Challenges in AI/ML at the Cosmic Frontier

Simone Ferraro

(Lawrence Berkeley National Lab)

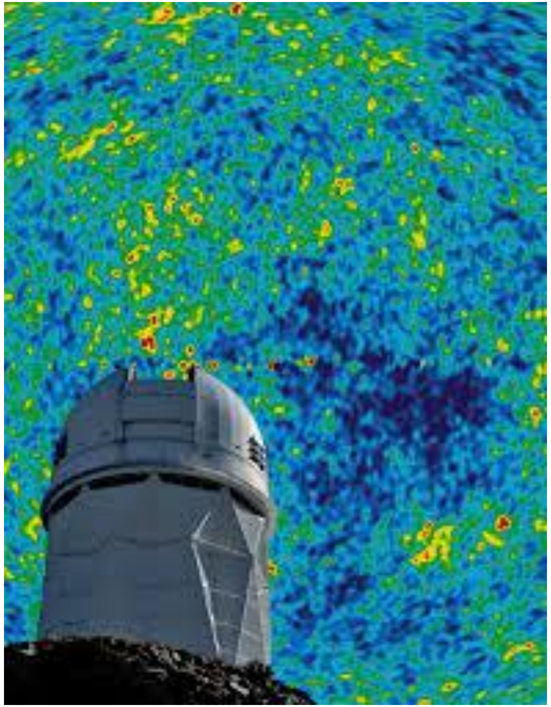
BERKELEY LAB

BERKELEY CENTER *for* COSMOLOGICAL PHYSICS

SLAC Summer Institute

Aug 8, 2023

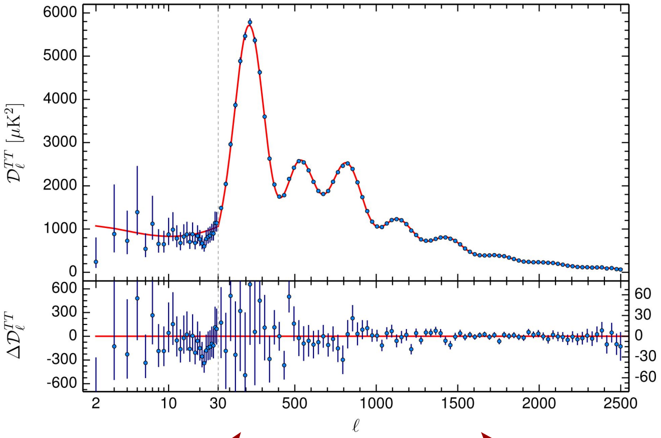
Plan for this talk

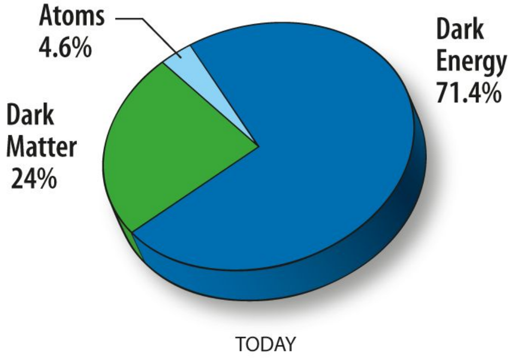


- ***Part I:*** What is the Cosmic Frontier and what are the observables?
- ***Part II:*** Are we being efficient (speed)?
- ***Part III:*** Are we extracting the whole information?
- ***Part IV:*** Do we understanding the data?

Part I:
What is the Cosmic Frontier and what are the observables?

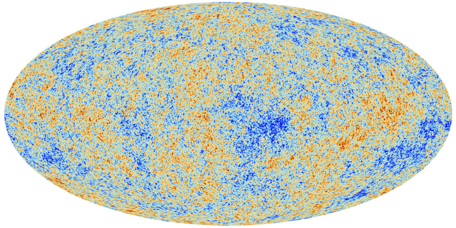
A simple yet strange Universe

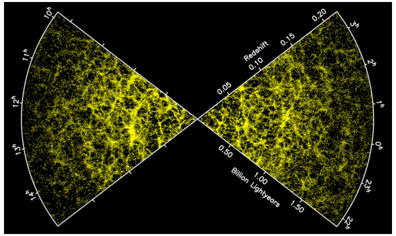




But the model is based on…

- **Dark Matter** (?)
- **Dark Energy** (?)
- **Inflation** (?)
- **Neutrinos and other light particles** (?)

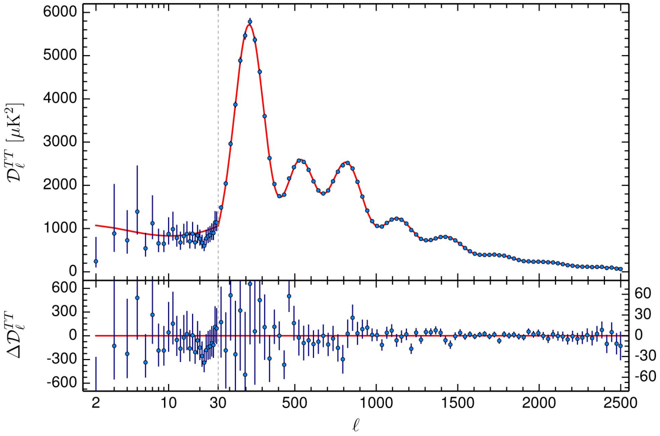


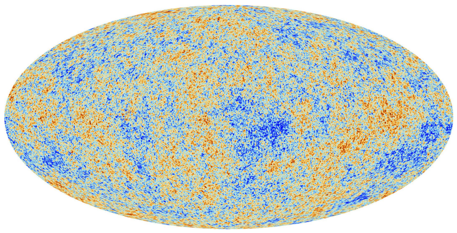


Planck, BOSS

A major goal of the Cosmic Frontier program is to understand these "ingredients"!

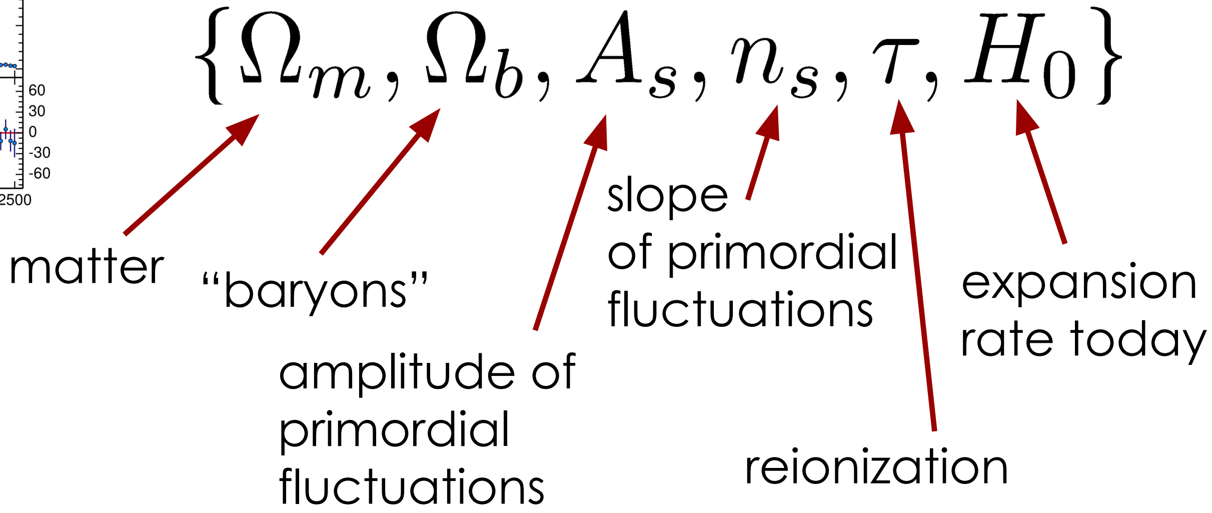
A simple yet strange Universe





Planck, BOSS

Fit fully characterized by 6 numbers*
(with no evidence of needing more):

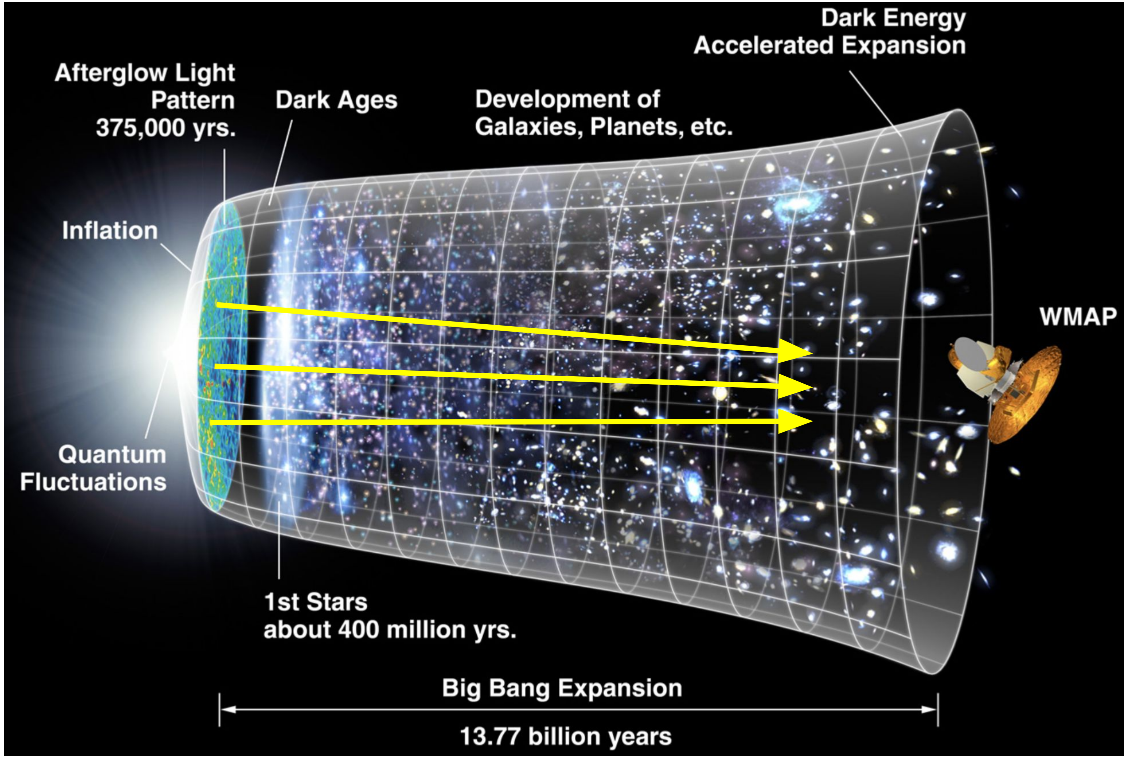


$$\{\Omega_m, \Omega_b, A_s, n_s, \tau, H_0\}$$

matter
"baryons"
amplitude of primordial fluctuations
slope of primordial fluctuations
reionization
expansion rate today

