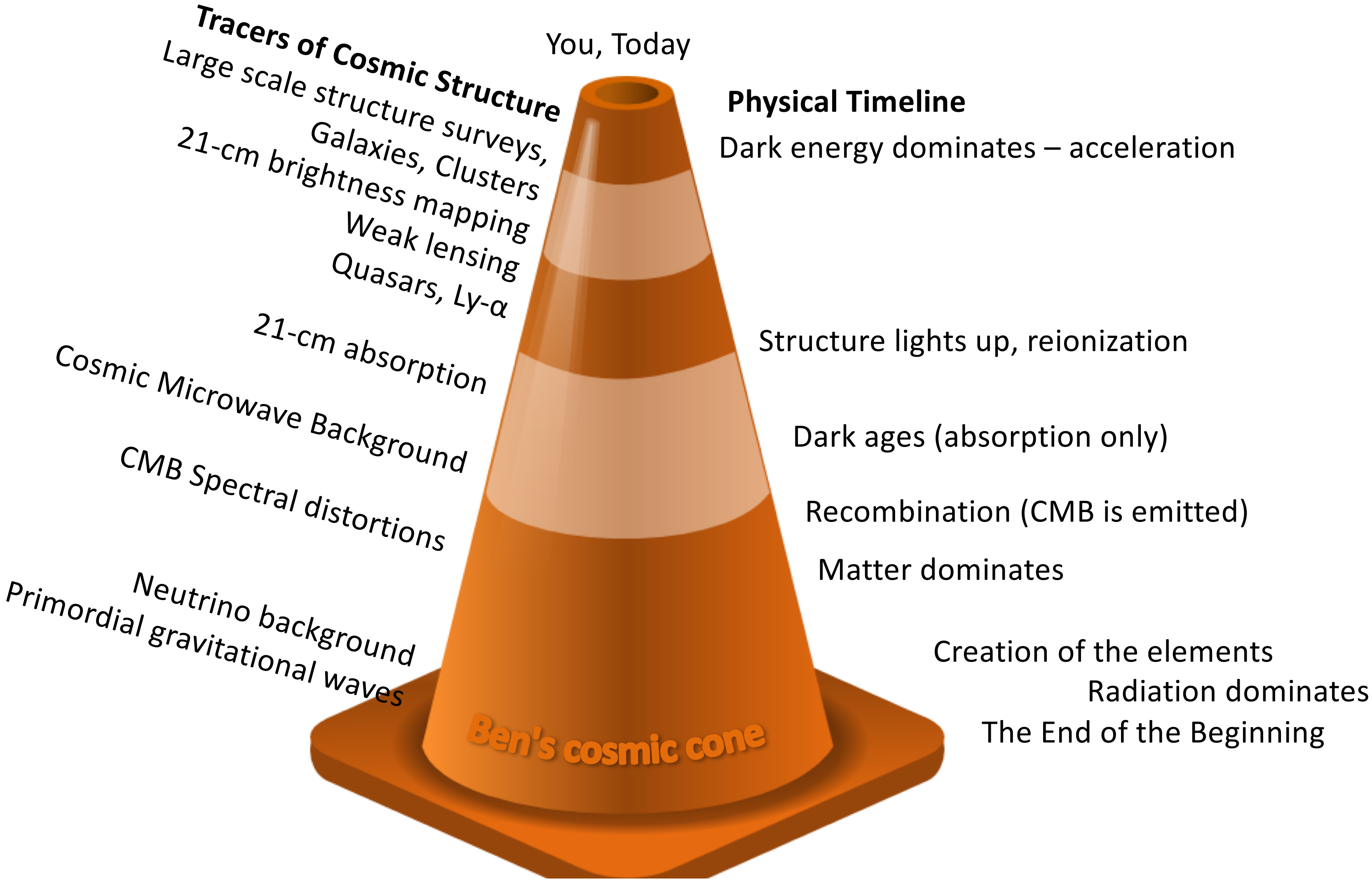
Bayesian Likelihood-Free Inference in Cosmology and Information Maximizing Neural Networks

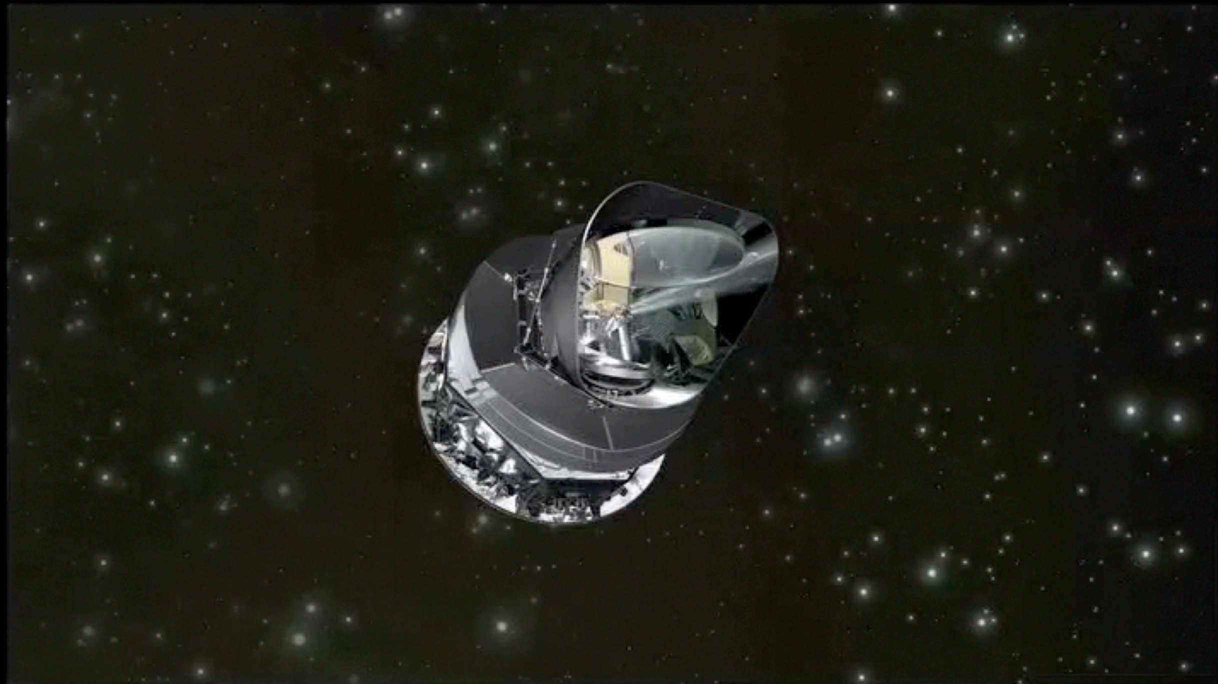
Benjamin Wandelt

with Justin Alsing, Tom Charnock, Guilhem Lavaux, Doogesh Ramanah, Stephen Feeney

Cosmology on one slide



A Journey of Light



through Space and Time

What makes learning from cosmic structure exciting?

We are awash in informative data – astrophysical data volumes have grown exponentially for decades; this will persist for at least another decade

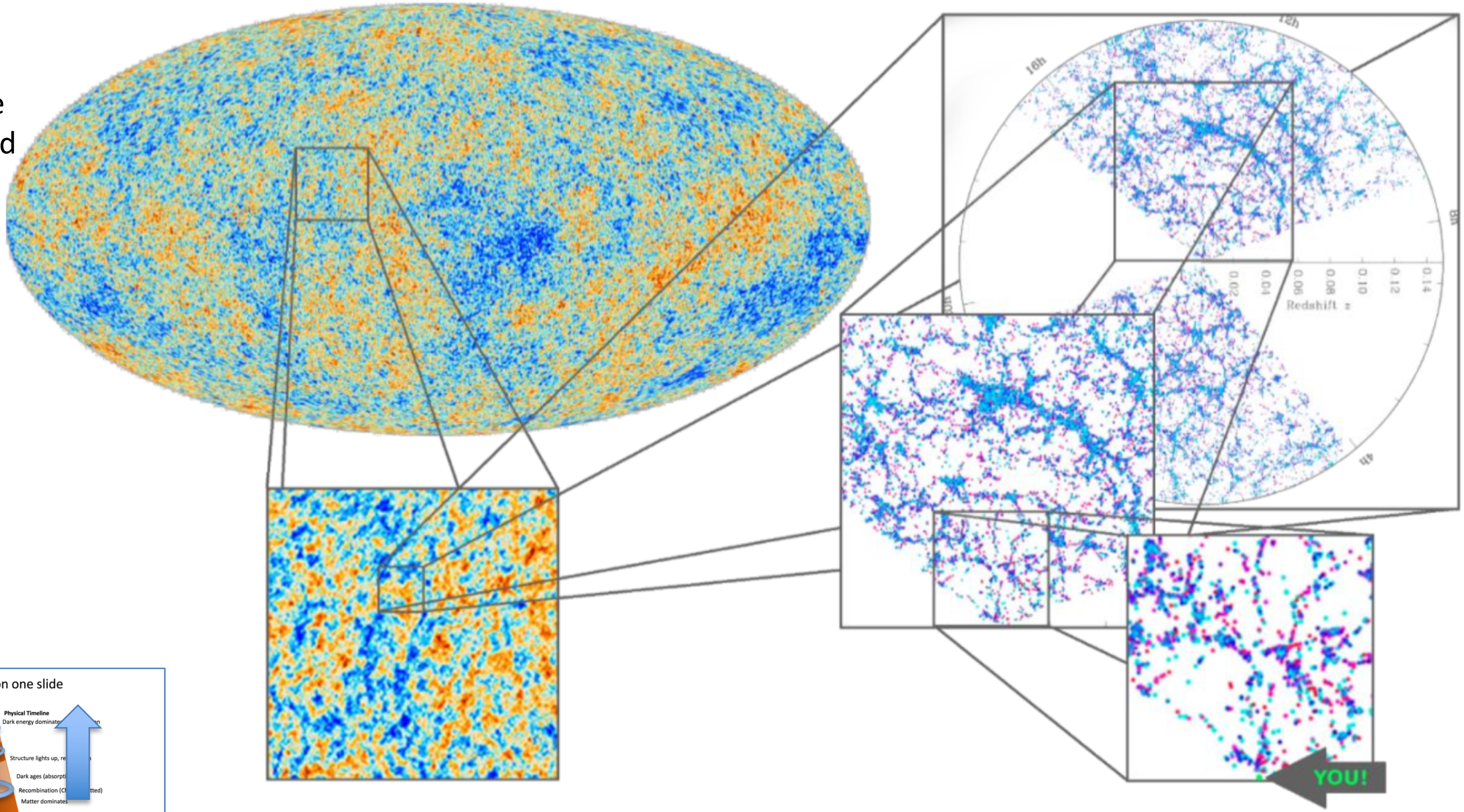
Solid theoretical foundations – can formulate very good priors, often in the form of hierarchical models with strong physical motivation. We have the power of physics on our side.

We live in the era of cosmological data

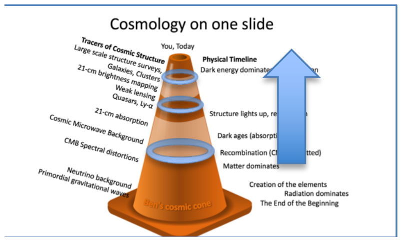
(Your favorite survey here)

