Bayesian analyses of galaxy surveys

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology
Imperial College London

Alan Heavens, Andrew Jaffe, George Kyriacou,
Arrykrishna Mootoovaloo, James Prideaux-Ghee (Imperial College),
Jens Jasche (U. Stockhom),
Guilhem Lavaux, Benjamin Wandelt (IAP),
Wolfgang Enzi (MPA), Will Percival (U. Waterloo)

and the Aquila Consortium
www.aquila-consortium.org

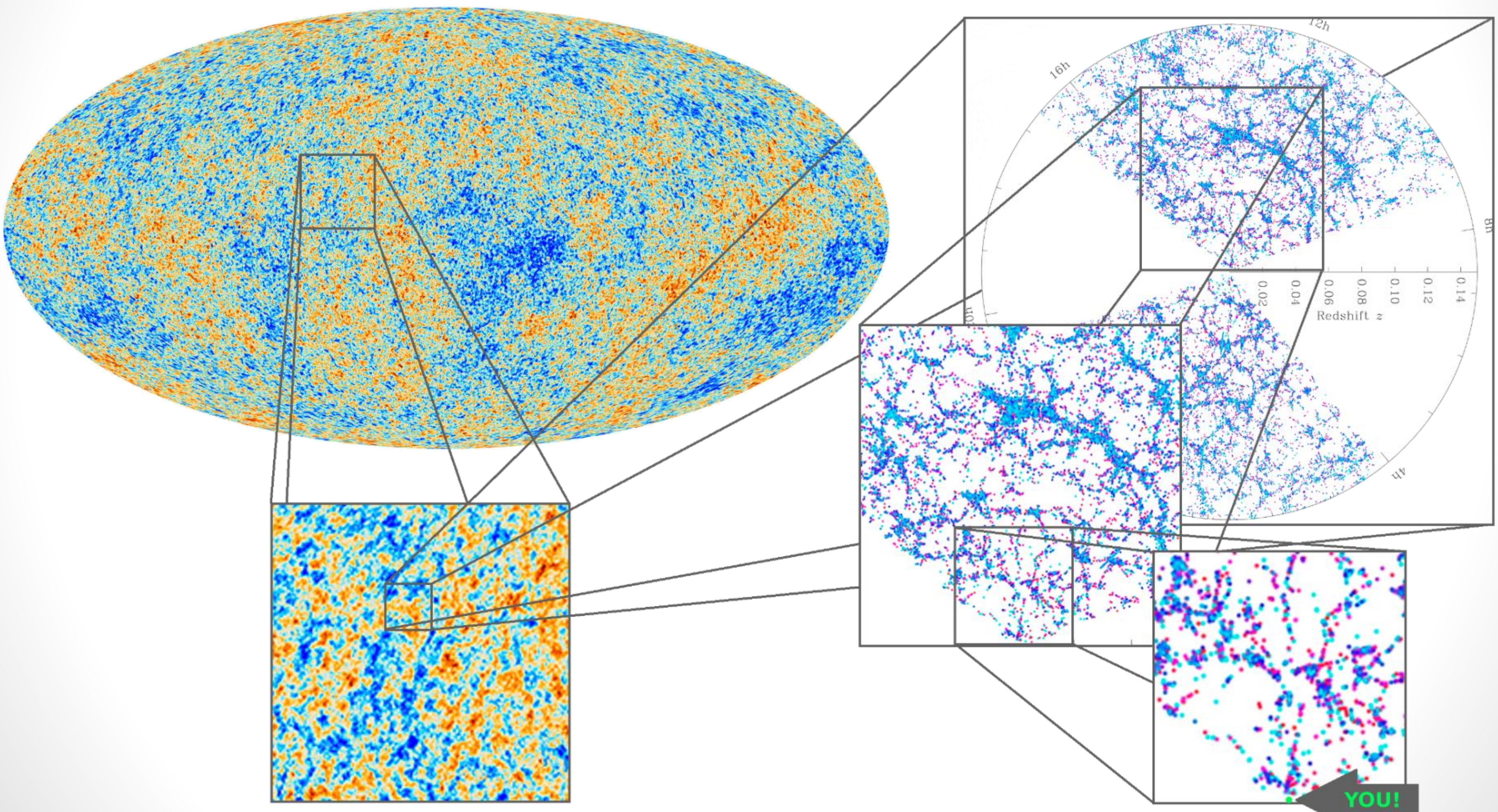
9 June 2021

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:

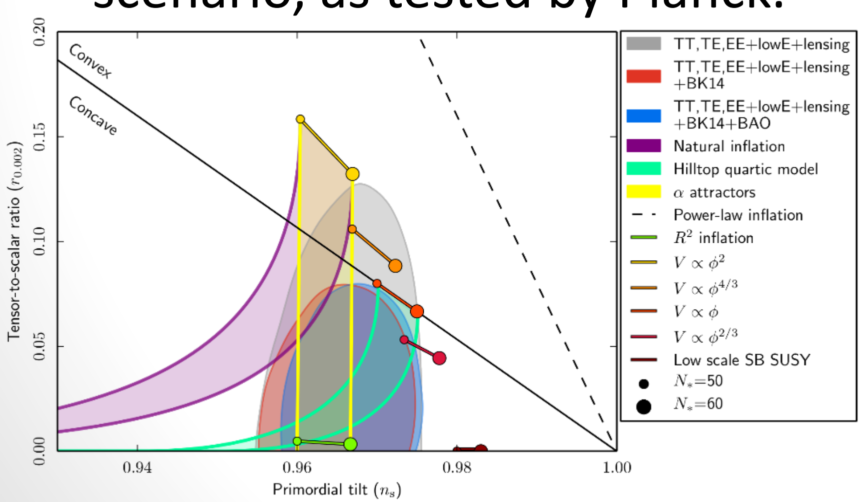
- Cosmology:
  - ΛCDM: cosmological parameters and tests against alternatives,
  - Physical nature of the dark components,
  - Neutrinos: number and masses,
  - Geometry of the Universe,
  - Tests of General Relativity,
  - Initial conditions and link to high energy physics
- Astrophysics: galaxy formation 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, 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)

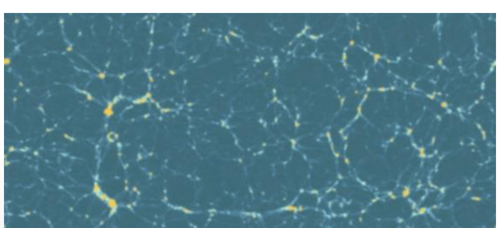
… how do we test these models against survey data?

J. Cham – PhD comics

Redshift range	Volume (Gpc^3)	$k_{\rm max}$ (Mpc/h)$^{-1}$	$N_{\rm modes}$
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 $k^3_{\rm max}$.



- 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: going beyond the linear and static analysis of the LSS.

Why Bayesian inference?

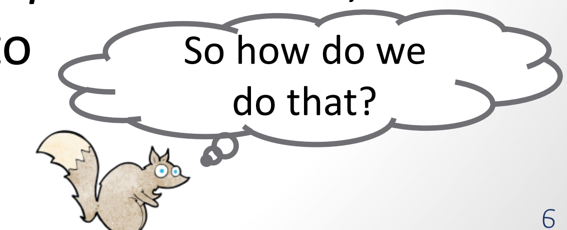
- Inference of signals = ill-posed problem
  - Incomplete observations: finite resolution, survey geometry, selection effects
  - Noise, biases, systematic effects
  - Cosmic variance

⇒ No unique recovery is possible!

