Hopes and challenges in data science for cosmology

ASNUM2022 : Journées de l'Action Spécifique Numérique de l'INSU

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

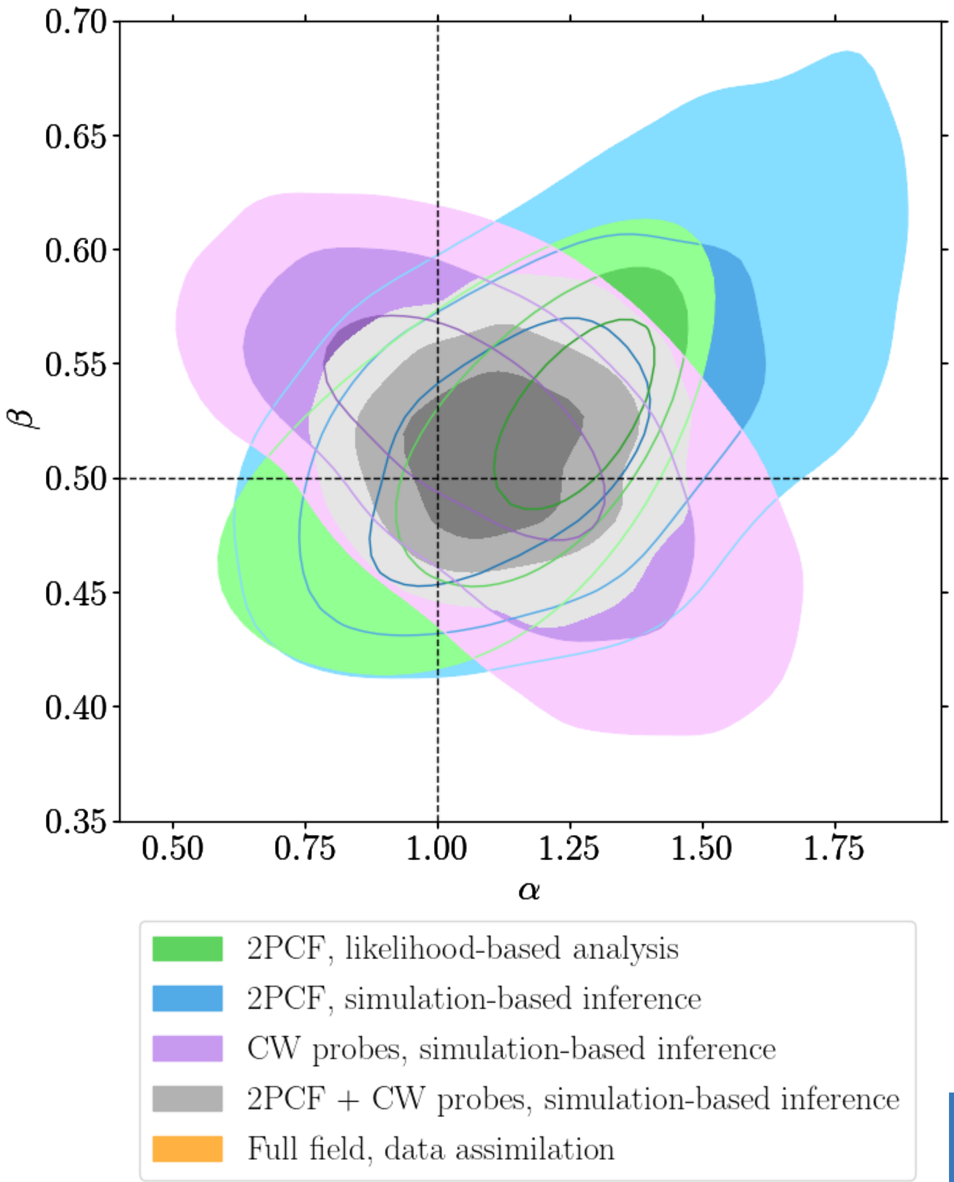
In collaboration with the Aquila Consortium

