Machine learning in Astrophysics

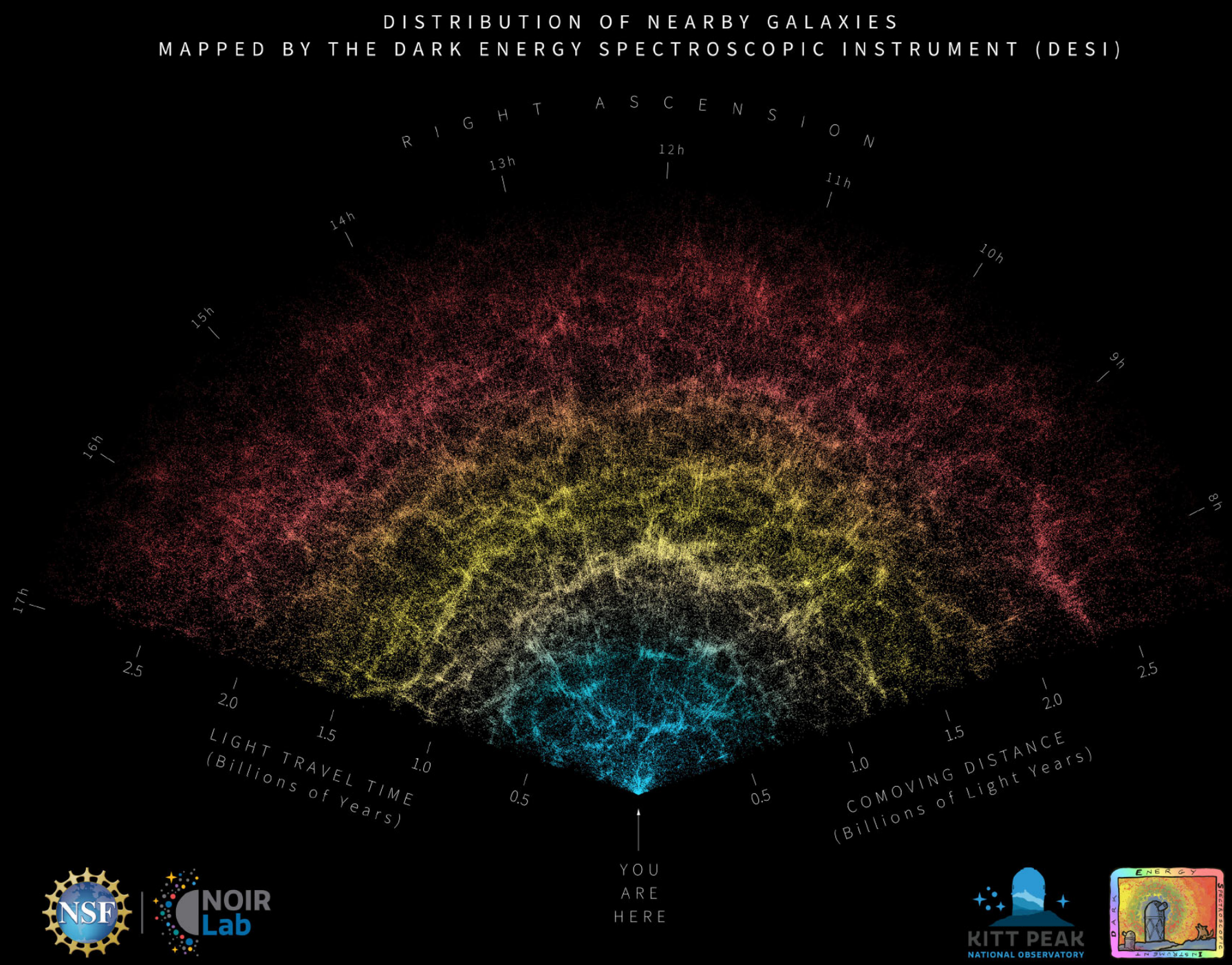
Dr David Harvey, Laboratoire d'Astrophysique

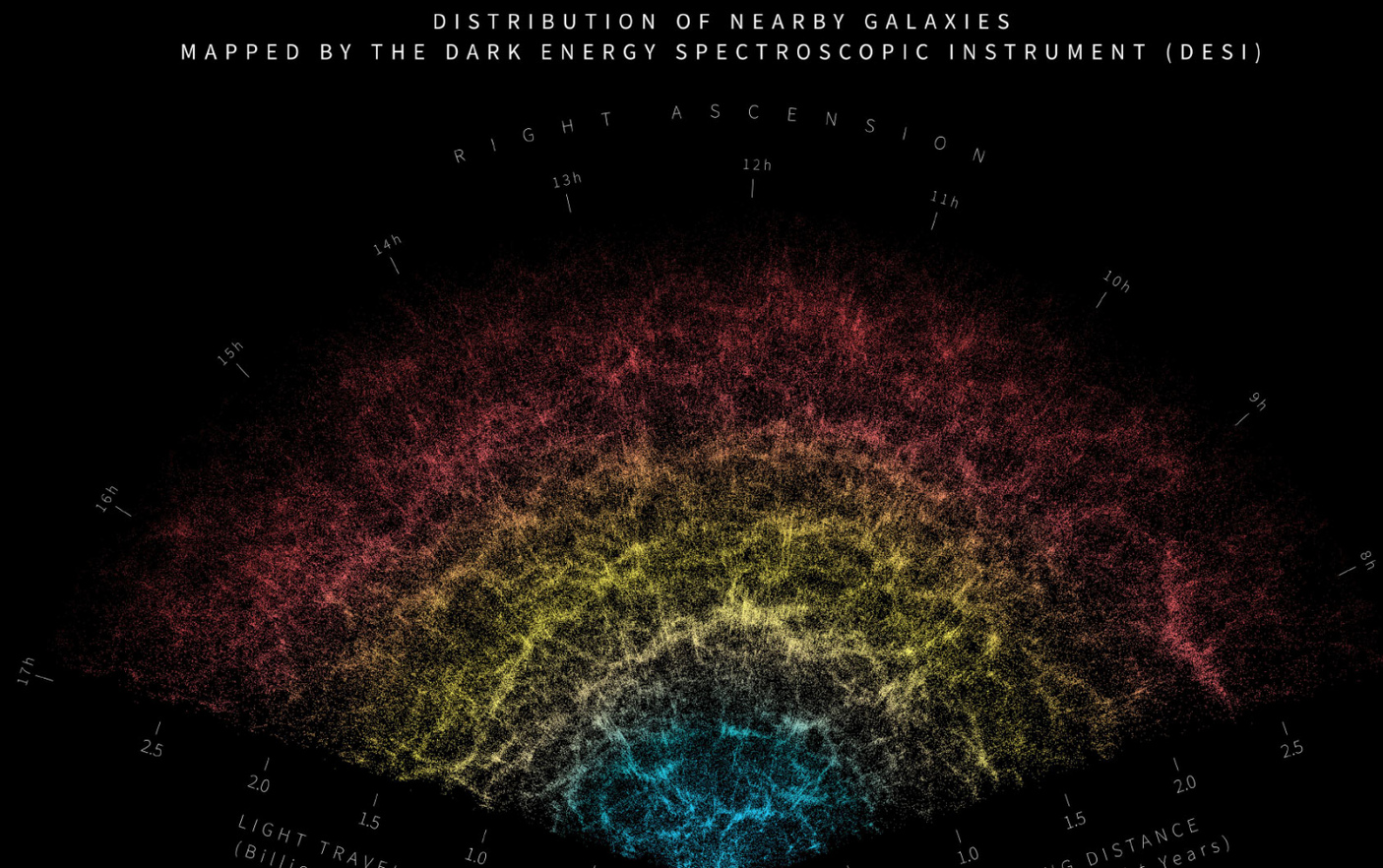
EPFL

A BIT ABOUT ME

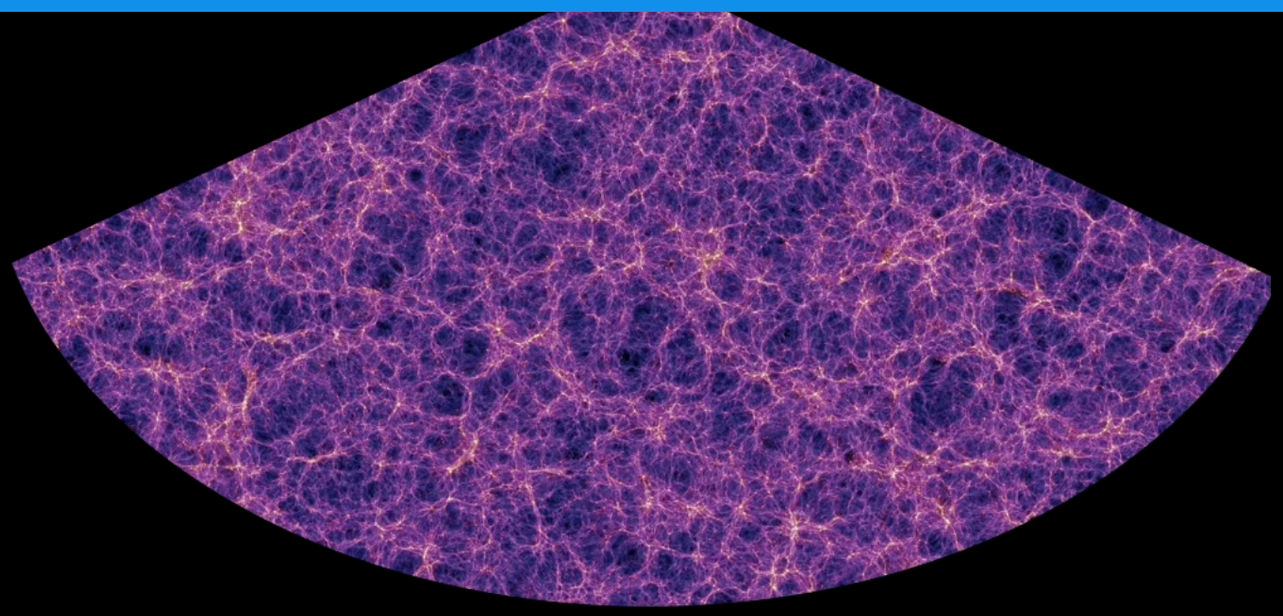
- PhD in Astrophysics from Edinburgh, UK
- Worked at Kaggle developing machine learning competitions, this included
    - Observing dark worlds
    - Galaxy Zoo
    - Titanic
- Post-doc Switzerland
- Fellowship Leiden, NL
- Machine Learning consultant at Terres des Hommes
- New position at EPFL
- Machine Learning consultant at Prophy
- Machine Learning consultant at TruthEngine

Astrophysics differs from almost any science



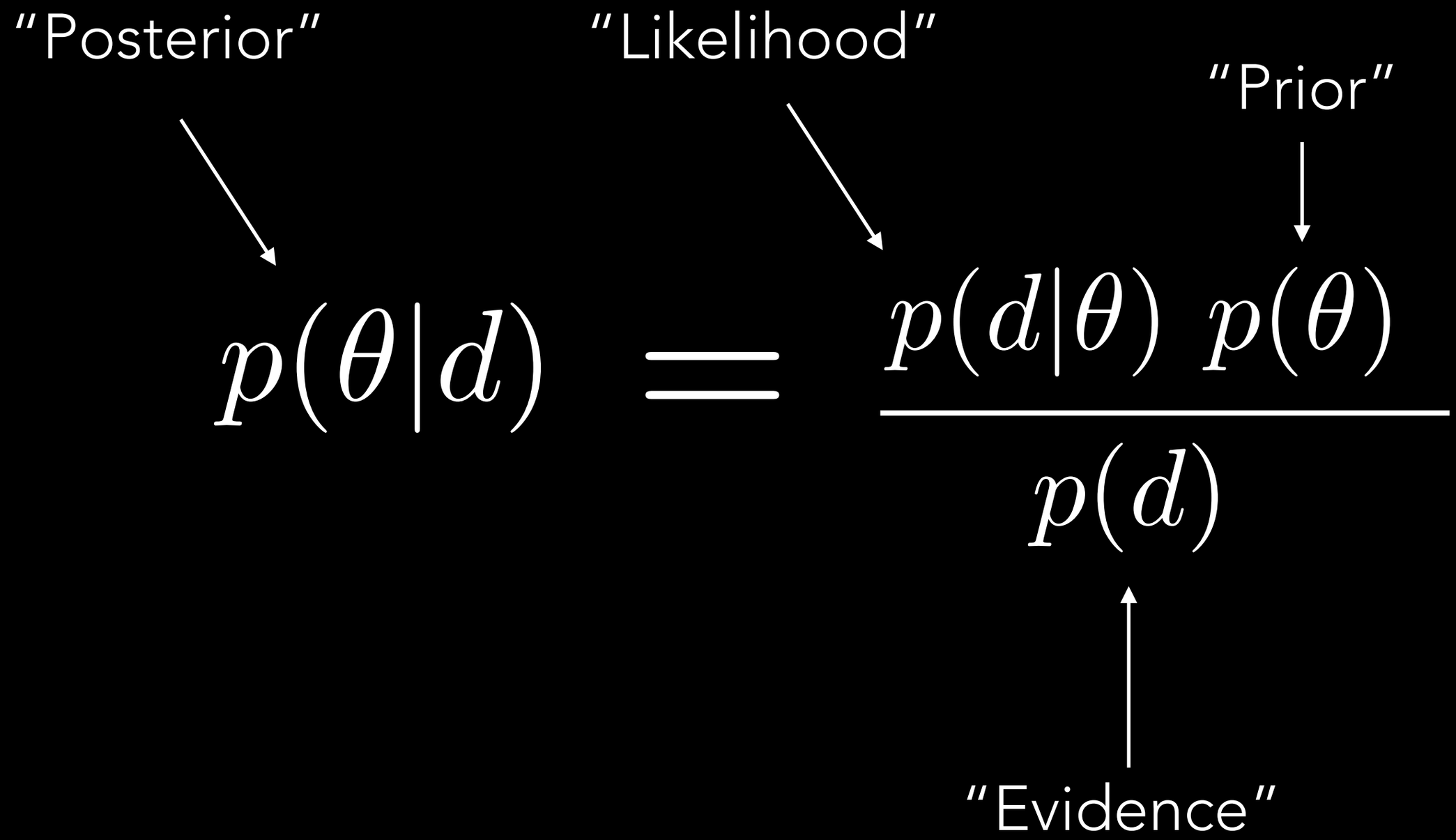


What is the probability of our model of the Universe given what we observe?



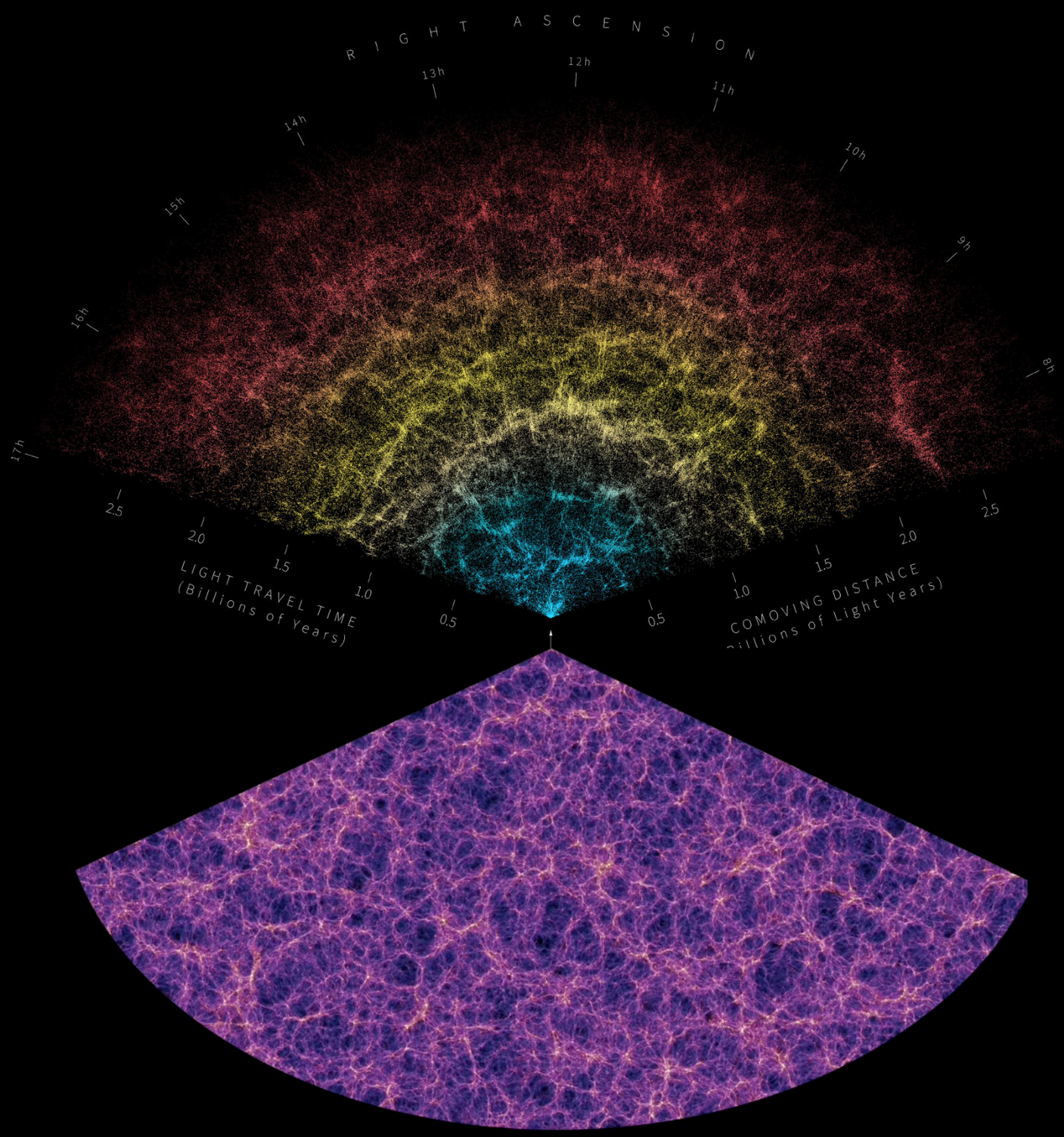
Distribution of matter as seen by the millennium simulation

Bayes theorem as a way to compare what we model to what we see.



