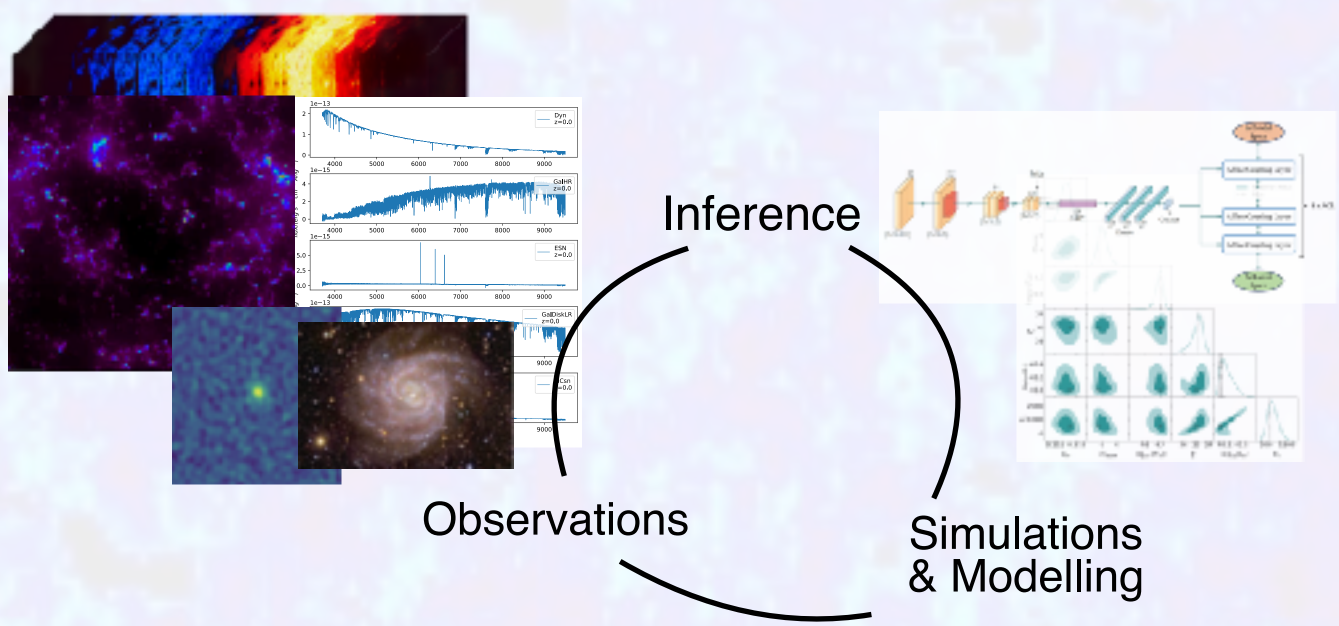
Machine Learning for Astrophysics



Caroline Heneka, ITP Heidelberg

Group ‘Computer Vision Astrophysics and Cosmology’

Advanced School on Applied Machine Learning, ICTP Trieste, May 28th 2024

Who am I? Caroline Heneka
Institute for Theoretical Physics, Heidelberg University

- B.Sc. and M.Sc. Physics in Heidelberg (+ Erasmus NPAC Paris)
- Ph.D. (2017) at Copenhagen University,
  DARK Cosmology Centre, Niels Bohr Institute
- 2017-2019: DFG Transregio 33 Fellowship (Heidelberg), Postdoc SNS Pisa
- ca. 1.5 yrs: DLR (German Aerospace Center) Headquarters Cologne,
  Executive Board Area Space, Programme Strategy Space
- 2020-2022: Senior Postdoc Hamburg University

- Since Oct 2022: back in HD
  Junior Group Leader & Freigeist (Volkswagen Foundation) Fellow
  'Computer Vision Astrophysics and Cosmology'

My **Research Interests**

- **Line Intensity Mapping**
  (also: radio galaxy clustering, cross-correlation studies, galaxy clusters, ..)

- **High-redshift astrophysics & Epoch of Reionization**:
  Modelling of 21cm background and further high-redshift lines (Lya, Ha, ..)

- The modern machine learning toolkit with '**Computer Vision Astrophysics + Cosmology**',
  specifically for intensity mapping of large-scale backgrounds:
  - Emulation, generation
  - Inference

  Also:
    - Classification (e.g. 4MOST spectra)
    - Detection (SKA preparation)

Simulations
Computational
Astrophysics
Reionisation

Data
Imaging
Spectral
3D cubes
Time series

Bayesian
Bias detection
Model comparison
Hierarchical

Modelling/
Theory
IM
Cosmological
signatures
Sources

Statistical Tools

Forecasting
Fisher

Sampling &
Inference
MCMC

Maps
Trees

Neural
Networks
Generative
Classification
Detection
Inference

Meta-
Heuristics
Genetic
algorithm

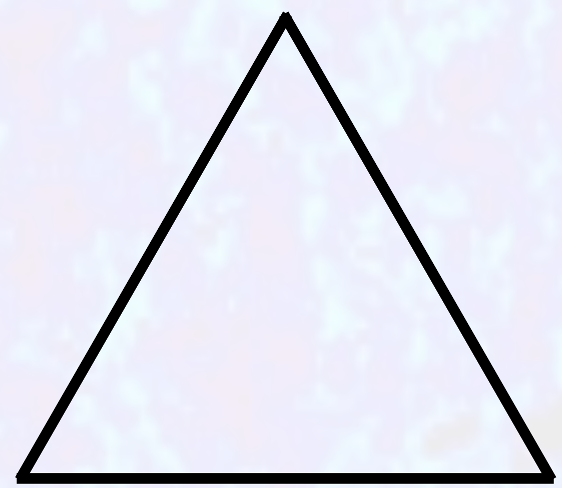
Machine learning

How will Astronomy and Astrophysics advance in coming years?

Understanding

Learn about evolution of structures

across cosmic time and scales



Survey Data

Large-scale structure

High-resolution

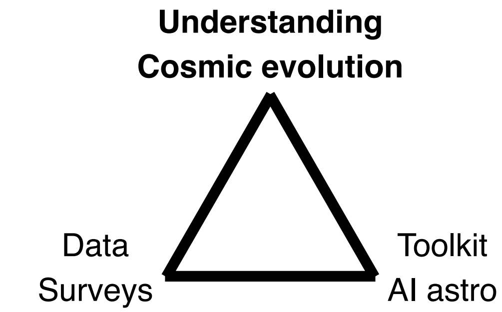
Data Science

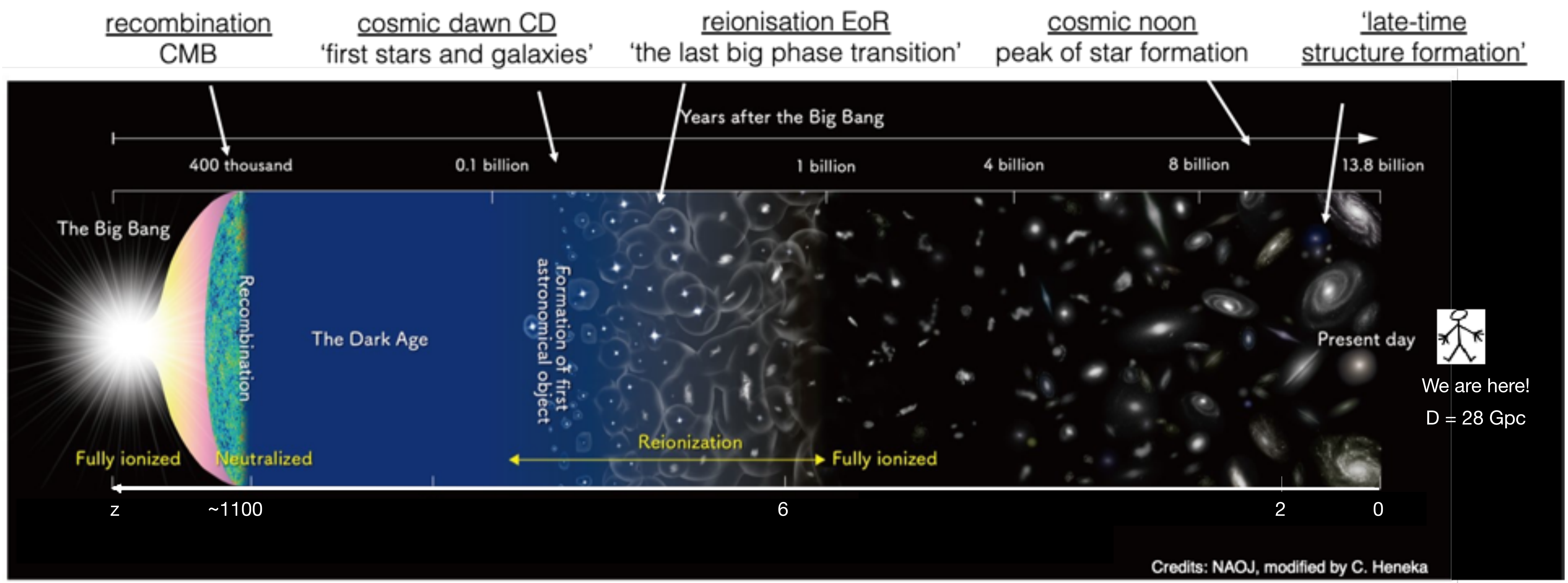
‘ML + AI Astrophysics’

Data Mining

Computational + Simulations

Where we stand: Data revolution and cosmic evolution

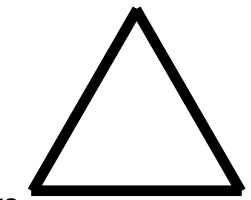


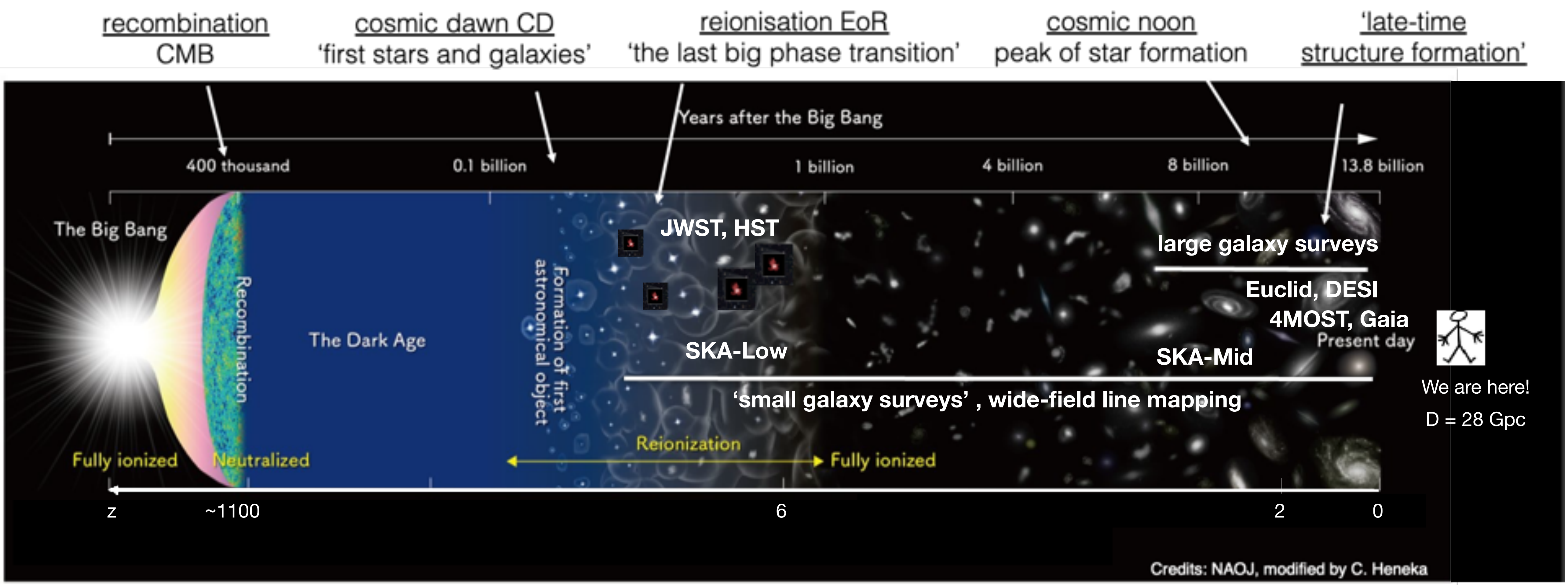


Our goal:

Learn about astrophysical & cosmological evolution across cosmic time and scales

Where we stand: Data revolution and cosmic evolution





Our goal:

Learn about astrophysical & cosmological evolution across cosmic time and scales

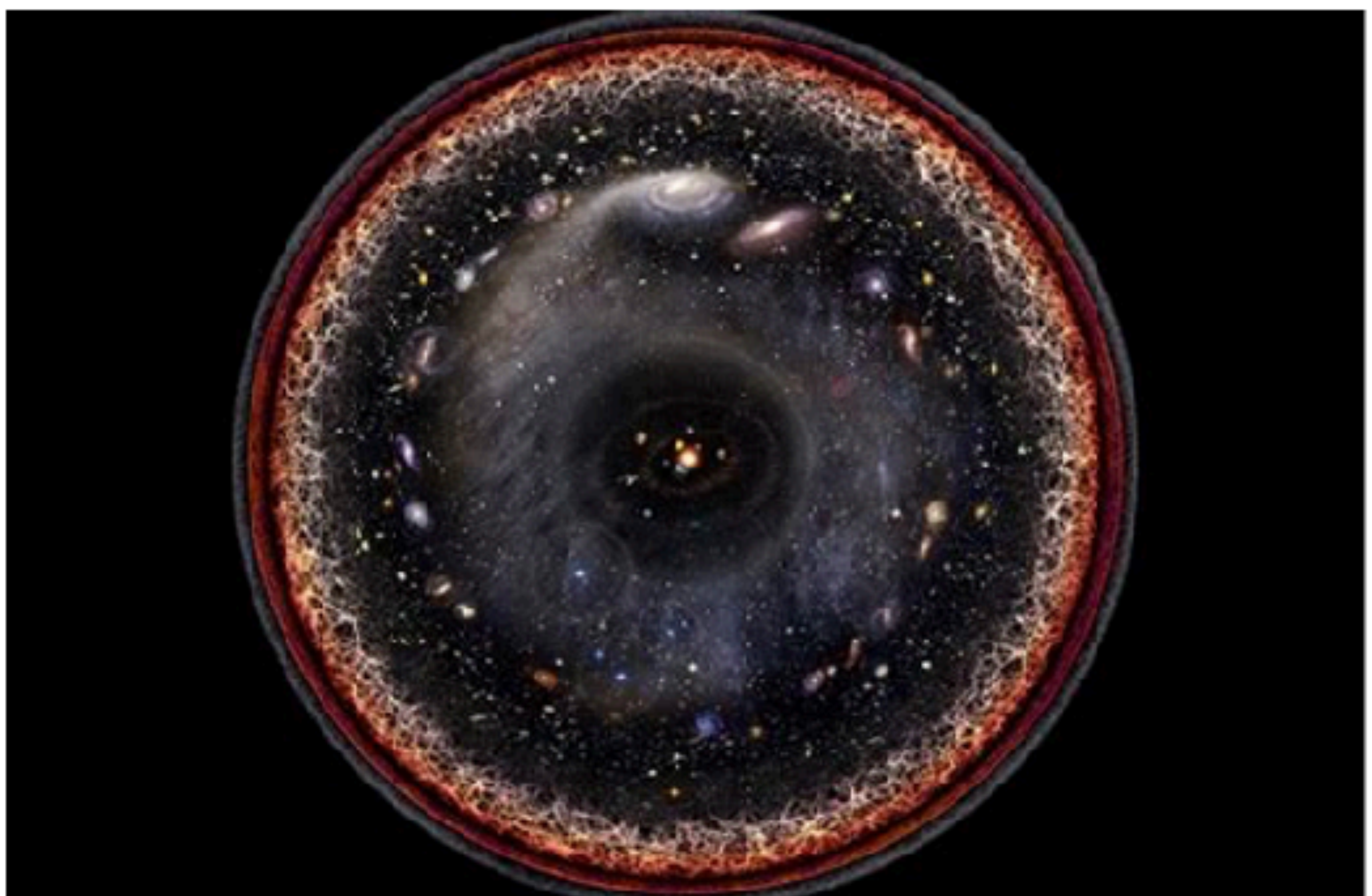
Coming decade: push to map up to **80% of the observable Universe**

… what does 80% of the observable Universe mean?

Modelling challenges

True LSS probes ⟶ orders of magnitude of scales up to the ultra-large

…what does 80% of the observable Universe even mean?



APOD, NASA, License & Credit: Wikipedia, Pablo Carlos Budassi

Observable Universe:

d ~ 28 Gpc (x3 Glyr)

80% if this:

d ~ 22 Gpc

Let's say we resolve (only) ~Mpc

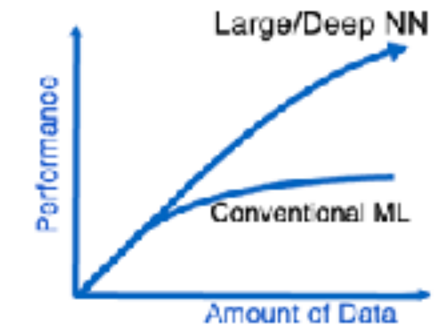
⟶ about 3-4 orders of magnitude

⟶ about 10^9-10^{10} modes!

… at some point we sub-grid model and/or change modelling approach

Advances in Data Science, for Science

Deep Learning
Driven by ability to improve
with large datasets



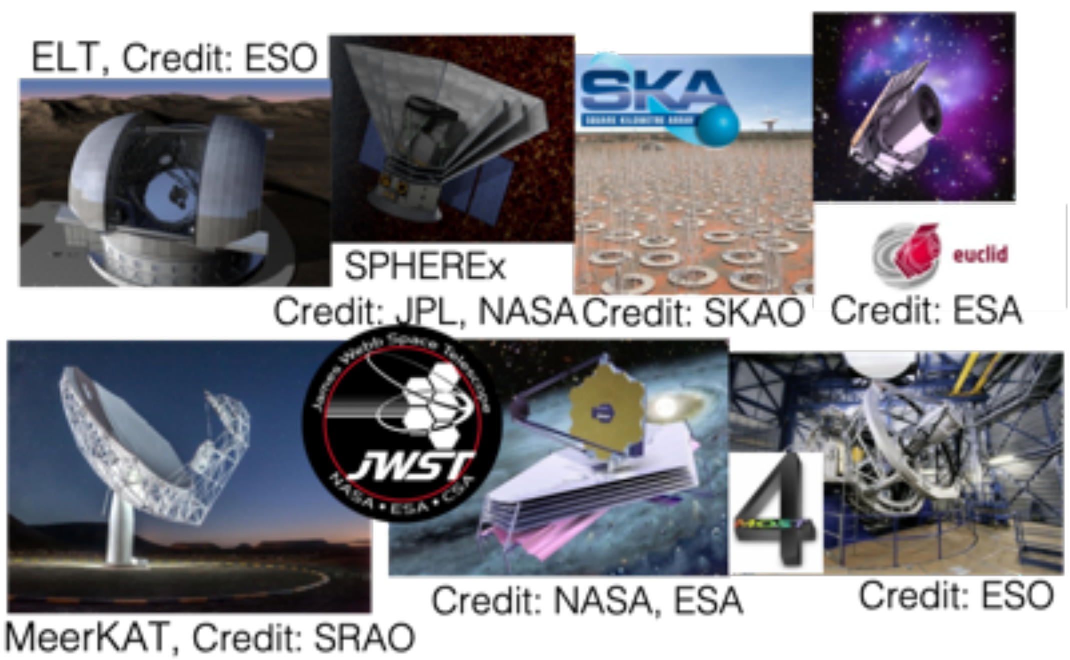
One big goal? Learn about astrophysical & cosmological evolution across cosmic time and scales

No, really - many goals, …

… and instruments

… types of data

… datarates, scales, signal-to-noise



** non-comprehensive list*

Large-scale surveys

High-resolution imaging and/or spectroscopy

Efficient data reduction
Automation

Extract more & less biased information
Data mining

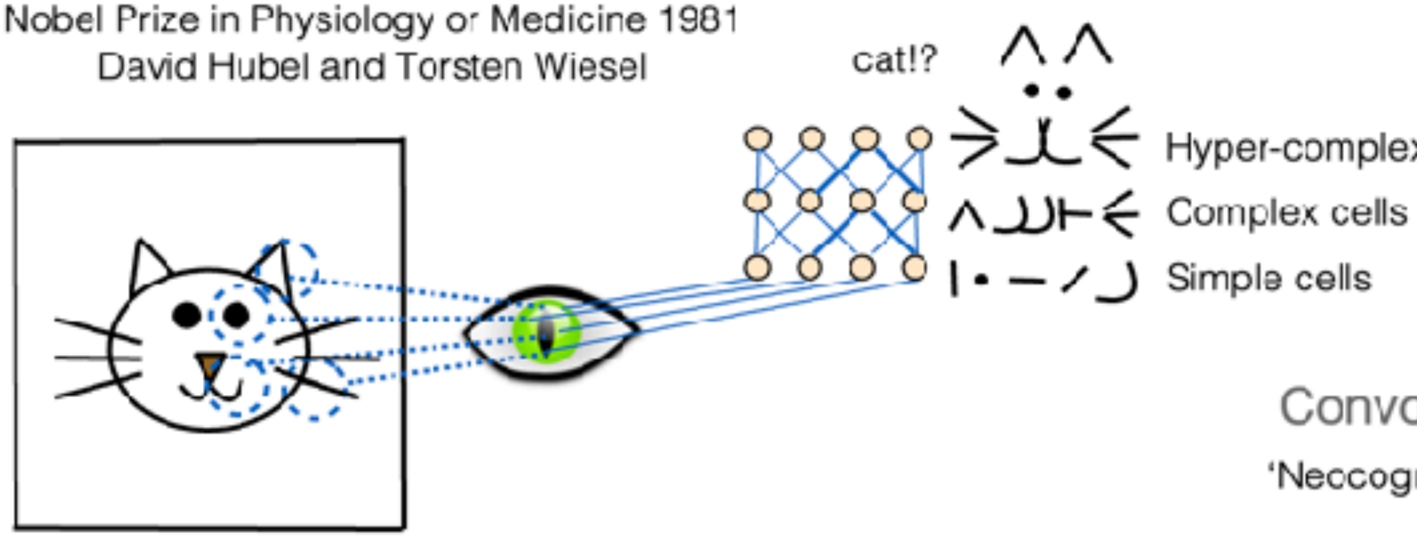
Why ML/DL for Astrophysics?

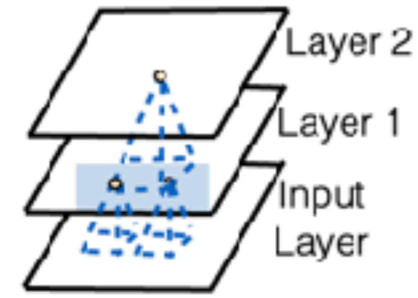
https://braintour.harvard.edu/archives/portfolio-items/the-first-neurobiology-department

Neuron response to visual stimuli

Nobel Prize in Physiology or Medicine 1981
David Hubel and Torsten Wiesel

Representation learning





Convolutional neural network

'Neocognitron' Kunihiko Fukushima 1979

Hierarchical learning

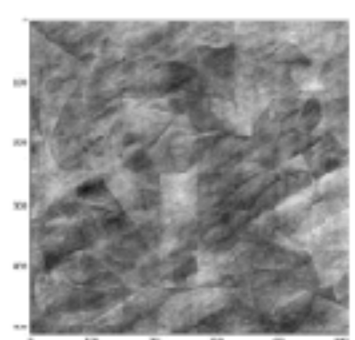
Non-linear, non-Gaussian

High-dim. correlations

duck

size
color
bkg

randomise phases



vs. "The famous Gaussian duck"

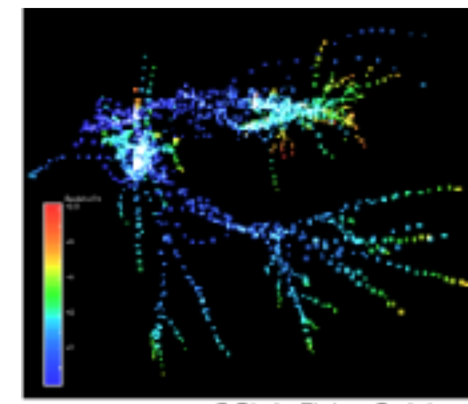
Same 2D power spectrum

Where we stand: ML and AI for Astronomy and Astrophysics

We need a versatile ML/AI toolkit, for:

- detection, segmentation
- classification
- regression, inference
- anomaly detection
- generation, emulation, …

Hierarchical

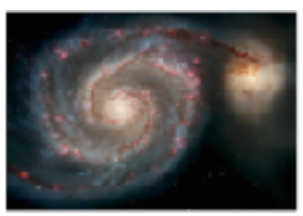


@Chris Fluke, Swinburne University of Technology

+ plenty of inverse problems

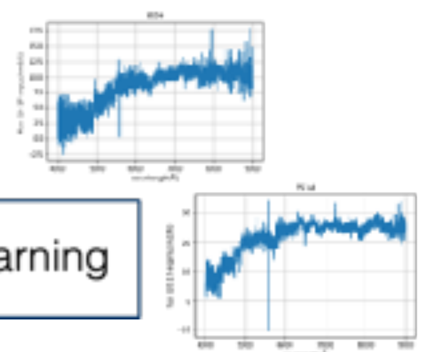
$$I^D(x, y) = R \times I(x, y) + n$$

@Hubble, NASA



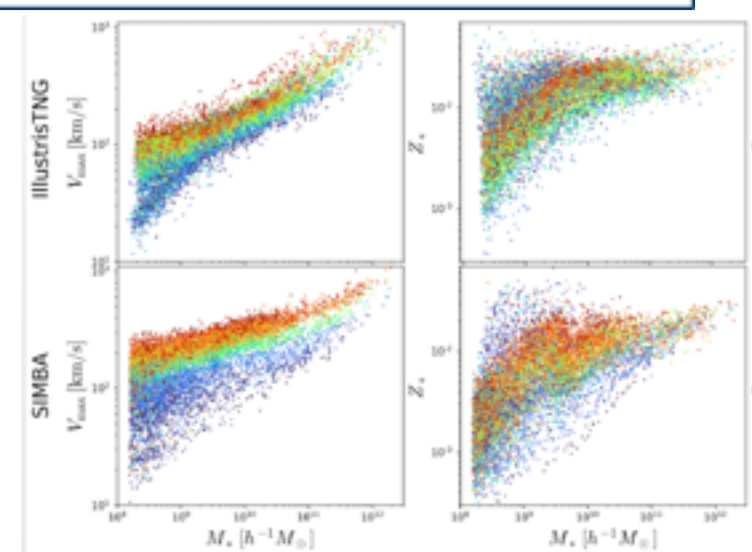
@Hubble, NASA

Representation learning

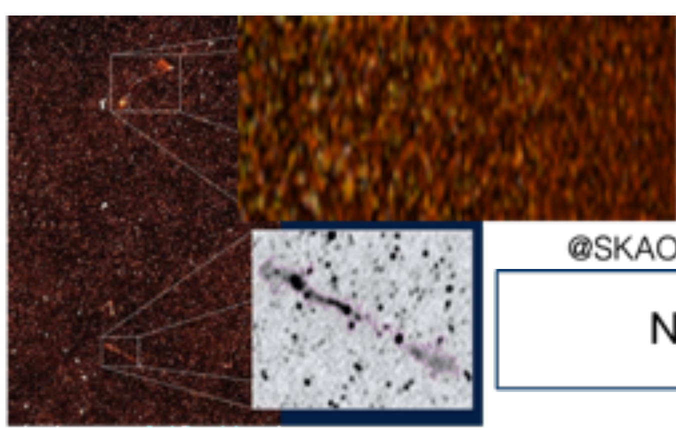


@SDSS

High-dim. correlations



arXiv:2201.02202



@SKAO

@MeerKAT

Non-linear, non-Gaussian

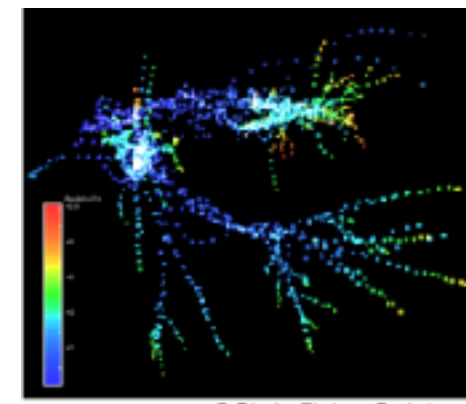
Where we stand: ML and AI for Astronomy and Astrophysics

We need a versatile ML/AI toolkit, for:

- detection, segmentation
- classification
- regression, inference
- anomaly detection
- generation, emulation, …

Examples:
Few sec: Classification 40.000 spectra
Few sec: 7-parameter inference ~100MB cube
Few sec: detection, segmentation & flux measurement O(100-1000) sources

Hierarchical

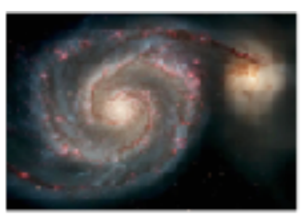


@Chris Fluke, Swinburne University of Technology

+ plenty of inverse problems

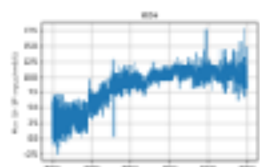
$$I^D(x, y) = R \times I(x, y) + n$$

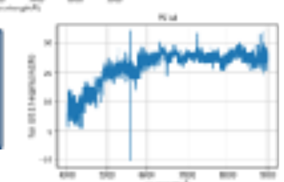
@Hubble, NASA



@Hubble, NASA

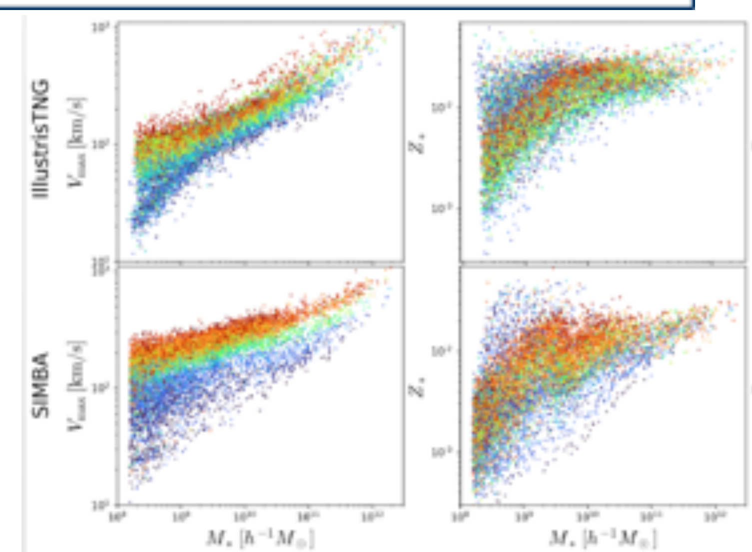
Representation learning



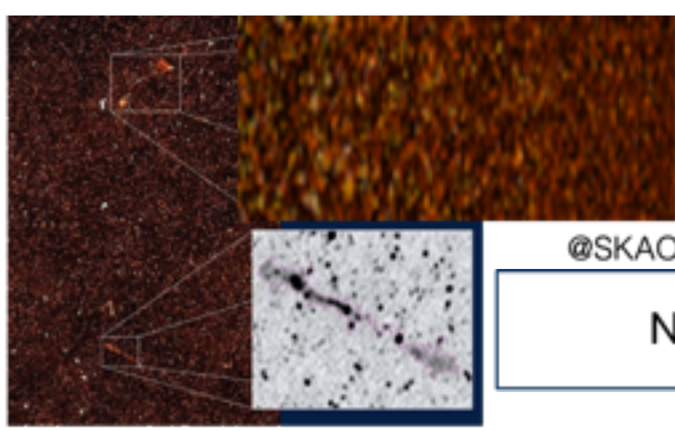


@SDSS

High-dim. correlations



arXiv:2201.02202



@SKAO

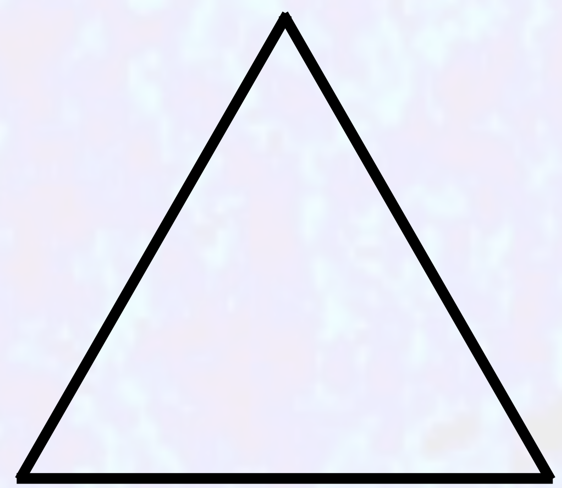
@MeerKAT

Non-linear, non-Gaussian

How will Astronomy and Astrophysics advance in coming years?

