Inference with generative cosmological models

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

June 20th, 2018

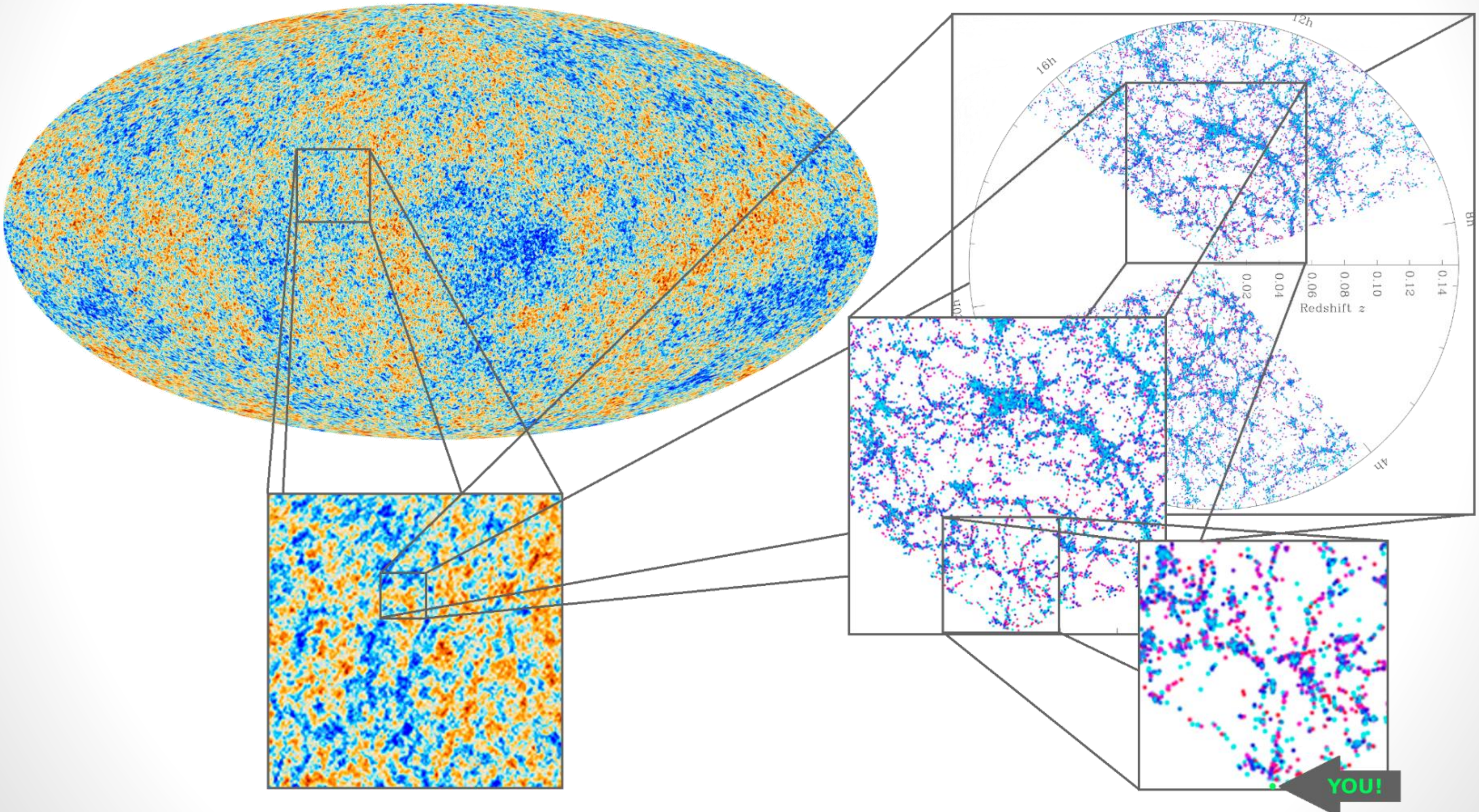
with the Aquila Consortium
www.aquila-consortium.org

