Bayesian analyses of galaxy surveys

Florent Leclercq
www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

Wolfgang Enzi, Alan Heavens, Jens Jasche,
Guilhem Lavaux, Will Percival, Benjamin Wandelt

and the Aquila Consortium
www.aquila-consortium.org

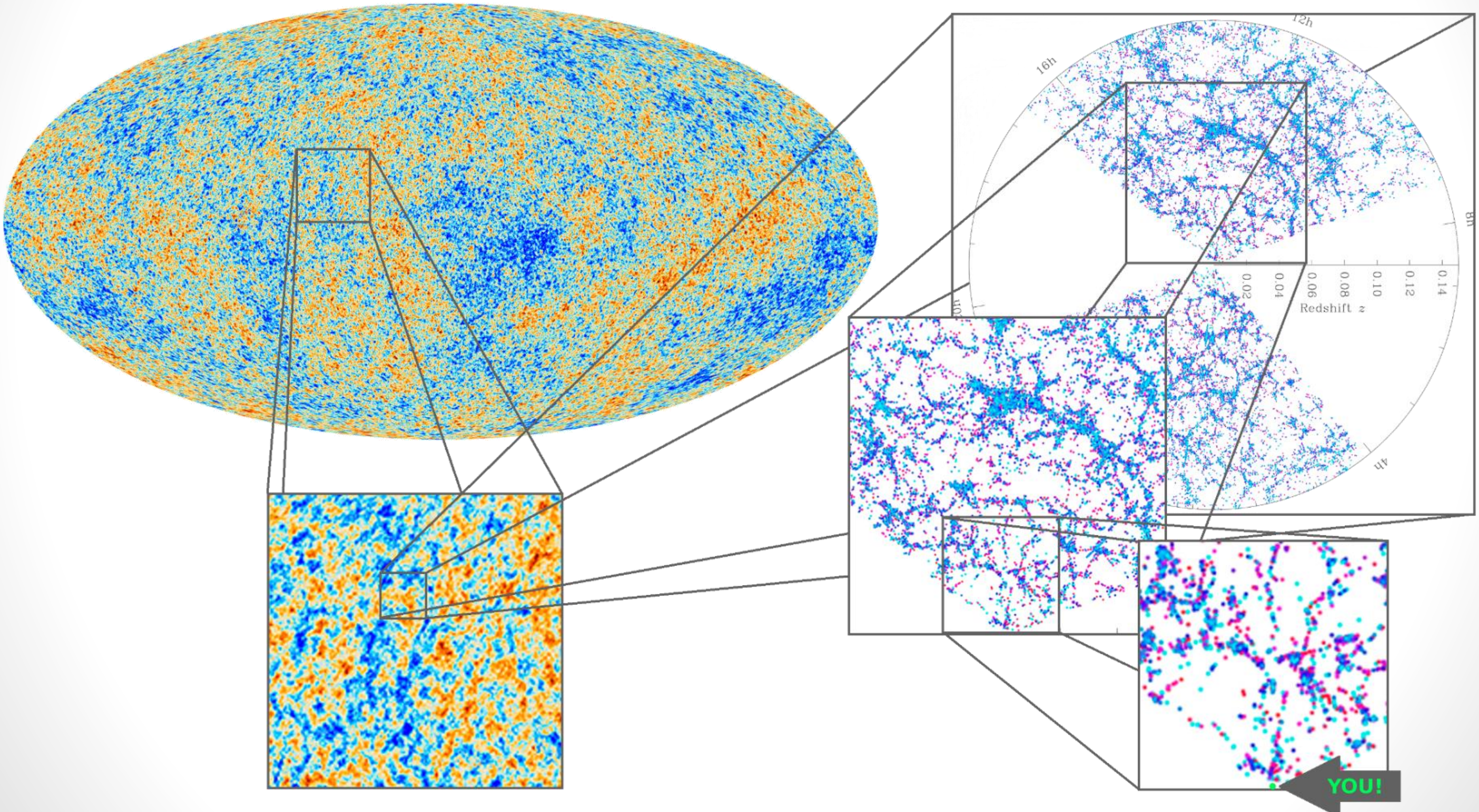
May 2nd, 2019

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 large-scale structure

The LSS is a vast source of knowledge:

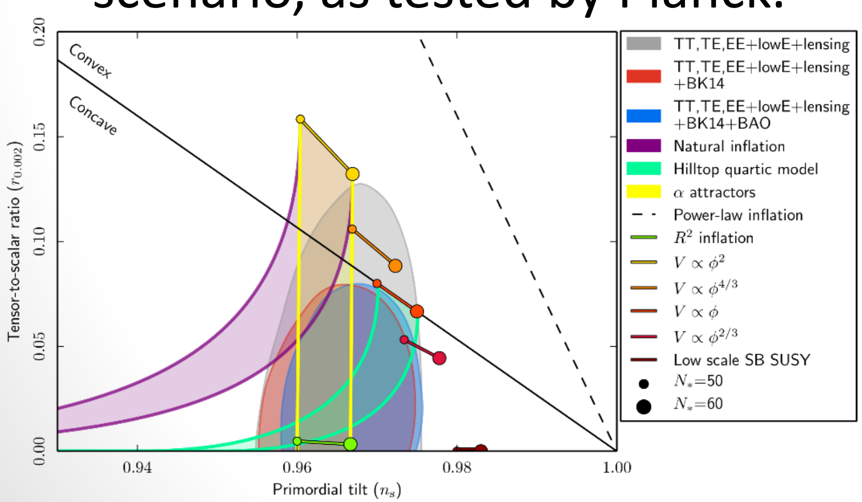
- Cosmology:
  - ΛCDM : cosmological parameters and tests against alternatives,
  - Physical nature of the dark components,
  - Neutrinos : number and masses,
  - Geometry of the Universe,
  - Tests of General Relativity,
  - Initial conditions and link to high energy physics
- Astrophysics: galaxy formation and evolution as a function of their environment
  - Galaxy properties (colours, chemical composition, shapes),
  - Intrinsic alignments, intrinsic size-magnitude correlations

We have theoretical and computer models…

- **Initial conditions:**
  a Gaussian random field

$$\mathcal{P}(\delta^{\mathrm{i}}|S) = \frac{1}{\sqrt{|2\pi S|}} \exp\left(-\frac{1}{2}\sum_{x,x'} \delta^{\mathrm{i}}_x S^{-1}_{xx'} \delta^{\mathrm{i}}_{x'}\right)$$

Everything seems consistent with the simplest inflationary scenario, as tested by Planck.



Planck 2018 X, arXiv:1807.06211

- **Structure formation:**
  numerical solution of the Vlasov-Poisson system for dark matter dynamics

$$\frac{\partial f}{\partial \tau} + \frac{\mathbf{p}}{ma} \cdot \nabla f - ma\nabla\Phi \cdot \frac{\partial f}{\partial \mathbf{p}} = 0$$

$$\Delta\Phi = 4\pi \mathrm{G} a^2 \bar{\rho}\delta$$

Y. Dubois & S. Colombi (IAP)

… how do we test these models against survey data?

J. Cham – PhD comics

Redshift range	Volume (Gpc^3)	k_{max} $(\mathrm{Mpc}/h)^{-1}$	N_{modes}
0-1	50	0.15	10^7
1-2	140	0.5	5×10^8
2-3	160	1.3	10^{10}

M. Zaldarriaga

- Precise tests require many modes.
- In 3D galaxy surveys, the number of modes usable scales as k_{max}^3.

- The challenge: non-linear evolution at small scales and late times.
- The strategy:
  - Pushing down the smallest scale usable for cosmological analysis
  - Using a numerical model linking initial and final conditions

In other words: go beyond the linear and static analysis of the LSS.

