Detección de exoplanetas con redes neuronales usando la base de datos de la misión TESS

Irene Delgado Borrego

Supervisor : Dr. César Augusto Guzmán Álvarez

8 de Julio de 2020

Indice

Introducción

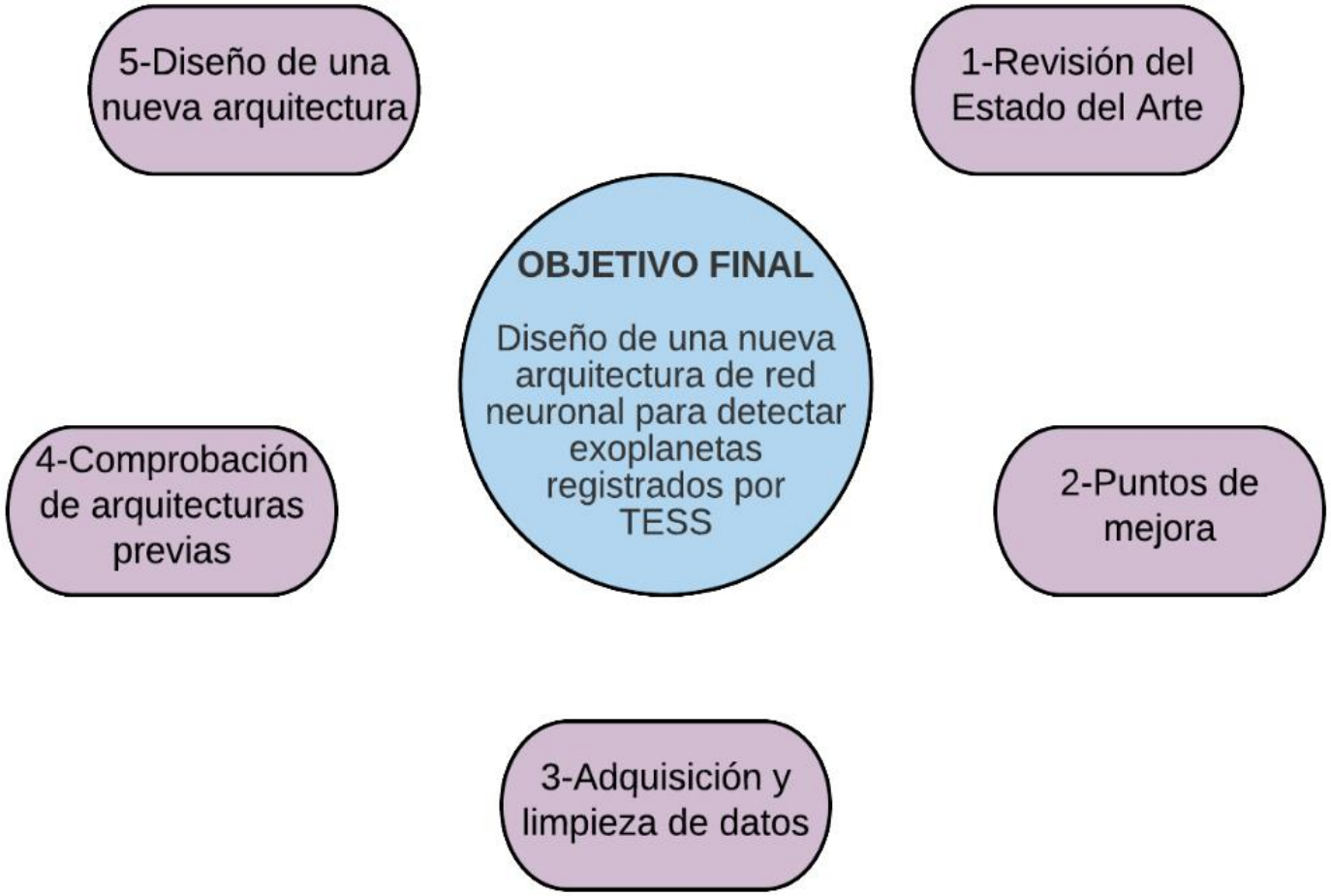
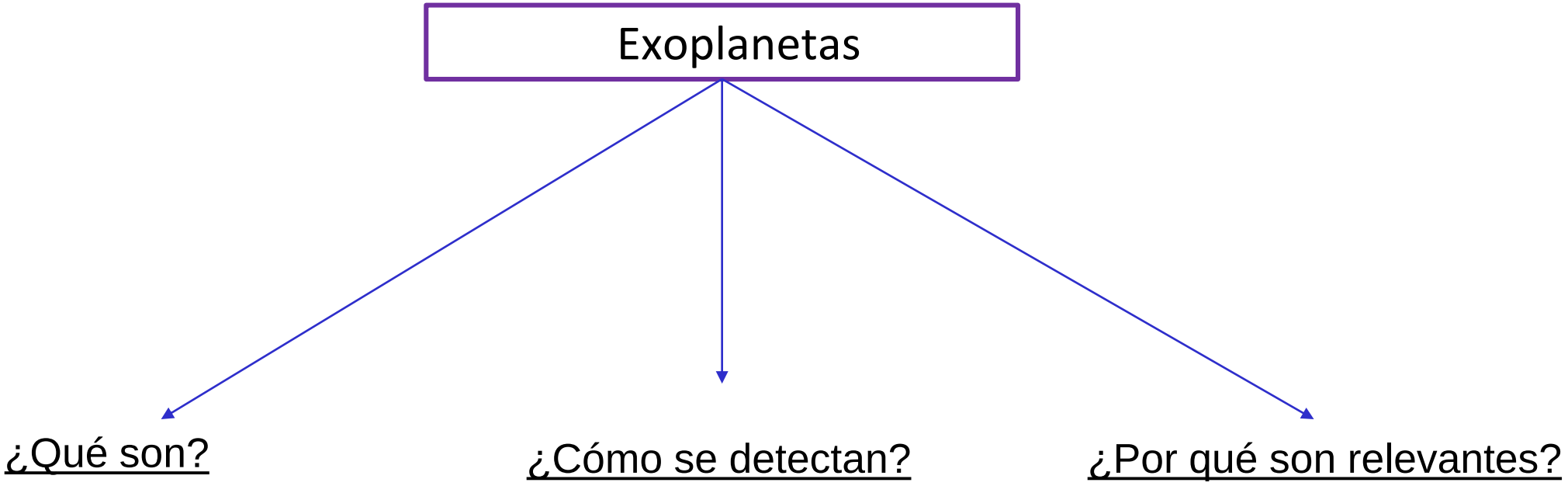
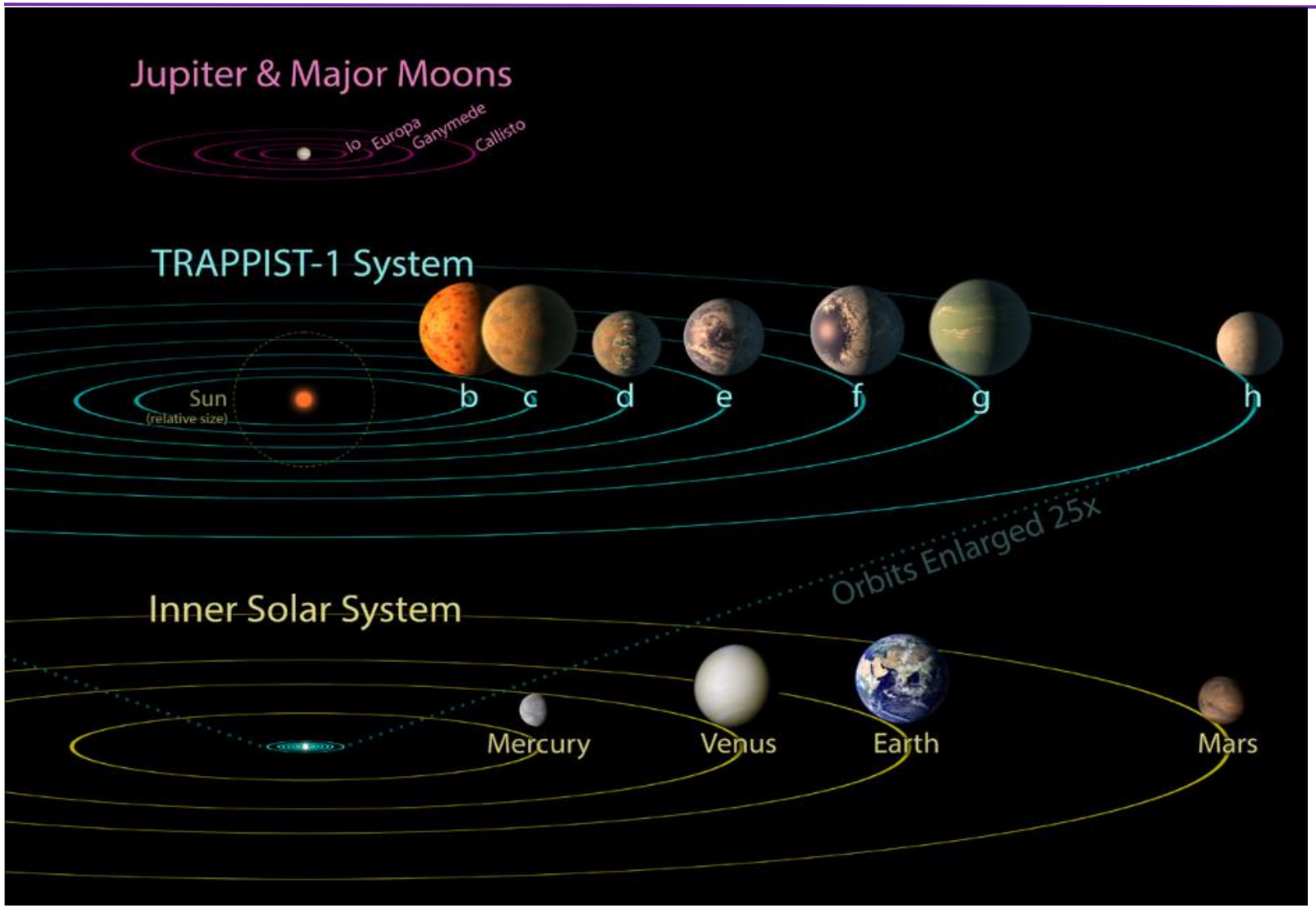


Figura 1. *Diagrama de objetivos*

Estado del Arte- Exoplanetas(I)



Estado del Arte- Exoplanetas(II)

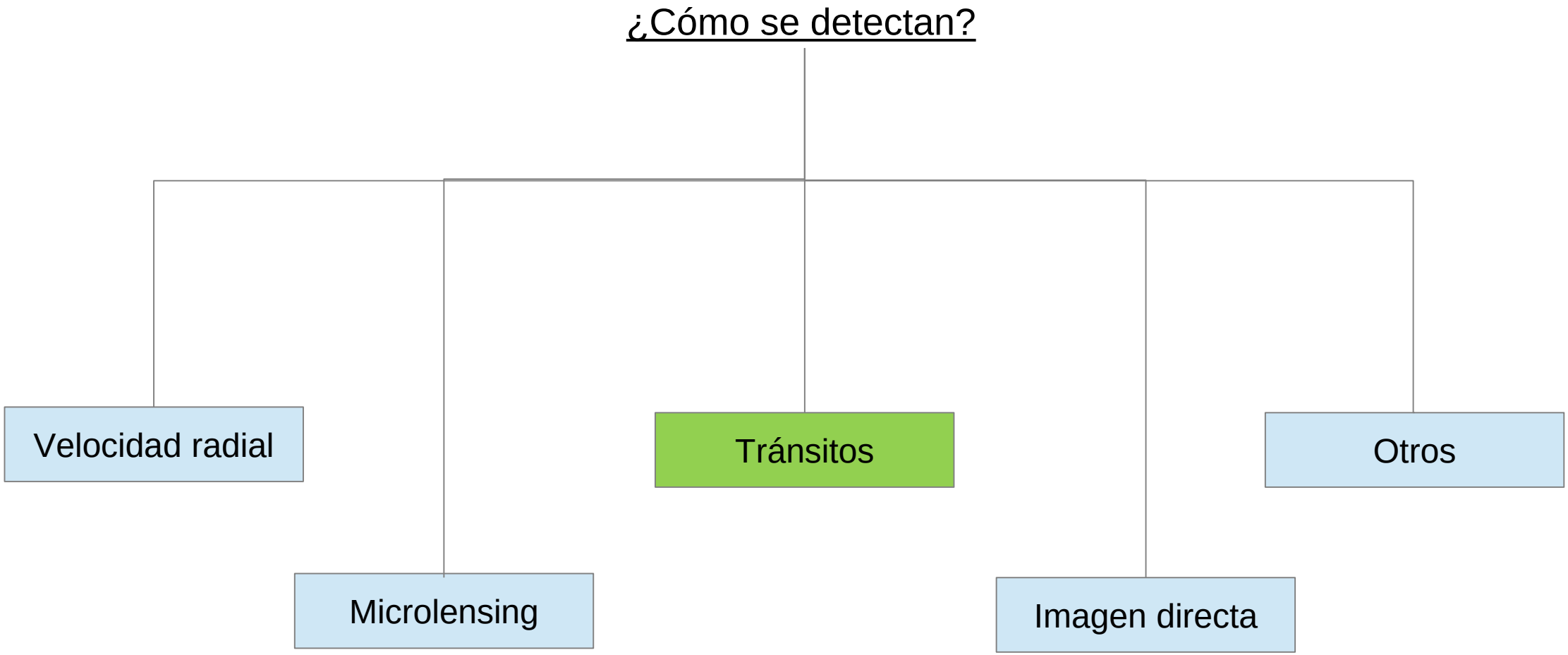


¿Qué son?

