Bayesian large-scale structure inference and cosmic web analysis

Florent Leclercq

Institut d'Astrophysique de Paris
Institut Lagrange de Paris
École polytechnique ParisTech

April 30th, 2015

In collaboration with:

Nico Hamaus (IAP), Jens Jasche (Excellence Cluster Universe Garching),
Alice Pisani (LAM/IAP), Paul M. Sutter (Trieste/IAP/Ohio State U.),
Benjamin Wandelt (IAP/U. Illinois), Matías Zaldarriaga (IAS Princeton)

How did structure appear in the Universe?

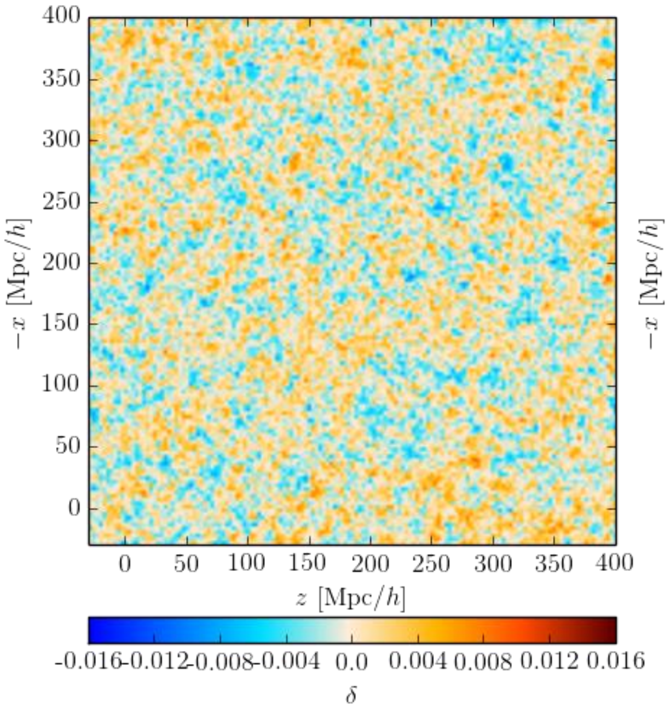
A joint problem!

- **How did the Universe begin?**
  - What are the statistical properties of the initial conditions?
- **How did the large-scale structure take shape?**
  - What is the physics of dark matter and dark energy?
- Usually these problems are addressed in isolation.
- This talk:
  - A case for physical inference of four-dimensional dynamic states
  - A description of methodology and progress towards enriching the standard for analysis of galaxy surveys
  - A round trip: from theory to data, from data to theory

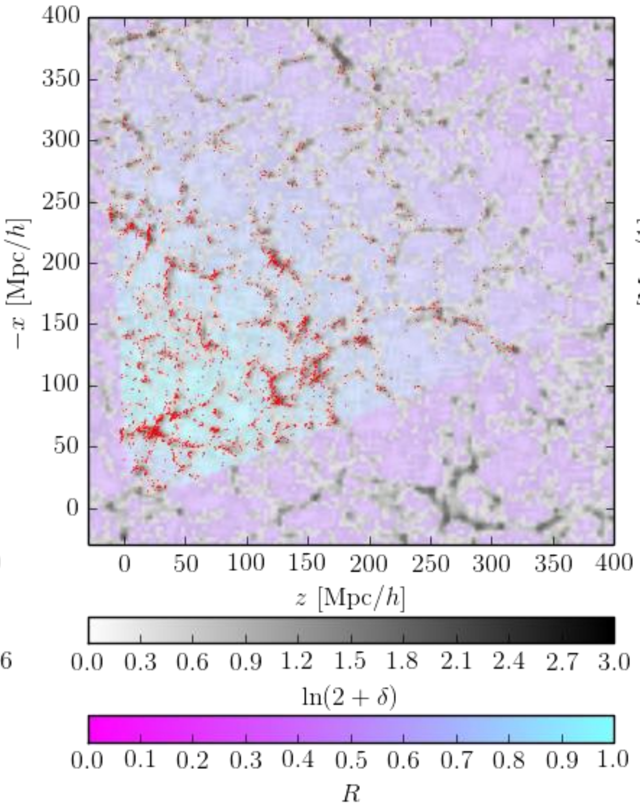
FL, Pisani & Wandelt 2014, arXiv:1403.1260

(Lectures Varenna 2013 and
Paris École Doctorale for Astronomy and Astrophysics)

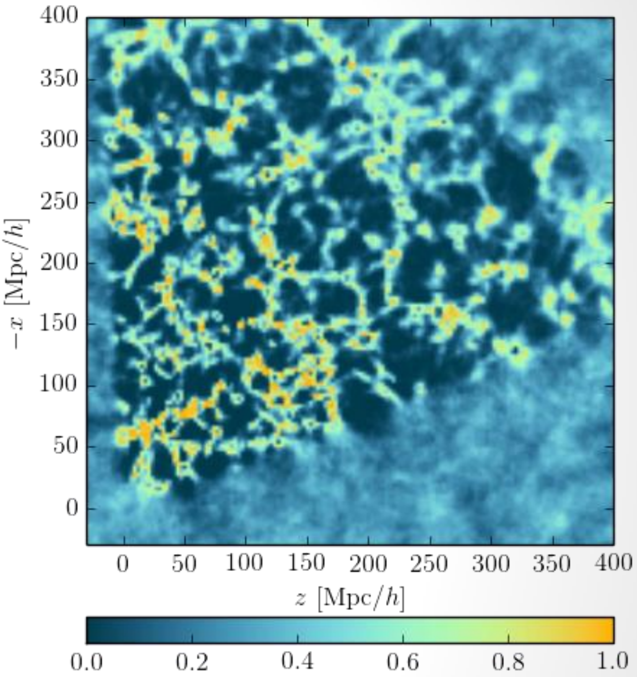
Maps of the large-scale structure



Initial conditions



Final conditions + Observations



Cosmic Web

Outline

1. Bayesian Inference
2. Chrono-Cosmography
3. Non-Linear Filtering
4. Cosmic Web Classification

1. BAYESIAN INFERENCE

- Data assimilation with BORG
- The BORG SDSS run

J. Jasche, B. Wandelt, arXiv:1203.3639.
Bayesian physical reconstruction of initial conditions from large scale structure surveys

J. Jasche, F. Leclercq, B. Wandelt, arXiv:1409.6308.
Past and present cosmic structure in the SDSS DR7 main sample

Why Bayesian inference?

- Why do we need Bayesian inference?

Inference of signals = ill-posed problem

- Incomplete observations: survey geometry, selection effects
- Noise, biases, systematic effects
- Cosmic variance

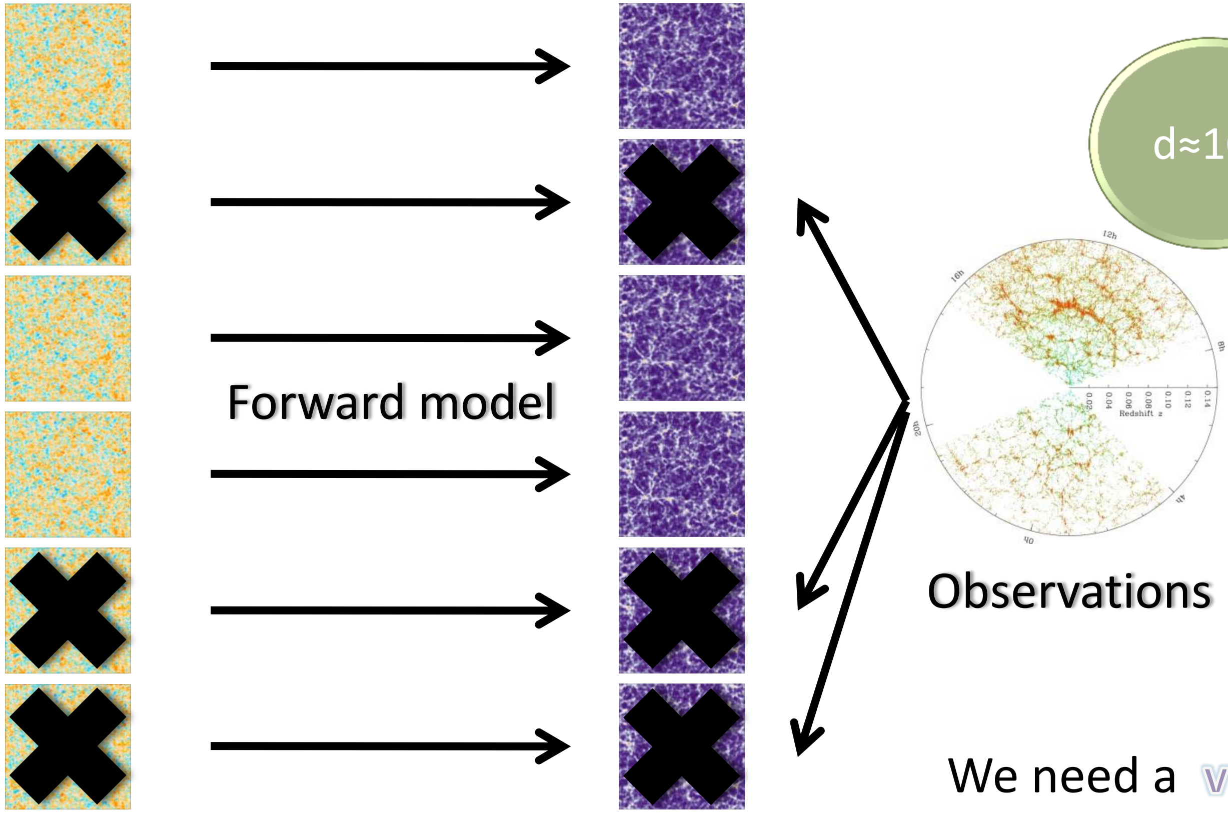
⇨ No unique recovery is possible!

“What is the formation history of the Universe?” ⇨ “What is the probability distribution of possible formation histories (signals) compatible with the observations?”

$$p(s|d)p(d) = p(d|s)p(s)$$

Bayesian forward modeling: the ideal scenario

Forward model = N-body simulation + Halo occupation +
Galaxy formation + Feedback + …



All possible ICs

All possible FCs

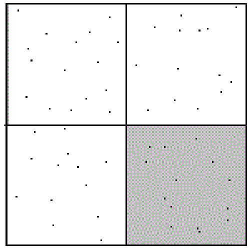
Observations

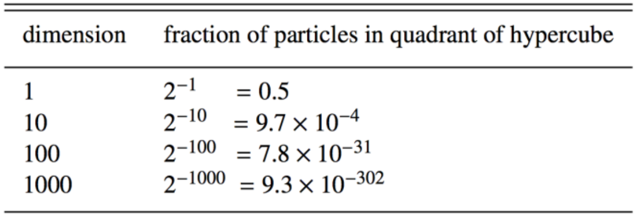
$d \approx 10^7$

We need a very, very, very big computer!

