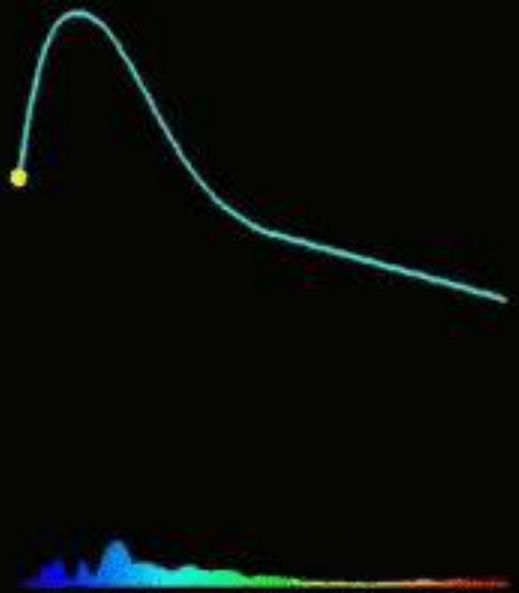
Time-domain Astrophysics in the Era of Big Data

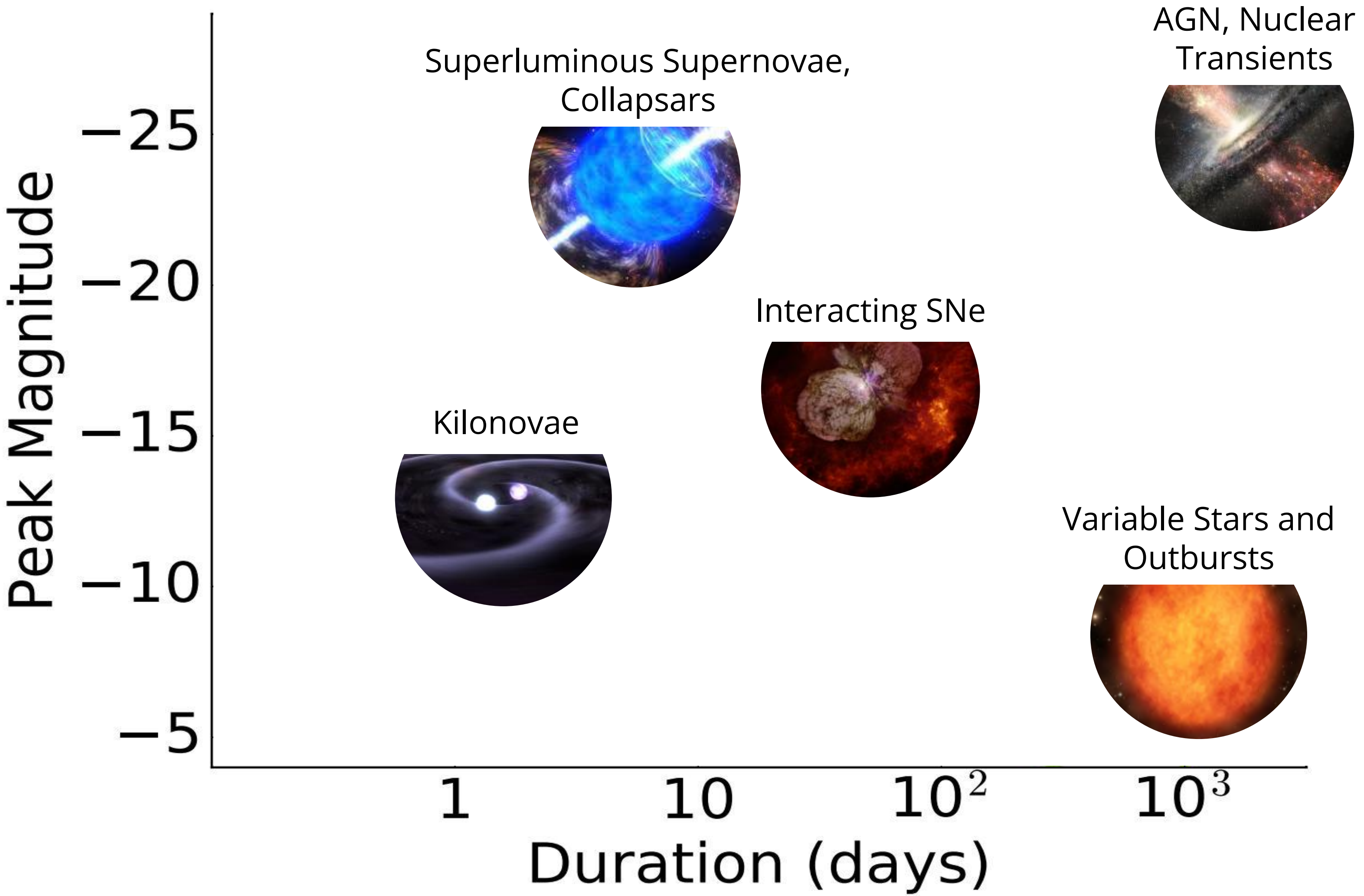
V. Ashley Villar
Harvard University, Assistant Professor

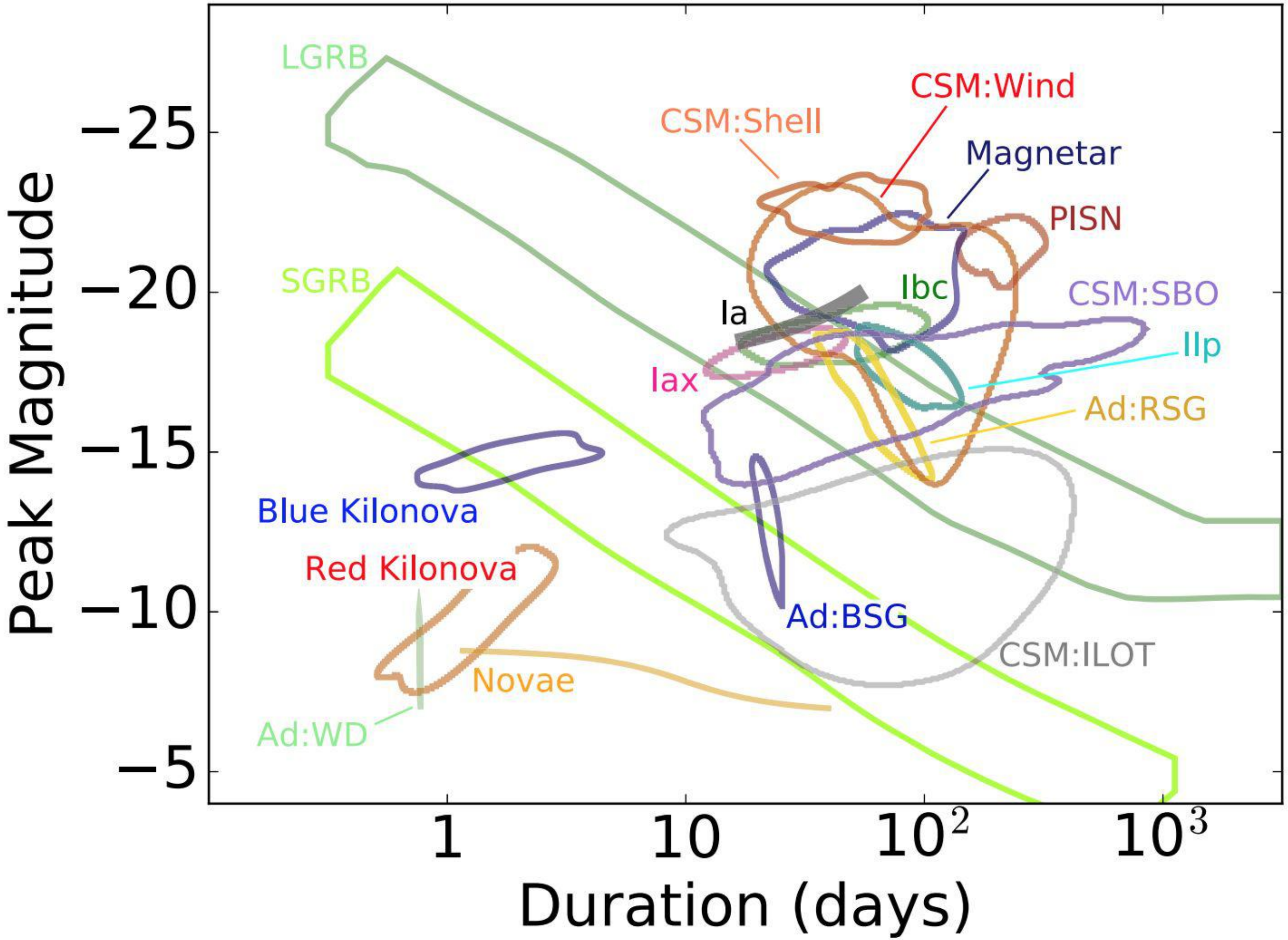
Today will be a talk on data-driven methodology,
time-domain astrophysics and the marriage of the two


VLT



Youtube: Magnetosheath

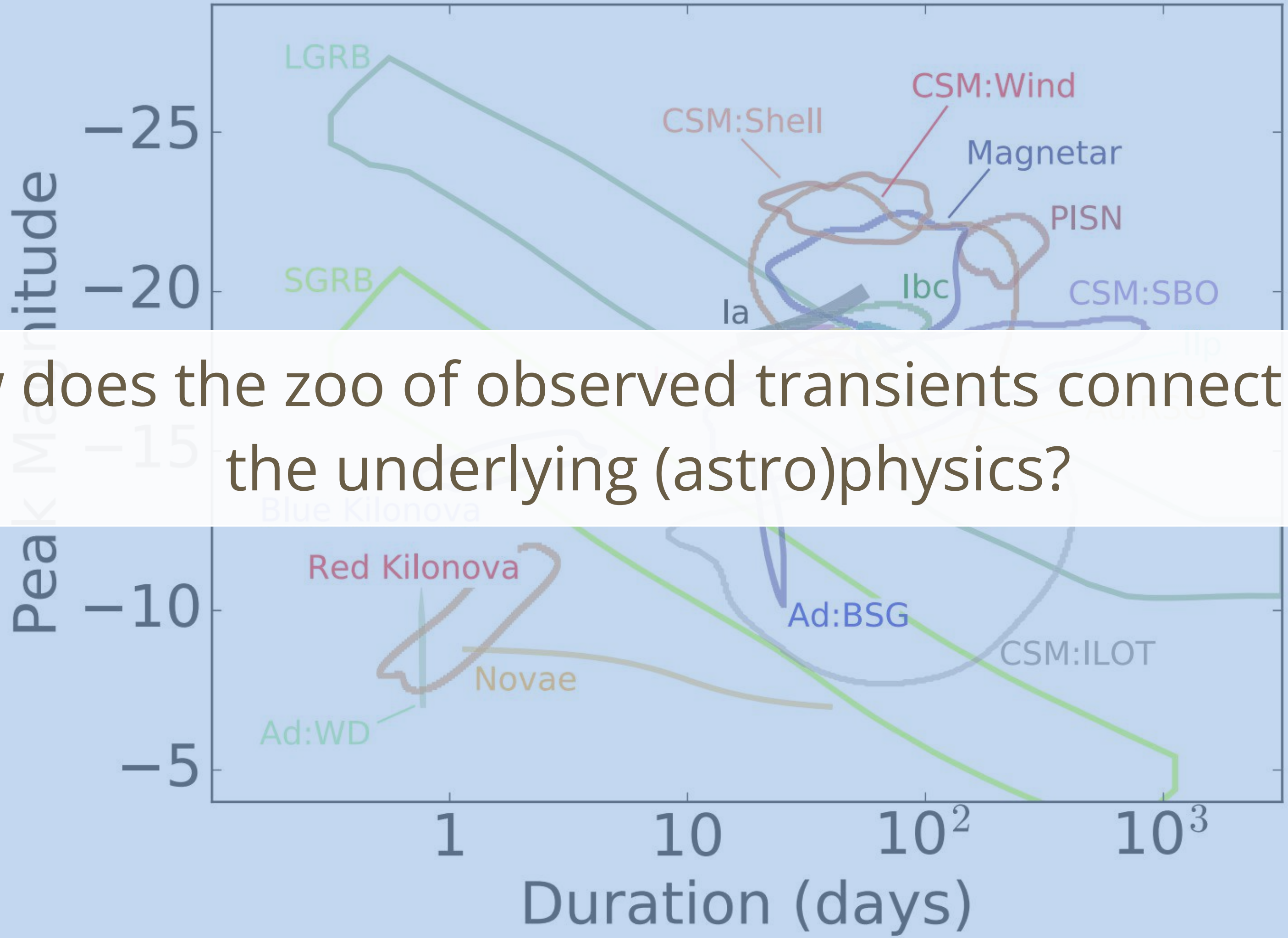

AGN, Nuclear Transients
Superluminous Supernovae, Collapsars
Interacting SNe
Kilonovae
Variable Stars and Outbursts
Peak Magnitude
−25
−20
−15
−10
−5
1
10
10^2
10^3
Duration (days)



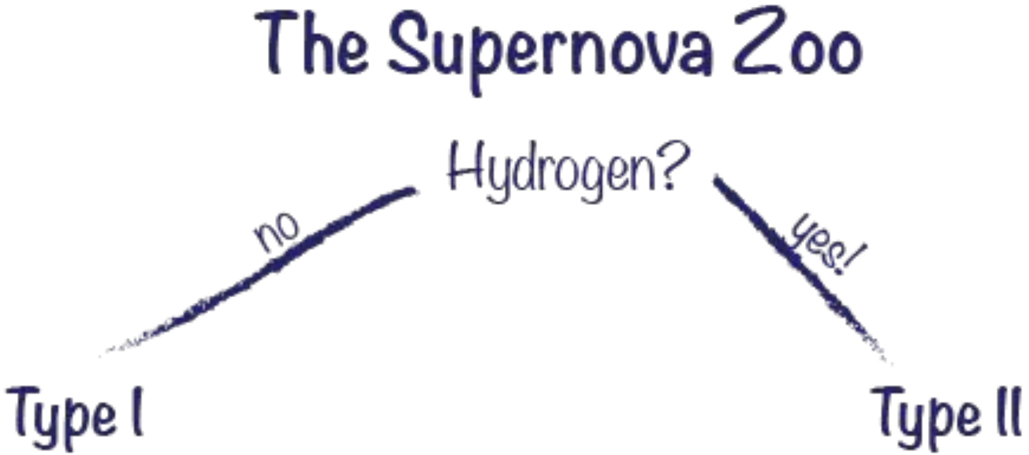
LGRB
SGRB
CSM:Wind
CSM:Shell
Magnetar
PISN
Ibc
Ia
CSM:SBO
IIp
Iax
Ad:RSG
Blue Kilonova
Red Kilonova
Ad:BSG
CSM:ILOT
Novae
Ad:WD
Peak Magnitude
−25
−20
−15
−10
−5
1
10
10^2
10^3
Duration (days)

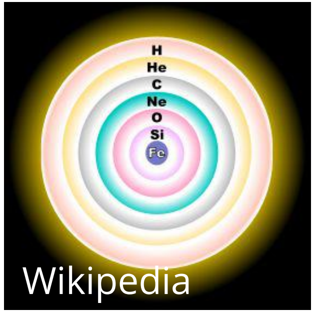
VAV+17a

How does the zoo of observed transients connect with the underlying (astro)physics?



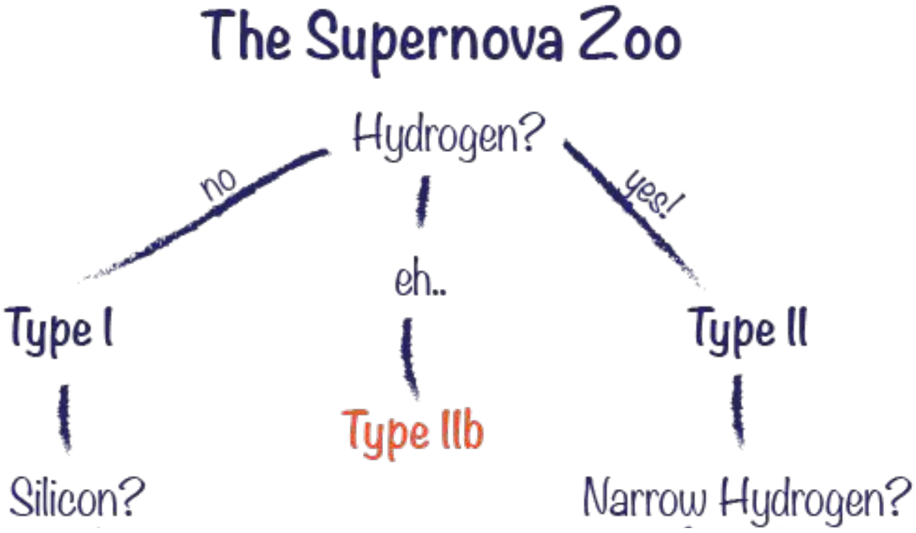
Transients are traditionally classified with spectra





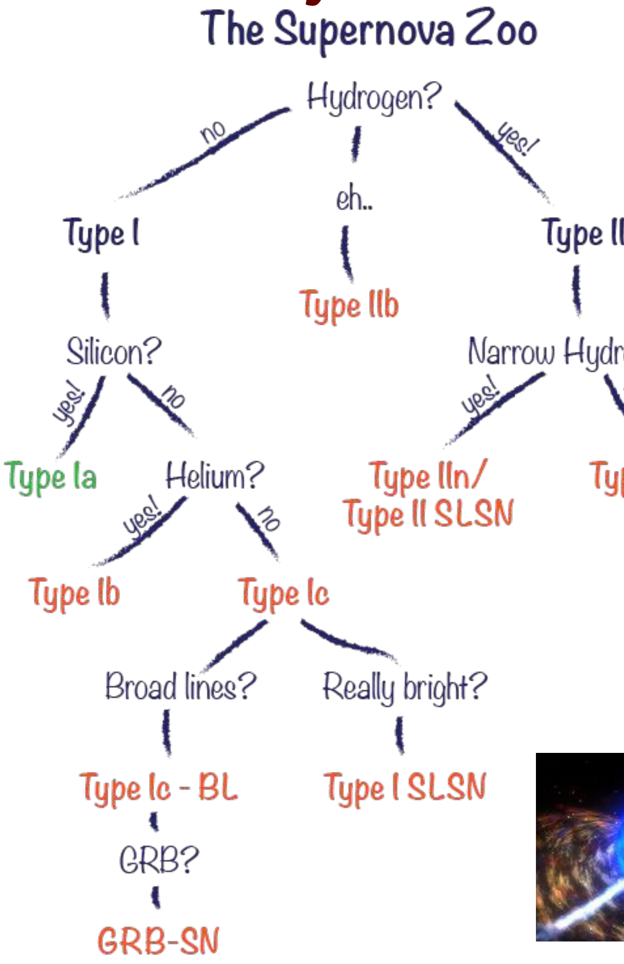
Credit: VAV for Astrobites

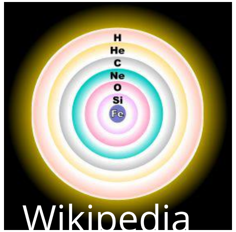
Transients are traditionally classified with spectra



Credit: VAV for Astrobites

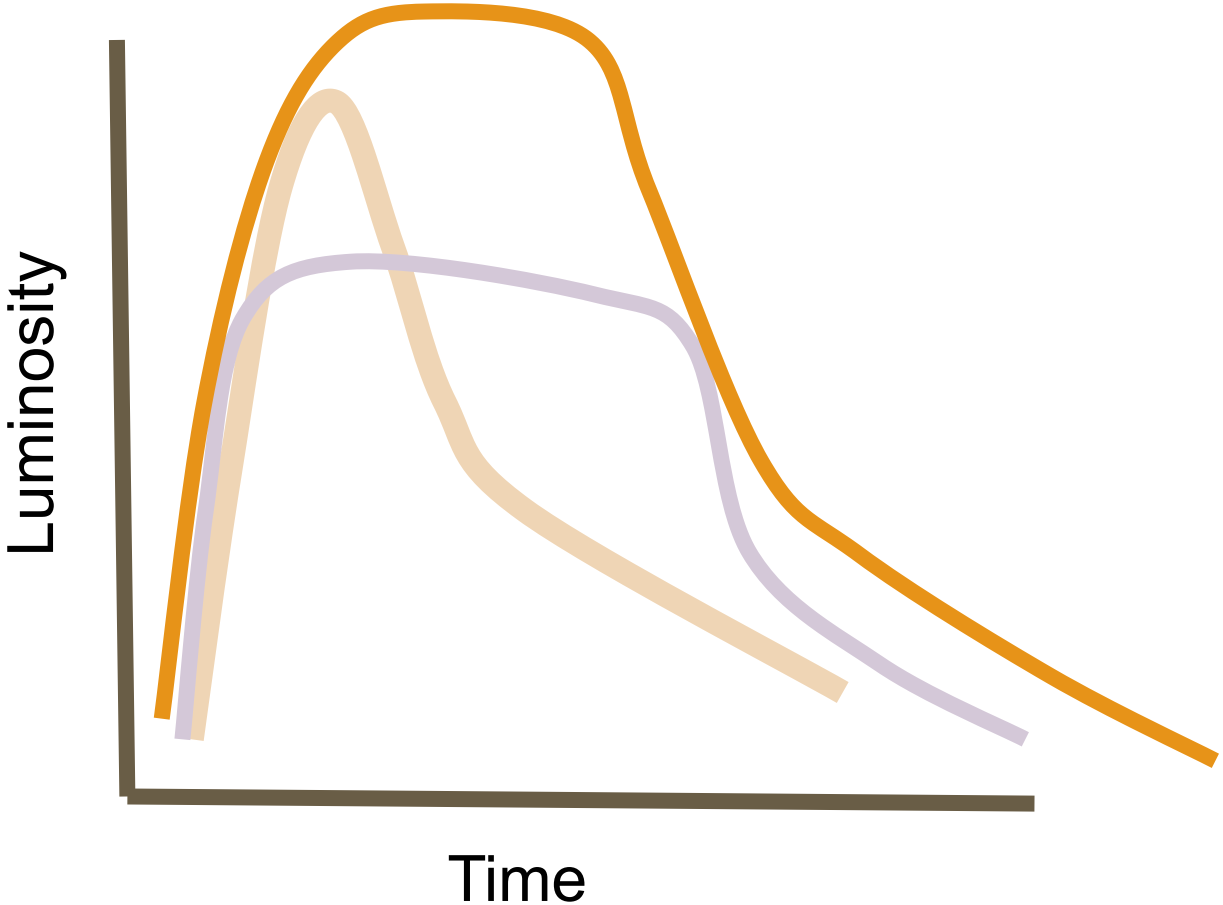
Transients are traditionally classified with spectra