- Planetas que orbitan estrellas diferentes al Sol

Figura 2. *Representación artística entre el sistema solar interno y el sistema TRAPPIST-1*
Fuente:Nasa's Jet Propulsion Laboratory

Estado del Arte- Exoplanetas(III)



Estado del Arte- Exoplanetas(IV)

¿Por qué son relevantes?

- Permite estudiar características de nuestros propio sistema Solar
- Búsqueda de vida
- Zona de habitabilidad

Estado del Arte- Misión TESS(I)

TESS: *Transiting Exoplanet Survey Satellite*

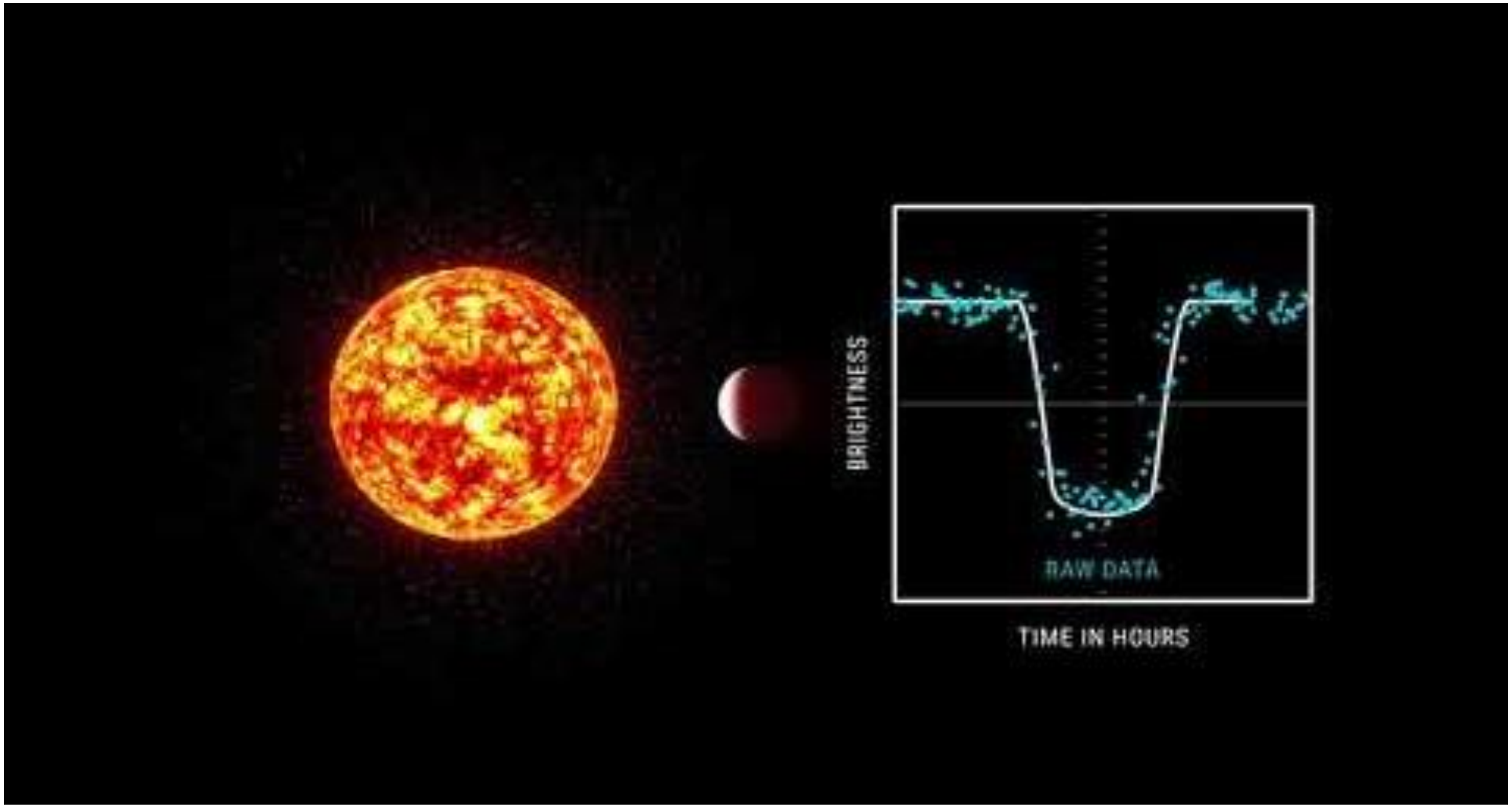
Misión espacial

Utiliza el método de los Tránsitos para la detección

Figura 3. *Imagen real del telescopio espacial TESS (Transiting Exoplanet Survey Satellite) Fuente: https://www.nasa.gov/content/tess-images*

Estado del Arte- Misión TESS(II)

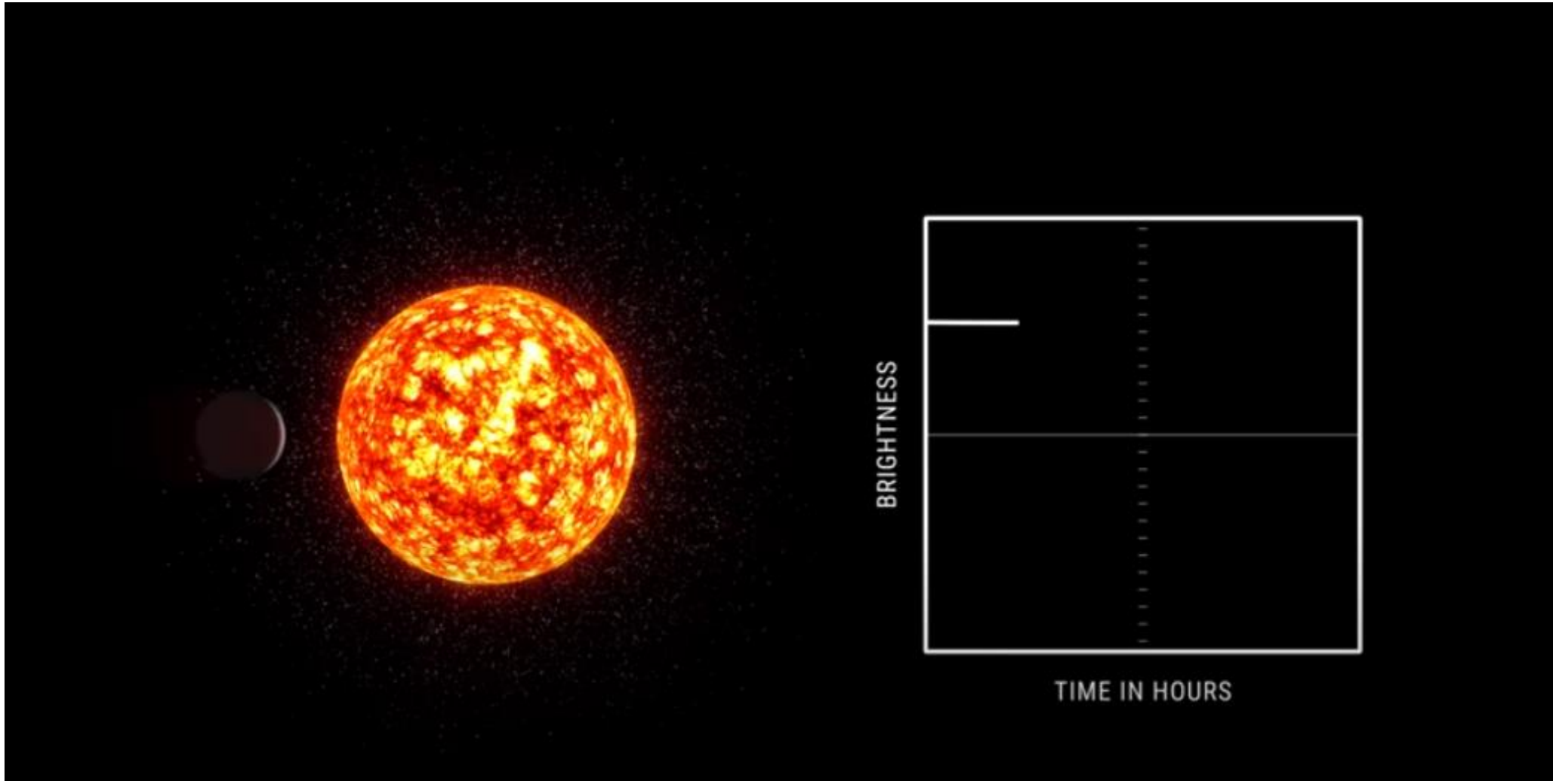
Método de los tránsitos



Video 1. *Animación por ordenador de un tránsito planetario. Fuente: Nasa*
https://youtu.be/BFi4HBUdWkk

Estado del Arte- Misión TESS(II)

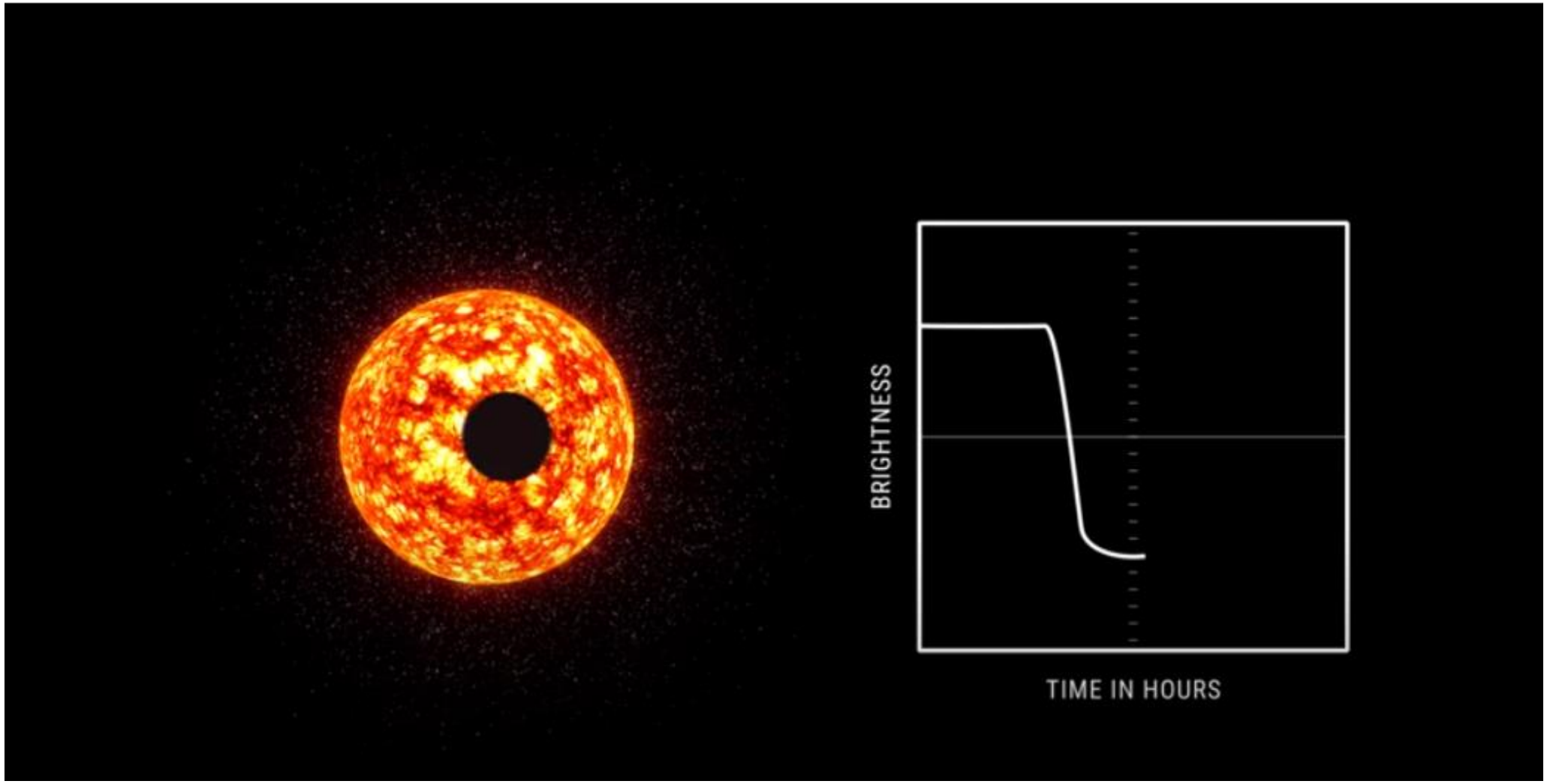
Método de los tránsitos



Video 1 Captura 1 *Animación por ordenador de un tránsito planetario. Fuente: Nasa*
https://youtu.be/BFi4HBUdWkk