www.aquila-consortium.org

15 December 2022

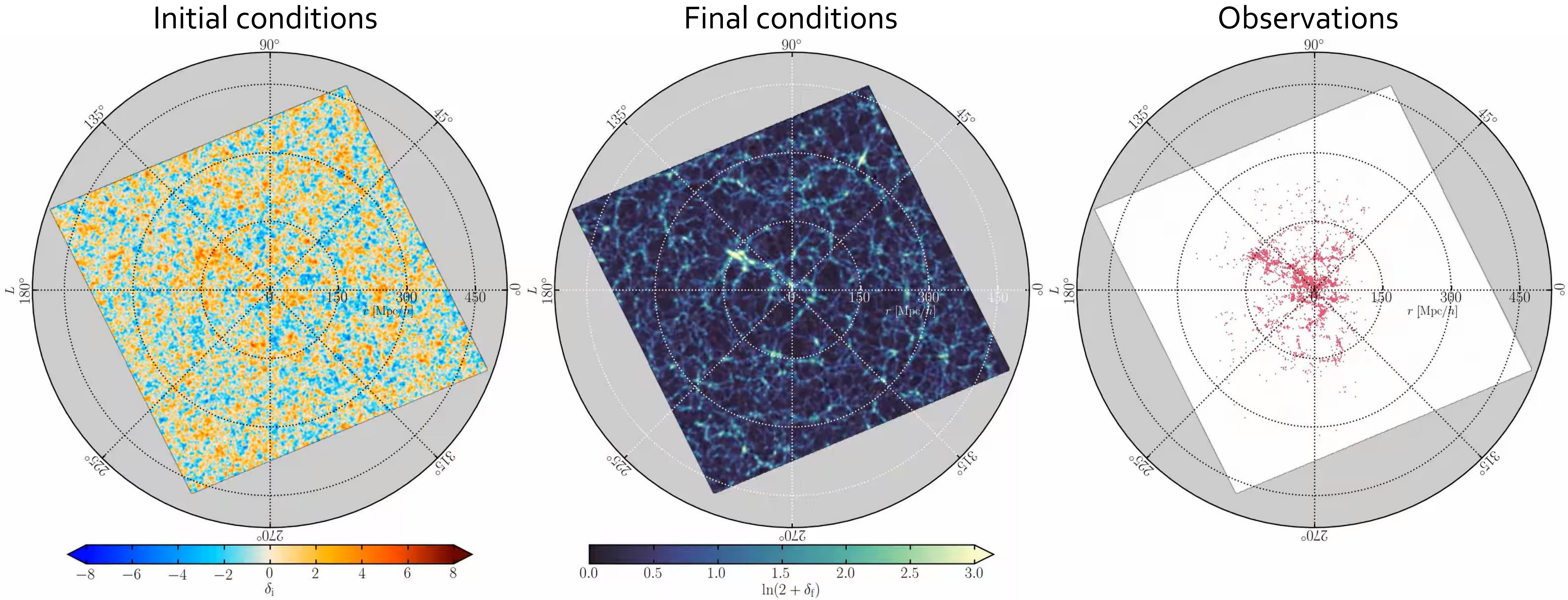
What there is to learn and how to get there

- A question of accuracy: first, avoid biases.
- A question of precision: can numerical forward models be used to push further than $k \gtrsim 0.15\ h/\mathrm{Mpc}$? The full field contains much more information.
- A question of 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?



FL & Heavens, 2103.04158

Field-level cosmological inference:
Bayesian Origin Reconstruction from Galaxies (BORG)