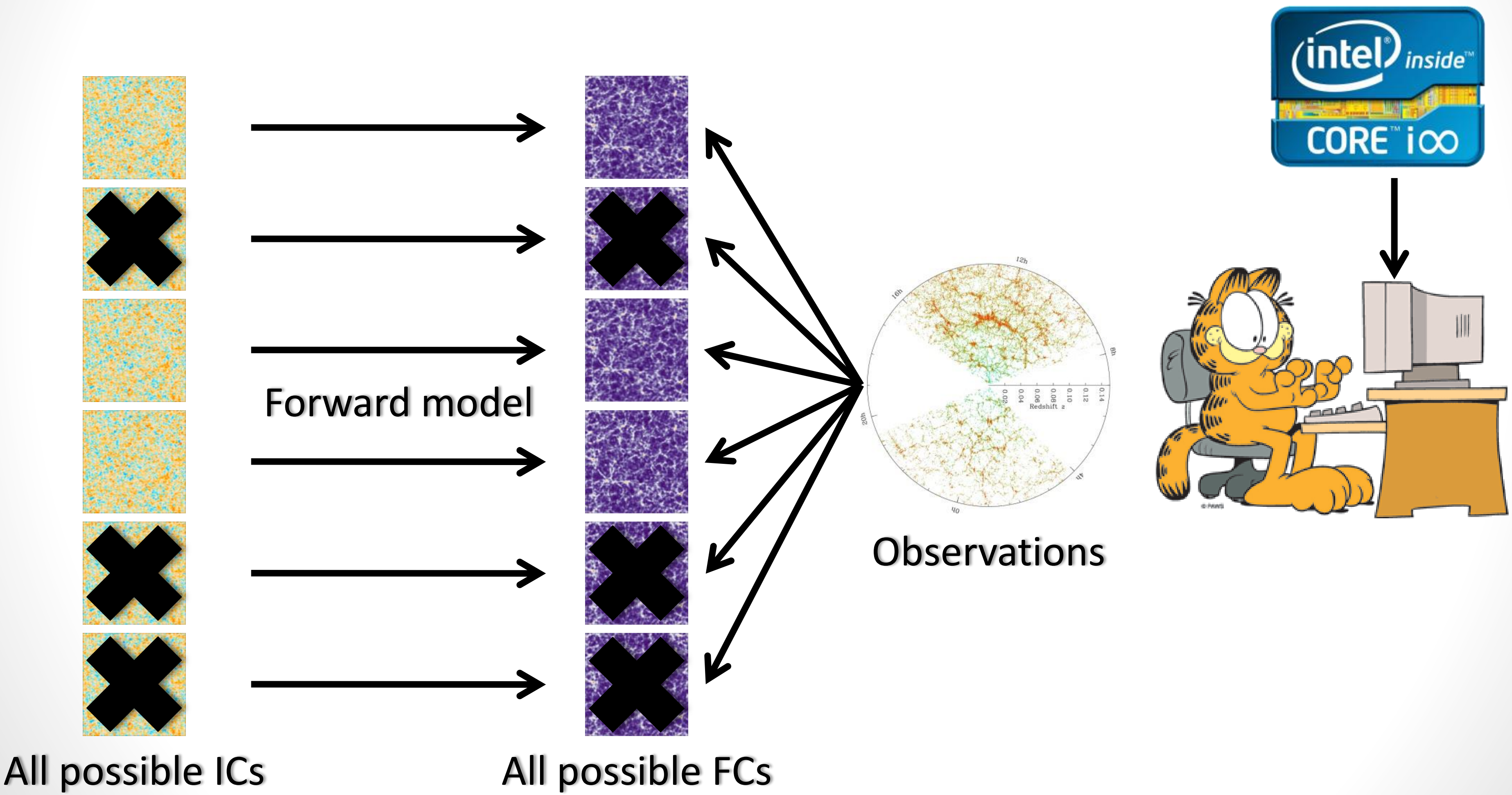
"What is the formation history of the Universe?" ⇒ "What is the probability distribution of possible formation histories (signals) compatible with the observations?"

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

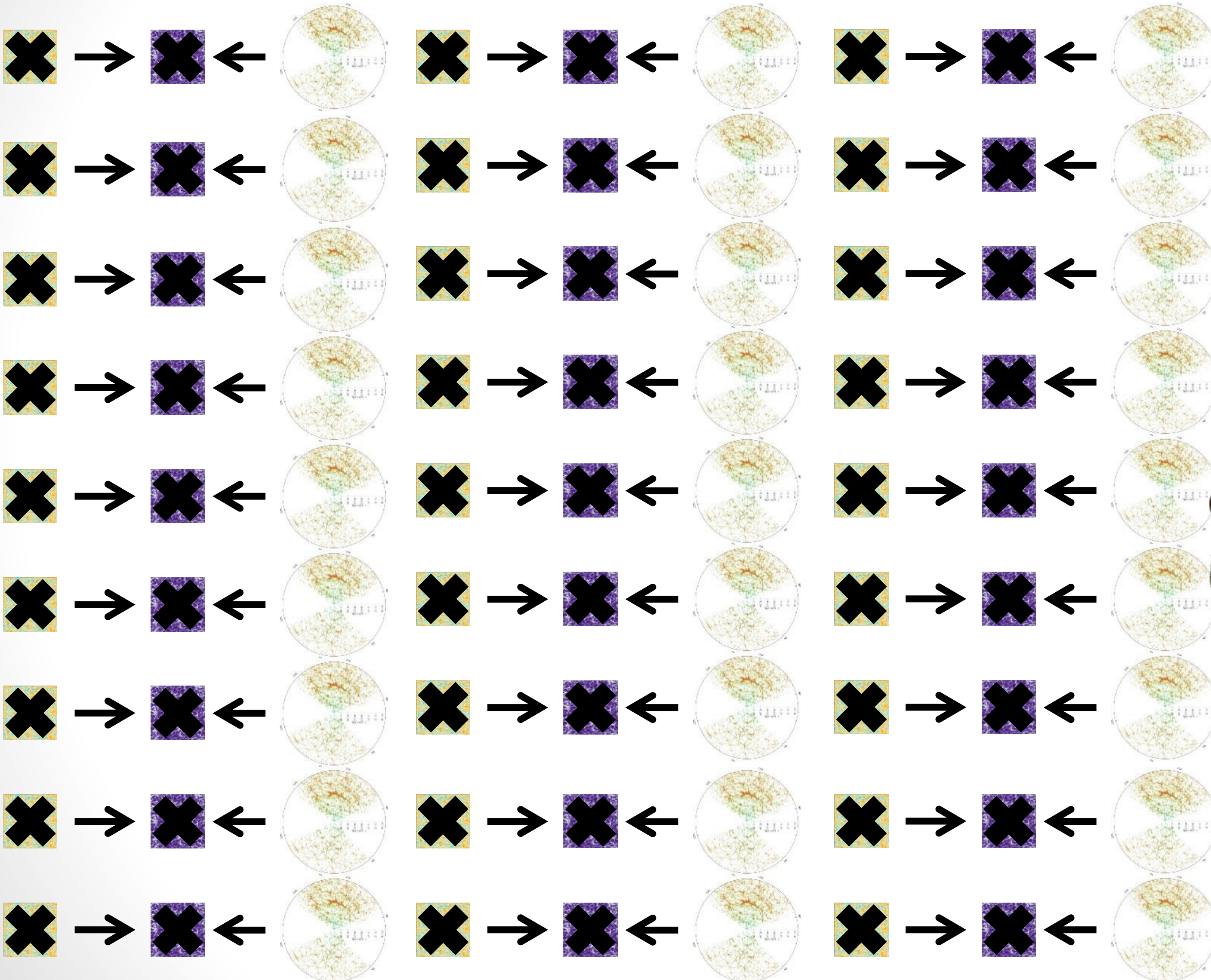
- Cox-Jaynes theorem: Any system to manipulate *"plausibilities"*, consistent with Cox's desiderata, is isomorphic to (Bayesian) probability theory



Bayesian forward modelling: the ideal scenario



Bayesian forward modelling: the challenge

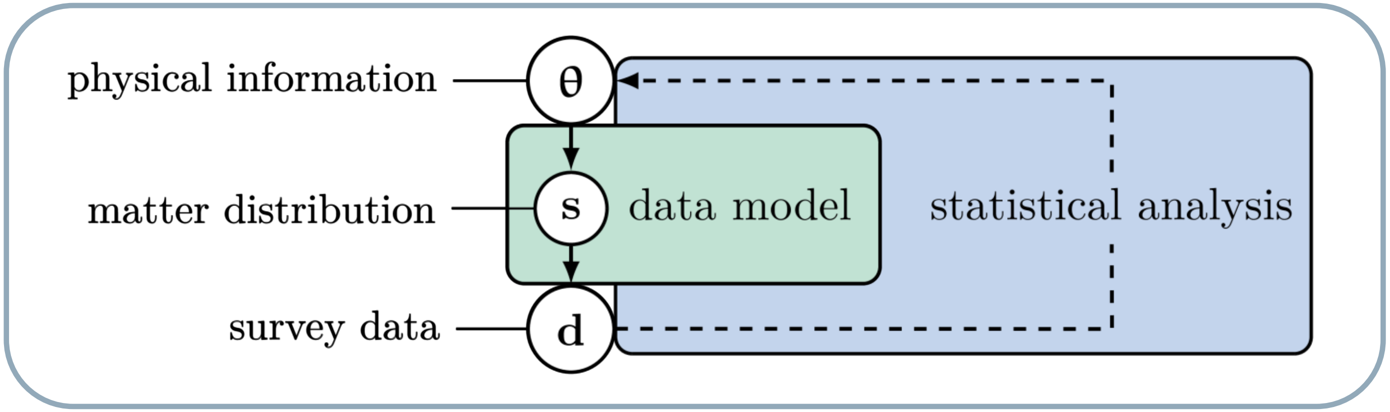


The (true) likelihood lives in

$d \approx 10^7$!

Making inferences requires advanced Bayesian techniques

- The physical computer models are incorporated into **Bayesian hierarchical models**.