* and a few assumptions such as flat geometry and minimum mass neutrinos

A brief history of the Universe



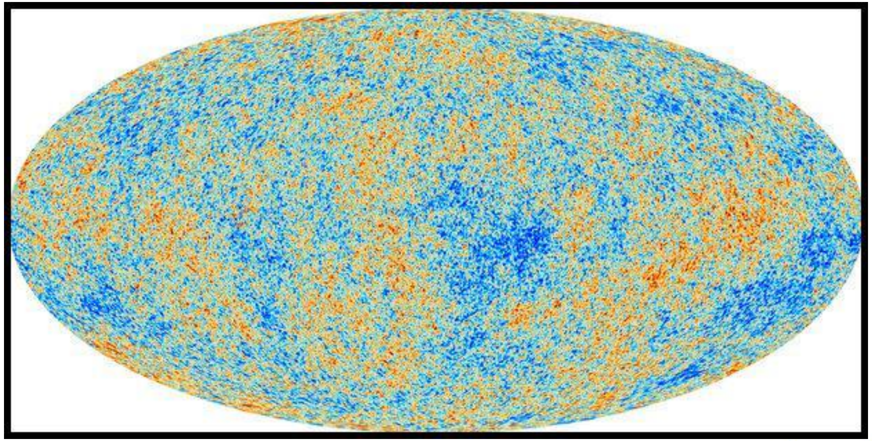
94% of photons travel from the CMB to us without scattering*

6% scatter with matter

On small scales, the Cosmic Microwave Background (CMB) contains a "map" of the entire observable universe

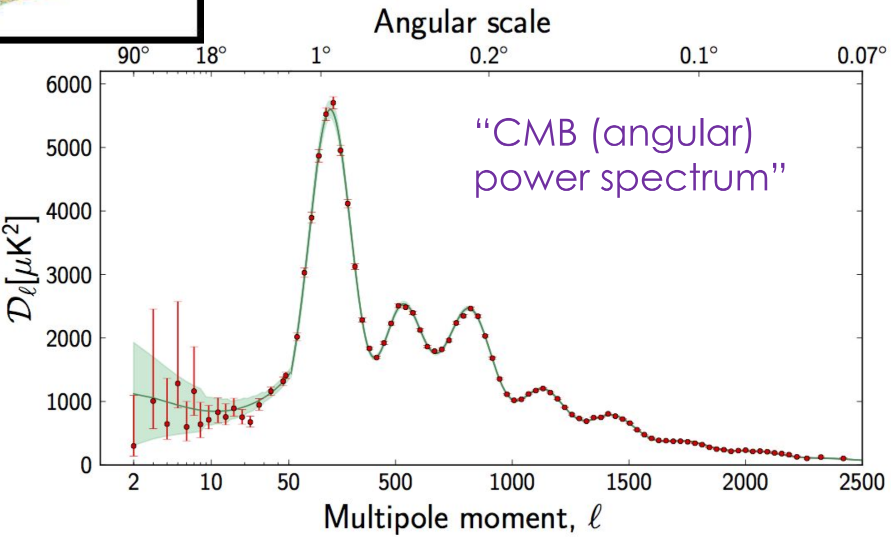
*path slightly deflected by gravitational lensing

Cosmic microwave background (CMB)



Planck Satellite (2018)

“primary fluctuations”



“CMB (angular) power spectrum”

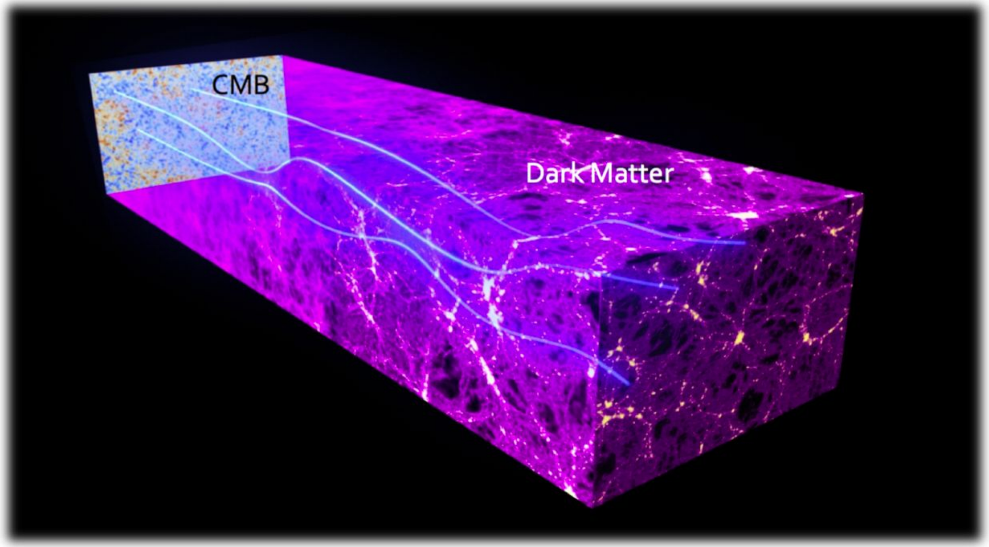
- Large scales (< 1 deg) □ primordial

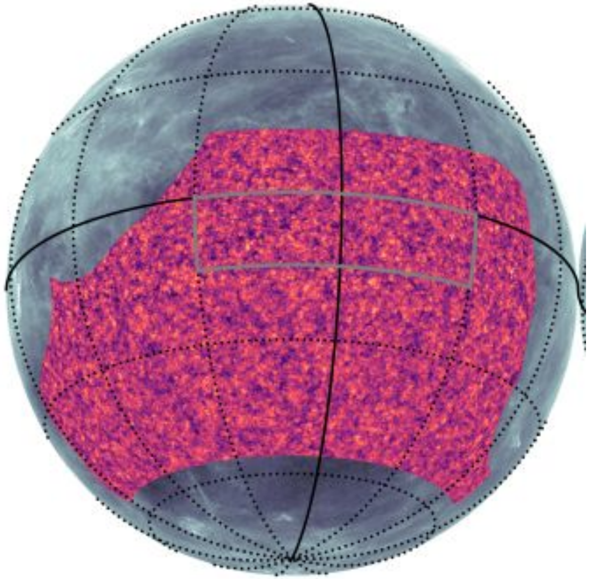
- Smaller scales (> 1 deg) □ processed by (known) plasma physics + gravity

CMB lensing

Paths of CMB photons deflected by matter □ create statistical anisotropy that can be measured

Can make maps of the projected matter density (including Dark Matter) to the CMB!

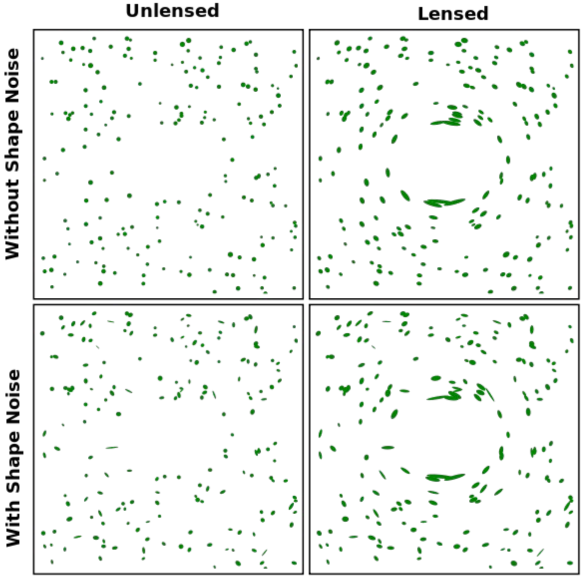




ACT DR6 lensing map

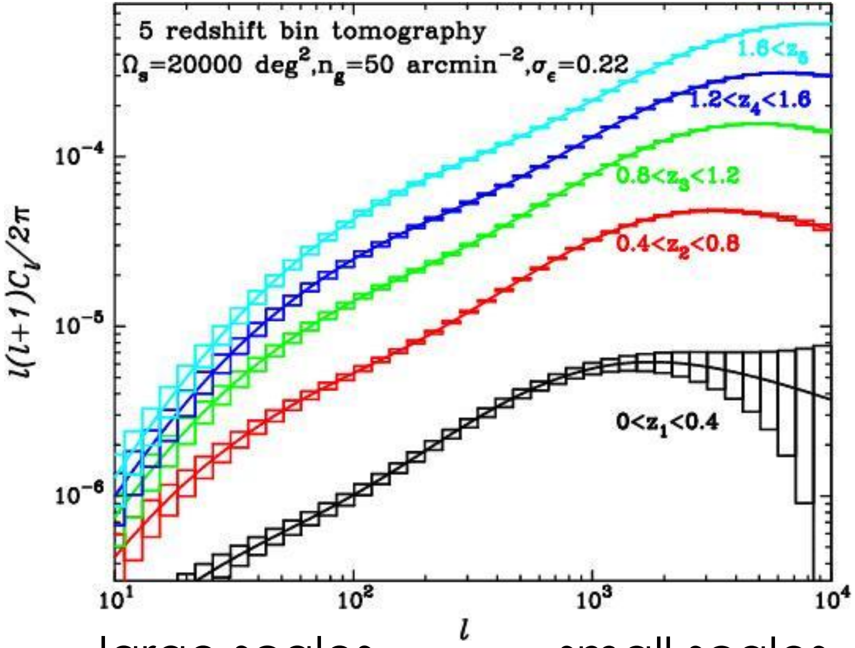
Galaxy (weak) lensing

Rubin Observatory
LSST



Wikipedia

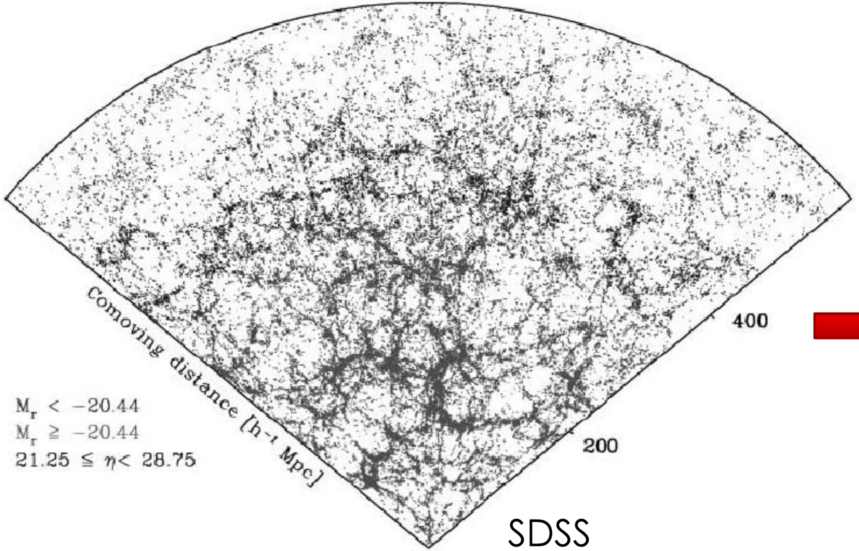
“cosmic shear power spectrum”



large scales

small scales

Large Scale Structure (LSS)

