# Statistical methods for large sky surveys: opportunities and challenges

Rachel Mandelbaum (Carnegie Mellon University)
May 2026, STAMPS workshop

# What is weak lensing, and why should you care?

# Illuminating a dark Universe

Our current cosmological paradigm, ΛCDM: the Universe is dominated by "dark" contents, which determine the Universe's expansion and rate of structure formation.

Testing this paradigm requires good ways to precisely measure the impacts of those dark components.

Revealed through gravitational influence on other matter

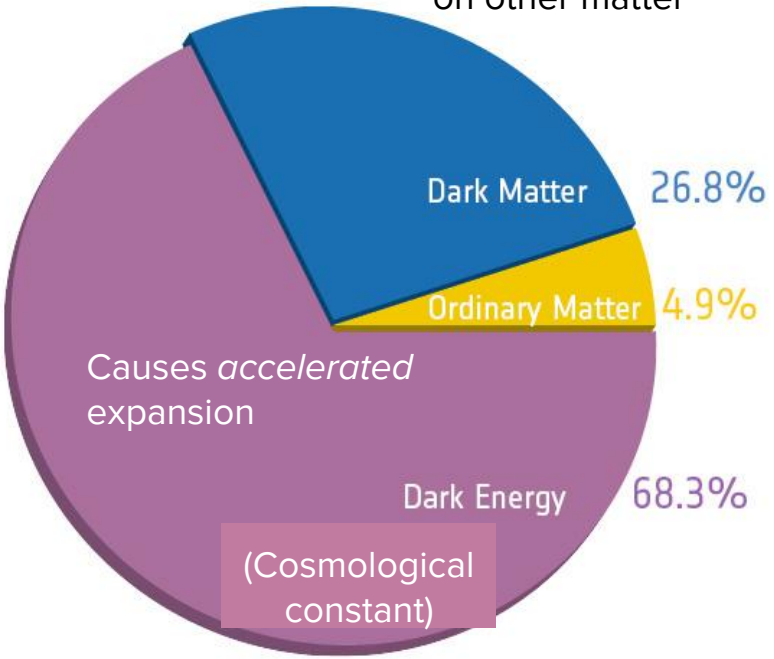
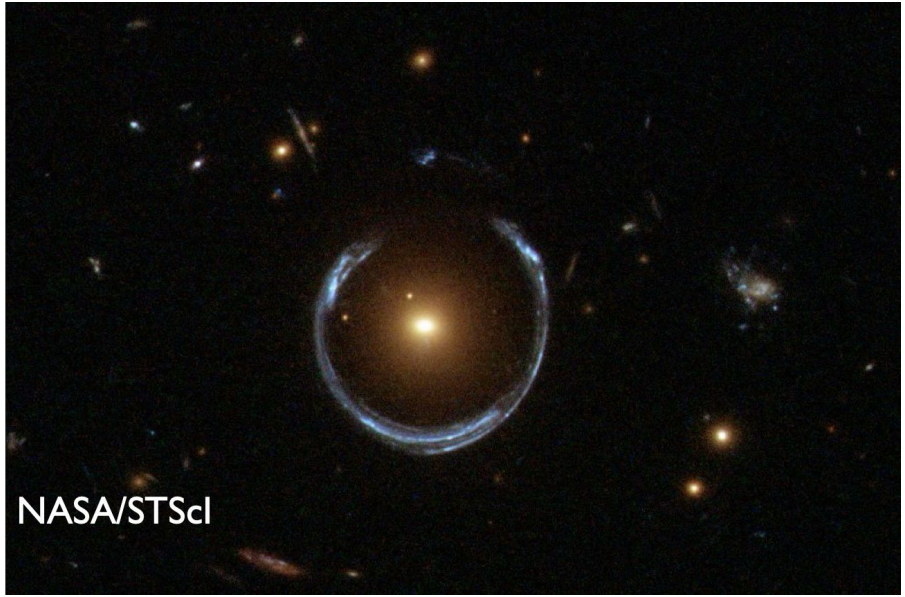



Image credit: ESA & the Planck collaboration

# Gravitational lensing: the deflection of light by (all) mass

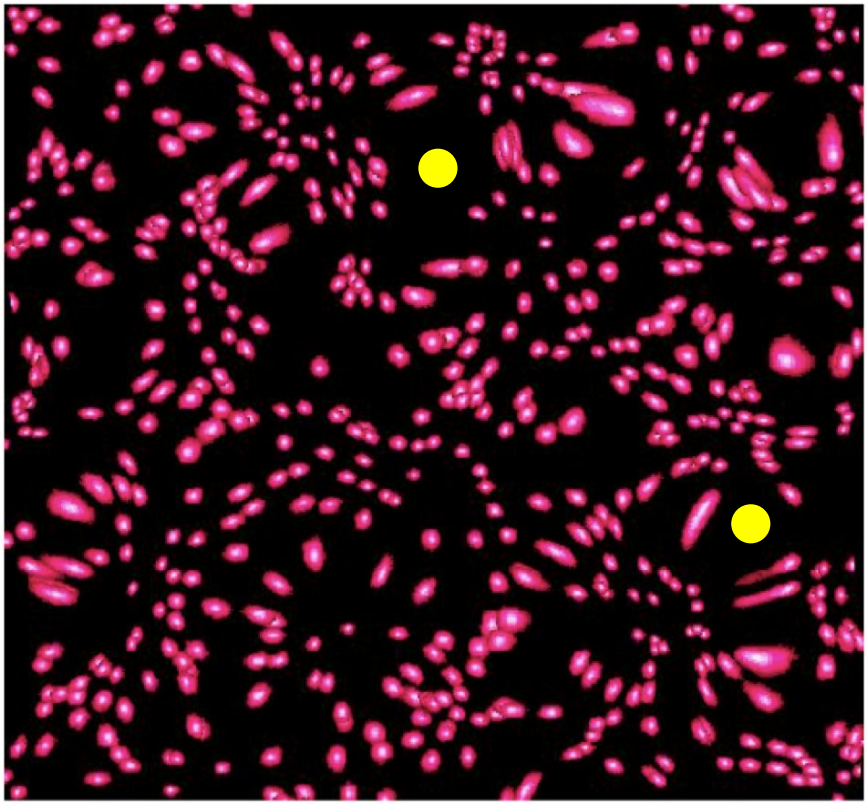
- Matter (baryonic & dark) distorts space-time, changes paths taken by light rays from distant objects
- Strong lensing depends on serendipitous alignment

