combine multi-wavelength data and/or –probes to obtain improved cosmological constraints, mitigating the systematic effects?
  - Weak lensing (WL), galaxies, X-ray, tSZ, kSZ, intensity mapping (e.g., HI), GW, …

Cosmology with galaxy surveys, compared to particle experiments

- Pros
  - Can use accurate theory, on large or quasi nonlinear scales: cosmological linear and perturbation theory (can recover the primordial information)
  - Can simulate the universe in computers; Can make mocks of a galaxy survey
  - Various datasets: multi probes & multi-wavelength data
  - **All data (even raw data) are public;** anyone can get great science with a better method
- Cons
  - Can't replicate the same survey (sample variance), unlike experiments in the lab
  - Simulations, especially hydrodynamical simulations, are still expensive (~trillion particle N-body sim takes ~10M CPU hours, for a single cosmological model)
  - (Unknown) systematic effects
  - No Feynman diagram …

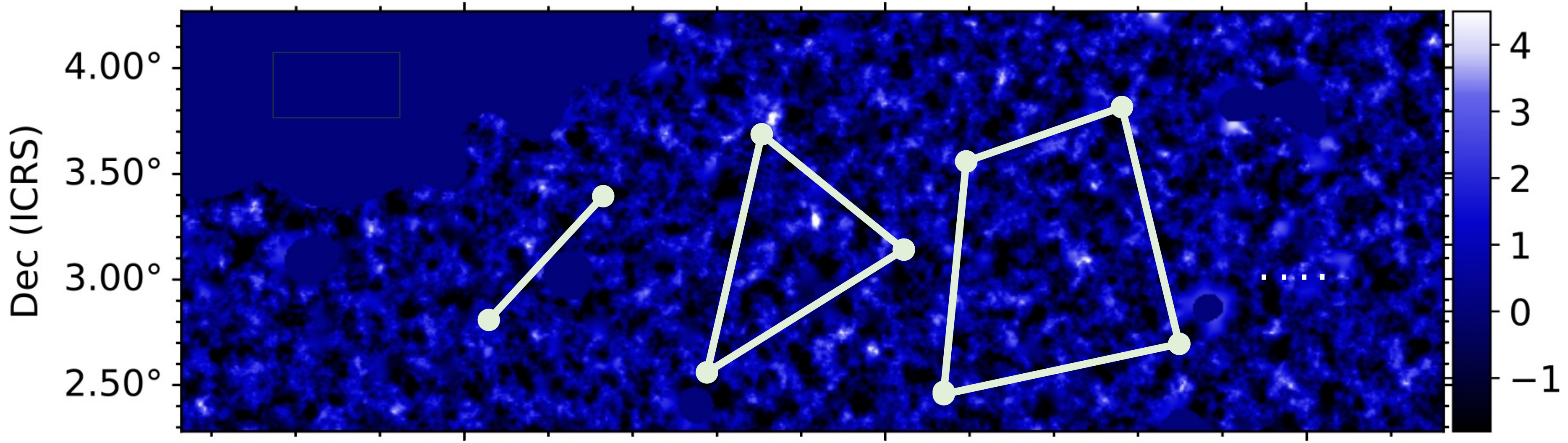


Summary statistics: conventional approach

- How can we extract the maximum amount of cosmological information?