Understanding

Learn about evolution of structures
across cosmic time and scales



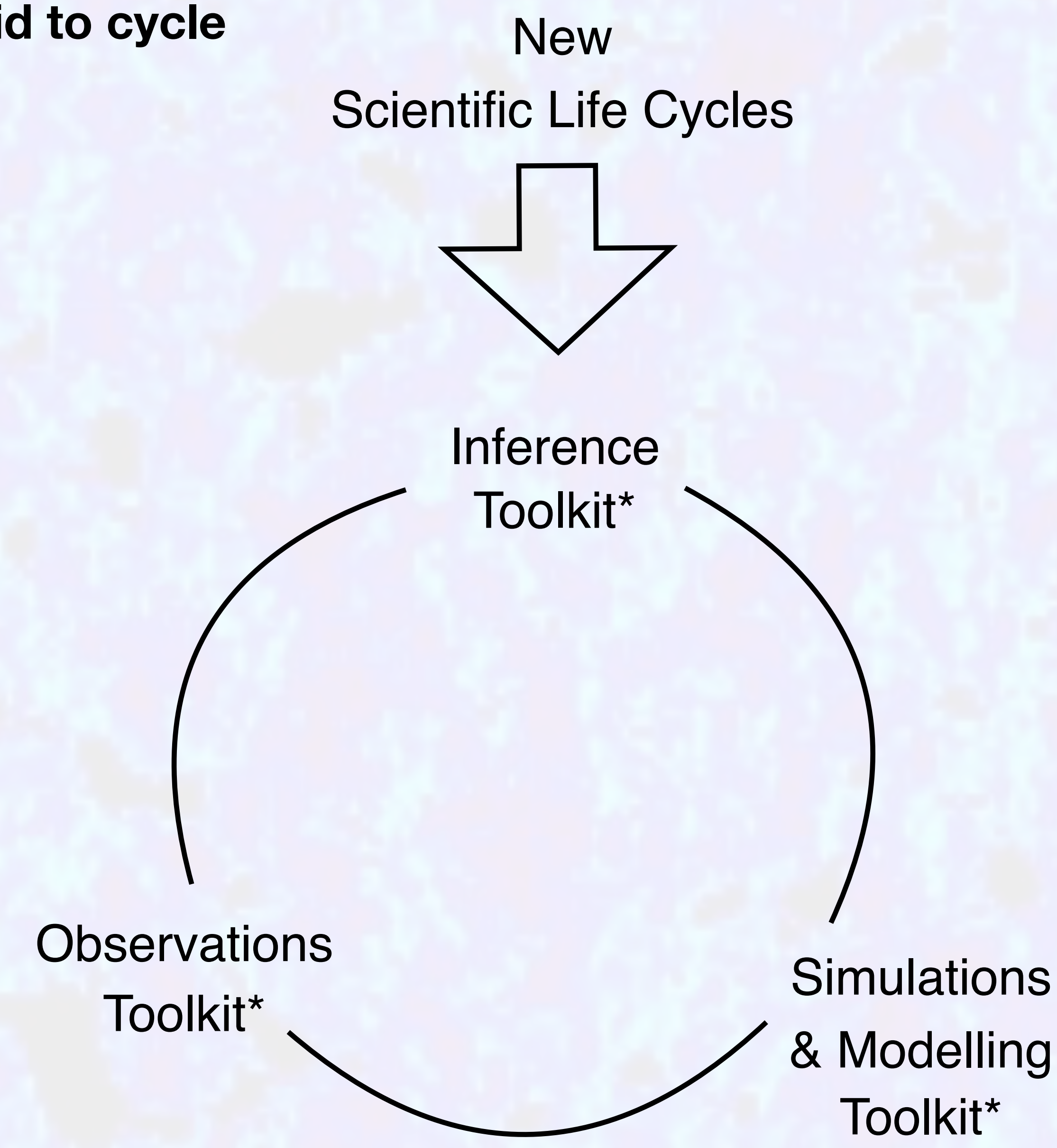
Survey Data

Large-scale structure
High-resolution

Data Science

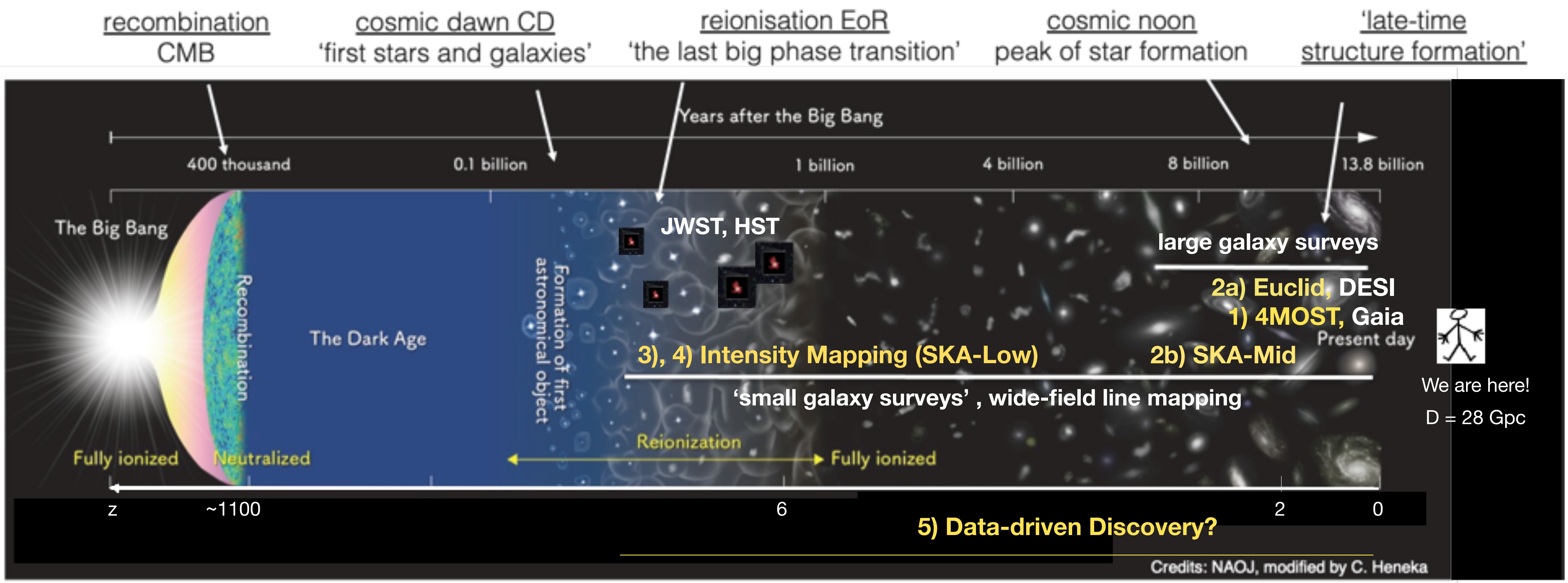
'ML + AI Astrophysics'
Data Mining
Computational + Simulations

In practice: From pyramid to cycle



*The Toolkit: Statistics, Machine & Deep Learning, Data Mining

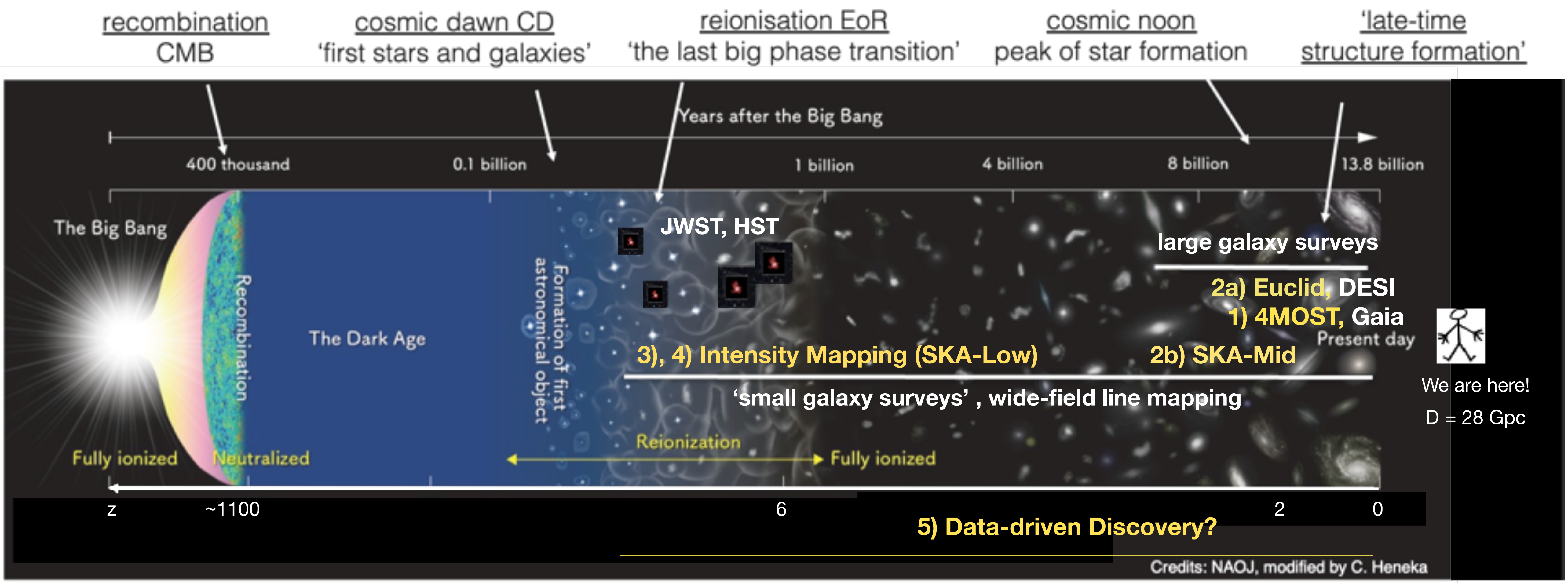
Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

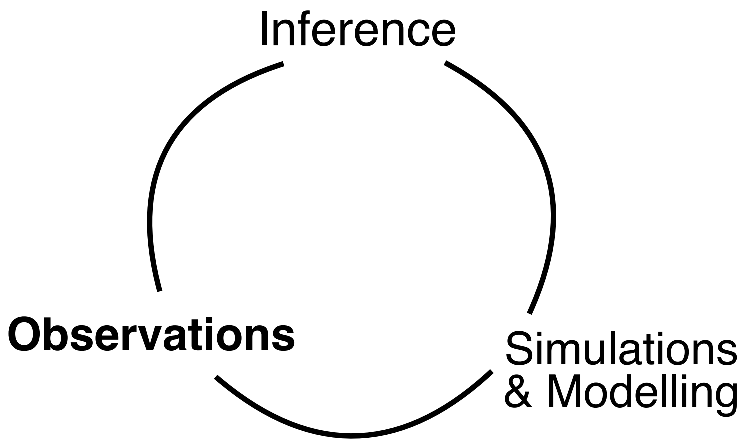
1) Classification / Triggering
2) Source detection & characterisation
3) Simulation-based inference (SBI) in 3D
4) Generative methods
5) Data-driven Discovery

Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

1) **Classification / Triggering**
2) Source detection & characterisation
3) Simulation-based inference (SBI) in 3D
4) Generative methods
5) Data-driven Discovery



1) Classification and triggering for large astronomical surveys

4MOST: On-the-fly classification of spectra (1D)

- 5-year survey
- wide-field, fibre-fed, optical spectroscopy
- on ESO's 4-m-class telescope VISTA
- 2.5-degree diameter field-of-view, 2436 fibres
- HRS R ≈ 18000 – 21000, LRS R ≈ 4000 – 7500
- 20mio. (LRS), 3mio. (HRS) sources

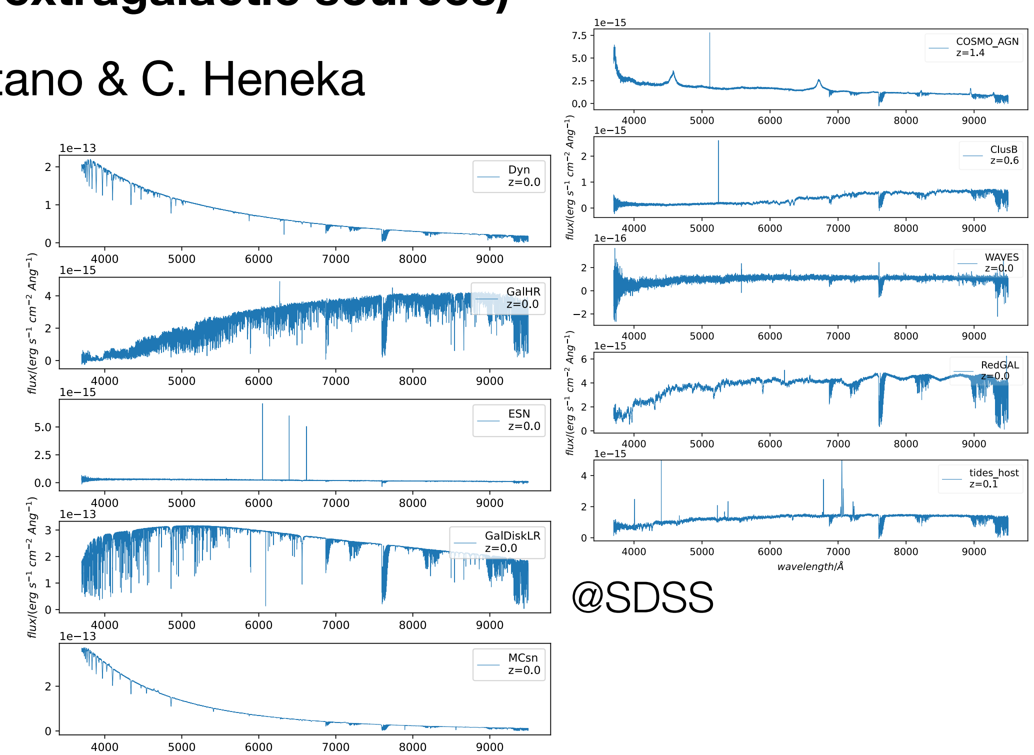
https://www.4most.eu Credit: ESO

<u>Goal:</u> Data-driven classification pipeline layer (galactic & extragalactic sources)

Classification infrastructure working group, led by: N. Napolitano & C. Heneka

→ *Benchmark with SDSS archival spectra:*

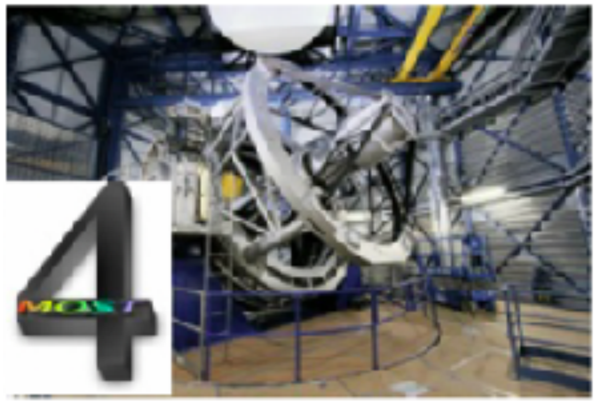
See also our tutorial!



@SDSS

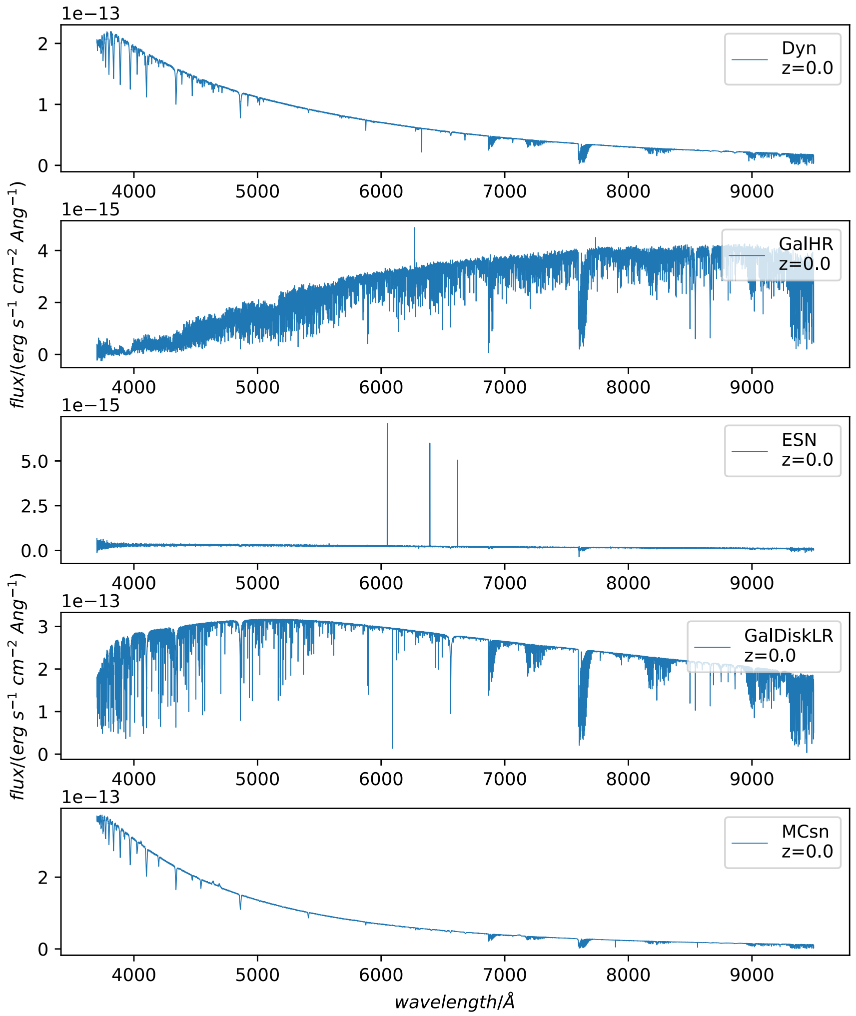
1) Classification and triggering for large astronomical surveys

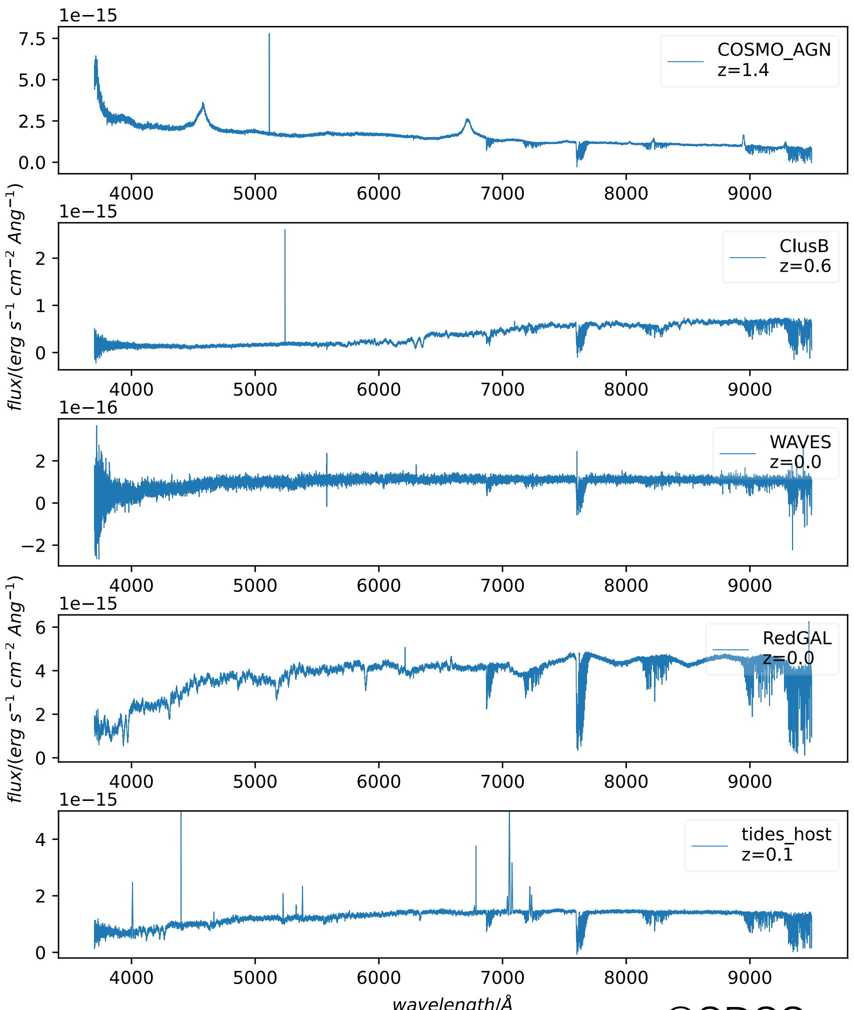
4MOST: On-the-fly classification of spectra (1D)



https://www.4most.eu Credit: ESO

- 5-y
- wid
- on
- 2.5
- HR
- 20r





@SDSS

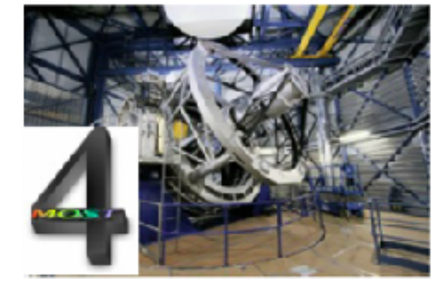
<u>Goal:</u> Data-driven classi

Classification infrastructu

Galactic vs. Extragalactic
See also our tutorial!

1) Classification and triggering for large astronomical surveys

4MOST: On-the-fly classification of spectra (1D)



https://www.4most.eu Credit: ESO

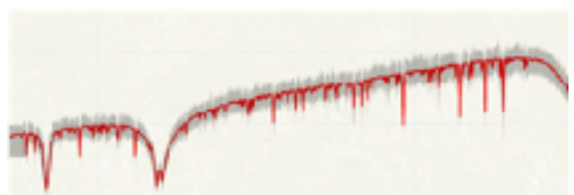
<u>Goal:</u> Data-driven classification pipeline layer (galactic & extragalactic sources)

Classification infrastructure working group, led by: N. Napolitano & C. Heneka

→ Probabilistic multi-classifier

<u>For class:</u>
Convolutional network variants

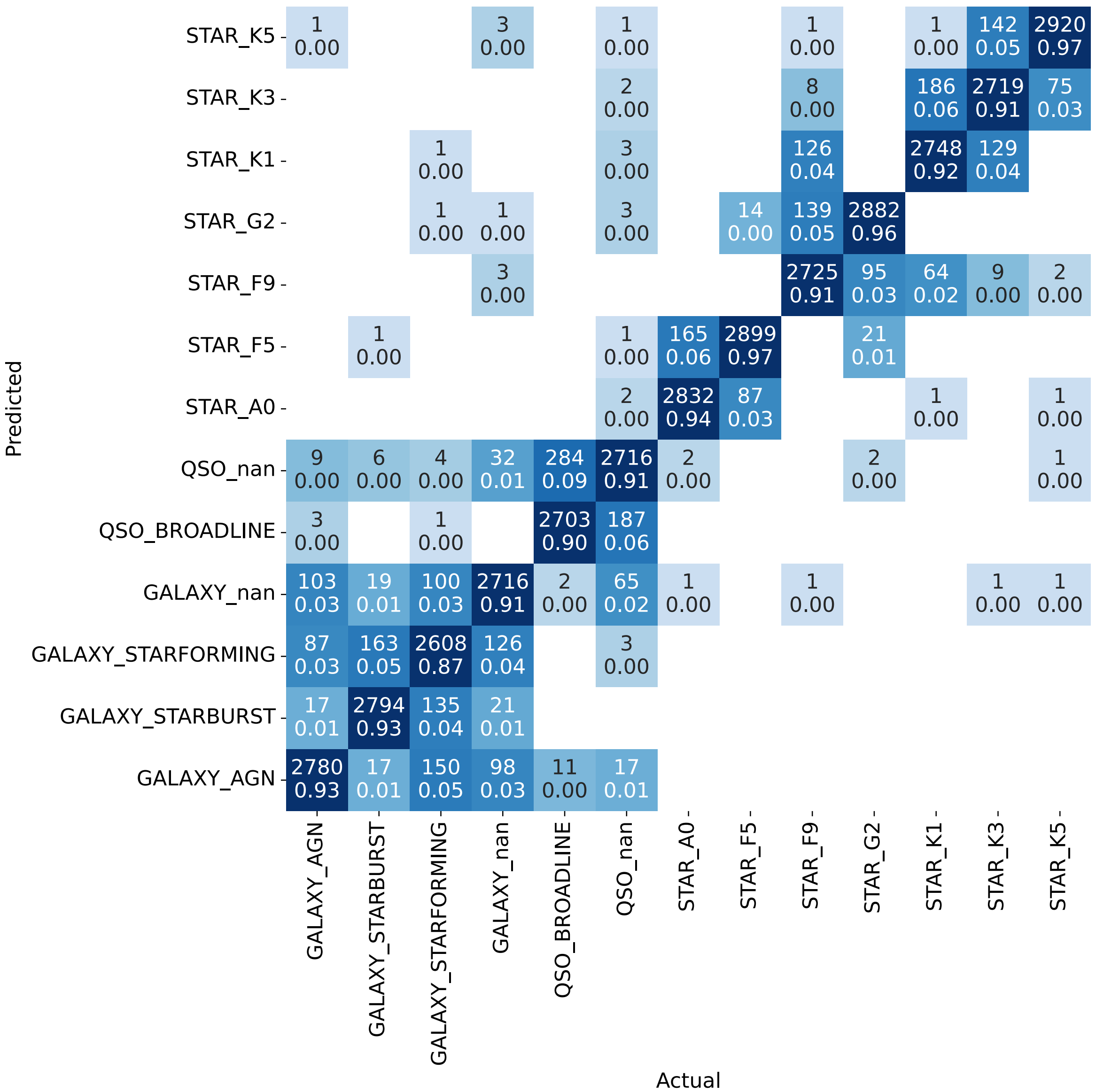
<u>For class uncertainties:</u>
Bayesian neural networks and contrastive learning