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…

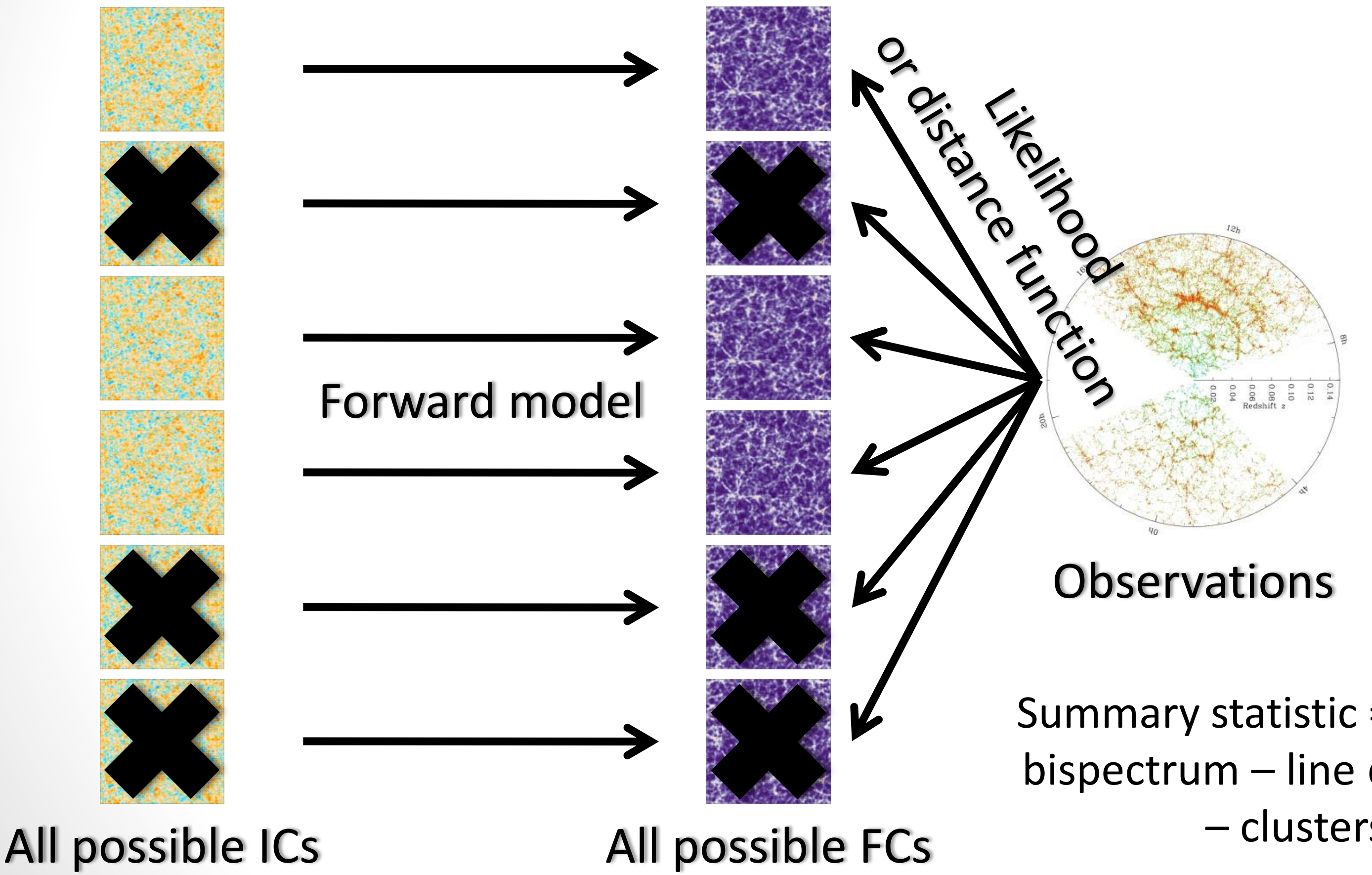


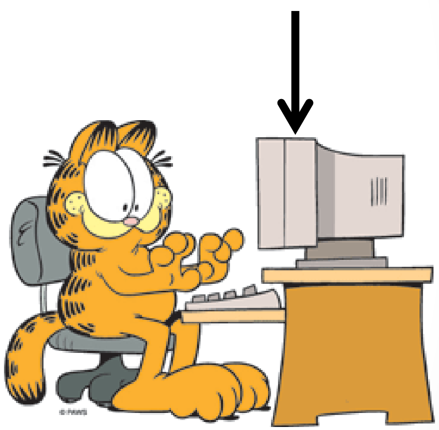
Planck collaboration (2013-2015)

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

Bayesian forward modeling: the ideal scenario

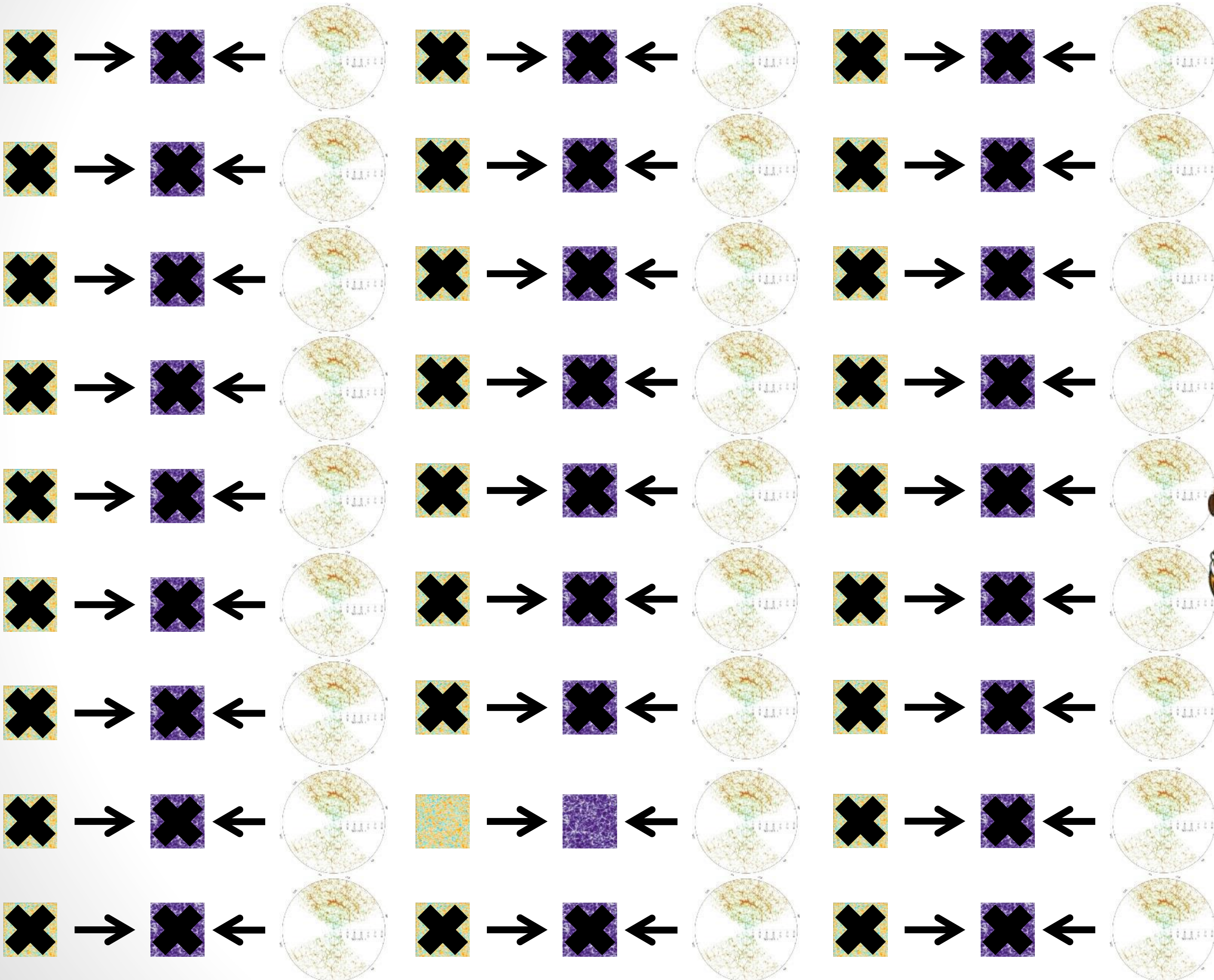
Forward model = N-body simulation + Halo occupation +
Galaxy formation + Feedback + …





