Forward modelling the large-scale structure: field-level and implicit likelihood inference

Rubin LSST-France meeting

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

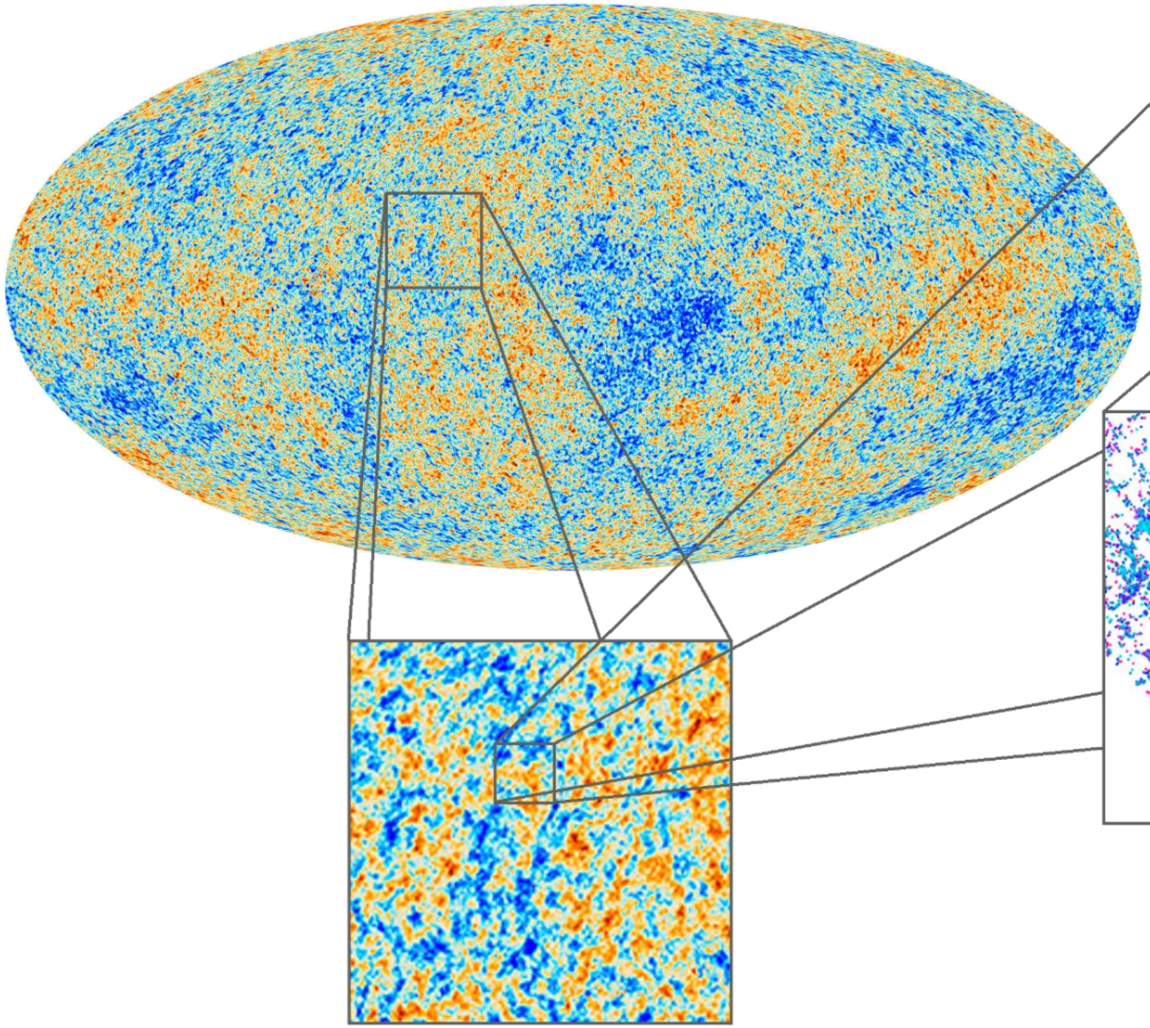
In collaboration with the Aquila Consortium

www.aquila-consortium.org

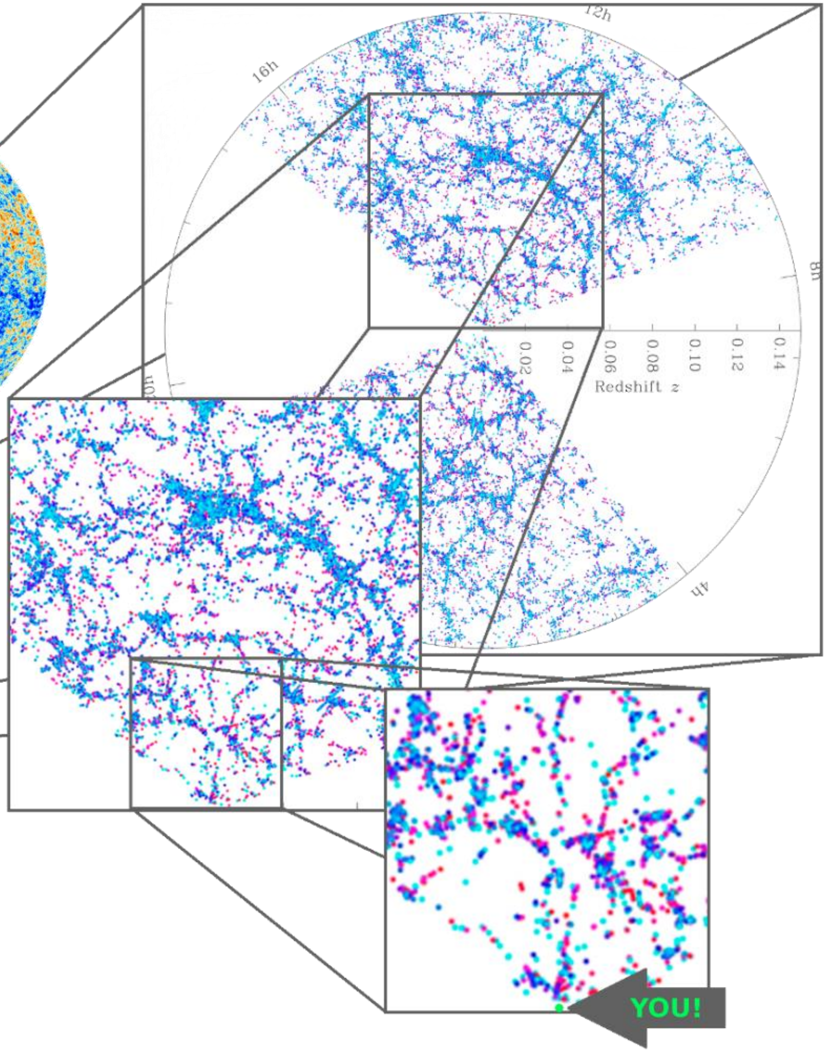
29 November 2022

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 large-scale structure

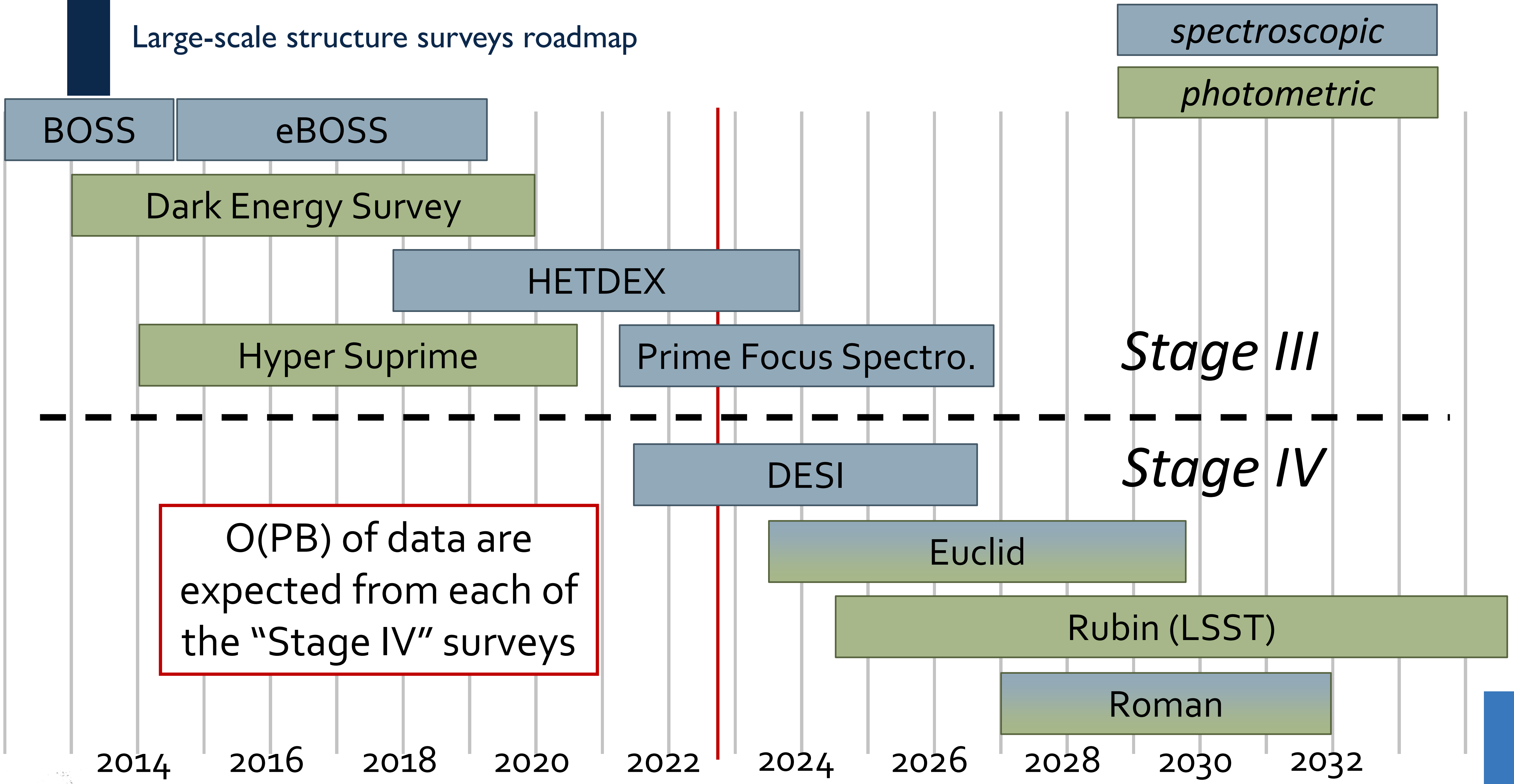
The LSS is a vast source of knowledge:

- Cosmology:
  - ΛCDM: cosmological parameters and tests against alternatives,
  - Physical nature of the dark components,
  - Neutrinos: number and masses,
  - 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, intrinsic size-magnitude correlations

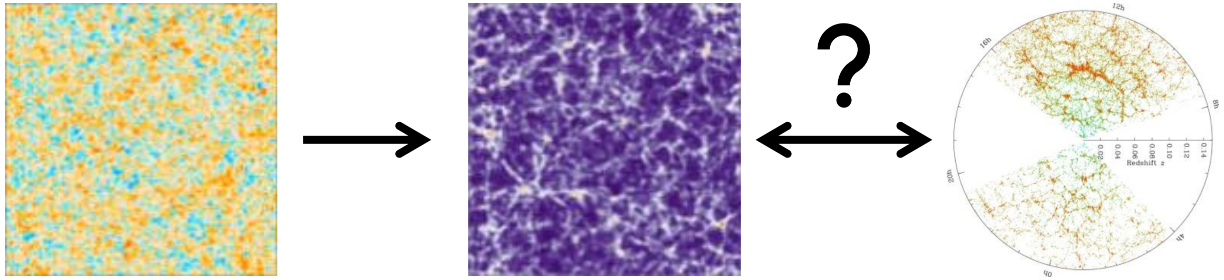
e.g. FL, Pisani & Wandelt 2014, 1403.1260

Large-scale structure surveys roadmap



Why Bayesian inference?

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

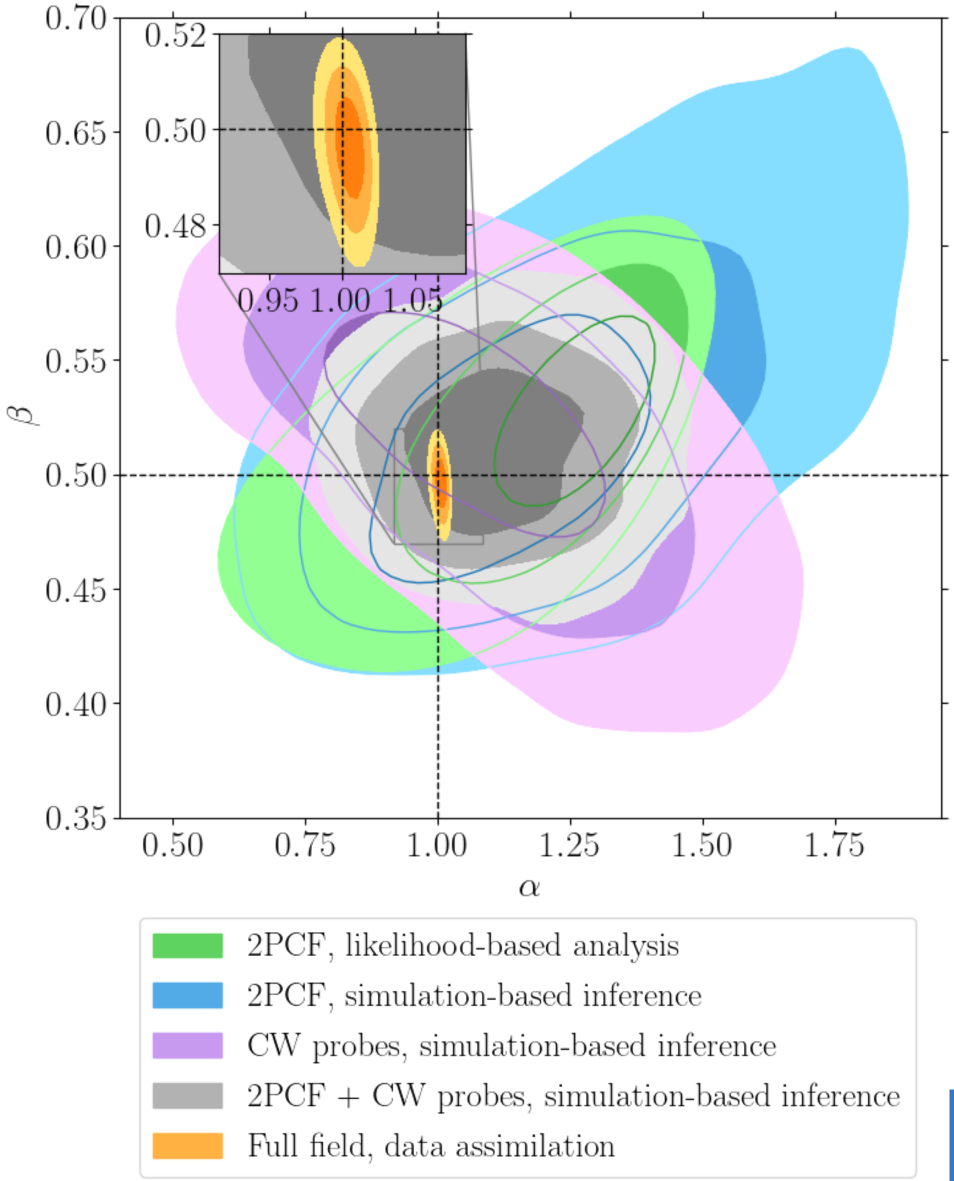


➡ No unique recovery is possible!

- A natural progression in cosmology:
  - Observations of the homogeneous and isotropic expansion (supernovæ)
  - Anisotropies of linear perturbations (CMB)
  - Non-linear cosmic structure at small scales and late times (galaxy surveys)
- Additional challenges for next-generation data:
  - Difficult data analysis questions and/or hints for new physics will first show up as tensions between measurements
  - Non-linearity: 80% of the total signal will come from non-linear structures
    e.g. LSST Science Book, 0912.0201
  - Model misspecification: Next-generation surveys will be dominated by (unknown) systematics

What there is to learn and how to get there

- A question of accuracy: first, avoid biases.
- A question of precision: can numerical forward models be used to push further than $k \gtrsim 0.15\ h/\mathrm{Mpc}$? The full field contains much more information.
- A question of scalability: the property of algorithms to handle a growing amount of data under computational resource constraints.
- The challenge is twofold:
  - in the data models: how can we best use modern computers and their architecture?
  - in the inference techniques: how can we perform rigorous Bayesian reasoning given a limited computational budget?



FL & Heavens, 2103.04158

What is forward modelling?

- Data analysis is the art of having the two ends meet…

Forward modelling → ← **"Backward modelling" assumptions, corrections, etc.**

Any step here is expensive and possibly complicated | Any step here is lossy and potentially biasing

Cosmological parameters → Initial conditions → Final conditions → Mock catalogues → Power spectrum ⟷ Power spectrum ← Observed catalogues ← Images / spectra ← Photons on instruments

- Ideally:

Forward modelling → Photons on instruments

- Less ideally, but still unrealistic:

Forward modelling → Images / spectra ← Photons on instruments

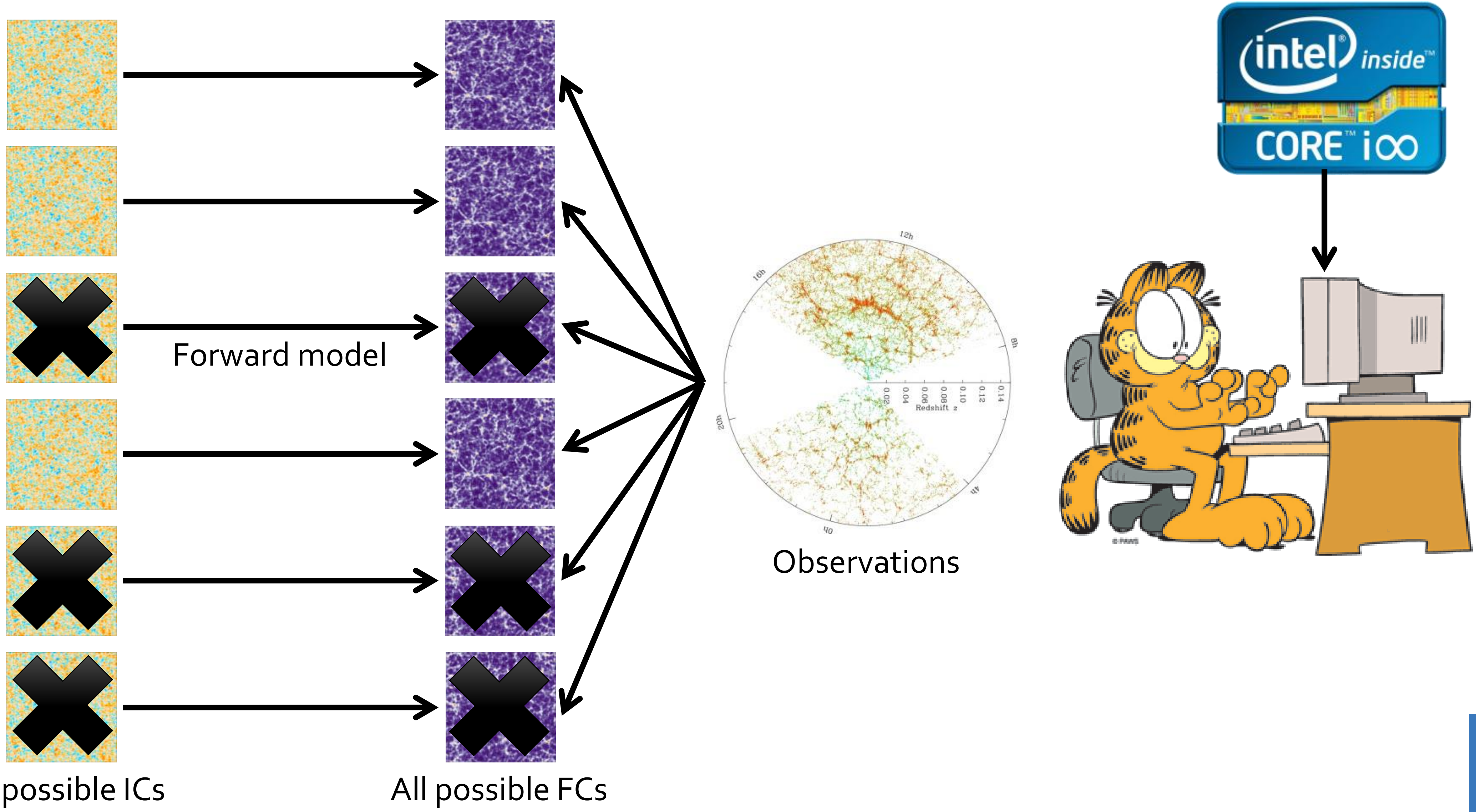
- This talk:

Forward modelling: (…) → Mock catalogues → Any summary ⟷ Any summary ← Observed catalogues ← (…)

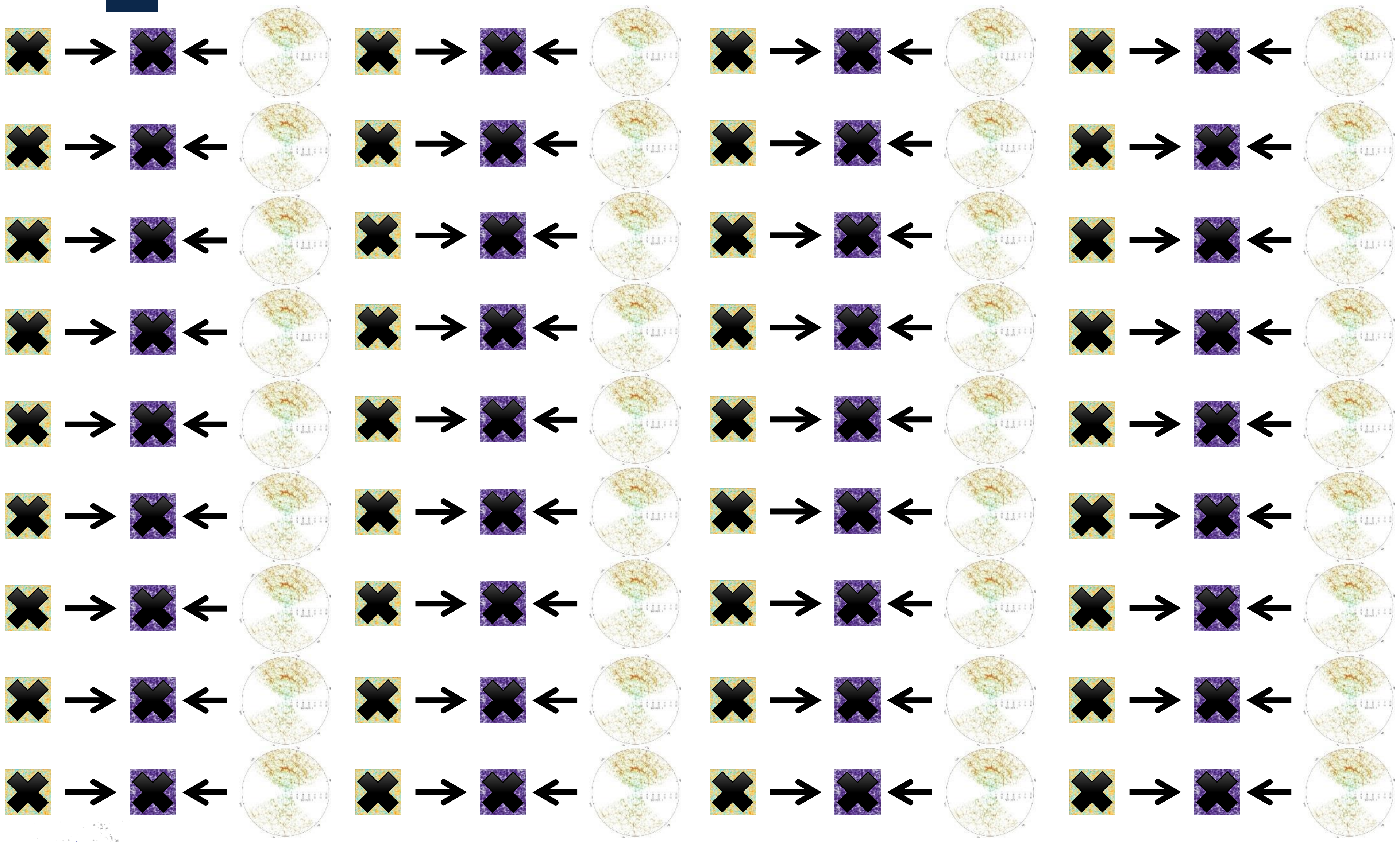
OR

Forward modelling: (…) → Mock catalogues ⟷ Observed catalogues ← (…)

Bayesian forward modelling: the ideal scenario



Bayesian forward modelling: the challenge

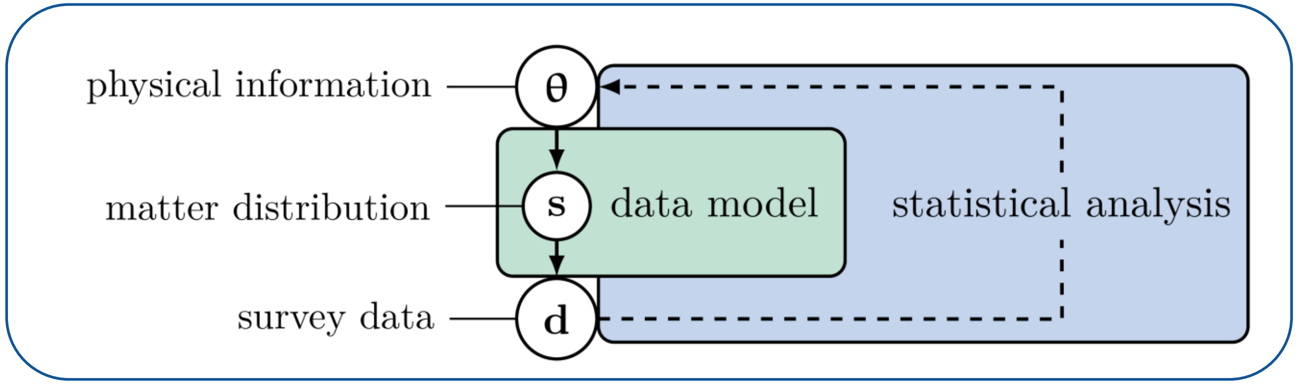


The (true) likelihood lives in

$d \approx 10^7$!

