

HITS

Heidelberg Institute for
Theoretical Studies

Perspective from RDS:

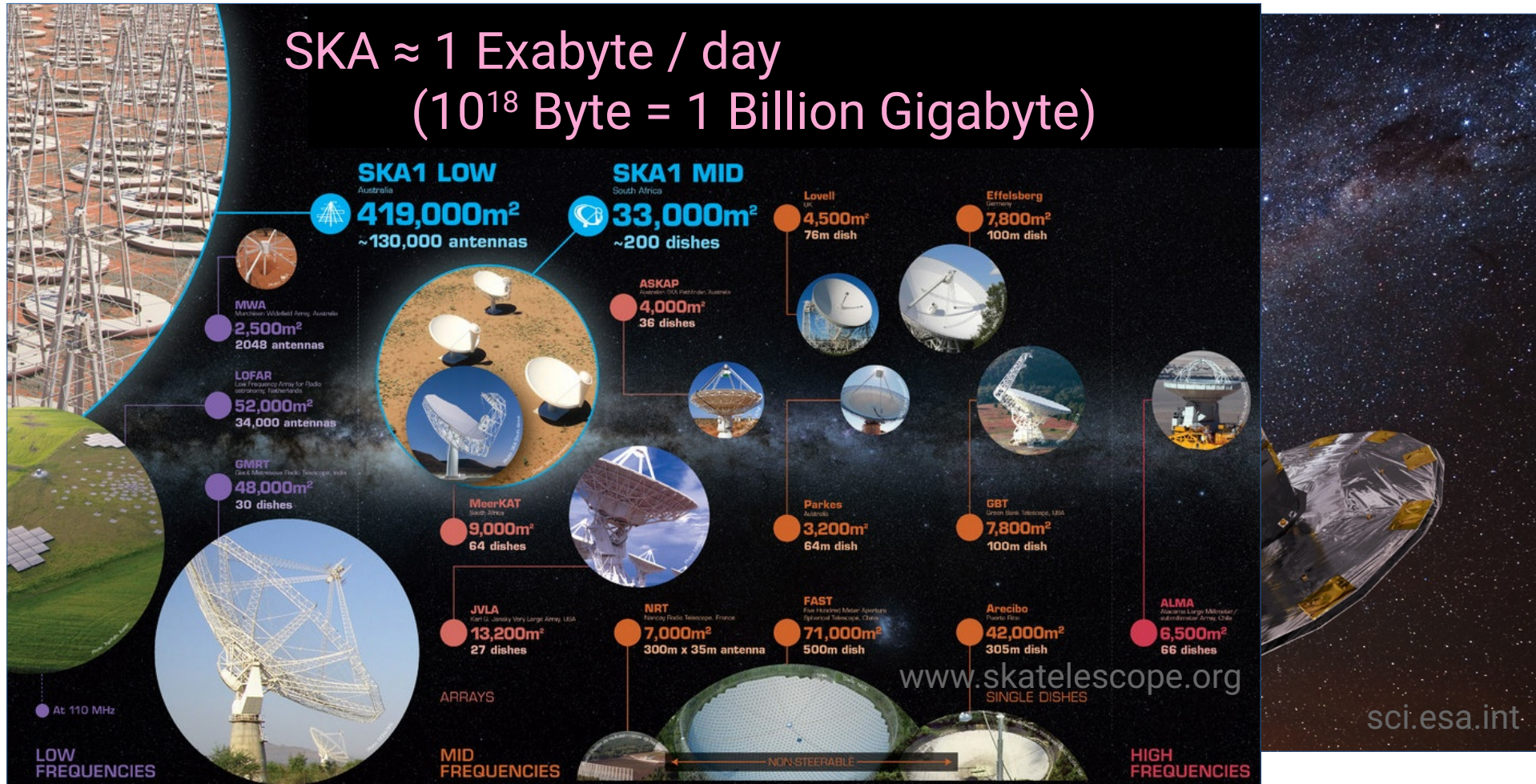
challenges w.r.t. observational and simulated data

Too much data?