(Parameter) Space: the final frontier

- The "curse of dimensionality" Bellman 1961



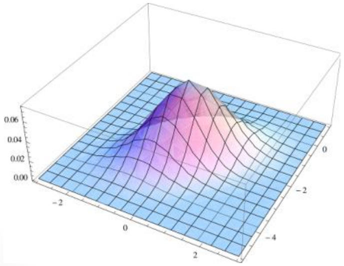


dimension	fraction of particles in quadrant of hypercube
1	$2^{-1} = 0.5$
10	$2^{-10} = 9.7 \times 10^{-4}$
100	$2^{-100} = 7.8 \times 10^{-31}$
1000	$2^{-1000} = 9.3 \times 10^{-302}$

Adding extra dimensions…

  - Exponential increase of the number of particles needed for uniform sampling
  - Exponential increase of sparsity given a fixed amount of particles

- High-dimensional probability distribution functions



⇨ Traditional sampling methods will fail

but gradients carry capital information

Hamiltonian Monte Carlo

- Use classical mechanics to solve statistical problems!
  - The potential: $\psi(\mathbf{x}) \equiv -\ln(\mathcal{P}(\mathbf{x}))$
  - The Hamiltonian: $H \equiv \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + \psi(\mathbf{x})$

$$(\mathbf{x},\mathbf{p}) \Rightarrow \left\{ \begin{array}{l} \dfrac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \dfrac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \\[2ex] \dfrac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\dfrac{\partial H}{\partial \mathbf{x}} = -\dfrac{\partial \psi(\mathbf{x})}{\partial \mathbf{x}} \end{array} \right\} \Rightarrow (\mathbf{x}',\mathbf{p}')$$

gradients

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

acceptance ratio unity

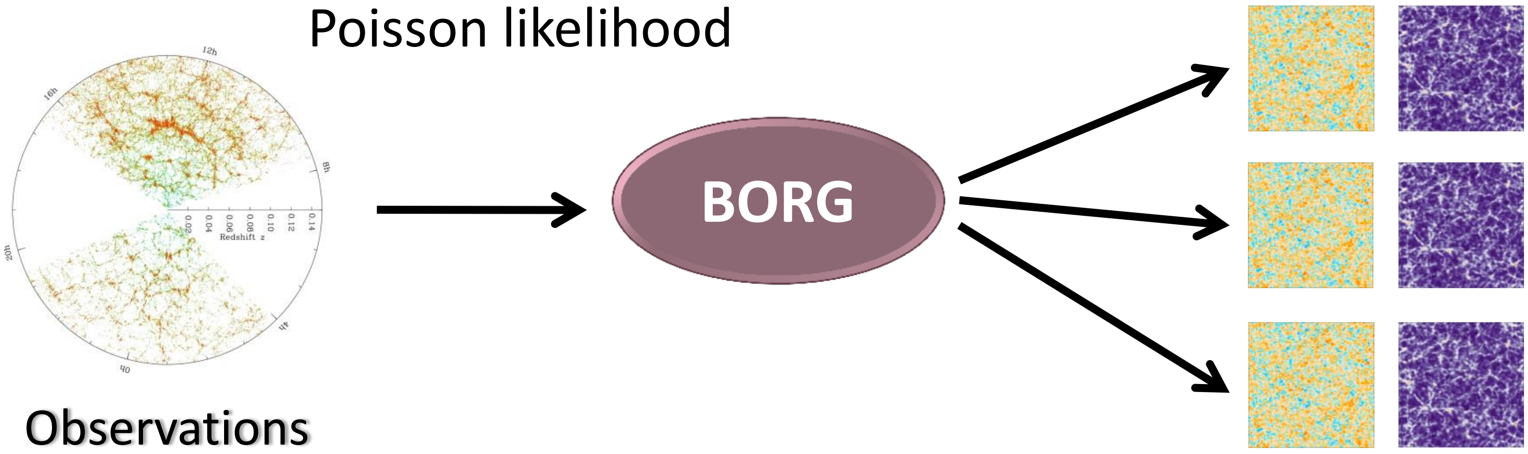
- HMC beats the curse of dimensionality by:
  - Exploiting gradients
  - Using conservation of Hamiltonian

Duane *et al.* 1987

BORG: *Bayesian Origin Reconstruction from Galaxies*

What makes the problem tractable:

- **Sampler**: Hamiltonian Markov Chain Monte Carlo method
- **Physical model**: Gaussian prior – Second-order Lagrangian perturbation theory (2LPT) – Poisson likelihood



Observations

BORG

Samples of possible 4D states

Jasche & Wandelt 2013, arXiv:1203.3639

see also:
Kitaura 2013, arXiv:1203.4184
Wang, Mo, Yang & van den Bosch 2013, arXiv:1301.1348

The BORG SDSS run

- 463,230 galaxies from the NYU-VAGC based on SDSS DR7
- Comoving cubic box of side length 750 Mpc/*h*, with periodic boundary conditions
- 256^3 grid, resolution 3 Mpc/*h* → ≈ 17 millions parameters
- 12,000 samples, four-dimensional maps
- ≈ 3 TB disk space
- 10 months wallclock time on 16-32 cores

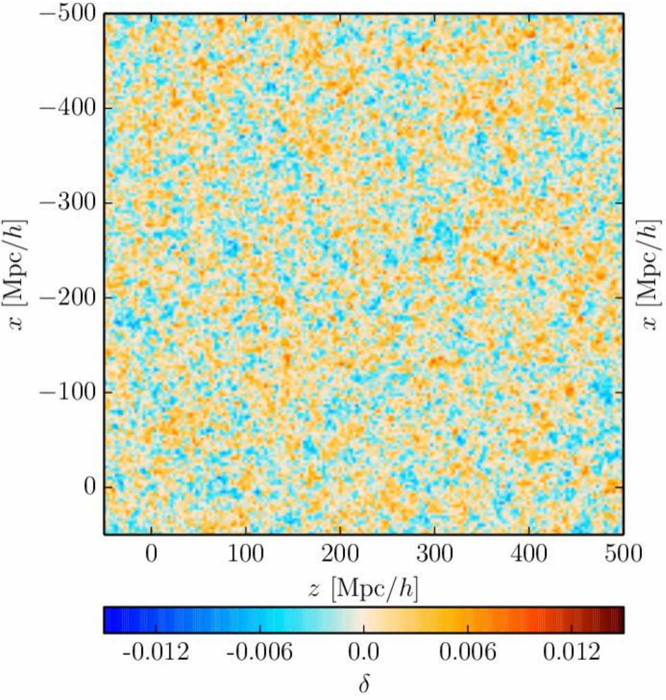
Jasche, FL & Wandelt 2015, arXiv:1409.6308

2. CHRONO-COSMOGRAPHY

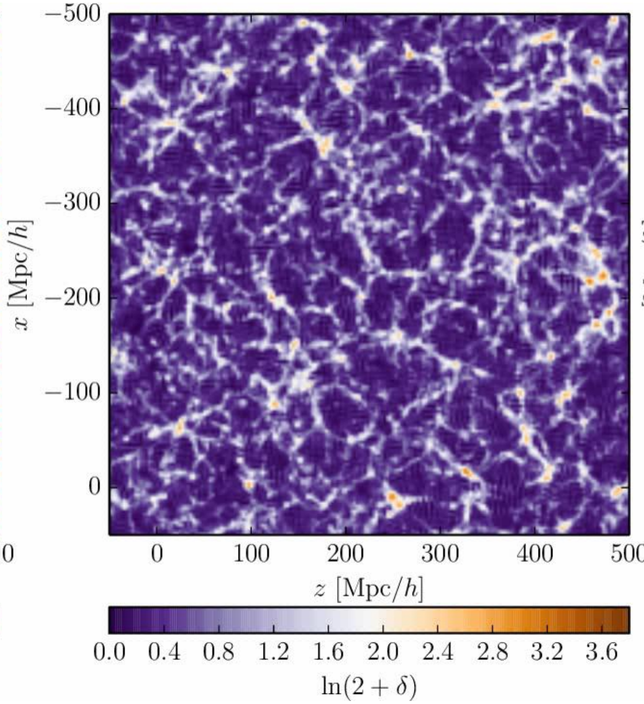
- Past and present cosmic structure in the Sloan volume

