

AI/ML Research in Astro/Cosmology in Japan

Where do discoveries happen?

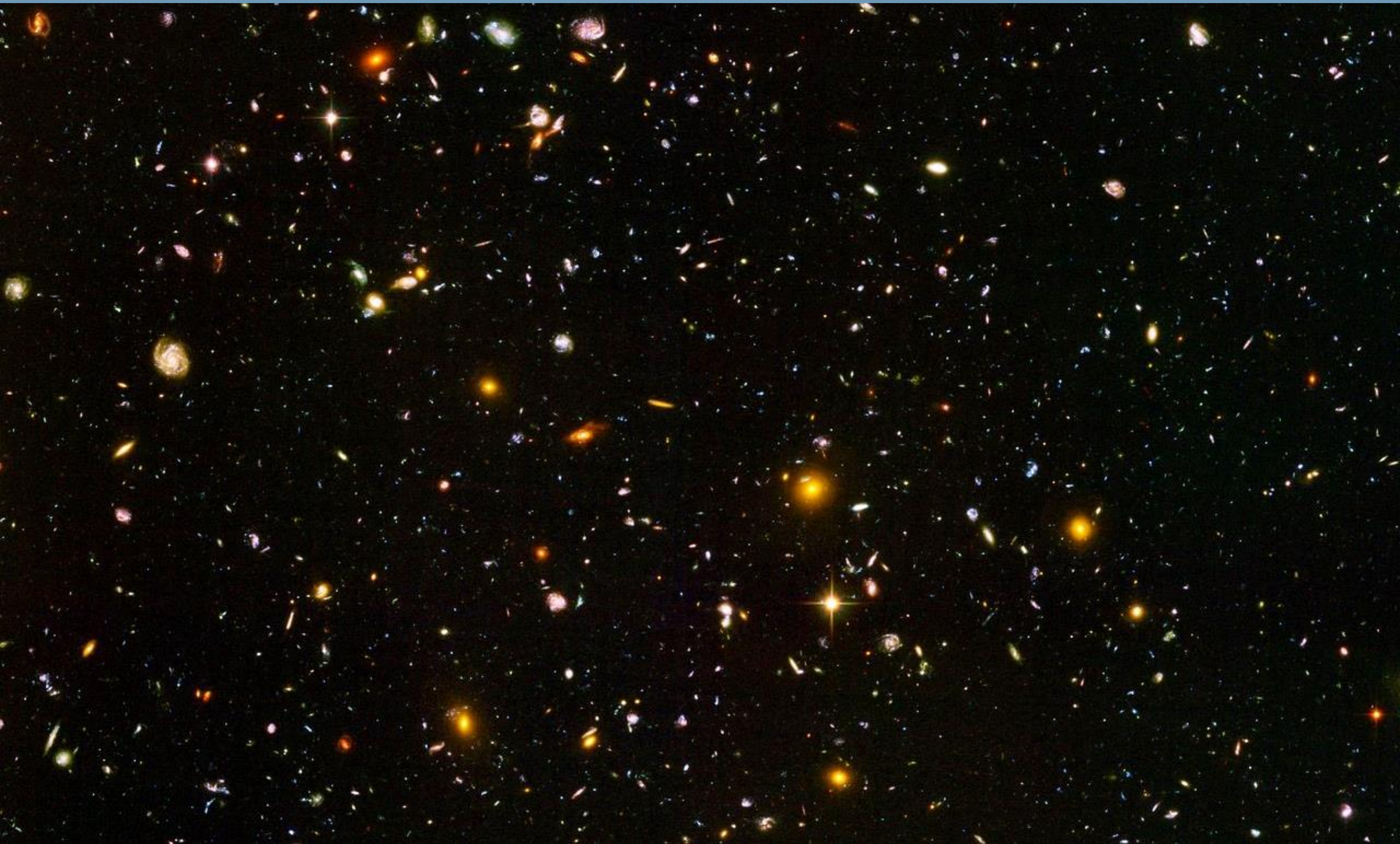
*FAIRS Japan 2026 @ Nagoya University
2026/10/05*

Suchetha Cooray

Stanford University -> Tohoku University

Observed Galaxy Fields (Our Target)

2



Transformation in the Data Landscape

We now have/soon have data that is
covering larger sky area, deeper,
and higher resolution
by orders of magnitude than the previous
generation data

Need for New Modeling & Inference Tools

Now!
The data tsunami in astronomy — ~~starting this summer!~~

Massive, heterogeneous, multi-modal datasets – arriving now through the 2030s.

Credit: Risa Wechsler

20 TB

per night

Rubin / LSST · 3.2 Gpx camera

100 M

spectra

DESI / DESI-II, Roman

600 M

galaxy shapes from space

Roman – 0.1", deep, public

$6 \times 10^9 +$

sources

Rubin + Roman + Euclid

+ JWST precision images · Simons Observatory (CMB) · Via high-res spectroscopy · multiwavelength cross-correlations

Data is not the bottleneck. Inference infrastructure is.

