Bayesian optimisation for likelihood-free cosmological inference

(work in progress)

Florent Leclercq

Institute of Cosmology and Gravitation, University of Portsmouth

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

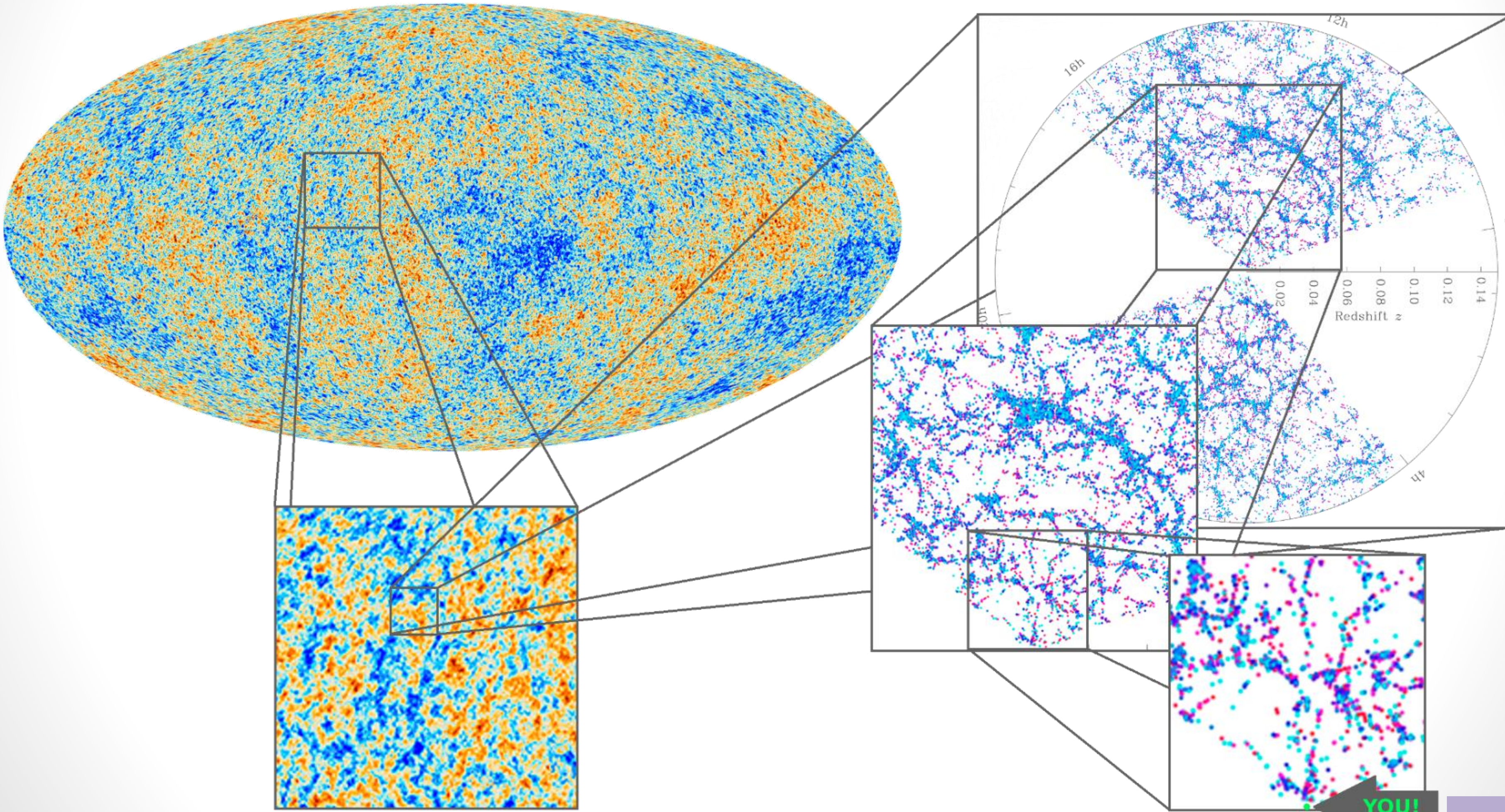
March 29th, 2017

In collaboration with

Wolfgang Enzi & Jens Jasche (ExC Universe, Garching)

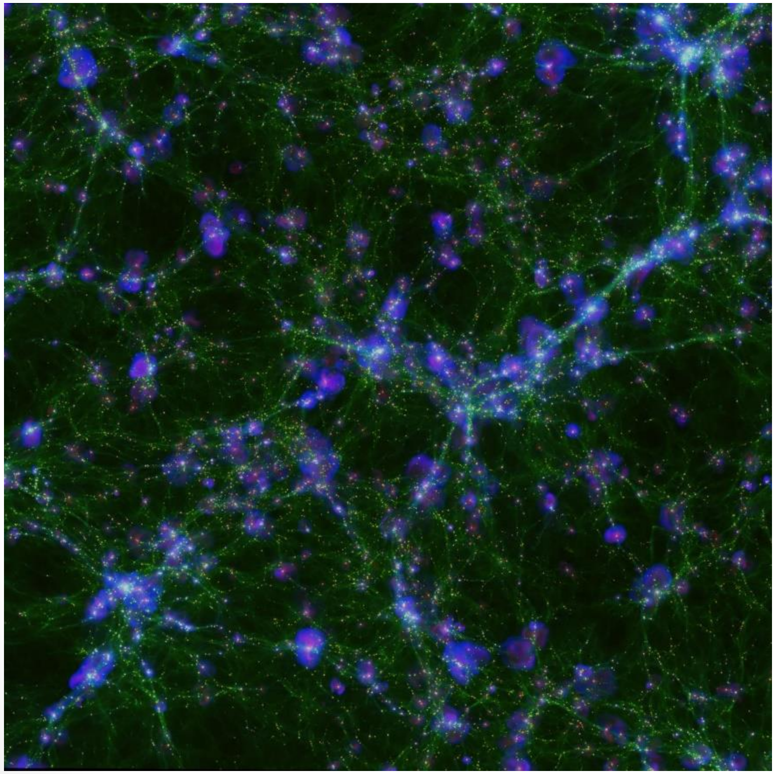
The big picture: the Universe is highly structured

