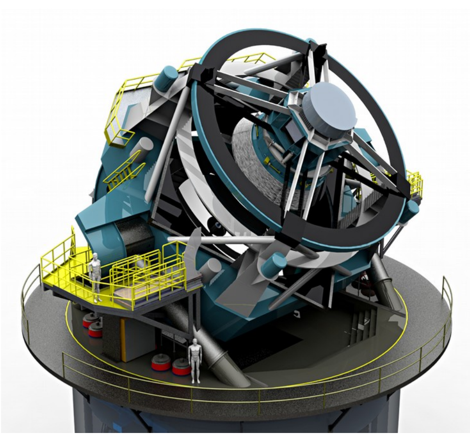
Exploration des données temporelles transitoires issues des relevés de prochaine génération

E. Gangler (LPC, Clermont-ferrand)



En collaboration avec
E. Mephu Nguifo (LIMOS, Clermont-Ferrand)
M. Moniez (LAL, Orsay)
J.-M. Petit, M. Scuturici (LIRIS, Lyon)
S. Lopes, L. Yeh, K. Zeitouni (DAVID, UVSQ)

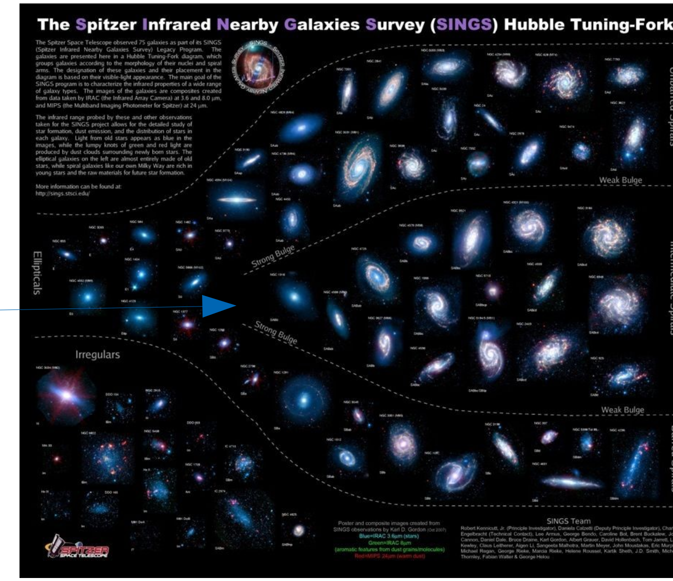
Classification des objets

- Données statiques

CFHT deep field

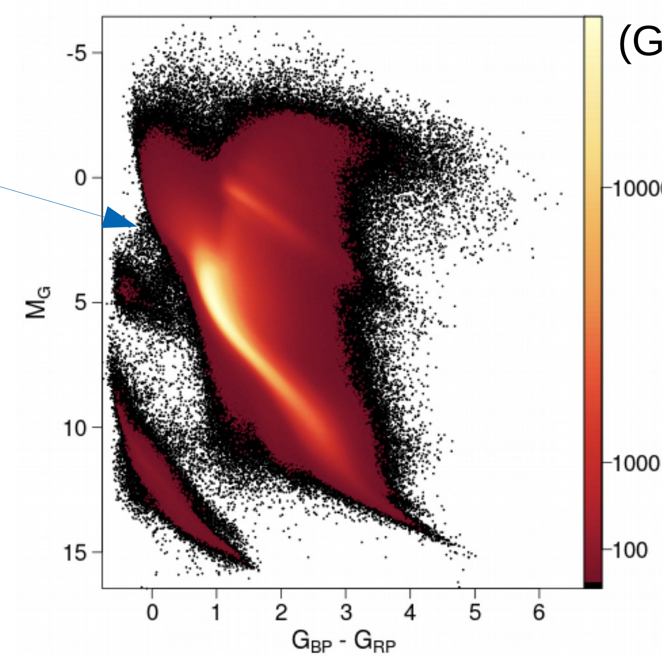
Morphologie (galaxies)

(Spitzer)

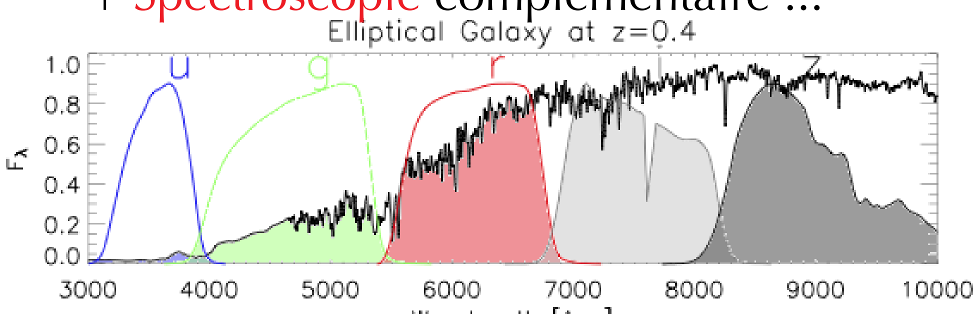


Magnitude, Couleur (étoiles, galaxies)

(Gaia)

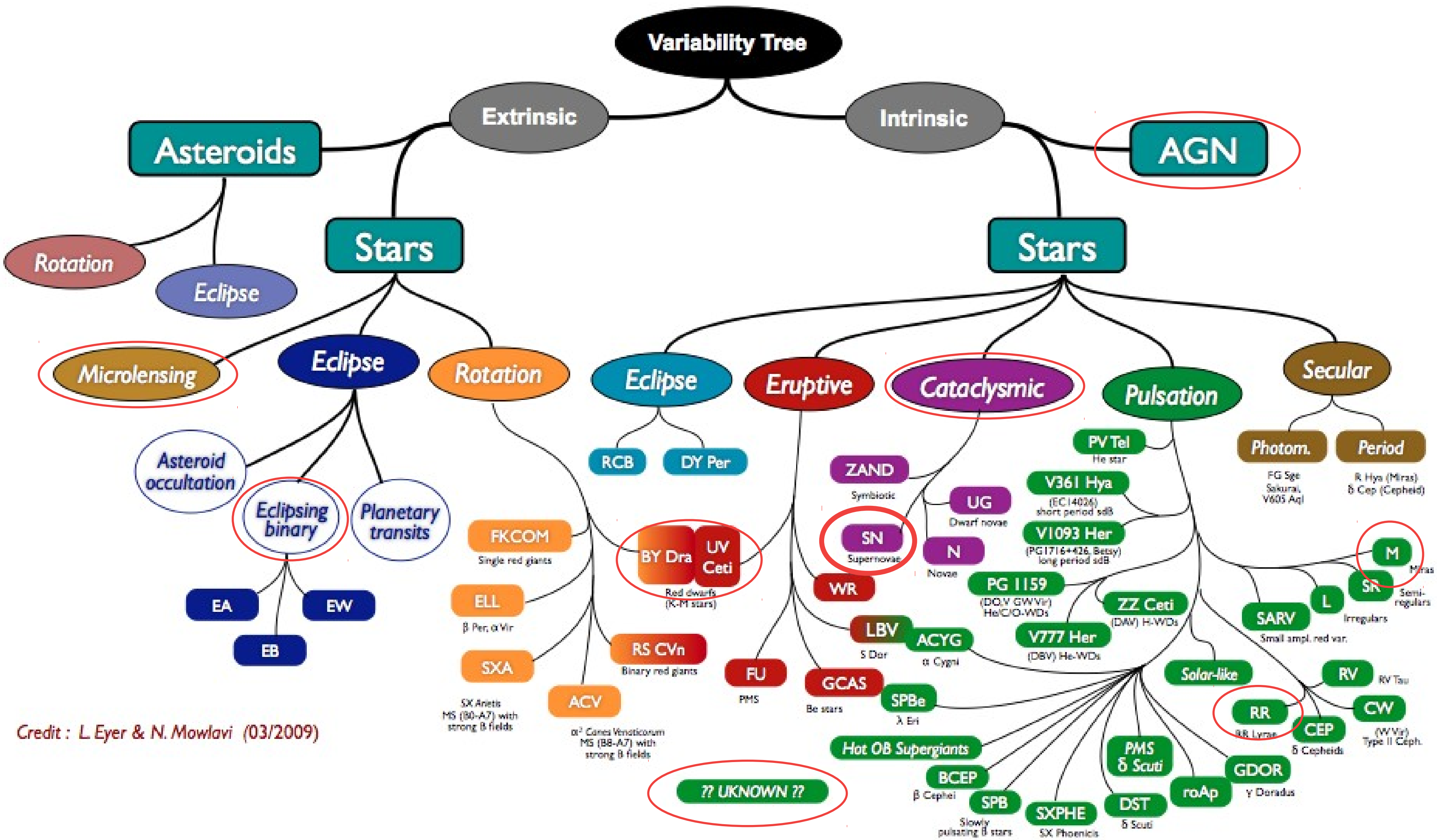


+ Spectroscopie complémentaire ...



