HITS
Heidelberg Institute for Theoretical Studies

SPACE

Interpreting the largest cosmological simulations using Representation Learning

Sebastian Trujillo-Gomez
Heidelberg Institute for Theoretical Studies (HITS)

team:
Kai Polsterer, Bernd Doser, Andreas Fehlner, Fenja Schweder, Romain Chazotte, Renuka Velu

Co-funded by the European Union
EuroHPC Joint Undertaking

MACHINE LEARNING FOR ASTROPHYSICS
2ND EDITION
CATANIA, 8-12 JULY, 2024

What are cosmological simulations?

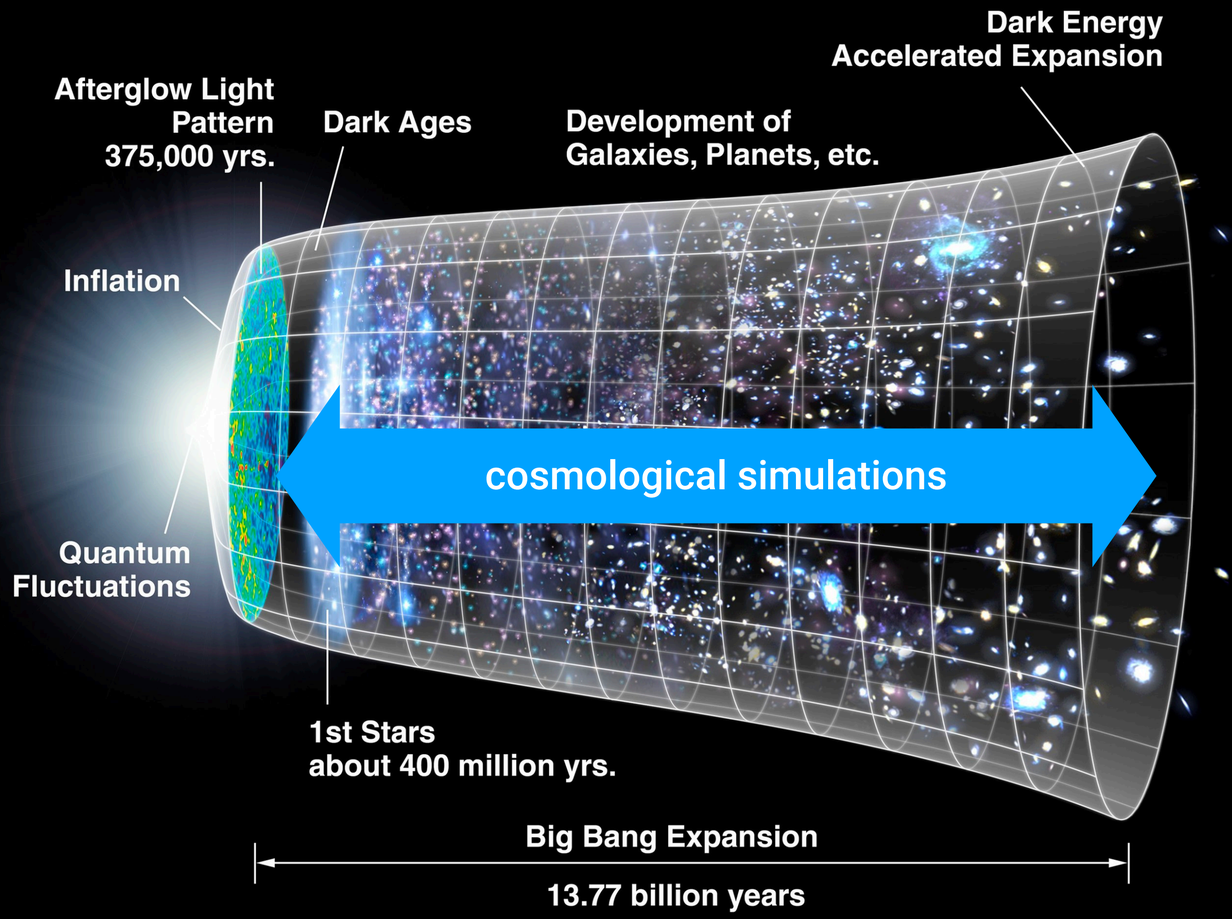
"… a disheartening number of ingredients must be assembled to produce a plausibly complete recipe for galaxy formation"

White & Frenk (1991)

"What I cannot create,
I do not understand"

Richard Feynman

cosmological simulations attempt to build a Universe in a computer using the known laws of physics