Cosmological data covers a hierarchy of scales

Cosmic Microwave Background

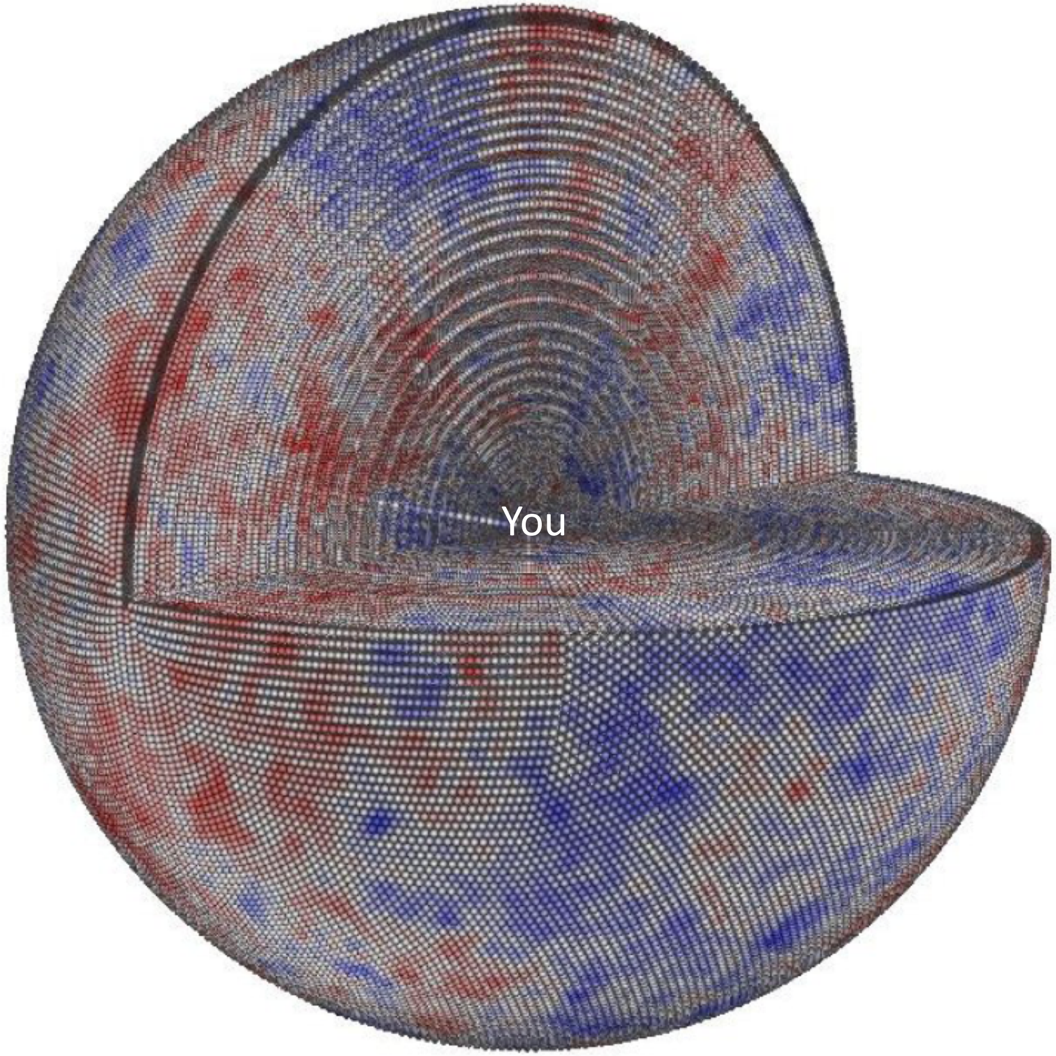


Galaxy surveys

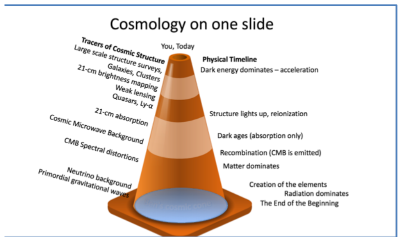


Origin of structure: quantum fluctuations in the primordial universe

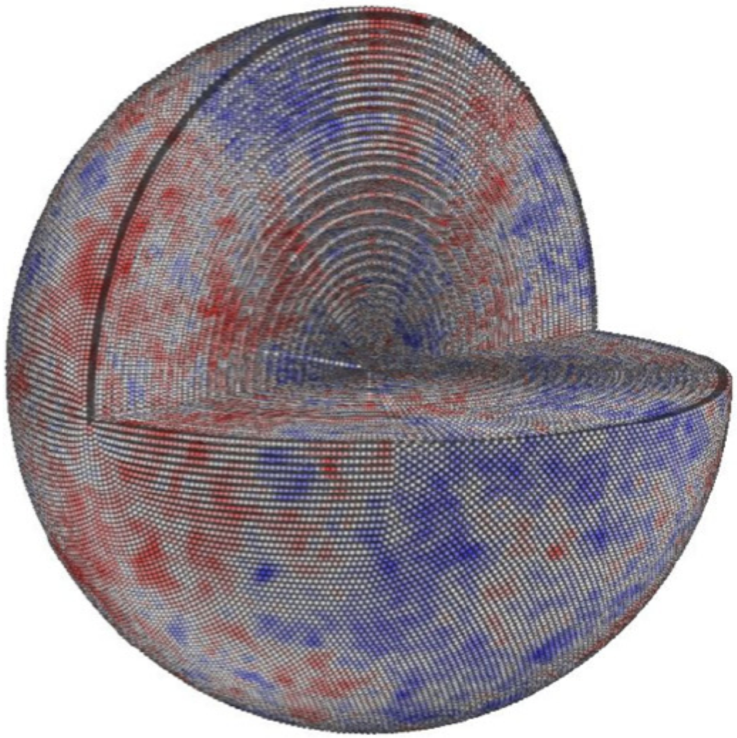
Primordial perturbation



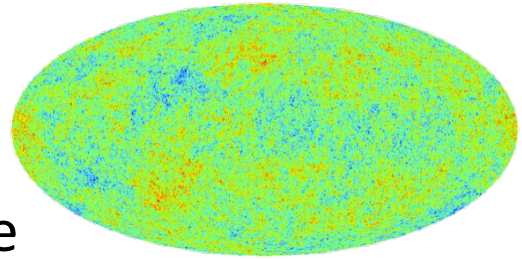
Benjamin Wandelt

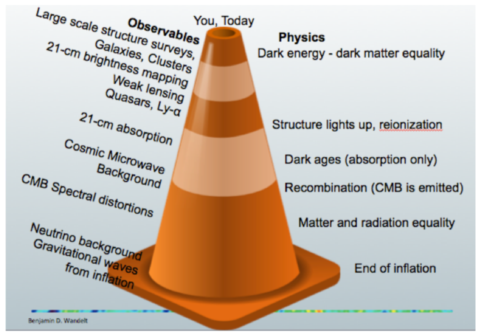


Primordial perturbations give rise to observations

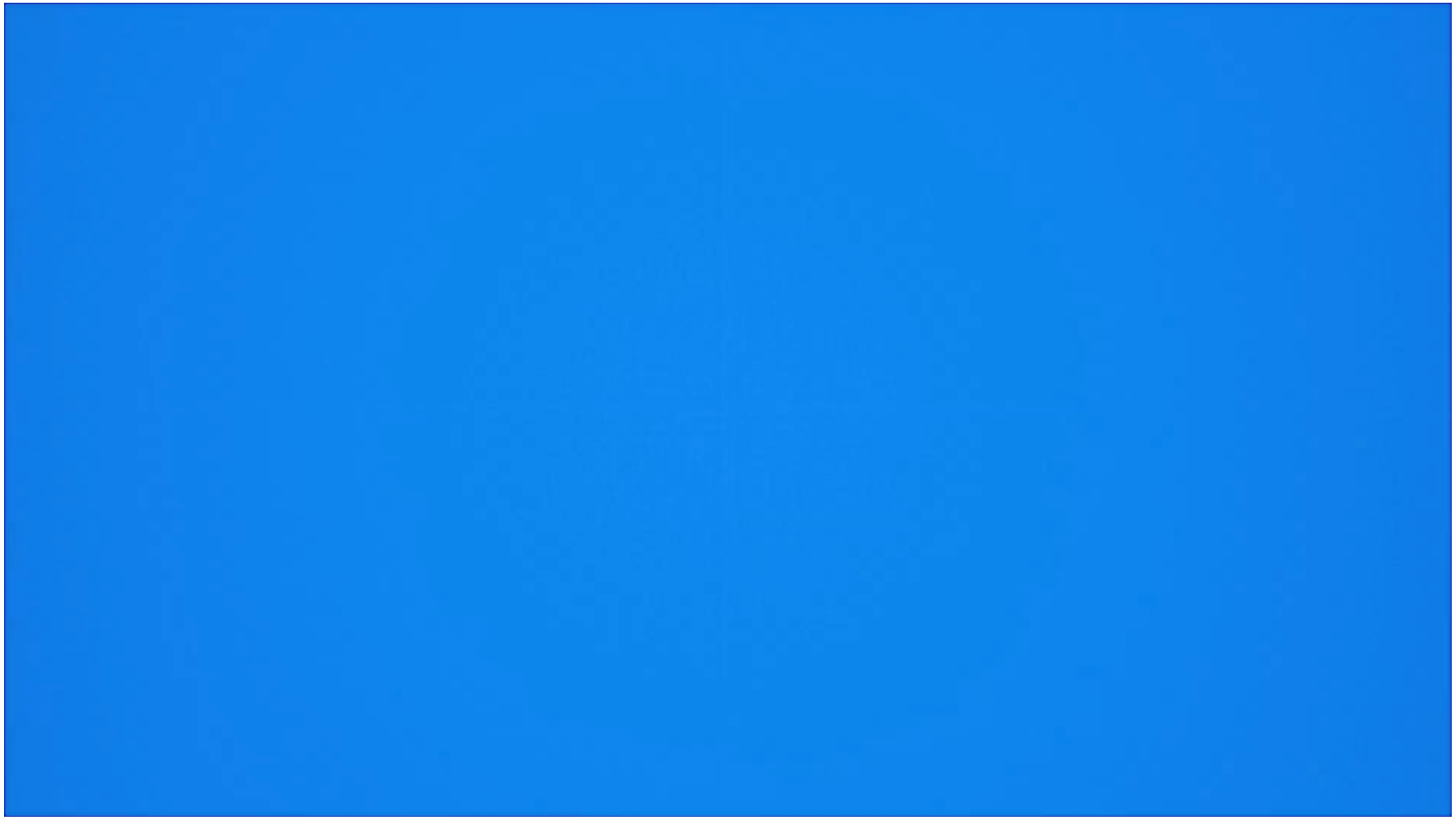


Radiative Transfer
+
Gravity
+
Astro-physics





Physical generative model of matter distribution



Benjamin Wandelt

Movie credit: Stéphane Colombi, IAP

What makes learning from cosmic structure challenging?

Limited information – only one universe!
Careful treatment of uncertainties

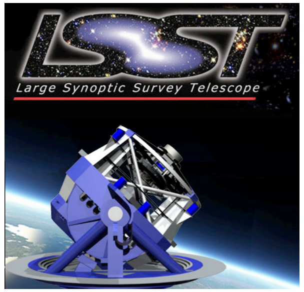
Non-linearity – affects most of the modes in the late universe

Large data sets – observational rather than experimental and often indirect

Systematics – astrophysical "contaminants," instrumental and observational effects

Why machine learning in **cosmology**?

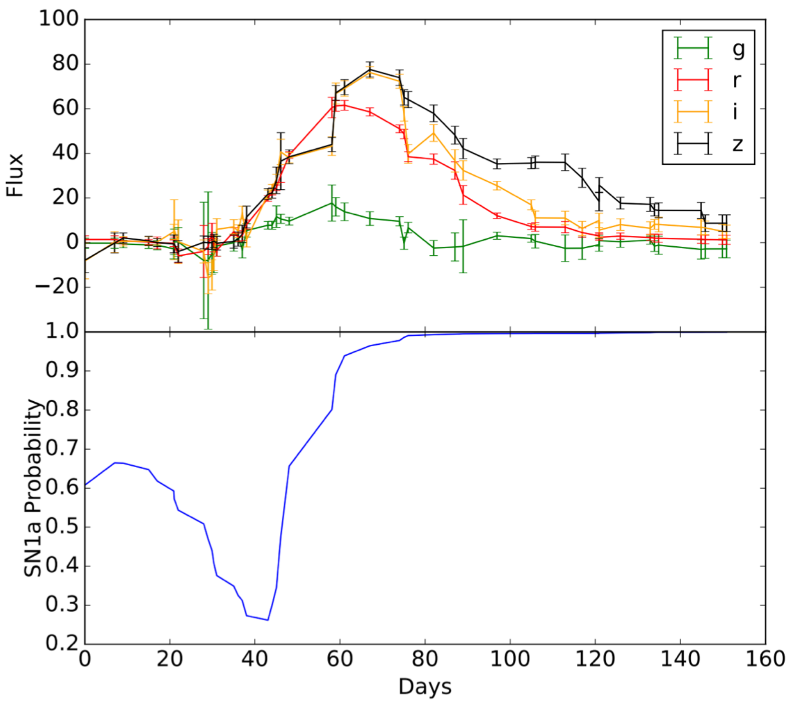
- **Automation**
  - Size of data sets (e.g. LSST: 20 TB per night) means manual intervention is only possible in highly exceptional circumstances. Many data analysis tasks (classification, regression) need to be automated.
- **Acceleration**
  - Machine learning can provide short cuts to costly physical simulations.
- **Superhuman performance**
  - Trained on physical models and data, “emergent” algorithms can sometimes exceed performance of “designed” algorithms.

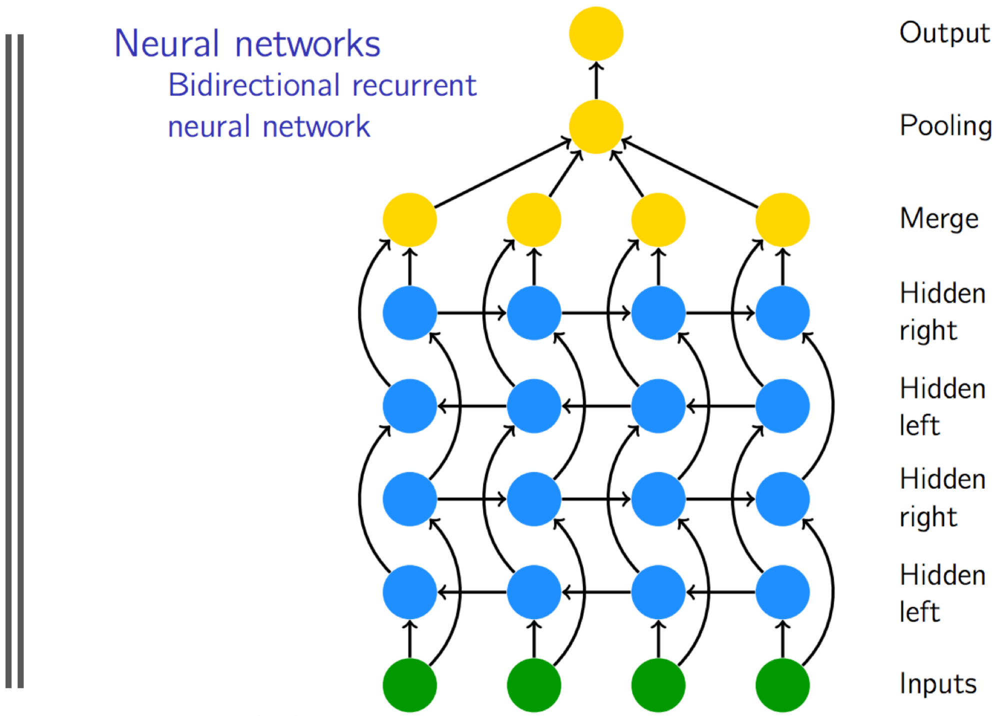


Example: supernova classification

- LSST will detect 10^5-10^6 supernovae (SNe) in 6 photometric filters.
- Type Ia SNe are standard candles, important for cosmology. Other types are not and can confuse analysis.
- Typing is done through expensive spectroscopic follow-up.
- But spectroscopic follow-up is far too costly for LSST.
- Can machine learning identify types?

Supernova classification

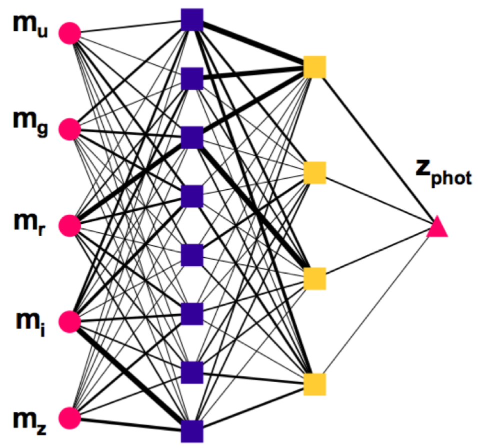


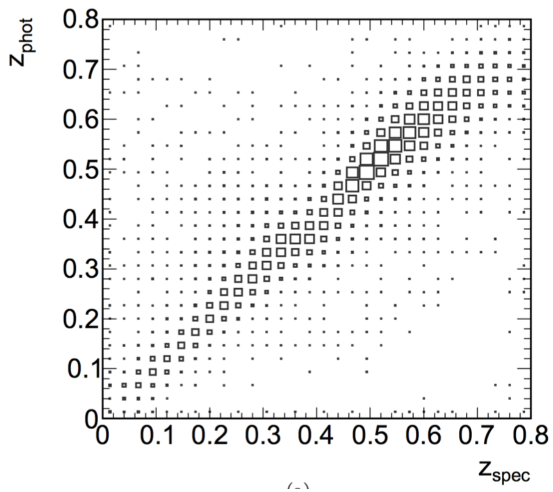


Charnock & Moss (2016), Möller *et al.* (2016)

Photo-z estimation

- Estimate the redshift of a galaxy spectrum due to recession speed of the object from wideband colors.
- First serious application of NN in astronomy.
- ANNz (Collister, Lahav 2004) was best in class. Now superseded by updated versions.
- Intrinsic difficulty: *how to train out-of class data?*