J. Jasche, F. Leclercq, B. Wandelt, arXiv:1409.6308.
Past and present cosmic structure in the SDSS DR7 main sample

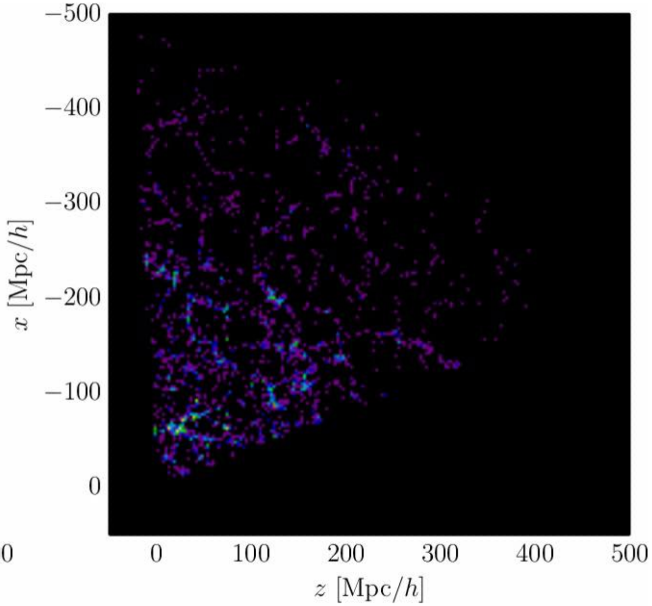
BORG at work – chronocosmography



Initial conditions



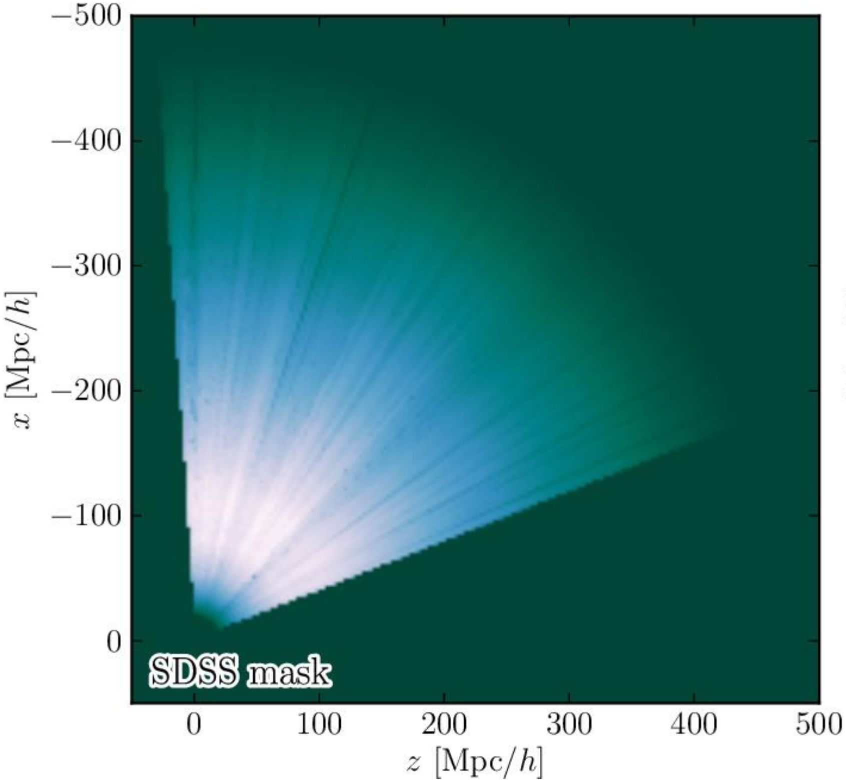
Final conditions

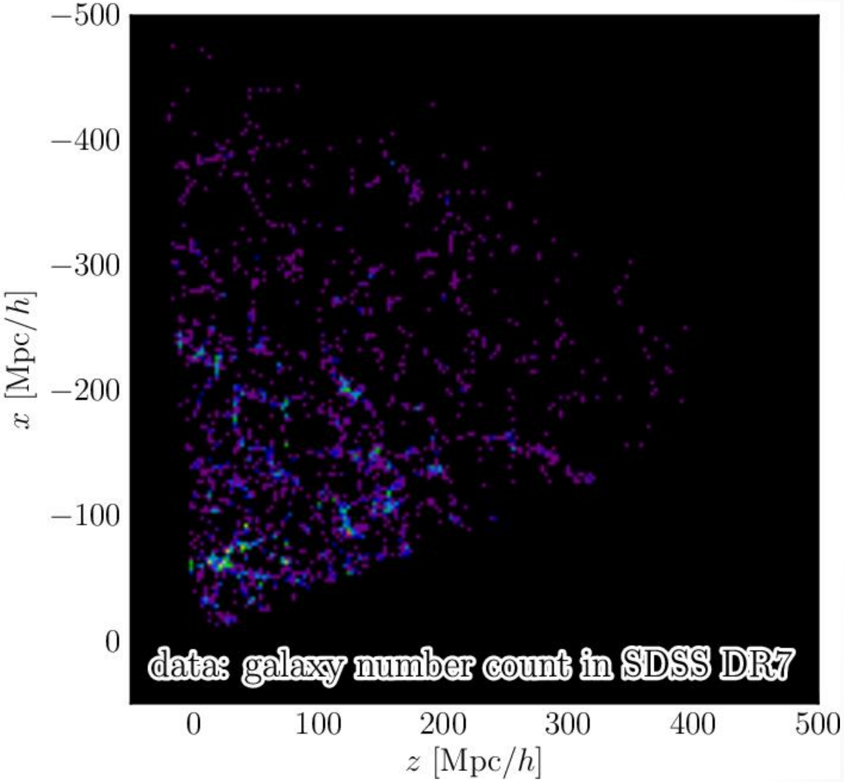


Observations

Jasche, FL & Wandelt 2015, arXiv:1409.6308

Bayesian chronocosmography from SDSS DR7

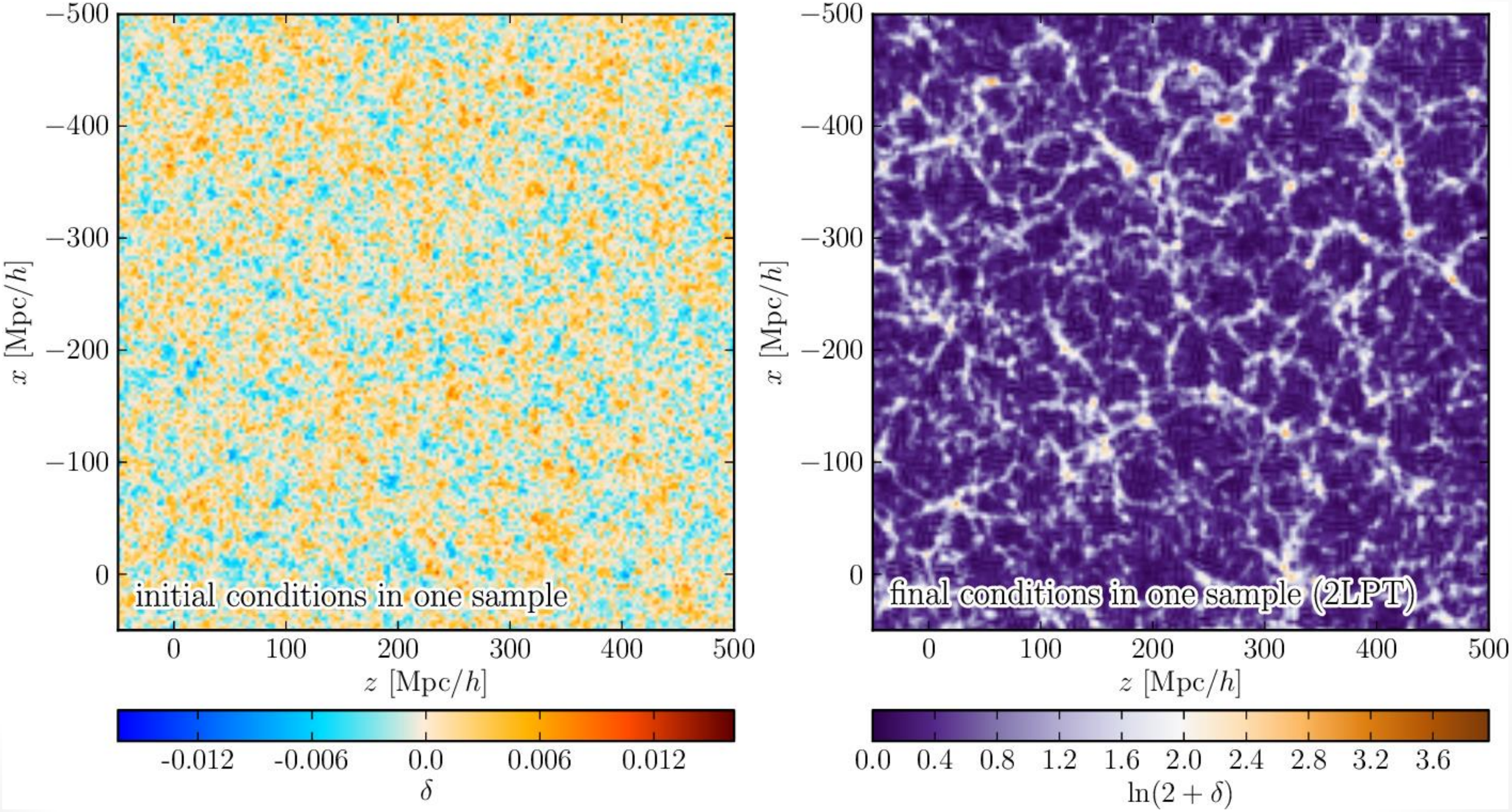




Jasche, FL & Wandelt 2015, arXiv:1409.6308

Data

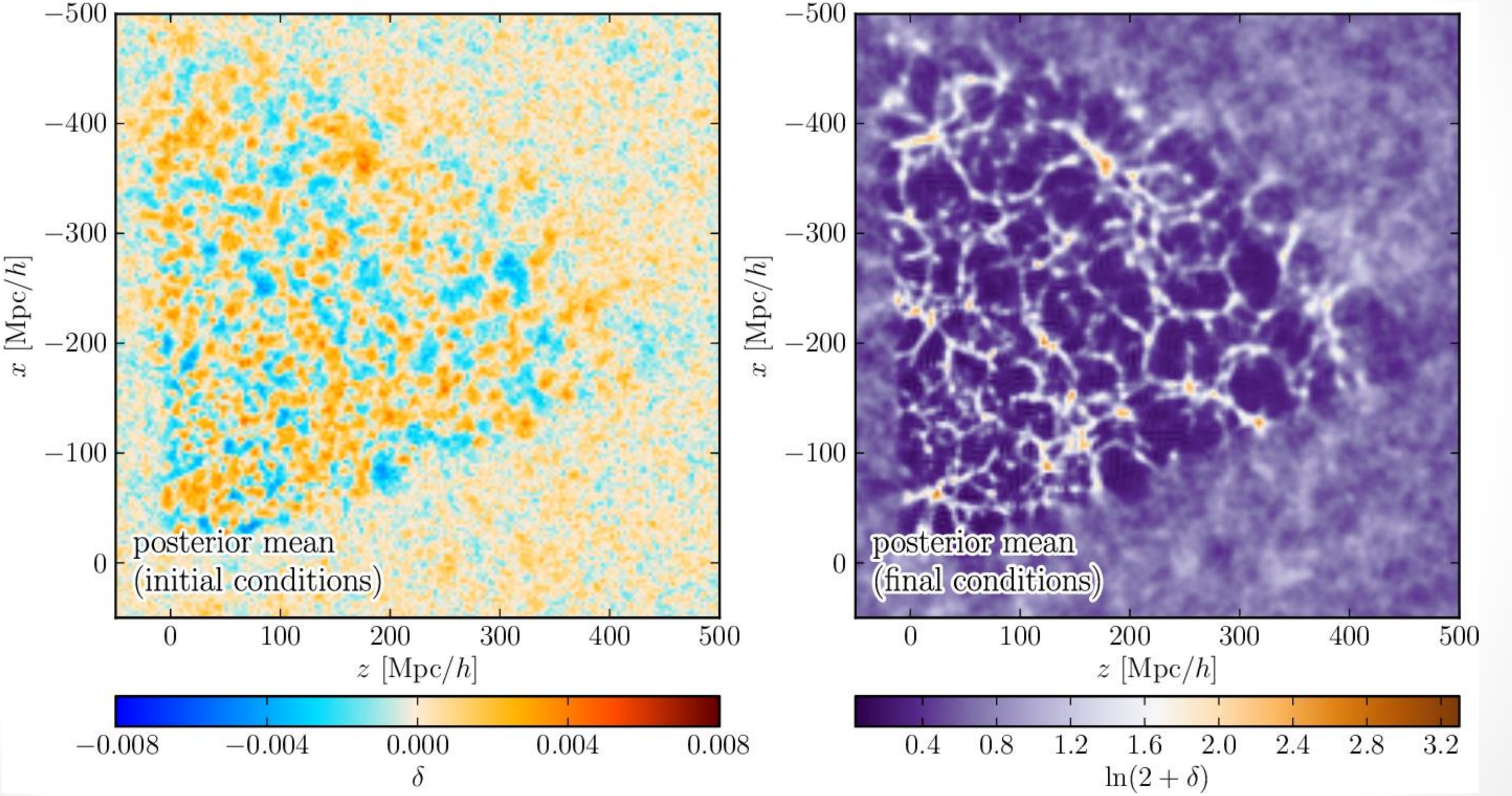
Bayesian chronocosmography from SDSS DR7



Jasche, FL & Wandelt 2015, arXiv:1409.6308

One sample

Bayesian chronocosmography from SDSS DR7



Posterior mean

Jasche, FL & Wandelt 2015, arXiv:1409.6308

3. NON-LINEAR FILTERING

- Non-linear filtering of BORG results
- The COLA method

NL Filtering

F. Leclercq, J. Jasche, P. M. Sutter, N. Hamaus, B. Wandelt, arXiv:1410.0355.
Dark matter voids in the SDSS galaxy survey

