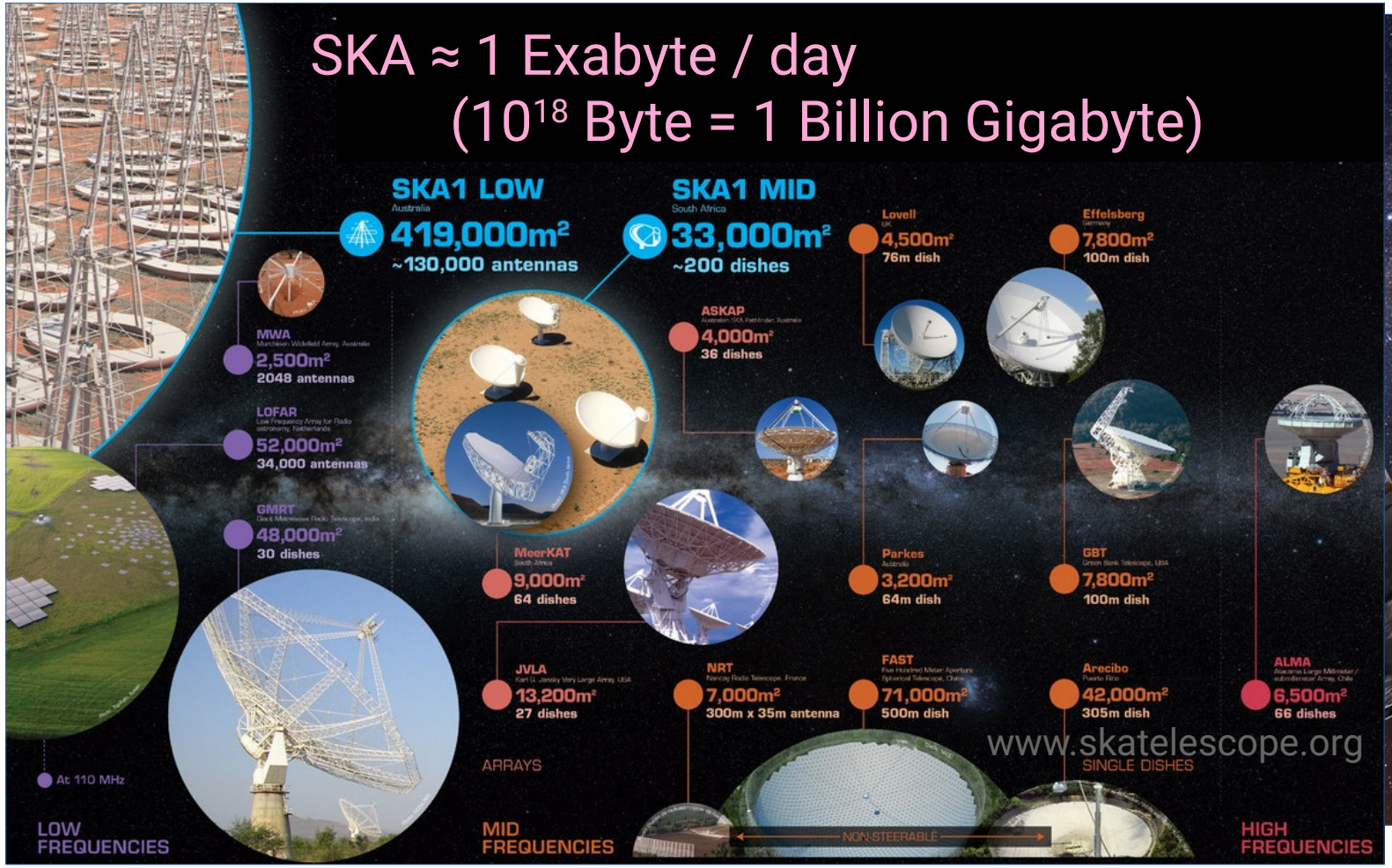
HITS
Heidelberg Institute for Theoretical Studies

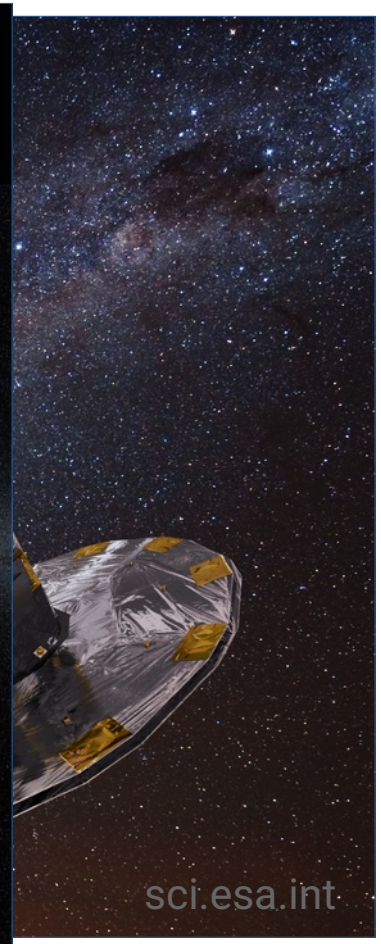
Perspective from RDS:

challenges w.r.t. observational and simulated data

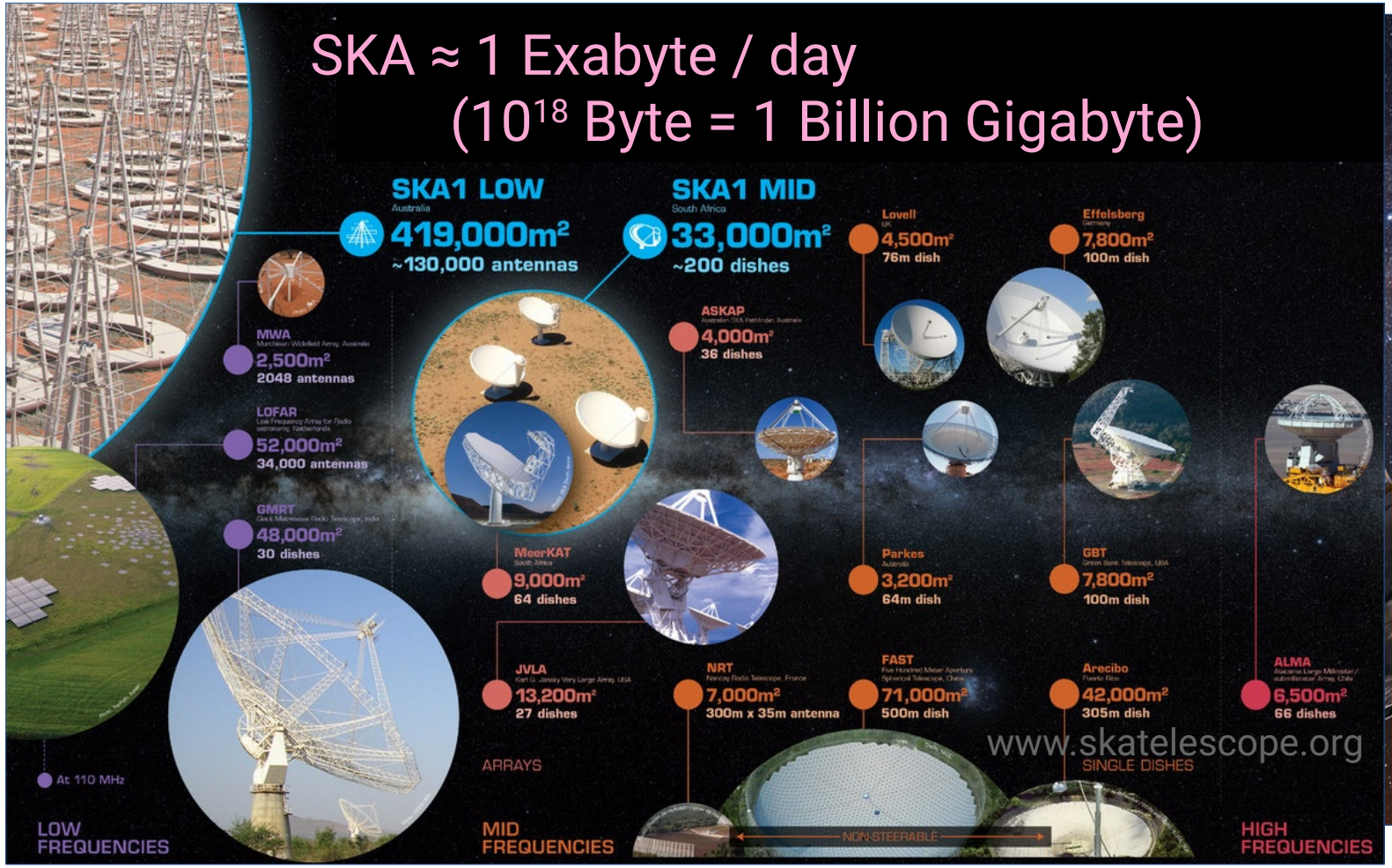
Too much data?

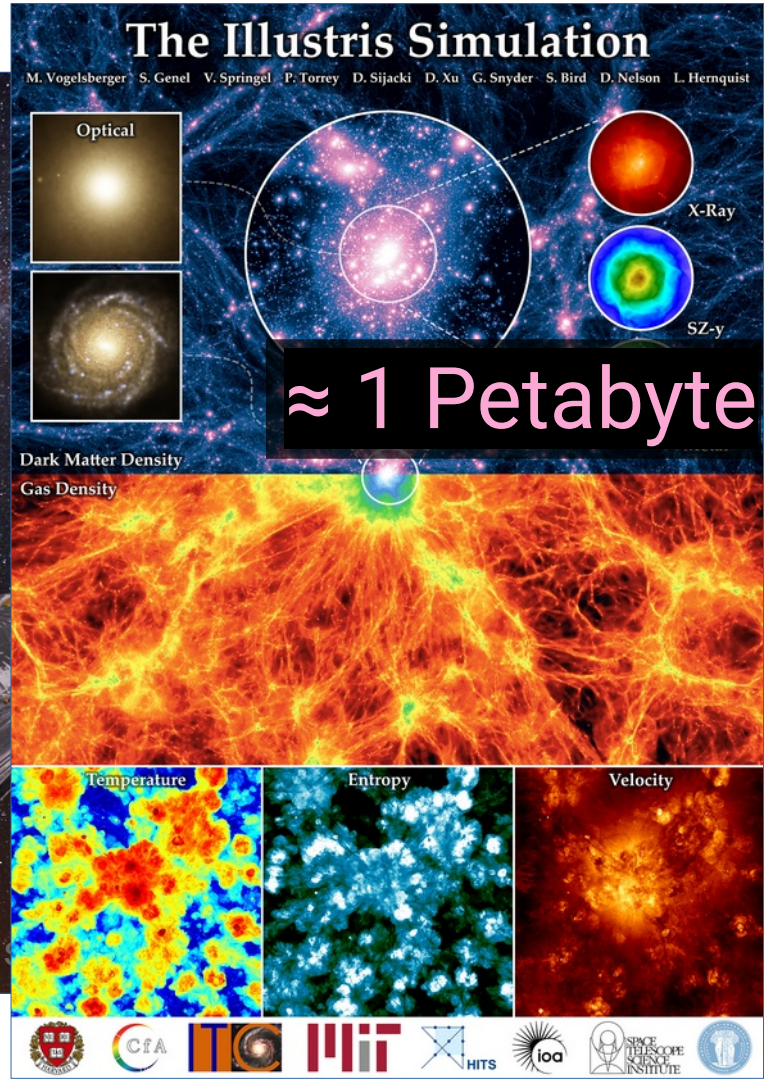
Too much data?