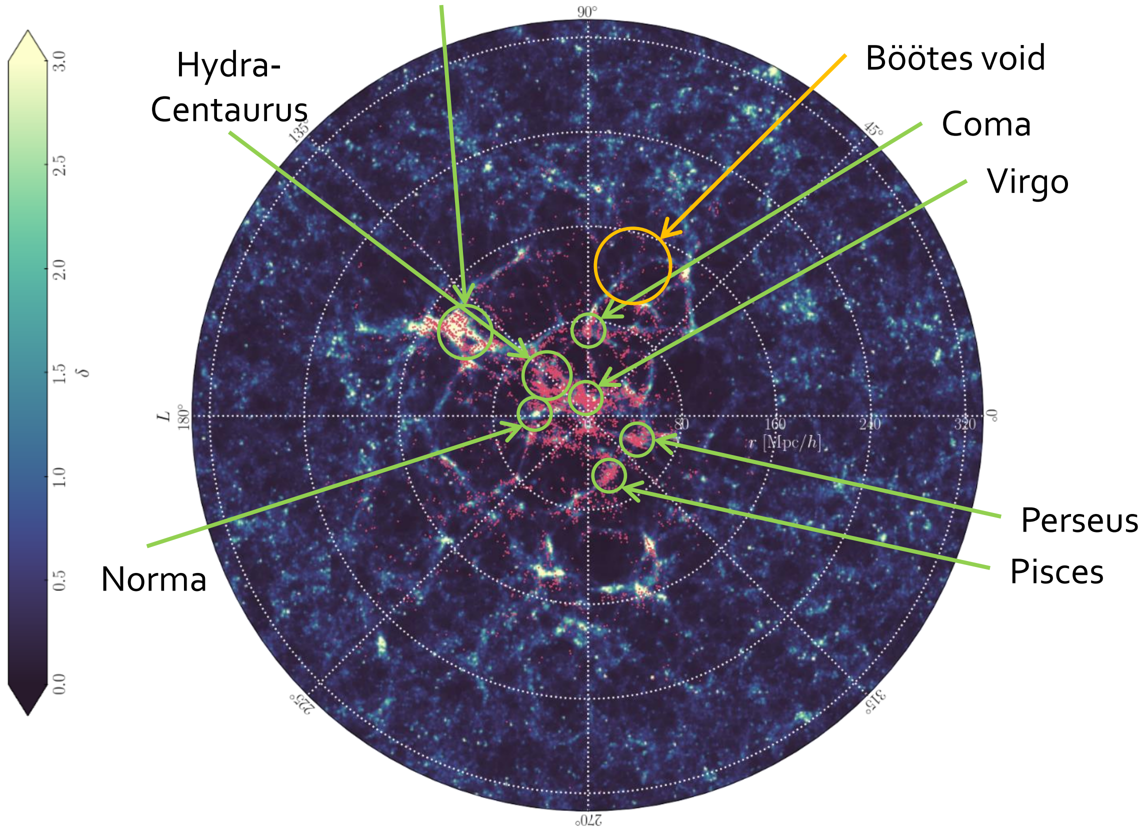


67,224 galaxies, ≈ 17 million parameters, 5 TB of primary data products, 10,000 samples, ≈ 500,000 forward and adjoint gradient data model evaluations, 1.5 million CPU-hours

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

BORG is beyond the proof-of-concept stage

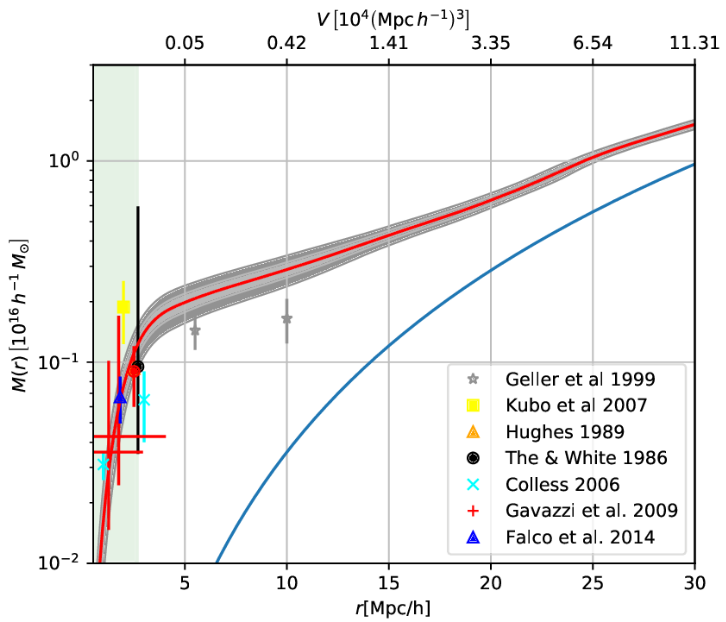
- Since 2014, BORG has been routinely applied to real state-of-the-art data.



