Learning the Universe: Setting sail to the first galaxies

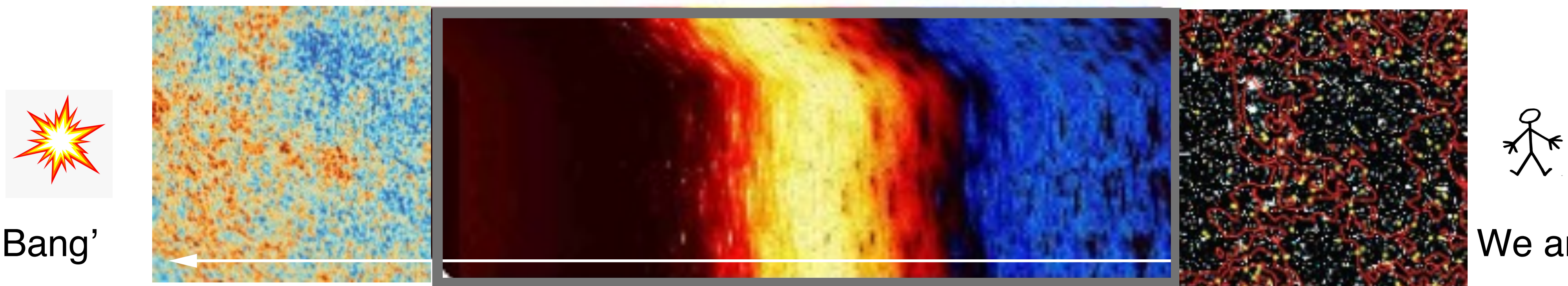
3D imaging to track the history of the Universe

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

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

(look-back time)
redshift z 20 15 10 7

galaxy surveys
e.g. LAEs at z~6



'Big Bang'

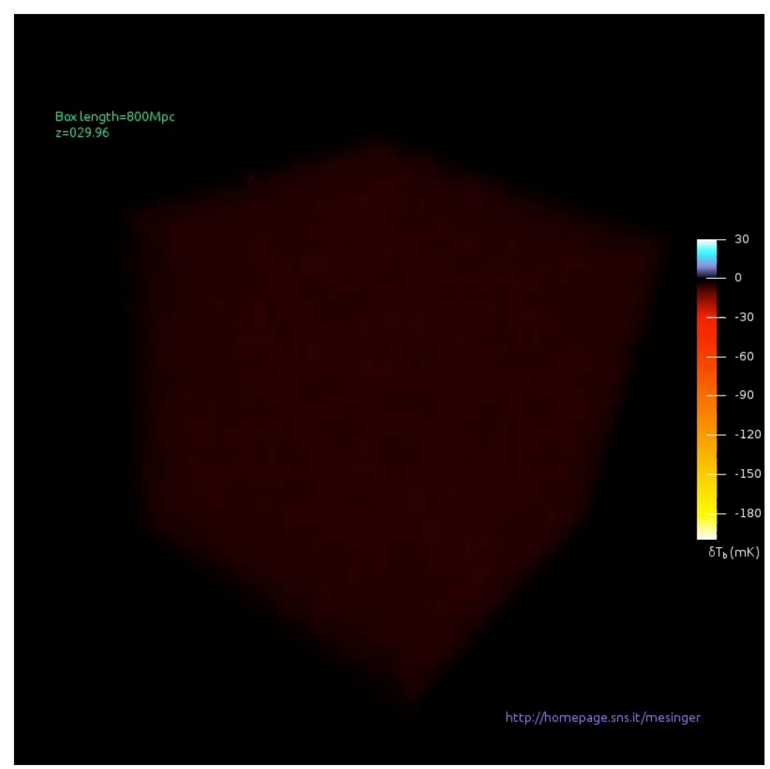
We are here

Credits: Planck, ESA

cosmological structure formation

Ouchi, Ono, Shibuya '20

+ multi-line intensity mapping
(21cm + Hα, Lyα, OIII, Hβ, CII, ..)



21-cm* 3D imaging**

* neutral hydrogen line

** blue = emission
red / yellow = absorption

Inference from 3D tomographic cubes

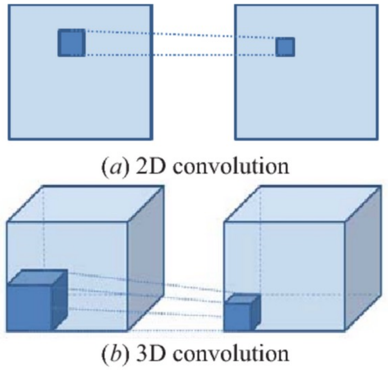
Direct likelihood-free inference from 3D tomographic cubes (21cm IM)

Various Options:

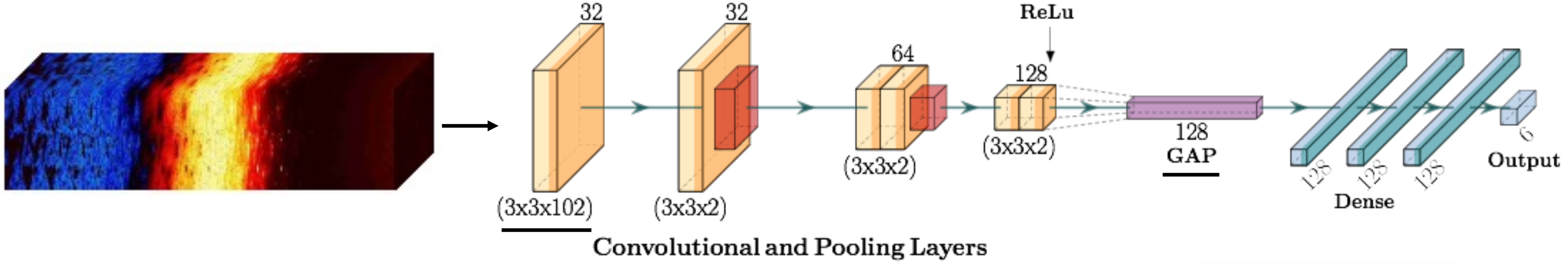
- Slicing and treatment with 2D CNN as image → 'standard' **2D CNN**, residual (skip connections) **ResNet**
- Time series (frequency) of co-eval images → **LSTM** network
- Full 3D convolution → 3D CNN:

[see e.g. Prelogovic+ arXiv: 2107.00018, Gillet+2019]

Moving from **2D to full 3D convolution**



Best-performing: simple Conv3D architecture



Database of ~5000 lightcones
140x140x2350 pix, 1.4 Mpc resolution
Training (single K80 GPU): ~20min/epoch, ~30 epochs

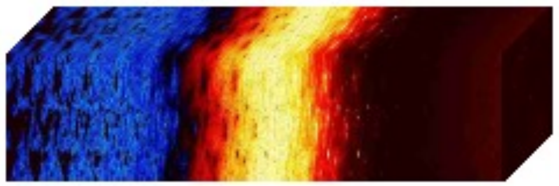
3D-21cmPIE-Net

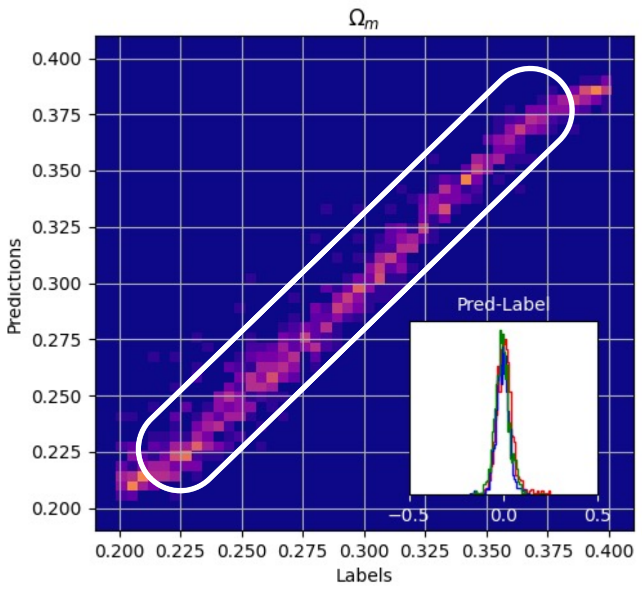
Neutsch, Heneka, Brüggen '22
arXiv:2201.07587

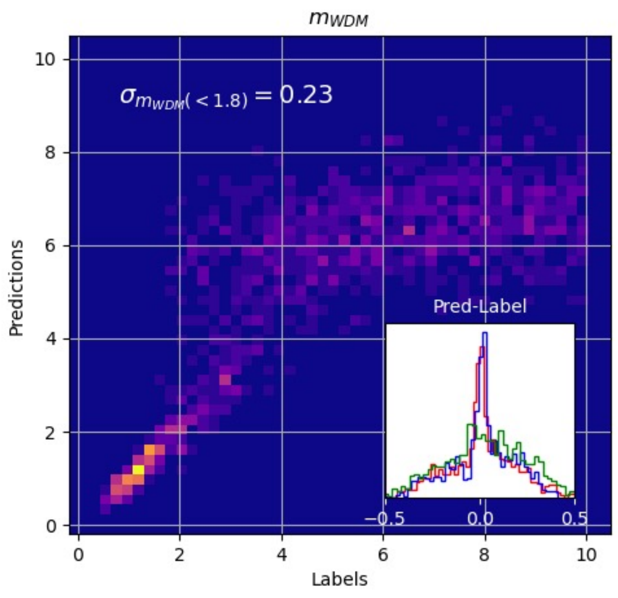
Inference from 3D tomographic cubes

Direct likelihood-free inference from 3D tomographic mock cubes (21cm IM)

