High-performance computing and high-performance data analysis in cosmology

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

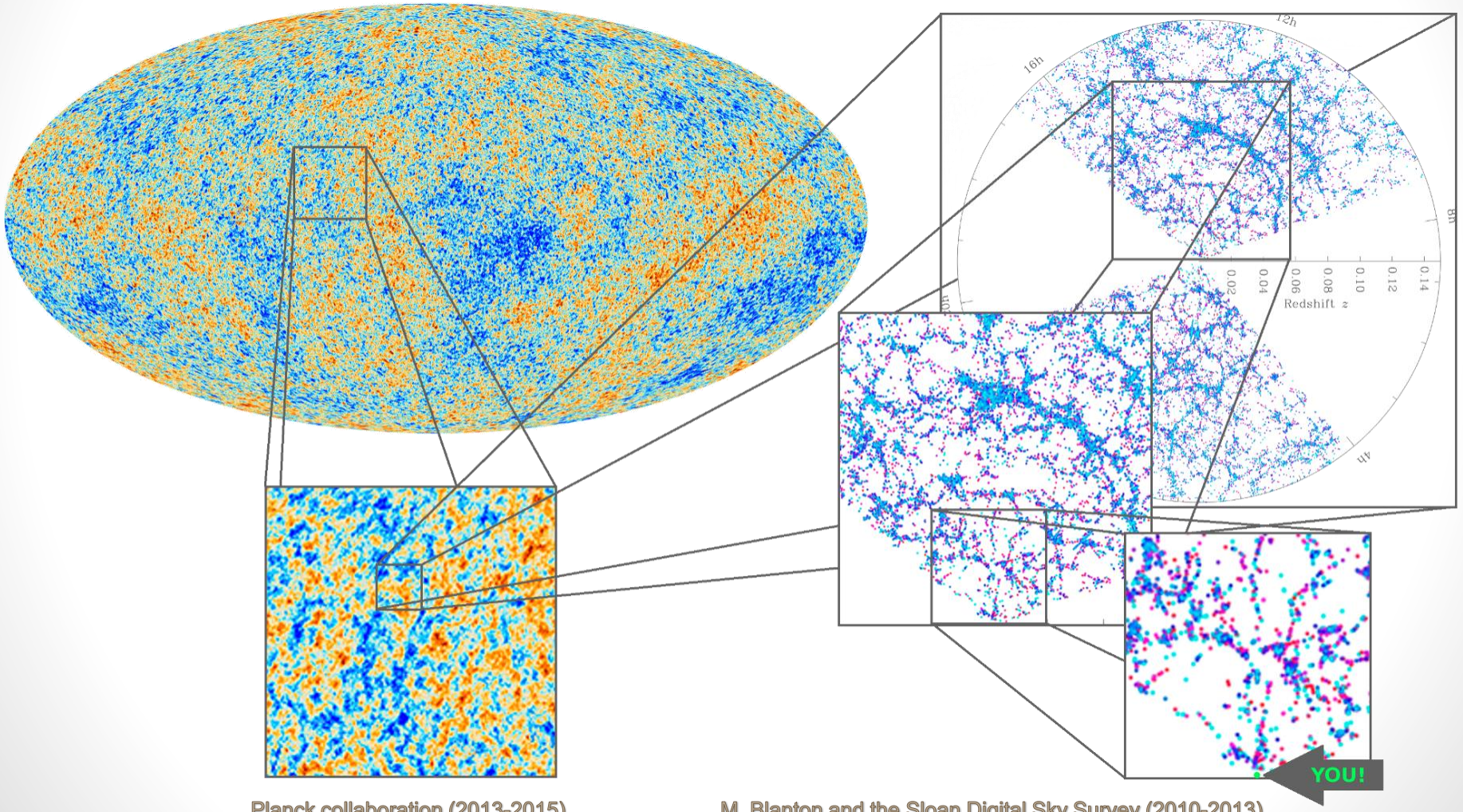
In collaboration with:
Alan Heavens, Andrew Jaffe (Imperial College),
Jens Jasche (U. Stockhom),
Guilhem Lavaux, Benjamin Wandelt (IAP),
Will Percival (U. Waterloo)

and the Aquila Consortium
www.aquila-consortium.org

6 July 2022

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)

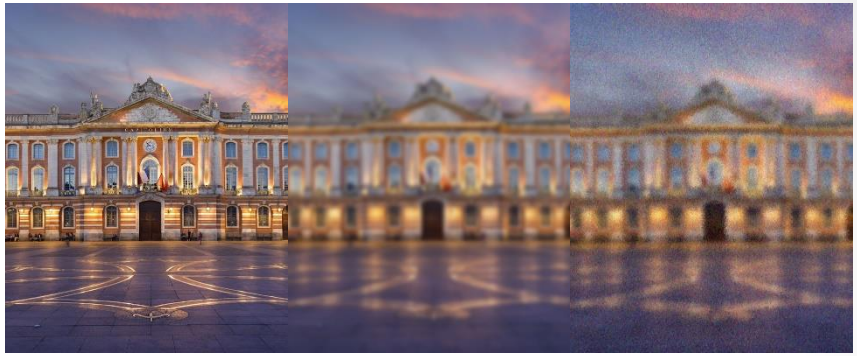
What we want to know from the large-scale structure

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

Why Bayesian inference?