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?"

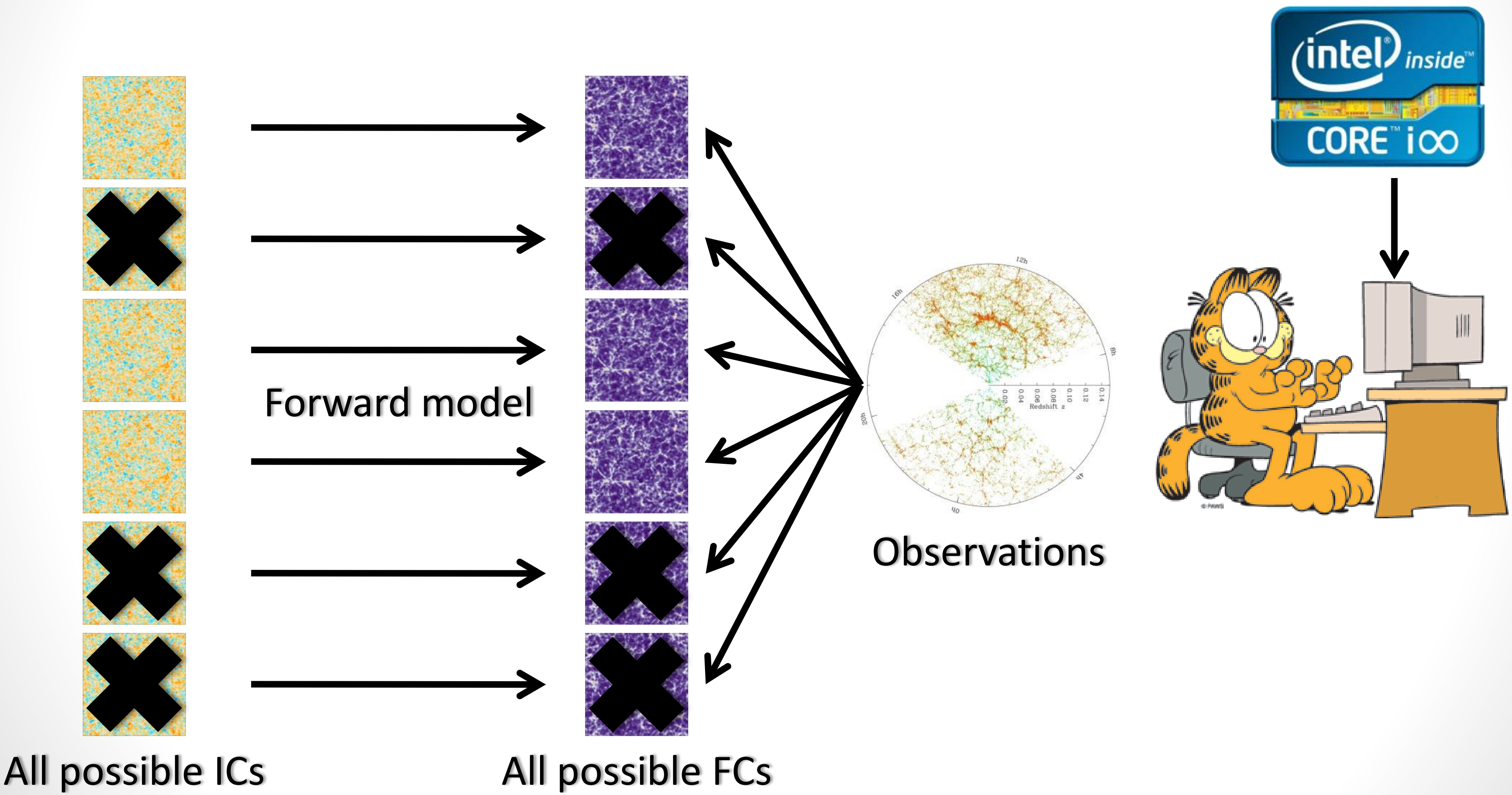
"What is the probability distribution of possible formation histories (signals) compatible with the observations?"

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

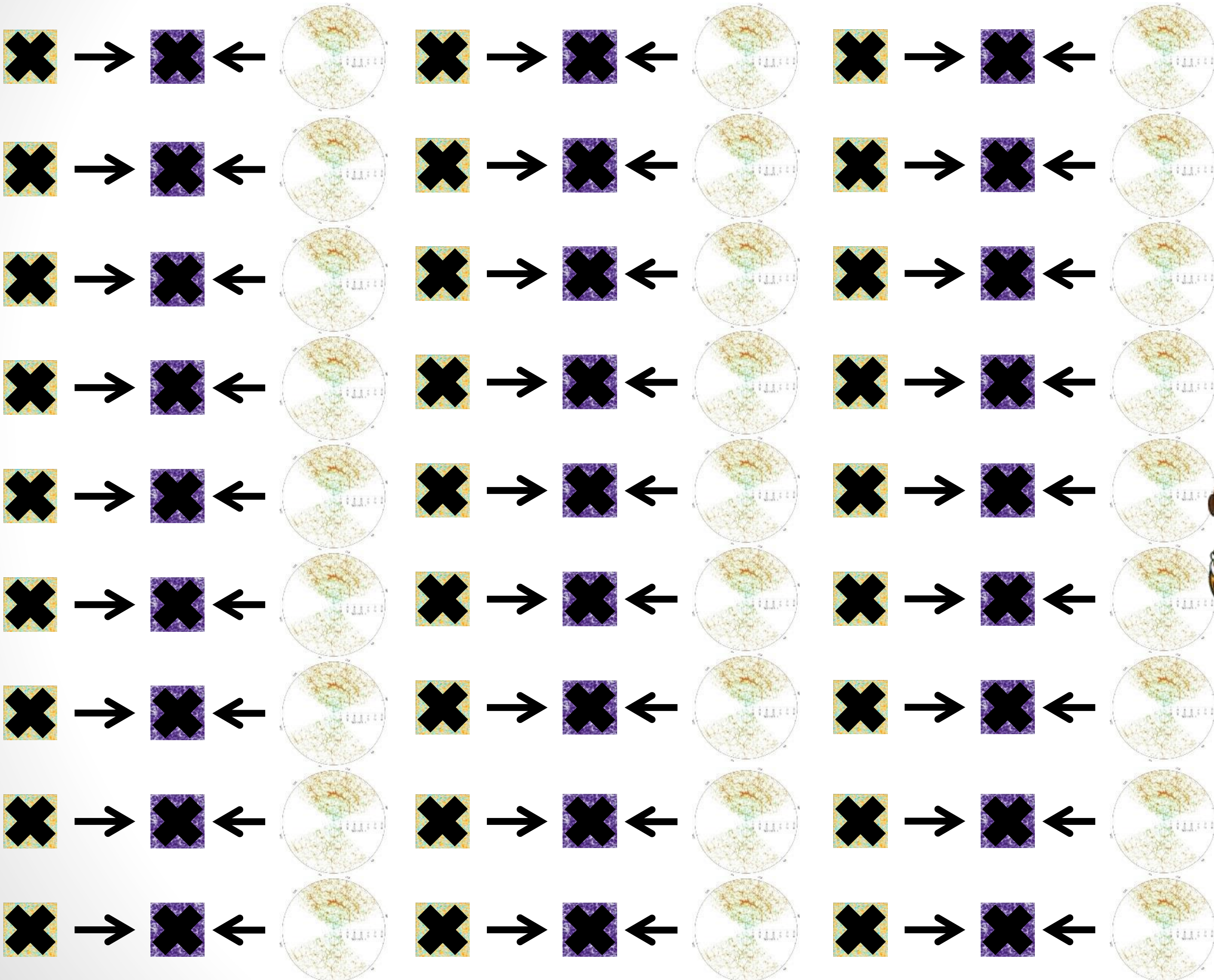
- Cox-Jaynes theorem: Any system to manipulate "*plausibilities*", consistent with Cox's desiderata, is isomorphic to (Bayesian) probability theory

So how do we do that?