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



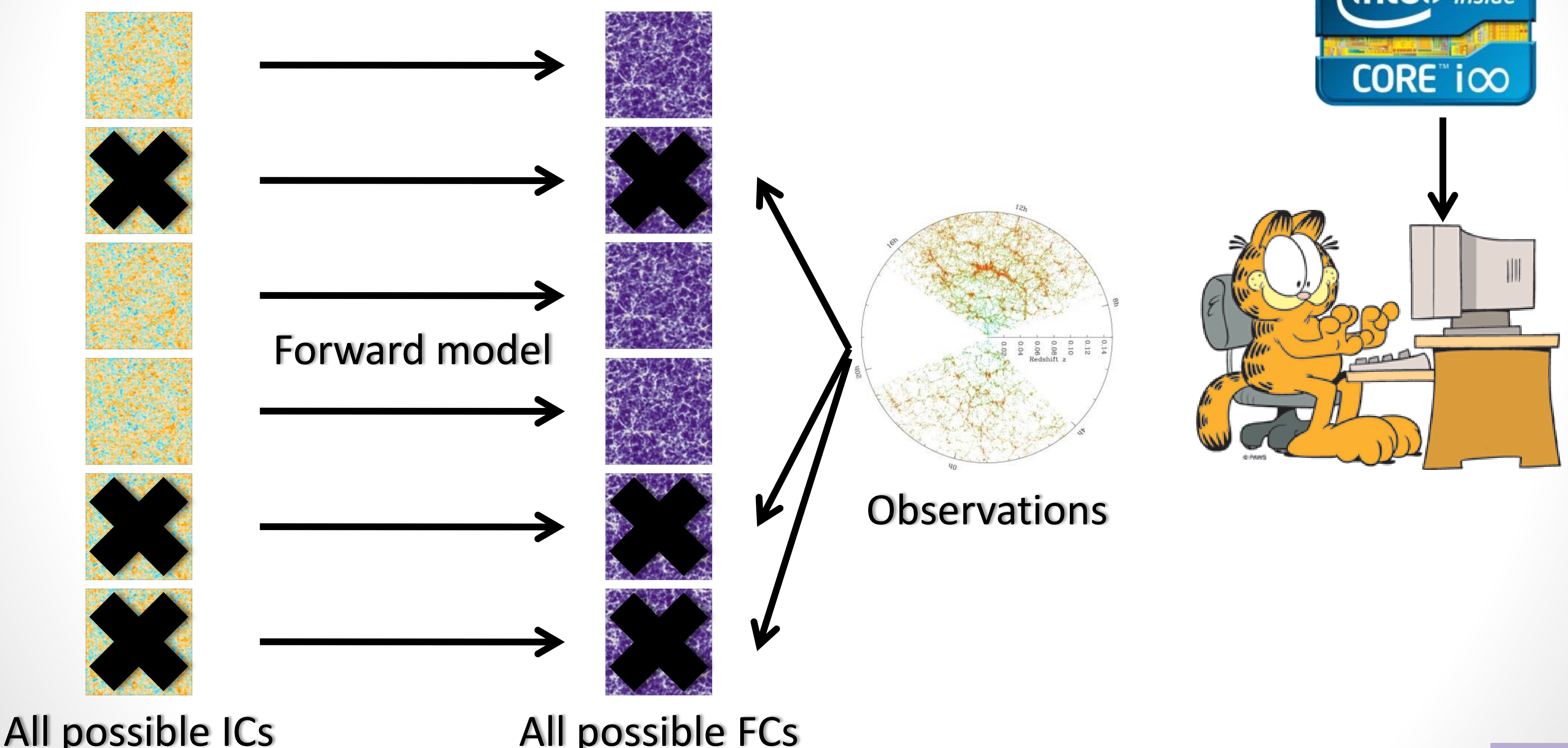
Y. Dubois (PI), Horizon AGN simulation (2014-2016)

The LSS is a vast source of knowledge:

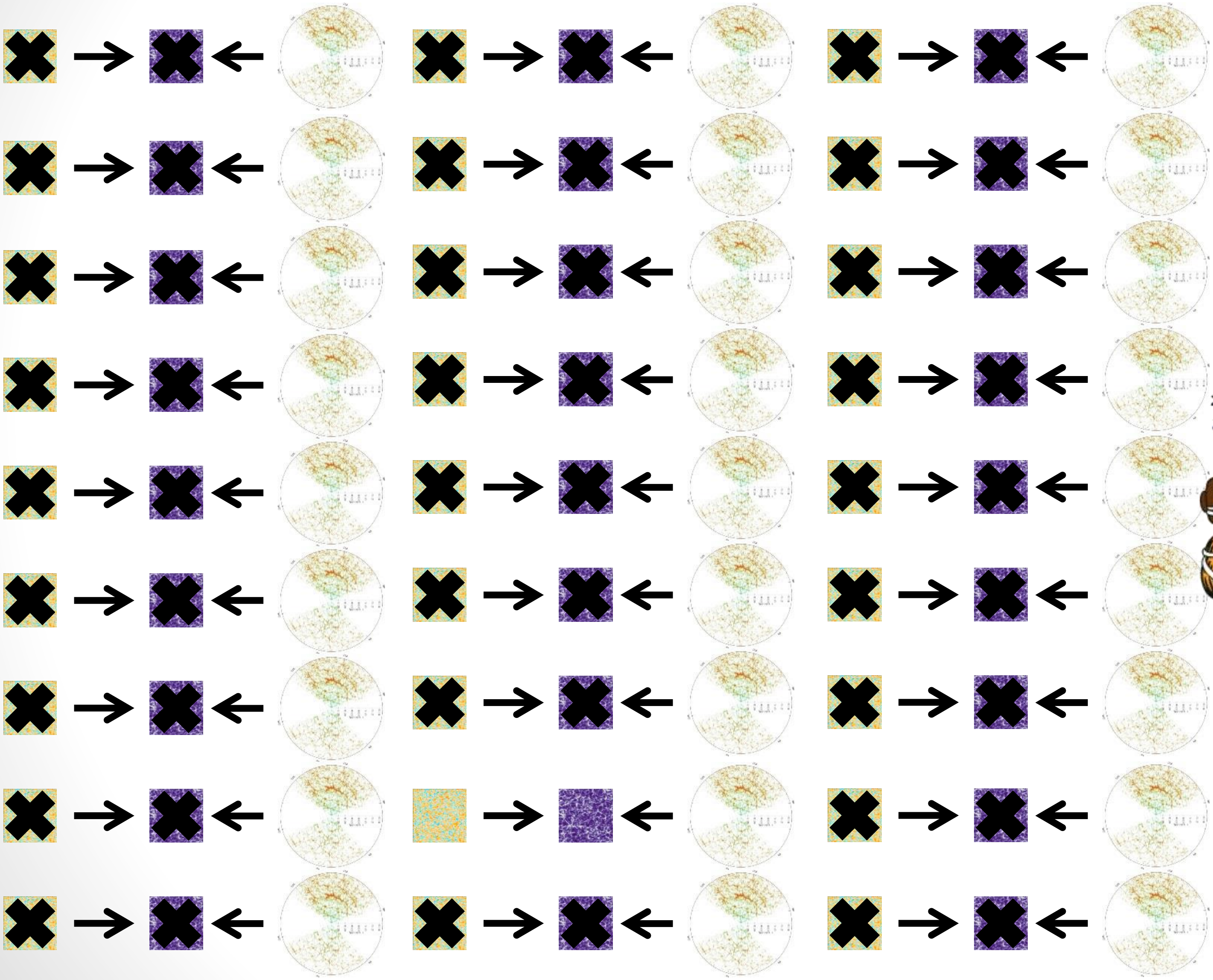
- **Cosmology**:
  - Cosmological parameters and tests of ΛCDM,
  - Physical nature of the dark components,
  - Geometry of the Universe,
  - Tests of General Relativity,
  - Initial conditions and link to high energy physics
- **Astrophysics**: galaxy formation and evolution as a function of their environment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

Bayesian forward modeling: the ideal scenario

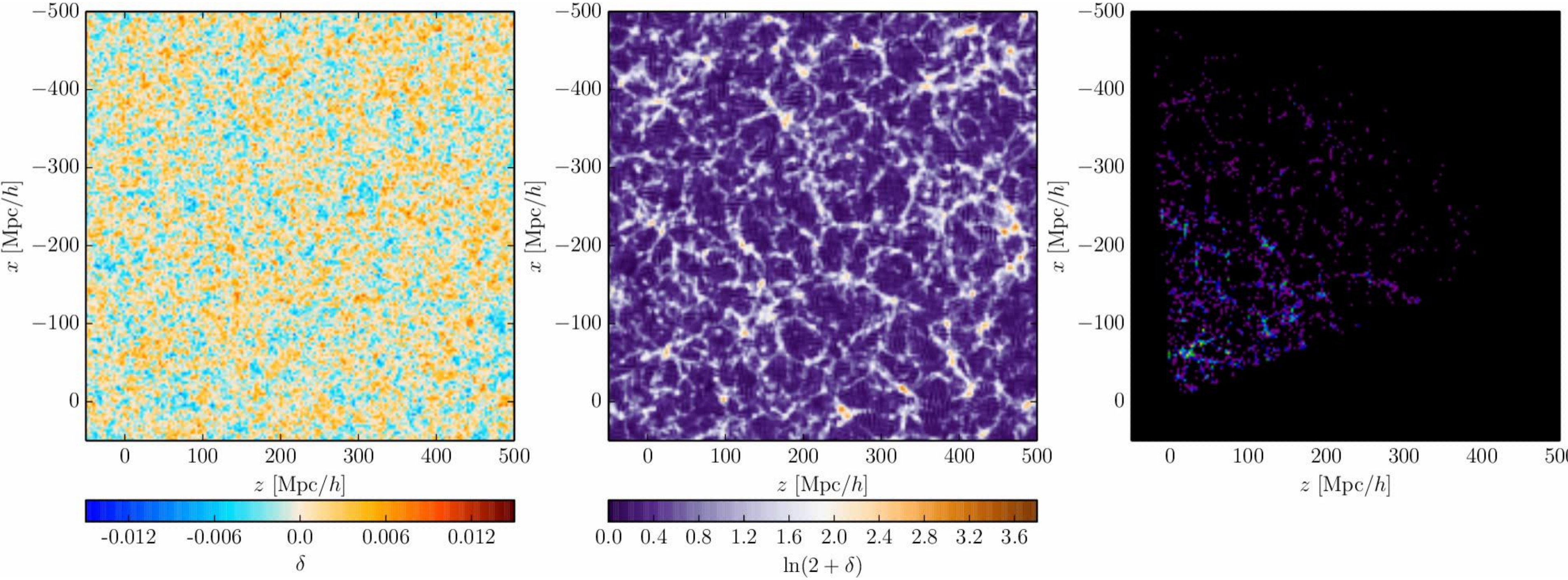
Forward model = N-body simulation + Halo occupation +
Galaxy formation + Feedback + …



