# 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

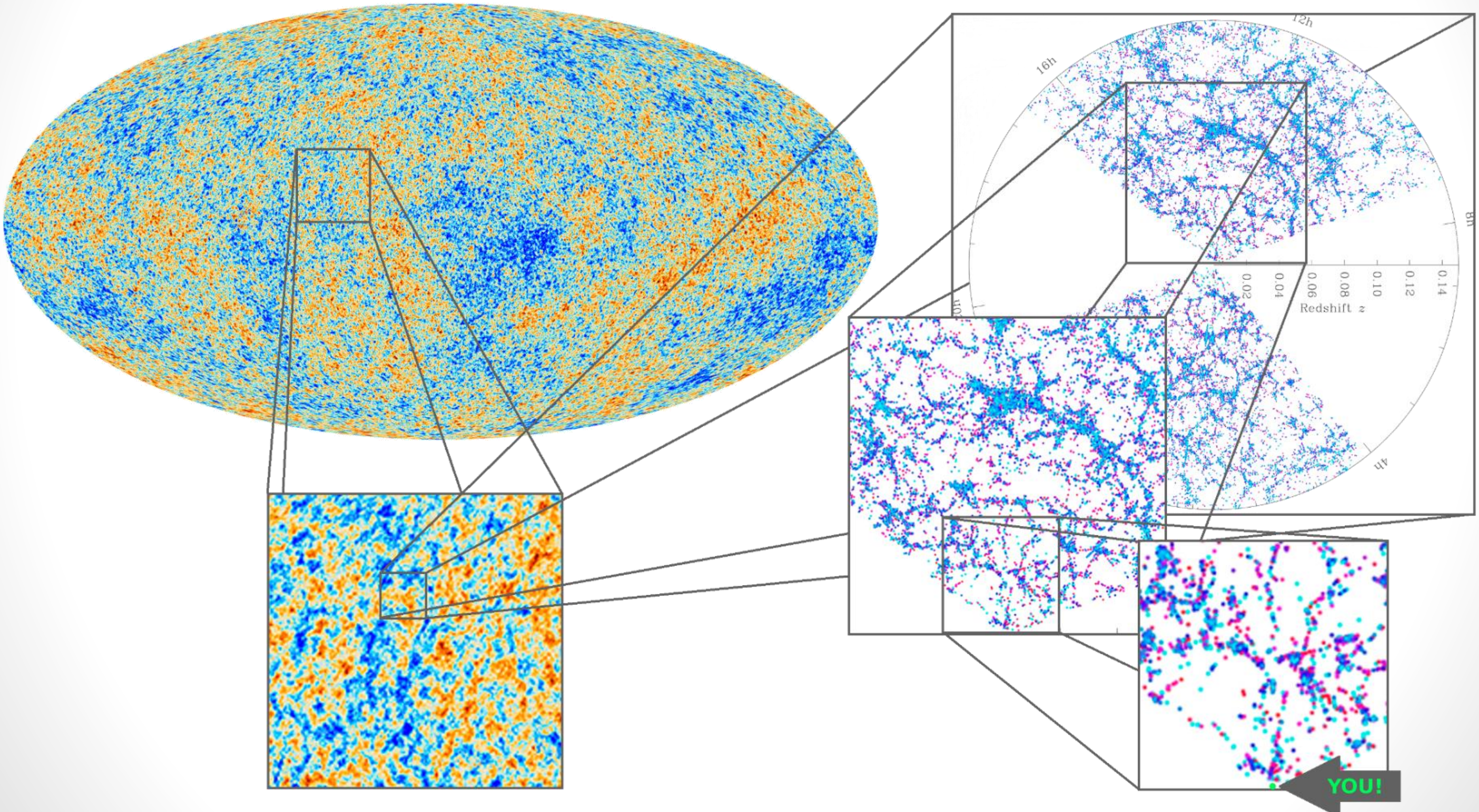
24 November 2021

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:

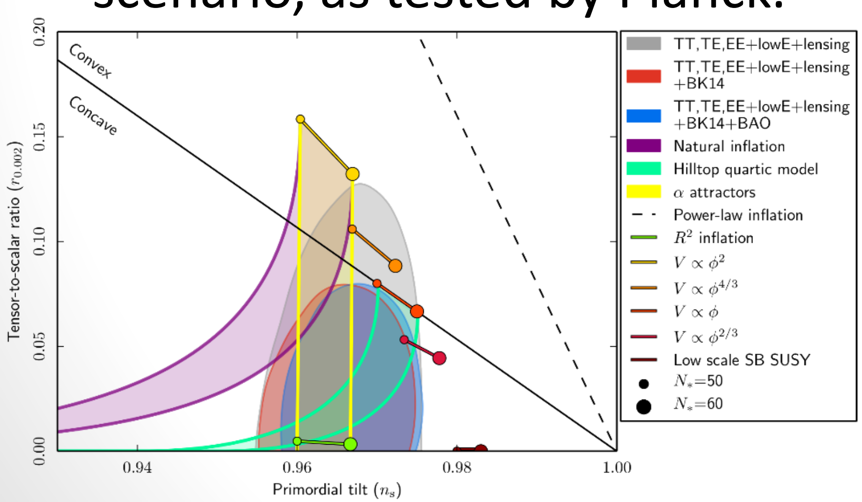
- Cosmology:
  - $\Lambda$CDM: cosmological parameters and tests against alternatives,
  - Physical nature of the dark components,
  - Neutrinos: number and masses,
  - Geometry of the Universe,
  - Tests of General Relativity,
  - Initial conditions and link to high energy physics
- Astrophysics: galaxy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

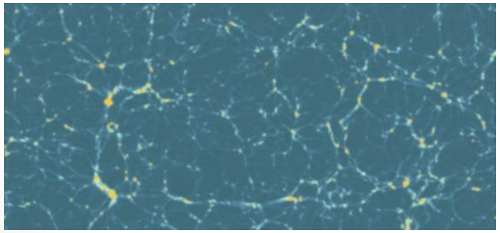
# … how do we test these models against survey data?


J. Cham – PhD comics

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



- 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 to 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

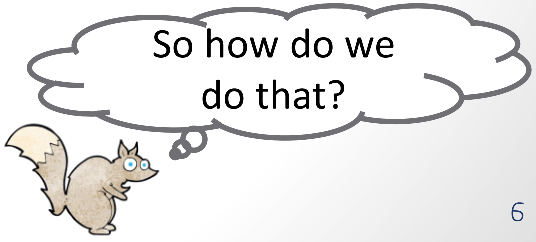
No unique recovery is possible!

"What is the formation history of the Universe?"

