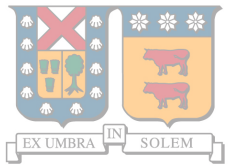
Refining Exoplanet Detection Using Supervised Learning and Feature Engineering

SLIOIA (CLEI) and CLEI-EJ

M. Bugueño, F. Mena and M. Araya

Departamento de Informática
Universidad Técnica Federico Santa María

April 2019



Summary

1 Introduction

2 Background
- State of the art

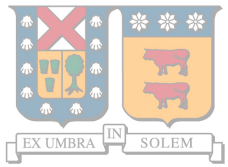
3 Data
- What and Where
- Inside the data
- Data generation

4 Models and Methods
- Feature extraction
- Learning models
- Evaluation

5 Experiments
- Results
- Detail of results
- Tagging

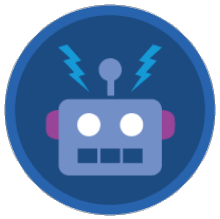
6 Working on

7 Conclusion



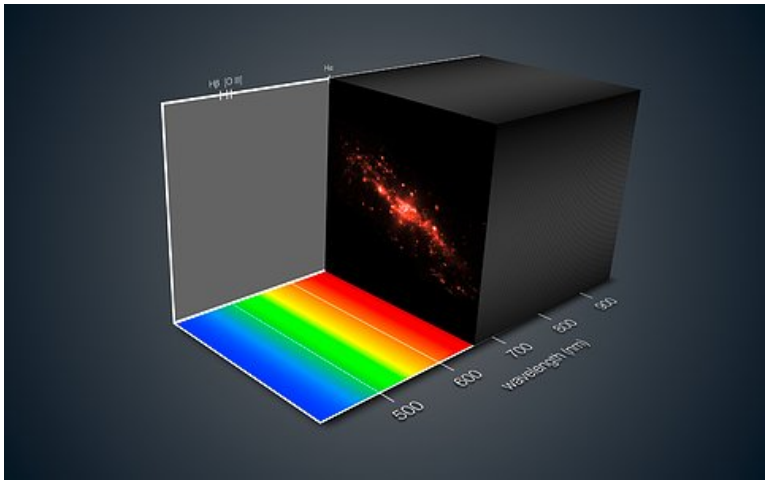
Motivation

A lot of manual process and analysis

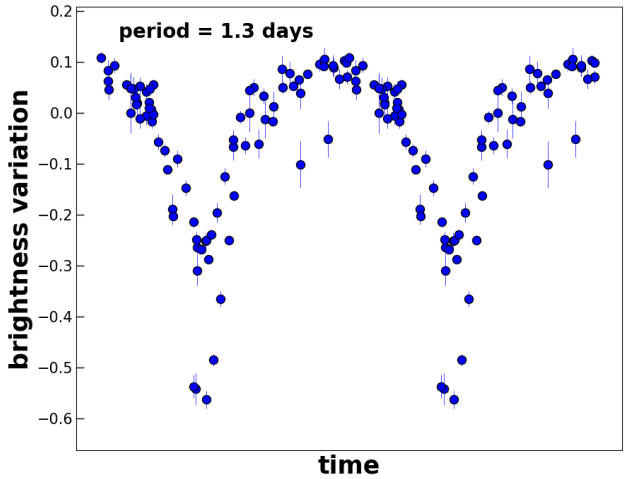


Differents fields of application

Data types in astronomy



(a) Data Cube



(b) Time series

source_id	ra	ra_error	dec	dec_error	parallax	parallax_error	phot_g_mean_mag	bp_rp
	deg	mas	deg	mas	mas	mas	mag	mag
4284585432346951424	277.0184221091102	0.46188309396256755	5.631563353695939	0.4862023067076275	0.5512842297805866	0.6189512459649872	19.239384	1.4463387
4284612885764705280	278.5055205273139	0.4026910740133187	4.869279454892372	0.4280067945557229	0.45571055041155867	0.3919193257674781	19.616207	2.373581
4284623816466582528	278.7796652692737	0.4421684796713564	5.072389376937544	0.6508133544034742	0.9780940030018949	0.8269676335062147	19.591053	
4284617288105228416	278.41751955947575	0.8859155082442142	4.937915872054388	0.9325826016333091	-0.5835659876031645	0.7321825803262654	20.520485	2.1895103
4284586325690985344	277.2424770752404	0.05535184499199222	5.5815214333073815	0.06184951904780605	0.2683497429786385	0.07385509146288734	16.142056	1.845604
4284613160642632064	278.4884061868495	0.19687204524175128	4.883988786277995	0.2130844399294118	0.2763052829931508	0.20188849524971694	18.643427	2.1965847
4284582550424225408	276.9239986836356	0.4880305524395156	5.5165784450770765	0.4647396167634752	-0.07380309135977635	0.5648456189964629	19.62505	1.3735962
4284607323794534528	277.15093468245533	1.233788359588574	5.860497263110219	1.7621337955589698	0.8687865723060184	1.5484880989128447	20.613918	1.6257915
4284626089721631360	278.40552511360397	0.3942340351107299	5.041117911416378	0.44380568123691505	1.1518533500147796	0.4289744337441624	19.706867	1.6361904
4284625019063919488	278.6795510178577	0.07861150429451681	5.107132089171365	0.08812810490449971	0.14191457890727033	0.09859248730278555	16.66429	2.309392
4284598351601693824	276.94422476030593	0.2198125509124454	5.6921797663250056	0.22191188665832123	-0.6465546094783365	0.25220086539158176	18.16406	1.3938389
4284582305599836544	277.00047964009906	0.28581032104433607	5.5481606610170875	0.2729011995435271	0.7701878632716691	0.3104060137312627	18.893915	1.5011101
4284623541588737152	278.71037006527007	0.9076172955280782	5.057887377634662	1.0390009489213046	0.22453398549395054	1.1150025835163397	20.433064	1.0395985
4284601104687970176	276.85332176745095	2.777810772376081	5.68663409929287	1.2060048830033478			19.623758	
4365477121264227584	253.70648592823147	1.9157537129191289	-4.815643678668132	1.9273872340008684			21.065166	
4334792500513544448	255.59229364173092	0.14641306644767804	-11.490641698930379	0.13258813061444458	0.3895549224298378	0.20095446249452514	17.65269	1.4059525
4334806691083238784	255.51754794398323	0.3477344180358136	-11.361887571012234	0.3109230868301414	0.5675282757305556	0.4407907319165841	19.196663	1.7098827
4334830639821736960	256.11797332199006	0.27832473760053106	-11.176205065500997	0.20587110300280598	0.8524347032125424	0.32455228003638825	18.763445	1.6481686
4334799681699369088	255.60579609642434	1.5214571774132257	-11.403624200072843	1.300801876595436	1.3396959771653827	1.627442461752697	20.617178	1.455513
4334807824957225472	255.346681367043	0.41616865305901213	-11.427488037140206	0.3594001816285756	-0.08984983377644823	0.5802020112350367	19.197062	1.0878944

(c) Catalog

(d) Image

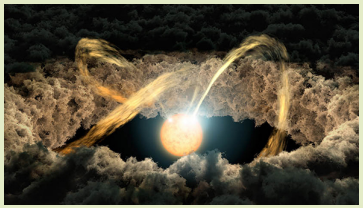
What is an exoplanet ?

Planets orbiting stars outside our solar systems are called **extra-solar planets** or **exoplanets**.

Birth of a planet

The theory of extrasolar planets is under development until these days (since the first confirmed detection of the giant planet **51 Pegasi b**)

- Current theories suggest that the dust particles of the protoplanetary disk begin to collapse by gravity forming larger grains.
- If these discs survive to stellar radiation and comets, the matter continues compacting giving way to a planetoid.



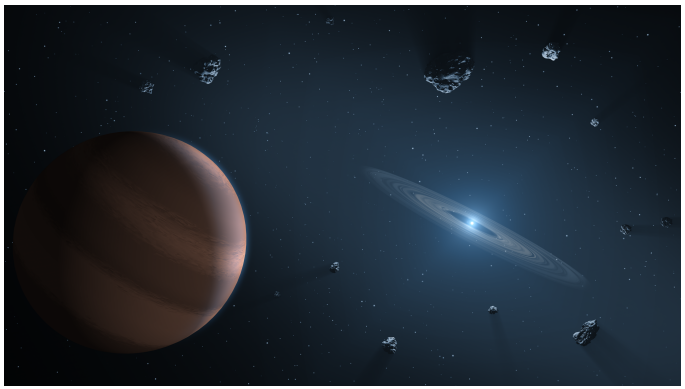
Images by [1] and [2]

1. http ://www.spitzer.caltech.edu/news/1876-feature16-07-Light-Echoes-Give-Clues-to-Protoplanetary-Disk
2. https ://www.space.com/19100-alien-planet-birth-alma-telescope.html

Why is difficult to detect exoplanets ?