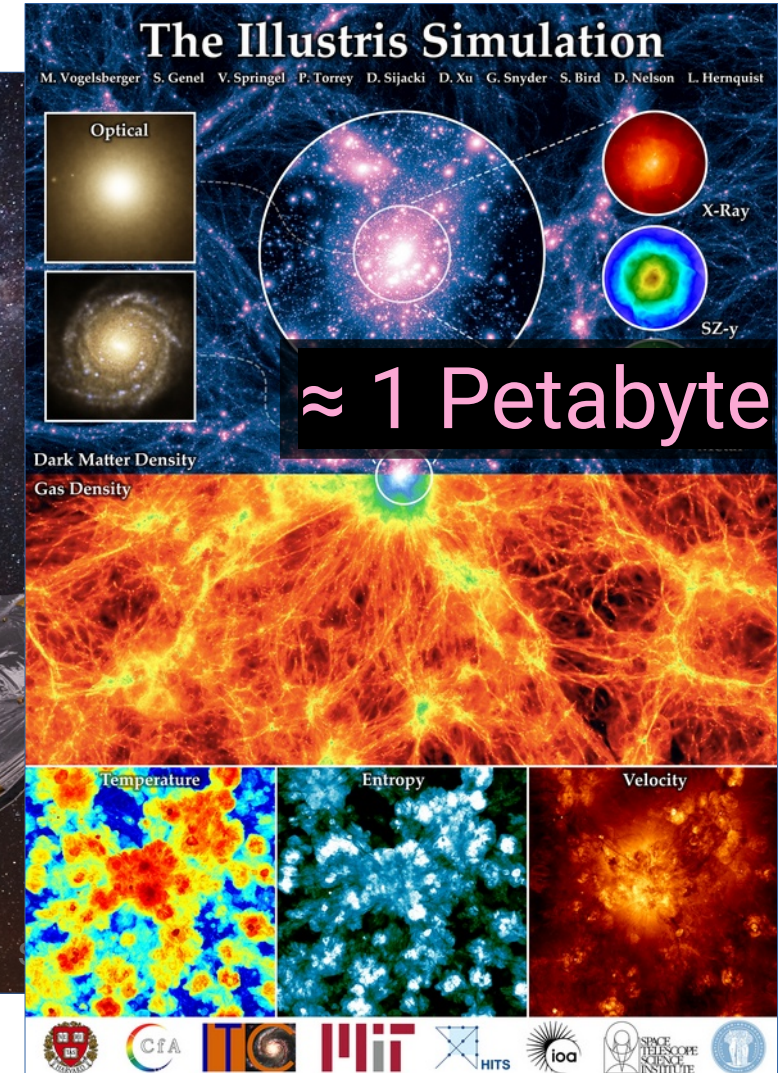
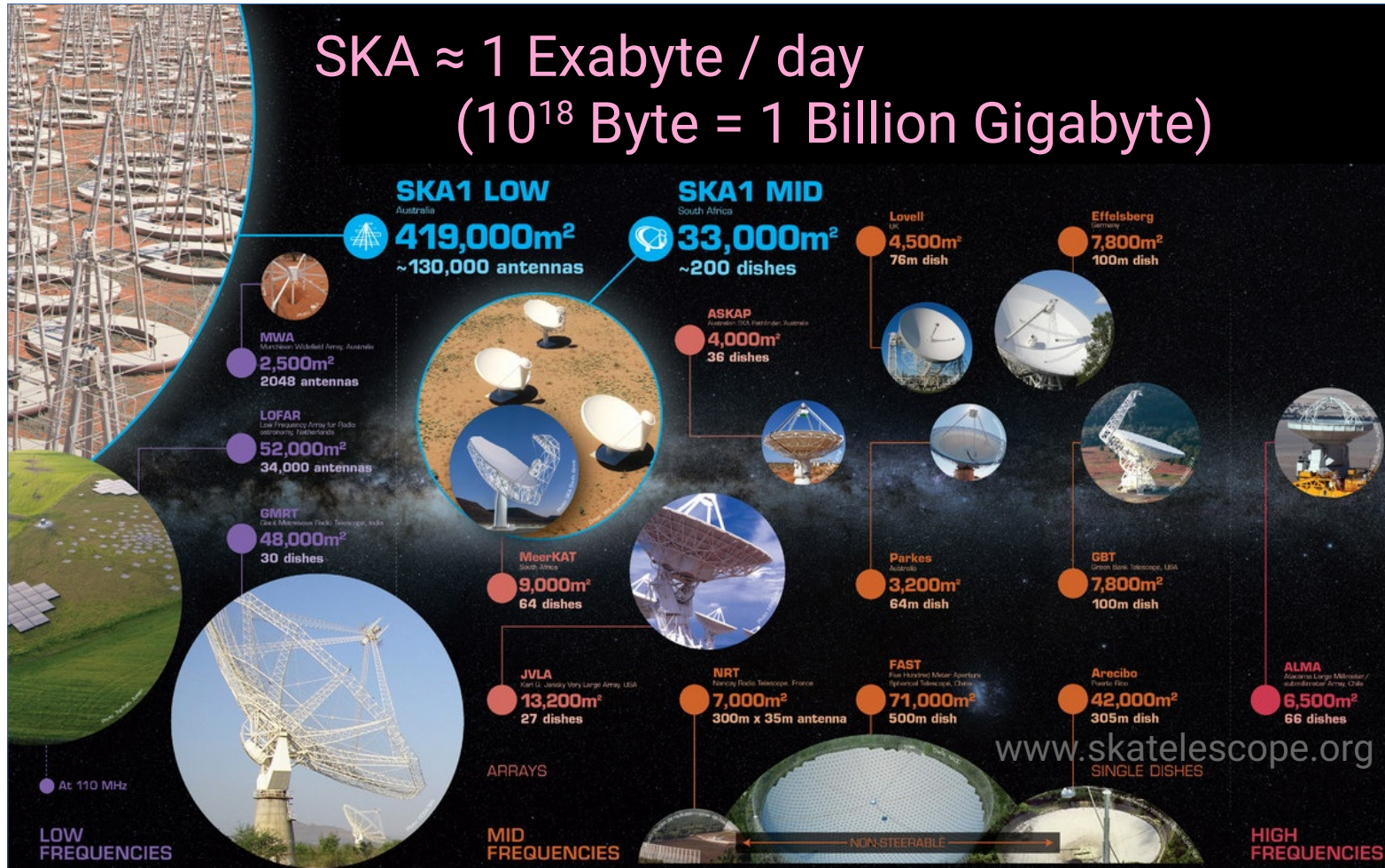
Too much data?