gravity

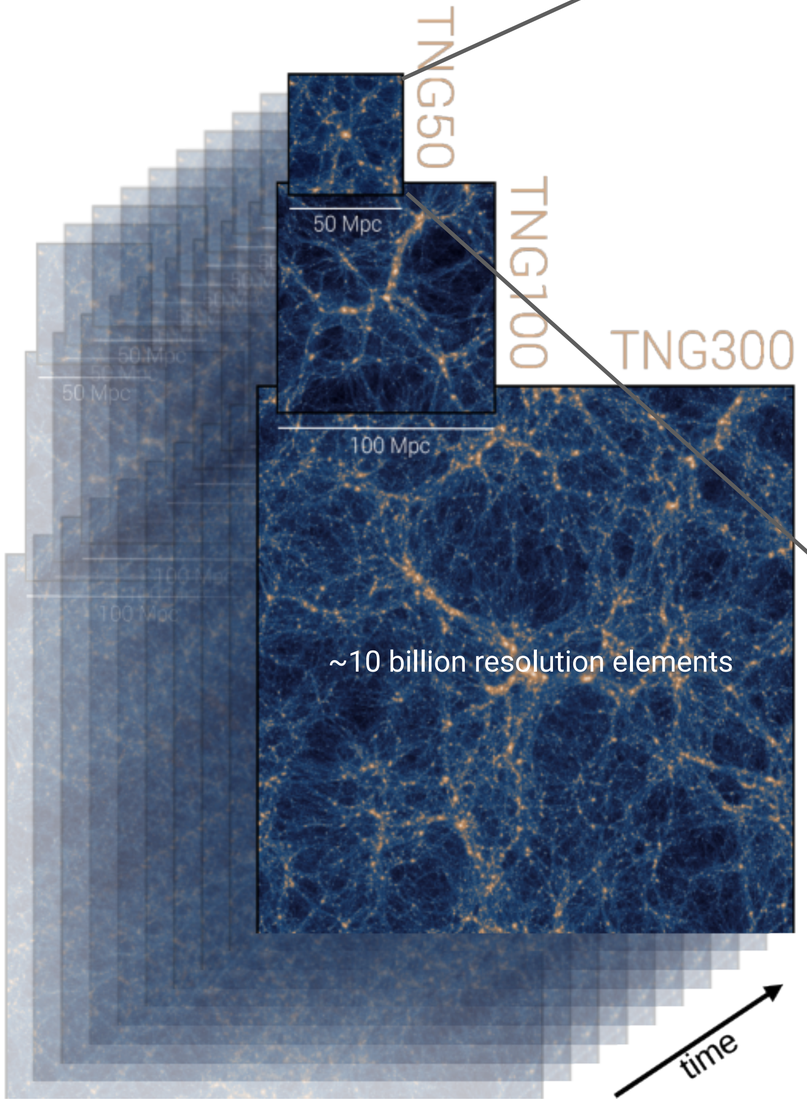
+

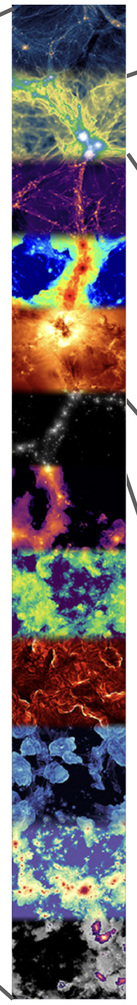
dark matter
dark energy
hydrodynamics
magnetic fields
atomic physics
radiation transport
chemical networks
molecular physics
dust physics
star formation
stellar evolution
nucleosynthesis
chemical enrichment
stellar feedback
black hole formation
black hole growth
black hole feedback
cosmic rays
.
.
.

… after running for 2 years on a supercomputer (~100M CPU-hours)

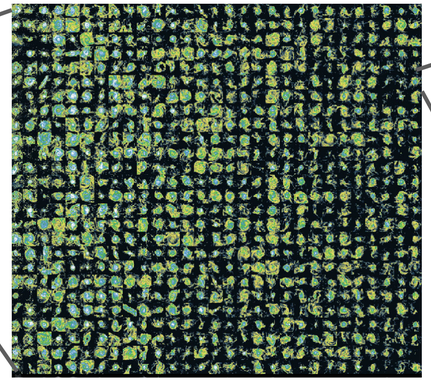
A multi-scale challenge

What does the data look like?





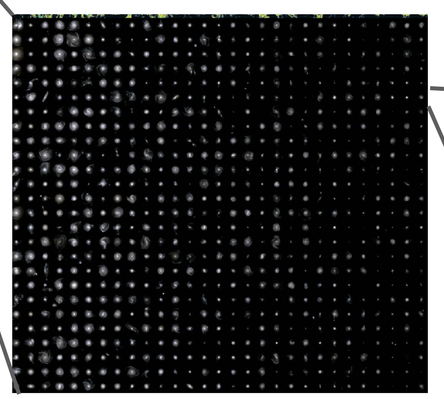
Physical properties:
Dark matter density
Gas density
Lyman alpha emiision
Gas temperature
Gas velocity
Stellar mass
X-ray emiision
Gas metallicity
Shock Mach number
Energy dissipation rate
Magnetic field
OVII column density
...
...
...



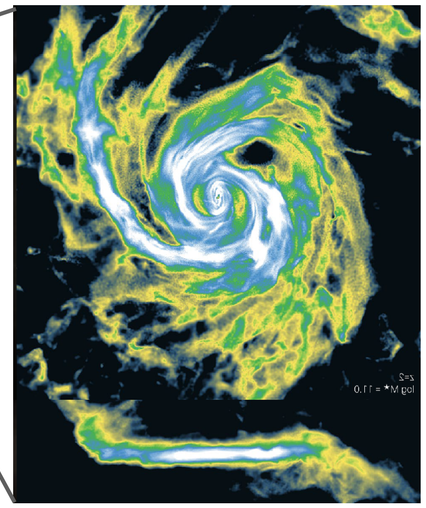
...

millions of objects

...



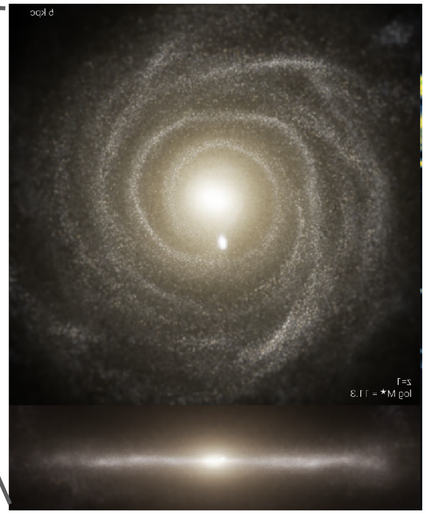
...
...
...
...
...



...

complex multiscale structures

...



...
...
...
...
...
...

images courtesy of the TNG Collaboration

If all the information is there, *why are these simulations so difficult to interpret?*

… let’s focus on a single object

The data volume and complexity problem

Today's largest projects evolve a universe in a box:

- ~100 billion resolution elements
- in a >100-dimensional feature space

} a feature space with $> 10^{12}$ dimensions

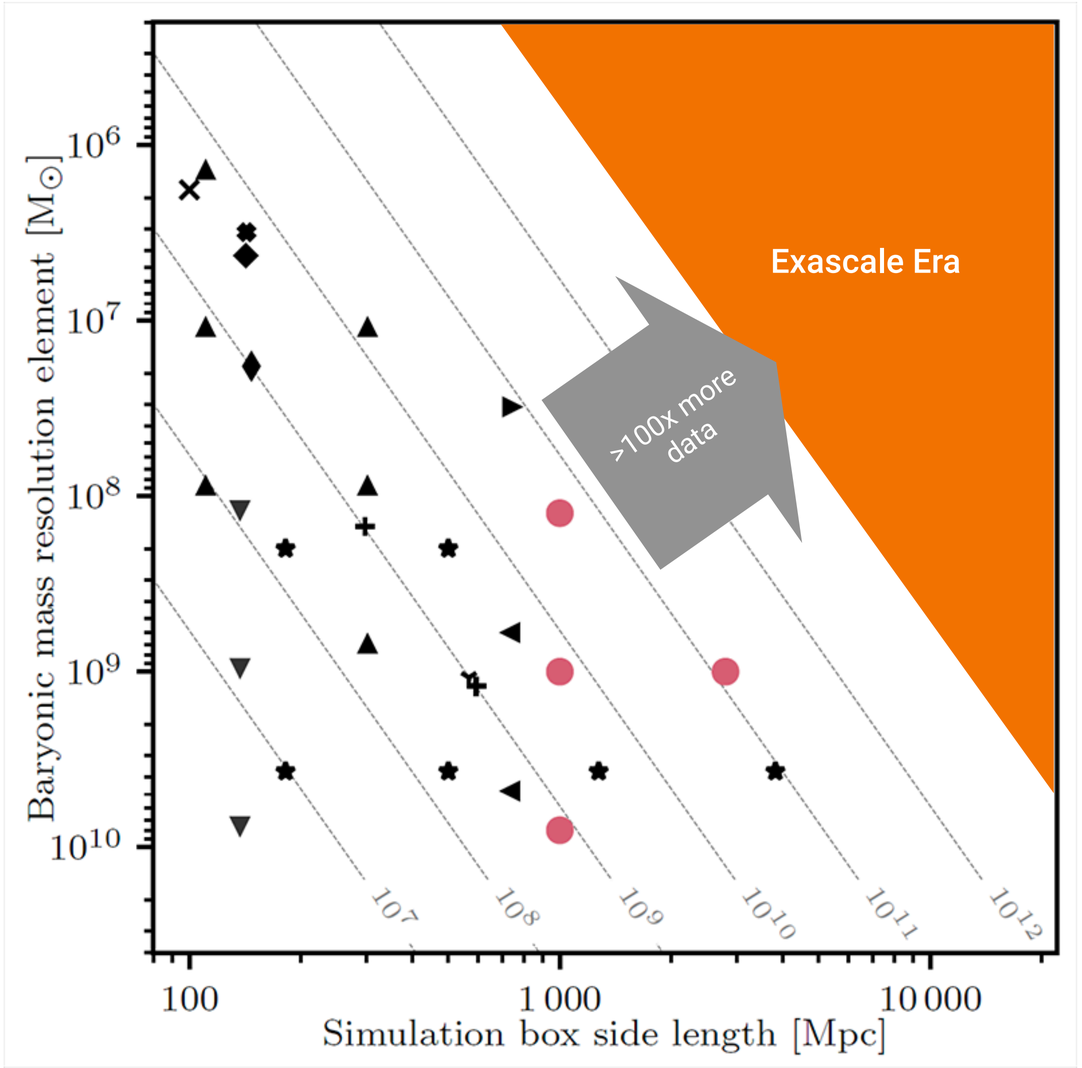
- in millions objects (galaxies)
- ~1 petabyte of data (petascale)

Exascale simulations: *>100 times larger* datasets -> ML essential

Hard-working students at a scientific sweatshop as imagined by DALL-E

LUMI, Europe's largest supercomputer

The largest cosmological simulations



Shaye et al. (2023)

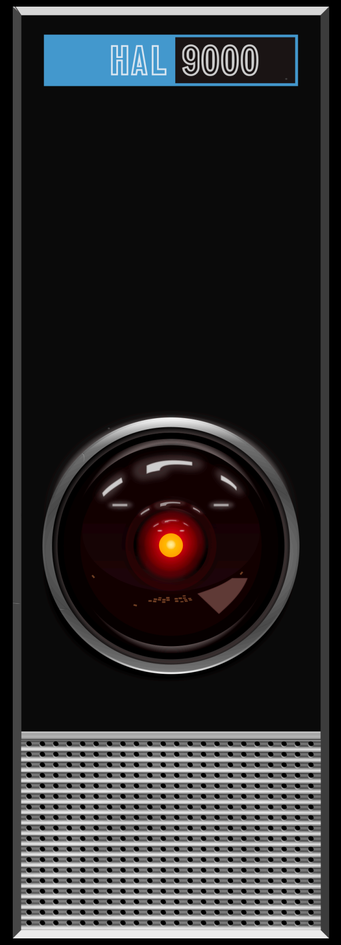
How can we approach this challenge?

The ideal tool should

1. learn a simple representation of all structures in the simulation *without labels*

3. enable exploration and interpretation using interactive visualization for *arbitrarily large datasets*

4. be code-agnostic (can be used on *any simulation* without expertise)



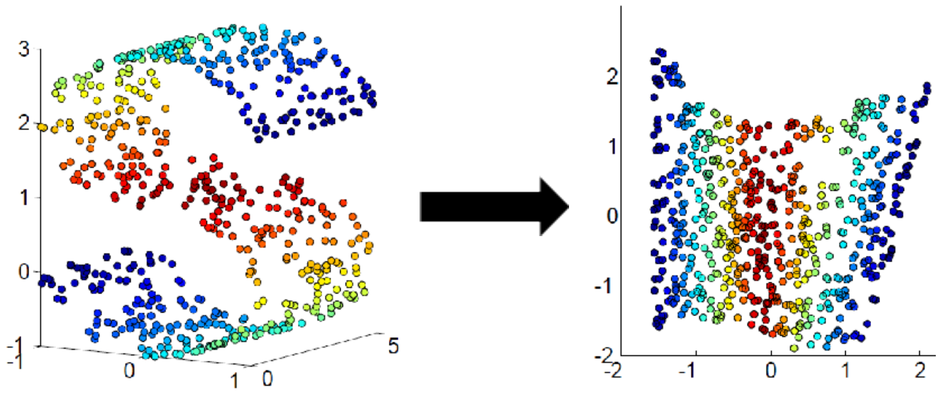
HAL9000 from
2001: A Space Odyssey

The power of Generative Deep Learning

"What I cannot create, I do not understand"

Richard Feynman

- Learns underlying distribution of dataset from samples *without labels*
- Projects data onto low-dimensional space, providing a powerful way to explore and interpret it