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



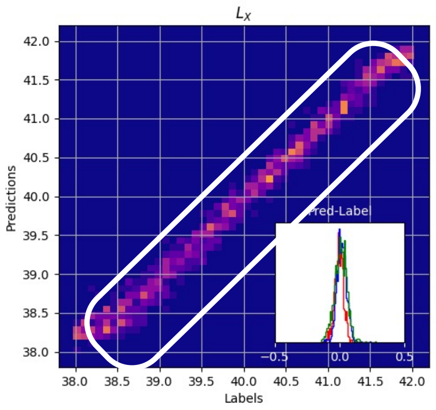


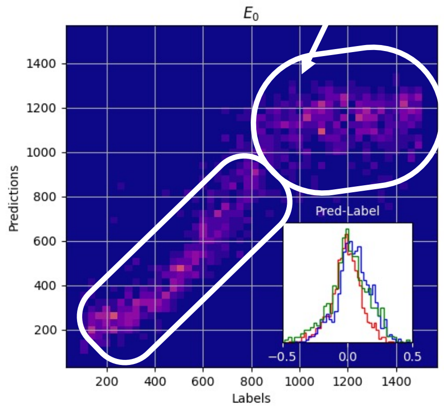


Cosmology

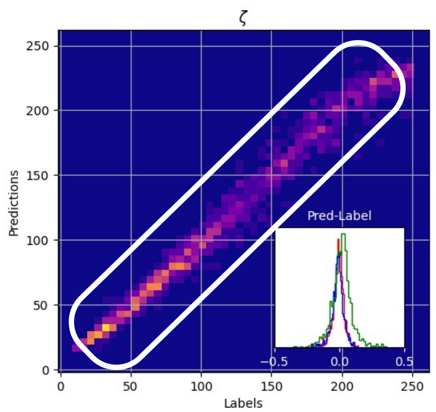
Cosmic Dawn

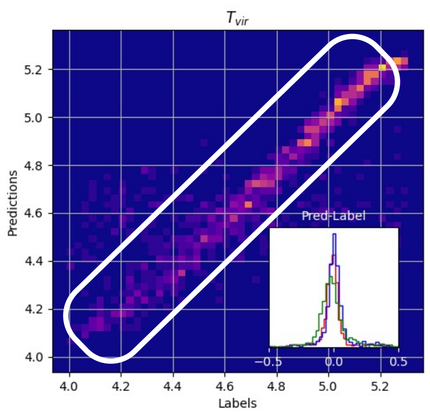
high E0, difficulty to escape





Reionisation





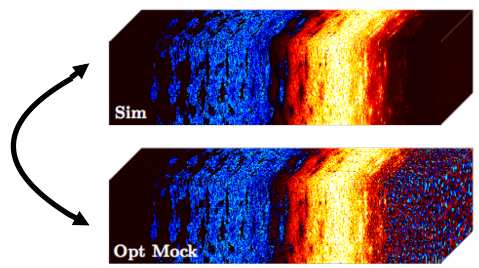
Directly constrain model parameters

Also tested: 'astro-only' inference

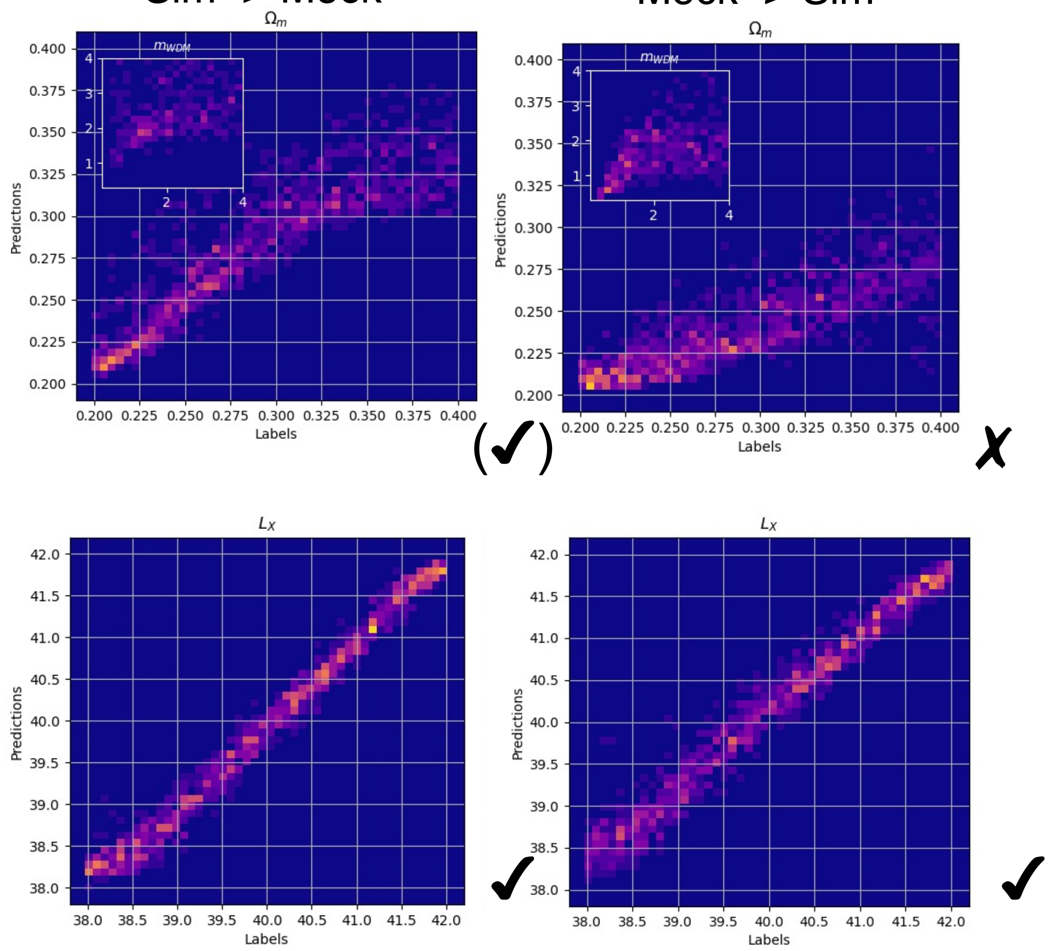
arXiv:2201.07587

Testing robustness & interpretability

1) Transfer learning with / without foregrounds



Sim -> Mock   Mock -> Sim

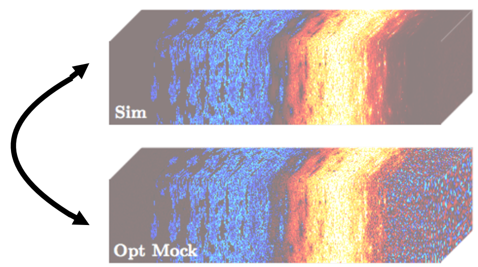


✓ Robust to foregrounds & systematics

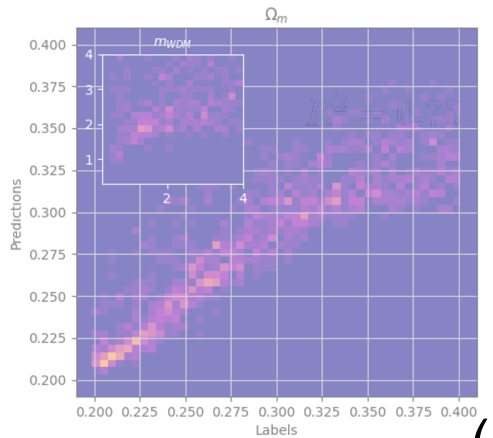
arXiv:2201.07587

Testing robustness & interpretability

1) Transfer learning Sims & Mocks

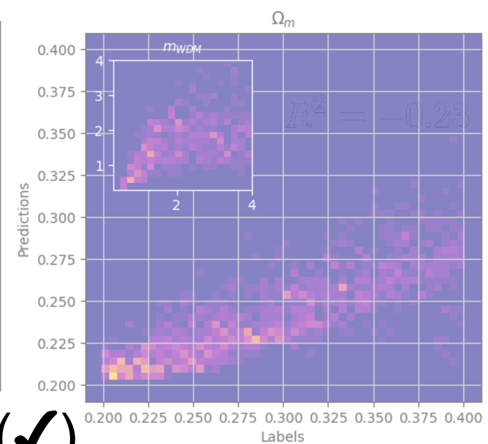


Sim -> Mock

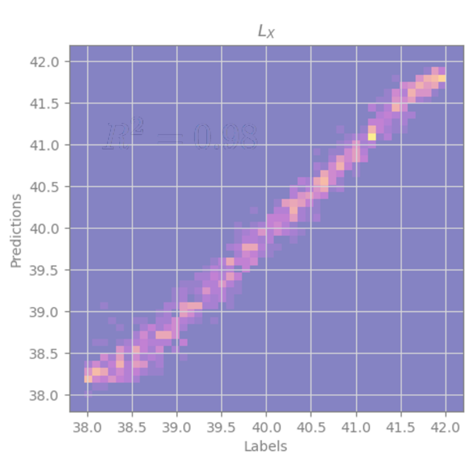


(✓)