Detecting these planets is a challenging problem !

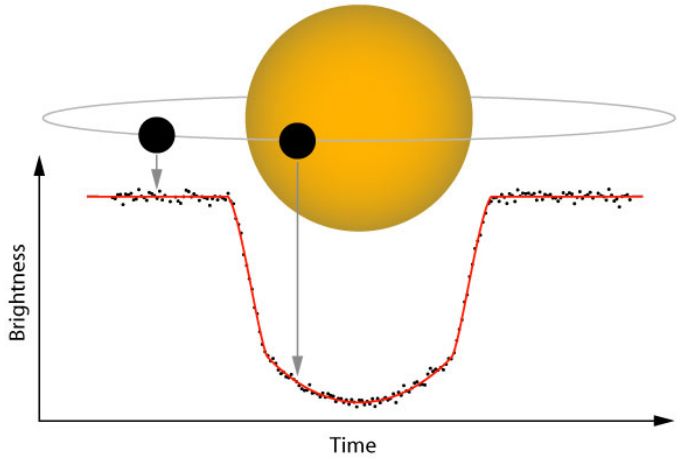
- They emit or reflect very dim magnitudes compared to their host stars
- They are very near to their host stars compared to the observation distance
- Fine-grained analysis is needed



Detection methods

The most successful detection mechanisms are the **indirect**, some of them :

- ♂ **Radial velocity**, which studies the speed variations of a star product of its orbiting planets, analyzing the spectral lines of this one through the Doppler effect.
  Successful, but only effective on giant planets near its star
- ♈ **Transit photometry**, photometric observation of the star and detection of variations in the light intensity when an orbiting planet passes in front of it.
  Efficient, detect high-volume planets independently of the proximity to its star

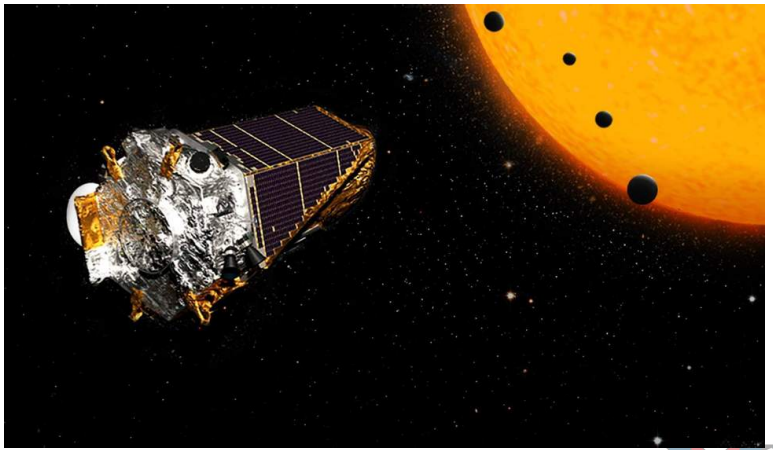


What is needed ?

Problem : A large volume of data is being generated today !

- The use of automated methods could reproduce the astronomer analysis to decide if the data supports the existence of an exoplanet or not.
- 400 billions of planets waiting to be discovered [3]

Fortunately, technological advances in photometry have allowed experiments like the space observatory Kepler to have sufficient sensitivity for detecting a greater range of exoplanets.

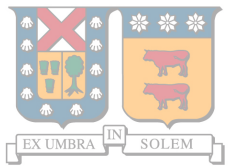


3. Looking for life in Exoplanets | Dawn Gelino | TEDxCERN

Previous work I

A typical research are detections of **RR Lyrae Stars**, because their intensity varies through time independent of a planet.

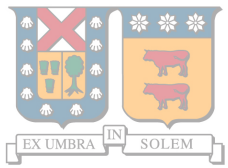
- Richards et al. (2011) presents a catalog of variable stars and manually extracted light curve specialized features from simple statistics and other features based on the period and frequency analysis of a LombScargle fitted model
- Donalek et al. (2013) classify variable stars from the Catalina Real-Time Transient Survey (CRTS) and the Kepler mission extracting similar features from the light curve to Richards et al.
- Mahabal et al. (2017) also with the propose of classify variable stars, transformed light curves into an image (grid) that represent the variations of magnitude through the variations of time intervals.



Previous work II

- Hinners et al. (2018) presents different machine learning techniques and models with the objective of classify and predict features over the same data as we use. They extract some statistical features from the light curve but they were not interested on exoplanet detection.
- Thompson et al. (2015) used unsupervised learning to reduce dimensionality of Kepler light curve and cluster together similar shapes with the focus of find candidates object of interest. They use k-NN model to predict.
- The inspiration come from the **Autovetter** project (McCauliff et al. 2015) : used a Random Forest model based on features derived from Kepler pipeline statistics to identify candidates.

Our Objective : Exoplanet detection through transit light curve with ad-hoc feature extraction and machine learning.



Where the data come from ?

Kepler space observatory

- Data collected by Kepler Mission (launched in 2009 by NASA) with the goal of searching similar planet to Earth.
  - Around 65% of exoplanet discoveries have been detected thanks to Kepler Mission [4]
- The Kepler Objects of Interest (KOI [5]) dataset is provided by MAST (Mikulski Archive for Space Telescopes) archive.
  - It containts 8054 KOI's.

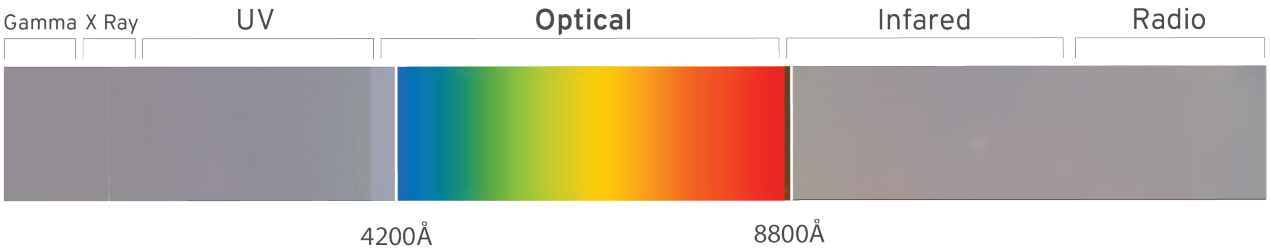


FIGURE – Kepler field of view of the spectrum.

4. https ://exoplanetarchive.ipac.caltech.edu/docs/counts_details.html
5. http ://archive.stsci.edu/search_fields.php ?mission=kepler koi

What are the data ?

The Labels

- Every record is associated to a Kepler Object of Interest labeled as :
  - *Confirmed* : those that have been confirmed as exoplanet, through extensive analysis.
  - *False Positive* : those that were initially selected as candidate exoplanets but there is additional evidence that shows they are not.
  - *Candidate* : those that are still under study.

  according to Nasa Exoplanet Science Institute[6]
- The reasons to catalog as a False Positive are observation that did not match with the star position on study. Also can be that the deep of the even transit was statistically different to the deep of the odd transits, showing a binary system.