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

Data assimilation:

exact statistical analysis

approximate data model

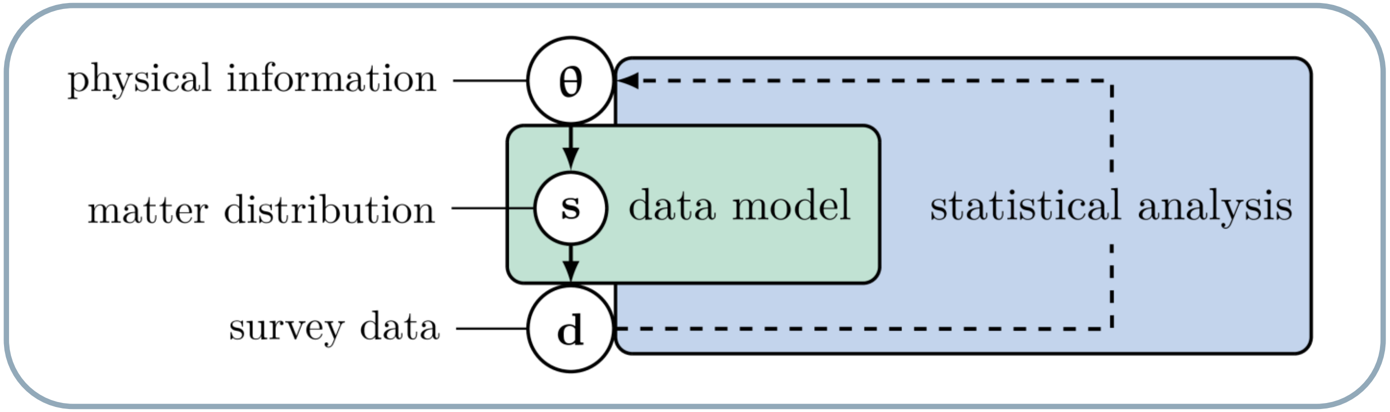
Simulation-based inference:

approximate statistical analysis

arbitrary data model

Likelihood-based solution: BORG

Bayesian Origin Reconstruction from Galaxies



<u>Data assimilation:</u>

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

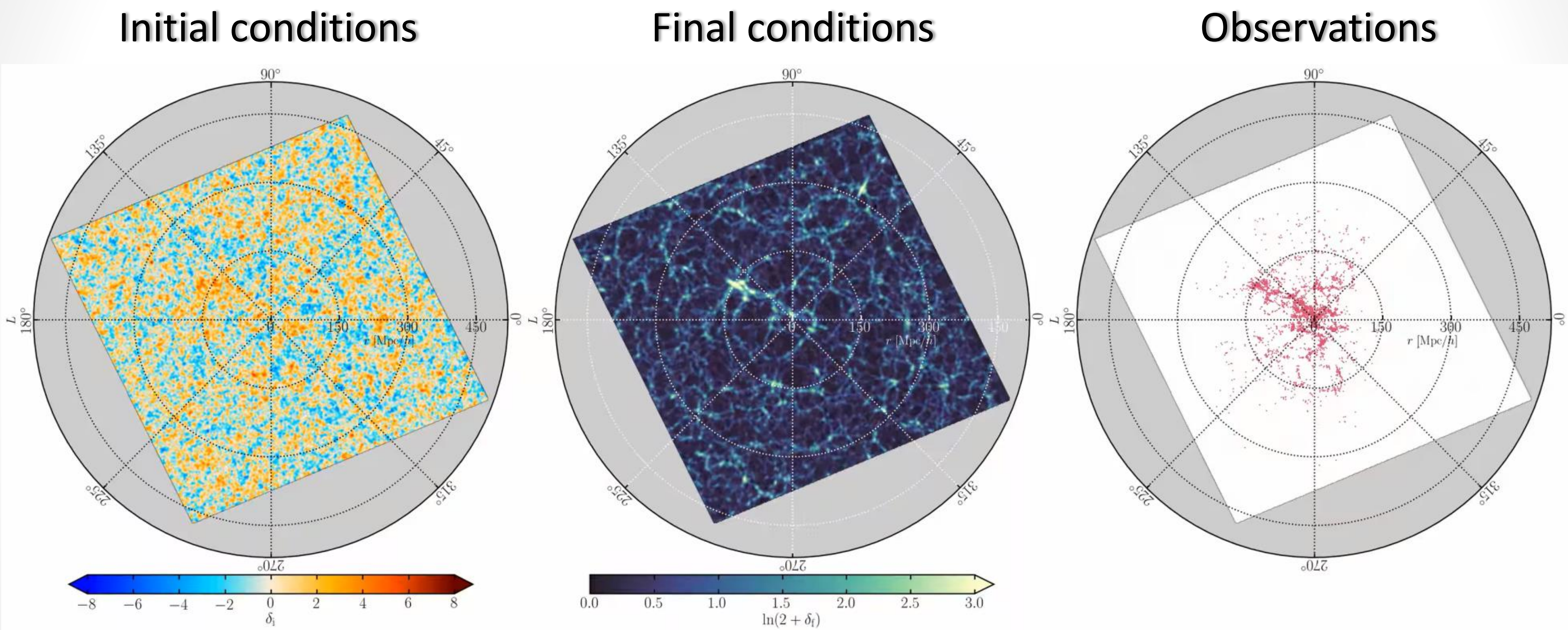
$$a(\mathbf{x}', \mathbf{x}) = \mathrm{e}^{-(H'-H)} = 1 \Leftarrow$$