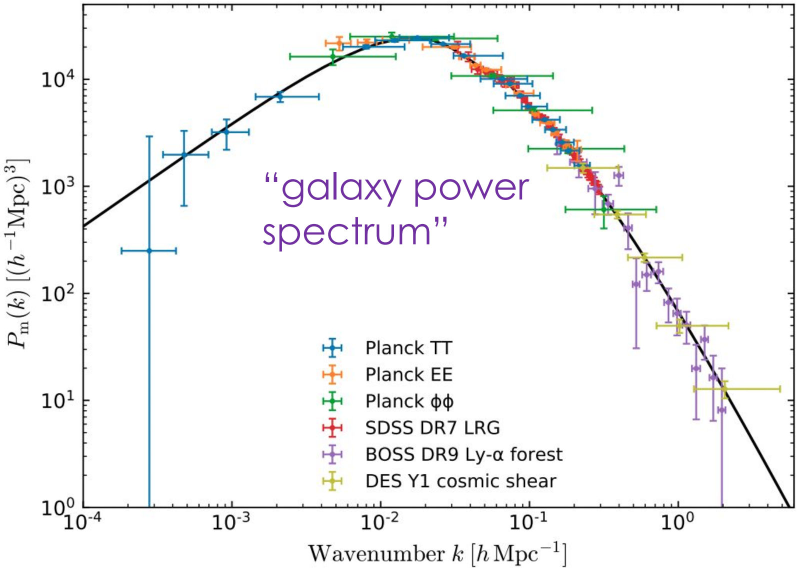


SDSS

Each dot is a real galaxy!

large scales

small scales



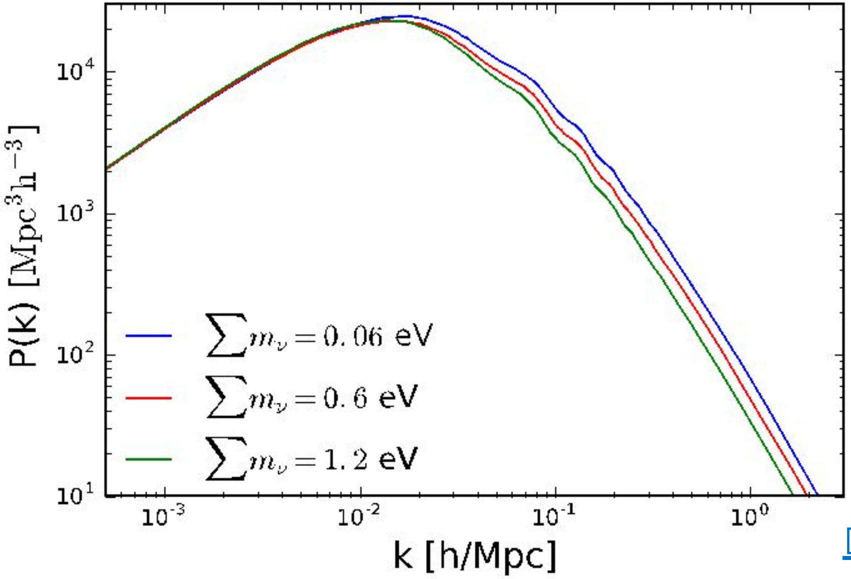
"galaxy power spectrum"

Neutrino mass effect on LSS

The small-scale matter power spectrum, k>k$_{fs}$, is reduced in presence of massive neutrinos:

- On larger scales νs cluster in the same way as cold dark matter
- Free-streaming νs do not cluster
- The growth rate of CDM and baryon fluctuations is reduced

$$k_{fs}\;'\;\;0.018\;□_m^{1/2}\;□\frac{M_□}{1\text{eV}}□^{1/2}\;h\;\text{Mpc}^{-1}$$



Power suppression at small scales

$$\frac{\Delta P(k)}{P(k)}\;□\;-\;8f_□$$

$f_\nu = \omega_\nu / \omega_m$
$= 0.5\%$

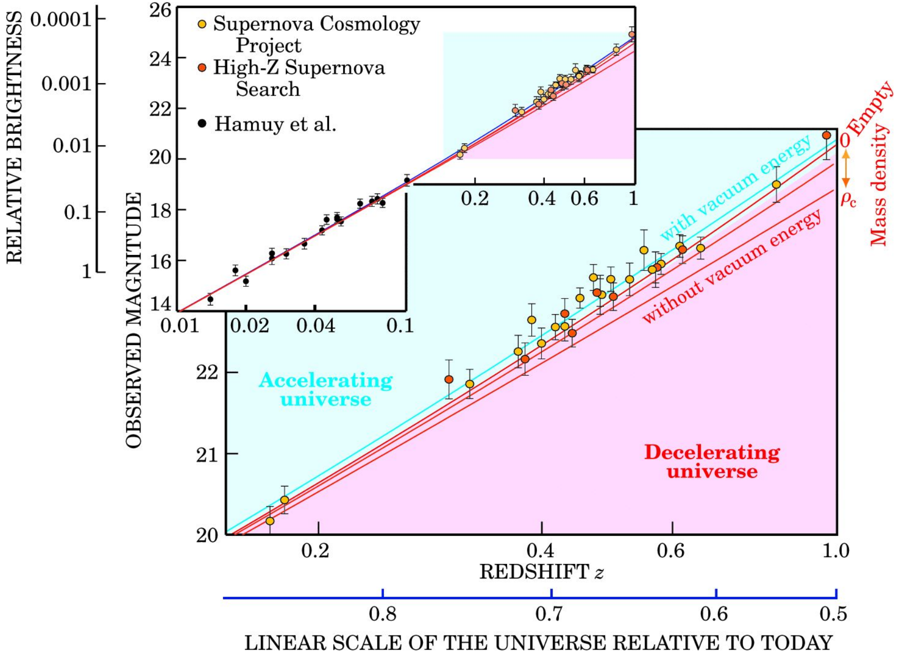
□ 4% suppression minimum!

+ non-linear calculations: additional suppression at large k see Villaescusa-Navarro et al. 2013

Transients

<u>Extragalactic:</u>

- Supernovae/kilonovae
- Fast Radio Bursts, gamma ray bursts
- Tidal disruption events
- Strong lensing time delays
- …

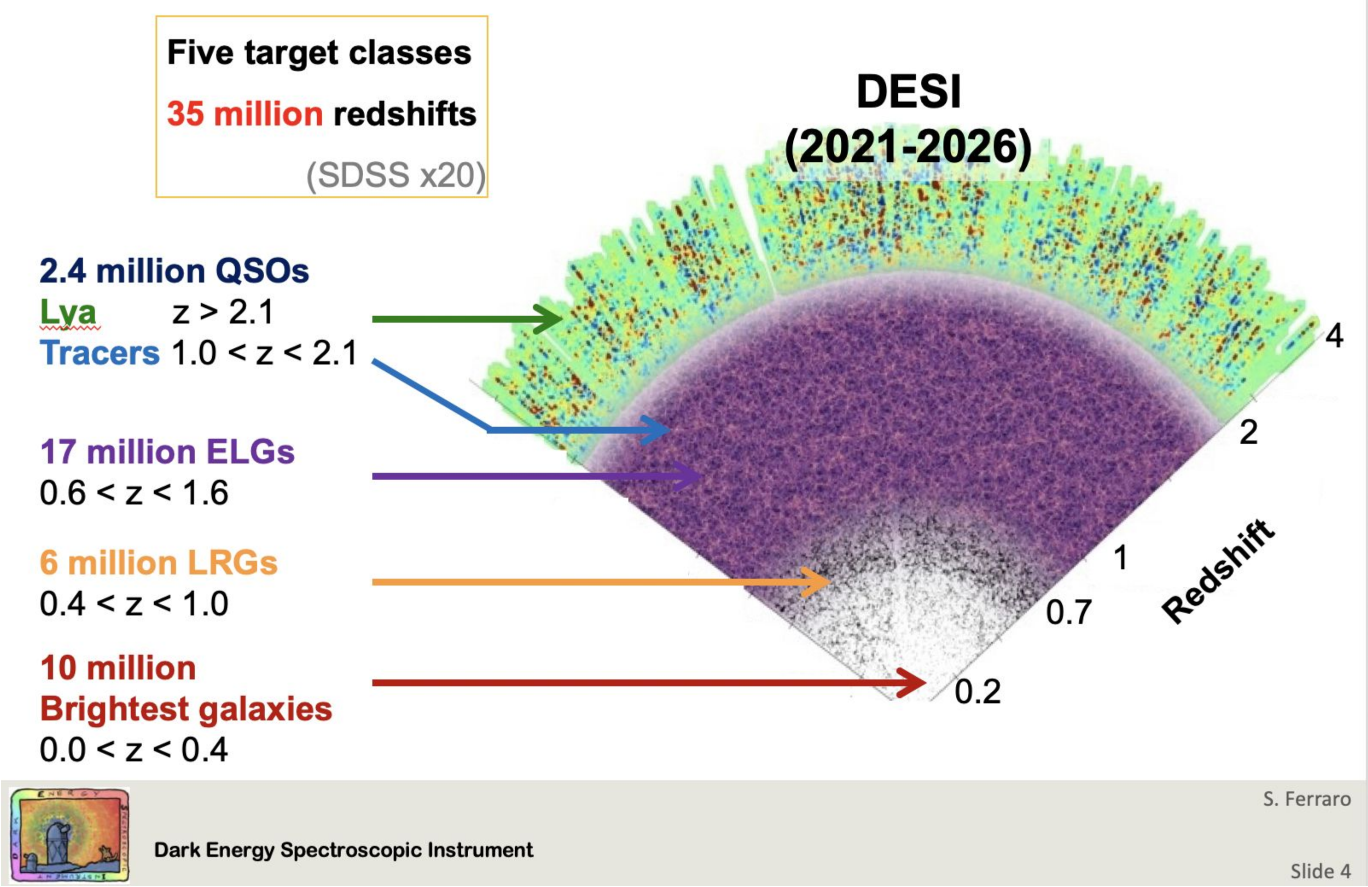


<u>Galactic:</u>

- Asteroids
- Interacting binaries
- Transiting exoplanets
- Microlensing
- Pulsars
- …

DESI

Dark **E**nergy **S**pectroscopic **I**nstrument: Massively multiplexed spectroscopic survey with 5000 robotic fibers, over ~14,000 sq. deg



Five target classes

35 million redshifts

(SDSS x20)