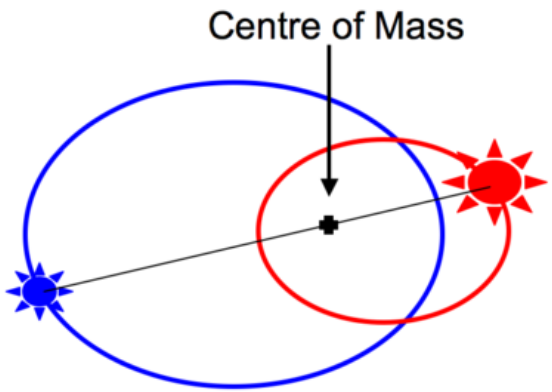
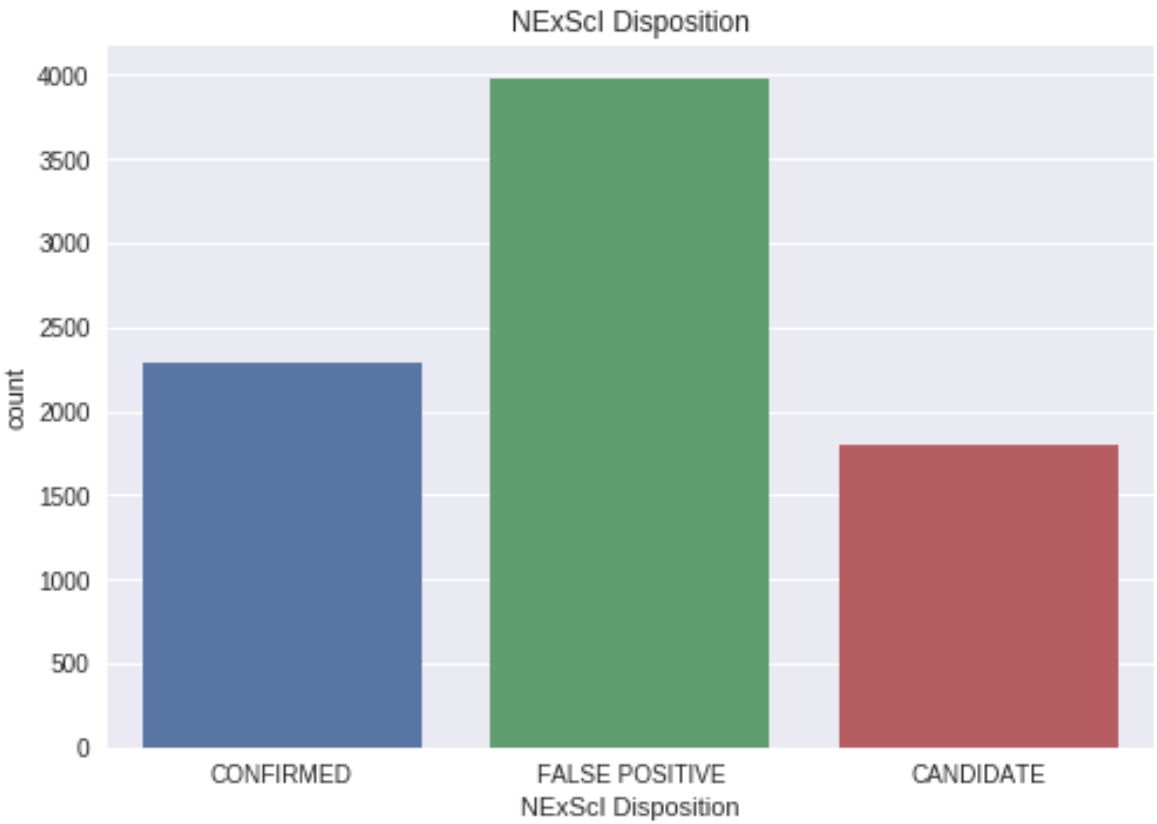


FIGURE – binary star system

What are the data ? - Class distribution



Tag	Data (&)
Confirmed False Positive	2281 (36.46) 3976 (63.54)
Candidate	1797

FIGURE – Categories distributions on collected data

What contain each KOI

Error

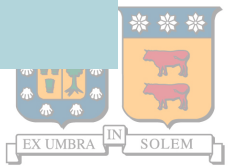
- The error associated to each measure

Time

- The time (in Julian Date - January 1, 4713 BC) when the measure was made

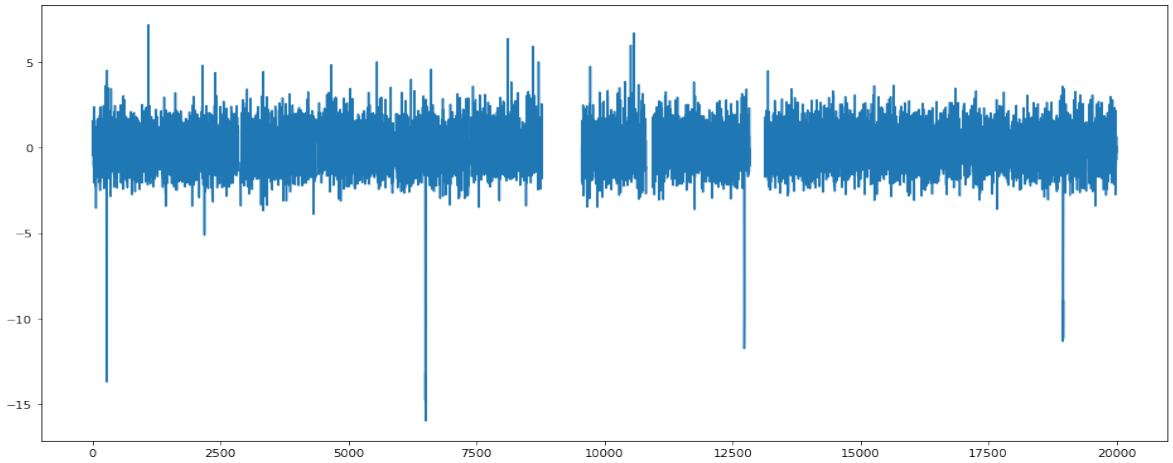
Raw light curve

- The raw measurements (intensity) of star light, with about 70000 measurements
- Frequency sampling rate of 30 minutes approximately (4 years of observations)
  - Measurements are not recorded uniform
- On average, the missing data is about 23% , this mean approximately about 55000 *effective* measurements.



Fill missing values

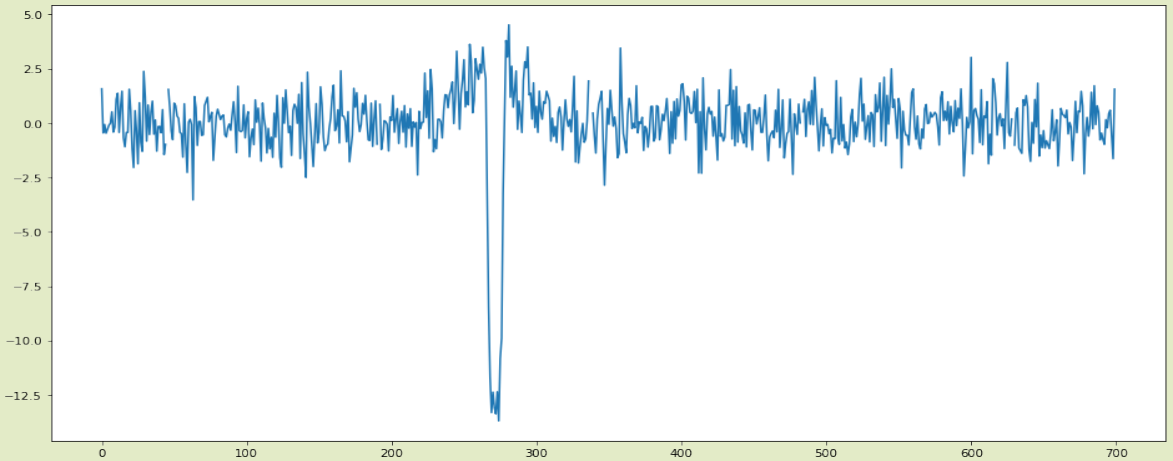
- Fill with zeros
  - Stationary state of the system (not eclipsing)
  - Expected value, $E[I] \approx 0$
- Fill the gaps with linear interpolation
  - To get a smooth version in extense continous missing data



