How is the cosmic web woven?

Inference with generative cosmological models

Florent Leclercq

http://icg.port.ac.uk/~leclercq/

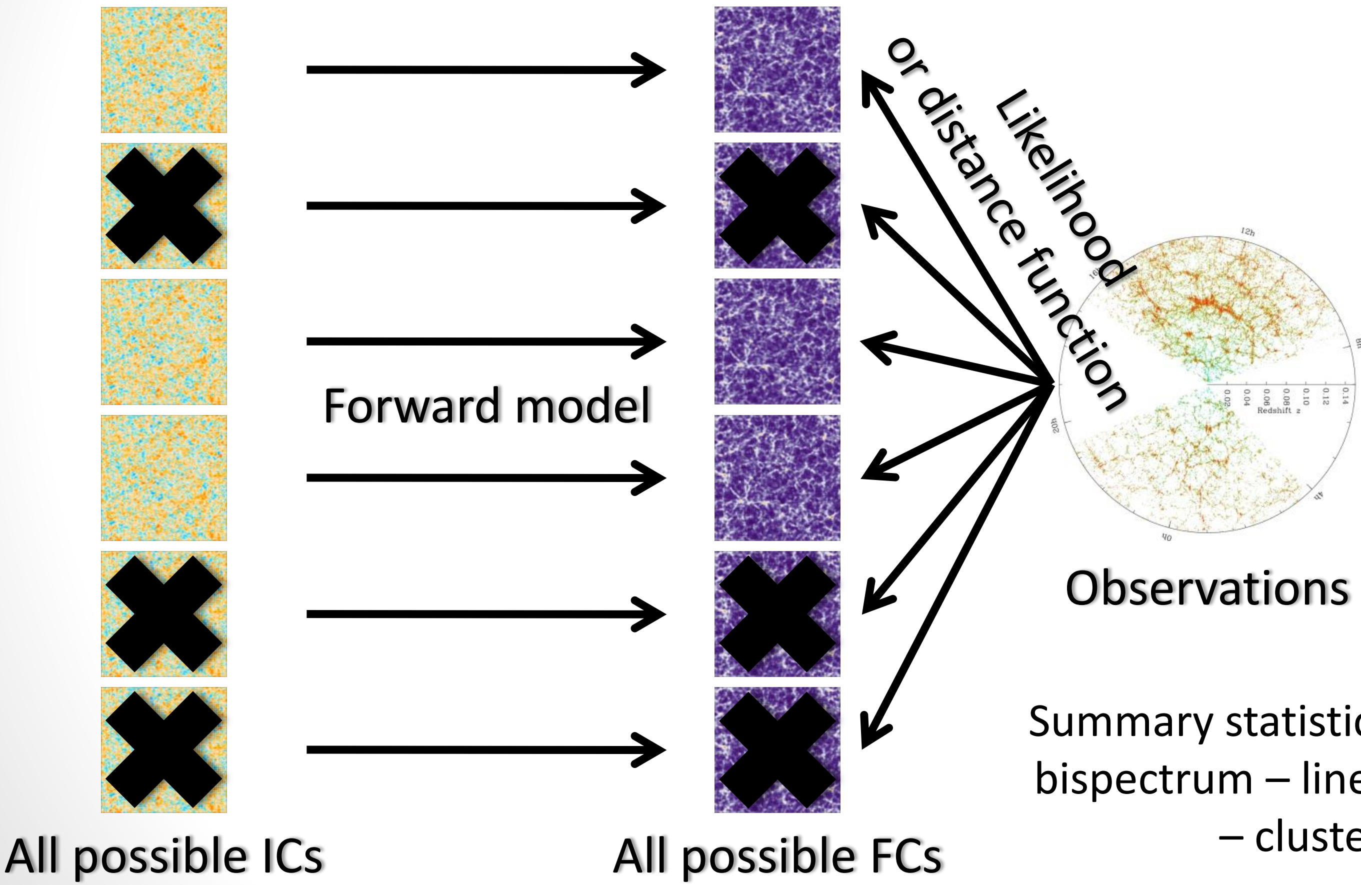
Institute of Cosmology and Gravitation, University of Portsmouth ⇒ Imperial Centre for Inference and Cosmology, Imperial College, London

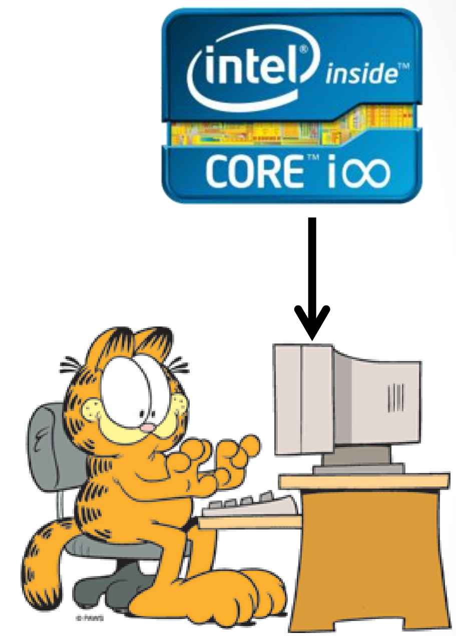
August 28th, 2017

Wolfgang Enzi (MPA, Garching), Baptiste Faure (École polytechnique & ICG),
Jens Jasche (ExC Universe, Garching), Guilhem Lavaux (IAP), Will Percival (ICG),
Benjamin Wandelt (IAP), Matías Zaldarriaga (IAS, Princeton)

