



DISCOVERY AND CHARACTERISATION OF EXTRASOLAR PLANETS

boosted by machine learning algorithms

Rodrigo F. Díaz

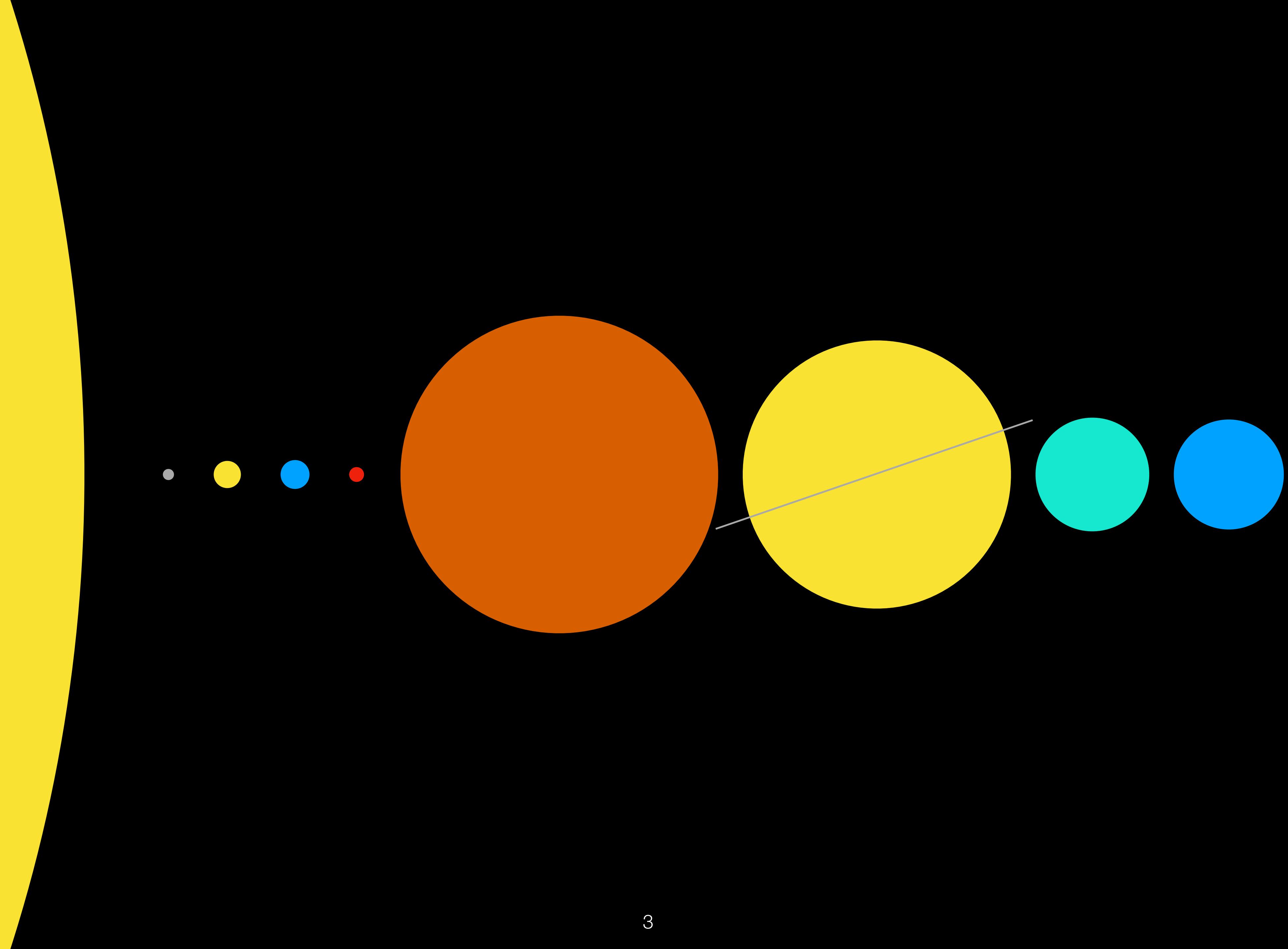
ICAS; Instituto de Ciencias Físicas (CONICET / UNSAM)

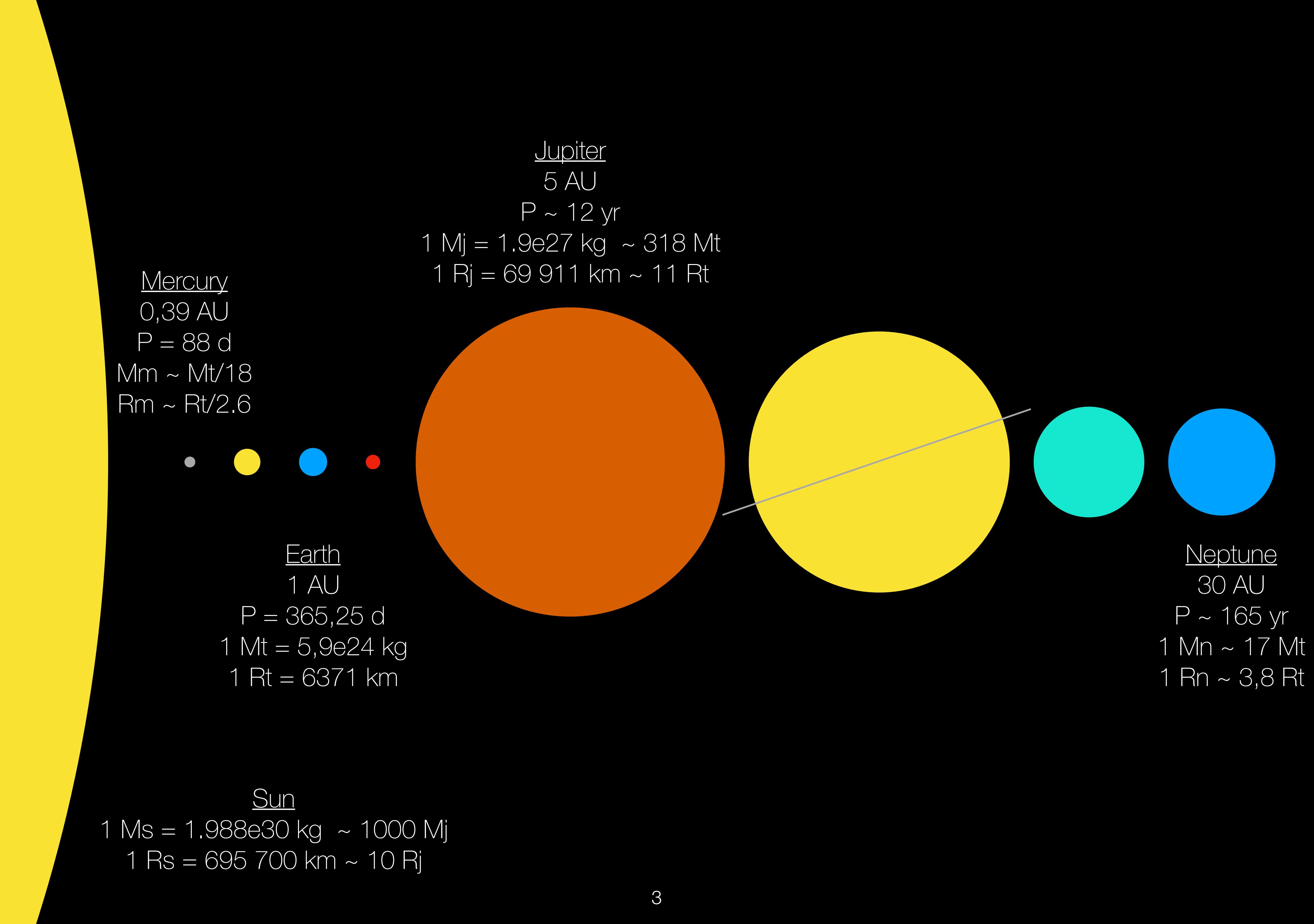
rdiaz@unsam.edu.ar
@RDeextrasolar

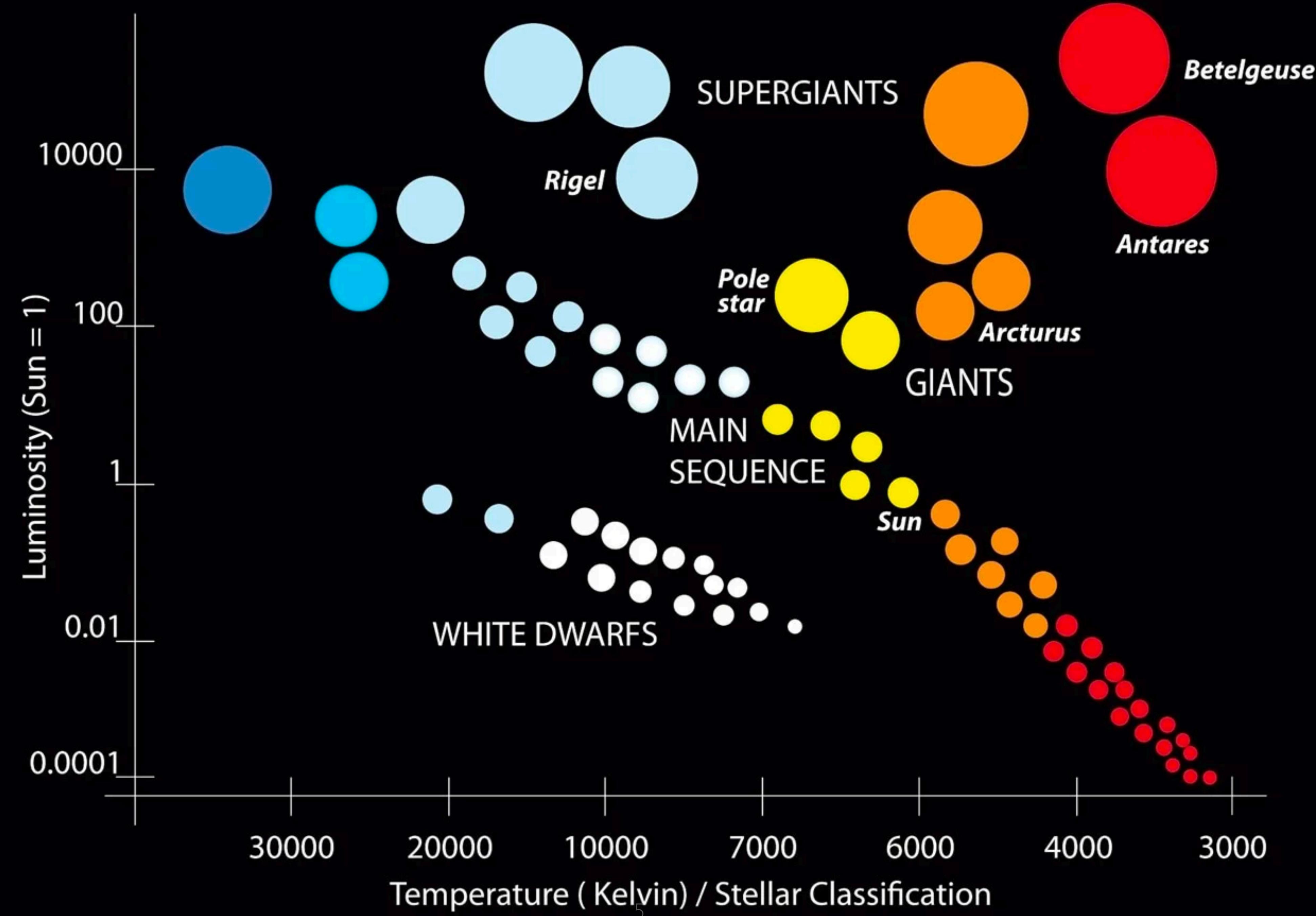


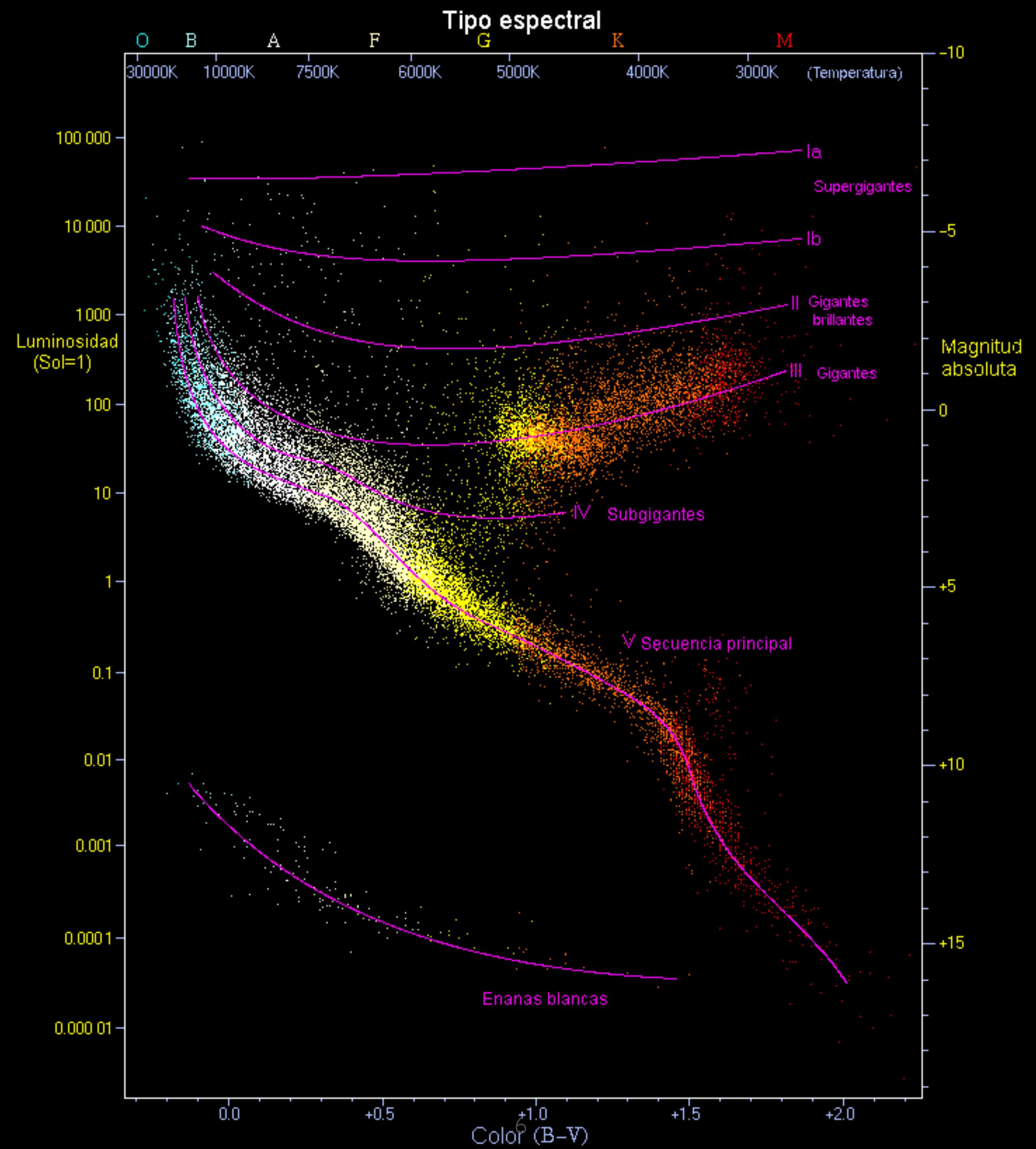
OUTLINE

- Overview of exoplanets
- Current limitations
- Our contribution using machine learning
- Bonus track: the advent of JWST









Some questions

What kind of planets are possible?

How common are planetary systems?

How common is the Solar System? And Earth?

What are planets made of?

What are their atmospheres like?

What can we learn about their formation and evolution by studying the architecture of systems?

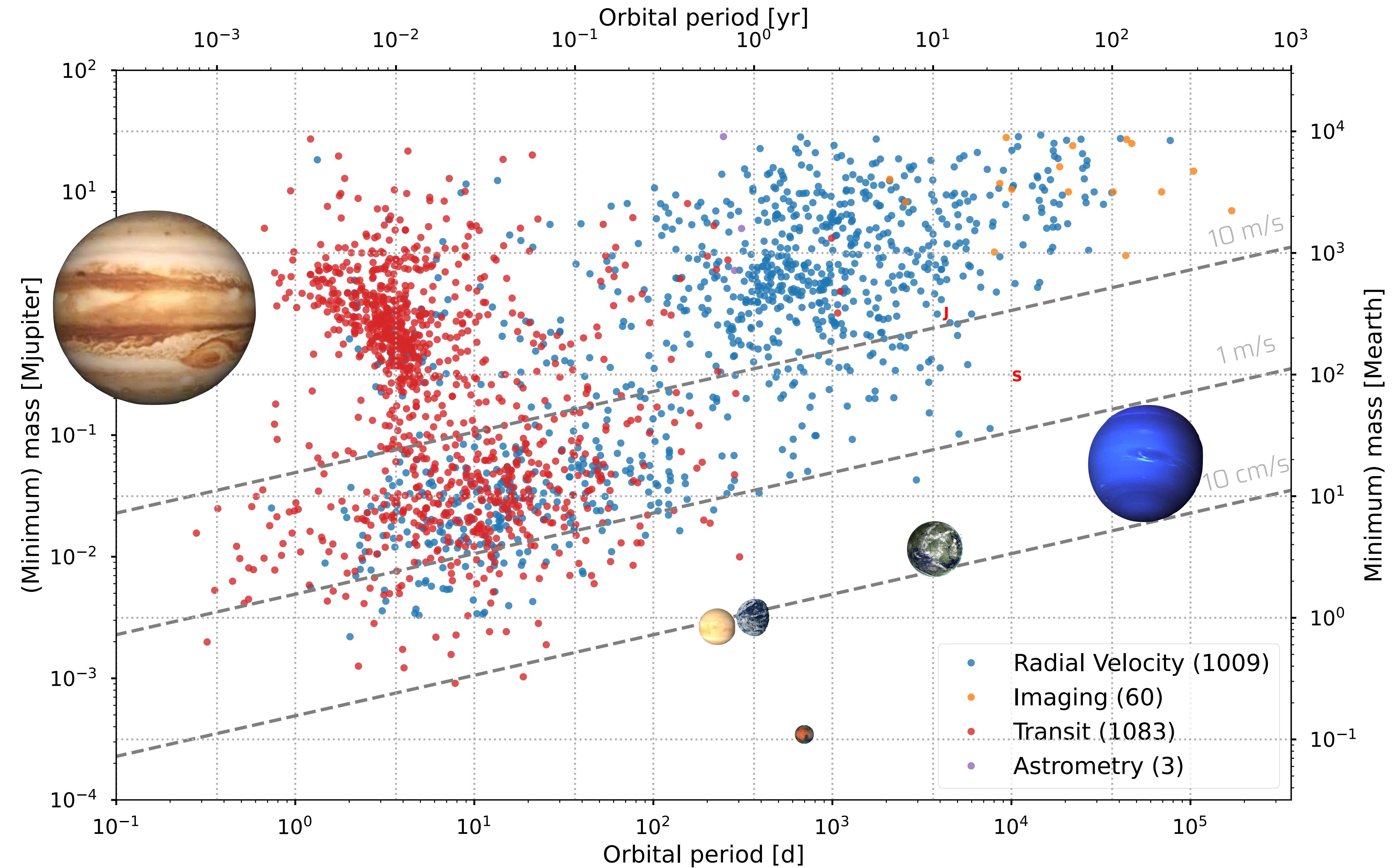
How do formation and evolution depend on stellar parameters?

...

Are we alone?

5539
known exoplanets

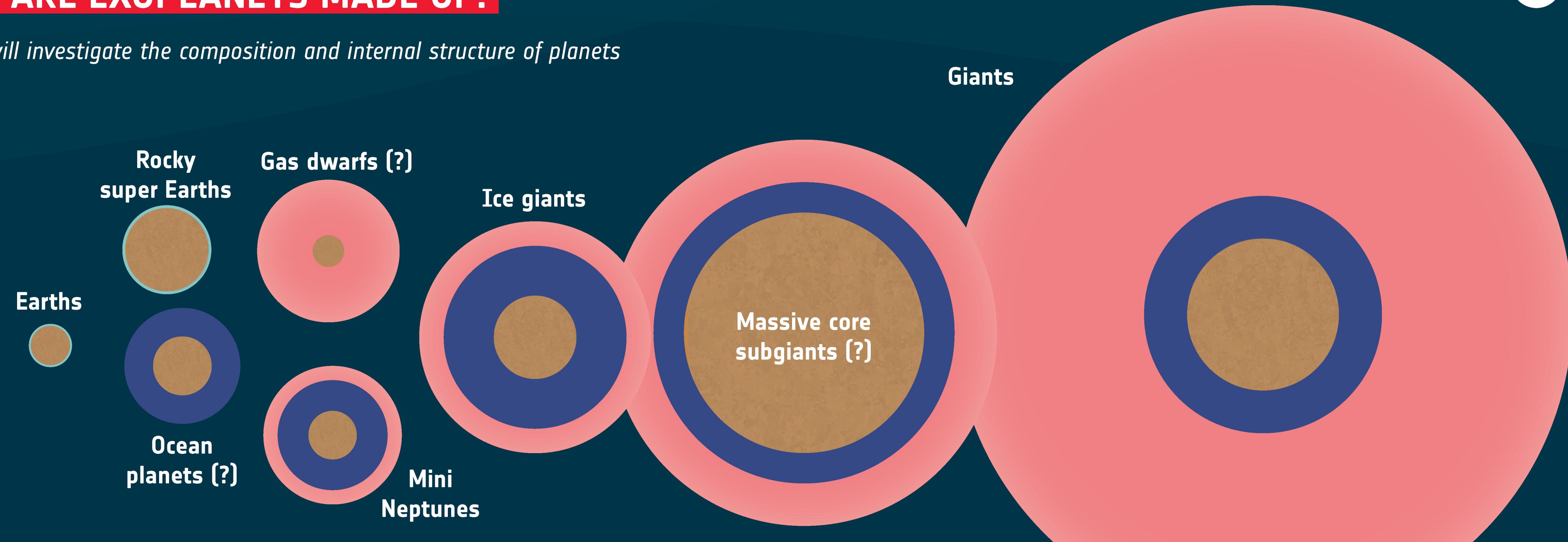
according to [NASA Exoplanet Archive](#) (on April 25 2021)



DIVERSE INTERNAL STRUCTURE

→ WHAT ARE EXOPLANETS MADE OF?

How Cheops will investigate the composition and internal structure of planets



- Hydrogen / helium envelope
- Thin atmosphere
- Ice mantle / volatile* envelope
- Solid core (rocks, metals)

* Planetary scientists call **volatiles** all chemical elements and compounds with low boiling points that are associated with a planet's or moon's crust or atmosphere. These include: nitrogen, water, carbon dioxide, ammonia, hydrogen, methane and sulphur dioxide.

Super Earths

1 M_{Earth}

Neptunes

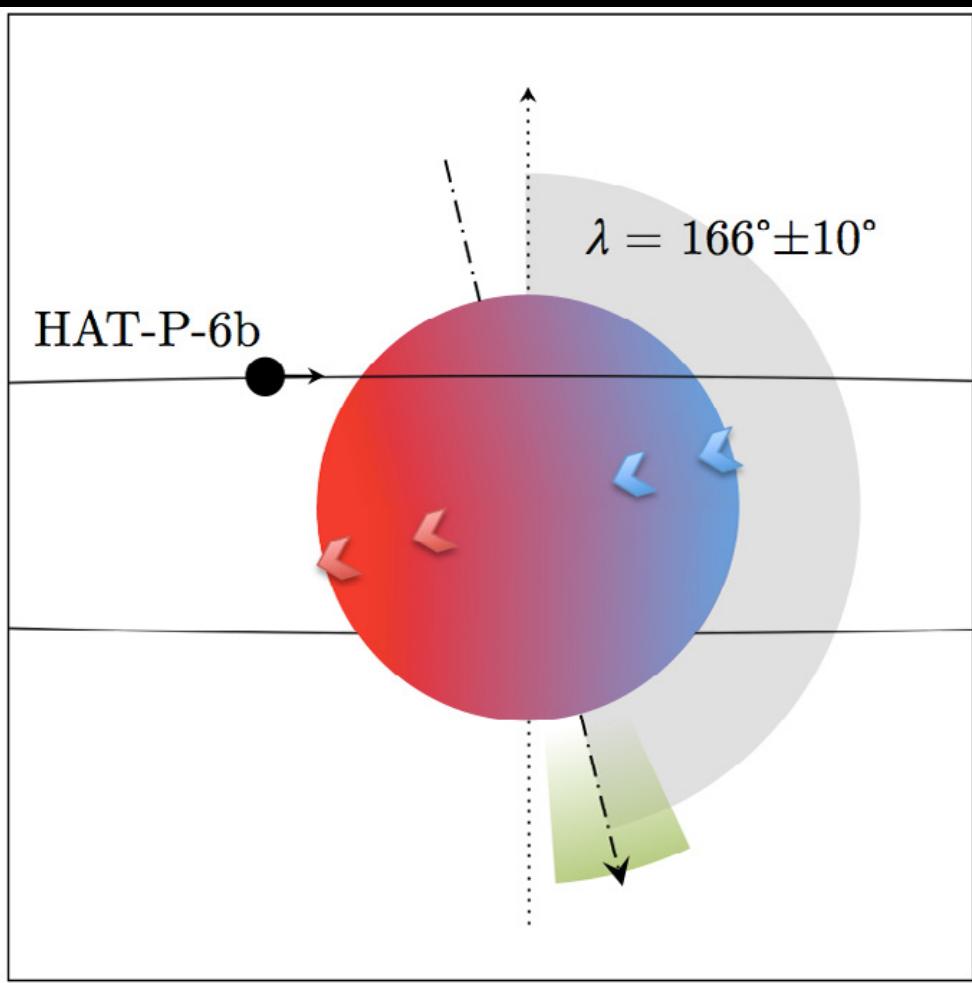
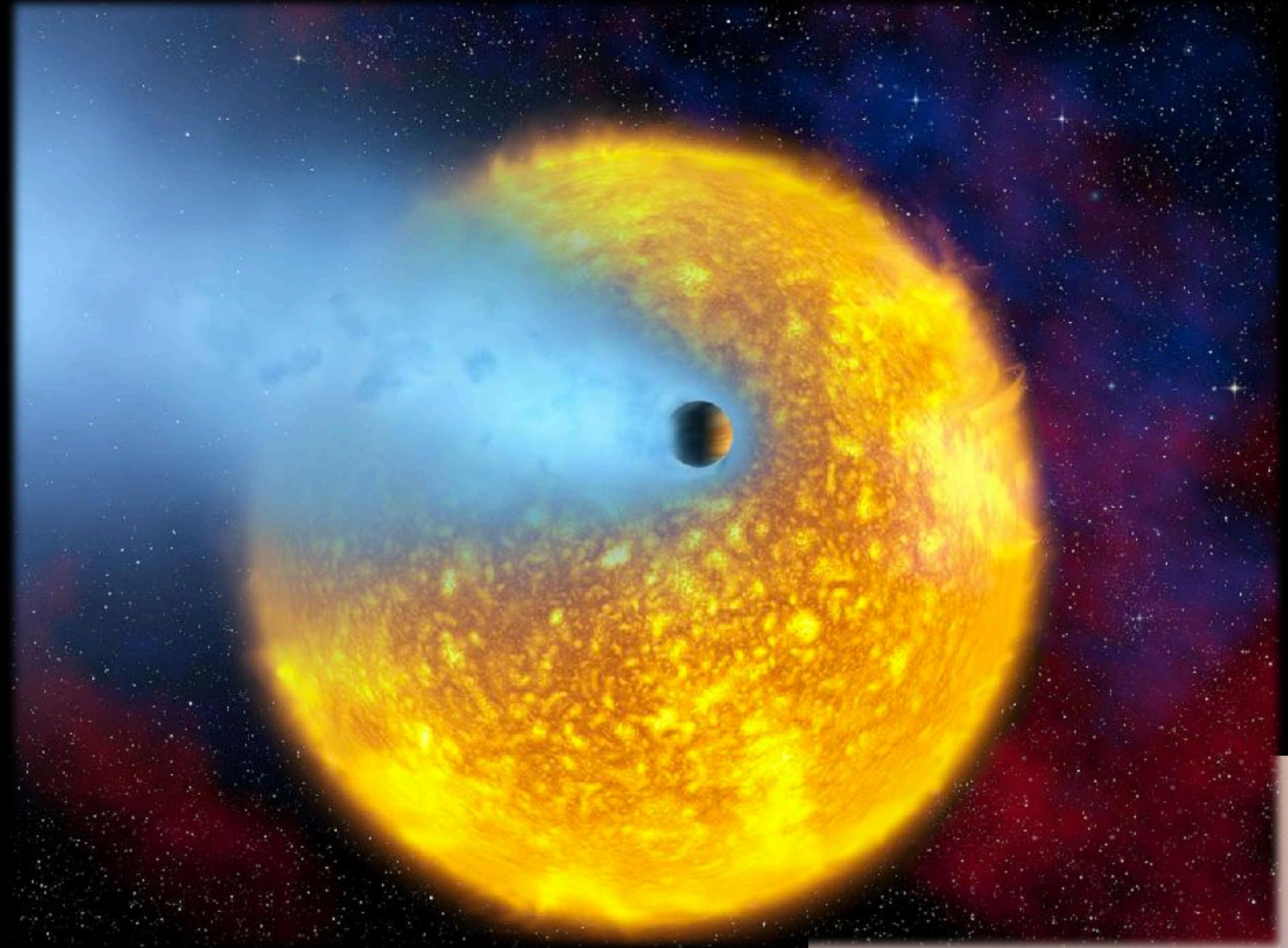
10 M_{Earth}

Jupiters

300 M_{Earth} ~ 1 M_{Jupiter}

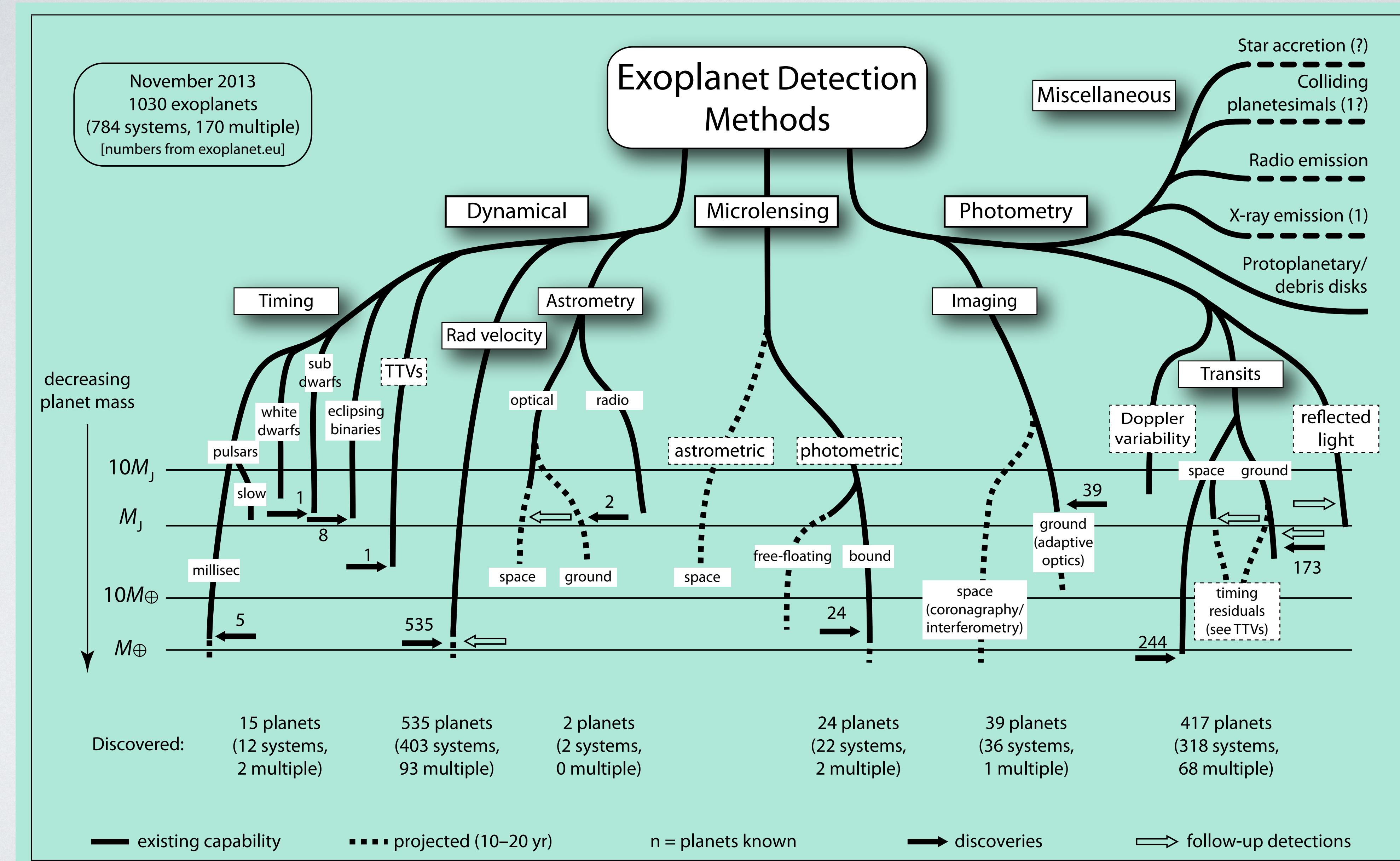
1000 M_{Earth}

(M=mass)

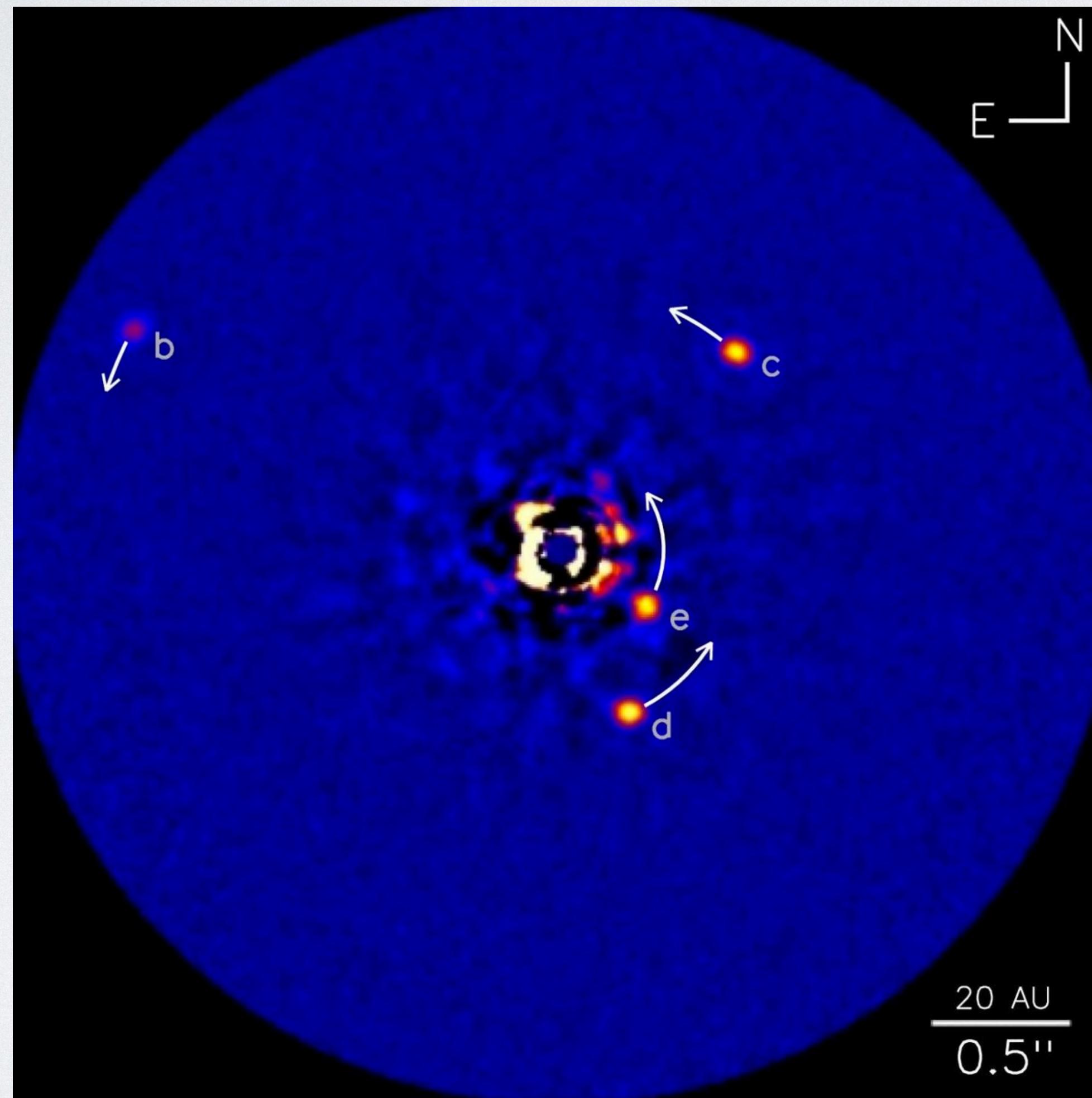


Hébrard, Ehrenreich, Bouchy, et al. (2011)



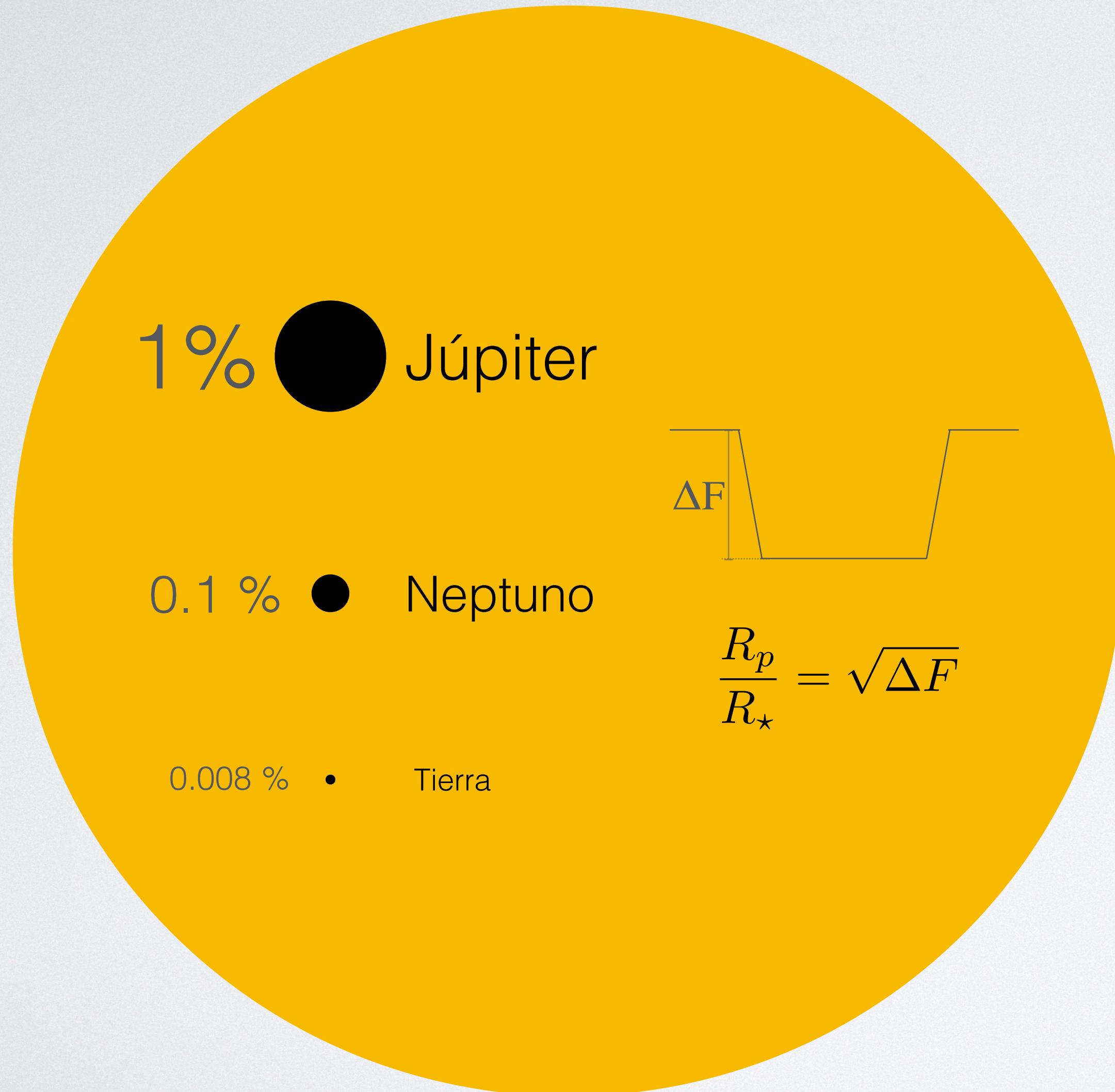


DIRECT IMAGING



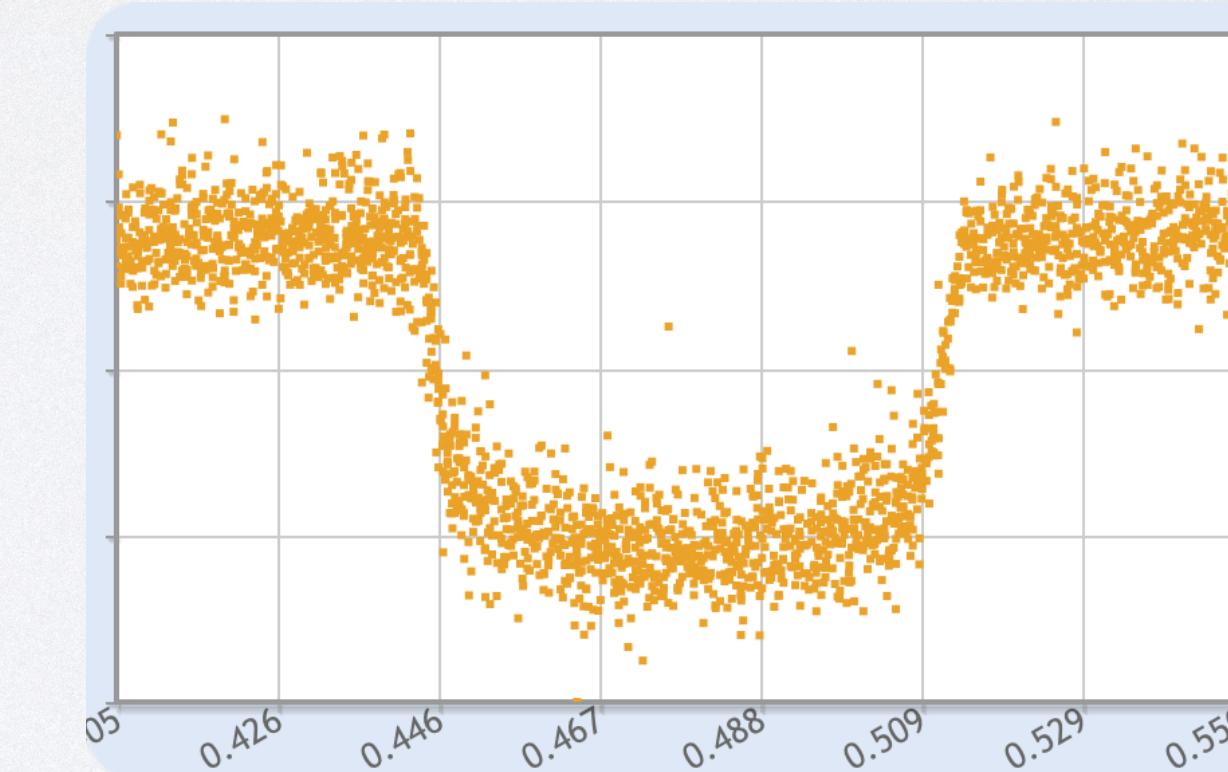
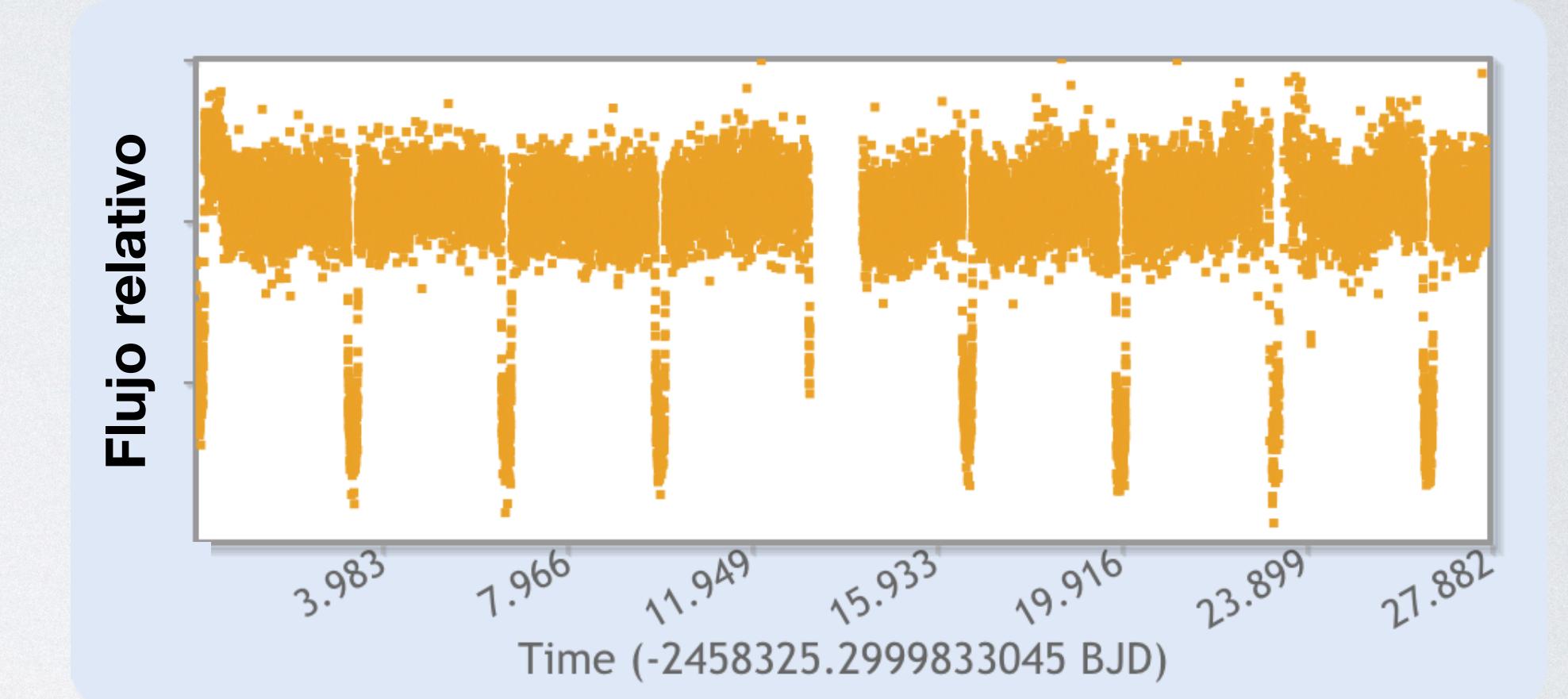
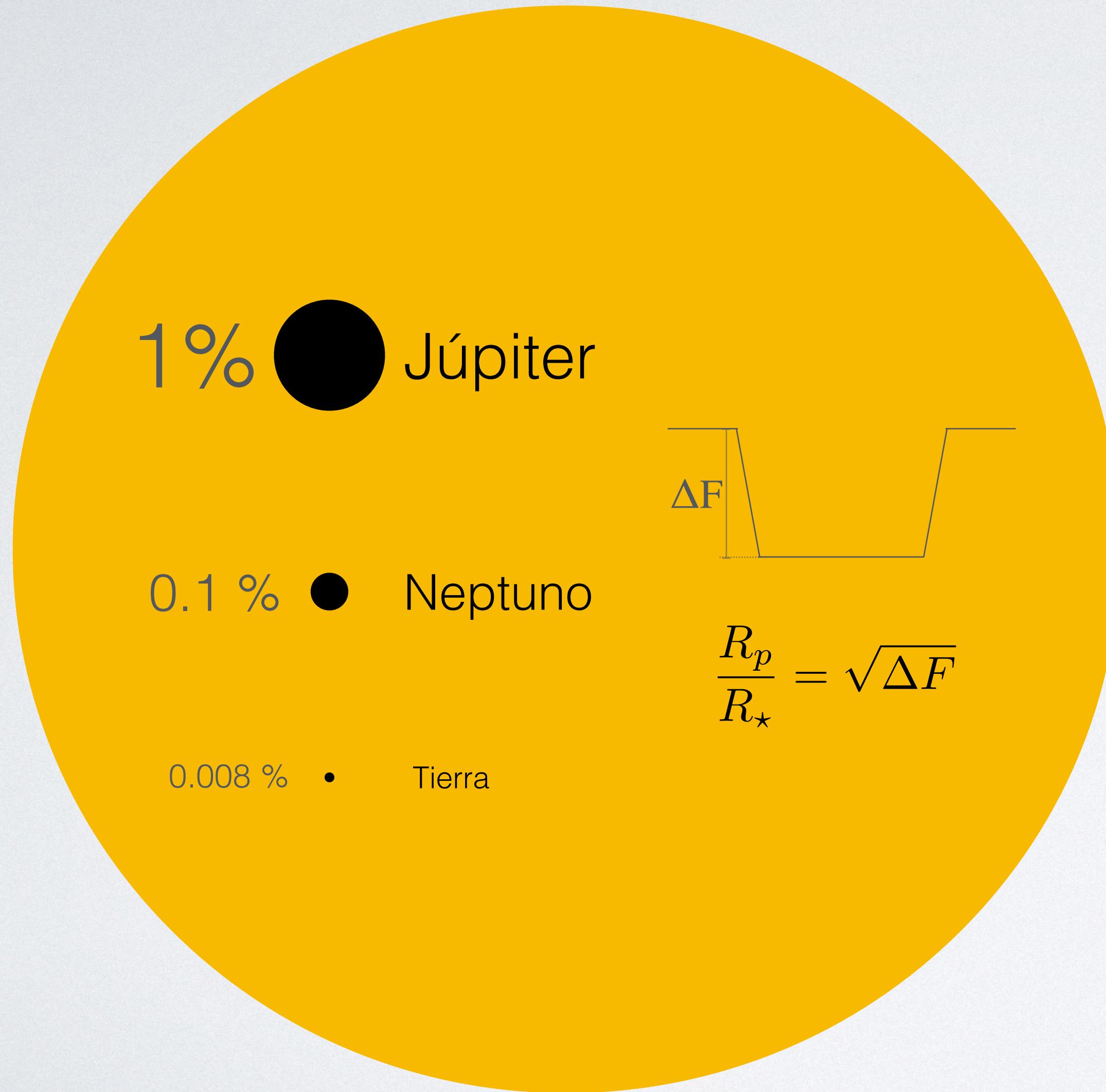
$$\frac{f_{\oplus}}{f_{\odot}} = 10^{-10}$$