Strong lensing




Disadvantage of weak lensing: it produces only one mildly distorted image
Advantage: it happens *everywhere*

# Weak lensing imprints coherent patterns (“shear”) in galaxy shapes



# Galaxies are not round (but that's OK)

Image credit: NASA / ESA / IPAC / Caltech / STScI / ASU

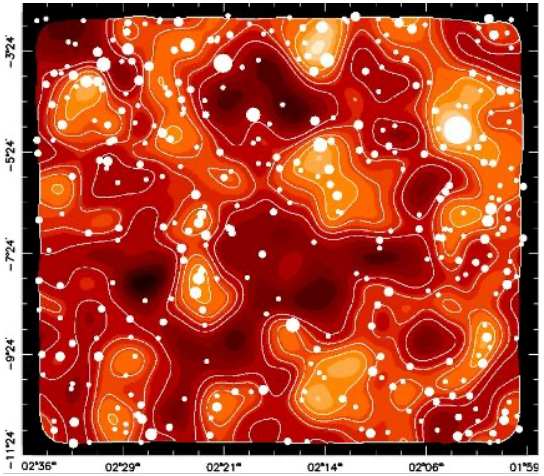
We measure weak lensing (WL) statistically:

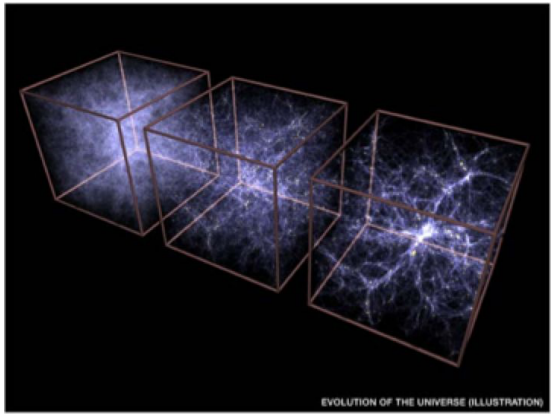
- measure coherent galaxy shape distortions (<~1%) ...
- underneath ~30%-level ellipticities (**statistical** error)...
- and non-lensing effects that cause comparable distortions (**systematic** error)

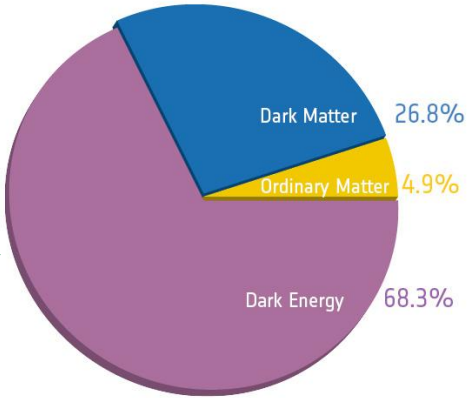
Assumption: we can model/remove non-lensing effects that cause coherent shape distortions.

# Weak lensing has exciting scientific applications

What are dark matter and dark energy?