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

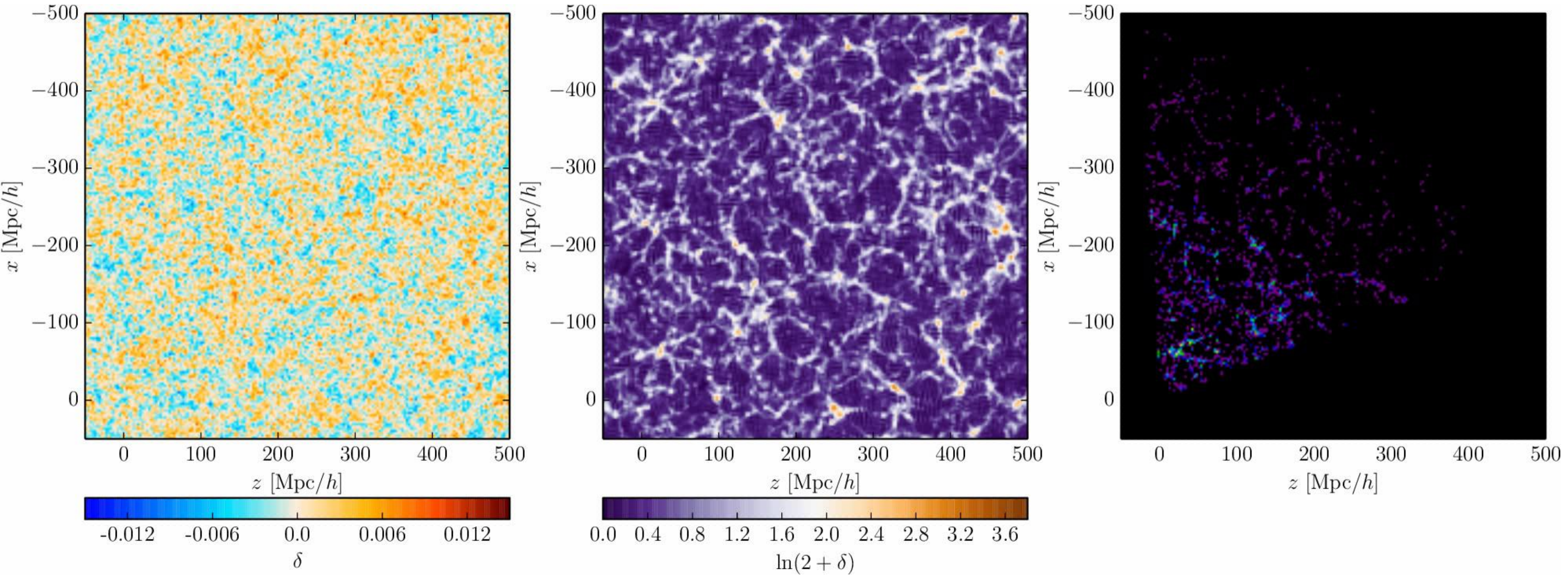
Bayesian forward modeling: the challenge



