Data-driven modeling in Cosmology

AI-at-scale approaches

Nesar Ramachandra (CPS)

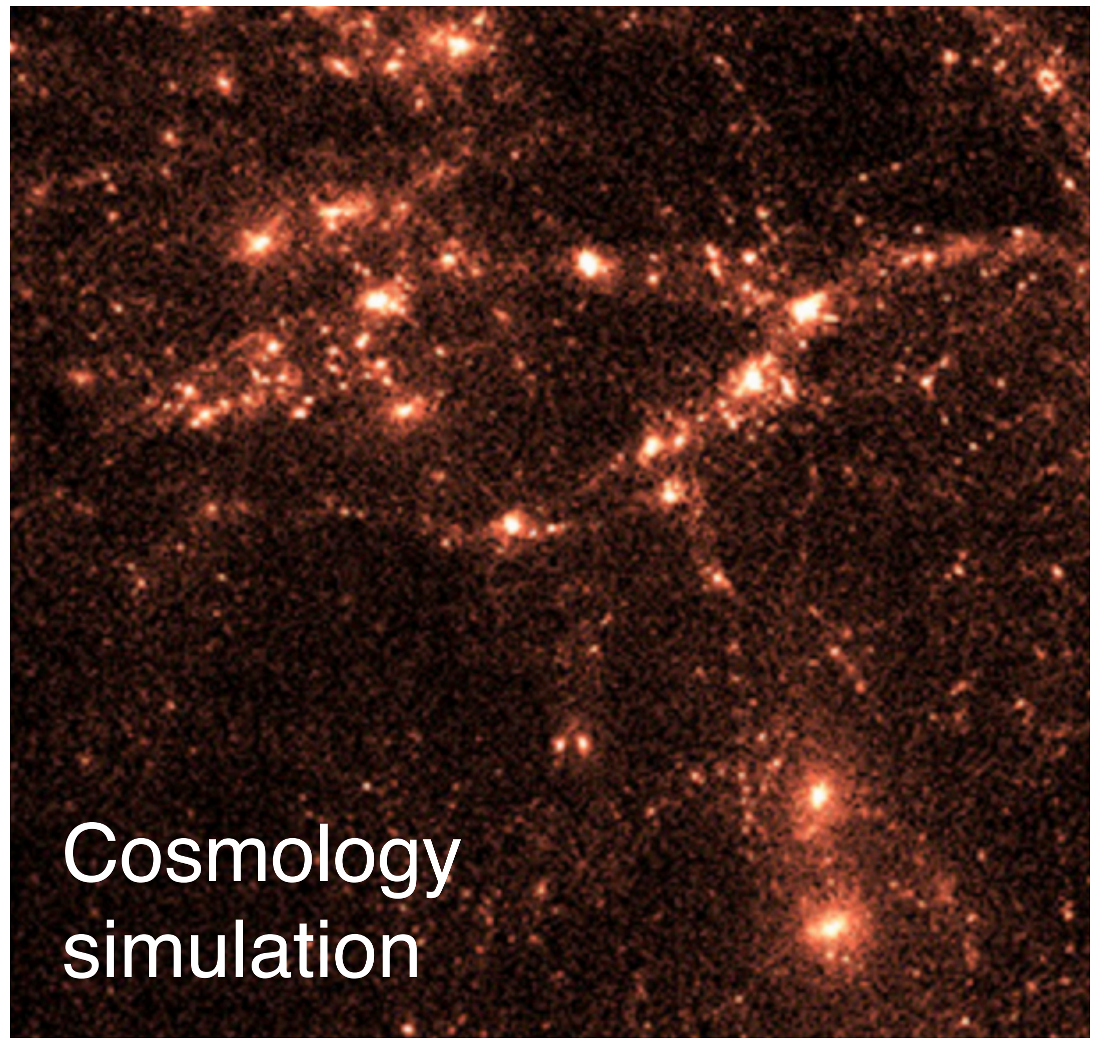
AI-for-Cosmology at Argonne (CPS, HEP, MCS and ALCF)