2.4 million QSOs
Lya z > 2.1
Tracers 1.0 < z < 2.1

17 million ELGs
0.6 < z < 1.6

6 million LRGs
0.4 < z < 1.0

10 million
Brightest galaxies
0.0 < z < 0.4

Dark Energy Spectroscopic Instrument

S. Ferraro

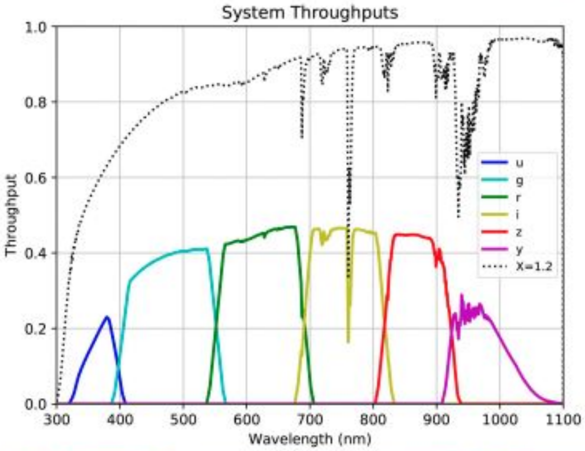
Slide 4

The Vera C. Rubin Observatory

Location: El Peñón, Cerro Pachon, Chile
(median seeing 0.67 arcsec)

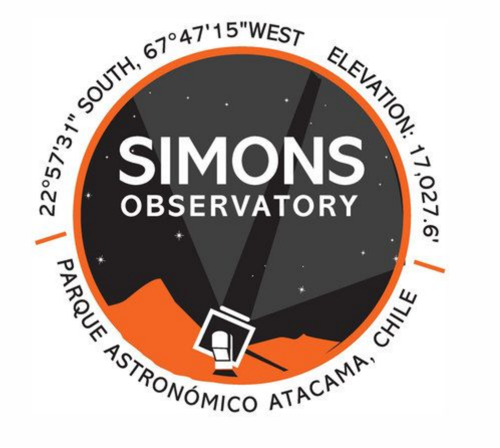
Specs: 8.4m mirror, 3.2 Gigapixels camera
9.6 sq. deg. field of view (~40 full moons),
6 broadband filters (*ugrizy*)

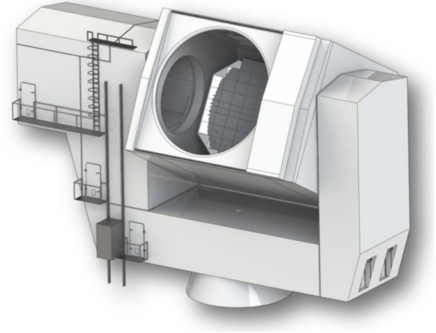
It will perform the 10 year LSST survey of the sky



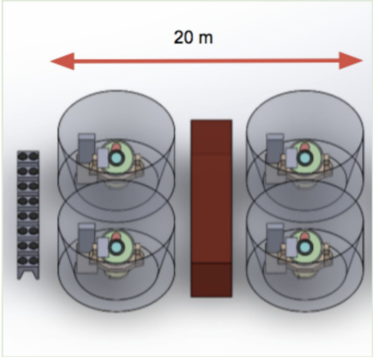
Optical elements seen by thrilled observers

The CMB landscape – mid 2020s


SIMONS
OBSERVATORY
22°57'31" SOUTH, 67°47'15"WEST
ELEVATION: 17,027.6'
PARQUE ASTRONÓMICO ATACAMA, CHILE



Large Aperture Telescope
one 6 meter in diameter



Small Aperture Telescopes
42 cm refractors

Large frequency coverage (30 – 270 GHz)

- 10 Countries
- 40+ institutions

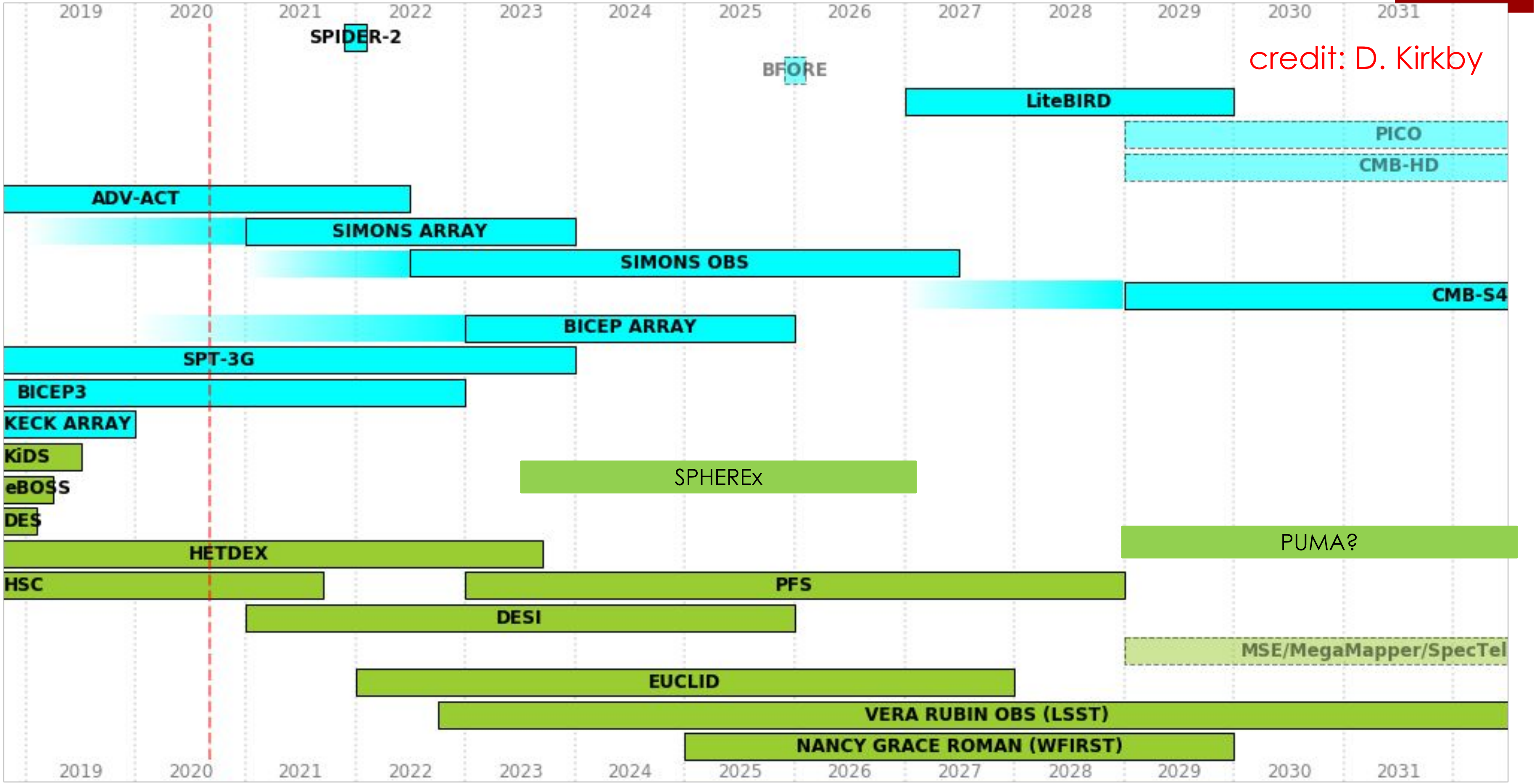
Fully funded
6-year program
First light in 2024!

CMB-S4

- **CMB S4**: next generation ground based experiment
- Factor of ~10 increase in sensitivity
- ETA ~late in this decade

- Multi-agency effort (DOE & NSF)

Looking ahead: the “explosion” of surveys



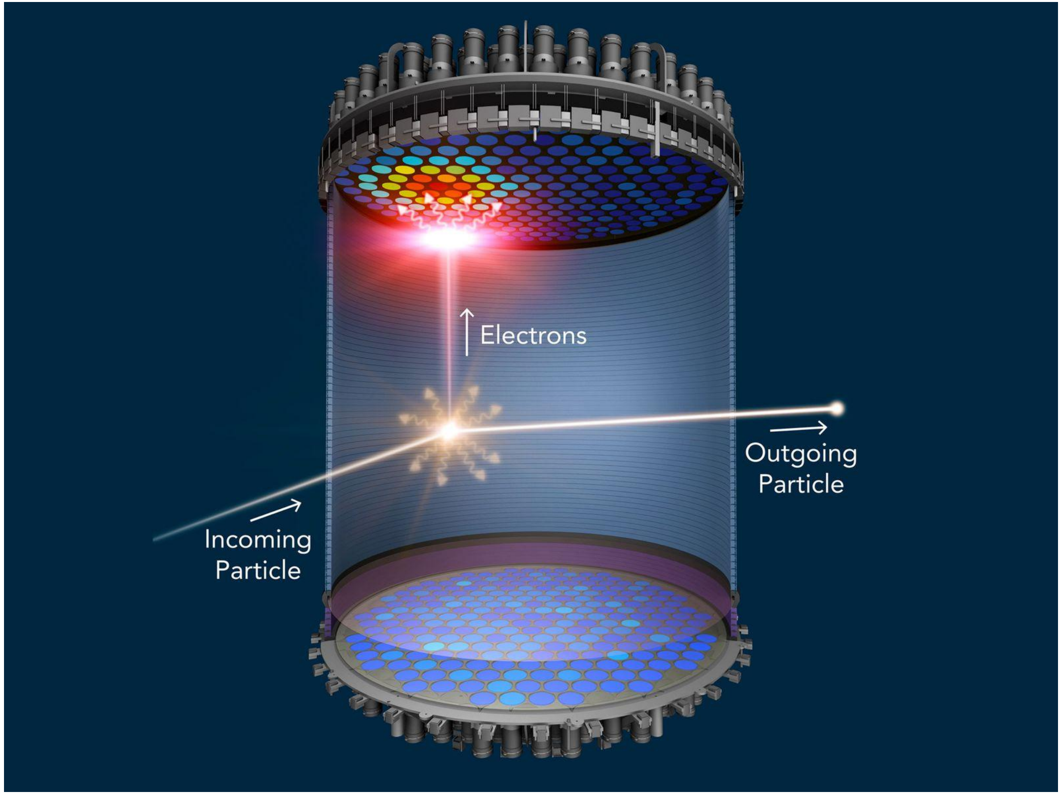
credit: D. Kirkby

Dark Matter direct and indirect detection

IF Dark Matter interacts (weakly) with the Standard Model, can look for scattering/recoil (*direct detection*).
Several targets: Xenon, Germanium, etc

Also: "*indirect detection*" in astrophysical systems (Eg. Fermi gamma ray satellite)

→ **Maria Elena Monzani's lecture on anomaly detection!**

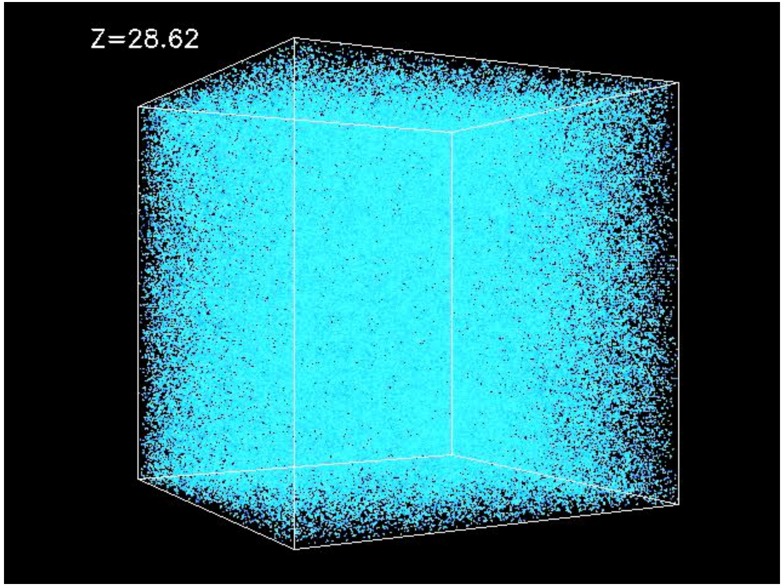


LZ/SLAC

Part II:
are we being efficient (speed)?

