Bayesian large-scale structure inference

Likelihood-based and likelihood-free approaches

Florent Leclercq

www.florent-leclercq.eu

Imperial College Research Fellow
Imperial Centre for Inference and Cosmology

January 31st, 2018

In collaboration with:

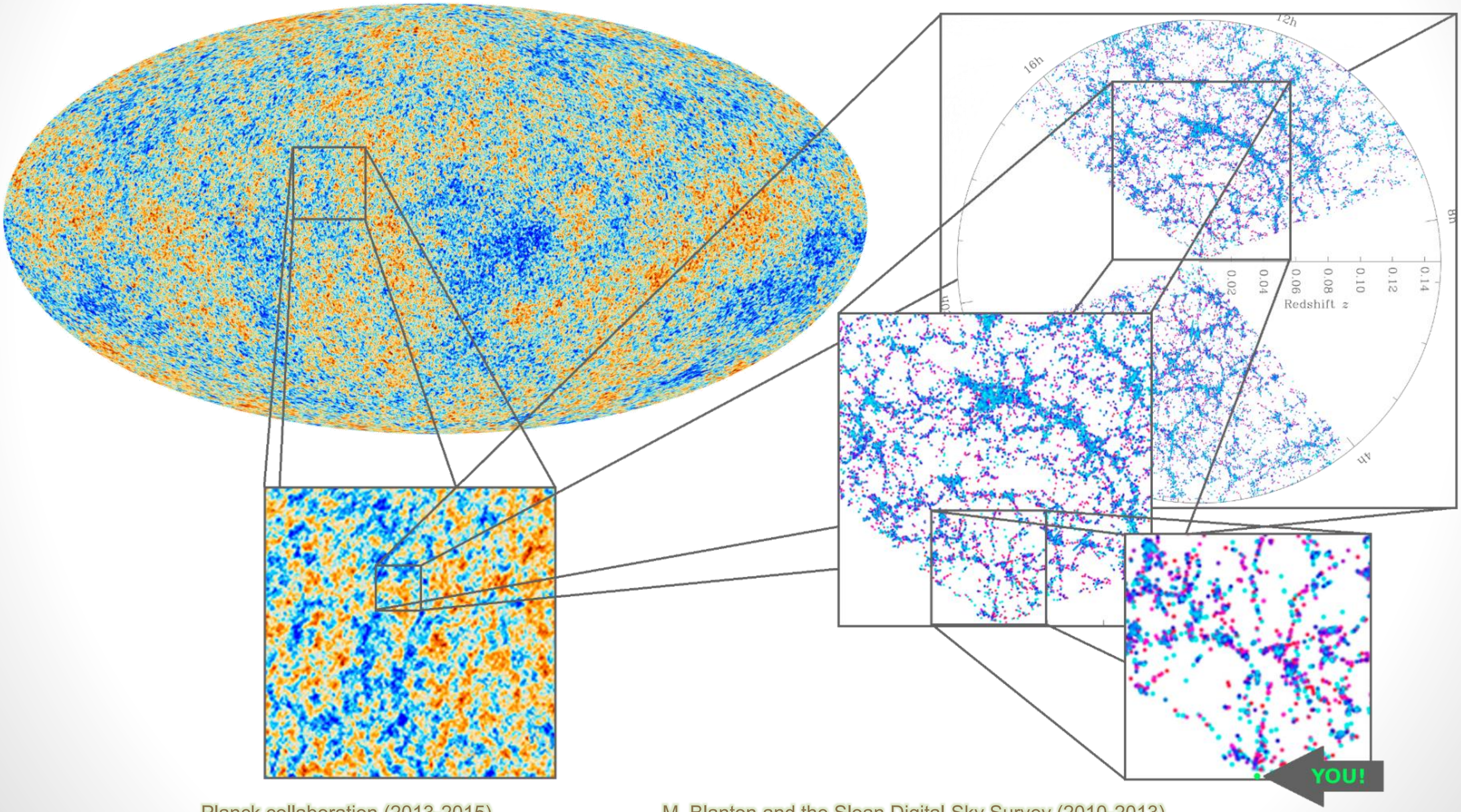
Wolfgang Enzi (MPA), **Jens Jasche** (ExC Garching/U. Stockholm), **Guilhem Lavaux** (IAP), **Will Percival** (U. Portsmouth), **Benjamin Wandelt** (IAP/CCA)

ICIC Imperial Centre for Inference & Cosmology

Imperial College London

The big picture: the Universe is highly structured

You are here. Make the best of it…



Planck collaboration (2013-2015)

M. Blanton and the Sloan Digital Sky Survey (2010-2013)

What we want to know from the LSS

Y. Dubois & S. Colombi (IAP)

The LSS is a vast source of knowledge:

- **Cosmology**:
  - Cosmological parameters and tests of ΛCDM,
  - Physical nature of the dark components,
  - Geometry of the Universe,
  - Tests of General Relativity,
  - Initial conditions and link to high energy physics
- **Astrophysics**: galaxy formation and evolution as a function of their environment
  - Galaxy properties (colours, chemical composition, shapes),
  - Intrinsic alignments

Why Bayesian inference?

- Inference of signals = ill-posed problem
  - Incomplete observations: finite resolution, survey geometry, selection effects
  - Noise, biases, systematic effects
  - Cosmic variance

➡ **No unique recovery is possible!**

"What is the formation history of the Universe?" ➡ "What is the probability distribution of possible formation histories (signals) compatible with the observations?"

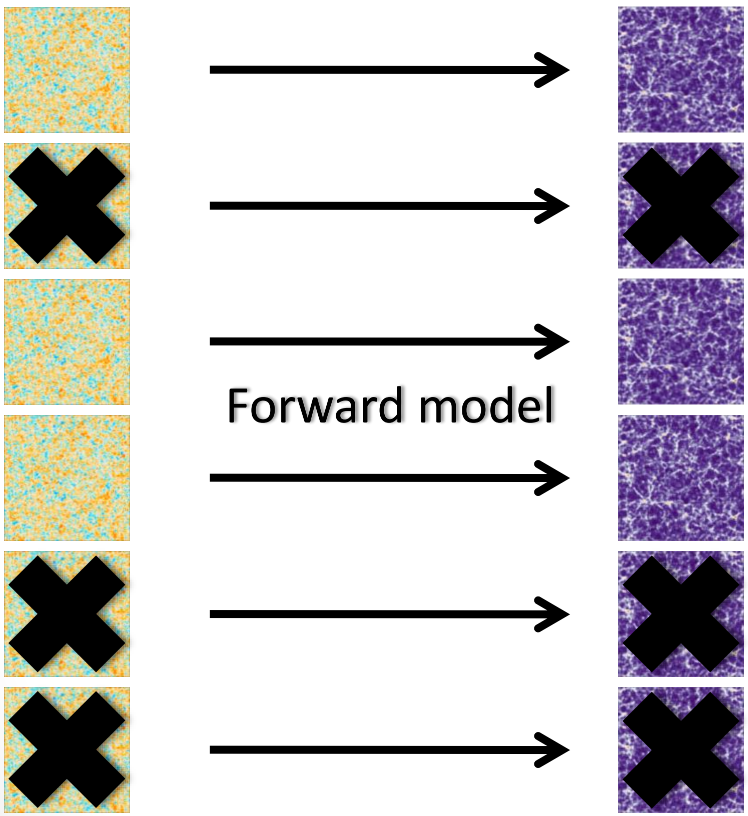
Bayes' theorem: $\mathcal{P}(s|d)\mathcal{P}(d) = \mathcal{P}(d|s)\mathcal{P}(s)$

- Cox-Jaynes theorem: Any system to manipulate "*plausibilities*", consistent with Cox's desiderata, is isomorphic to **(Bayesian) probability theory**

➡ How to do that?

