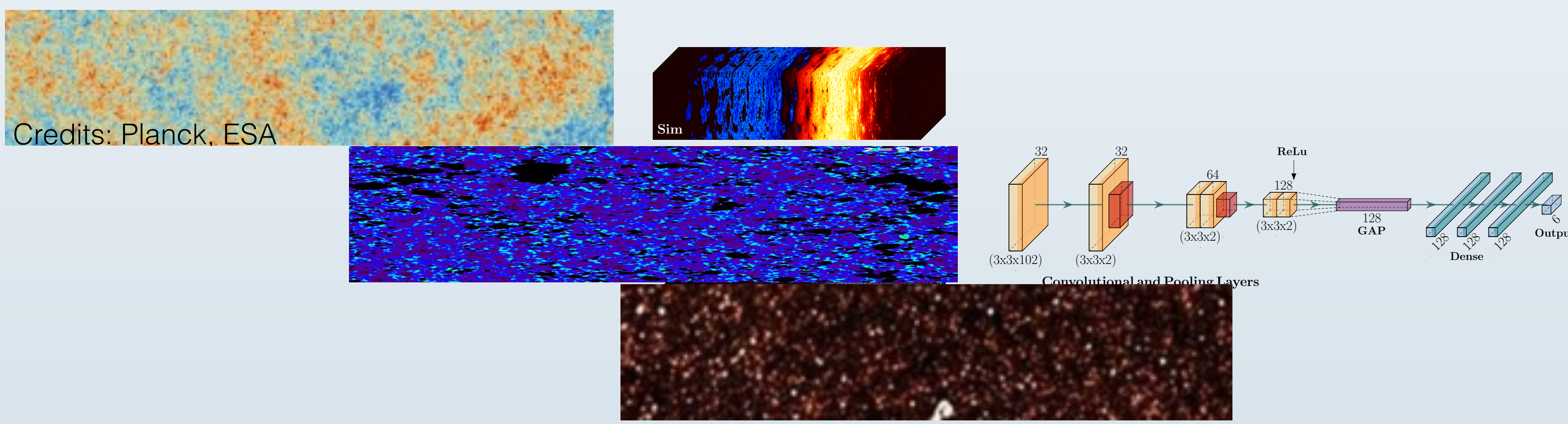
Astrophysics & ML/DL

(At least some examples)

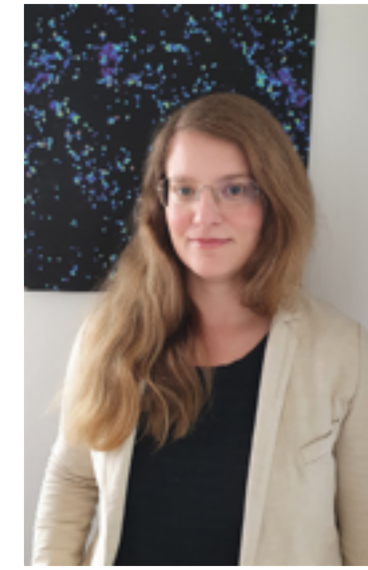


Caroline Heneka
ITP Heidelberg

Machine Learning in Particle Theory, Mainz/Oppenheim, July 17th 2023

FREIGEIST FELLOWSHIP DER VOLKSWAGENSTIFTUNG
ITP
Universität Heidelberg
Zukunft. Seit 1386.

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)

Statistical Tools

Bayesian
Bias detection
Model comparison
Hierarchical

Forecasting
Fisher

Sampling & Inference
MCMC

Simulations
Computational
Astrophysics
Reionisation

Data
Imaging
Spectral
3D cubes
Time series

Modelling/ Theory
IM
Cosmological
signatures
Sources

Machine learning

Neural Networks
Convolutional
Generative
Classification
Detection
Inference

Maps
Trees

Meta-Heuristics
Genetic
algorithm

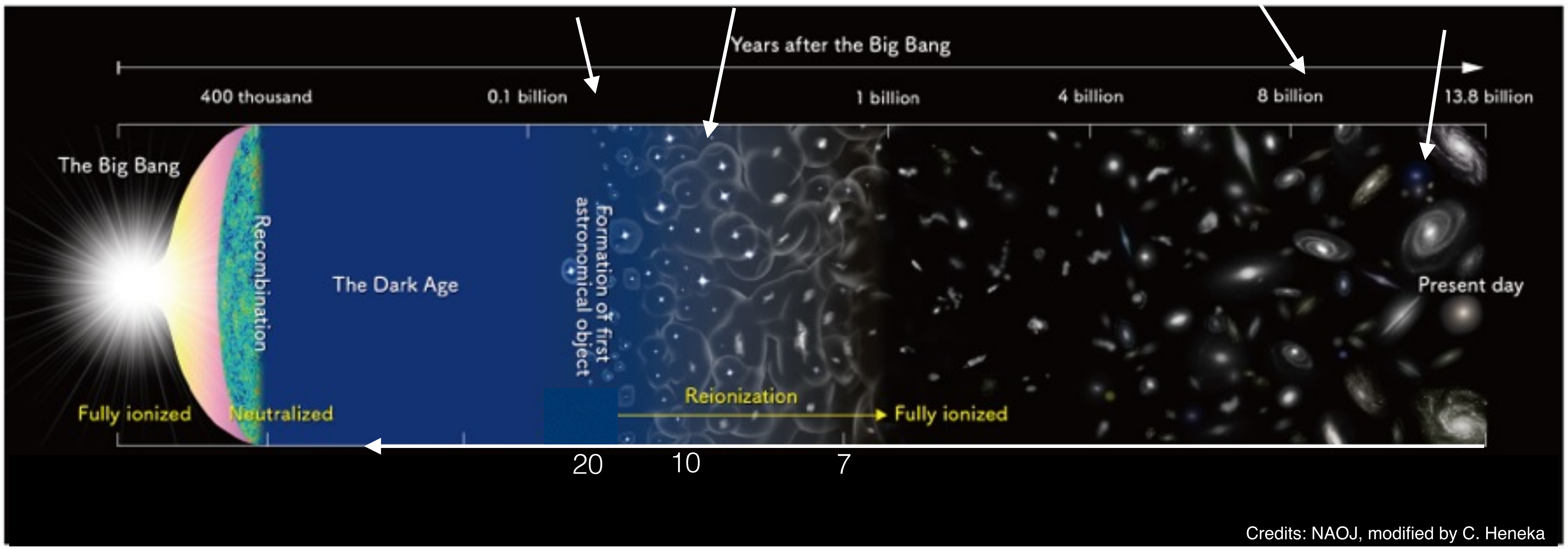
Where we stand: New frontiers in astronomy and astrophysics

<u>cosmic dawn **CD**</u>
'first stars and galaxies'

<u>reionisation **EoR**</u>
'the last big phase transition'

<u>peak of star formation</u>

<u>'late-time structure formation'</u>



We are here

Where we stand: New frontiers in astronomy and astrophysics

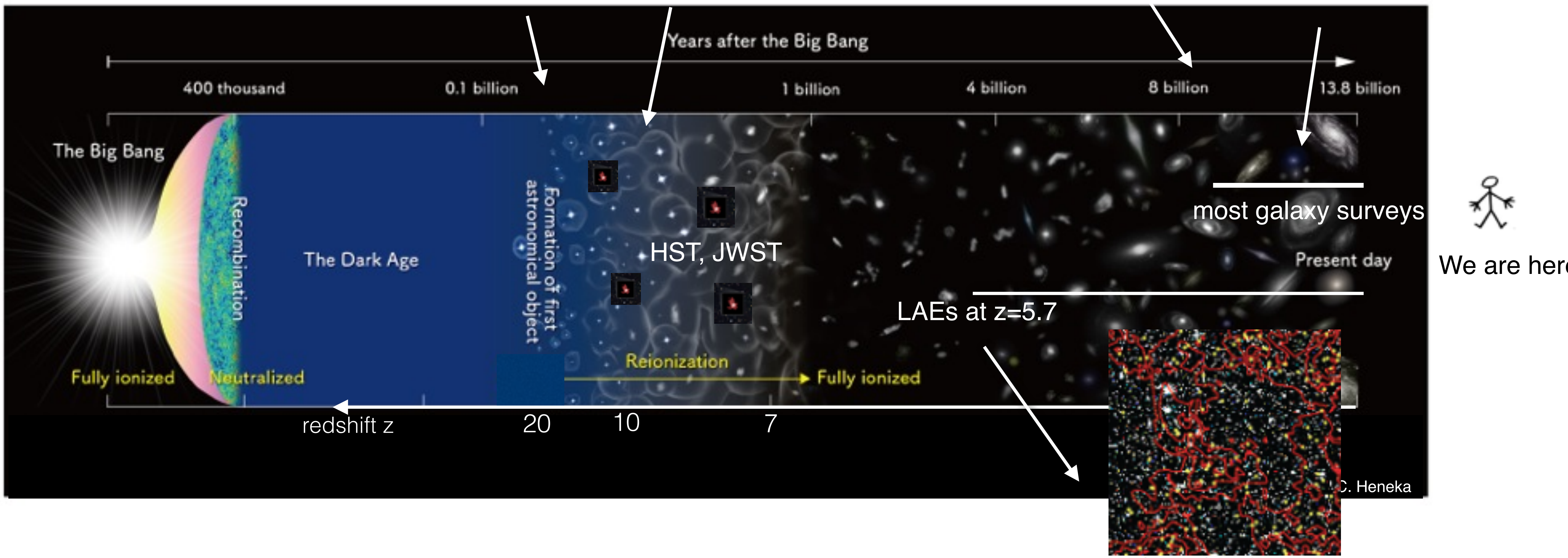
Galaxy surveys as an astronomical probe

cosmic dawn **CD**
‘first stars and galaxies’

reionisation **EoR**
‘the last big phase transition’

peak of
star formation

‘late-time
structure formation’



Ouchi, Ono, Shibuya ‘20

large-scale structure formation — map
80% of the observable Universe

Test astrophysics of galaxy evolution, the intergalactic medium (IGM) as well as cosmological structure formation from the Epoch of Reionisation (EoR) to the present.

…80% of observable Universe

Modelling challenges

True LSS probes $\longrightarrow$ 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

$\longrightarrow$ about 3-4 orders of magnitude

$\longrightarrow$ about 10^9-10^{10} modes!

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

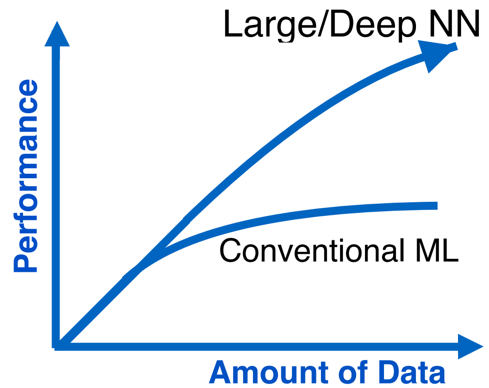
Why ML/DL for astrophysics

Driven by large-scale + high-resolution surveys

SPHEREx
Credit: SKAO Credit: ESA
Credit: NASA, ESA, CSA
Credit: ESO

+ multi-messenger

Deep Learning
Driven by ability to improve with large datasets



Examples:

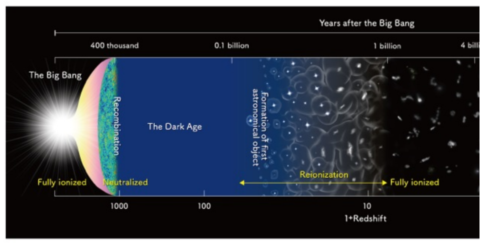
<3 sec: Classification 40.000 spectra
3 sec: 7-parameter regression O(100)MB cube
<4 sec: detection, segmentation & flux measurement on O(100) sources

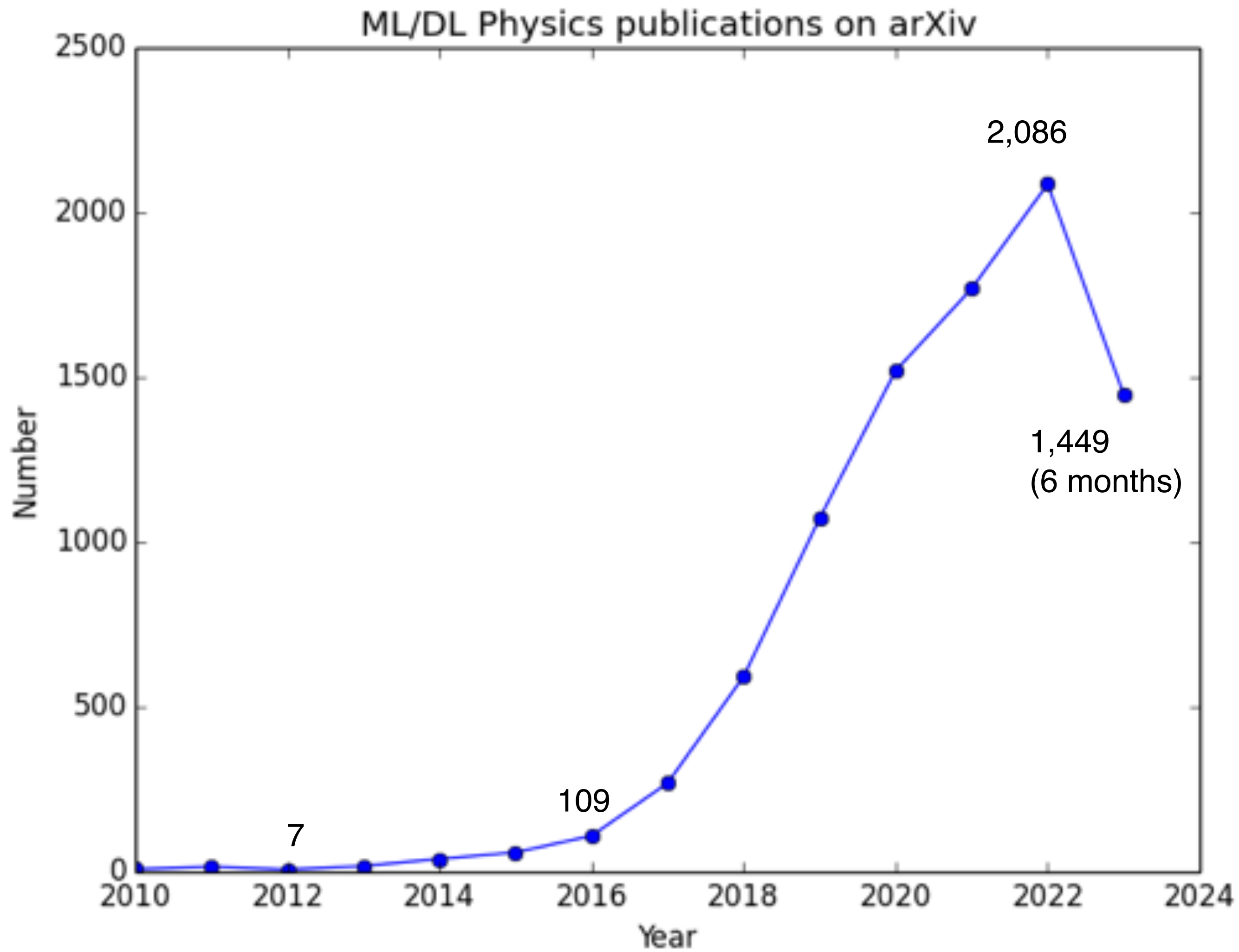
Extract more & less biased information
Data mining

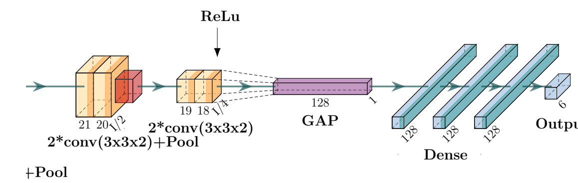
Efficient data reduction
Automation

Representation learning

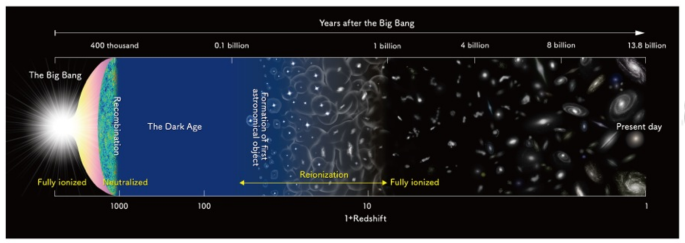
ML/DL for astrophysics

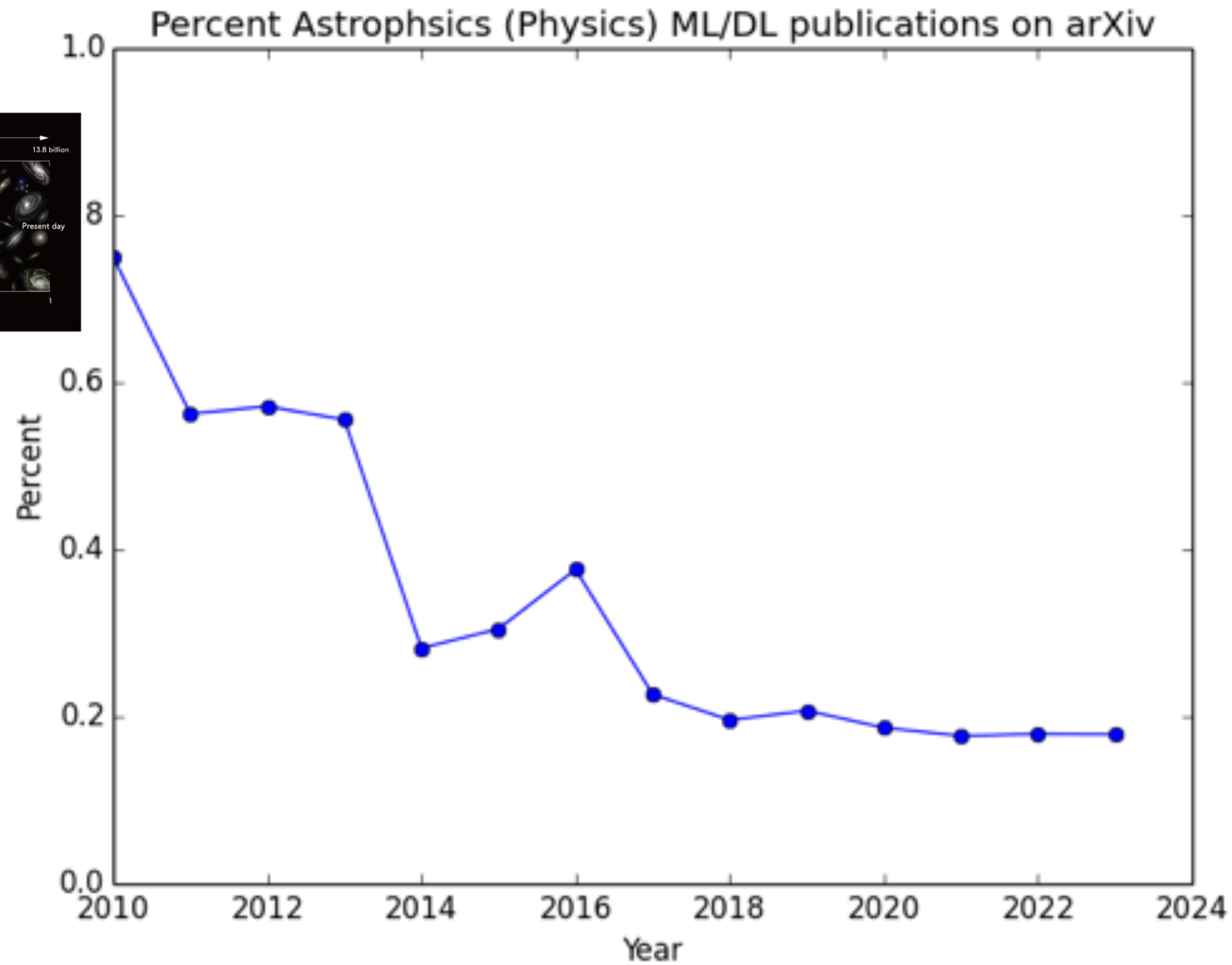


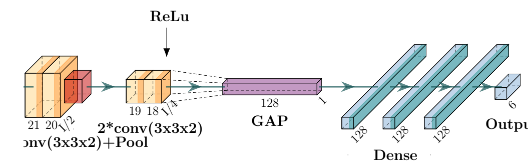




ML/DL for astrophysics





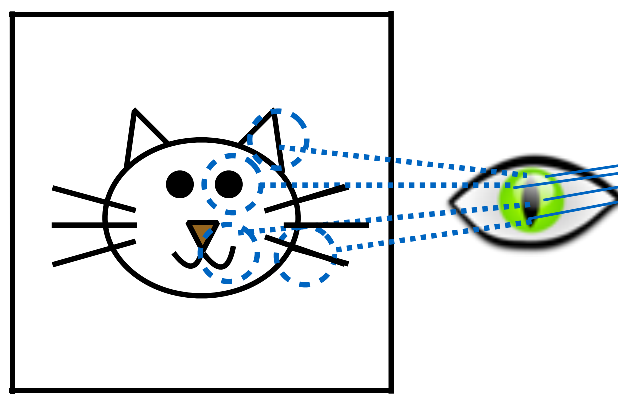


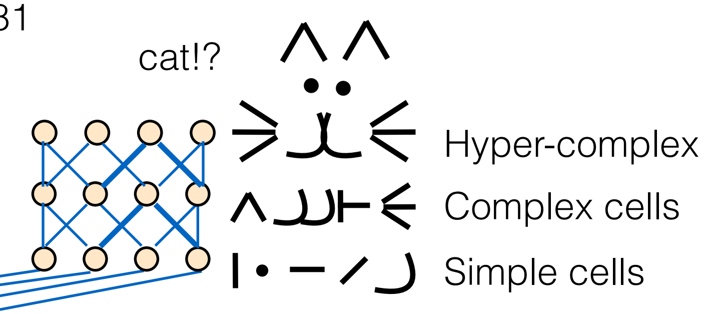
Why ML/DL

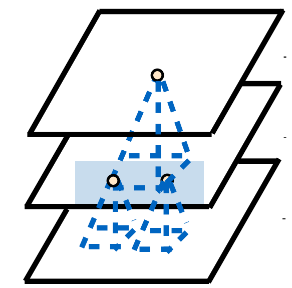
Neuron response to visual stimuli

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

Representation learning







Convolutional neural network

'Neocogniton' Kunihiko Fukushima 1979

Hierarchical learning

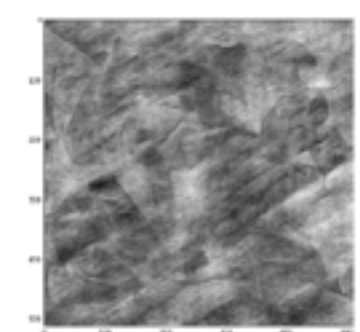
Non-linear, non-Gaussian

High-dim. correlations

duck

size
color
bkg

randomise phases →



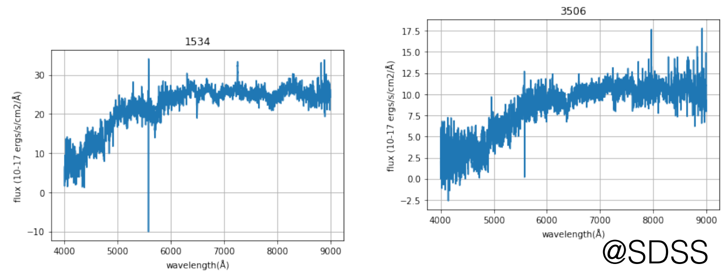
vs. "The famous Gaussian duck"

Same 2D power spectrum

Why ML/DL **for astrophysics**

+ plenty of inverse problems

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

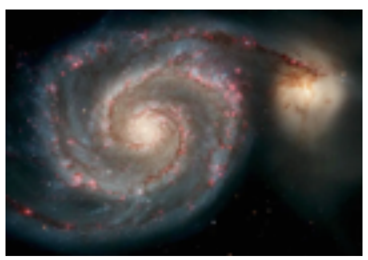


@SDSS

star vs. galaxy

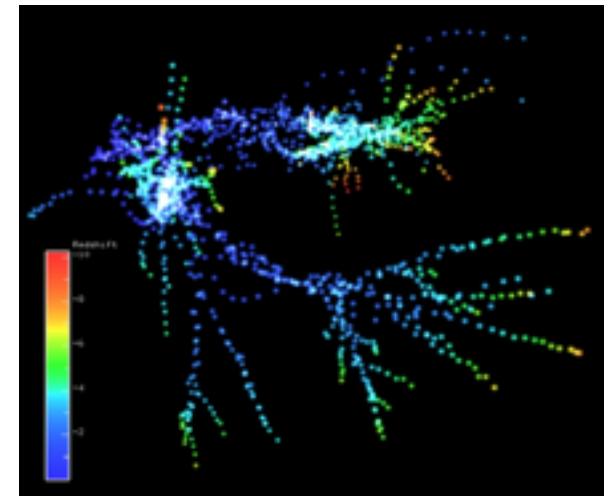
Representation learning

galaxy



@Hubble, NASA

@Hubble, NASA

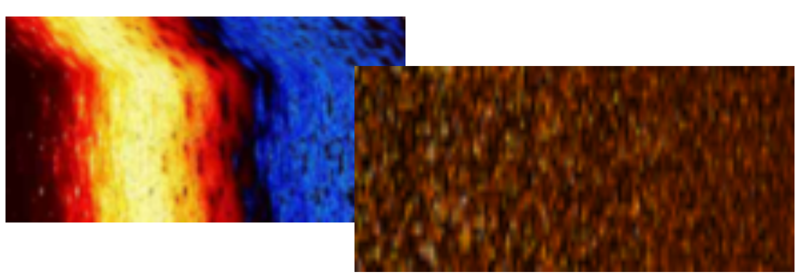


@Chris Fluke, Swinburne University of Technology

Hierarchical learning

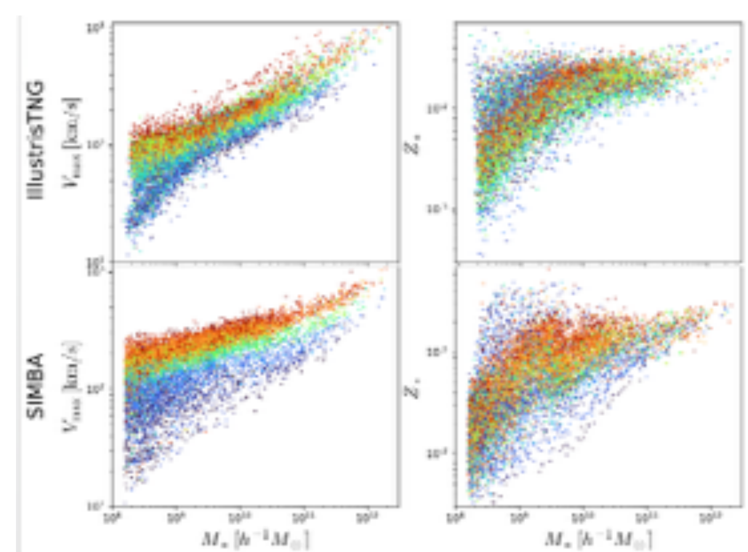
Non-linear, non-Gaussian

High-dim. correlations



fluctuation fields

@SKAO



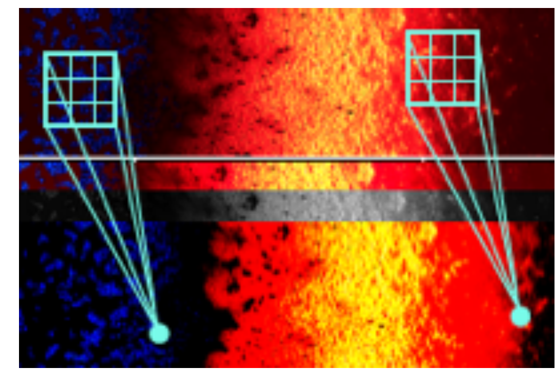
arXiv:2201.02202

ML/DL **for astrophysics**

scientific life cycles

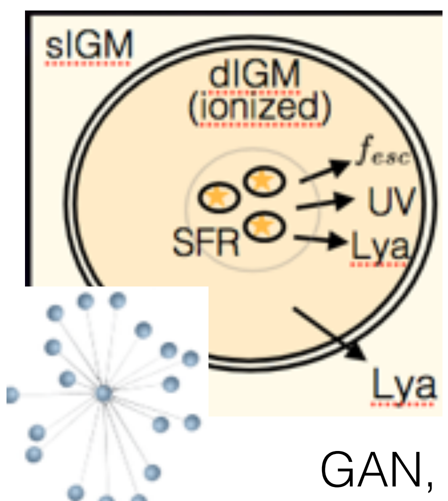
CNN, U-Net,
region-proposal, YOLO, ..