Credit: VAV for Astrobites

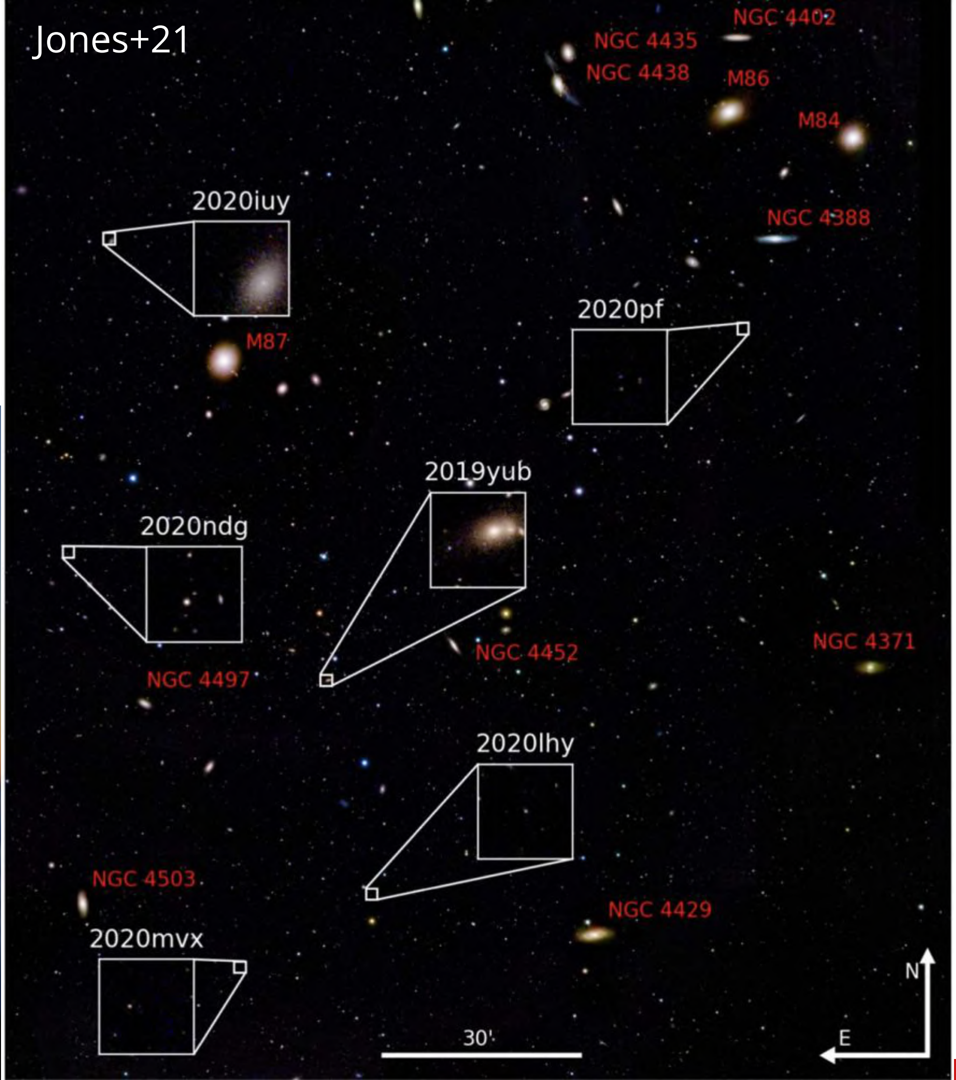
The shapes of light curves encode physics



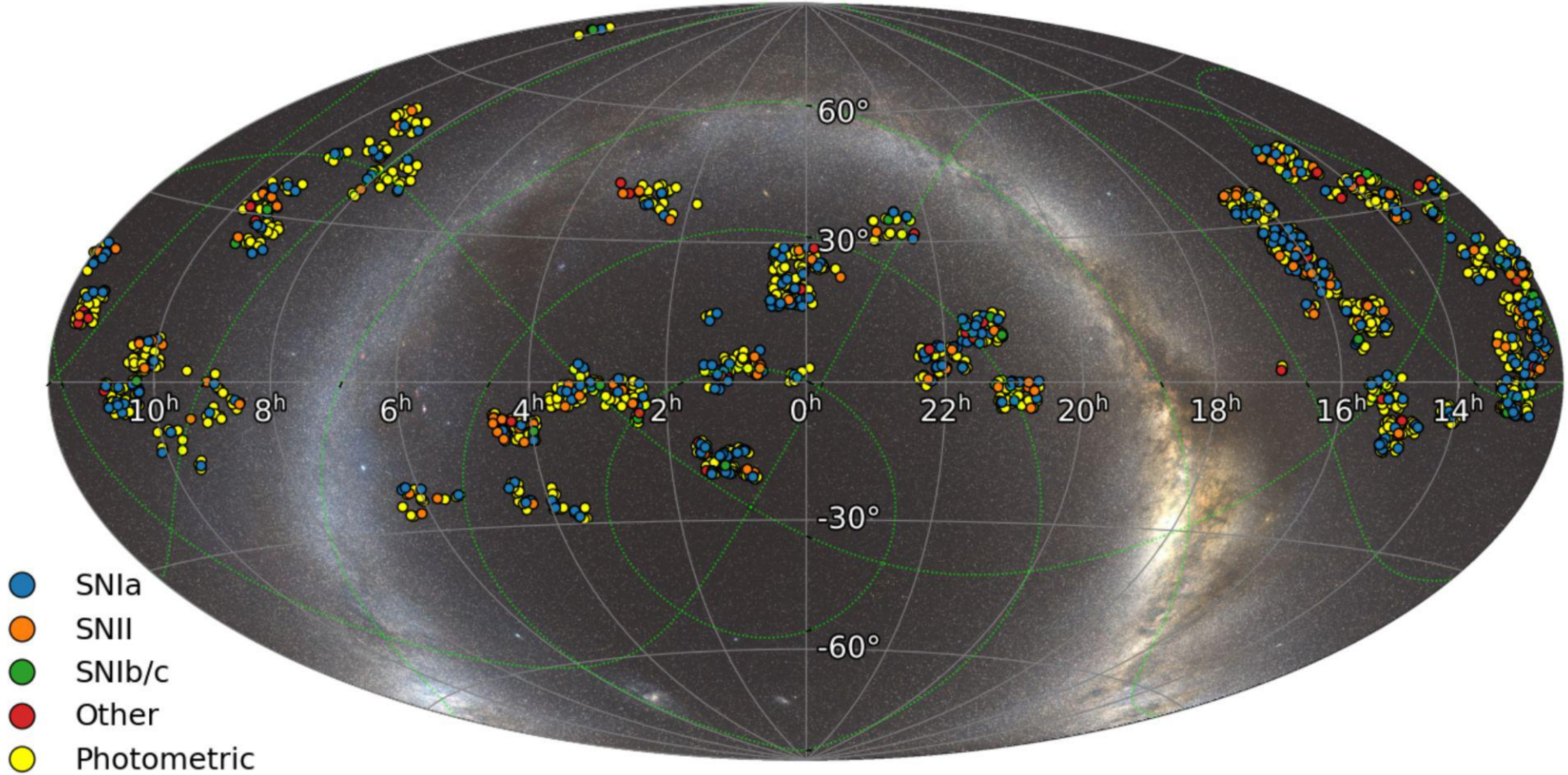
Young Supernova Experiment

Area: 1,500 deg^2

Depth: m_r~21.5

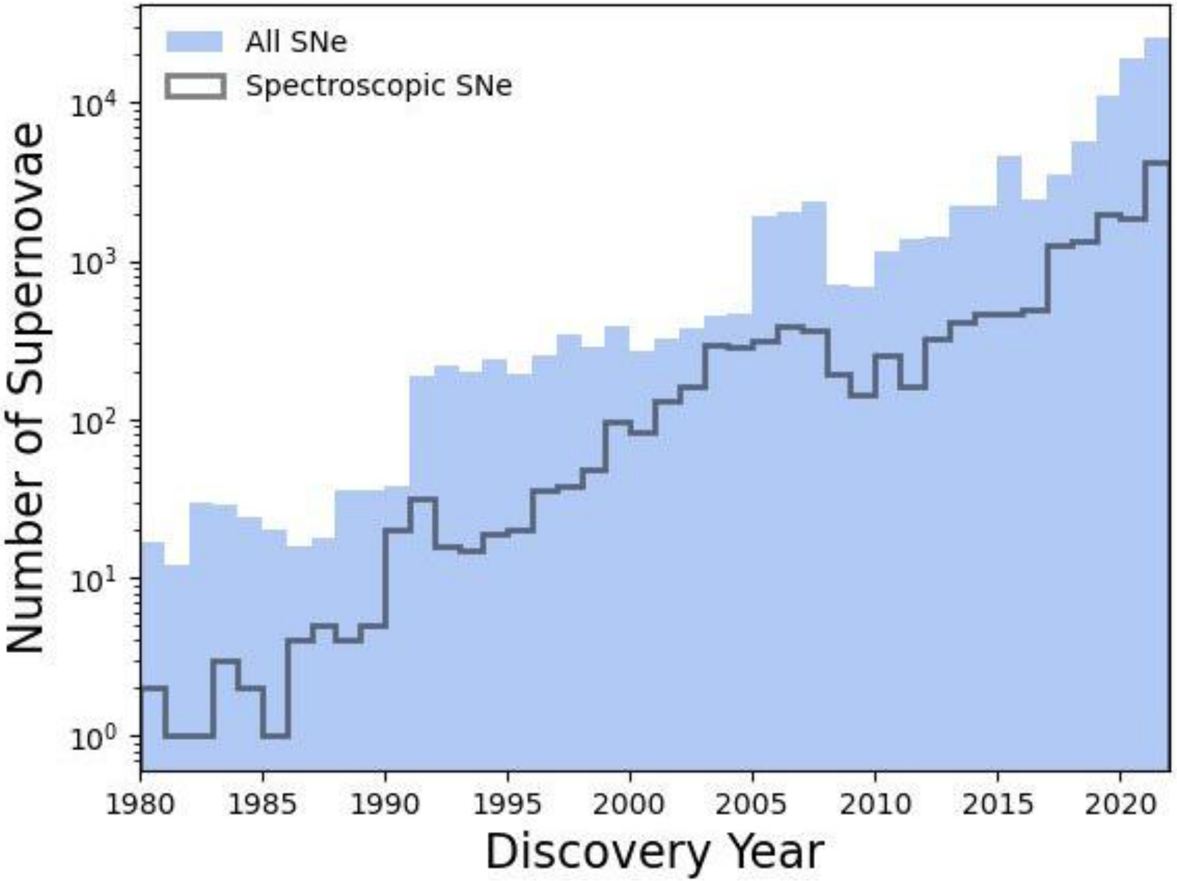


First data release now available - 1,975 supernovae!



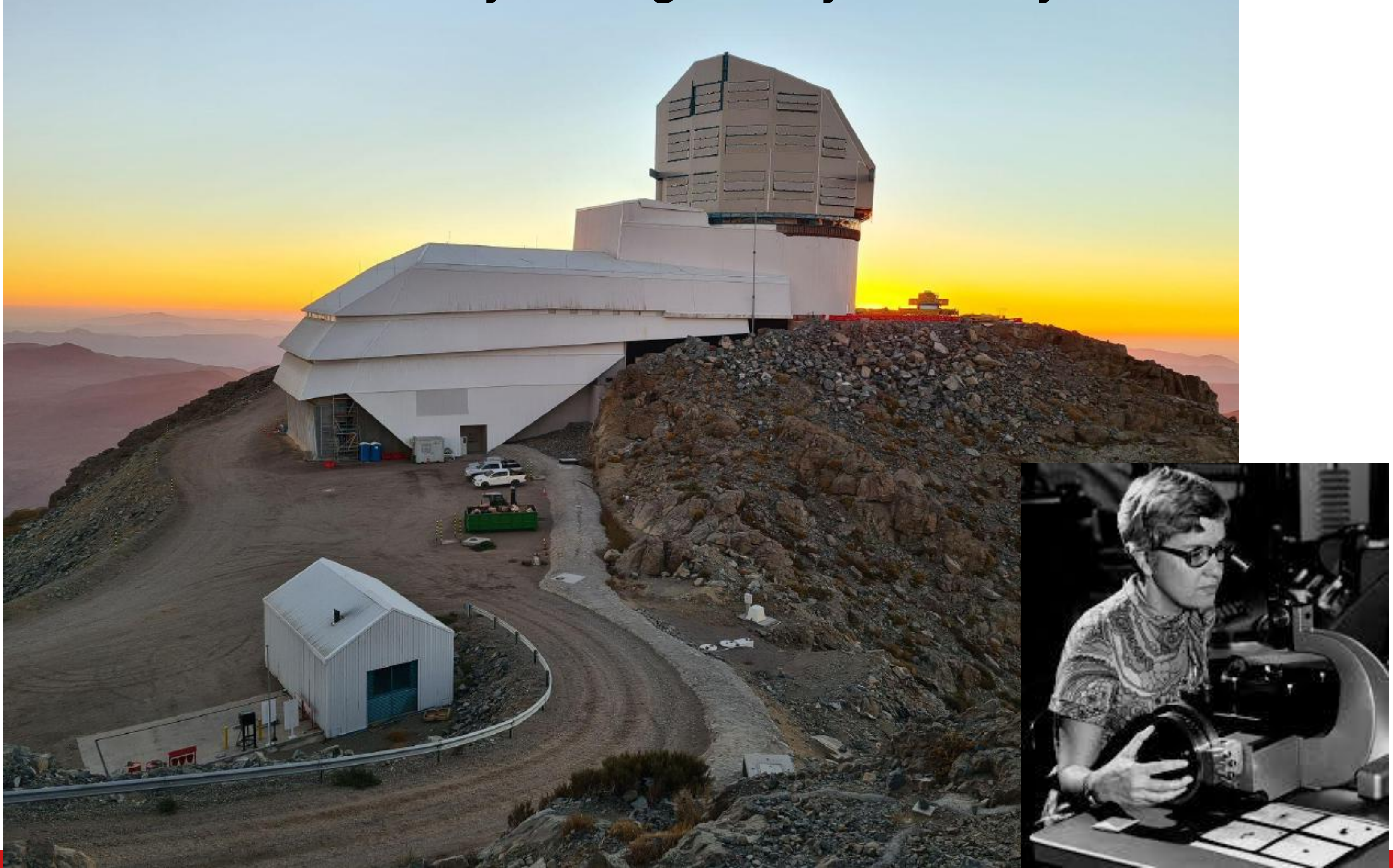
Aleo+23

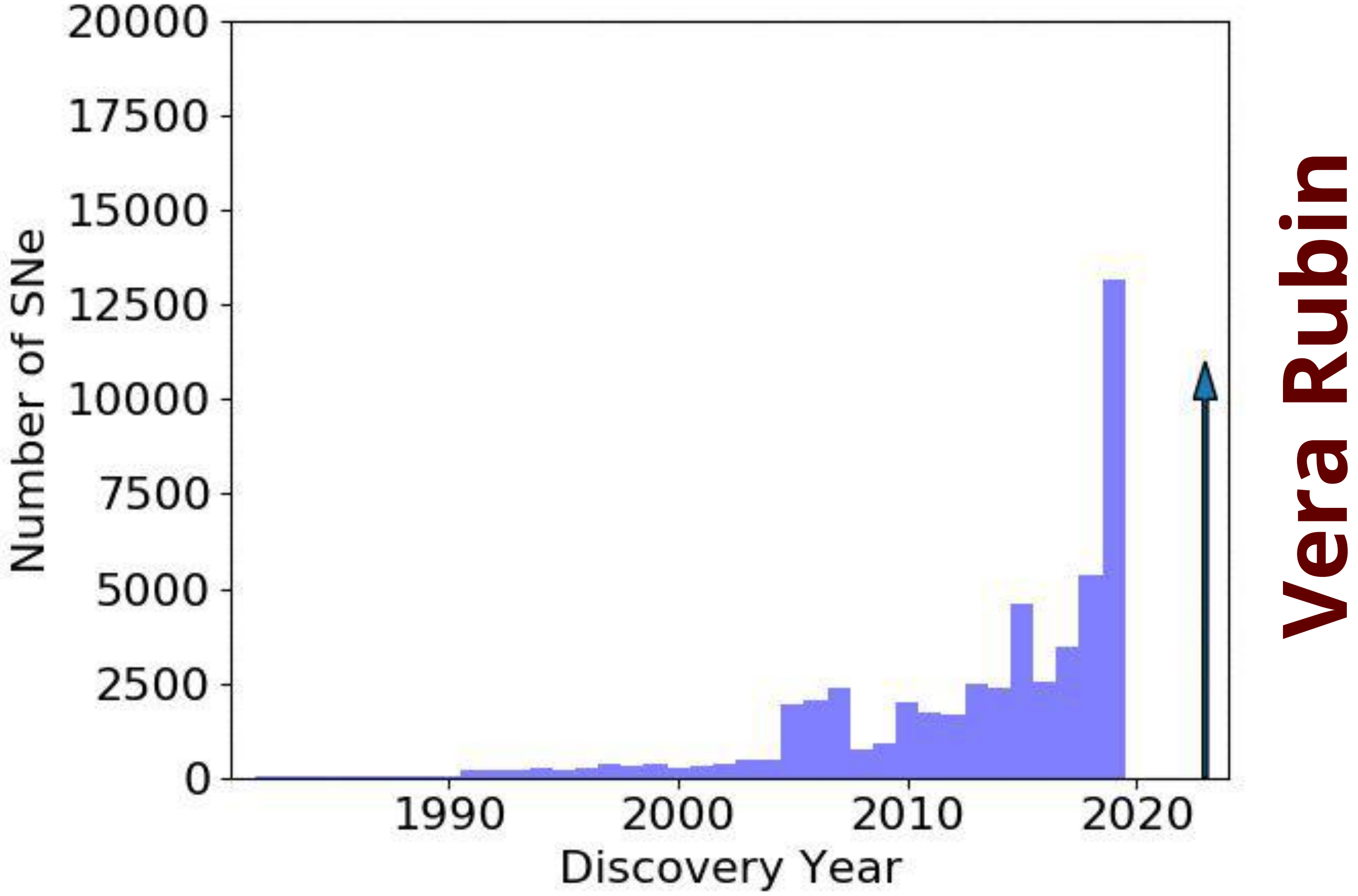
We currently discover ~20,000 supernovae annually

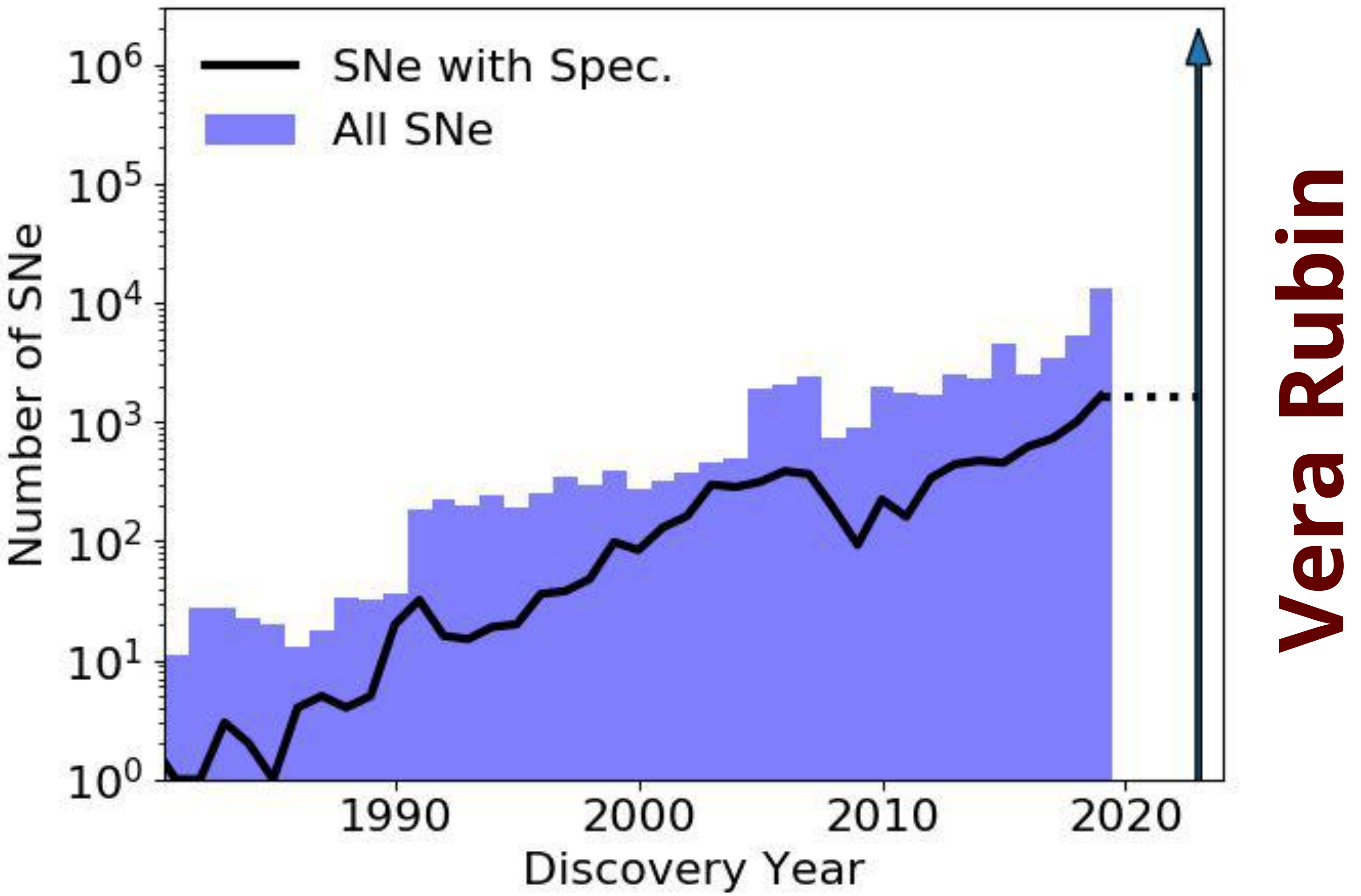


Data from OSC, TNS

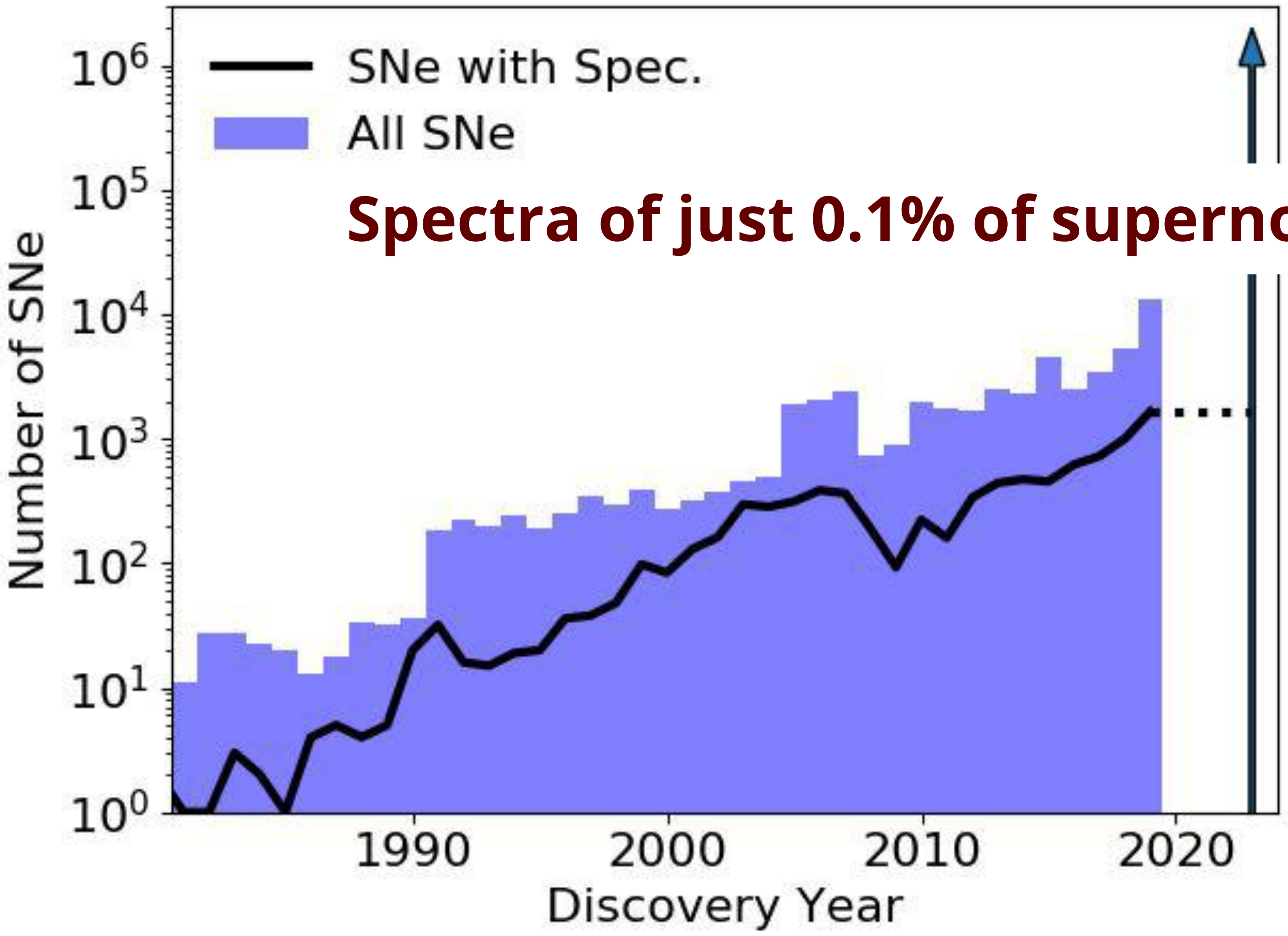
Vera Rubin Observatory will begin a 10-year survey in 2025