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

BORG at work: Bayesian chrono-cosmography

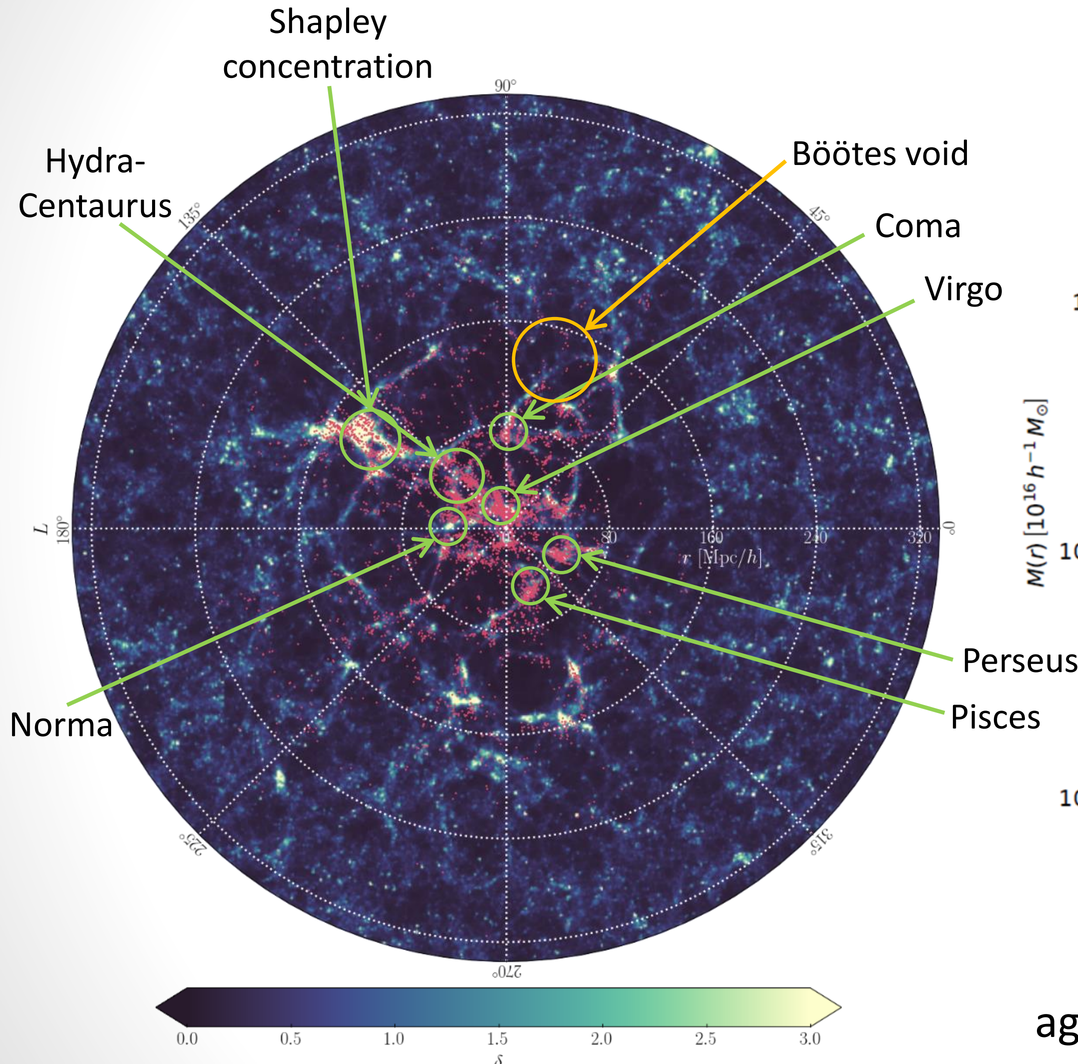


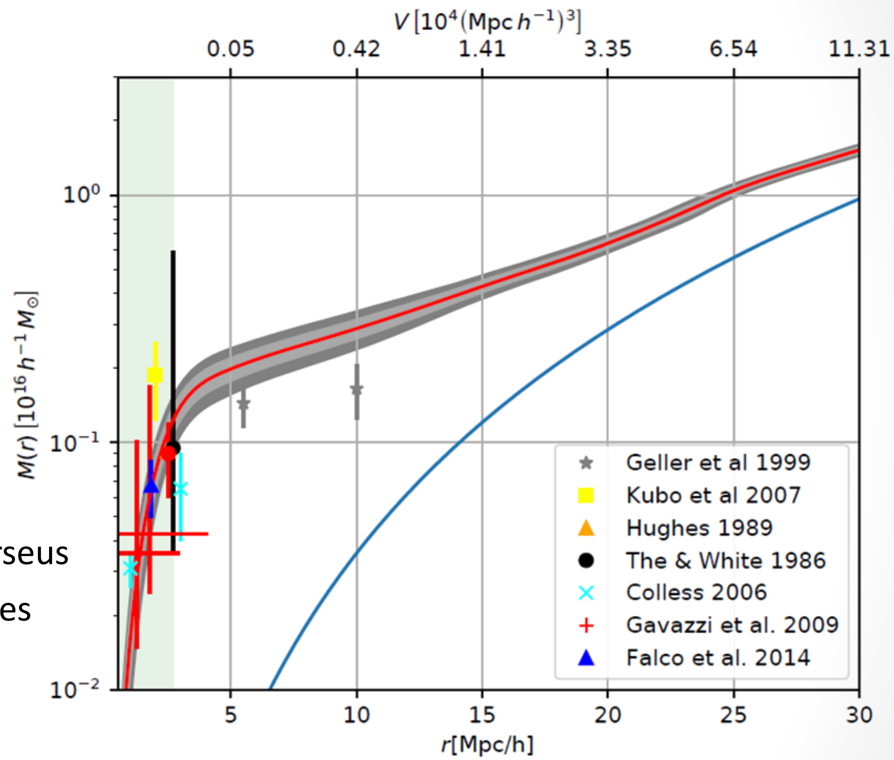
Supergalactic plane

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

BORGPM density field: full non-linear dynamics

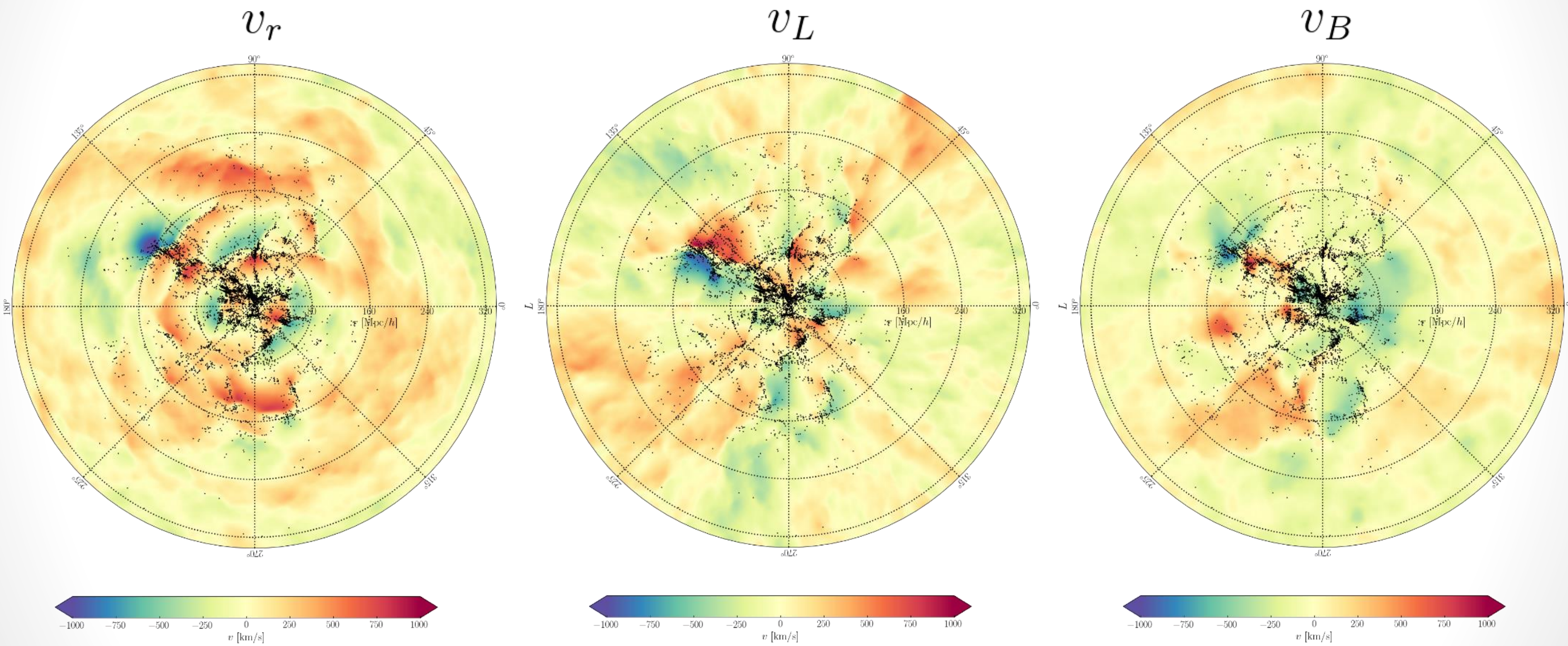




Mass profile of the Coma cluster, in agreement with gravitational lensing and X-ray observations down to a few Mpc.

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

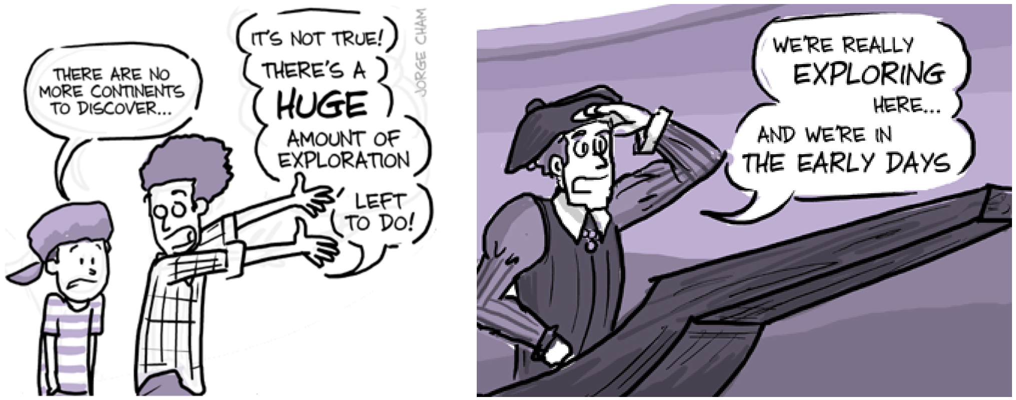
Velocity field in the supergalactic plane



The gravitational infall of known structures can be observed.

with James Prideaux-Ghee (PhD student), Guilhem Lavaux & Alan Heavens

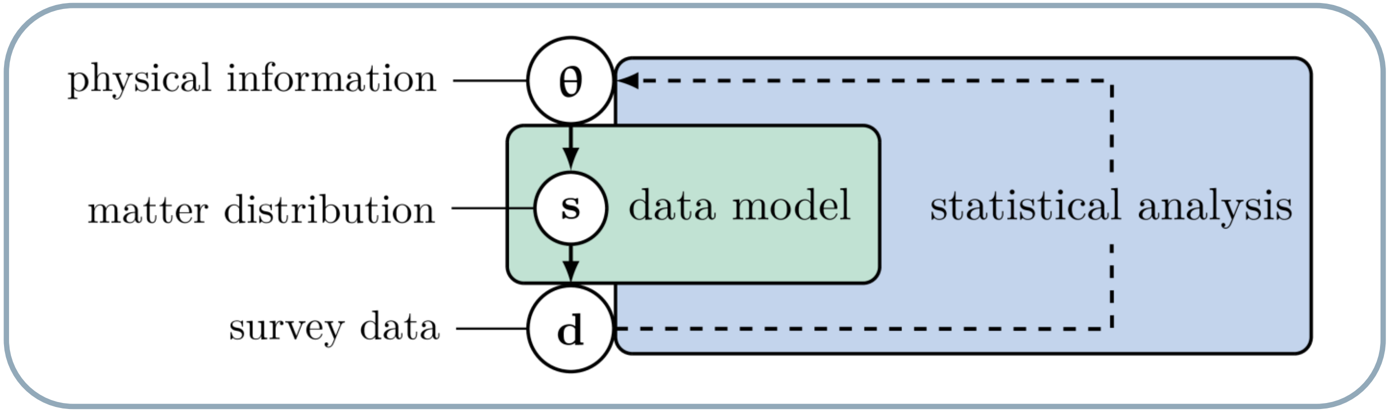
Mapping the Universe: epilogue?