Astro/Cosmo @ FAIRS 2024 vs. Now

2024 Community White Paper: astro/cosmo

- Simulation-based and field-level inference
- Fast, differentiable forward modeling and surrogates
- Compression, calibration, uncertainty propagation
- Anomaly detection and discovery
- Reproducible, shared infrastructure

Stressed: interpretability, robustness to misspecification

2024 talks

- Takada: information extraction from big-data cosmology (field-level inference, SBI)
- Thiele: interpolation, emulation, inference methods

This talk

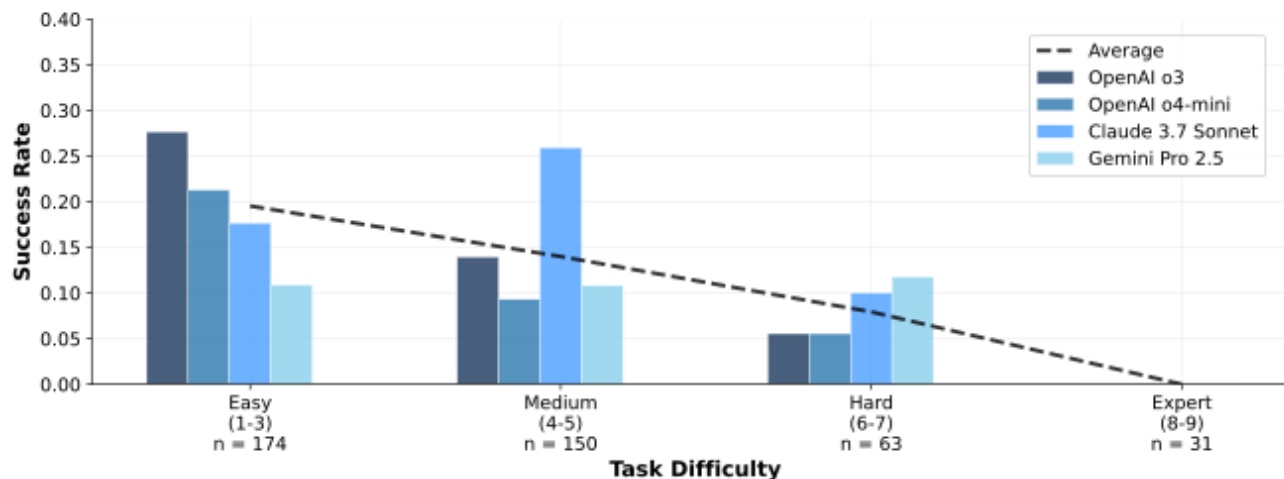
- LLM/Agents (briefly)
- Broader astro/cosmo: stars, planets, galaxies, transients
- The medium-N regime
- Differentiable forward models
- Selection, model uncertainty, validation, and calibration

Agents have transformed scientific workflows

ResearchBench (Ye et al. incl. Cooray 2025) gives an agent real scientific data, requires planning and execution, and grades a reproducible result.