TRANSIT METHOD

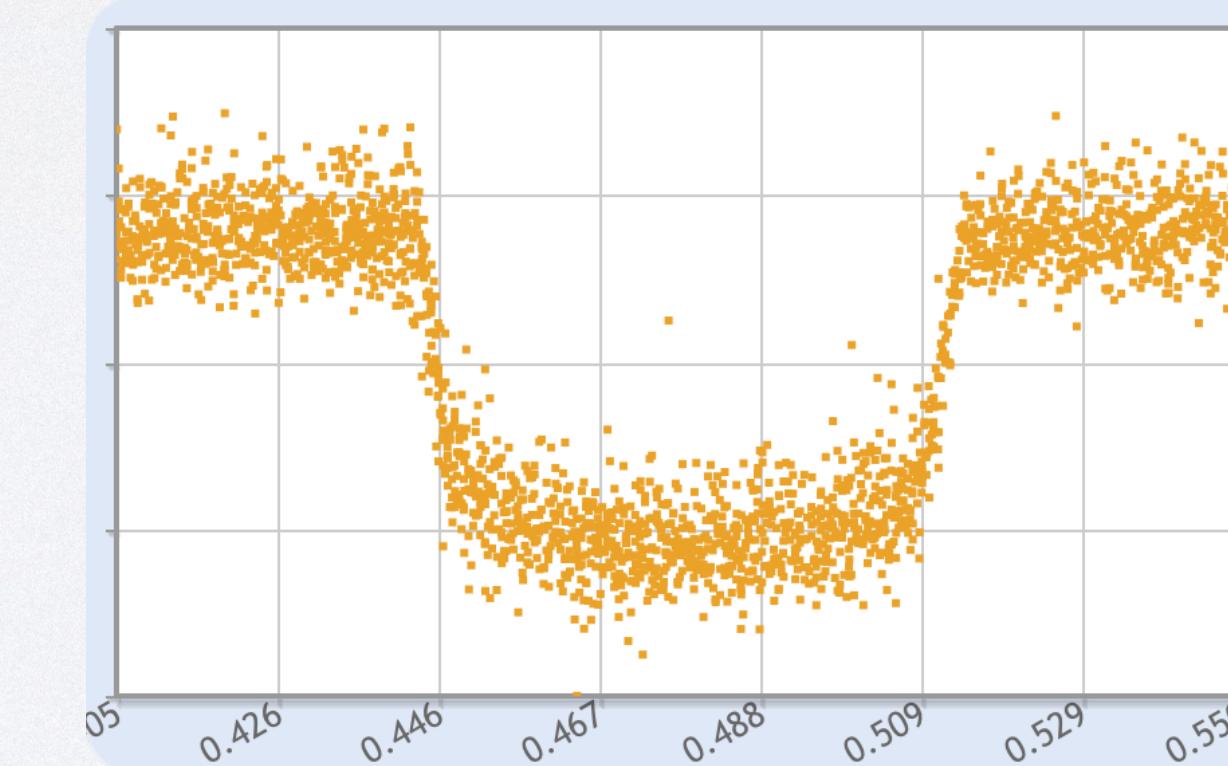
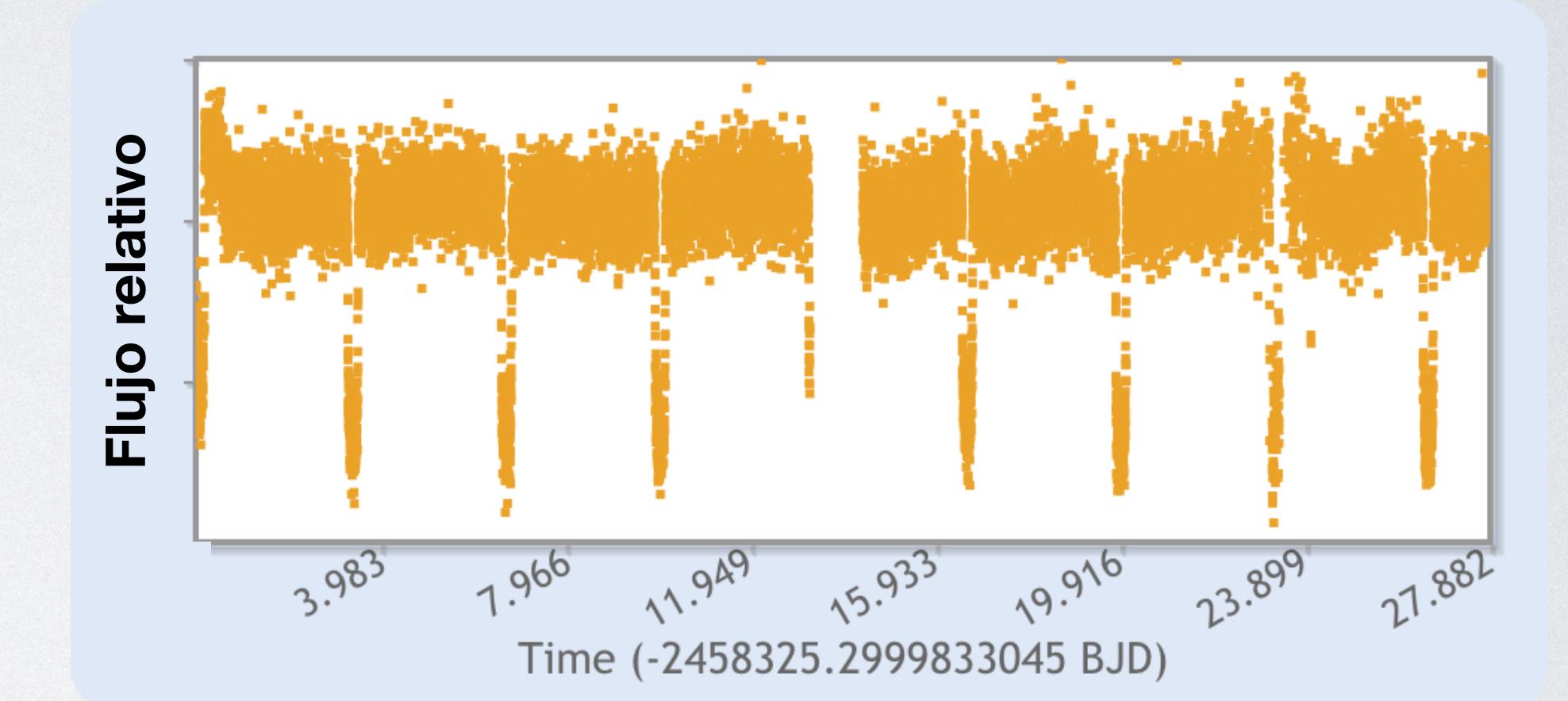
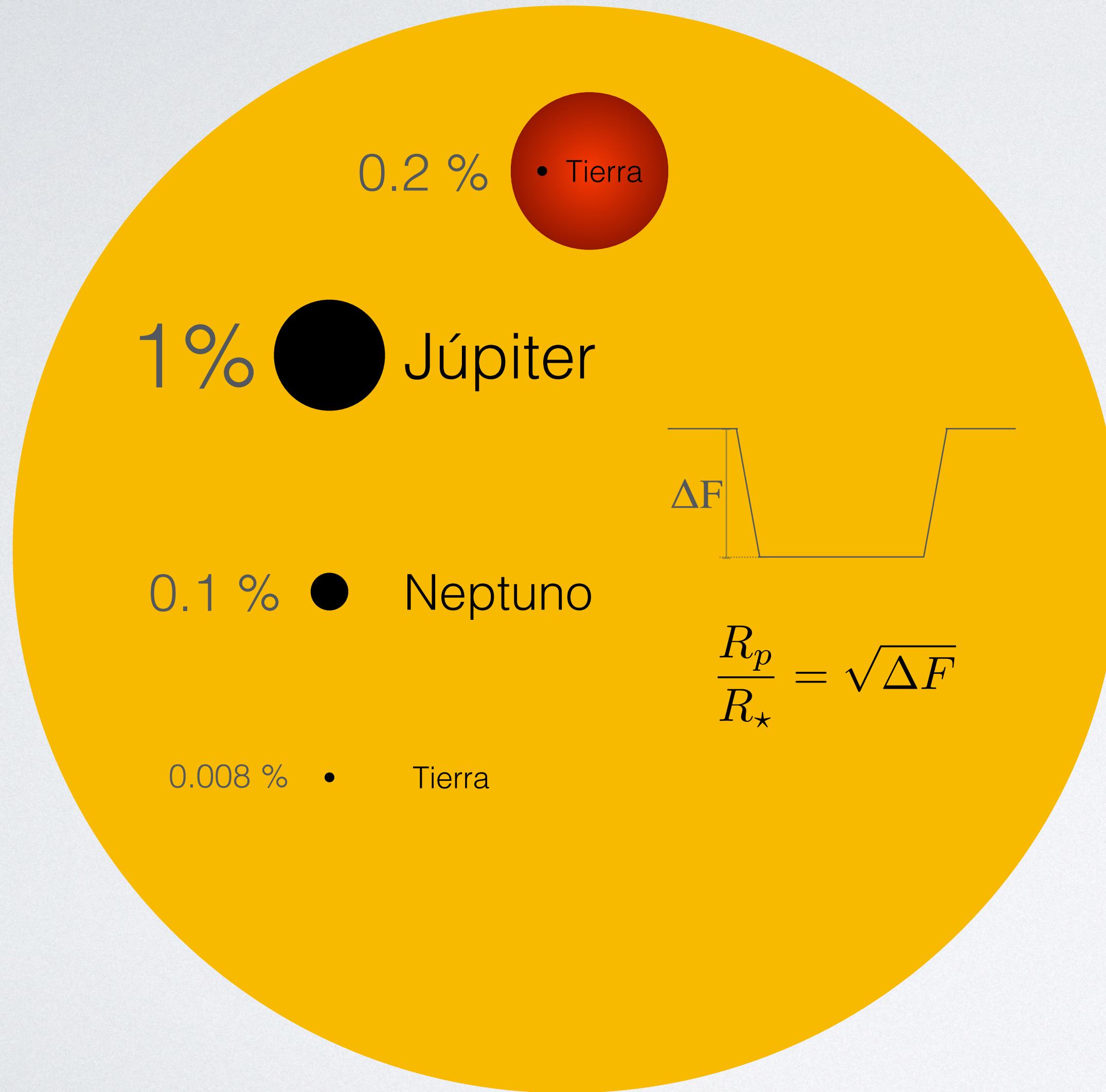


$$\frac{R_p}{R_\star} = \sqrt{\Delta F}$$

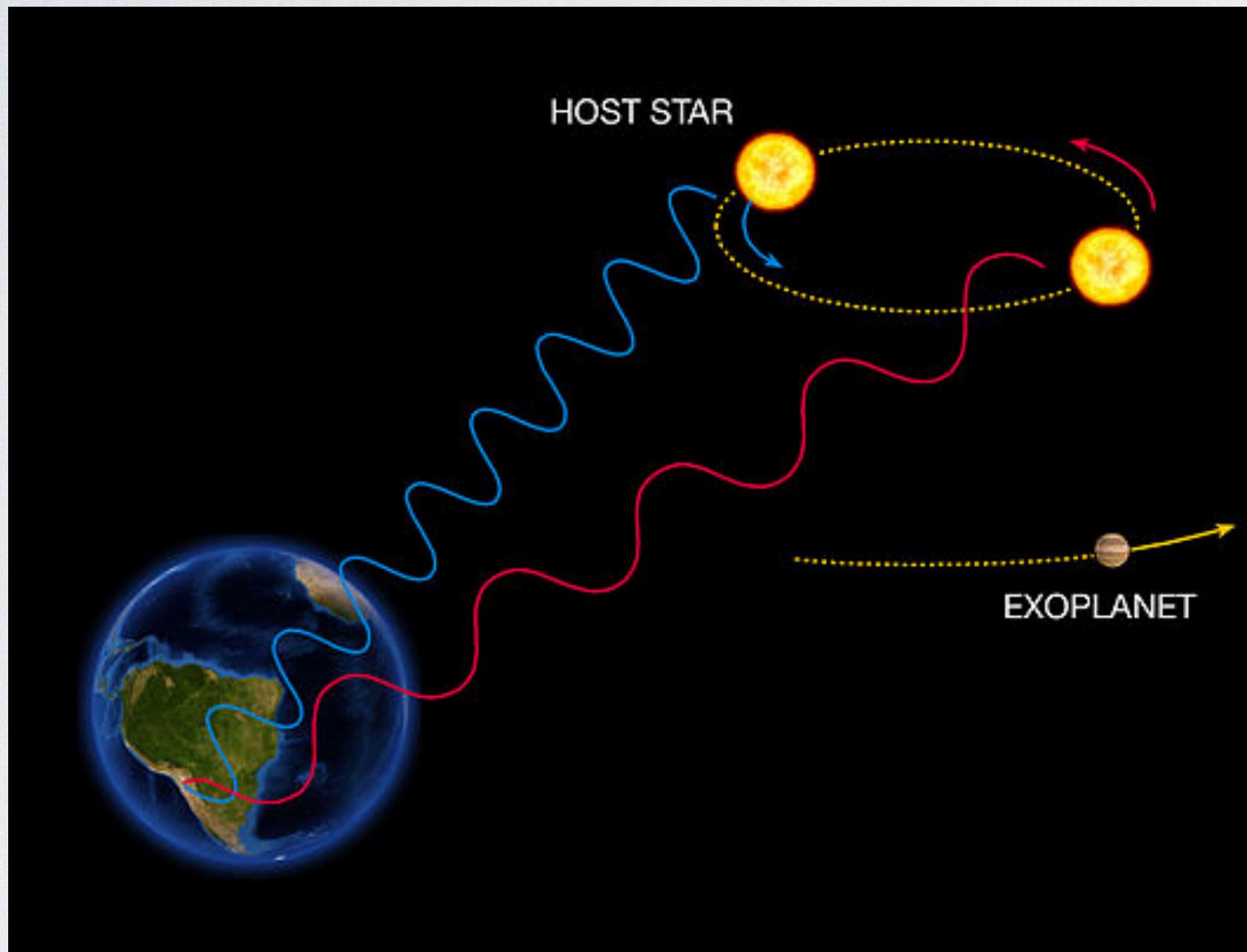
TRANSIT METHOD



TRANSIT METHOD

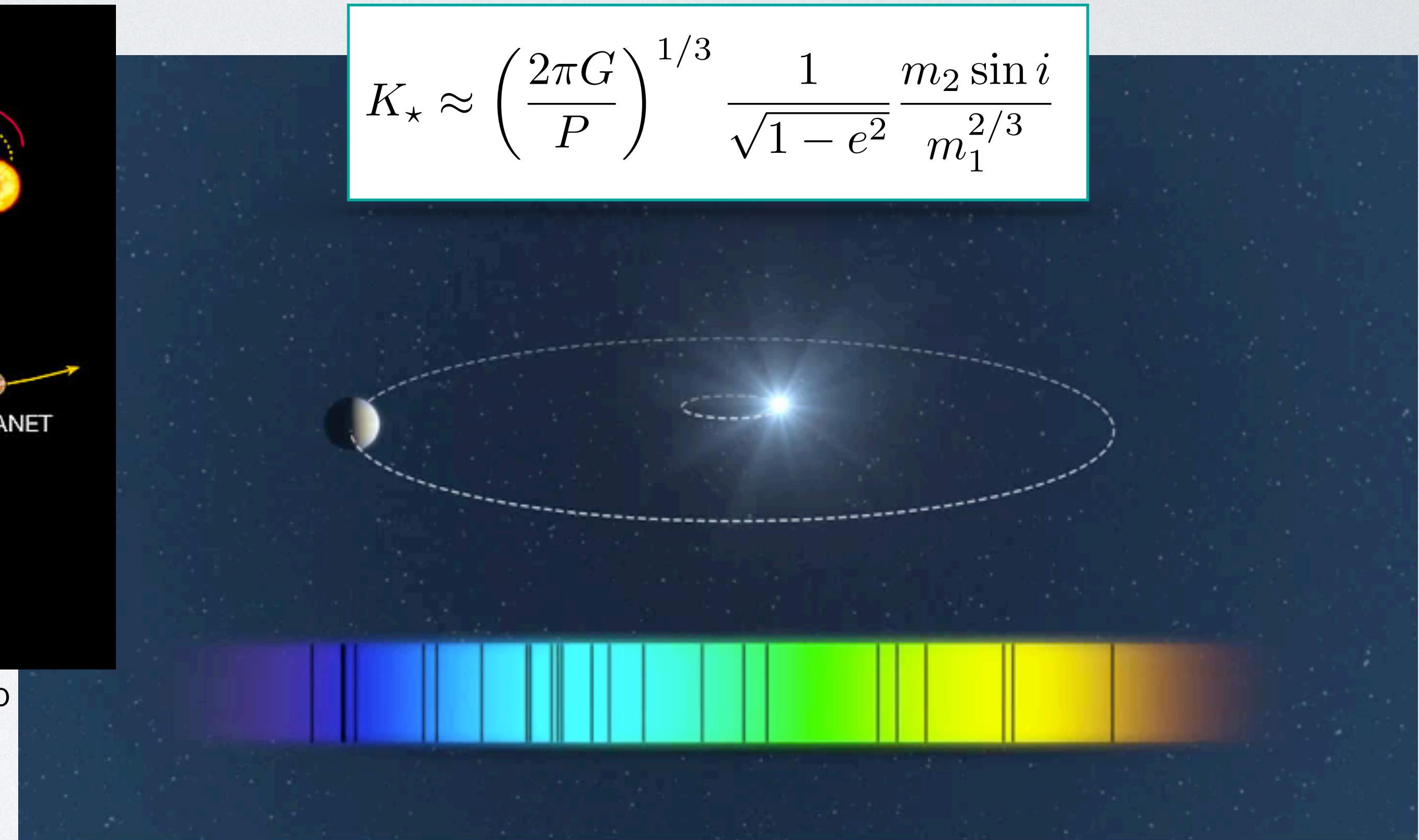


RADIAL VELOCITY METHOD

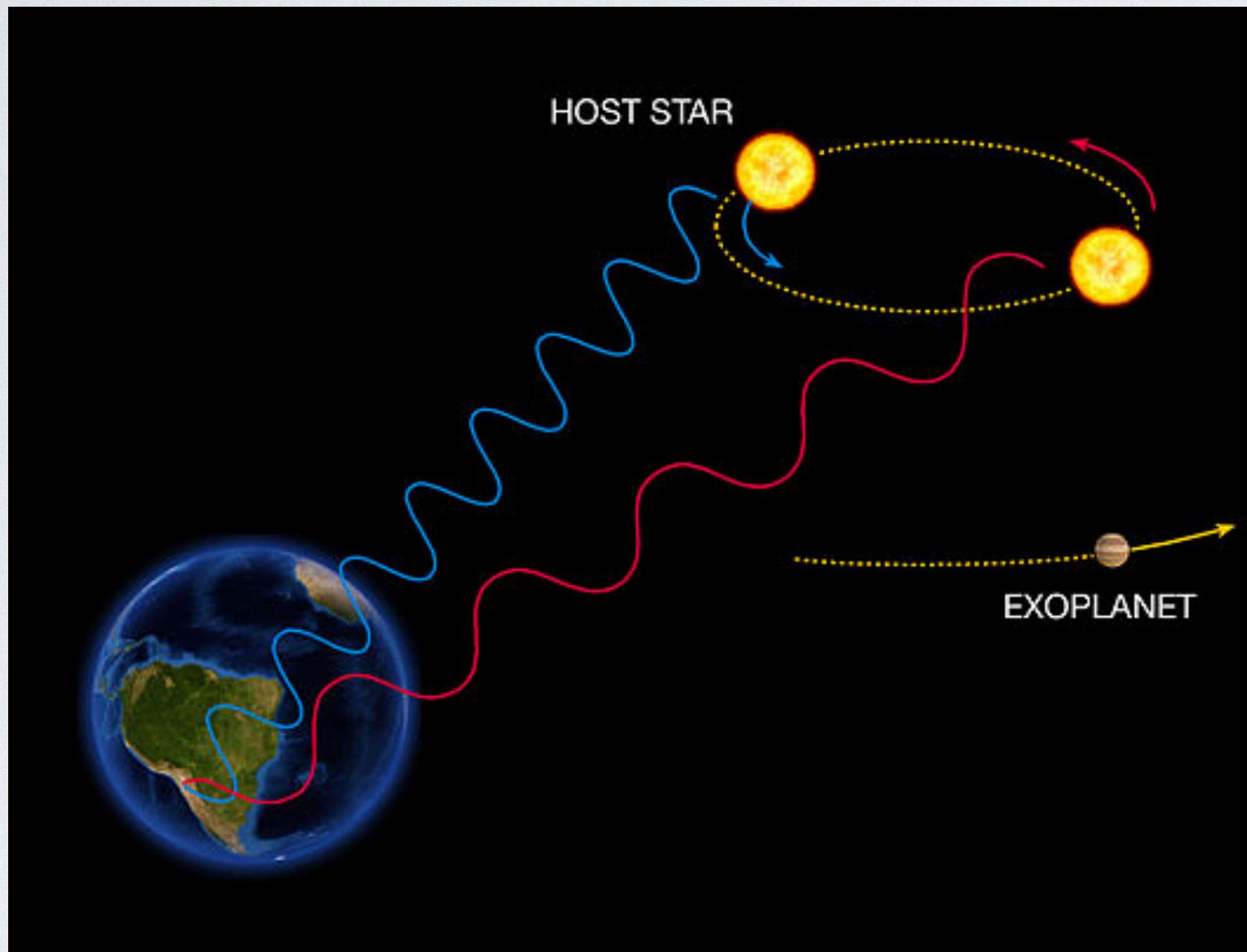


Credit: ESO

$$K_{\star} \approx \left(\frac{2\pi G}{P} \right)^{1/3} \frac{1}{\sqrt{1 - e^2}} \frac{m_2 \sin i}{m_1^{2/3}}$$

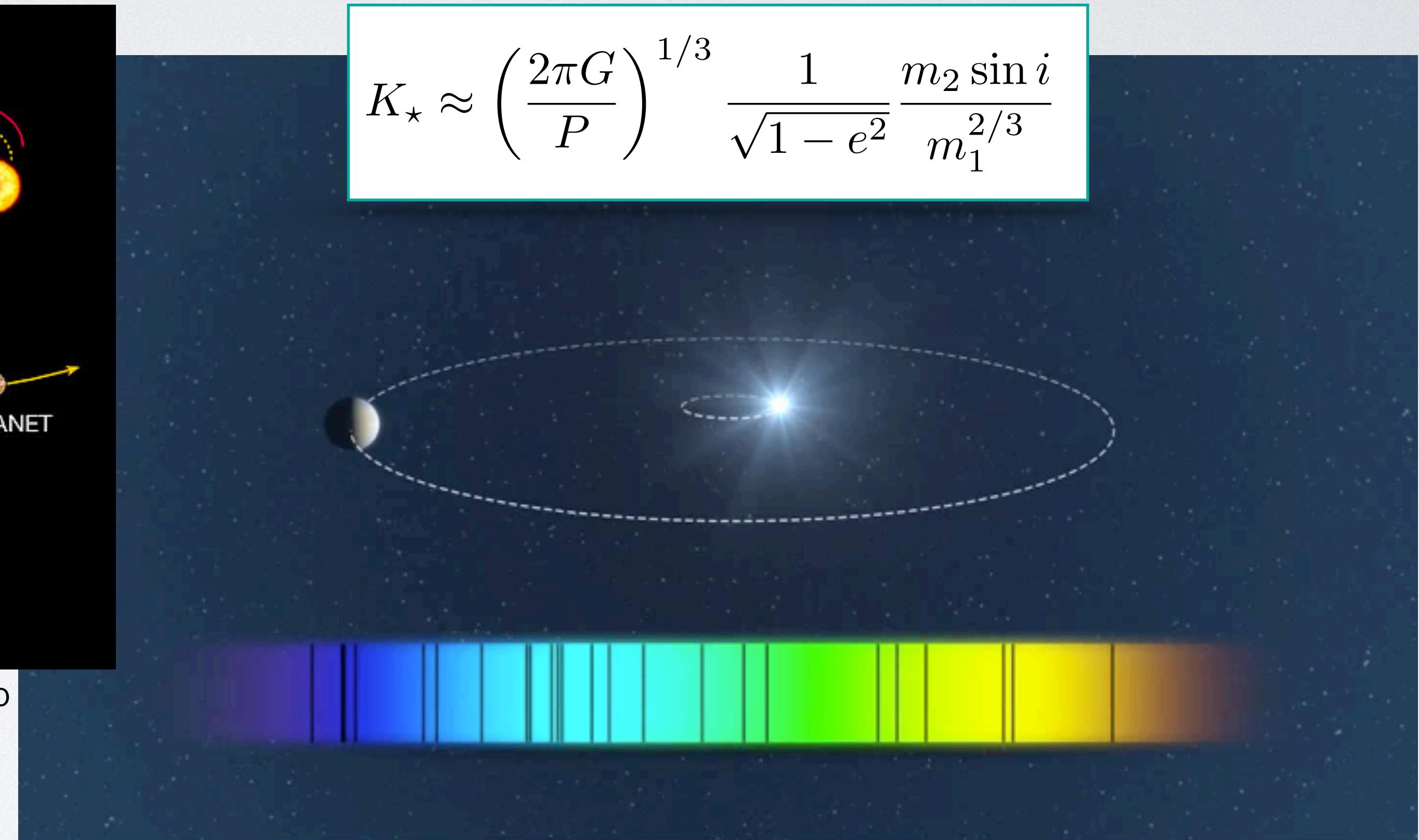


RADIAL VELOCITY METHOD

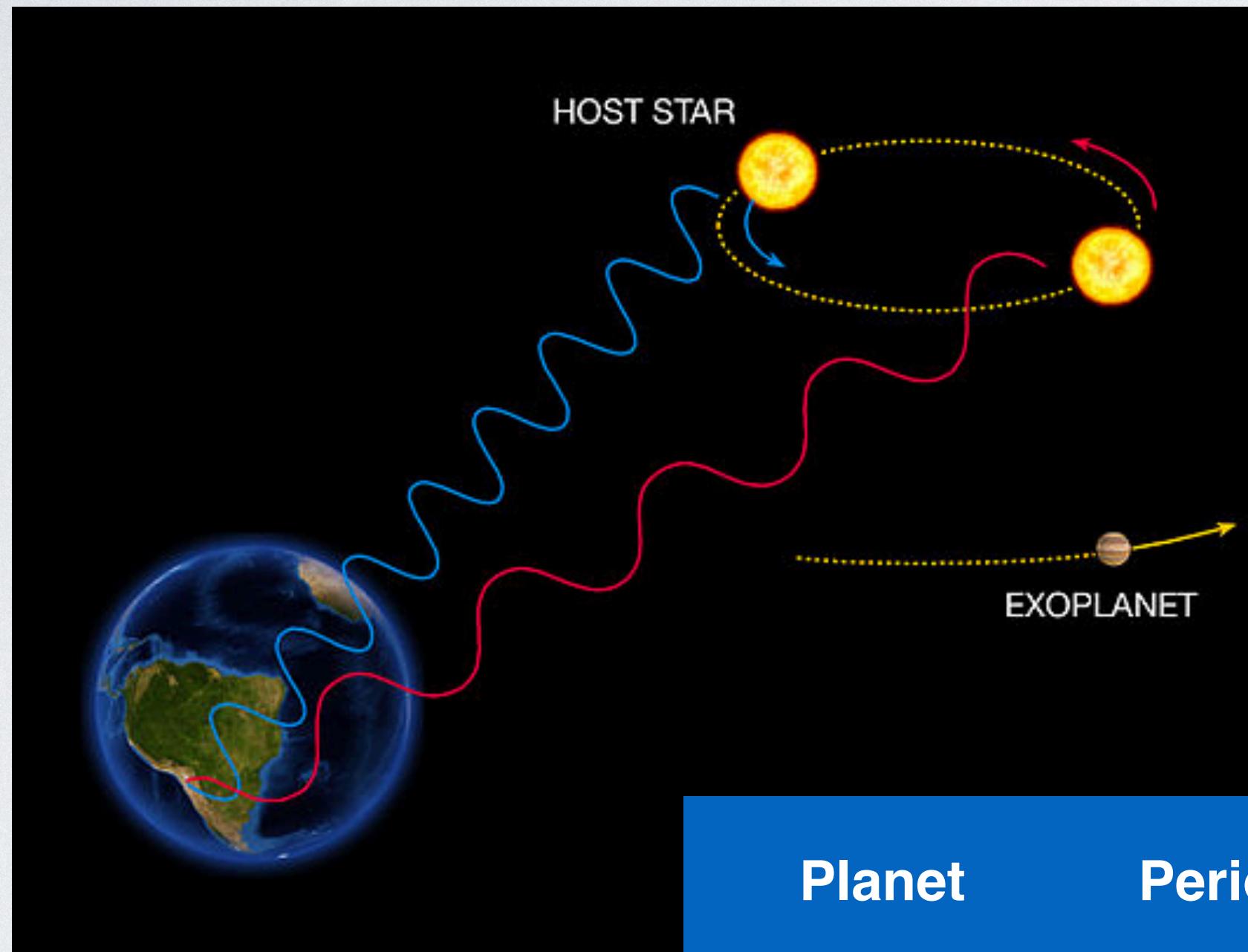


Credit: ESO

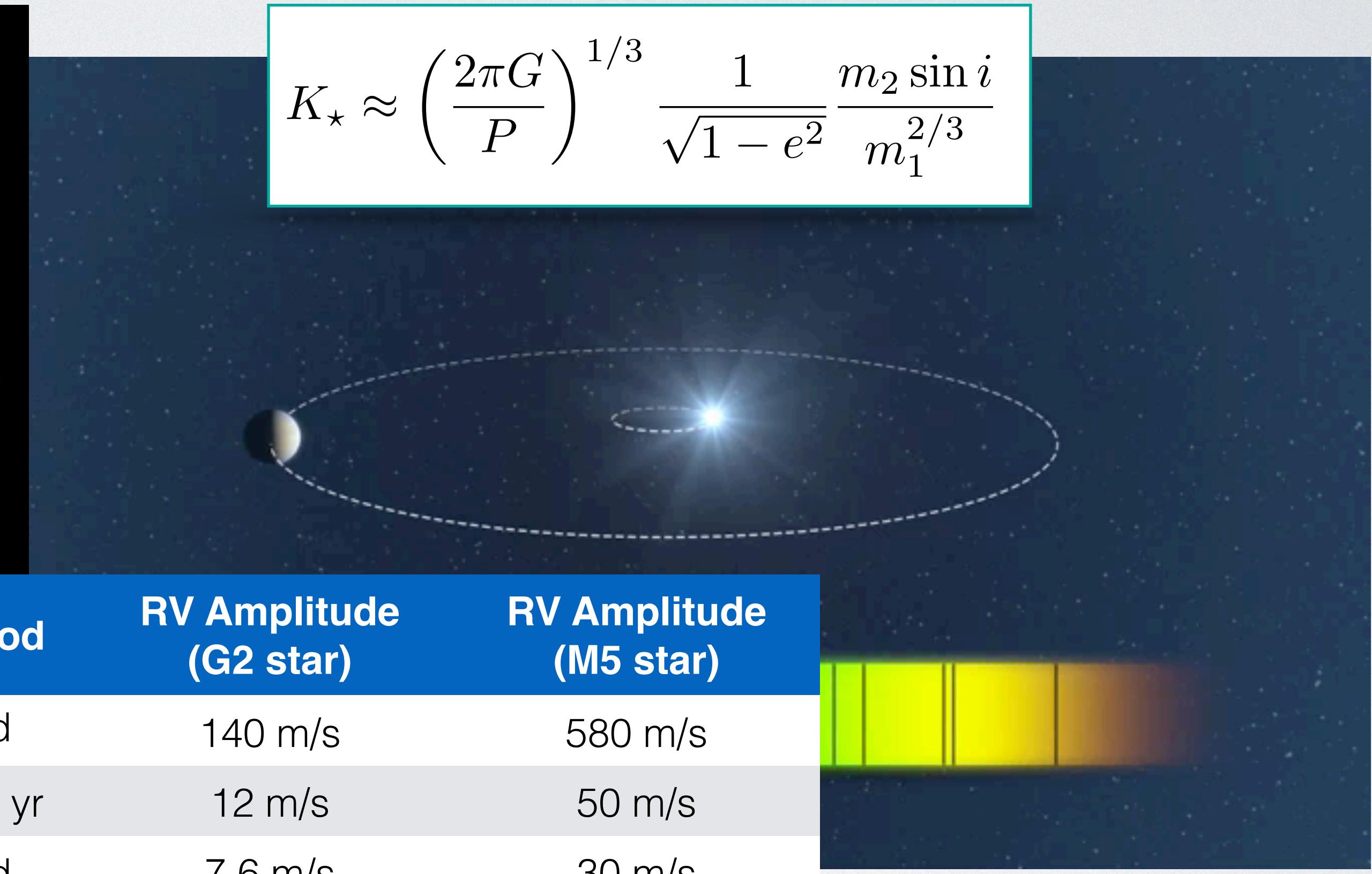
$$K_{\star} \approx \left(\frac{2\pi G}{P} \right)^{1/3} \frac{1}{\sqrt{1 - e^2}} \frac{m_2 \sin i}{m_1^{2/3}}$$



RADIAL VELOCITY METHOD

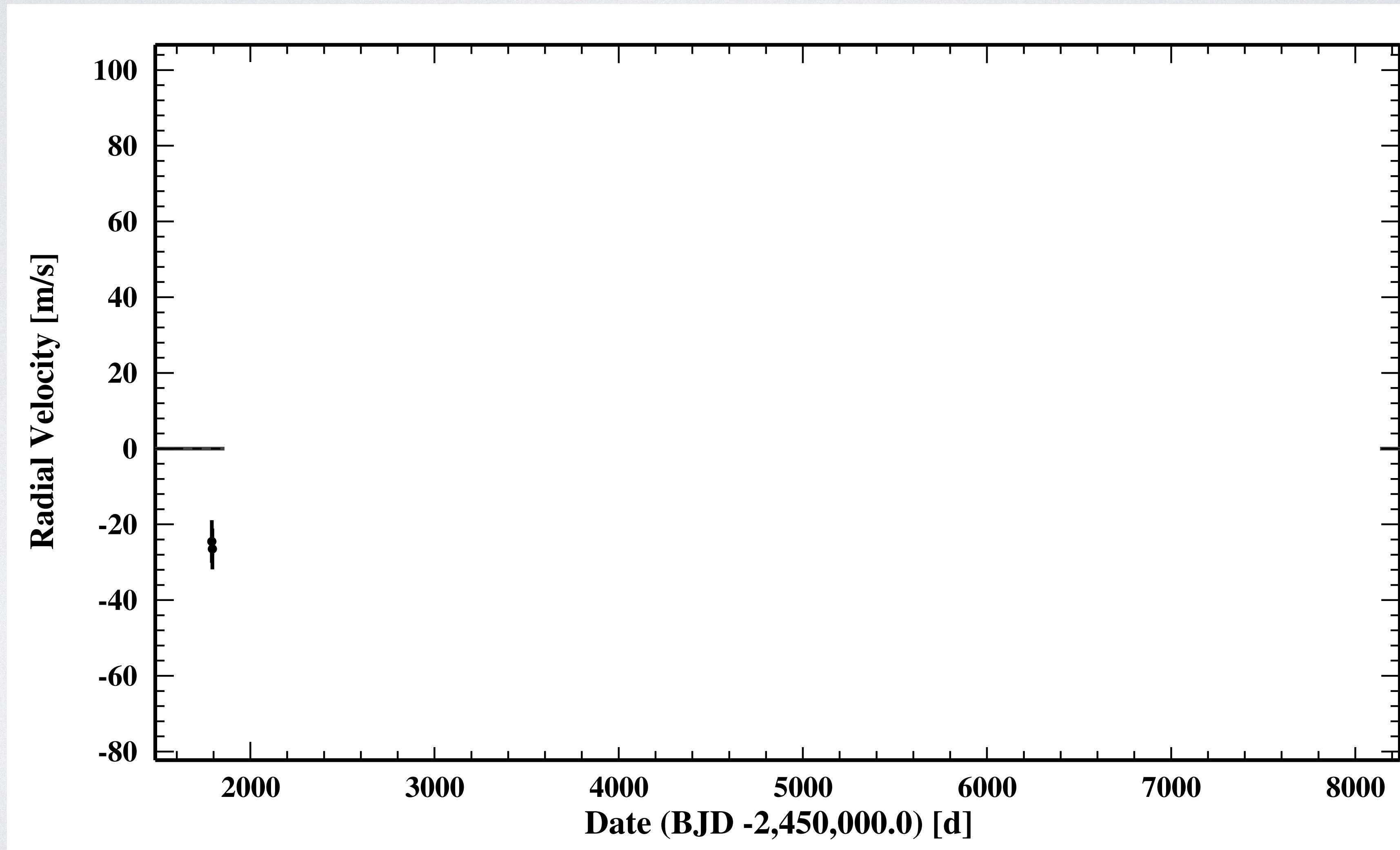


$$K_{\star} \approx \left(\frac{2\pi G}{P} \right)^{1/3} \frac{1}{\sqrt{1 - e^2}} \frac{m_2 \sin i}{m_1^{2/3}}$$

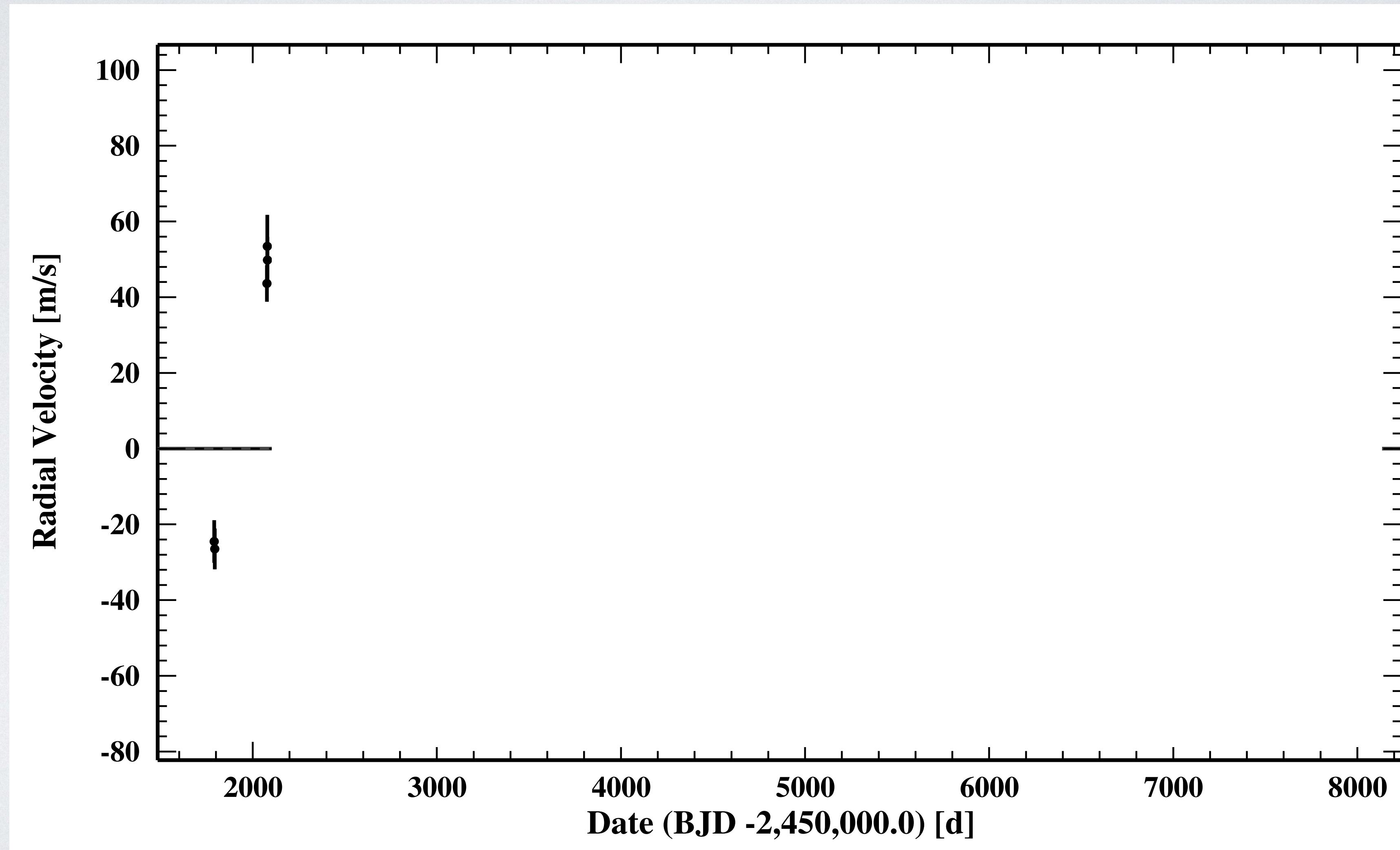


Planet	Period	RV Amplitude (G2 star)	RV Amplitude (M5 star)
Jupiter	3 d	140 m/s	580 m/s
Jupiter	11.9 yr	12 m/s	50 m/s
Neptune	3 d	7.6 m/s	30 m/s
Earth	3 d	44 cm/s	1.8 m/s
Earth	1 yr	9 cm/s	40 cm/s