Bayesian forward modeling: the ideal scenario

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

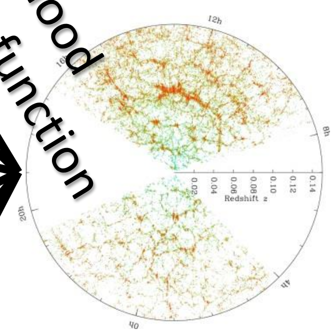


Forward model

All possible ICs

All possible FCs

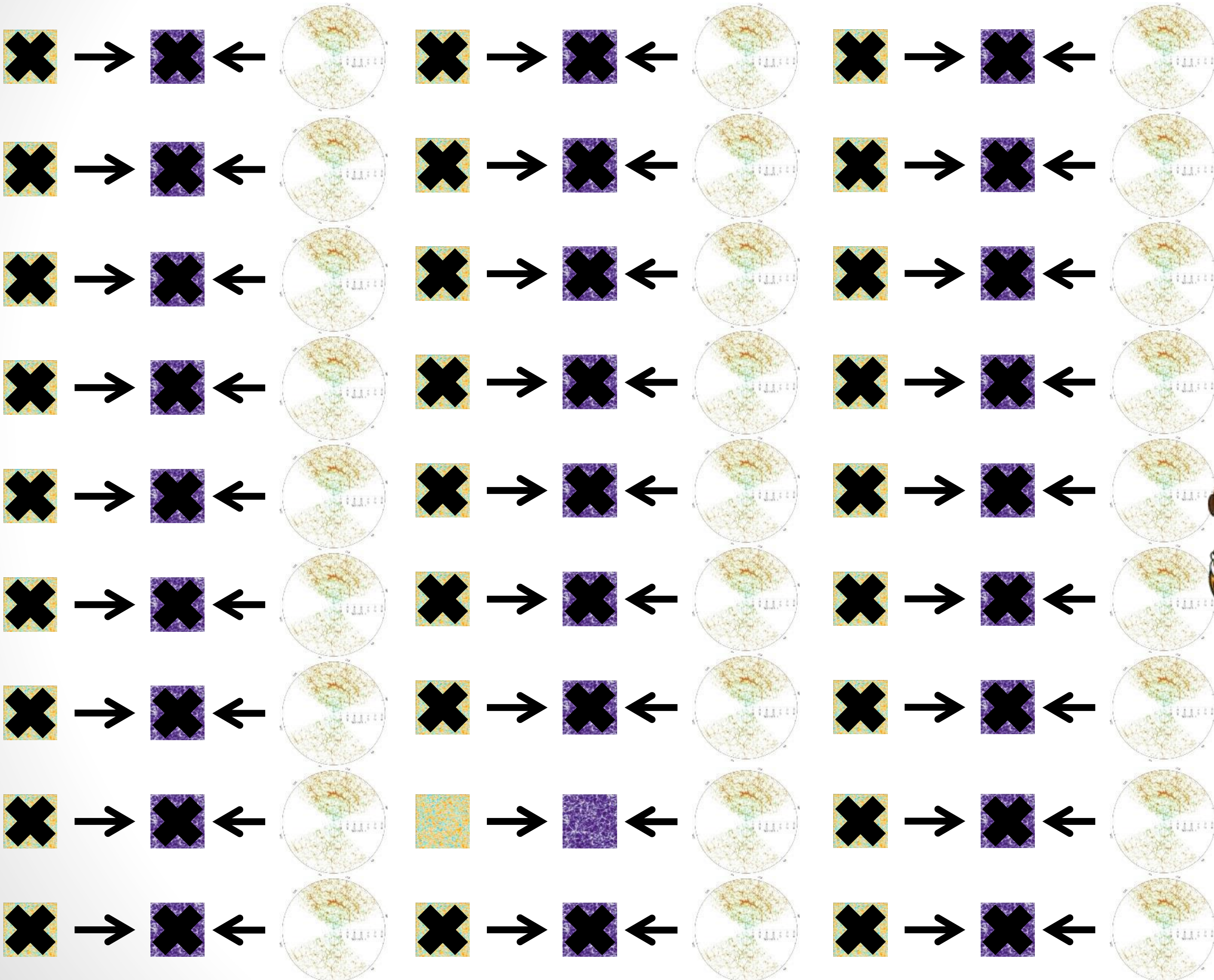
Likelihood or distance function



Observations

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

Bayesian forward modeling: the challenge



$d \approx 10^7$

Likelihood-Based Solution: BORG

Exact statistical inference
Approximate physical model

?

Hamiltonian (Hybrid) Monte Carlo

- Use classical mechanics to solve statistical problems!
  - The potential: $\psi(\mathbf{x}) \equiv -\ln p(\mathbf{x})$
  - The Hamiltonian: $H(\mathbf{x},\mathbf{p}) \equiv \frac{1}{2}\mathbf{p}^\intercal \mathbf{M}^{-1}\mathbf{p} + \psi(\mathbf{x})$

$$(\mathbf{x},\mathbf{p}) \Rightarrow \left\{ \begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\frac{\partial H}{\partial \mathbf{x}} = -\frac{\mathrm{d}\psi(\mathbf{x})}{\mathrm{d}\mathbf{x}} \end{aligned} \right\} \Rightarrow (\mathbf{x}',\mathbf{p}')$$

gradients of the pdf

$$a(\mathbf{x}',\mathbf{x}) = \mathrm{e}^{-(H'-H)} = 1$$

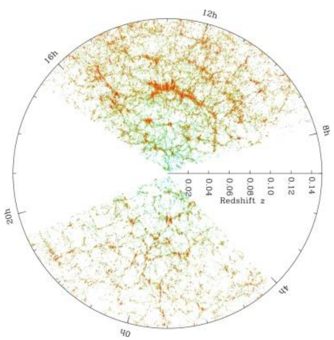
acceptance ratio unity

- HMC **beats the curse of dimensionality** by:
  - Exploiting gradients
  - Using conservation of the Hamiltonian

Duane *et al*. 1987, Phys. Lett. B **195**, 2

BORG: *Bayesian Origin Reconstruction from Galaxies*

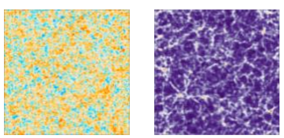
- **Sampler**: Hamiltonian Monte Carlo
- **Data model**:
  - Gaussian prior for the initial conditions
  - Second-order Lagrangian perturbation theory (2LPT)
  - Poisson likelihood



Observations

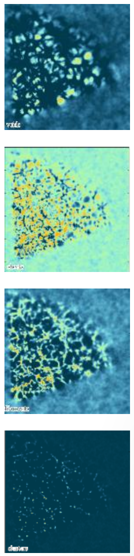
(galaxy catalog + meta-data: selection functions, completeness...)

→ BORG →



Inferred dark matter density

→



Cosmic web analysis

see also:

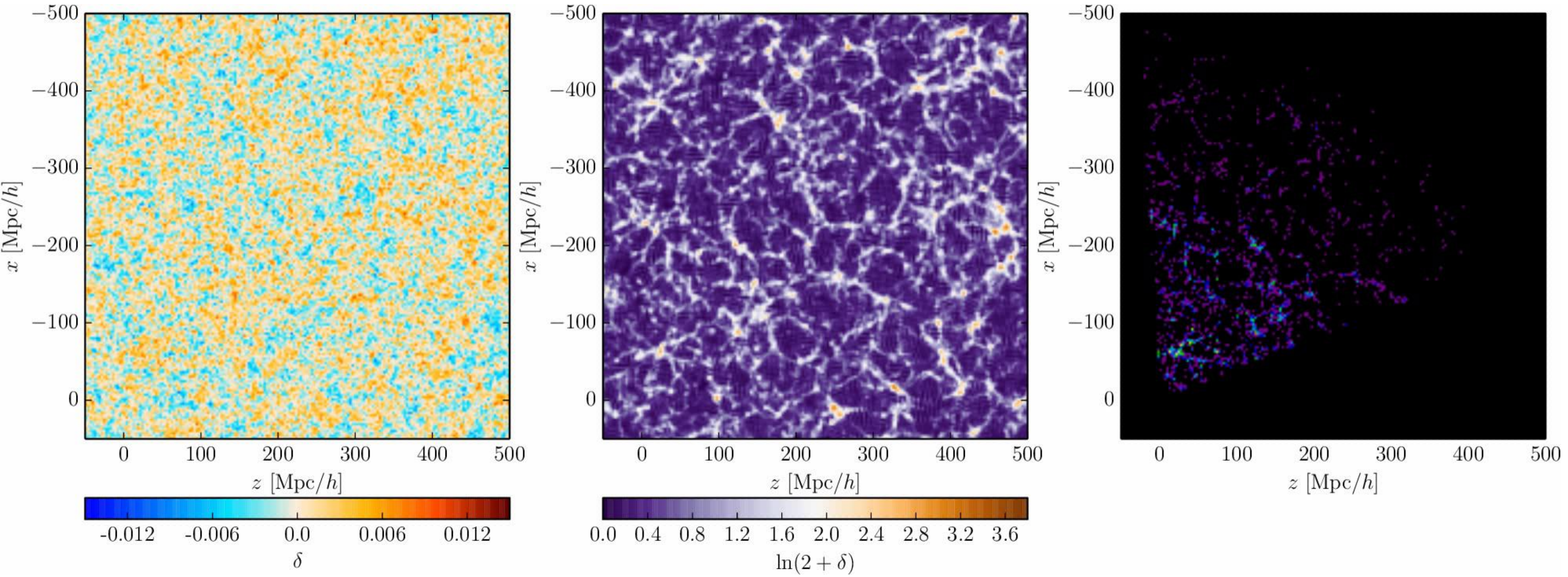
Kitaura 2013, arXiv:1203.4184

Wang, Mo, Yang & van den Bosch 2013, arXiv:1301.1348

Jasche & Wandelt 2013, arXiv:1203.3639

Likelihood-based solution: BORG at work

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

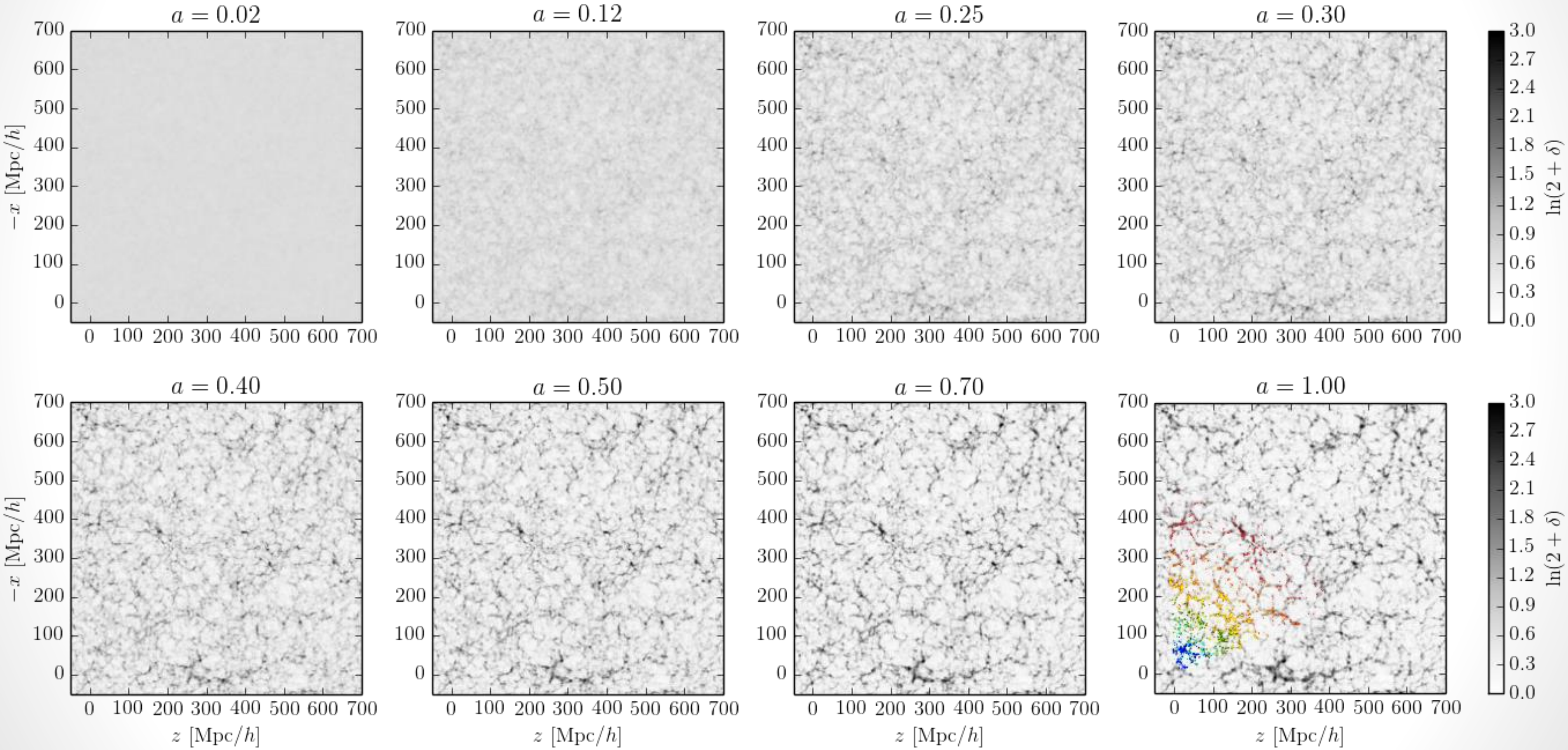


Initial conditions   Final conditions   Observations

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

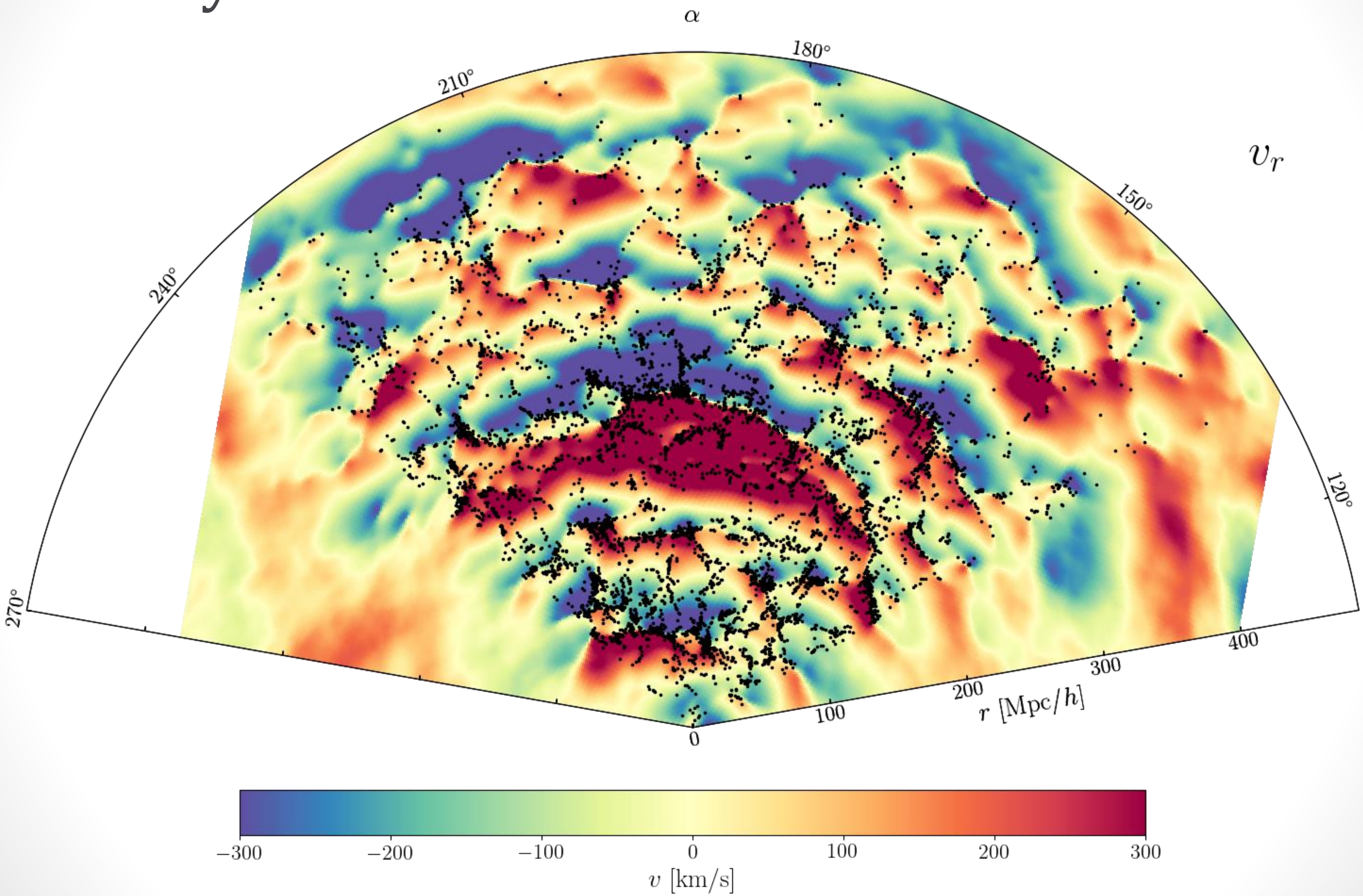
Jasche, FL & Wandelt 2015, arXiv:1409.6308

