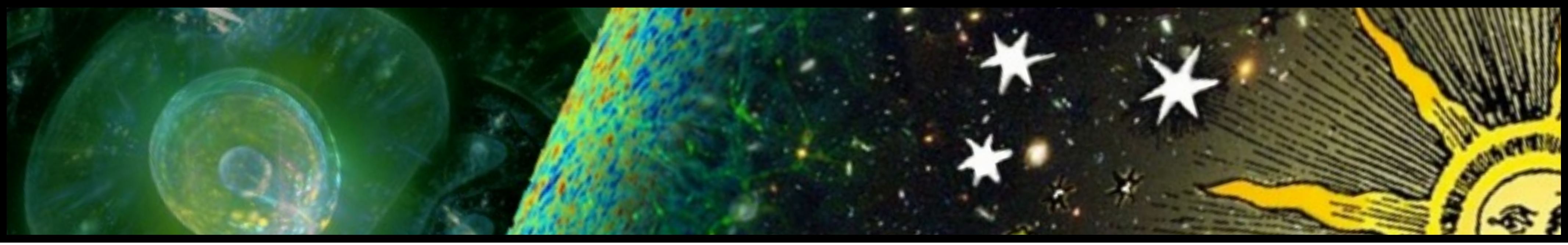


Machine learning for classification in cosmology

Hiranya V. Peiris
UCL and Stockholm

European Research Council

Oskar Klein centre

UCL

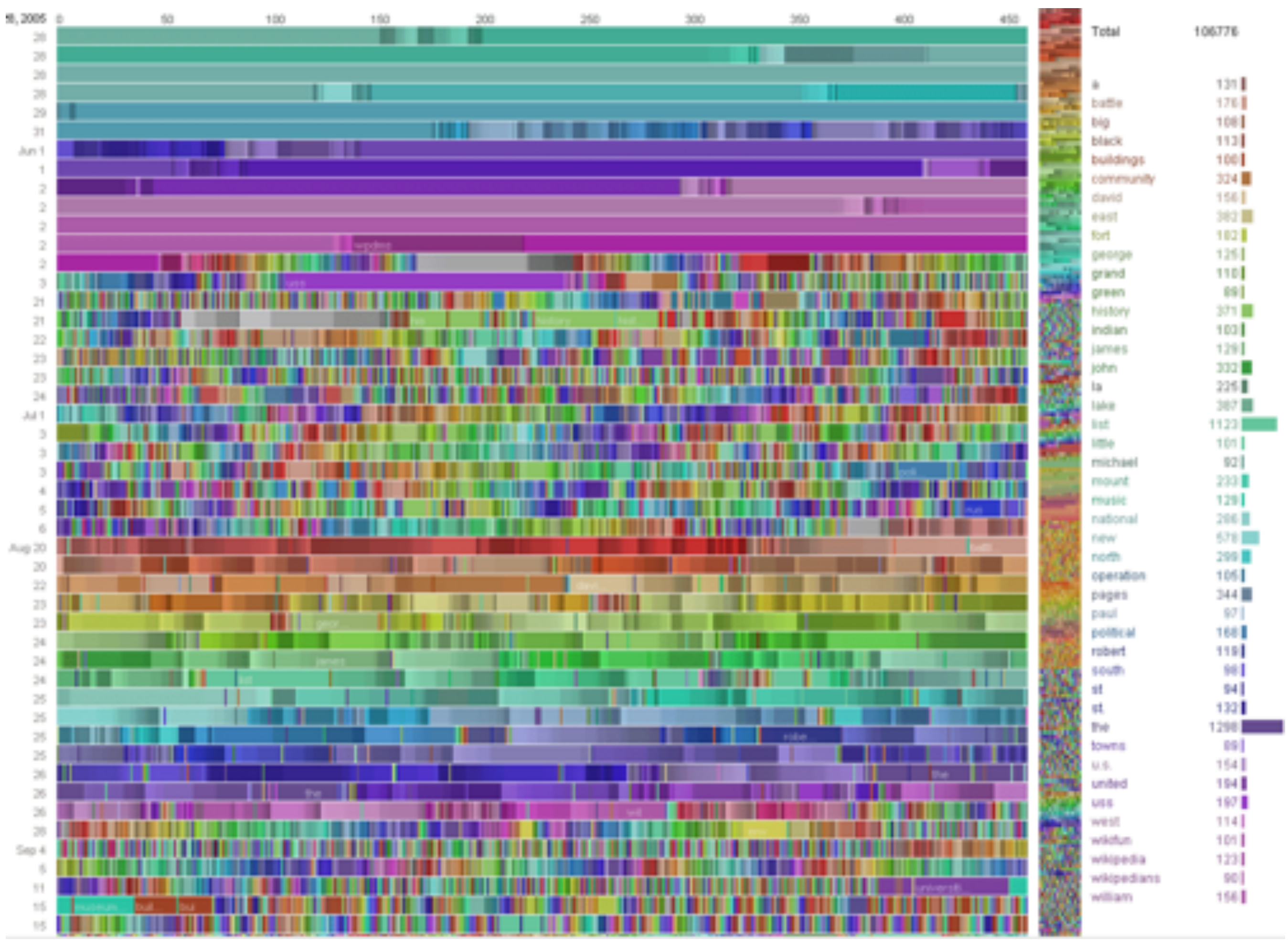
What is Machine Learning?

- *Automatically building a (usually highly nonlinear) model that maps a given input to output.*

- *Different algorithms use different prescriptions for building the model.*

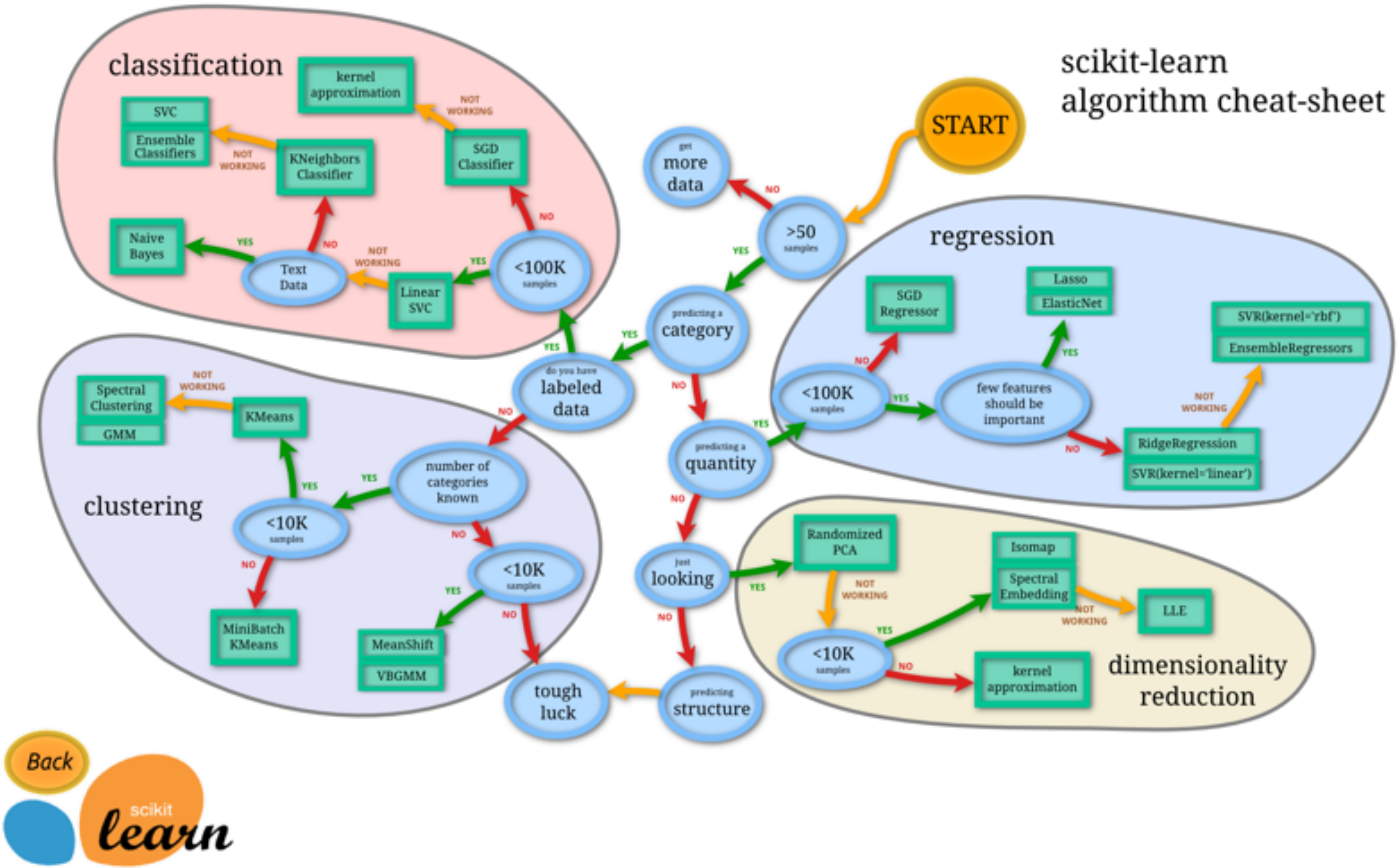
When to use Machine Learning?

- *When your data are too complex for traditional model development and fitting with statistics*



Activity on Wikipedia (Wikimedia Commons)

This is not a tutorial



Machine Learning terminology

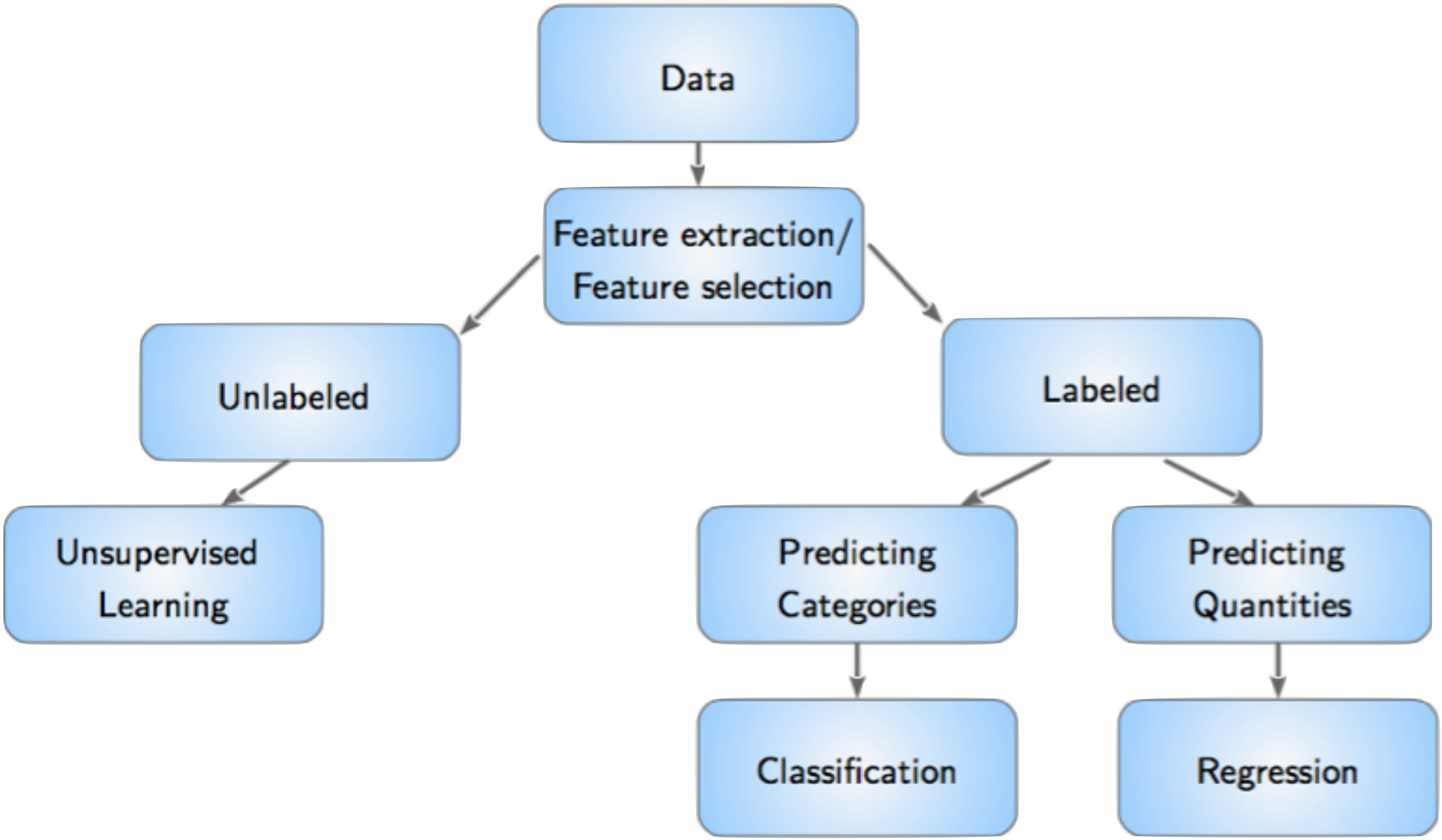
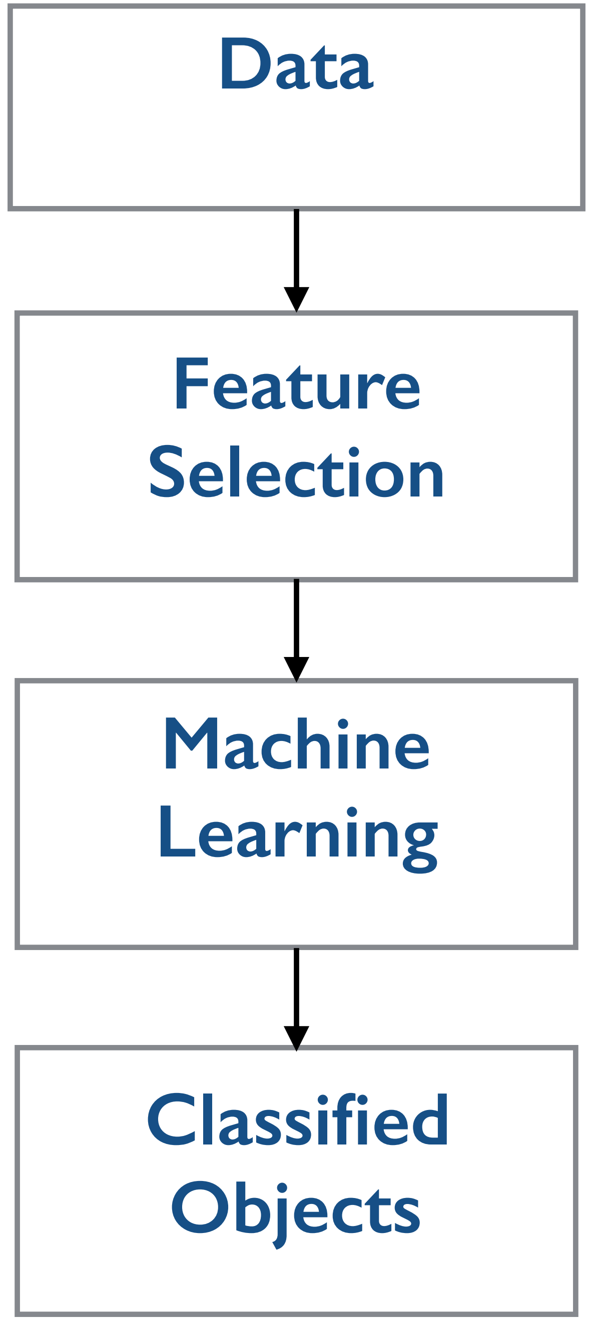


IMAGE: MICHELLE LOCHNER

A typical ML classification workflow



Decision Trees

- *At each leaf node, find best feature that split the data (i.e., best separation between classes), and the best split value of that feature.*

