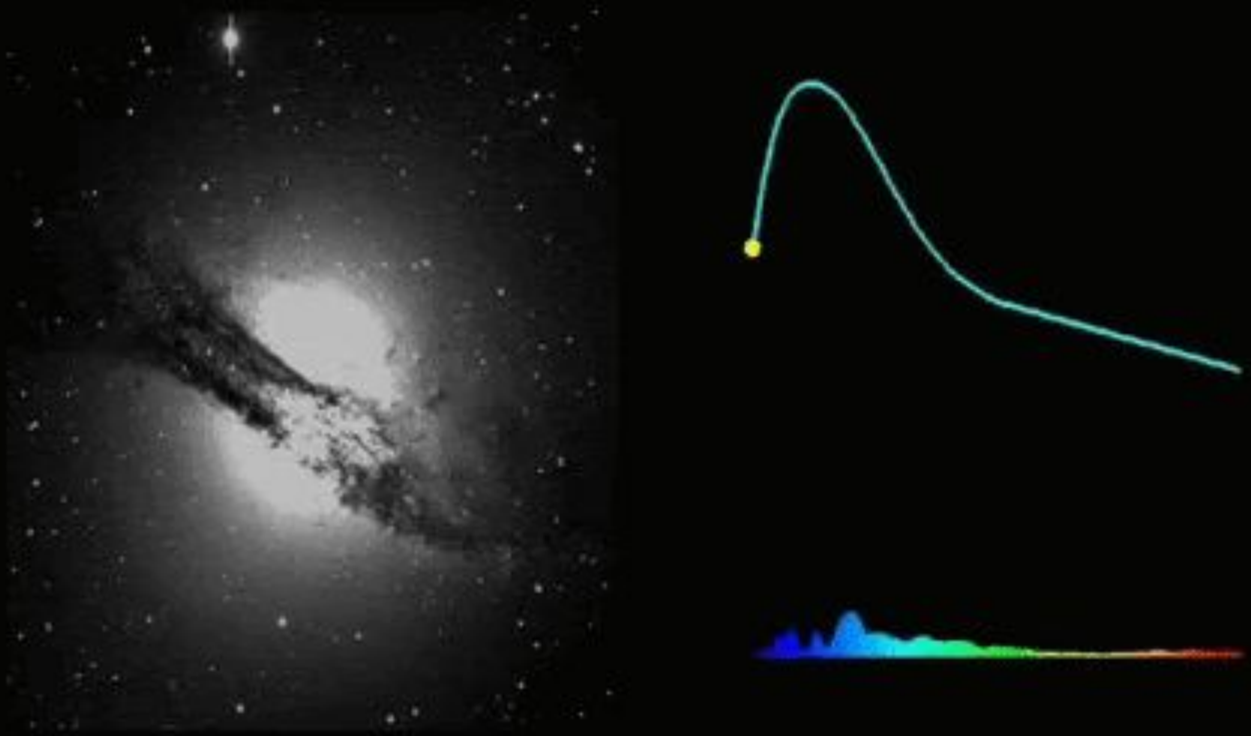
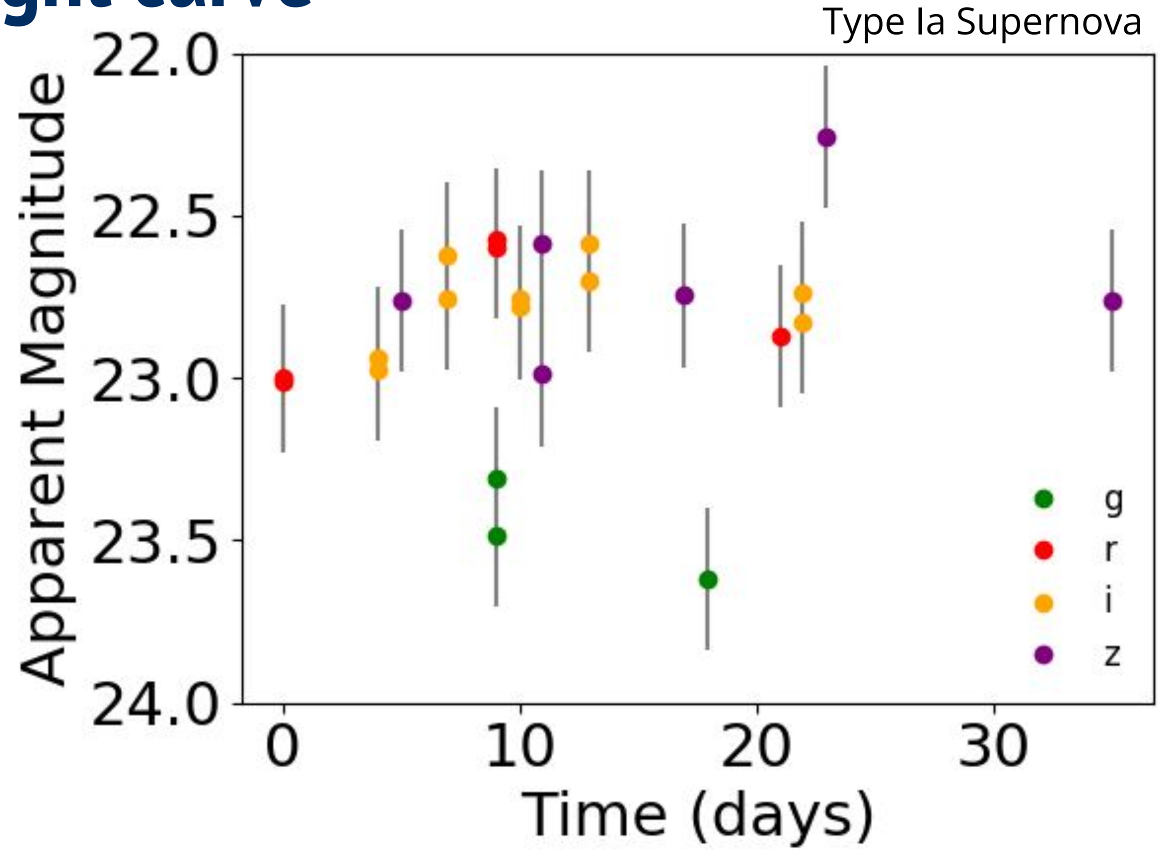
The Challenges of Identification, Classification and Inference for Time-domain Astrophysics with Big Data

V. Ashley Villar
Penn State, Astronomy Department
Institute for Computational and Data Sciences



Youtube: Magnetosheath

A *real* light curve



**sparse,
noisy,
irregularly-sampled**

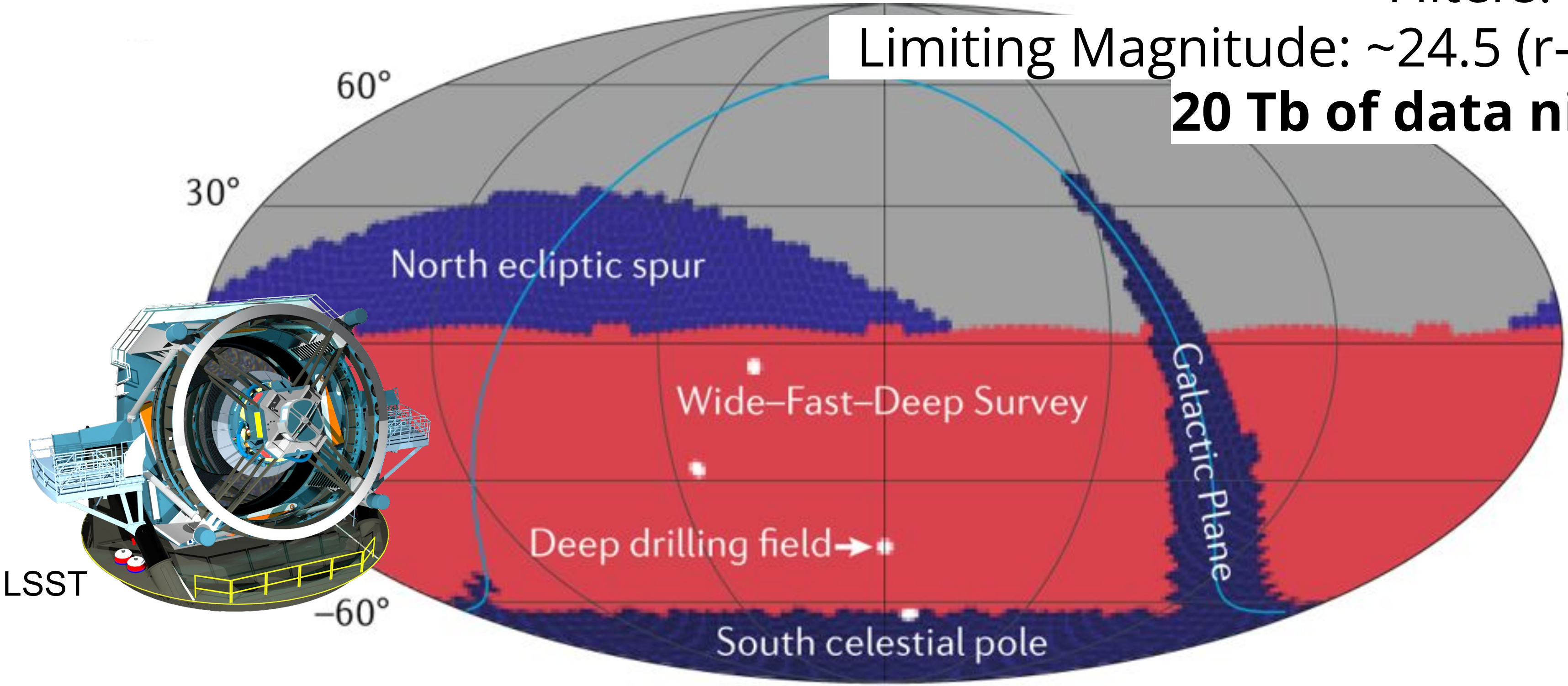
Vera Rubin Observatory, 2021

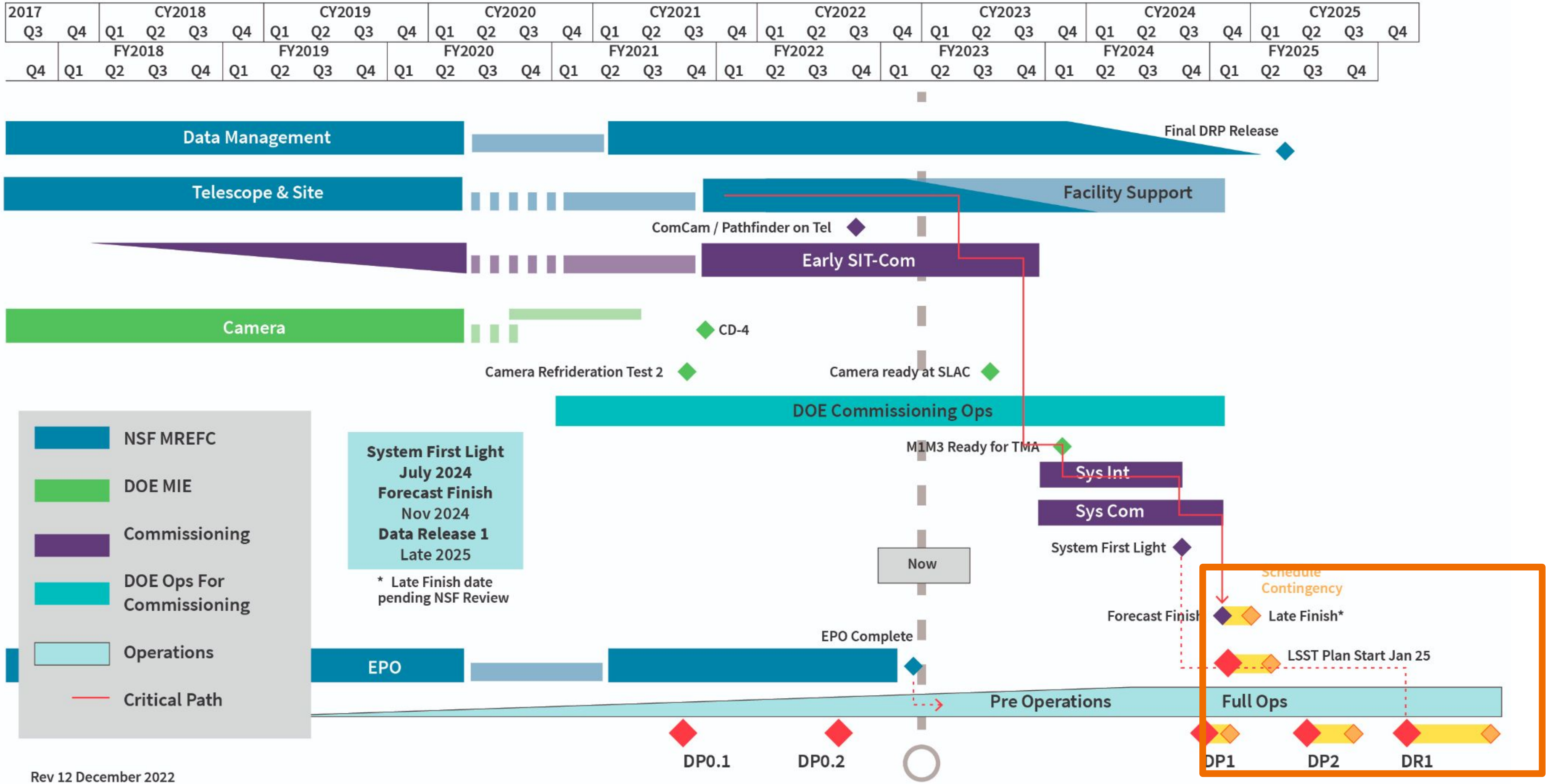
Legacy Survey of Space and Time

Filters: ugrizy

Limiting Magnitude: ~24.5 (r-band)