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



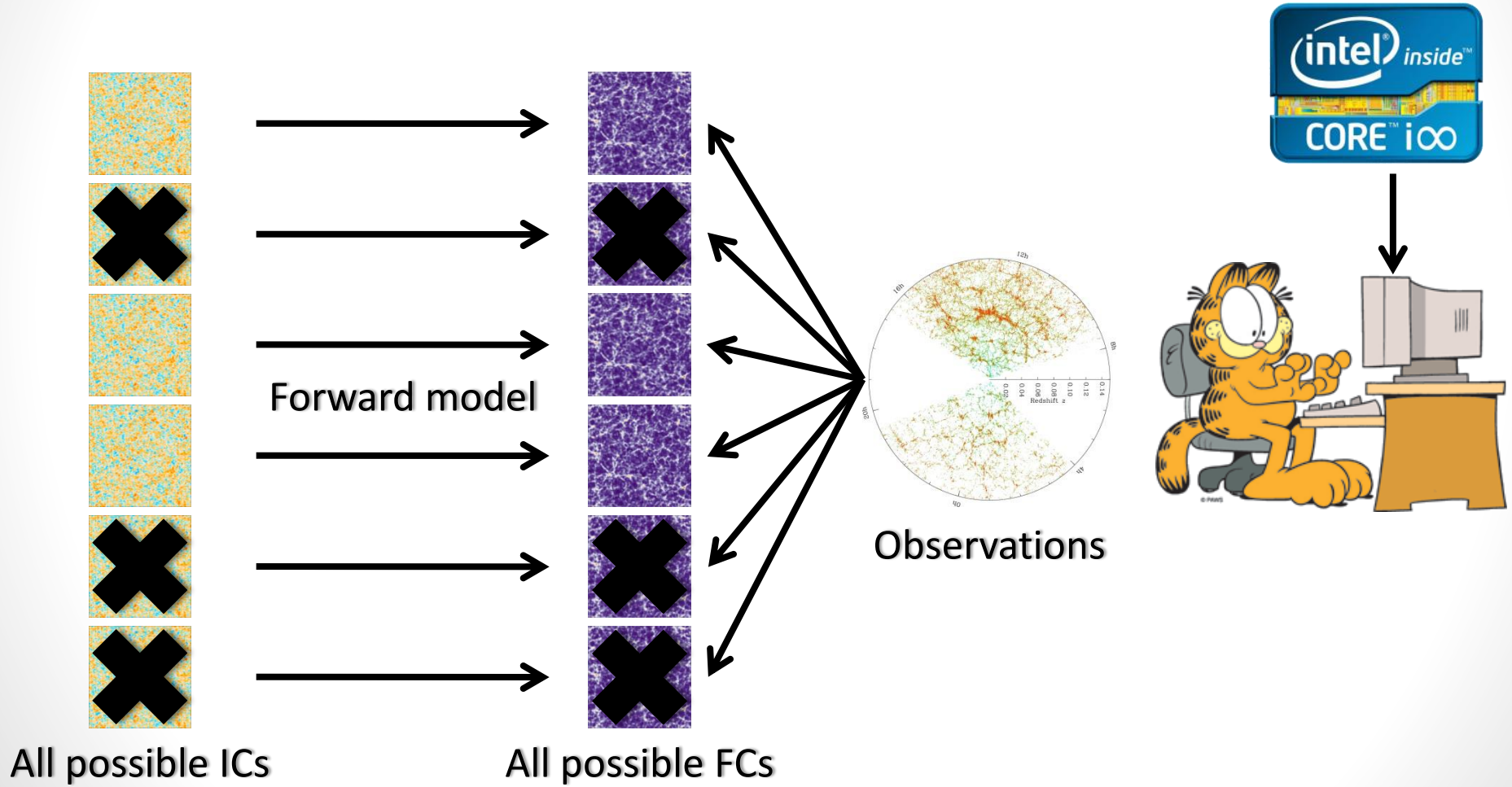
➡ No unique recovery is possible!

