# Galaxy Clustering with Likelihood-Free Inference

## Prospects and Forecasts for Euclid


**Florent Leclercq**

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

Wolfgang Enzi, Baptiste Faure, Alan Heavens,
Andrew Jaffe, Jens Jasche, Guilhem Lavaux,
Will Percival, Benjamin Wandelt

and the Aquila Consortium
www.aquila-consortium.org


20 April 2020

Euclid consortium

ICIC Imperial Centre for Inference & Cosmology

Imperial College London

# Forward modelling within the Euclid Consortium

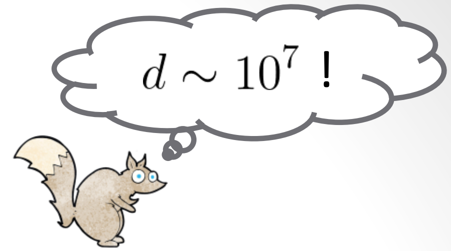
- Galaxy Clustering Science Working Group (**GC SWG**)
  - WP Likelihood + IST:L:
    - "Methods to speed up the GC likelihood computation" (includes BOLFI)
  - WP Additional Probes:
    - "Density reconstruction via Bayesian large-scale structure inference" (ARES, HADES, BORG)
    - "Primordial power spectrum from black-box galaxy surveys" (SELFI)
- Weak Lensing Science Working Group (**WL SWG**):
  - WP Forward-modelling (DELFI, BOLFI, SELFI)
- Theory Science Working Group (**TH SWG**):
  - WP Initial conditions (BORG with $f_{\mathrm{NL}}$)
- Cosmological Simulations Science Working Group (**SIM SWG**)
  - WP Machine Learning (emulators, neural networks)

---

- Galaxy clustering with likelihood-free inference (GCLFI) will be proposed as a SP.

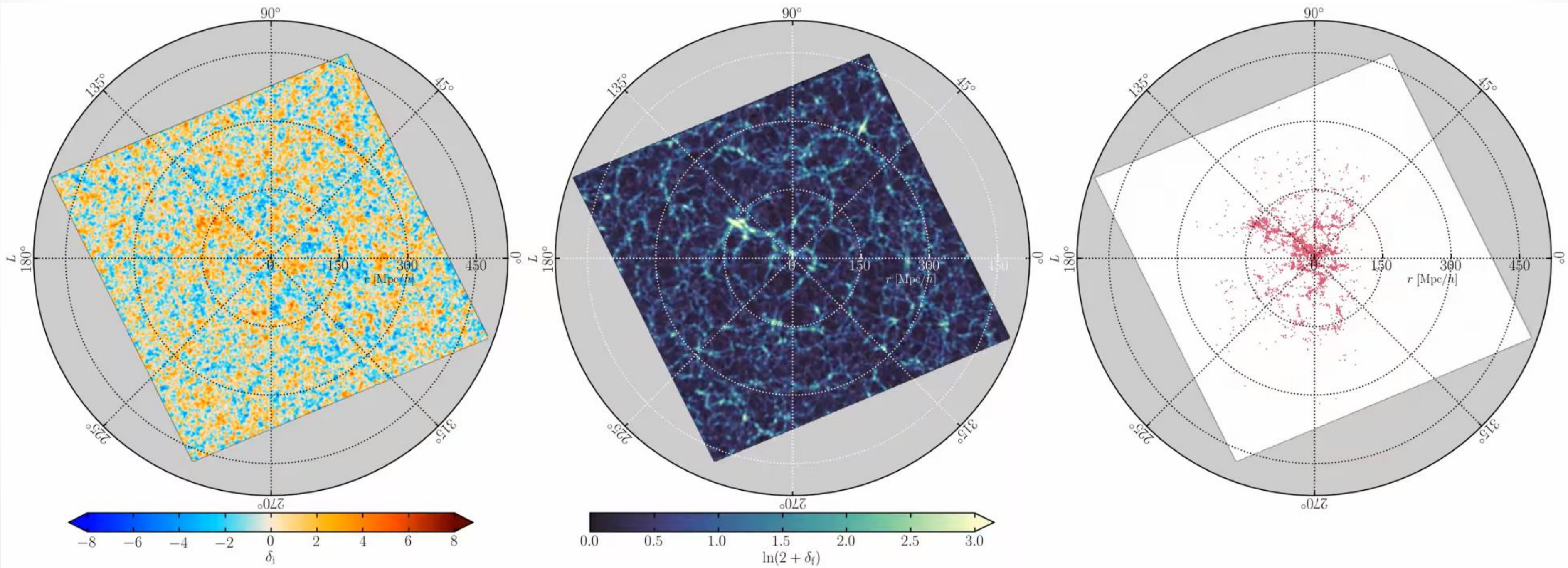
# Vocabulary consideration:

*What is the likelihood?*




In cosmology, the (true?) likelihood should live at the level of the **map** of the CMB or LSS.
e.g. Wiener filtering for the CMB, BORG for the LSS (a $256^3$-dimensional Poisson likelihood):

**Initial conditions** **Final conditions** **Observations**



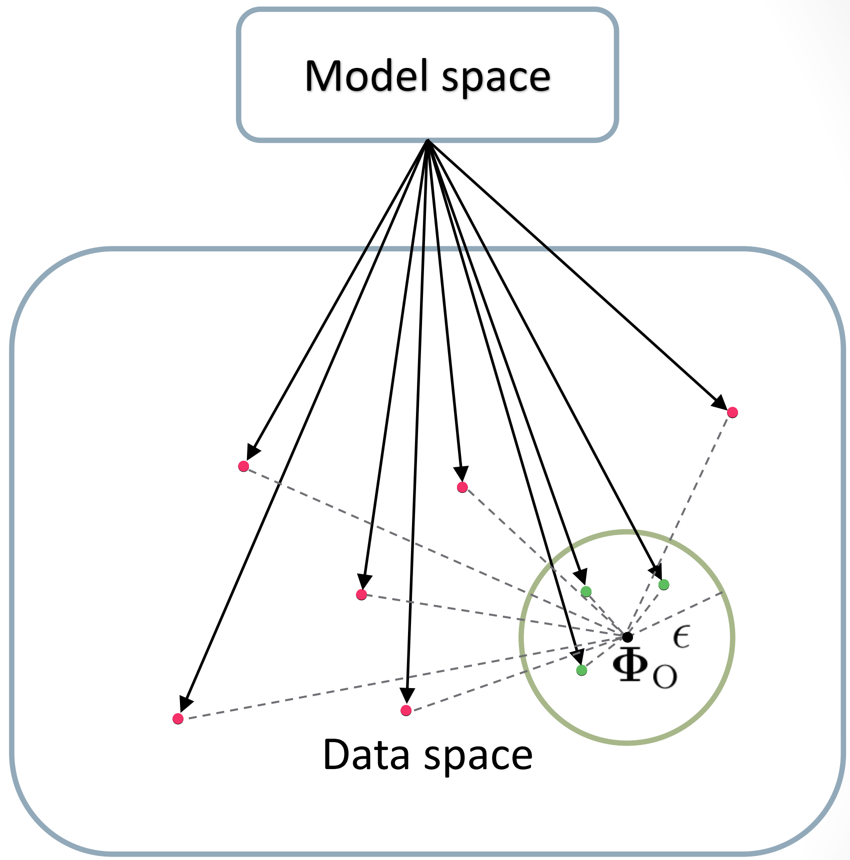

Jasche & Lavaux 2019, 1806.11117 – FL, Lavaux & Jasche, in prep.

