Recurrent VAE for anomaly detection in supernova time series

V. Ashley Villar
Simons Society Junior Fellow,
Columbia University/Flatiron Institute (CCA)

With thanks to CCA group: S. Ho, M. Cranmer, G. Contardo, J. Lin
How does the zoo of observed transients connect with the underlying (astro)physics?

Soares-Santos+2017

HST

NASA

Soares-Santos+2017

J. Guillochon
Extragalactic transients, like *supernovae*, are visible for days to months.
Bulk properties, like the black body temperature/radius

“Fingerprints” of elements
Supernova light curves encode physics
A real light curve

Type Ia Supernova

sparse, noisy, irregularly-sampled
Spectra of just 0.1% of supernovae!
The Rubin Needles & the Haystack

~Million supernovae / year

~1000s / year with traditional (spectroscopic) classification

~100 supernovae we actively follow with other resources
The Rubin Needles & the Haystack

~Million supernovae / year

~1000s / year with traditional (spectroscopic) classification

~100 supernovae we actively follow with other resources

The rarest needles: the unknown unknowns
A data-driven, unsupervised anomaly detection method using a variational recurrent autoencoder
Measured / Estimated $ugrizY$ Fluxes, Errors and Time

Encoded Layer

Repeated Encoded Layer with Concatenated Phases

Output Light Curve
Recurrent neurons

**RNN**

**LSTM**

**GRU**
Decoded light curve updated with new data

Type Ia SN

VAE estimate is a little odd, thinks it is short and dim.
Decoded light curve updated with new data

Type Ia SN

VAE estimate hits the “correct” peak flux for this type of supernova
Decoded light curve updated with new data

Type Ia SN

VAE estimate correctly predicts the ‘bump’ in z-band (again a distinct feature for this supernova type)
Decoded light curve updated with new data

Type Ia SN

Absolute Magnitude

Days Relative to Peak

Final estimate
Step 3: Isolation forest to look for outliers

[Image: https://donghwa-kim.github.io/iforest.html]
Look at the encoded space for “needles”

Variational AE pushes events into a “nice” distribution

“Looking” with isolation forest to find anomalies
Look at encoded space as the event evolves!

In progress: what is a useful metric for anomaly detection vs. time?
Our problem: Need to identify unknown ‘needles’ (anomalies) in a ‘haystack’ of transient events in real time.

Solution:
- Step 1: Pre-process (interpolate) time series with a GP
- Step 2: Train RVAE on the data to learn an encoded space
- Step 3: Search for anomalies in the encoding space using an isolation forest

Thank you!