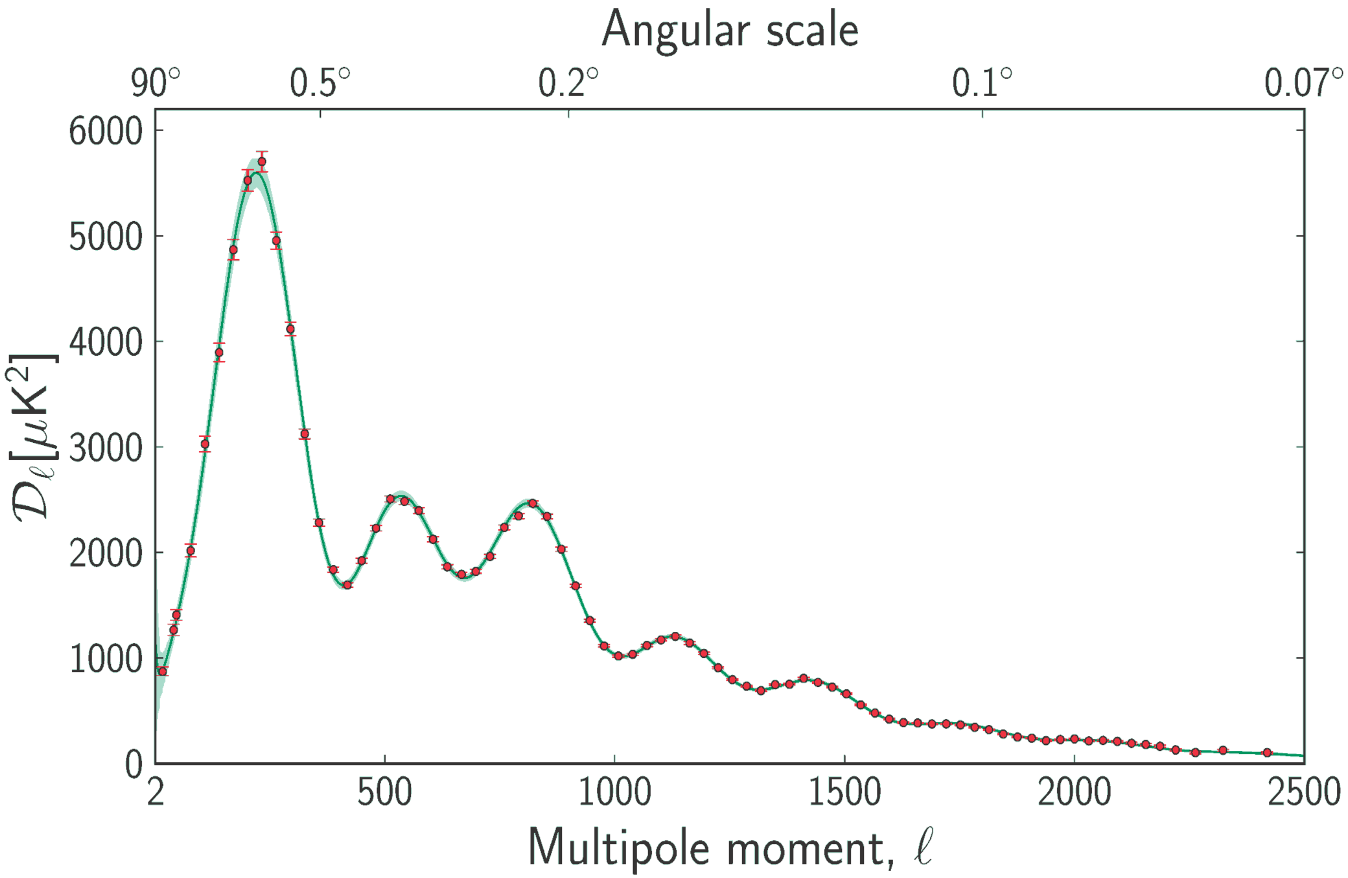




Acceleration

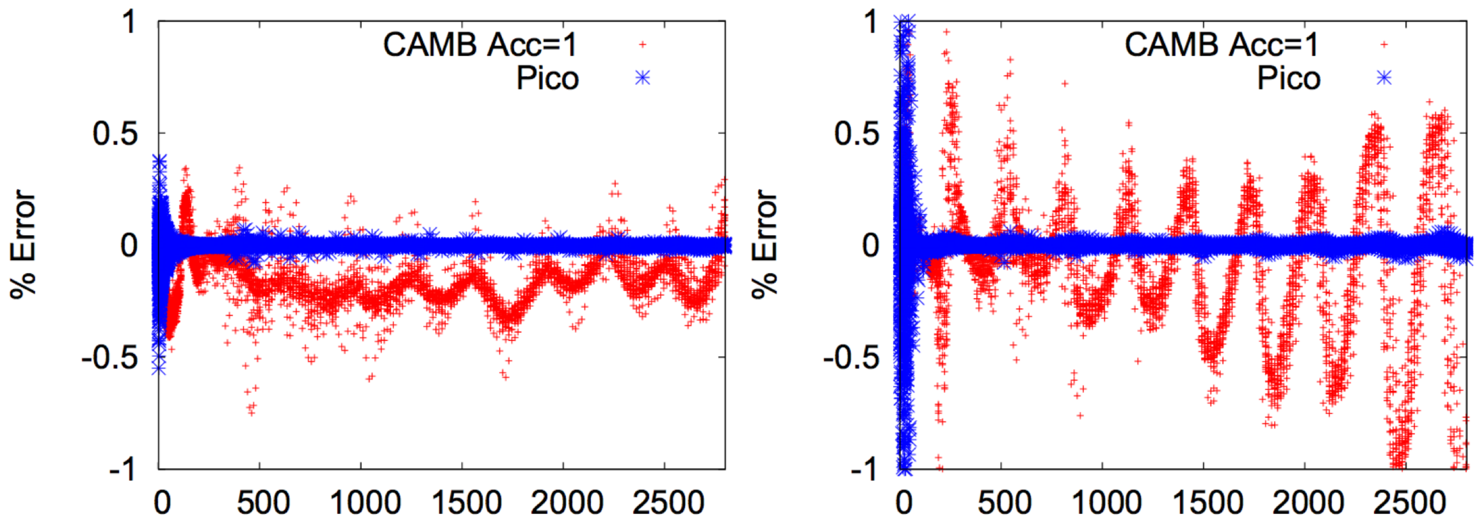
- **Emulators**
- Solve the problem of model predictions that involve costly computations
- First high-accuracy emulator in cosmology: **Pico** (Fendt, Wandelt 2007)
  - Acceleration through Parallel Precomputation and LEarning (APPLE)
  - PCA pre-compression of outputs – then predict compressed quantities as a function of parameters using regression.

Theory confronts Planck data – extreme accuracy required



Need O(10^5) sequential evaluations per analysis (for MCMC.) Up to 1 hour per high precision evaluation. 2 ms for the emulator.

Emulators for Speed and Accuracy (PICO)



High accuracy and 10000x speed-up for Planck-accuracy power spectra

Fendt & Wandelt 2007, 2008

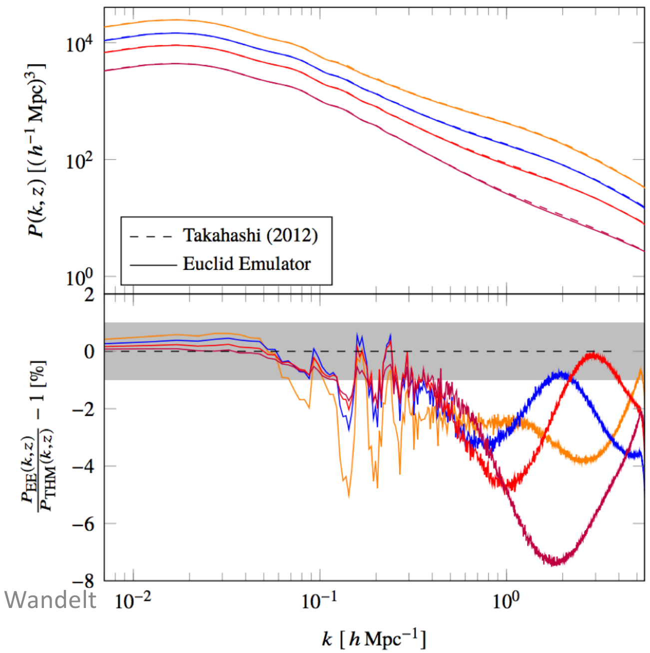
Machine learning ≠ Neural Networks

- PICO does *not* use NNs – actually it's embarrassingly simple!
- Global approach for this problem is much more accurate than locally adaptive techniques.
- Better generalization, because the target function is very smooth.
- Accuracy requirement very high.

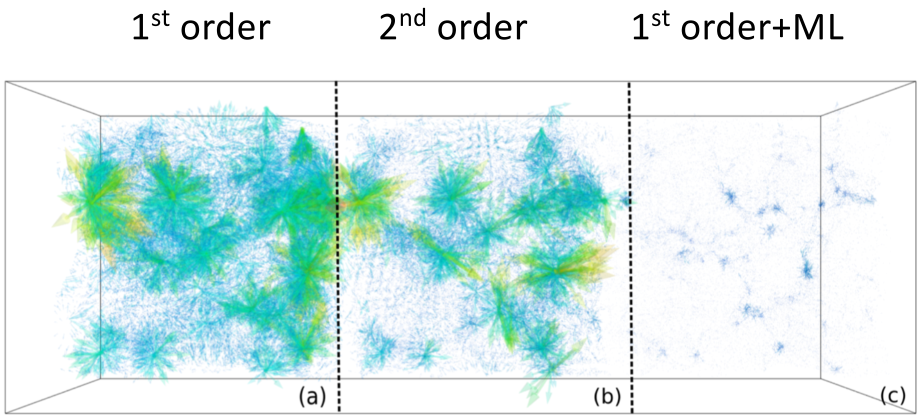
- So let's not forget trees/forests, regression, clustering, Gaussian Processes etc.

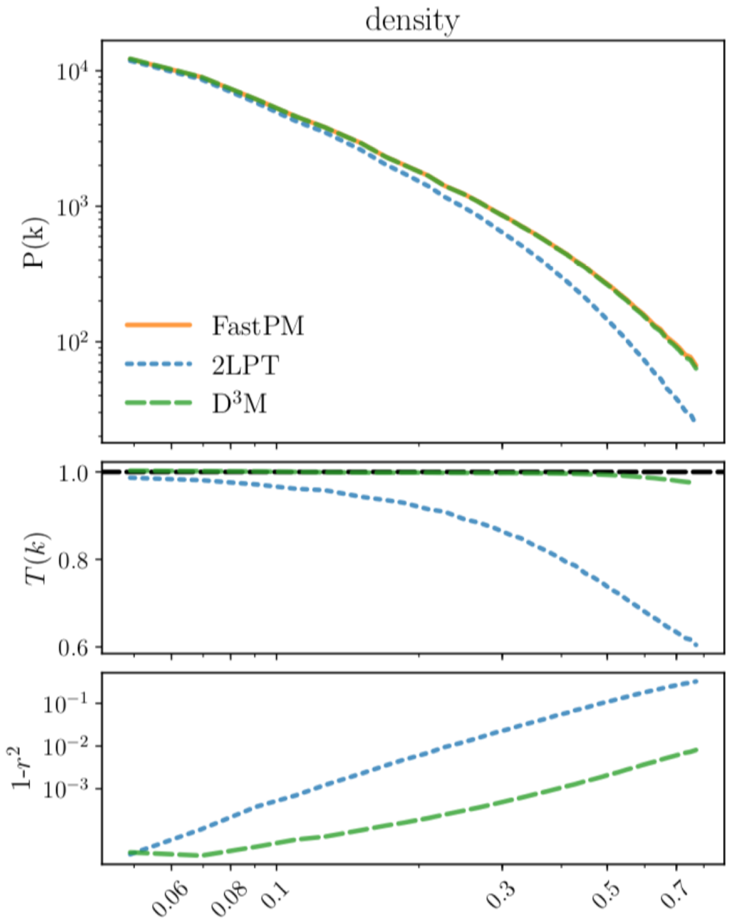
Emulators now

- **Emulators**
- Now training emulators for ESA's Euclid space mission using essentially the same technique. *(Euclid Collab., Knabenhans et al., arXiv: 1809.0469)*
- Emulating non-linear correction (*boost*) to spatial two-point correlations.



Machine learning cosmological physics



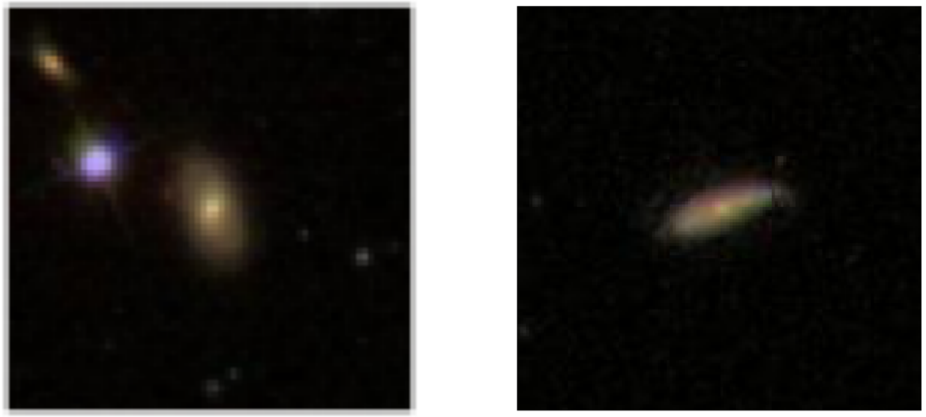


- He et al. (arXiv: 1811.06533)

- Learning non-linear correction to particle displacement (U-net)
- Find some ability to generalize beyond trained parameters

Generative Adversarial Networks to simulate galaxy images

Real or fake?



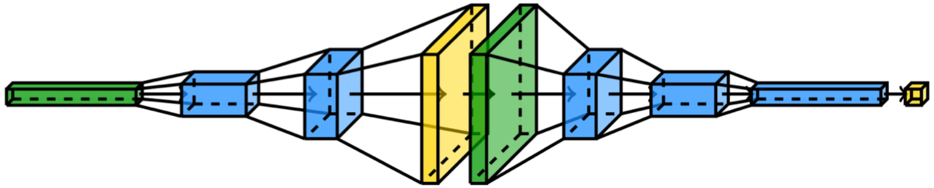
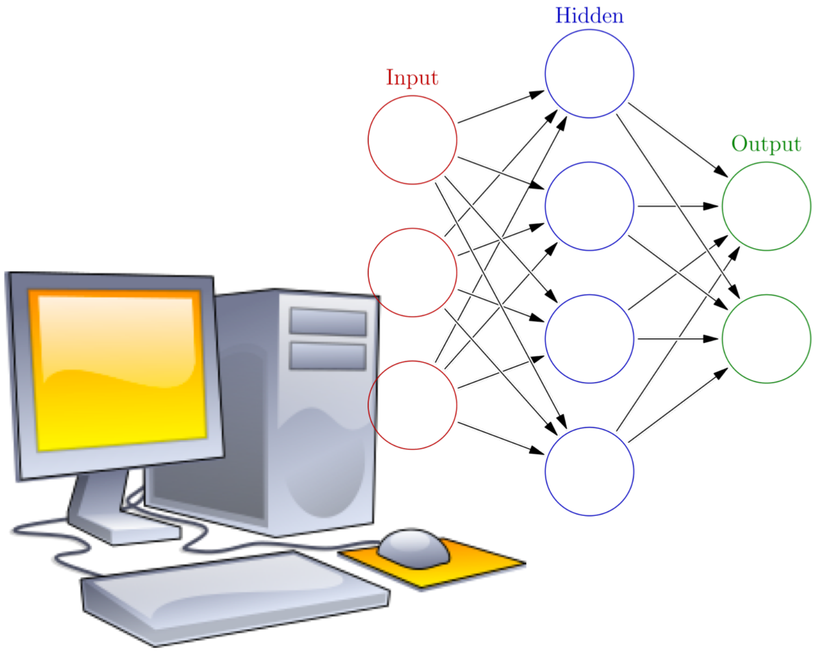
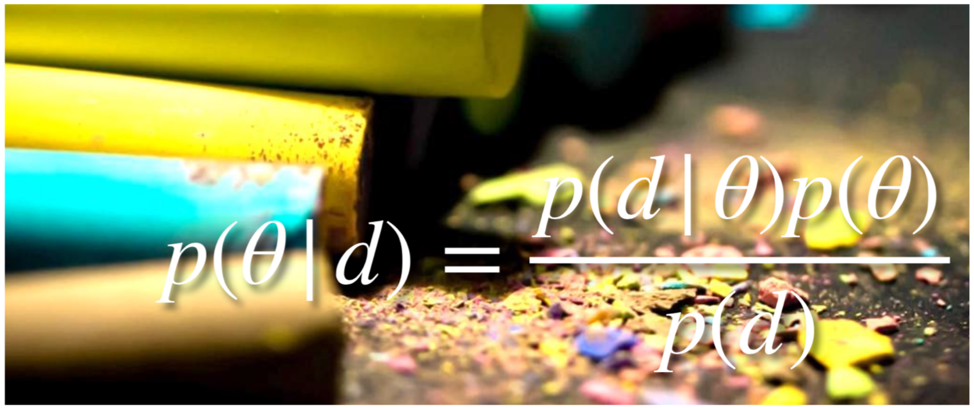