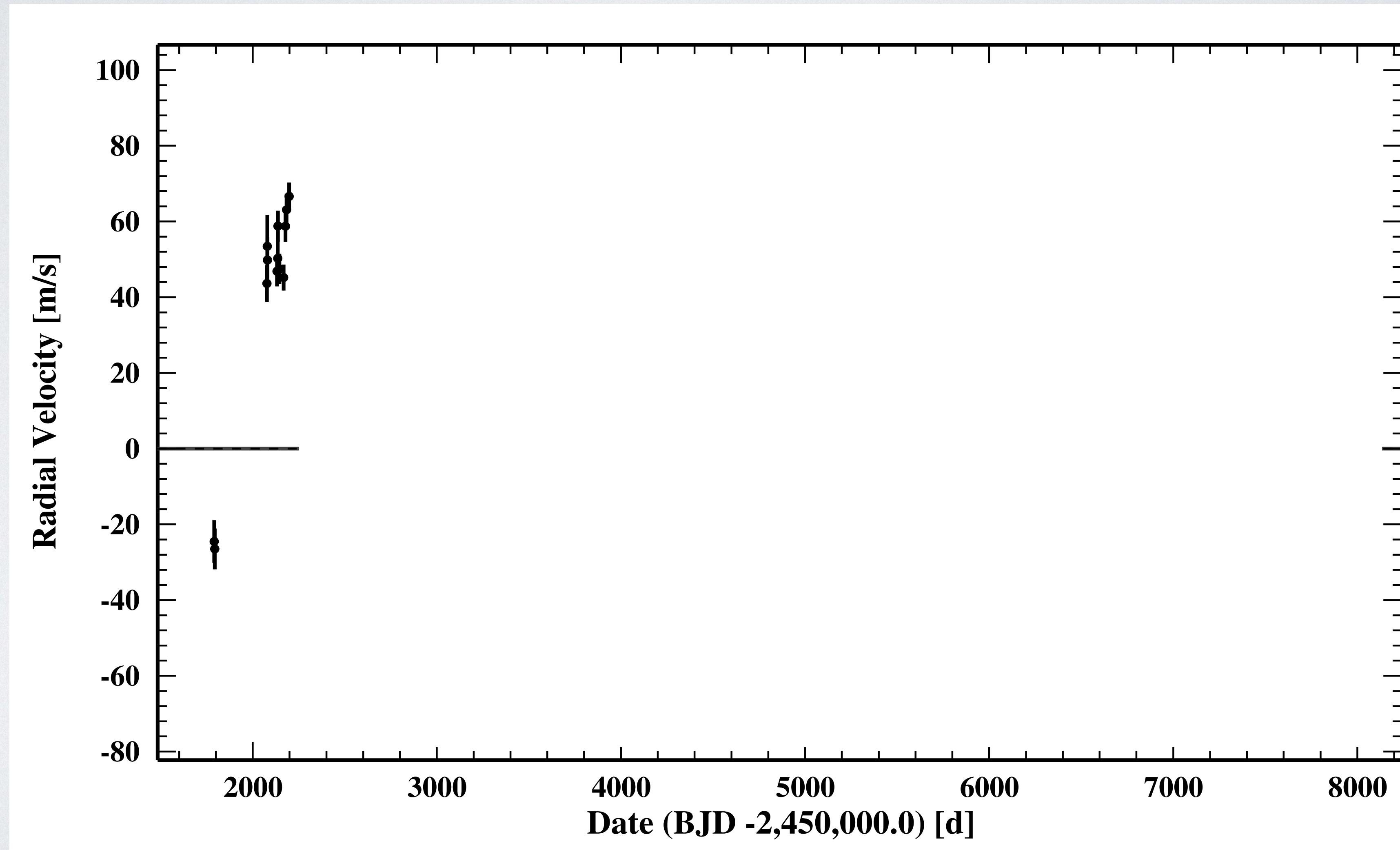
RADIAL VELOCITY TIME SERIES



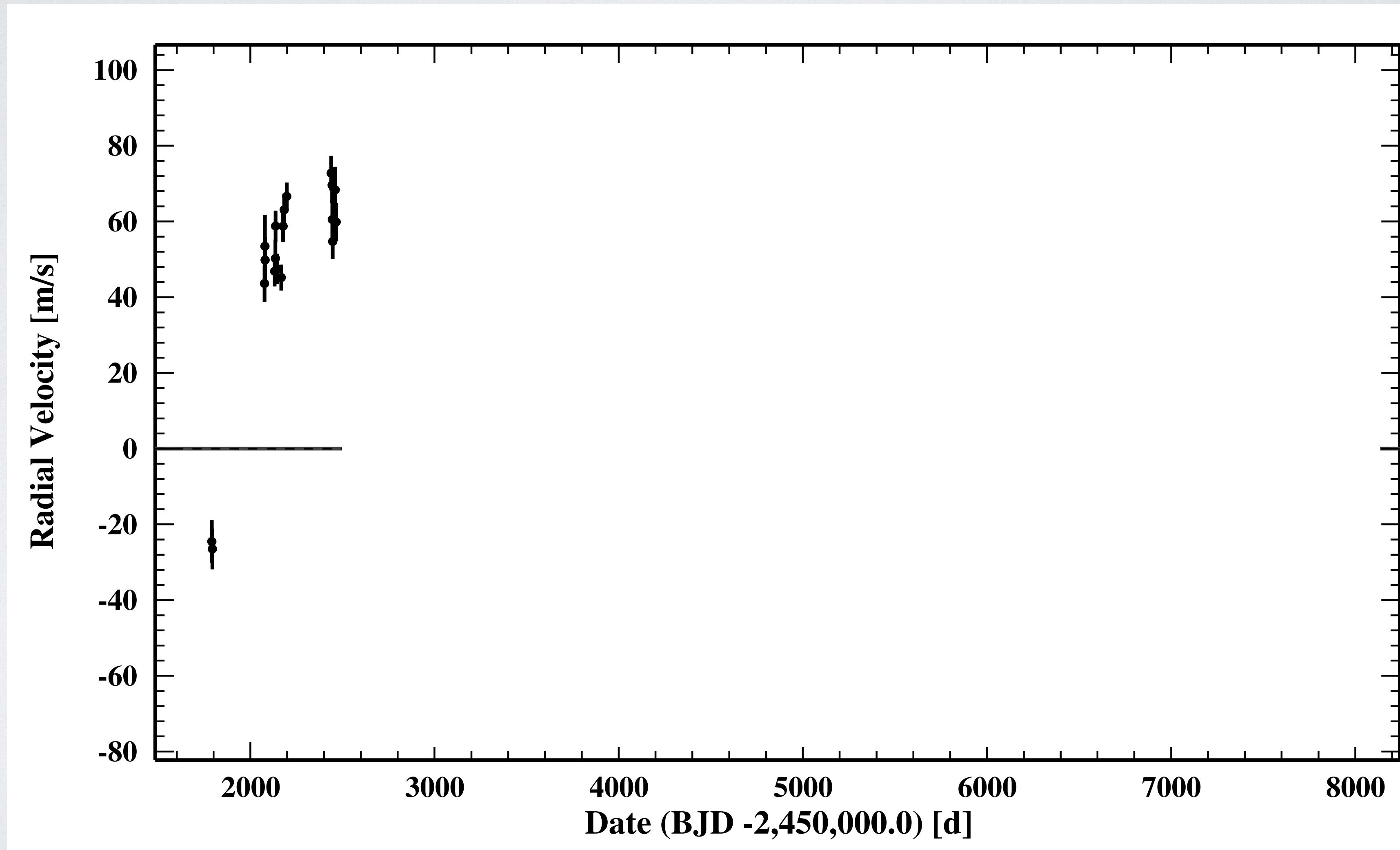
RADIAL VELOCITY TIME SERIES



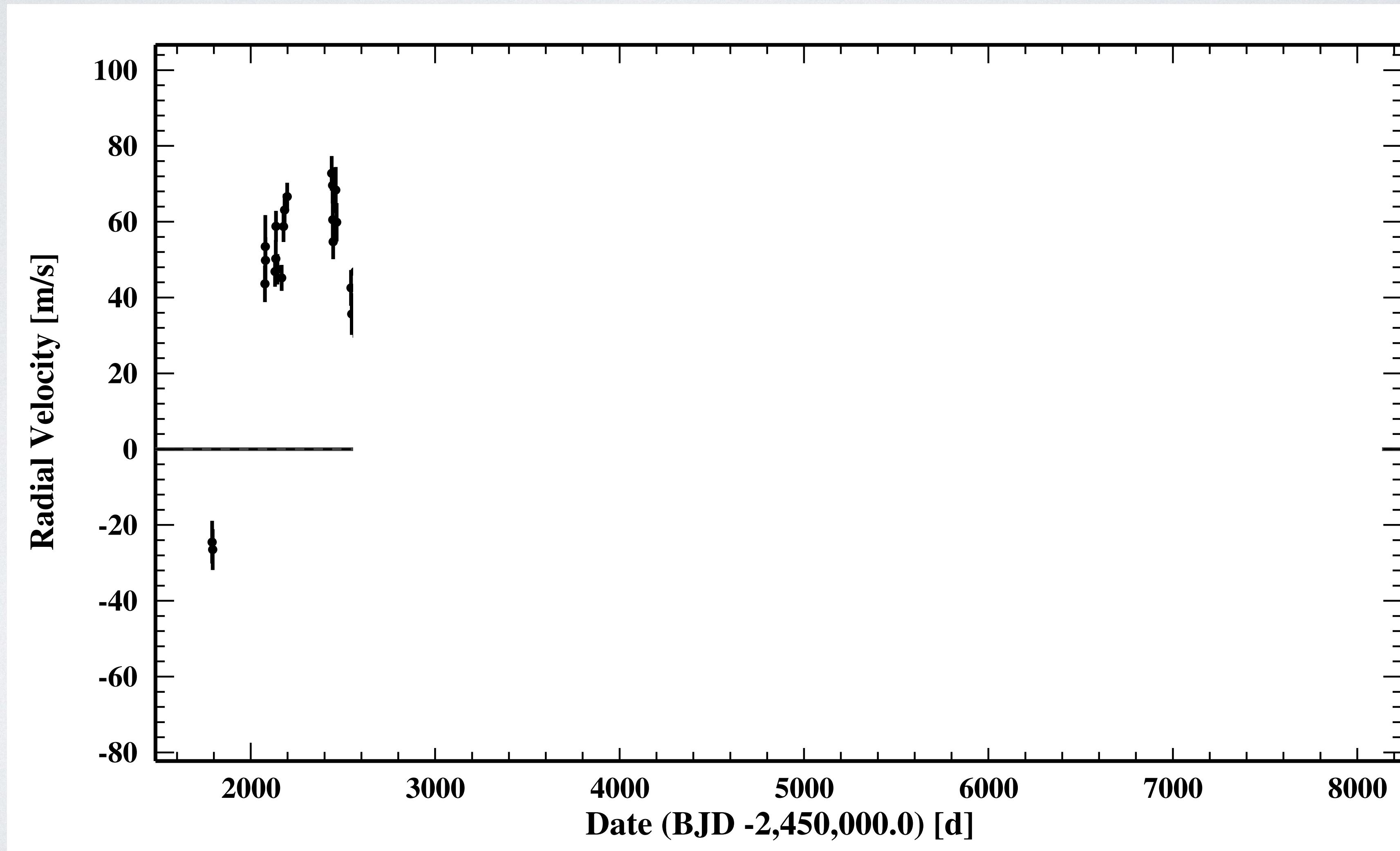
RADIAL VELOCITY TIME SERIES



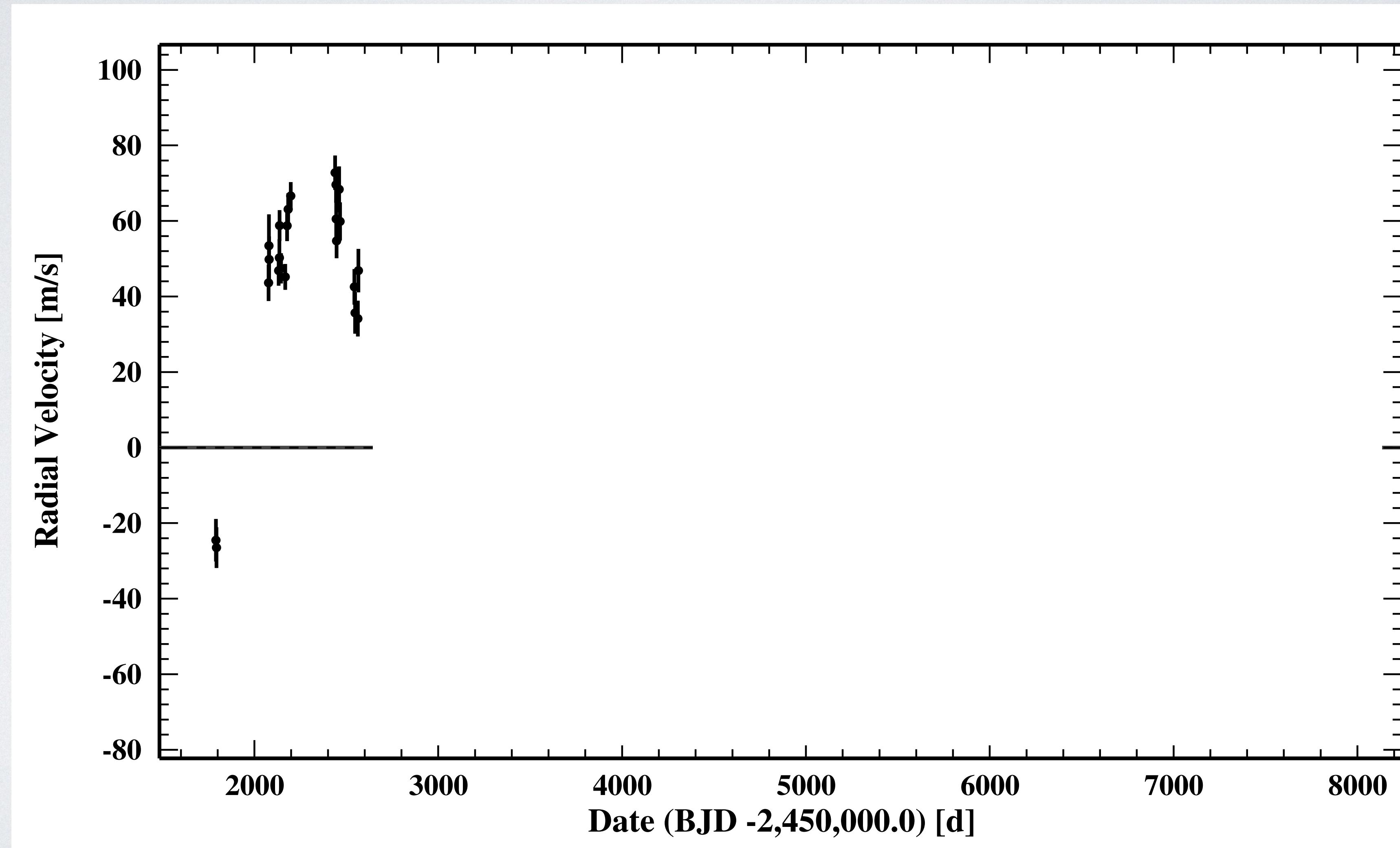
RADIAL VELOCITY TIME SERIES



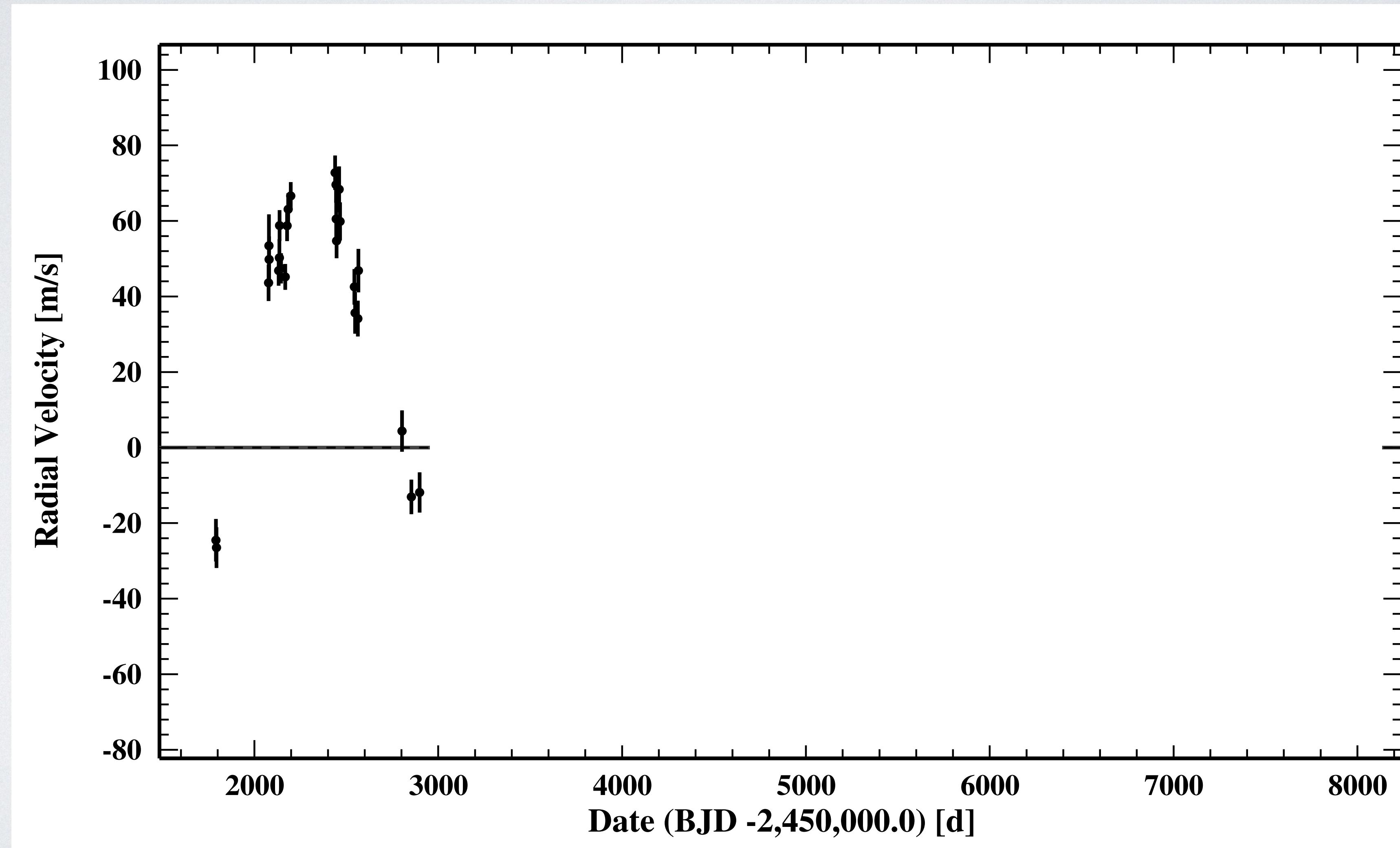
RADIAL VELOCITY TIME SERIES



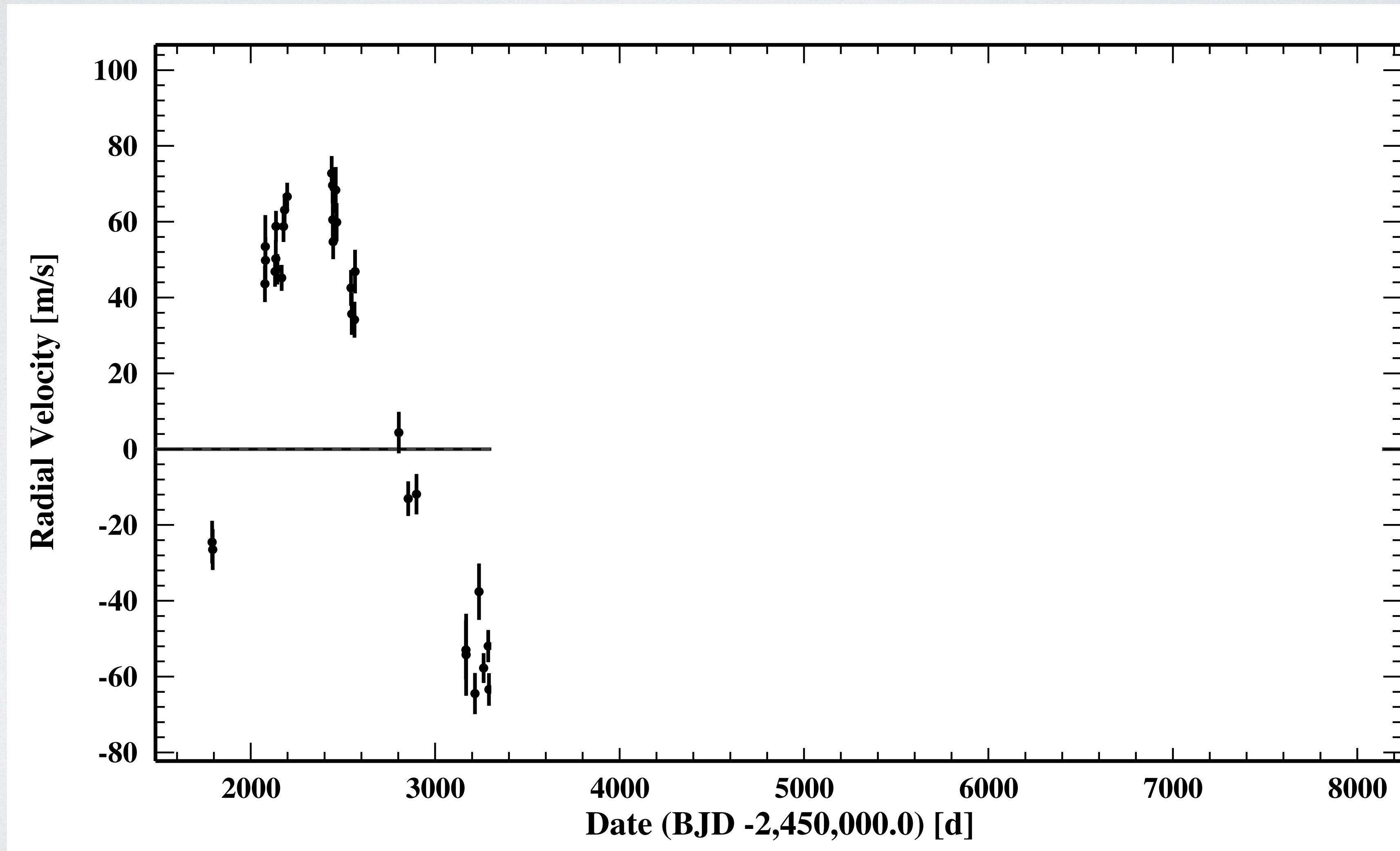
RADIAL VELOCITY TIME SERIES



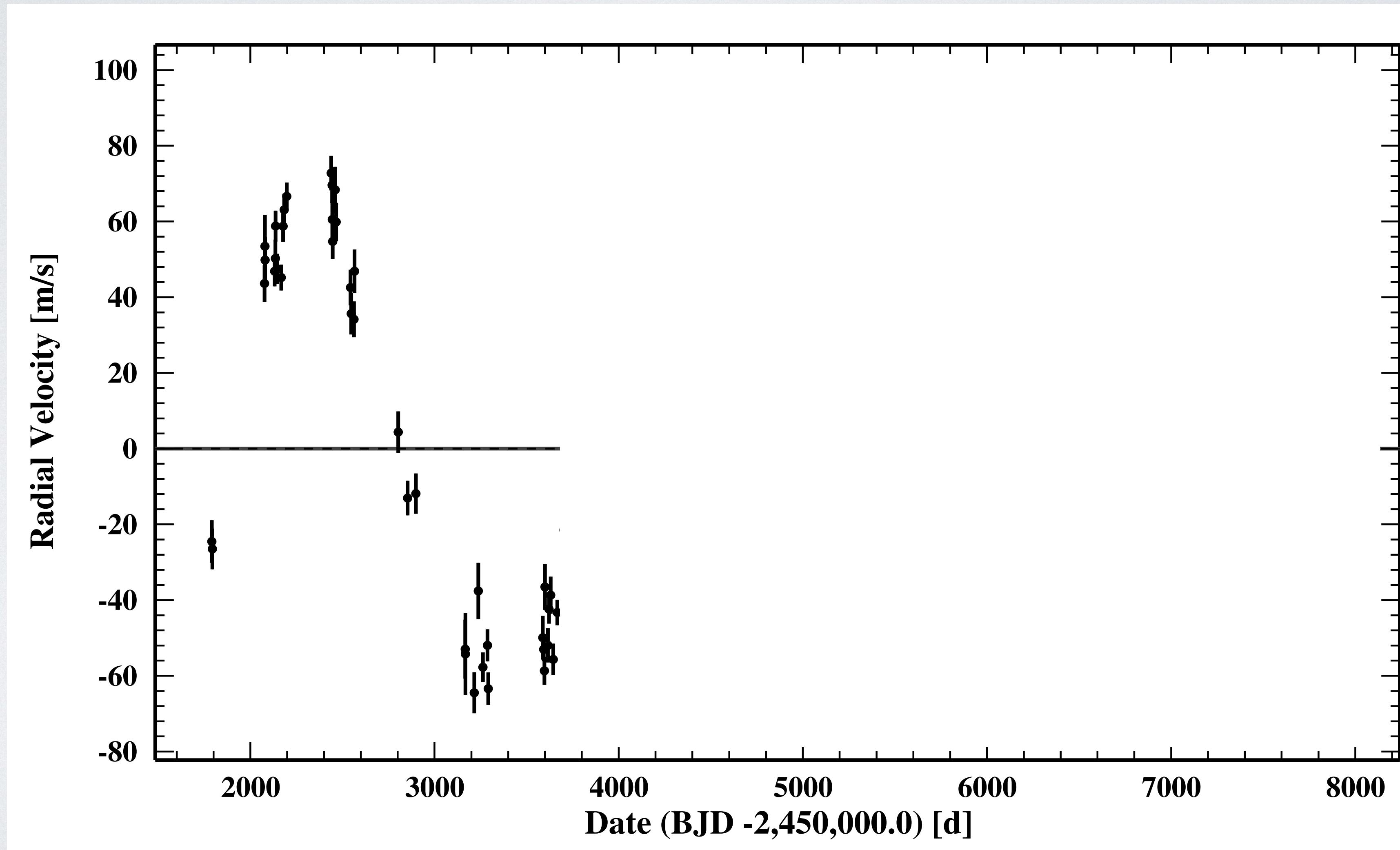
RADIAL VELOCITY TIME SERIES



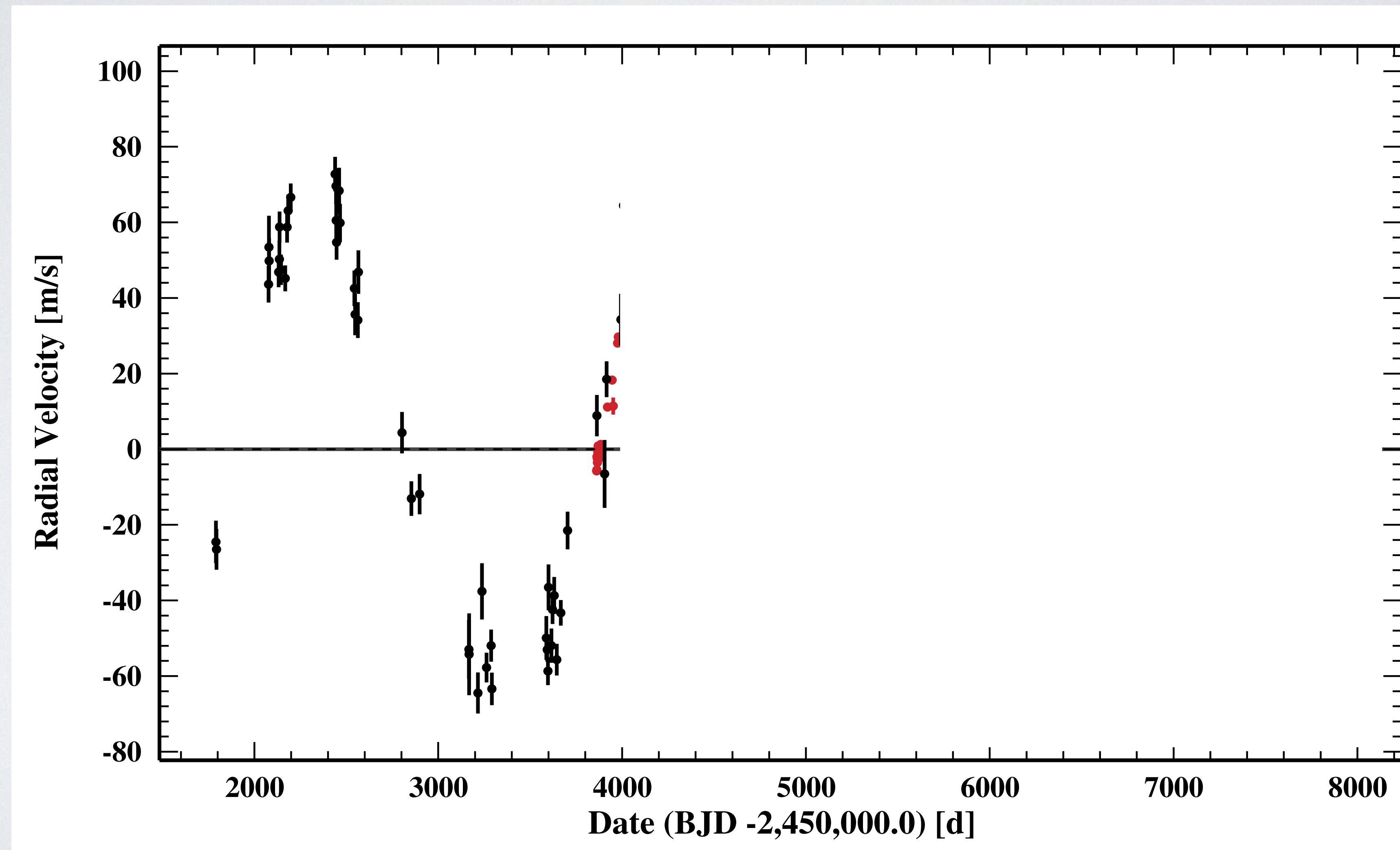
RADIAL VELOCITY TIME SERIES



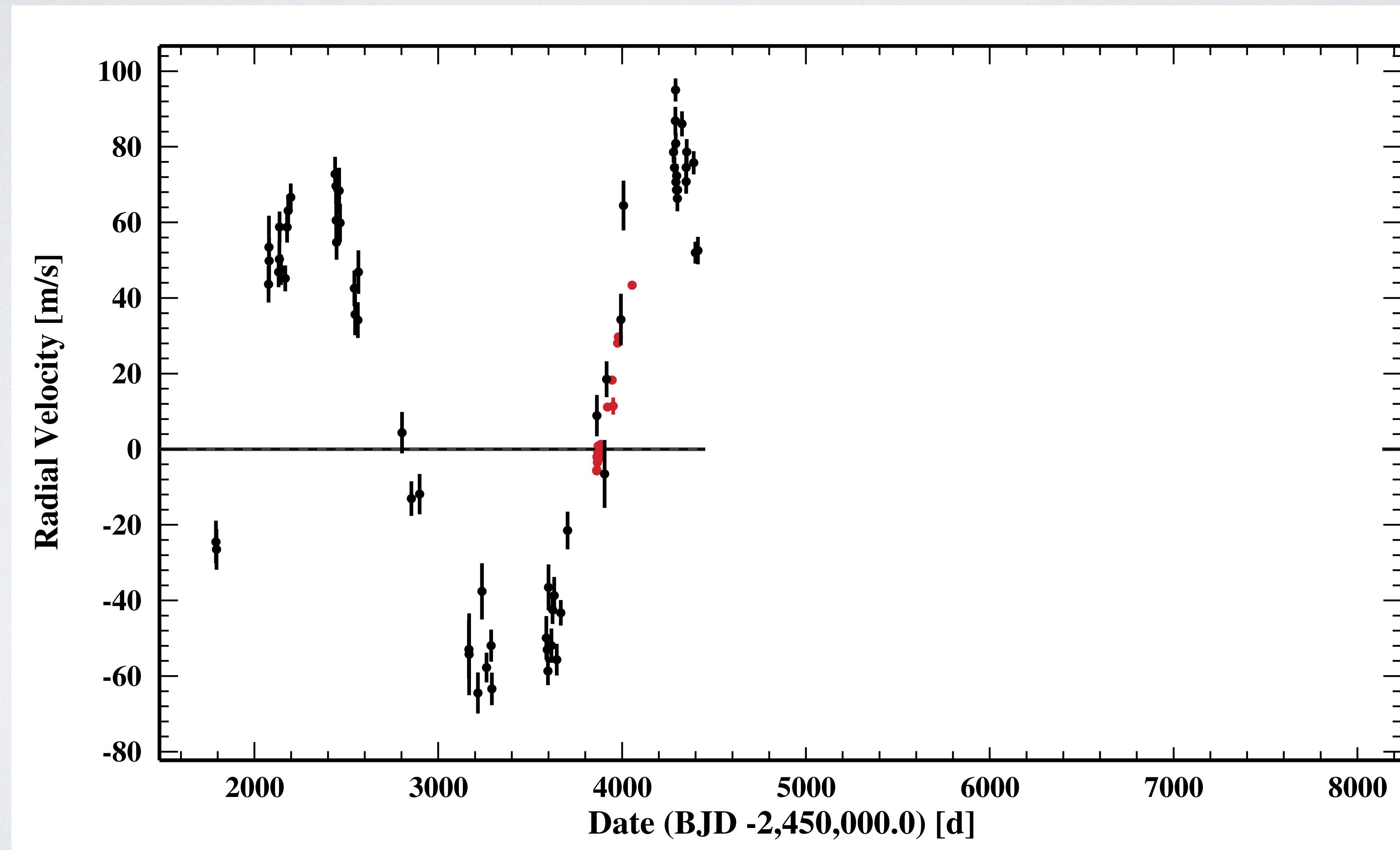
RADIAL VELOCITY TIME SERIES



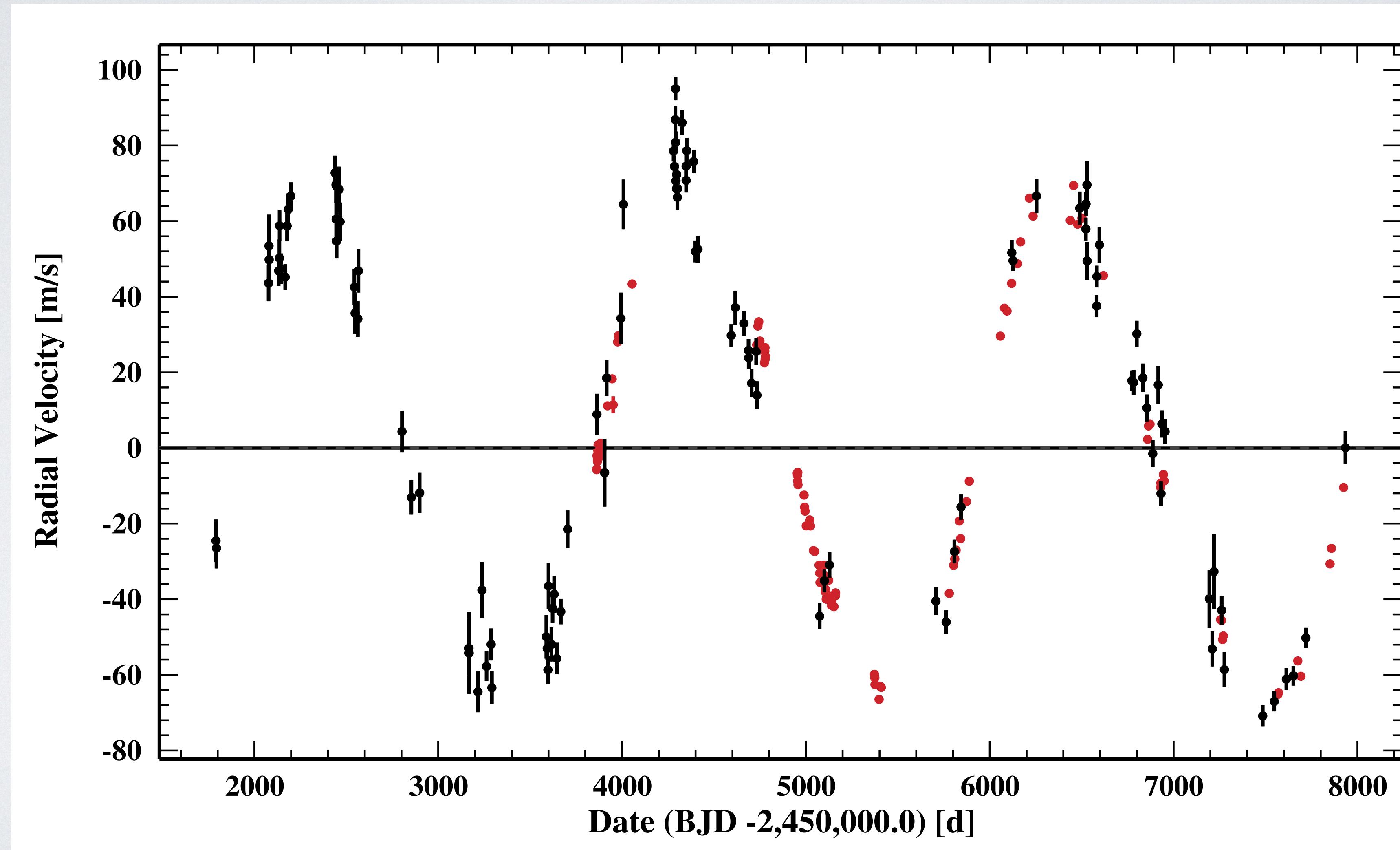
RADIAL VELOCITY TIME SERIES



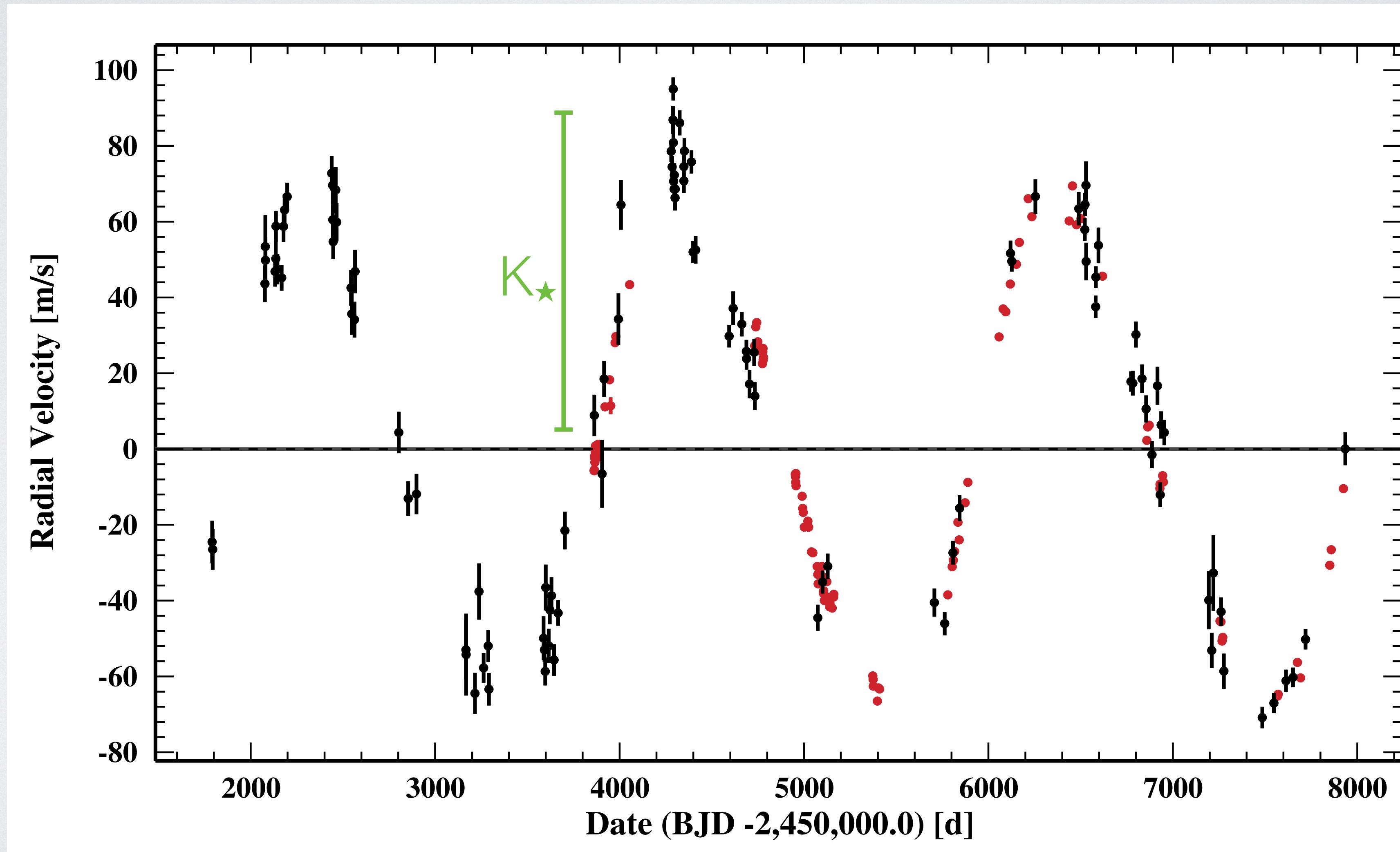
RADIAL VELOCITY TIME SERIES



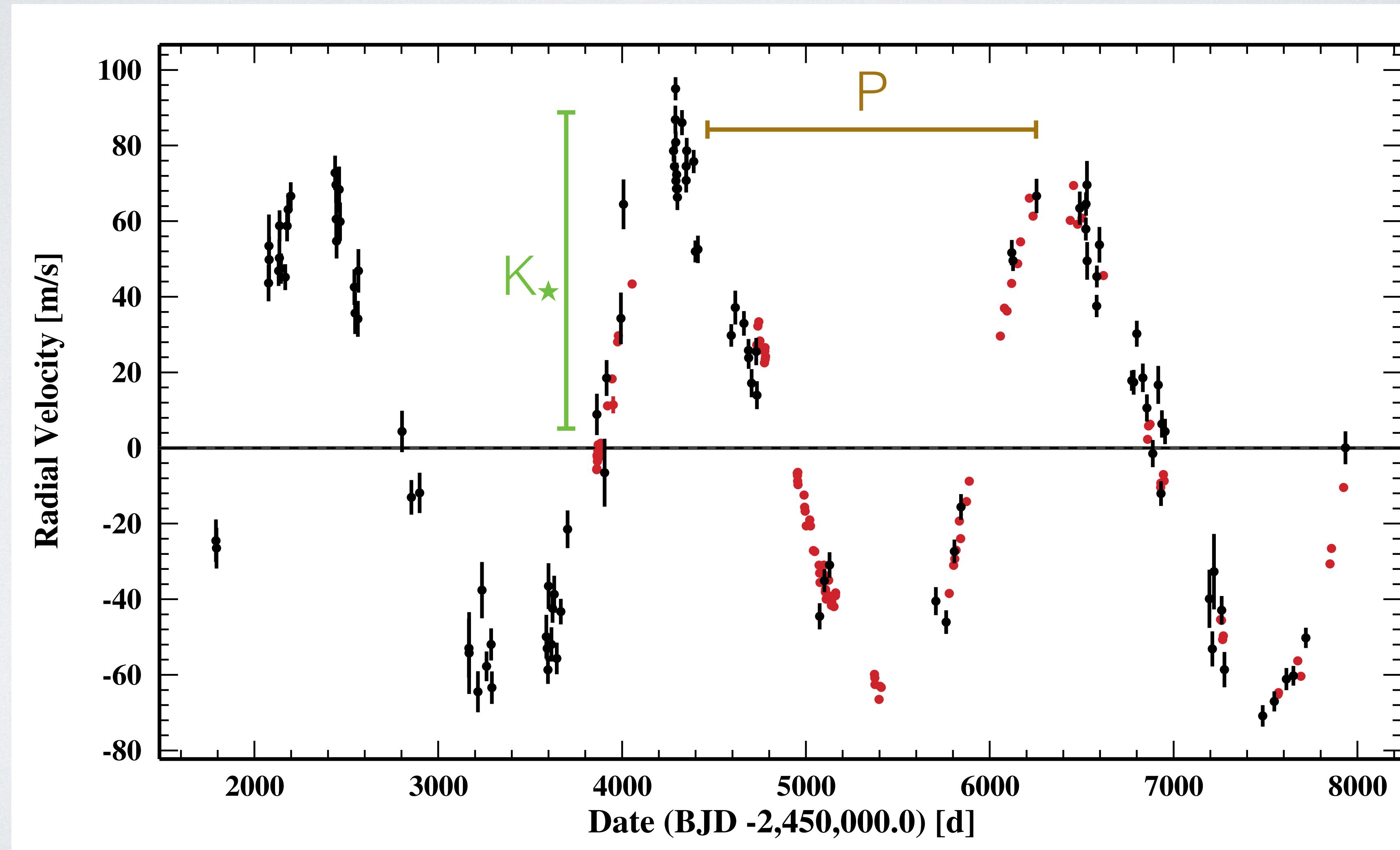
RADIAL VELOCITY TIME SERIES



RADIAL VELOCITY TIME SERIES



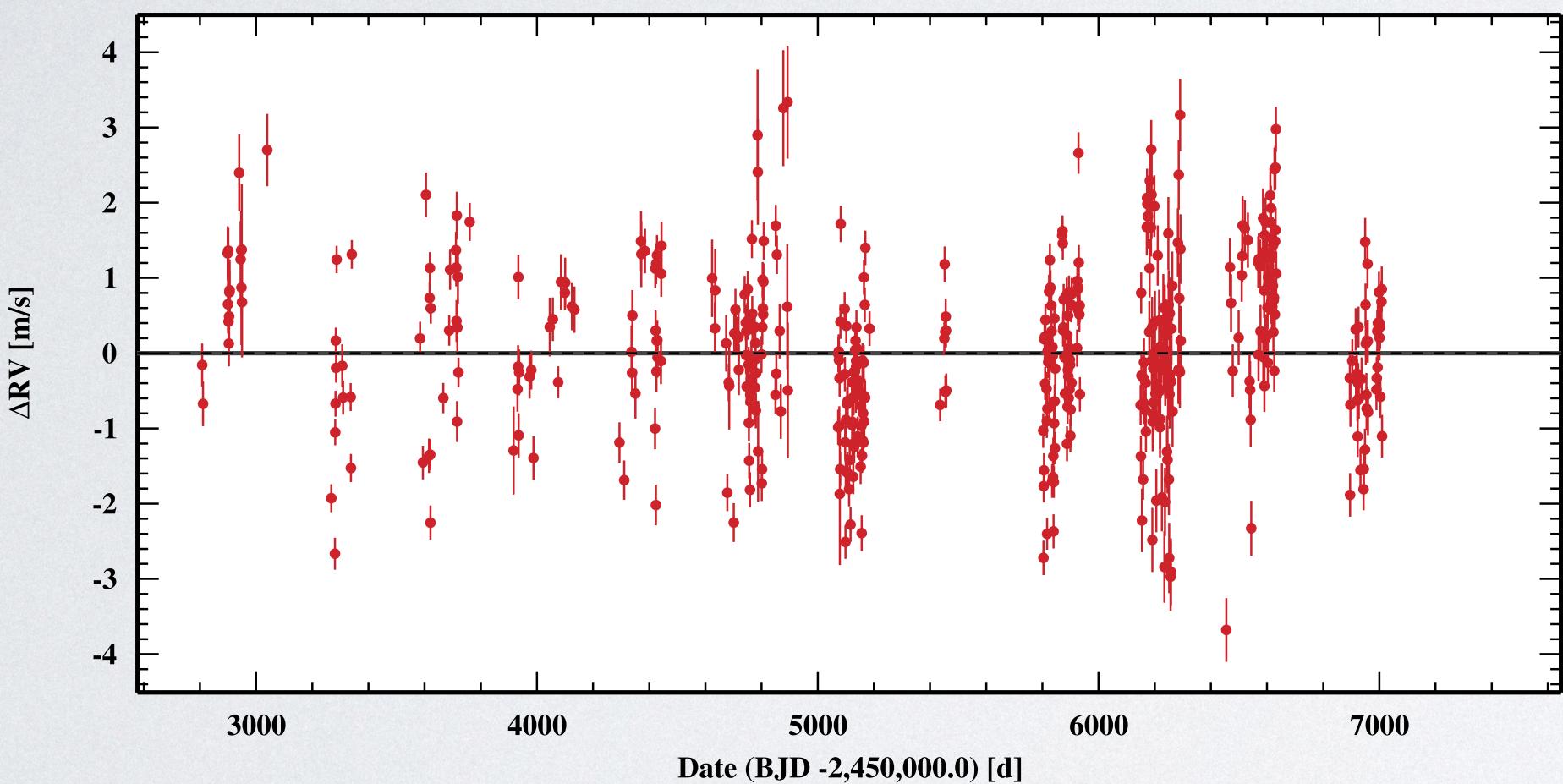
RADIAL VELOCITY TIME SERIES



THE GLS PERIODOGRAM

The classical approach to detecting signals in RV time series

Generalised Lomb Scargle (GLS): Sort of Fourier Transform for unevenly sampled data.



$$p(\omega) = \frac{\chi_0^2 - \chi_\omega^2}{\chi_0^2}$$

Lomb (1976)

Scargle (1982)

Baluev (2008, 2013)

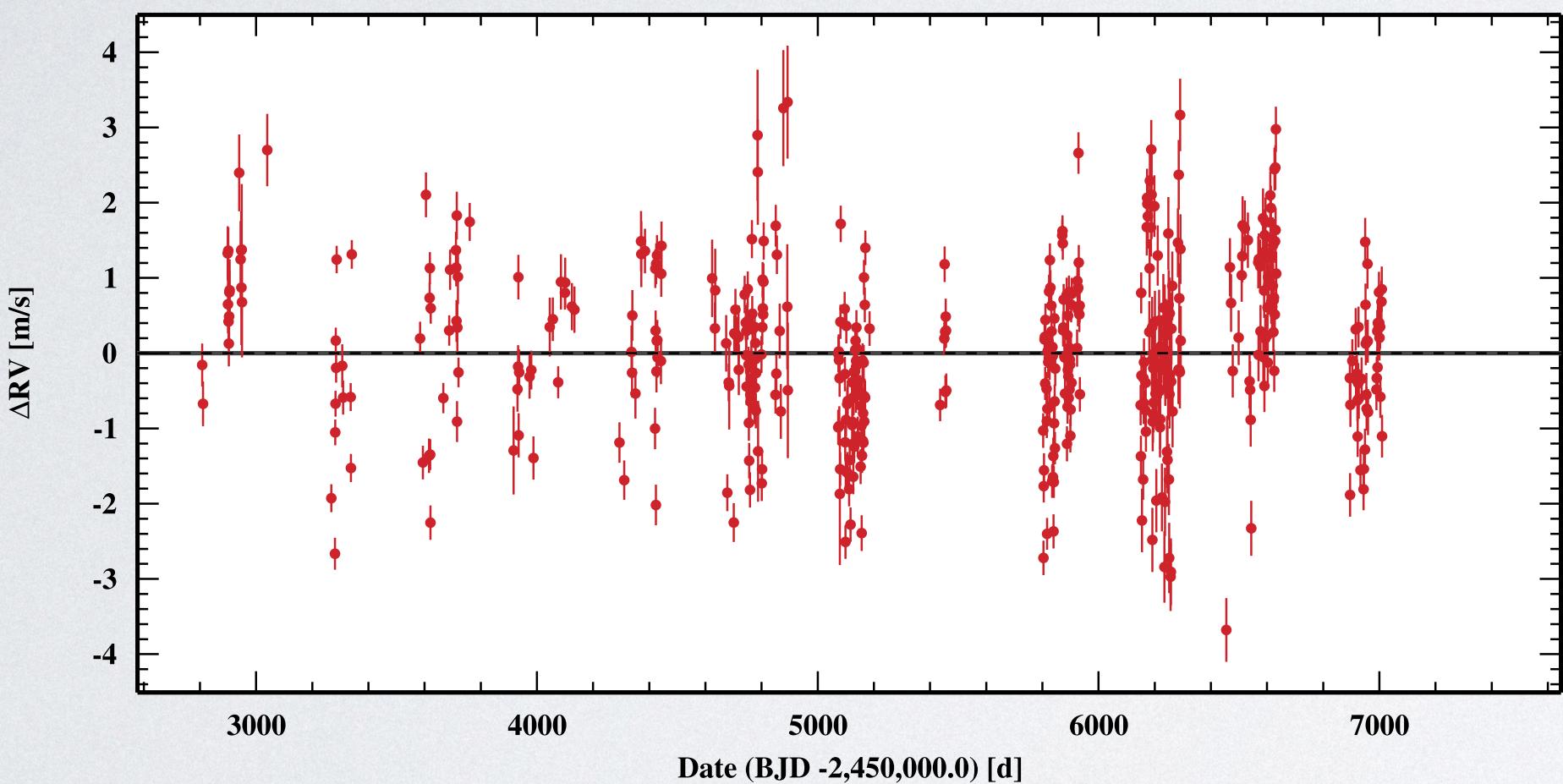
Zehcmeister & Kürster (2009)

Delisle, Ségransan & Hara (2020)

THE GLS PERIODOGRAM

The classical approach to detecting signals in RV time series

Generalised Lomb Scargle (GLS): Sort of Fourier Transform for unevenly sampled data.



$$p(\omega) = \frac{\chi_0^2 - \chi_\omega^2}{\chi_0^2}$$

χ^2 for a linear sinusoidal model

χ^2 for the weighted mean

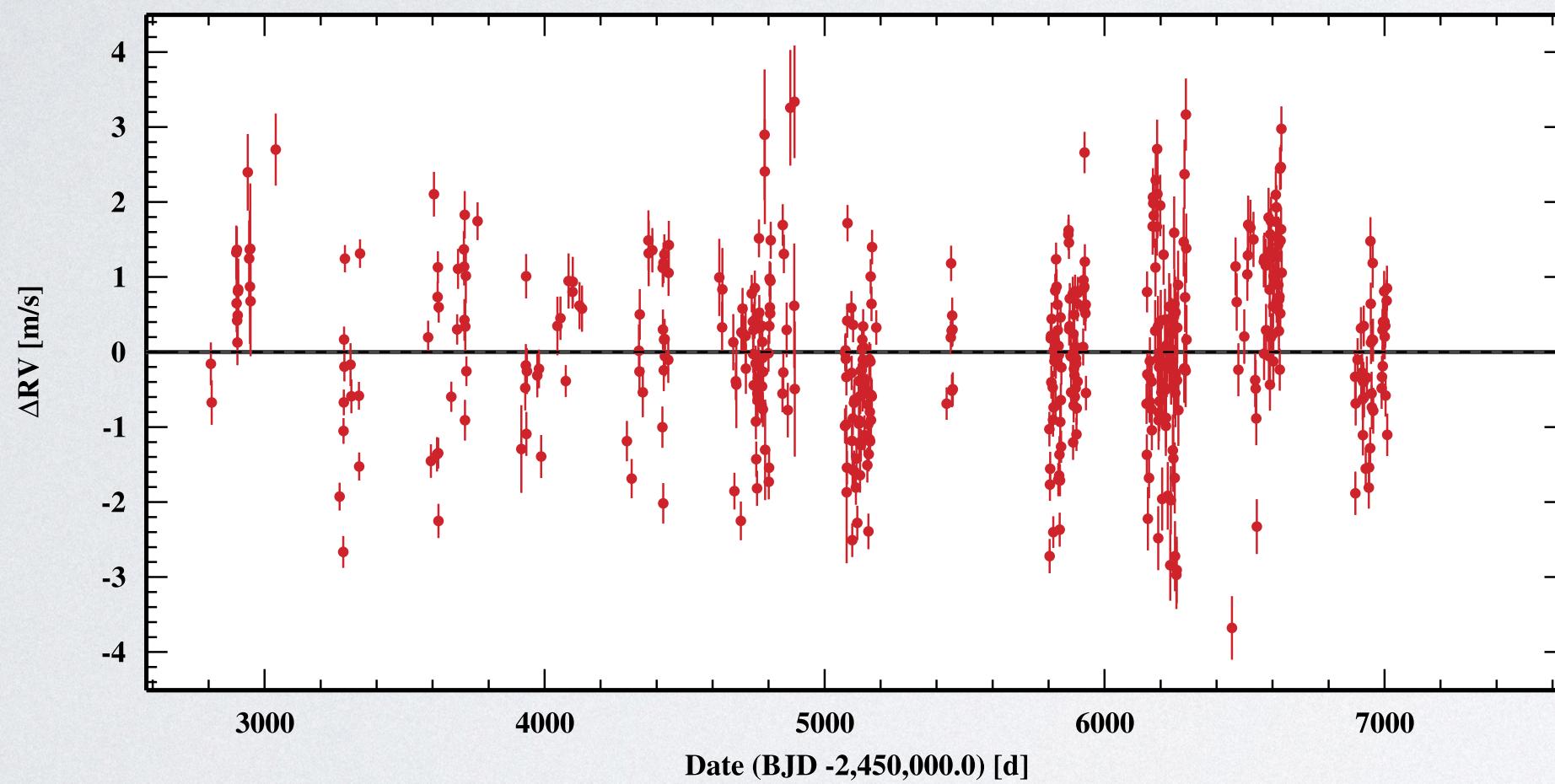
$$A \sin(\omega t) + B \cos(\omega t) + C$$

- Lomb (1976)
- Scargle (1982)
- Baluev (2008, 2013)
- Zehcmeister & Kürster (2009)
- Delisle, Ségransan & Hara (2020)

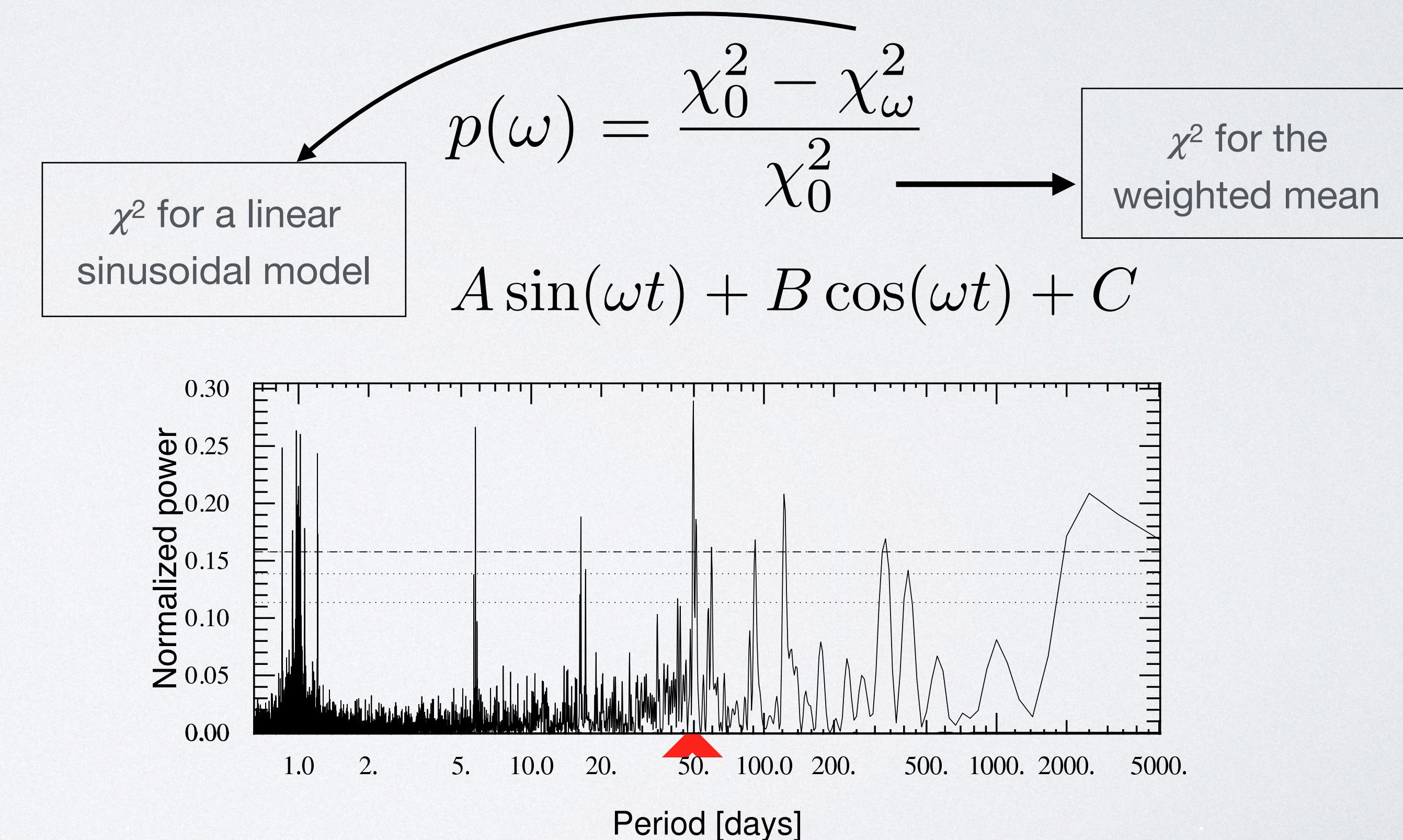
THE GLS PERIODOGRAM

The classical approach to detecting signals in RV time series

Generalised Lomb Scargle (GLS): Sort of Fourier Transform for unevenly sampled data.



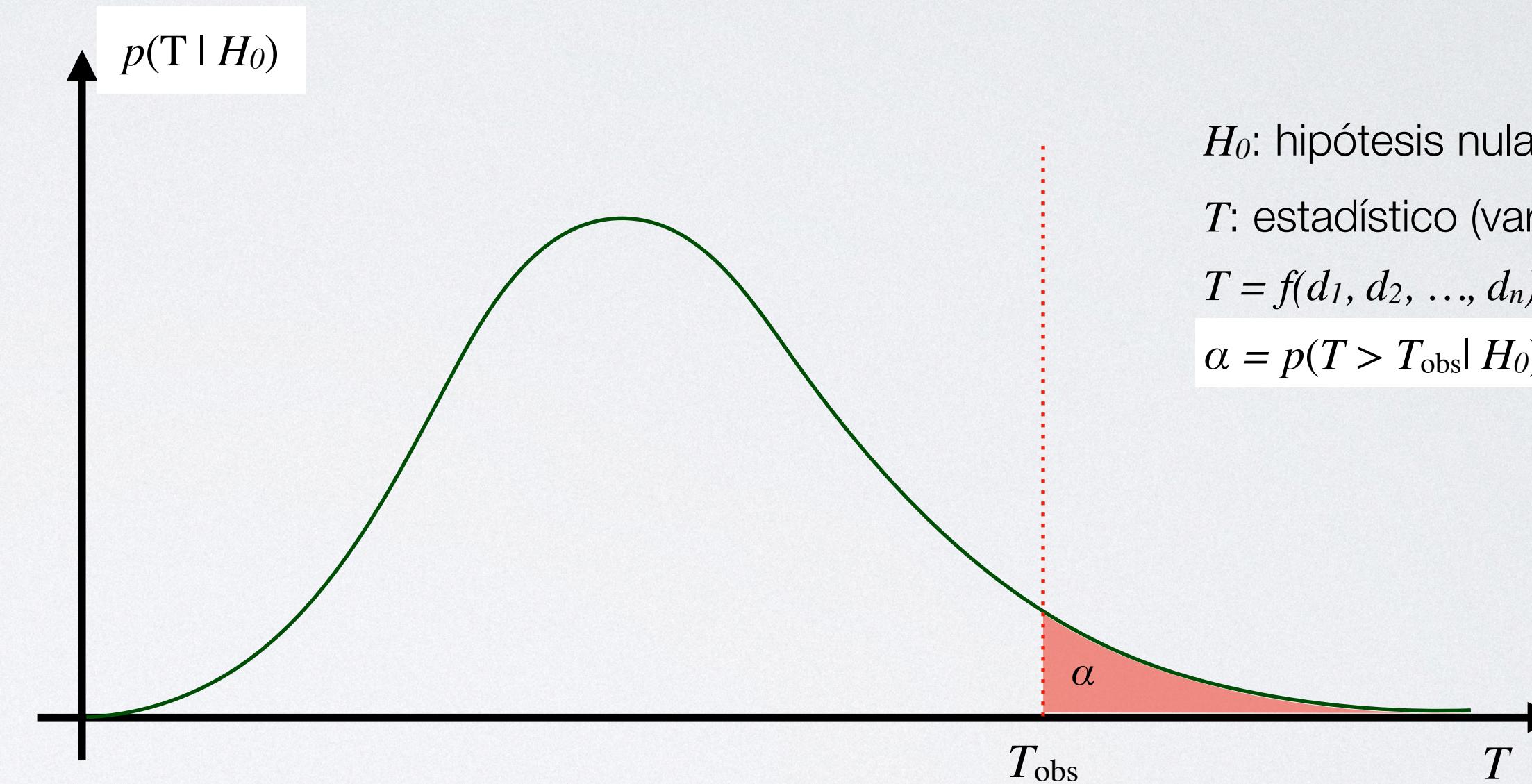
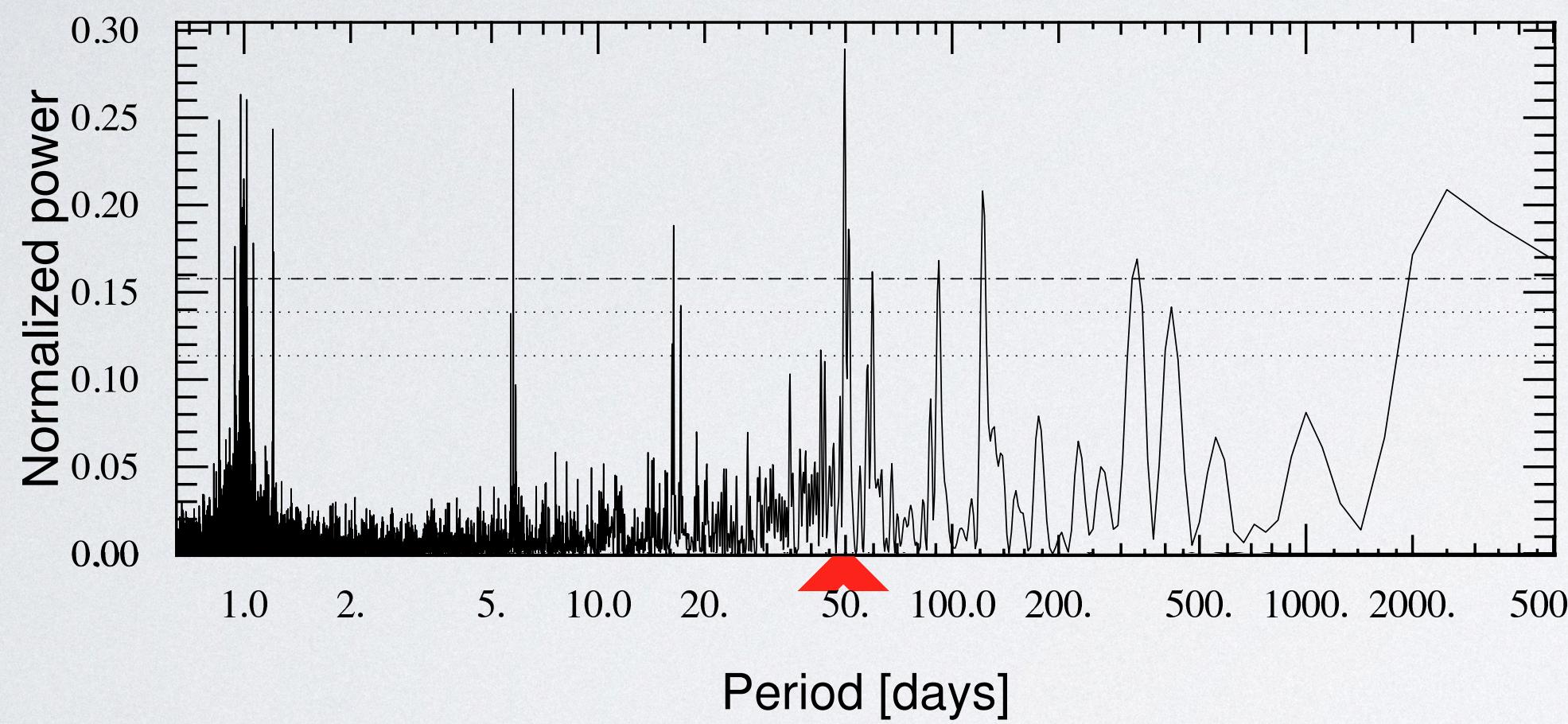
Lomb (1976)
Scargle (1982)
Baluev (2008, 2013)
Zehcmeister & Kürster (2009)
Delisle, Ségransan & Hara (2020)



THE GLS PERIODOGRAM

The classical approach to detecting signals in RV time series

Statistical significance evaluated via Null Hypothesis Significance Testing



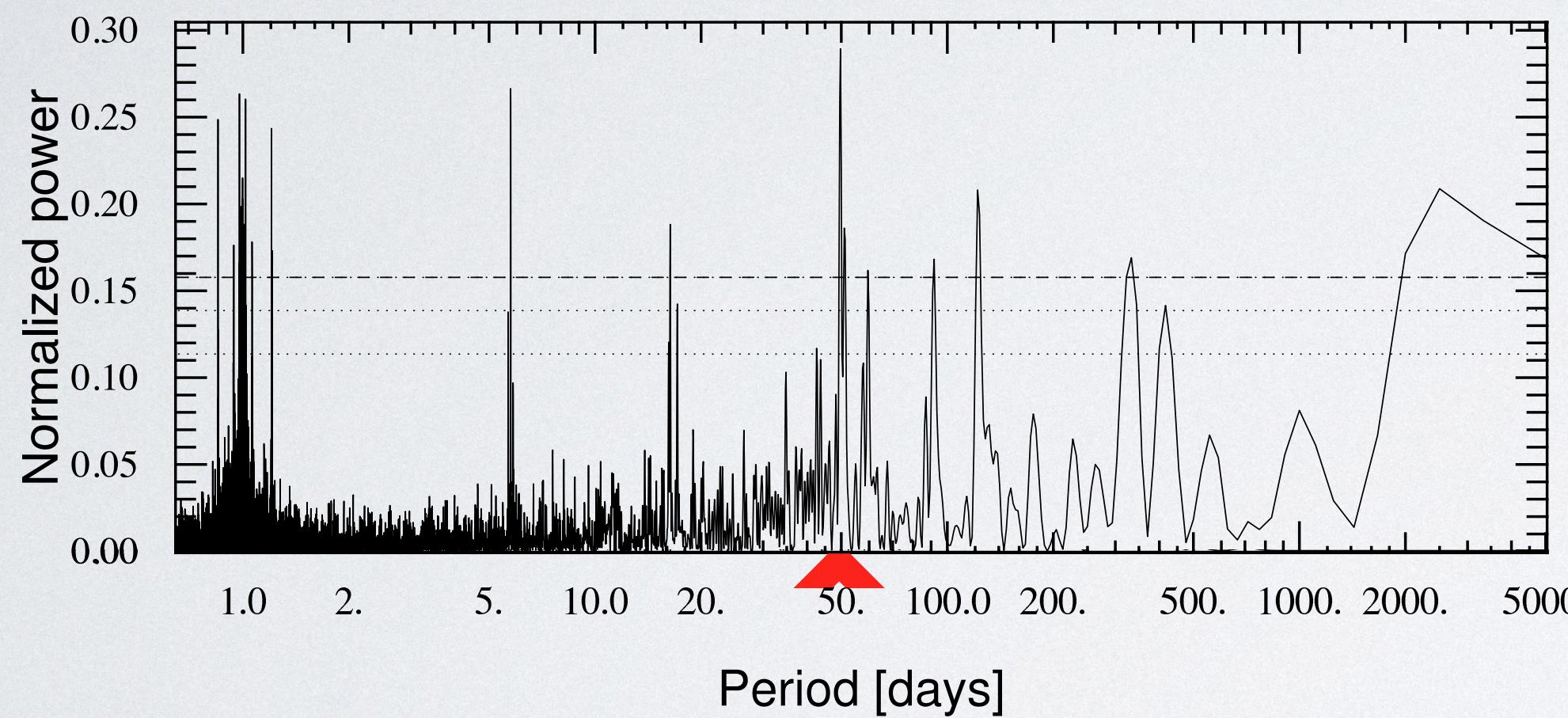
H_0 : hipótesis nula
 T : estadístico (variable aleatoria)
 $T = f(d_1, d_2, \dots, d_n)$
 $\alpha = p(T > T_{\text{obs}} \mid H_0)$, p-value

Lomb (1976)
Scargle (1982)
Baluev (2008, 2013)
Zehcmeister & Kürster (2009)
Delisle, Ségransan & Hara (2020)

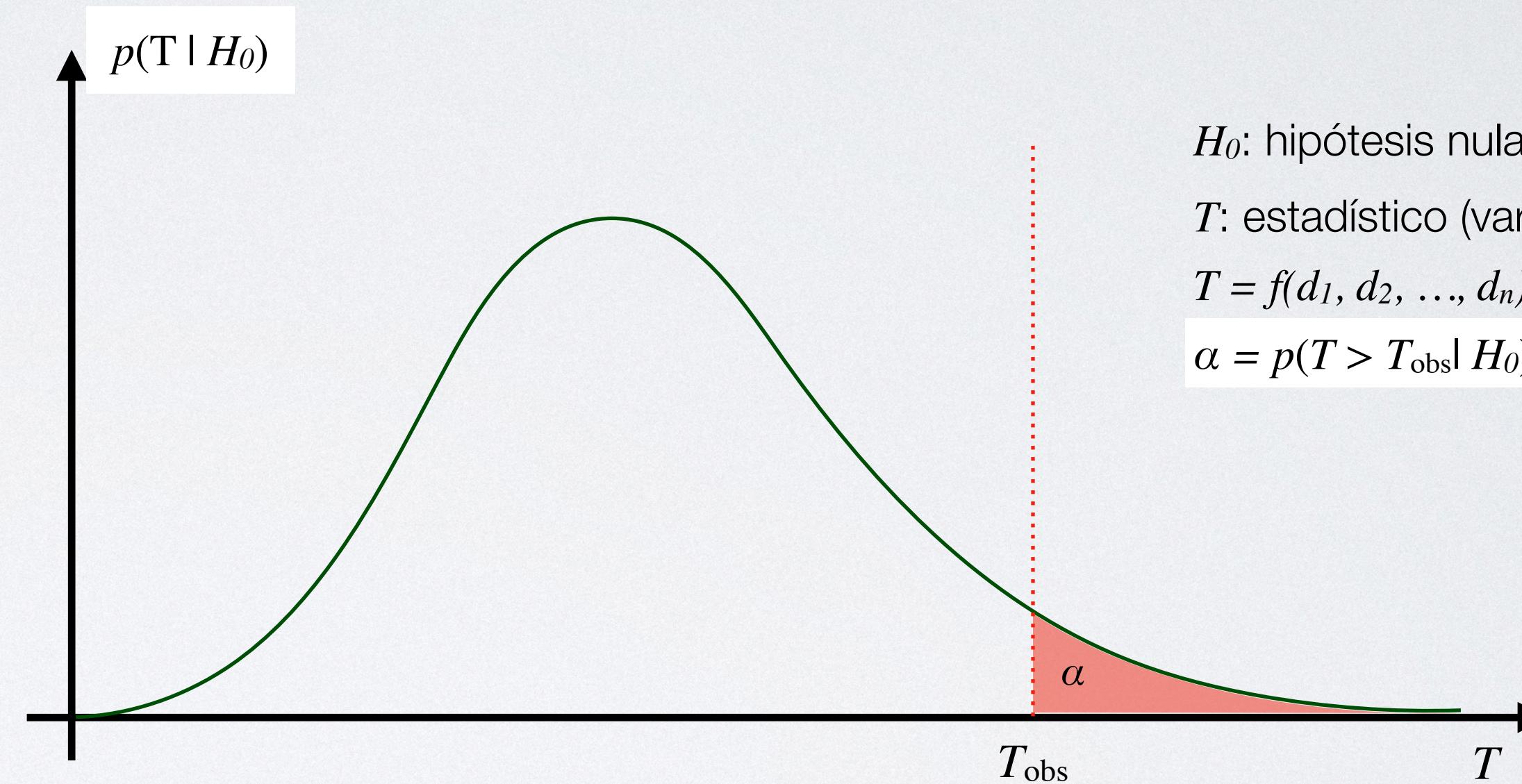
THE GLS PERIODOGRAM

The classical approach to detecting signals in RV time series

Statistical significance evaluated via Null Hypothesis Significance Testing



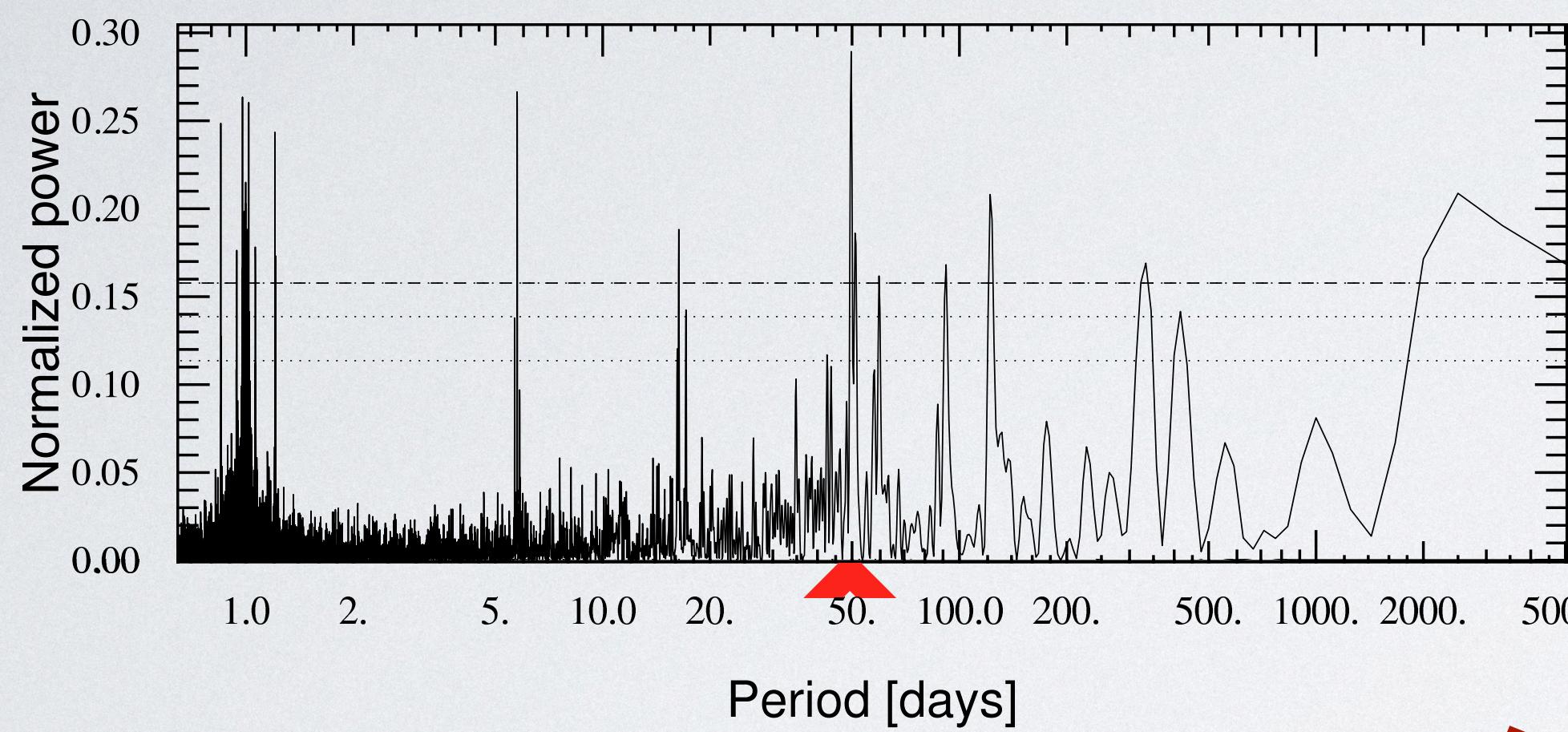
Lomb (1976)
Scargle (1982)
Baluev (2008, 2013)
Zehcmeister & Kürster (2009)
Delisle, Ségransan & Hara (2020)



The most commonly used statistics is the power of the largest peak in the periodogram.

