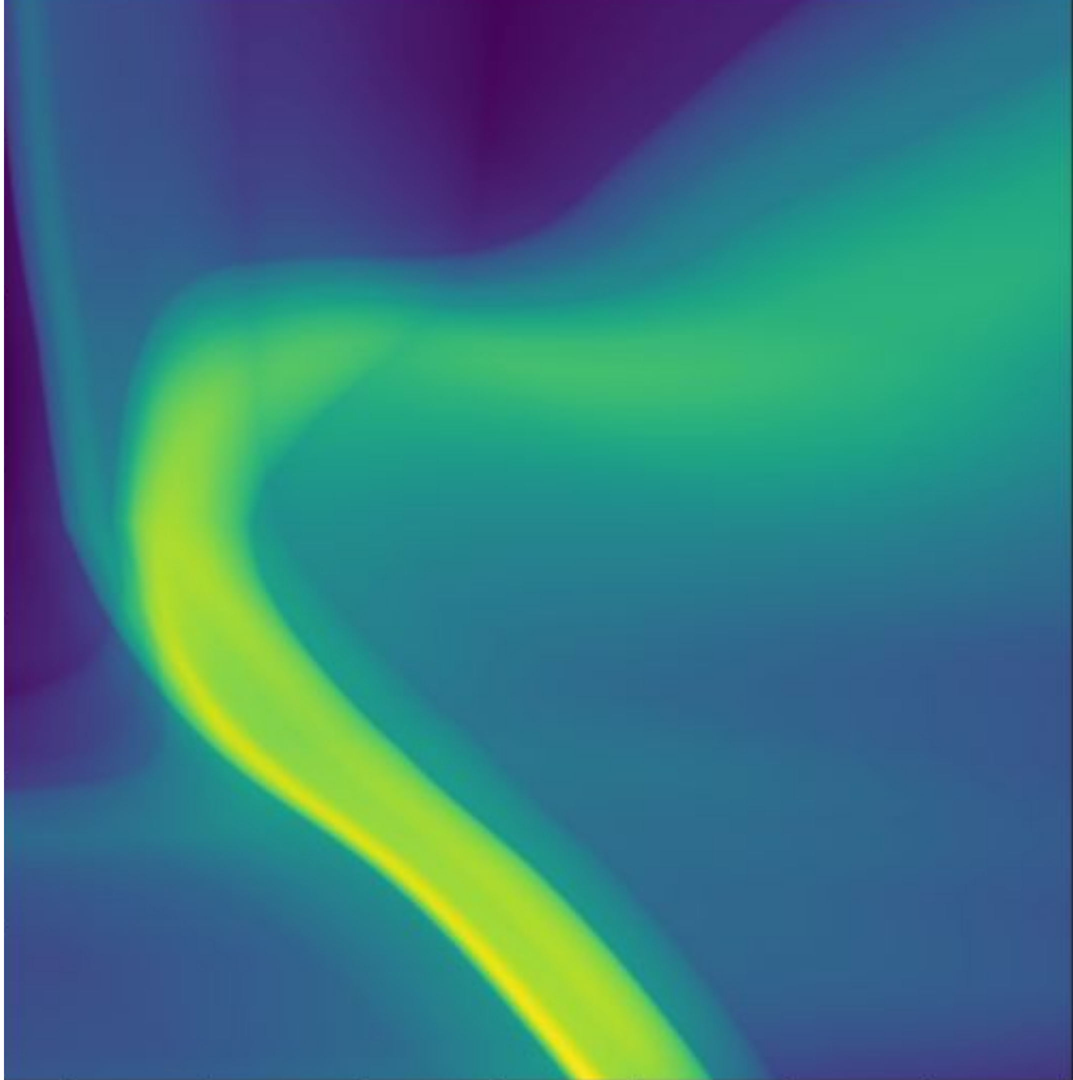
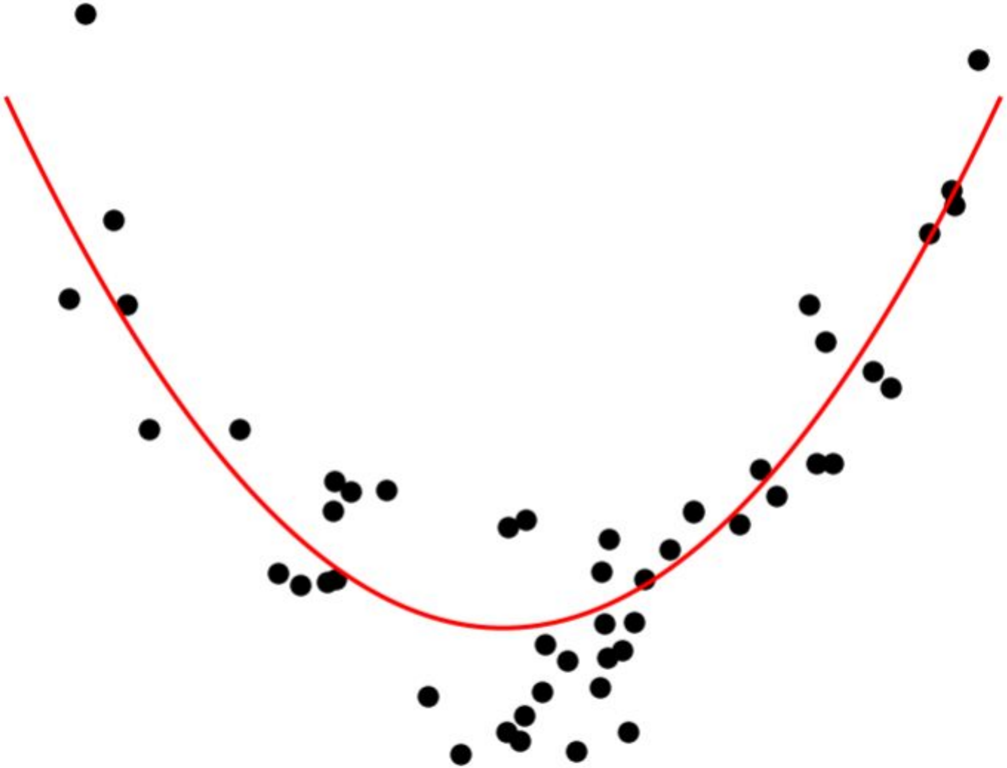
Machine Learning in Astronomy



Kyle Boone
University of Washington
SOMACHINE, April 20, 2021

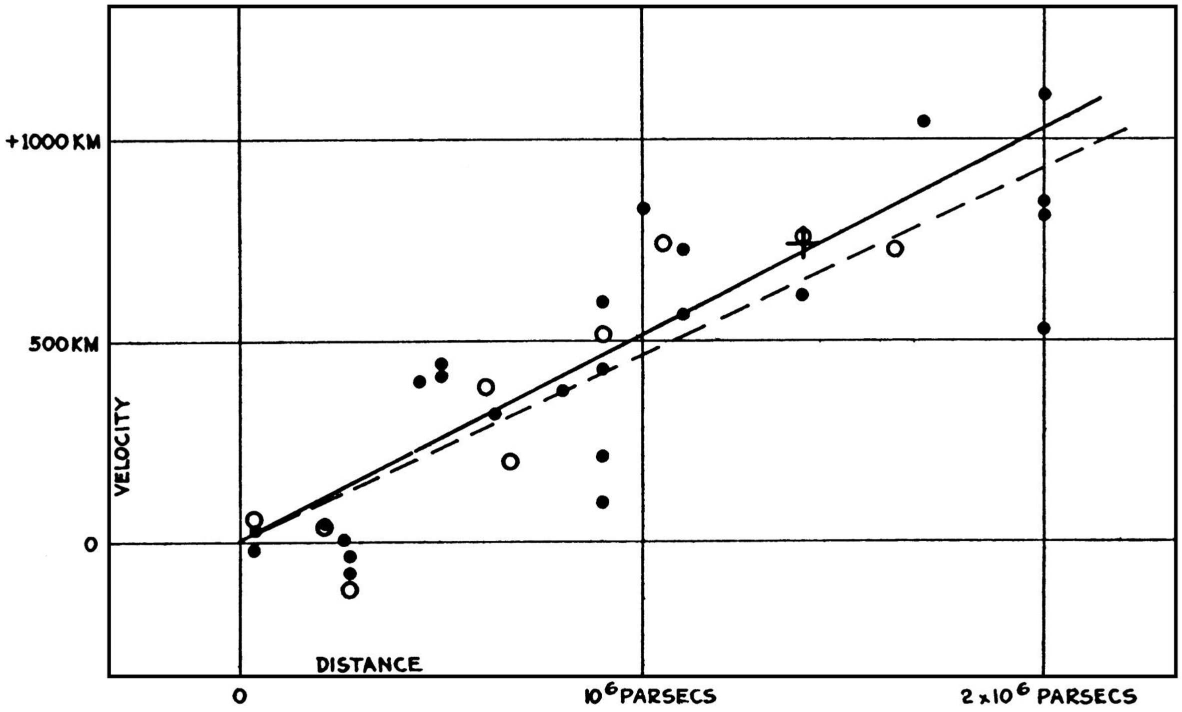
1. Regression

Learn a model that maps some inputs to a continuous output.



Linear Regression

Hubble 1929



e.g. the Hubble Diagram

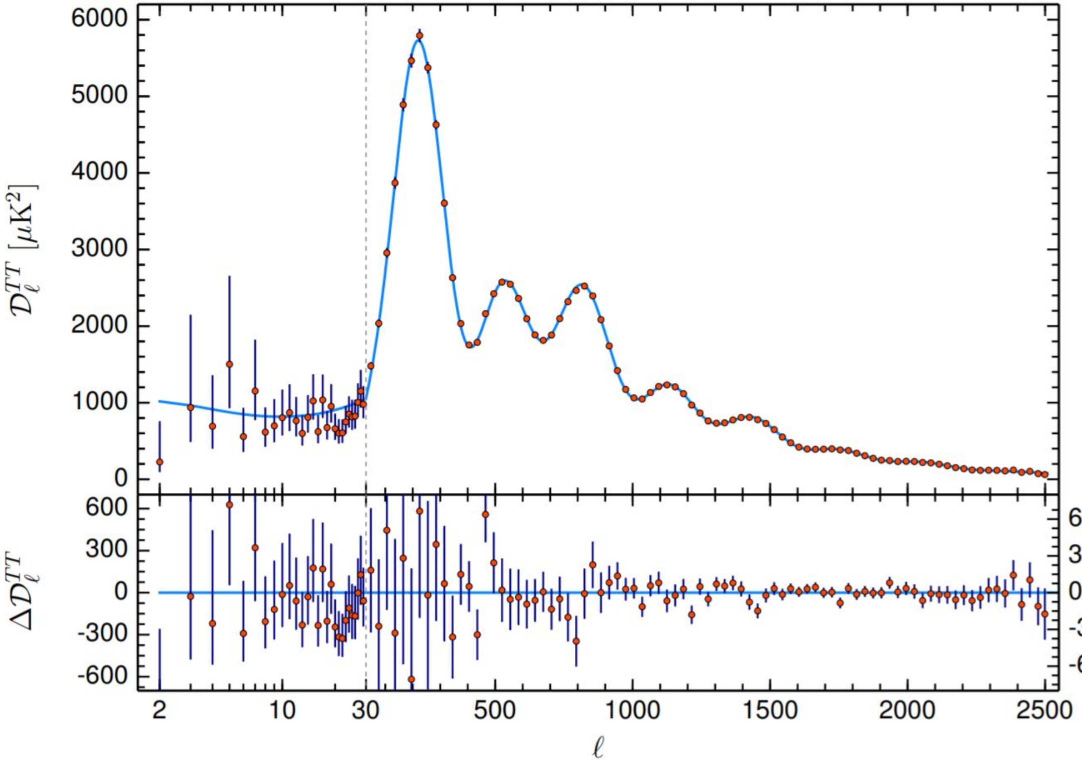
$$v = Hd$$

Regression

Planck Collaboration 2018

e.g. fitting the cosmic microwave background.