Image credit: T. Charnock

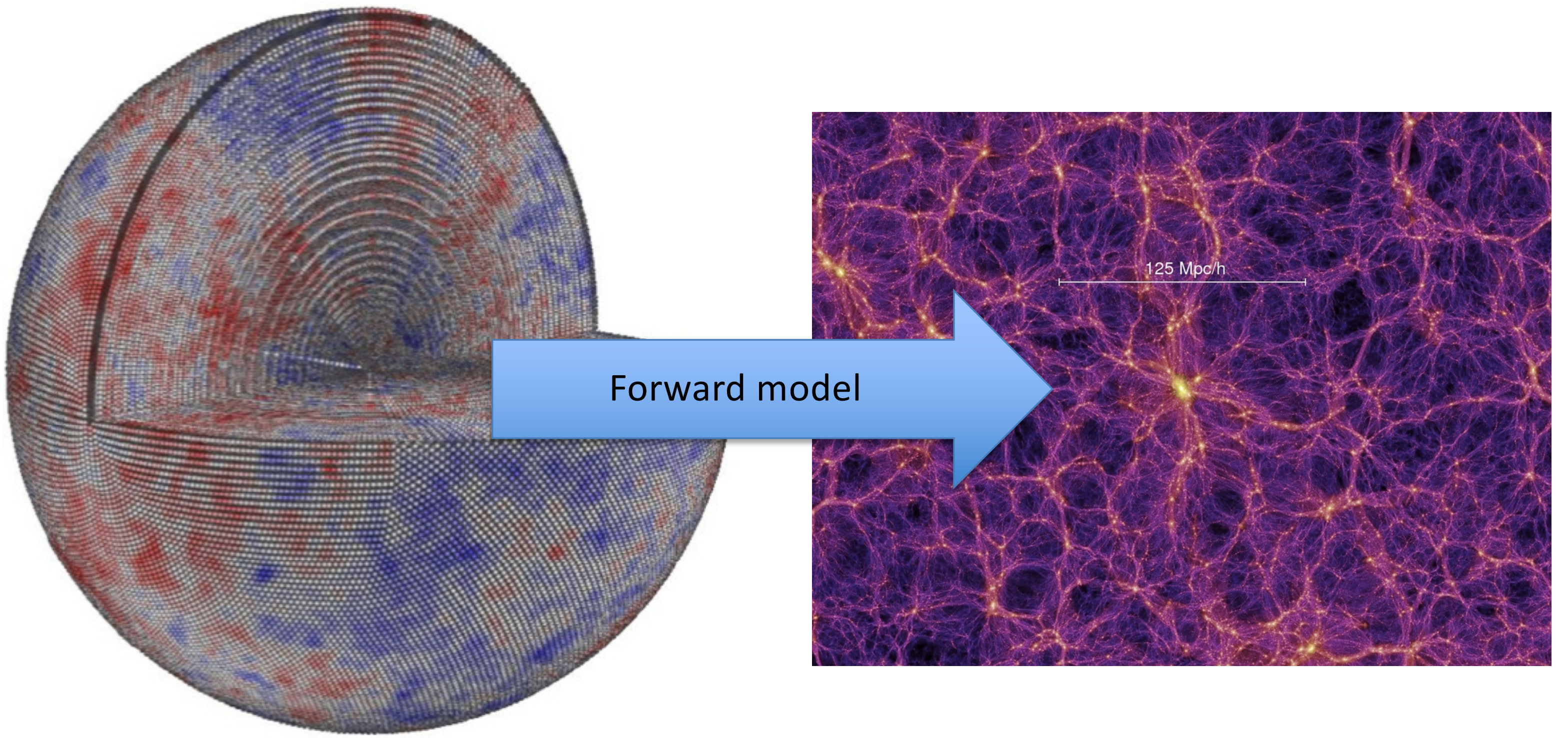
Machine learning vs Physics?



$$W = \int_{k<\Lambda} [Dg][DA][D\psi][D\Phi] \exp\left\{ i \int d^4x \sqrt{-g} \left[\frac{m_p^2}{2} R \right.\right.$$
$$\left.\left. - \frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} + i\bar{\psi}^i \gamma^\mu D_\mu \psi^i + \left(\bar{\psi}^i_L V_{ij} \Phi \psi^j_R + \text{h.c.} \right) - |D_\mu \Phi|^2 - V(\Phi) \right] \right\}$$



Physics-based machine learning



Initial conditions of the universe

The universe today

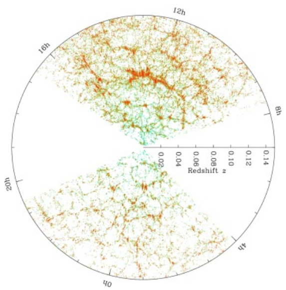
A fully generative *probabilistic* model of galaxy surveys with $O(10^7)$ parameters

BORG: *Bayesian Origin Reconstruction from Galaxies*

- Gaussian prior + **Gravity** + likelihood for galaxies

  (includes survey model, bias model, automatic noise level calibration, selection function, mask, …)

- Hamiltonian Markov Chain Monte Carlo in $O(10^7)$-D



Observations

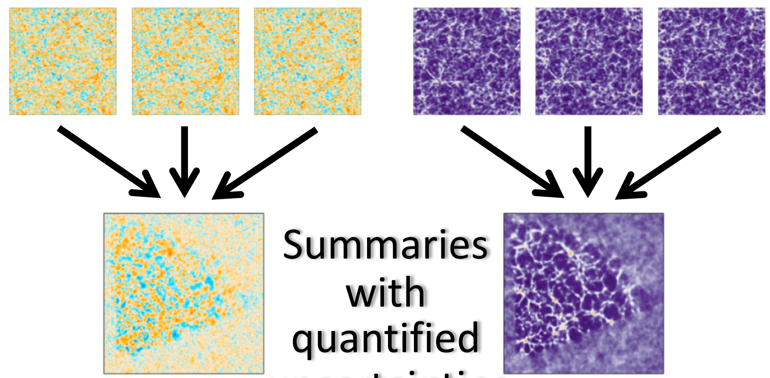
(galaxy catalog + meta-data: selection functions, completeness…)

→ BORG →

E.g. inferred dark matter densities

z=100

z=0



Summaries with quantified uncertainties

Jasche & Wandelt 2013, arXiv:1203.3639

Jasche, Leclercq & Wandelt 2015, arXiv:1409.6308

A fully generative *probabilistic* model of galaxy surveys with $O(10^7)$ parameters

BORG: *Bayesian Origin Reconstruction from Galaxies*

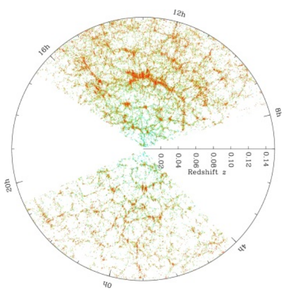
- Gaussian prior + **Gravity** + likelihood for galaxies

(includes survey model, bias
selection function, mask, …)

- Hamiltonian Mark

No NNs here – except for gravity: 20-layer 3D NN respecting symplectic Hamiltonian flow, energy and mass conservation…

… i.e., an N-body code with ~20 time steps

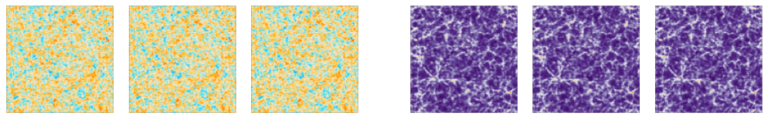


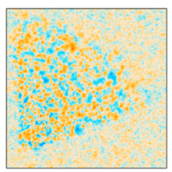
Observations

(galaxy catalog + meta-data: selection functions, completeness…)

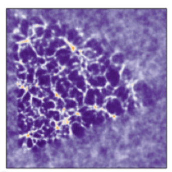
→ BORG →

z=



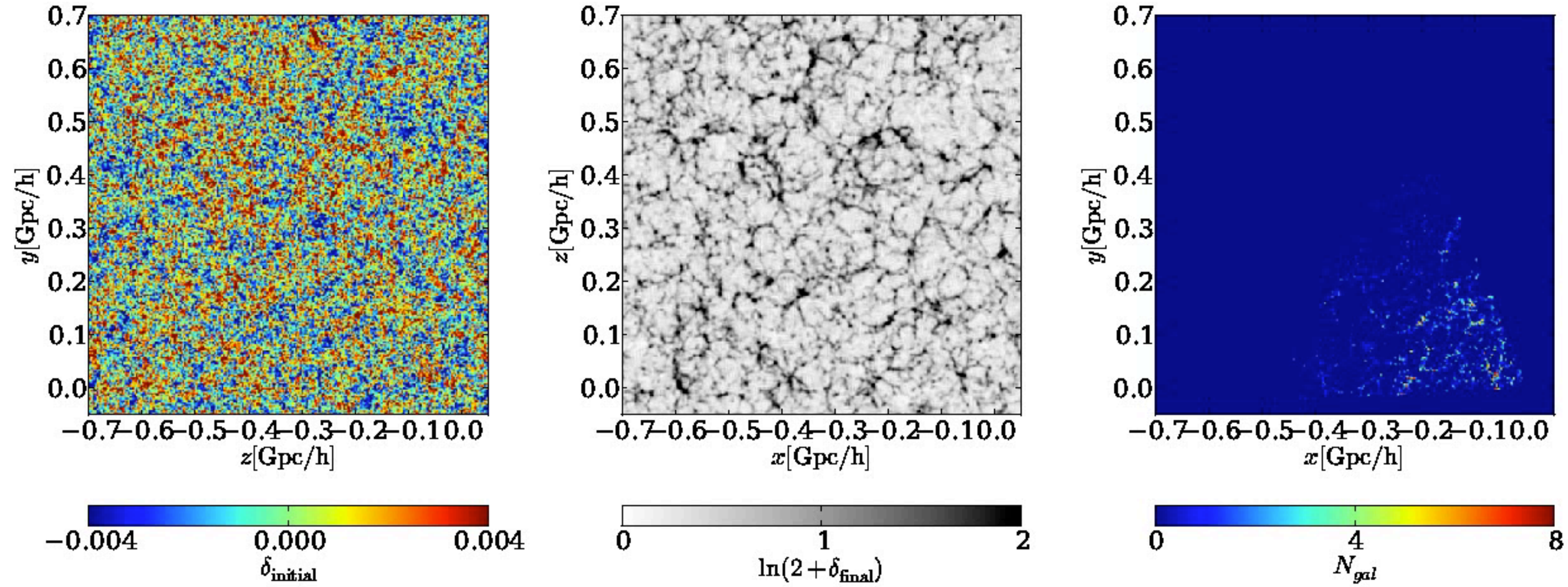


Summaries with quantified uncertainties



Jasche & Wandelt 2013, arXiv:1203.3639
Jasche, Leclercq & Wandelt 2015, arXiv:1409.6308