Estado del Arte- Misión TESS(II)

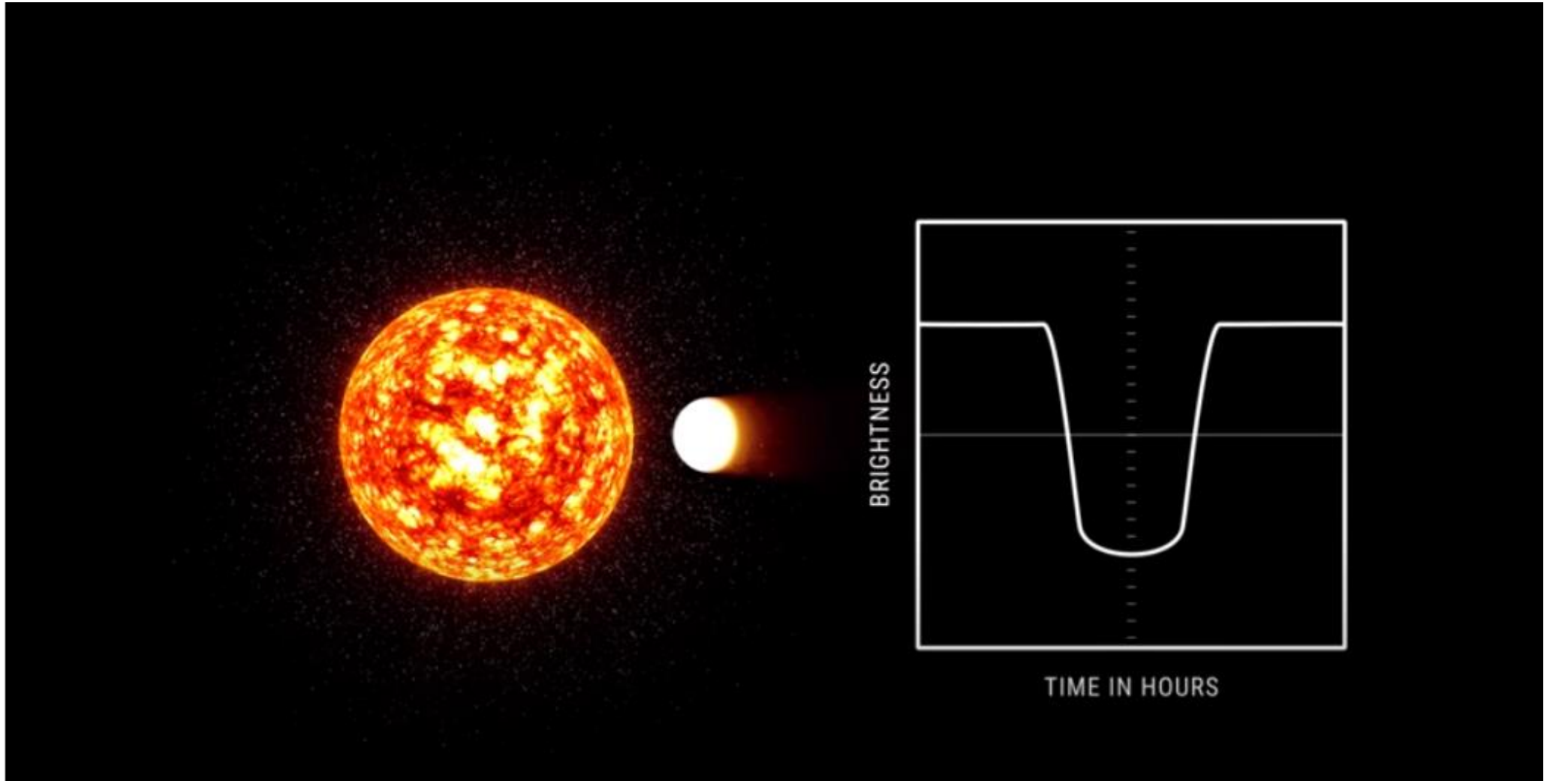
Método de los tránsitos



Video 1 Captura 2 *Animación por ordenador de un tránsito planetario. Fuente: Nasa*
https://youtu.be/BFi4HBUdWkk

Estado del Arte- Misión TESS(II)

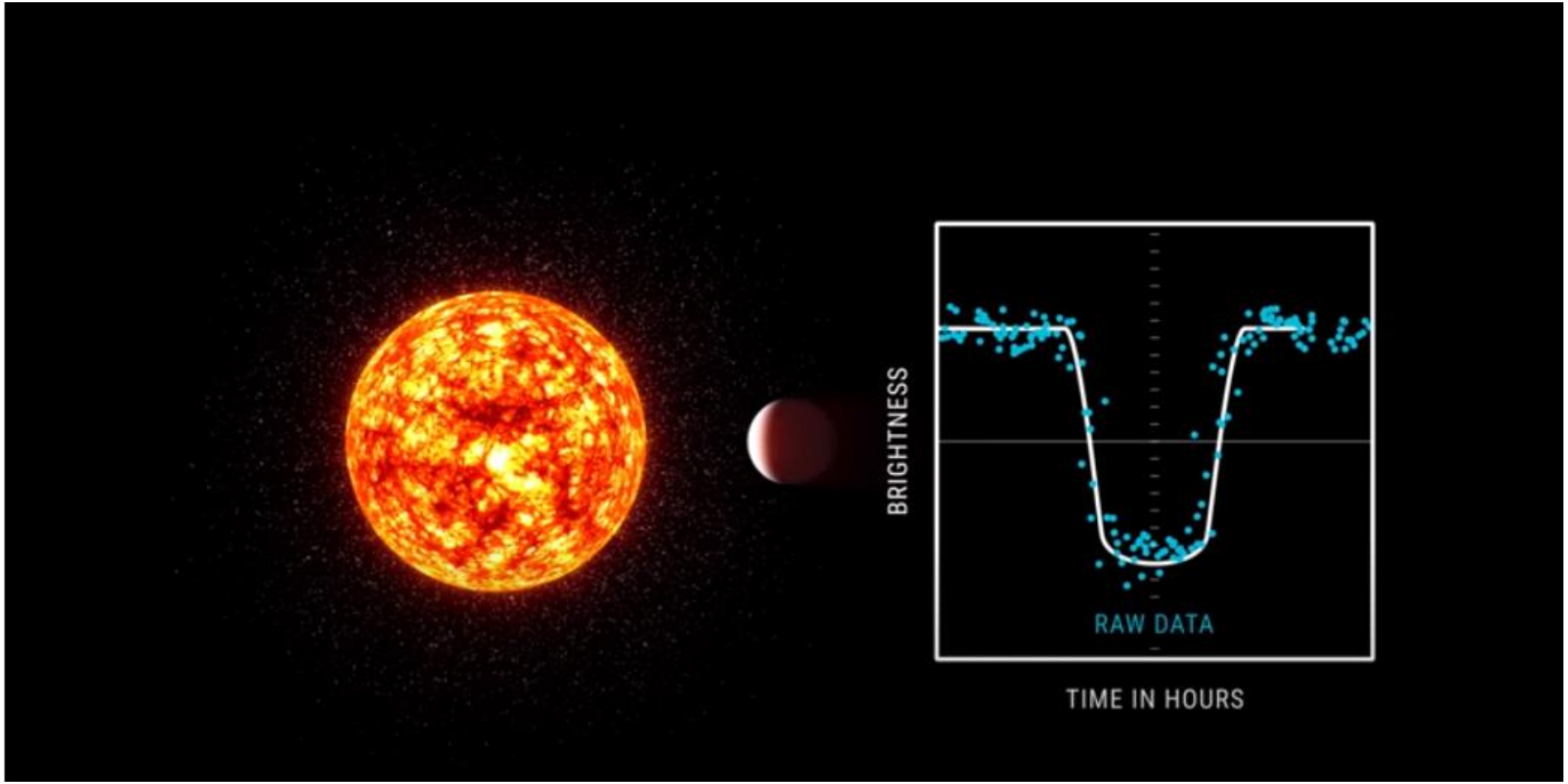
Método de los tránsitos



Video 1 Captura 3 *Animación por ordenador de un tránsito planetario. Fuente: Nasa*
https://youtu.be/BFi4HBUdWkk

Estado del Arte- Misión TESS(II)

Método de los tránsitos

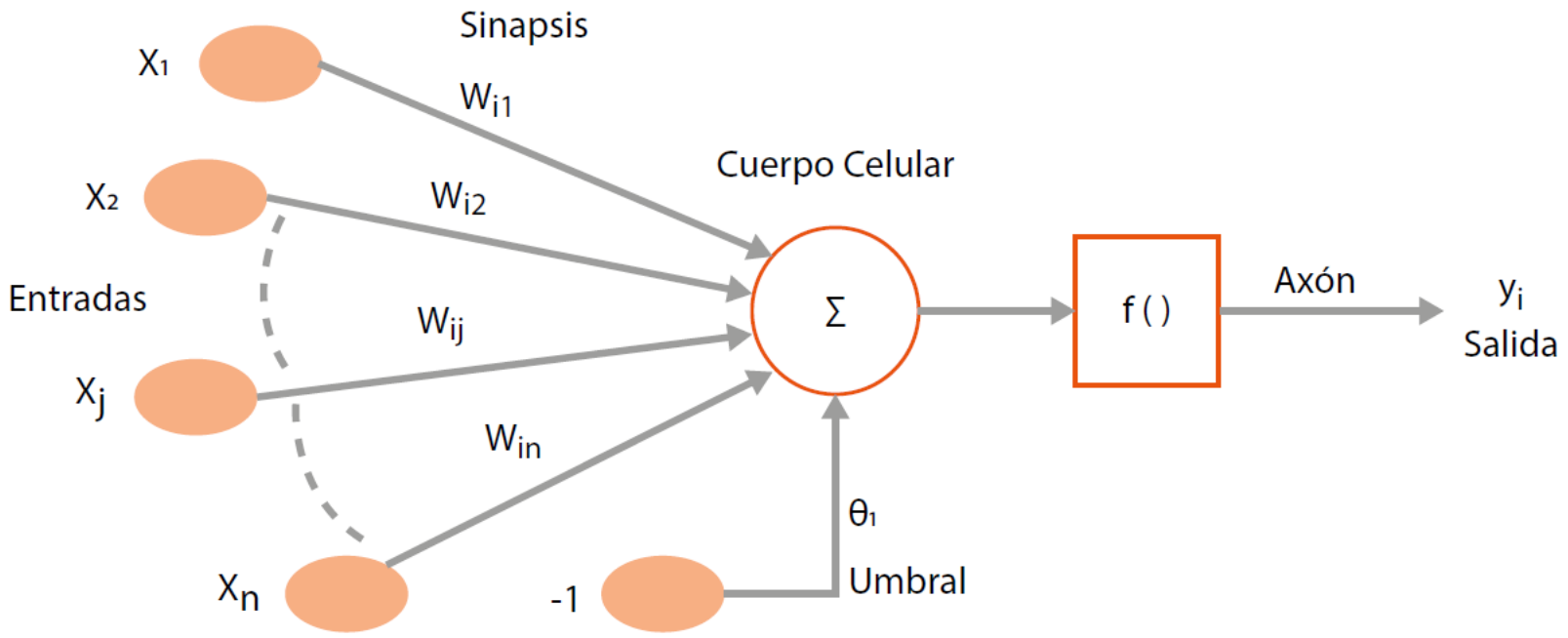


Video 1 Captura 4 *Animación por ordenador de un tránsito planetario. Fuente: Nasa*
https://youtu.be/BFi4HBUdWkk

Estado del Arte- Aprendizaje Automático(I)

Figura 4. *Esquema de los diferentes tipos de aprendizajes que existen y las técnicas más representativas asociadas a ellos. Fuente: Universitat Oberta de Catalunya*

Estado del Arte- Red Neuronal(I)