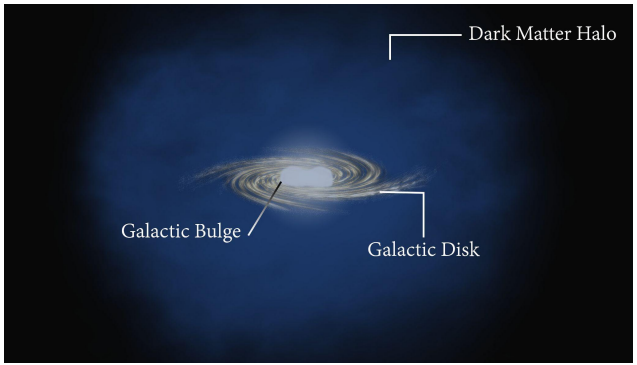



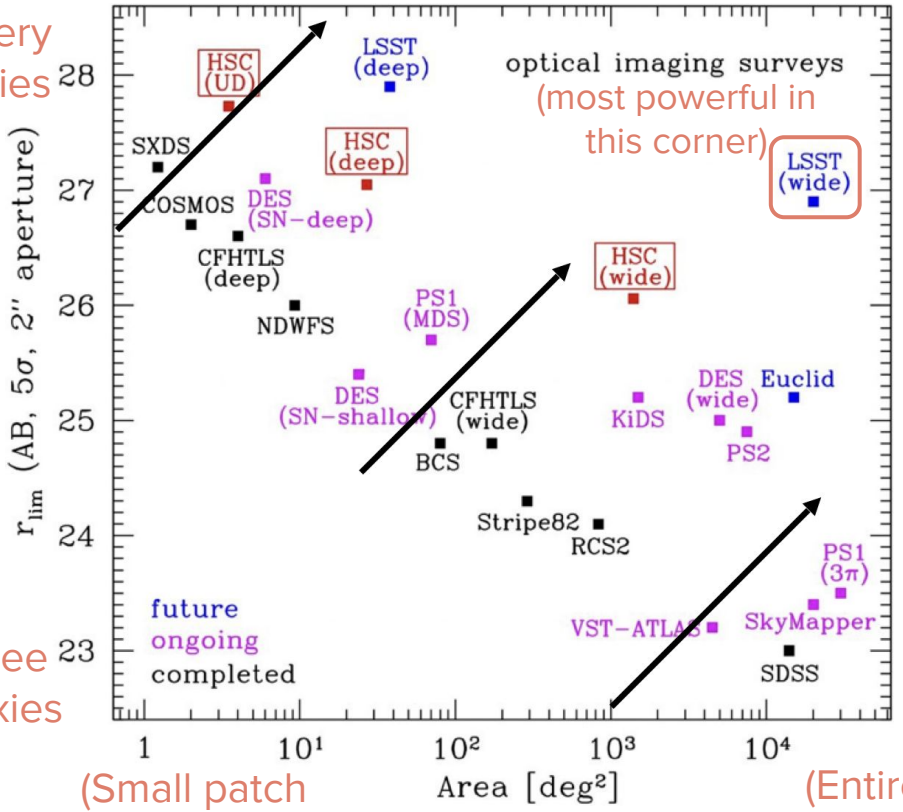

How has cosmic structure grown over time?

How is matter distributed in the Universe?

How do the visible parts of galaxies relate to their dark matter halo?



Can see very faint galaxies




Can only see bright galaxies

(Small patch of sky)

(Entire sky)

In the past two decades, imaging surveys have produced increasingly powerful and complex datasets...
Not just images but also spectra, time series, and more

Image credit: the HSC survey collaboration

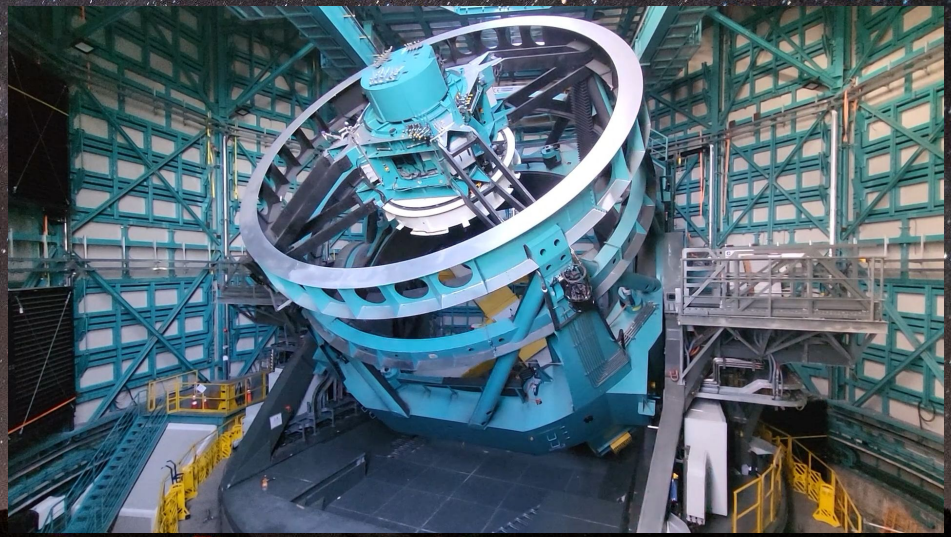
# Vera C. Rubin Observatory's Legacy Survey of Space and Time

**What is Dark Matter and Dark Energy?**

**How is dark matter distributed throughout the universe?**

**Can we find Planet 9?**

**How are stars born and how do they die?**



**Challenge: how do scientists make discoveries** using data sets of 40 billion stars and galaxies, 30 trillion observations, 500 PB of images?

# Areas where statistical methods can help

# Areas in weak lensing where new statistical and AI approaches could help

Statistical uncertainties:

 Extracting all the information possible from these rich datasets

Systematic uncertainties: data analysis challenges where advances are needed beyond what traditional modeling/analysis approaches can do

 Estimating distances to galaxies

- Producing realistically complex simulations of our theoretical models to test analysis methods and/or for use in likelihood-free inference methods (Generative AI)

For other science with LSST, like time domain science, there are other opportunities