20 Tb of data nightly



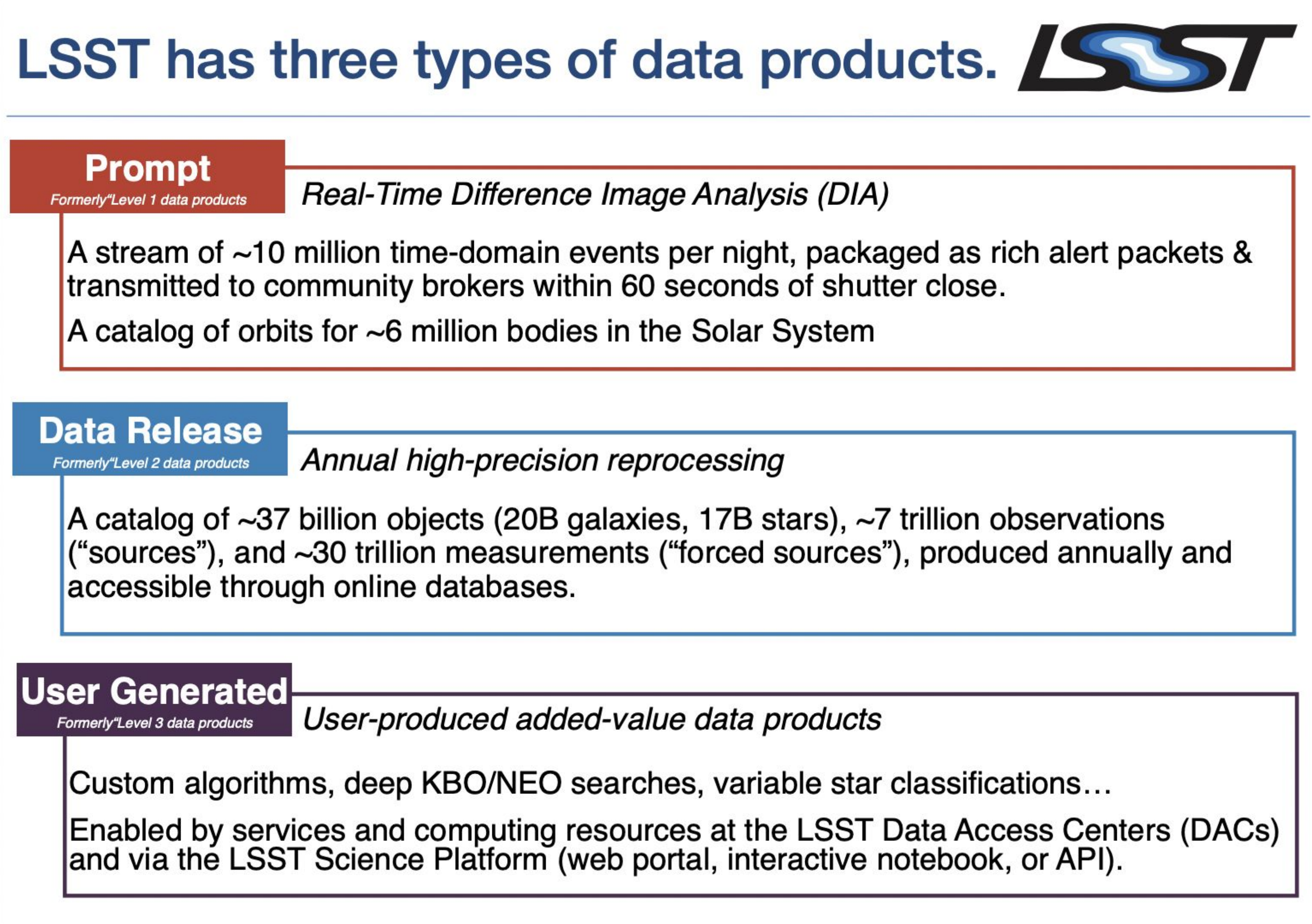


🟦	NSF MREFC
🟩	DOE MIE
🟪	Commissioning
🟦	DOE Ops For Commissioning
□	Operations
—	Critical Path

System First Light
July 2024
Forecast Finish
Nov 2024
Data Release 1
Late 2025

* Late Finish date pending NSF Review

Rev 12 December 2022



LSST has three types of data products.

Prompt

Formerly“Level 1 data products

Real-Time Difference Image Analysis (DIA)

A stream of ~10 million time-domain events per night, packaged as rich alert packets & transmitted to community brokers within 60 seconds of shutter close.

A catalog of orbits for ~6 million bodies in the Solar System

Data Release

Formerly“Level 2 data products

Annual high-precision reprocessing

A catalog of ~37 billion objects (20B galaxies, 17B stars), ~7 trillion observations (“sources”), and ~30 trillion measurements (“forced sources”), produced annually and accessible through online databases.

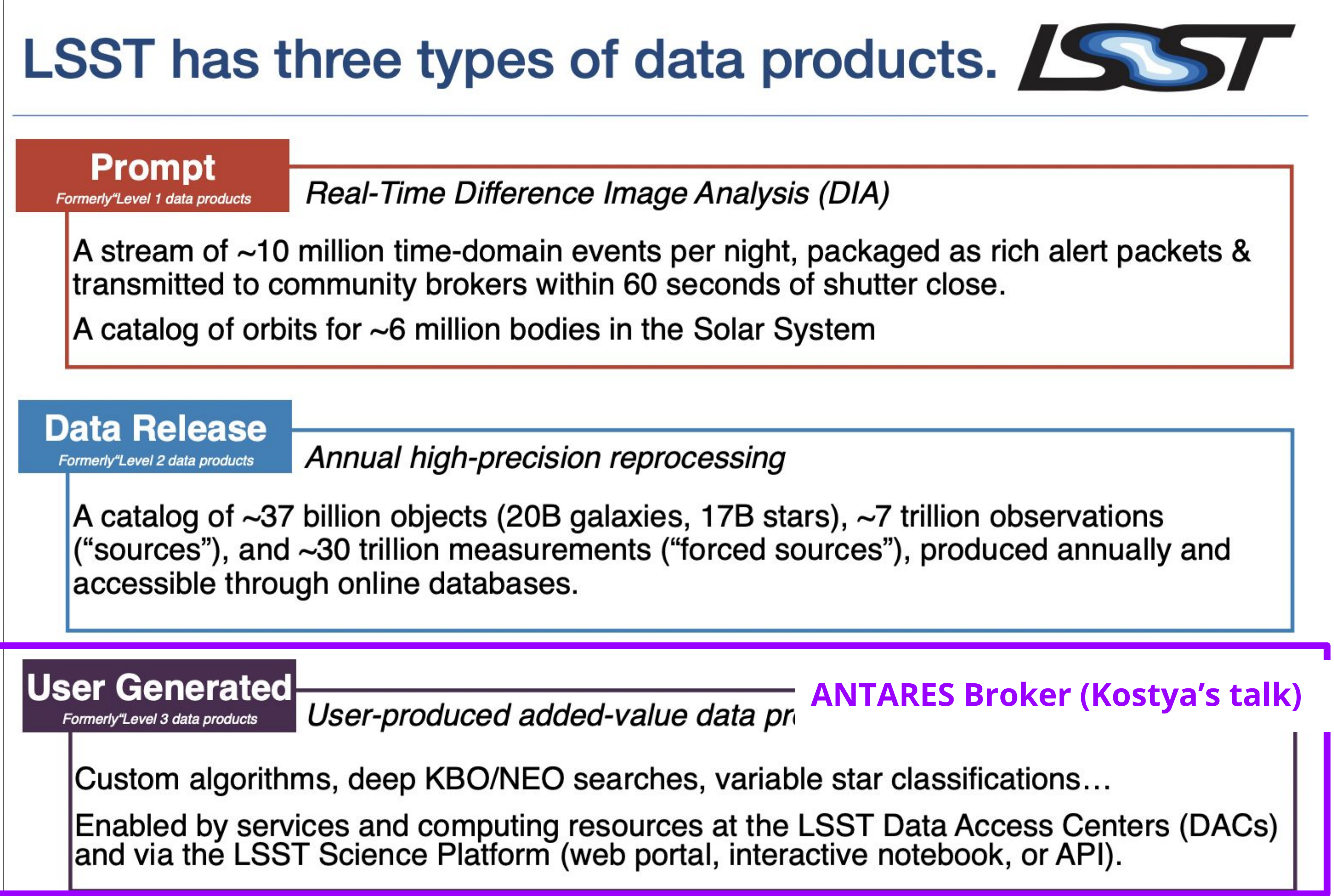
User Generated

Formerly“Level 3 data products

User-produced added-value data products

Custom algorithms, deep KBO/NEO searches, variable star classifications…

Enabled by services and computing resources at the LSST Data Access Centers (DACs) and via the LSST Science Platform (web portal, interactive notebook, or API).



LSST has three types of data products.

Prompt
Formerly“Level 1 data products

Real-Time Difference Image Analysis (DIA)

A stream of ~10 million time-domain events per night, packaged as rich alert packets & transmitted to community brokers within 60 seconds of shutter close.

A catalog of orbits for ~6 million bodies in the Solar System

Data Release
Formerly“Level 2 data products

Annual high-precision reprocessing

A catalog of ~37 billion objects (20B galaxies, 17B stars), ~7 trillion observations (“sources”), and ~30 trillion measurements (“forced sources”), produced annually and accessible through online databases.