**Expert knowledge** of the likelihood is needed to beat the curse of dimensionality: conditionals/gradients of the likelihood are required by the samplers (Gibbs/Hamiltonian).

# Likelihood-free rejection sampling (LFRS)

- Iterate many times:
  - Sample $\theta$ from a proposal distribution $q(\theta)$
  - Simulate $\Phi_\theta$ using the black-box
  - Compute the distance $\Delta(\Phi_\theta, \Phi_\mathrm{O})$ between simulated and observed data
  - Retain $\theta$ if $\Delta(\Phi_\theta, \Phi_\mathrm{O}) \leq \epsilon$, otherwise reject

$\epsilon$ can be adaptively reduced (Population Monte Carlo)



# Beyond LFRS: two scenarios

The "number of simulations" route:

- Specific cosmological models ($d \lesssim 10$), general exploration of parameter space
- Density Estimation for Likelihood-Free Inference (DELFI)
  Papamakarios & Murray 2016, 1605.06376
  Alsing, Feeney & Wandelt 2018, 1801.01497
- Bayesian Optimisation for Likelihood-Free Inference (BOLFI)
  Gutmann & Corander 2016, 1501.03291
  FL 2018, 1805.07152

The "number of parameters" route:

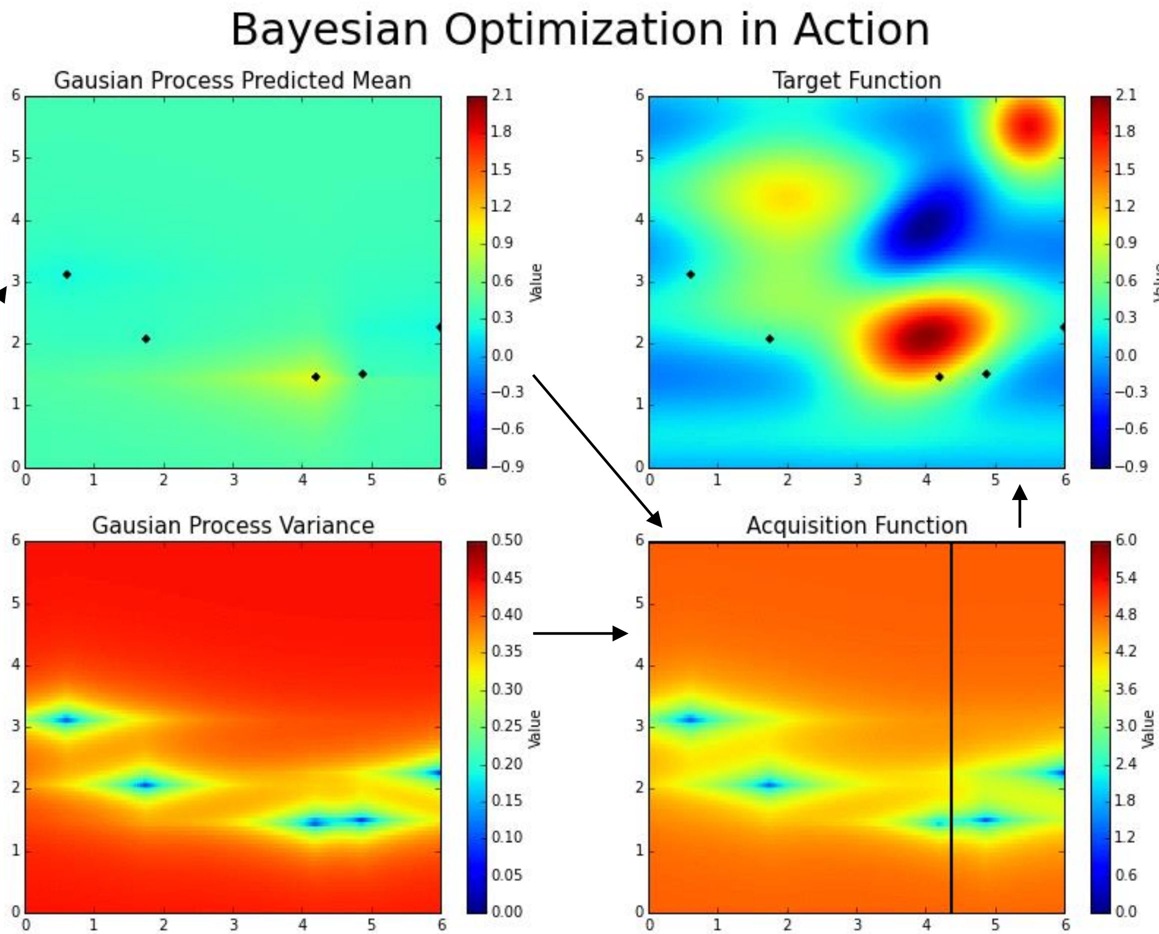
- Model-independent theoretical parametrisation ($d \gtrsim 100$), strong existing constraints in parameter space
- Simulator Expansion for Likelihood-Free Inference (SELFI)
  FL, Enzi, Jasche & Heavens 2019, 1902.10149

# The “number of simulations” route: BOLFI

*Bayesian Optimisation for Likelihood-Free Inference*

# BOLFI: Data acquisition

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




Regressed effective likelihood

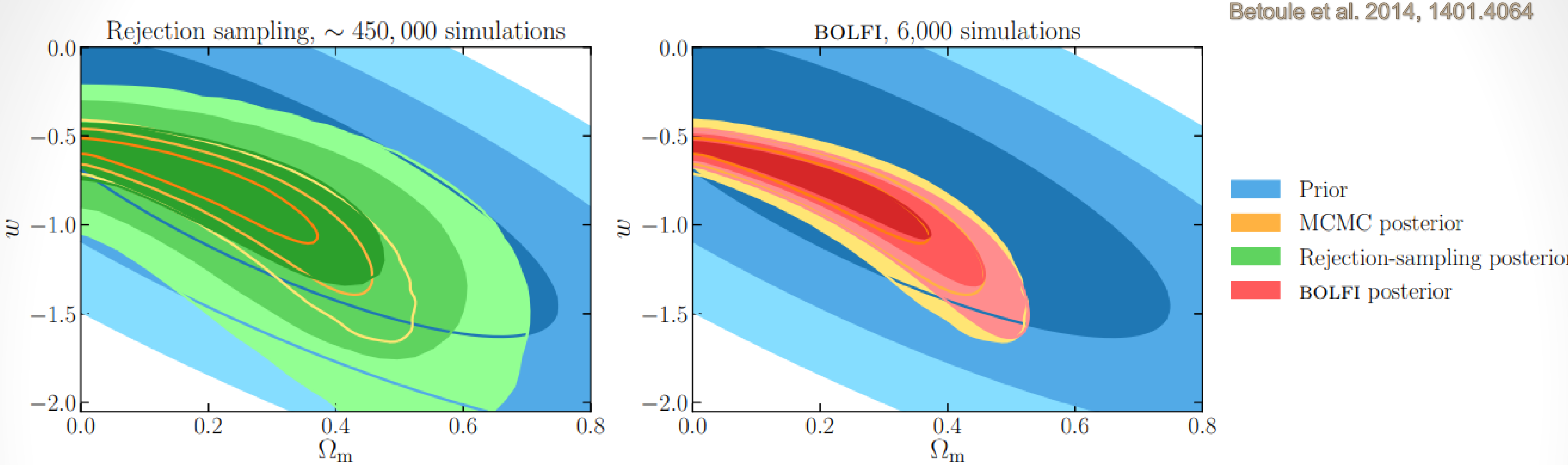
True (unknown) effective posterior

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. 2014, 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