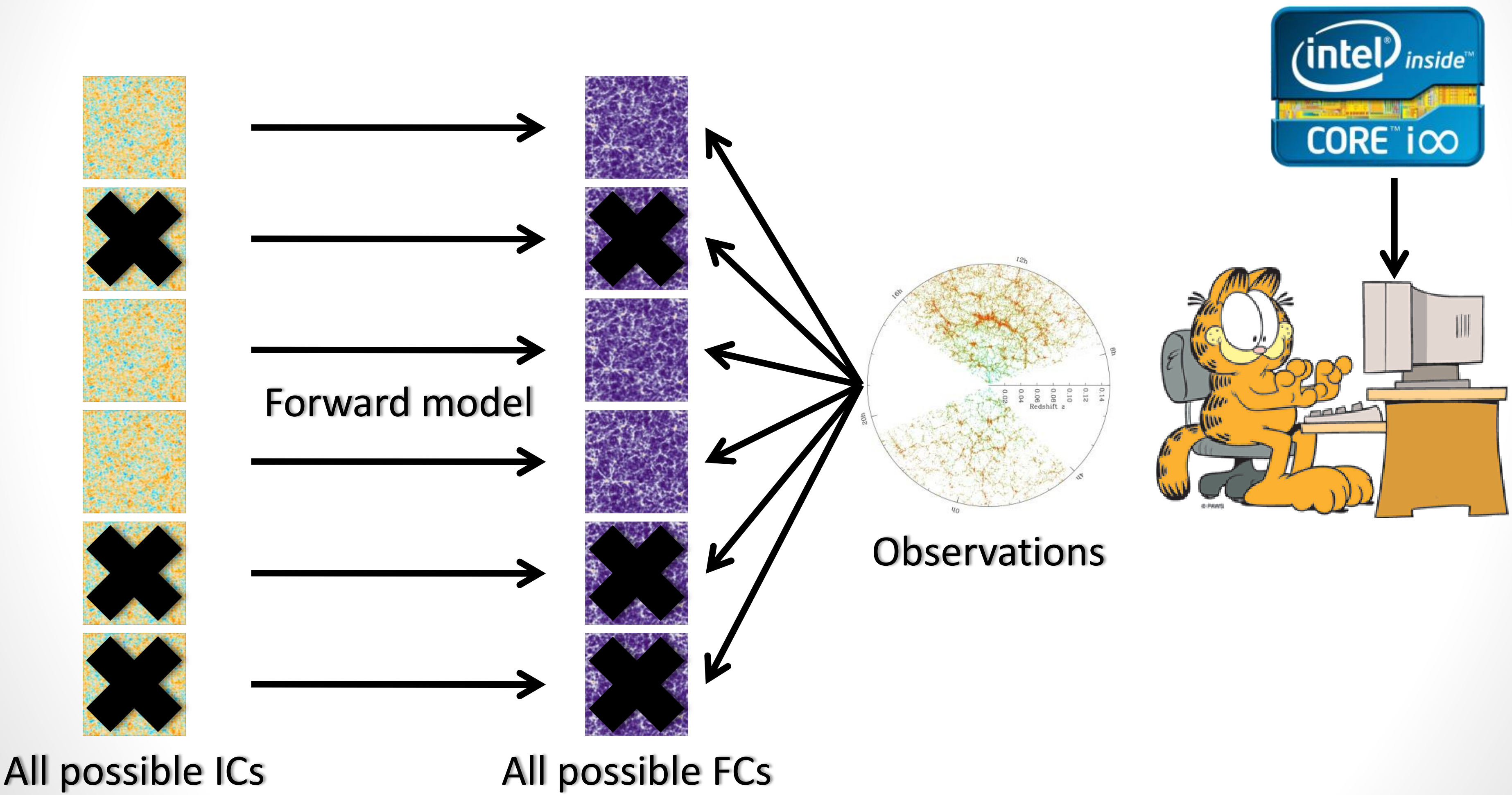
"What is the probability distribution of possible formation histories (signals) compatible with the observations?"

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

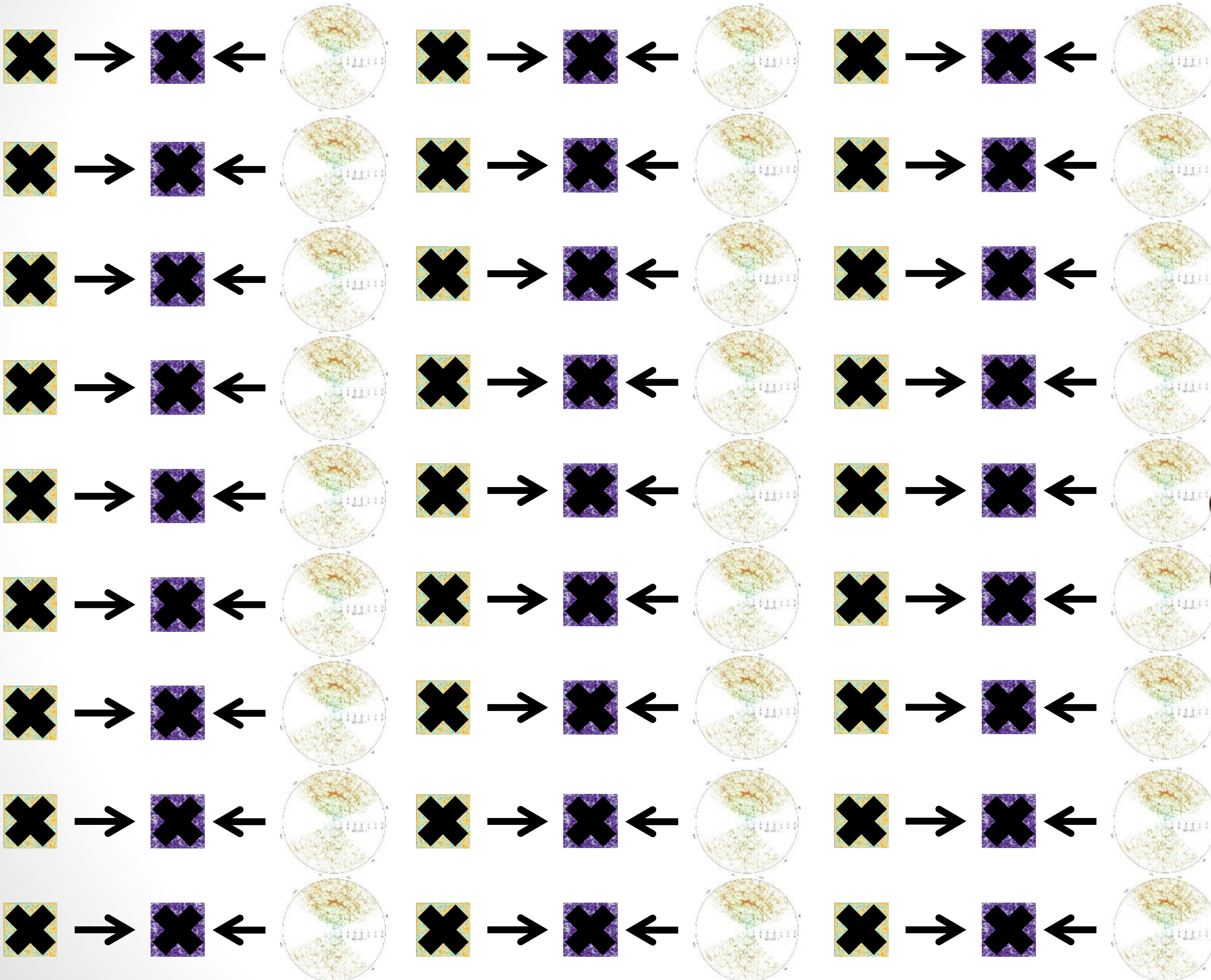
- Cox-Jaynes theorem: Any system to manipulate *"plausibilities"*, consistent with Cox's desiderata, is isomorphic to (Bayesian) probability theory