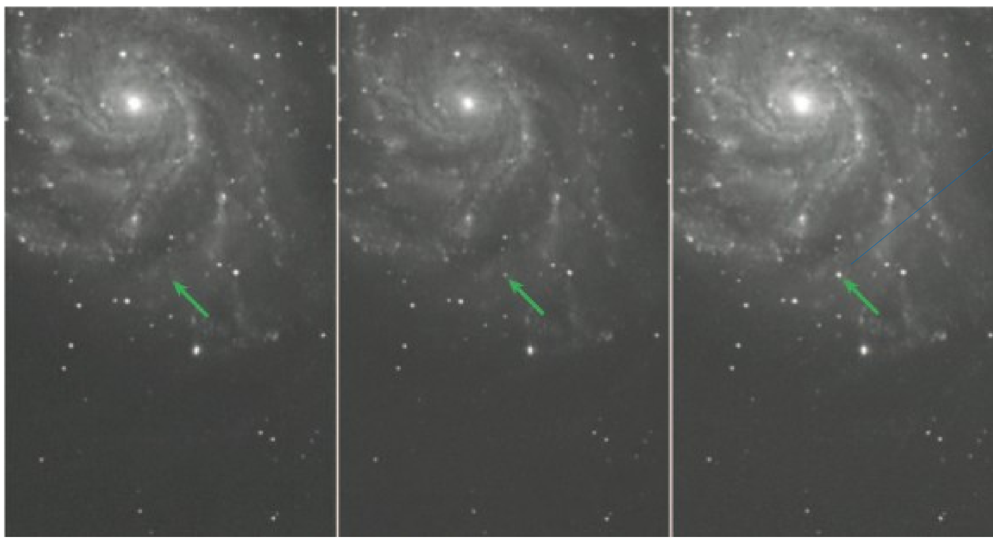
Classification des transitoires

- Une zoologie complexe :

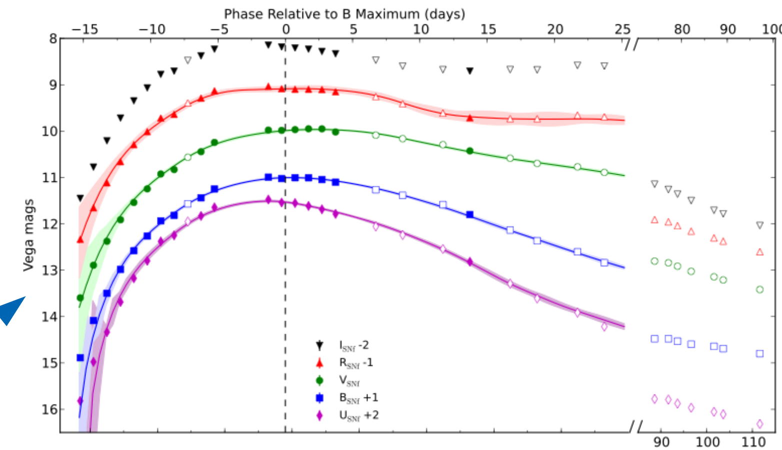


Classification des objets

- Données temporelles

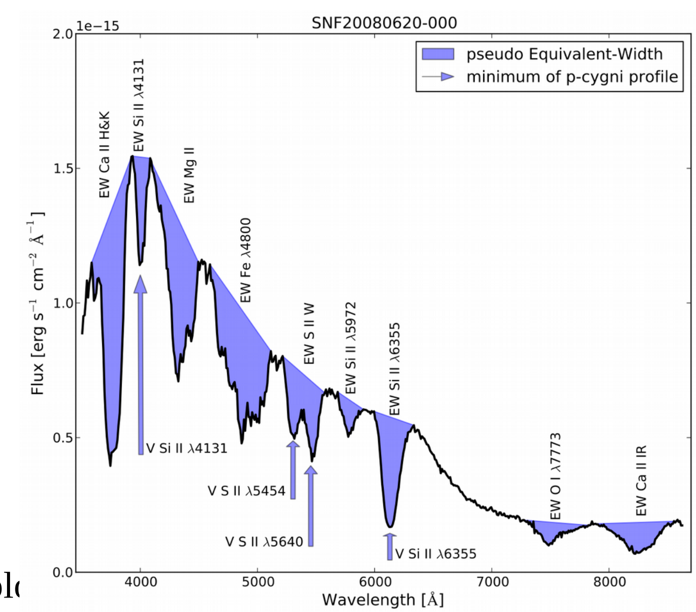


SN 2011 fe



Courbe de lumière
(évolution temporelle du flux en plusieurs bandes)

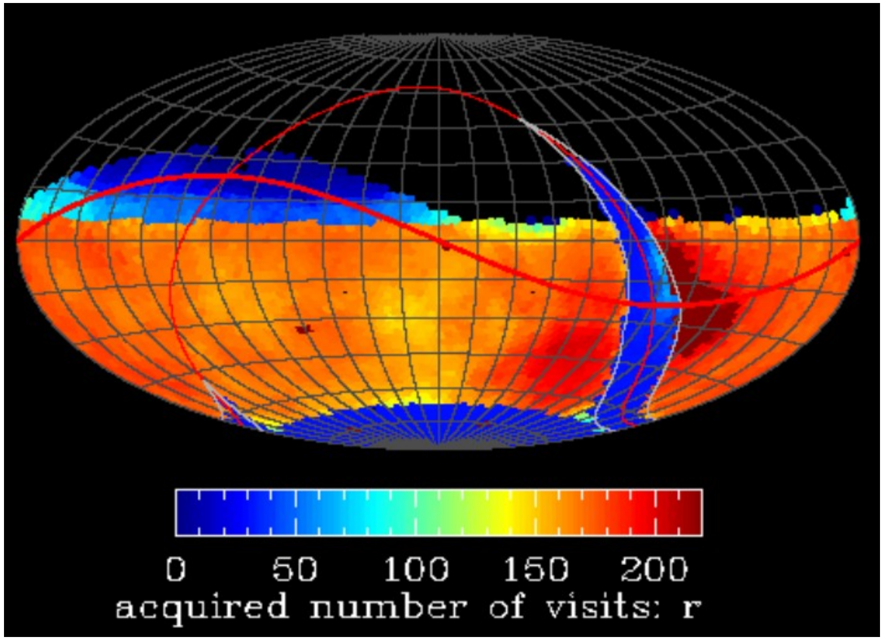
Spectroscopie d'identification:
trop coûteuse
seuls quelques objets
→ faible lot d'entraînement



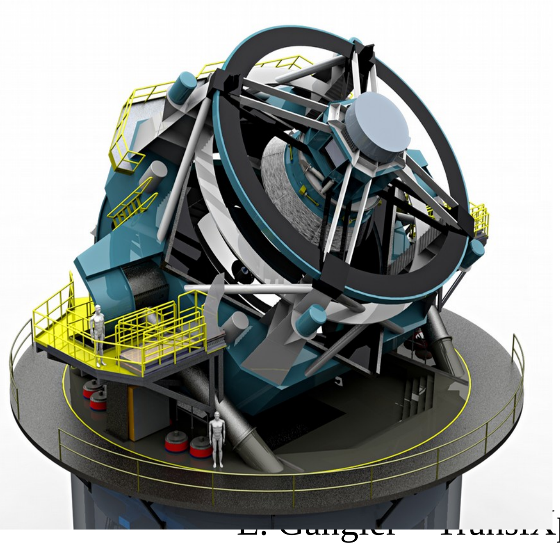
LSST
Large Synoptic Survey Telescope

capabilities :