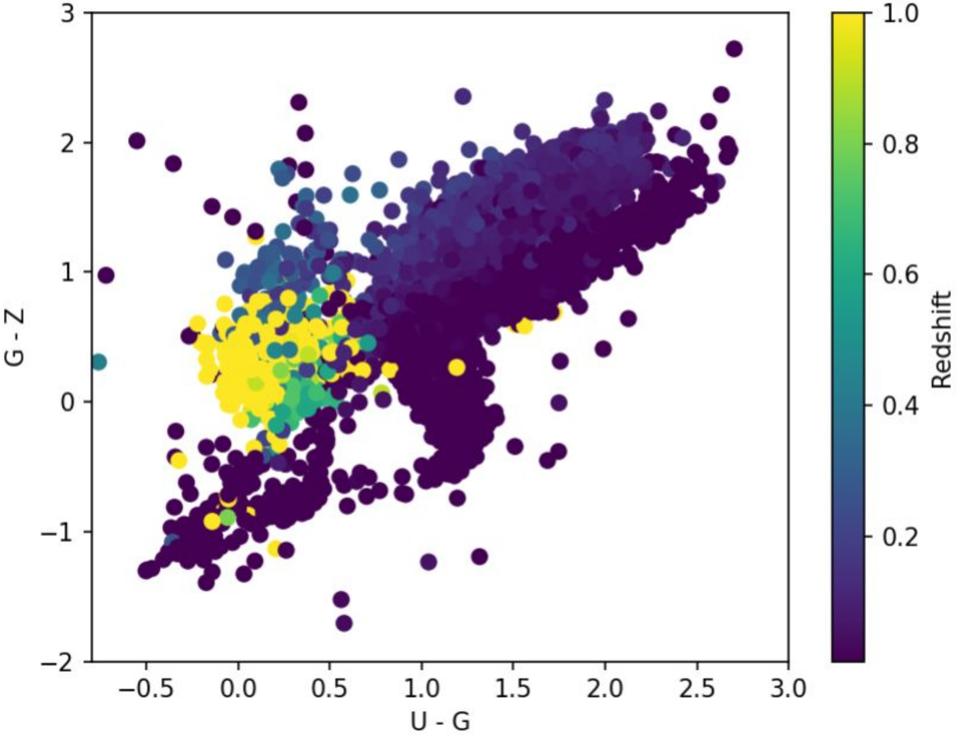
$$\mathcal{D}_l^{TT} = f(H_0, \Omega_m, \sigma_8, ...)$$

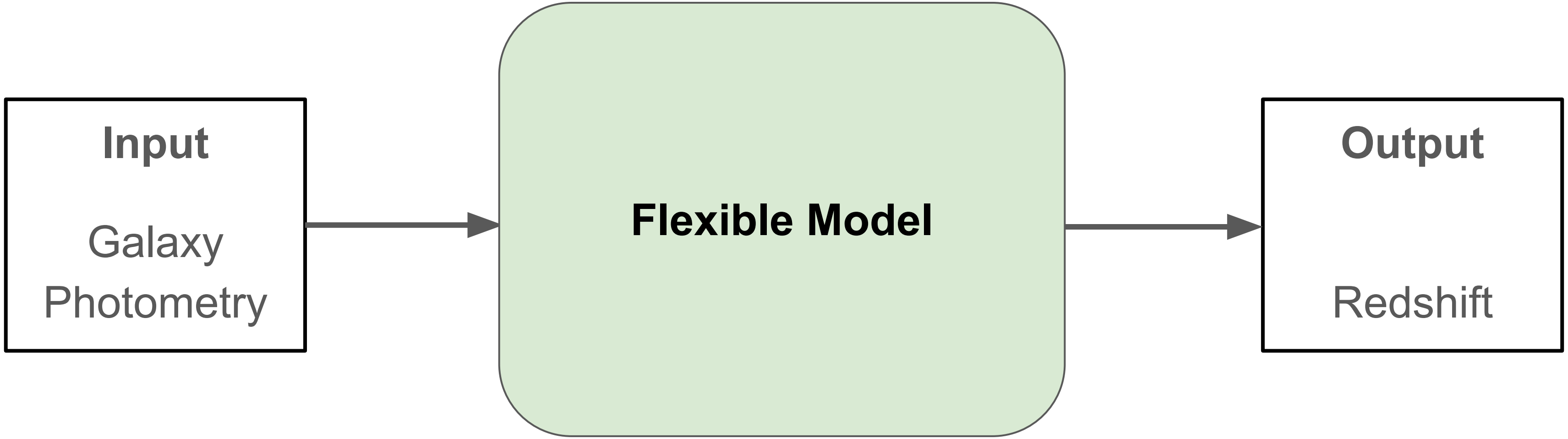


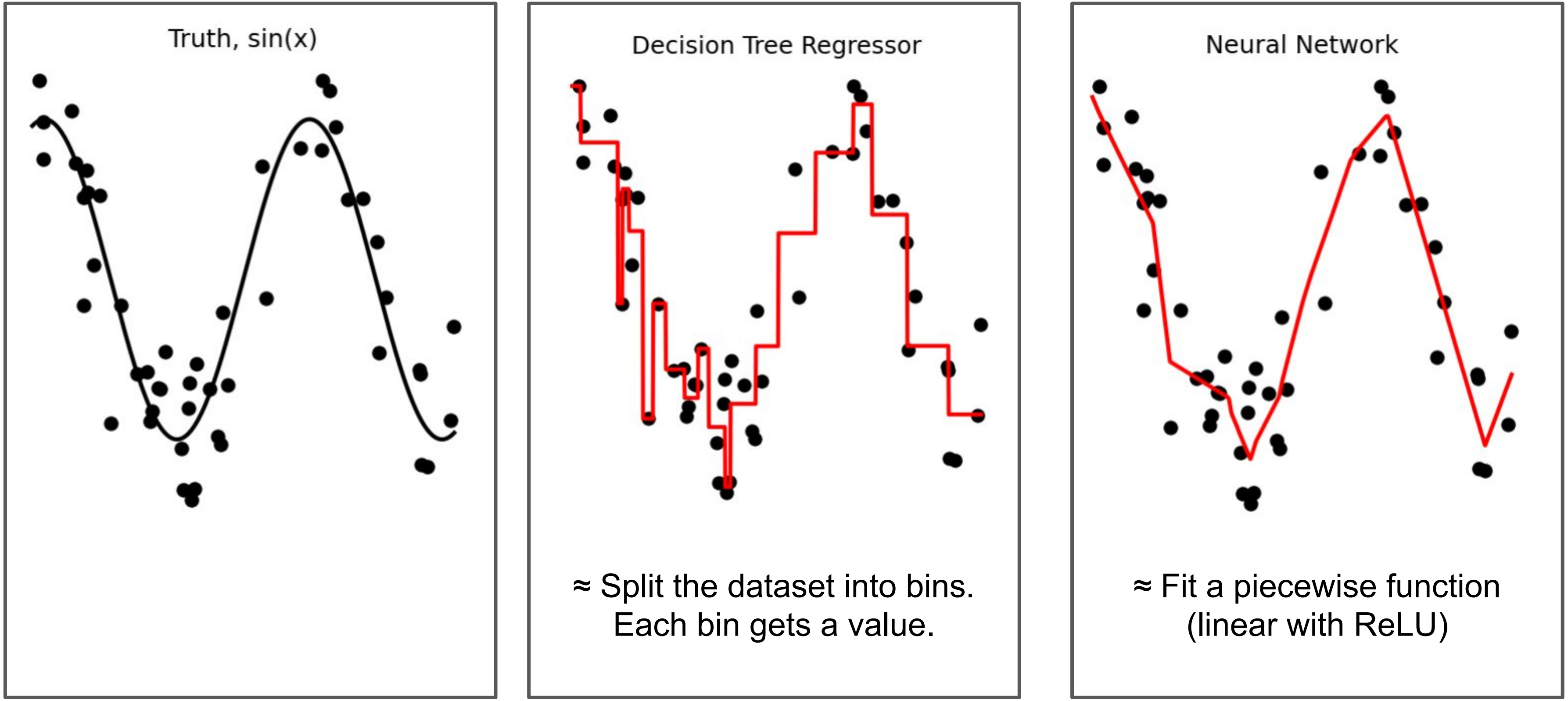
Photometric redshifts

Can we predict the redshift of a galaxy from its photometry?

What functional form should we assume?



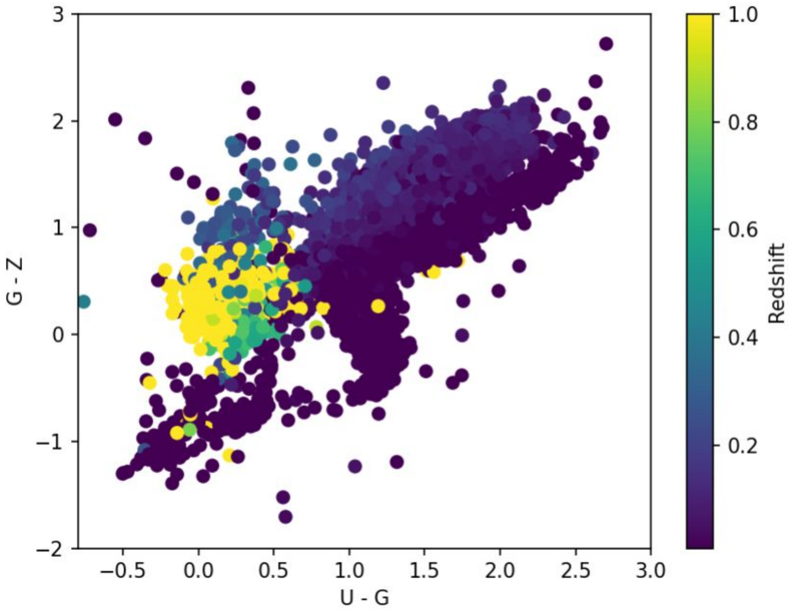

Input
Galaxy
Photometry
Flexible Model
Output
Redshift



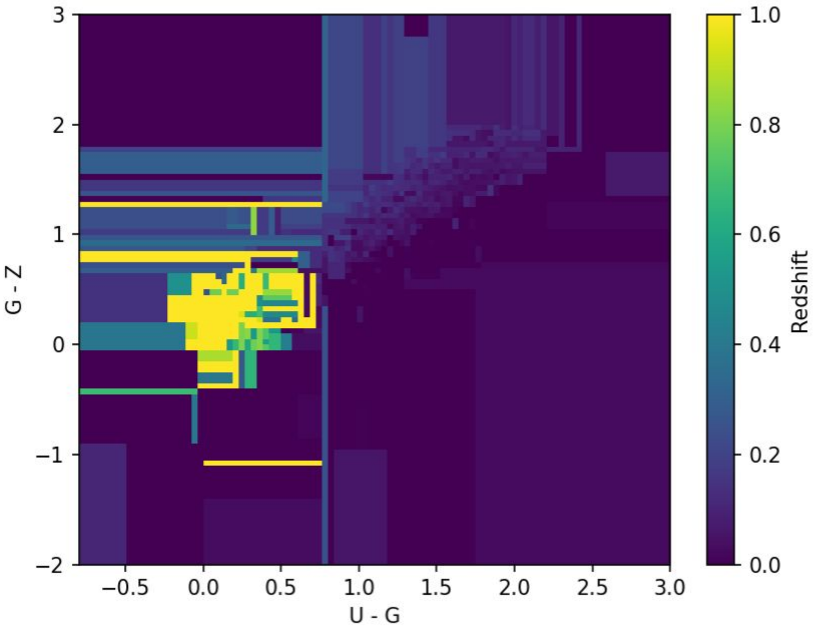
With enough parameters, we can approximate any function!

