Transient photometric classification: An astronomical data challenge

IN2P3 Workshop in Machine Learning , Lyon - 29 March 2018

Emille E. O. Ishida

CNRS/Laboratoire de Physique de Clermont - Université Clermont Auvergne
Clermont Ferrand, France

Overview

1. Astronomical transients

2. How are they classified ?

3. Data issues

4. The road so far

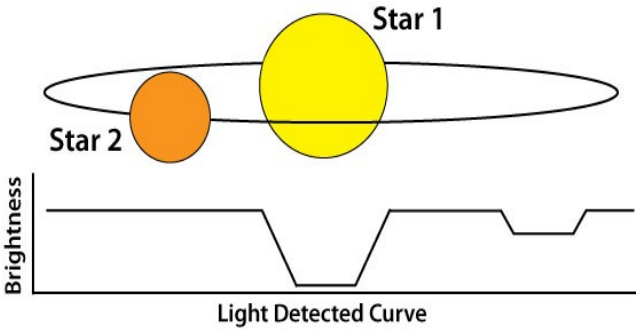
5. The Large Synoptic Survey Telescope (LSST)

6. PLAsTiCC – Photometric LSST Astronomical Time-series Classification Challenge

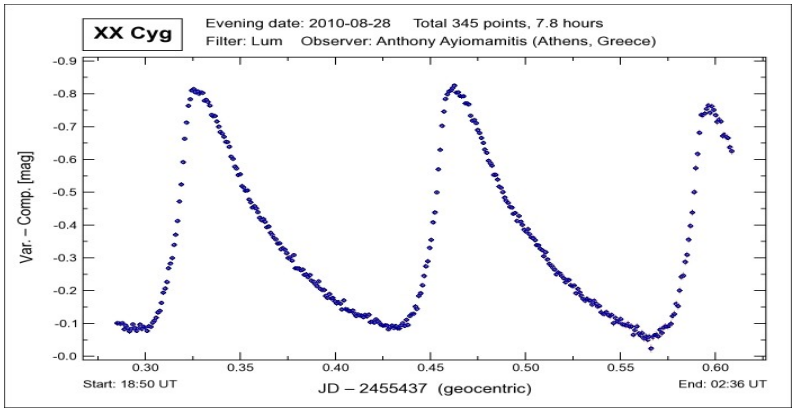
Astronomical Transients

1.Transients: time evolution from observer frame

Eclipsing binaries

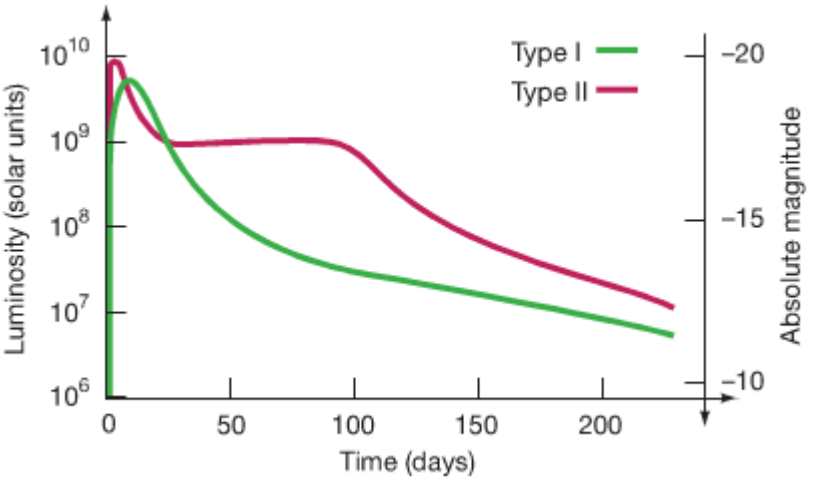


Pulsating

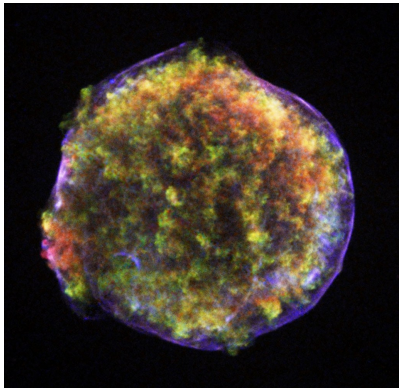


Cataclysmic

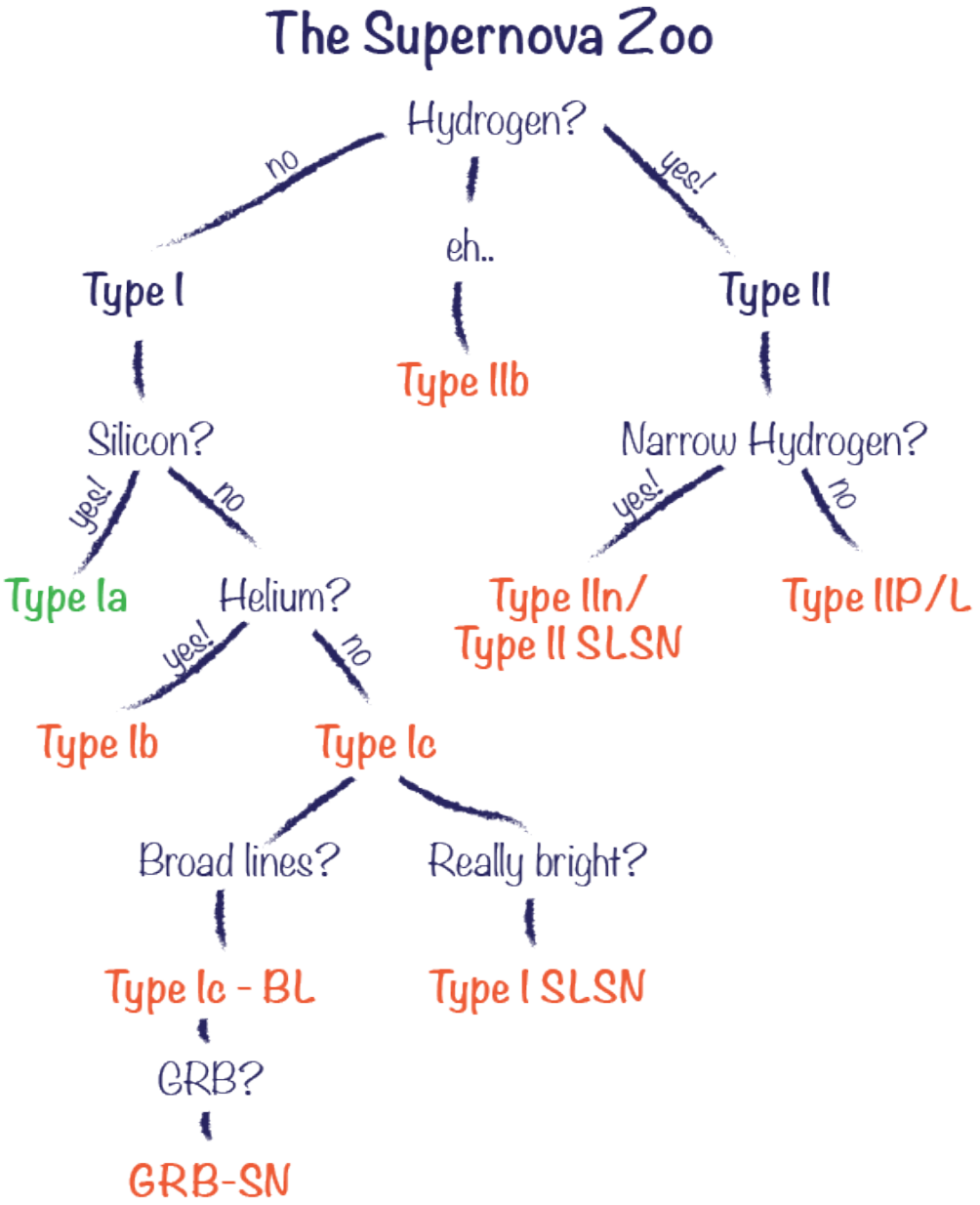
Supernovae: Type Ia



Supernovae: Type II



1.Classical example: Supernovae



it's complicated

https://astrobites.org/2016/12/02/classifying-supernovae/

1. Supernova: cosmological application

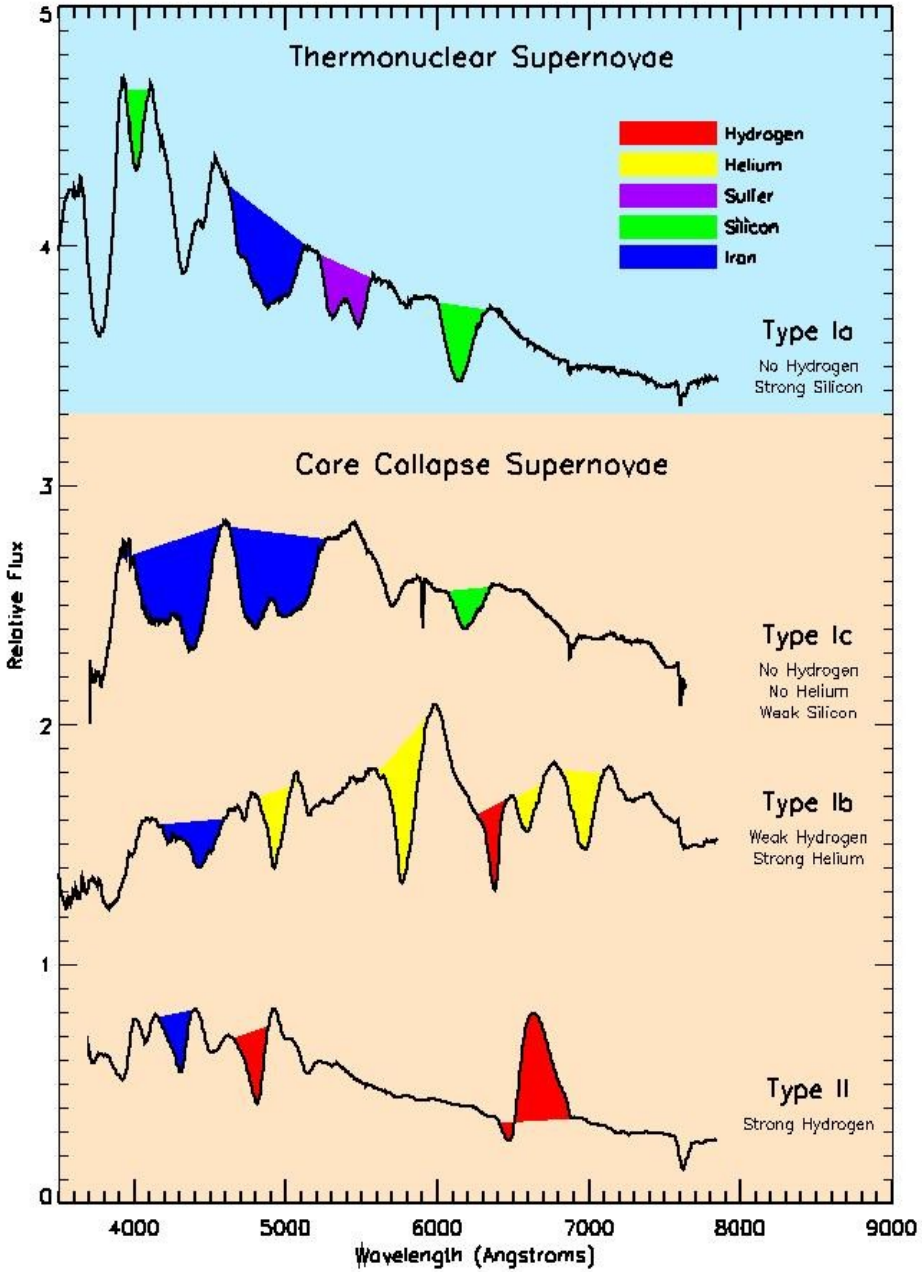
Only Ia can be used as distance indicators

2011

How are they classified?