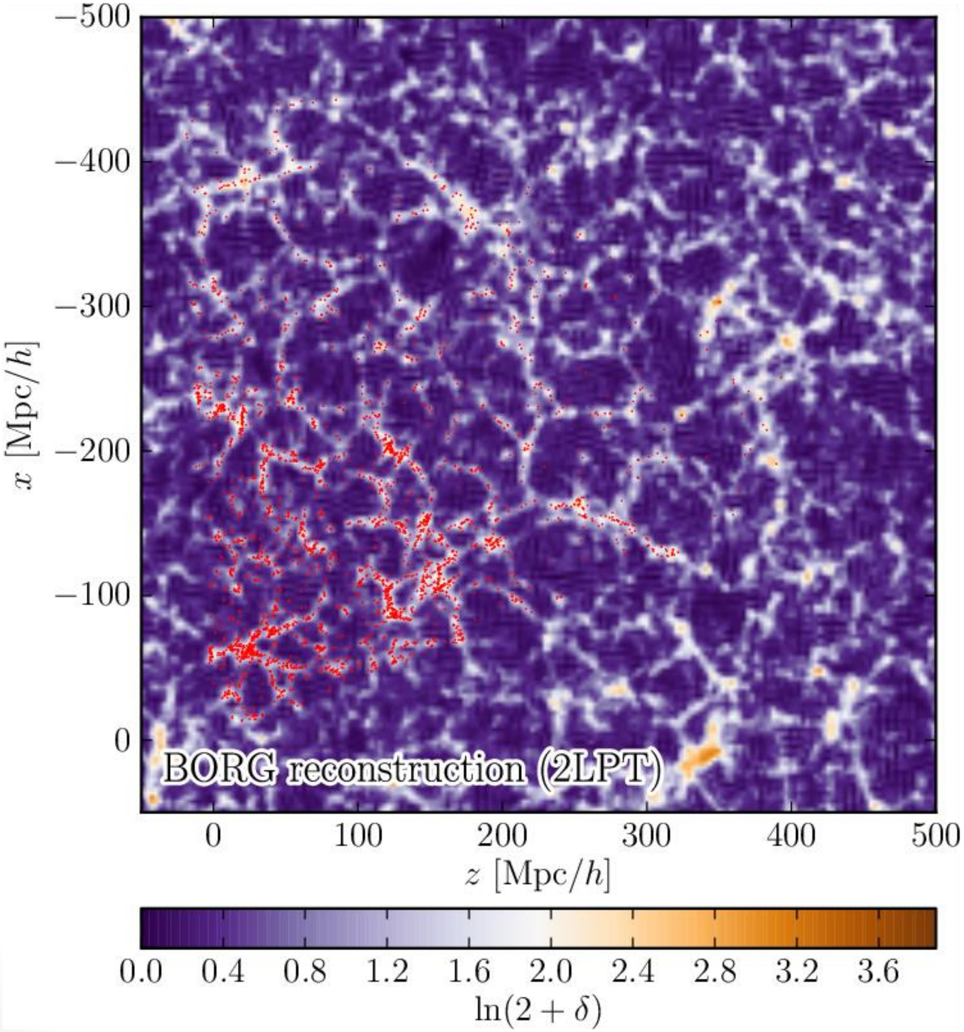
F. Leclercq, J. Jasche, B. Wandelt, arXiv:1502.02690.
Bayesian analysis of the dynamic cosmic web in the SDSS galaxy survey

COLA

S. Tassev, M. Zaldarriaga, D. Eisenstein, arXiv:1301.0322.
Solving Large Scale Structure in Ten Easy Steps with COLA

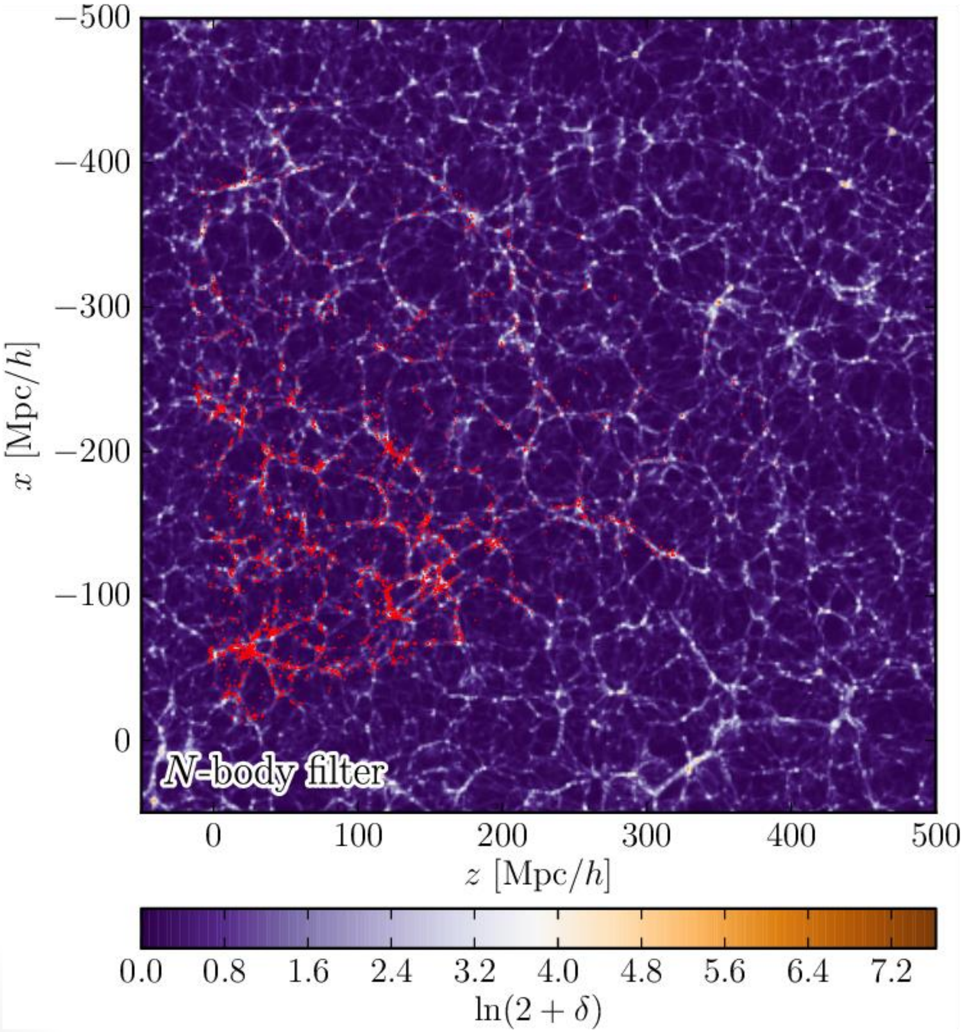
S. Tassev, D. Eisenstein, B. Wandelt, M. Zaldarriaga, arXiv:1502.07751; F. Leclercq, B. Wandelt, M. Zaldarriaga *et al.*, in prep.
sCOLA: The N-body COLA method extended to the Spatial Domain

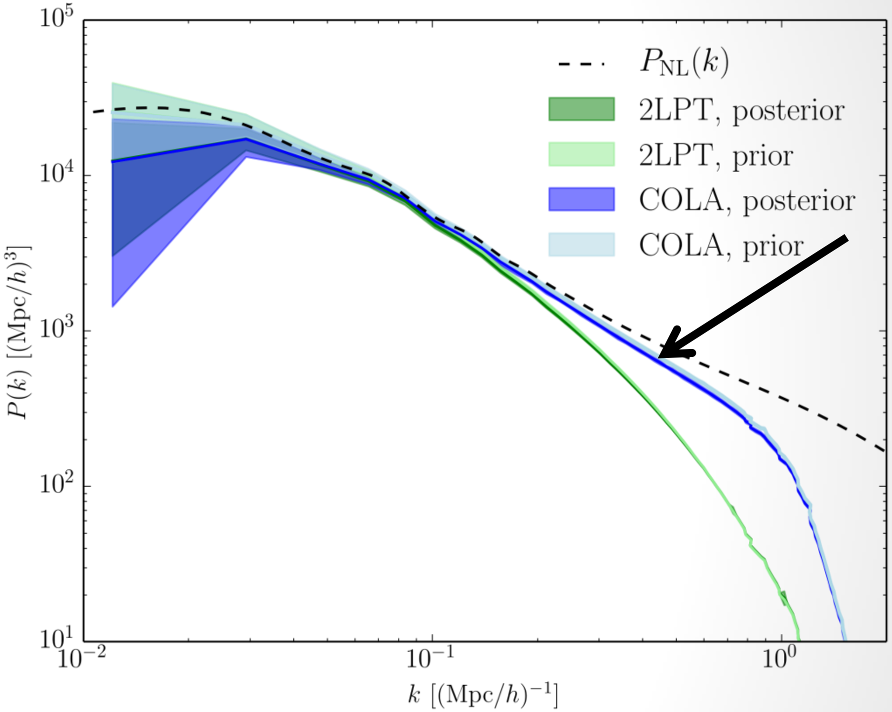
Non-linear filtering



FL, Jasche, Sutter, Hamaus & Wandelt 2015, arXiv:1410.0355

Non-linear filtering





FL, Jasche, Sutter, Hamaus & Wandelt 2015, arXiv:1410.0355

COLA: *COmoving Lagrangian Acceleration*

- Write the displacement vector as: $\mathbf{s} = \mathbf{s}_{\mathrm{LPT}} + \mathbf{s}_{\mathrm{MC}}$

Tassev & Zaldarriaga 2012, arXiv:1203.5785

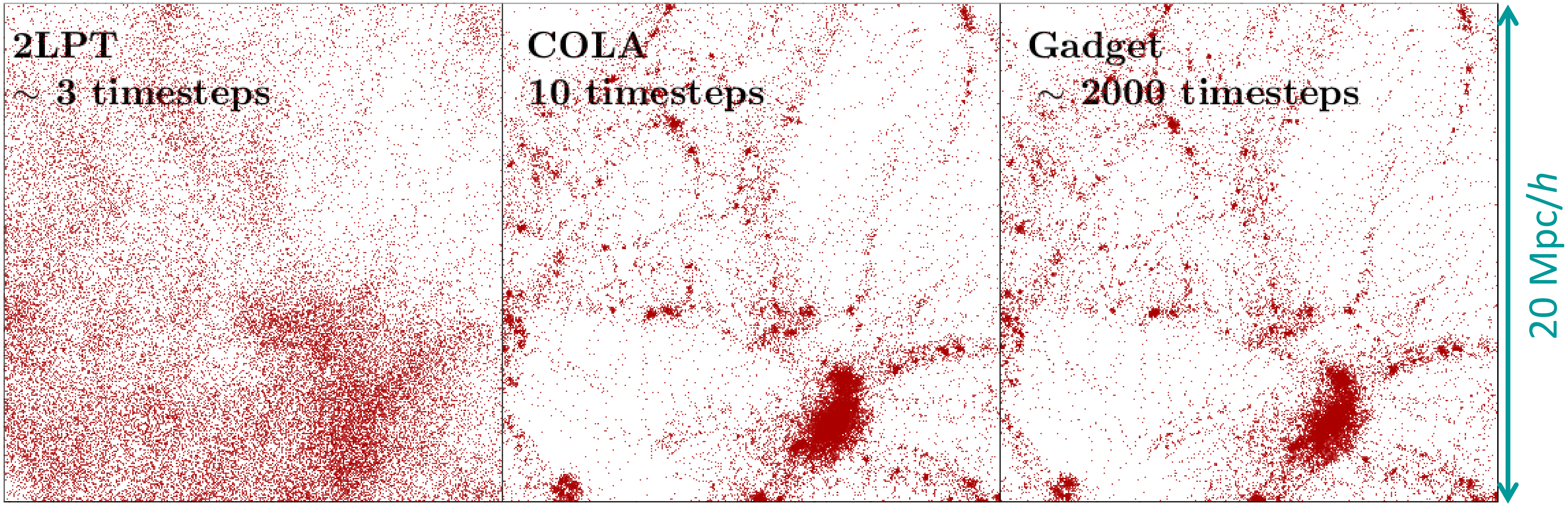
- Time-stepping (omitted constants and Hubble expansion):