SKA \approx 1 Exabyte / day
(10^{18} Byte = 1 Billion Gigabyte)



Too much data?

SKA \approx 1 Exabyte / day
(10^{18} Byte = 1 Billion Gigabyte)



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)



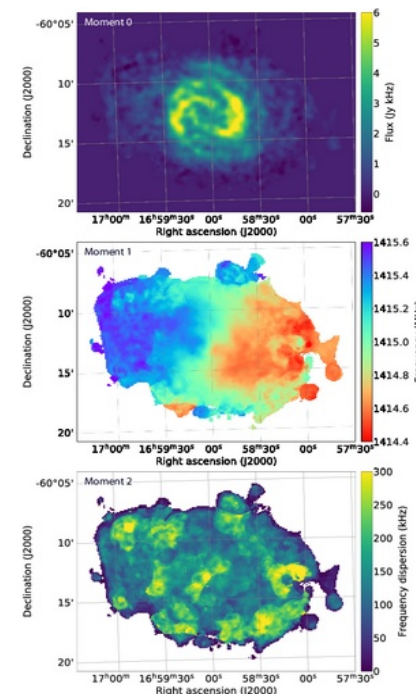
Euclid ERO

Square Kilometer Array (SKA)

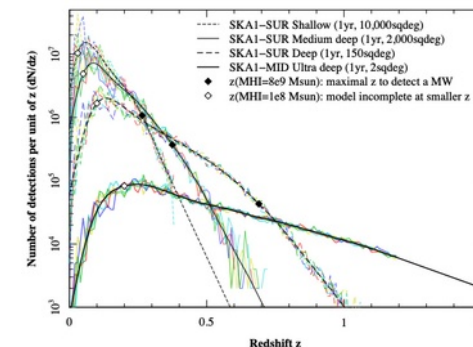
~3/4 of the sky up to

~1 billion galaxies with spectra

>1 million resolved galaxies in HI



WALLABY (Westmeier+ 2022)



Blyth+ (2014)