$d \approx 10^7$

Likelihood-based solution: BORG at work

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

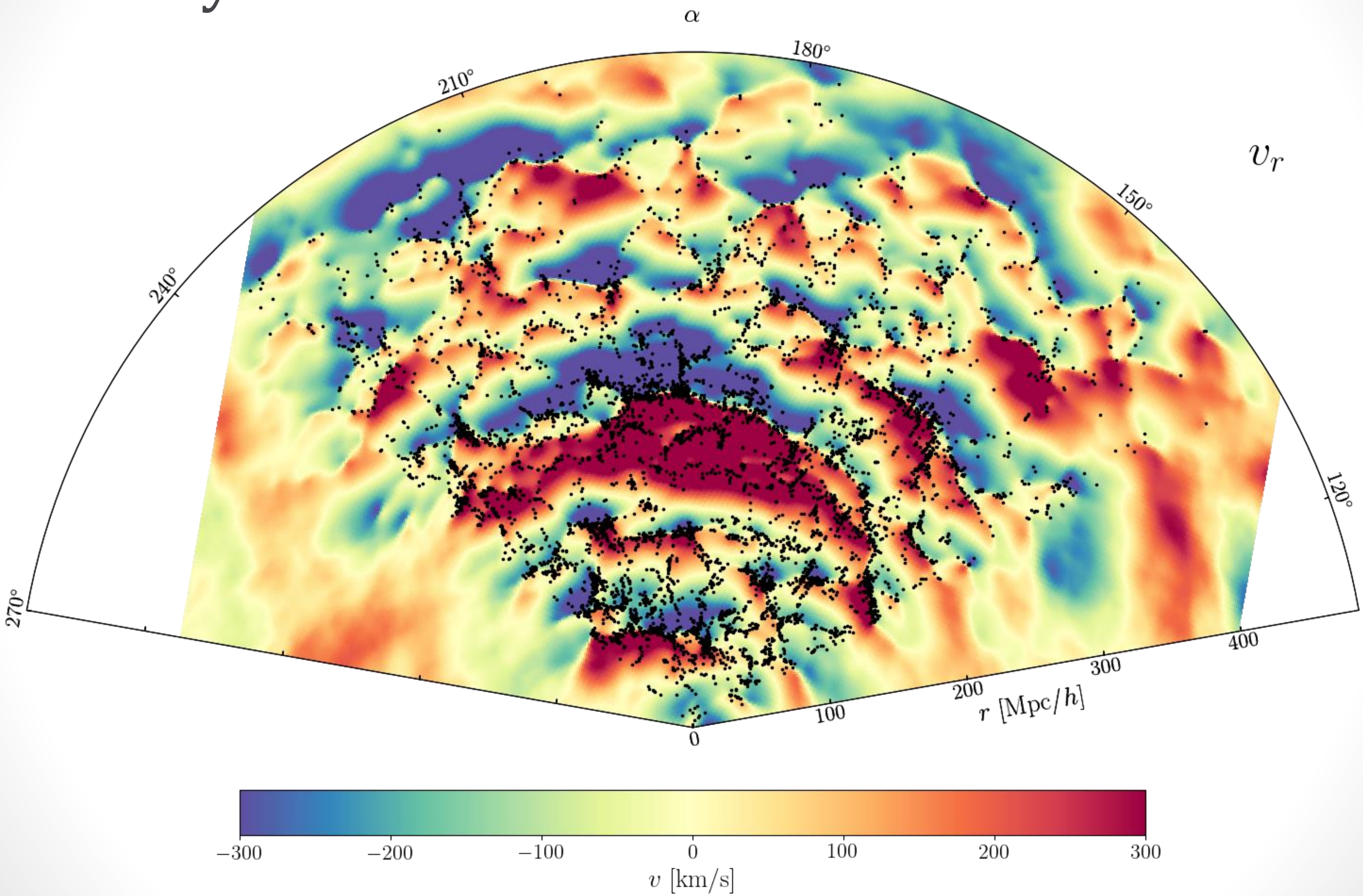


Initial conditions | Final conditions | Observations

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

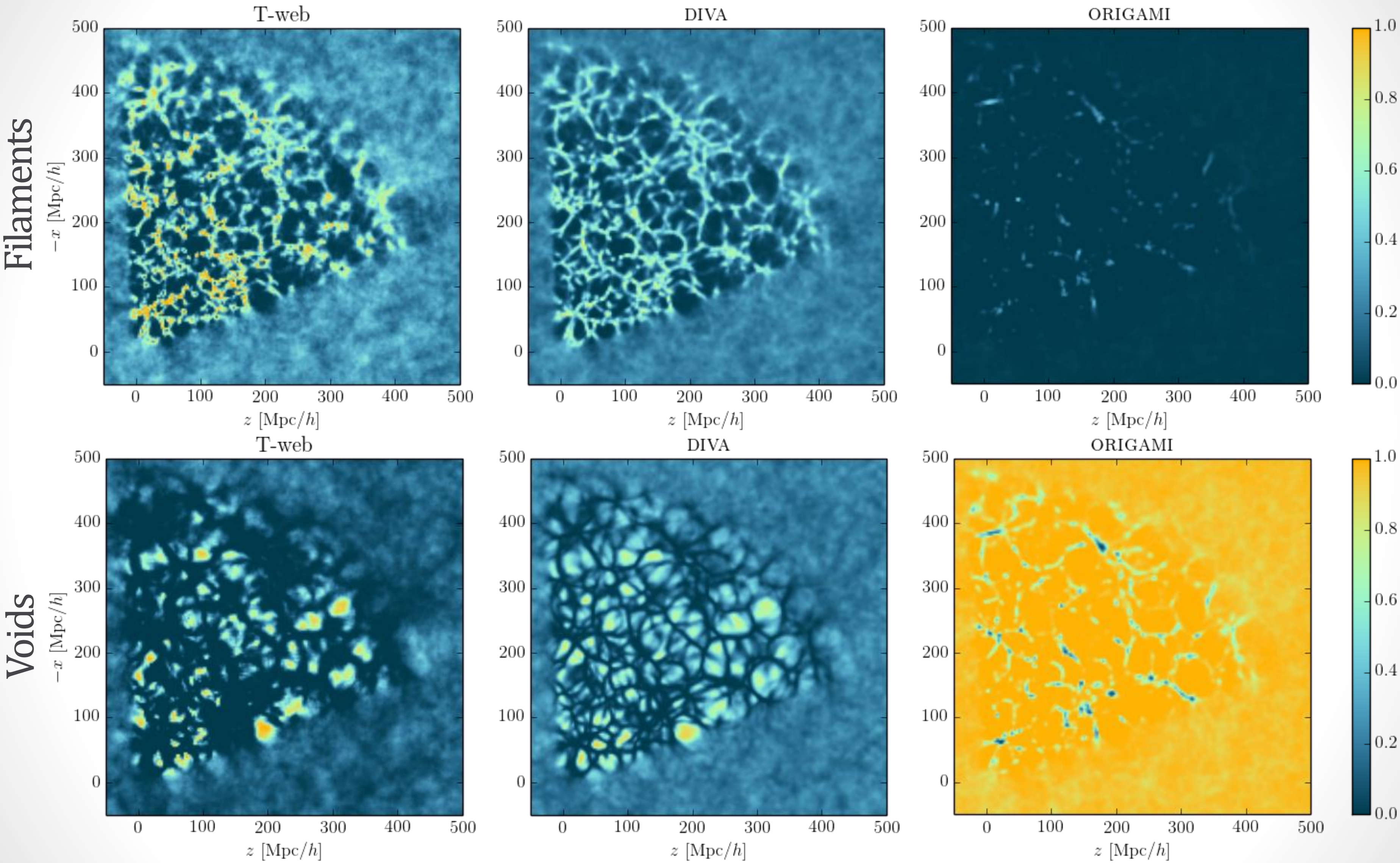
Jasche, FL & Wandelt 2015, arXiv:1409.6308

Velocity field



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

Cosmic web classifications



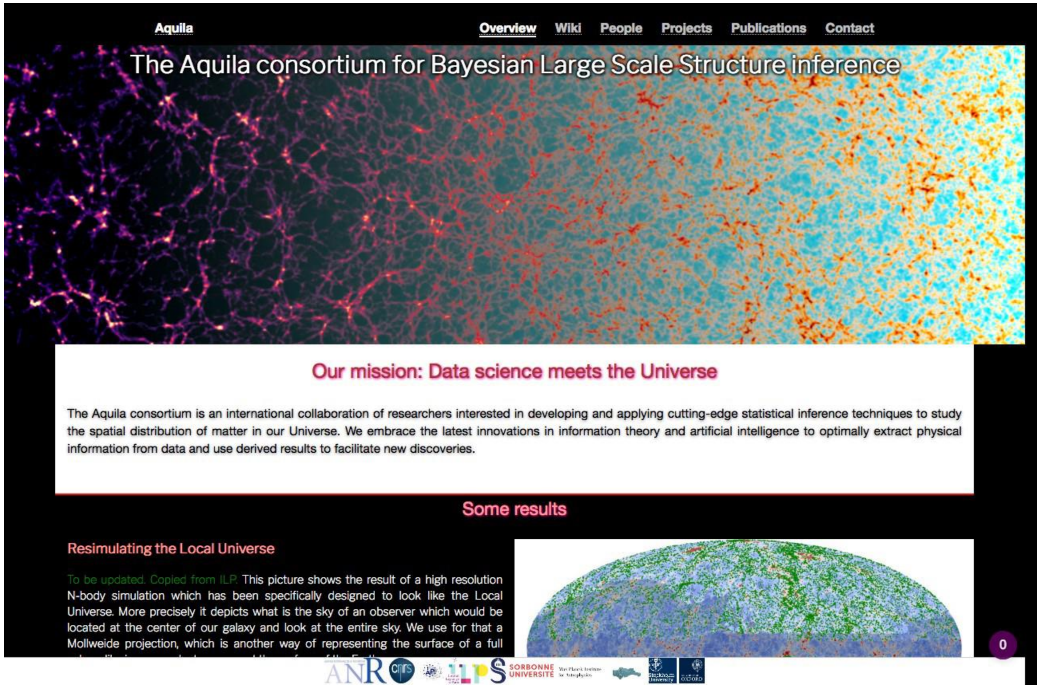
FL, Jasche & Wandelt 2015a, arXiv:1502.02690