Bayesian forward modelling: the ideal scenario



Bayesian forward modelling: the challenge



The (true) likelihood
lives in

$d \approx 10^7$!

Likelihood-based solution: BORG

Bayesian Origin Reconstruction from Galaxies

Likelihood-based solution:
Exact statistical analysis
Approximate data model

Data assimilation

?

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

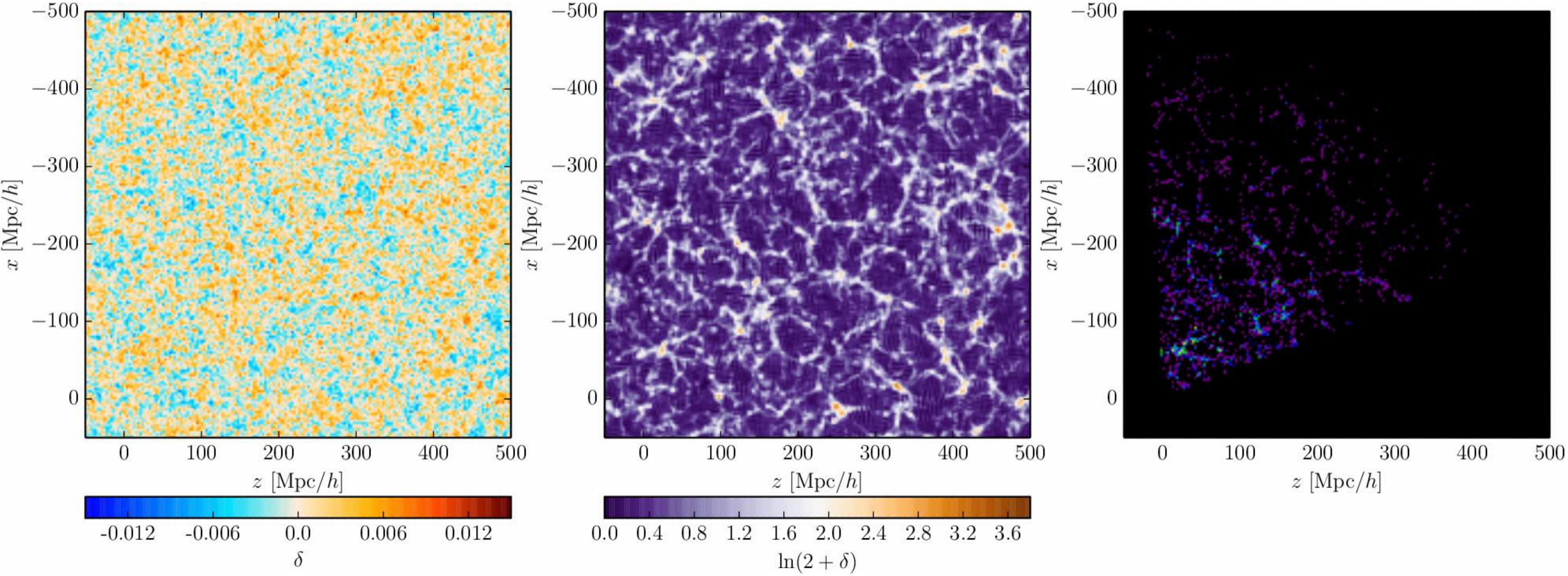
$$a(\mathbf{x}',\mathbf{x}) = \mathrm{e}^{-(H'-H)} = 1$$

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 at work: chrono-cosmography from SDSS data



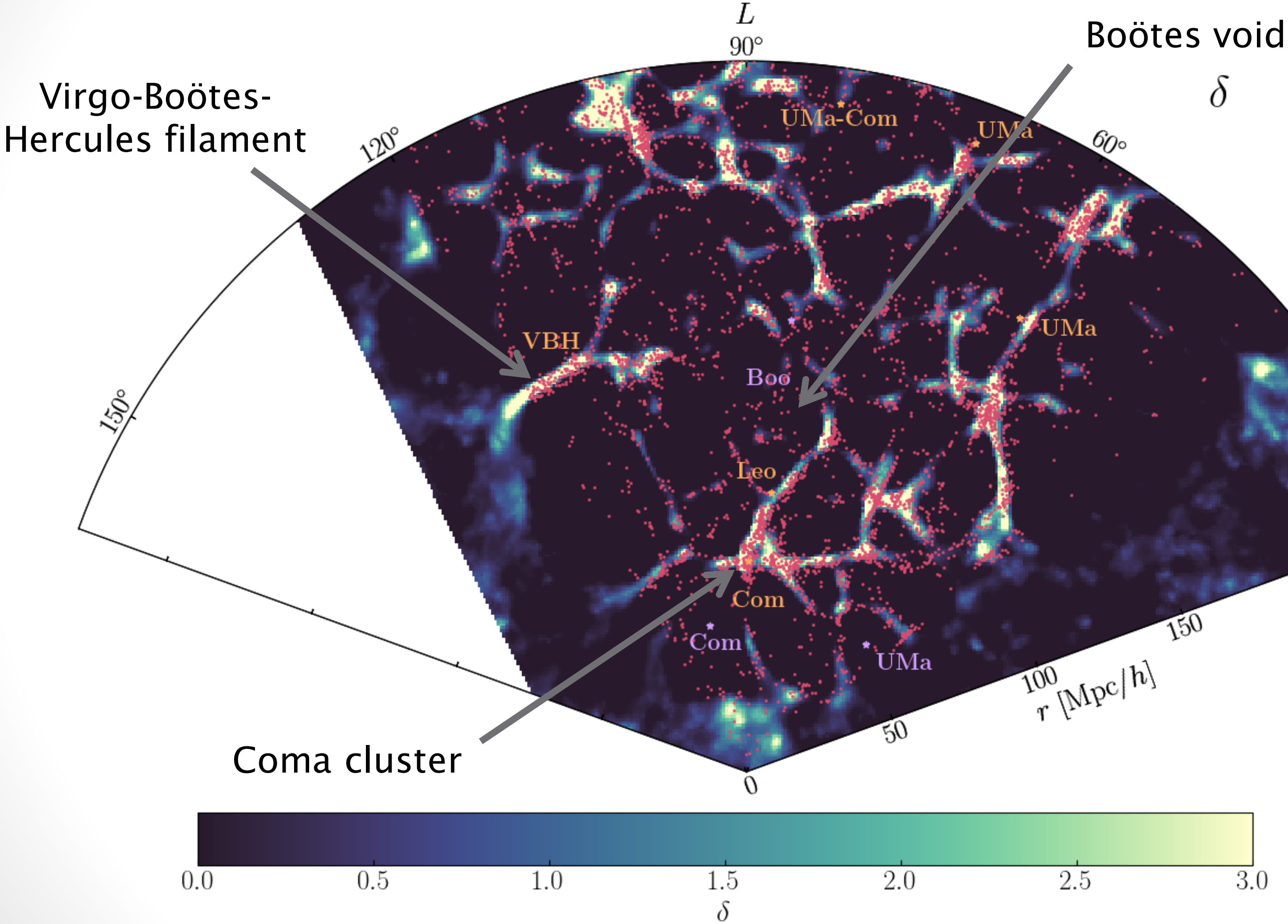
Initial conditions — Final conditions — Observations

334,074 galaxies, ≈ 17 million parameters, 3 TB of primary data products,
12,000 samples, ≈ 250,000 data model evaluations, 10 months on 32 cores