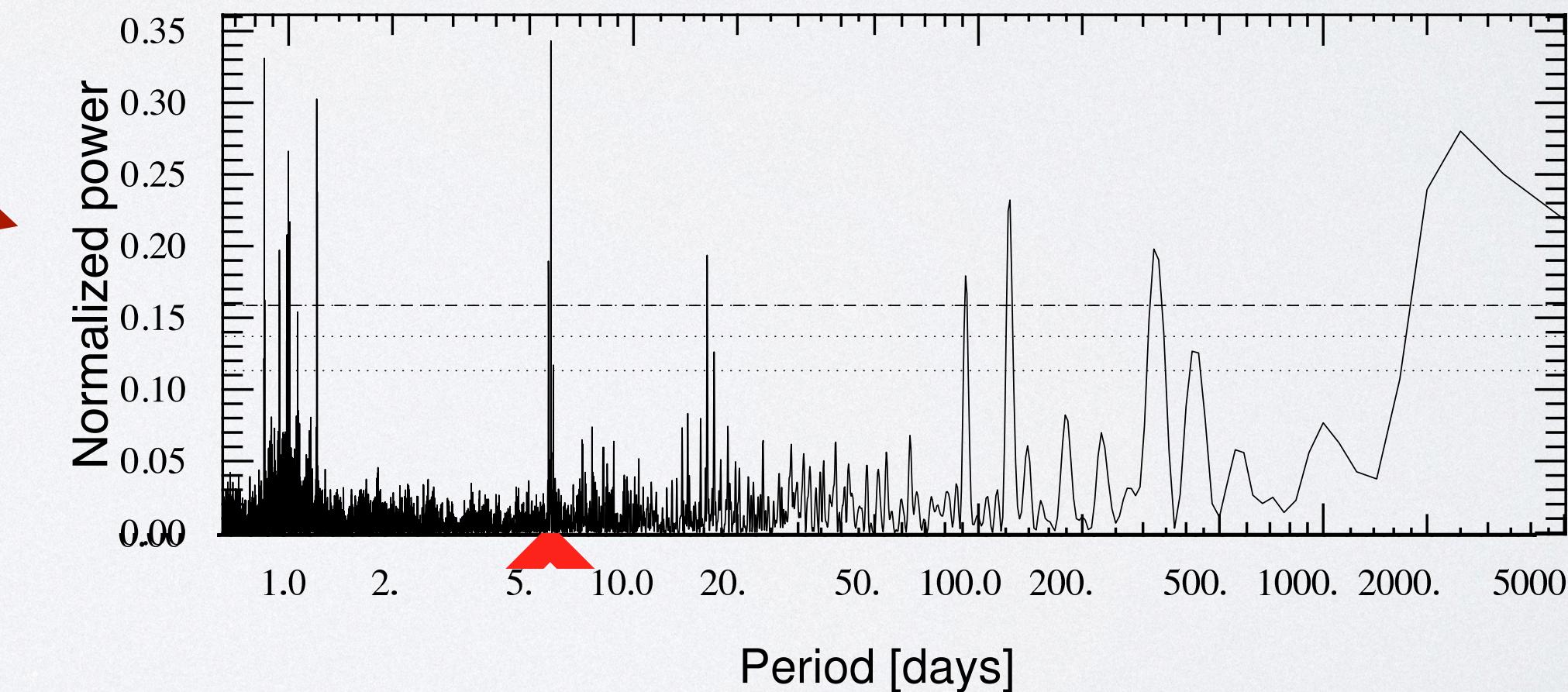
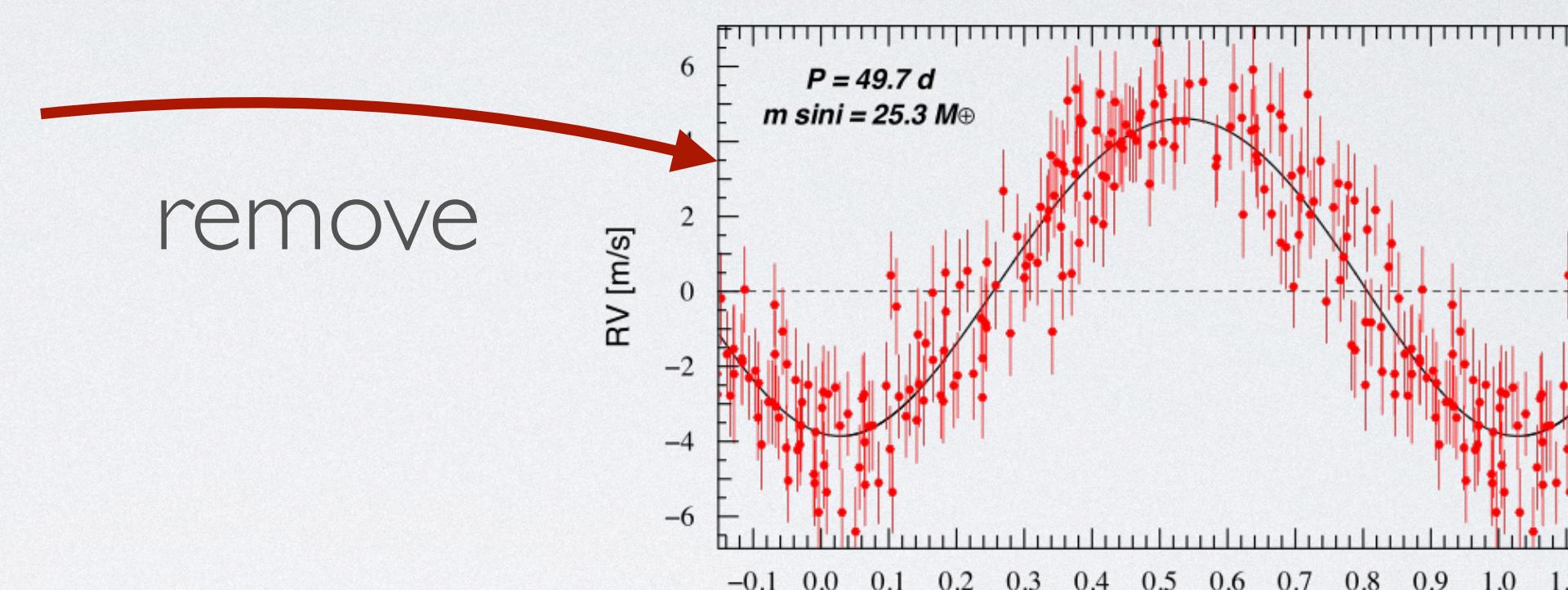
THE GLS PERIODOGRAM

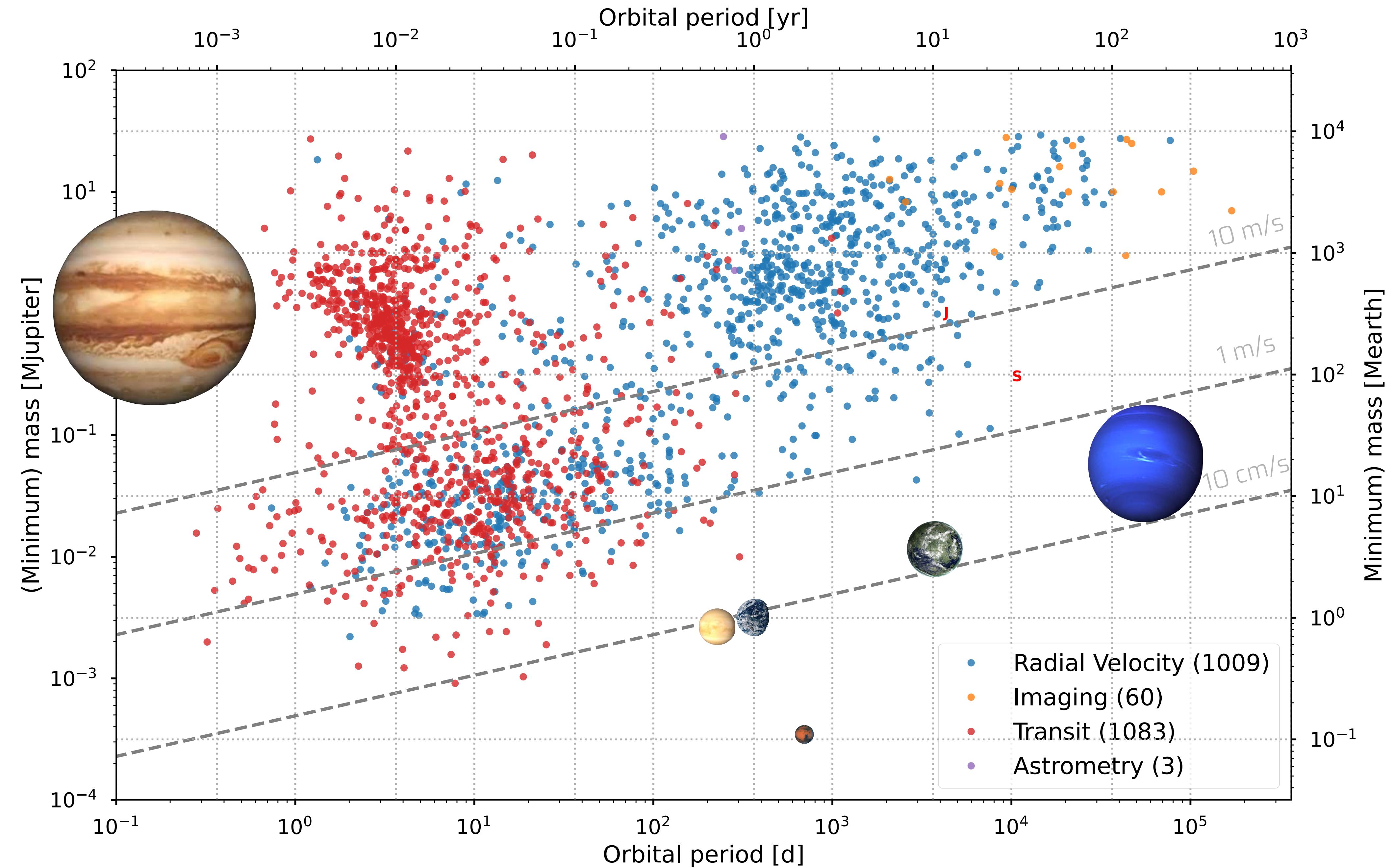
The classical approach to detecting signals in RV time series



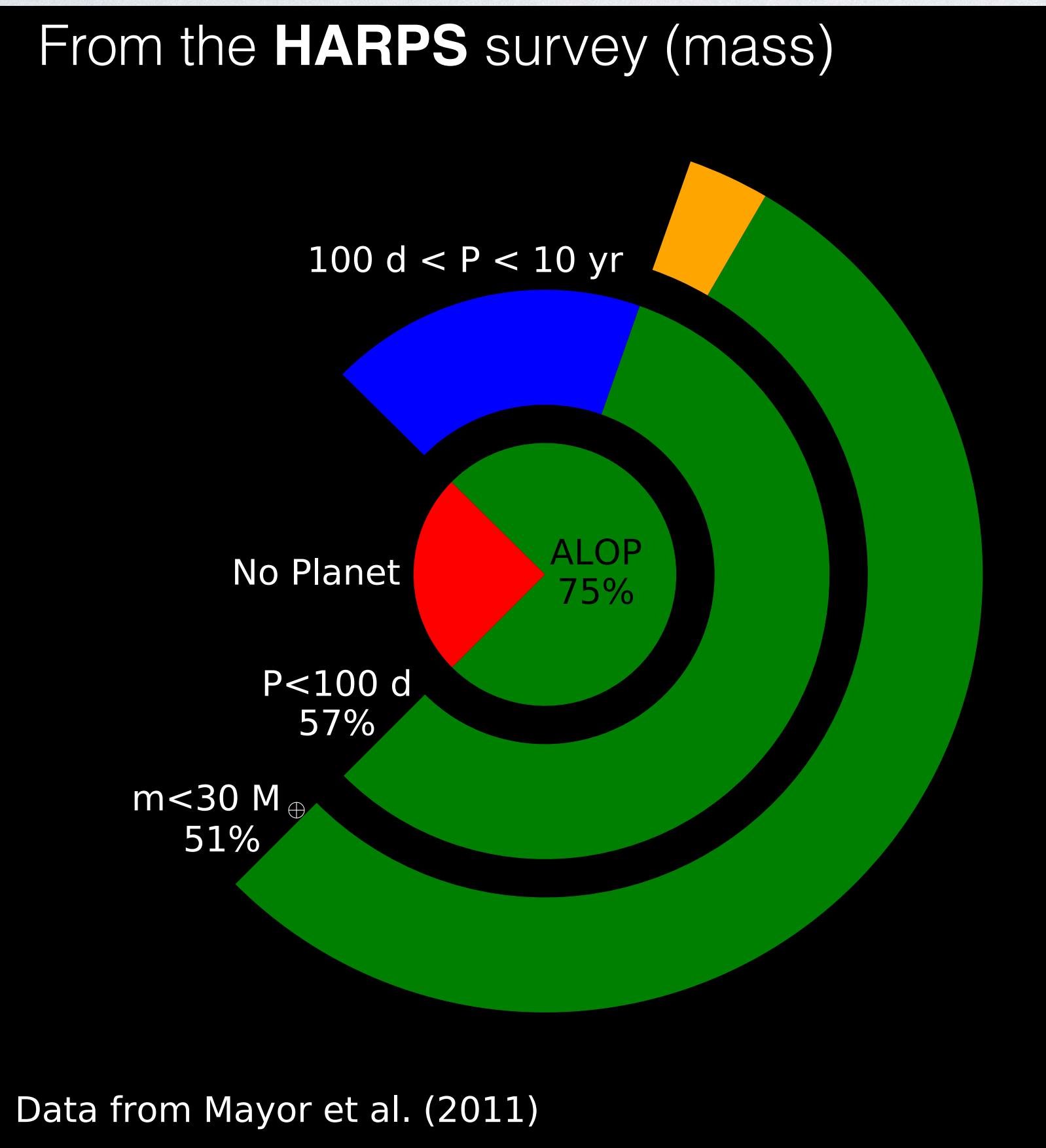
If significant, remove largest peak, which is now considered a real (planet) frequency and continue with the residuals.

Stop when no significant peak is present.

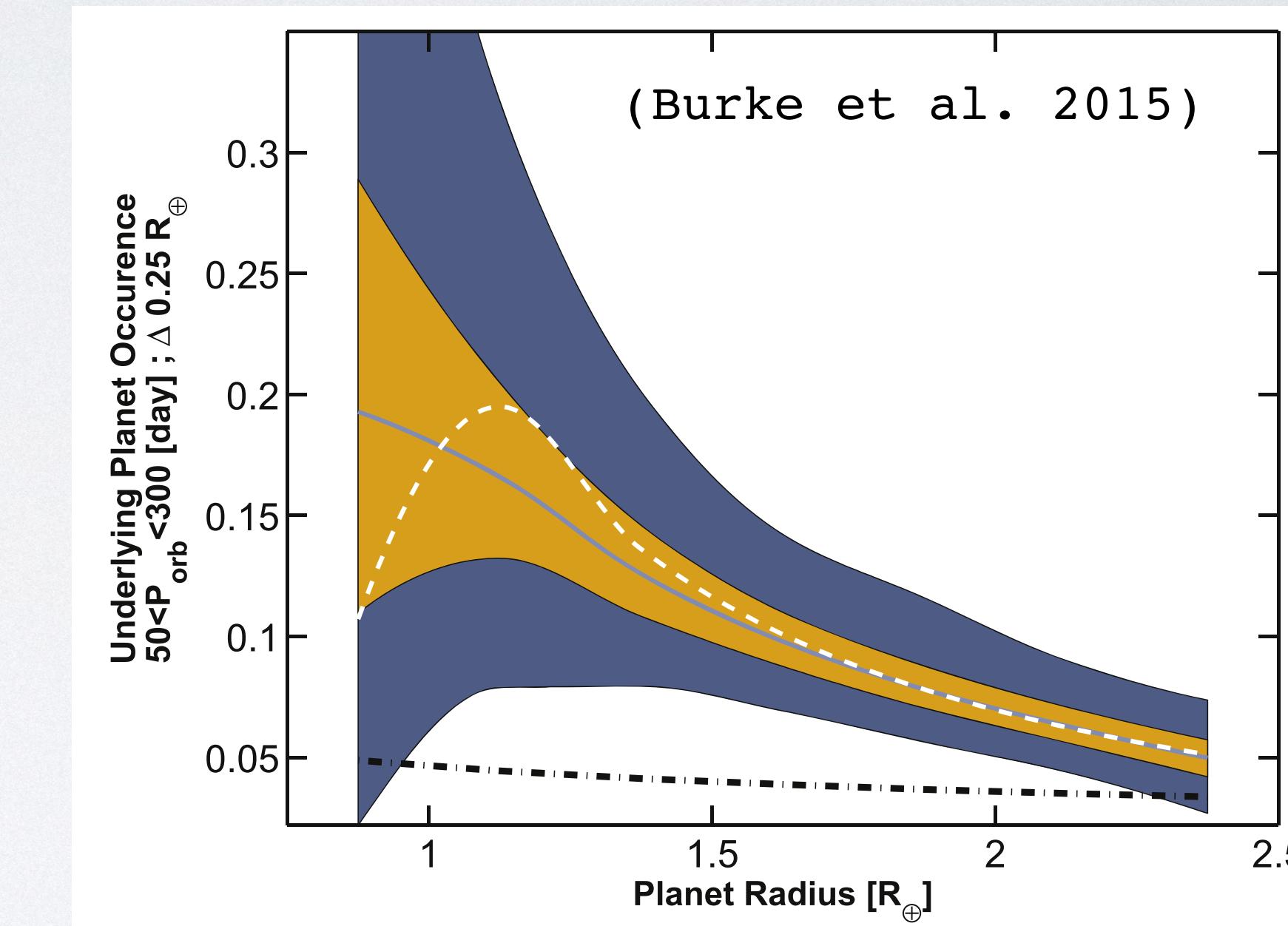




PLANET OCCURRENCE RATES



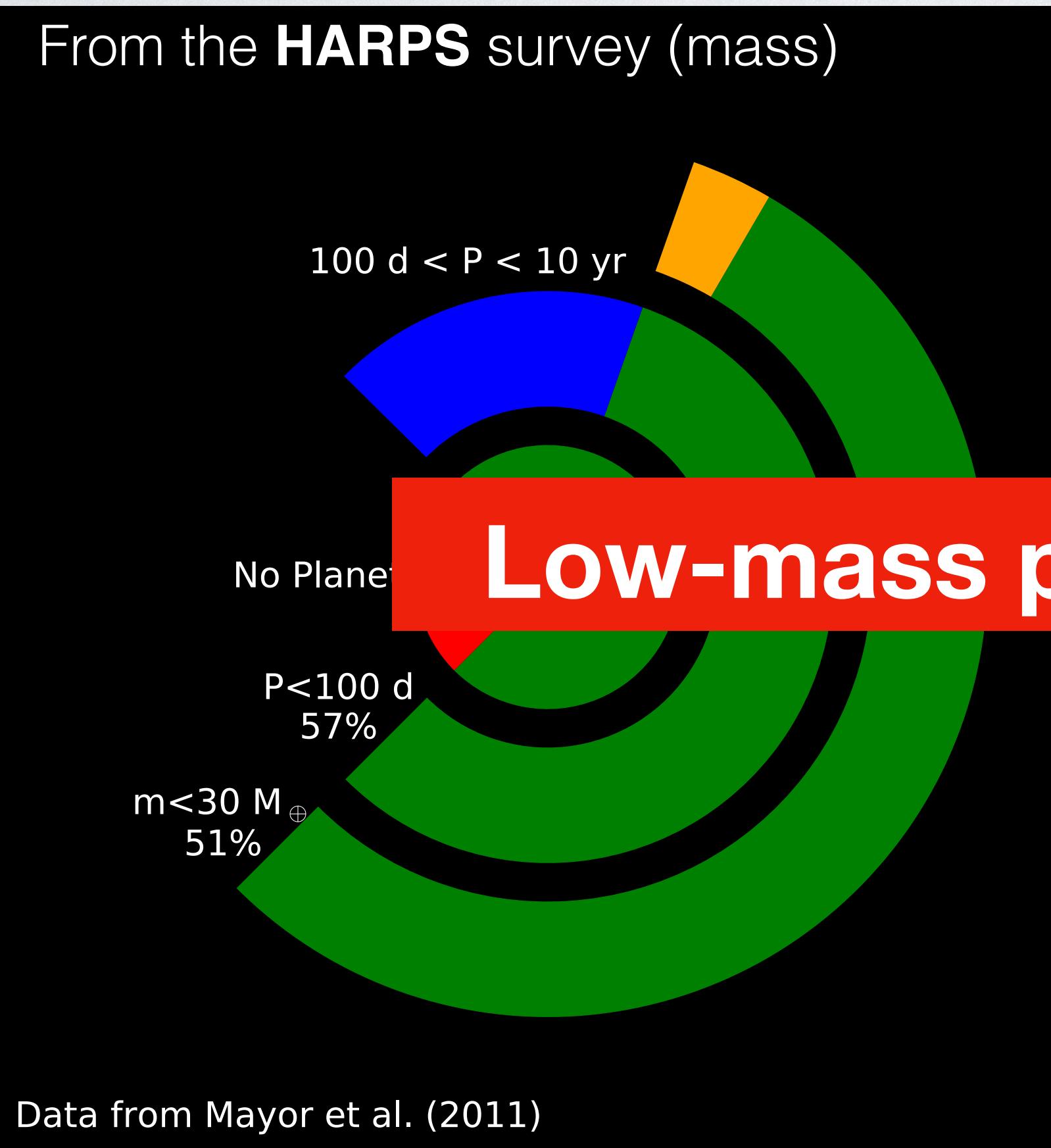
From the **Kepler** mission (radius)



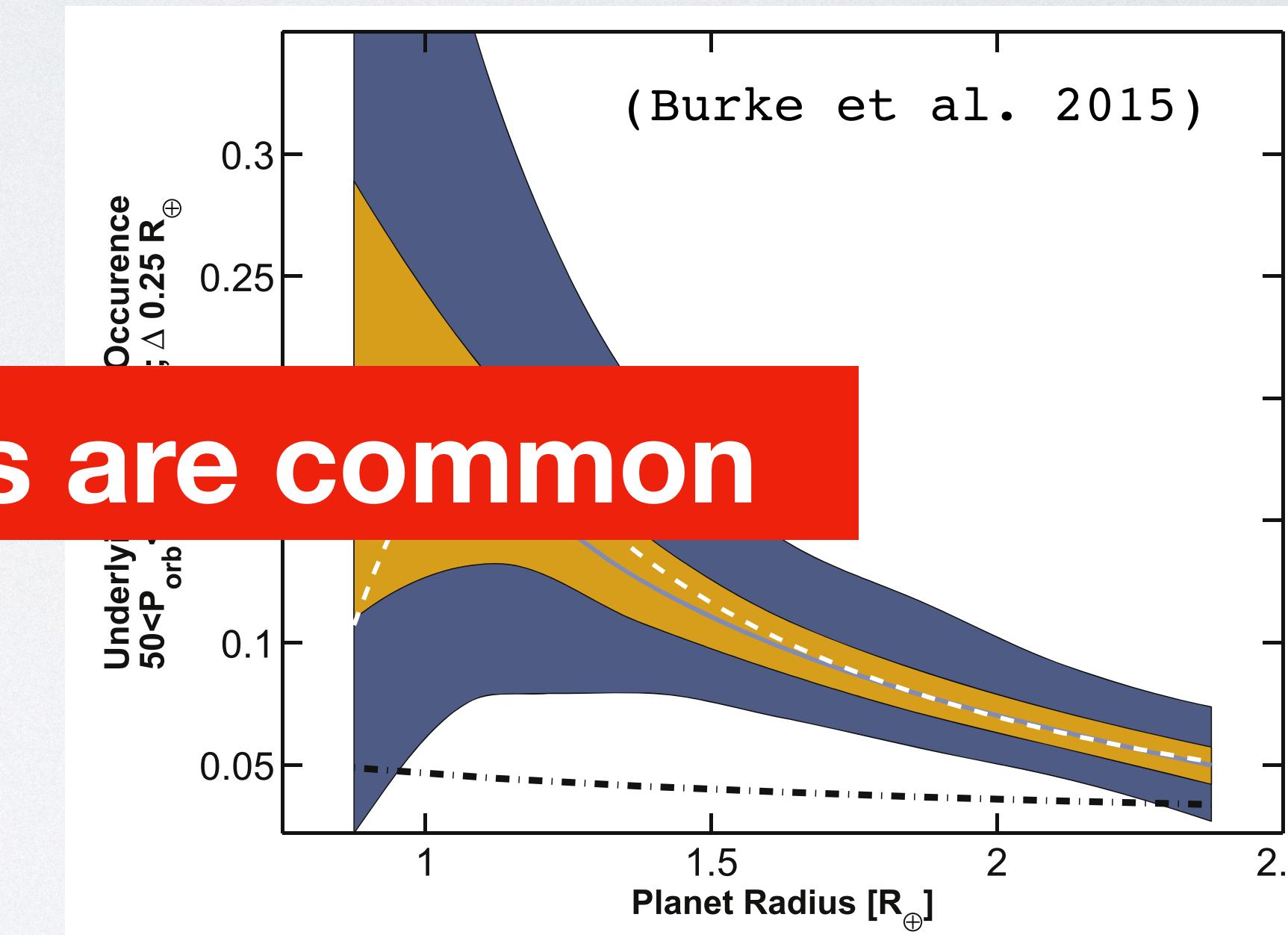
0.77 planets per star [0.49 - 1.3]

See also Youdin (2011), Howard et al. (2012b), Farr et al. (2014), among others

PLANET OCCURRENCE RATES



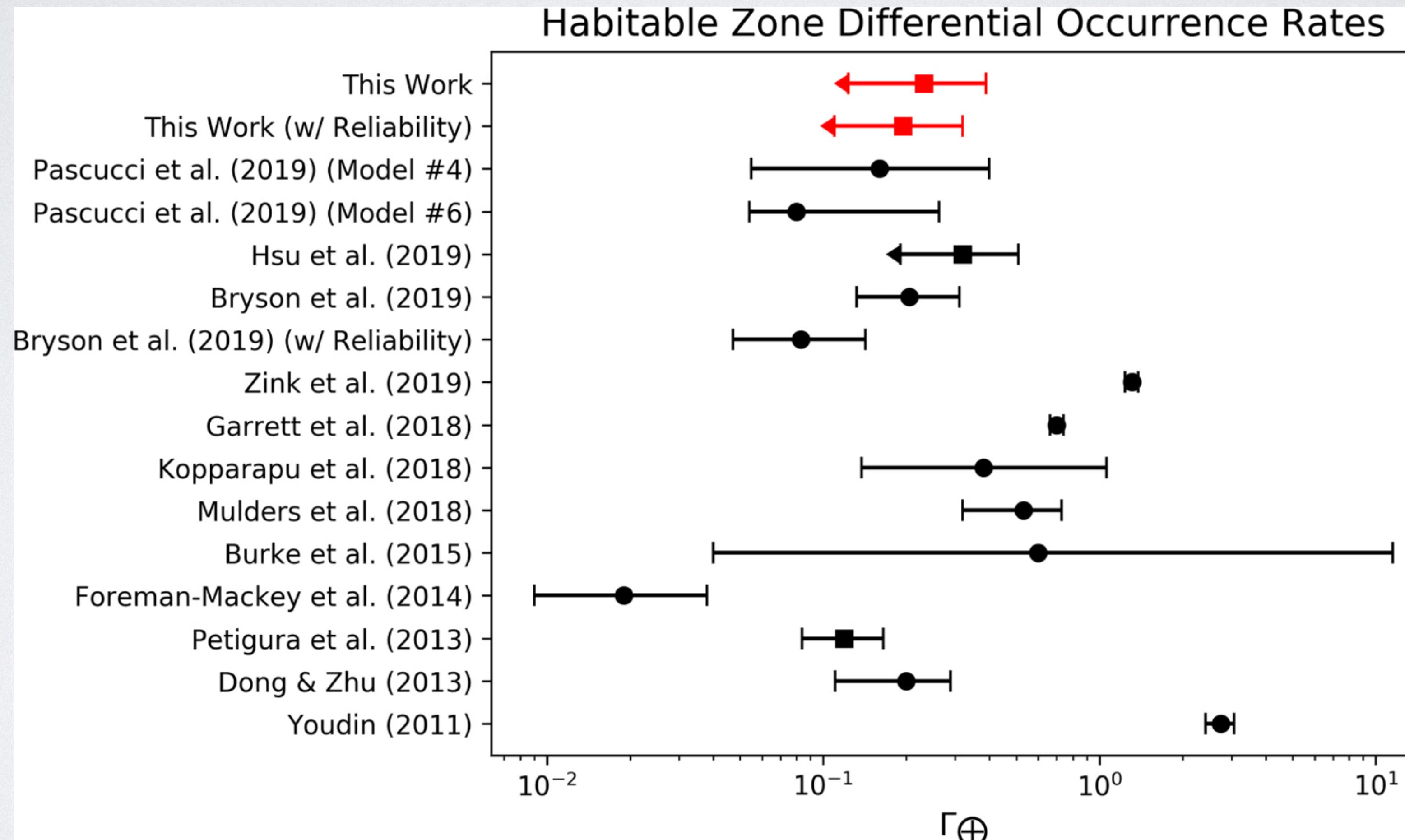
From the **Kepler** mission (radius)



0.77 planets per star [0.49 - 1.3]

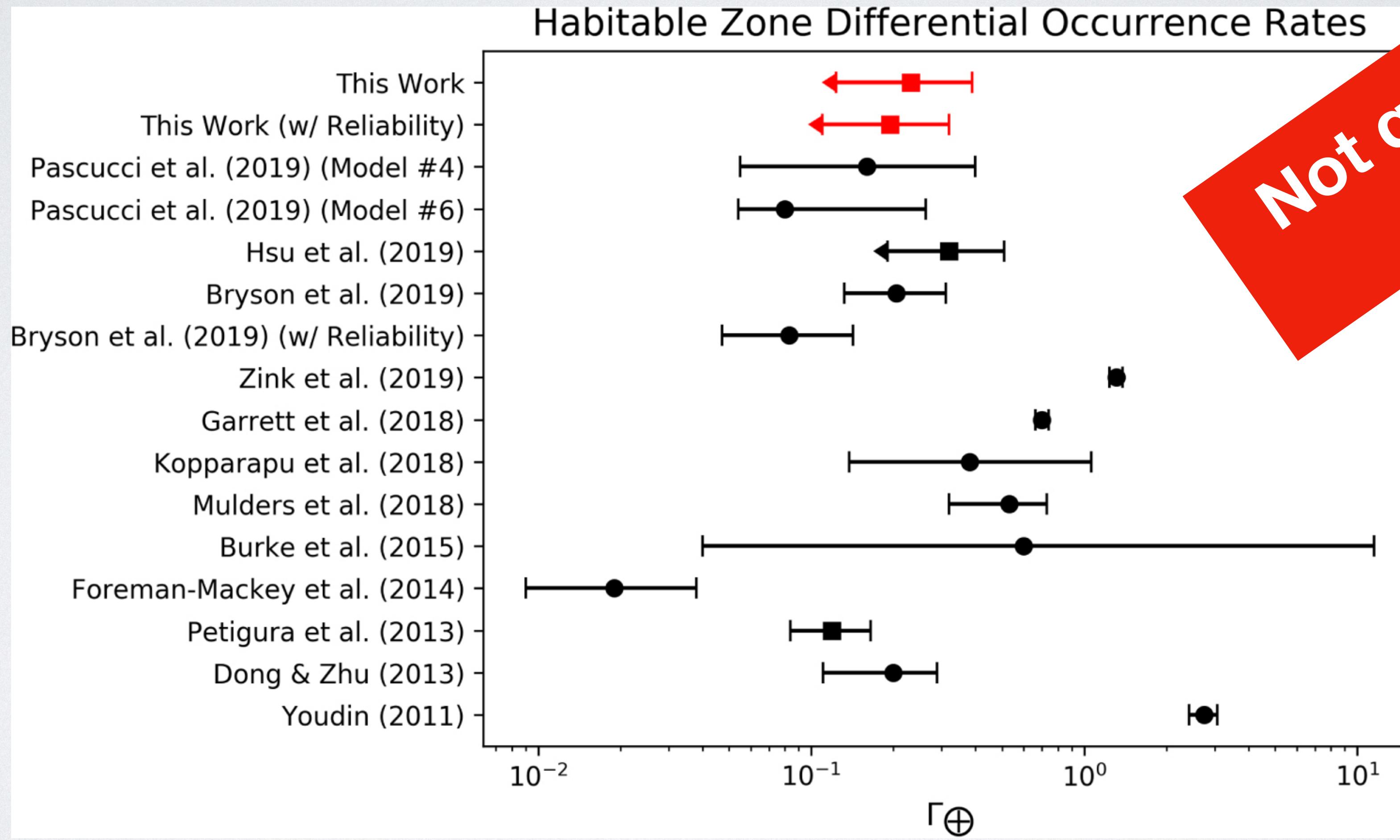
See also Youdin (2011), Howard et al. (2012b), Farr et al. (2014), among others

ETA EARTH, THE HOLY GRAIL

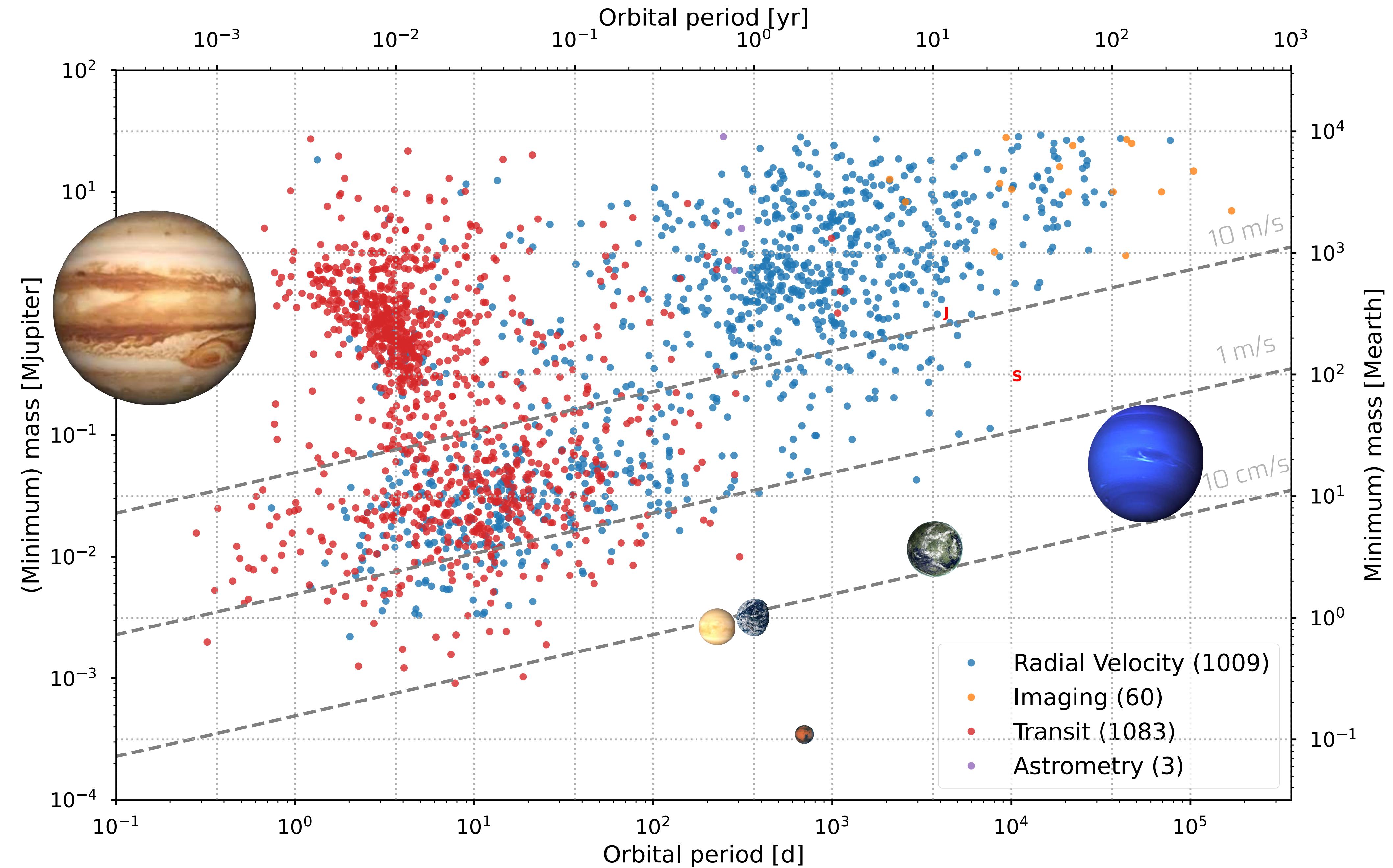


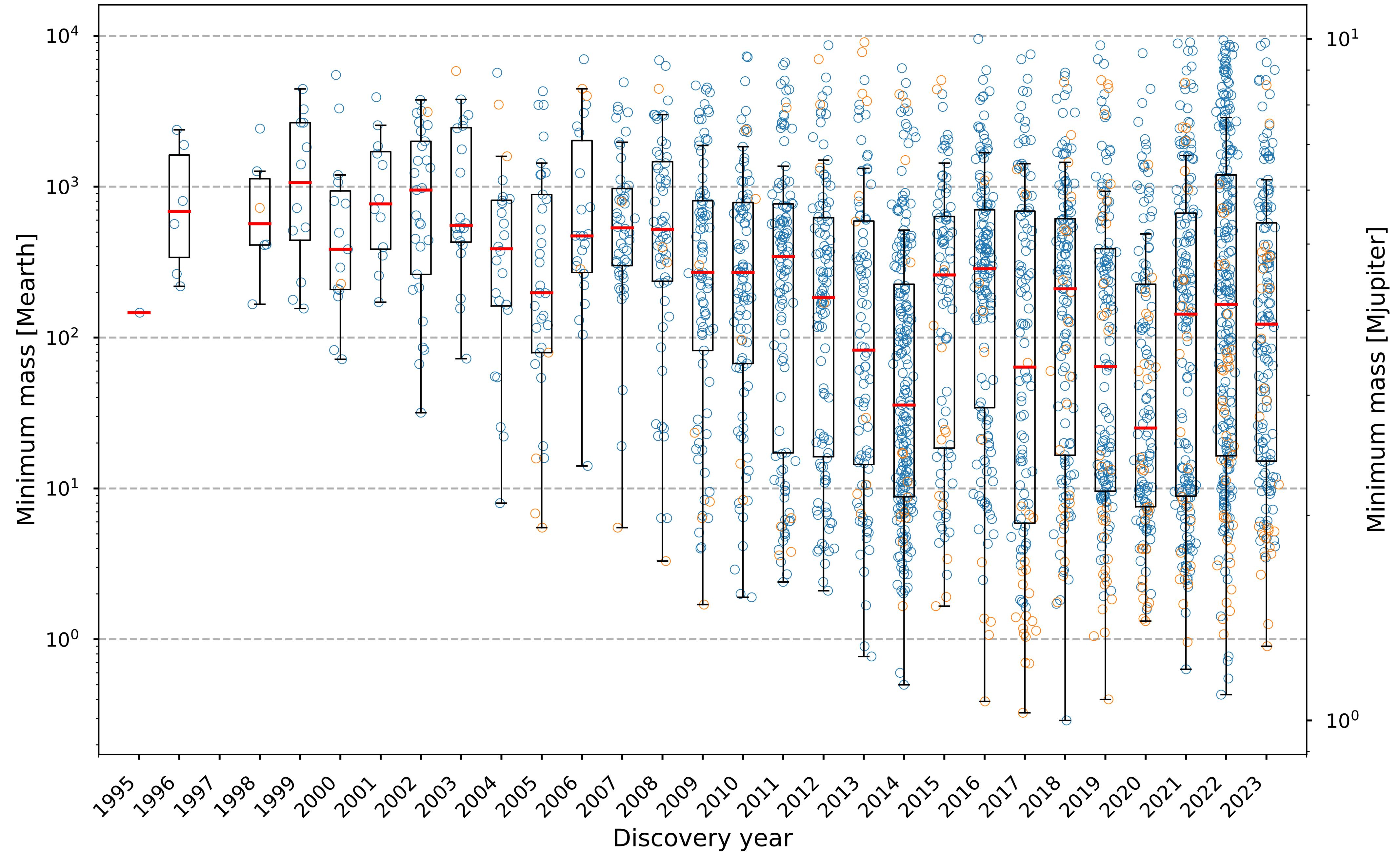
Kunimoto & Matthews (2020)

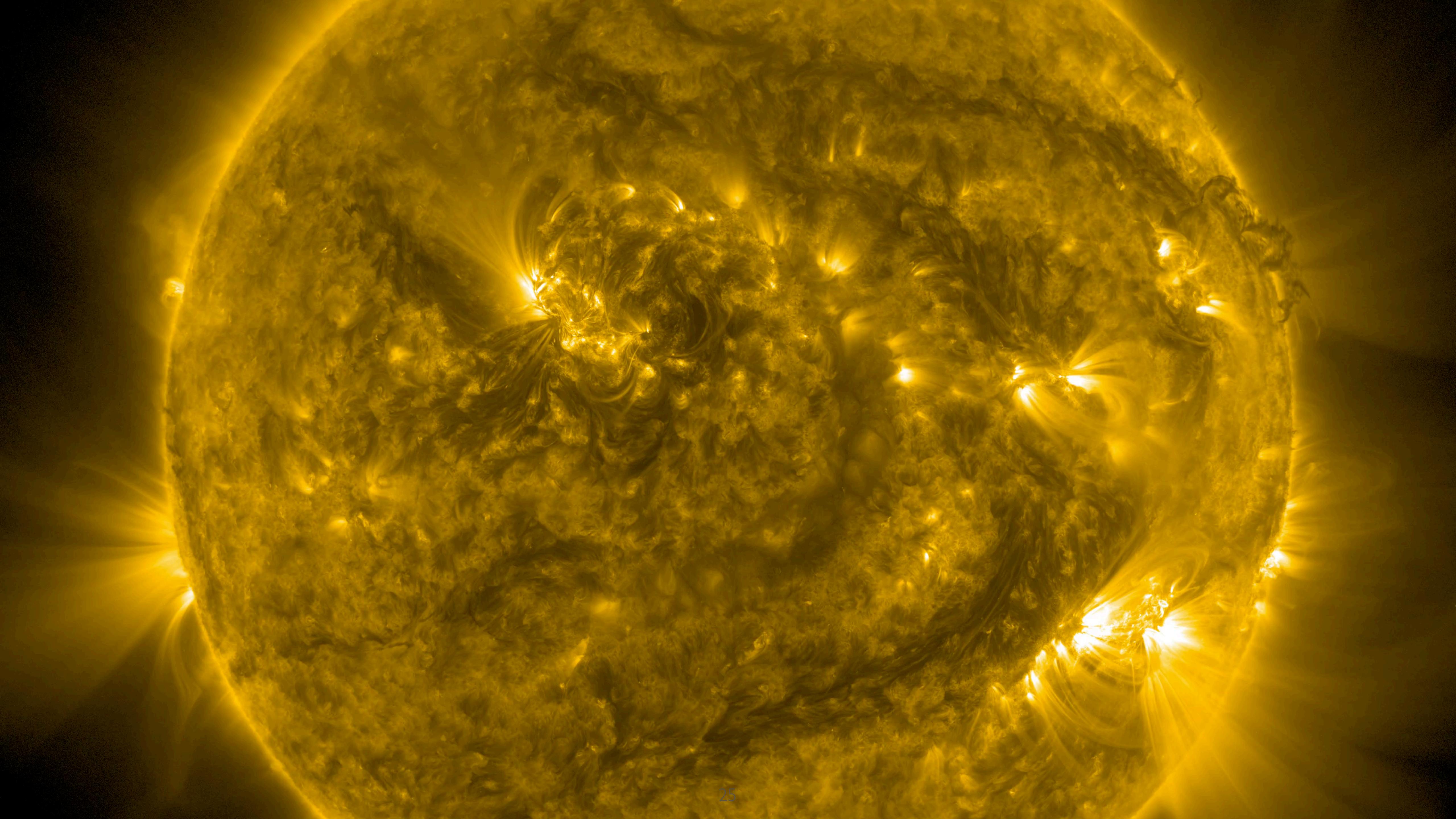
ETA EARTH, THE HOLY GRAIL

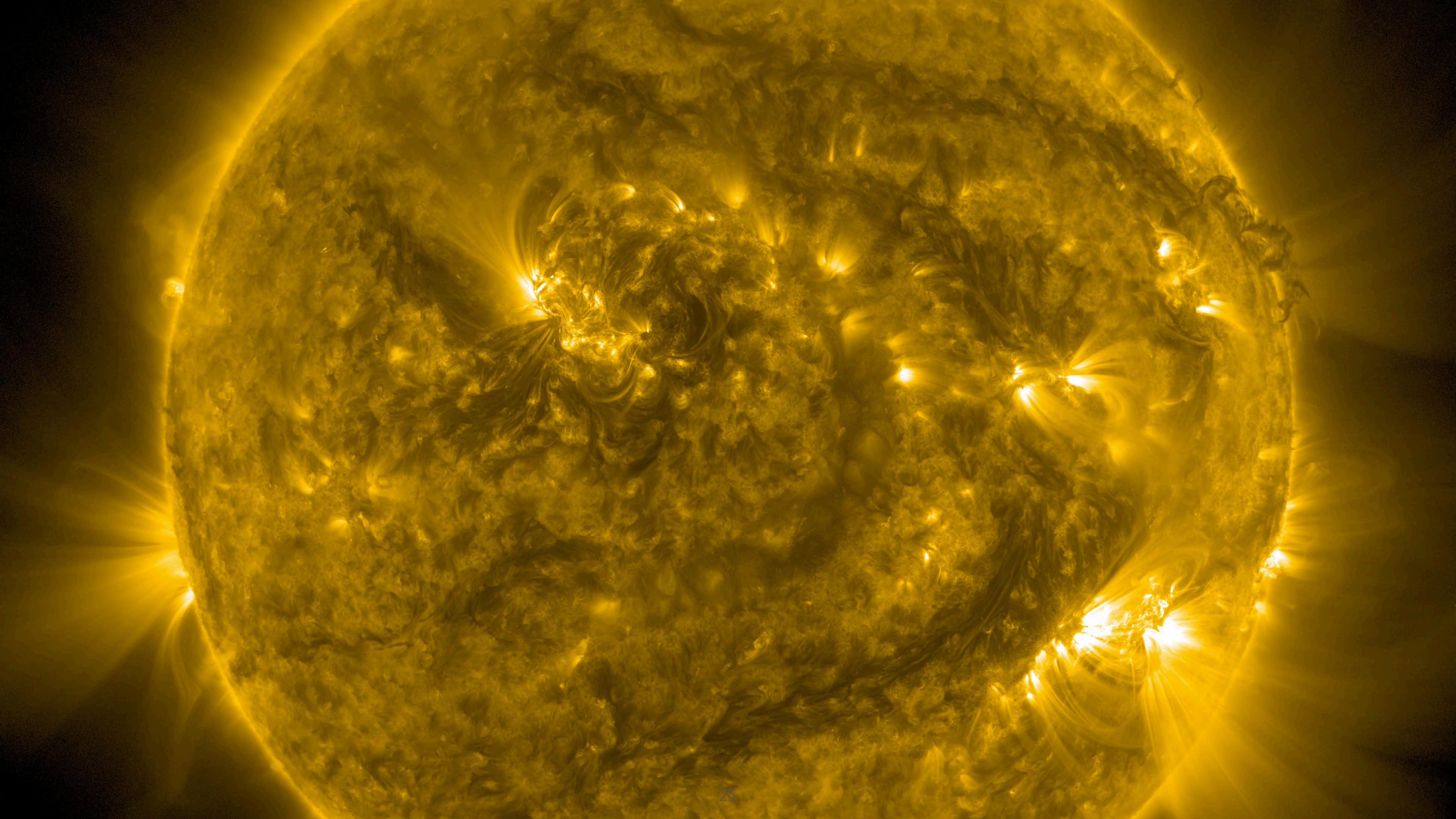


Kunimoto & Matthews (2020)