- Entradas: $\bar{X} = (x_1, x_2, \ldots, x_n)$
- Sinapsis: w_{ij}
- Cuerpo celular: $\sum$
- Umbral de activación: θ_1
- Función de activación: f( )
- Salida: $y_i = f\left(\sum_j w_{ij} x_n - \theta_1\right)$

Figura 5. *Representación de las partes de una neurona artificial. Fuente: (Brío y Sanz, 1997*

Estado del Arte- Red Neuronal(II)

Recurrente

- Todas las capas están conectadas entre si.
- El parámetro más importante es w_{ij}
- Útiles en resolución de patrones temporales

Convolucional

- Red tridimensional
- Añaden dos tipos de capa extra; de convolución y de *pooling*
- Útiles en procesamiento de audio e imágenes

Estado del Arte- Estudios Previos(I)

- *Backpropagation Applied to Handwritten Zip Code Recognition, LeCun 1989*
- *Identifying Exoplanets with Deep Learning, Shallue 2018*
- *Rapid Classification of TESS Planet Candidates with Convolutional Neural Networks, Osborn 2019*

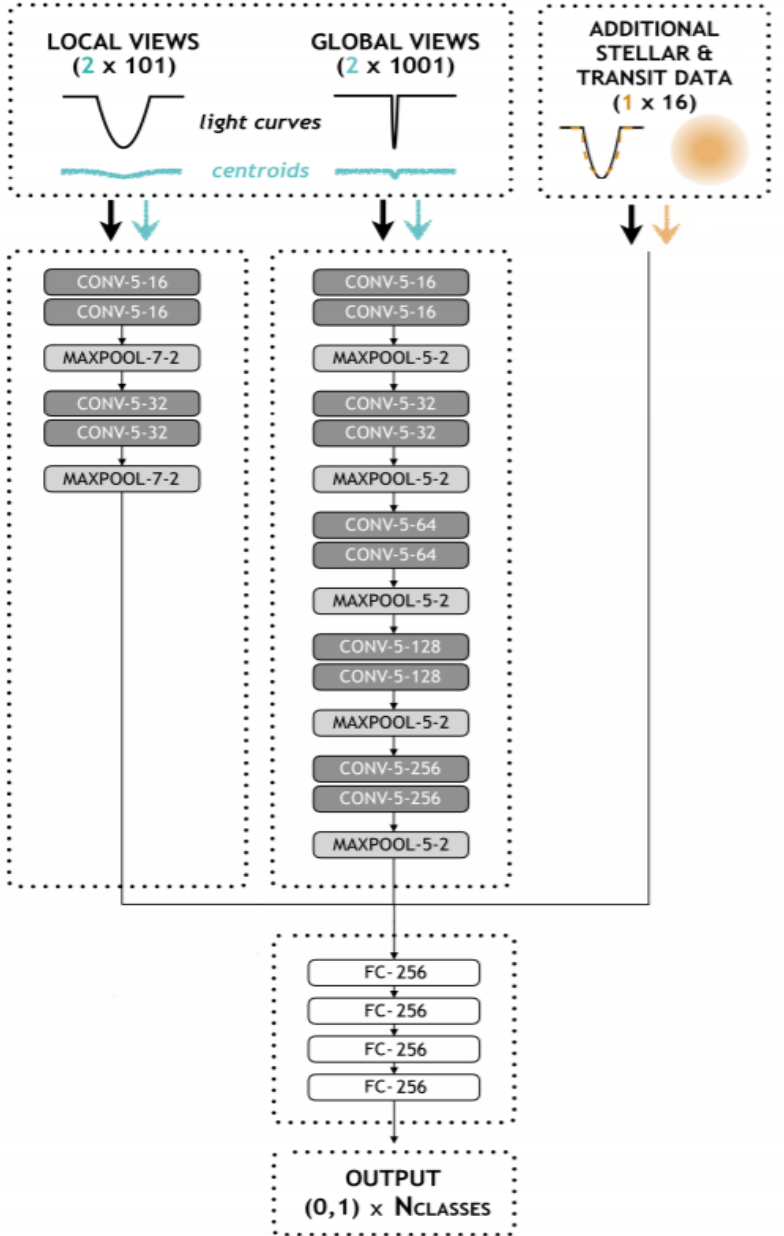


Figura 6. *Esquema de la CNN Exonet.Fuente: Osborn y col, 2019*

Desarrollo- Introducción(I)

BigML

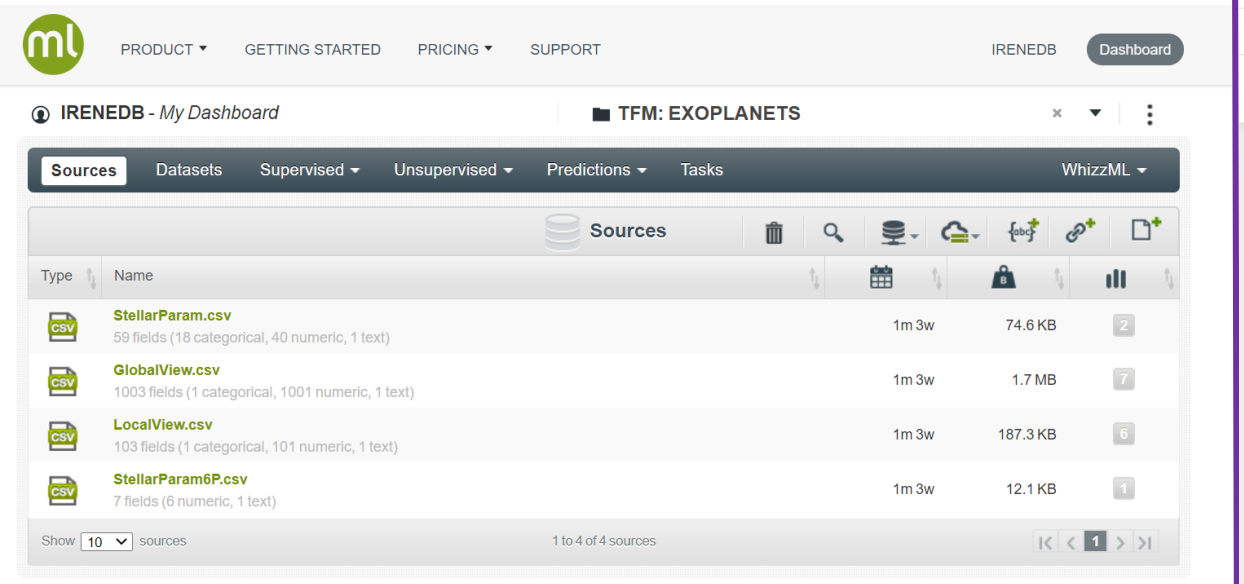


Figura 7. *Dashboard de la herramienta BigML.*

Programación en Python en *Jupyter Notebook*

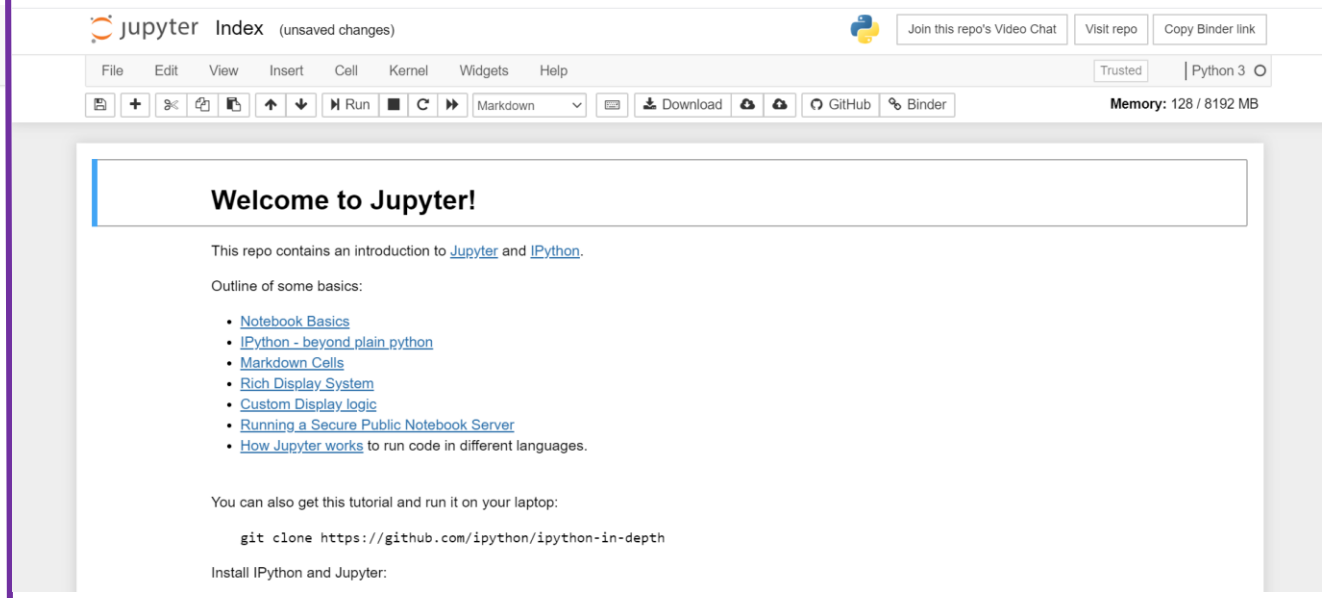


Figura 8 *Interfaz de Jupiter Notebook*

Desarrollo– Introducción(II)

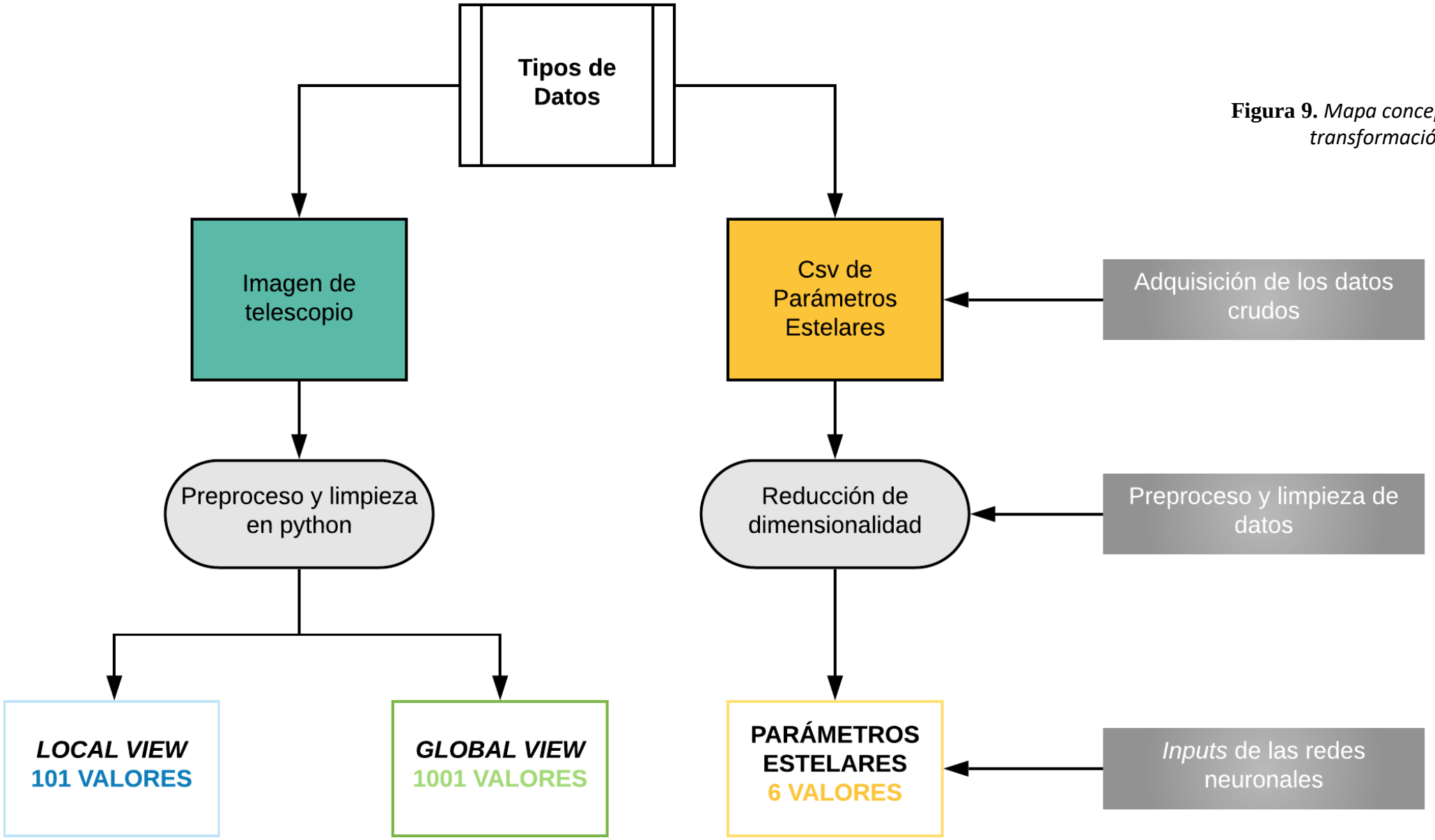
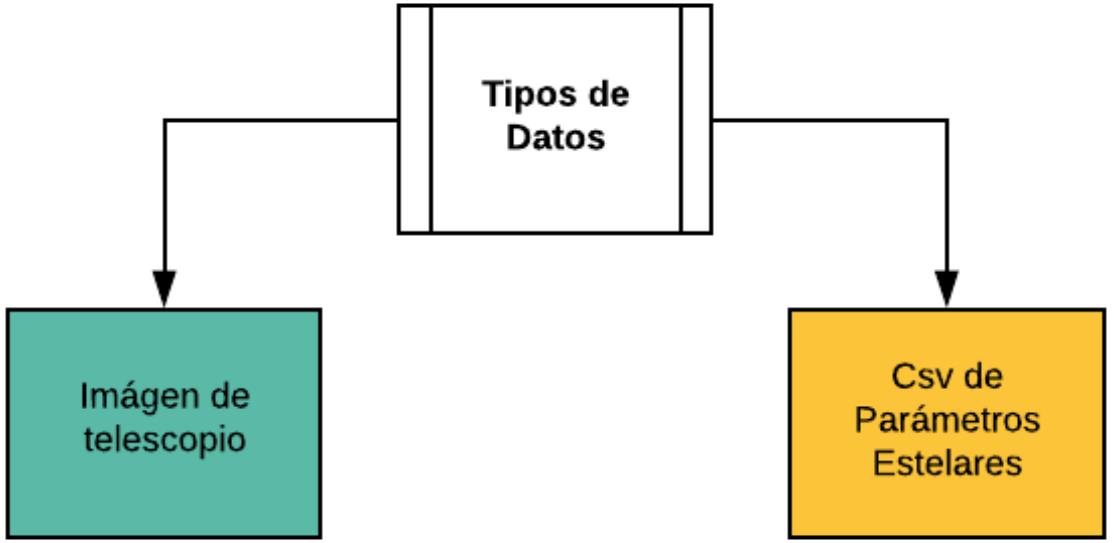