All data products are publicly available:

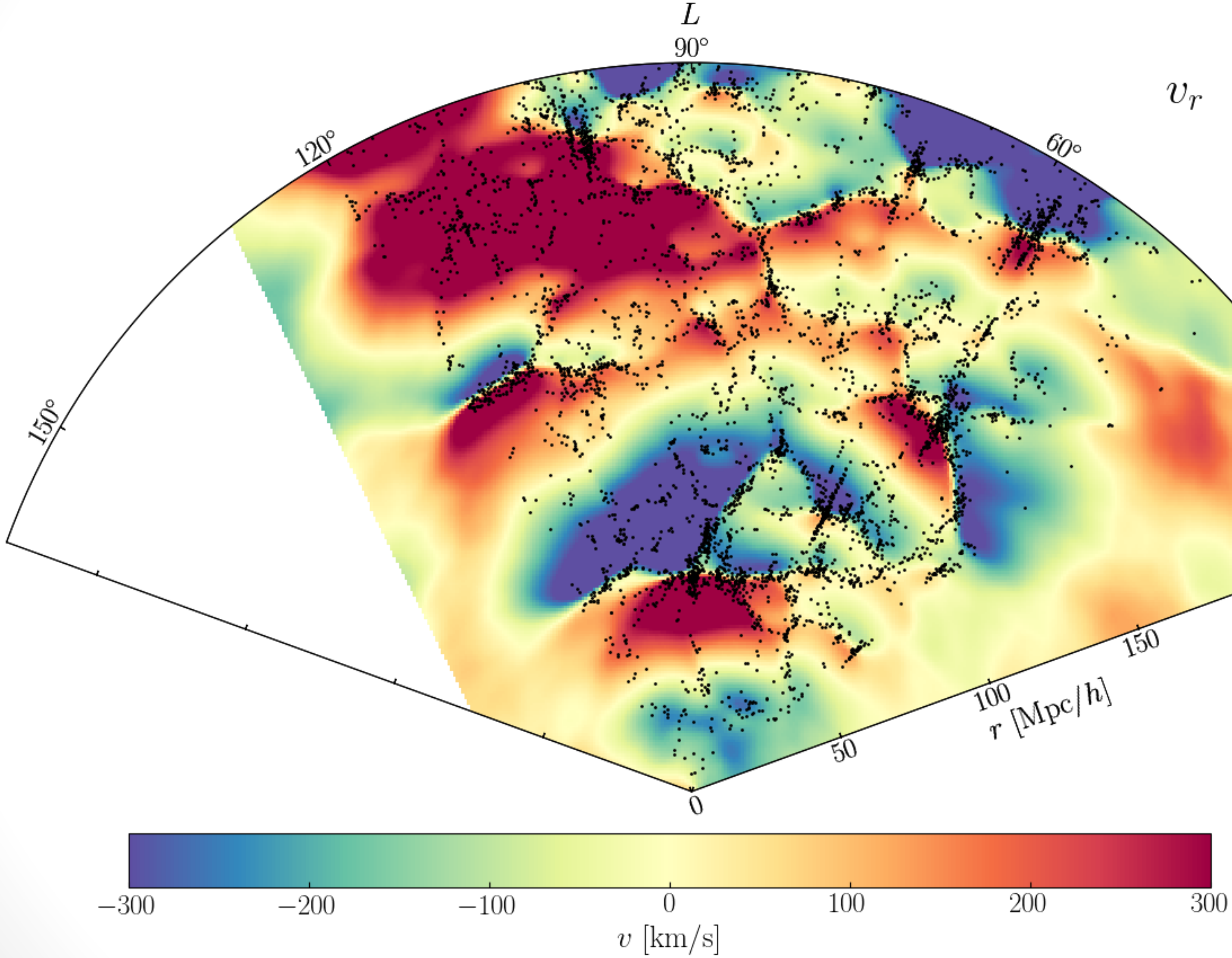
Jasche, FL & Wandelt 2015, arXiv:1409.6308 — https://github.com/florent-leclercq/borg_sdss_data_release, doi: 10.5281/zenodo.1455729

Cosmography in the supergalactic plane



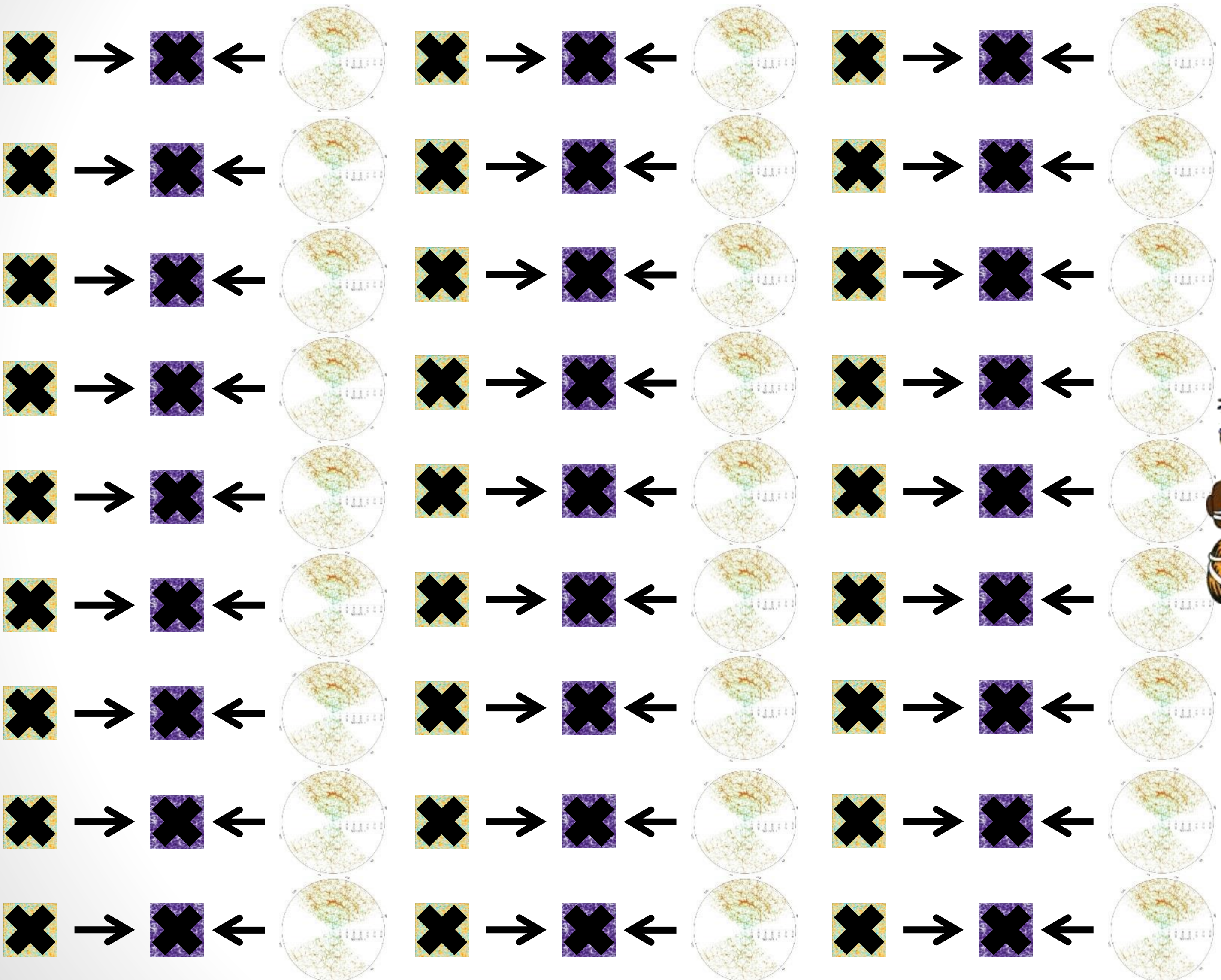
FL, Jasche, Lavaux, Wandelt & Percival 2017, arXiv:1601.00093