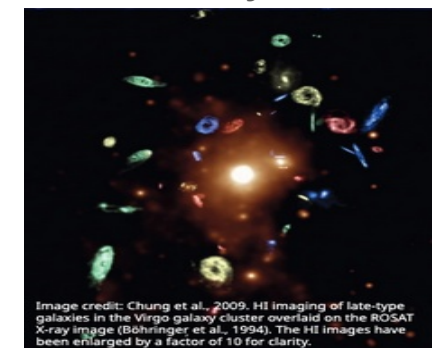
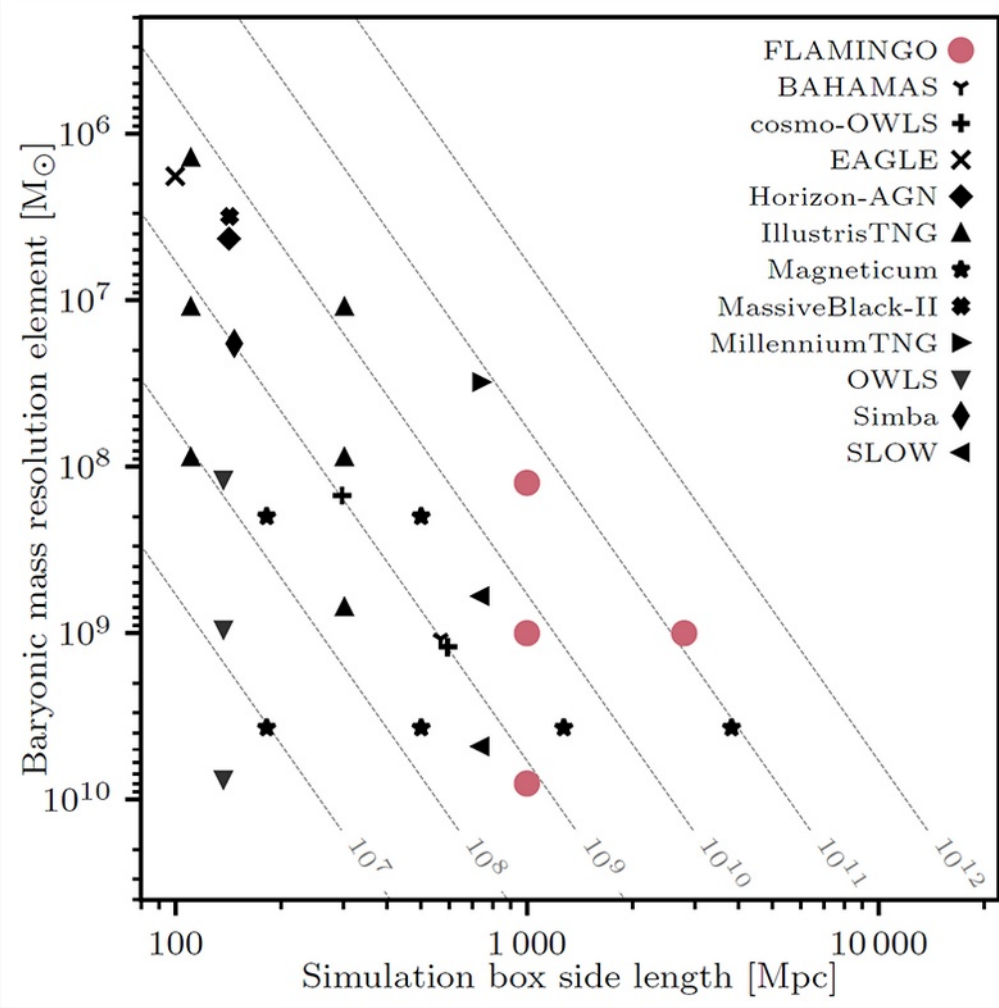
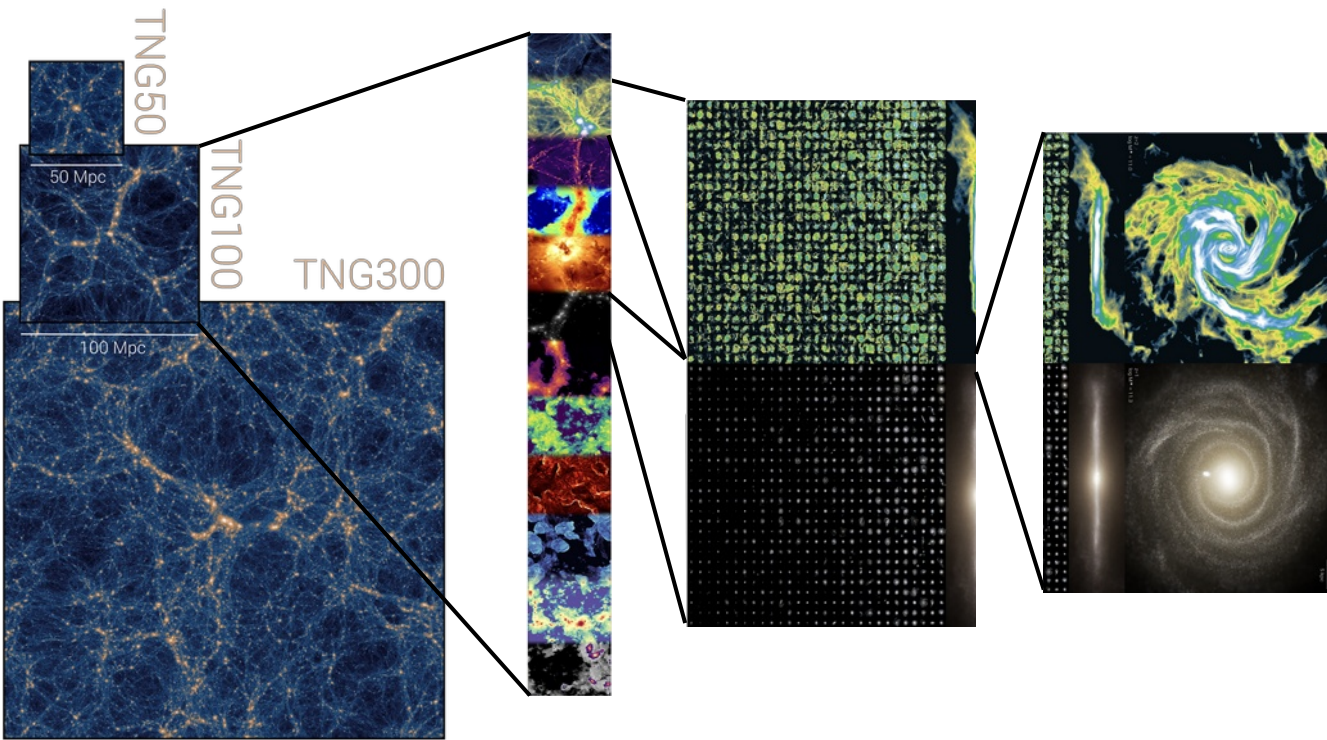


Image credit: Chung et al., 2009. HI imaging of late-type galaxies in the Virgo galaxy cluster overlaid on the ROSAT X-ray image (Böhringer et al., 1994). The HI images have been enlarged by a factor of 10 for clarity.