Making inferences requires advanced Bayesian techniques

- Complex computer models are incorporated into Bayesian hierarchical models:



- The challenge: using new statistical methods is necessary. Two approaches are possible:

<u>Data assimilation:</u>

exact statistical analysis

approximate data model

<u>Implicit likelihood inference:</u>

approximate statistical analysis

arbitrary data model

Implicit likelihood inference

Implicit likelihood inference:

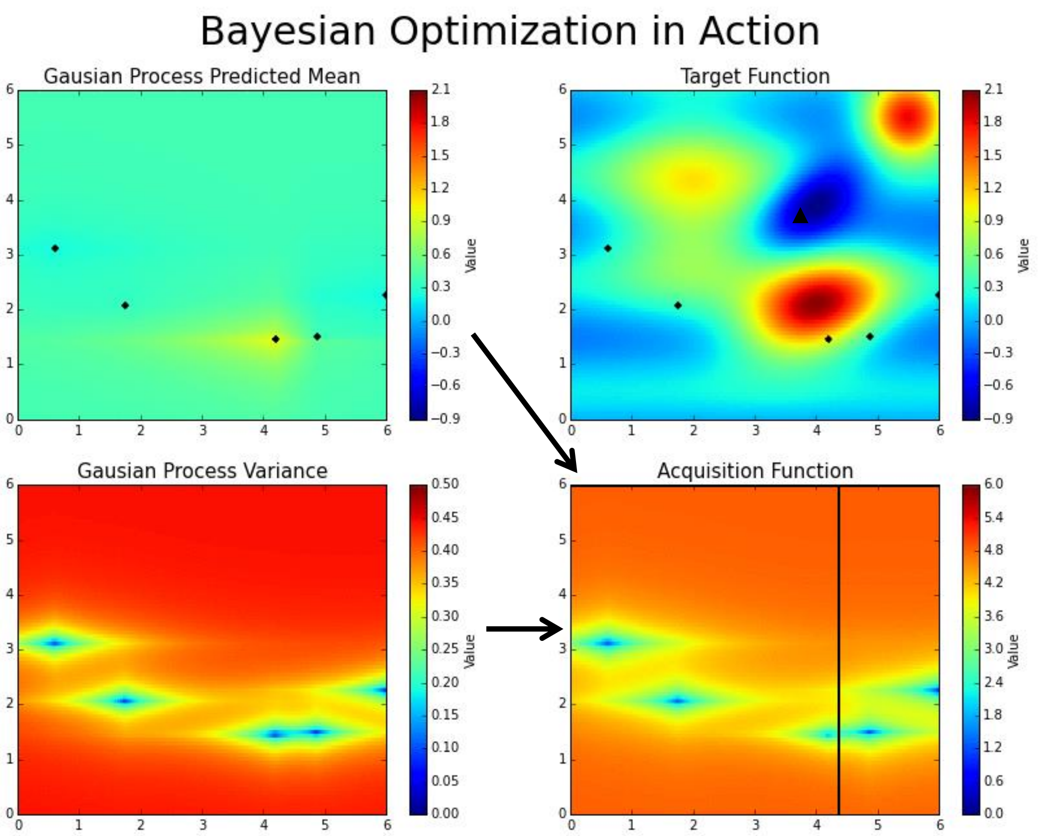
approximate statistical analysis

arbitrary data model

Bayesian Optimisation for Likelihood-Free Inference (BOLFI): An active data acquisition procedure to efficiently place simulations in parameter space

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

Regressed effective likelihood

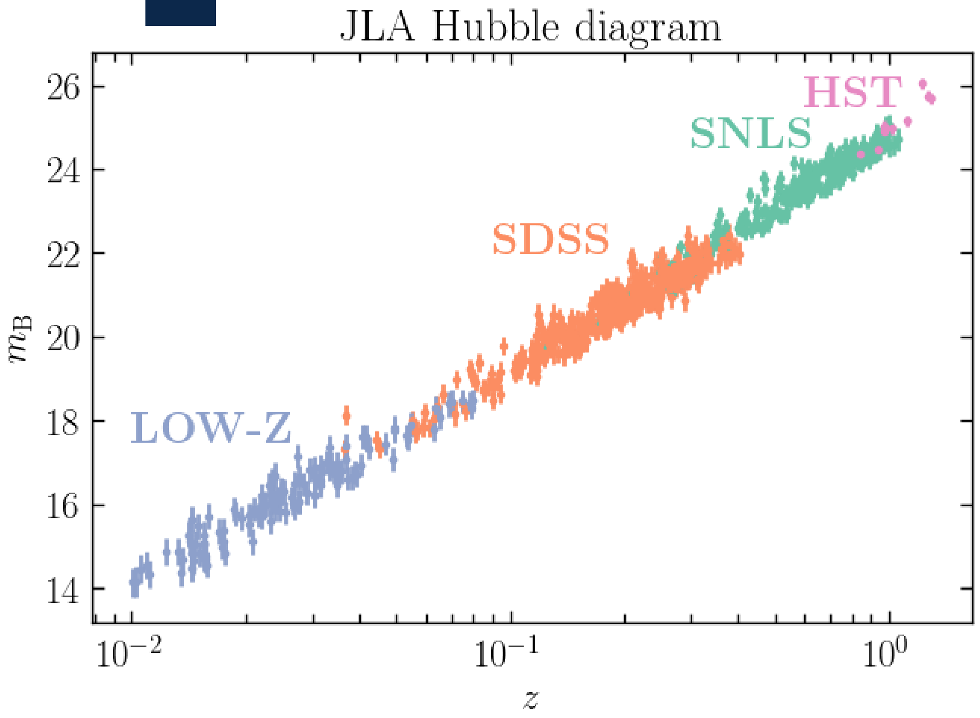


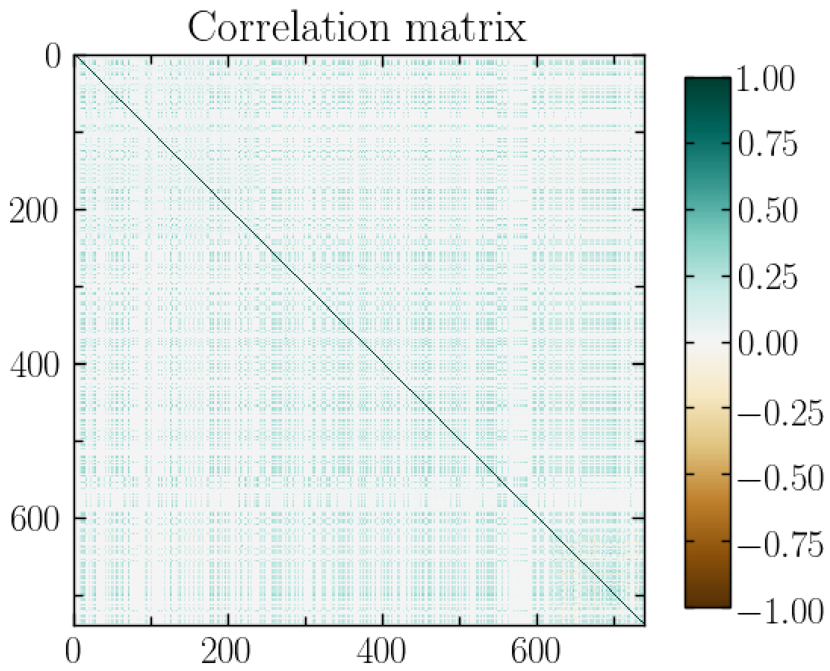
True (unknown) effective posterior

The acquisition function finds a compromise between exploration & exploitation

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

BOLFI: Re-analysis of the JLA supernova sample (Betoule *et al.*, 1401.4064)





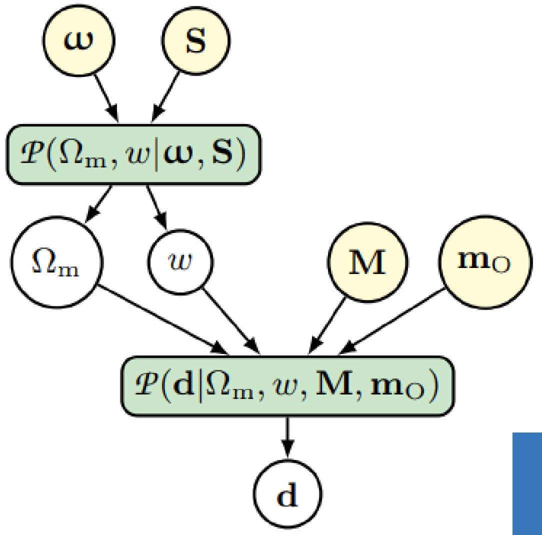
- 6-parameter model:
  2 cosmological parameters +
  4 nuisance parameters

$$m_B = 5 \log_{10}\left[\frac{D_L(z)}{10\ \mathrm{pc}}\right] + \widetilde{M}_B(M_\mathrm{stellar}, M_B, \delta M) - \alpha X_1 + \beta C$$

$$\widetilde{M}_B(M_\mathrm{stellar}, M_B, \delta M) = M_B + \delta M\, \Theta\left(M_\mathrm{stellar} - 10^{10} M_\odot\right)$$