Cosmography in the supergalactic plane



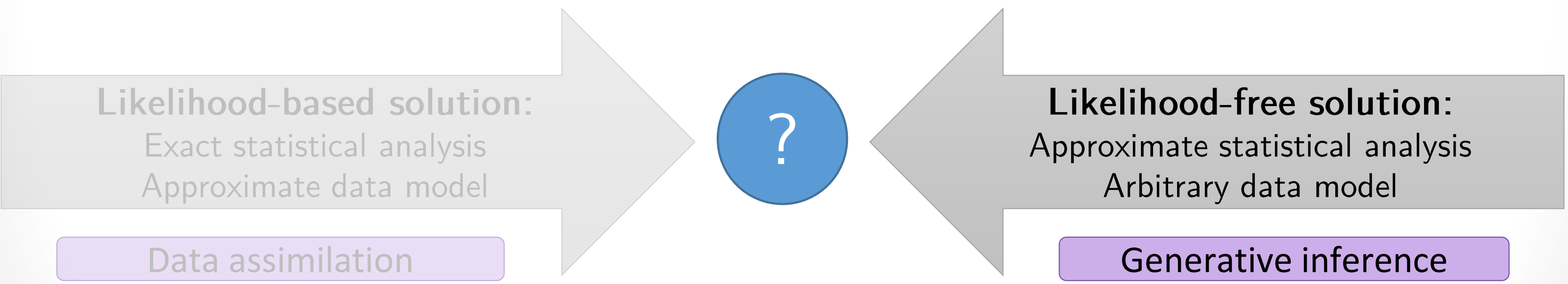
FL, Jasche, Lavaux, Wandelt & Percival 2017, arXiv:1601.00093

Let's go back to the challenge…



Likelihood-free solution: BOLFI & SELFI

Bayesian Optimisation for Likelihood-Free Inference
Simulator Expansion for Likelihood-Free Inference



Likelihood-free inference: 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, arXiv:1605.06376
  Alsing, Feeney & Wandelt 2018, arXiv:1801.01497
- Bayesian Optimisation for Likelihood-Free Inference (BOLFI)

  Gutmann & Corander 2016, arXiv:1501.03291
  FL 2018, arXiv:1805.07152

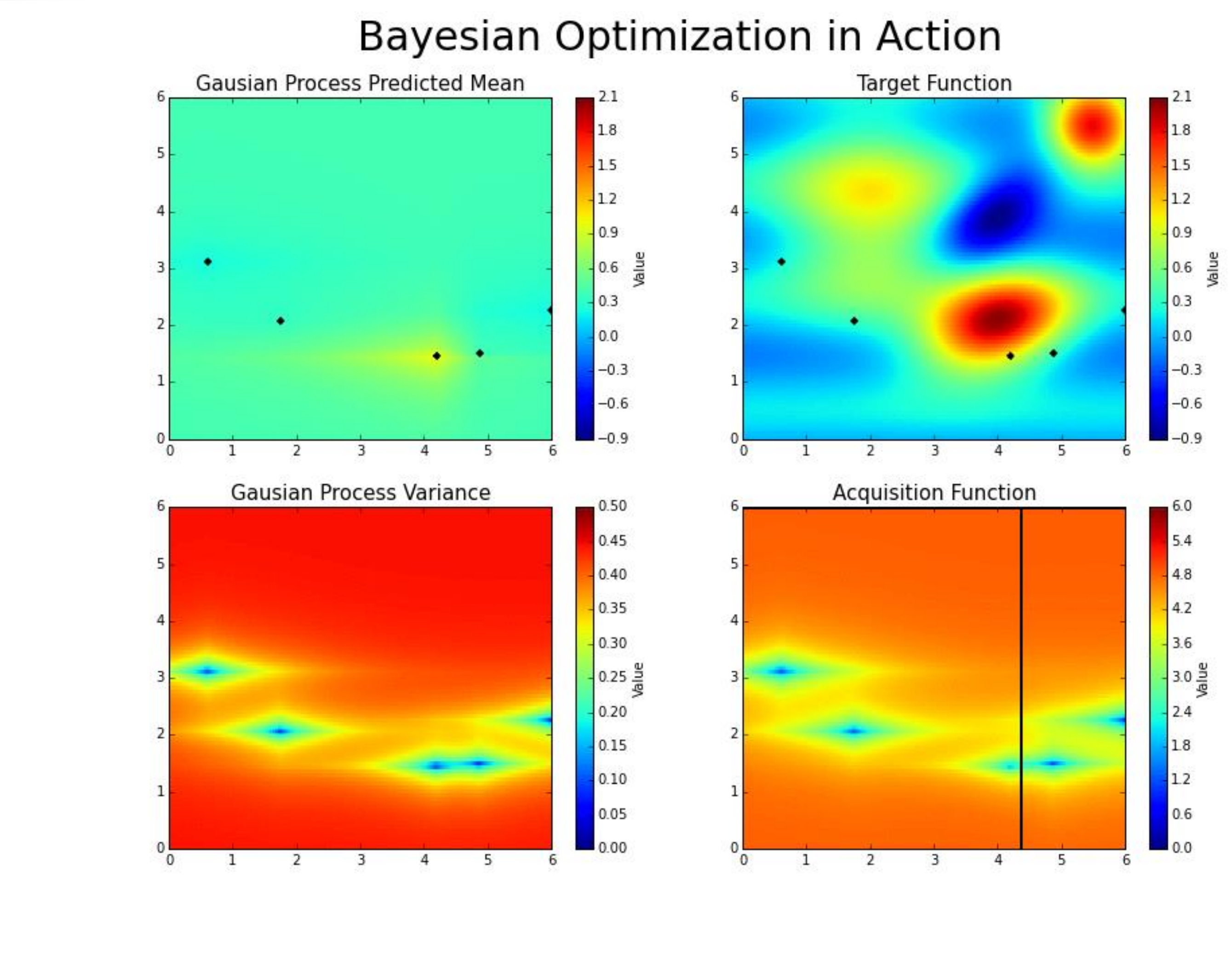
The "number of parameters" route:

- 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, arXiv:1902.10149

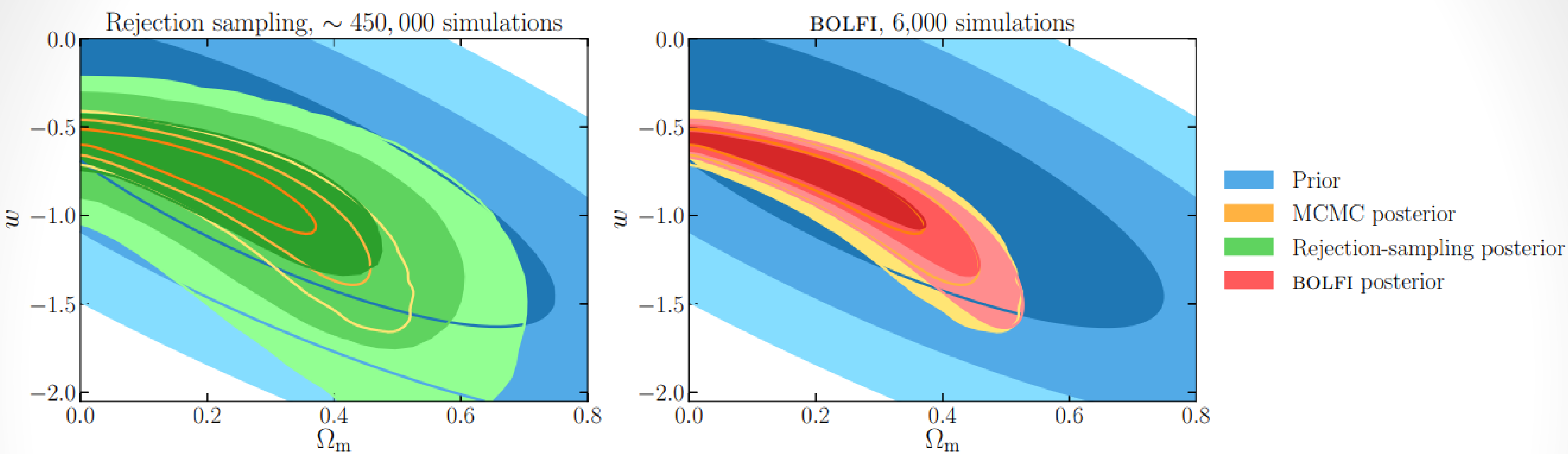
I thought of the name **after** developing the method!