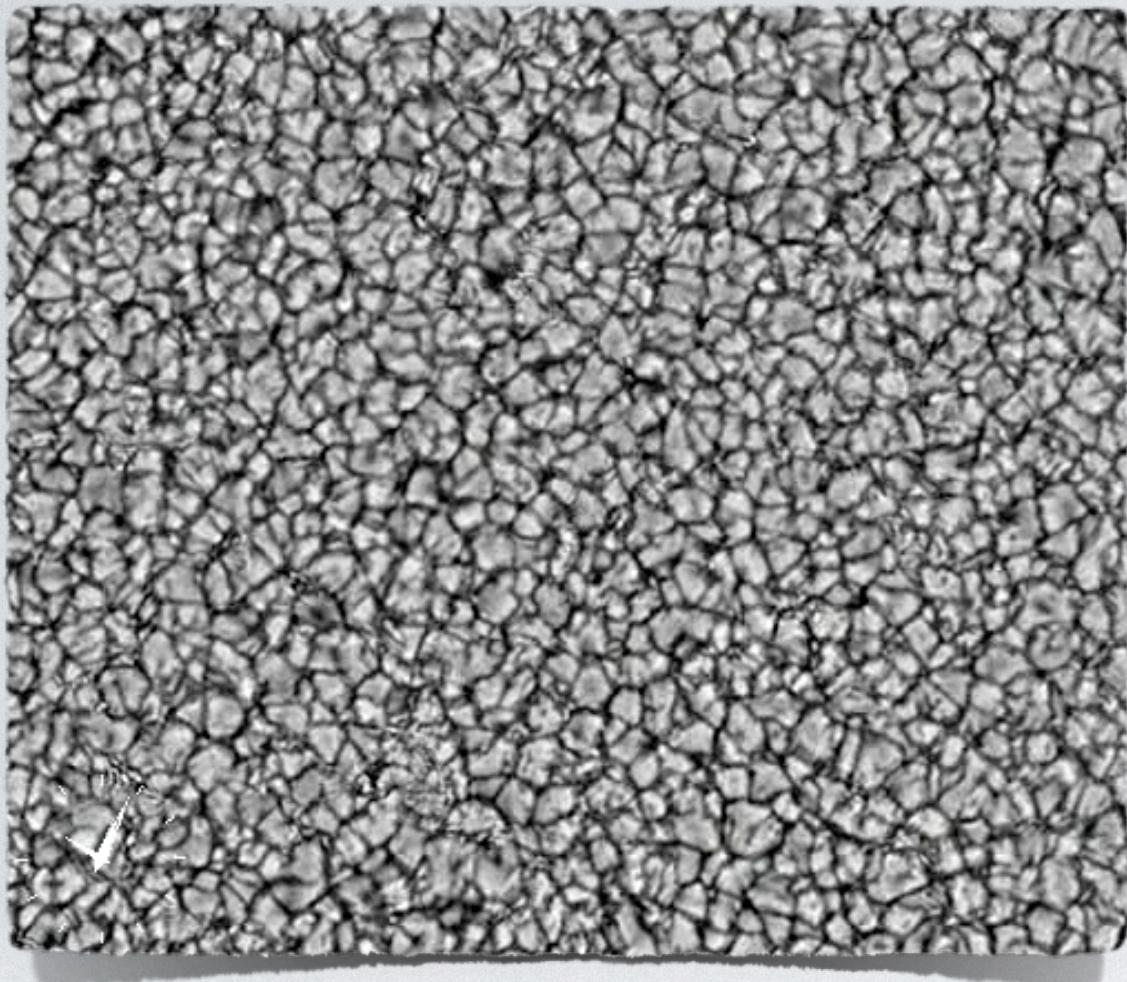




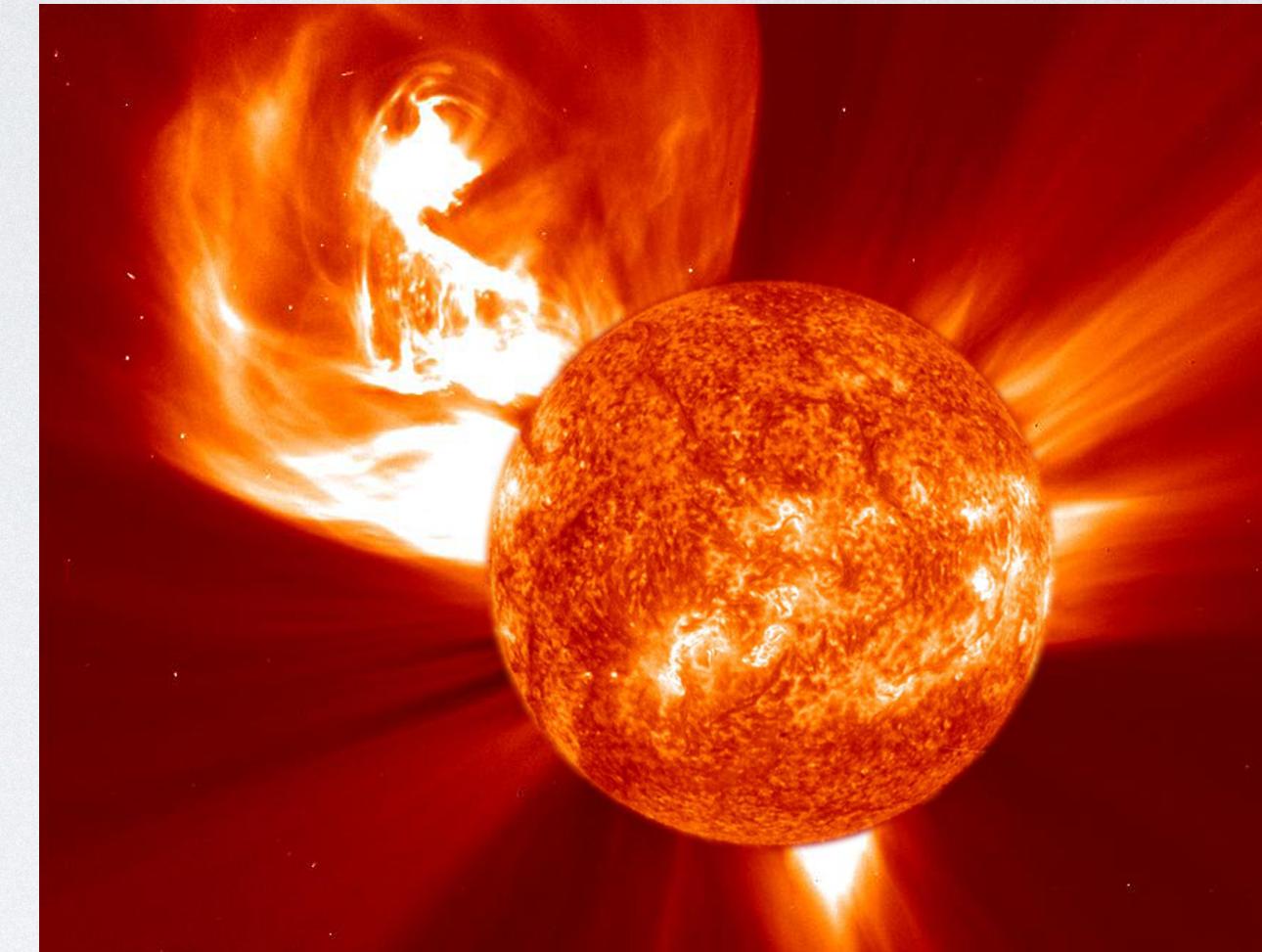




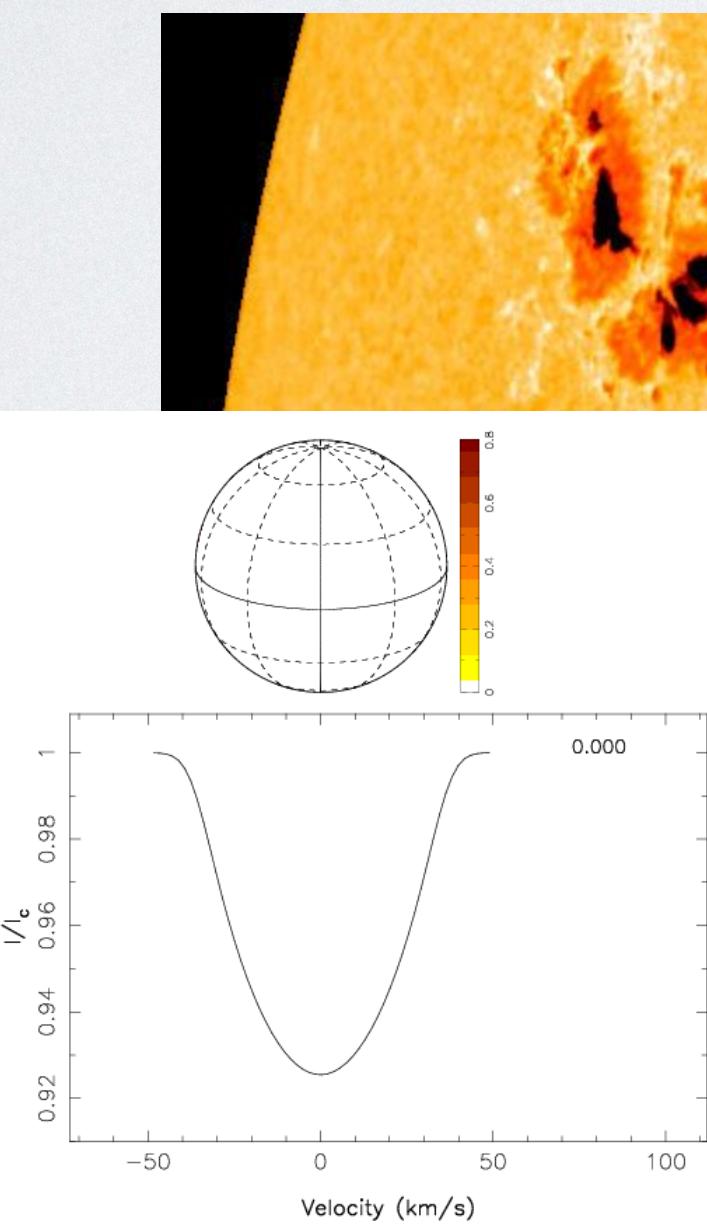
THE MANY FACES OF STELLAR ACTIVITY



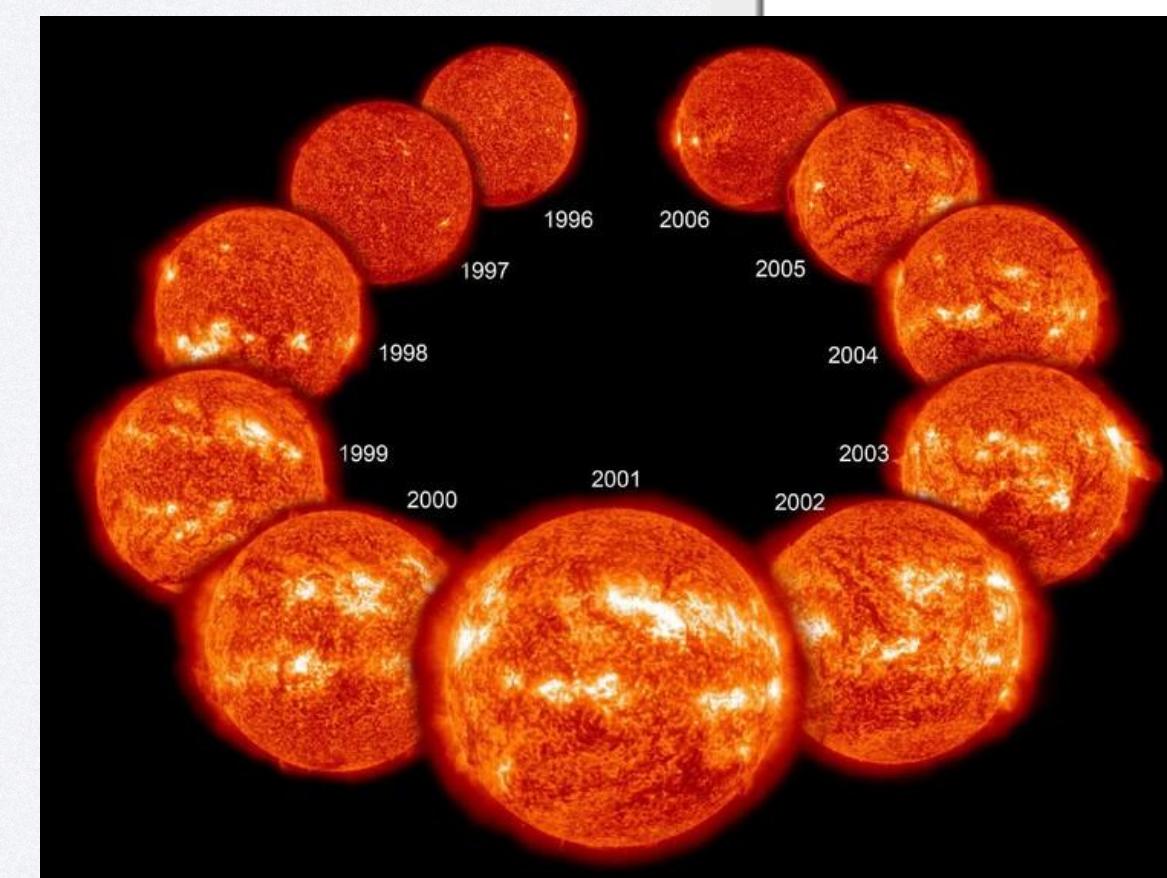
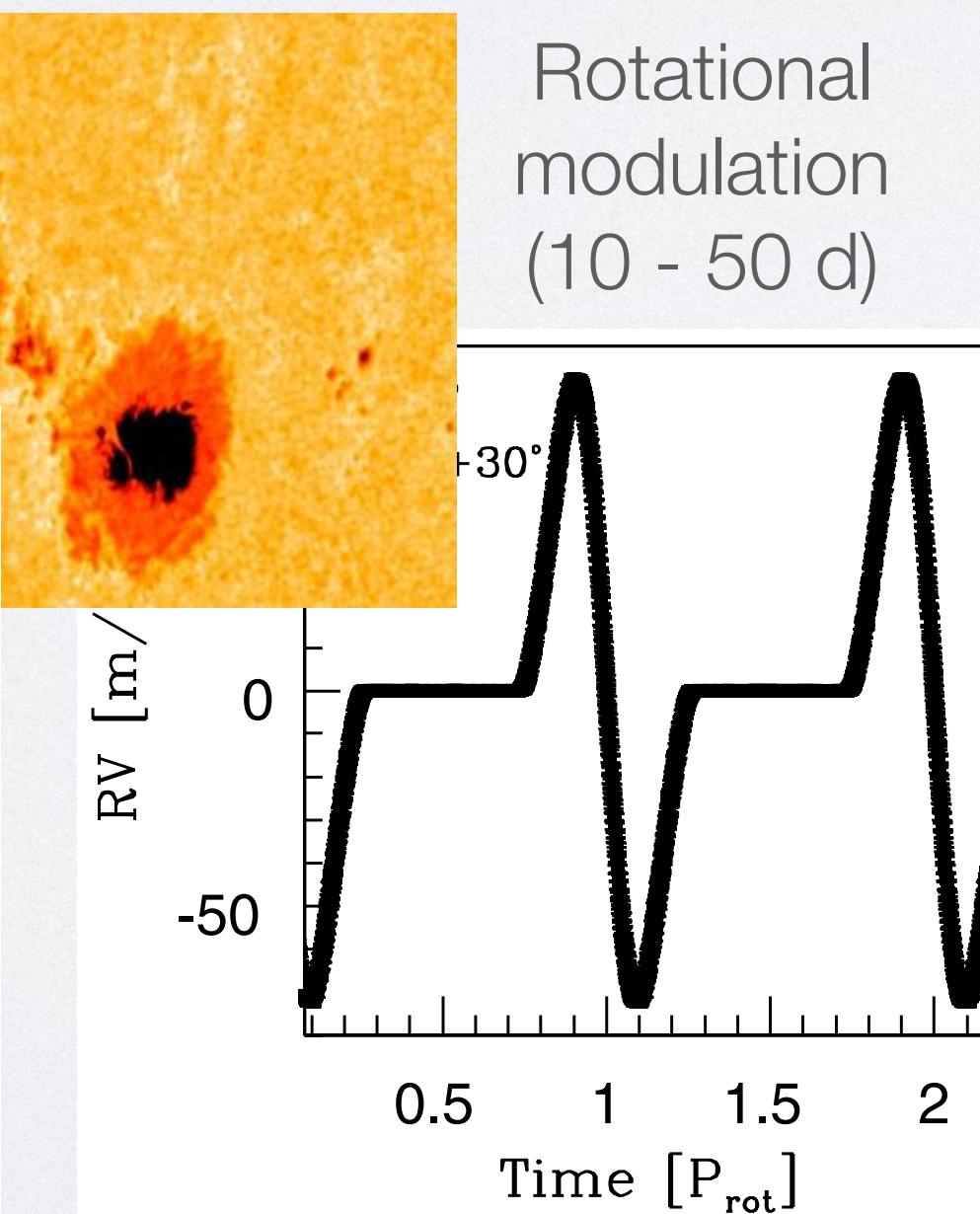
Granulation
(15 m - 2 d)



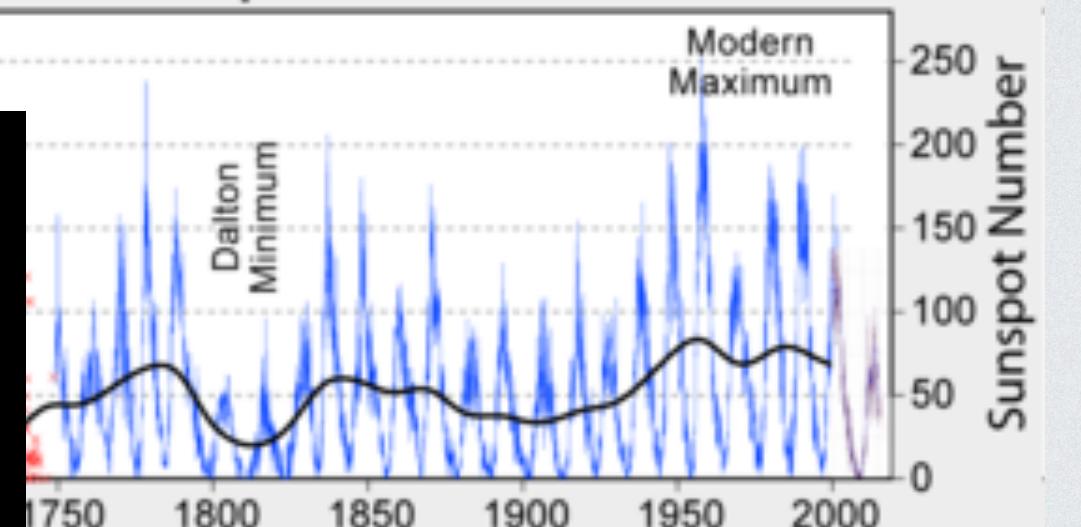
Flares & CMEs
(~ 1 h)



Rotational
modulation
(10 - 50 d)

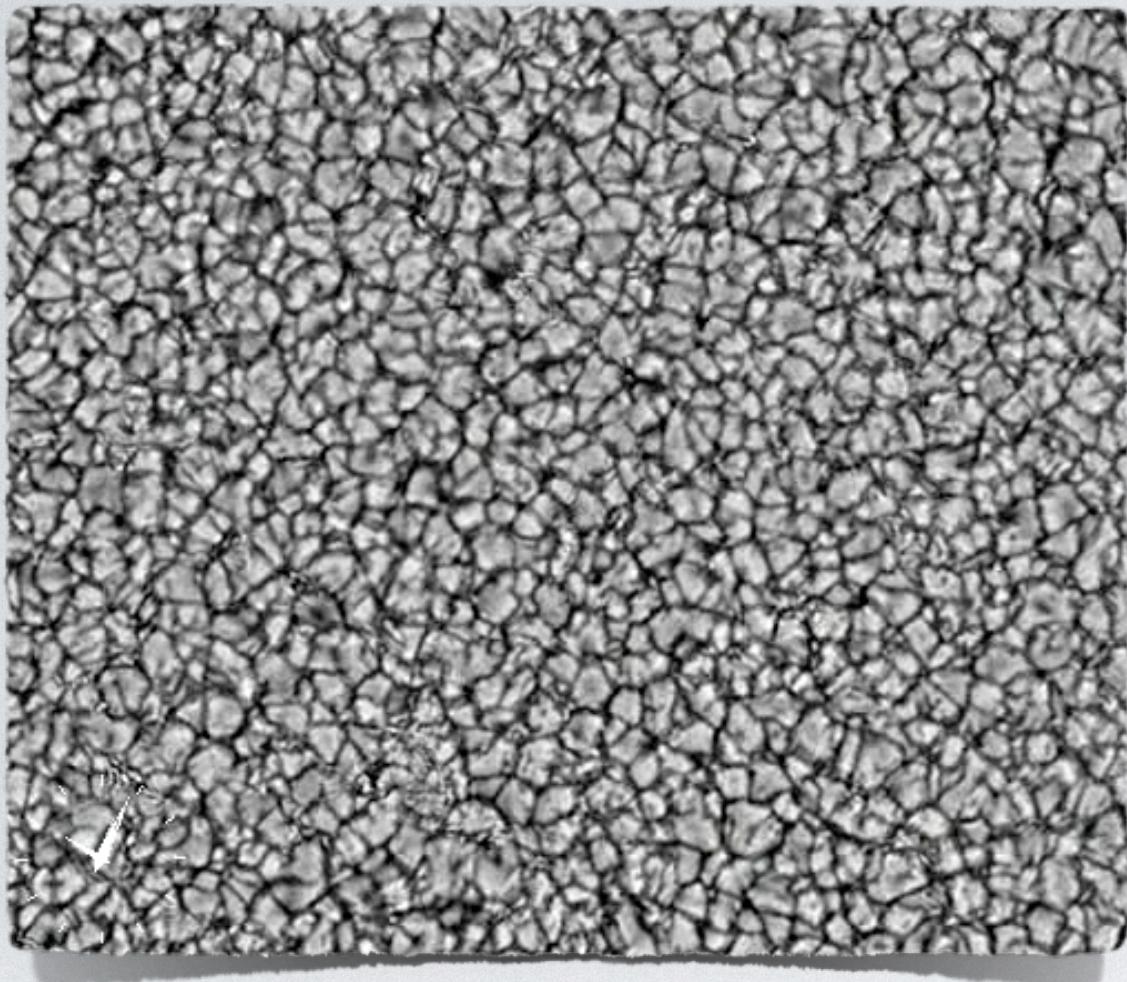


400 Years of Sunspot Observations

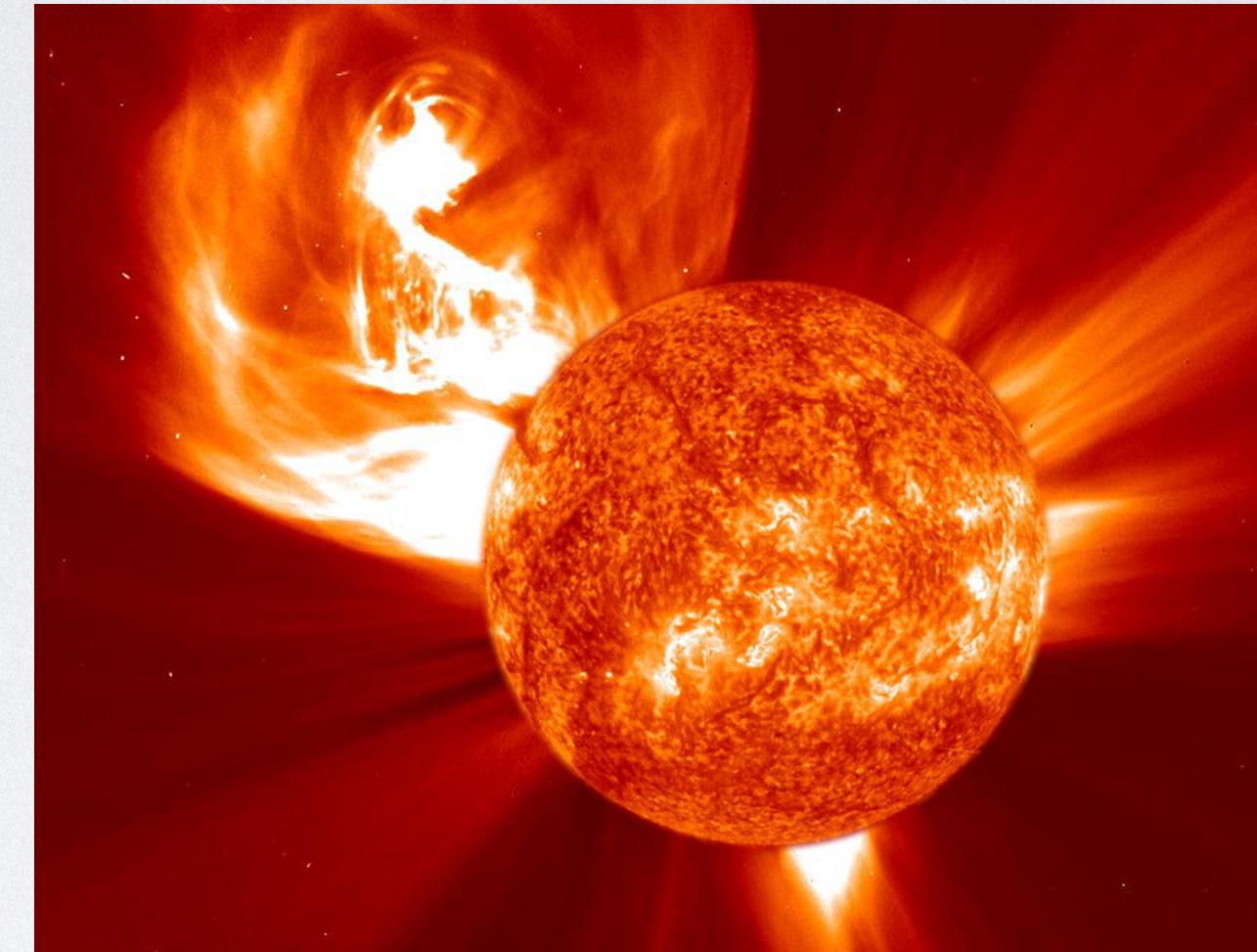


Activity cycles
(5 - 20 yr)

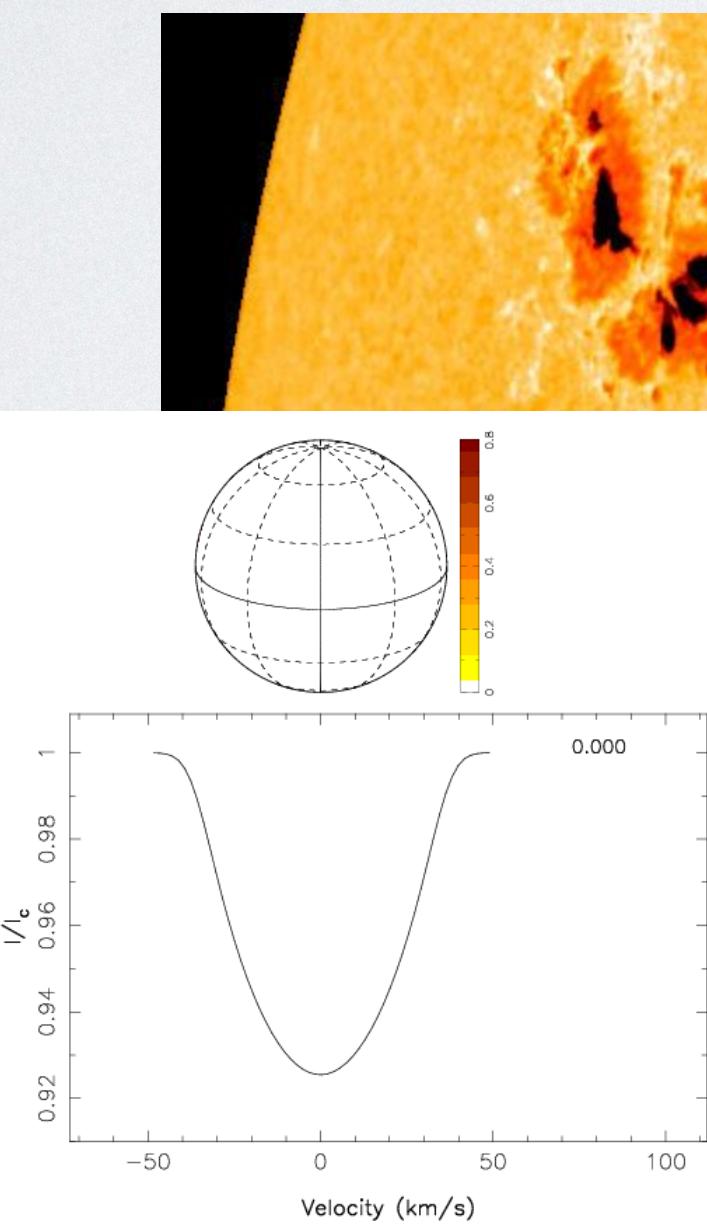
THE MANY FACES OF STELLAR ACTIVITY



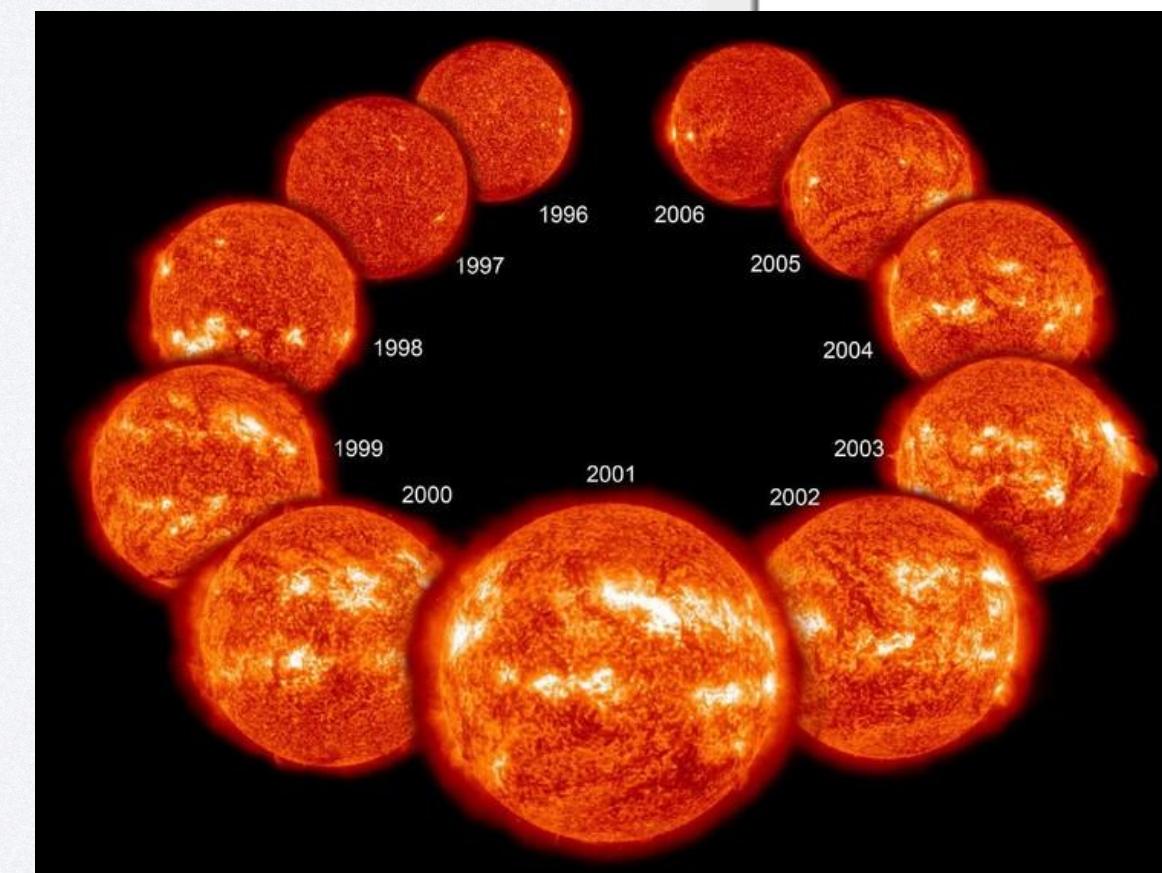
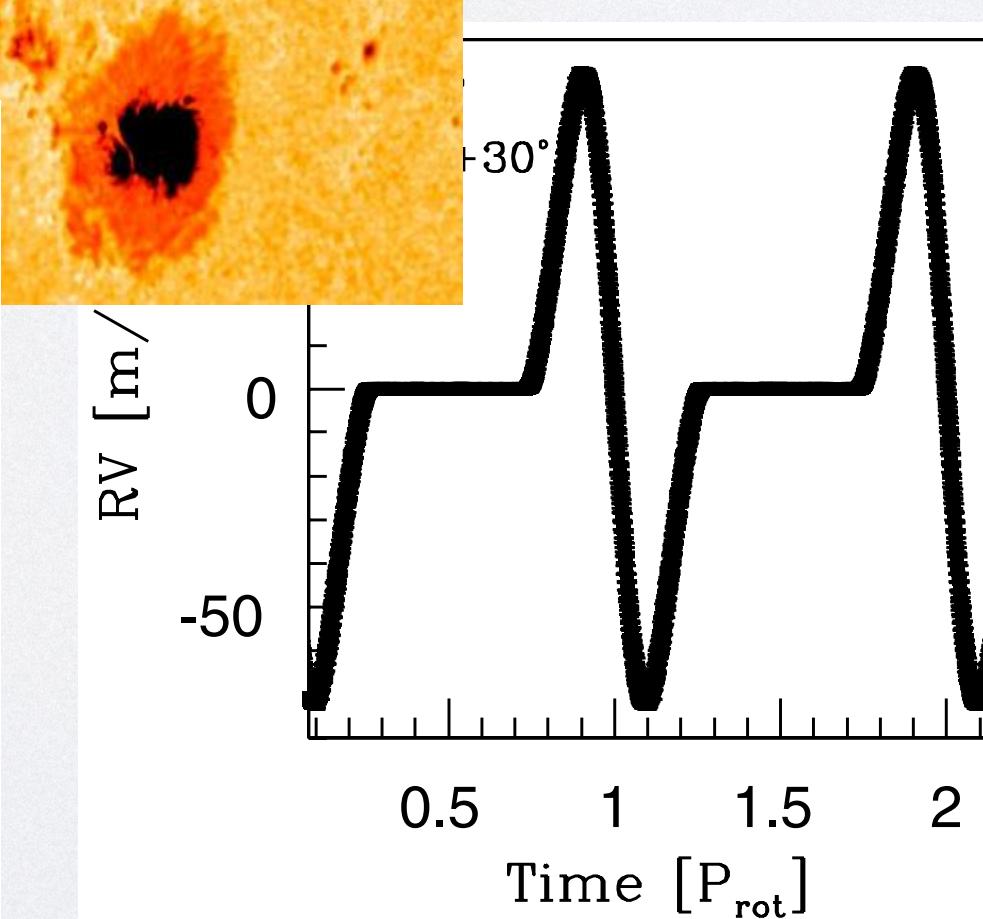
Granulation
(15 m - 2 d)



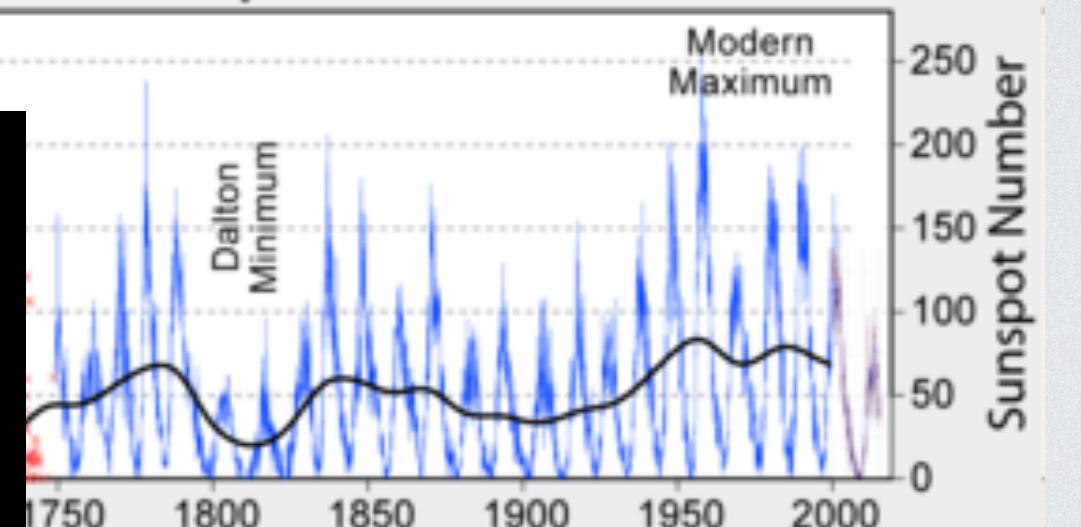
Flares & CMEs
(~ 1 h)



Rotational
modulation
(10 - 50 d)



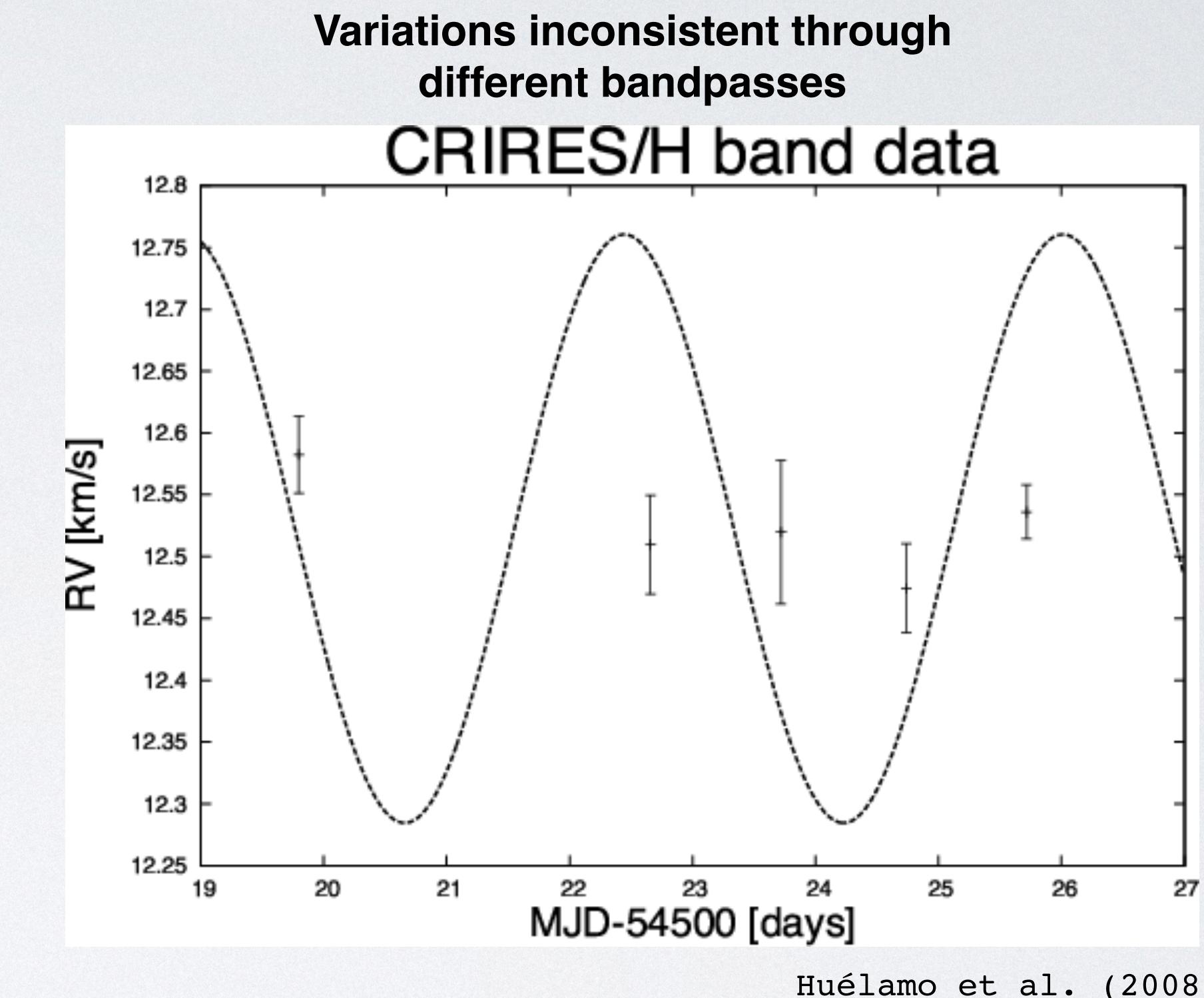
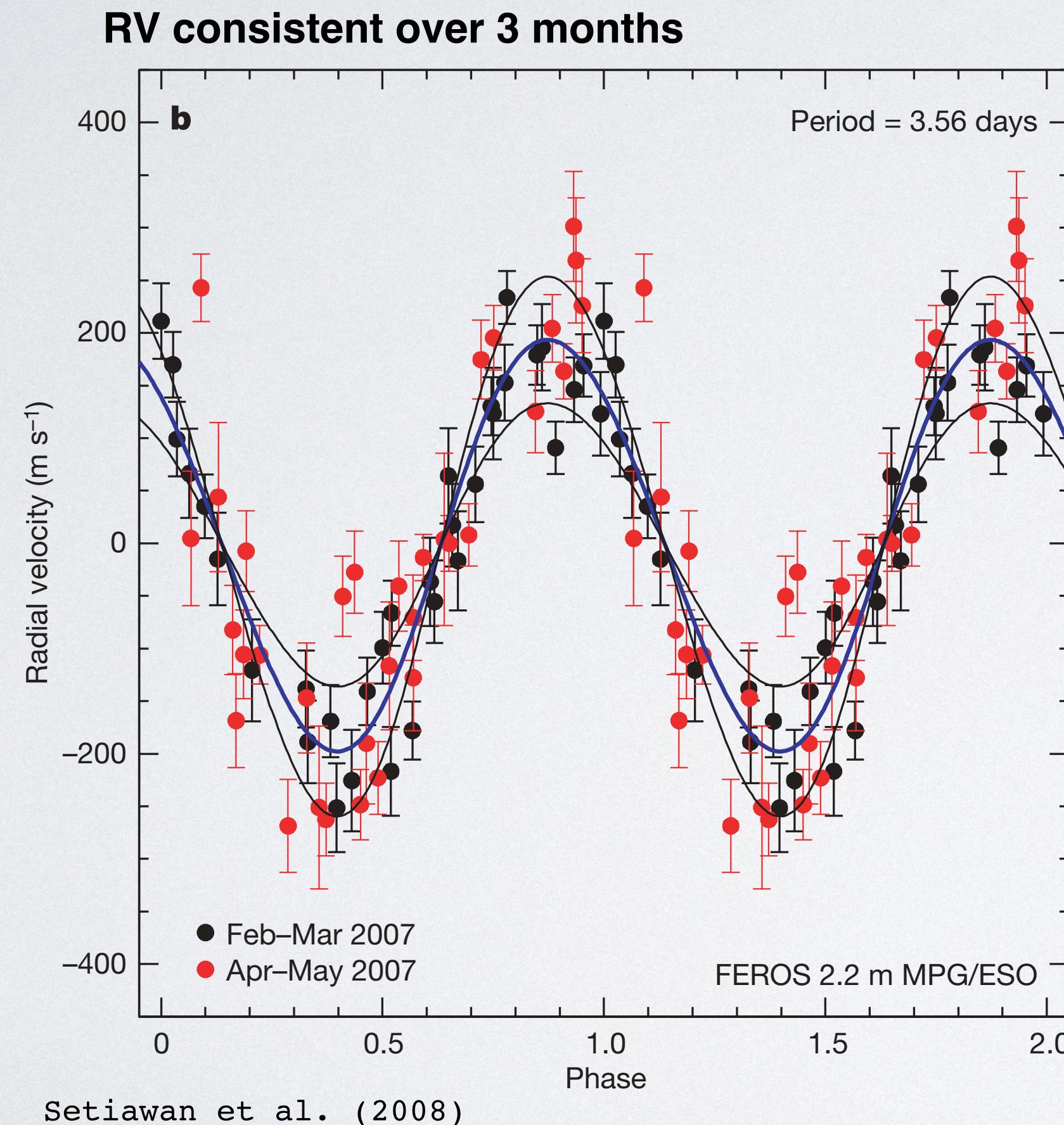
400 Years of Sunspot Observations



Activity cycles
(5 - 20 yr)

FALSE DETECTIONS

TW Hya has only 8-10 Myr of age.



A long-lived active region producing coherent RV effect over months.

NOISE MODELS

Astronomy & Astrophysics manuscript no. RV_challenge_paper_II_v4
October 16, 2018

©ESO 2018

Radial-Velocity Fitting Challenge ★

II. First results of the analysis of the data set

X. Dumusque^{1,2} **, F. Borsa³, M. Damasso⁴, R. Díaz¹, P. C. Gregory⁵, N.C. Hara⁶, A. Hatzes⁷, V. Rajpaul⁸, M. Tuomi⁹, S. Aigrain⁸, G. Anglada-Escudé^{9,10}, A.S. Bonomo⁴, G. Boué⁶, F. Dauvergne⁶, G. Frustagli³, P. Giacobbe⁴, R. D. Haywood², H. R. A. Jones⁹, M. Pinamonti^{11,12}, E. Poretti³, M. Rainer³, D. Ségransan¹, A. Sozzetti⁴, and S. Udry¹

"The most efficient methods to recover planetary signals take into account the different activity indicators, use **red-noise models** to account for stellar RV signals and a **Bayesian framework** to provide model comparison in a robust statistical approach."

$$\mathbf{v} = f(\mathbf{x}|\boldsymbol{\theta}) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

NOISE MODELS

Astronomy & Astrophysics manuscript no. RV_challenge_paper_II_v4
October 16, 2018

©ESO 2018

Radial-Velocity Fitting Challenge ★

II. First results of the analysis of the data set

X. Dumusque^{1,2} **, F. Borsa³, M. Damasso⁴, R. Díaz¹, P. C. Gregory⁵, N.C. Hara⁶, A. Hatzes⁷, V. Rajpaul⁸, M. Tuomi⁹, S. Aigrain⁸, G. Anglada-Escudé^{9,10}, A.S. Bonomo⁴, G. Boué⁶, F. Dauvergne⁶, G. Frustagli³, P. Giacobbe⁴, R. D. Haywood², H. R. A. Jones⁹, M. Pinamonti^{11,12}, E. Poretti³, M. Rainer³, D. Ségransan¹, A. Sozzetti⁴, and S. Udry¹

"The most efficient methods to recover planetary signals take into account the different activity indicators, use **red-noise models** to account for stellar RV signals and a **Bayesian framework** to provide model comparison in a robust statistical approach."

$$\mathbf{v} = f(\mathbf{x}|\boldsymbol{\theta}) + \epsilon$$

~~$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$~~

NOISE MODELS

Astronomy & Astrophysics manuscript no. RV_challenge_paper_II_v4
October 16, 2018

©ESO 2018

Radial-Velocity Fitting Challenge *

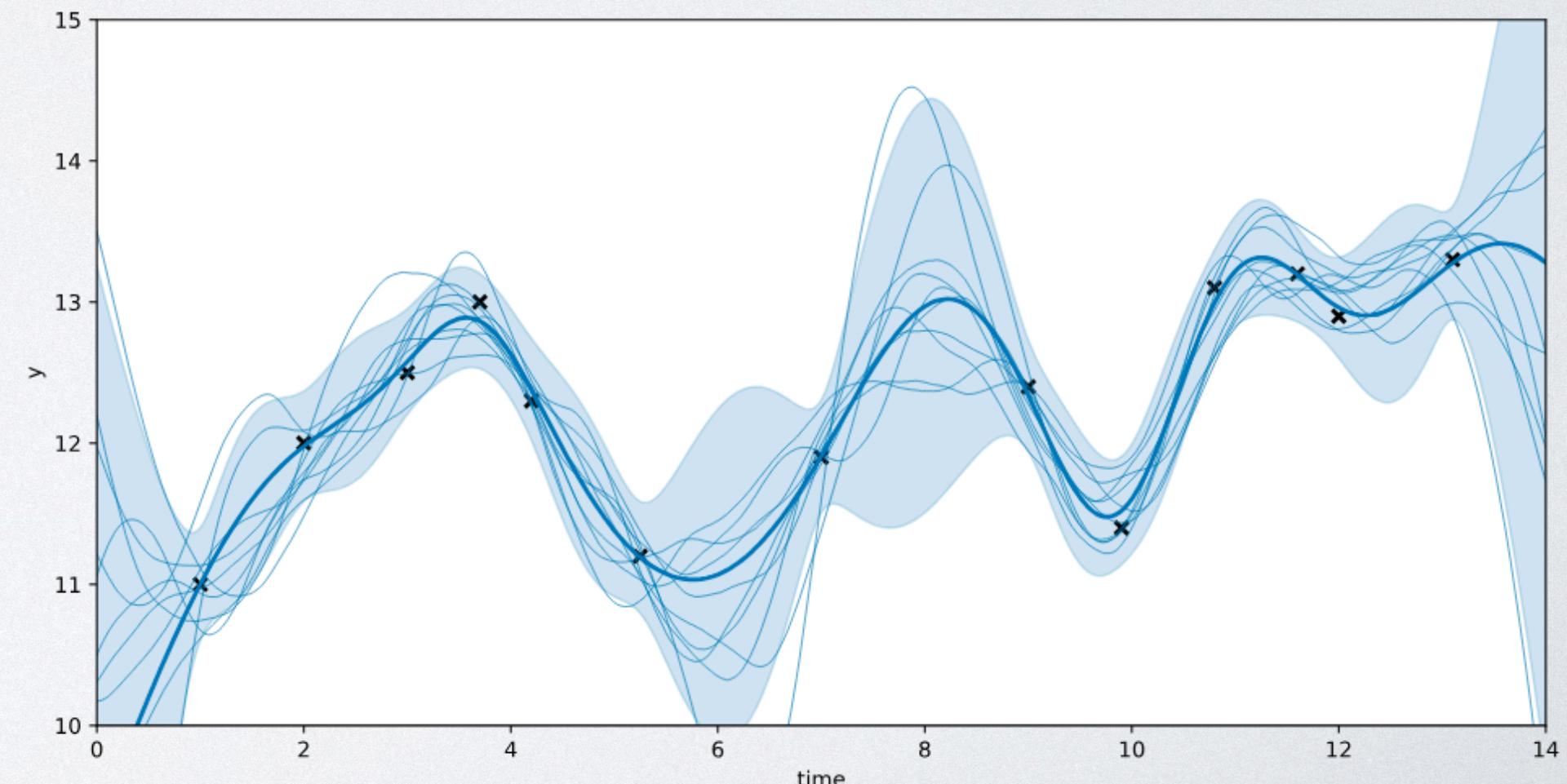
II. First results of the analysis of the data set

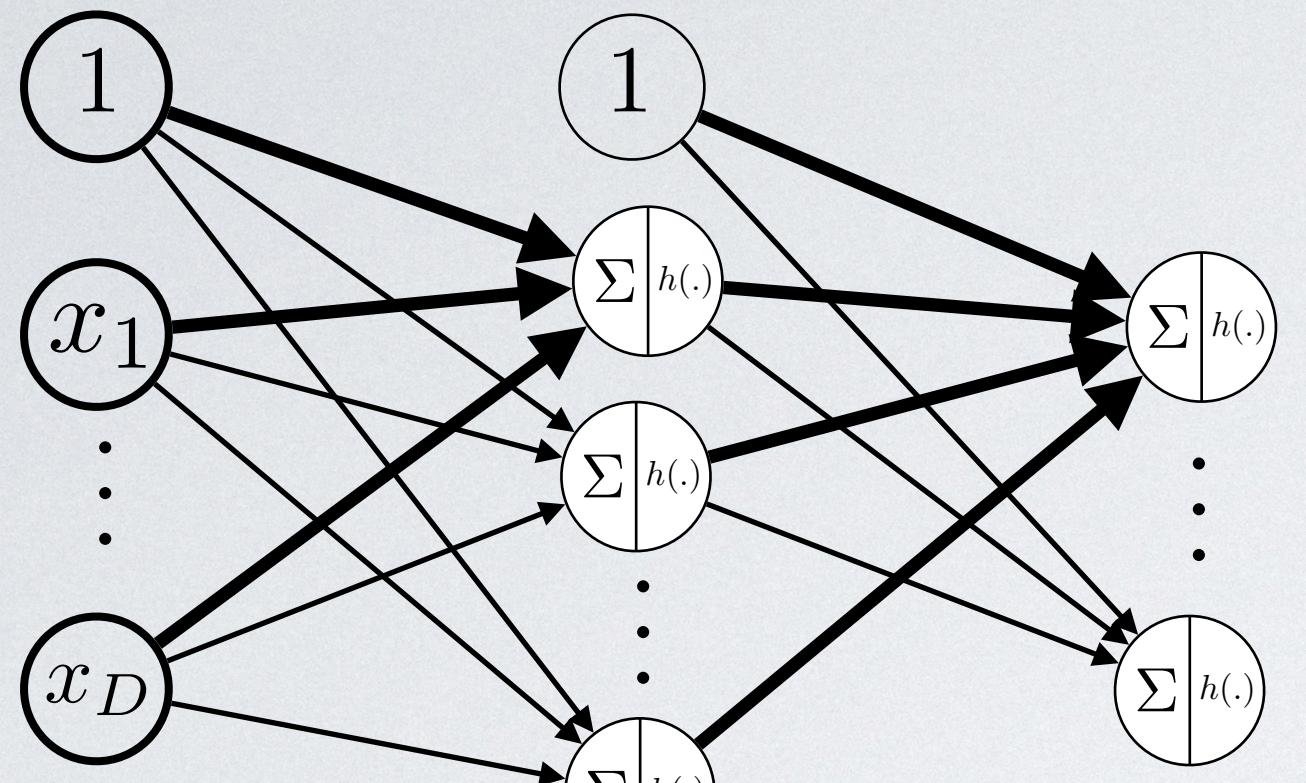
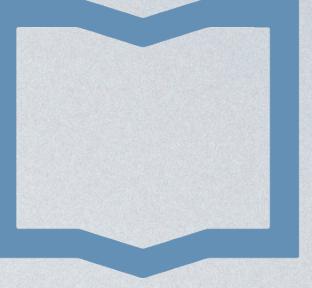
X. Dumusque^{1,2} **, F. Borsa³, M. Damasso⁴, R. Díaz¹, P. C. Gregory⁵, N.C. Hara⁶, A. Hatzes⁷, V. Rajpaul⁸, M. Tuomi⁹, S. Aigrain⁸, G. Anglada-Escudé^{9,10}, A.S. Bonomo⁴, G. Boué⁶, F. Dauvergne⁶, G. Frustagli³, P. Giacobbe⁴, R. D. Haywood², H. R. A. Jones⁹, M. Pinamonti^{11,12}, E. Poretti³, M. Rainer³, D. Ségransan¹, A. Sozzetti⁴, and S. Udry¹

"The most efficient methods to recover planetary signals take into account the different activity indicators, use **red-noise models** to account for stellar RV signals and a **Bayesian framework** to provide model comparison in a robust statistical approach."

$$\mathbf{v} = f(\mathbf{x}|\theta) + \epsilon$$

~~$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$~~



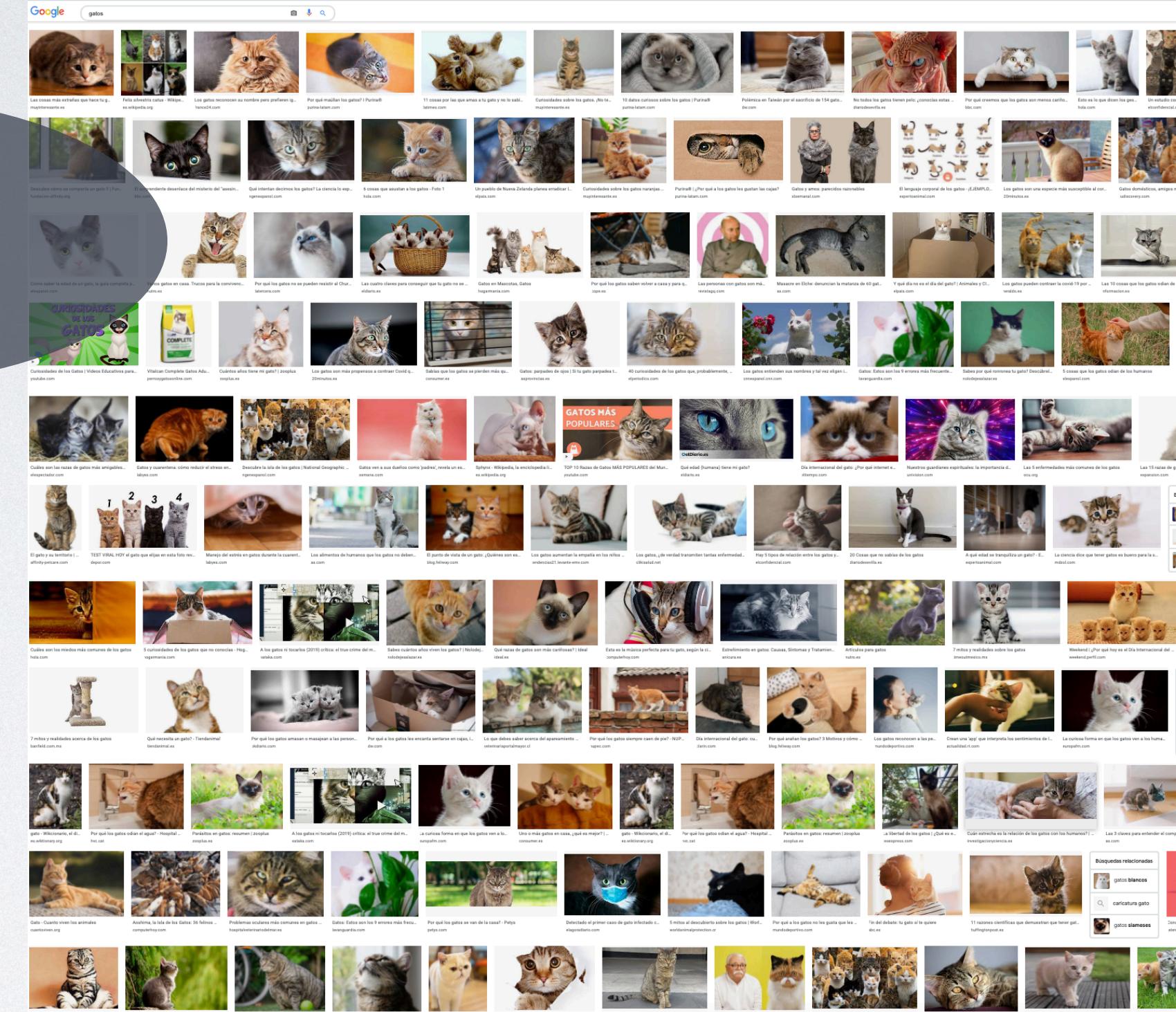
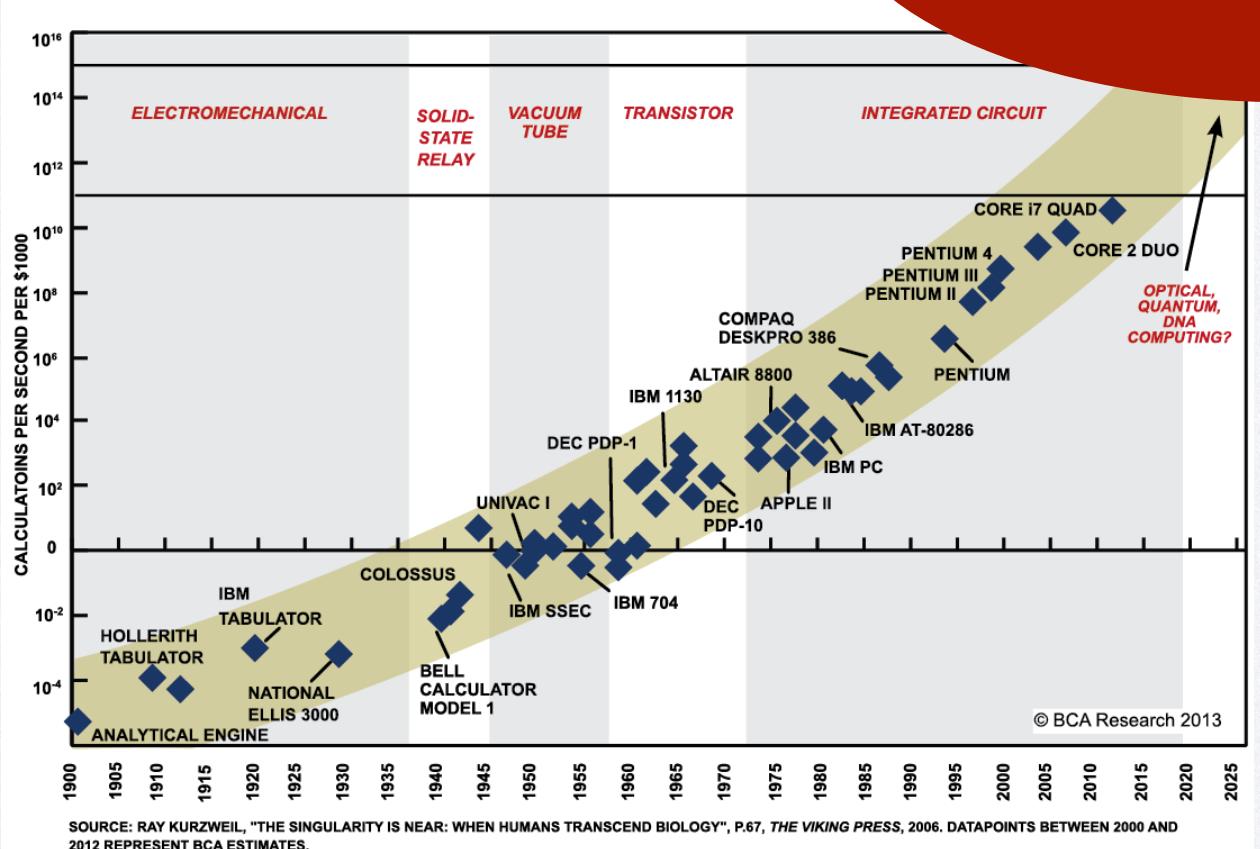