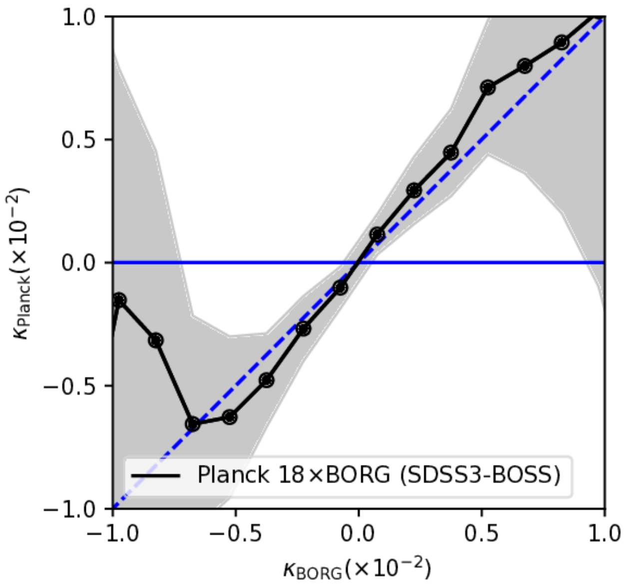
Jasche & Lavaux, 1806.11117; FL, Lavaux & Jasche, in prep.

- Density field reconstructions are in agreement with gold standard complementary data (lensing, X-ray, CMB).

Coma cluster mass profile



Cross-correlation with CMB lensing



Lavaux, Jasche & FL, 1909.06396

Some technical considerations

- BORG is a complex framework (~80,000 lines of C++ code, 10 developers over the last ten years),
  - It is compatible with modern popular tools such as Julia and JAX.
  - But it has been designed to the core for MPI multi-CPU capability, with multi-GPU capability currently under development.
  - The forward and adjoint gradient models show strong scaling on up to 1,000 cores.
- The barrier for entry is high (challenging for a ~3 year PhD), but the scientific reward is correspondingly high, especially for real data applications.

- Over the last few years, several cosmological codes with features common to BORG (e.g. differentiable *N*-body simulator, high-dimensional sampler/optimiser) have been written.
- BORG vs out-of-the-shelf (PyTorch, TF, JAX)
  - Typical memory overconsumption that limits the resolution/scalability.
  - Challenging lack of homogeneity of frameworks (e.g. TF1 → TF2 → JAX).
  - Difficult multi-node capability.
  - Complex management of dependencies, possible subsequent issues with reproducibility.
  - Lack of language flexibility (e.g. incompatibility with Julia, C++).

My point of view: "*no free lunch*" –
Algorithms and codes will always need to be adapted to problems.

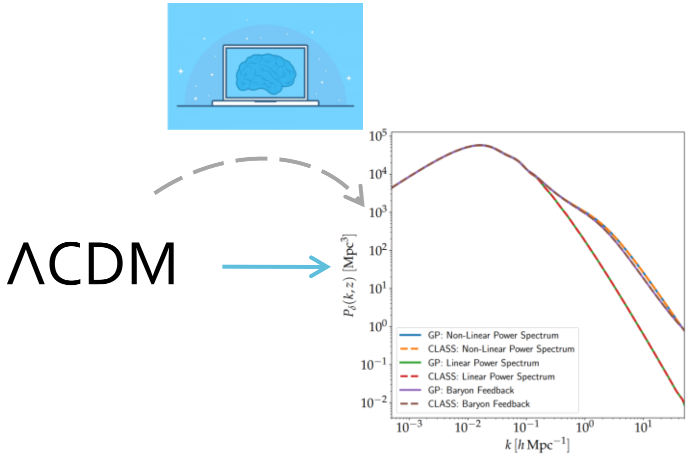
AI algorithms: metaphors & methodology

Symbolic AI, explainable but costly
Numerical AI/ML, automatic but "black-box"