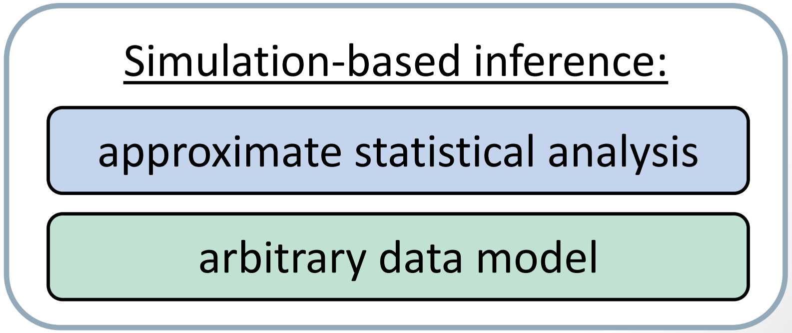


J. Cham – PhD comics

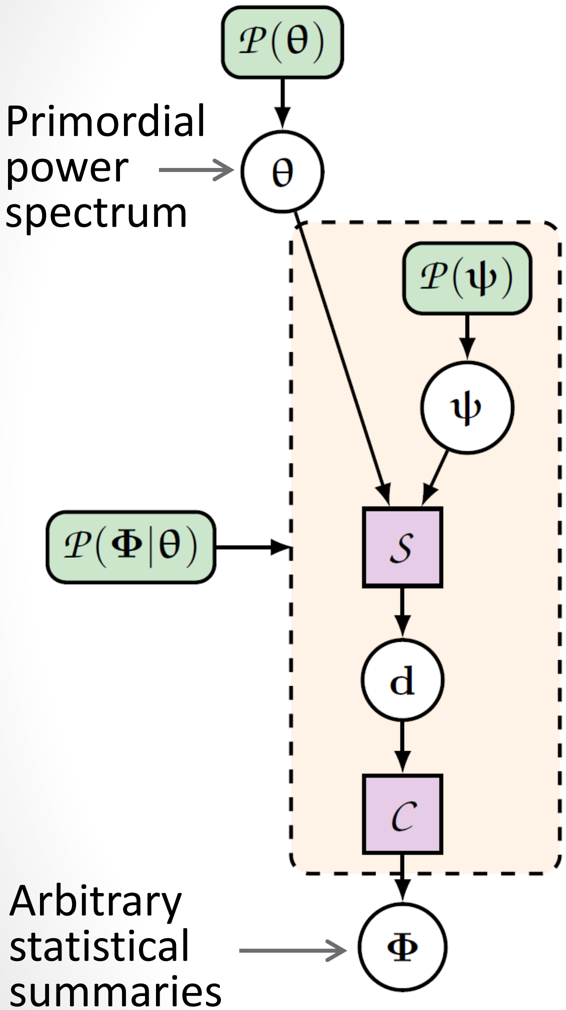
Likelihood-free solution: SELFI

Simulator Expansion for Likelihood-Free Inference

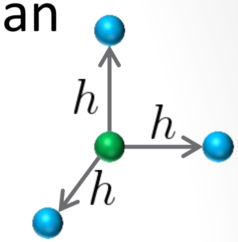




SELFI: Method



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



$$\hat{\boldsymbol{\Phi}}_{\theta} \approx \mathbf{f}_0 + \nabla \mathbf{f}_0 \cdot (\theta - \theta_0)$$

⇒ The posterior is Gaussian and analogous to a Wiener filter:

expansion point

observed summaries

$$\boldsymbol{\gamma} \equiv \theta_0 + \boldsymbol{\Gamma}\,(\nabla \mathbf{f}_0)^\intercal\, \mathbf{C}_0^{-1}(\boldsymbol{\Phi}_\mathrm{O} - \mathbf{f}_0)$$
$$\boldsymbol{\Gamma} \equiv \left[(\nabla \mathbf{f}_0)^\intercal\, \mathbf{C}_0^{-1} \nabla \mathbf{f}_0 + \mathbf{S}^{-1}\right]^{-1}$$

prior covariance

covariance of summaries

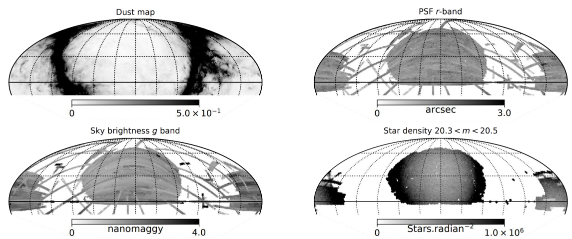
gradient of the black-box

$\mathbf{f}_0, \mathbf{C}_0$ and $\nabla \mathbf{f}_0$ can be evaluated through simulations only.
The number of required simulations is fixed *a priori*.

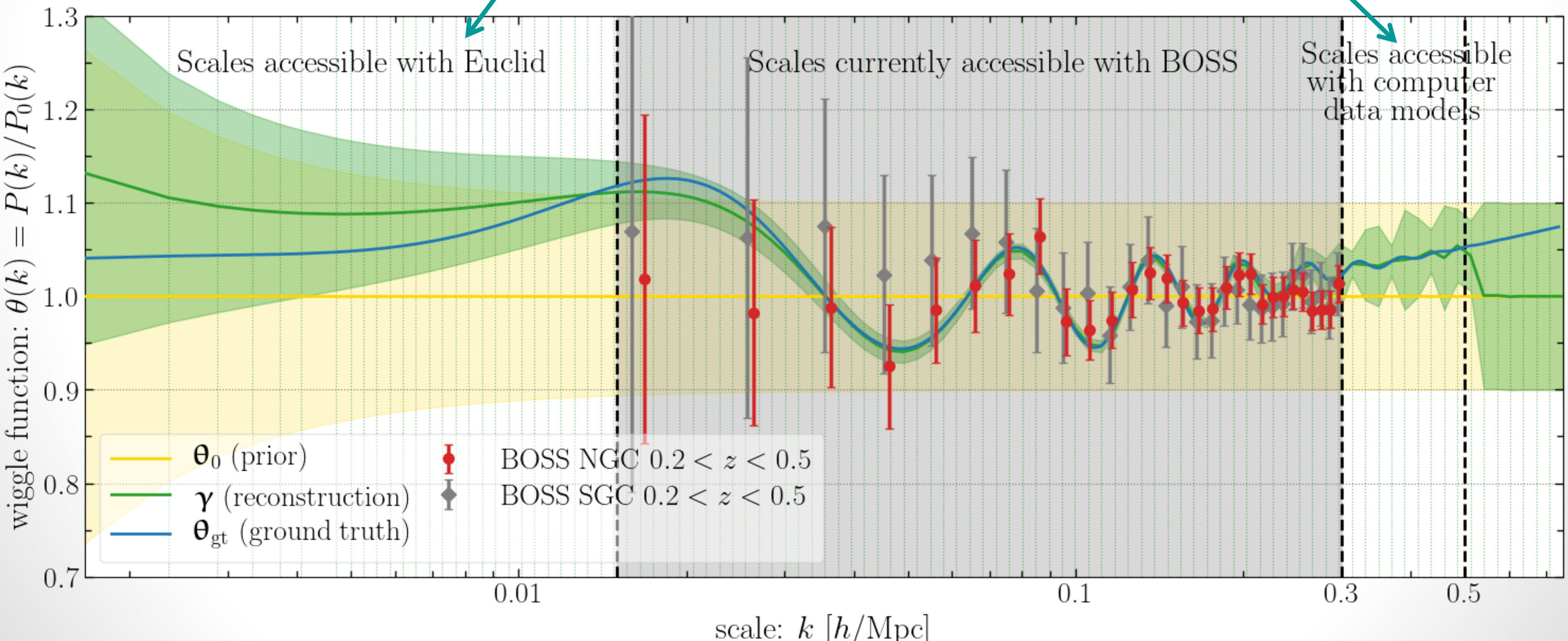
FL, Enzi, Jasche & Heavens 2019, 1902.10149

Euclid GC-LFI forecast (SELFI-1 Euclid versus BOSS)

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



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



Perfectly parallel simulations using sCOLA

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

Tassev, Zaldarriaga & Einsenstein 2013, 1301.0322
Tassev, Eisenstein, Wandelt & Zaldarriaga 2015, 1502.07751

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}_{\mathrm{LPT}} + \boldsymbol{\Psi}_{\mathrm{res}} \quad (\mathbf{x} = \mathbf{q} + \boldsymbol{\Psi})$$

Analytical solutions!

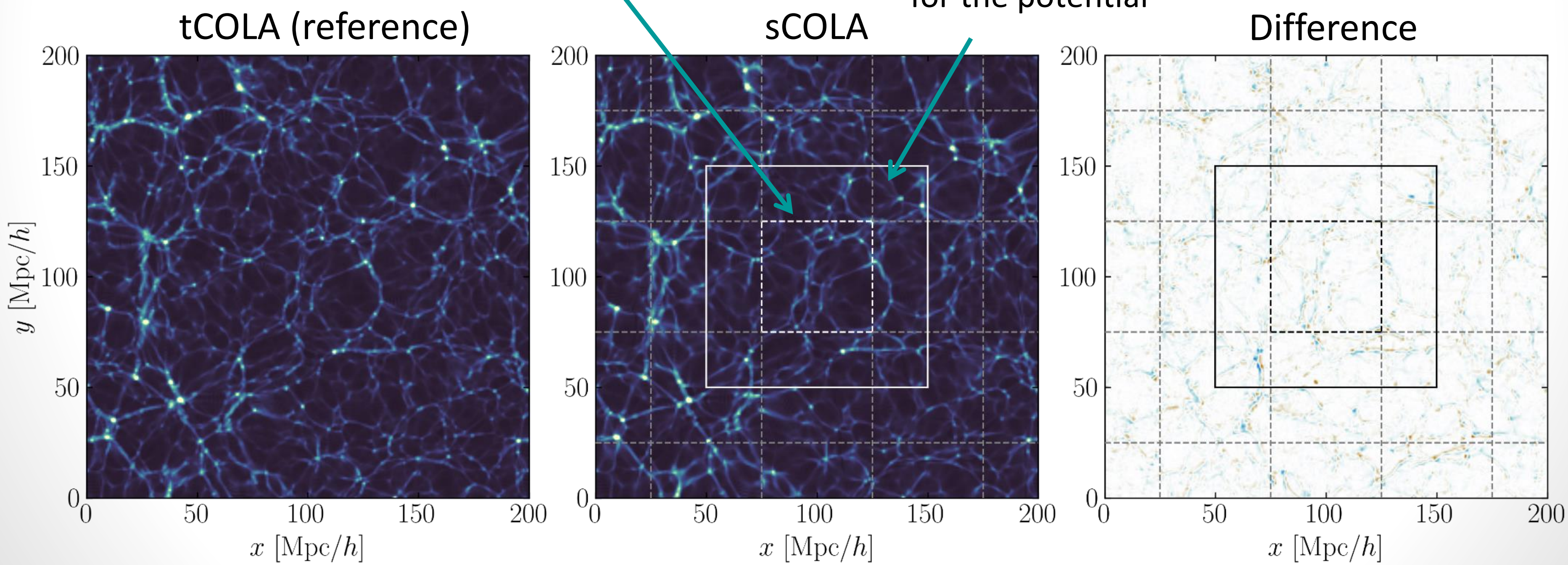
Standard:

$$\partial_a^2 \boldsymbol{\Psi} = -\boldsymbol{\nabla}_{\mathbf{x}} \Phi$$

Modified:

$$\partial_a^2 \boldsymbol{\Psi}_{\mathrm{res}} = \partial_a^2 (\boldsymbol{\Psi} - \boldsymbol{\Psi}_{\mathrm{LPT}}) = -\boldsymbol{\nabla}_{\mathbf{x}} \Phi - \partial_a^2 \boldsymbol{\Psi}_{\mathrm{LPT}}$$