data in physical space

data in intrinsic
low-dimensional space

galaxy structure

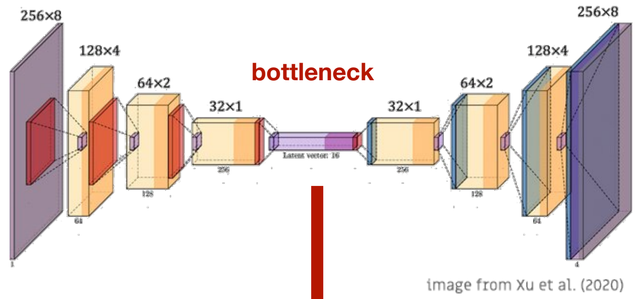
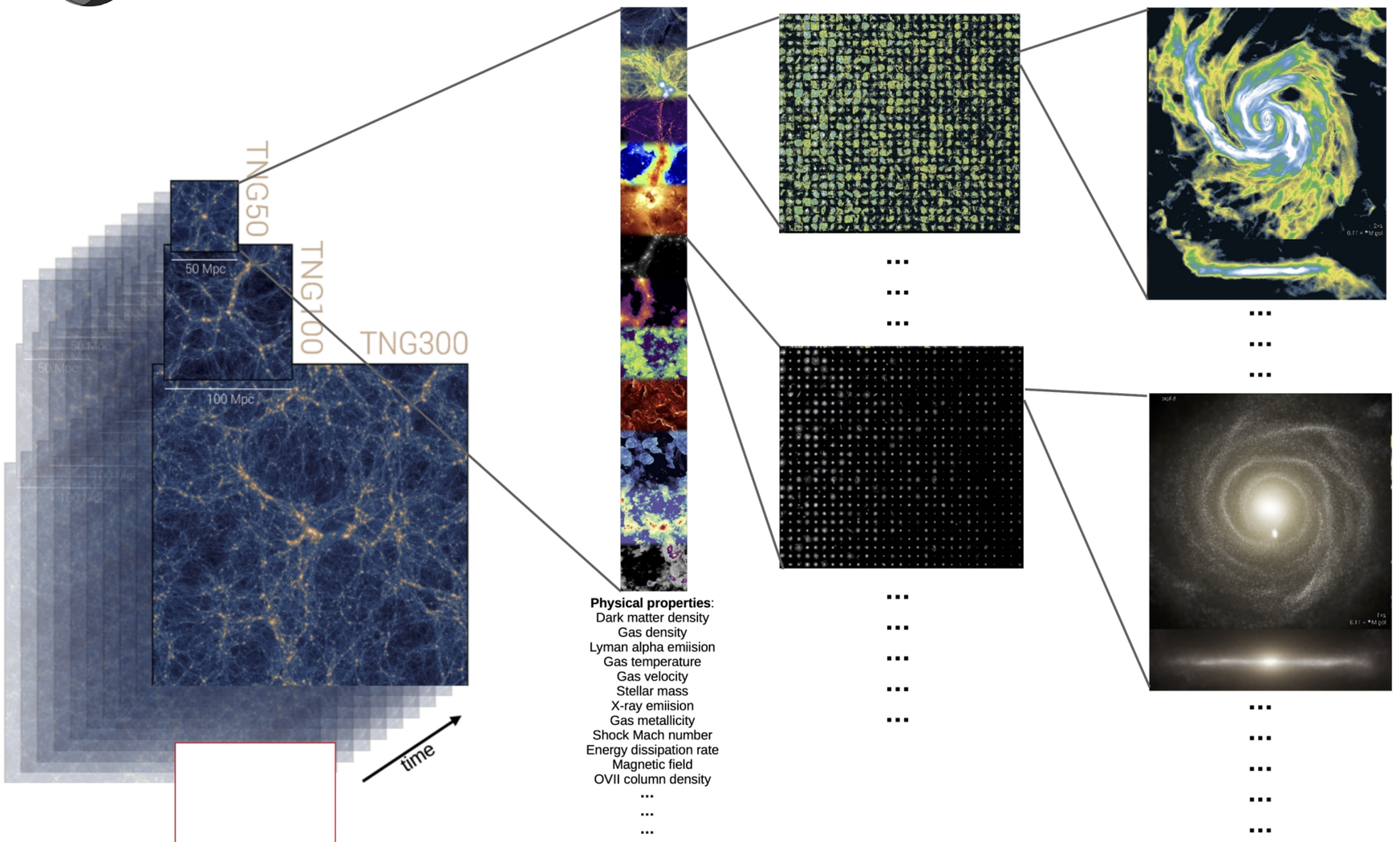


image from Xu et al. (2020)

reconstruction

$\vec{z} = (z_1, z_2, z_3)$
compression of structure
into few numbers



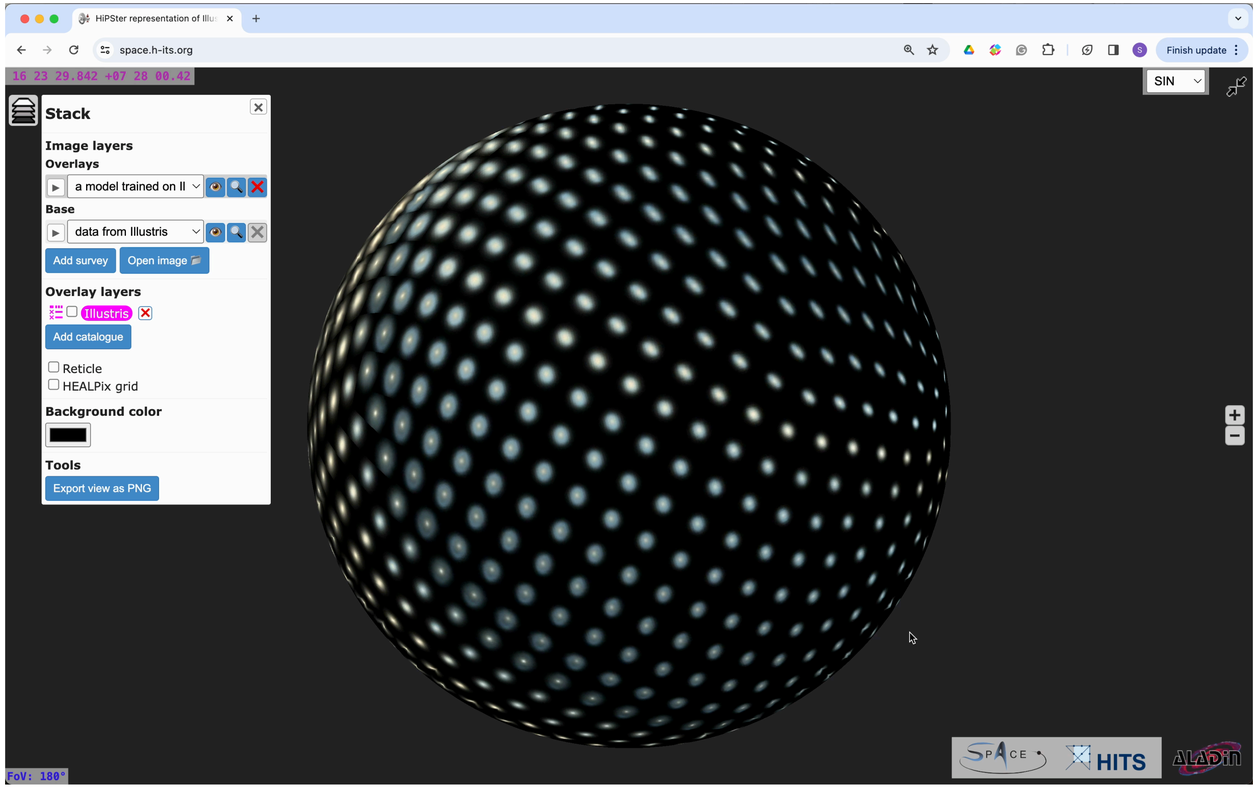
images courtesy of the TNG Collaboration