Task Success Rate by Difficulty Level



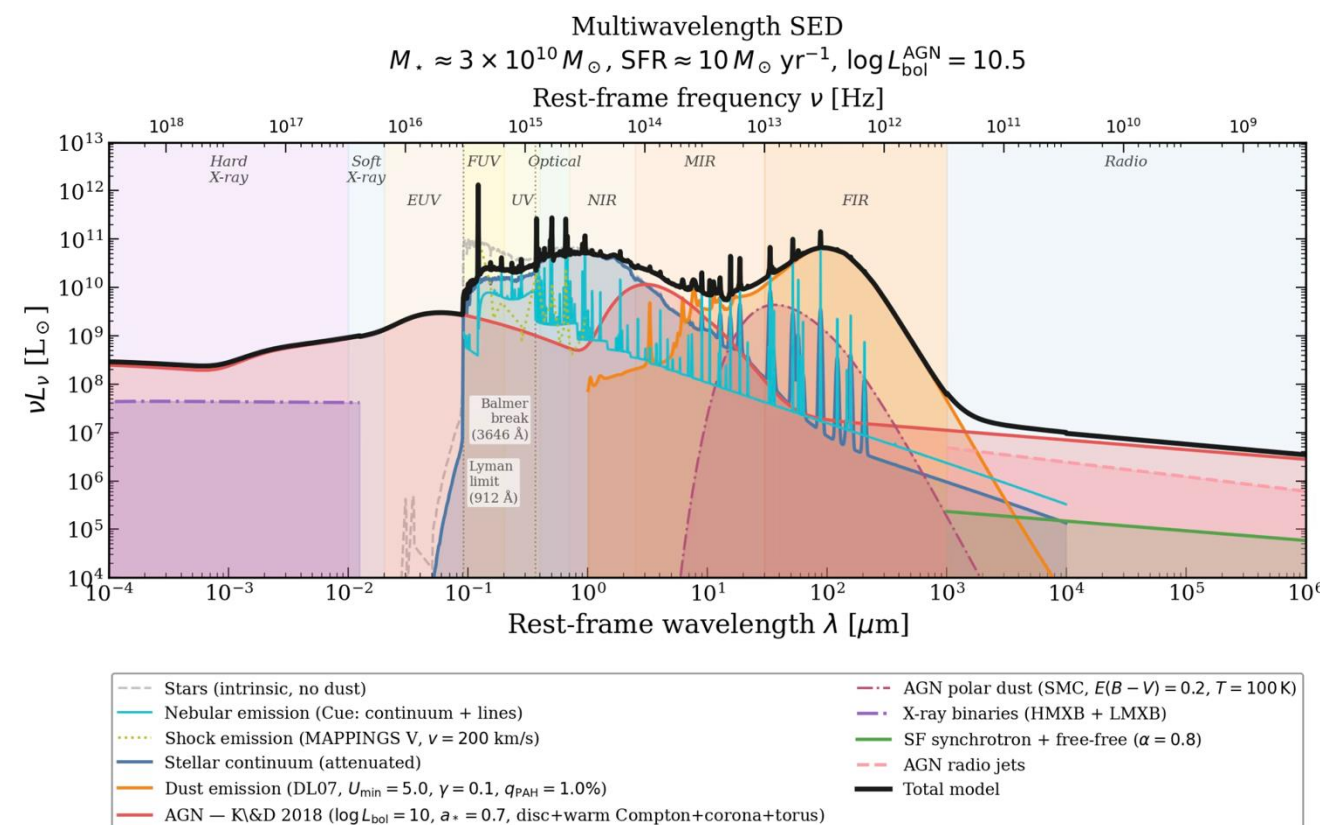
Suchetha Cooray

Tengri (Cooray et al. in prep.) is a differentiable framework for high-dimensional Bayesian inference from galaxy SEDs.



Primarily agent-assisted development.
Designed to be used with agents
(includes a AGENTS.md)

Only 7 months old. Would have taken years.



Inference problems differ across physical scales ⁷

	Cosmology	Galaxies	ISM & star formation	Stars	Planets	Transients
Theory	ΛCDM: few parameters; analytic at linear scales, N-body beyond ● ● ●	Effective models with many parameters; sub-grid physics ● ○ ○	Non-equilibrium, multiphase, turbulent, magnetized ● ○ ○	1D evolution and atmospheres; convection, rotation, mass loss, binaries uncertain ● ● ○	Atmospheric radiative transfer; formation less predictive ● ● ○	Explosion mechanisms and opacities uncertain ● ○ ○
Observable	CMB, lensing, clustering ● ● ●	SEDs need stellar populations, dust, AGN ● ○ ○	Lines need chemistry and radiative transfer ● ○ ○	Spectra from model atmospheres ● ● ○	Spectra; clouds and chemistry uncertain ● ● ○	Light curves and spectra ● ○ ○
Typical inference problem	Precision parameters from large-N fields	Per-object properties and population models	Physical conditions per region	Large catalogs with per-object models	Per-object retrievals; demographics	Per-event models; populations

● ● ● more predictive ● ○ ○ less predictive (qualitative)

Three questions shape the inference problem:

How predictive is the theory? How directly does it predict the relevant observable? **How many objects are there?**

In which scenarios do discoveries
tend to happen in
astronomy/cosmology?

How Many Objects Did It Take?

21

galaxies

Flat rotation curves

Rubin, Ford & Thonnard 1980

42 + 16

high-z SNe Ia

Accelerating expansion

Perlmutter+99; Riess+98

67

pulsars timed for 15 years

nHz GW background

NANOGrav 2023

13

candidates (6 massive)

**Early massive galaxies &
Little Red Dots**

Labbé+23 (JWST)