20000
17500
15000
12500
10000
7500
5000
2500
0
Number of SNe
1990
2000
2010
2020
Discovery Year
Vera Rubin

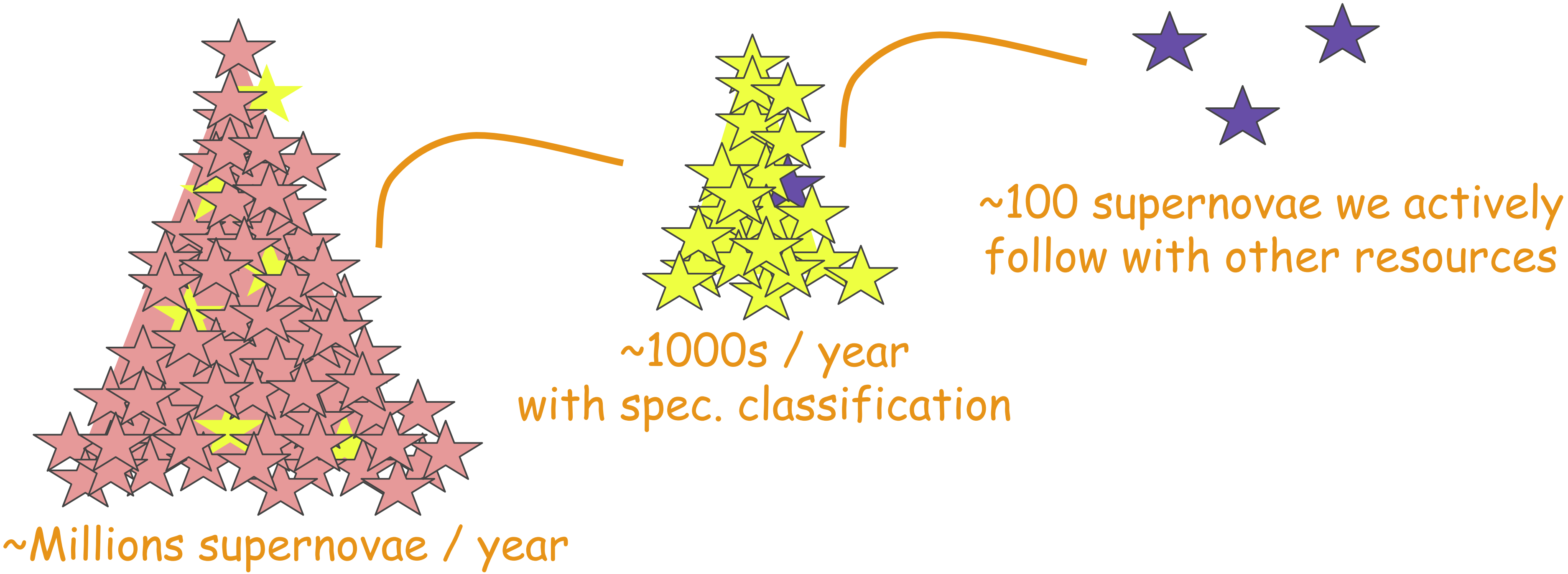


SNe with Spec.
All SNe
Number of SNe
10^0
10^1
10^2
10^3
10^4
10^5
10^6
1990
2000
2010
2020
Discovery Year
Vera Rubin



Spectra of just 0.1% of supernovae!

The VRO Needles & the Haystack



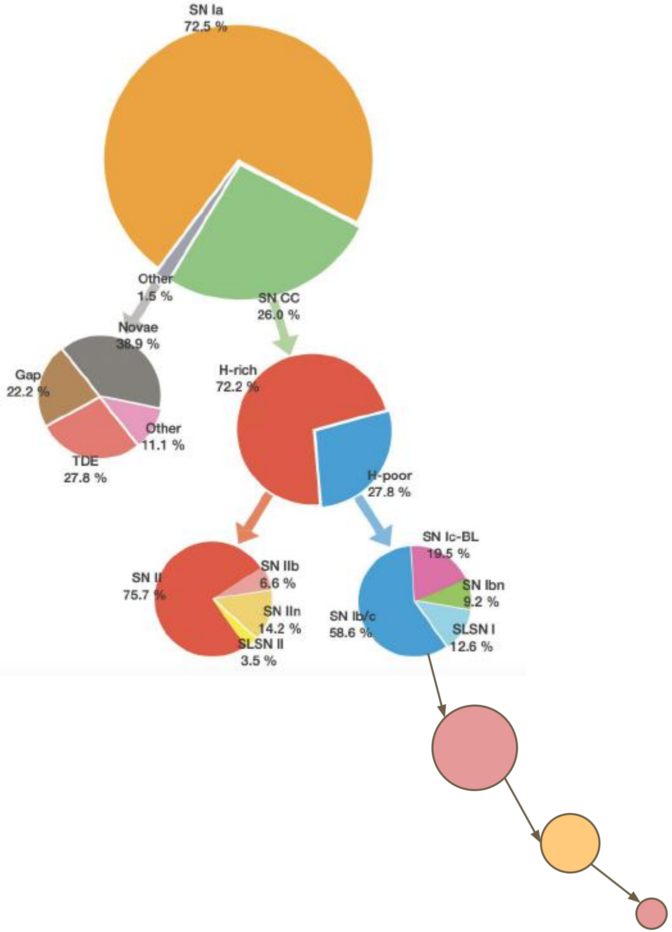
The VRO Needles & the Haystack

Only 1 in 10,000 SNe can be studied in real time

Perley+20

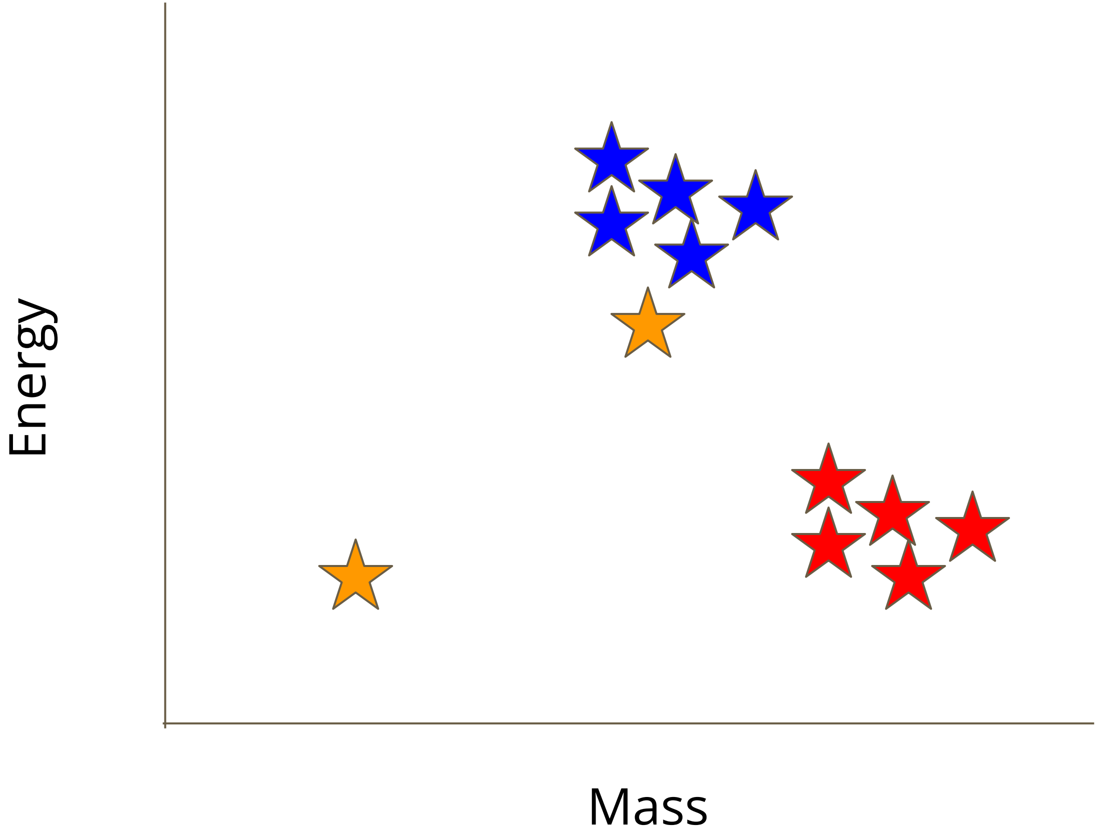
Even the MOST RARE classes of supernovae will be incredibly common in the era of the Vera Rubin Observatory!

We need to be ready for the "unknown unknowns"

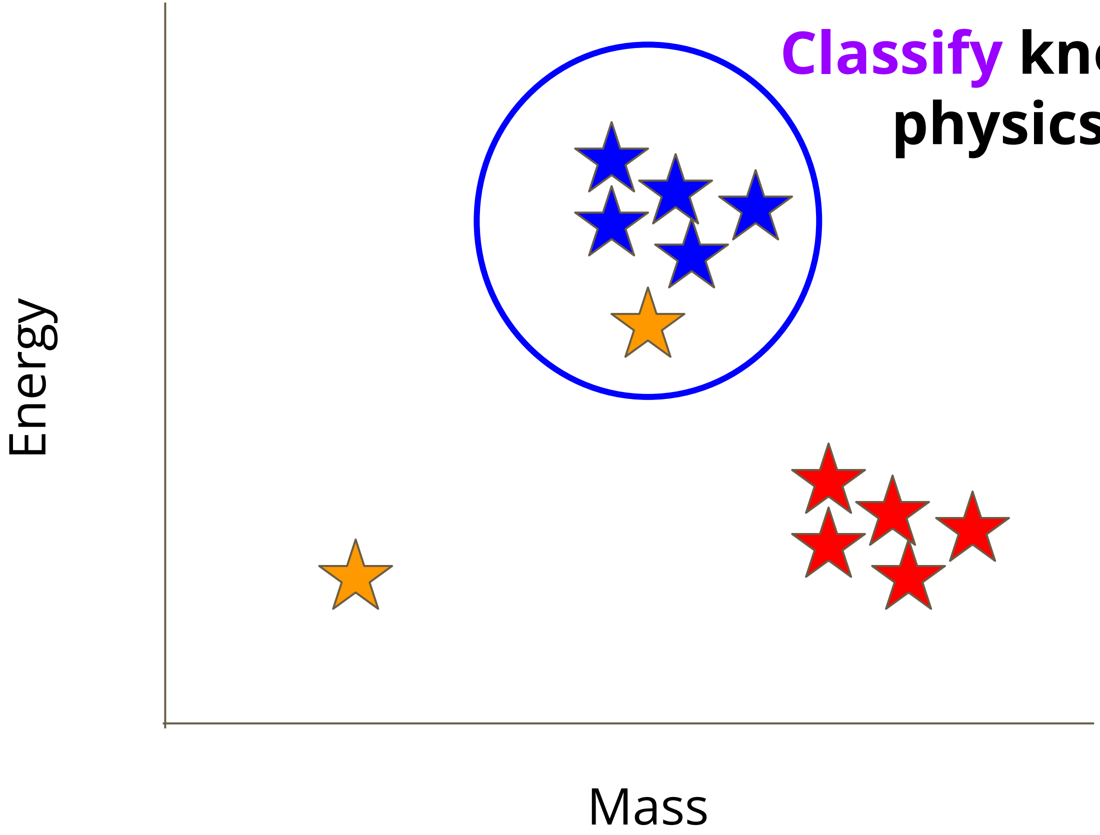


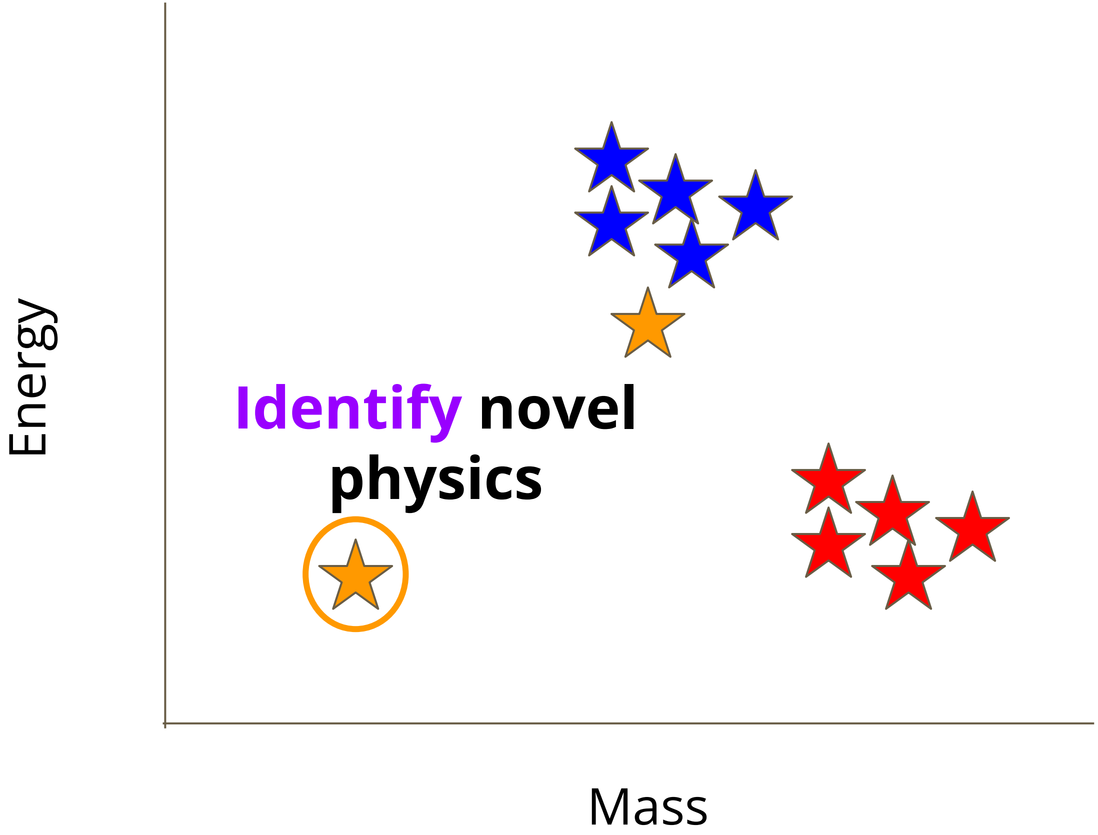
We need to go deeper!