Evolution of cosmic structure



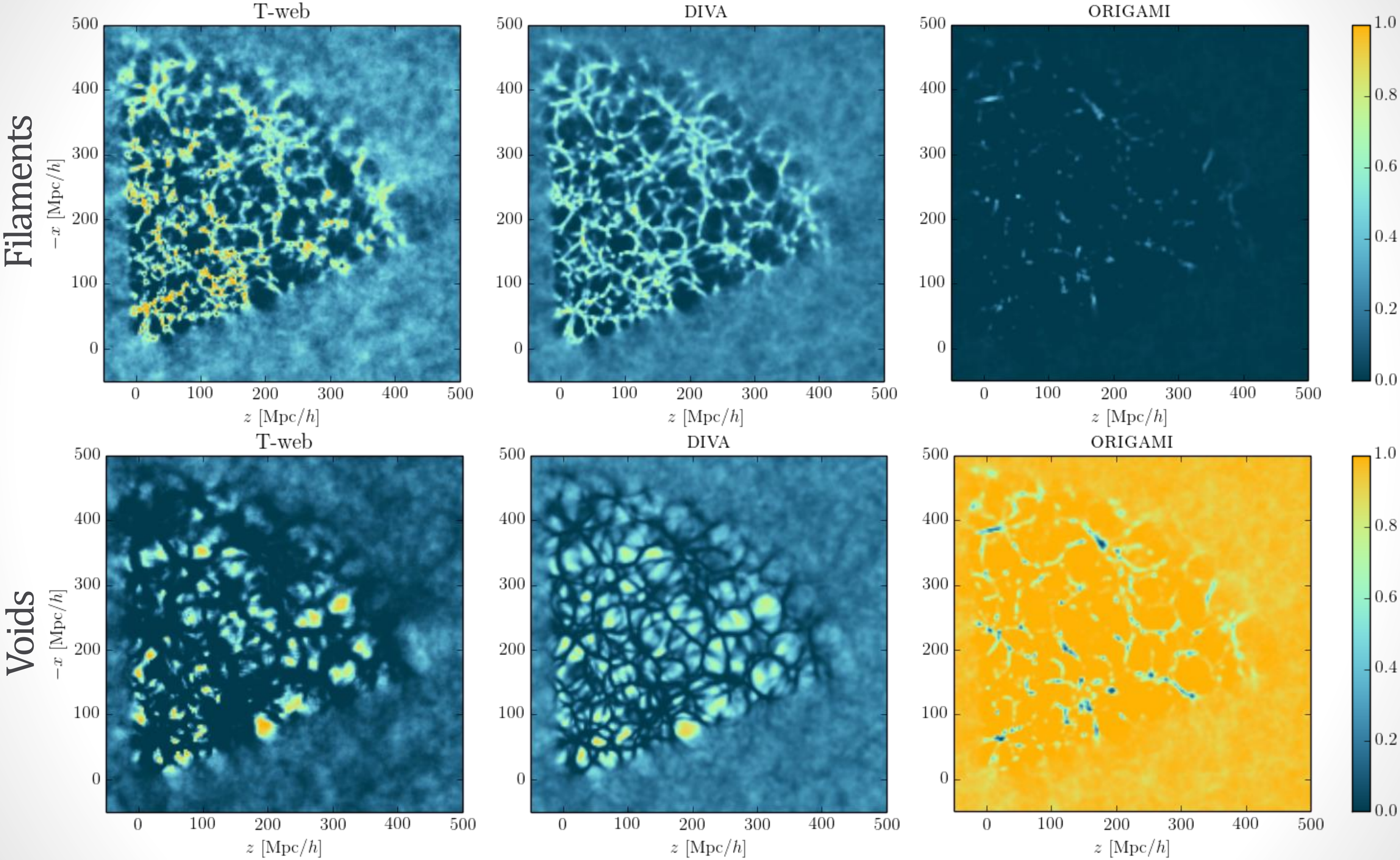
Jasche, FL & Wandelt 2015, arXiv:1409.6308

Velocity field



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

Cosmic web classifications



FL, Jasche & Wandelt 2015a, arXiv:1502.02690

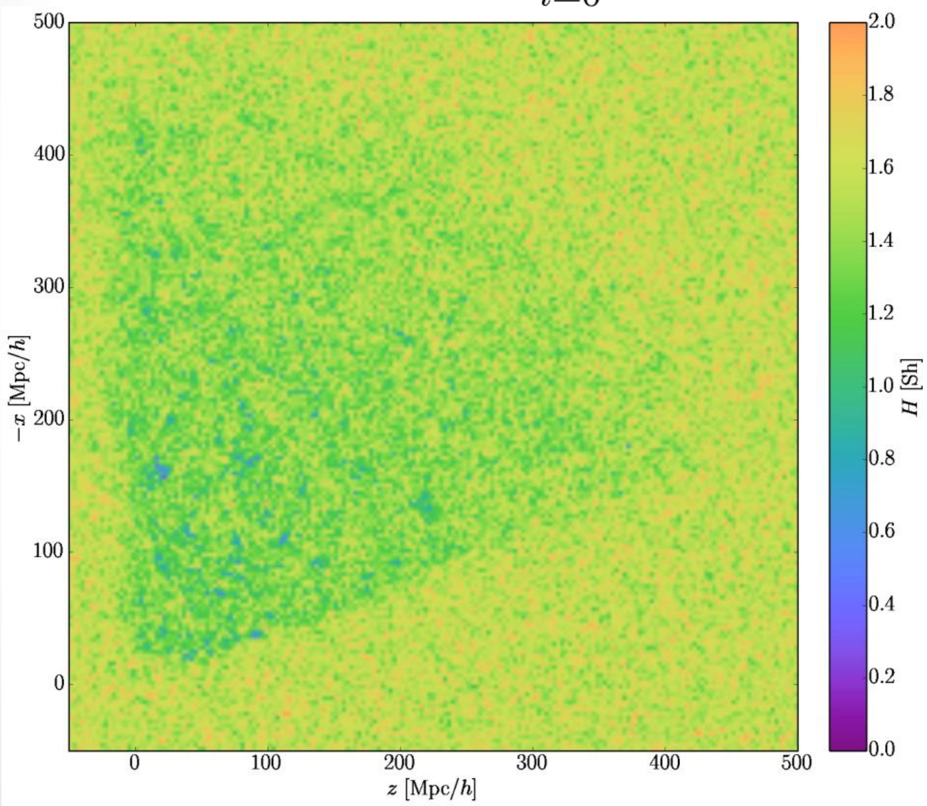
FL, Lavaux, Jasche & Wandelt 2016, arXiv:1606.06758

How is information propagated?