# Beyond two-point statistics: there is more information!

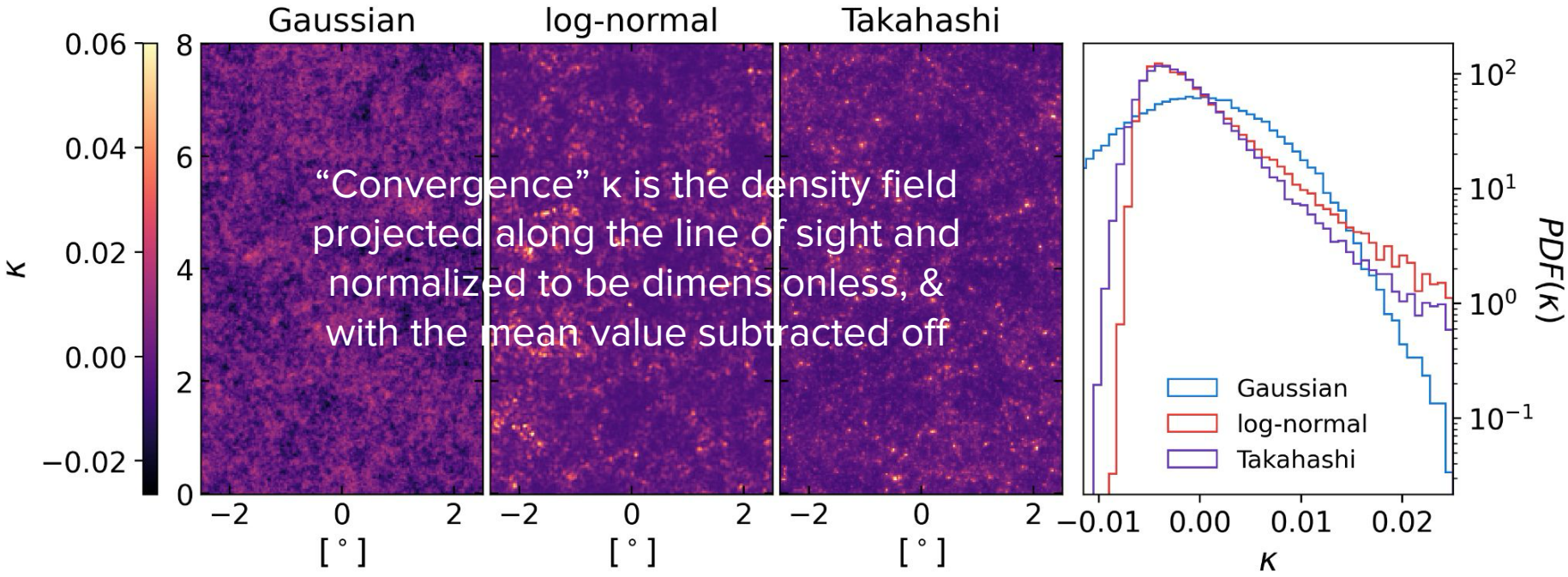



Figure from Alan Zhou (UChicago), Zhou+24

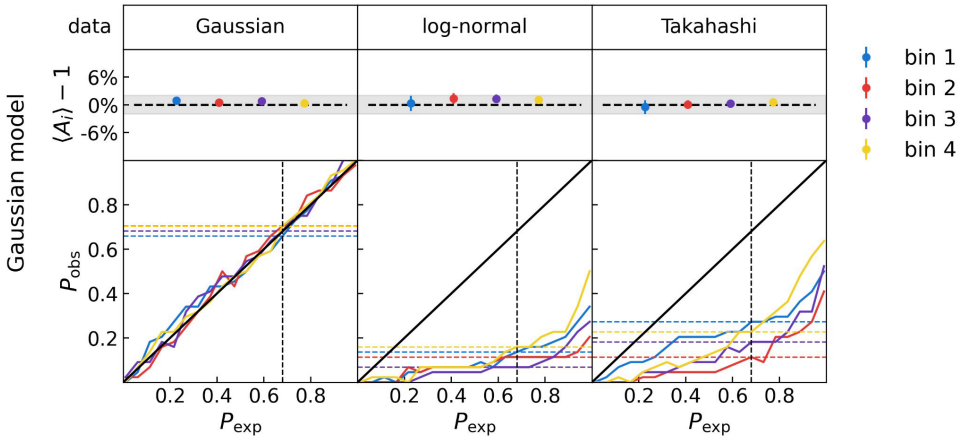
# Key takeaways from recent work

Zhou+24 (Bayesian hierarchical inference):

Model misspecification can bias inferred cosmological parameters *and* uncertainties

Identified approach that can work at ~2% level: Gaussian map prior for cosmological parameter estimates with uncertainties from lognormal map prior

Top row: is the estimate of the power spectrum amplitude unbiased?



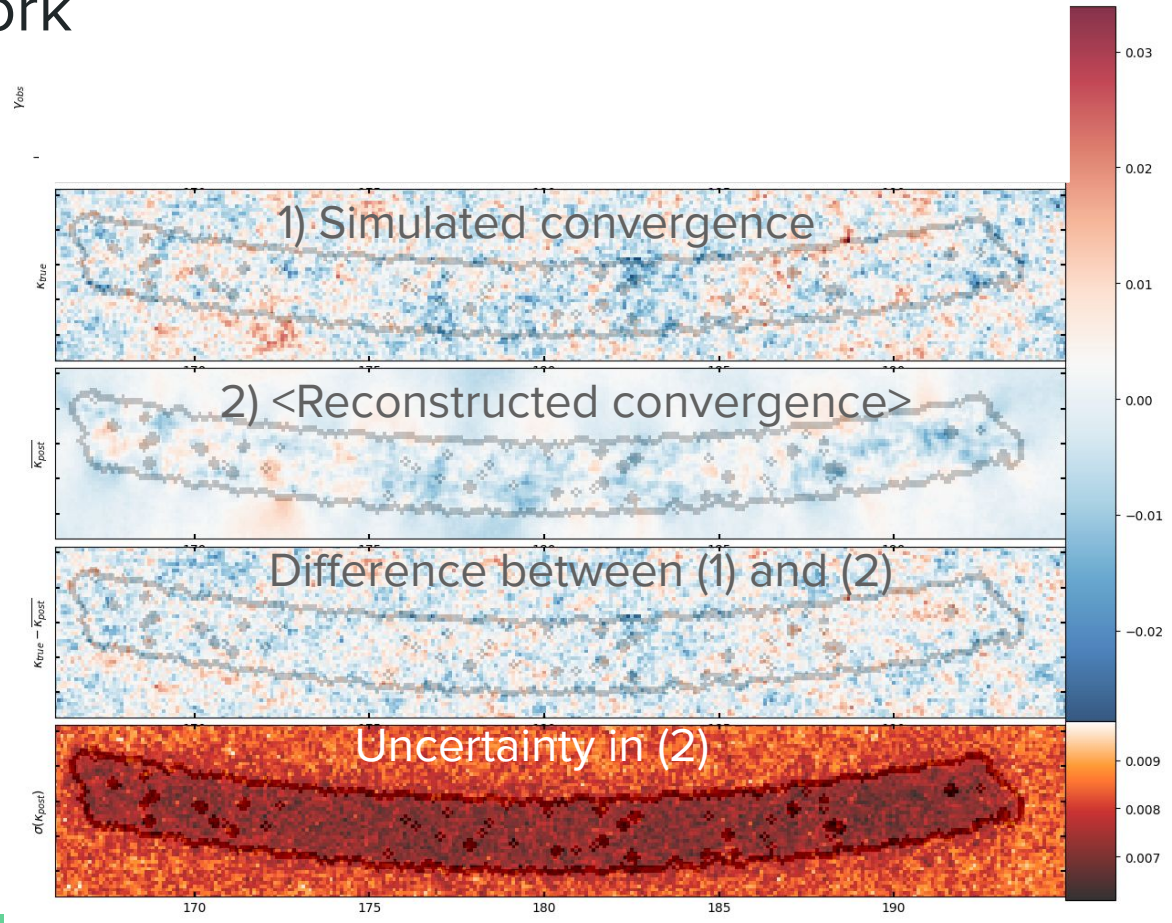

Bottom row: is the estimated uncertainty in the power spectrum amplitude unbiased?