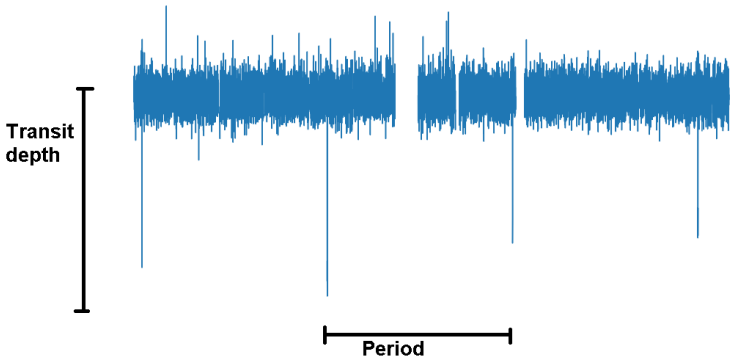
What contain each KOI

Filtered light curve

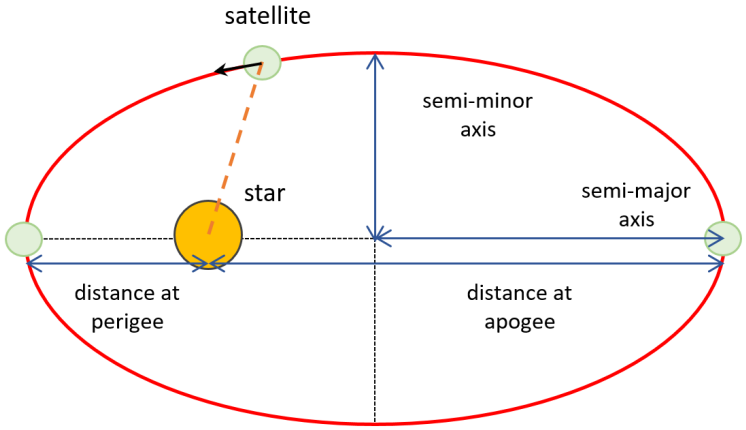
- Light curve with a **whitened filter**. This transform led a constant white noise on the light curve, i.e., the higher signal get amplified and give a more uniform signal.
- The white noise is a random signal that have the same intensity on different frequencies, which gives a spectral density of constant power.



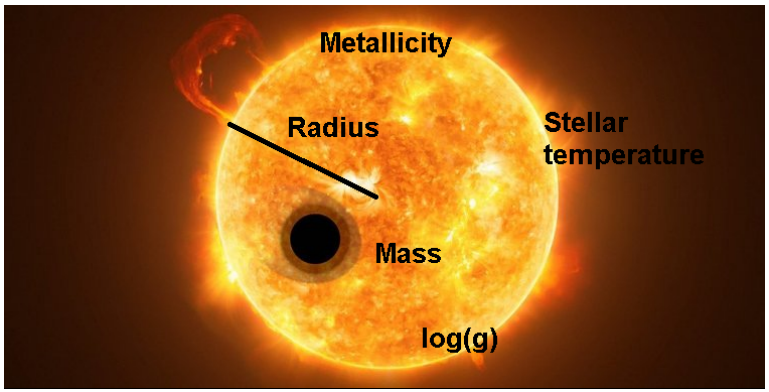
What does each KOI contain ? - Metadata



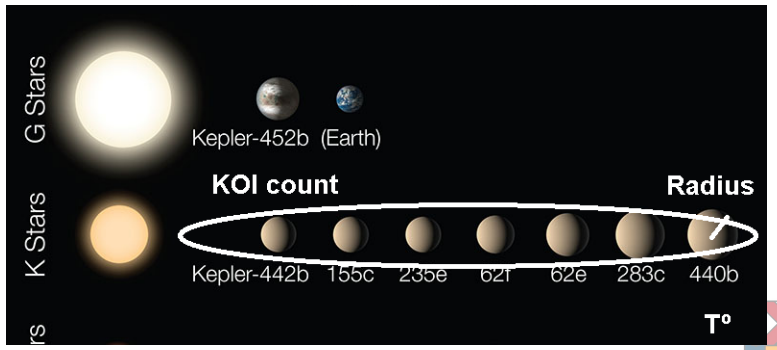
(a) Planet metadata



(b) Transit metadata



(c) Stellar metadata



(d) Planet metadata

Data process

- A system can contain various exoplanet orbiting the host star.
- There are systems that have from one to seven objects under study.

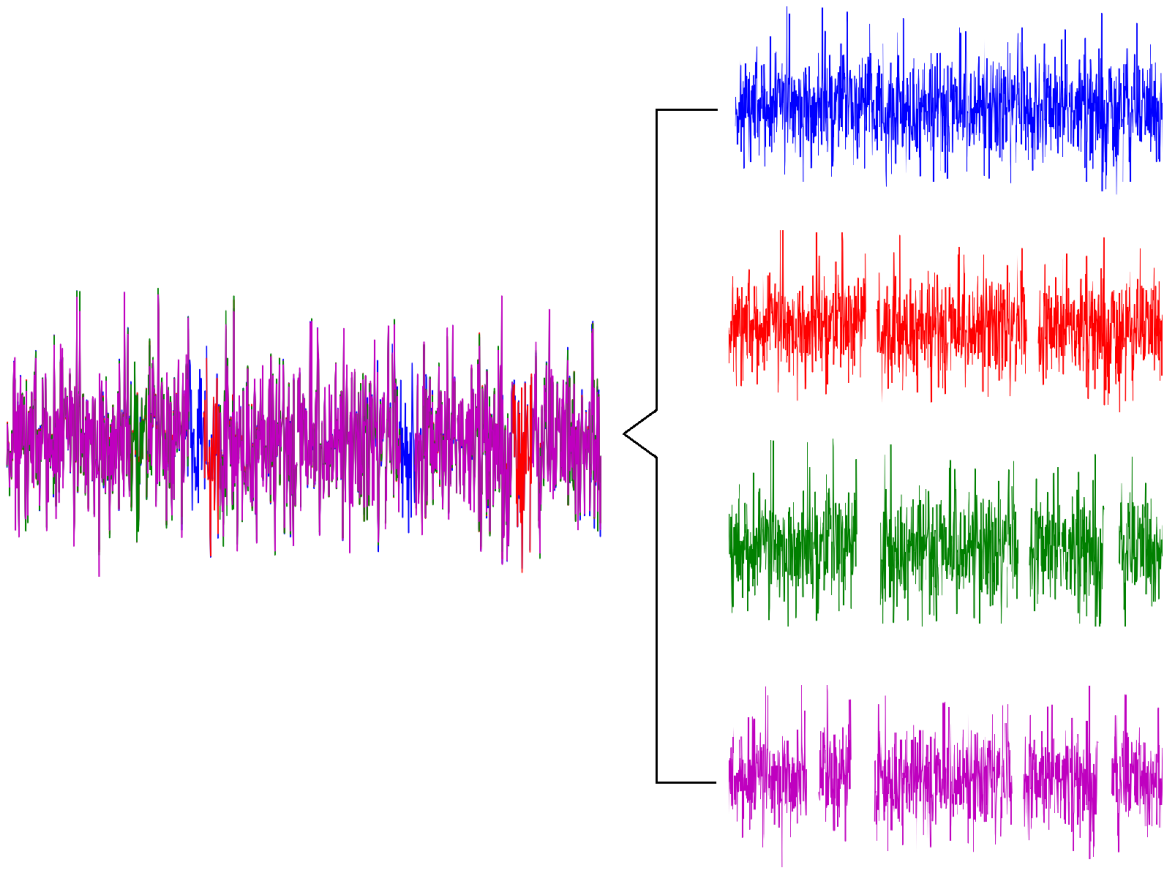
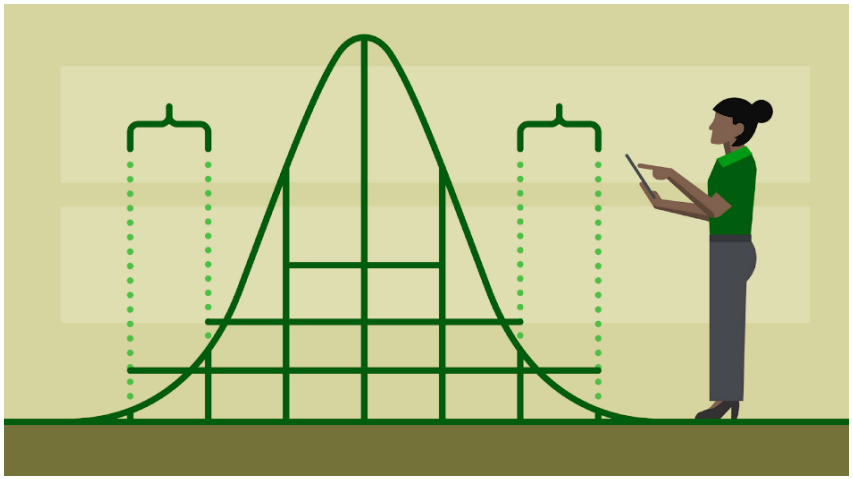


FIGURE – Light curve separation

Manual - baseline

We used extraction techniques specialized on time series, which in this case corresponds to measurements of intensity of the light along time, inspired on the library Feature Analysis for Time Series (**FATS**[7]) for Python.