Classify, Identify, Analyze


Energy
Mass

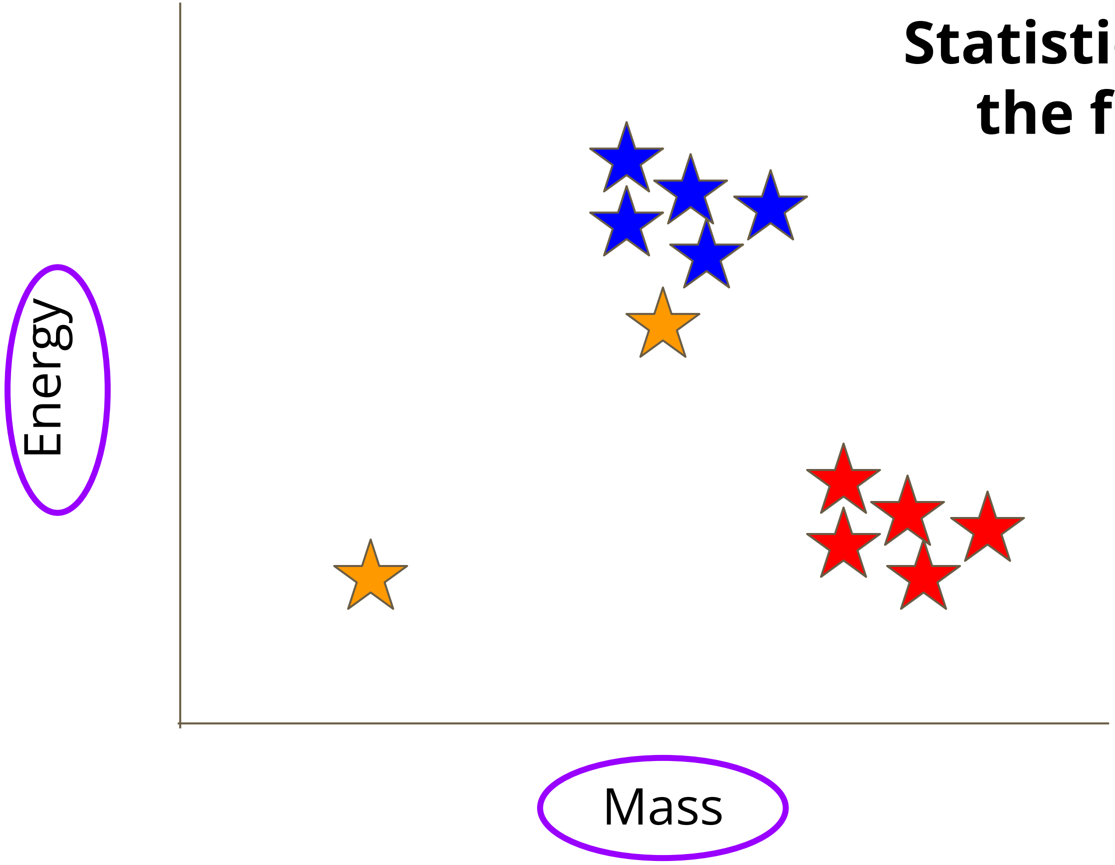
Classify known physics





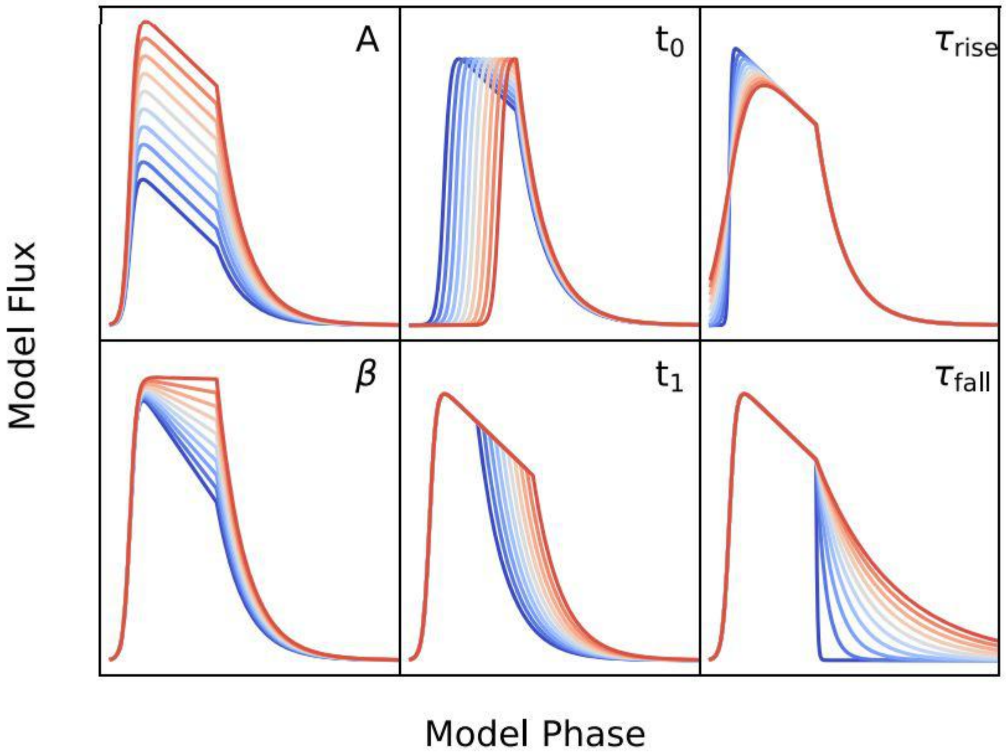
Energy
Identify novel physics
Mass

Statistically analyze the full sample

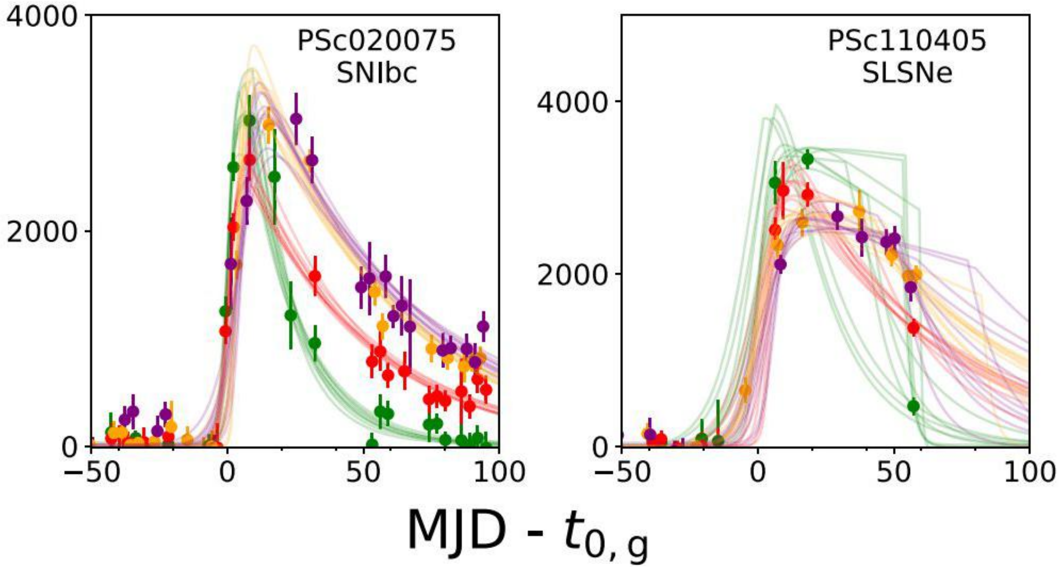


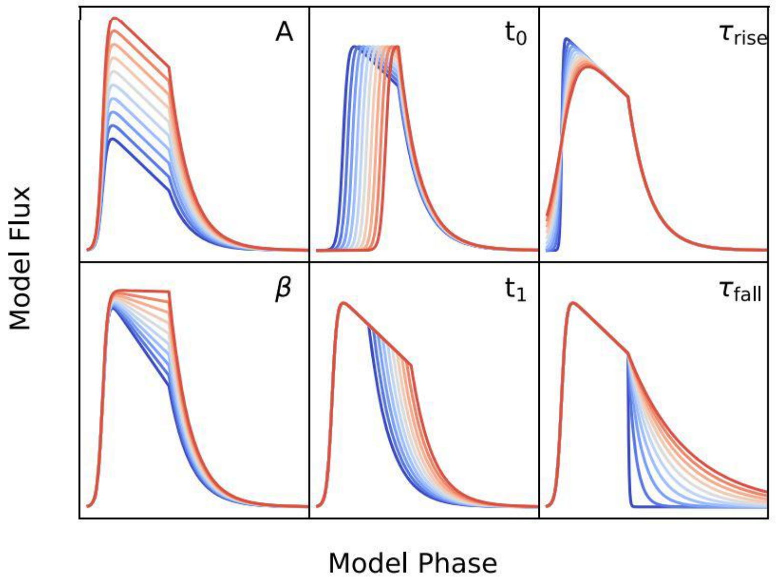
We will extract features from transient light curves and use them to classify events

A surefire way to extract meaningful features: fit a model



A surefire way to extract meaningful features: fit a model



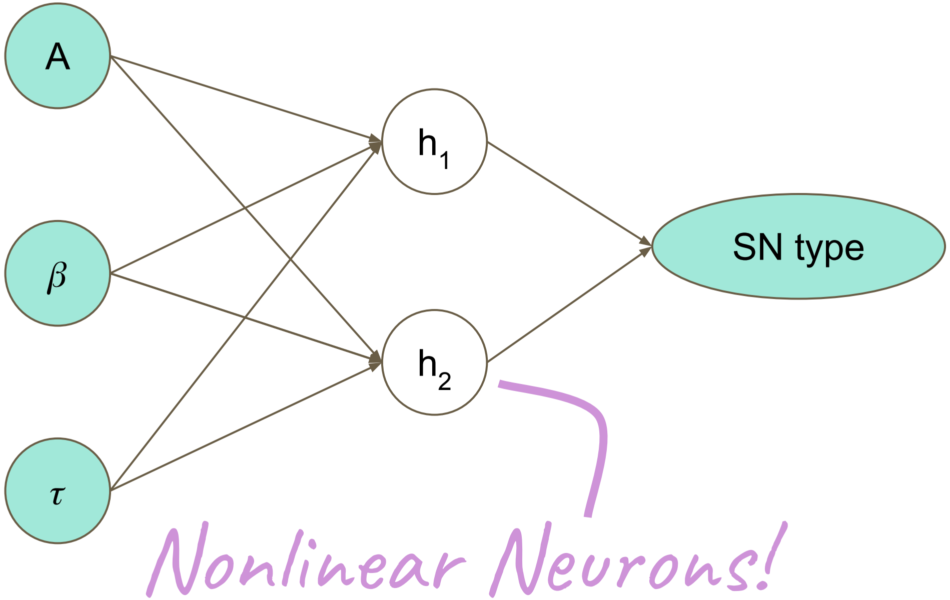


We don’t know the best combination of parameters to estimate a class probability

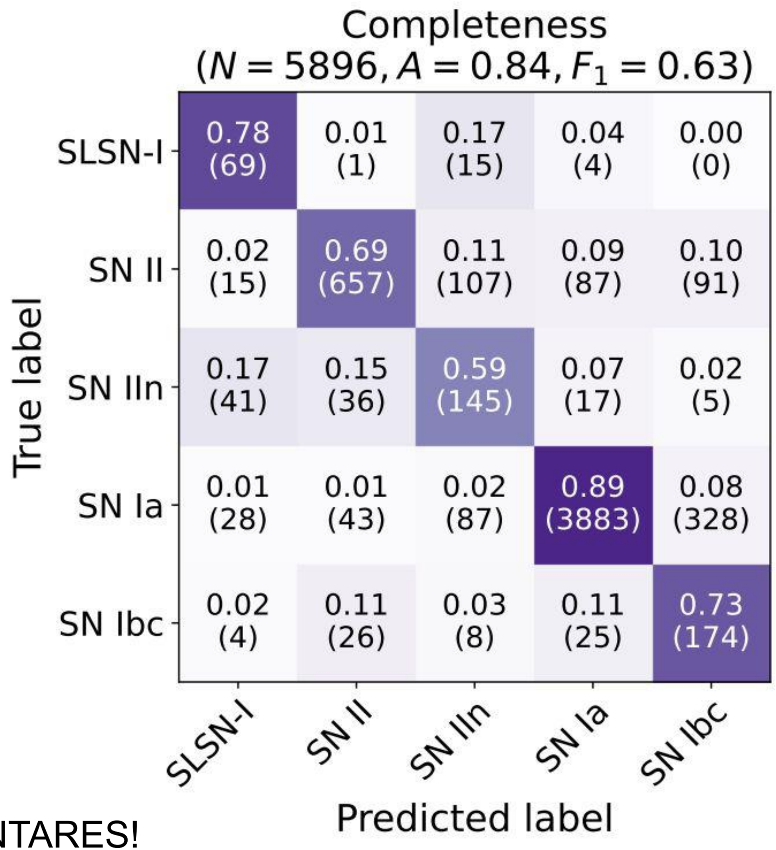
$$A * \tau_{\text{Rise}} + \beta \,/\, \tau_{\text{Fall}} = \text{probability of Type Ia?}$$

$$A * \beta + t_1 \,/\, \tau_{\text{Fall}} = \text{probability of Type II?}$$

A neural network will give us an approximate guess of this nonlinear function



Using supervised methods, we classify supernovae



de Soto*, VAV+ in prep - on ANTARES!
VAV, Gagliano, de Soto 2023

Our classification methods have been applied to...

Pan-STARRS Medium Deep Survey

(Villar+19, Villar+20, Hosseinzadeh+20)

Zwicky Transient Facility

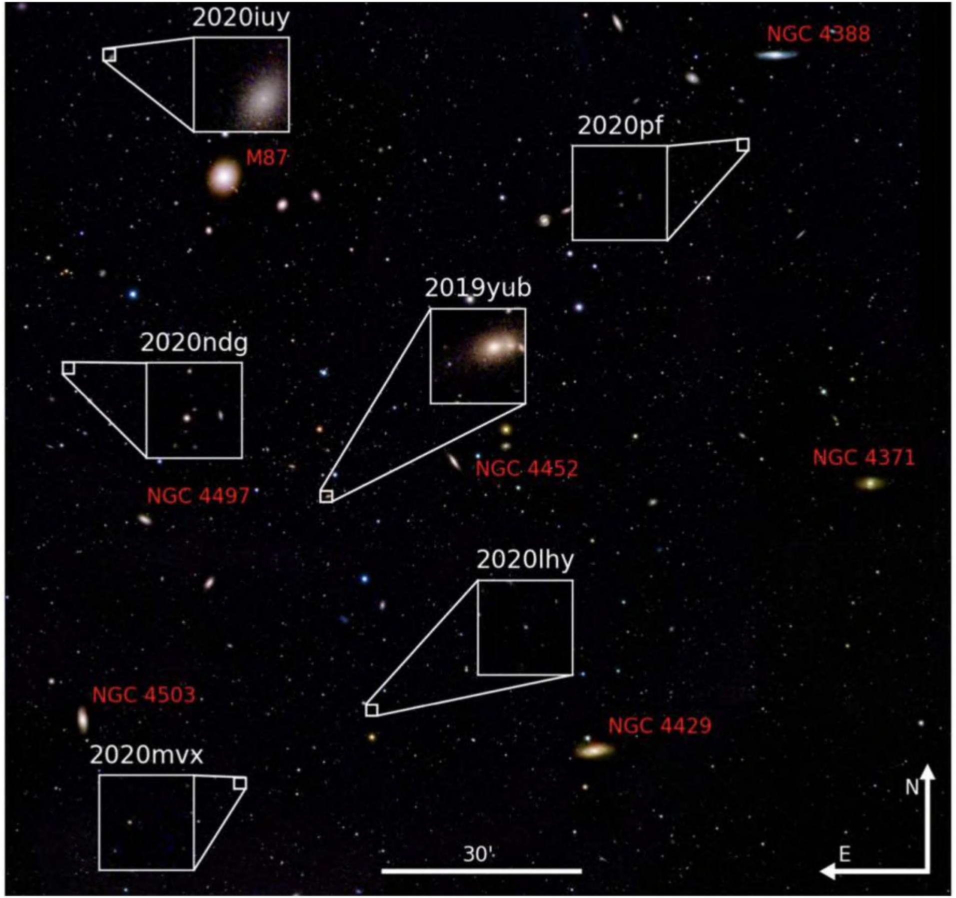
(de Soto* et al. in prep - filter in ANTARES)

Young Supernova Experiment

(Aleo+23)

Pan-STARRS

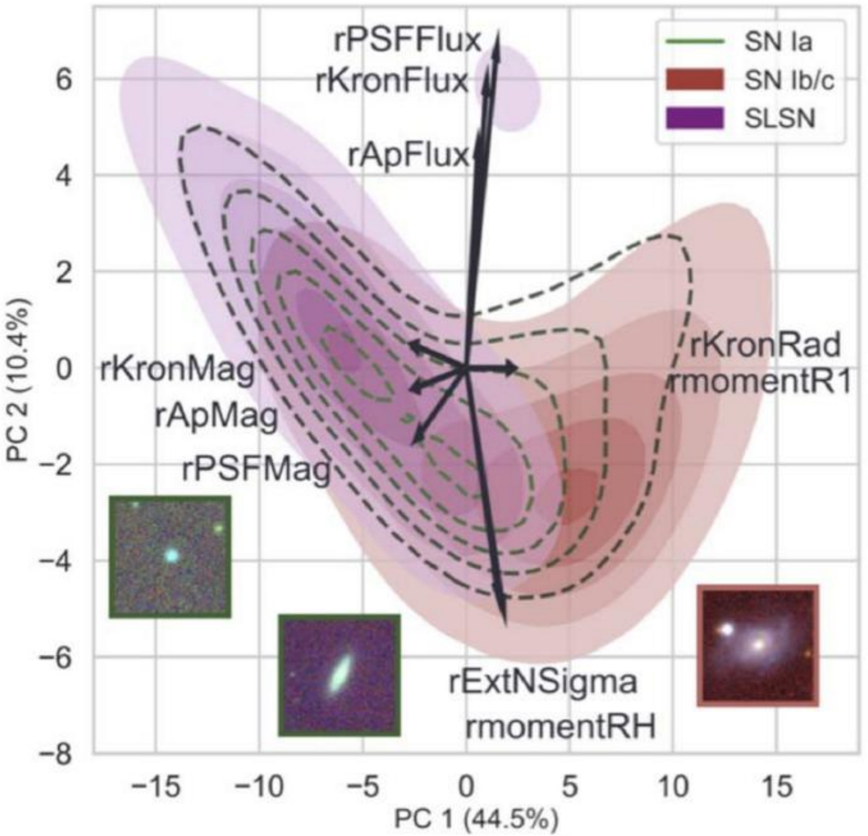
ZTF



We can also classify with 0 SN photons!

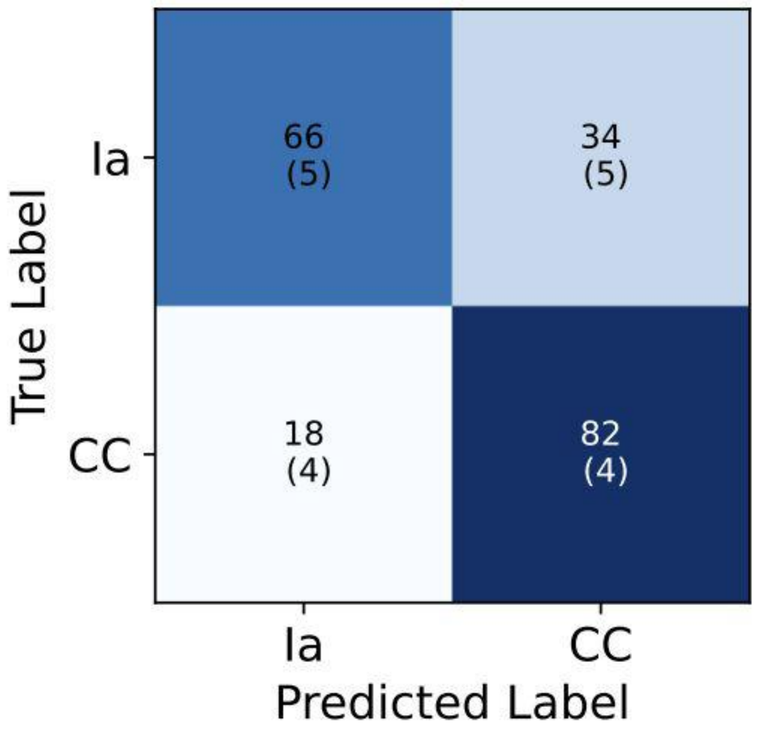
Host-galaxy classification

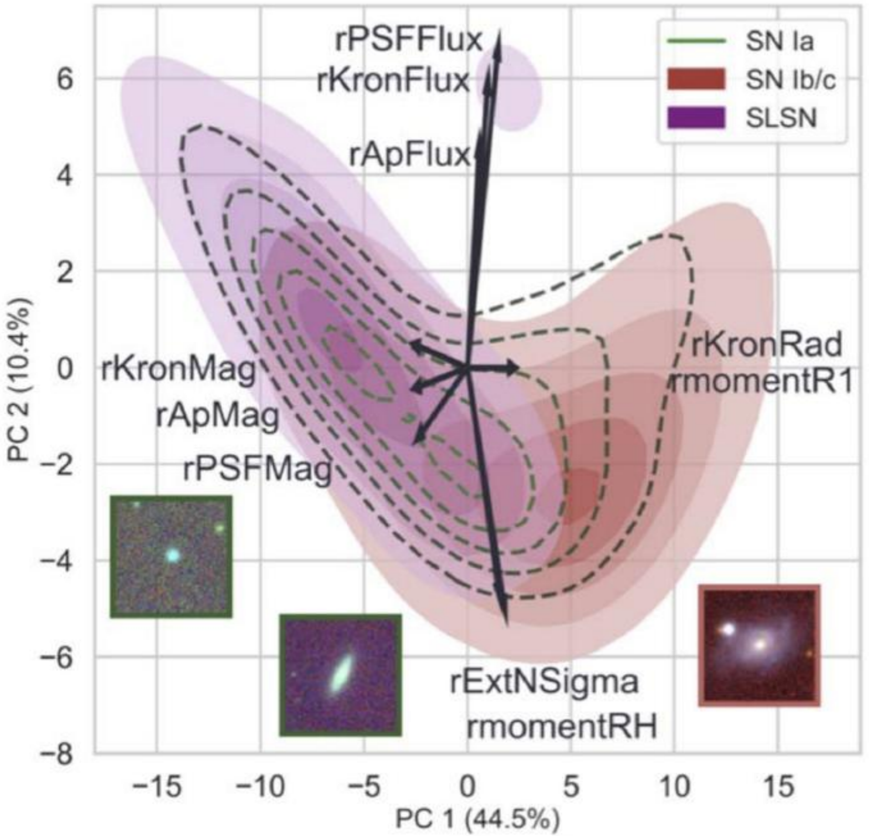
Supernovae know where they are born



VAV+ in prep; Gagliano+21

Host galaxy classification



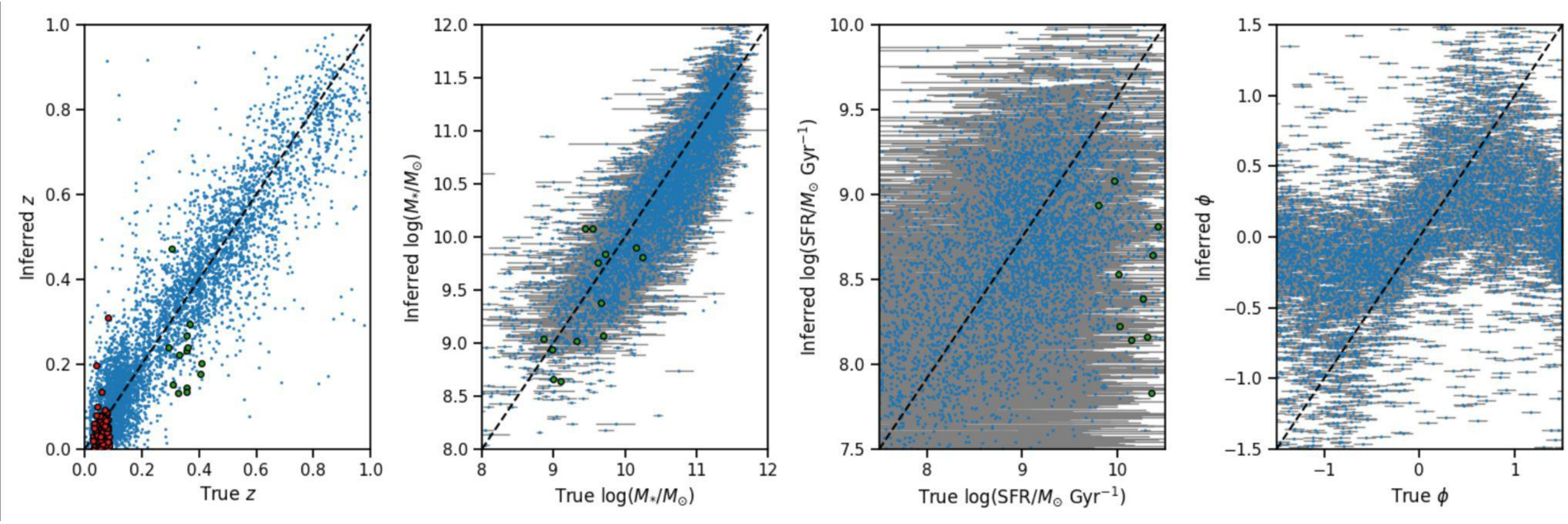


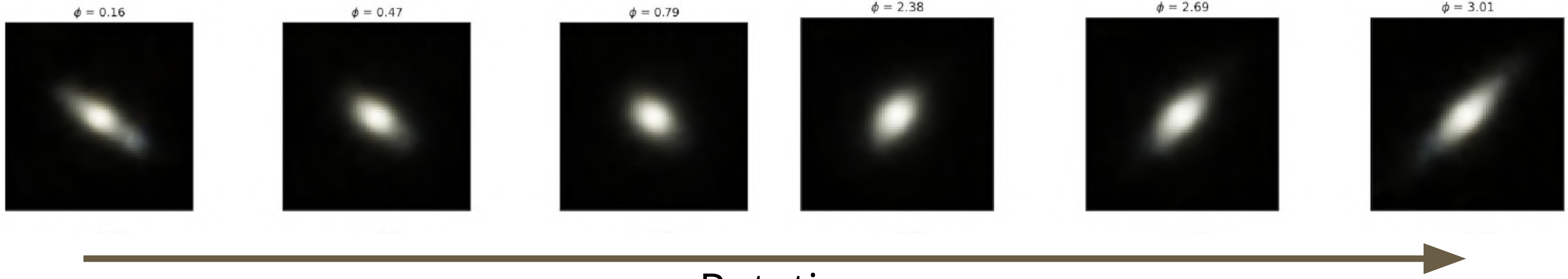
VAV+ in prep; Gagliano+21

Optimize a neural network to do the following:

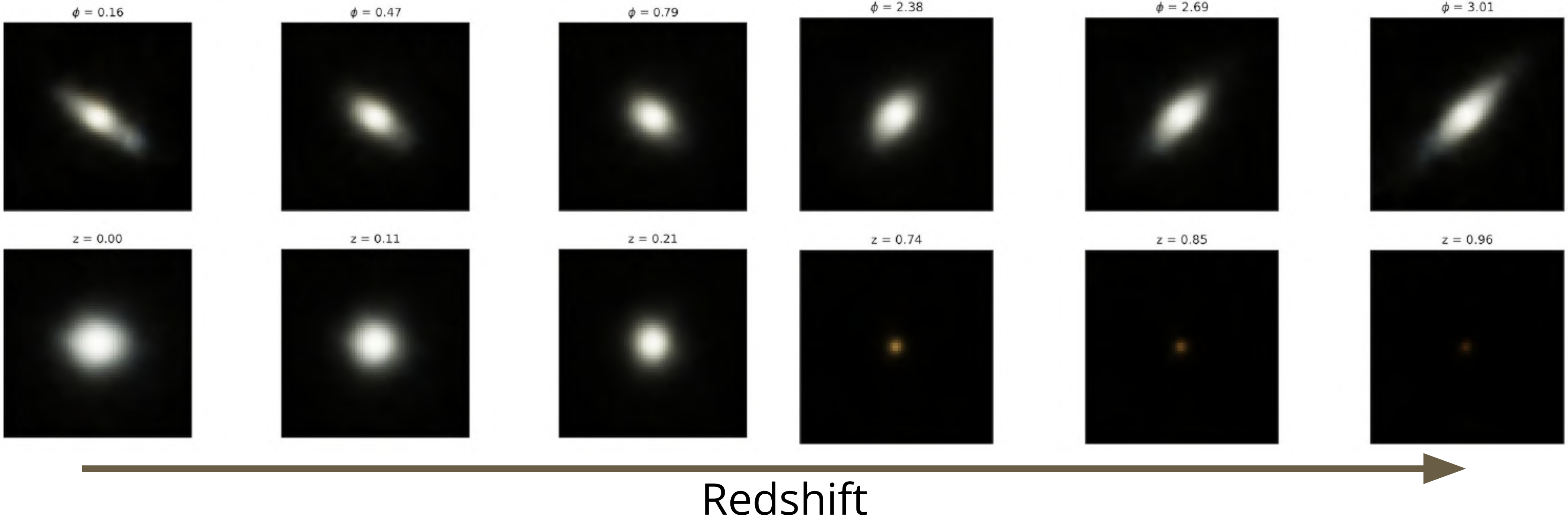
1. Predict the physical parameters of a galaxy
2. Be able to compress and then regenerate a galaxy image
3. (Make sure that the “representation space” of the galaxies is continuous –we’ll come back to this!)