Too much data?





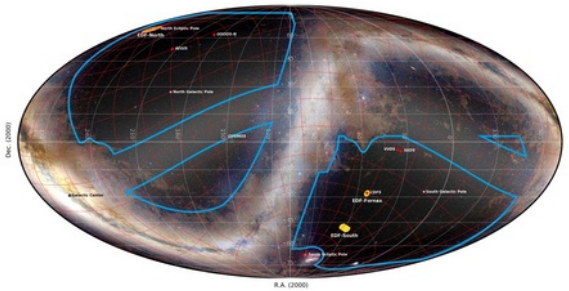
Next-generation survey data

Euclid

~1/3 of the sky up to

Wide survey: 1.5 billion galaxies (photo-z)

Deep survey: 35 million galaxies (spectro-z)



Perseus field (~100k galaxies)

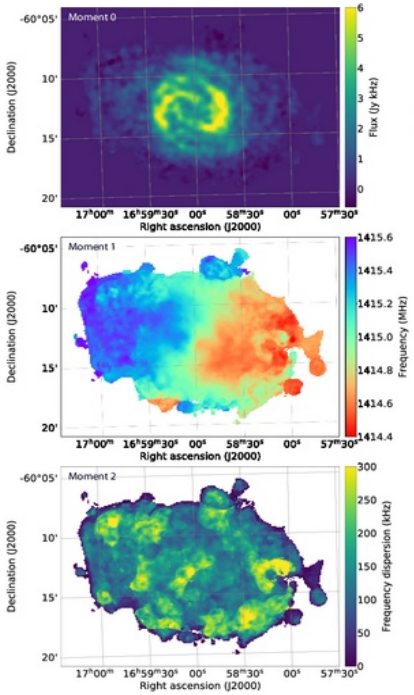
Euclid ERO

Square Kilometer Array (SKA)

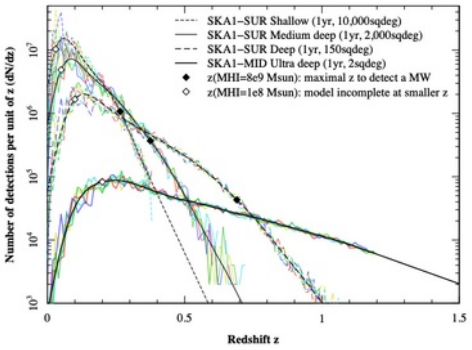
~3/4 of the sky up to

~1 billion galaxies with spectra

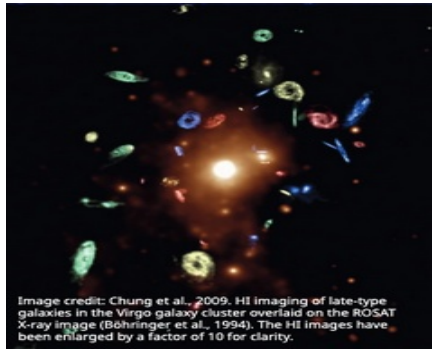
>1 million resolved galaxies in HI



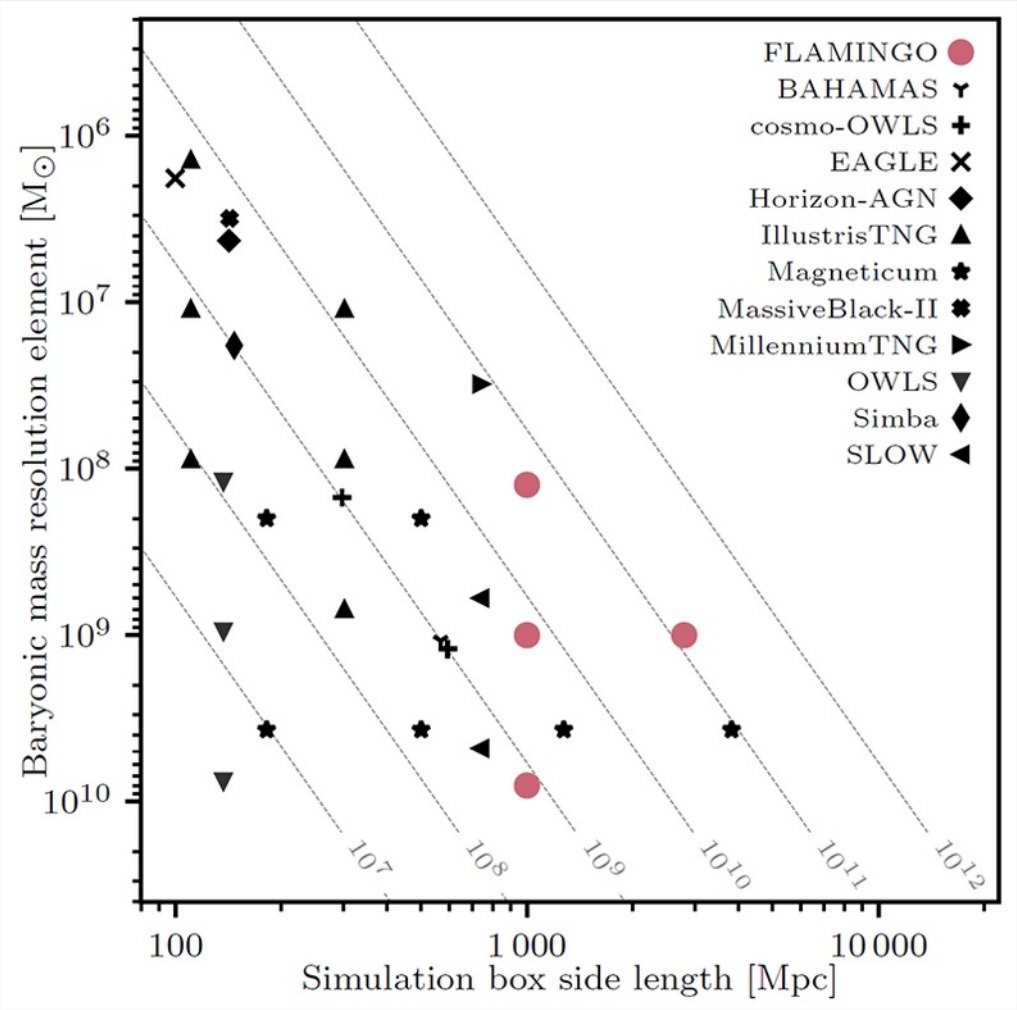
WALLABY (Westmeier+ 2022)