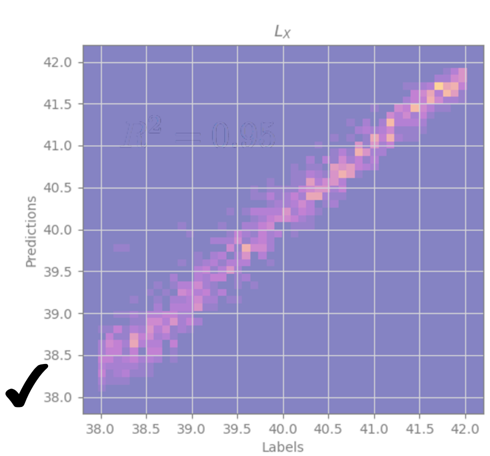
Mock -> Sim



✗



✓



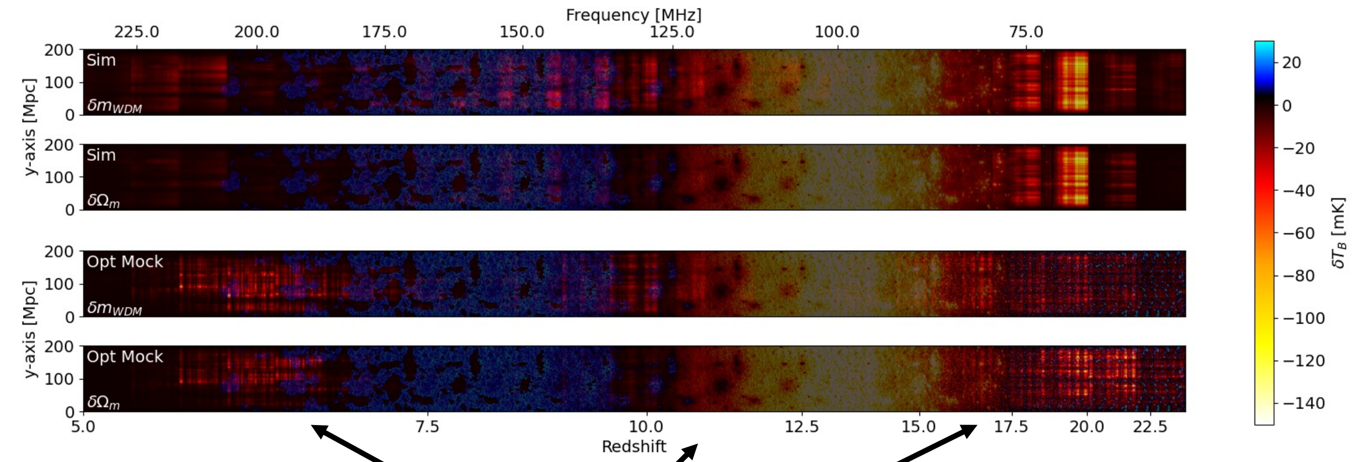
✓

✓Robust to foregrounds & systematics

2) Gradient-based saliency maps

Sim

Mock



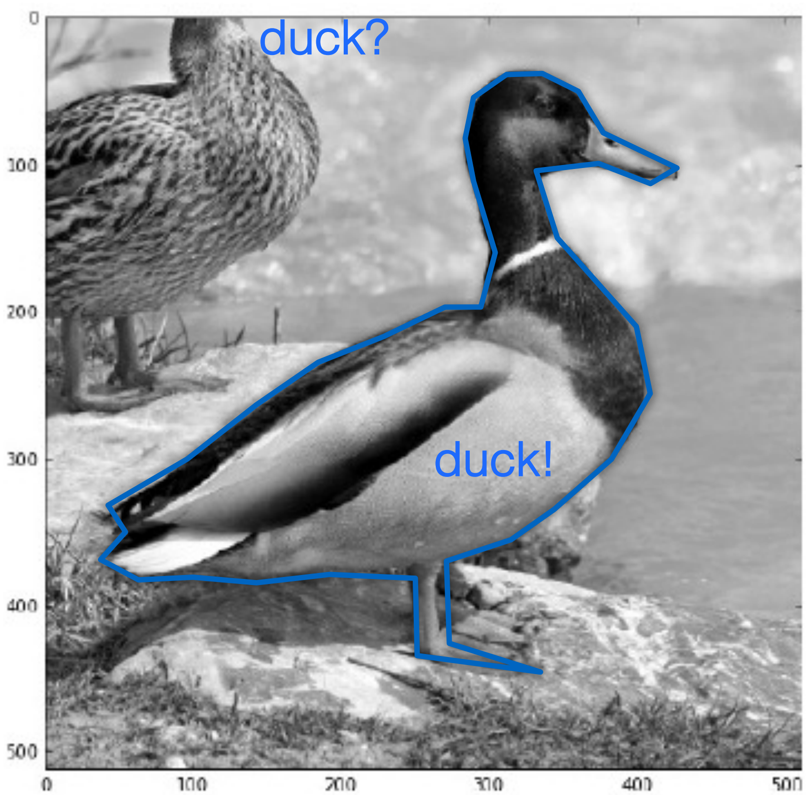
‘transitions’

also seen for Fisher forecasts:
Heneka & Amendola 2018
Liu, Heneka, Amendola 2020

arXiv:2201.07587

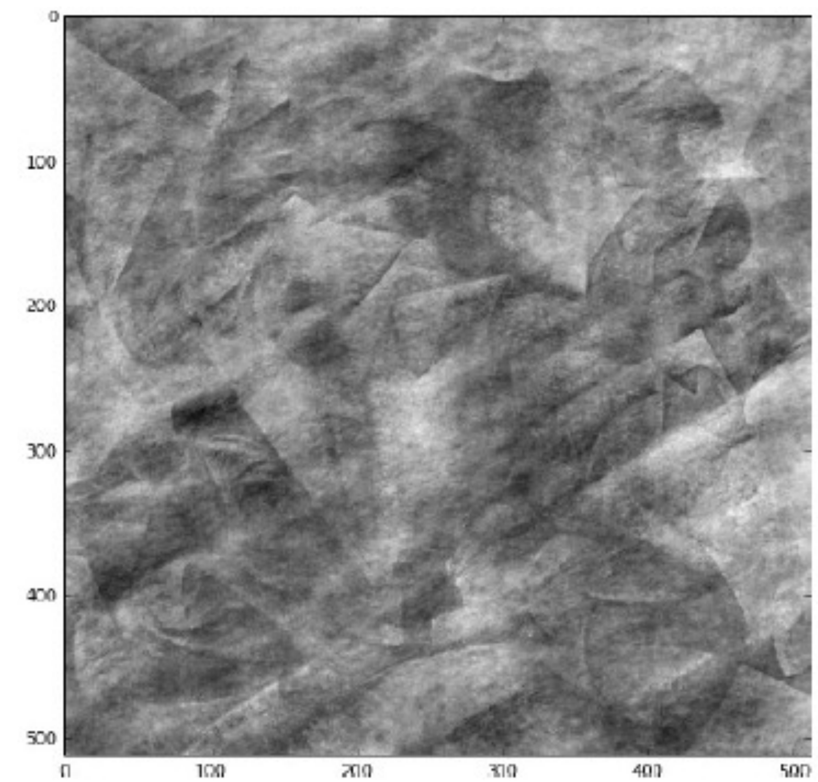
Why deep learning?

The duck example of (Non-)Gaussianity



The duck: highly non-Gaussian

randomise phases



Credit: G. Bernardi

The Gaussian duck

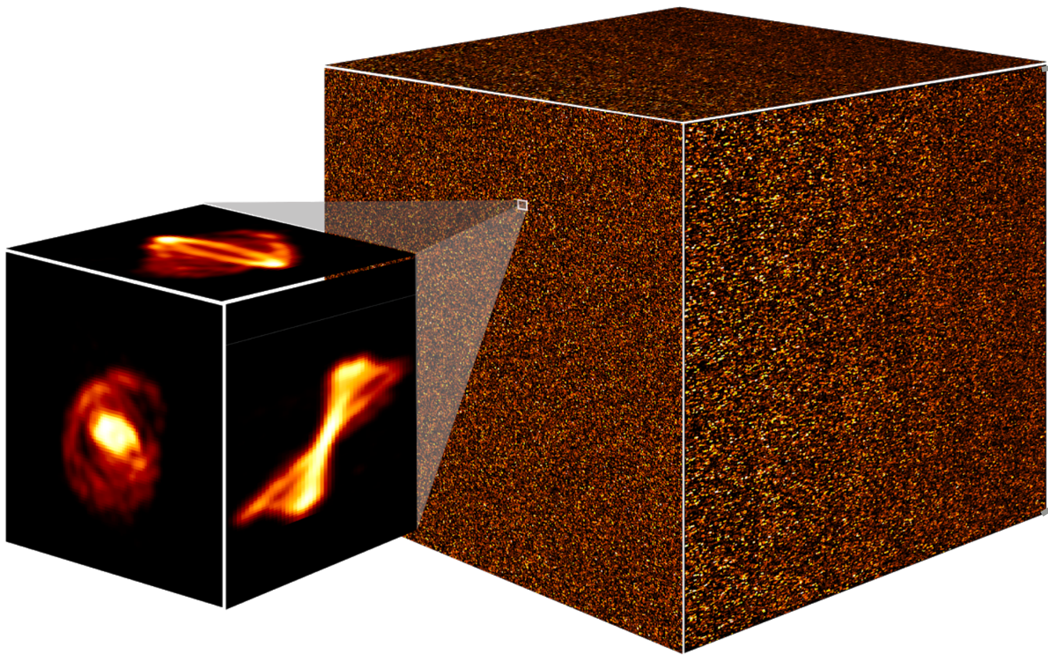
Same 2D
power spectrum

- Picks up non-Gaussian information
- Representation learning

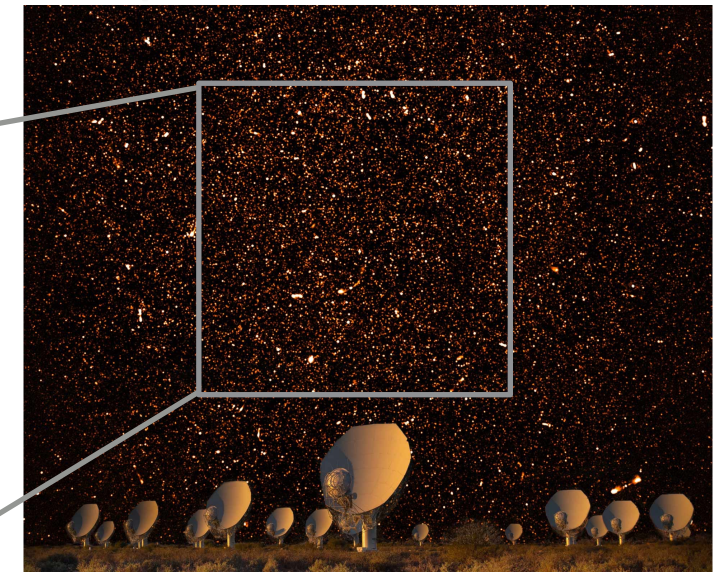
Applications:

1. Detect the duck (or galaxy, or signature)
2. Inference (what duck? what properties? what shapes?)

Detection in 3D: SKA Science Data Challenge



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



Composite MeerKAT dishes and observations.
Credit: South African Radio Astronomy Observatory (SARAO)

SKA -
The Square Kilometre Array
An international effort to build the world's largest radio telescope

Expected data rate in full operation:
1 TB/s

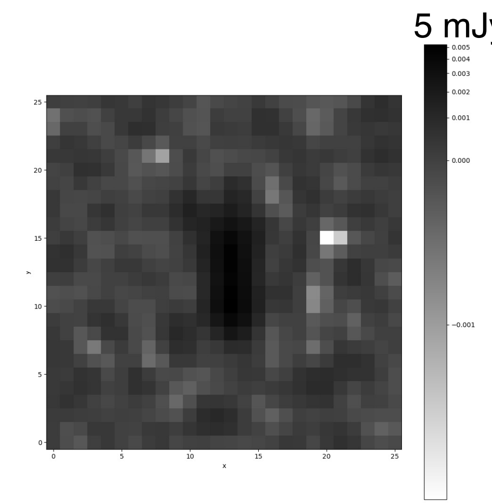
Key science goals include:
Galaxy Evolution, Reionisation, Cosmology, Astroparticles

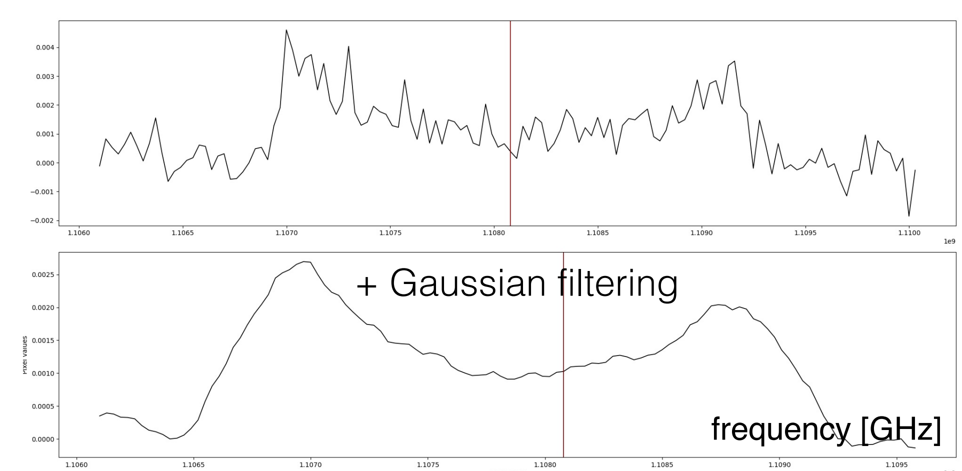
Goal is both source finding and characterisation
& test for new SKAO Regional Data Centers

The challenging HI sources:

- low S/N
- small spatial size
- systematics

The brightest HI source





Detection in 3D: SKA Science Data Challenge

Machine learning and deep learning come together?