# Challenges & future work

Follow-up work includes additional systematics modeling needed to apply model to real HSC data

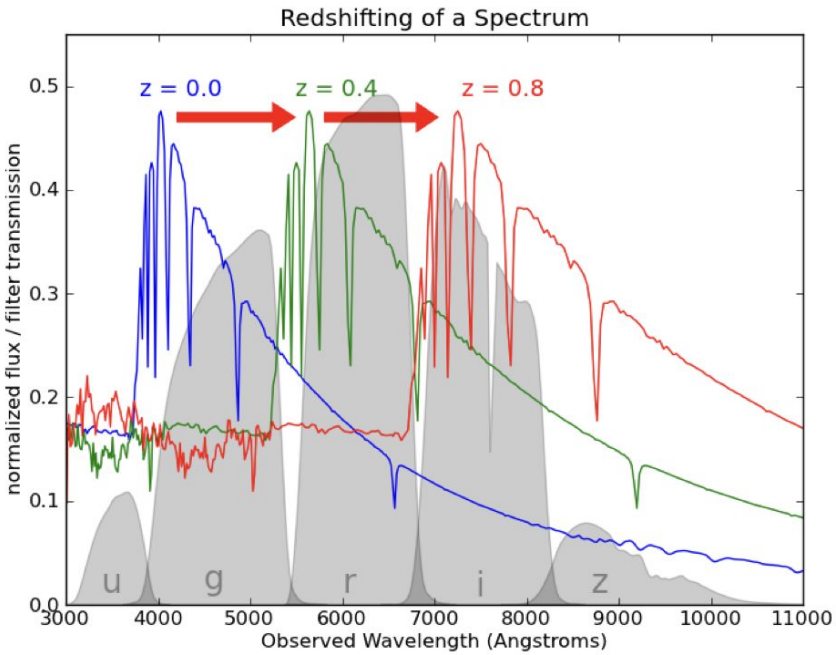
Translating decades of experience on two-point statistics to the field level!

Could also be helpful for non-homogeneous survey data

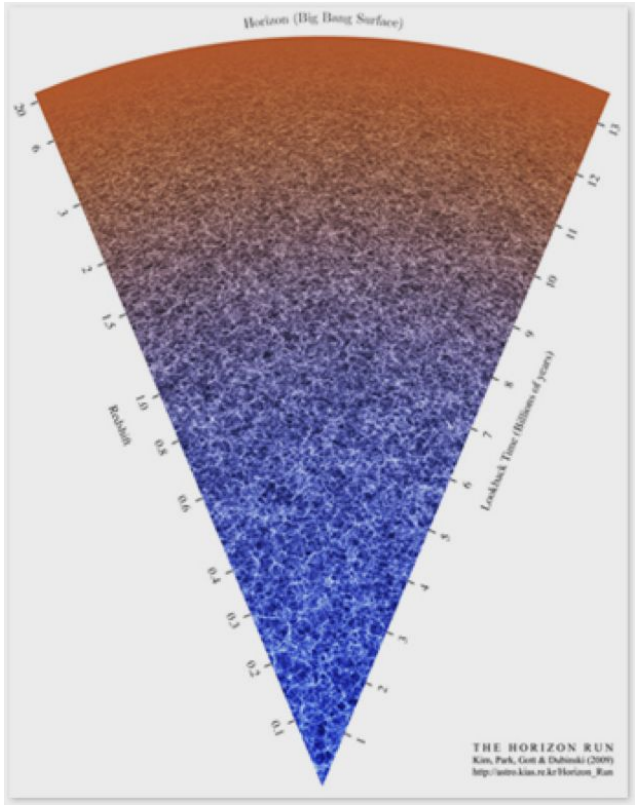


# Redshift estimation

We need to know the redshift distribution of the galaxy population used to measure weak lensing.

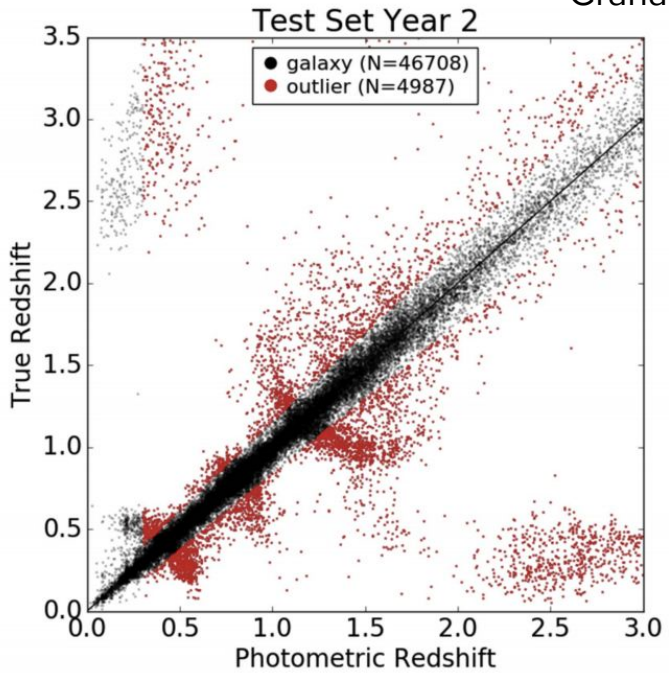


# The challenge of *photometric redshifts*: often resolved with ML




Imaging data

Graham+18




Photometric redshifts: redshift estimates from broad-band photometry rather than from spectra.