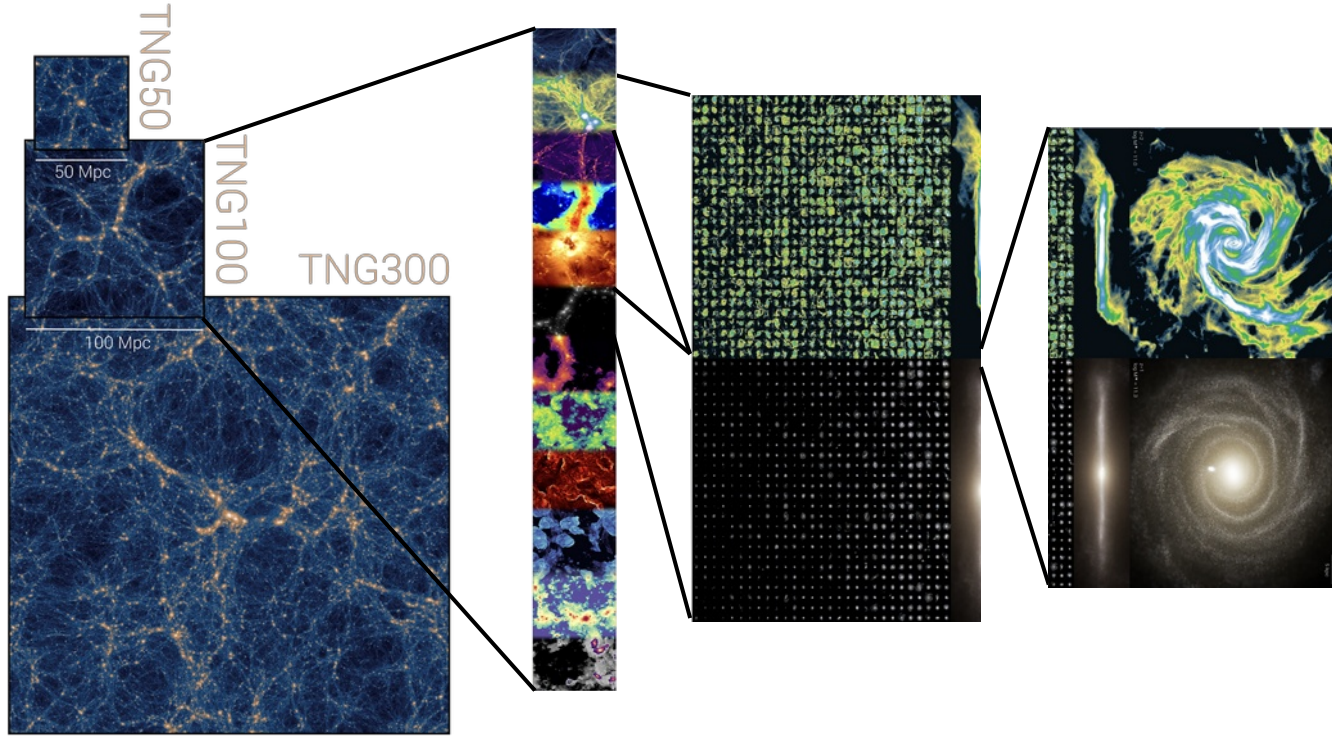
Blyth+ (2014)