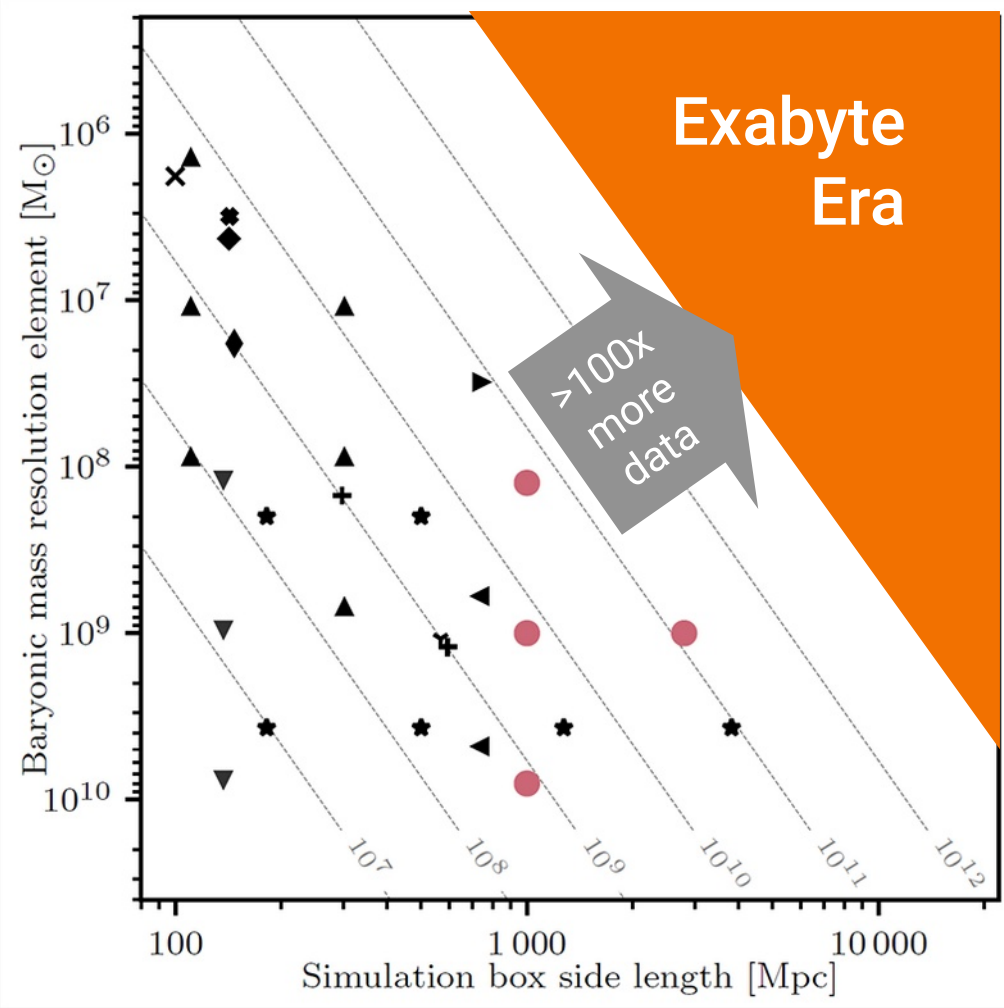
Simulation data



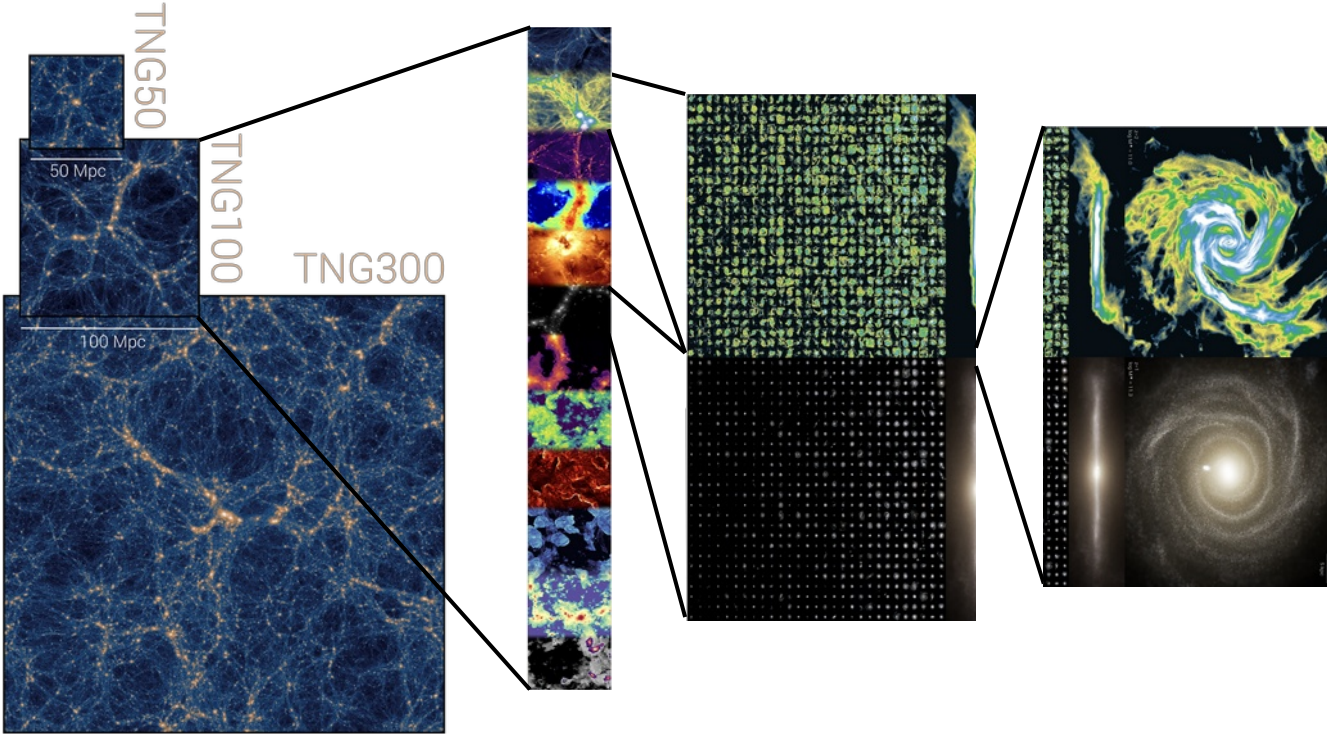
Shaye et al. (2023)



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

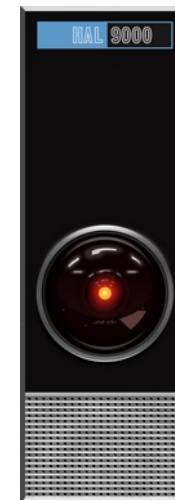
The screenshot shows the VizieR web interface. At the top is a navigation bar with logos for CDS, PORTAL, SIMBAD, VizieR, ALADIN, X-MATCH, and others. Below this is a search bar with the text "Find catalogs among 24690 available". To the left of the search bar is a "Search Criteria" sidebar with sections for "Preferences" (max: 50, HTML Table, All columns, Compute), "Mirrors" (CDS, France), and "Wavelength Mission Astronomy" table. The "Wavelength Mission Astronomy" table lists various astronomical data types and their corresponding missions. Below the search bar is a "Search by Position across 27747 tables" section with fields for "Target Name (resolved by Sesame) or Position", "Target dimension" (2 arcmin), and "Go!" button. A note at the bottom states "NB: The epoch used for the query is the original epoch of the table(s)".