8

ultra-faint dwarf galaxies

Dark matter dominated

Simon & Geha 2007

129

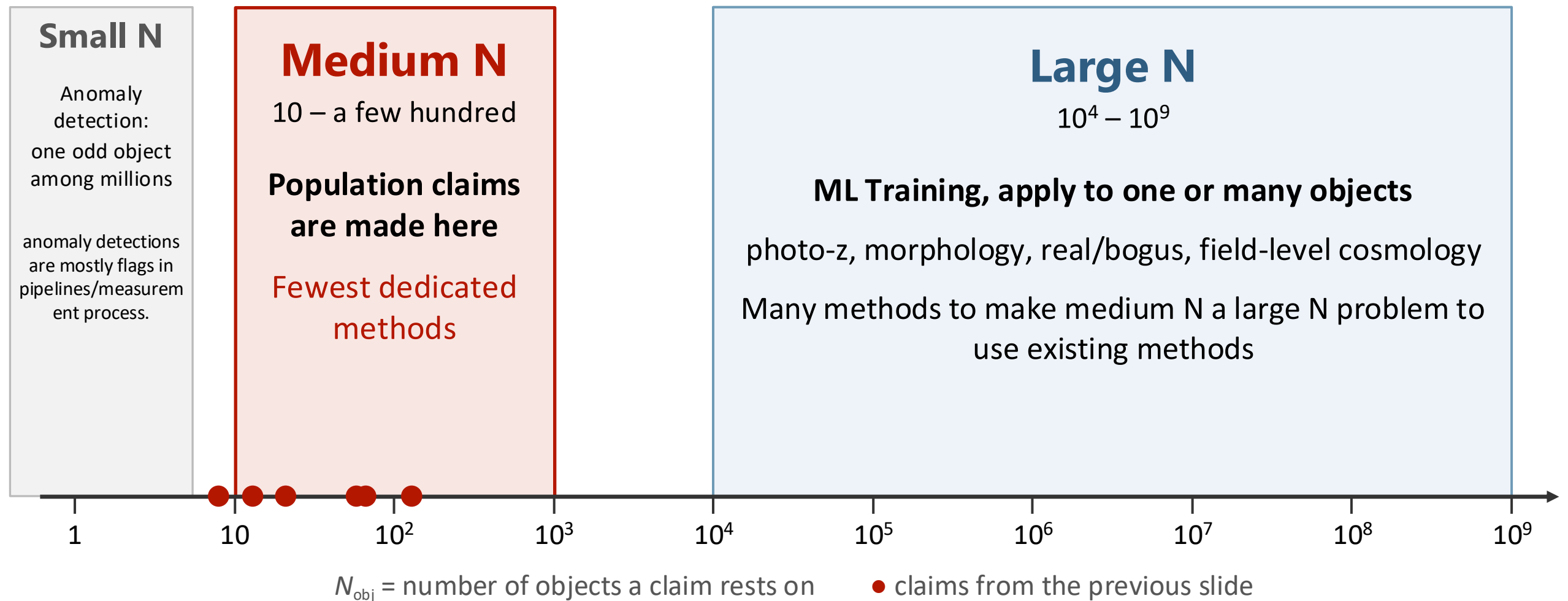
shear-selected clusters

HSC cluster cosmology

Chiu+ (HSC-SSP Y3)

Each discovery/claim was made with $N \sim 10\text{--}10^2$!

The Number-of-Objects Axis (Simplification)