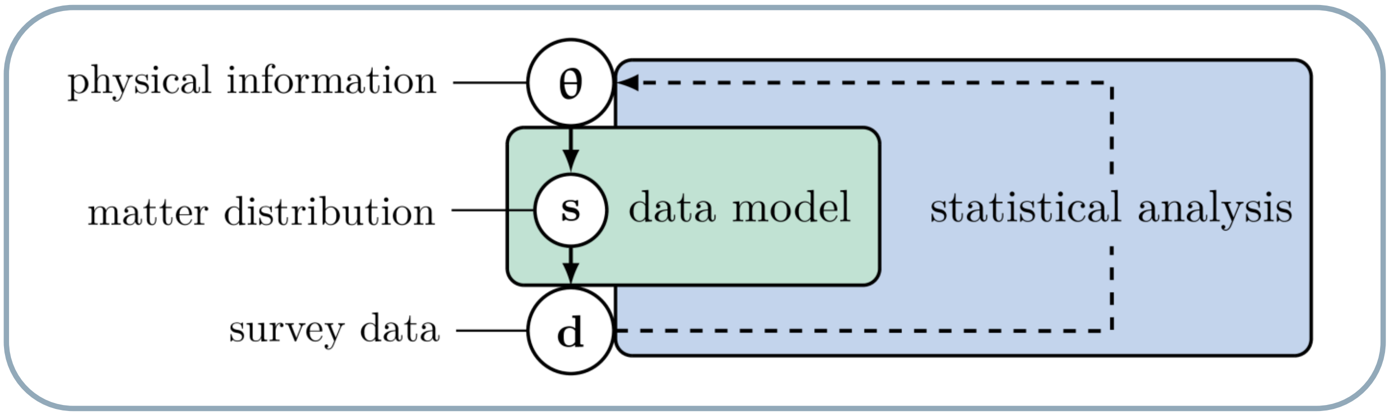
1. A buffer region around each tile
2. Appropriate Dirichlet boundary conditions for the potential



Publicly available implementation:
https://bitbucket.org/florent-leclercq/simbelmyne/

FL, Faure, Lavaux, Wandelt, Jaffe, Heavens, Percival & Noûs 2020, 2003.04925

So, which one is the best?



Data assimilation:

exact statistical analysis

approximate data model

Simulation-based inference:

approximate statistical analysis

arbitrary data model

Correlation functions versus field-level inference

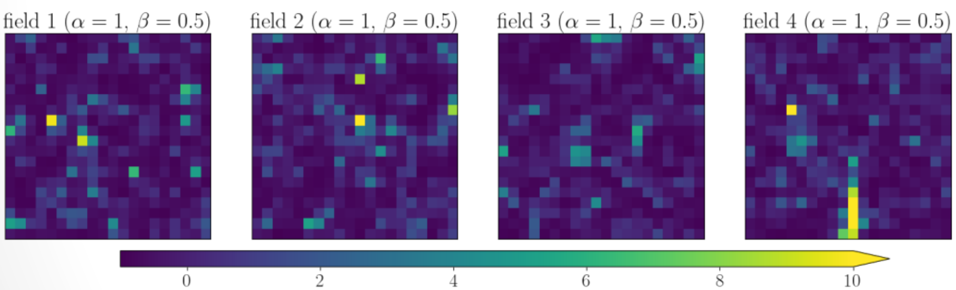
- We checked accuracy and precision of different methods for a log-normal model:

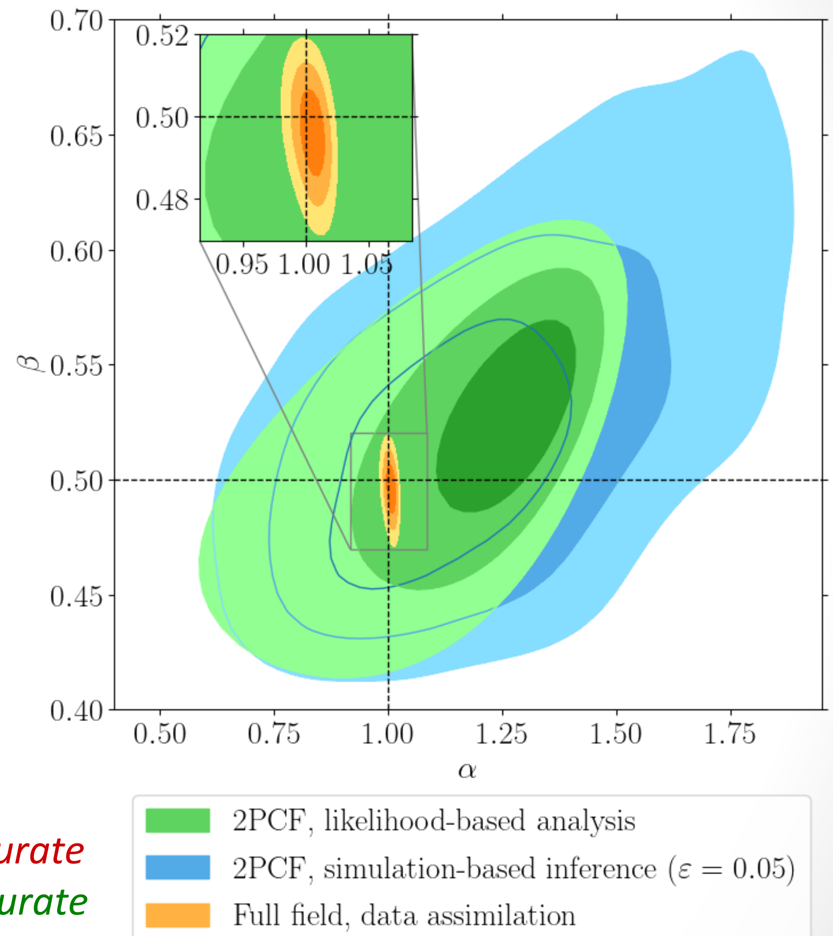
$$f = \frac{1}{\alpha}\left[\boxed{\exp\left(\alpha g\right.}\left. - \frac{1}{2}\alpha^2\right) - 1\right]$$

Log-normal field

Gaussian field with 2PCF:

$$\xi_g(r) = \exp\left(-\frac{1}{4}\frac{r^2}{\beta^2}\right)$$





- 2PCF likelihood-based analysis is *imprecise* and *inaccurate*
- 2PCF simulation-based inference is *imprecise* but *accurate*
- Full-field data assimilation is *precise* and *accurate*