Shannon entropy

$$H\left[\mathcal{P}(\mathrm{T}(\vec{x}_k)|d)\right] \equiv -\sum_{i=0}^{3} \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)\log_2(\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d))$$

in shannons (Sh)



More about cosmic web analysis:

FL, Jasche & Wandelt 2015a, arXiv:1502.02690

(T-web, entropy, relative entropy)

FL, Jasche & Wandelt 2015b, arXiv:1503.00730

(decision theory for structure classification)

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

(mutual information, classifier utilities)

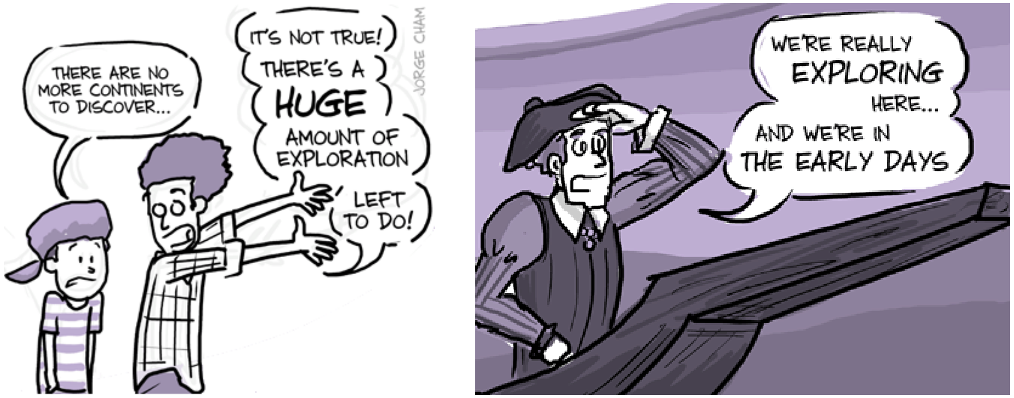
FL, Jasche, Lavaux, Wandelt & Percival 2017

(phase-space structure of dark matter)

FL, Jasche & Wandelt 2015a, arXiv:1502.02690

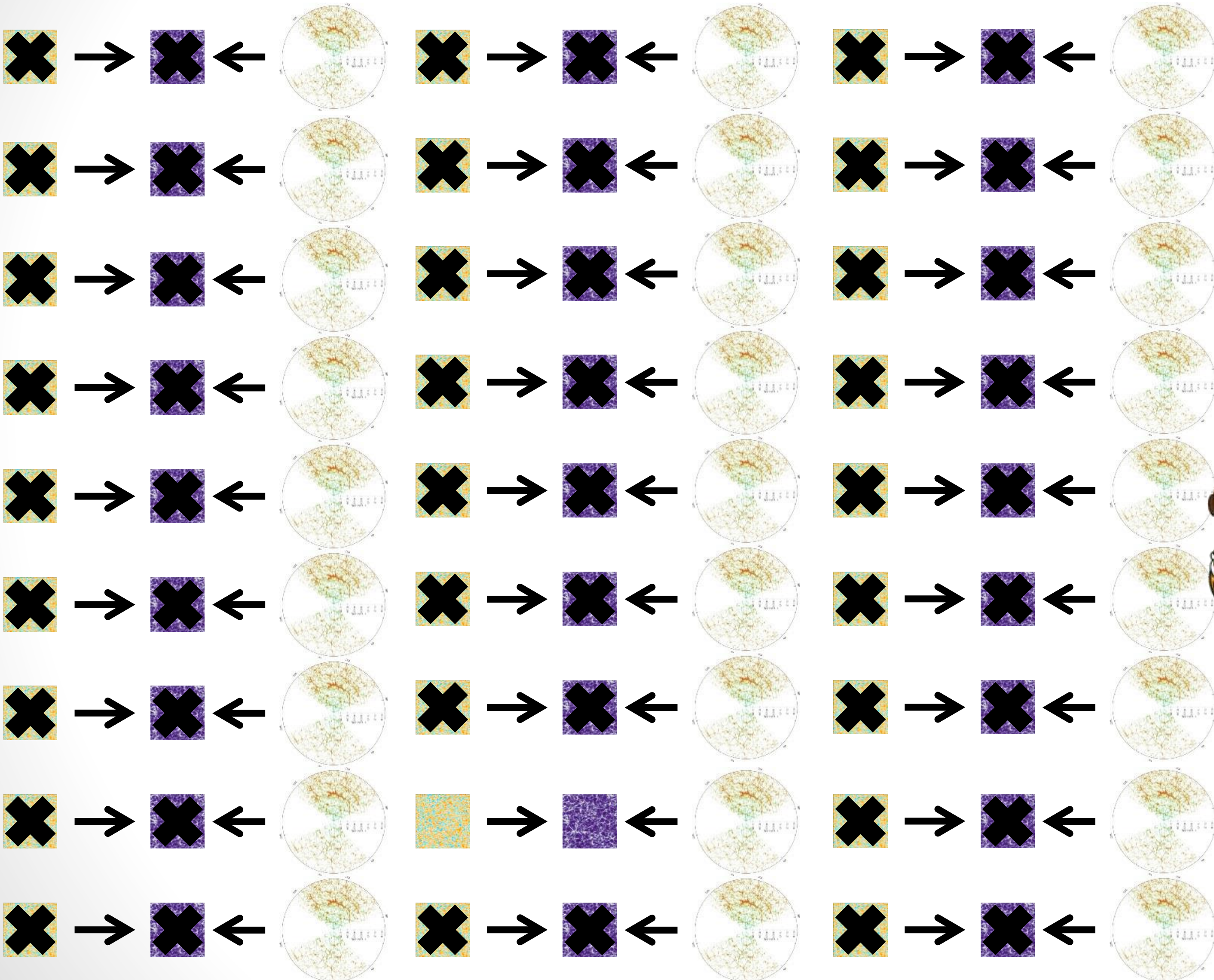
Interlude

Mapping the Universe: epilogue?



J. Cham – PhD comics

Let's go back to the challenge…



$d \approx 10^7$

LIKELIHOOD-FREE SOLUTION: BOLFI

?

Approximate statistical inference
Exact physical model

Approximate Bayesian Computation (ABC)

- Statistical inference for models where:
  1. The likelihood function is intractable
  2. Simulating data is possible

- **General idea**: find parameter values for which the distance between simulated data and observed data is small