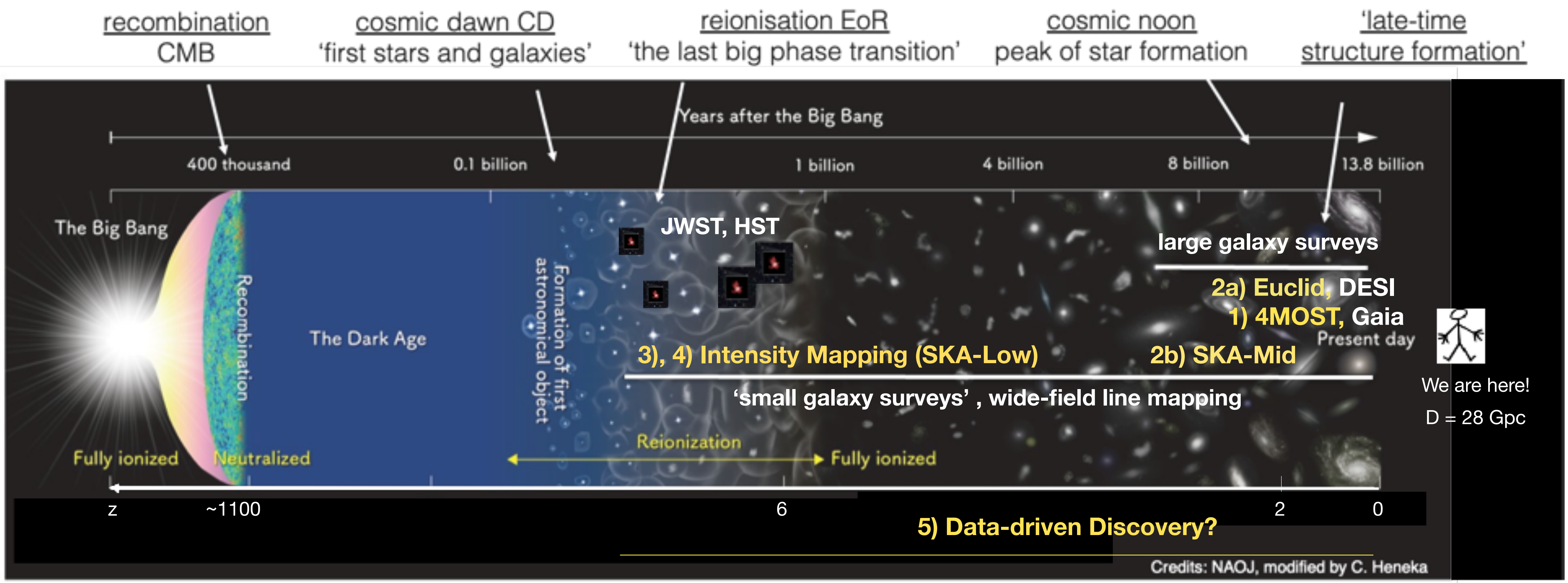
++ competitive with template fitting

Examples:
<u>Few sec:</u> Classification 40.000 spectra
<u>Few sec:</u> 7-parameter inference ~100MB cube
<u>Few sec:</u> detection, segmentation & flux measurement O(100-1000) sources

Zhong, Napolitano, Heneka+ arXiv:2311.04146

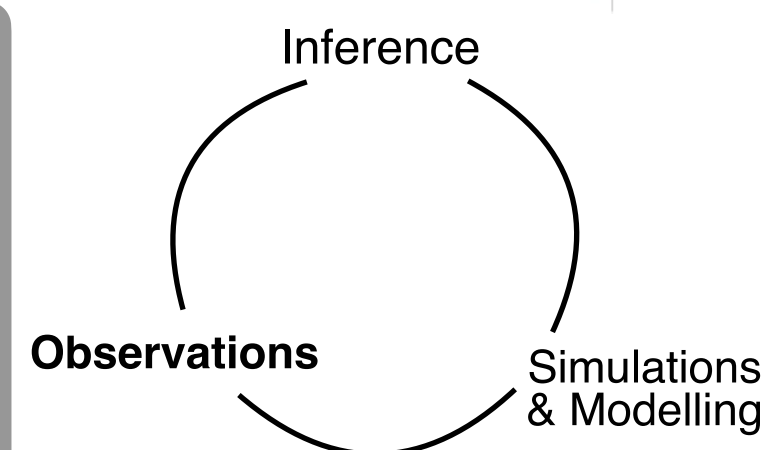


Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

1) Classification / Triggering
2) **Source detection & characterisation**
3) Simulation-based inference (SBI) in 3D
4) Generative methods
5) Data-driven Discovery



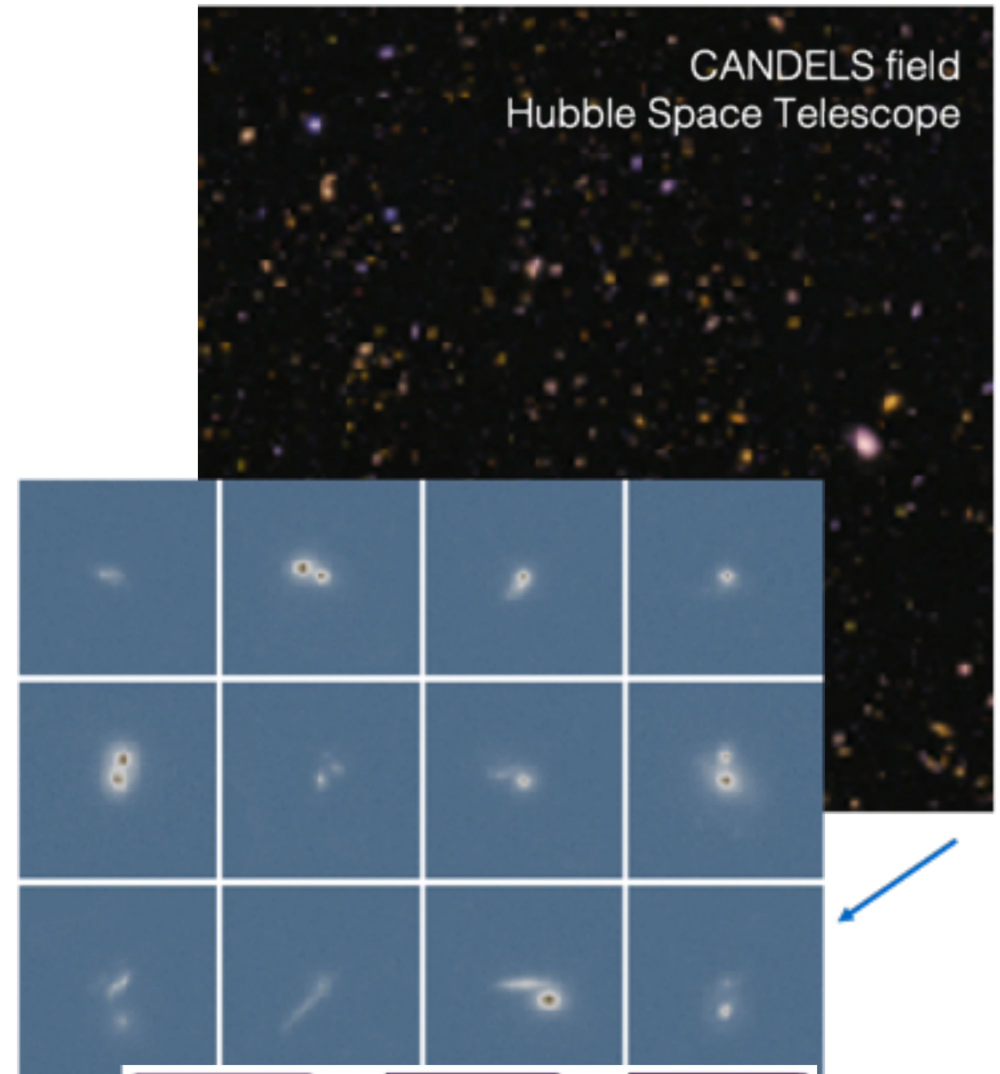
2a) The deblending problem

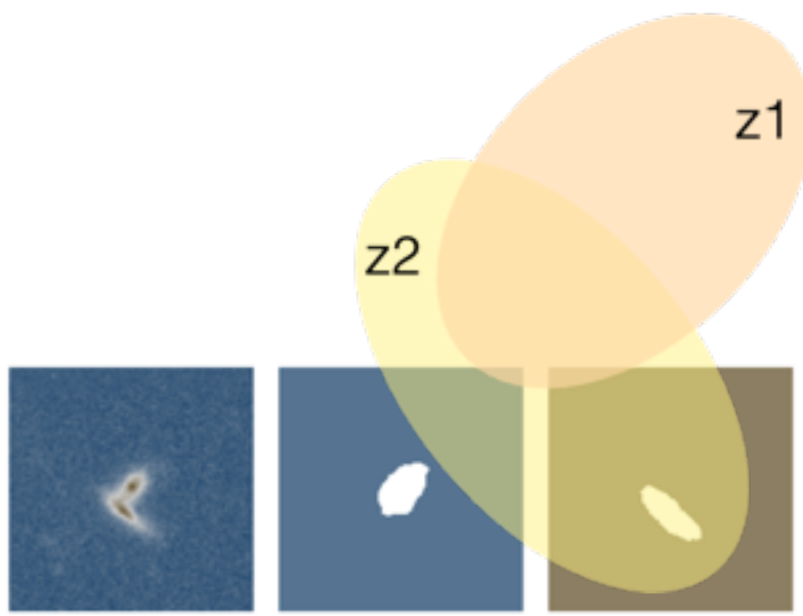
Example: Optical source detection & characterisation

Goal: ‘Good’ photometry for surveys with high blended fraction - avoid bias!

Galaxy morphology

Challenge: Galaxies are ‘transparent’





Lily Hu+ 2017

Similar challenge:
Overlapping chromosomes

coindeblend

Boucaud, Huertas-Company, Heneka+ 20, arXiv:1905.01324

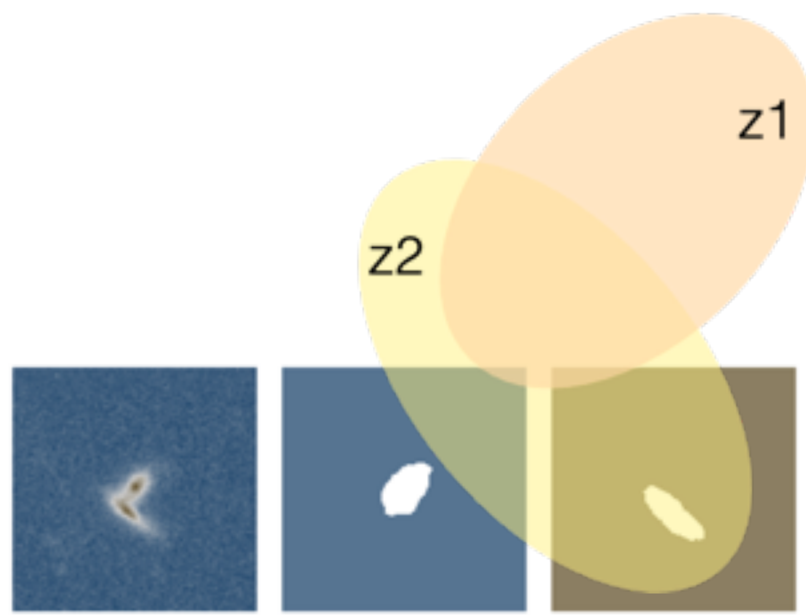
2a) The deblending problem

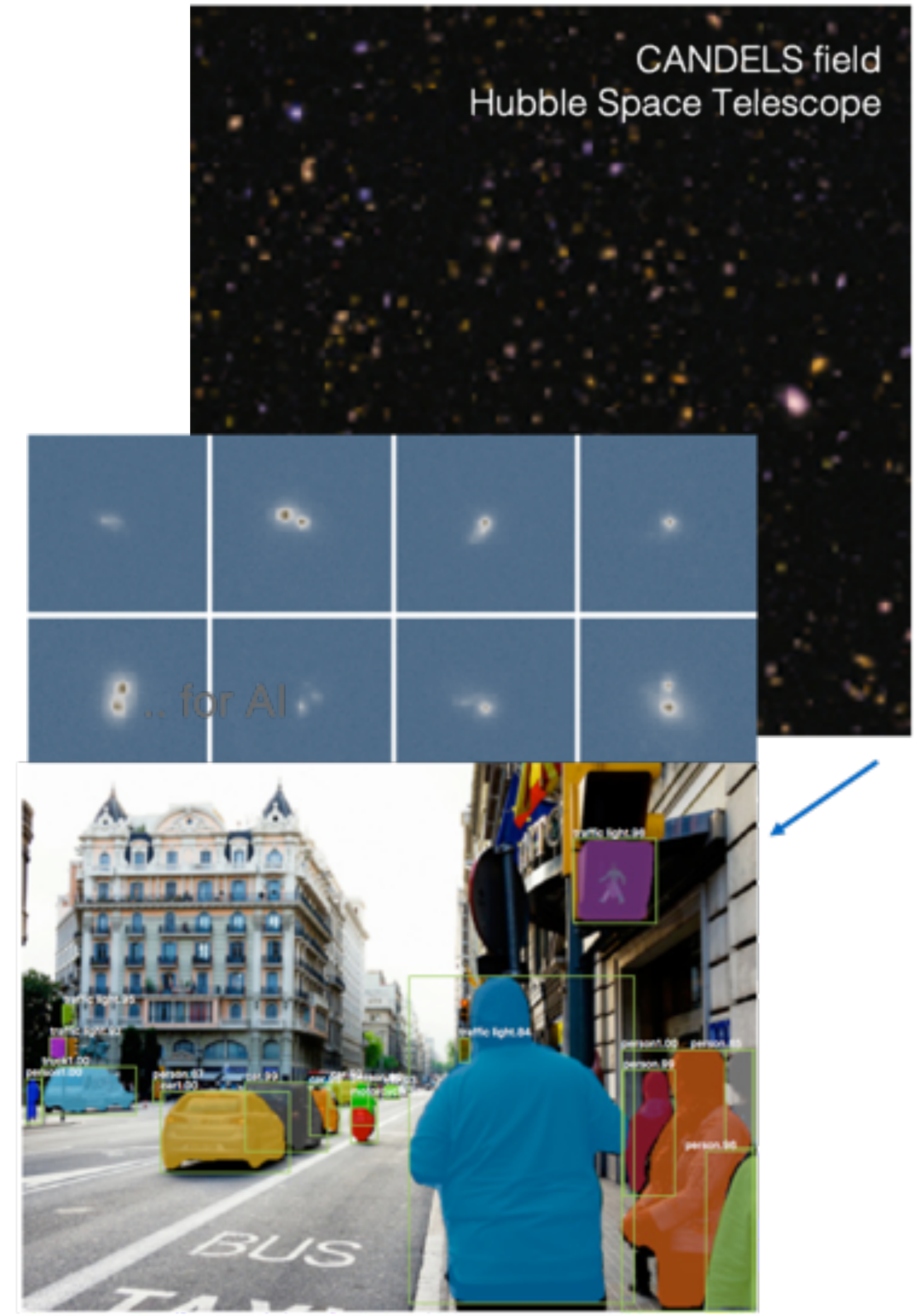
Example: Optical source detection & characterisation

Goal: 'Good' photometry for surveys with high blended fraction
- avoid bias!

Galaxy morphology

Challenge: Galaxies are 'transparent'





https://medium.com/@umerfarooq_26378/from-r-cnn-to-mask-r-cnn-d6367b196cfd

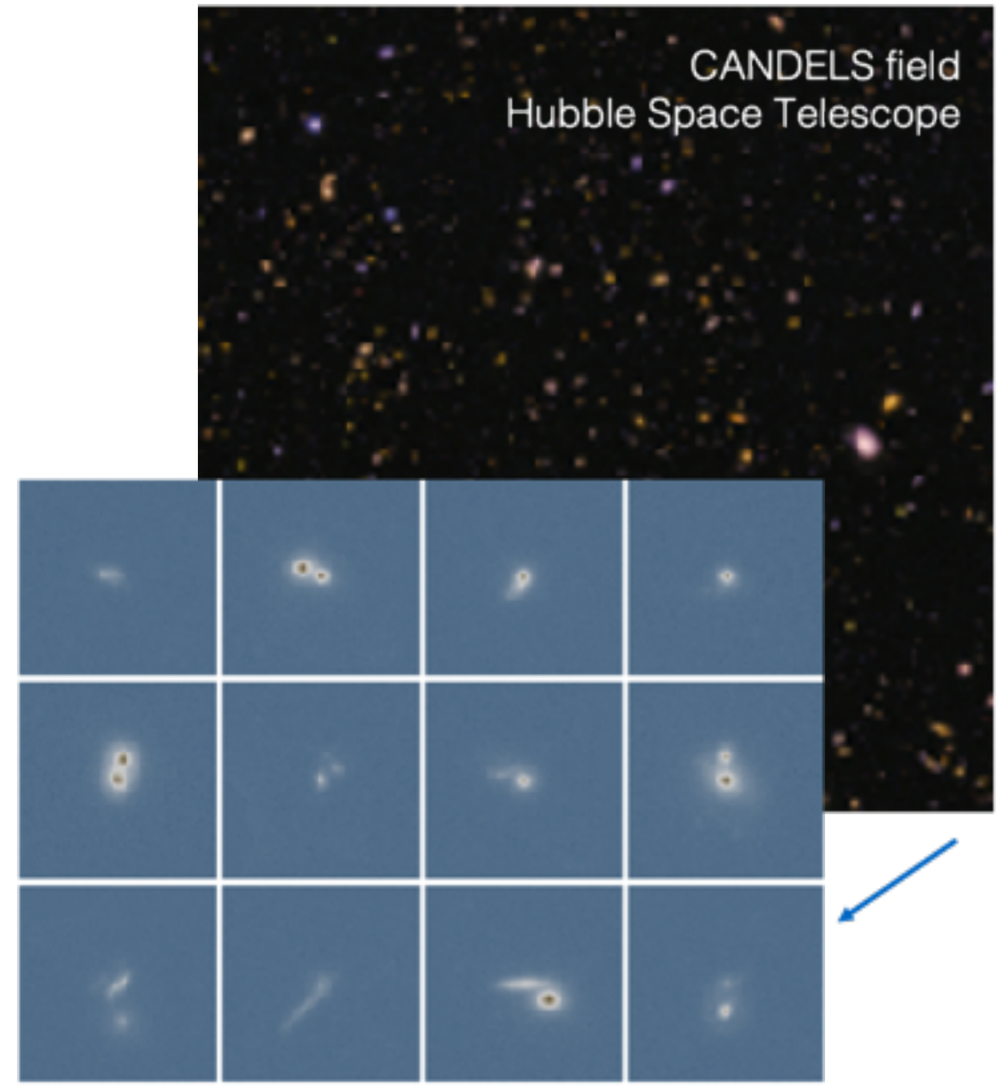
Another challenge: Object detection

coindeblend

Boucaud, Huertas-Company, Heneka+ 20, arXiv:1905.01324

2a) The deblending problem

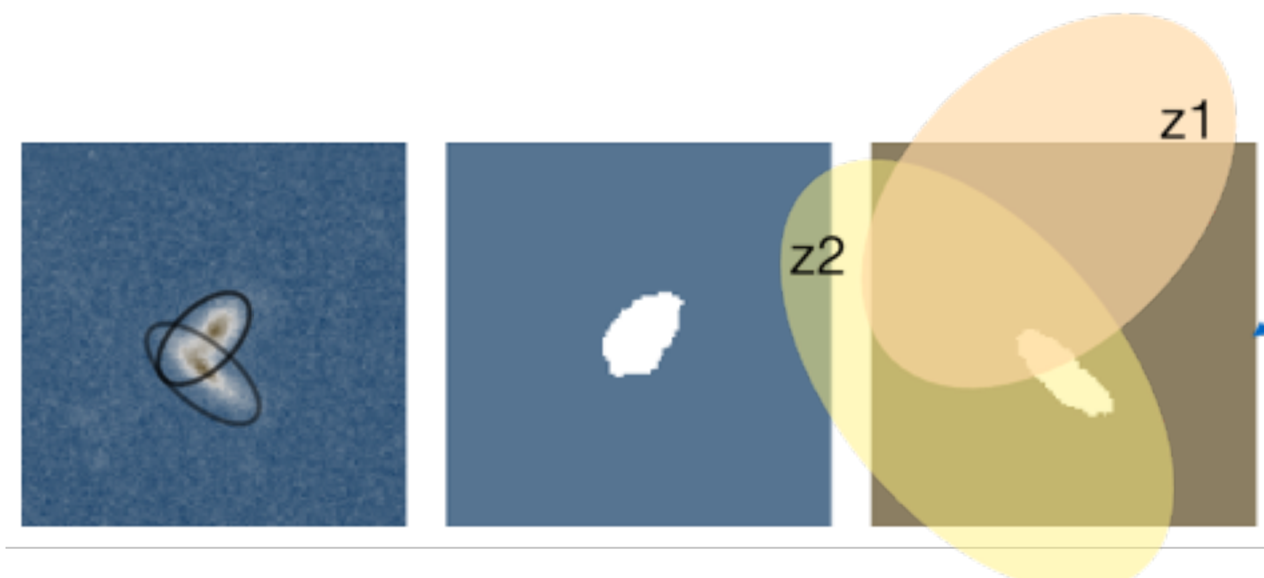
Example: Optical source detection & characterisation



Goal: 'Good' photometry for surveys with high blended fraction
- avoid bias!

Galaxy morphology

Challenge: Galaxies are 'transparent'



<u>'Classic'</u>:
Fit ellipse(s)
and profile(s)

e.g. Einasto ('65):

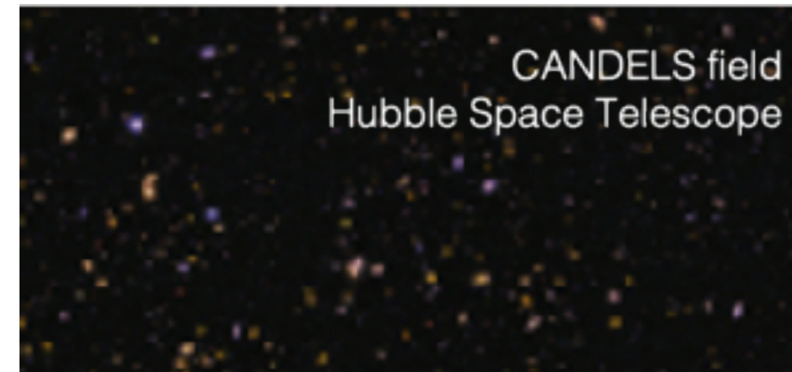
$$\frac{d\log(\rho)}{d\log(r)} = -2\left(\frac{r}{r_s}\right)^{\alpha}$$

coindeblend

Boucaud, Huertas-Company, Heneka+ 20,
arXiv:1905.01324

2a) The deblending problem

Example: Optical source detection & characterisation



Goal: ‘Good’ photometry for surveys with high blended fraction
- avoid bias!

