Machine Learning in Cosmology

Benjamin Wandelt (IAP, ILP, Sorbonne U.)

With Justin Alsing, Tom Charnock, Guilhem Lavaux, Stephen Feeney

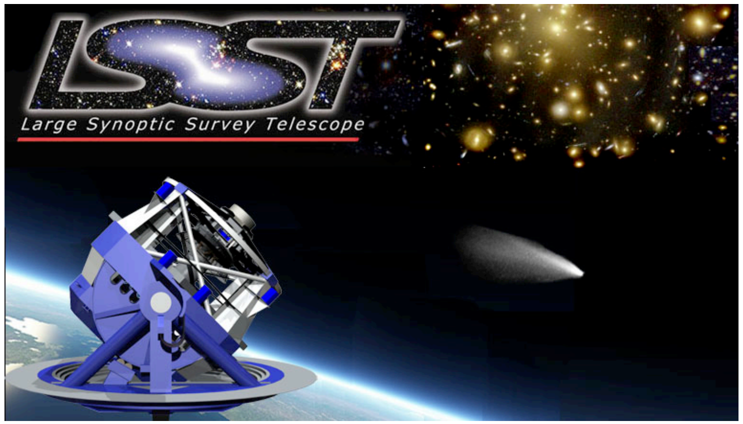
We live in the era of cosmological data

esa PLANCK

EUCLID
MAPPING THE GEOMETRY OF THE DARK UNIVERSE

MegaPrime CFH MegaCam
Legacy Survey
CFHTLS

CMB-S4



THE DARK ENERGY SURVEY

HYPER SUPRIME-CAM

WFIRST

Your favorite survey here

What makes learning from cosmic structure exciting?

We are awash in informative data – forms of Moore's law have held for astrophysical data for decades and 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.

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

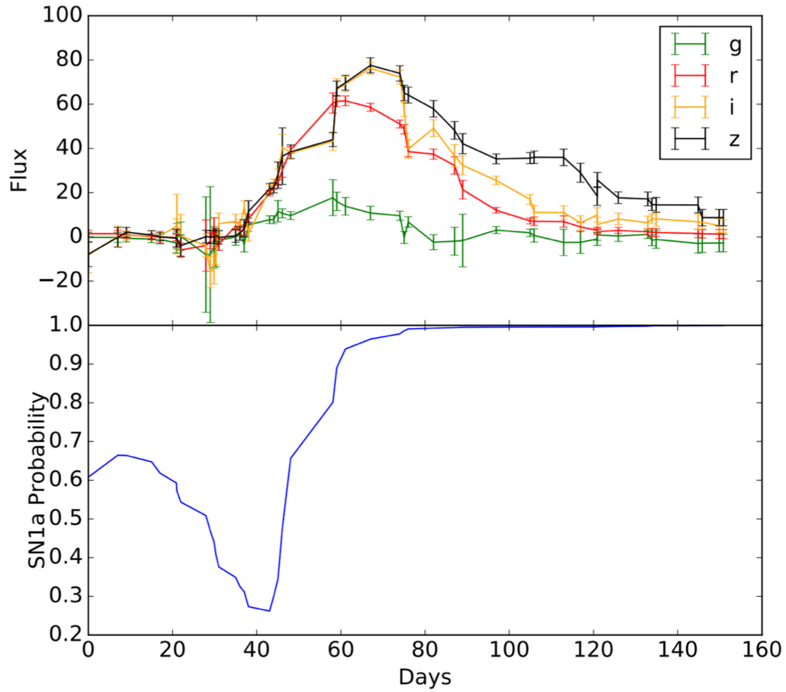
Why machine learning in **cosmology**?

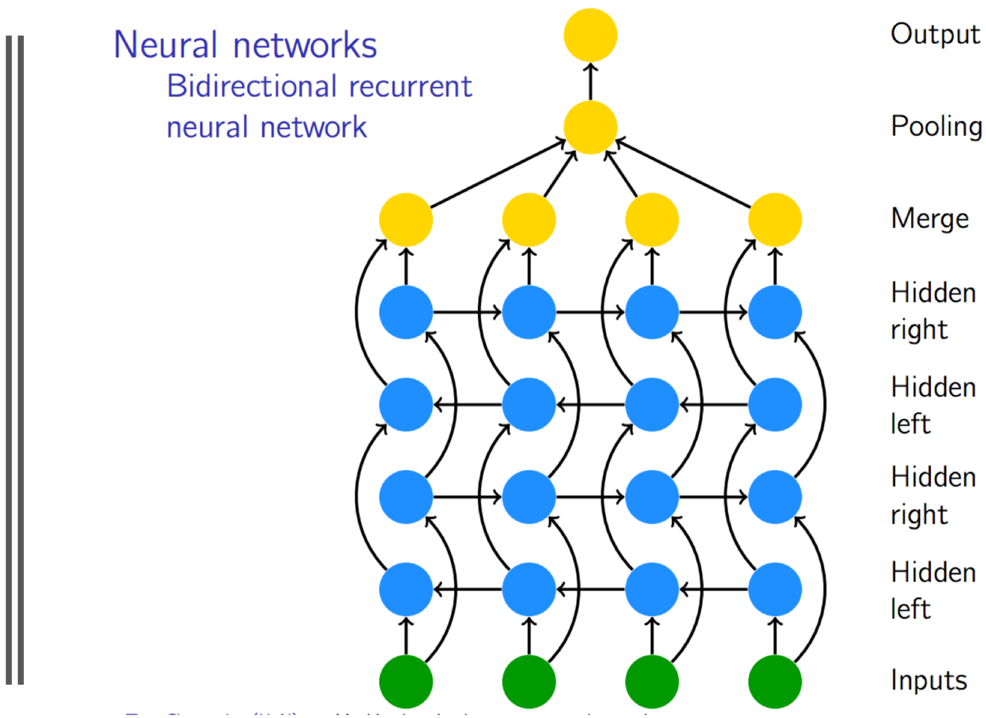
- **Automation**
  - Size of data sets (e.g. LSST will be on SDSS per week) means manual intervention is only possible in highly exceptional circumstances. Many data analysis tasks need to be automated.
- **Acceleration**
  - Machine learning can provide short cuts to costly physical simulations.
- **Superhuman 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 standard candles. Traditionally typing is done through spectroscopic follow-up.
- But spectroscopic follow-up is far too costly for LSST.
- Can machine learning identify types?

Supernova classification

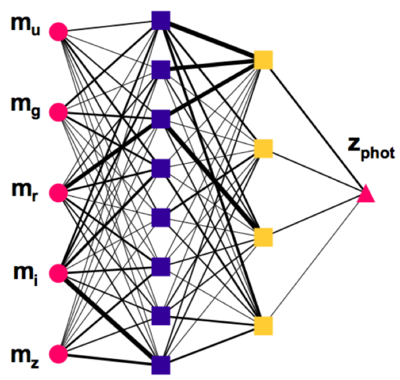




Charnock & Moss (2016), Möller *et al.* (2016)

Photo-z estimation

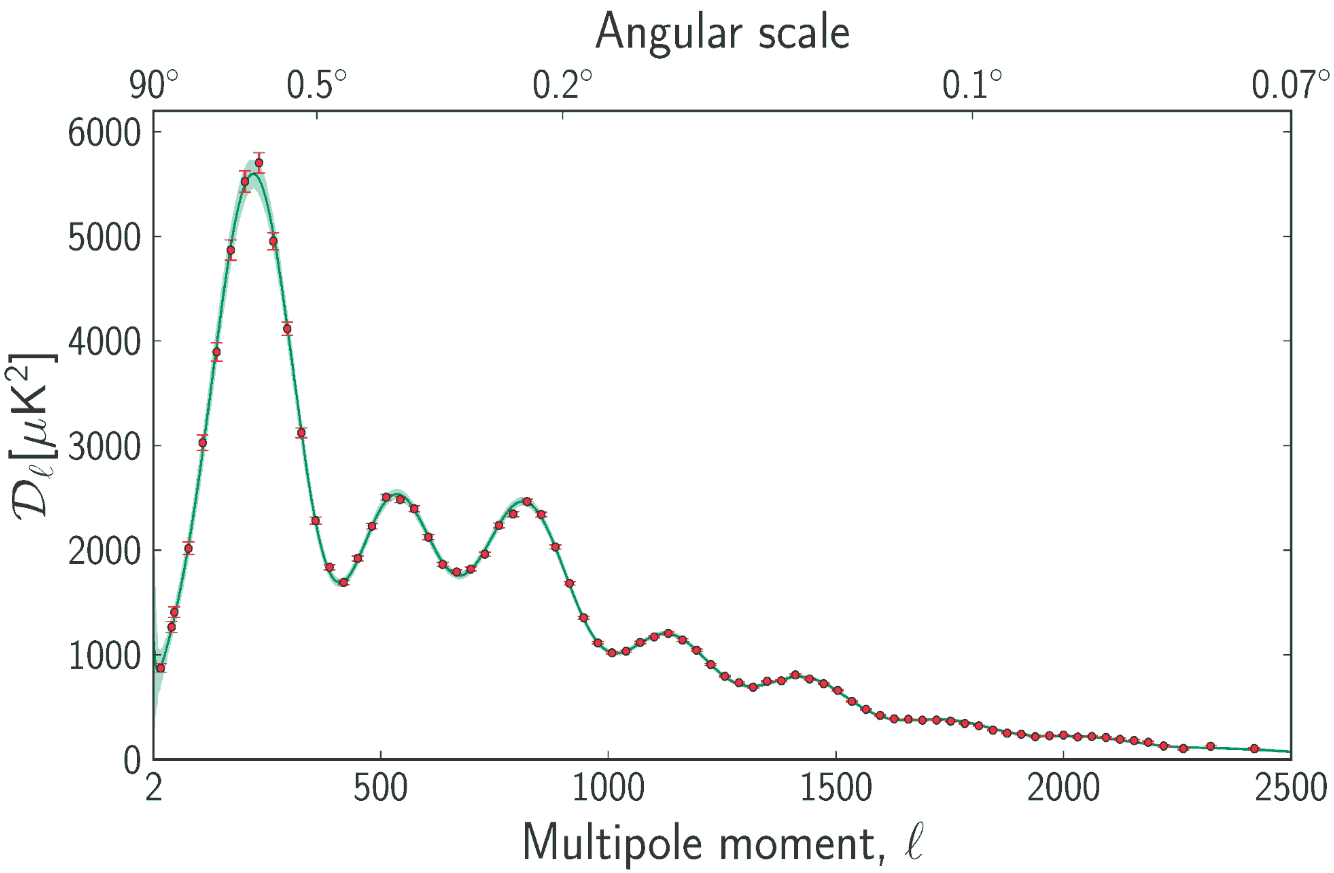
- Estimate the redshift of the spectrum of an object from its colors.
- First serious application of NN in astronomy.
- ANNz (Collister, Lahav 2004) was best in class. Now superseded by updated versions.
- Intrinsic difficulty: *how to train out-of class data? (Discuss!)*



