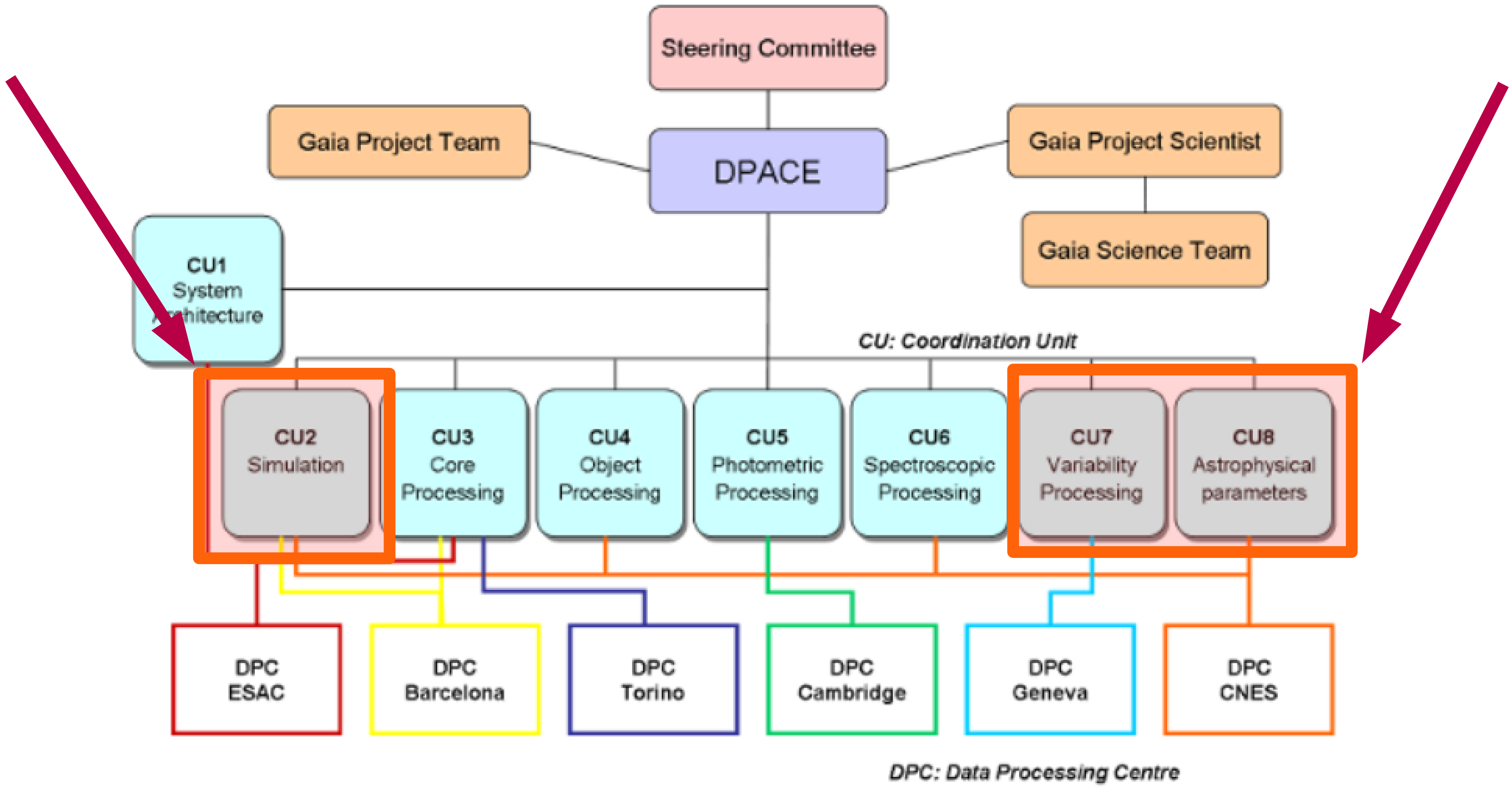
Minería de Datos y Astroestadística: Perspectivas futuras

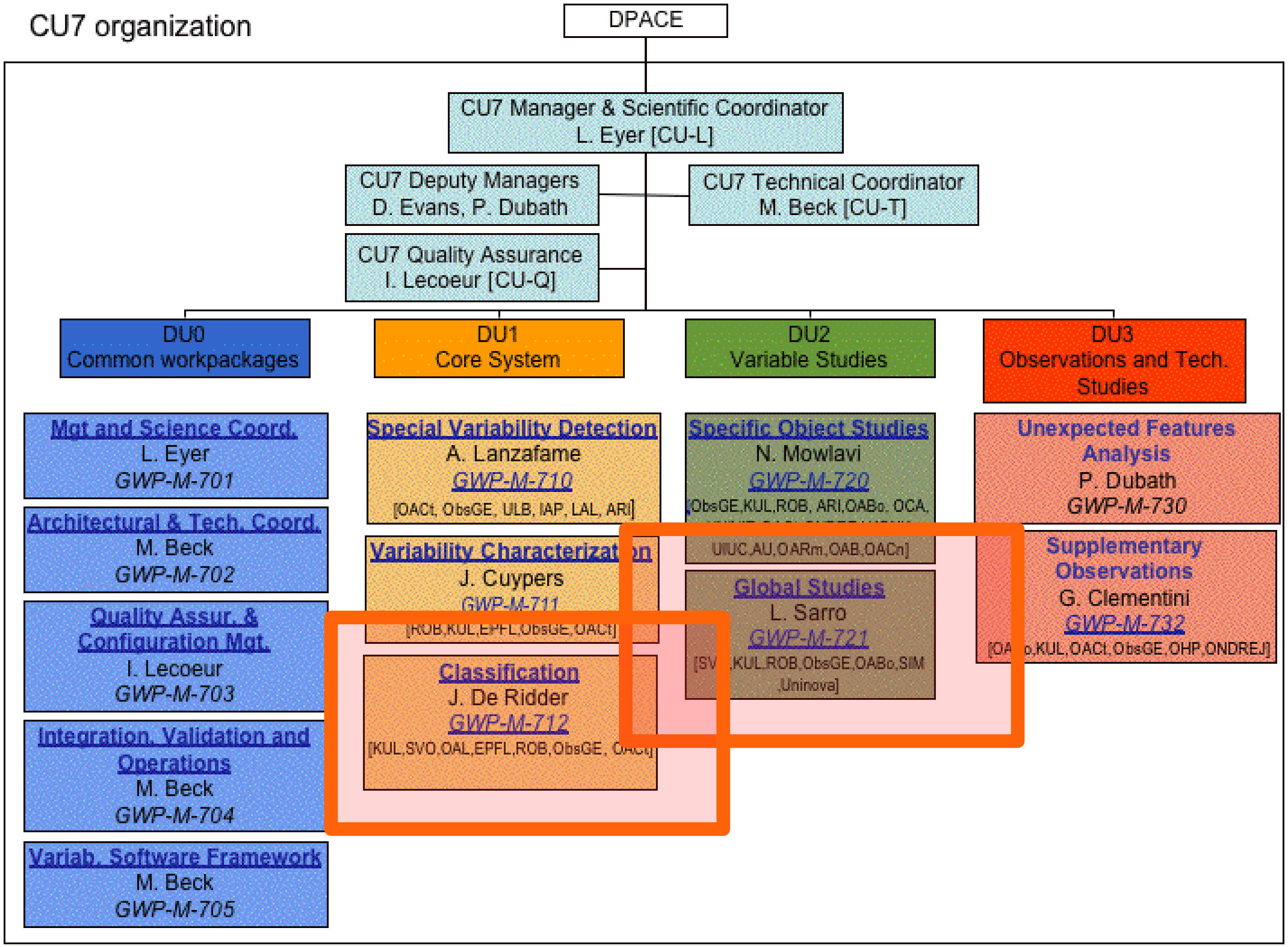
Luis M. Sarro – Dpt. Inteligencia Artificial (UNED)

