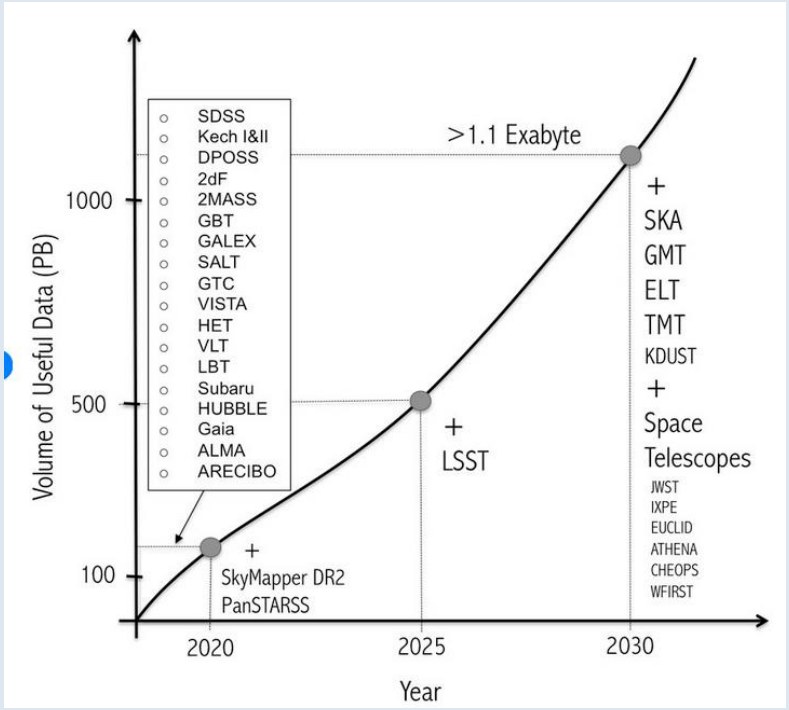
Clasificación fotométrica de cuasares y galaxias usando aprendizaje automático

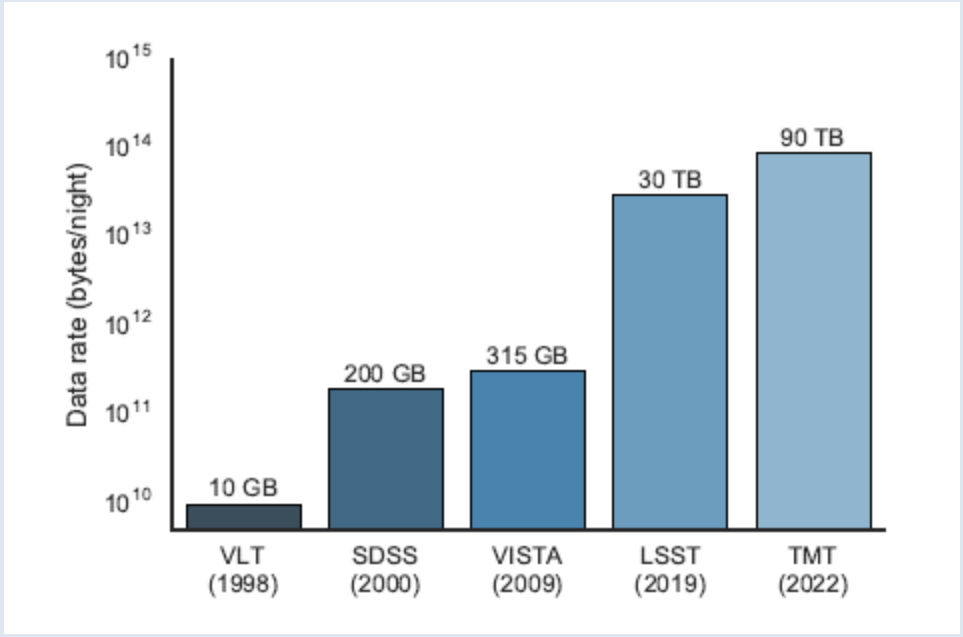
Nestor Sanchez (1) & Benjamín Arroquia-Cuadros (2)

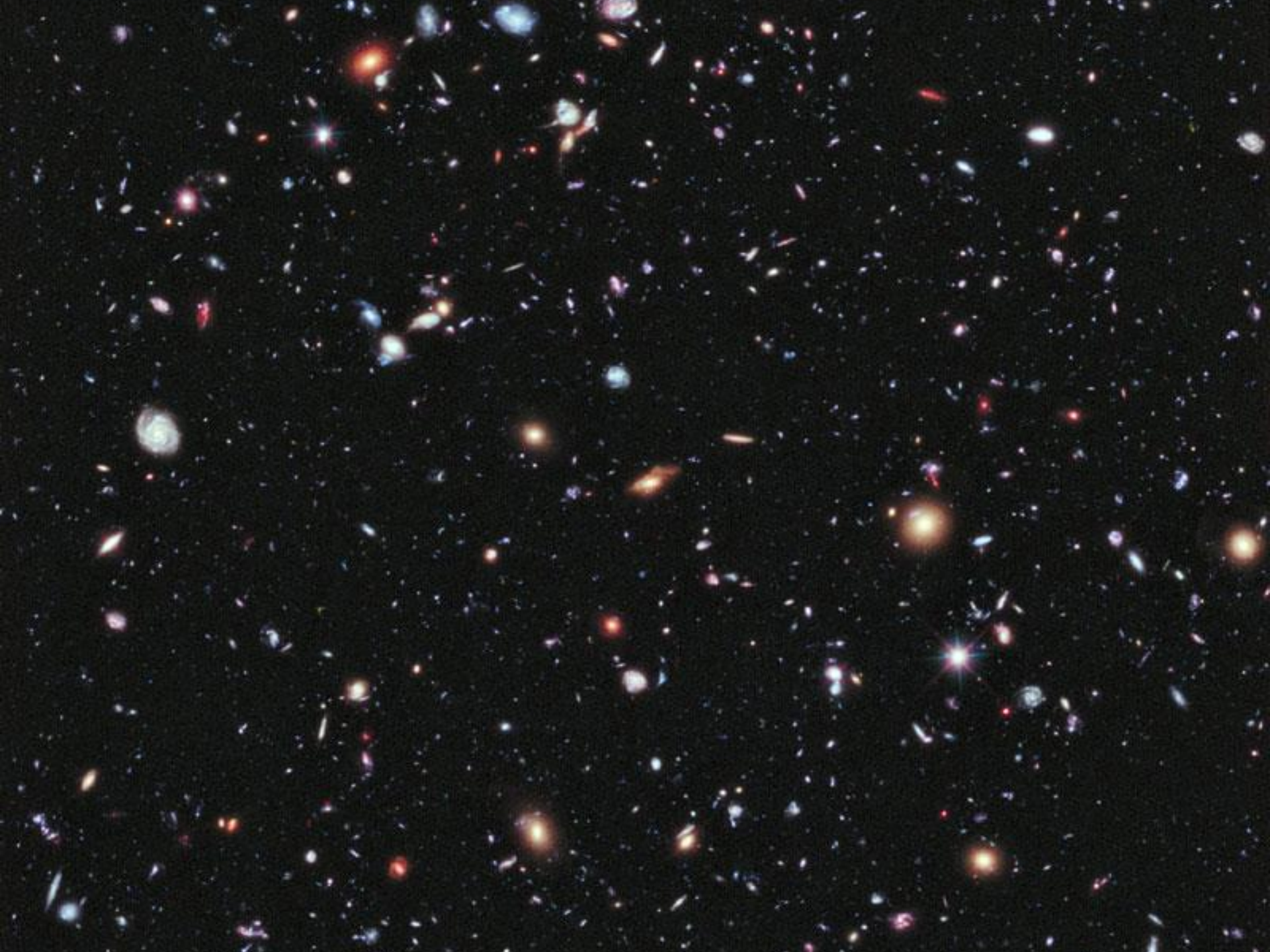
(1) ASGARD: AStronomy Group for Academic Research and Dissemination
(2) GRID: Grupo de investigación en Ciencia de Datos

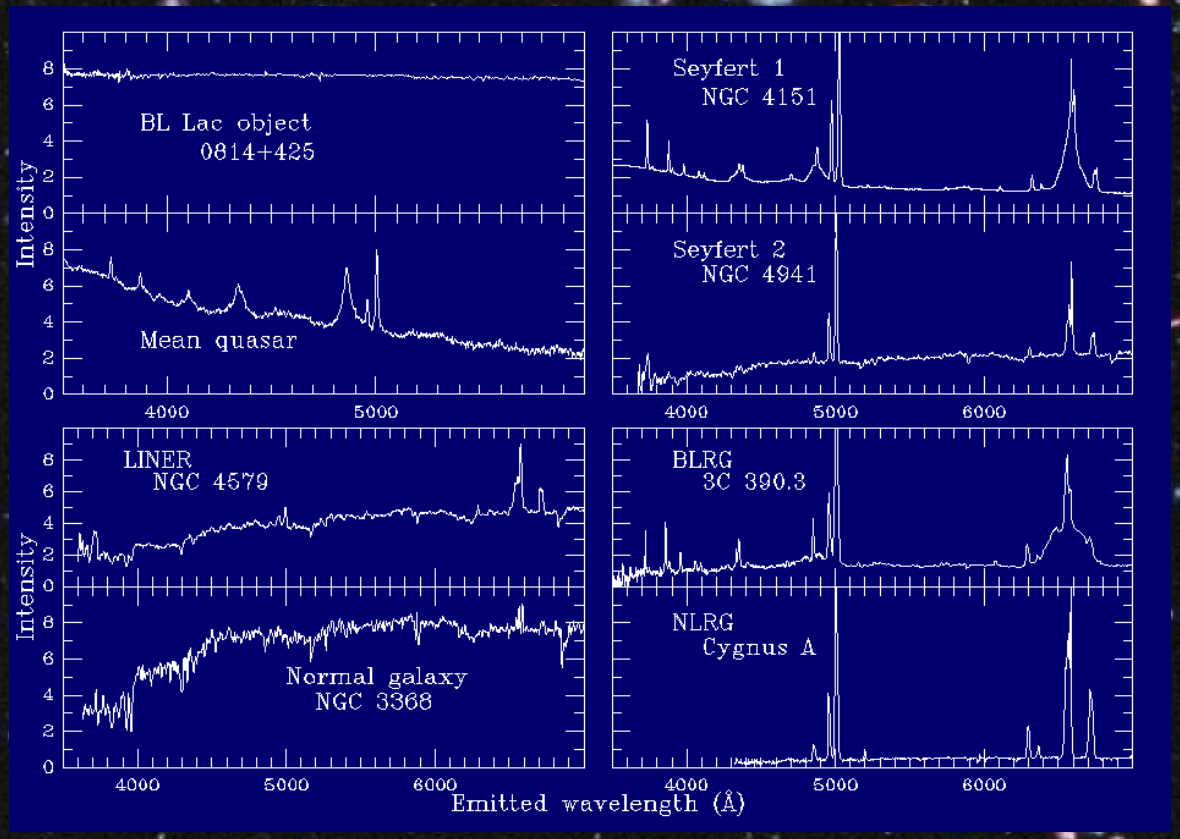
Universidad Internacional de Valencia (VIU)