observation, signal extraction



Gillet+19
Prelogovic+22
Neutsch, Heneka, Brüggen 2022
arXiv:2201.07587

$(\Omega_{\rm m},\, \zeta,\, T_{\rm vir},\, L_{\rm X},\, E_0,\, m_{\rm WDM})$

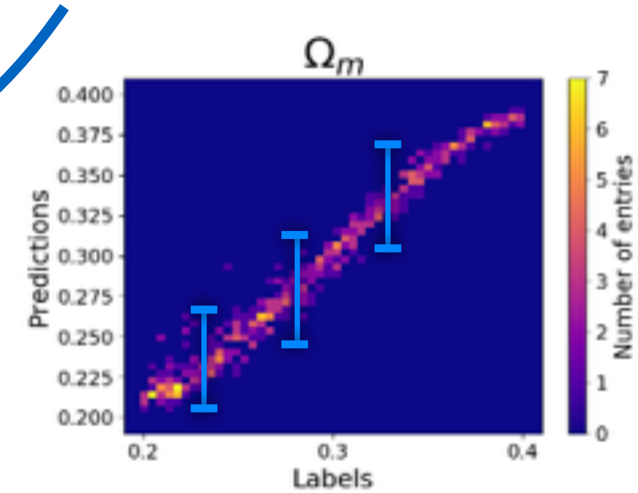


GAN, VAE,
GNN, NFs, ..

modelling, simulation

inference

BNN, MDN,
ensembles, NFs

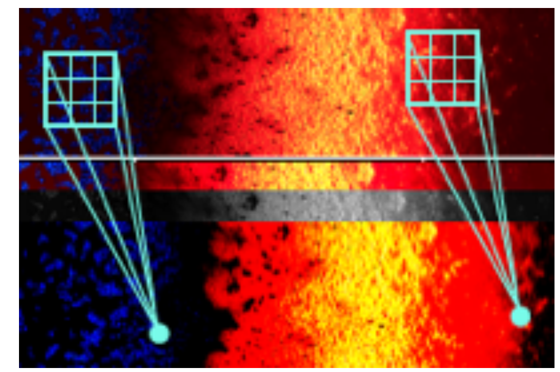


ML/DL **for astrophysics**

scientific life cycles

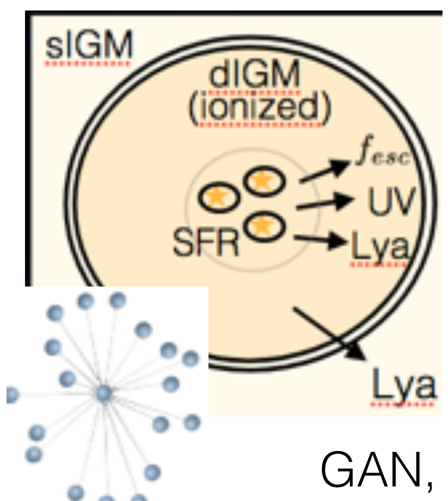
CNN, U-Net,
region-proposal, YOLO, ..

observation, signal extraction



Gillet+19
Prelogovic+22
Neutsch, Heneka, Brüggen 2022
arXiv:2201.07587

$(\Omega_{\rm m},\, \zeta,\, T_{\rm vir},\, L_{\rm X},\, E_0,\, m_{\rm WDM})$

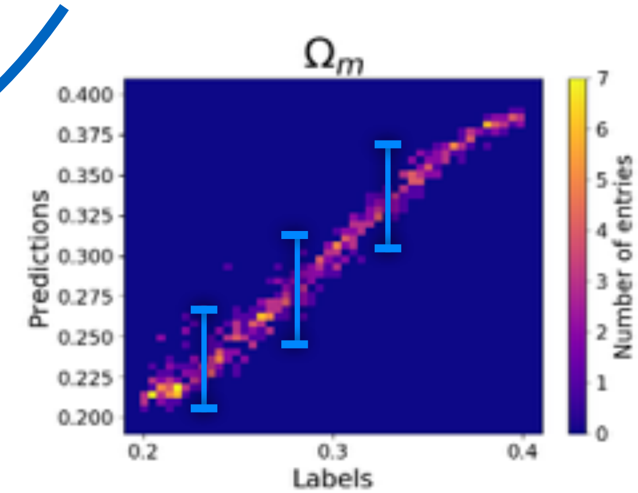


GAN, VAE,
GNN, NFs, ..

modelling, simulation

inference

BNN, MDN,
ensembles, NFs



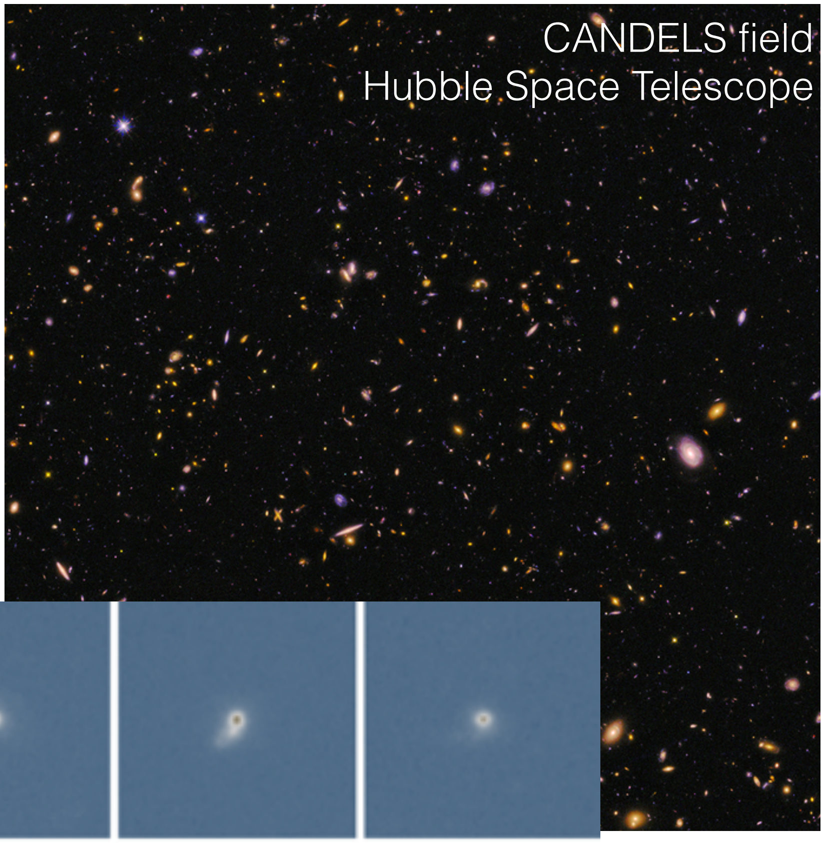
Example: Segmentation & regression, representations

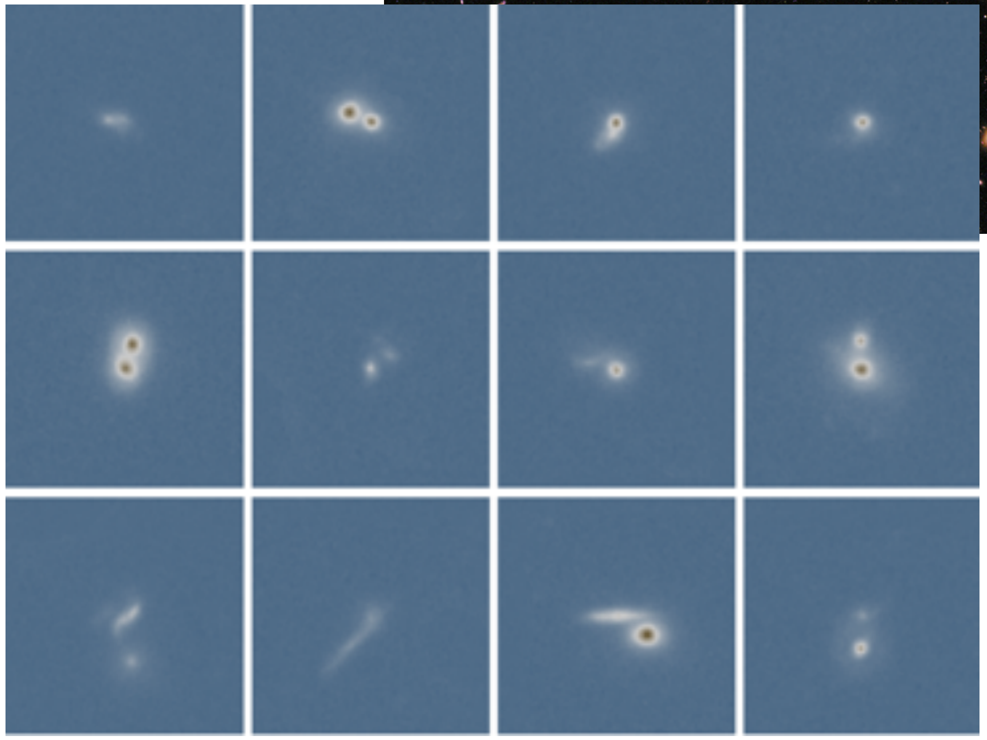
The deblending problem

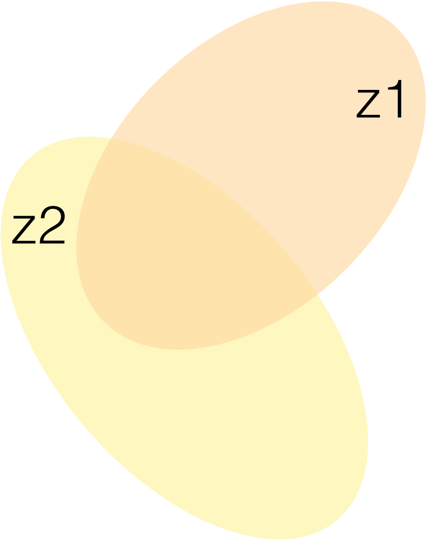
Goal: ‘Good’ photometry for surveys with high blended fraction (expected), e.g. SDSS, LSST and Euclid

Add-on: Galaxy segmentation and morphology / shape (also prior for ‘classic’ methods)

Challenge: Galaxies are ‘transparent’, separating flux in overlapping regions is difficult





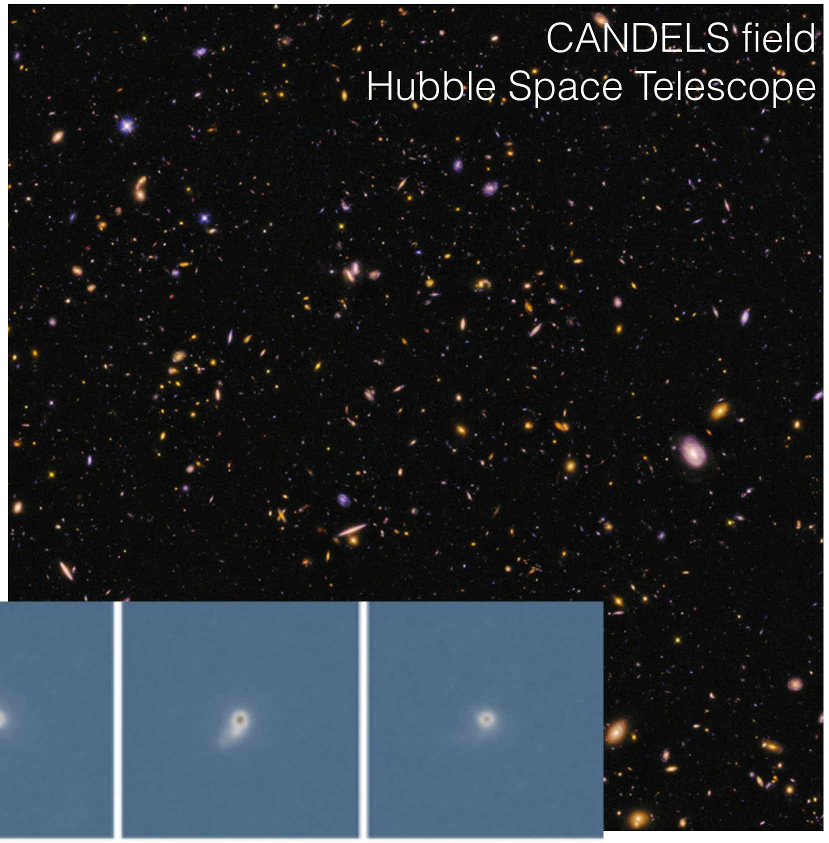


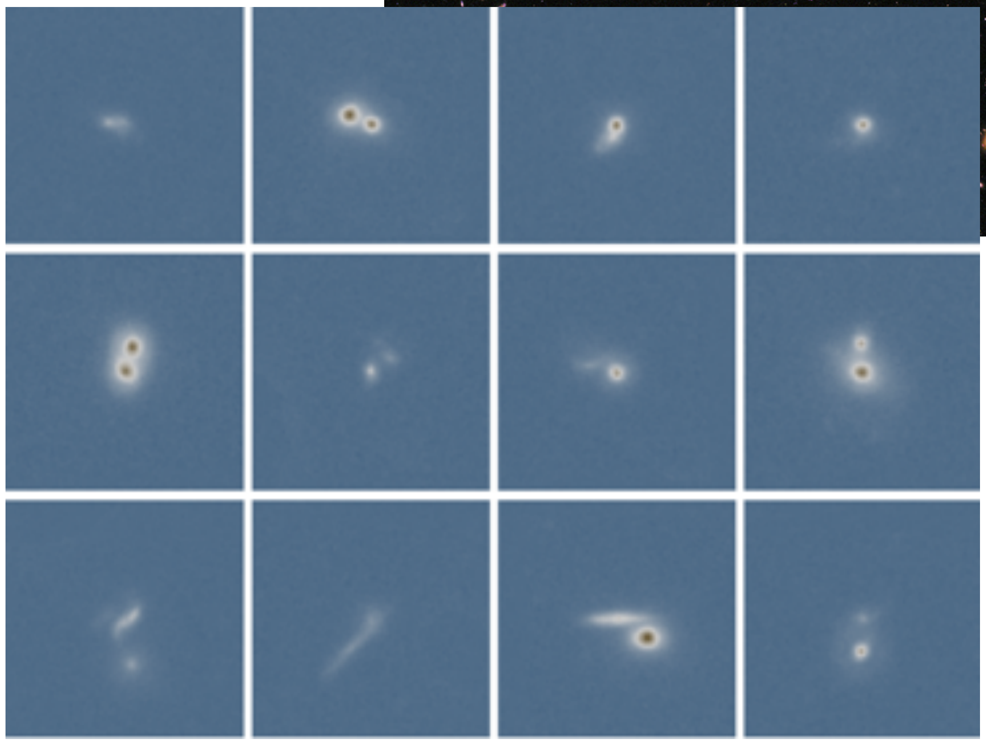
The deblending problem

Goal: ‘Good’ photometry for surveys with high blended fraction (expected), e.g. SDSS, LSST and Euclid

Add-on: Galaxy segmentation and morphology / shape (also prior for ‘classic’ methods)

Challenge: Galaxies are ‘transparent’, separating flux in overlapping regions is difficult

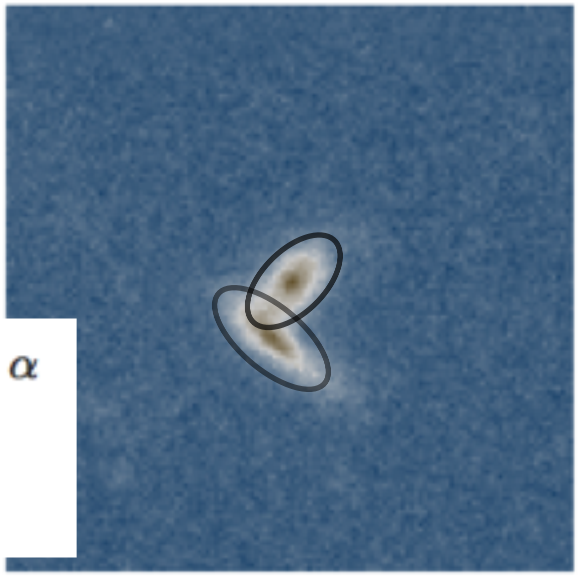




‘Classic’:
Fit ellipse(s)
and profile(s)

e.g. Einasto (‘65):

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

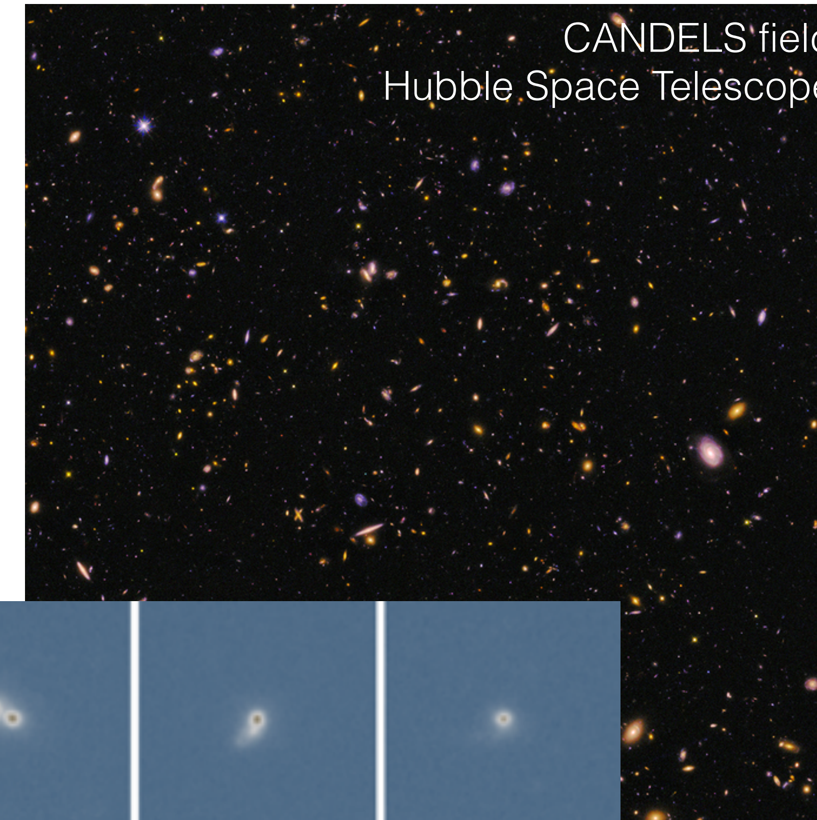


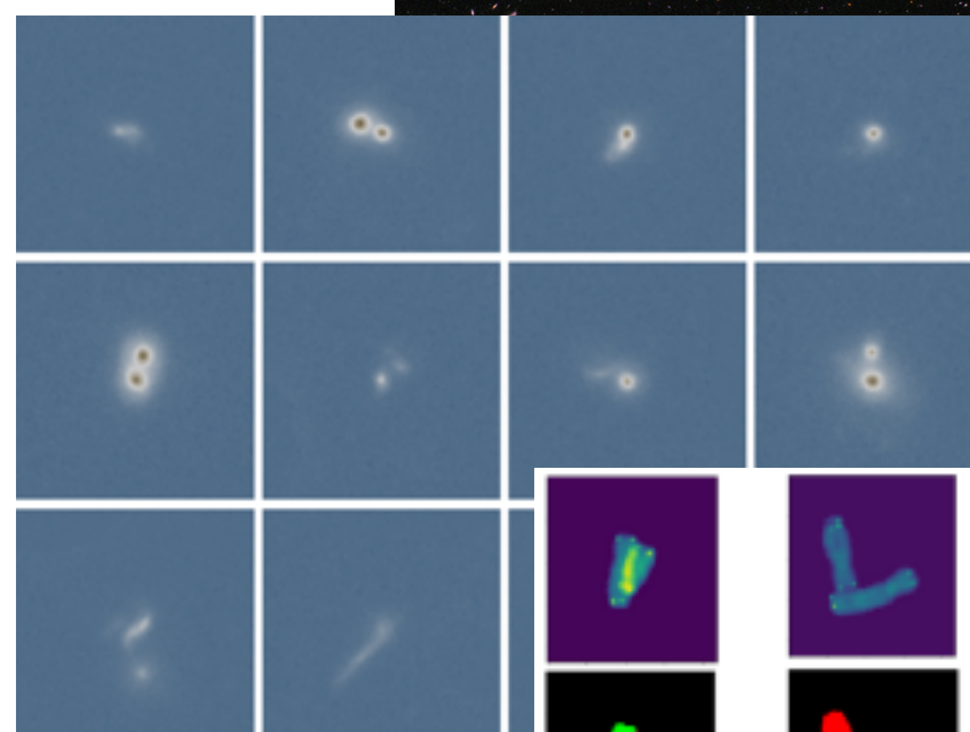
The deblending problem

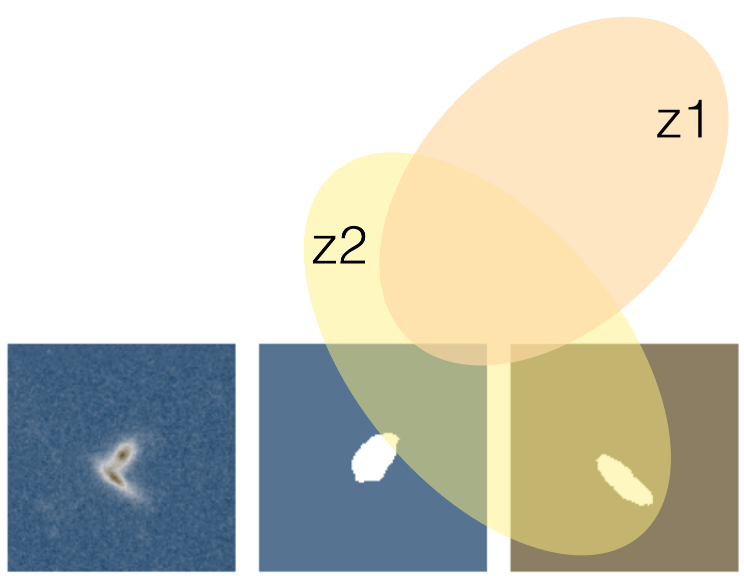
Goal: 'Good' photometry for surveys with high blended fraction (expected), e.g. SDSS, LSST and Euclid - avoid bias!

low stellar density (Ross et al. 2012a). The correlation of galaxy density with stellar density is the most significant known bias on measured clustering, likely caused by incomplete deblending of detected objects in crowded fields of the SDSS imaging data. On the other hand, no significant correlation is seen between number density and potential

Dawson et al. 2016







Similar challenge:
Overlapping chromosomes

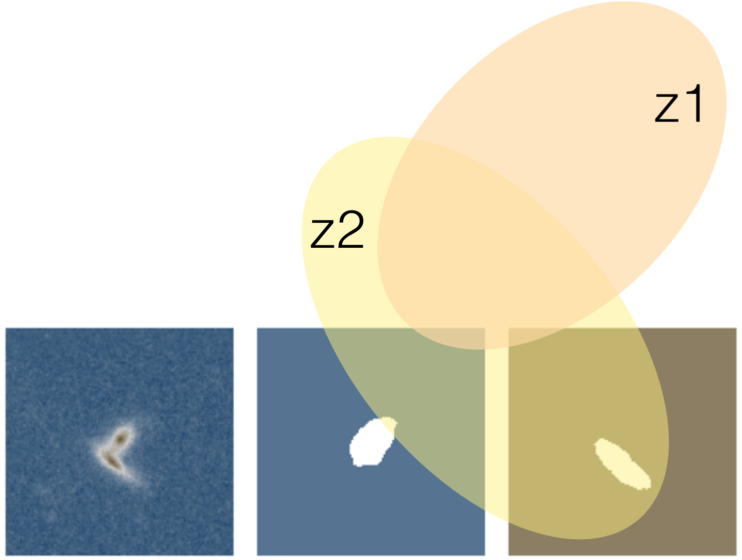
Lily Hu et al. 2017

The deblending problem

Goal: 'Good' photometry for surveys with high blended fraction (expected), e.g. SDSS, LSST and Euclid - avoid bias!