Challenge 1: theoretical model

- Complex & non-linear dependence of theory on cosmological parameters even for power spectrum (2 point function). Often no analytical form, and prediction relies on expensive numerical simulations.



Calculating

$$\text{theory}(\Omega_m, \Omega_{DE}, A_s, n_s, \tau, \ldots, \{\text{nuisance parameters}\})$$

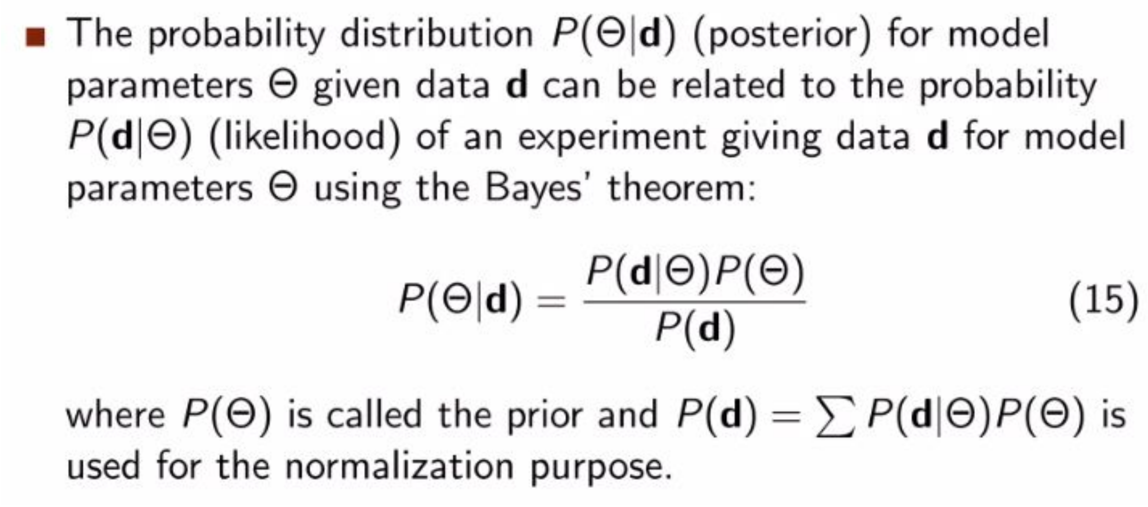
Can take minutes to hours (or more).
Often too slow for parameter inference!

SOLUTION: Build emulators!
Reduce theory calculation to O(ms) per call

□ **Joe DeRose's lecture on emulators**

Challenge 2: parameter inference

- High dimensional problem: typically > 100 parameters (dimensions) inference. Slow or impossible.



- The probability distribution $P(\Theta|\mathbf{d})$ (posterior) for model parameters Θ given data $\mathbf{d}$ can be related to the probability $P(\mathbf{d}|\Theta)$ (likelihood) of an experiment giving data $\mathbf{d}$ for model parameters Θ using the Bayes' theorem:

$$P(\Theta|\mathbf{d}) = \frac{P(\mathbf{d}|\Theta)P(\Theta)}{P(\mathbf{d})} \tag{15}$$

where $P(\Theta)$ is called the prior and $P(\mathbf{d}) = \sum P(\mathbf{d}|\Theta)P(\Theta)$ is used for the normalization purpose.

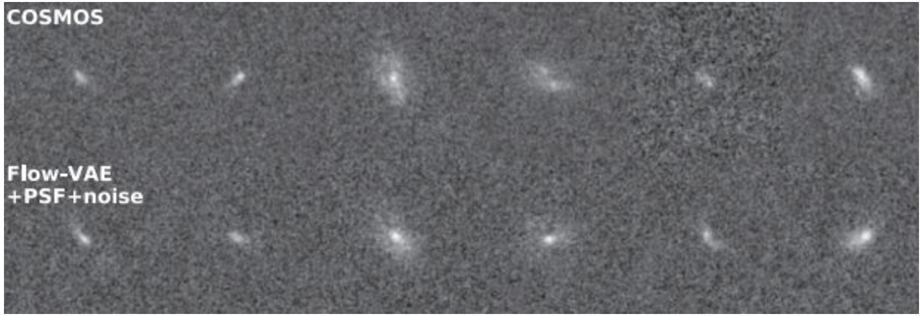
Typically you would sample the posterior by **Monte-Carlo**.
In high-dimension (>100 parameters), algorithms involving ***gradients*** are more efficient (eg. Hamiltonian Monte Carlo).

SOLUTION: Build a differentiable likelihood + differential emulators!

□ **Joe DeRose's lecture**

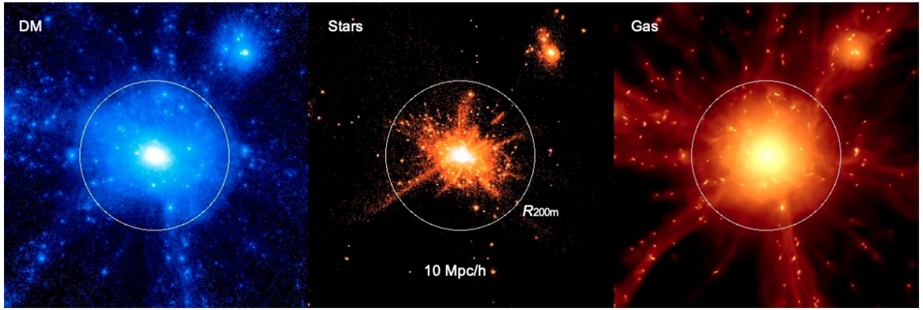
Mock data generation

Lensing data



arXiv:2008.03833

"Gas pasting" on Dark-Matter only simulations



arXiv:2110.02232

Both cases based on (conditional) Variational Auto-Encoder (VAE)

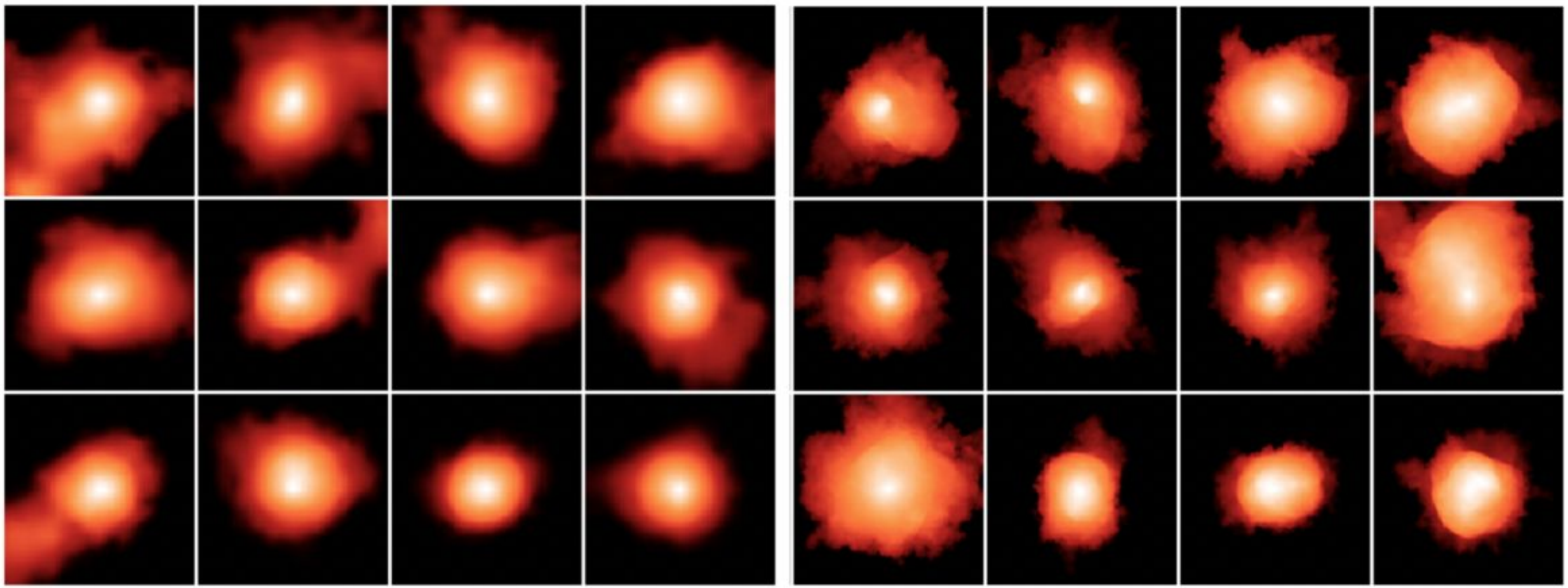
Similar applications with Generative Adversarial Networks (GANs)

Diffusion models?

□ **See lectures on generative models applications**

Mock data generation

Can you tell which one is generated by an (expensive) hydrodynamical simulation and which is generated by CVAE?

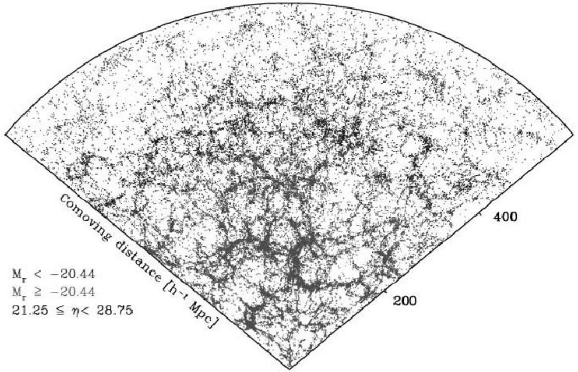


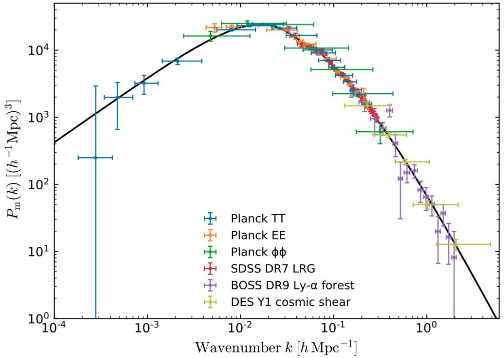
Part III:
are we extracting the whole information?