Bayesian forward modeling: the ideal scenario

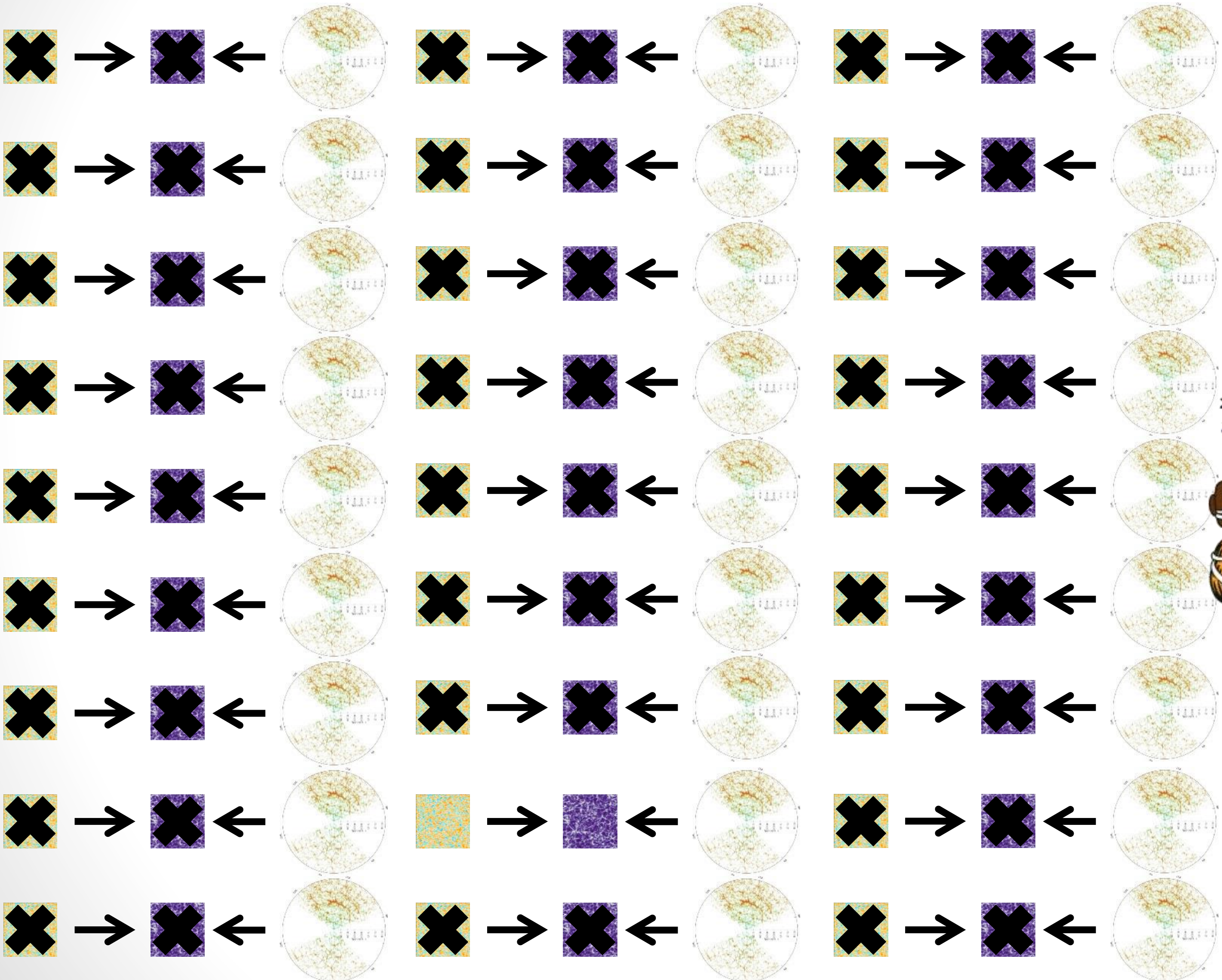
Forward model = N-body simulation + Halo occupation + Galaxy formation + Feedback + …





Summary statistic = power spectrum – bispectrum – line correlation function – clusters – voids…