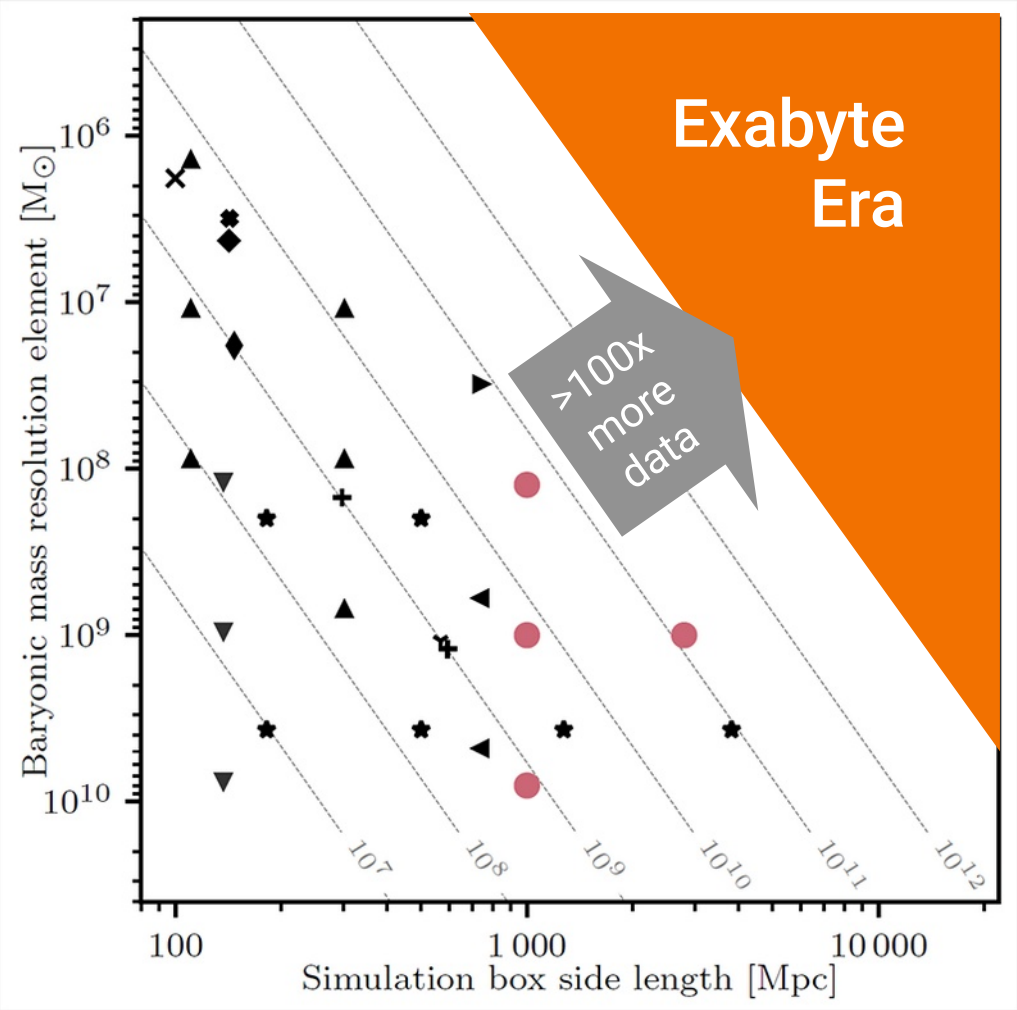
Simulation data



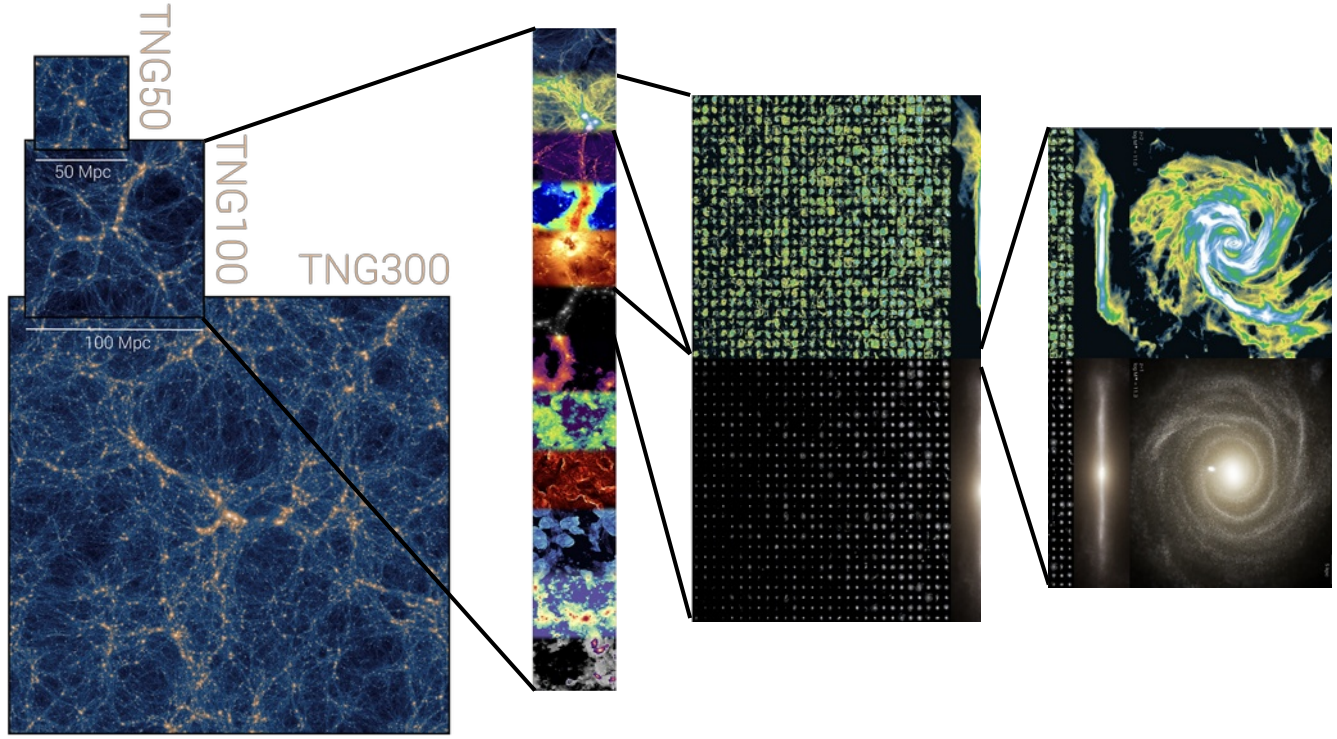
Shaye et al. (2023)