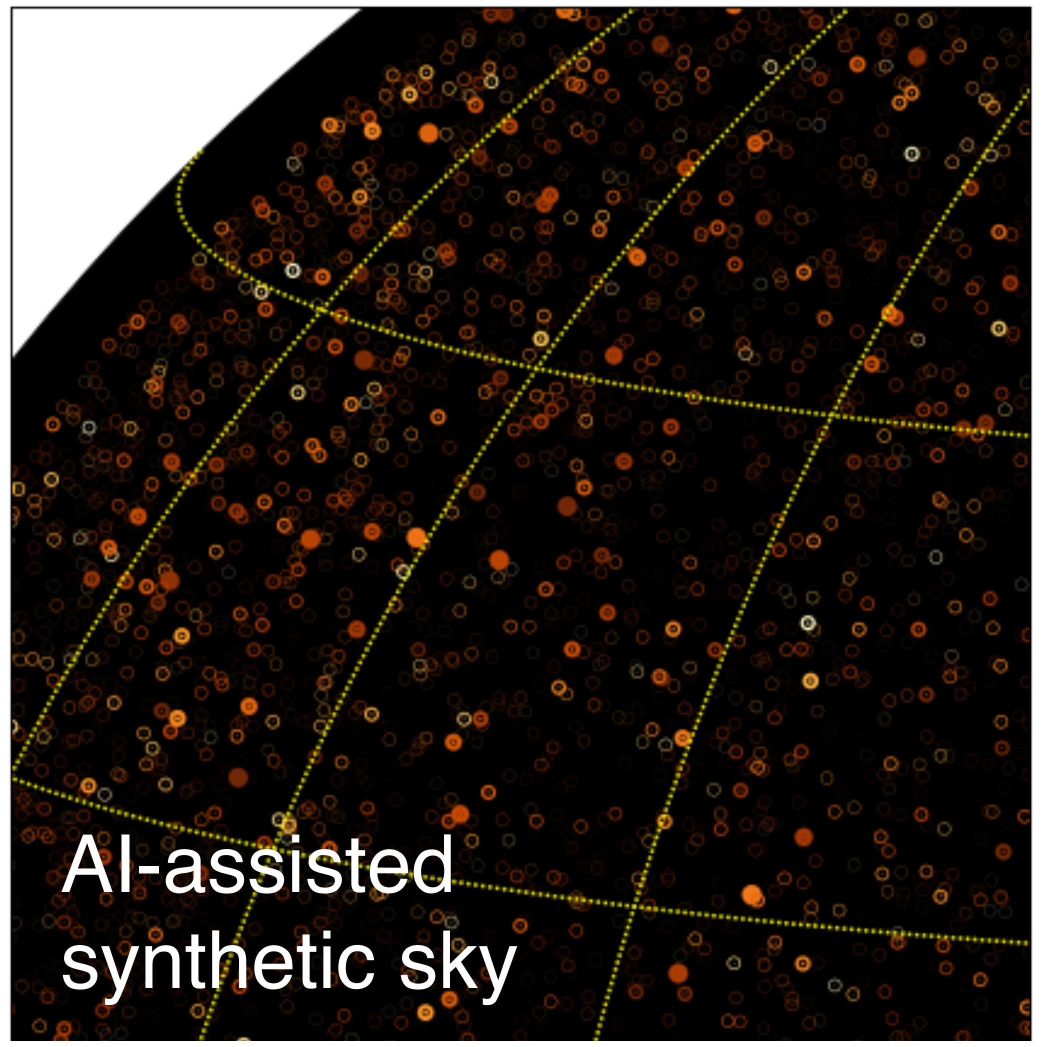
Common themes

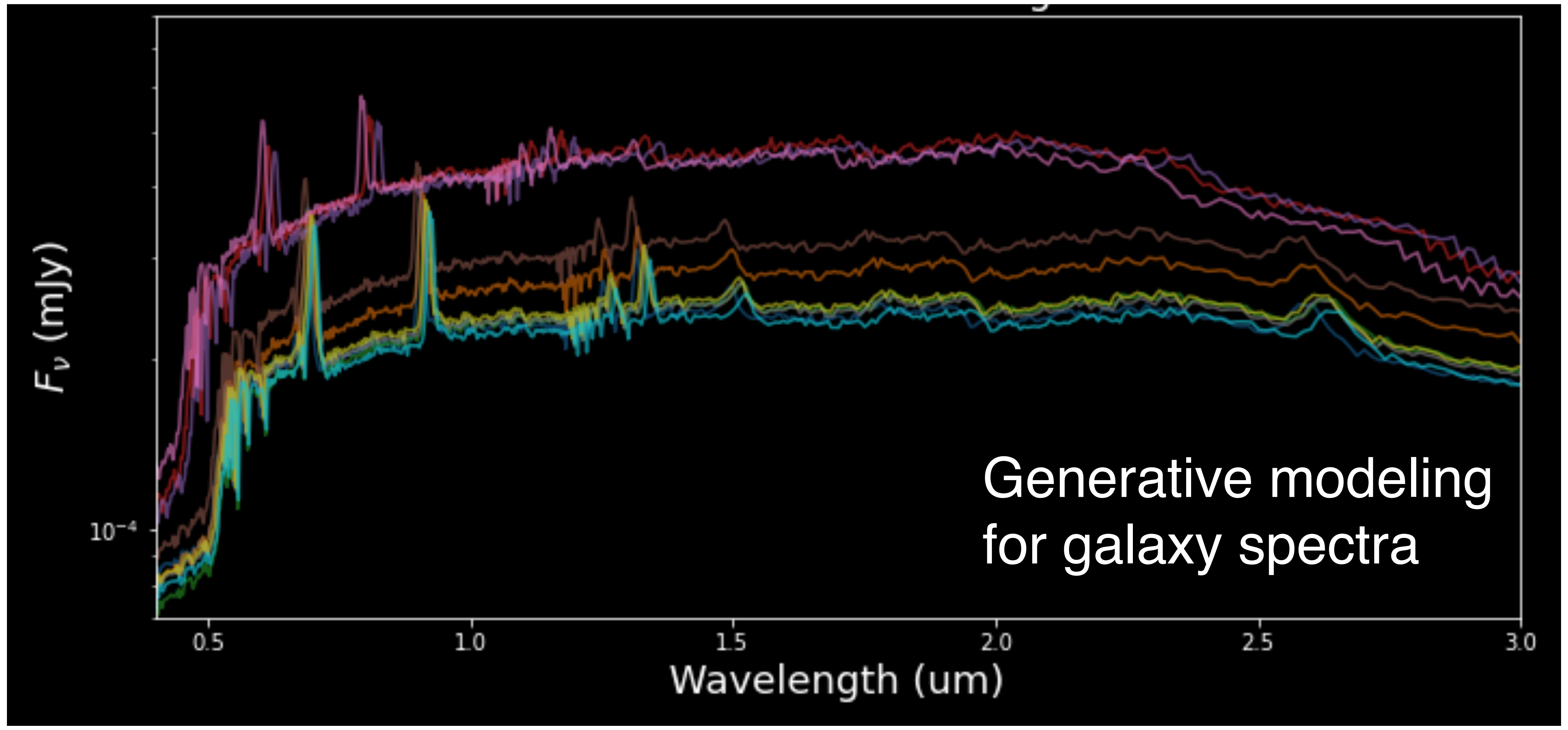
- High-fidelity simulation data supplements real astronomical observations.
  - Expensive simulations, expensive modeling
- Bayesian/probabilistic schemes.
- Explainability of the AI algorithms, physics inclusion in optimizations and benchmarks