low stellar density (Ross et al. 2012a). The correlation of galaxy density with stellar density is the most significant known bias on measured clustering, likely caused by incomplete deblending of detected objects in crowded fields of the SDSS imaging data. On the other hand, no significant correlation is seen between number density and potential

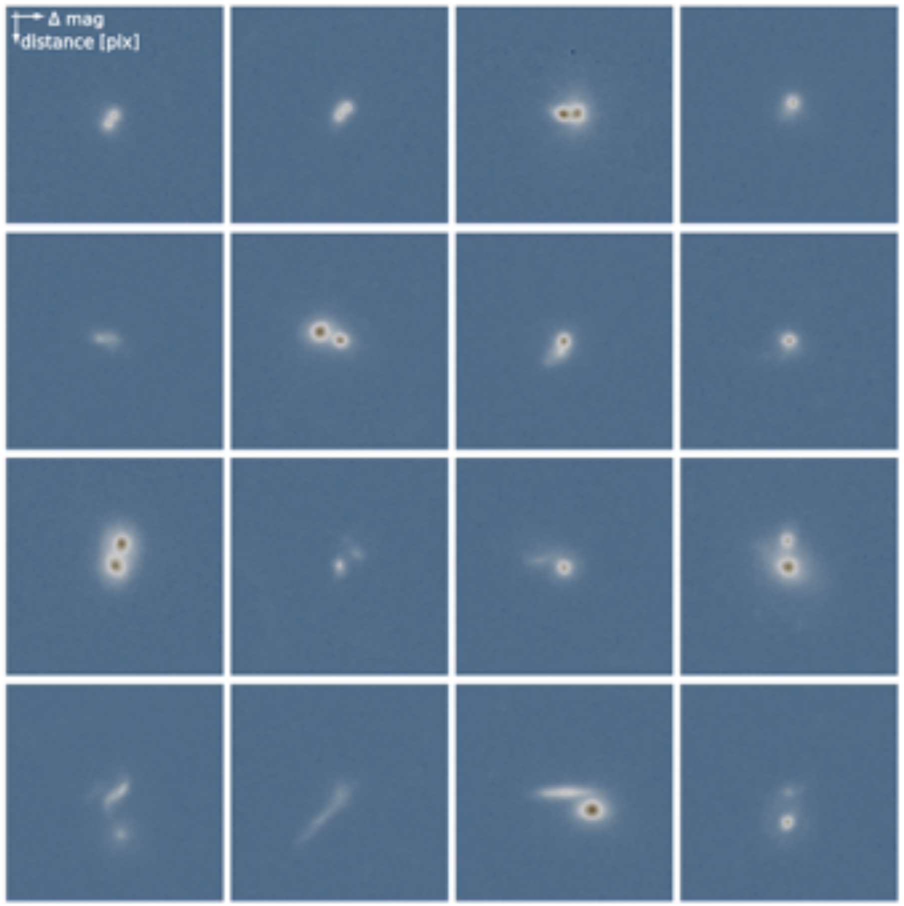
Dawson et al. 2016



Similar challenge:
Object detection

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

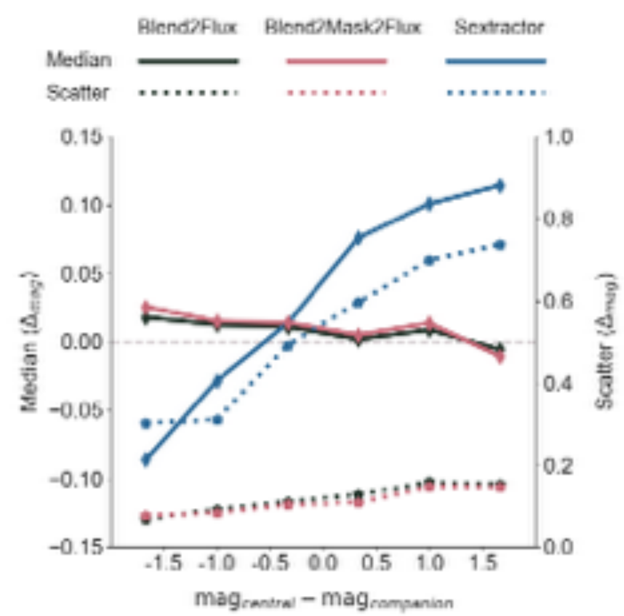
Application of Deep Neural Networks: Galaxy photometry and deblending, shape measurements



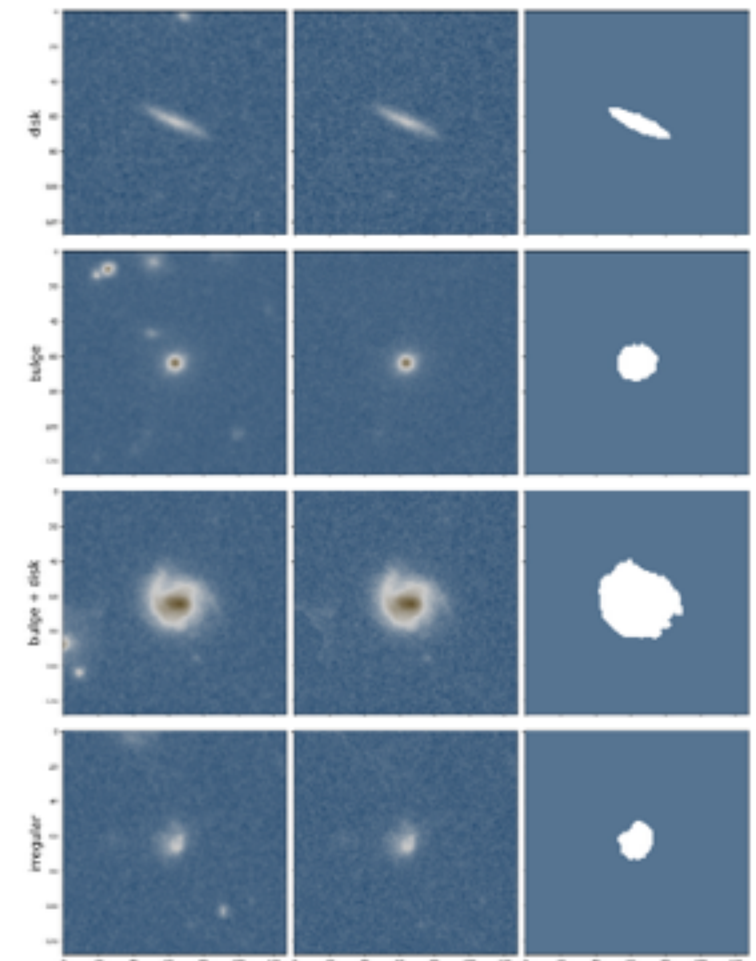
Artificially blended CANDELS data
18 < mag < 24
(https://github.com/aboucaud/candels-blender)

Get photometry of blended galaxies..

Goals for our deep NNs



..do so bias-free

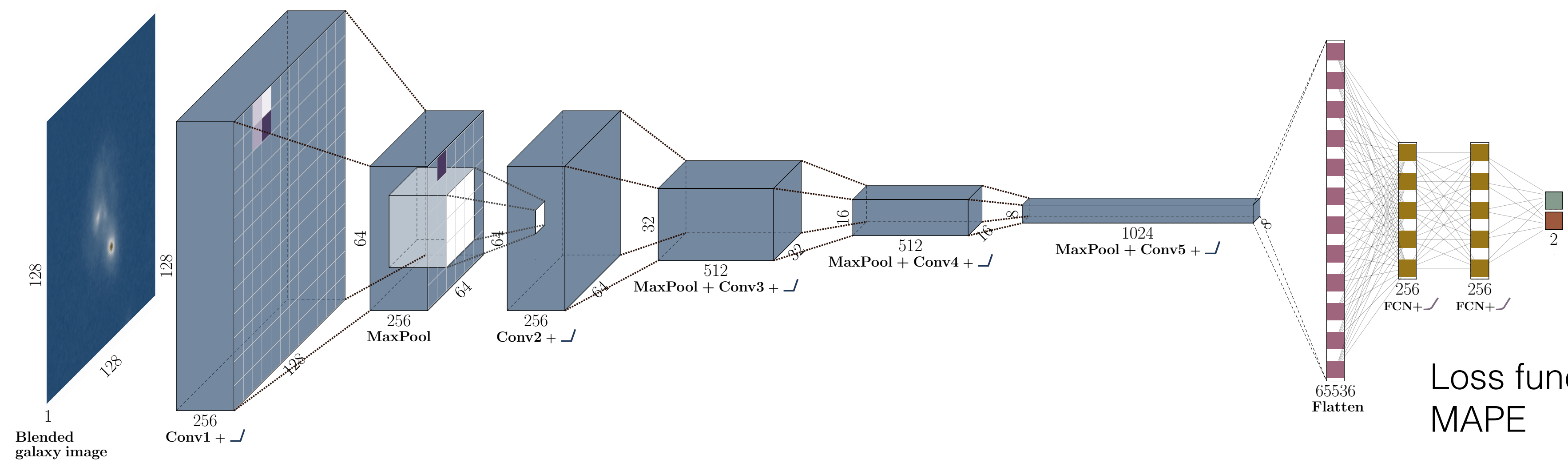


..Derive galaxy masks (shape measurements)

Boucaud et al.,
COIN collaboration
MNRAS 491,2 (2020)

Application of Deep Neural Networks: Galaxy photometry and deblending, shape measurements

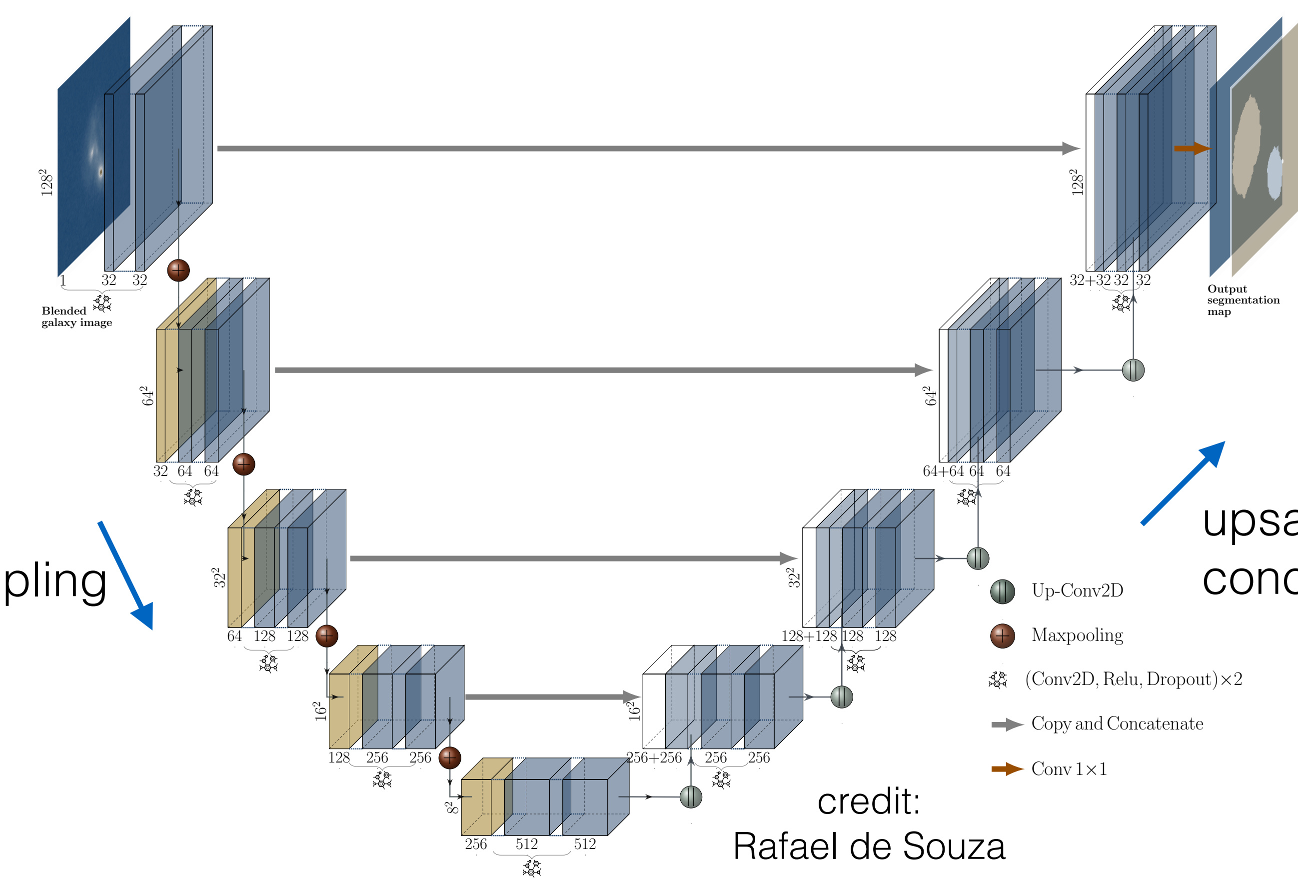
1) *blend2flux*
a CNN for photometry



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

github.com: coindeblend

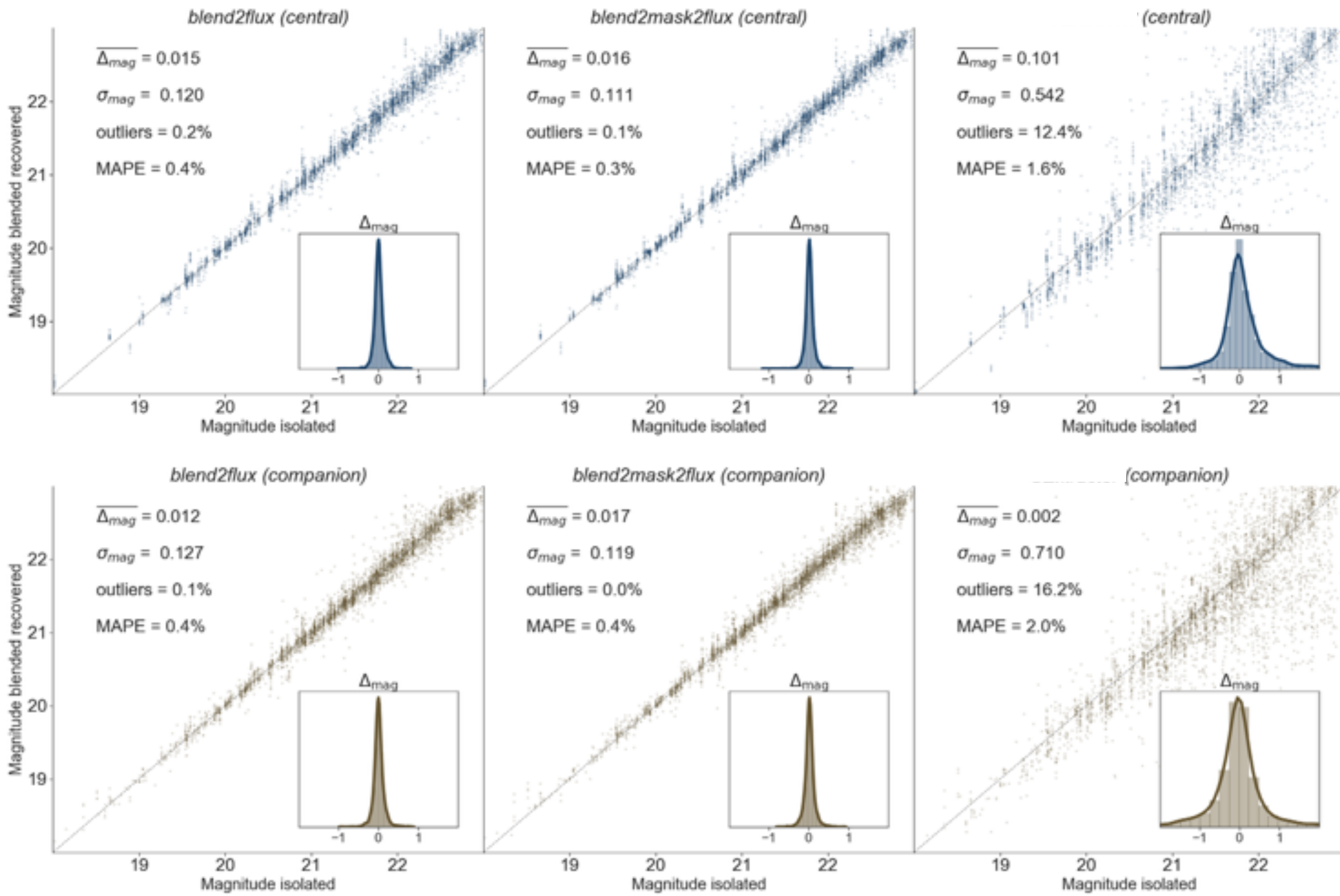
Boucaud et al.,
COIN collaboration
MNRAS 491,2 (2020)

Application of Deep Neural Networks: Galaxy photometry and deblending, shape measurements

Histograms of photometric errors

central

companion



→ small bias and scatter

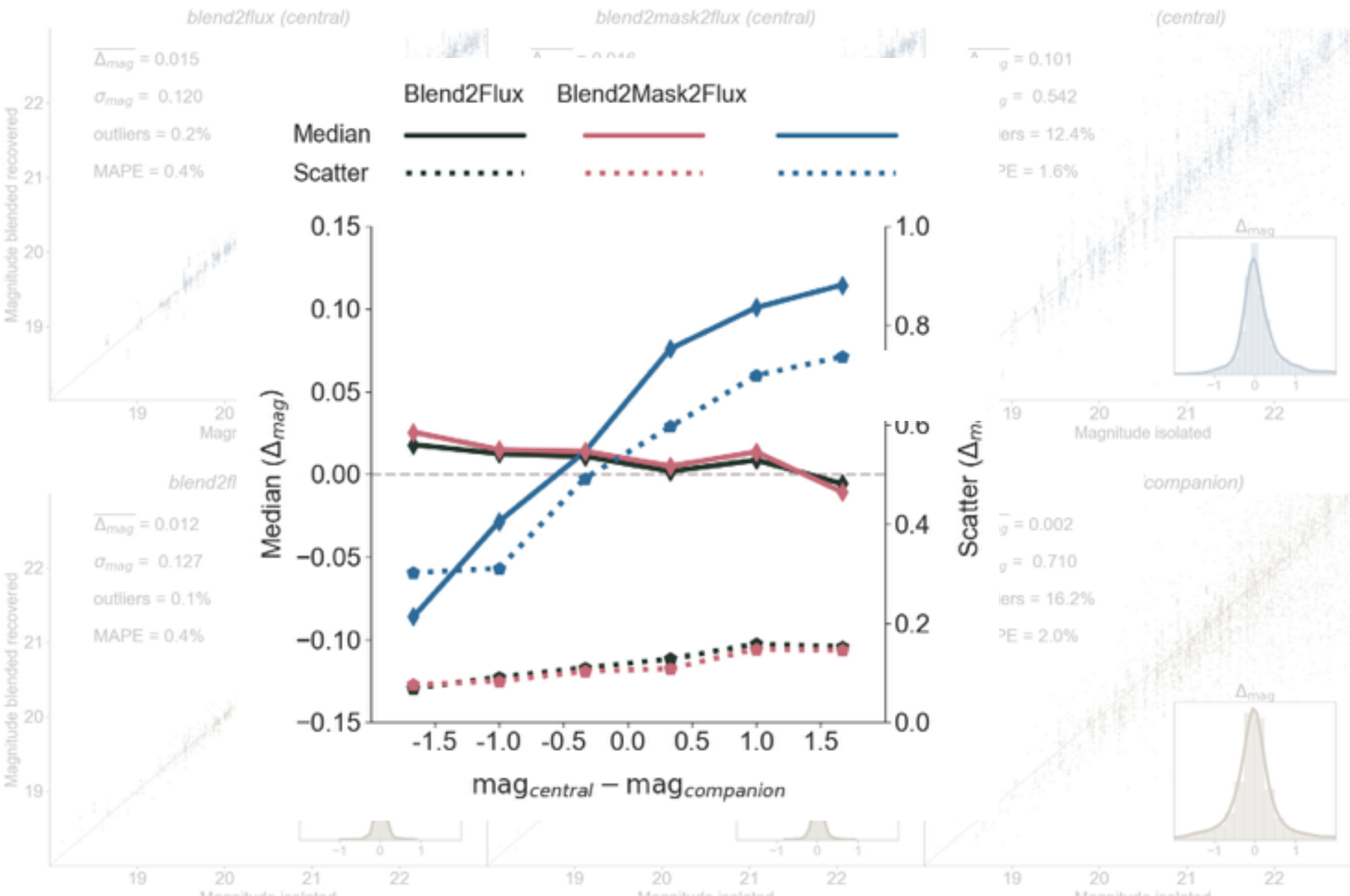
Boucaud et al.,
COIN collaboration
MNRAS 491,2 (2020)

Application of Deep Neural Networks:
Galaxy photometry and deblending, shape measurements

Histograms of photometric errors

central

companion



→ small bias and scatter

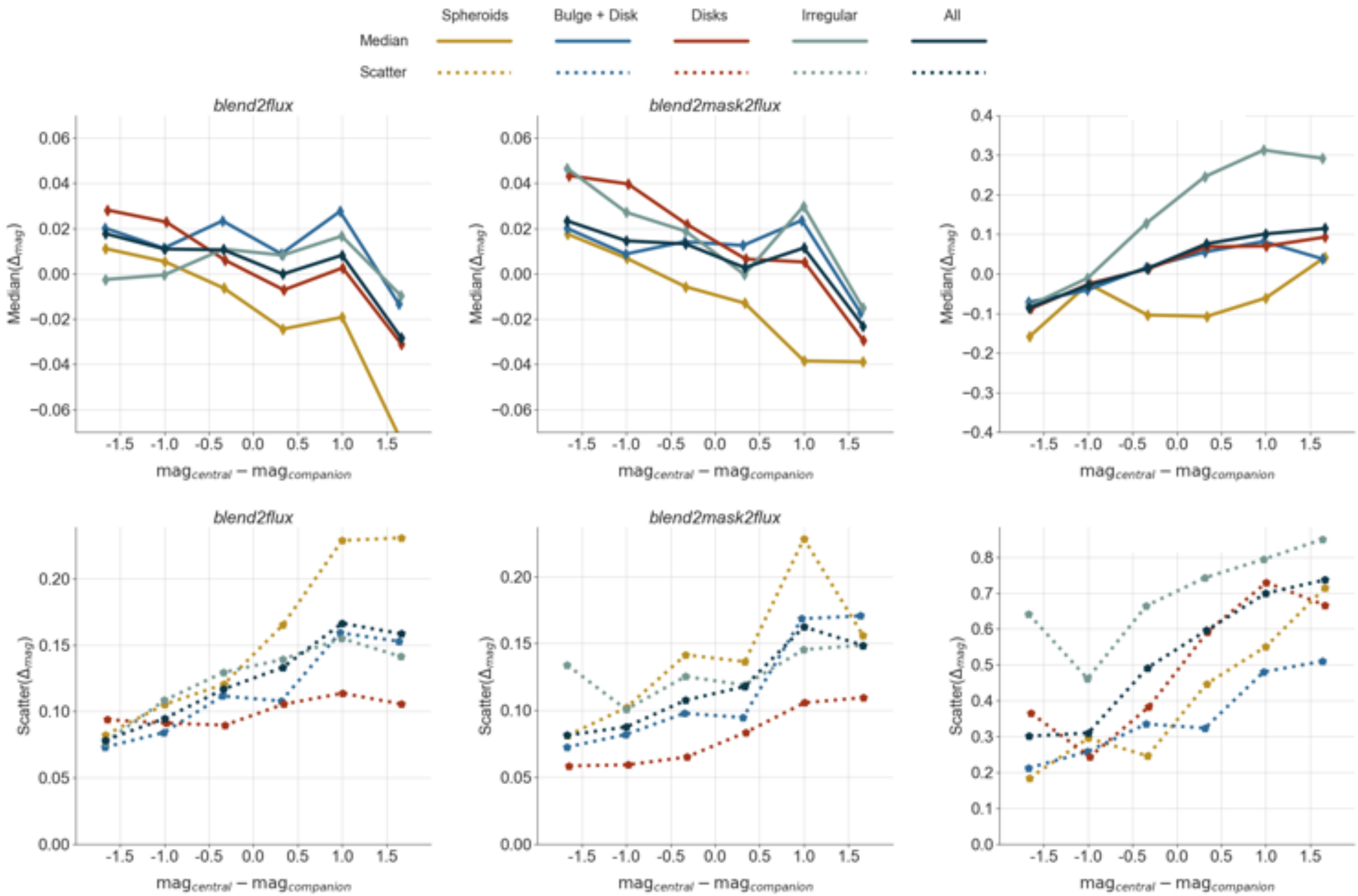
Boucaud et al.,
COIN collaboration
MNRAS 491,2 (2020)

Application of Deep Neural Networks:
Galaxy photometry and deblending, shape measurements