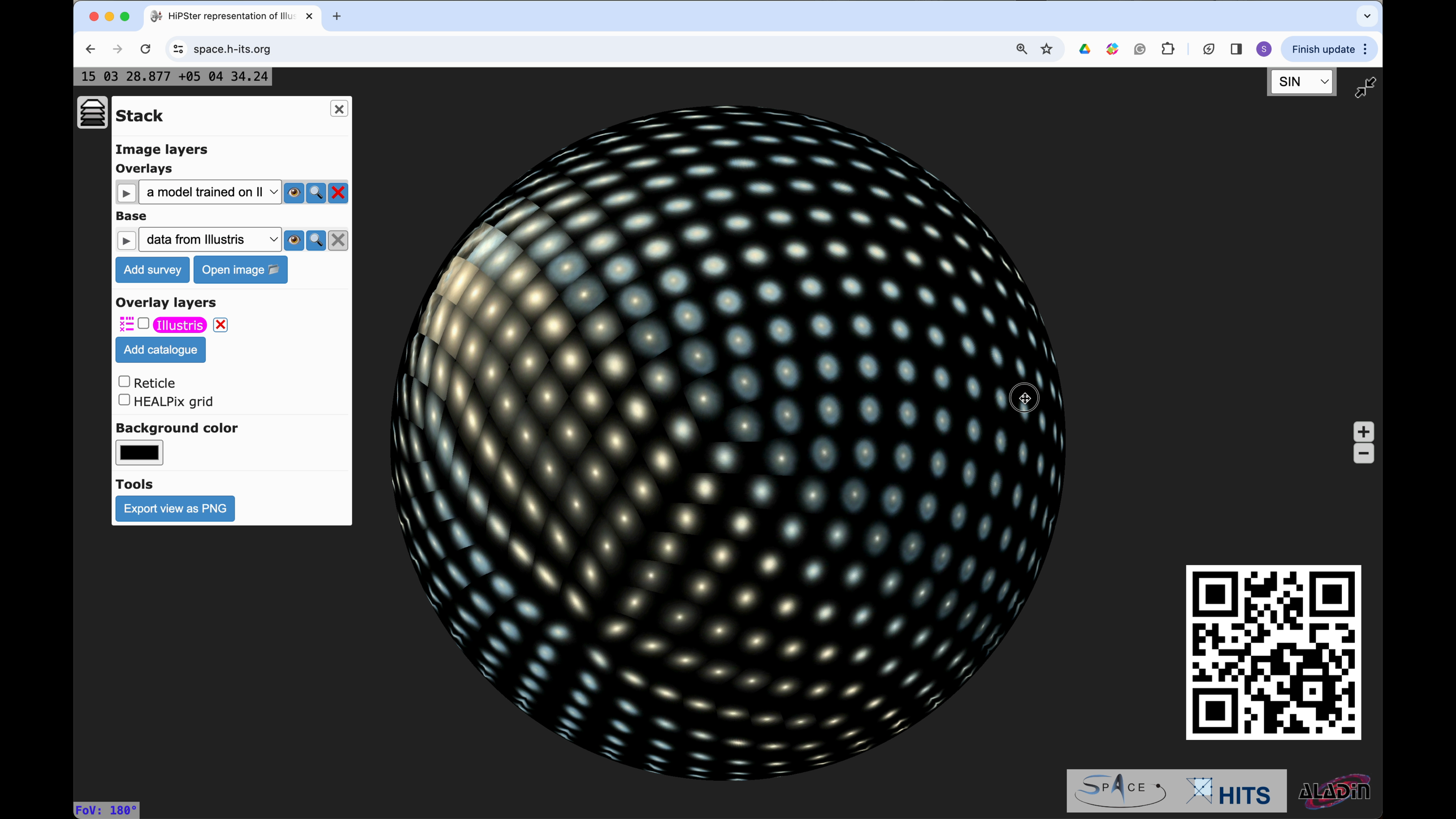
First results

Rule #4:
Explore limits and scope of models

Rule #3:
Interpret and visualize models

- Trained prototype on
  - ~51,000 synthetic galaxy images from the IllustrisTNG simulations (Nelson et al. 2019; Rodriguez-Gomez et al. 2019), and
  - ~150,000 real galaxy images from the GalaxyZoo (Lintott et al. 2008)
- Learned the high-level galaxy features (i.e. color, size, morphology, inclination, and interactions)



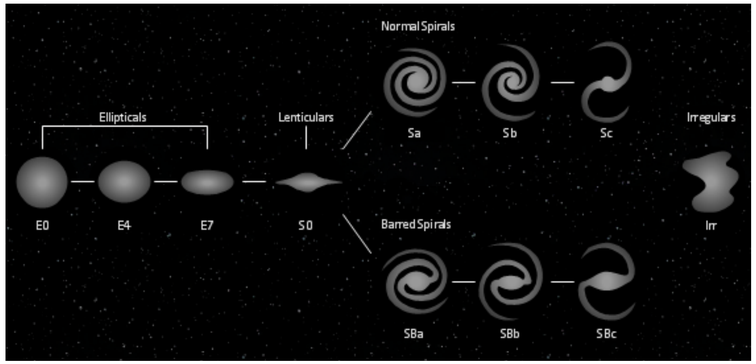

HiPSter representation of Illus
space.h-its.org
Finish update
15 03 28.877 +05 04 34.24
SIN
Stack
Image layers
Overlays
a model trained on Il
Base
data from Illustris
Add survey
Open image
Overlay layers
Illustris
Add catalogue
Reticle
HEALPix grid
Background color
Tools
Export view as PNG
FoV: 180°
SPACE
HITS
ALADIN