User Generated
Formerly“Level 3 data products

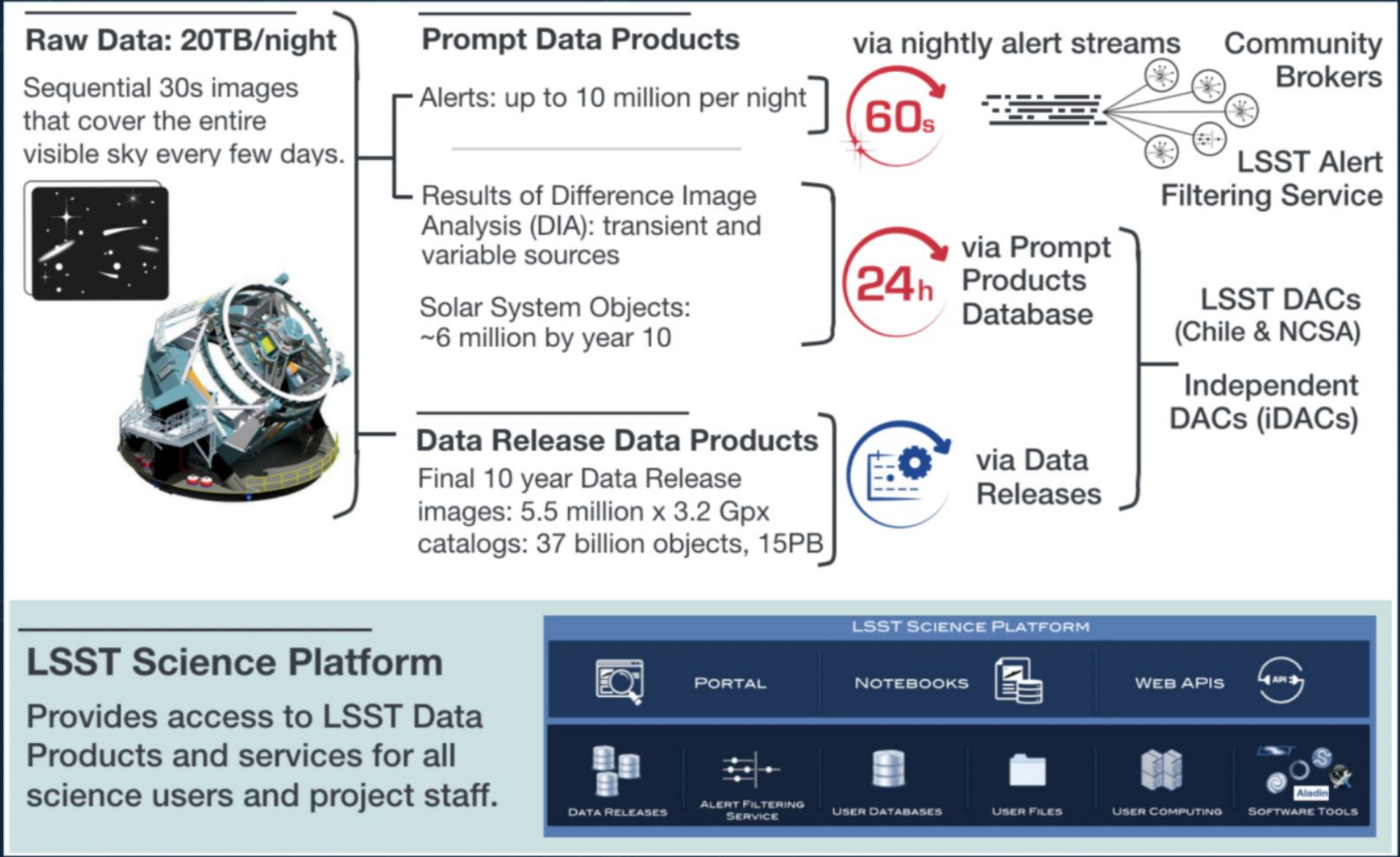
User-produced added-value data pr

ANTARES Broker (Kostya’s talk)

Custom algorithms, deep KBO/NEO searches, variable star classifications…

Enabled by services and computing resources at the LSST Data Access Centers (DACs) and via the LSST Science Platform (web portal, interactive notebook, or API).

Data Management (5m)



Raw Data: 20TB/night

Sequential 30s images that cover the entire visible sky every few days.

Prompt Data Products

Alerts: up to 10 million per night

Results of Difference Image Analysis (DIA): transient and variable sources

Solar System Objects: ~6 million by year 10

Data Release Data Products

Final 10 year Data Release
images: 5.5 million x 3.2 Gpx
catalogs: 37 billion objects, 15PB

LSST Science Platform

Provides access to LSST Data Products and services for all science users and project staff.

Find all of this in the:

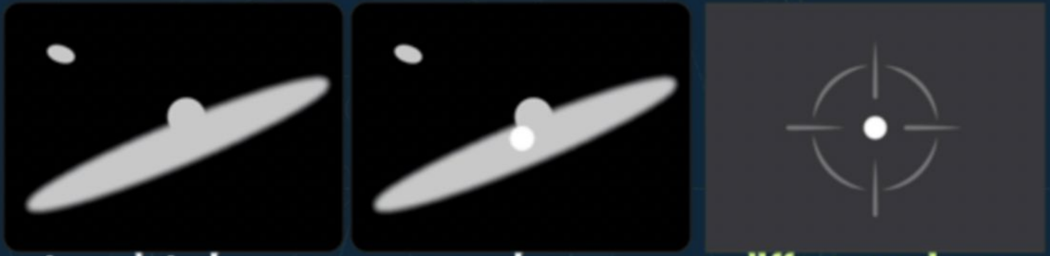
Data Products Definitions Document

"The DPDD"

ls.st/dpdd

Data Management (5m)

Prompt Processing is based on Difference Image Analysis (DIA)



template image

new image

difference image

DIASources: detections in difference image.
DIAObjects: are DIASources linked by coordinate.

Catalogs are stored in the **Prompt Products Database (PPDB)**

60s

Stream of Alerts is released to Alert Brokers and to the LSST Alert Filtering Service.

Alerts: packets of LSST data for a DIASource.
Brokers: receive & process Alerts (*external to LSST*).

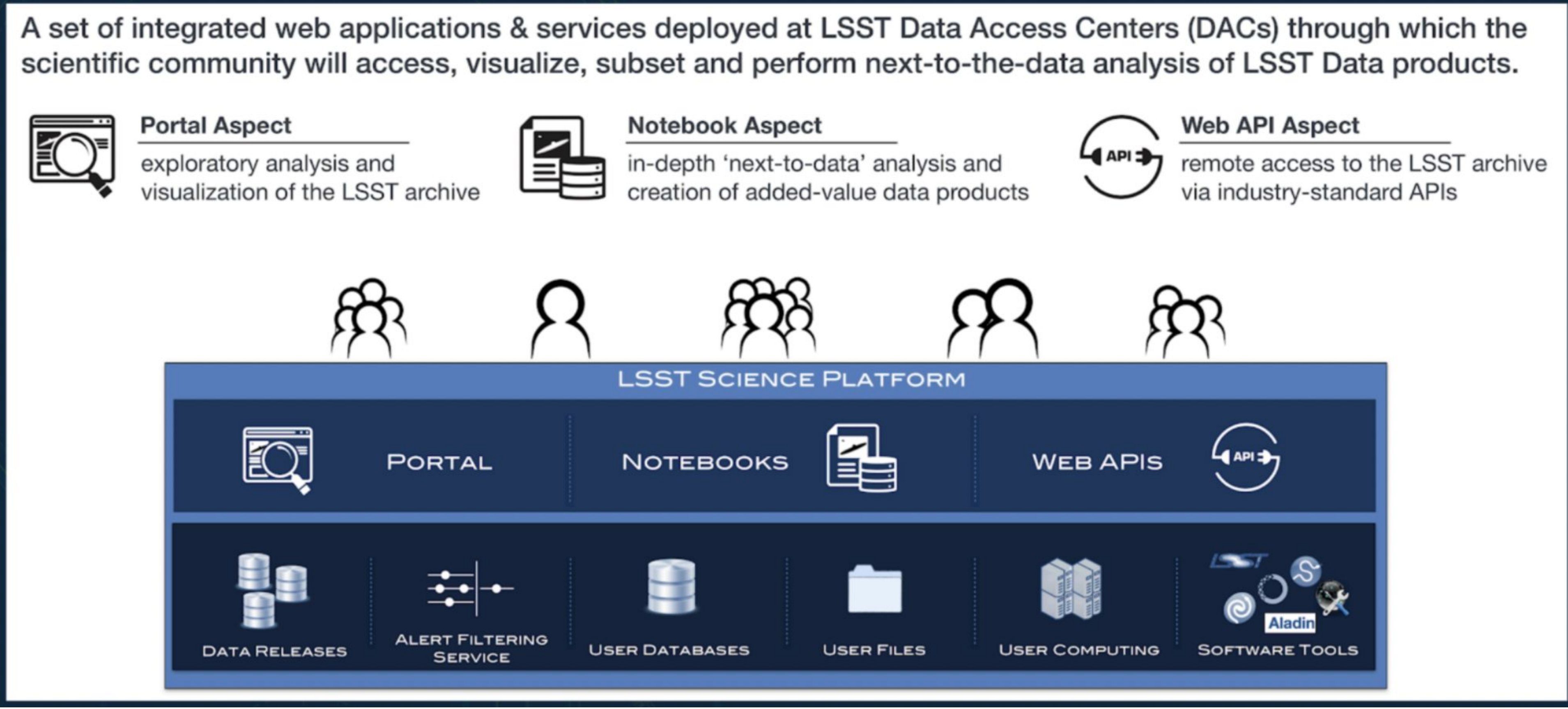
24h

DIASource and DIAObject catalogs, and direct and difference images, available in the LSST Science Platform.

Data Management (5m)

The LSST Science Platform

If you want early access to simulated data - let me know!

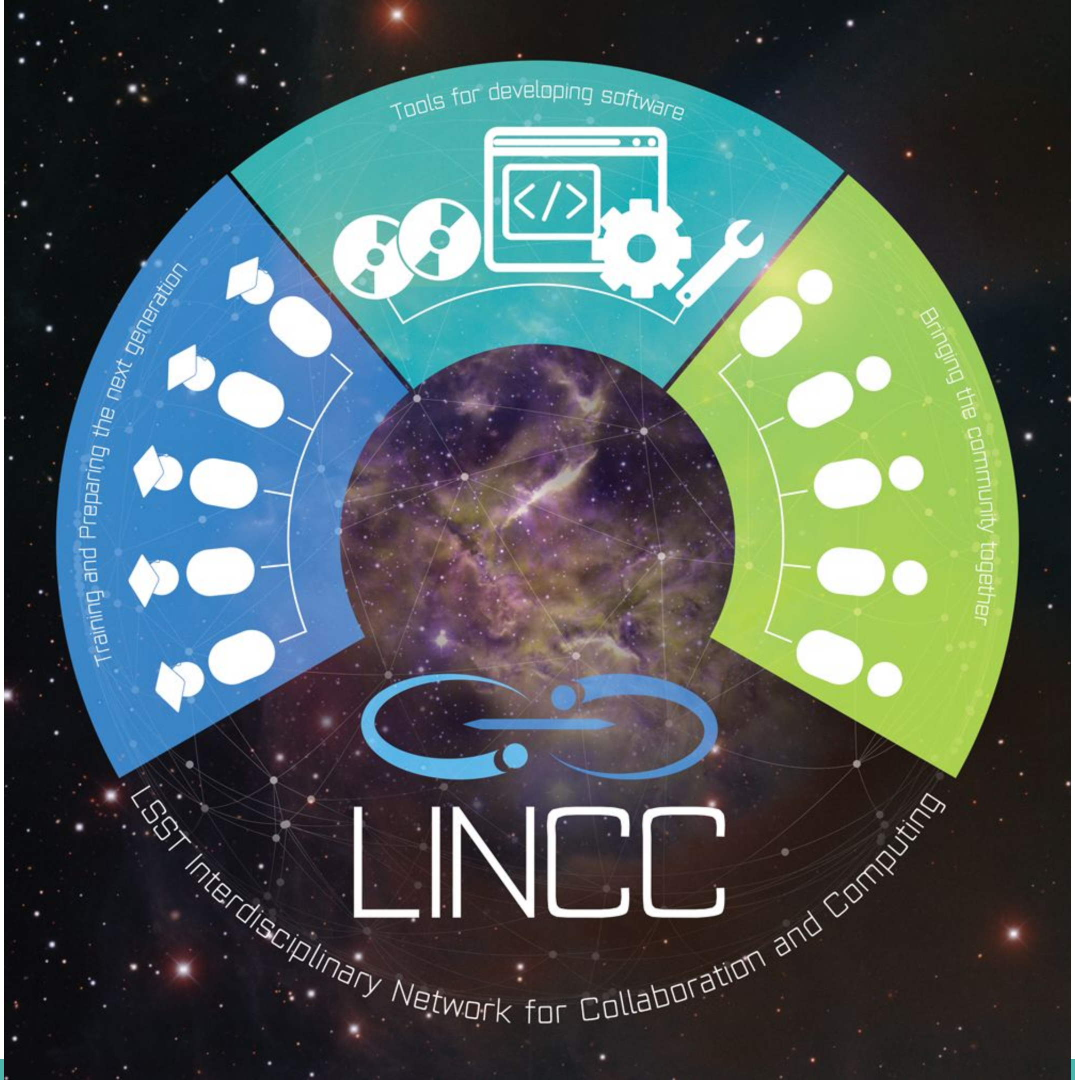


A set of integrated web applications & services deployed at LSST Data Access Centers (DACs) through which the scientific community will access, visualize, subset and perform next-to-the-data analysis of LSST Data products.

