Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Optimal compression and Simulation-Based Inference of the cosmic 21-cm signal

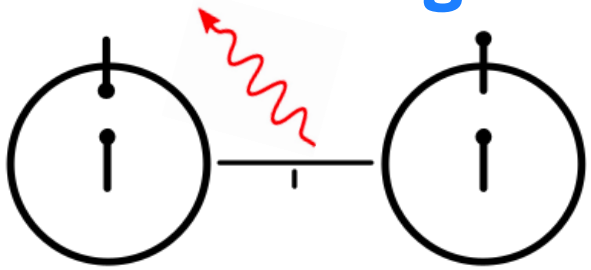
David Prelogović, Andrei Mesinger

Scuola Normale Superiore

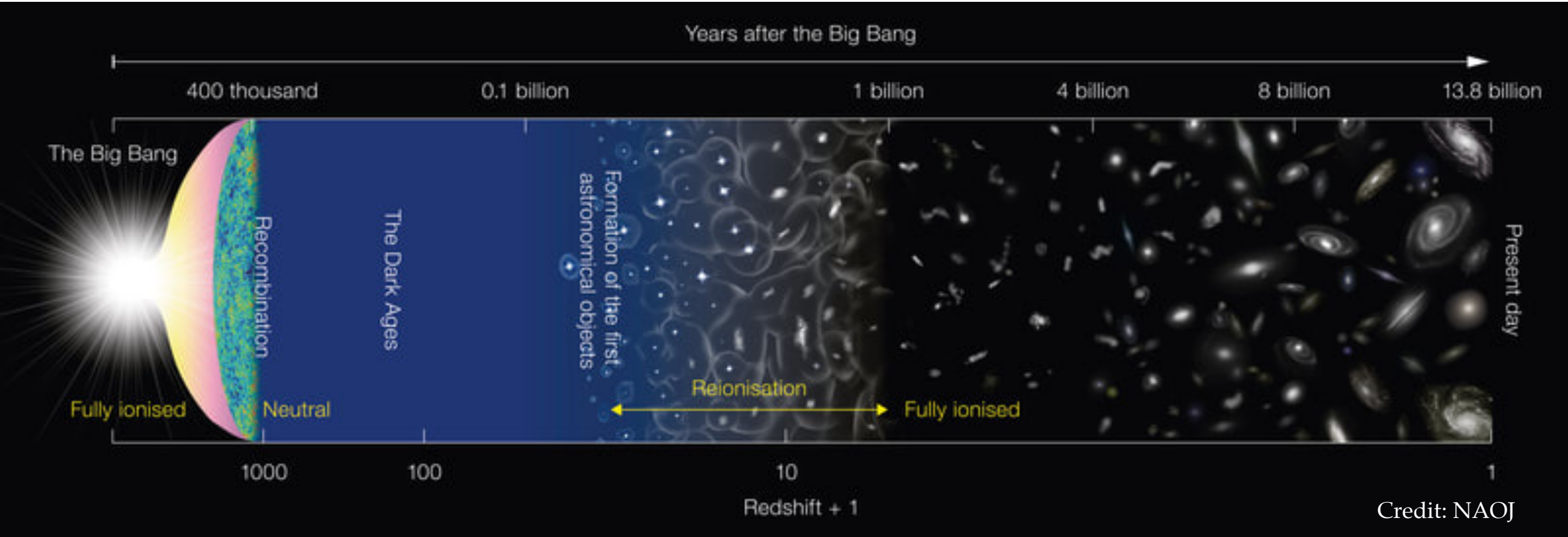
Spoke 3 Technical Workshop, Trieste October 9 / 11, 2023