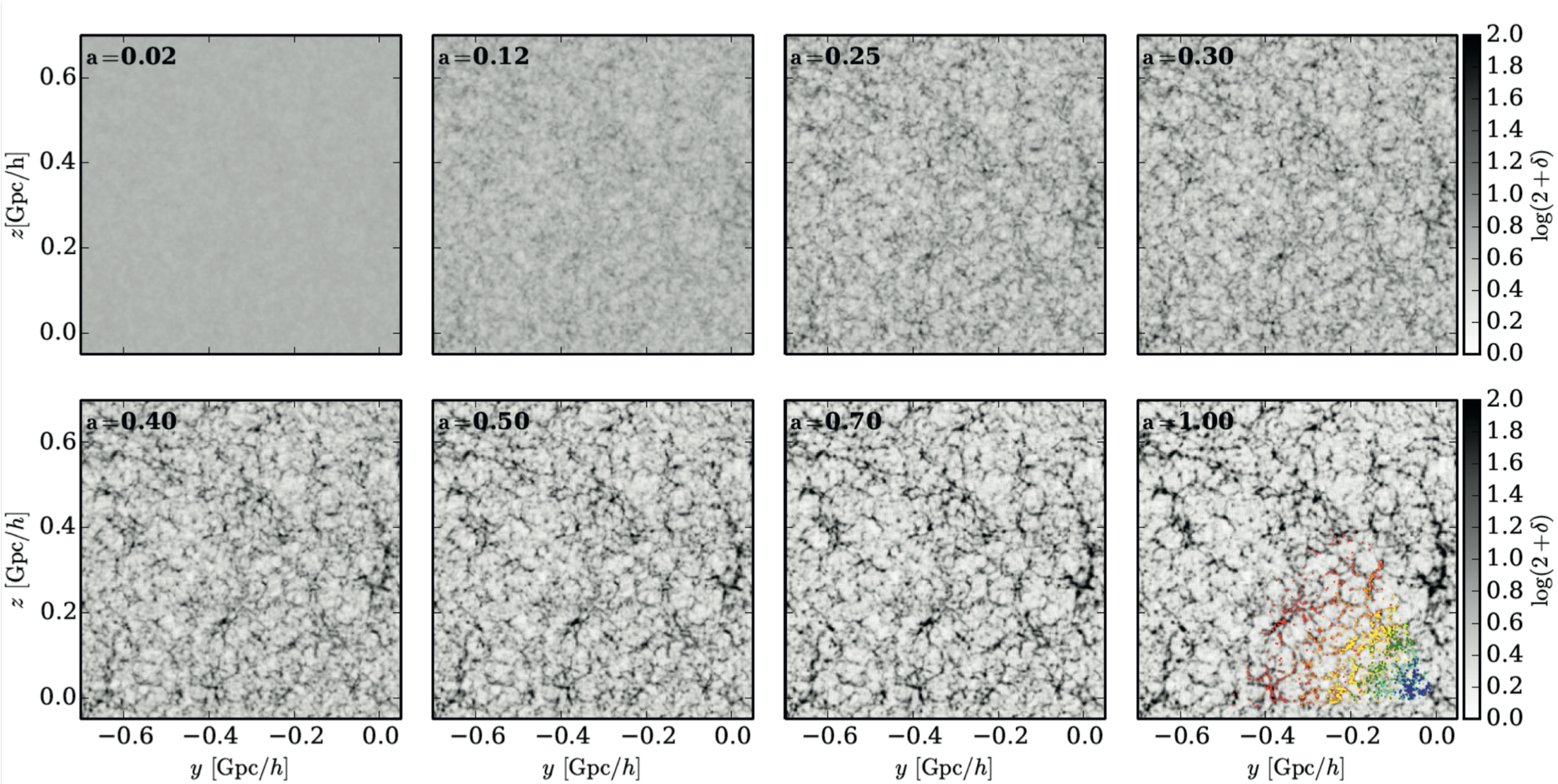
Bayesian LSS sampling – the movie



Initial conditions Final conditions Observations

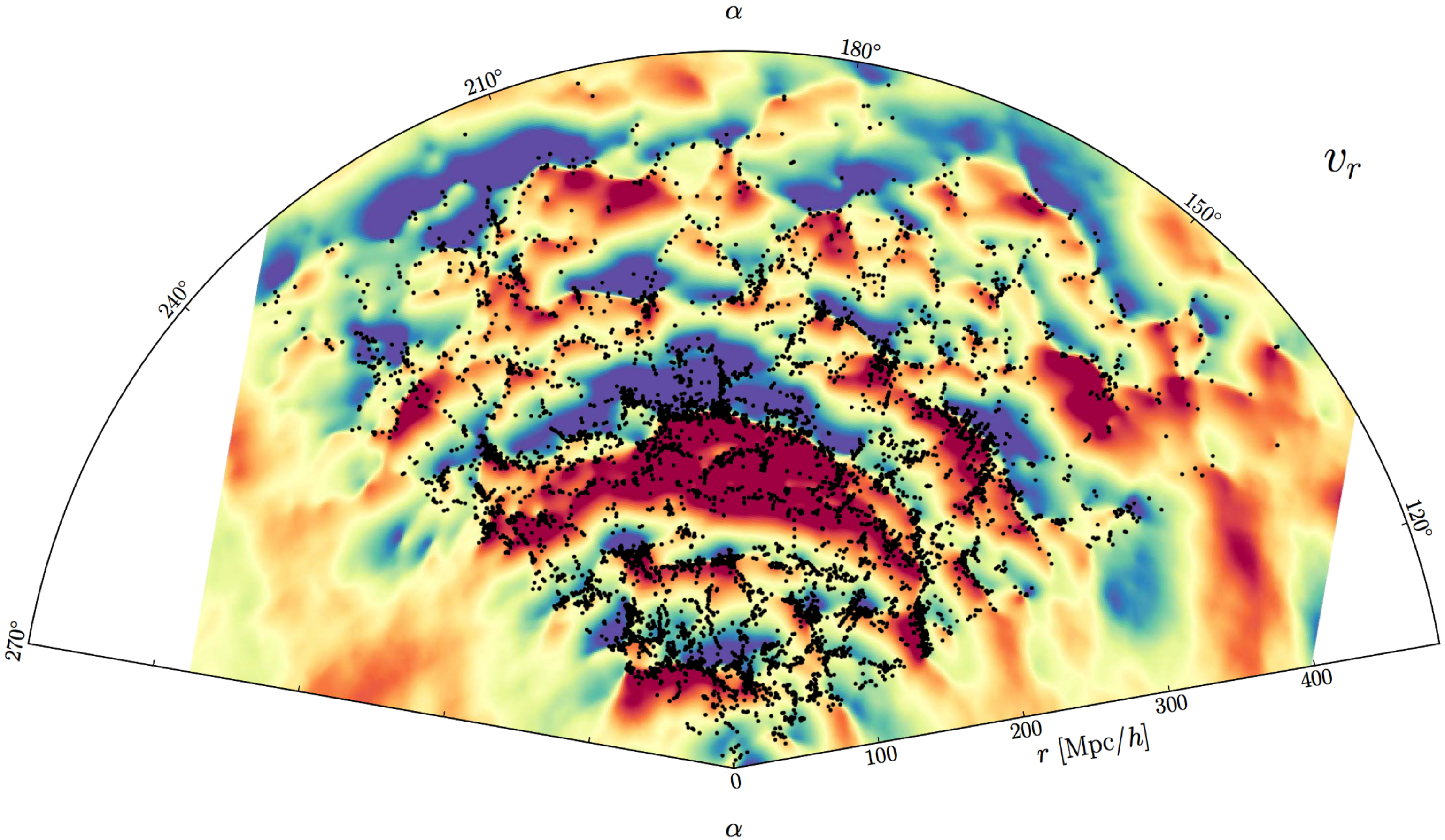
Jasche, Leclercq & Wandelt 2014, arXiv:1409.6308

A posterior sample of the formation history of our Universe



Jasche, Leclercq & BDW 2014, arXiv:1409.6308

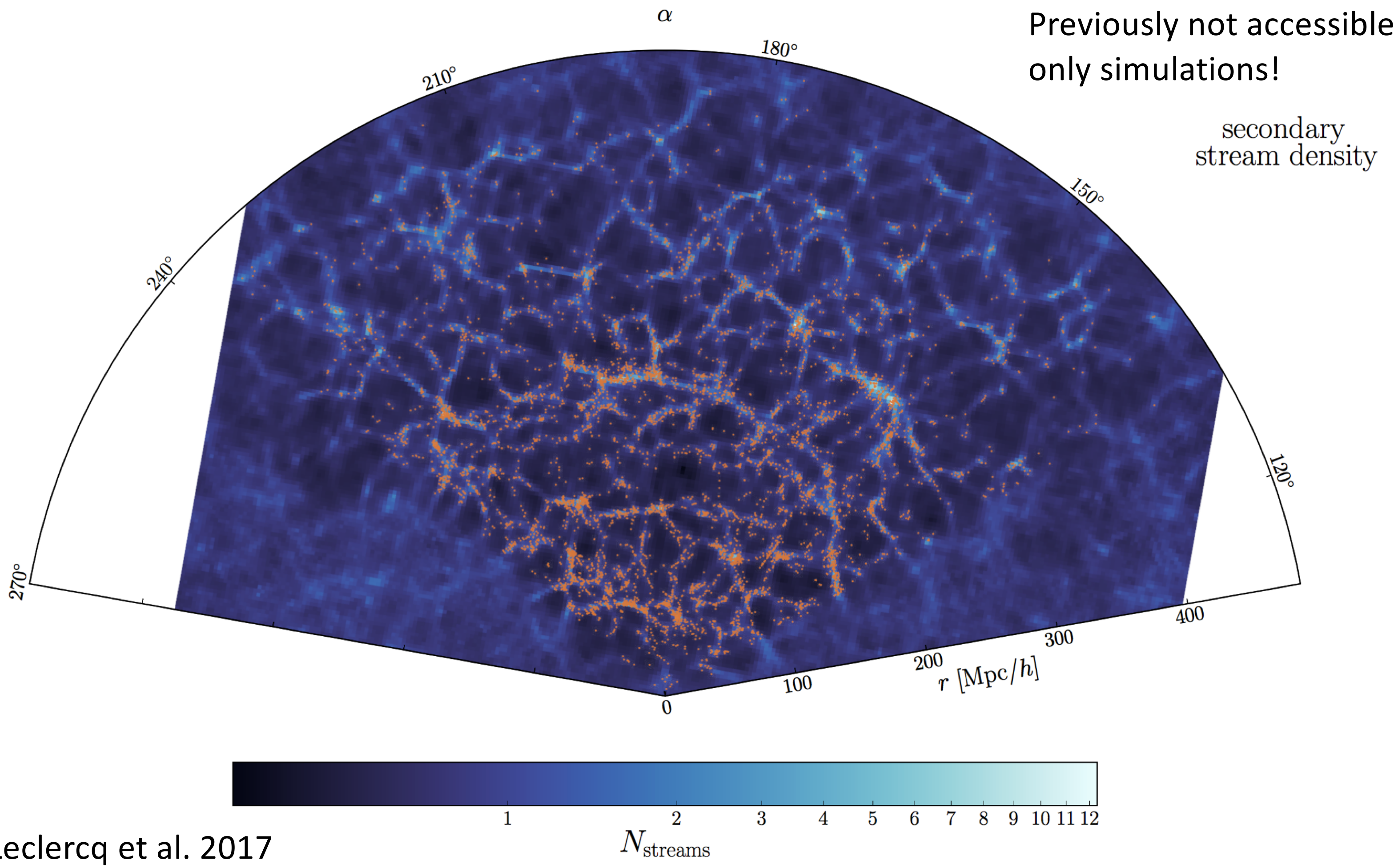
Example Bayesian LCDM predictions: dynamical velocities



Leclercq et al. 2017

Posterior mean of Lagrangian stream density

Previously not accessible in data, only simulations!



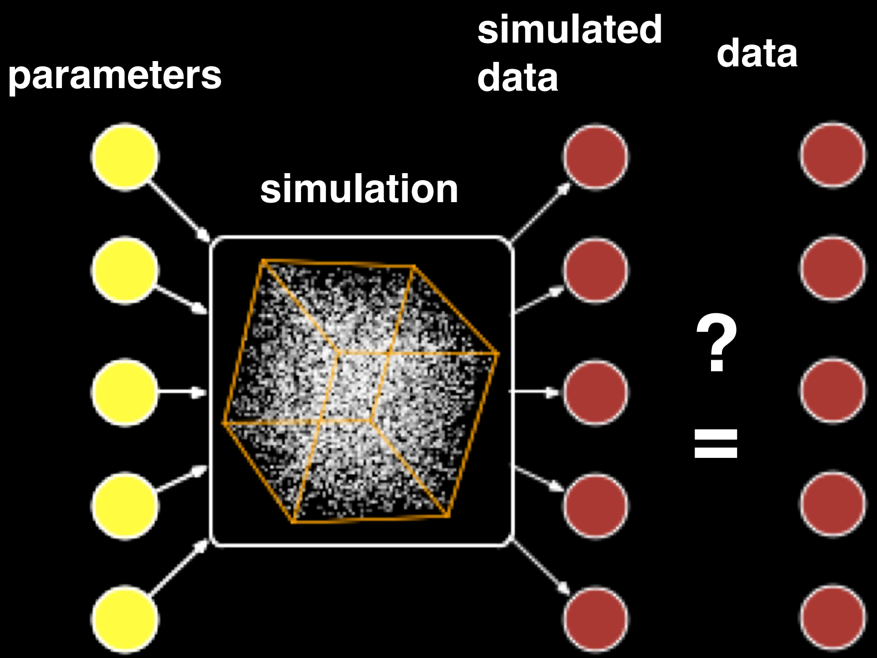
Leclercq et al. 2017

What if we can <u>only</u> do simulations?

$$P(\boldsymbol{\theta}|\mathbf{d}) = \frac{P(\mathbf{d}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{d})}$$

$$\mathbf{d}^* \leftarrow \mathrm{simulation}(\mathbf{d}^*|\boldsymbol{\theta})$$

Likelihood-free inference 101



Draw from prior:

$$\boldsymbol{\theta} \leftarrow P(\boldsymbol{\theta})$$

Simulate data:

$$\mathbf{d}^* \leftarrow P(\mathbf{d}^*|\boldsymbol{\theta})$$

If $\rho(\mathbf{d}^*, \mathbf{d}) < \epsilon$

accept;

else:

reject;

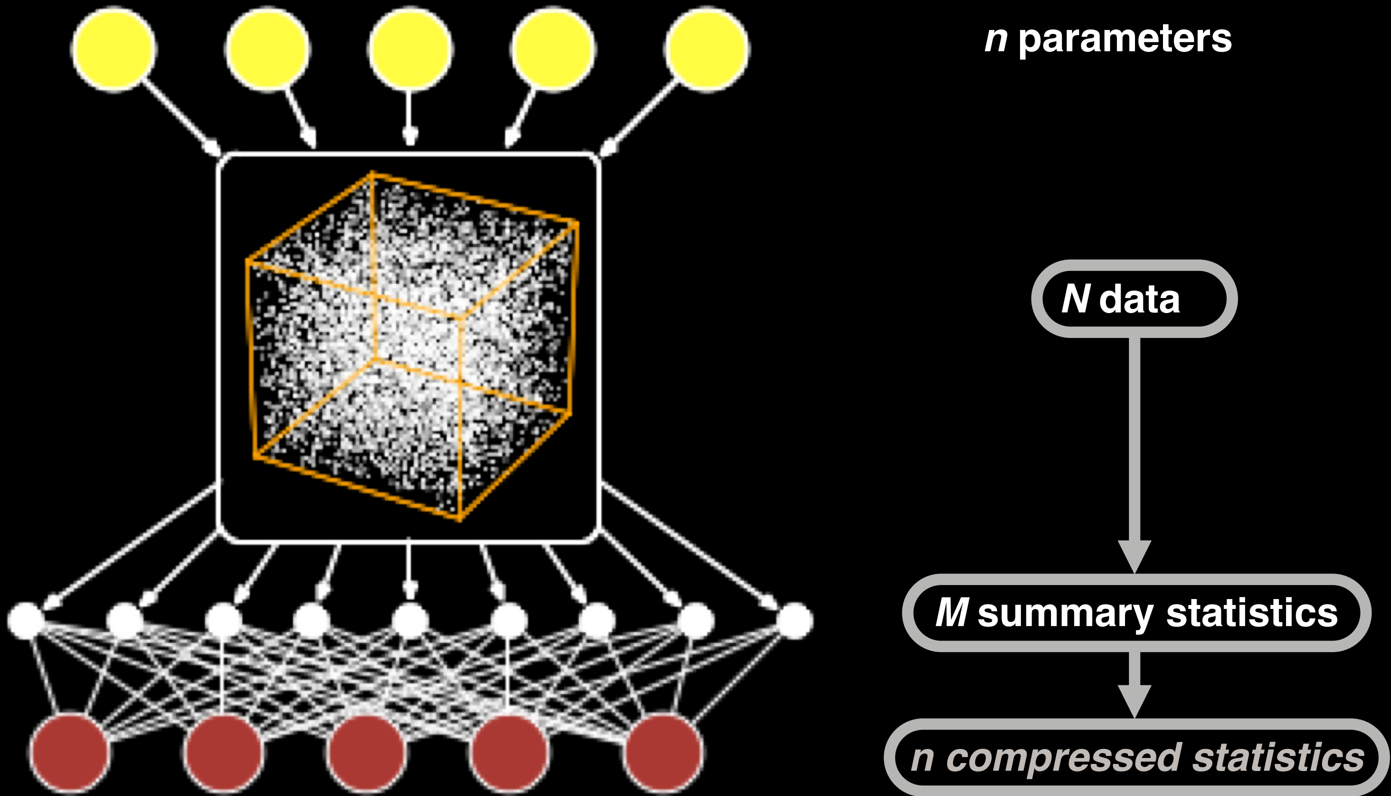
In the limit $\epsilon \rightarrow 0, \; \{\boldsymbol{\theta}\} \leftarrow P(\boldsymbol{\theta}|\mathbf{d})$

Likelihood-free inference 101

How to reduce data-space?

How to explore parameter-space?

Reducing data space: massive data compression



Massive data compression

Fisher information

$$\mathbf{F} \equiv -\mathbb{E}_{\boldsymbol{\theta}}(\nabla\nabla^T\mathcal{L})$$

Information inequality

$$\mathbb{V}_{\boldsymbol{\theta}}(t_\alpha) \geq [\nabla\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t})^T\mathbf{F}^{-1}\nabla\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t})]_{\alpha\alpha}$$

Can derive *n* compressed quantities that contain all the Fisher information!