Acceleration

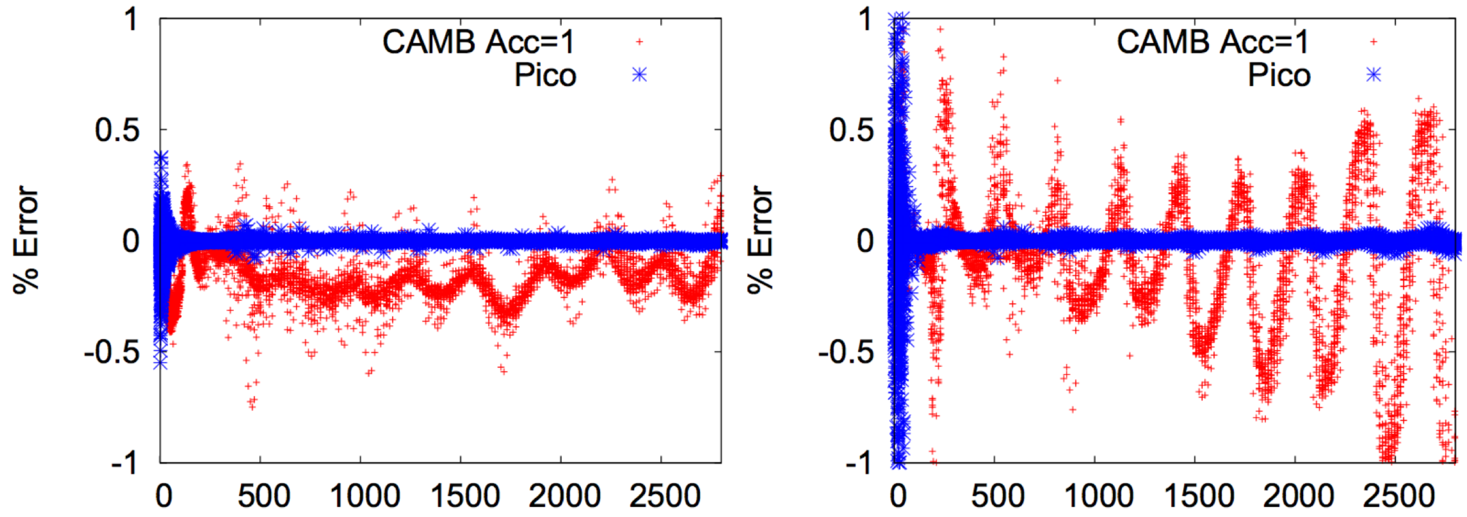
- **Emulators**
- Solve the problem of model predictions that involve costly computations
- First high-accuracy emulator in cosmology: **Pico** (Fendt, Wandelt 2007)
  - Acceleration through Parallel Precomputation and LEarning (APPLE)
  - PCA pre-compression of outputs – then predict compressed quantities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

Theory confronts Planck data – extreme accuracy required



Need 10^5 sequential evaluations per analysis.
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

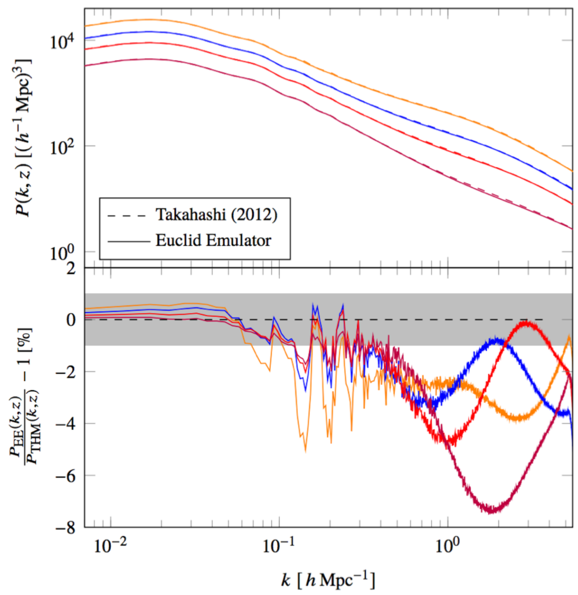
Machine learning $\neq$ Neural Networks

- 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.

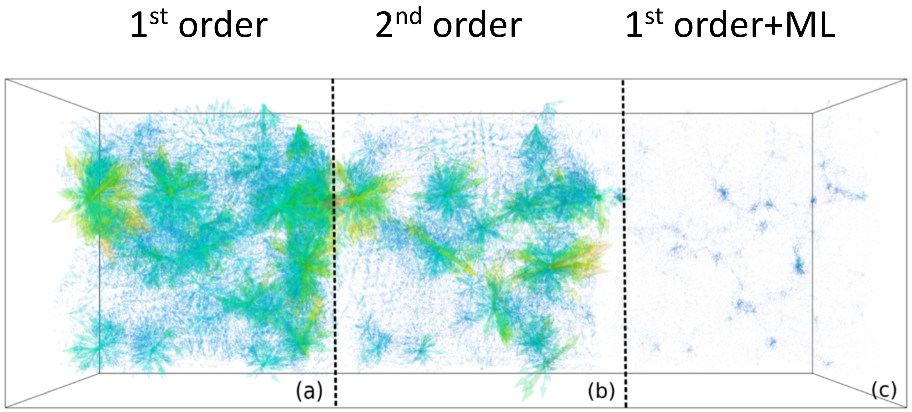
- So let's not forget trees/forests, regression, clustering, Gaussian Processes etc.

Emulators now

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

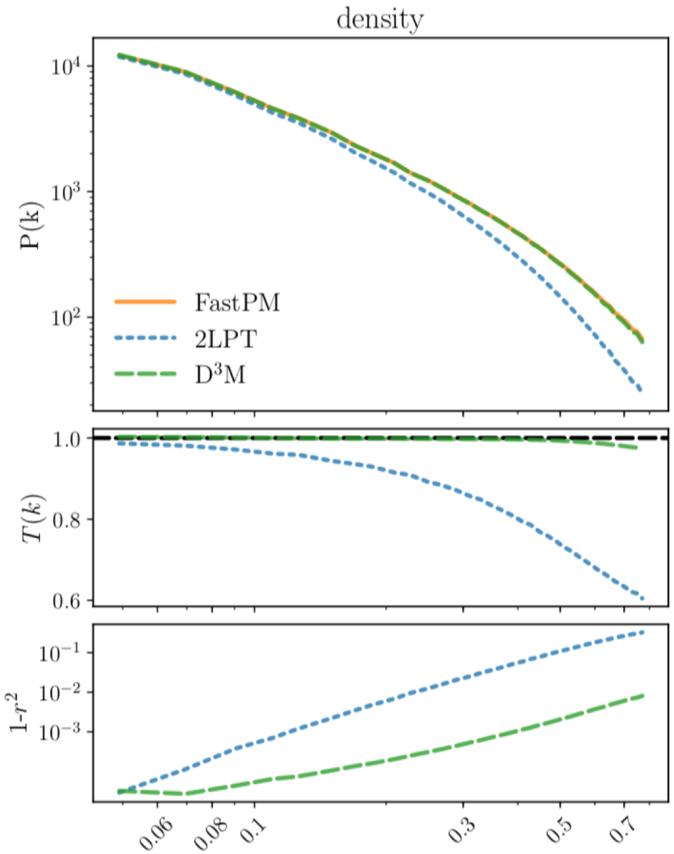


Machine learning cosmological physics



- He et al. (arXiv: 1811.06533)