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Cosmic 21-cm signal

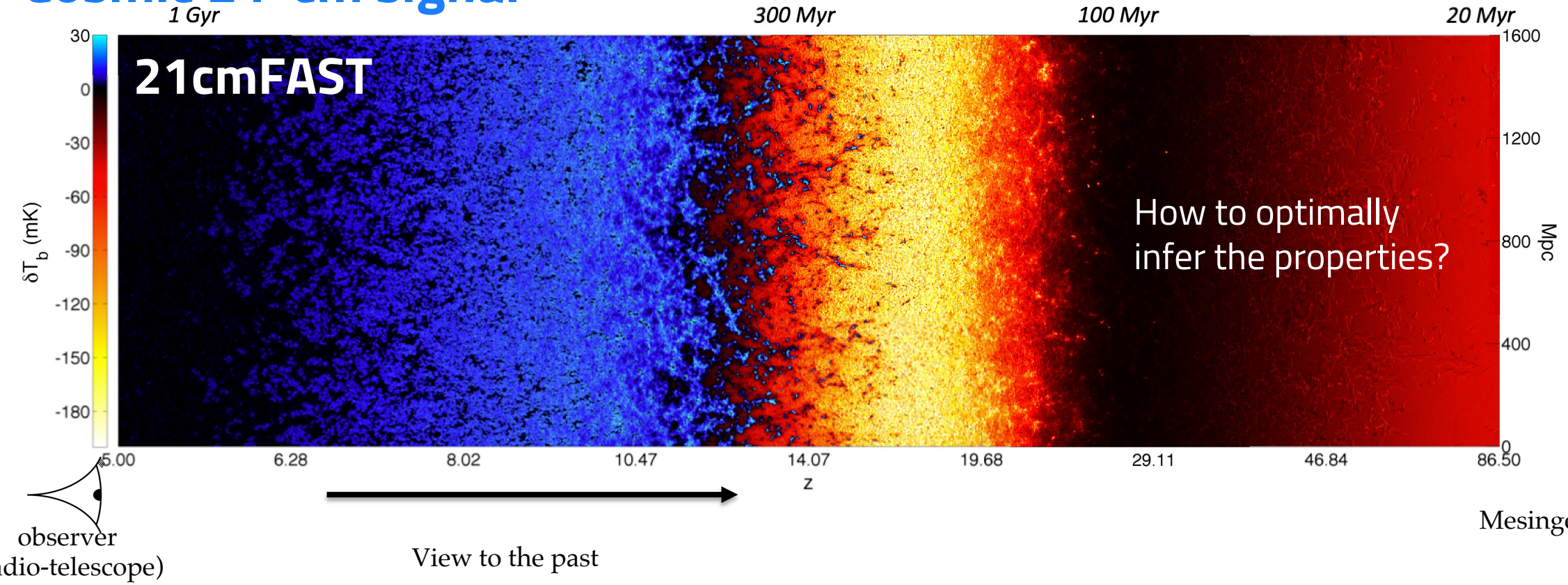


- Over 90% of the "normal" matter in the Universe is ***hydrogen***



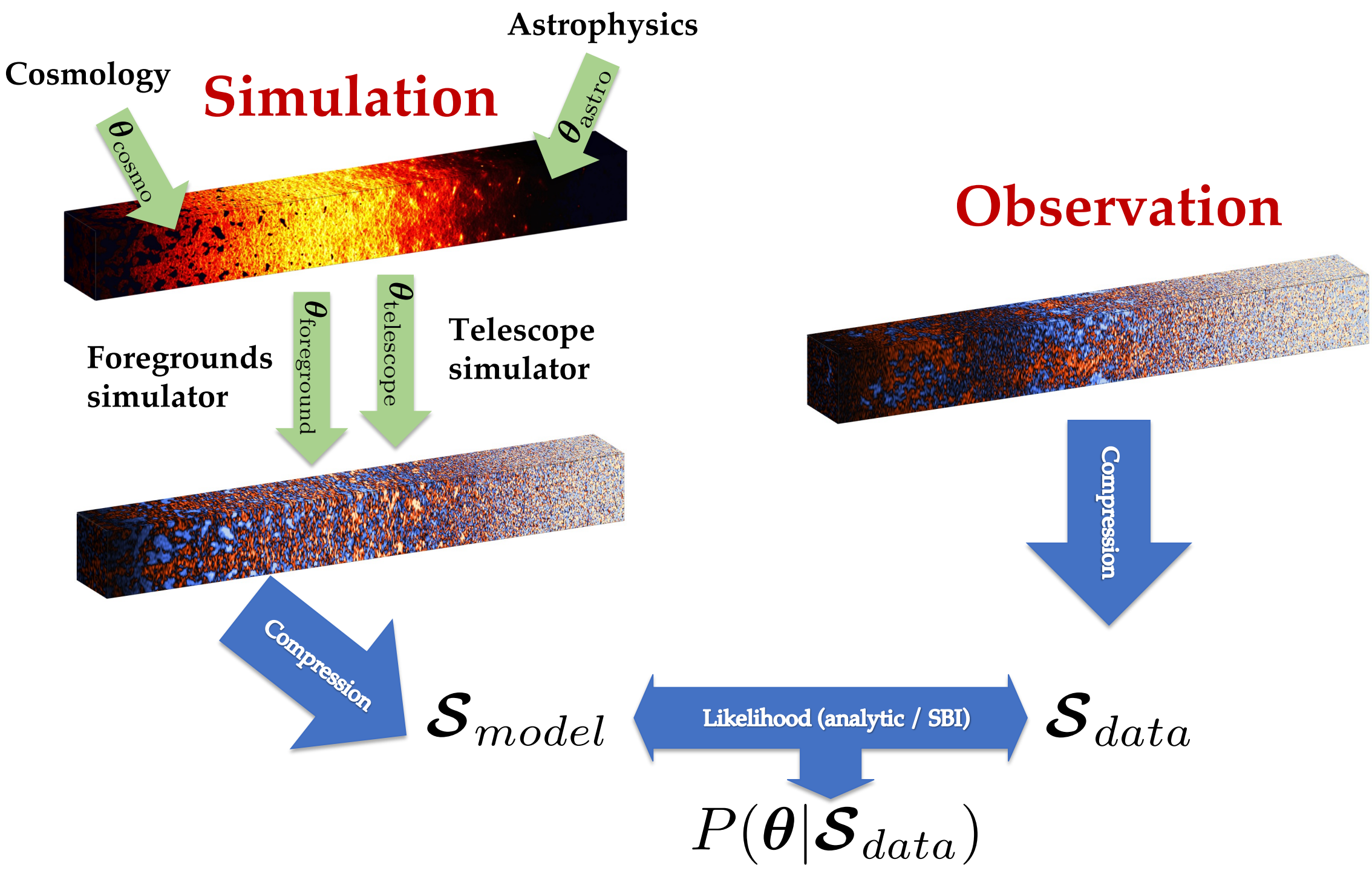
Credit: NAOJ

Cosmic 21-cm signal



Mesinger+2016

- The properties of the ***unseen first galaxies*** are encoded in the timings and patterns of this signal!


Astrophysics
Cosmology
Simulation
$\theta_{\rm cosmo}$
$\theta_{\rm astro}$
Observation
$\theta_{\rm foreground}$