Classical astrophysical inference

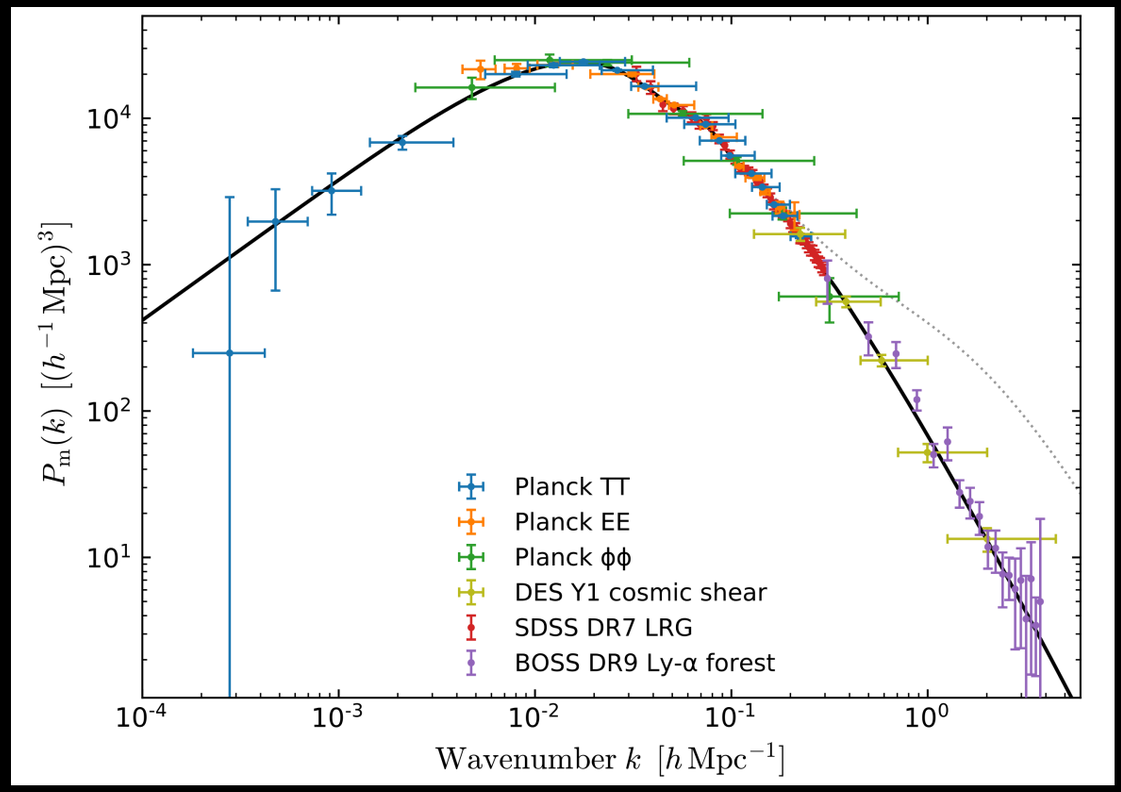
d



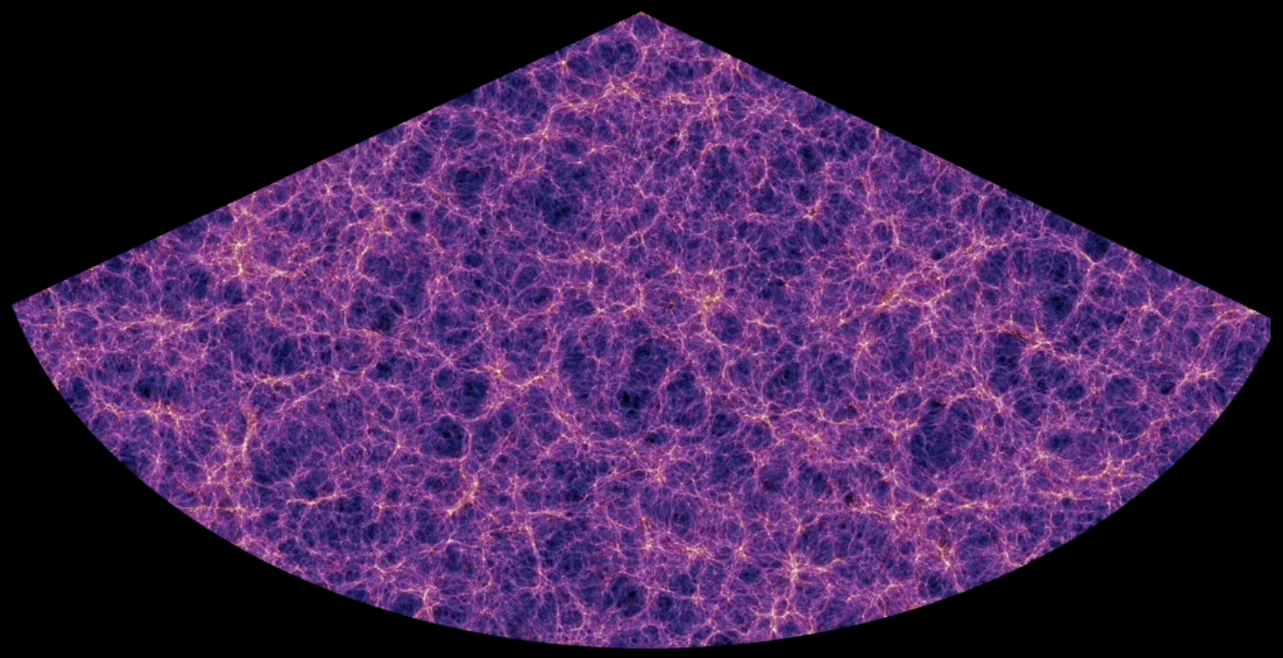
θ

Distribution of matter as seen by the millennium simulation

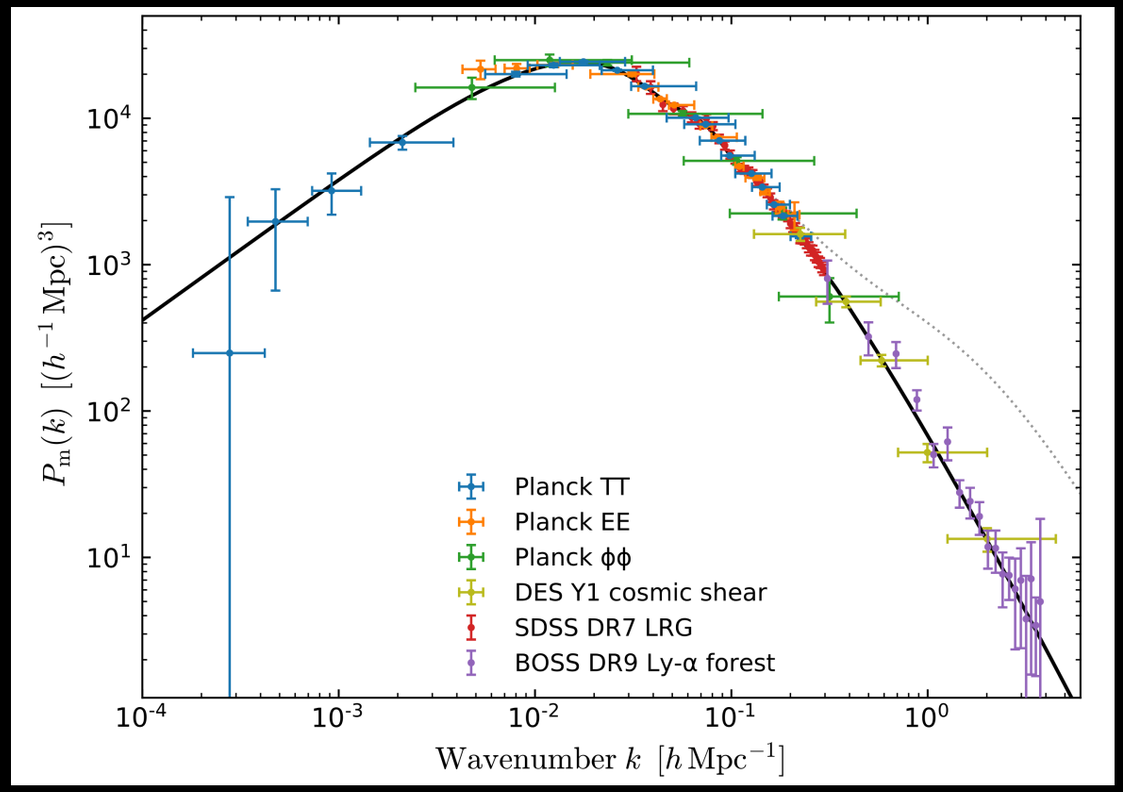
d


$P_{\rm m}(k)$ $[(h^{-1}\,{\rm Mpc})^3]$
10^4
10^3
10^2
10^1
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
Wavenumber k $[h\,{\rm Mpc}^{-1}]$
Planck TT
Planck EE
Planck ϕϕ
DES Y1 cosmic shear
SDSS DR7 LRG
BOSS DR9 Ly-α forest

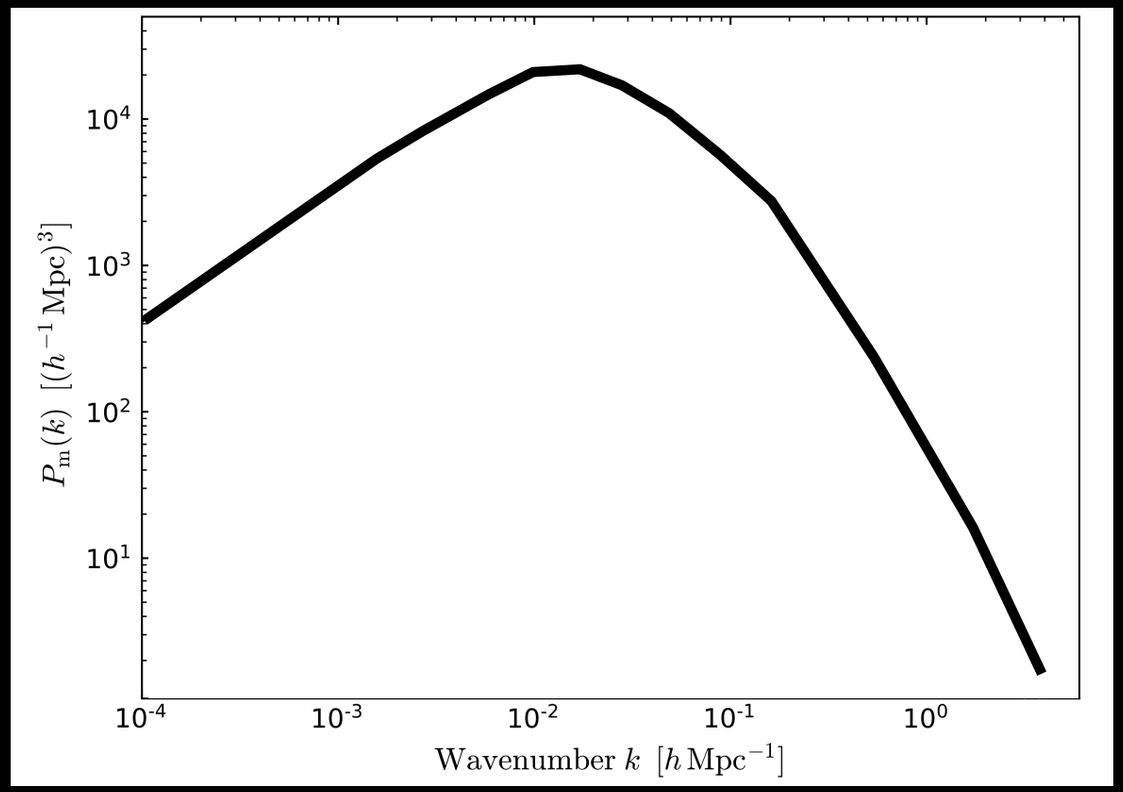
θ



d

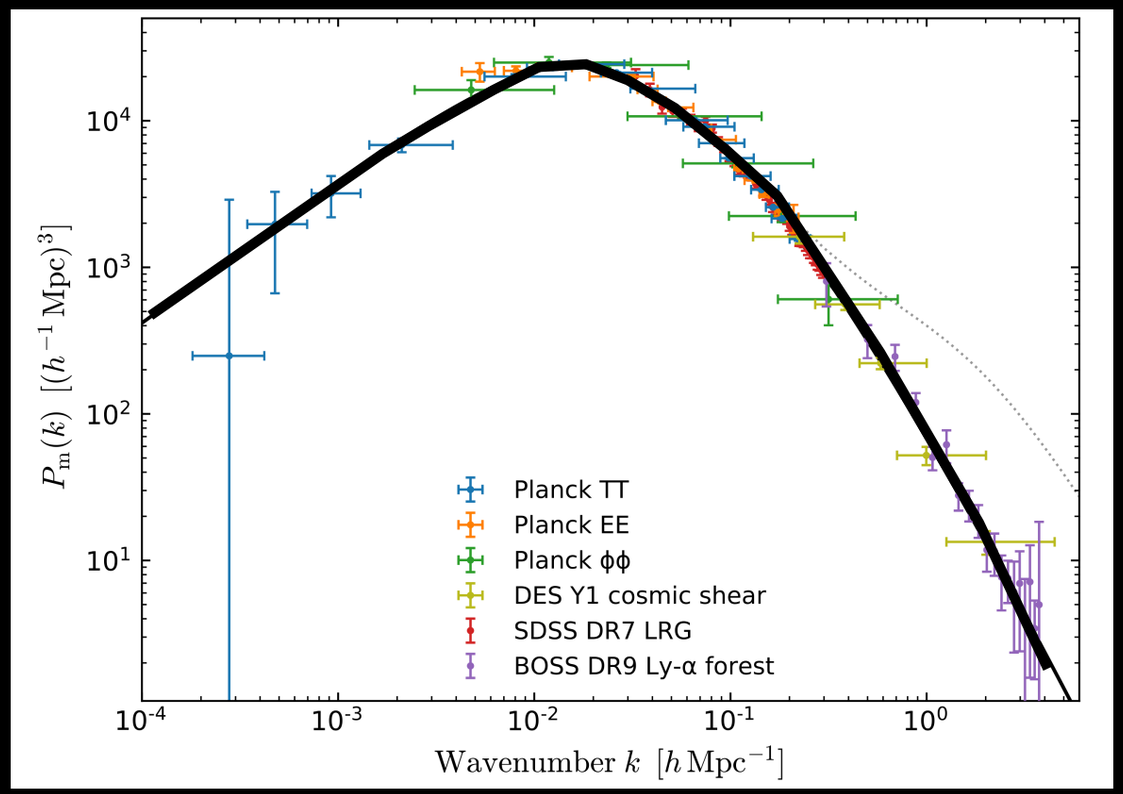

$P_{\rm m}(k)$ $[(h^{-1}\,{\rm Mpc})^3]$
10^4
10^3
10^2
10^1
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
Wavenumber k $[h\,{\rm Mpc}^{-1}]$
Planck TT
Planck EE
Planck ϕϕ
DES Y1 cosmic shear
SDSS DR7 LRG
BOSS DR9 Ly-α forest

θ


$P_{\rm m}(k)$ $[(h^{-1}\,{\rm Mpc})^3]$
10^4
10^3
10^2
10^1
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
Wavenumber k $[h\,{\rm Mpc}^{-1}]$

Compare the likelihood of the model given the data to find the best fitting model

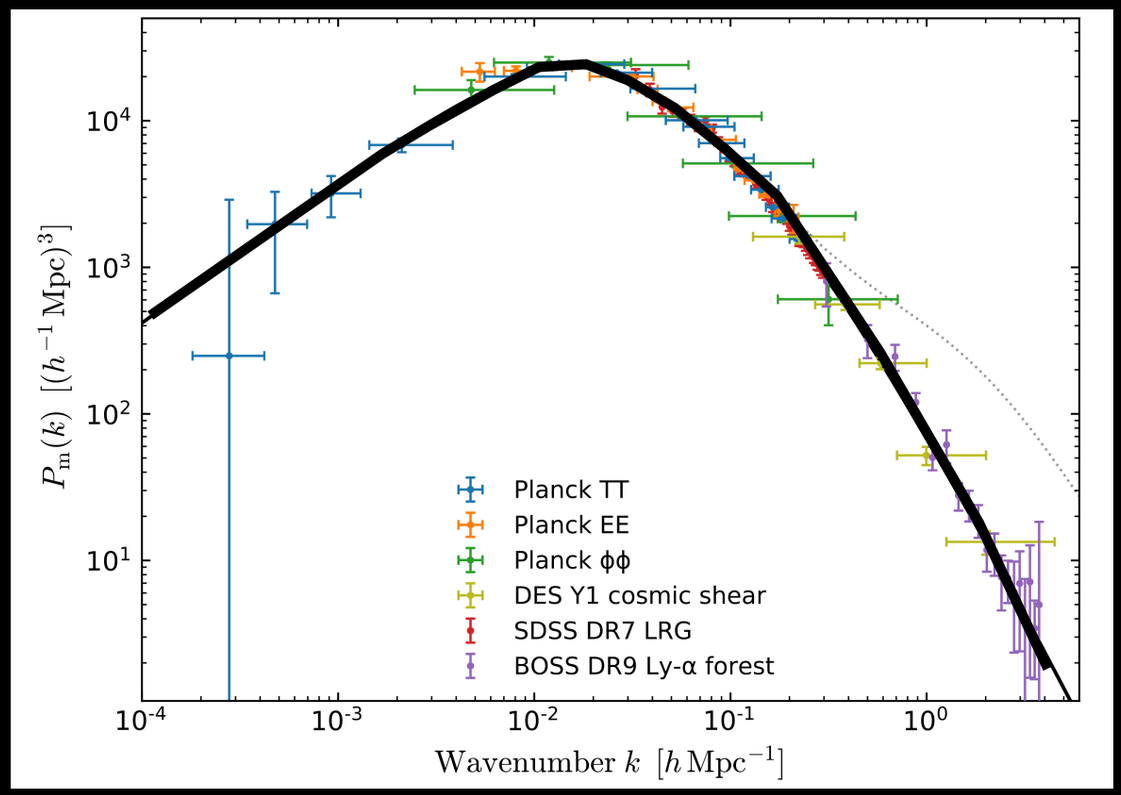
d



θ

Compare the likelihood of the model given the data to find the best fitting model

d



θ

Assuming an analytical model
Assuming a Gaussian likelihood function.
With covariances from theory or simulations

Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

- The number of posterior calls is normally very large.

- Likelihoods are not Gaussian.

- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

- The number of posterior calls is normally very large.

- Likelihoods are not Gaussian.

- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

- The number of posterior calls is normally very large.

- Likelihoods are not Gaussian.

- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

- The number of posterior calls is normally very large.

- Likelihoods are not Gaussian.

- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

- The number of posterior calls is normally very large.

- Likelihoods are not Gaussian.

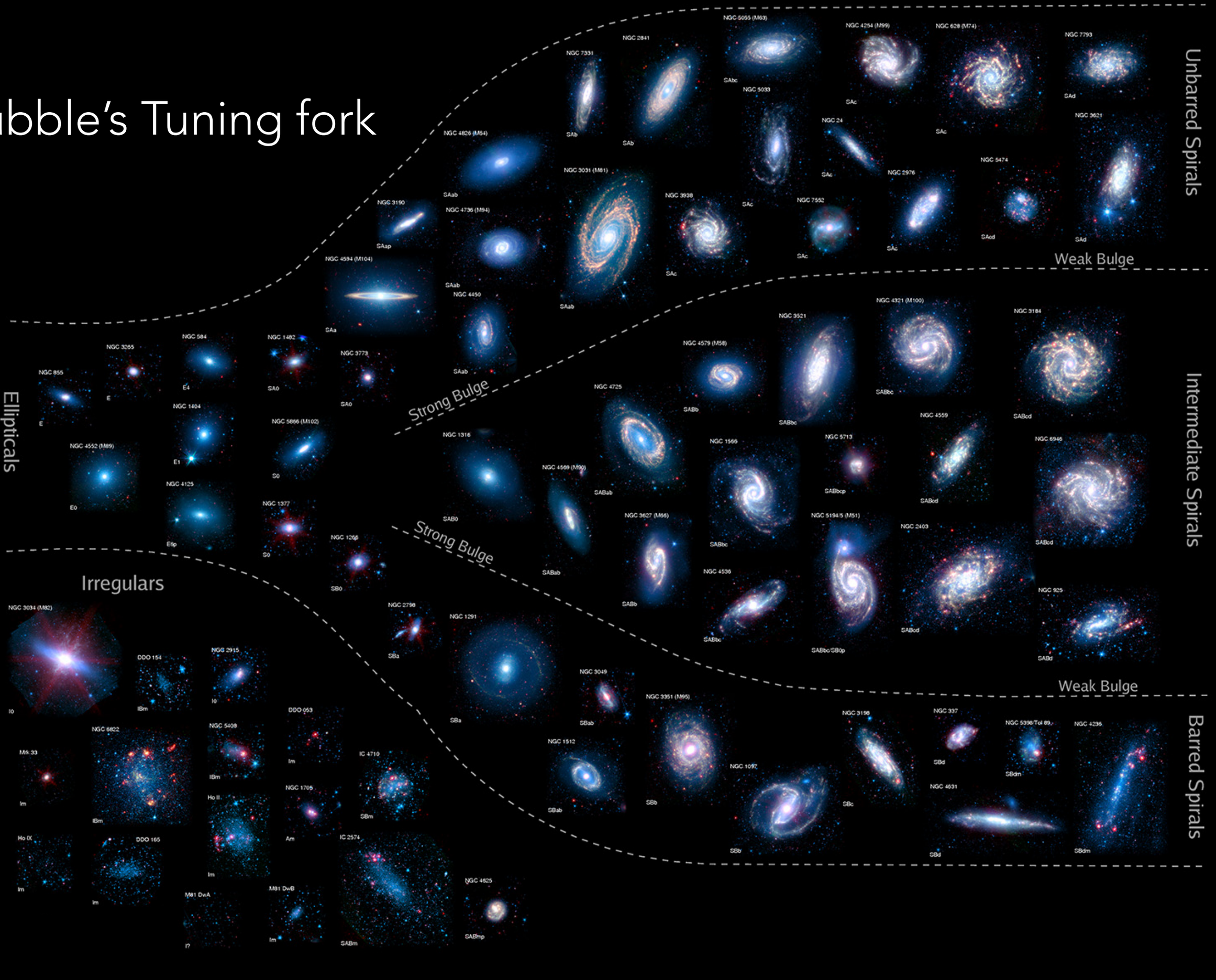
- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