FL 2018, 1805.07152

- Bayesian optimisation can also be applied to the “true” likelihood (if known) or to build an emulator of the data model: see derivative work by

Rogers *et al.* 2019, 1812.04631 – Takhtaganov *et al.* 2019, 1905.07410 – Pellejero-Ibañez *et al.* 2019, 1912.08806 and in the WP:Lik

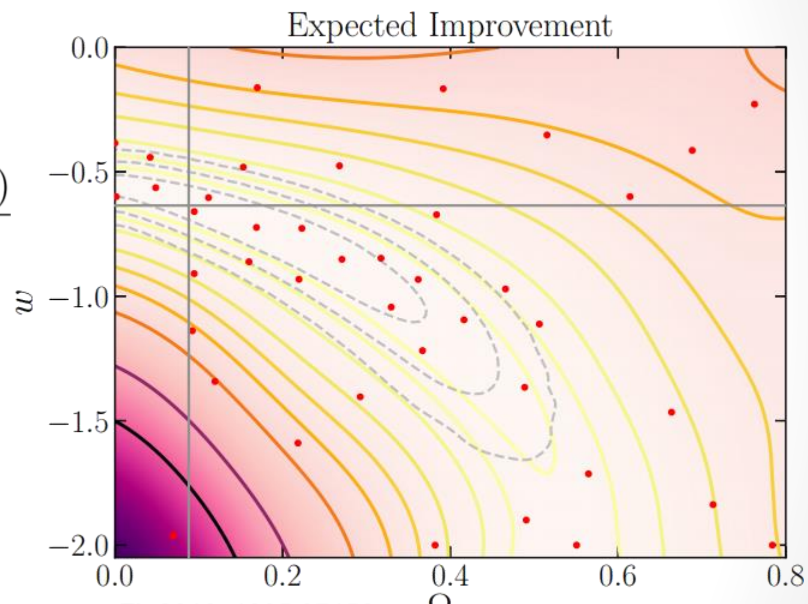
# Standard acquisition functions are suboptimal (at best)

- Goal for Bayesian optimisation: finding the optimum (assumed unique) of a function (≠ exploring a parameter space)
- Examples of acquisition functions :
  - The Expected Improvement

Gaussian cdf — Gaussian pdf

$$\mathrm{EI}(\boldsymbol{\theta}_\star) \equiv \underbrace{\sigma(\boldsymbol{\theta}_\star)}_{\text{Exploration}} \underbrace{\left[z\Phi(z) + \phi(z)\right]}_{\text{Exploitation}} \qquad z \equiv \frac{\min(\mathbf{f}) - \mu(\boldsymbol{\theta}_\star)}{\sigma(\boldsymbol{\theta}_\star)}$$

e.g. Brochu, Cora & de Freitas 2010, 1012.2599 – Takhtaganov *et al*. 2019, 1905.07410



FL 2018, 1805.07152

  - The Upper Confidence Bound

$$\mathrm{UCB}(\boldsymbol{\theta}_\star) = \mu(\boldsymbol{\theta}_\star) + \alpha\sigma(\boldsymbol{\theta}_\star)$$

Rogers *et al*. 2019, 1812.04631 – Pellejero-Ibañez *et al*. 2019, 1912.08806

- Drawbacks of these acquisition rules:
  - Do not take into account prior information
  - Local evaluation rules
  - Too greedy for parameter inference

⇨ In most cases these standard rules will lead to a suboptimal exploration of parameter space. But in some situations (e.g. posteriors with non-trivial degeneracies or multiple modes) they can even yield a biased inference.

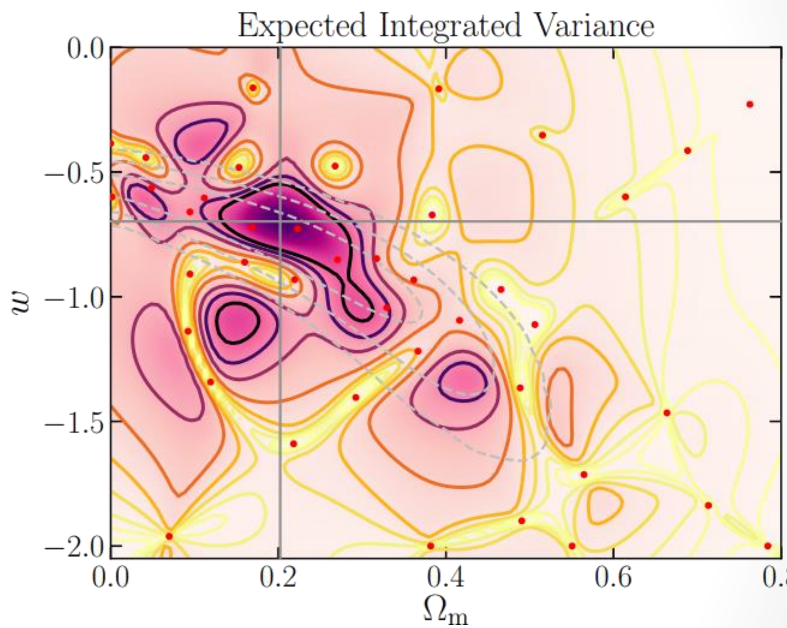
# The optimal acquisition function for parameter inference

- Goal for parameter inference: minimise the expected uncertainty in the estimate of the (approximate) posterior over the future evaluation of the simulator
- The optimal acquisition function : the **Expected Integrated Variance**

$$\mathrm{EIV}(\boldsymbol{\theta}_\star) = \underbrace{\int}_{\text{Integral}} \underbrace{\frac{\mathcal{P}(\boldsymbol{\theta})^2}{4}}_{\text{Prior}} \underbrace{\exp\left[-\mu(\boldsymbol{\theta})\right]}_{\text{Exploitation}} \underbrace{\left[\sigma^2(\boldsymbol{\theta}) - \tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)\right]}_{\text{Exploration}} \, \mathrm{d}\boldsymbol{\theta}$$

$$\tau^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) \equiv \frac{\mathrm{cov}^2(\boldsymbol{\theta}, \boldsymbol{\theta}_\star)}{\sigma^2(\boldsymbol{\theta}_\star)}$$

- Advantages:
  - Takes into account the prior
  - Non-local (integral over parameter space): more expensive… but much more informative
  - Exploration of the posterior tails is favoured when necessary
  - Analytic gradient