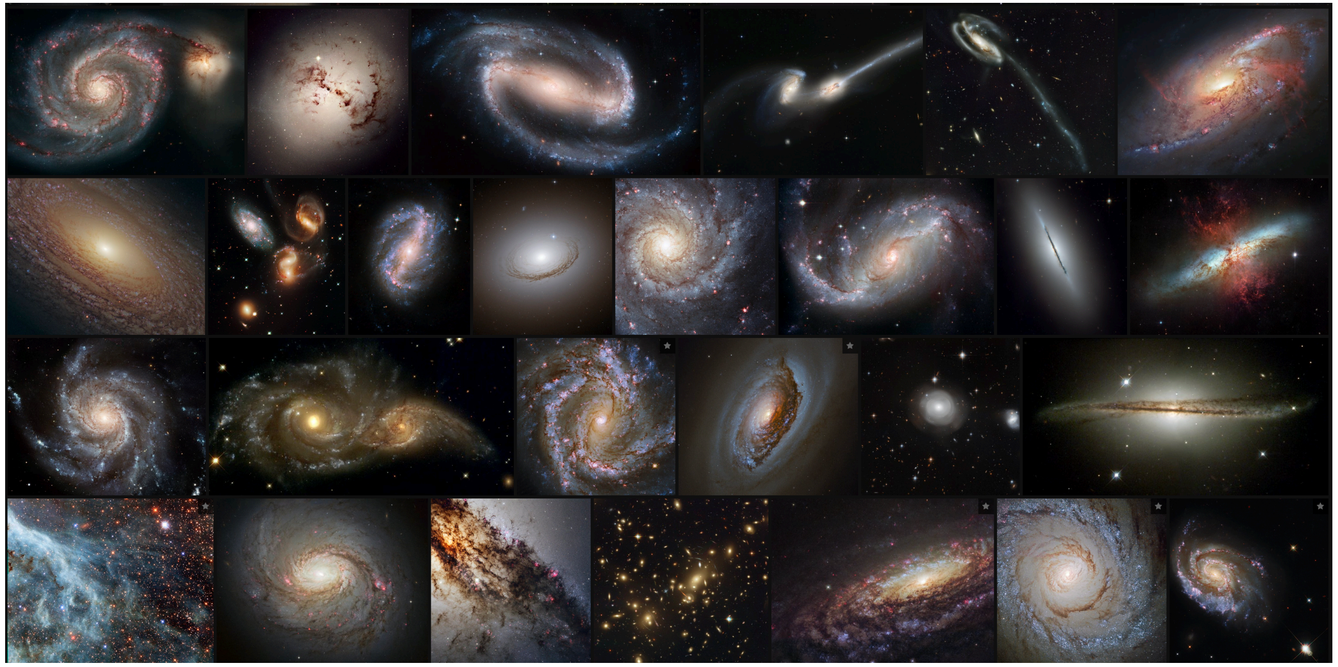
Scientific discovery with SPHERINATOR + HiPSter

Rule #1:
Compare to domain reference and broader context

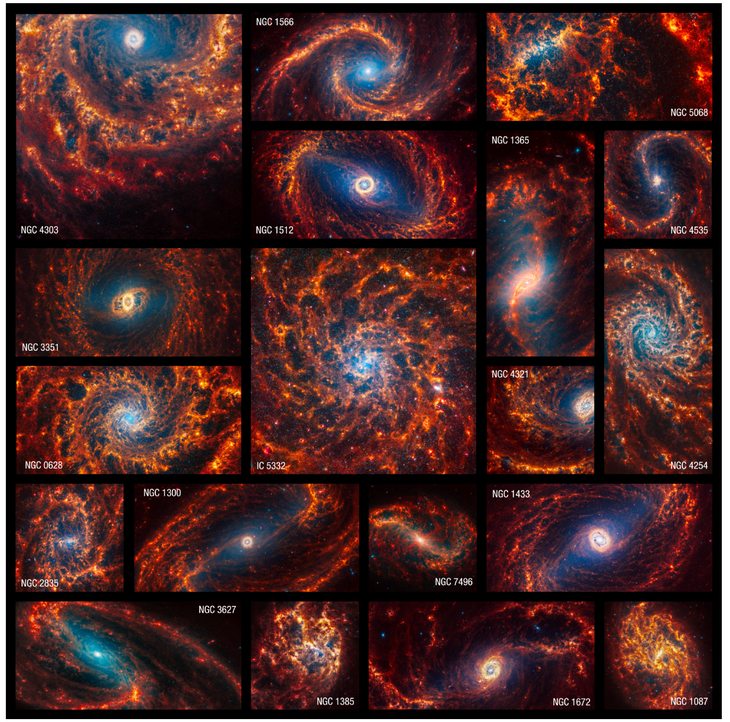
- Astronomers describe galaxies using single numbers based on intuition (e.g. luminosity, size, shape, etc.)
- *But galaxies ARE NOT points!*
  - they are complex objects emerging from many physical processes