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



Benjamin Wandelt

Generative Adversarial Networks to simulate galaxy images

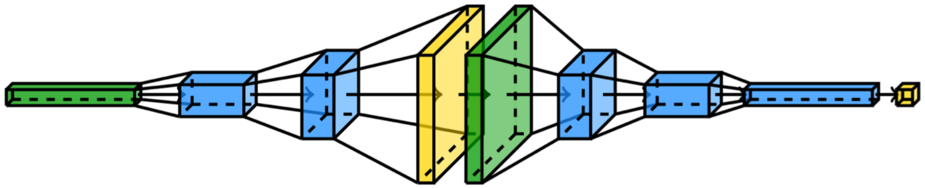
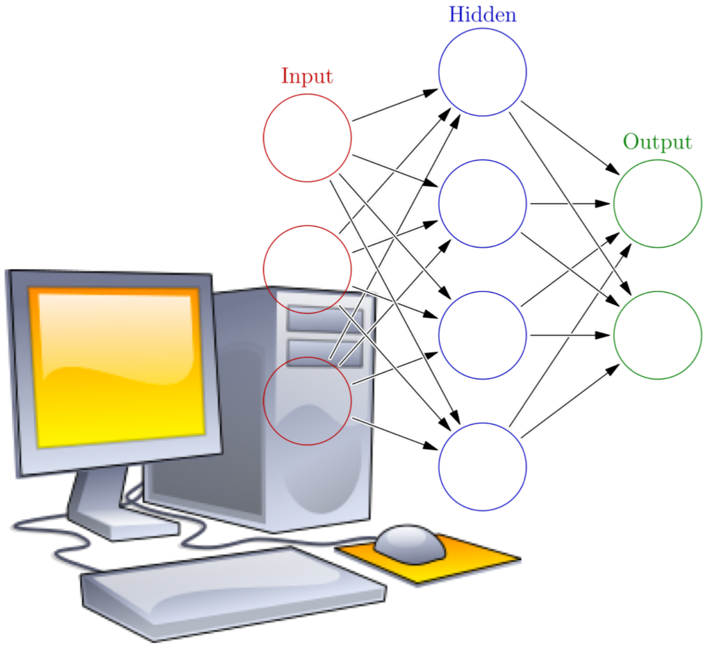
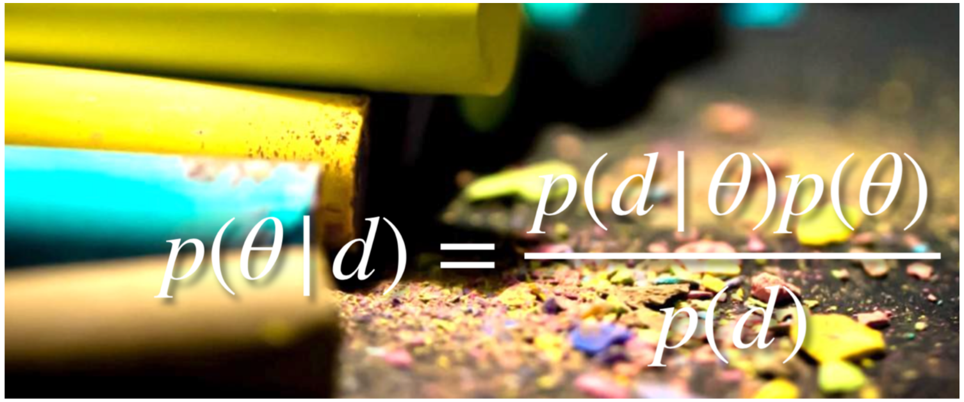


Image credit: T. Charnock

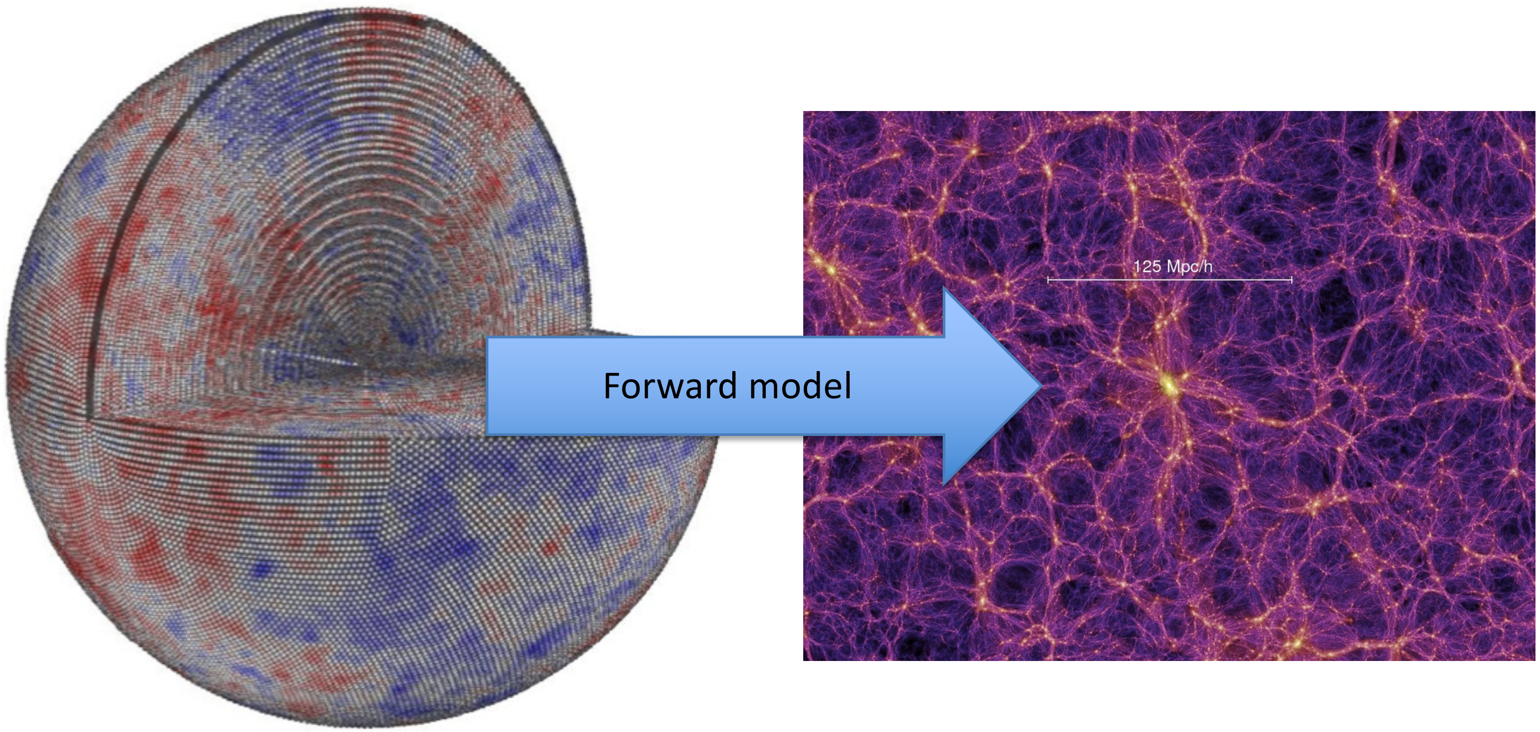
Machine learning vs Physics?



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



Physics-based machine learning



Initial conditions of the universe

The universe today

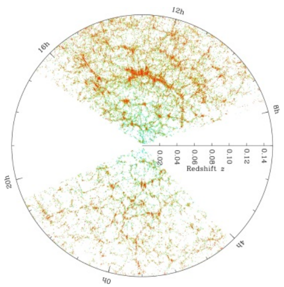
A fully *probabilistic* model of galaxy surveys