Portal Aspect
exploratory analysis and visualization of the LSST archive

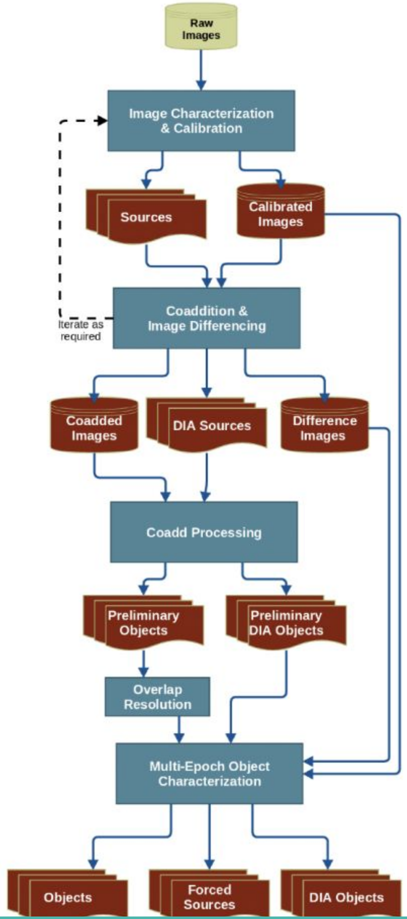
Notebook Aspect
in-depth 'next-to-data' analysis and creation of added-value data products

Web API Aspect
remote access to the LSST archive via industry-standard APIs


Tools for developing software
Training and Preparing the next generation
Bringing the community together
LINCC
LSST Interdisciplinary Network for Collaboration and Computing

https://www.lsstcorporation.org/lincc/

Annual Data Releases enable deep and high-precision science.



Raw Images

Image Characterization & Calibration

Sources

Calibrated Images

Iterate as required

Coaddition & Image Differencing

Coadded Images

DIA Sources

Difference Images

Coadd Processing

Preliminary Objects

Preliminary DIA Objects

Overlap Resolution

Multi-Epoch Object Characterization

Objects

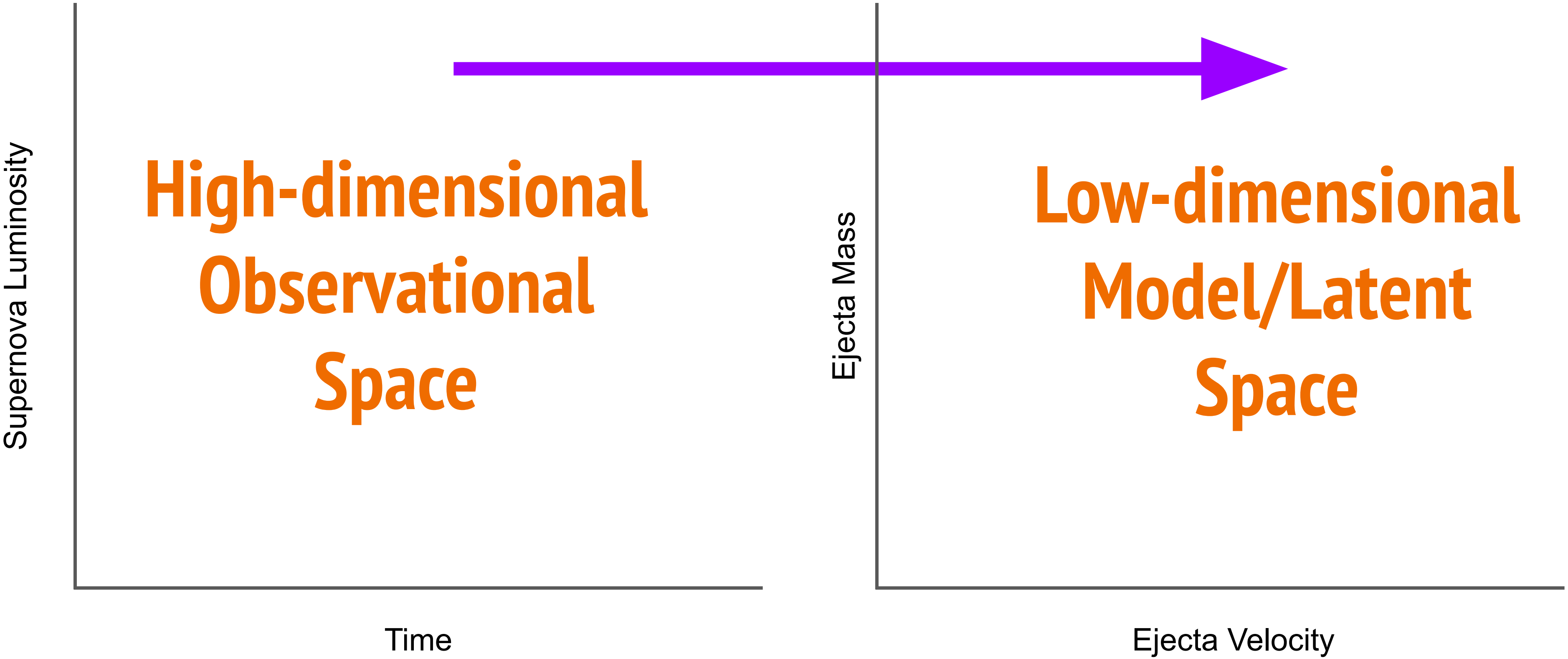
Forced Sources

DIA Objects

- Well calibrated, consistently processed catalogs and images
  - Catalogs of objects, detections, detections in difference images, etc.
  - Combine information from many exposures
- Made available via an annual ***Data Release***
  - Performed yearly (DR2..DR11)
  - …with an additional data release for first 6 months of survey data.(DR1
  - Complete reprocessing of all data to date for each DR with latest pipelines
  - Including fully reprocessed prompt data products
- Catalog Access
  - Relational database and via the LSST Science Platform (LSP)
  - Remote access APIs, VO Protocols (TAP)
- Projected catalog sizes are:
  - 18 billion objects (DR1) -> 37 billion (DR11)
  - 750 billion observations (DR1) -> 30 trillion (DR11)
  - Few PB (DR1) -> 70 PB (DR11)