2. Spectroscopy

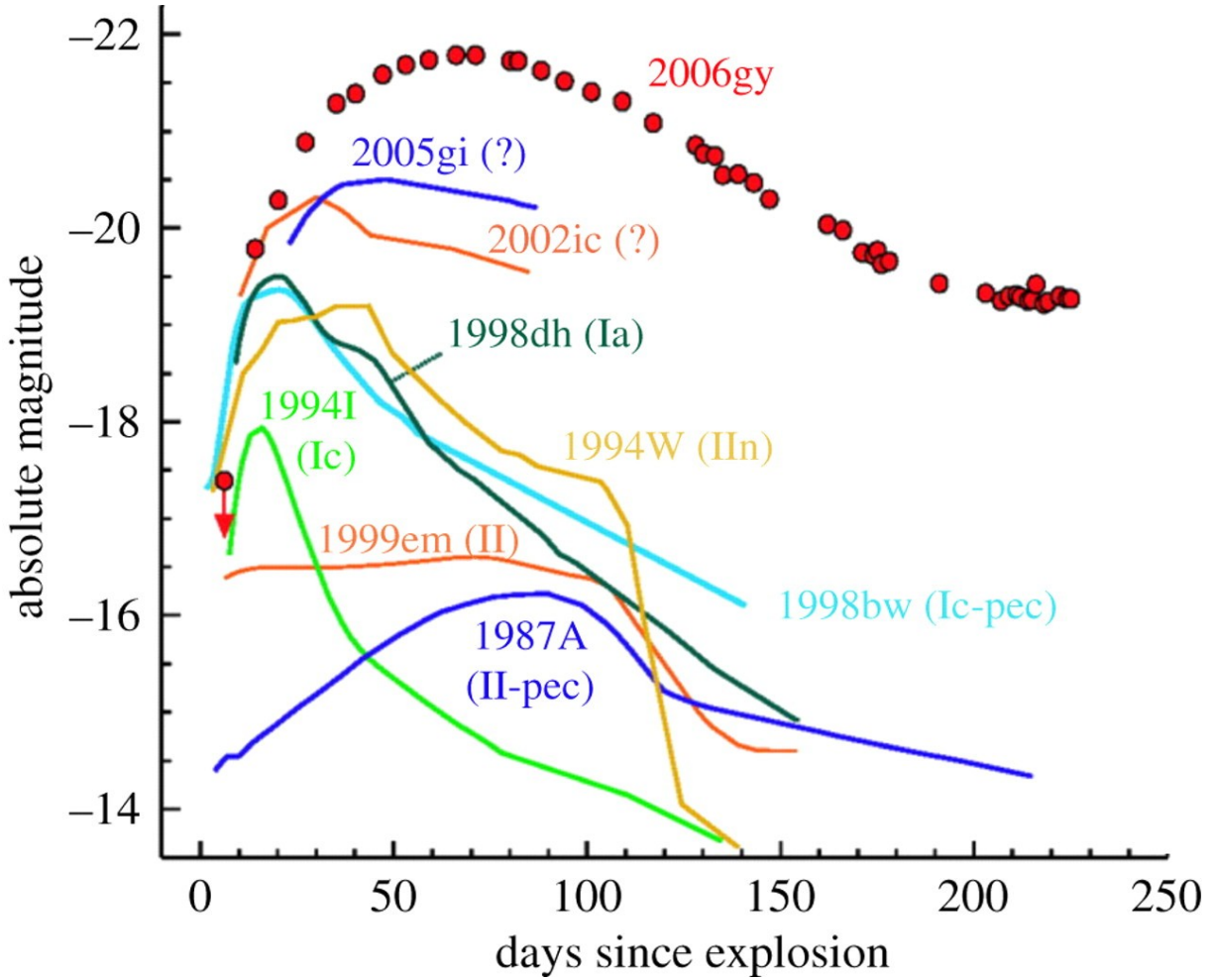
Spectra (high resolution snap-shot)



http://supernova.lbl.gov/~dnkasen/tutorial/

2. Photometry

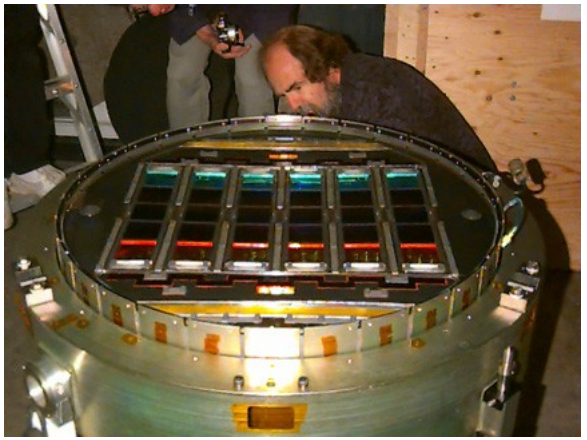
Light Curves (low resolution time evolution)



http://rsta.royalsocietypublishing.org/content/370/1960/774

Photometry x Spectroscopy

An example from SDSS



Exposure time 2 x 54s

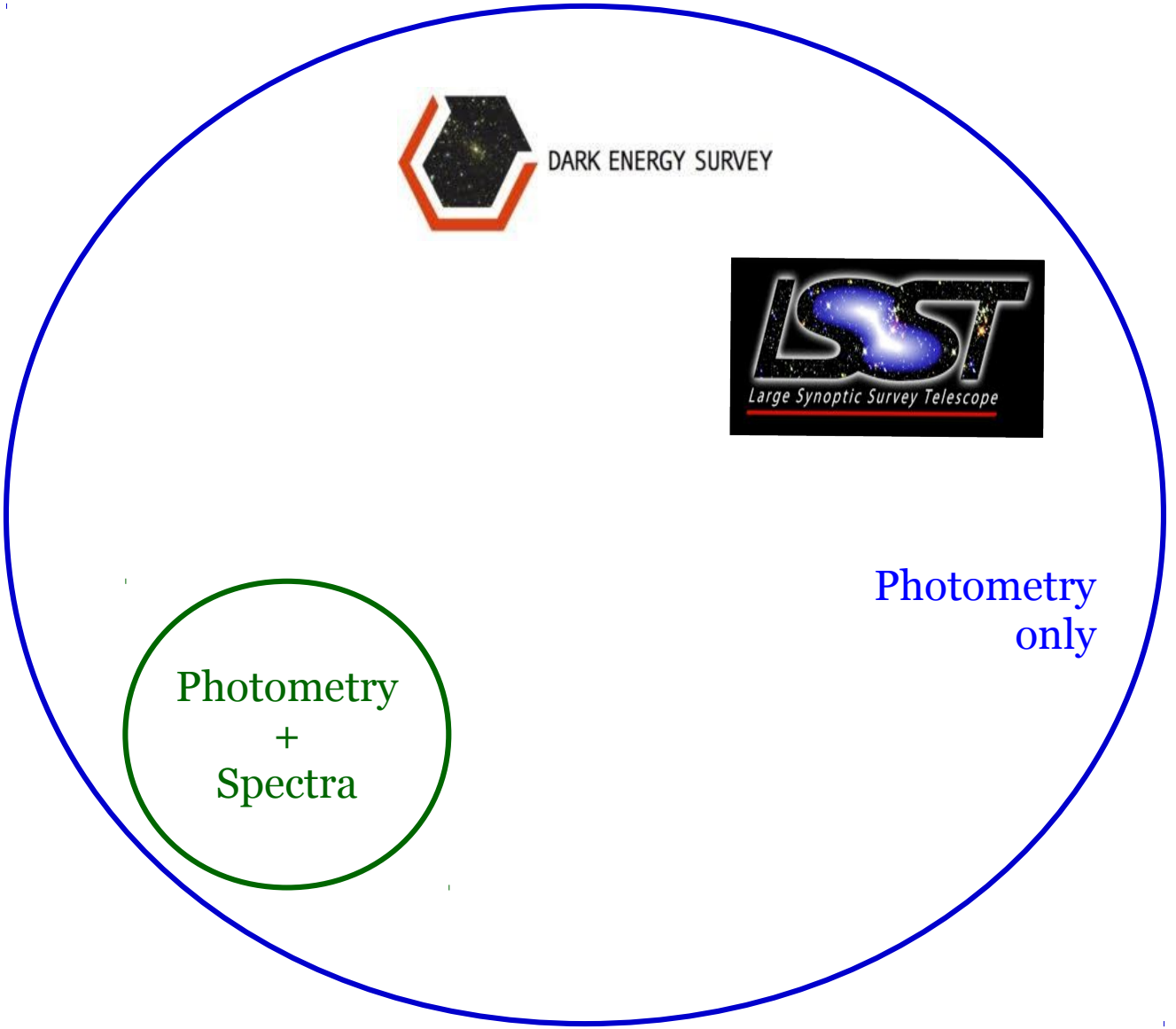
~ Integration at least 45 minutes

://www.preposterousuniverse.com/blog/2009/10/06/practicality-and-the-universe/

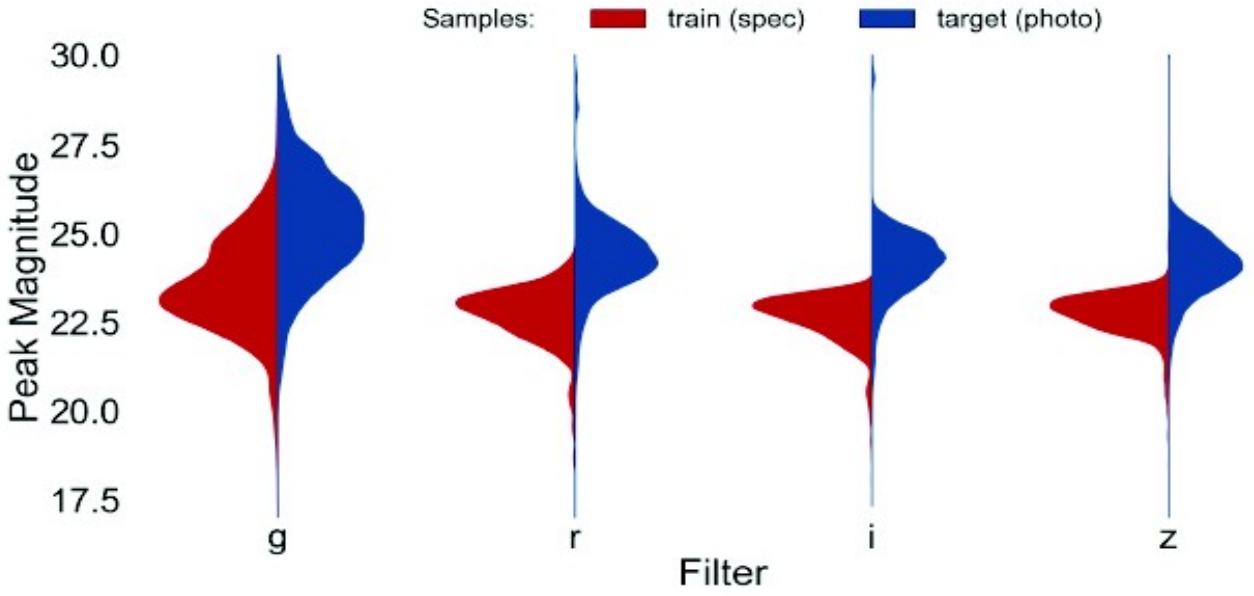
http://www.stsci.edu/~inr/bdpics/bd5.htm