CLASSICAL INFERENCE REQUIRES ANALYTICAL MODELS THAT CAN BE PASSED IN TO A MCMC WITH MANY CALLS TO A LIKELIHOOD FUNCTION

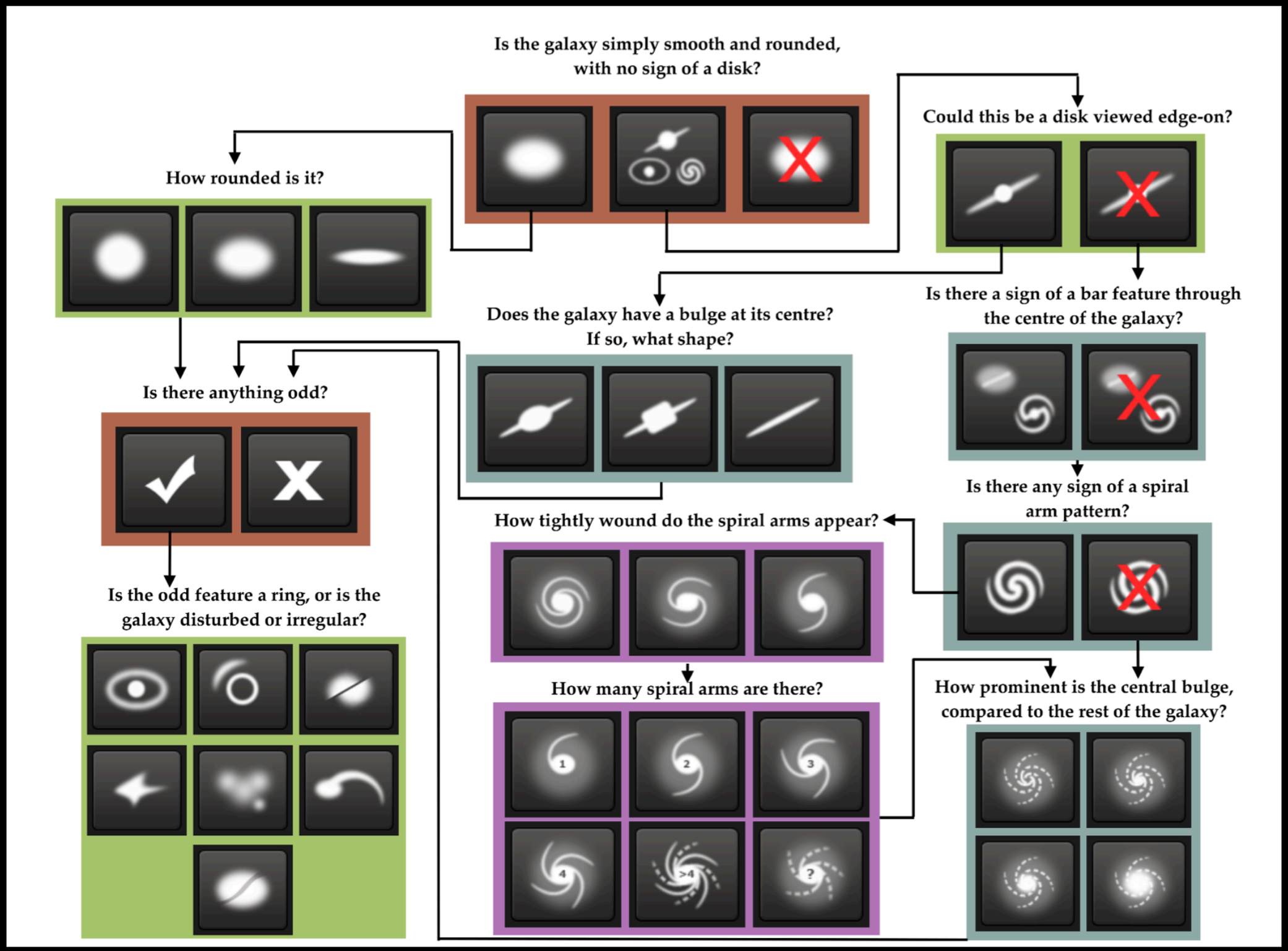
EXAMPLE 1: GALAXY CLASSIFICATION

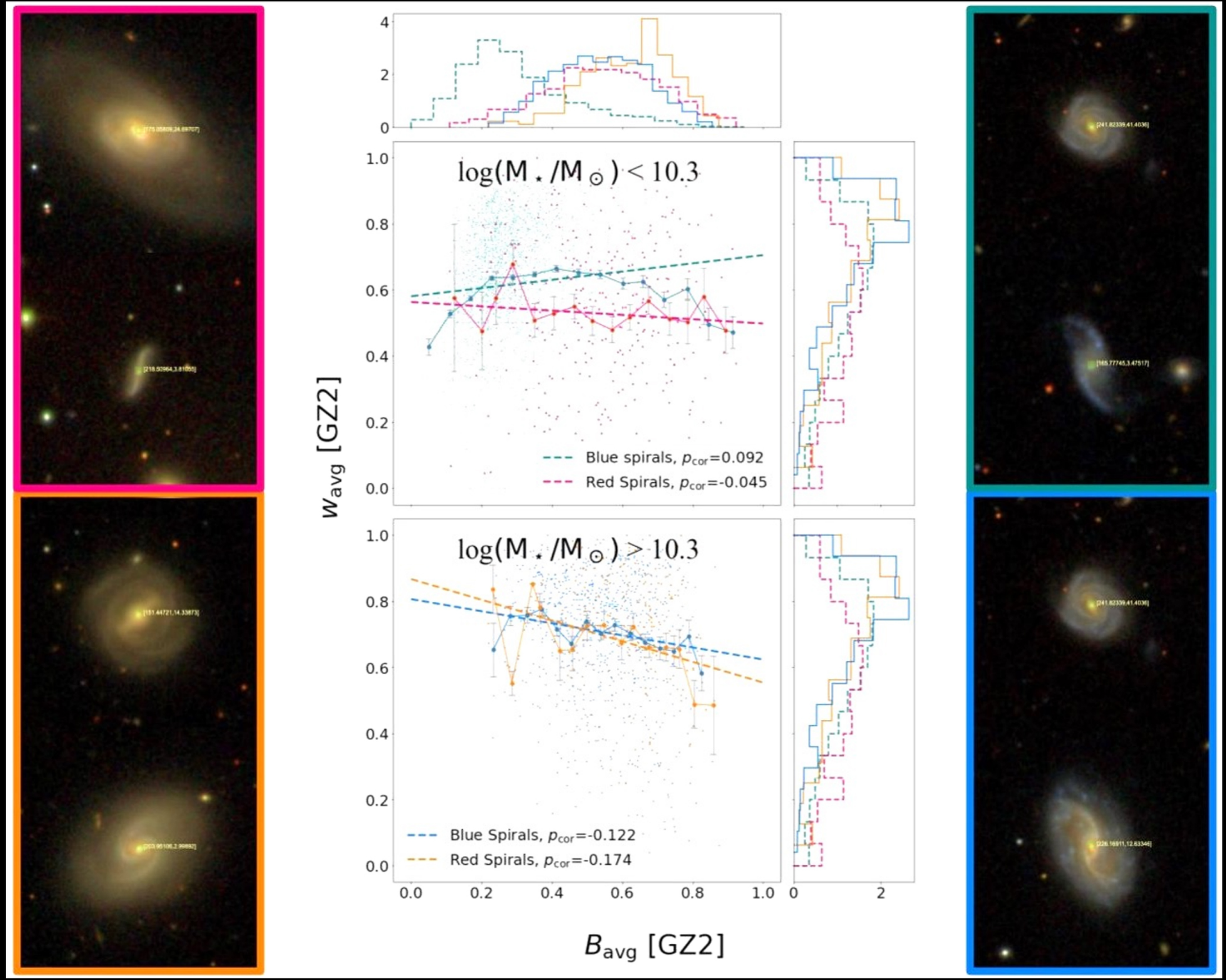
Hubble's Tuning fork



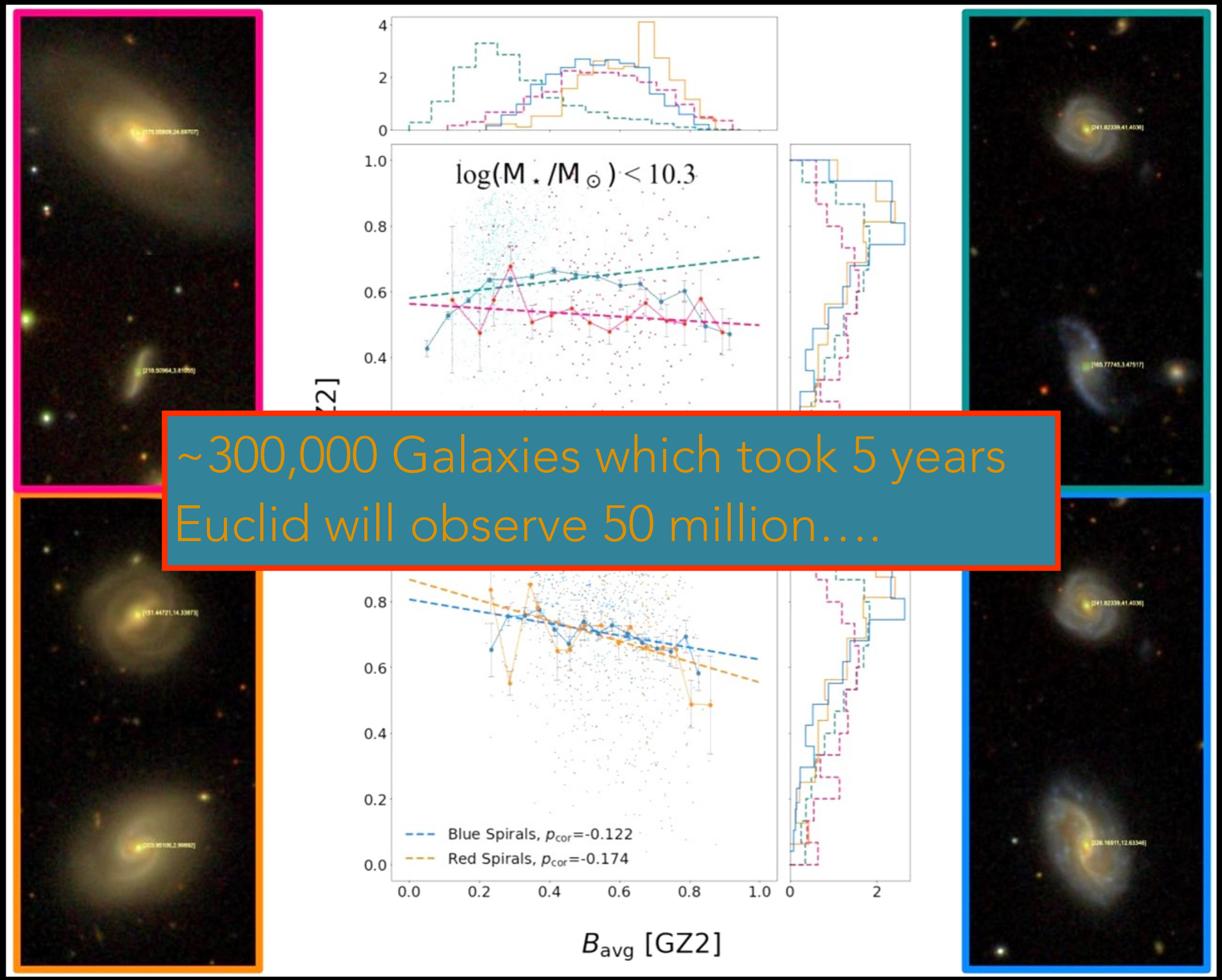
Lintott et al 2008

The Galaxy Zoo decision tree.





Mengistu & Master 2023



~300,000 Galaxies which took 5 years

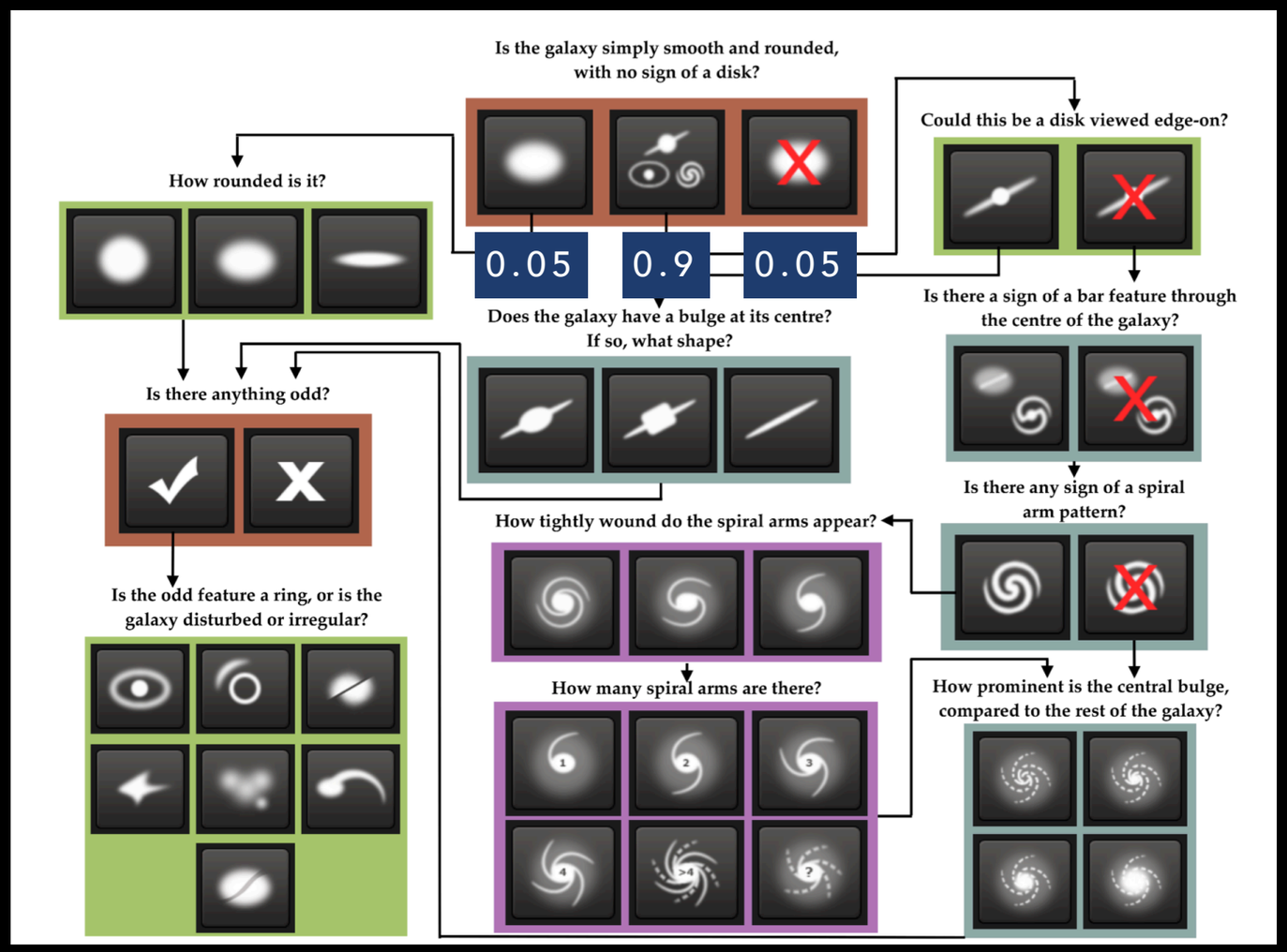
Euclid will observe 50 million….

Mengistu & Master 2023

kaggle

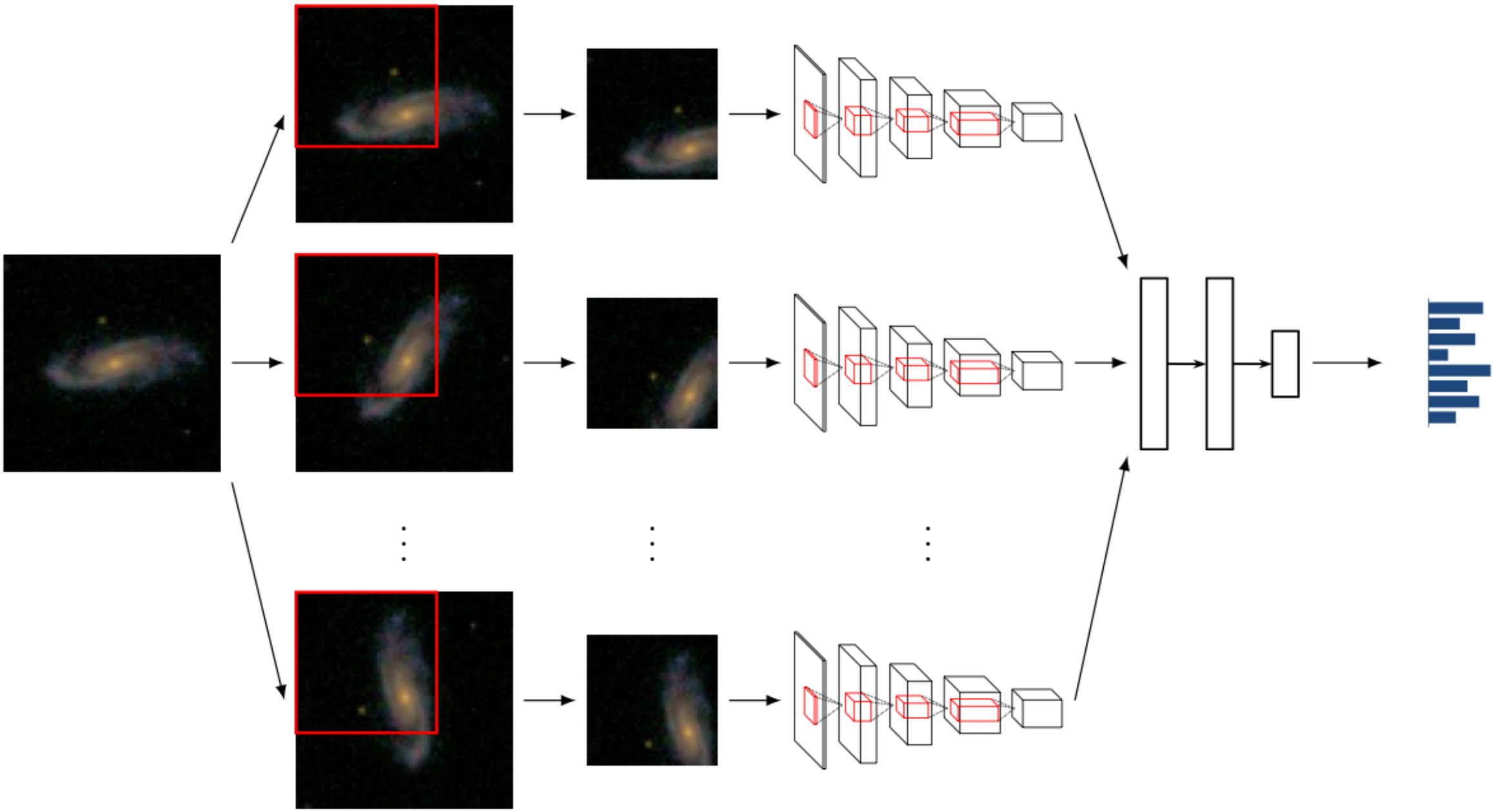
https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/data

Rotationally invariant convolutional neural networks to predict what a classifier would measure.

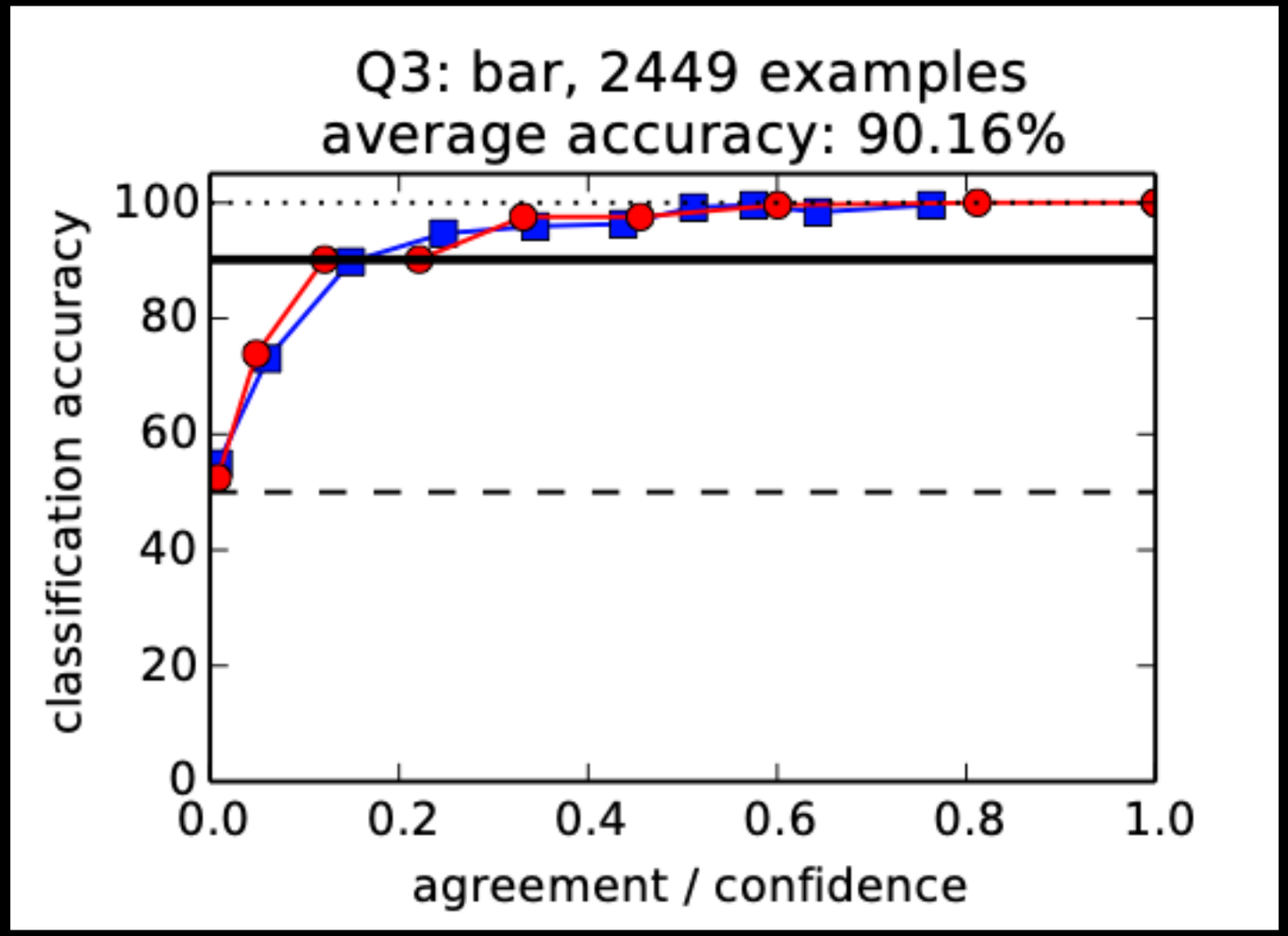


$$e(\hat{p}, p) = \sqrt{\sum_{k=1}^{37} (\hat{p} - p)^2}$$

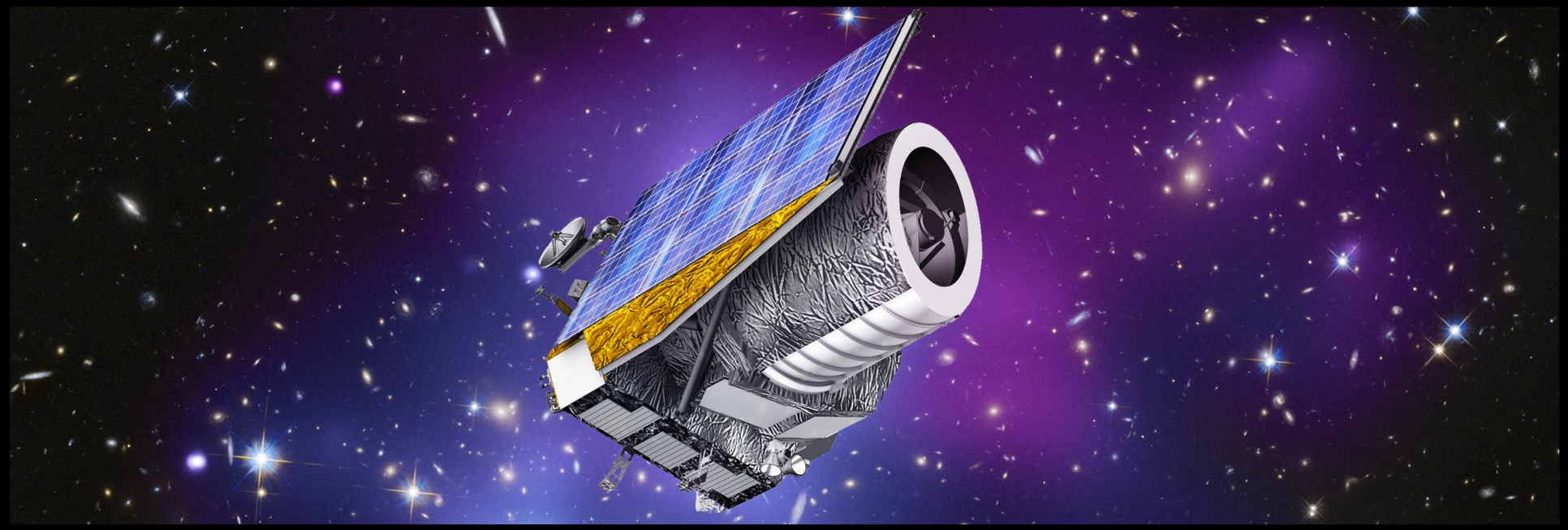
The winning algorithm



Winning the game or doing science?