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

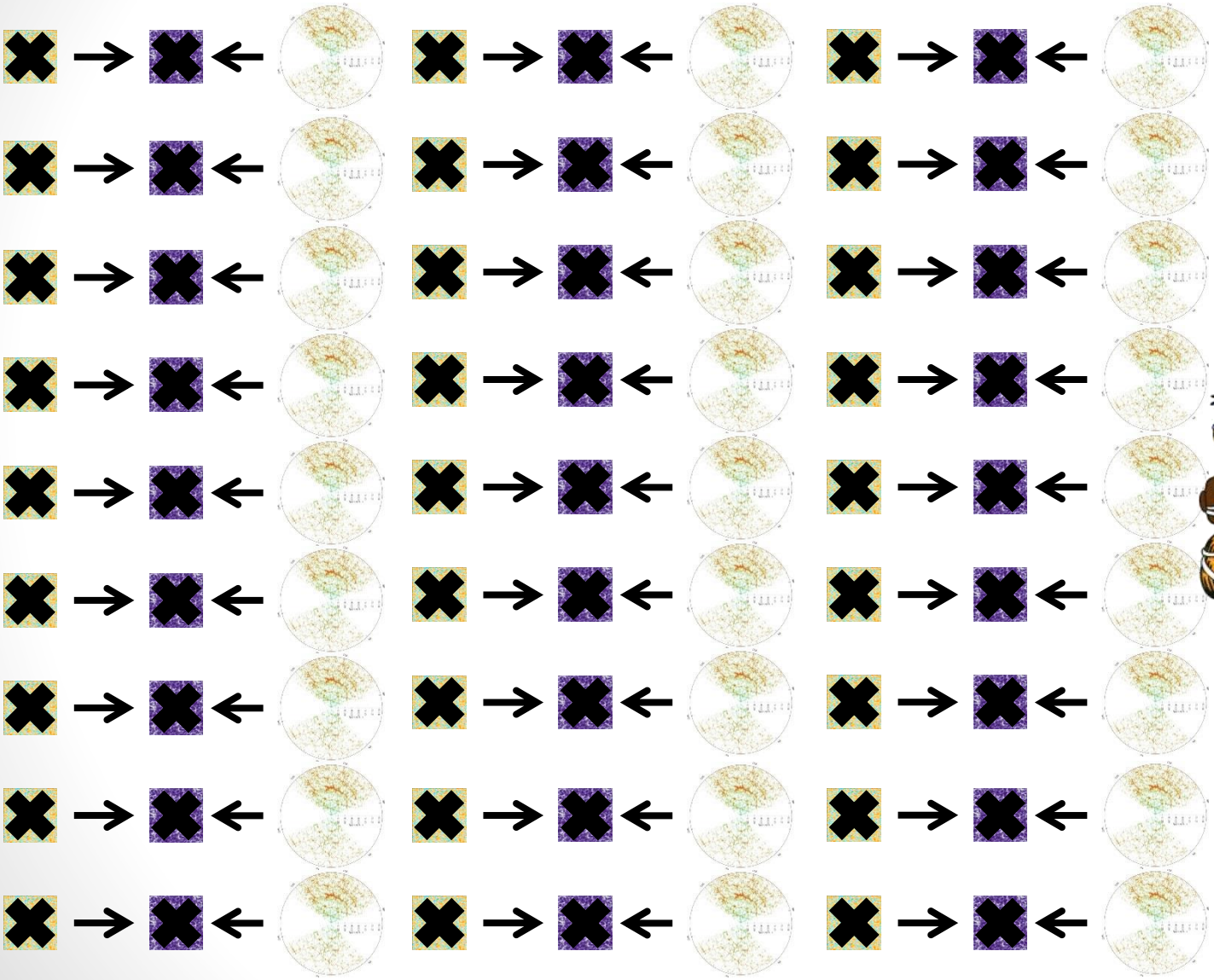
Bayes' theorem: $$\mathcal{P}(s|d)\mathcal{P}(d) = \mathcal{P}(d|s)\mathcal{P}(s)$$

- 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

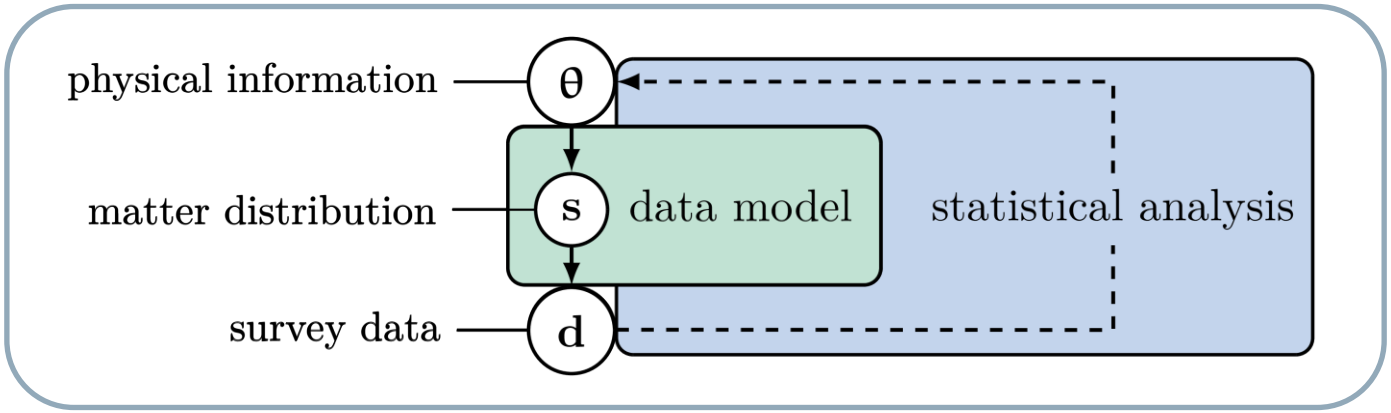


The (true) likelihood lives in

$d \approx 10^7$!

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

Hamiltonian (Hybrid) Monte Carlo

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

$$(\mathbf{x}, \mathbf{p}) \Rightarrow \left\{ \begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\frac{\partial H}{\partial \mathbf{x}} = -\frac{\mathrm{d}\psi(\mathbf{x})}{\mathrm{d}\mathbf{x}} \end{aligned} \right\} \Rightarrow (\mathbf{x}', \mathbf{p}')$$