Simulation data

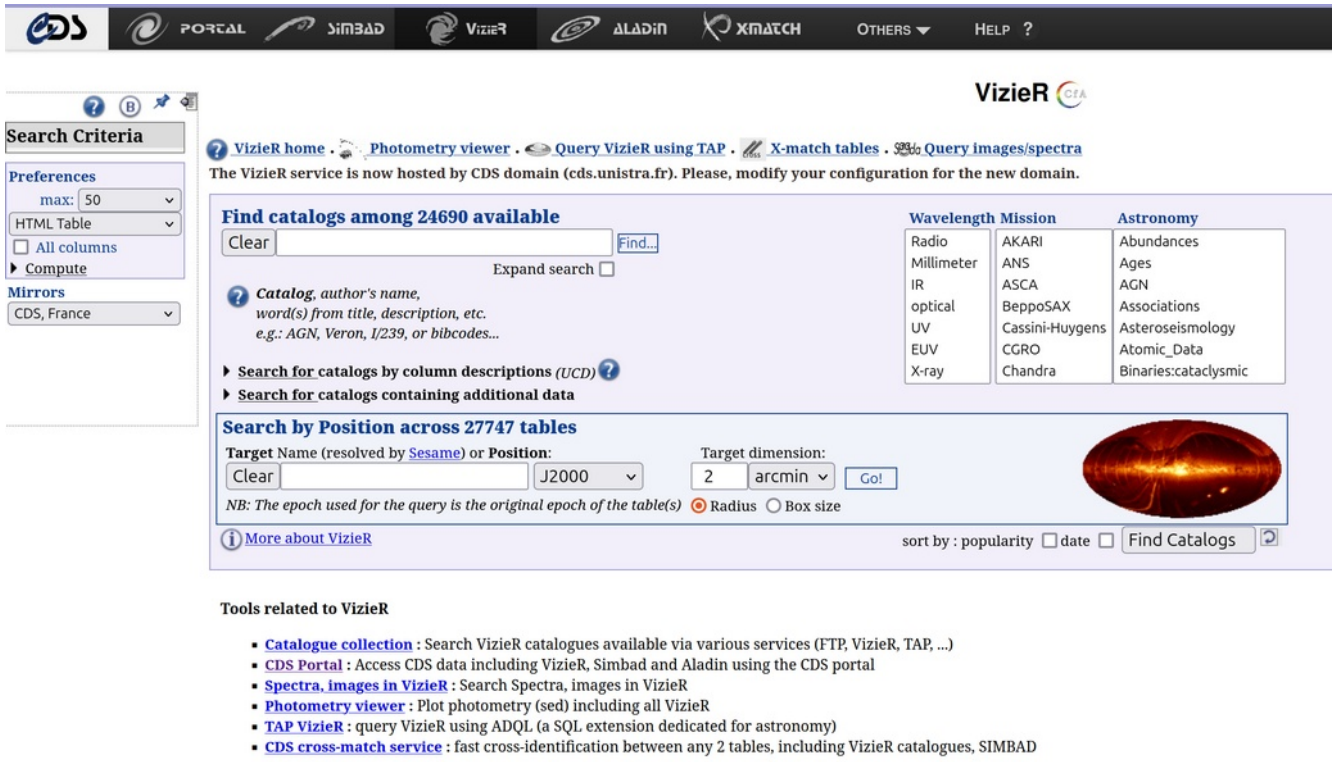


Shaye et al. (2023)



Data access patterns

- Extracting features
  → compressing data
- Storing in databases
  → table oriented representation
- Explicit criteria based access
  → limited in flexibility & UX
- Often single source science
  → no access to complex structures