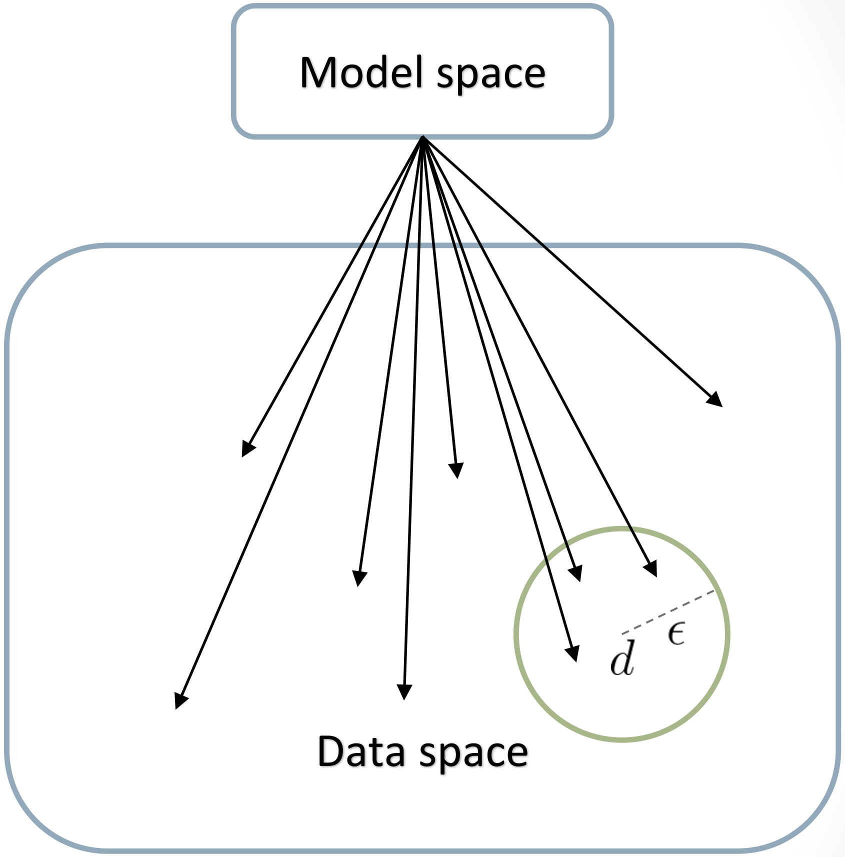
$$p(\theta|d) \Longrightarrow p(\theta|\tilde{d}) \quad \text{where } \mathrm{d}(\tilde{d}(\theta), d) \text{ is small}$$

- **Assumptions**:
  - Only a small number of parameters are of interest
  - But the process generating the data is very general: a noisy non-linear dynamical system with an unrestricted number of hidden variables

Likelihood-free rejection sampling

- Iterate many times:
  - Sample θ from a proposal distribution $q(\theta)$
  - Simulate $\tilde{d}(\theta)$ according to the data model
  - Compute distance $\mathrm{d}(\tilde{d}(\theta), d)$ between simulated and observed data
  - Retain θ if $\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon$, otherwise reject
- Effective likelihood approximation:

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



ϵ can be adaptively reduced (Population Monte Carlo)

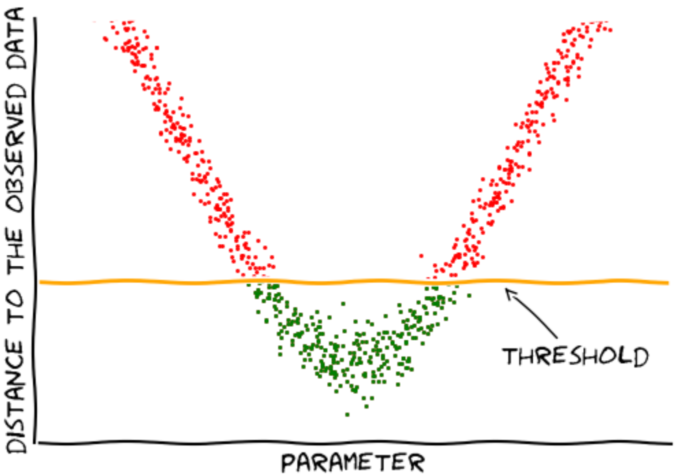
Why is likelihood-free rejection so expensive?

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



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

Proposed solution:

BOLFI: *Bayesian Optimisation for Likelihood-Free Inference*

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**

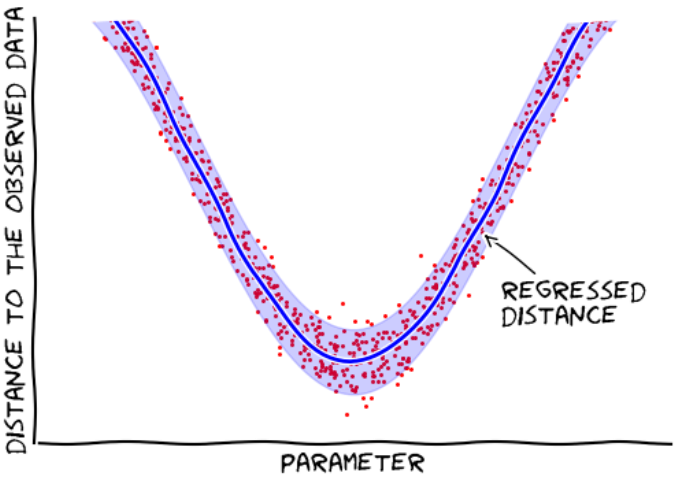
3. It uses only a fixed proposal distribution, not all information available

   ⇨ **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



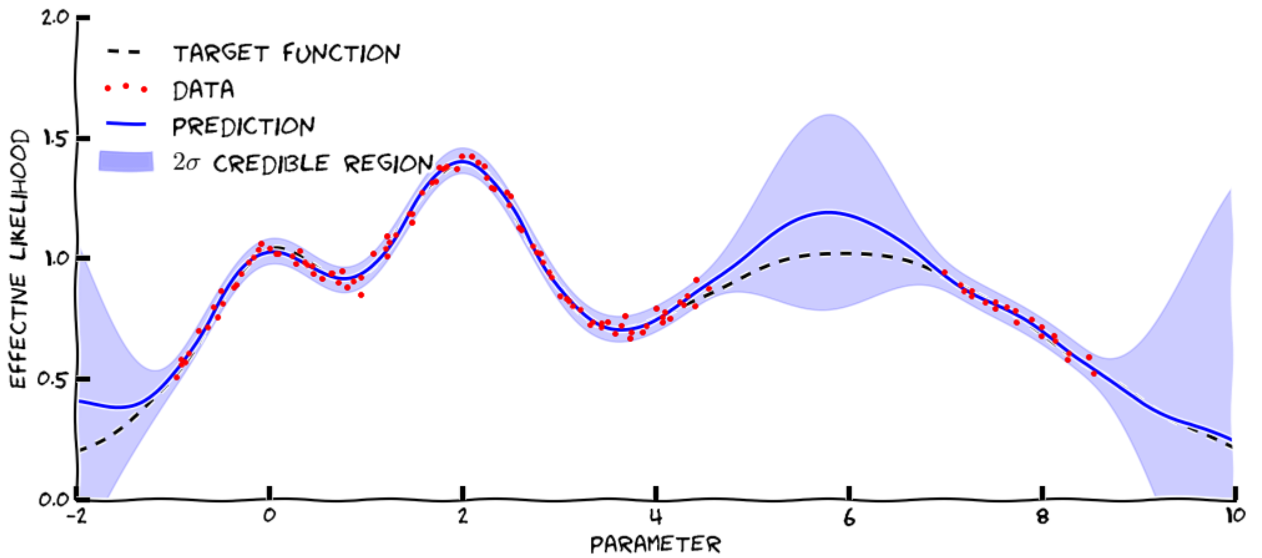
Related work in cosmology:

Alsing & Wandelt 2017, arXiv:1712.00012
(data compression for ABC)

Alsing, Wandelt & Feeney 2018, arXiv:1801.01497
(density estimation for ABC – DELFI)

Enzi, Jasche & FL 2018, to be submitted
(ABC with linear expansion of the effective likelihood)

Regressing the effective likelihood (points 1 & 2)



1. "It rejects most samples when ϵ is small"

- Keep all values (θ_i, d_i) $\quad$ $\mathrm{d}_i = \mathrm{d}(\tilde{d}(\theta_i), d)$

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

- Model the conditional distribution of distances given this training set

Data acquisition (points 3 & 4)

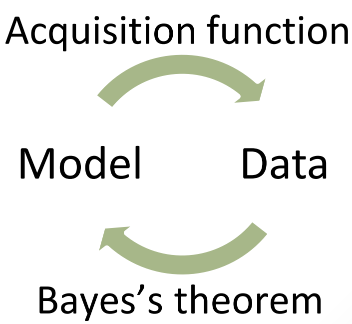
3. "It uses only a fixed proposal distribution, not all information available"