From galaxy images alone, we can predict key parameters

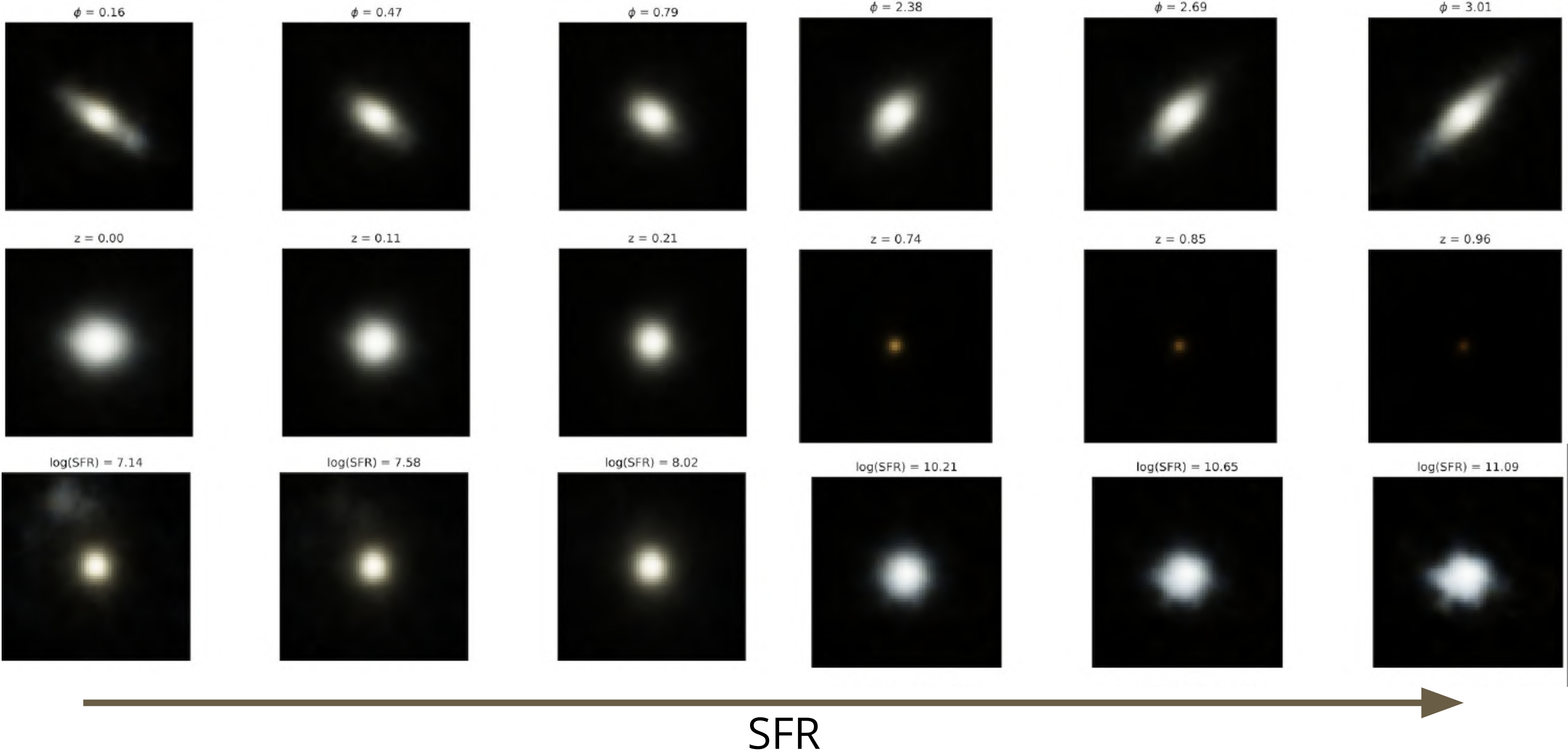


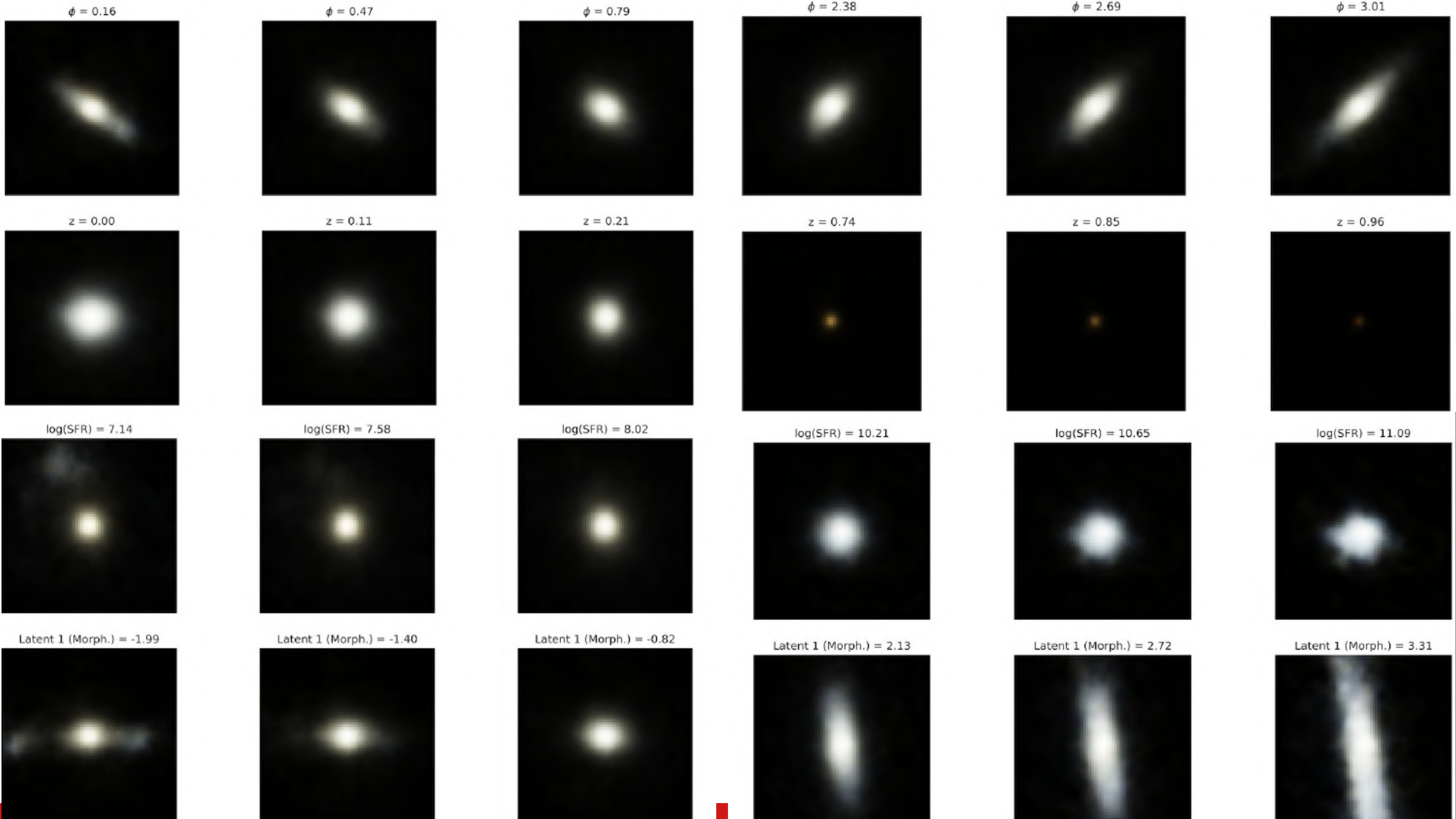


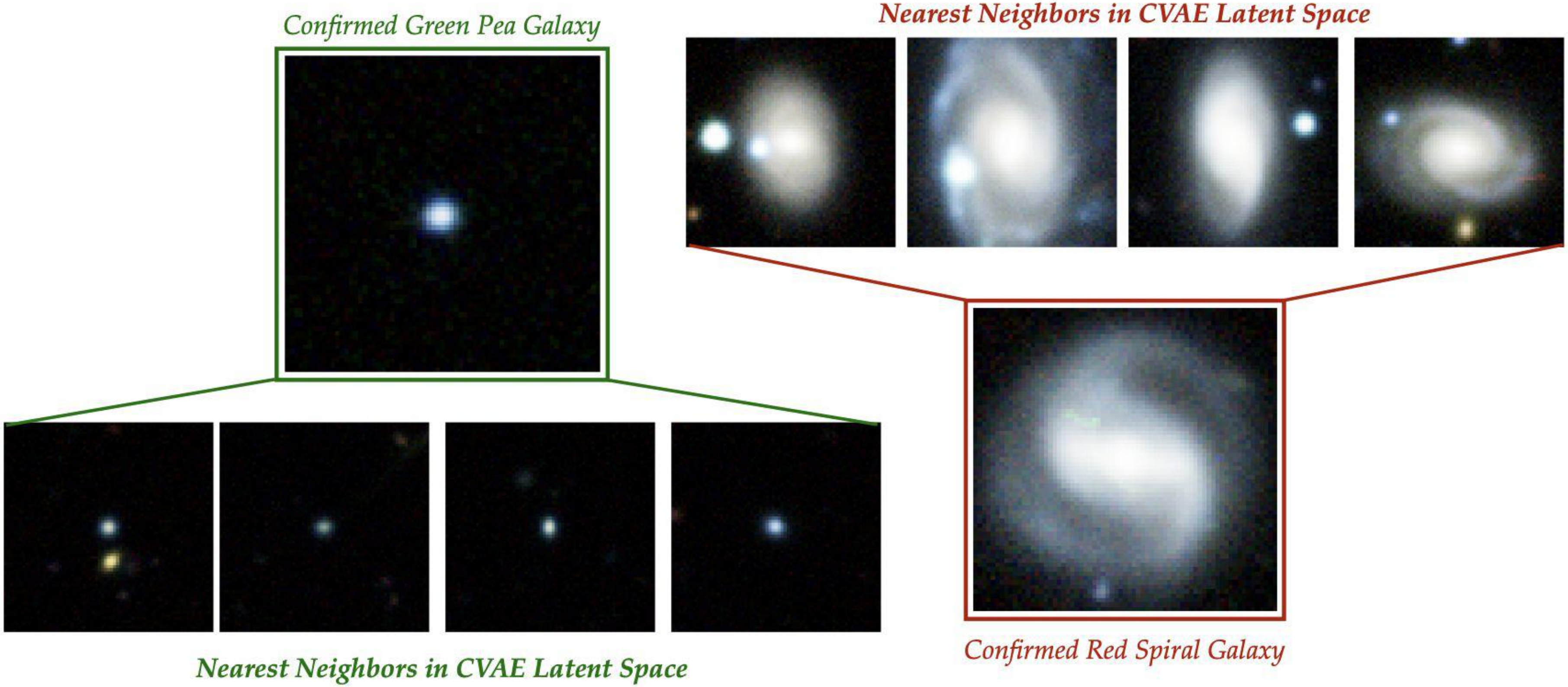
Rotation



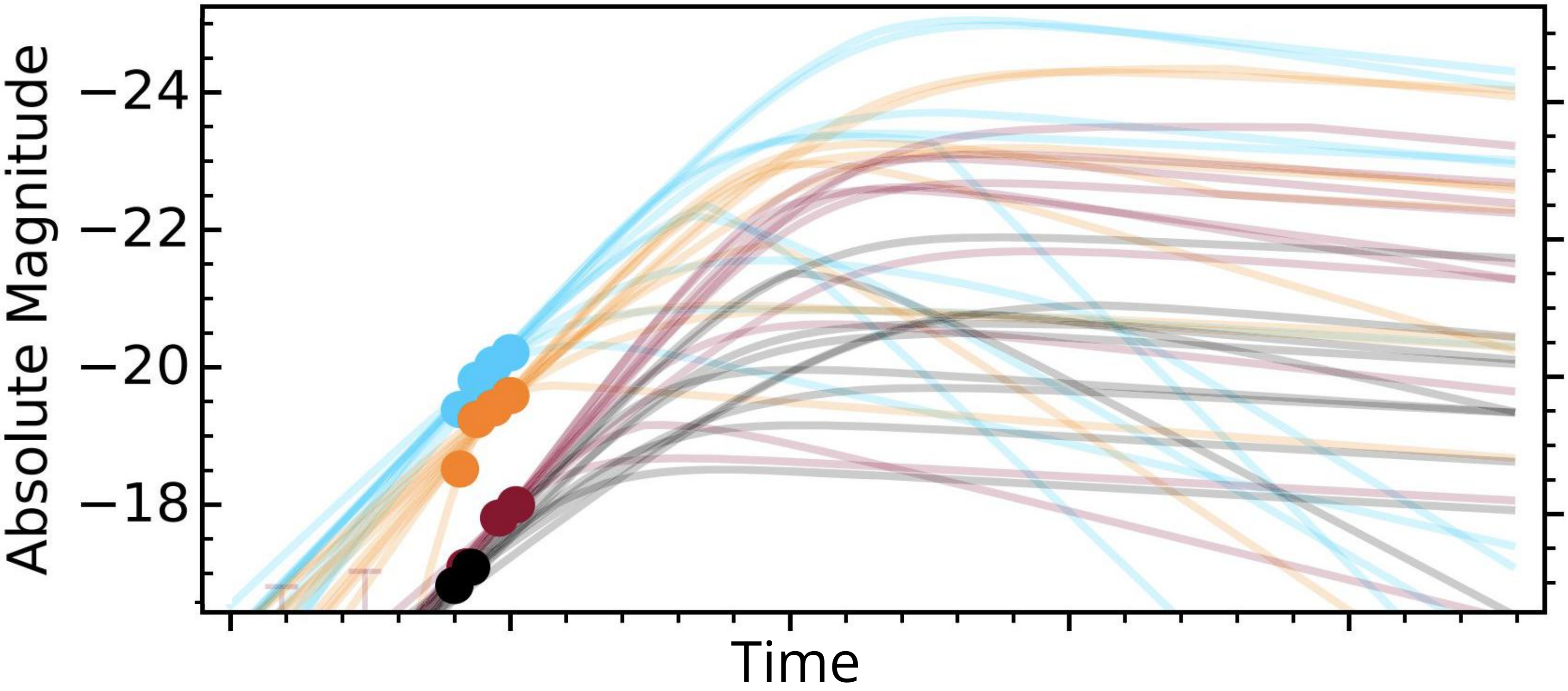
ϕ = 0.16
ϕ = 0.47
ϕ = 0.79
ϕ = 2.38
ϕ = 2.69
ϕ = 3.01
z = 0.00
z = 0.11
z = 0.21
z = 0.74
z = 0.85
z = 0.96
Redshift


ϕ = 0.16
ϕ = 0.47
ϕ = 0.79
ϕ = 2.38
ϕ = 2.69
ϕ = 3.01
z = 0.00
z = 0.11
z = 0.21
z = 0.74
z = 0.85
z = 0.96
log(SFR) = 7.14
log(SFR) = 7.58
log(SFR) = 8.02
log(SFR) = 10.21
log(SFR) = 10.65
log(SFR) = 11.09
SFR


ϕ = 0.16
ϕ = 0.47
ϕ = 0.79
ϕ = 2.38
ϕ = 2.69
ϕ = 3.01
z = 0.00
z = 0.11
z = 0.21
z = 0.74
z = 0.85
z = 0.96
log(SFR) = 7.14
log(SFR) = 7.58
log(SFR) = 8.02
log(SFR) = 10.21
log(SFR) = 10.65
log(SFR) = 11.09
Latent 1 (Morph.) = -1.99
Latent 1 (Morph.) = -1.40
Latent 1 (Morph.) = -0.82
Latent 1 (Morph.) = 2.13
Latent 1 (Morph.) = 2.72
Latent 1 (Morph.) = 3.31


Confirmed Green Pea Galaxy
Nearest Neighbors in CVAE Latent Space
Nearest Neighbors in CVAE Latent Space
Confirmed Red Spiral Galaxy

But what about identifying interesting events in real time?



Hosseinzadeh+ 2020

A data-driven, **unsupervised** method using a **variational, recurrent neuron-based autoencoder**

Aside: Data-driven methods require *data*

Real:

Pan-STARRS Medium Deep Survey

Zwicky Transient Facility

Young Supernova Experiment

Sim: PLAsTiCC

Community effort, with ~20 classes of transients

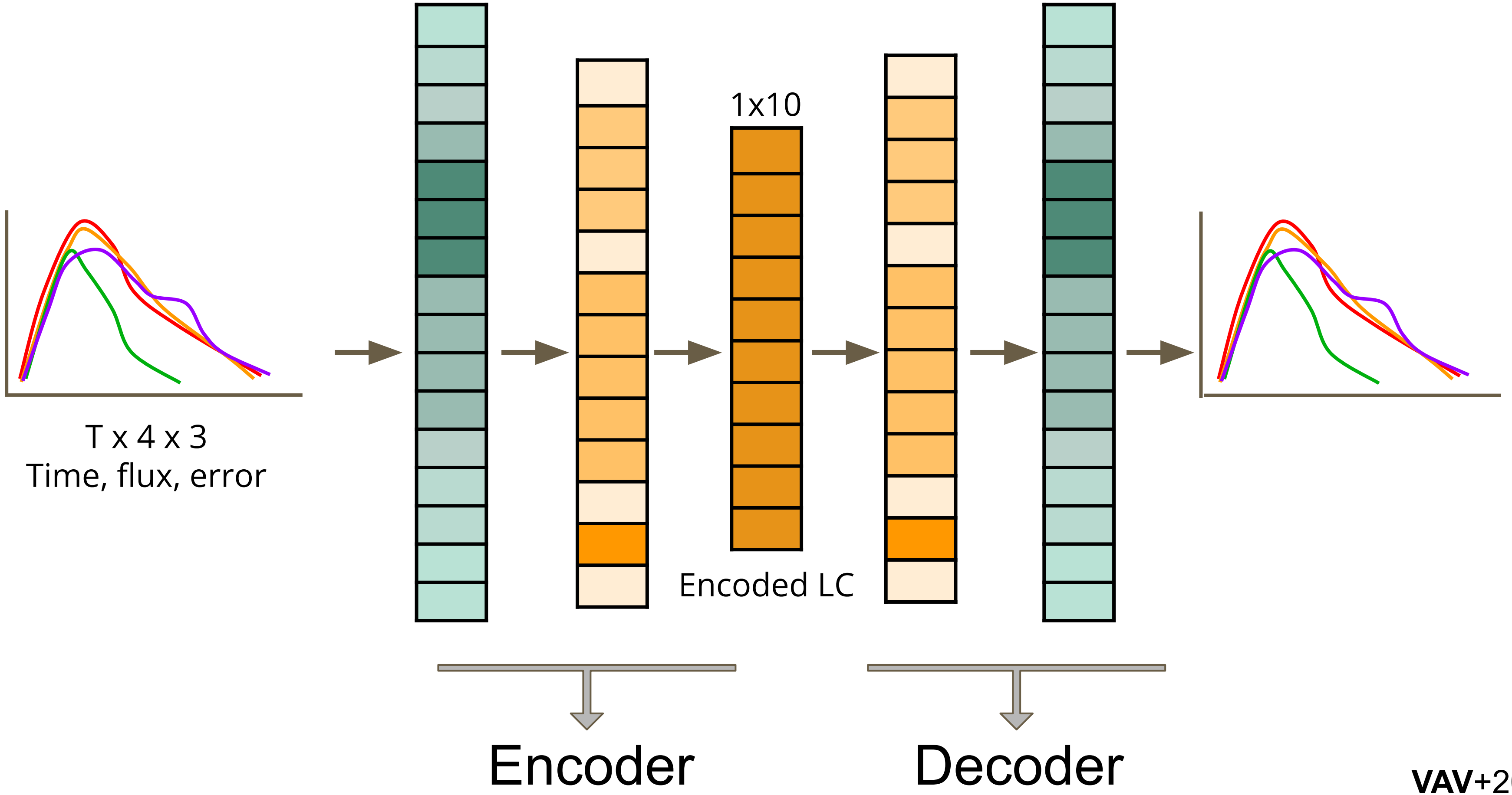
~1 million SN-like transients in 3 years of LSST