Photometric redshifts

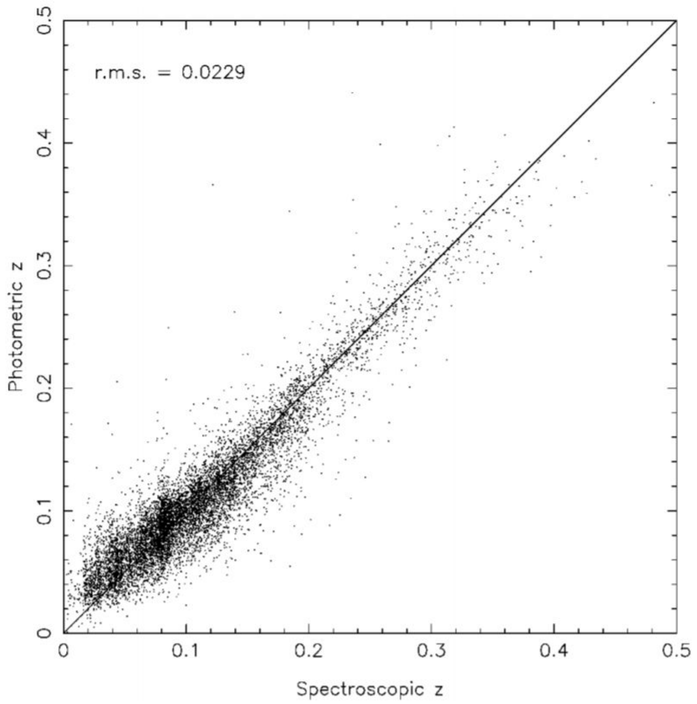
Training data



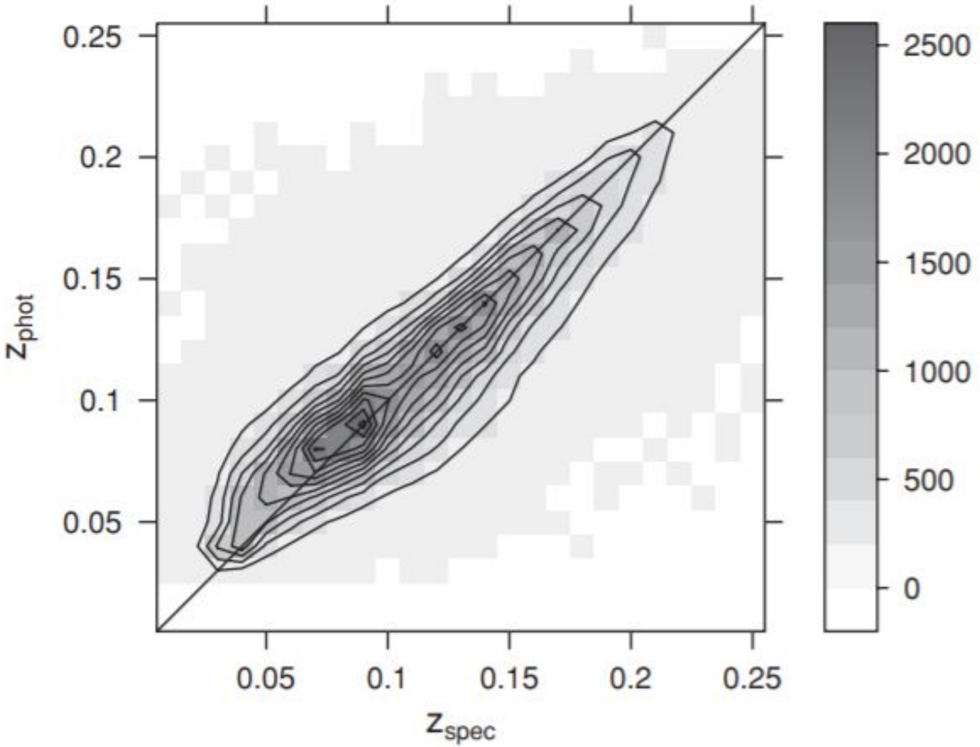
Decision tree prediction



ANNz, Collister & Lahav 2004



Random Forests, Carliles et al. 2010

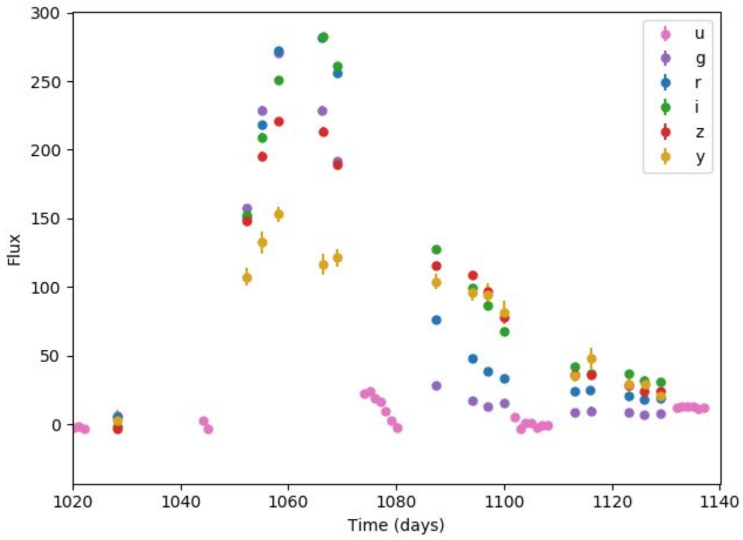


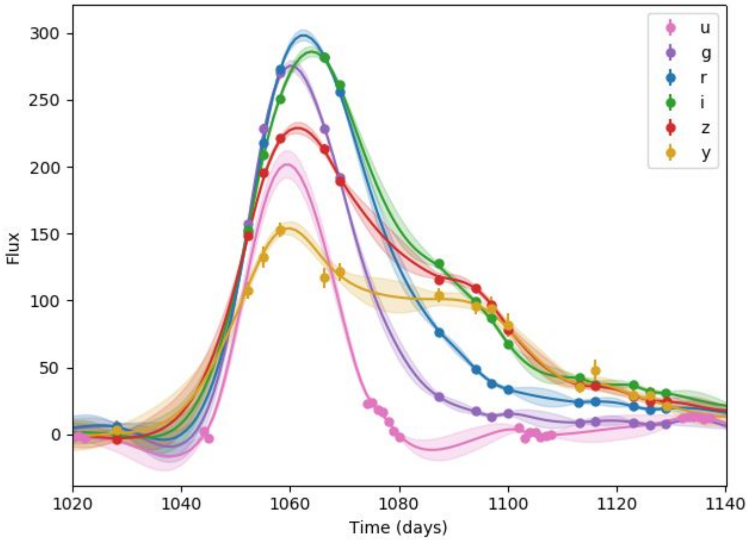
For problems like this, typically get similar performance for different algorithms.

Training set is very important.

Gaussian Process Regression

Boone 2019





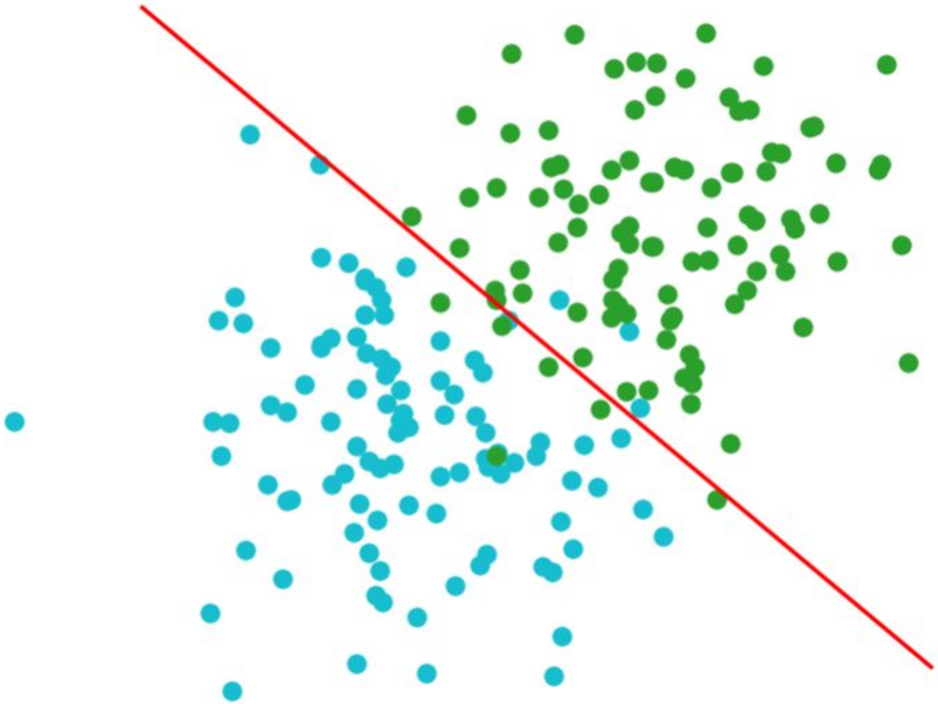
Produces smooth models with uncertainties from discrete data.

Very useful for timeseries analysis.

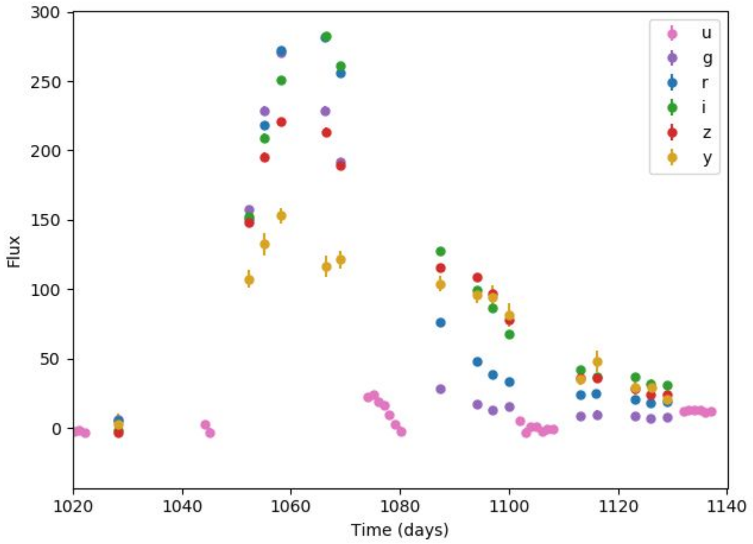
2. Classification

Predict which class an observation belongs to.

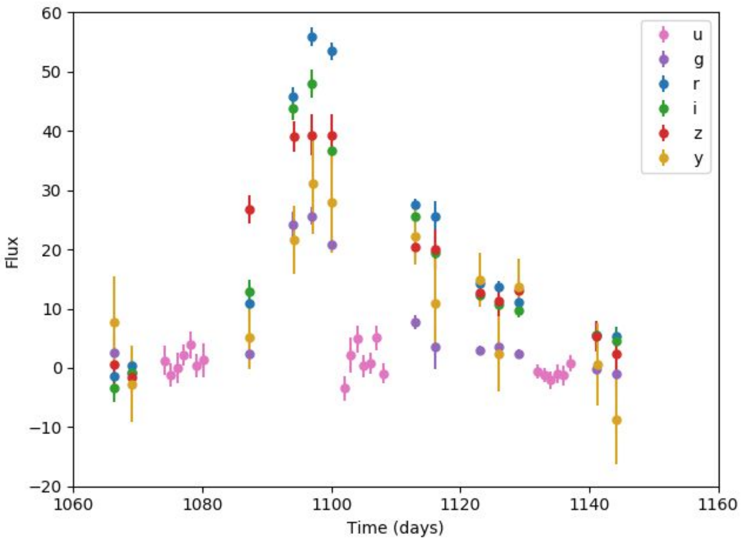
Equivalent to regression with discrete outputs.



Light curve classification



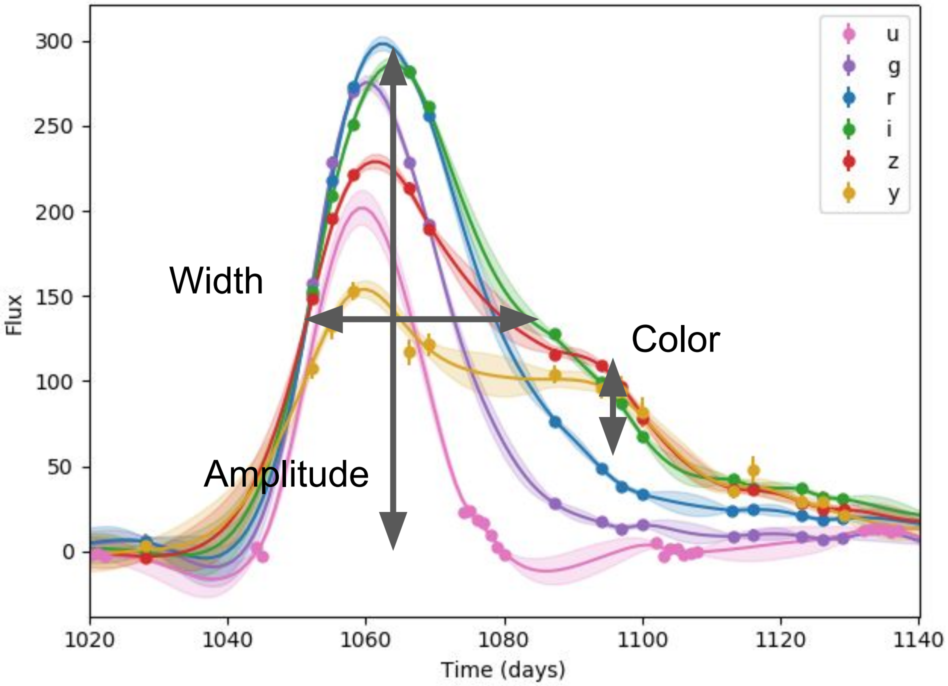
Type Ia supernova lightcurve



Type II supernova lightcurve

How do we train a model to tell these apart?

Feature extraction



Extract a number of expert-chosen features.

Use these as inputs to some ML algorithm.

Lots of different approaches!

Performance mostly depends on the dataset.