Figura 9. *Mapa conceptual del proceso de obtención, transformación y carga de los datos.*

Desarrollo– Adquisición de los datos(I)

Características

- Fuente: La NASA
- Datos reales de la misión TESS
- Datos de exoplanetas confirmados y falsos positivos
- Número de registros muy reducido
- Dos tipos diferentes de datos



Desarrollo– Preproceso y limpieza de datos(I)

CSV de Parámetros Estelares

- Está en el formato adecuado
- Aplicamos criterio experto

↓

Pasamos de 58 a 9 dimensiones

↓

- Aplicamos PCA y Correlaciones en BigML

↓

Pasamos de 9 a 6 dimensiones

→

Input de la red neuronal

PARÁMETROS ESTELARES
6 VALORES

Desarrollo– Preproceso y limpieza de datos(II)

Imagen de telescopio

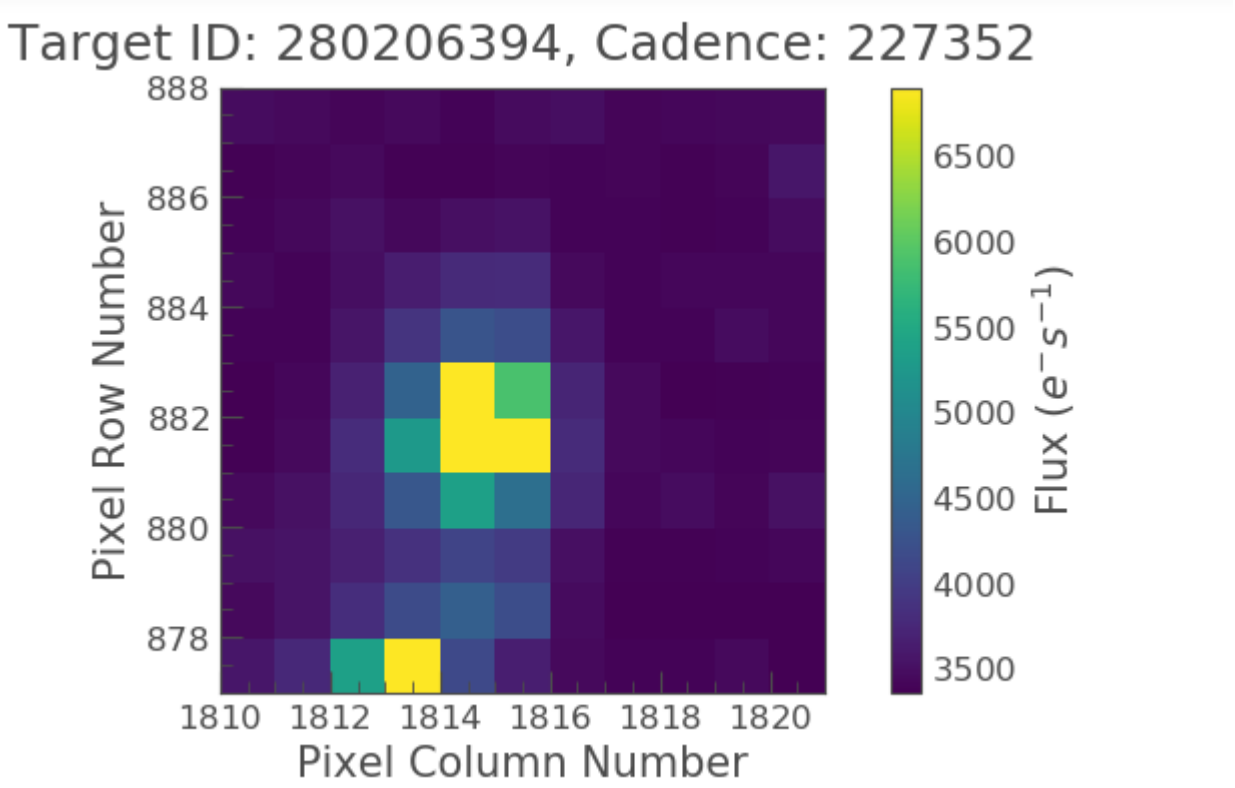


Figura 10. *Imagen .fits del exoplaneta "TOI 677.01".*

- Imagen de la estrella anfitriona
- Píxeles de una cámara CCD
- Formato .fits
- Necesitamos la variación del flujo de luz

↓

Hay que procesarlo

Desarrollo– Preproceso y limpieza de datos(III)

Imagen de telescopio

- Se procesa en Python en Jupyter Notebook
- Procesamos registros positivos y negativos
- Nos basamos en el código recogido en Shallue 2018→ Lo adaptamos a los datos de TESS

Con el código conseguimos:

→ Extraer desde las imágenes la curva de luz

→ Limpiar *outliers*

→ Normalizar y centrar el tránsito

→ Descargar los valores del flujo de luz en formato CSV

Desarrollo– Preproceso y limpieza de datos(IV)