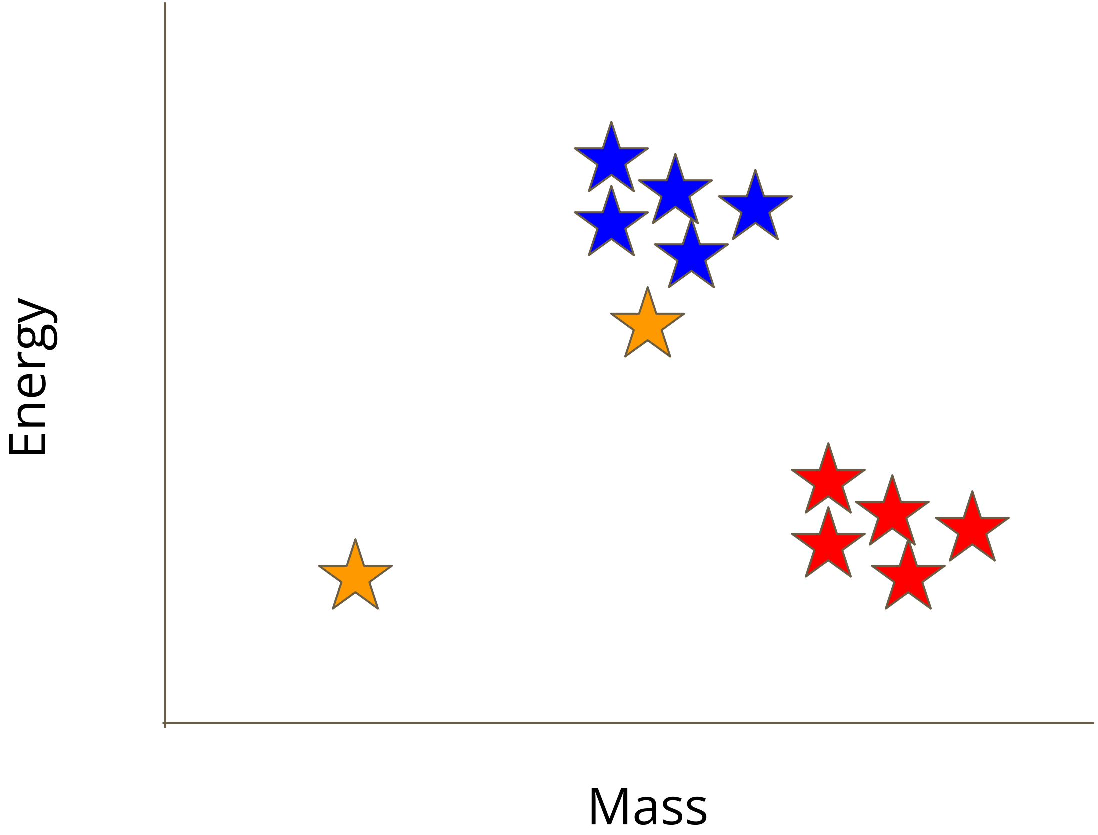
Classify, Identify, Analyze

Want to transform

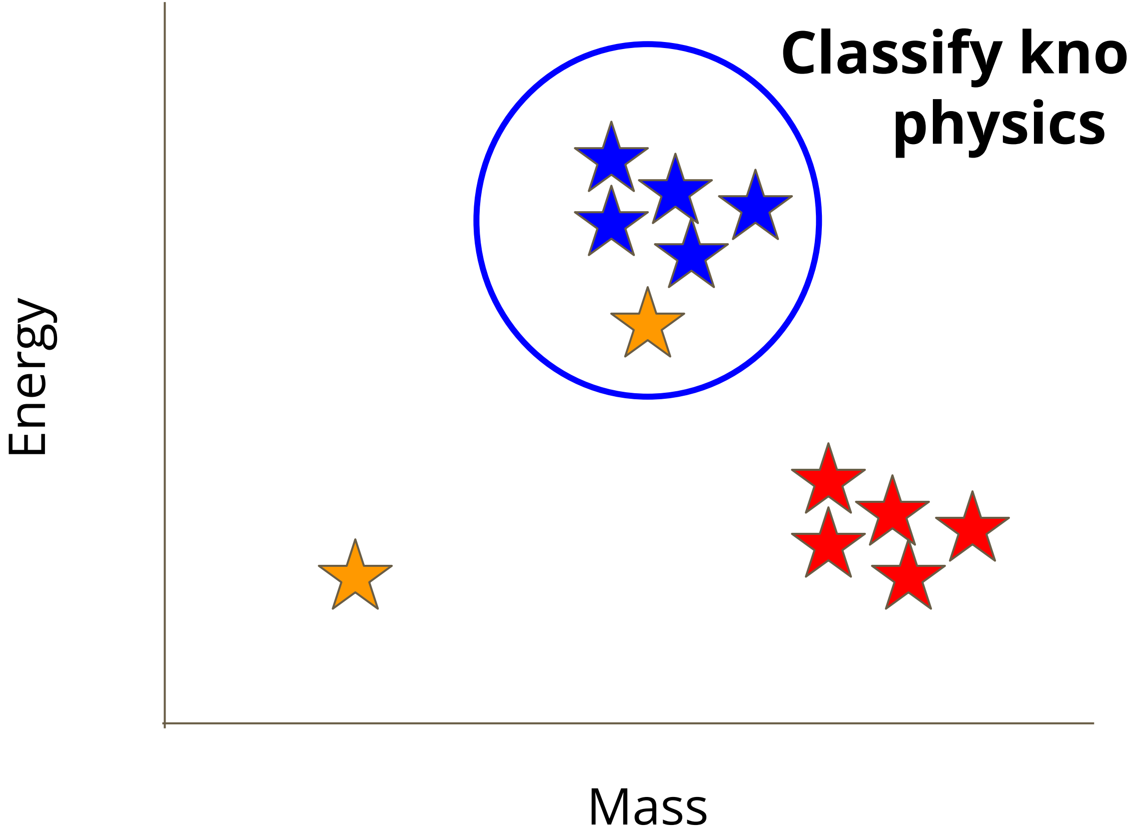


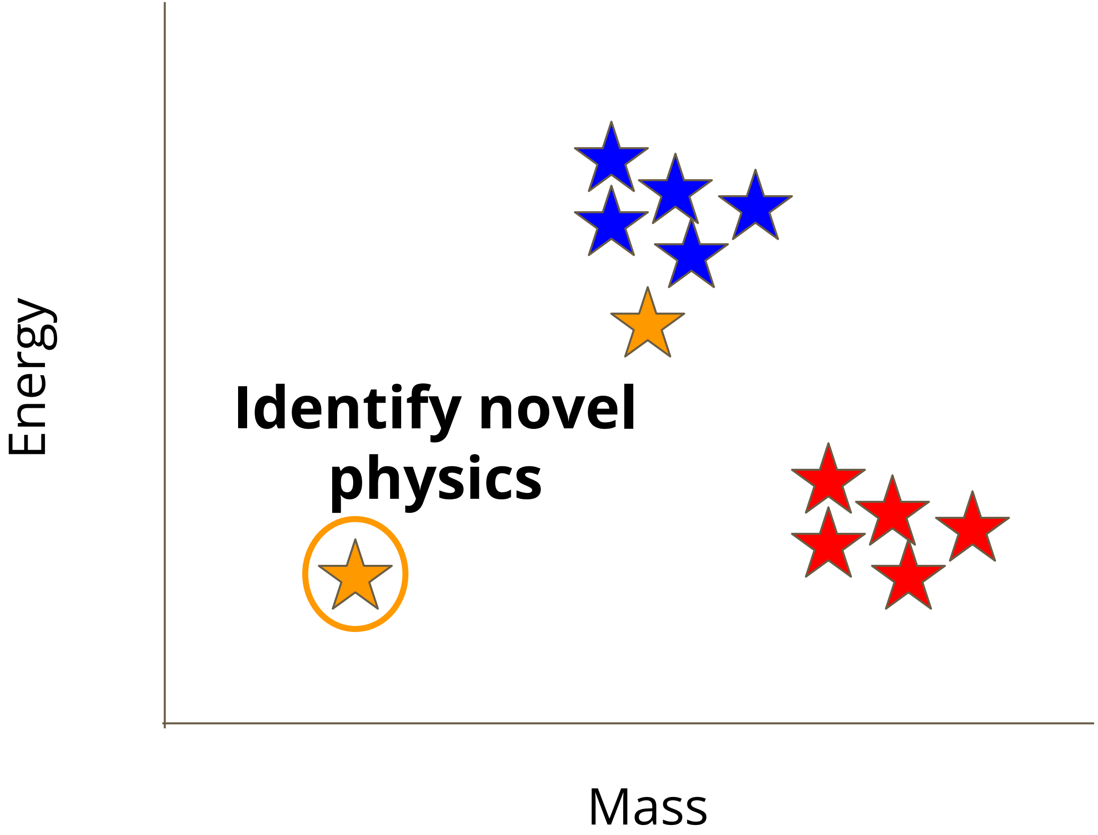
High-dimensional Observational Space

Low-dimensional Model/Latent Space

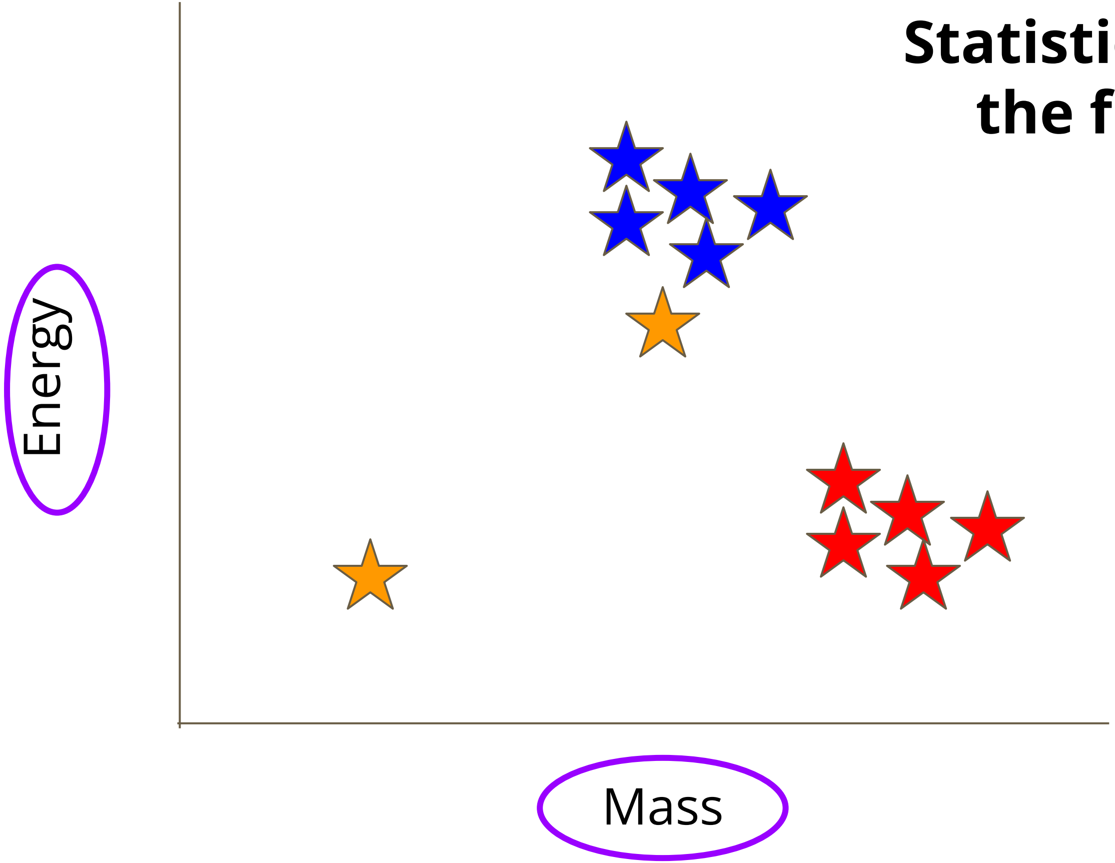

Energy
Mass

Classify known physics



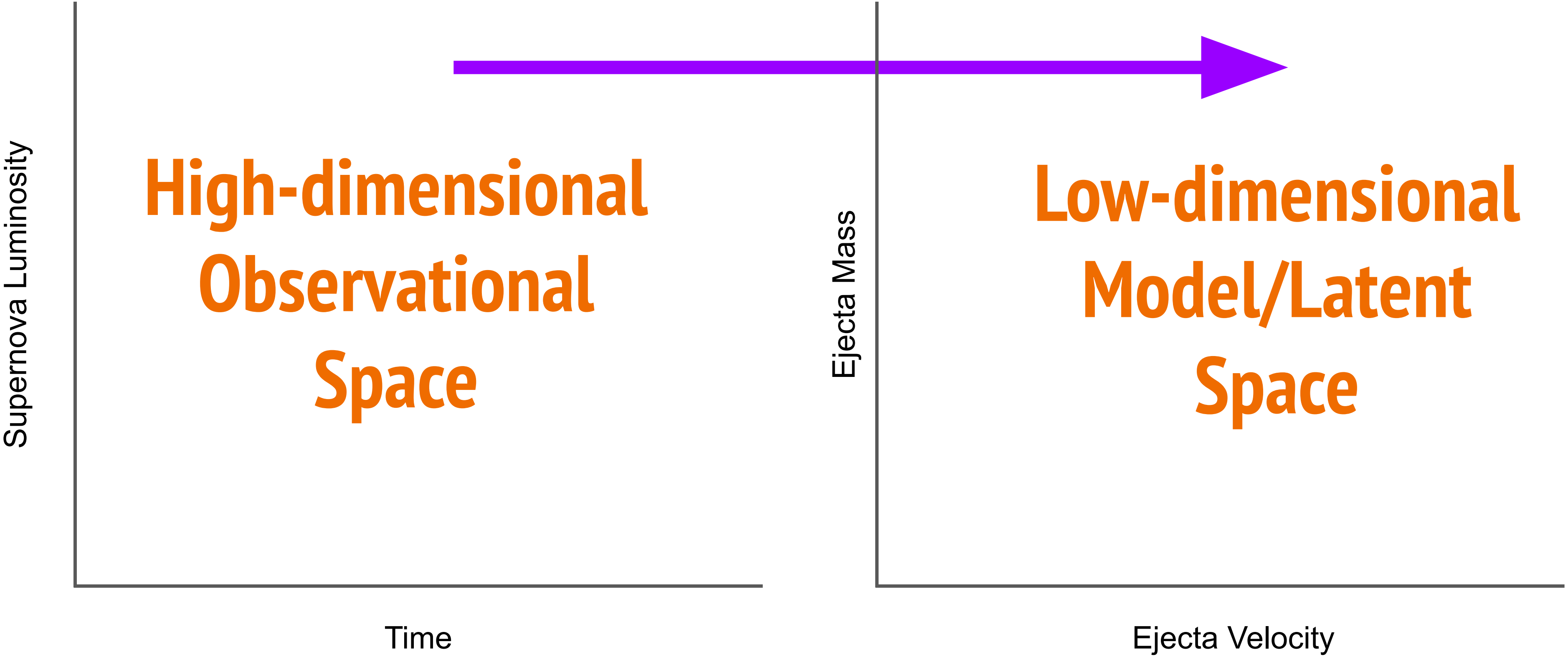


Identify novel physics



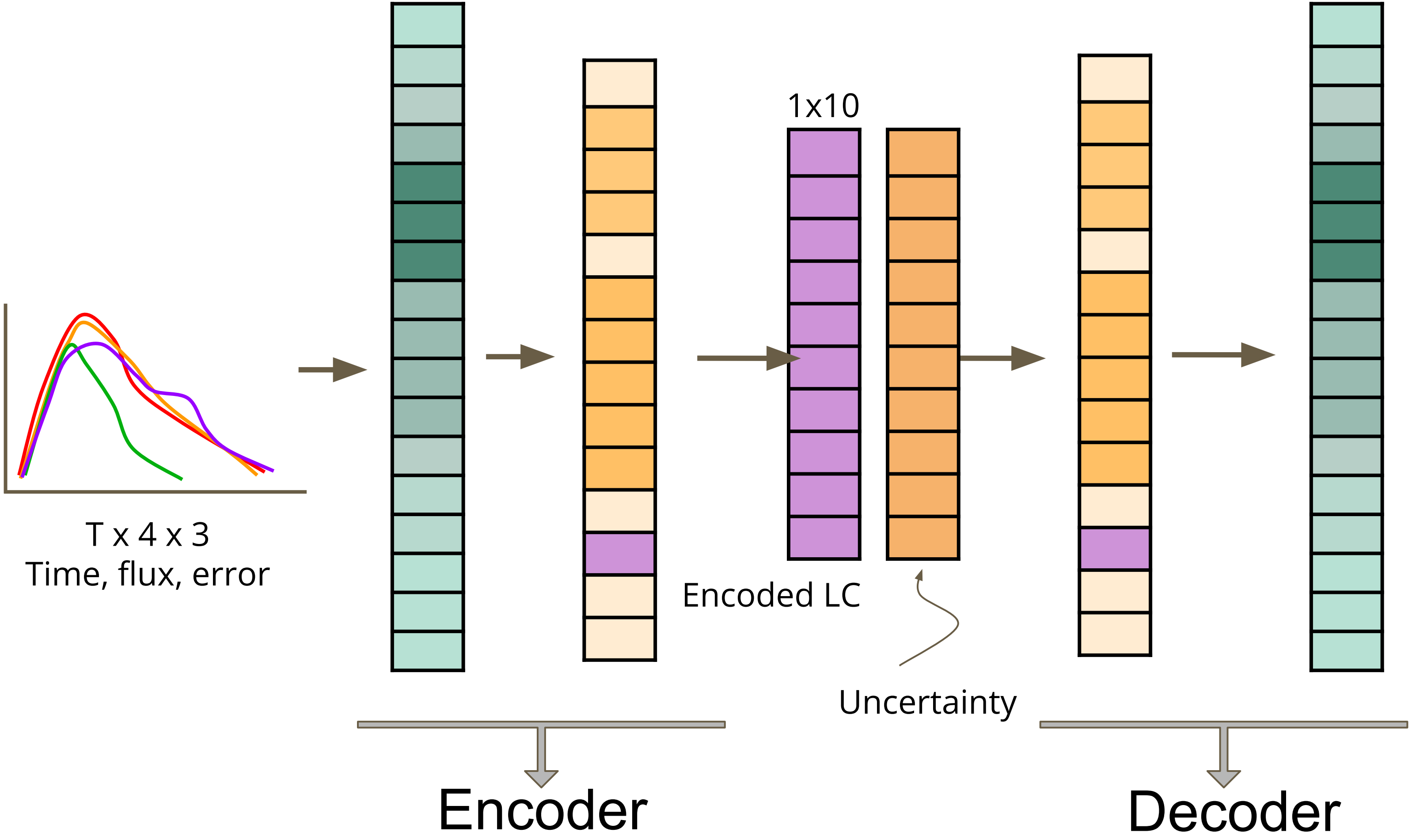
Statistically analyze the full sample

This transformation can be complicated!