$\theta_{\rm telescope}$
Foregrounds simulator
Telescope simulator
Compression
Compression
$\mathcal{S}_{model}$
Likelihood (analytic / SBI)
$\mathcal{S}_{data}$
$P(\theta|\mathcal{S}_{data})$

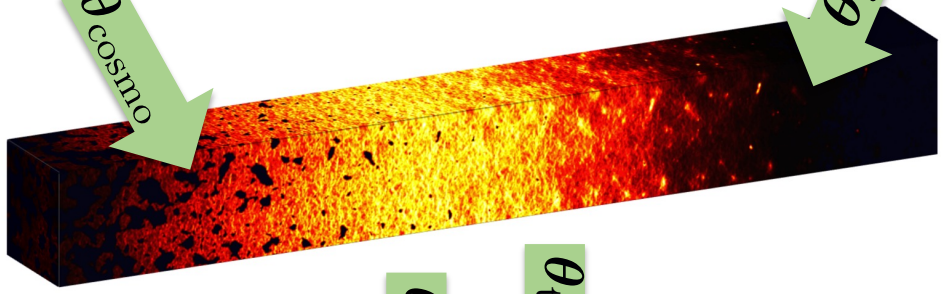
Astrophysics

Cosmology

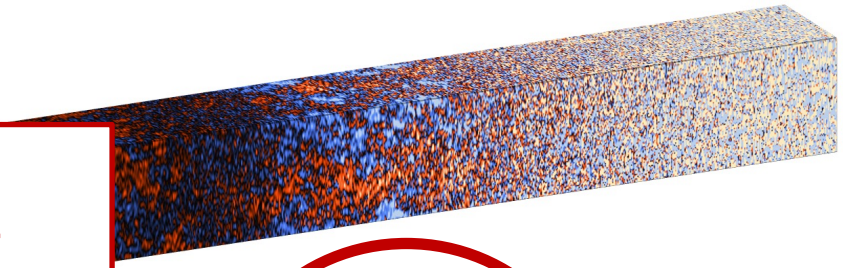
Simulation

θ_{cosmo}

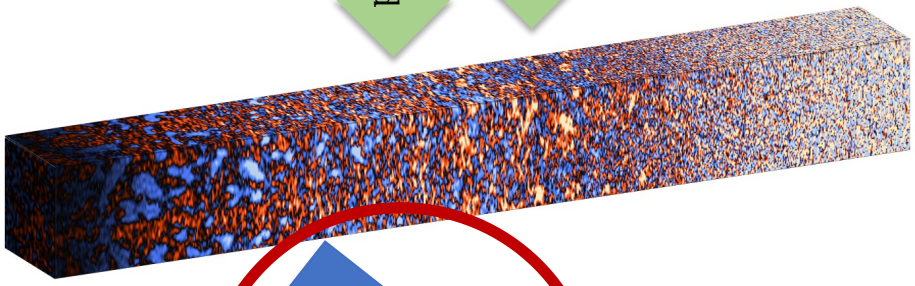
θ_{astro}



Observation



#1 – optimal compression



Compression

Compression

$\boldsymbol{\mathcal{S}}_{model}$

Likelihood (analytic / SBI)