- Samples are obtained from sampling an **adaptively-constructed proposal distribution**, using the regressed effective likelihood

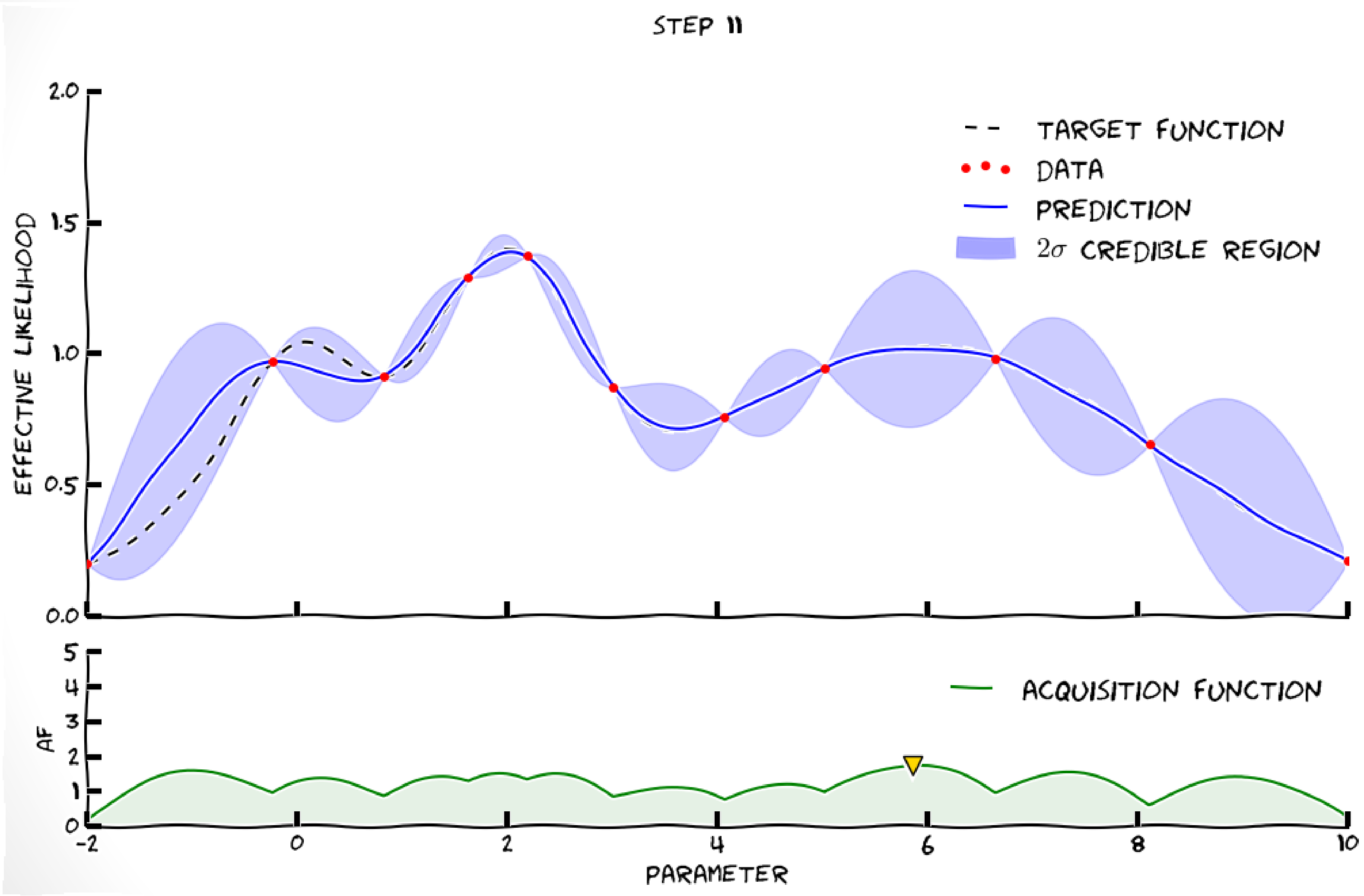
4. "It aims at equal accuracy for all regions in parameter space"

- The **acquisition function** finds a compromise between exploration (trying to find new high-likelihood regions) & exploitation (giving priority to regions where the distance to the observed data is already known to be small)

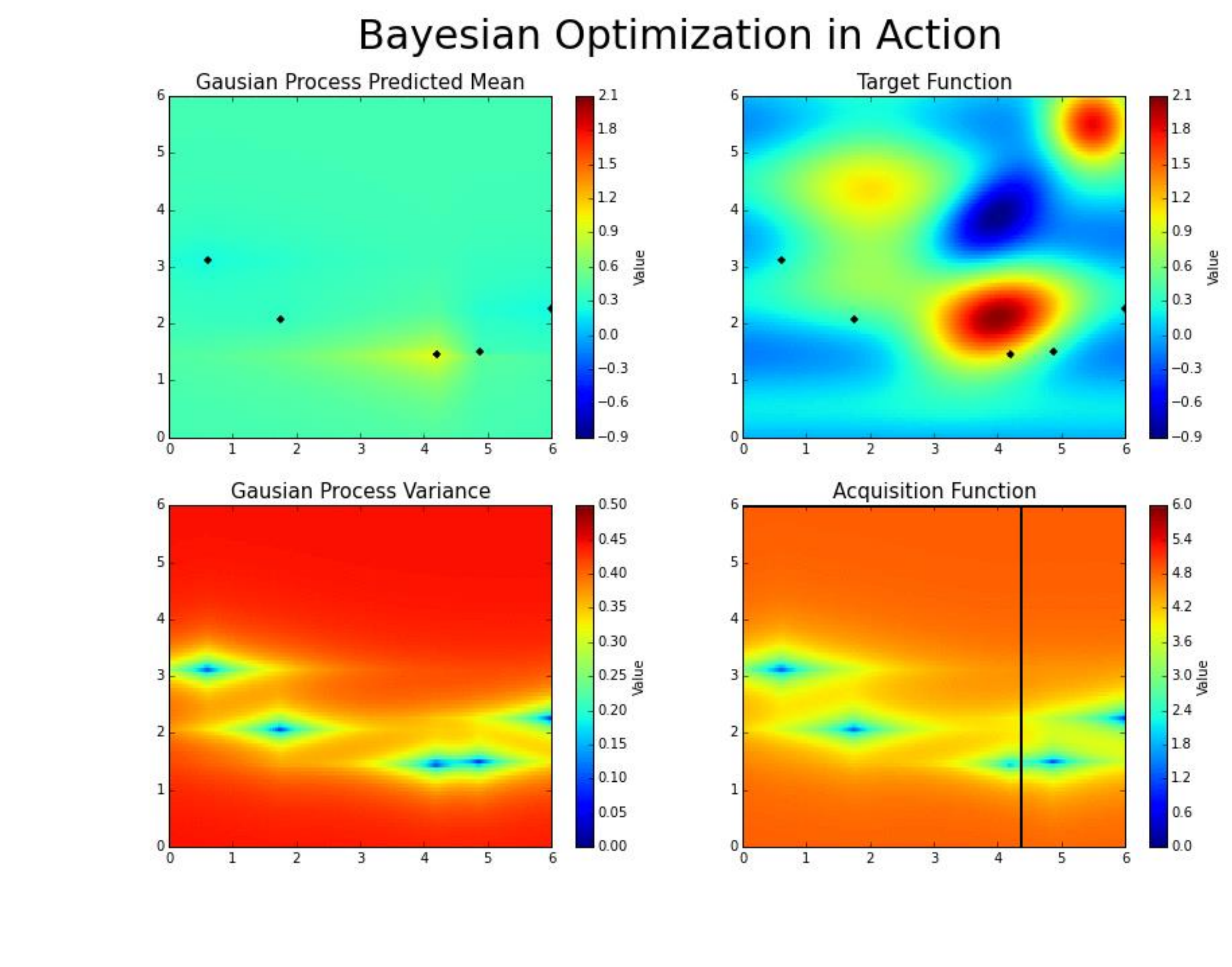
- **Bayesian optimisation** (decision making under uncertainty) can then be used