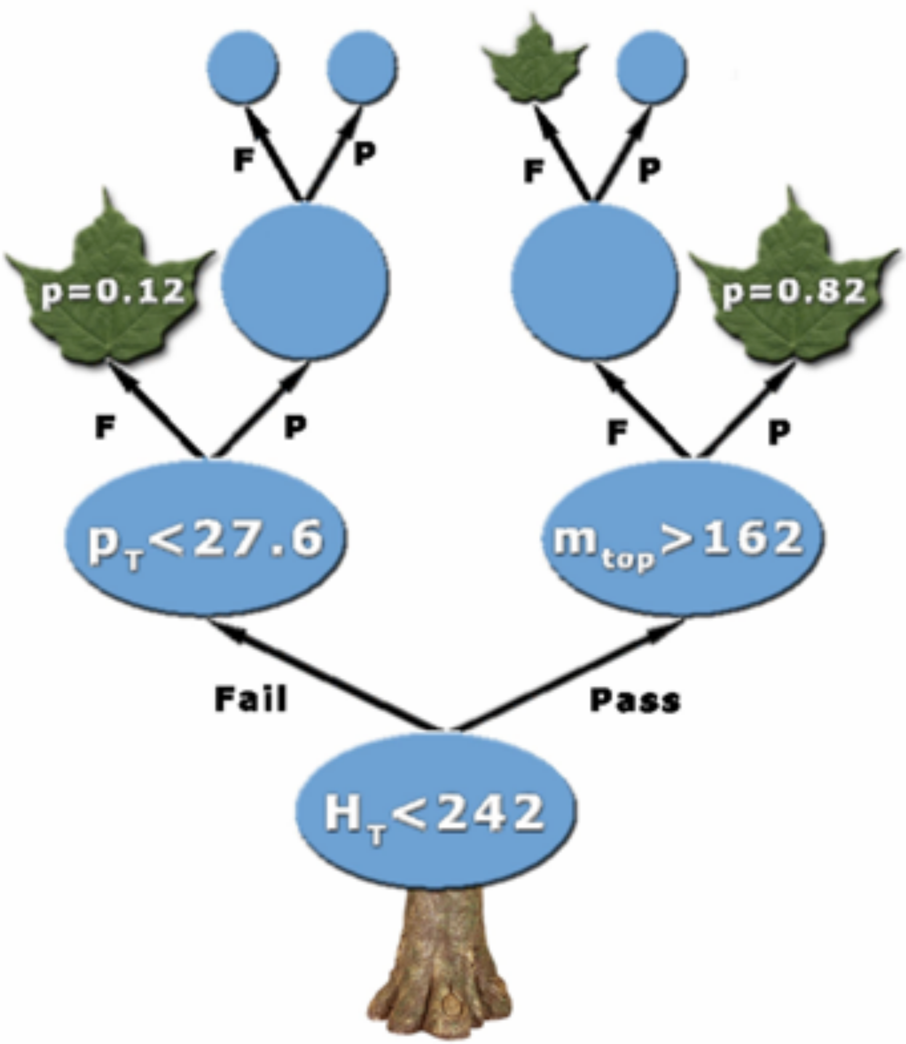
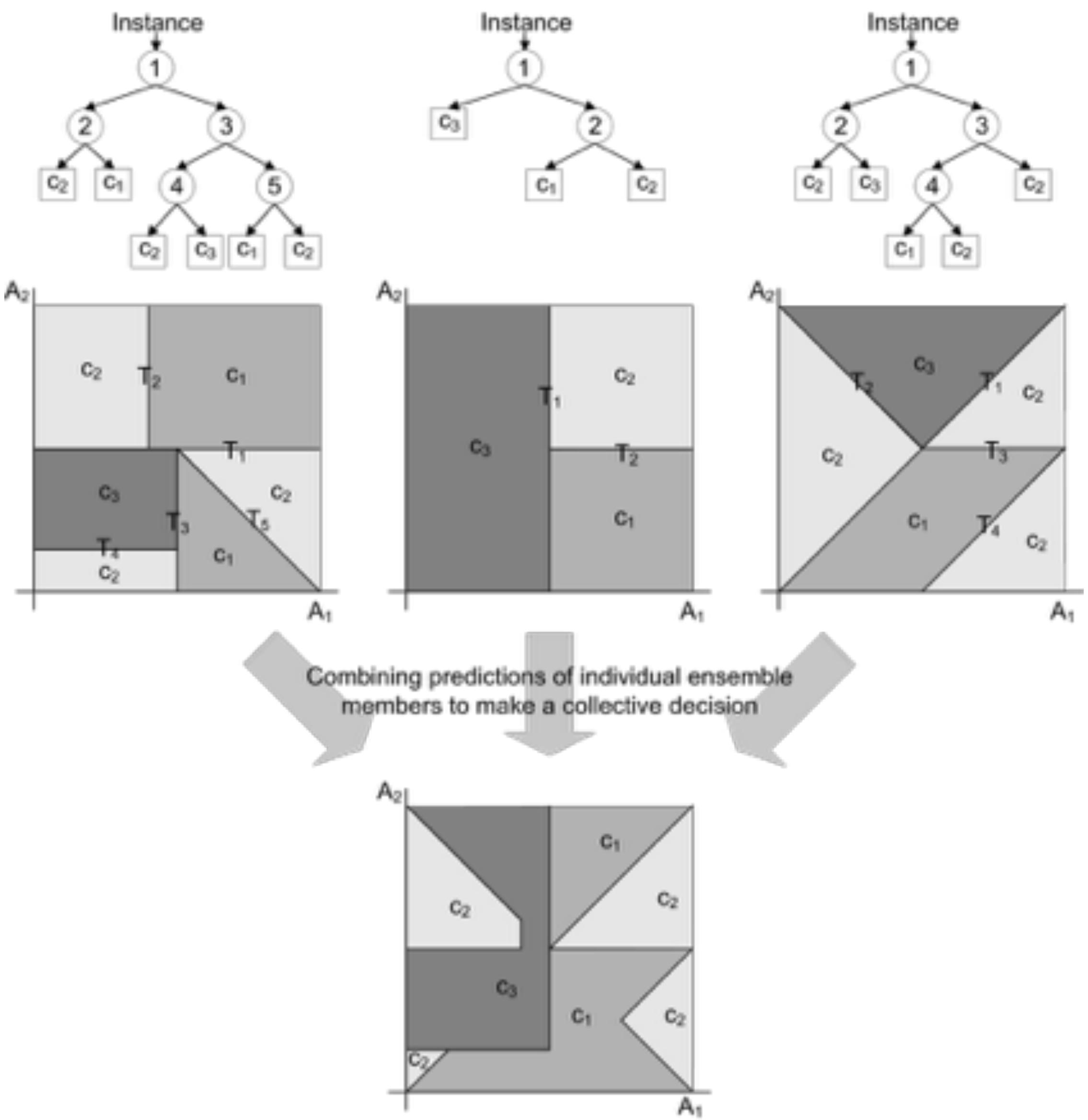


IMAGE: WIKIMEDIA COMMONS

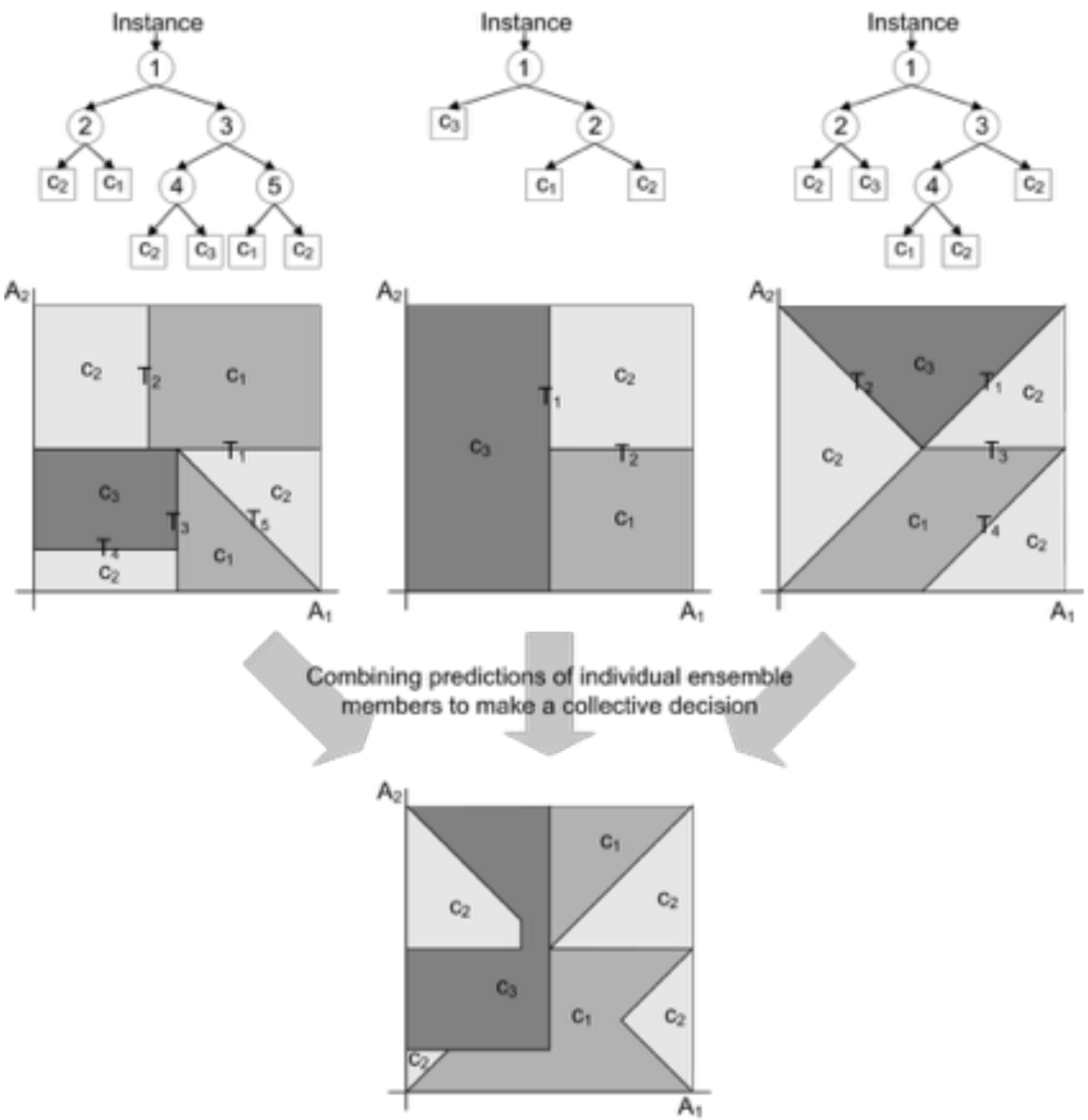
Ensemble methods



- *Ensemble methods average weak classifiers to create robust classifier.*

IMAGE: WIKIMEDIA COMMONS

Ensemble methods with decision trees



- *Robust (low variance)*
- *Allows mixed feature types*
- *Robust to high dimensionality*
- *Can rank feature importance*
- ***Random Forests:*** *my classification algorithm of choice.*

IMAGE: WIKIMEDIA COMMONS

LSST survey of 18,000 sq deg

(half the sky)

- 4 billion galaxies (with photo-z)
- Time domain:
  - 5 million asteroids
  - 1 million supernovae
  - 1 million gravitational lenses
  - 100 million variable stars

\+ new phenomena

survey of 37 billion objects in space and time

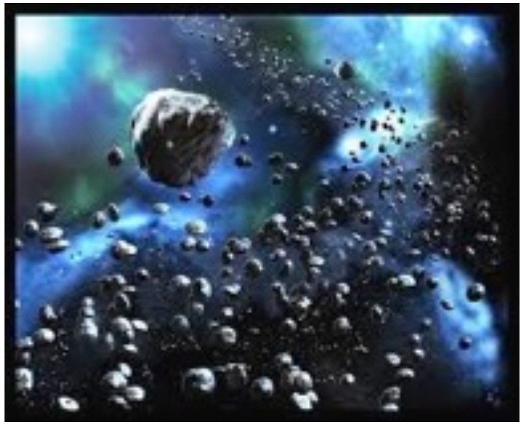
30 trillion measurements

LSST 4 science missions

Dark matter-Dark energy

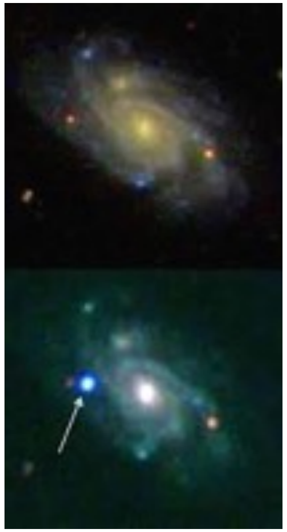
Multiple investigations into the nature of the dominant components of the Universe.

Solar system inventory



Find 90% of hazardous NEOs down to 140m over 10 years; test theories of Solar System formation.

"Movie of the Universe"



Discovering the transient and unknown over time scales days to years

Mapping the Milky Way

Map the rich and complex structure of the Milky Way in unprecedented detail [test-beds for dark matter physics]

All missions conducted in parallel.

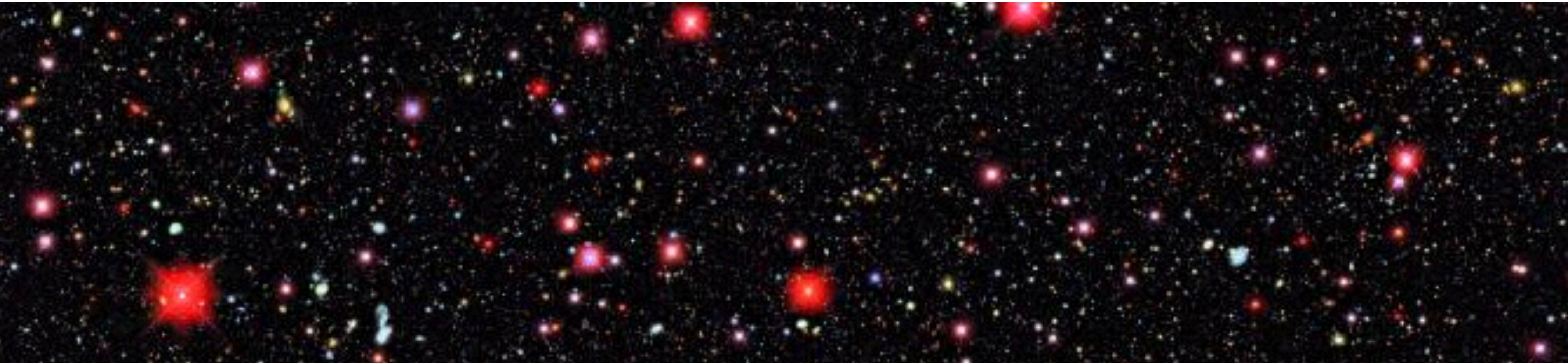
Adapted from Ian Shipsey

LSST is a "datascope".

Due to its deep/wide imaging and high cadence, LSST enables unprecedented data-driven astronomical discovery, including:

- new classes of objects and processes;
- new attributes of known classes;
- rare events and objects;
- novel temporal behaviour.

Making discoveries using LSST's 100 PB-petascale database (10000-D with 40-billion entries) requires classification, statistical inference, clustering, outlier-detection and multi-resolution algorithms.



From Zeljko Ivezic

The era of surveys…

"Ask Not What Data You Need To Do Your Science, Ask What Science You Can Do With Your Data."

LSST From the User's Perspective: A Data Stream, a Database, and a (small) Cloud

Nightly Alert Stream

- A stream of ~10 million time-domain events per night, detected and transmitted to event distribution networks within 60 seconds of observation.
- A catalog of orbits for ~6 million bodies in the Solar System.

Level 1

Yearly Data Releases

- A catalog of ~37 billion objects (20B galaxies, 17B stars), ~7 trillion single-epoch detections ("sources"), and ~30 trillion forced sources, produced annually, accessible through online databases.
- Deep co-added images.