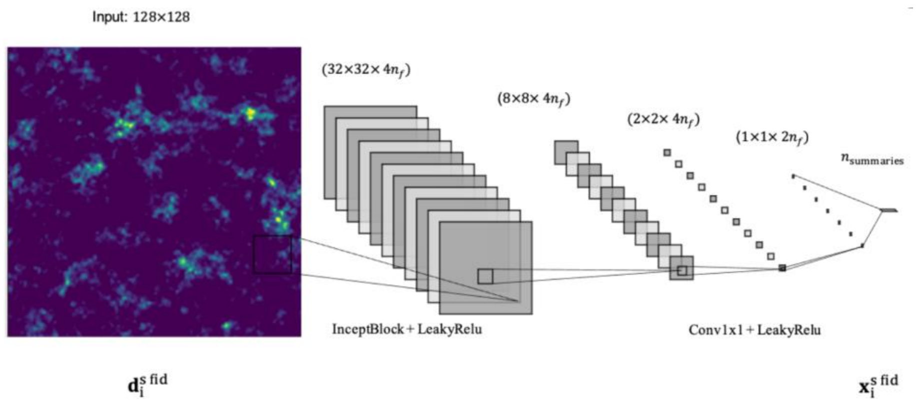
- Humanity: classical theories of learning
  - Rule-based models, case-based reasoning (Aamodt & Plaza 1994)
  - Learning by practice, "chunking" (Newell & Rosenbloom 1981)
  - Reinforcement learning (Samuel 1959)
  - Non-supervised learning (Feigenbaum 1963), e.g. auto-encoders (Kramer 1991)
- Physiology: the brain
  - Artificial neuron (McCullogh & Pitts 1943), perceptron (Rosenblatt 1958)
  - Multi-layer perceptrons (Rumelhart *et al.* 1986, Rumelhard & McClelland 1987), gradient back-propagation (Rumelhart *et al.* 1986)
  - Deep learning & convolutional neural networks (LeCun et al. 2015, Goodfellow et al. 2016)
- Nature: evolution
  - Genetic algorithms (Holland 1975)
- Culture: epistemology
  - Scientific discovery (Langley et al. 1987)
  - Ontologies (Powers & Turk 1989), semantic web
- Physics: statistical mechanics, thermodynamics, quantum physics
  - Decision trees (Quinlan 1975), Bayesian networks, graphs
  - Hamiltonian Monte Carlo (Duane *et al.* 1987)
  - Information theory, distributed AI (Demazeau & Müller 1989)
  - Hidden Markov Models (Baum 1966)

Why machine learning for cosmology?

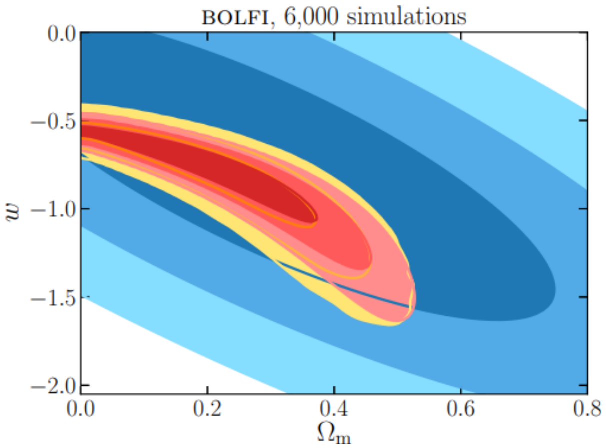
Speed up & go beyond approximations	Find the information content	Build a posterior/evidence approximator
Emulators	Automatic data compression	Implicit likelihood inference



emuPK: Mootoovaloo, Jaffe, Heavens & FL, 2105.02256



Information Maximising Neural Networks (IMNN): Charnock, Lavaux & Wandelt, 1802.03537; Makinen *et al.*, 2107.07405