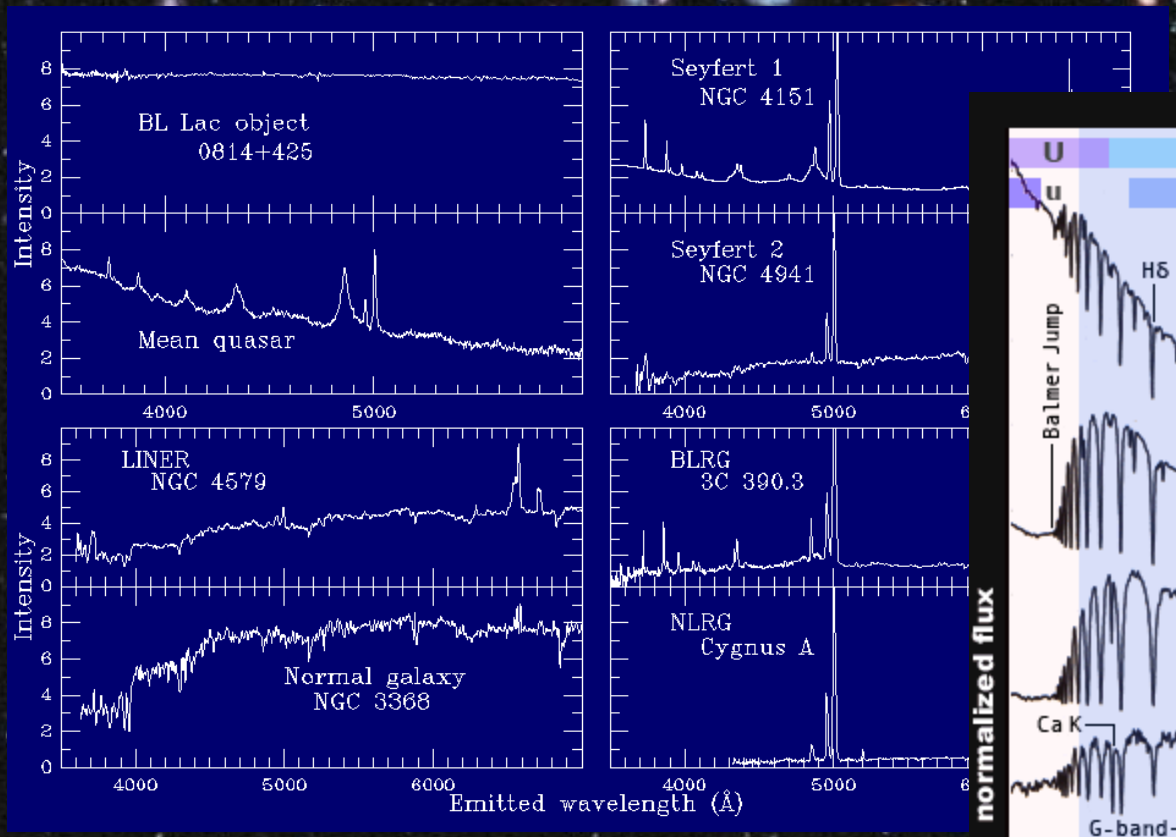
La era de Big Data en Astronomía

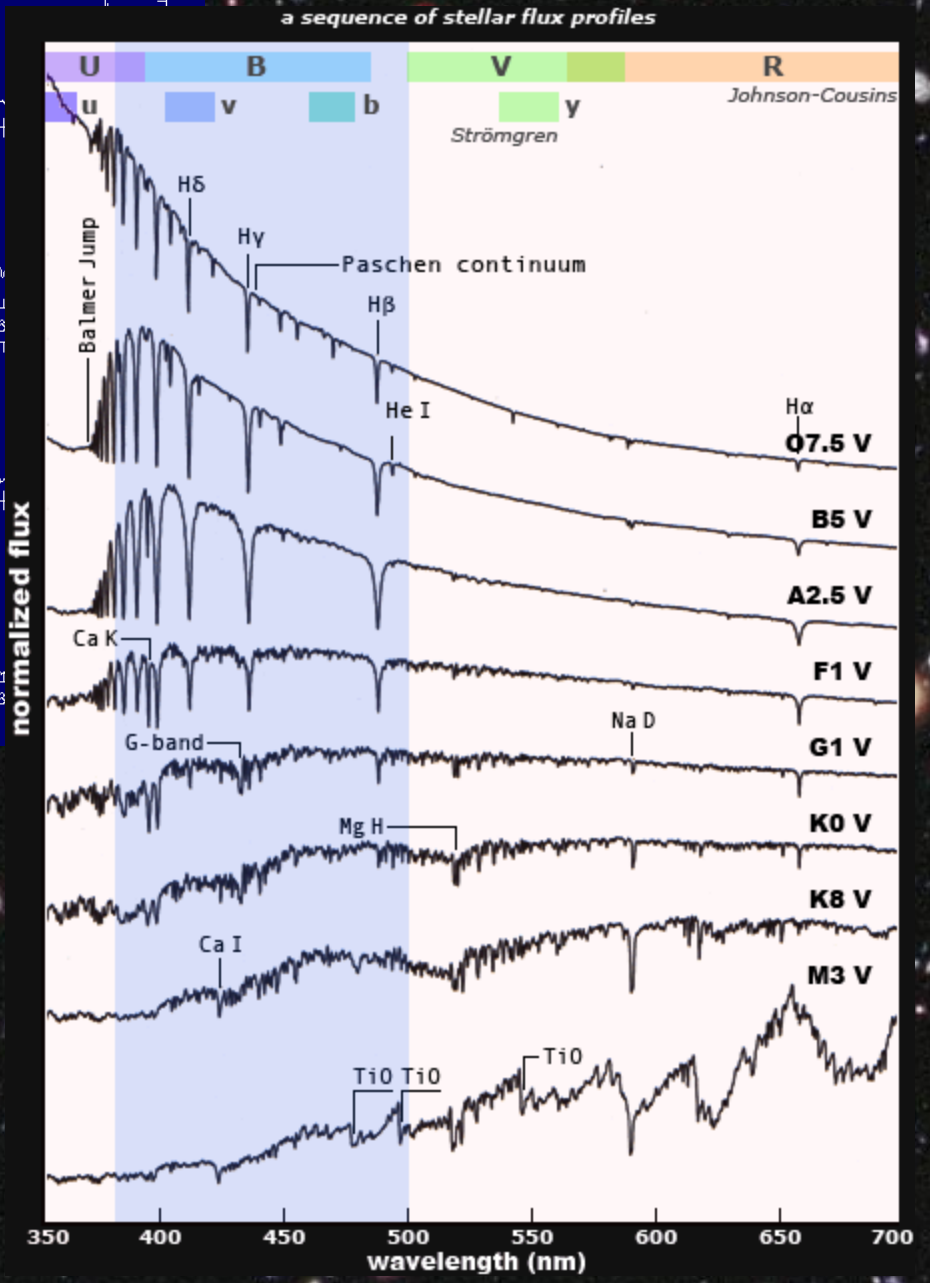


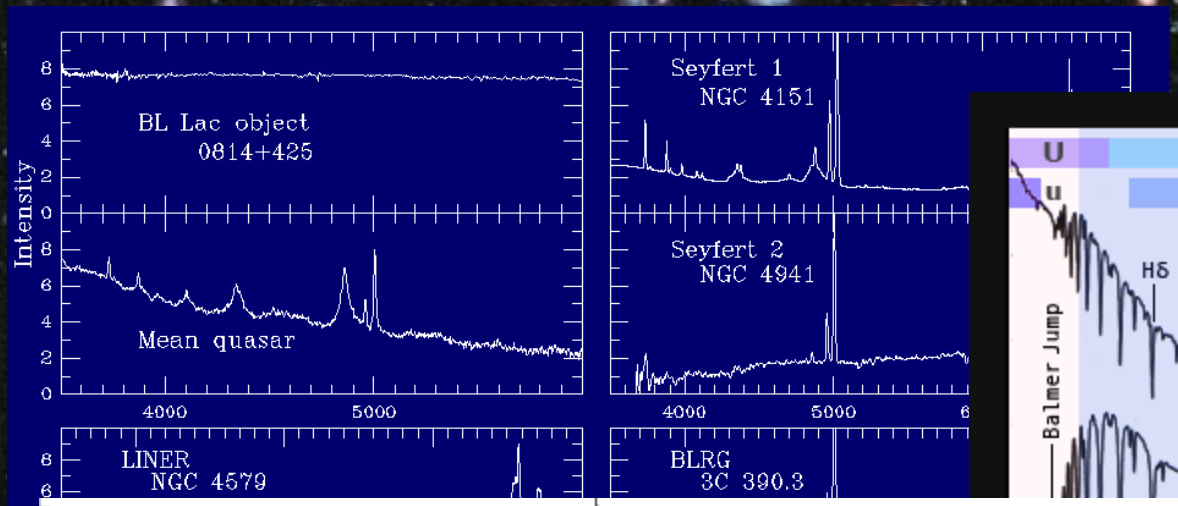


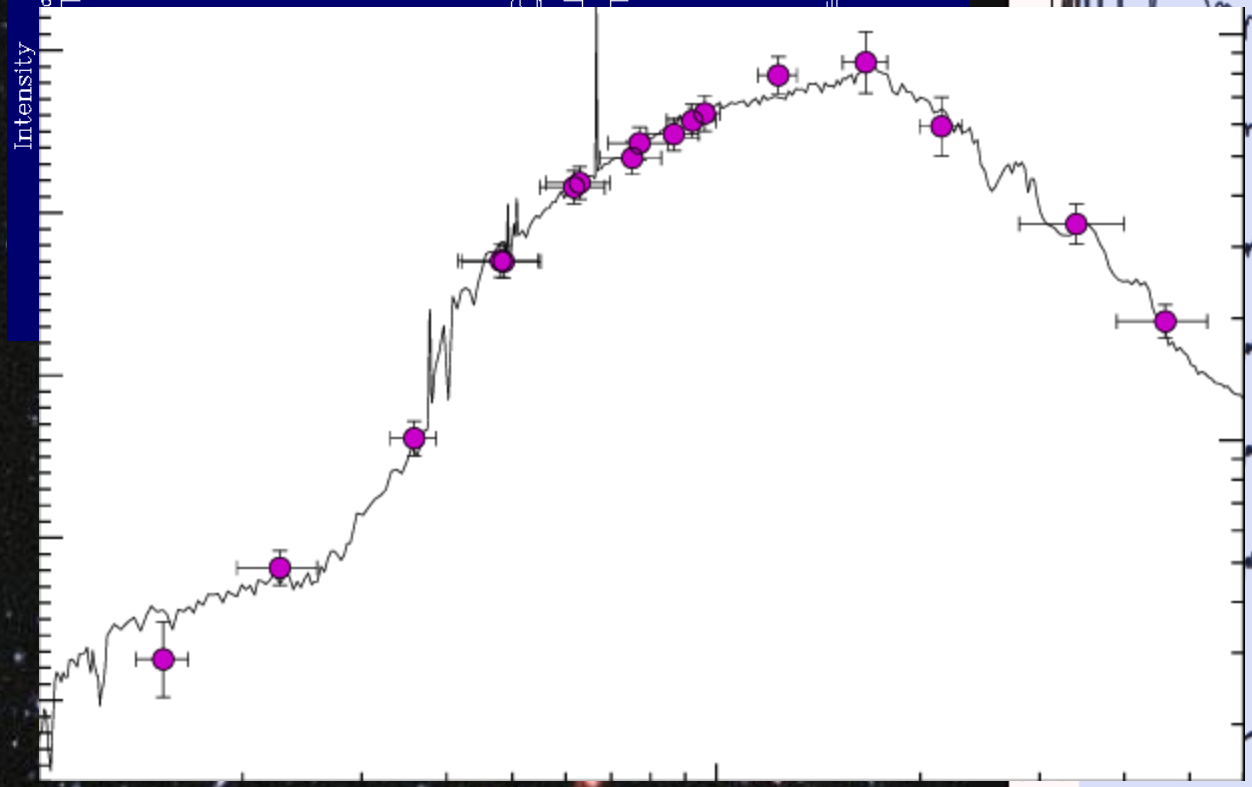


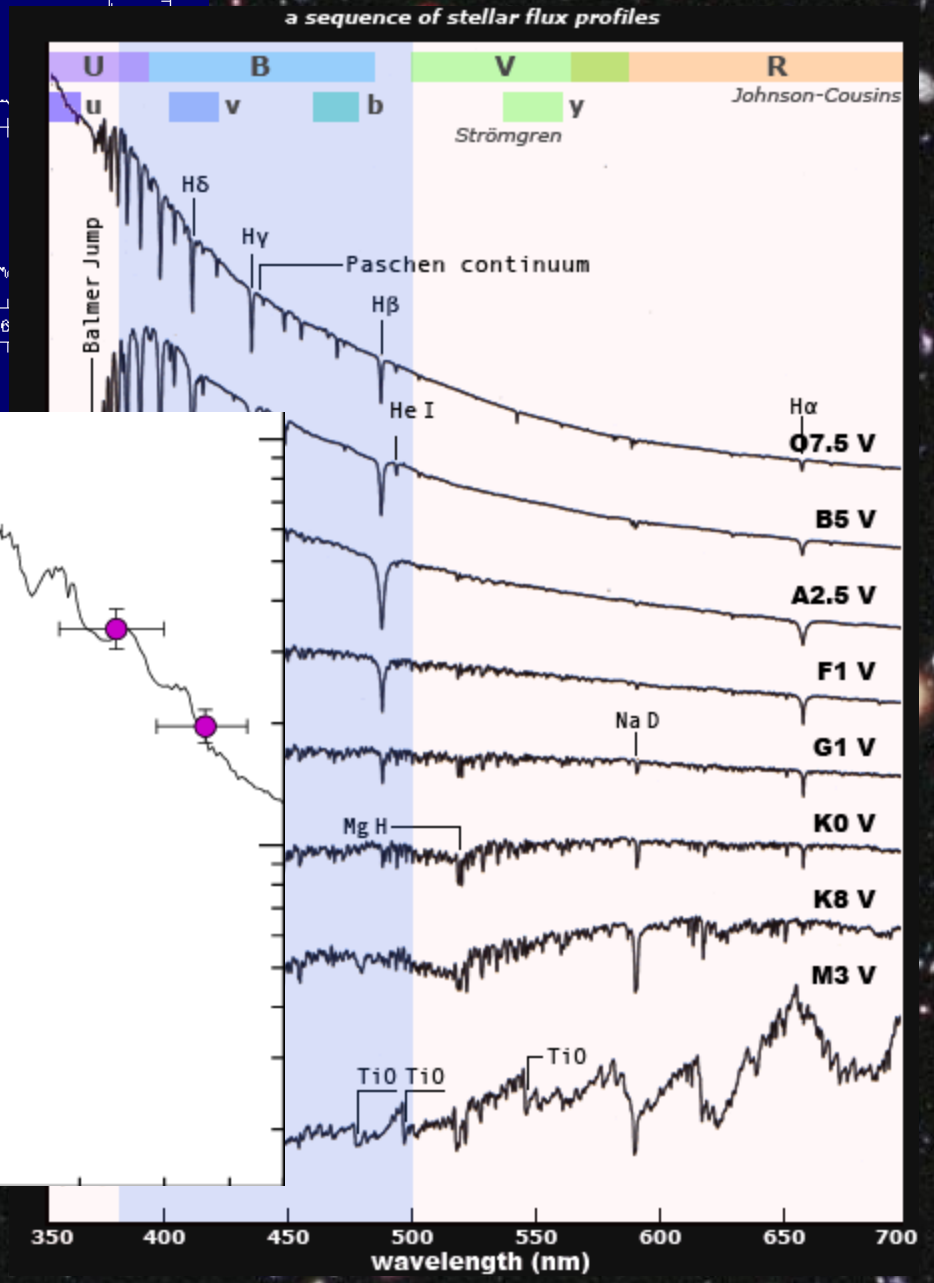

BL Lac object
0814+425
Seyfert 1
NGC 4151
Mean quasar
Seyfert 2
NGC 4941
LINER
NGC 4579
BLRG
3C 390.3
Normal galaxy
NGC 3368
NLRG
Cygnus A
Intensity
Intensity
Emitted wavelength (Å)
0
2
4
6
8
4000
5000
6000


BL Lac object
0814+425
Seyfert 1
NGC 4151
Mean quasar
Seyfert 2
NGC 4941
LINER
NGC 4579
BLRG
3C 390.3
Normal galaxy
NGC 3368
NLRG
Cygnus A
Intensity
Emitted wavelength (Å)


a sequence of stellar flux profiles
U
B
V
R
Johnson-Cousins
u
v
b
y
Strömgren
Hδ
Hγ
Paschen continuum
Hβ
Balmer Jump
He I
Hα
O7.5 V
B5 V
A2.5 V
Ca K
F1 V
Na D
G-band
G1 V
Mg H
K0 V
K8 V
Ca I
M3 V
TiO
TiO TiO
normalized flux
wavelength (nm)


Intensity
BL Lac object
0814+425
Mean quasar
Seyfert 1
NGC 4151
Seyfert 2
NGC 4941
LINER
NGC 4579
BLRG
3C 390.3
4000
5000


Intensity


a sequence of stellar flux profiles
U
B
V
R
Johnson-Cousins
u
v
b
y
Strömgren
Hδ
Hγ
Paschen continuum
Hβ
Balmer Jump
He I
Hα
O7.5 V
B5 V
A2.5 V
F1 V
Na D
G1 V
Mg H
K0 V
K8 V
M3 V
TiO TiO
TiO
350
400
450
500
550
600
650
700
wavelength (nm)

Introducción

Meta/deseo:

- Identificar/caracterizar millones de objetos usando métodos automáticos.
- Usar información fotométrica (magnitudes y/o colores) de, por ejemplo:
  - Sloan Digital Sky Survey (SDSS; York et al. 2000)
  - Wide-Field Infrared Survey Explorer (WISE; Wright et al. 2010)
  - Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006)

Introducción

Meta/deseo:

- Identificar/caracterizar millones de objetos usando métodos automáticos.