Bayesian forward modeling: the ideal scenario



Likelihood-based solution: BORG



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

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

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

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

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

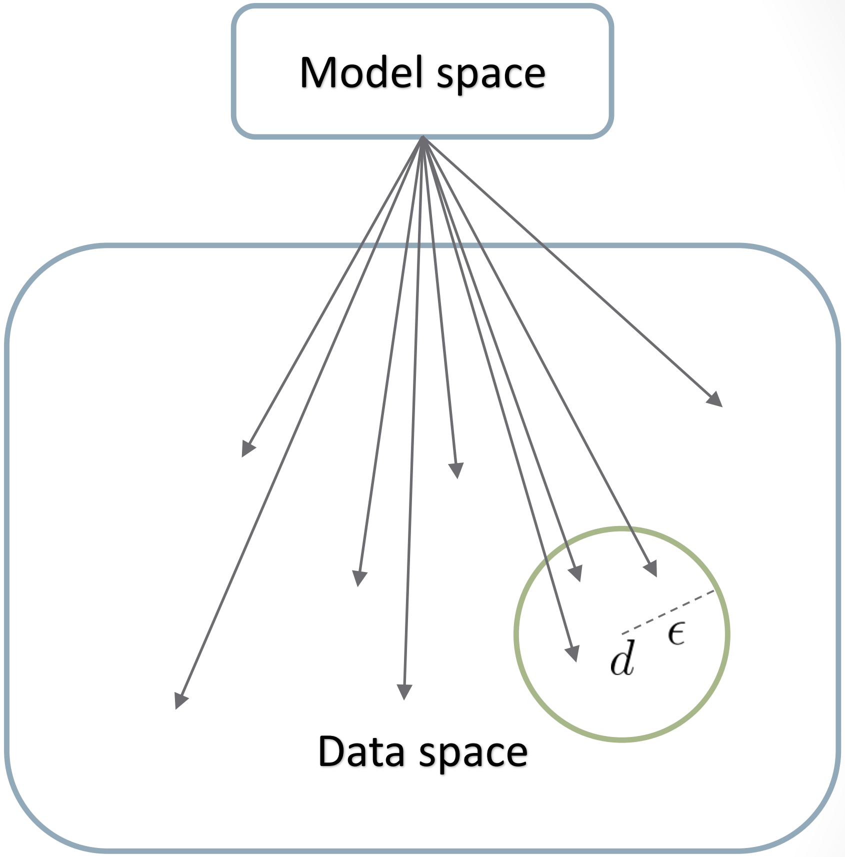
$$p(\theta|d) \Longrightarrow p(\theta|\tilde{d}) \quad \text{where } \mathrm{d}(\tilde{d}(\theta), d) \text{ is small}$$

- **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

- ϵ can be adaptively reduced (Population Monte Carlo)



Effective likelihood approximation:

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

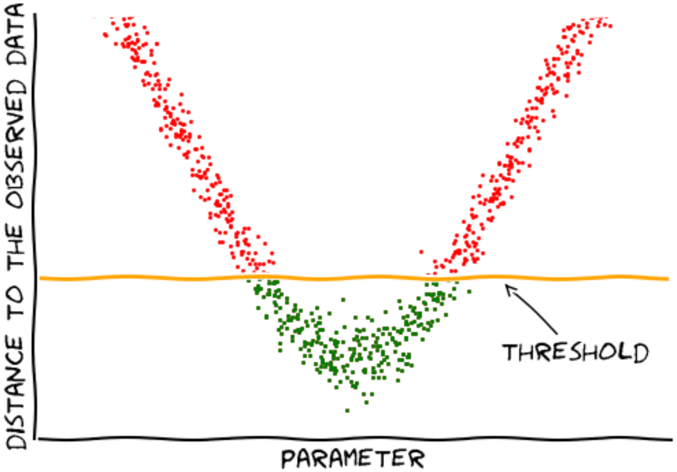
Why is likelihood-free rejection so expensive?

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

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 $L(\theta)$

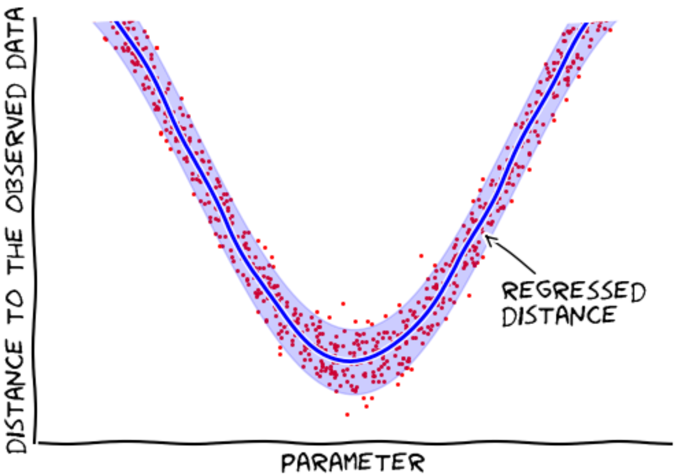
⇨ **Model the distances, assuming the average distance is smooth**