Reference	Light Curve fit	Dimensionality Reduction	Classification algorithm	Use redshift	Source Code
Poznanski *et al*, 2006	---	---	Template fit	Yes	pSNiD II
Newling *et al*, 2010	parametric	parameters from fit	Kernel Density Estimation Boosting	Yes	No
Richards *et al*, 2011	spline	diffusion maps	Random Forest	Yes	No
Karpenka *et al*, 2012	parametric	parameters from fit	Neural Network	No	No
Ishida & de Souza, 2013	spline	kernel PCA	Nearest Neighbor	No	github
Mislis *et al*, 2015	---	descriptive statistics	Random Forest	No	No
Varughese *et al*, 2015	spline	Wavelets	Nearest Neighbor Support Vector Machine	No	No
Hernitschek *et al*, 2016	χ^2	---	Random Forest	No	No
Lochner *et al*, 2016	parametric Gaussian Process	Wavelets PCA Model Fit	Naive Bayes Nearest Neighbor Support Vector Machine Boosted Decision Trees	No	No
Moller *et al*, 2016	parametric	parameters from fit	Boosted Decision Trees Random Forest	Yes	No
Charnok and Moss, 2017	---	---	Recurrent Neural Network	No	github
Mahabal *et al*, 2017	rate of change	---	Neural Network	No	No
Narayan *et al*, 2018	parametric Gaussian Process	Wavelets PCA	Random Forest	No	No
Revsbech *et al*, 2018	Gaussian Process	Diffusion Maps	Random Forest	Yes	github
Dai *et al*, 2018	parametric	parameters from fit	Random Forest	No	No

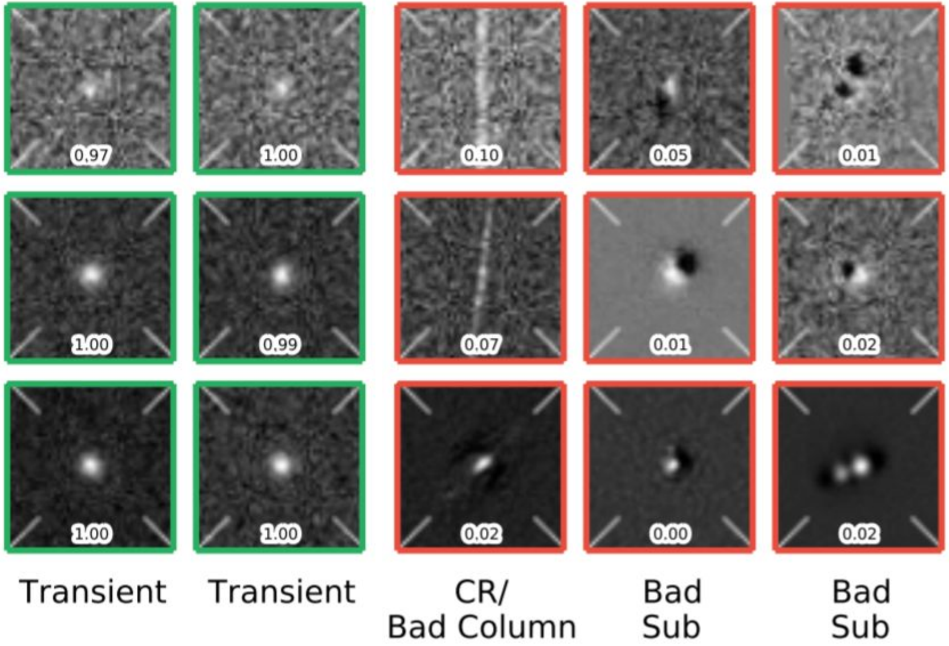
https://www.kaggle.com/michaelapers/the-plasticc-astronomy-classification-demo

Difference Imaging - Image Classification

Goldstein et al. 2015

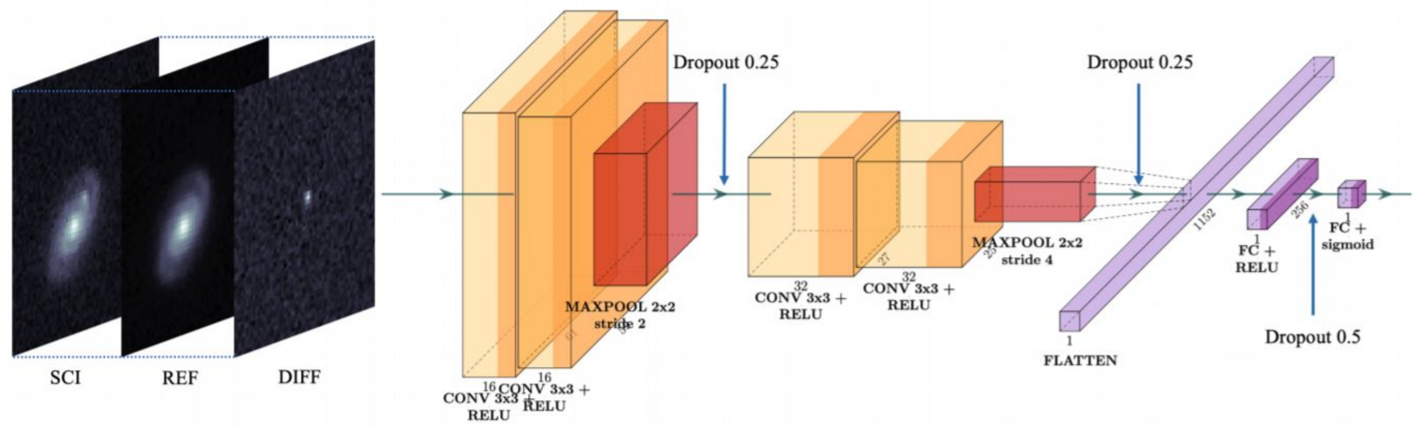
Use machine learning to identify transients/variables in difference images.

Previously work used feature extraction + decision trees.



Difference Imaging - Deep Learning

Duev et al. 2019



- Instead of manually selecting features, input the raw data.
- Requires a much more complex model (> millions of parameters).
- For images, convolutional neural networks work very well.

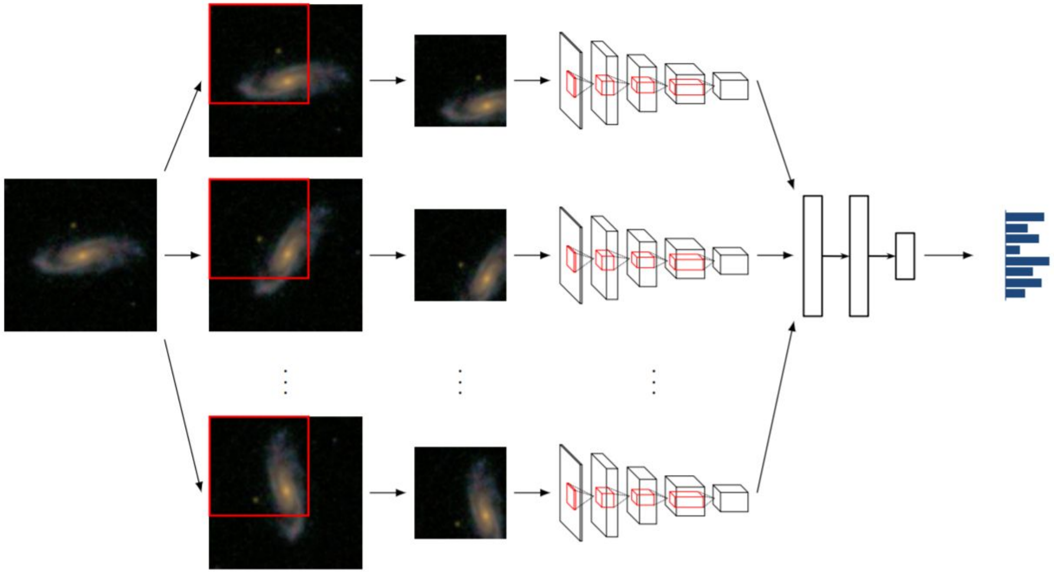
Galaxy morphology - Deep Learning

Dieleman et al. 2014

Input: galaxy cutout

Output: galaxy properties

See next talk

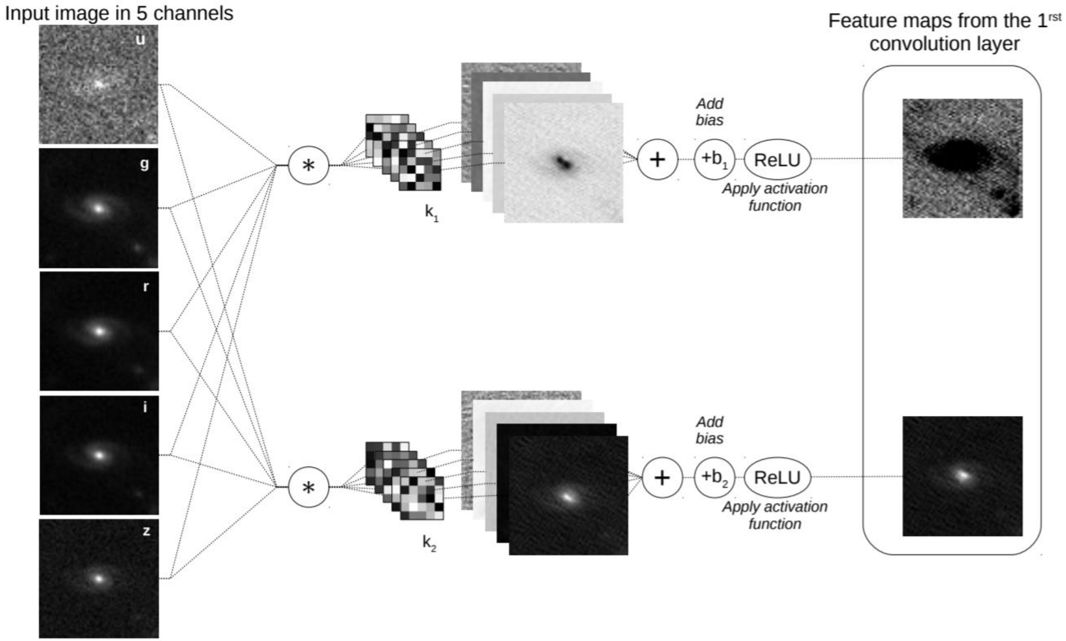


Photometric redshifts - Deep Learning

Pasquet et al. 2019

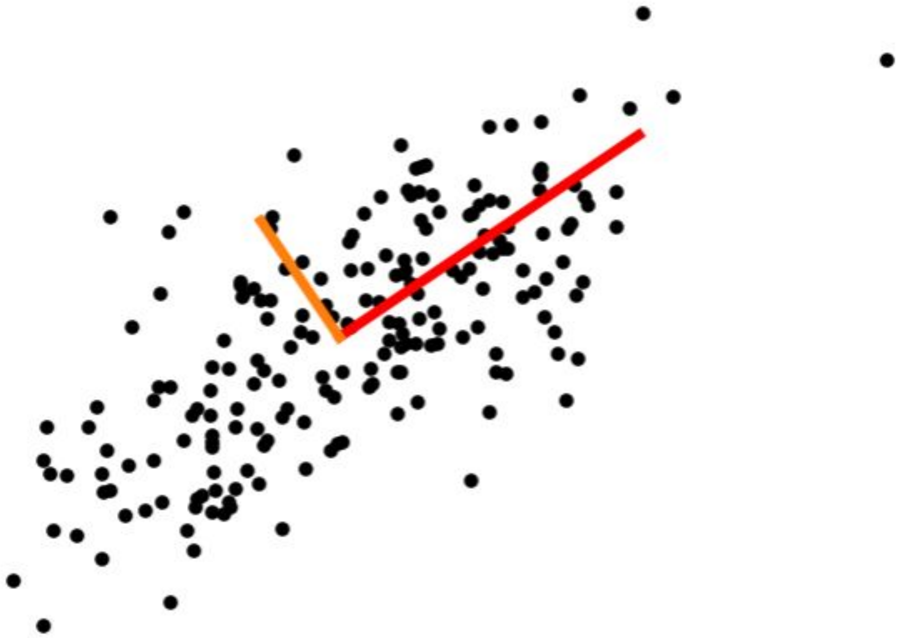
Previously only used photometry.

With deep learning, can use spatial information to improve photometric redshifts.