How we have described galaxies for the last 100 years





Real galaxies as seen by HST …

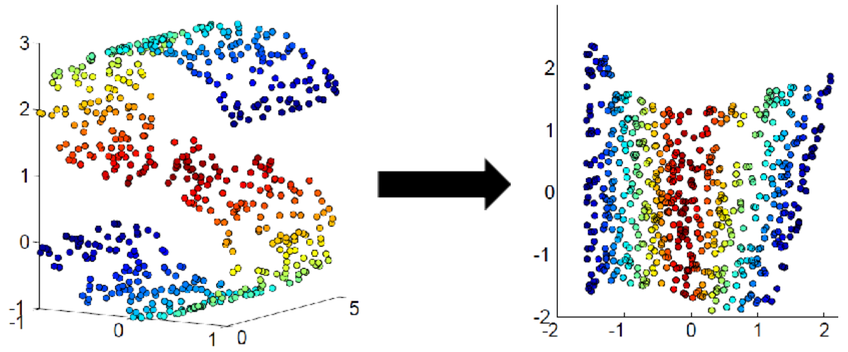


… and JWST

Scientific discovery with SPHERINATOR + HiPSter

Rule #5:
Share and discuss lessons learned

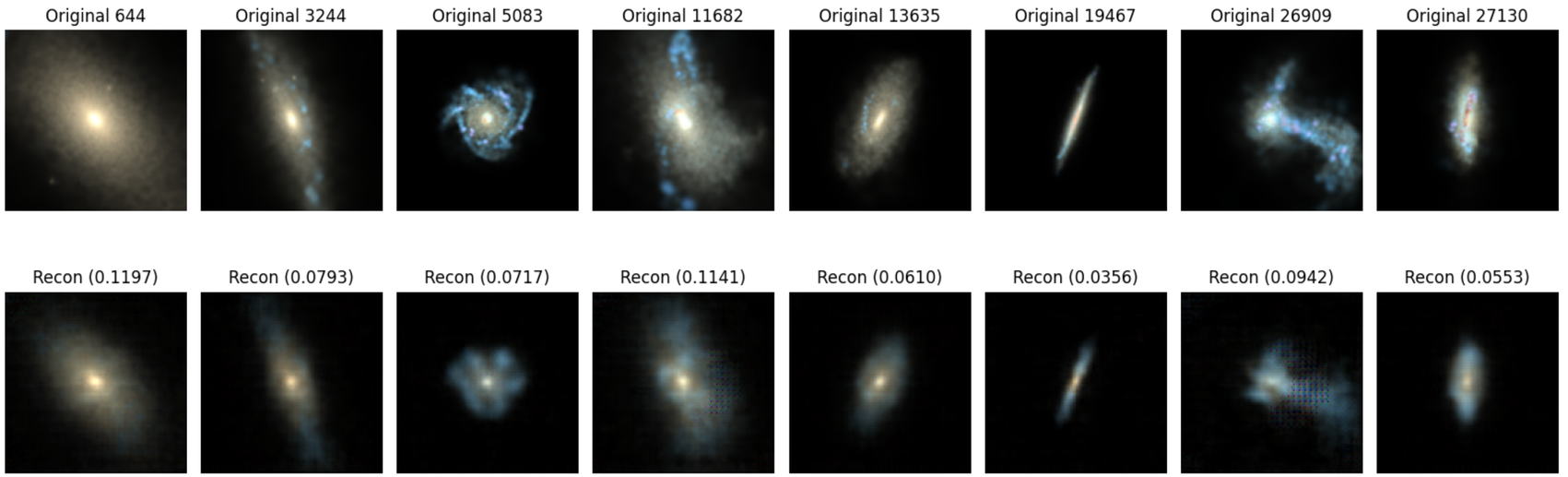
- Astronomers describe galaxies using single numbers based on their intuition (e.g. luminosity, size, shape, etc.)
- *But galaxies ARE NOT points!*
  - they are complex objects emerging from many physical processes
- Ask the machine: how many numbers are needed to describe a galaxy?



> 10,000-dimensional pixel space

32-dimensional manifold

32 dimensions are enough to capture galaxy structure (at ~2% level)



Scientific discovery with SPHERINATOR + HiPSter

More applications…

1. Exploration

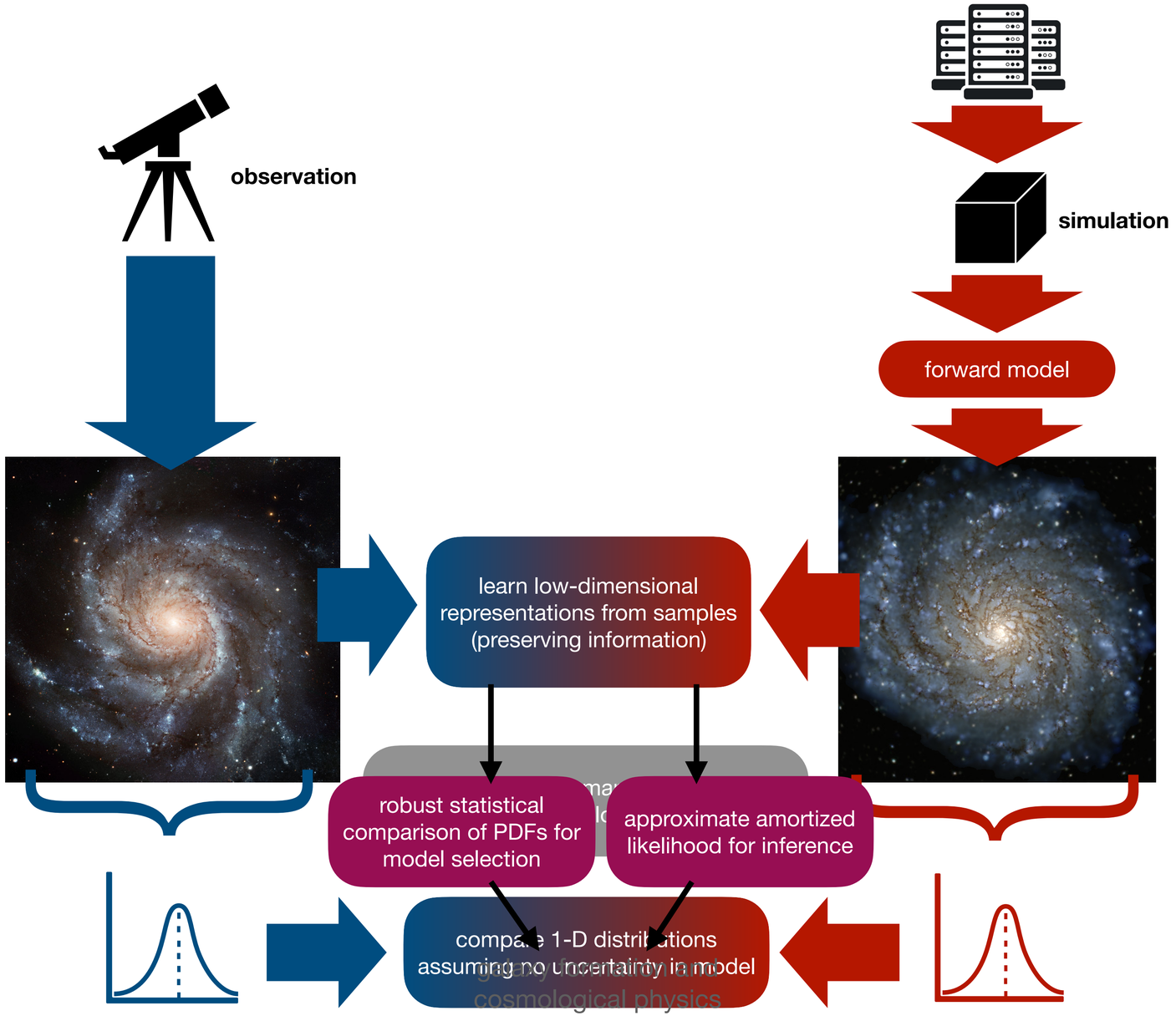
   - Correlations between all physical properties of structures
   - Interpret using known labels (e.g. galaxy morphology or environment)

2. Simulation Based Inference + Model selection

   - Galaxies are chaotic: cannot compare objects directly!
   - Instead compare structural distributions non-parametrically
   - Likelihood without need for summary statistics

simulation
or
real?

Scientific discovery with SPHERINATOR + HiPSter



Rule #2:
Adopt best practices from ML community

Rule #6:
Publish software and data

Try it!

- arXiv:2406.03810 (Polsterer, Doser, Fehlner & Trujillo-Gomez 2023, in press)
- github.com/HITS-AIN/Spherinator

Can we help you explore and interpret your simulations and/or data?

sebastian.trujillogomez@h-its.org

Play with a demo!

space.h-its.org

Funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Belgium, Czech Republic, France, Germany, Greece, Italy, Norway, and Spain under grant agreement No 101093441.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European High Performance Computing Joint Undertaking (JU) and Belgium, Czech Republic, France, Germany, Greece, Italy, Norway, and Spain. Neither the European Union nor the granting authority can be held responsible for them.