Bayesian forward modeling: the challenge

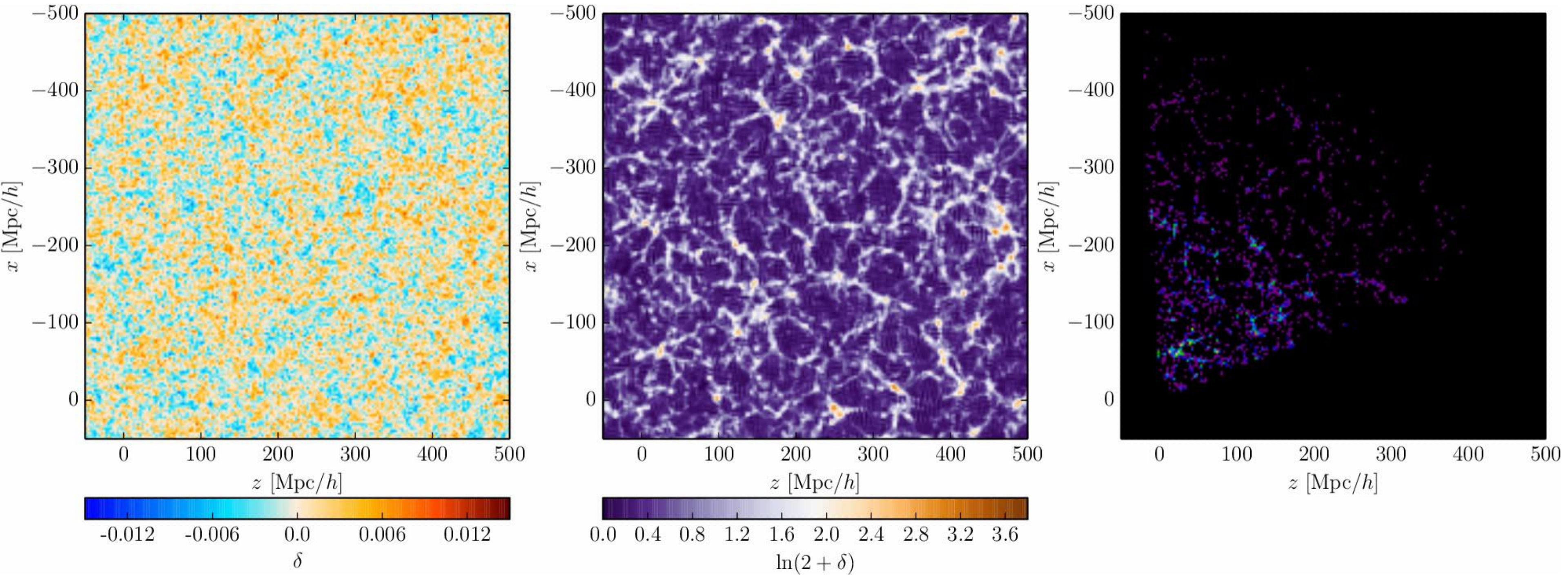


$d \approx 10^7$

Likelihood-Based Solution: BORG

Likelihood-based solution: BORG

uses Hamiltonian Monte Carlo (HMC) to explore the exact posterior

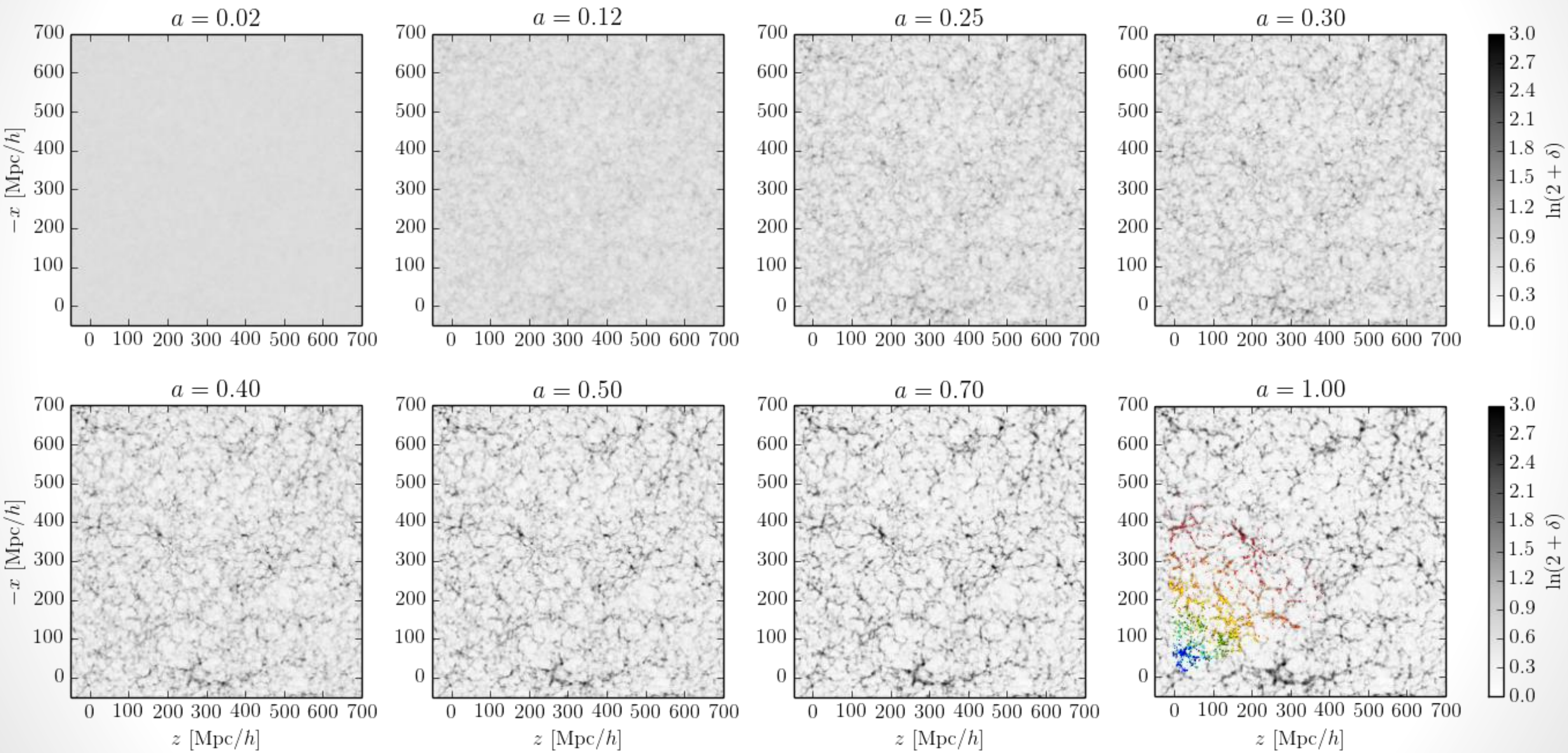


Initial conditions | Final conditions | Observations

334,074 galaxies, ≈ 17 millions parameters, 3 TB of primary data products, 12,000 samples, ≈ 250,000 data model evaluations, 10 months on 32 cores

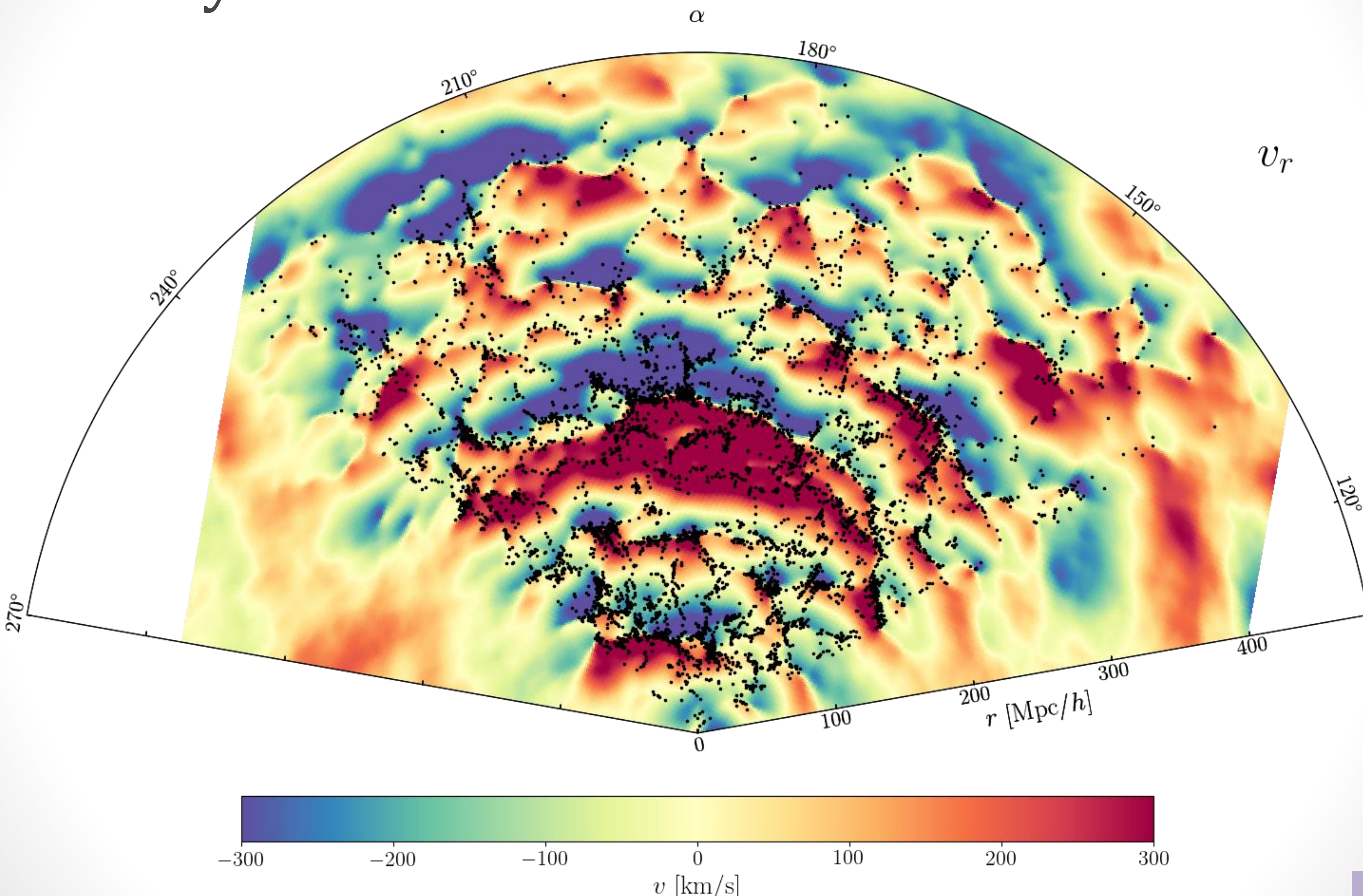
Jasche, FL & Wandelt 2015, arXiv:1409.6308