Wavelength	Mission	Astronomy
Radio	AKARI	Abundances
Millimeter	ANS	Ages
IR	ASCA	AGN
optical	BeppoSAX	Associations
UV	Cassini-Huygens	Asteroseismology
EUV	CGRO	Atomic_Data
X-ray	Chandra	Binaries:catclysmic

Tools related to VizieR

- [Catalogue collection](#) : Search VizieR catalogues available via various services (FTP, VizieR, TAP, ...)
- [CDS Portal](#) : Access CDS data including VizieR, Simbad and Aladin using the CDS portal
- [Spectra, images in VizieR](#) : Search Spectra, images in VizieR
- [Photometry viewer](#) : Plot photometry (sed) including all VizieR
- [TAP VizieR](#) : query VizieR using ADQL (a SQL extension dedicated for astronomy)
- [CDS cross-match service](#) : fast cross-identification between any 2 tables, including VizieR catalogues, SIMBAD

Analysis

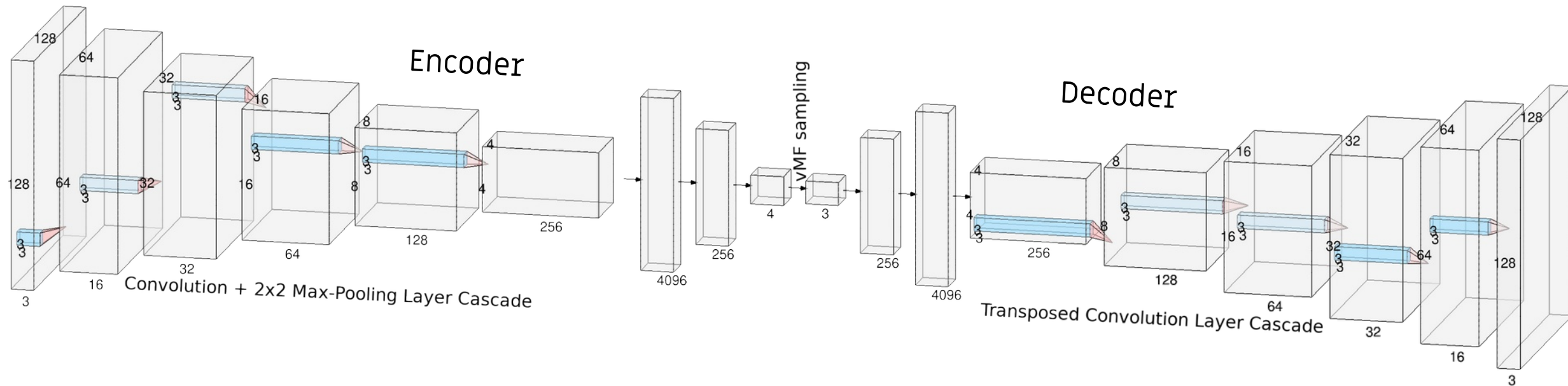


ML
tools



HPC systems

Hyperspherical Variational Convolutional Autoencoder

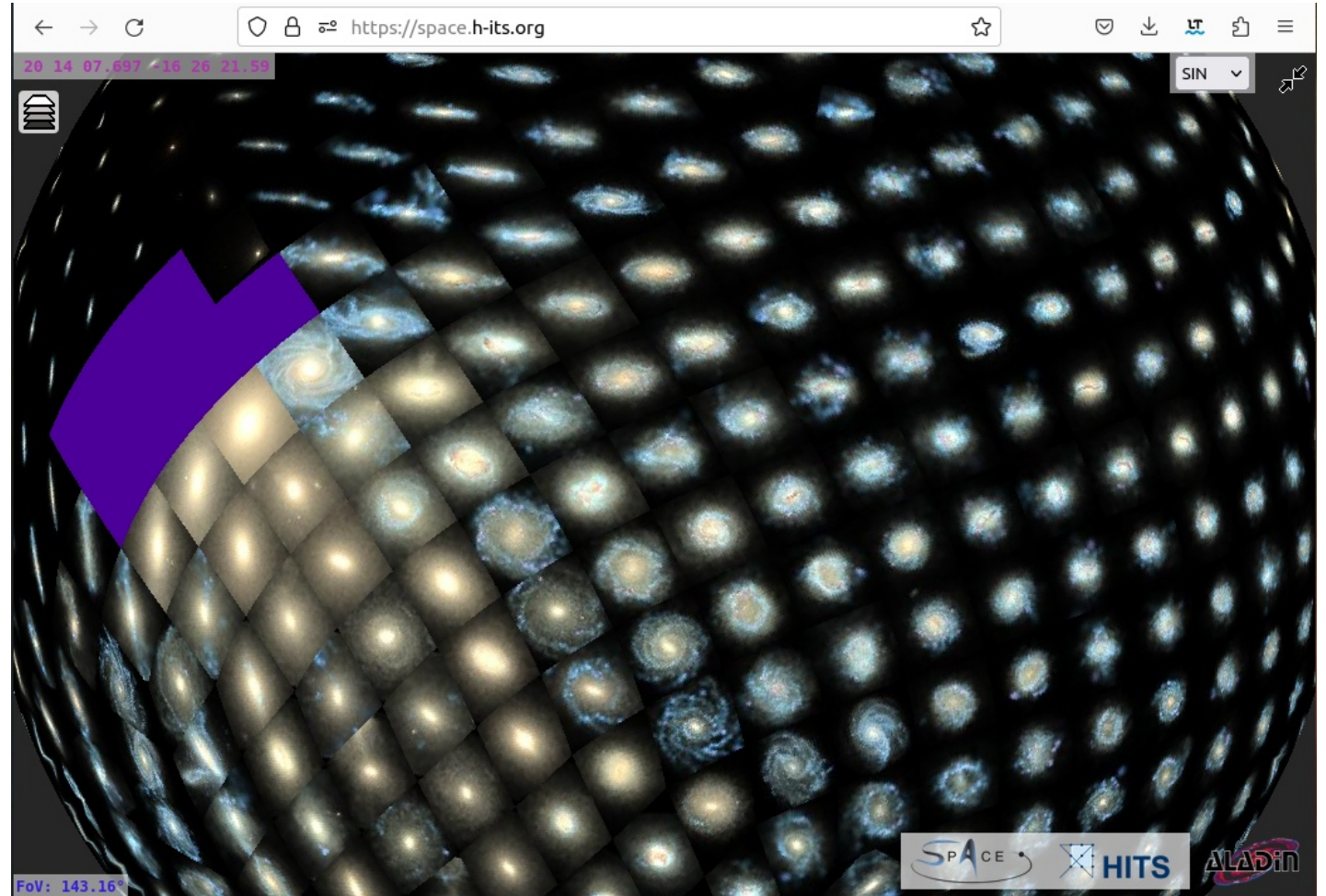


Representing the Model/Data with HiPS



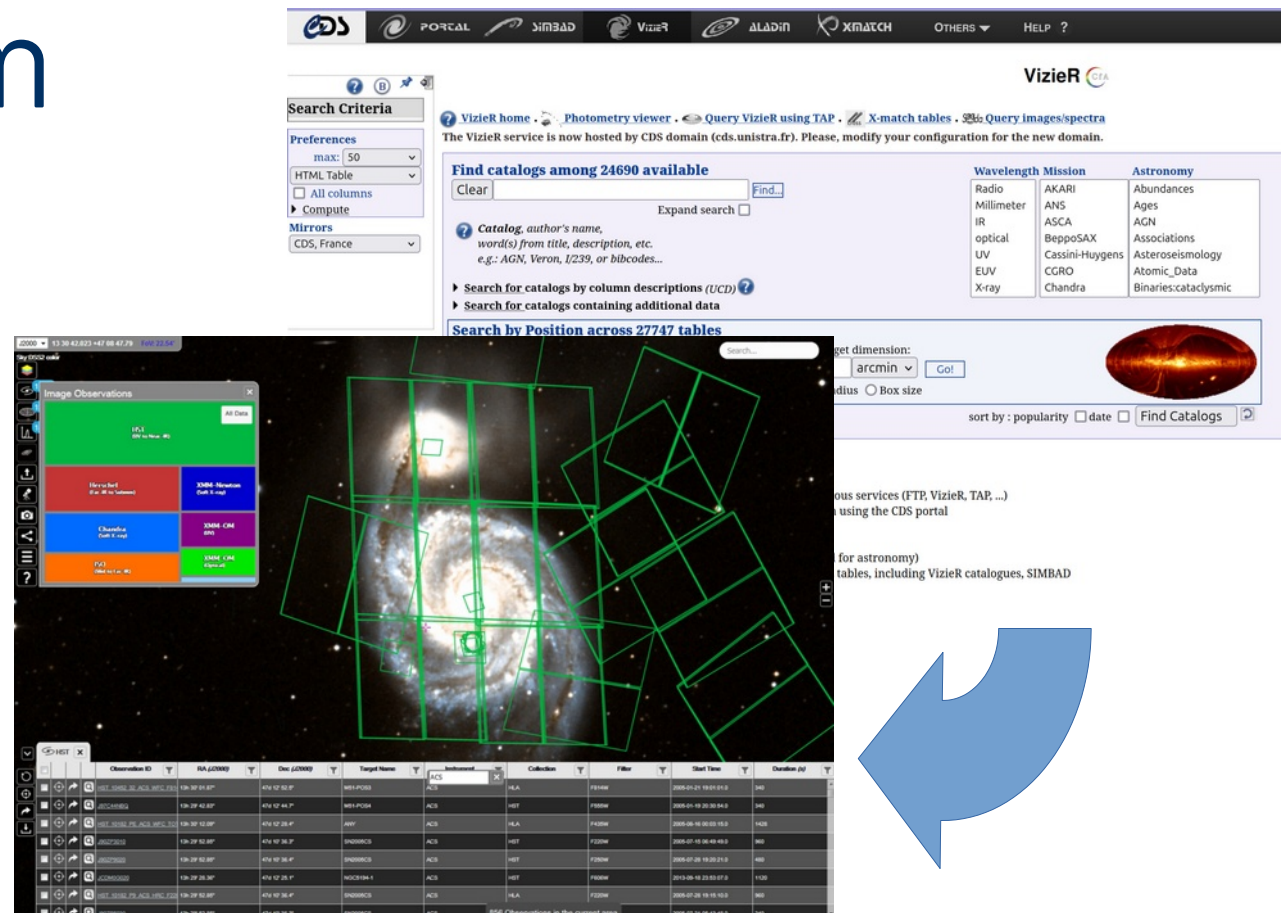
Demo with Aladin Lite :