Level 2

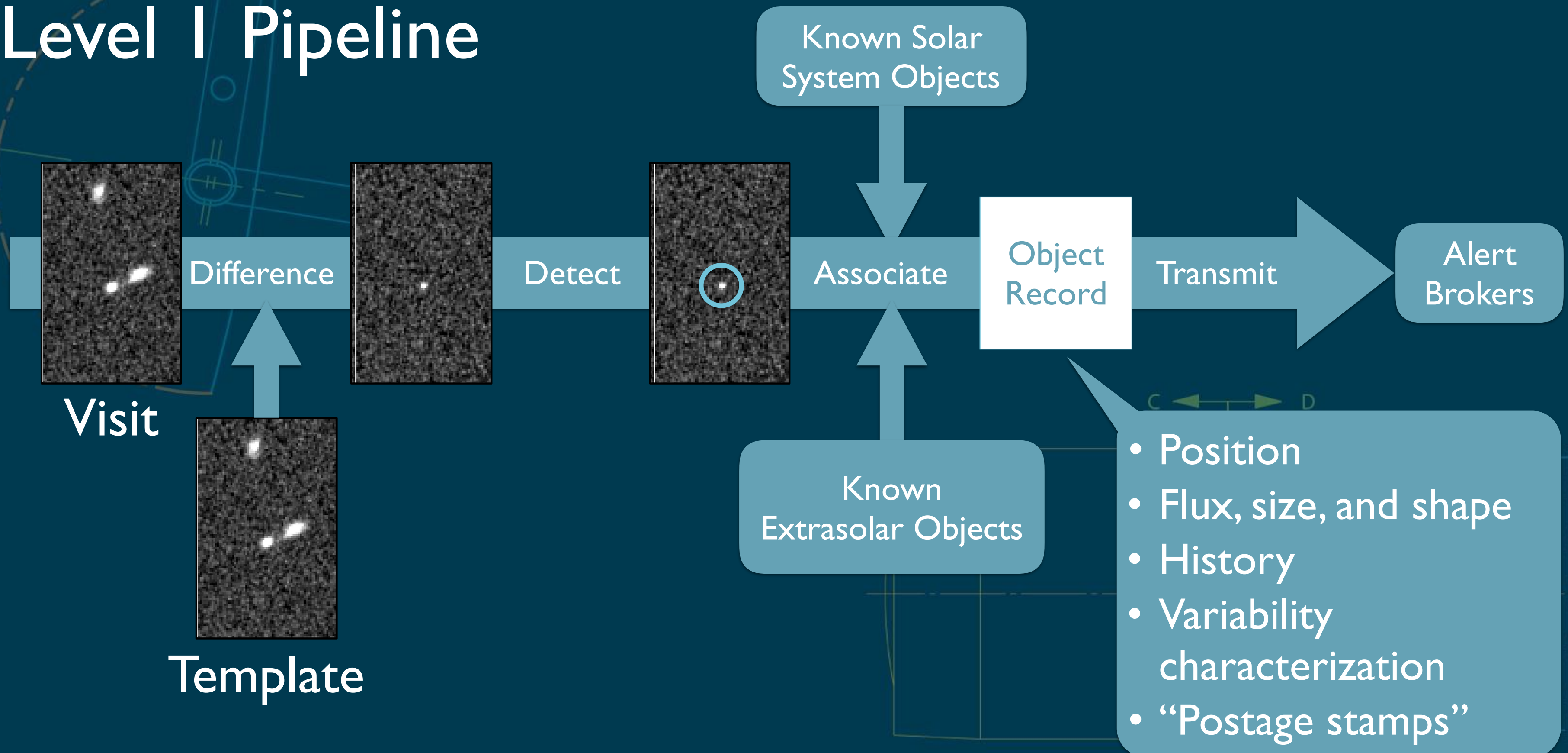
Community Services

- Services and computing resources at the Data Access Centers to enable user-specified custom processing and analysis.
- Software and APIs enabling development of analysis codes.

Level 3

LSST Data Products: see http://ls.st/dpdd

Level 1 Pipeline



0 seconds

60 seconds

Meredith Rawls • @merrdiff

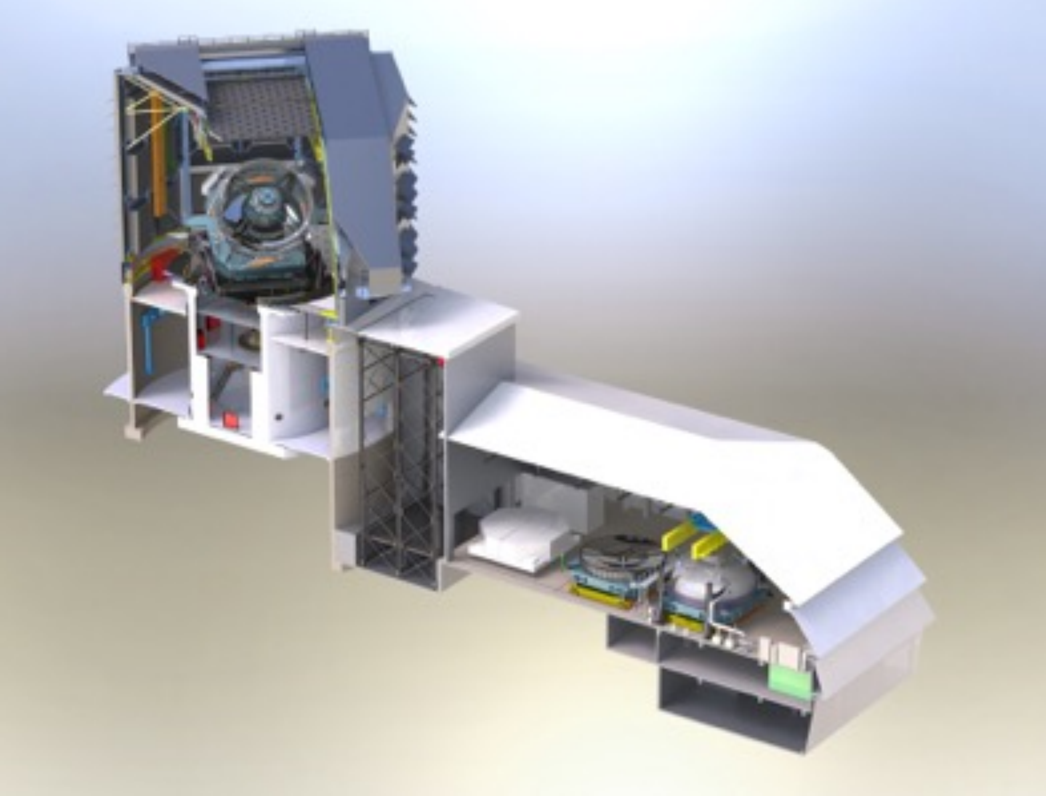
Shields up, red alert!

Nightly LSST alert stream	All of twitter
⬇	⬇
Alert brokers	Filtered search by hashtag
⬇	⬇
Find follow-up objects	Find interesting content

~60 kB/alert
~60 GB/night

- Alerts will include metadata, historical observations, and an image "postage stamp"
- Hierarchy of access systems via brokers
- Broadcast in a stream; archived in a database

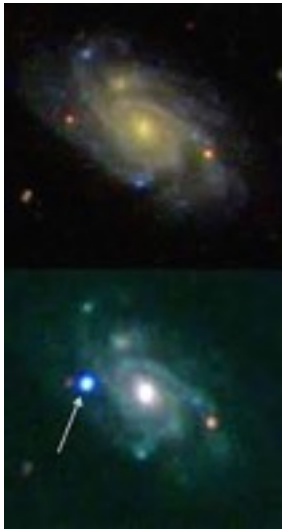
First light: 2019



BACKGROUND IMAGE: UCL / PONTZEN

Transient science

Movie of the Universe



Discovering the transient and unknown over time scales days to years

Known unknowns
Unknown unknowns

LSST will extend time-space volume a thousand times over current surveys

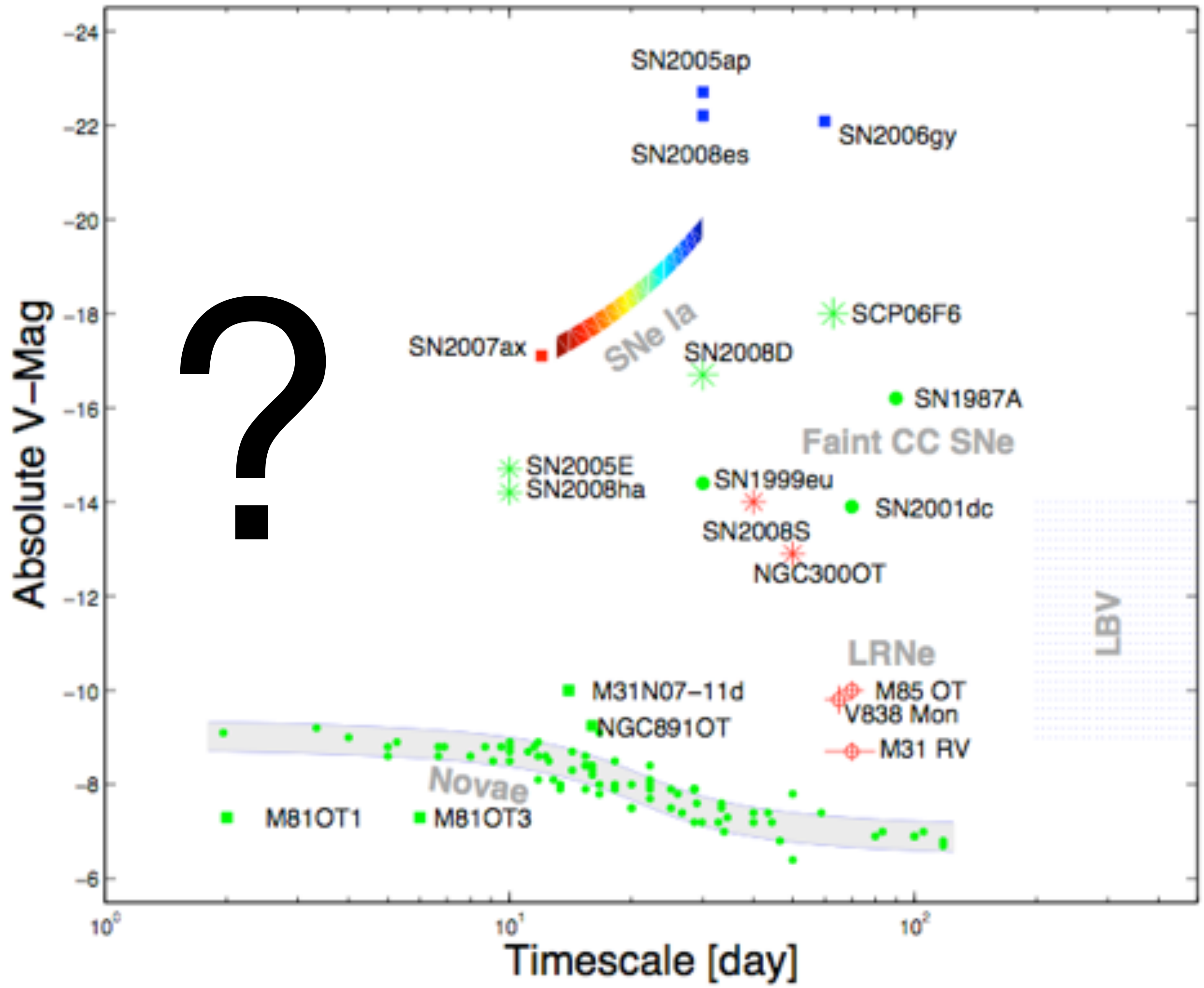


FIG. 29.— The phase space of cosmic explosive and eruptive transients as represented by their absolute V band peak brightness and the event timescale (adapted from Kulkarni et al. 2007).

Adapted from Zeljko Ivezic

SNMachine: Photometric Supernovae Classification with Machine Learning

Michelle Lochner

Jason McEwen

Robert Schuhmann

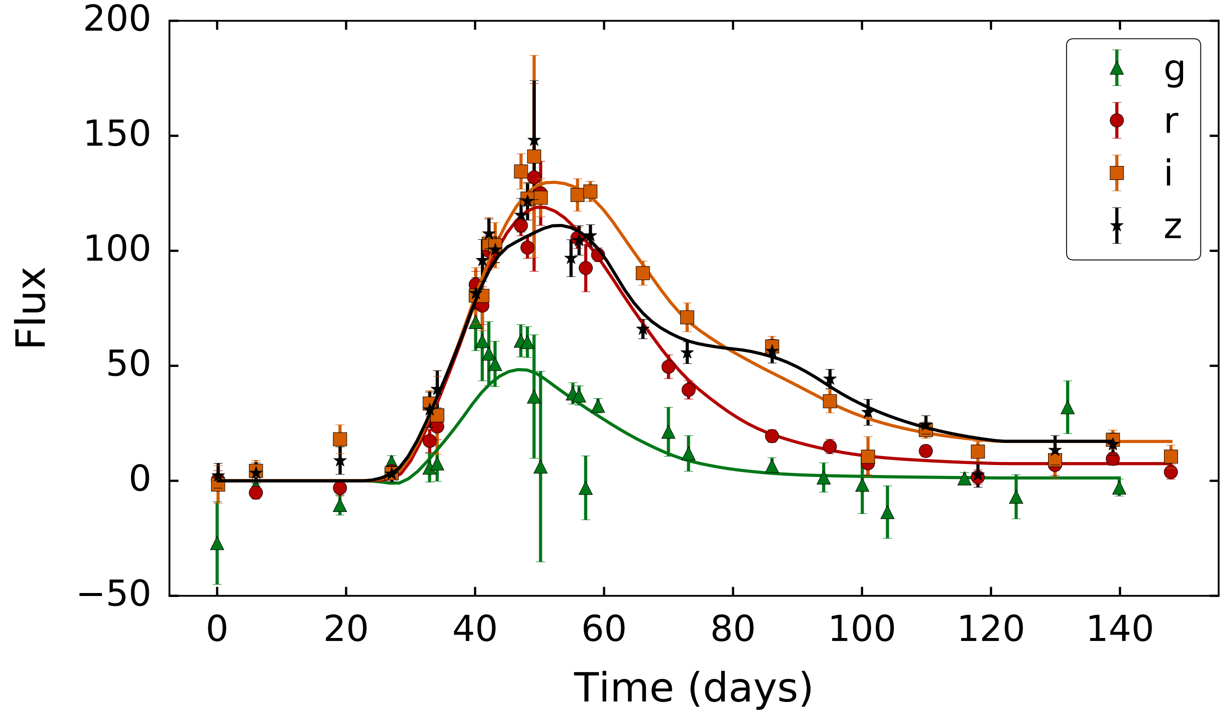
Lochner, McEwen, Peiris, Lahav, Winter (ApJ Suppl. 2016)

https://github.com/LSSTDESC/snmachine
(will be publicly released)

Why photometric classification?

- *In the past spectroscopic followup for majority of sample possible to determine SN type. Not scalable.*

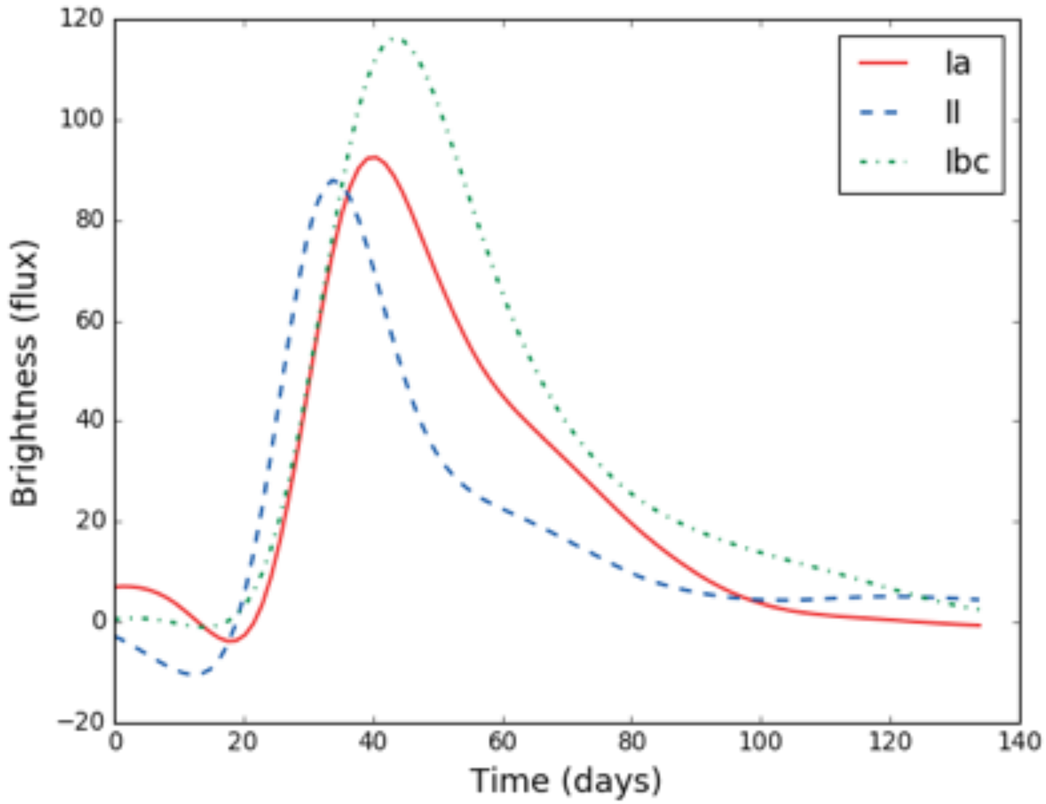
- JLA (Betoule et al 2014): 740 SNe
- DES: 1000s of SNe
- LSST: 100000s of SNe



Simulated DES type Ia supernova light curve at redshift 0.42, from Supernova Photometric Classification Challenge (Kessler et al. 2010).