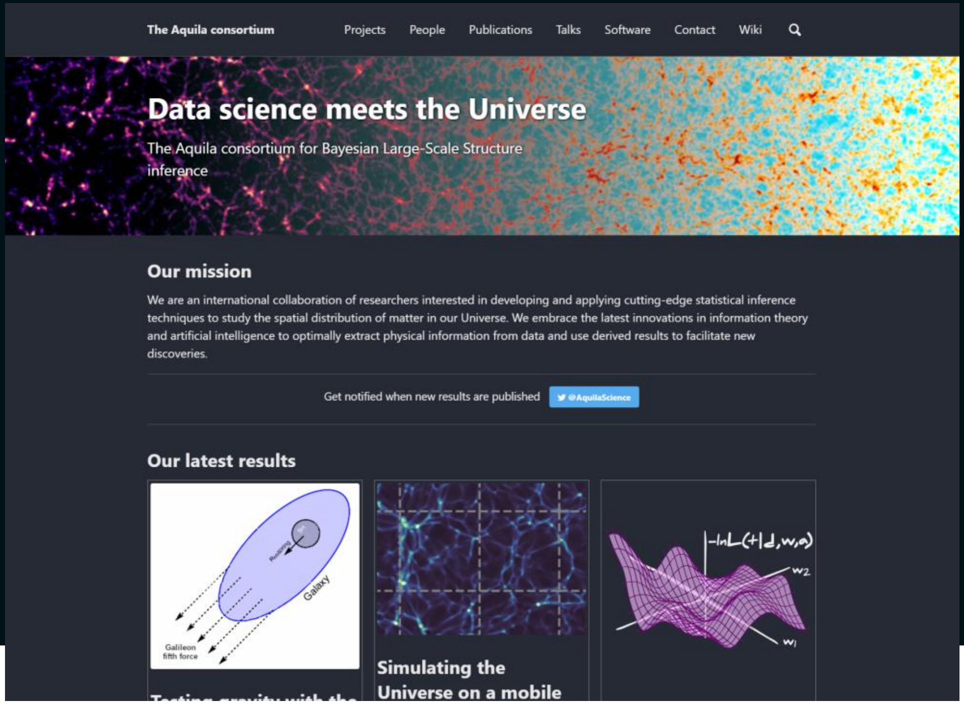


Bayesian Optimisation for Likelihood-Free Inference (BOLFI): FL, 1805.07152

My point of view: "*If you have a hammer, everything looks like a nail.*" – Deep learning is not the solution to all problems.

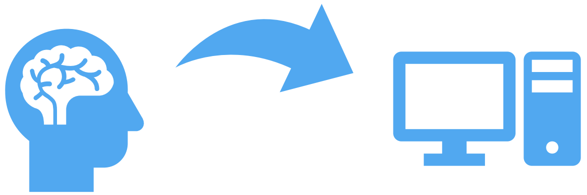
The Aquila Consortium

- Created in 2016. Currently 38 members from 8 countries (Europe & Americas).
- Gathers people interested in developing Bayesian pipelines and running analyses on cosmological data.



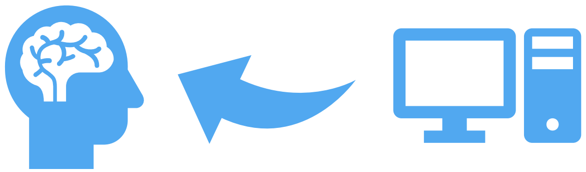
Visit us at www.aquila-consortium.org

Conclusion:
Hopes and challenges in data science for cosmology



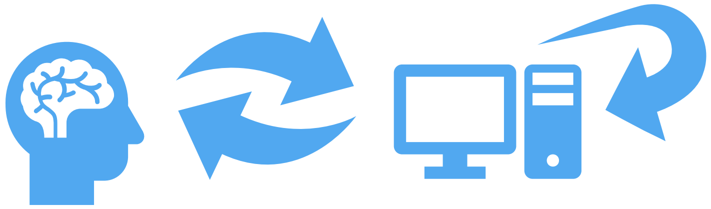
The forward problem

- Hopes: Numerical models are the new way to formulate theory in data analysis.
- Challenges: Scalability & design choices in implementations



The inverse problem

- Hopes: Field-level inference is established and validated on real survey data.
- Challenges: Control of external components in modern Bayesian models (in addition to likelihood and prior) : training data, posterior approximator…



The imitation problem

- Hopes: Machine-driven scientific discovery becomes conceivable.
- Challenges: Interpretability & explainability