Supernova Luminosity
High-dimensional Observational Space
Time
Ejecta Mass
Low-dimensional Model/Latent Space
Ejecta Velocity

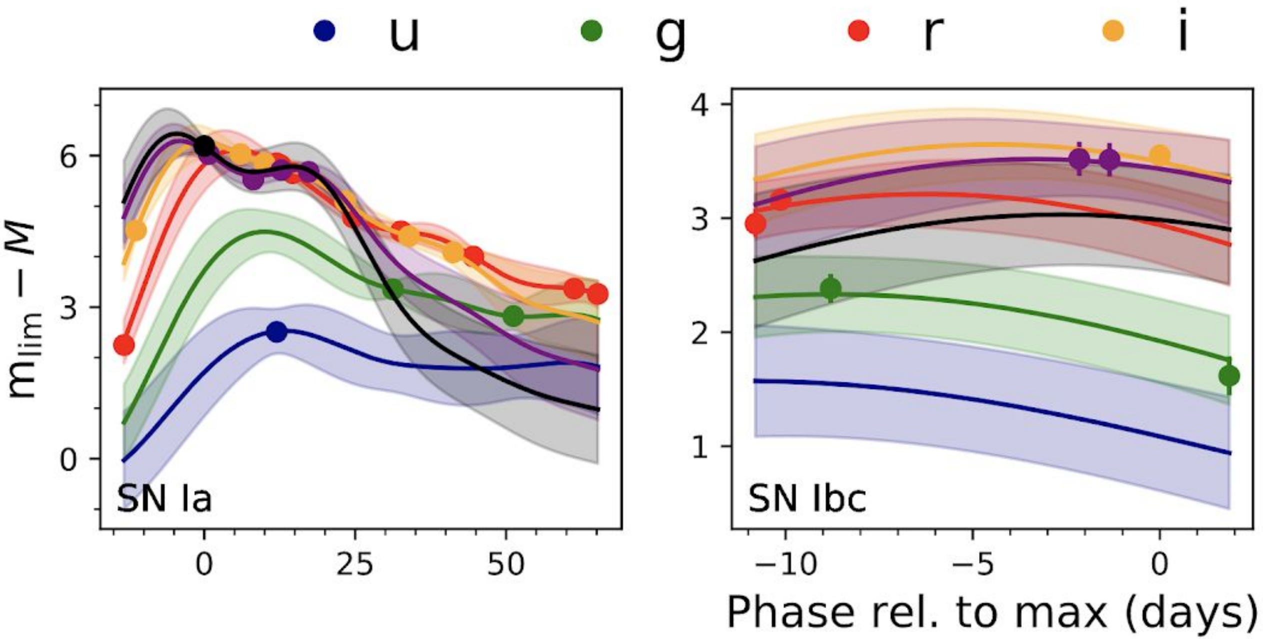
We can make this transformation in a data-driven way....

Use a *variational* autoencoder to *encode* full sample of transient light curve



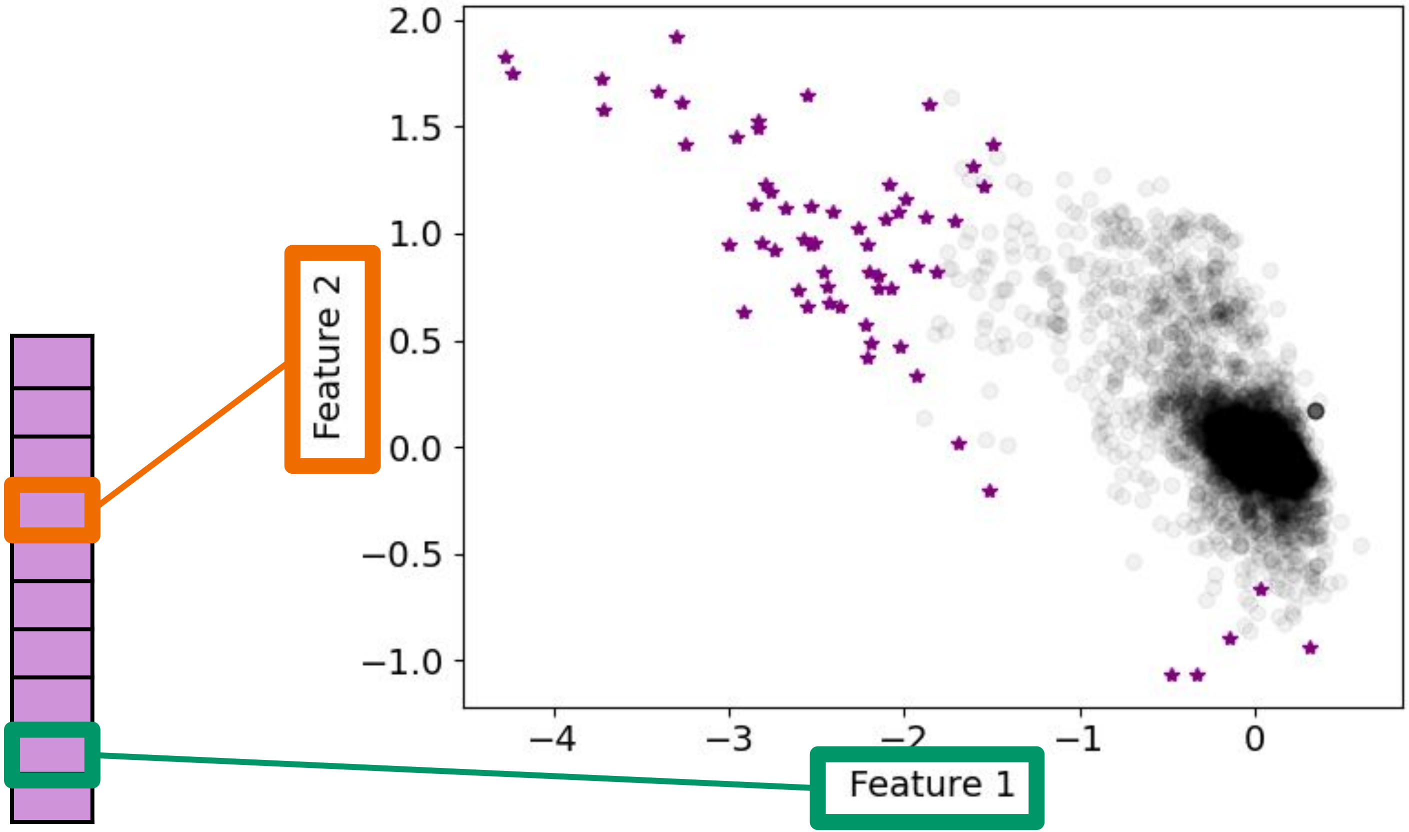
VAV+20,21

Preprocess light curves with 2D Gaussian Processes



*2D -- interpolate in time and filter

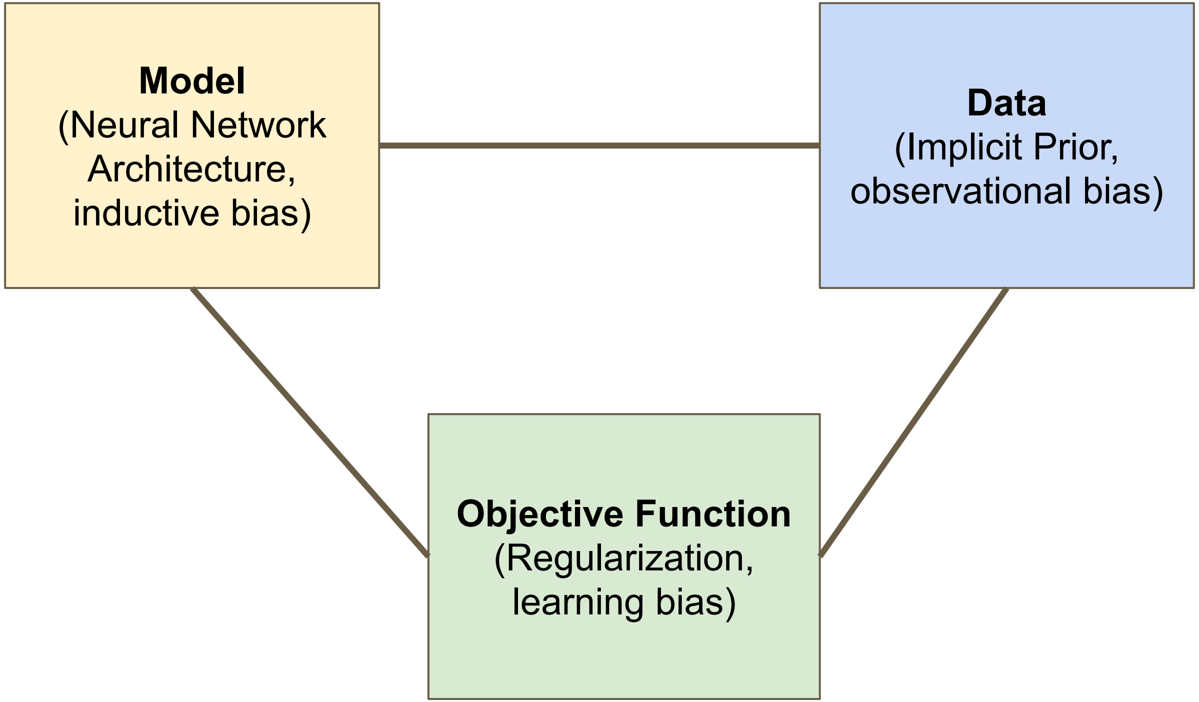
Our learned latent space make it easy to classify or search for anomalies



VAV+ 20,21

We can include some physics in this data-driven method...

Where can I put my physics?