The frontier stays medium N even as experiments become better

Each deeper survey finds a new population at small N

**Most ML methods target large N or single anomalies;
population/physics claims are made at $N \sim 10-10^3$.**

There are actually 3 numbers to consider

N_{obj}

Objects behind a claim

From a single object to
 10^9

N_{sim}

**Simulations or model
predictions**

Cost per prediction
varies by many orders of
magnitude

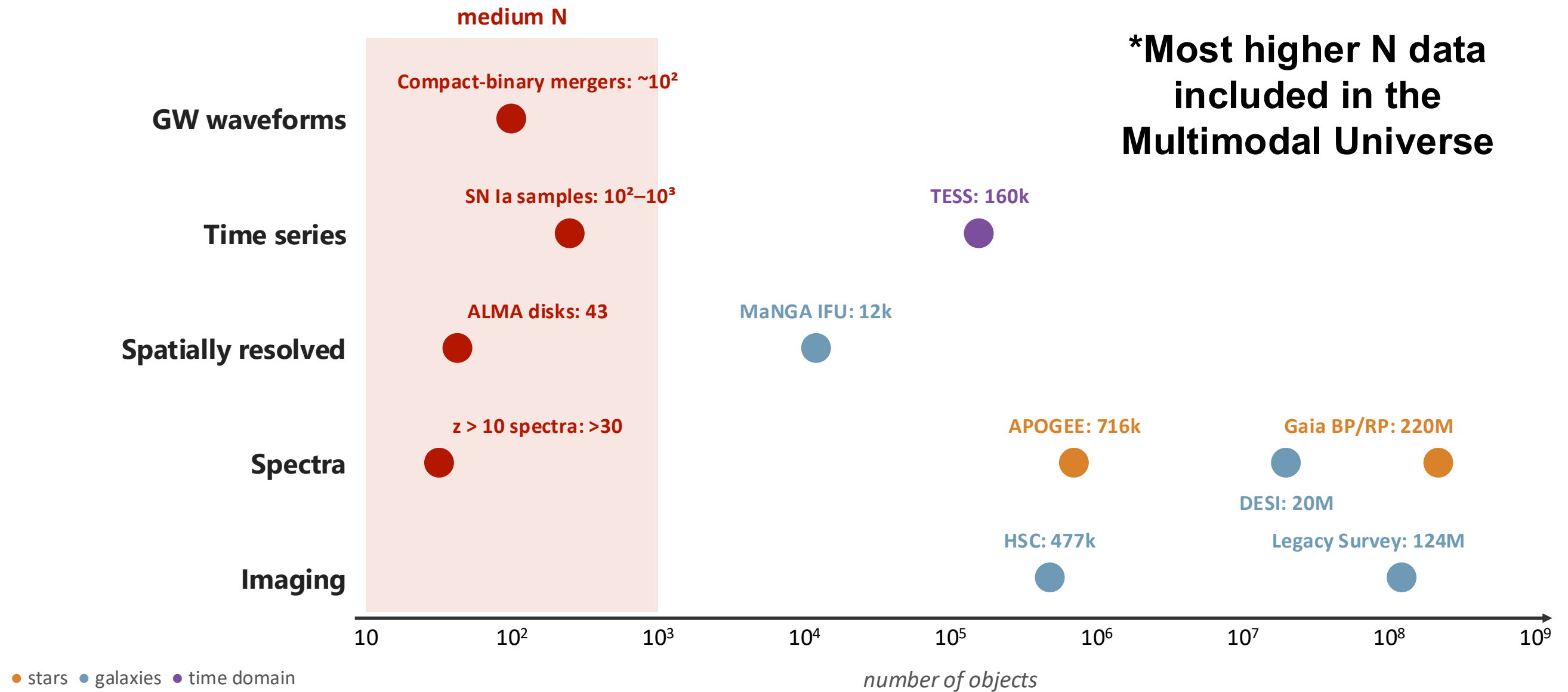
D

Information per object

From a few broad bands
to spatially resolved
spectra and time series

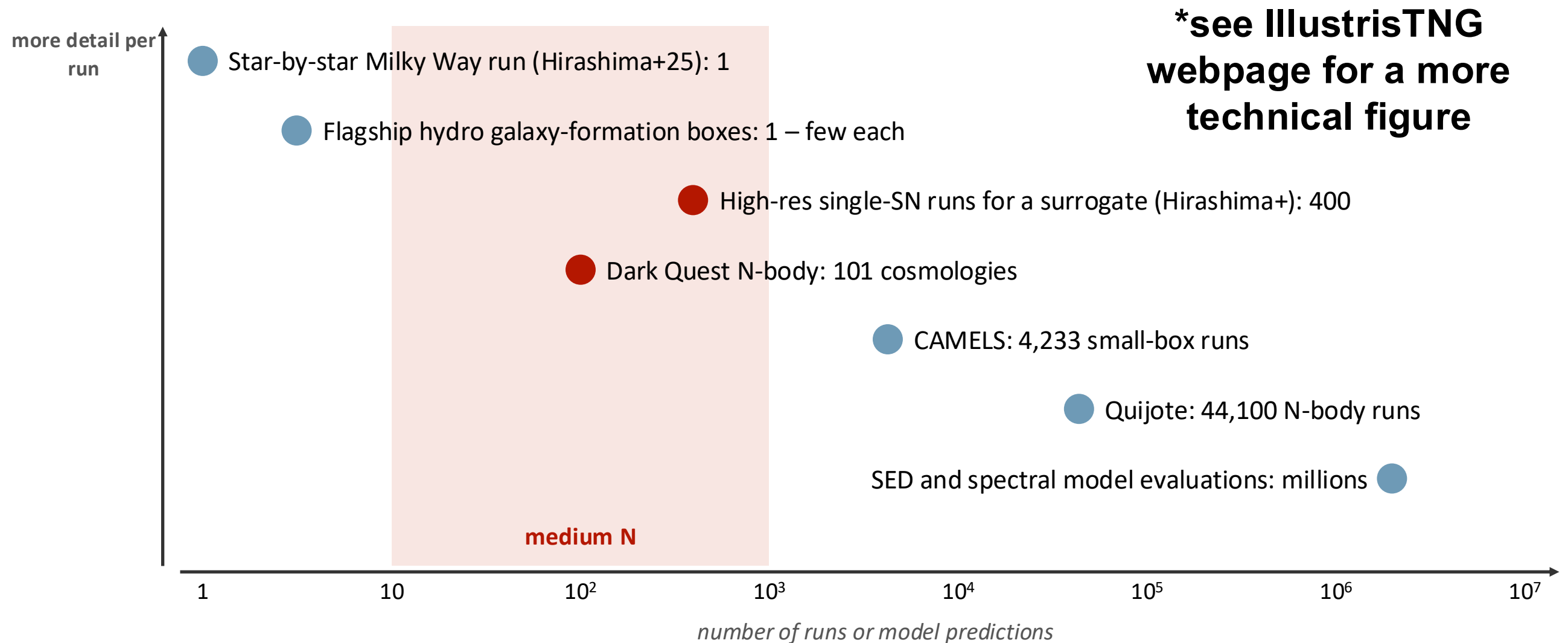
**Medium-N problems: few objects, rich data per object,
and few high-fidelity model predictions.**

The Data Space



**Information per object rises as the number of objects falls;
the information-rich samples are medium-N.**

The Simulation and Model Space



Why ML took off here: Quijote and CAMELS provide 10^3 – 10^4 runs in a form existing ML tools can use directly.

Trade-offs: a subset of the physics; small volumes. CAMELS boxes from 25 to 50 $h^{-1}\text{Mpc}$ are already hard; robust cosmology needs $\sim 10\times$ the length, $\sim 10^3\times$ the volume.

Multi-fidelity SBI: 100 high- + 10^4 low-fidelity runs come close to $\sim 28\text{k}$ high-fidelity runs (Thiele+25).

**High-fidelity, large-volume simulations stay at medium N.
Combining them with many cheap low-fidelity runs for ML**

Japanese contribution summarized

Large N observations

train once, apply widely

- HSC photo-z
- Morphology: Tadaki+20; GALAXY CRUISE classifier (~690k galaxies)
- Real/bogus and transient classification for HSC and Tomo-e (Morii+16; Takahashi+20, 22)
- SBI of galaxy physical properties from images and photometry (Alfonzo+)
- LIM, CosmoGLINT (Moriwaki, Yoshida)

Expensive simulations of medium N

emulators, surrogates, multi-fidelity

- Dark Emulator: 101 N-body cosmologies, used in HSC-Y3 (Nishimichi, Miyatake, Takada+)
- SN-feedback surrogate from 300 high-res runs (Hirashima+)
- Multi-fidelity SBI: 100 high- + 10^4 low-fidelity runs (Thiele, Bayer & Takeishi)

Few object observations

Often physics-heavy inference

- Sparse modeling with ISM statisticians: M87* (EHT); 43 ALMA disks (Honma, Akiyama, Ikeda; Yamaguchi+)
- ExoJAX: differentiable spectra for exoplanets and brown dwarfs (Kawahara+)
- High dimensional low sample statistics (Takeuchi+)

Large N, few-simulation emulators, single-object inference.