gradients of the pdf

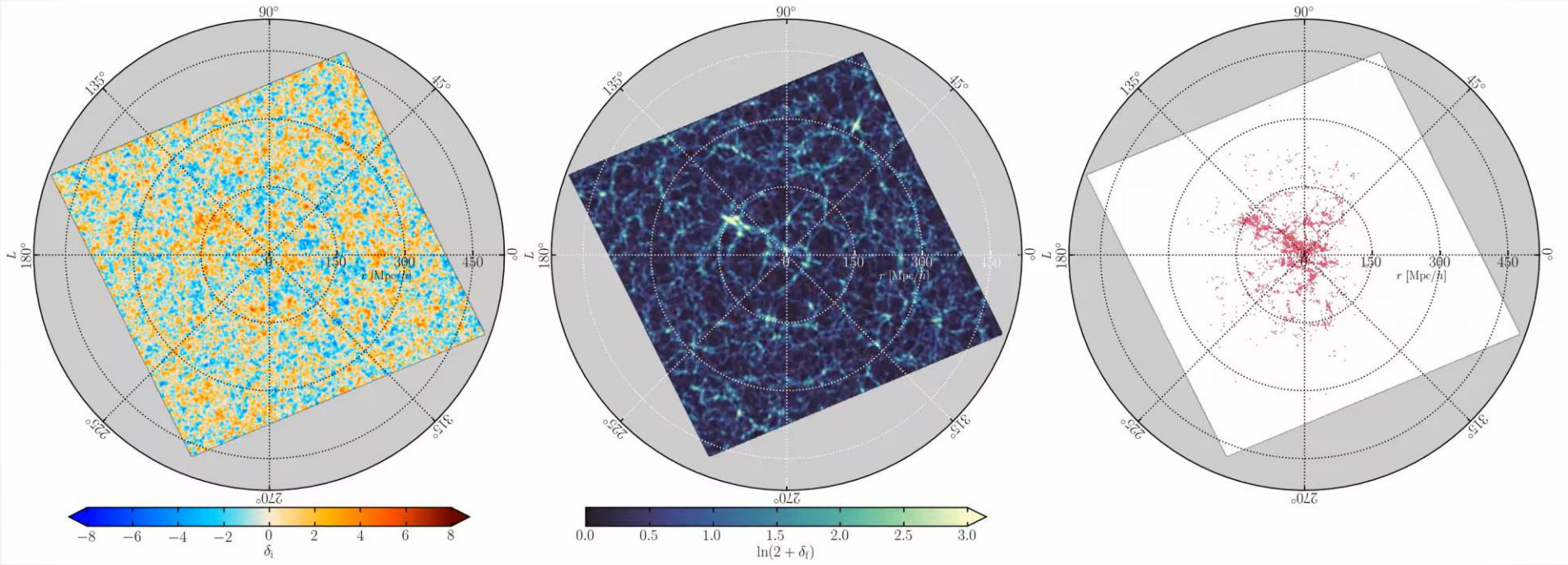
$$a(\mathbf{x}', \mathbf{x}) = \mathrm{e}^{-(H'-H)} = 1 \Leftarrow$$ acceptance ratio unity