- Amplitude
- Slope
- Max
- Mean
- Median
- Median absolute deviation
- Min
- Q1
- Q2
- Q31
- Residual bright faint ratio
- Skew
- Kurtosis
- Std

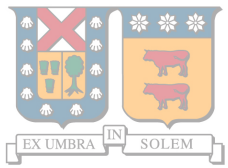


To this we decide to add the metadata.

7. https ://github.com/isadoranun/FATS

Selected Metadata

Based on previous small knowledge	
Planet+Transit	**Stellar**
Radius	Radius
Temperature	Temperature
KOI count	Metallicity
Period	Mass
Transit Depth	Log(g)



Automatic

- We use : **Unsupervised Learning Methods**
  the objective is to find intrinsic patterns among all the data independently from task.

Pre-process

- Firstly we applied a Discrete Fourier Transform[1] to the light curve.
  - It transforms the data from the time domain in which the measurements where obtained, to the frequency domain where the signal was generated.

1. *this method is designed to analyze periodic signals, which is exactly the case of transit light curves.*

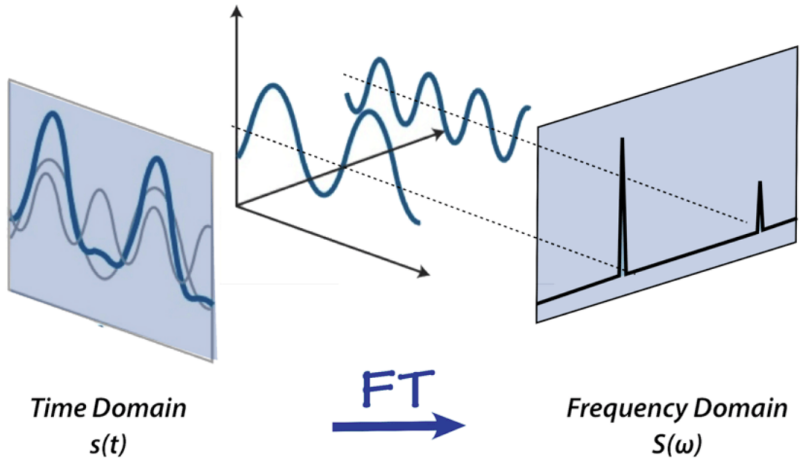


FIGURE – image by http ://mriquestions.com/fourier-transform-ft.html

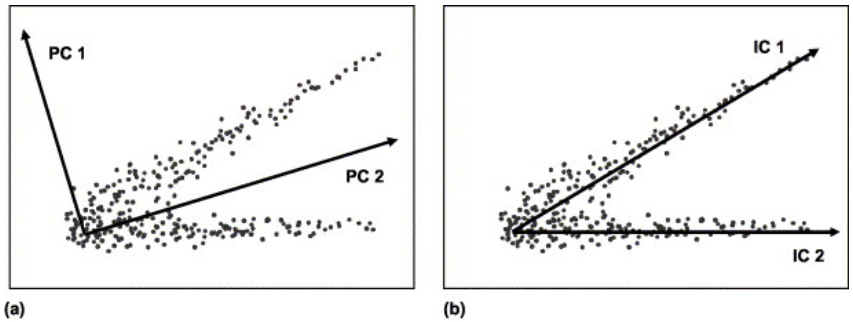
Automatic

PCA

- The **Principal Component Analysis**, as a linear method that projects data into a lower dimensional space (the higher variance vector).
- Why PCA ?
  - Has been applied to several applications, obtaining particularly good results on time series
  - Its great efficiency from specific optimizations over linear algebra methods with high dimensional data.

ICA

- The Fast iterative algorithm for **Independent Component Analysis** that finds statistically independent components of the data.
- Why ICA ?
  - Is focused on the signal abstraction, since it tries to detect the independently sources that, mixed, produce the observed data.
  - Differs to the uncorrelated componentes that finds PCA.



Machine Learning models

k-NN, regularization parameter : k

- Based on memory (i.e., non-parametric). Simple, but with good performance

Logistic Regression, regularization parameter : C

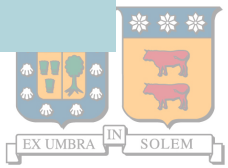
- Classify based on a probabilistic (logistic - sigmoid function) linear model.

SVM, regularization parameter : C

- A linear margin-based model... but with Kernel : RBF (Radial Basis Function)

Random Forest, regularization parameters : *depth* and T

- A ensemble of decision trees.. not linear !



Metrics

Classification

The exoplanet problem is an instance of **unbalanced binary** classification problem.

- Precision : ability to label one class when the object effectively was from that class. (inverse to contamination)

$$P = \frac{T_p}{T_p + F_p}$$

- Recall : ability to include all object that effectiveley are from one class. (similar to completeness)

$$R = \frac{T_p}{T_p + F_n}$$

- F_1-score : as harmonic mean between P and R

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

All these metrics reach their best values at 1 and worst at 0

Experiments

- Large amount of data to be processed ! So it was necessary to use a cluster provided by **ChiVO** [8] (Chilean Virtual Observatory)
- We have 6257 labeled data, corresponding to 121*GB*, and 1797 not labeled data (candidates), corresponding to another 33*GB*.
  - We used the (64/18/18)% of the labeled data for set-split. Approximately 4000, 1000 and 1000 registers were grouped as training, validation and testing sets. This last one represents the actual *target* unlabeled data that is unknown (Candidates).

Note : The selection of the hyper-parameters of the classifiers was not expensive in computational terms (dimensionality d much smaller than the original), unlike the feature extraction process.

8. http ://www.chivo.cl

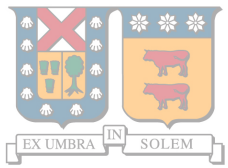
Experiments

- For automatic feature extraction techniques, fixed dimensions were experimented

	5	**10**	**15**	**20**	**25**	**50**
ICA	**0.711**	0.709	0.709	0.686	0.679	0.675
	5	**10**	**25**	**55**	**100**	**255**
PCA	**0.713**	0.701	0.701	0.699	0.702	0.689

TABLE – F1 score of the best classifier, Random Forest, in function of dimensionality.

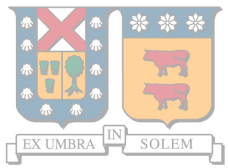
- Surprisingly enough, completing missing data with zeros produces a consistent improvement of ~ 0.1 in the F_1-score, while linear interpolation produced worse.



What about unbalanced data ?

1. **Undersampling technique** ; the majority class was subsampled

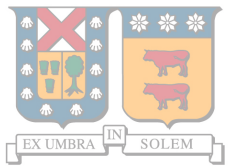
2. **Weighting** ; weights the different classes
   - This improved the results by ~ 0.1 on F_1 score metric.



Performance Results

Learners performance				
	k-NN	*Logistic Regression*	*SVM RBF*	*Random Forest*
Fourier + PCA	0.679	0.493	0.486	**0.713**
Fourier + ICA	0.679	0.493	0.486	**0.711**
OwnFATS	**0.666**	0.583	0.575	0.658
Planet meta-data	0.825	0.848	0.848	**0.870**
Stellar meta-data	0.766	0.718	0.751	**0.766**
OwnFATS + stellar & planet metadata	0.844	0.864	0.876	**0.883**

TABLE – F1 score on the classification of different models (learners) over the test set on the different representations generated.