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

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

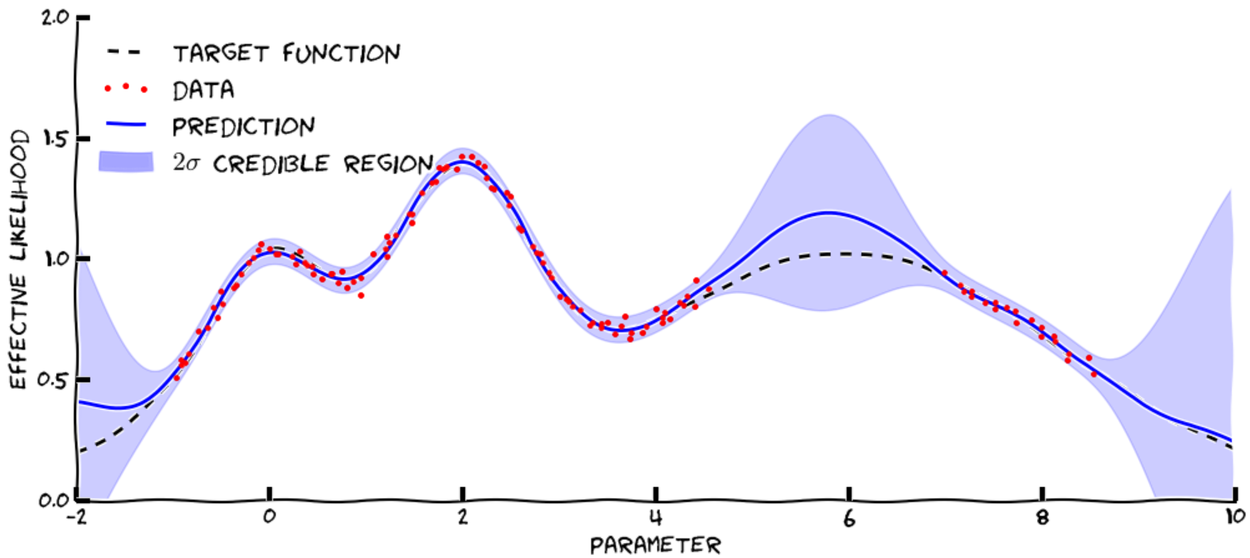
4. It aims at equal accuracy for all regions in parameter space

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



Gutmann & Corander JMLR 2016, arXiv:1501.03291

Regressing the effective likelihood (points 1 & 2)



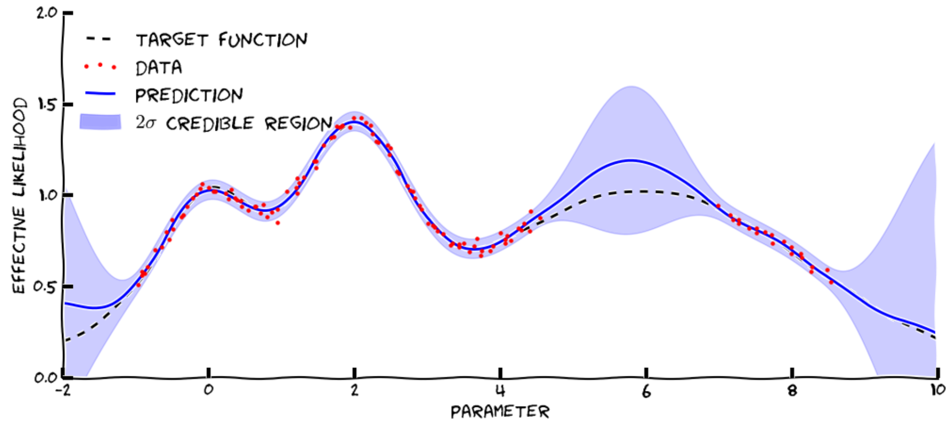
1. “It rejects most samples when ϵ is small”

- 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

Gaussian process regression (a.k.a. kriging)



- Why?
  - It is a **general purpose regressor**: it will be able to deal with a large variety of complex/non-linear features of likelihood functions.
  - It provides not only a prediction, but also the **uncertainty of the regression**.
  - It allows to **extrapolate** in regions where we have no data points.

$$p(\mathbf{f}|\mathbf{X}) \propto \exp\left[-\frac{1}{2}\sum_{mn}(f(\mathbf{x}_m)-\mu(\mathbf{x}_m))^\intercal K(\mathbf{x}_m,\mathbf{x}_n)(f(\mathbf{x}_n)-\mu(\mathbf{x}_n))\right]$$

$$K(\mathbf{x}_m,\mathbf{x}_n) = \underbrace{C_1}_{K_{\mathrm{C}}(C_1)} \times \underbrace{\exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}_m-\mathbf{x}_n}{C_2}\right)^2\right]}_{K_{\mathrm{RBF}}(C_2)} + \underbrace{C_3\delta_{\mathrm{K}}^{mn}}_{K_{\mathrm{GN}}(C_3)}$$

The prediction and uncertainty for a new point is:

$$p(f_\star|\mathbf{x}_\star,\mathbf{X},\mathbf{f}) \propto \exp\left[-\frac{1}{2}\left(\frac{f_\star-\alpha(\mathbf{x}_\star)}{\sigma(\mathbf{x}_\star)}\right)^2\right]$$

