

# Modelos Jerárquicos

*Limitaciones:*

-Si tiramos 25 veces la moneda y sale 25 veces cara.. raro

...Pero con el prior de 100 y 100 que usamos obtenemos:

$$p(0|25, 25) = 125/225 \simeq 0.56 \text{ ¡no tan raro!}$$

# Modelo

*likelihood:*

$$p(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$k \sim \text{Binomial}(\theta, n)$$

*prior:*

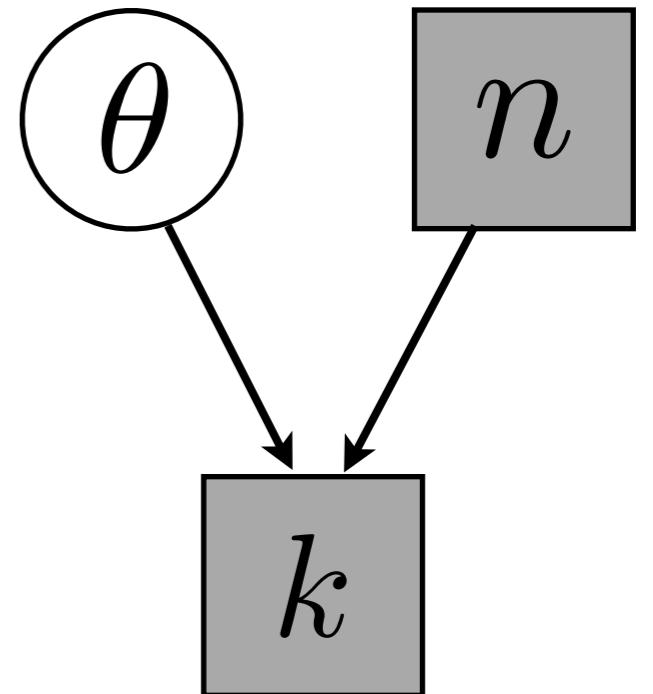
$$\theta \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1)$$

$$\theta \sim \text{Beta}(100, 100)$$

*posterior:*

$$p(\theta|D) = \text{Beta}(k+1, n-k+1)$$

$$p(\theta|D) = \text{Beta}(k+100, n-k+100)$$

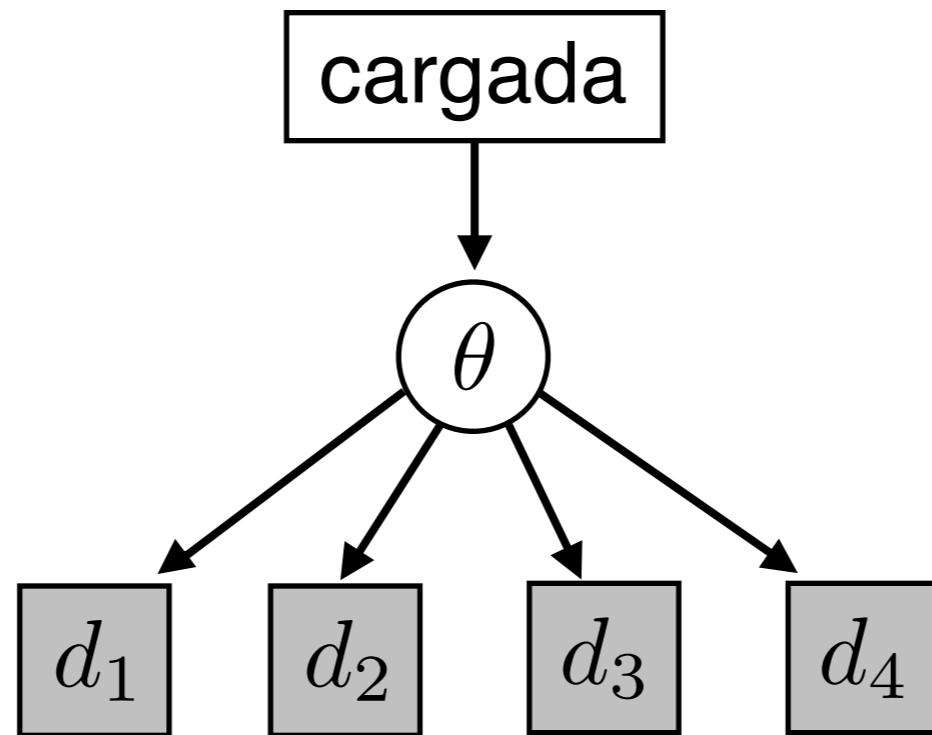


# Modelos Jerárquicos

*prior* débil:  
poco razonable

Beta(1, 1)

*prior* fuerte:  
¡poco flexible!  
Beta(100, 100)