Results analysis

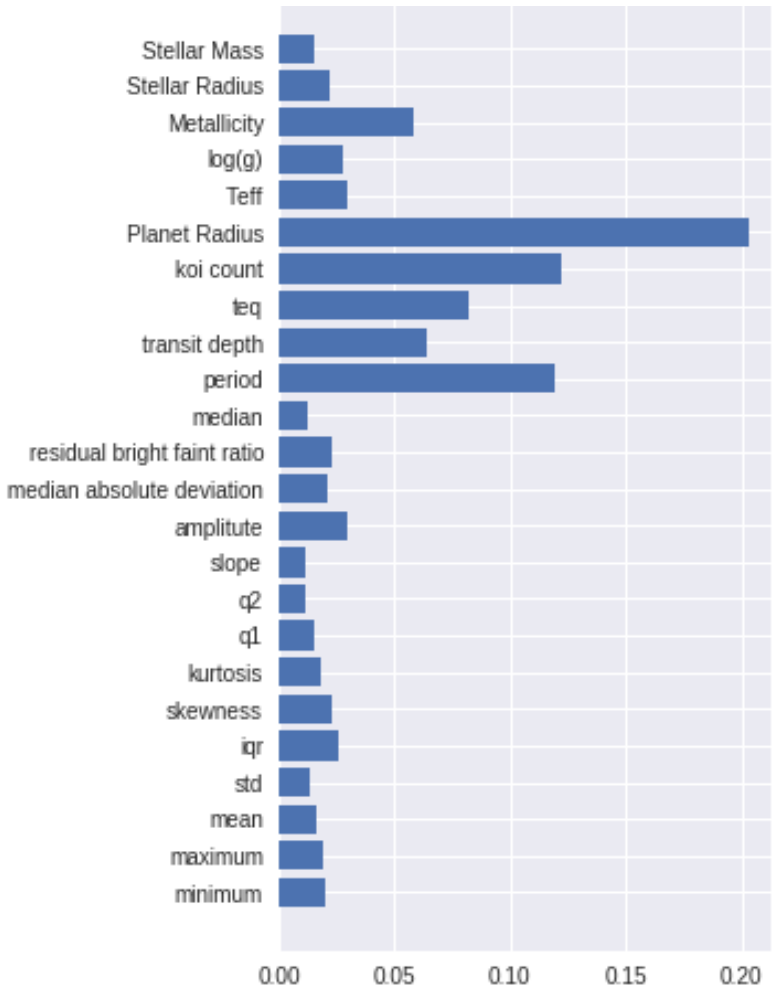


FIGURE – Random Forest feature importance

Performance analysis

- The best trained model, was Random Forest
- Performance of 88.3% for future classification by F_1 score metric
- In detail :
  - Radius, period and number of objects in the system as the most relevant features
  - The less important features are the extracted from the light curve (slope and second quartile)

Precision-Recall detail

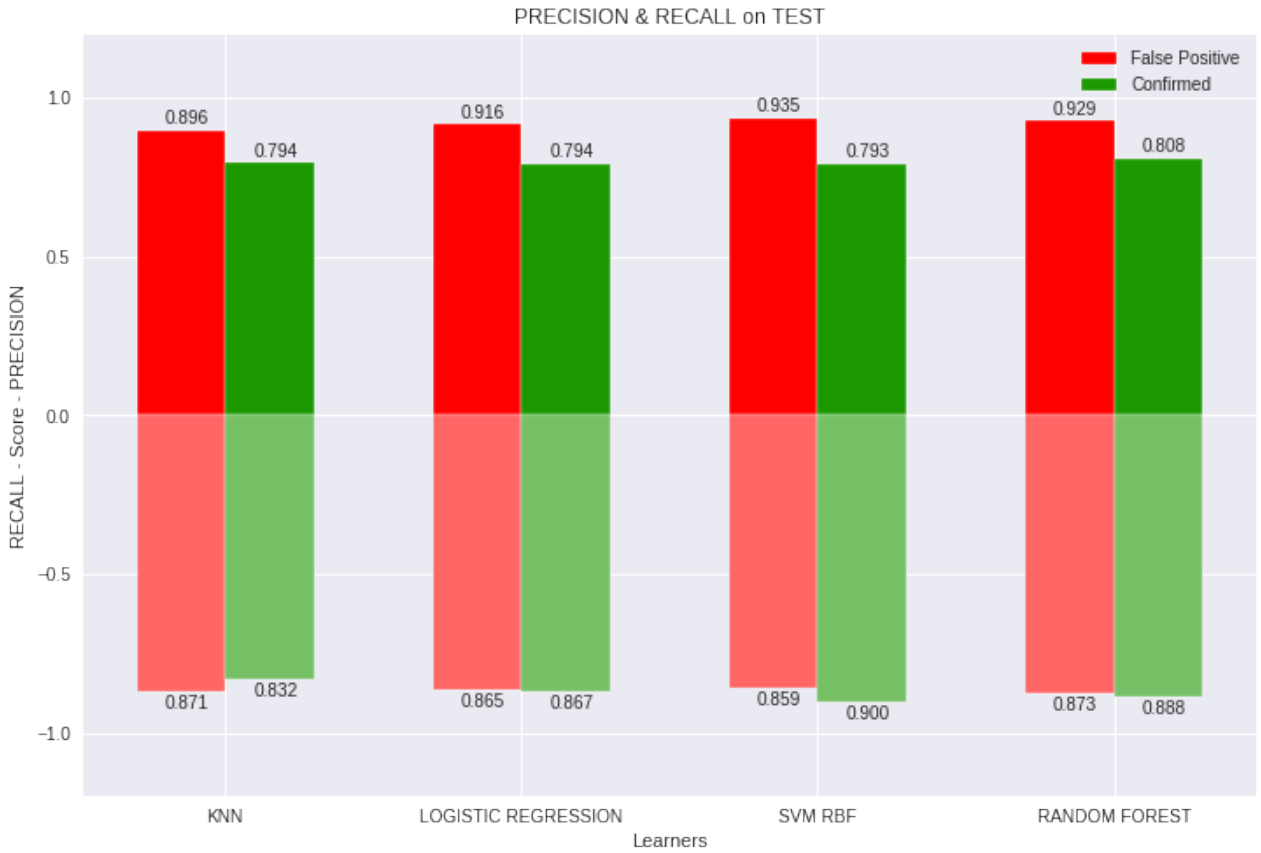


FIGURE – Manual extraction features plus metadata

- The best model that can identify correctly the False Positive Class (FP) based on Precision and Recall is the SVM RBF.
- Classification on the Confirmed class shows lower scores than FP suggesting the difficulty on the exoplanet prediction.
  - Why ? Could be because all of them do not have very similar features on the light curve
- The best model in the task of confirmed exoplanets on the different representation of the data was Random Forest.

Final Results

Selected representation :

- OwnFATS + stellar & planet metadata

Selected model for *confirmed* class :

- Random Forest ($depth = 15$, $T = 15$)

Selected model for *false positive* class :

- SVM RBF ($C = 100$)

We show the classification over the Kepler Object of Interest that are still being studied by the staff of NexScI on September 2017 :

Total CANDIDATE	1791
Subtotal CONFIRMED	975
Subtotal FALSE POSITIVE	434
Unclassified	382

KOI name	**Disposition**	**Confirmed on system**	**Star**
K00601.02	*False Positive*	2/3	Kepler 619
K00750.02	*Unclassified*	1/3	Kepler 662
K01082.01 K01082.02 K01082.04	*Confirmed* *False Positive* *Confirmed*	1/4	Kepler 763
K01236.04	*Confirmed*	2/3	Kepler 279
K01358.01 K01358.02 K01358.03 K01358.04	*Confirmed* *Confirmed* *Confirmed* *Confirmed*	0/4	-
K01750.02	*Confirmed*	1/2	Kepler 948
K02064.01	*Unclassified*	0/1	-
K02420.02	*Confirmed*	1/2	Kepler 1231
K02578.01	*False Positive*	0/1	-
K02828.02	*False Positive*	1/2	Kepler 1259
K03444.03	*Unclassified*	0/4	-
K03451.01	*Unclassified*	0/1	-
K04591.01	*False Positive*	0/1	-
K05353.01	*False Positive*	0/1	-
K06267.01	*Confirmed*	0/1	-

github.com/FMena14/ExoplanetDetection

Publication

- SLIOIA 182772, CLEI publication :

Refining Exoplanet Detection Using Supervised Learning and Feature Engineering

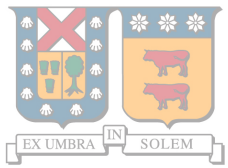
Margarita Bugueño
Departamento de Informática
Univ. Técnica Federico Santa María
Santiago, Chile
margarita.bugueno.13@sansano.usm.cl

Francisco Mena
Departamento de Informática
Univ. Técnica Federico Santa María
Santiago, Chile
francisco.mena.13@sansano.usm.cl

Mauricio Araya
Departamento de Informática
Univ. Técnica Federico Santa María
Valparaíso, Chile
maray@inf.utfsm.cl

***Abstract*—The field of astronomical data analysis has experienced an important paradigm shift in the recent years. The automation of certain analysis procedures is no longer a desirable feature for reducing the human effort, but a must have asset for coping with the extremely large datasets that new instruments**