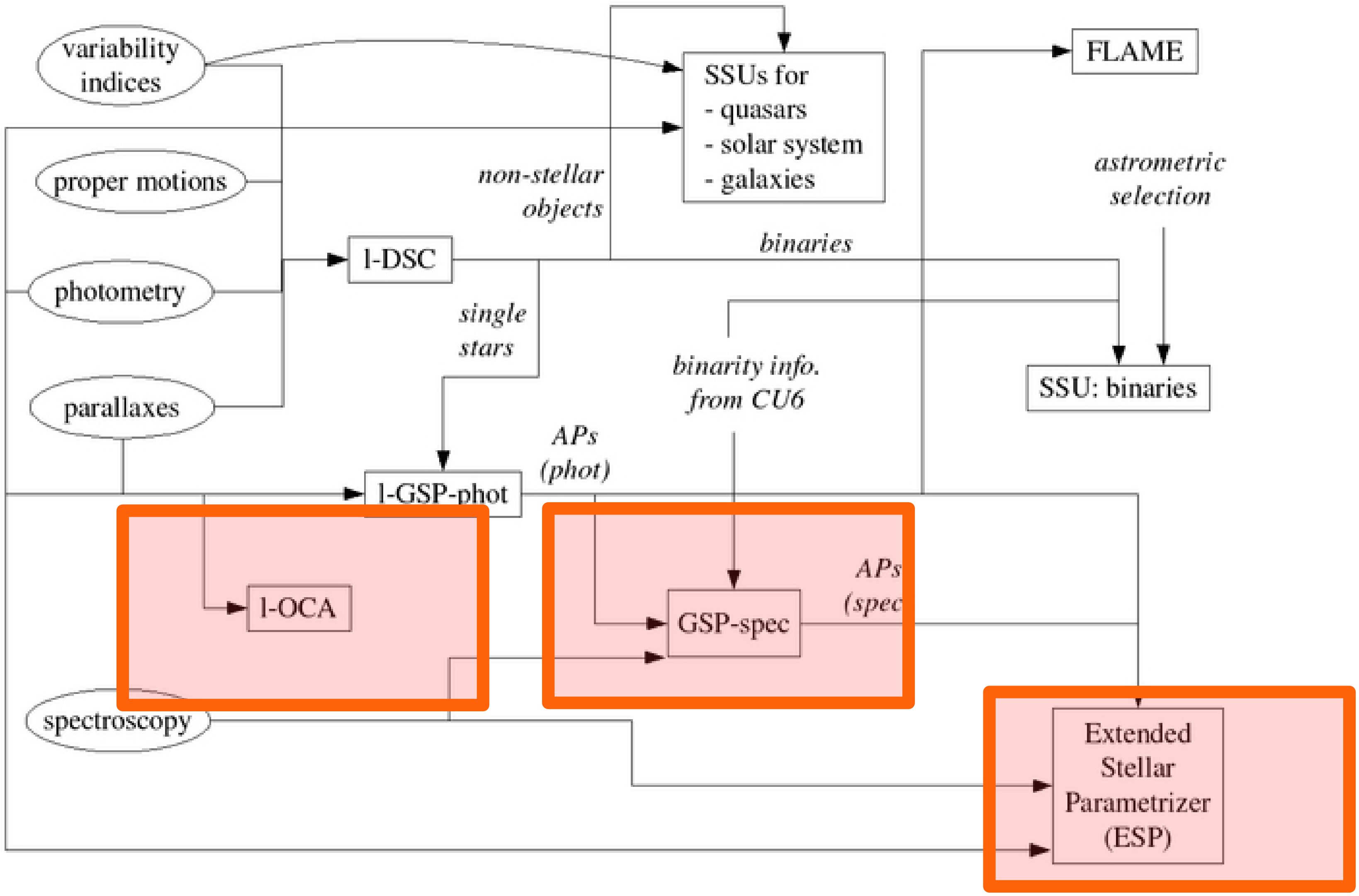
1.- Implicación española en el Gaia DPAC
2.- Gaia-ESO survey
3.- Modelos jerárquicos



Steering Committee
Gaia Project Team
DPACE
Gaia Project Scientist
Gaia Science Team
CU1
System
Architecture
CU: Coordination Unit
CU2
Simulation
CU3
Core
Processing
CU4
Object
Processing
CU5
Photometric
Processing
CU6
Spectroscopic
Processing
CU7
Variability
Processing
CU8
Astrophysical
parameters
DPC
ESAC
DPC
Barcelona
DPC
Torino
DPC
Cambridge
DPC
Geneva
DPC
CNES
DPC: Data Processing Centre

CU7 organization




variability indices
proper motions
photometry
parallaxes
spectroscopy
SSUs for
- quasars
- solar system
- galaxies
FLAME
non-stellar objects
astrometric selection
binaries
1-DSC
single stars
binarity info. from CU6
SSU: binaries
APs (phot)
1-GSP-phot
1-OCA
GSP-spec
APs (spec
Extended Stellar Parametrizer (ESP)

Desarrollos:

1.- Clasificación supervisada

2.- Predicción de parámetros físicos

3.- Detección de objetos exóticos y nuevas clases (clasificación no supervisada)

4.- Control de calidad

WG9 Clasificación espectral
WG10 Determinación de parámetros físicos

- FERRE
- MATISSE
- BIQUINI

WG12 PMS?
WG13 OBA?