Photometric bias and scatter - galaxy type

bias

scatter

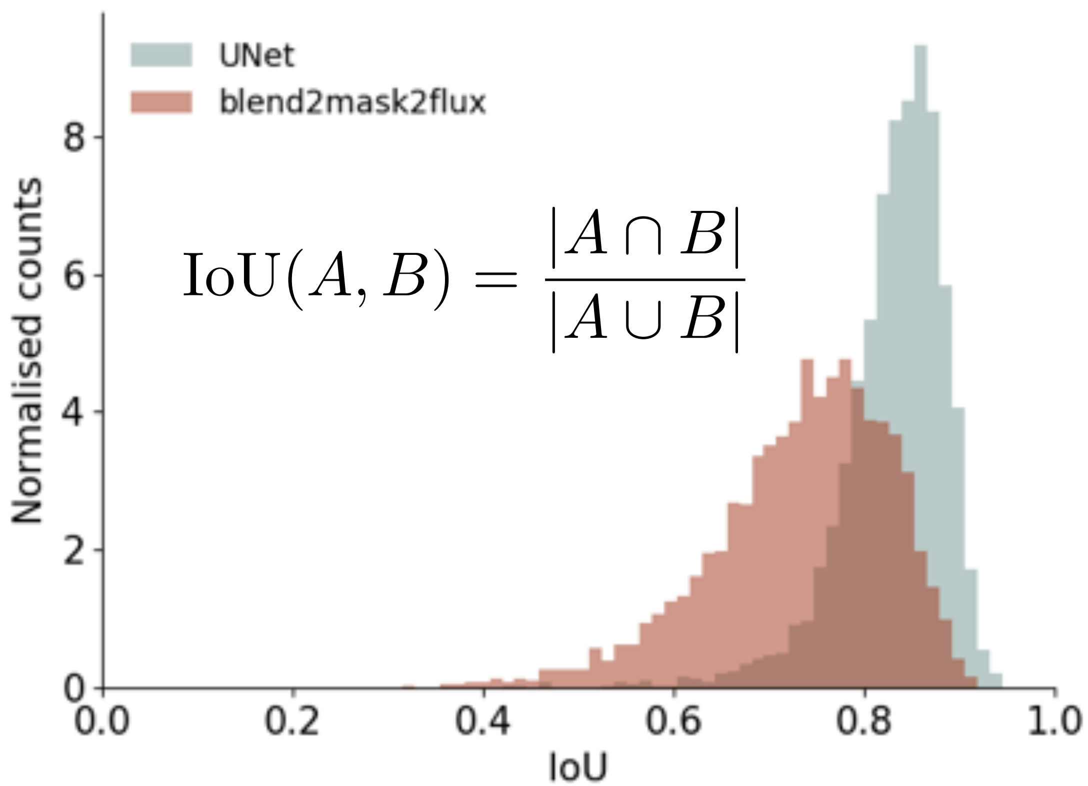


→ small bias and scatter

Boucaud et al.,
COIN collaboration
MNRAS 491,2 (2020)

Application of Deep Neural Networks:
Galaxy photometry and deblending, shape measurements

Histogram of IoU
(Intersection over Union - Jaccard index)

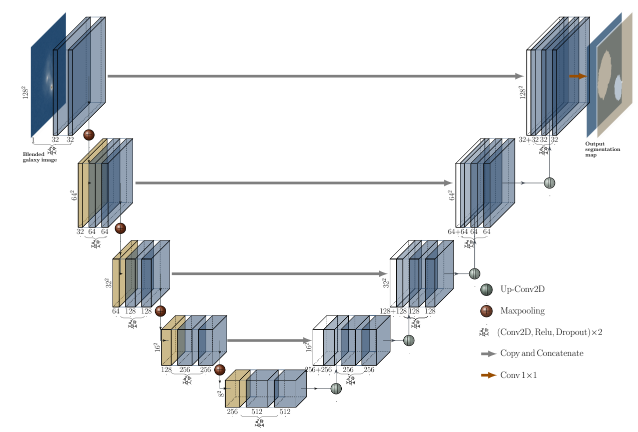


$$\mathrm{IoU}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

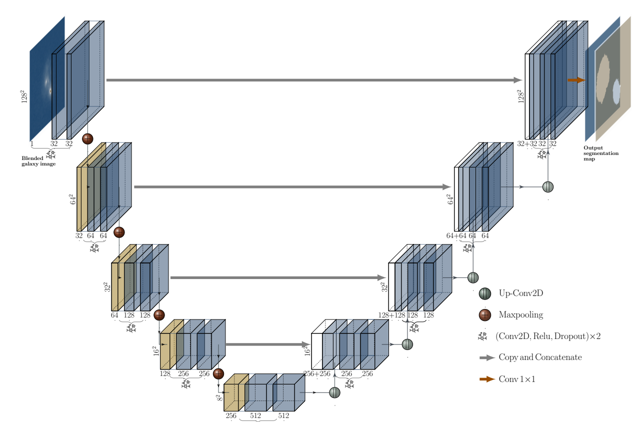
ground truth *U-Net* *blend2mask2flux*

→ Dispersion broadens when optimised for photometry

2 a) *U-net*
masks, no photometry



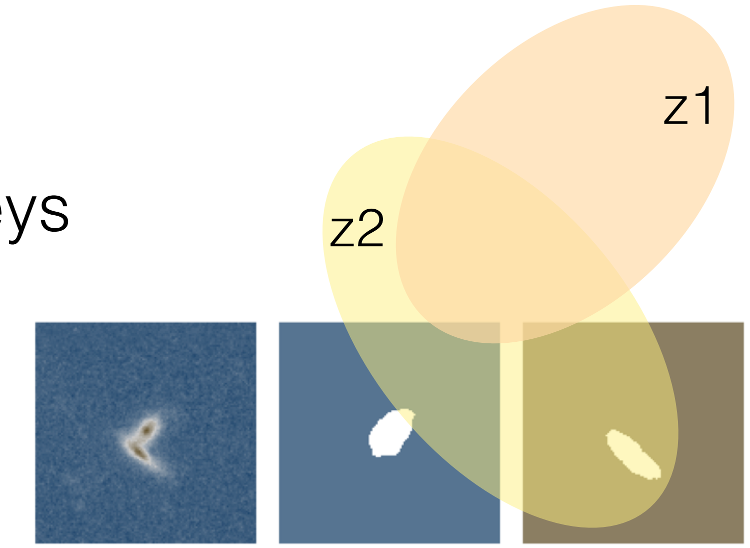
2 b) *blend2mask2flux*
photometry + masks



+ a CNN type *blend2flux*

Boucaud et al.,
COIN collaboration
MNRAS 491,2 (2020)

Galaxy deblending with nets: Take-aways



The deblending problem

- Large fraction of blends for deep photometric surveys
- Non-trivial to disentangle single galaxies
- Causes bias

Photometry

- Nets recover flux: low bias and high precision
- A ‘simple’ CNN *blend2flux* performs well
- Slight improvement when simultaneously constraining masks

Mask Segmentation

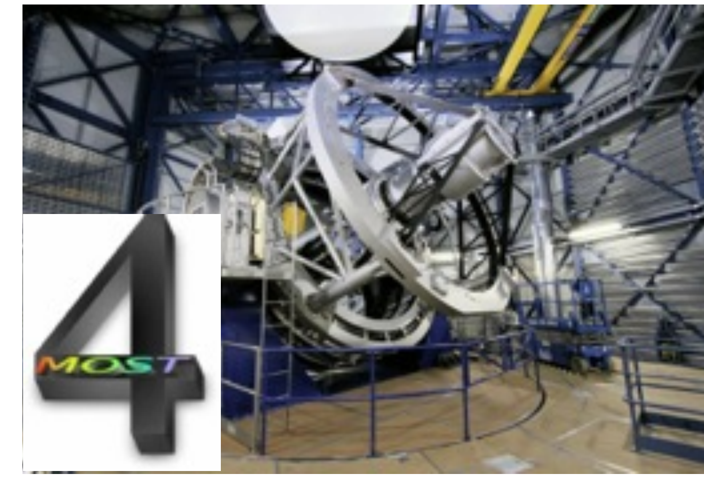
- *U-Net* architecture suitable to recover shapes
- Pitfall: Train photometry + shapes end-to-end

Needed for ‘precision-cosmology’

… prepared for Euclid satellite

Example: Classification

Classification in 1D — spectroscopy



Credit: ESO

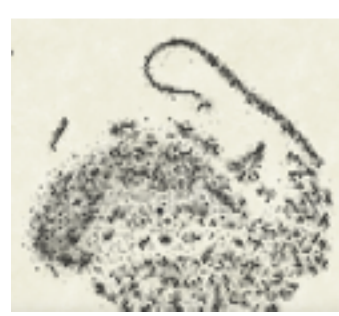
https://www.4most.eu

Building a classifier for 4MOST - Classification IWG9

- 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

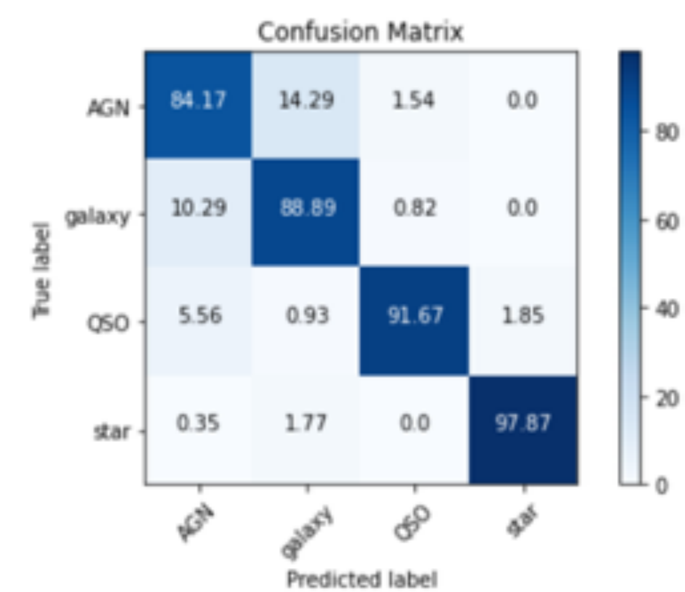
Goal: Data-driven classification layer between L1 and L2 pipelines

- **Basic target classification.** → Probabilistic multiclassifier
- **Galactic & extragalactic source classification.**
- **Feedback on a) targets, b) 'unknown' class**
  Currently set-up:
  4MOST explorer t-SNE
  (Gregor Traven, Gal Matijevic)
  arXiv: 1612.02242



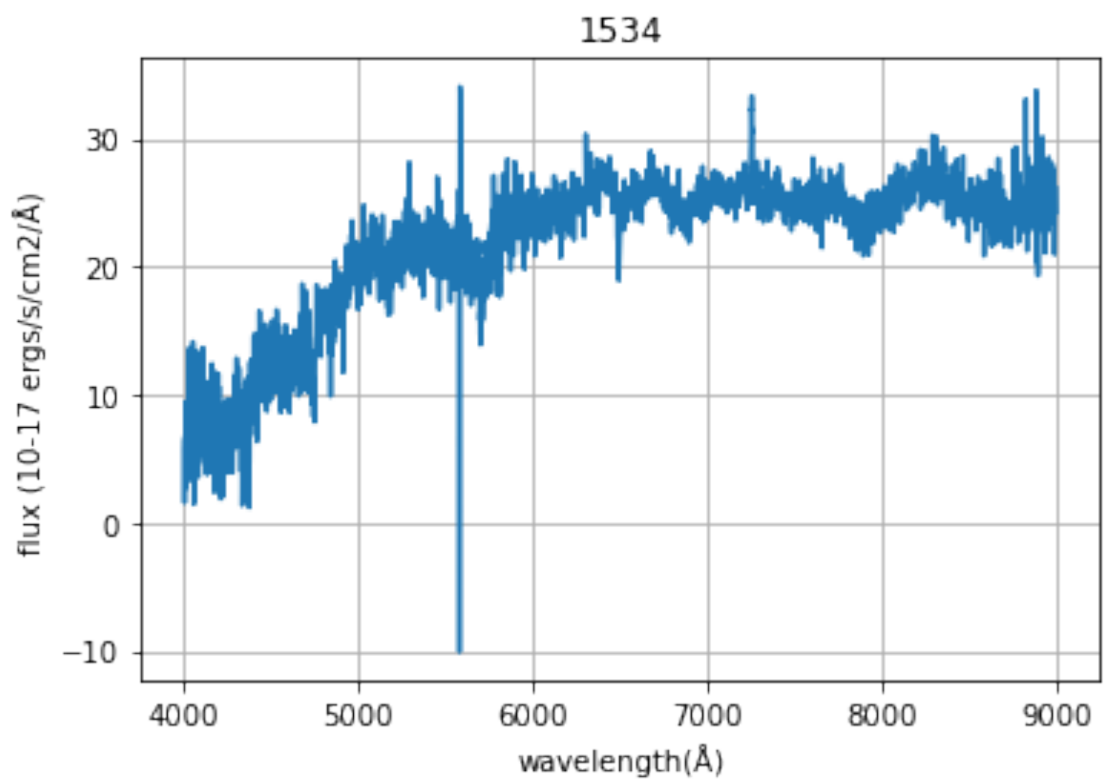
Benchmark tests with SDSS spectra

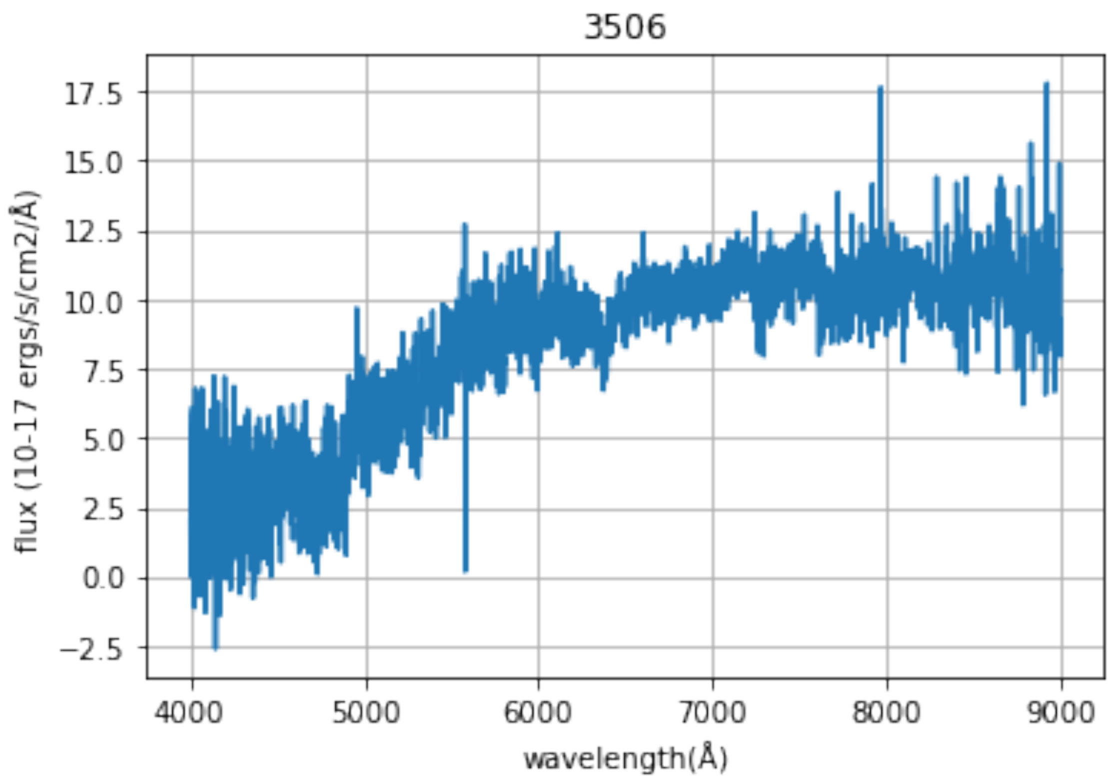
CNN variants
Random Forest
Support Vector Machine
Logistic Regression
Gaussian Naive Bayes



Classification in 1D — spectroscopy

Benchmark tests with SDSS spectra





Star or galaxy? Which type?

ML-for-Astro-tutorial

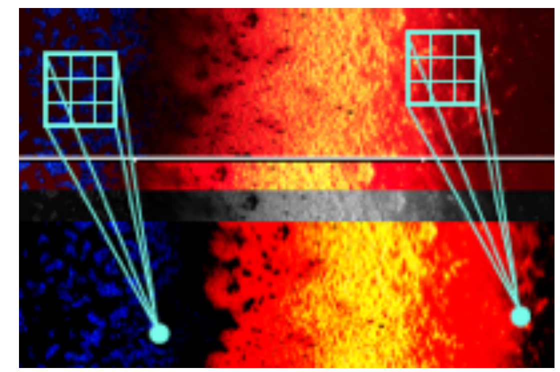
Computer Vision Astrophysics & Cosmology

Monday: Introduction, Deblending (regression, segmentation), classification

novel scientific life cycles

CNN, U-Net, region-proposal, YOLO, ..

observation, signal extraction

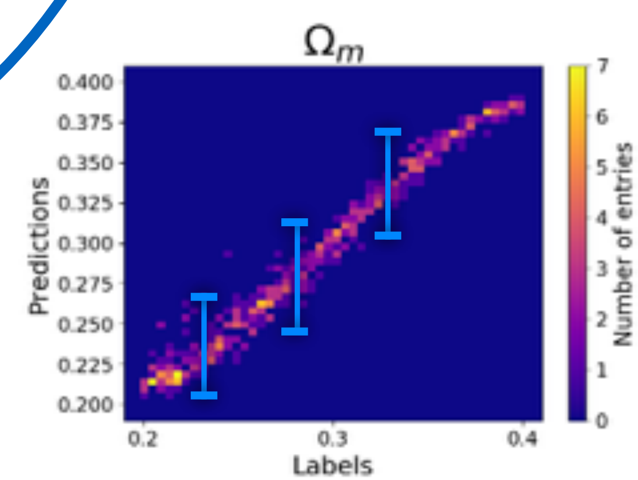


Gillet+19
Prelogovic+22
Neutsch, Heneka, Brüggen 2022
arXiv:2201.07587

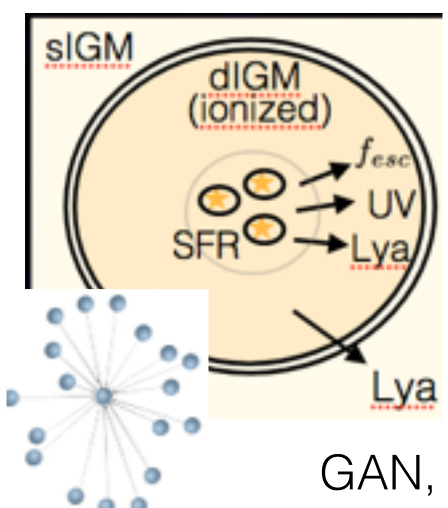
$(\Omega_{\rm m},\, \zeta,\, T_{\rm vir},\, L_{\rm X},\, E_0,\, m_{\rm WDM})$

inference

BNN, MDN, ensembles, NFs



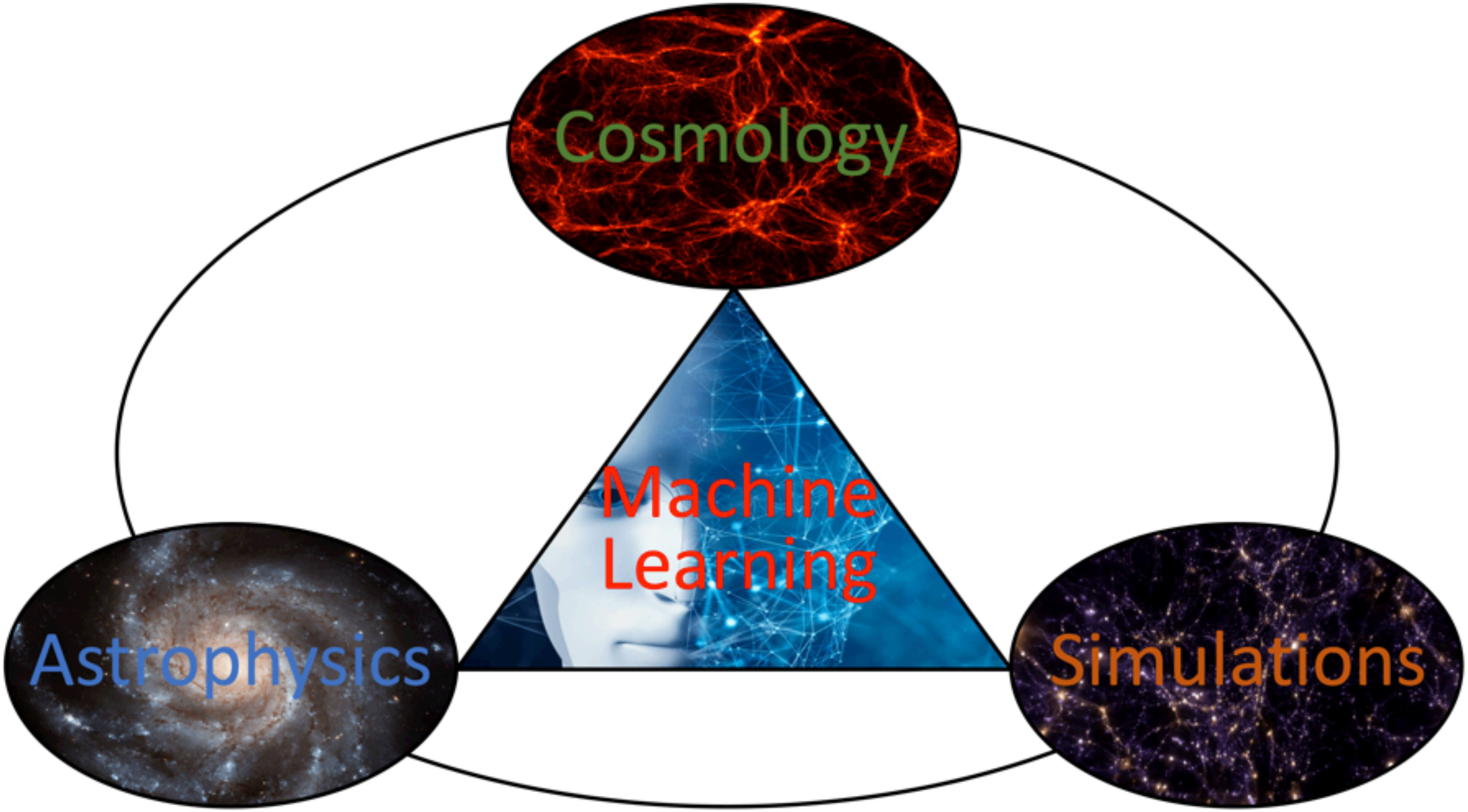
modelling, simulation



GAN, VAE, GNN, NFs, ..

Example: Inference, higher-order correlations

Our (ML) motivation



https://camels.readthedocs.io

Simulations: Camels

CAMELS

= Cosmology and Astrophysics with MachinE Learning Simulations

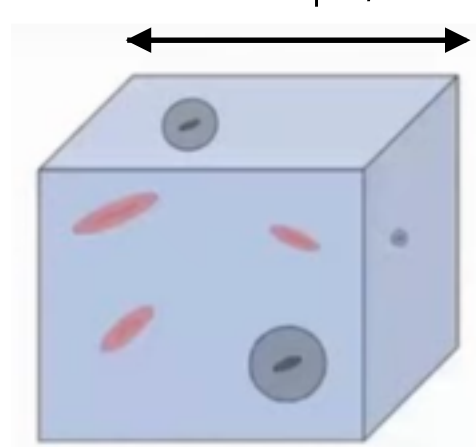
Type	Code	Subgrid model	Simulations
Hydrodynamic	Arepo	IllustrisTNG	1,092
Hydrodynamic	Gizmo	SIMBA	1,092
Hydrodynamic	MP-Gadget	Astrid	1,092
N-body	Gadget-III	–	3,049

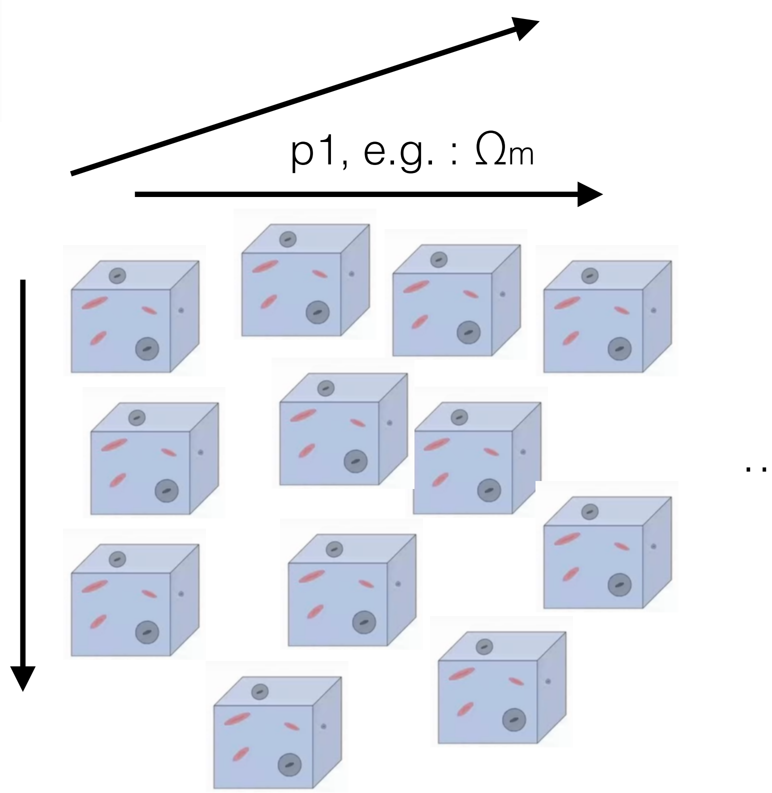
https://camels.readthedocs.io

Parameter set:

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

25cMpc/h

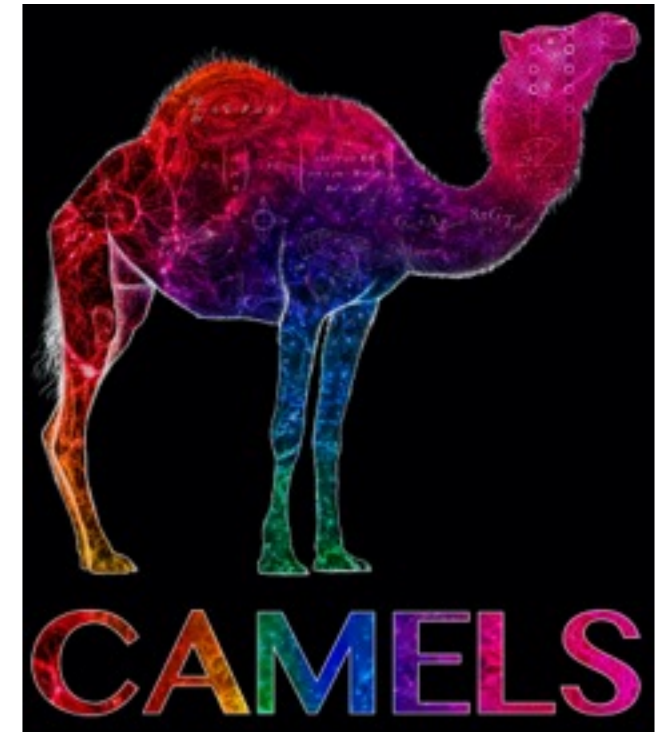




Simulations: Camels

CAMELS

= Cosmology and Astrophysics with MachinE Learning Simulations

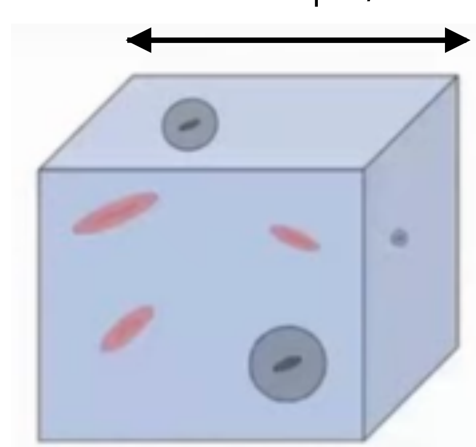


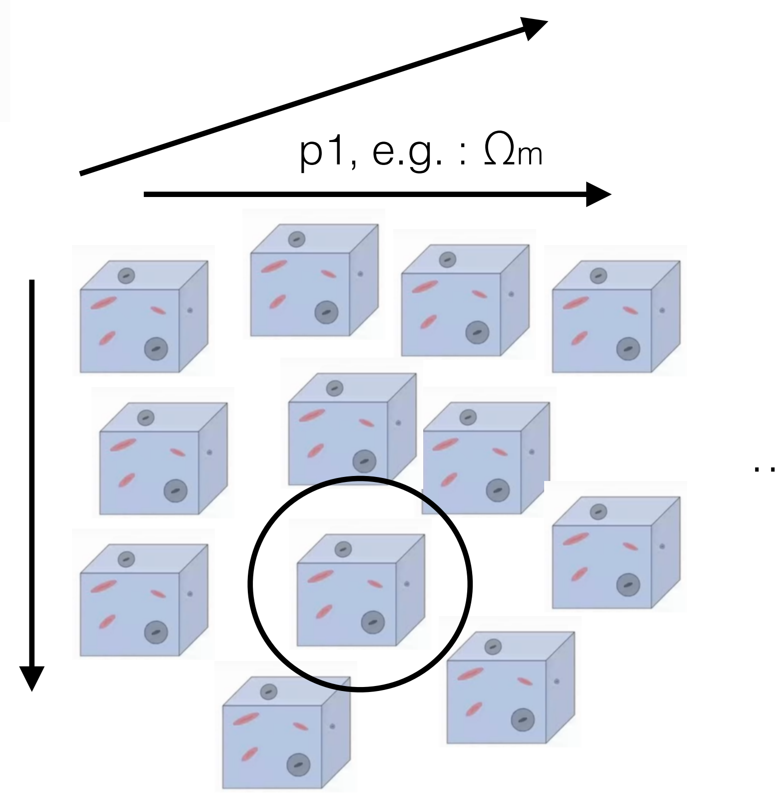
https://camels.readthedocs.io

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:

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





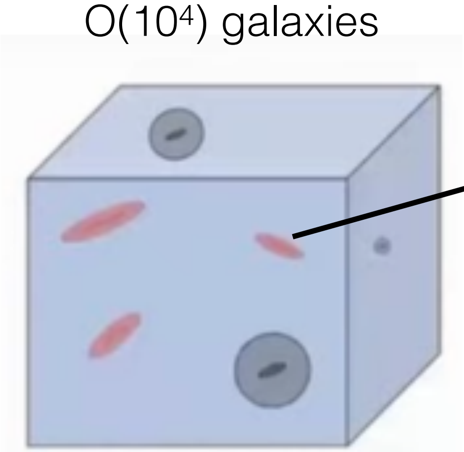
From one galaxy to global parameters

for each galaxy:

$$\left[\begin{array}{c} M_* \\ \mathrm{K} \\ M_g \\ Z_g \\ V_{max} \\ Z_* \\ \mathrm{g} \\ \sigma_v \\ R_* \\ M_t \\ \mathrm{U} \\ R_t \\ R_{max} \\ \mathrm{SFR} \\ J \\ \mathrm{V} \\ M_{bh} \end{array}\right.$$

(SIMBA)

O(10^4) galaxies



$\Omega_m = 0.32$
$\sigma_8 = 0.79$
$A_{SN1} = 1.2$
$A_{SN2} = 0.8$
$A_{AGN1} = 1.5$
$A_{AGN2} = 0.7$

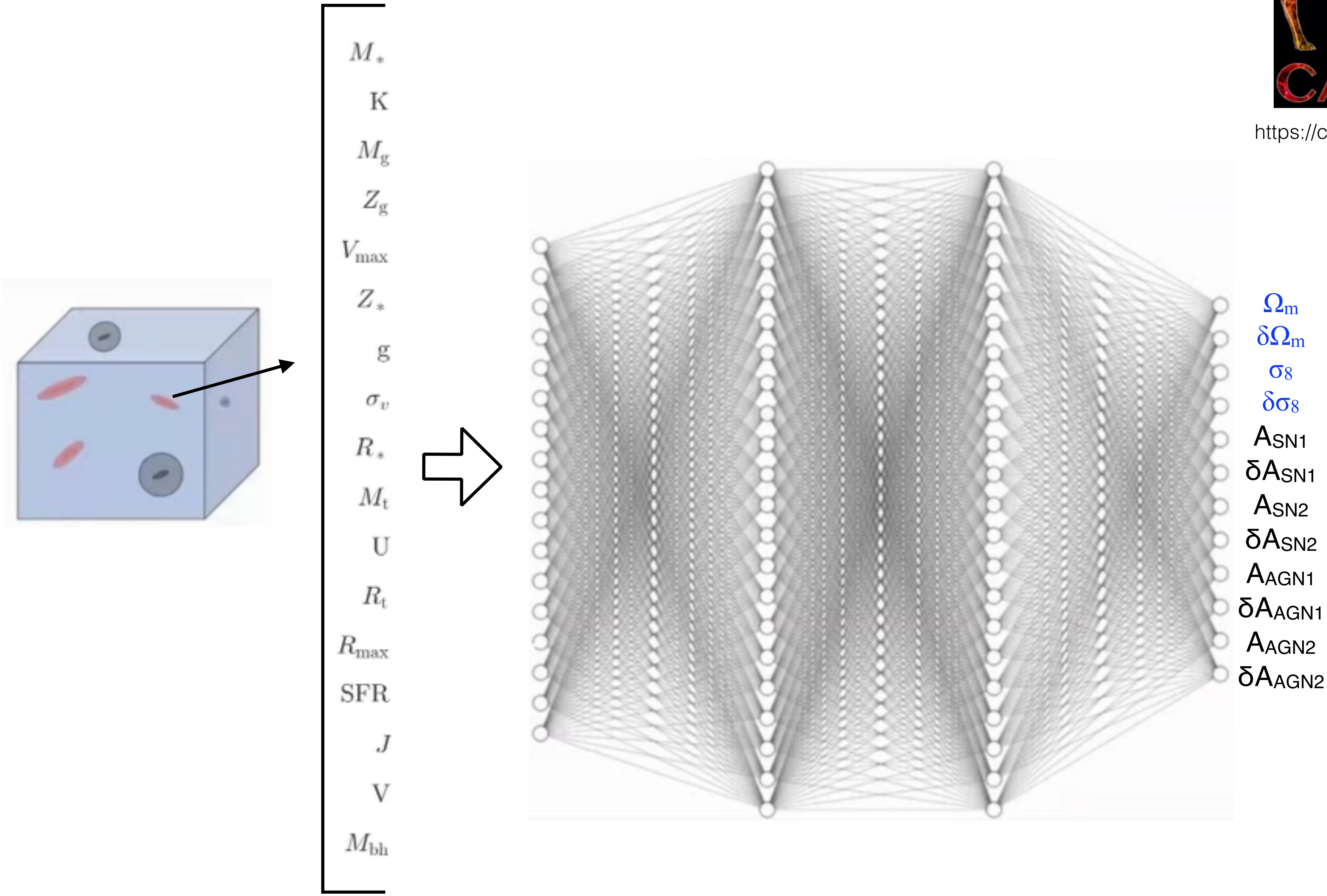
https://camels.readthedocs.io

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

Let's start with one!

From one galaxy to global parameters

https://camels.readthedocs.io



(SIMBA)

This model: Likelihood-free marginal inference

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:

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

Our model F(x):

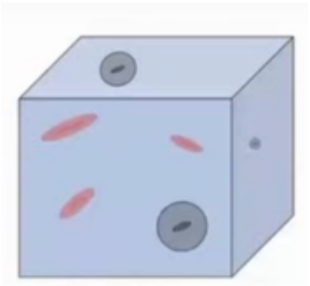
CNN layers (19)

+ dense (2)

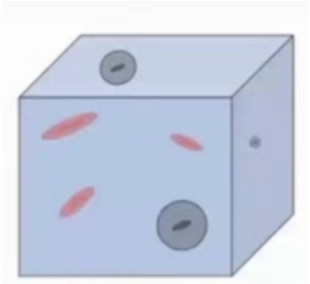
Villaescusa-Navarro+ arXiv:2109.10915

From one galaxy to global parameters

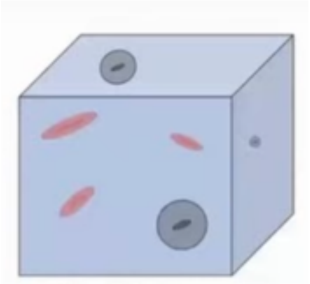
$\Omega_m = 0.32$
$\sigma_8 = 0.79$
$A_{SN1} = 1.2$
$A_{SN2} = 0.8$
$A_{AGN1} = 1.5$
$A_{AGN2} = 0.7$



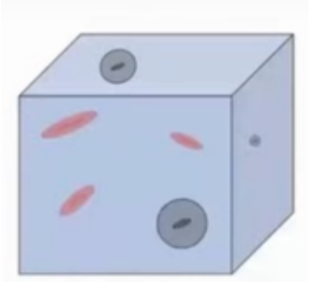
$\Omega_m = 0.27$
$\sigma_8 = 0.85$
$A_{SN1} = 0.7$
$A_{SN2} = 0.9$
$A_{AGN1} = 1.4$
$A_{AGN2} = 0.8$



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



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

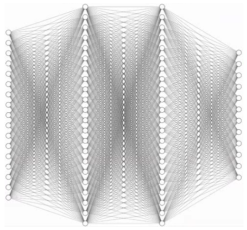


random galaxy: $\longrightarrow$

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

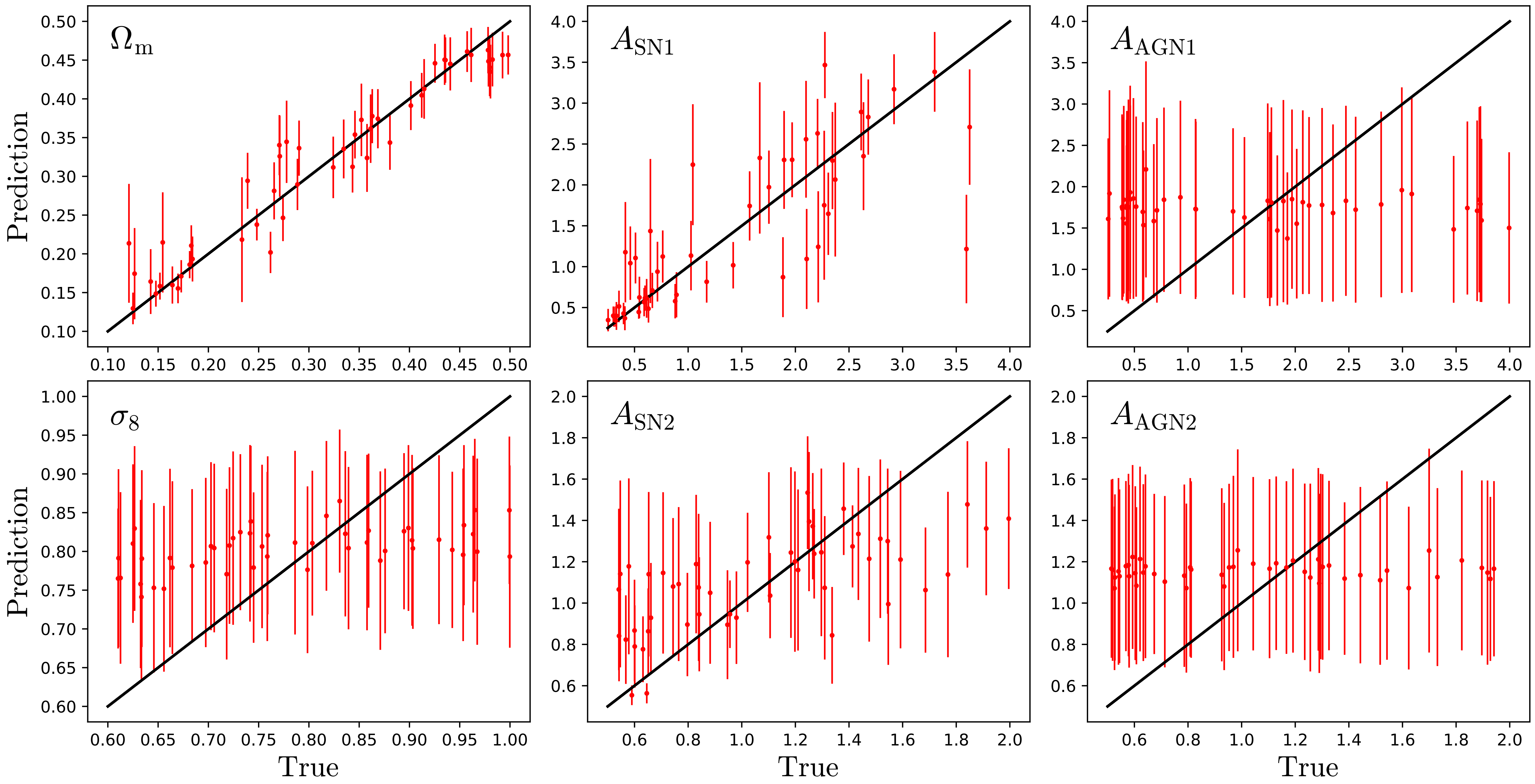
(SIMBA)

$\xrightarrow{?}$ Ω_m



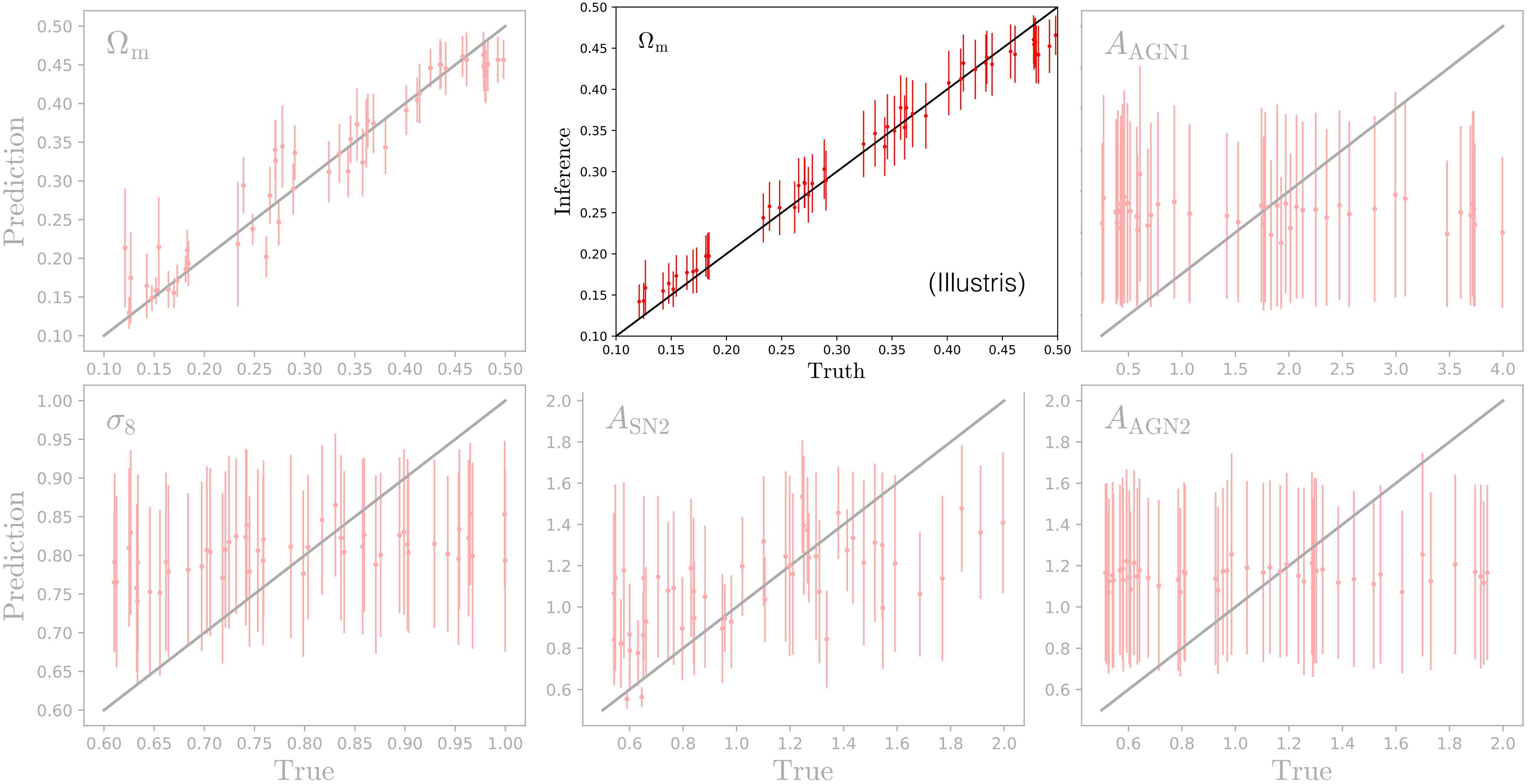
https://camels.readthedocs.io

From one galaxy to the matter density



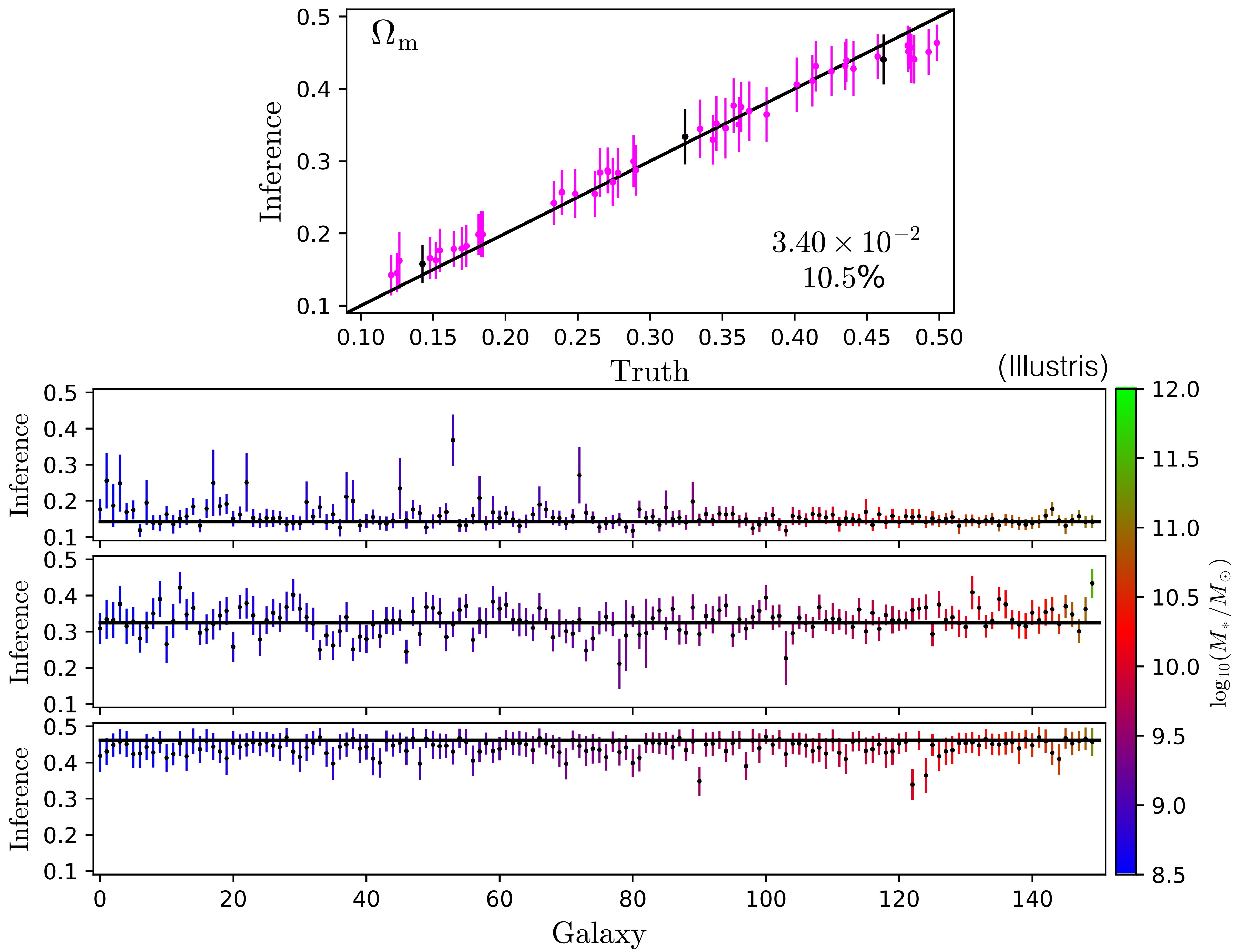
(SIMBA)

From one galaxy to the matter density



→ recovery of matter density for random galaxy in random simulation

Robustness: Matter density and galaxy mass



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

Importance of properties

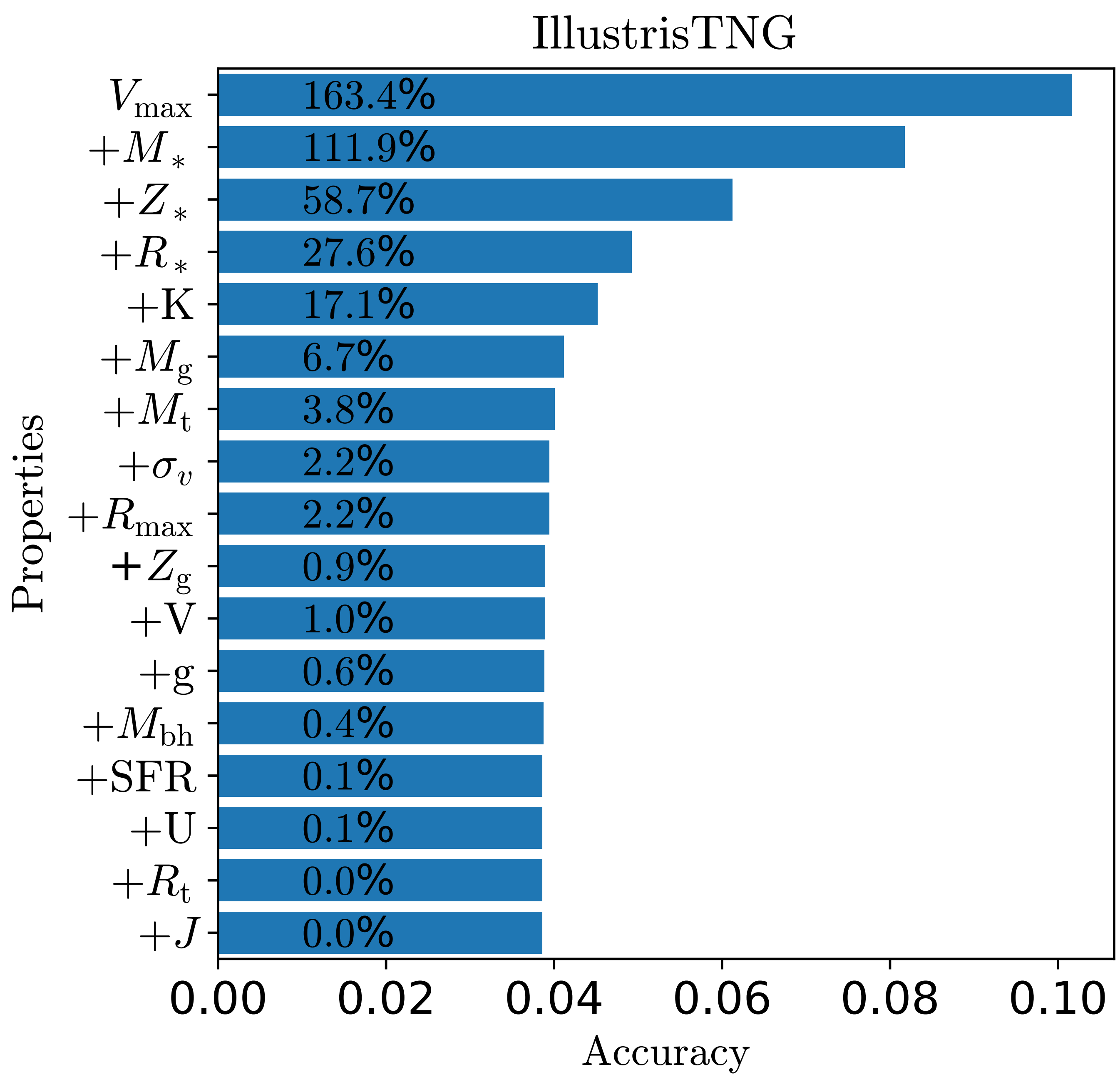
Only take into account relation input & predictions for attribution.
Show importance of feature on predictions via one-by-one retraining.

$$\{V_{\max}, M_*, Z_*, R_*, \mathrm{K}\}$$

Illustris

$$\{V_{\max}, M_*, R_{\max}, Z_*, R_*\}$$

SIMBA



Model-agnostic eXplainable AI (XAI)

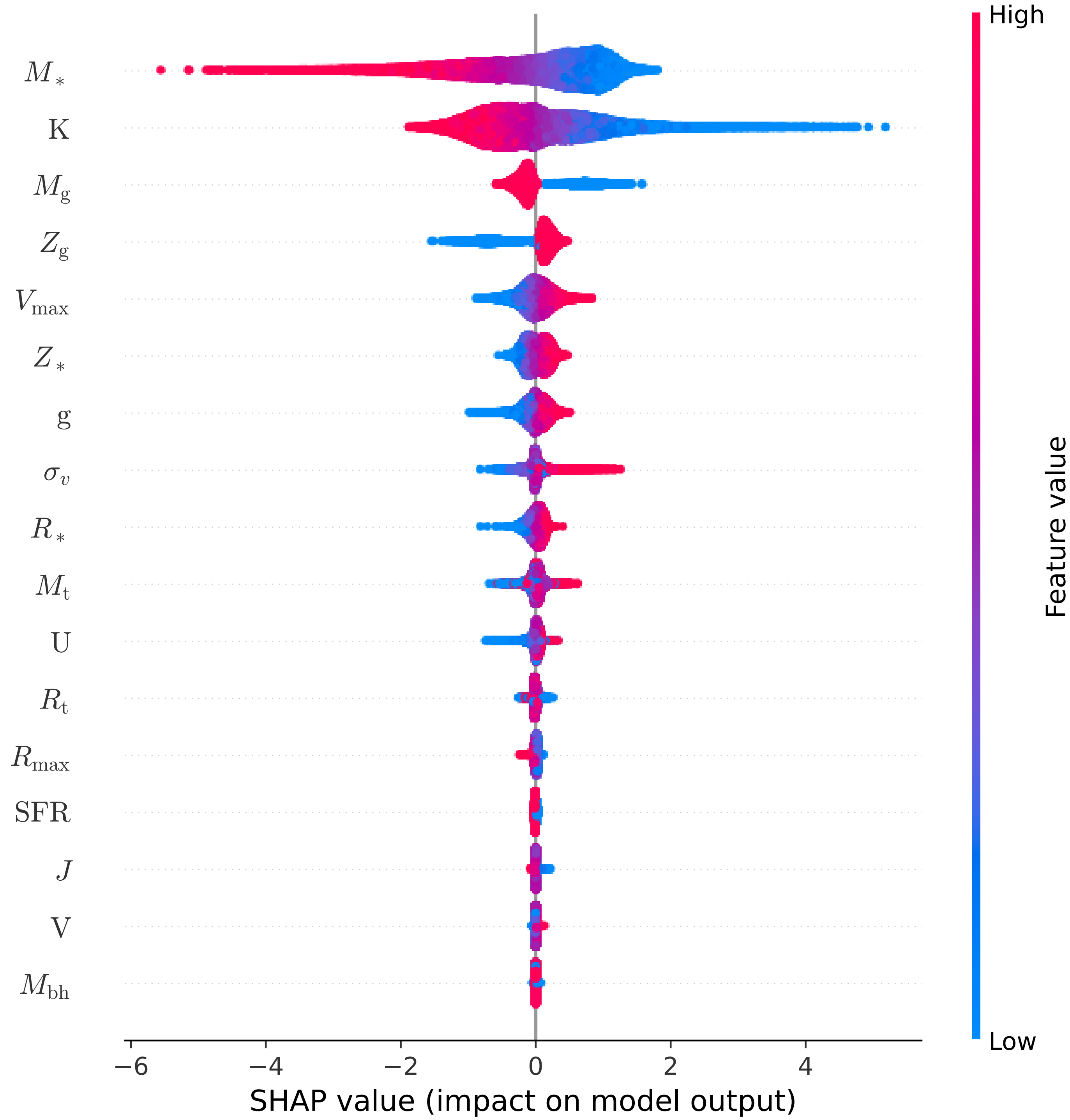
Example: SHAP values https://arxiv.org/abs/1705.07874
SHAP = SHapley Additive exPlanations

Approach from Game Theory:
Assign to each input feature a value.
A larger value indicates higher importance for the output prediction.