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 method



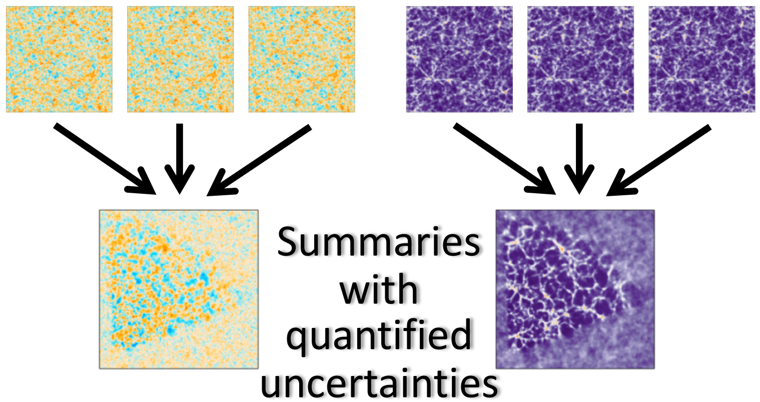
Observations

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

→ BORG →

E.g. inferred dark matter densities

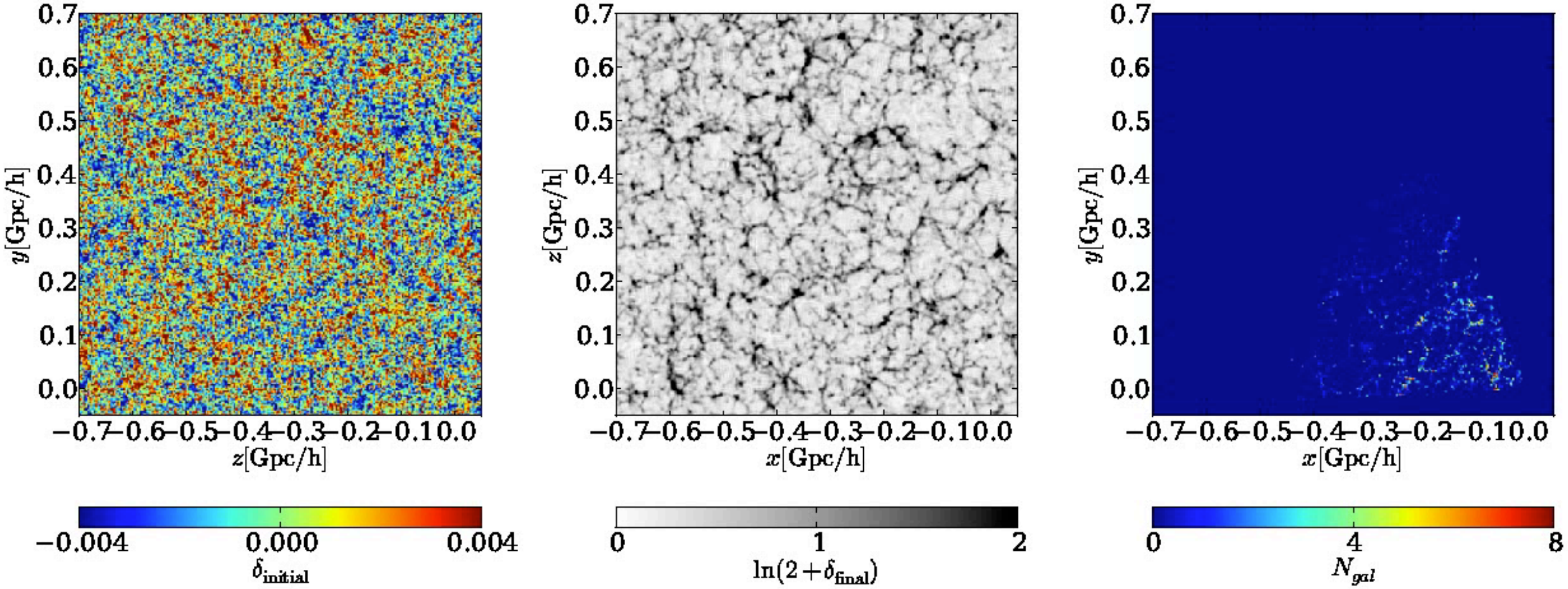
z=100 z=0



Jasche & Wandelt 2013, arXiv:1203.3639

Jasche, Leclercq & Wandelt 2015, arXiv:1409.6308

Bayesian LSS sampling - movie



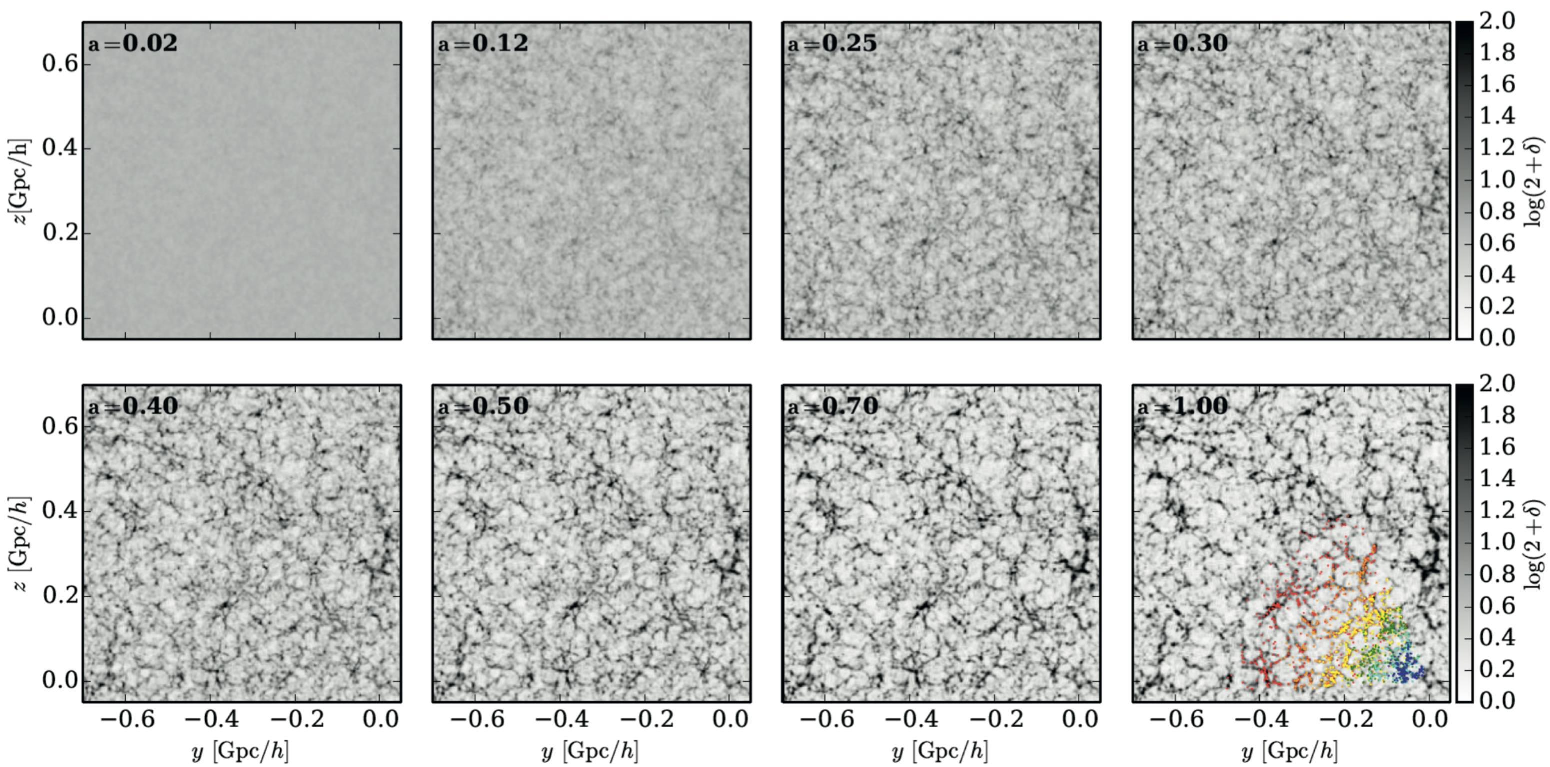
Initial conditions

Final conditions

Observations

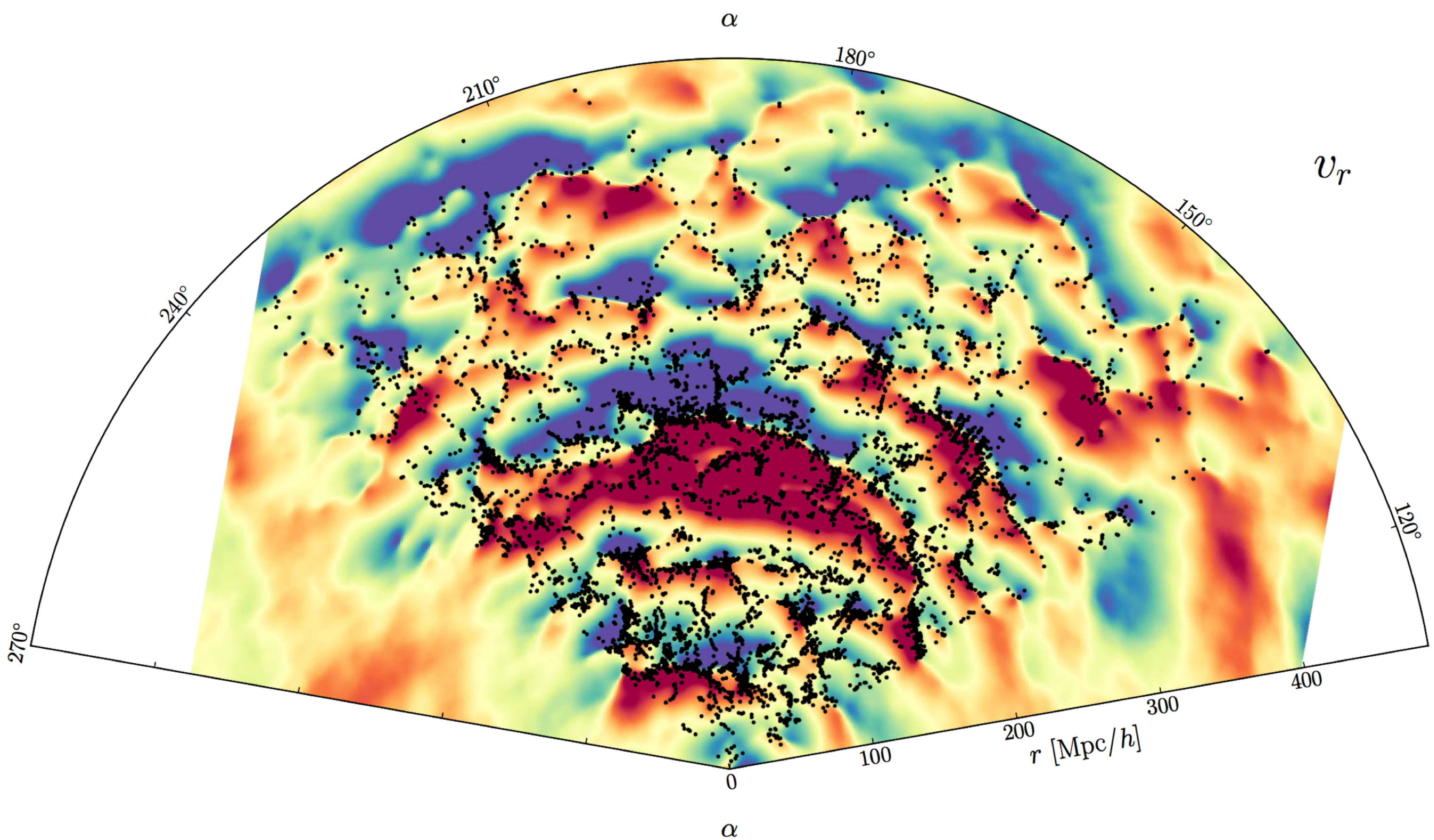
Jasche, Leclercq & Wandelt 2014, arXiv:1409.6308