Standard:

$$\partial_\tau^2 \mathbf{s} = -\nabla\Phi$$

Modified:

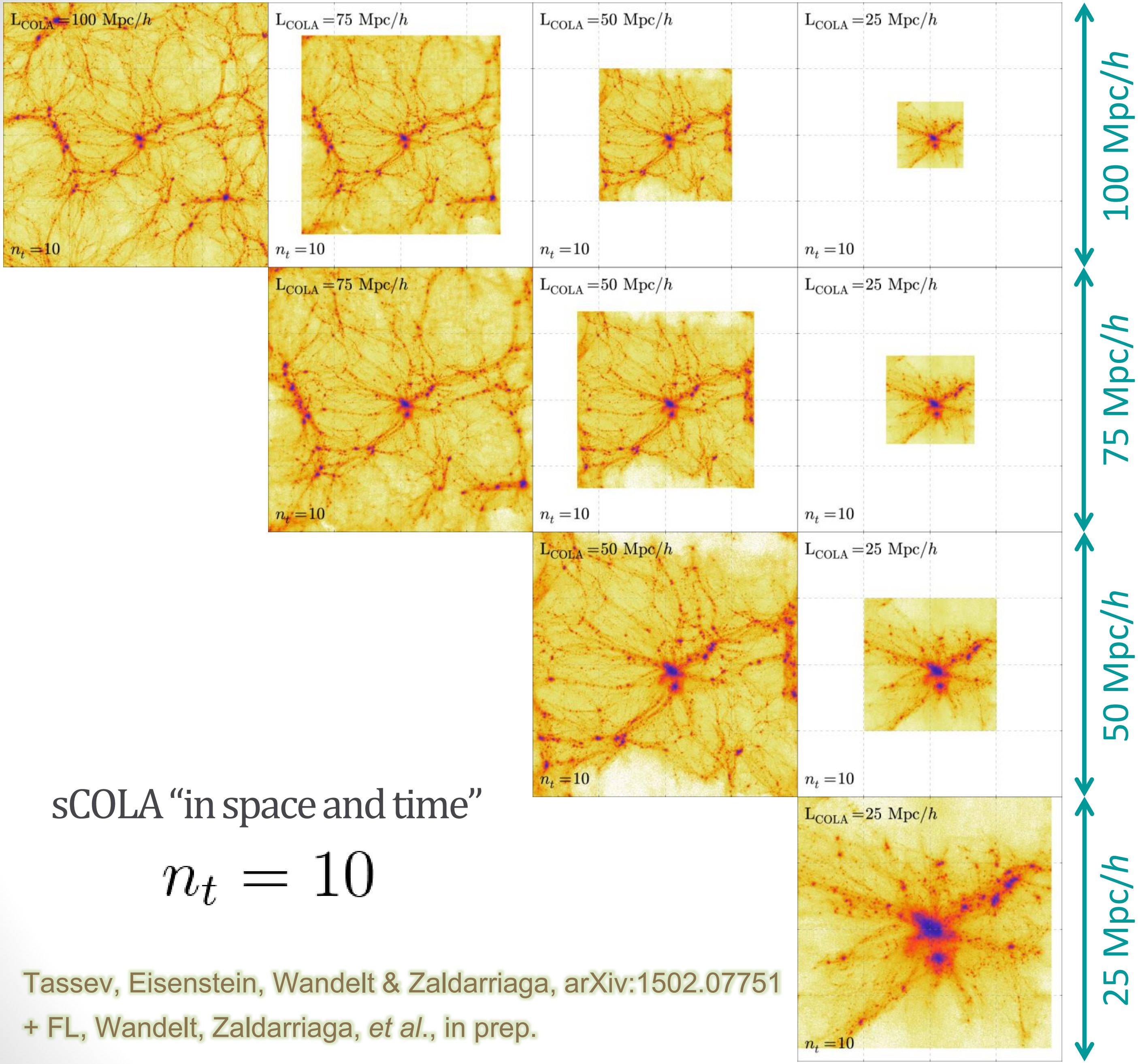
$$\partial_\tau^2 \mathbf{s}_{\mathrm{MC}} = \partial_\tau^2 (\mathbf{s} - \mathbf{s}_{\mathrm{LPT}}) = -\nabla\Phi - \partial_\tau^2 \mathbf{s}_{\mathrm{LPT}}$$



Original COLA "in time"

Tassev, Zaldarriaga & Einsenstein 2013, arXiv:1301.0322

Extending COLA



sCOLA "in space and time"

$$n_t = 10$$

Tassev, Eisenstein, Wandelt & Zaldarriaga, arXiv:1502.07751
+ FL, Wandelt, Zaldarriaga, *et al.*, in prep.

4. COSMIC WEB CLASSIFICATION

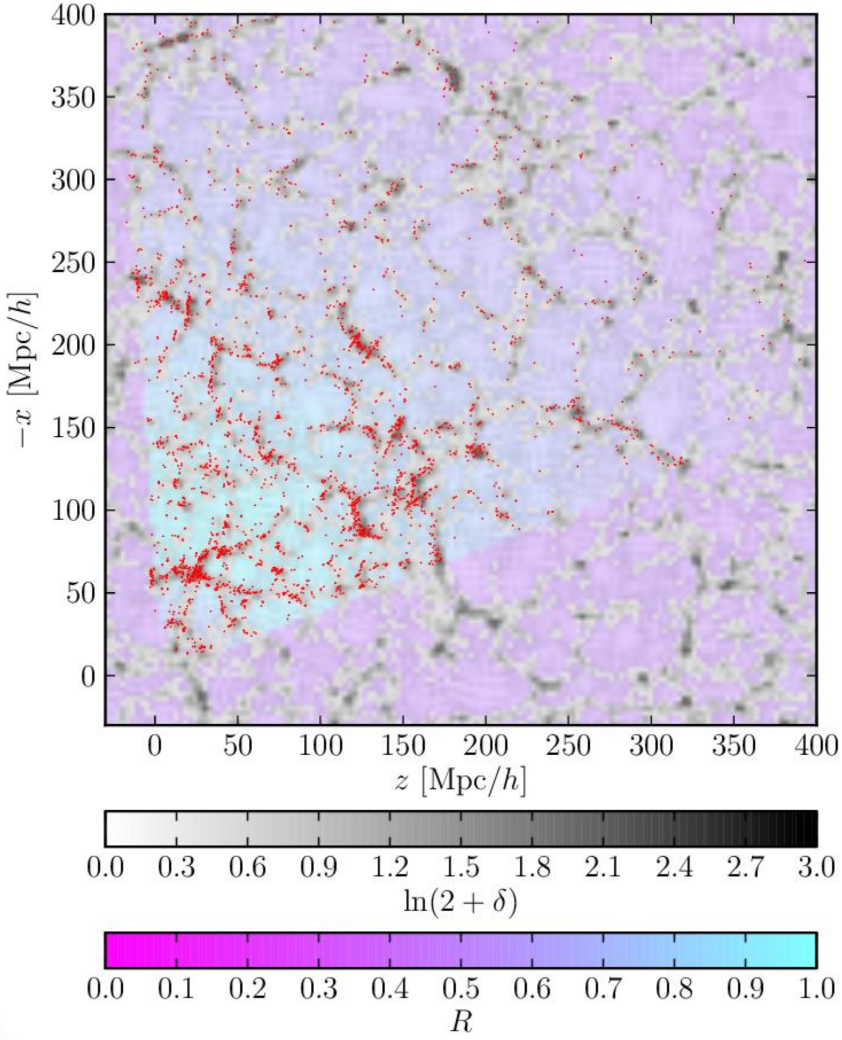
- Dark matter voids in the SDSS
- Tidal shear analysis in the SDSS, dynamic structure type classification
- Cosmic-web classification and Bayesian decision theory

F. Leclercq, J. Jasche, P. M. Sutter, N. Hamaus, B. Wandelt, arXiv:1410.0355.
Dark matter voids in the SDSS galaxy survey

F. Leclercq, J. Jasche, B. Wandelt, arXiv:1502.02690.
Bayesian analysis of the dynamic cosmic web in the SDSS galaxy survey

F. Leclercq, J. Jasche, B. Wandelt, arXiv:1503.00730.
Cosmic web-type classification using decision theory

Dark matter voids in the SDSS



FL, Jasche, Sutter, Hamaus & Wandelt 2015, arXiv:1410.0355

- Why?

Sparsity & Bias

Sutter *et al*. 2014, arXiv:1309.5087

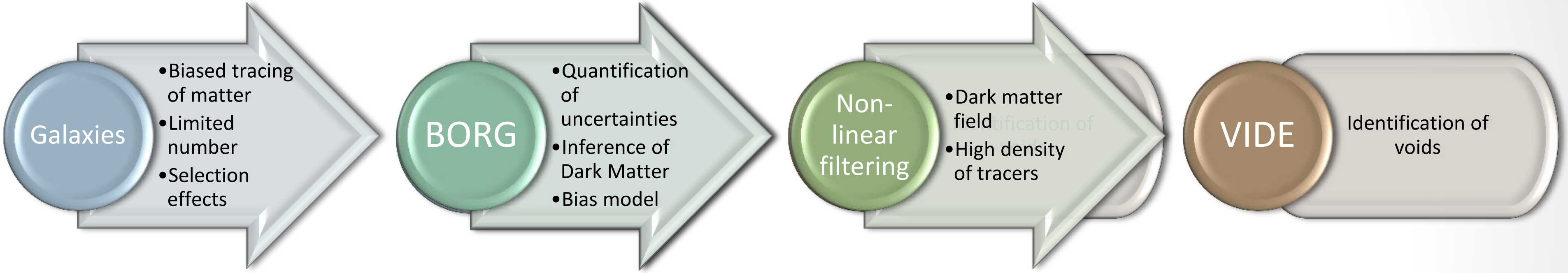
Sutter *et al*. 2014, arXiv:1311.3301

- How?