# For WL, we care about ensemble redshift distributions

It’s OK if individual photometric redshifts are not so great!

Statistical inference of the redshift distributions of galaxies based on cross-correlations with reference (spectroscopic) samples is a way forward

- Original idea from Newman (2008)
- Two groups have pioneered Bayesian hierarchical implementations with different methodology: Sanchez & Bernstein (2019); Rau, Wilson, RM (2020) and Rau et al (2022) - clustering cross-correlations joint with photometry

# Solving the inference problem for redshift distributions with machine learning

Statistical inference of the redshift distributions of galaxies based on cross-correlations with reference (spectroscopic) samples is promising:

- Reference sample doesn't have to be representative.
- It works even if blending affects photometric galaxy measurements.

Bayesian hierarchical implementations: Sanchez & Bernstein (2019); Rau, Wilson, RM (2021).

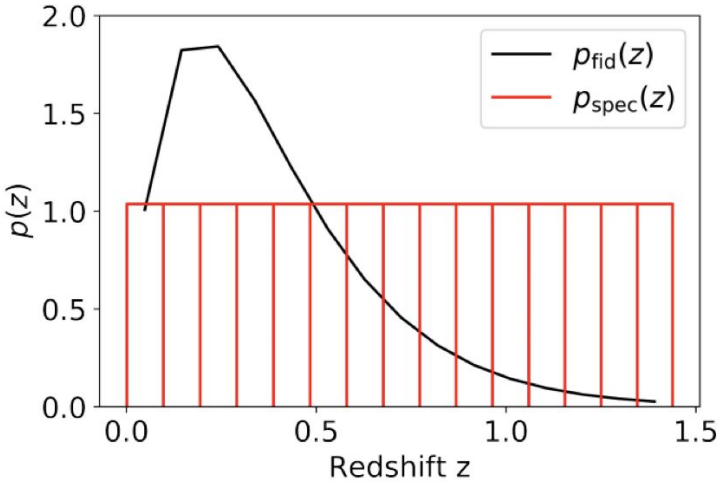
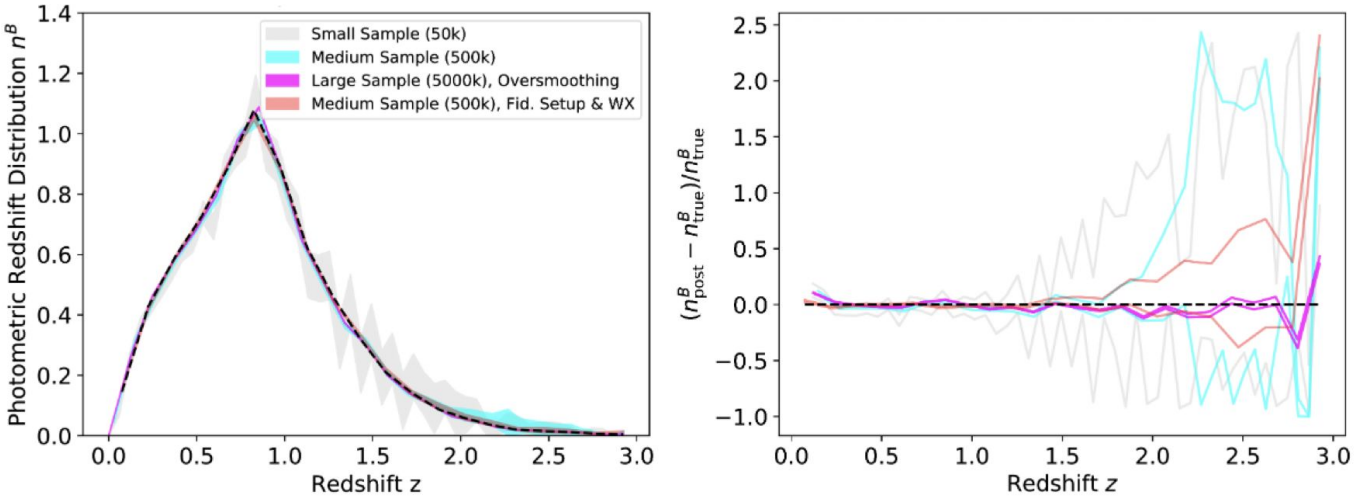