- A stage-IV survey :
  - 8.4 (6.7) m telescope
  - Cerro Pachon (Chili)
  - 3.2 Gpix 9.6□ FoV camera
  - 0.2'' pixel / 0.7'' median FWHM
  - First light 2020, Survey 2022



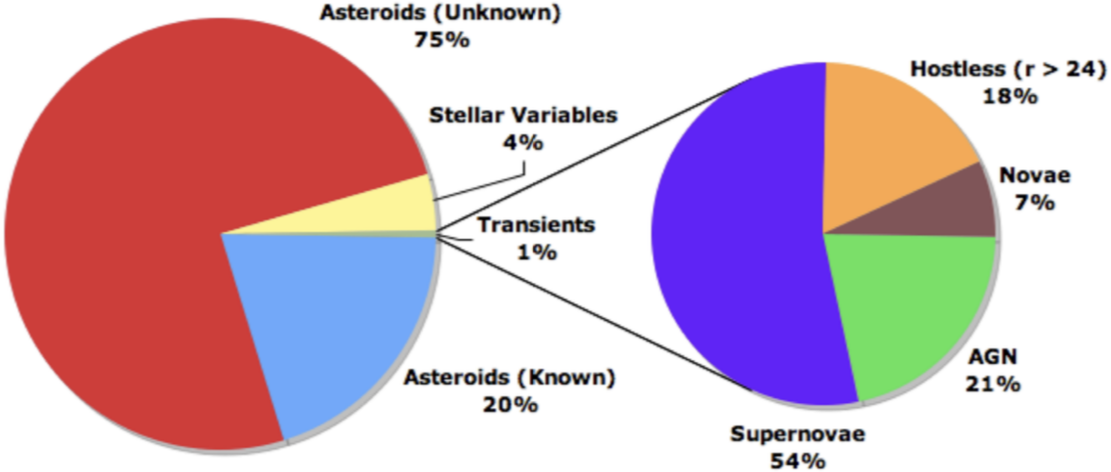
- All visible sky in 6 bands (ugrizy) (~18000□)
- 2x15 s exposure, 1 visit / 3 days
  r ~24.4 / visit
- During 10 years !
  → ~825 visits (all bands)
- 20TB/day 60 PB/10 years



LSST Data in short

Simulation 1 CCD
4k x 4k

- Huge data flow
  - Images : 2x6.4 Gpix/39 seconds
  - 20 TB/night, 60 PB image archive
  - 40. 10^9 objets (100-200 TB catalog)
  - 5 000. 10^9 detections ''sources'' (3-5 PB catalog)
  - 32 000. 10^9 measurements ''forced sources'' (1-2 PB catalog)
    - → **Post-observation transients classification**
  - Nightly alerts: **10 millions per night**
    - *Live transient classification*



https://www.kaggle.com/c/PLAsTiCC-2018

Featured Prediction Competition

PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

LSST Project · 1,094 teams · a month ago

1,094	1,384	22,895
Teams	Competitors	Entries

$25,000
Prize Money

Overview | Data | Kernels | Discussion | Leaderboard | Rules

Launched Sep 28th 2019
Due dec 14th 2019
147 public kernels
279 discussion topics

Overview

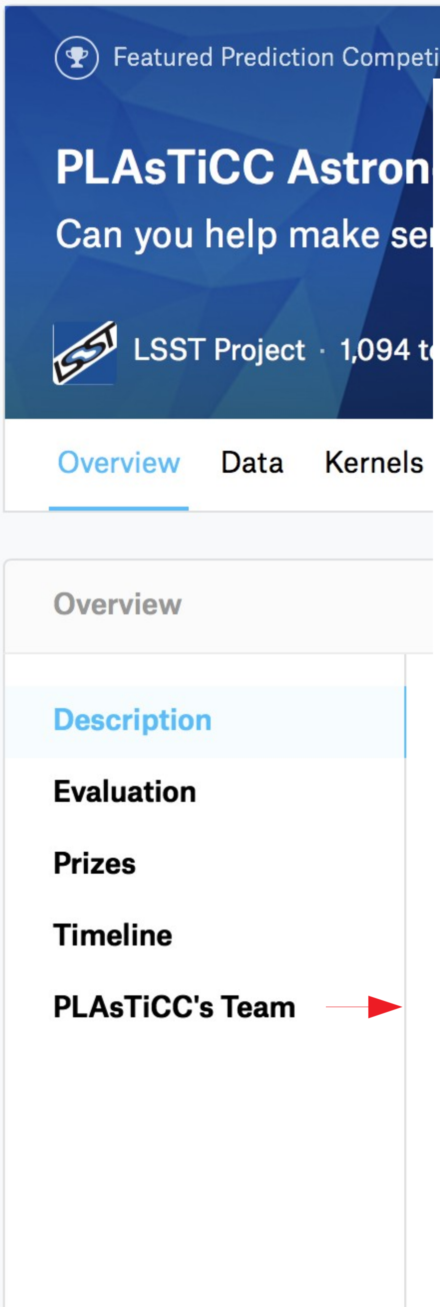
- **Description**
- **Evaluation**
- **Prizes**
- **Timeline**
- **PLAsTiCC's Team**

Help some of the world's leading astronomers grasp the deepest properties of the universe.

The human eye has been the arbiter for the classification of astronomical sources in the night sky for hundreds of years. But a new facility -- the Large Synoptic Survey Telescope (LSST) -- is about to revolutionize the field, discovering 10 to 100 times more astronomical sources that vary in the night sky than we've ever known. Some of these sources will be completely unprecedented!

The Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) asks Kagglers to help prepare to classify the data from this new survey. Competitors will classify astronomical sources that

https://www.kaggle.com/c/PLAsTiCC-2018

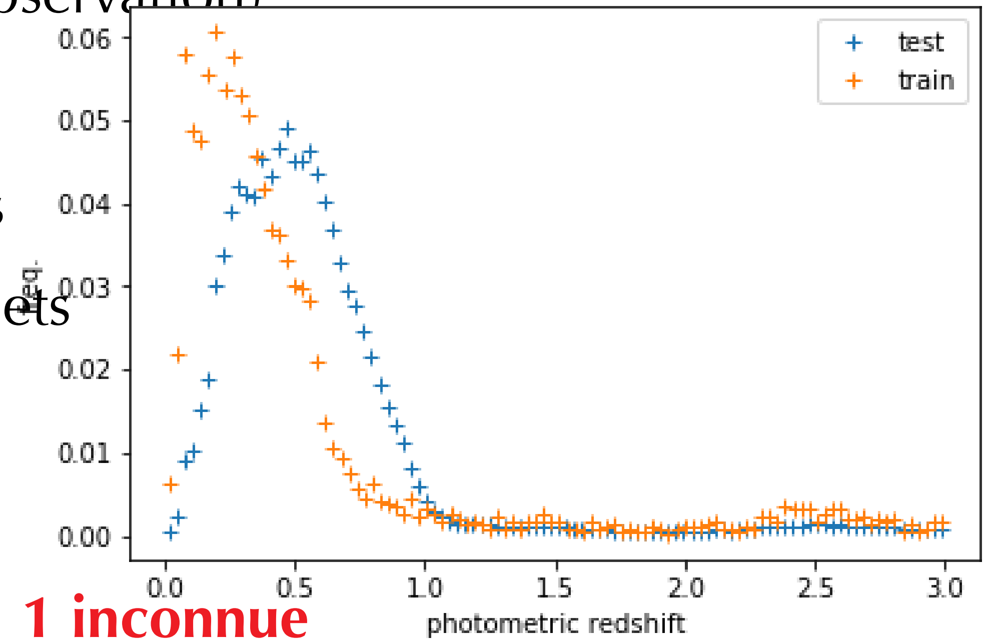


Tarek Allam Jr.* (University College of London, UK)
Anita Bahmanyar (University of Toronto, Canada)
Rahul Biswas (Stockolm University, Sweden)
Mi Dai (Rutgers, The State University of New Jersey, USA)
Lluís Galbany (University of Pittsburgh, USA)
Renée Hložek (Univeristy of Toronto, Canada)
Emille E. O. Ishida (CNRS/Universite Clermont Auvergne, France)
Saurabh W. Jha (Rutgers, The State University of New Jersey, USA)
David O. Jones (University of California Santa Cruz, USA)
Michelle Lochner* (African Institute for Mathematical Sciences, South Africa)
Ashish A. Mahabal (California Institute of Technology, USA)
Alex I. Malz* (New York University, USA)
Kaisey S. Mandel (Cambridge University, UK)
Juan Rafael Martínez-Galarza (Harvard-Smithsonian Center for Astrophysics, USA)
Jason D. McEwen (University College of London, UK)
Daniel Muthukrishna (Kavli Institute for Cosmology, UK)
Gautham Narayan (Space Telescope Science Institute, USA)
Hiranya Peiris* (Stockolm University, Sweden)
Christina M. Peters (University of Toronto, Canada)
Kara Ponder (University of California Berkeley, USA)
Christian N. Setzer* (Stockolm University, Sweden)