VIDE toolkit: Sutter *et al*. 2015, arXiv:1406.1191
www.cosmicvoids.net

based on ZOBOV: Neyrinck 2008, arXiv:0712.3049

Dark matter voids: pipeline



Galaxies

- Biased tracing of matter
- Limited number
- Selection effects

BORG

- Quantification of uncertainties
- Inference of Dark Matter
- Bias model

Non-linear filtering

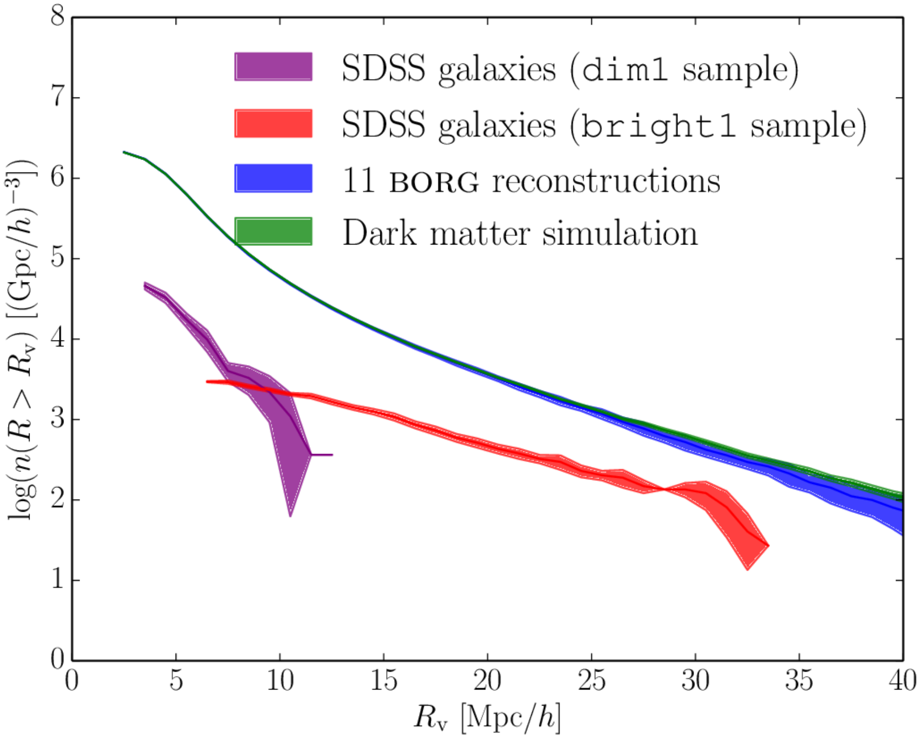
- Dark matter field
- High density of tracers

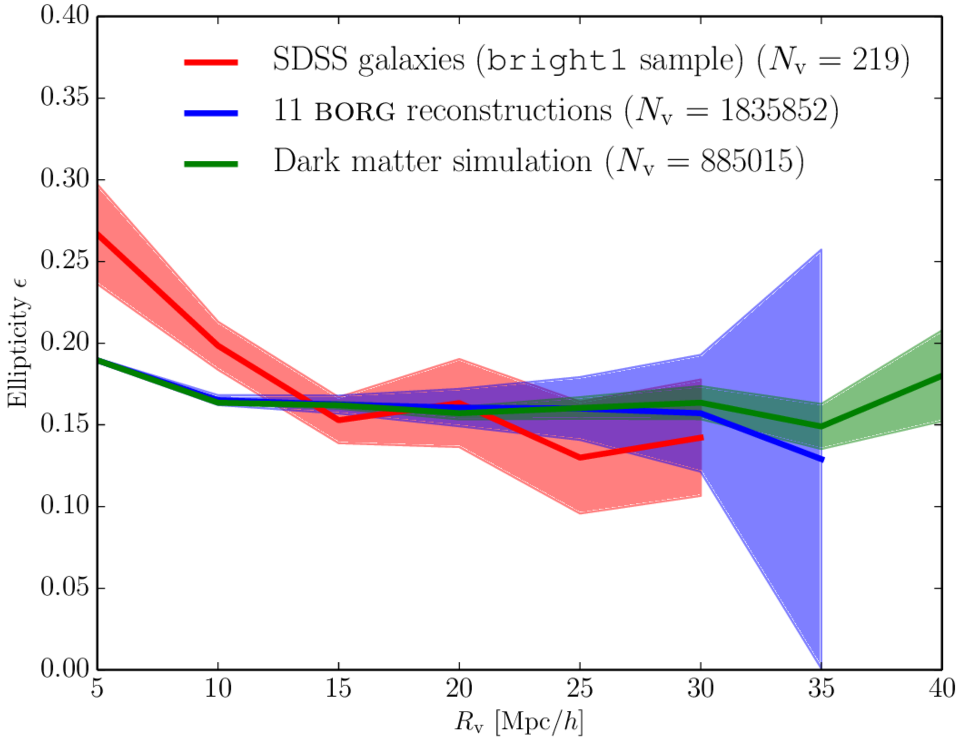
VIDE

Identification of voids

FL, Jasche, Sutter, Hamaus & Wandelt 2015, arXiv:1410.0355

Dark matter void properties





All catalogs are publicly available at www.cosmicvoids.net

FL, Jasche, Sutter, Hamaus & Wandelt 2015, arXiv:1410.0355

Tidal shear analysis

- $\lambda_1, \lambda_2, \lambda_3$: eigenvalues of the tidal field tensor, the Hessian of the gravitational potential: $T_{ij} = \partial_i \partial_j \Phi$ $\lambda_1 + \lambda_2 + \lambda_3 = \delta$
  - Voids: $\lambda_1, \lambda_2, \lambda_3 < 0$
  - Sheets: $\lambda_1 > 0 \text{ and } \lambda_2, \lambda_3 < 0$
  - Filaments: $\lambda_1, \lambda_2 > 0 \text{ and } \lambda_3 < 0$
  - Clusters: $\lambda_1, \lambda_2, \lambda_3 > 0$

Hahn *et al.* 2007, arXiv:astro-ph/0610280

see also:

- Extensions:

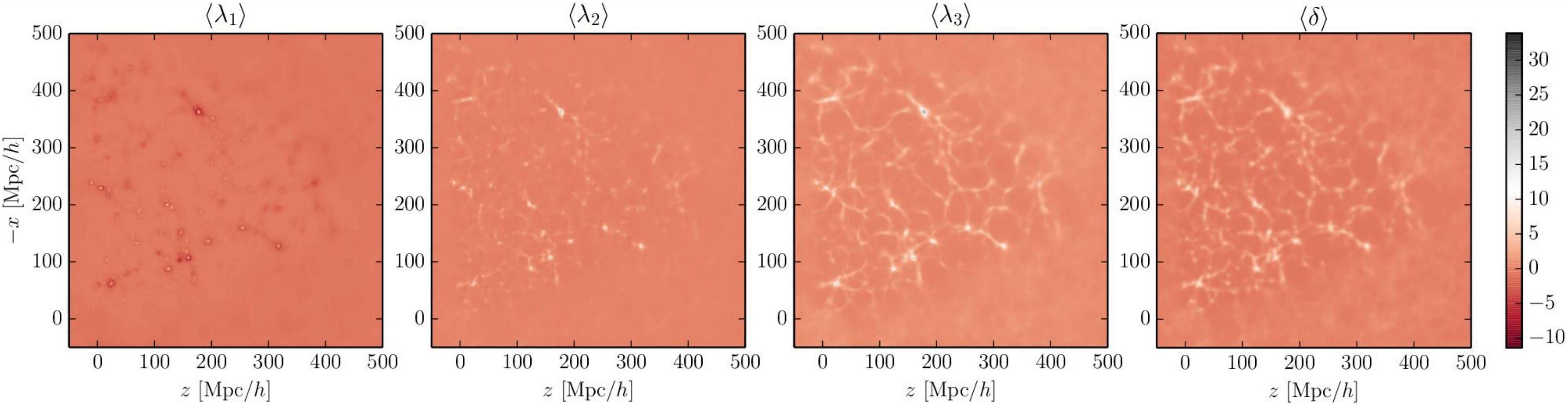
Forero-Romero *et al.* 2009, arXiv:0809.4135
Hoffman *et al.* 2012, arXiv:1201.3367

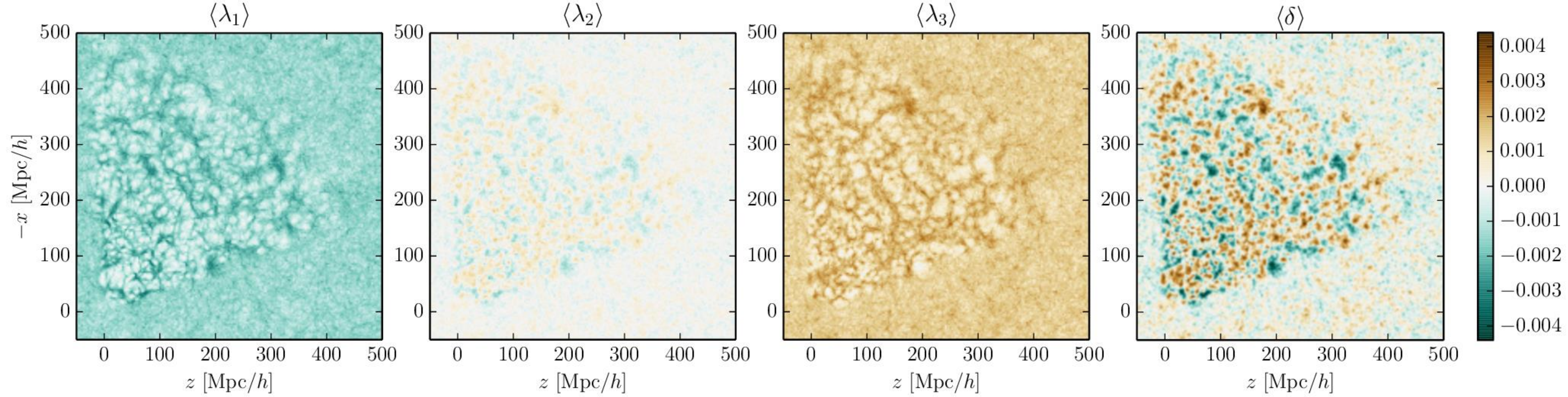
- Similar web classifiers:

DIVA, Lavaux & Wandelt 2010, arXiv:0906.4101
ORIGAMI, Falck, Neyrinck & Szalay 2012, arXiv:1201.2353