The goal

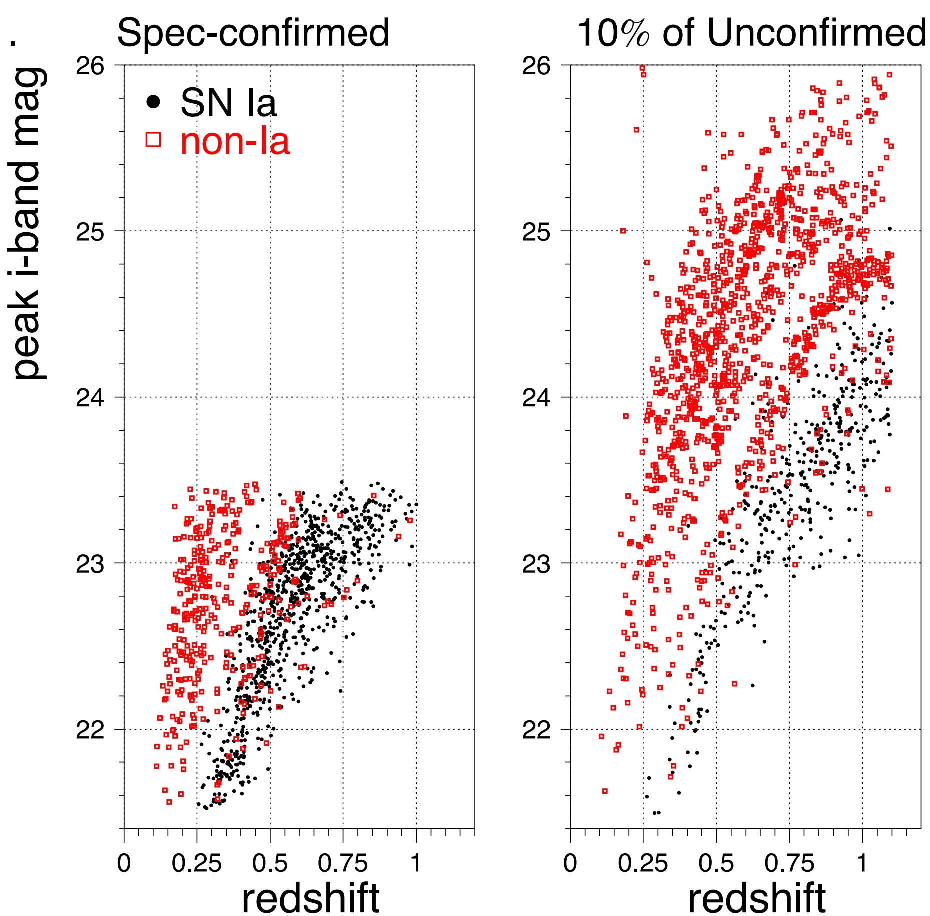
- *Maximise use of photometric data (for cosmology / SN science)*
- *Classify SNe based on their multi-band light curves*
- *Produce probability that SN is Ia, Ibc, etc*
- *Inform LSST observing strategy using realistic simulations*



Workflow: SNMachine

Data

Supernova Photometric Classification Challenge (Kessler et al. 2010).



Feature Selection

Machine Learning

Classified SNe

SNmachine pipeline overview

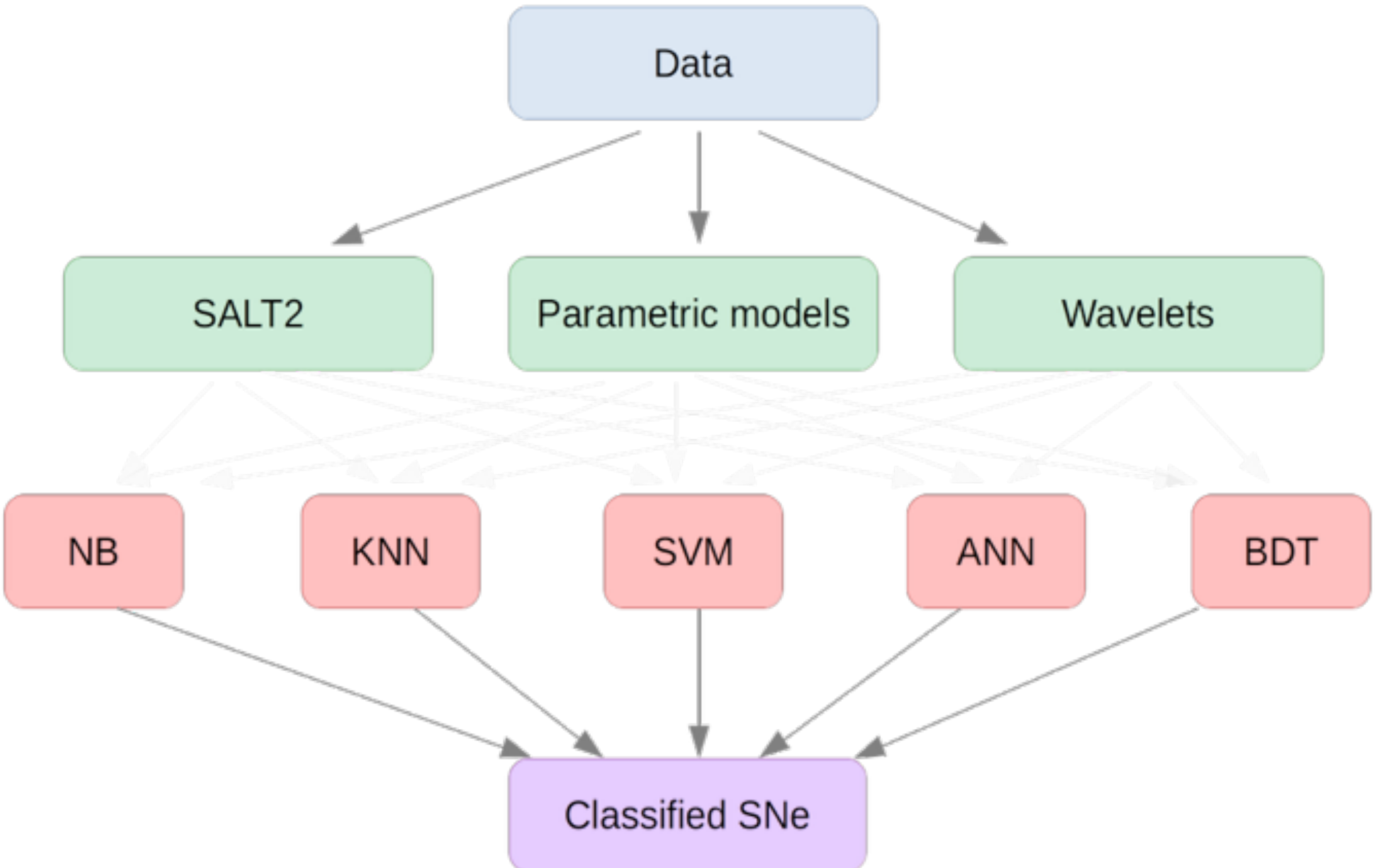
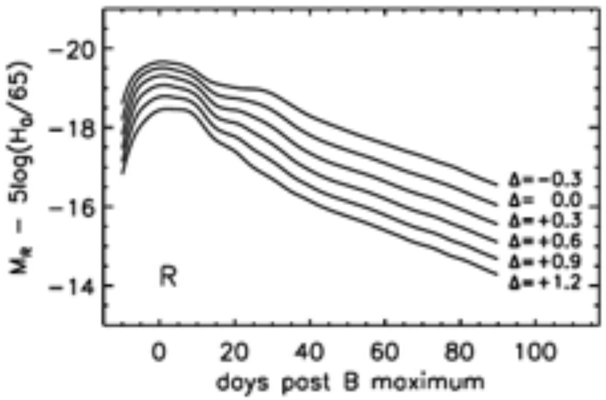
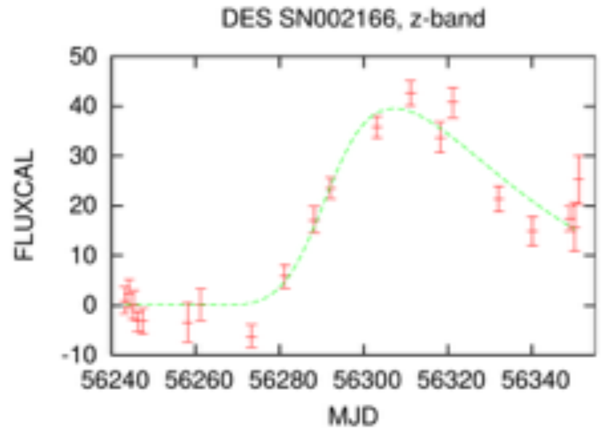
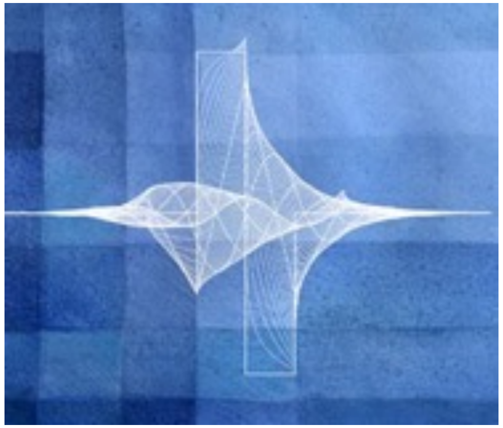


IMAGE: LOCHNER

Feature Selection







Template Fitting

SALT2 templates fitted with SNCosmo +MultiNest

General parameterisations

Karpenka et al (2014)
Newling et al. (2010)
fitted with MultiNest

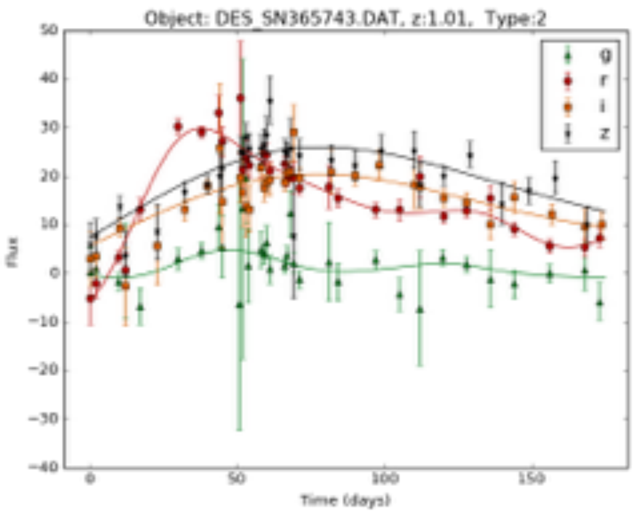
Wavelets

Gaussian Process fit to light curves; wavelet decomposition; PCA

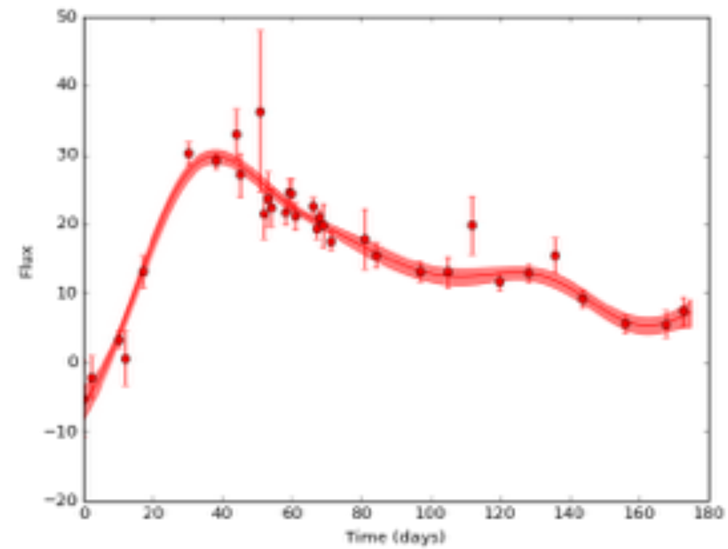
Model Independence

Wavelets

- *Decompose light curve into wavelets, then apply PCA to select most important wavelet coefficients from training set*