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

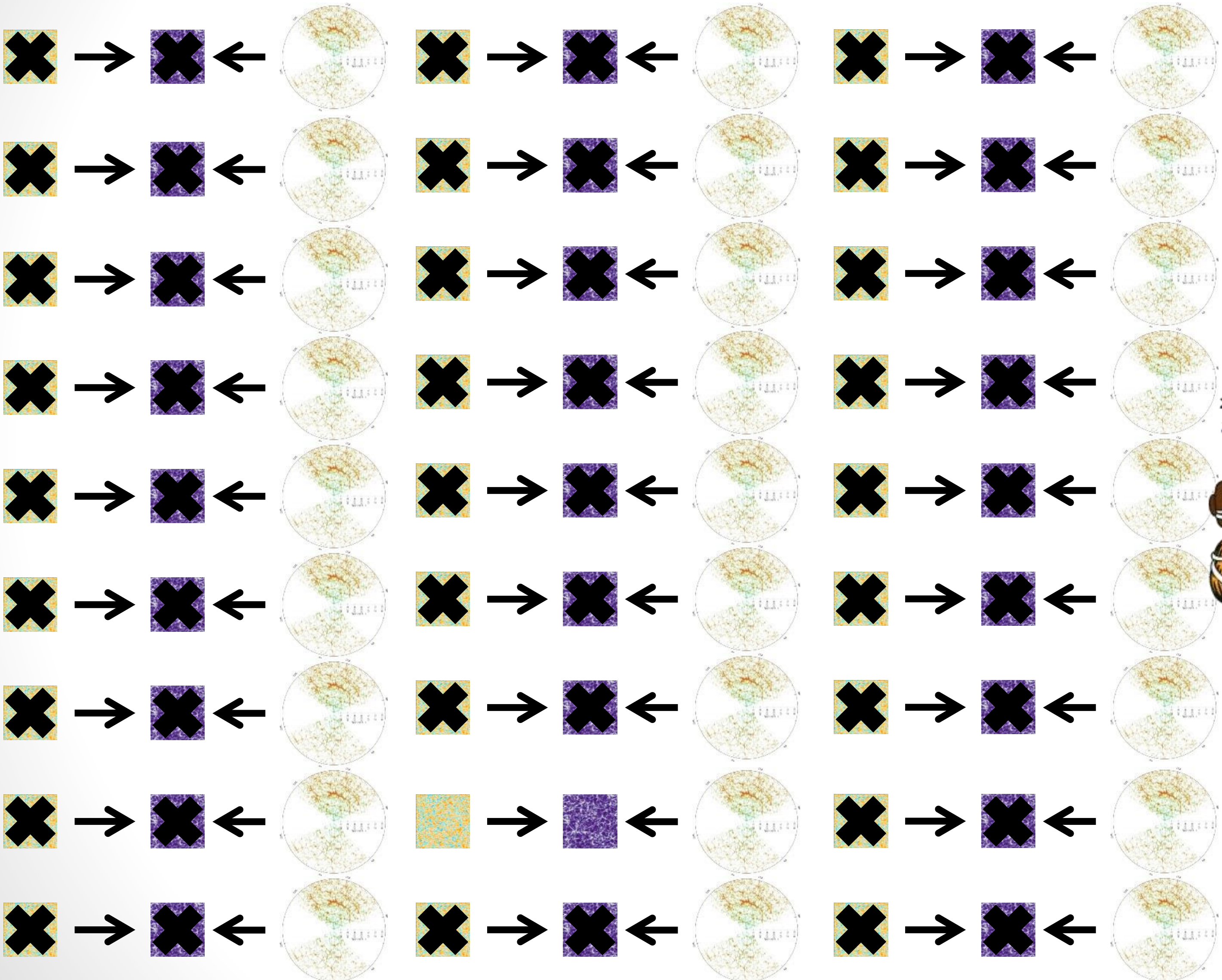
The Aquila Consortium
for Bayesian large-scale structure inference

- Created in 2016. Members from the UK, France, Germany & Sweden.
- Gathers people interested in developing the Bayesian pipelines and running analyses on cosmological data.

www.aquila-consortium.org



Let's go back to the challenge…



$d \approx 10^7$

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

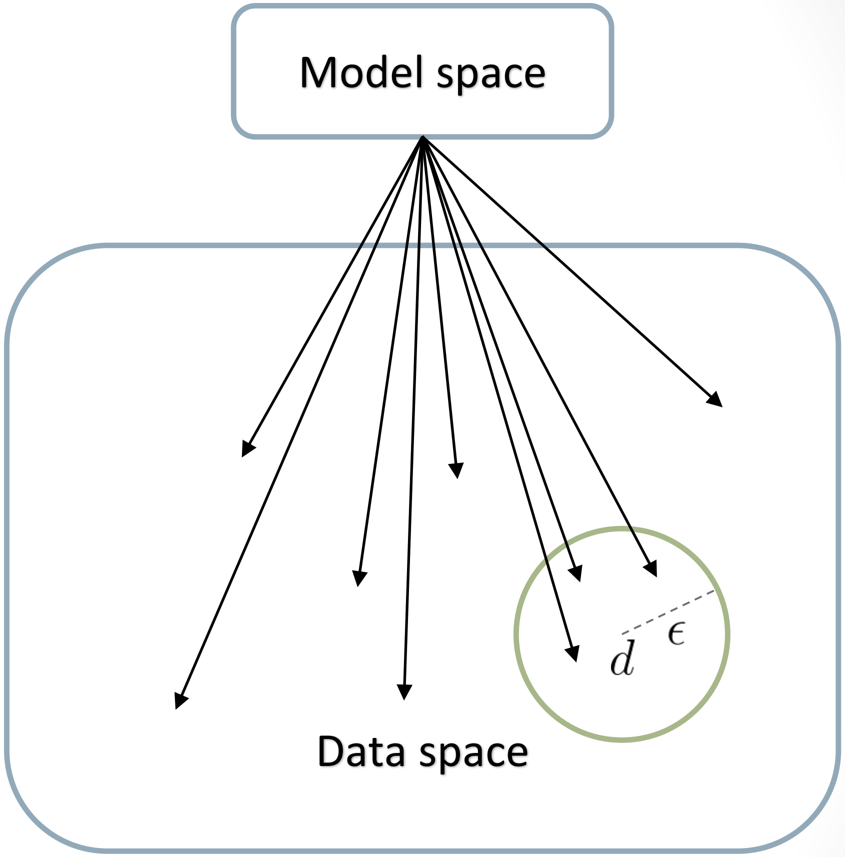
$$p(\theta|d) \Longrightarrow p(\theta|\tilde{d}) \qquad \text{where } \mathrm{d}(\tilde{d}(\theta), d) \text{ is small}$$

- **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)

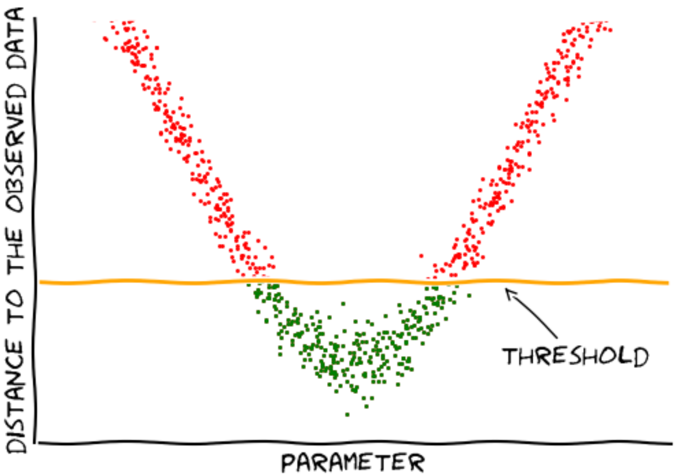
Why is likelihood-free rejection so expensive?