Données de PLAsTiCC

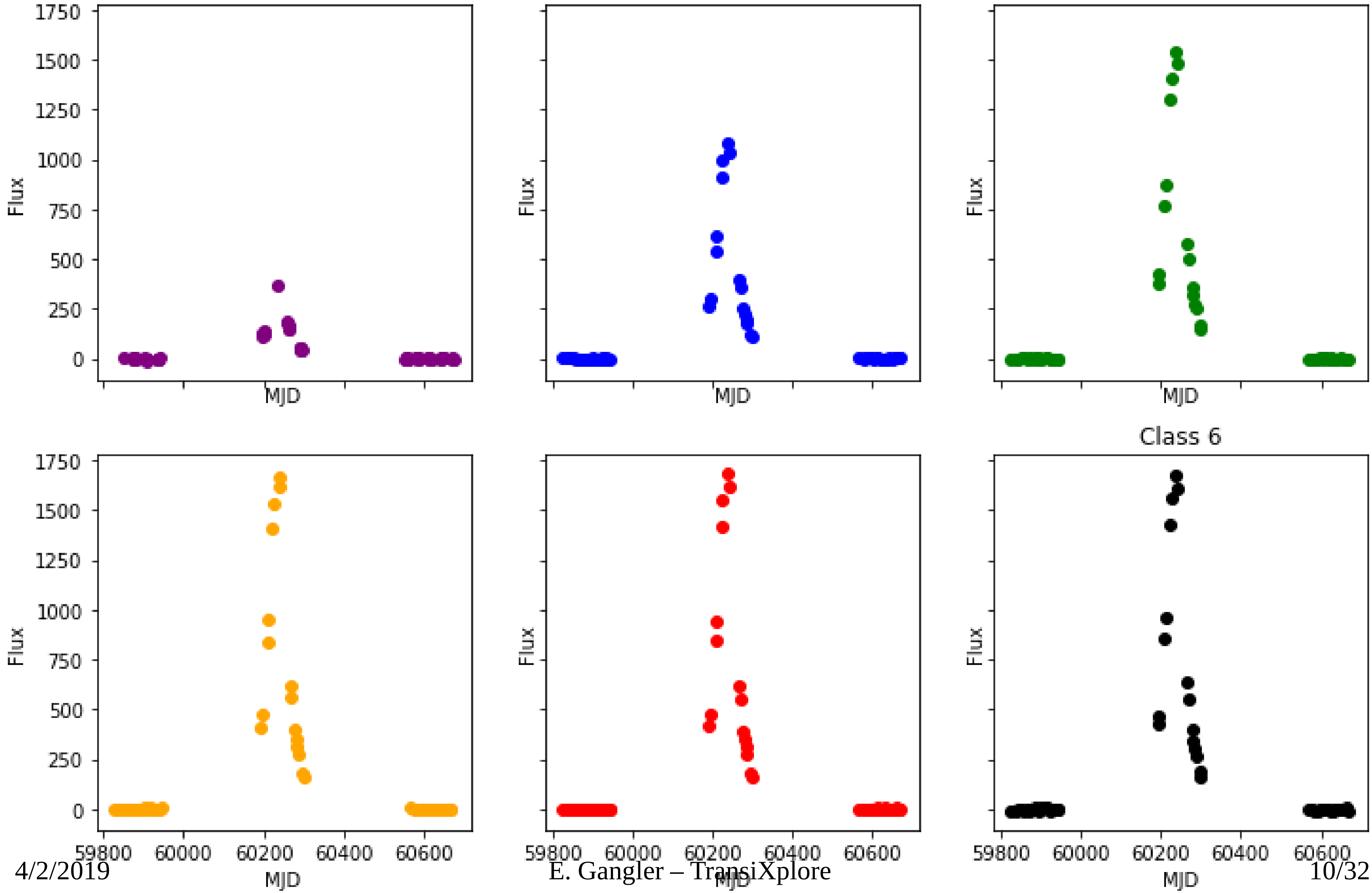
- Données ***réalistes*** simulées
  - 3 ans de fonctionnement de LSST, 1/150ème de l'empreinte
    - **Volume** : ~450 000 000 mesures

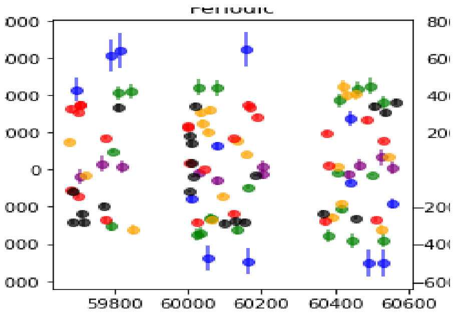
      (équivalent à ~7 jours de transitoires hors astéroïdes)
    - Trous dans les séries (saisons d'observation)
  - Problèmes de **représentativité**
    - Lot d'entraînement : 7 848 objets
    - Lot d'application : 3 492 890 objets
    - Entraînement mieux mesuré
  - 15 classes d'objets
    - 5 galactiques, 9 extragalactiques, **1 inconnue**