Evolution of cosmic structure



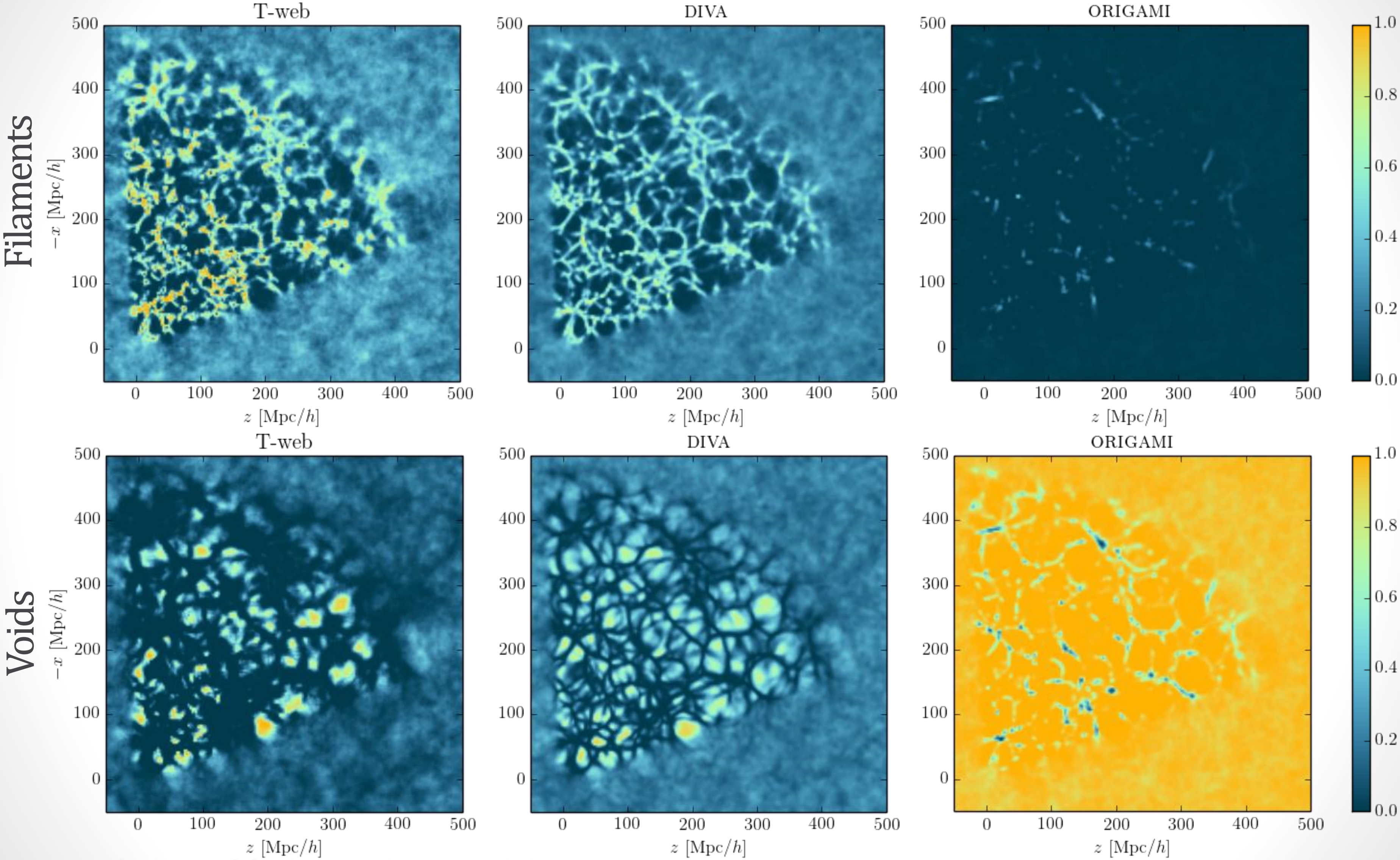
Jasche, FL & Wandelt 2015, arXiv:1409.6308

Velocity field



FL, Jasche, Lavaux, Wandelt & Percival 2017, arXiv:1601.00093

Cosmic web classifications



FL, Jasche & Wandelt 2015a, arXiv:1502.02690

FL, Lavaux, Jasche & Wandelt 2016, arXiv:1606.06758

Likelihood-Free Solution

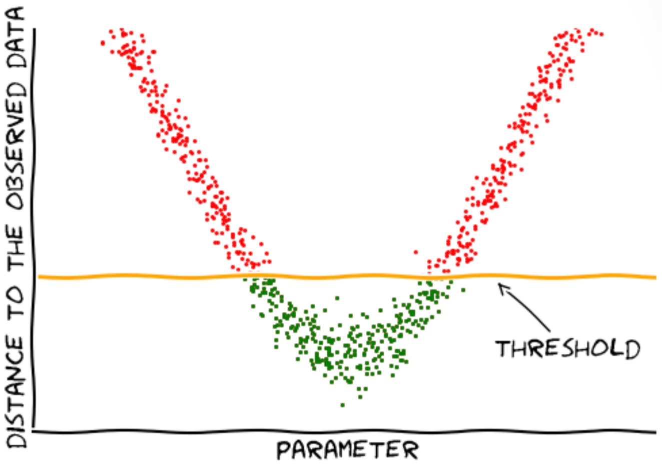
Why is likelihood-free rejection so expensive?