# Bayesian forward modelling: the ideal scenario



# Bayesian forward modelling: the challenge

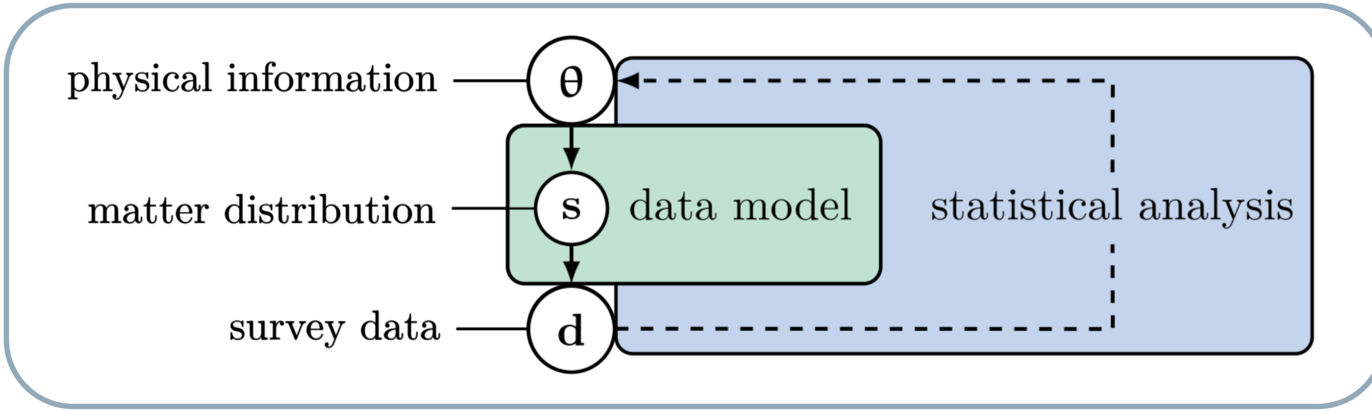


The (true) likelihood lives in

$d \approx 10^7$ !