- Usar información fotométrica (magnitudes y/o colores) de, por ejemplo:
  - Sloan Digital Sky Survey (SDSS; York et al. 2000)
  - Wide-Field Infrared Survey Explorer (WISE; Wright et al. 2010)
  - Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006)

Trabajos previos han ensayado diferentes algoritmos para clasisifcar fuentes puntuales en estrellas, galaxias y cuasares (QSOs):

- decision trees o random forests (RF)
- support vector machines (SVM)
- artificial neural networks (ANN)
- k-nearest neighbours (KNN)

REFs: Suchkov et al. 2005; Ball et al. 2006; Yèche et al. 2010; Vasconcellos et al. 2011; Peng et al. 2012; Carrasco et al. 2015; Kovács & Szapudi 2015; Krakowski et al. 2016; Bai et al. 2019; Makhija et al. 2019; Schindler et al. 2019; Nakoneczny et al. 2019, 2021; Clarke et al. 2020; Khramtsov et al. 2021; Li et al. 2021; Nakoneczny et al. 2021; Guarneri et al. 2021; Nakazono et al. 2021; Wenzl et al. 2021; Cunha & Humphrey 2022; Wang et al. 2022; ...

Introducción

- Análisis y comparación de diferentes métodos, ejemplos:
  - Bai et al. (2019): RF tiene mejores resultados (*accuracy* o precisión) que KNN y que SVM.
  - Wang et al. (2022): SVM es "mejor" que RF.

- OJO: Las diferencias parecen estar mas relacionadas con las muestras (especialmente el tamaño y calidad de las muestras de aprendizaje) y con las características (features) usadas: magnitudes, colores, o combinaciones de éstas y otras.

Introducción

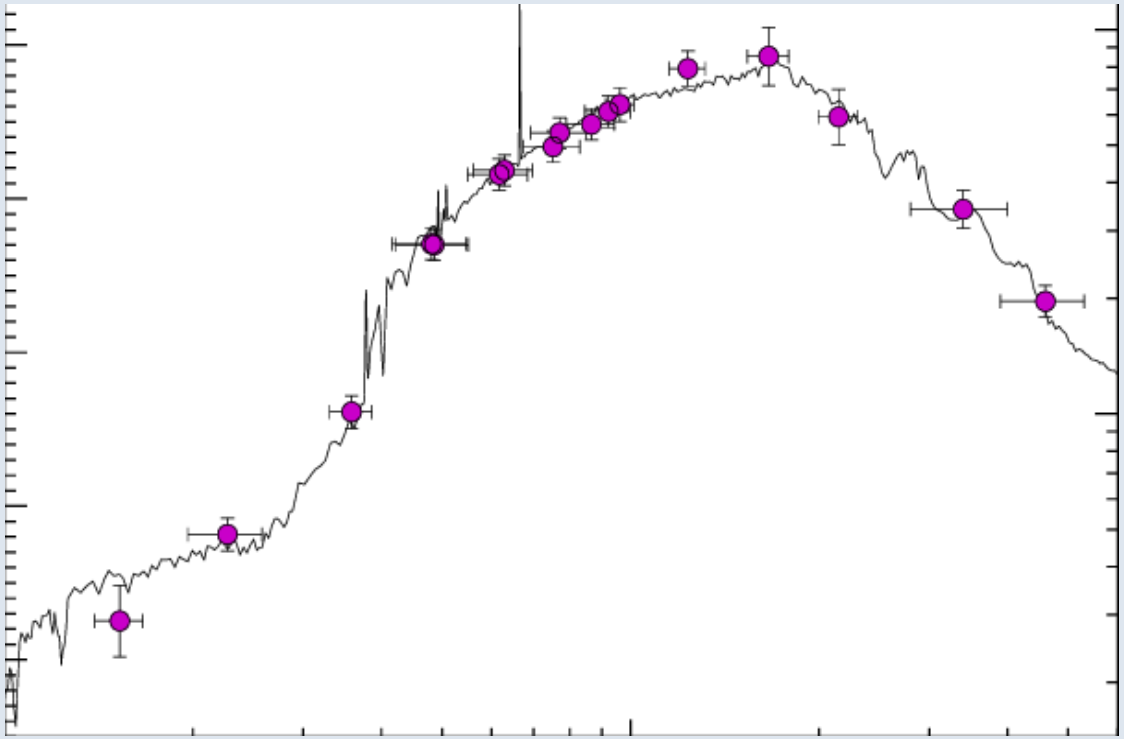
- Análisis y comparación de diferentes métodos, ejemplos:
  - Bai et al. (2019): RF tiene mejores resultados (*accuracy* o precisión) que KNN y que SVM.
  - Wang et al. (2022): SVM es "mejor" que RF.