A posterior sample of the formation history of our Universe



Jasche, Leclercq & BDW 2014, arXiv:1409.6308

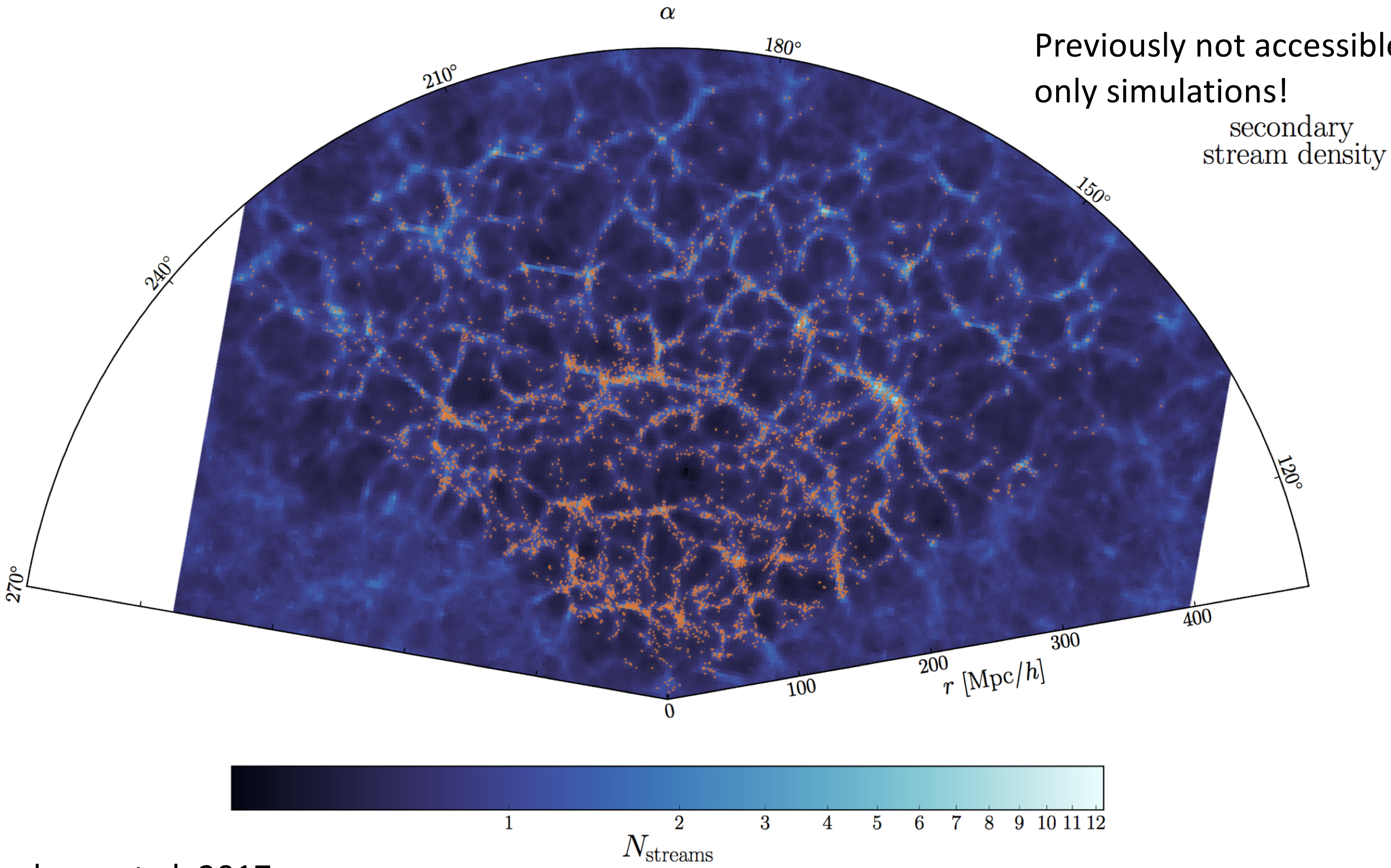
Bayesian LCDM predictions: Dynamical velocities



Leclercq et al. 2017

Posterior mean of Lagrangian stream density

Previously not accessible in data, only simulations!



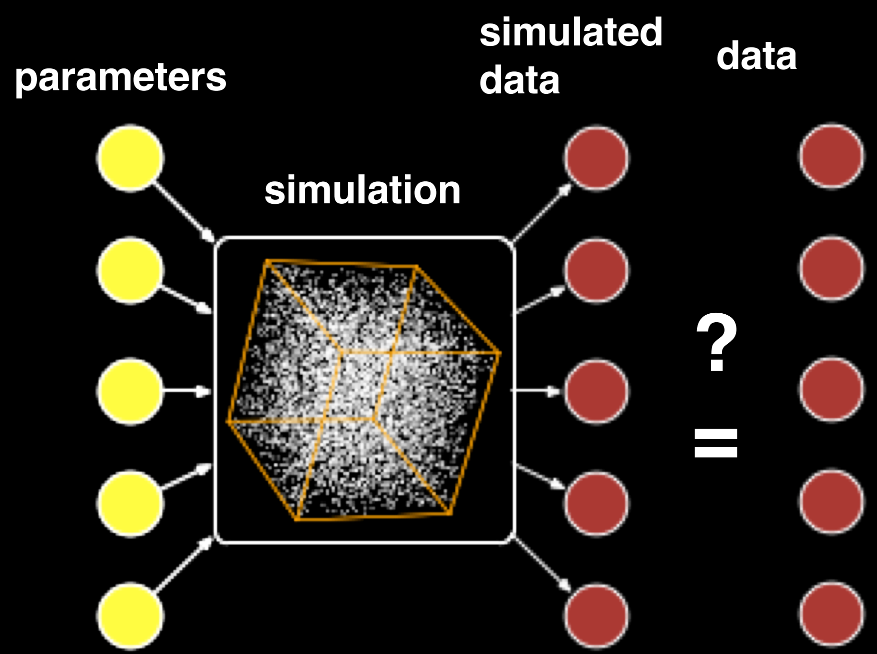
Leclercq et al. 2017

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:

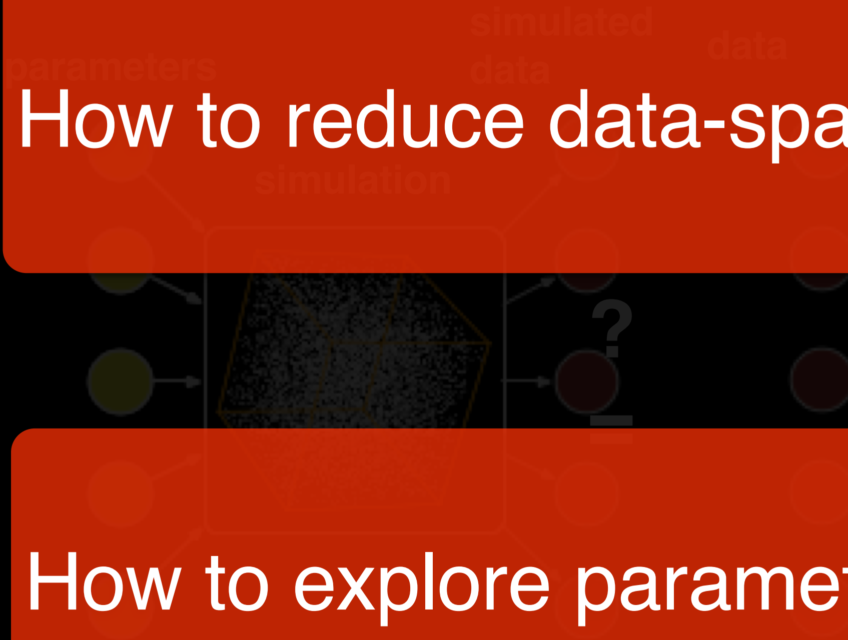
reject;

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?