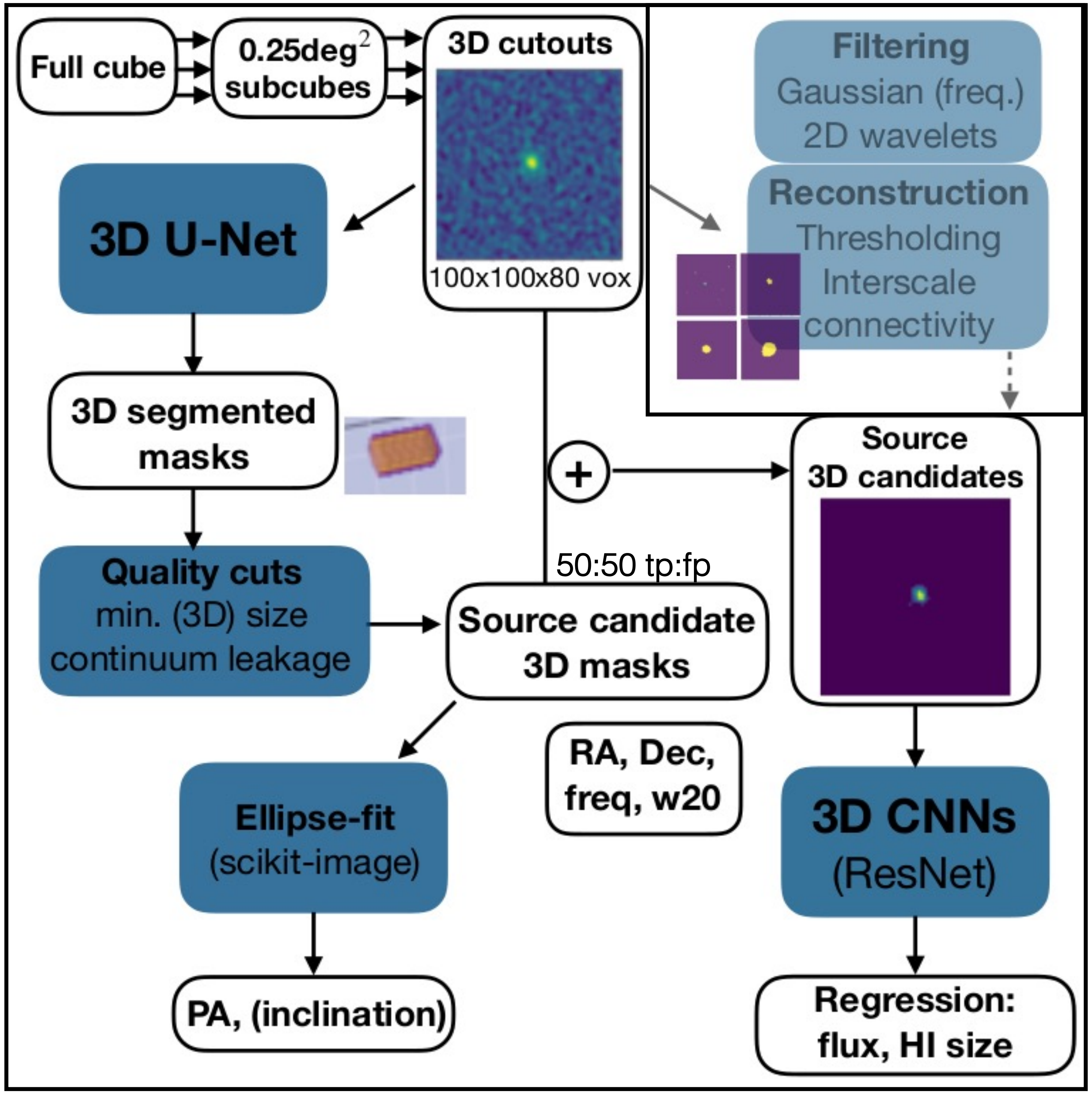
Team: Michelle delle Veneri, Andrew Soroka, Bernardo Fraga, Fedor Gobanov, Clecio de Bom, Alex Meshcheryakov

DL source detection & characterisation:
Best performing: full 3D approaches (U-Net type)
Trials: 2D/3D variants of U-Net, R-CNN, inception network

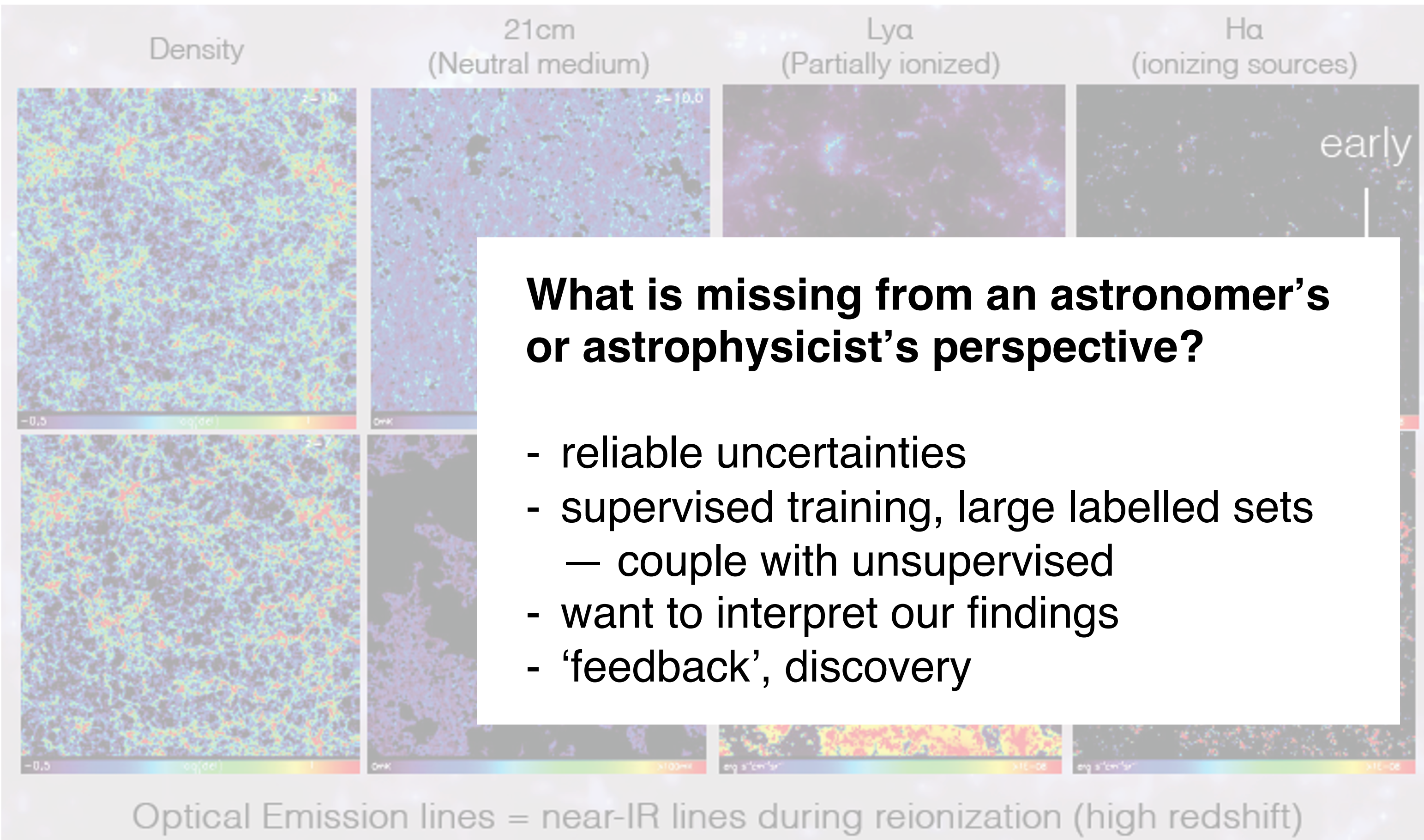
+ Trial source detection baseline
Wavelet denoising & Multi-scale model

Pitfalls & Take-aways:

- Pre-processing, noise model(s)
- High sparsity
- Choice of training set
  -> pre-training, let the network choose?
- Multi-step and/or ensemble decision



A scientific life-cycle of 3D (cosmic) inference

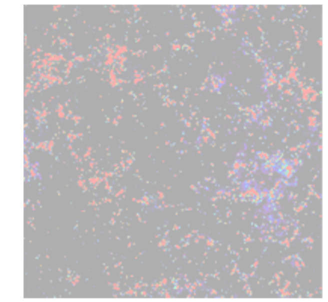


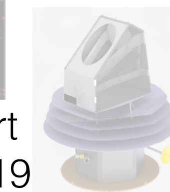
What is missing from an astronomer's or astrophysicist's perspective?

- reliable uncertainties
- supervised training, large labelled sets
  — couple with unsupervised
- want to interpret our findings
- 'feedback', discovery

+noise

denoise





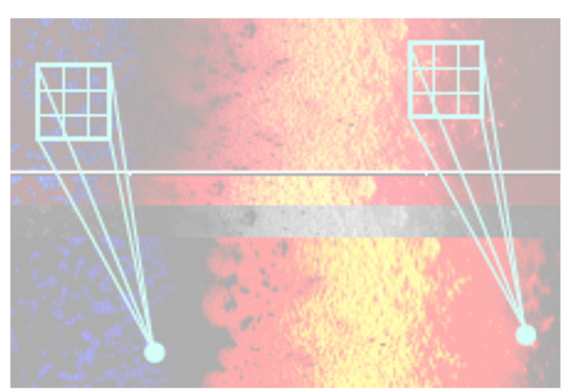
CDIM Science Report
Cooray, Heneka+ 2019