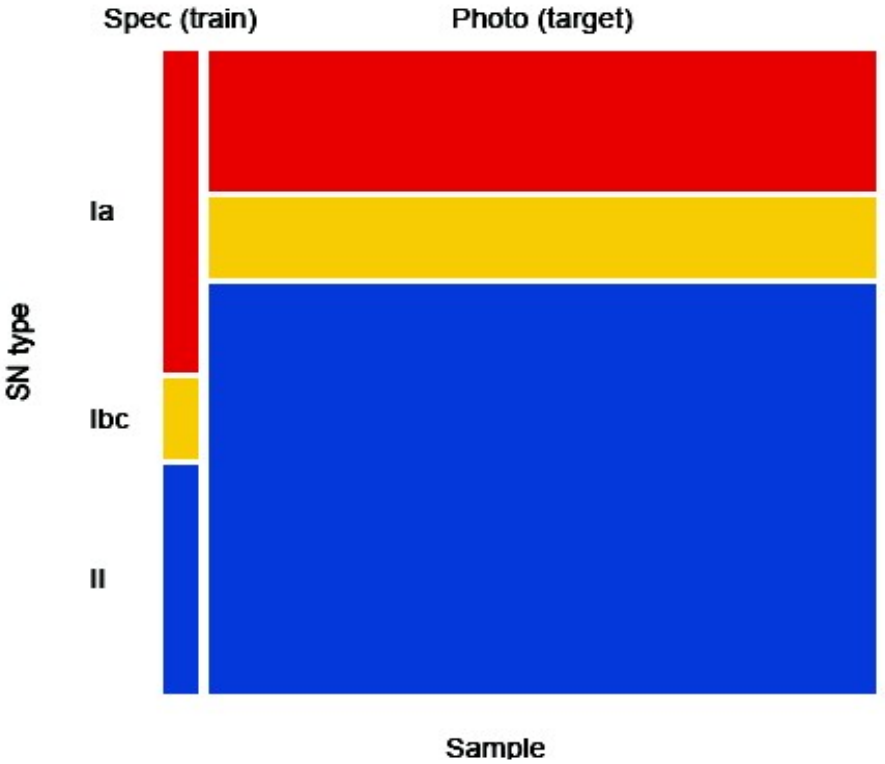
Data issues

3. Representativeness



3. Representativeness



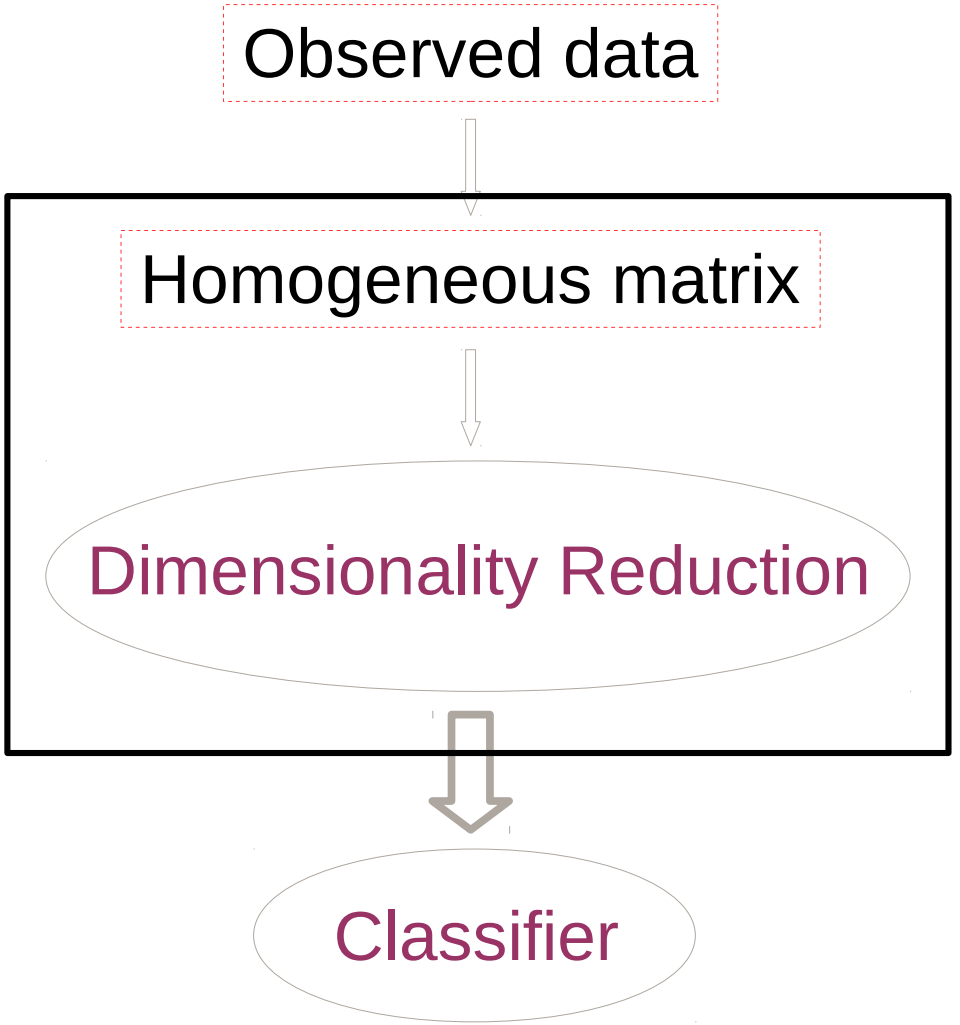


Data: Kessler et al., 2010 - SNPCC
Figure: From CRP #4 – Ishida et al, 2018 – in prep

The road so far

4. Strategies

Pre-processing



Parametric fit

4. Strategies: previous attempts

The first SN photometric classification challenge...

Kessler et al., 2010, PASP, Volume 122, Issue 898, pp. 1415

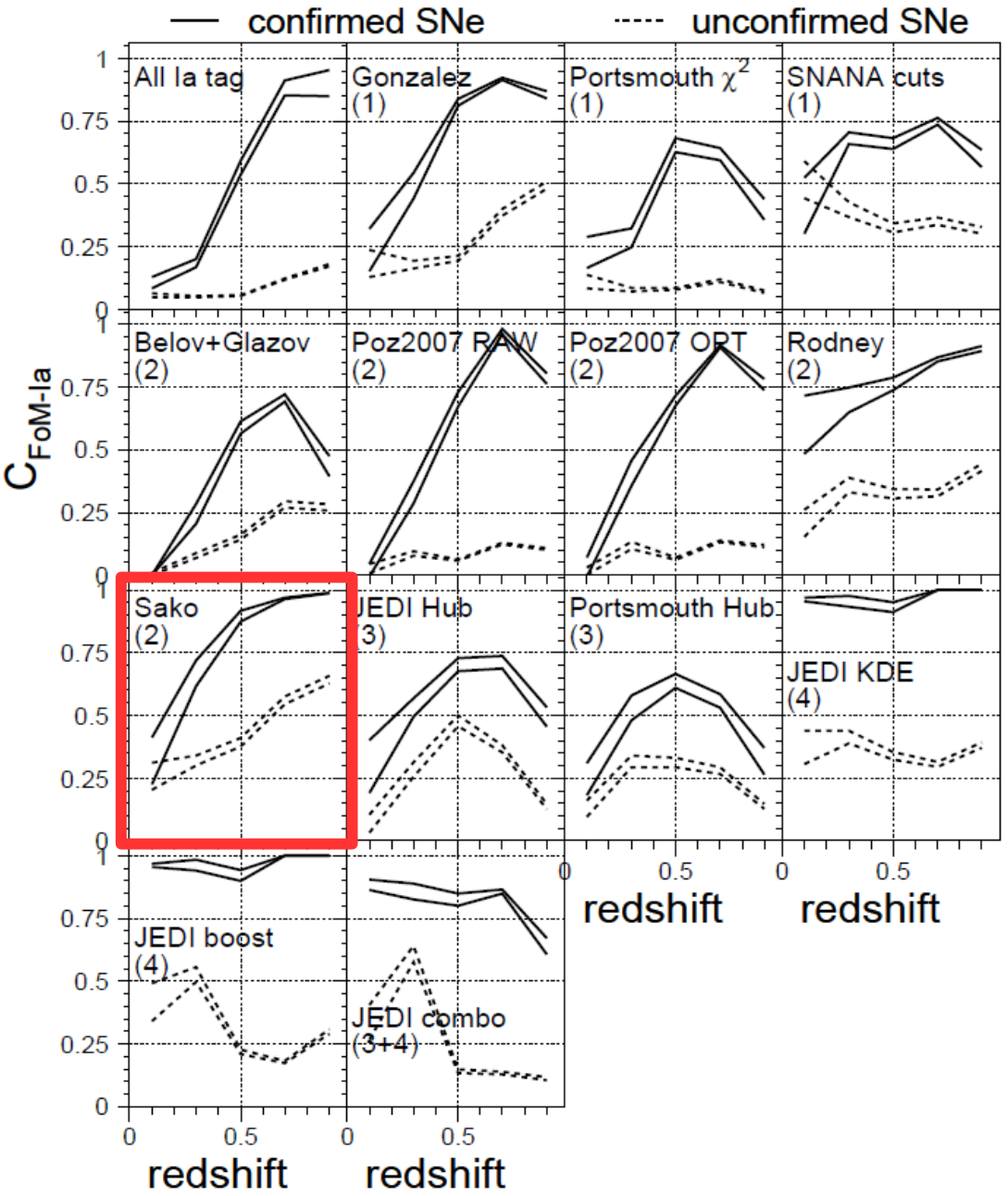
Simulations: Ia, Ib/c, II as expected from DES