Beyond the power spectrum

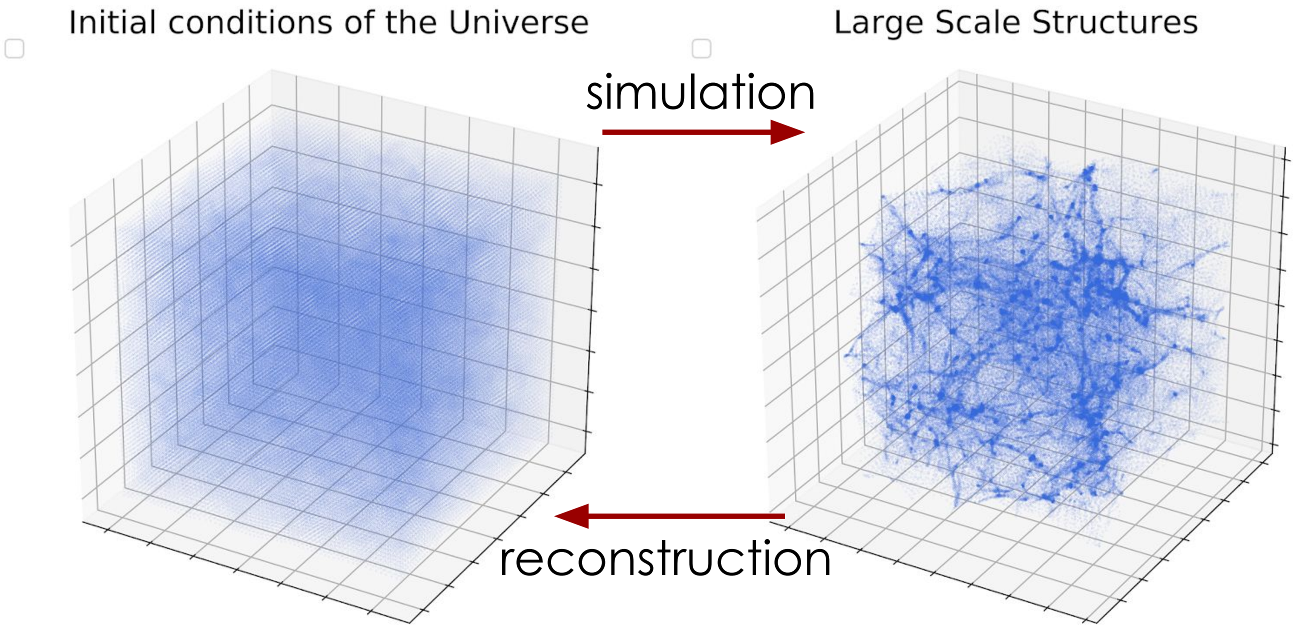
- Power spectrum (2pt function) contains the whole information only for *Gaussian* fields
- Can perturbatively consider 3pt function and higher, but (in general), limited information available.
- But in general, no guidance on what's the most informative statistic…

- Several options available:
  - Field-level inference
  - Compression in a "small" number of summary statistics
  - In both cases, likelihood may not be known analytically □ likelihood-free/simulation-based inference)





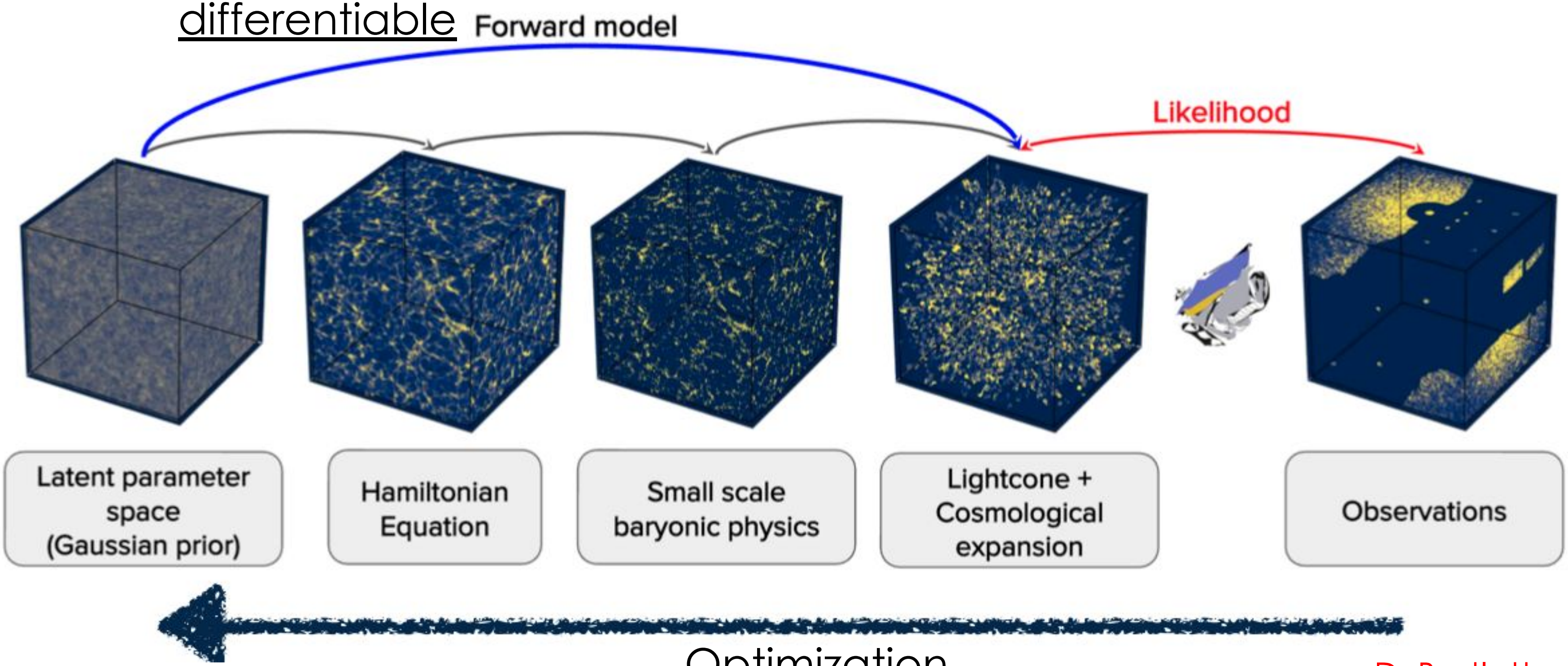
Field-level inference



The initial conditions are very close to Gaussian: they contain the whole information. Can we reconstruct them?

Field-level inference

differentiable


Forward model

Likelihood

Latent parameter space (Gaussian prior)

Hamiltonian Equation

Small scale baryonic physics

Lightcone + Cosmological expansion

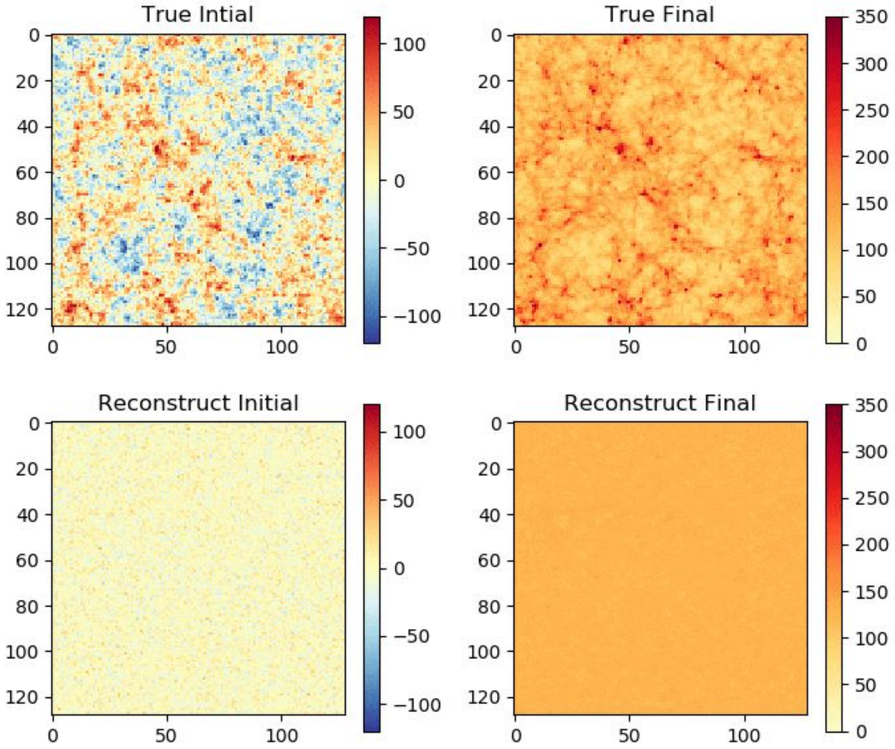
Observations

Optimization

D. Bartlett

“solving the inverse problem by optimization”

Field-level inference



Optimization converges in O(20 steps), even though $N_{dim} = N_{pix} \sim 10^6$ or more!

But: want to marginalize over the initial conditions to extract cosmological parameters. Active area of research and questions remain!

Can use Laplace approximation or MUSE (see arXiv:2112.09354).

Or… full HMC sampling (eg. BORG https://www.aquila-consortium.org/)

□ See "Introduction to Differentiable Programming in Jax" (F. Lanusse)

https://blog.tensorflow.org/2020/03/simulating-universe-in-tensorflow.html (Modi, Lanusse et al)