$\boldsymbol{\mathcal{S}}_{data}$

$P(\boldsymbol{\theta}|\boldsymbol{\mathcal{S}}_{data})$

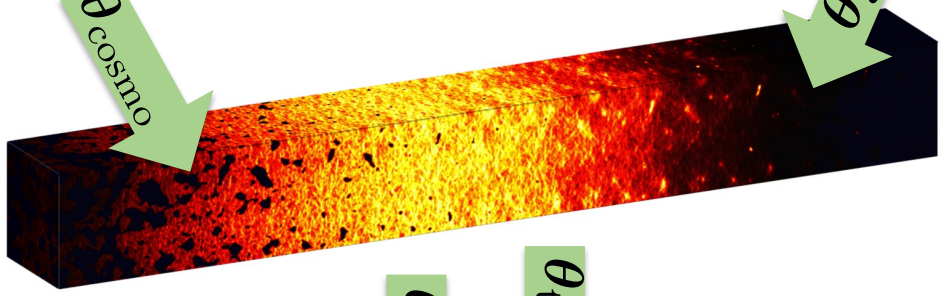
Astrophysics

Cosmology

Simulation

θ_{cosmo}

θ_{astro}



Observation

#1 – optimal compression

#2 – "optimal" likelihood

Compression

Compression

$\boldsymbol{s}_{model}$

Likelihood (analytic / SBI)

$\boldsymbol{s}_{data}$

$P(\boldsymbol{\theta}|\boldsymbol{s}_{data})$

Scientific Rationale

- The 21-cm signal has non-Gaussian formation with unknown likelihood
- Future telescopes such as SKA will be able to record images of the signal
- Advanced Inference techniques are necessary:
  I. to extract as much information as possible form the signal
  II. to reduce computational cost of classical inference techniques

Main objectives:

#1 – finding the optimal compression of the signal

#2 - ”writing” the likelihood of the signal – Simulation-Based Inference

Simulation-Based Inference for the 21-cm PS

Likelihood of the 21-cm

- Analytically not available
- Numerically non-tractable

Likelihood of the 21-cm

- Analytically not available
- Numerically non-tractable
- Eg. – power spectrum

$$\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{\mu}; \mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\Pi_i \sigma_i^2}} e^{-\frac{1}{2}\sum_i (x_i - \mu_i)^2/\sigma_i^2}$$

$$\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu}; \mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Likelihood of the 21-cm

- Analytically not available
- Numerically non-tractable
- Eg. – power spectrum

$$\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{\mu}; \mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\Pi_i \sigma_i^2}} e^{-\frac{1}{2}\sum_i (x_i - \mu_i)^2/\sigma_i^2}$$

$$\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu}; \mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Classical inference simplifications

1) μ estimated from one sim.
2) $\Sigma = \Sigma_{\theta \text{fid}}$
3) No higher order terms

–> Simulation-Based Inference

Simulation-Based Inference

- Pull from the prior $\boldsymbol{\theta}^* \sim P(\boldsymbol{\theta})$
- Pull from the likelihood