# 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:

<u>Data assimilation:</u>

exact statistical analysis

approximate data model

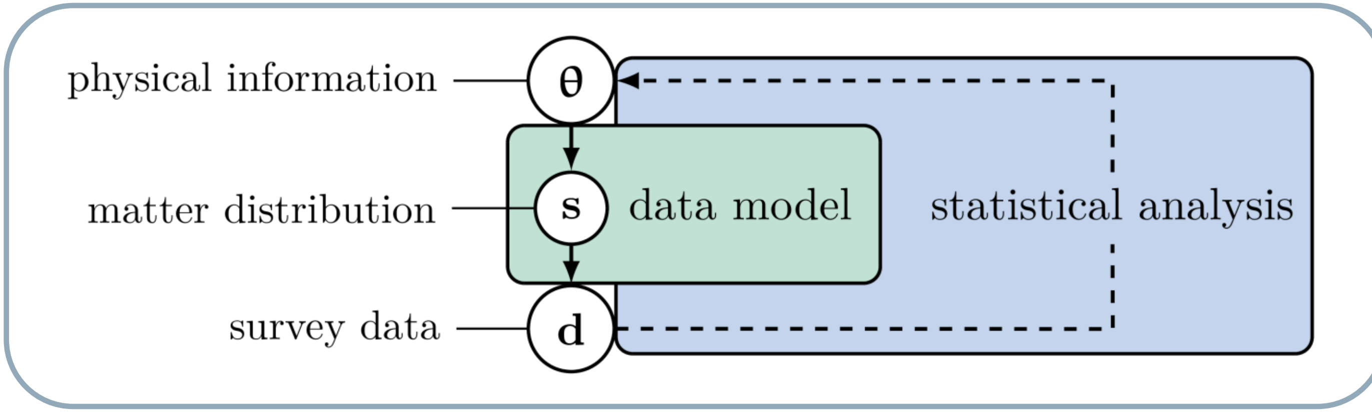
<u>Simulation-based inference:</u>

approximate statistical analysis

arbitrary data model

# Likelihood-free solution: SELFI

*Simulator Expansion for Likelihood-Free Inference*



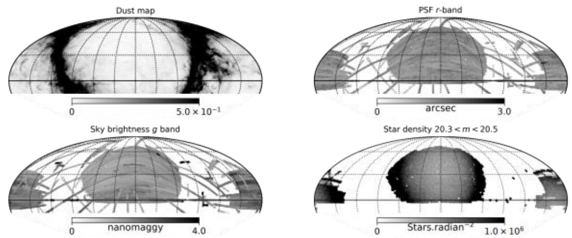

<u>Simulation-based inference:</u>

approximate statistical analysis

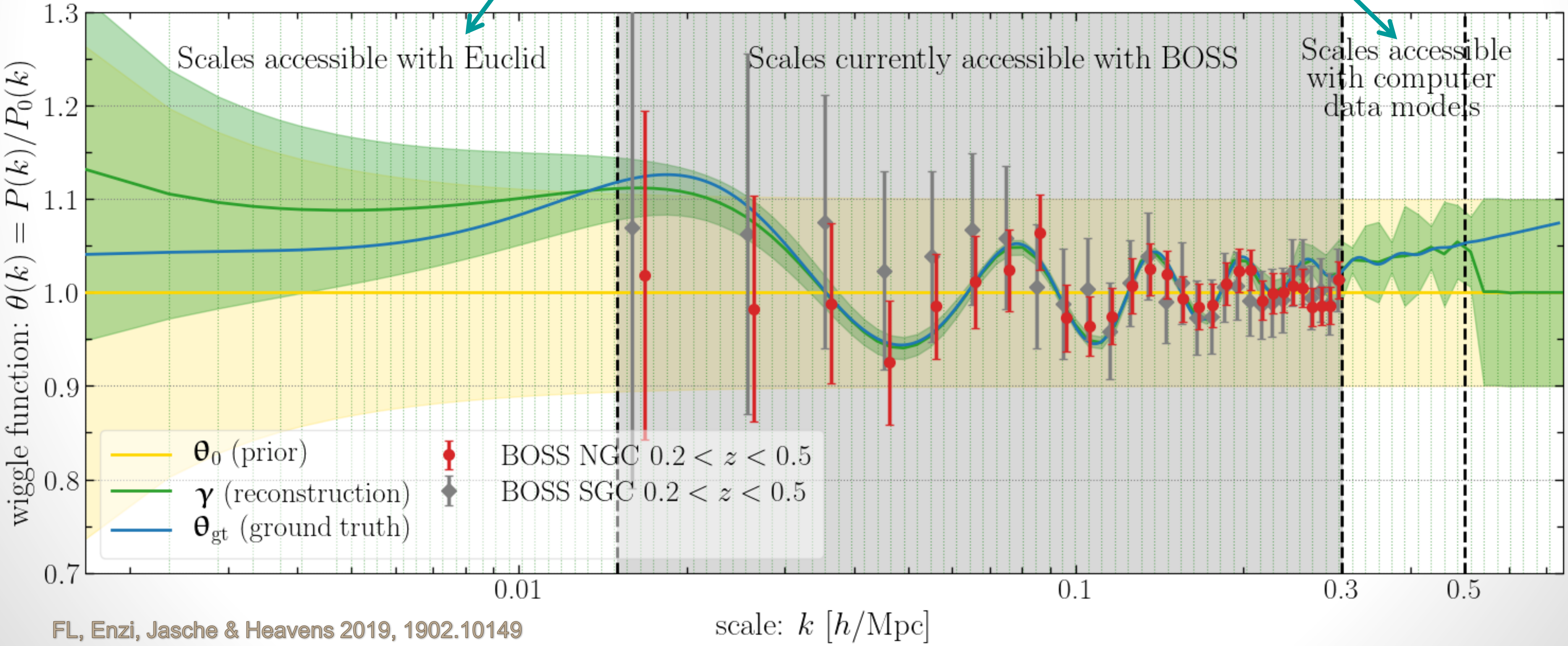
arbitrary data model

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

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



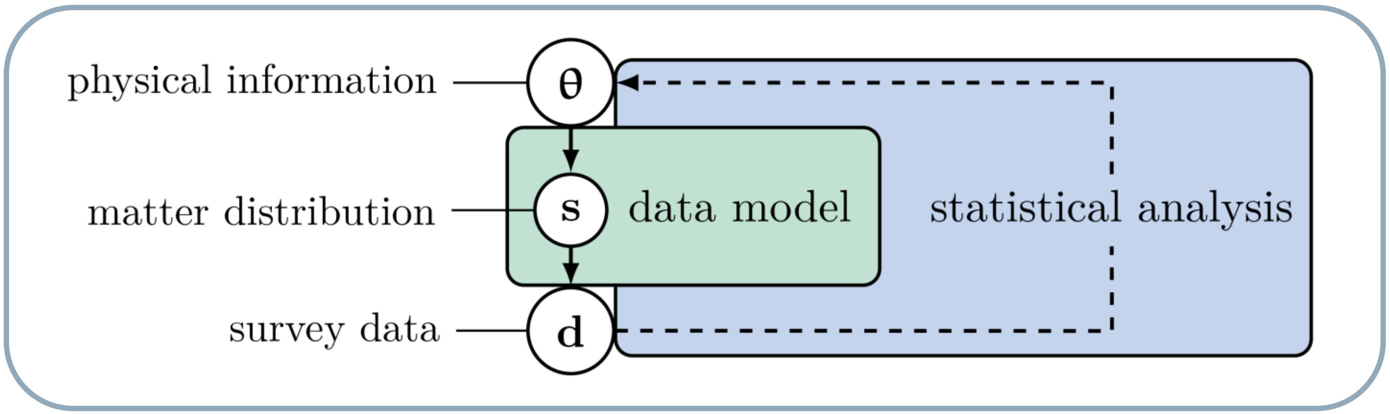

$N_{\mathrm{modes}} \propto k^3$: 5 times more modes are used in the analysis




FL, Enzi, Jasche & Heavens 2019, 1902.10149

# Likelihood-based solution: BORG

*Bayesian Origin Reconstruction from Galaxies*



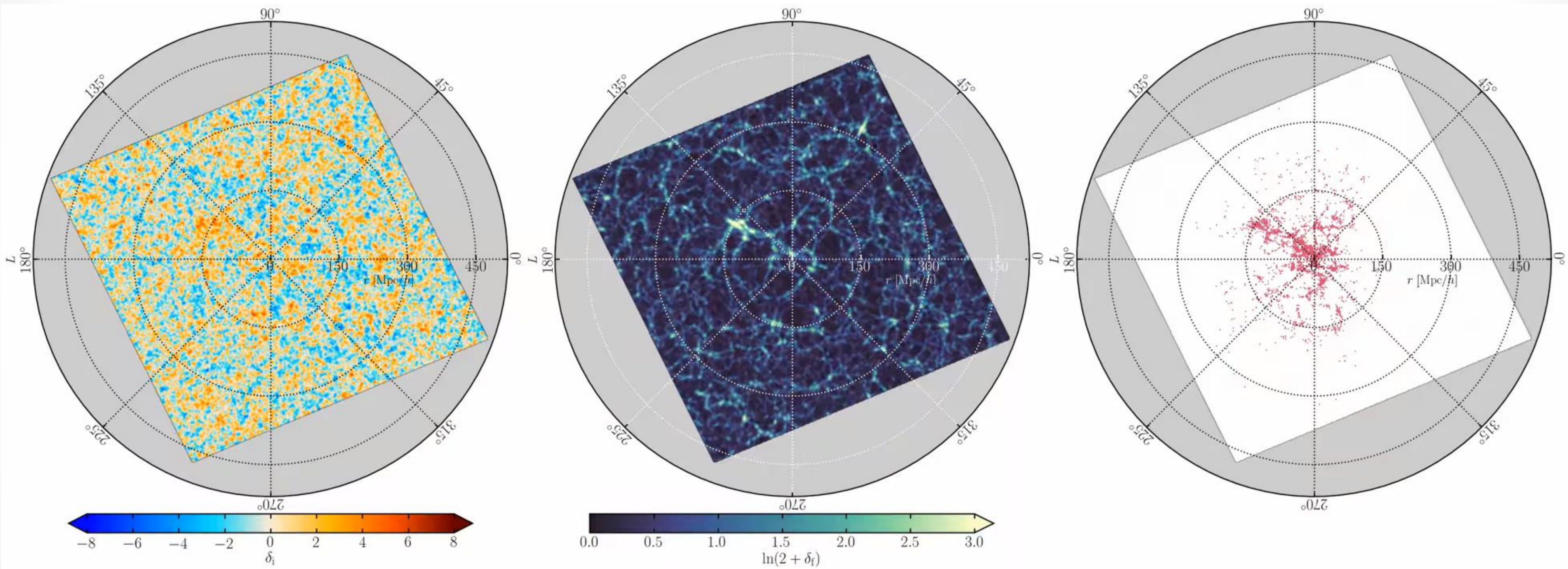

<u>Data assimilation:</u>

exact statistical analysis

approximate data model

# BORG at work: Bayesian chrono-cosmography

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