Congreso/Escuela de Astroestadística y Minería de Datos

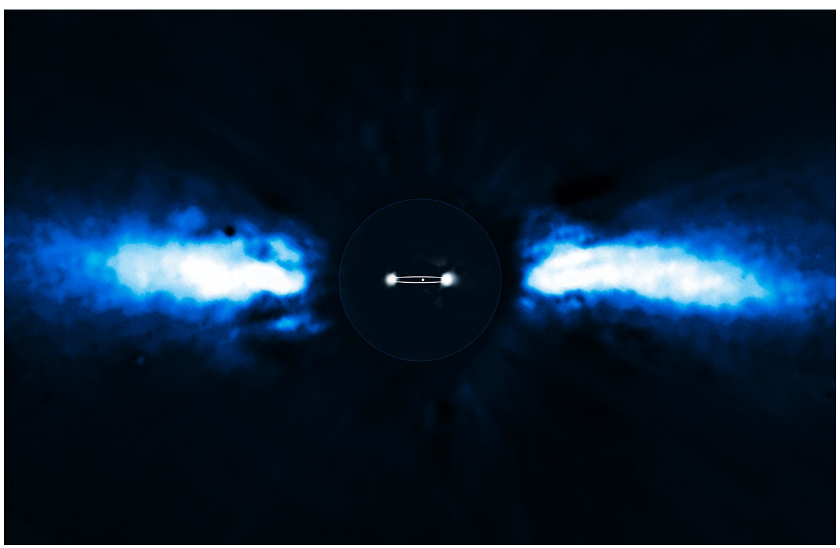
Dos problemas centrales en la explotación científica de Gaia:

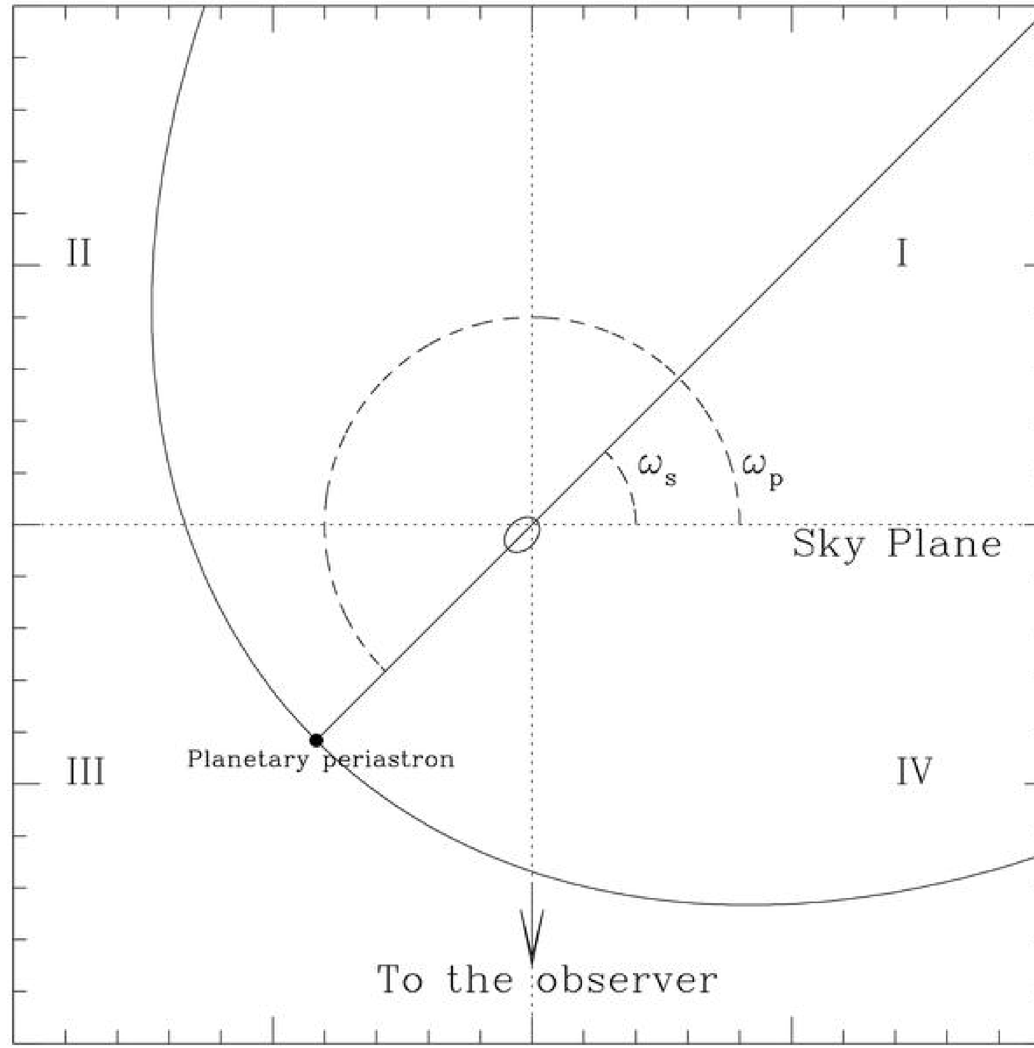
1. Determinar para una fuente observada (conjunto de datos D) sus parámetros físicos ω
2. Determinar, para un conjunto de fuentes (muestra), las distribuciones estadísticas (población $p(\{\omega_j\})$) que generaron la muestra.

Ejemplos:
[1]
Parámetros físicos: Teff, logg, [M/H], A_V,P,...
Membresía a grupo móvil, población, cúmulo...
[2]
La función inicial de masas
La tasa de formación estelar
Los parámetros que definen la barra, el bulbo, un grupo móvil...