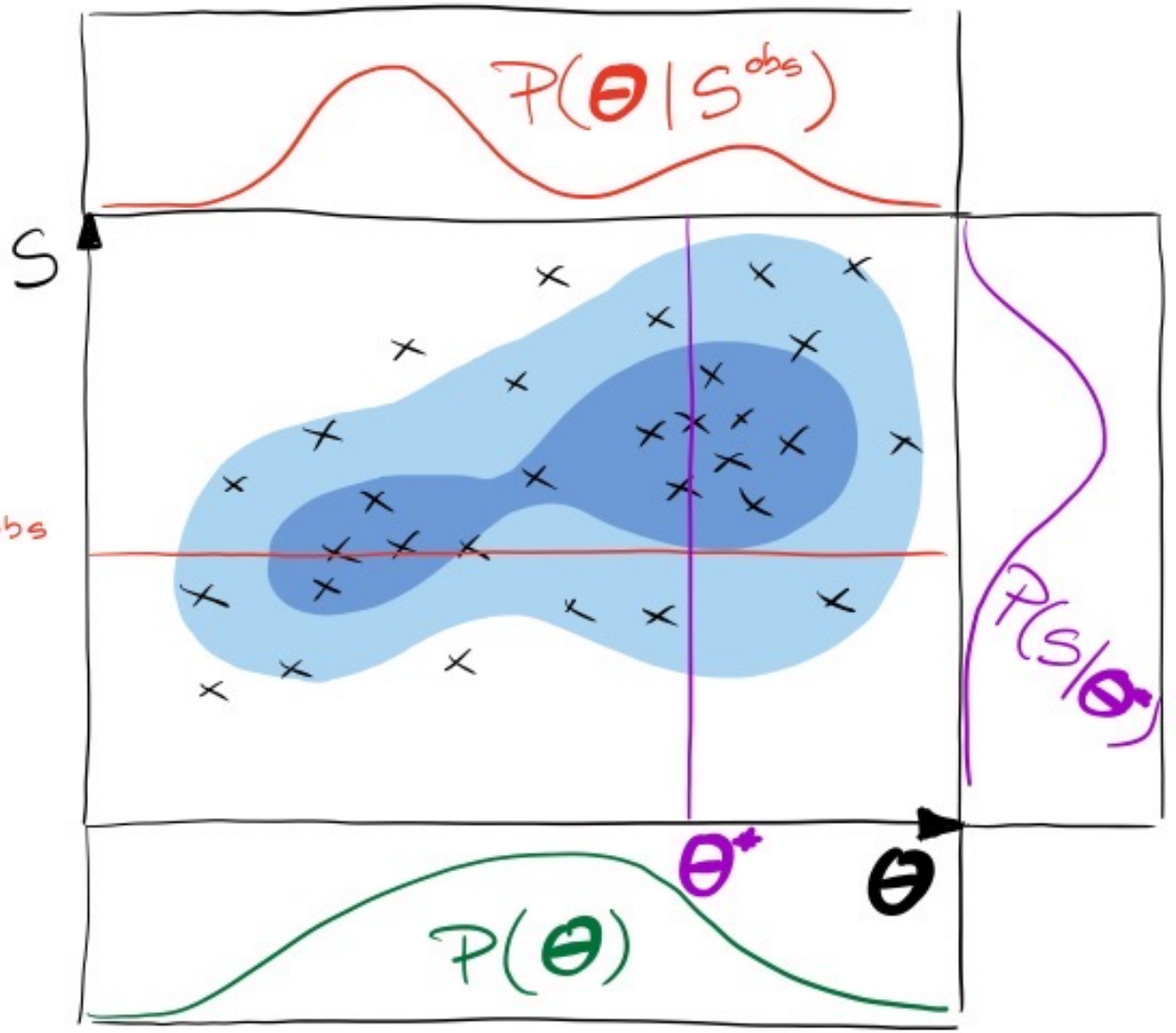
$$\mathcal{S}^* \sim P(\mathcal{S}|\boldsymbol{\theta}^*)$$

i.e. simulate and compress

$$\mathcal{S}^* = \text{compress}(\text{simulate}(\boldsymbol{\theta}^*))$$

- Repeat many times
- Fit the distribution with

Neural Density Estimators (NDE)