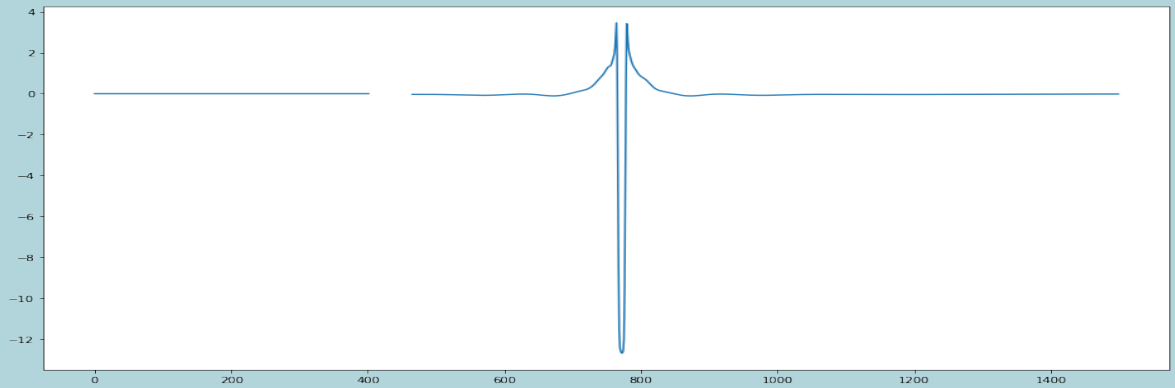
performance of supervised classification methods. To compare them, we used a few standard statistical learning criteria, yet the results motivated an hybrid selection of techniques that reduces the number of unconfirmed light curves.



Working on

Mandel-Agol fit

- The Mandel-Agol model the transit of a stratospheric planet around a stratoshperic star, like an eclipse, assuming a uniform source.
  - It requires the distance from the planet to the parent star as well as the radius of each one of the bodies (ratio) and the period of the orbit. For a better fit could provide eccentricity and limb-darkening coefficients.
- The closeness of the planet to star (eclipse) is modeled as a quadratic polynomial.



Selected Metadata II

Based on previous not so small knowledge		
Transit	**Stellar**	**Planet**
Period	Radius	Radius
Transit Depth	Temperature	Temperature
Eccentricity	Metallicity	KOI count
Planet-Star Radius Ratio	Mass	
Orbit Semi-Major Axis	Log(g)	
Limb Darkening Coefs		
Transit Signal-to-Noise		
+**Upper and lower uncertainty limit**		

Dataset split

We selected a new test set which share certain similarity with the unlabeled objects