generate maps

observations

GAN, VAE,
U-Net, GNN

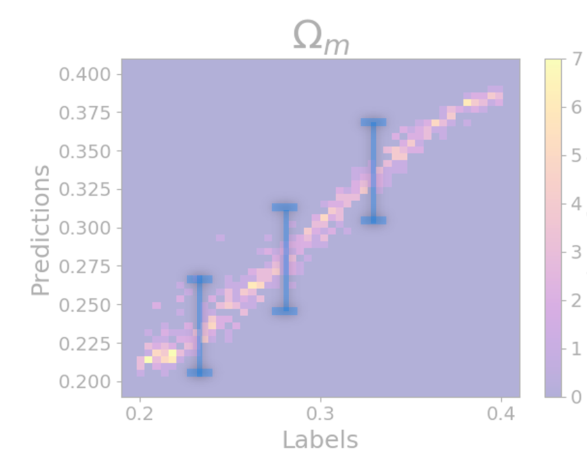
2) direct derivation
& detect



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

parameter studies

forecast + inference

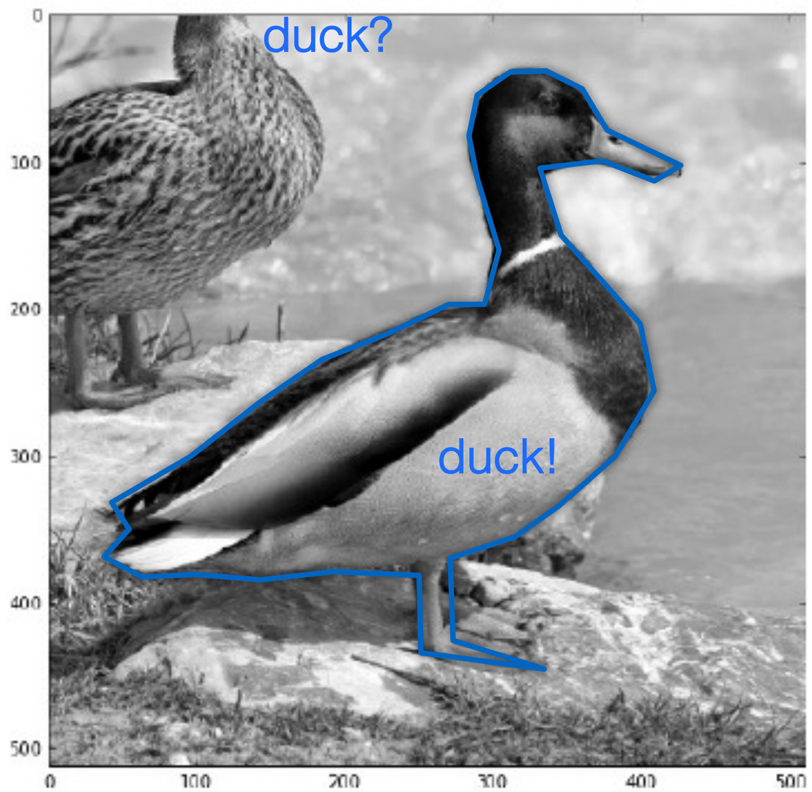


→BNNs, DELFI, ..

Gillet+, MNRAS 484 (2019)
Neutsch, Heneka,
Brüggen MNRAS 511 (2022)

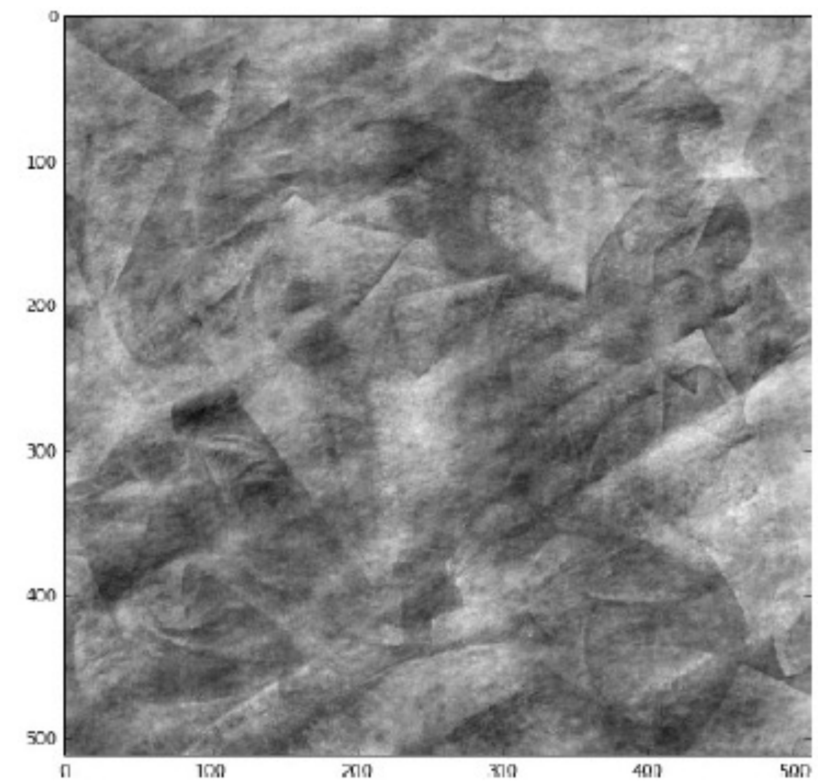
Why deep learning?

The duck example of (Non-)Gaussianity



The duck: highly non-Gaussian

randomise phases



Credit: G. Bernardi

The Gaussian duck

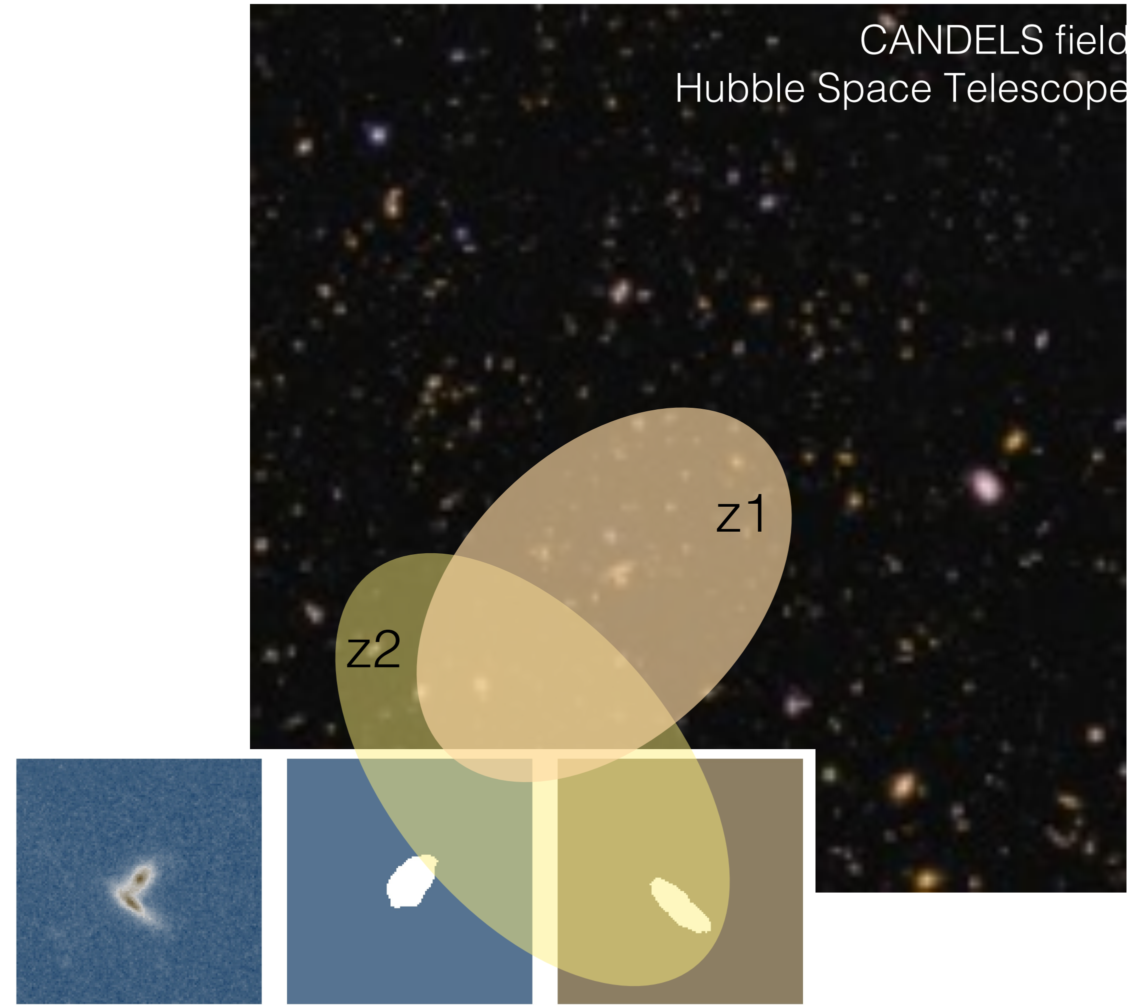
Same 2D
power spectrum

- Picks up non-Gaussian information
- Representation learning

Applications:

1. Detect the duck (or galaxy, or signature)
2. Inference (what duck? what properties? what shapes?)

Detection in 2D: The deblending problem



Deblending.. in Astrophysics

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

Galaxy morphology

Challenge: Galaxies are 'transparent'

Deblending.. in Biology

Lily Hu+ 2017

Goal: Separate e.g. DNA for for medical diagnostics, drug development, and biomedical research

Similar challenge:
Overlapping chromosomes are mostly transparent

Detection in 2D: The deblending problem

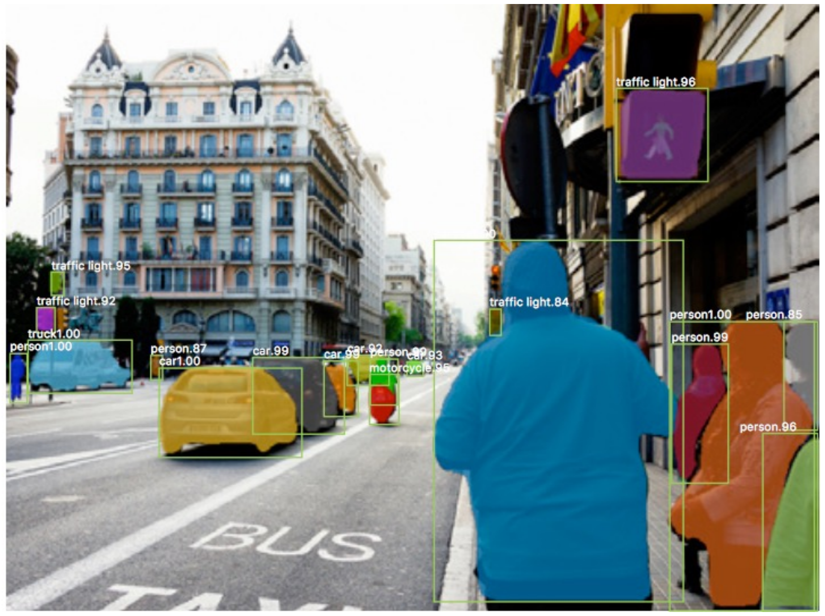
Deblending.. in Astrophysics

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

Galaxy morphology

Challenge: Galaxies are 'transparent'

.. for AI



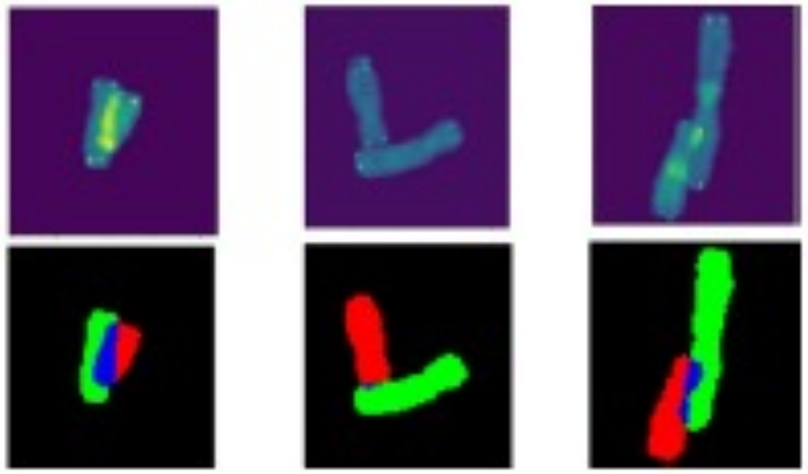
https://medium.com/@umerfarooq_26378/from-r-cnn-to-mask-r-cnn-d6367b196cfd