Modelos jerárquicos

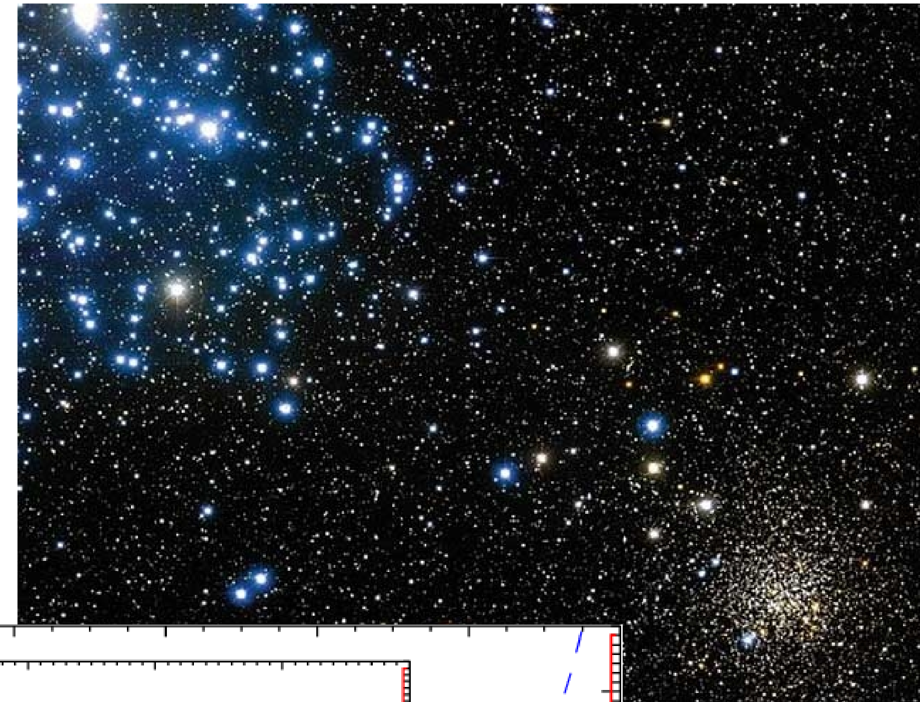
	Ejemplo 1
Fuente	Estrella+exoplaneta
Datos ($\{D_i\}$)	Vr(t)
Parámetros físicos ($\{\omega_j\}$)	e, T, ω, φ, κ, i
Distribuciones	IEF (Función Inicial de Excentricidades)=p(e)

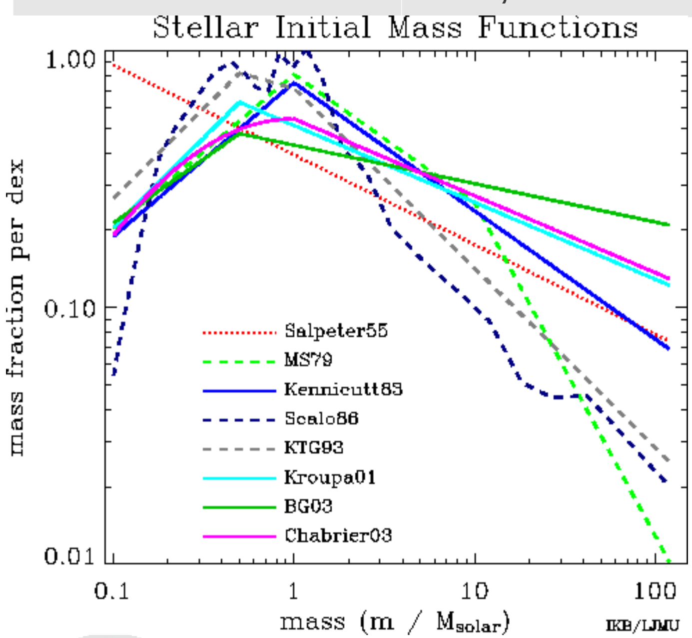


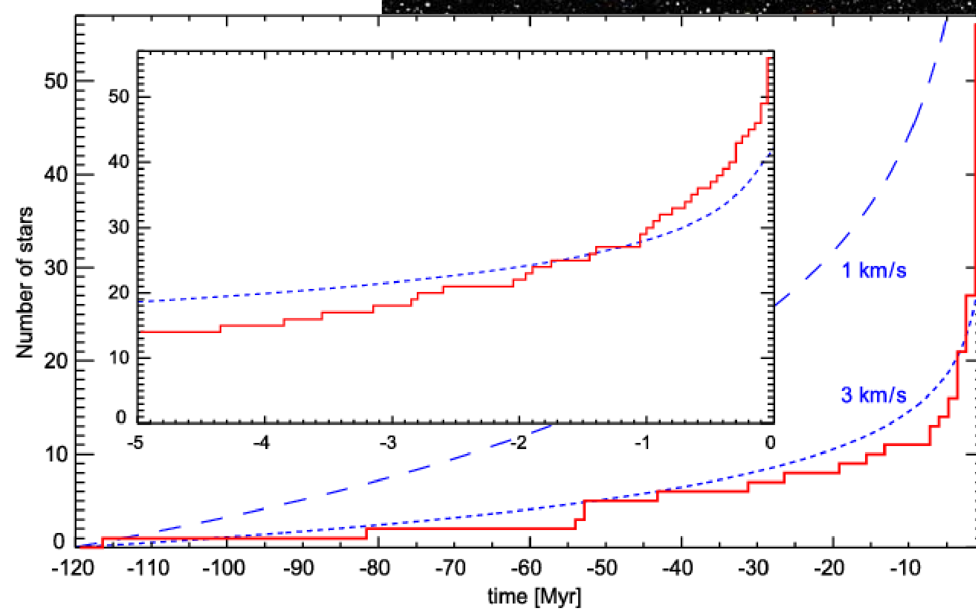


Modelos jerárquicos

	Ejemplo 2
Fuente	Estrella en cúmulo
Datos ({D_i})	BP/RP, astrometría,RVS
Parámetros físicos ({ω_j})	Probabilidad de pertenencia al cúmulo y parámetros físicos
Distribuciones	IMF, SRF