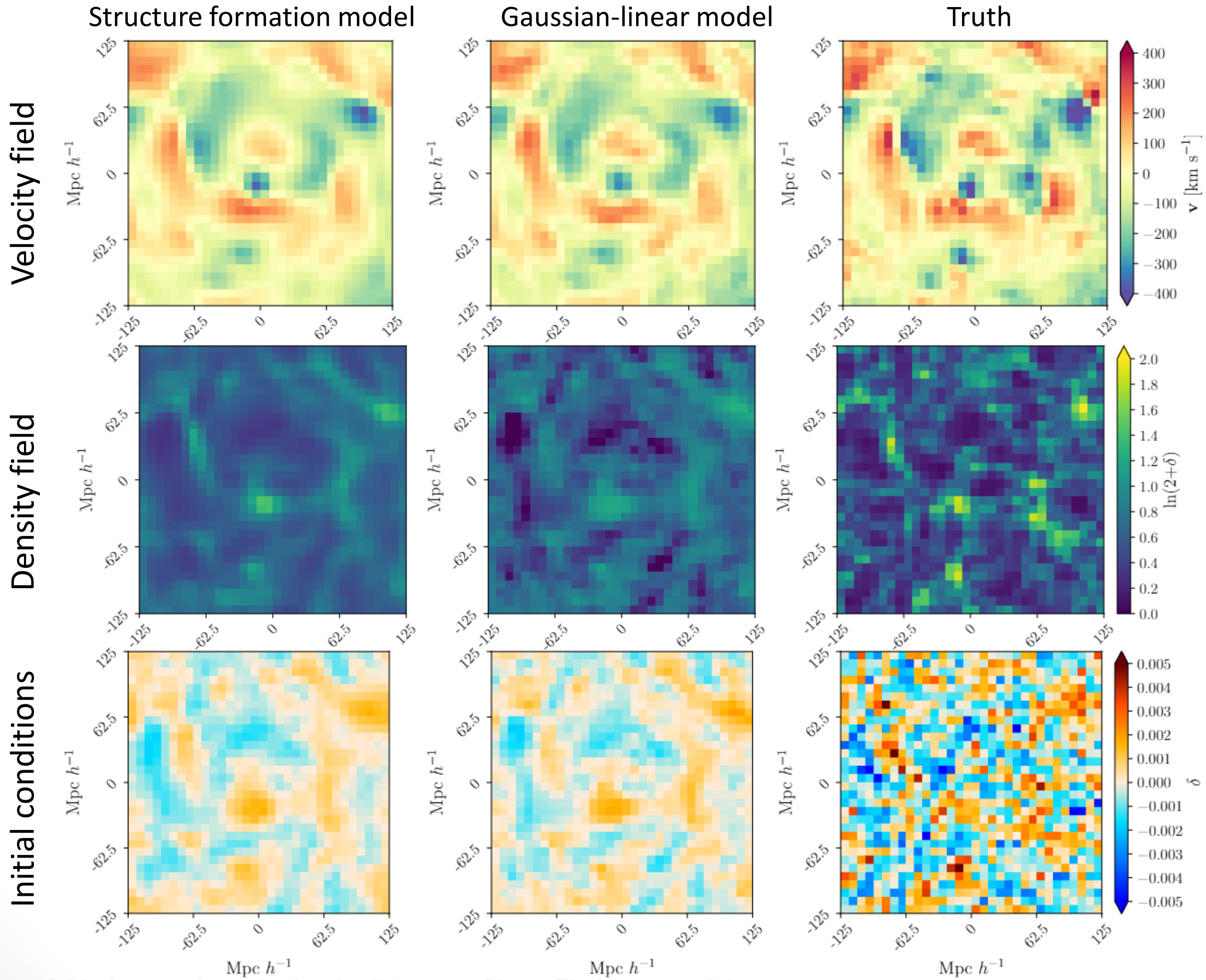

Supergalactic plane

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.

# Reconstructing dark matter with peculiar velocities

- Redshift + distance information ⇒ peculiar velocity information
- Distance tracers can constrain the initial conditions without assumptions on galaxy bias (up to inhomogeneous Malmquist bias).



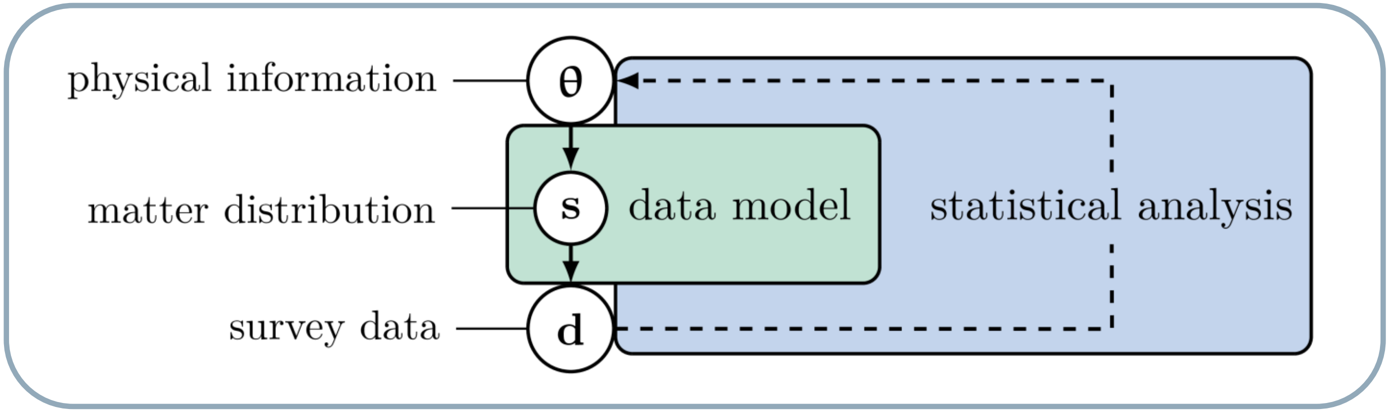

Boruah, Lavaux & Hudson, to be submitted – Prideaux-Ghee, FL, Heavens & Lavaux, in prep.

# Mapping the Universe: epilogue?



J. Cham – PhD comics

# So, which one is the best?




**<u>Data assimilation:</u>**

exact statistical analysis

approximate data model

**<u>Simulation-based inference:</u>**

approximate statistical analysis

arbitrary data model

# Correlation functions versus field-level inference

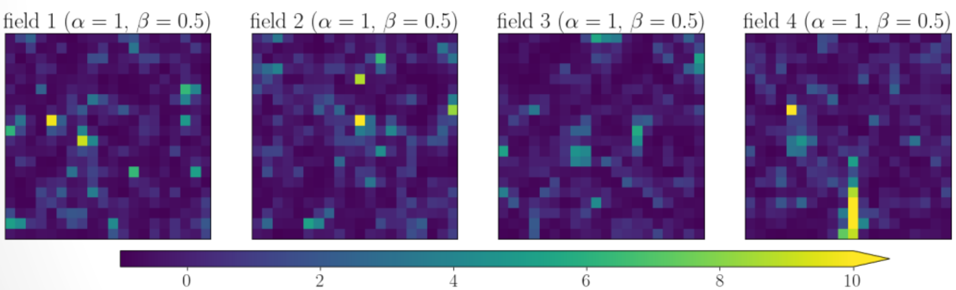
- We checked accuracy and precision of different methods for a log-normal model:

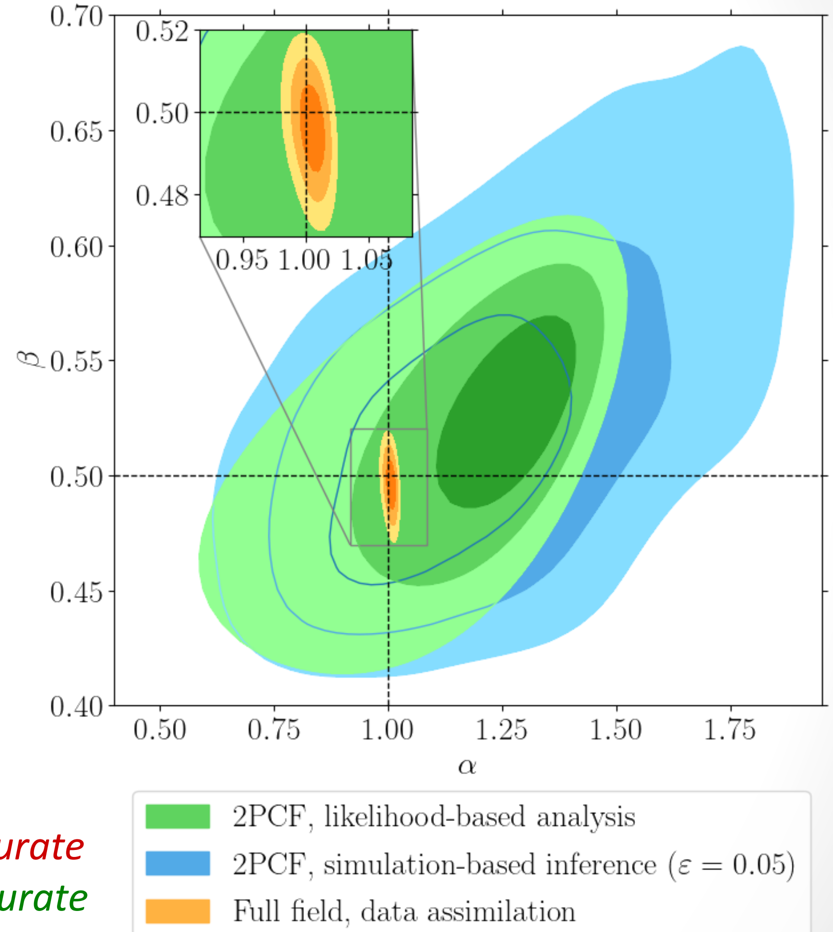
$$f = \frac{1}{\alpha}\left[\exp\left(\alpha g - \frac{1}{2}\alpha^2\right) - 1\right]$$