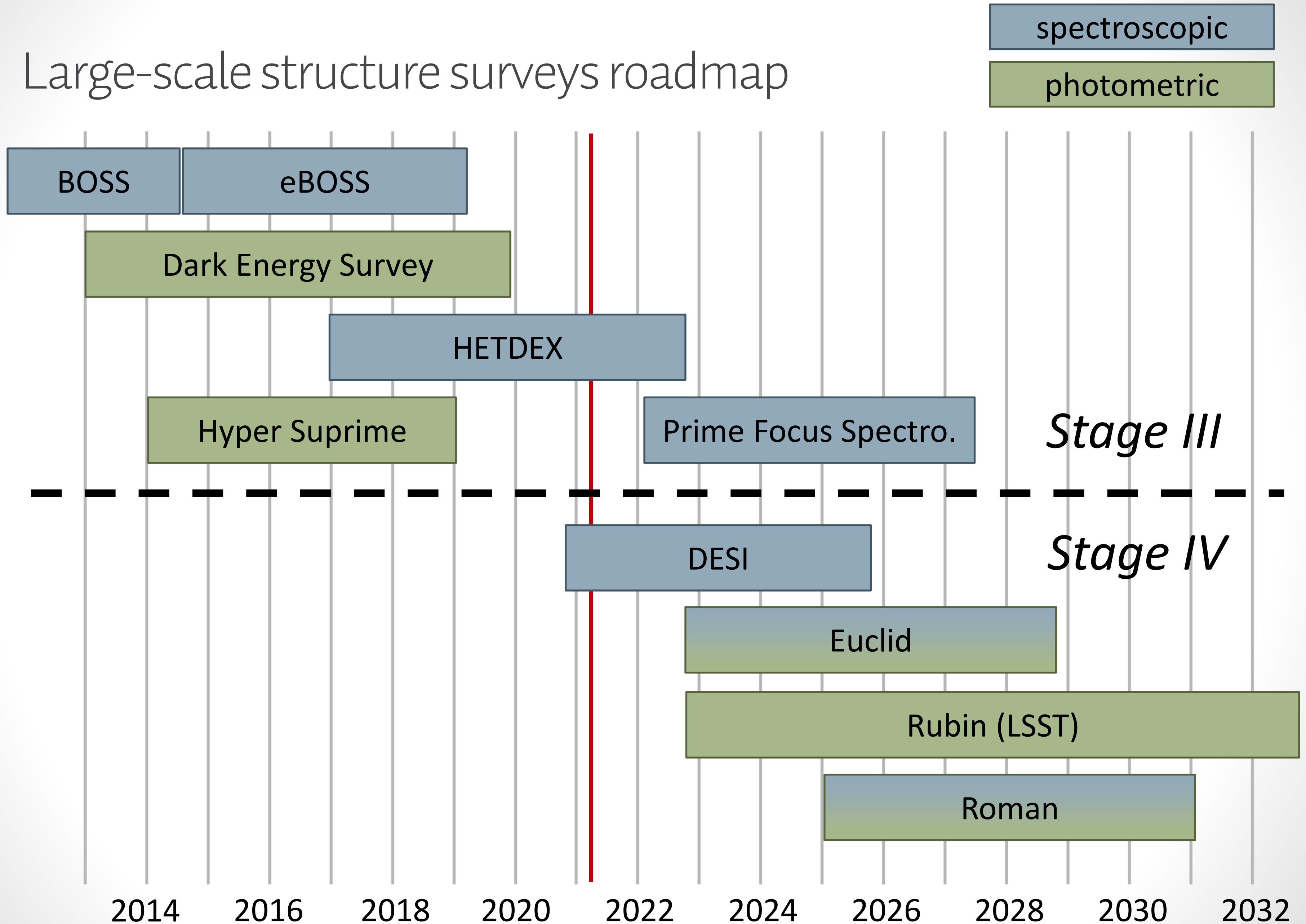
Companion repository:
https://github.com/florent-leclercq/correlations_vs_field

FL & Heavens, 2103.04158

The Future: Opportunities & Challenges

DESI, Euclid, LSST, WFIRST, and more…

Large-scale structure surveys roadmap



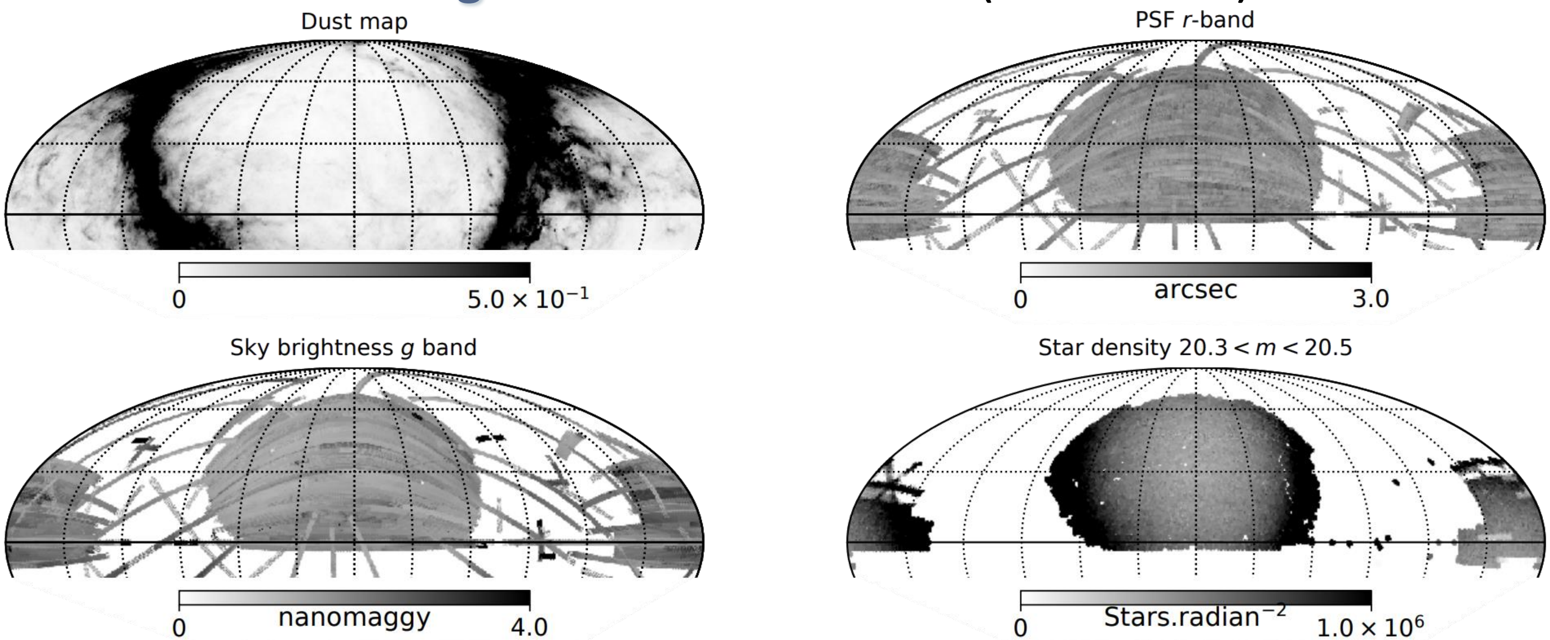
Data-intensive scientific discovery from galaxy surveys

- Next-generation surveys will be dominated by systematics
- 80% of the total signal will come from non-linear structures
- Challenging data analysis questions and/or hints for new physics will first show up as tensions between measurements

- Can data analysts keep pace?

Accounting for known and unknown systematics

- Some known foreground contaminants (11 in total)

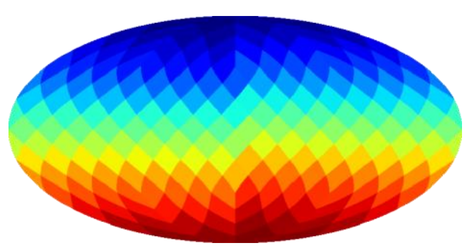


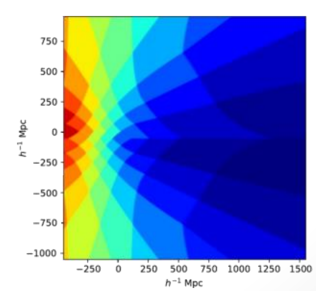
Forward model introduced by Jasche & Lavaux 2017, 1706.08971

- A procedure to marginalise over unknown foreground contaminations

Robust likelihood introduced by Porqueres, Ramanah, Jasche & Lavaux 2018, 1812.05113

Map of patches on the sky…



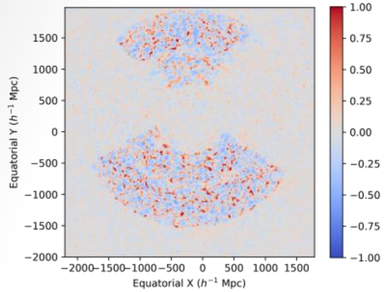


… extruded in 3D

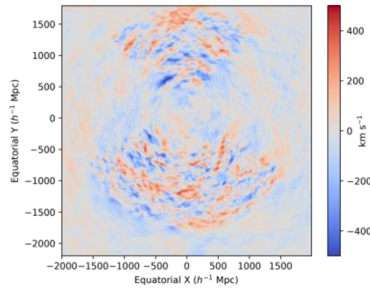
Lavaux, Jasche & FL 2019, 1909.06396

Application to SDSS-III/BOSS (LOWZ+CMASS)