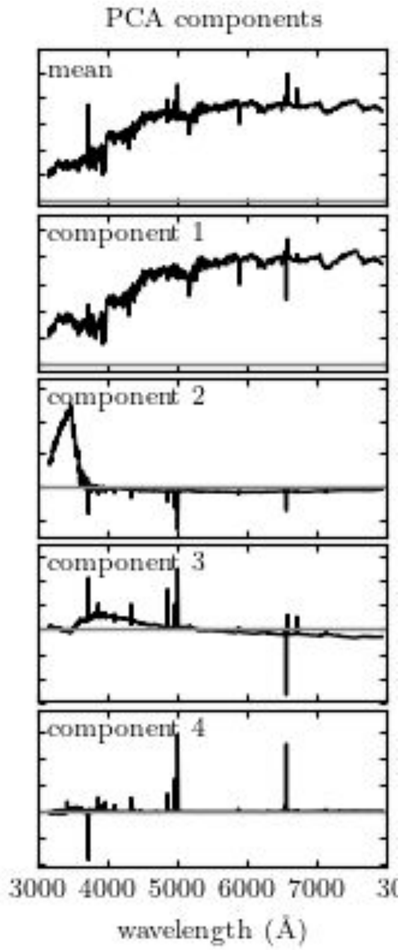


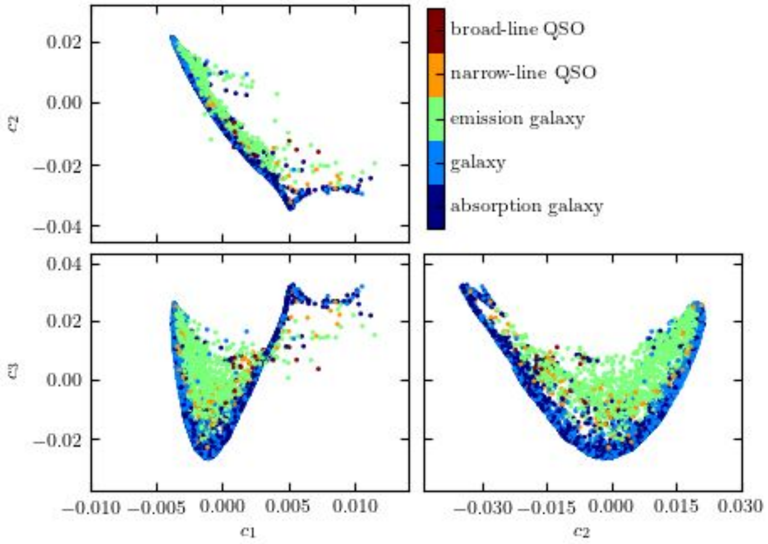
3. Unsupervised Learning

Finding patterns in unlabeled data.



Galaxy Spectra - PCA



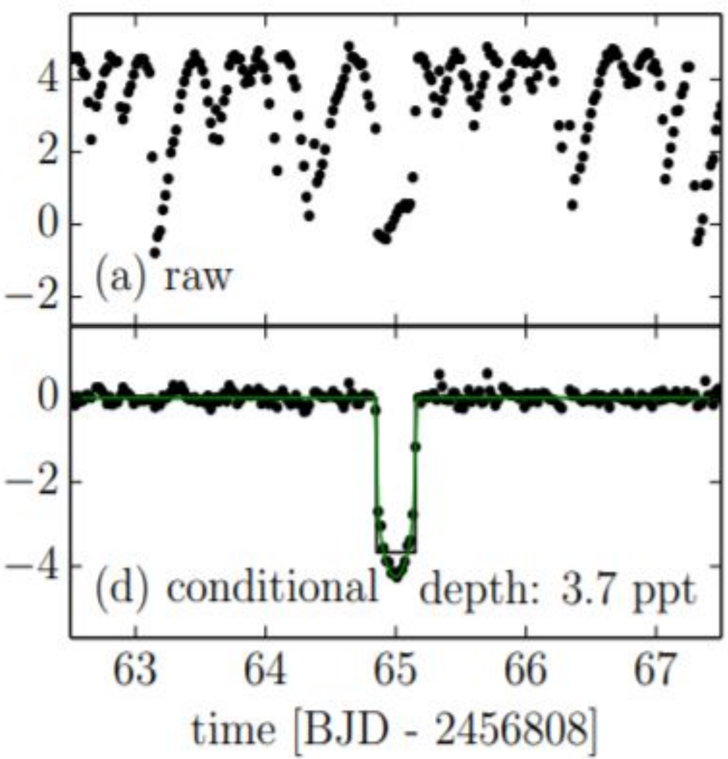


VanderPlas et al. 2014

Model galaxy spectra as a sum of linear components.

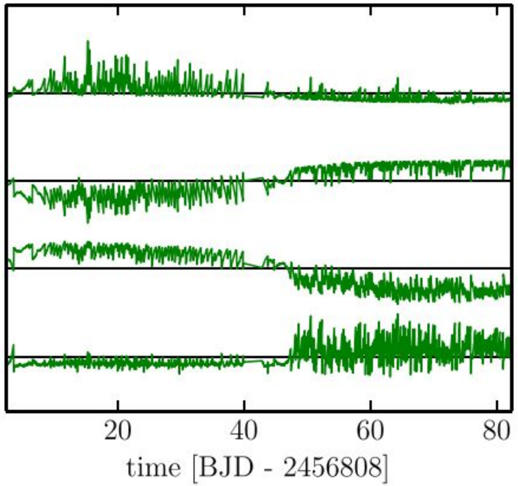
Automatically separates different kinds of galaxies.

Detrending with Kepler - PCA



Find noise in common between observations.

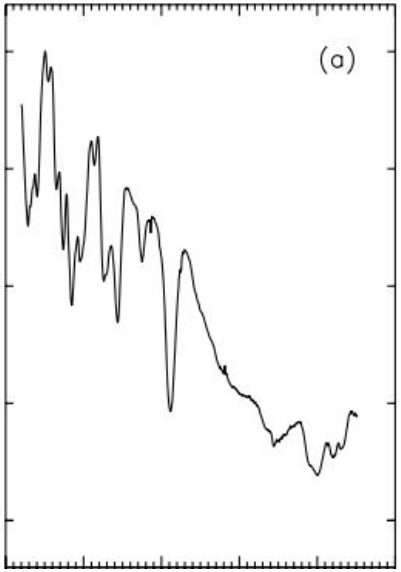
Foreman-Mackey et al. 2015

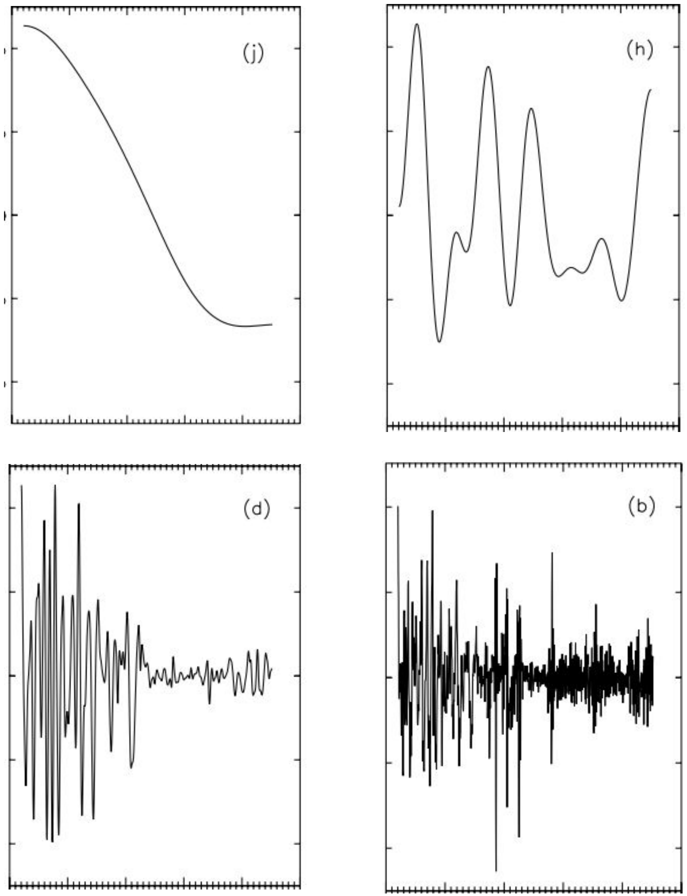


Remove common noise

Wagers et al. 2010

SN Spectra - Wavelets

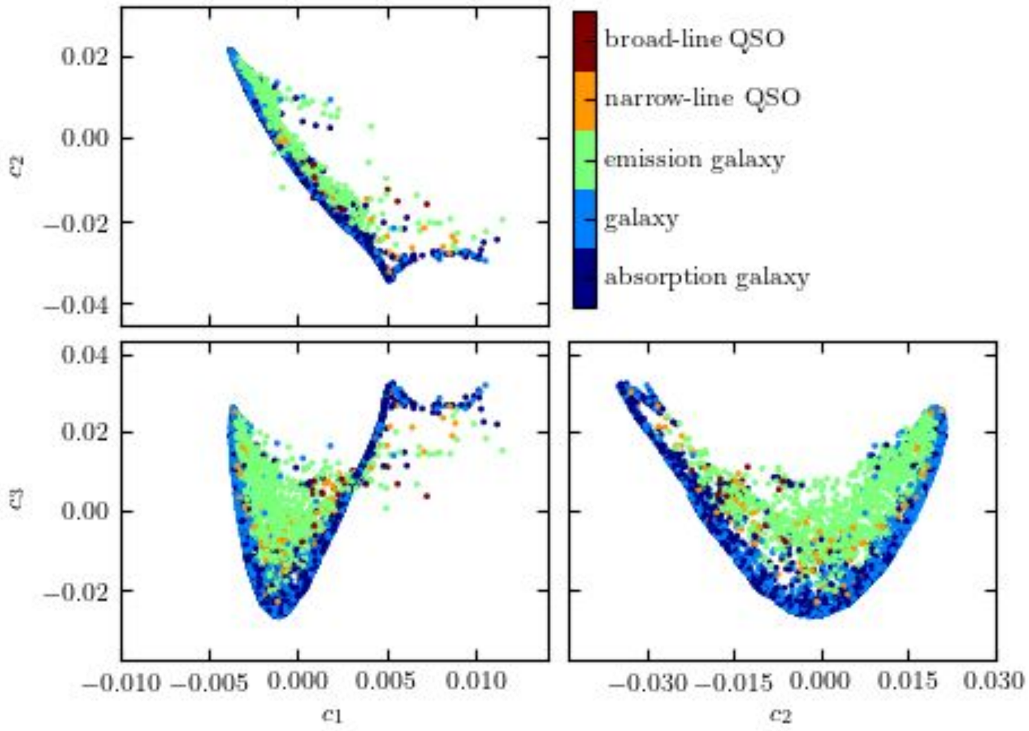




Split a signal into different scales. Often used for feature extraction.

Galaxy Spectra - PCA

VanderPlas et al. 2014



Why is there such strange structure?

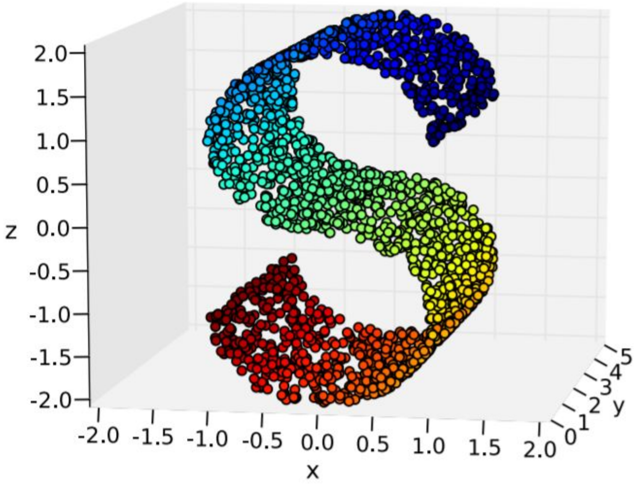
Linear models (like PCA) struggle to capture nonlinear behavior (e.g. line widths, velocities, etc.)

Manifold Learning

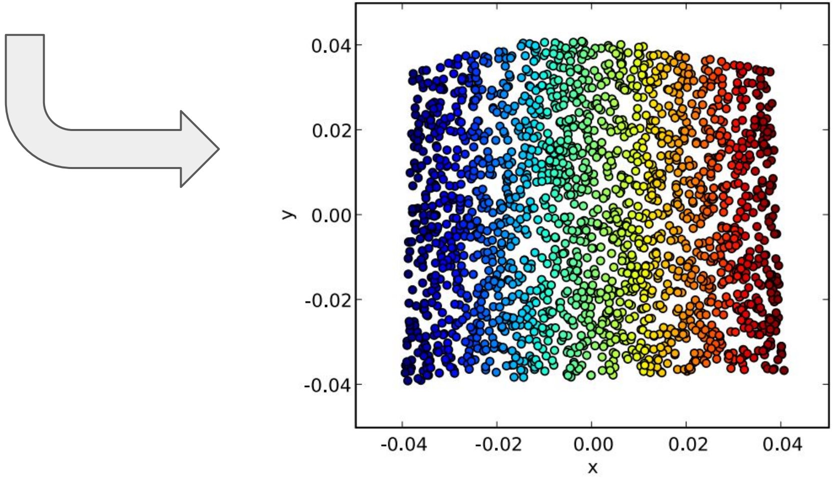
Often times our data has complex structure, and PCA won't work.