Figure credit: Markus Rau

# But wait, there's more (information)

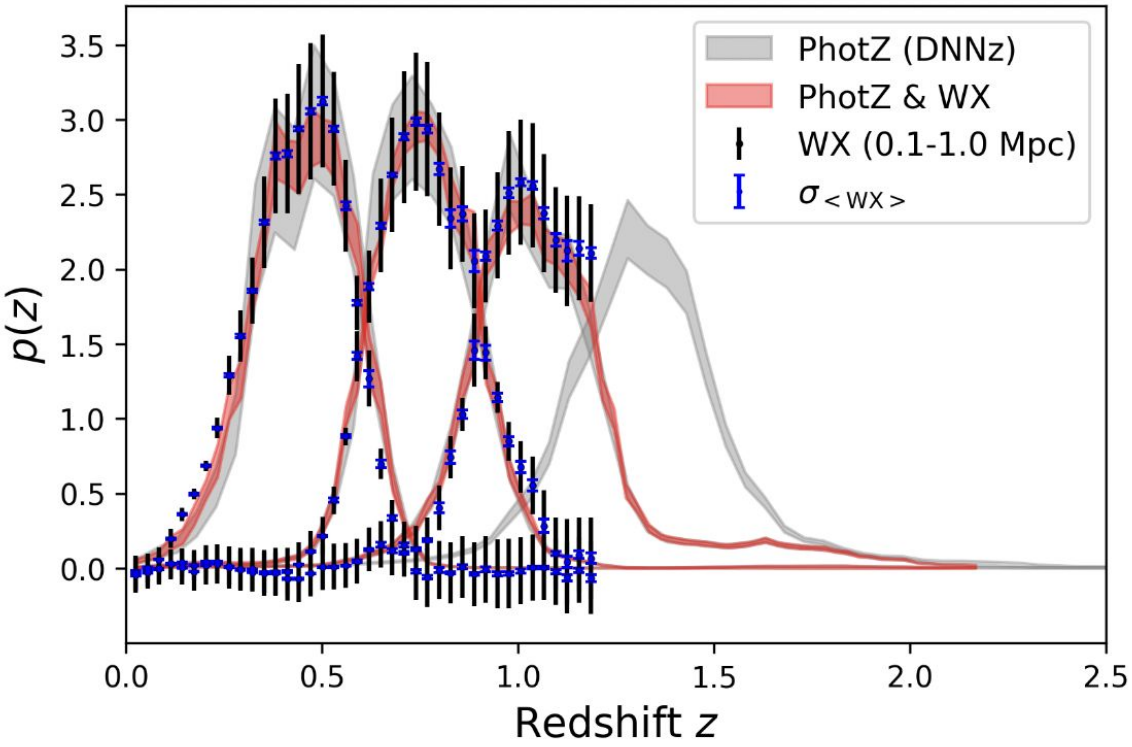
Galaxy colors provide photo-z information: we can build a composite likelihood including clustering and photo-z information.




Rau+2022 (arXiv:2101.01184): proof of concept using simulations, with efficient sampling techniques.

# Trying this out on HSC, in preparation for LSST

Results are promising: the two sources of information are complementary but still give largely consistent results




Main challenges: appropriately encapsulating *systematic* uncertainties in p(z) by building the model appropriately; and acquiring reference samples at high z

Rau+2023

# Software infrastructure for robust application of statistical analysis and AI at scale

# LINCC Frameworks: led by CMU, UW, LSST-DA

https://lsstdiscoveryalliance.org/programs/lincc-frameworks/

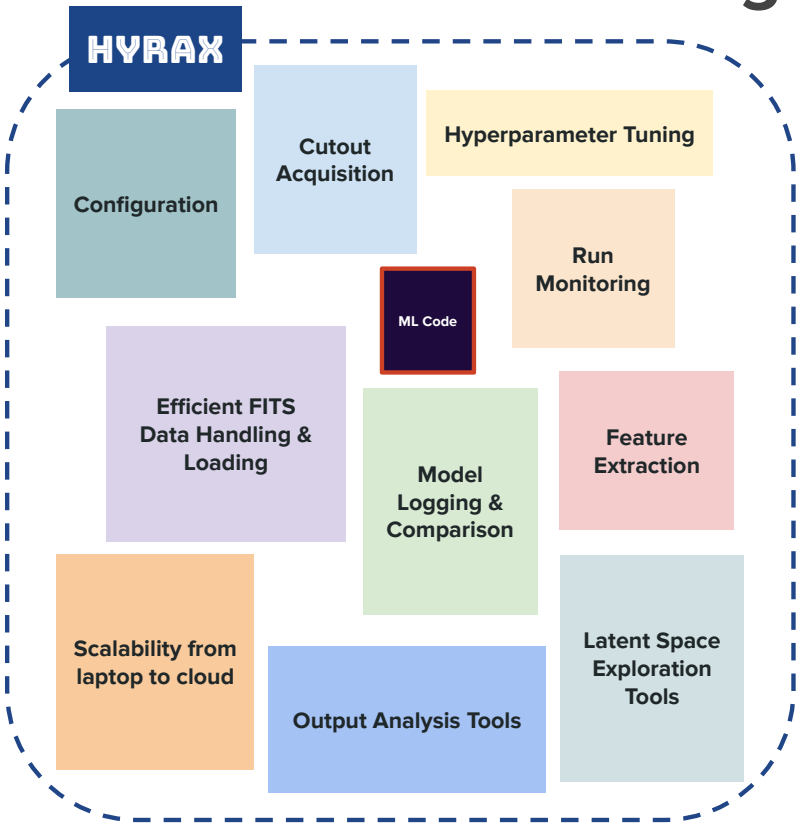
**Mission**: LINCC Frameworks' mission is to enable scientists by developing scalable and productionized software/algorithms in collaboration with broader Rubin community.

Specifically focused on supporting the Rubin Observatory's Legacy Survey of Space and Time (Rubin LSST) data.

Enabling ML/AI applications is a strong interest of the team!

Credit: NSF-DOE Rubin Observatory

# Hyrax: we bring the infrastructure, you bring the machine learning algorithms




✅ **Effortless Scaling**
Seamlessly move from laptop to HPC, from zero to many GPUs, with no code changes.

🌌 **Survey-Aware Data Downloaders**
Easily acquire cutouts from HSC, LSST, and more to come.

📦 **Comprehensive Data Organization**
Data, configurations and results are tracked and stored automatically.

🔧 **Extensible and Open Source by Design**
Allows users to create custom models, datasets, metrics, visualizations, etc.

📊 **Built-In Visualizations and Diagnostics**
Diagnostic tools, embedding plots, and more.

Slide provided by Drew Oldag (UW)

# Spatial Analysis Software

## HATS (Hierarchical Adaptive Tiling System)

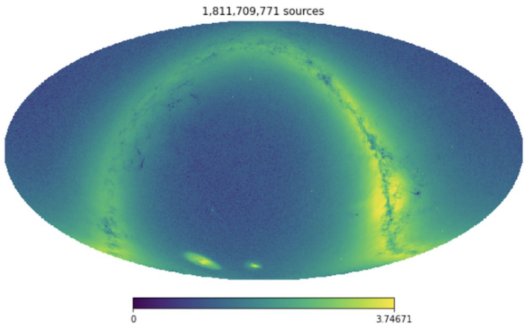
https://github.com/astronomy-commons/hats

Parquet-based format for storing files of spatial data.