- OJO: Las diferencias parecen estar mas relacionadas con las muestras (especialmente el tamaño y calidad de las muestras de aprendizaje) y con las características (features) usadas: magnitudes, colores, o combinaciones de éstas y otras.

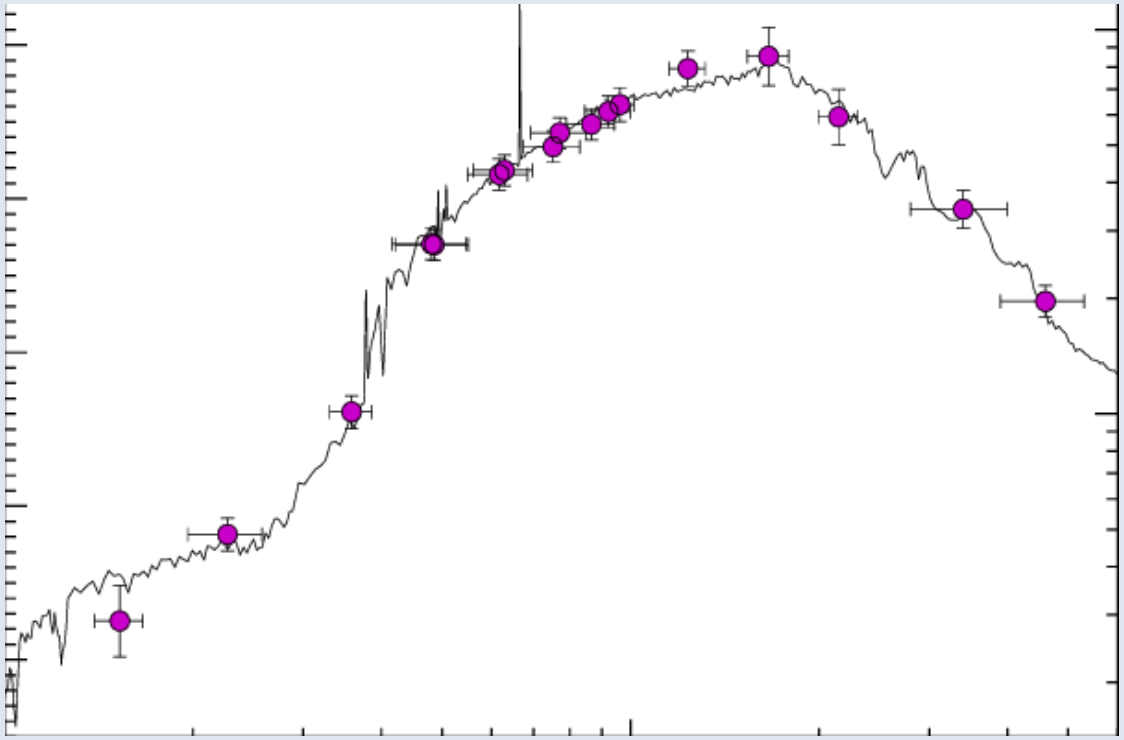
- **La búsqueda entre astrónomos de la estrategia "óptima" produce enfoques excesivamente desiguales**: usar solo magnitudes de banda ancha, usar 83 features (que incluyen magnitudes, colores, ratios de magnitudes), o combinar 32 modelos de machine-learning basados en diferentes algoritmos.

REFs: Nakazono et al. 2021; Nakoneczny et al. 2021; Khramtsov et al. 2021.

Magnitudes versus colores



Magnitudes versus colores

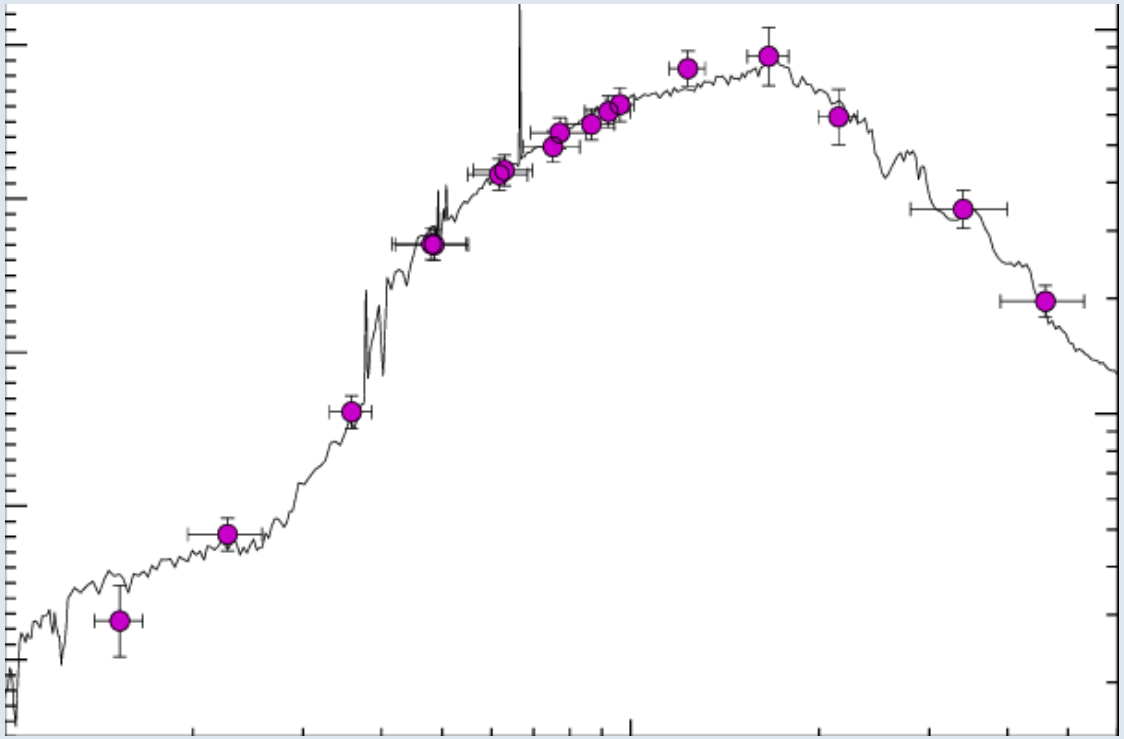


Obetivo

Evaluar **sistemáticamente** el desempeño de **RFs** para clasificar QSOs usando:

- magnitudes
- colores
- filtros anchos
- filtros estrechos

Magnitudes versus colores



Obetivo

Evaluar **sistemáticamente** el desempeño de **RFs** para clasificar QSOs usando:

- magnitudes
- colores
- filtros anchos
- filtros estrechos

A&A 673, A48 (2023)
https://doi.org/10.1051/0004-6361/202245531

Astronomy & Astrophysics

Photometric classification of quasars from ALHAMBRA survey using random forest

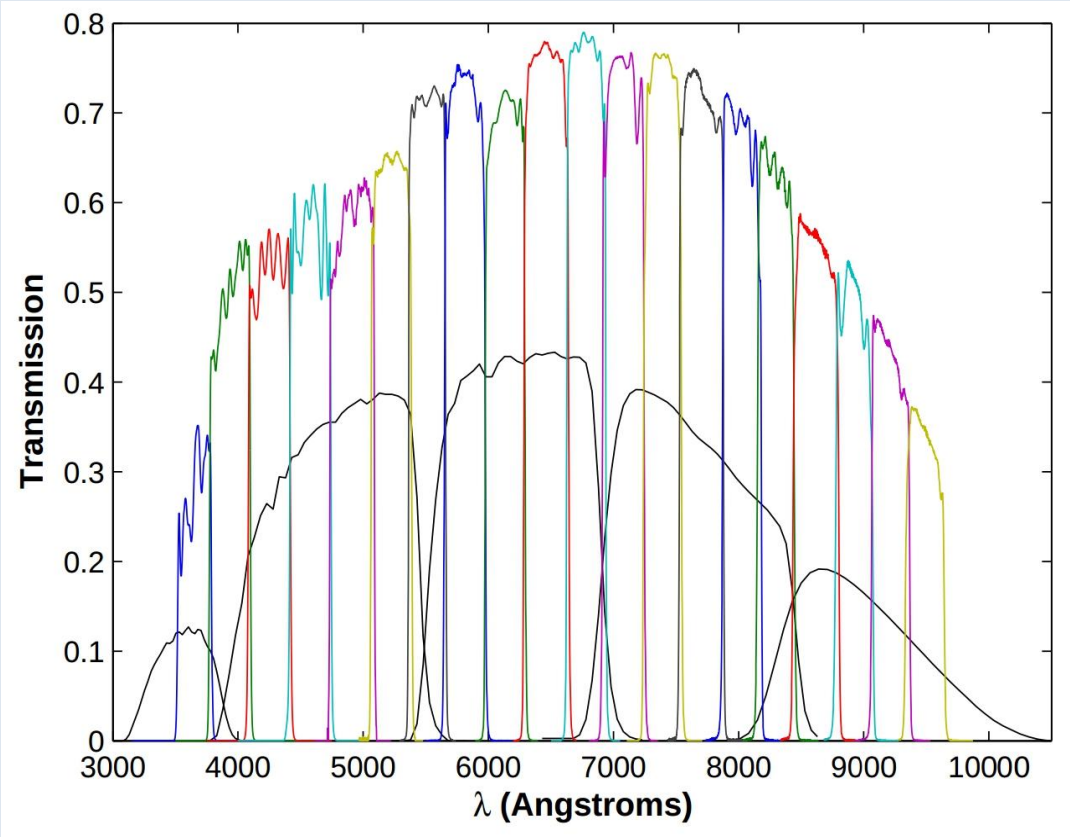
Benjamín Arroquia-Cuadros, Néstor Sánchez, Vicent Gómez, Pere Blay, Vicent Martinez-Badenes and Lorena Nieves-Seoane

Universidad Internacional de Valencia (VIU), C/Pintor Sorolla 21, 46002 Valencia, Spain
e-mail: nestor.sanchezd@campusviu.es