→ Use the local structure to recover a low-dimensional manifold.

e.g. UMAP, t-SNE, Isomap, LLE, ...

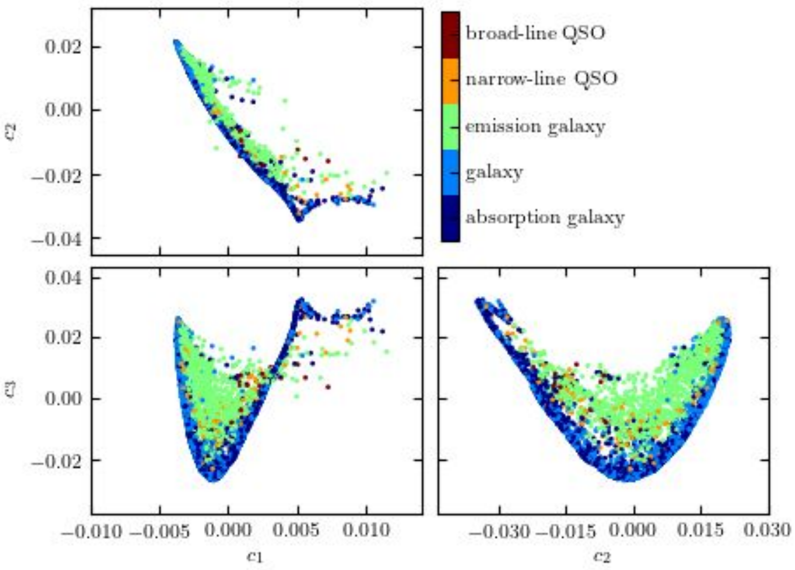


VanderPlas 2009

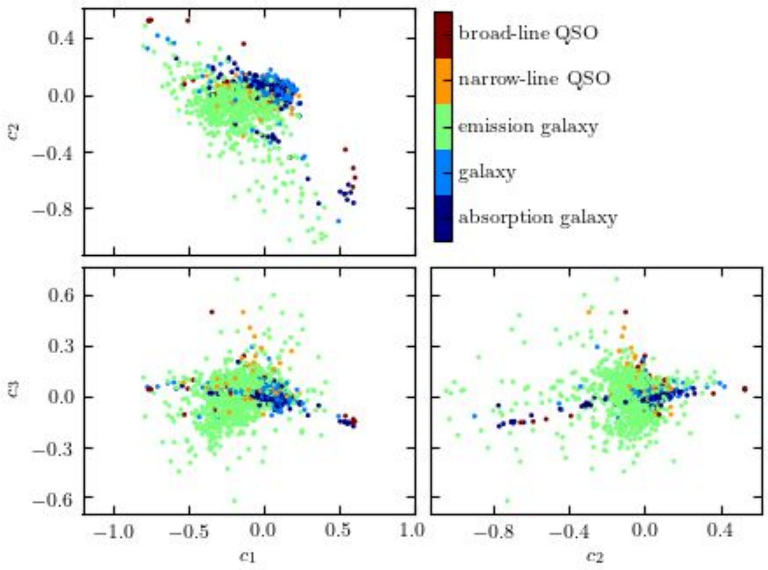


Galaxy Spectra - Locally Linear Embedding

PCA



Locally Linear Embedding



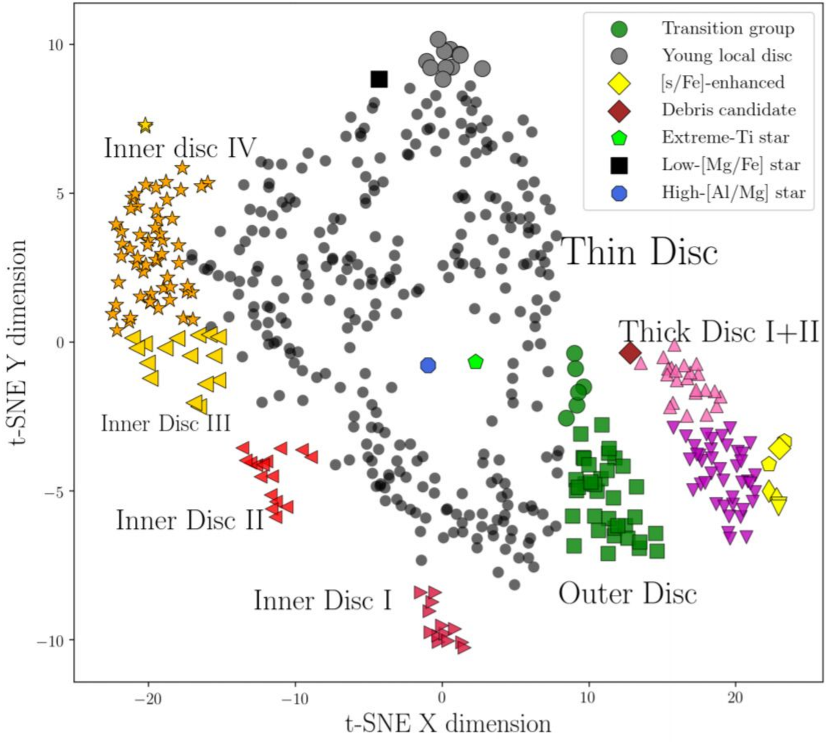
VanderPlas et al. 2014

Stellar Chemical Abundance - t-SNE

Anders et al. 2018

Generate a two-dimensional embedding from 13 different element abundances.

Very useful for visualizing subgroups.

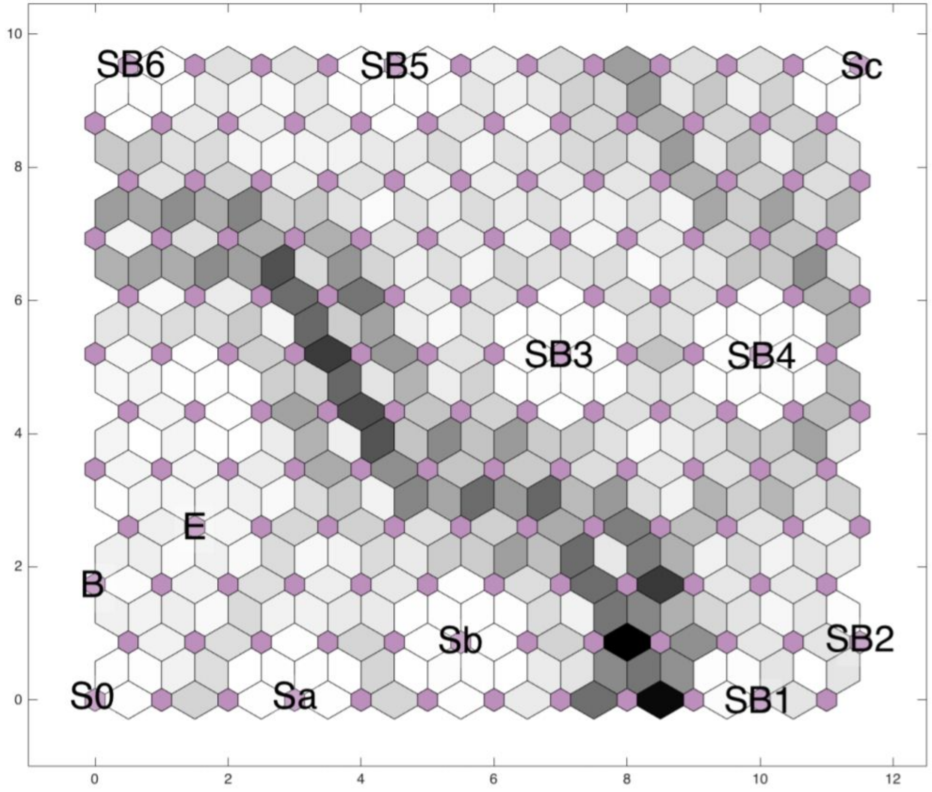


Galaxy Spectra - Self Organizing Maps

Rahmani et al. 2018

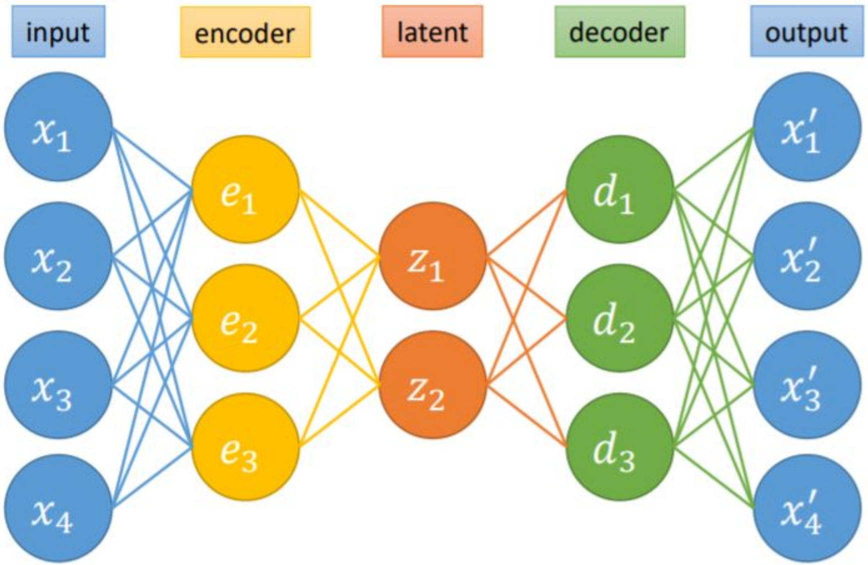
Fit a mesh of neurons to observations.

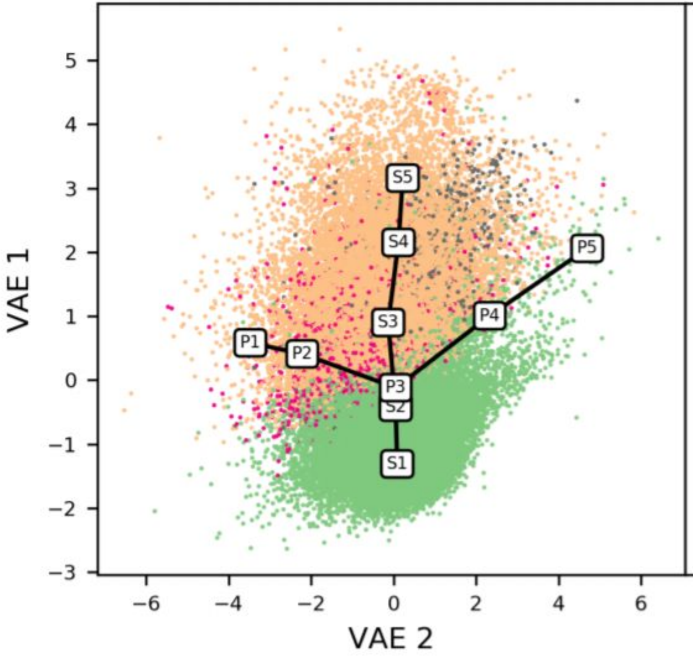
Used for photometric redshifts.