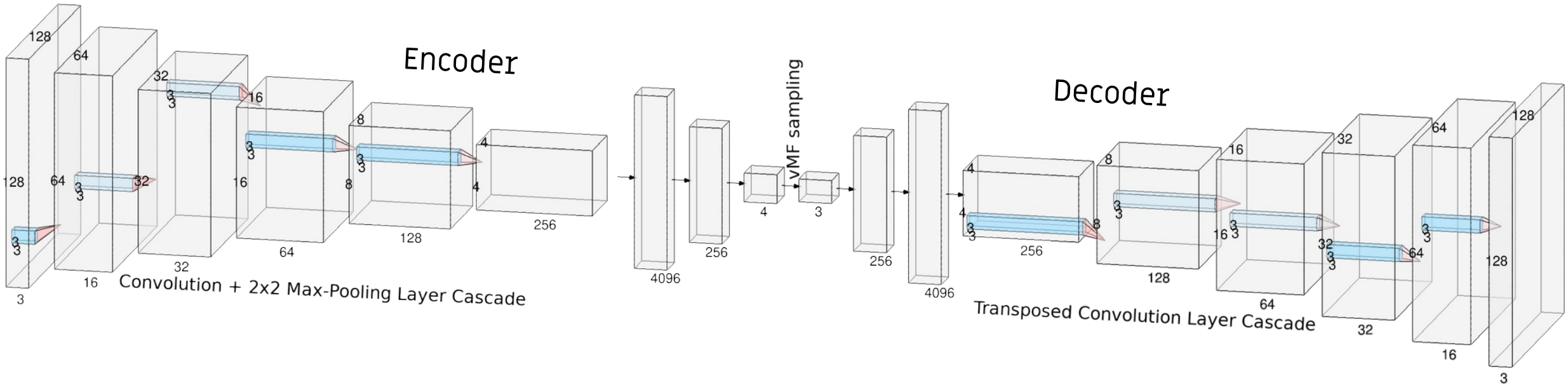
Analysis

ML tools

+

HPC systems

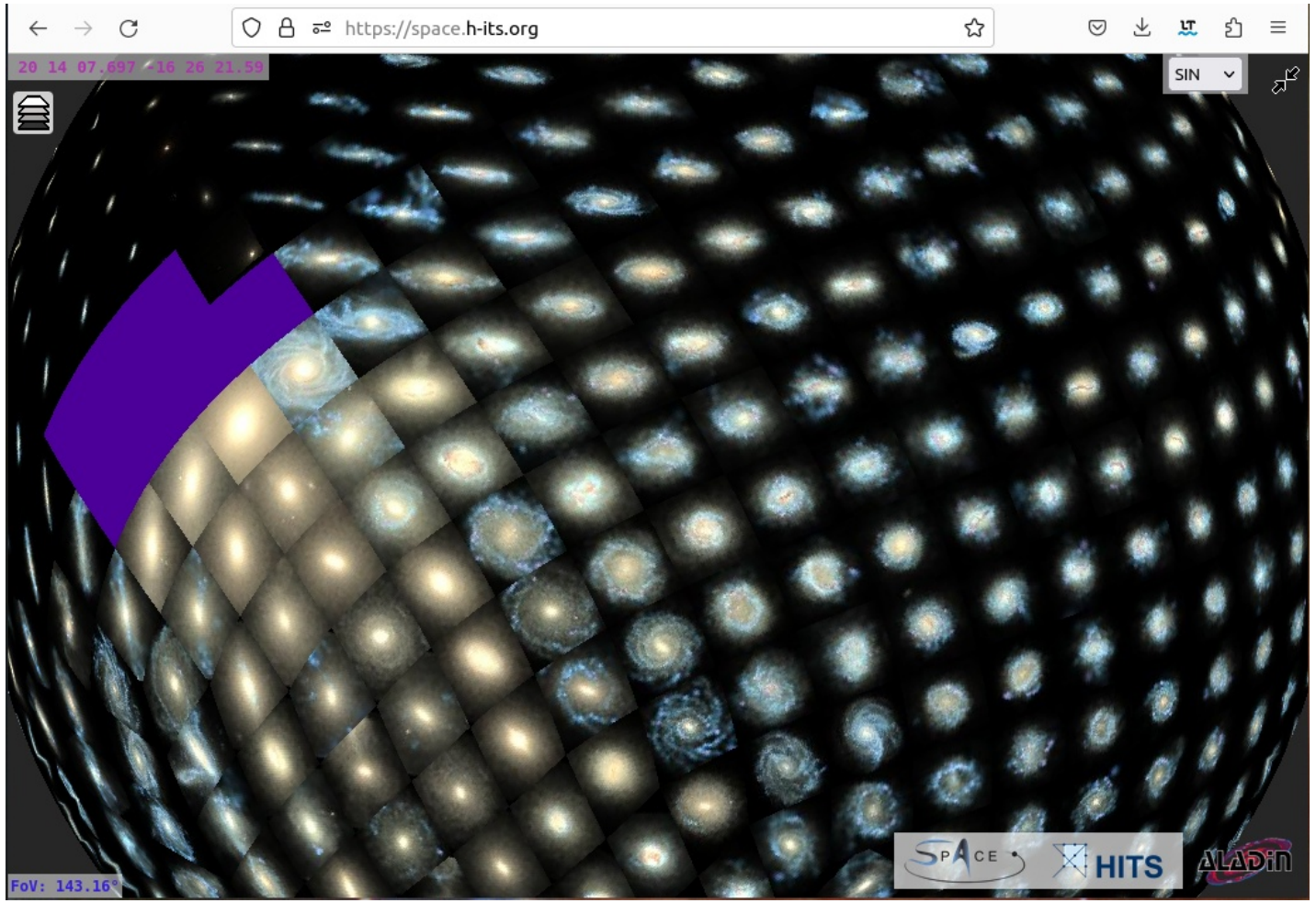
Hyperspherical Variational Convolutional Autoencoder



Representing the Model/Data with HiPS

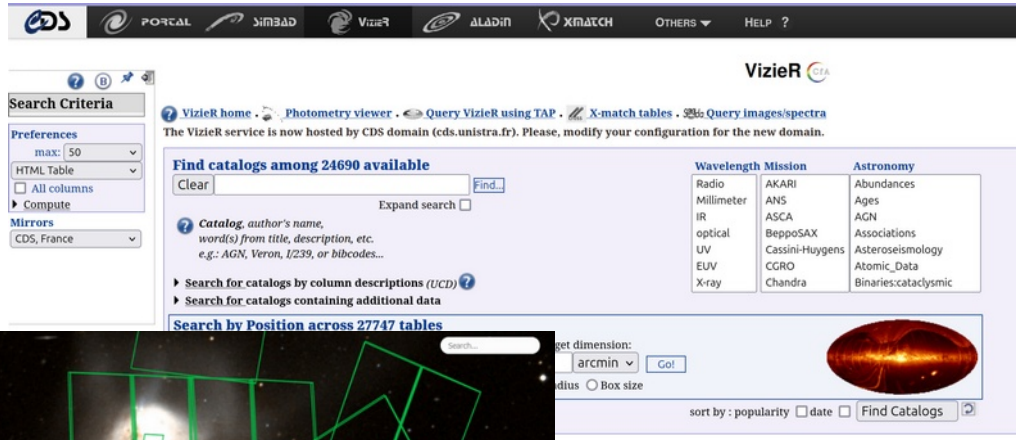
Demo with Aladin Lite :