Hipótesis de más alto nivel: *overhypothesis*, o teoría

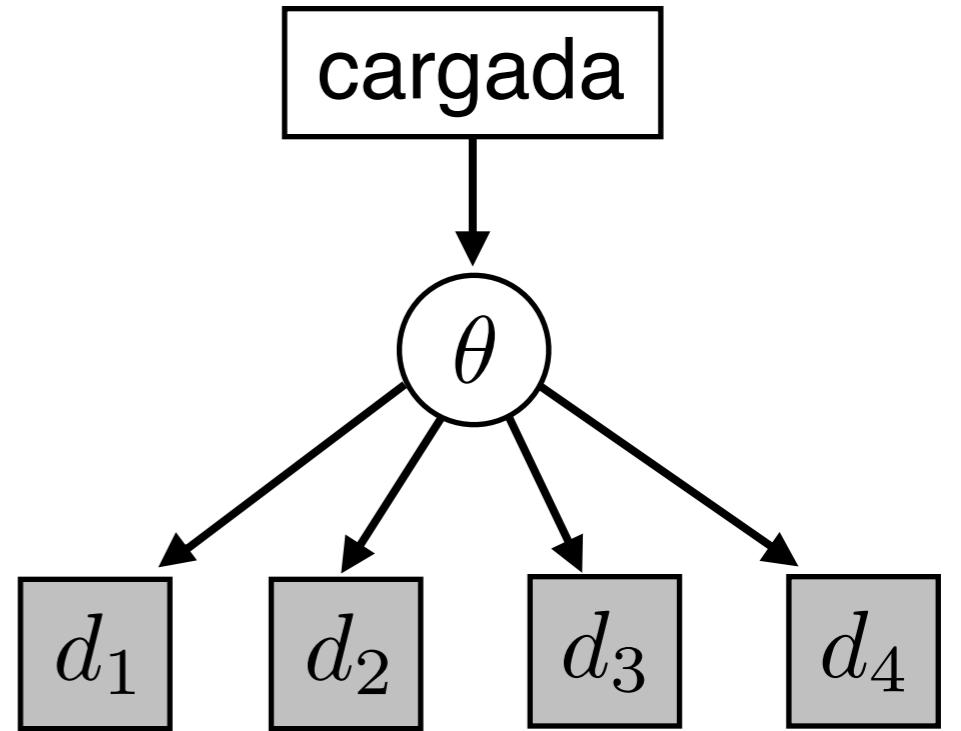
# Modelo jerárquico

$$k \sim \text{Binomial}(\theta, n)$$

$$\underline{p(\theta|\text{no cargada}) = \text{Beta}(100, 100)}$$

$$\underline{p(\theta|\text{cargada}) = \text{Beta}(1, 1)}$$

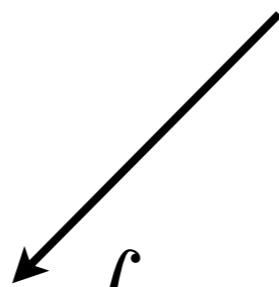
$$p(\text{cargada}) = 0.01$$



La información de los datos se “propaga” hacia arriba

Hacemos inferencia sobre la teoría...  $p(\text{cargada}|25 \text{ caras})$

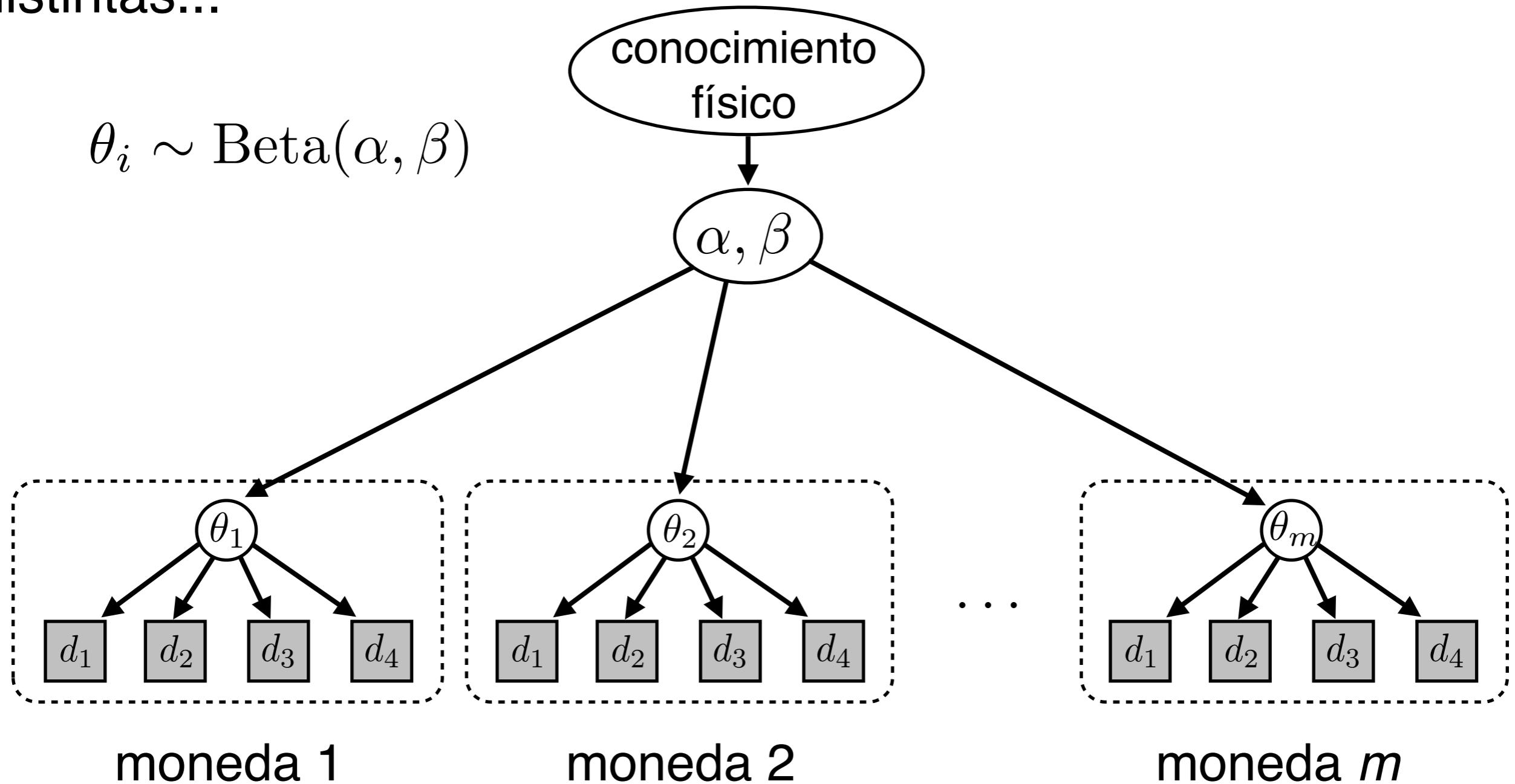
$$\frac{p(\text{cargada}|25 \text{ caras})}{p(\text{no cargada}|25 \text{ caras})} = \frac{p(25 \text{ caras}|\text{cargada})}{p(25 \text{ caras}|\text{no cargada})} \frac{p(\text{cargada})}{p(\text{no cargada})}$$



$$p(25 \text{ caras}|\text{cargada}) = \int p(25 \text{ caras}|\theta)p(\theta|\text{cargada})d\theta$$

$$\frac{p(\text{cargada}|25 \text{ caras})}{p(\text{no cargada}|25 \text{ caras})} \simeq \frac{0.038}{1.14 \times 10^{-7}} \frac{0.01}{0.99} \simeq 3367$$

...No es lo mismo ver 200 tiradas de una moneda que 20 de 10 distintas...



Inferencia sobre los *hiperparámetros*..  
caracterizan la población de monedas

# Intercambiabilidad (Exchangeability)

$$p(\theta_1, \theta_2, \dots, \theta_m)$$

invariante ante permutaciones de los índices  $1, 2, \dots, m$

útil en general, puede ser para *parámetros* o *datos*  
ignorancia implica intercambiabilidad

$$p(\theta|\phi) = \prod_{j=1}^m p(\theta_j|\phi) \quad \text{parámetros } iid$$

$$p(\theta) = \int \left( \prod_{j=1}^m p(\theta_j|\phi) \right) p(\phi) d\phi \quad \text{mezcla de distribuciones } iid$$

# Inversa: teorema de representación de *de Finetti*

En el límite del número de variables yendo a infinito, toda distribución intercambiable puede escribirse como mezcla de *iid*'s

Para número finito: Gaussiana bivariada

$$\mu_x = \mu_y = 0 \quad \sigma_x = \sigma_y = 1 \quad \text{correlación} = \rho$$

$$p(x, y) \propto e^{-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)}$$

¡Intercambiables pero no mezcla de independientes!