- 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

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

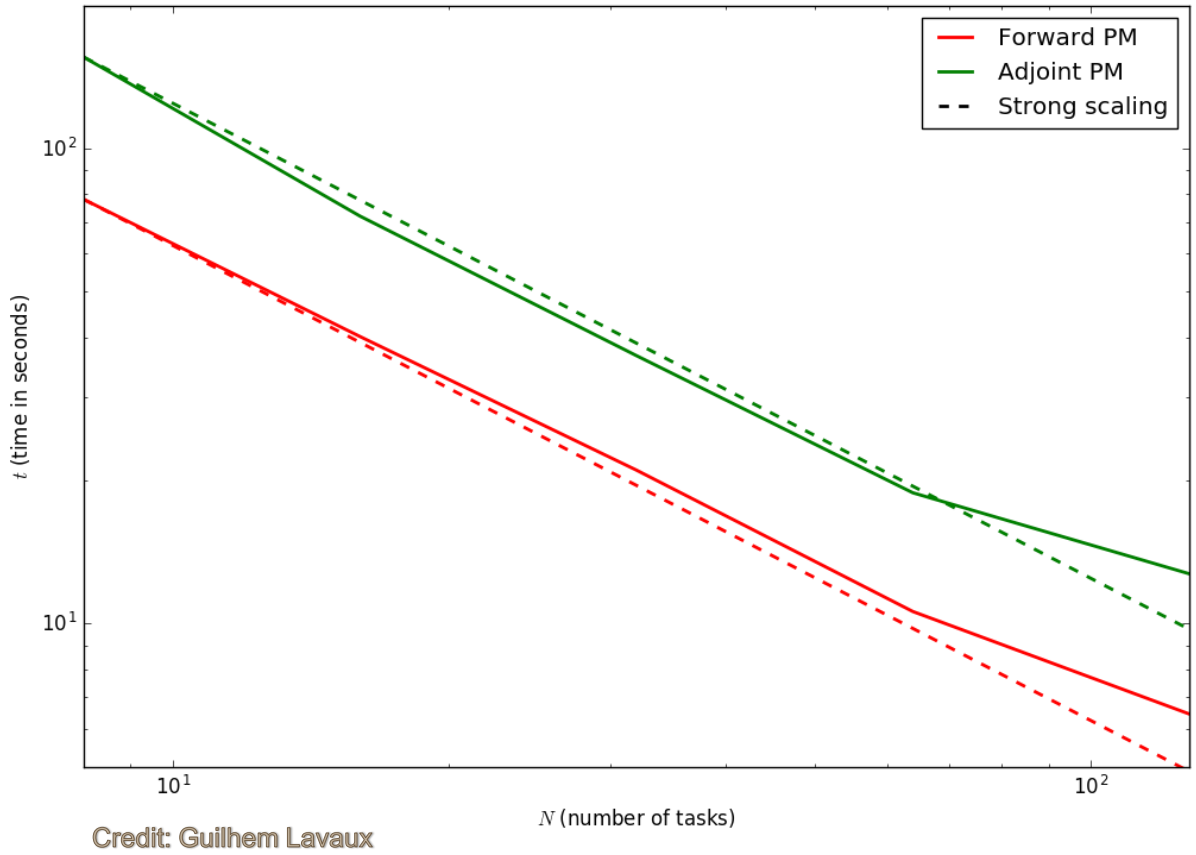


Supergalactic plane

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 & Wandelt, 1203.3639; Jasche, FL & Wandelt, 1409.6308; Jasche & Lavaux, 1806.11117; Lavaux, Jasche & FL, 1909.06396

BORG scaling test



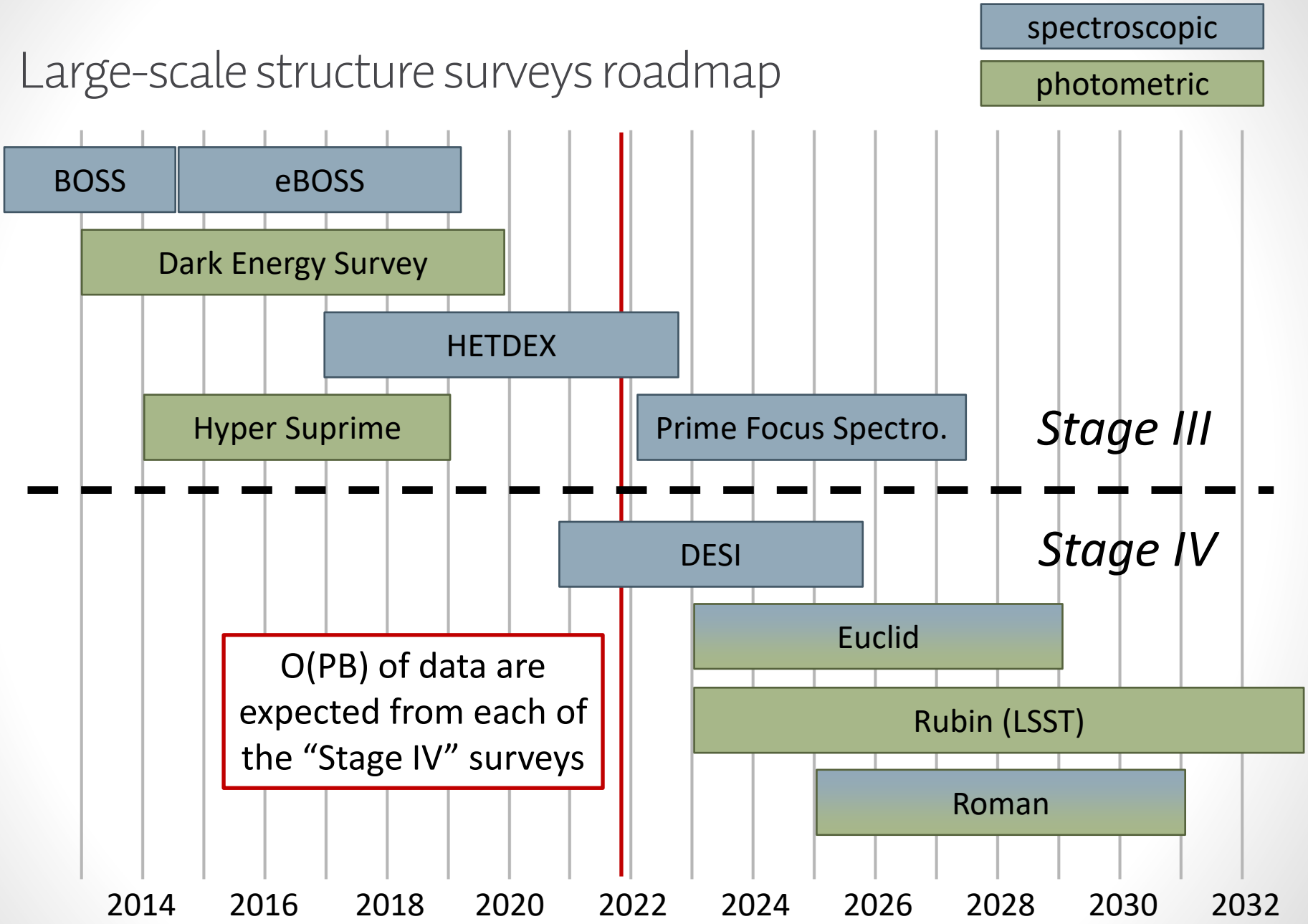
Credit: Guilhem Lavaux

- BORG scales up to 1024 cores (128 MPI tasks of 8 CPUs)
- Hyperthreading explains the saturation at 128 tasks