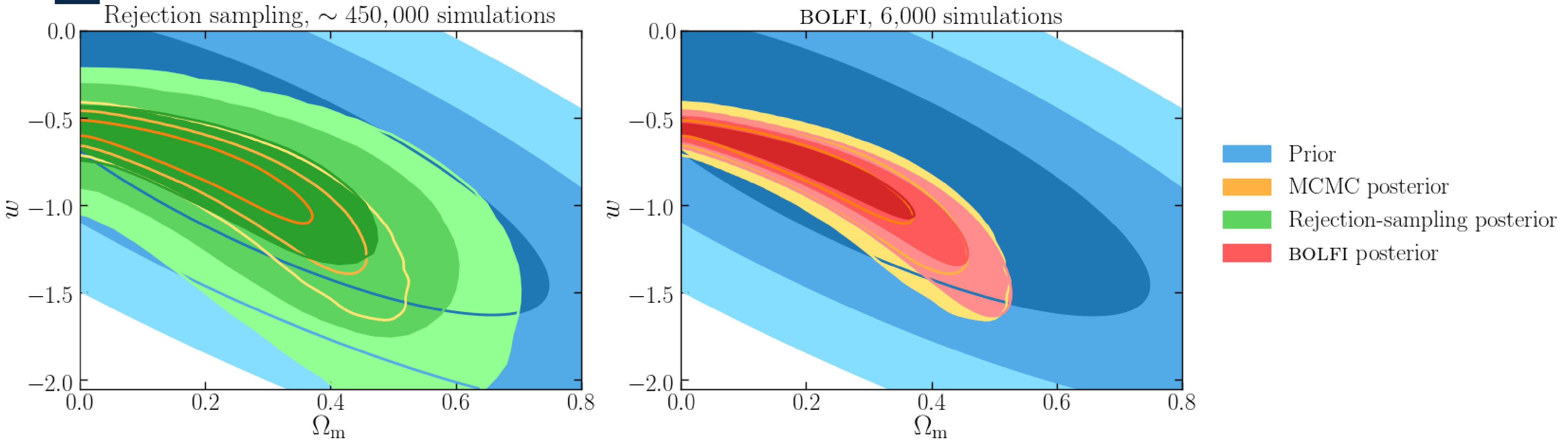
$$D_L(z) = \frac{(1+z)\,c}{H_0} \int_0^z \frac{\mathrm{d}z'}{E(z')}$$

$$E(z) \equiv \sqrt{\Omega_m (1+z)^3 + (1-\Omega_m)(1+z)^{3(w+1)}}$$



FL, 1805.07152

BOLFI: Re-analysis of the JLA supernova sample (Betoule *et al.*, 1401.4064)



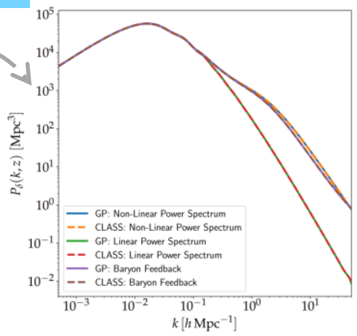
- The number of required simulations is reduced by:
  - 2 orders of magnitude with respect to likelihood-free rejection sampling (for a much better approximation of the posterior),
  - 3 orders of magnitude with respect to exact Markov Chain Monte Carlo sampling.
- Bayesian optimisation can also be applied to the "true" likelihood (if known) or to iteratively build an emulator of the data model.

FL, 1805.07152

Speed up & go beyond approximations

Emulators

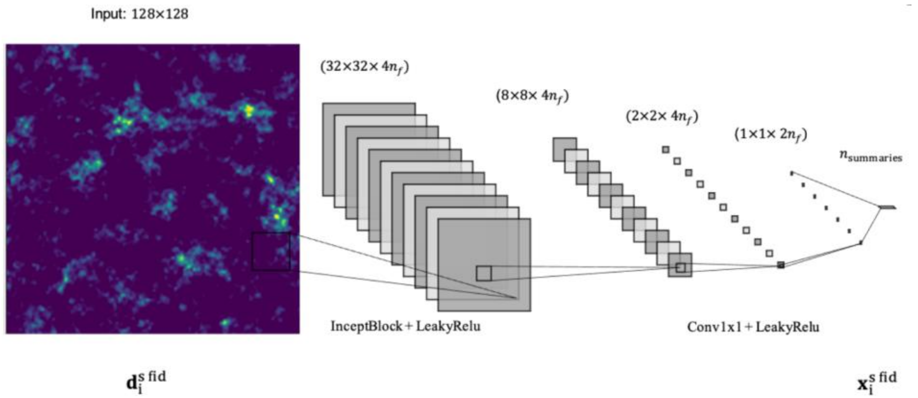
ΛCDM



emuPK: Mootoovaloo, Jaffe, Heavens & FL, 2105.02256

Find the information content

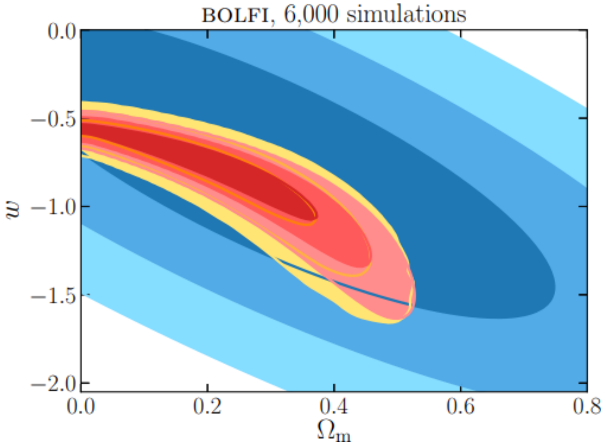
Automatic data compression



Information Maximising Neural Networks (IMNN): Charnock, Lavaux & Wandelt, 1802.03537; Makinen *et al.*, 2107.07405

Build a posterior/evidence approximator

Implicit likelihood inference



Bayesian Optimisation for Likelihood-Free Inference (BOLFI): FL, 1805.07152

Field-level inference via data assimilation

Data assimilation:

exact statistical analysis

approximate data 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

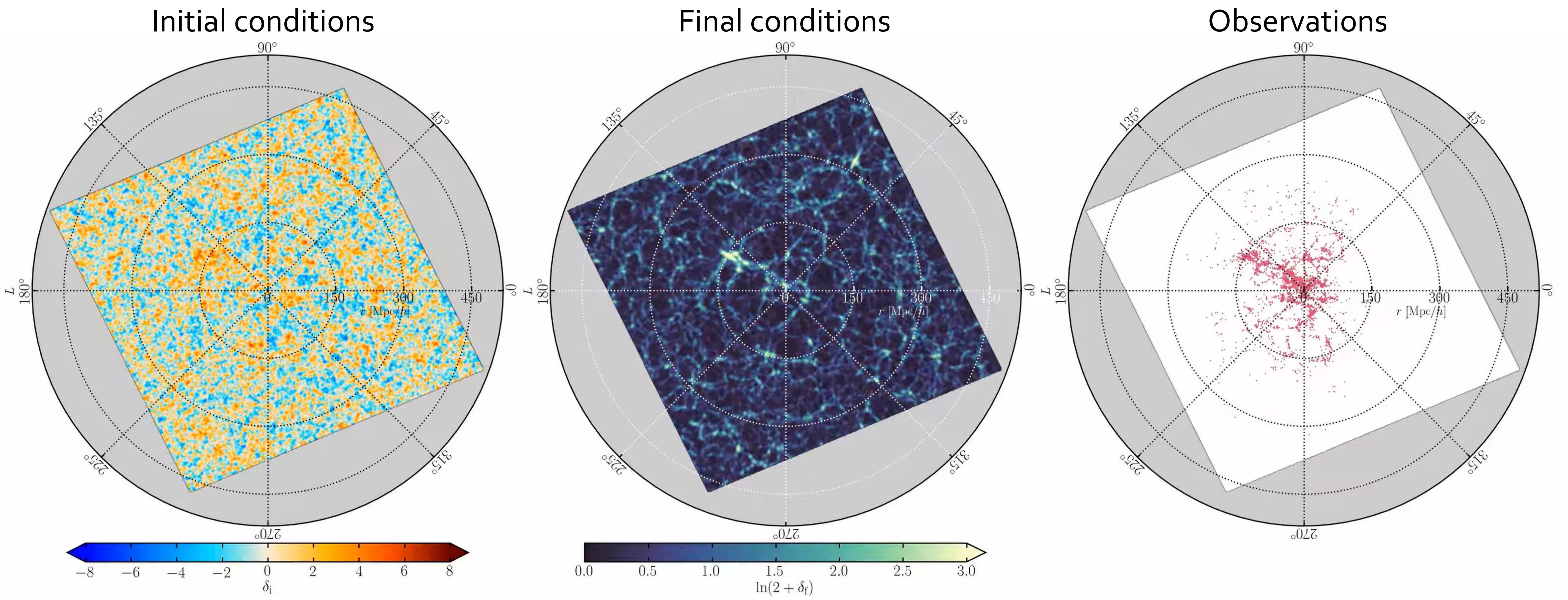
$$a(\mathbf{x}',\mathbf{x}) = \mathrm{e}^{-(H'-H)} = 1$$

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

Field-level inference in practice:
Bayesian Origin Reconstruction from Galaxies (BORG)

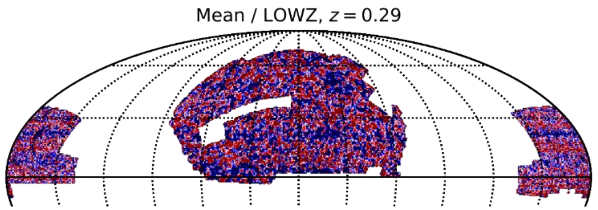


67,224 galaxies, ≈ 17 million parameters, 5 TB of primary data products, 10,000 samples, ≈ 500,000 forward and adjoint gradient data model evaluations, 1.5 million CPU-hours