Another challenge: Object detection

Deblending.. in Biology

Goal: Separate e.g. DNA for for medical diagnostics, drug development, and biomedical research

Similar challenge:
Overlapping chromosomes are mostly transparent

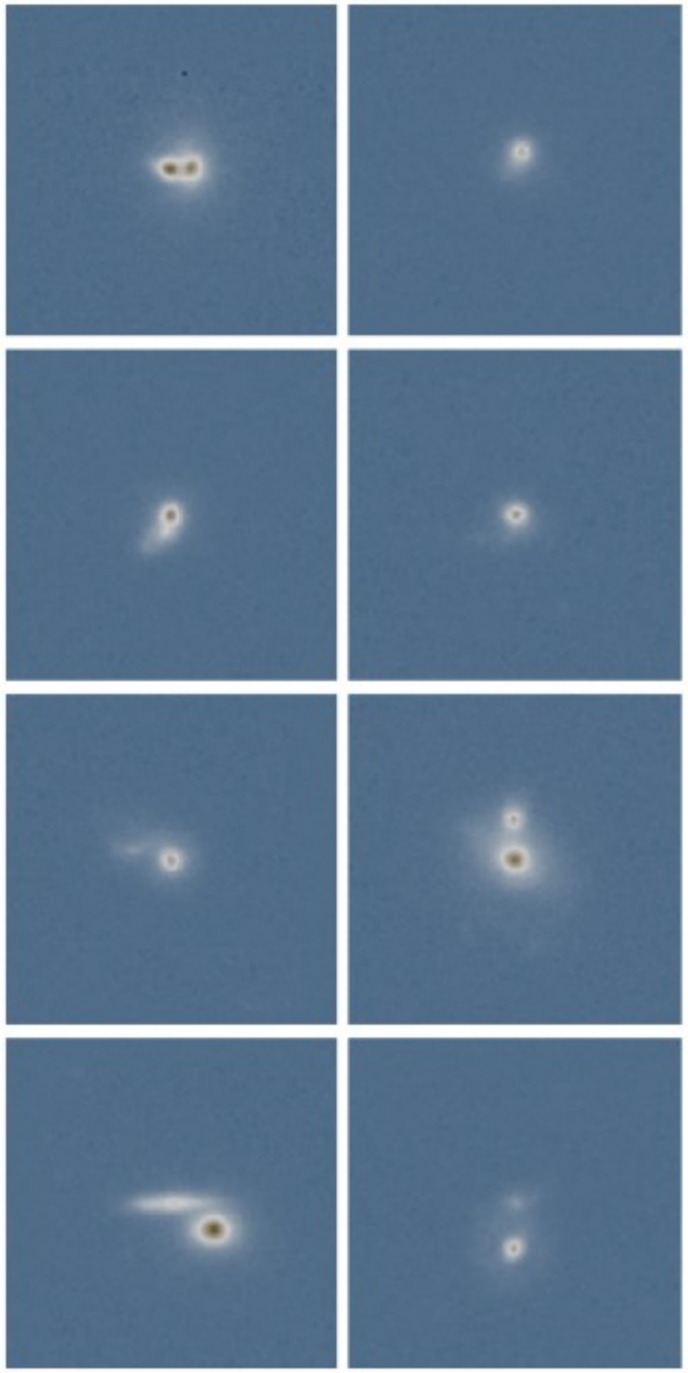


Lily Hu+ 2017

Detection in 2D: The deblending problem

COIN -
Cosmostatistics Initiative
A worldwide endeavour to create an interdisciplinary community around data-driven problems in Astronomy