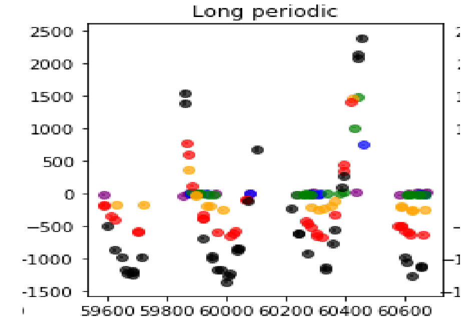
Quelques exemples

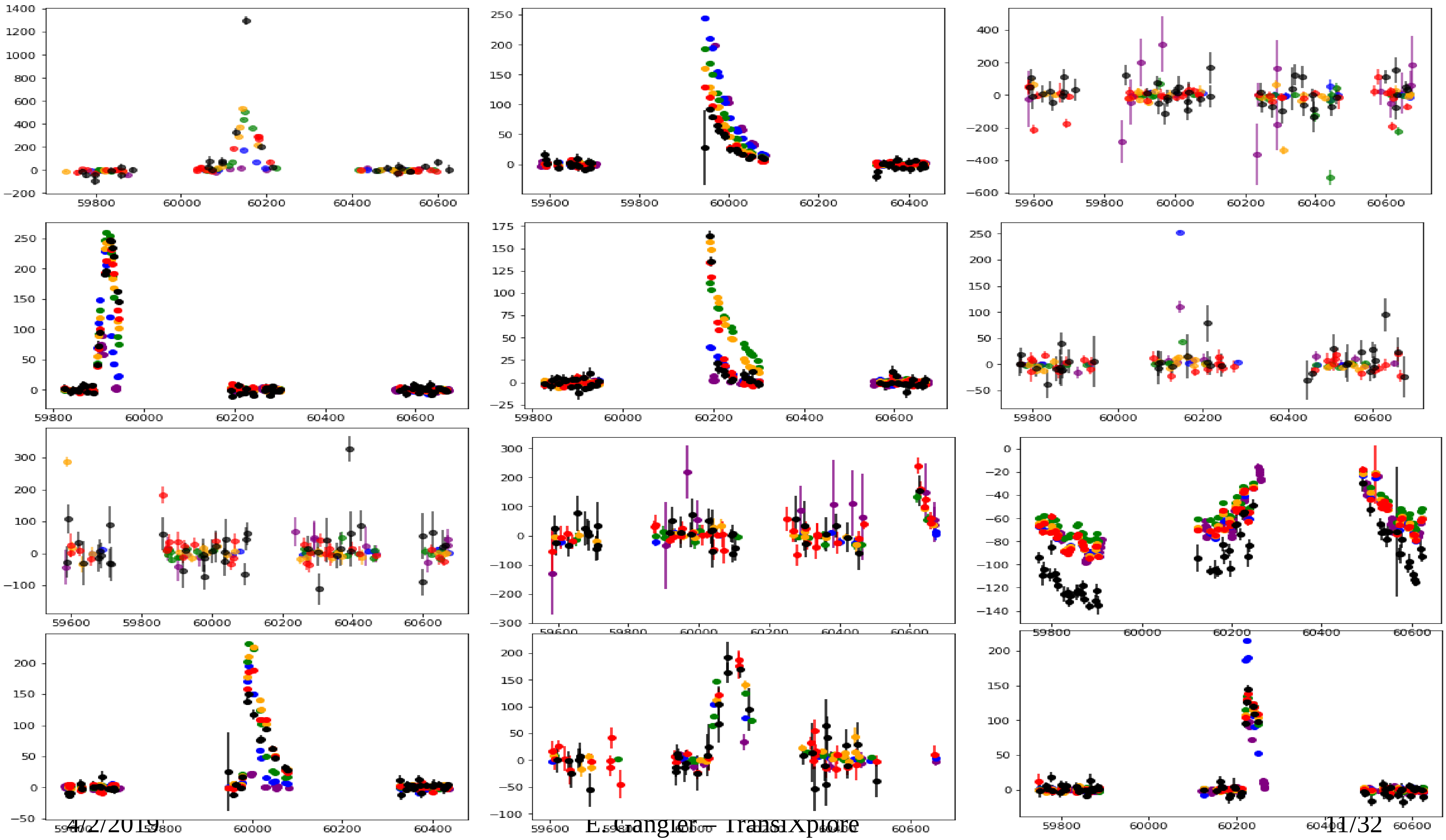
1/2 000 000 000 de LSST





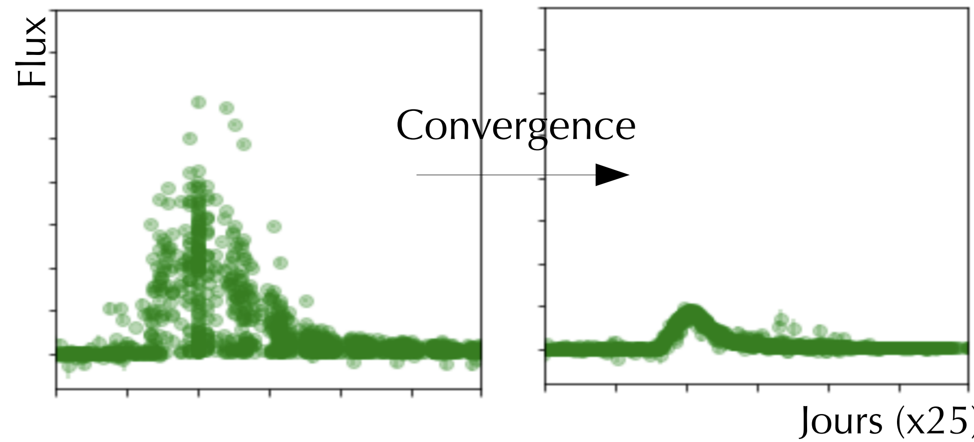
Les 14 classes...





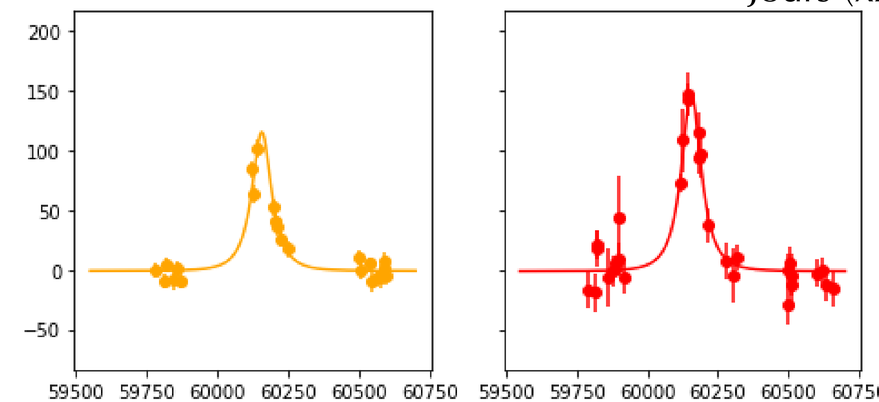
Données PLAsTiCC

- *Modèle complet !* (corrélation entre bandes, variabilité interne aux classes)
- Séries longues ; alignement des données → *temps de convergence* du modèle
- Temps de développement…



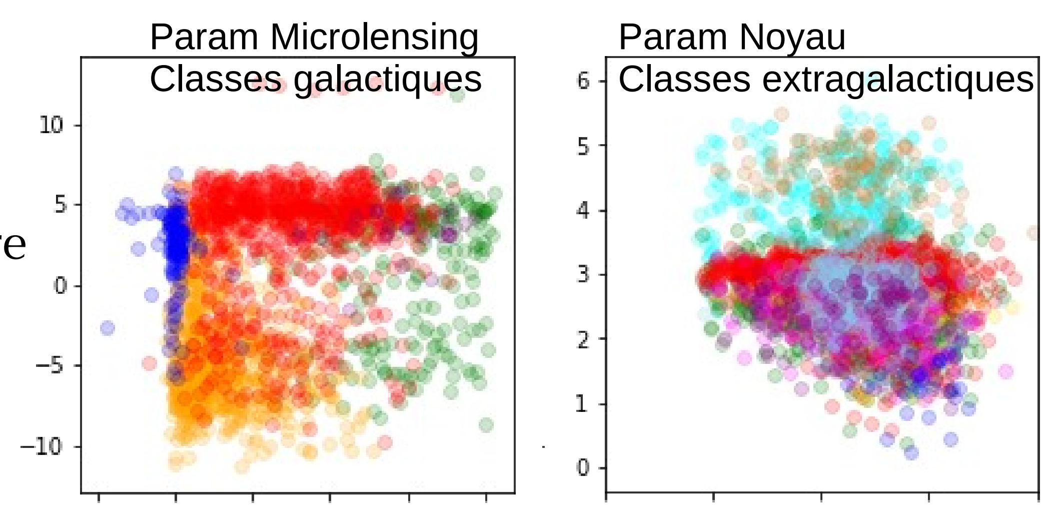
- *Templates exacts* pour le *microlensing*

  → kernel public, utilisé par les soumissions dans le top #10



- *Extraction de motifs pour les étudiants*

- Code disponible sous github/TransiXplore



Plasticc : retour sur expérience

Astronomes

1	—	Kyle Boone	0.68503	104	1mo
2	▲ 2	Mike & Silogram	0.69933	176	1mo
3	▼ 1	Major Tom	0.70016	366	1mo
4	▼ 1	AhmetErdem	0.70423	233	1mo
5	—	SKZ Lost in Translation	0.75229	343	1mo
6	▲ 2	Stefan Stefano			
7	▲ 3	hklee			
8	▼ 1	rapids.ai			
9	▼ 3	Three Muskete			
10	▲ 3	J&J			

Thx #9 !

1	**Representative light curves for hard classes** Raman 4 months ago	last comment by Raman 4mo ago	0
25	**Results on the unblinded test set** Kyle Boone 9 months ago	last comment by Sergey Zlobin 1mo ago	2
4	**Error Importing Sklearn Cross_Validation** Danny a year ago	last comment by Misba 1mo ago	22
3	**How to install a library in a kernal** Manoj 10 months ago	last comment by Mezianek 4mo ago	5
33	**20th Place Solution** Giba 10 months ago	last comment by Akhilesh 5mo ago	8

Discussions
Encore un peu actives !