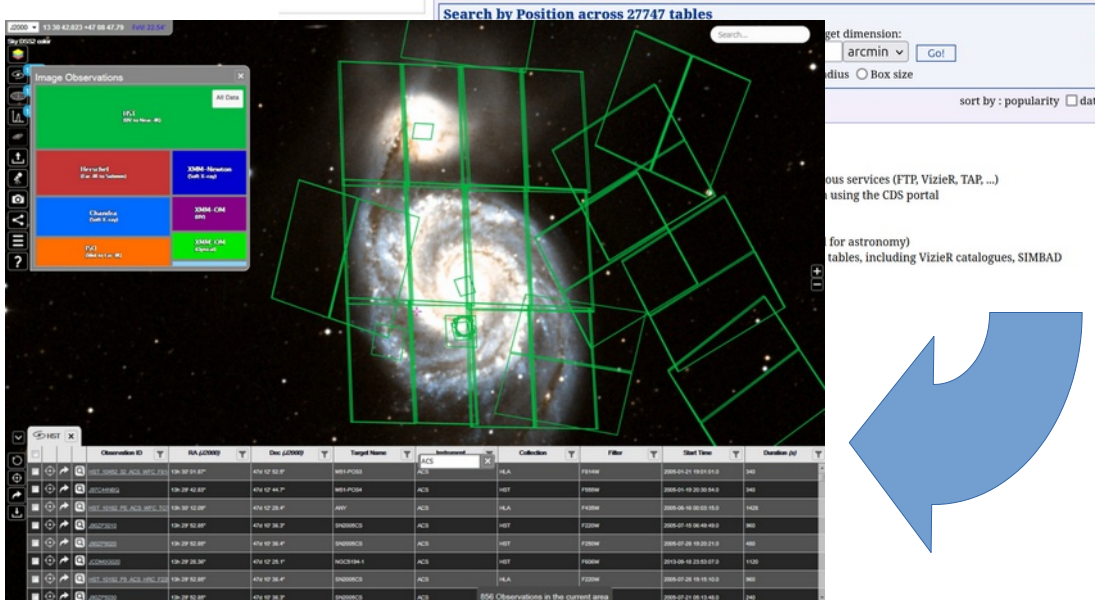
https://space.h-its.org



Access beyond a form

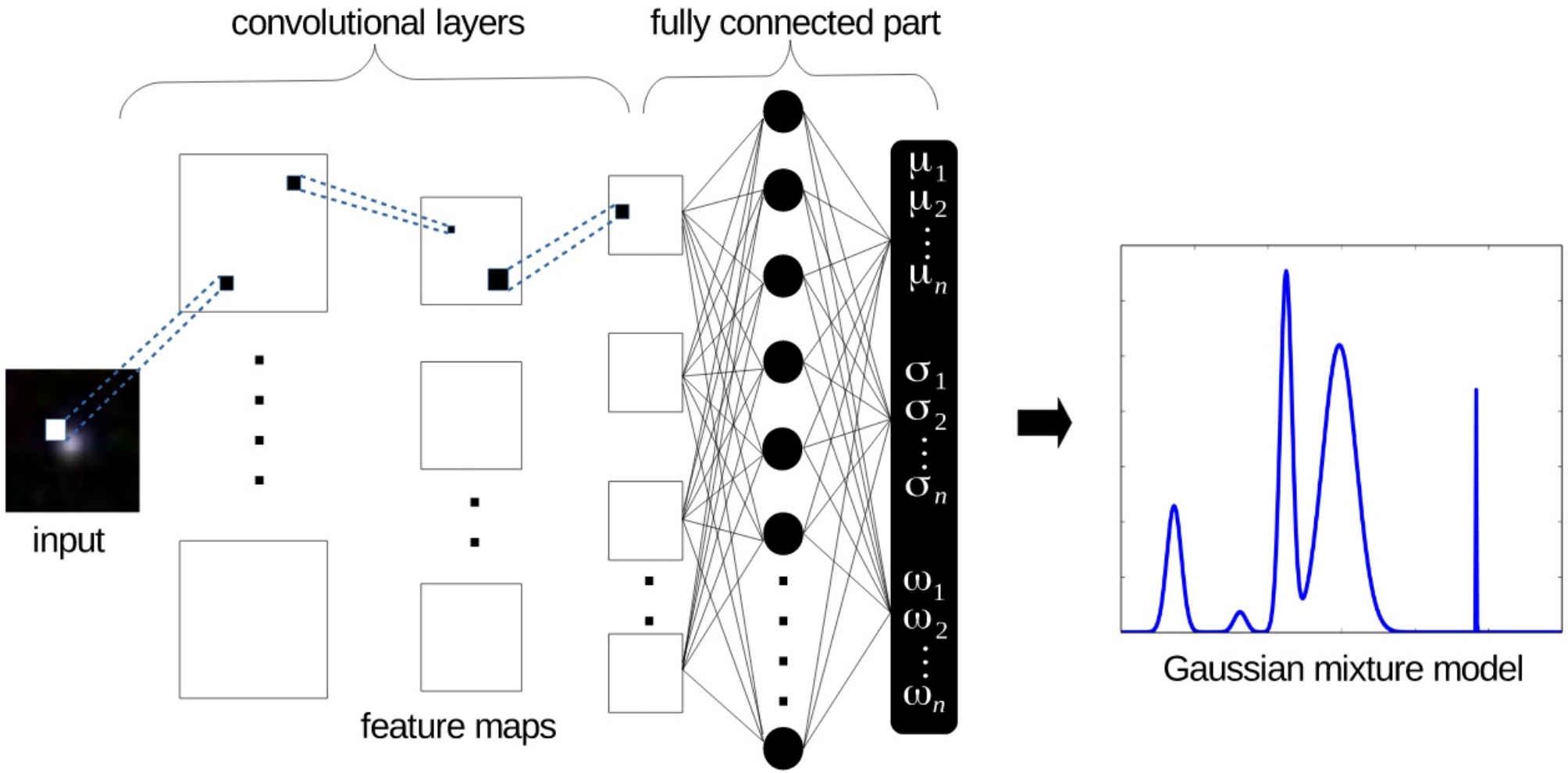
- Intuitive and exploratory
- Search by similarity
- Define and train your own agents
- Linking data with text





What is next?

ML on data → uncertainties



How to preserve uncertainties?

- Store uncertainties/likelihood functions in databases
- Represent uncertainties that allows for indexing
- Allow for queries with uncertainty, e.g.

  Give me all object,
  classified as quasar (with a probability of more than 90%),
  with a redshift larger than 4 (with a probability of more than 80%),
  sorted by joint probability

→ develop database extension / modify ADQL to enable such queries

Workflows / reproducibility

- Jupyter is great for documenting
  → semi-optimal for efficient workflows
- How to preserve data provenance
  → everyone own solution?
- How to exchange/preserve workflows
  → Common Workflow Language?
- How to reproduce higher data generation process
  → (the click/click/click problem)

How can NFDI support GSP?

Critical comments

data volumes → tasks / goals

hardware infrastructure → humans / services

providers perspective on data → scientists perspective / UX

open data / FAIR → just labels? We should be honest

we need more compute and storage → we should talk

sustainability → reuse existing data / less instruments

look at fields independently → identify synergies better

HITS
Heidelberg Institute for Theoretical Studies

Thanks for your attention!

follow us online

@Astroinformatics.bsky.social
@Hitster.bsky.social

@Astroinformatix
@HITStudies

/TheHITSters

/HITStudies

/the_hitsters