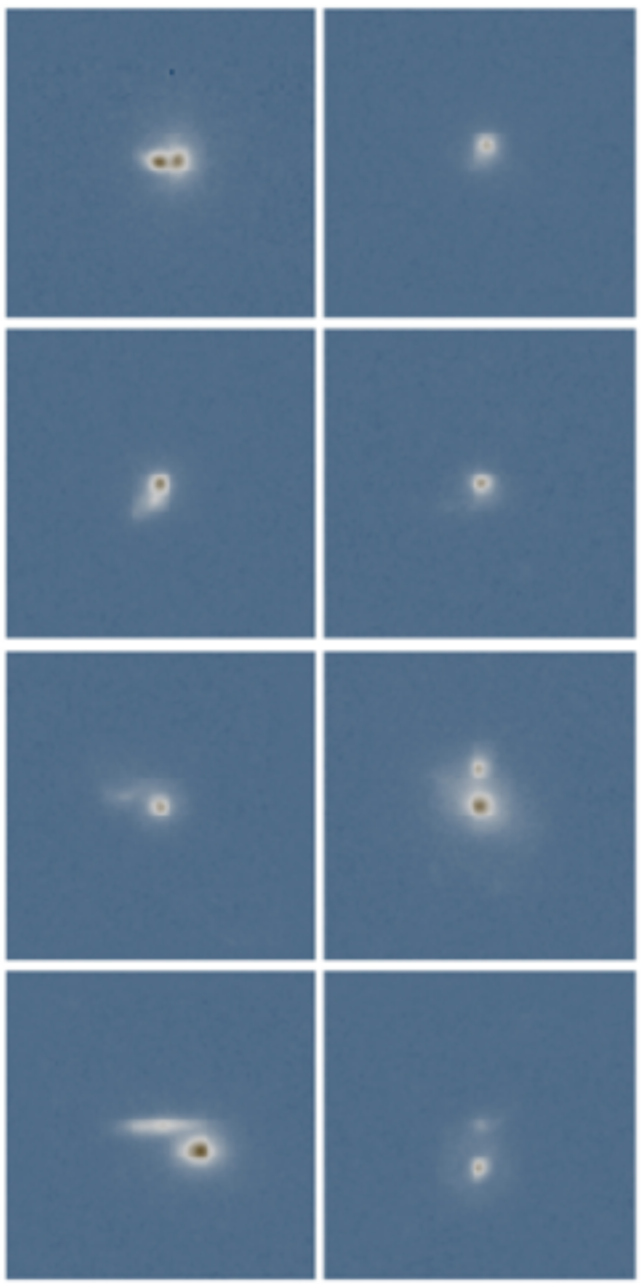
Galaxy morphology

Credit: Euclid, ESA

2a) The deblending problem

Example: Optical source detection & characterisation

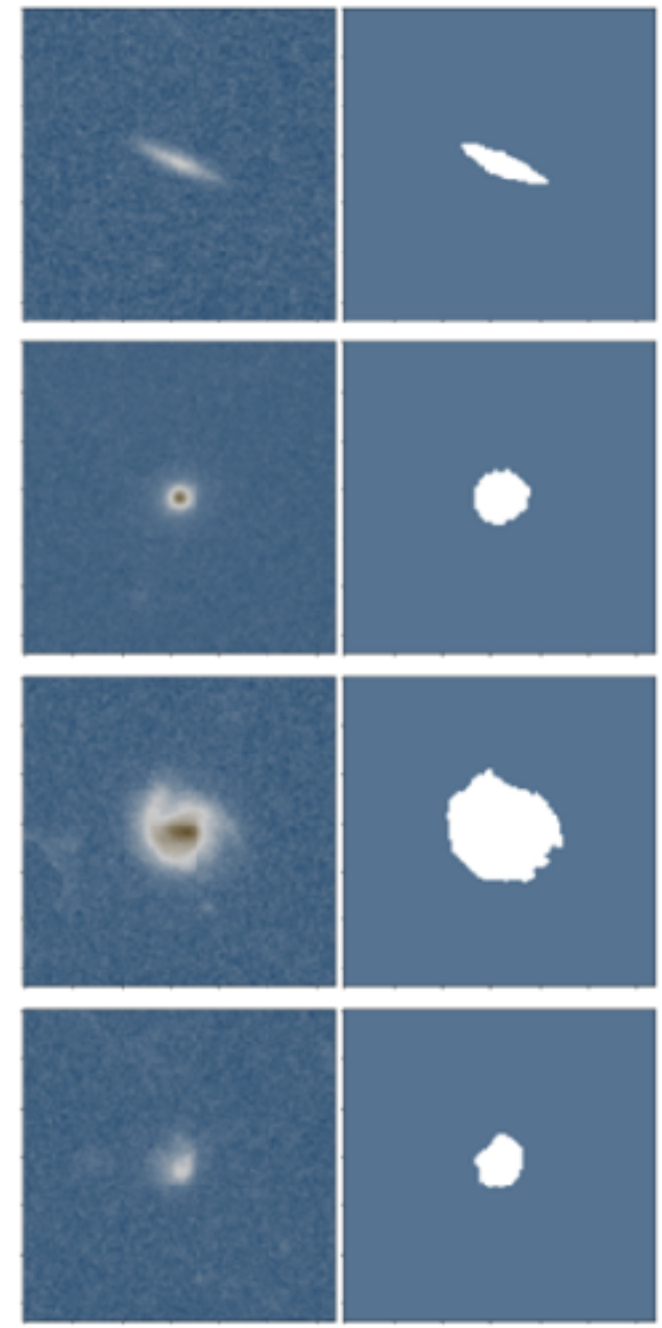
CANDELS field
Hubble Space Telescope



Get precise Photometry

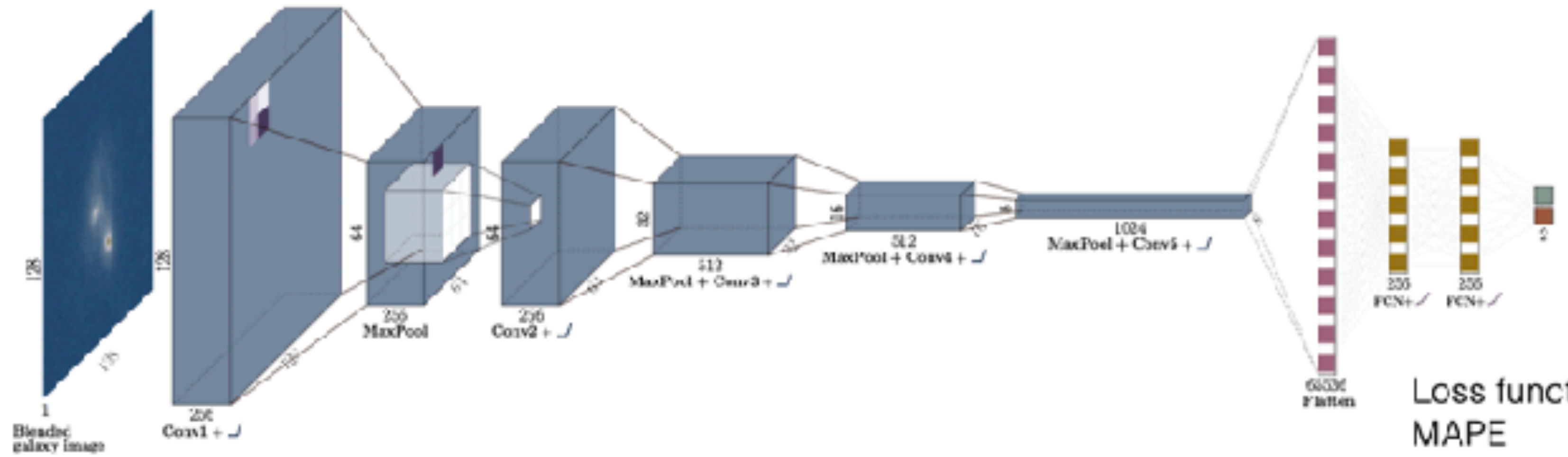
Derive masks

..and do so bias-free



2a) The deblending problem

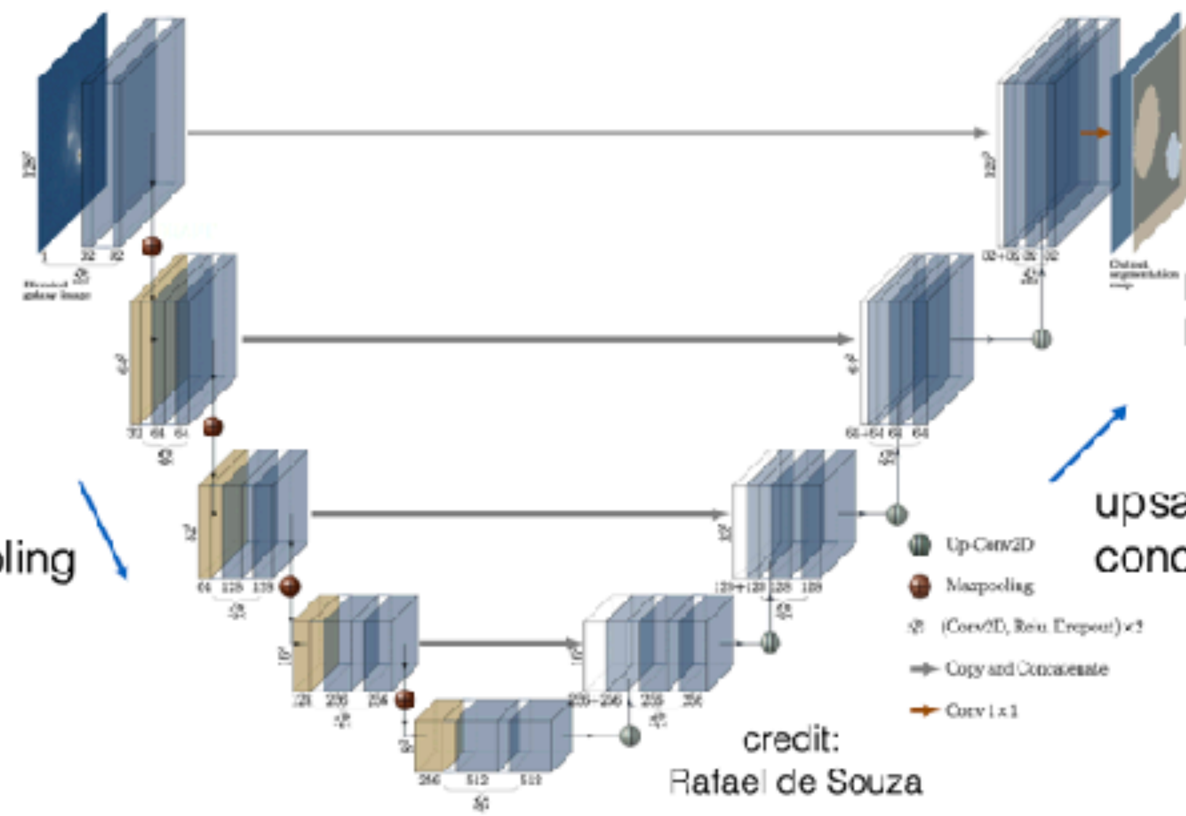
1) *blend2flux*
a CNN for photometry



Loss function:
MAPE

2) *blend2mask2flux*
photometry + masks

downsampling



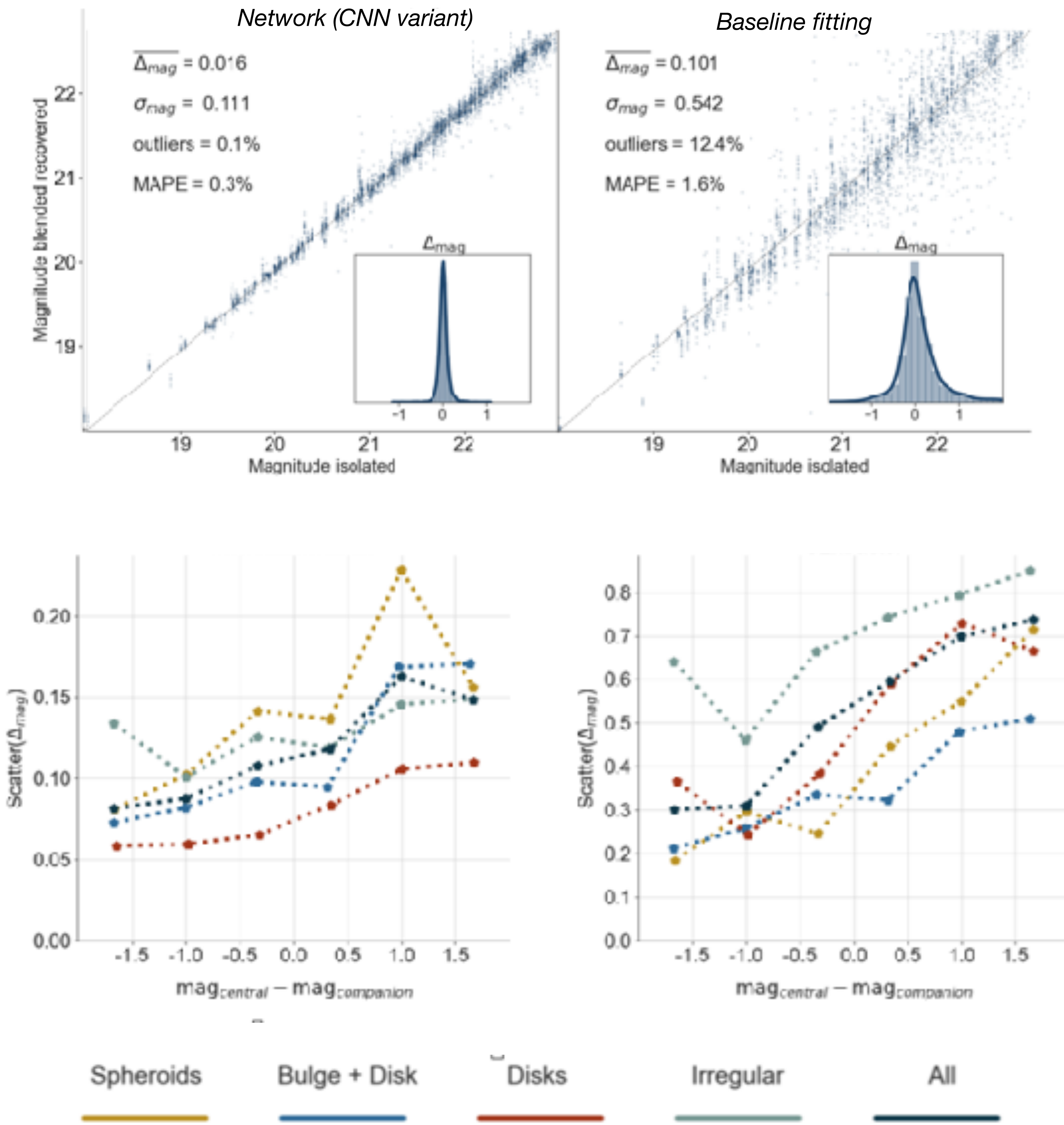
+ a CNN type *blend2flux*
(but 3-channel input)

Loss function:
Binary cross-entropy

upsampling
concatenation

credit:
Rafael de Souza

2a) Optical source detection and characterisation



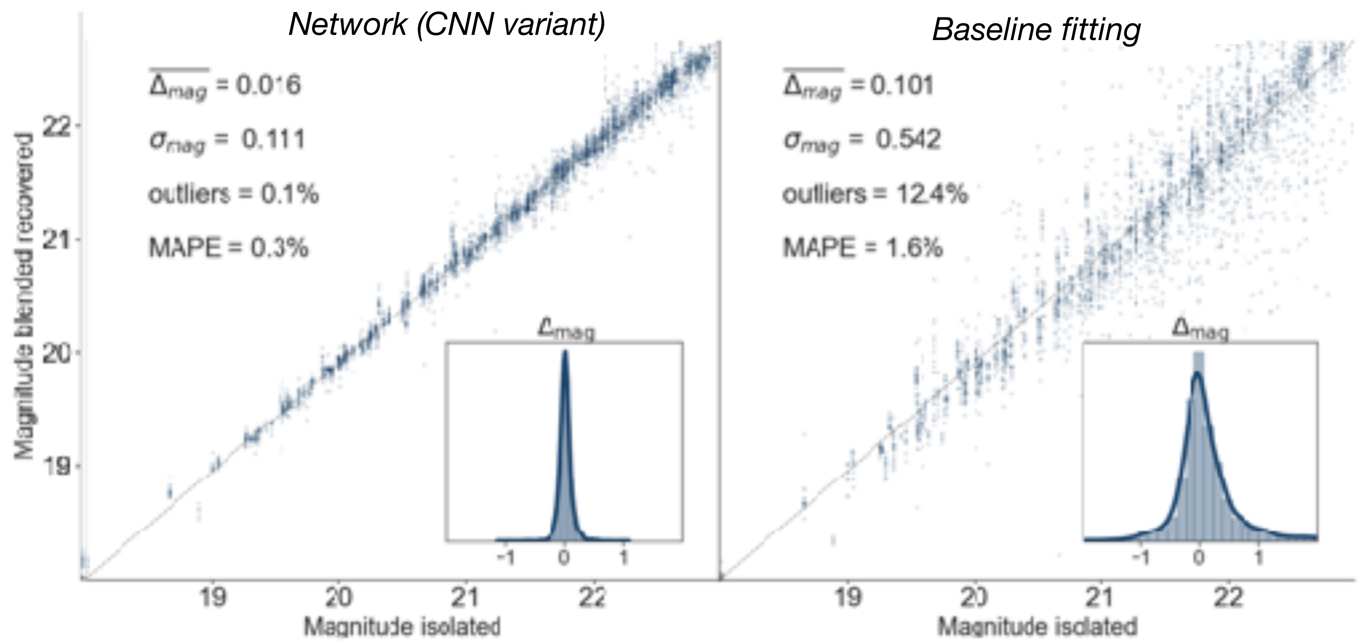
Get precise Photometry

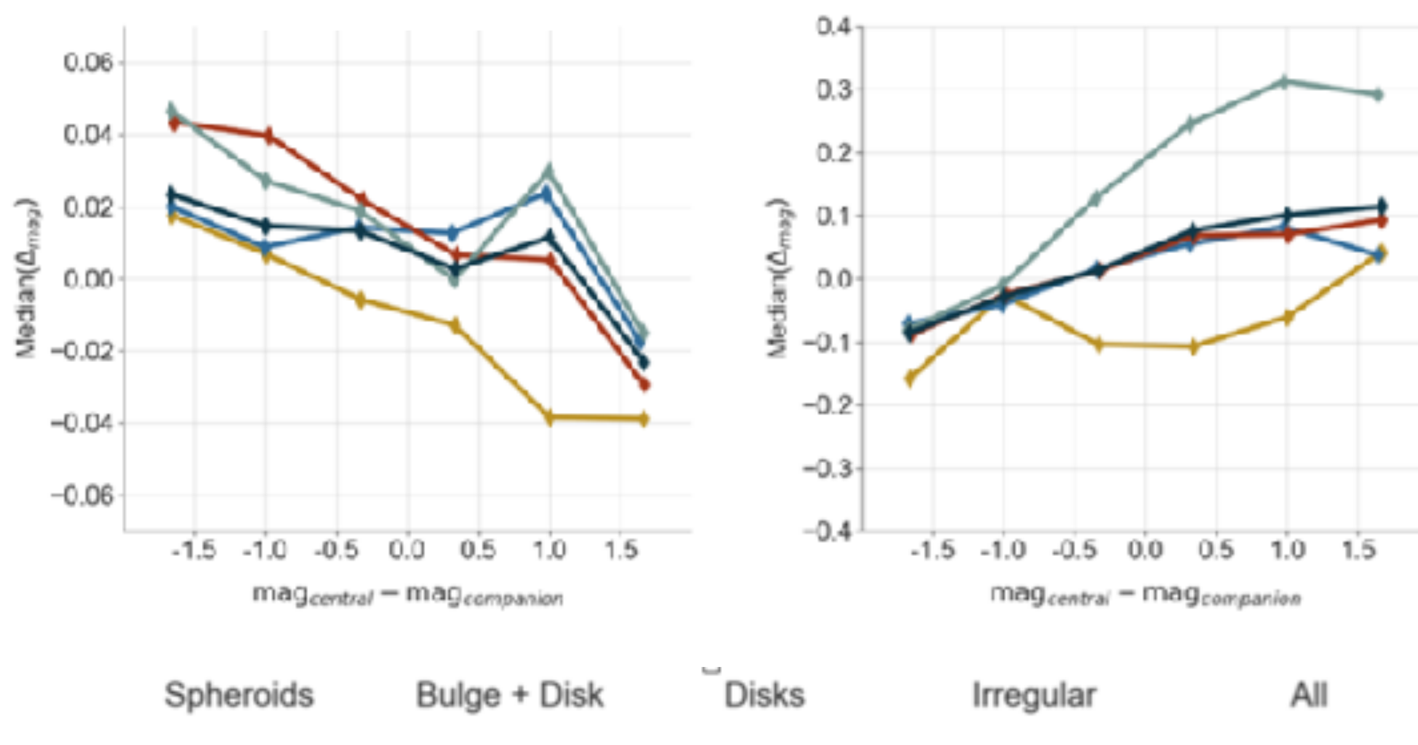
For irregular shapes

coindeblend

Boucaud, Huertas-Company, Heneka+ 20

2a) Optical source detection and characterisation





Get precise Photometry ✔

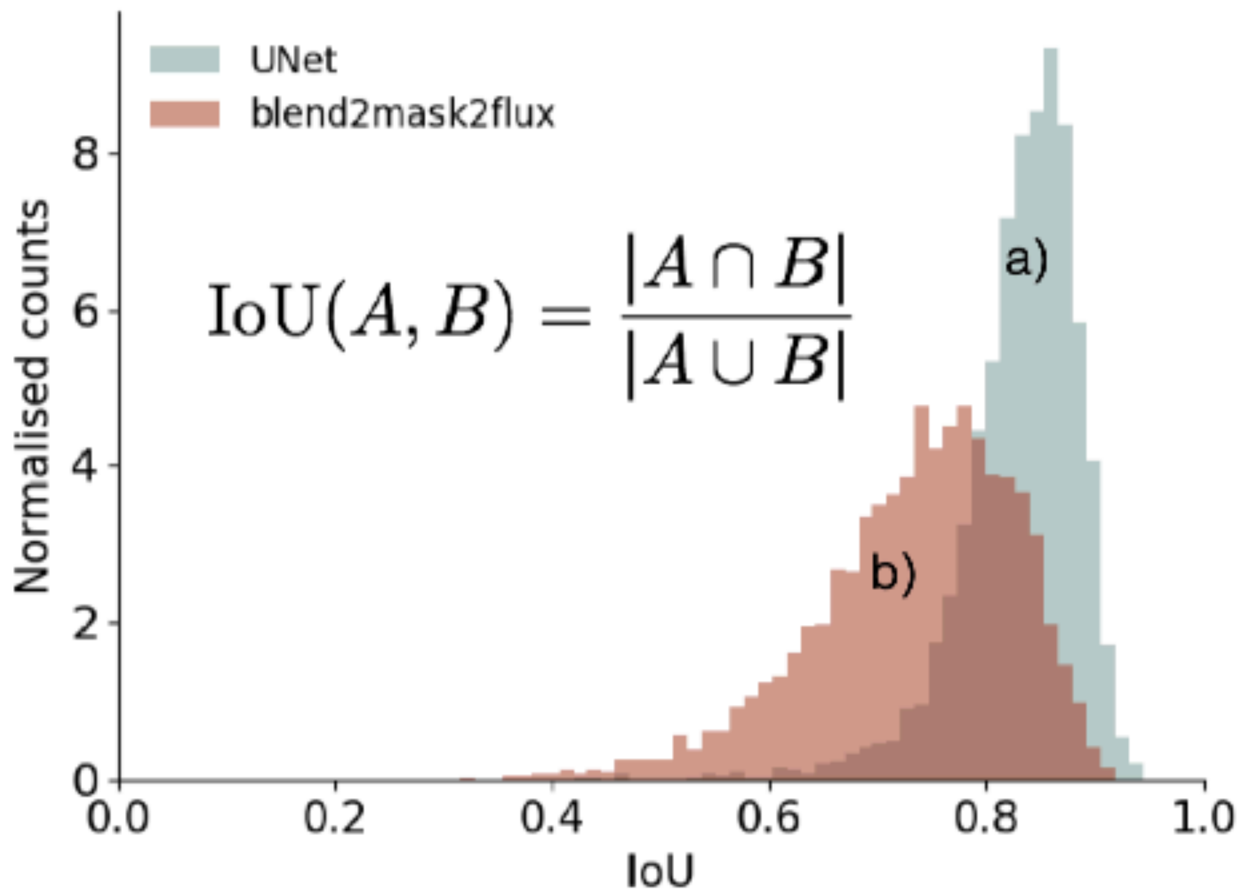
..and do so bias-free ✔

coindeblend

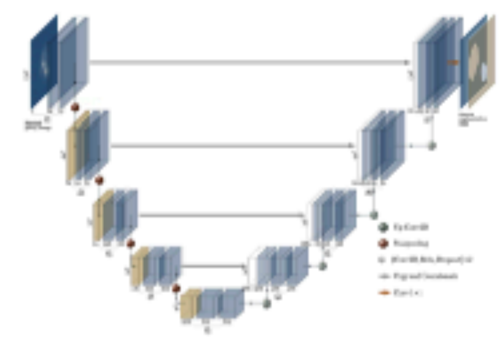
Boucaud, Huertas-Company, Heneka+ 20

2a) Optical source detection and characterisation

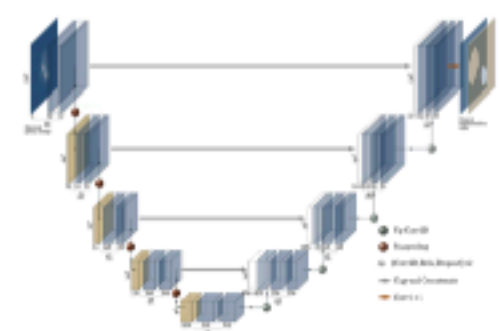
Optimal design is goal-dependent



a) *U-net*
masks, no photometry



b) *blend2mask2flux*
photometry + masks



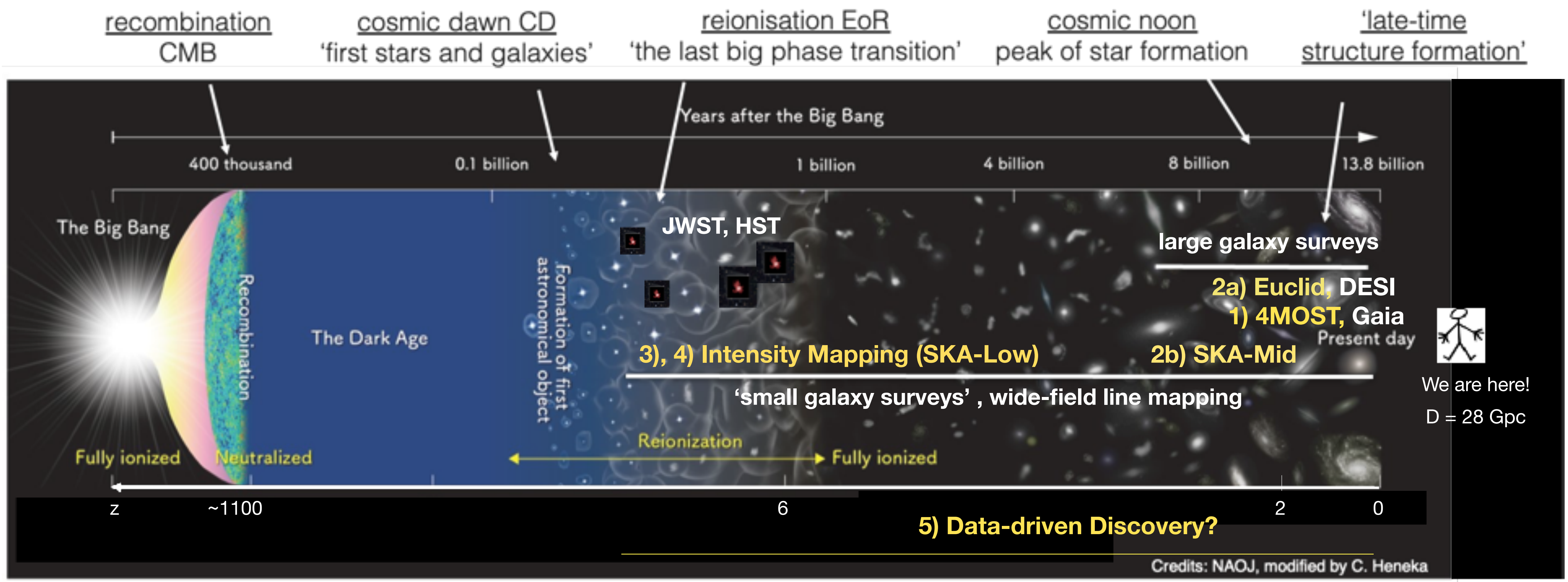
+ a CNN type *blend2flux*

ground truth | *U-Net* | *blend2mask2flux*

→ Dispersion broadens when optimised for photometry

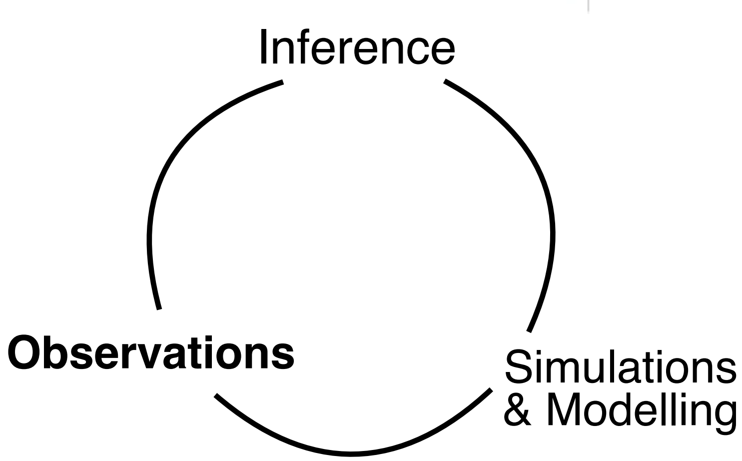
→ Tailor to your research question

Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

1) Classification / Triggering
2) **Source detection & characterisation**
3) Simulation-based inference (SBI) in 3D
4) Generative methods
5) Data-driven Discovery



The Square Kilometre Array (SKA) in one slide

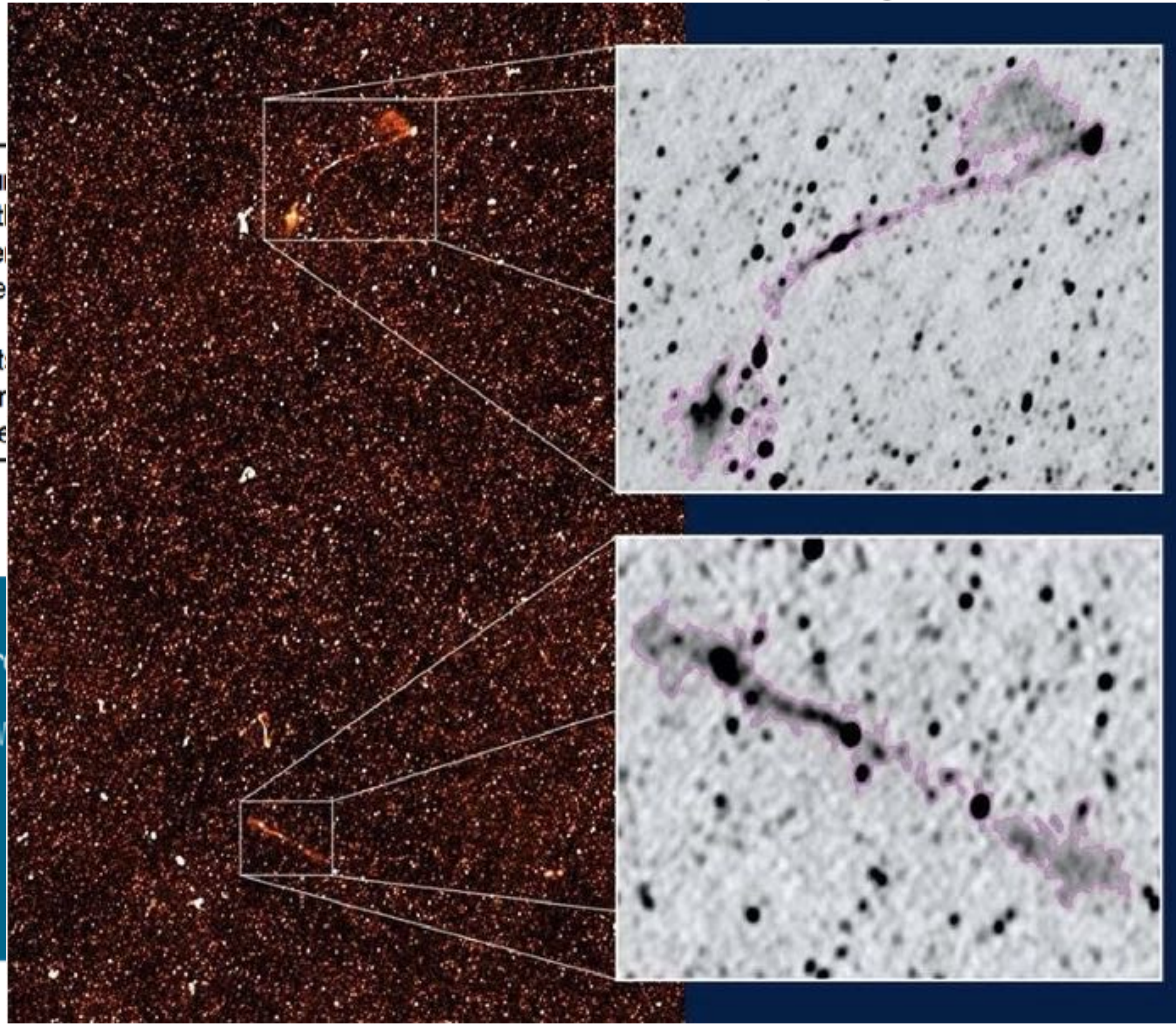
SKA in numbers
- Currently 16 member countries, >100 member organisations
- Routine science observations are expected to start in the late 2020s
- Consists of thousands of dishes and up to 1 million antennas, >1km^2 collecting area
- Expected data rate in full operation: 1 TB/s

SKA-LOW SKA-MID Credits: SKAO

The Square Kilometre Array (SKA) in one slide

https://doi.org/10.1093/mnras/staa3837

SKA in nu
- Current
  membe
- Routine
  to start
- Consist
  1 million
- Expecte



2b) Radio source detection and characterisation

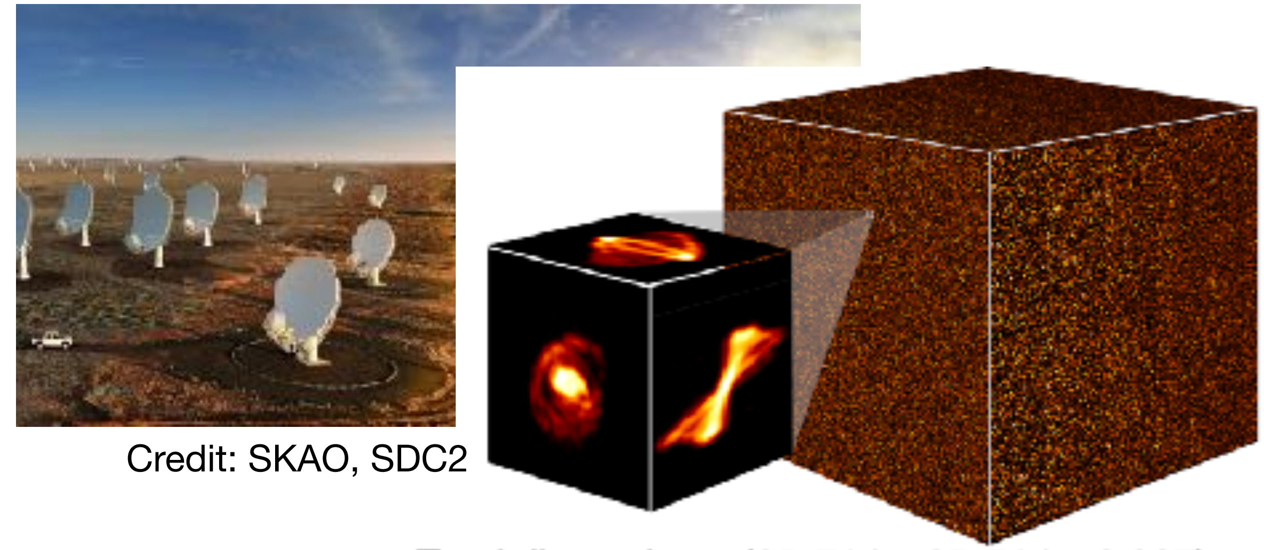
Example: Source detection in 3D tomographic data