BOLFI: Data acquisition



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

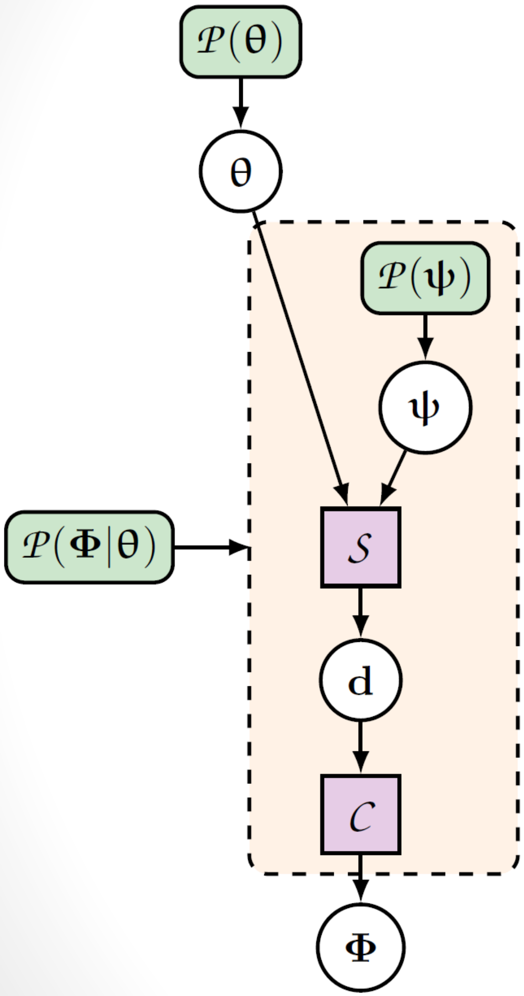
BOLFI: Re-analysis of the JLA supernova sample



- 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

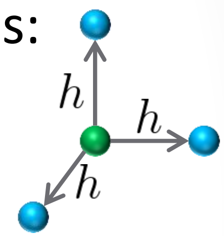
FL 2018, arXiv:1805.07152

SELFI: Method



- Gaussian prior + Gaussian effective likelihood
- Linearisation of the black-box around an expansion point + finite differences:

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



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

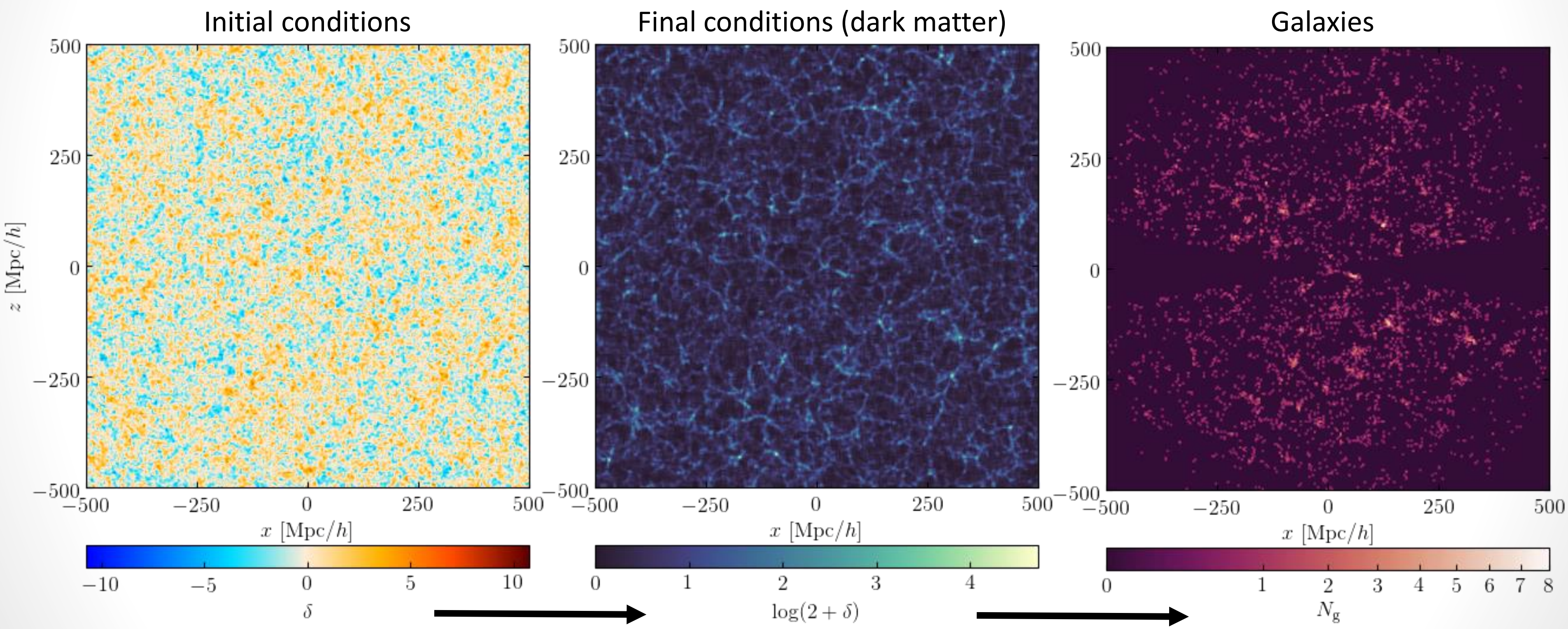
FL, Enzi, Jasche & Heavens 2019, arXiv:1902.10149

A black-box: Simbelmynë

I'm happy to explain the name later today…

Publicly available code:

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

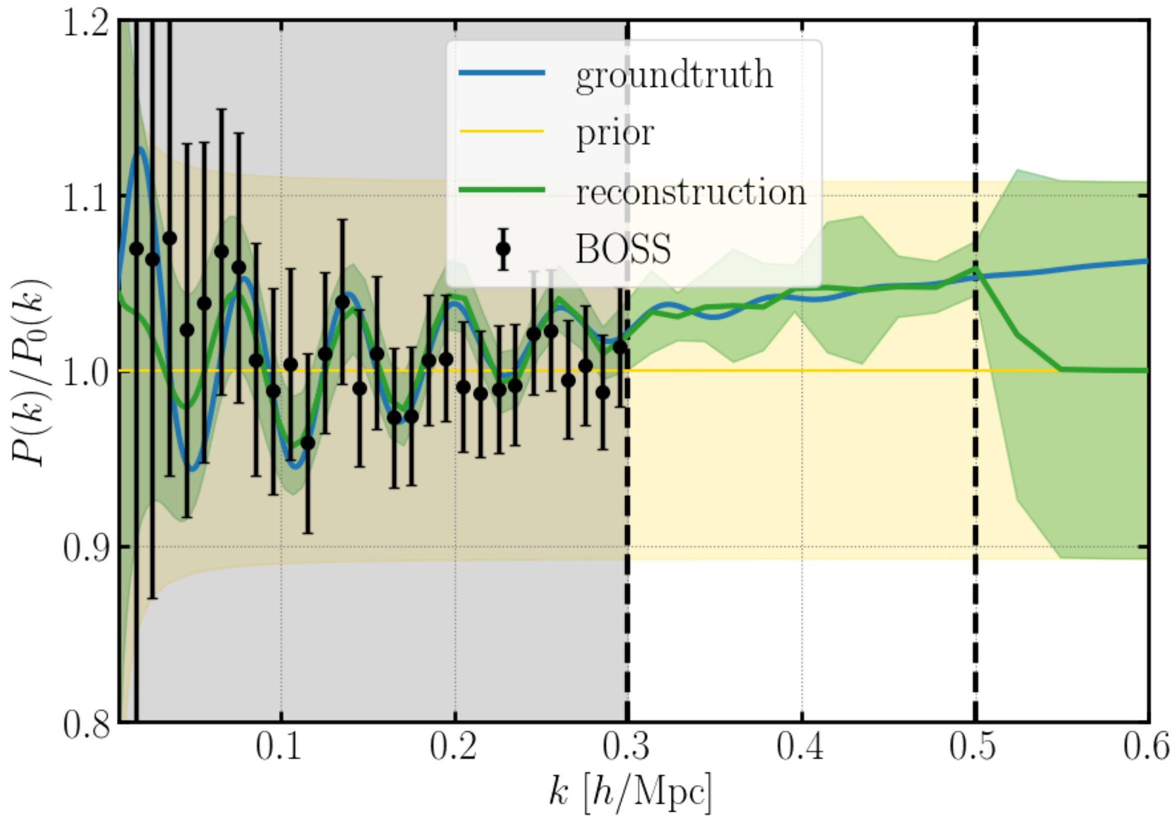


Dark matter simulation with COLA

Tassev, Zaldarriaga & Eisenstein 2013, arXiv:1301.0322

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

SELFI + Simbelmynë: Proof-of-concept



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

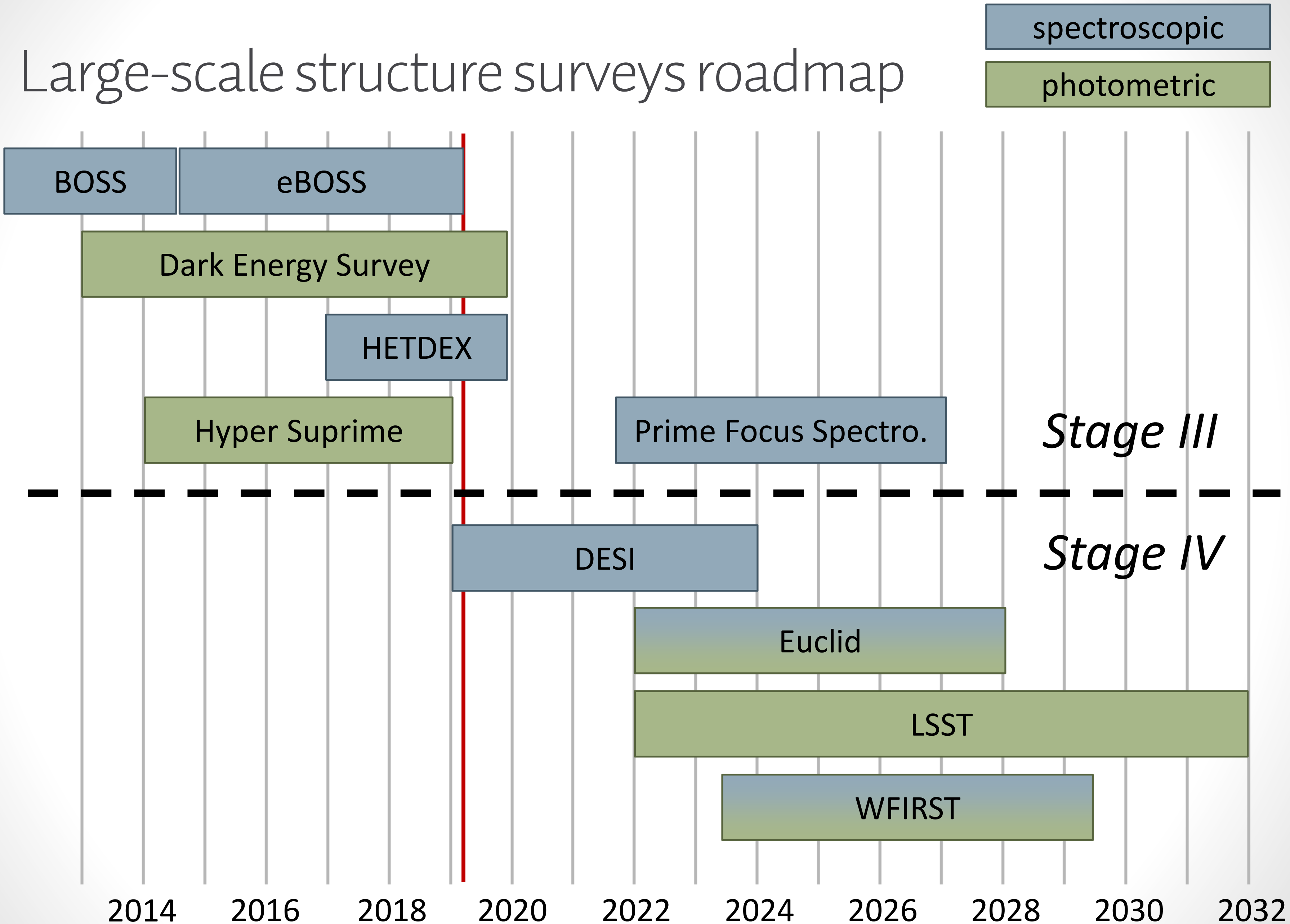
$N_\mathrm{modes} \propto k^3$: **5** times more modes are used in the analysis

FL, Enzi, Jasche & Heavens 2019, arXiv:1902.10149

The Future: Opportunities & Challenges

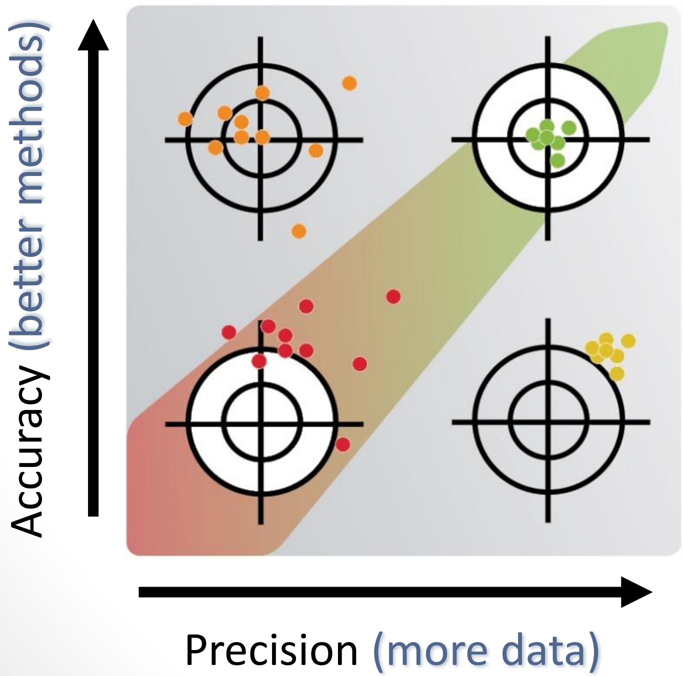
DESI, Euclid, LSST, WFIRST, and more…

Large-scale structure surveys roadmap



Data-intensive scientific discovery from galaxy surveys

- Next-generation surveys will be dominated by systematics
- 80% of the total signal will come from non-linear structures