Simulations to synthetic sky catalogs

- Science goal: Emulate the astronomical observations of galaxies with high realism. Replicate the density functions, correlations, spectra of the galaxies as seen by future cosmic surveys (ground-based telescopes and space telescopes).
- Data: Trained on millions of galaxies, to be deployed on some of the largest cosmological simulations and billions of objects.
- Methodology: Surrogate models trained to map from simulation products to telescope outputs. Pipeline with several AI-modules



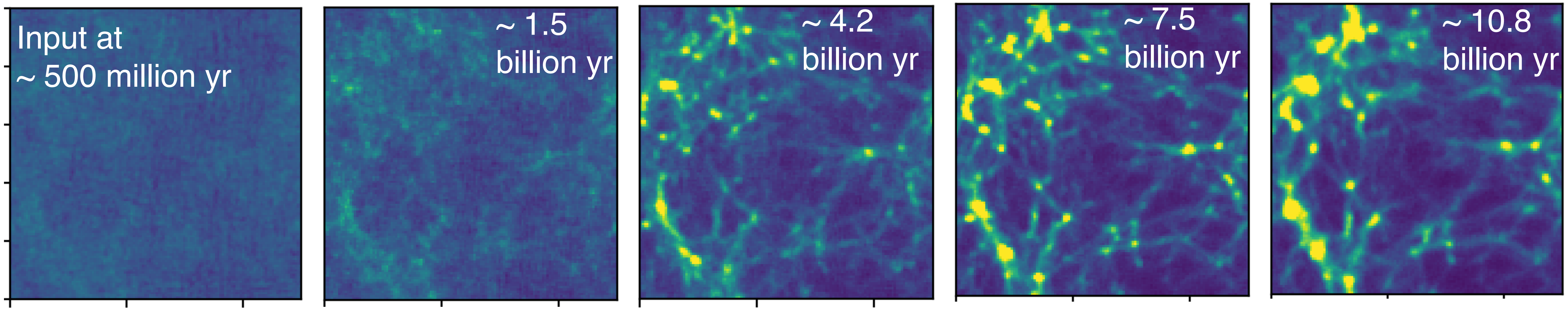




AI efforts within Cosmology@Argonne (CPS, HEP, MCS and ALCF)

Generative modeling of evolution of the Universe

- Science goal: Generate dark matter distributions at various stages of evolution of the Universe.
- Data: Trained on 1000s of low-to-medium resolution 3D simulations
- Methodology: Image-to-image mapping via U-nets and diffusion models. Sharded data-parallel training on multi-GPUs, SambaNova testbed.



Scaling challenges in AI-for-Cosmology:

- Data-intensive:
  - Rubin Observatory will see **37 billion** astronomical objects, producing **15 terabytes** of data per night.
  - Raw data in cosmological simulations are in **O(petabytes)**
- Often data sizes of individual datapoints is too large for GPU memory.
- Physics benchmarking and domain specific loss functions are expensive (compared to out-of-the-box loss metrics)
- Integration of trained AI-surrogates within simulation steps.