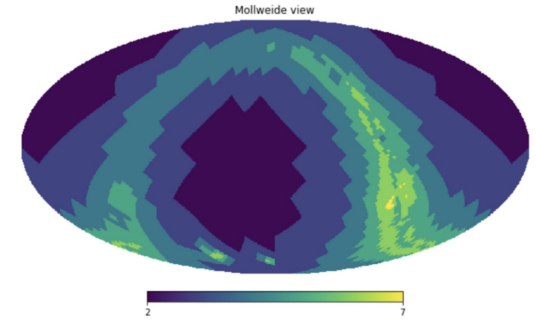
## LSDB

https://github.com/astronomy-commons/lsdb

Dask-based framework for parallel distributed computation over spatial data




*Gaia DR2 Catalog Counts (log scale)*




*Visualization of file storage (color = healpix level)*
*3933 partitions of similar size (128-256 MB)*

# Times Series (Nested-Pandas)

https://github.com/lincc-frameworks/nested-pandas

Provides the ability to nest dataframes within dataframes

**Goal**: Provide an easy to use (“pandas for light curves”), distributed framework for analyzing time series

- Supports user defined algorithms
- Distributed computation with Dask
- Built into LSDB

Supports non-time series data (e.g. spectra)

[15]:
```
import nested_pandas as npd
from nested_pandas.datasets import generate_data

nf = generate_data(5,3, seed=1).rename({"nested":"sources"}, axis=1)
nf
```

[15]:

| | a | b | sources |
|---|---|---|---|
| 0 | 0.417022 | 0.184677 | t flux band 0 8.383890 89.... |
| 1 | 0.720324 | 0.372520 | t flux band 0 13.704390 8.... |
| 2 | 0.000114 | 0.691121 | t flux band 0 4.089045 3.... |
| 3 | 0.302333 | 0.793535 | t flux band 0 17.562349 16.... |
| 4 | 0.146756 | 1.077633 | t flux band 0 0.547752 87.... |

[16]:
```
nf.query("sources.band == 'g'")
```

[16]:

| | a | b | sources |
|---|---|---|---|
| 0 | 0.417022 | 0.184677 | t flux band 0 8.38389 89.4606... |
| 1 | 0.720324 | 0.372520 | t flux band 0 8.346096 42.11... |
| 2 | 0.000114 | 0.691121 | t flux band 0 11.173797 95.... |
| 3 | 0.302333 | 0.793535 | t flux band 0 17.562349 16.... |
| 4 | 0.146756 | 1.077633 | t flux band 0 0.547752 87.... |

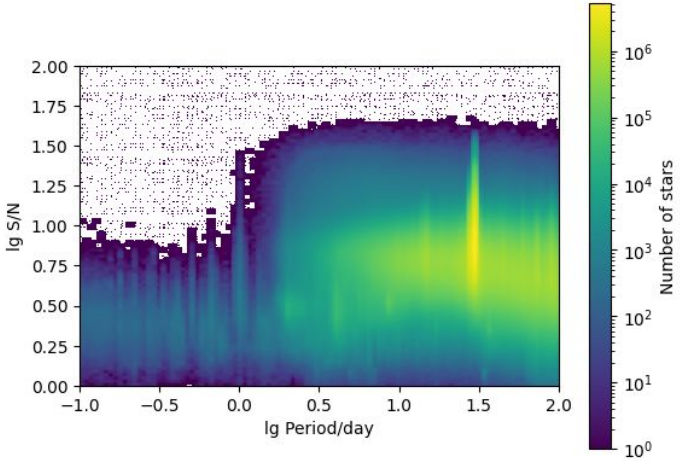
[17]:
```
nf.iloc[0]["sources"]
```

[17]:

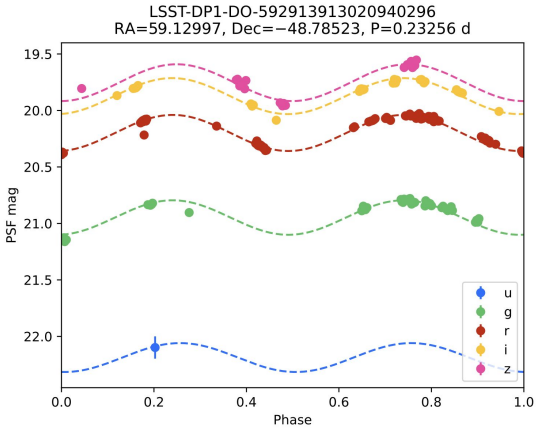
| | t | flux | band |
|---|---|---|---|
| 0 | 8.383890 | 89.460666 | g |
| 1 | 13.409350 | 9.834683 | r |
| 2 | 16.014891 | 31.551563 | r |

# LSDB: Applications

Distributed approach with spatial partitioning can scale up a range of analysis algorithms. Examples:




Light curve feature extraction (Lomb–Scargle)




Finding transient and variable sources
Source: Malanchev et. al. 2025


Estimating stellar distances via dust extinction (Palaversa et. al. 2025)

# Summary

- Weak lensing enables precise tests of our cosmological model, which push the limits of our understanding of astrophysics and observational systematics
    - Statistical approaches can help extract information beyond traditional approaches – especially hierarchical Bayesian networks
    - Care is needed in uncertainty quantification (especially systematic!) and avoiding model misspecification biases
- There are many other opportunities for improved scientific discovery through novel statistical approaches to complex, high-volume datasets
- Improved data infrastructure will be essential for working robustly at scale