Tidal shear analysis

Final conditions



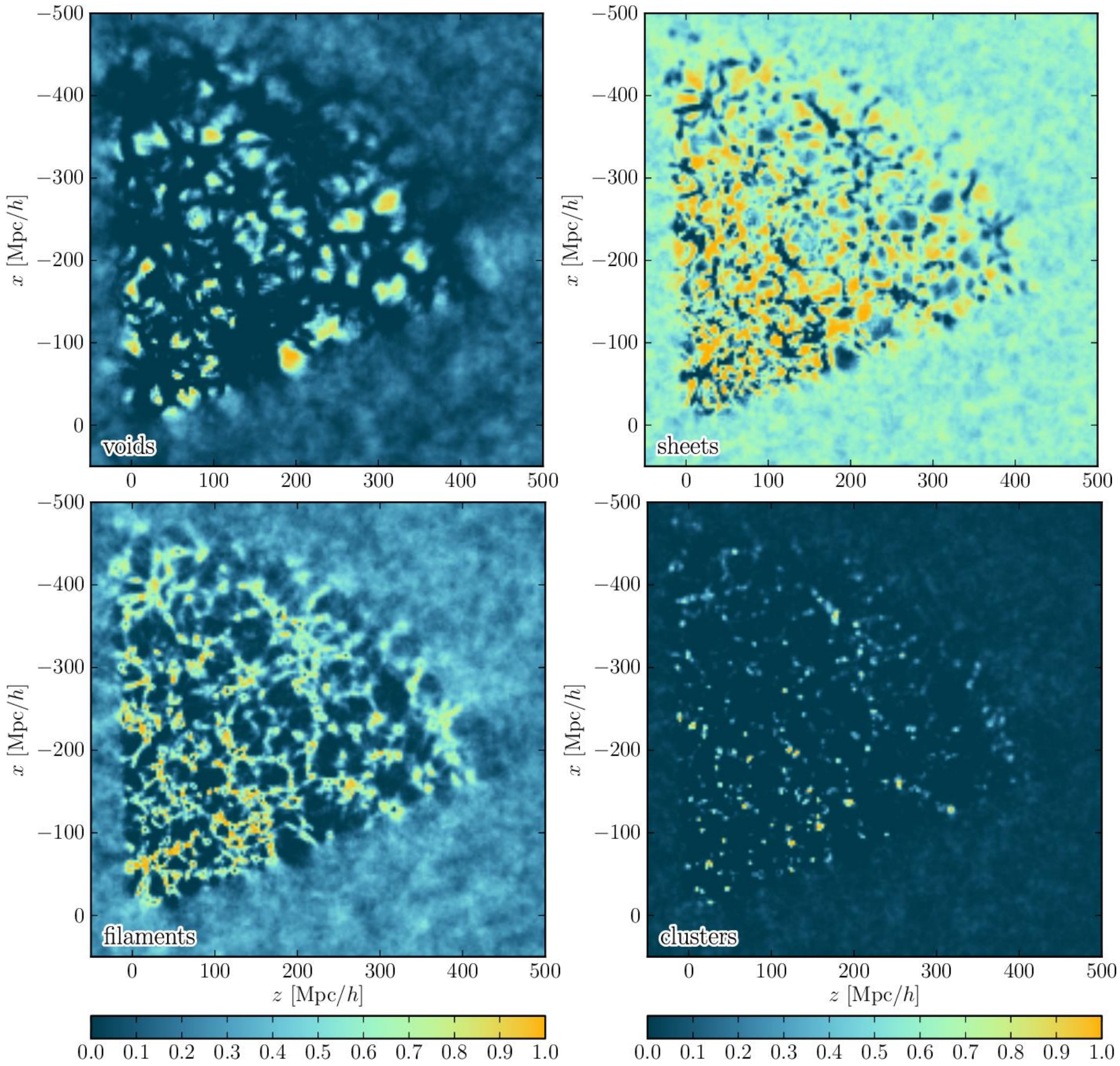


FL, Jasche & Wandelt 2015, arXiv:1502.02690

Initial conditions

Dynamic structures inferred by BORG

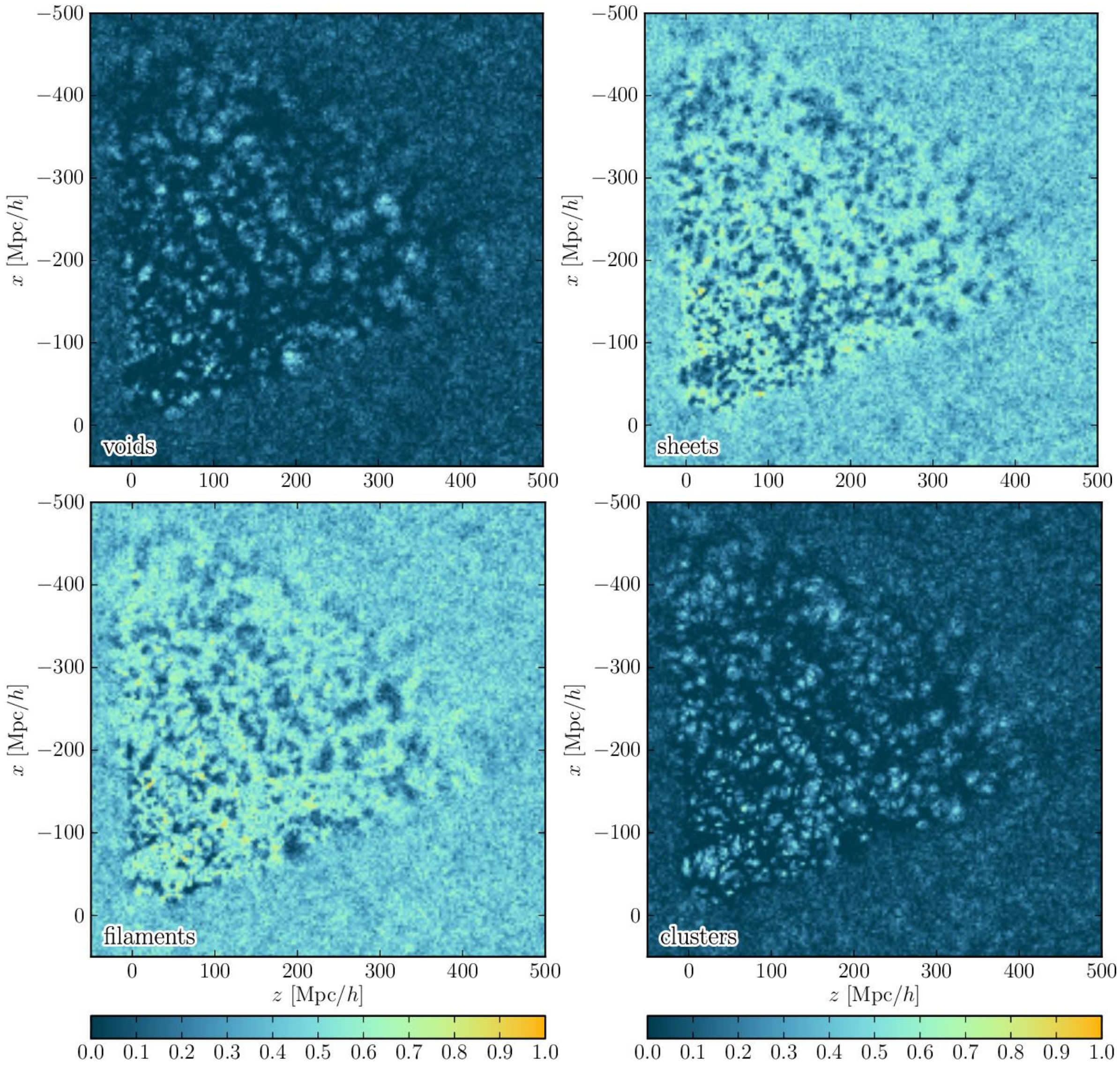
Final conditions



FL, Jasche & Wandelt 2015, arXiv:1502.02690

Dynamic structures inferred by BORG

Initial conditions

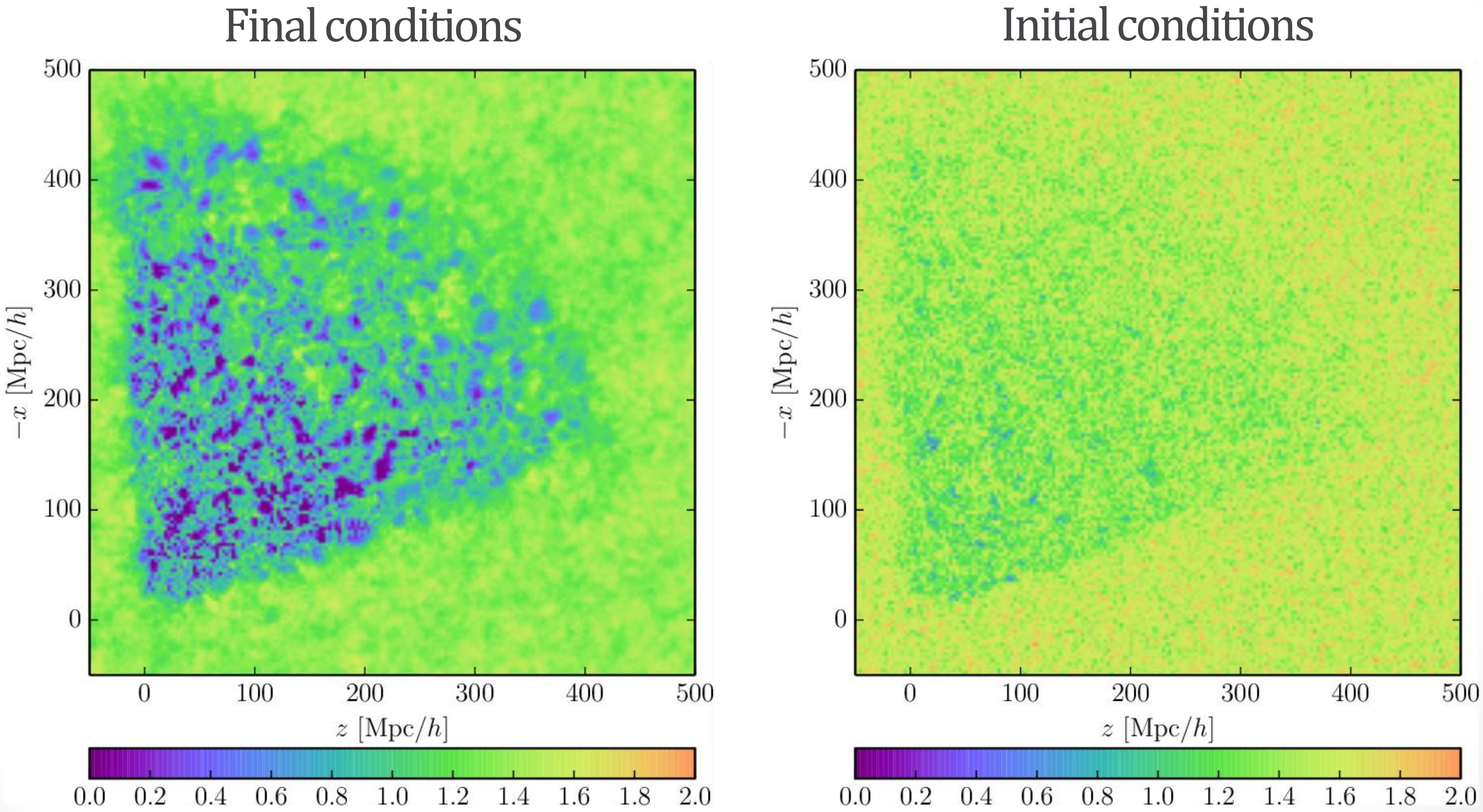


FL, Jasche & Wandelt 2015, arXiv:1502.02690

Entropy of the structure types pdf

$$H\left[\mathcal{P}(\mathrm{T}(\vec{x}_k)|d)\right] \equiv -\sum_{i=0}^{3} \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)\log_2(\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d))$$

in shannons (Sh)