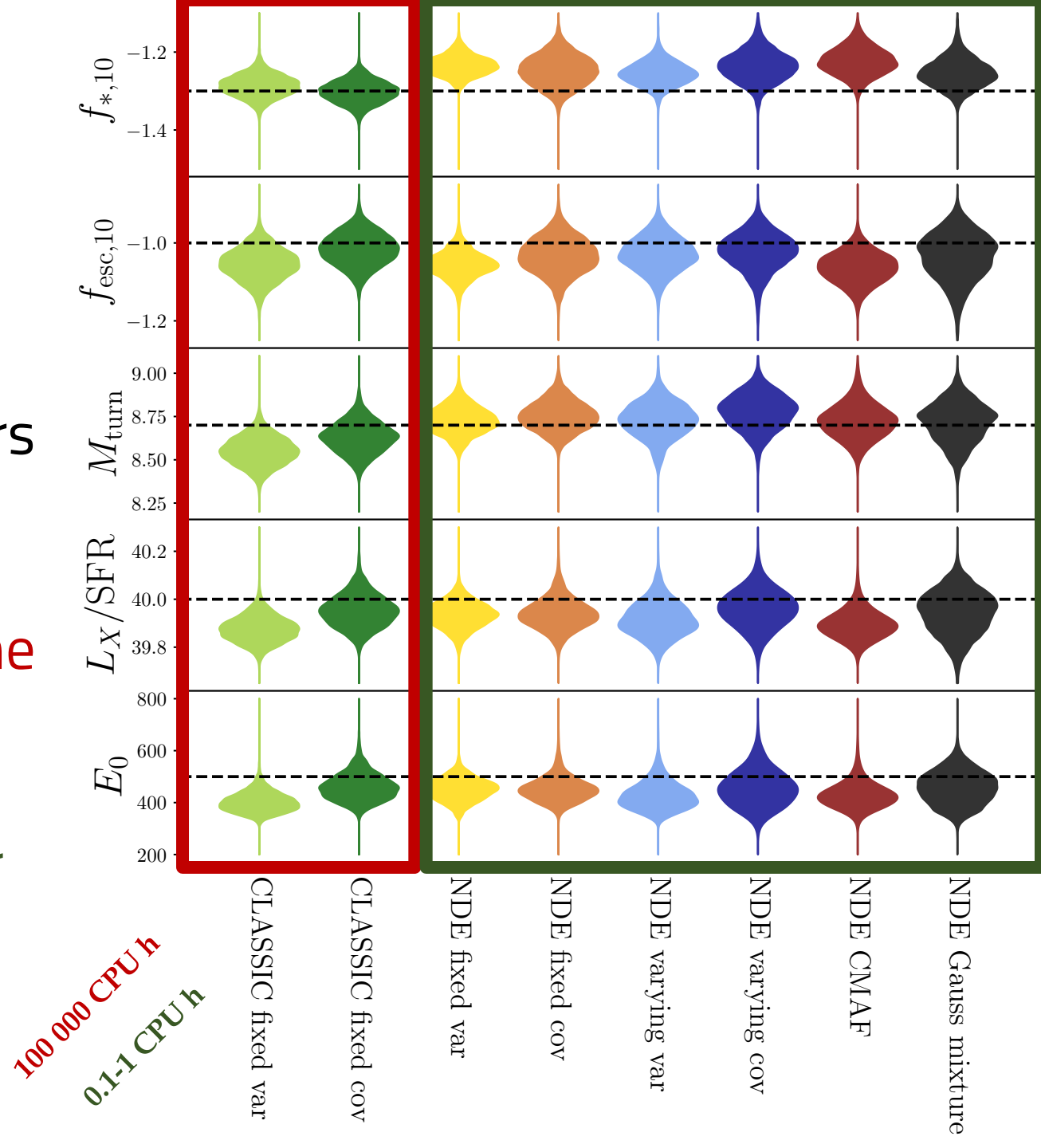
Credit: Tom Charnock

SBI for the 21-cm PS

- Able to infer the parameter posteriors
- At a fraction of computational cost

- Caveat: large database needed for the NDE training!

- Once constructed, reusable for other inferences, other summaries



Prelogović+2023

Optimal Compression of the 21-cm lightcones

Ad-hoc compressions

- Without good a-priori physical motivation, we cannot know what is THE optimal compression/summary, i.e. providing tightest recovery of astrophysics and cosmology

Ad-hoc compressions

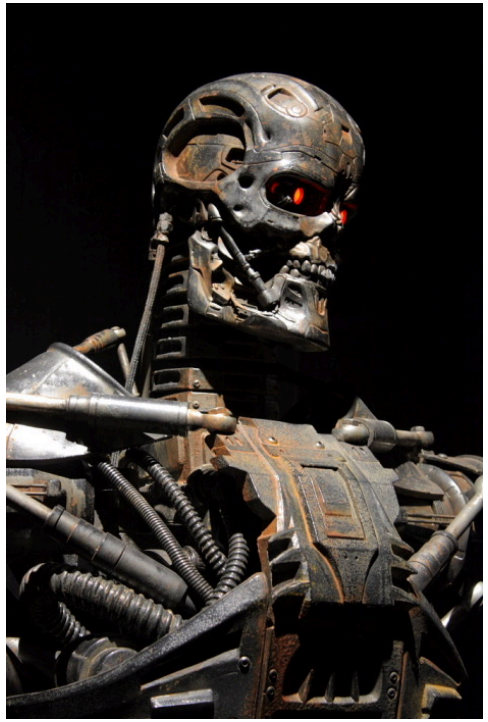
- Without good a-priori physical motivation, we cannot know what is THE optimal compression/summary, i.e. providing tightest recovery of astrophysics and cosmology

Solution:

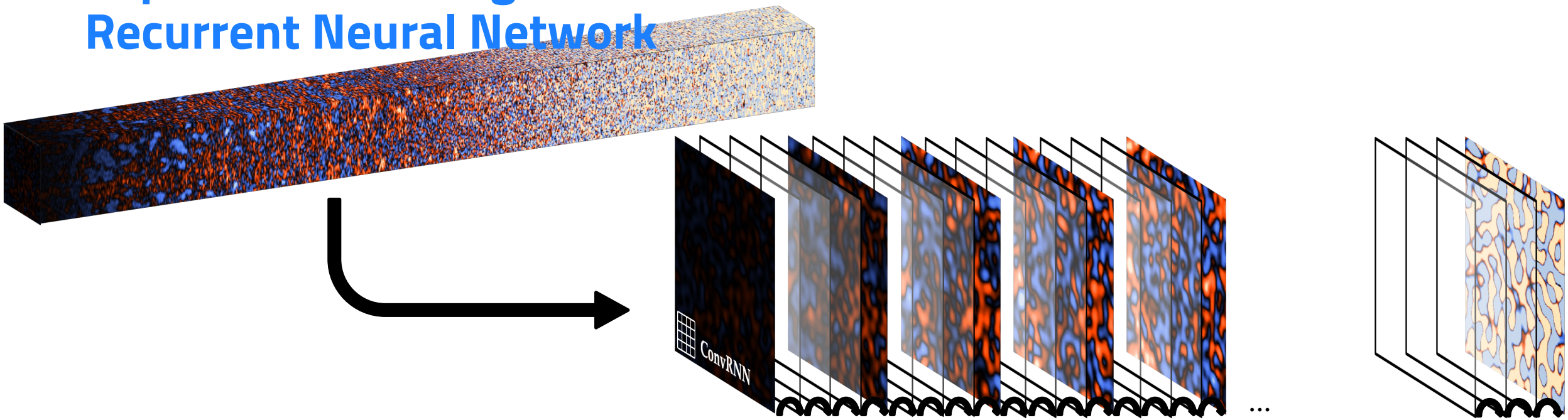
Let the machines figure it out for us!