ML Inference methods that intrinsically work at medium N needs to be developed.

So far, medium-N solutions (ML?)
appears to be domain-specific.

Can there be a general domain-
agnostic framework?

Hierarchical Models for Optimal Inference

Objects share a physical process

- Measured objects are often not statistically independent. They are often from same physical process with connected histories
- Each object borrows constraints from the others.

Heterogeneous data, new physics

- Model many low-fidelity objects together with a few high-quality ones
- SEDUniverseMachine: trend of nebular log U with sSFR, from population photometry alone
- Stacked SDSS spectra: a few points in the same plane (Kaasinen+18)

Forward-model selection and observation

- Each dataset enters with its own selection and noise model
- Photometry, spectroscopy and different surveys in one likelihood
- Non-detections and upper limits enter directly

Posteriors do not simply multiply

- A product of per-object posteriors counts the prior N times
- Reweight by the interim prior (Hogg+10), or fit the population directly

When the question is about a population, a hierarchical model is the optimal use of the data.

Hierarchical models tend to be very high-dimensional.

Hierarchical Models for Medium N Inference

$$p(\theta | \{d_i\}) \propto \pi(\theta) \prod_i [\int L(d_i | x_i) p(x_i | \theta) dx_i] / \xi(\theta)^N$$

$$\xi(\theta) = \int P_{\text{det}}(x) p(x | \theta) dx$$

$$L(d_i | x_i)$$

Per-object likelihood

measurement uncertainty;
each object can come from a
different survey

$$p(x | \theta)$$

Population model

physics and priors; model
uncertainty enters here

$$\xi(\theta)$$

Selection function

detection, candidate choice,
follow-up success

$$\pi(\theta)$$

Prior

on population parameters

**Combining objects and surveys is a natural path for
inference at medium N.**

**Open problems: selection, per-object likelihoods, model
uncertainty, computation.**

Challenge 1: Selection Functions

$$\xi(\theta) = \int P_{\text{det}}(\mathbf{x}, \text{data quality}) p(\mathbf{x} | \theta) d\mathbf{x}$$

The detection probability depends on each object's properties and on the data it was observed with.

Depth and data quality

- Depth, seeing and S/N vary across fields and programs
- Samples are often merged from several surveys

Candidate Selection

- Color and photo-z cuts
- ML labels, anomaly scores, foundation-model similarity
- **These are part of the selection too**

Follow-up process

- Spectra confirm objects with strong lines more easily
- Time allocation decides which candidates get observed

At $N \sim 10\text{--}10^3$, errors in the selection function can be comparable to the statistical uncertainty.

At large N , completeness can be measured directly with injections.

ML methods can be used to learn the selection function too.

Challenge 2: Model Uncertainty

Cosmology

- Baryonic feedback on small scales
- Intrinsic alignments
- Photo-z calibration

Galaxies

- Star-formation histories
- Dust, IMF
- AGN contribution

Stars

- Convection and mixing
- Rotation, mass loss
- Binarity, atmospheres

Planets

- Clouds and hazes
- Chemistry
- Line lists

e.g. early JWST galaxies: ~ 2 dex in stellar mass from SED modeling alone (Wang+24)

Propagate model uncertainty into $p(\mathbf{x} \mid \theta)$

model averaging, flexible components, misspecification tests.

Challenge 3: Validation & Calibration

“when expressive regressions are used to label datasets, those labels cannot be used in downstream joint or ensemble analyses without taking on uncontrolled biases.”

“when ML models are used to emulate physical (or first-principles) simulations, they amplify confirmation biases.”

Hogg & Villar (ICML 2024)

Large N: a truth set exists

- Photo-z checked against spectroscopic redshifts
- Classifiers checked against labeled data
- **Error bars can be tested on real objects**