Source finding

Location in RA, Dec,
central frequency (Hz)

Characterisation

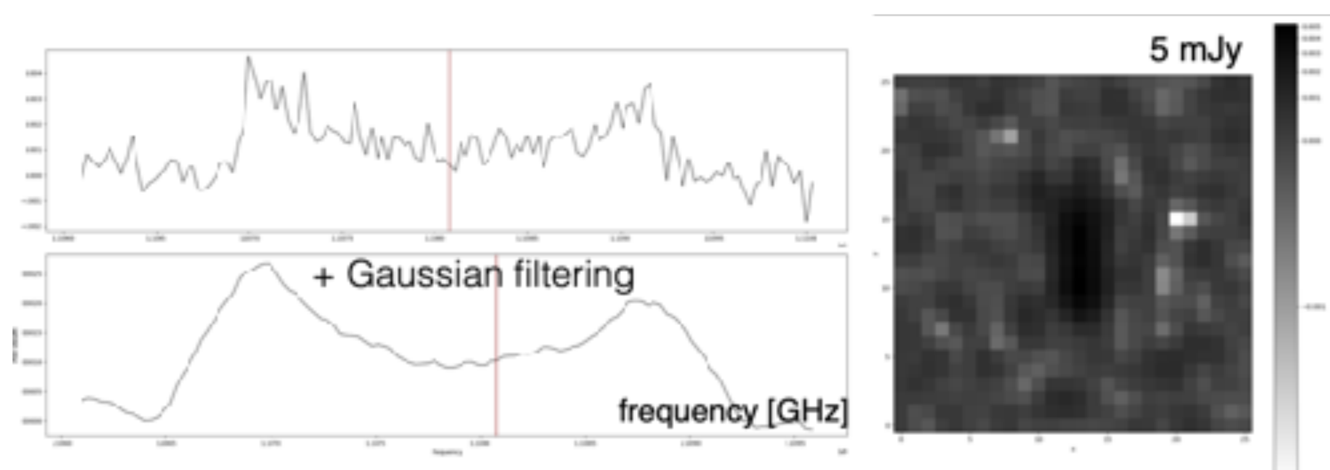
- Integrated line flux (Jy Hz)
- Line width (km/s)
- HI major axis diameter (arcsec)
- Position angle (degrees)
- Inclination angle (degrees)



Credit: SKAO, SDC2

Total dimensions: (25,714 x 25,714 x 6,667) vox

The brightest HI source



The challenging HI sources:

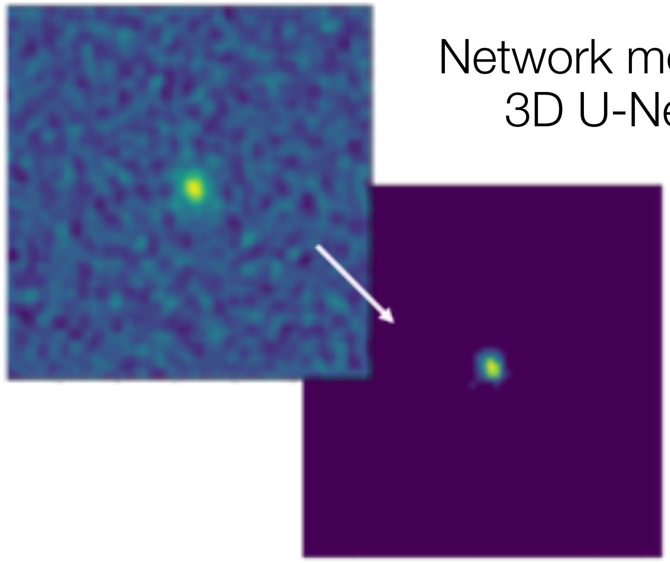
- low S/N
- small spatial size
- systematics

Hartley+ 23 (incl. Heneka), arXiv:2303.07943
Heneka 23, arXiv:2311.17553

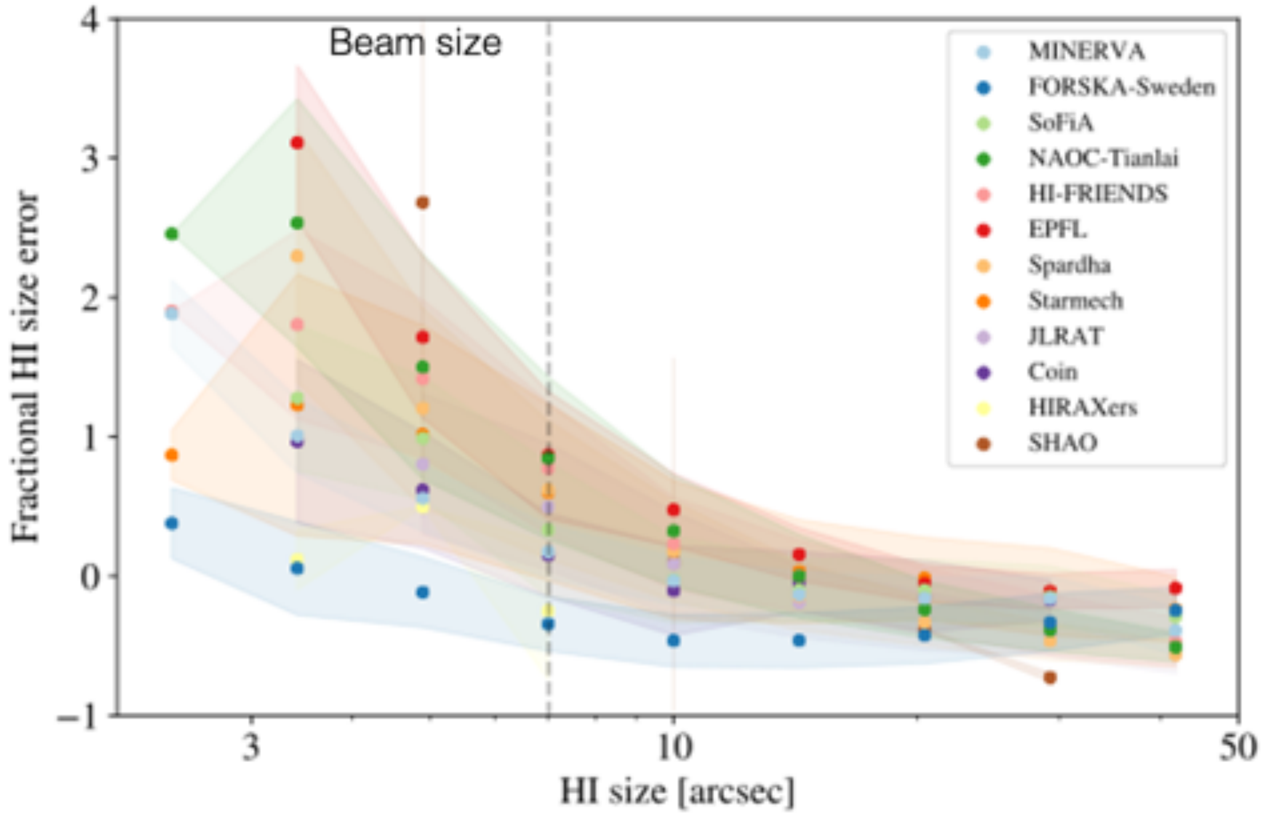
2b) Radio source detection and characterisation

Example: Source detection in 3D tomographic data

- 3D better than stitching of 2D + 1D
- High-fidelity 3D reconstructions
- **Good prior for characterisation tasks via nets:**



Network model:
3D U-Net



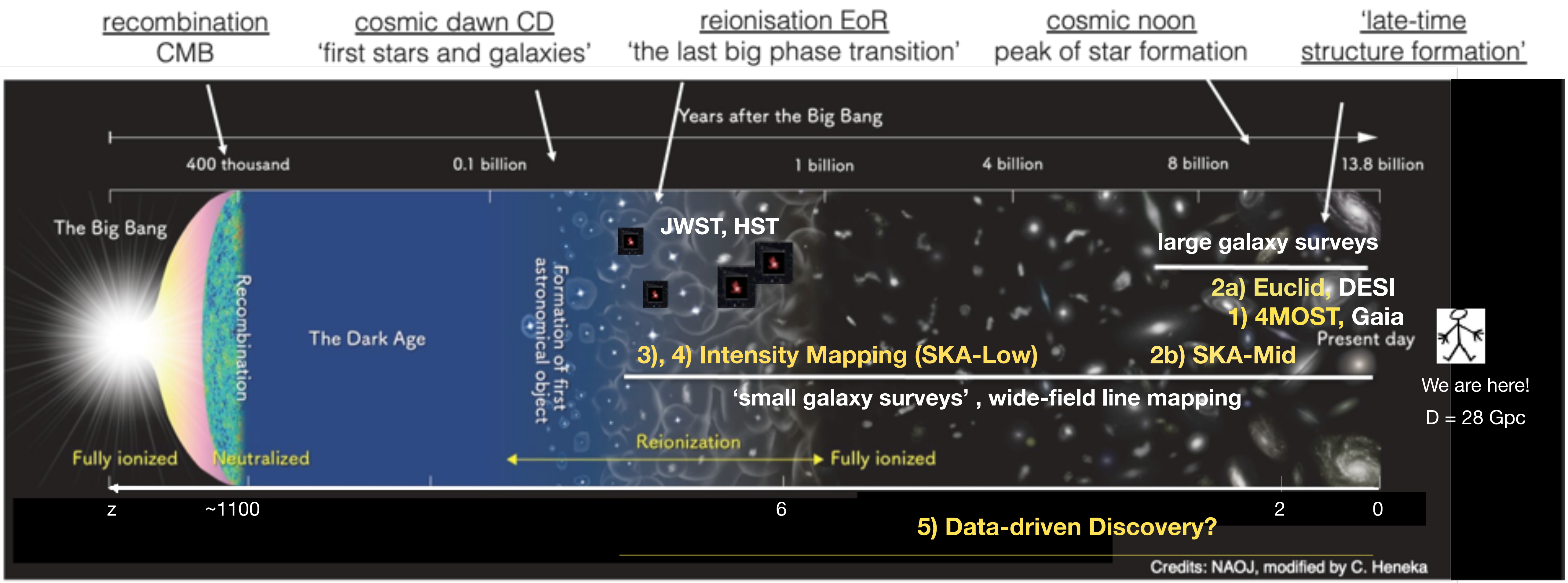
Recovery across wide range in HI flux and size

Pushing to low S/N recovery came at a cost (FPs)

Uncertainty?

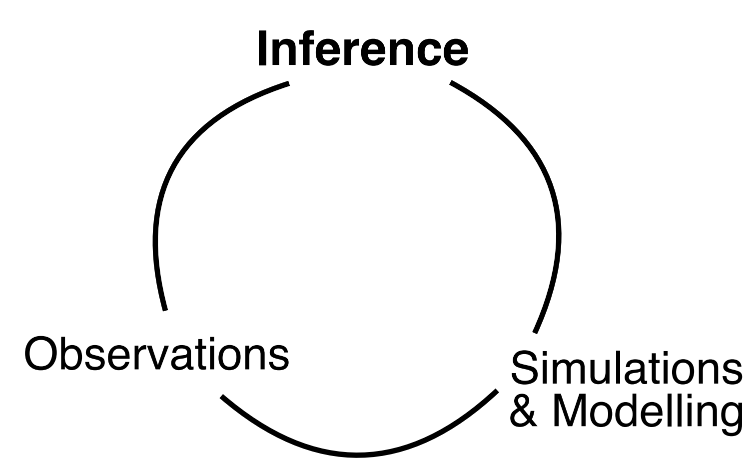
Hartley+ 23 (incl. Heneka), arXiv:2303.07943
Heneka 23, arXiv:2311.17553

Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

1) Classification / Triggering
2) Source detection & characterisation
3) **Simulation-based inference (SBI) in 3D**
4) Generative methods
5) Data-driven Discovery



3) Simulation-based inference (SBI) for intensity mapping (3D)

21cm signal

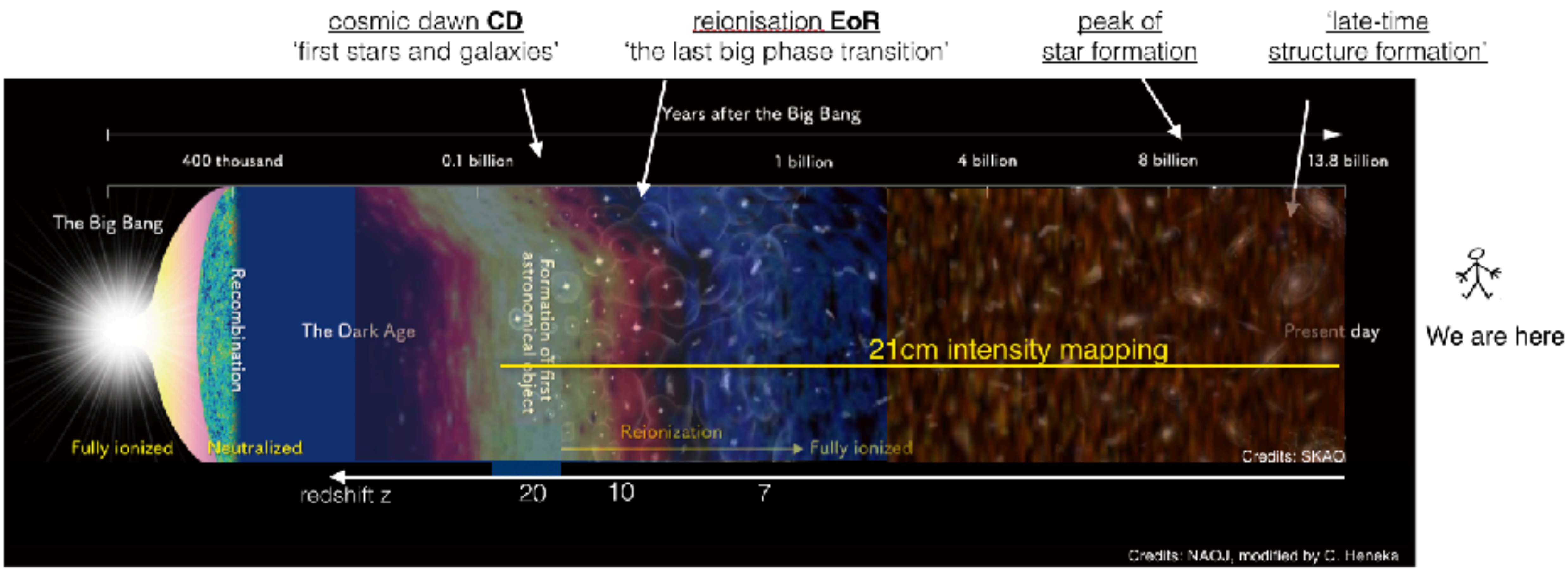
a tracer of neutral hydrogen:

$$\delta T_{\rm b}(\nu) = \frac{T_{\rm S} - T_\gamma}{1+z}\left(1 - e^{-\tau_{\nu_0}}\right)$$

$$\propto x_{\rm HI}\,(1+\delta_{\rm nl})\left(\frac{H}{{\rm d}v_{\rm r}/{\rm d}r + H}\right)$$

Why care?

Tomography of >80% of the Universe



3) Simulation-based inference (SBI) for intensity mapping (3D)

21cm signal

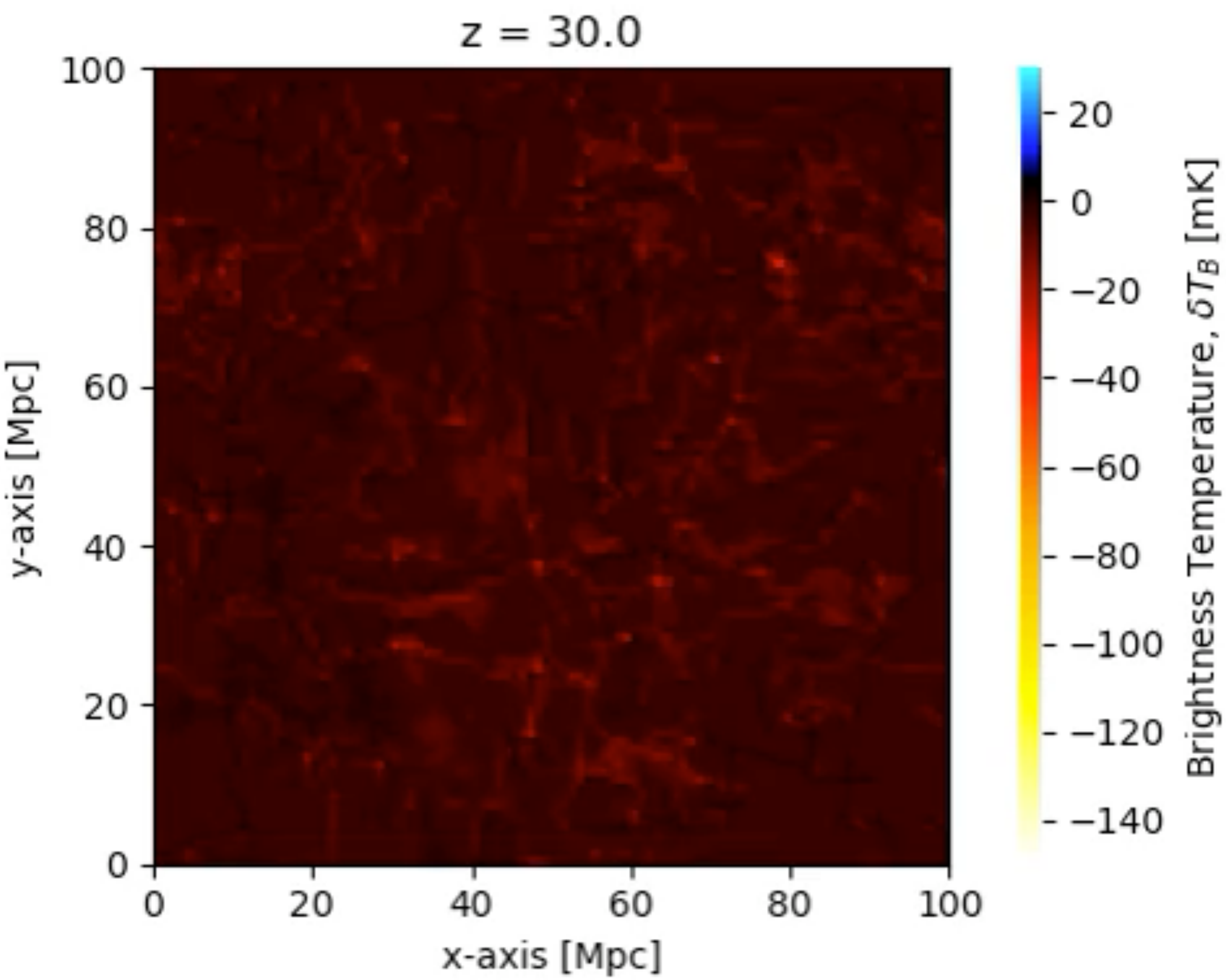
a tracer of neutral hydrogen:

Why care?

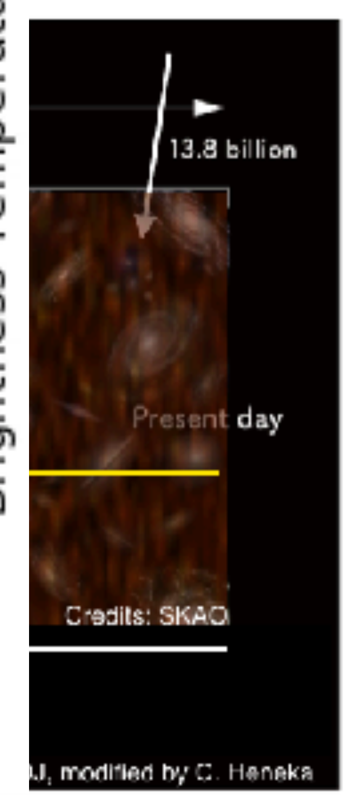
Tomography of >80% of the Universe

$\delta T_b(\nu) =$

$\propto$



'late-time structure formation'



We are here

3) Simulation-based inference (SBI) for intensity mapping (3D)

Expected data SKA rate:
TB/s, few EB/day
Archive: ~700 PB/yr

Why care?

Tomography of >80% of the Universe

Square Kilometre Array - true ‘Big Data’

non-linear, non-Gaussian signal

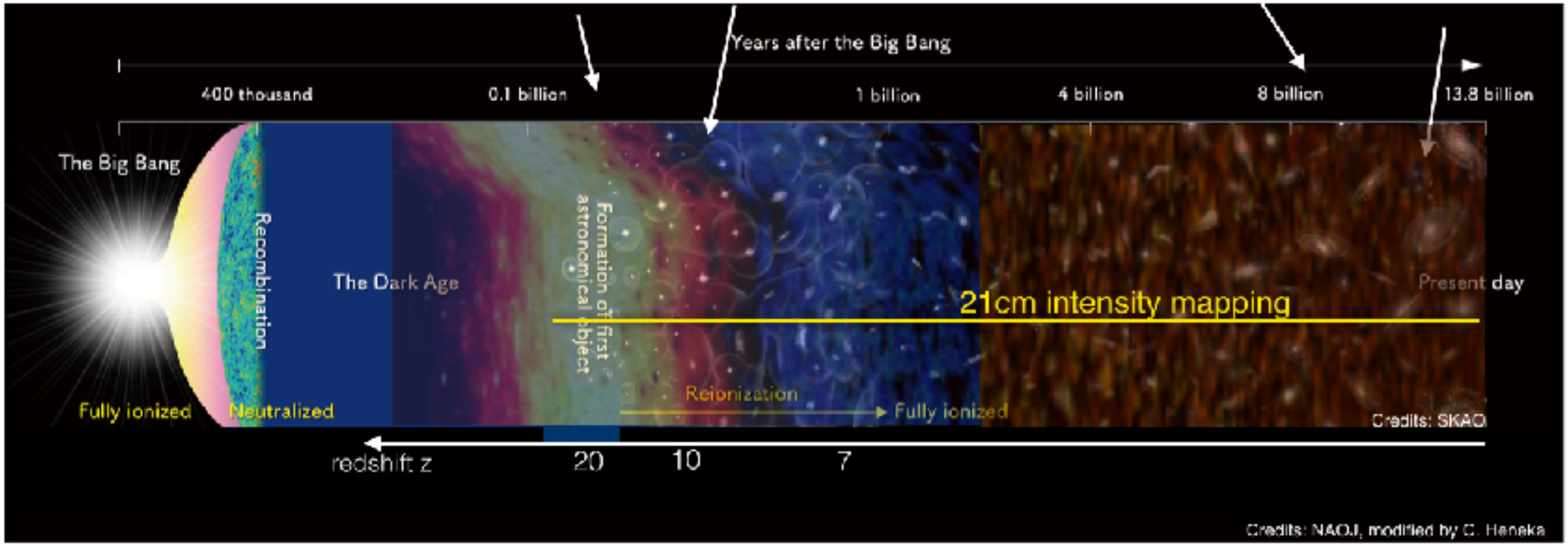
→ Move to full likelihood inference with networks

cosmic dawn **CD**
'first stars and galaxies'

reionisation **EoR**
'the last big phase transition'

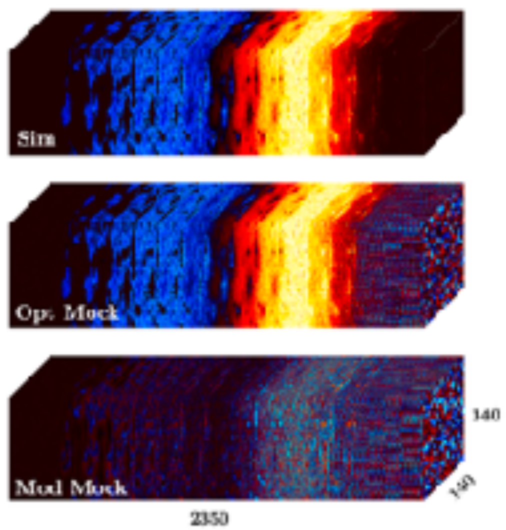
peak of
star formation

'late-time
structure formation'

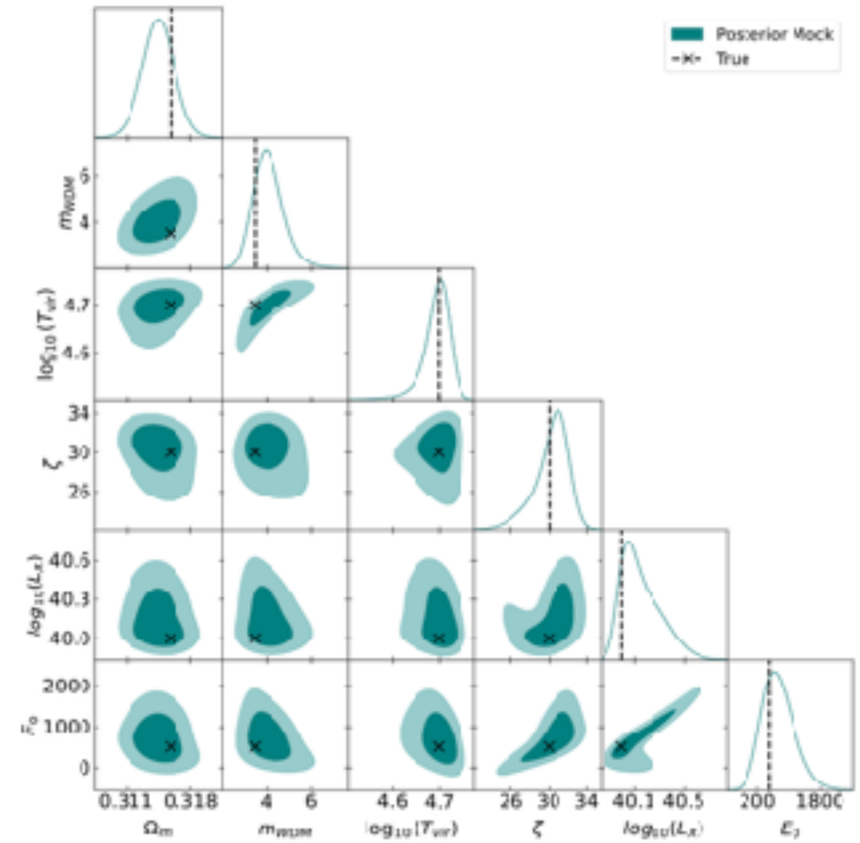


We are here

3) Simulation-based inference (SBI) for intensity mapping (3D)



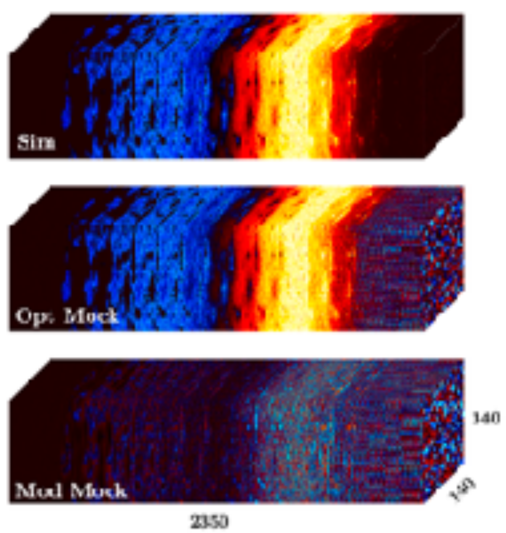
?