(Neural Network)

- Gillet+2018
- La Plante & Ntampaka 2019
- Makinen+2020
- Mangena+2020
- Hortúa+2020
- Prelogović+2021
- +++



Supervised learning: Recurrent Neural Network



- RNN – encoding correlations across all frequency bins at once
- 2D Convolutional NN– local correlations in sky-plane

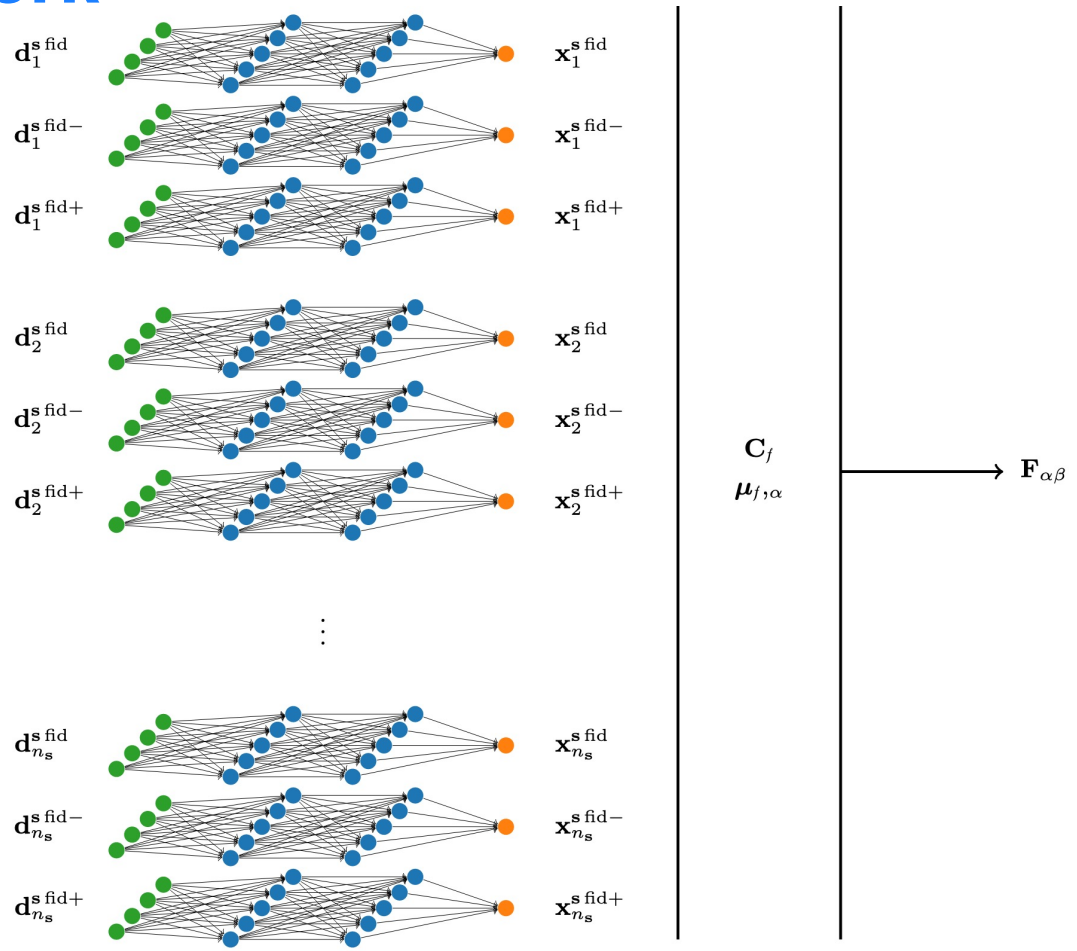
Unsupervised learning: Information Maximizing Neural Network