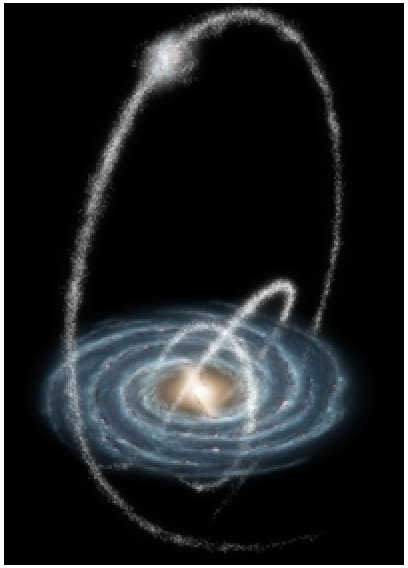


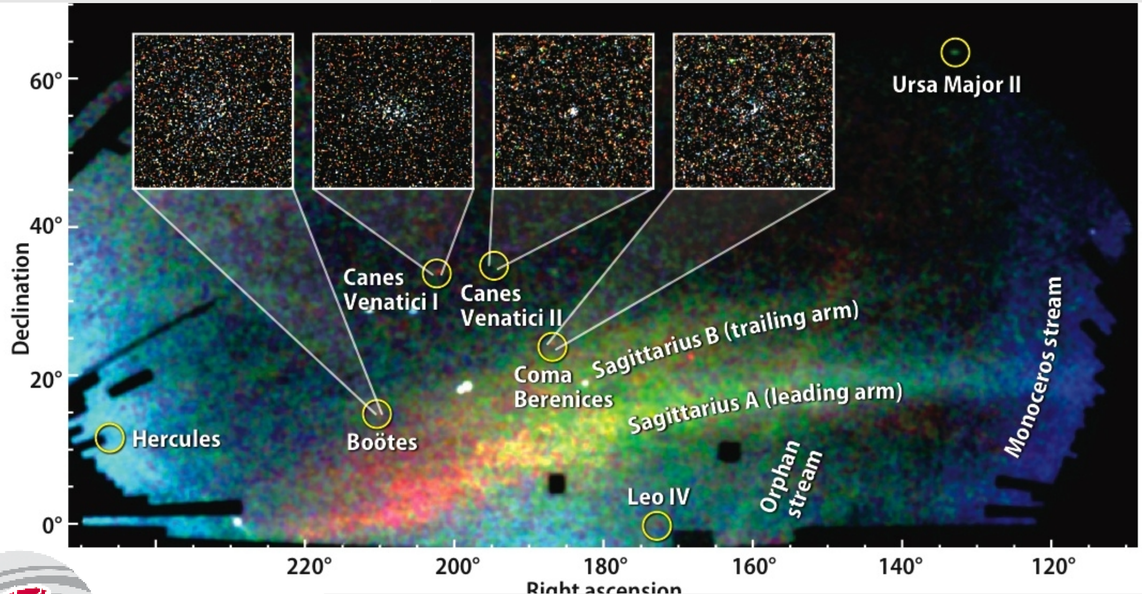




Modelos jerárquicos

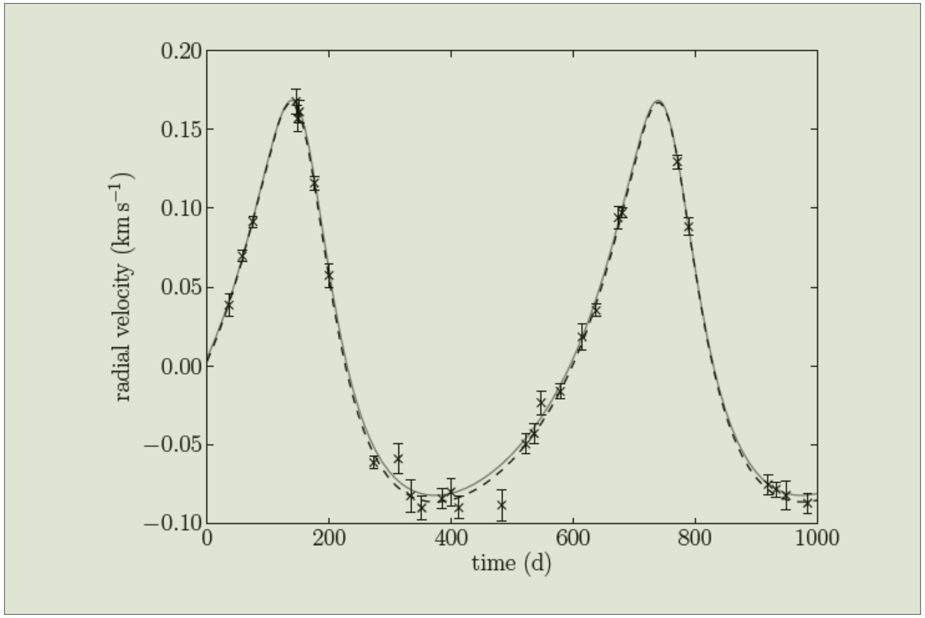
	Ejemplo 3
Fuente	Estrella Galaxia
Datos ($\{D_i\}$)	Astrometría, BP/RP, RVS, espectro alta resolución
Parámetros físicos ($\{\omega_j\}$)	Temperaturas, gravedades, metalicidades, edades...
Distribuciones	IMF(t,**r**), episodios de fusión, Ω(t,**r**), componentes de la Galaxia, número y distribución

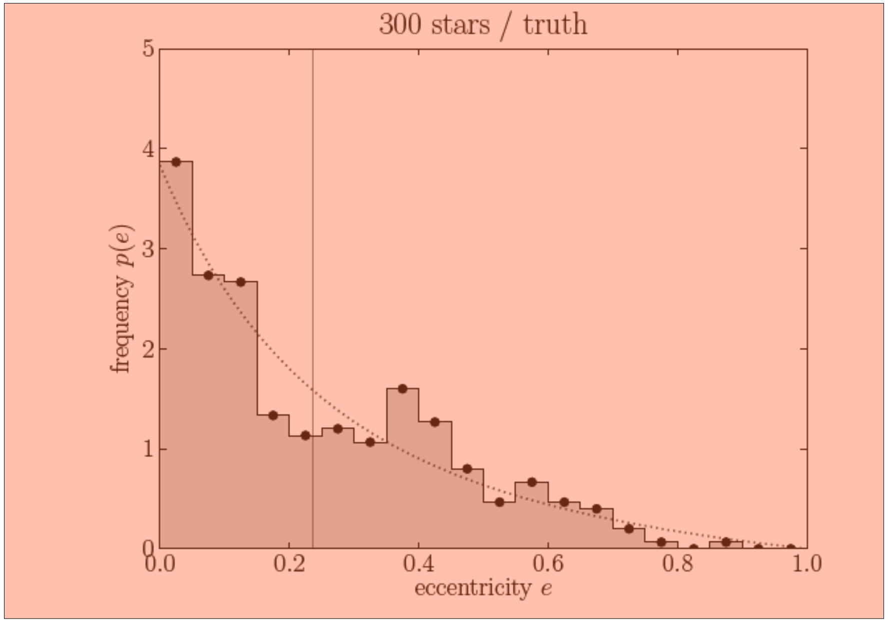




La inferencia Bayesiana clásica procede a inferir los parámetros (**ω**) a partir la verosimilitud y del conocimiento a priori:

$$p(\boldsymbol{\omega}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\omega}) \cdot p(\boldsymbol{\omega})}{p(\boldsymbol{D})}$$