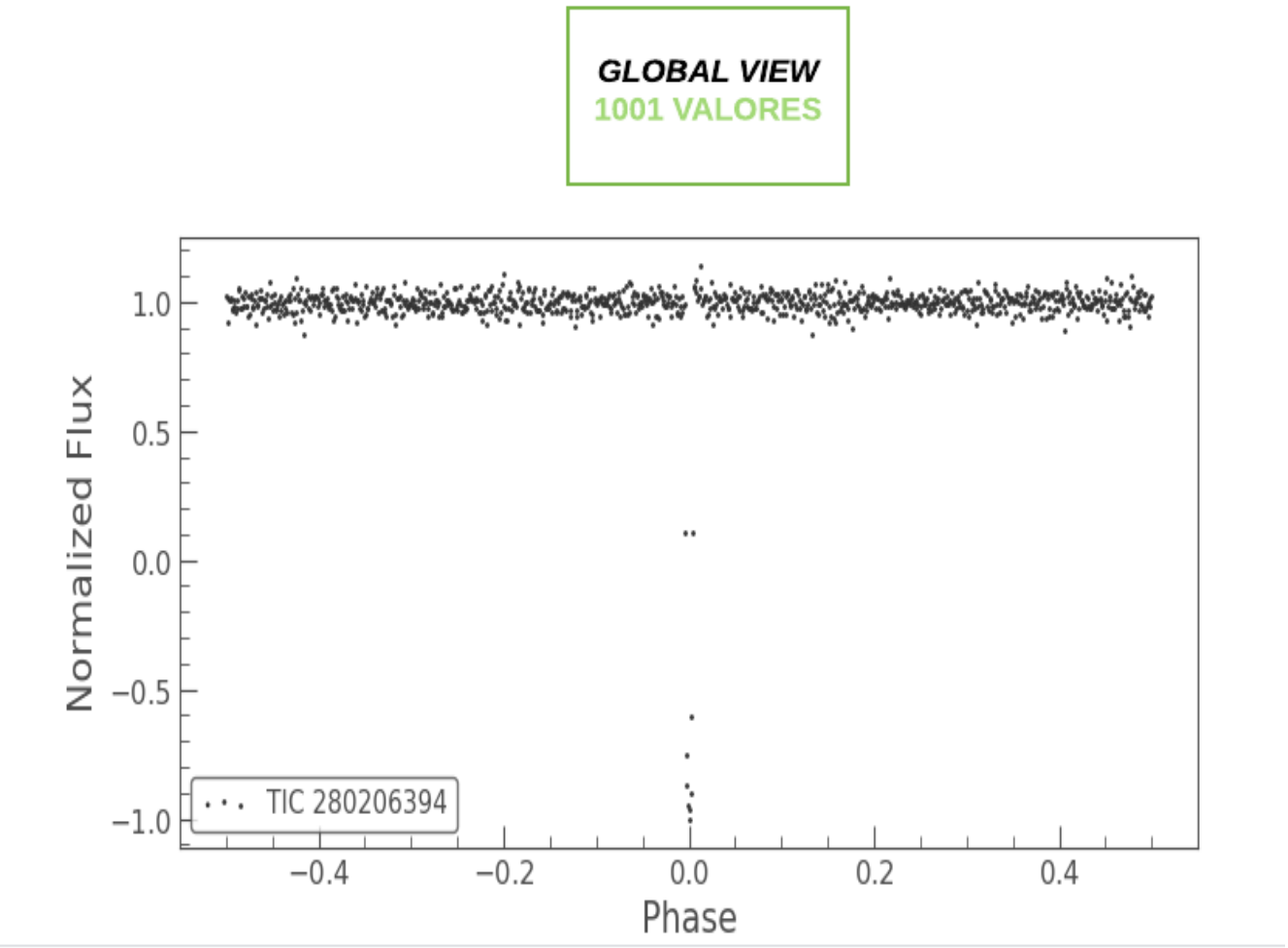


Figura 11. *Representación gráfica del Global View exoplaneta "TOI677.01."*

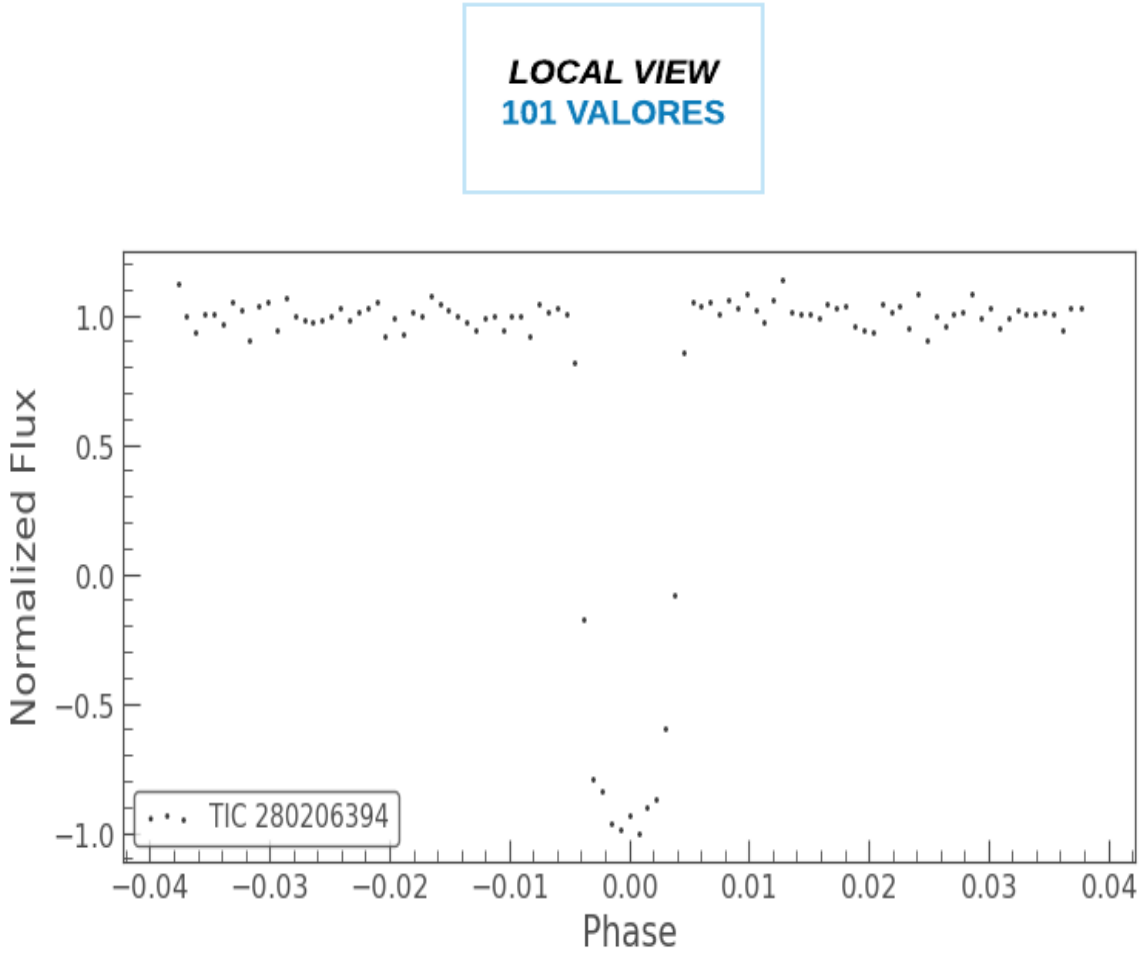
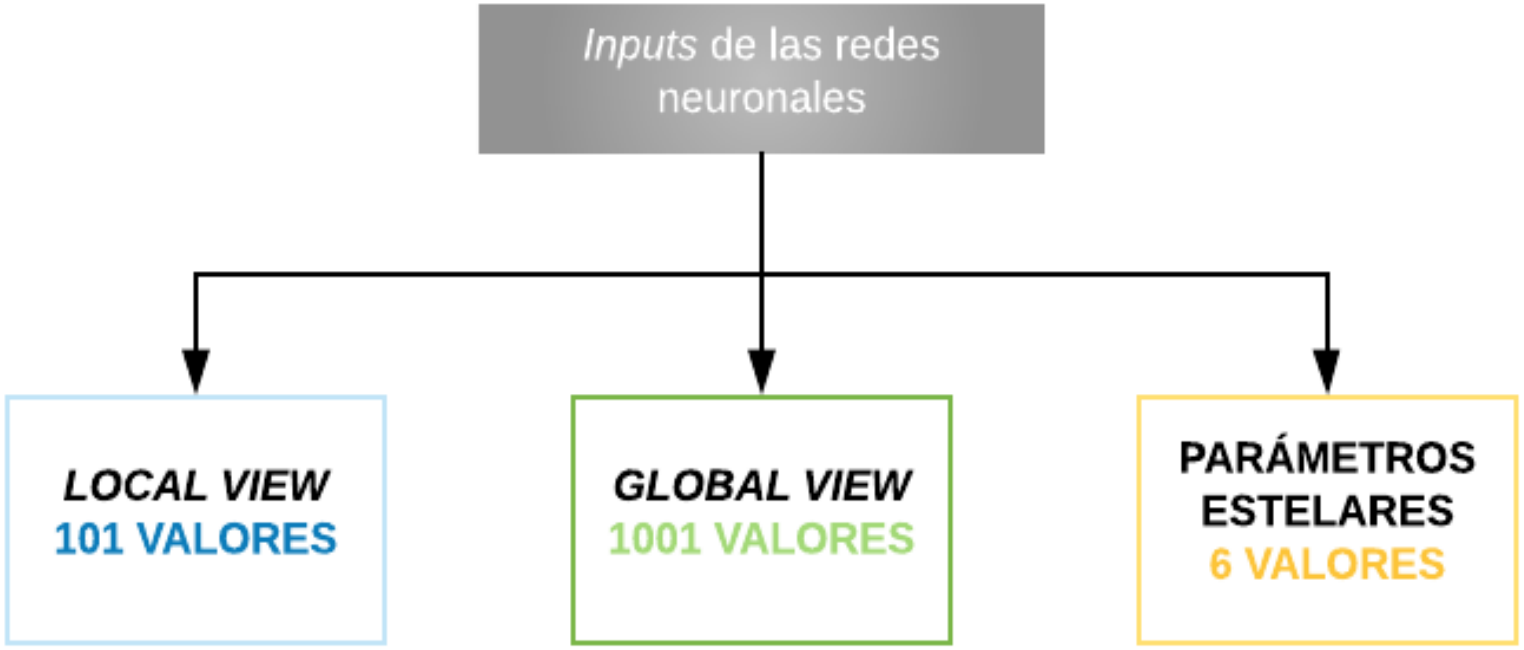


Figura 12. *Representación gráfica del Local View exoplaneta "TOI677.01."*

Desarrollo– Inputs(I)



- Hay 3 entradas diferentes en la red
- Valores numéricos en formato CSV
- Unificamos criterios y obtenemos para cada Input:

  33 muestras positivas
  116 muestras negativas

Desarrollo- Implementación(I)

1ª Fase: Red recurrente:

- Más simple
- Solo usamos capas conectadas
- Procesamos los inputs por igual
- Estructura de las capas conectadas de la red ExoNet
- Probamos el compartamiento de la red y de las entradas
- Se realiza en BigML

2ª Fase: Red convolucional

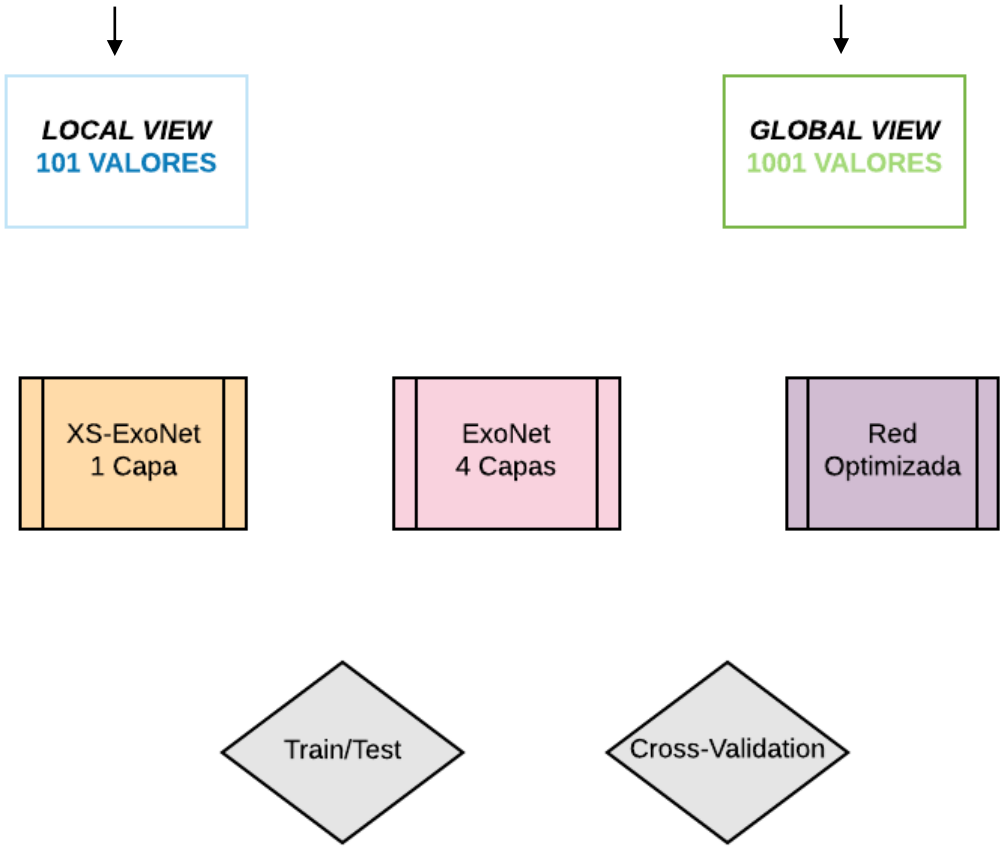
- Más compleja
- Usamos capas convolucionales, de pooling y conectadas
- Cada input se procesa de diferente forma en las capas convolucionales
- Estructura completa de la red ExoNet
- Se realiza mediante código en python

Desarrollo– 1ªFase(I)

Realizamos pruebas al Local View y al Global View por separado

Aplicamos tres tipos de arquitecturas diferentes

Comparamos resultados con dos métodos de validación



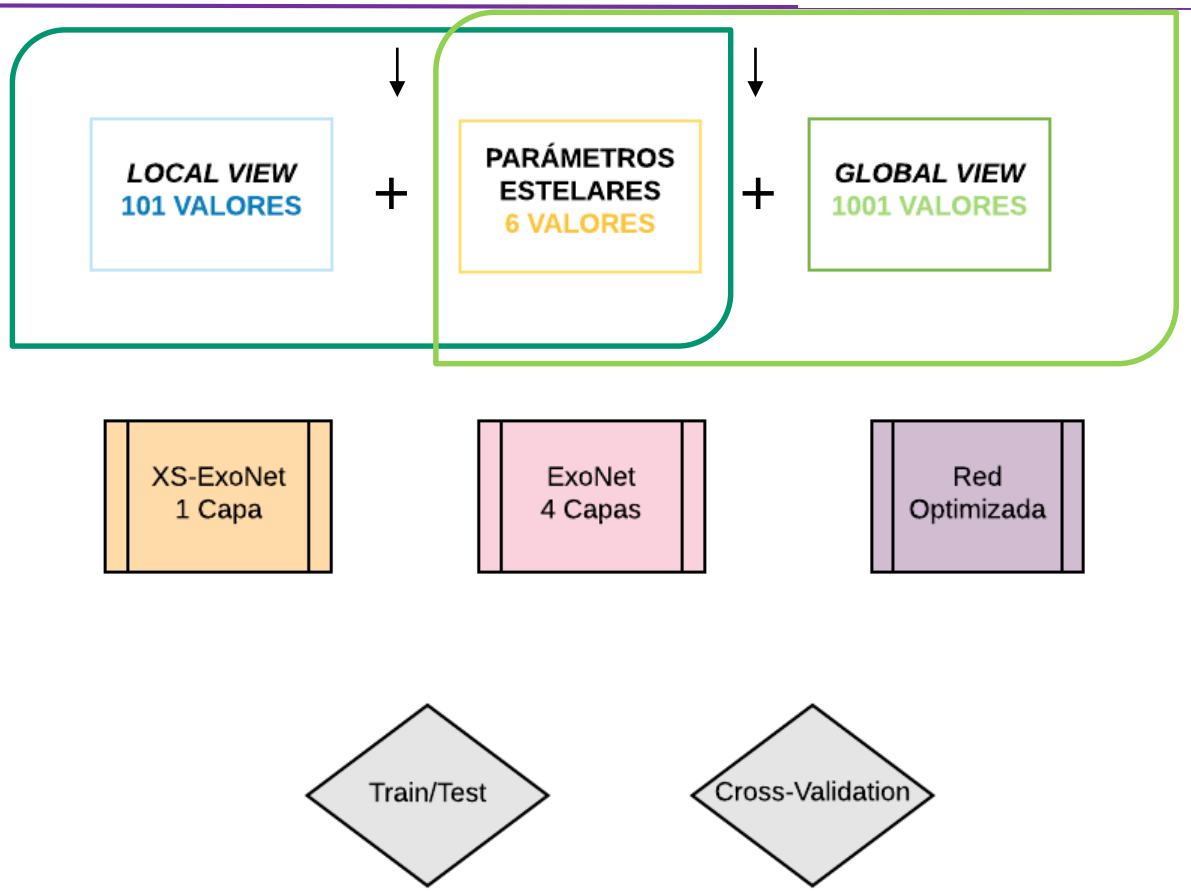
Desarrollo– 1ªFase(II)

Realizamos pruebas al Local View y al Global View por separado

Aplicamos tres tipos de arquitecturas diferentes

Comparamos resultados con dos métodos de validación

Añadimos los parámetros estelares y repetimos las pruebas



Desarrollo– 2ªFase(I)

- Utilizamos los tres Inputs

- Cada Input se procesa diferente:
  -Local View: Procesamiento mediante capas convolucionales
  -Global View: Procesamiento mediante capas convolucionales
  -Parámetros : No se procesan en capas convolucionales

- Se concatenan los datos anteriores, que sirven como nuevo Input para las capas conectadas.