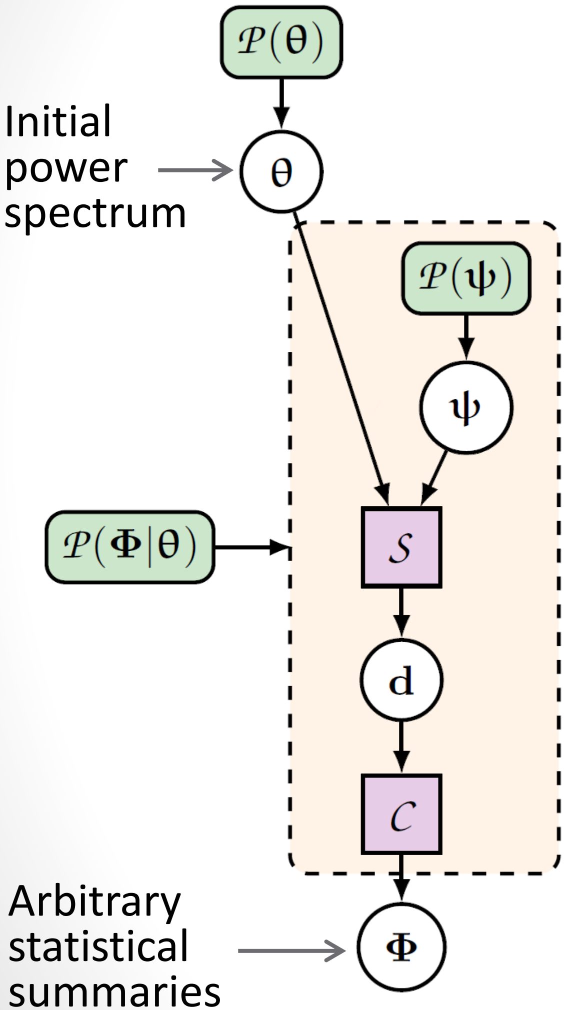



Järvenpää et al. 2017, 1704.00520 (expression of the EIV in the non-parametric approach)
FL 2018, 1805.07152 (expression of the EIV in the parametric approach)

# The "number of parameters" route: SELFI

*Simulator Expansion for Likelihood-Free Inference*

# SELFI: expansion of black-box data models

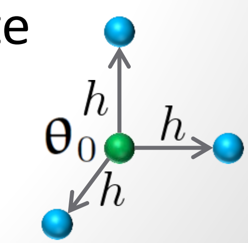



- We aim at inferring the initial power spectrum, which contains (almost?) all of the information
- This requires doing LFI in $d = \mathcal{O}(100) - \mathcal{O}(1,000)$
- If we trust the results of earlier experiments, we can Taylor-expand the black-box around an expansion point $\theta_0$:

$$\hat{\boldsymbol{\Phi}}_\theta \approx \mathbf{f}_0 + \nabla \mathbf{f}_0 \cdot (\theta - \theta_0) \boxed{+ \frac{1}{2} (\theta - \theta_0)^\intercal \cdot \mathbf{H} \cdot (\theta - \theta_0)} + \dots$$

SELFI-2 (second-order): coming soon!

- Gradients, Hessian matrix, etc. of the black-box can be evaluated via finite differences in parameter space



# SELFI-1: linearization of the black-box

- Linearization of the black-box:

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{\theta}} \approx \mathbf{f}_0 + \nabla\mathbf{f}_0 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

- Gaussian prior + Gaussian effective likelihood

The posterior is Gaussian and analogous to a Wiener filter:

expansion point

observed summaries

$$\boldsymbol{\gamma} \equiv \boldsymbol{\theta}_0 + \boldsymbol{\Gamma}\,(\nabla\mathbf{f}_0)^\intercal\,\mathbf{C}_0^{-1}(\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0)$$

$$\boldsymbol{\Gamma} \equiv \left[(\nabla\mathbf{f}_0)^\intercal\,\mathbf{C}_0^{-1}\nabla\mathbf{f}_0 + \mathbf{S}^{-1}\right]^{-1}$$

prior covariance

covariance of summaries

gradient of the black-box

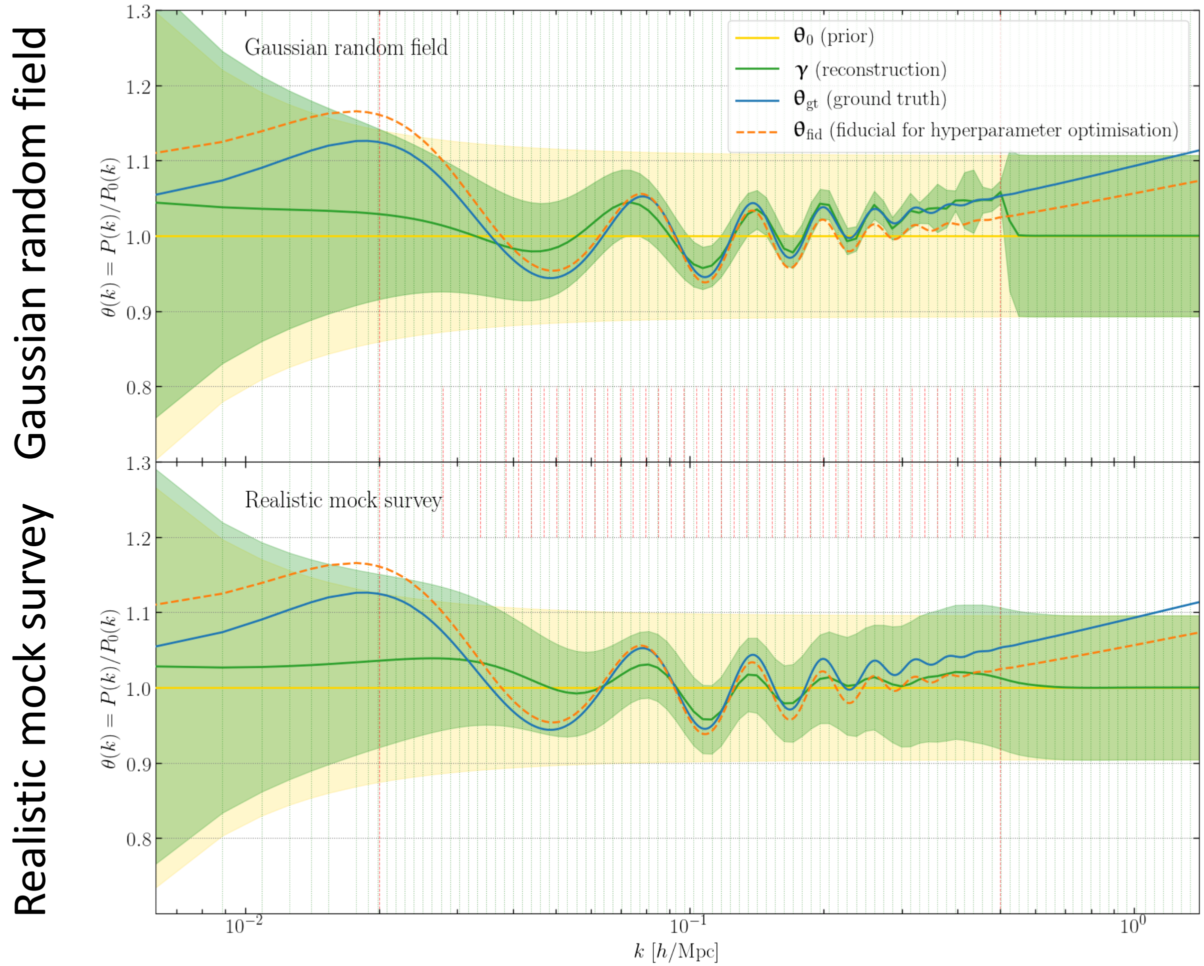
$\mathbf{f}_0$, $\mathbf{C}_0$ and $\nabla\mathbf{f}_0$ can be evaluated through simulations only.