Every event tagged with physical parameters

Chambers+16, VAV+20, Hosseinzadeh+20

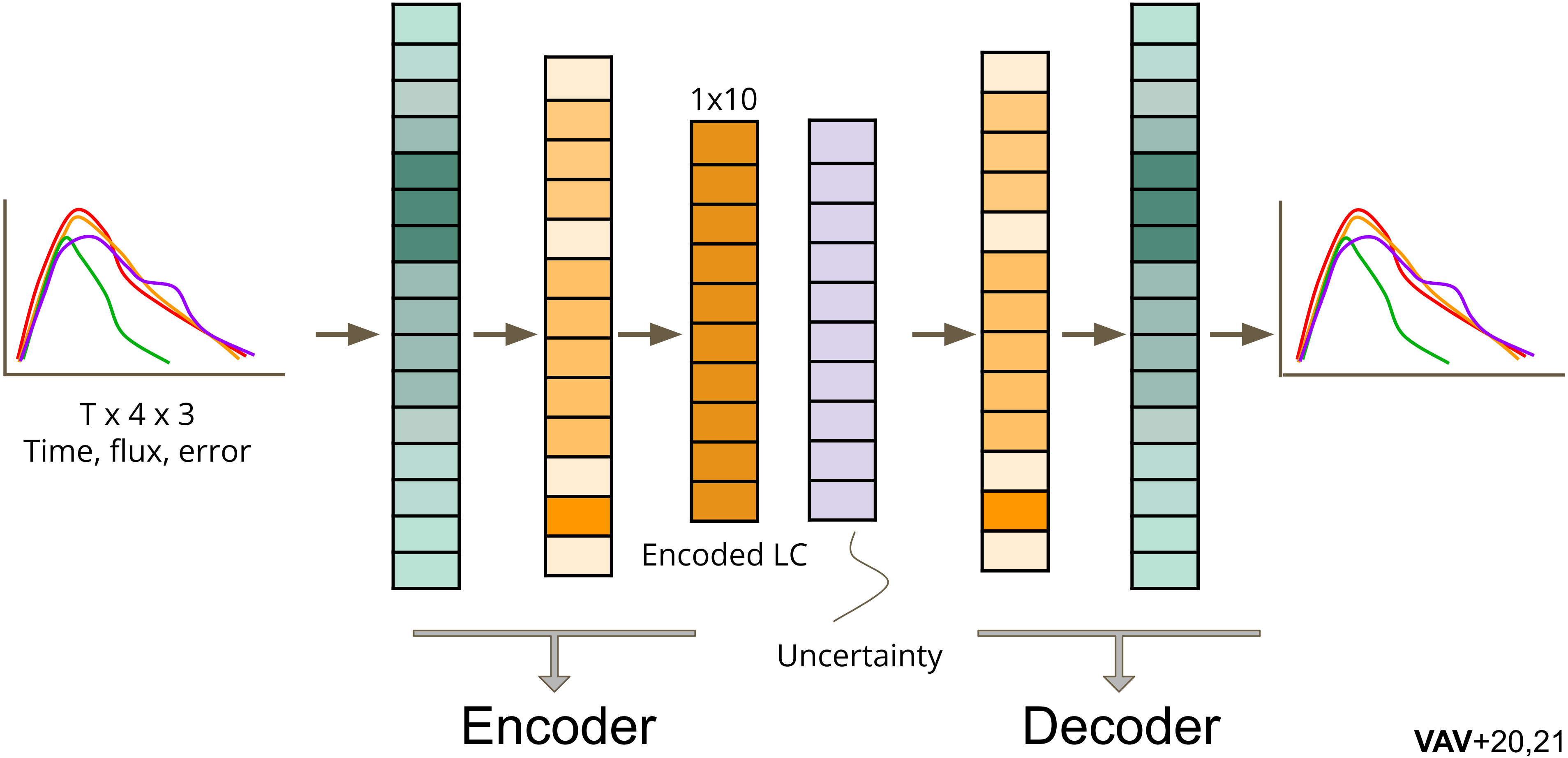
Kessler+19, Hložek+20

Use an autoencoder to *encode* the full sample



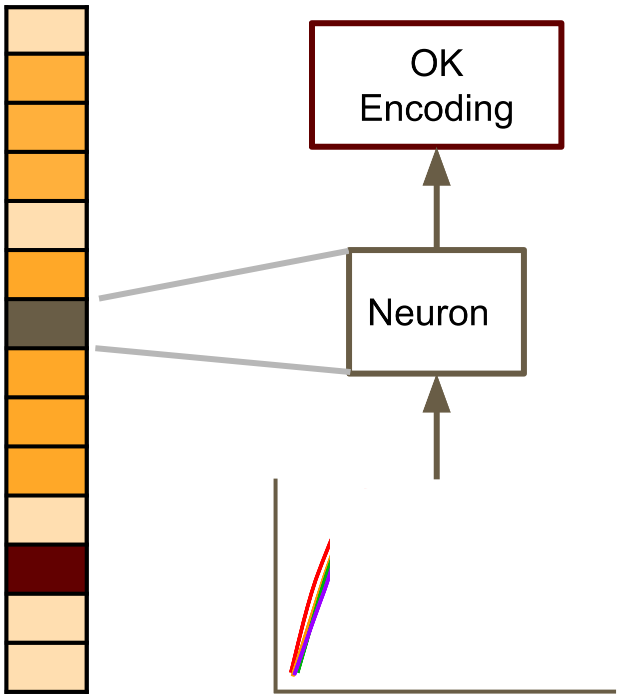
VAV+20,21

Use a <u>variational</u> autoencoder to *encode* the full sample

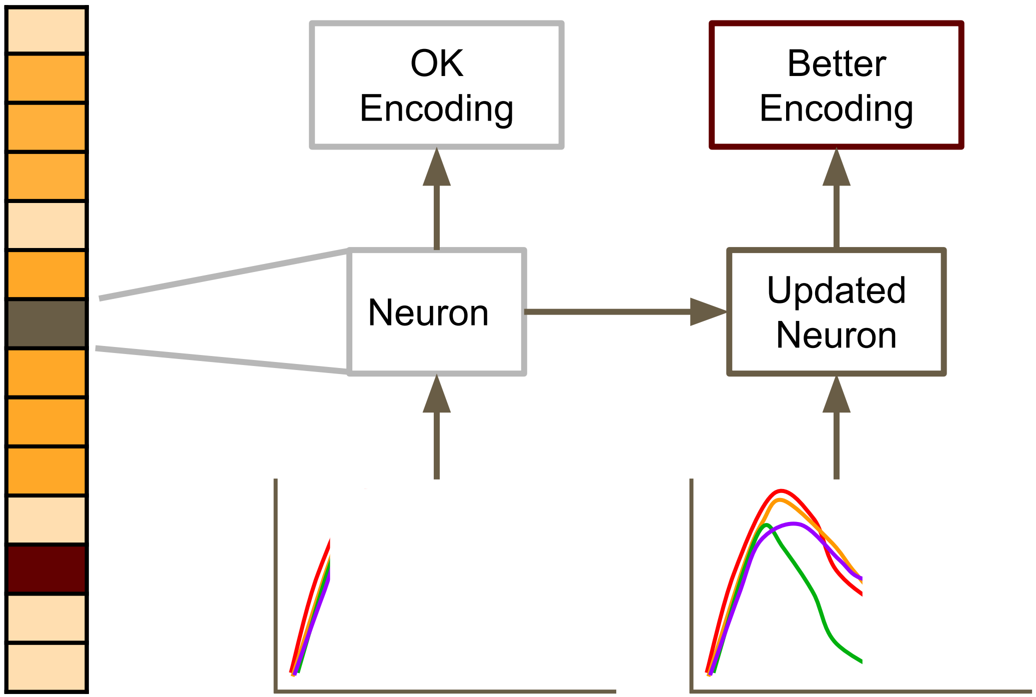


VAV+20,21

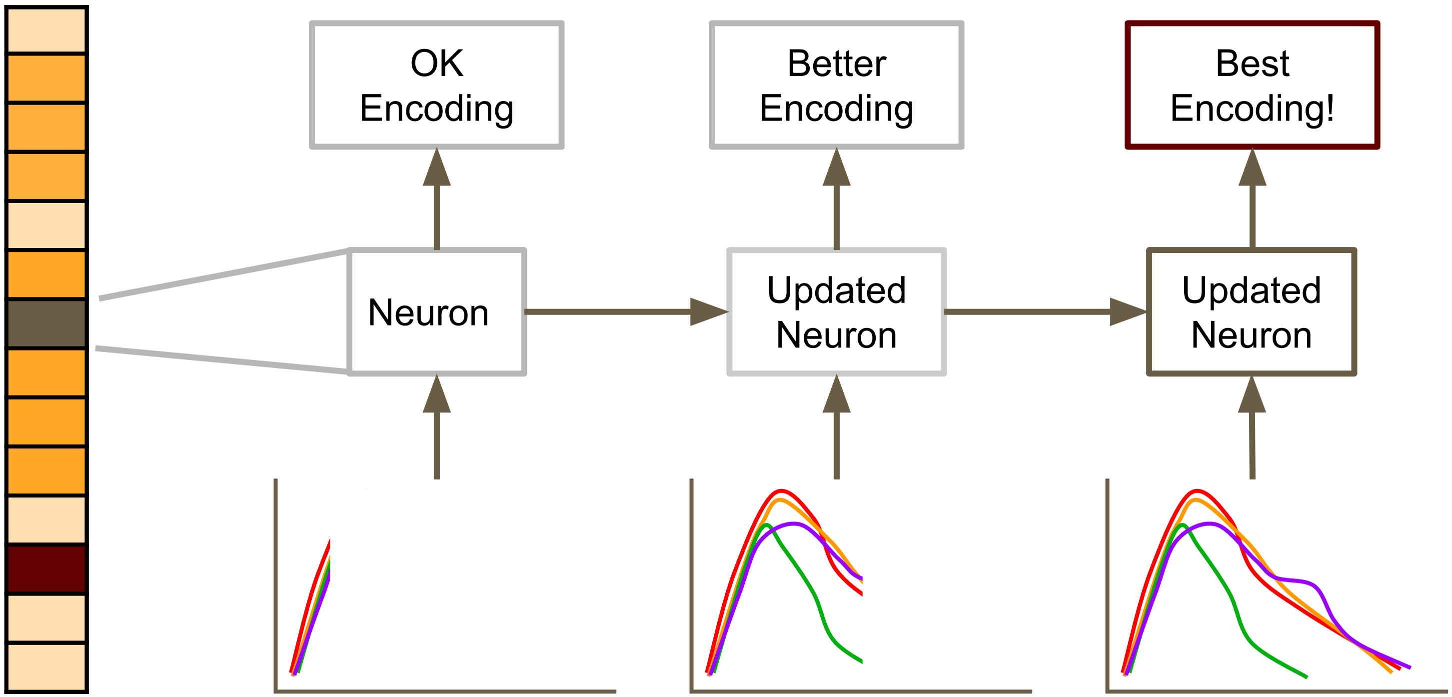
Use recurrent neurons to utilize new data



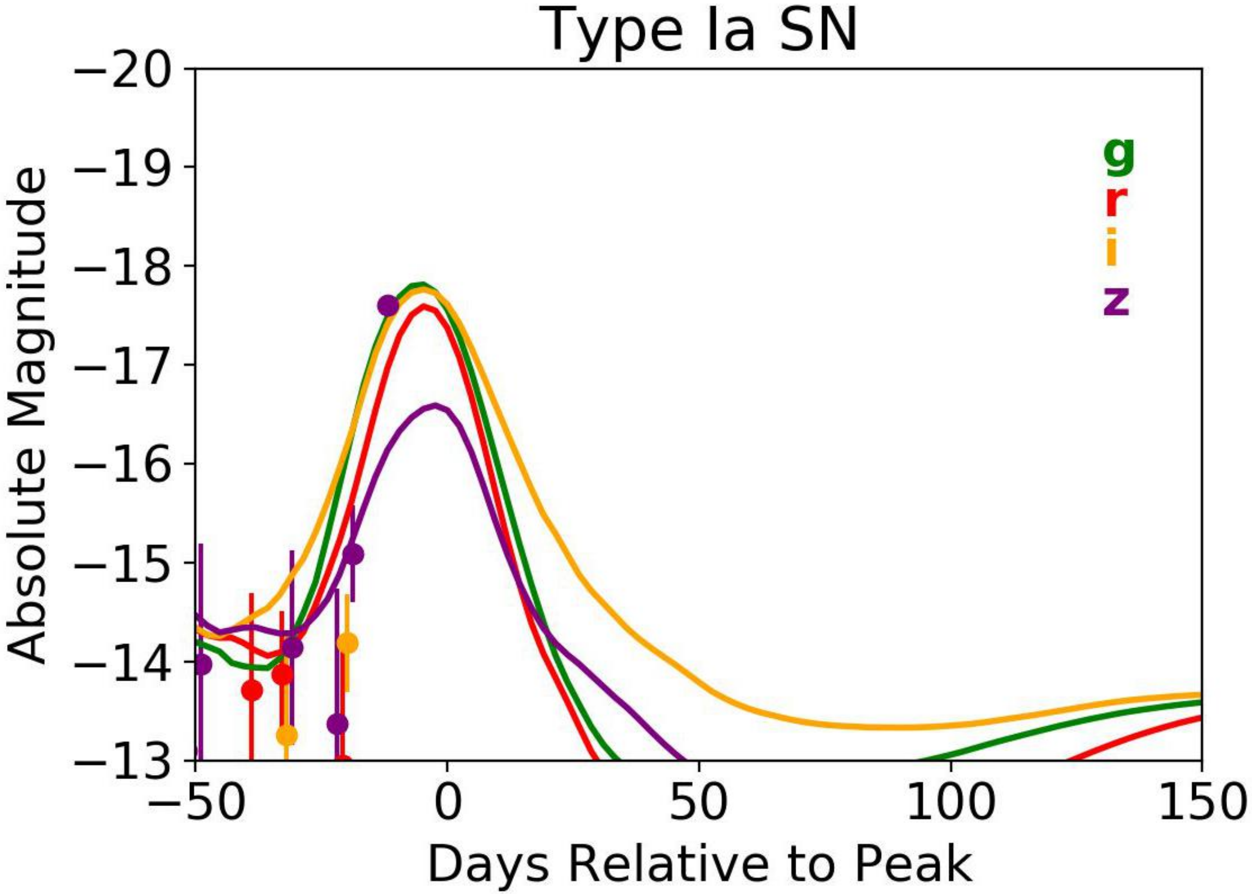
Use recurrent neurons to utilize new data



Use recurrent neurons to utilize new data



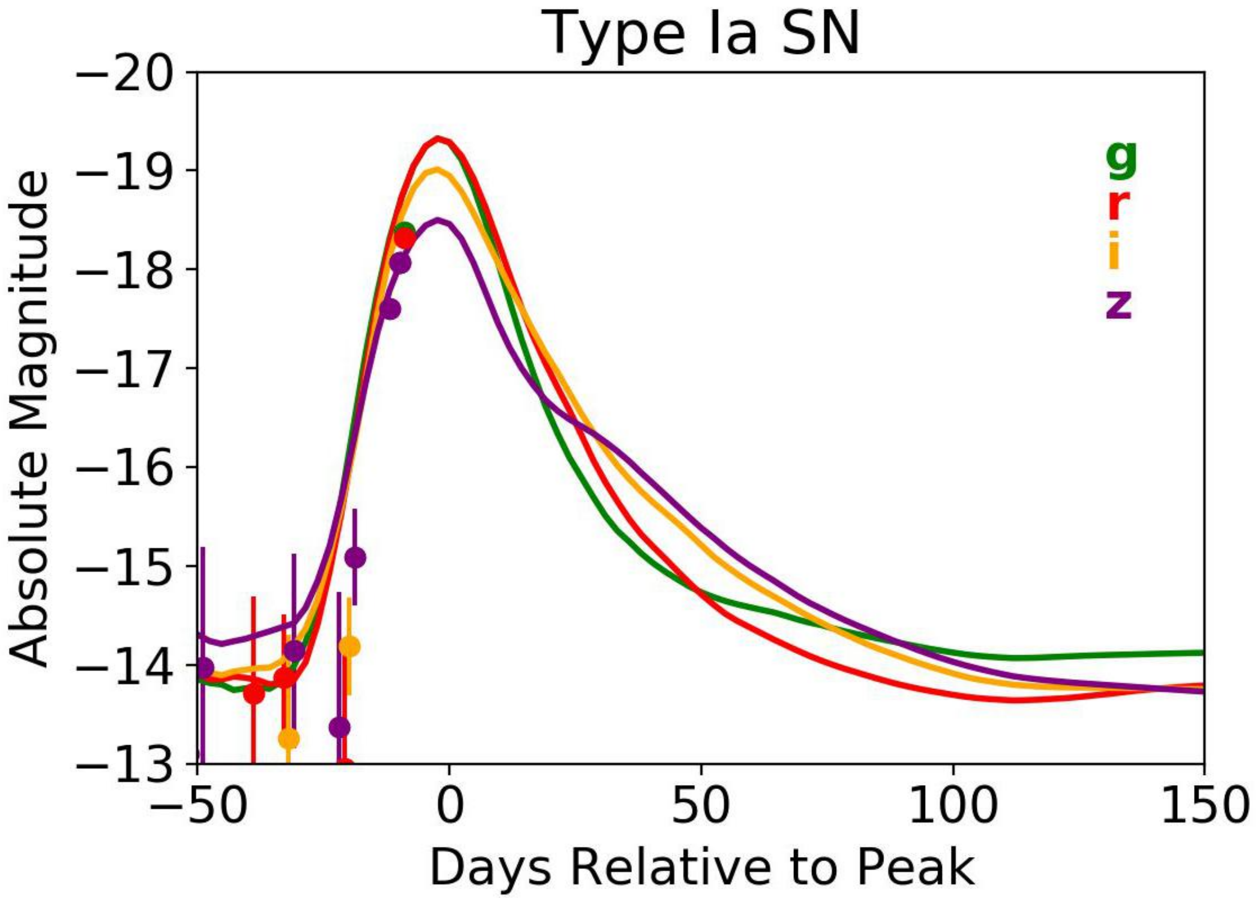
Decoded light curve updated with new data



VAE estimate is a little odd, thinks it is short and dim.

VAV+ 20,21

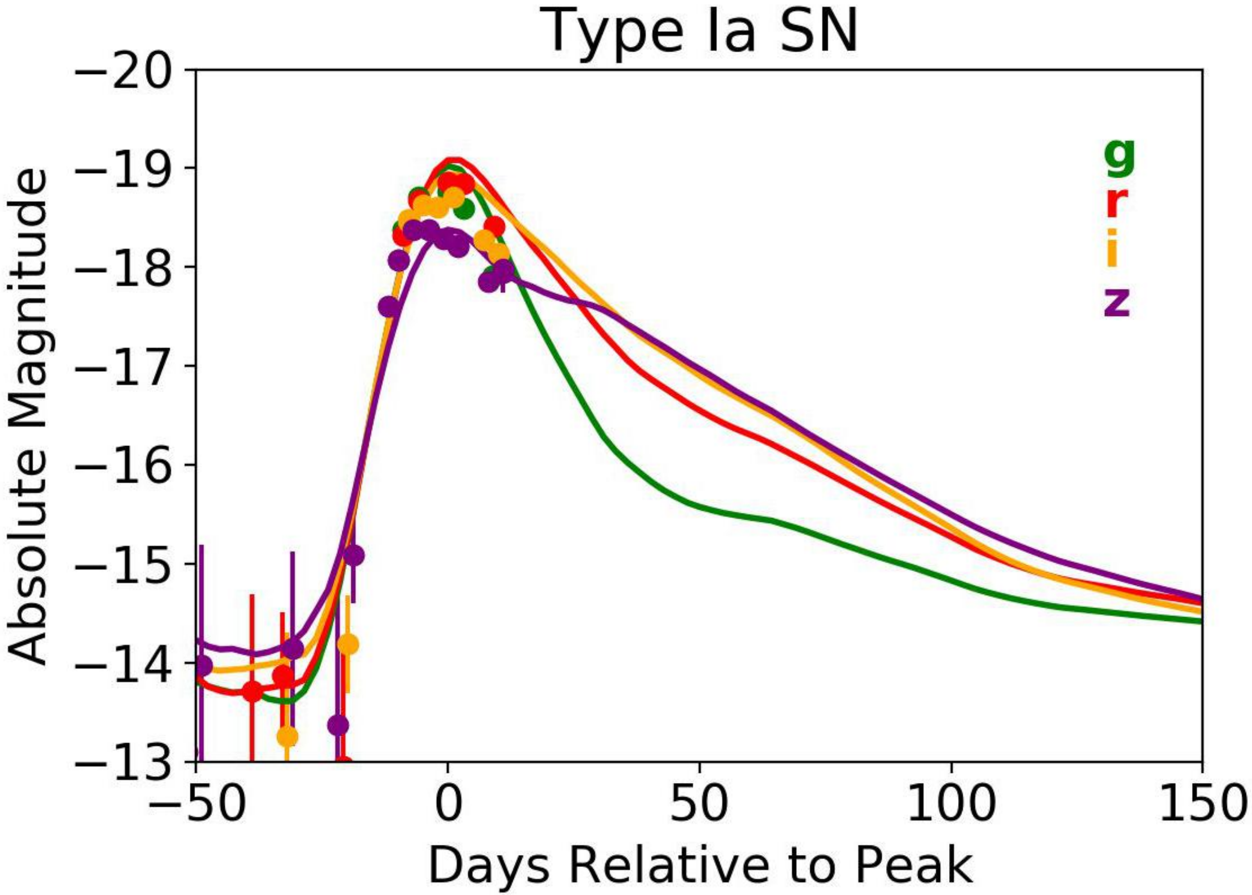
Decoded light curve updated with new data



VAE estimate hits the "correct" peak flux for this type of supernova

VAV+ 20,21

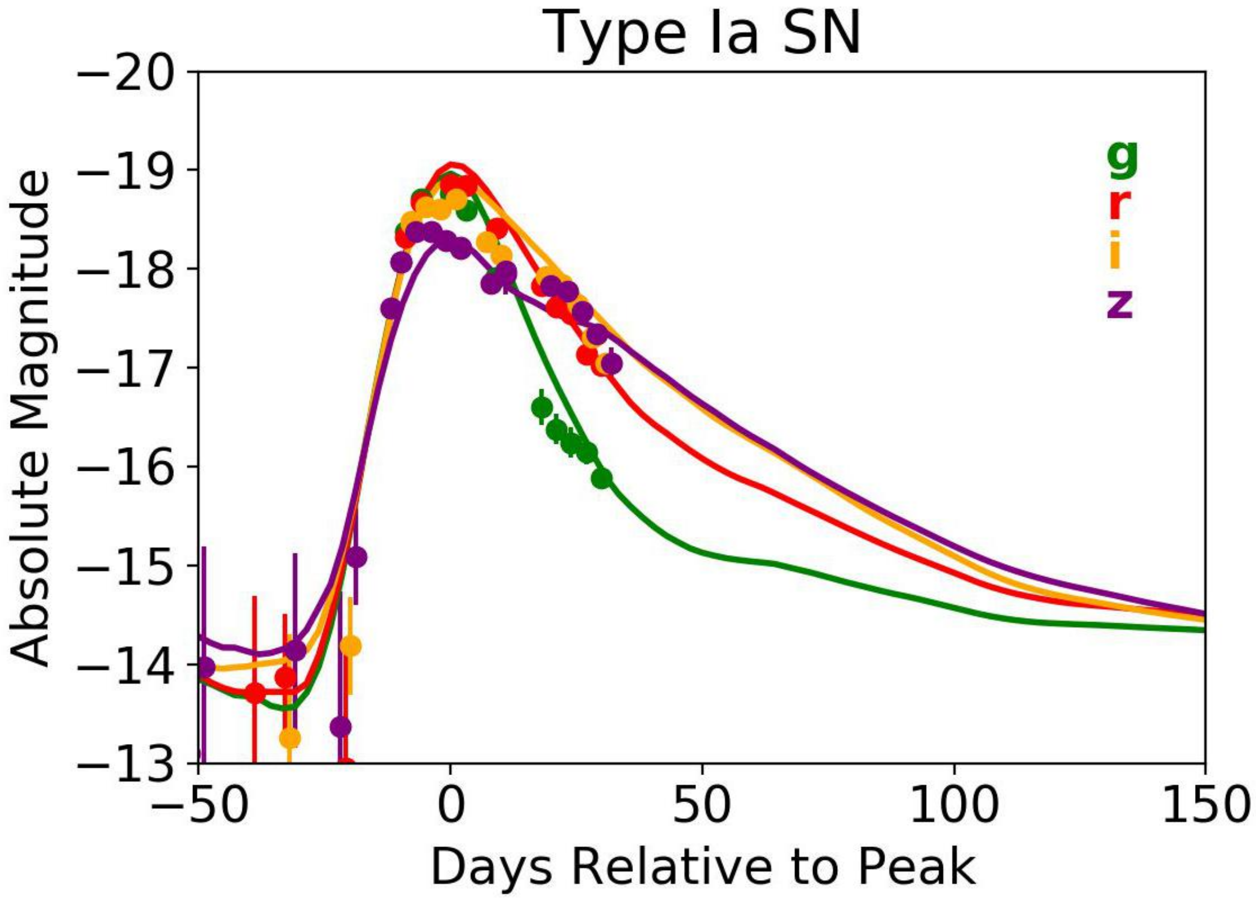
Decoded light curve updated with new data