More generally: the full problem

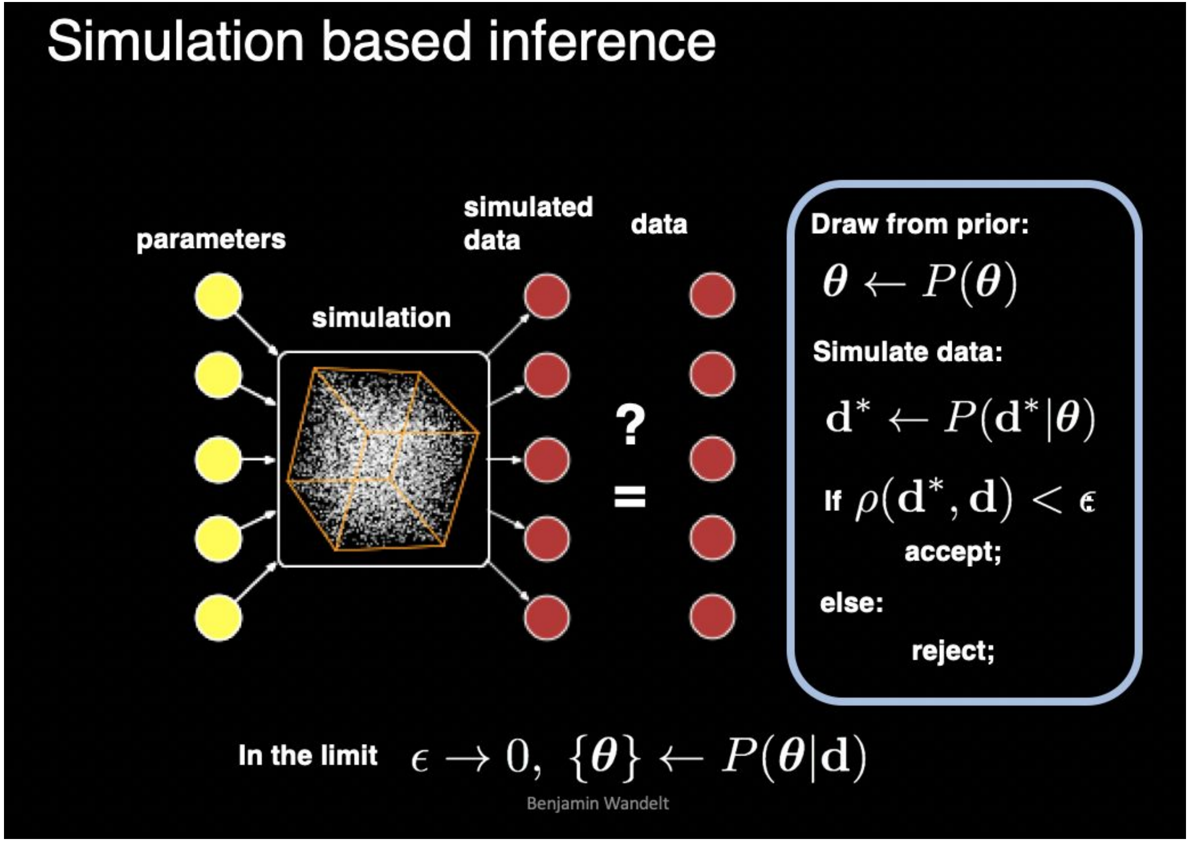
Often we need more freedom than a traditional likelihood approach:

- We may not know what the likelihood is (Gaussian approximation is often a bad one!)
- We may summarize, cut, mask the data any way we want
- Observational or instrumental effects are hard to treat analytically but easy to simulate.

Simulating data is often much easier than deriving an accurate likelihood
▯ **Simulation-Based Inference** (SBI)

SBI = Inference "engine" when explicit likelihood is intractable or unknown, but simulation is possible.

Simulation based inference introduction



Simulation based inference

```
Draw from prior:
    $\theta \leftarrow P(\theta)$
Simulate data:
    $d^* \leftarrow P(d^*|\theta)$
If $\rho(d^*, d) < \epsilon$
    accept;
else:
    reject;
```

In the limit $\epsilon \to 0, \; \{\theta\} \leftarrow P(\theta|d)$

Benjamin Wandelt

Simplest implementation of "**Approximate Bayesian Computation**" (ABC)

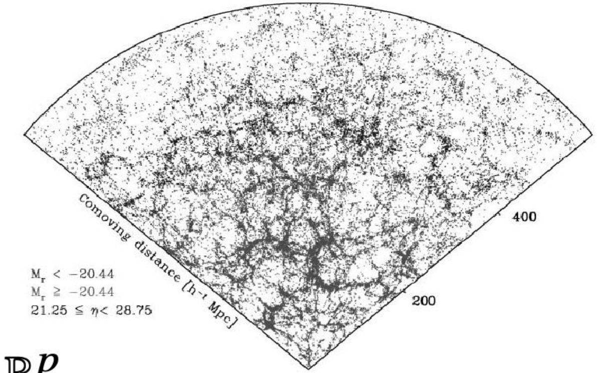
Suffers from severe "Curse of dimensionality"

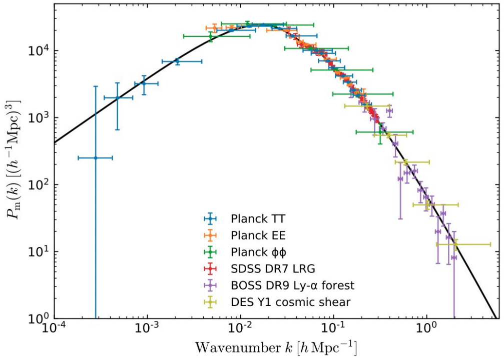
Compression beyond the power spectrum

- Beyond the power spectrum, we have little guidance on what's the most informative statistic…

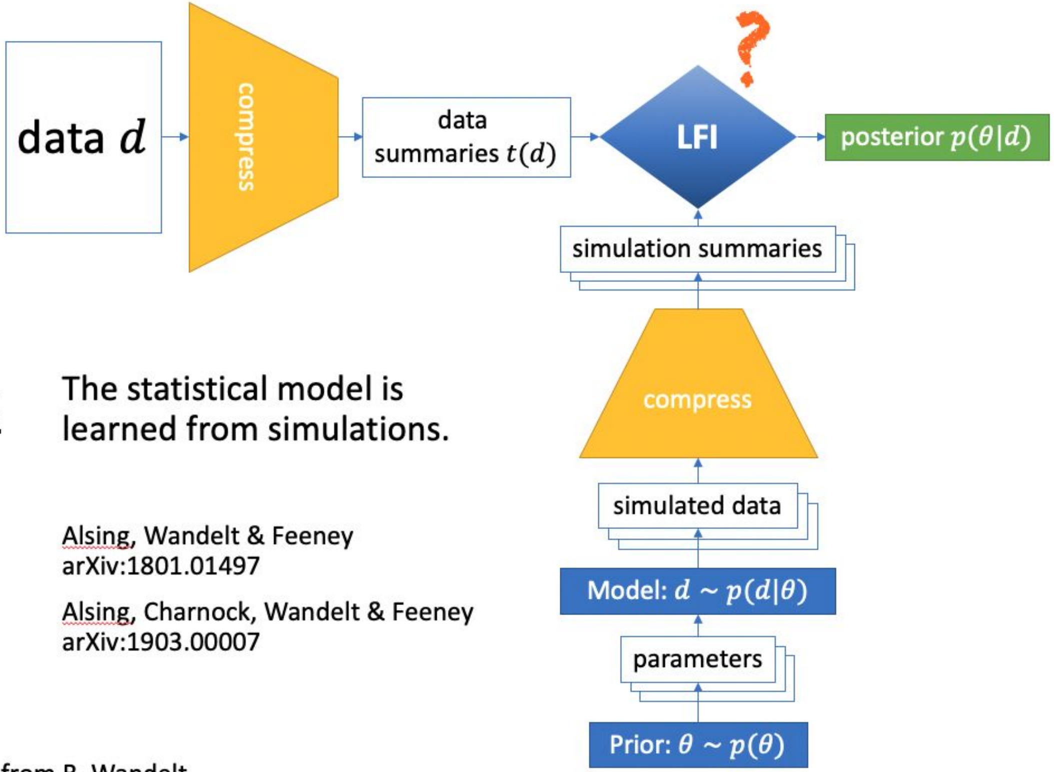
- *Score compression* and *Information Maximizing Neural Networks* (IMNN): $\mathbf{t}(\mathbf{d}) : \mathbb{R}^N \rightarrow \mathbb{R}^p$ (p < N) produce a small number of summary statistic that maximize the retained Fisher information.

- See arXiv:1802.03537 for more info.





Likelihood-free inference (LFI)



Likelihood-Free Inference (LFI) with summaries

The statistical model is learned from simulations.

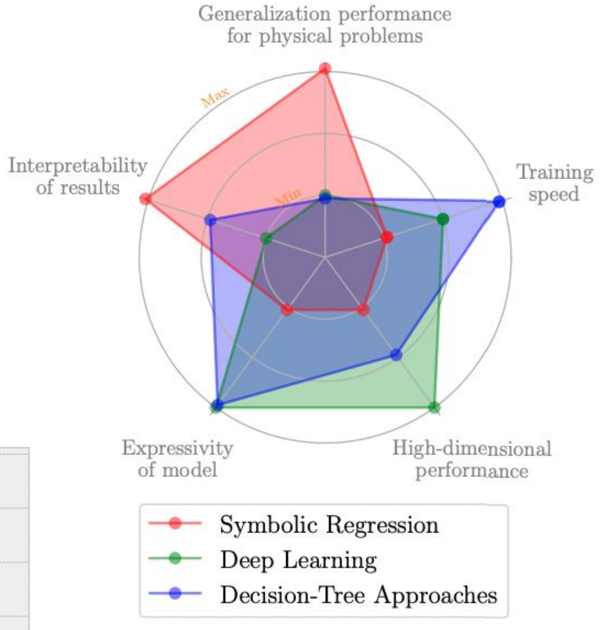
Alsing, Wandelt & Feeney
arXiv:1801.01497

Alsing, Charnock, Wandelt & Feeney
arXiv:1903.00007

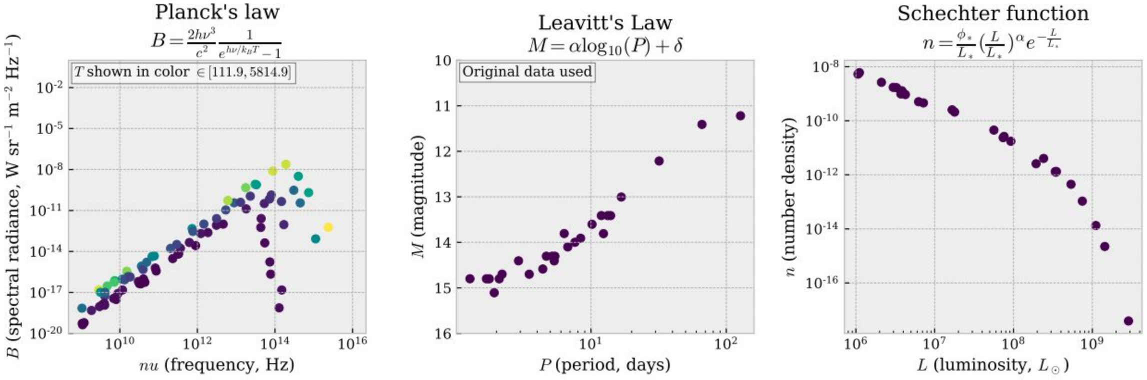
Slide from B. Wandelt

Discovering new relations: symbolic regression

- Search the space of analytic equations to fit some data
- Often done by hand but efficient algorithms exist!
- Concise
- Interpretable



arXiv:2201.01305



arXiv:2305.01582

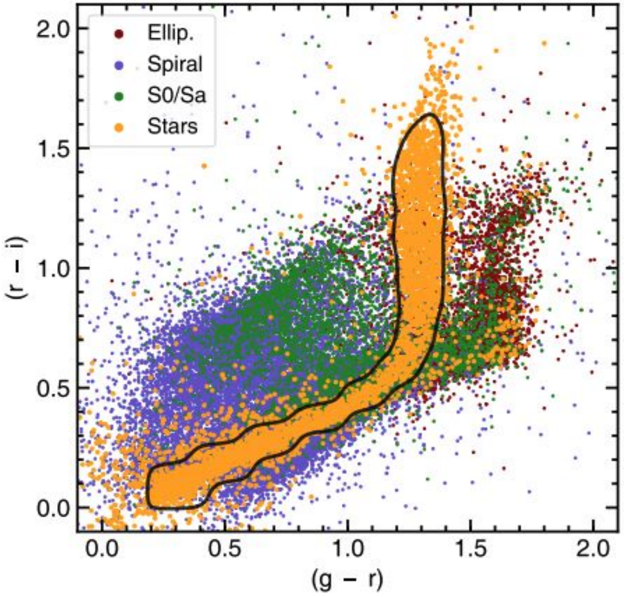
https://github.com/MilesCranmer/PySR

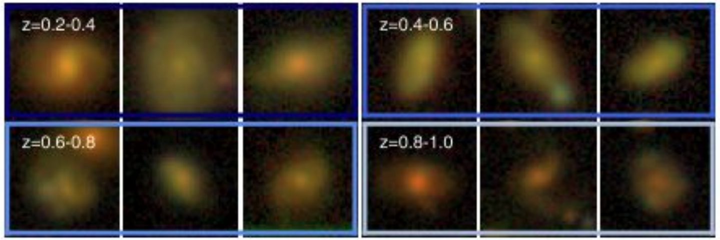
Part IV:
Do we understand the data?