- First of all I would like to point to @manugangler's kernel, where he provided a method for fitting light curves to microlensing events. The features obtained from there worked wonderfully for us, especially after taking passband-wise ratios and differences for the microamp and microbase features, respectively. Einstein time is the most important feature of my model.

Plasticc : retour sur expérience

- **Soumission gagnante :**
  - *Adaptation de domaine et augmentation de données*
  - Processus gaussiens (**temps de calcul !**)
  - Extraction de motifs
  - Classificateur *Light Gradient Boost Model*

Step	Approximate time to run
Augmenting the training set	1 hour
Fitting the GP and calculating features	**10 objects/s -> 100 hours for test set**
Training the model	20 minutes
Generating predictions	40 minutes

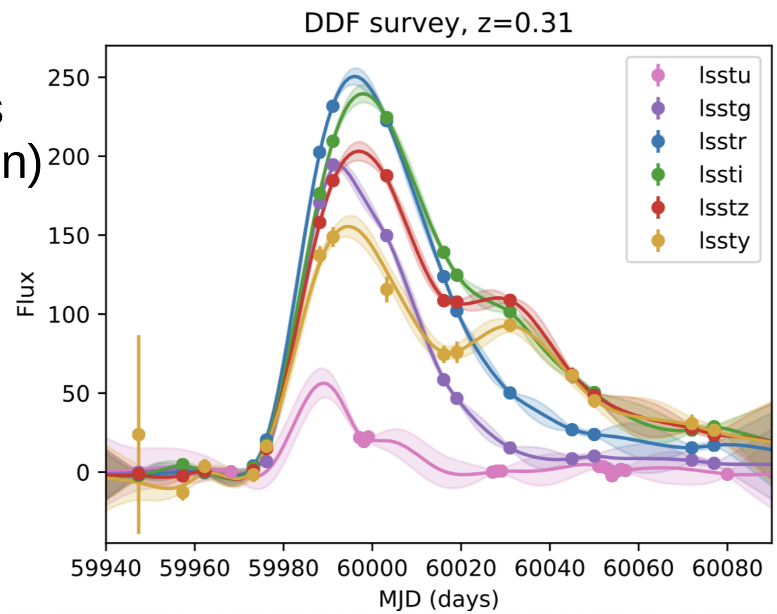
- #2-9 : peu de diversité…
  - *Ensemble de classificateurs*, 2 *RNN*
  - #5 : Ajustement par templates (sous-optimaux) pour déterminer les motifs
- #10 : PELICAN (Johanna Itam-Pasquet, CPPM, invitée au workshop TransiXplore)
  - Autoencodeurs sous-performants → piste de collaboration
- **Remarques :**
  - Kernels disponibles, beaucoup d'information à exploiter
  - « triche » concernant la classe 99
  - Résultats tributaires du modèle de photo-redshift
  - Notre approche est pertinente vis-à-vis des résultats

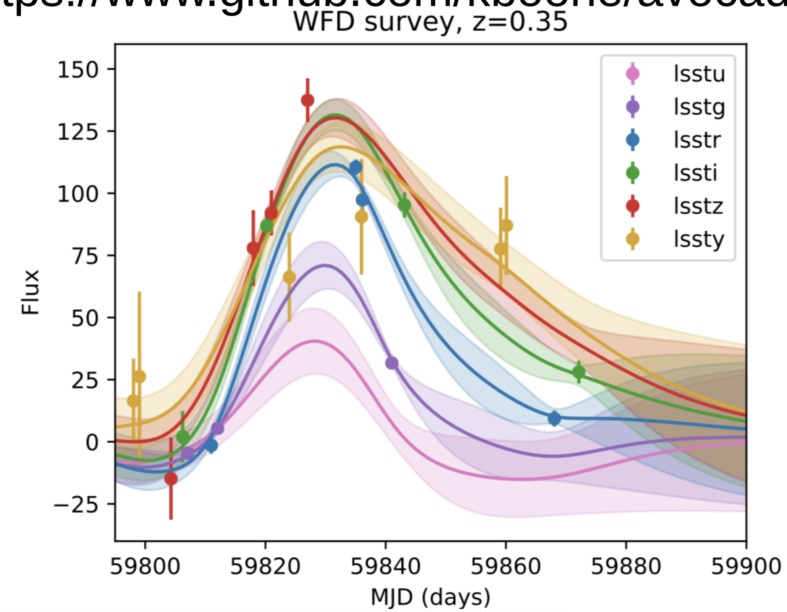
Quid des shapelets ?

Avocado (Kyle Boone)

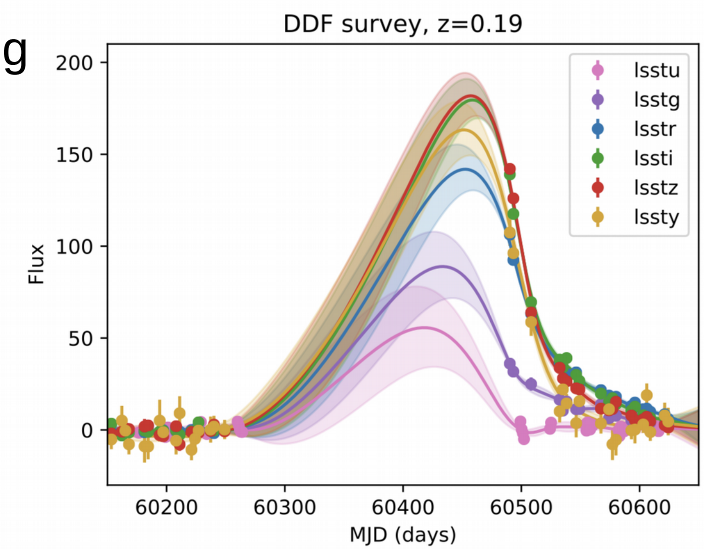
https://arxiv.org/pdf/1907.04690.pdf
https://www.github.com/kboone/avocado

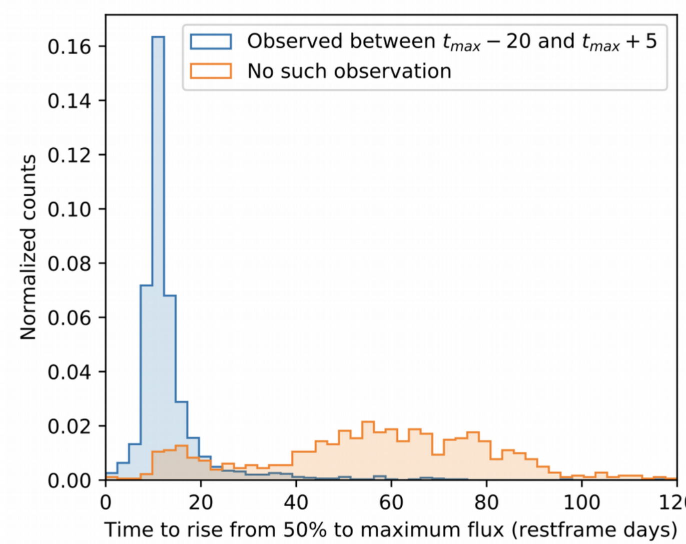
Gaussian processes
(with band correlation)