VAE estimate correctly predicts the 'bump' in z-band (again a distinct feature for this supernova type)

VAV+ 20,21

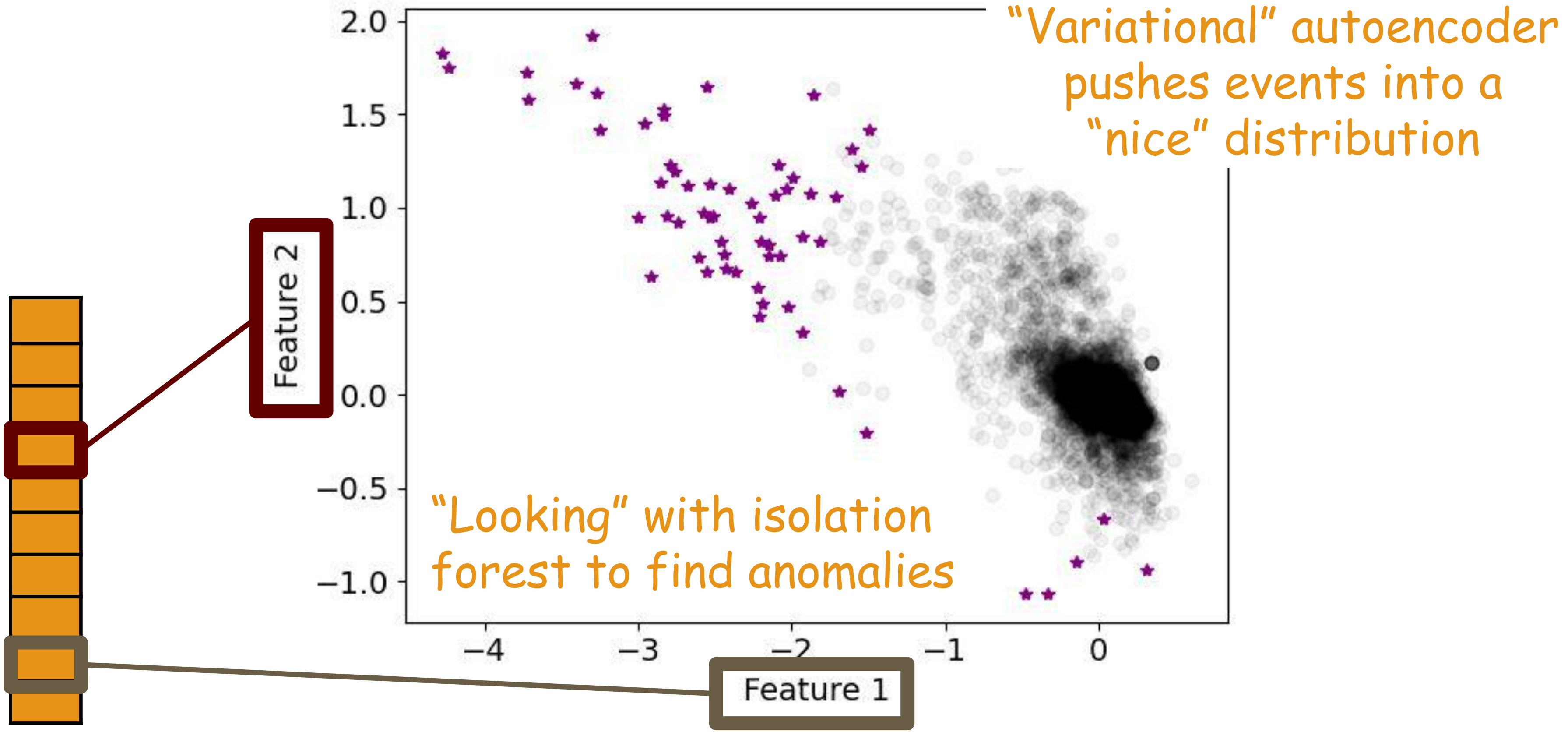
Decoded light curve updated with new data



Final estimate

VAV+ 20,21

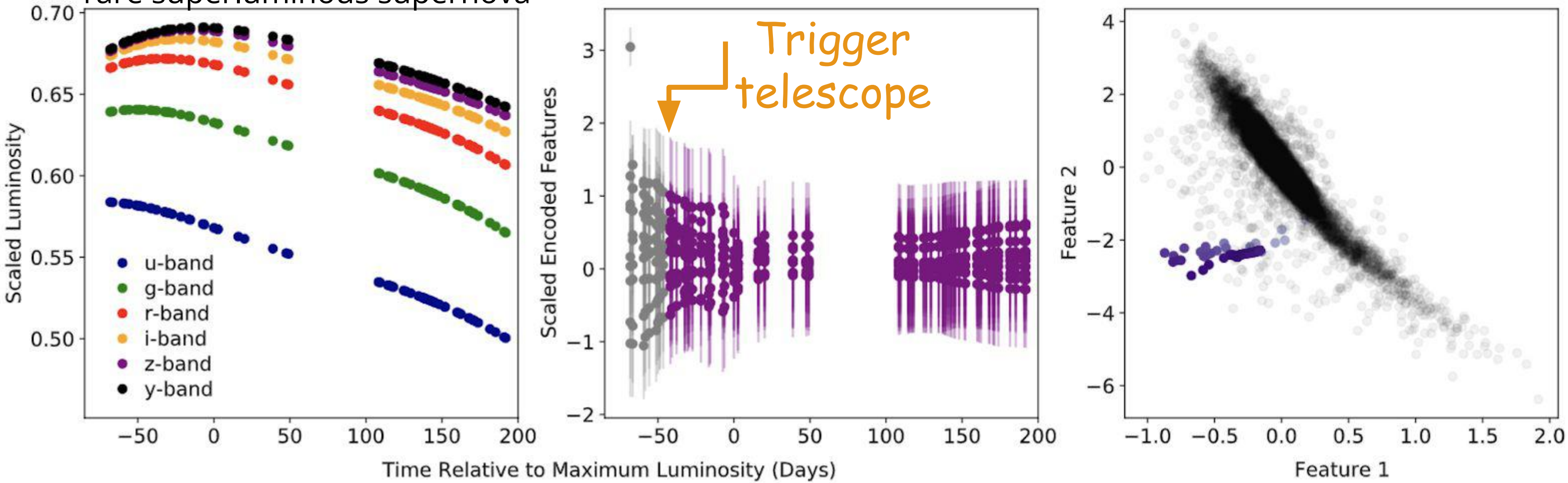
Look at the encoded space for "needles"



VAV+ 20,21

Look at encoded space as the event evolves!

rare superluminous supernova



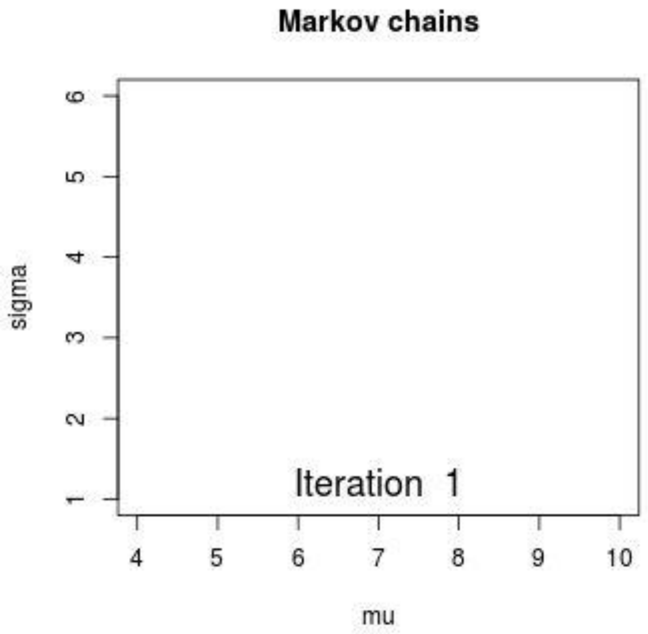
VAV+ 21

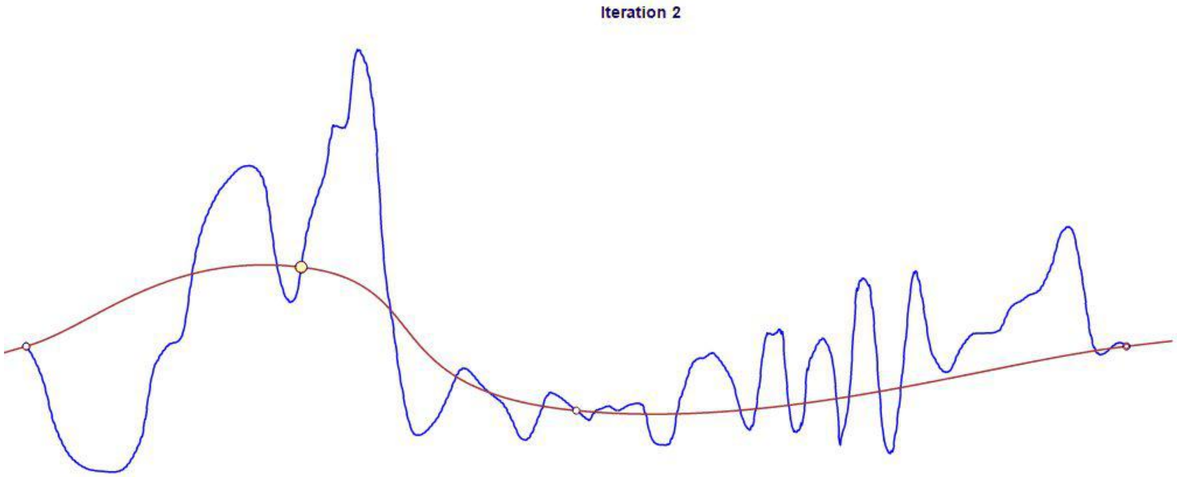
✓ Classify,

✓ Identify,

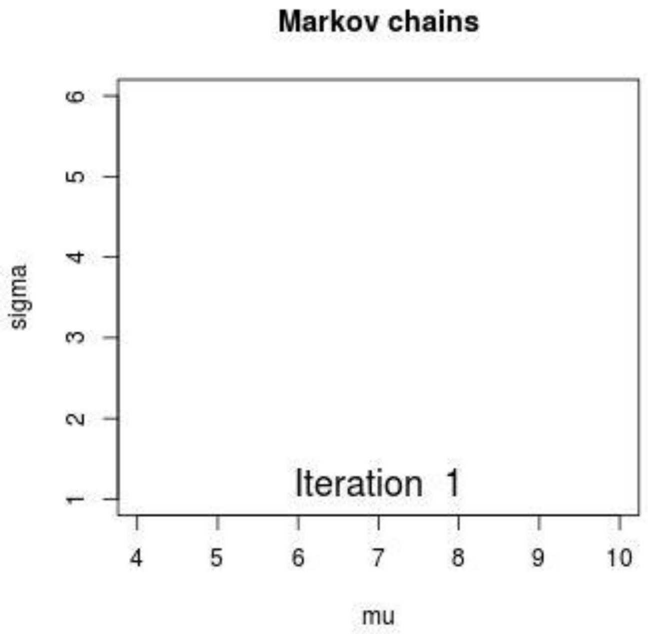
Analyze

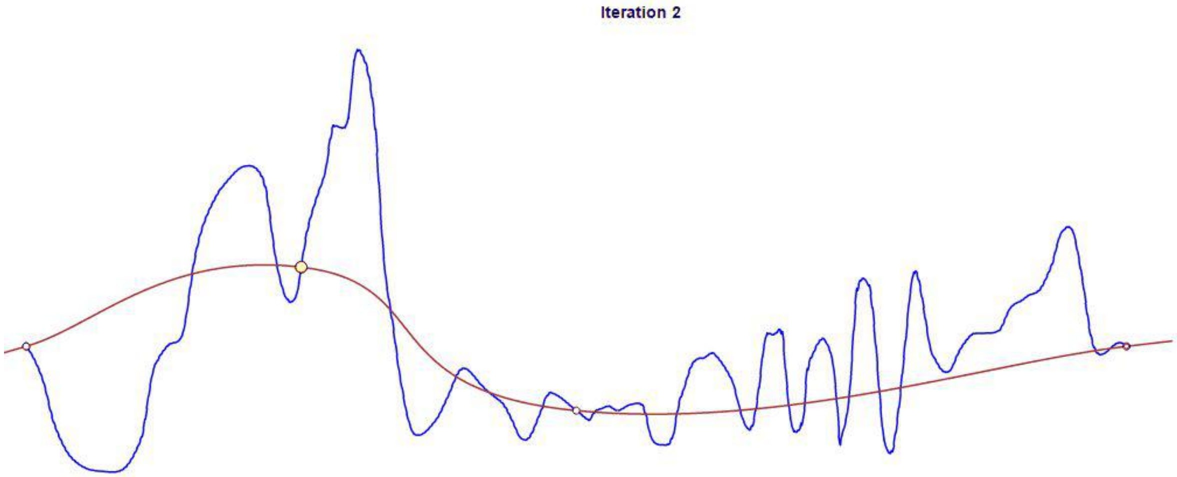
Traditional fitting takes ~10s of minutes to hours for one SN





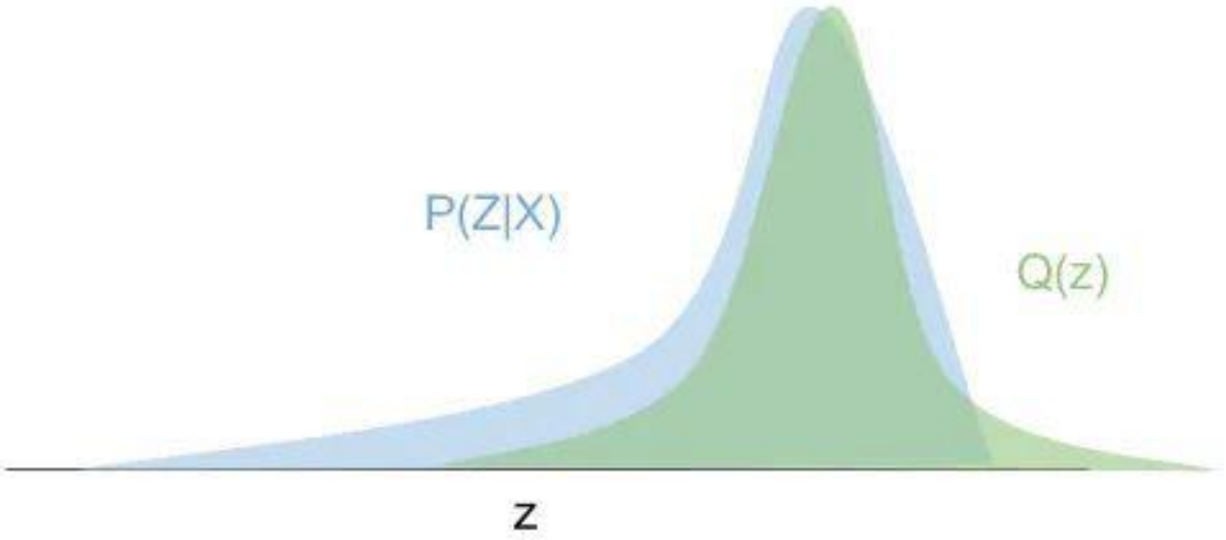
Traditional fitting takes ~10s of minutes to hours for one SN



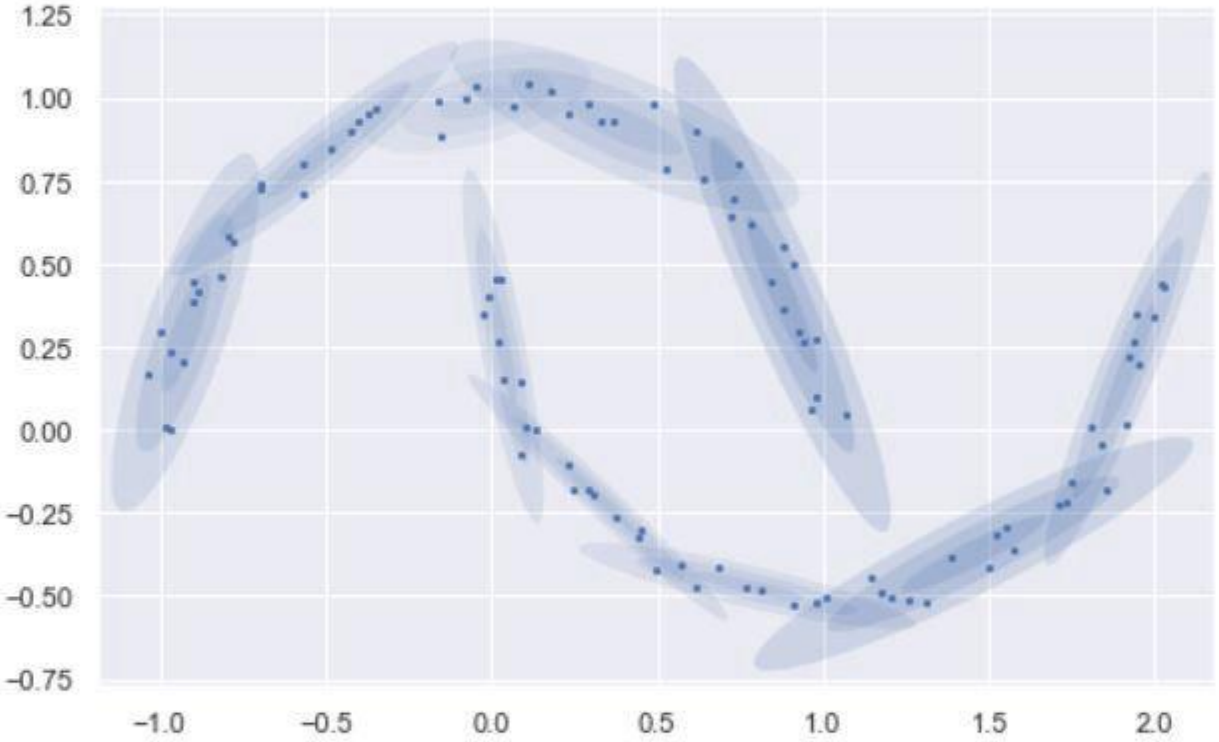


So the sample of 10 million SNe from Rubin will cost ~10 million CPU hours!