Received 22 November 2022 / Accepted 15 March 2023

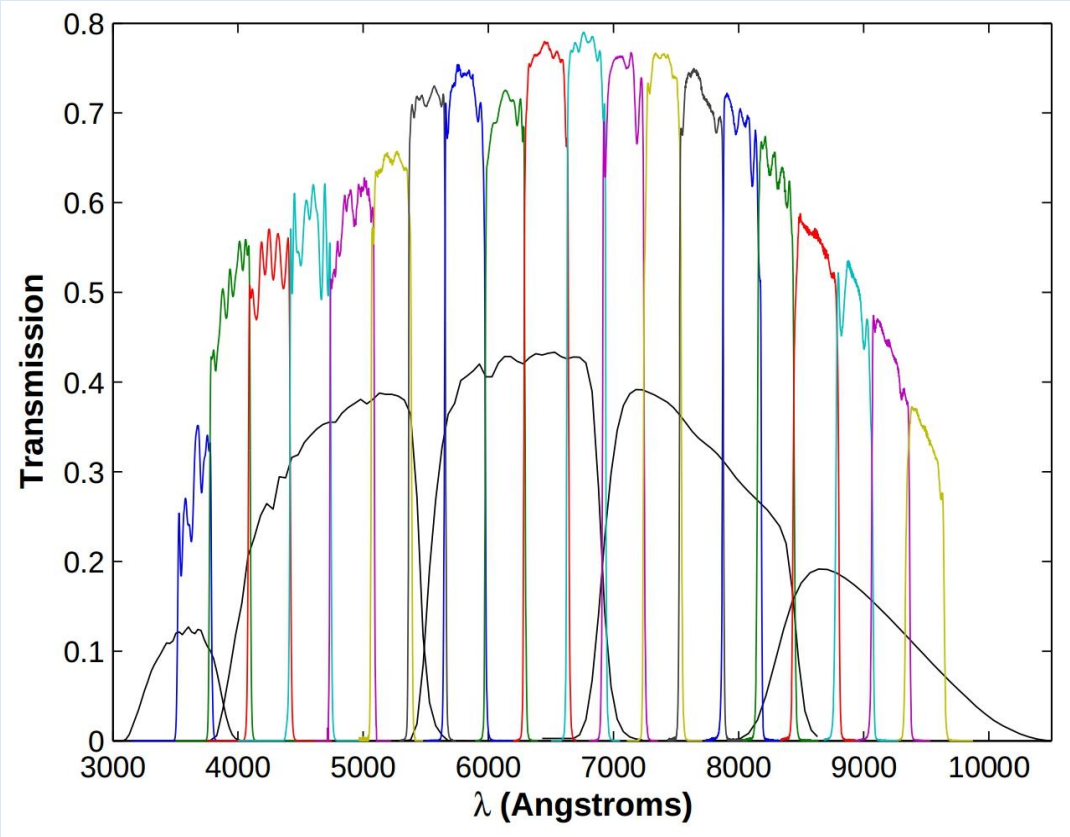
Los datos

Figure: The ALHAMBRA filter set with superimposed Sloan Digital Sky Survey filter system.



Los datos

Figure: The ALHAMBRA filter set with superimposed Sloan Digital Sky Survey filter system.



Total de fuentes: 441303

Survey ALHAMBRA combinado con Sloan Survey + Milliquas catalogue:

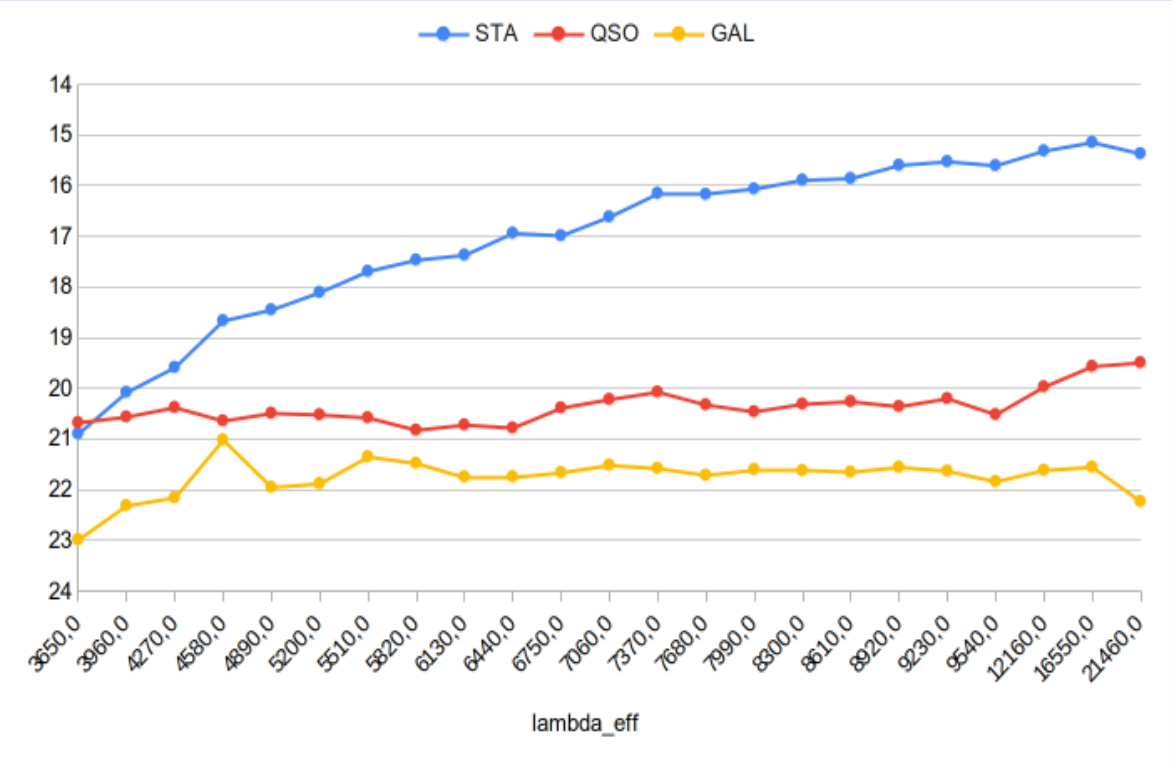
---> 2621 fuentes con clasificación espectroscópica

STA: 516

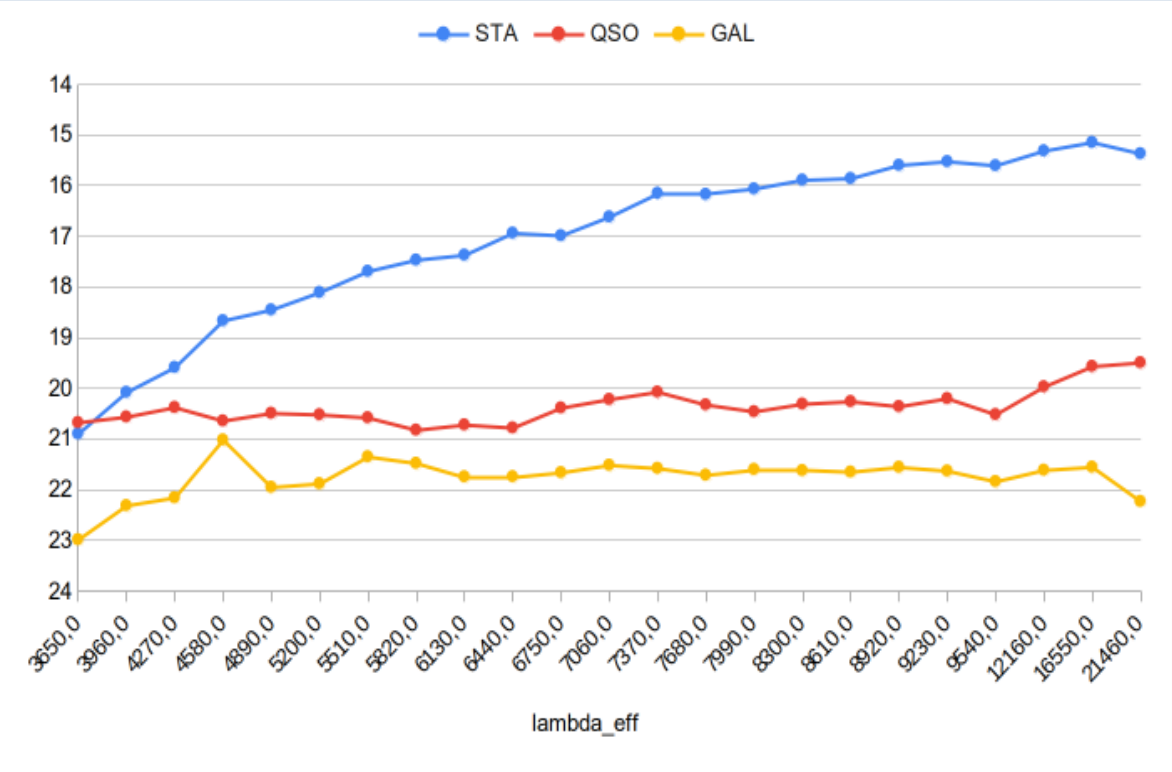
GAL: 1347

QSO: 758

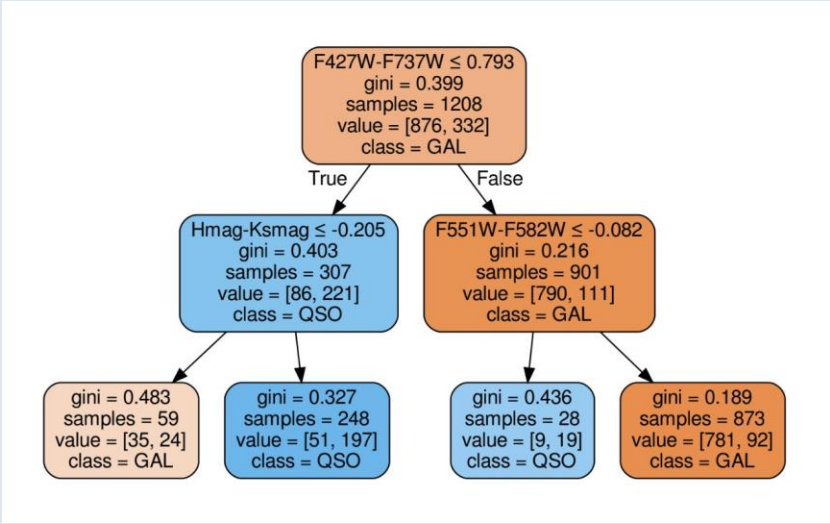
Ejemplo: 3 primeros objetos



Ejemplo: 3 primeros objetos



n_estimators = número de árboles
max_features = subset aleatorio features
max_depth = profundidad máxima de árbol



Resultados: efecto de los parámetros libres

Caso de referencia:

n_estimators = 100
max_features = 5
max_depth = 10

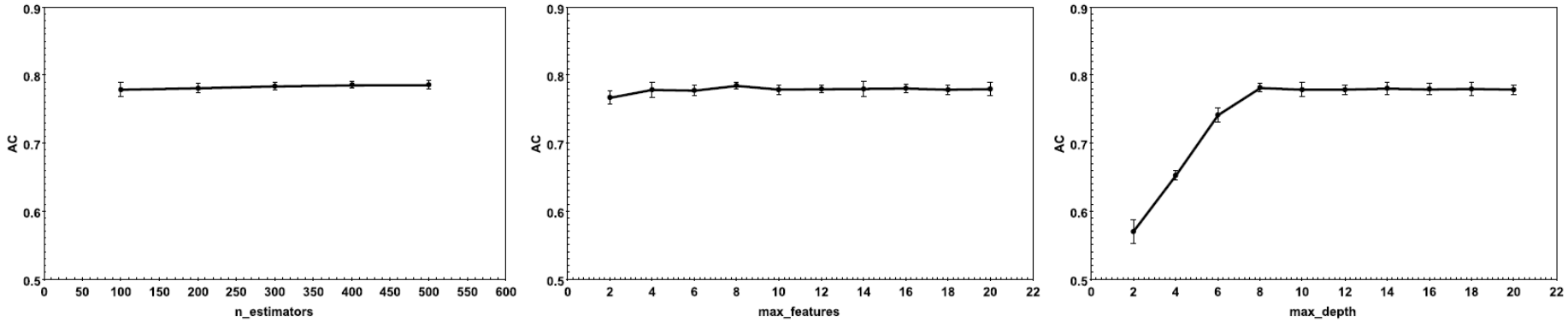


Fig. 2. Global accuracy (*AC*) as a function of free parameters *n_estimators* (left panel), *max_features* (central panel) and *max_depth* (right panel) when the RF classifier is applied with magnitudes from ALHAMBRA as features. In each case, the rest of the free parameters are fixed to their reference values (see text). Error bars correspond to three times the estimated standard deviations.