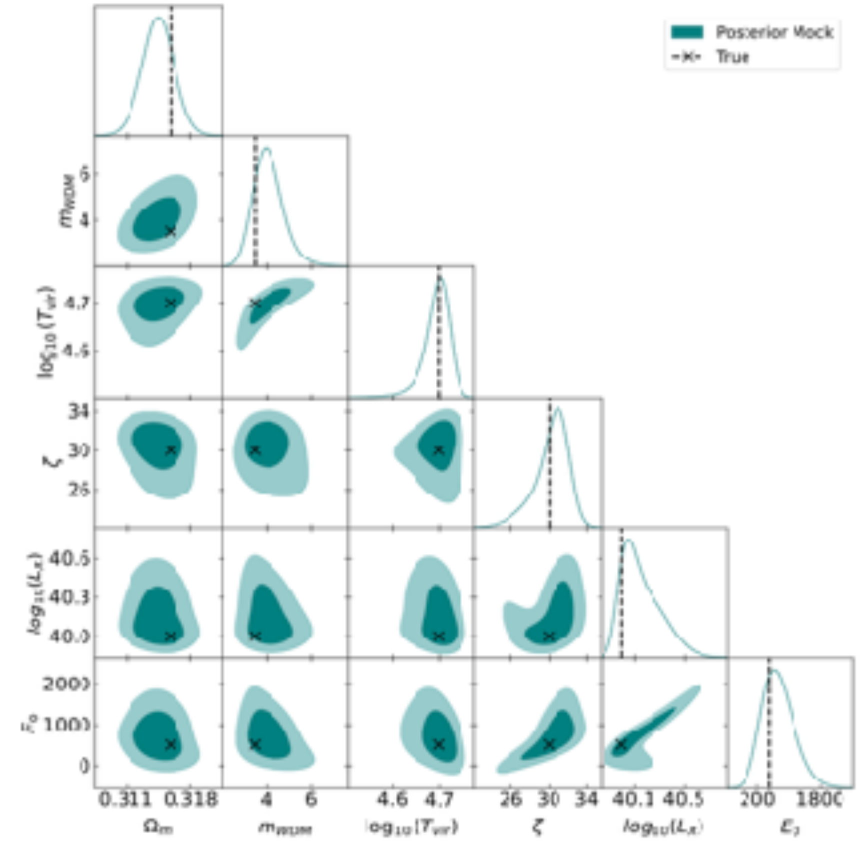
Neutsch, Heneka, Brüggen (2022), arXiv:2201.07587

Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for intensity mapping (3D)



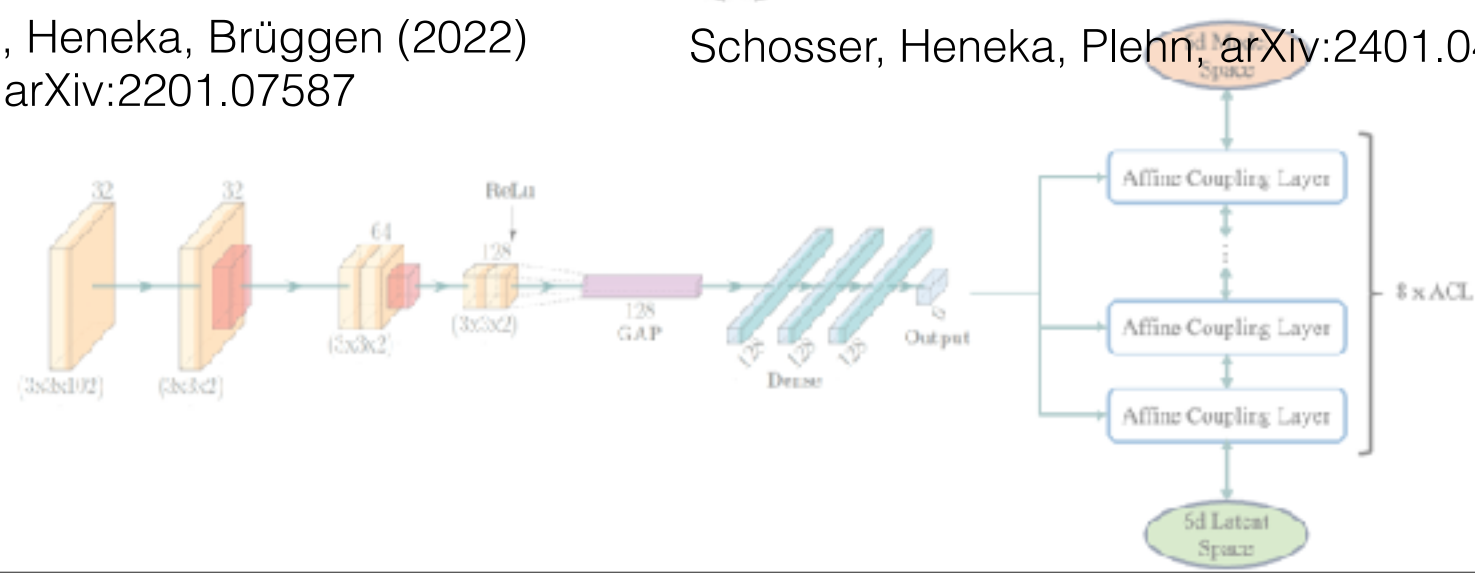
SBI with flows/cINN
based on BayesFlow
(arXiv:2003.06281)



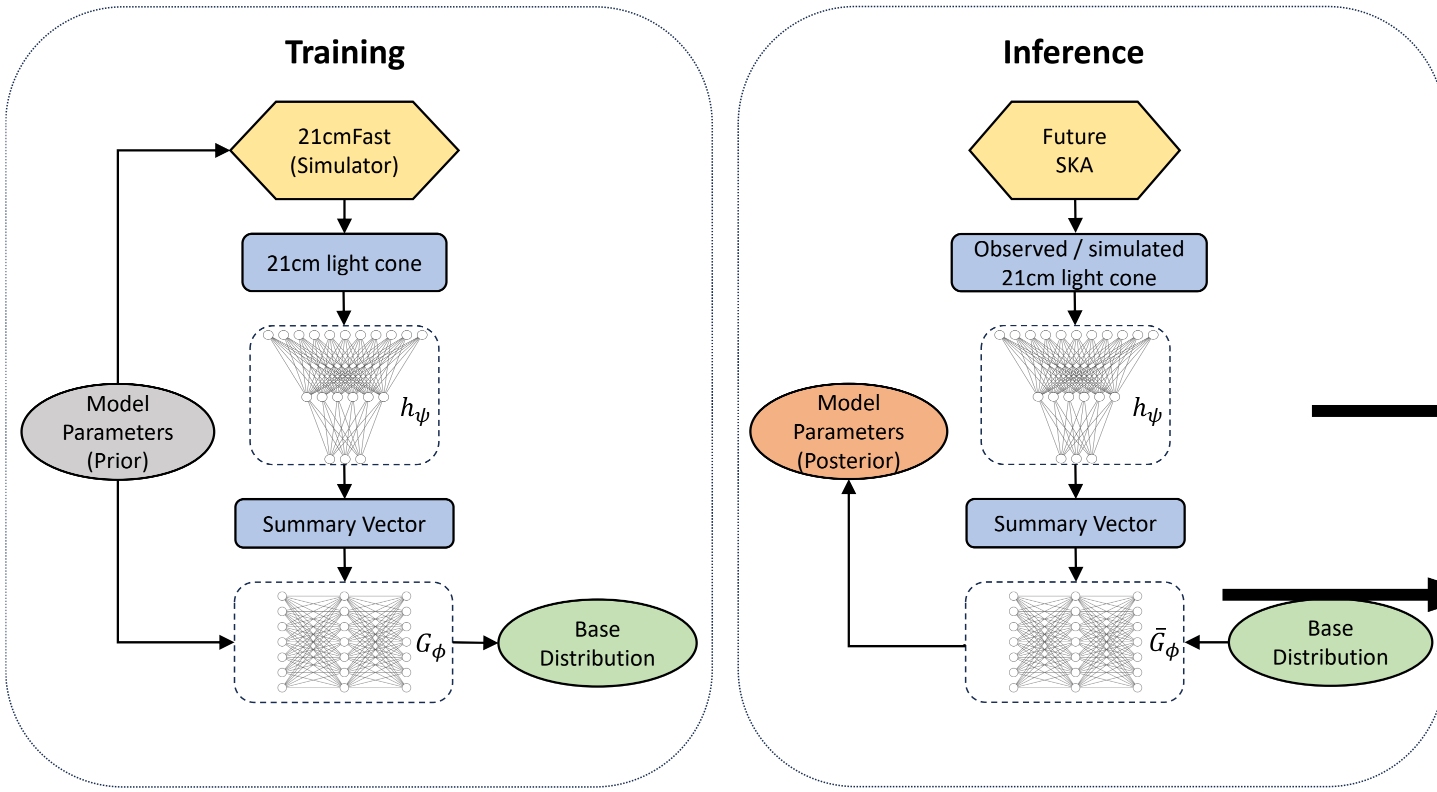
3D-21cmPIE-Net (public) + 21cm-cINN (public when published)

Neutsch, Heneka, Brüggen (2022)
arXiv:2201.07587

Schosser, Heneka, Plehn, arXiv:2401.04174



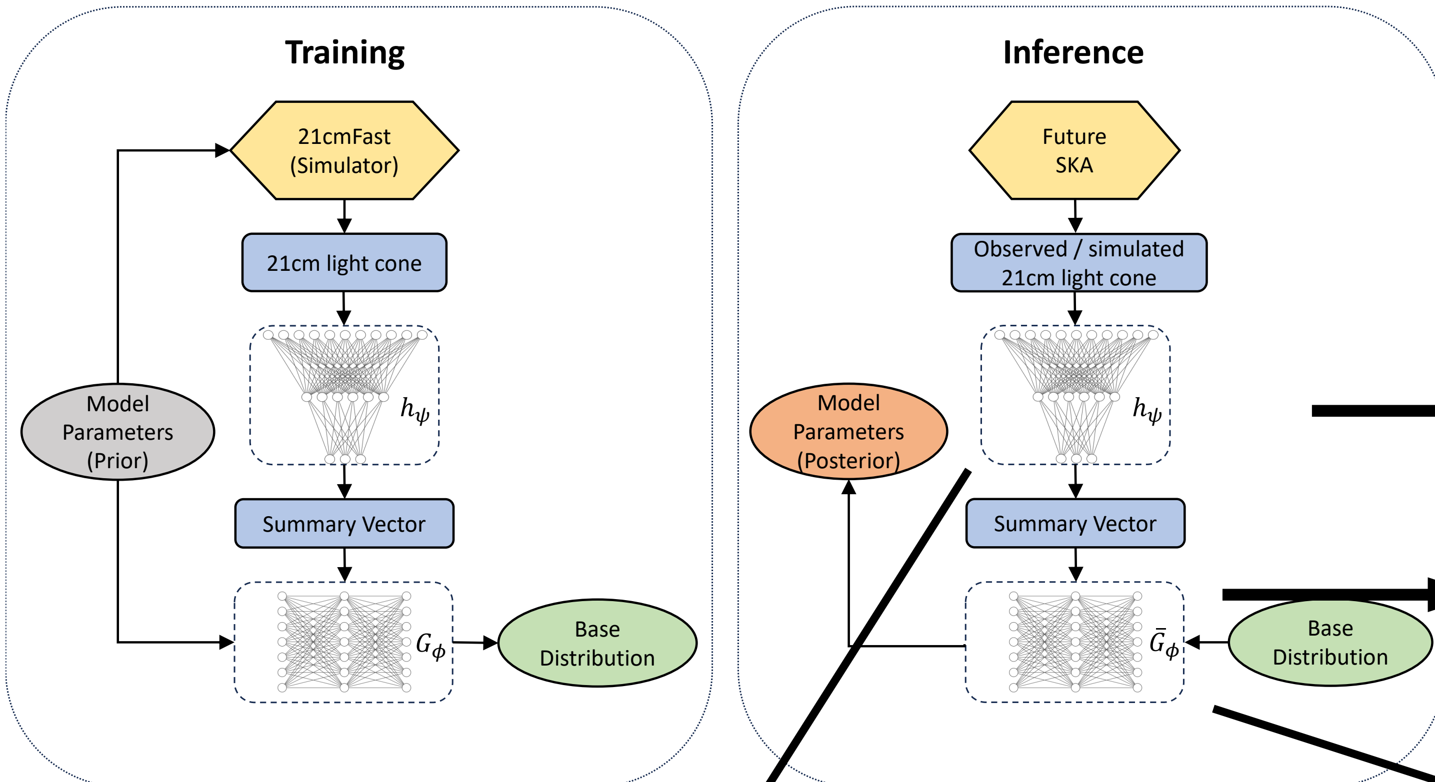
3) Simulation-based inference (SBI) for intensity mapping (3D)



NN1: Dimension reduction

NN2: Learning of distribution

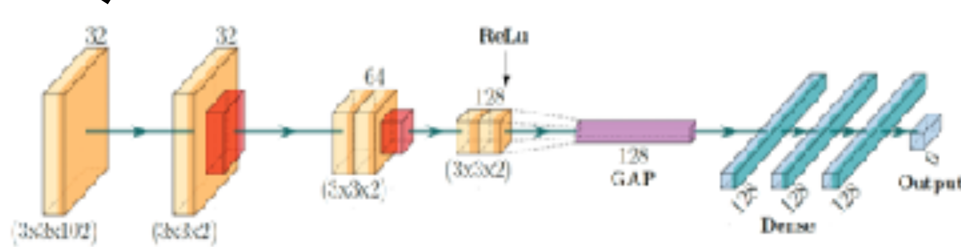
3) Simulation-based inference (SBI) for intensity mapping (3D)

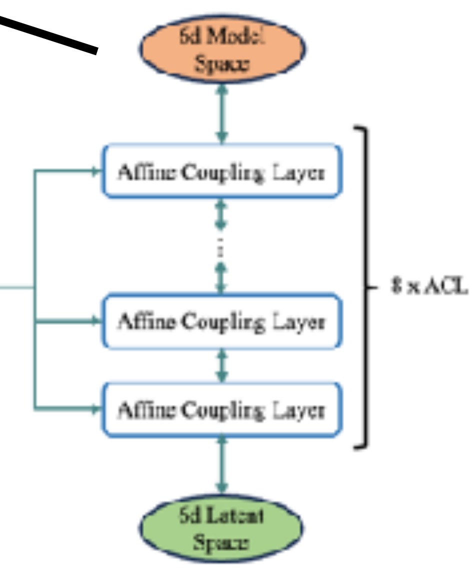


NN1: Dimension reduction

NN2: Learning of distribution

Network Model:





NN1: 3D-CNN

Small networks,
comparably few parameters
= fast training and convergence

3D-21cmPIE-Net (public)

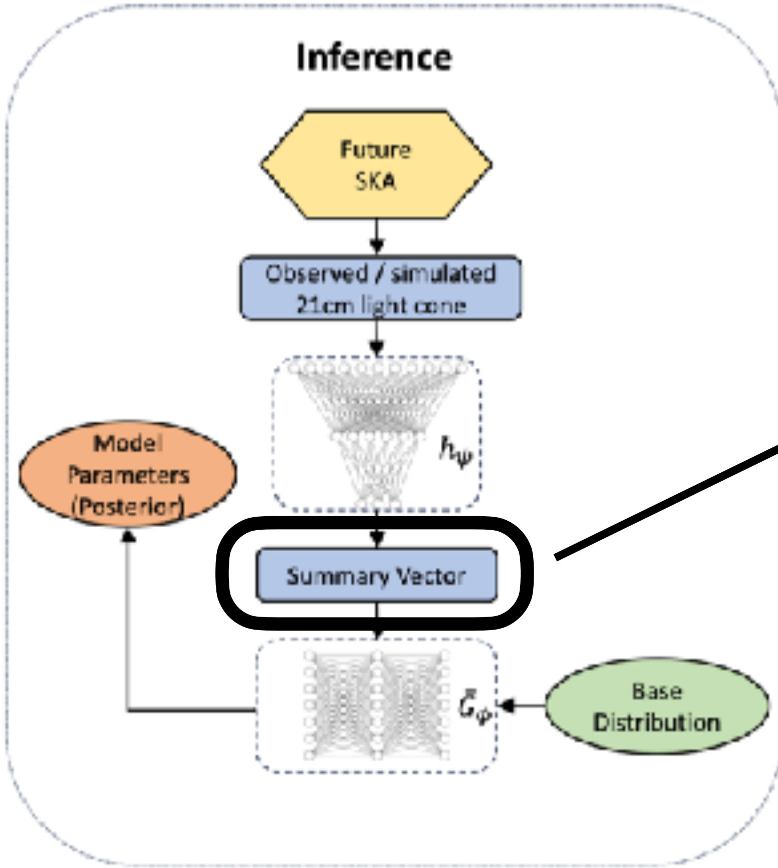
Neutsch, Heneka, Brüggen (2022)
arXiv:2201.07587

NN2: Invertible Network

21cm-cINN (public when published)

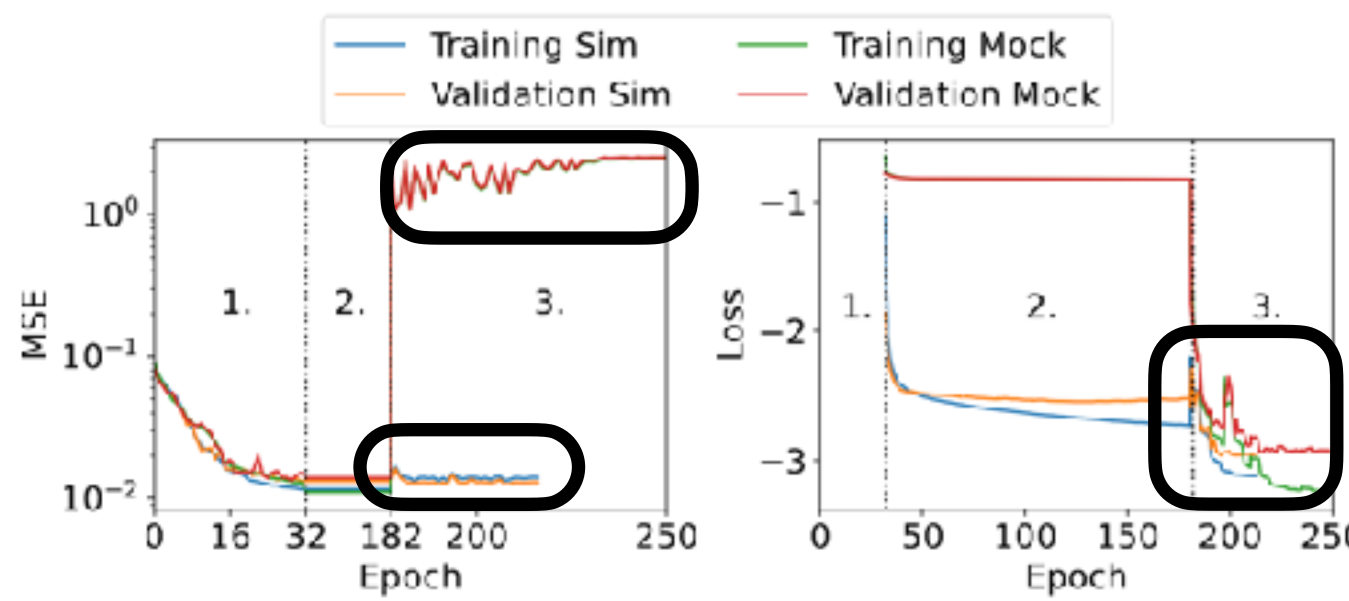
Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for intensity mapping (3D)



Summary vector versus summary statistics:
Free adjustment with scheduled training
= move away from 'parameter interpretation' of vector

Key difference between Sim and Mock:



We profit from learned summary in presence of noise (more).

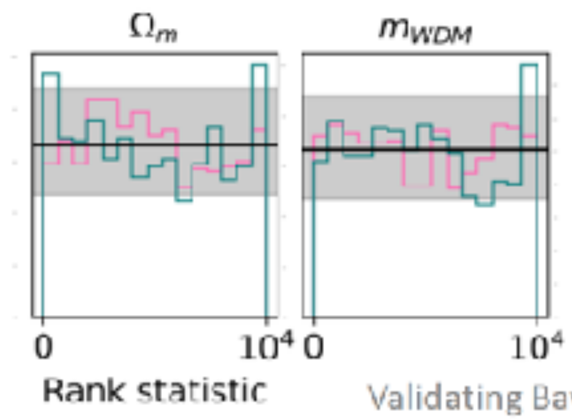
Sim: Summary stays close to original
Mock: Heavy adjustment of summary vector

Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for intensity mapping (3D)

Performance validation via:

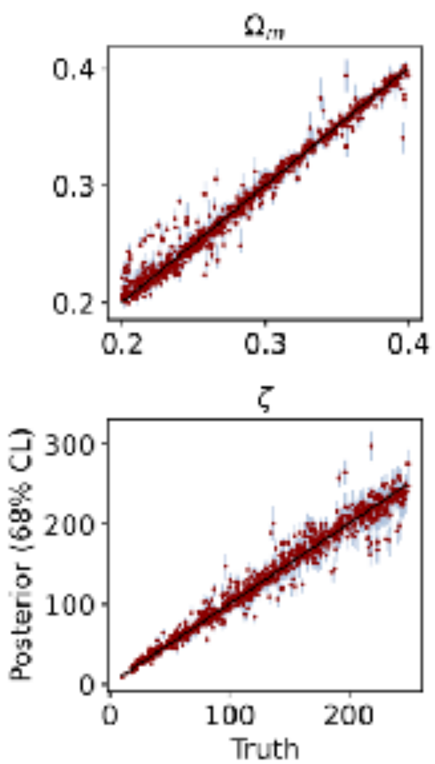
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery



Uniform distribution

Self-consistent sampling

Validating Bayesian Inference Algorithms with Simulation-Based Calibration, arXiv:1804.06788



Check Posterior vs. True label

Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for intensity mapping (3D)

Performance validation via:

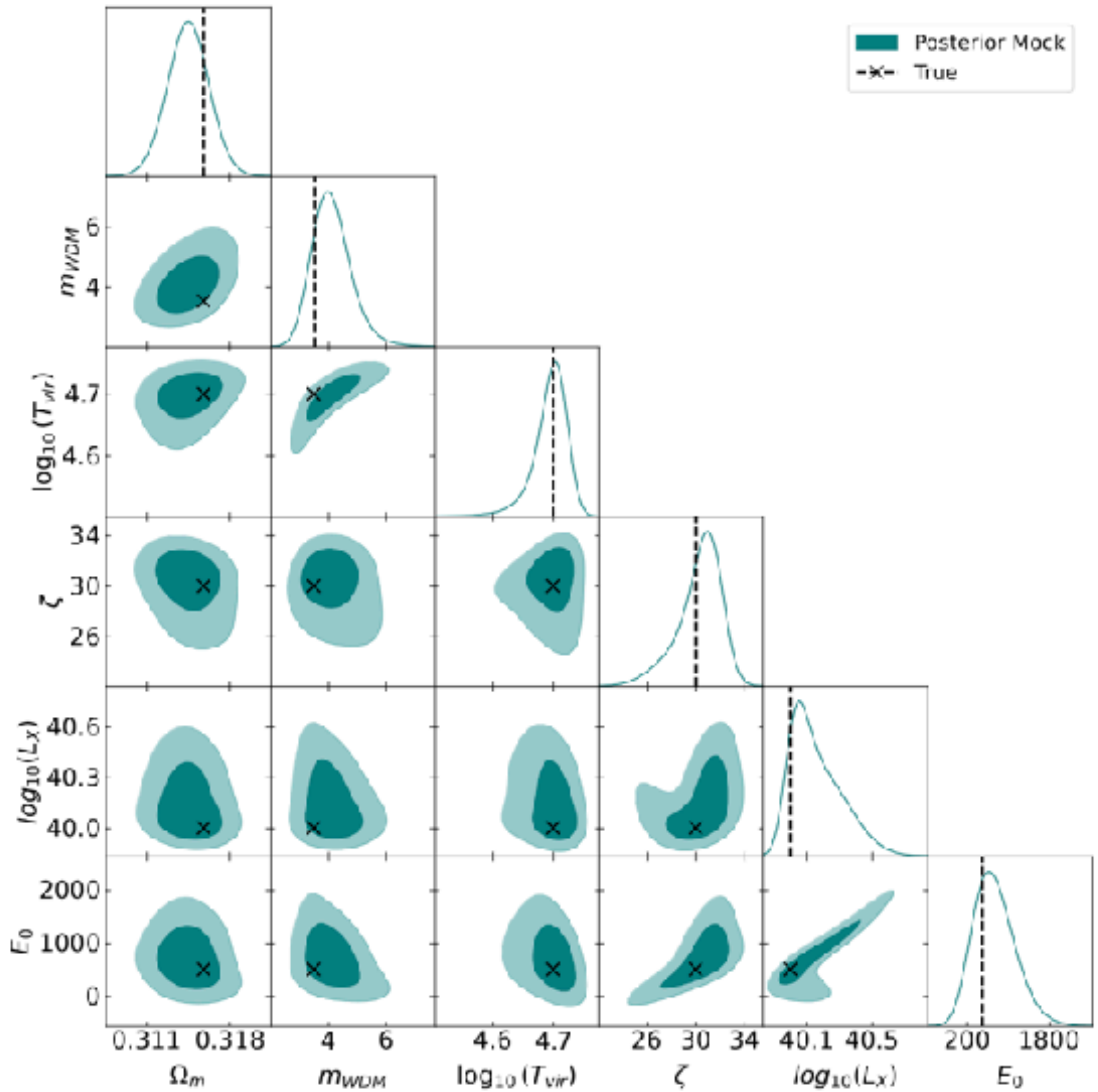
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery

Uniform distribution

<u>Likelihood contours:</u>

Explore any model in prior range!

+ similar performance
  Mock vs. Sim
  (except for E_0)



Schosser, Heneka, Plehn, arXiv:2401.04174

Check Posterior vs. True label

3) Simulation-based inference (SBI) for intensity mapping (3D)

Performance validation via:

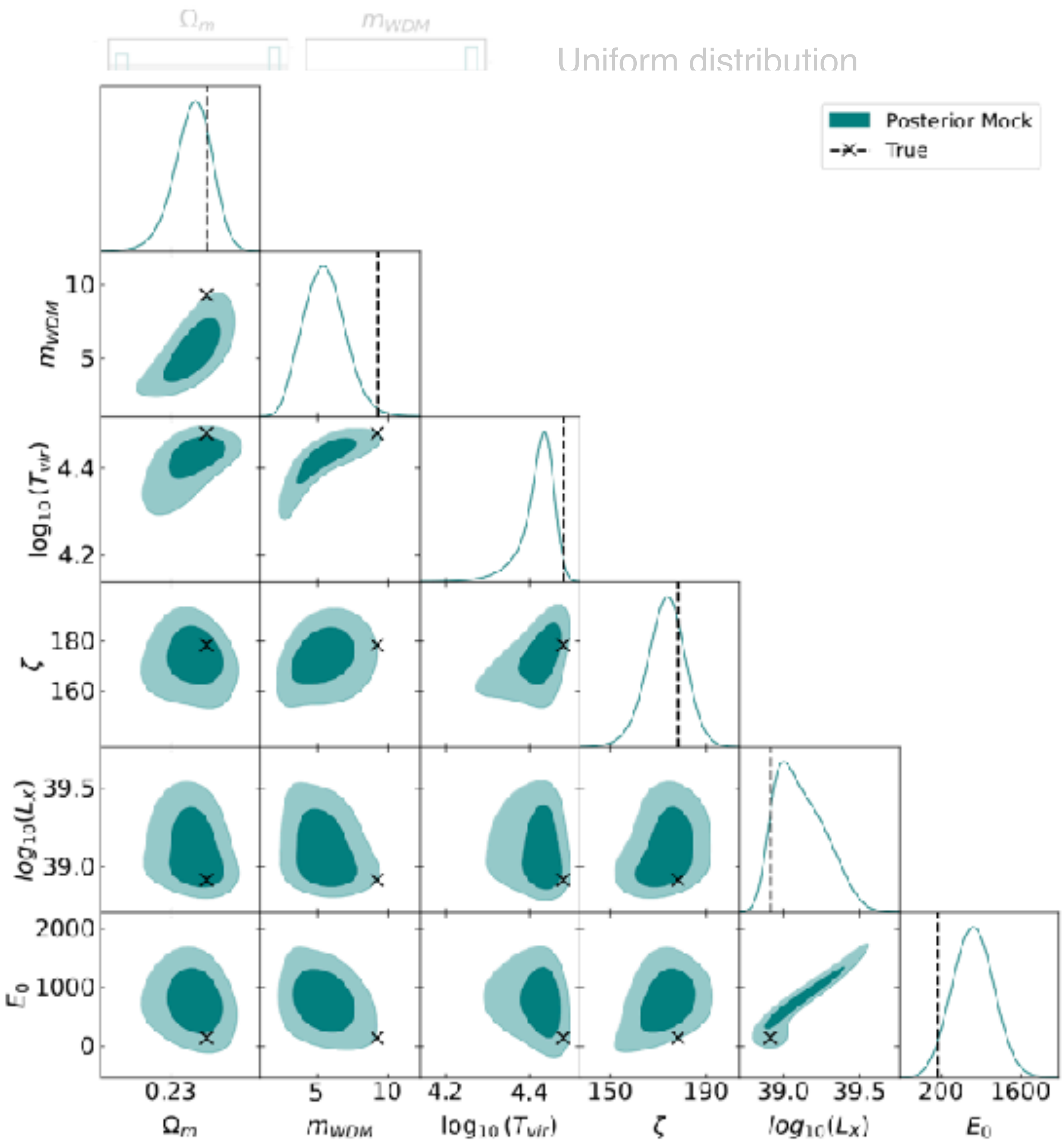
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery

<u>Likelihood contours:</u>

Explore any model in prior range!