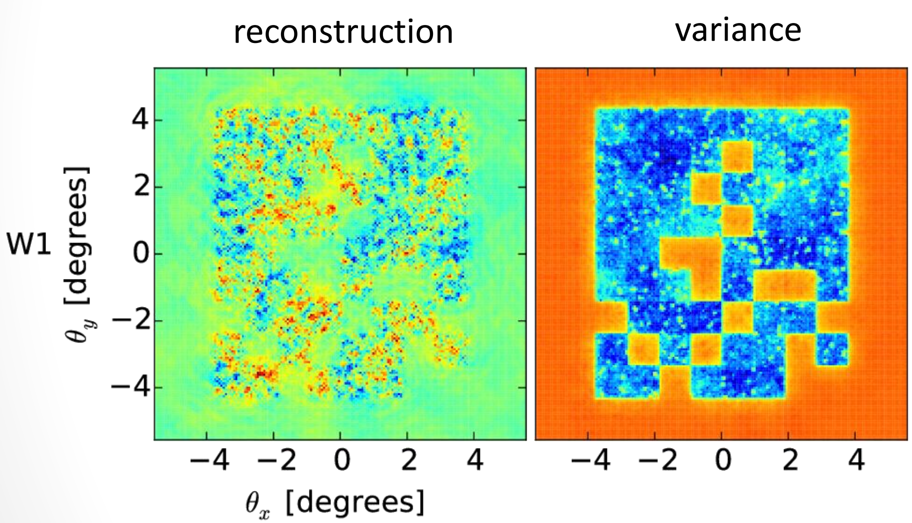


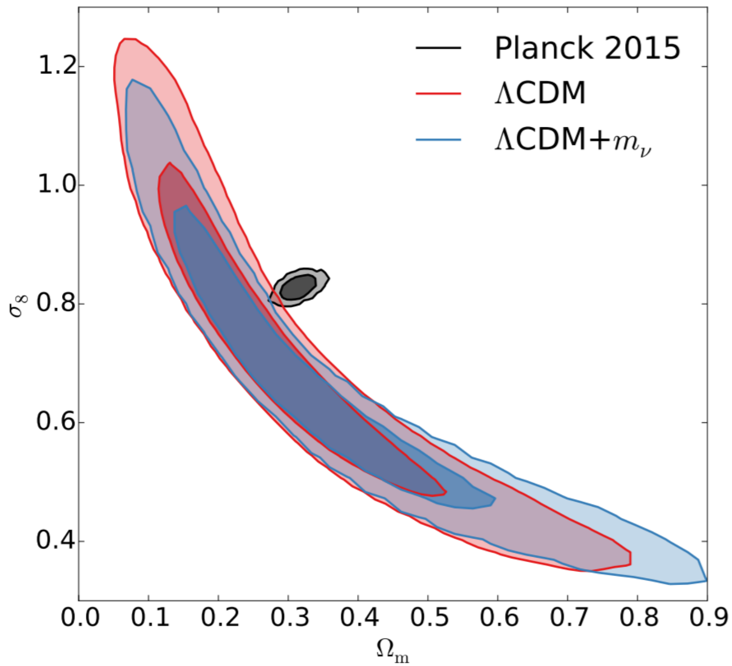
- Can data analysts keep pace?

The Imperial weak lensing inference framework

with George Kyriacou, Arrykrishna Mootoovaloo, Alan Heavens & Andrew Jaffe

Joint inference of cosmic shear maps and power spectra/cosmology from CFHTLenS



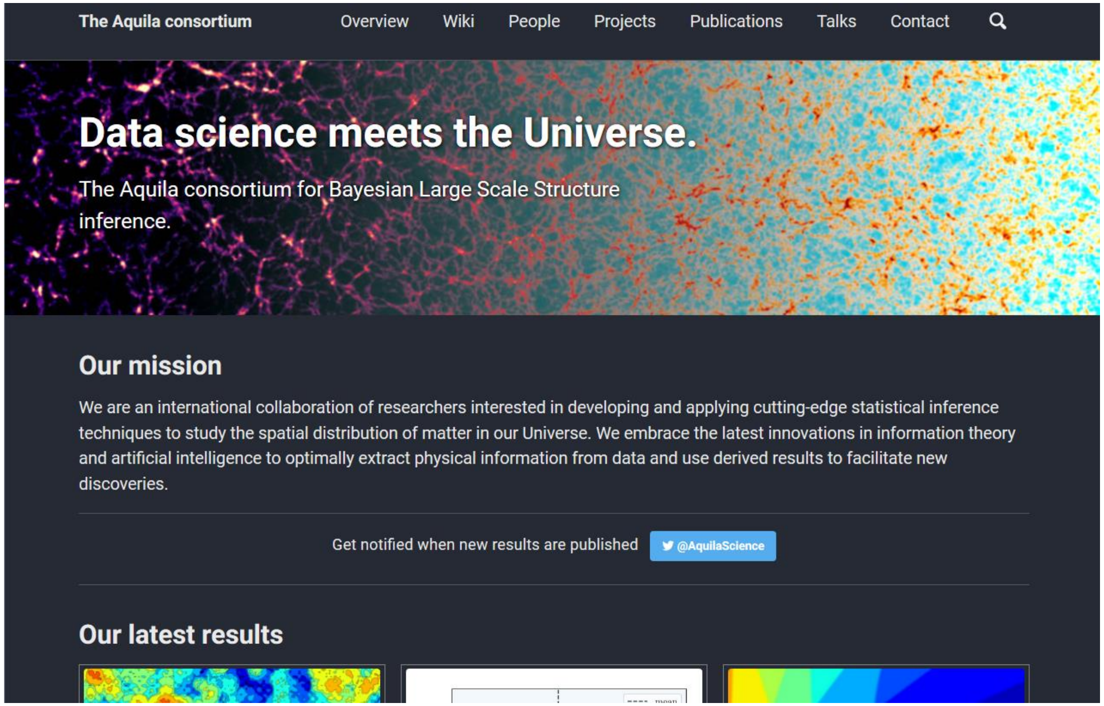


$\sum m_\nu < 4.6$ eV from lensing data alone

Alsing, Heavens & Jaffe 2016, arXiv:1607.00008

The Aquila Consortium

- 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

Concluding thoughts

Likelihood-based solution:
Exact statistical analysis
Approximate data model

Data assimilation

?

Likelihood-free solution:
Approximate statistical analysis
Arbitrary data model

Generative inference

- Bayesian analyses of galaxy surveys with fully non-linear numerical models is not an impossible task!
- A likelihood-based solution (BORG): general purpose reconstruction of dark matter from galaxy clustering, providing new measurements and predictions
- A likelihood-free solution (BOLFI/SELFI): algorithms for targeted questions, allowing the use of simulators including all relevant physical and observational effects

Concluding thoughts

- The future: great science and challenges

Galaxy formation: bias model & likelihood
Large volume, photometric redshifts
Instruments modelling

DESI, Euclid, LSST, WFIRST

Bayesian large-scale structure analyses

Cosmological measurements:

- Cosmic expansion
- Power spectrum (and governing parameters)
- Gaussianity tests of the initial conditions
- Dark energy from the growth of structure

Predictive cosmology:

- Velocity field
- X-ray cluster emission
- Gravitational lensing
- CMB secondary effects
- Dark matter?