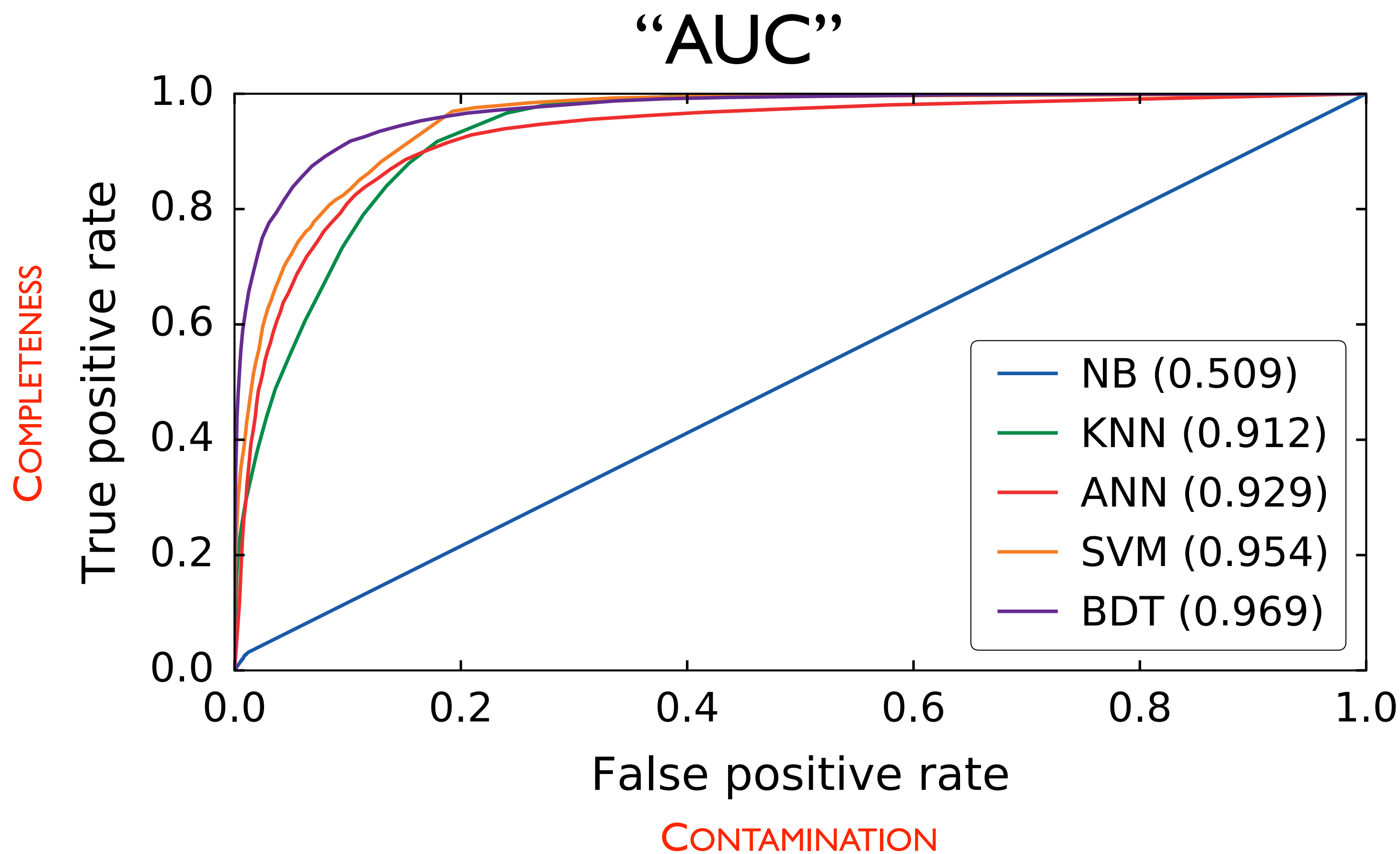
Gaussian Process fit



Wavelet decomposition

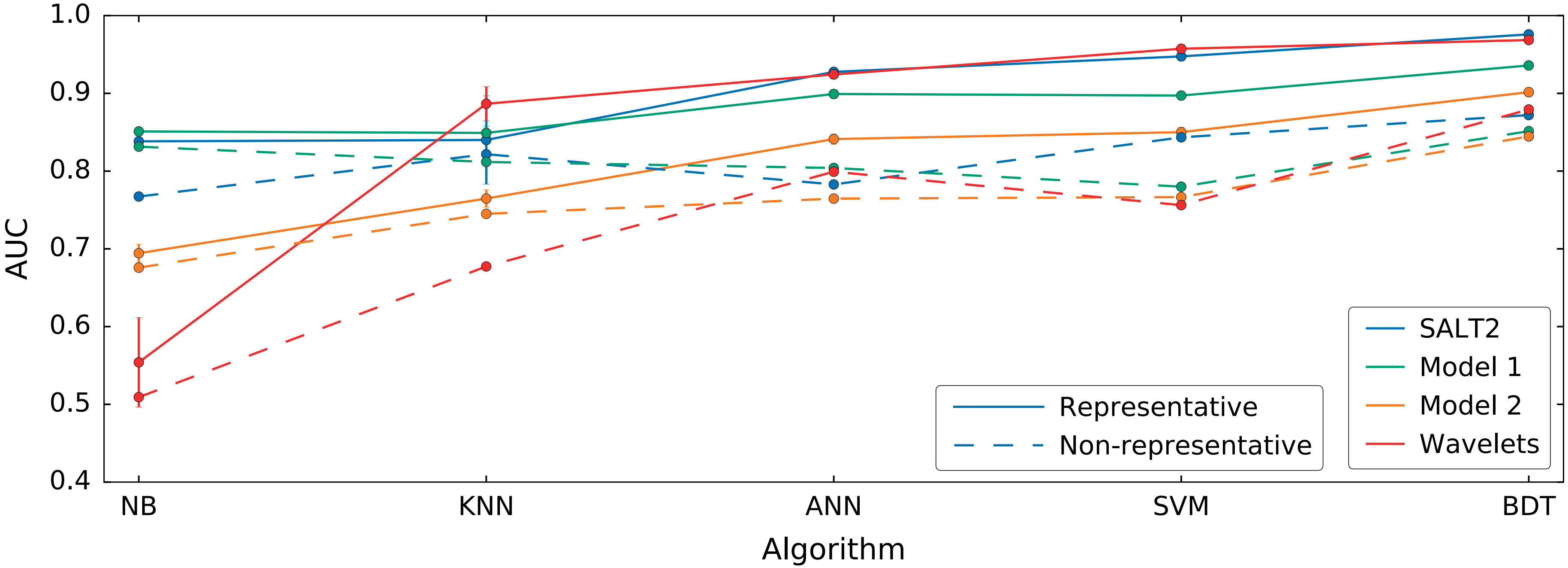
PCA

Wavelets ROC curves



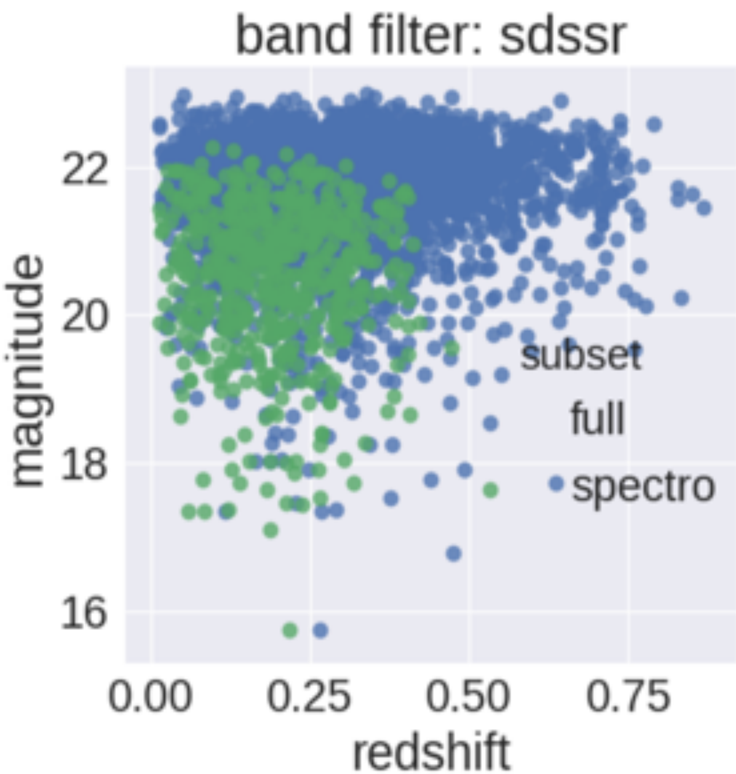
- *Naive Bayes (NB)*
- *K-nearest neighbours (KNN)*
- *Support vector machines (SVM)*
- *Artificial neural networks (ANN)*
- *Boosted decision trees (BDT)*

Effect of non-representative training sets



All feature extraction methods / machine learning algorithms sensitive to non-representativeness in training set; investigate domain adaptation techniques (e.g. data augmentation)

Non-representative data



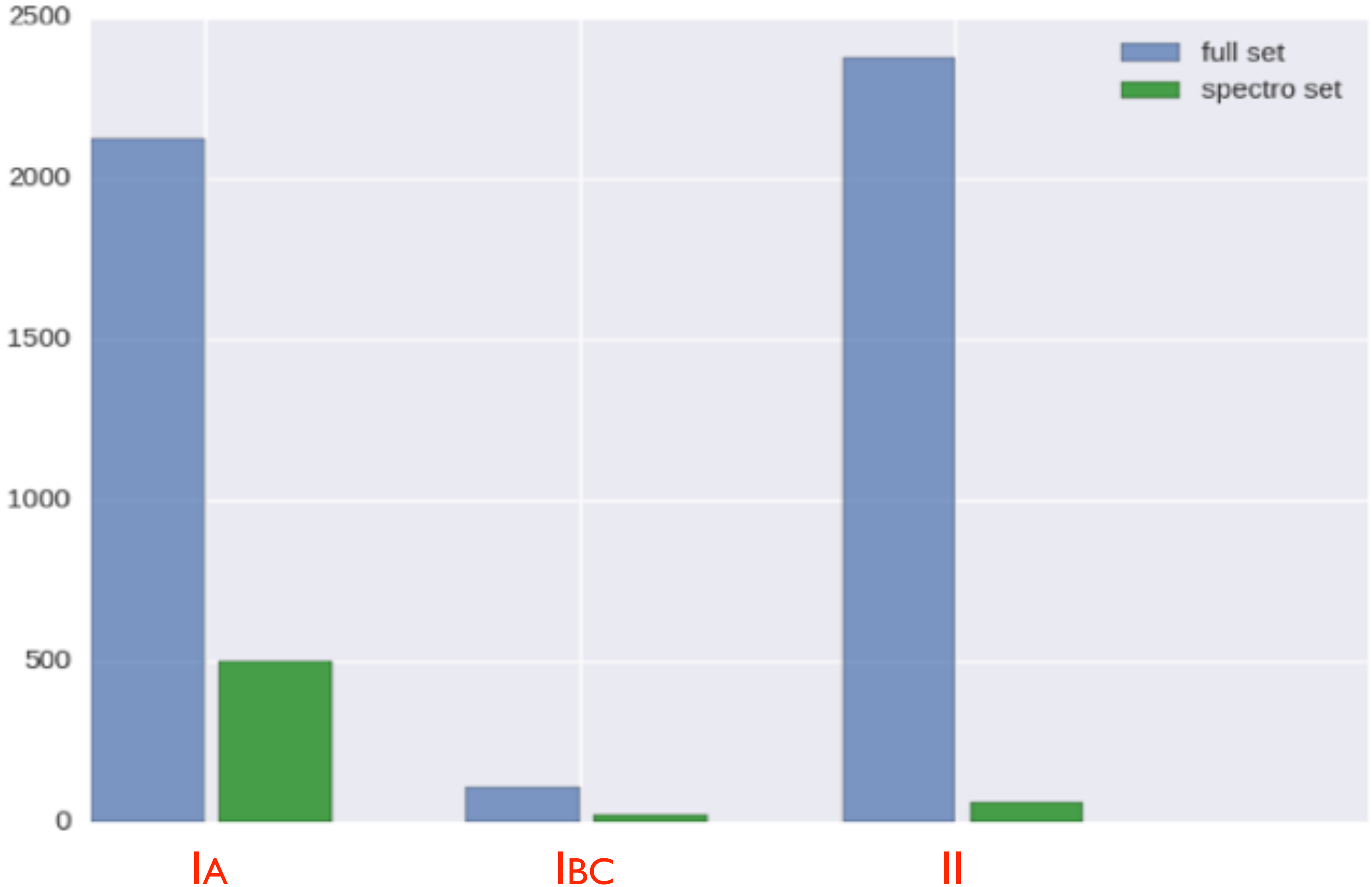
type	# in spectro	fraction	# in photo	fraction
Ia	500	85.9%	1625	40.4%
II	62	10.7%	2311	57.5%
Ibc	20	3.4%	86	2.1%
total	582	100%	4022	100%

Most training sets are non-representative in some way. Spectroscopic follow-up is always biased!

Example: SDSS supernova survey (classification: spectro followup or pSNId)

SCHUHMANN ET AL (IN PREP)

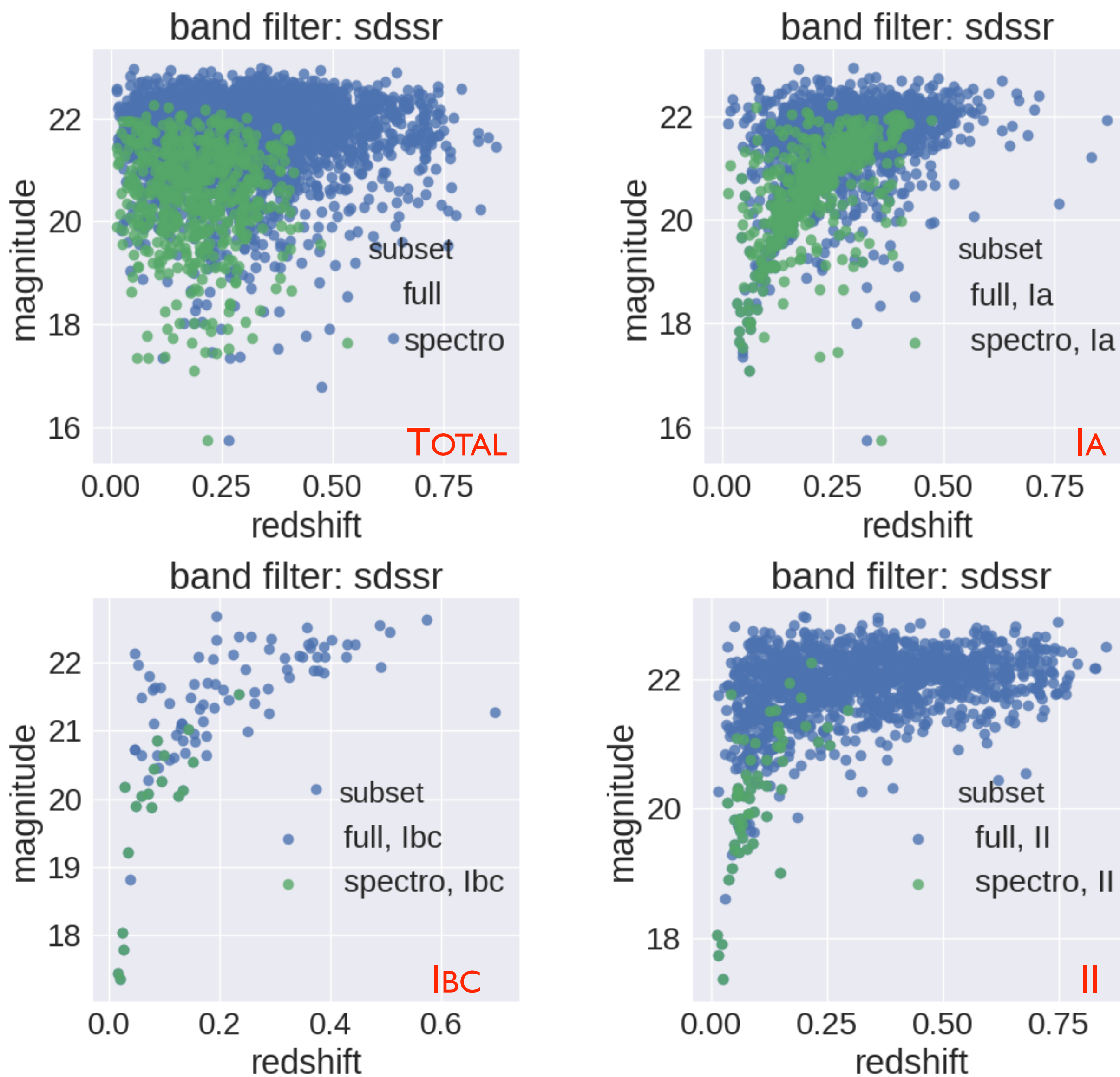
Class non-representativity



Followup bias (TACs tend not to award spectro time to non-Ia followup)

Schuhmann et al (in prep)

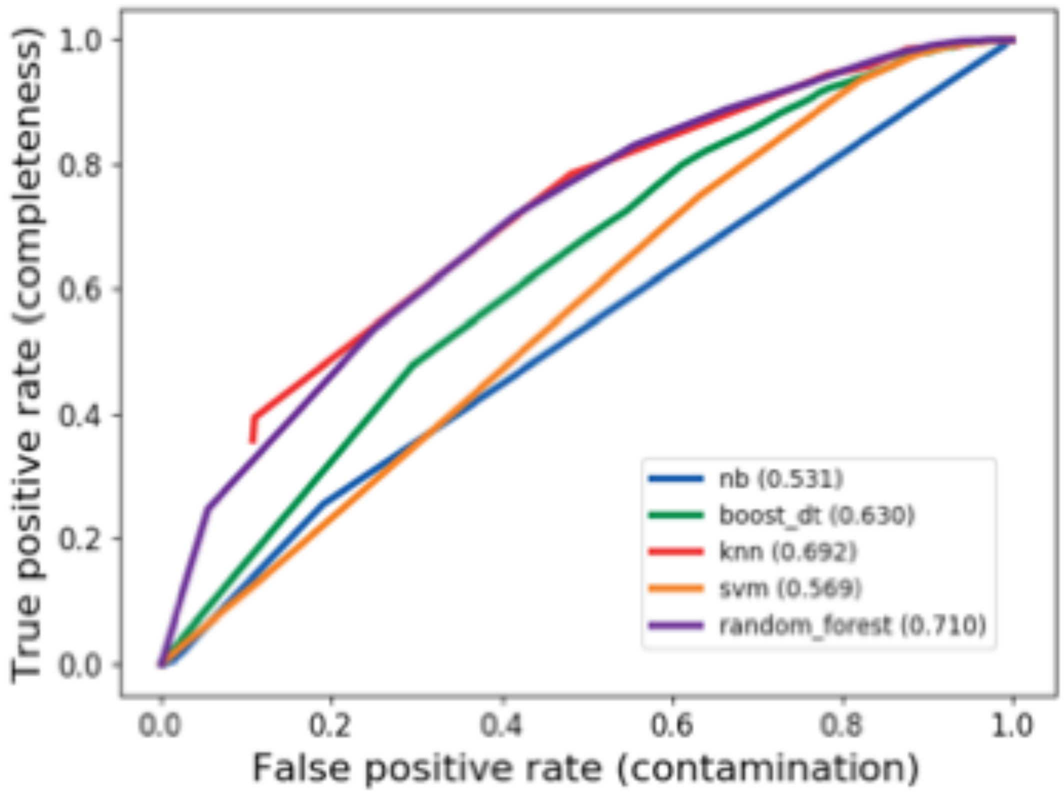
Feature non-representativity



Malmquist bias

Schuhmann et al (in prep)

Application of SNmachine to SDSS



Photometric classification performance limited by the spectro set.