\+ similar performance
Mock vs. Sim
(except for E_0)

Check Posterior vs. True label



Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for intensity mapping (3D)

Performance validation via:

- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery

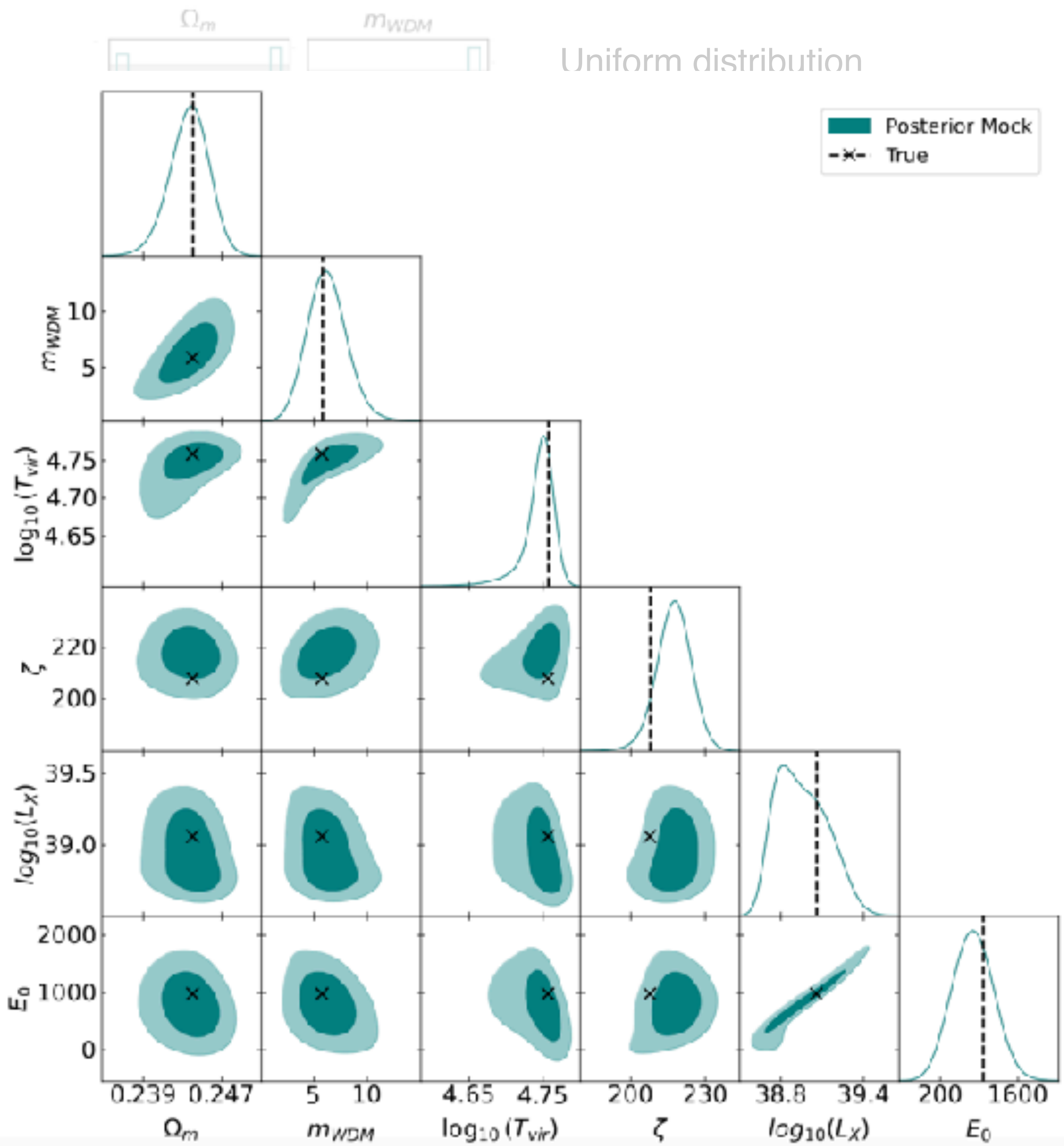
<u>Likelihood contours:</u>

Explore any model in prior range!

+ similar performance
  Mock vs. Sim
  (except for E_0)

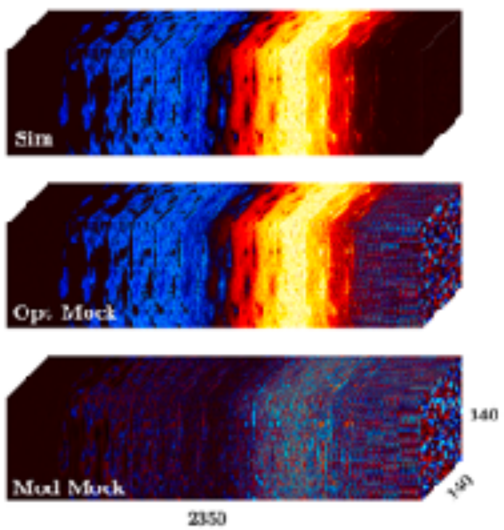
Check Posterior vs. True label

Uniform distribution

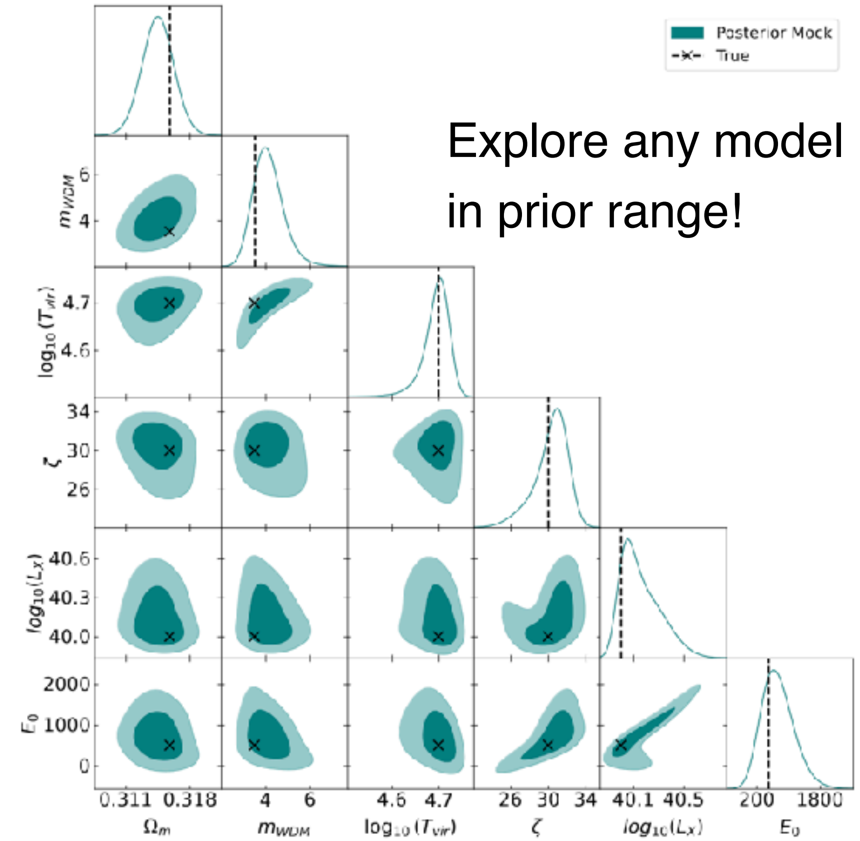


Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for intensity mapping (3D)



SBI with flows/cINN
based on BayesFlow
(arXiv:2003.06281)



Explore any model
in prior range!

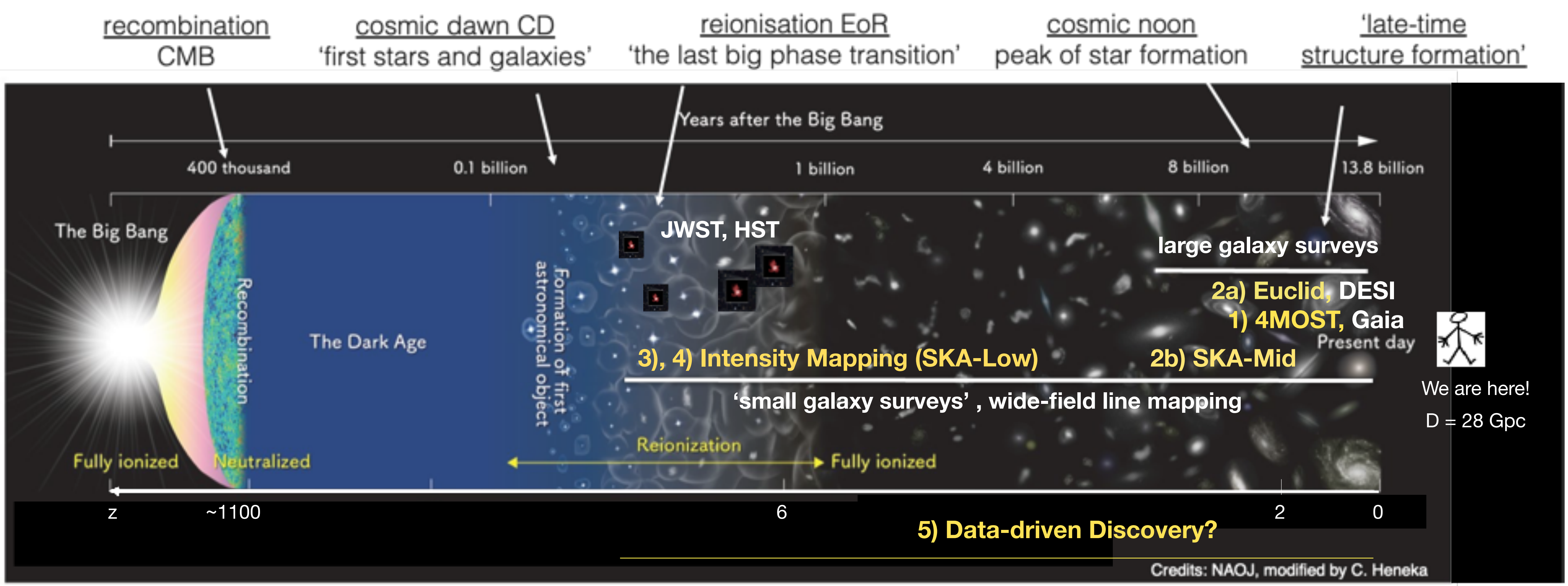
3D-21cmPIE-Net (public) + 21cm-cINN (public when published)

Neutsch, Heneka, Brüggen (2022)
arXiv:2201.07587

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

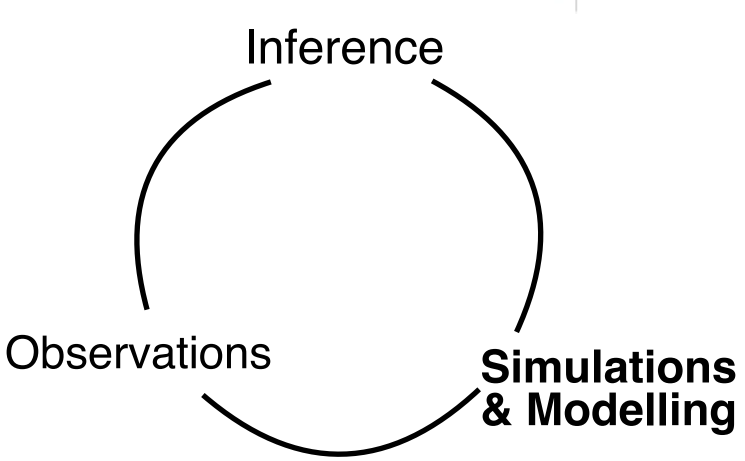
‘Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

1) Classification / Triggering
2) Source detection & characterisation
3) Simulation-based inference (SBI) in 3D
4) **Generative methods**
5) Data-driven Discovery



3) Generative methods for simulation

Reionization simulations are costly

Is there a fast way to emulate whole simulations?

+ Radiative transfer

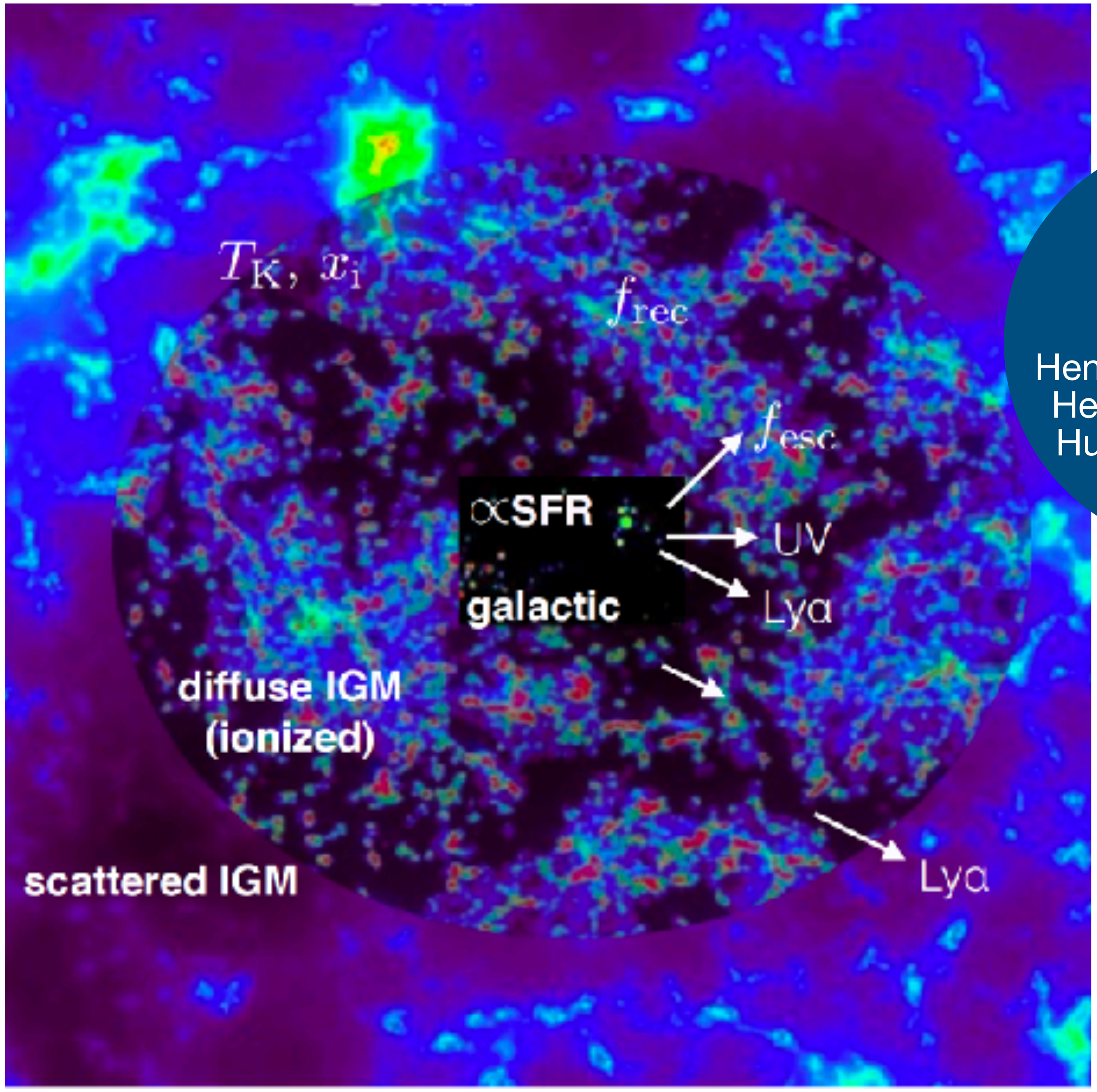
Hydrodynamical

Semi-numerical

Semi-analytical

Empirical scaling

Simulation Cost



Multi-line modelling

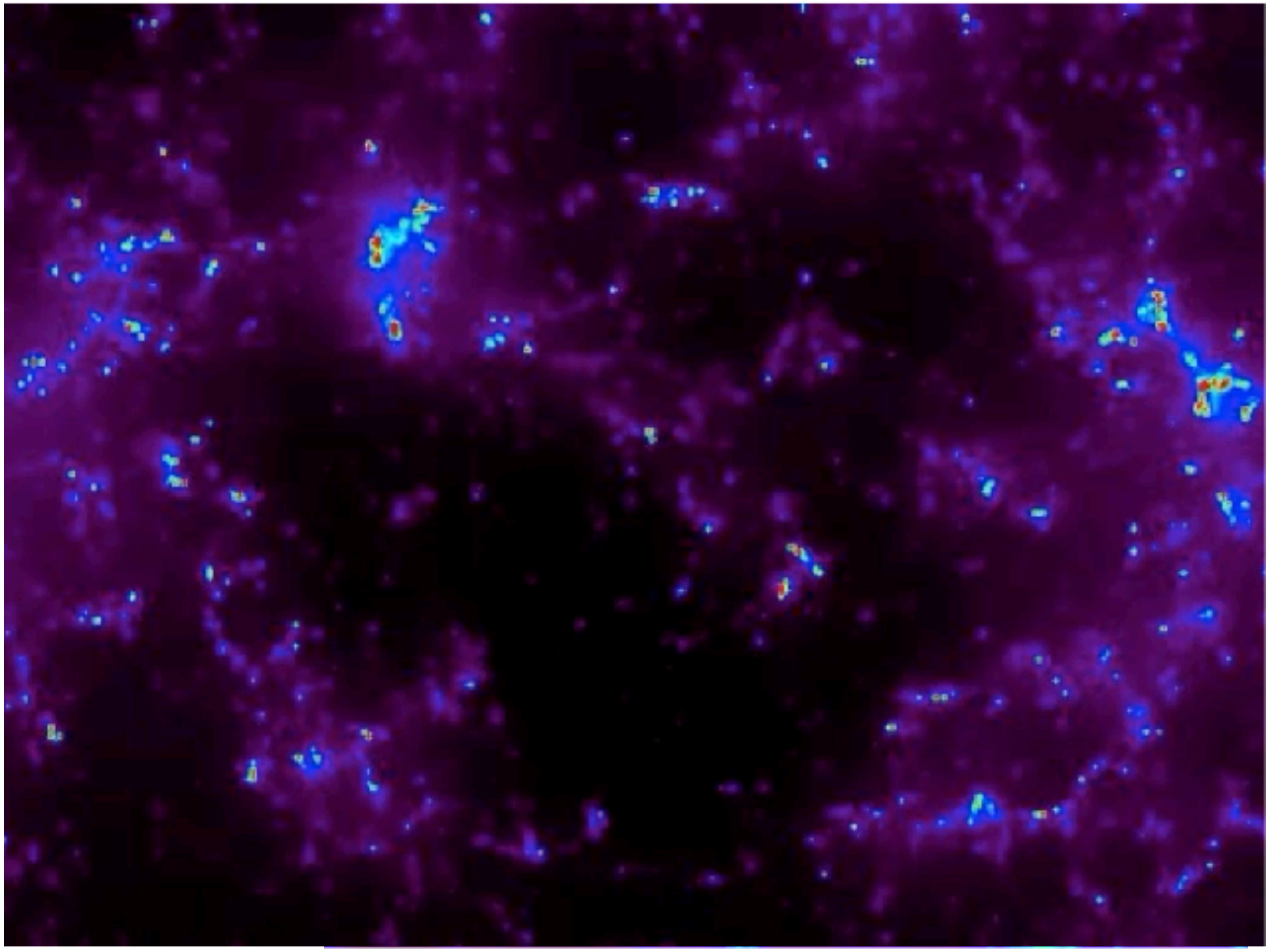
Heneka+17
Heneka, Mesinger 20
Heneka, Cooray 21
Hutter, Heneka+23

Example: Lyman-alpha

3) Generative methods for simulation

Reionization simulations are costly … and model a rich signal

Is there a fast way to emulate whole simulations?



Example: Lyman-alpha to 21cm

+ Radiative transfer

Hydrodynamical

Semi-numerical

Semi-analytical

Empirical scaling

Simulation Cost

Multi-line modelling

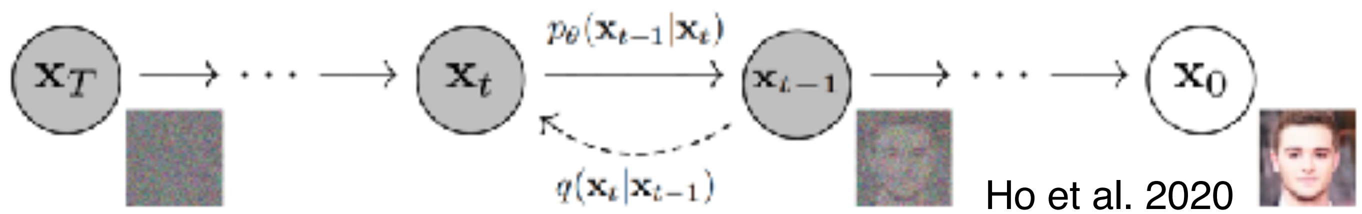
Heneka+17
Heneka, Mesinger 20
Heneka, Cooray 21
Hutter, Heneka+23

3) Generative methods for simulation (of the 21cm signal)

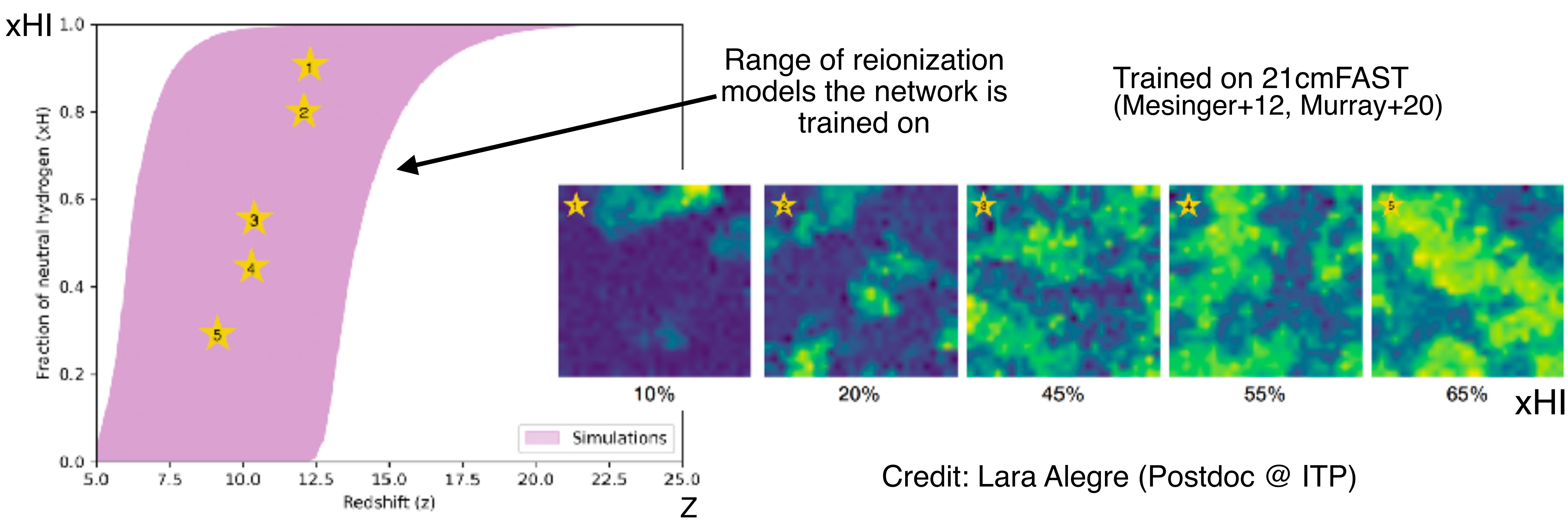
Reionization simulations are costly

Is there a fast way to emulate whole simulations?

Our DL solution: Generation of slices of the 21cm brightness temperature using diffusion models



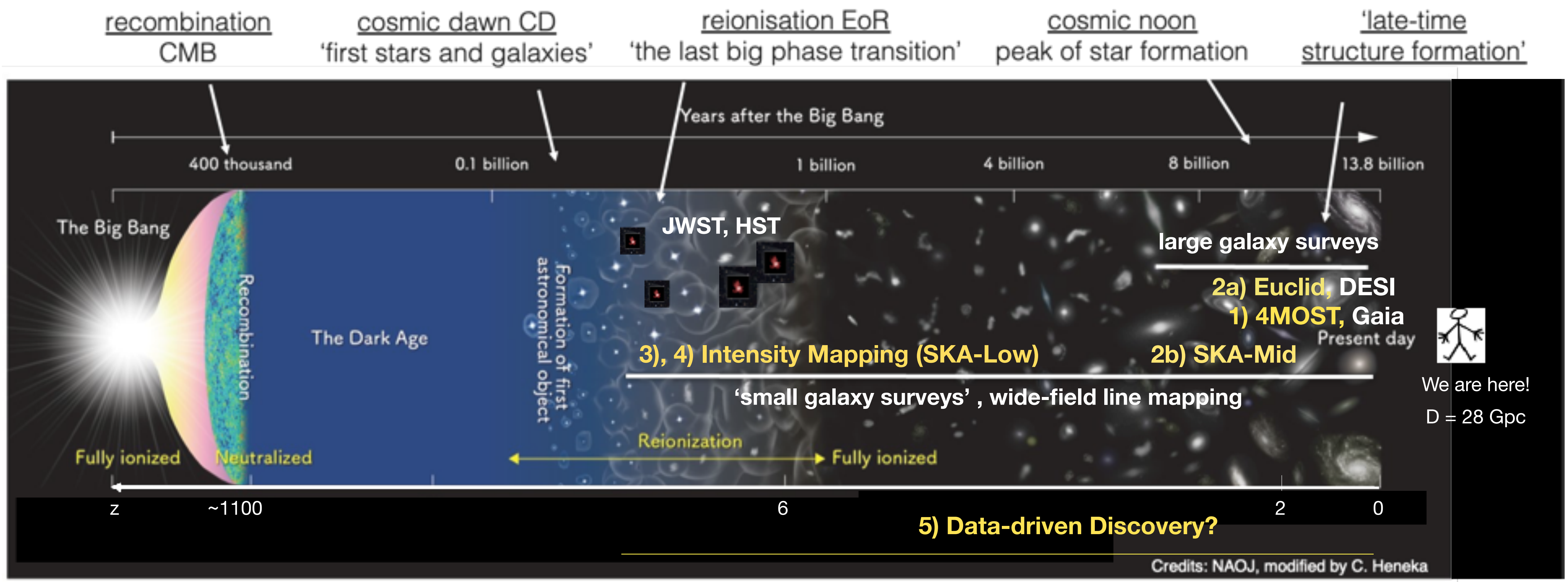
Ho et al. 2020



Trained on 21cmFAST (Mesinger+12, Murray+20)

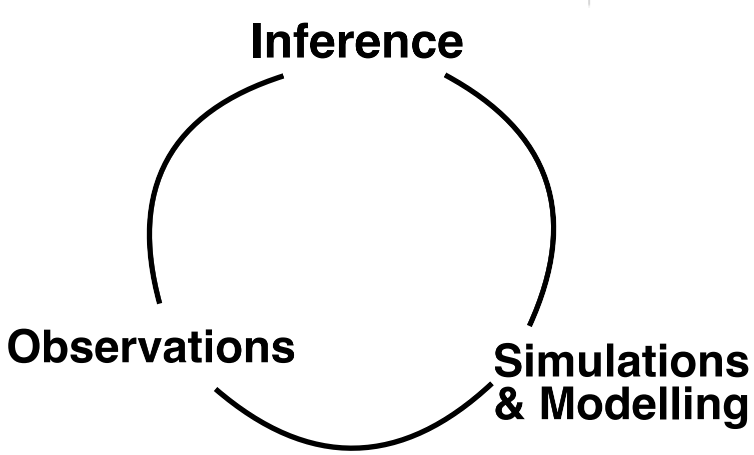
Credit: Lara Alegre (Postdoc @ ITP)

Astronomical and Astrophysical Machine Learning



Highlights in this Lecture

1) Classification / Triggering
2) Source detection & characterisation
3) Simulation-based inference (SBI) in 3D
4) Generative methods
5) **Data-driven Discovery**



5) Already now: Data-driven discovery

CAMELS

= Cosmology and Astrophysics with MachinE Learning Simulations

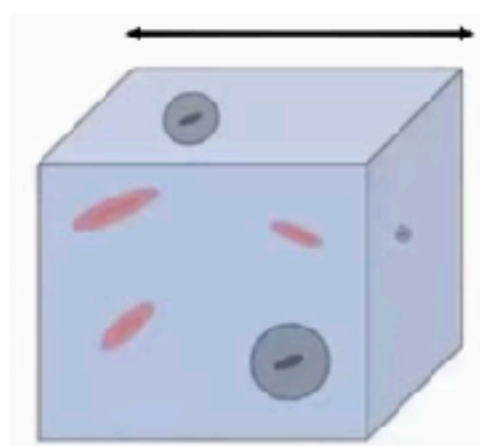
https://camels.readthedocs.io

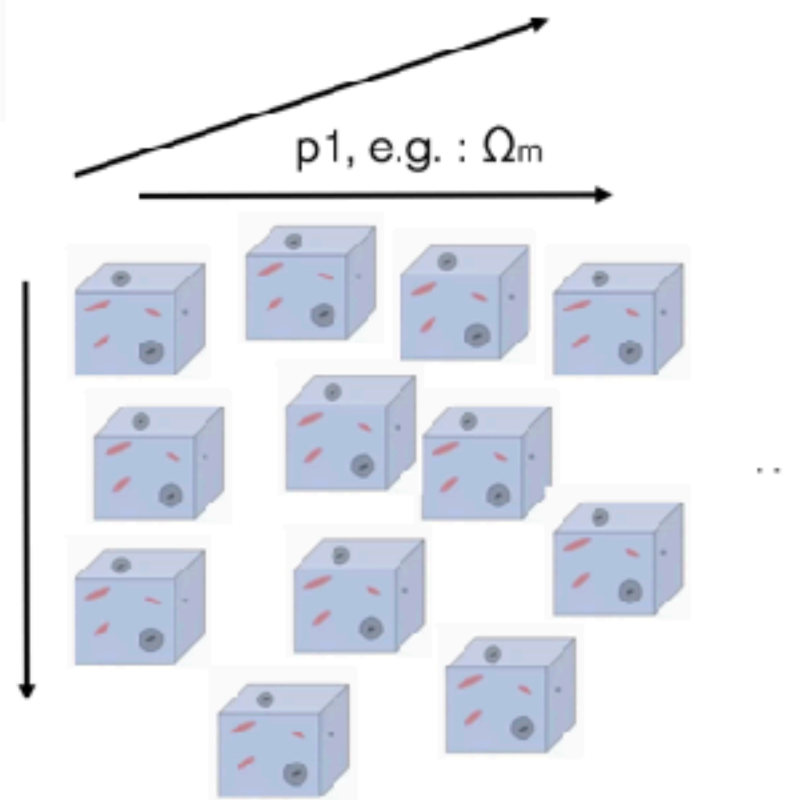
Type	Code	Subgrid model	Simulations
Hydrodynamic	Arepo	IllustrisTNG	1,092
Hydrodynamic	Gizmo	SIMBA	1,092
Hydrodynamic	MP-Gadget	Astrid	1,092
N-body	Gadget-III	–	3,049

Parameter set:

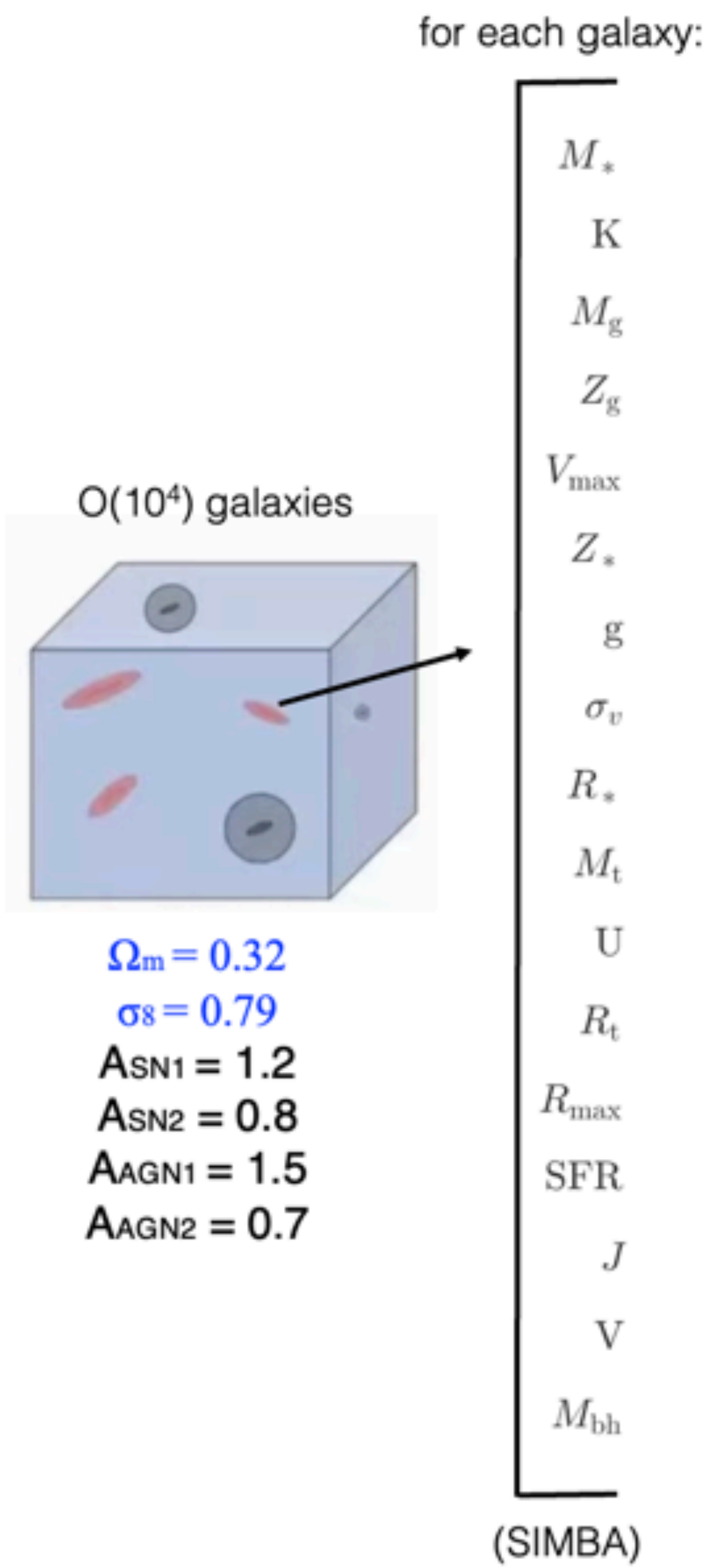
$$
\begin{aligned}
0.1 \leq \Omega_{\mathrm{m}} &\leq 0.5 \\
0.6 \leq \sigma_8 &\leq 1.0 \\
0.25 \leq A_{\mathrm{SN1}} &\leq 4.0 \\
0.50 \leq A_{\mathrm{SN2}} &\leq 2.0 \\
0.25 \leq A_{\mathrm{AGN1}} &\leq 4.0 \\
0.50 \leq A_{\mathrm{AGN2}} &\leq 2.0
\end{aligned}
$$

25cMpc/h





5) Already now: Data-driven discovery



How many galaxies do we need to constrain e.g. Ω_m?