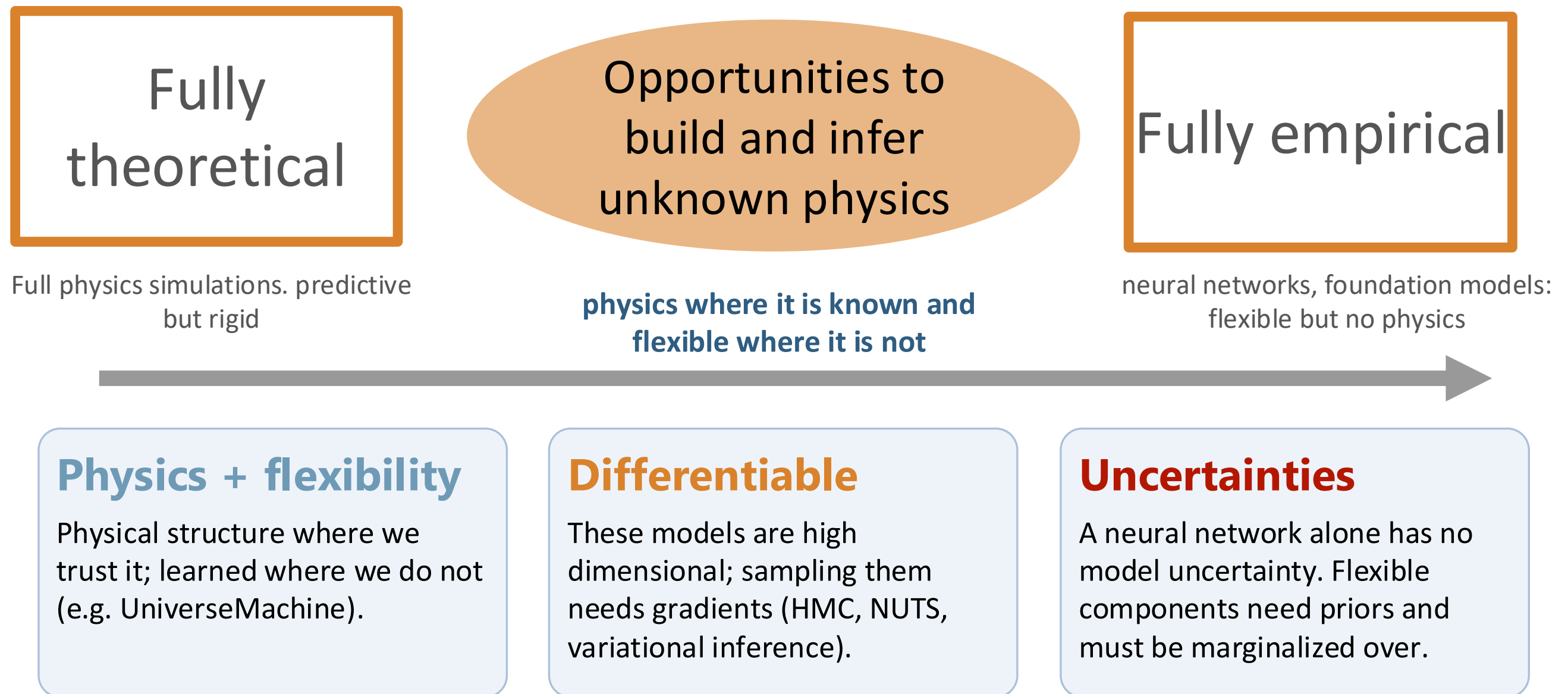
Medium N: often no truth set

- e.g. stellar masses of early JWST galaxies
- Checks are done on simulations
- ...which share input assumptions
- **Domain adaptation and calibration**

Tools today: injection–recovery through the real pipeline · checks across different simulators · posterior predictive checks · sensitivity to model choices

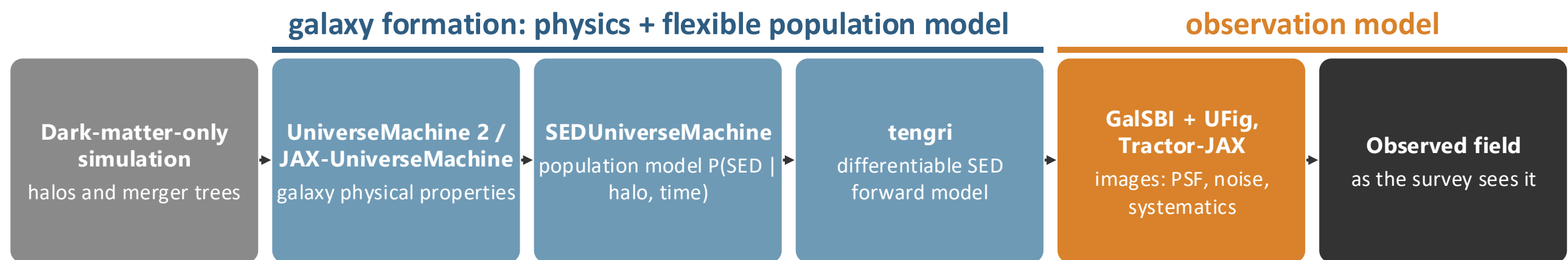
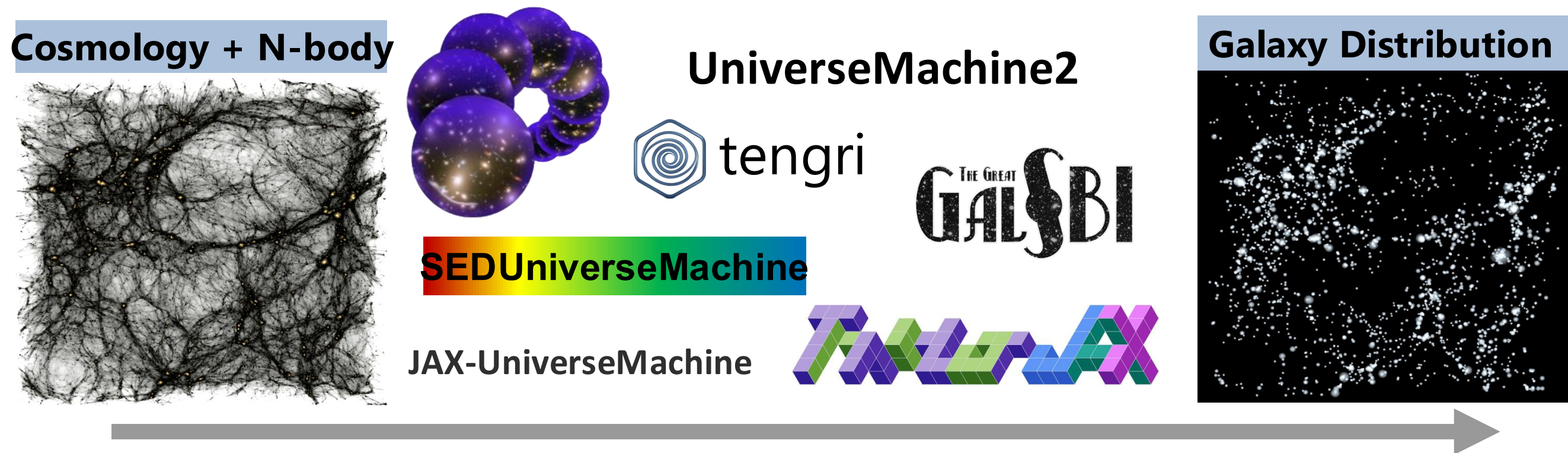
**Validation on simulations tests the method, not the model.
coverage tests on real data need a truth set.**