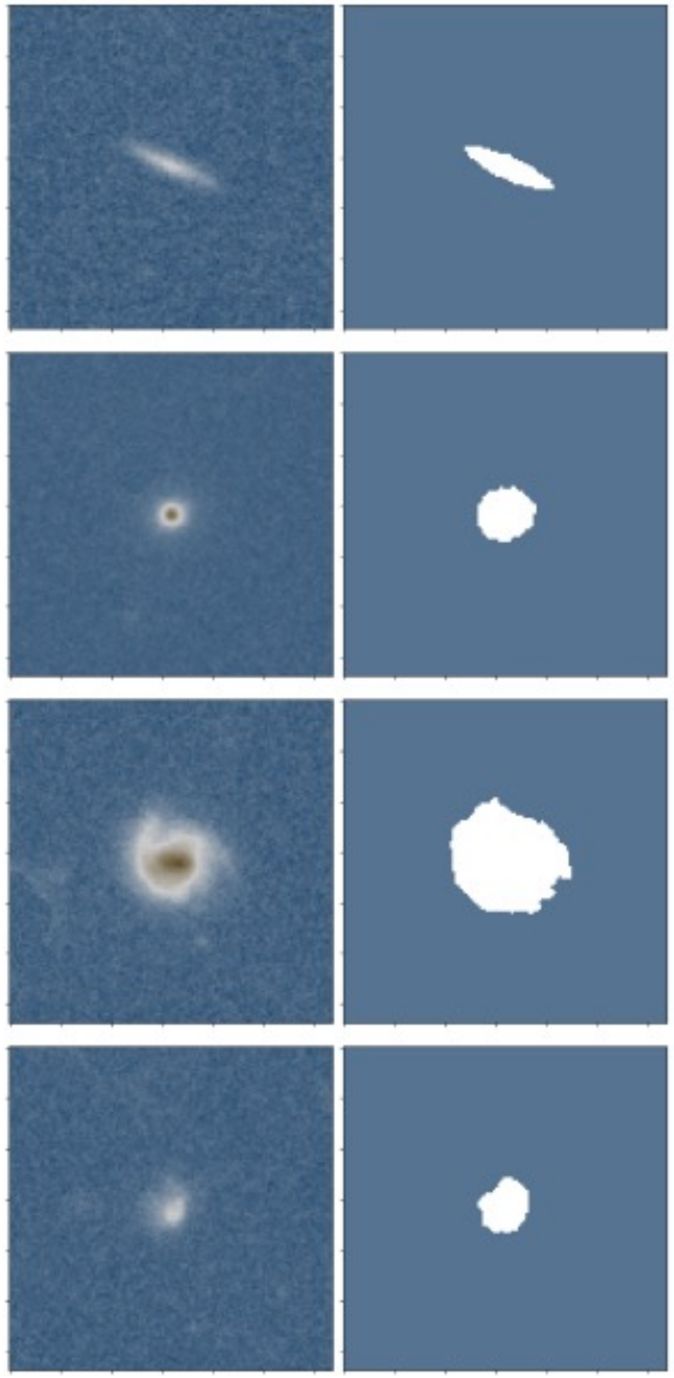
Goals for our method



Get the correct photometry

Derive masks

..and do so bias-free

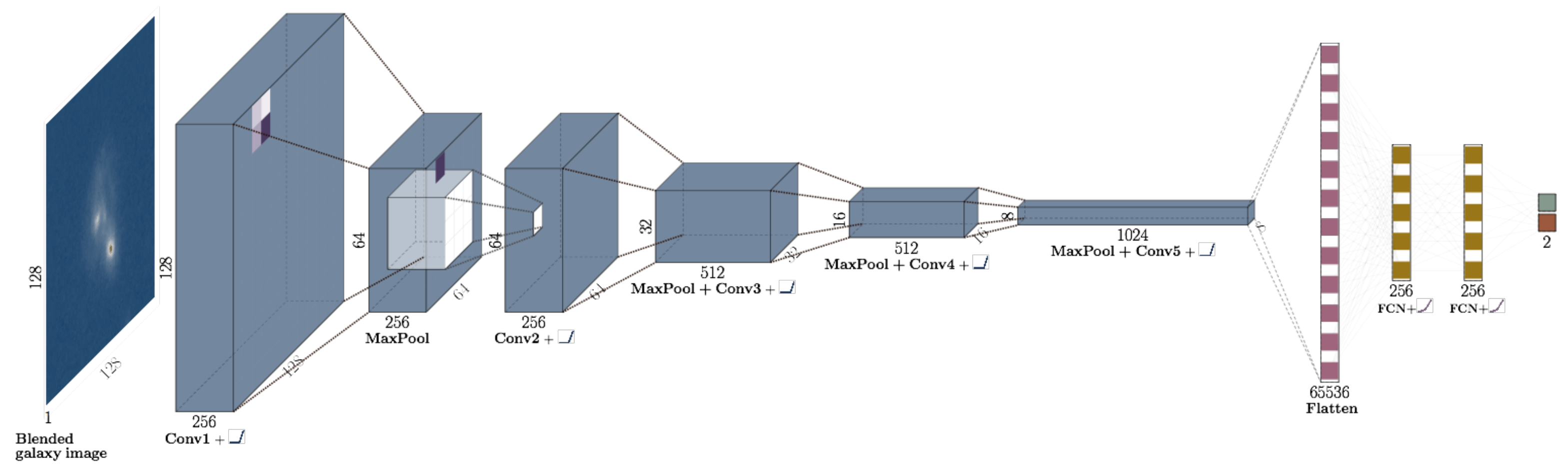


Boucaud, Huertas-Company, Heneka+ 20

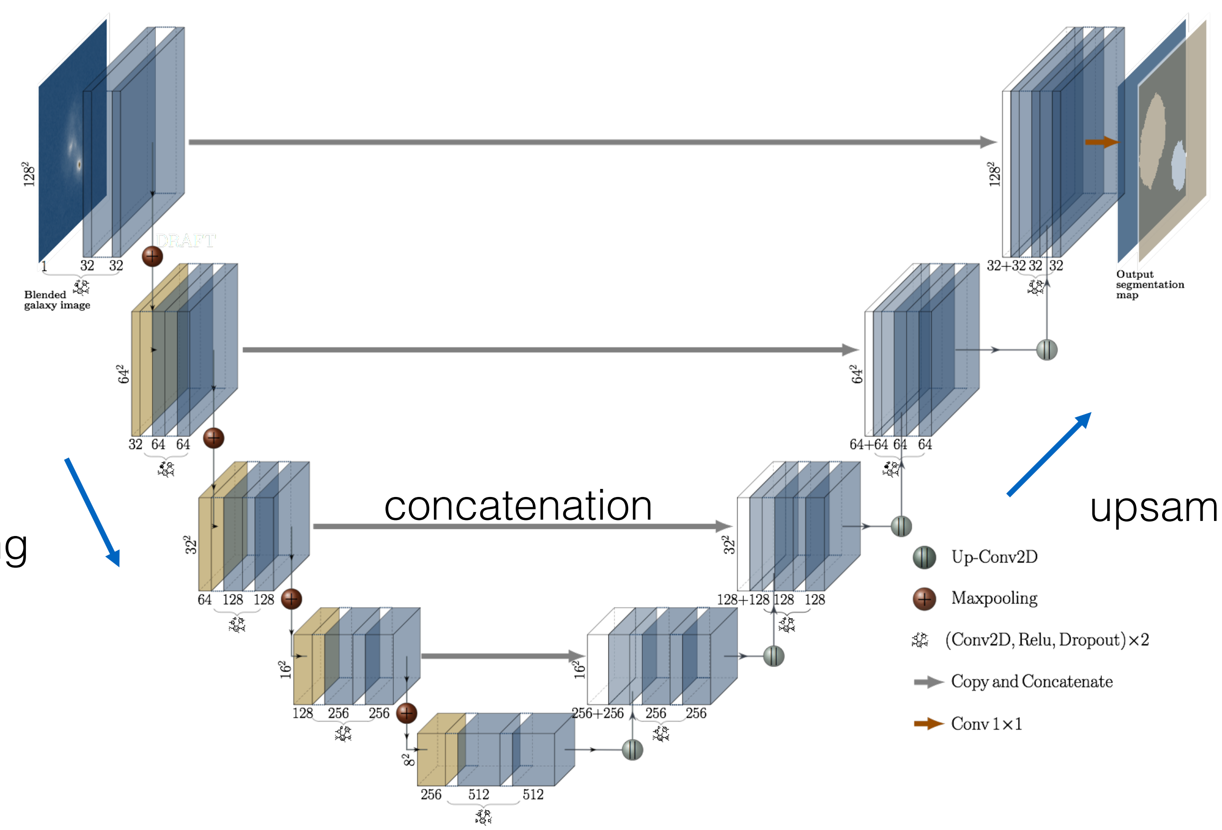
Detection in 2D: The deblending problem

For photometry, we let two architectures compete:

1) *blend2flux*
a CNN for photometry



2) *blend2mask2flux*
photometry + masks



downsampling

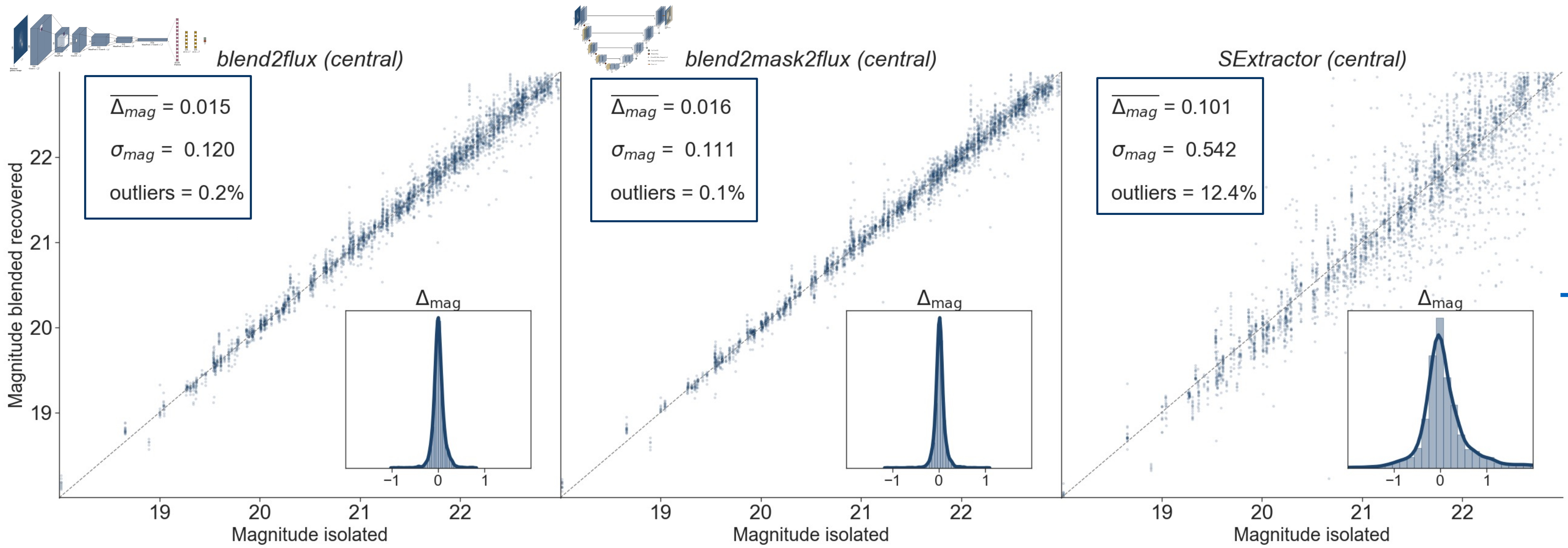
upsampling

+ a *blend2flux*-type net

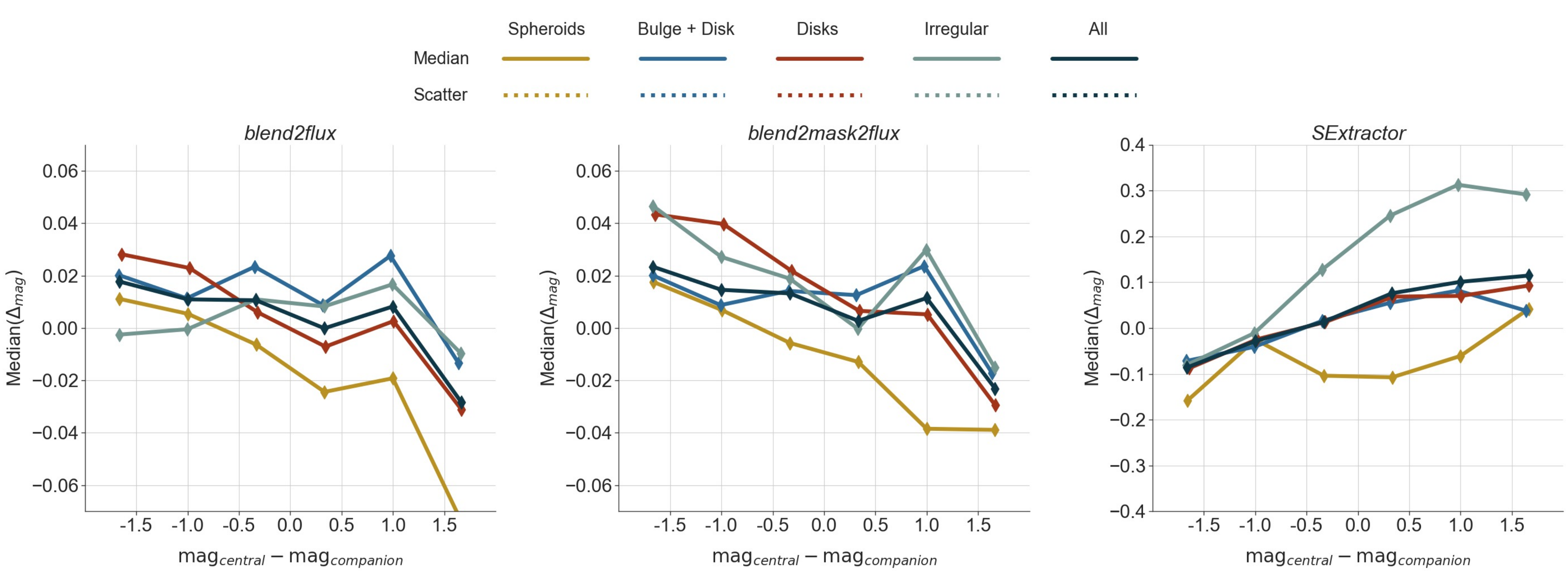
Boucaud, Huertas-Company, Heneka+ 20
github.com: coindeblend

Detection in 2D: The deblending problem

Machine learning outperforms traditional baselines



uncertainty 0.1mag << conventional

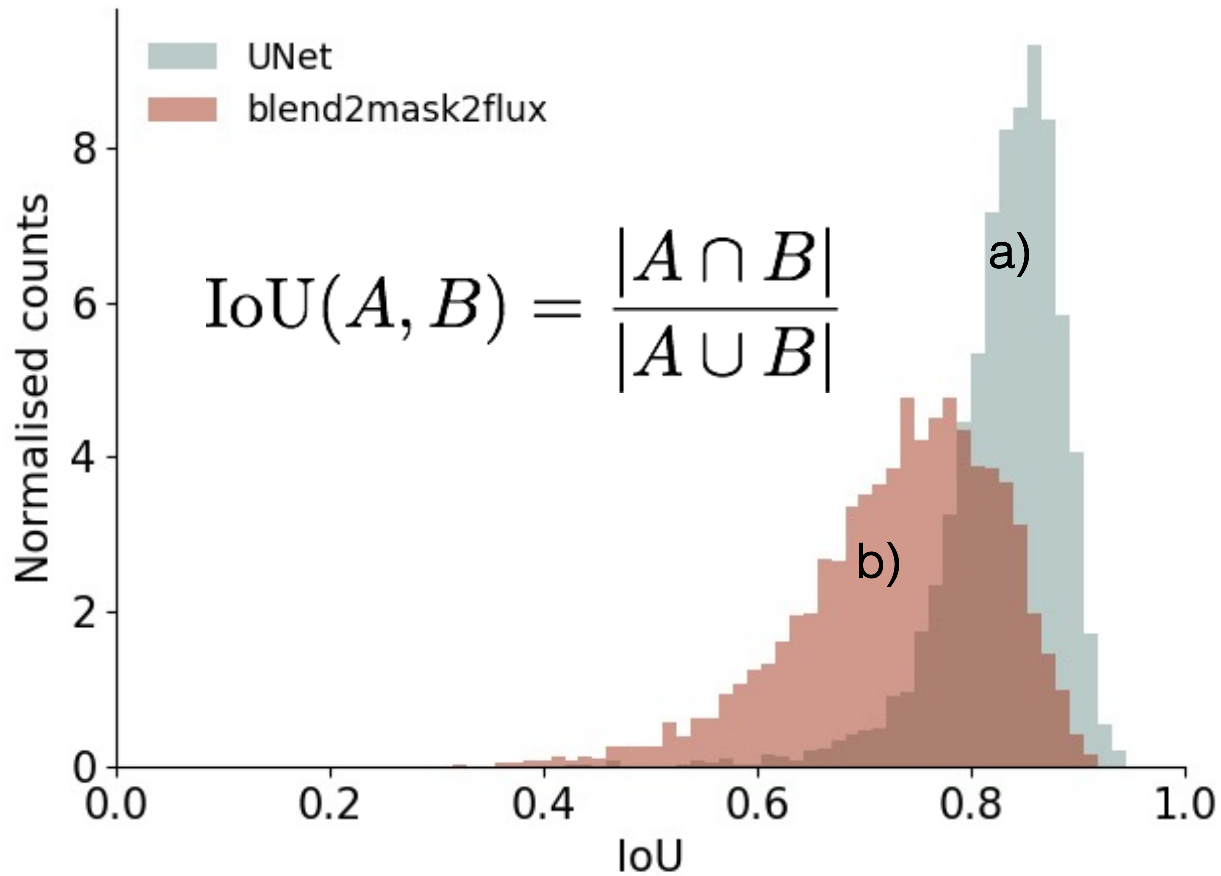


small bias

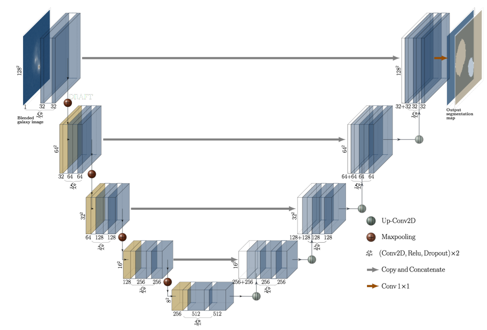
Boucaud, Huertas-Company, Heneka+ 20
github.com: coindeblend