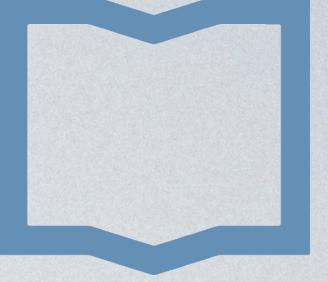
Algorithms

Data

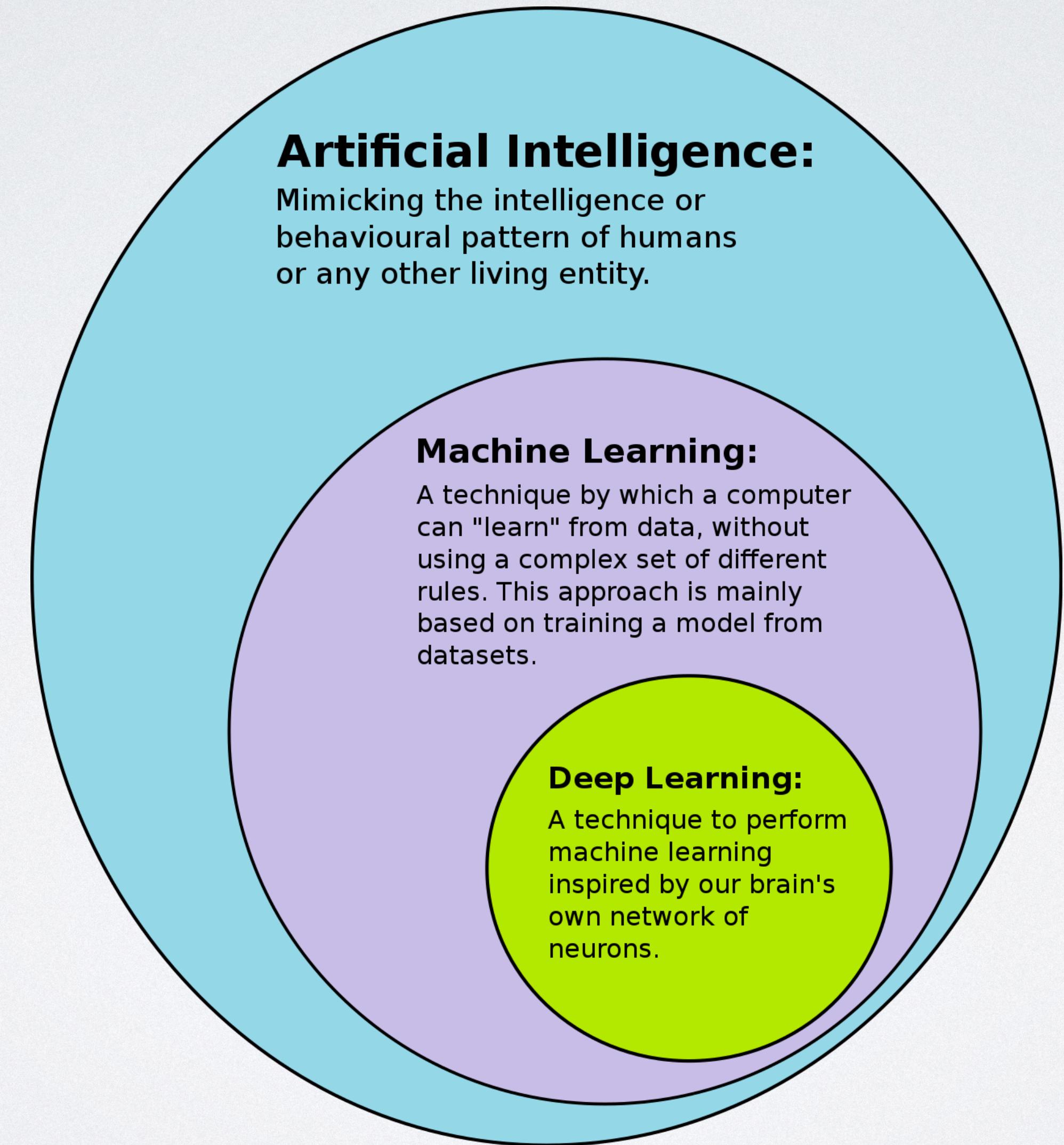
A.I.
revolution

Computing power

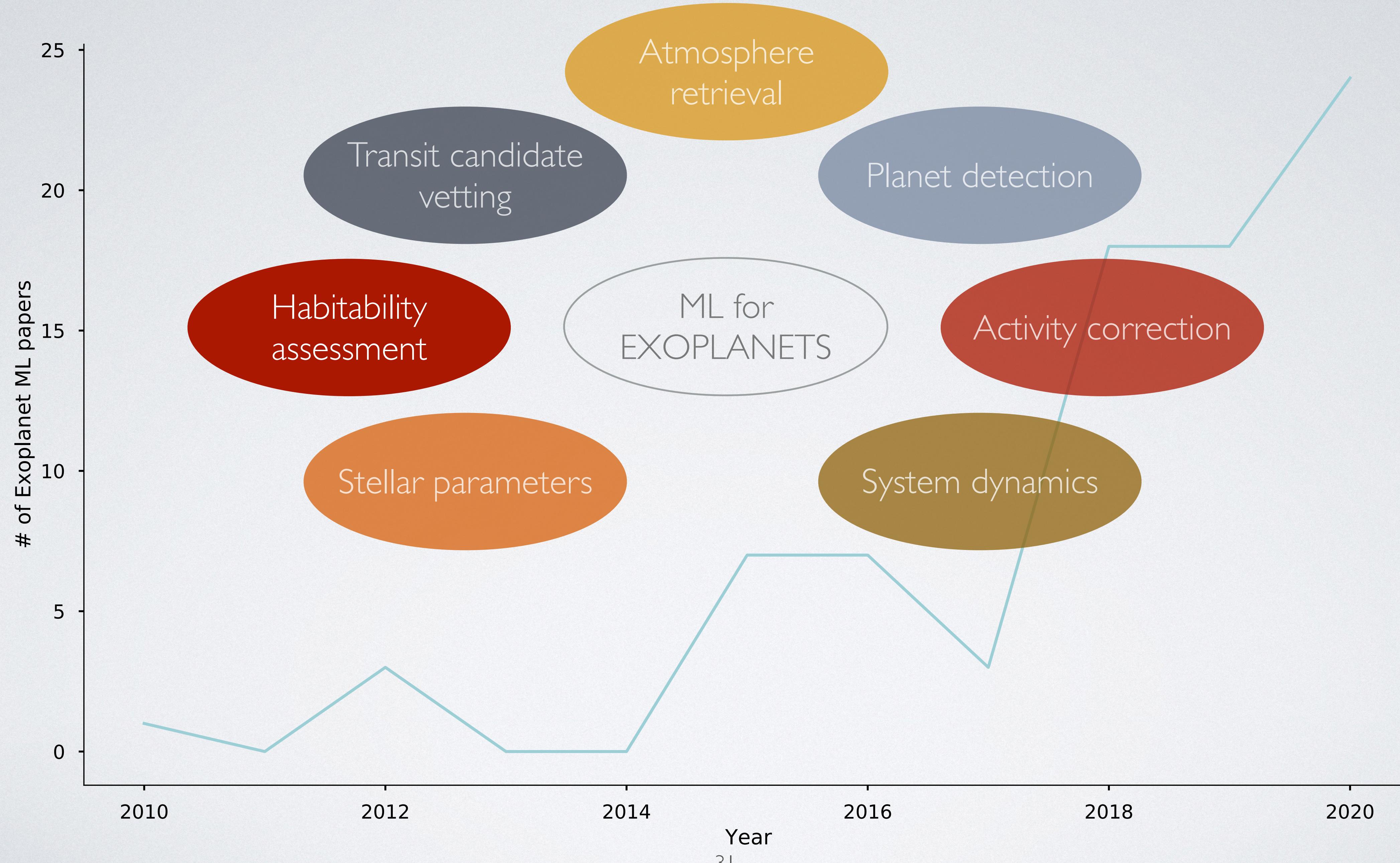




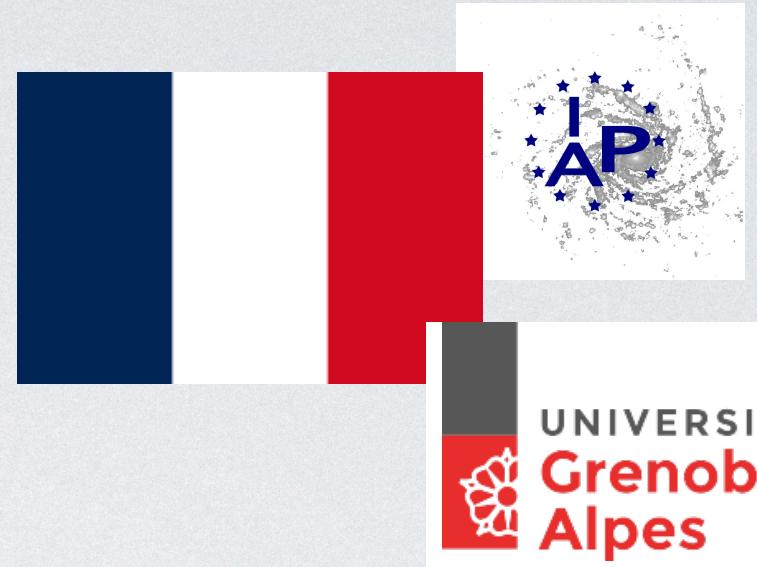
ARTIFICIAL INTELLIGENCE







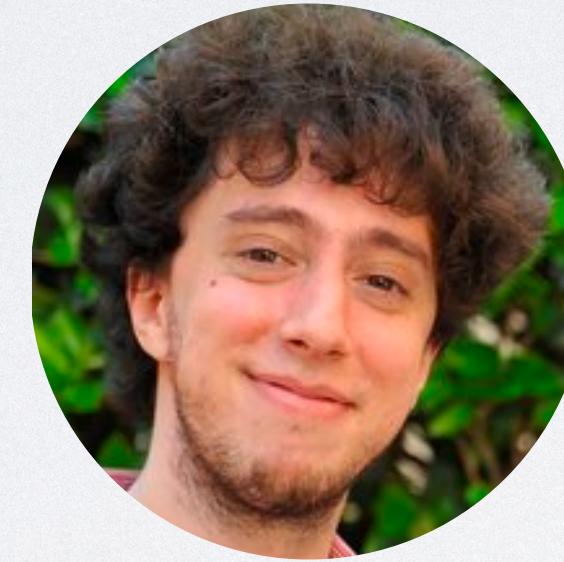
ICIFI



Luis Agustín Nieto
(Computer science)

Exoplanet detection in RV
data.

Deep Learning



Alejandro Hacker
(Physics)

TESS follow-up

Population parameter
inference

Hierarchical Bayesian
Models



Juan Serrano
(Astronomy)

ML for instrument
optimisation

TESS follow-up
observations

Blind source separation

Andrea Buccino
(Physics)

Stellar Activity in low-
mass stars

Dynamo theory



Carla Oviedo
(Astronomy)

Effect of activity on RVs



Leila Asplanato
(Physics)

Dropout studies in
University

Causal inference



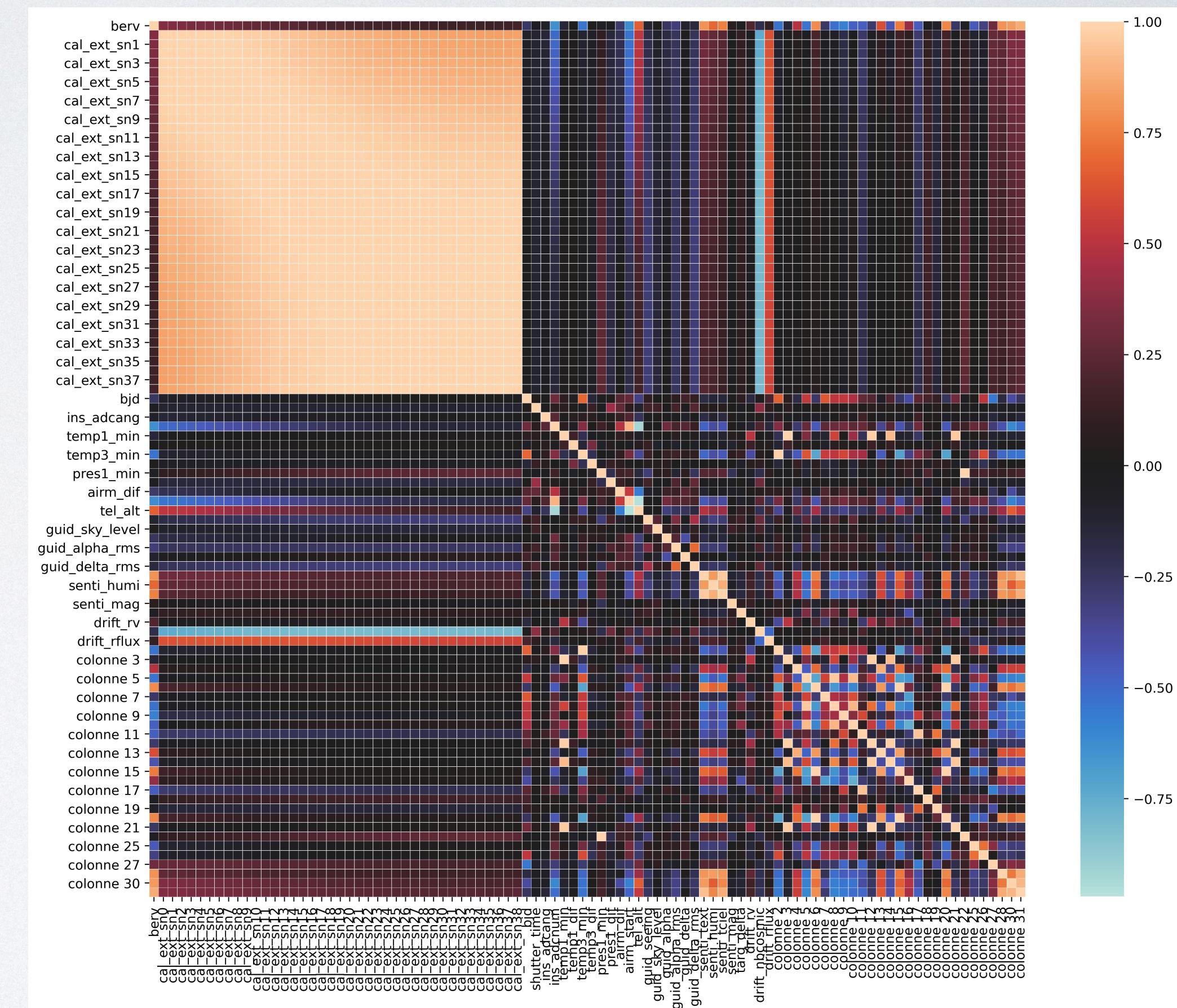
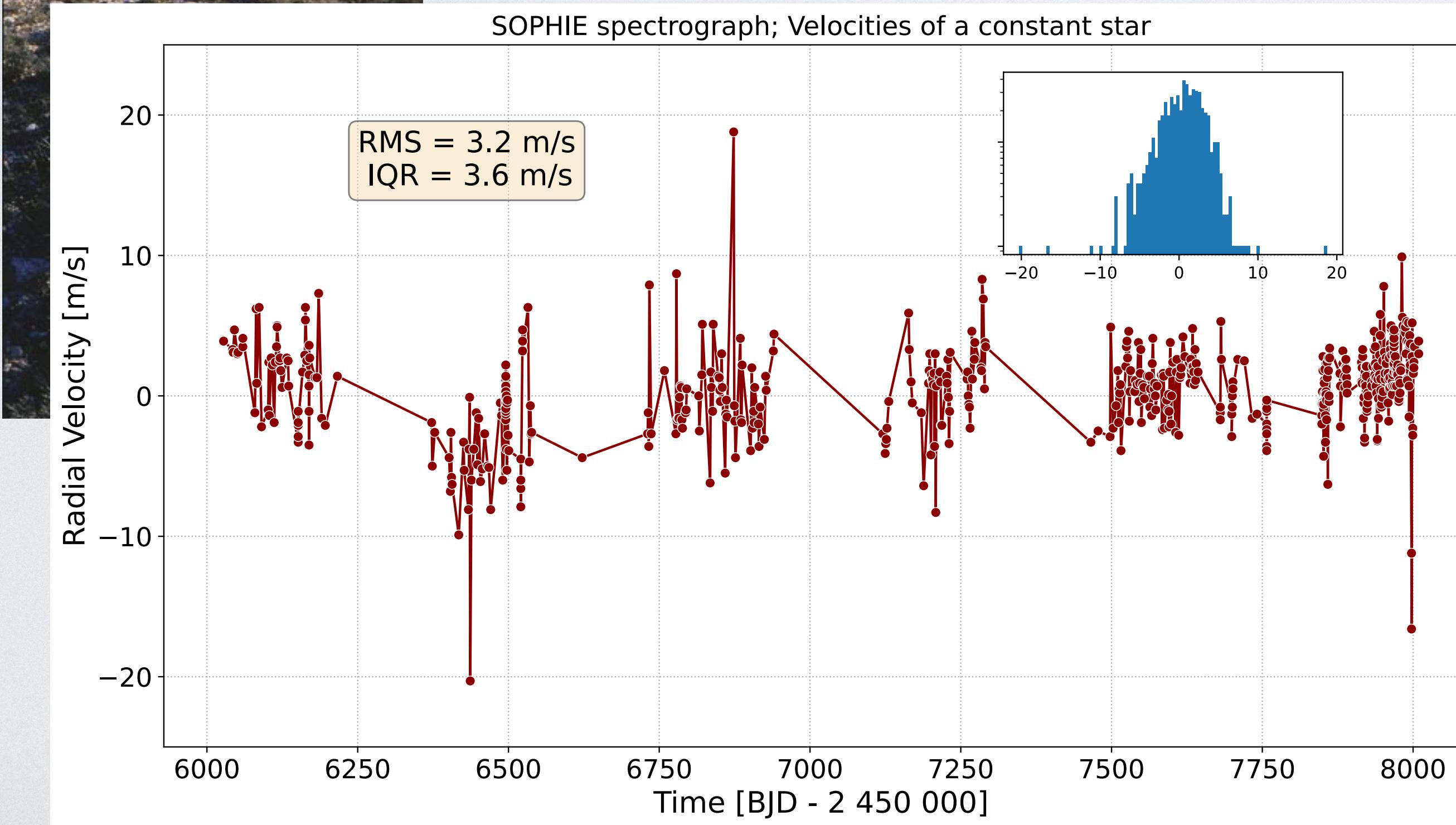


Observatoire de Haute-Provence, Francia

ML FOR INSTRUMENT OPTIMISATION

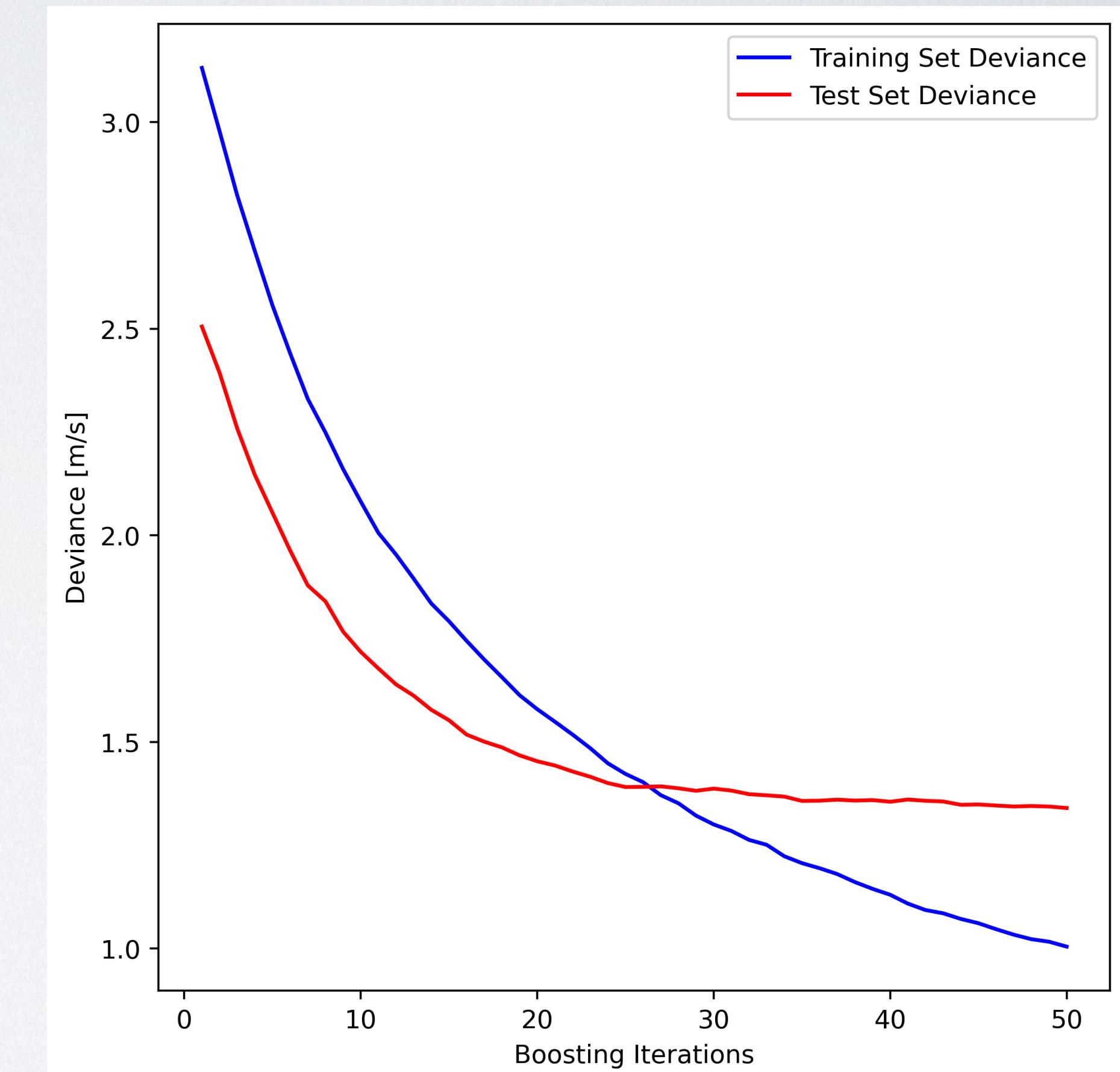
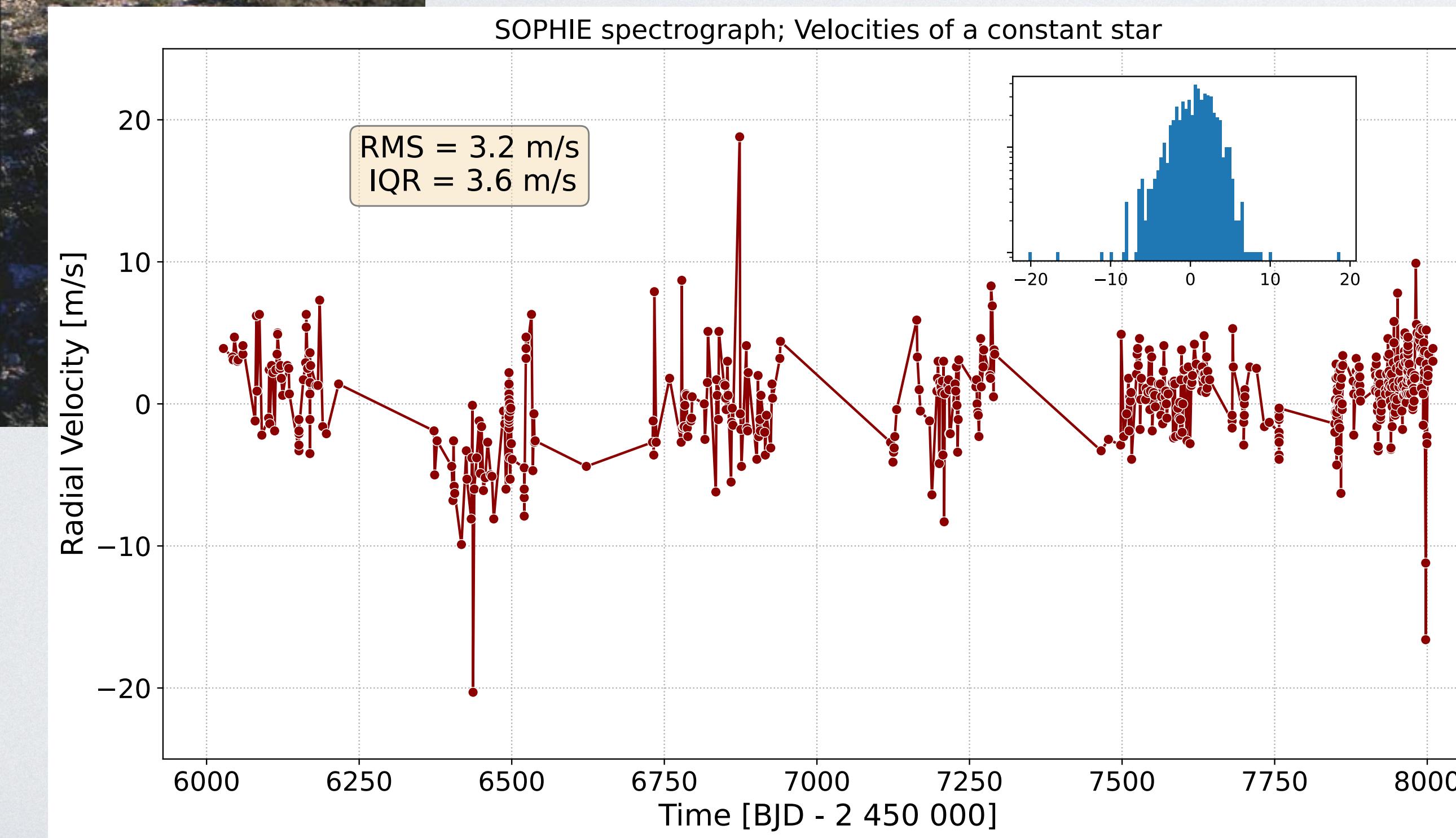


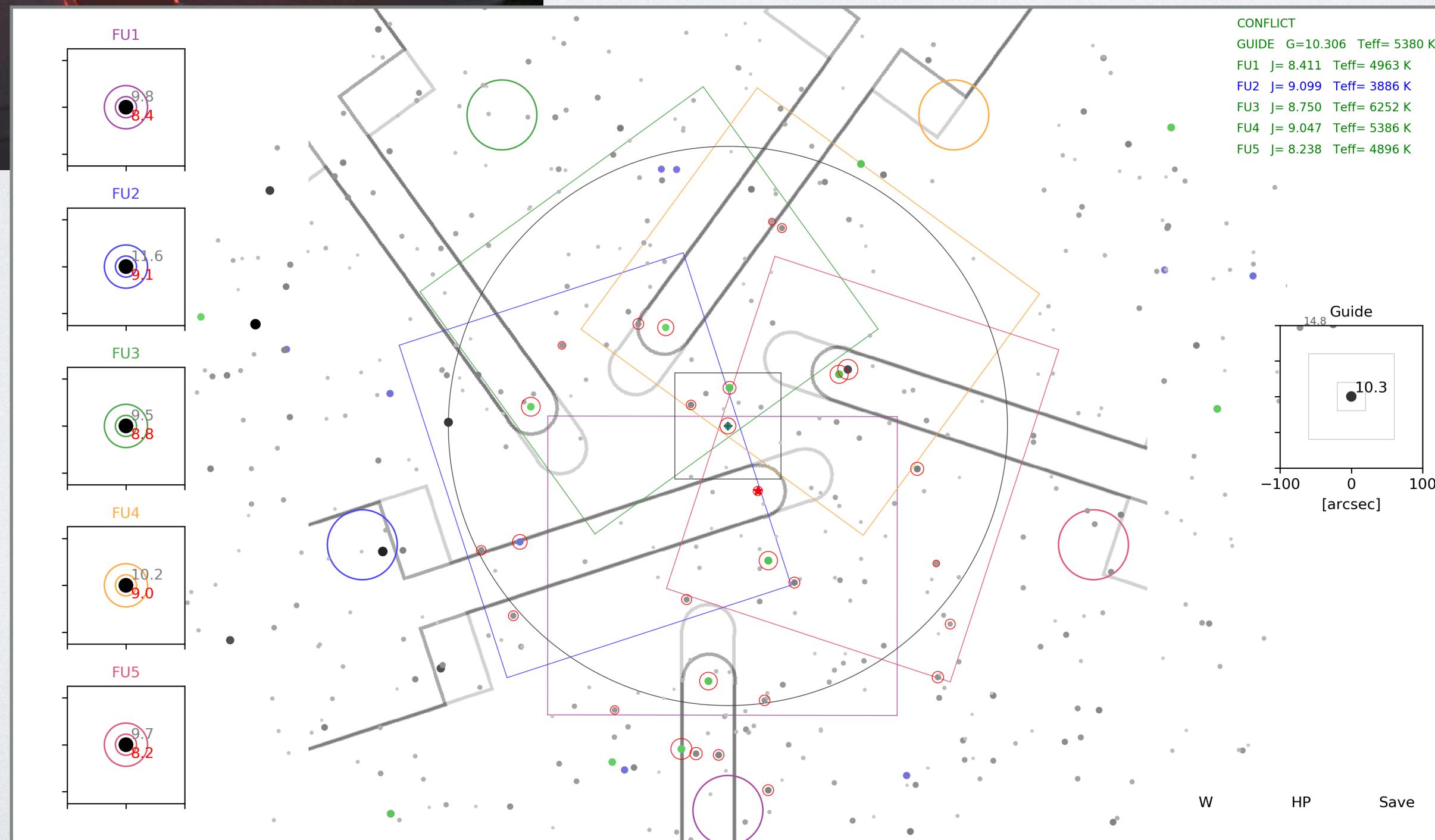
SOPHIE spectrograph; Velocities of a constant star



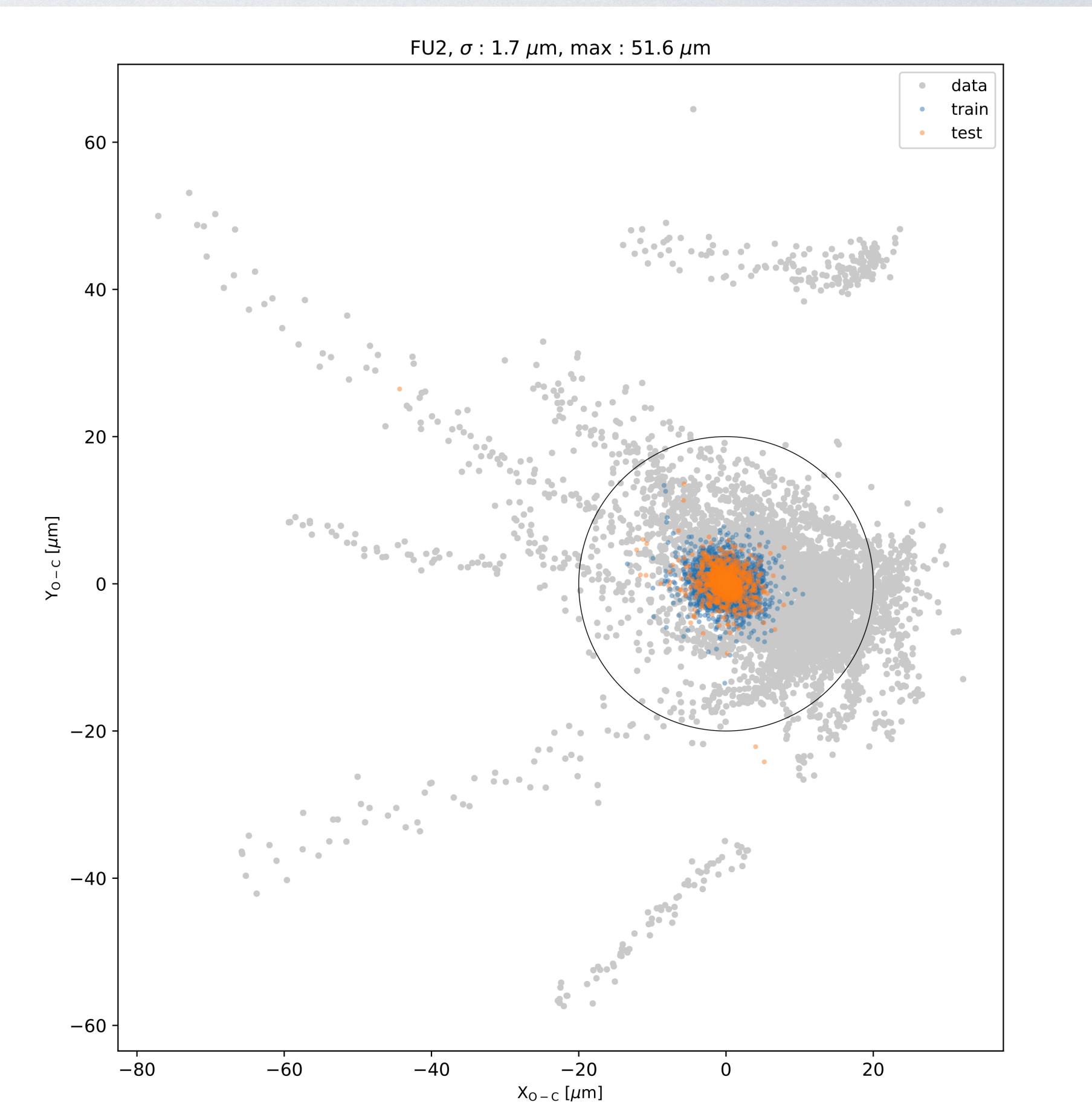


ML FOR INSTRUMENT OPTIMISATION

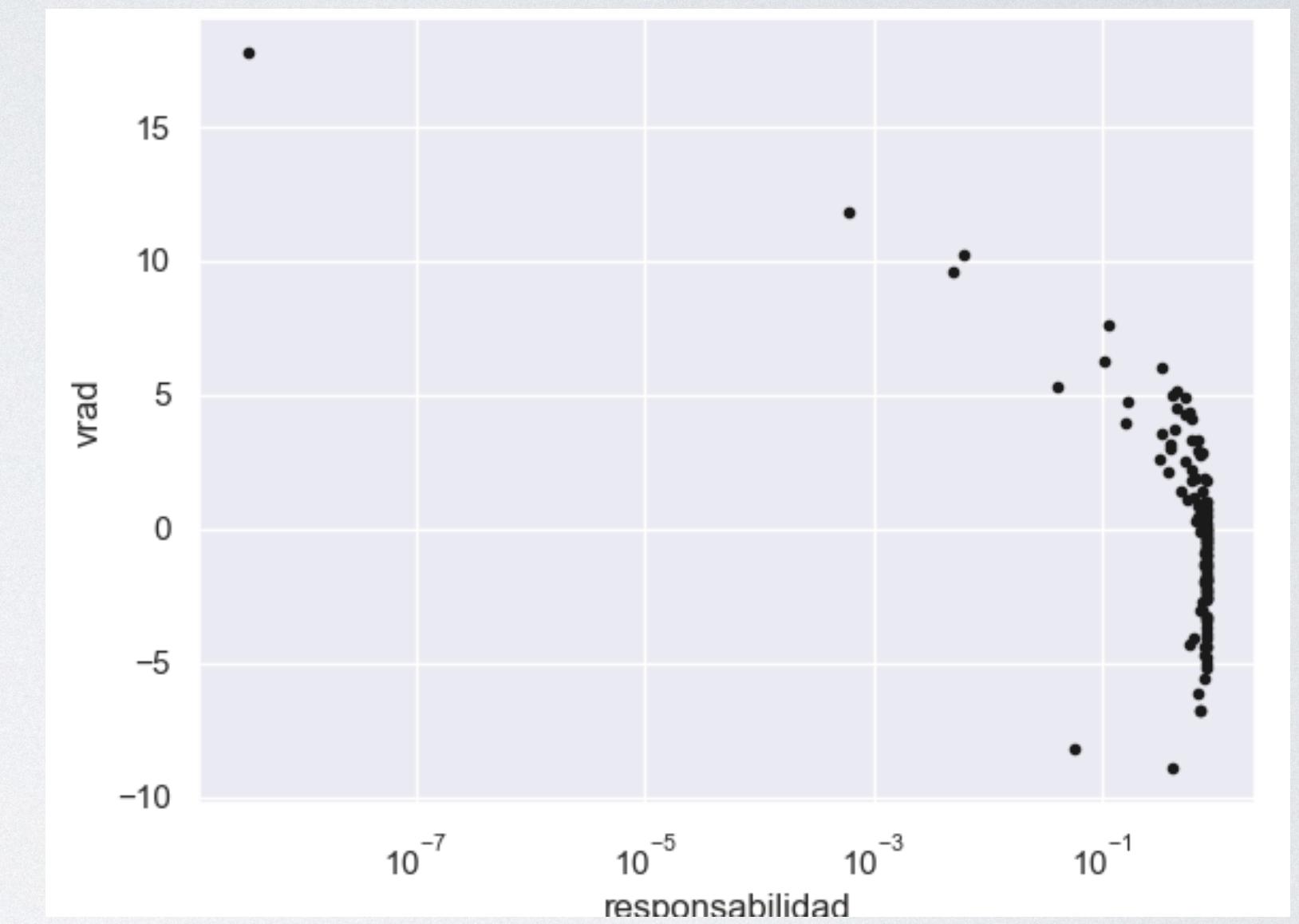
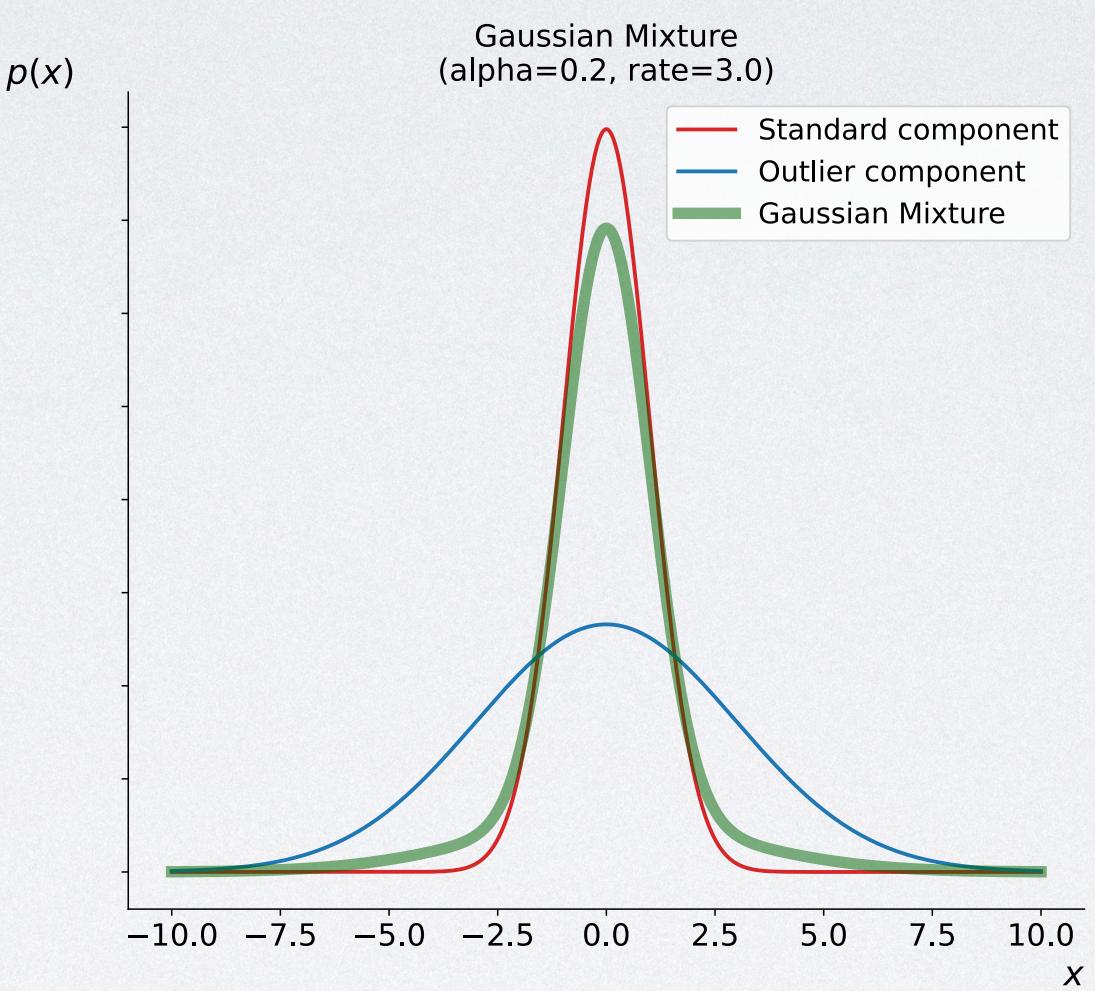
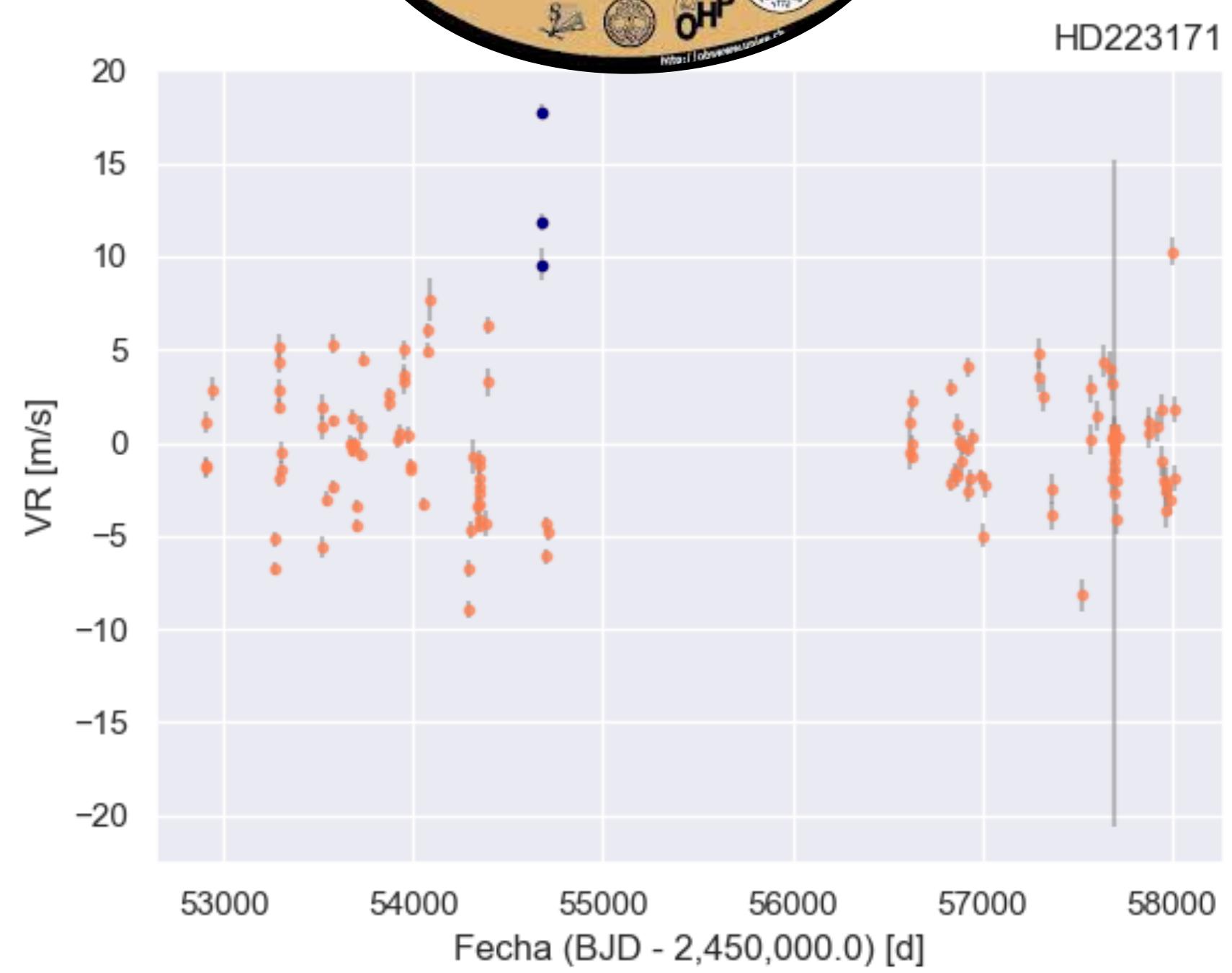
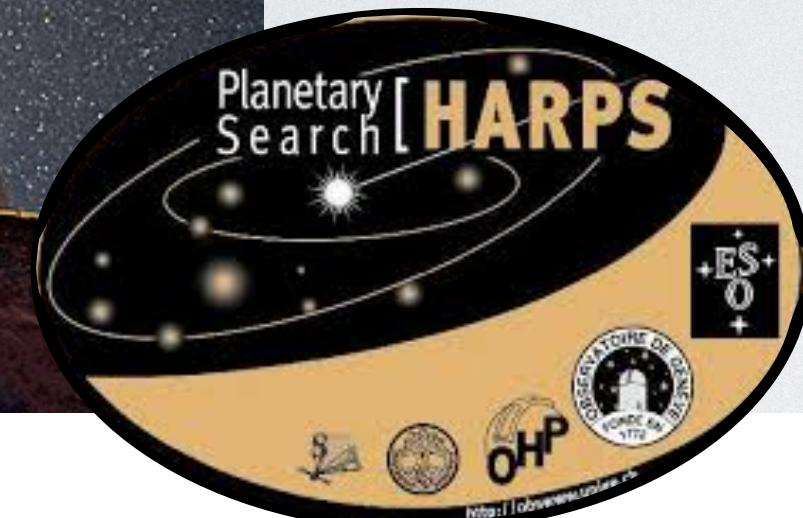
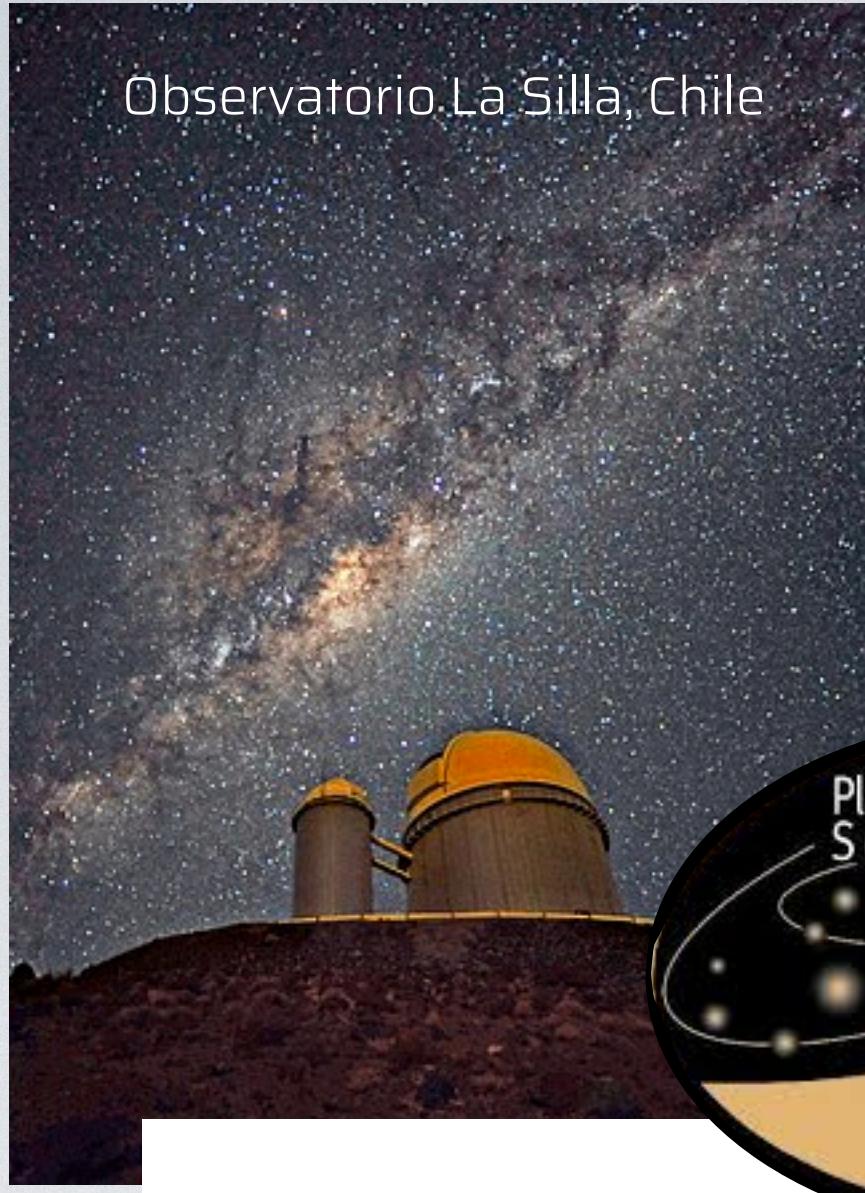




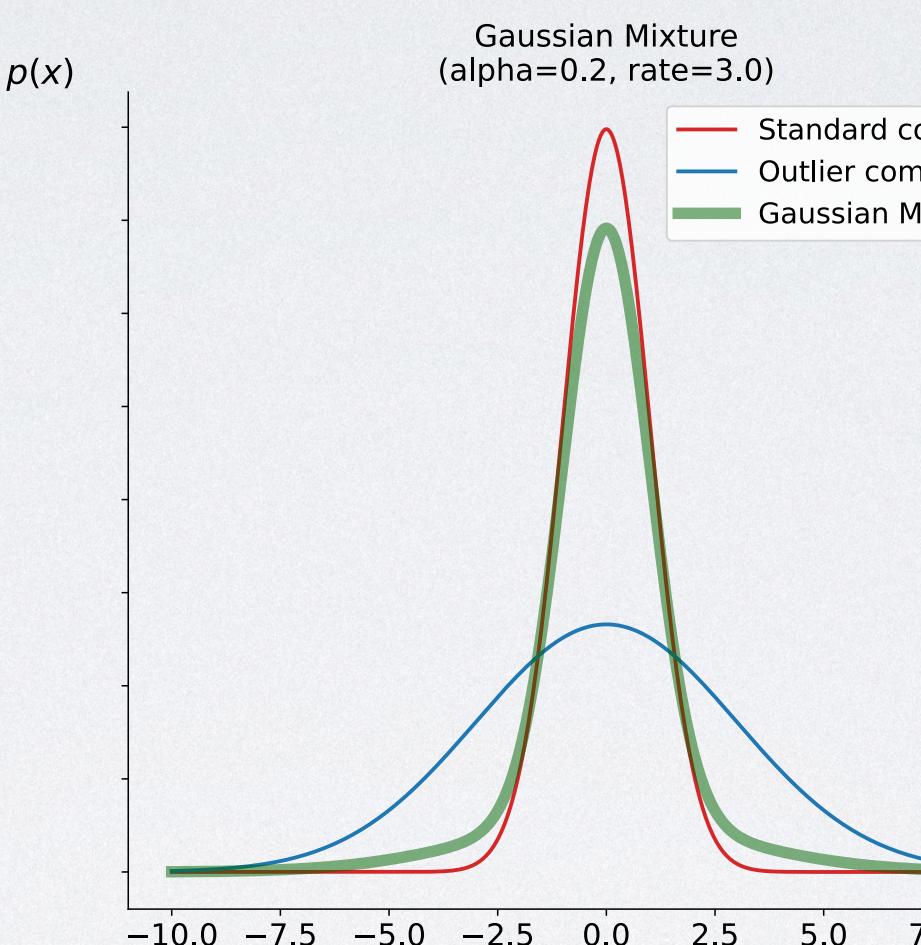
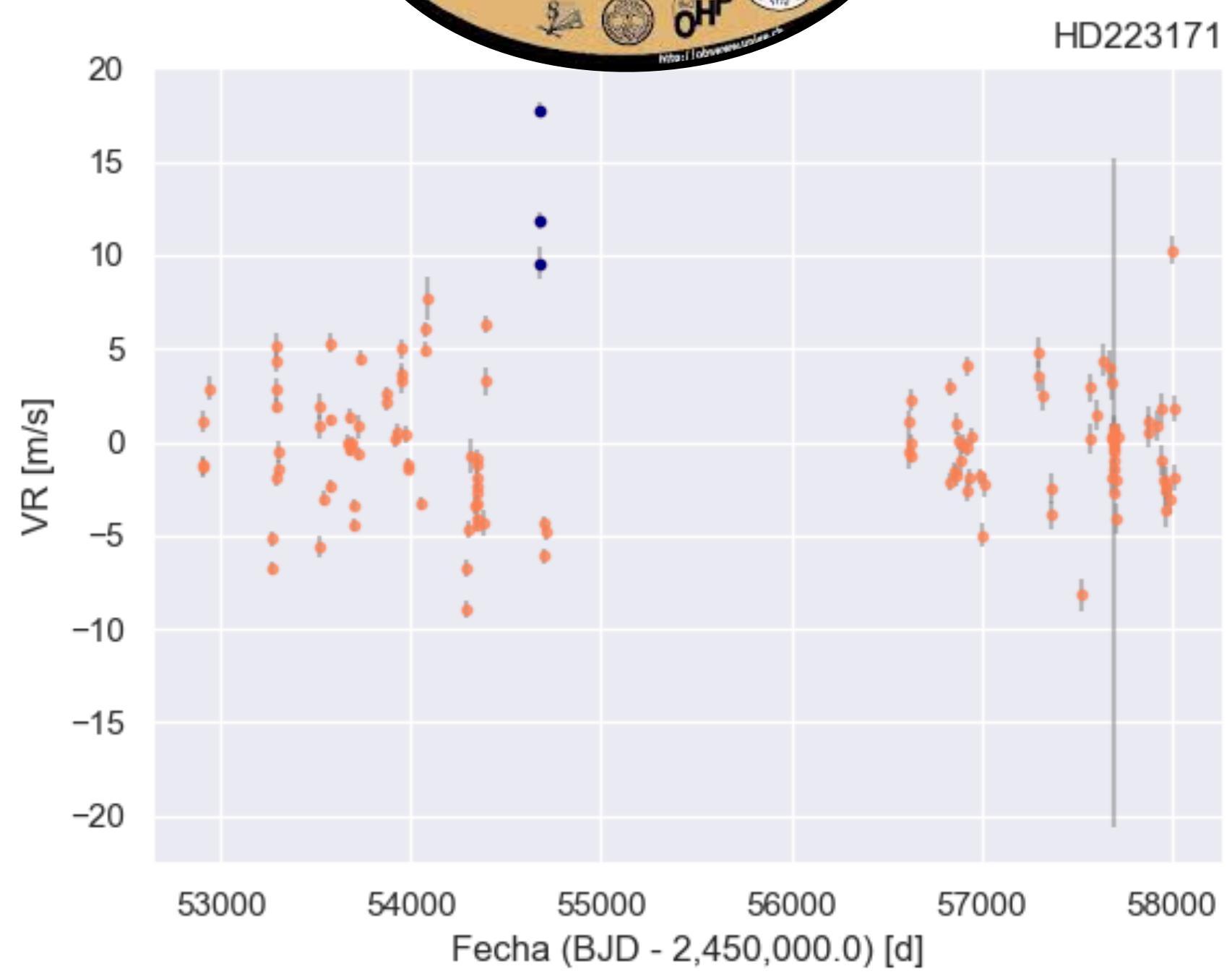
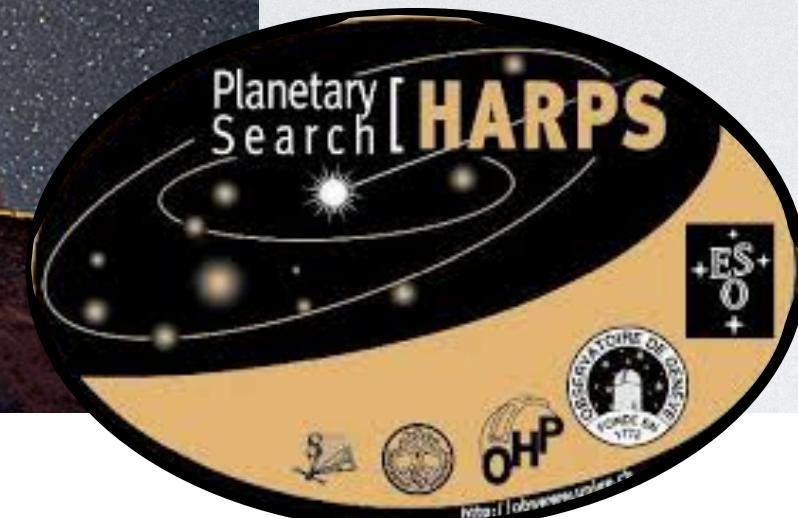
ML TO IMPROVE INSTRUMENT OPERATION