Alsing & Wandelt arXiv:1712.00012

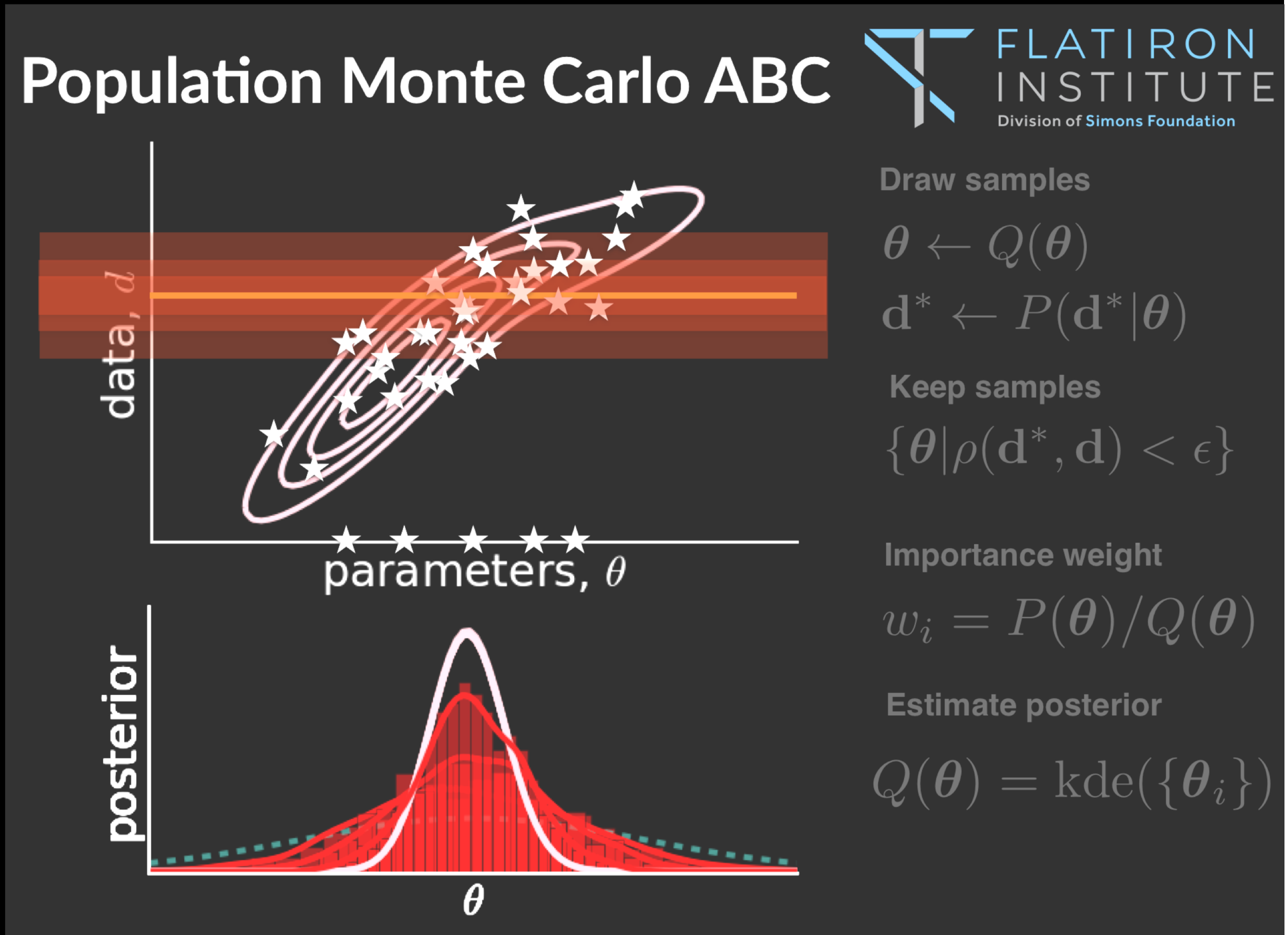
Exploration of parameter space

Density estimation Likelihood free inference (DELFI)

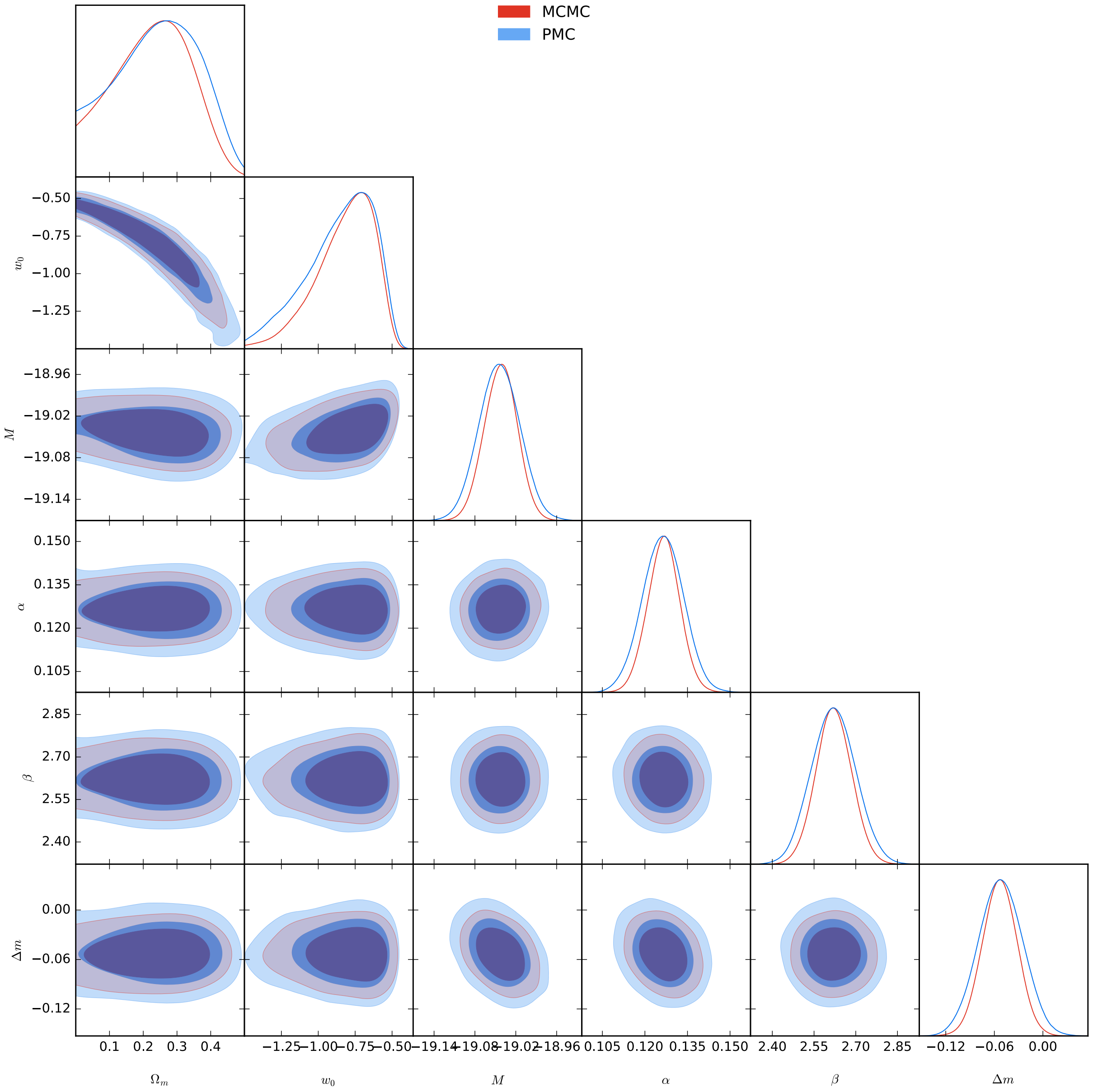
Learn *joint* probability density of parameters and compressed data using a Gaussian Mixture Model

Alsing, Feeney &Wandelt arXiv: 1801.01497

Exploration of parameter space (I)



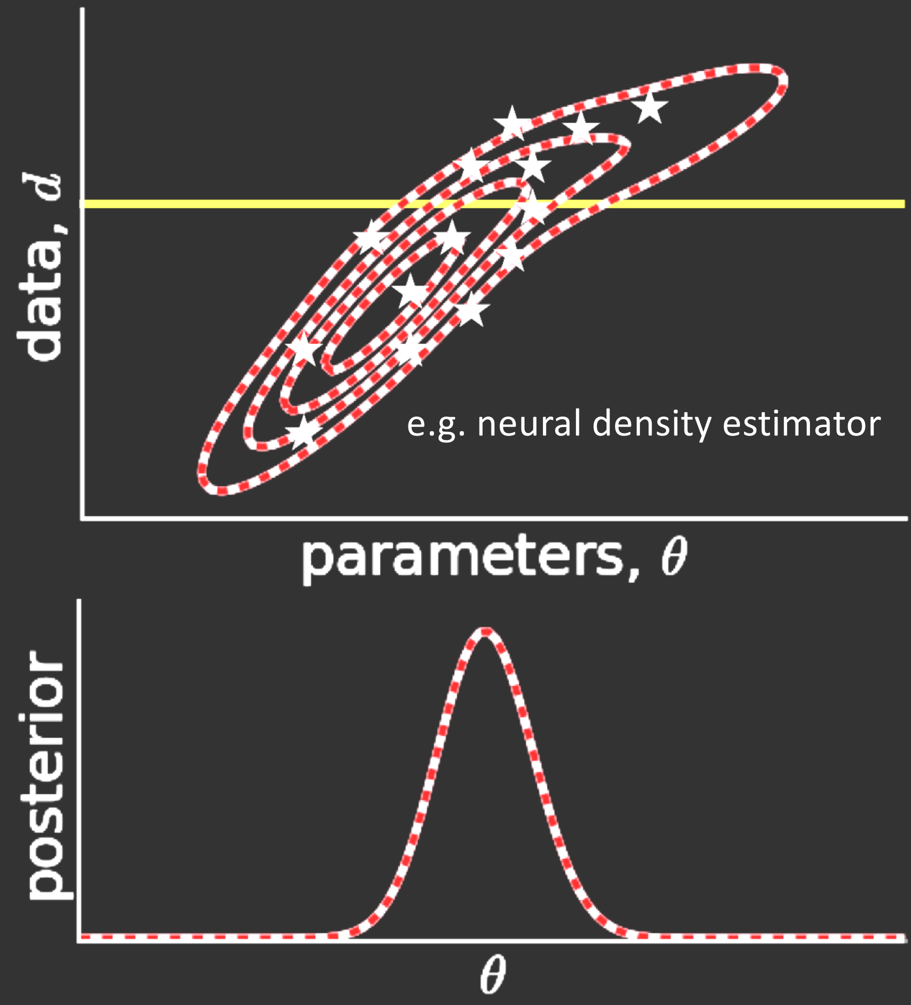
PMC Posterior inference



(>3 million simulations)

Exploration of parameter space (II)

Bayesian density estimation



Draw samples

$$\boldsymbol{\theta} \leftarrow P(\boldsymbol{\theta})$$

$$\mathbf{d}^* \leftarrow P(\mathbf{d}^*|\boldsymbol{\theta})$$

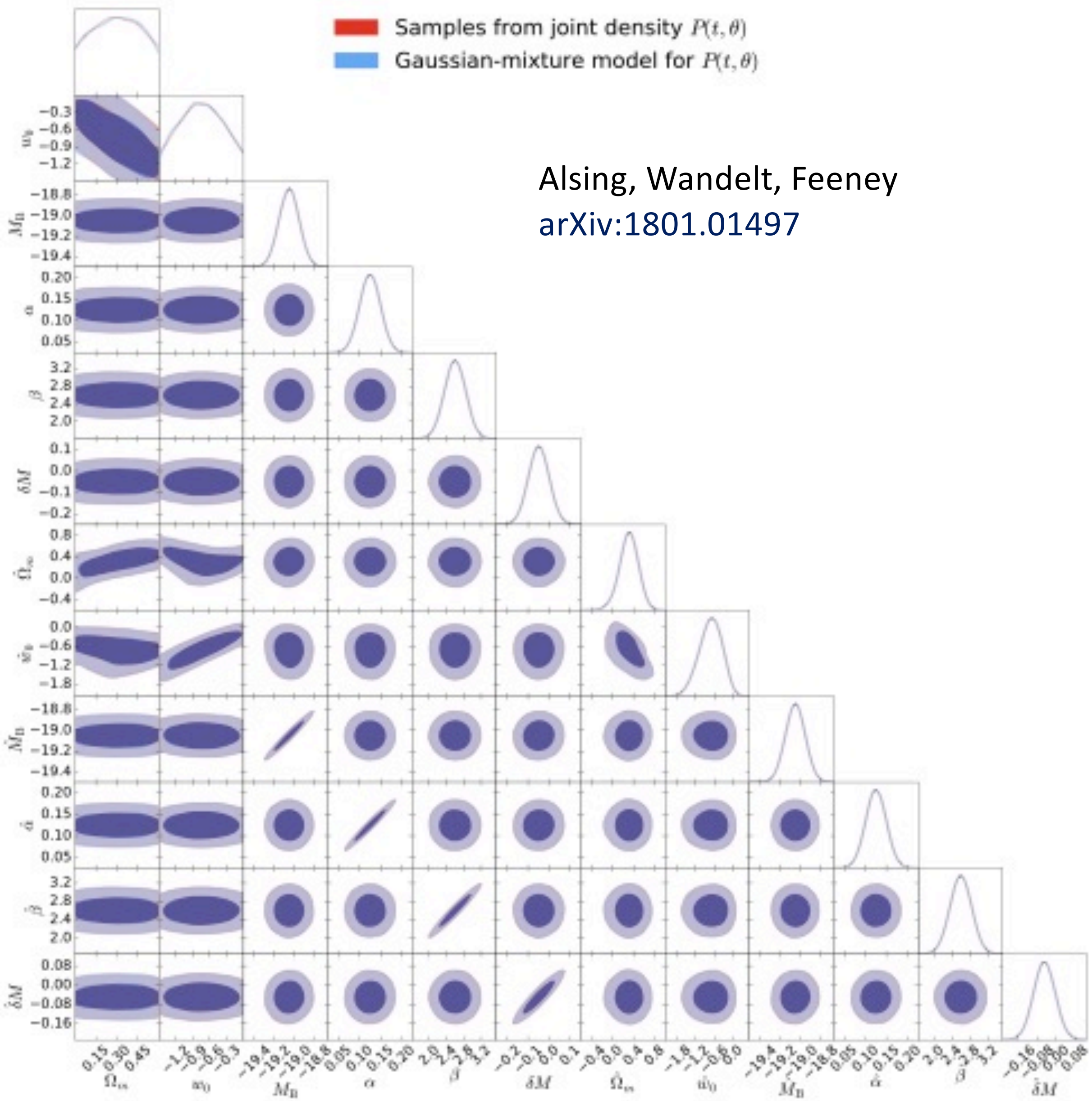
Fit model

$$P(\boldsymbol{\theta}, \mathbf{d}^*) = P(\boldsymbol{\theta}, \mathbf{d}^*; \boldsymbol{\eta})$$

Take slice

$$P(\boldsymbol{\theta}|\mathbf{d}) = P(\boldsymbol{\theta}, \mathbf{d}^* = \mathbf{d})$$