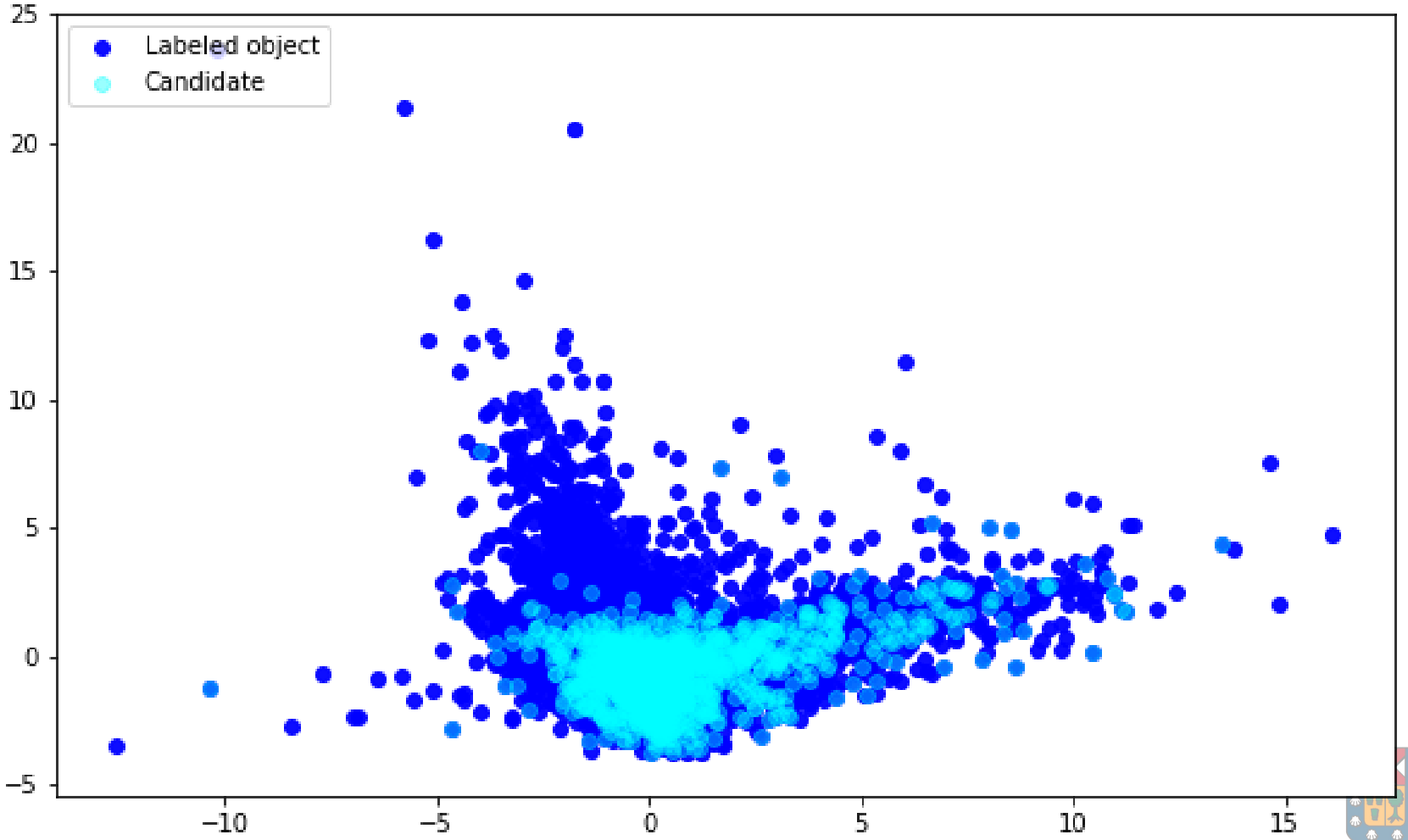


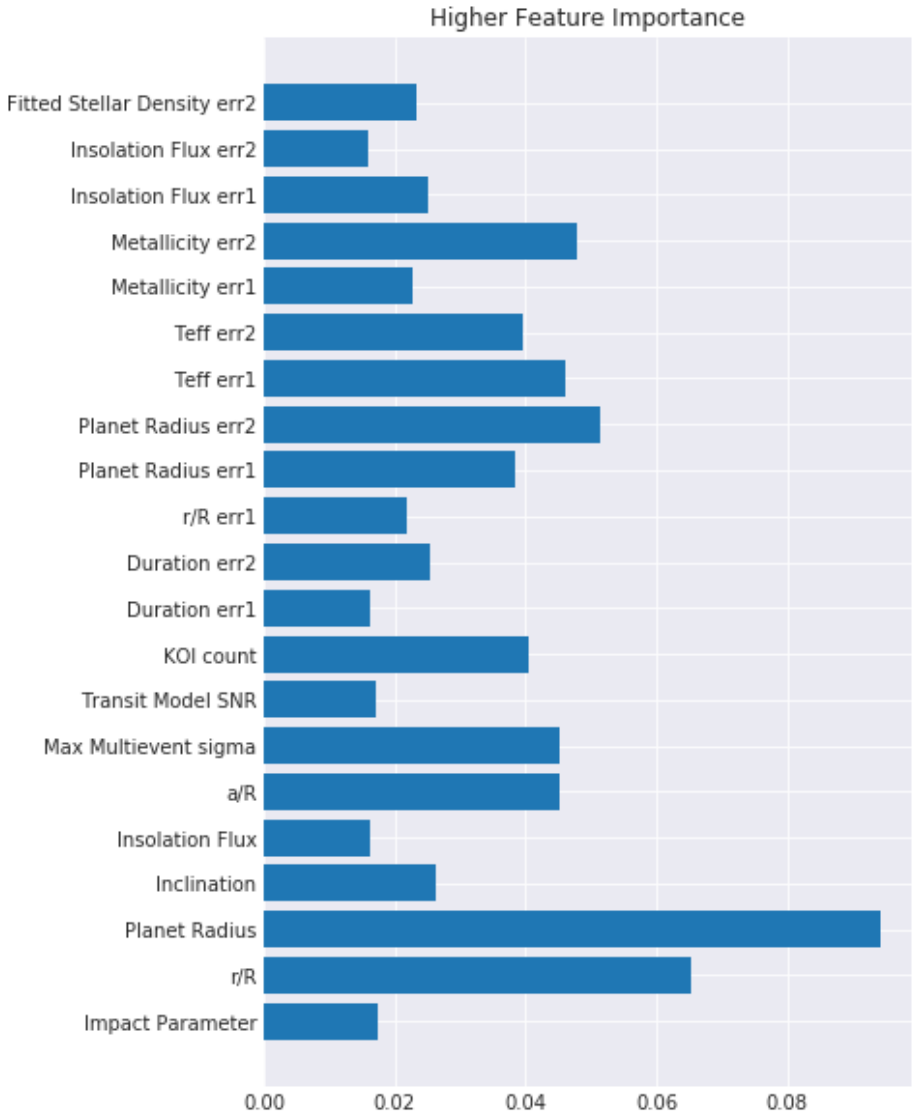
FIGURE – PCA over metadata

New and working on results

Learners performance new			
	k-NN	*SVM RBF*	*Random Forest*
Fourier-PCA (25) on Mandel Agol	0.764	0.440	**0.793**
Fourier-ICA (25) on Mandel Agol	0.773	0.440	**0.804**
All metadata + OwnFATS	0.889	0.916	**0.932**
(FSS) All metadata + OwnFATS			
All metadata + OwnFATS-MA	0.883	0.922	**0.930**
(FSS) All metadata + OwnFATS-MA			

TABLE – F1 score (weighted) on the classification of different models (learners) over the test set on the different representations generated.

Details



Performance analysis

- On the image 20 most important features for Random Forest.
- Different distribution on Planet Radius :

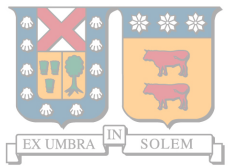
Statistic	Confirmed	False Positive
mean	2.470	9.415
std	2.027	53.021
min	0.440	0.220
25%	1.350	1.090
50%	1.940	1.580
75%	2.745	3.330
max	13.600	1543.740

Varying the dimensionality

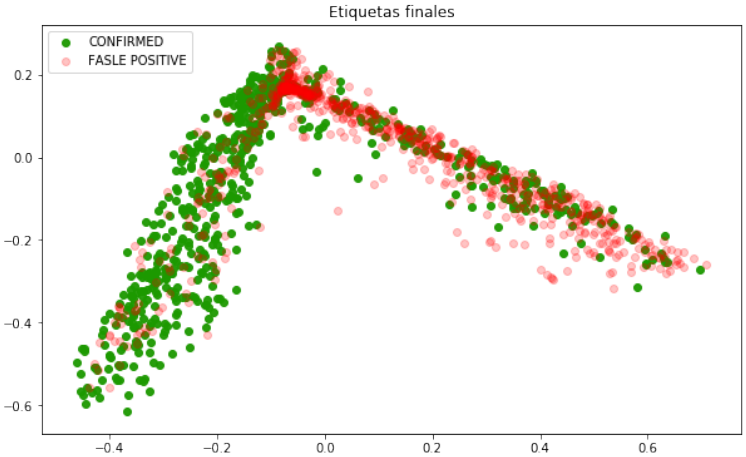
- Change on best dimensionality

	5	**10**	**25**	**50**
ICA	0.751	0.780	**0.804**	0.803
	5	**10**	**25**	**50**
PCA	0.748	0.764	**0.793**	0.793

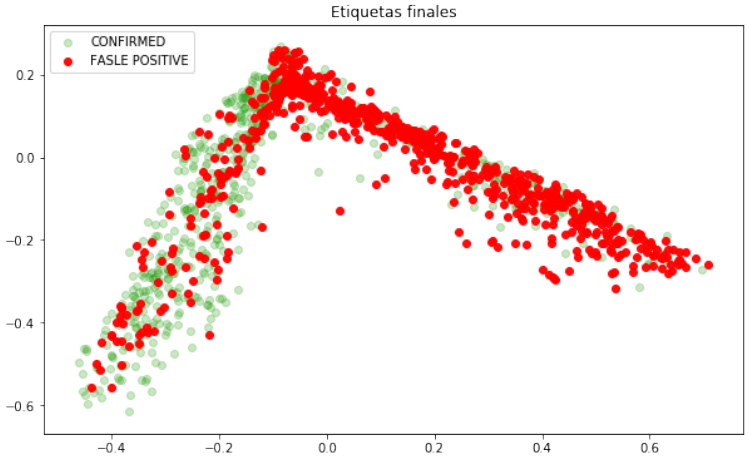
TABLE – F1 score of the best classifier, Random Forest, in function of dimensionality.



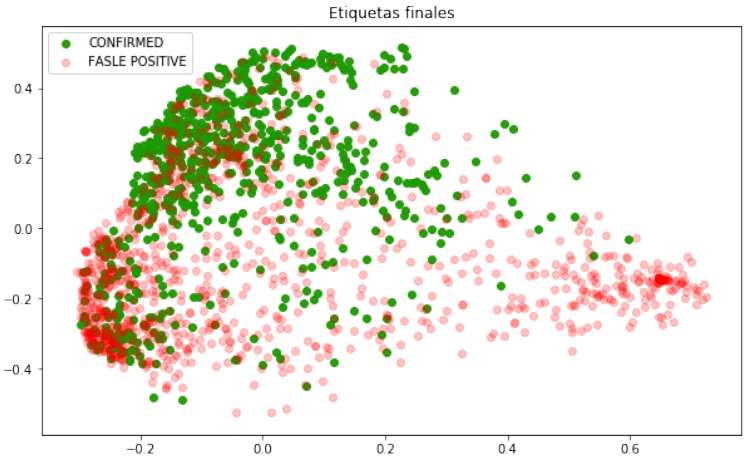
Tagging Visualization



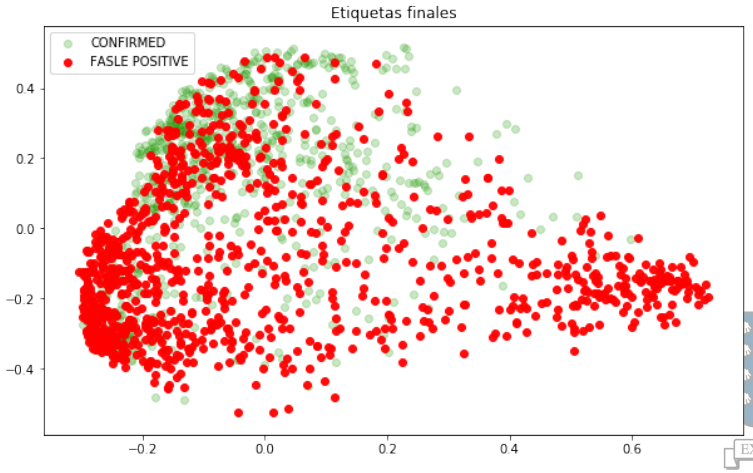
(a) One projection, Confirmed highlighted



(b) One projection, False Positive highlighted



(c) Another projection, Confirmed highlighted

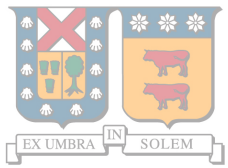


(d) Another projection, False Positive highlighted

Conclusion

- We introduced a new refining method to decide if an object on study (KOI) is really an exoplanet using automatic learning and handling raw data
  - We reproduce the arduous and extensive work that experts perform on detection
- The results show that the automatic techniques used to extract information from the light curve was not good enough compared to the metadata
  - Maybe the not suitable methods for the feature extraction, or well, too simple for the complex problem that we faced
- Also the problem was complex regarding the execution time **on feature extraction**
- The assistance of an expert in this area could be of great value
- **Possible improvements** : Change input representation.

Acknowledgments : Thanks to Chilean Virtual Observatory, ChiVO. Also we thanks to the academic Ricardo Ñanculef.



Refining Exoplanet Detection Using Supervised Learning and Feature Engineering

SLIOIA (CLEI) and CLEI-EJ

M. Bugueño, F. Mena and M. Araya

Departamento de Informática
Universidad Técnica Federico Santa María

April 2019