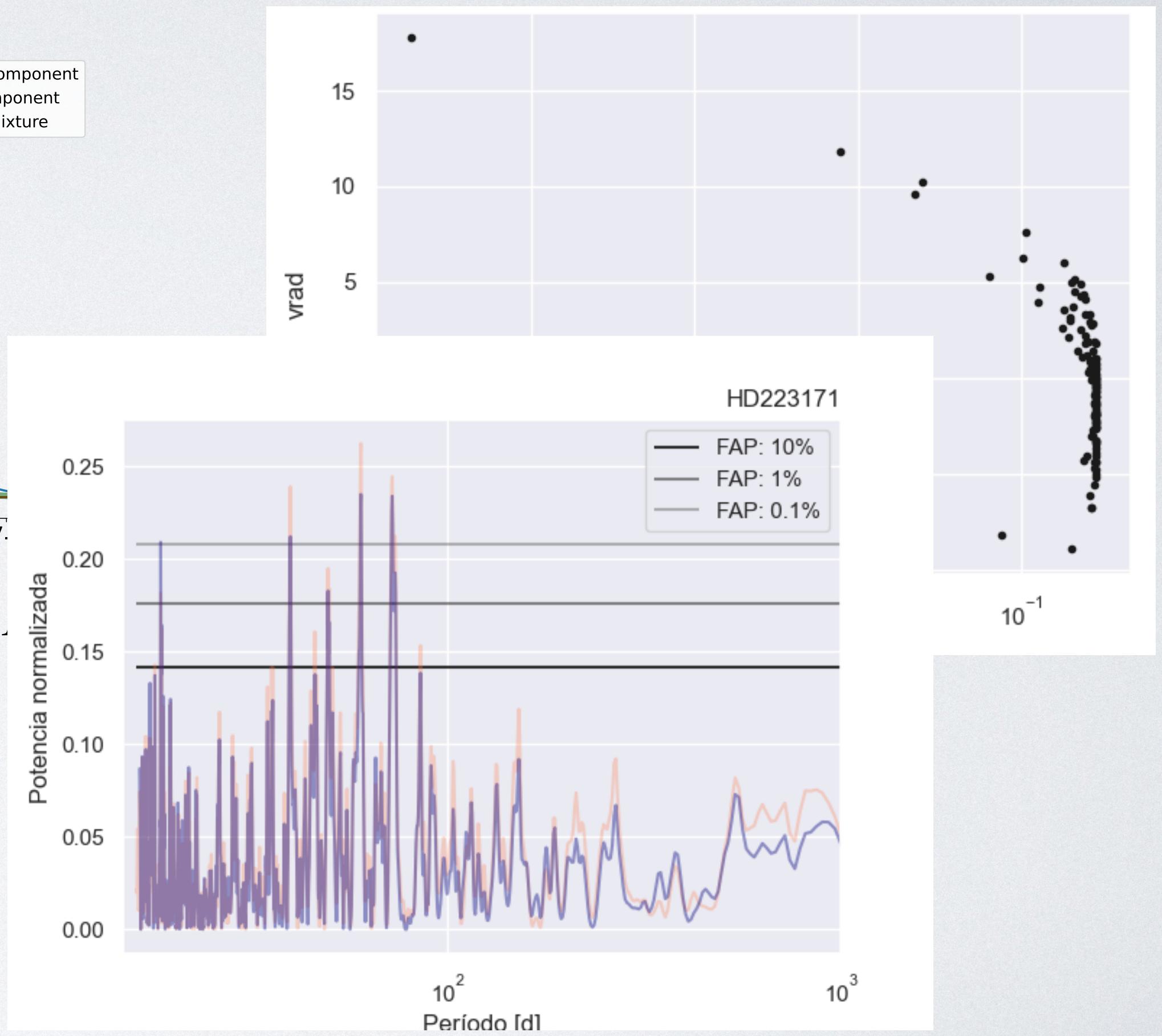
MIXTURE MODELS TO ACCOUNT FOR OUTLIERS



MIXTURE MODELS TO ACCOUNT FOR OUTLIERS



$$p(x) = (1 - \alpha)N(0, \sigma) + \alpha$$



Astronomy & Astrophysics manuscript no. ExoplANNET-arvix-2
July 4, 2023



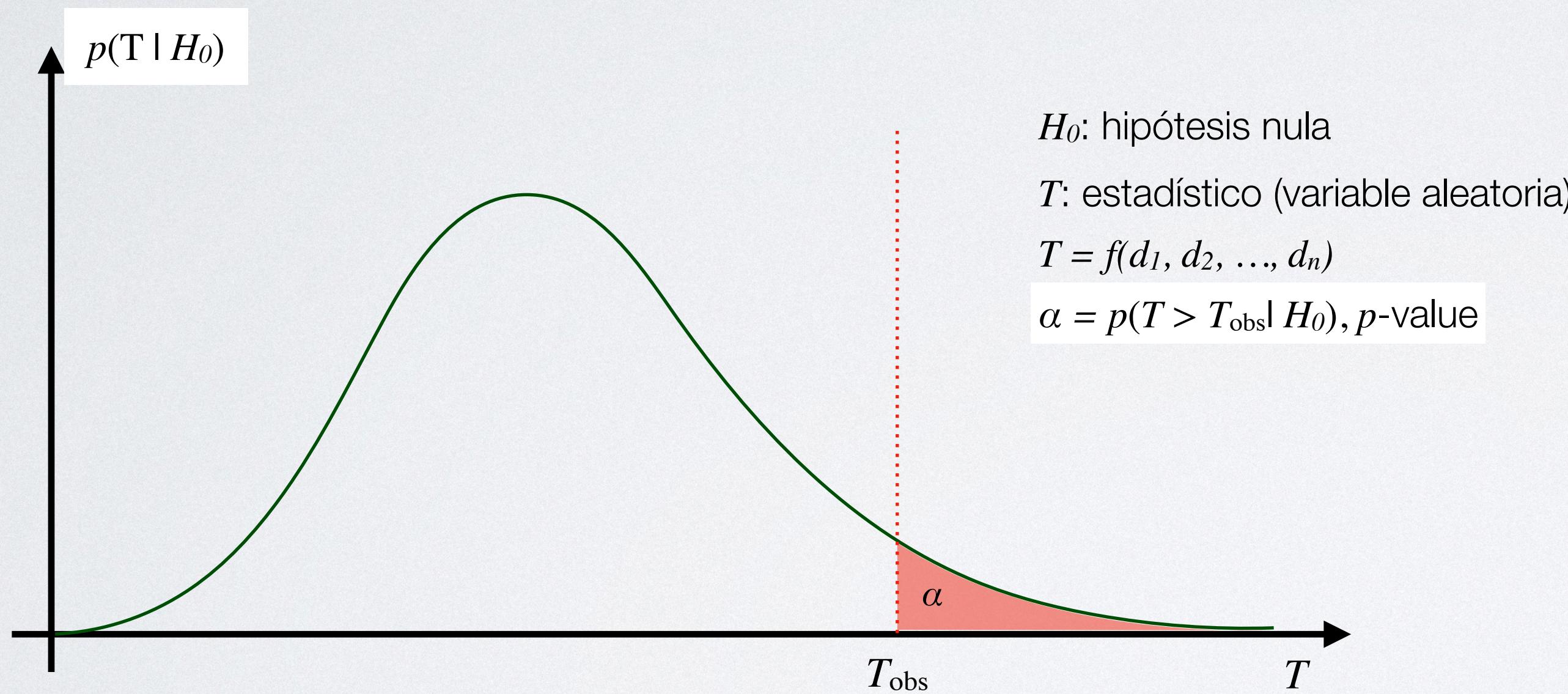
Luis Agustín Nieto

ExoplANNET: A deep learning algorithm to detect and identify planetary signals in radial velocity data

L. A. Nieto^{1,2} & R. F. Díaz²

OUR BASELINE MODEL

The classical approach to detecting signals in RV time series



Analytical distributions depend strongly on hypotheses that are **rarely satisfied**.

Simulations (*bootstrapping*) under the null are performed to alleviate this. This is **computationally expensive**.

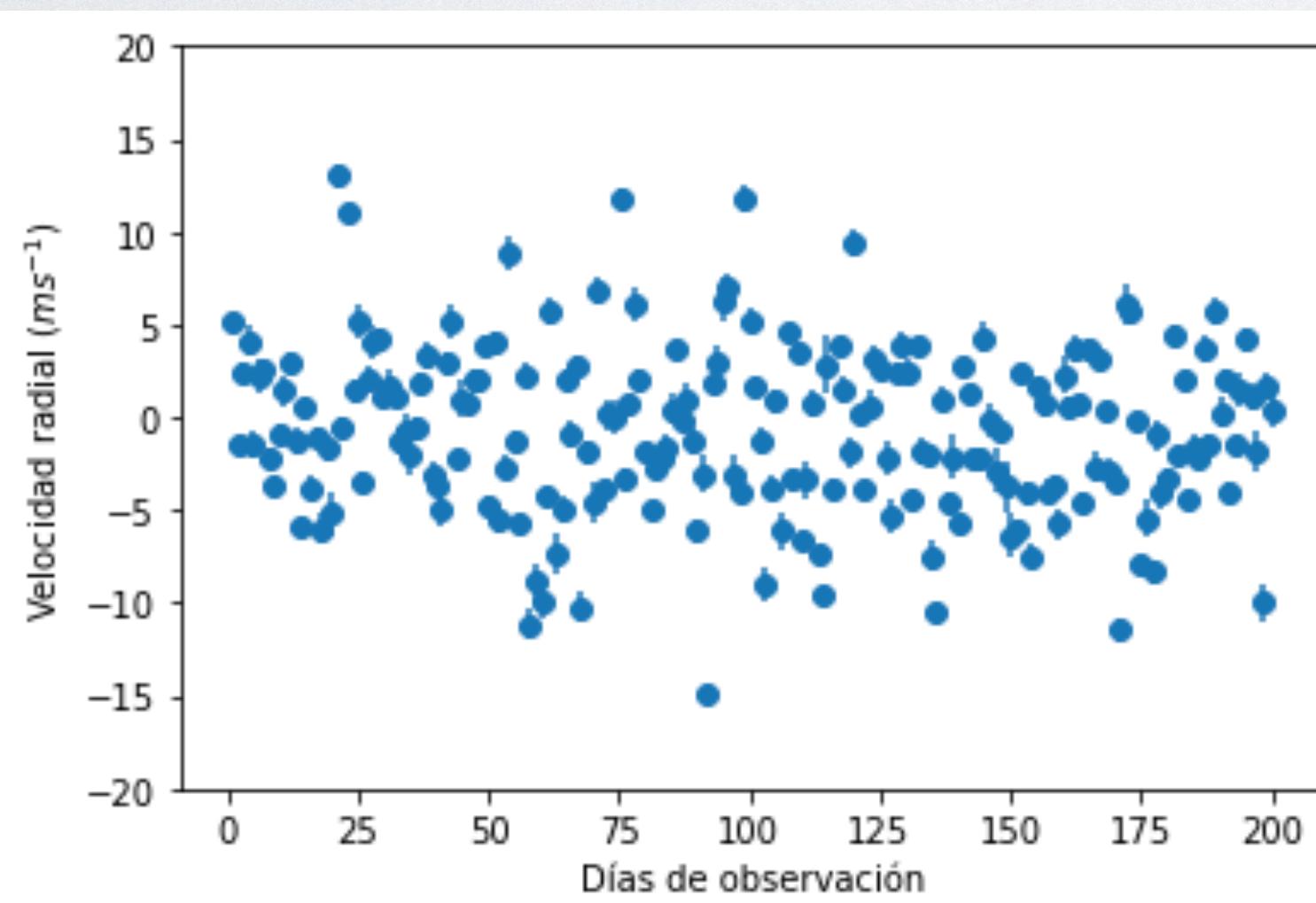
Theoretical issues with p -values in general
—> Bayesian statistics.

DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.

Realistic noise

- White (photon) noise.
- Pulsations, oscillations.
- Rotational modulation.

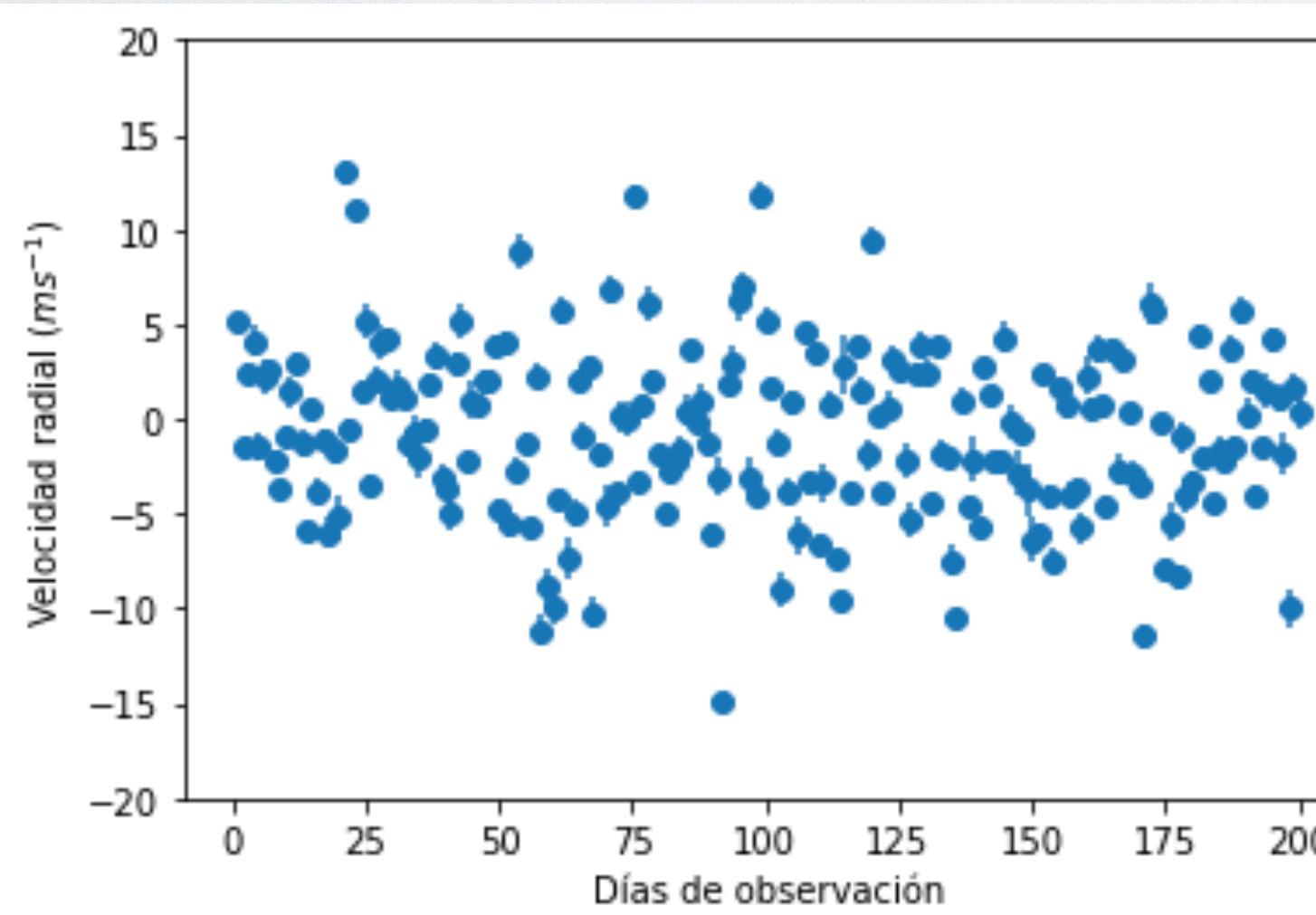


DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.

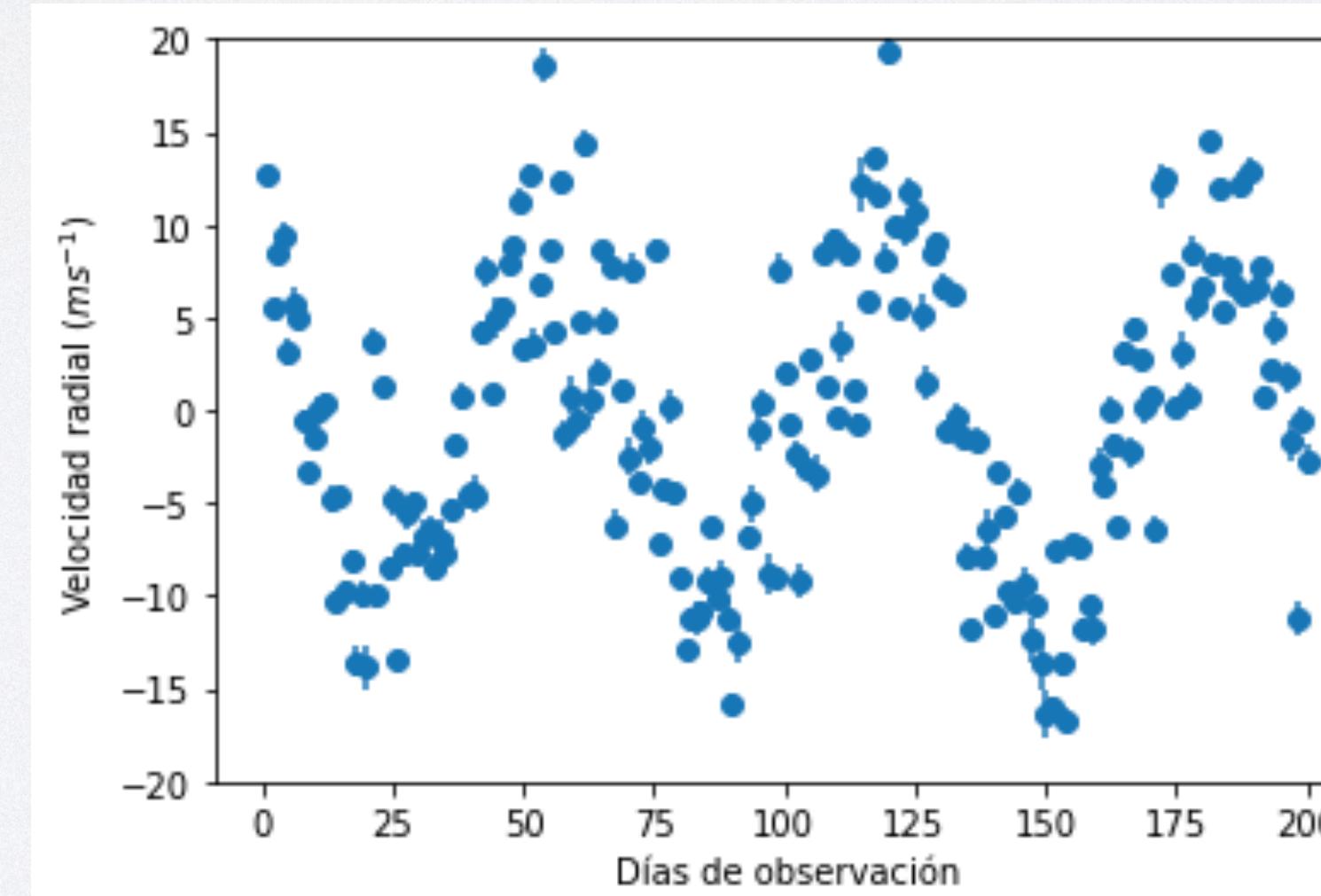
Realistic noise

- White (photon) noise.
- Pulsations, oscillations.
- Rotational modulation.



Circular planets

- Period $\sim U[10 \text{ d}, 100 \text{ d}]$
- Amplitudes $\sim \text{log-flat}[0.1 \text{ m/s}, 10 \text{ m/s}]$
- Nplanets in $\{0, 1, 2, 3, 4\}$

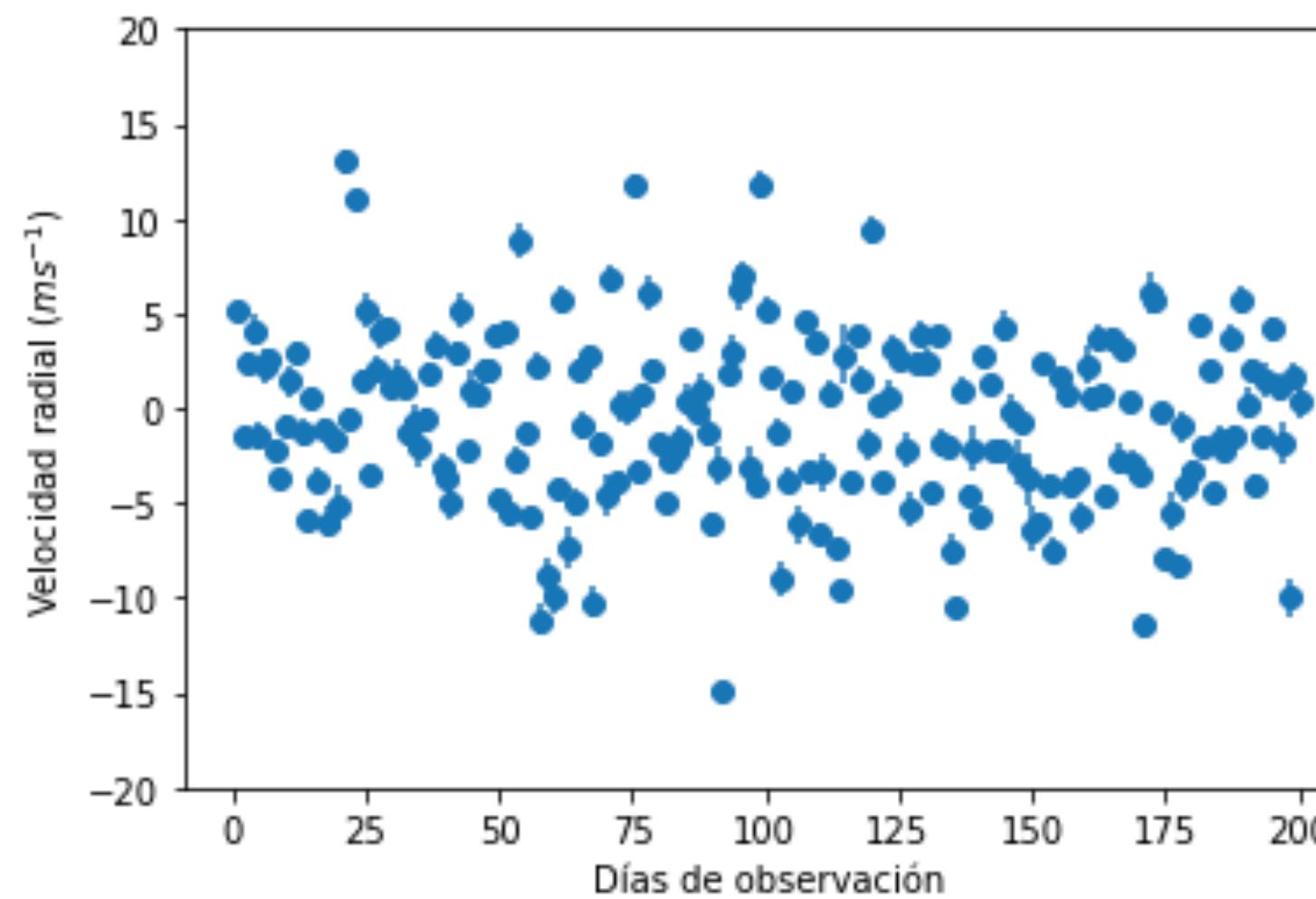


DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.

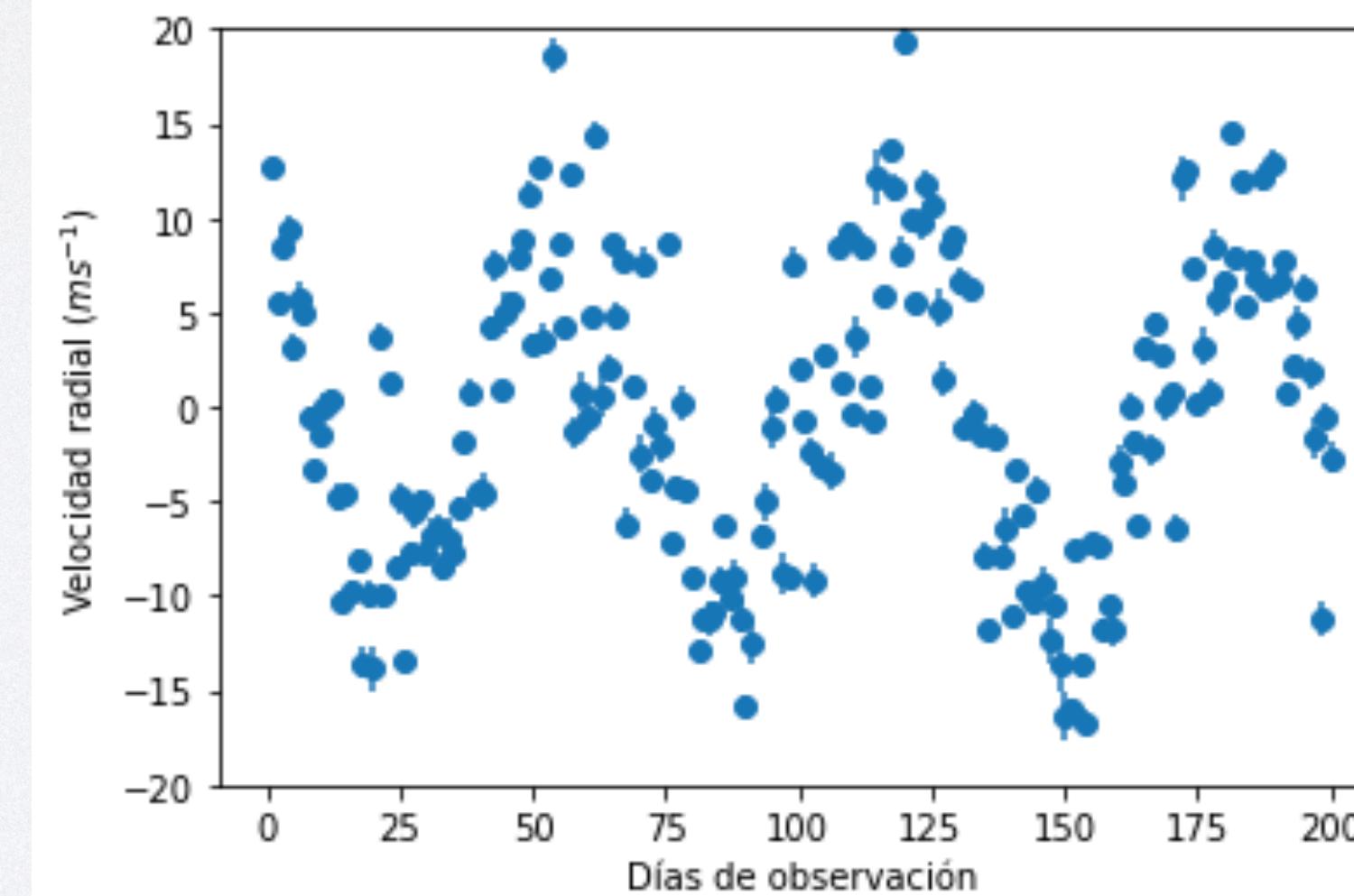
Realistic noise

- White (photon) noise.
- Pulsations, oscillations.
- Rotational modulation.



Circular planets

- Period $\sim U[10 \text{ d}, 100 \text{ d}]$
- Amplitudes $\sim \text{log-flat}[0.1 \text{ m/s}, 10 \text{ m/s}]$
- Nplanets in $\{0, 1, 2, 3, 4\}$



Parameters for noise simulations taken from real RV survey:

* HARPS high precision programme
(PI: Mayor —> Udry —> Díaz)

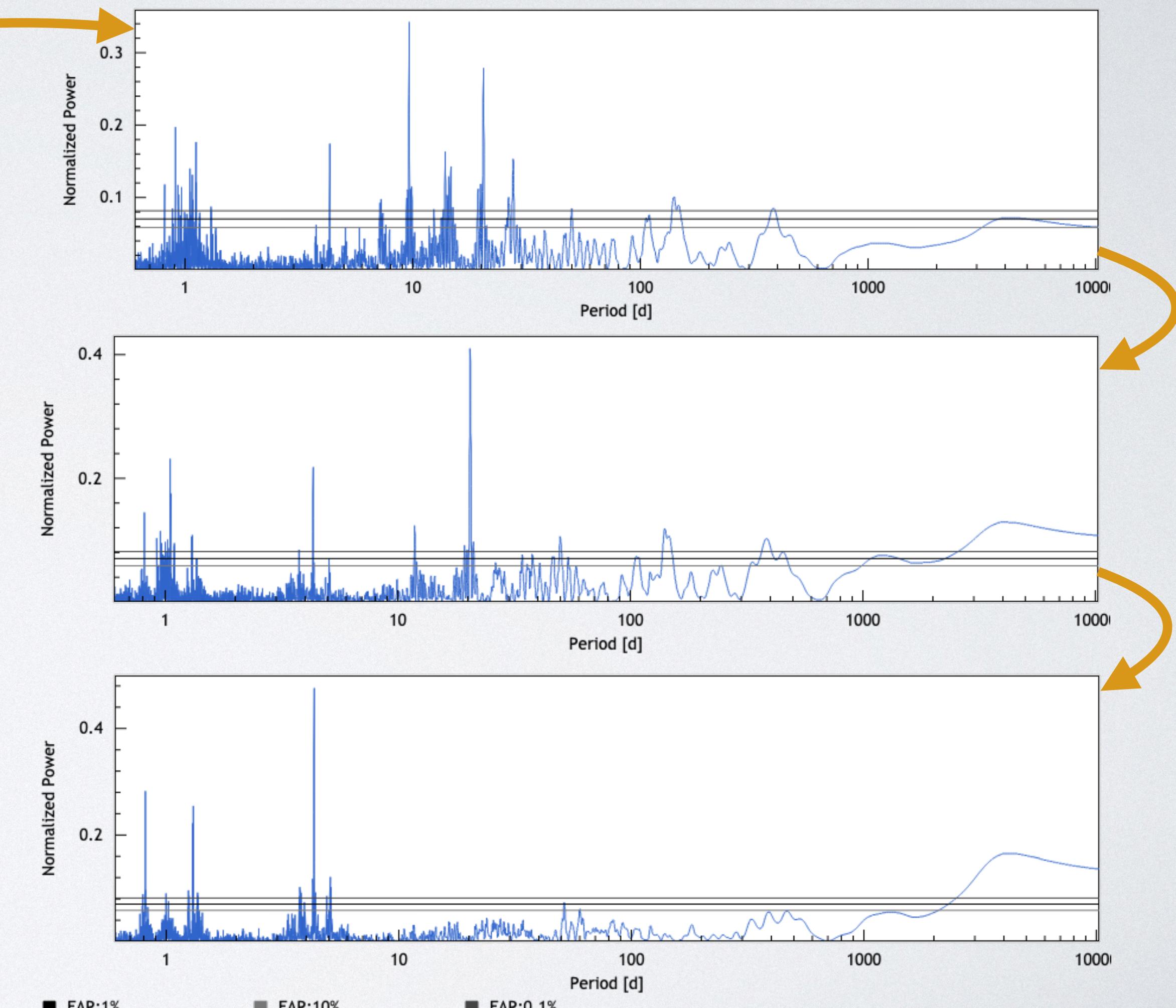
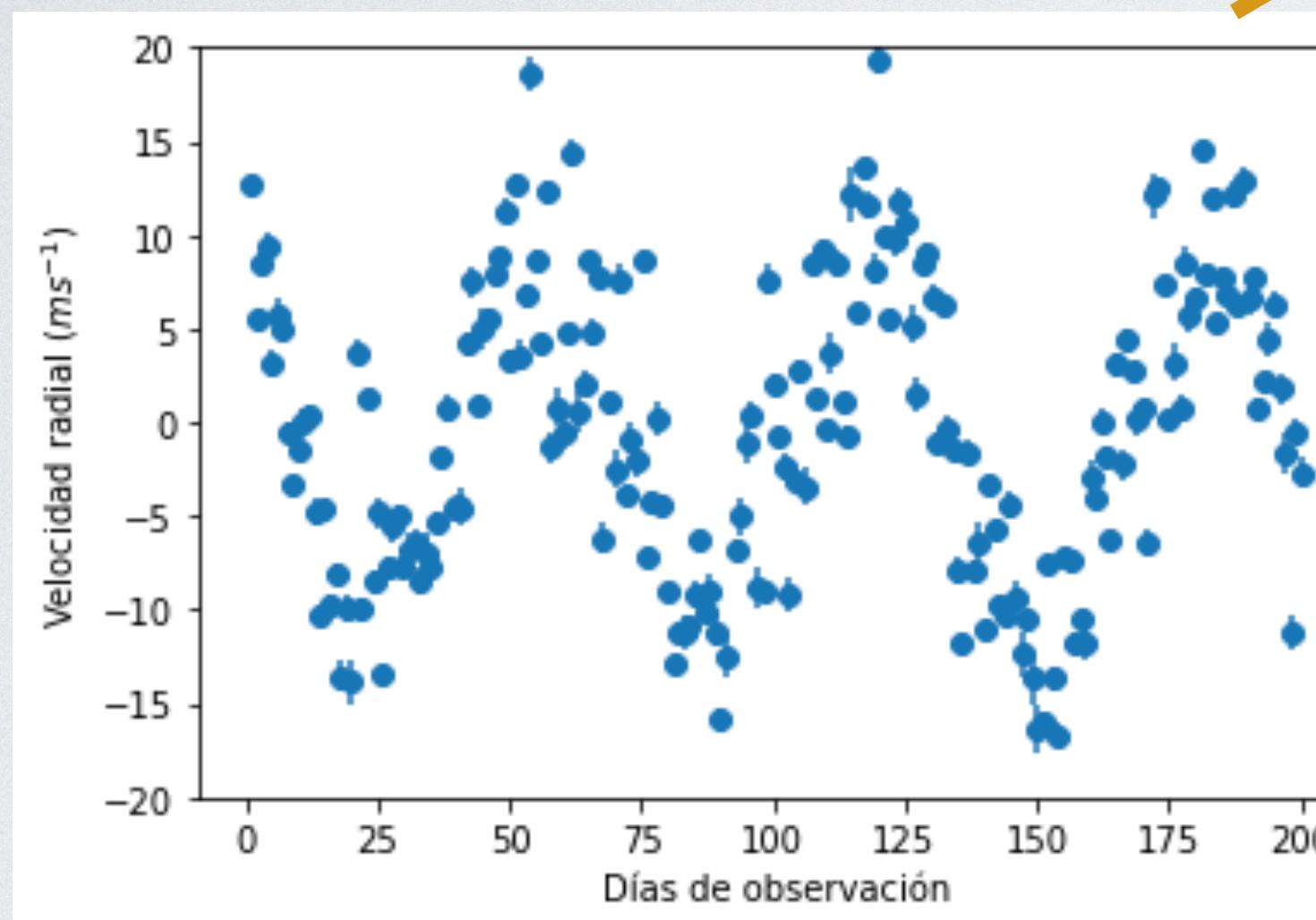
(ask me about noise simulations if you're interested!)

Time sampling

Pseudo-uniform (for historical reasons; not very realistic...)

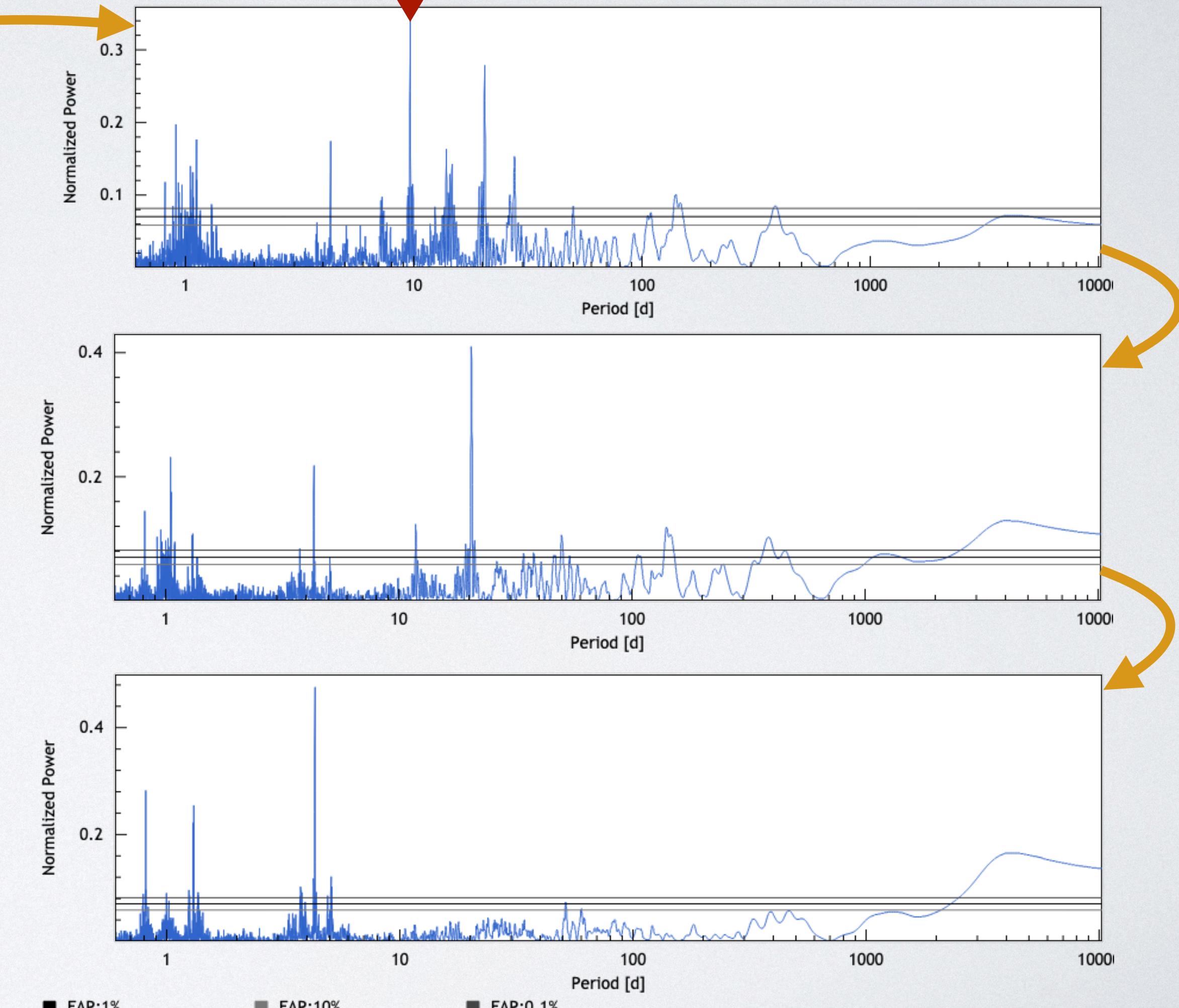
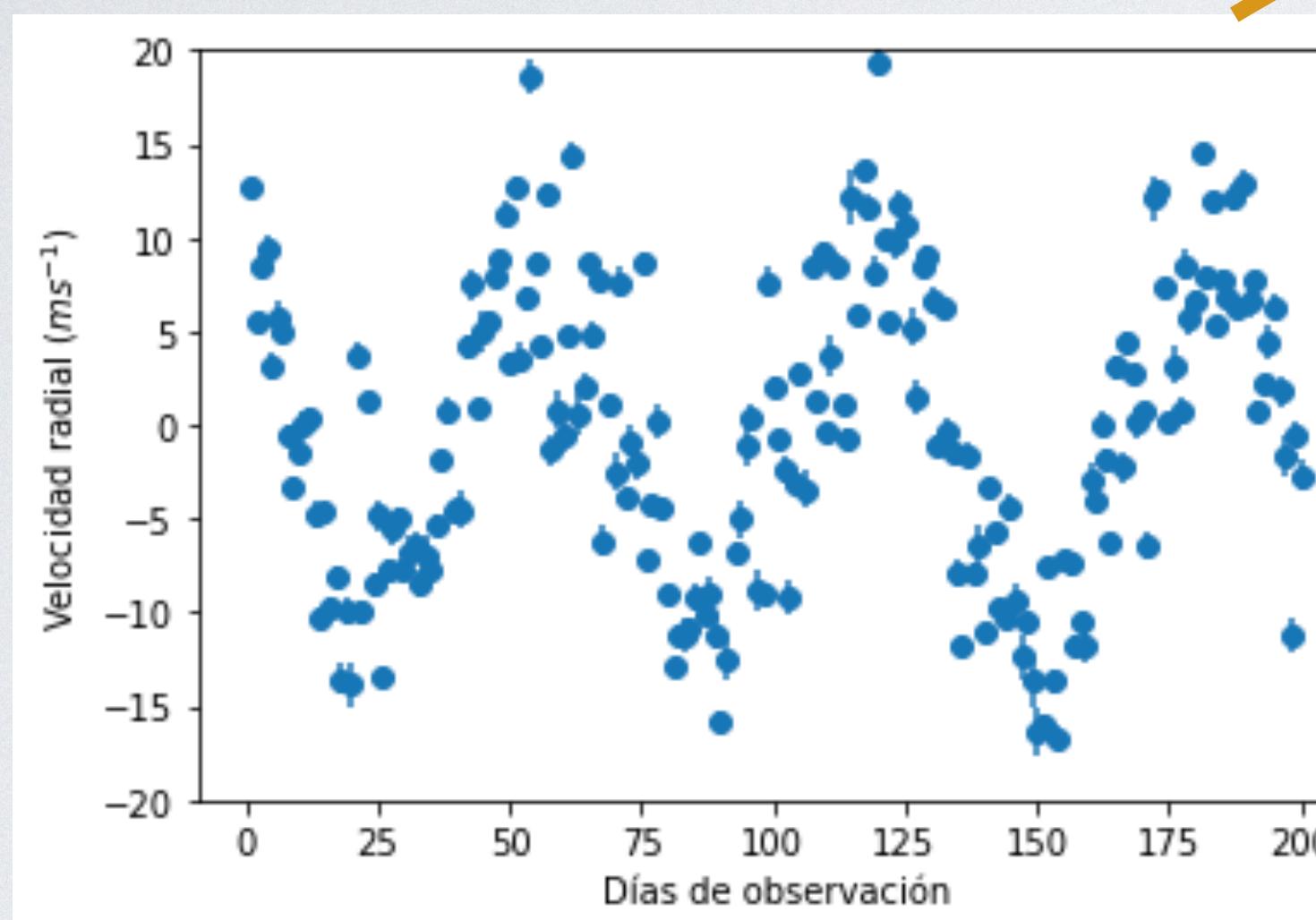
DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.



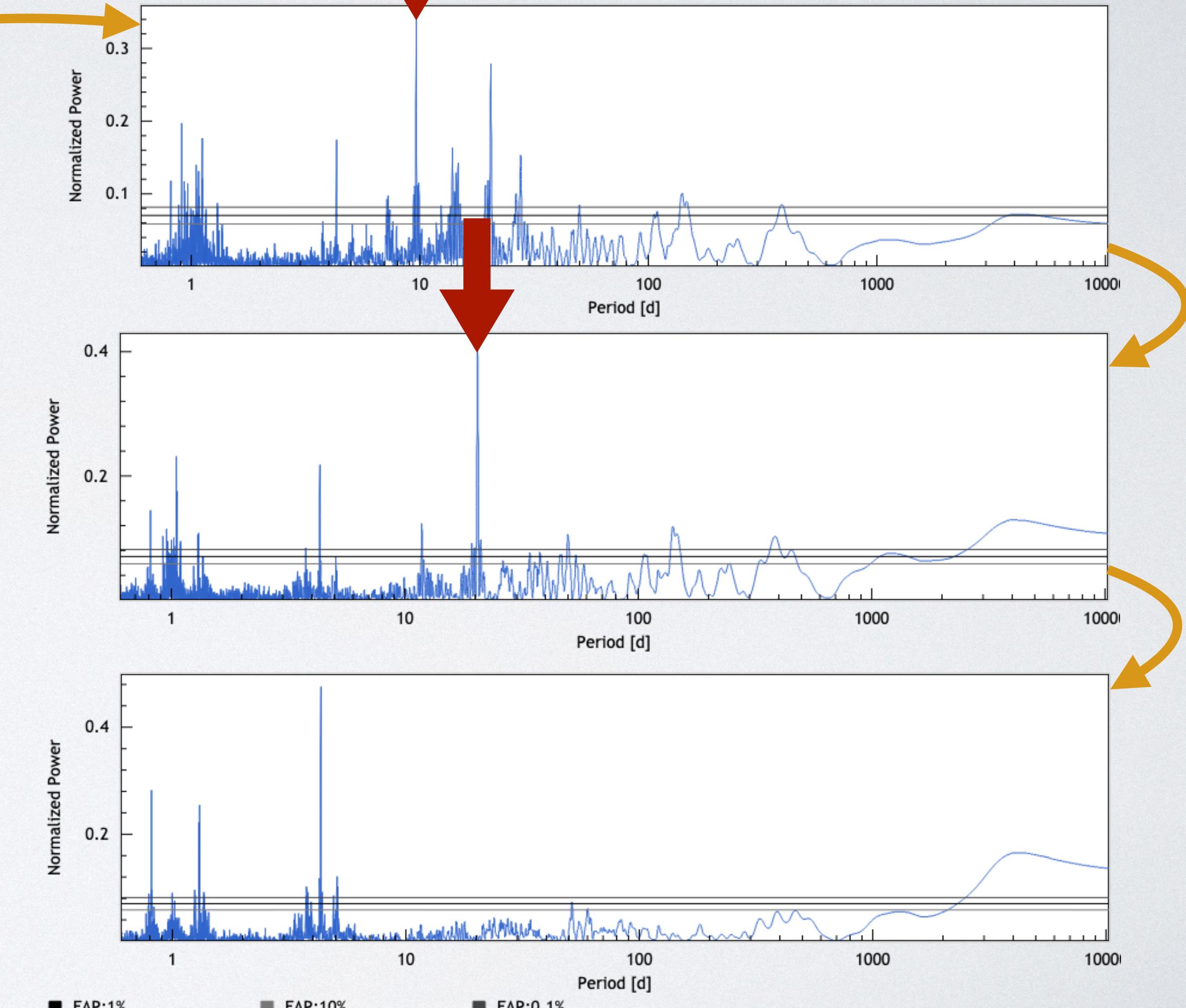
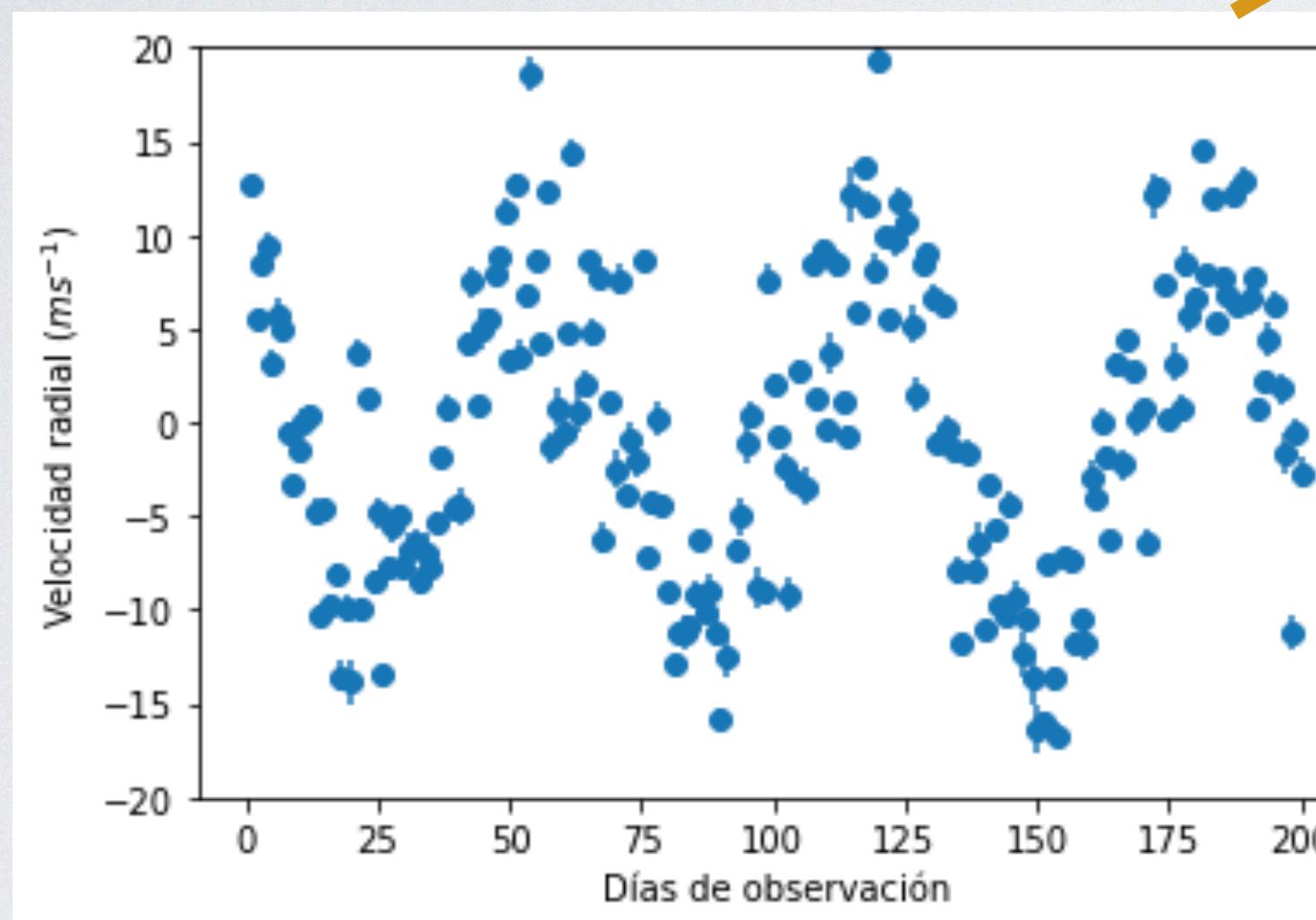
DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.



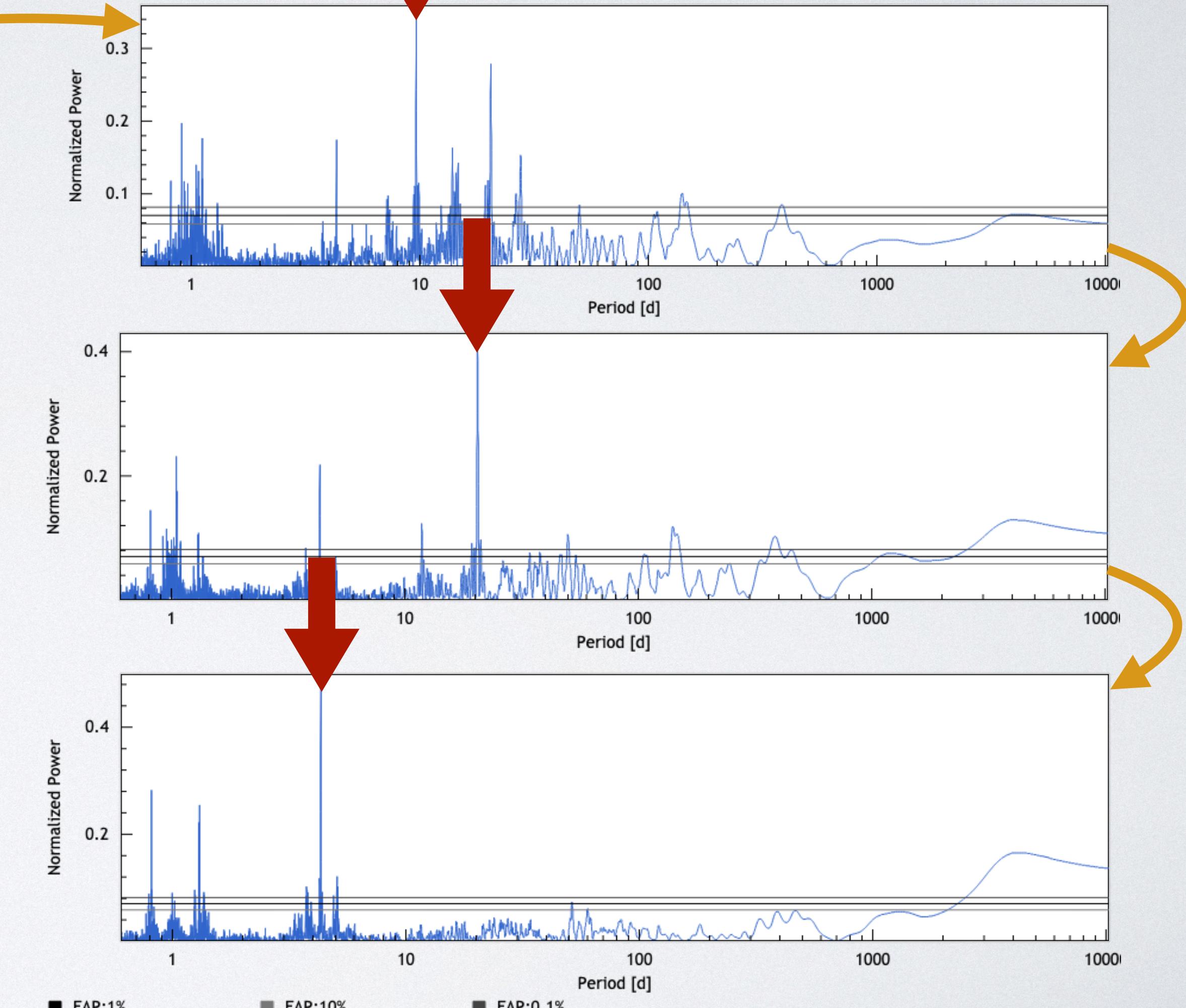
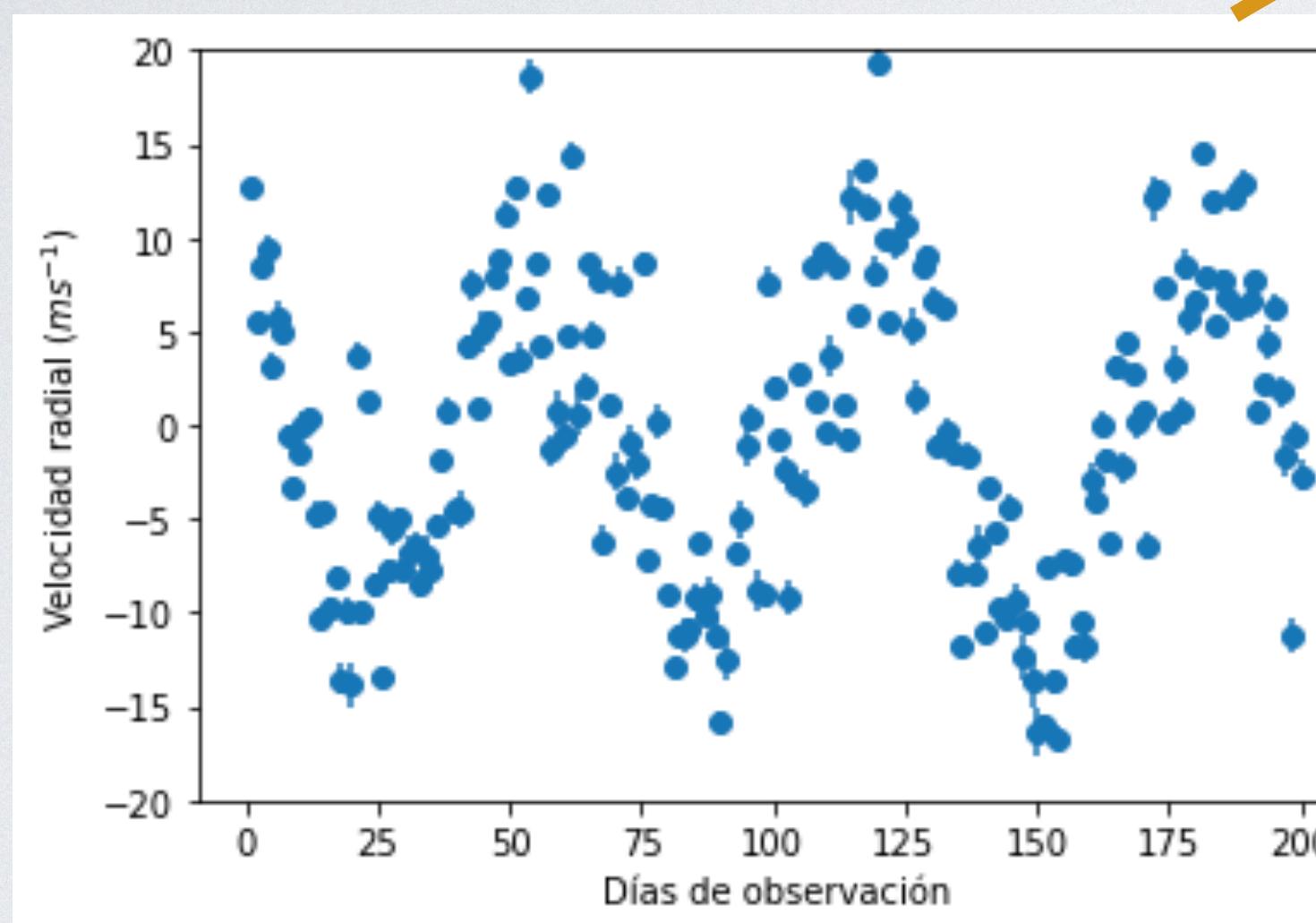
DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.