Data size: ~20.000 objects, 1103 for training

Participants: 10 groups, 13 entries

4. Strategies: SNPCC results

Kessler et al., 2010, PASP, Volume 122, Issue 898, pp. 1415



Outcomes:

1 – None of the methods obviously outperformed the others

2 – SNID had better overall metric

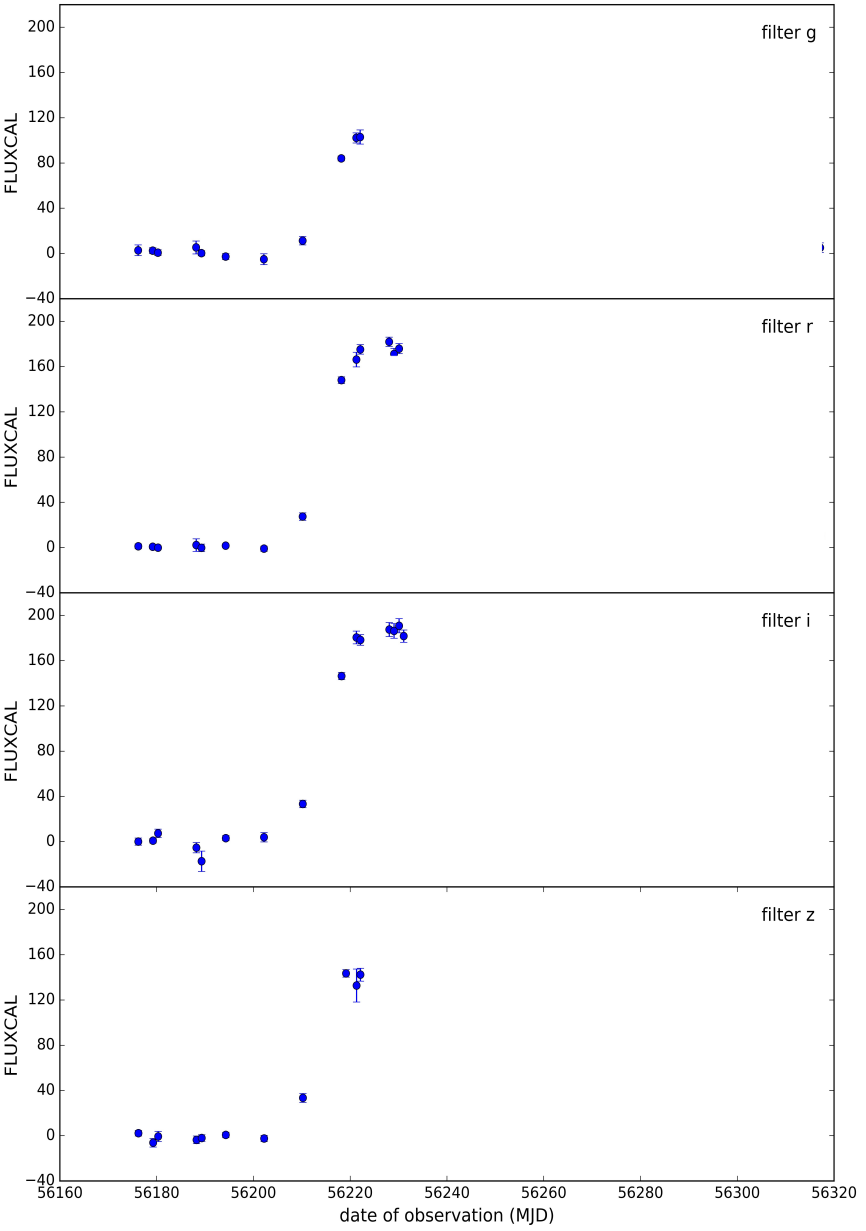
3- An updated data set was released to the community

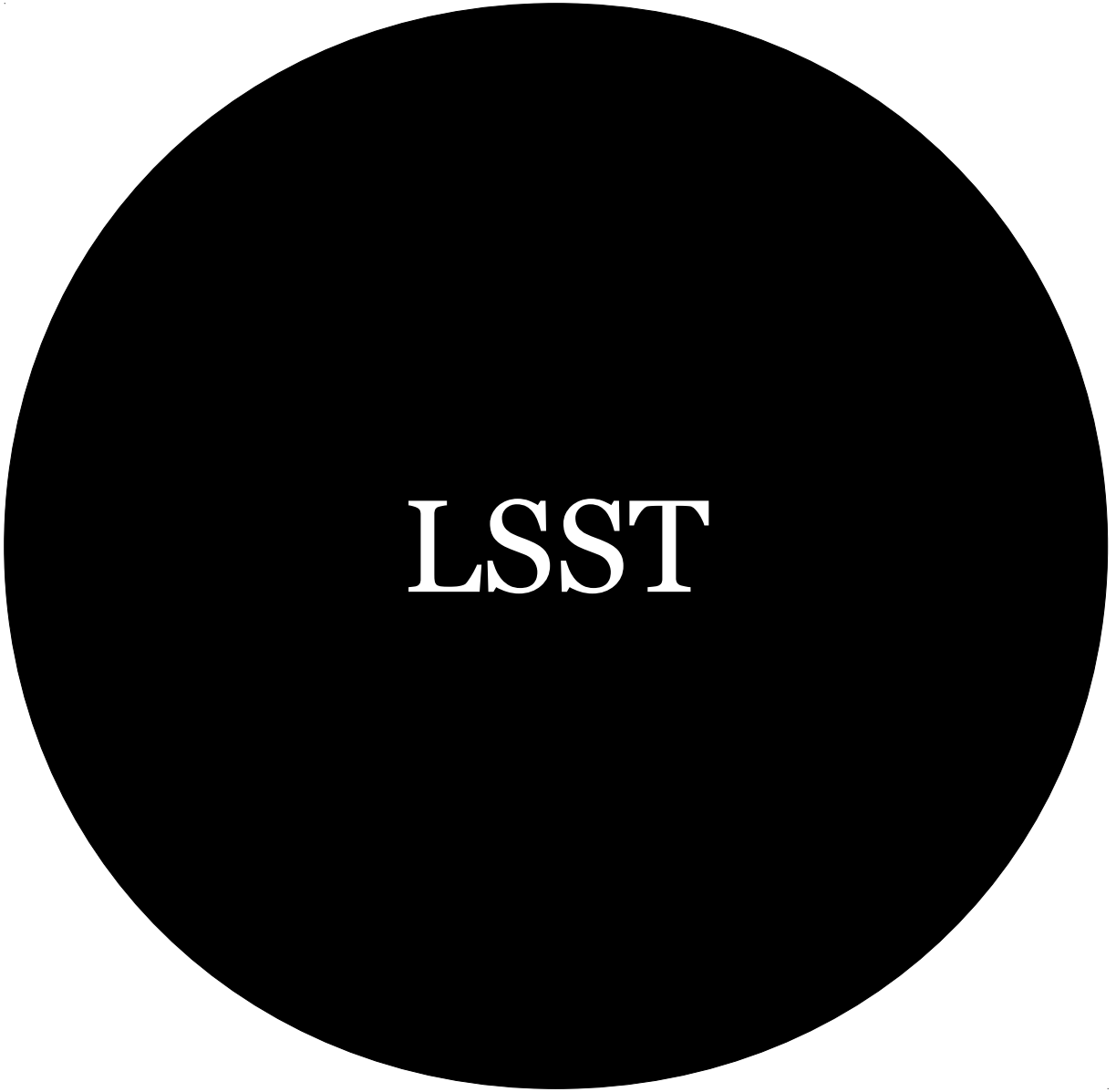
4. Strategies: the non-existing example

Early epoch classification

Crucial for spectroscopic follow-up planning

0 entries in the SNPCC




LSST

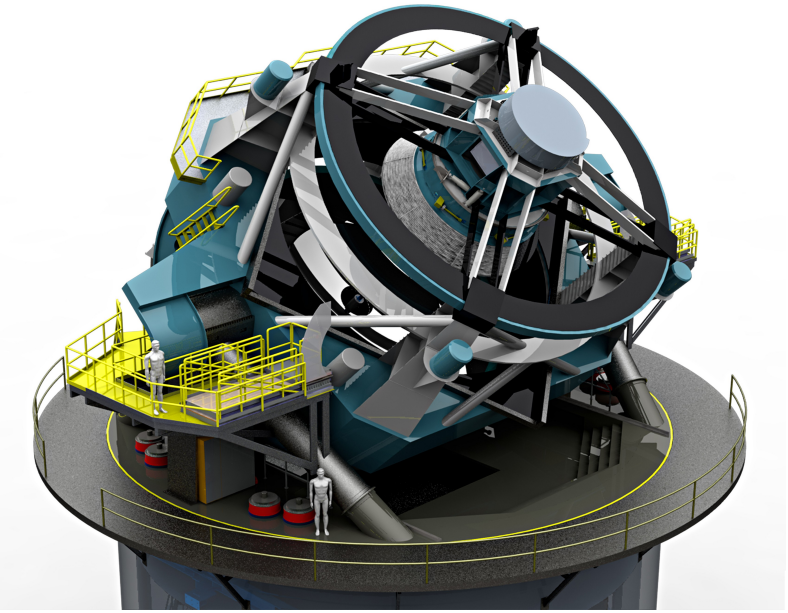
5. The Large Synoptic Survey Telescope (LSST)

year	Number of supernova
1998	42
2014	740
2025	> 10 000

2 million alerts/day
15 TB/day

40 nights of LSST

entire Google database



https://www.lsst.org/


PLAsTiCC
Photometric LSST
Astronomical Time-series
Classification Challenge

For LSST...

A data challenge aimed to prepare a larger community for the LSST data paradigm

2017ClassificationChallenge: an LSST Photometric Classification Challenge

Renée Hložek, Tina Peters, Rick Kessler, Dan Scolnic, Saurabh Jha, Ashish Mahabal, Federica Bianco, Hiranya Peiris, Michelle Lochner, Robert Schumann, Rob Firth, Mark Sullivan, Alex Malz, Lluís Galbany, Emille Ishida, Rahul Biswas, Bob Nichol

Enabling Science Grant

Goals:

1. Up to date view on ML performances on LSST data
2. Engage non-astronomers in the classification task
3. Boost the development of new algorithms
4. **Provide a data set for future studies**

PLAsTiCC

Photometric LSST Astronomical Time-series Classification Challenge

A data challenge aimed to prepare a larger community for the LSST data paradigm

Ensuring realistic simulations - not a SN challenge!!

Call for notice of intent for transient model inputs for PLAsTiCC: Photometric LSST Astronomical Time-series Classification Challenge

May 1, 2017

Renée Hložek, Rick Kessler, Anita Bahmanyar, Federica Bianco, Rahul Biswas, Mi Dai, Seth Digel, Jason McEwen, Rob Firth, Dominique Fouchez, Lluís Galbany, Philippe Gris, Emille Ishida, Saurabh Jha, Michelle Lochner, Ashish Mahabal, Alex Malz, Bob Nichol, Johanna Pasquet, Tina Peters, Hiranya Peiris, Chad Schafer, Robert Schuhmann, Dan Scolnic, Mark Sullivan, Elizabeth Swann.