Do so via optimal credit allocation of possible 'coalitions'.

· difference in output by including/excluding feature
· average over N! orderings
· repeat: all subsets of remaining features

before via re-training: $\{V_{\max}, M_*, Z_*, R_*, \mathrm{K}\}$



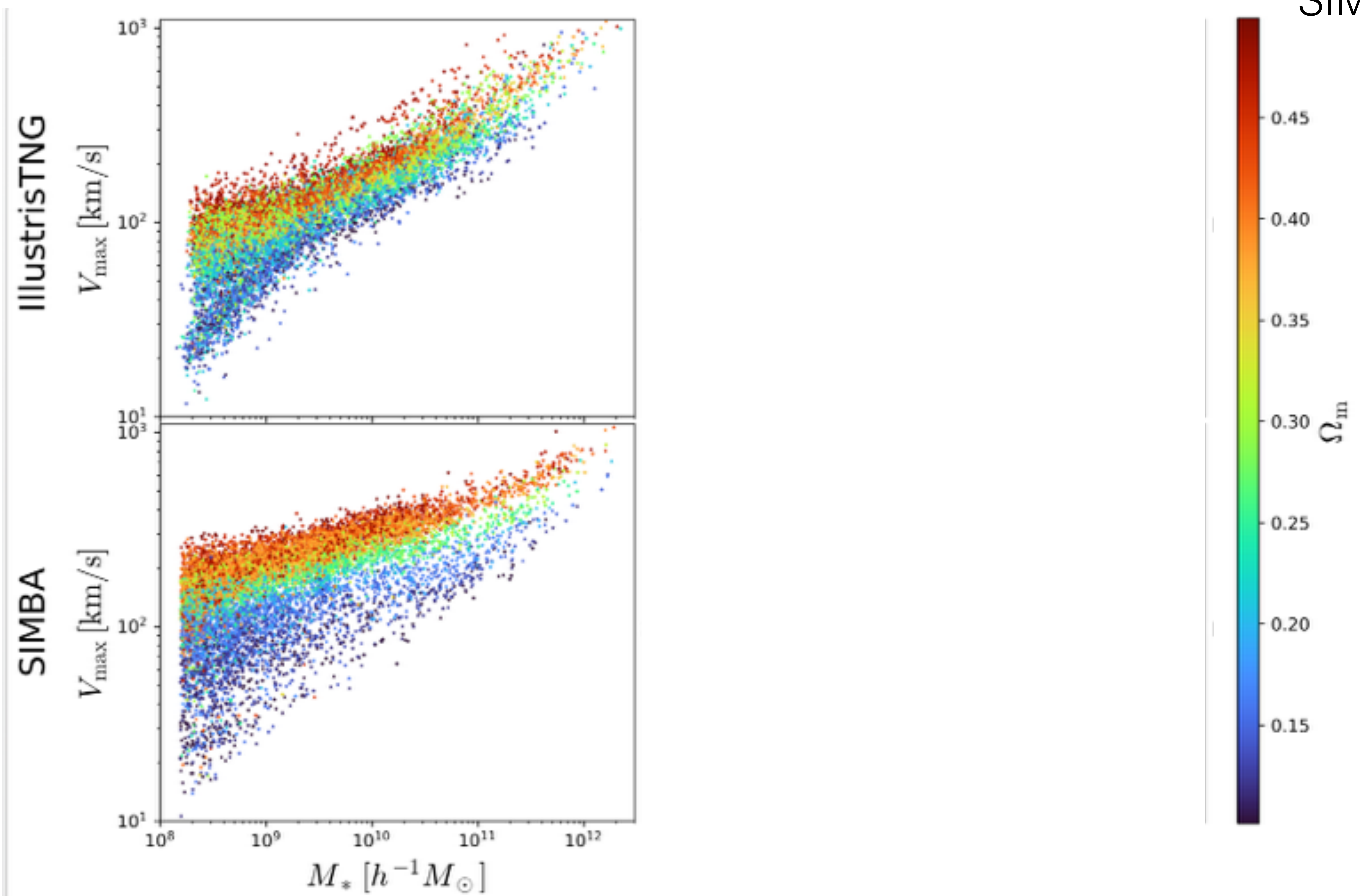
Importance of properties

$$\{\underline{V_{\max}}, M_*, Z_*, R_*, \mathrm{K}\}$$

Illustris

$$\{\underline{V_{\max}}, M_*, R_{\max}, Z_*, R_*\}$$

SIMBA



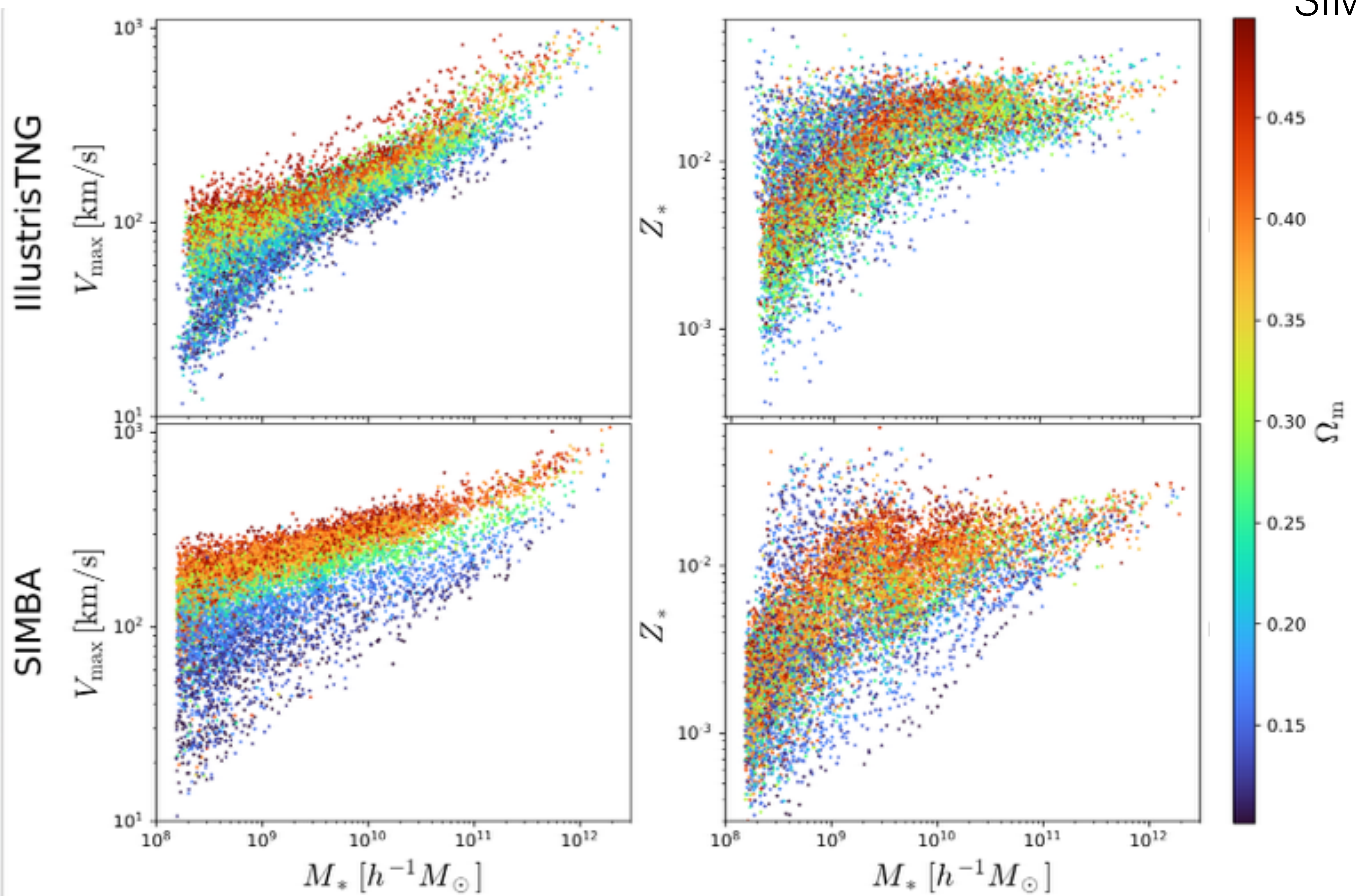
Importance of properties

$\{\underline{V_{\max}}, M_*, Z_*, R_*, \mathrm{K}\}$

Illustris

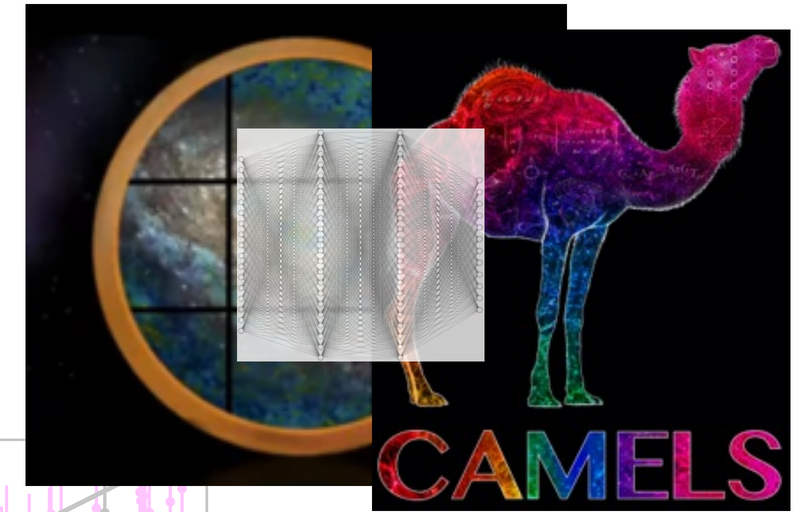
$\{\underline{V_{\max}}, M_*, R_{\max}, Z_*, R_*\}$

SIMBA



Also: PCA

Cosmology with one galaxy



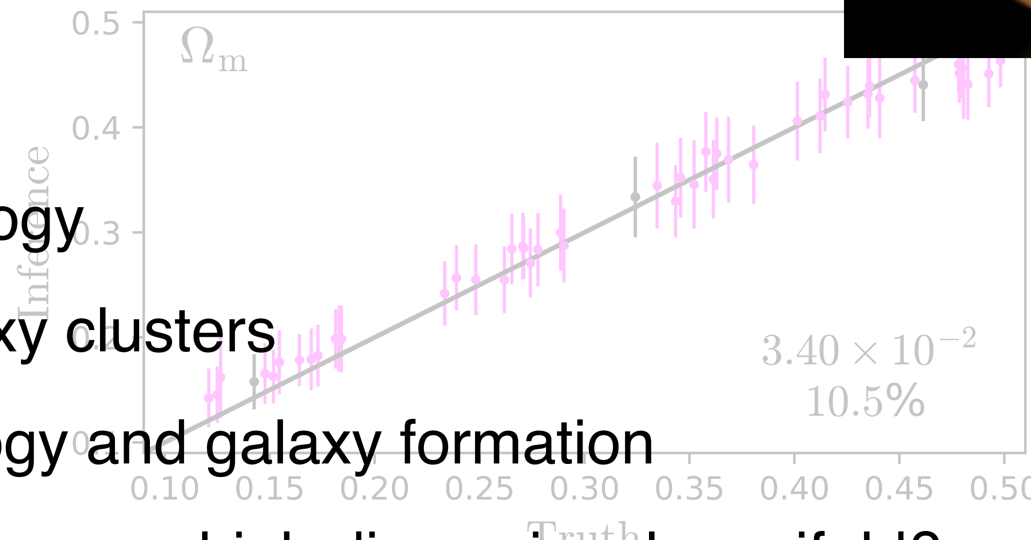
- Avenue to constrain cosmology

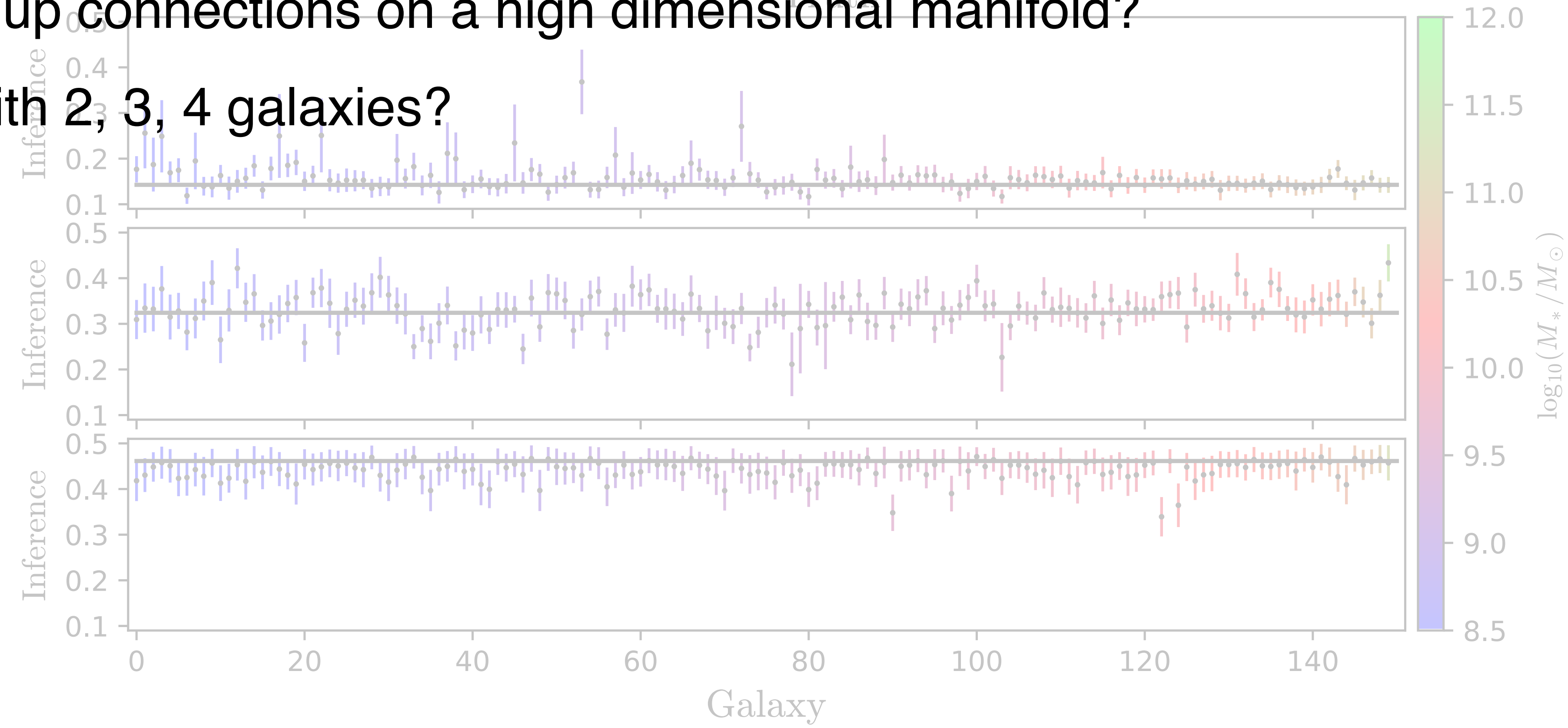
  Interesting if true, think galaxy clusters

  A new link between cosmology and galaxy formation

  Network picks up connections on a high dimensional manifold?

- Cosmology with 2, 3, 4 galaxies?



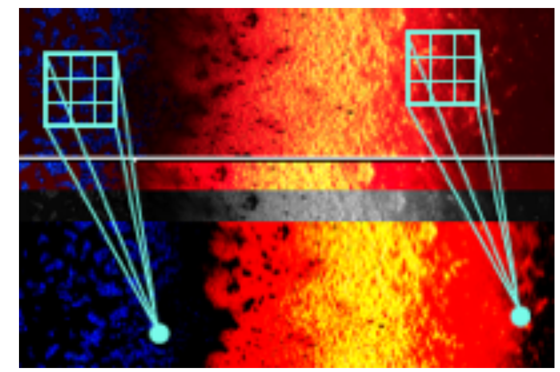


Computer Vision Astrophysics & Cosmology

novel scientific life cycles

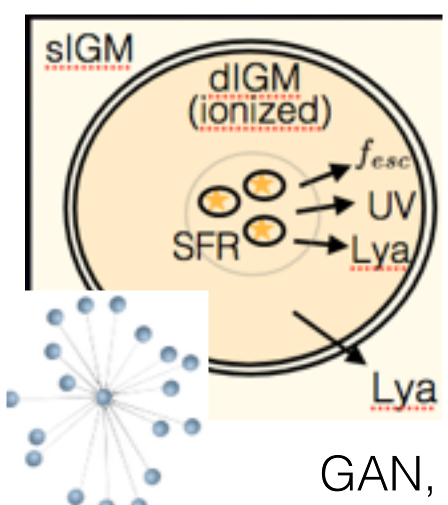
CNN, U-Net,
region-proposal, YOLO, ..

observation, signal extraction



Gillet+19
Prelogovic+22
Neutsch, Heneka, Brüggen 2022
arXiv:2201.07587

$(\Omega_{\rm m},\, \zeta,\, T_{\rm vir},\, L_{\rm X},\, E_0,\, m_{\rm WDM})$

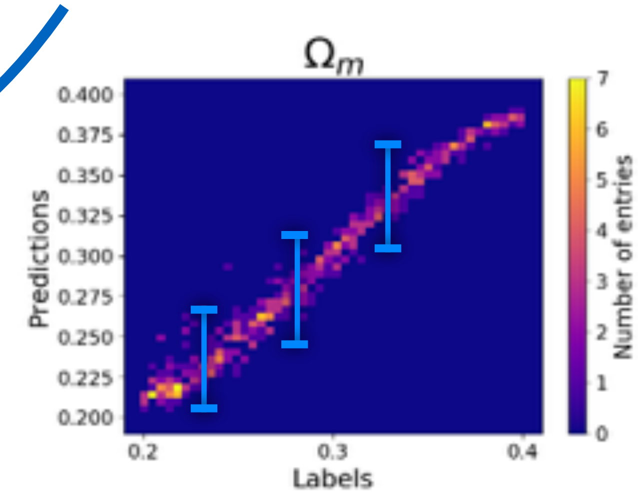


modelling, simulation

GAN, VAE,
GNN, NFs, ..

inference

BNN, MDN,
ensembles, NFs



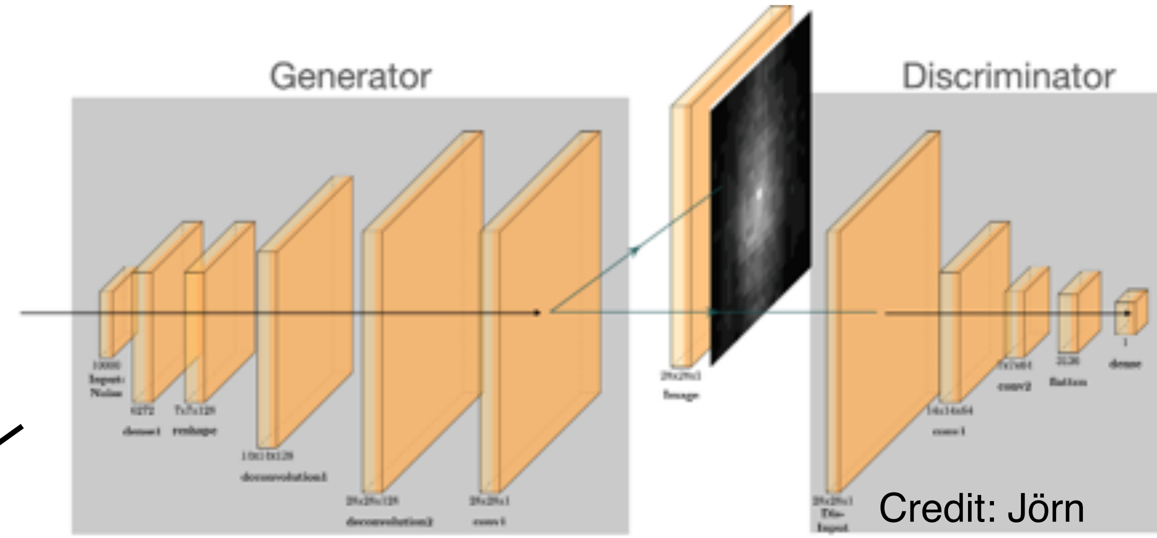
Example: Generation / Simulation

How to GAN galaxy clusters

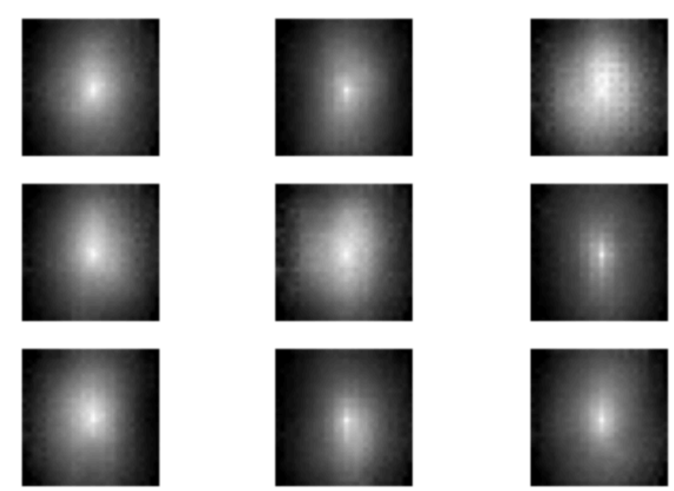
Scientific use case:
Fast, reliable way to simulate galaxy cluster images (in X-rays, at different cosmologies)

The WGAN - Wasserstein GAN

Seeking convergence instead of equilibrium



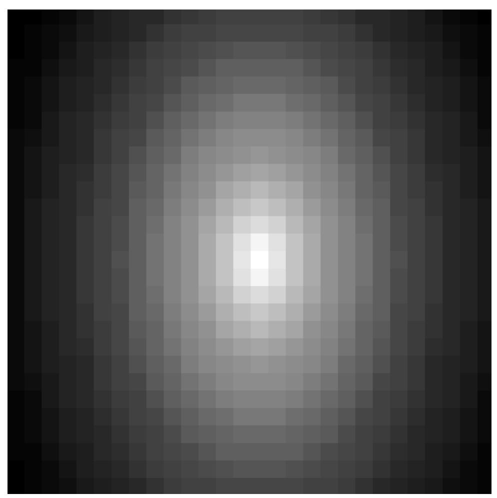
Training data from Comparat et al. 2020 (arXiv:2008.08404)