The Future: Opportunities & Challenges

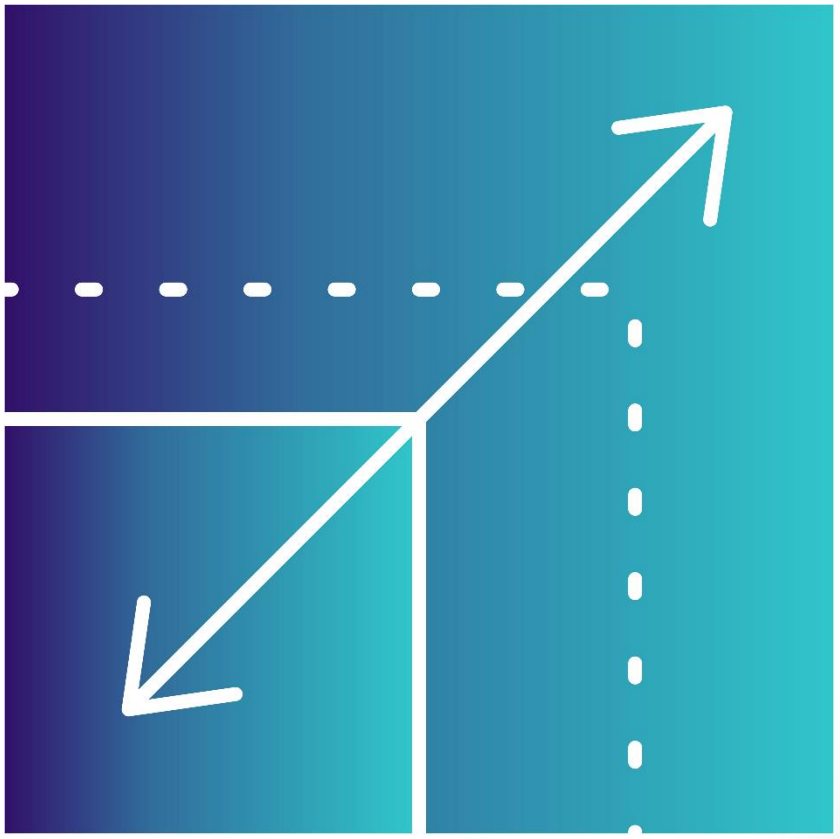
DESI, Euclid, Rubin, Roman, and more…

Large-scale structure surveys roadmap



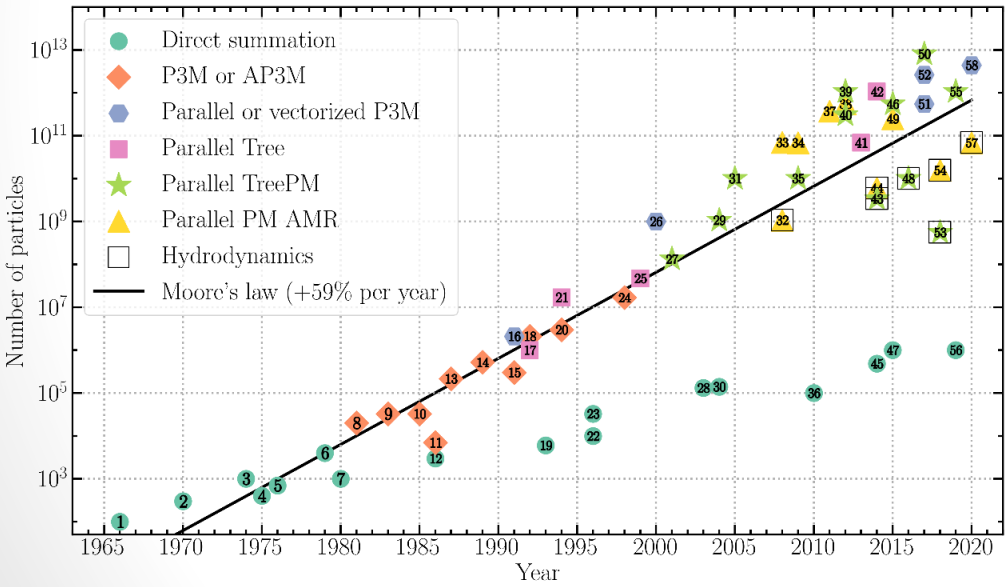
Our main challenge is not scaling, but scalability

- Scalability: the property of algorithms to handle a growing amount of data under computational resource constraints.
- The challenge is twofold:
  - in the data models: how can we best use modern computers and their architecture?
  - in the inference techniques: how can we perform rigorous Bayesian reasoning given a limited computational budget?
    (➡ a good subject for another talk)