PLAsTiCC

Photometric LSST Astronomical Time-series Classification Challenge

A data challenge aimed to prepare a larger community for the LSST data paradigm

Participation is **open to everyone**:
no need to be LSST member!

https://plasticcblog.wordpress.com/

PLAsTiCC

Photometric LSST Astronomical Time-series Classification Challenge

A data challenge aimed to prepare a larger community for the LSST data paradigm

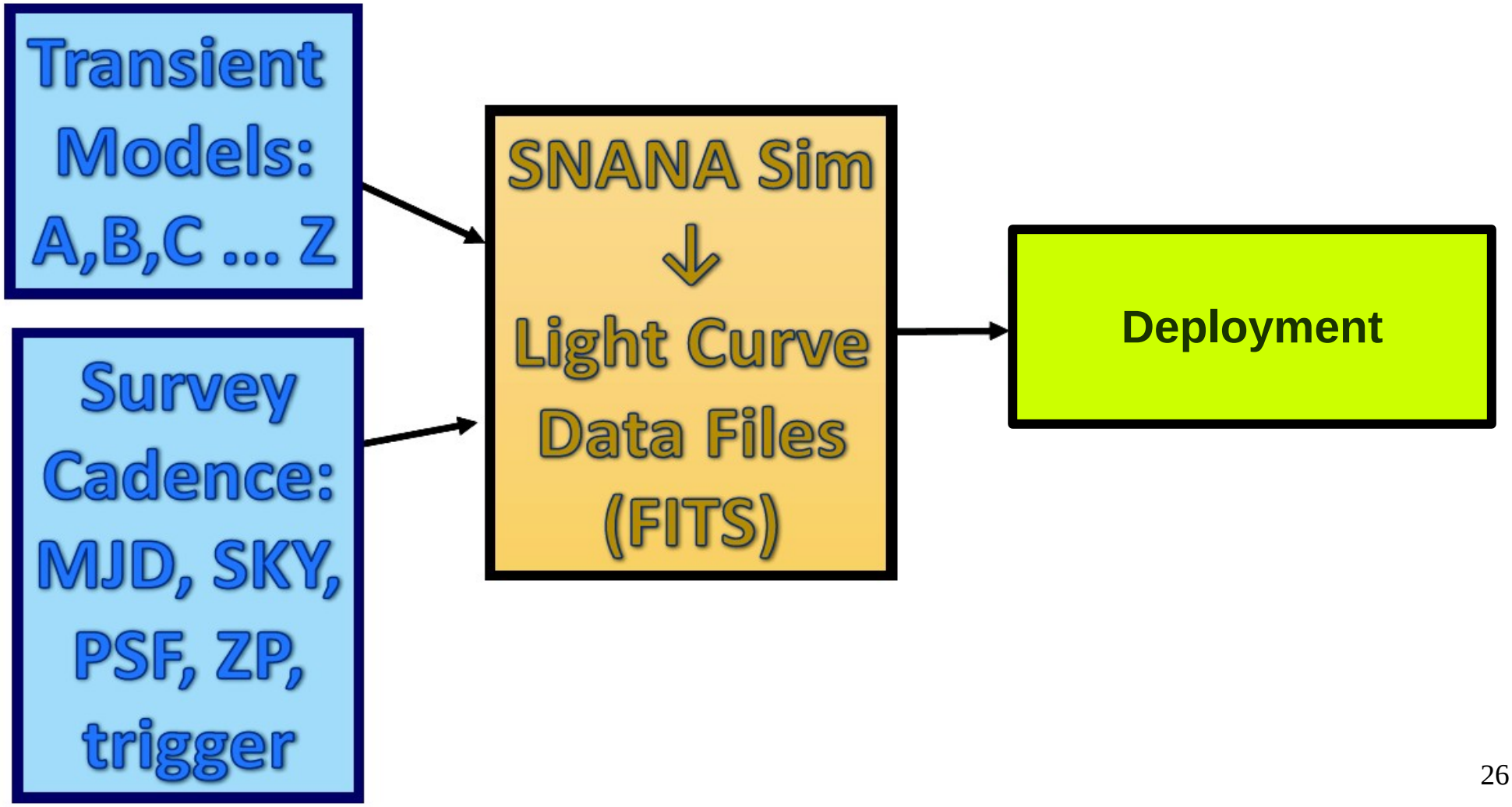
- SNANA simulations → Light curves in observer-frame (no images!)
- 3 years worth of LSST data, ~ 100 MB
- ~ 10^7 objects
- Around 20 transient models
  (galactic and extra-galactic, periodic and non-periodic)

- Please respect model-information policy:
  ``don't ask, don't tell"

- Training sample will be small and biased
- Not all models will be present in the training sample

PLAsTiCC - *Current status*

Simulation Architecture



Slide by Rick Kessler, I PLAsTiCC workshop, July 2017

PLAsTiCC

How it will be deployed?

kaggle™ ← Answer-based submissions

Code-based submissions + Collaborative phase →

RAPID ANALYTICS AND MODEL PROTOTYPING (RAMP)

COLLABORATIVE CHALLENGE WITH CODE SUBMISSION

Paris-Saclay Center for Data Science

See A. Boucaud talk

PLAsTiCC

Tasks

Supervised and/or Semi-supervised learning

Novelty and/or Anomaly detection

PLAsTiCC:

Get involved!

Play with the first SNPCC data!

Here is a peak at the data.

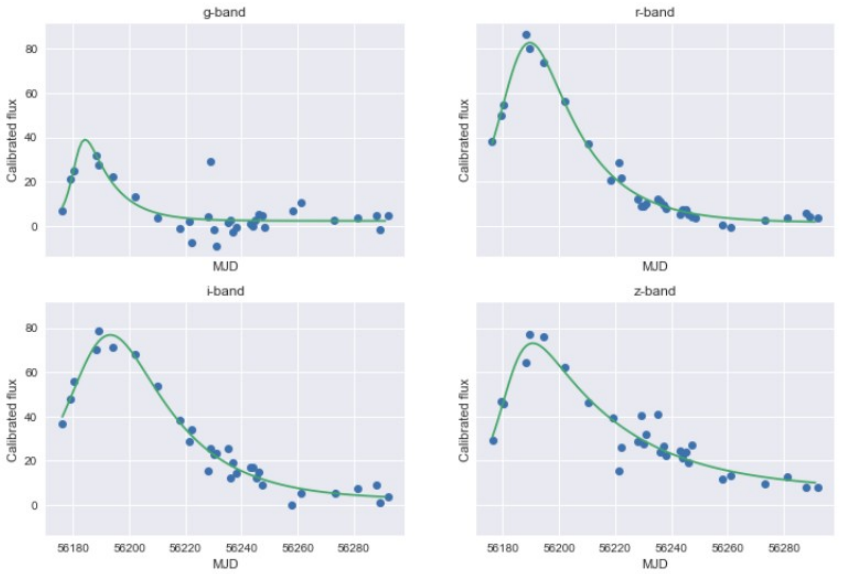
```
In [2]: X, y = read_data('data/des_train.pkl')

In [3]: # 5 first rows of the dataframe
        X.head()
```

Out[3]:

	mjd_g	fluxcal_g	fluxcalerr_g	mjd_r	fluxcal_r	fluxcalerr_r	mjd_i	fluxcal_i	f
1186	[56283.203, 56288.199, 56292.098, 56304.051, 5...	[3.182, -13.07, -7.177, -0.4645, 0.2906, 16.63...	[6.285, 11.79, 11.65, 2.479, 7.569, 6.195, 3.6...	[56283.211, 56288.215, 56292.102, 56304.055, 5...	[-4.973, -6.427, 9.861, -1.663, 6.724, 23.82, ...	[3.326, 6.181, 4.83, 1.657, 2.966, 2.553, 3.07...	[56283.215, 56289.109, 56292.191, 56304.07, 56...	[-0.615, -9.294, 0.8371, 0.6895, -1.593, 24.51...	
	[56177.172, 56179.172,	[27.18, 34.83, 24.15,	[4.837, 2.372, 1.805,	[56177.188, 56179.312,	[33.02, 34.65, 26.92,	[5.375, 3.109, 1.365,	[56177.203, 56179.328,	[49.55, 42.79, 30.72,	

```
In [14]: plot_lightcurves_with_fit(2)
```



Constructing the data matrix for classification

Now that you can fit each individual light curve, you are ready to build your low dimension representation of the data.

universite PARIS-SACLAY

Paris-Saclay Center for Data Science

PLAsTiCC - RAMP: classification of astronomical transients

Alexandre Boucaud (Paris-Saclay Center for Data Science)

Emille E. O. Ishida (Universite Clermont-Auvergne)

https://github.com/ramp-kits/supernovae

TransiXplore :

Exploration des données temporelles transitoires issues des relevés de prochaine génération

Talk to Emmanuel Gangler for further informations

Target data release:
Summer 2018!

PLAsTiCC

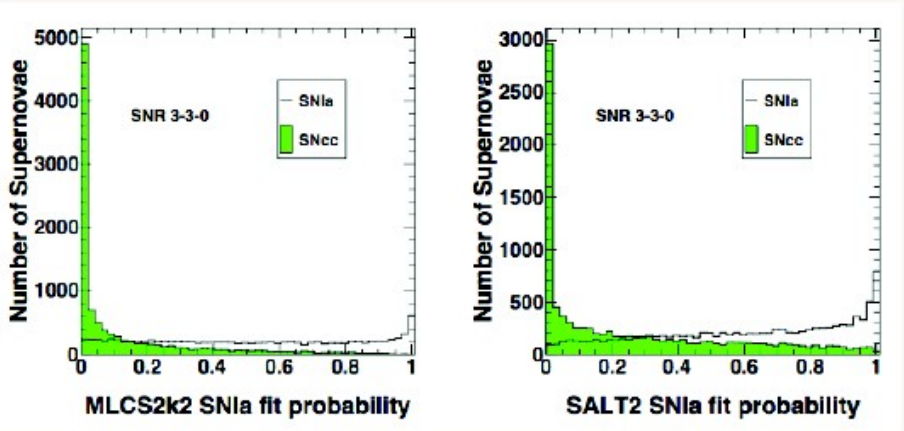
Photometric LSST Astronomical Time-series Classification Challenge

Thank you!

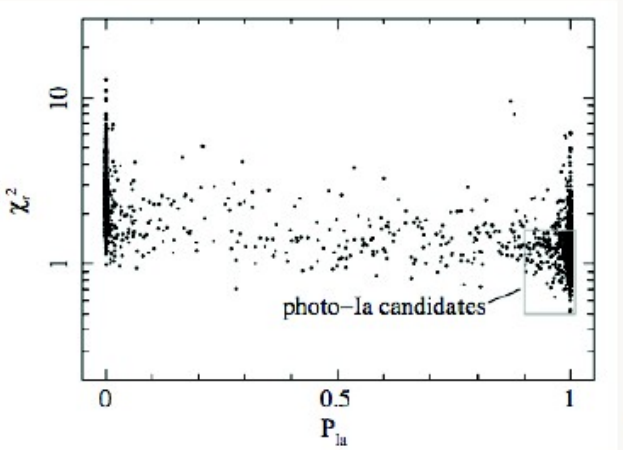
Extra Slides

After the first SN photometric classification challenge...