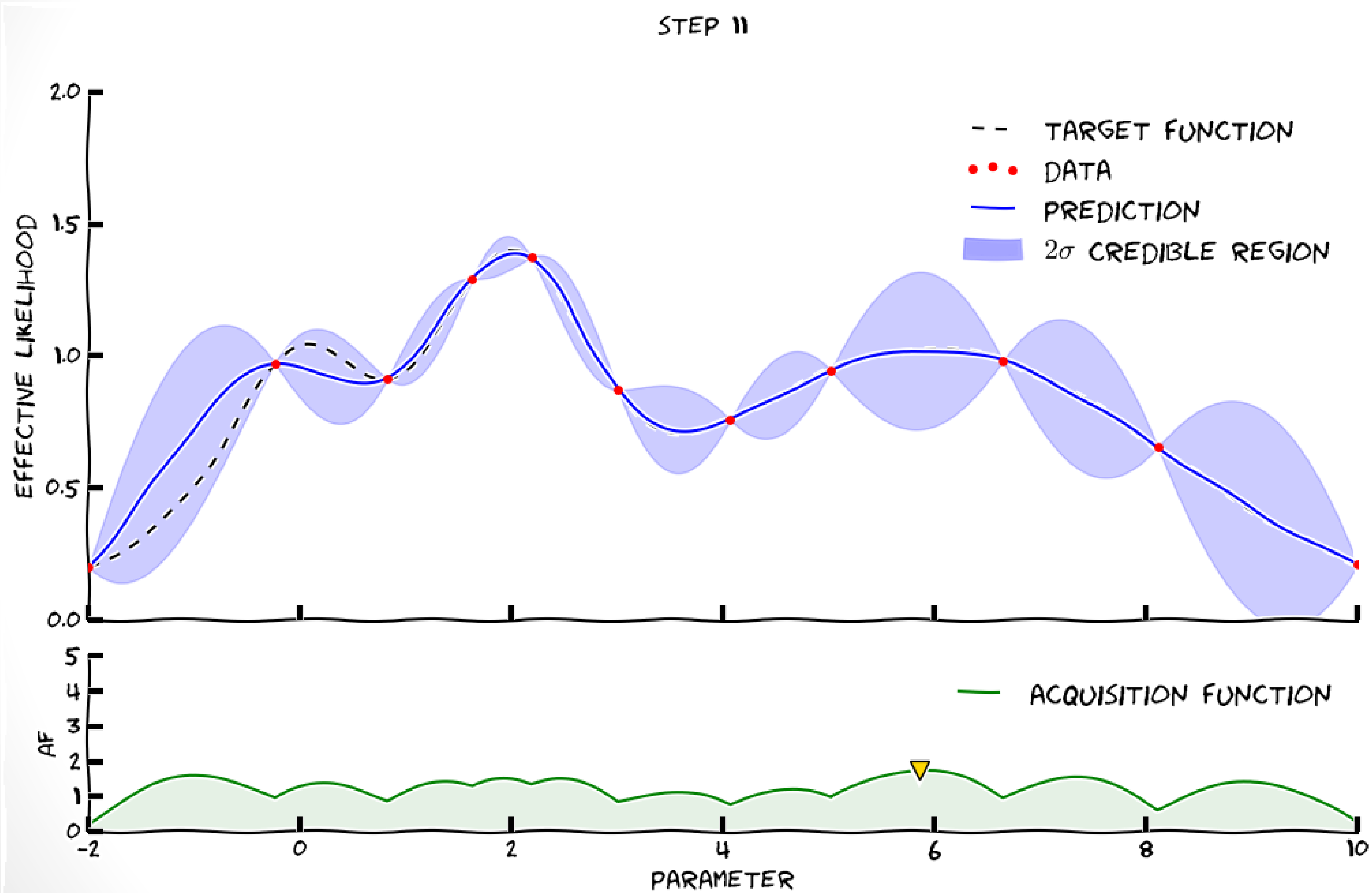
$$\alpha(\mathbf{x}_\star) = \mu(\mathbf{x}_\star) + K(\mathbf{x}_\star,\mathbf{x}_m)^\intercal K^{-1}(\mathbf{x}_m,\mathbf{x}_n)(\mathbf{f}-\mu(\mathbf{X}))_n$$

$$\sigma(\mathbf{x}_\star)^2 = K(\mathbf{x}_\star,\mathbf{x}_\star) - K(\mathbf{x}_\star,\mathbf{x}_m)^\intercal K^{-1}(\mathbf{x}_m,\mathbf{x}_n)K(\mathbf{x}_\star,\mathbf{x}_n)$$

Hyperparameters C_1, C_2, C_3 are automatically adjusted during the regression.

Rasmussen & Williams 2006

Data acquisition



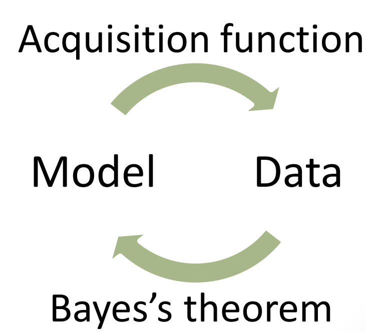
Data acquisition (points 3 & 4)