Resultados: distintos features y datasets

Table 1. Accuracy (AC) and its standard deviation (in brackets) obtained for different combinations of features, databases and free parameter sets.

Features	Database	Parameter set	AC (σ)
Magnitudes	ALHAMBRA	Reference	0.779 (0.007)
Magnitudes	ALHAMBRA	Optimal	0.779 (0.005)
Magnitudes	SDSS	Reference	0.769 (0.006)
Magnitudes	SDSS	Optimal	0.772 (0.007)
Colours	ALHAMBRA	Reference	0.878 (0.004)
Colours	ALHAMBRA	Optimal	0.900 (0.002)
Colours	SDSS	Reference	0.792 (0.006)
Colours	SDSS	Optimal	0.795 (0.004)

Resultados: distintos features y datasets

Table 2. Precision, recall and F1 metrics with their standard deviations (in brackets) obtained for different combinations of features, databases and free parameter sets, and for each of the classes in the sample: quasars (QSO), galaxies (GAL) and stars (STA).

Features	Database	Parameter set	Class	Precision (σ)	Recall (σ)	F1 (σ)
Magnitudes	ALHAMBRA	Reference	QSO	0.743 (0.009)	0.712 (0.020)	0.727 (0.012)
			GAL	0.727 (0.016)	0.743 (0.008)	0.735 (0.009)
			STA	0.864 (0.005)	0.884 (0.006)	0.874 (0.004)
Magnitudes	ALHAMBRA	Optimal	QSO	0.752 (0.008)	0.706 (0.012)	0.728 (0.009)
			GAL	0.730 (0.007)	0.748 (0.012)	0.738 (0.007)
			STA	0.850 (0.009)	0.885 (0.007)	0.867 (0.004)
Magnitudes	SDSS	Reference	QSO	0.749 (0.008)	0.701 (0.013)	0.724 (0.010)
			GAL	0.695 (0.009)	0.760 (0.011)	0.726 (0.008)
			STA	0.867 (0.006)	0.848 (0.007)	0.858 (0.006)
Magnitudes	SDSS	Optimal	QSO	0.757 (0.014)	0.702 (0.015)	0.728 (0.011)
			GAL	0.699 (0.011)	0.758 (0.009)	0.727 (0.009)
			STA	0.864 (0.007)	0.859 (0.011)	0.862 (0.007)
Colours	ALHAMBRA	Reference	QSO	0.860 (0.007)	0.839 (0.009)	0.849 (0.005)
			GAL	0.847 (0.007)	0.835 (0.009)	0.841 (0.005)
			STA	0.923 (0.005)	0.959 (0.004)	0.941 (0.003)
Colours	ALHAMBRA	Optimal	QSO	0.885 (0.006)	0.881 (0.004)	0.883 (0.002)
			GAL	0.885 (0.003)	0.853 (0.007)	0.869 (0.004)
			STA	0.928 (0.004)	0.964 (0.005)	0.945 (0.003)
Colours	SDSS	Reference	QSO	0.750 (0.012)	0.683 (0.012)	0.715 (0.009)
			GAL	0.719 (0.008)	0.775 (0.014)	0.746 (0.008)
			STA	0.904 (0.003)	0.922 (0.006)	0.913 (0.004)
Colours	SDSS	Optimal	QSO	0.763 (0.008)	0.675 (0.012)	0.716 (0.008)
			GAL	0.714 (0.008)	0.773 (0.006)	0.742 (0.004)
			STA	0.904 (0.003)	0.942 (0.005)	0.922 (0.003)

Resultados: importancia de los colores

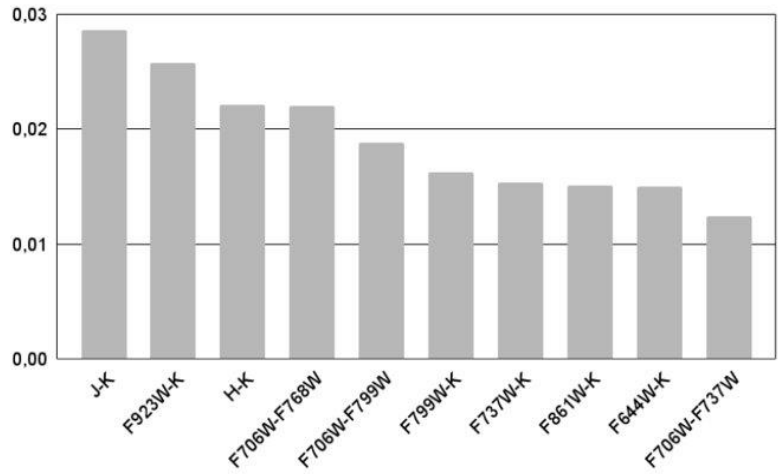


Fig. 4. Relative importances of the features (colours) for the same execution shown in Figure 3. For clarity, only the ten most important features are presented.

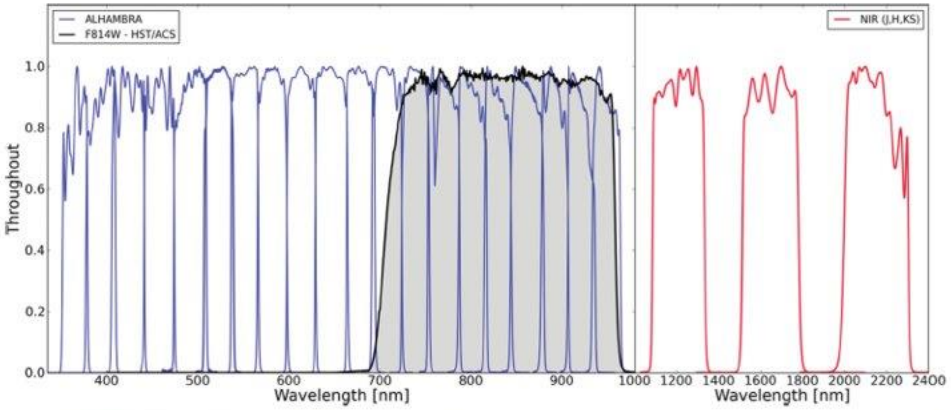


Figure 2. The ALHAMBRA survey filter set. On the left-hand side, solid blue lines represent the optical filter system composed of 20 contiguous, equal-width, non-overlapping, medium-band (~300 Å) filters. The solid black line corresponds to the synthetic *F814W* filter used to define a constant observational window across fields. On the right-hand side, solid red lines represent the standard *JHKs* near-infrared broad bands. All transmission curves are normalized to the maximum value.

Resultados: importancia de los colores

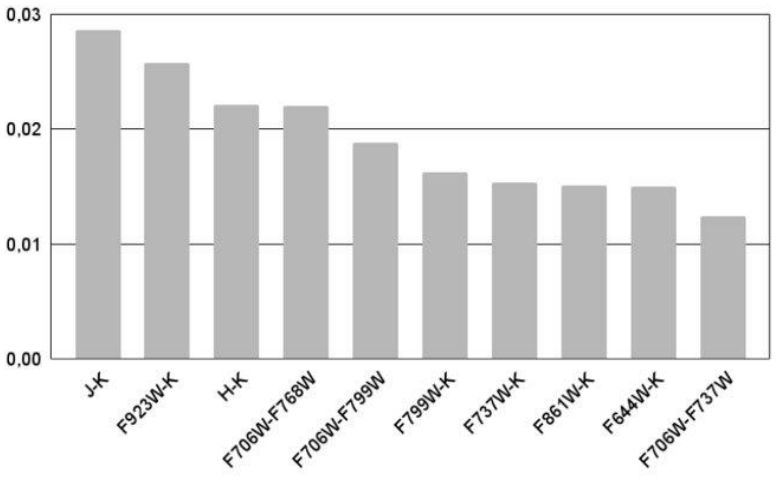


Fig. 4. Relative importances of the features (colours) for the same execution shown in Figure 3. For clarity, only the ten most important features are presented.

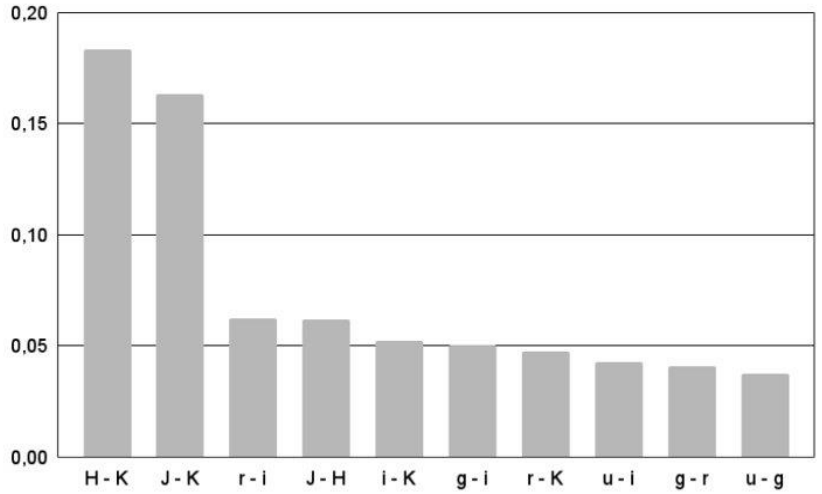


Fig. 6. Relative importances of the SDSS colours for the reference case. Only the ten most relevant colours are shown.