- Simulate the data at a fiducial parameter set: $\mathbf{d}(\boldsymbol{\theta}_{\rm fid})$
- Simulate around the fiducial parameters: $\mathbf{d}(\boldsymbol{\theta}^+_{\rm fid}), \mathbf{d}(\boldsymbol{\theta}^-_{\rm fid})$
- Calculate compressed summary:

$$\mathbf{s}(\boldsymbol{\theta}) = \mathrm{NN}(\mathbf{d}(\boldsymbol{\theta}))$$

- Maximize Fisher information:

$$L = -\ln|F|$$



Methods and results

Database:

- for ~150 points pulled from the prior, construct finite—difference "Fisher database"
  –> 50 000 samples

Methodology:

- distribution of Fisher information (ln|F|) as a proxy for summary quality

Summaries:

- RNN & IMNN for deep learning summaries
- Wavelets, 1DPS and 2DPS for classical summaries

Methods and results

Database:

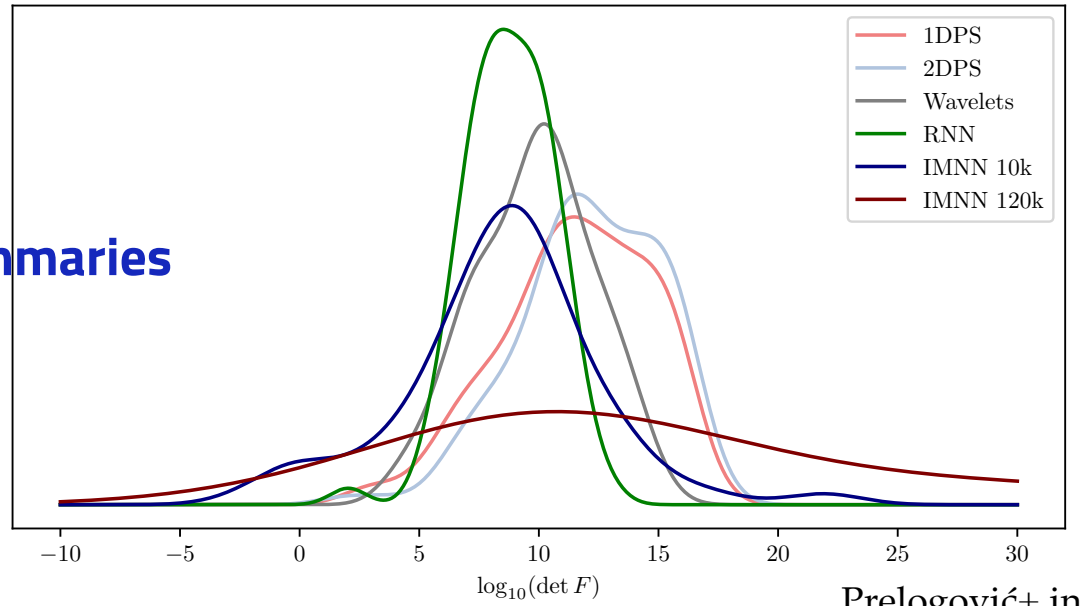
- for ~150 points pulled from the prior, construct finite—difference "Fisher database"
  –> 50 000 samples

Methodology:

- distribution of Fisher information (ln|F|) as a proxy for summary quality

Summaries:

- RNN & IMNN for deep learning summaries
- Wavelets, 1DPS and 2DPS for classical summaries



Prelogović+ in.prep.

Summary

Milestones and KPIs

Bottlenecks:

Construction of databases:
- SBI database ~100 000 21cmFAST simulations
- IMNN database ~ 20 000
- Fisher database ~ 50 000

} CPU resources

NN trainings:
- IMNN – 5 000 GPU h per model
- RNN – 1 000 GPU h per model
- SBI NDEs – relatively cheap (~few GPUh per model)

} GPU resources

Code release:
- two publicly available Python libraries:
  - -> 21cmLikelihoods – SBI library for exploring 21cmPS likelihoods
    https://github.com/dprelogo/21cmLikelihoods
  - -> conditionalKDE – helper library for prior sampling
    https://github.com/dprelogo/conditional_kde

Publications:

Exploring the likelihood of the 21-cm power spectrum with simulation-based inference
arXiv:2305.03074

Next Steps and Expected Results

Code release:

summaries21cm: library for exploration of different summaries of the 21cm signal

Publications:

How informative are summaries of the cosmic 21-cm signal?

Database:

start a construction of a 9-dim database, for multi-wavelength SBI (talk tomorrow)

(granted a fraction of CPU resources)

Projects:

constrained realizations of the 21-cm signal from galaxy probes

(granted small Cineca C call – Leonardo, more resources needed)