Huge data compression

Discovery potential: DE, PNG, neutrino mass

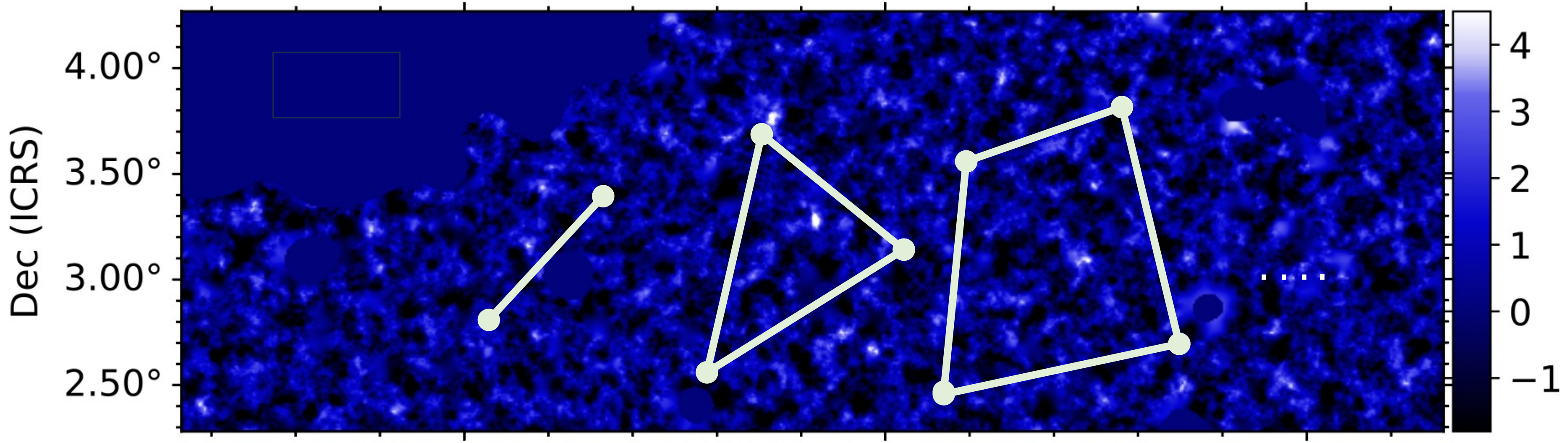


Oguri & Miyazaki 24 (HSC weak lensing map)

2-point correlation function	$\langle f(\mathbf{x}_1) f(\mathbf{x}_2) \rangle$	Power spectrum (Fourier space)
3pt	$\langle f(\mathbf{x}_1) f(\mathbf{x}_2) f(\mathbf{x}_3) \rangle$	Bispectrum
4pt	$\langle f(\mathbf{x}_1) f(\mathbf{x}_2) f(\mathbf{x}_3) f(\mathbf{x}_4) \rangle$	Trispectrum

Summary statistics: conventional approach

- How can we extract the maximum amount of cosmological information?
- Q: Up to what n-point correlations should we measure?



Oguri & Miyazaki 24 (HSC weak lensing map)

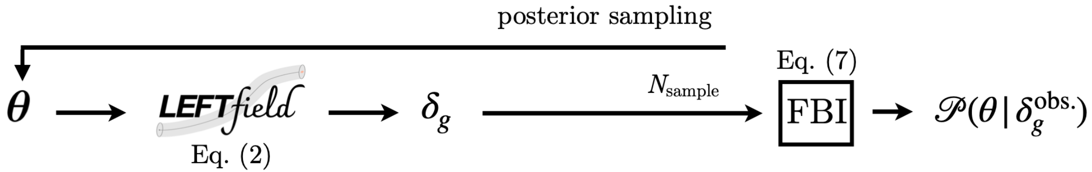
- Challenges
  - The measurements of >3pt functions are not easy (WL bispectrum hasn't yet measured
  - Constructing accurate modes of >3pt functions are expensive (but not impossible)
  - Need to include observational effects (survey window, masks, …)

Field level inference

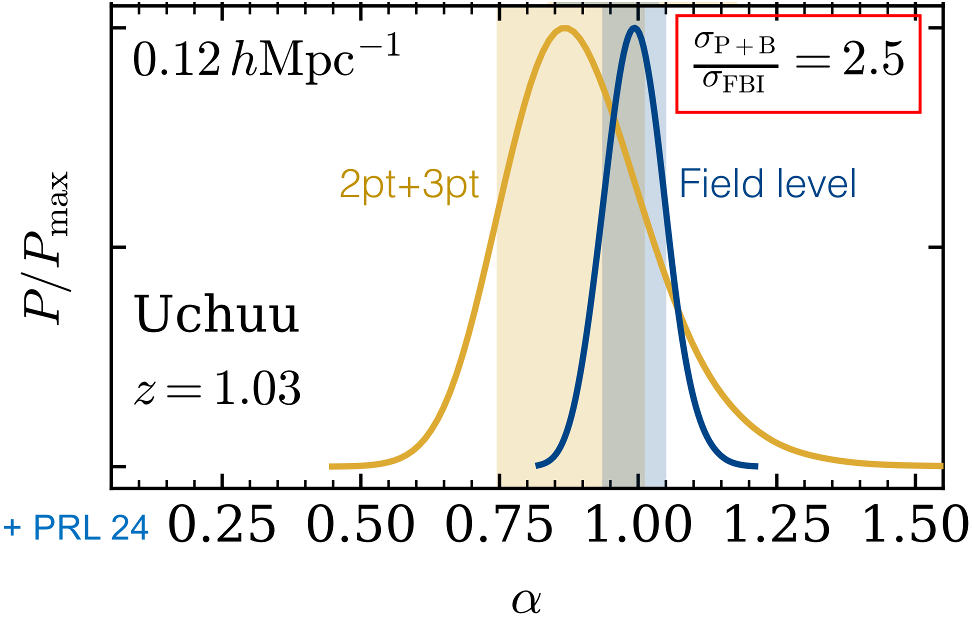
DE, PNG, neutrino mass

- How can we extract the maximum amount of cosmological information?