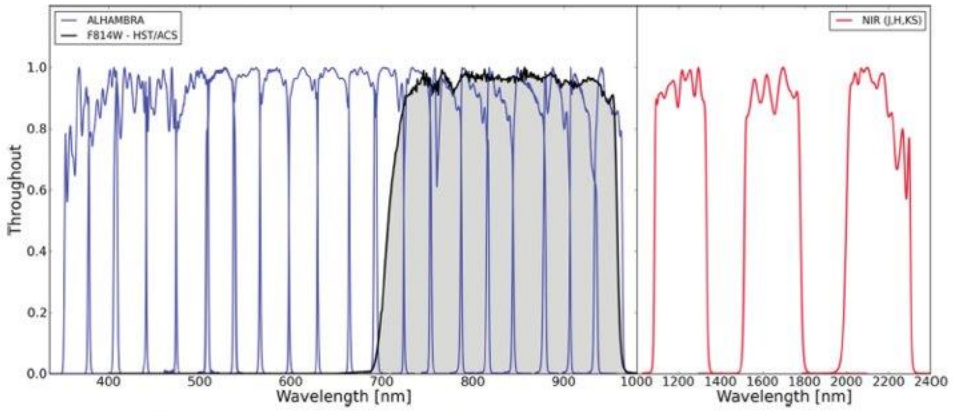


Figure 2. The ALHAMBRA survey filter set. On the left-hand side, solid blue lines represent the optical filter system composed of 20 contiguous, equal-width, non-overlapping, medium-band (~300 Å) filters. The solid black line corresponds to the synthetic *F814W* filter used to define a constant observational window across fields. On the right-hand side, solid red lines represent the standard *JHKs* near-infrared broad bands. All transmission curves are normalized to the maximum value.

Resultados: usando colores de Alhambra

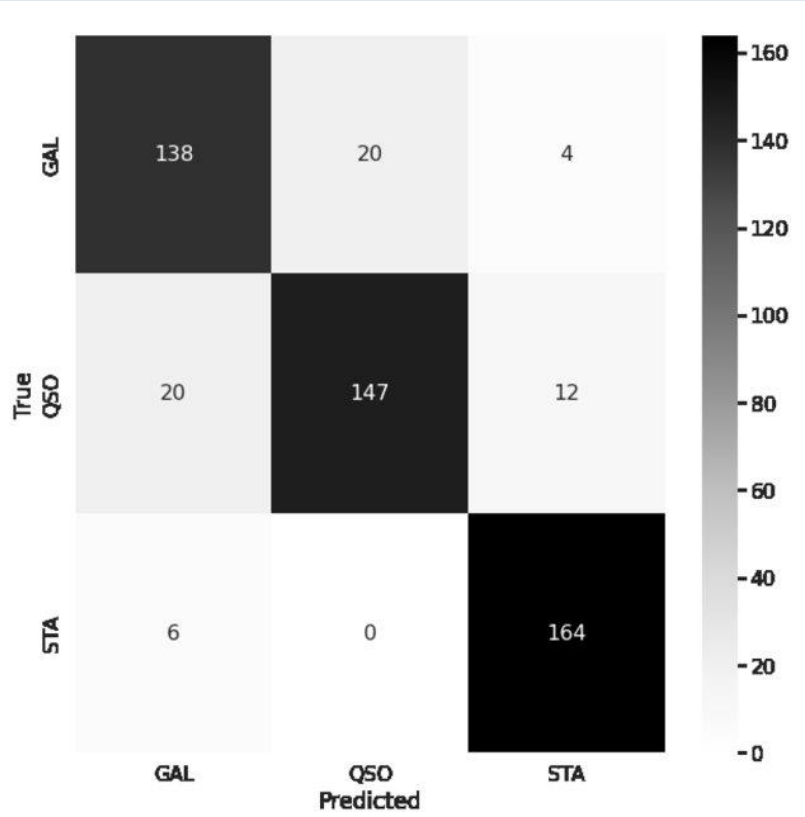


Fig. 3. Resulting confusion matrix of predicted and true classes for the first random execution of the RF using ALHAMBRA colours as features and the standard set of free parameters.

Resultados: usando colores de Alhambra

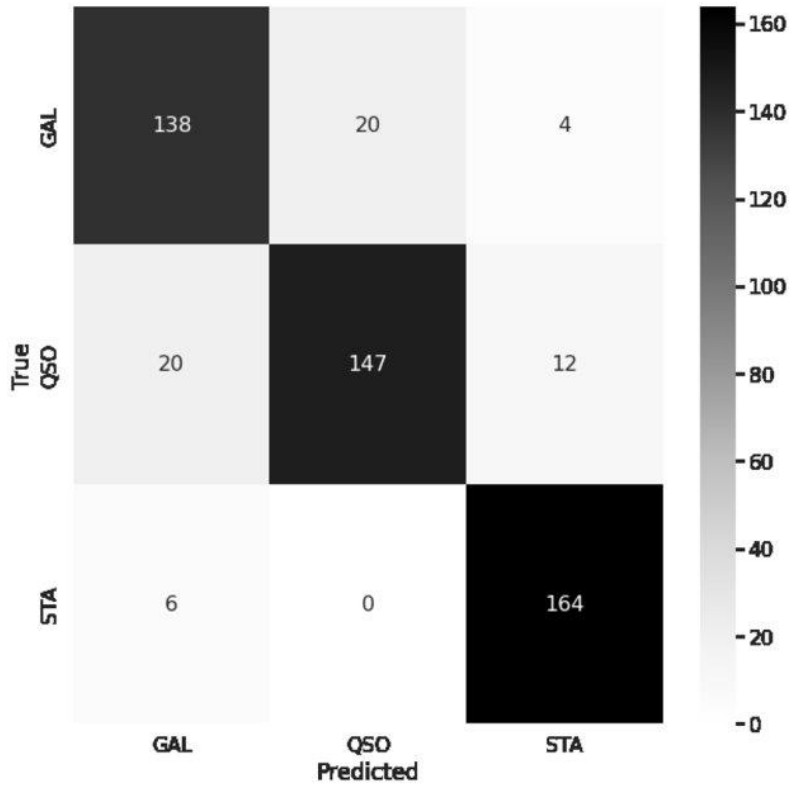


Fig. 3. Resulting confusion matrix of predicted and true classes for the first random execution of the RF using ALHAMBRA colours as features and the standard set of free parameters.

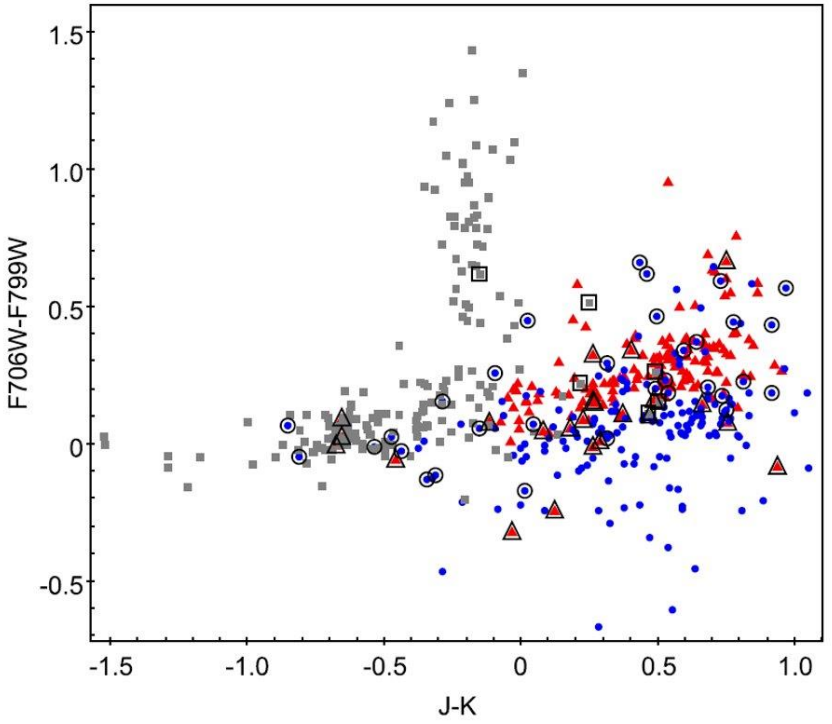


Fig. 5. Colour-colour diagram, $F706W - F799W$ versus $J - K$, for all stars (grey solid squares), galaxies (red solid triangles) and QSOs (blue solid circles) in the sample, for the same execution shown in Figure 3. Misclassified sources are also plotted as open symbols for stars (squares), galaxies (triangles) and QSOs (circles).

Resultados: comparación diferentes algoritmos

– K-nearest neighbours (KNN)
– gradient boosting (GBoost)
– support vector classifier (SVC)
– feedforward neural networks (FNN)

Resultados: comparación diferentes algoritmos

– K-nearest neighbours (KNN)
– gradient boosting (GBoost)
– support vector classifier (SVC)
– feedforward neural networks (FNN)

Table 3. Global accuracy (AC), precision, recall, and F1 metrics with their standard deviations σ (in brackets) obtained with different algorithms using either ALHAMBRA or SDSS colours.

Database	Classifier	AC (σ)	Class	Precision (σ)	Recall (σ)	F1 (σ)
ALHAMBRA	KNN	0.840 (0.025)	QSO	0.869 (0.073)	0.708 (0.091)	0.774 (0.043)
			GAL	0.773 (0.065)	0.858 (0.060)	0.810 (0.035)
			STA	0.893 (0.061)	0.962 (0.025)	0.925 (0.031)
ALHAMBRA	GBoost	0.864 (0.020)	QSO	0.845 (0.043)	0.801 (0.039)	0.821 (0.025)
			GAL	0.820 (0.058)	0.839 (0.044)	0.828 (0.030)
			STA	0.923 (0.053)	0.954 (0.023)	0.937 (0.027)
ALHAMBRA	SVC	0.828 (0.018)	QSO	0.797 (0.061)	0.733 (0.056)	0.761 (0.031)
			GAL	0.789 (0.068)	0.801 (0.049)	0.792 (0.041)
			STA	0.892 (0.054)	0.954 (0.023)	0.921 (0.025)
ALHAMBRA	FNN	0.838 (0.014)	QSO	0.690 (0.040)	0.440 (0.050)	0.540 (0.030)
			GAL	0.860 (0.020)	0.914 (0.008)	0.884 (0.011)
			STA	0.859 (0.018)	0.930 (0.030)	0.892 (0.013)
SDSS	KNN	0.784 (0.020)	QSO	0.709 (0.054)	0.701 (0.038)	0.704 (0.037)
			GAL	0.722 (0.046)	0.709 (0.039)	0.714 (0.029)
			STA	0.912 (0.046)	0.945 (0.032)	0.927 (0.026)
SDSS	GBoost	0.795 (0.017)	QSO	0.725 (0.049)	0.704 (0.046)	0.713 (0.035)
			GAL	0.732 (0.062)	0.737 (0.055)	0.734 (0.053)
			STA	0.916 (0.032)	0.943 (0.032)	0.929 (0.021)
SDSS	SVC	0.789 (0.026)	QSO	0.738 (0.075)	0.662 (0.060)	0.696 (0.055)
			GAL	0.714 (0.070)	0.747 (0.059)	0.729 (0.058)
			STA	0.901 (0.039)	0.959 (0.020)	0.928 (0.017)
SDSS	FNN	0.871 (0.015)	QSO	0.730 (0.040)	0.650 (0.050)	0.680 (0.030)
			GAL	0.901 (0.017)	0.915 (0.013)	0.908 (0.012)
			STA	0.880 (0.017)	0.916 (0.016)	0.898 (0.015)

Conclusiones (principales)

(a) El efecto de variar (dentro de valores razonables) parámetros libres como el número de árboles, profundidad máxima o número de features, es despreciable: los valores son los mismos (dentro de las incertidumbres).