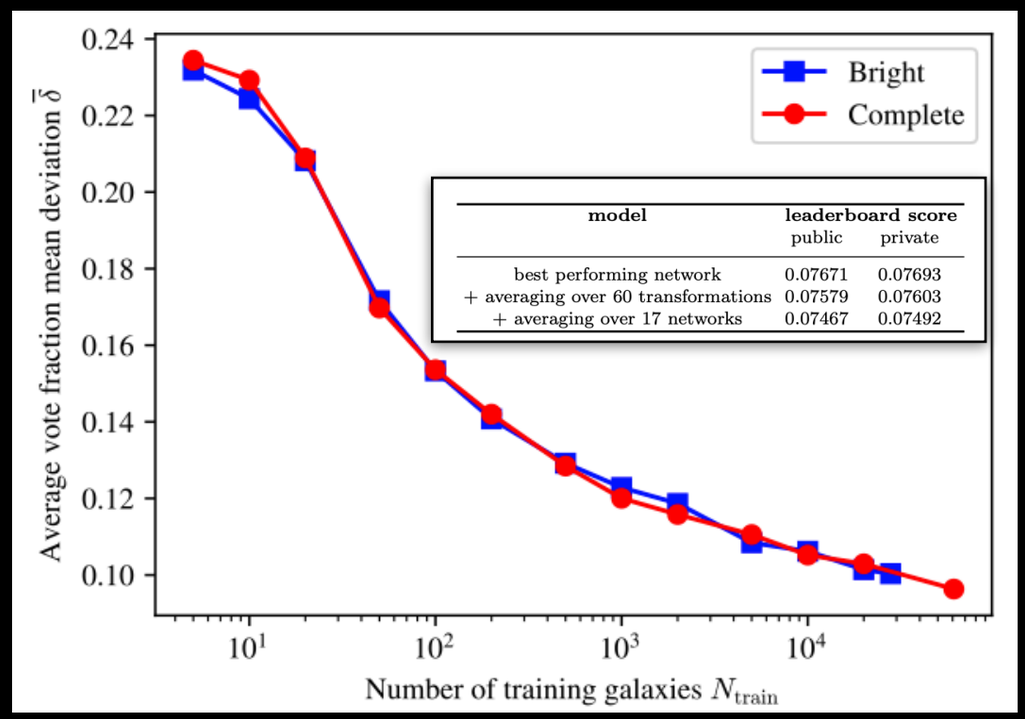


Labelled data for Euclid



Euclid preparation

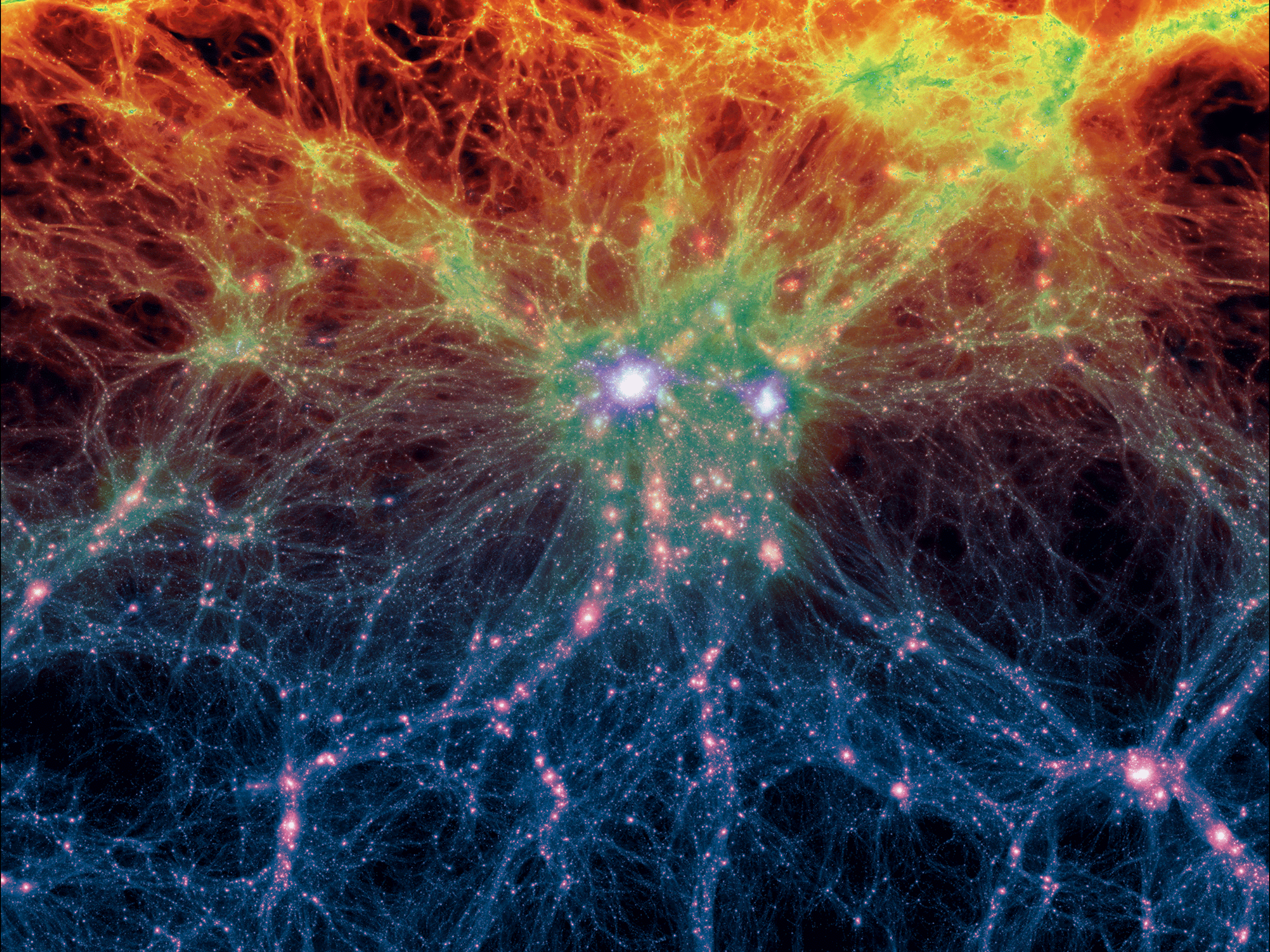
XLIII. Measuring detailed galaxy morphologies for *Euclid* with machine learning



model	leaderboard score public	leaderboard score private
best performing network	0.07671	0.07693
+ averaging over 60 transformations	0.07579	0.07603
+ averaging over 17 networks	0.07467	0.07492

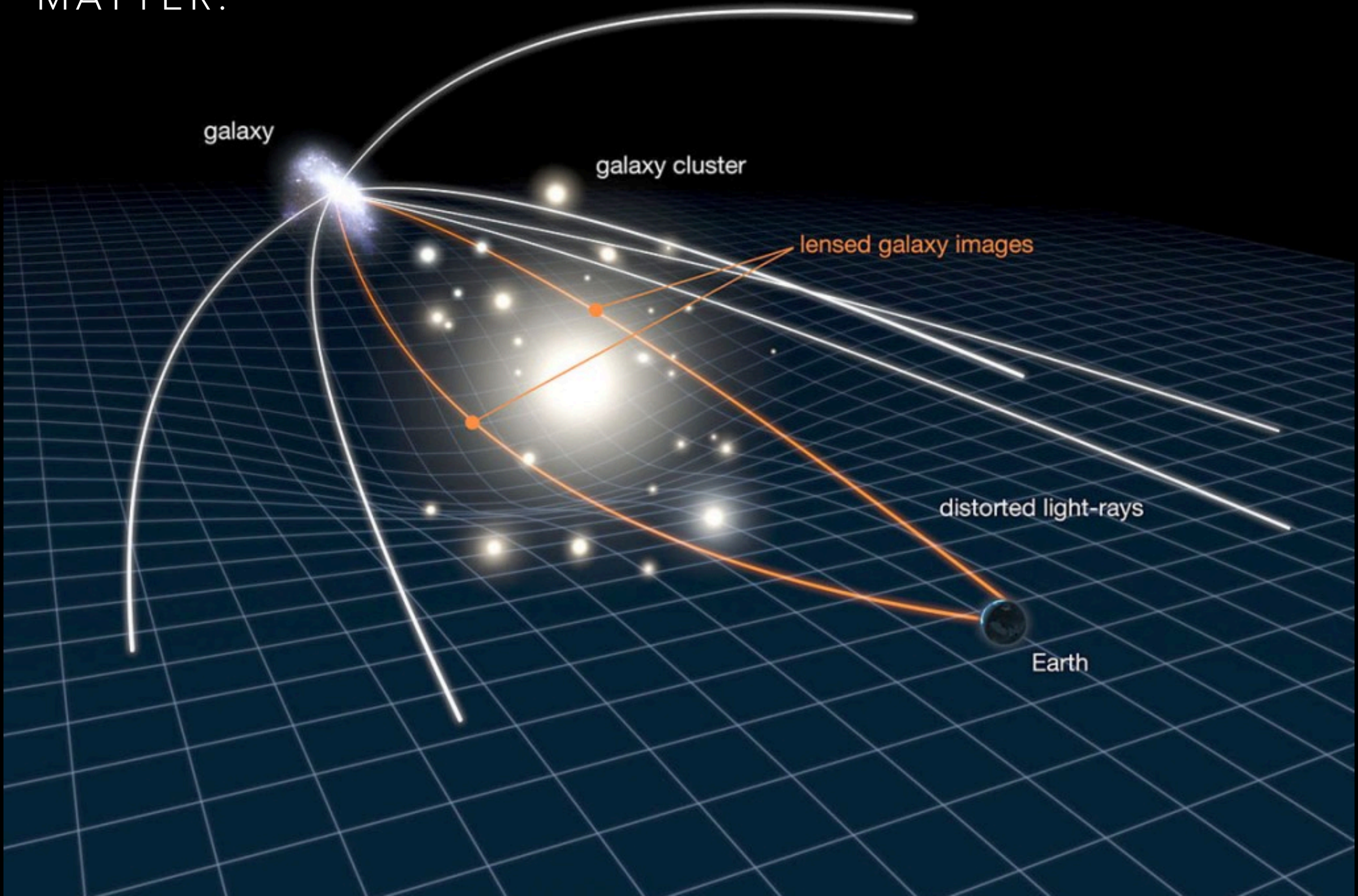
CLASSICAL INFERENCE REQUIRES ANALYTICAL MODELS THAT CAN BE PASSED IN TO A MCMC WITH MANY CALLS TO A LIKELIHOOD FUNCTION

EXAMPLE 2: DARK MATTER



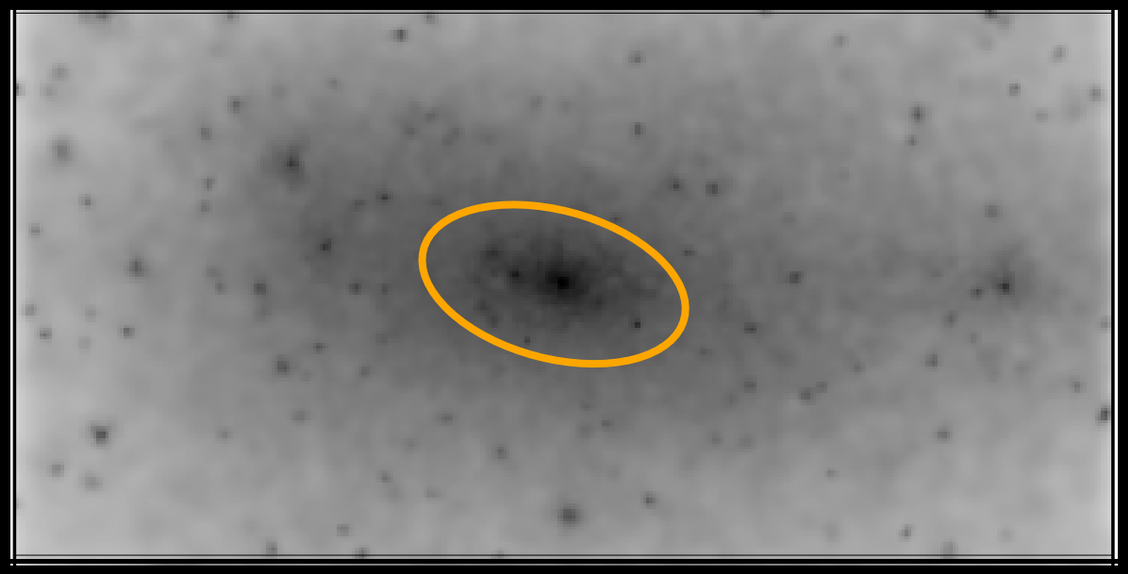
GALAXY CLUSTERS ARE DOMINATED BY DARK MATTER

GRAVITATIONAL LENSING HELPS US TRACE THE DARK MATTER.

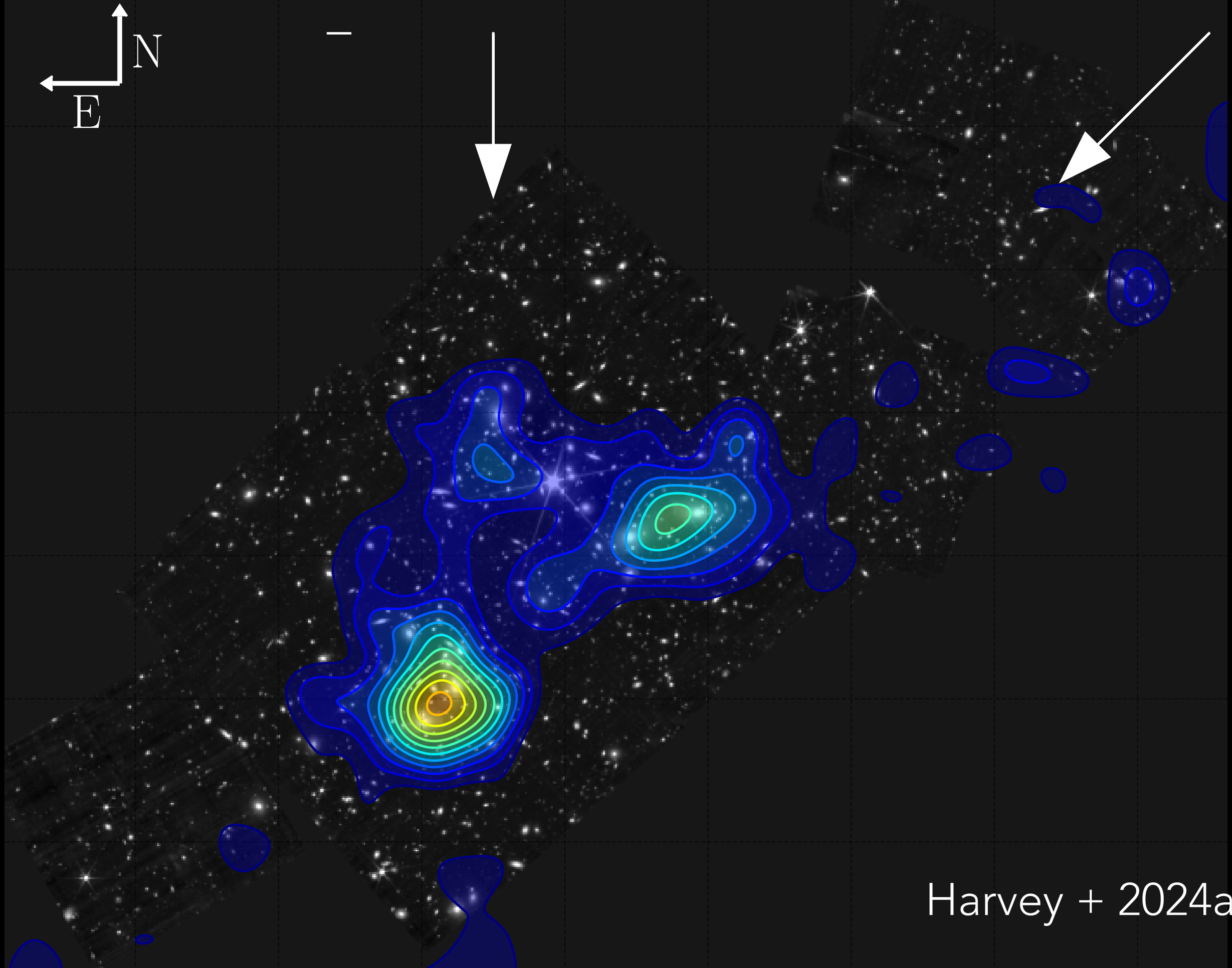


CLASSICAL MODELLING TO INFER PARAMETERS

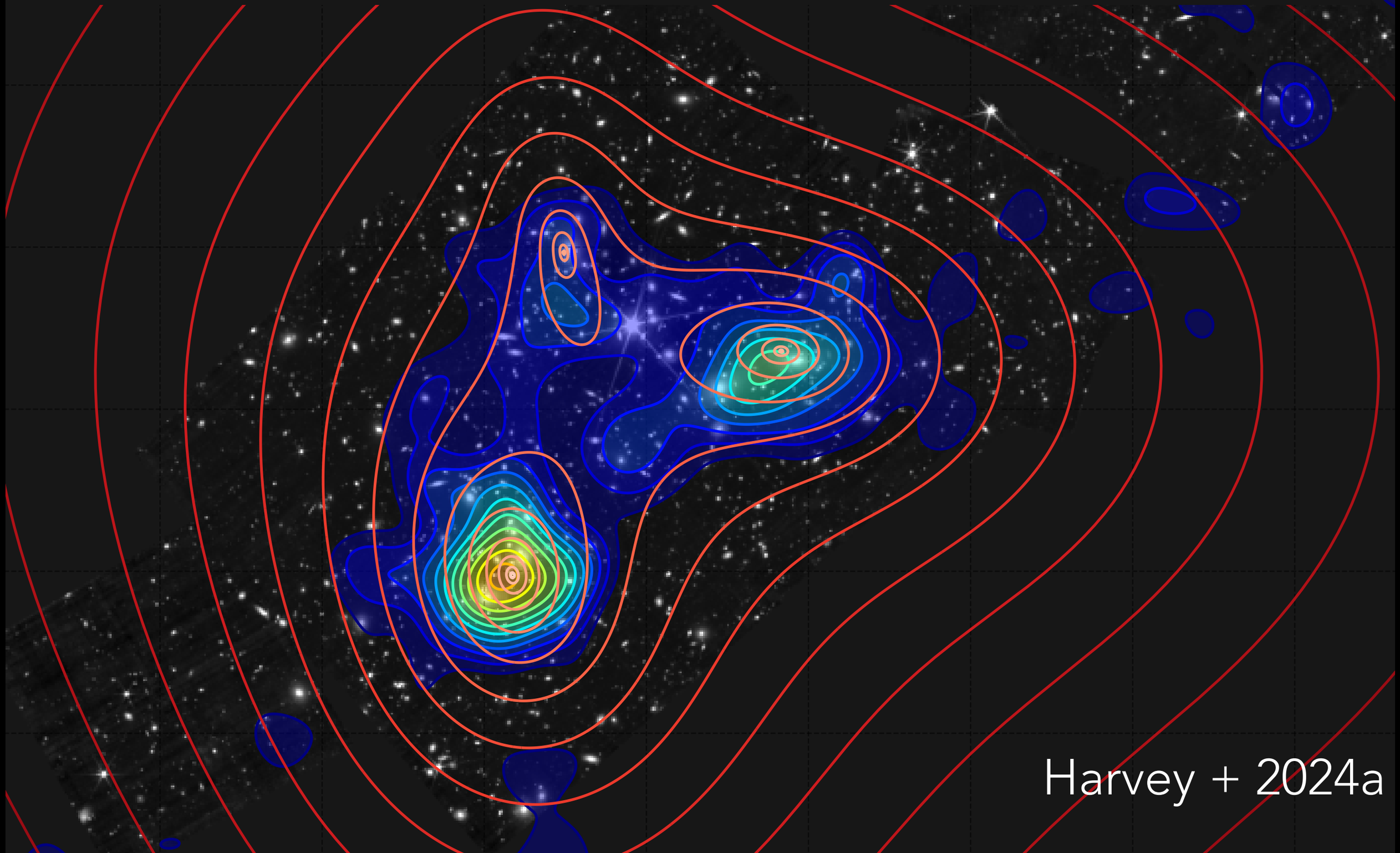
COMPARE THE PARAMETERS TO SIMULATIONS OF GALAXY CLUSTERS WITH DIFFERENT MODELS OF DARK MATTER



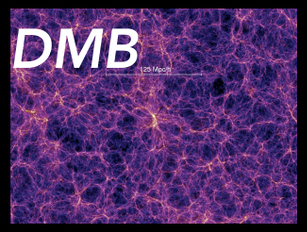
MODELLING IS SLOW

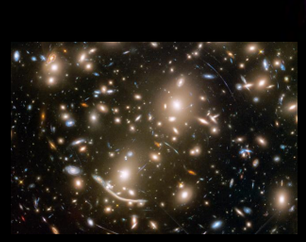

N
E
Harvey + 2024a

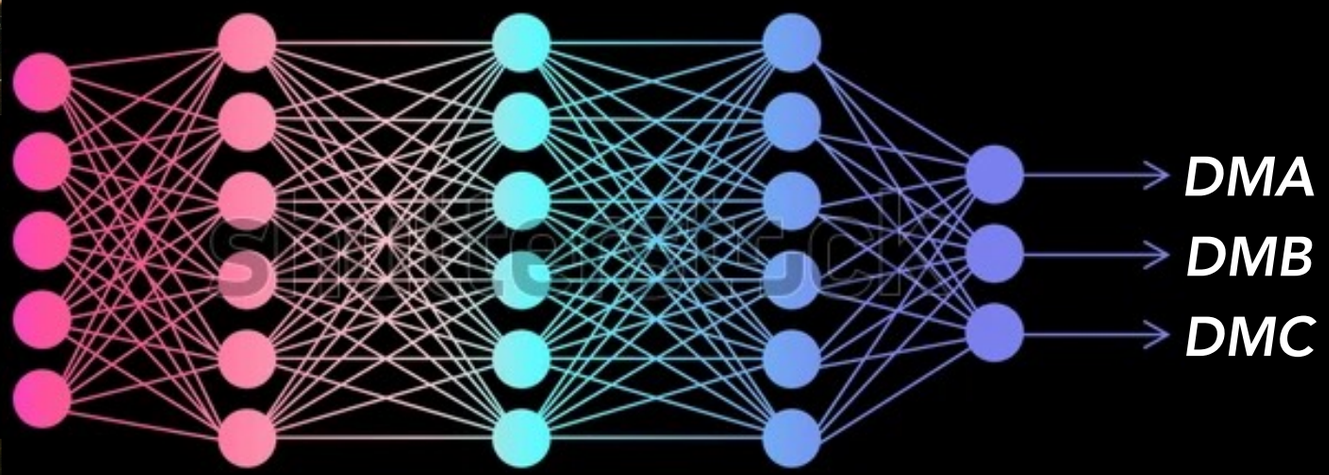
BUT INFORMATION CAN BE
LOST IN SIMPLIFICATION.