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*

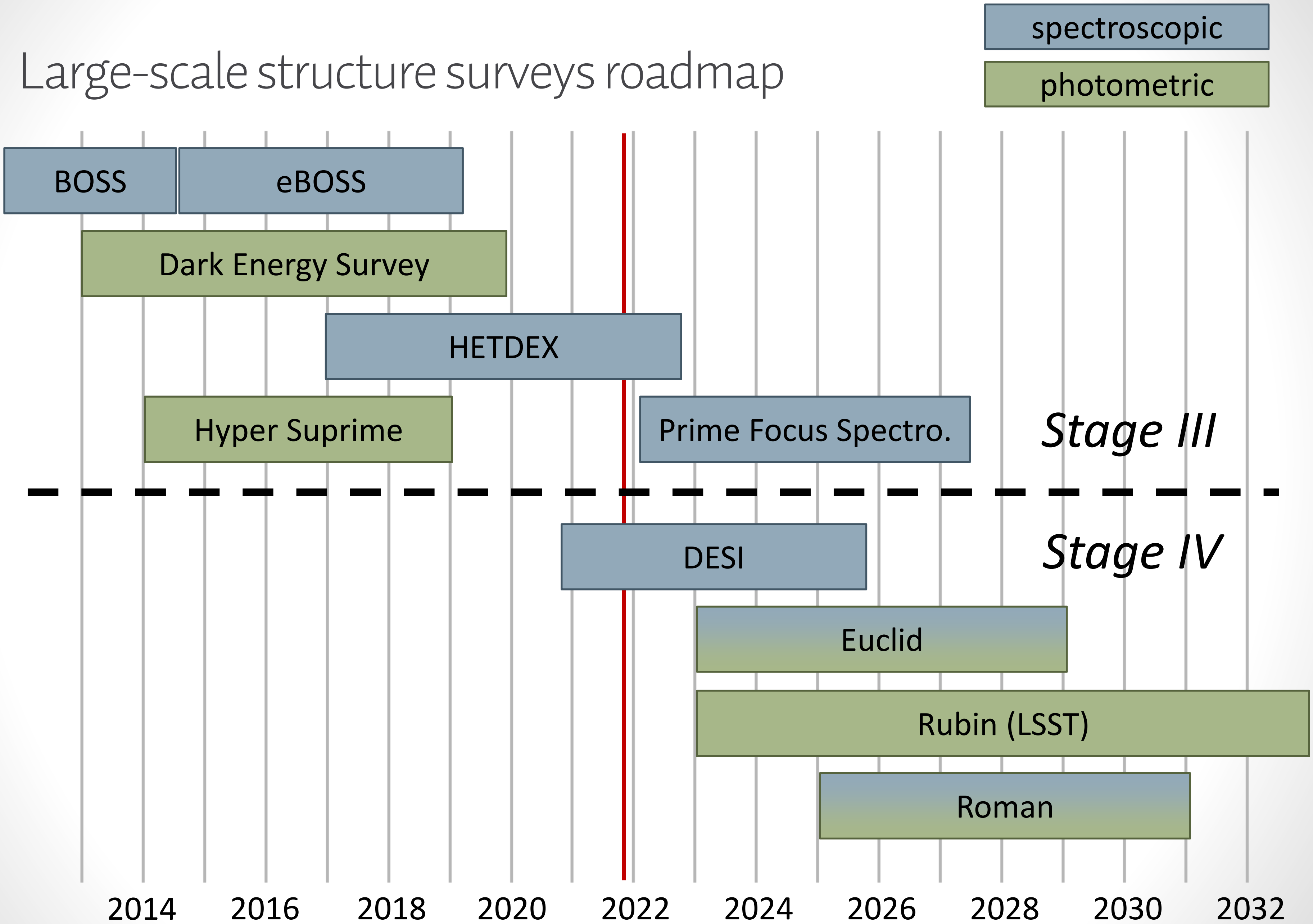
Companion repository:
https://github.com/florent-leclercq/correlations_vs_field

FL & Heavens, 2103.04158

# The Future: Opportunities & Challenges

*DESI, Euclid, Rubin, Roman, and more…*

# Large-scale structure surveys roadmap

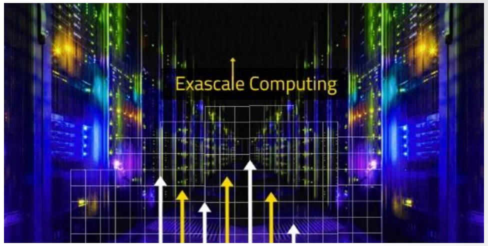


# Data-intensive scientific discovery from galaxy surveys

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

- Can data analysts keep pace?

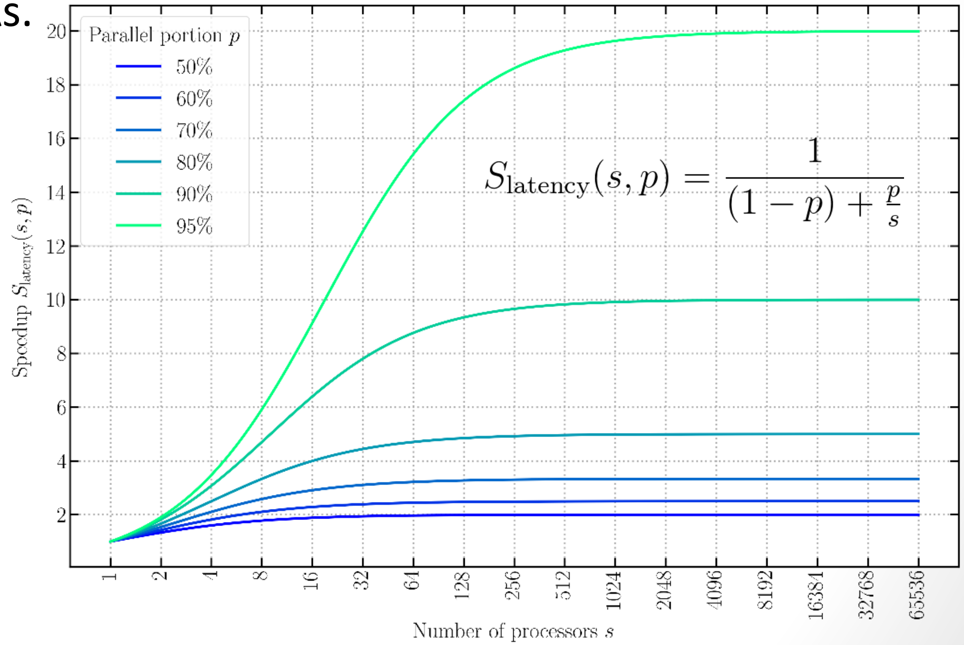
# Numerical data models in the exascale world




- Traditional hardware architectures are reaching their physical limit.
- Current hardware development focuses on:
  - Packing a larger number of cores into each CPU: currently $\mathcal{O}(10^5)$, soon $\mathcal{O}(10^{6-7})$ in systems that are currently being built.
  - Developing hybrid architectures with cores + accelerators: GPUs and reconfigurable chips such as FPGAs.