Detection in 2D: The deblending problem

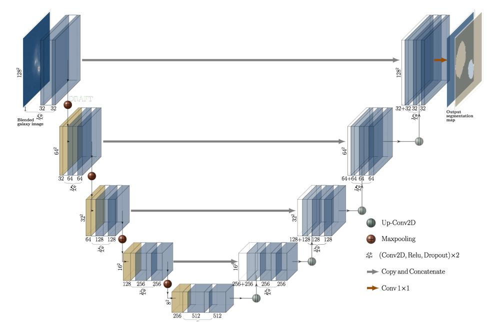
Optimal design is goal-dependent



a) *U-net*
masks, no photometry

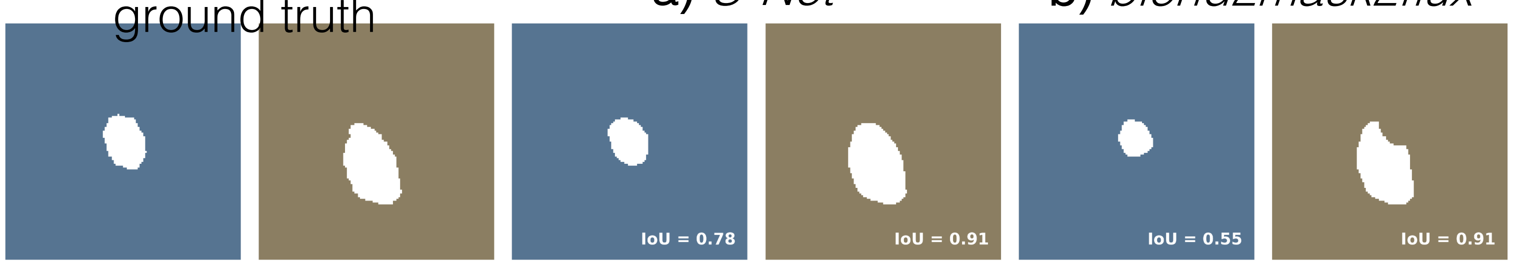


b) *blend2mask2flux*
photometry + masks



+ a CNN type *blend2flux*

ground truth   a) *U-Net*   b) *blend2mask2flux*



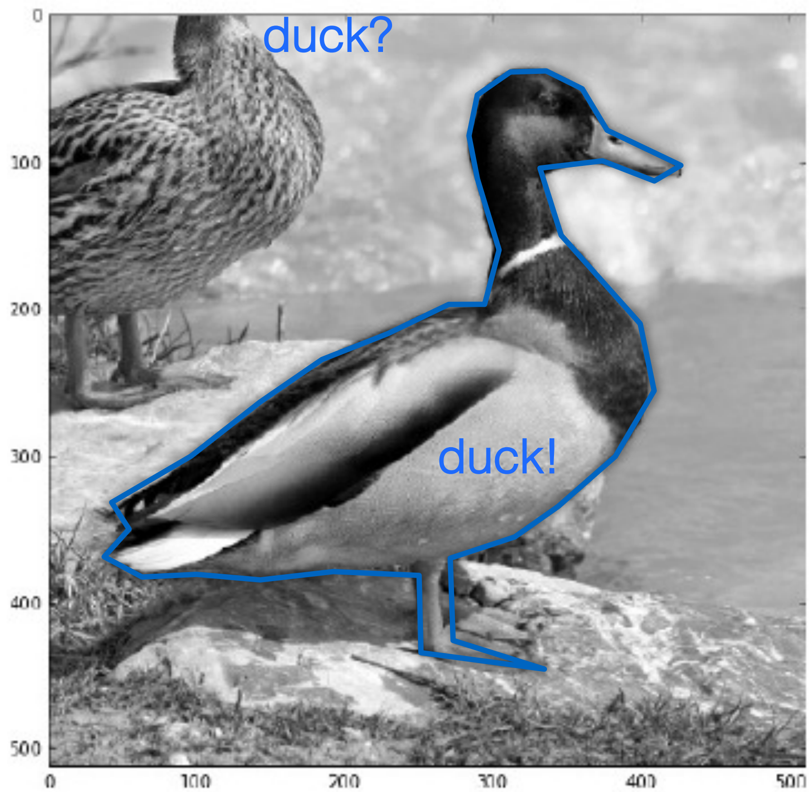
→ Dispersion broadens when optimised for photometry

→ Tailor to your research question

Boucaud, Huertas-Company, Heneka+ 20

Why deep learning?

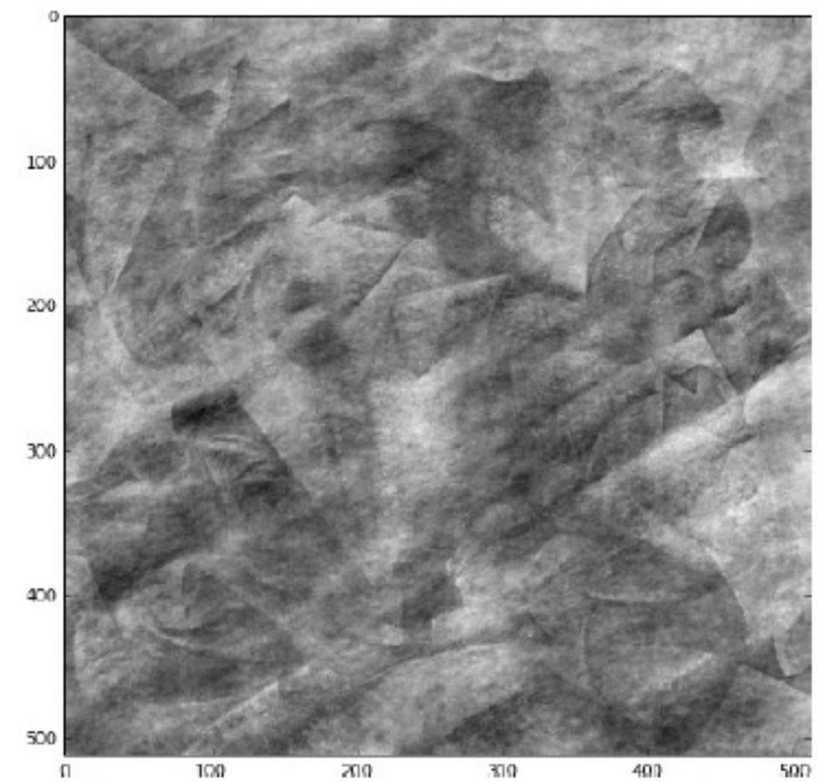
The duck example of (Non-)Gaussianity



The duck: highly non-Gaussian

randomise phases

Same 2D
power spectrum



Credit: G. Bernardi

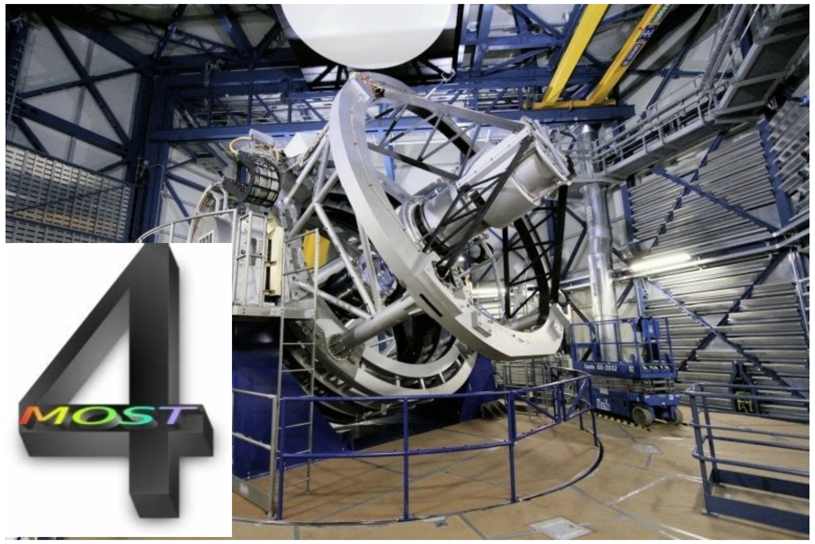
The Gaussian duck

- Picks up non-Gaussian information
- Representation learning

Applications:

1. ~~Detect~~ the duck (or galaxy, or signature)
   duck! classify.

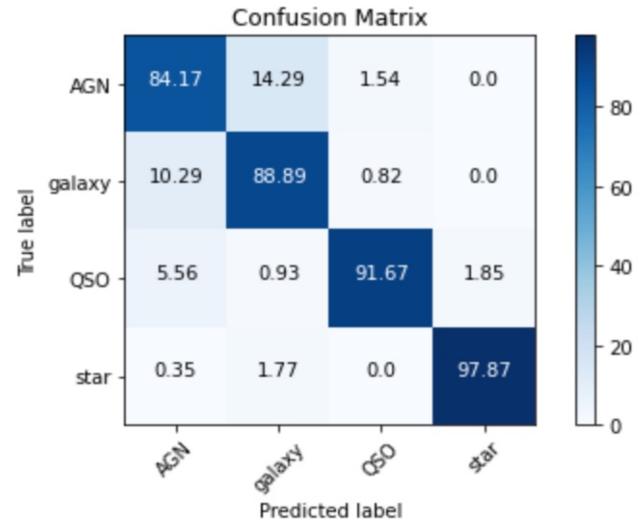
Classification in 1D: Spectroscopy



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

Building a classifier for 4MOST

- 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



Classification tasks:

- **Basic target classification.**
  Classes: star, galaxy, AGN, unknown. → Probabilistic multiclassifier
  also: lowres vs. highres
  (low S/N vs. high S.N)

- **Galactic source classification.**
  Sub-classes matching galactic pipeline
  (as of now) FGKM, OBA, WD sub-pipelines
  supplement to e.g. Gaia metadata based decision → Probabilistic multiclassifier II
  (sub-classes)

- **Extra-galactic.**

- **Feedback on a) targets, b) 'unknown' class** →
  a) match with expectation
  b) clustering, dimensionality reduction

  Currently set-up in galactic:
  4MOST explorer t-SNE
  (Gregor Traven, Gal Matijevic)
  arXiv: 1612.02242

Back to tutorial slides