Draw from prior:

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

Simulate data:

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

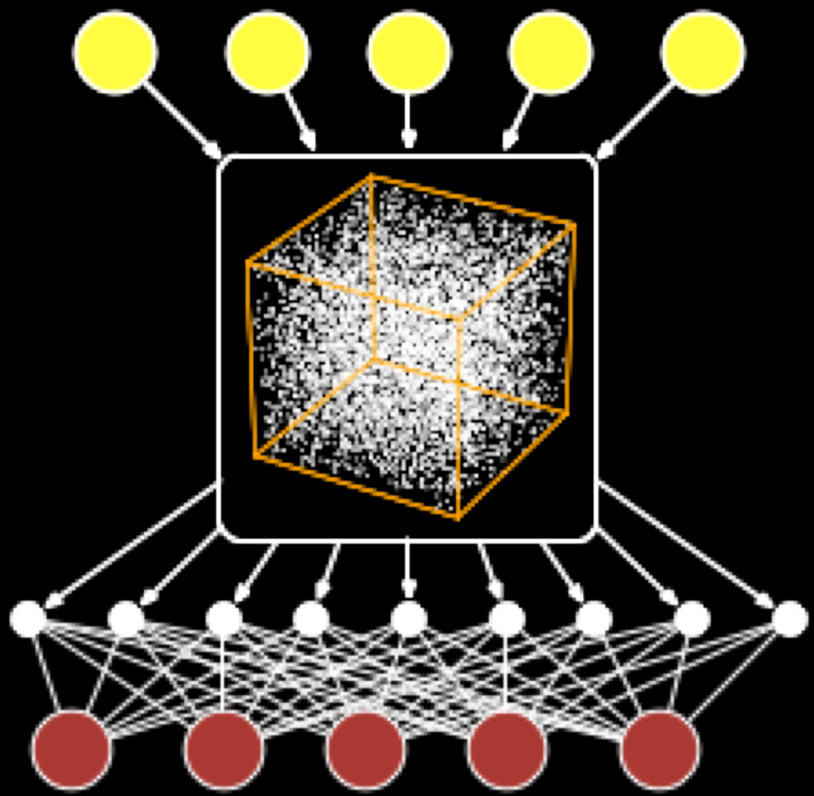
If $\rho(\mathbf{D}^*, \mathbf{D}) < \epsilon$:

accept;

reject;

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

Massive data compression



n **parameters**

N **data**

M **summary statistics**

n compressed statistics

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

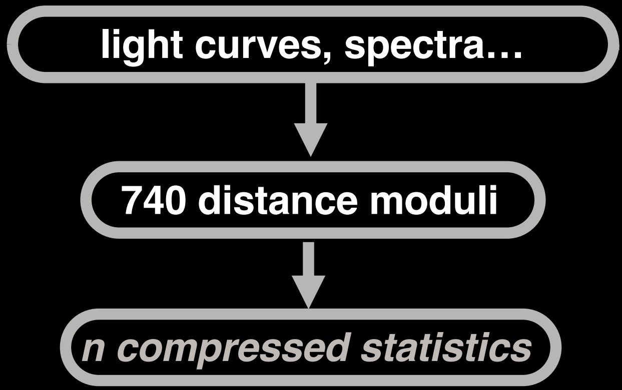
Exploration of parameter space

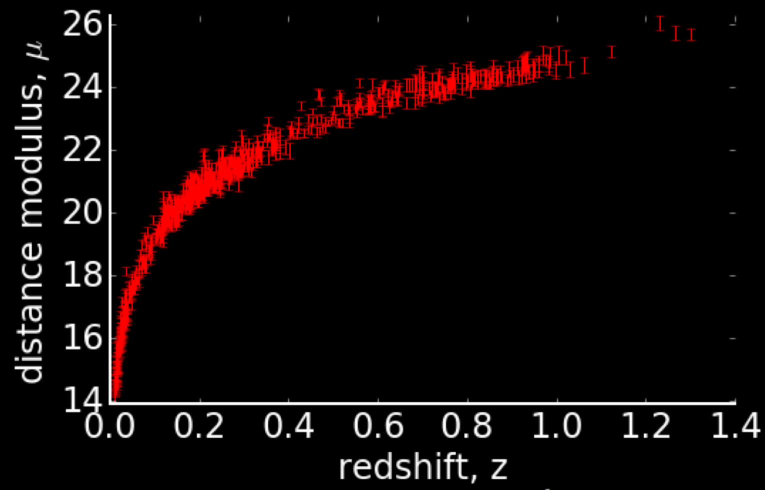
Density estimation Likelihood free inference (DELFI)

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

Alsing, Feeney &Wandelt arXiv: 1801.01497

Case study: JLA SNe





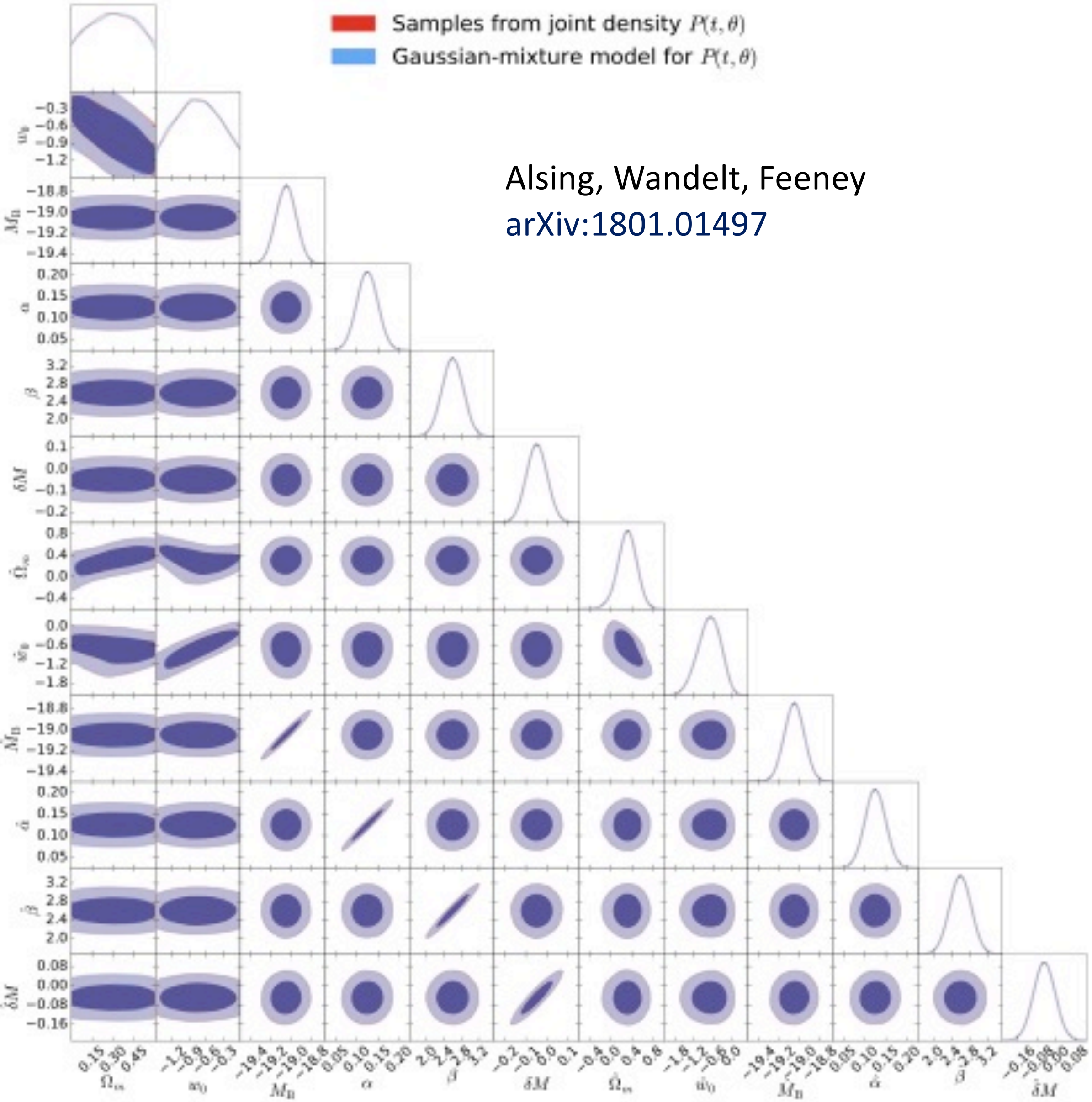
$$\mathcal{L} = -\frac{1}{2}(\mathbf{d} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{d} - \boldsymbol{\mu}) - \frac{1}{2}\ln|\mathbf{C}|$$

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\Omega_{\mathrm{m}}, w_0, M, \alpha, \beta, \delta m)$$