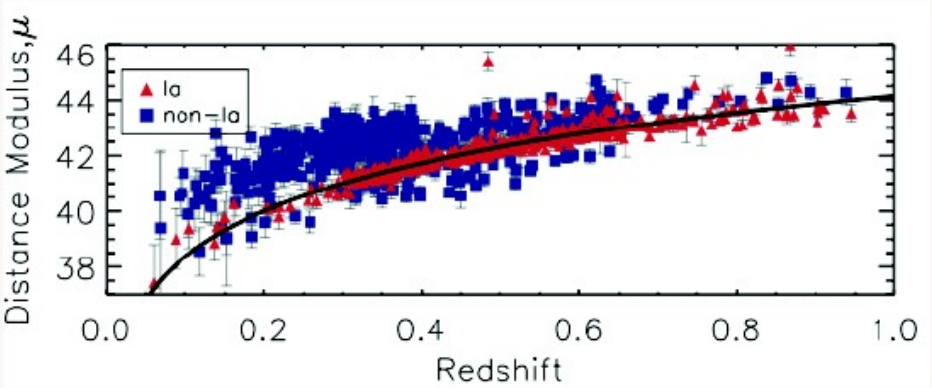
Fit Ia template to determine type
Bayesian inference using Ia and non-Ia templates
Define Hubble diagram >> Ia are those that are close
Perform parametric light curve fits, then ML/statistics



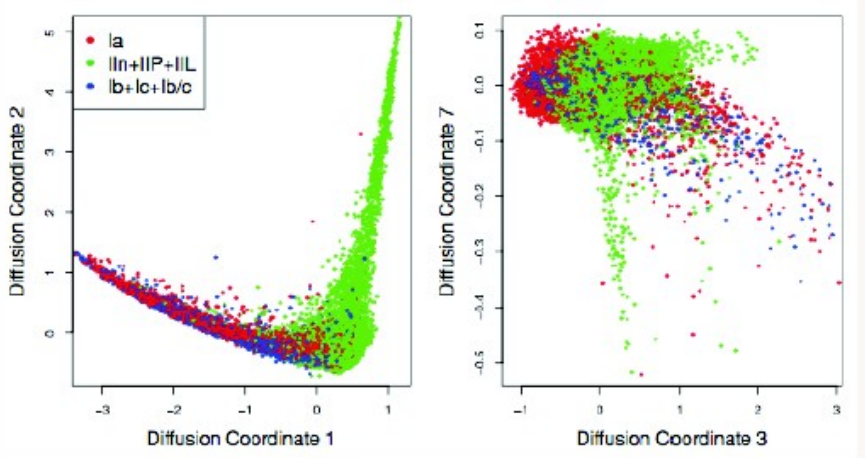
Gjergo+13



Sako+11



Newling+11



Richards+12

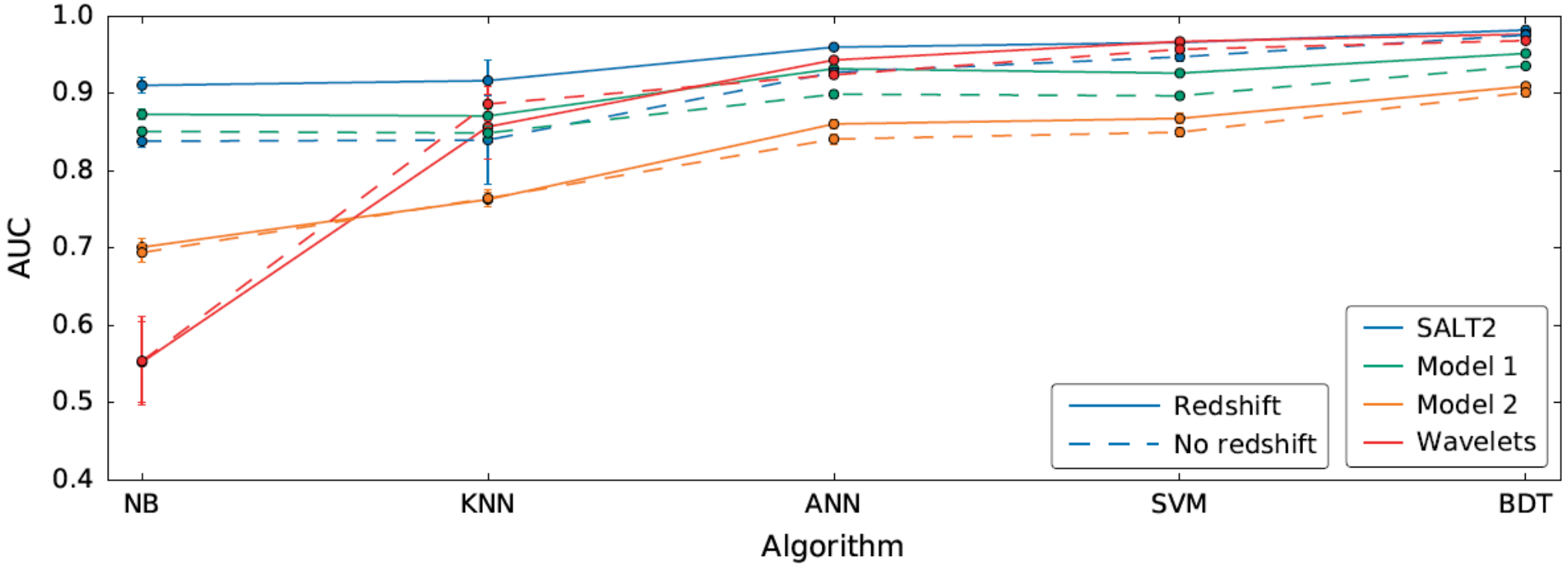
Slide from Adam A. Miller, CIERA-LSST SN Workshop, 2 June 2017

More recently …

Lochner et al., 2016, APJ 225, 2

Comparison of different feature extraction + Machine Learning techniques

Post-SNPCC data



SNANA simulation:

Header

```
SURVEY: DES
SNID:   2553
IAUC:    UNKNOWN
PHOTOMETRY_VERSION: DES
SNTYPE:  -9
FILTERS: griz
RA:      36.750000  deg
DECL:    -4.500000  deg
MAGTYPE: LOG10
MAGREF:  AB
FAKE:    2   (=> simulated LC with snlc_sim.exe)
MWEBV:   0.0283    MW E(B-V)
REDSHIFT_HELIO:   0.34470 +- 0.03150  (Helio, z_best)
REDSHIFT_FINAL:   0.34470 +- 0.03150  (CMB)
REDSHIFT_SPEC:    -9.00000 +- 9.00000
REDSHIFT_STATUS: OK

HOST_GALAXY_GALID:   12339
HOST_GALAXY_PHOTO-Z:   0.3447  +- 0.0315



SIM_MODEL:  NONIA  10  (name index)
SIM_NON1a:      43   (non1a index)
SIM_COMMENT:  SN Type = II , MODEL = SDSS-018834
SIM_LIBID:  2
SIM_REDSHIFT:  0.3283
SIM_HOSTLIB_TRUEZ:  0.3300  (actual Z of hostlib)
SIM_HOSTLIB_GALID:  12339
SIM_DLMU:      41.182091  mag   [ -5*log10(10pc/dL) ]
SIM_RA:        36.750000 deg
SIM_DECL:      -4.500000 deg
SIM_MWEBV:   0.0270   (MilkyWay E(B-V))
SIM_PEAKMAG:   22.98  23.13  23.60  23.05  (griz obs)
SIM_EXPOSURE:     1.0    1.0    1.0    1.0  (griz obs)
SIM_PEAKMJD:   56208.320312  days
SIM_SALT2x0:   3.366e-17
SIM_MAGDIM:    0.000
SIM_SEARCHEFF_MASK:  3  (bits 1,2=> found by software,humans)
SIM_SEARCHEFF:  1.0000  (spectro-search efficiency (ignores pipelines))
SIM_TRESTMIN:   -22.69   days
SIM_TRESTMAX:    67.56   days
SIM_RISETIME_SHIFT:   0.0 days
SIM_FALLTIME_SHIFT:   0.0 days

SEARCH_PEAKMJD:   56207.977
```

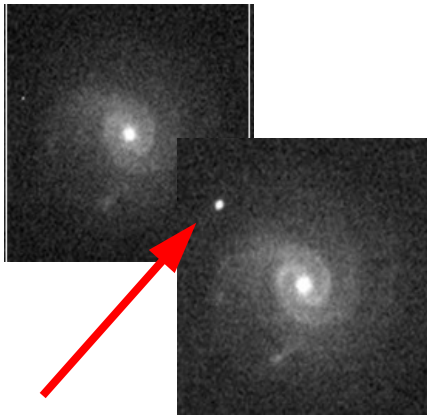
SNANA simulation:

Light curve

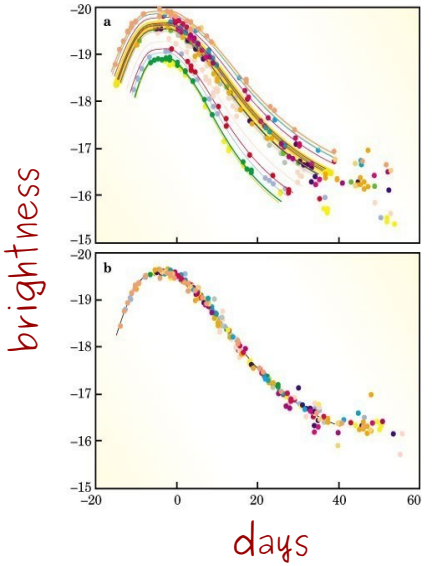
```
# TERSE LIGHT CURVE OUTPUT:
#
NOBS: 99
NVAR: 9
VARLIST:  MJD  FLT FIELD   FLUXCAL   FLUXCALERR   SNR    MAG     MAGERR  SIM_MAG
OBS:   56178.188  g NULL  -8.405e+00   9.406e+00  -0.89   99.000    5.000   98.968
OBS:   56178.195  r NULL   3.721e+00   7.172e+00   0.52   99.000    5.000   99.005
OBS:   56178.203  i NULL  -8.865e-01   6.833e+00  -0.13   99.000    5.000   99.065
OBS:   56178.211  z NULL  -4.081e+00   7.919e+00  -0.52   99.000    5.000   98.907
OBS:   56180.172  g NULL  -7.214e-01   4.012e+00  -0.18   99.000    5.000   98.968
OBS:   56180.188  r NULL   3.867e+00   3.175e+00   1.22   99.000    5.000   99.005
OBS:   56180.195  i NULL  -3.010e+00   4.395e+00  -0.68   99.000    5.000   99.065
OBS:   56180.203  z NULL   7.996e+00   4.965e+00   1.61   99.000    5.000   98.907
OBS:   56188.148  g NULL  -2.440e+01   1.796e+01  -1.36  128.000    0.000   45.106
OBS:   56188.289  r NULL   1.617e+00   3.995e+00   0.40   26.978  101.022   46.679
OBS:   56188.297  i NULL  -2.510e+00   8.132e+00  -0.31  128.000    0.000   47.554
OBS:   56188.312  z NULL   7.324e+00   5.394e+00   1.36   25.338    1.447   47.995
OBS:   56194.145  g NULL   8.978e-01   4.678e+00   0.19   27.617  100.383   40.917
OBS:   56194.156  r NULL   2.316e+00   2.751e+00   0.84   26.588  101.412   42.664
OBS:   56194.172  i NULL   3.518e+00   4.628e+00   0.76   26.134  101.866   43.390
OBS:   56194.188  z NULL  -7.580e+00   4.465e+00  -1.70  128.000    0.000   43.464
OBS:   56207.188  g NULL   5.977e+01   4.637e+00  12.89   23.059    0.088   22.989
OBS:   56207.195  r NULL   6.437e+01   3.691e+00  17.44   22.978    0.063   23.135
OBS:   56207.203  i NULL   3.419e+01   6.290e+00   5.44   23.665    0.221   23.672
OBS:   56207.219  z NULL   6.978e+01   6.101e+00  11.44   22.891    0.099   23.080
OBS:   56215.195  g NULL   5.340e+01   3.035e+00  17.59   23.181    0.064   23.136
OBS:   56215.203  r NULL   4.733e+01   2.390e+00  19.80   23.312    0.056   23.301
OBS:   56215.211  i NULL   4.553e+01   2.798e+00  16.27   23.354    0.069   23.411
OBS:   56215.227  z NULL   7.188e+01   3.696e+00  19.45   22.859    0.058   22.828
```