The number of required simulations is fixed *a priori* (contrary to MCMC).

The workload is perfectly parallel.

FL, Enzi, Jasche & Heavens 2019, 1902.10149

# SELFI + numerical model: Proof-of-concept



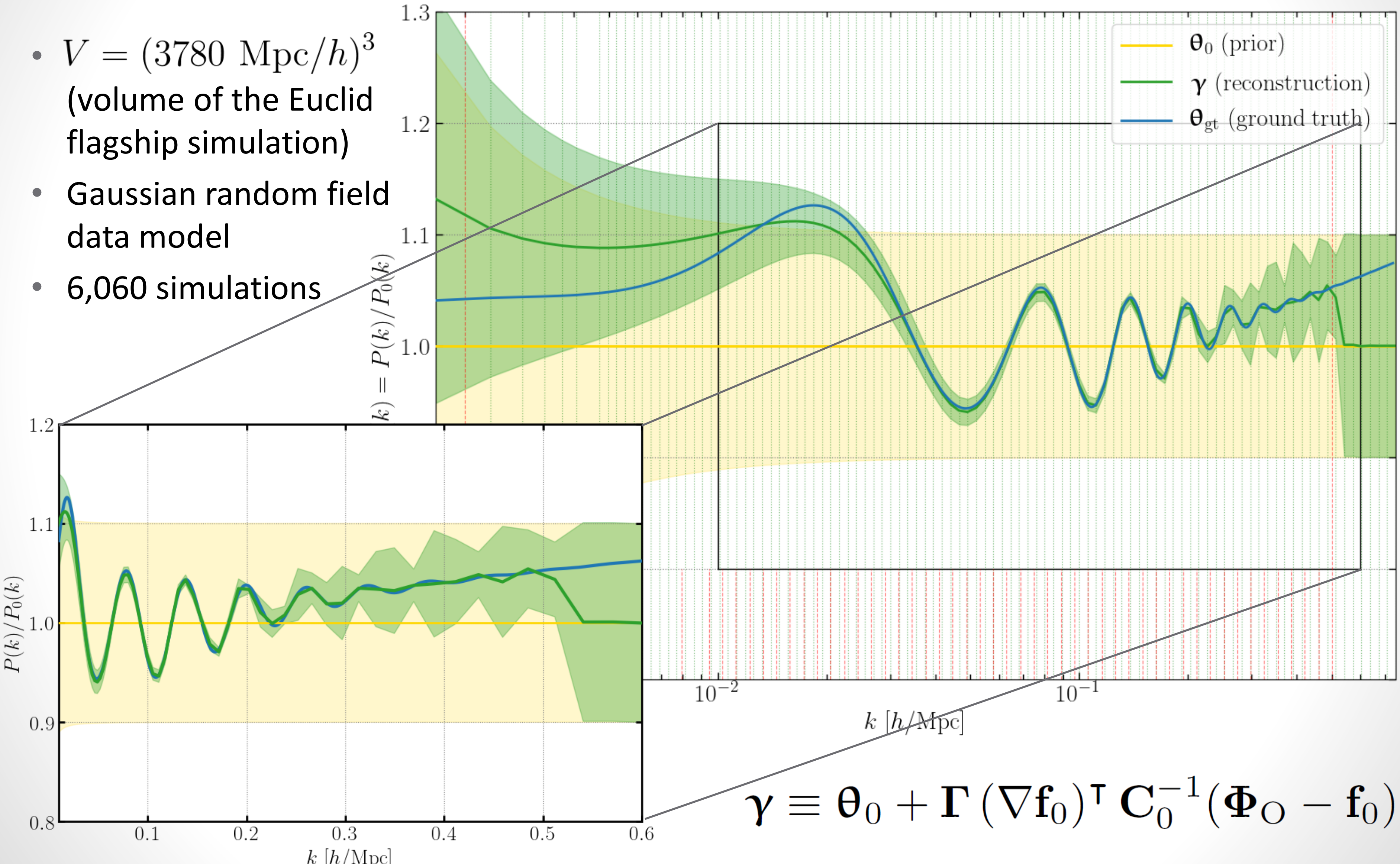

**100** parameters are simultaneously inferred from a black-box data model

**1** (Gpc/$h$)$^3$ only! Much more potential for upcoming data…

FL, Enzi, Jasche & Heavens 2019, 1902.10149

# SELFI-1 Euclid forecast (cosmic variance limit)

- $V = (3780\ \mathrm{Mpc}/h)^3$ (volume of the Euclid flagship simulation)
- Gaussian random field data model
- 6,060 simulations