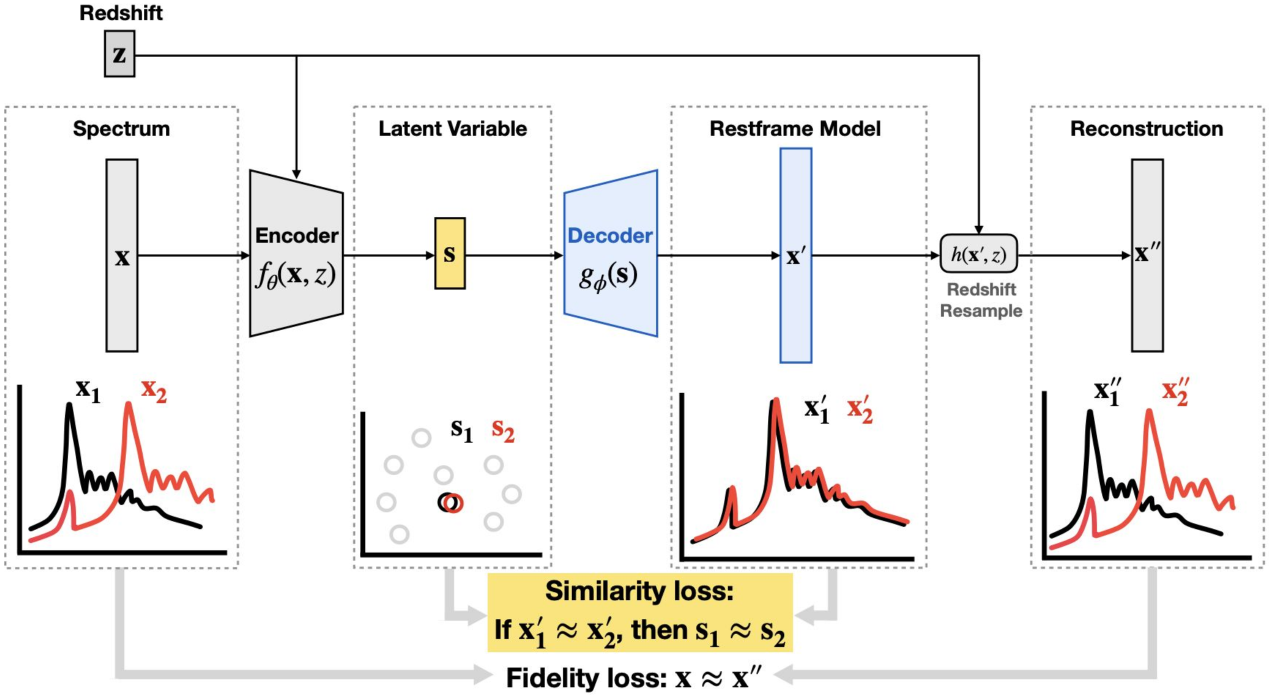
Model
(Neural Network Architecture, inductive bias)

Data
(Implicit Prior, observational bias)

Objective Function
(Regularization, learning bias)

See review:
Karniadakis+ 21

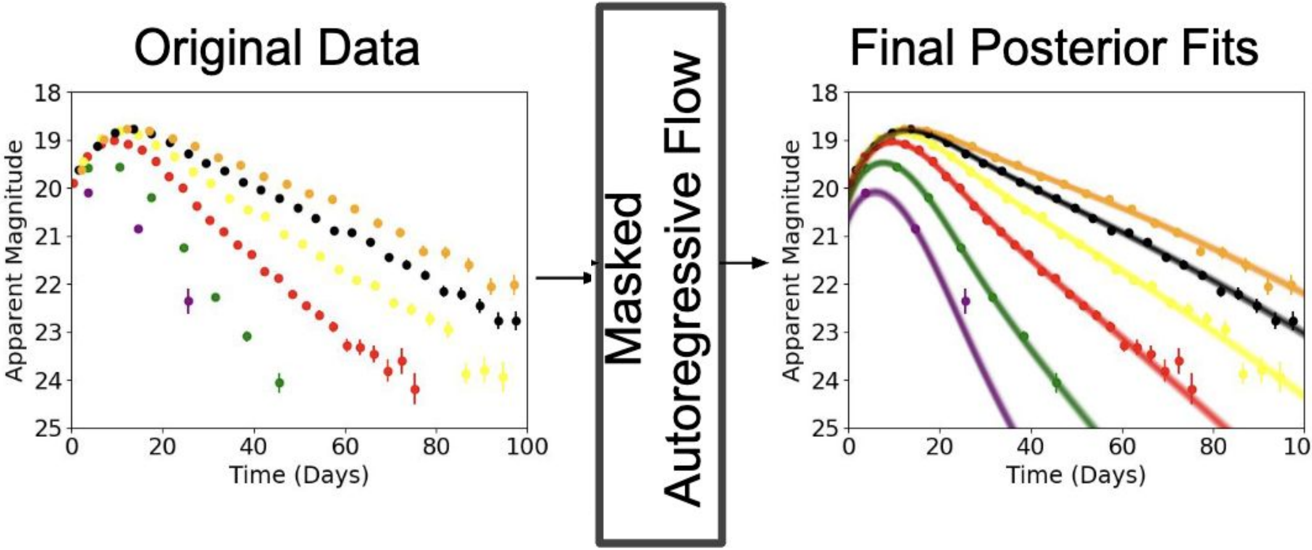
Spotlight: Modifying the objective function – somewhat inconsistent results!



Liang+ 2022
See also Chen, VAV+ 2022

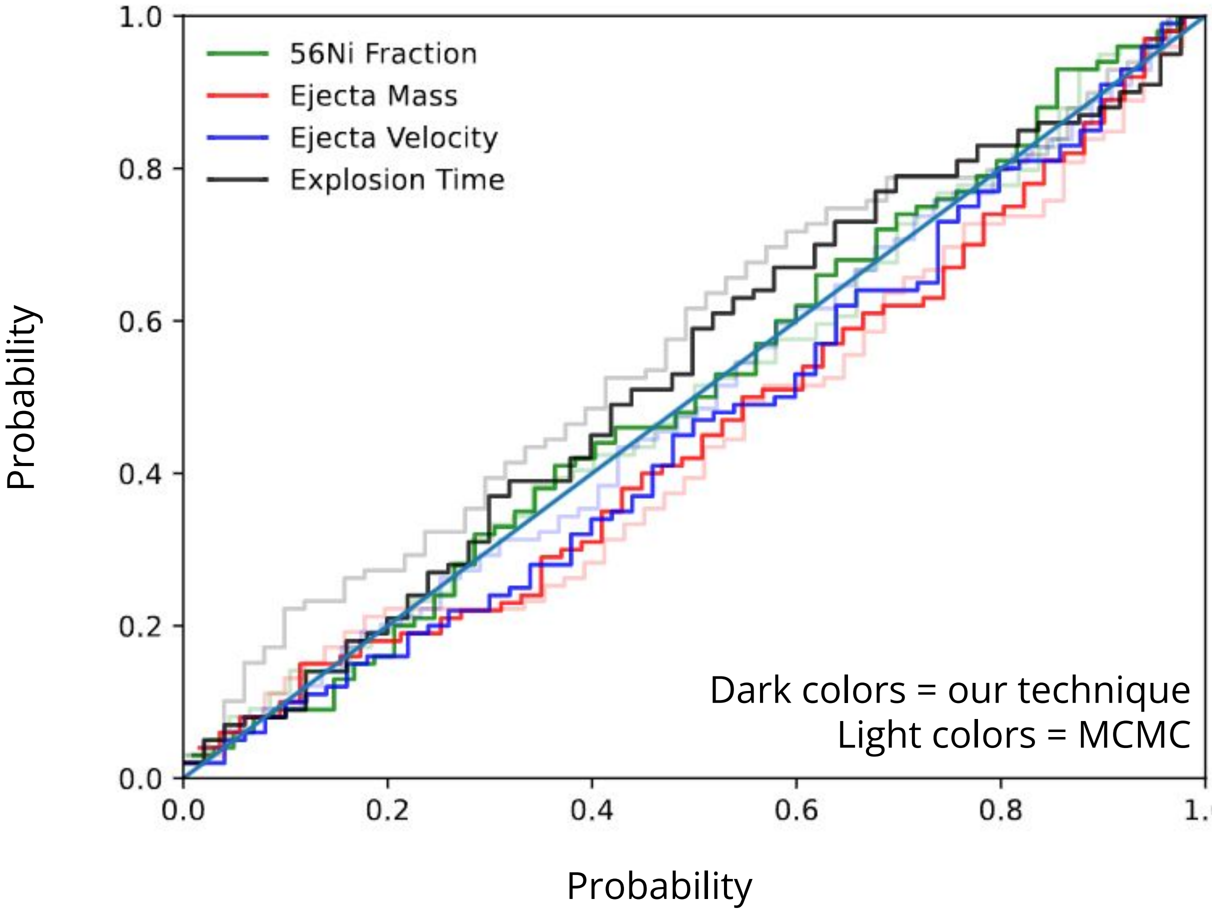
Why not make the latent space entirely physics based?

New method takes 10ms per SN - so about 1 day on a single CPU!



VAV in ML4Physics, Neurips 2022

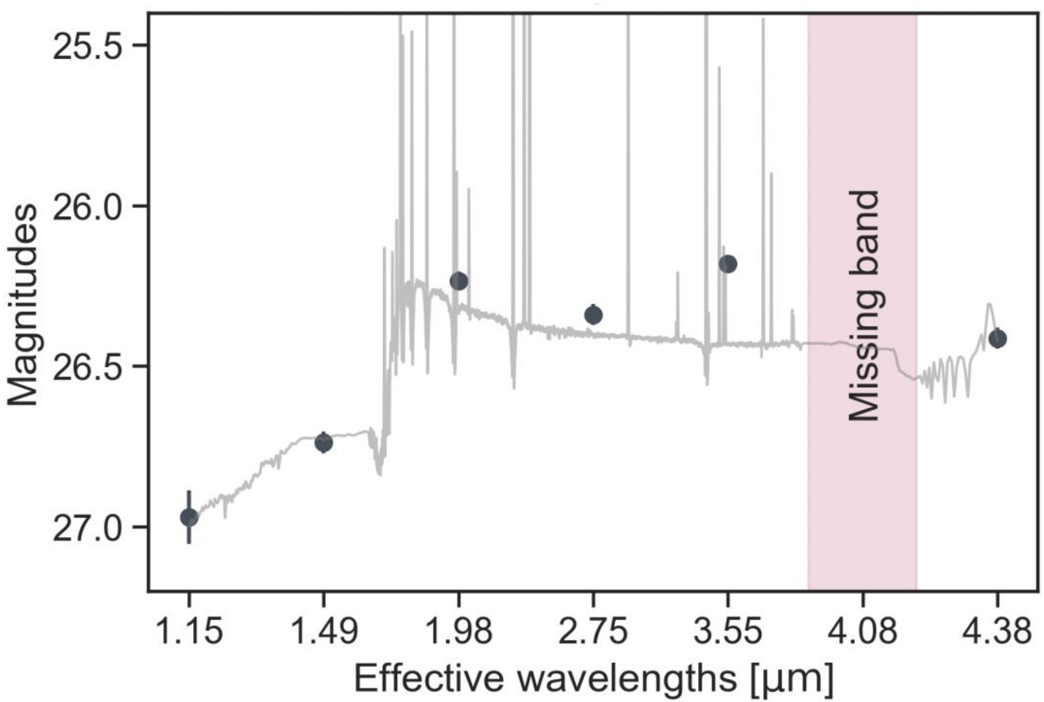
Our probabilities are well-calibrated compared to traditional methods