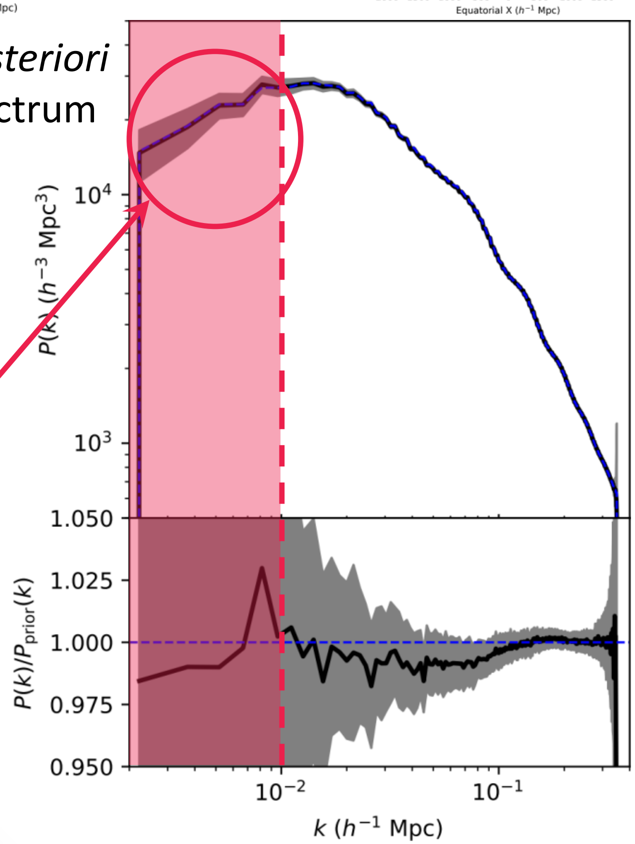
Density field



Velocity field



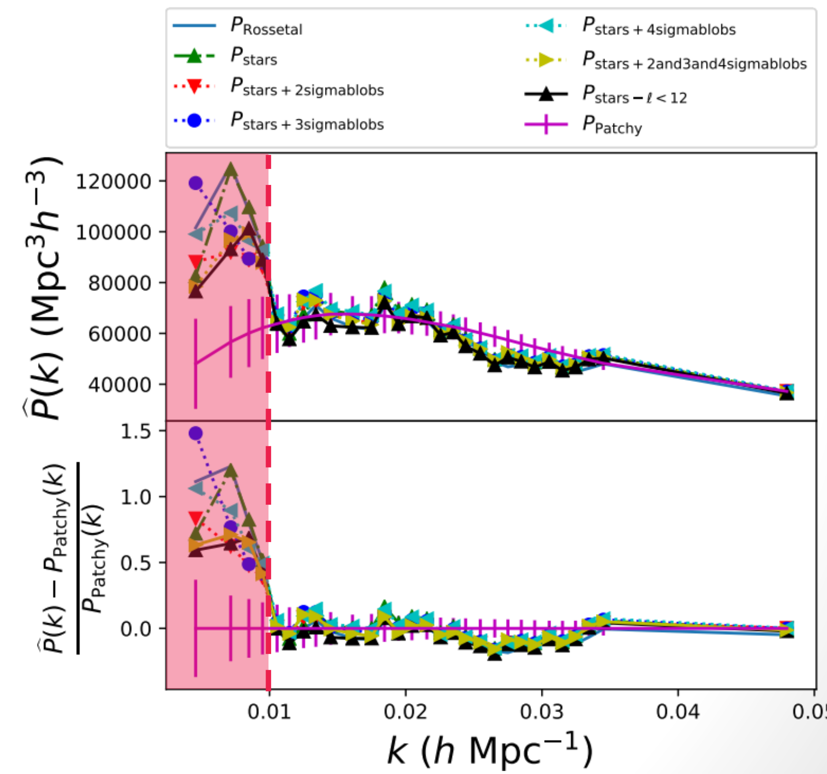
BORG *a posteriori* power spectrum



No apparent contamination, even well beyond the turn-over

Lavaux, Jasche & FL 2019, 1909.06396

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

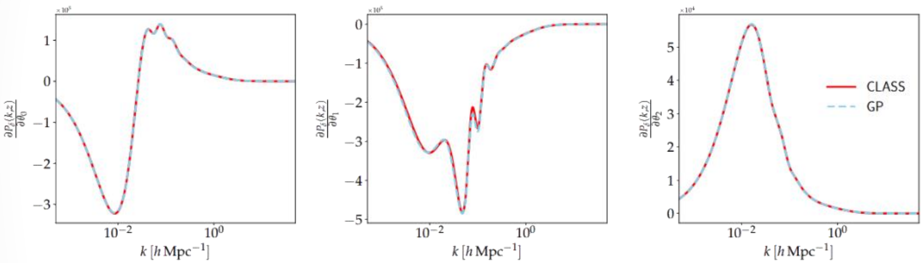


Kalus, Percival *et al.* 2018, 1806.02789

The Imperial weak lensing inference framework

with George Kyriacou (PhD student), Arrykrishna Mootoovaloo (PhD student), Natàlia Porqueres, Alan Heavens & Andrew Jaffe

- Gaussian process emulation and massive data compression for weak lensing cosmology



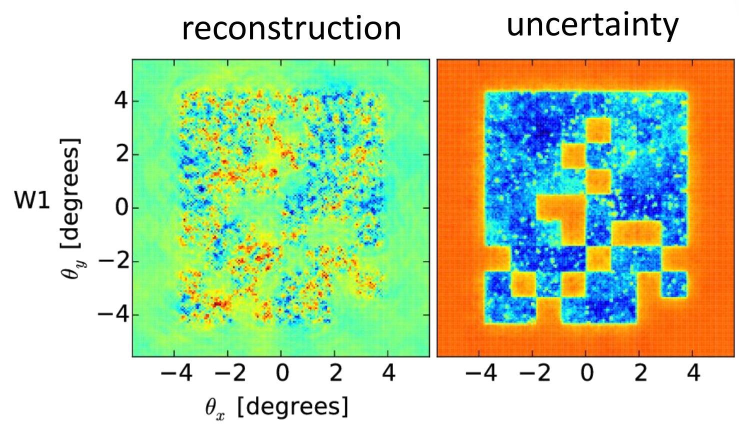
Mootoovaloo, Heavens, Jaffe & FL, 2005.06551

Mootoovaloo, Jaffe, Heavens & FL, 2105.02256

- Bayesian hierarchical inference of galaxy redshift distributions $n(z)$

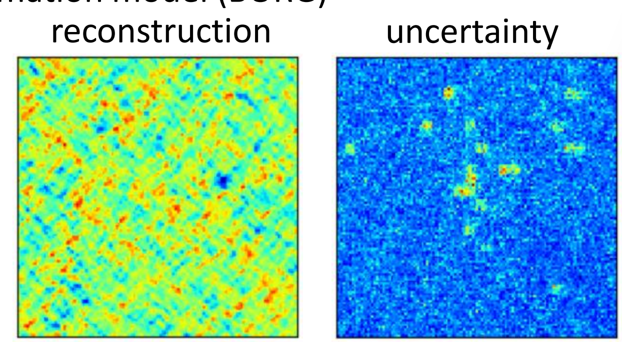
with G. Kyriacou, A. Jaffe & A. Heavens

- Joint inference of Gaussian cosmic shear maps and power spectra/cosmological parameters



Alsing, Heavens & Jaffe 2016, 1607.00008

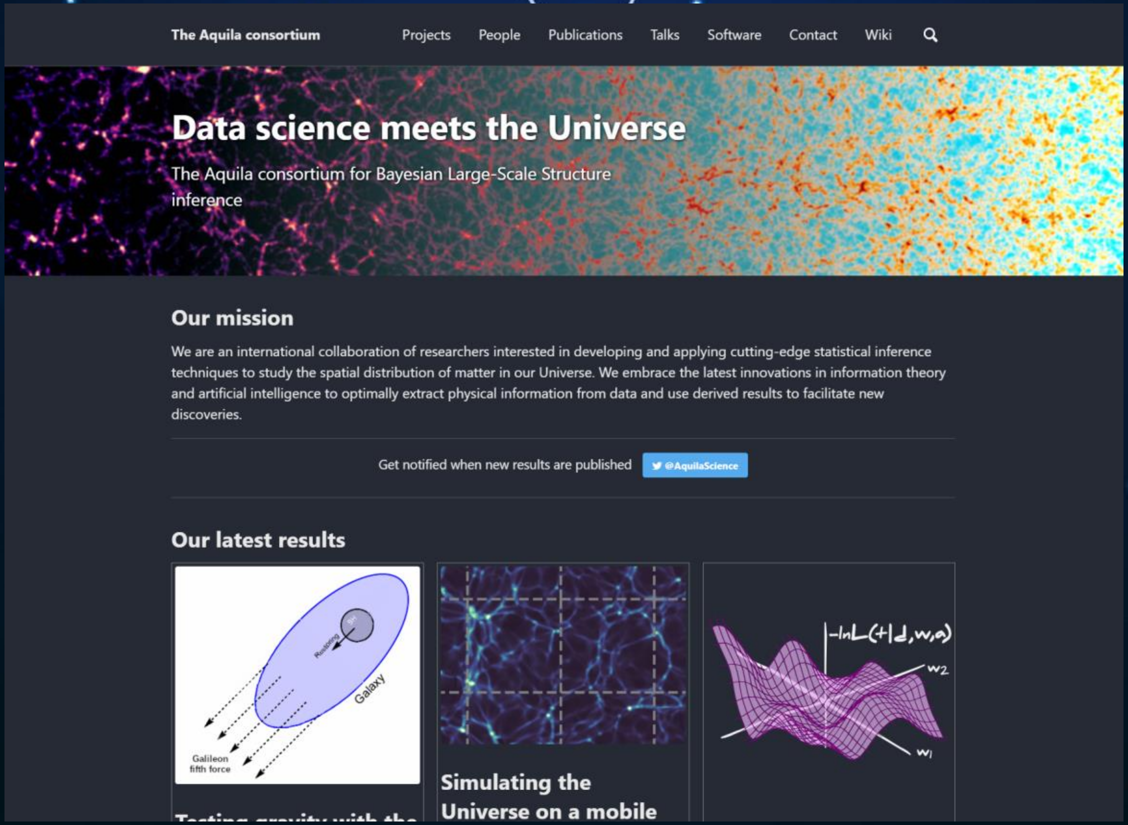
- Cosmic shear map inference with a structure formation model (BORG)



Porqueres, Heavens, Mortlock & Lavaux, 2011.07722

The Aquila Consortium

- Created in 2016. Currently 27 members from 9 countries (Europe & North America).
- Gathers people interested in developing the Bayesian pipelines and running analyses on cosmological data.



Visit us at www.aquila-consortium.org

Concluding thoughts

Data assimilation:	Simulation-based 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!
- A likelihood-based solution (BORG): general purpose reconstruction of dark matter from galaxy clustering, providing new measurements and predictions
- A likelihood-free solution (SELFI): algorithm 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)
- Tests of Gaussianity 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?