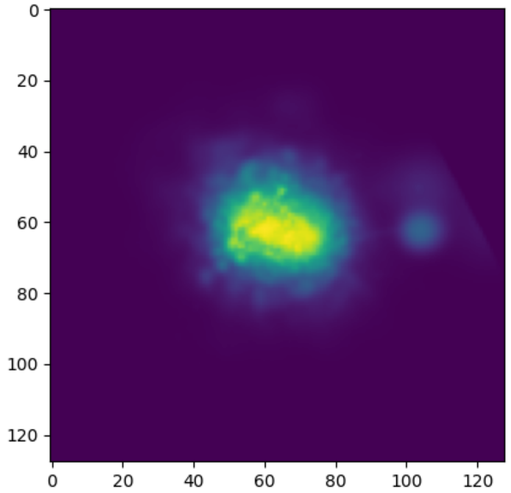
Co-funded by the European Union

EuroHPC Joint Undertaking

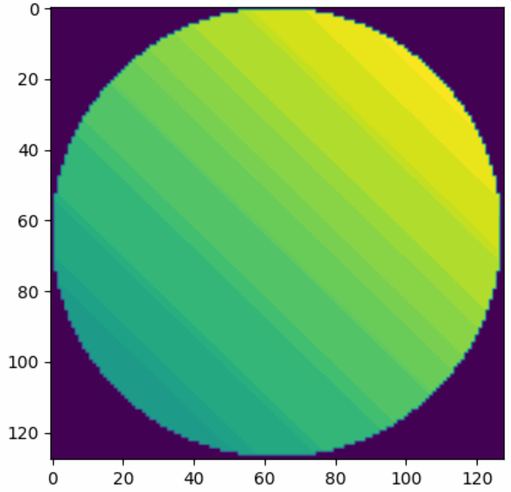
What's next?

1. Physical symmetries: rotational equivariance
2. Beyond images to full 3D structures:
   - described using Geometric Deep Learning
   - explore 3D structures interactively

original galaxy image



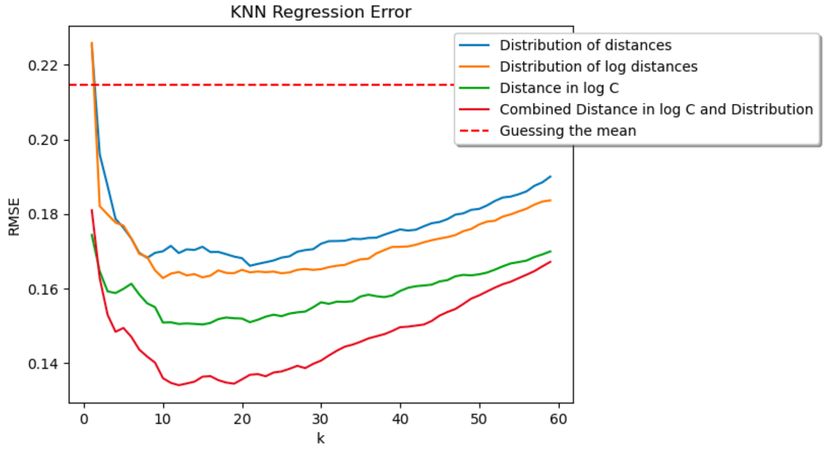
equivariant representation using Zernike polynomials



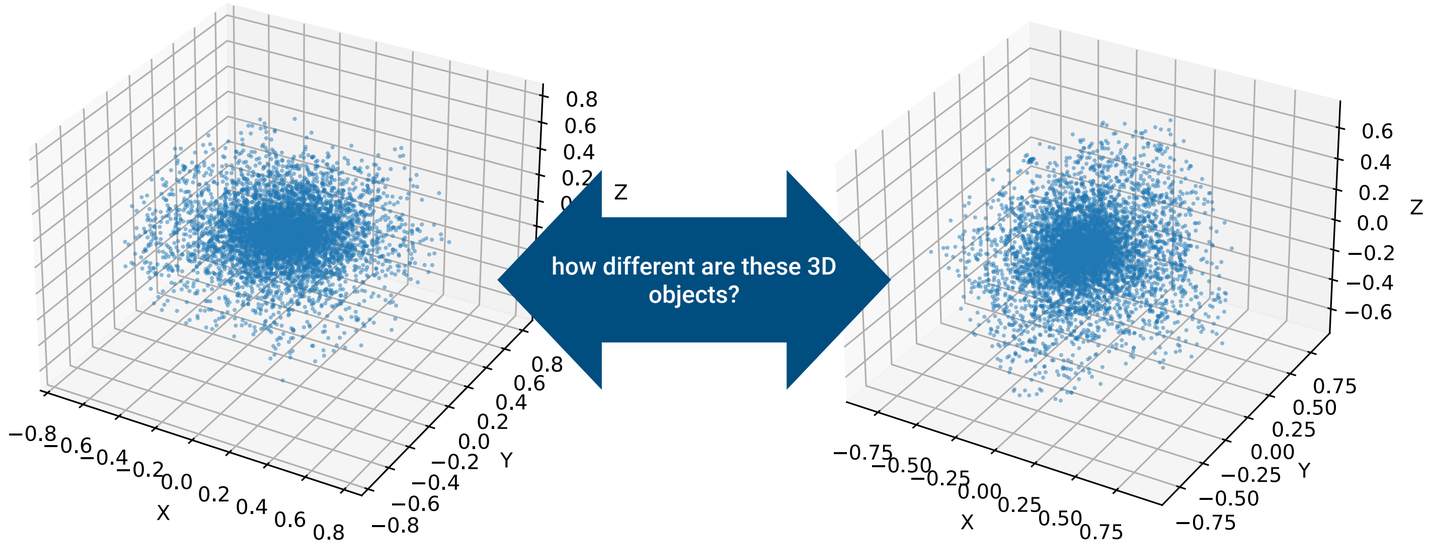
Chazotte et al. (in prep)

Scientific discovery with SPHERINATOR + HiPSter

- Astronomers describe galaxies using single numbers based on their intuition (e.g. luminosity, size, shape, etc.)
- *But galaxies ARE NOT points!*
  - they are complex objects emerging from many physical processes
- Ask the machine: how many numbers are needed to describe a galaxy?



Shape of the galaxy halo predicts stellar mass
(Packer, Villar, Hogg, STG in prep.)



The dark halos of two simulated galaxies