- Compute cycles are no longer the scarce resource. The cost is driven by interconnections.
- Amdahl's law: latency kills the gains of parallelisation

Amdahl 1967, doi:10.1145/1465482.1465560

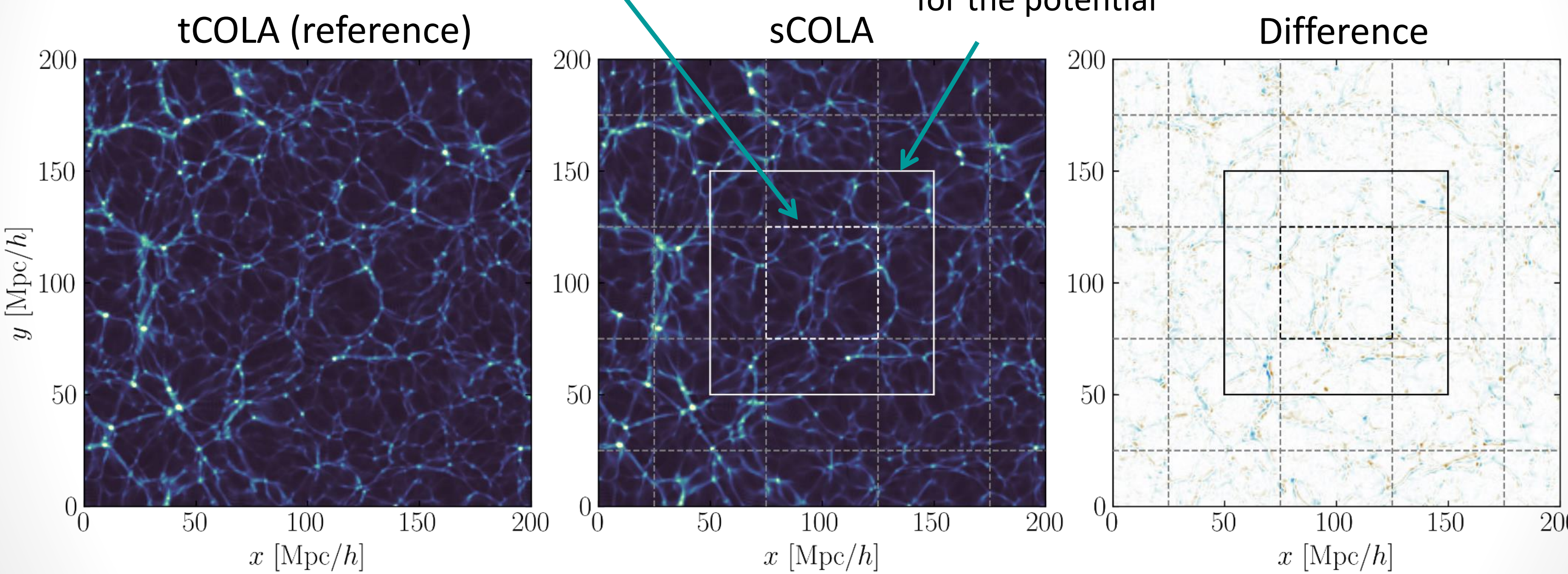



Cosmological simulations cannot merely rely on computers becoming faster to reduce the computational time.

# 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



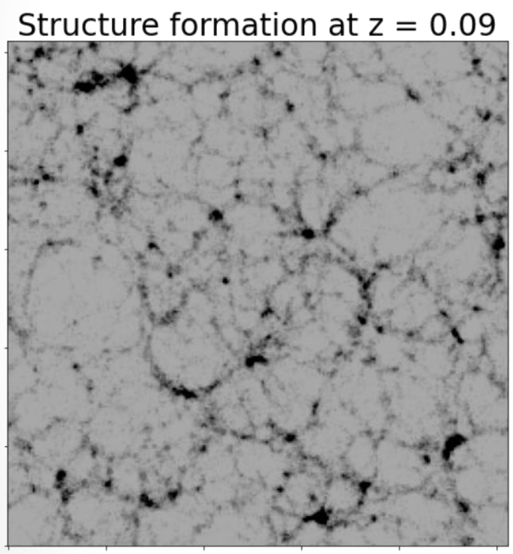

"Computer, enhance!" – John Wise (@AstroAhura) on Twitter

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

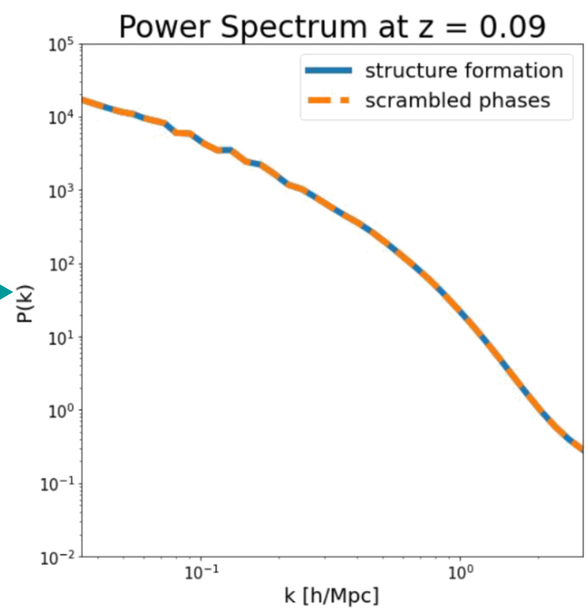
Publicly available implementation:
https://bitbucket.org/florent-leclercq/simbelmyne/

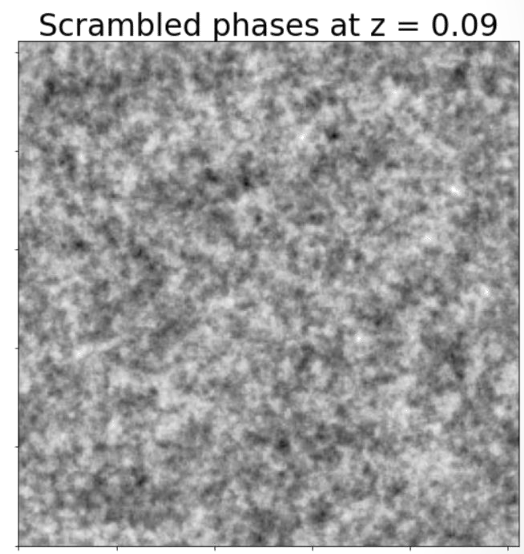
# Associative versus causal reasoning in scientific research

- With traditional machine-learning, we obtain associative links between a latent space and data.
- But this doesn't mean we understand how nature works!