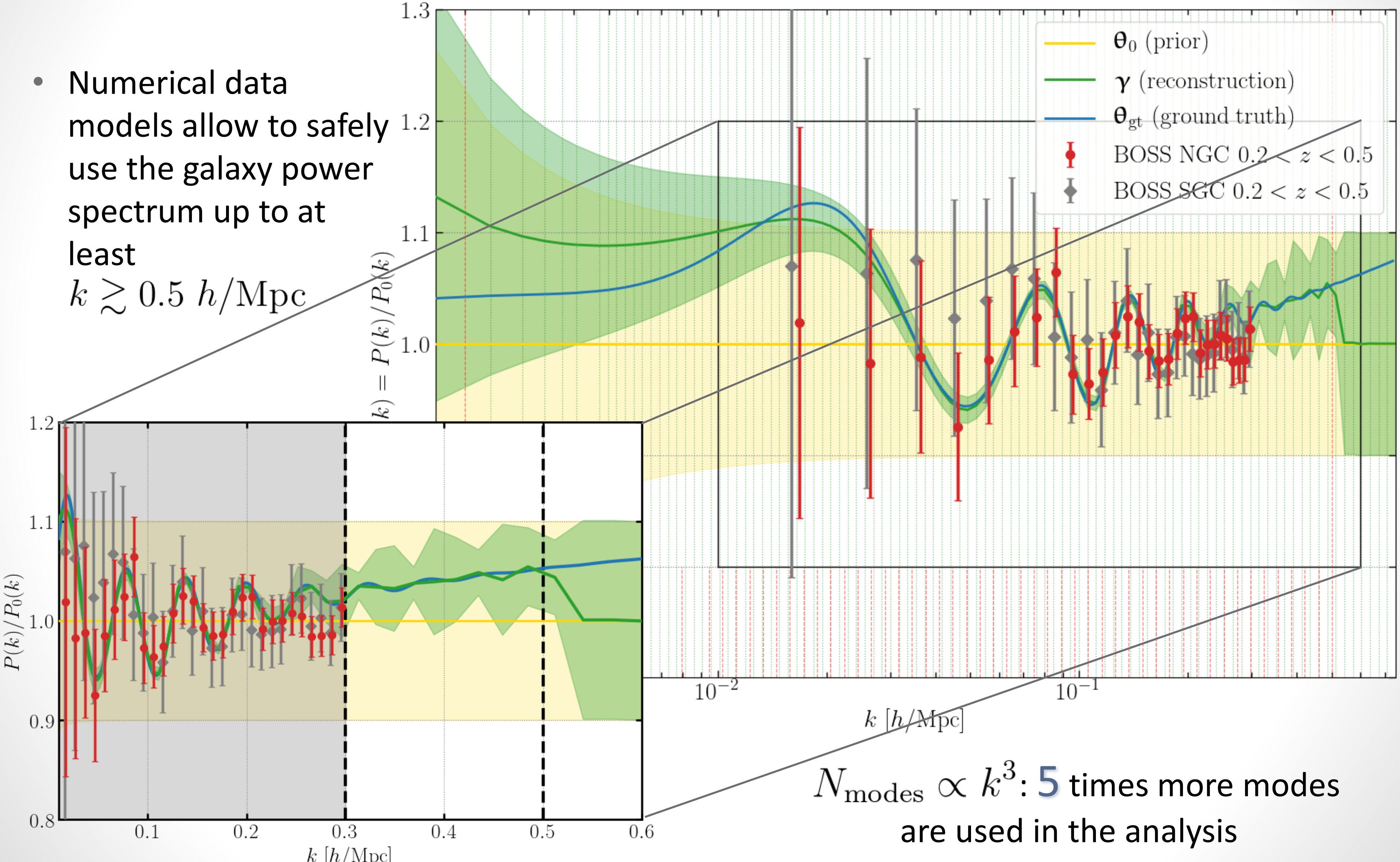



$$\boldsymbol{\gamma} \equiv \boldsymbol{\theta}_0 + \boldsymbol{\Gamma}\,(\nabla \mathbf{f}_0)^\intercal\,\mathbf{C}_0^{-1}(\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0)$$

# SELFI-1 Euclid versus BOSS

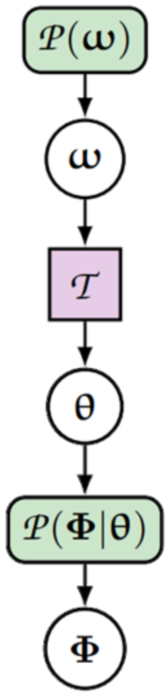
Data points from
Beutler *et al.* 2016, 1607.03149

- Numerical data models allow to safely use the galaxy power spectrum up to at least
  $k \gtrsim 0.5\ h/\mathrm{Mpc}$




$N_{\mathrm{modes}} \propto k^3$: 5 times more modes are used in the analysis

# From initial power spectrum to cosmology

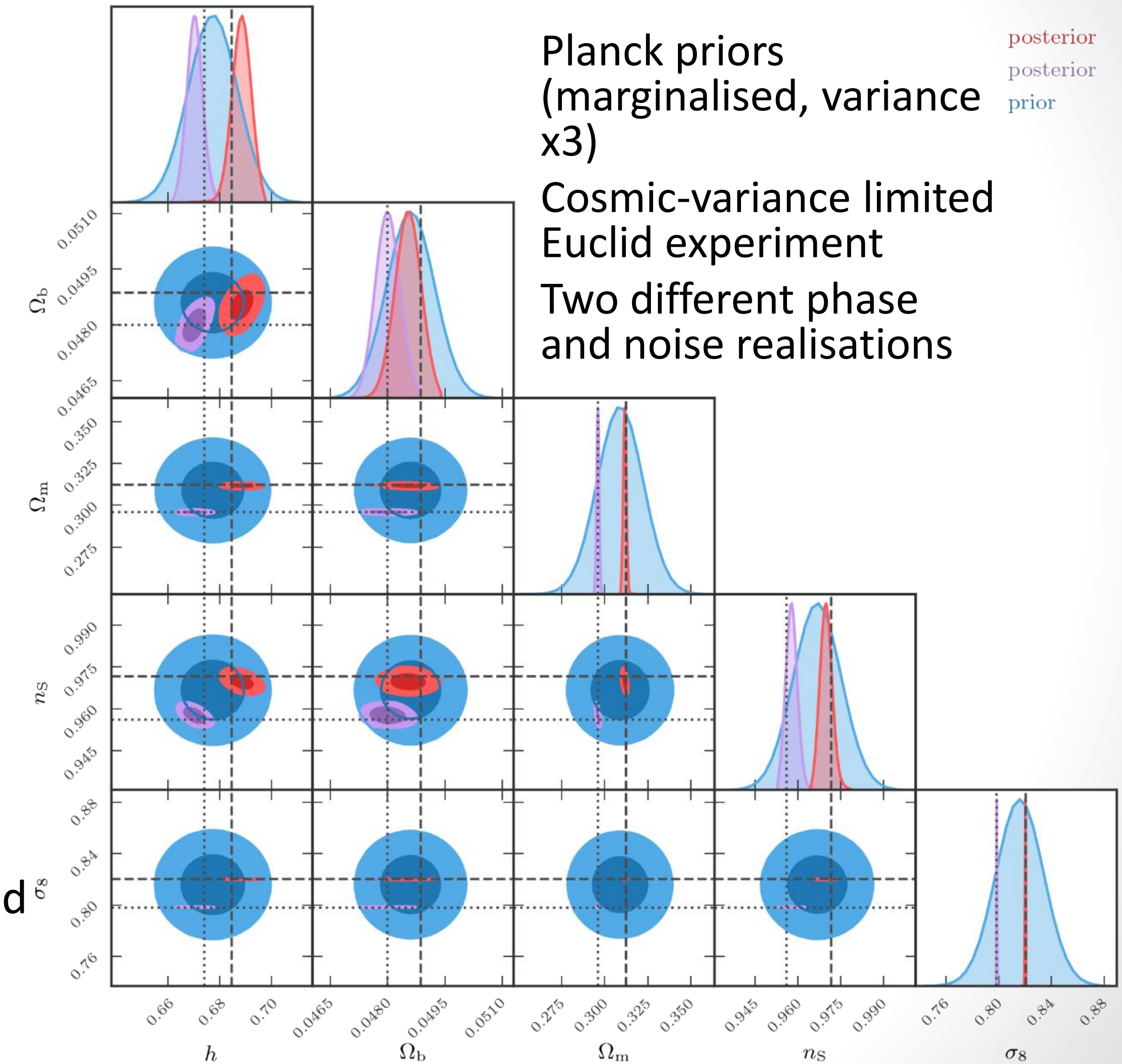



Cosmological parameters

Initial power spectrum

Taylor-expanded black-box

Arbitrary statistical summaries




Planck priors (marginalised, variance x3)