Dark colors = our technique
Light colors = MCMC

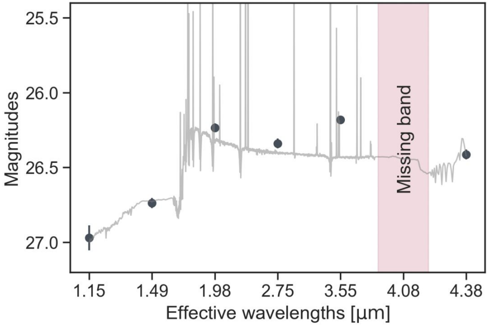
VAV in ML4Physics, Neurips 2022

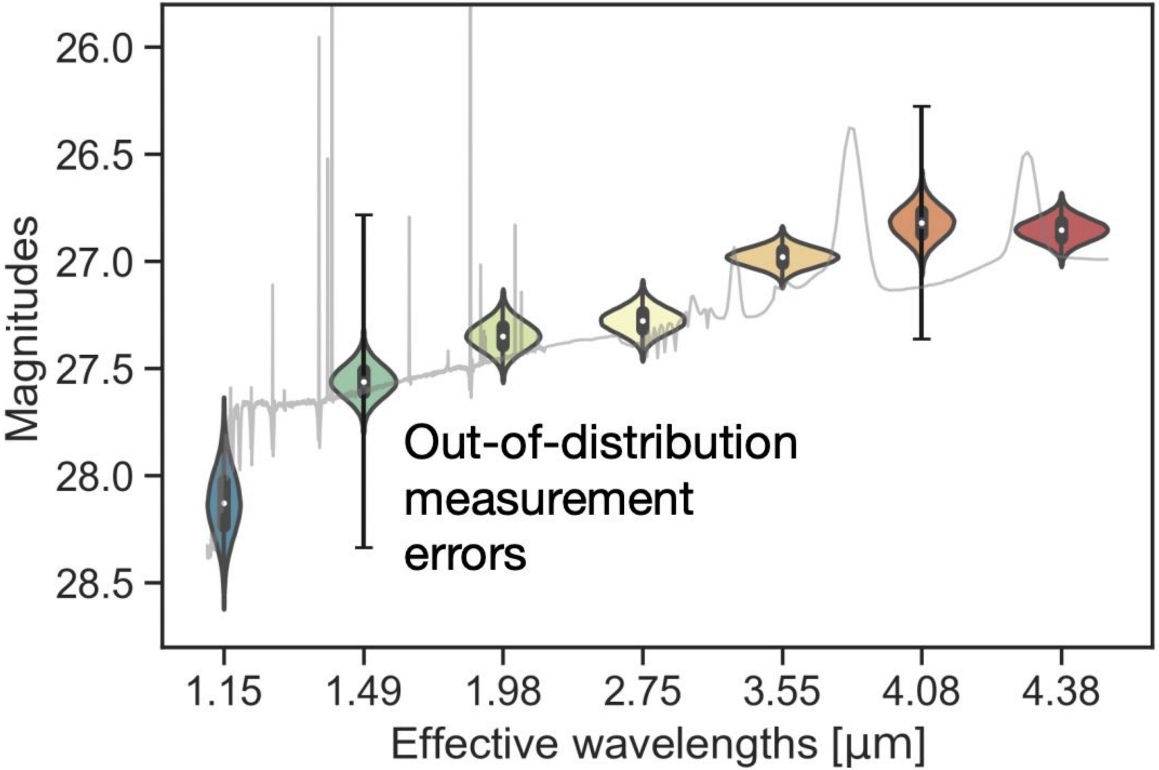
What if your data is not so perfect?



Wang+ in ML4Physics, Neurips 2022

What if your data is not so perfect?

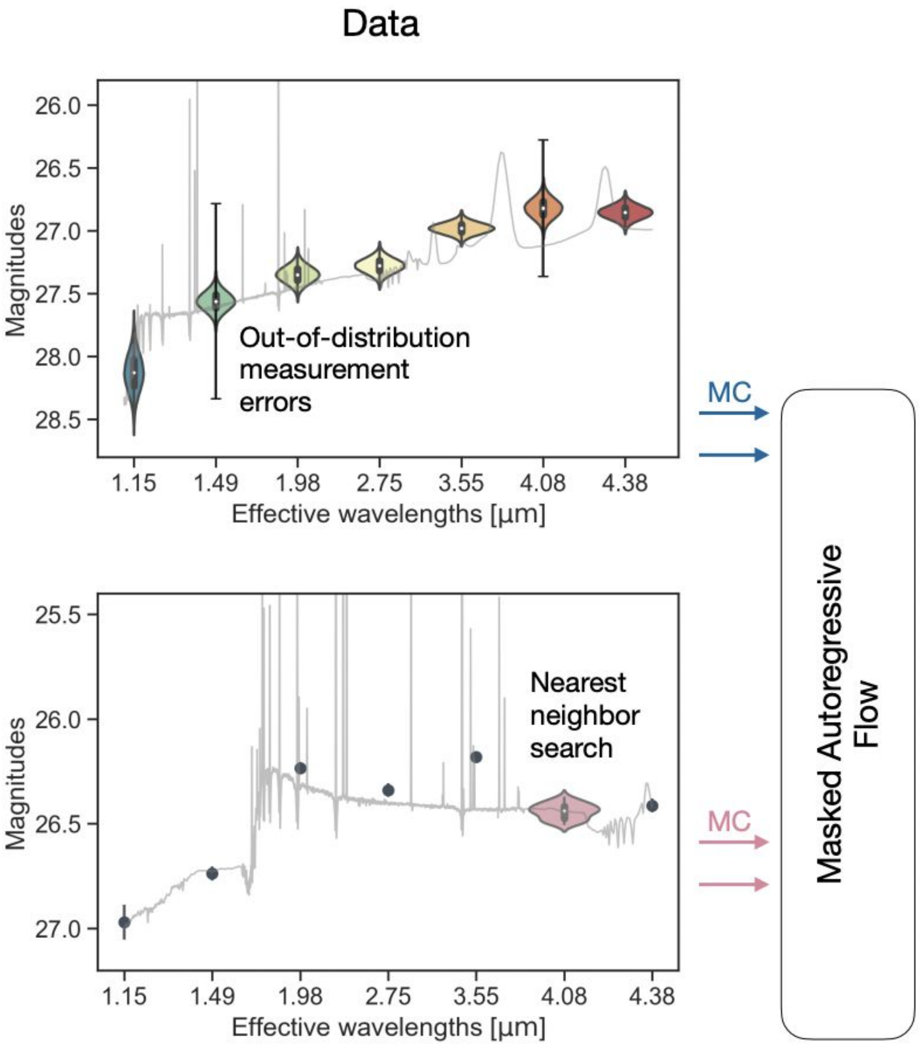




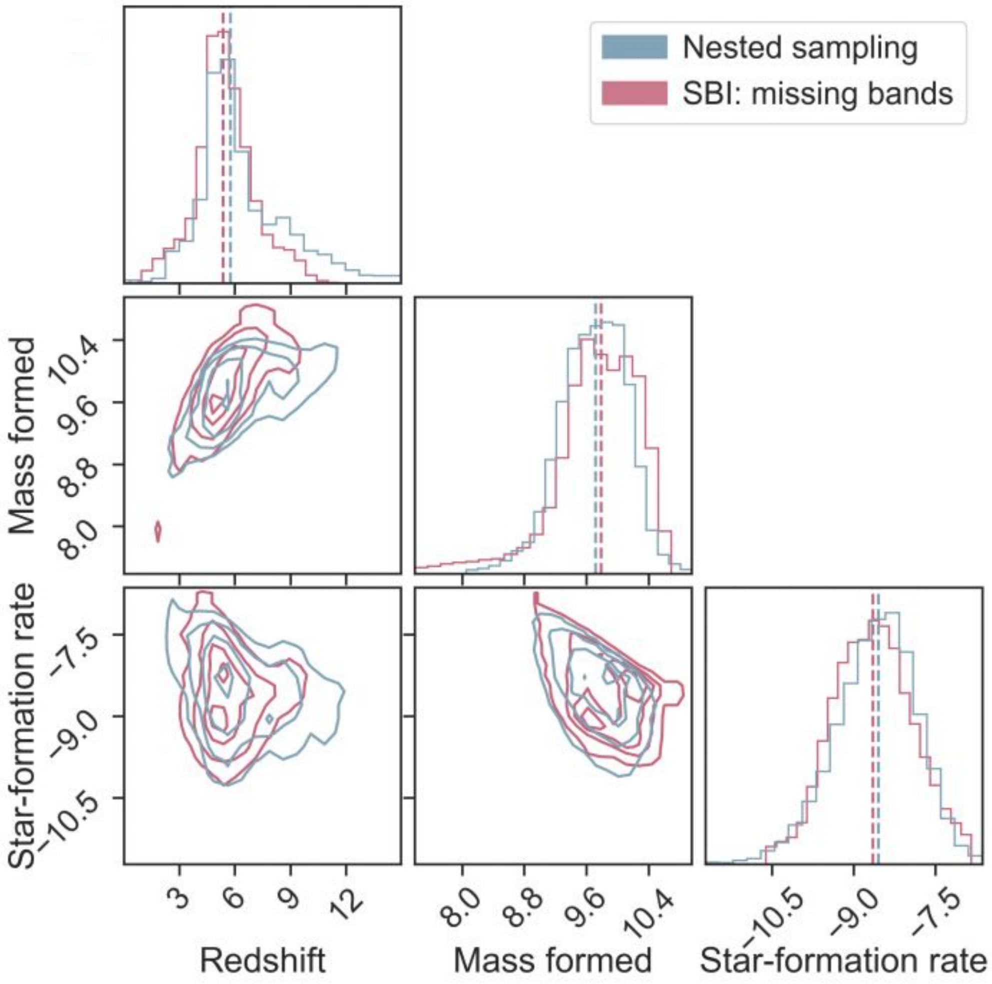
Out-of-distribution measurement errors

Presented new methods to deal with the “reality” of messy data!

Wang+ in ML4Physics, Neurips 2022

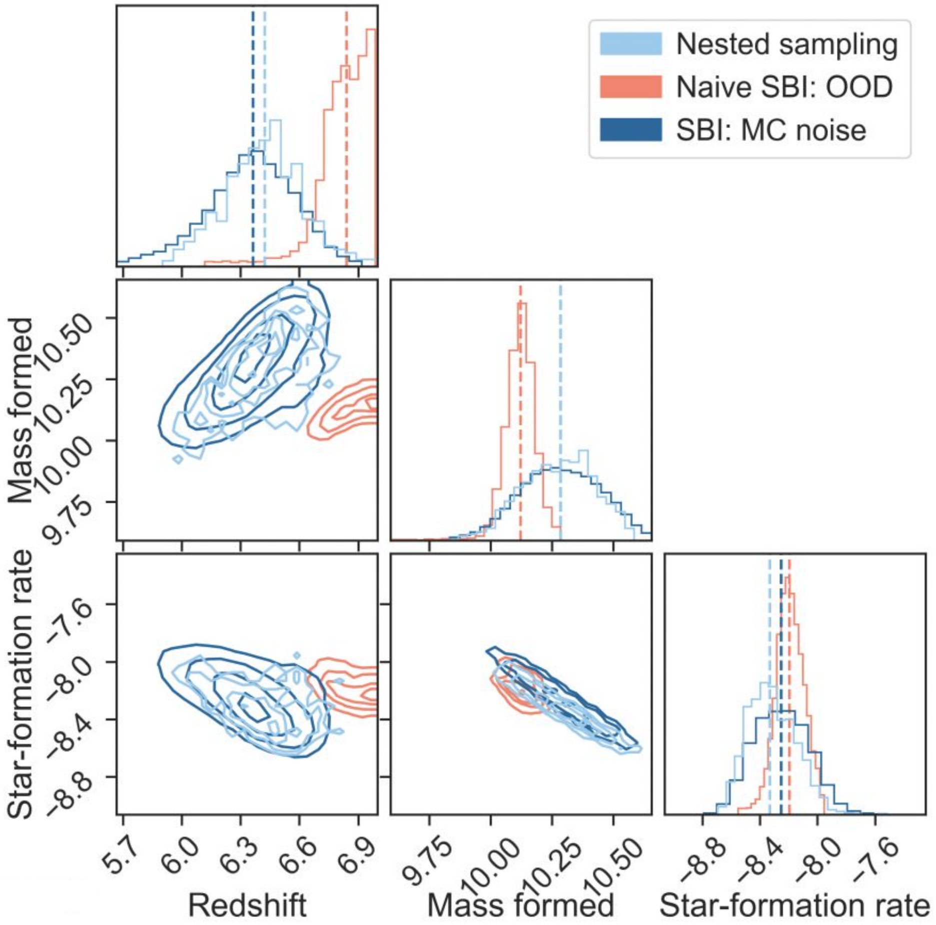


For missing bands: We reproduce results from standard inference techniques in just ~seconds of time!



For out-of-distribution (“weird”) noise: We reproduce results from standard inference techniques in ~10s of seconds.

Results



Wang+ in ML4Physics, Neurips 2022

Concluding remarks

- Simulation-based inference (SBI) is a new technique to rapidly approximate traditional statistical methods
- SBI can lead to factors of >1000x potential savings in computational time
- We have presented two applications (VAV22 and Wang+22) with solutions for realistic datasets

Really excited to chat about other applications!!