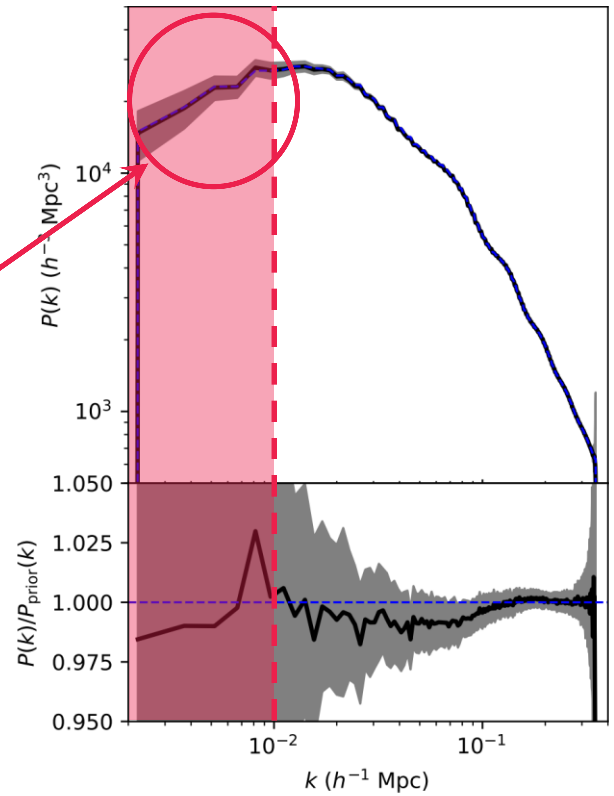
Jasche & Wandelt, 1203.3639; Jasche, FL & Wandelt, 1409.6308; Jasche & Lavaux, 1806.11117; Lavaux, Jasche & FL, 1909.06396

Machine-aided report of unknown data contaminations
Application to SDSS-III/BOSS (LOWZ+CMASS)

Map of unknown foreground contaminant

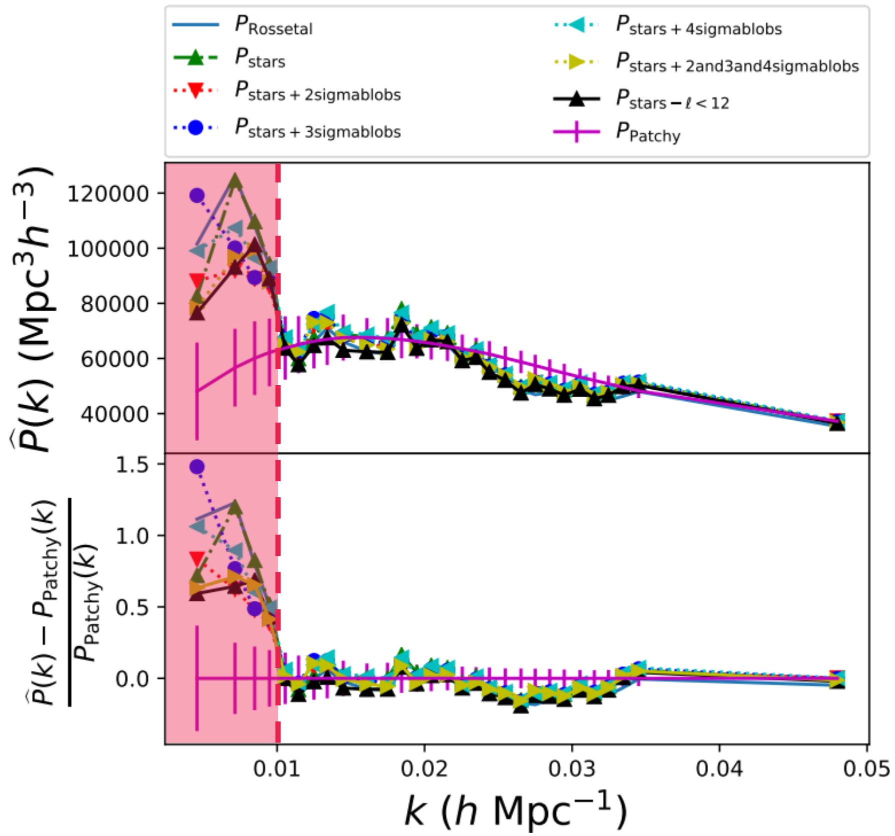


BORG *a posteriori* power spectrum



No apparent contamination, even well beyond the turn-over

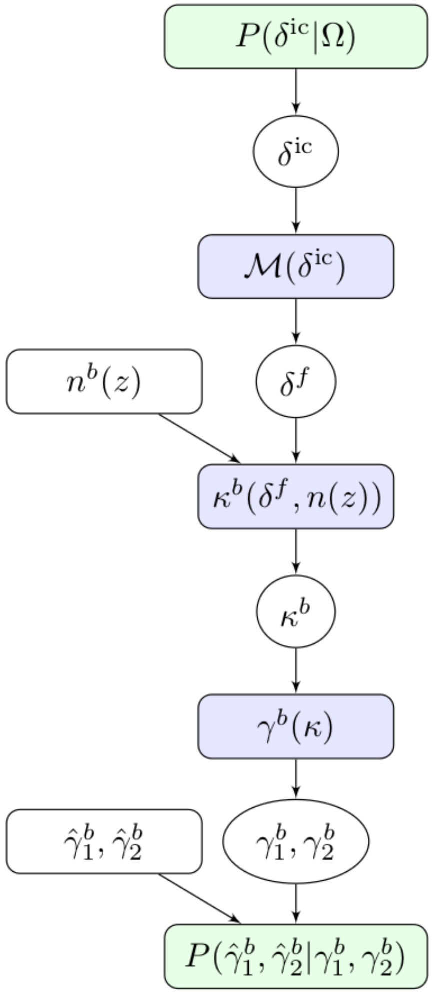
State-of-the-art with backward-modelling technique (mode subtraction)

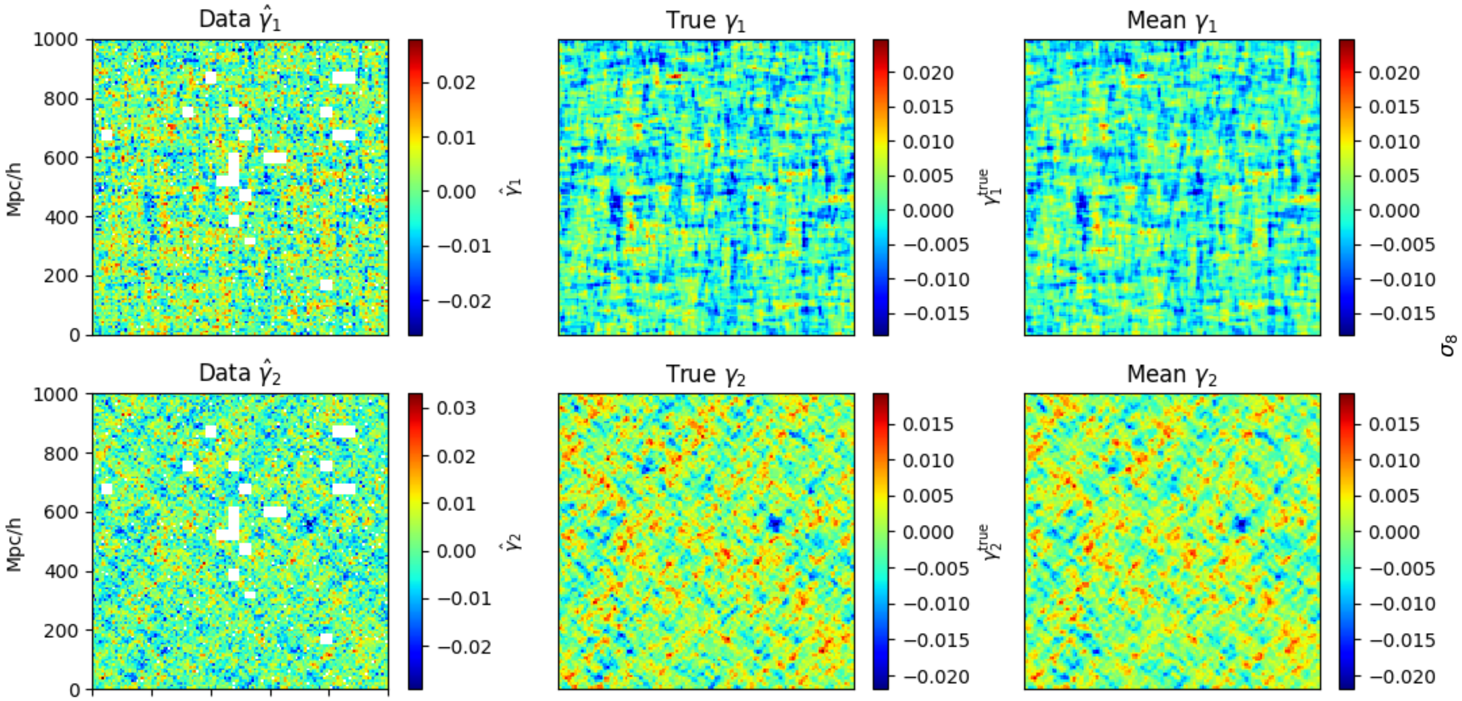


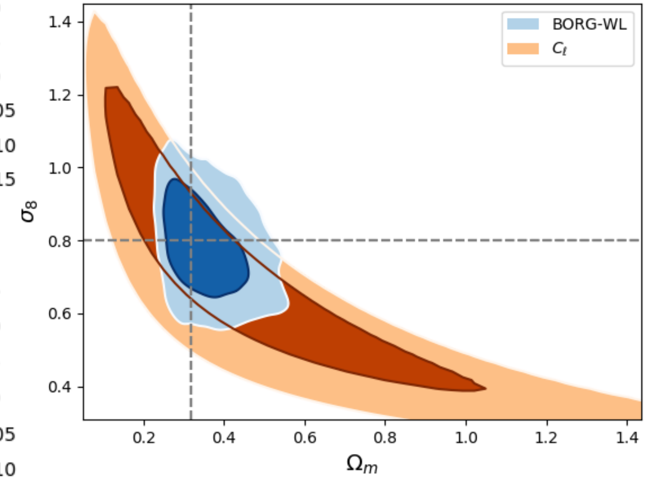
Porqueres, Ramanah, Jasche & Lavaux, 1812.05113
Lavaux, Jasche & FL, 1909.06396