$$\delta_g^{\mathrm{model}}(\mathbf{k}|\boldsymbol{\theta}) \leftrightarrow \delta_g^{\mathrm{data}}(\mathbf{k})$$



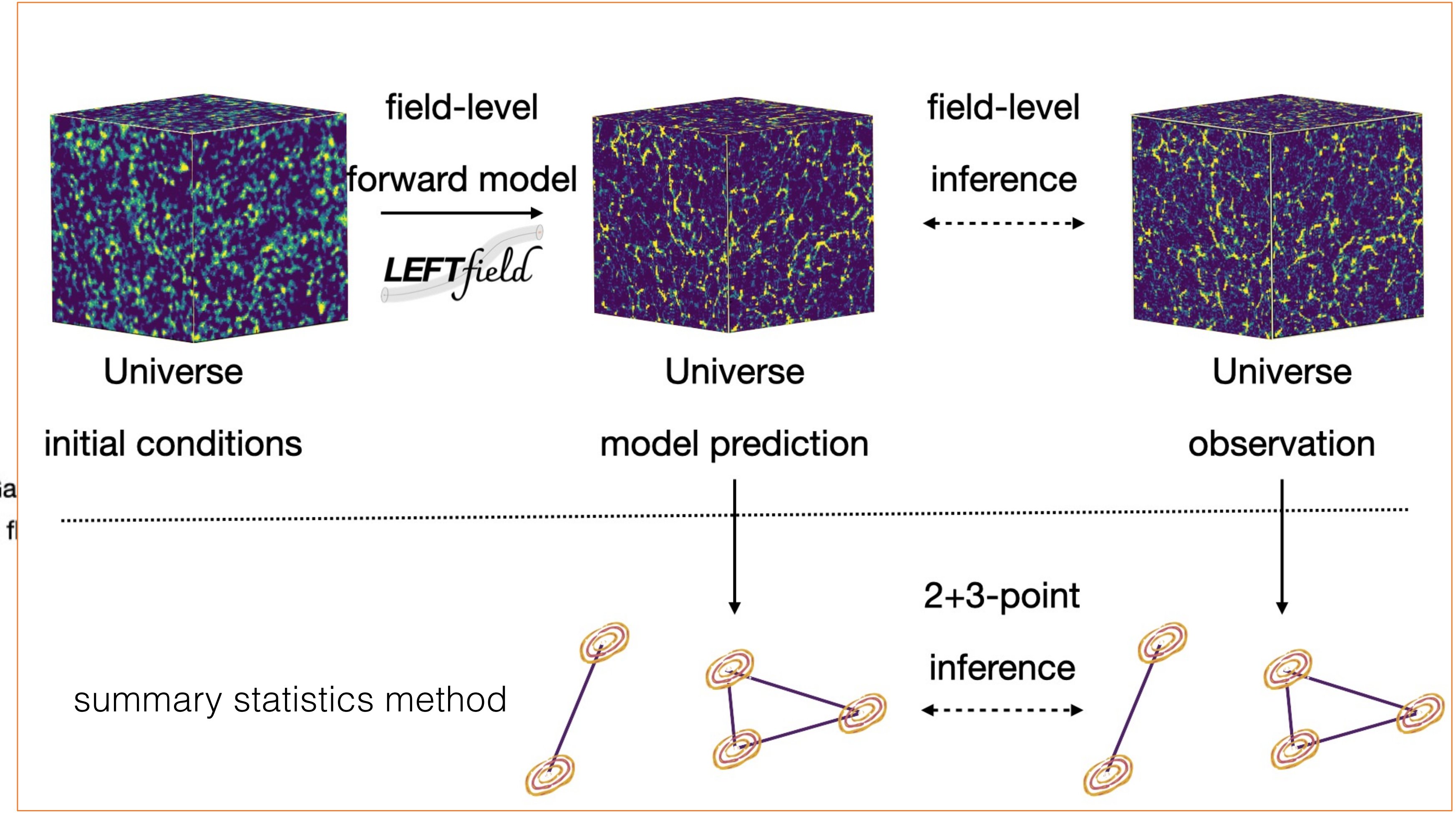
- **Hope**: we can extract the full information, instead of using an infinite series of n-point correlation functions
- Can extend FLI to multi-probe and/or wavelength data
- Note claim in Cabass+24 & Akitsu+ (in prep.)