$$\mathbf{C} = \mathbf{C}(\alpha, \beta)$$

Betoule et al 2014

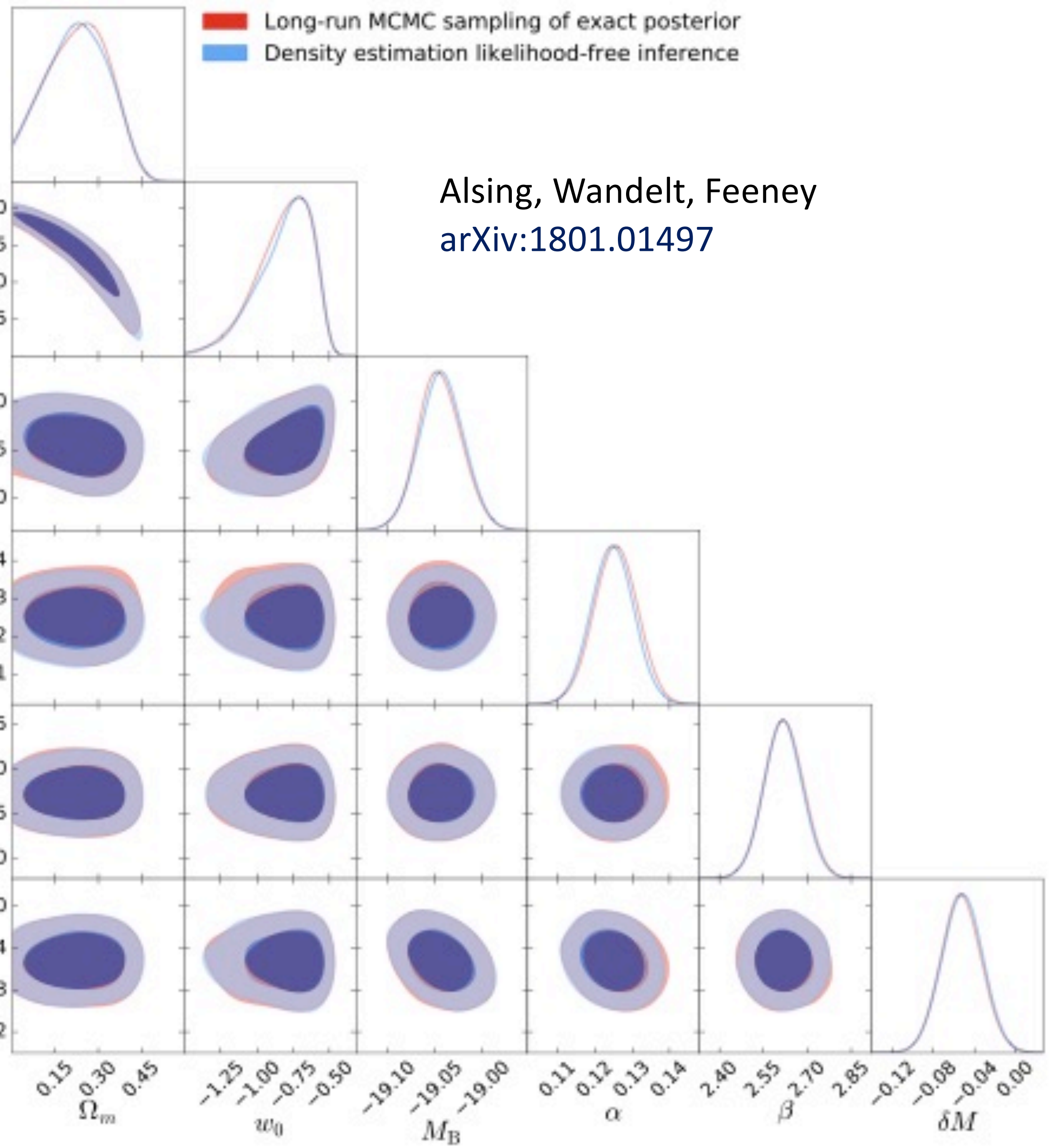
Fit to joint density $P(\boldsymbol{\theta}, \mathrm{d}^*)$



Alsing, Wandelt, Feeney
arXiv:1801.01497

(10000 simulations)

Posterior inference



Alsing, Wandelt, Feeney
arXiv:1801.01497

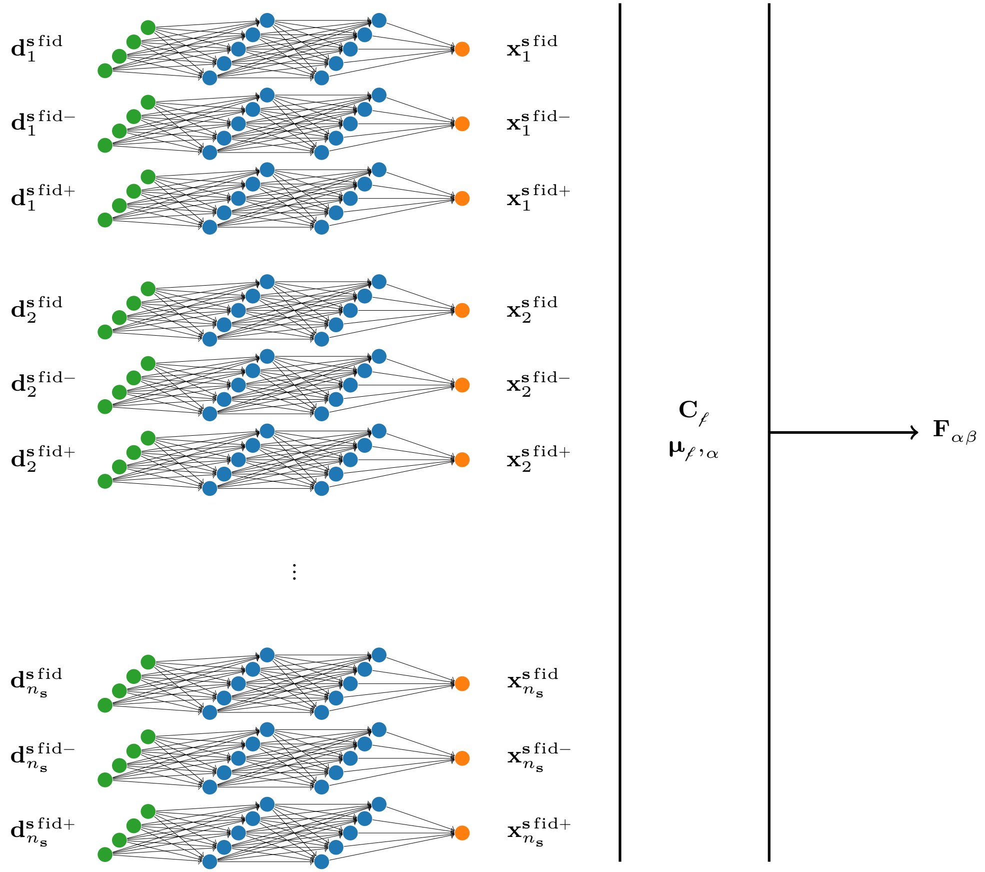
But what if you don’t know how to compute informative summaries of your data?

Automatic Physical Inference

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

- Obviates the need to “guess” heuristic, informative summaries of the data
- An neural network is trained through reinforcement learning to compute functions of the data (*summaries)* that maximize the information about the parameters of the model.
- The network generates non-linear informative summaries that constrain model parameters

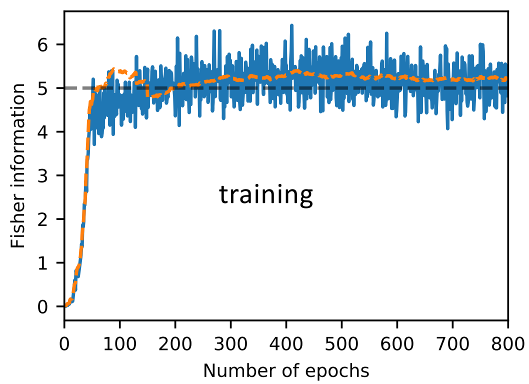
Information maximizing neural network

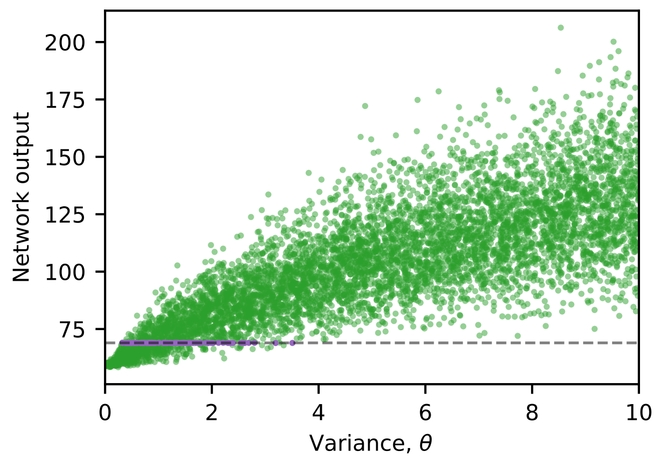


Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 1: variance inference

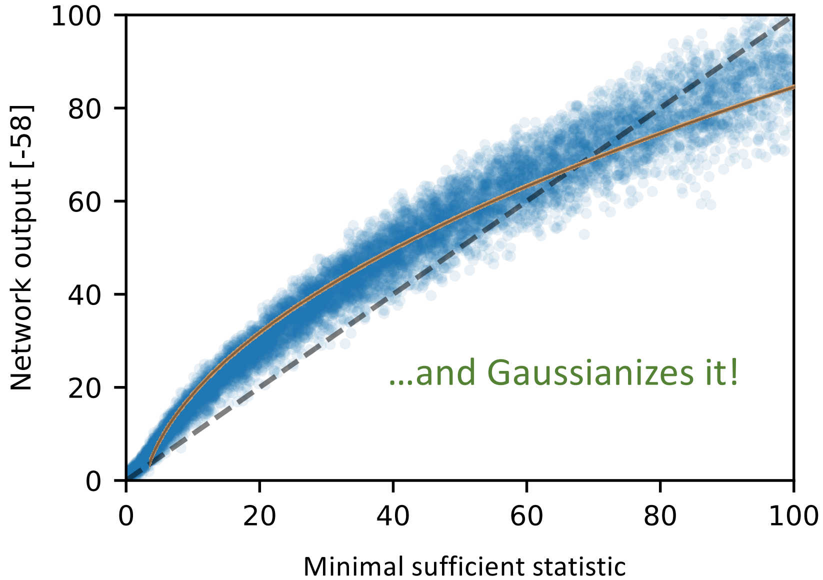
- Perfect information $F = 5$ in this problem
- Linear summaries give $\mathrm{F} = 0.5$