Supernova Cosmology

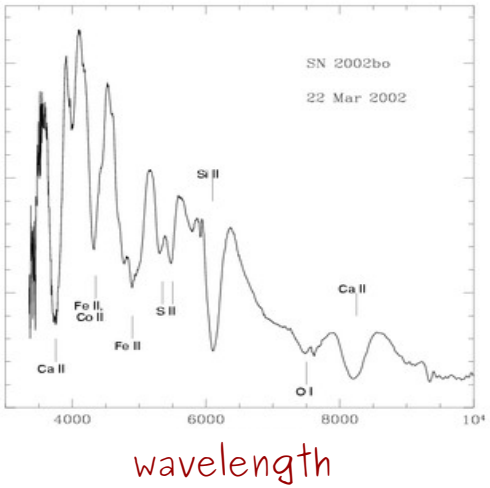
1. detection



2. photometry

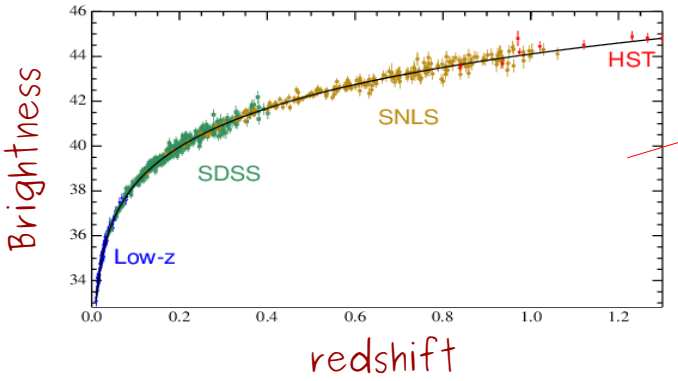


3. spectroscopy



Distance (redshift) + classification

4. standardization + cosmological fit

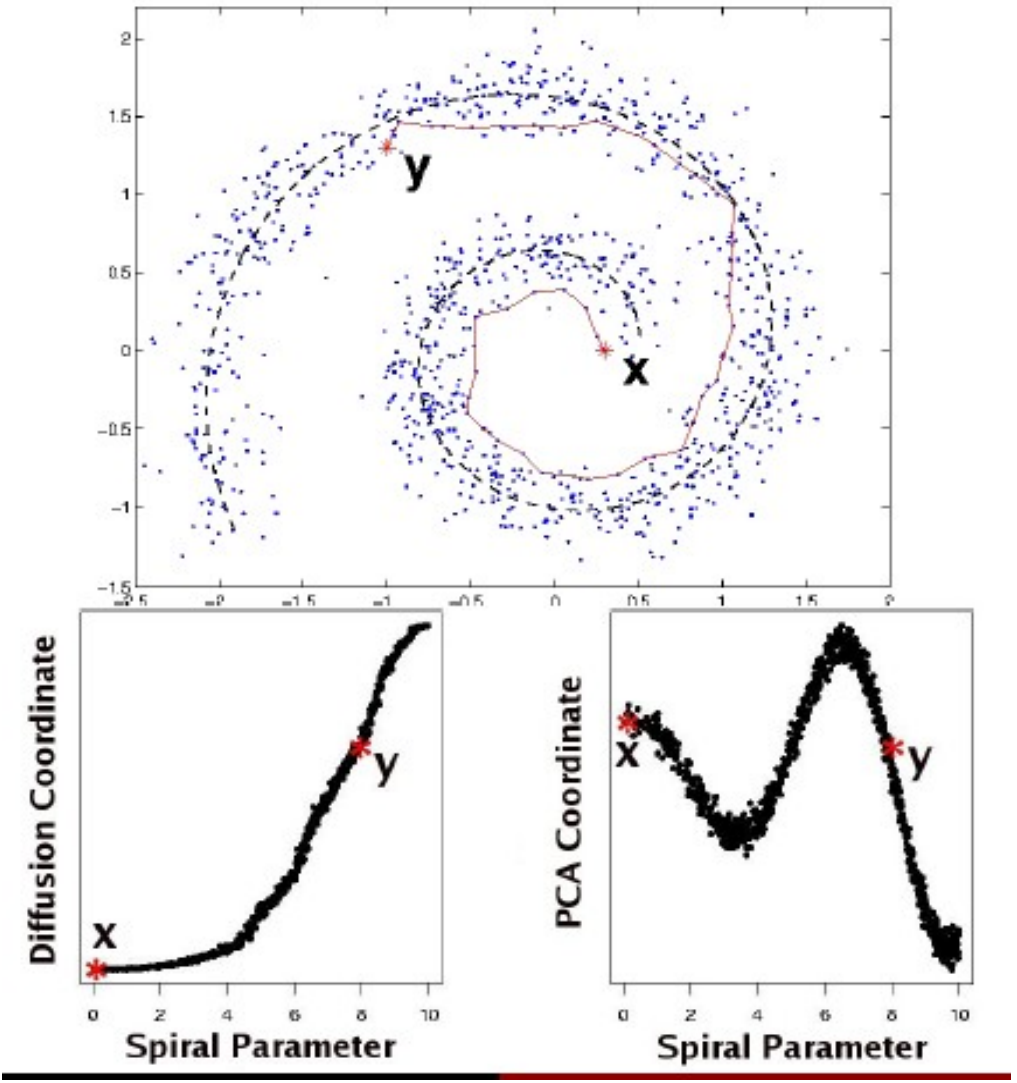


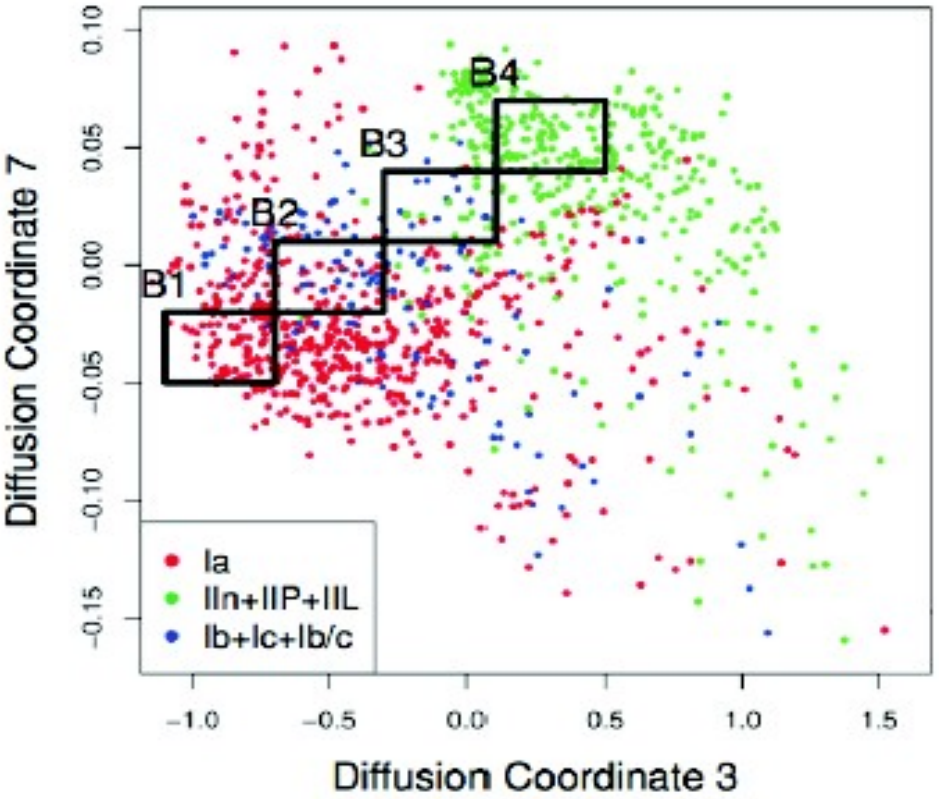
year	Number of supernova
1998	42
2014	740
2025	> 10 000

4. Strategies: example

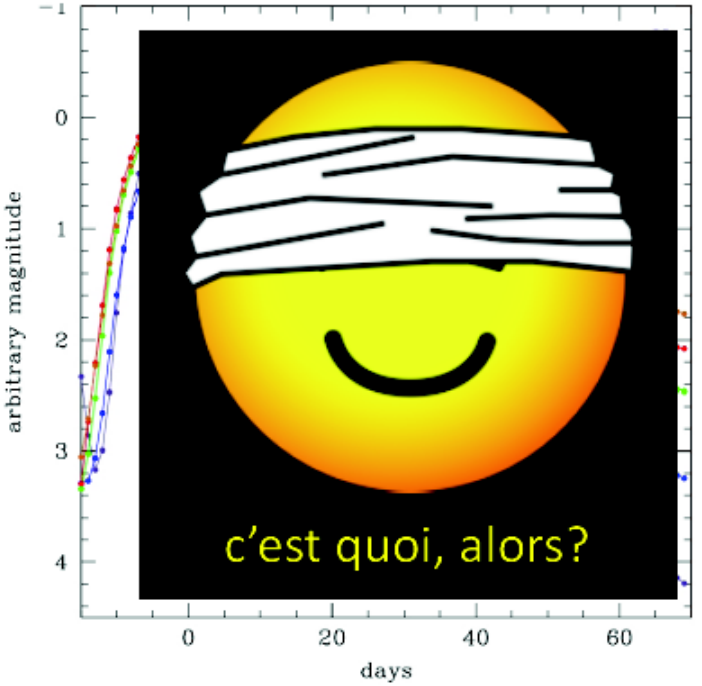
Diffusion maps + random forests

Richards et al, 2011





Early-Epoch Challenge



- Important for LSST startup: classify transients with few epochs → enables spectroscopic follow-up.
- Full light-curve analysis is fun for science, but less critical for LSST operations.
- In 2010 SN(Ia+CC) challenge, **NOBODY** tried early-epoch classification !
- Should we ***require*** early-epoch classification to accept a submission ?
- How is early-epoch challenge defined ? Nobs past trigger ? Ndays past trigger ?
- How many early-epoch ranges to include ?