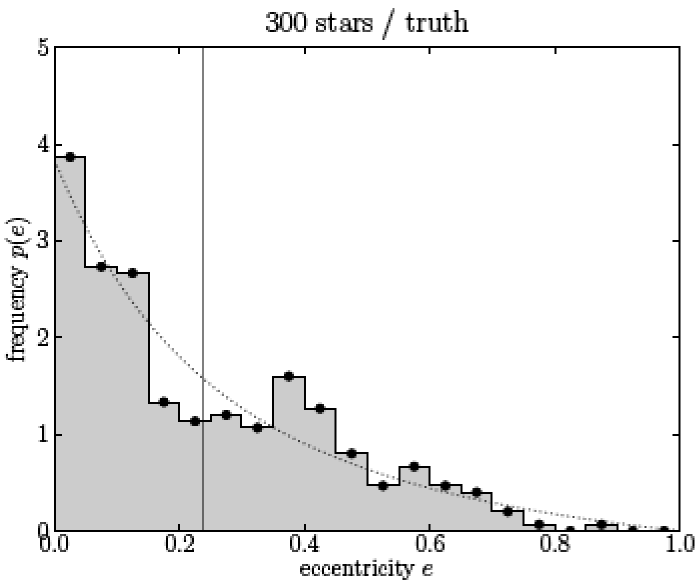


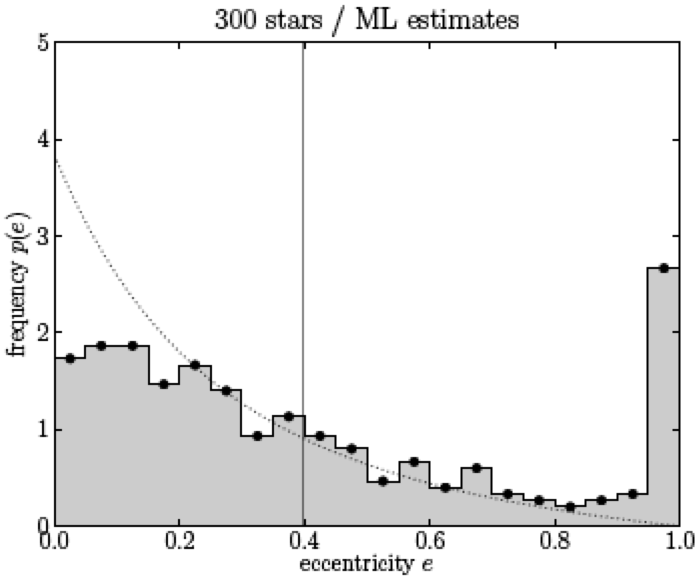
Supongamos que no estamos tan interesados en las propiedades particulares de un sistema estelar, sino en la distribución de una o varias propiedades (posiblemente en función de algún parámetro)
≡ Análisis estadístico

Alternativa 1: Obtenemos una estimación para cada uno de los sistemas de estudio y elaboramos un histograma o pdf equivalente.

Problema: los estimadores puntuales no sesgados son muy escasos. Máxima verosimilitud, MAP o las medianas de muestreos no lo son en general.

Además, no propagamos las incertidumbres.





Análisis estadístico

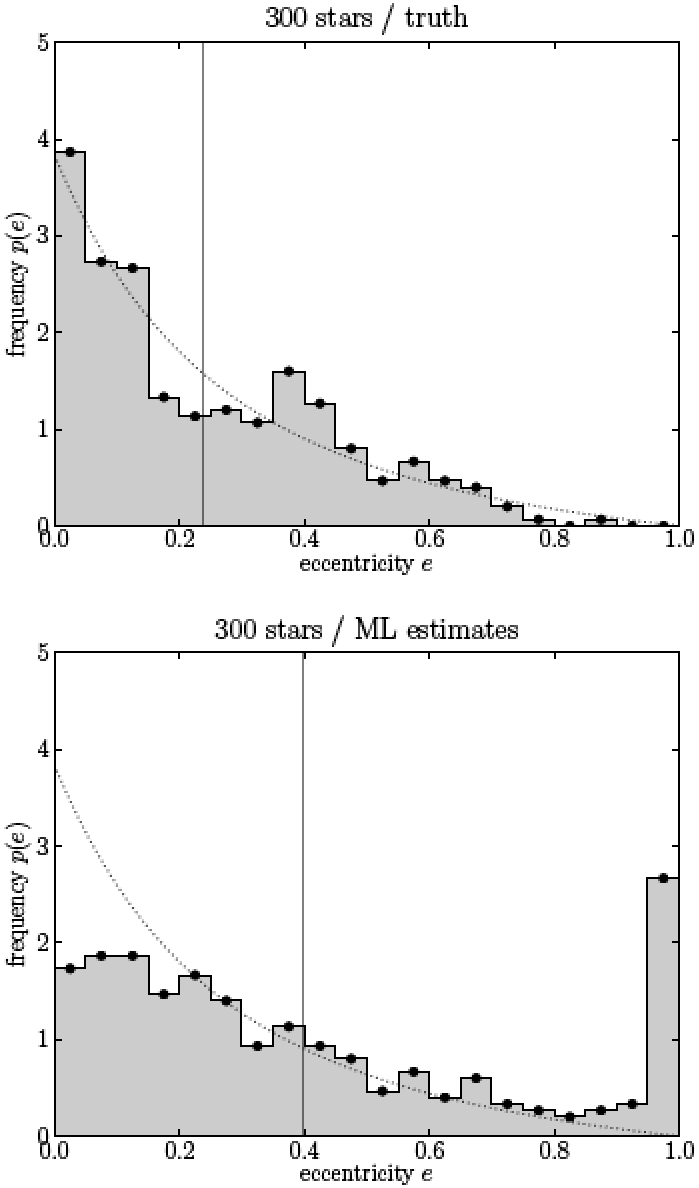
Alternativa 2:

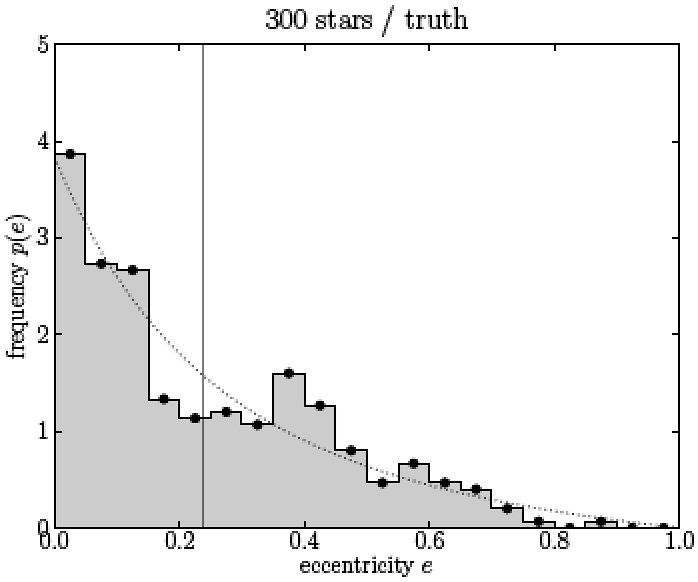
Asumir que la distribución observada es el resultado de convolucionar la distribución real con las incertidumbres.

Implica modelar **de forma general** la relación entre los parámetros reales y los observados.

No utiliza las incertidumbres reales de las observaciones.

Es poco robusto.

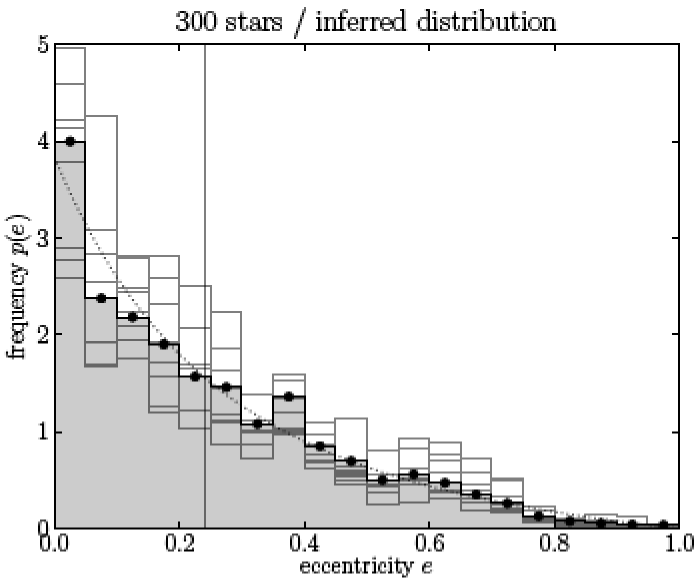




Alternativa 3: Proponer un modelo parametrizado por un vector **α** e inferir la distribución a posteriori de **α**

$$p(\boldsymbol{\alpha}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\alpha}) \cdot p(\boldsymbol{\alpha})}{p(\boldsymbol{D})}$$

$$p(\boldsymbol{D}|\boldsymbol{\alpha}) = \int p(\boldsymbol{D}|\boldsymbol{\omega}) \cdot p(\boldsymbol{\omega}|\boldsymbol{\alpha}) \, \boldsymbol{d}\,\boldsymbol{\omega}$$



Recordad, **D** siguen siendo los datos (por ejemplo, un espectro o la curva de velocidad radial)