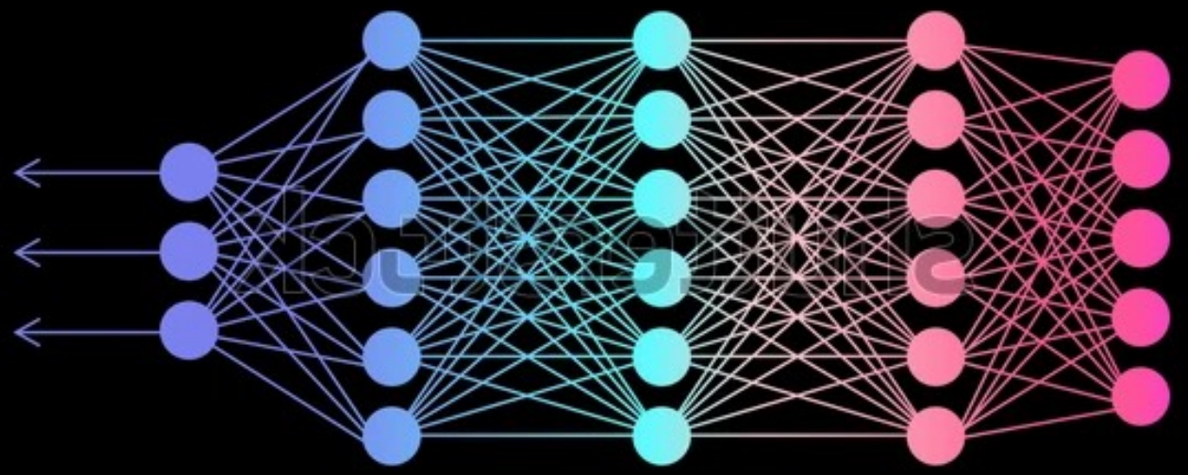


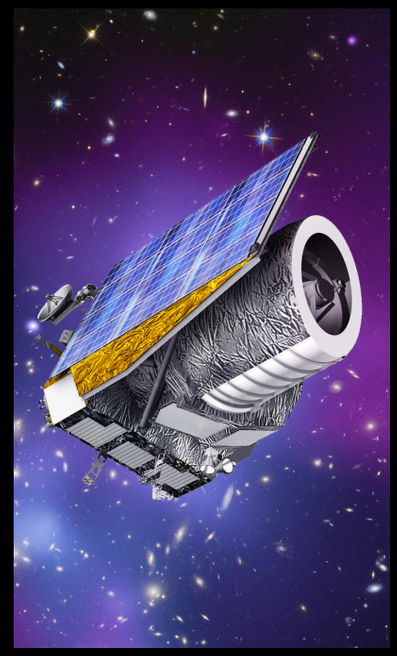
DEEP NETS CAN HELP US AGNOSTICALLY PROBE DARK MATTER WHILST SPEEDING UP THE PROCESS



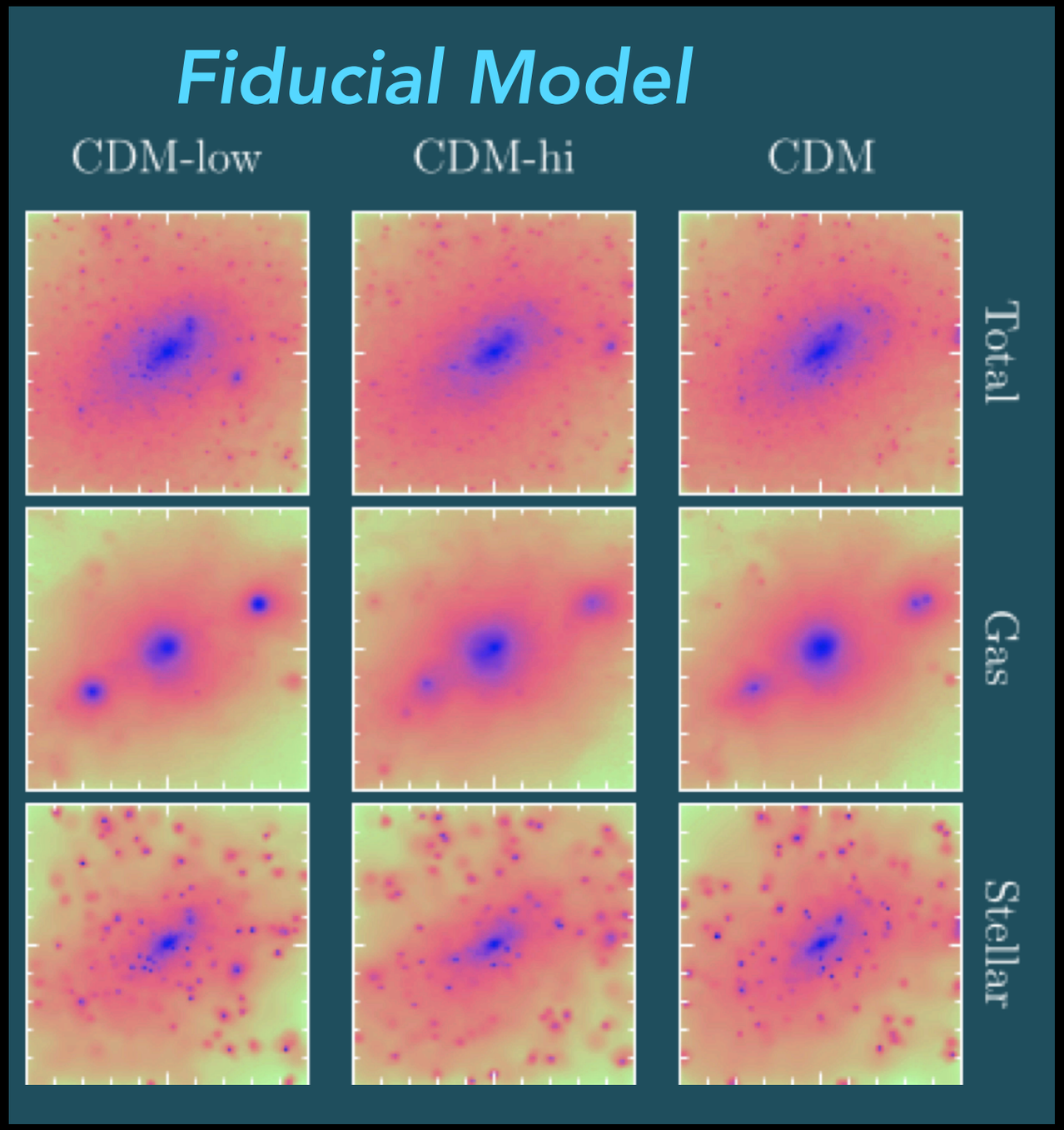




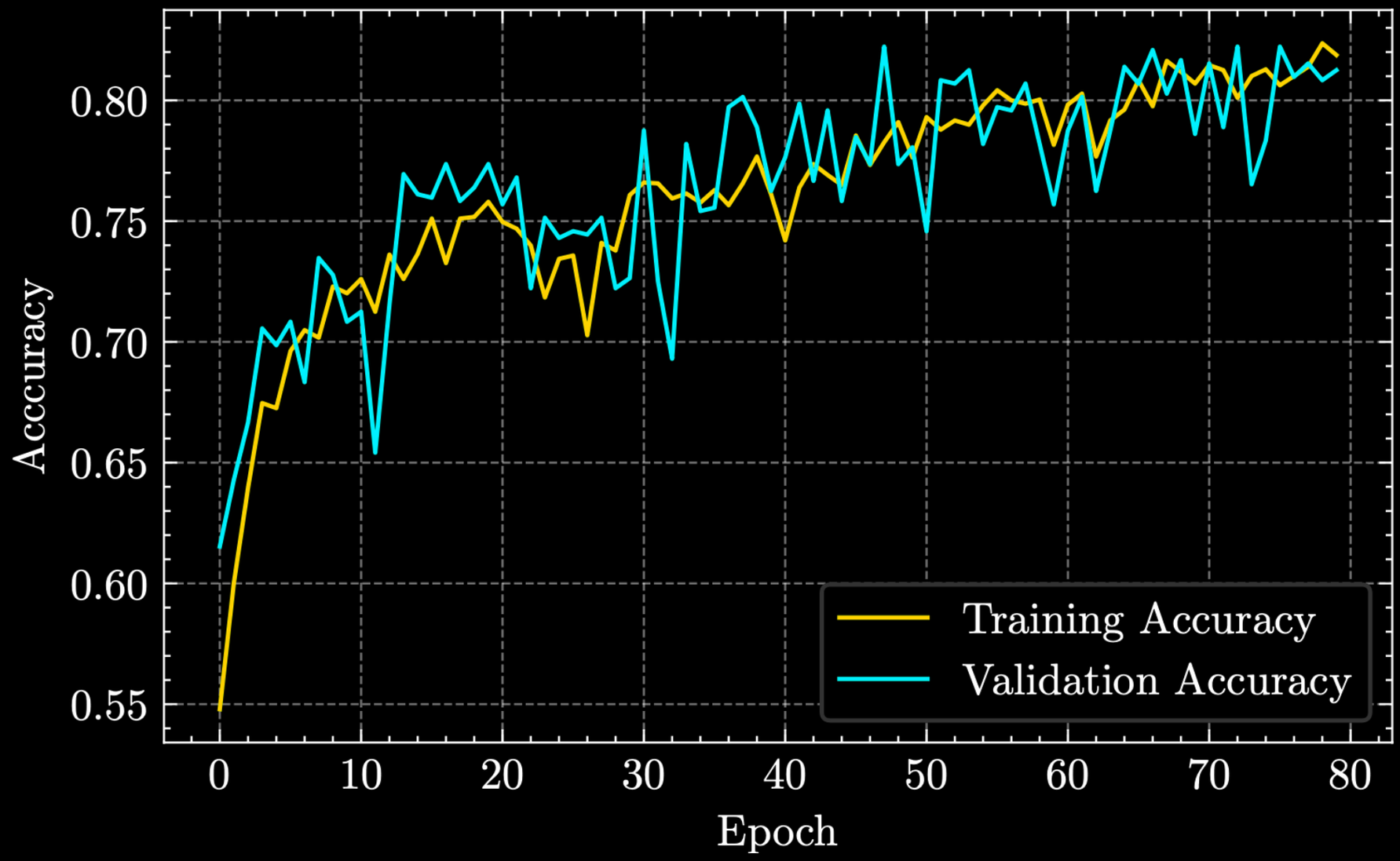




FIRST WE NEED OUR SAMPLES

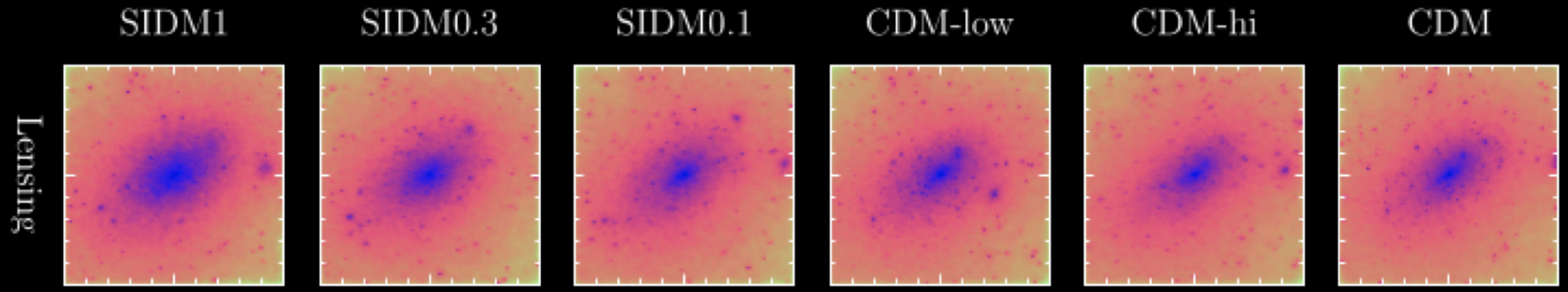


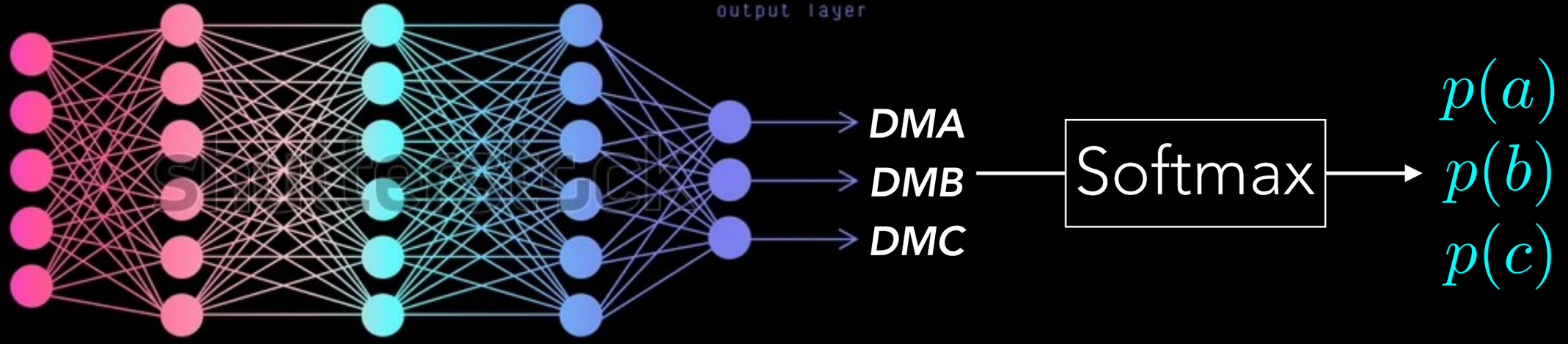
WE CAN REACH 80% ACCURACY



Harvey 2024

CLASSIFICATION TO REGRESSION

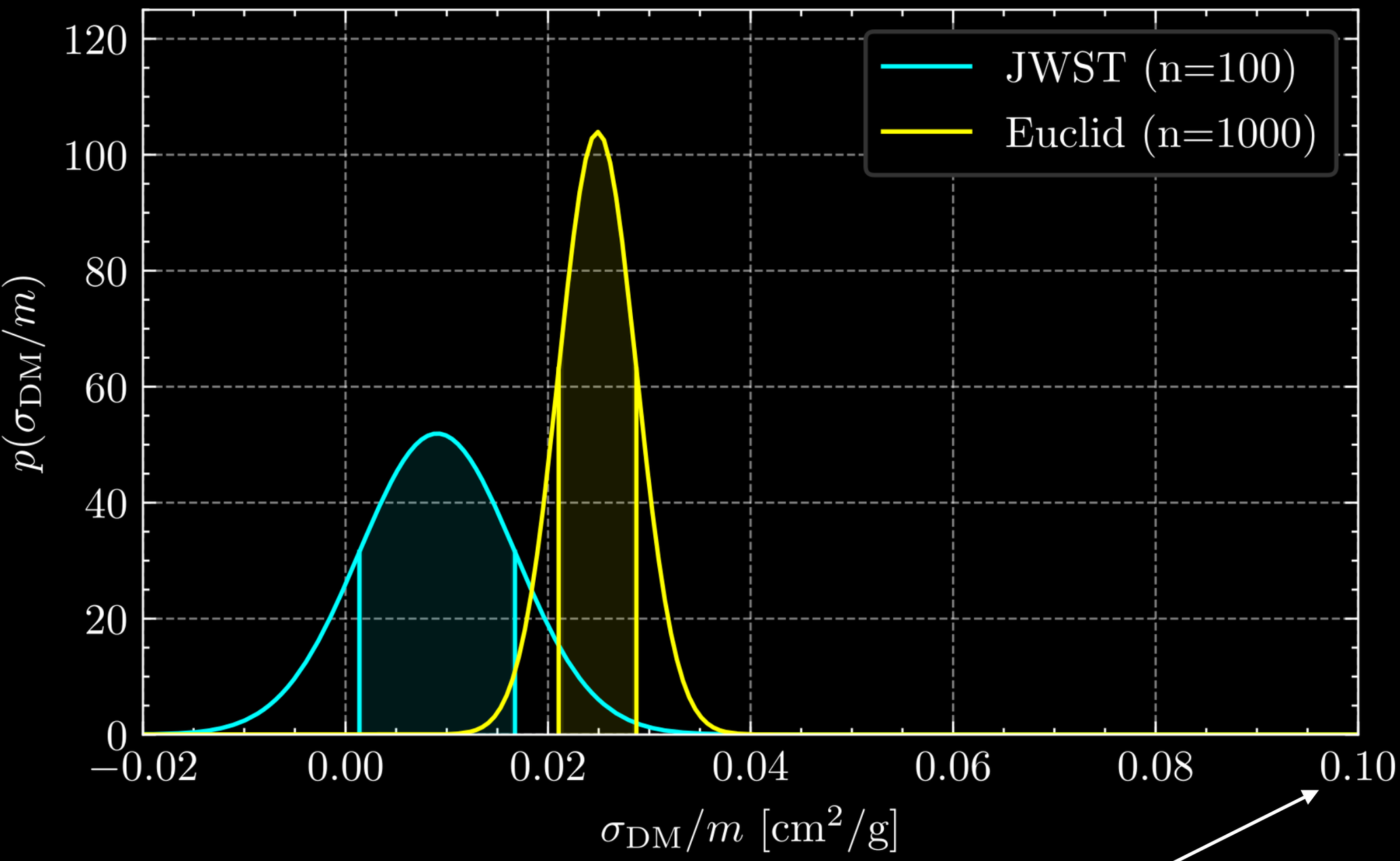




$$\hat{x} = \sum p(x)x$$

Harvey 2024

Predictions after forward modelling on upcoming telescopes



Limit with classic methods

Harvey + 2024b

Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

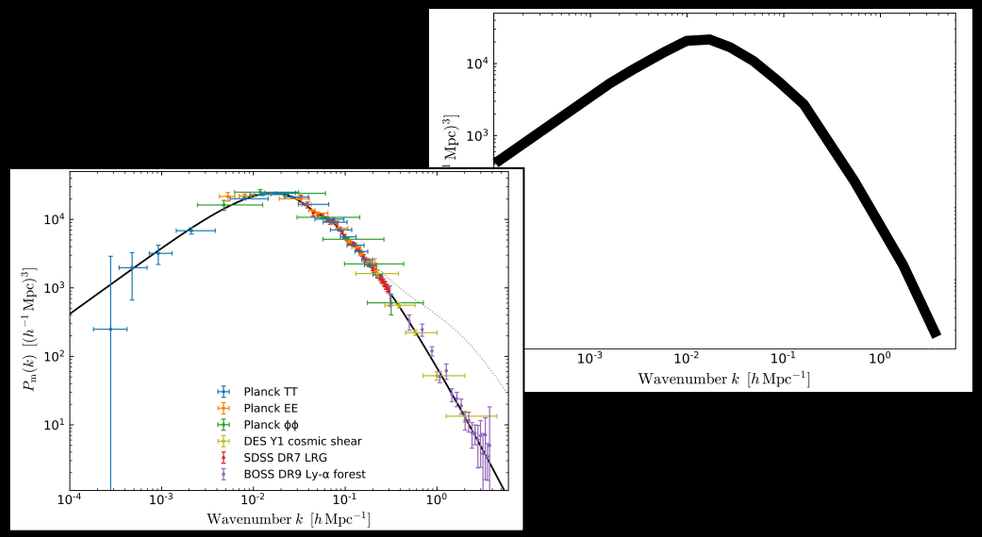
- The number of posterior calls is normally very large.