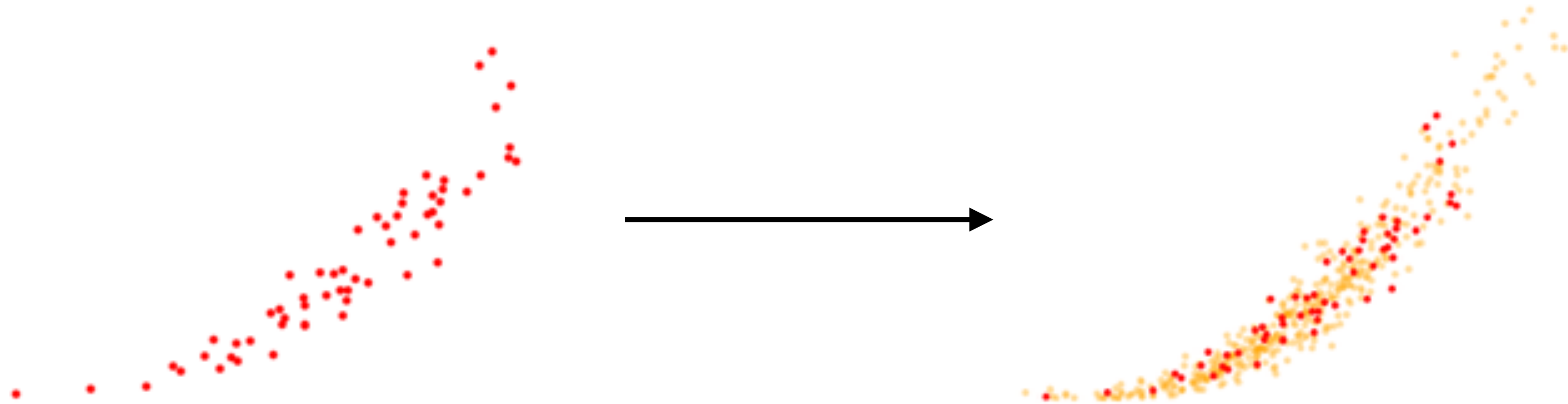
- ***class non-representativity:*** *non-Ia fraction underrepresented in spectro set, cannot map their intrinsic variability*
- ***feature non-representativity:*** *magnitude and z cutoff*

SCHUHMANN ET AL (IN PREP)

Data augmentation

Standard technique in (supervised) machine-learning.

- *avoid overfitting*
- *increase robustness of classifier to training data*
- *improve coverage of training data*

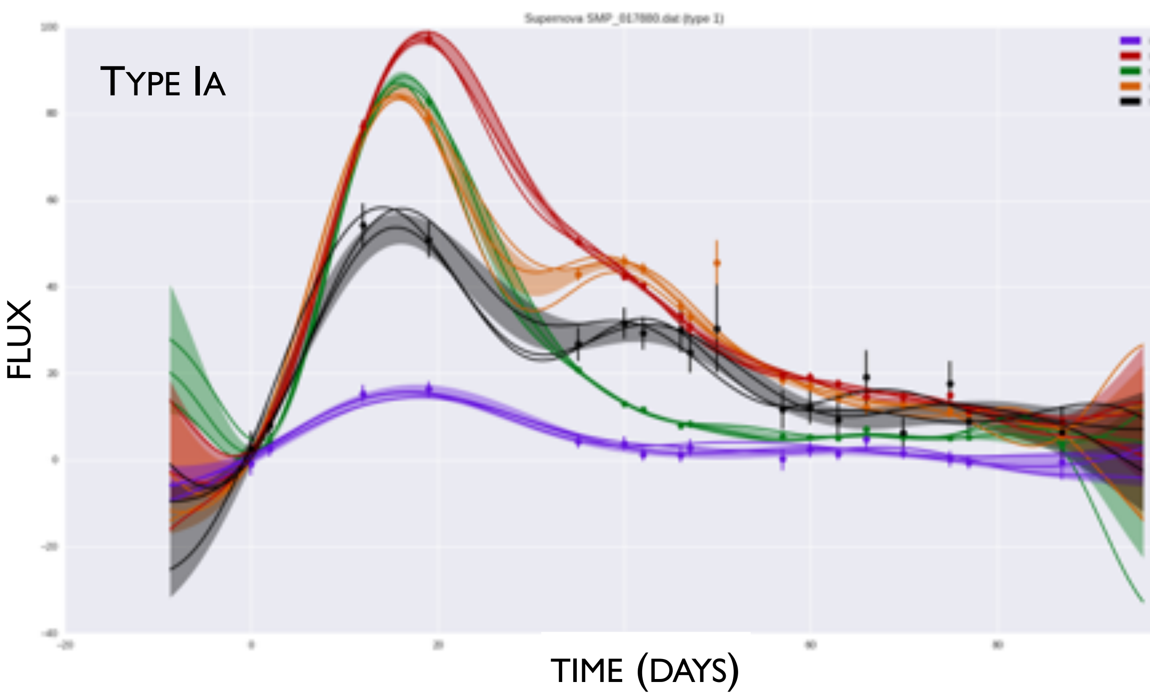


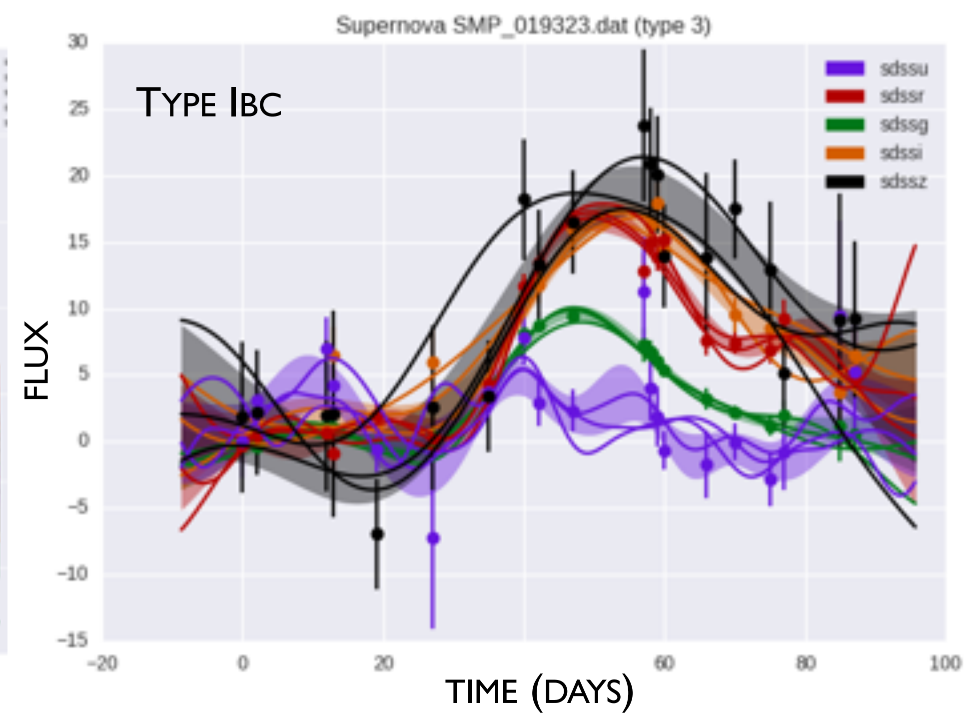
SCHUHMANN ET AL (IN PREP)

Example 1: Gaussian Process augmentation

***Idea**: bootstrap existing data to generate an augmented training dataset. Resample directly from light curves.*

- *fit GP to each band; draw samples with desired cadence*
- *cure class non-representativity without assuming non-Ia model?*
- *likely cannot solve feature non-representativity*





SCHUHMANN ET AL (IN PREP)

Example 2: Pure simulation augmentation

- *Can use our extensive knowledge of supernovae to augment the training data purely with simulations (correct both class and feature bias)*

IMAGE: SDSS

Data augmentation with simulations

- *Training set: pure simulations! interpolation and extrapolation of training sets. Control over:*
  - relative cluster size
  - absolute cluster size
  - intrinsic variability of every class
  - selection cuts
- *Accurate classifier training requires:*
  - reliable simulations of Ia **and** non-Ia lightcurves
  - representative targeting of classes in spectroscopy (e.g. 4MOST in LSST era)

Non-Ia simulations

- *Cadence simulations incl. accurate Core Collapse SNe templates*

Collaborating with Rob Firth, Szymon Prajs, Mark Sullivan at Southampton

https://github.com/UoS-SNe/CoCo (will be publicly released)

Approach: CoCo

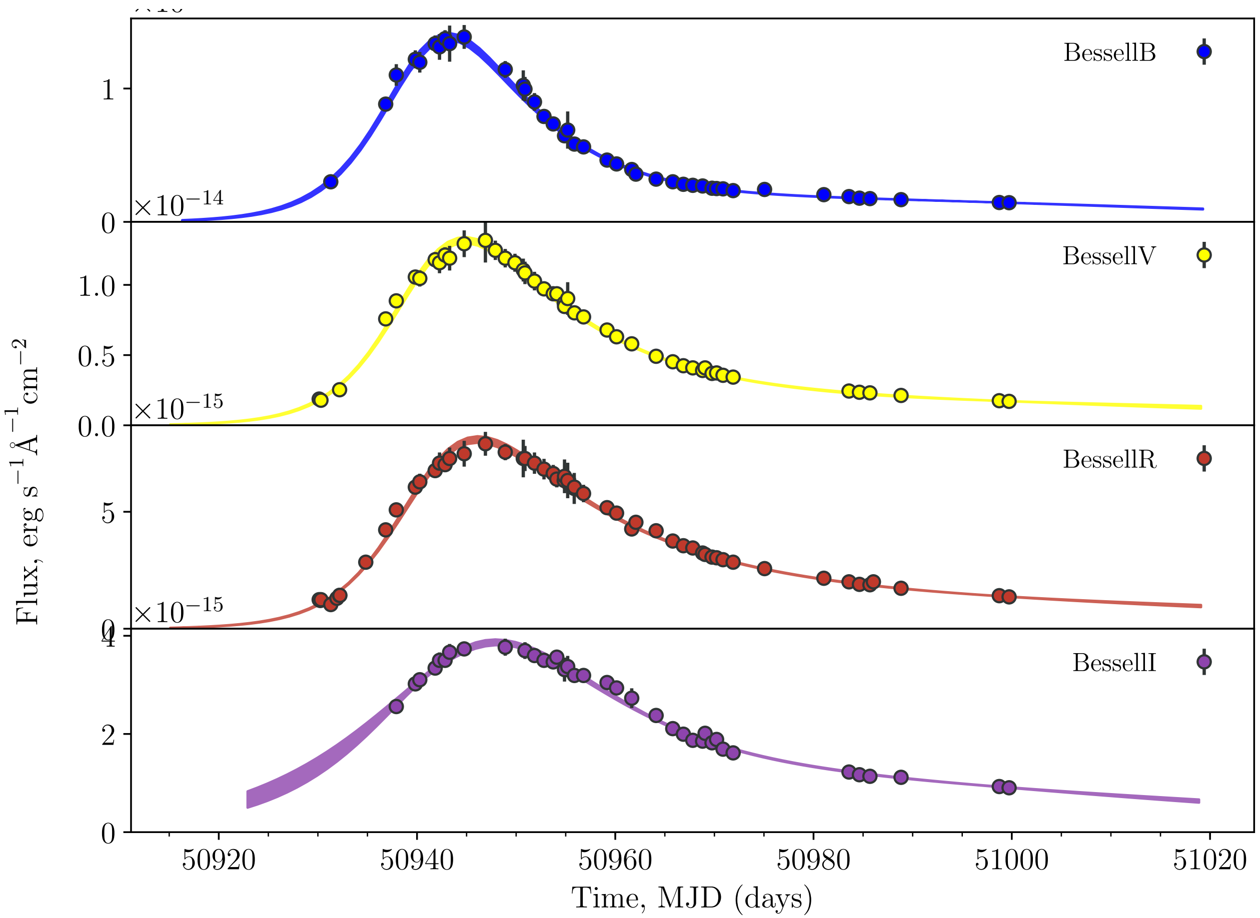
- Assemble sample of SESNe – Spectra and Photometry
- Fit light curves SN-by-SN, filter-by-filter
- Mangle spectra (see eg. Hsiao et. al. 2007, Conley et. al. 2008)
- Spline order is Nfilters+2
- Correct for MW extinction
- Use adjusted spectra to generate spectrophotometry
- Fit this synthetic data with LC function to cover all epochs
- Preserve (normalised) z=0 template and mangling function
- Use Luminosity Function to generate LCs (currently Li et. al. 2011)

Data sample

- 29 Stripped Envelope SNe, split into:
  - *12 SNe Ib*
  - *9 SNe Ic*
  - *6 SNe IIb*
  - *2 SNe with intermediate classifications (Ib/c & II/Ib)*
- $9 \leq$ N(spectra) ≤ 59
- 17/29 use data from the CfA sample

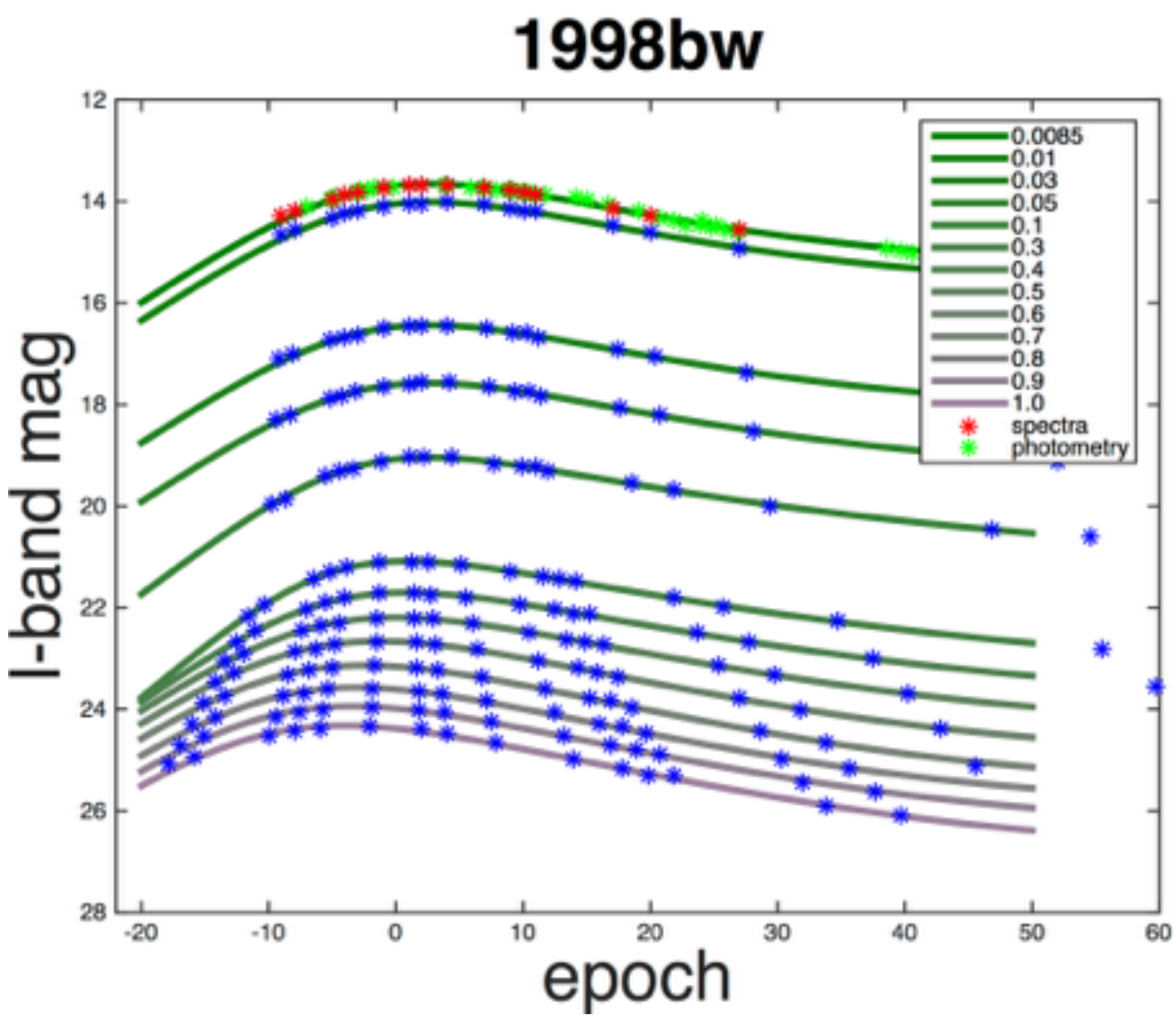
  (Modjaz et. al. 2014, Bianco et. al. 2014)
- That's all the data there is!