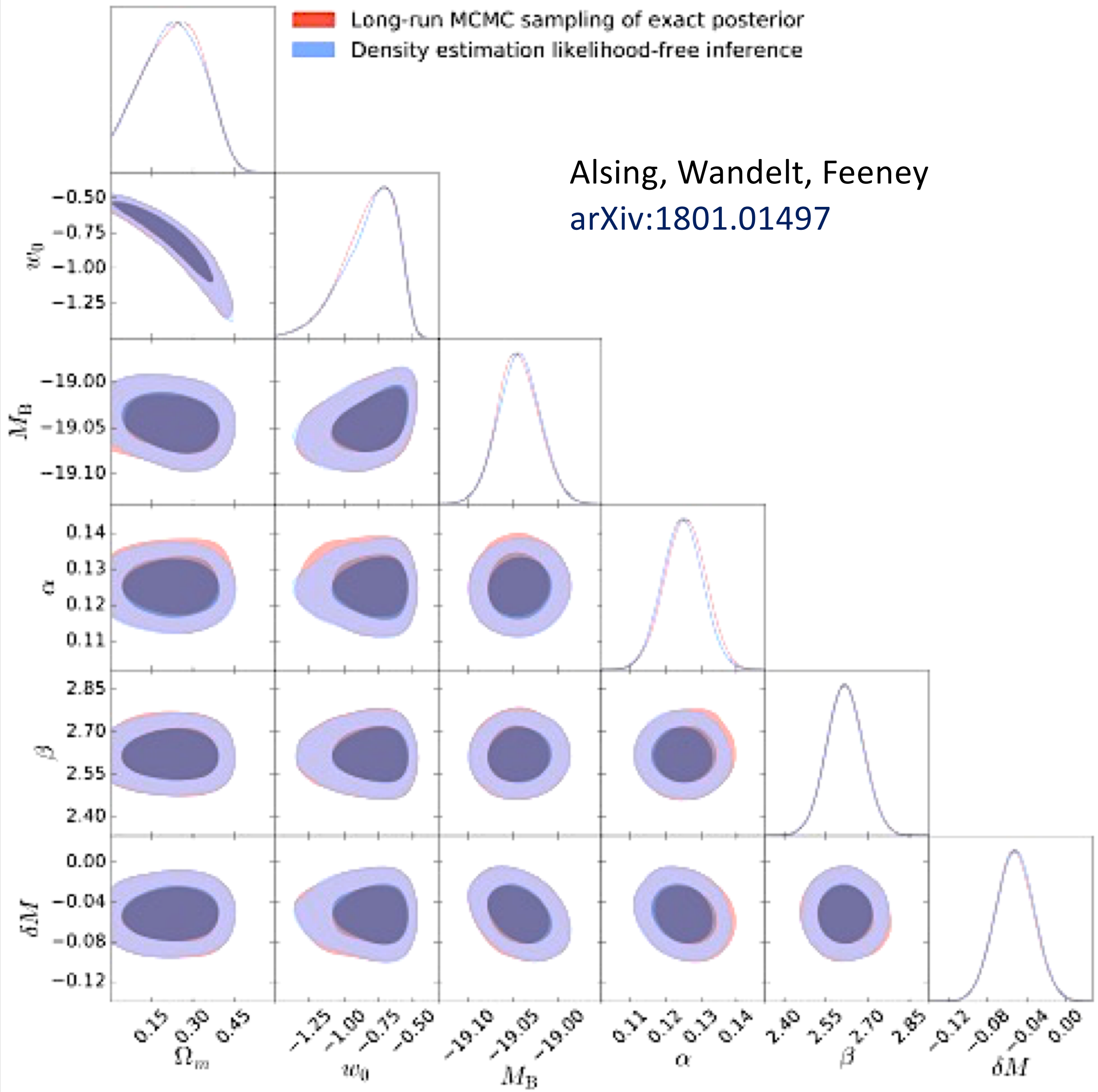
Fit to joint density $P(\boldsymbol{\theta}, \mathrm{d}^*)$



Alsing, Wandelt, Feeney
arXiv:1801.01497

(O(1000) simulations)

DELFI Posterior inference



Alsing, Wandelt, Feeney
arXiv:1801.01497

(O(1000) simulations)

But what if you don’t know how to compute informative summaries of your data?

Automatic Physical Inference with Information Maximizing Neural Networks

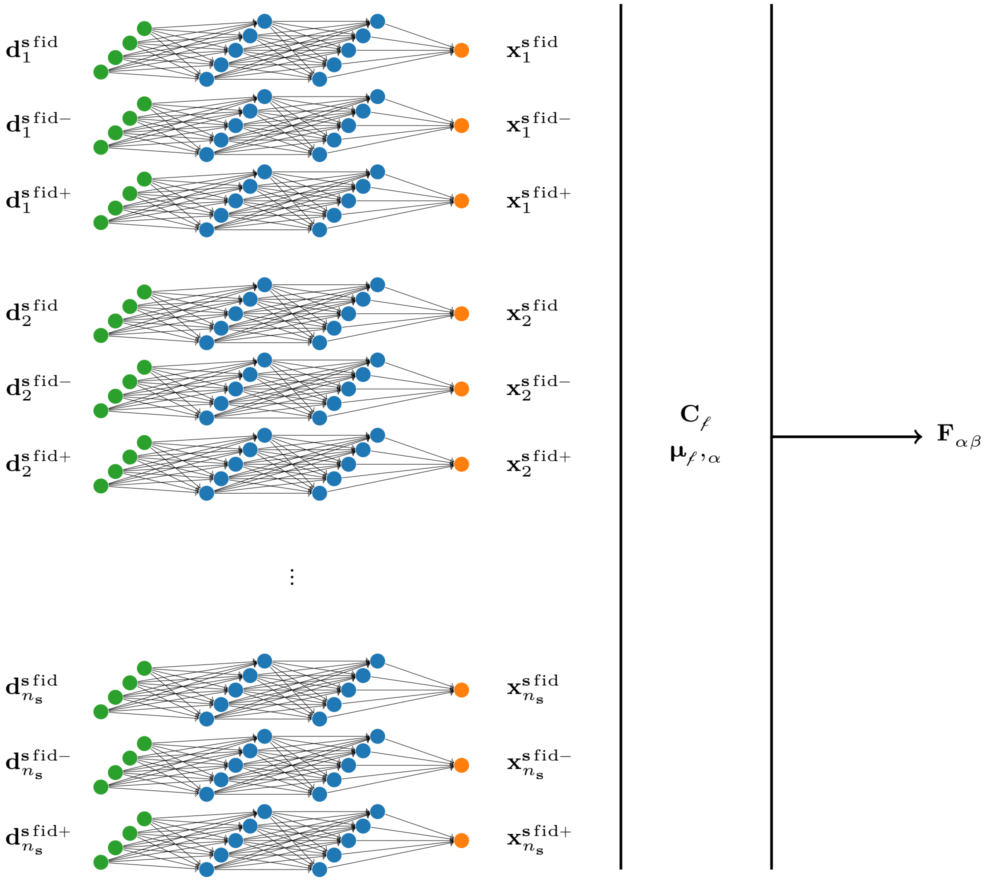
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

- Goal: obviate the need to "guess" heuristic, informative summaries of the data
- Setup: design a neural network that maps the data into a small set of informative *summaries*.
- The loss is (– the Fisher information) under an assumed simple likelihood for the summaries.

- Training uses physical simulations of the model to maximize the information in the summaries about the parameters of the model.
- The achieved loss on a test set is meaningful – it's the information content of the data.

- Note: paper in very similar spirit to *Brehmer et al. arXiv: 1805.12244*

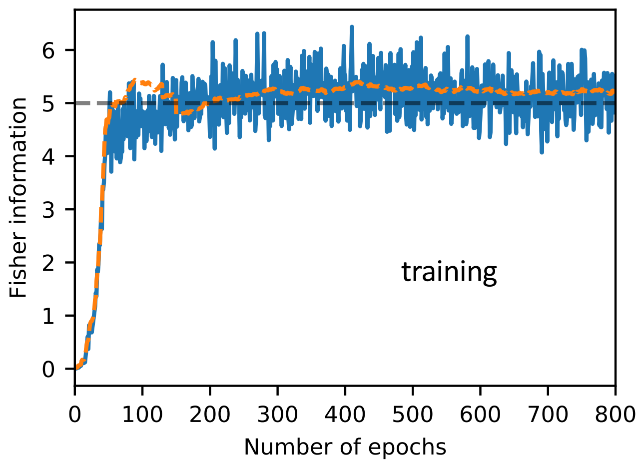
Information maximizing neural network

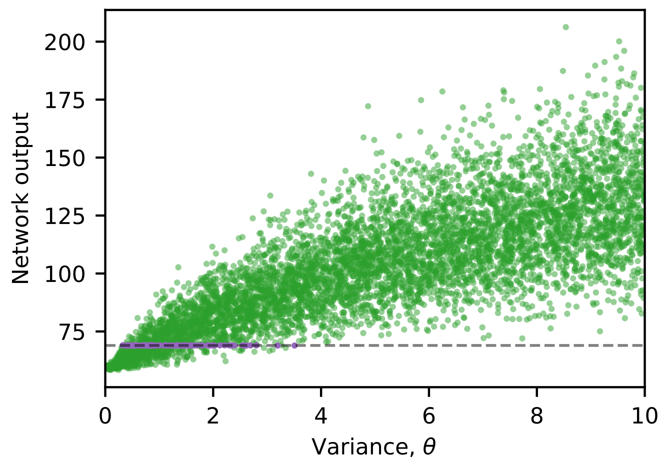


Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 1: inference of variance

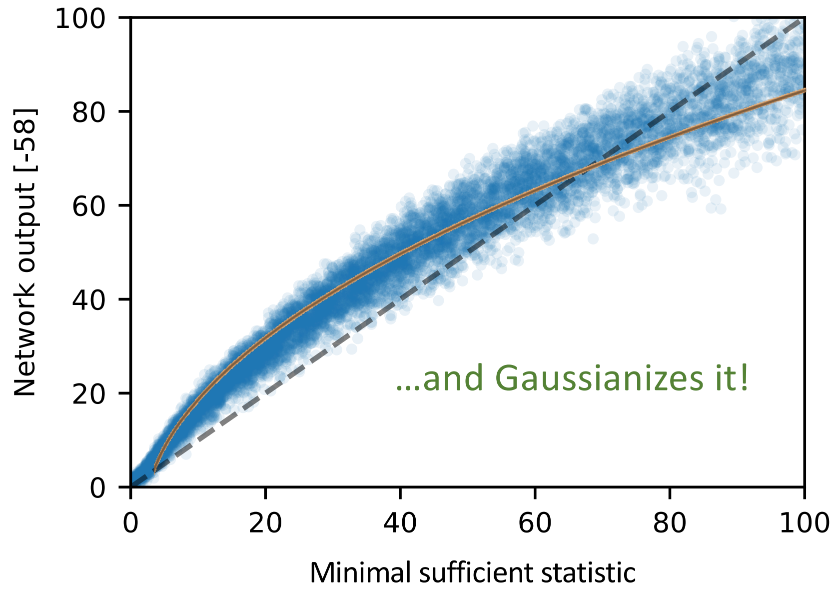
- Perfect information gives $|F| = 5$ in this problem
- Any linear summary gives $|\mathrm{F}| = 0.5$





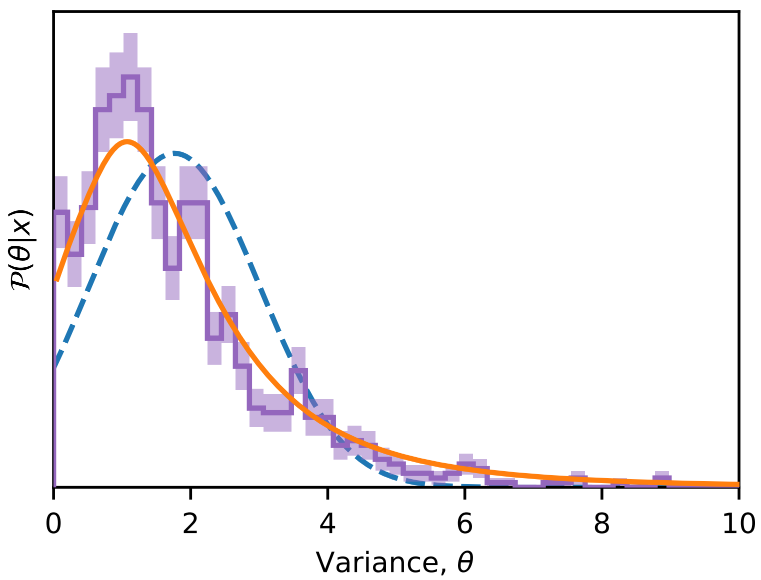
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

The IMNN finds a minimal sufficient statistic for this inference problem



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

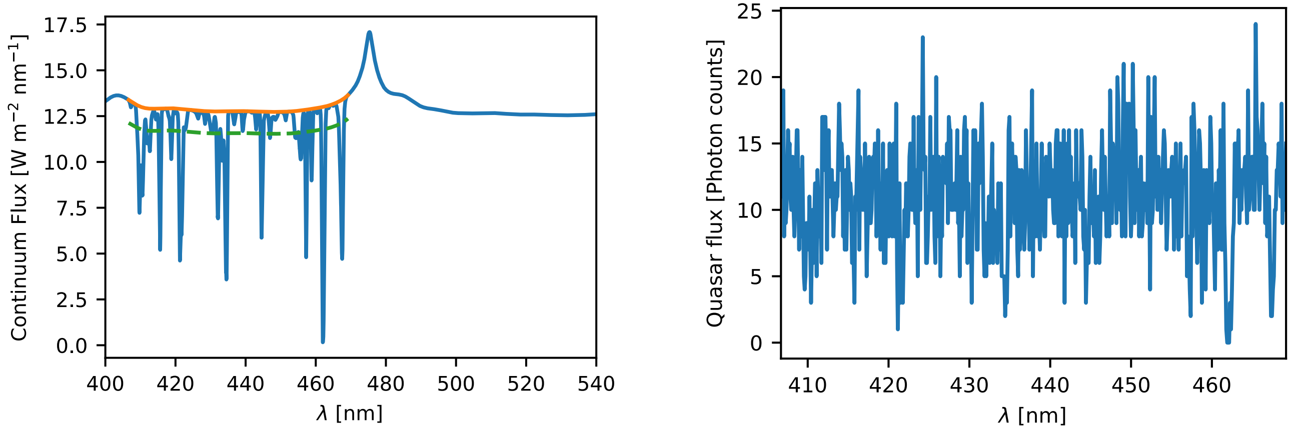
Example 2: Automatic physical inference with unknown noise



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

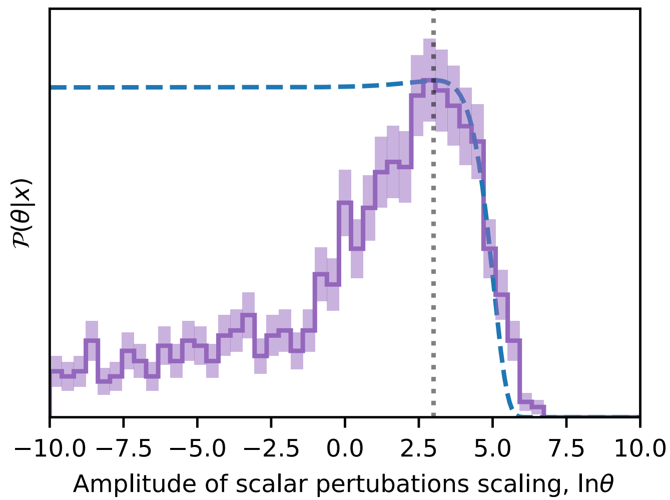
Example 3: Lyman-α forest inference

- The idea is to infer the variance of the underlying density field from a non-linearly transformed, photon-noise dominated Lyman-α forest spectrum



Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 3: Lyman-α forest inference

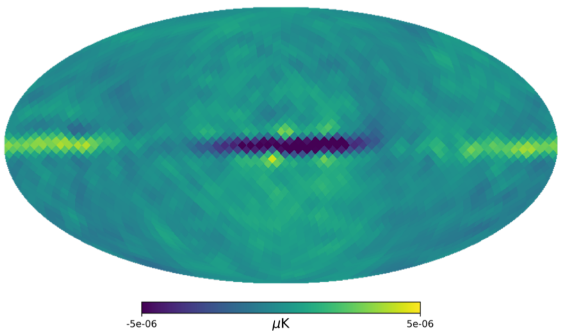


Charnock, Lavaux, Wandelt (arXiv:1802:03537)

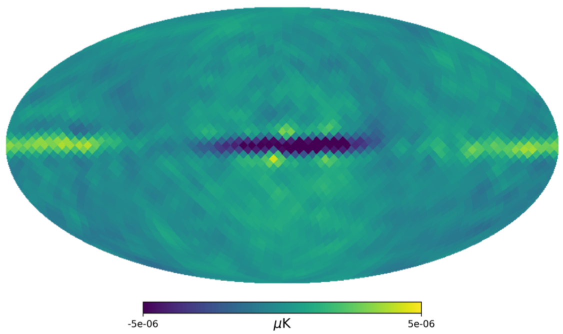
How transparent is our Universe?

planck

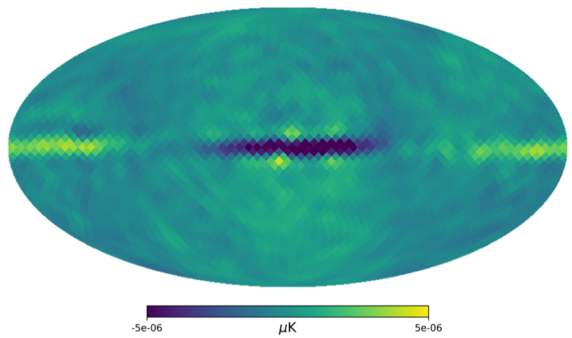
$\tau = 0.5$



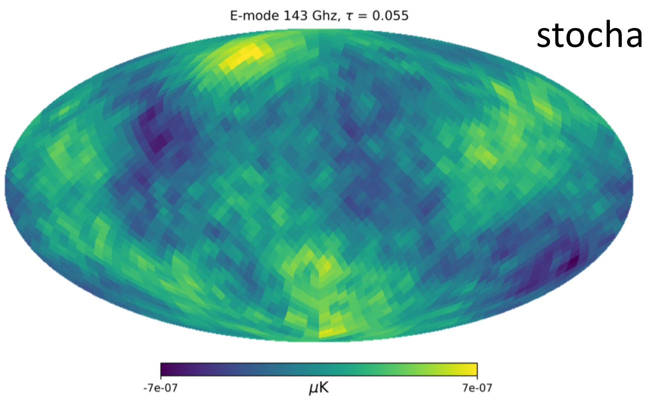
$\tau = 0.55$



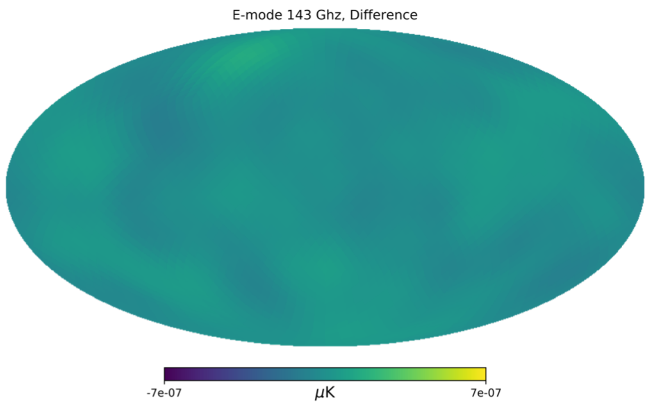
$\tau = 0.6$



simulations



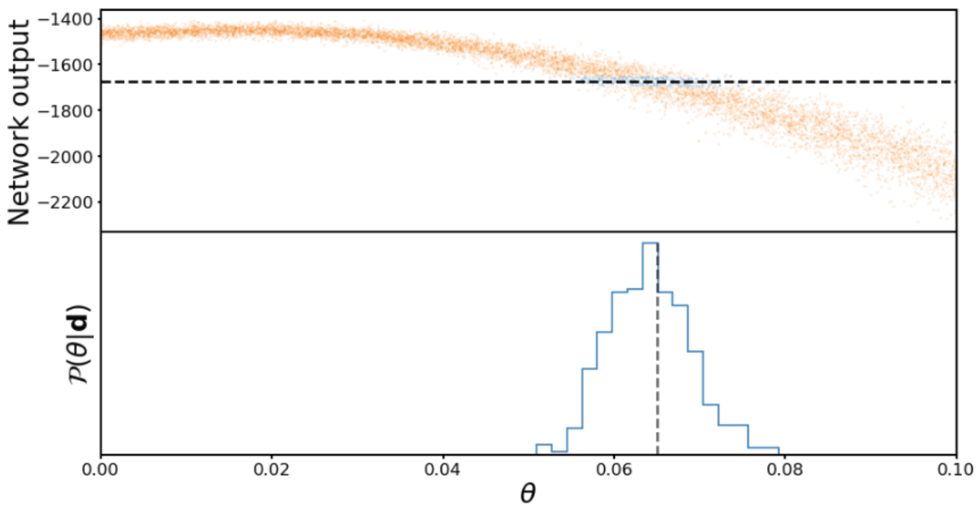
stochastic signal



signal difference
(same seed)

Data on spherical graph!

IMNN inference from Planck (simulated)



Total mission cost: $\$10^9$

Fully optimal analysis in O(1) postdoc days vs

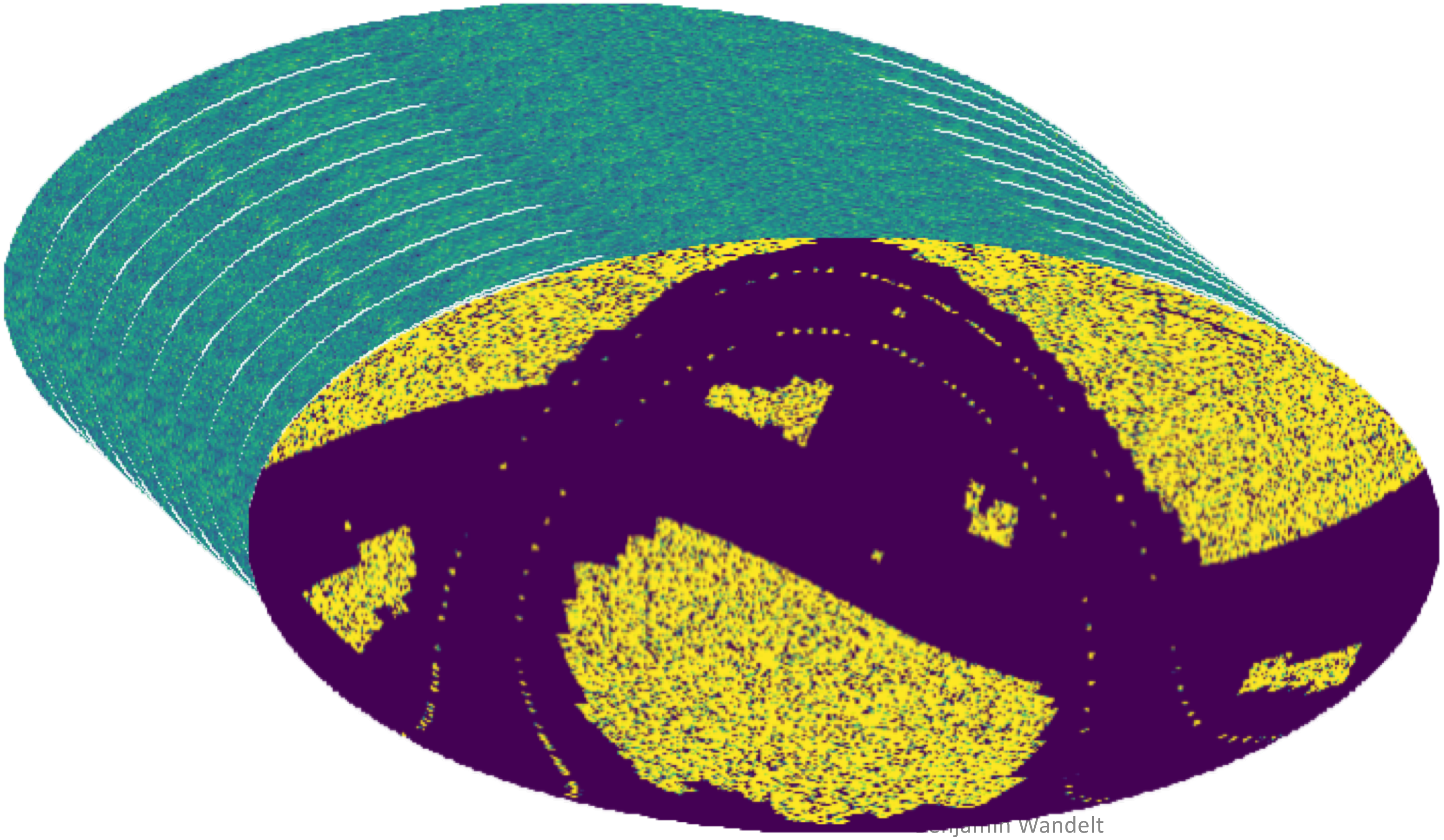
4 Planck mission teams working for O(1) years.

The latest

- Include nuisance-hardened compression technique to ”pre-marginalize” nuisance parameters (Alsing & Wandelt 2019, on arXiv in the next days). Greatly reduces required number of simulations.
- Now use a deep ensemble of Masked Autoregressive Flows to fit the likelihood (Alsing et al. to be submitted). Includes active learning strategy for choosing where to run simulations

Gravitational lensing

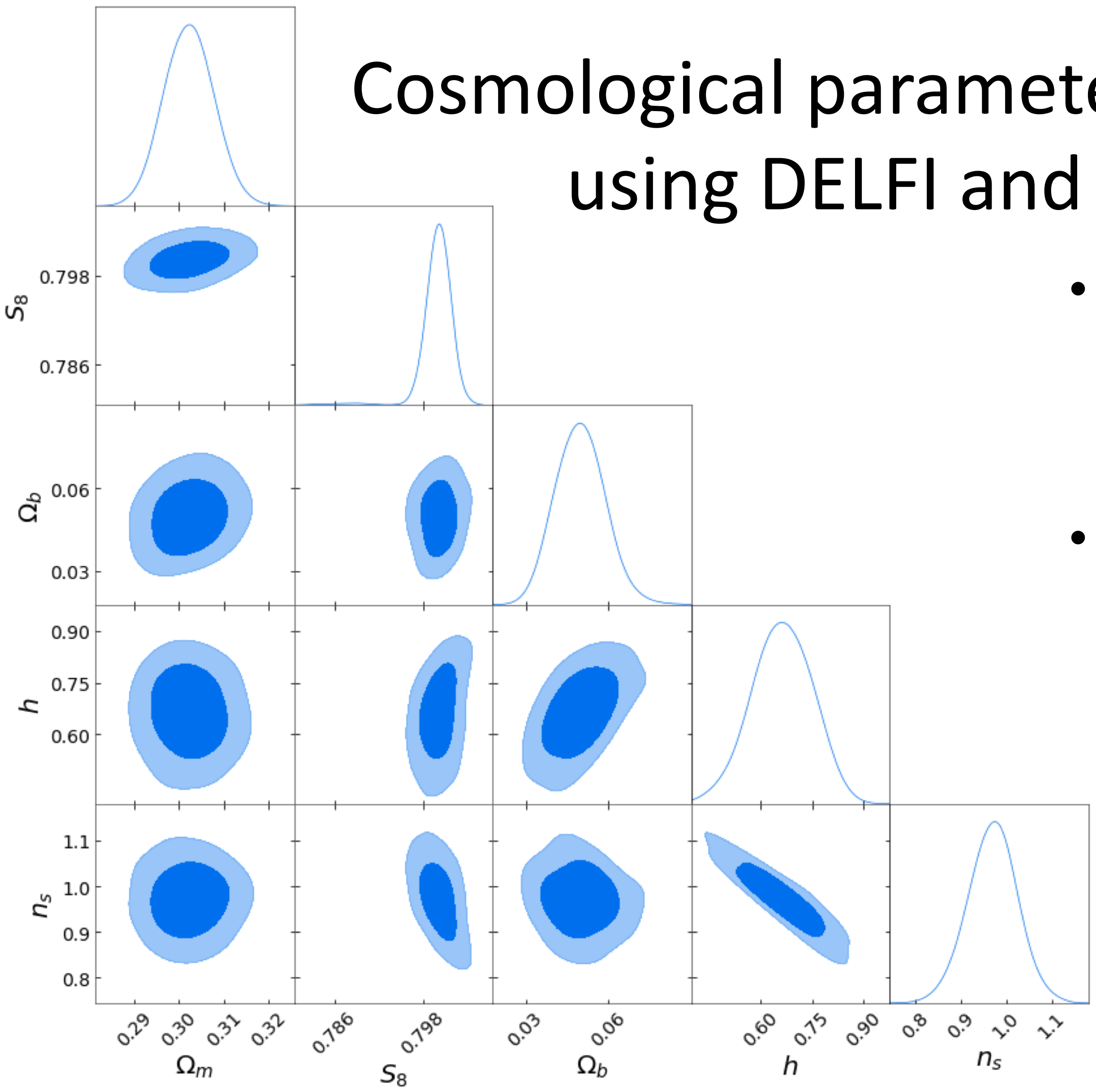
Weak lensing tomography: the data



Mask and noise

10 spherical shells of correlated signal simulations

Cosmological parameter inferences using DELFI and IMNN



- The IMNN compresses hundreds of statistics based on the 10 masked sky maps into 5 sufficient statistics.
- DELFI uses six neural density estimators: 5 MDNs with 1-5 Gaussian components, each with two hidden layers of 50 hidden units and a MAF containing five MADEs.

Summary

A broad spectrum of machine learning ideas are at play in cosmology and astrophysics

- With non-linear inference techniques we can now reconstruct our cosmic history and unlock a vast range of scales to probe dark matter and dark energy
- Automatic physical inference using information-maximizing neural networks and tractable likelihood-free inference is a new approach to extracting scientific, information from complex data and physical simulations.
- Cosmology adds some unique challenges:
  - observational science (e.g. out of class training)
  - Information is in “correlations” between all cosmic messengers (photons, particles, gravitons)
  - Precision science with only one universe!

To reproduce the results in the IMNN paper the code version used is archived on zenodo

https://doi.org/10.5281/zenodo.1175196

The most current development version is on github:

IMNN:

https://github.com/tomcharnock/IMNN

DELFI:

https://github.com/justinalsing/pydelfi