1. It rejects most samples when ϵ is small

2. It does not make assumptions about the shape of $L(\theta)$

3. It uses only a fixed proposal distribution, not all information available

4. It aims at equal accuracy for all regions in parameter space



Effective likelihood approximation:

$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left(\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon \right)$$

Proposed solution

Bayesian optimisation for likelihood-free inference (BOLFI)

1. It rejects most samples when ϵ is small

   ⇨ **Don't reject samples: learn from them!**

2. It does not make assumptions about the shape of $L(\theta)$

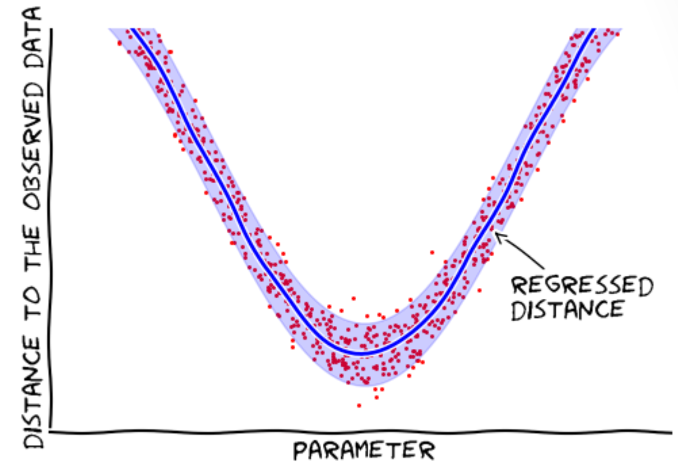
   ⇨ **Model the distances, assuming the average distance is smooth**

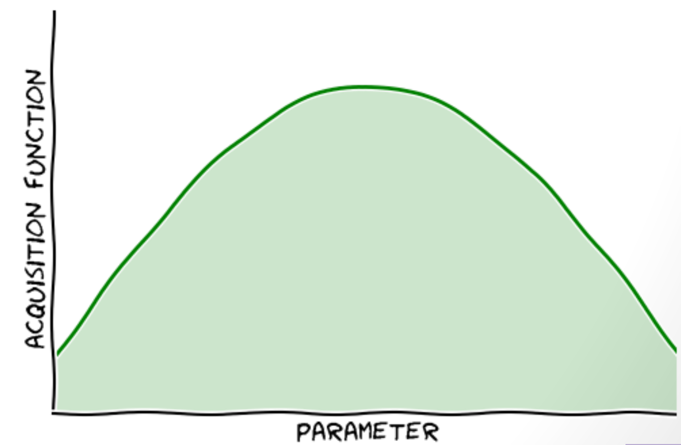
3. It uses only a fixed proposal distribution, not all information available

   ⇨ **Use Bayes' theorem to update the proposal of new points**

4. It aims at equal accuracy for all regions in parameter space

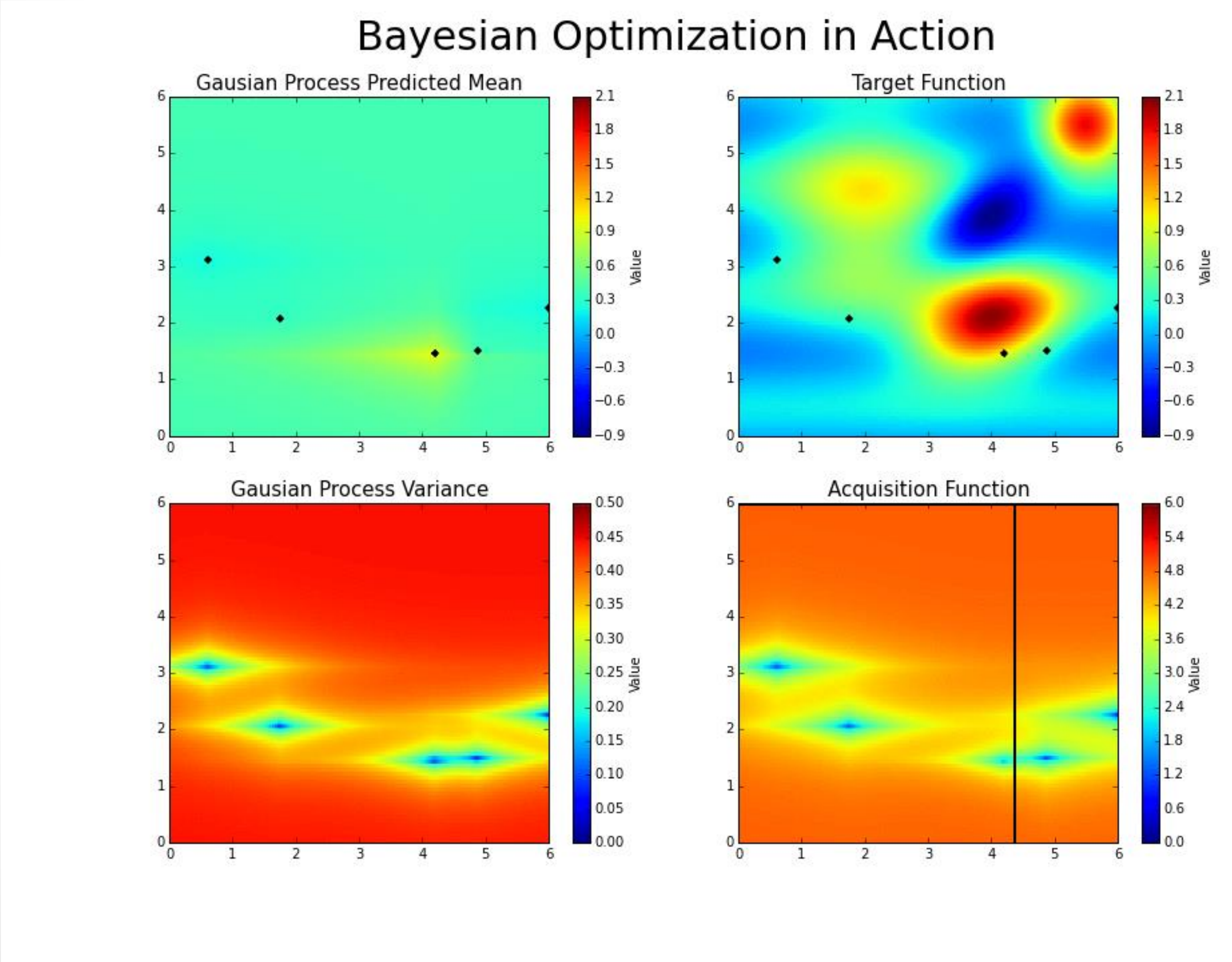
   ⇨ **Prioritize parameter regions with small distances to the observed data**