Final conditions

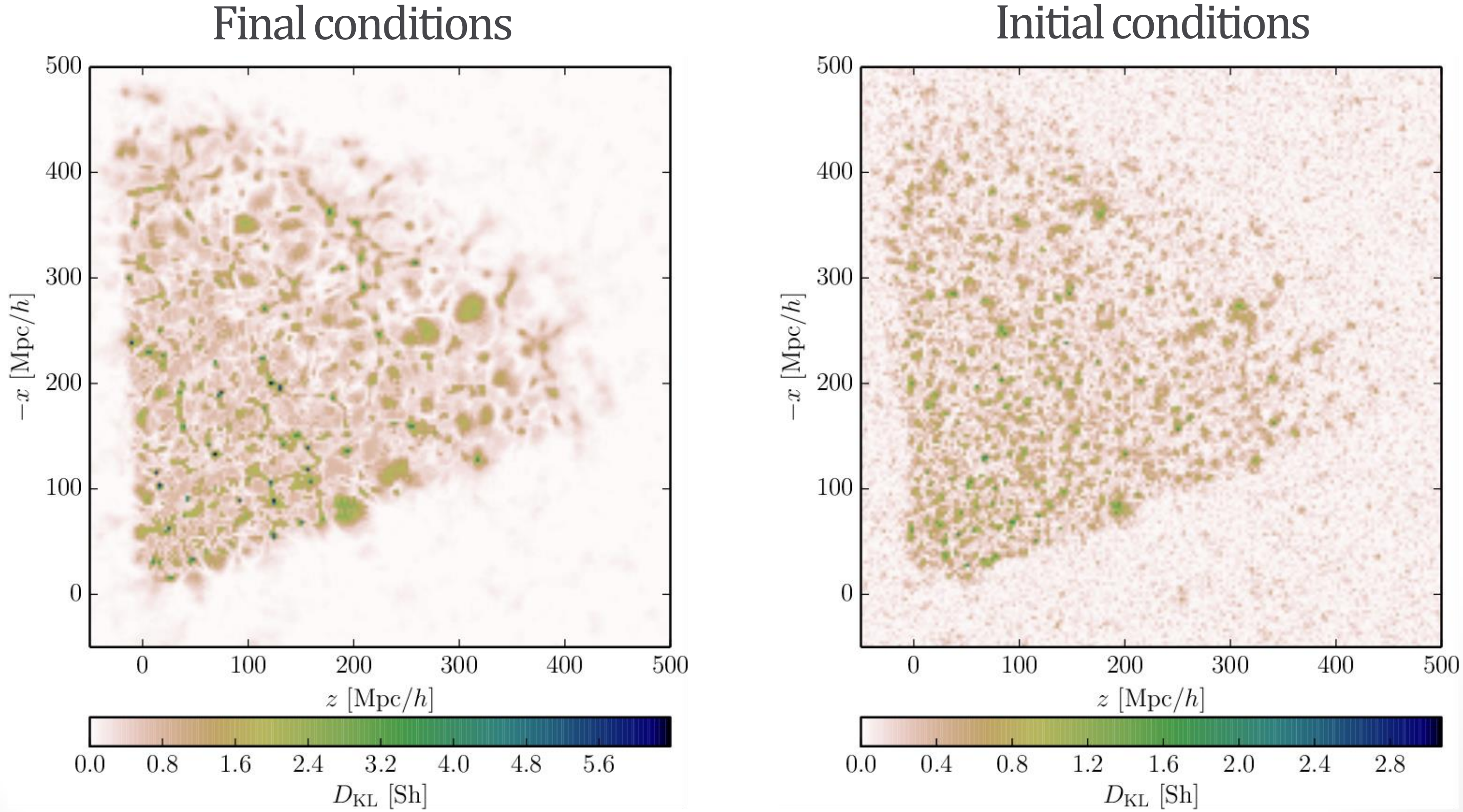
Initial conditions

FL, Jasche & Wandelt 2015, arXiv:1502.02690

Kullback-Leibler divergence posterior/prior

$$D_{\mathrm{KL}}(\mathcal{P}(\mathrm{T}(\vec{x_k})|d)||\mathcal{P}(\mathrm{T})) \equiv \sum_i \mathcal{P}(\mathrm{T}_i(\vec{x_k})|d) \log_2 \left(\frac{\mathcal{P}(\mathrm{T}_i(\vec{x_k})|d)}{\mathcal{P}(\mathrm{T}_i)} \right)$$

in Sh



FL, Jasche & Wandelt 2015, arXiv:1502.02690

A decision rule for structure classification

- Space of “input features”:

$$\{\mathrm{T}_0 = \text{void}, \mathrm{T}_1 = \text{sheet}, \mathrm{T}_2 = \text{filament}, \mathrm{T}_3 = \text{cluster}\}$$

- Space of “actions”:

$$\{a_0 = \text{“decide void”}, a_1 = \text{“decide sheet”}, a_2 = \text{“decide filament”}, a_3 = \text{“decide cluster”}, a_{-1} = \text{“do not decide”}\}$$

⇨ A problem of **Bayesian decision theory**:
one should take the action which maximizes the utility

$$U(a_j(\vec{x}_k)|d) = \sum_{i=0}^{3} G(a_j|\mathrm{T}_i)\,\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$$

- How to write down the gain functions?

FL, Jasche & Wandelt 2015, arXiv:1503.00730

Gambling with the Universe

- One proposal:

$$G(a_j|\mathrm{T}_i) = \begin{cases} \dfrac{1}{\mathcal{P}(\mathrm{T}_i)} - \alpha & \text{if } j \in [\![0,3]\!] \text{ and } i = j \qquad \text{“Winning”} \\ -\alpha & \text{if } j \in [\![0,3]\!] \text{ and } i \neq j \qquad \text{“Loosing”} \\ 0 & \text{if } j = -1. \qquad \text{“Not playing”} \end{cases}$$

- Without data, the expected utility is

$$U(a_j) = 1 - \alpha \quad \text{if } j \neq 1 \qquad \text{“Playing the game”}$$

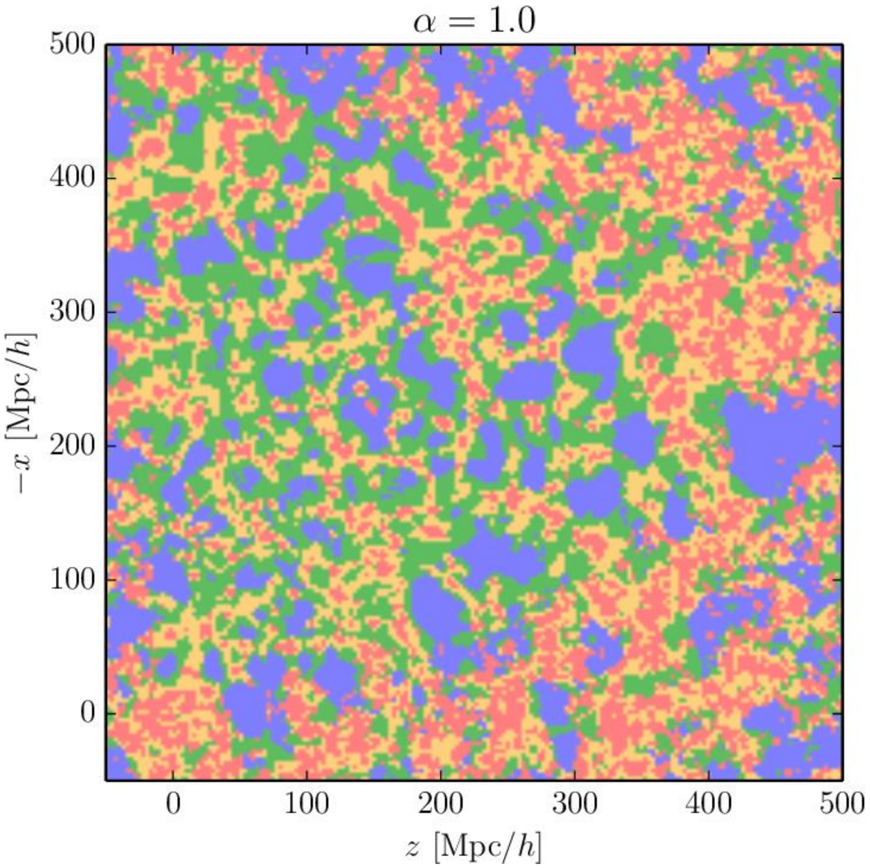
$$U(a_{-1}) = 0 \qquad \text{“Not playing the game”}$$

- With $\alpha = 1$, it’s a *fair game* → always play → “speculative map” of the LSS
- Values $\alpha > 1$ represent an *aversion for risk* → increasingly “conservative maps” of the LSS

FL, Jasche & Wandelt 2015, arXiv:1503.00730

Playing the game…

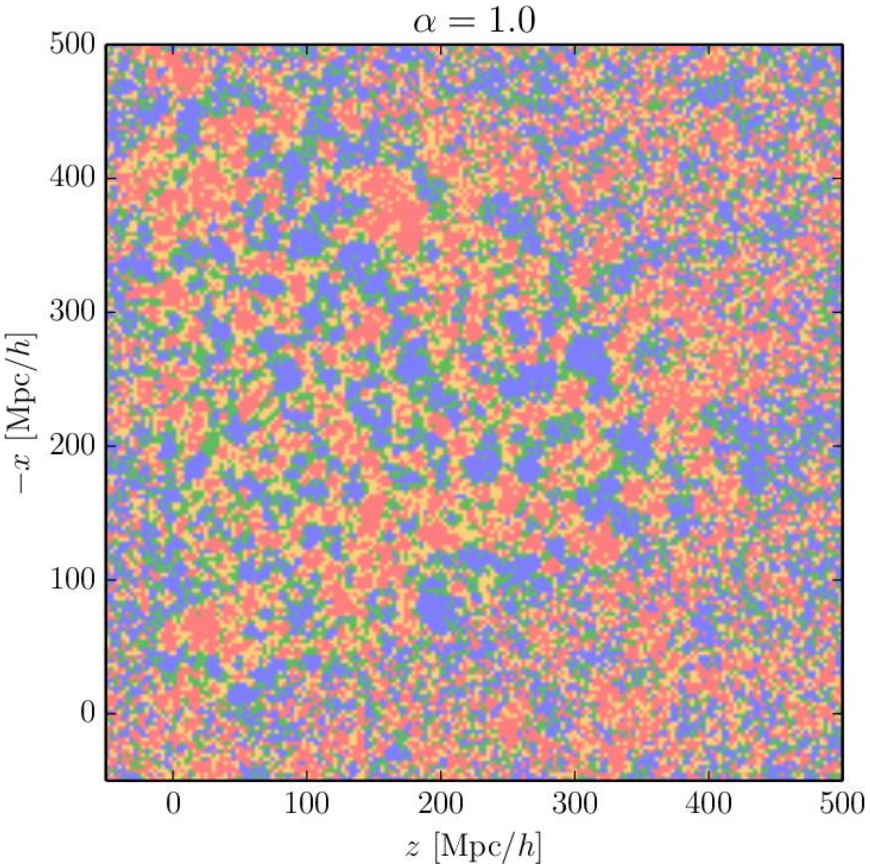
Final conditions



FL, Jasche & Wandelt 2015, arXiv:1503.00730

Playing the game…

Initial conditions



FL, Jasche & Wandelt 2015, arXiv:1503.00730

Summary & Conclusions

- Bayesian large-scale structure inference in 10 millions dimensions is possible!
  - Uncertainty quantification (noise, survey geometry, selection effects and biases)
  - Non-linear and non-Gaussian inference with improving techniques

- Application to data: four-dimensional chronocosmography
  - Simultaneous analysis of the morphology and formation history of the large-scale structure
  - Physical reconstruction of the initial conditions
  - Inference of cosmic voids at the level of the dark matter distribution
  - Characterization of the dynamic cosmic web underlying galaxies