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

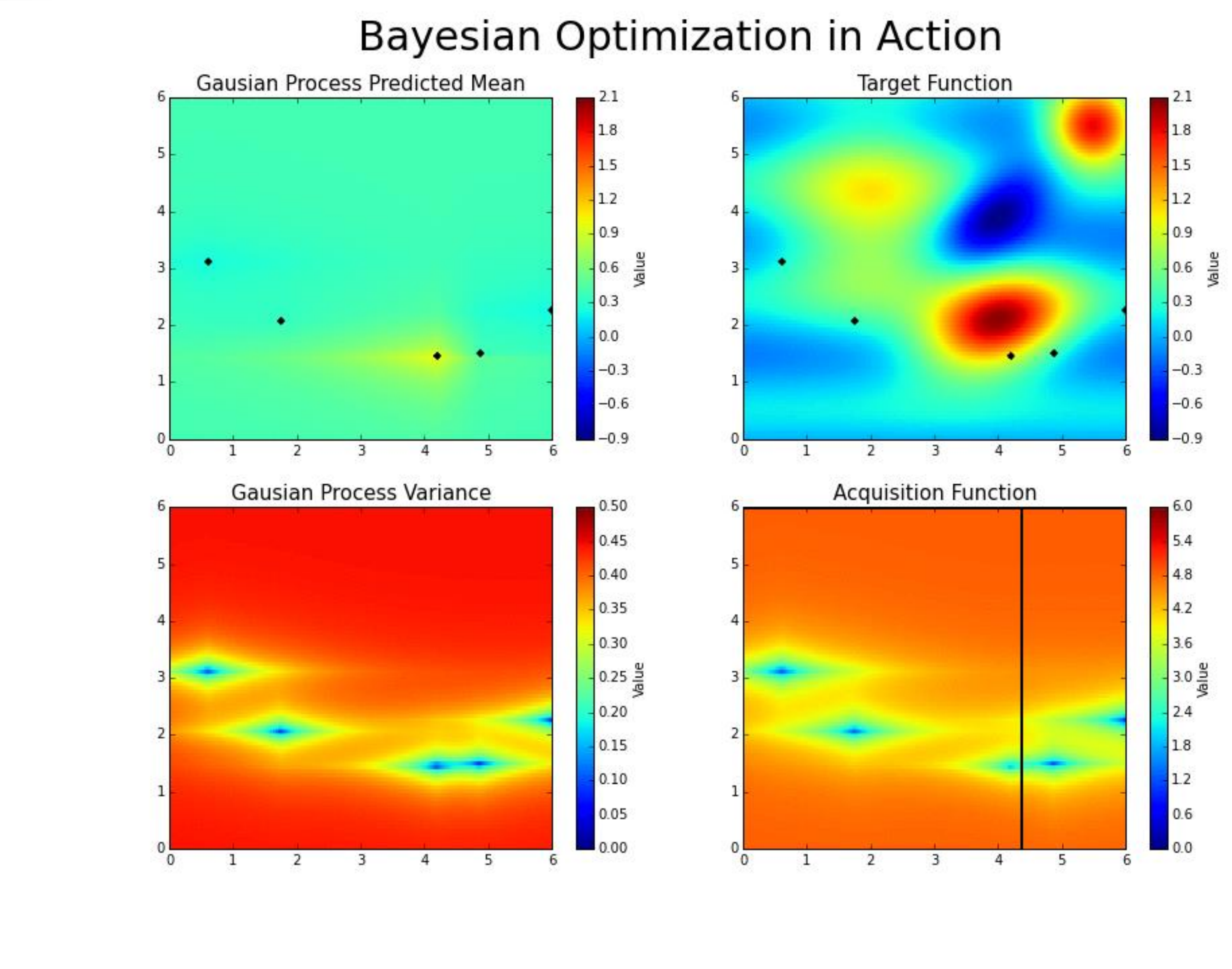
- Samples are obtained from sampling an **adaptively-constructed proposal distribution**, using the regressed effective likelihood