<https://space.h-its.org>



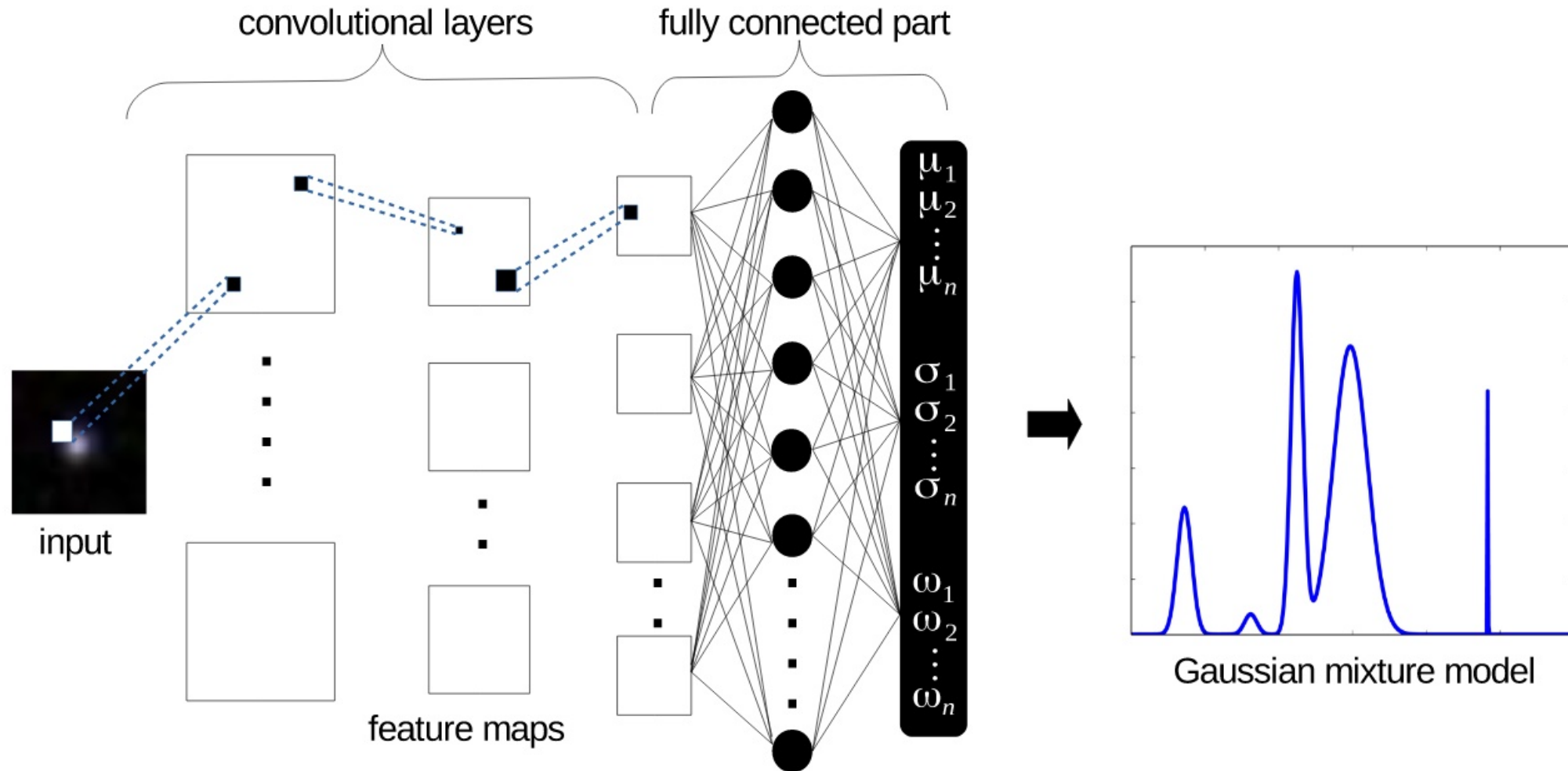
Access beyond a form

- 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





ϕ with $d(A, \phi(B)) = 0$ and thus $A = \phi(B)$ (i.e., $A \sim B$) because d is a metric.

Symmetry: $\Delta(A, B) = \min\{d(A, \phi(B)) | \phi \in \Phi\}$

Thanks for your attention!

follow us online

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

 @Astroinformaticx
@HITStudies

 /TheHITSters

Thanks for your attention!

follow us online



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



@Astroinformatix
@HITStudies



/TheHITSters



/HITStudies



/the_hitsters

follow us online