Process 1




Process 2

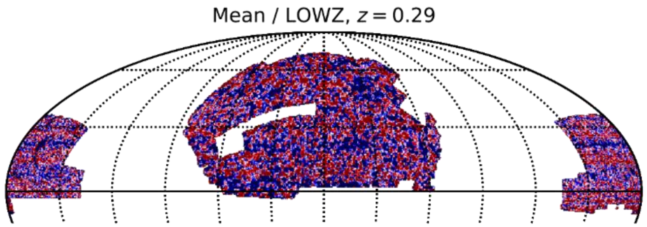
J. Jasche

Purely data-driven machine learning may not be sufficient for research!
(causal explanation, hypothesis generation, discovery)

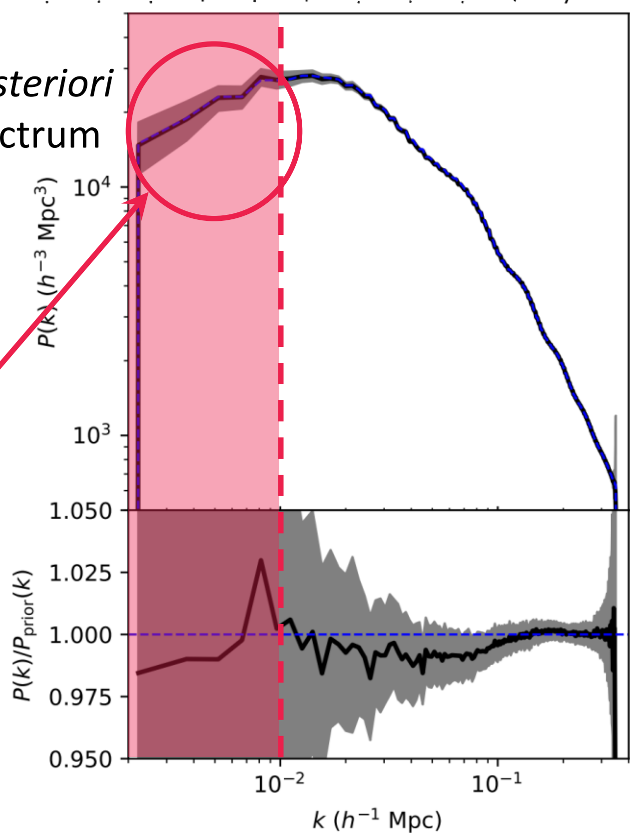
# Machine-aided report of unknown data contaminations

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

Map of unknown foreground contaminant




BORG *a posteriori* power spectrum

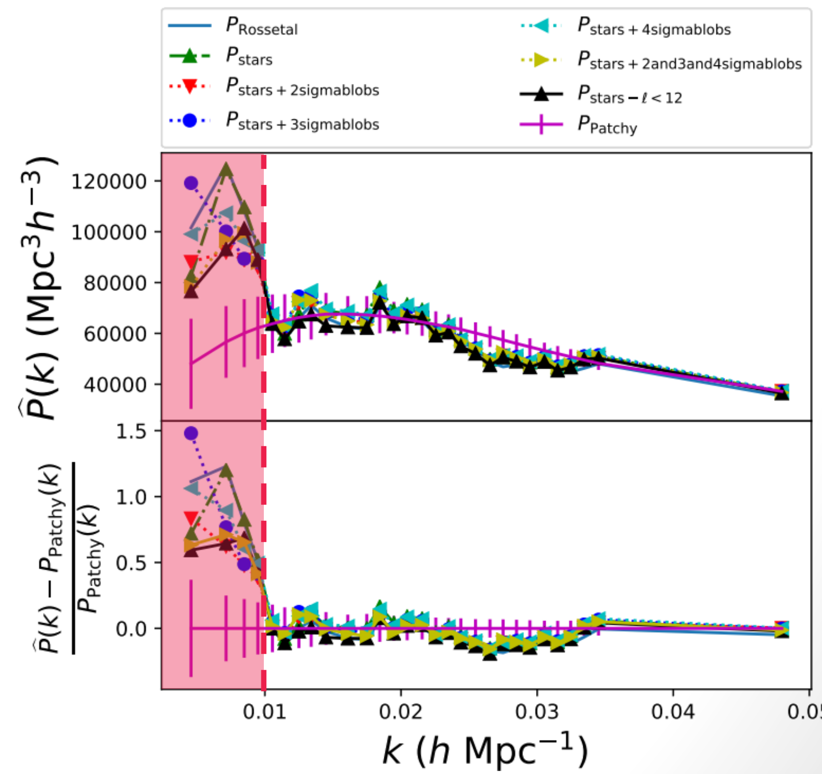



No apparent contamination, even well beyond the turn-over

Porqueres, Ramanah, Jasche & Lavaux 2018, 1812.05113

Lavaux, Jasche & FL 2019, 1909.06396

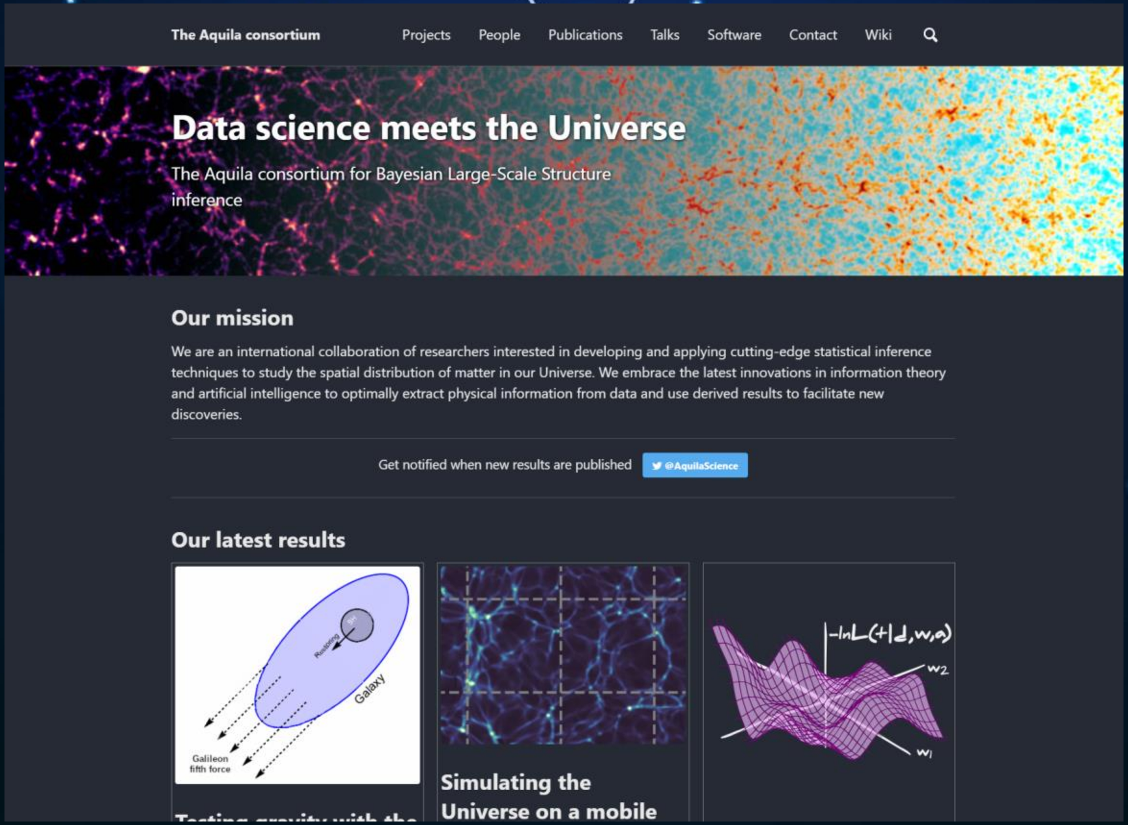
State-of-the-art with backward-modelling technique (mode subtraction)




Kalus, Percival *et al.* 2018, 1806.02789

# The Aquila Consortium

- Created in 2016. Currently 51 members from 16 countries (Europe & Americas).
- 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:**

- exact statistical analysis
- approximate data model

**Simulation-based inference:**

- approximate statistical analysis
- arbitrary data model

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