Gutmann & Corander JMLR 2016, arXiv:1501.03291

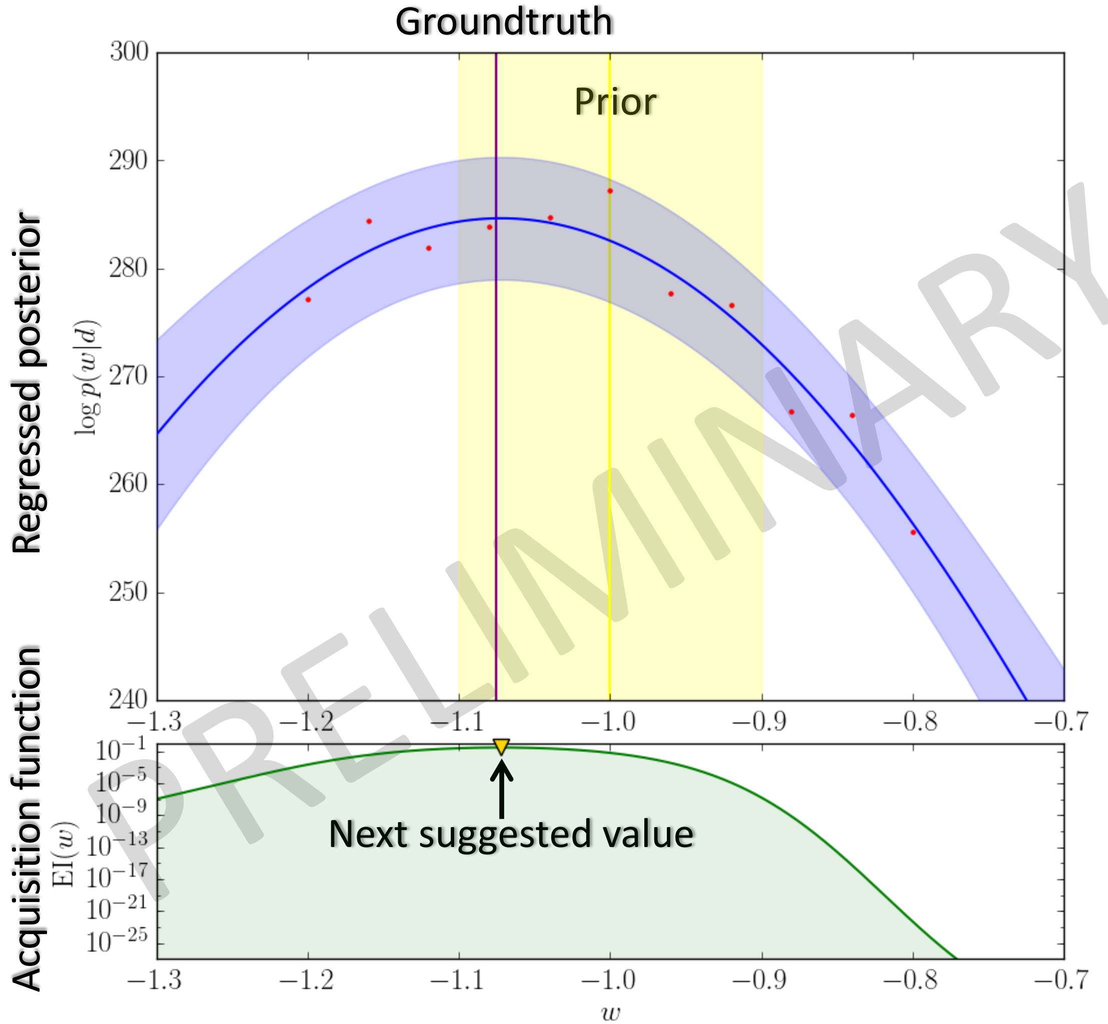
Demonstration in 2D



F. Nogueira, https://github.com/fmfn/BayesianOptimization

Likelihood-free large-scale structure inference

- Inference of the equation of state of dark energy
- Distance function using the power spectrum
- 1100 large-scale structure simulations
- $\approx 10^7$ hidden variables



This proof-of-concept has been performed completely blindly.

with W. Enzi & J. Jasche

Optimising the Data Model with sCOLA

tCOLA: *COmoving Lagrangian Acceleration (temporal domain)*

- Write the displacement vector as: $\mathbf{s} = \mathbf{s}_{\mathrm{LPT}} + \mathbf{s}_{\mathrm{MC}}$

Tassev & Zaldarriaga 2012, arXiv:1203.5785

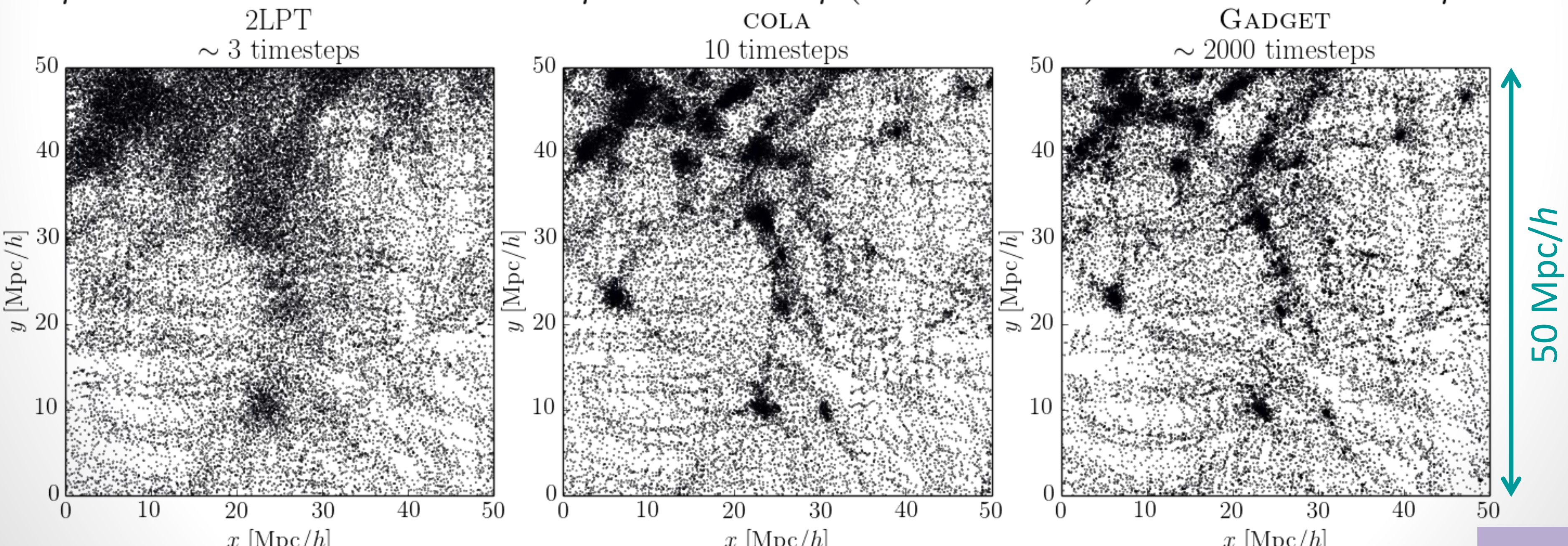
- Time-stepping (omitted constants and Hubble expansion):