Kalus, Percival *et al.*, 1806.02789

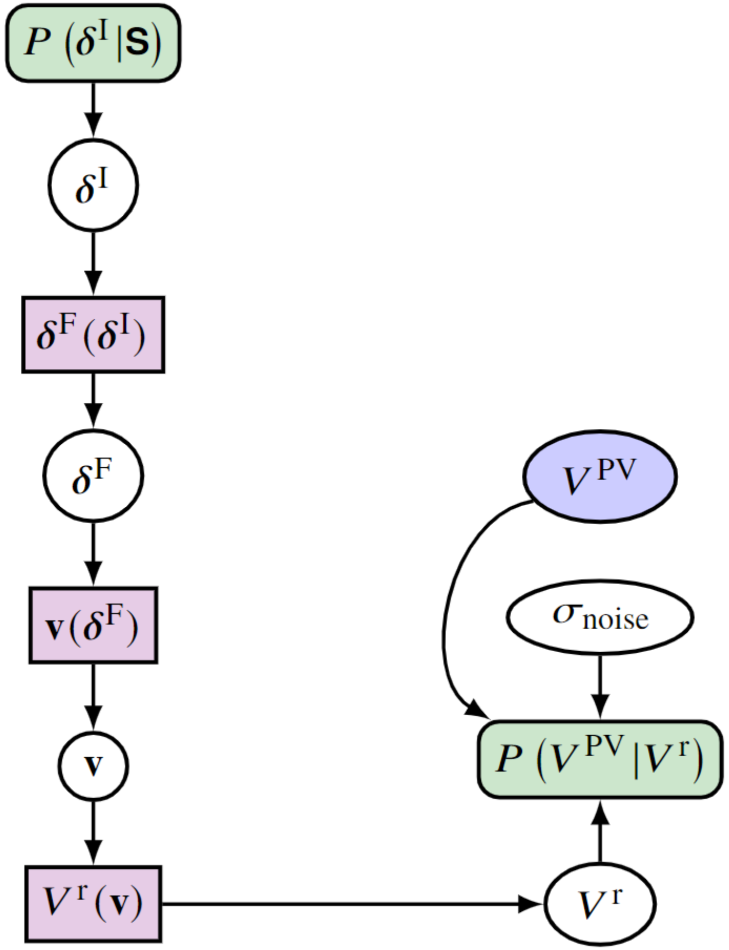
Extending BORG: weak lensing field-level inference using shear and convergence data

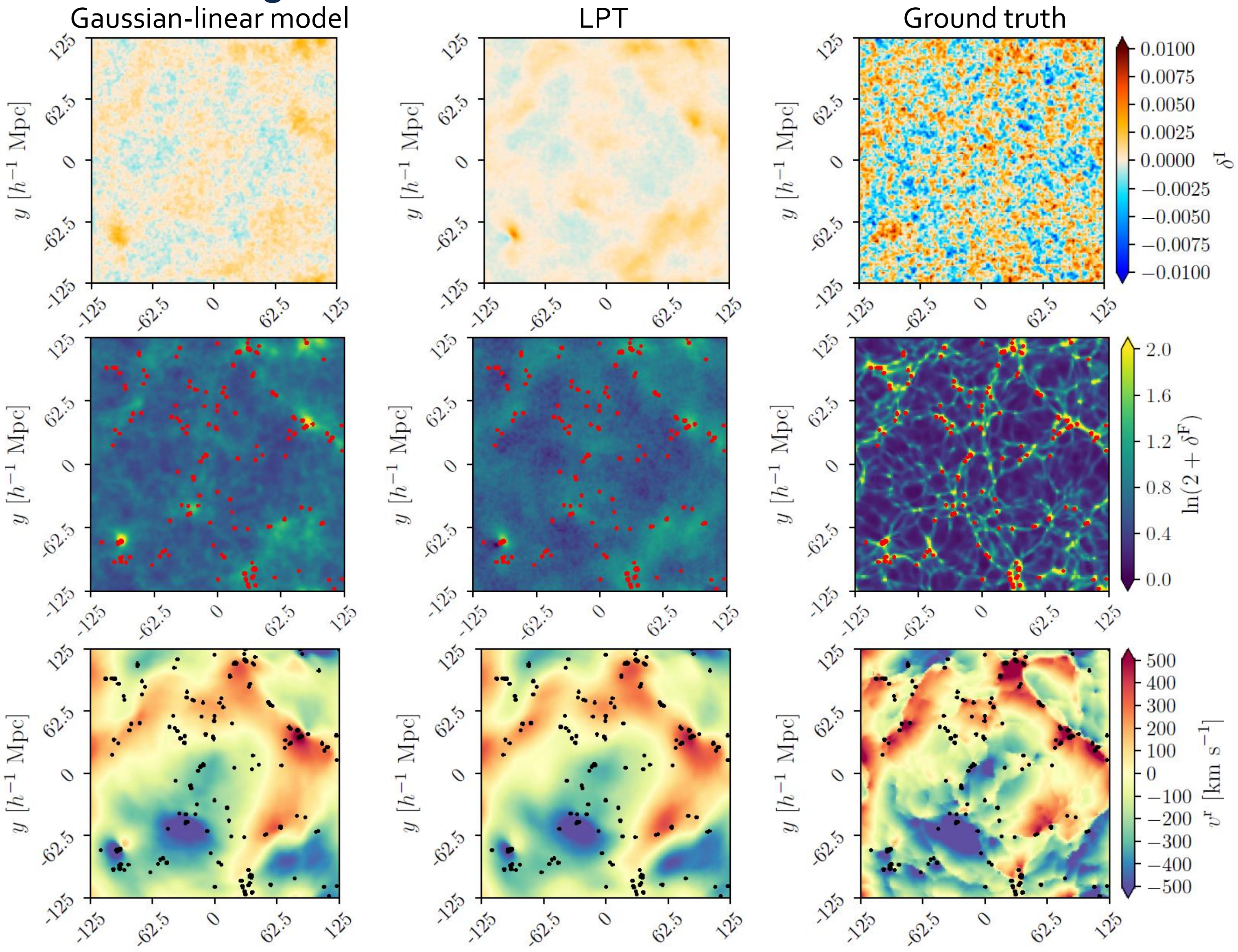






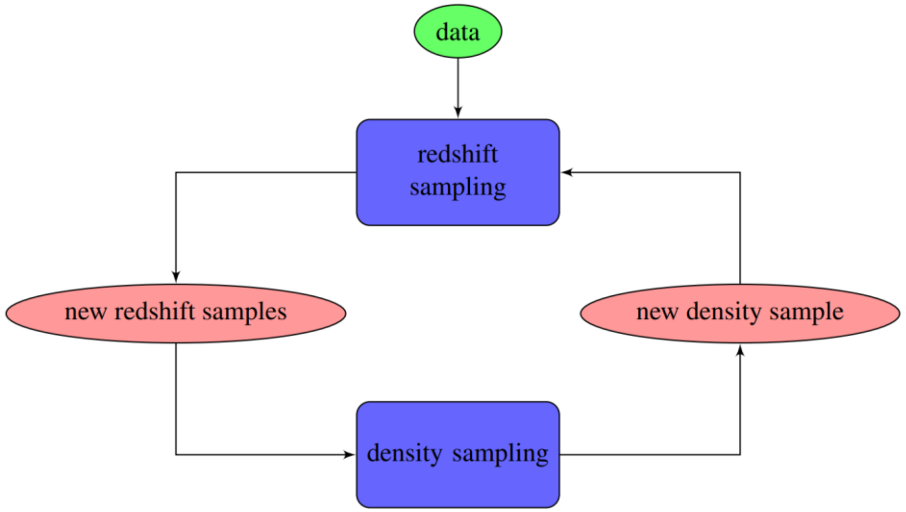
Porqueres, Heavens, Mortlock & Lavaux, 2011.07722; Porqueres, Heavens, Mortlock & Lavaux, 2108.04825