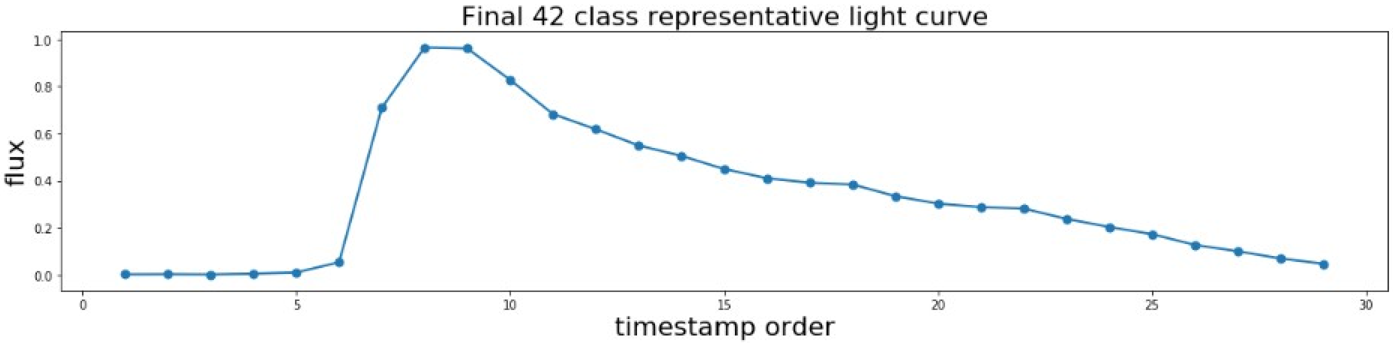


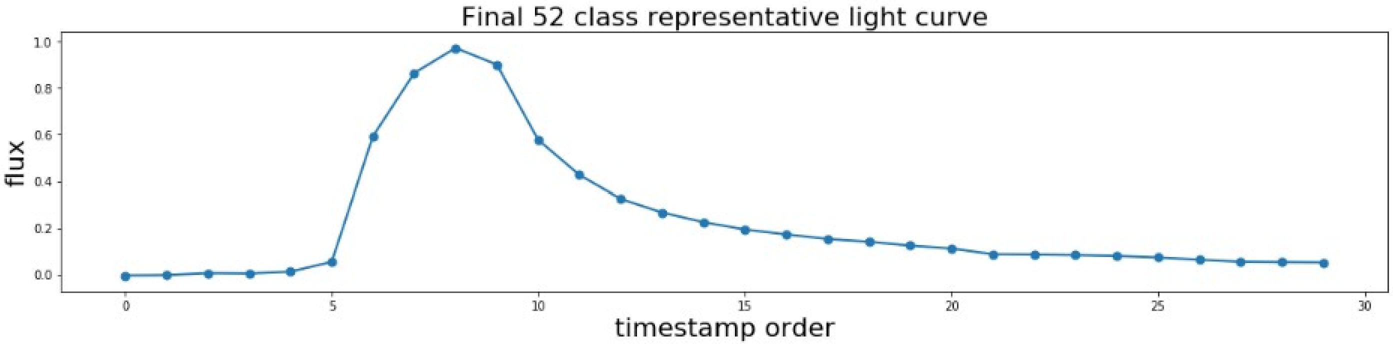
Feature engineering
(data quality)