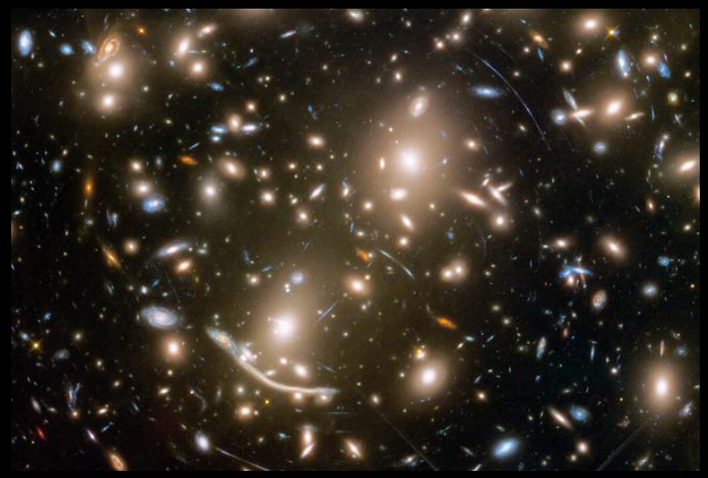
- Likelihoods are not Gaussian.

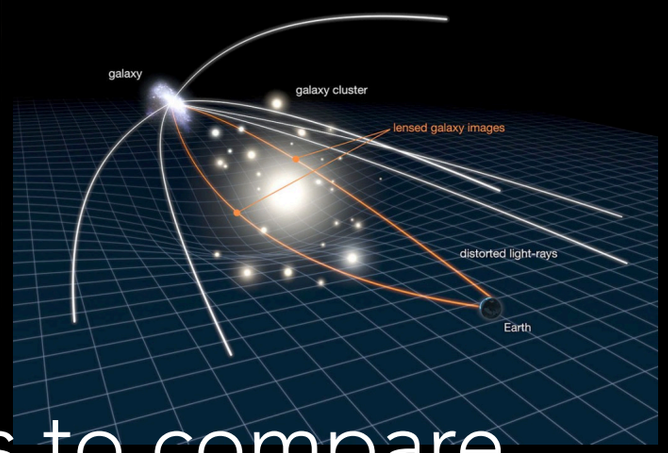
- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

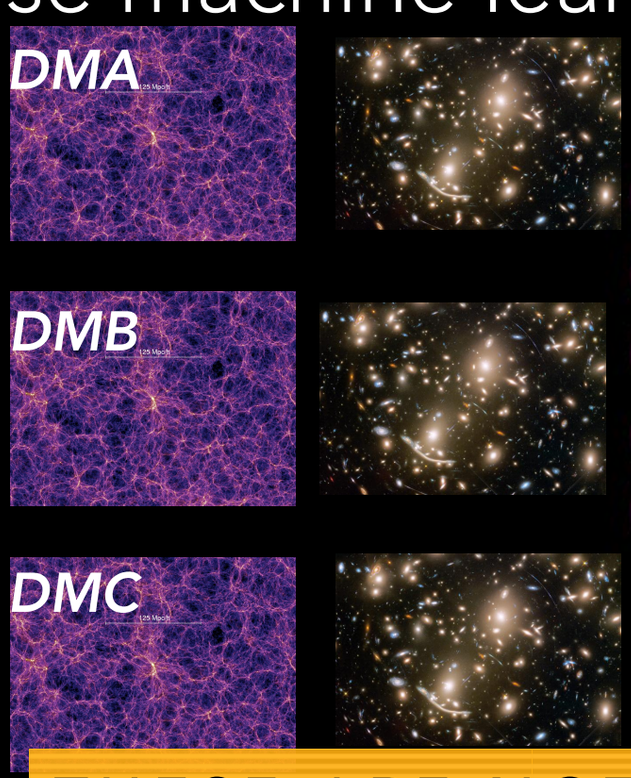
Classical inference is slow and requires analytical models that often lose information resulting in loose and biased inference.

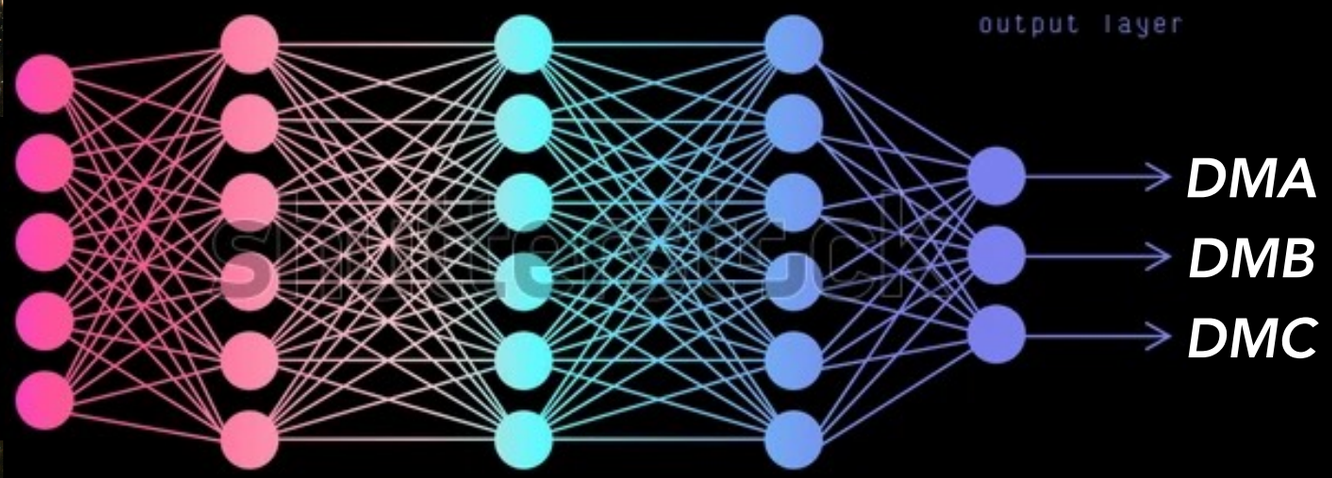




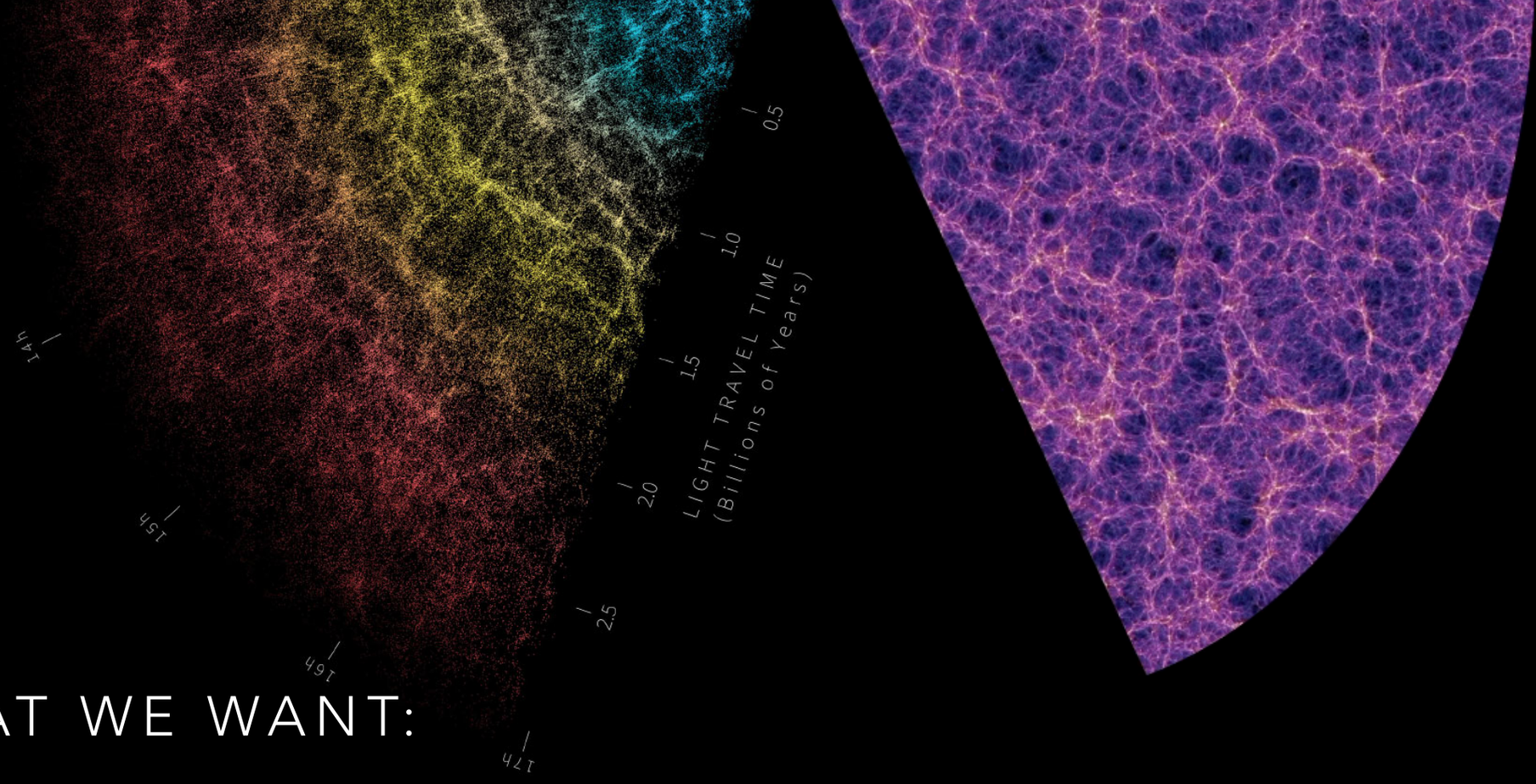


Use machine learning directly on simulations to compare.





THESE ARE NOT STATISTICALLY ROBUST WITH UNREGULATED UNCERTAINTY



WHAT WE WANT:

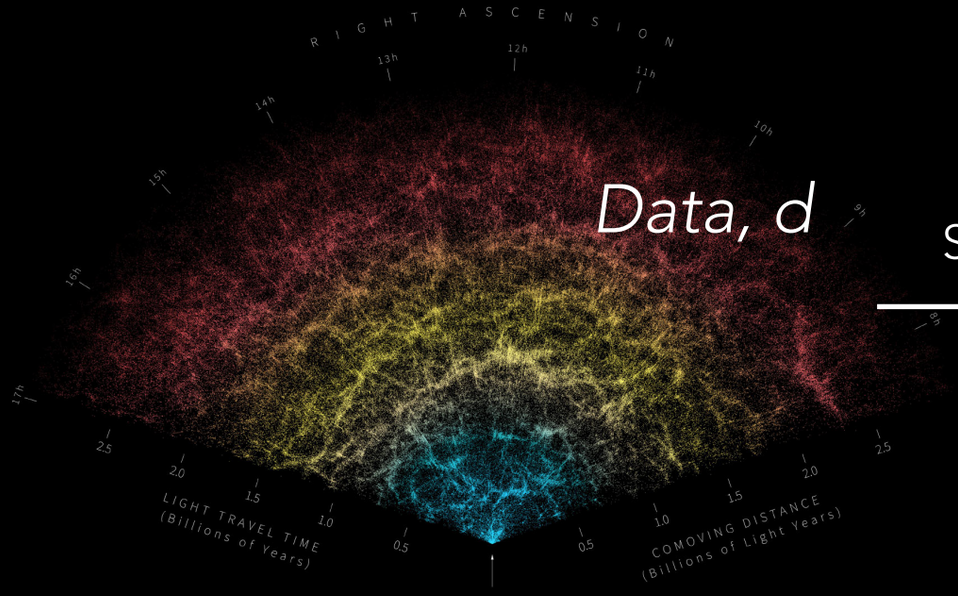
- TO DIRECTLY COMPARE OUR COMPLEX FORWARD MODELS TO THE DATA

- THAT DO NOT ASSUME HOW THE LIKELIHOOD OF THE PARAMETERS OF THE MODEL ARE DISTRIBUTED

- QUICKLY, EFFICIENTLY AND UNBIASED

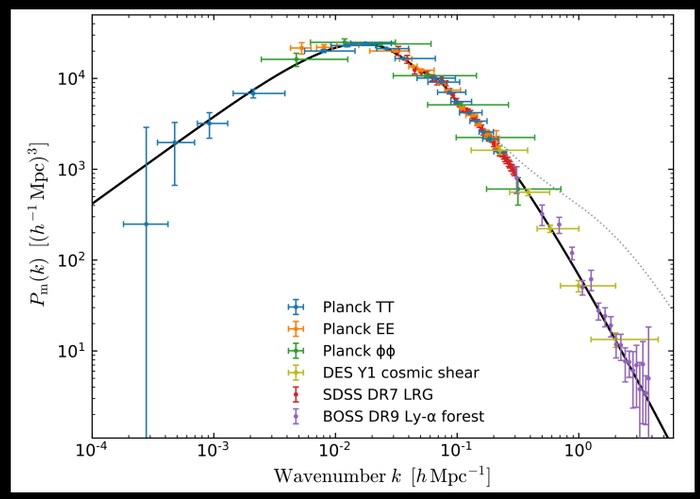
Simulation Based Inference

BASIC PRINCIPLES OF SBI

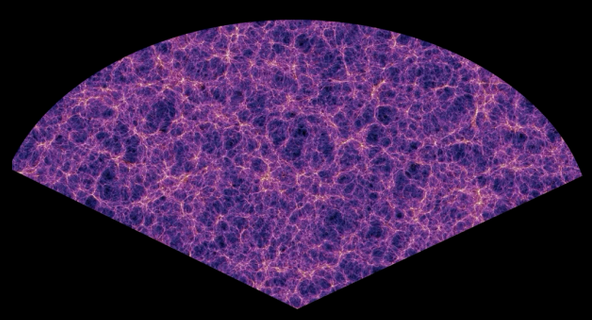


Data, d

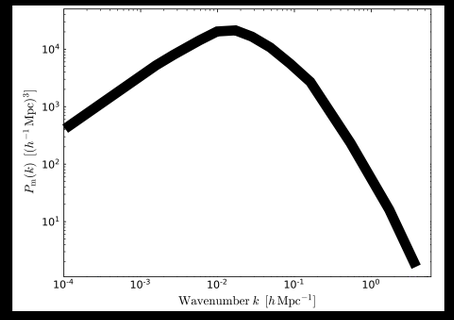
Summaries t_o



θ_1

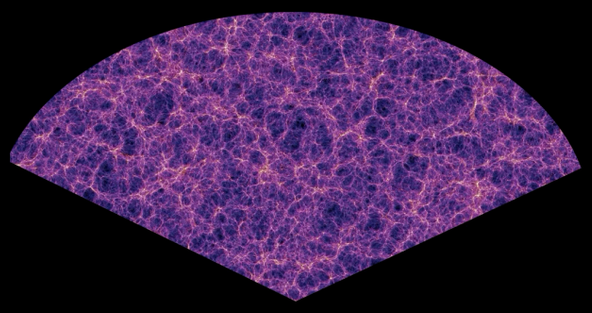


Summaries

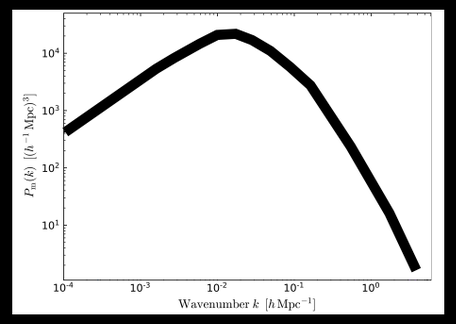


$\rho(t_1, t_o)$

θ_2

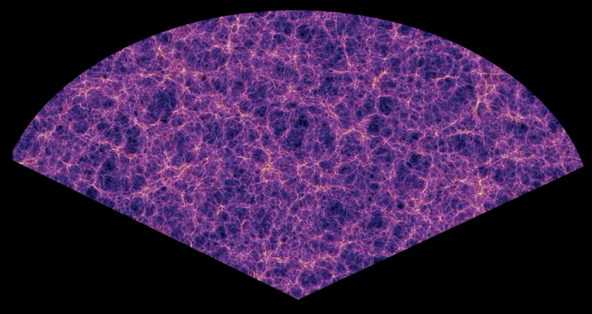


Summaries

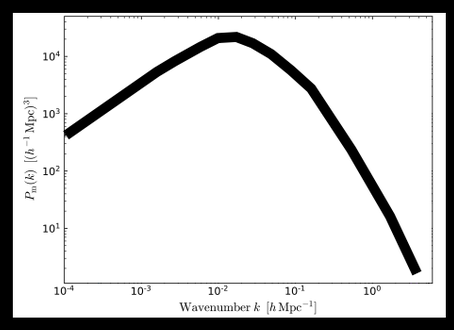


$\rho(t_2, t_o)$

θ_n

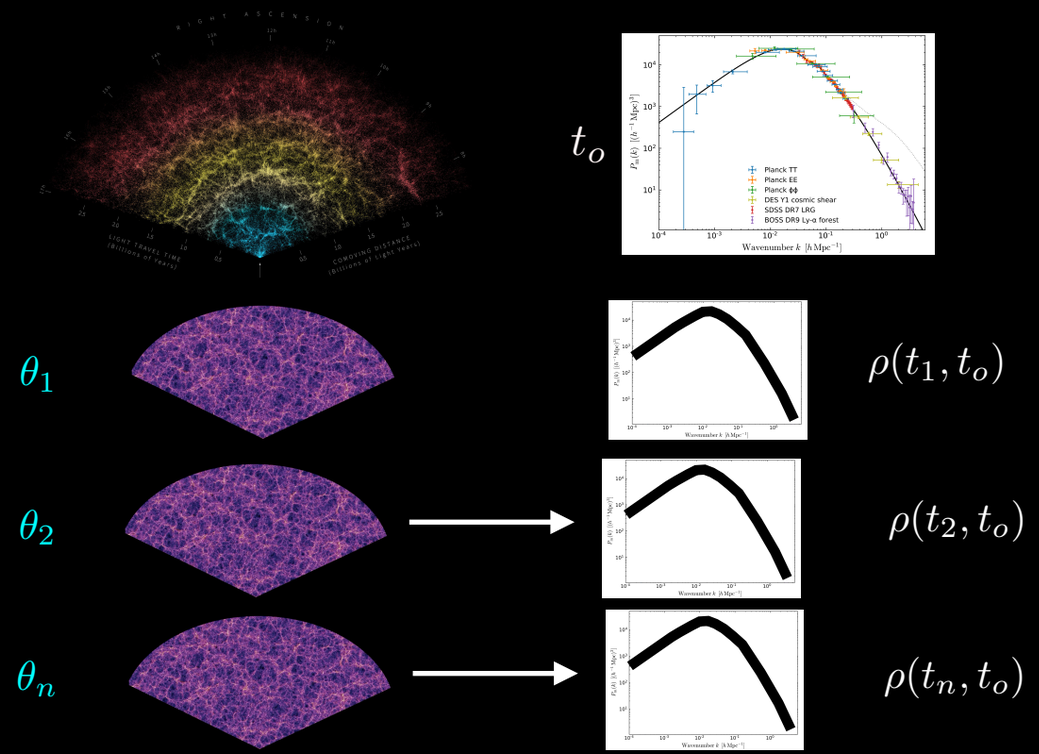


Summaries



$\rho(t_n, t_o)$

APPROXIMATE BAYESIAN COMPUTATION

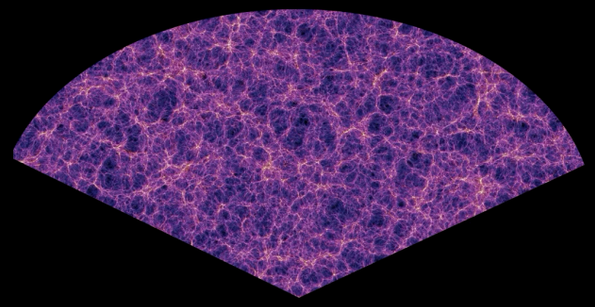


- Directly compare complex non-linear models
- Forward modelled to resemble the data
- Unbiased, conservative
- Evidence is calculated

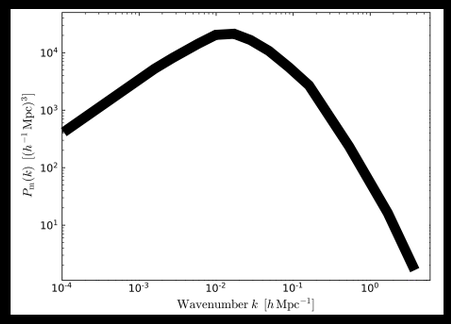
- Take some delta in “agreement” - not robust still
- Requires a lot of simulations
- Scales exponentially with number of parameters

What I would rather: Estimate the true posterior or likelihood of the parameters in a model free way.