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 $L(\theta)$

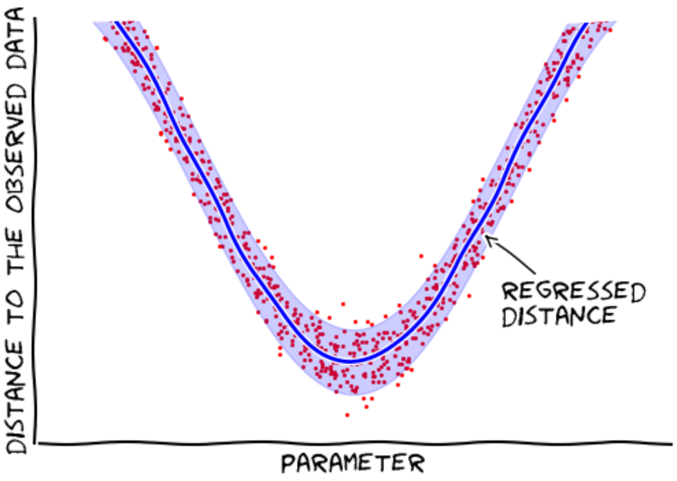
⇨ **Model the distances, assuming the average distance is smooth**

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

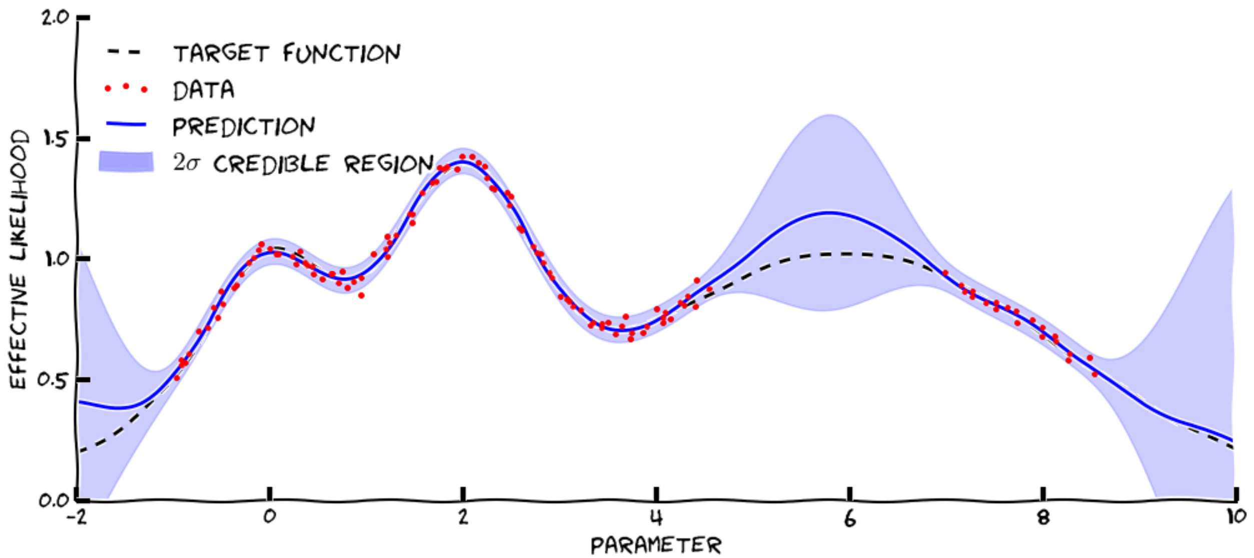
⇨ **Prioritize parameter regions with small distances to the observed data**



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)

Gutmann & Corander JMLR 2016, arXiv:1501.03291

Regressing the effective likelihood (points 1 & 2)



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

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

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

- Model the conditional distribution of distances given this training set

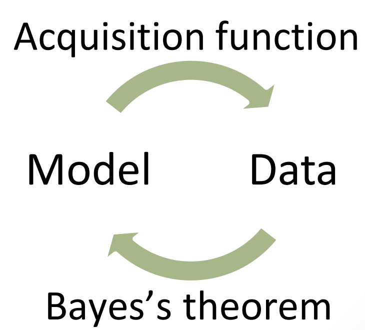
Data acquisition (points 3 & 4)

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

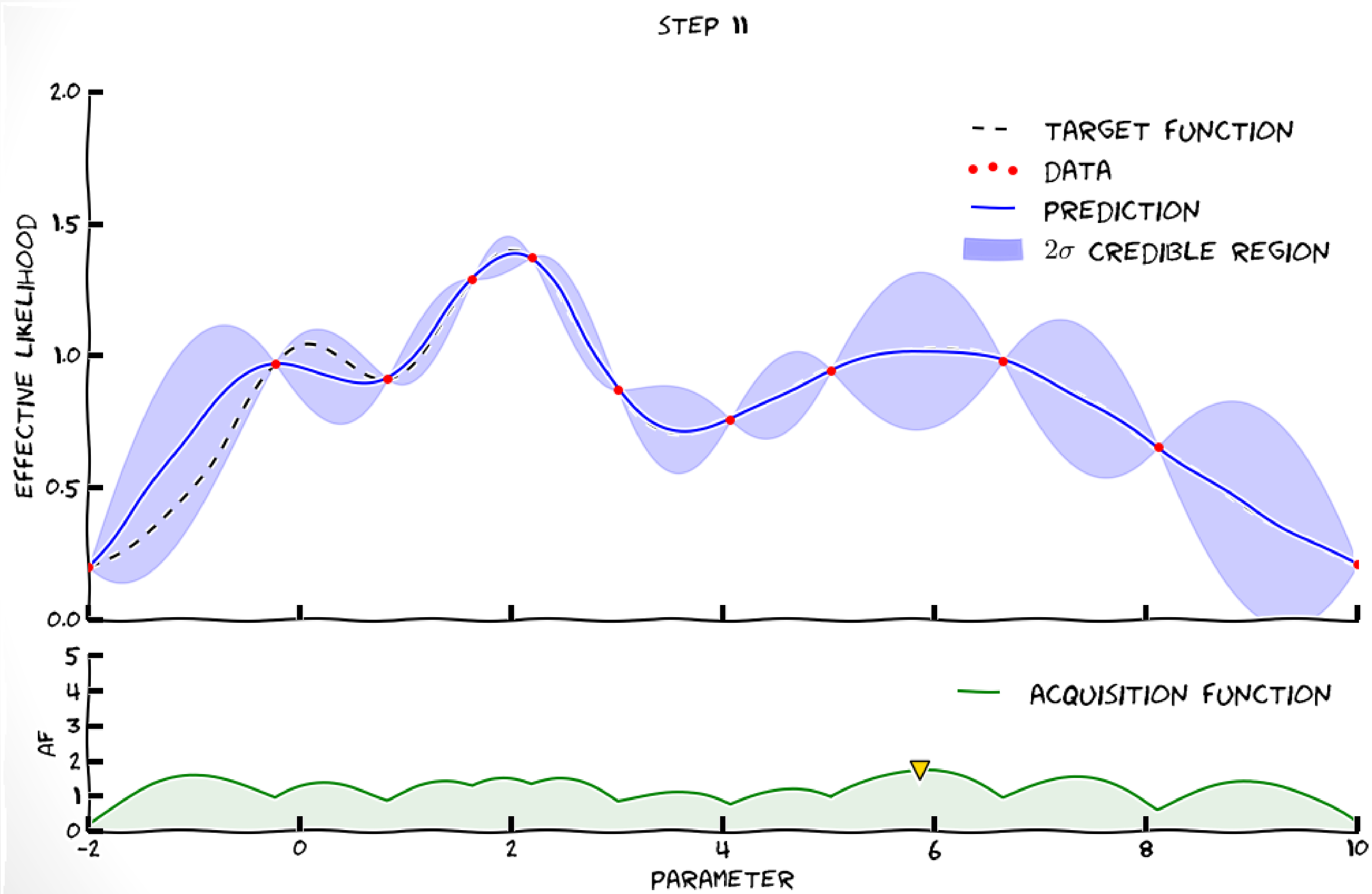
- Samples are obtained from sampling an **adaptively-constructed proposal distribution**, using the regressed effective likelihood