Galaxy Spectra - Autoencoders

Portillo et al. 2020





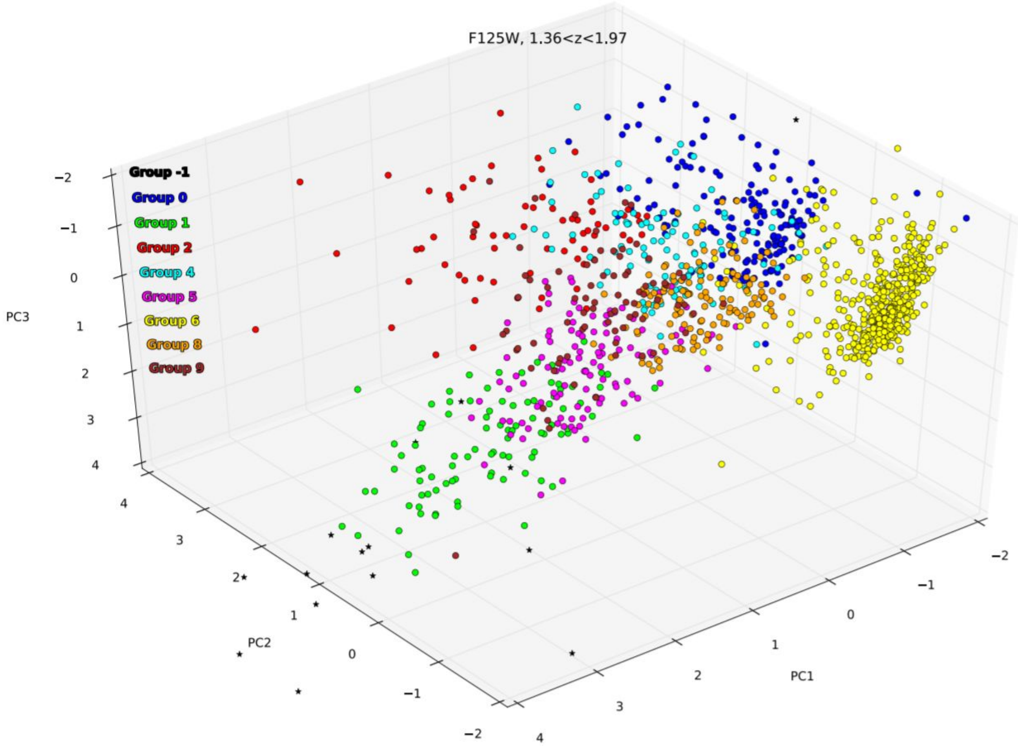
Neural network with a bottleneck layer that encodes a low-dimensional representation.

Galaxies - Clustering

Peth et al. 2016

Often times we want to identify subgroups in a dataset. Many ways to do this:

- K-means
- Hierarchical clustering
- Mixture models

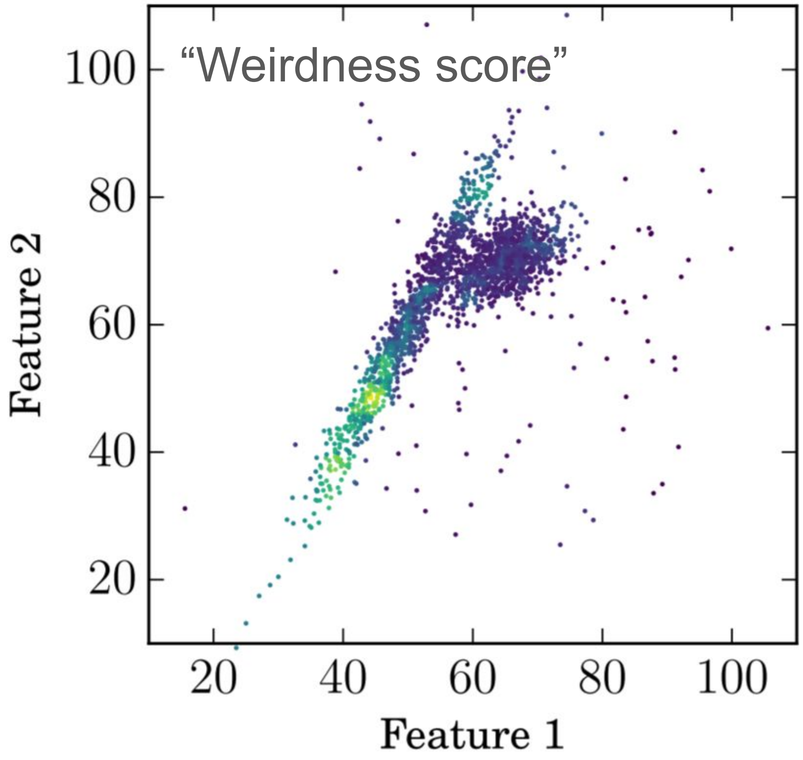


Galaxy Spectra - Anomaly Detection

Baron and Poznanski, 2017

Identify spectra that are very different from the rest of the sample or from known objects.

e.g. density modeling, isolation forest

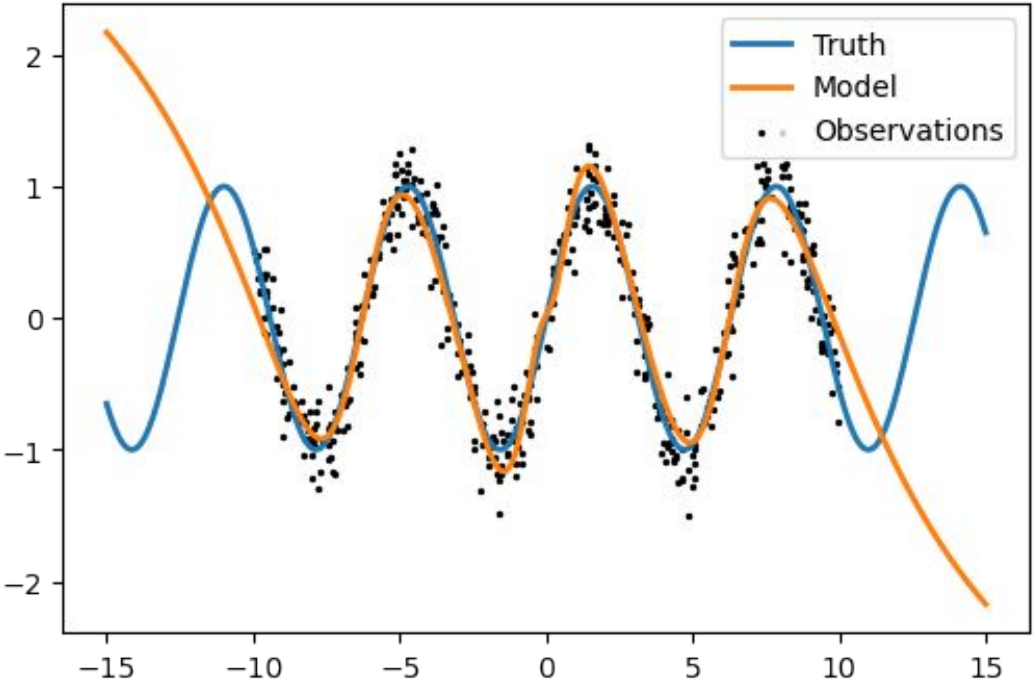


4. Challenges of ML in Astronomy

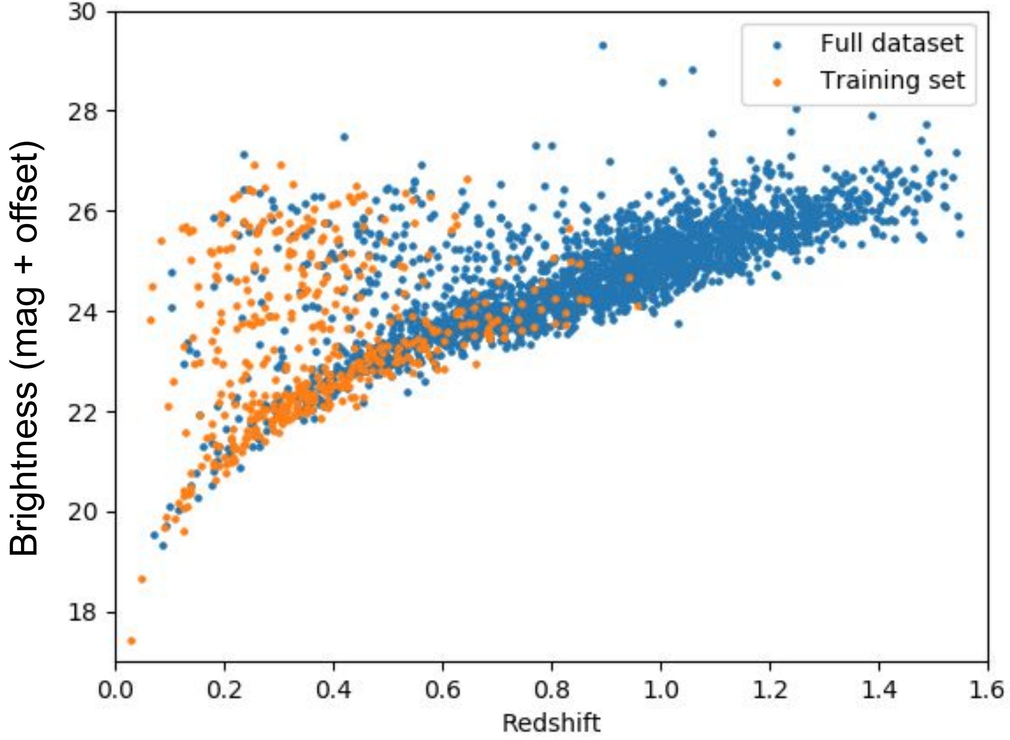
ML algorithms interpolate.
They don't extrapolate.

The training set must be representative of the test set.

Garbage in, garbage out!



Training sets

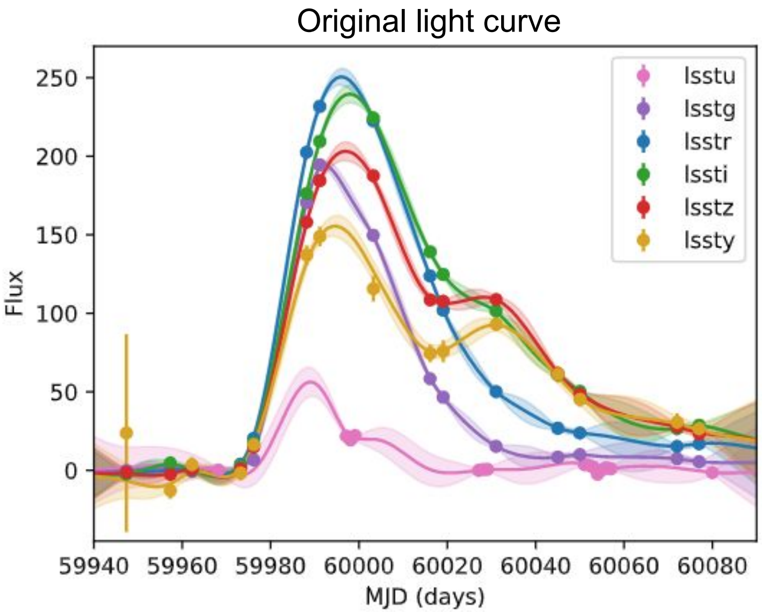


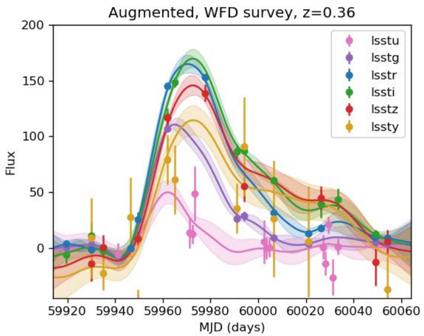
Training sets in astronomy tend to be very biased!

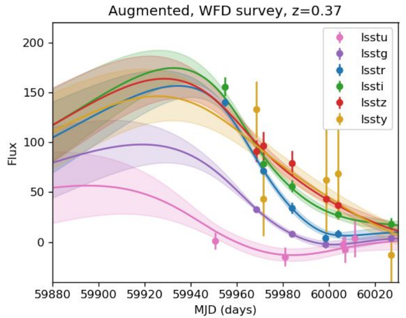
e.g. transient classification. It is easier to follow up brighter objects.

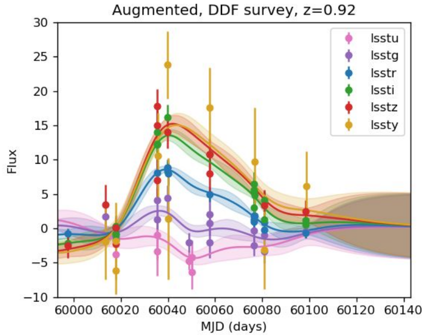
Augmentation

Boone 2019









Modify our training set to look like the full dataset.

e.g. make things fainter and add noise.