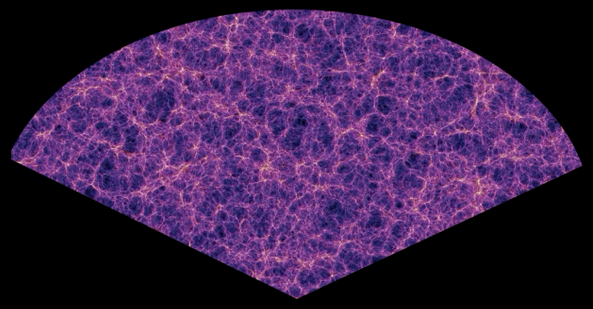
θ_1



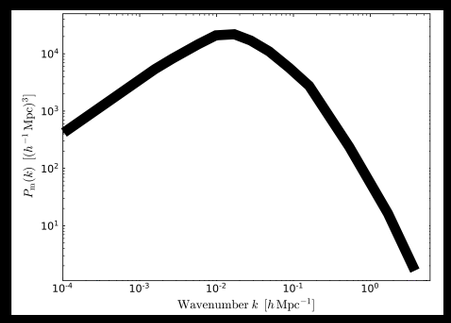
Summaries t_1



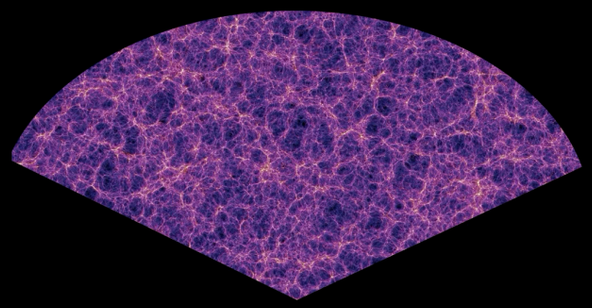
θ_2



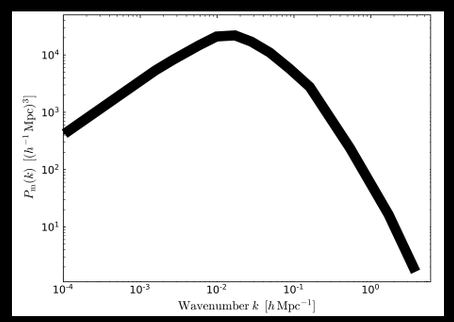
Summaries t_2



θ_n



Summaries t_n



$$p(\theta|t)$$

$$p(t|\theta)$$

Then train some density estimator to predict the likelihood / posterior at non-sampled points in the parameter space -

Then evaluate at the location of the observations..

$$p(\theta|t)$$

$$p(\theta|t = t_o)$$

Does not scale with parameters very well

$$p(t|\theta)$$

$$p(t_o|\theta)p(\theta)$$

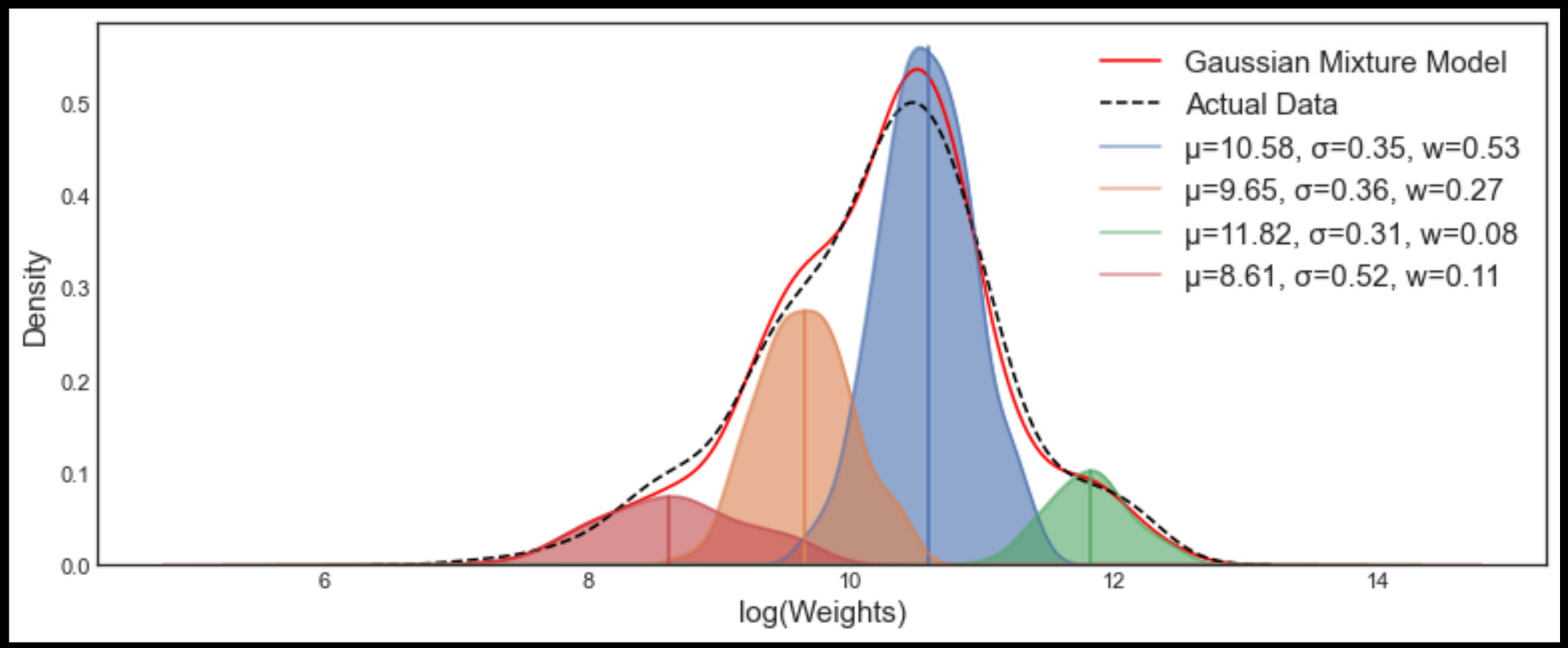
Nice that different priors can be placed

How do we estimate the likelihood for the sampled distribution

Neural Density Estimators

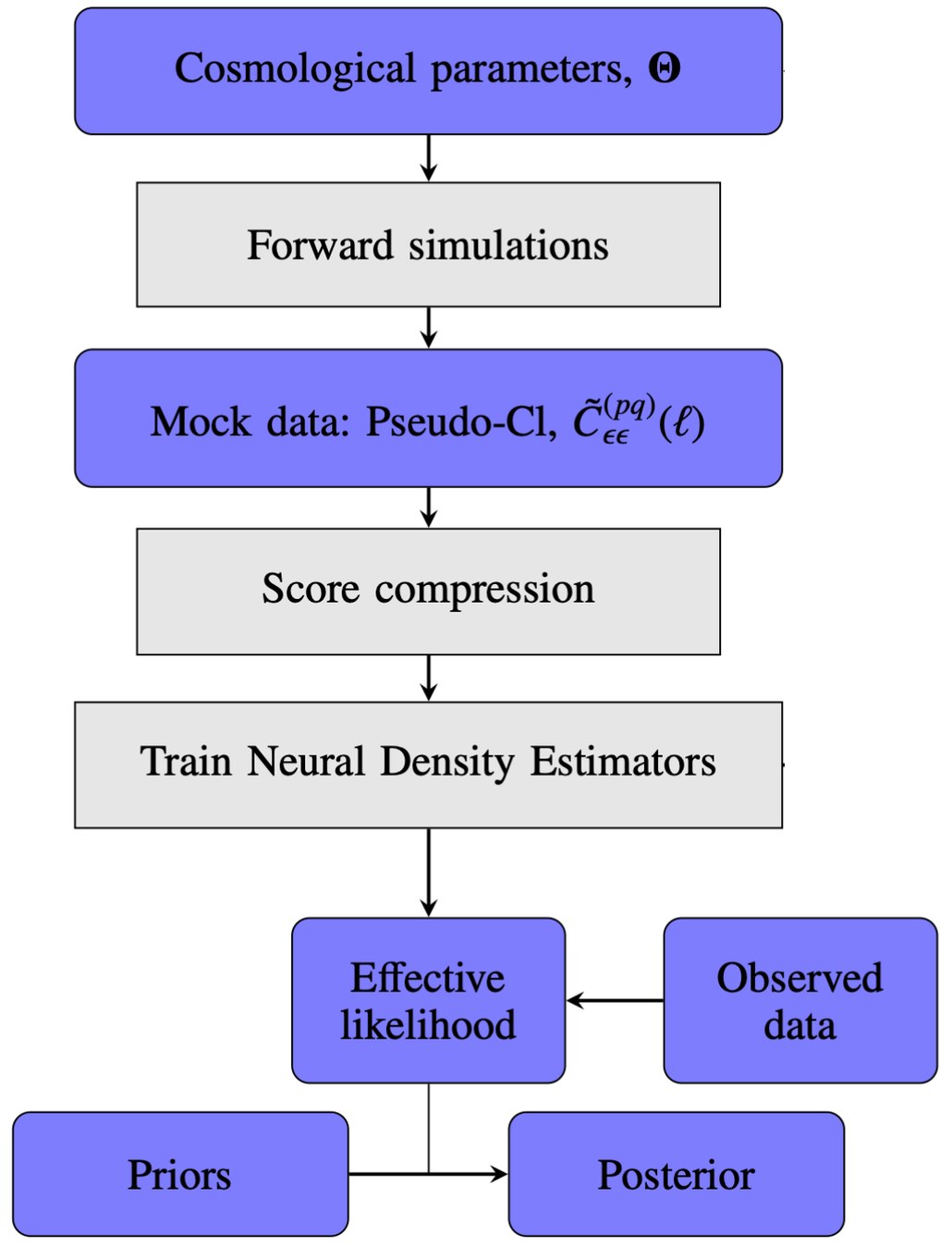
$$p(t|\theta; w)$$

Gaussian mixture model



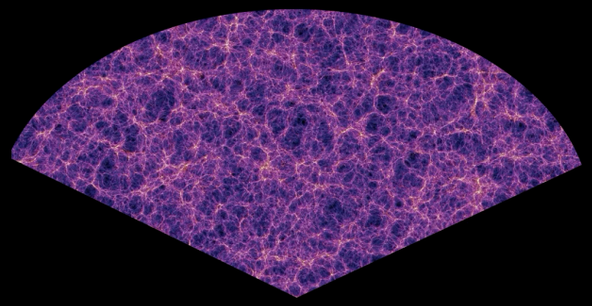
Train using a discrete, monte carlo estimation of the Kullback-Leibler divergence.

HOW THIS WORKS IN PRACTICE

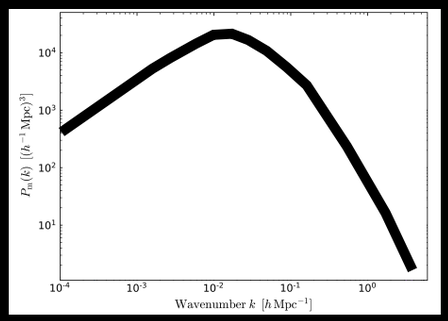


Simulations are very expensive to run.

θ_n



Summaries t_n



Why astronomy needs machine learning

- Classical inference requires analytical models that can be passed in to a MCMC

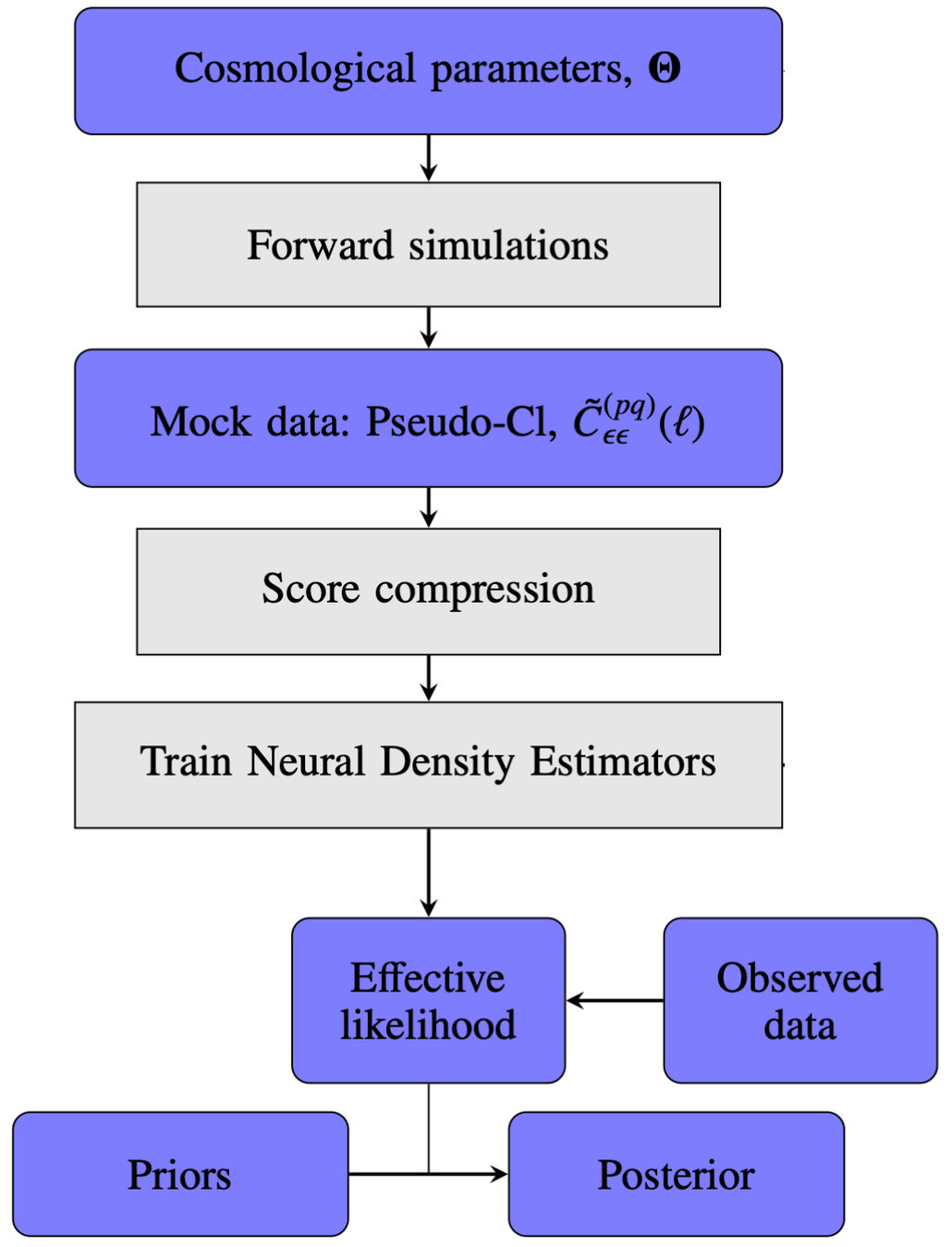
- The number of posterior calls is normally very large.

- Likelihoods are not Gaussian.

- Direct comparison of simulations to observations would be impossible in the current situation.

- Astrophysical simulations are often computationally expensive

HOW THIS WORKS IN PRACTICE



Active learning (with Sequential Neural Likelihood)

First -> more neural density estimators the better

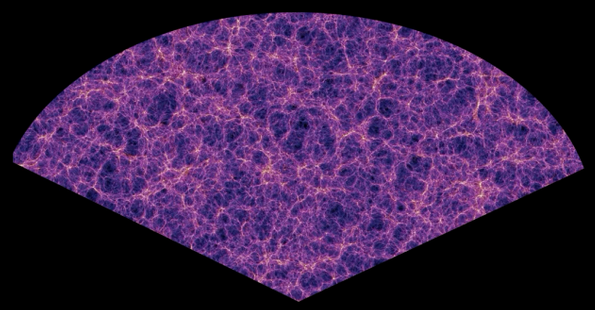
Ensembles NDEs

$$p(t|\theta;w) = \sum^{N_{\mathrm{NDE}}} \beta_\alpha p_\alpha(t|\theta;w)$$

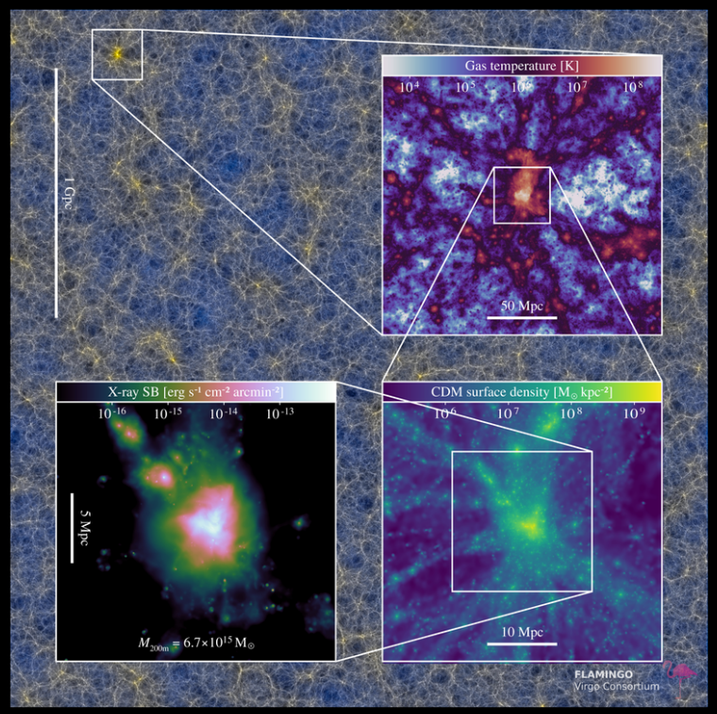
1

Secondly -> this tells you where the likelihood is important and where there is uncertainty.

SPEED UP SIMULATIONS WITH GENERATIVE METHODS.

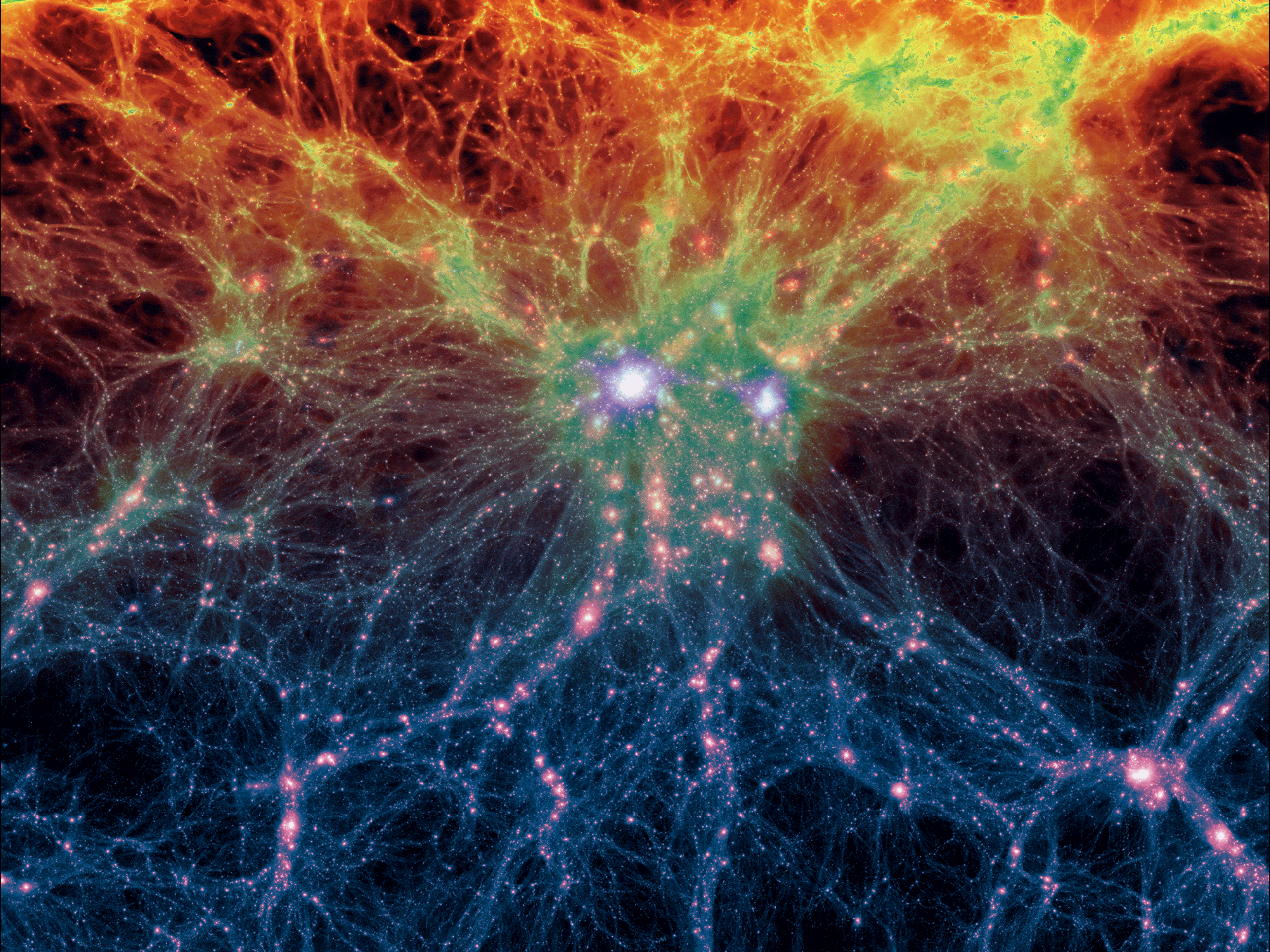


Dark matter only simulations are quick(ish)



Full simulations are slow

TRAIN A METHOD TO LEARN THE CONNECTION BETWEEN DARK MATTER AND ASTROPHYSICS



SPEED UP SIMULATIONS WITH GENERATIVE METHODS: VARIABLE AUTOENCODERS

X-ray emission

Dark matter distribution

$$p(x|y)$$

$$p(x|y) = \int \mathrm{d}z \; p(x|y,z)p(z|y)$$

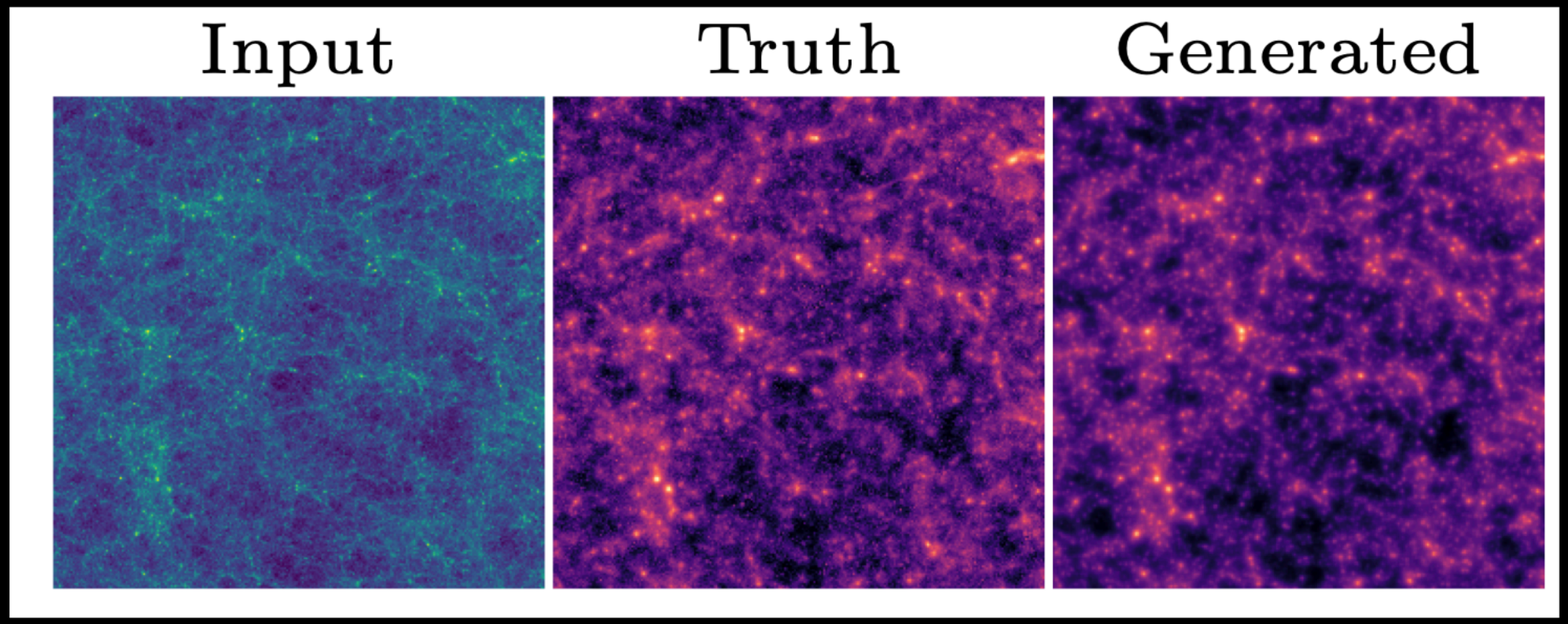
Some parameterised model

Prior information on the weights.