Standard:

$$\partial_\tau^2 \mathbf{s} = -\nabla\Phi$$

Modified:

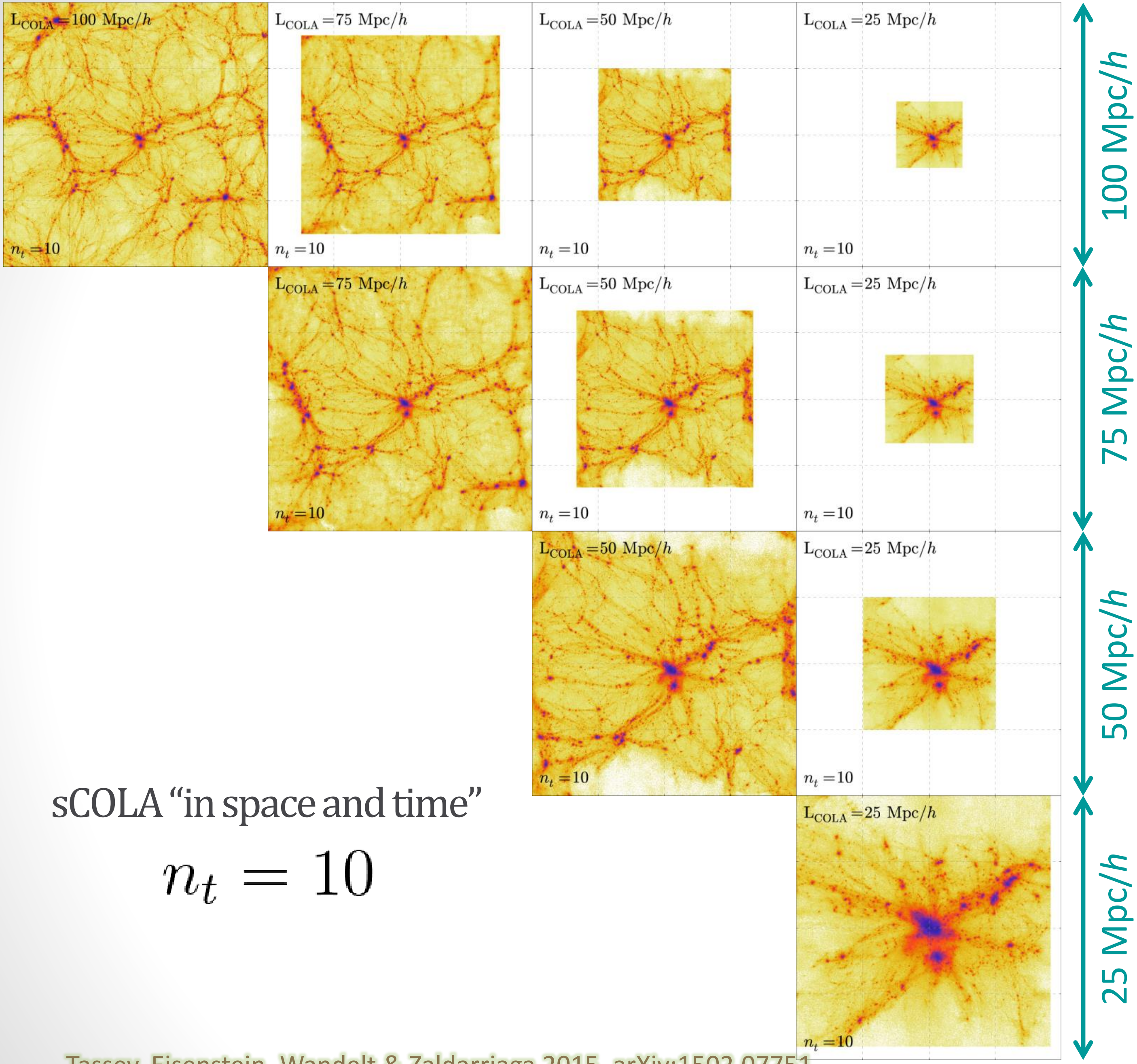
$$\partial_\tau^2 \mathbf{s}_{\mathrm{MC}} = \partial_\tau^2 (\mathbf{s} - \mathbf{s}_{\mathrm{LPT}}) = -\nabla\Phi - \partial_\tau^2 \mathbf{s}_{\mathrm{LPT}}$$



Tassev, Zaldarriaga & Einsenstein 2013, arXiv:1301.0322

sCOLA:

Extension to the spatial domain

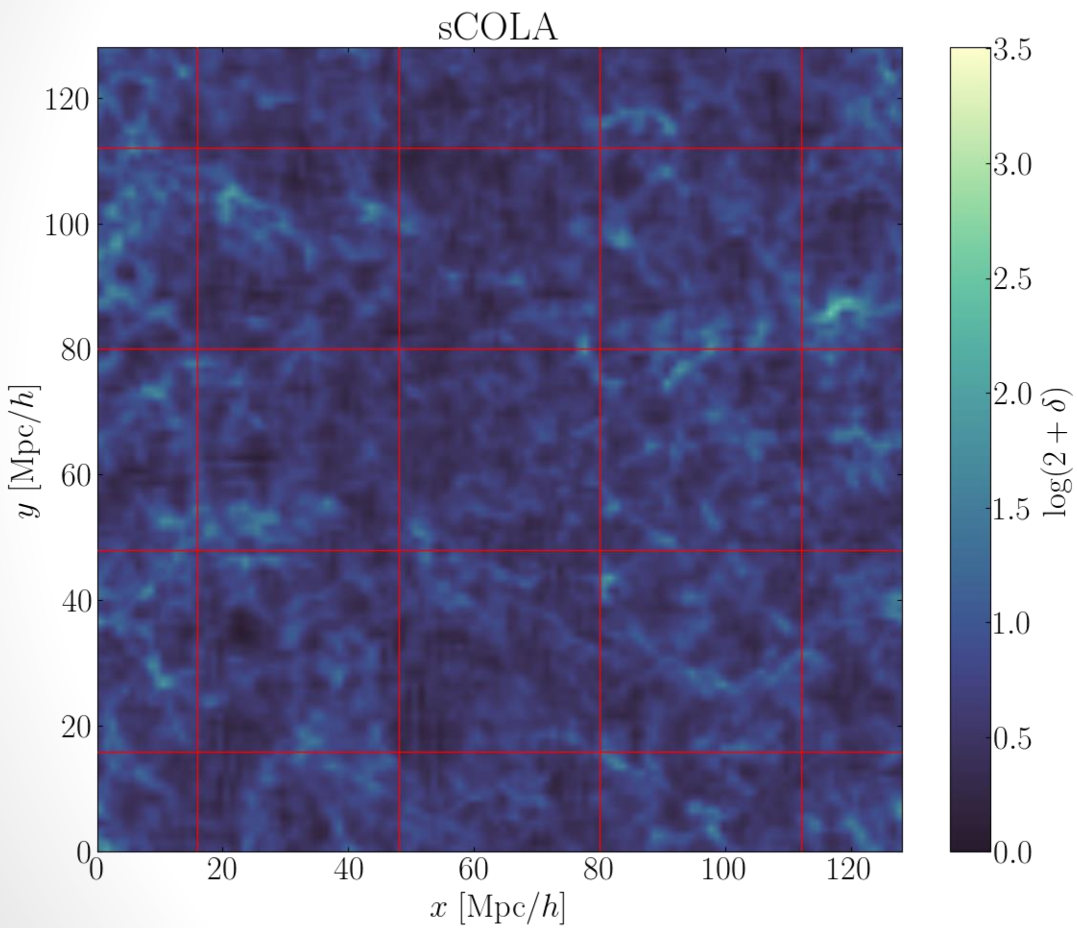


sCOLA "in space and time"

$$n_t = 10$$

Tassev, Eisenstein, Wandelt & Zaldarriaga 2015, arXiv:1502.07751

Using sCOLA to parallelize N-body sims



Parallelisation potential:

- Subvolumes…
  - do not need to communicate,
  - can even be run out of order!
- Factor ~ 8 overhead due to boundary regions.
- But $\sim 50\ \mathrm{Mpc}/h$ N-body sims can be done **in cache or on a GPU**.

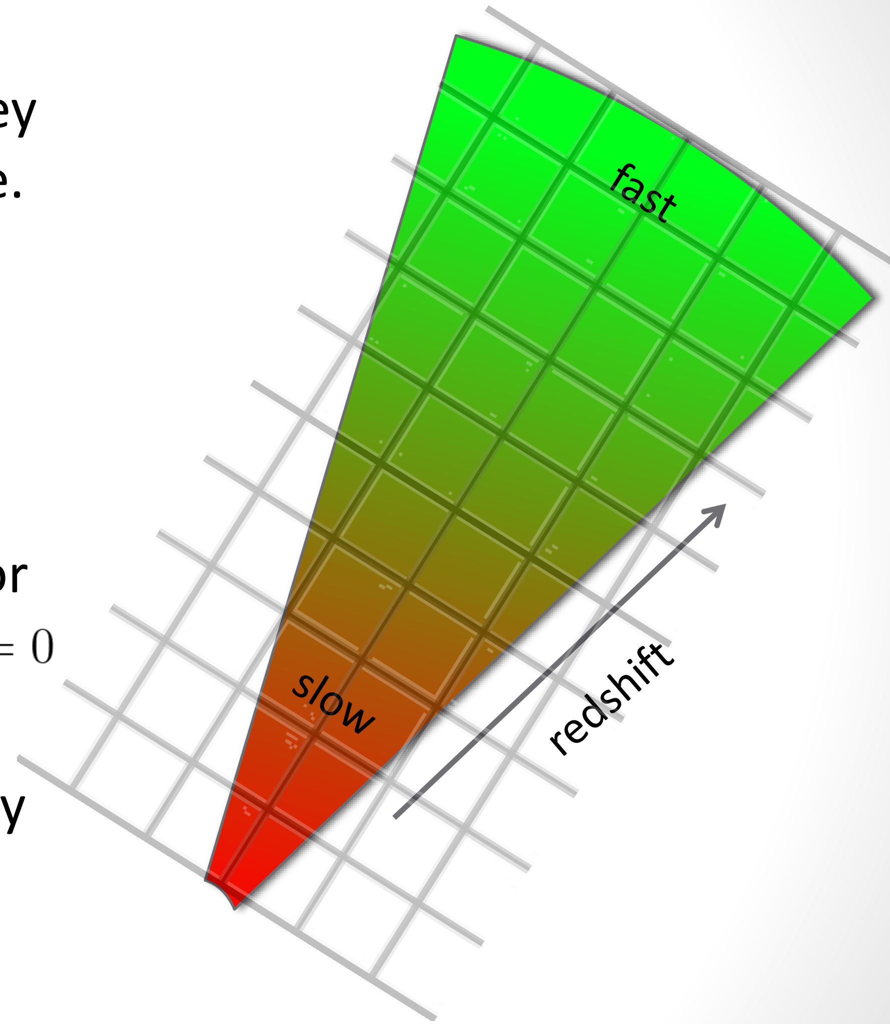
  ⇨ speed-up of s
- Potential parallelisation speed-up:

$$\frac{1}{8} \times s \times \left(\frac{10\ \mathrm{Gpc}/h}{50\ \mathrm{Mpc}/h}\right)^3 = s \times 10^6$$

with B. Faure (master project), B. Wandelt, W. Percival & M. Zaldarriaga

Constructing lightcones

- Subvolumes only need to run until they intersect the observer's past lightcone.
- Most of the high-z volume will be faster than $z = 0$.
- Many unobserved subvolumes do not even have to run!
- The wall-clock time limit is the time for running a single $\sim 50\ \mathrm{Mpc}/h$ box to $z = 0$ at the observer position.
- Leads to **further speed-up**, especially for deep surveys.



with B. Faure (master project), B. Wandelt, W. Percival & M. Zaldarriaga

Summary

- A likelihood-based method for principled analysis of galaxy surveys:
  Bayesian large-scale structure inference (BORG)
  - Simultaneous analysis of the morphology and formation history of the large-scale structure.
  - Characterization of the dynamic cosmic web underlying galaxies.
- A likelihood-free method for models where the likelihood is intractable but simulating is possible:
  Regression of the distance + Bayesian optimisation
  - Number of required simulations reduced by several orders of magnitude.
  - The approach will allow to **ask targeted questions to cosmological data**, including all relevant physical and observational effects.
- Optimisation of the data model using **tCOLA + sCOLA**
  - Enormous parallelisation potential for dark matter simulations.
  - Further speed-up expected for realistic synthetic observations.