- **Innovaciones:**

  - Nueva estructura en la rama convolucional del Global View
  - Modificación del umbral
  - Modificación de la función de activación
  - Modificación del algoritmo de optimización

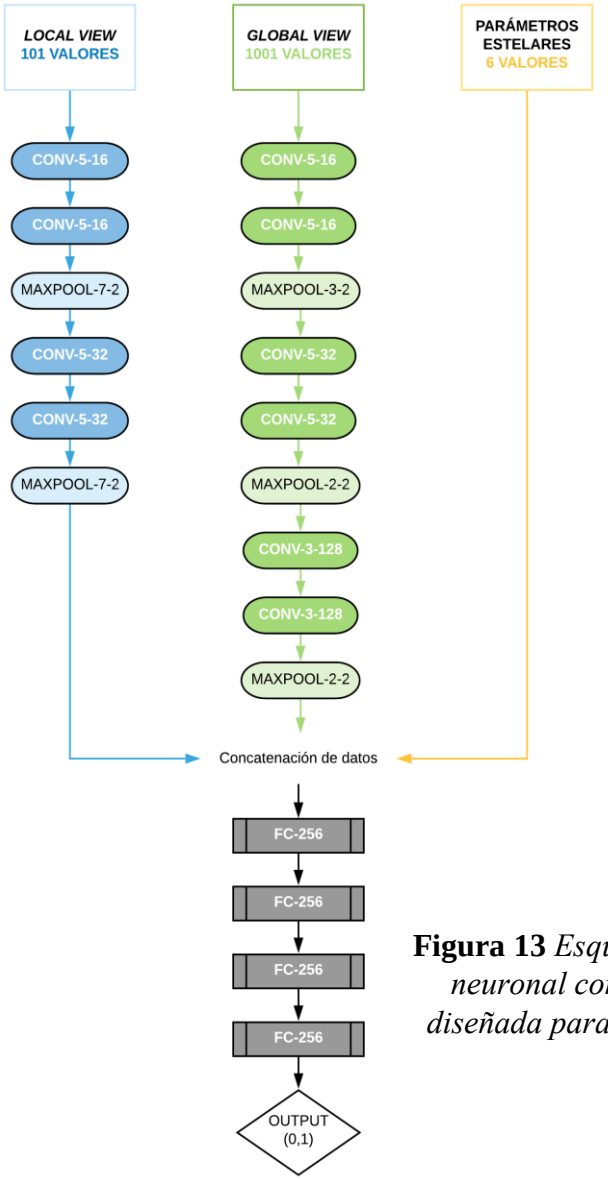


Figura 13 *Esquema de la red neuronal convolucional diseñada para este trabajo.*

Resultados y Discusión– Introducción(I)

Para medir la eficacia de los modelos utilizaremos diferentes métricas basadas principamente en la matriz de confusión:

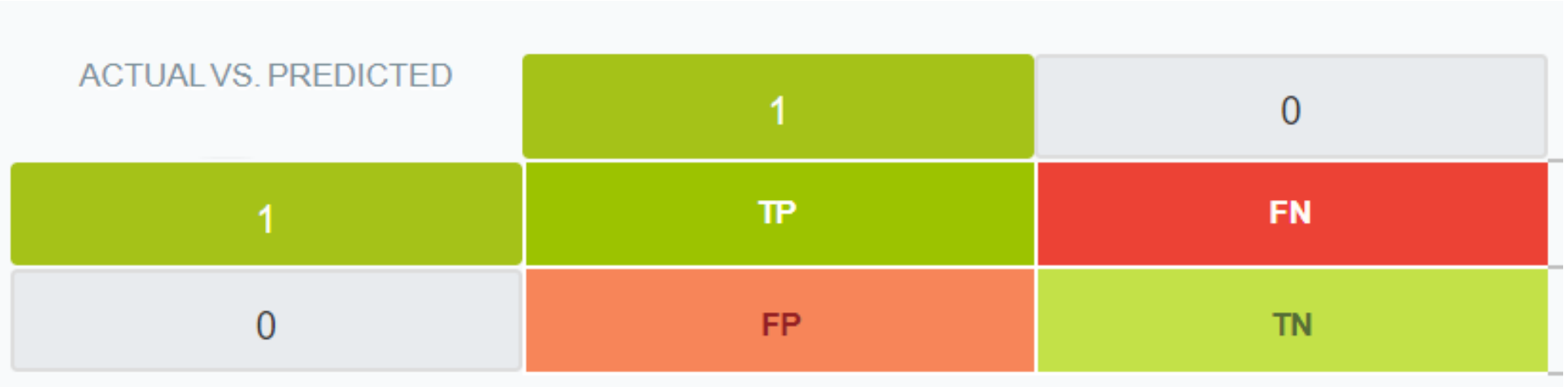


Figura 14 *Ejemplo de una matriz de confusión para una clasificación binaria*

Resultados y Discusión– Introducción(II)

Para medir la eficacia de los modelos utilizaremos diferentes métricas basadas principamente en la matriz de confusión:

- *Accuracy* = $\frac{TP+TN}{TP+TN+FP+FN}$
- *Recall* = $\frac{TP}{TP+FN}$
- *Precision* = $\frac{TP}{TP+FP}$
- *F-measure* = 2 x $\frac{Precision\ x\ Recall}{Precision+\ Recall}$
- Curva *ROC-AUC:*

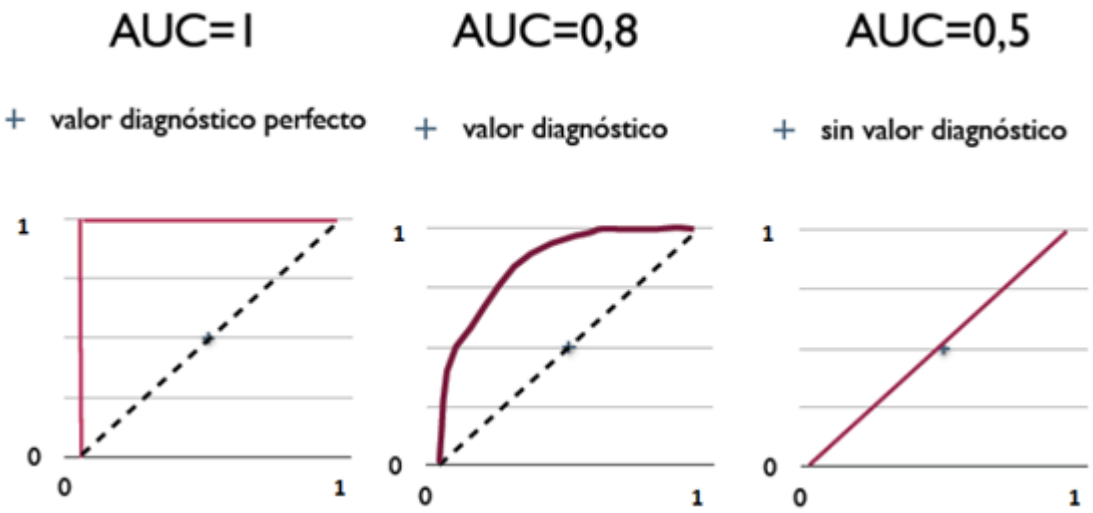


Figura 15 *Posibles valores de curvas ROC*
Fuente: UPO649 1112 prod-gom (2011)

Resultados y Discusión–1ªFase(I)

Sin parámetros Estelares:

Local View

	TRAIN/TEST		CROSS-VALIDATION	
	F-measure	*ROC AUC*	*F-measure*	*ROC AUC*
XS-Exonet	0.619	0.718	0.44	-
Exonet	0.625	0.666	0.44	-
Optimización	0.538	0.761	0.54	-

Tabla 1 *Resumen de los resultados obtenidos para el dataset Local View en la plataforma BigML sin tener en cuenta los parámetros estelares en la redes recurrentes*

Global View

	TRAIN/TEST		CROSS-VALIDATION	
	F-measure	*ROC AUC*	*F-measure*	*ROC AUC*
XS-Exonet	0.444	0.599	0.45	-
Exonet	0.211	0.679	0.4	-
Optimización	0.378	0.518	0.46	-

Tabla 2 *Resumen de los resultados obtenidos para el dataset Global View en la plataforma BigML sin tener en cuenta los parámetros estelares en la redes recurrentes*

Resultados y Discusión–1ªFase(II)

Con parámetros Estelares:

Local View

	TRAIN/TEST		CROSS-VALIDATION	
	F-measure	*ROC AUC*	*F-measure*	*ROC AUC*
XS-Exonet	0.422	0.722	0.50	-
Exonet	0.511	0.572	0.52	-
Optimización	0.770	0.734	0.64	-

Tabla 3 *Resumen de los resultados obtenidos para el dataset Local View en la plataforma BigML teniendo en cuenta los parámetros estelares en la redes recurrentes*

Global View

	TRAIN/TEST		CROSS-VALIDATION	
	F-measure	*ROC AUC*	*F-measure*	*ROC AUC*
XS-Exonet	0.505	0.561	0.48	-
Exonet	0.423	0.553	0.44	-
Optimización	0.459	0.465	0.56	-

Tabla 4 *Resumen de los resultados obtenidos para el dataset Global View en la plataforma BigML teniendo en cuenta los parámetros estelares en la redes recurrentes*

Resultados y Discusión–2ªFase(I)

- **Recordamos las innovaciones introducidas:**

  - Nueva estructura en la rama convolucional del Global View

    - Modificación del umbral
    - Modificación de la función de activación
    - Modificación del algoritmo de optimización

→Obtenemos diferentes resultados al ir aplicando las anteriores modificaciones

Resultados y Discusión–2ªFase(II)

- **Estructura Semi-ExoNet con modificación de umbral**

Validación Train/Test

	TRAIN 70/TEST 30		TRAIN 80/TEST 20	
	F-measure	*ROC AUC*	*F-measure*	*ROC AUC*
Umbral Estándar	0.714	0.919	0.625	0.895
Umbral Óptimo	0.824	0.919	0.8	0.895

Tabla 5 *Comparación de los resultados obtenidos para la Red Neuronal Convolucional mediante el umbral estándar y el umbral óptimo*

Comparación con Umbral Óptimo

	TRAIN/TEST		CROSS-VALIDATION	
	F-measure	*ROC AUC*	*F-measure*	*ROC AUC*
SemiExonet	0.824	0.919	0.768	0.908

Tabla 6 *Comparación de resultados obtenidos en las pruebasTrain/Test y Cross-Validationpara la red convolucional SemiExonet*

Resultados y Discusión–2ªFase(III)

- **Estructura Semi-ExoNet con modificación de umbral**

→Mejor resultado: Umbral Óptimo y Cross-Validation

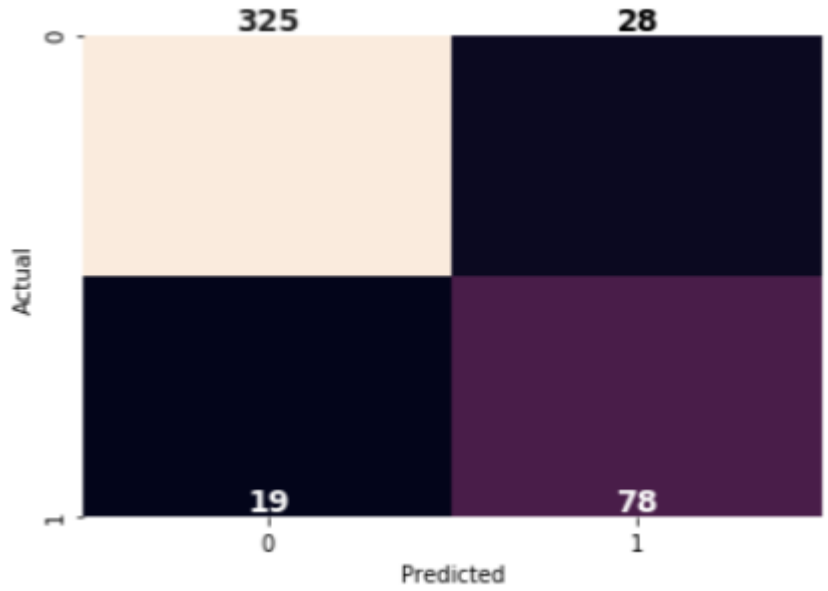


Figura 16 *Matriz de confusión obtenida para la CNN realizada mediante Cross-Validation y valor óptim odel umbral*