Minh Nguyen+ PRL 24

Demonstration using simulations (not data)

Credit: Minh Nguyen



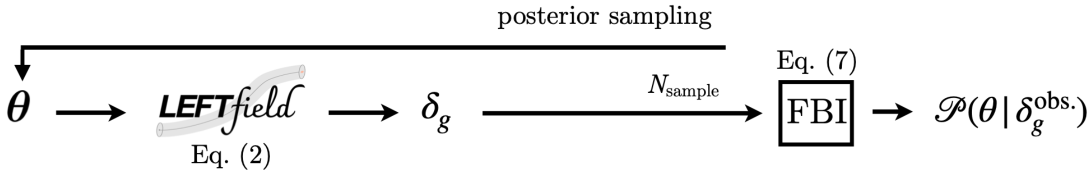
Ga
initial fi

Field level inference

DE, PNG, neutrino mass

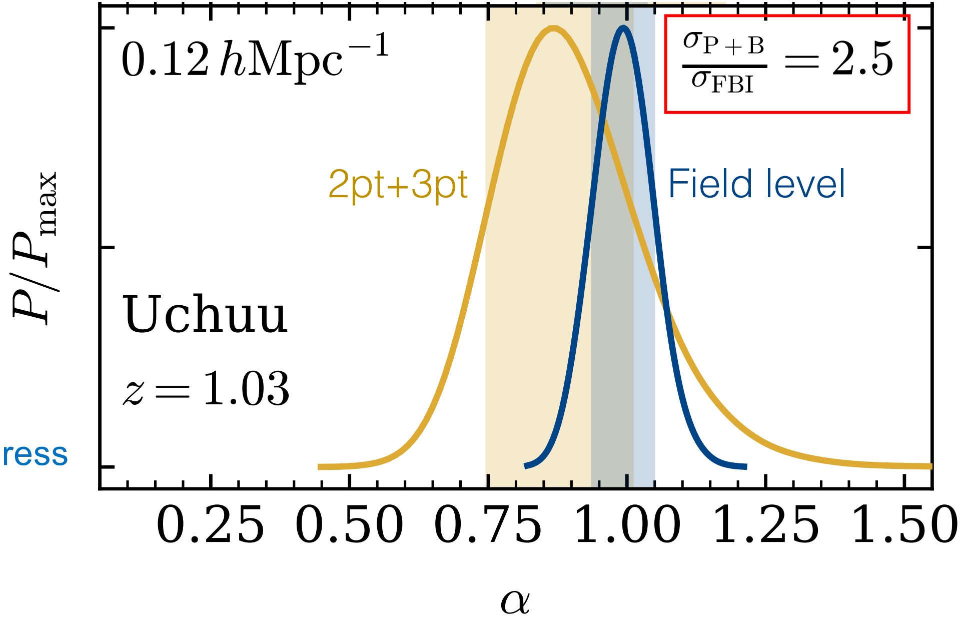
- How can we extract the maximum amount of cosmological information?

$$\delta_g^{\text{model}}(\mathbf{k}|\boldsymbol{\theta}) \leftrightarrow \delta_g^{\text{data}}(\mathbf{k})$$



- Challenges
  - Not yet applied to data
  - Need to include observation effects (selection function, systematic errors, …)
  - Interpretability/reproducibility?

Minh Nguyen+ PRL in press



Demonstration using simulations (not data)

Simulation based inference (SBI)