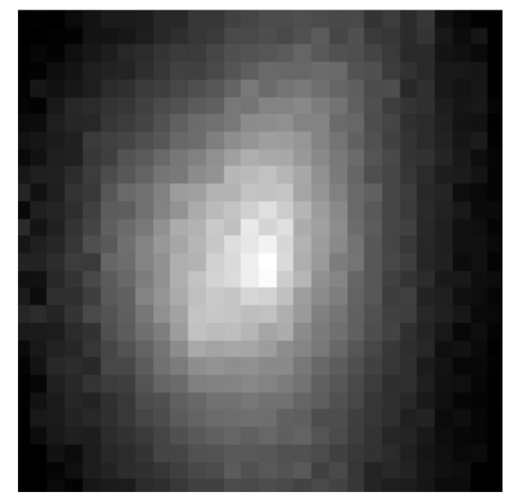


@Jörn Bach, Caroline Heneka, Marcus Brüggen

See arXiv:1701.07875 for a comprehensive mathematical introduction of the WGAN

How to GAN galaxy clusters





Real vs. simulated vs. WGAN-generated?

csheneka / GAN-tutorial Public

..explore different failure modes of GANs

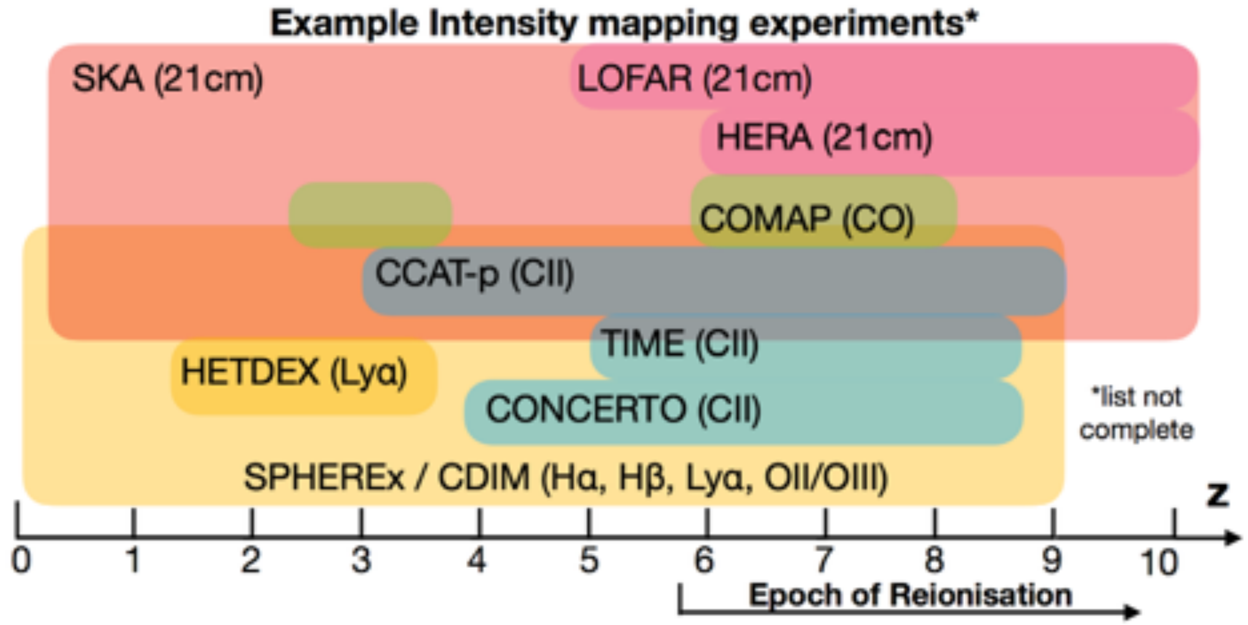
Radio interferometry - imaging in Fourier space

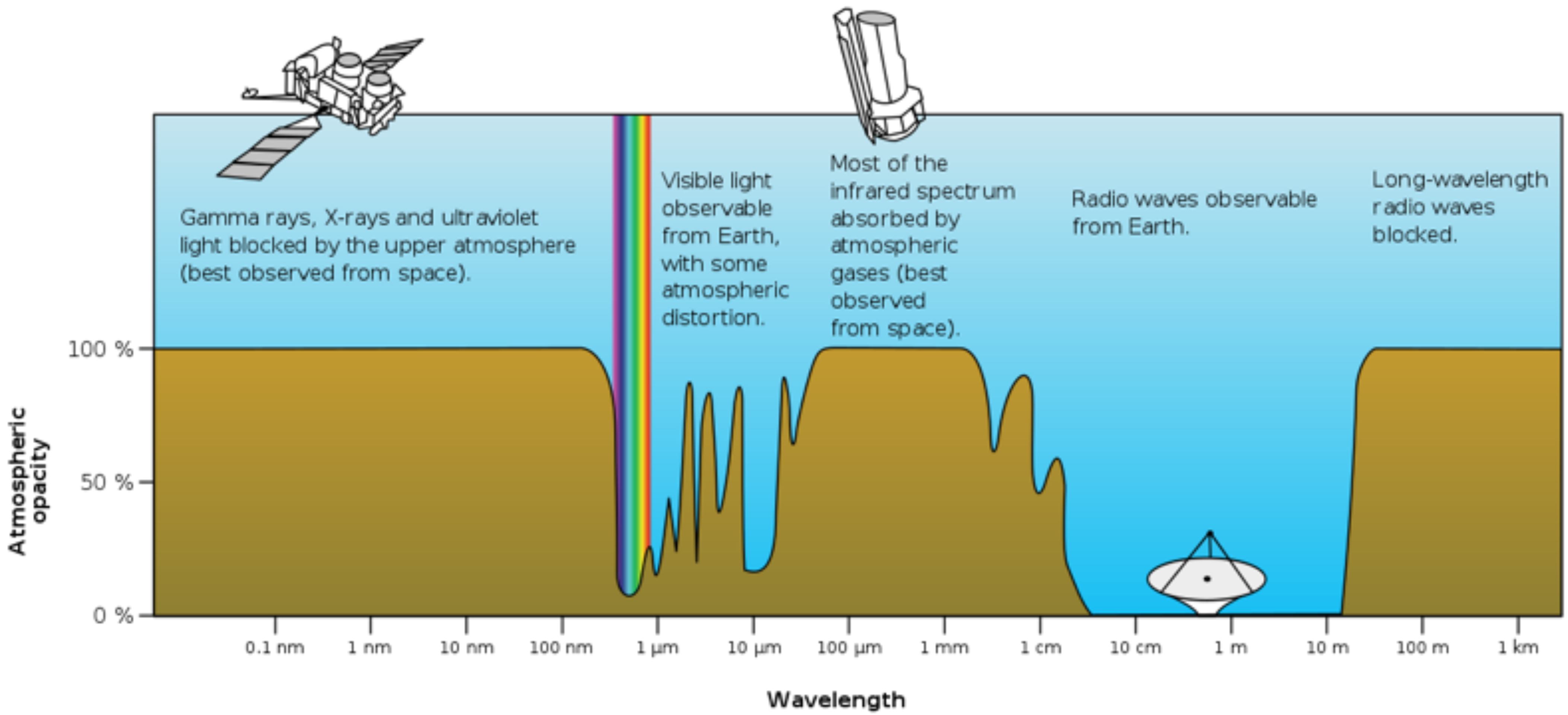
‘The’ inverse problem in astronomy/astrophysics

New frontiers in astronomy and astrophysics

3D imaging to track the history of the Universe

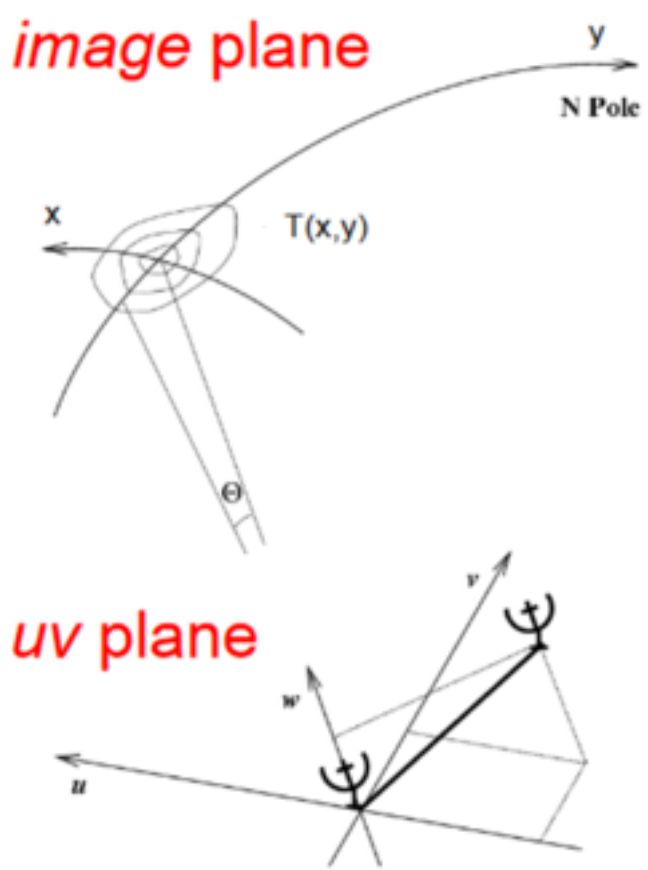
Radio and IR legacy surveys
(Hα, Lyα, OIII, Hβ, CII, .., CIB, ..)

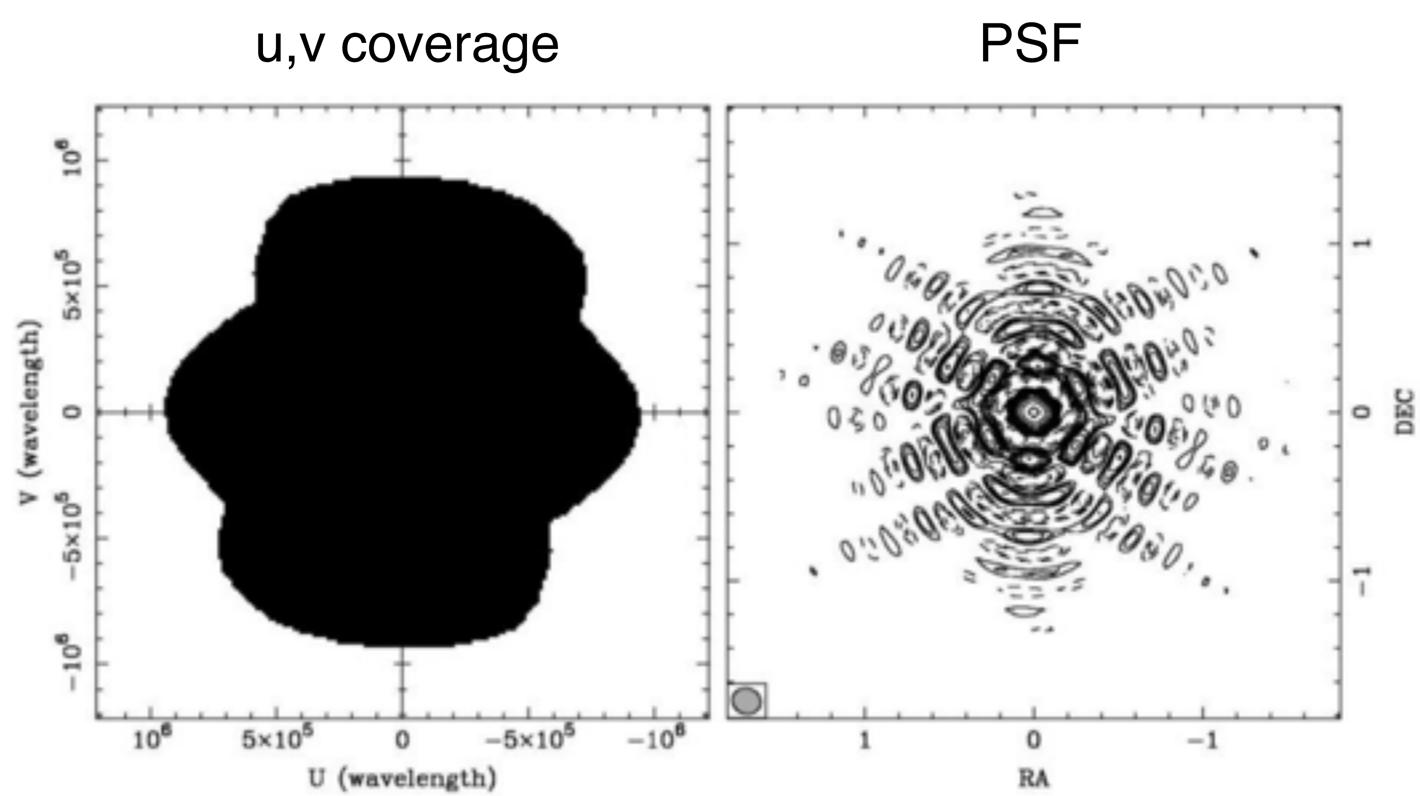




Windows in the atmospheric absorption. (Image: NASA)

Radio interferometry





@Tim Cornwell

$$T(x,y) = \int\int V(u,v)e^{-2\pi i(ux+vy)}dudv$$

complex visibility V(u,v) = interferometer response

Radio interferometry: an inverse problem

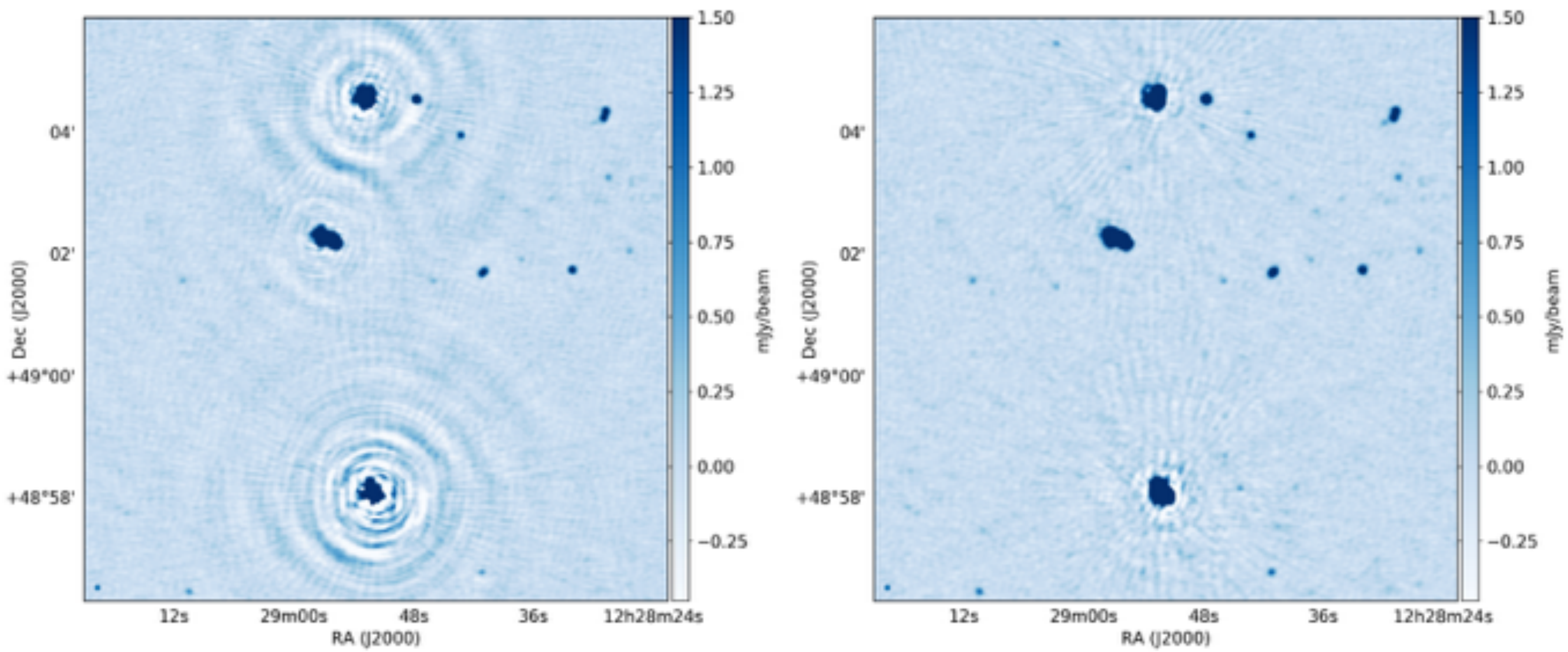
Goal: reconstruct I(x,y)

PSF (~known)

noise

visibility V(u,v) $\xrightarrow{\text{FT}}$ $$I^D(x,y) = R \times I(x,y) + n$$

diffuse, off-beam sources
(?)

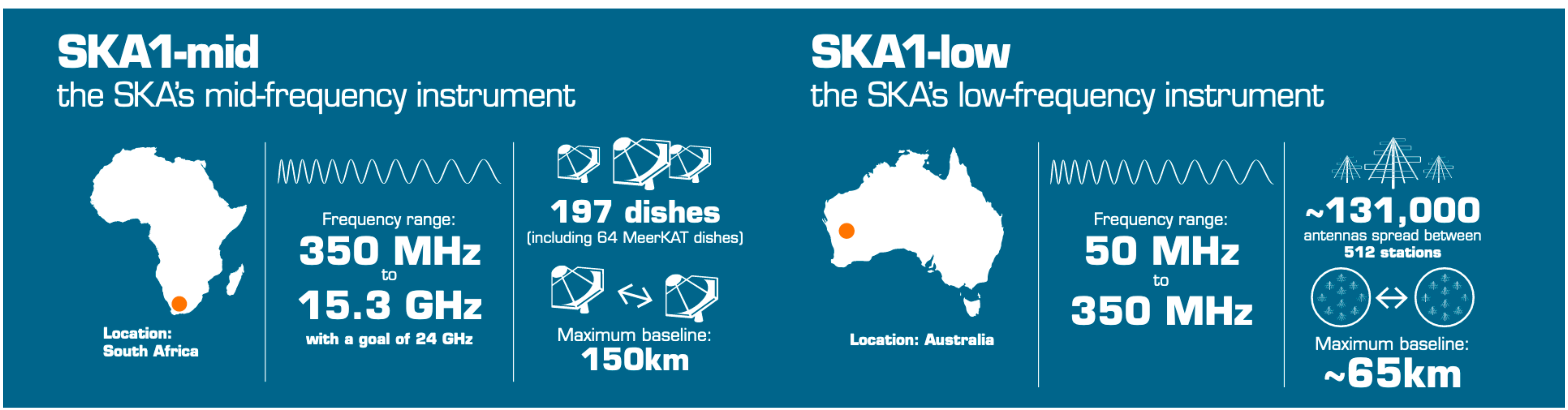


LOFAR DR2, Shimwell+22

The Square Kilometre Array (SKA)

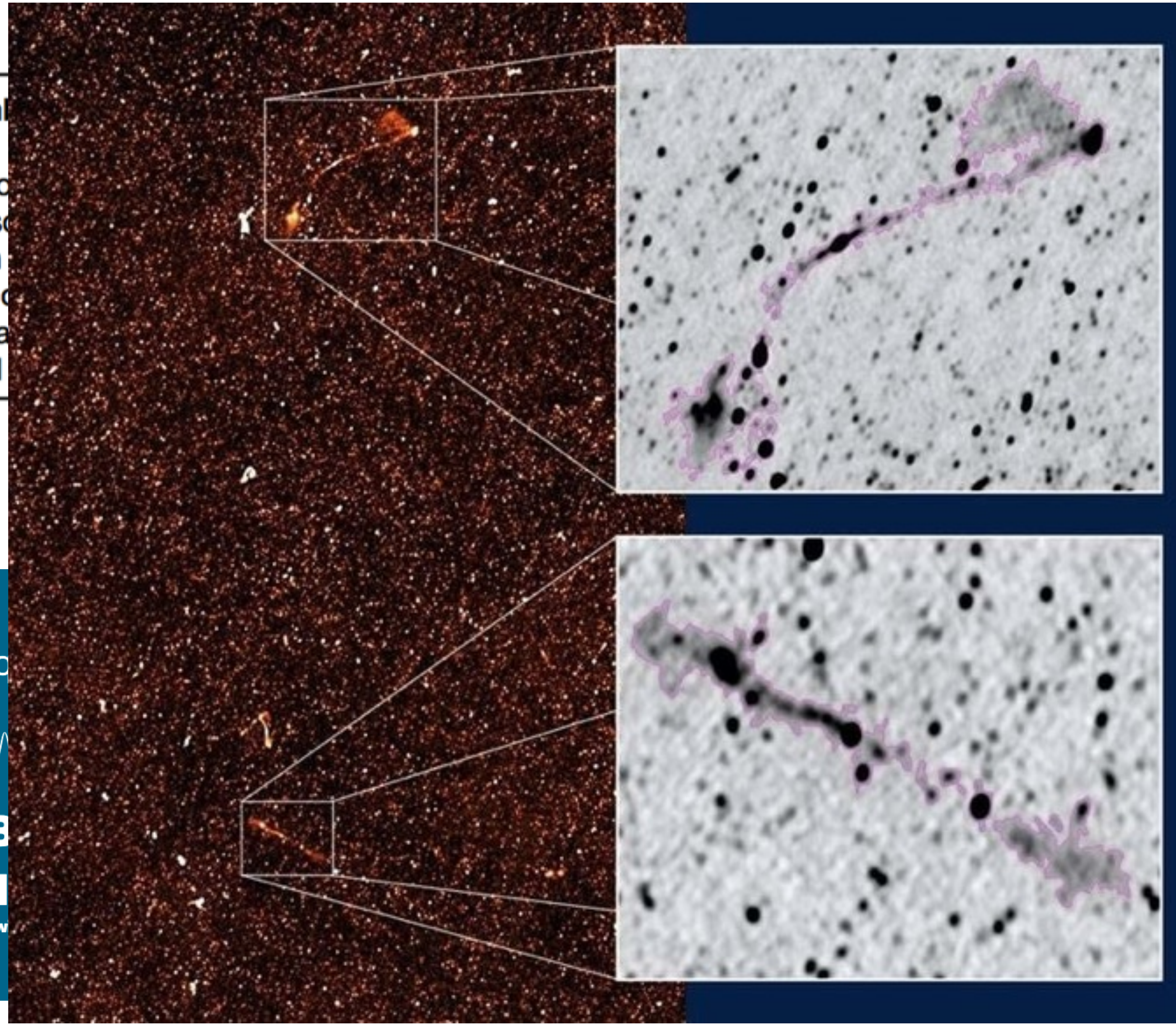
SKA in numbers
- Currently 16 member countries, >100 member organisations
- Routine science observations are expected to start in the late 2020s
- Consists of thousands of dishes and up to 1 million antennas, >1km^2 collecting area
- Expected data rate in full operation: 1 TB/s

SKA-LOW SKA-MID Credits: SKAO

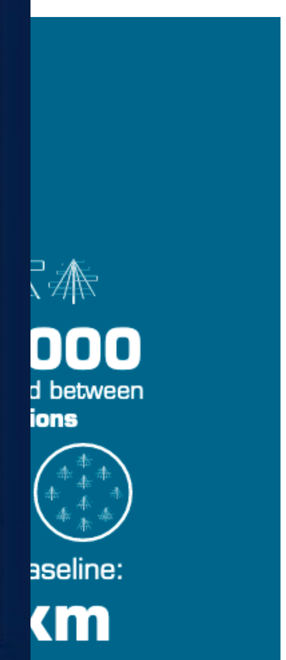


The Square Kilometre Array (SKA)

SKA in num
- Currently
  member c
- Routine sc
  to start in
- Consists c
  1 million a
- Expected



SKAO



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

Example: Denoising, detection

Detection & Inference in 3D: SDC2

Goal for individual HI sources:

Source finding

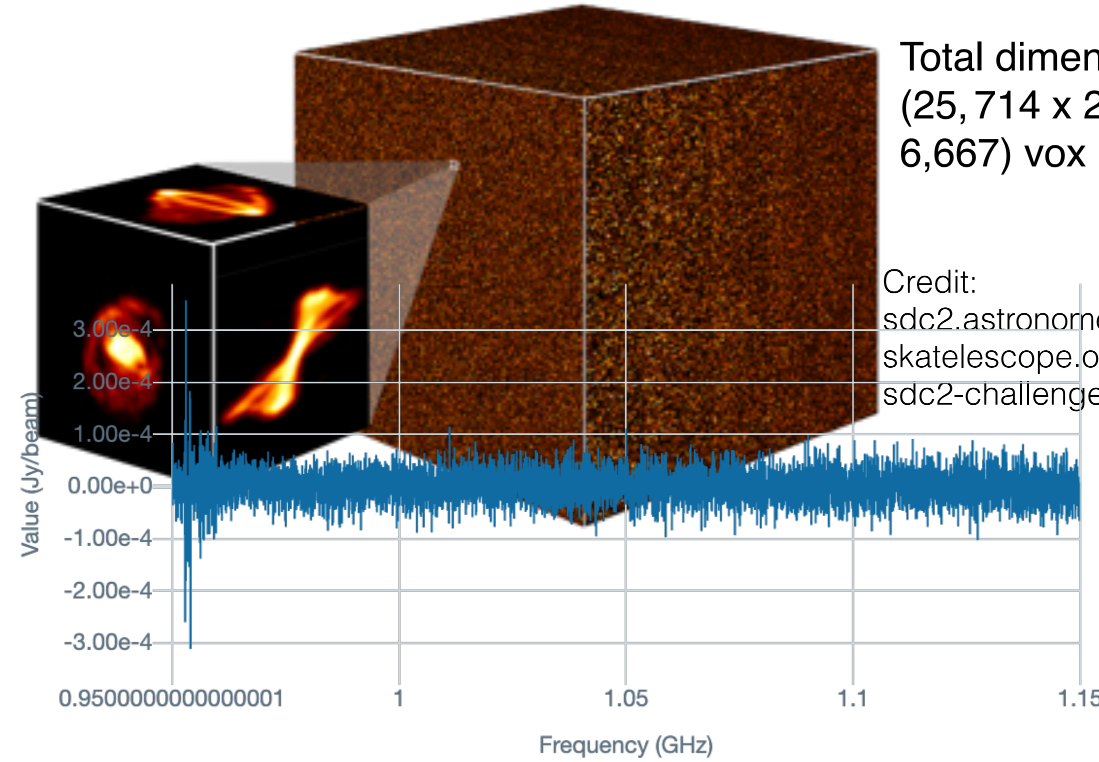
Location in RA, Dec,
central frequency (Hz)