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



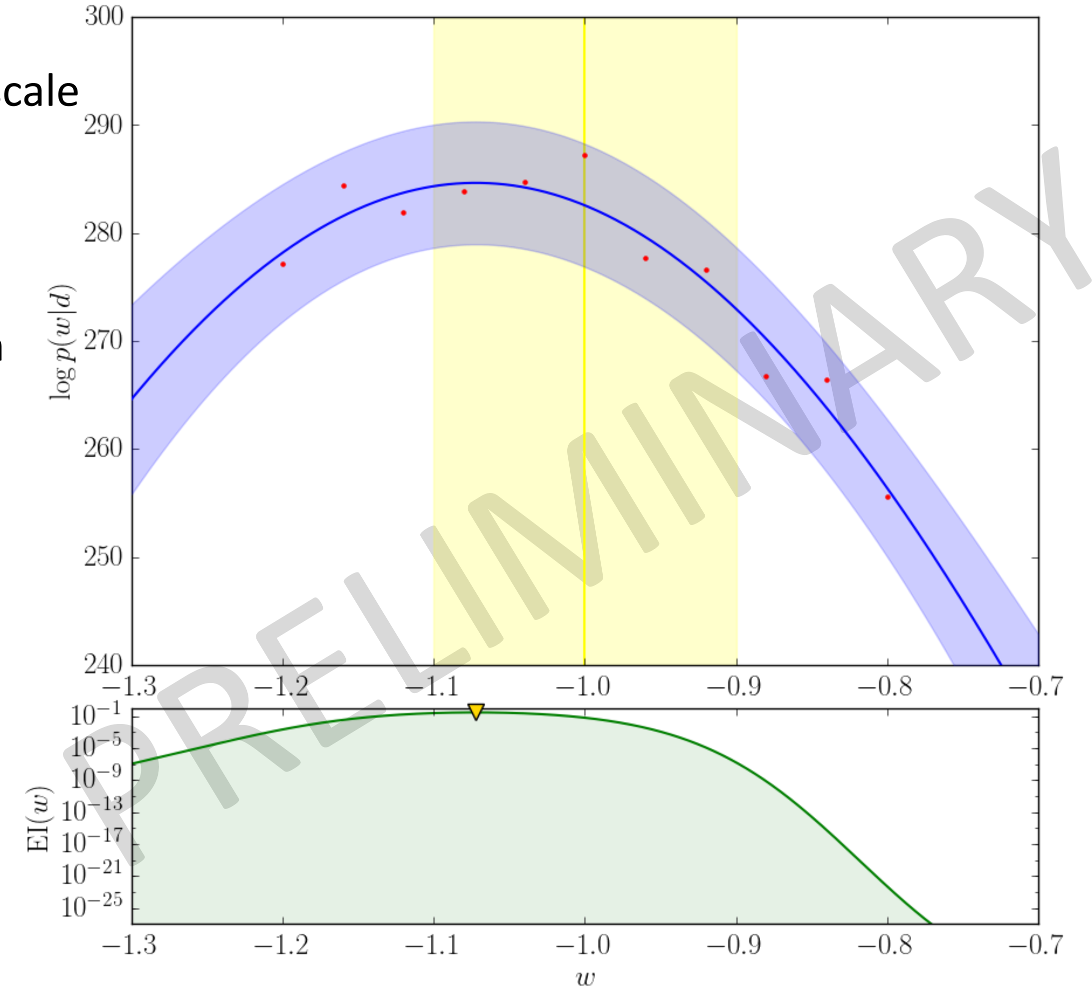
In higher dimension…



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

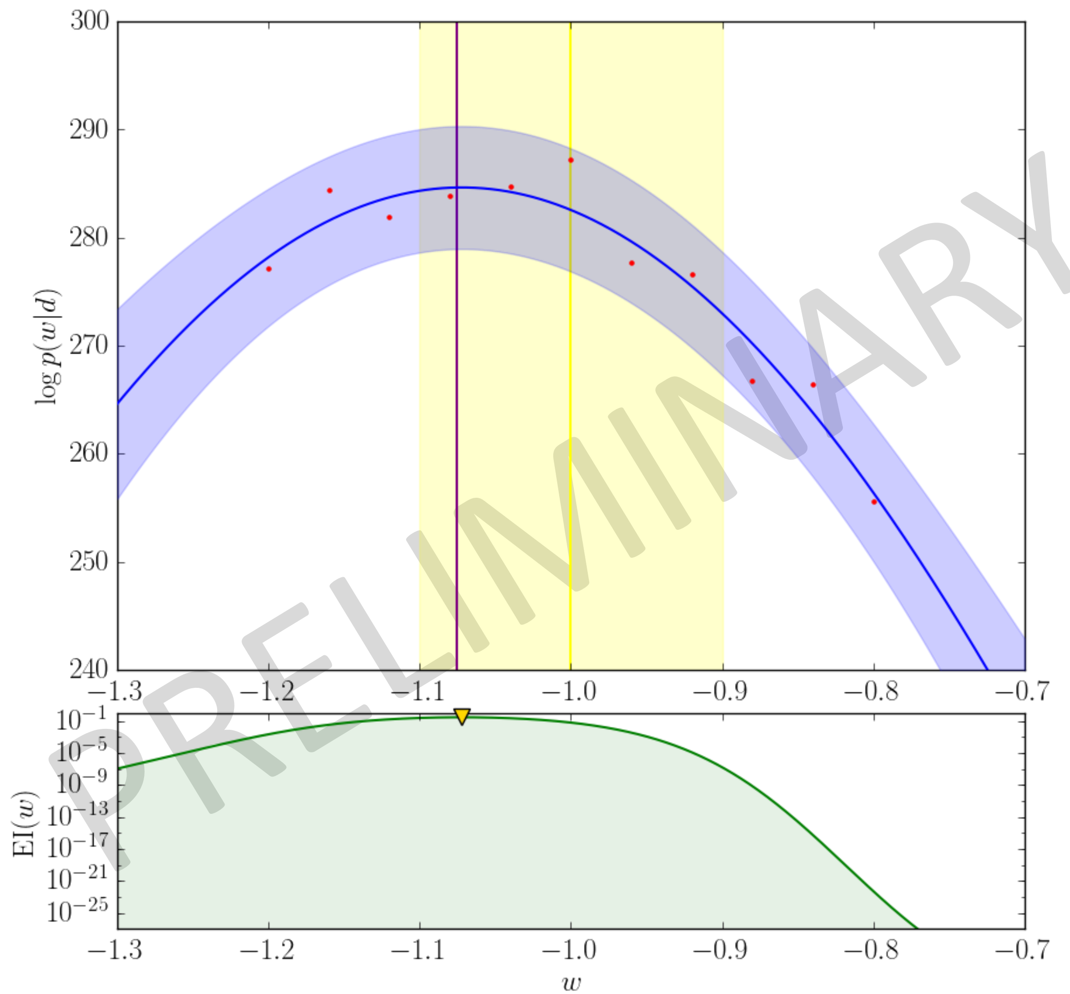
Likelihood-free large-scale structure inference

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



FL, Enzi & Jasche (in prep.)

Likelihood-free large-scale structure inference



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

FL, Enzi & Jasche (in prep.)

Summary

- **Problem considered**: inference for models where the likelihood is intractable but simulating is possible.
- **Approach**: combination of statistical modelling of the distance with Bayesian optimisation.
- **Outcome**: efficiency of the inference is increased by several orders of magnitude.

- The approach will allow to **ask targeted question to cosmological data**, including all relevant physical and observational effects.

- Open questions:
  - Summary statistics: how to "automatically" model the distance between simulated and observed data?

    Fearnhead & Prangle 2011, arXiv:1004.1112, Prangle et al. 2013, arXiv:1302.5624
  - Acquisition function: Can we find strategies that are optimal for cosmological problems?

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