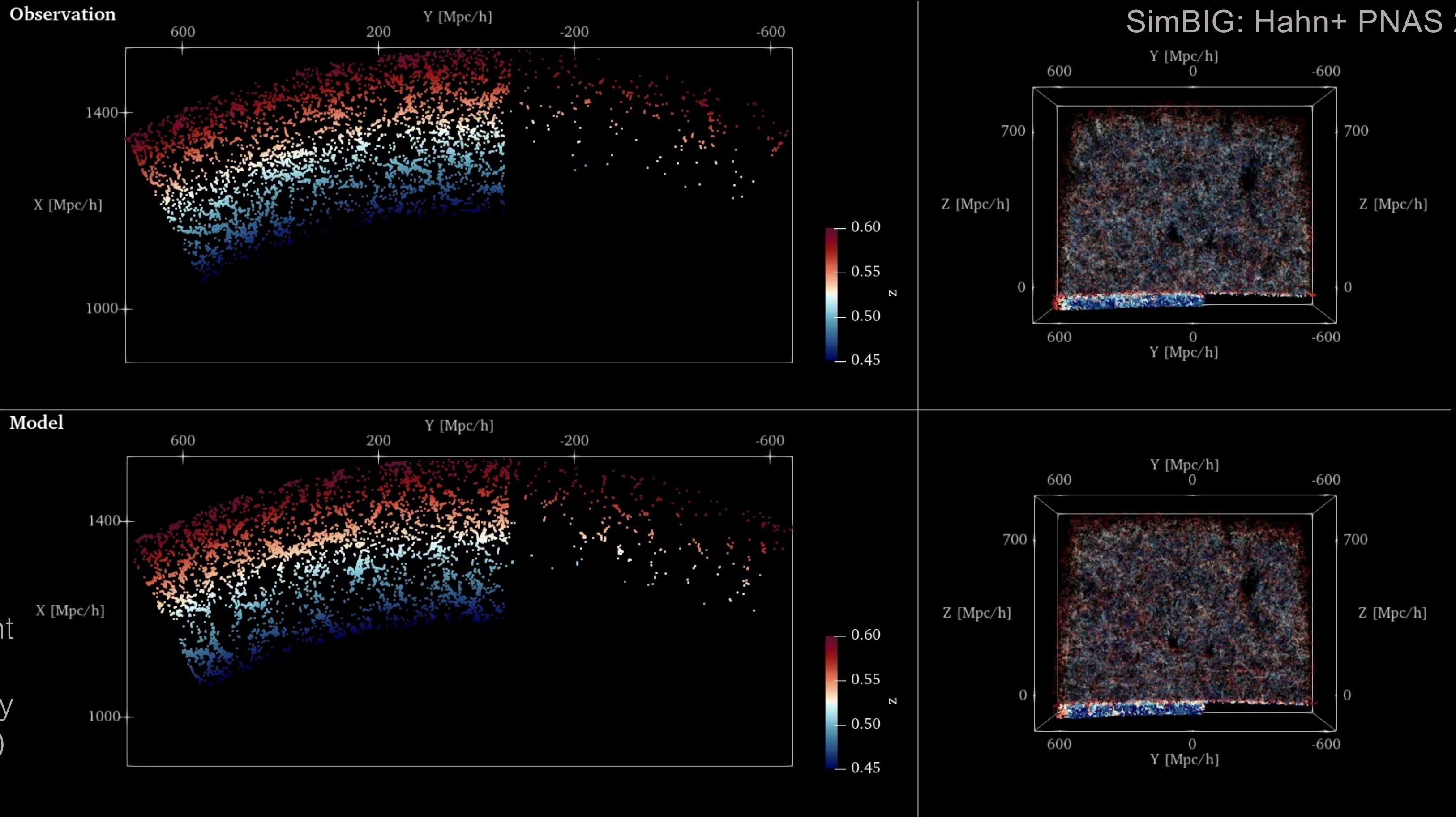
- How can we extract the maximum amount of cosmological information?

SimBIG: Hahn+ PNAS 23

Data

Model

Obs. effects included (light cone effect, masks, survey geometry, ...)