Example: SN1998bw



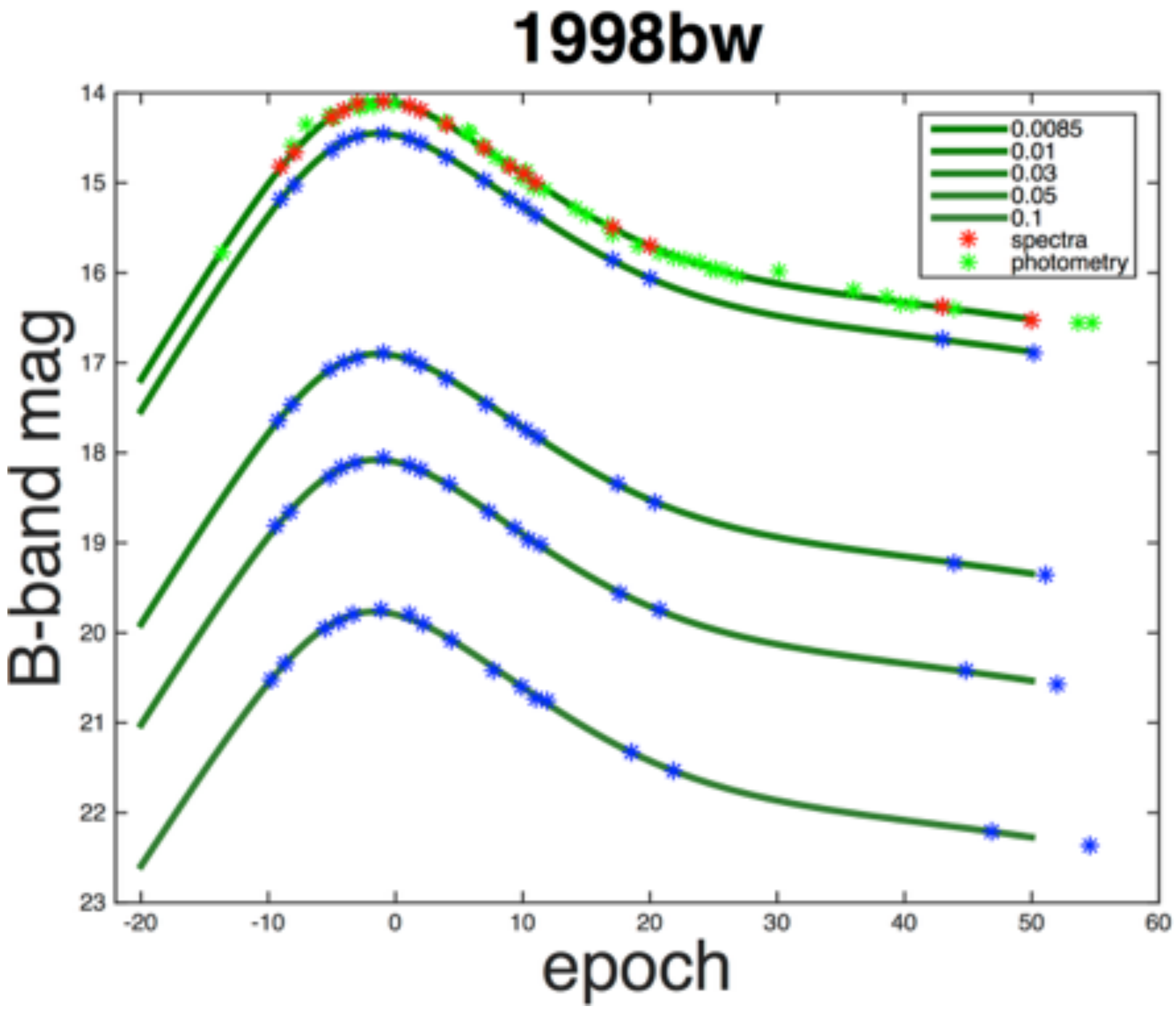
- BVRI Light curves
- Fit in flux-space
- Fit using MultiNest
- Full propagation of uncertainty
- Each band fitted independently
- Very good spectral coverage
- Do not need all bands covered in all spectra
- Need at least 2 for mangle

Usage: Simulation



- Can take SN1998bw to $z \approx 1.0$ in I-band

Usage: Simulation



- Can only get to z≈0.1 in B-band
- Need more Blue-Optical and UV spectral data!

Work in progress

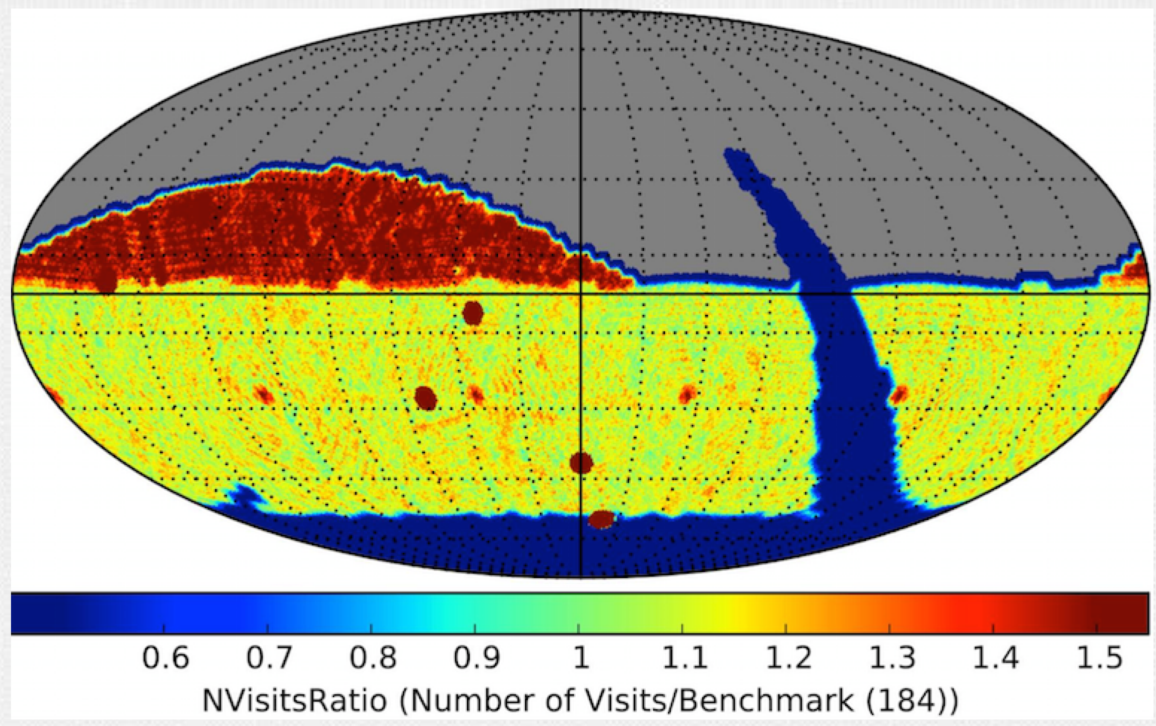
- *Do the simulated training sets represent the data well?*
- *What has higher impact - class bias or feature bias?*
- *How much non-Ia spectroscopy does GP augmentation need to eliminate class bias?*
- *Does our final strategy need GPs at all?*

Solving non-representativity problem in training data will likely require strategy with multiple ingredients.

The Survey

- Deep, Wide, Fast
  - Starting 2022
  - 18000+ deg^2
  - 10 years
  - 30s exposure per visit
  - ~825 visits per point
  - r~24.5/visit; r~27.5 total

About 0.000000000000000008 times the brightness of the full moon.



Figures: Ivezic et al, arXiv:0805.2366

ML in LSST survey strategy design

Baseline Wide-Fast-Deep Uniform cadence:

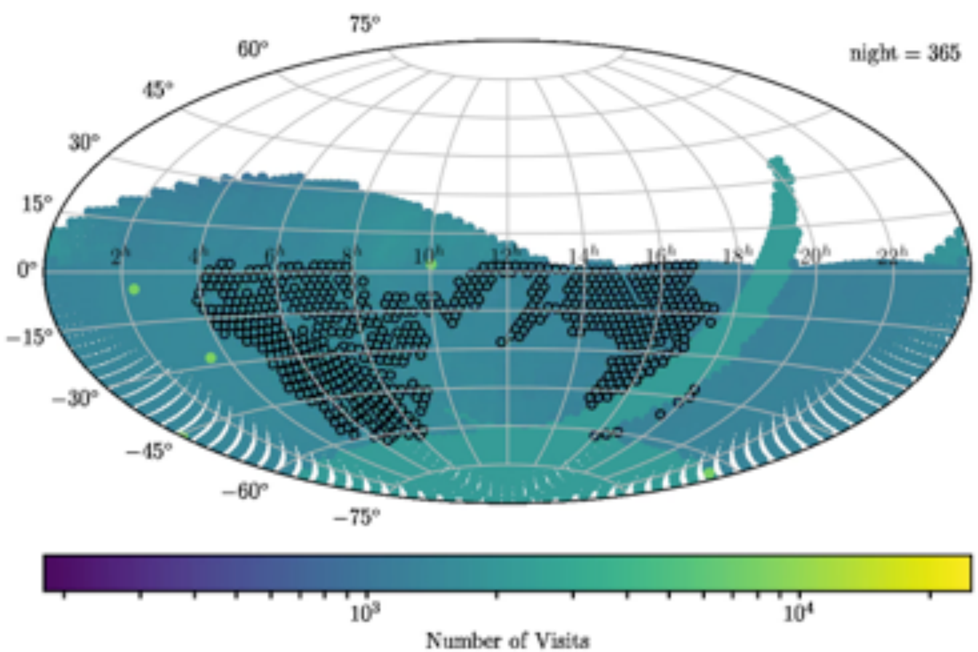
With ~ 800 visits spaced approximately uniformly over 10 years (distributed among 6 filters), not clear that LSST can offer sufficiently dense time sampling for study of transients with typical durations less than or ≃ 1week. Particularly a concern for key science requiring well-sampled SNIa light curves.

WFD Rolling cadence:

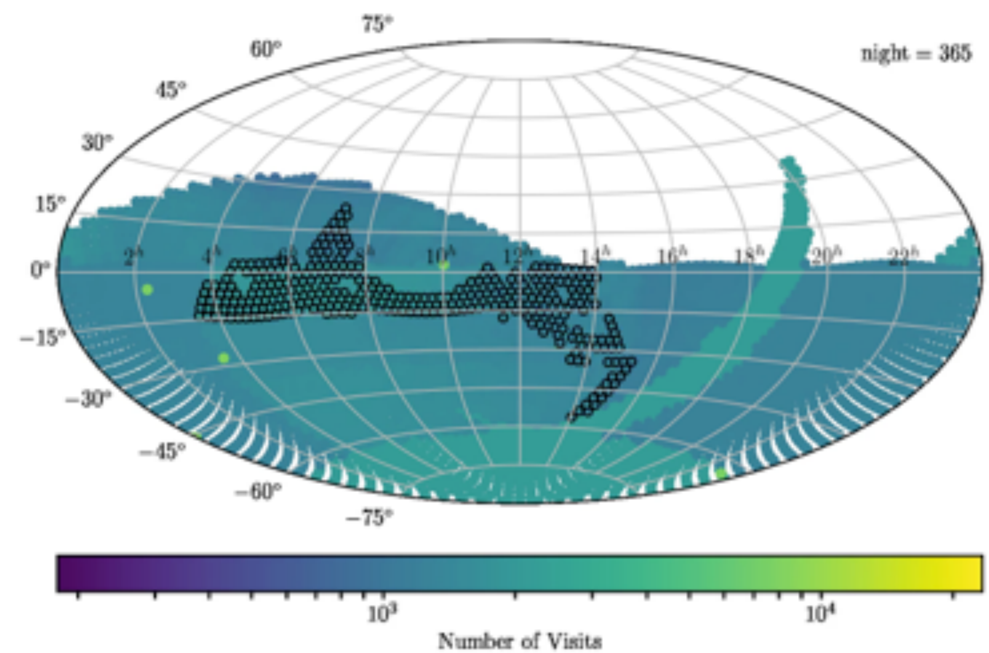
Enhanced sampling over selected sky area, rotating selected area in to exercise enhanced sampling over all the survey area part of the time, returning to balance at end of survey.

WFD Rolling Cadence proposal

Sampling rate about three times higher than uniform sampling implemented in baseline cadence (revisit time scale of about one day), and lasting 3-4 months, is suggested by SNe.



WFD baseline strategy



A rolling WFD proposal

MOVIES: ROB FIRTH

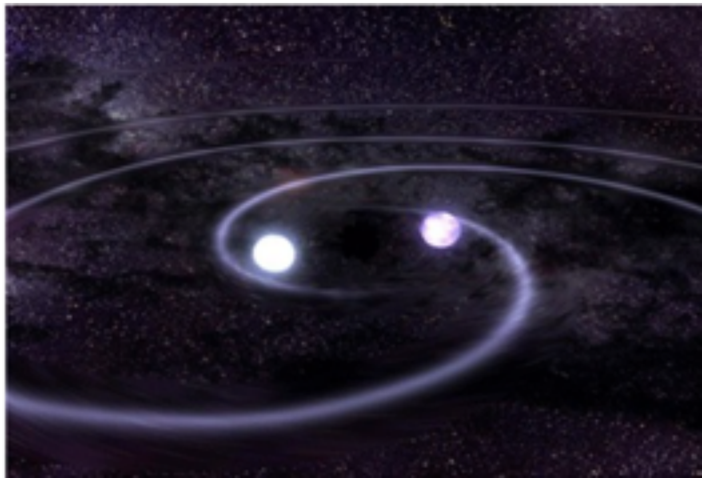
Machine learning for EM counterparts of GW sources

- *Large sky localisation means many potential electromagnetic counterparts, esp. in LSST / SKA era (known unknowns).*

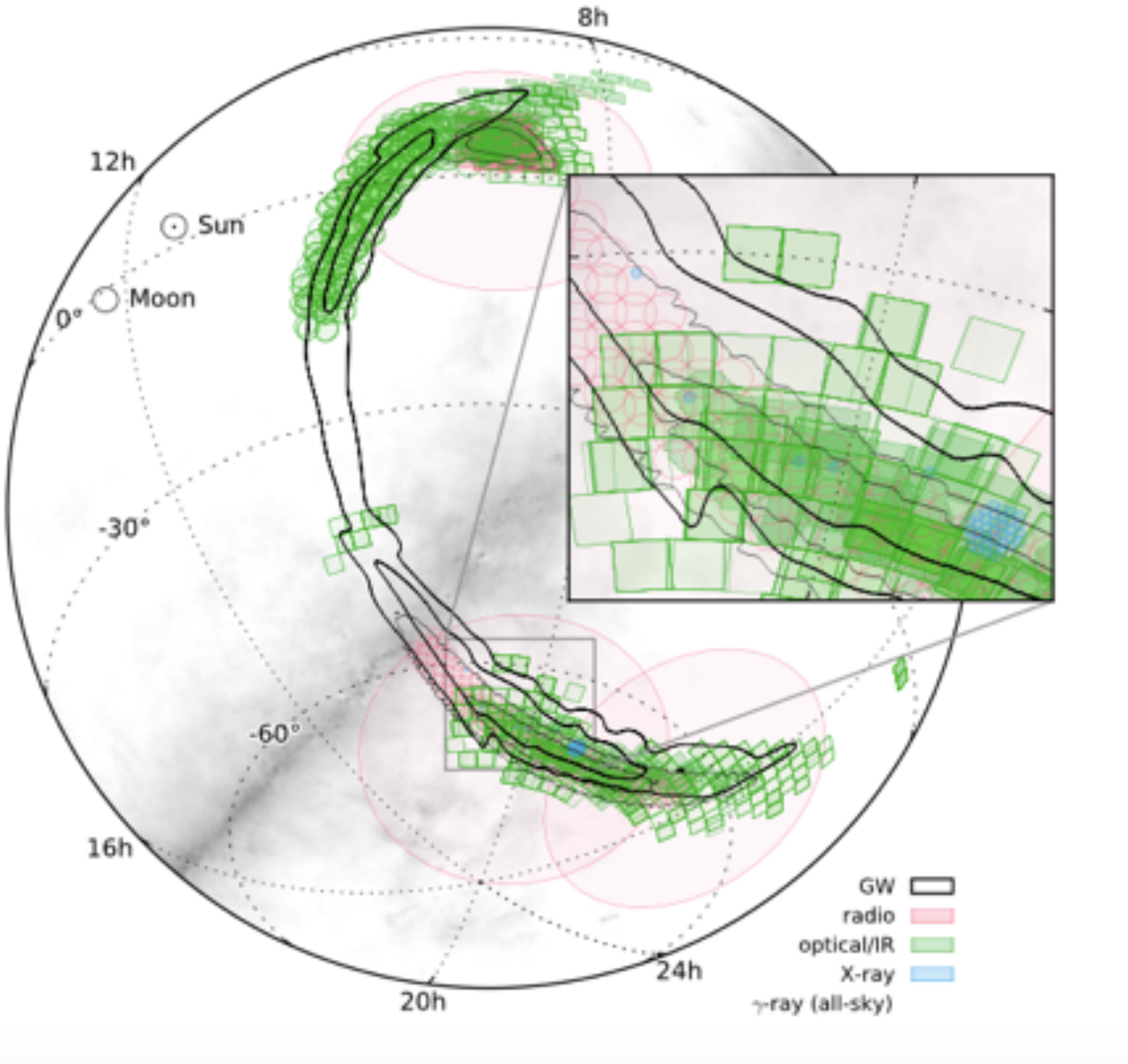
- *Uncertainty in what kind of counterparts to expect (unknown unknowns).*