Slide by Rick Kessler, I PLAsTiCC workshop, July 2017

Kessler et al., 2010, PASP, Volume 122, Issue 898, pp. 1415

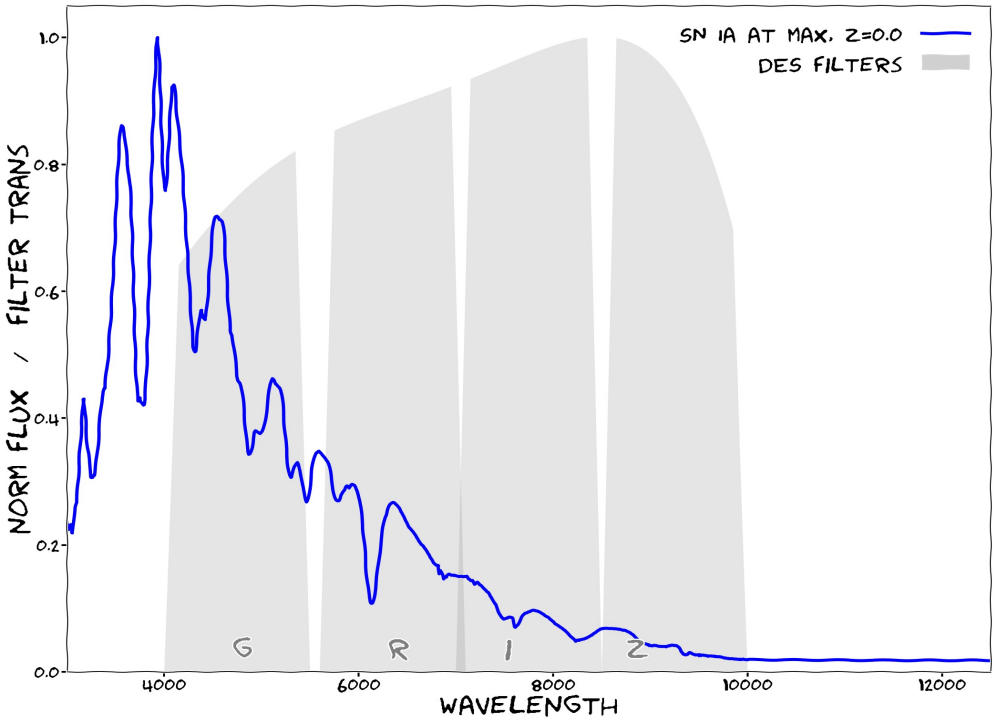
LIST OF PARTICIPANTS IN THE SNPhotCC.

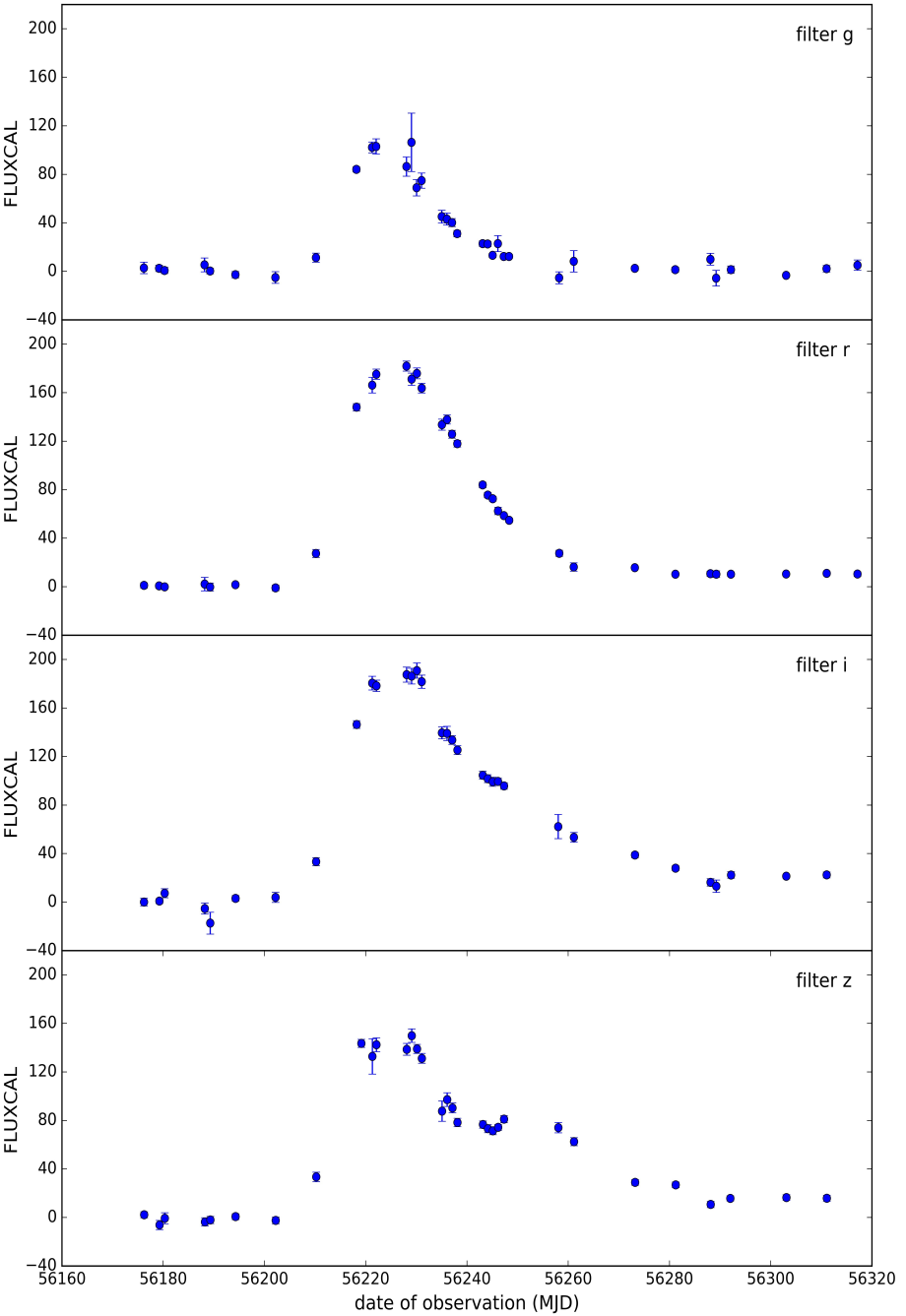
Participants	Abbreviation[a]	Classified +Z[b]/noZ[c]	SN z_{ph}[d]	CPU[e]	Description (strategy class[f])
P. Belov and S. Glazov	Belov & Glazov	yes/no	no	90	light curve χ^2 test against Nugent templates (2)
S. Gonzalez	Gonzalez	yes/yes	no	120	cuts on SiFTO fit χ^2 and fit parameters (1)
J. Richards, Homrighausen, C. Schafer, P. Freeman	InCA[g]	no/yes	no	1	Spline fit & nonlinear dimensionality reduction (4)
J. Newling, M. Varuguese,	JEDI-KDE	yes/yes	no	10	Kernel Density Evaluation with 21 params (4)
B. Bassett, R. Hlozek,	JEDI Boost	yes/yes	no	10	Boosted decision trees (4)
D. Parkinson, M. Smith,	JEDI-Hubble	yes/no	no	10	Hubble diagram KDE (3)
H. Campbell, M. Hilton, H. Lampeitl, M. Kunz, P. Patel (JEDI group[h])	JEDI Combo	yes/no	no	10	Boosted decision trees + Hubble KDE (3+4)
S. Philip, V. Bhatnagar,	MGU+DU-1[i]	no/yes	no	< 1	light curve slopes & Neural Network (2)
A. Singhal, A. Rai, A. Mahabal, K. Indulekha	MGU+DU-2	no/yes	no	< 1	light curve slopes & Random Forests (2)
H. Campbell, B. Nichol,	Portsmouth χ^2	yes/no	no	1	SALT2–χ^2_r & False Discovery Rate Statistic (1)
H. Lampietl, M .Smith	Portsmouth-Hubble	yes/no	no	1	Deviation from parametrized Hubble diagram (3)
D. Poznanski	Poz2007 RAW	yes/no	yes	2	SN Automated Bayesian Classifier (SN–ABC) (2)
	Poz2007 OPT	yes/no	yes	2	SN–ABC with cuts to optimize $\mathcal{C}_{\mathrm{FoM-Ia}}$ (2).
S. Rodney	Rodney	yes/yes	yes	230	SN Ontology with Fuzzy Templates (2)
M. Sako	Sako	yes/yes	yes	120	χ^2 test against grid of Ia/II/Ibc templates (2)
S. Kuhlmann, R. Kessler	SNANA cuts	yes/yes	yes	2	Cut on MLCS fit probability, S/N & sampling (1)

[a]Groups are listed alphabetically by abbreviation.
[b]Classifications included for SNPhotCC/HOSTZ.
[c]Classifications included for SNPhotCC/noHOSTZ.
[d]photo-z estimates included.
[e]Average processing time per SN (seconds) using similar 2-3 GHz cores.
[f]From §3, strategy classes are 1) selection cuts, 2) Bayesian probabilities, 3) Hubble-diagram parametrization and 4) statistical inference.
[g]International Computational Astrophysics Group: http://www.incagroup.org
[h]Joint Exchange and Development Initiative: http://jedi.saao.ac.za
[i]MGU=Mahatma Gandhi University, DU=Delhi University.

3. The data: light curve

broad brand filters





4. Strategies: previous attempts

The first SN photometric classification challenge...

Kessler et al., 2010, PASP, Volume 122, Issue 898, pp. 1415

$$\mathrm{eff} = \frac{N_{\mathrm{Ia}}^{\mathrm{SC}}}{N_{Ia}^{\mathrm{tot}}}$$

$$\mathrm{pur} = \frac{N_{\mathrm{Ia}}^{\mathrm{SC}}}{N_{\mathrm{nonIa}}^{\mathrm{WC}} + N_{\mathrm{Ia}}^{\mathrm{SC}}}$$

$$\mathrm{SC} = \frac{N_{Ia}^{\mathrm{SC}} + N_{\mathrm{nonIa}}^{\mathrm{SC}}}{N^{\mathrm{TOT}}}$$

$$\mathrm{ppur} = \frac{N_{\mathrm{Ia}}^{\mathrm{SC}}}{N_{\mathrm{Ia}}^{\mathrm{SC}} + W N_{\mathrm{nonIa}}^{\mathrm{WC}}},$$

$$\mathrm{FoM} = \mathrm{eff} \times \mathrm{ppur},$$