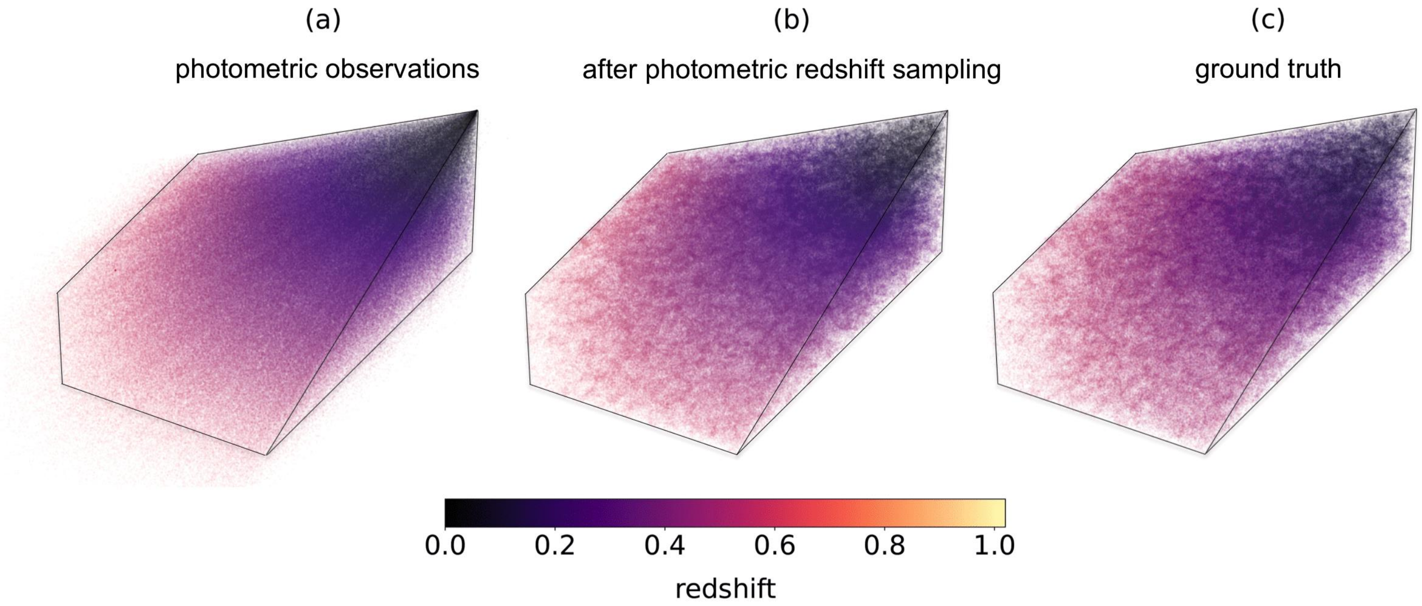


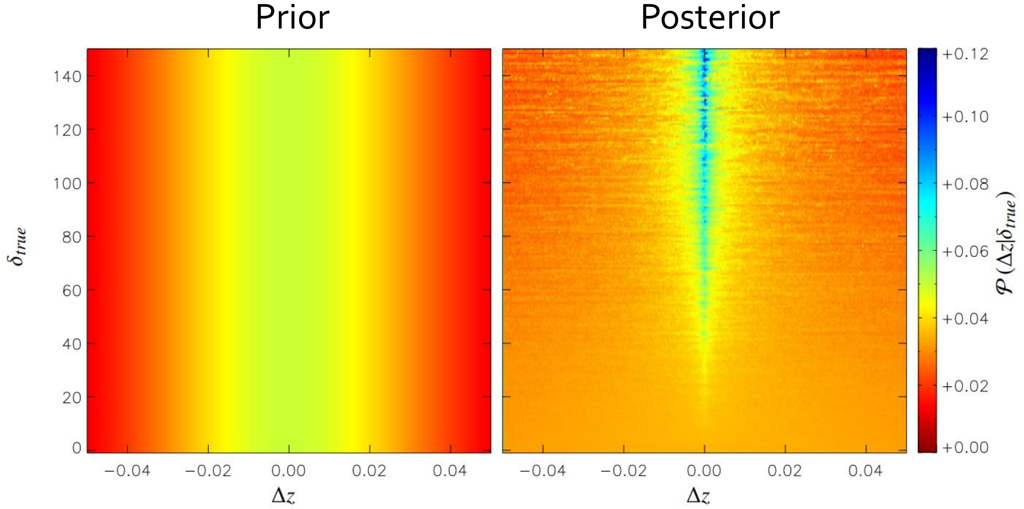


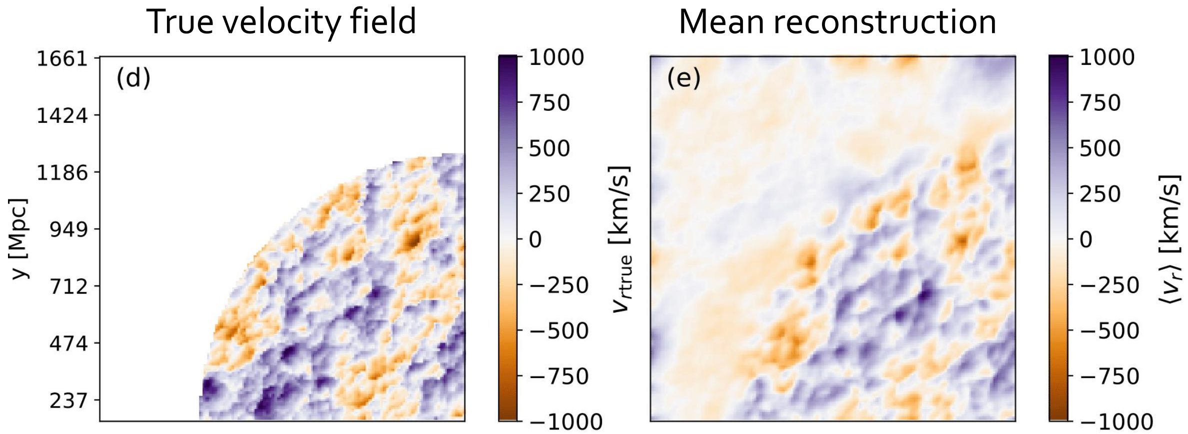
Lavaux, 1512.04534; Boruah, Lavaux & Hudson, 2111.15535; Prideaux-Ghee, FL, Lavaux, Heavens & Jasche, 2204.00023

Extending BORG: joint inference of fields and photometric redshifts





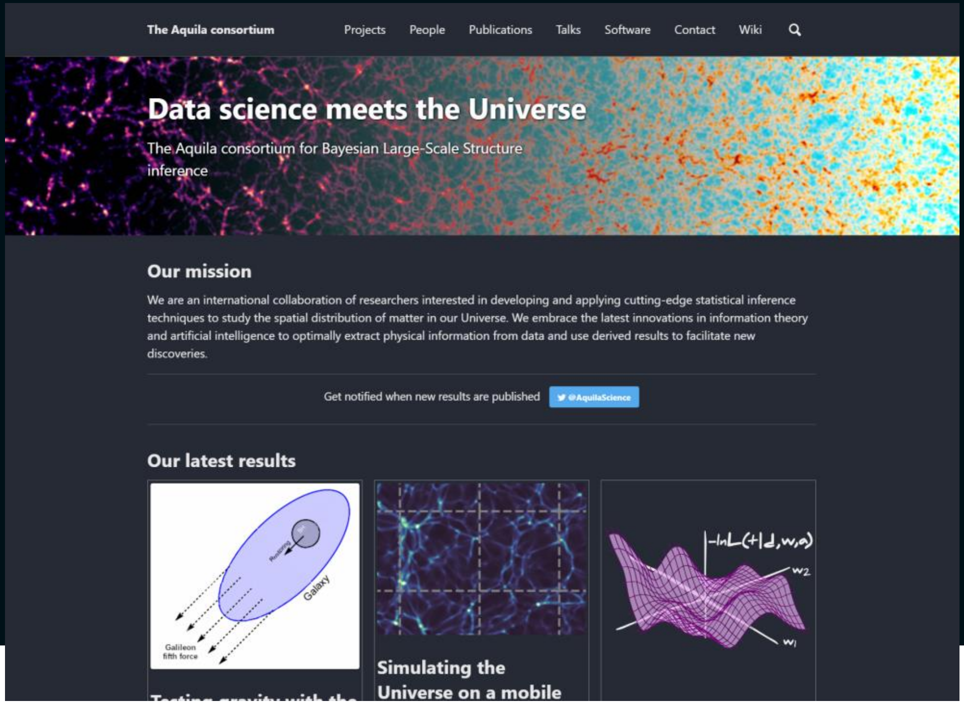




Jasche & Wandelt, 1106.2757; Tsaprazi, Jasche, Lavaux & FL, in prep.

The Aquila Consortium

- Created in 2016. Currently 38 members from 8 countries (Europe & Americas).
- Gathers people interested in developing Bayesian pipelines and running analyses on cosmological data.



Visit us at www.aquila-consortium.org

Concluding thoughts

Data assimilation:	Implicit likelihood inference:
exact statistical analysis	approximate statistical analysis
approximate data model	arbitrary data model

- Bayesian analyses of galaxy surveys with fully non-linear numerical models is not an impossible task!
- Implicit likelihood inference – a likelihood-free solution (BOLFI): algorithm for targeted questions, allowing the use of accurate simulators including all relevant physical and observational effects.
- Field-level inference via data assimilation – a likelihood-based solution (BORG): general purpose inference of the initial conditions from cosmological observables (galaxy clustering, weak lensing, distance tracers), providing new measurements and predictions.