Classification

- Source classification (stars vs galaxies vs quasars etc)
- Transient classification
- Classification of the cosmic web (voids, sheets, filaments etc)
- Photometric redshifts
- …

Both supervised and unsupervised methods. Need to allow for the existence of unexpected patterns!





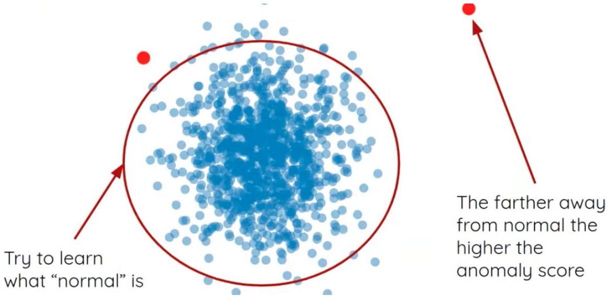
(a) *Spiral galaxies.*

arXiv:1909.10537

Anomaly detection

Several goals:

- Eliminate the influence of outliers or contaminants
- Find new signals when the signal is rare
  - *known unknowns*: supernovae, strong lenses, transients, …
  - *unknown unknowns*: eg. discovery of pulsars



Try to learn what “normal” is

The farther away from normal the higher the anomaly score

credit: M. Lochner

anomaly

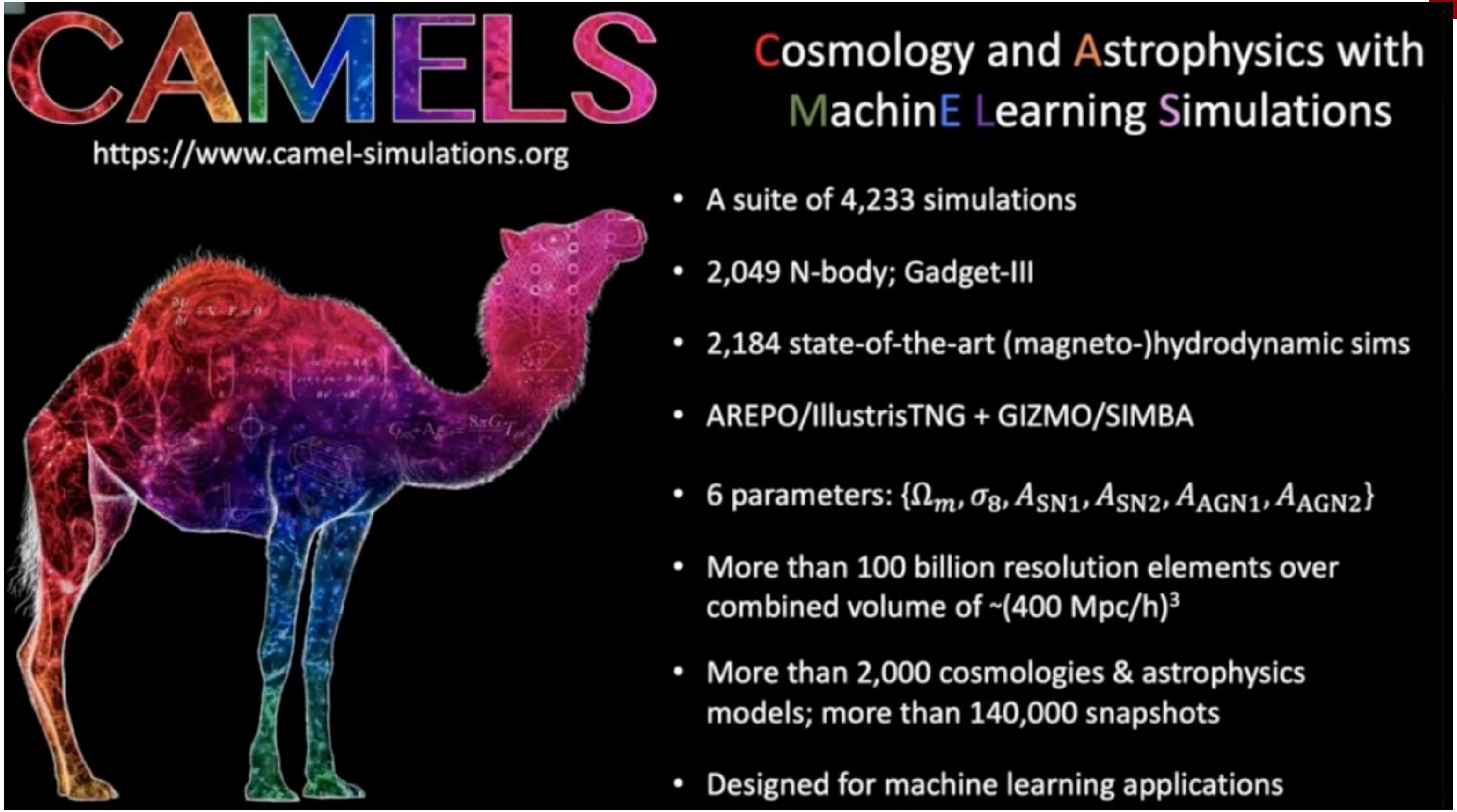
vs

real

https://github.com/MichelleLochner/astronomaly

□ Maria Elena Monzani’s lecture

CAMELS: an “MNIST” dataset for ML in Cosmology



https://www.camel-simulations.org

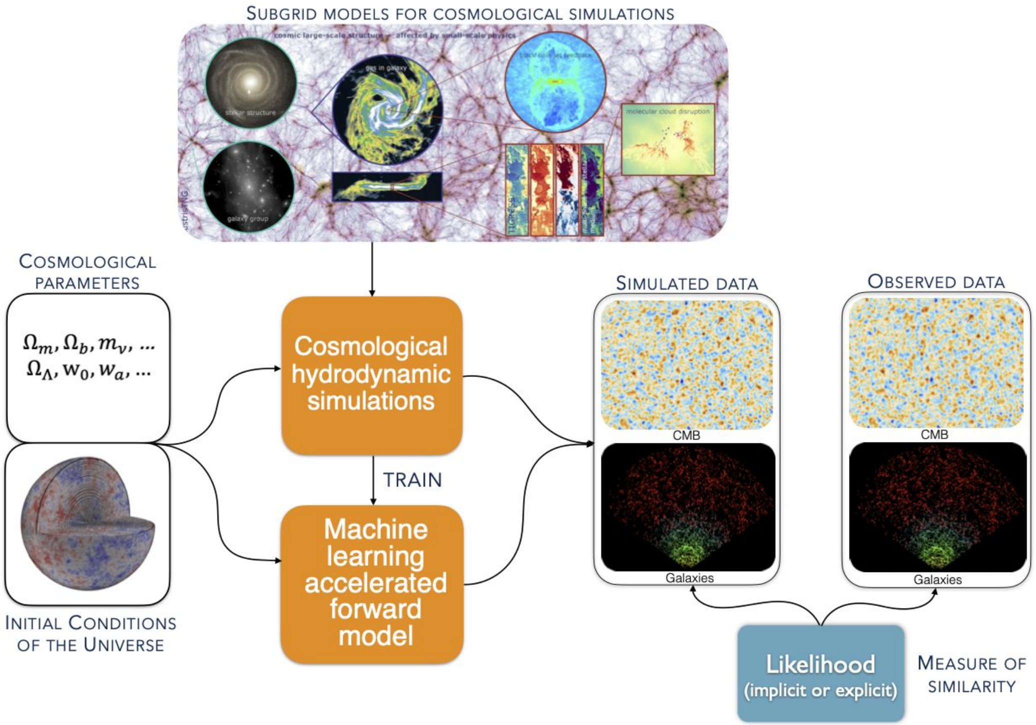
Cosmology and Astrophysics with MachinE Learning Simulations

- A suite of 4,233 simulations
- 2,049 N-body; Gadget-III
- 2,184 state-of-the-art (magneto-)hydrodynamic sims
- AREPO/IllustrisTNG + GIZMO/SIMBA
- 6 parameters: $\{\Omega_m, \sigma_8, A_{\mathrm{SN1}}, A_{\mathrm{SN2}}, A_{\mathrm{AGN1}}, A_{\mathrm{AGN2}}\}$
- More than 100 billion resolution elements over combined volume of ~(400 Mpc/h)^3
- More than 2,000 cosmologies & astrophysics models; more than 140,000 snapshots
- Designed for machine learning applications

F. Villaescusa-Navarro

Learning the Universe

Simons Collaboration on "Learning the Universe"


SUBGRID MODELS FOR COSMOLOGICAL SIMULATIONS
cosmic large-scale structure — affected by small-scale physics
stellar structure
galaxy group
IllustrisTNG
gas in galaxy
black hole jet feedback
TIGRESS
multi-phase interstellar medium
molecular cloud disruption
COSMOLOGICAL PARAMETERS
$\Omega_m, \Omega_b, m_\nu, \ldots$
$\Omega_\Lambda, w_0, w_a, \ldots$
INITIAL CONDITIONS OF THE UNIVERSE
Cosmological hydrodynamic simulations
TRAIN
Machine learning accelerated forward model
SIMULATED DATA
CMB
Galaxies
OBSERVED DATA
CMB
Galaxies
Likelihood (implicit or explicit)
MEASURE OF SIMILARITY

https://www.learning-the-universe.org/

Conclusions

- ML will help us solve cosmological problems that are intractable today
- With great power, come great responsibility! Astrophysical systems are complex and often not fully understood. Model misspecification can lead to issues and great care needs to be taken.
- Finally, for a comprehensive list of ML applications to cosmology, see https://github.com/georgestein/ml-in-cosmology

Thanks!