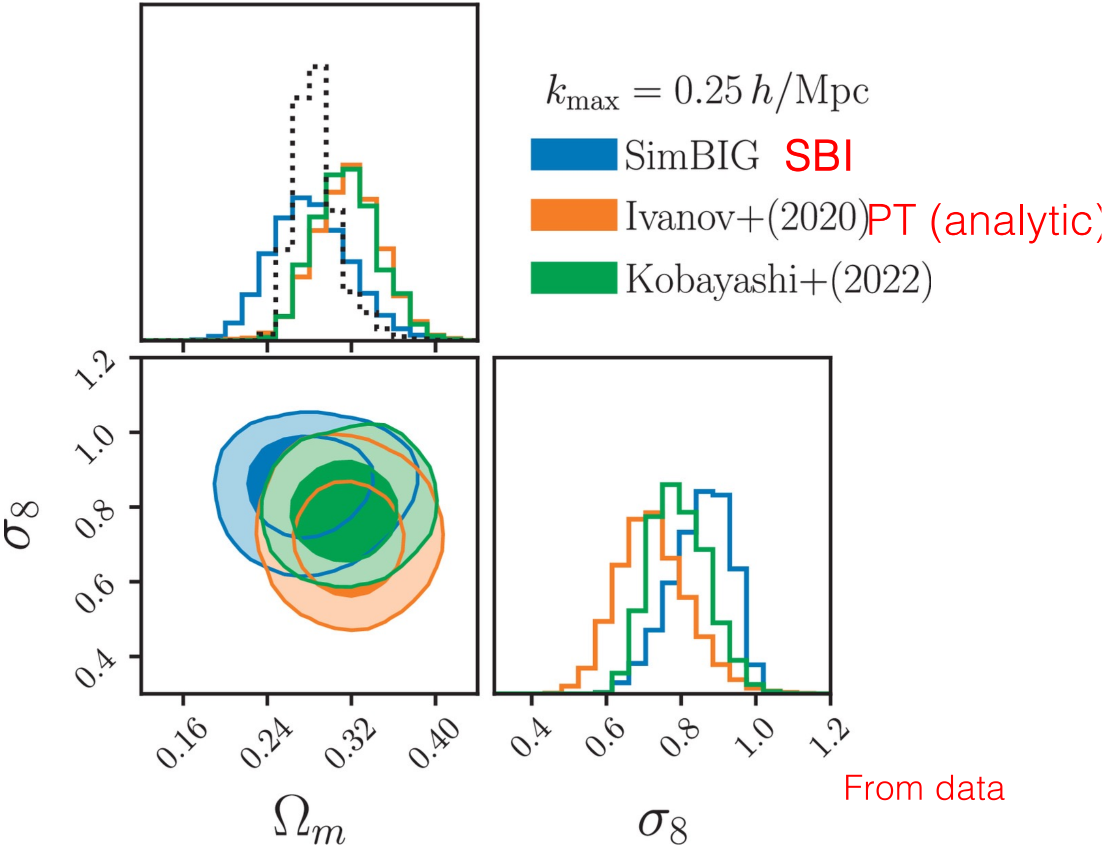
Simulation based inference

DE, PNG, neutrino mass

- How can we extract the maximum amount of cosmological information?

Hahn+21

- Hope: we can use the data on small (higher kmax) scales to obtain cosmological constraints
- Can extend the method to multi-probe and/or wavelength data
- The SBI method is now feasible for some datasets, e.g., the SDSS covering 1 (Gpc/h)^3



From data

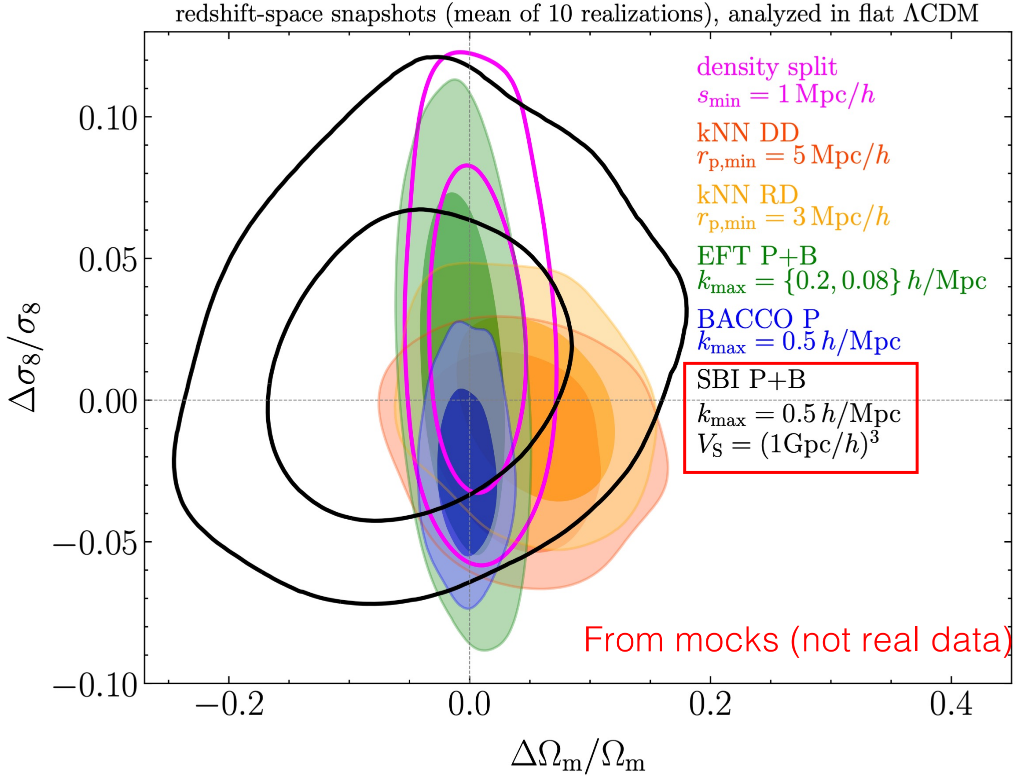
Simulation based inference

DE, PNG, neutrino mass

- How can we extract the maximum amount of cosmological information?