Conclusiones (principales)

(a) El efecto de variar (dentro de valores razonables) parámetros libres como el número de árboles, profundidad máxima o número de features, es despreciable: los valores son los mismos (dentro de las incertidumbres).

(b) Usar colores en vez de magnitudes da mejores resultados, especialmente usar colores de Alhambra.

Conclusiones (principales)

(a) El efecto de variar (dentro de valores razonables) parámetros libres como el número de árboles, profundidad máxima o número de features, es despreciable: los valores son los mismos (dentro de las incertidumbres).

(b) Usar colores en vez de magnitudes da mejores resultados, especialmente usar colores de Alhambra.

(c) Los colores que contribuyen más son JHK (infrarrojo: IR).

Conclusiones (principales)

(a) El efecto de variar (dentro de valores razonables) parámetros libres como el número de árboles, profundidad máxima o número de features, es despreciable: los valores son los mismos (dentro de las incertidumbres).

(b) Usar colores en vez de magnitudes da mejores resultados, especialmente usar colores de Alhambra.

(c) Los colores que contribuyen más son JHK (infrarrojo: IR).

Resumen: para identificar fotométricamente QSOs es clave un conjunto de **colores** que incluya **bandas IR** y, por supuesto, un buen dataset...
--- mientras que el modelo de clasificación específico y el valor exacto de sus parámetros NO es determinante.

Próximo?

Clasificación de YSOs (Objetos Estelares Jóvenes) usando Machine Learning...

Próximo?

Clasificación de YSOs (Objetos Estelares Jóvenes) usando Machine Learning...

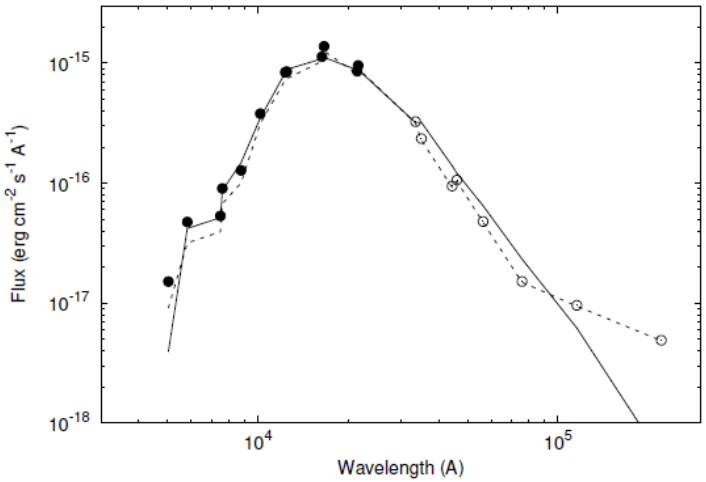


Fig. 6. Observed and best-fitted flux densities for one example source, J120022.63-631523.2, for which we obtained $T_{\rm eff}$ = 1700 K and $A_{\rm V}$ = 0.5, which is a transition disk object according to Koenig & Leisawitz (2014)'s criteria. The dashed line indicates the observed photometric data. Circles represent dereddened data, where solid circles denote data points that have been considered in the fitting process by VOSA. The solid lines indicate the best-fitted BT-Settl model. Some infrared excess is evident at wavelengths larger than ~10 μm.

Próximo?

Clasificación de YSOs (Objetos Estelares Jóvenes) usando Machine Learning...

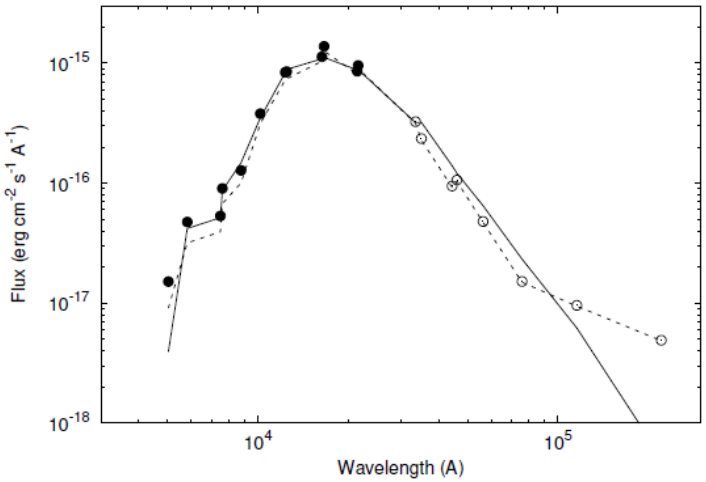


Fig. 6. Observed and best-fitted flux densities for one example source, J120022.63-631523.2, for which we obtained $T_{\rm eff}$ = 1700 K and $A_{\rm V}$ = 0.5, which is a transition disk object according to Koenig & Leisawitz (2014)'s criteria. The dashed line indicates the observed photometric data. Circles represent dereddened data, where solid circles denote data points that have been considered in the fitting process by VOSA. The solid lines indicate the best-fitted BT-Settl model. Some infrared excess is evident at wavelengths larger than ~10 μm.

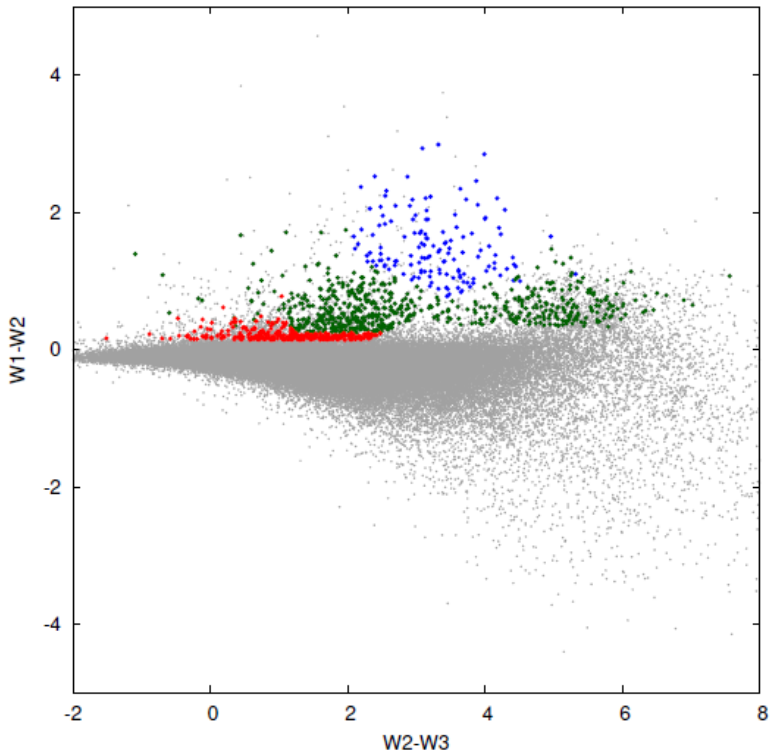


Fig. 3. Colour-colour diagram, $W1 - W2$ vs. $W2 - W3$, for all stars in the sample (grey dots) and for sources fulfilling criteria of Class I (blue dots), Class II (green dots), and transition disk (red dots) objects.

Próximo?

Clasificación de YSOs (Objetos Estelares Jóvenes) usando Machine Learning...

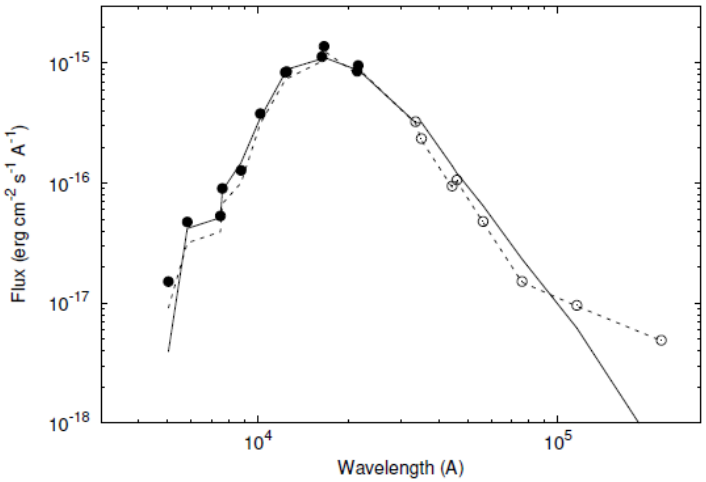


Fig. 6. Observed and best-fitted flux densities for one example source, J120022.63-631523.2, for which we obtained $T_{\rm eff}$ = 1700 K and $A_{\rm V}$ = 0.5, which is a transition disk object according to Koenig & Leisawitz (2014)'s criteria. The dashed line indicates the observed photometric data. Circles represent dereddened data, where solid circles denote data points that have been considered in the fitting process by VOSA. The solid lines indicate the best-fitted BT-Settl model. Some infrared excess is evident at wavelengths larger than ~10 μm.

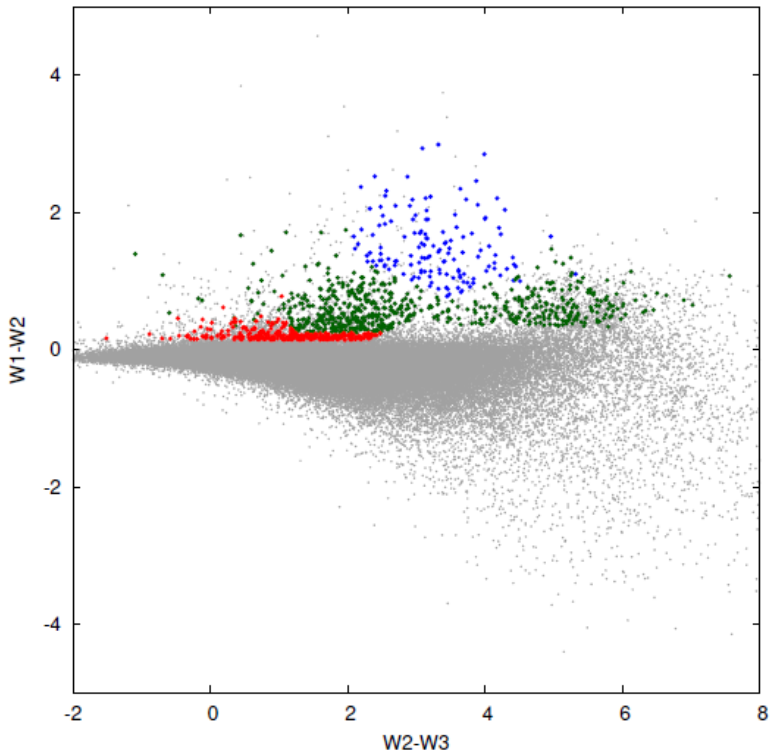


Fig. 3. Colour-colour diagram, $W1 - W2$ vs. $W2 - W3$, for all stars in the sample (grey dots) and for sources fulfilling criteria of Class I (blue dots), Class II (green dots), and transition disk (red dots) objects.

Primer (y principal) reto:

- LOS DATOS !!!