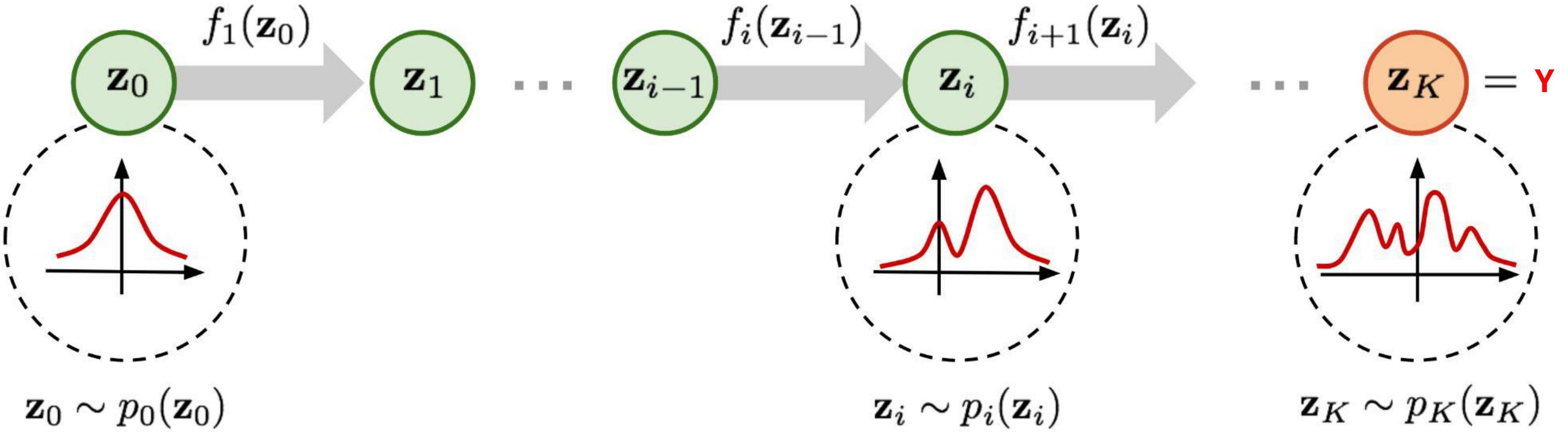
Replace traditional methods with variational inference

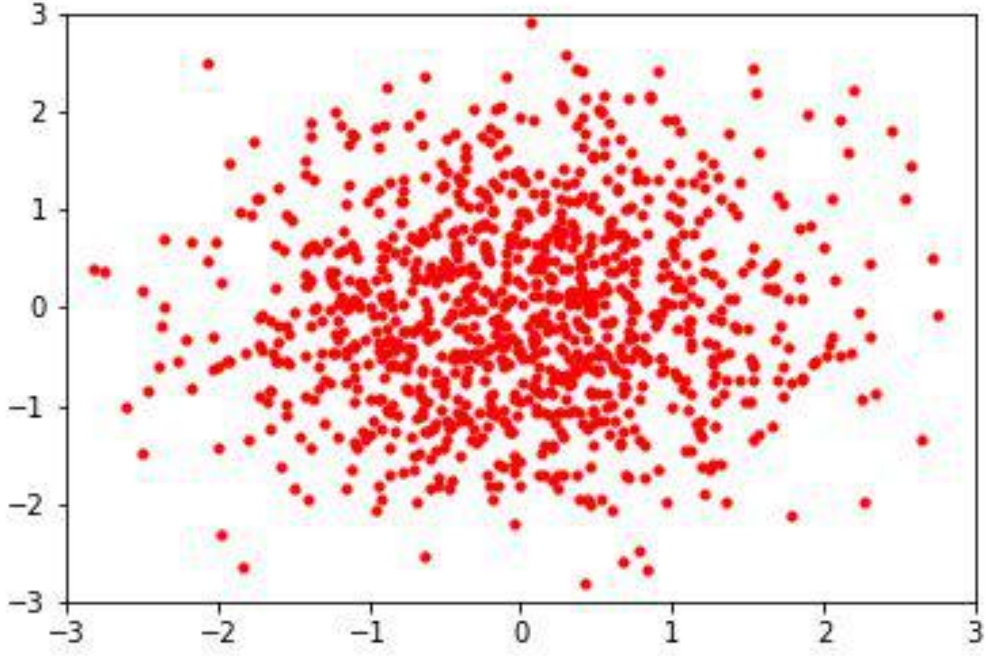


But if our samples have a complex distribution, it may take *many* Gaussians to estimate the density

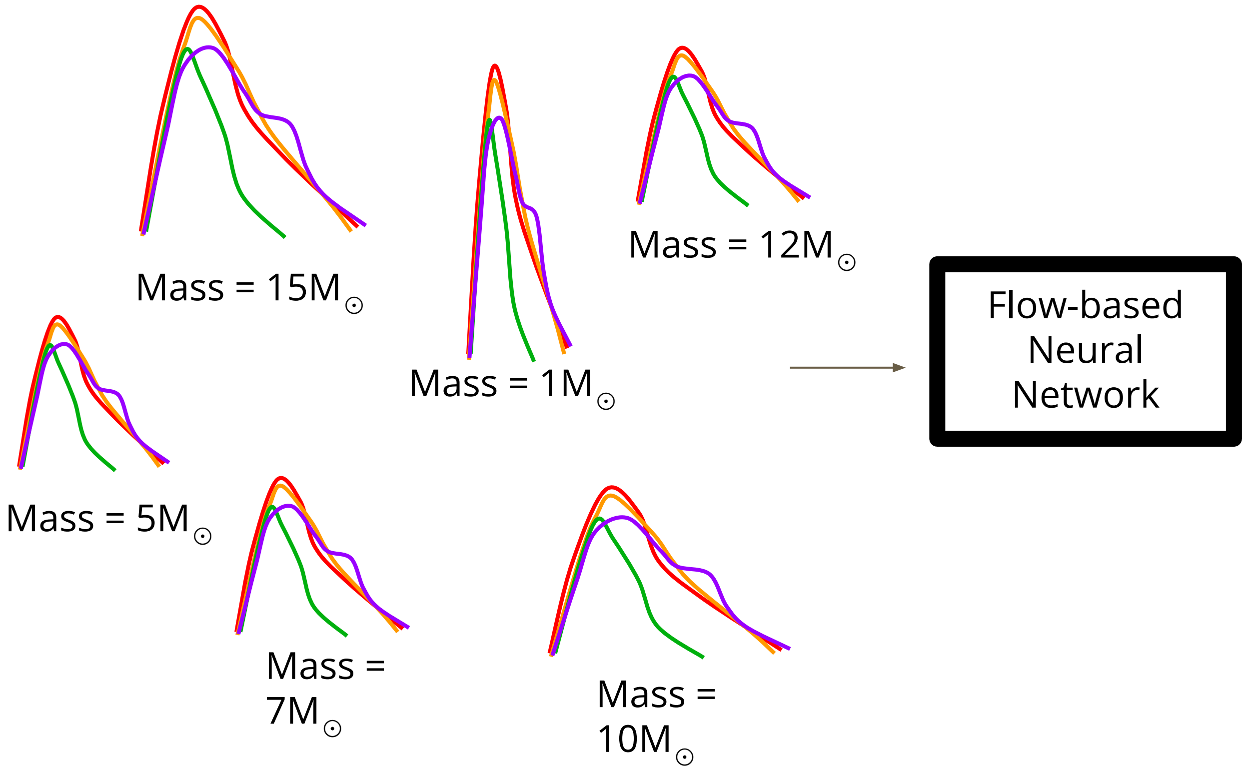


We are going to learn a (simple!) transformation to take a Gaussian to a complex distribution




3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

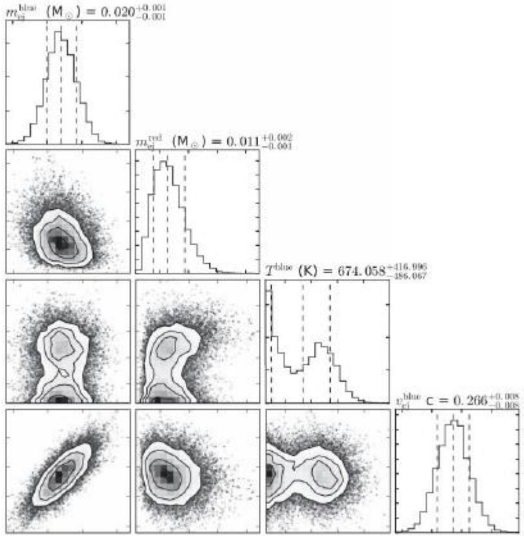
Simulation-based inference: Bypassing statistics via deep learning



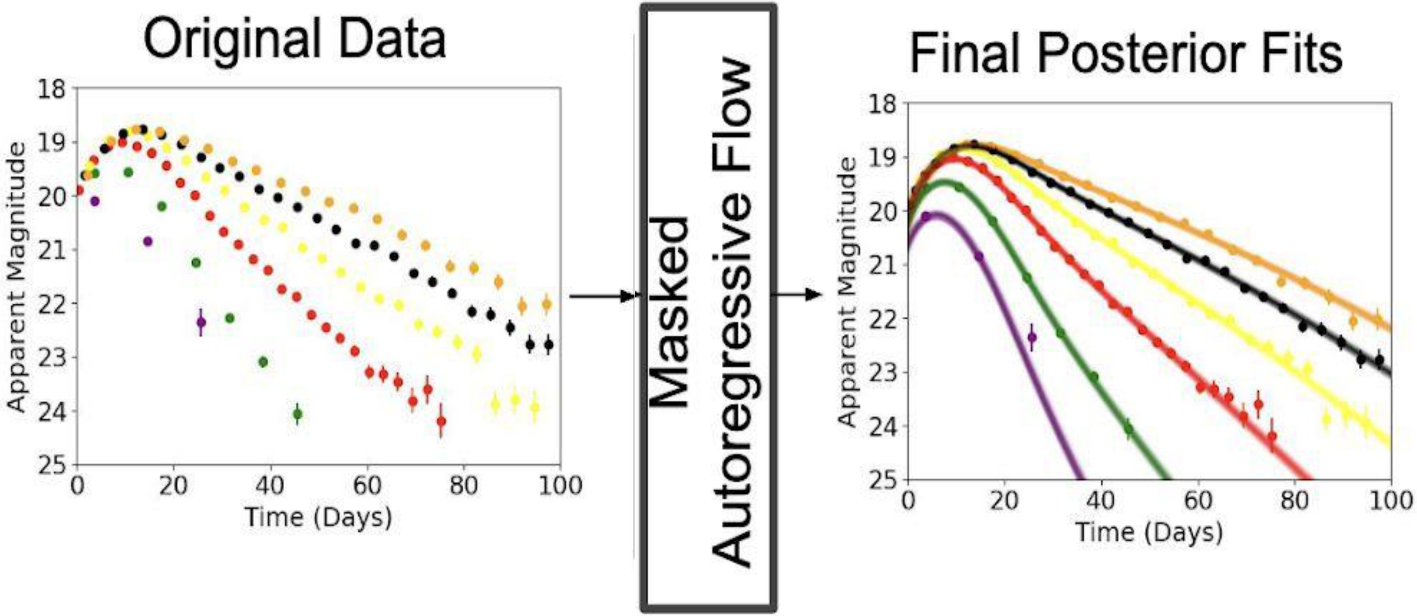
Simulation-based inference: Bypassing statistics via deep learning

Mass = $1M_\odot$

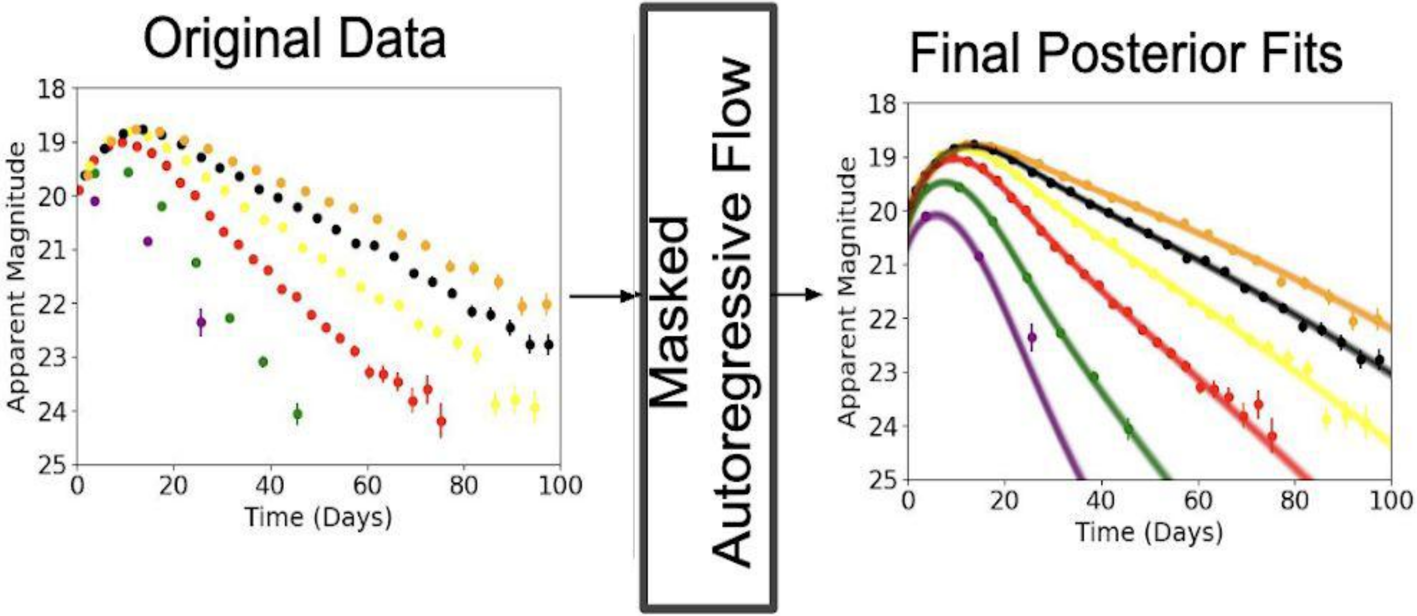
Flow-based
Neural
Network



New method takes 10ms per SN...

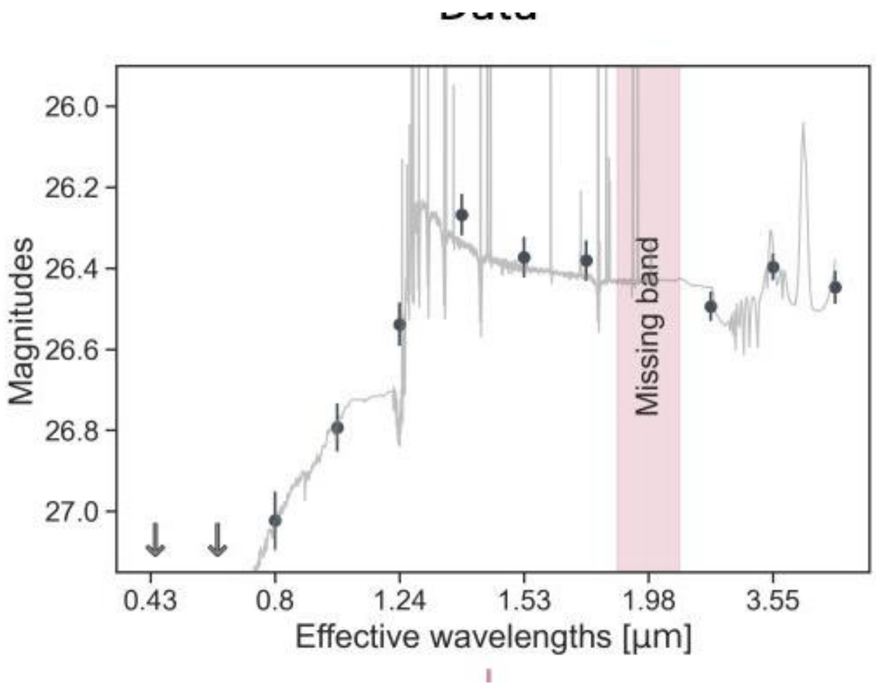


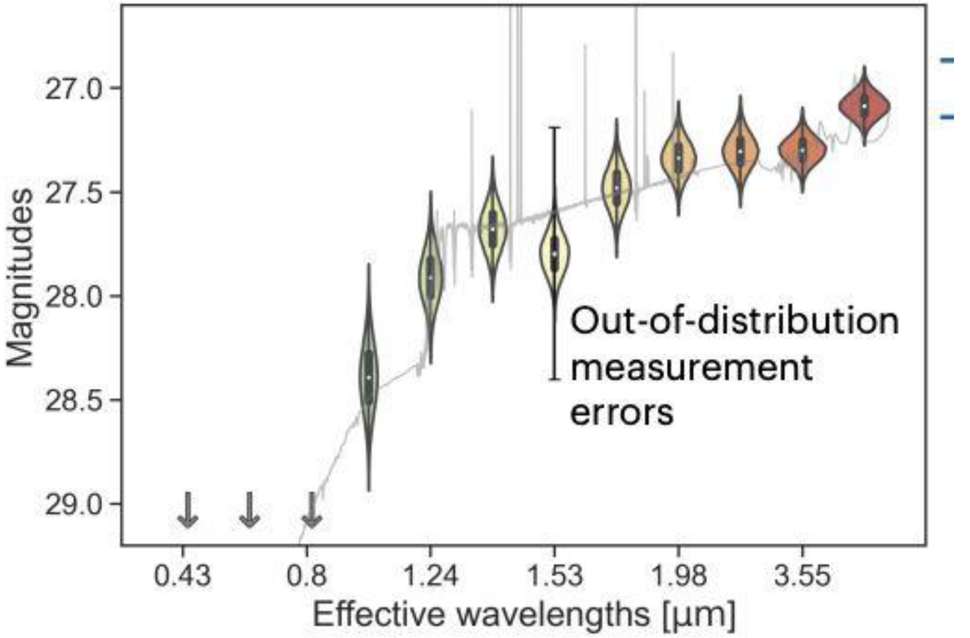
New method takes 10ms per SN… so about 1 day on a single CPU for the full set of Rubin SNe!



VAV 2022

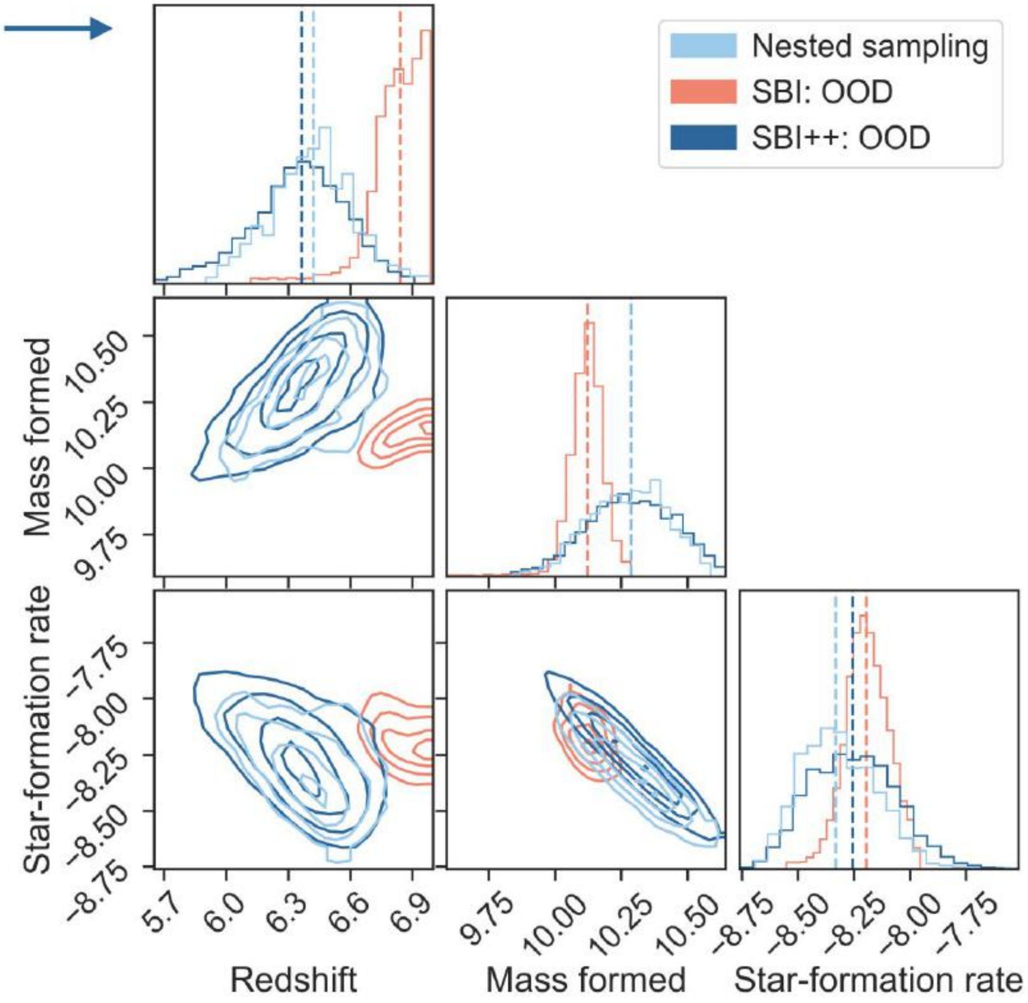
But real data is messy!





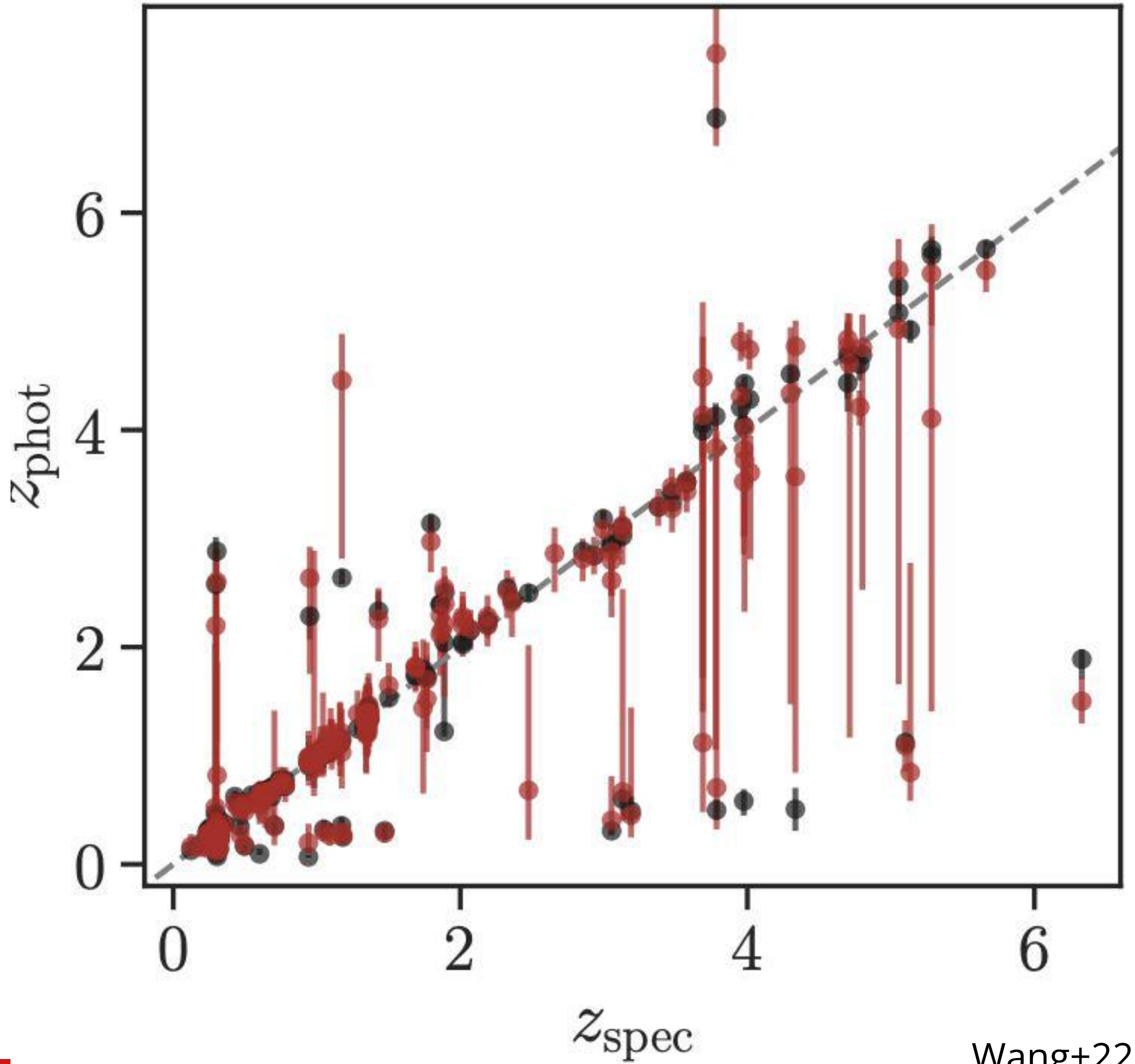
Wang+22

What if I have a poor understanding of the underlying noise?



Wang+22

SBI++ is (seemingly) better calibrated than standard nested sampling techniques in the literature!



Wang+22

Welcome to a new era for time-domain astrophysics!

- LSST will push our discovery rate of extragalactic transients to over 1 million objects per year
- By intertwining machine learning and our physical understanding of transients, we will be able to:
  - classify SNe into known classes
  - identify needles (new, exciting physics) in real time
  - fully analyze the haystack at a computationally reasonable cost

Thank you!