Data acquisition



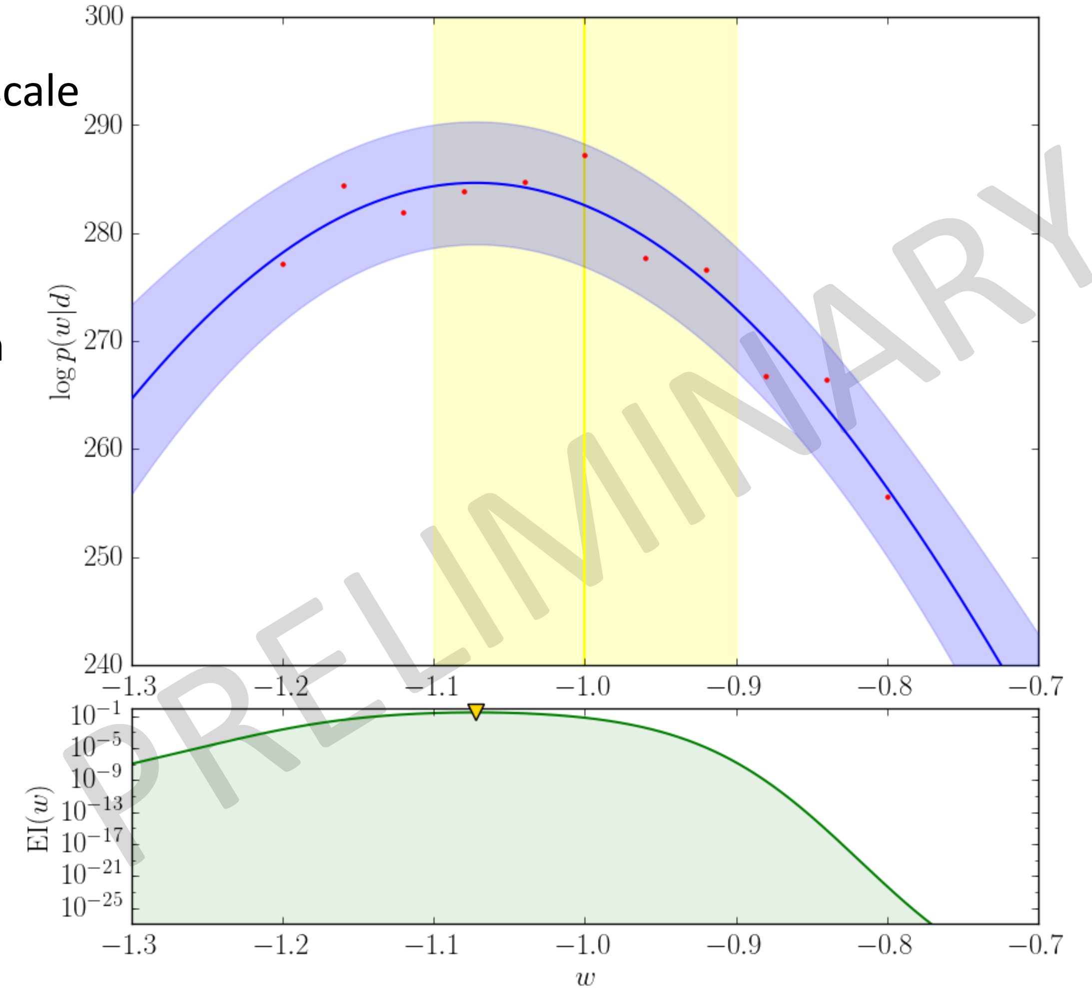
In higher dimension…



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

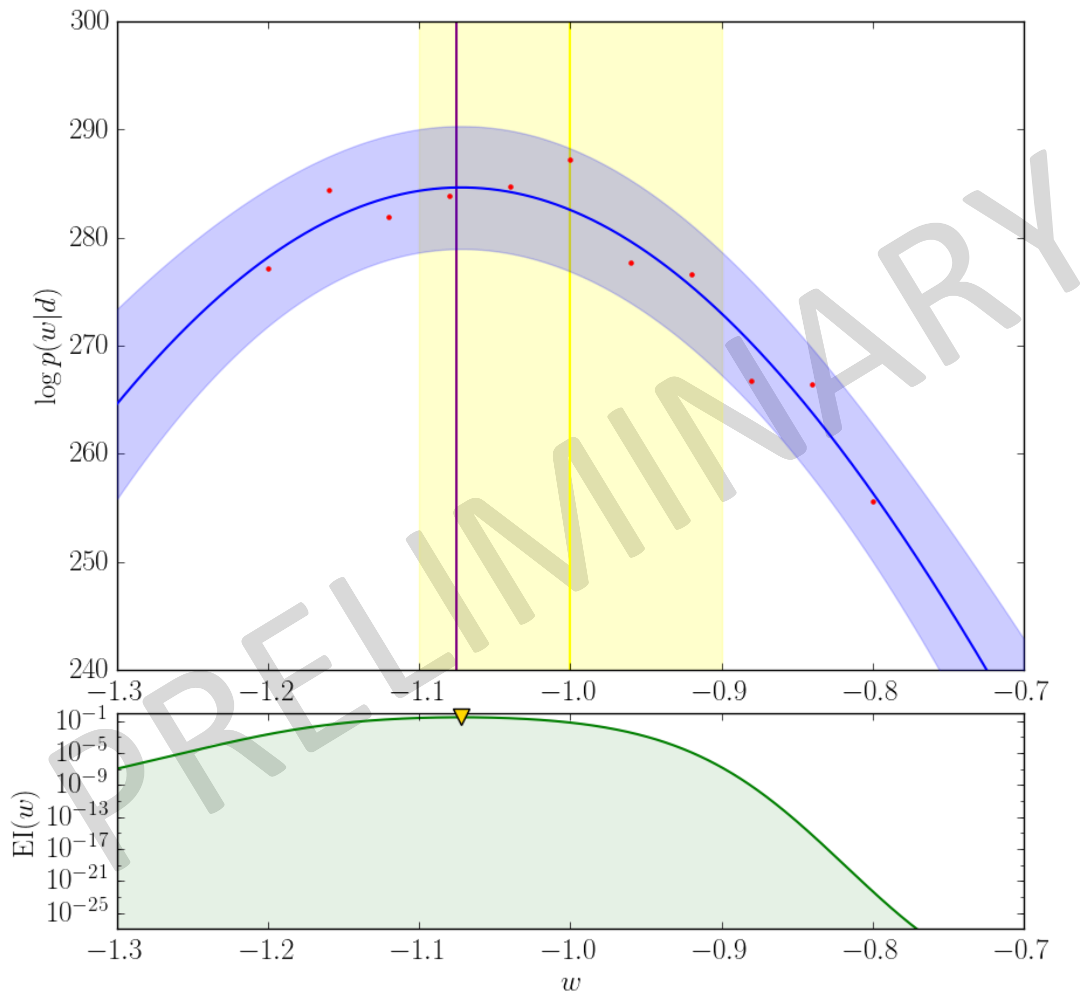
Likelihood-free large-scale structure inference

- 1100 large-scale structure simulations using COLA
- ≈10^7 hidden variables



FL, Jasche & Enzi (in prep.)

Likelihood-free large-scale structure inference



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

FL, Jasche & Enzi (in prep.)

Summary

Bayesian large-scale structure inference

Exact statistical inference
Approximate physical model

?

Approximate statistical inference
Exact physical model

- A likelihood-based method for principled analysis of galaxy surveys:
  Hamiltonian Monte Carlo (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 (BOLFI)
  - 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.