- *Need to trigger follow-up based purely on photometry.*

- *Machine learning: work on SNe classification directly transferable to both scenarios.*



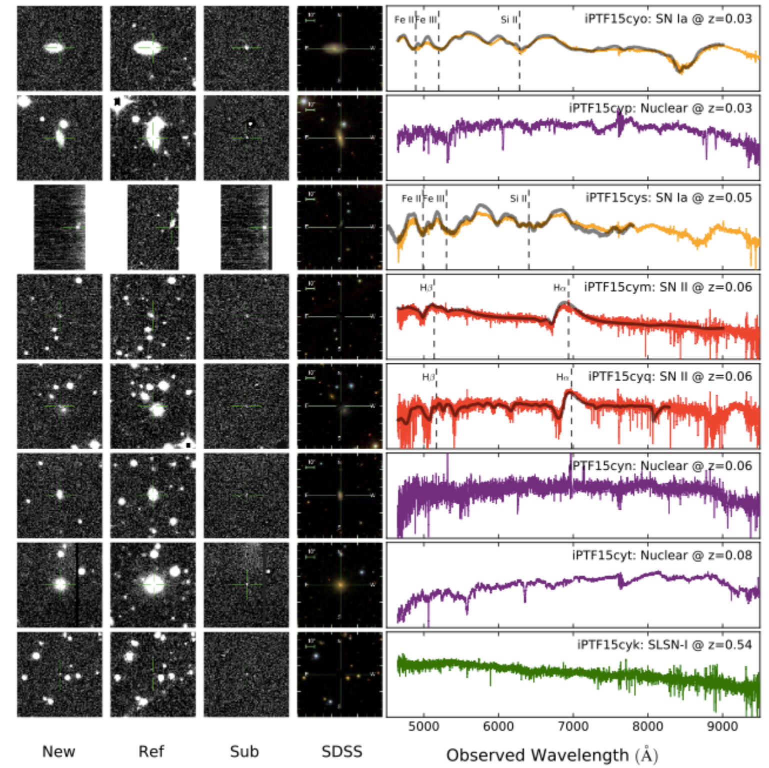
Sky localisation and followup for GW150914



- *90% credible region 630 sq. deg.*
- *10^5 galaxies > $0.1L_*$ within the comoving volume of 10^{-2} Gpc^3 within this region + 90% CL source distance.*
- *aLIGO+VIRGO would have given localisation to 10s of sq.deg.*
- *VIRGO joined "observation run 2" (O2) on 1 Aug 2017!*

ABBOTT ET AL (2017) APJL

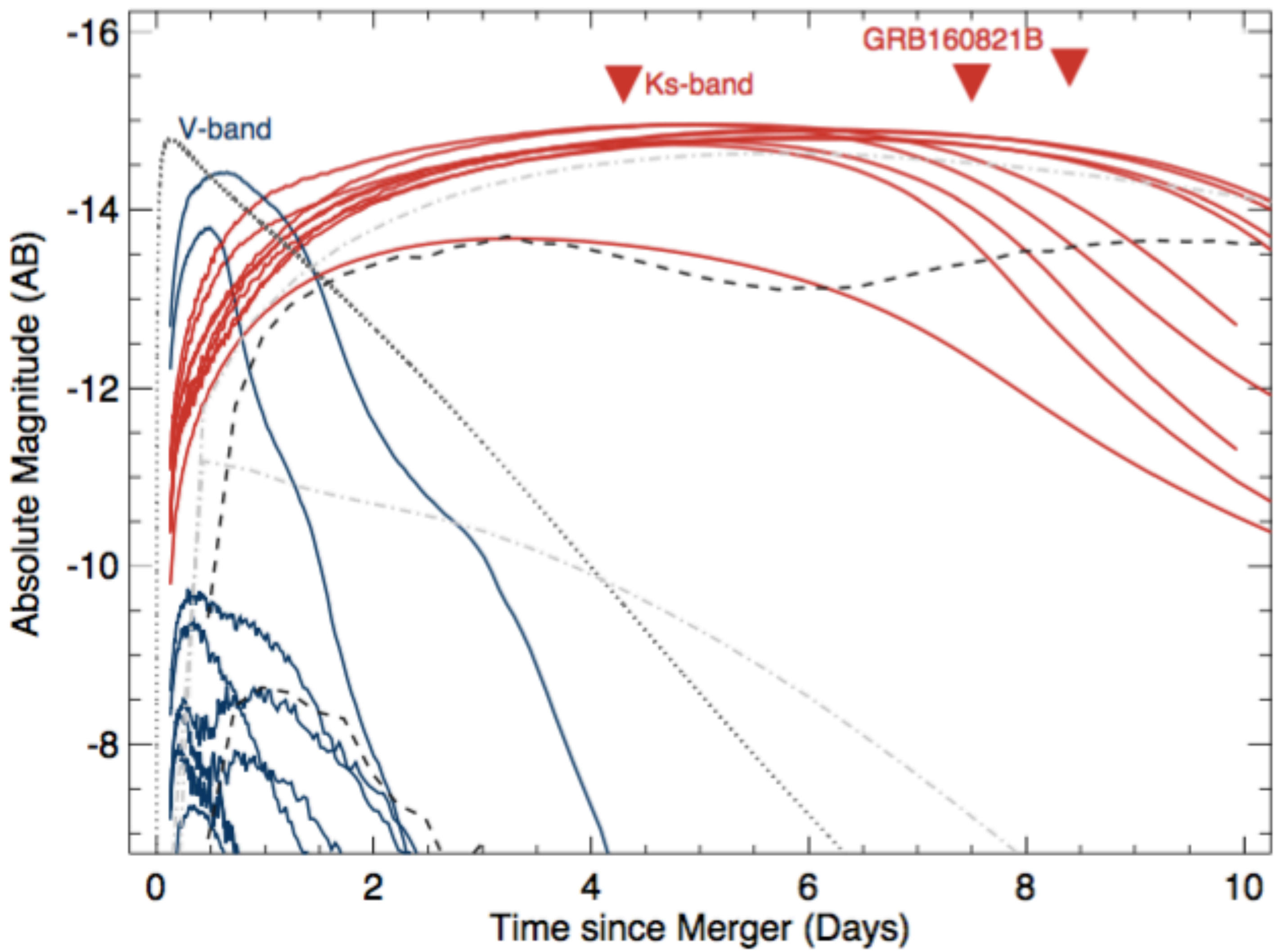
GW150914: needle in a haystack



- *127676 candidates in subtraction images*
- *78951 do NOT have a quiescent stellar source*
- *15624 are detected twice and NOT asteroids*
- *5803 pass machine learning threshold*
- *1007 are coincident with a nearby galaxy*
- *13 were vetted by human scanners*
- *8 were scheduled for follow-up spectroscopic observations*
- *0 were associated with the gravitational wave*

KASLIWAL ET AL (2016)

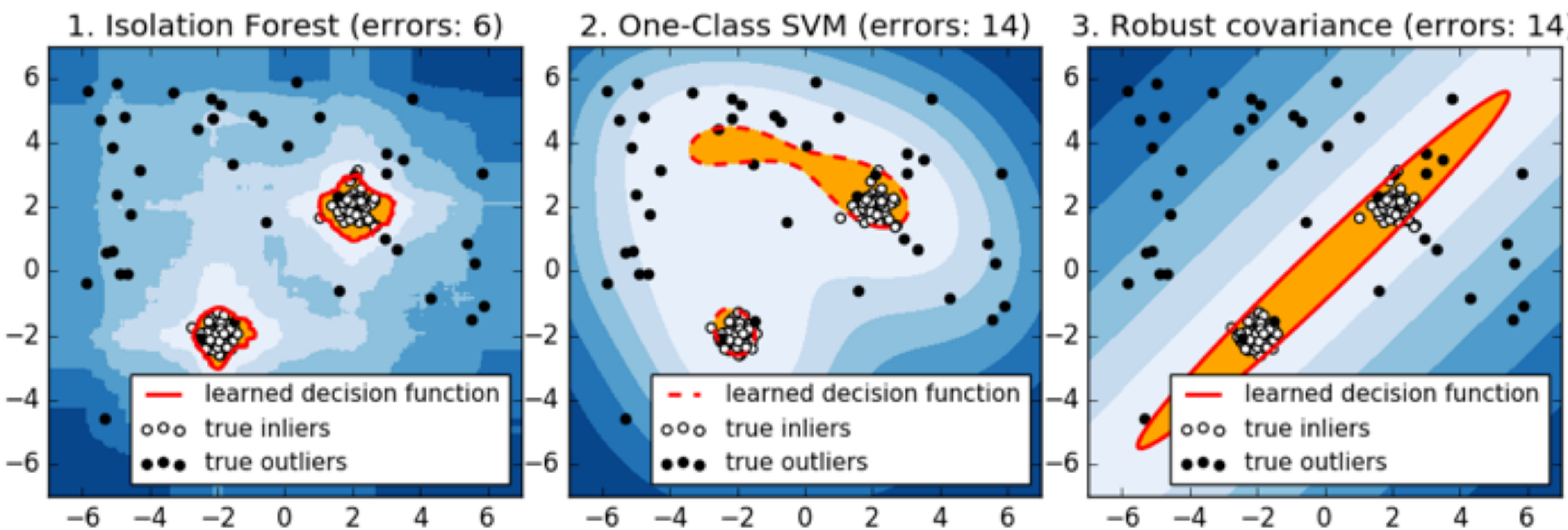
Fast Blue + Slow Red



Models: Wollaeger+ (2017), Metzger+ (2015), Barnes+ (2016), Rosswog+ (2016)

KASLIWAL ET AL (2017)

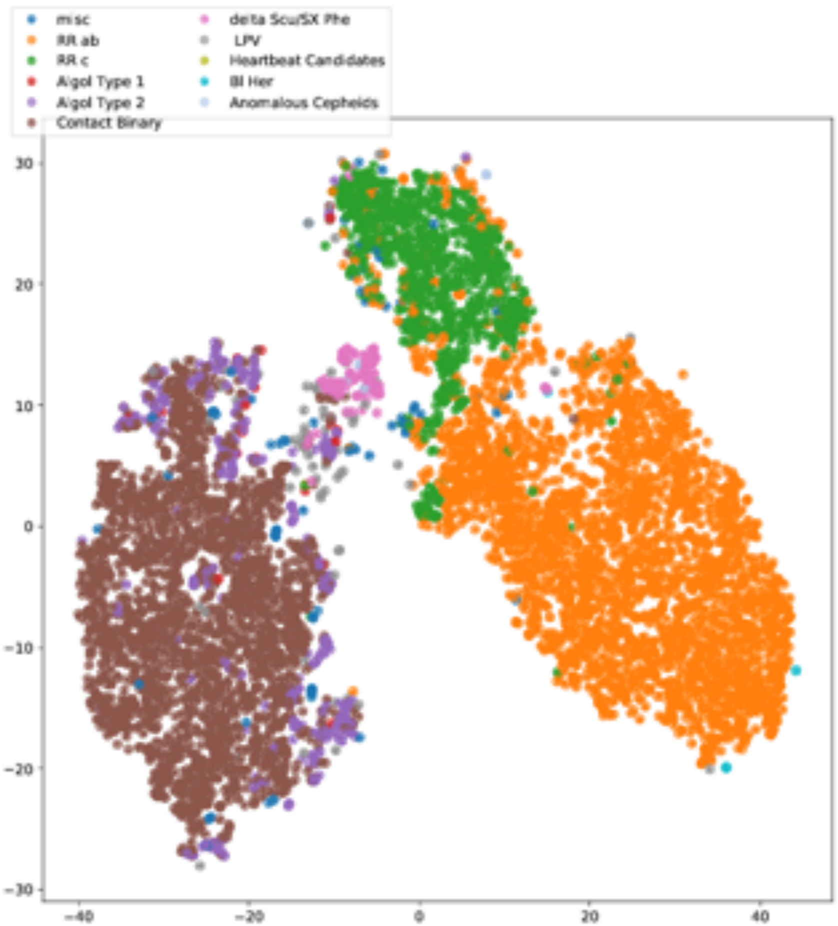
Classification and anomaly detection using ML



- *The methods developed for SNmachine can be used for general transient classification, relying on wavelet decomposition.*
- *Once you have a good description of your data (i.e. features), you can use machine learning for rapid anomaly detection.*
- *Both classification and anomaly detection critical for GW EM counterpart searches.*

FIGURE: SCIKIT-LEARN

General transient classification



- *t-SNE plot for different types of variable stars, decomposed using Gaussian processes, using SNmachine as classifier (Tayeb Zaidi & Gautham Narayan, private communication)*

G.R.E.A.T. @ Stockholm

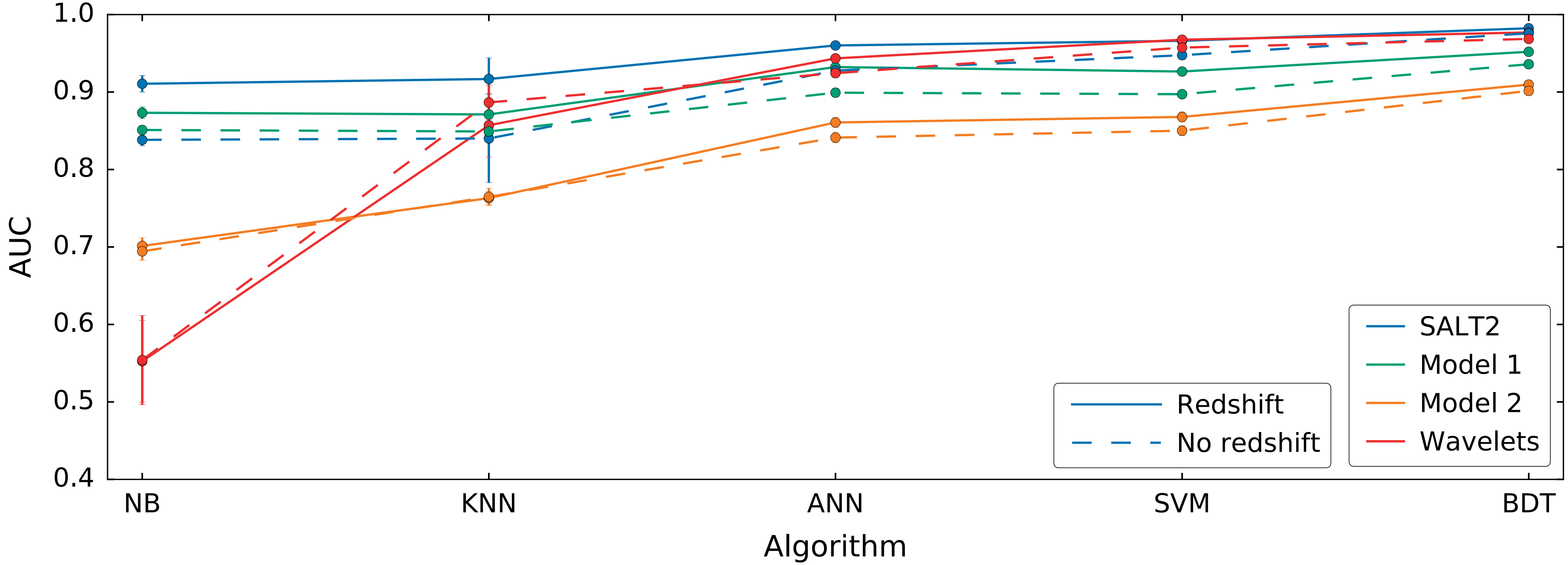
Gravitational Radiation and Electromagnetic Astrophysical Transients

- *6 year programme.*
- *Create end-to-end simulations of EM signals from compact object mergers.*
- *Use to optimize search strategies and perform searches for electromagnetic counterparts of GW events in ZTF and LSST.*
- *Join us! https://www.great.cosmoparticle.com*

HIRANYA PEIRIS, JESPER SOLLERMAN, STEPHAN ROSSWOG, AND ARIEL GOOBAR

COSMOPARTICLE, WWW.PENELOPEROSECOWLEY.COM

Effect of redshift information



When using BDT, SALT2 and wavelet features able to classify equally well with or without redshift.