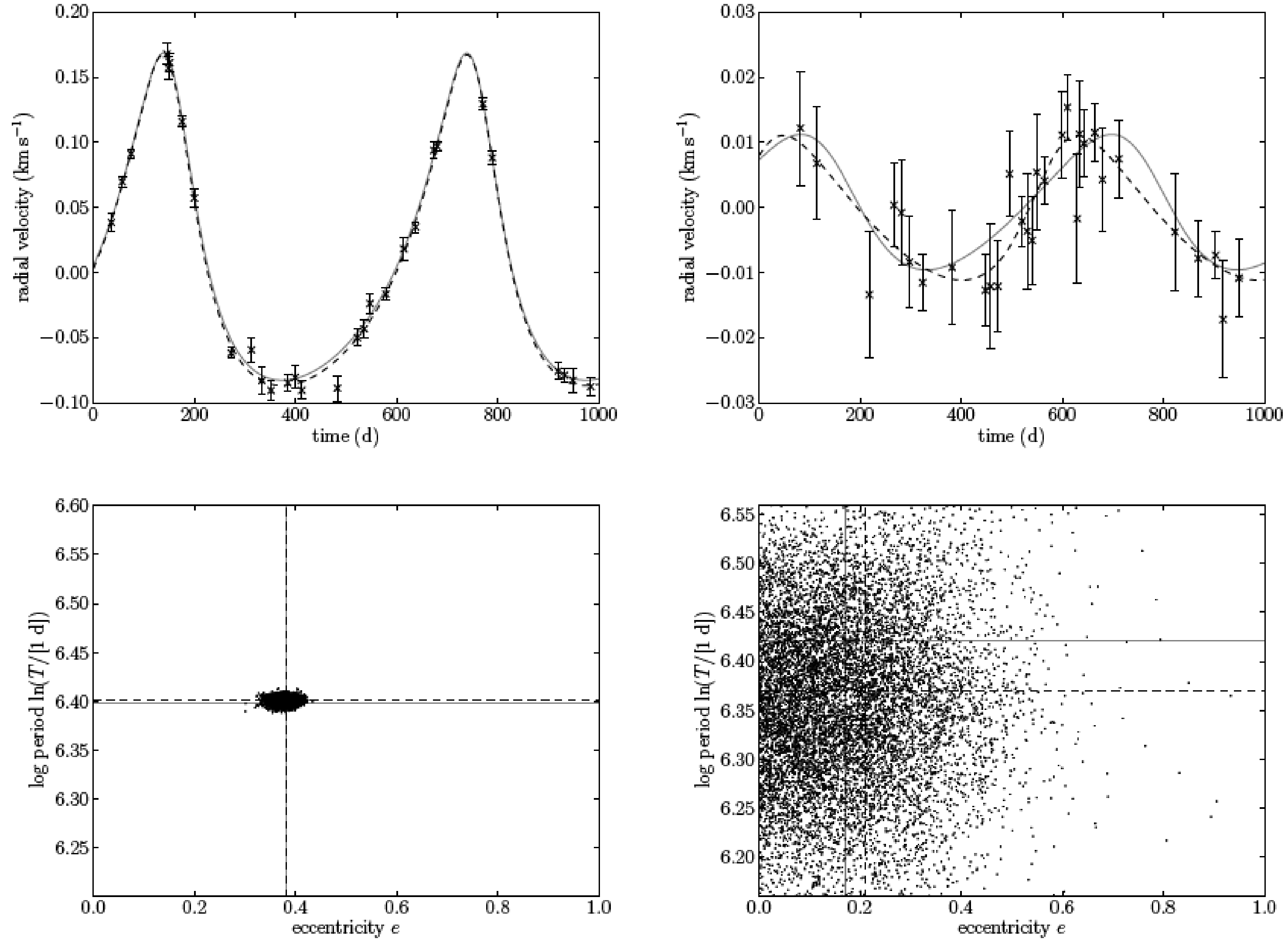
Modelos jerárquicos

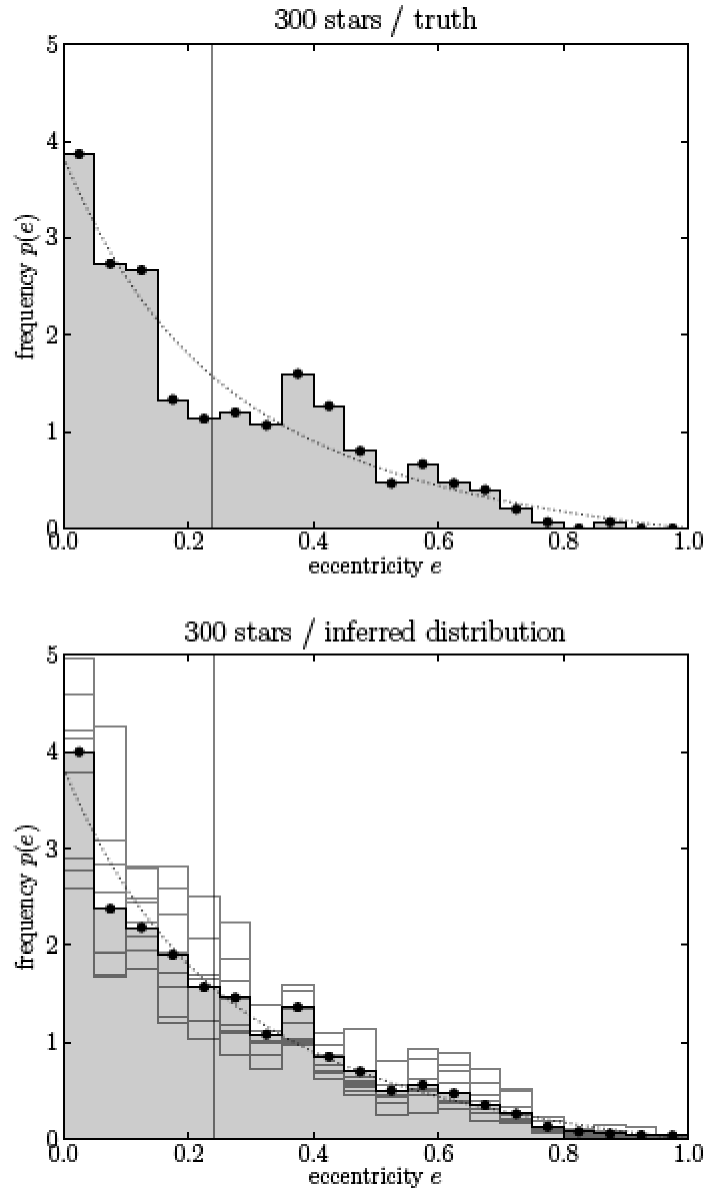
$$p(\boldsymbol{\omega}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\omega}) \cdot p(\boldsymbol{\omega})}{p(\boldsymbol{D})}$$

En espacios de muy alta dimensionalidad, hay que evaluar la verosimilitud en un número muy elevado de puntos.

Ejemplo: en astrosismología, hay que elaborar un modelo estelar para cada conjunto de parámetros (T_{eff}, t, [Fe/H],...). Si tenemos 5 dimensiones y 10 puntos por eje hay que evaluar la verosimilitud en 10^5 puntos.
Soluciones: Sampling y forward modelling.

Sampling:


radial velocity (km s^{-1})
time (d)
radial velocity (km s^{-1})
time (d)
log period ln(T/[1 d])
eccentricity e
log period ln(T/[1 d])
eccentricity e



Alternativa 3: Proponer un modelo parametrizado por un vector **α** e inferir la distribución a posteriori de **α**

$$p(\boldsymbol{\alpha}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\alpha}) \cdot p(\boldsymbol{\alpha})}{p(\boldsymbol{D})}$$

$$p(\boldsymbol{D}|\boldsymbol{\alpha}) = \int p(\boldsymbol{D}|\boldsymbol{\omega}) \cdot p(\boldsymbol{\omega}|\boldsymbol{\alpha})\, \boldsymbol{d}\,\boldsymbol{\omega}$$

Recordad, **D** siguen siendo los datos (por ejemplo, un espectro o la curva de velocidad radial)

Ventajas:

- Correcta propagación de las incertidumbres
- Posibilidad de hacer selección de modelos
- Selección de la complejidad óptima del modelo

¡Cuidado! El conjunto de parámetros e, T, ω, φ, κ, I tiene una complejidad contínua.

Posibles aplicaciones en Gaia:

1. Determinación de la distancia a las Nubes de Magallanes
2. Determinación de las distribuciones correspondientes a diferentes componentes de la Galaxia
3. IMFs, SFRs...