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

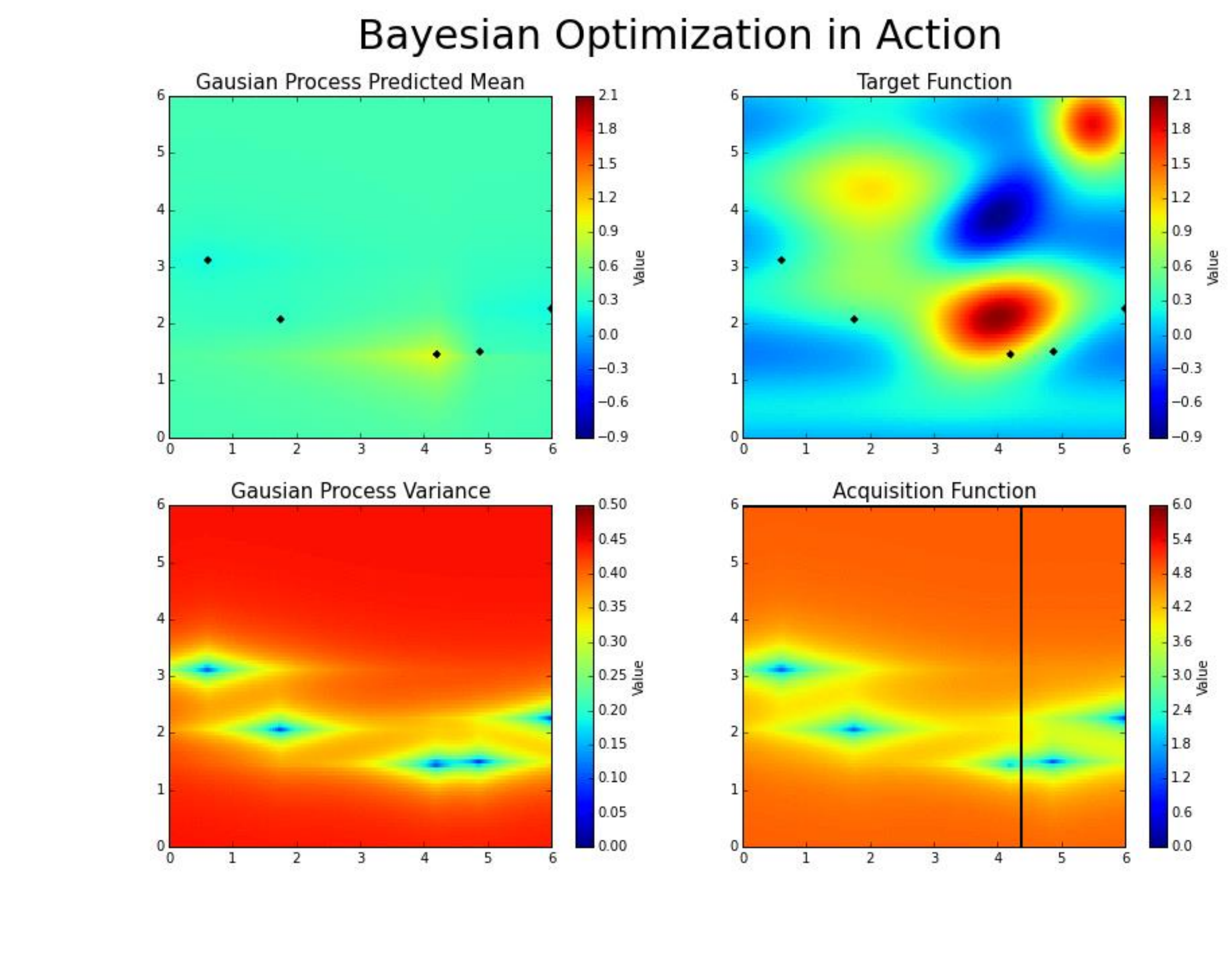
- The **acquisition function** finds a compromise between exploration (trying to find new high-likelihood regions) & exploitation (giving priority to regions where the distance to the observed data is already known to be small)
- **Bayesian optimisation** (decision making under uncertainty) can then be used