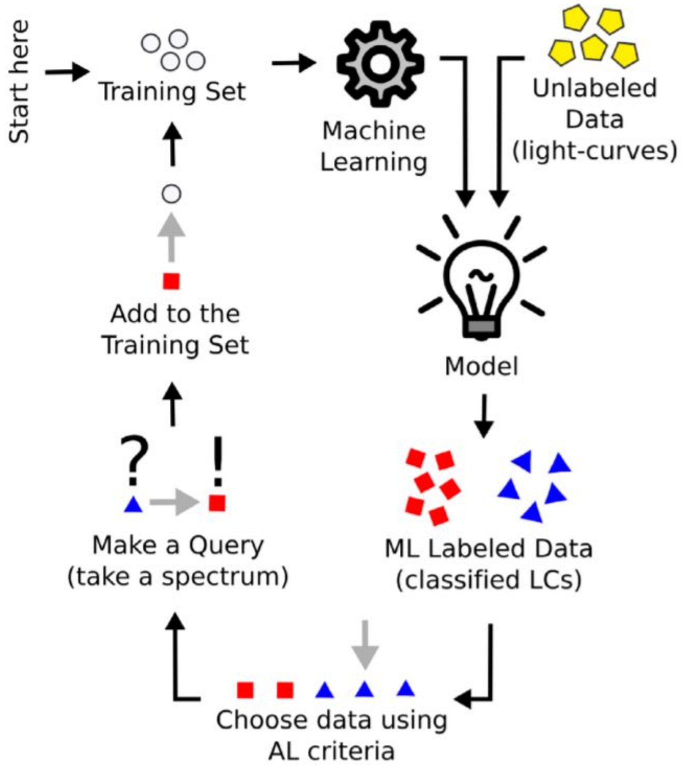
Take advantages of symmetries!

Active Learning

What is the optimal way to use our telescope resources?

→ Train a ML algorithm to tell us what it is missing in the training set.

Ishida et al. 2019

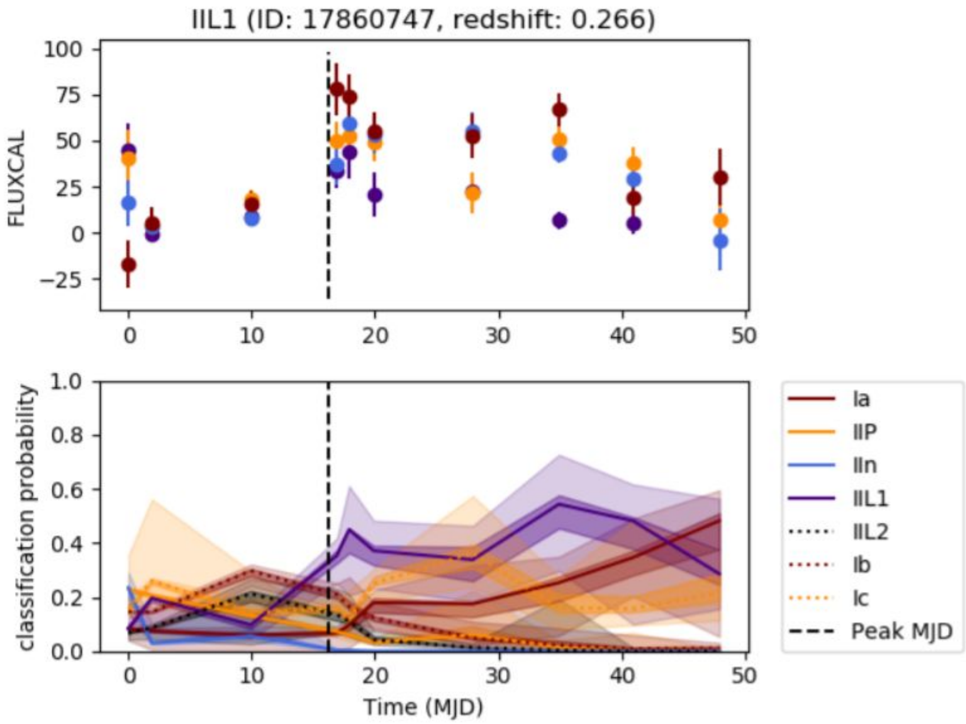


Uncertainties and Biases

Moller & de Boissiere 2019

Can modify ML algorithms to provide uncertainties on estimates, e.g. MC dropout.

Warning: modeling errors don't capture training set differences!

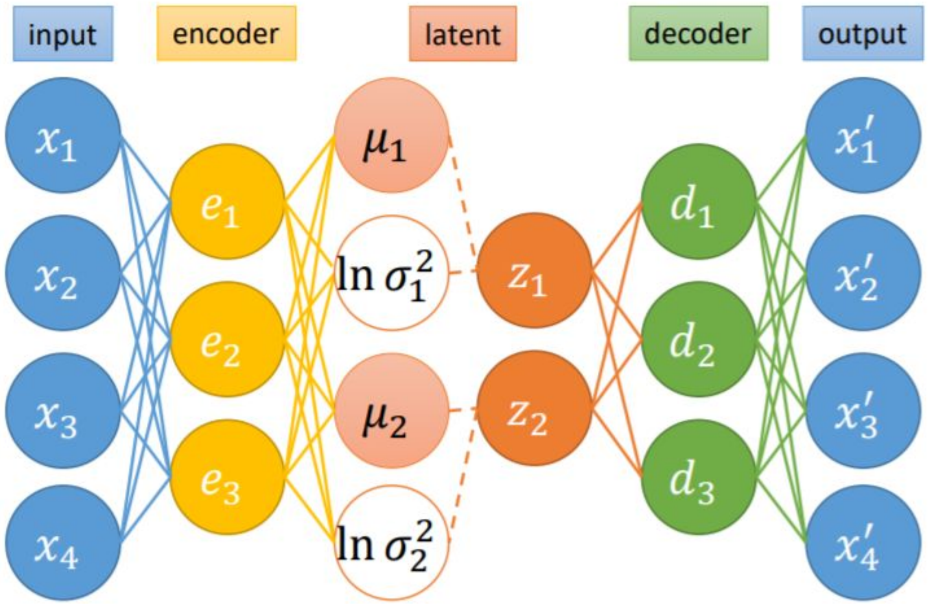


Variational Inference

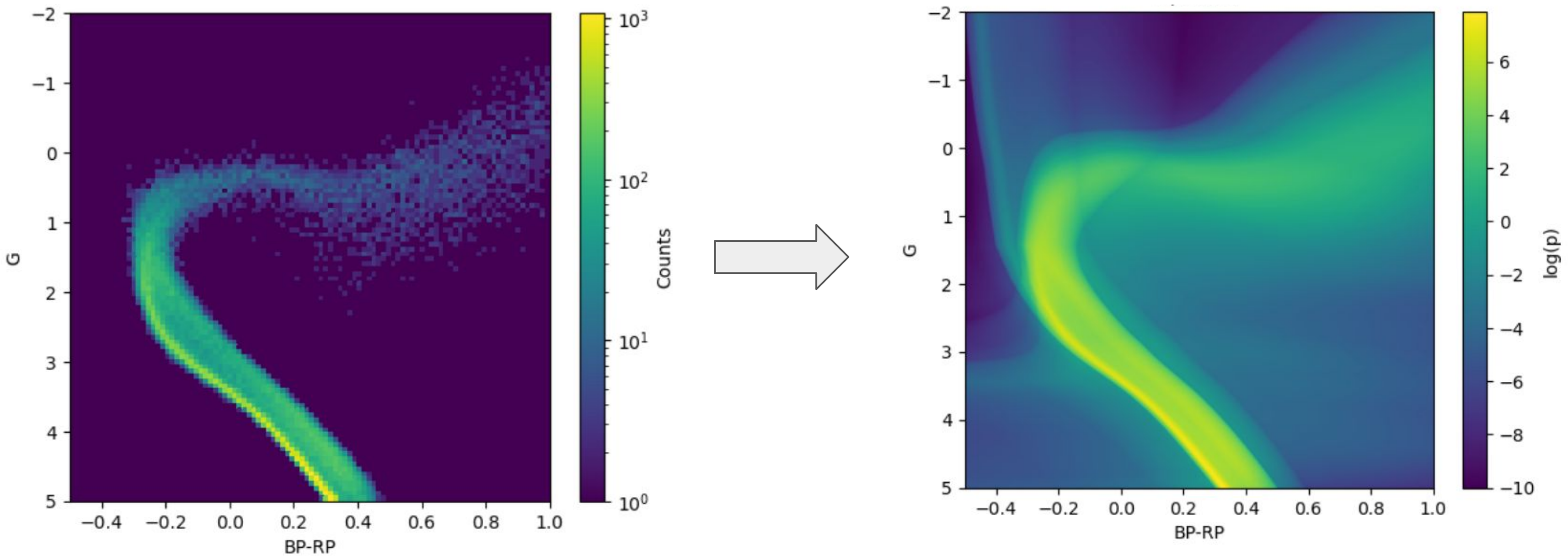
Portillo et al. 2020

Predict the parameters of a distribution instead of a point estimate.

e.g. Variational Autoencoders



HR Diagram - Normalizing Flows



Model a probability distribution with a neural network using a continuous transformation of a known distribution.

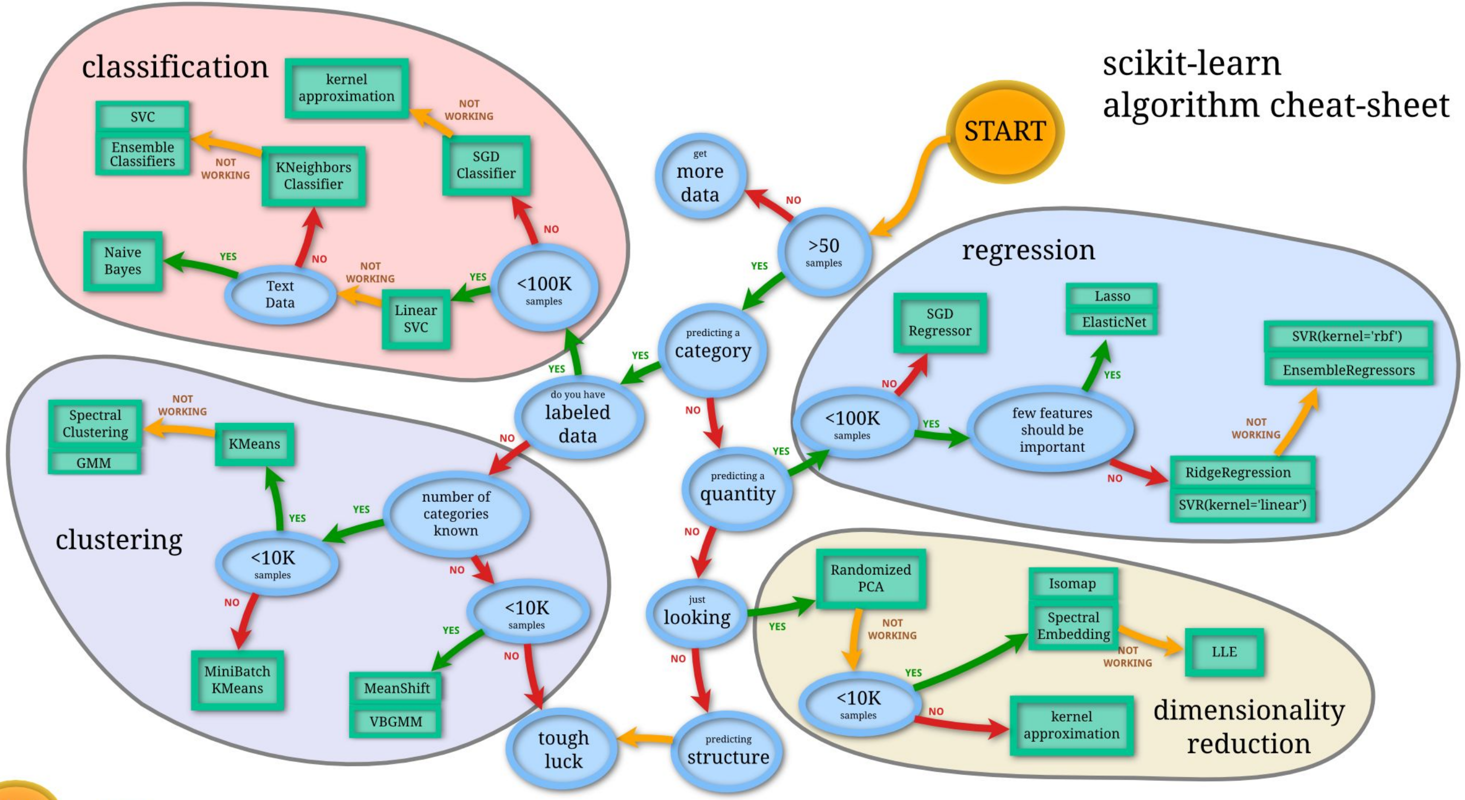
Togootogtokh and Amartuvshin, 2018

(a) (b)

Figure 1. (a) Chihuahua and muffin, (b) Labradoodle and fried chicken

Machine learning is not a magic solution!

Need to think carefully about the problem, algorithm, and training data.



Back

Lots of algorithms out there, almost all are used in astronomy!

More examples: https://www.astroml.org