ASSUMING GAUSSIAN MIXTURES

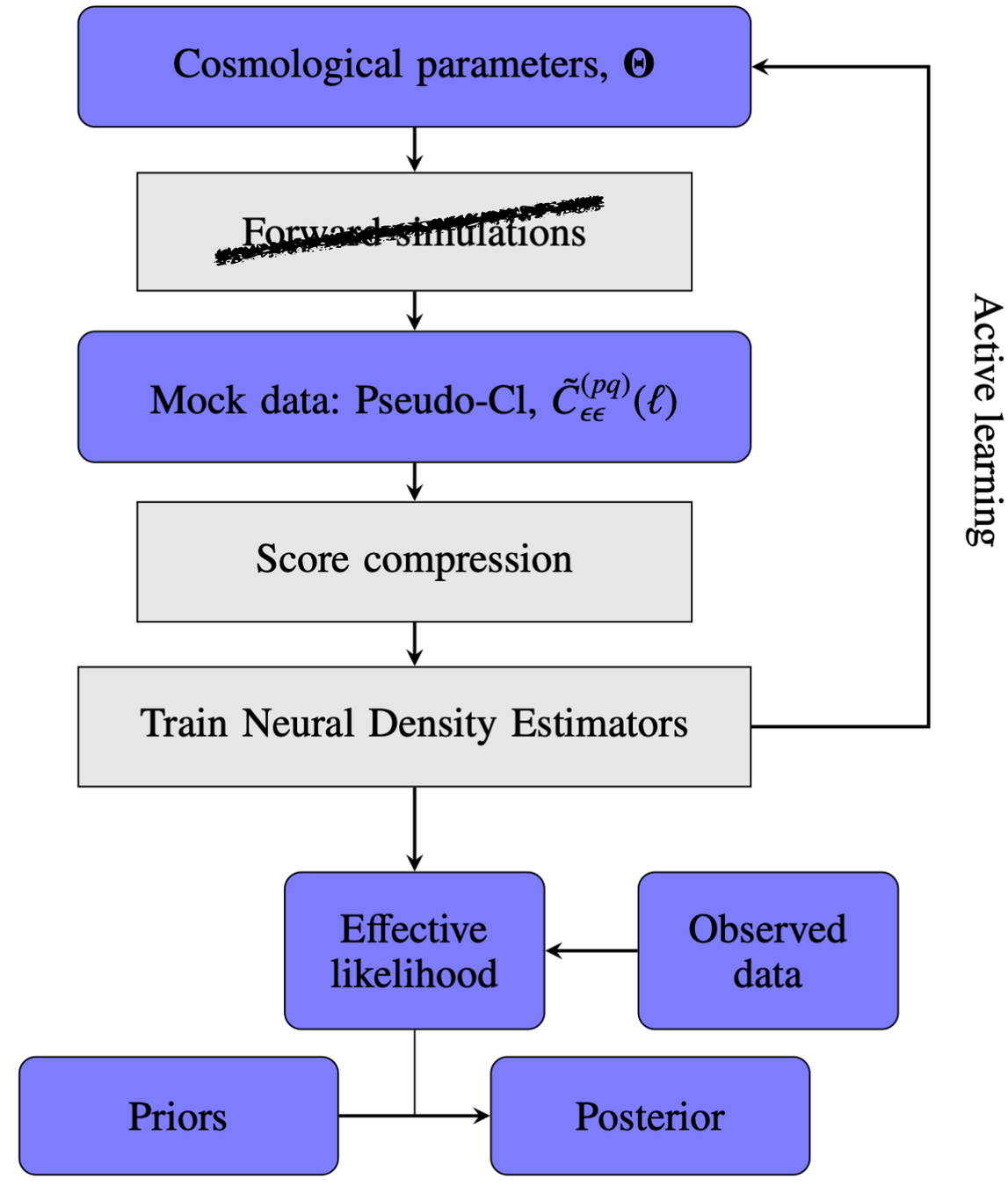
$$p(x|y) = \int \mathrm{d}z \; p(x|y,z)p(z|y)$$

Predict the weights of the mixture model using a CNN then sampled from the resulting distributions to create new samples



HOW THIS WORKS IN PRACTICE

Generative models

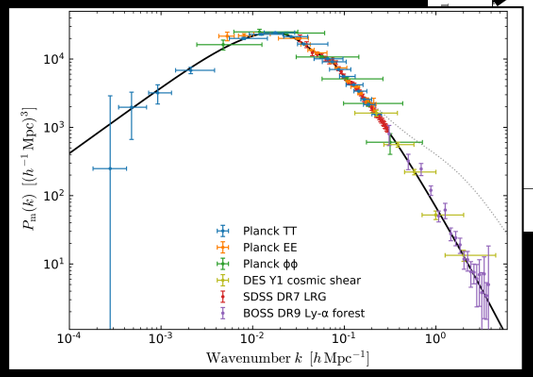


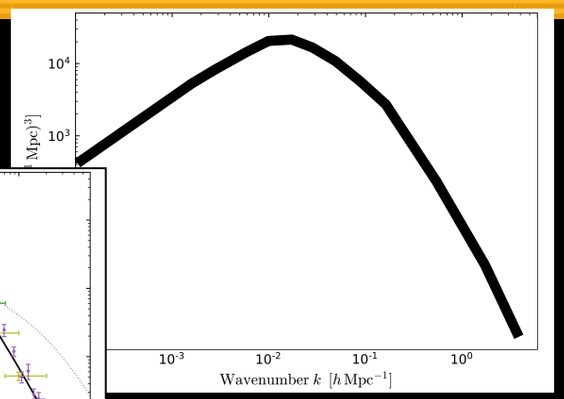
OTHER METHODS FOR GENERATIVE MODELS IN ASTRONOMY:

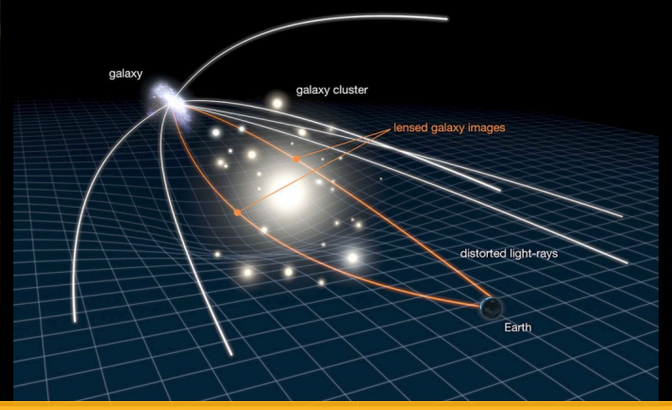
NORMALISING FLOWS (E.G. HTTPS://ARXIV.ORG/PDF/2211.15161, HTTPS://ARXIV.ORG/ABS/2105.12024)

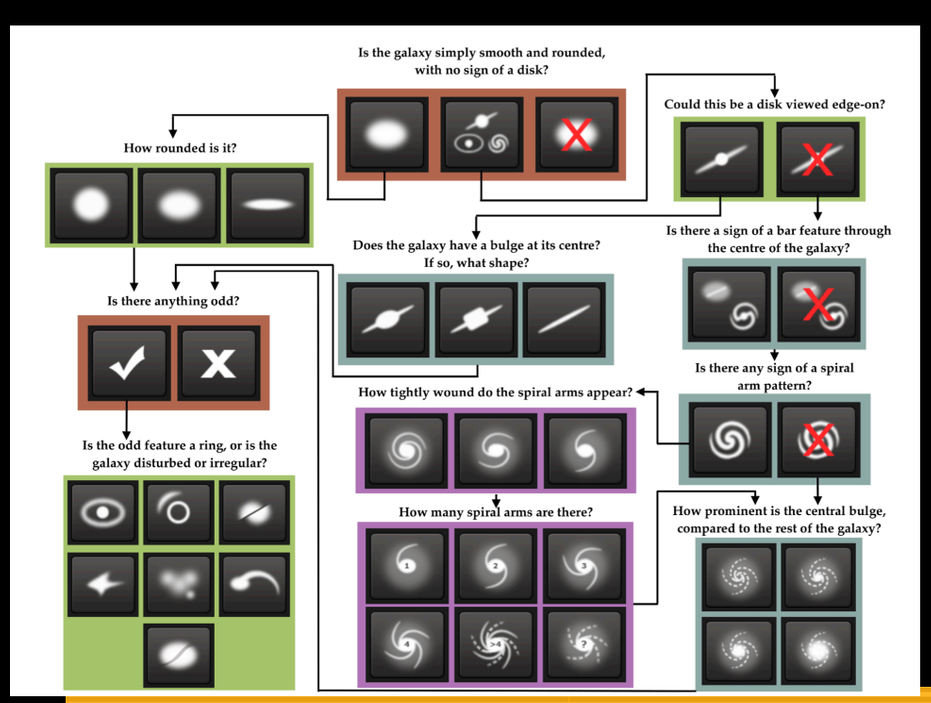
DIFFUSION MODELS (E.G. HTTPS://ARXIV.ORG/PDF/2311.05217)

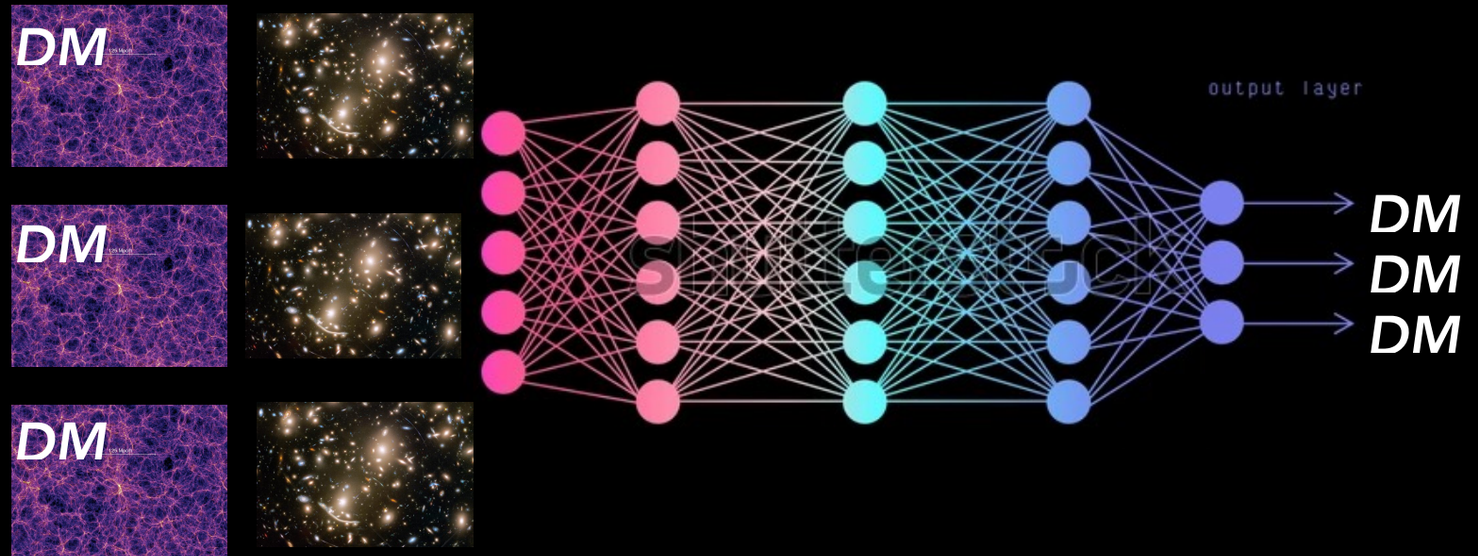
CLASSICAL INFERENCE IS SLOW AND REQUIRES ANALYTICAL MODELS THAT OFTEN LOSE INFORMATION RESULTING IN LOOSE AND BIASED INFERENCE.





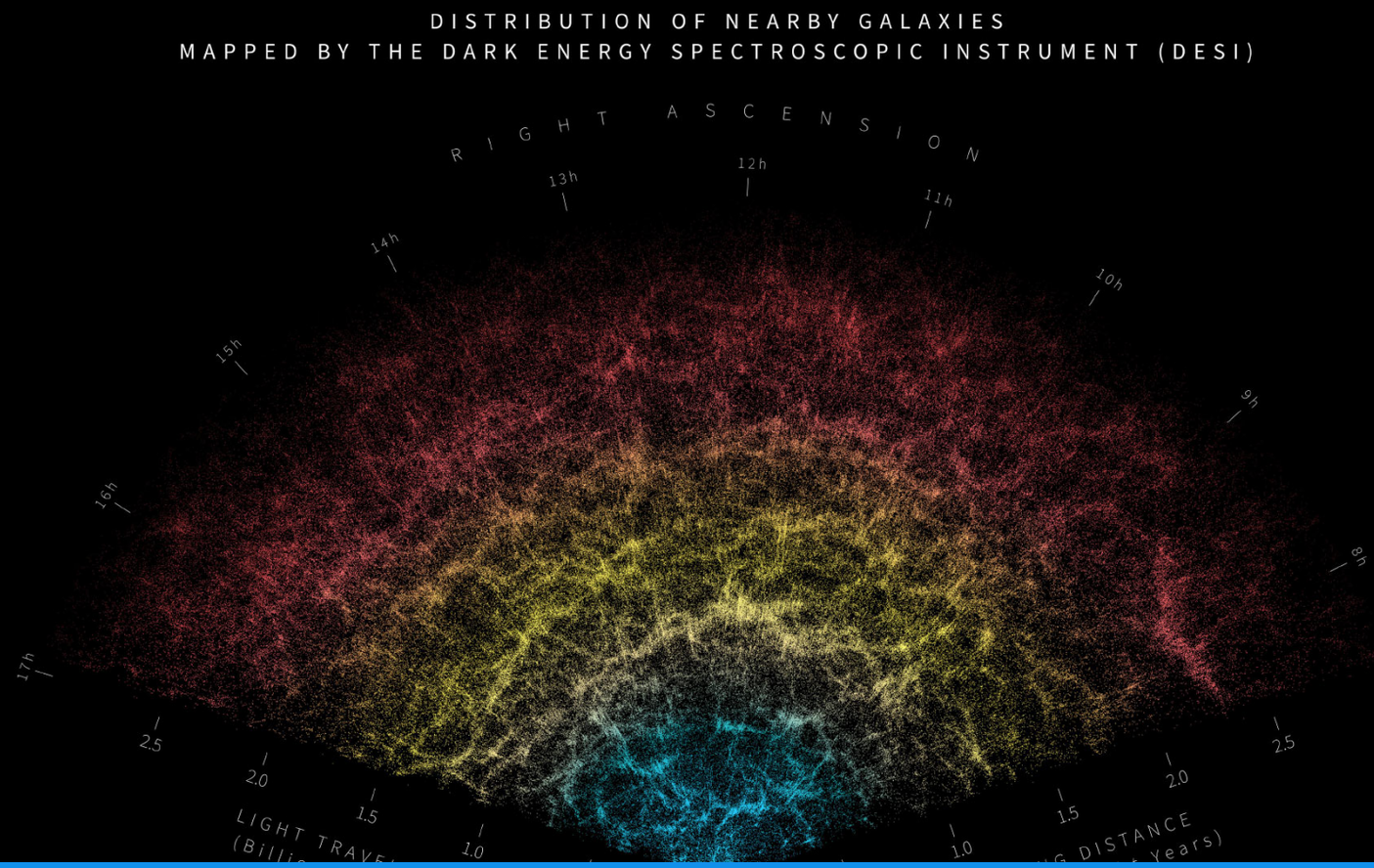




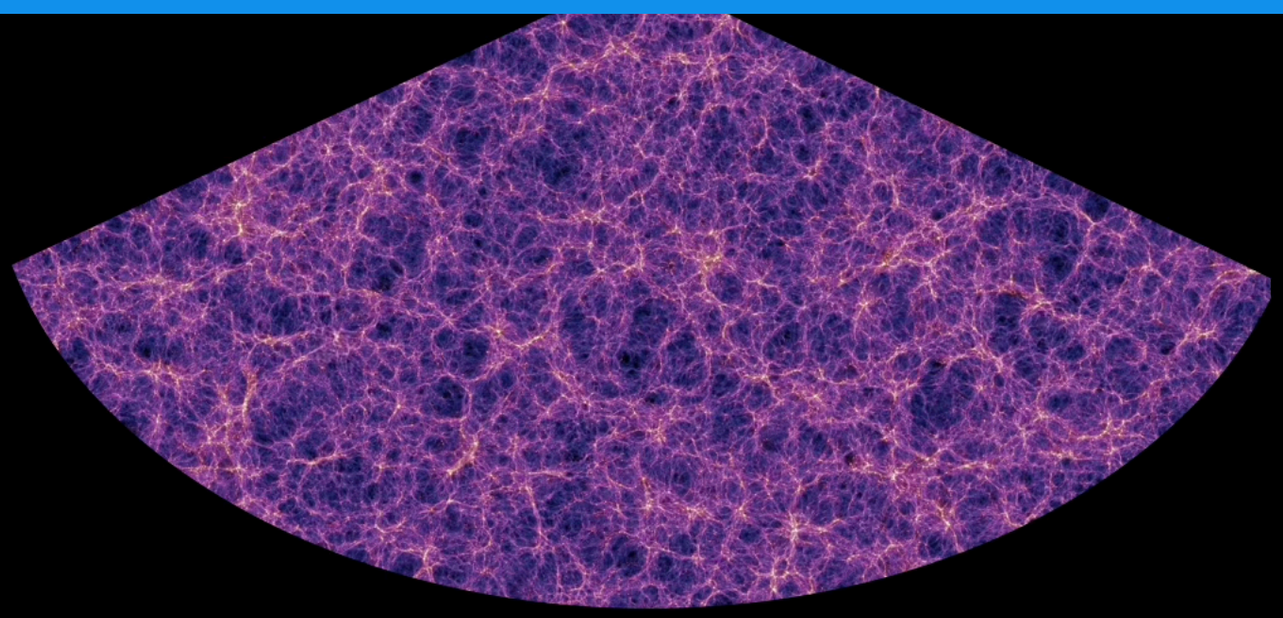


SIMULATION BASED INFERENCE PRESENTS A WAY TO COMPARE DIRECTLY TO SIMULATIONS IN A ROBUST STATISTICAL WAY WITH ADVANCE IN GENERATIVE NETWORKS AIDING INFERENCE

ONE CAN USE MACHINE LEARNING DIRECTLY ON SIMULATIONS TO COMPARE AVOIDING LOSING INFORMATION.



What is the probability of the our model of the Universe given what we observe?



Distribution of matter as seen by the millennium simulation