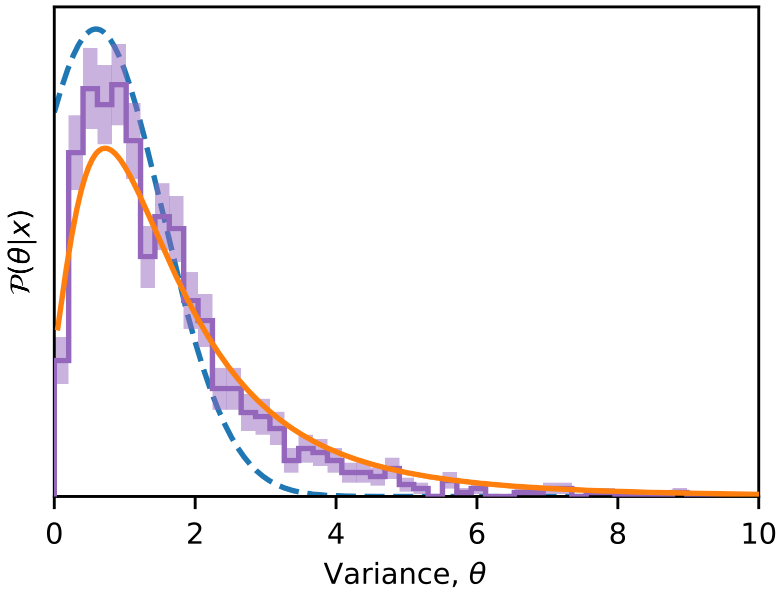
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

The IMNN finds a minimal sufficient statistic for this inference problem



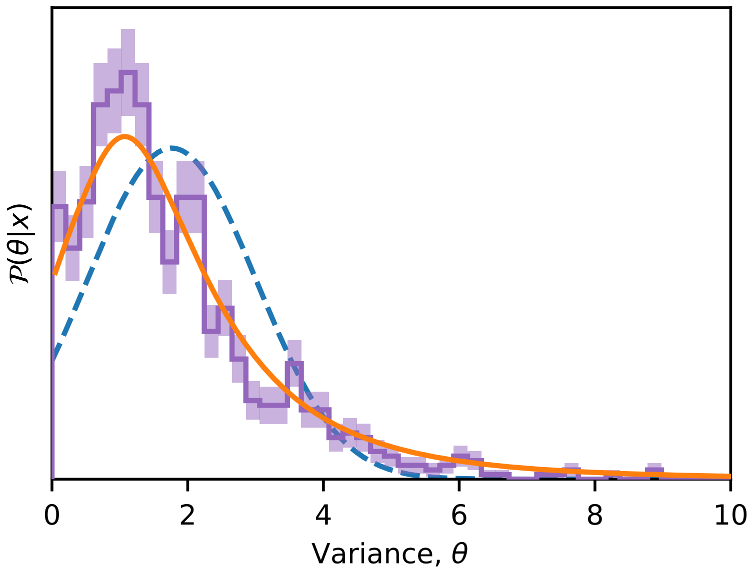
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 2: Automatic physical inference with noisy data



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

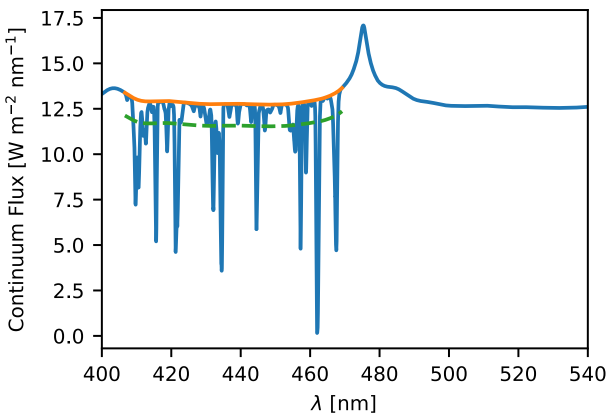
Example 3: Automatic physical inference with unknown noise

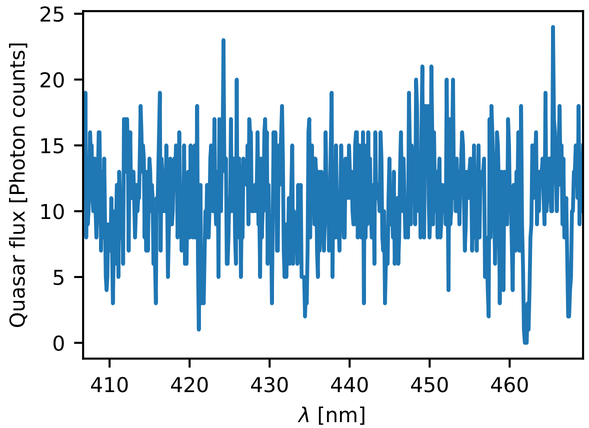


Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 4: Lyman-α forest inference

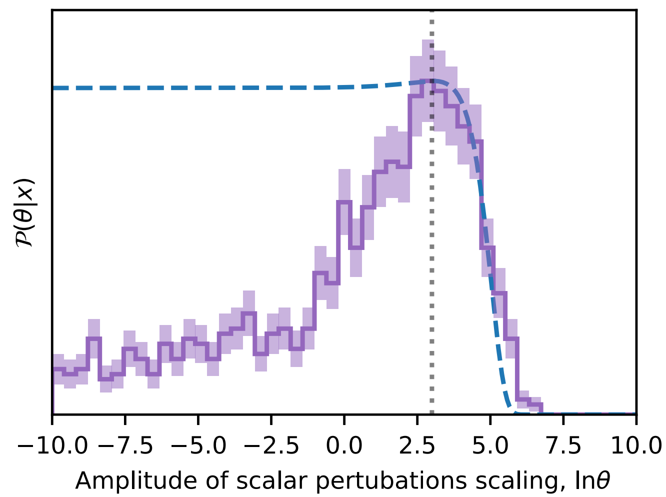
- The idea is to infer the variance of the underlying density field from a non-linearly transformed, photon-noise dominated Lyman-α forest spectrum





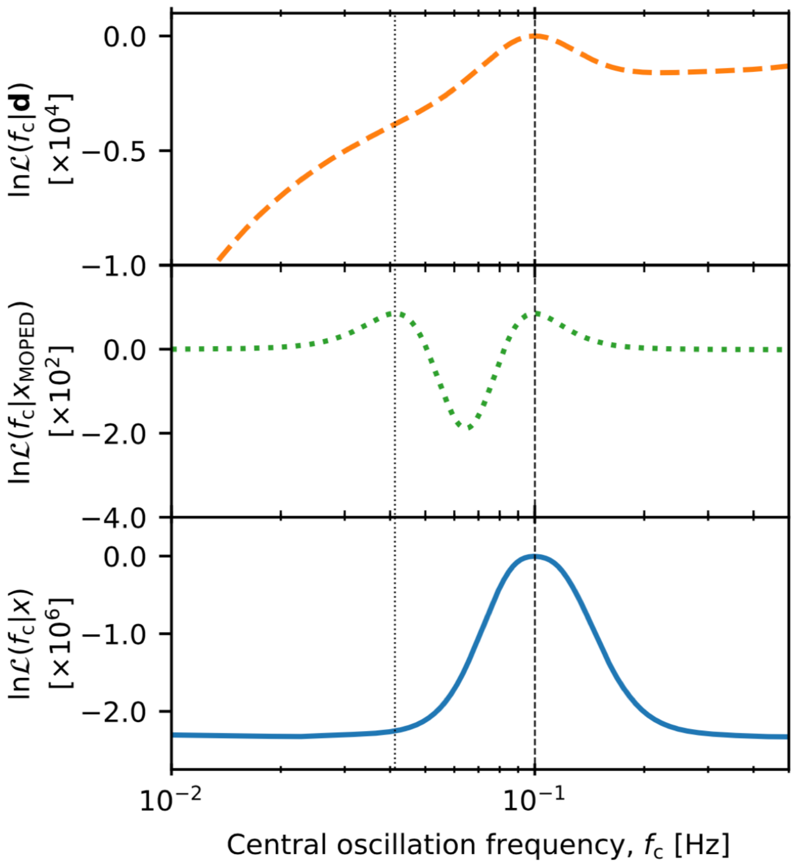
Charnock, Lavaux, Wandelt (arXiv:1802:03537)

Example 4: Lyman-α forest inference



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

Example 5: LISA gravitational wave chirp



Full likelihood (LH)

Gaussian likelihood based on linear compression

Gaussian LH based on IMNN compression

The Information Maximizing Network summary gives the correct unique likelihood peak.

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

The latest

- Include nuisance-hardened compression technique to ”pre-marginalize” nuisance parameters
- Use another NN, a *mixture density network* to fit the likelihood rather than Gaussian mixture.

- Larger applications (see Alessandro Manzotti’s talk this afternoon!)

- Note: HEP paper in very similar spirit a few months later by Brehmer et al. arXiv: 1805.12244

Summary

- A broad spectrum of machine learning ideas are at play in cosmology and astrophysics
- Need to find good ways to balance:
  - the *power of physical modeling* with the *power of machine learning*
  - *ad hoc analysis* with *principled statistics*
  - *interpretability* with *black box machines*
- Cosmology adds some unique challenges:
  - observational science (e.g. out of class training)
  - Information is in “correlations” between all cosmic messengers (photons, particles, gravitons)
  - Only one universe!

Let’s go beyond putting machine learning into physics: let’s put physics into machine learning!

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 recent version can be downloaded from github:

https://github.com/tomcharnock/information_maximiser