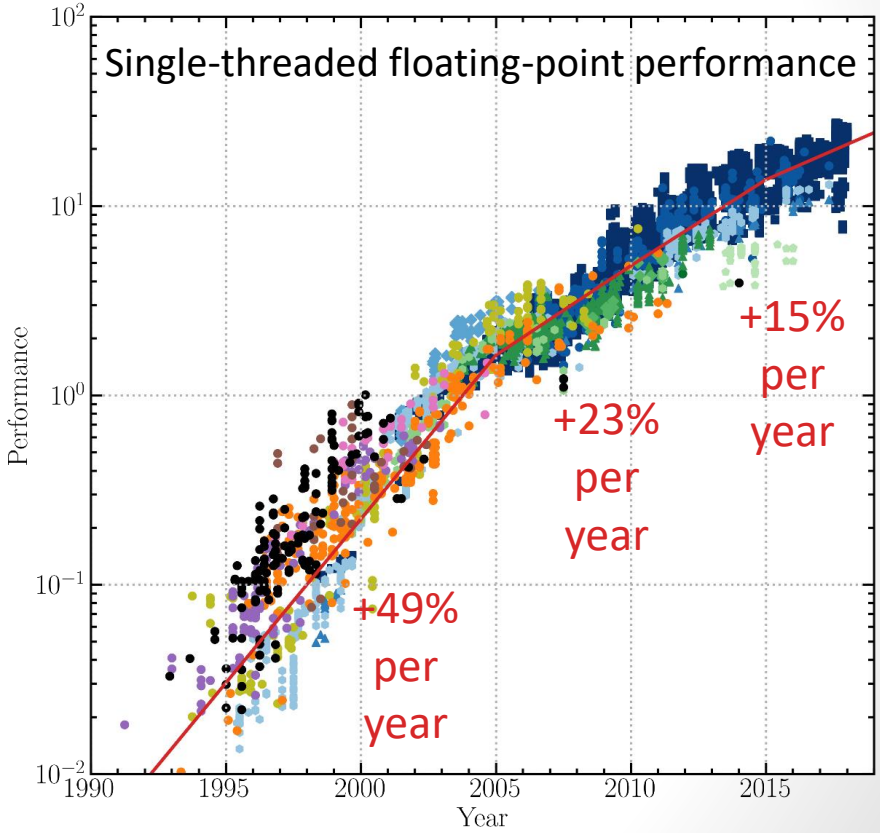


Parallelisation of N-body codes: the challenge

- Most of the work on numerical cosmology so far has focused on algorithms (such as tree, multipole, and mesh methods) that reduce the need for communications across the full computational volume
- But per-core compute performance is slowing down

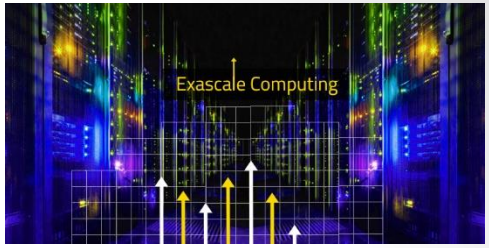


References for all points available at http://florent-leclercq.eu/blog.php?page=2



Based on adjusted SPECfp® results, http://spec.org

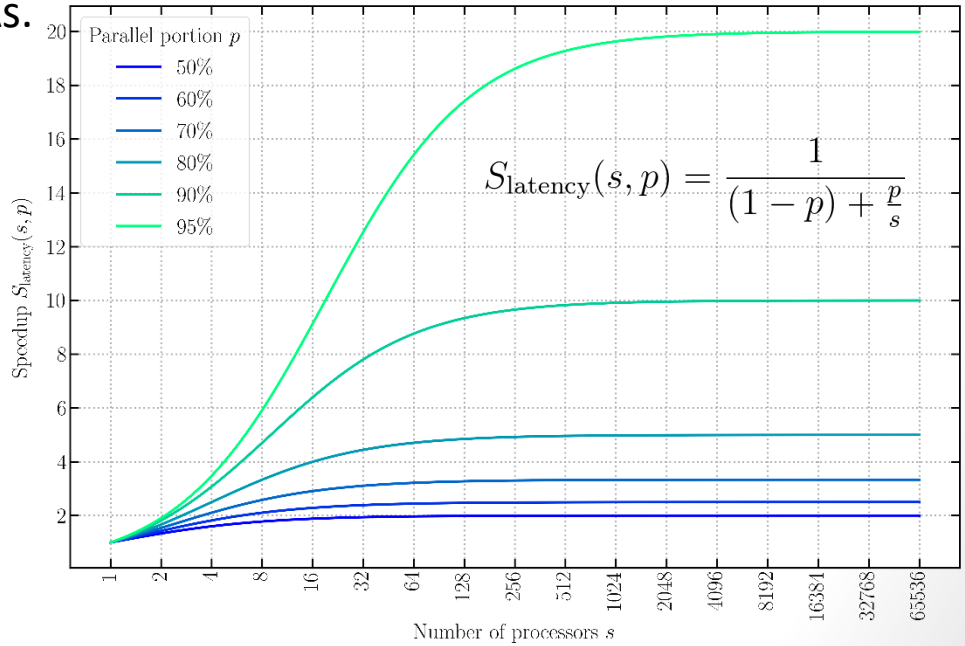
Numerical data models in the exascale world



- 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.