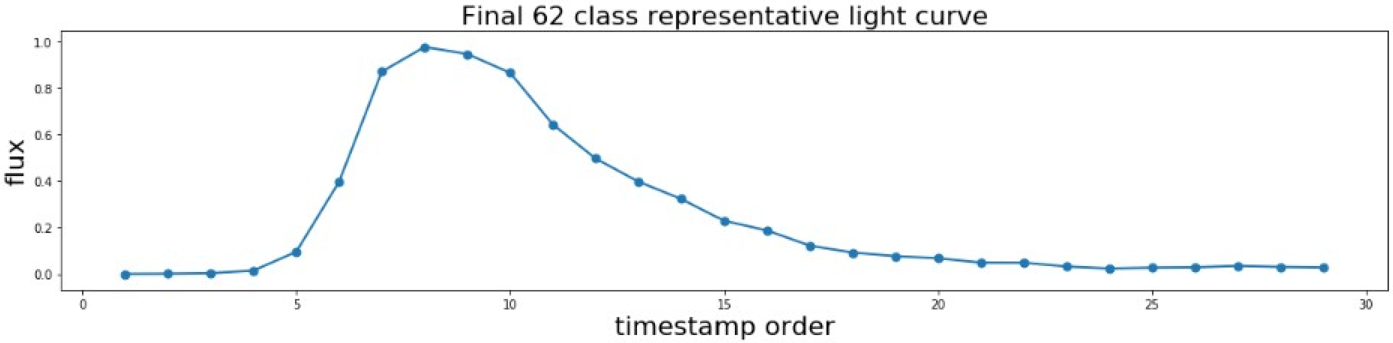




Templates approches

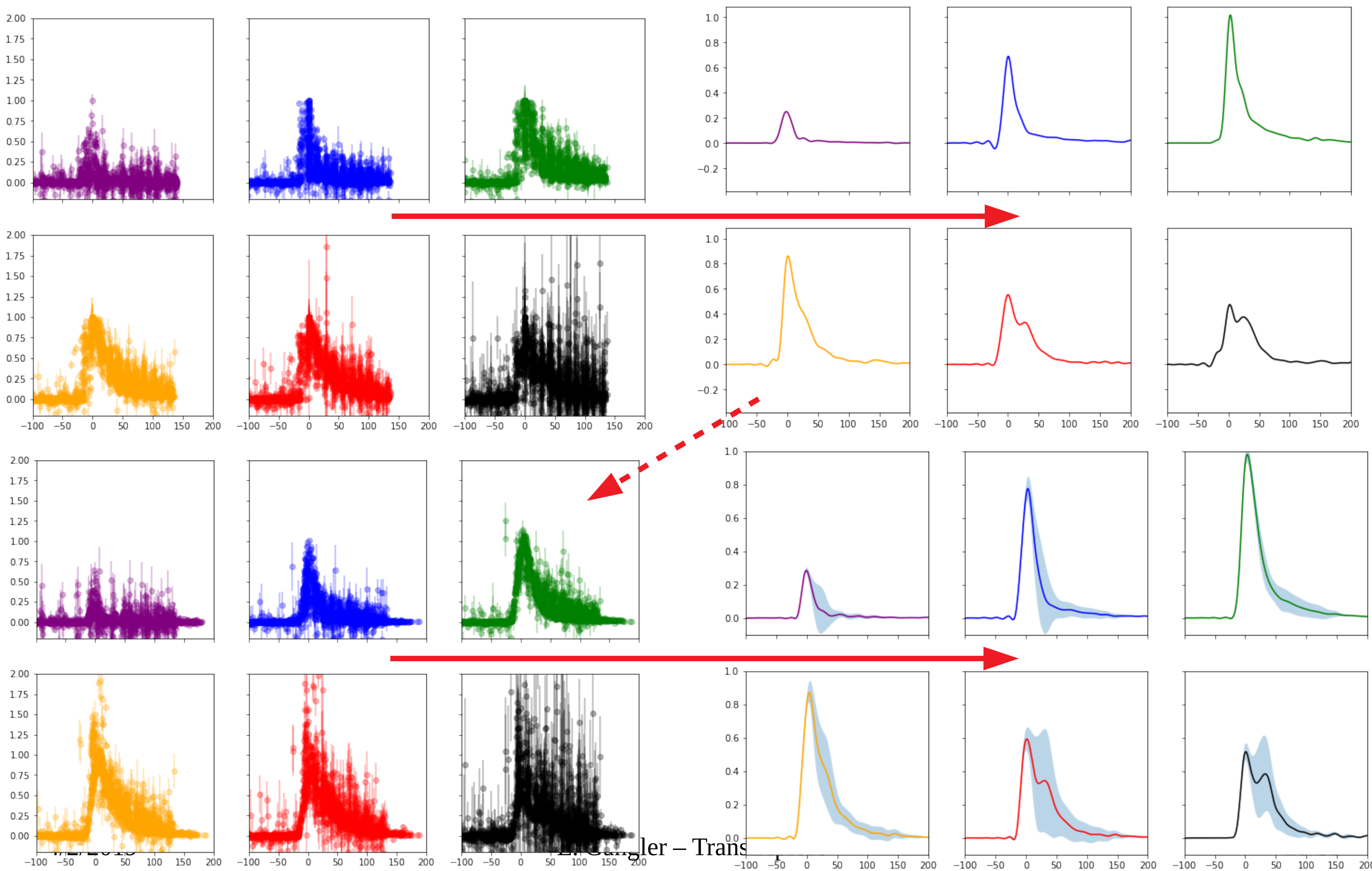




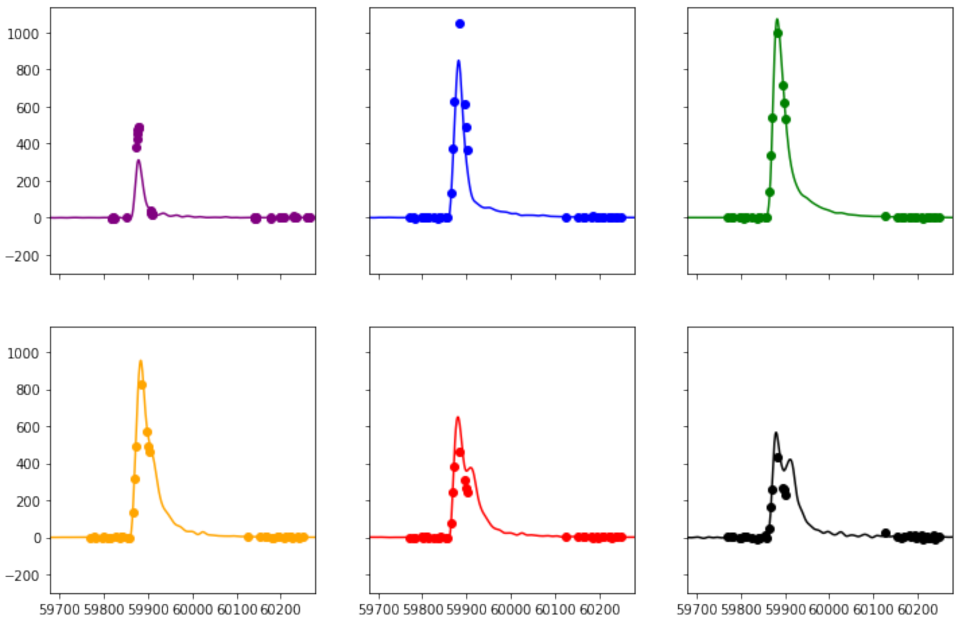


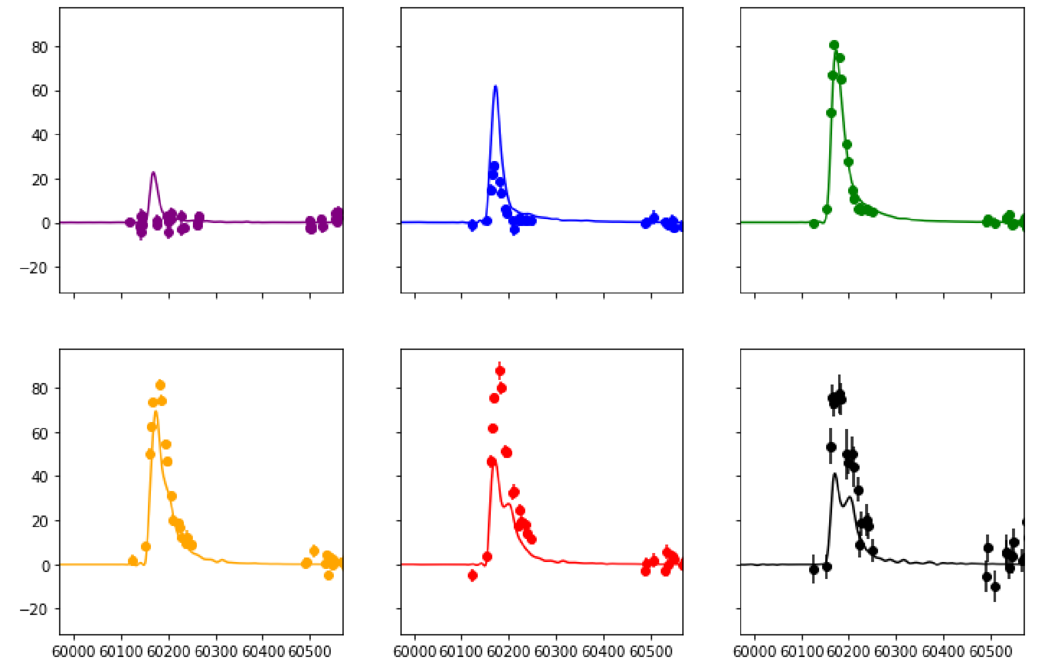
Template fitting approach

- Time interpolation :
  - → choose a 3rd order B-spline model
  - Fit model by minimizing the distance to data
  - Optimize time step by cross-validation → ~10 days on SNIa
- Initialization and convergence tricky
  - Align on max data for t0 guess.
  - Best scale approach : initial scale given by maximum → set to 1.
- Convergence
  - Alternate model fitting and object properties fitting
  - 1st step : Fit for t0, not Scale (gives good initial template)
  - 2nd step : fit for t0 and scale → converge the model
  - 3rd step : build an error model, and refit supernovae
    - (not yet done : reconverge the model with error model on!)



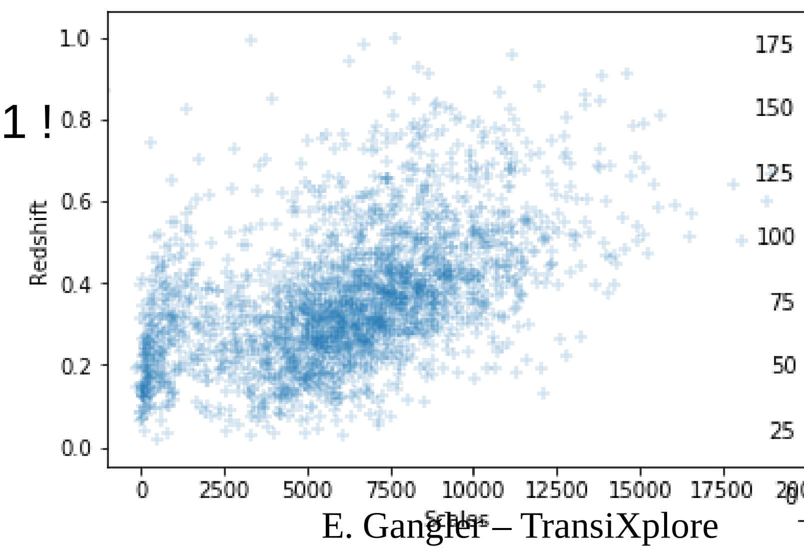
A few bad fits

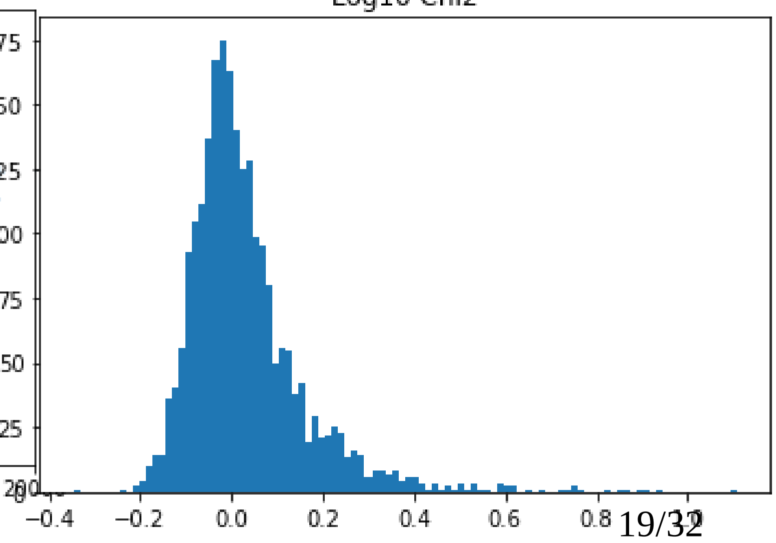




Reasons :

- Redshift effect
- Multi-classes in 1 !





Some conclusions

- Fitting a template :
  - Shared idea in the community
  - Provides restricted set of features : chi2 & scale vs redsfhit
    - Model can be conplexified ad lib (variational parameter, K-means based sub-models)
  - But takes time to get good results
  - To come : fit other models
- Goal for december:
  - Compare model-based selection (and models) to output of a shapelet-based approach

Application aux données SN2010PCC

- *Data challenge précurseur*
  - 1100 supernovas
  - 2 classes : SNIa et Non Ia
- 2 Patrons
  - Discrimination par rapport de vraisemblances
    - Aucun autre motif utilisé !
    - AUC 0.843 : comparable à Lochner 2016

→ **preuve de principe !**