Characterisation

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

The challenging HI sources:

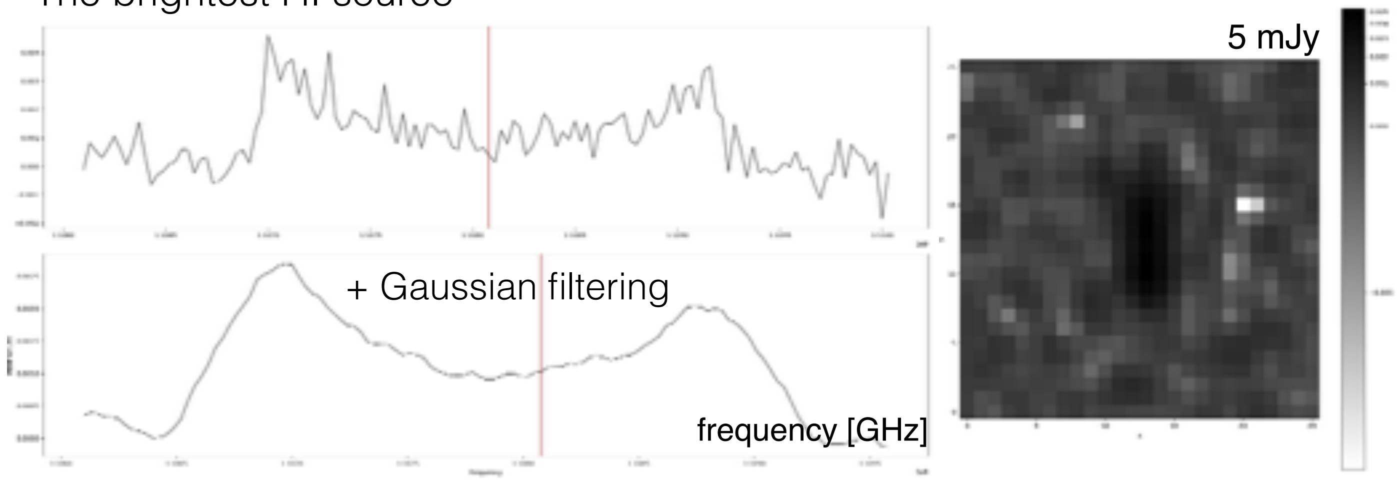
- **low S/N**
- **small spatial size**
- **systematics**



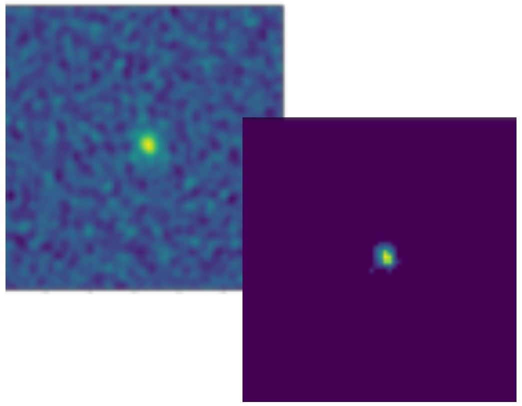
Total dimensions:
(25, 714 x 25 ,714 x
6,667) vox

Credit:
sdc2.astronomers.
skatelescope.org/
sdc2-challenge

The brightest HI source



Detection & Inference in 3D: SDC2

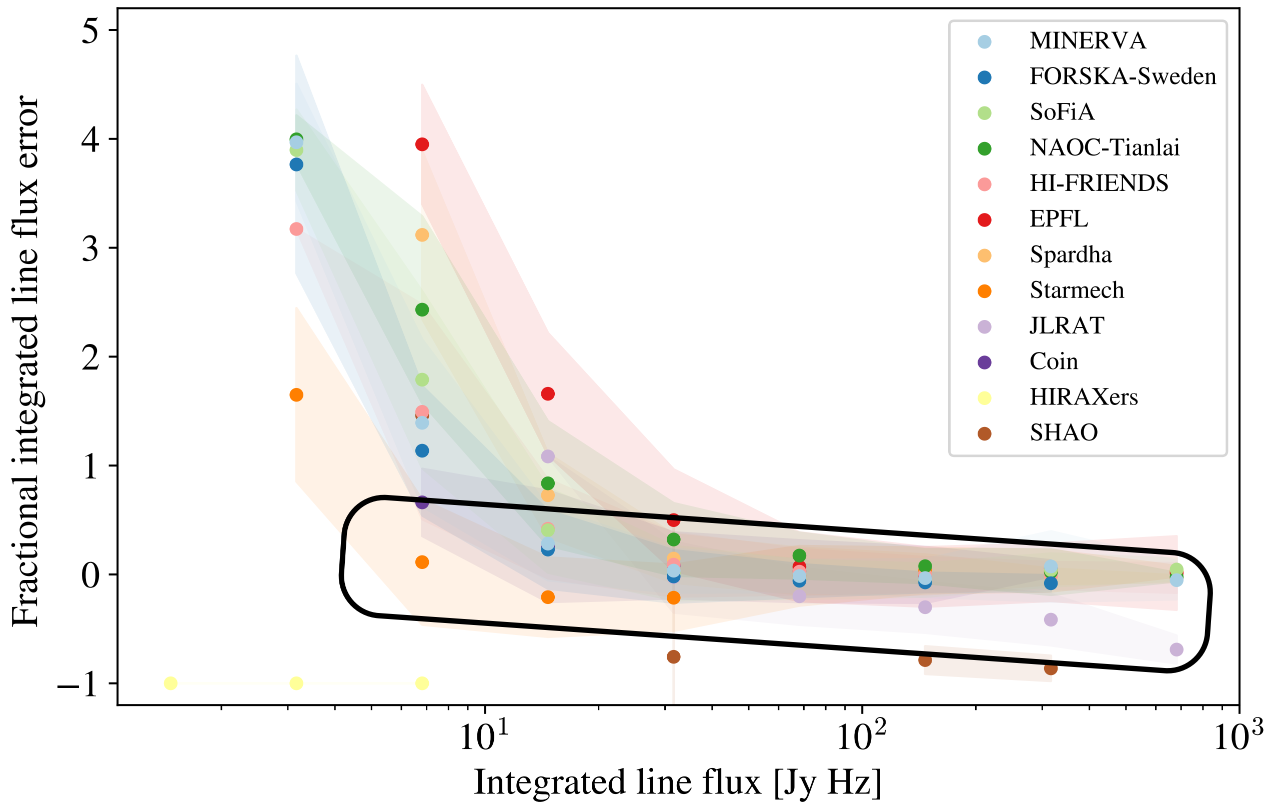


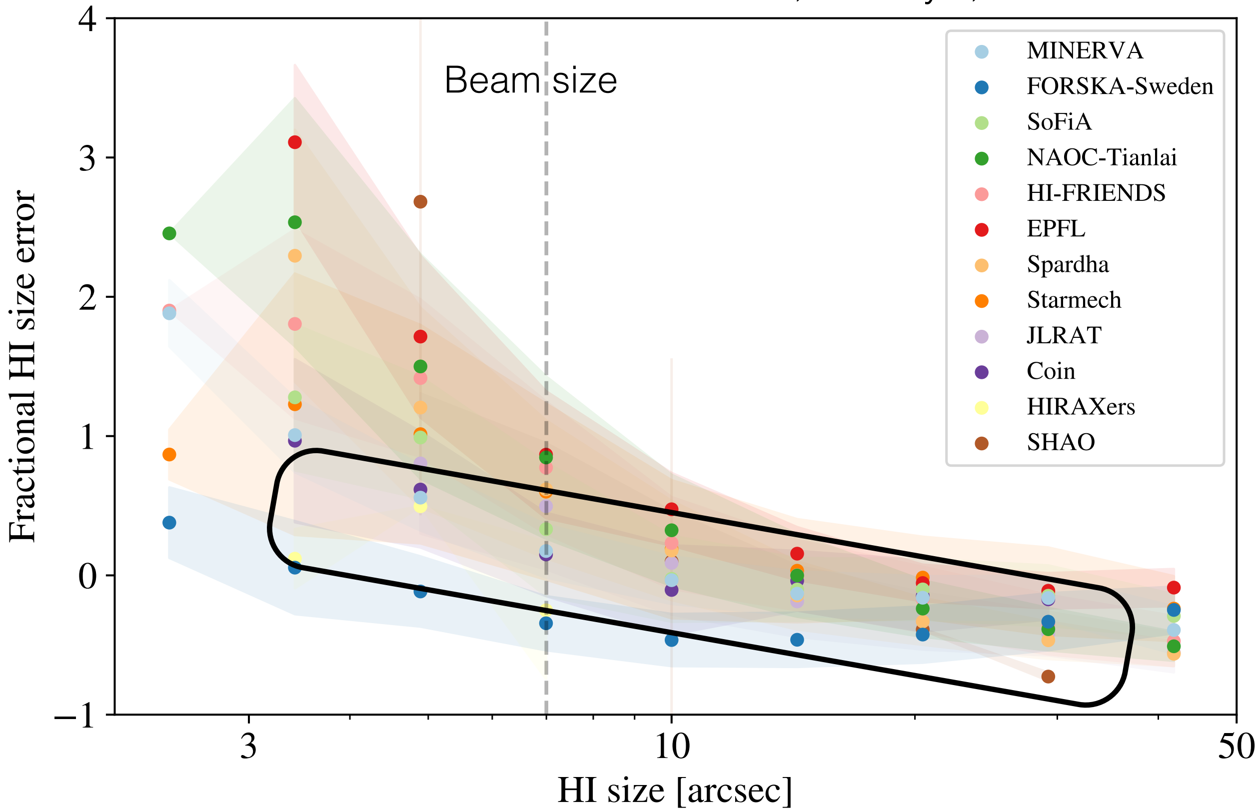
Deep learning for source detection and characterisation

3D U-Net

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

SDC2 results, Hartley+, under review





→ Recovery across wide range in HI flux and size

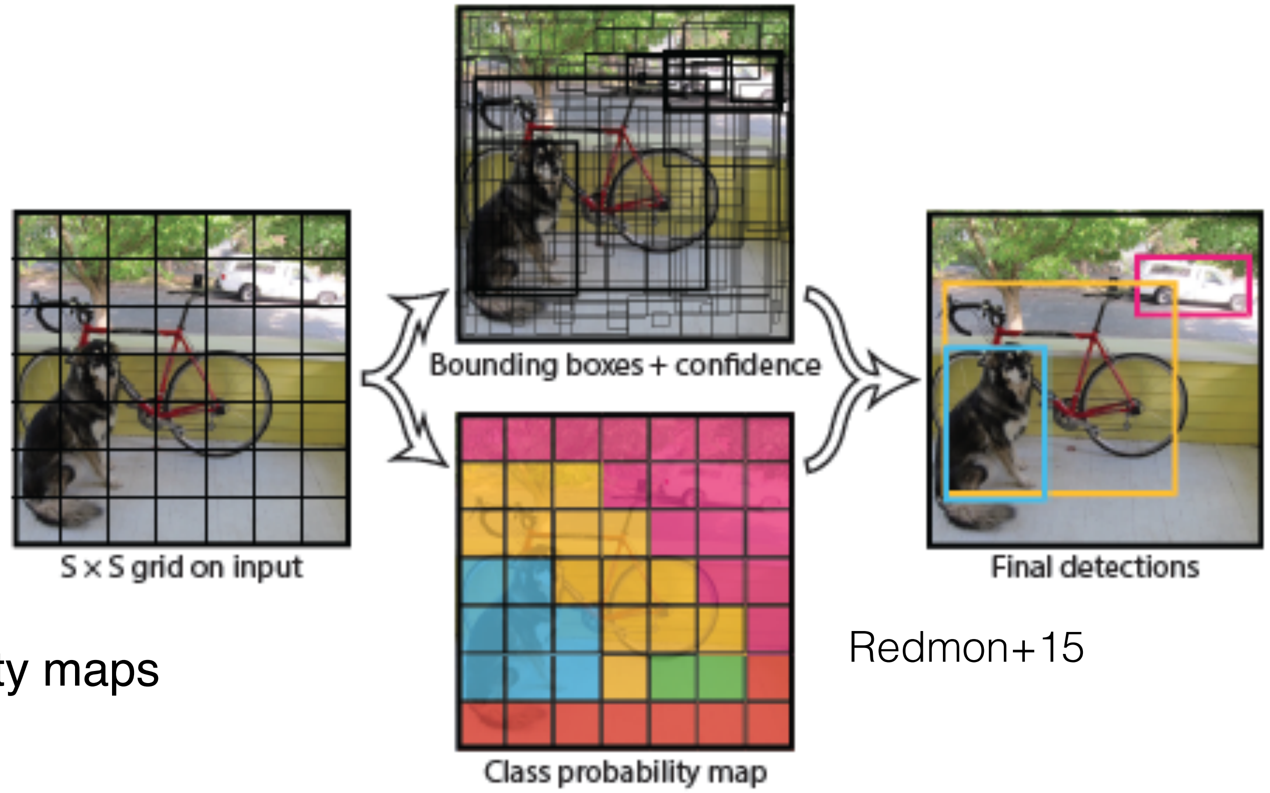
(also in terms of S/N)

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

Detection & Inference in 3D: SDC2

Deep learning for source detection and characterisation

Pitfalls:
- High sparsity
- **Choice of training set**
  choose high (enough) S/N sub-sample
- Avoid or mitigate false positives



Redmon+15

Top 3 SDC2 models

- *Minerva*: ensemble of
  1)YOLO (you-only-look-once), Redmon+15,16
  Turning classification into regression with probability maps
  2) CHADHOC, 'traditional' detection + CNN(s)

- *FORSKA-Sweden*: U-Net segmentation, Ronneberger+15,
  pre-trained + SoFiA

- *Team SoFia*: SoFiA (Source Finding Application, in 3D),
  Serra+15, Westmeier+21
  Thresholding of peaks and reliability estimate

Take-aways:
- **Choice of training set**
  pre-training (see also YOLO-approach)
  let the network choose!?
- Multi-step and/or ensemble decision

Ongoing: SDC3 foreground challenge

The complete SDC3a dataset consists of a visibility measurement set (MS) and imaging products derived from it, as well as ancillary data. Teams can decide whether they want to analyse the MS or the imaging products, depending on their pipelines.

Details of the data products are as follows:

General

- Observation track length HA = -2 to +2 hours
- Thermal noise equivalent 1000 [h]
- Field of View: one SKA1-Low pointing at RA, Dec = 0h, -30deg

Measurement sets

- Size 5.4 [TB] (3.3 TB if concatenated)
- Integration time 10 [s]
- Channel width 100 [kHz]
- Frequency coverage 106 - 196 [MHz]

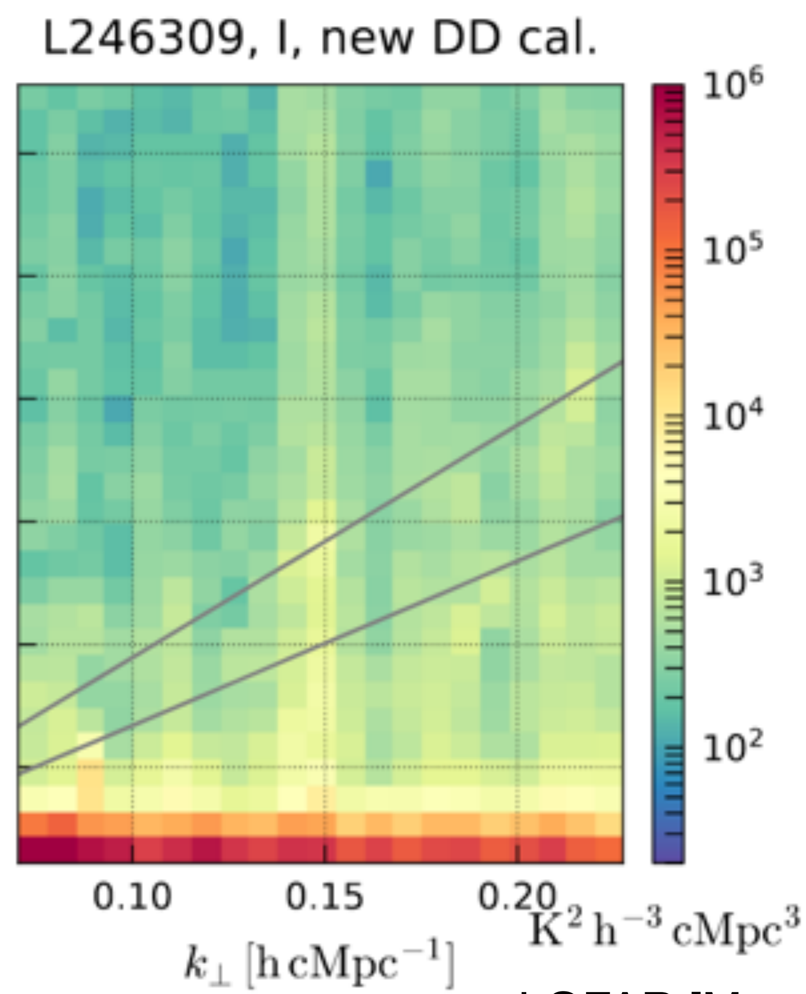
Image cube

- Weighting: Natural
- Pixel size [arcsec]: 16x16 arcsec
- Number of pixels in RA/Dec 2048x2048

Ancillary data

- Synthesised beam and primary beam at each frequency

Non-blind validation data

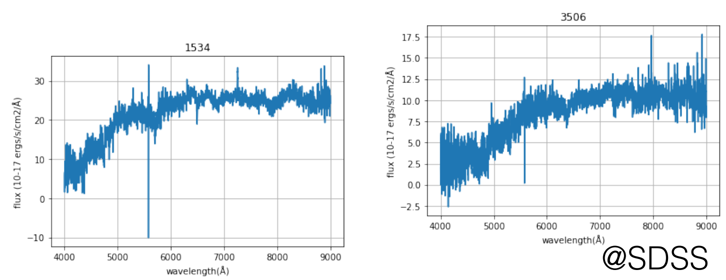


LOFAR [Mondal+20]

Why ML/DL **for astrophysics - Tutorials**

+ plenty of inverse problems

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

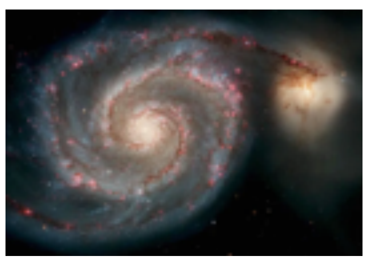


@SDSS

star vs. galaxy

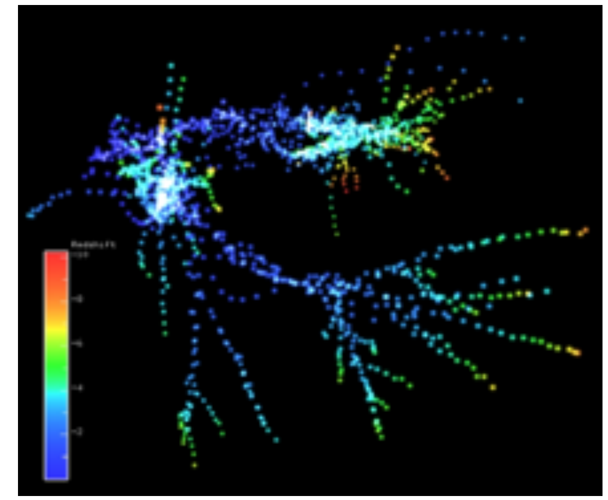
Representation learning

galaxy



@Hubble, NASA

@Hubble, NASA

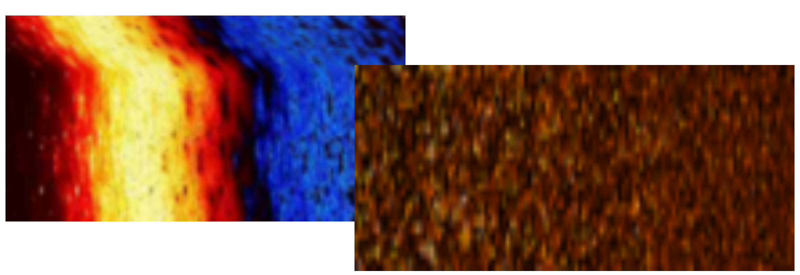


@Chris Fluke, Swinburne University of Technology

Hierarchical learning

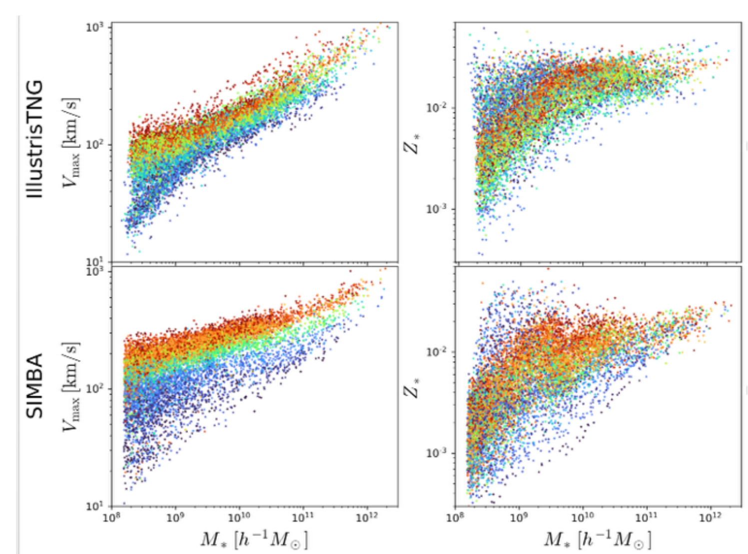
Non-linear, non-Gaussian

High-dim. correlations



fluctuation fields

@SKAO



arXiv:2201.02202

Example: Discovery

Example clustering: Open Cluster identification

Open Clusters (OCs) as tracers of the Galactic disk

Astrometry:
Gaia DR2

k-*d* tree
(spatial,
density-aware)

2048
tiles
→

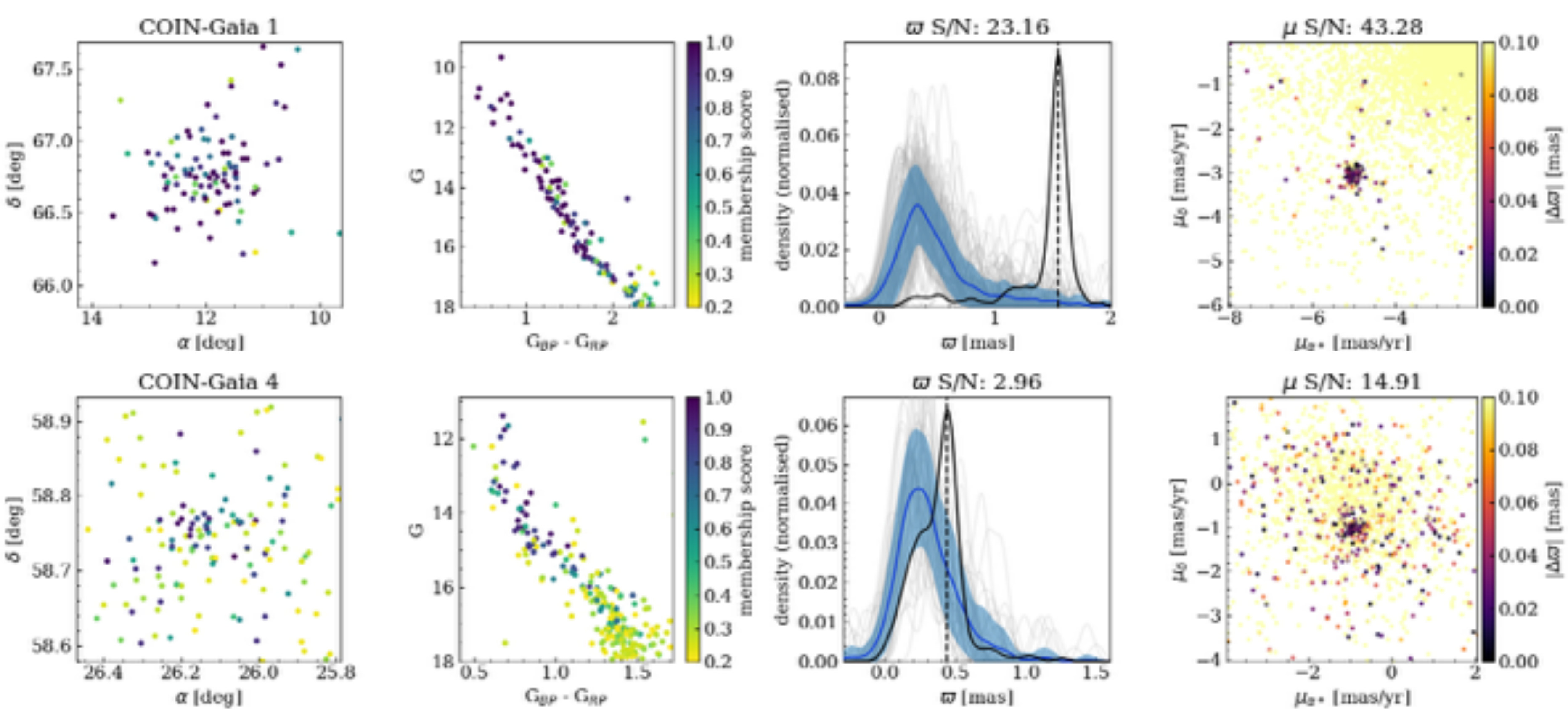
Gaussian mixture model
(proper motion)

mean: bulk motion OC
variance < field

candidate region
> 10 members
→

unsupervised
membership
assignment

create groups:
k-means clustering $(\mu_{\alpha,\delta}, \bar{\omega})$
compare to uniform:
minimum spanning tree



+ Validation:
Colour-magnitude
diagrams

↓

41 new OCs
33 within 2kpc

Cantat-Gaudin et al.
A&A, 624 (2019)