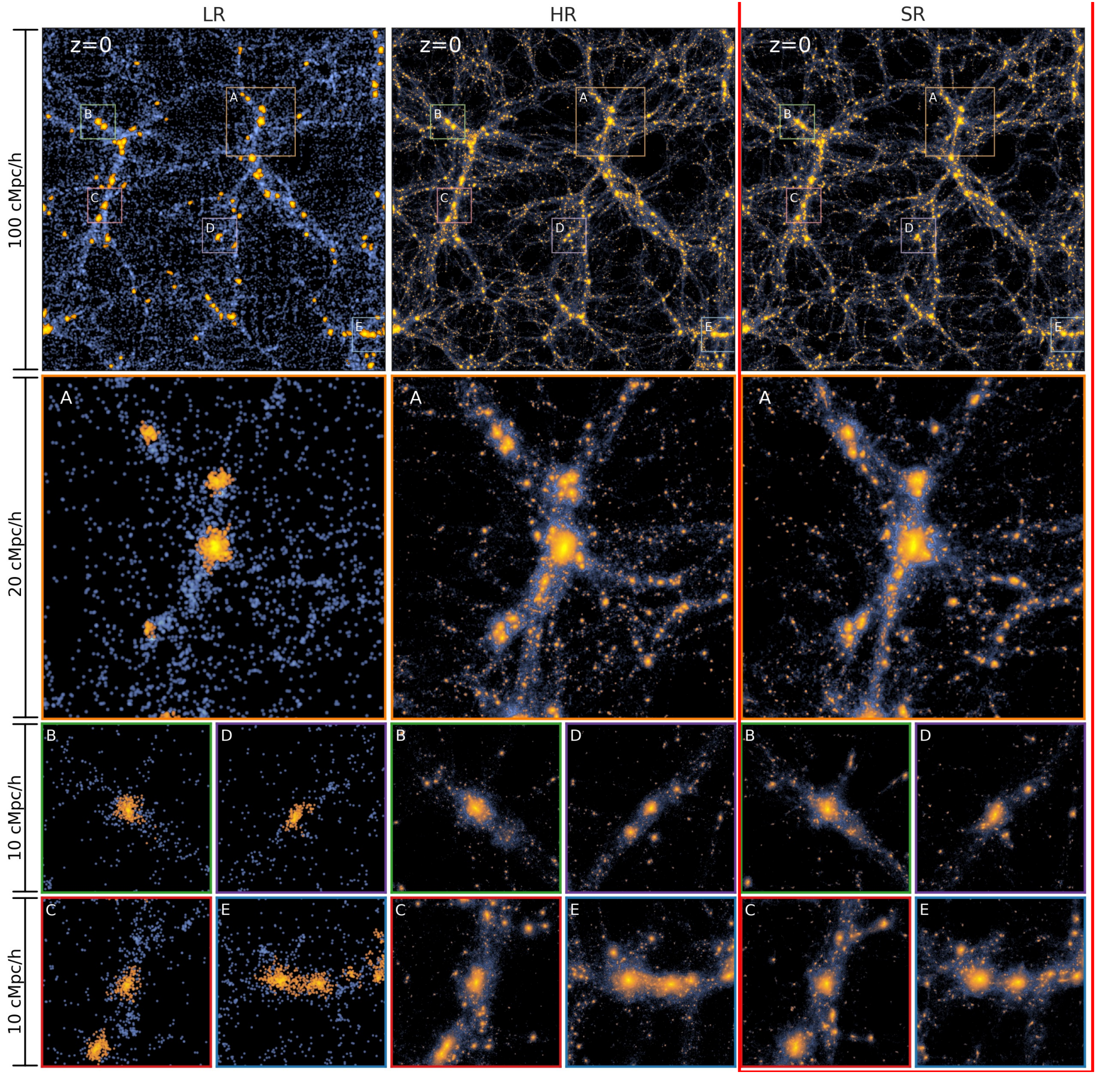
- Challenges
  - SBI needs to simulate galaxy survey, **with the same volume**, and to use **many realizations** to model the sample variance effect
  - Still computationally very expensive: **impossible to simulate a DESI-like survey of >100 (Gpc/h)^3** with the required spatial resolution
  - Interpretability and reproducibility