Cosmic-variance limited Euclid experiment

Two different phase and noise realisations

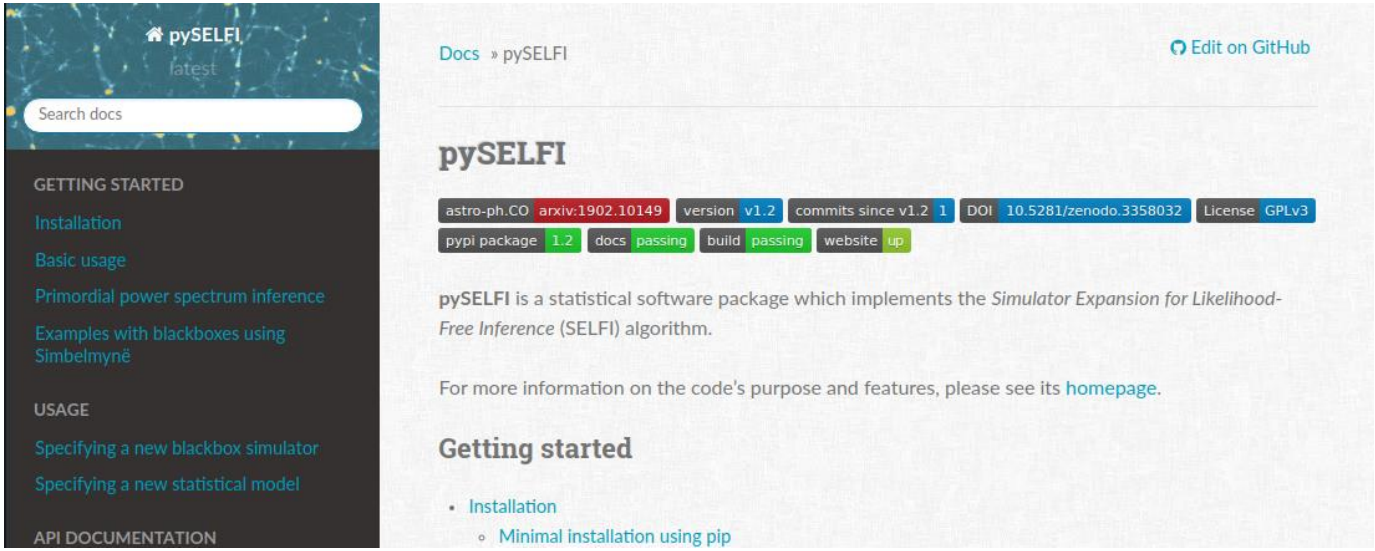
- Robust **inference of cosmological parameters** can be easily performed *a posteriori* once the linearized data model is learnt

FL, Enzi, Jasche & Heavens 2019, 1902.10149

# pySELFI is publicly available

- Code homepage: http://pyselfi.florent-leclercq.eu/
- Source on GitHub: https://github.com/florent-leclercq/pyselfi/
- Documentation on ReadtheDocs: https://pyselfi.readthedocs.io/en/latest/

(with templates to use your own black-box)



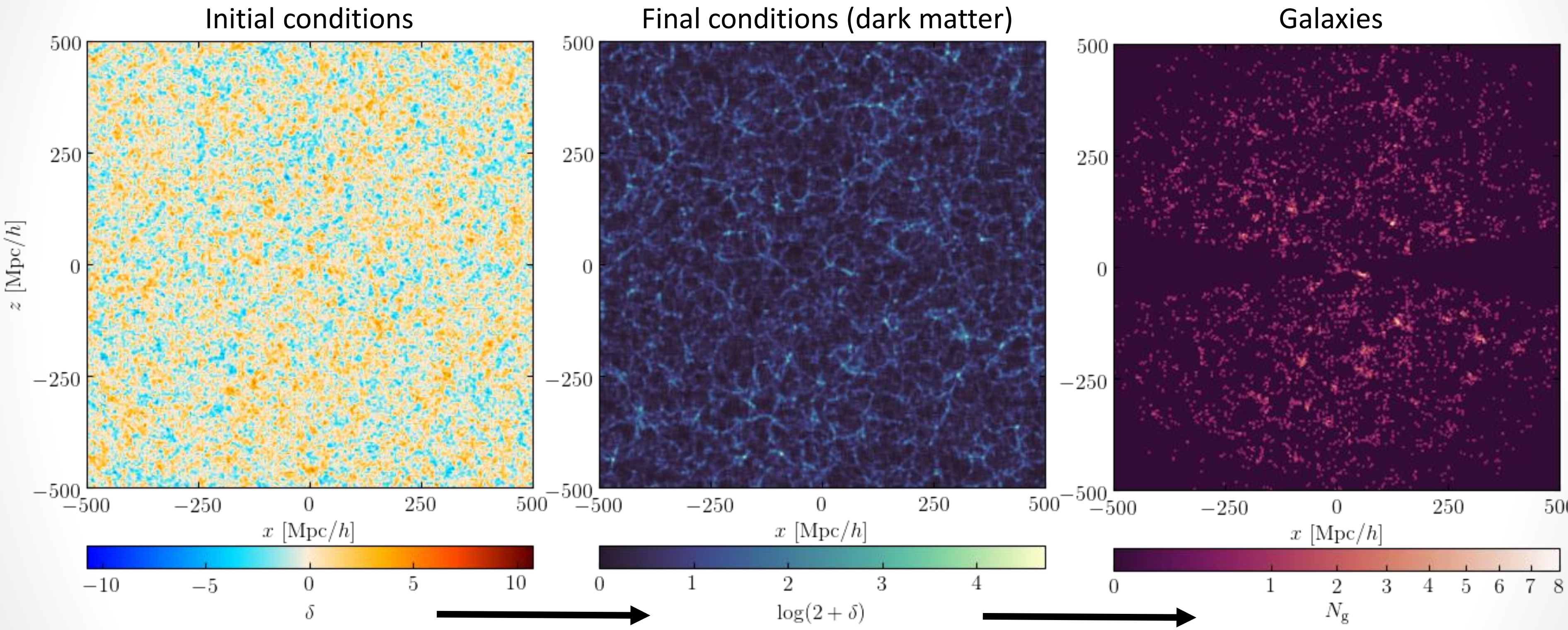

```
pip install pyselfi
```

# Numerical data models: Galaxy Clustering and beyond

# A black-box: Simbelmynë

Publicly available code:

https://bitbucket.org/florent-leclercq/simbelmyne/




Dark matter simulation with PM/tCOLA/sCOLA

Tassev, Zaldarriaga & Eisenstein 2013, 1301.0322

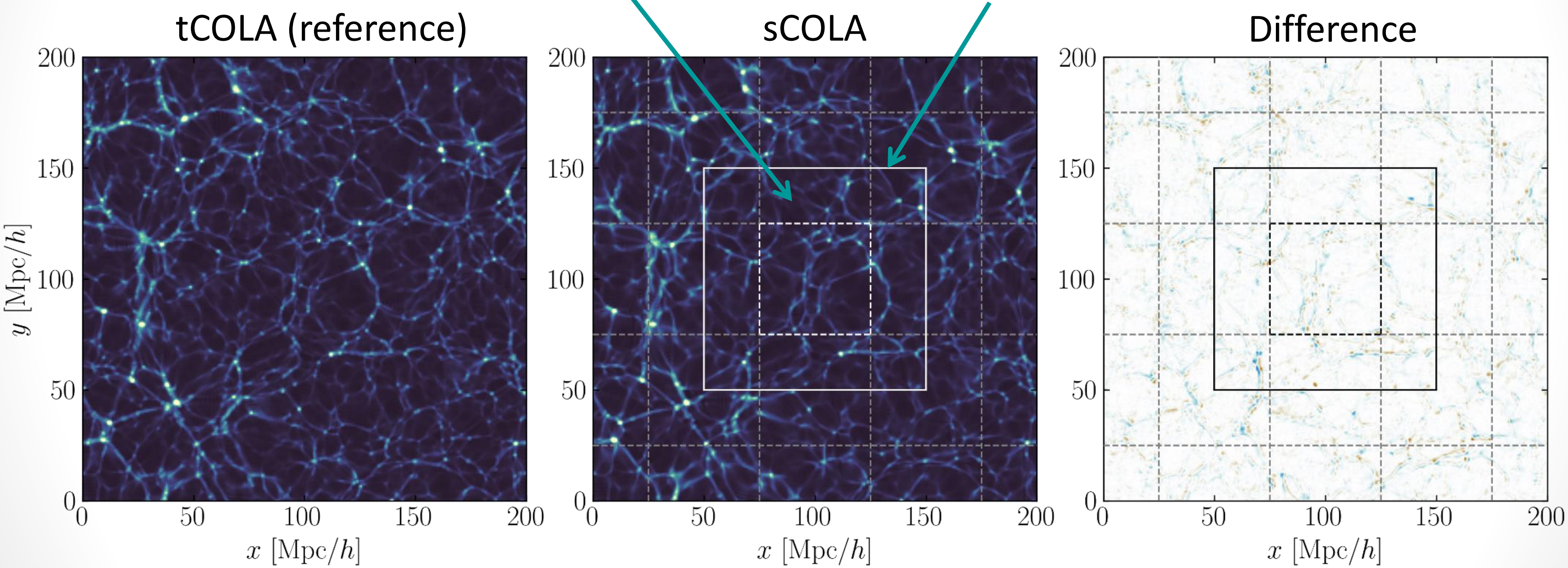
Tassev, Eisenstein, Wandelt & Zaldarriaga 2015, 1502.07751

Survey simulation: Redshift-space distortions, galaxy bias, selection effects, survey geometry, instrumental noise

# Perfectly parallel cosmological simulations using sCOLA

Can we decouple sub-volumes by using the large-scale analytical solution?

1. A buffer region around each tile
2. Appropriate Dirichlet boundary conditions for the potential



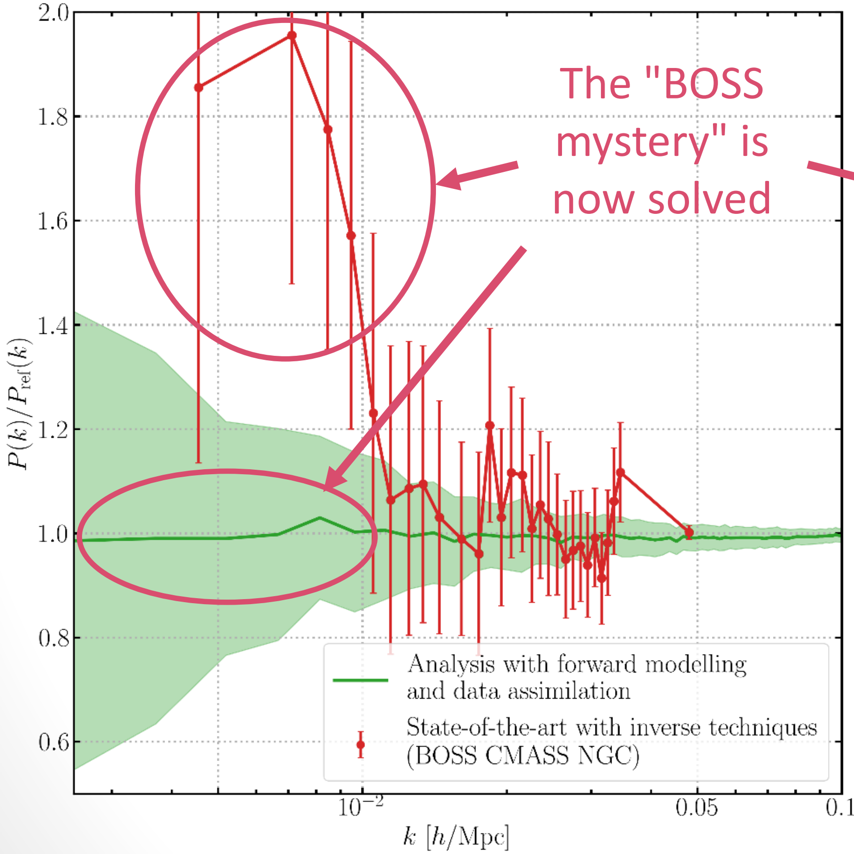

Our new perfectly parallel algorithm unlocks profoundly new possibilities of computing larger and higher-resolution cosmological simulations, taking advantage of a variety of hardware architectures (e.g. Cosmology@Home).

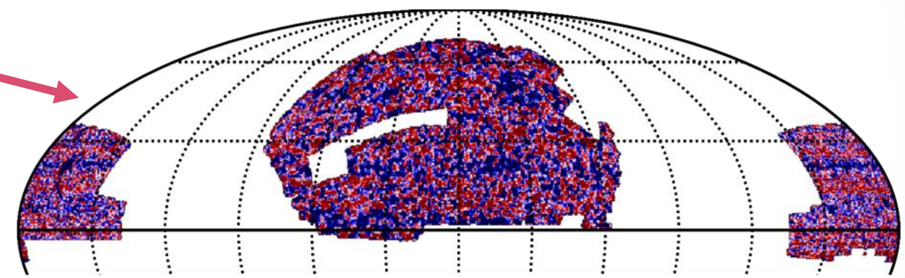
FL, Faure, Lavaux, Wandelt, Jaffe, Heavens, Percival & Noûs, 2003.04925

# Forward modelling of known and unknown systematics

- **Effects currently accounted for in our numerical data models:**

  Non-local galaxy biases, redshift-space distortions, light-cone effects, survey geometry, selection effects, foregrounds (including stars, dust, atmosphere, and unknown foregrounds)



Residual unexplained foreground in BOSS data

Lavaux, Jasche & FL 2019, 1909.06396

# Summary and concluding thoughts

- Goal: developing and using algorithms for targeted science questions, allowing the use of simulators including all relevant physical and observational effects.
- Bayesian analyses of galaxy surveys with fully non-linear numerical black-box models is not an impossible task!
- BOLFI allows inference within specific cosmological models with a very limited simulation budget. The optimal acquisition function shall be used.
- SELFI allows inference of the initial power spectrum and cosmological parameters.
- Our numerical data models are being refined and optimised to prepare for upcoming data, including Euclid.