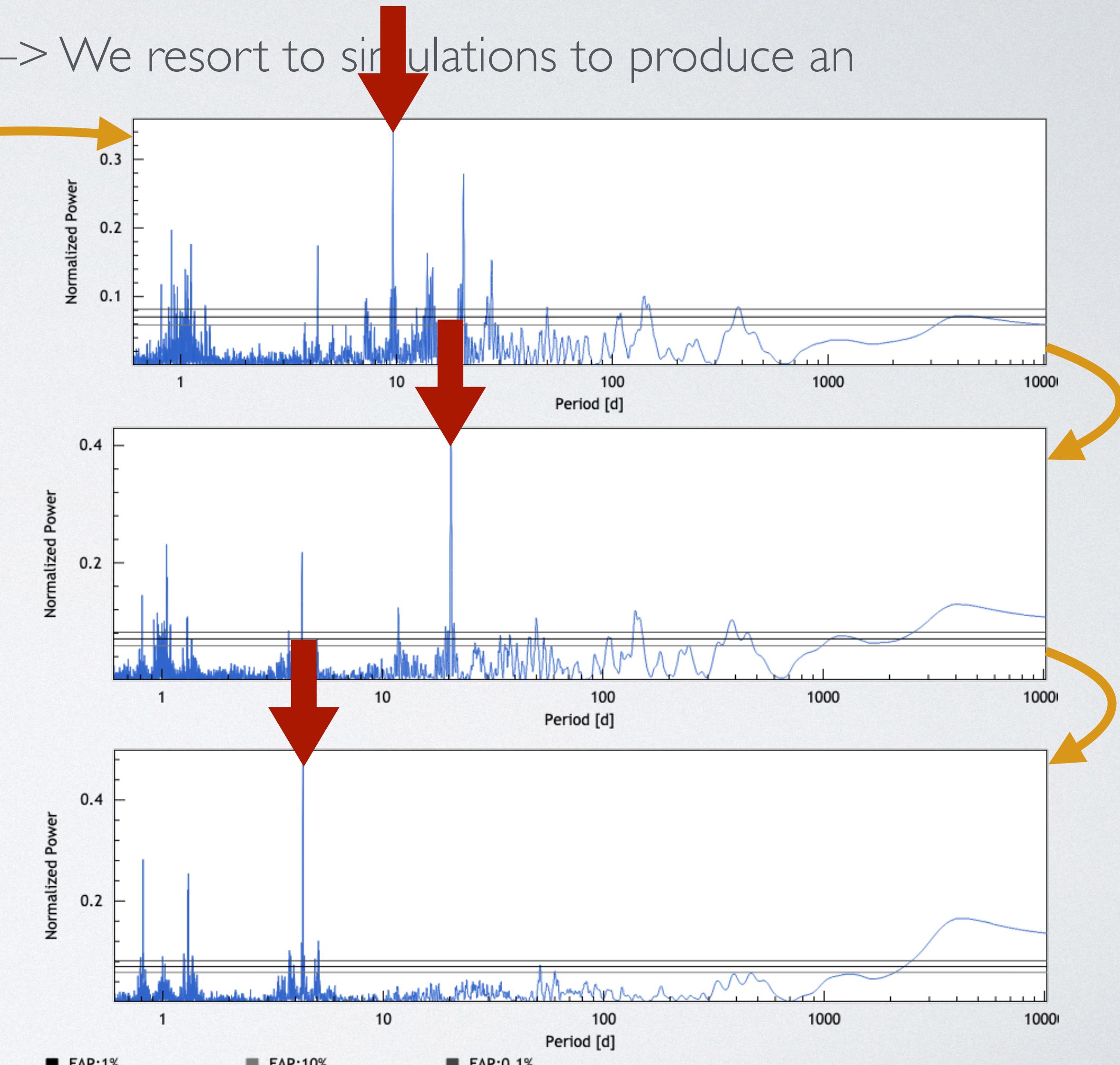
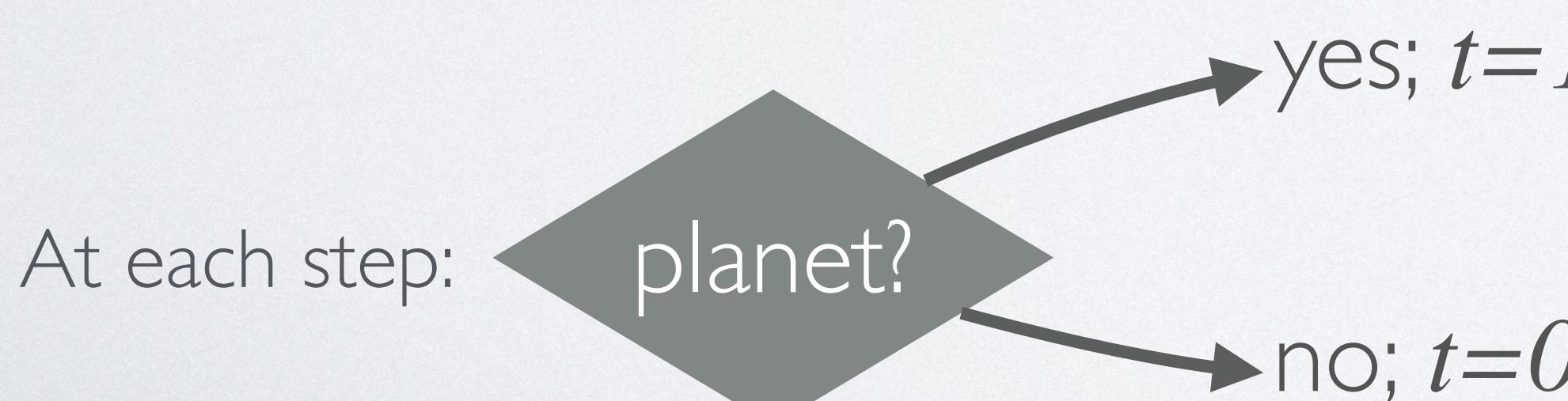
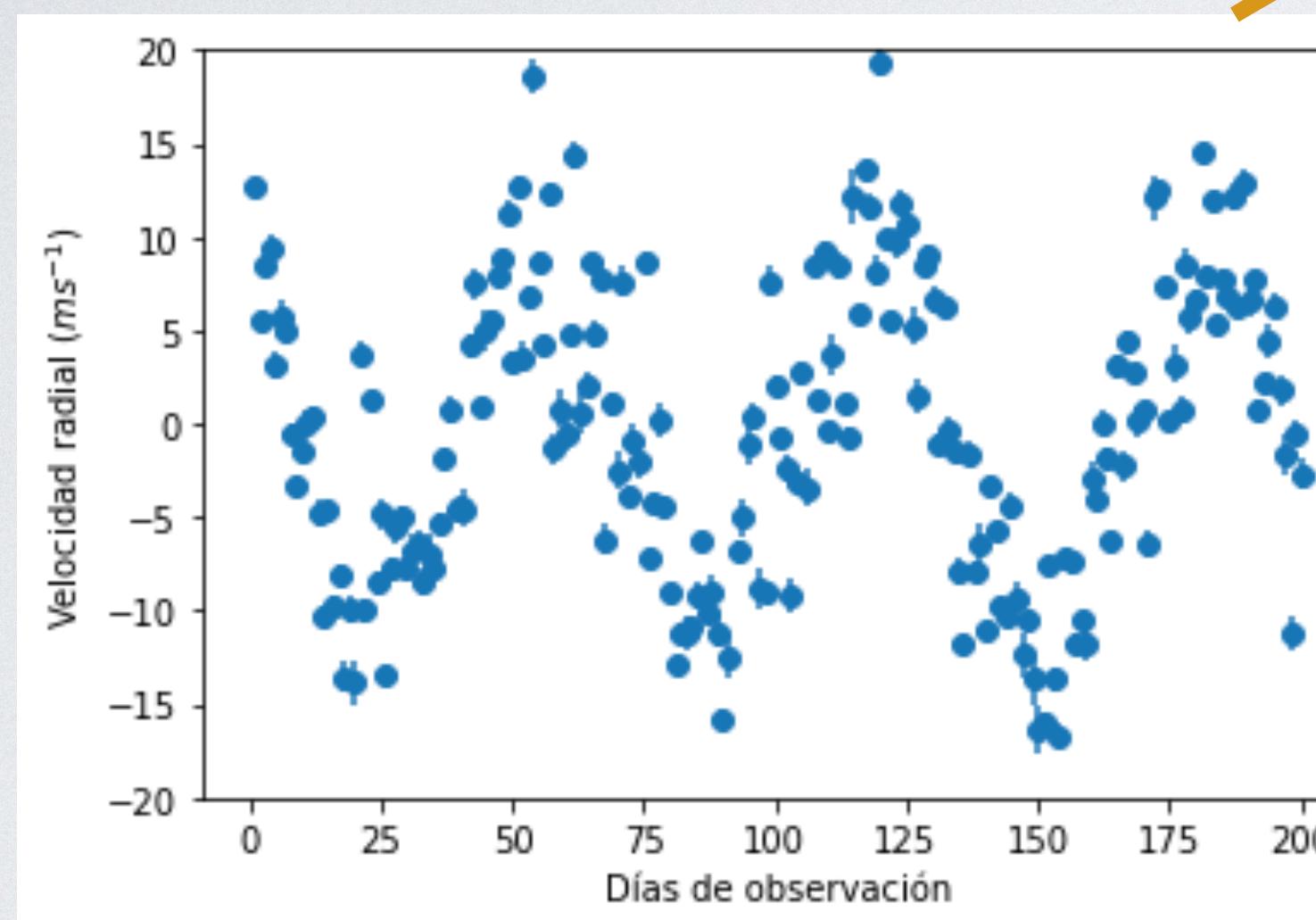
DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.



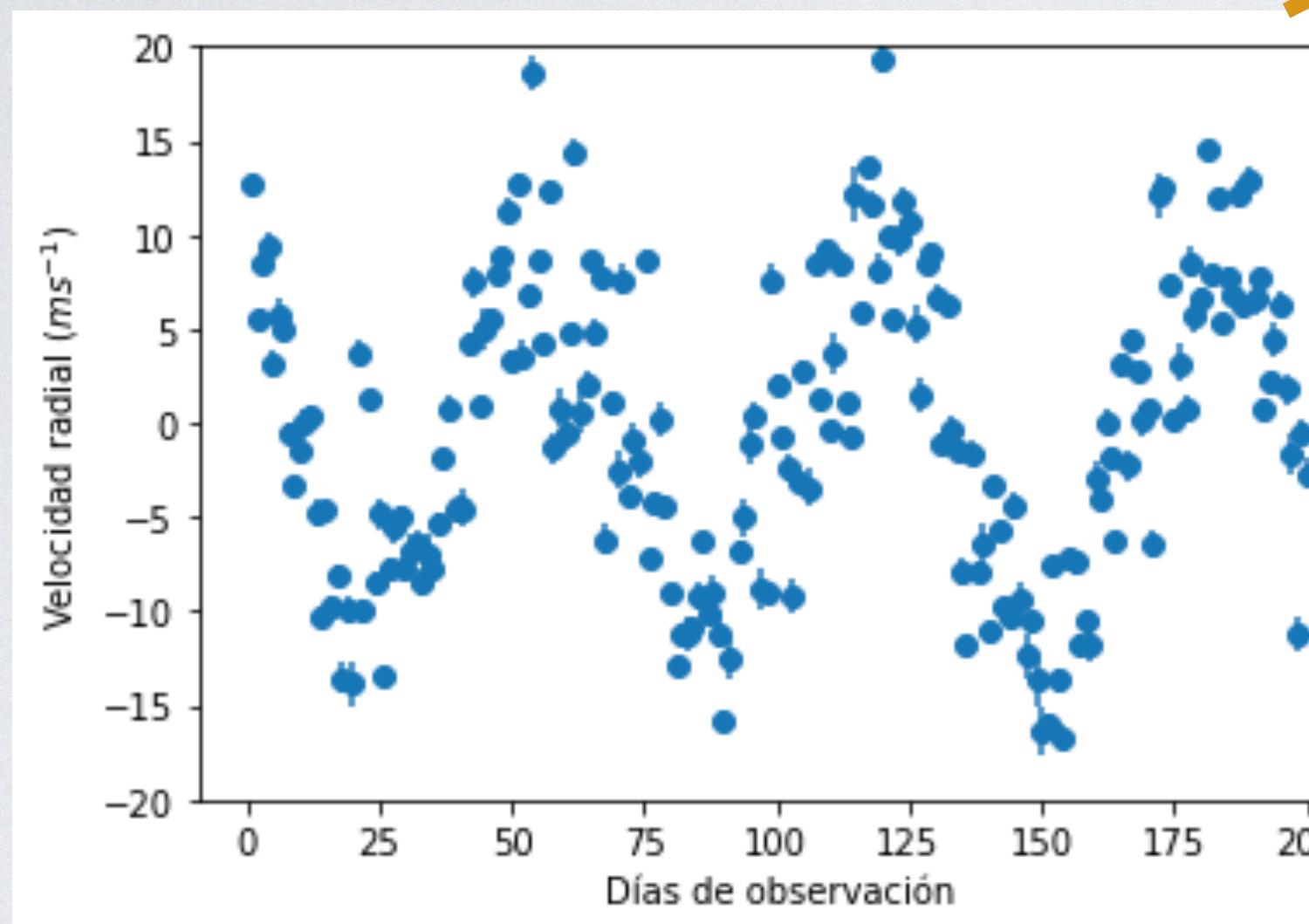
DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.



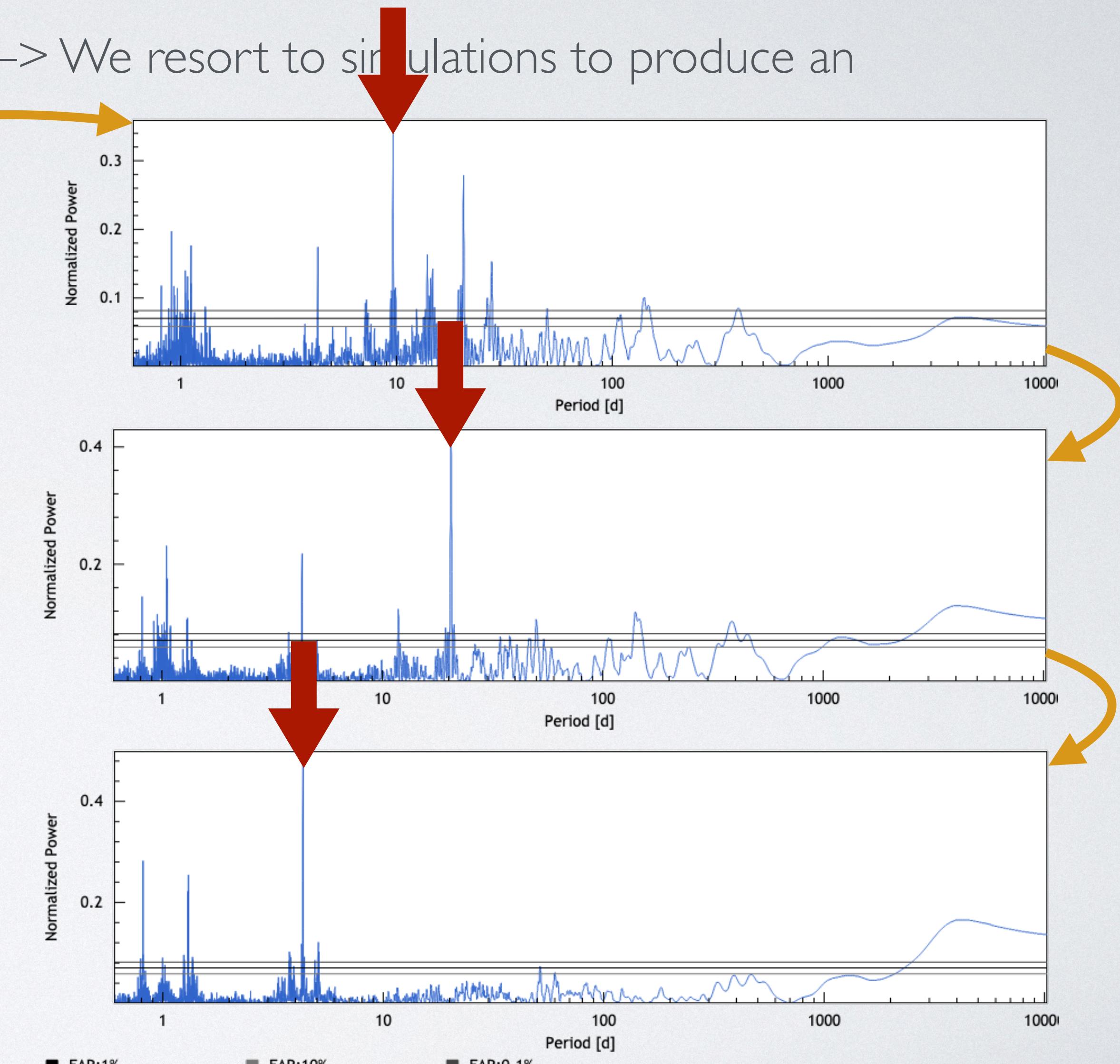
DATA

RV surveys do not provide LARGE amounts of data. —> We resort to simulations to produce an appropriate training set.

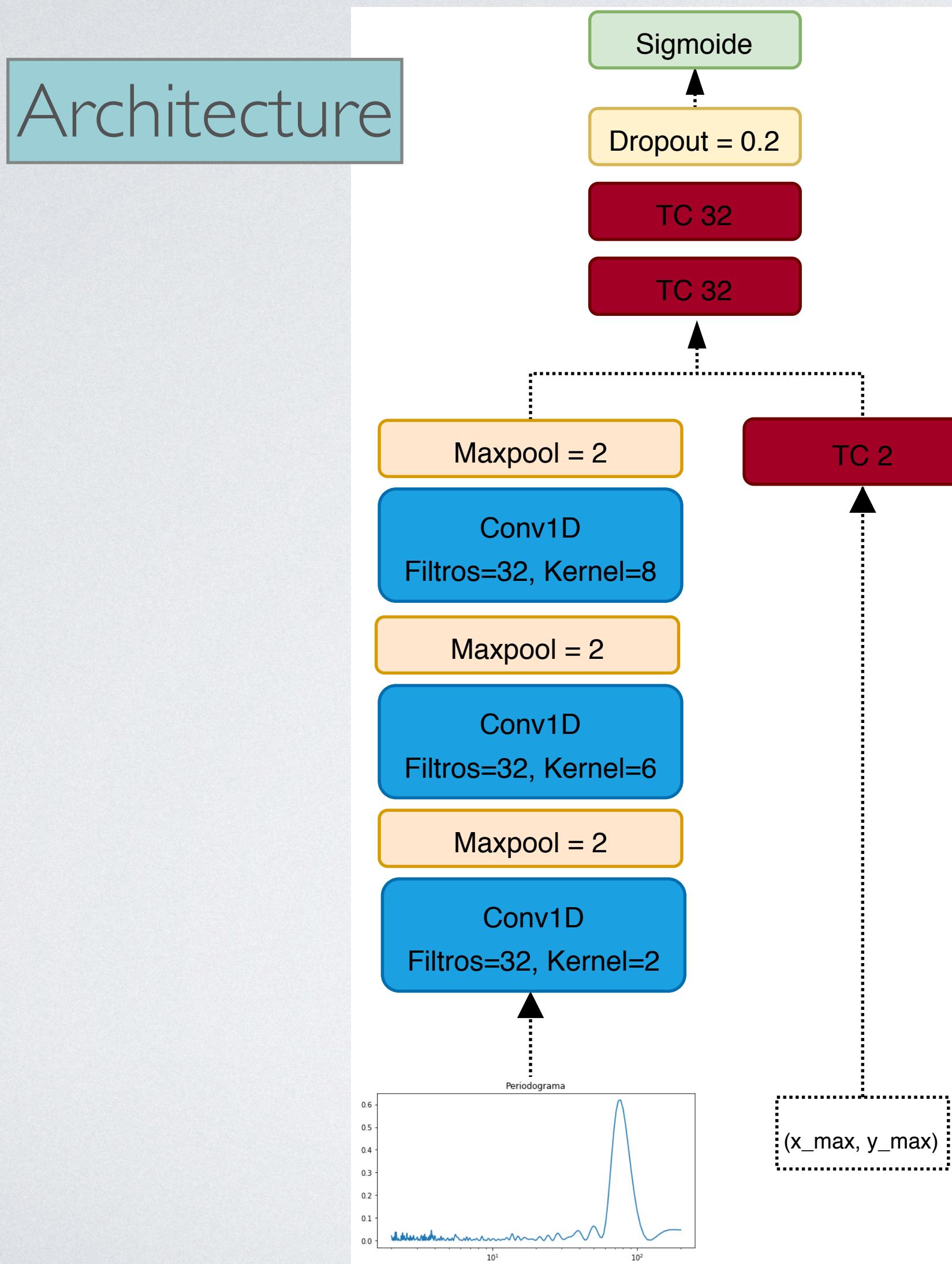


Training set: 13700 periodograms (3425 stars)
Unbalanced: around 40% of positive cases.

Test sets (2):
2500 stars (1e4 GLS) + 5000 stars (2e4 GLS)



OUR ML MODEL



- implemented in Tensorflow / Keras
- reLU activation functions

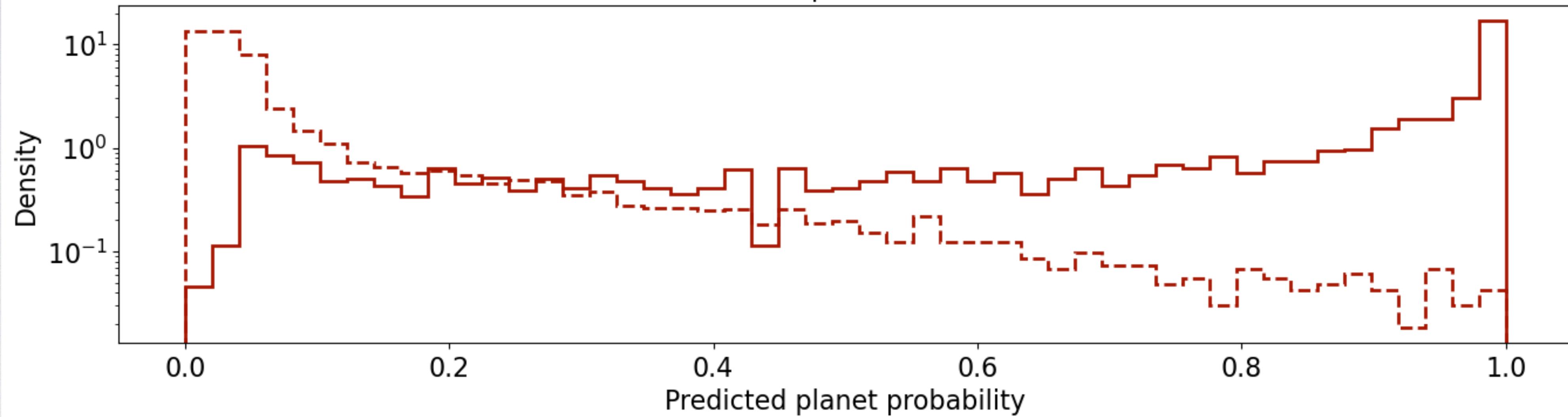
Training

- binary crossentropy loss function.
- Adam optimiser
- batch size = 16

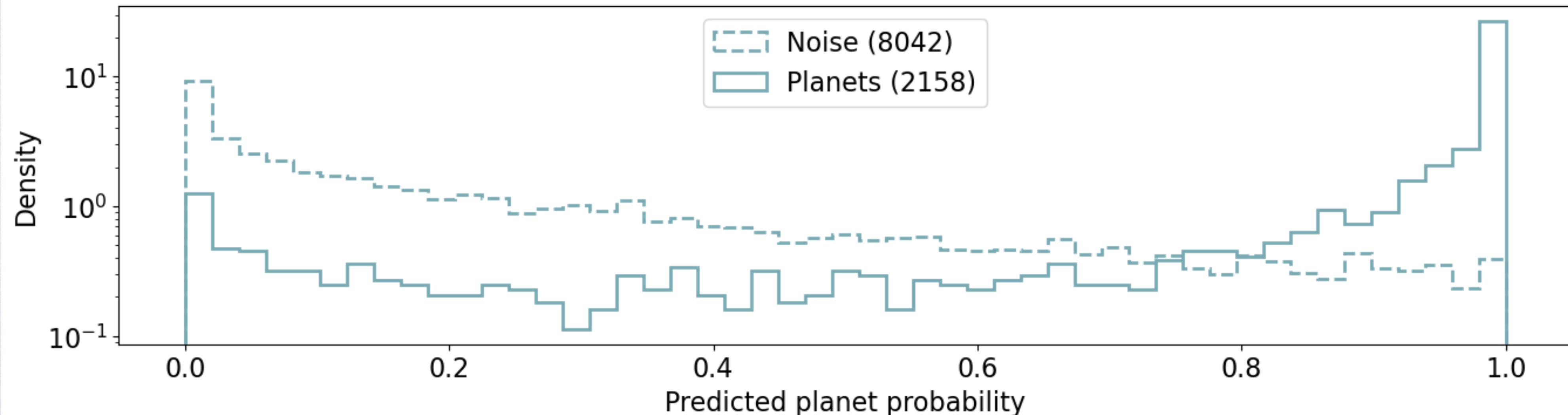
Name....

Predicted probabilities

ExoPIANNET



FAP



$$\text{precisión} = \frac{TP}{TP + FP}$$

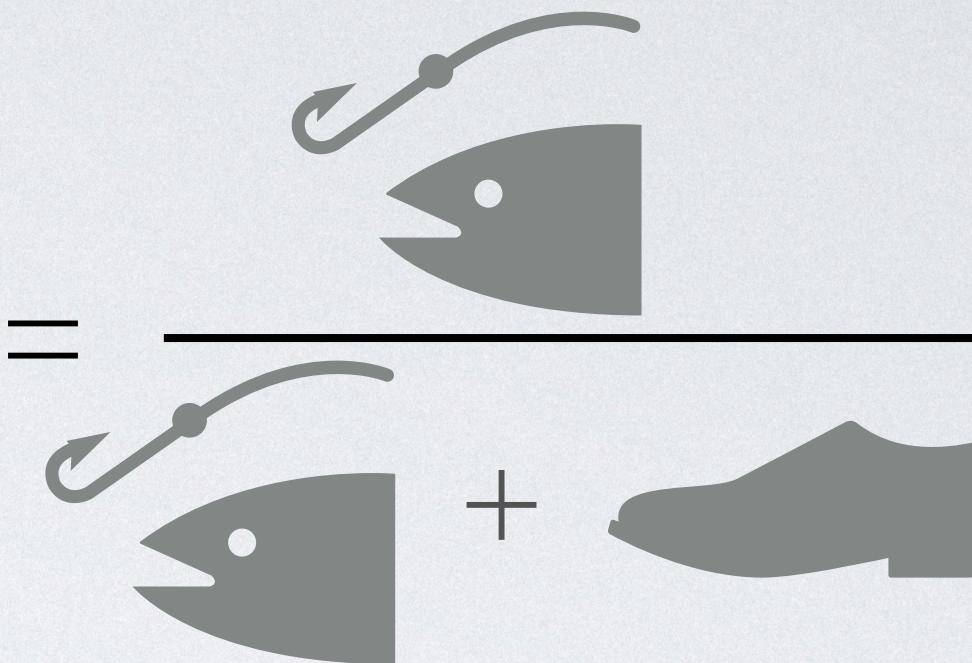
$$\text{recall} = \text{exhaustividad} = \frac{TP}{TP + FN}$$

$$\text{precisión} = \frac{TP}{TP + FP}$$

instances classified as
positive (planets)

$$\text{recall} = \text{exhaustividad} = \frac{TP}{TP + FN}$$

true positive
instances

$$\text{precisión} = \frac{TP}{TP + FP} = \frac{\text{ }$$


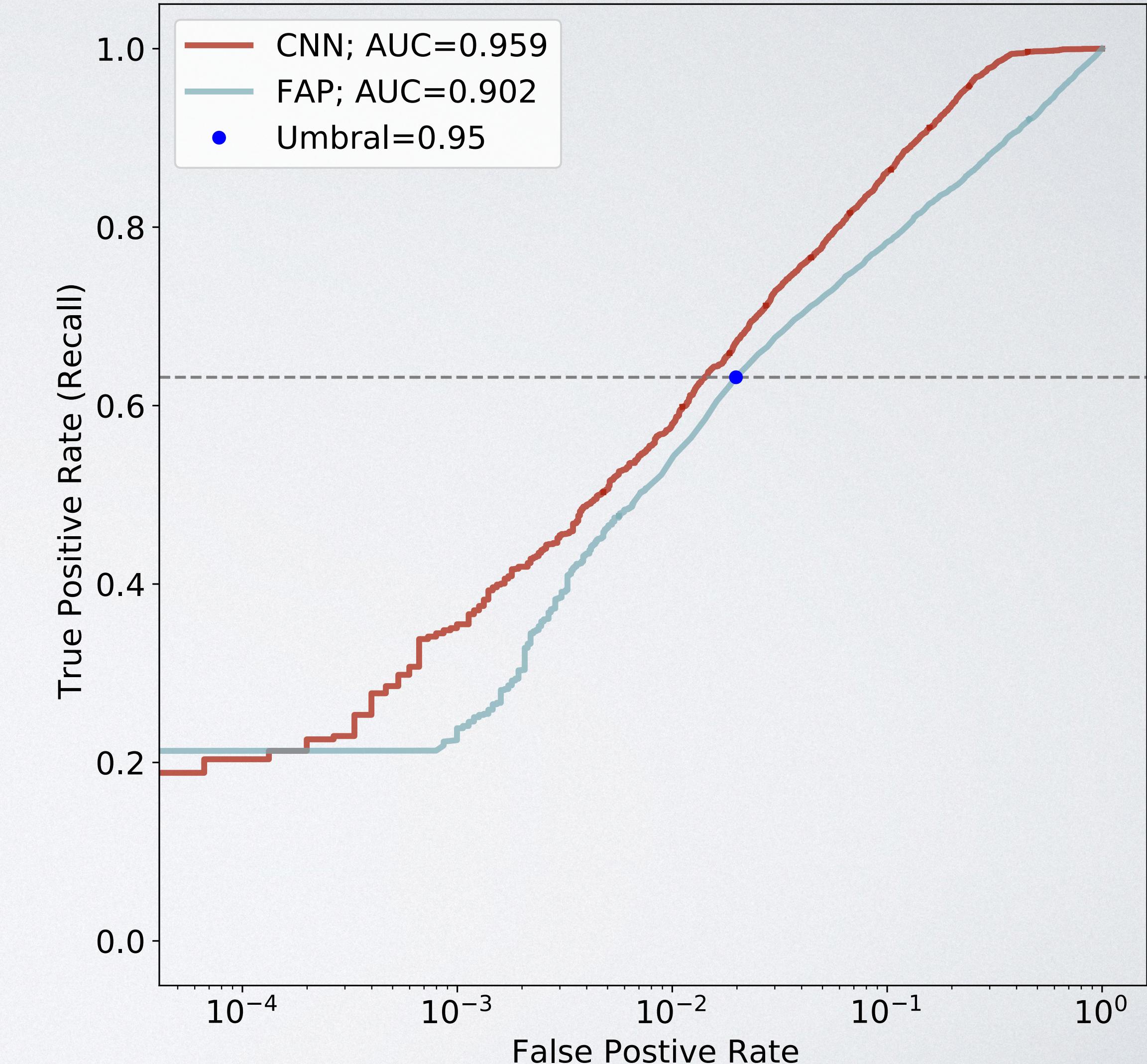
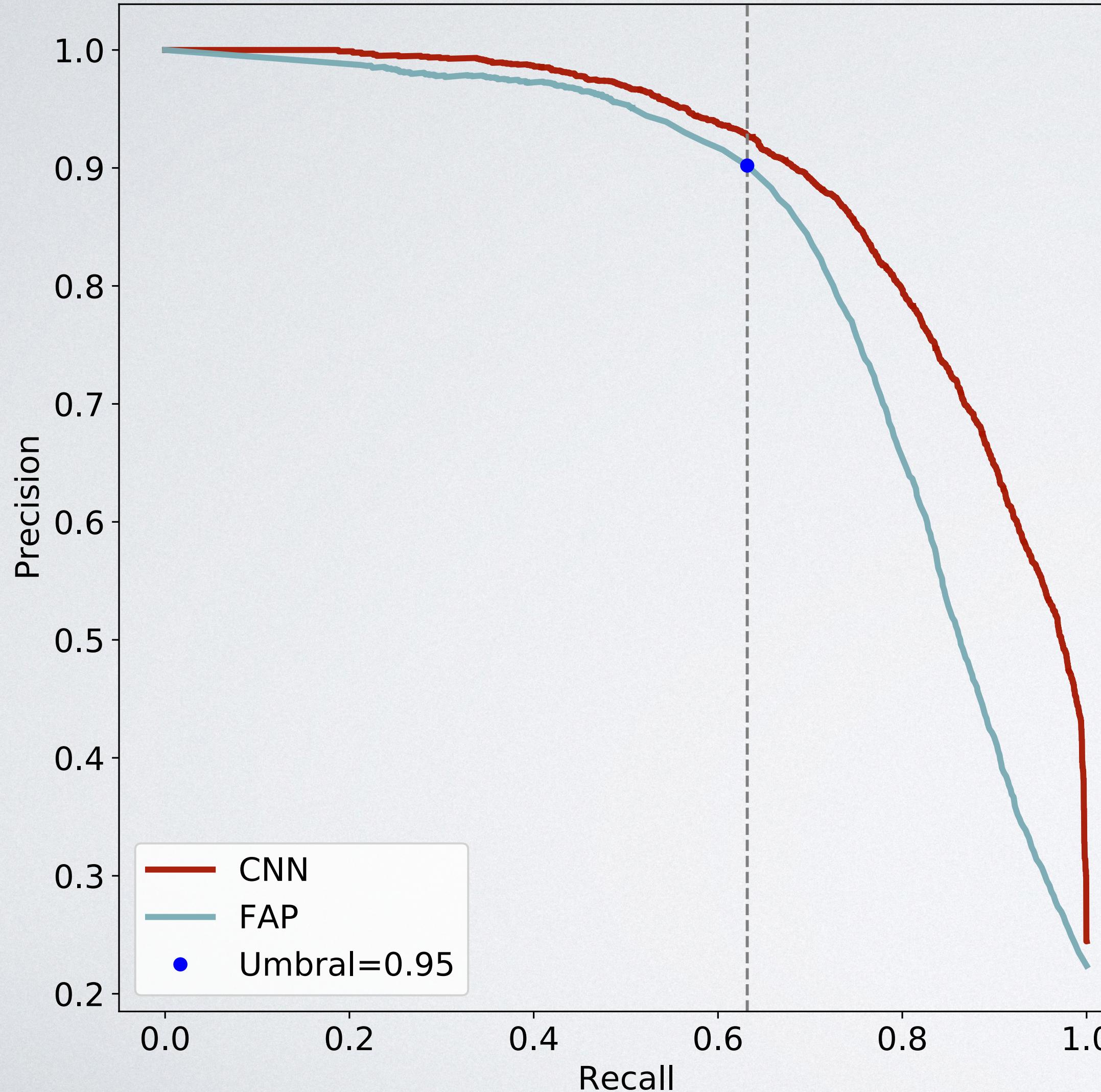
$$\text{recall} = \text{exhaustividad} = \frac{TP}{TP + FN}$$

true positive
instances

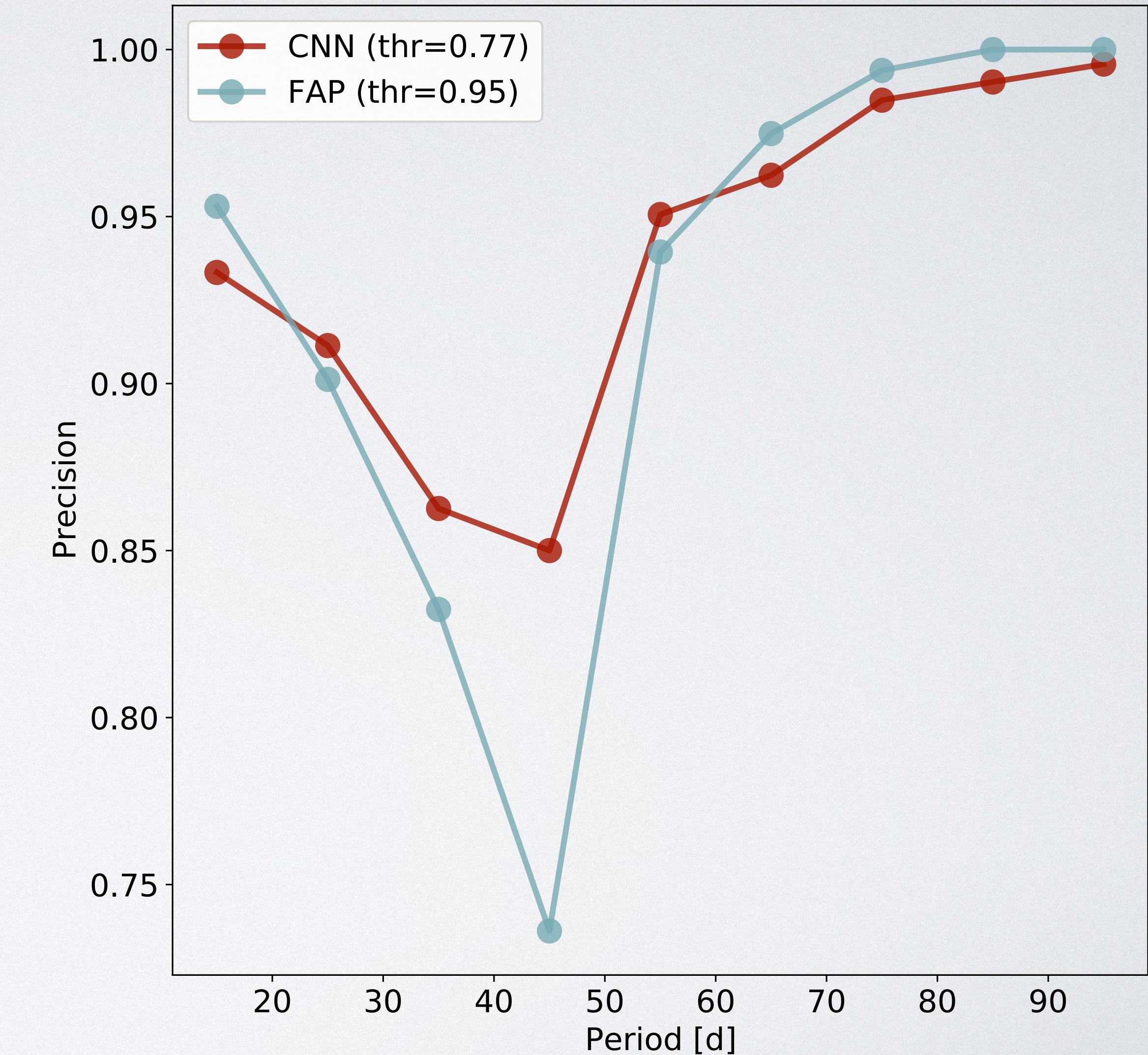
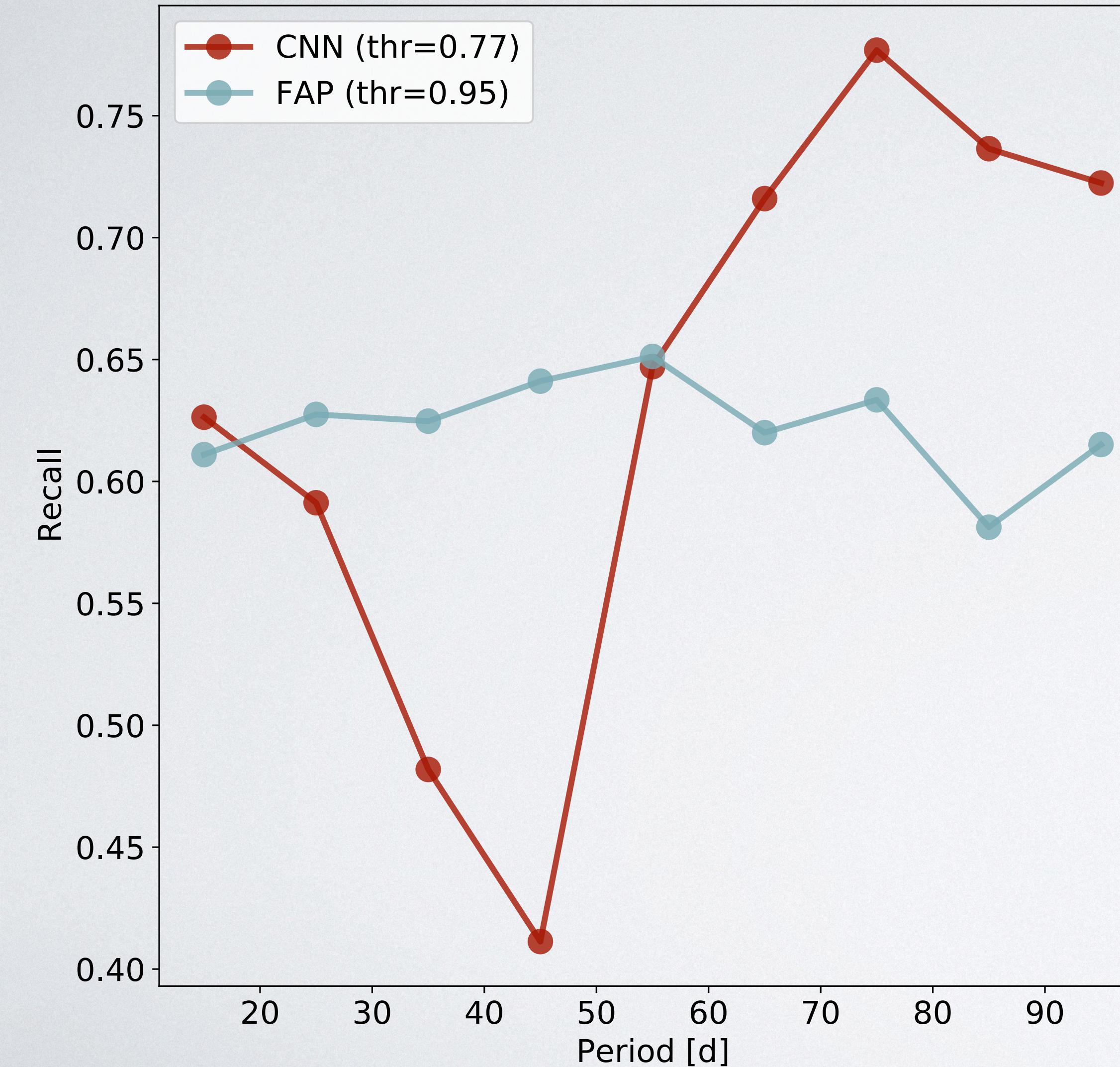
$$\text{precisión} = \frac{TP}{TP + FP} = \frac{\text{fish}}{\text{fish} + \text{shoe}}$$

$$\text{recall} = \text{exhaustividad} = \frac{TP}{TP + FN} = \frac{\text{fish}}{\text{fish} + \text{trout}}$$

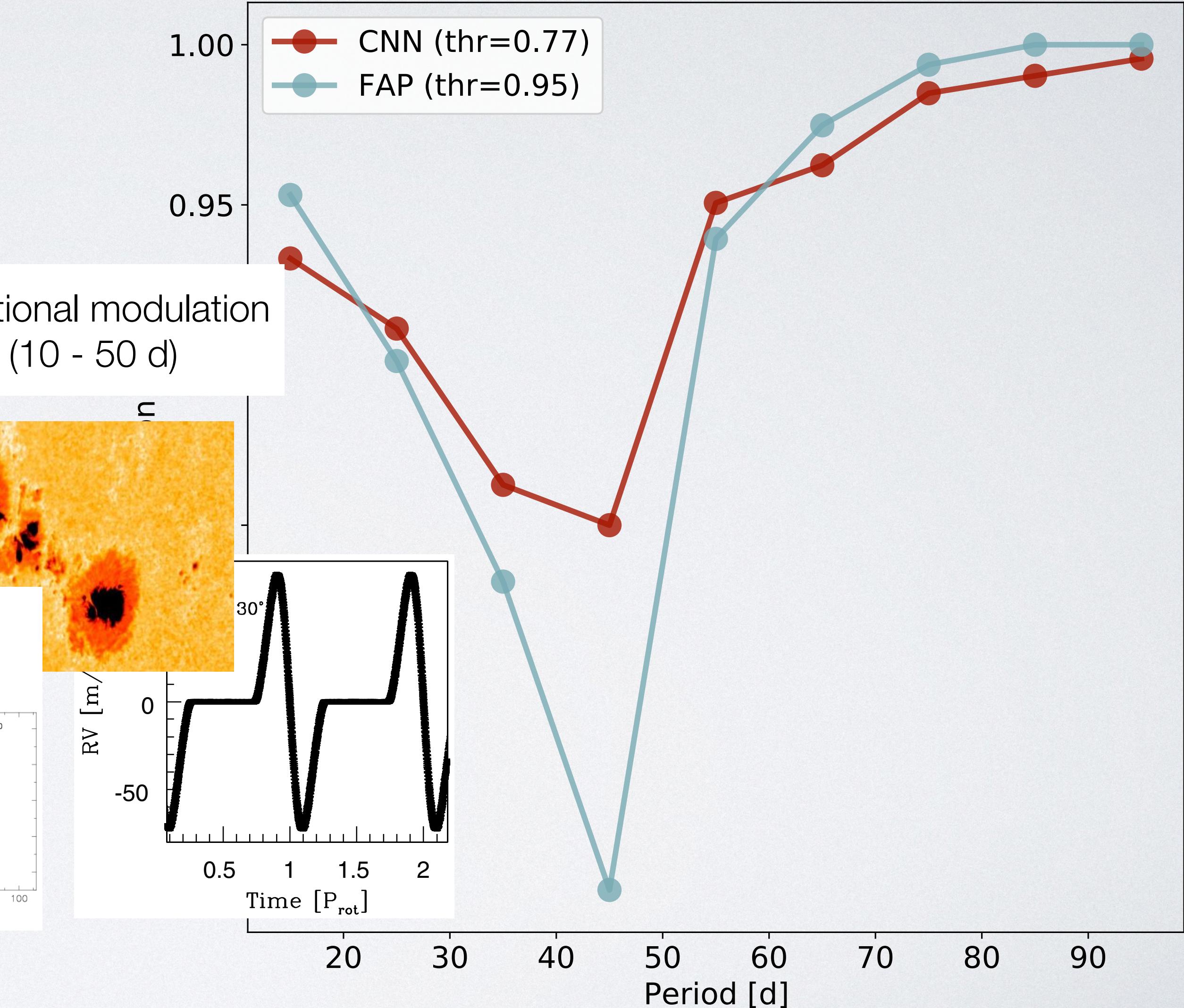
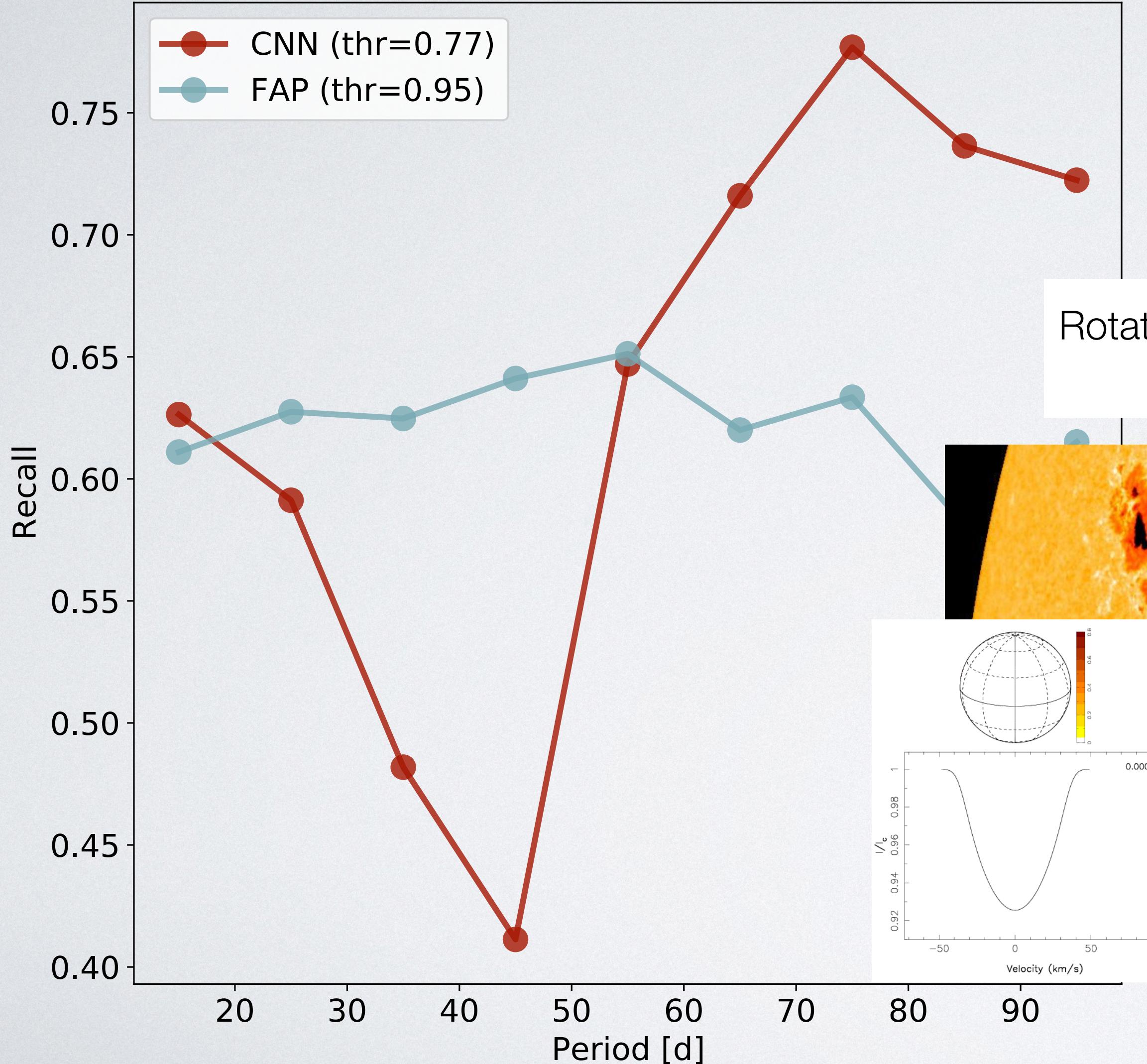
ExoplANNNet outperforms traditional method (FAP)

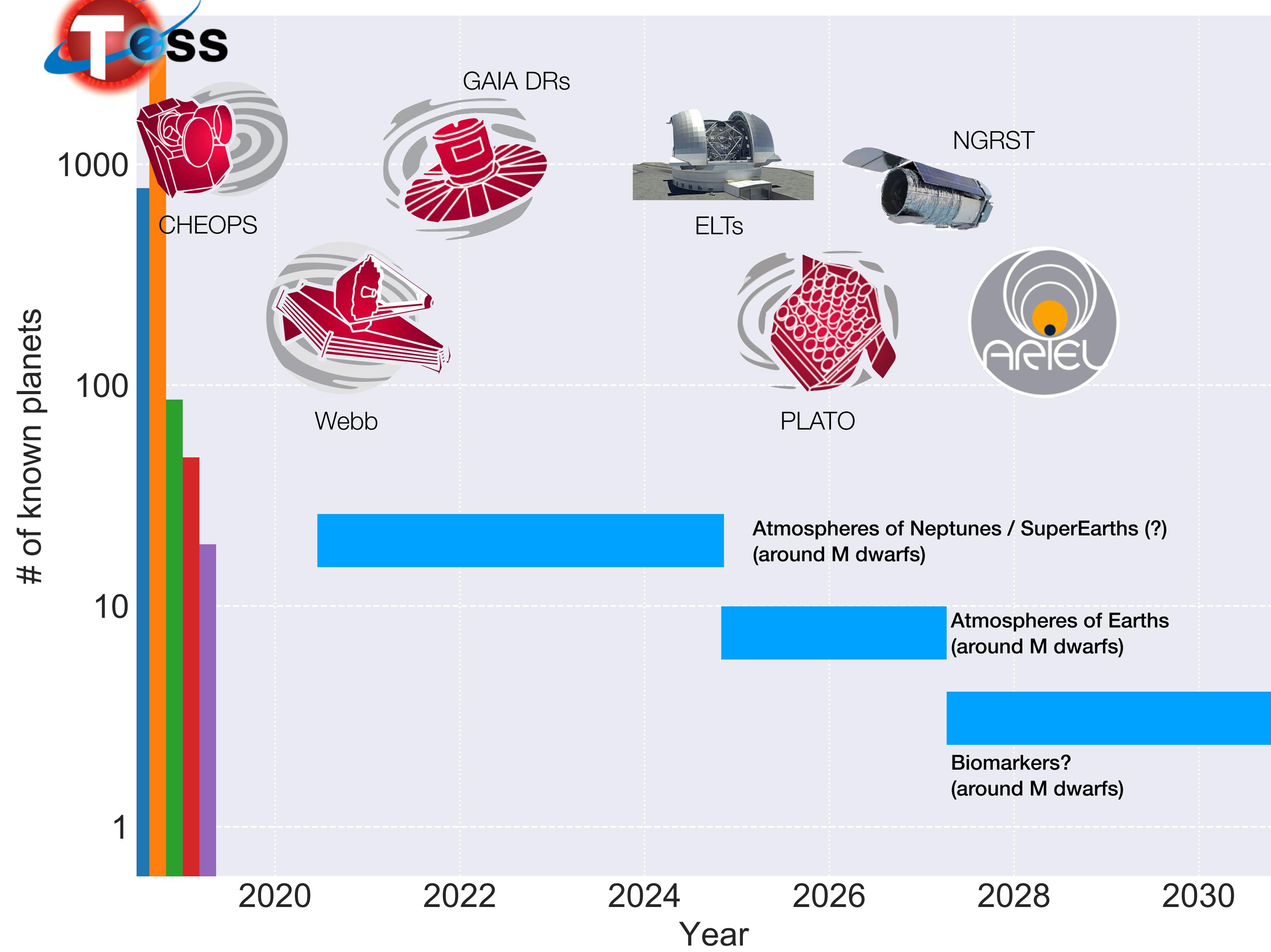


Dependence with data input parameters



Is ExoplANNNet avoiding problems close to the rotational period?





SUMMARY

- The last 25 years brought a wealth of information about planets outside the Solar system.
- Many questions remain open. Chief among them, is the occurrence of planets like Earth.
- Our exoplanet team at UNSAM uses data science and machine learning techniques to solve some of the outstanding questions in exoplanet science, by improving instrument performance, operation efficiency and / or detection power.