Krause, Kobayashi+24



An example of using ML for cosmology

AI-assisted super-resolution cosmological simulations

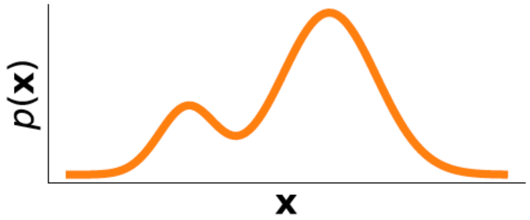


Li, Ni+ PNAS 21

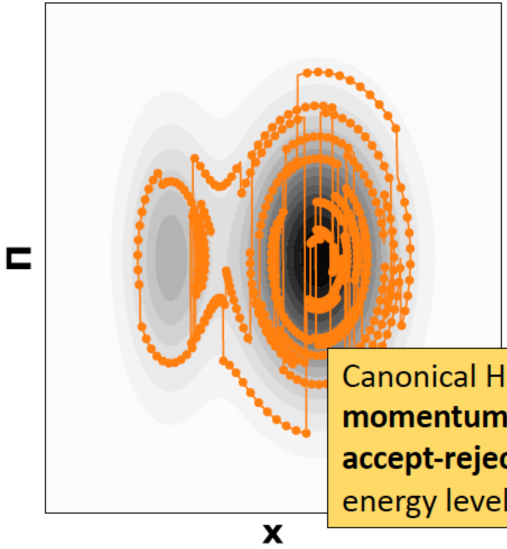
We also need a faster sampler for cosmology inference (~100 parameters for LSST)

For example, MCLMC

Canonical HMC

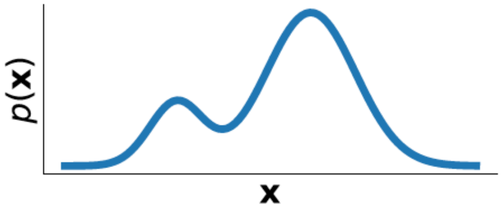


$$p(\mathbf{x}, \mathbf{\Pi}) \propto e^{-H(\mathbf{x}, \mathbf{\Pi})}$$



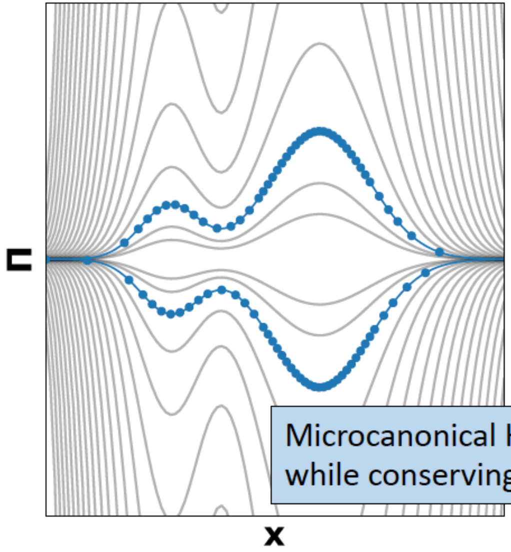
Canonical HMC requires **momentum resampling** and **accept-reject step** to change energy levels and converge

Microcanonical HMC



$$p(\mathbf{x}, \mathbf{\Pi}) \propto \delta(H(\mathbf{x}, \mathbf{\Pi}) - E)$$

$$\nabla \mathcal{L}(\mathbf{x})$$



Microcanonical HMC converges while conserving energy

Robnik+2022

Many more opportunities for the use of ML/AI in astrophysics

- E.g., see the following websites
  - Machine Learning for Astrophysics
  - Center for Computational Astrophysics
  - The NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI) (at MIT)
  - SkIA
  - Smsharma: A community sourced list of papers and resources on simulations-based inference (cosmology/astrophysics dominates the papers)
  - …

Summary (discussion items)

- Fundamental cosmology: many exciting opportunities & discovery potential
  - Dark energy (modified gravity), primordial non-Gaussianity (inflation physics), neutrino mass, dark matter …
- Big data cosmology or data driven cosmology: we don't yet have an optimal methodology for extracting the full information
  - Summary statistics ⇔ Field-level inference (explicit LF inference ⇔ LF-free inference)
  - Simulation based inference
  - Faster sampler for cosmological parameter inference
  - Emulation of simulation data or posterior distribution
- ML/AI methods are clearly needed for big data cosmology
  - Many opportunities and Many challenges