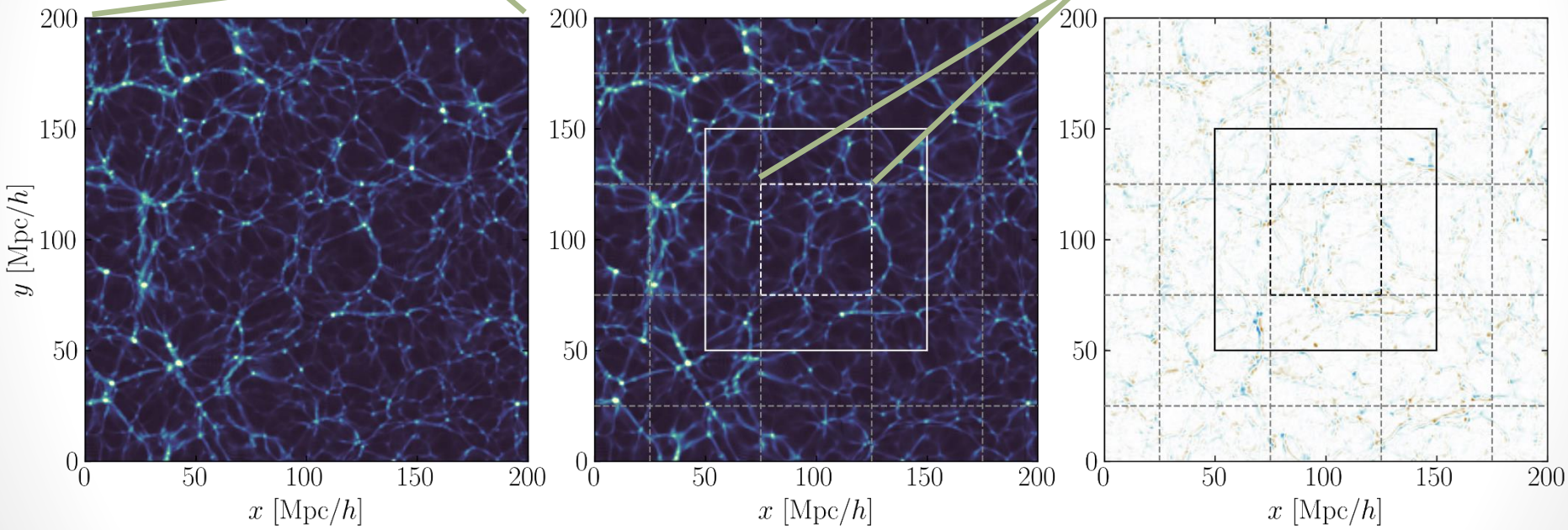
Perfectly parallel cosmological simulations using sCOLA

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

$$\frac{\partial^2 \mathbf{x}}{\partial t^2} = -\nabla \left[\Delta^{-1}\delta\right] \qquad \frac{\partial^2}{\partial t^2}\left(\mathbf{x} - \mathbf{x}_{\mathrm{l.s.}}\right) = -\nabla \left[\Delta^{-1}\left(\delta - \delta_{\mathrm{l.s.}}\right)\right]$$

non-local

local



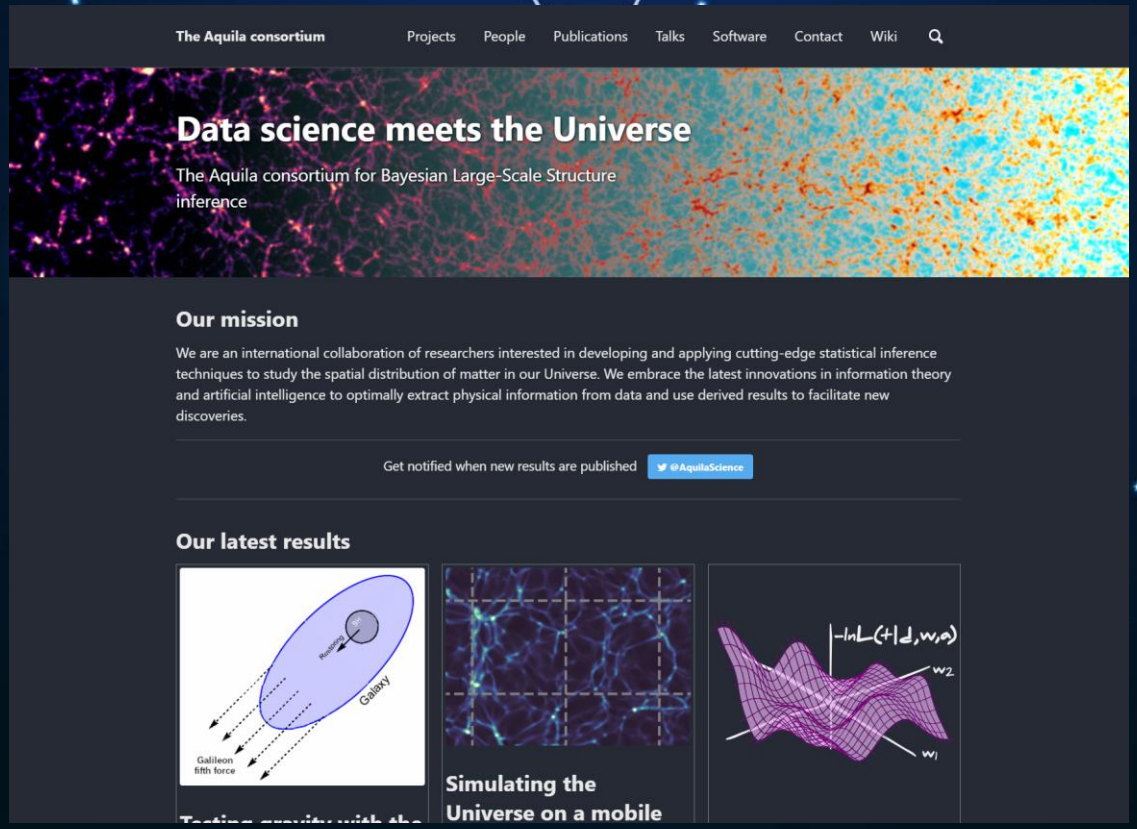
tCOLA (reference) sCOLA Difference

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

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

The Aquila Consortium

- Created in 2016. Currently 40 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

- Bayesian analyses of galaxy surveys with fully non-linear numerical models is not an impossible task!
- Cosmological analyses triggers studies in high-performance data analysis (HPDA) and high-performance computing (HPC)
- The future: data-intensive scientific discovery from galaxy surveys. Can data analysts keep pace?