Data acquisition

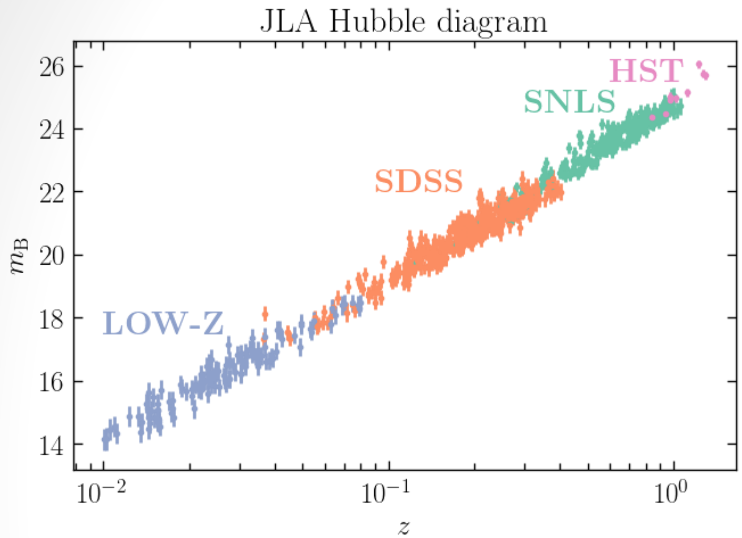


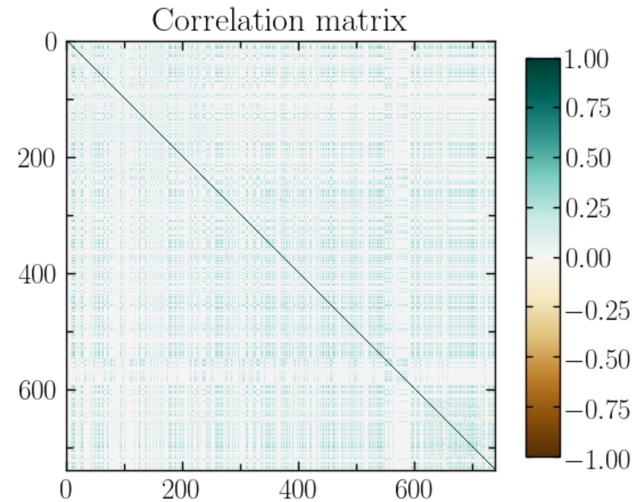
In higher dimension…



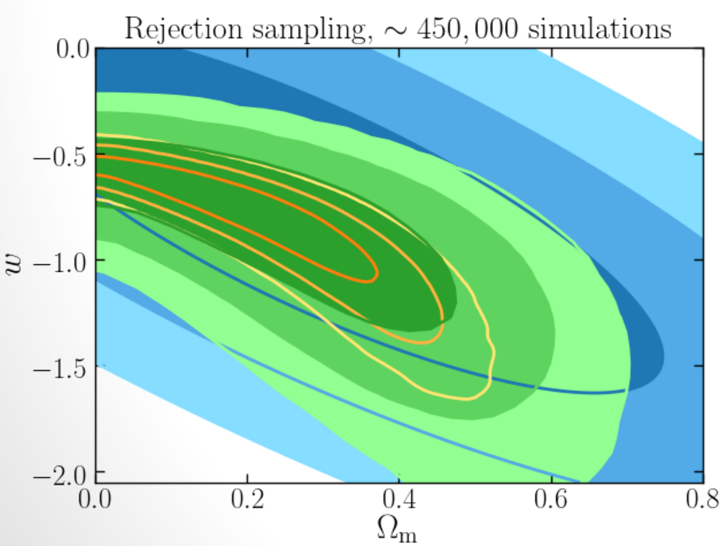
F. Nogueira, https://github.com/fmfn/BayesianOptimization

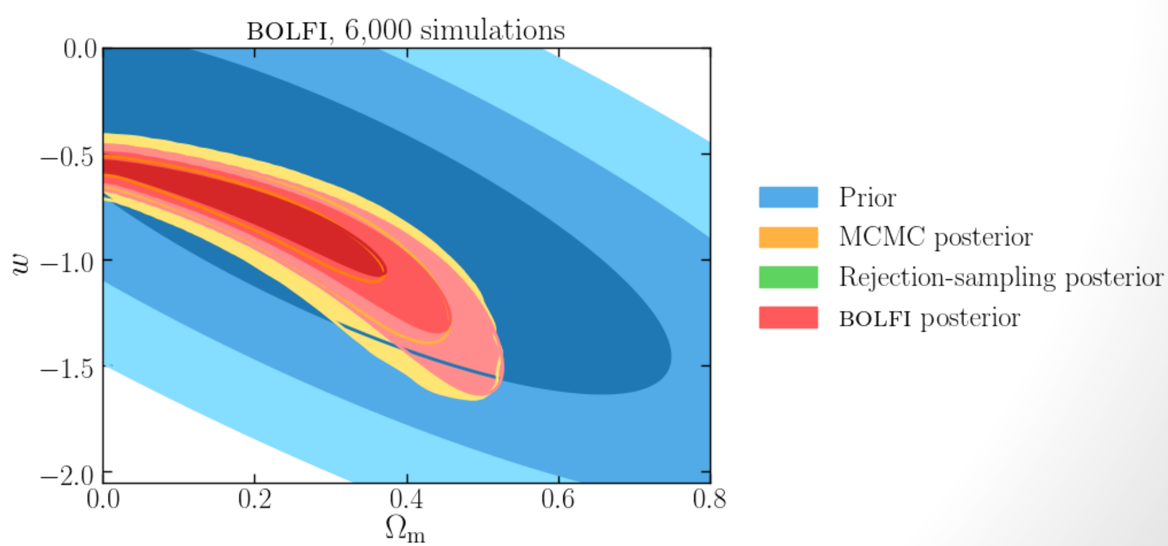
Application: Analysis of the JLA supernova sample





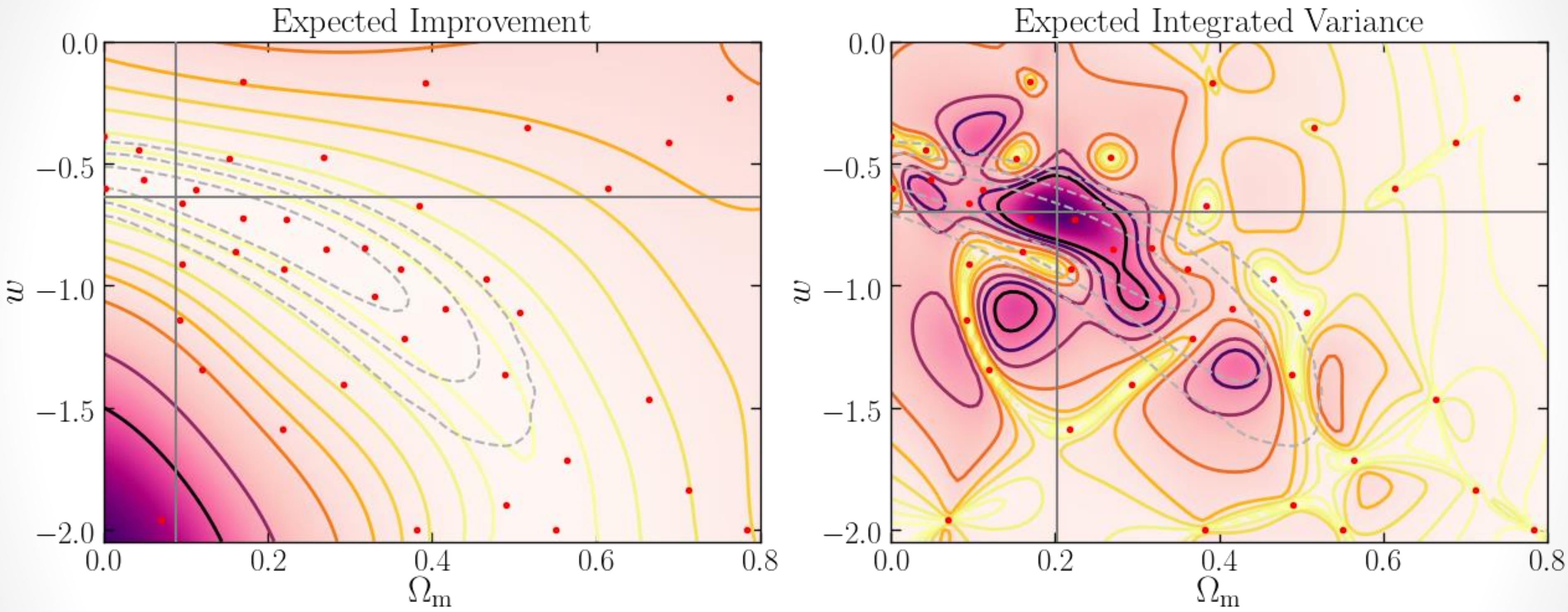
Betoule *et al.* 2014, arXiv:1401.4064





FL 2018, arXiv:1805.07152

An acquisition function designed for ABC



Järvenpää *et al.* 2017, arXiv:1704.00520

FL 2018, arXiv:1805.07152

Summary

Inference with generative cosmological models

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.