Let's start with one!

5) Already now: Data-driven discovery

= Cosmology and Astrophysics with MachinE Learning Simulations

https://camels.readthedocs.io

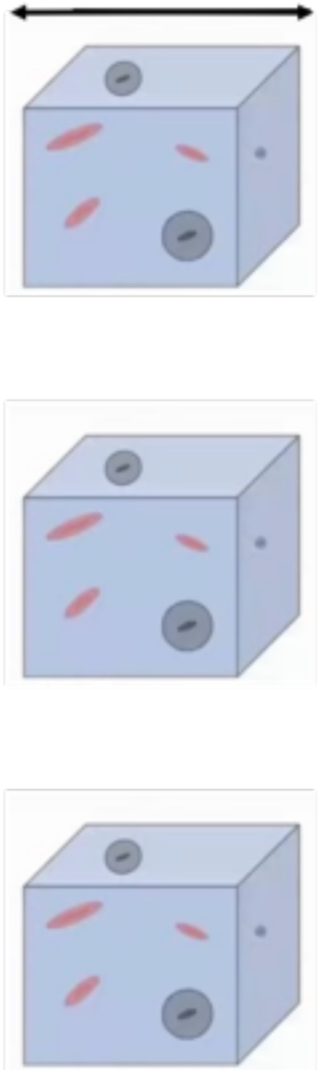
How many galaxies do we need to constrain e.g. Ω_m?

25cMpc/h

$\Omega_m = 0.27$
$\sigma_8 = 0.85$
$A_{SN1} = 0.7$
$A_{SN2} = 0.9$
$A_{AGN1} = 1.4$
$A_{AGN2} = 0.8$

$\Omega_m = 0.4$
$\sigma_8 = 0.65$
$A_{SN1} = 1.1$
$A_{SN2} = 0.9$
$A_{AGN1} = 1.1$
$A_{AGN2} = 0.9$

$\Omega_m = 0.2$
$\sigma_8 = 0.95$
$A_{SN1} = 1.8$
$A_{SN2} = 0.5$
$A_{AGN1} = 3.0$
$A_{AGN2} = 1.5$

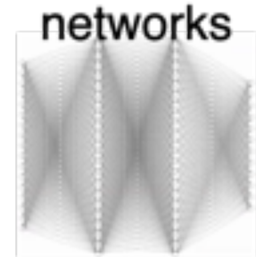


O(10^4) galaxies per cube

random galaxy: → galaxy properties

$$\begin{bmatrix} M_* \\ \mathrm{K} \\ M_g \\ Z_g \\ V_{max} \\ Z_* \\ g \\ \sigma_v \\ R_* \\ M_t \\ \mathrm{U} \\ R_t \\ R_{max} \\ \mathrm{SFR} \\ J \\ \mathrm{V} \\ M_{bh} \end{bmatrix}$$

Moment density networks



arXiv:2011.05991

→ ?

Ω_m
$\delta\Omega_m$ ~10% uncertainty

recovery of matter density for very different masses (and environments)
also: holds for redshifts other than z=0

connections on a high dimensional manifold?

5) Already now: Data-driven discovery

Loss based on moment density networks (MDN) Jeffrey & Wandelt 2011 arXiv:2011.05991

MDN idea: hierarchy of neural regression models (mean → variance → skewness →...)

We begin by noting that if we find some function of our data $\mathcal{F}(\boldsymbol{x})$ that minimizes an L_2 loss over the distribution of possible training examples $\{\boldsymbol{x}_i, \boldsymbol{\theta}_i\}$,

$$J_0 = \int ||\boldsymbol{\theta} - \mathcal{F}(\boldsymbol{x})||^2 p(\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\theta} \quad , \tag{4}$$

then $\mathcal{F}$, which we represent as a neural network, evaluated for the observed data is the mean of the posterior distribution $\mathcal{F}(\boldsymbol{x}_{obs}) = \langle \boldsymbol{\theta} \rangle_{\boldsymbol{\theta}|\boldsymbol{x}_{obs}}$. It is therefore possible to create a hierarchy of networks

In practice we minimise the following loss function:

$$\mathcal{L} = \sum_{i=1}^{6} \log\left(\sum_{j\in\mathrm{batch}} (\theta_{i,j} - \mu_{i,j})^2\right) + \sum_{i=1}^{6} \log\left(\sum_{j\in\mathrm{batch}} \left((\theta_{i,j} - \mu_{i,j})^2 - \sigma_{i,j}^2\right)^2\right)$$

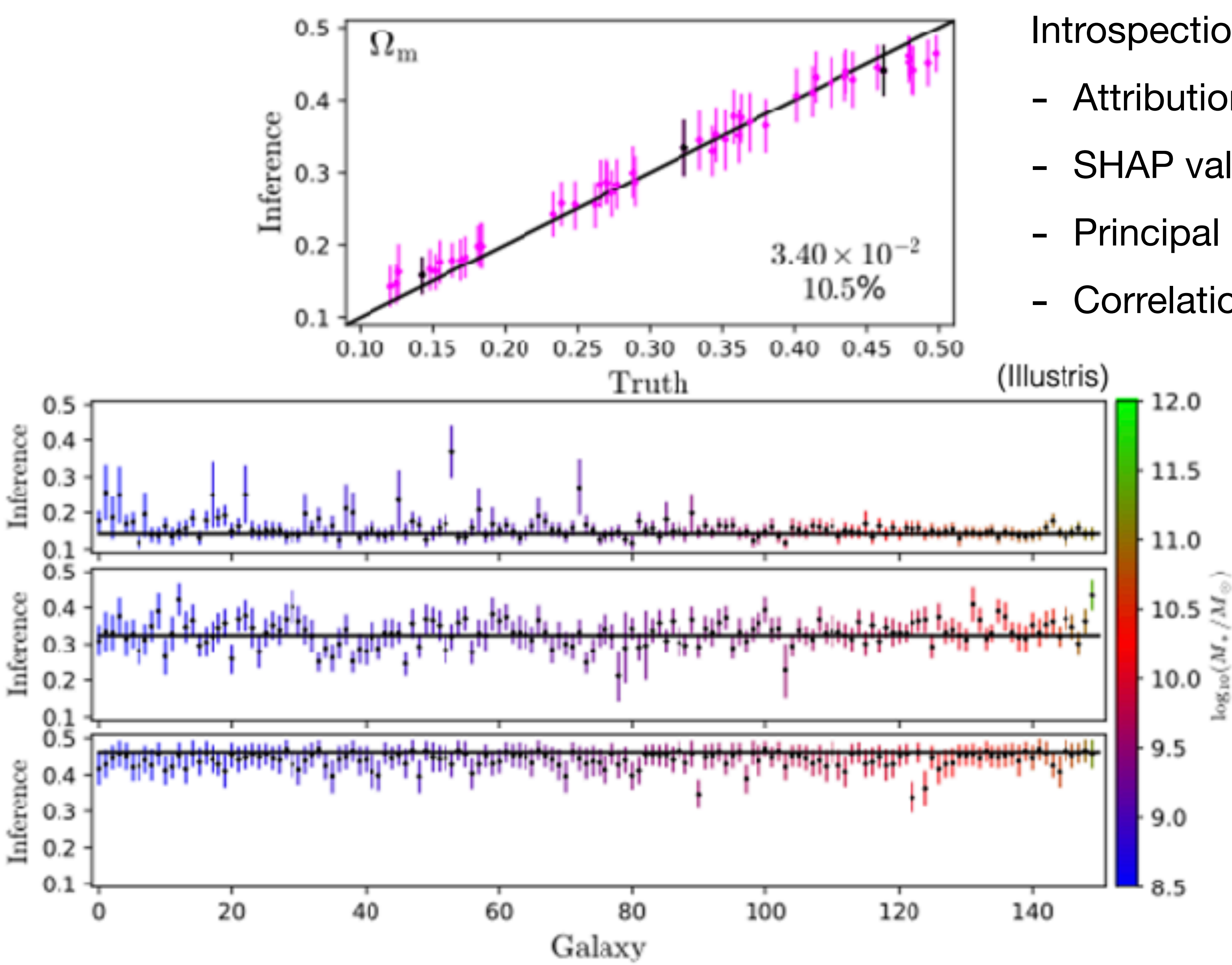
Our model F(x):

CNN layers (19)

+ dense (2)

Villaescusa-Navarro+ arXiv:2109.10915

5) Already now: Data-driven discovery



Introspection via:

- Attribution, one-by-one retraining
- SHAP values (Shapley Additive exPlanations)
- Principal Component Analysis (PCA)
- Correlation strengths

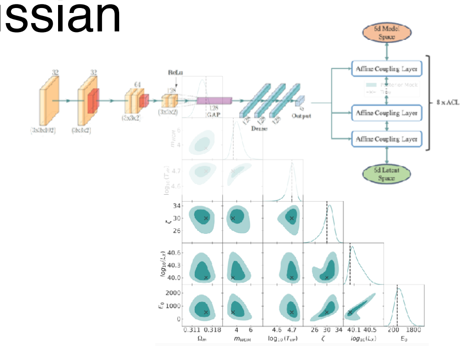
→ recovery of matter density for very different masses (and environments)

also: holds for redshifts other than z=0

Summary: Where we stand

Goal: Understanding & discovery

Inference
SBI in 3D from non-Gaussian
tomographic data



Inference
& Analysis

Data-driven **Discovery**

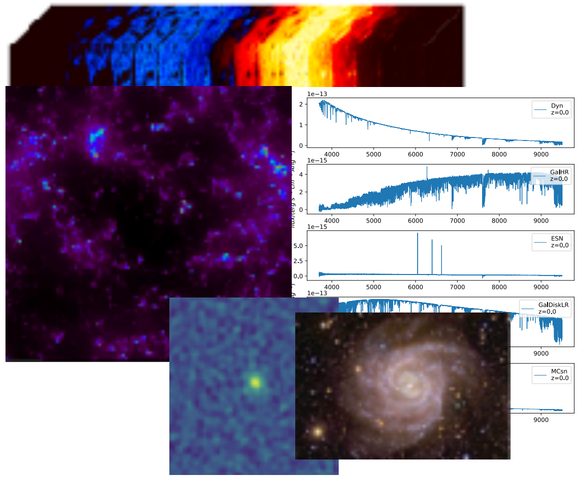
In astronomy & astrophysics

Observations
& Data

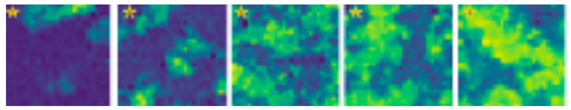
Simulations
& Modelling

Detection & Characterisation
Unbiased measurements from
diverse sources (galaxies)

Classification
Online classifier and triggering



Simulation
Produce large range of
reionization topologies

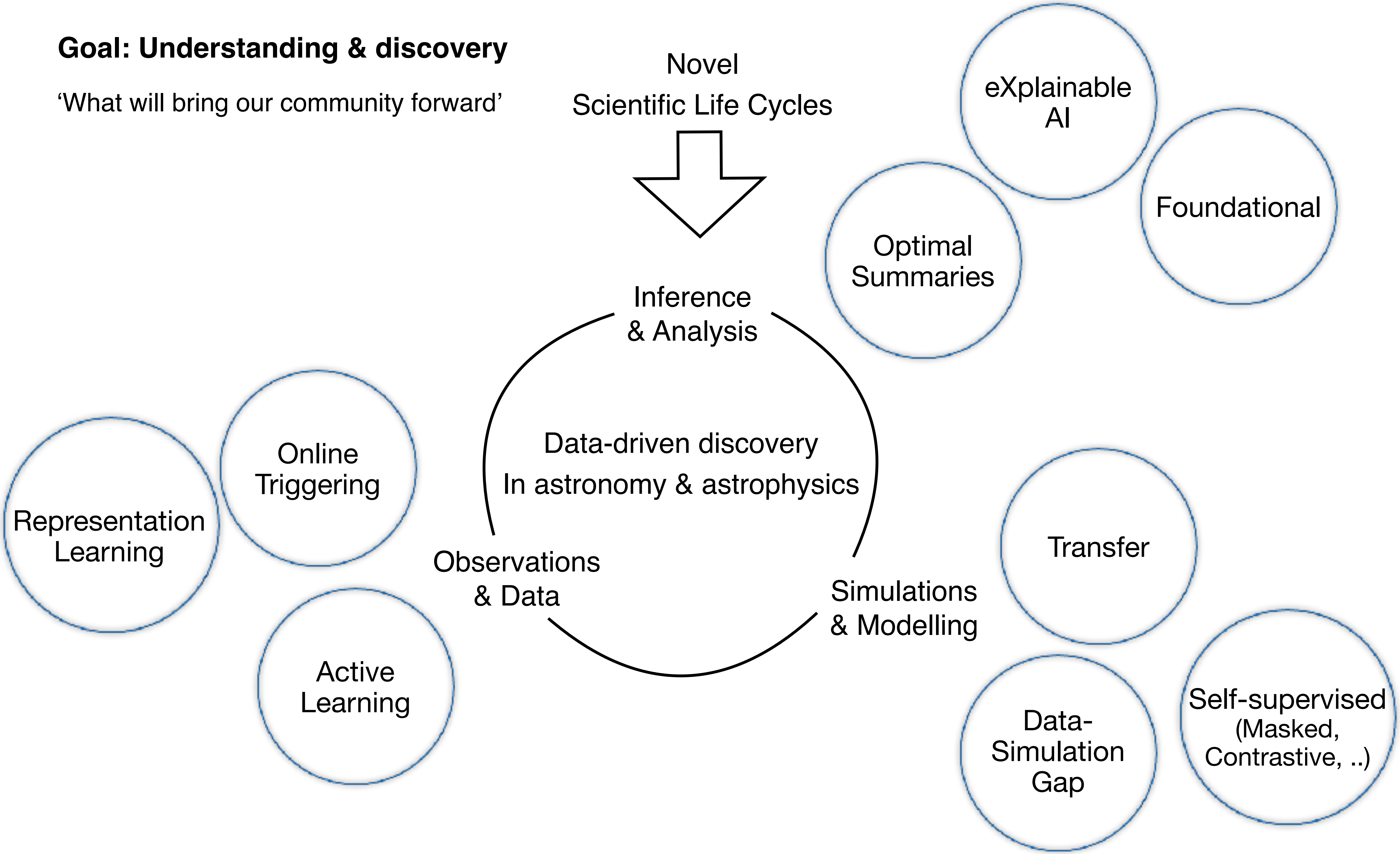


Select publications:
Neutsch, Heneka, Brüggen 22, arXiv:2201.07587
Schosser, Heneka, Plehn, arXiv:2401.04174
Hartley+ 23 (incl. Heneka), arXiv:2303.07943
Heneka 23, arXiv:2311.17553
Boucaud, Huertas-Company, Heneka+ 20, arXiv:1905.01324
Zhong, Napolitano, Heneka+24, arXiv:2311.04146

Research plans: Where we are going

Goal: Understanding & discovery

‘What will bring our community forward’



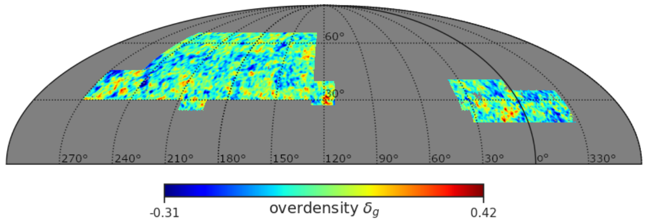
Example: Robust Foundational models for inference

Data-Simulation Gap

Generation & modelling

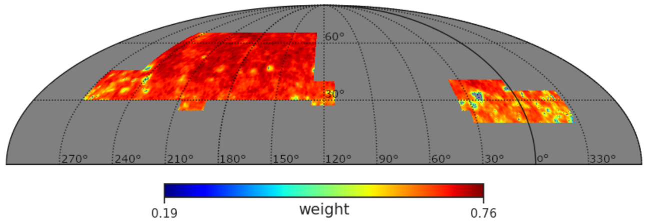
LOFAR observations

1) simulations, mock catalogues + empirical relations



Mock observations

LoTSS maps



Currently:

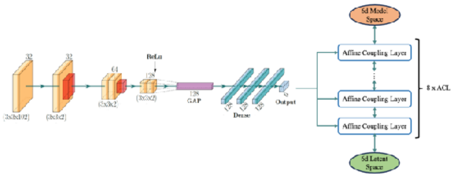
Augmentation, semi-supervised, importance weighing

2) Transfer

Comparison to 'random mocks'

Optimal Summaries

Foundational ?



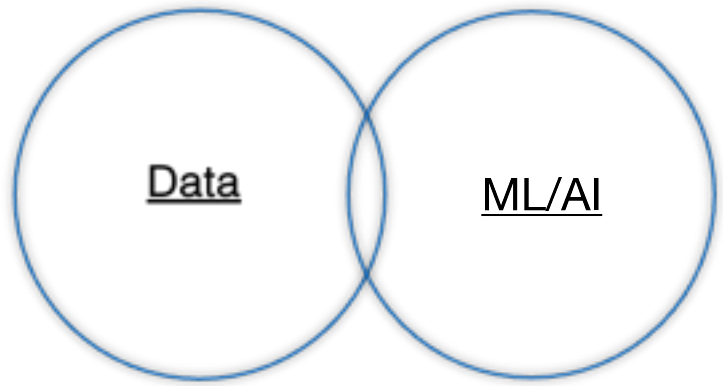
Simulation-based inference

3) Inference

C_ℓ^{gg} + jackknife + MCMC

Research plans: where we are moving towards

Use data from astronomical surveys for **data-driven discovery.**



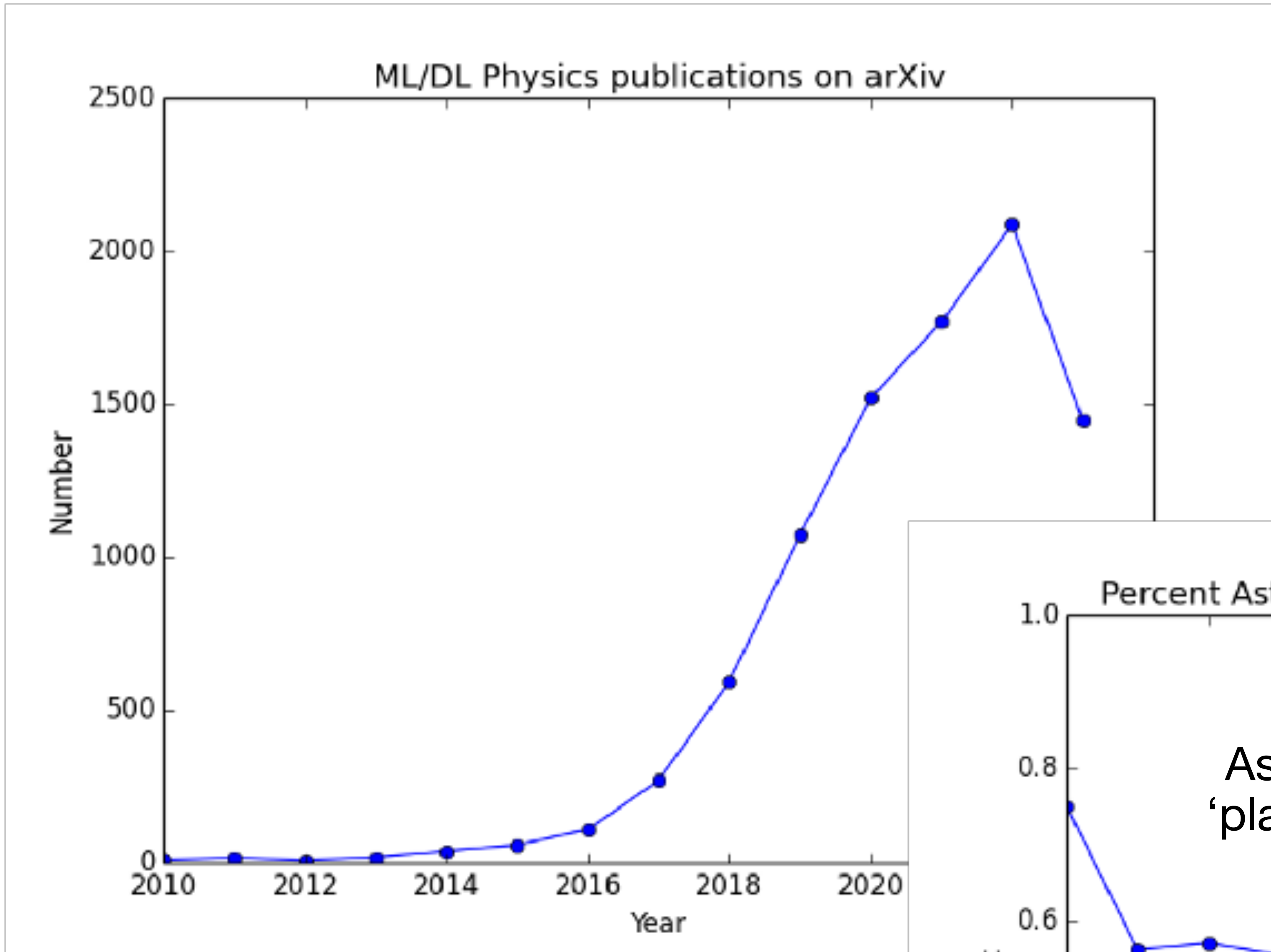
Goals:
Map-based, multi-channel, ‘3D’ approach to astrophysics
Representation learning, beyond summaries
Automated data mining, anomaly search
New signatures & discovery

Novel
Scientific
Life Cycles

Based on
foundational models

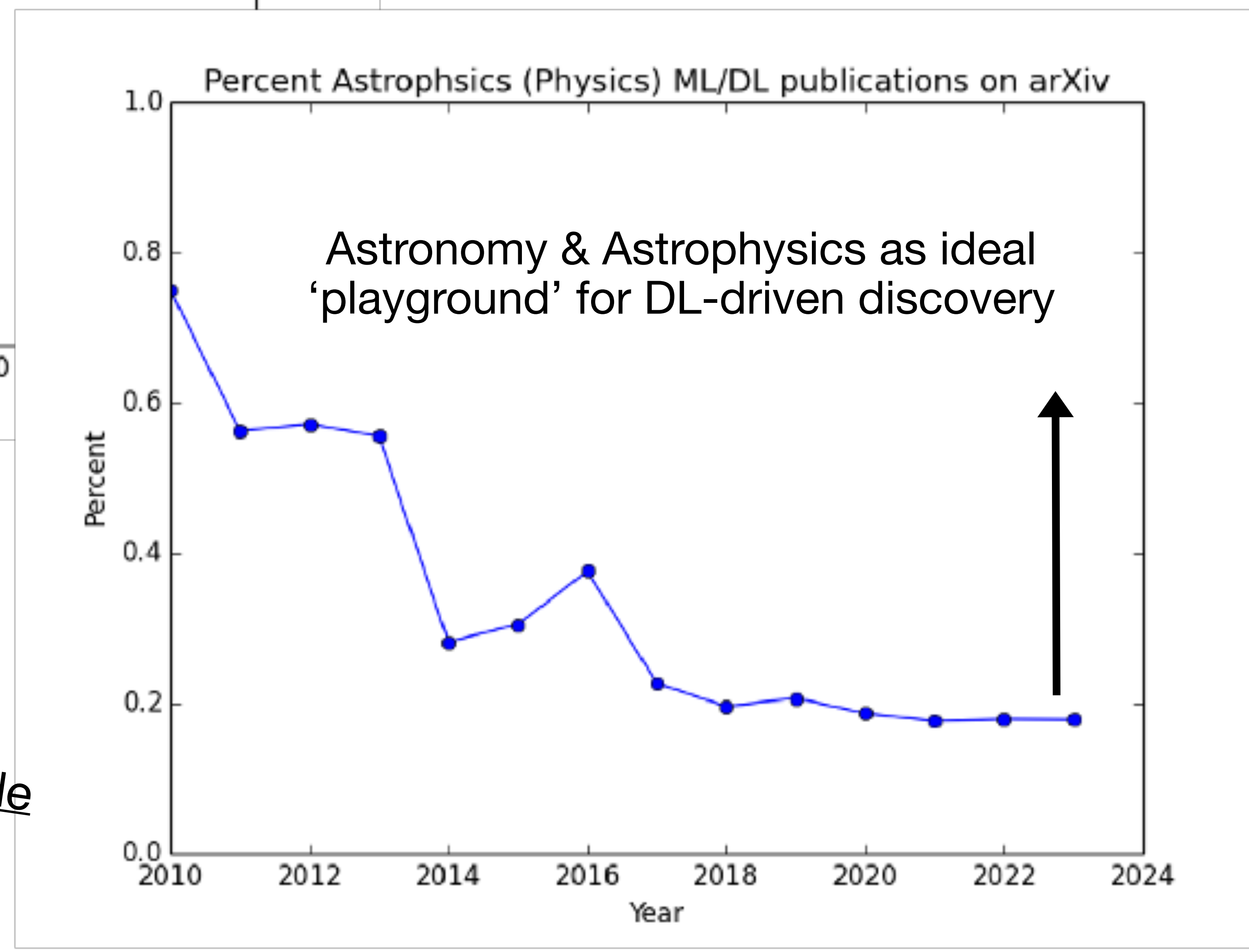
Outcomes:
Fast and expressive simulators
Reliable error estimates & inference
Interpretability
Active interaction and learning
‘human in the loop’
New signatures & discovery

ML/DL/AI has come to stay when dealing with astronomical data.



Some advertisement:

We have ‘endless’ open-source public data



Astronomy & Astrophysics as ideal ‘playground’ for DL-driven discovery

Thank you for your attention!

heneka@thphys.uni-heidelberg.de