Physics-Informed, Flexible Models



Possible goal for the next 5 years are the models in the middle. Physics-informed, flexible, differentiable, with uncertainties.

An example of physics-informed forward models²²



Goal: infer galaxy formation from the observed field itself, combine different observations to obtain observational constraints in a theoretical space (Digital Twin)

Complementary Routes to Medium N Inference

Physics-informed forward models and priors

- Encode what we know: physical and empirical models, priors
- Fewer objects for the same constraint
- Uncertainties and selection modeled explicitly
- Differentiability hugely helps as hierarchical models are high dimensional
- *e.g. empirical galaxy models, SED and radiative-transfer models, ExoJAX, tengri*
- **Risk: model misspecification**
- **Difficult to model multimodal data**

Foundation models

- Transfer from large-N to medium-N data
- Multiscale, multimodal: pixels → galaxy → environment; images + spectra
- *AION-1: 200M+ objects; Legacy Survey, HSC, SDSS, DESI, Gaia; 39 modalities; open weights*
- **Risks: cannot know what was missed; selecting with them adds a selection step; no calibrated uncertainties by themselves**

**Learned compression of high-dimensional data +
a physical forward model of population, selection and uncertainties.**

World Models? Models with language based inductive biases?

Key suggestions for the near future

1

Physical, flexible, differentiable models

fewer objects needed and explicit uncertainties.

2

Hierarchical model with selection functions

including ML- and foundation-model-based selection and learning the selection function.

3

End-to-end uncertainty propagation

including model uncertainty: averaging, flexible components, misspecification tests

4

Efficient use of expensive simulations

multi-fidelity inference, emulators, choosing which runs to do. BO-type approach?

5

Foundation models coupled to forward models

learned compression and likelihood surrogates.

6

Shared validation practice

standard injection–recovery test. Fidelity-flexibility balance. Posterior predictive tests.

Most of these apply across surveys and subfields.

Japan's position among large international projects

The landscape

- Largest surveys and AI compute led by the US, Europe and China (Rubin, Roman, DESI, Euclid, SKA)
- Collaborations are becoming larger and more international
- **Surveys find candidates at large N; characterization is medium N**

1

Subaru HSC, PFS, and other surveys

deep imaging + wide-field multiplexed spectroscopy for characterizing rare candidates

2

International partnerships

ALMA, LiteBIRD, Hyper-Kamiokande, KAGRA, PRIMA, etc. without much need for alignment

3

High-performance computing

Fugaku and successors for high-fidelity simulations that enable efficient inference.

4

Statistics and methods

Historical strength in theoretical statistics. Including application in astro; sparse modeling, emulators

Opportunity: lead development and applications of methods for medium-N inference with multimodal, cross-survey data

Bringing Methods Researchers/Resources In

Center for Decoding the Universe

Stanford, since 2024: joint KIPAC + Stanford Data Science center pairing astrophysicists with statisticians and ML researchers

NSF–Simons AI institutes

SkAI (Northwestern-led, with Chicago, Illinois) and CosmicAI (UT Austin-led): \$20M / 5 yr each; surrogates, inverse problems, time-domain foundation models, agents

Research credits and compute

AI labs and cloud providers offer research credits; a coordinated community request carries more weight

In Japan: Nagoya KMI, Kavli IPMU CD3, RIKEN AIP, ISM, and other institutions as seeds of a joint astro/ fundamental physics × statistics/CS collaboration?

Methods researchers engage when our problems are posed in their language.

Some questions for us

1. Which open problems in your area are truly limited by the number of objects/data and not by systematics/modeling/inference?
2. Which published conclusions might change with a proper hierarchical model that includes proper uncertainty propagation?
3. Where should the boundary between physics and learned components sit in your models?
4. Which of your models is "fully theoretical", which "fully empirical", and what would the middle version look like?
5. How should we validate results based on a few objects, with no truth set? What statistical methods need to be developed?
6. What problem statement would make a statistician or computer scientist want to work with us?