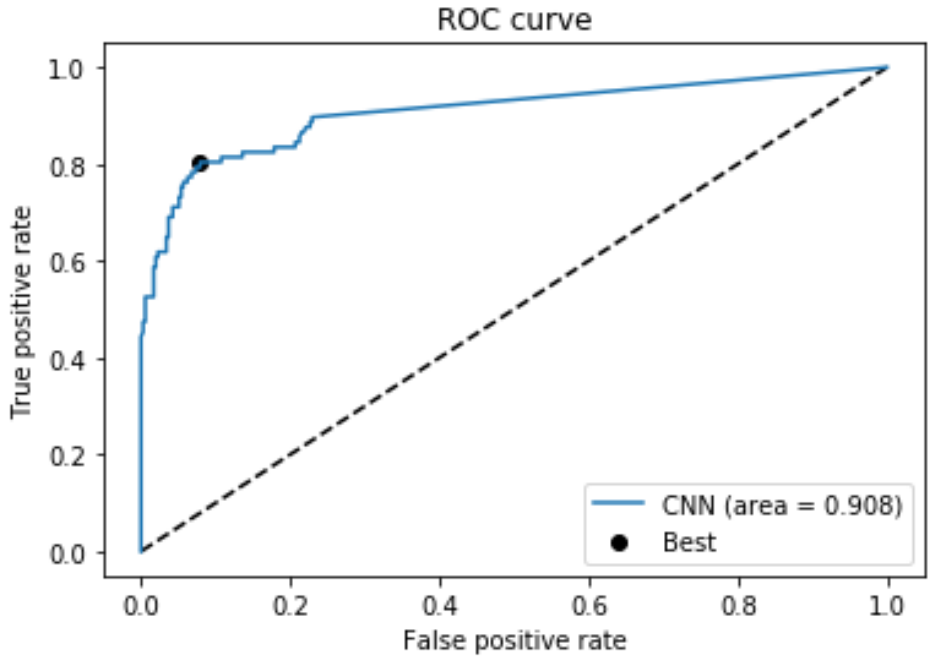


Figura 17 *Curva ROC obtenida para la CNN realizada medianteCross-Validation. El punto señalado en el gráfico muestra el valor óptimodel umbra*

Resultados y Discusión–2ªFase(IV)

- **Nuevas configuraciones:**

Modificación de la Función de Activación

Sigmoidea / Tanh

Modificación del Algoritmo de Optimización

Adam / AdaDelta

	Sigmoidea		Tanh	
	Adam	*Adadelta*	*Adam*	*Adadelta*
F-measure	0.768	0.818	0.866	0.878
ROC AUC	0.908	0.939	0.971	0.964

Tabla 7 *Resumen de los resultados obtenidos para las diferentes pruebas llevadas a cabo en la red convolucional*

Resultados y Discusión–2ªFase(IV)

- **Nuevas configuraciones:**

→Mejor resultado :

Tanh /Adadelta

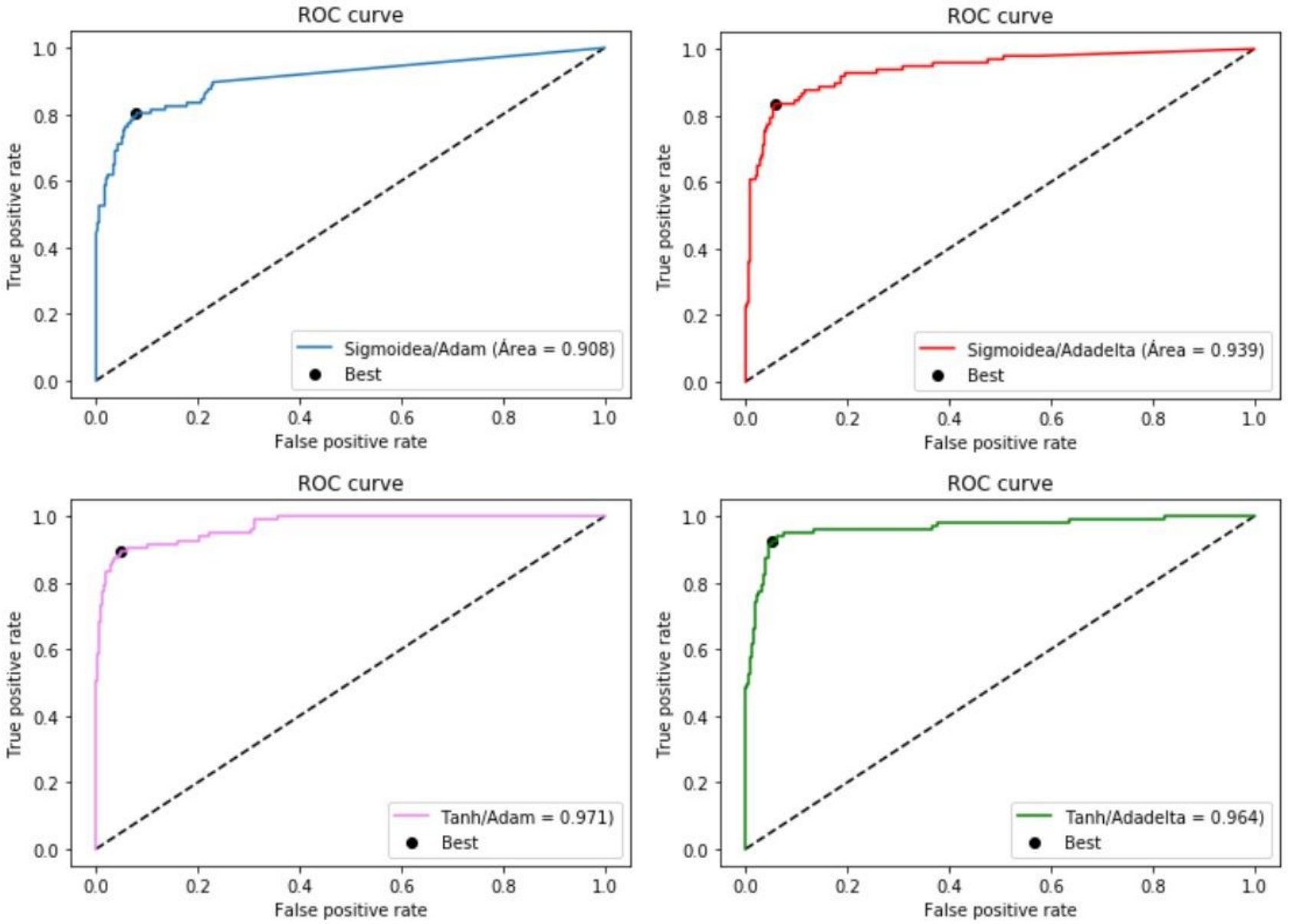


Figura 18 *Comparación de las 4 curvas ROC para las diferentes configuraciones de parámetros*

Resultados y Discusión–2ªFase(V)

- **Comparación de Inputs:**

	INPUTS		
	LocalView+Parámetros	GlobalView+Parámetros	LocalView+GlobalView
F-measure	0.887	0.545	0.753
ROC AUC	0.954	0.718	0.903

Tabla 8 *Resumen de los resultados obtenidos al combinar los diferentes inputs, llevadas a cabo en la red convolucional*

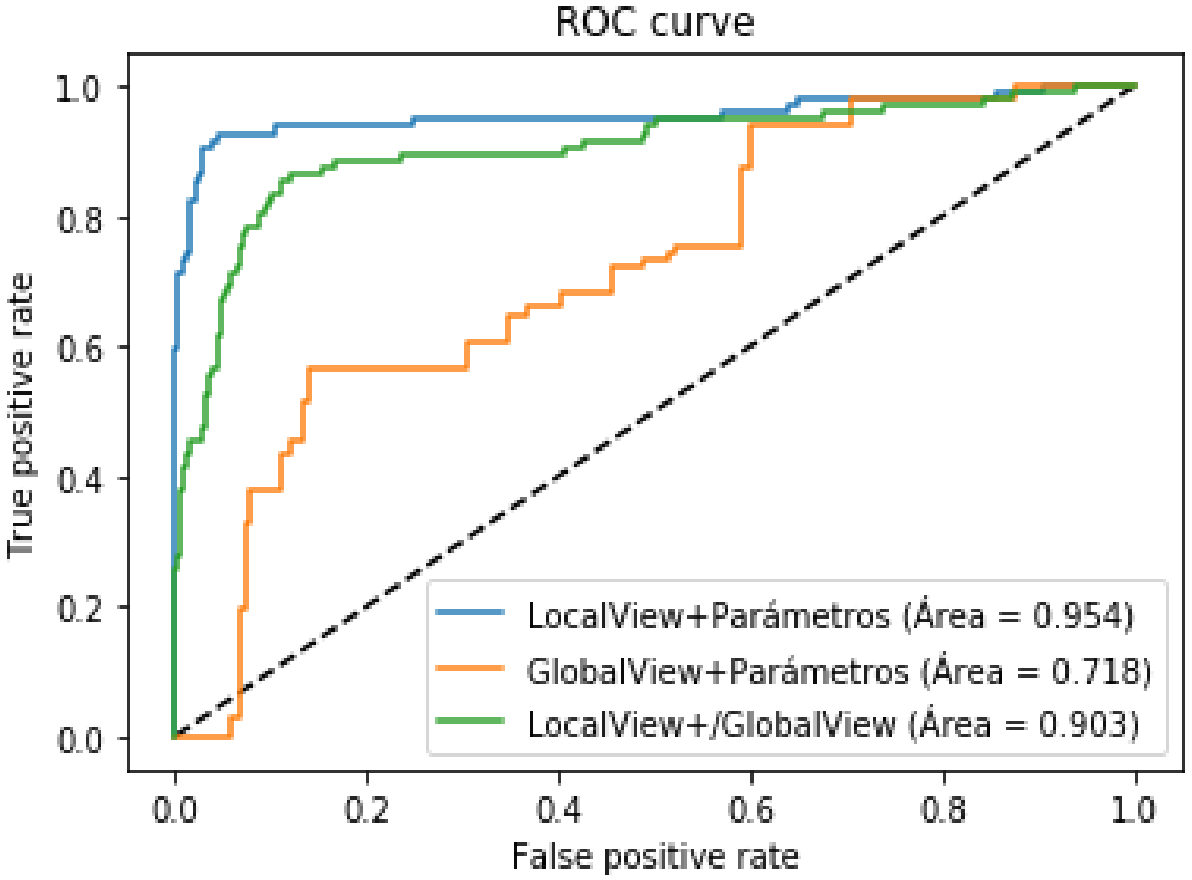


Figura 19 *Comparación de las 3 curvas ROC para las diferentes combinaciones de inputs*

Resultados y Discusión–Otros trabajos(I)

Shallue 2018

- En Shallue no utilizan Parámetros Estelares
- Mejores resultados generales en Shallue 2018→ Datos de Kepler y registros negativos
- Resultados similares para los Inputs Local View+ Parámetros Estelares

	Accuracy		
	Local/Local+Param	Global/Global+Param	Local+Global
TFM	0.842/0.917	0.642/0.8	0.878
Shallue,2018	0.924/-	0.954/-	0.960

Tabla 9 *Resumen de los resultados obtenidos al combinar los diferentes inputs, llevadas a cabo en la red convolucional*

Osborn 2019

- Mejor resultado de *Average Precision* en Osborn 2019→ Tenemos un dataset que no esta balanceado
- *Average Accuracy similares*
- *Recall Accuracy levemente mejor→ Nuestra arquitectura detecta más exoplanetas*

	Parámetros		
	Average Accuracy	*Average Recall*	*Average Precision*
TFM	0.944	0.939	0.789
Osborn,2019	0.946	0.932	0.973

Tabla 10 *Resumen de los resultados obtenidos al combinar los diferentes inputs, llevadas a cabo en la red convolucional*

Conclusiones

La utilización de la red convolucional mejora notablemente el resultado:

Las redes convolucionales ofrecen gran rendimiento al gestionar y reducir grandes volúmenes de datos perdiendo la mínima cantidad de información.

Los parámetros estelares no son determinantes en la red convolucional:

Al contrario que en las redes recurrentes, al procesar conjuntamente las dos curvas de luz los resultados no son lo bastante buenos como para que los parámetros elegidos no ofrezcan una mejora significativa.

El dataset con el que se alcanzan resultados más veraces es el Local View:

Esto resulta especialmente interesante.
Al centrarnos en positivos y falsos positivos las curva de luz son muy similares, por tanto un menor número de valores permiten descubrir con mayor facilidad las variaciones de las curvas de luz.

Trabajo Futuro

- Obtener un mayor número de muestras: Datos de otras misiones, datos simulados o esperar nuevos datos de TESS
- Aplicar más conocimiento experto tanto en el preproceso como las dimensiones introducidas en la red
- Profundizar en el aspecto matemático
- Desarrollo de nuevas arquitectura

Muchas gracias por su atención