# Modelos jerárquicos

$\phi \rightarrow$  hiperparámetros

$\theta \rightarrow$  parámetros

$y \rightarrow$  datos

$p(\phi, \theta) \rightarrow$  prior conjunto

$p(\phi, \theta|y) \rightarrow$  posterior conjunto

$$p(\phi, \theta|y) \propto p(y|\phi, \theta)p(\phi, \theta) = p(y|\theta)p(\phi, \theta)$$

Nos pueden interesar tanto los parámetros como los hiperparámetros para la *posterior*, y tanto los parámetros como los datos para la *predicción*

# Procedimiento numérico/analítico para modelos conjugados

$p(\phi|\theta)$  conjugada a  $p(y|\theta)$

1) Descomponer la *posterior* conjunta

$$p(\phi, \theta|y) \propto p(y|\theta)p(\theta|\phi)p(\phi)$$

2) Determinar la *posterior* sobre los parámetros, dados  
los hiperparámetros  $\mathbb{P}$

$$p(\theta|y, \phi) \propto \prod_i p(\theta_i|y, \phi)$$

3) Computar la *posterior* sobre los hiperparámetros

$$p(\phi|y) = \int p(\theta, \phi|y)d\theta$$

fuerza bruta

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}$$

cálculo analítico  
(¡normalización!)

Obteniendo muestras de la *posterior* conjunta  $p(\phi, \theta|y)$

1) Obtener muestras de los hiperparámetros

$$p(\phi|y)$$

2) Obtener muestras de los parámetros, independientemente para cada componente, usando las muestras anteriores

$$p(\theta|y, \phi) \propto \prod_i p(\theta_i|y, \phi)$$

3) Obtener muestras *predictivas* usando las muestras obtenidas

Limitado a modelos muy sencillos..  
En general: algoritmos de muestreo

# Midiendo la predictibilidad de palabras

*“Voy a sacudir el mantel porque está lleno de \_\_\_\_\_”*

# Midiendo la predictibilidad de palabras

2

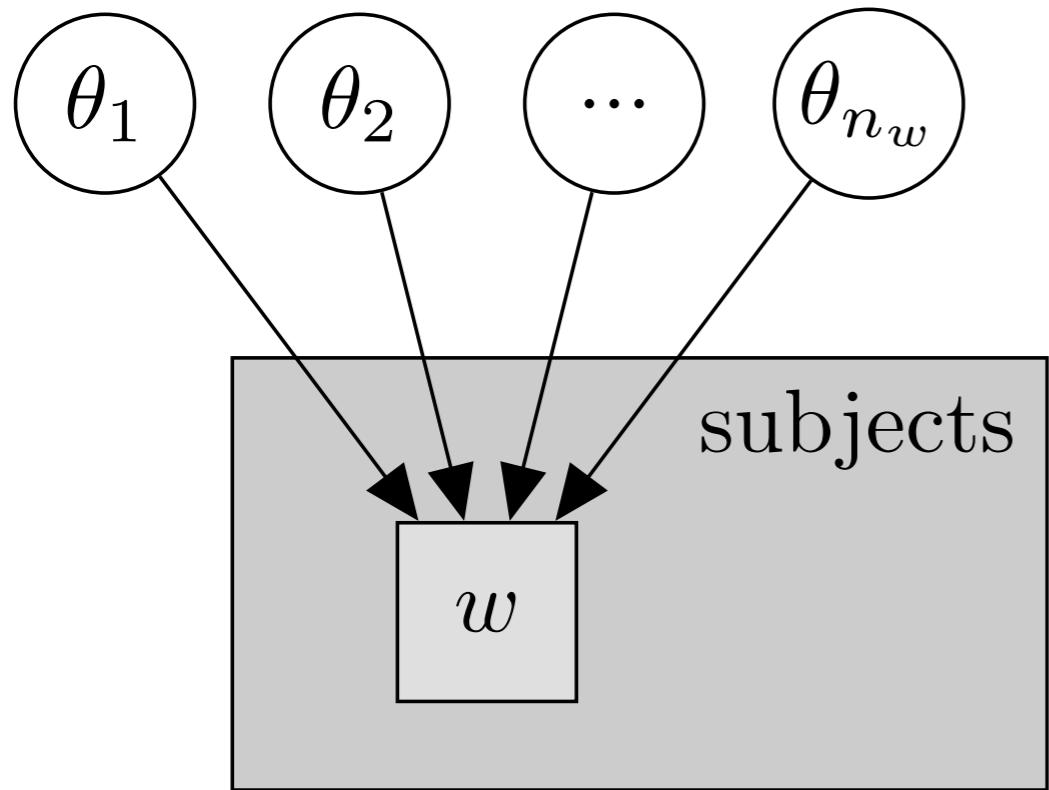
Indagando más allá:

*“Voy a sacudir el mantel porque está lleno de \_\_\_\_\_”*

\_\_\_\_\_

\_\_\_\_\_

# Modelo de una palabra

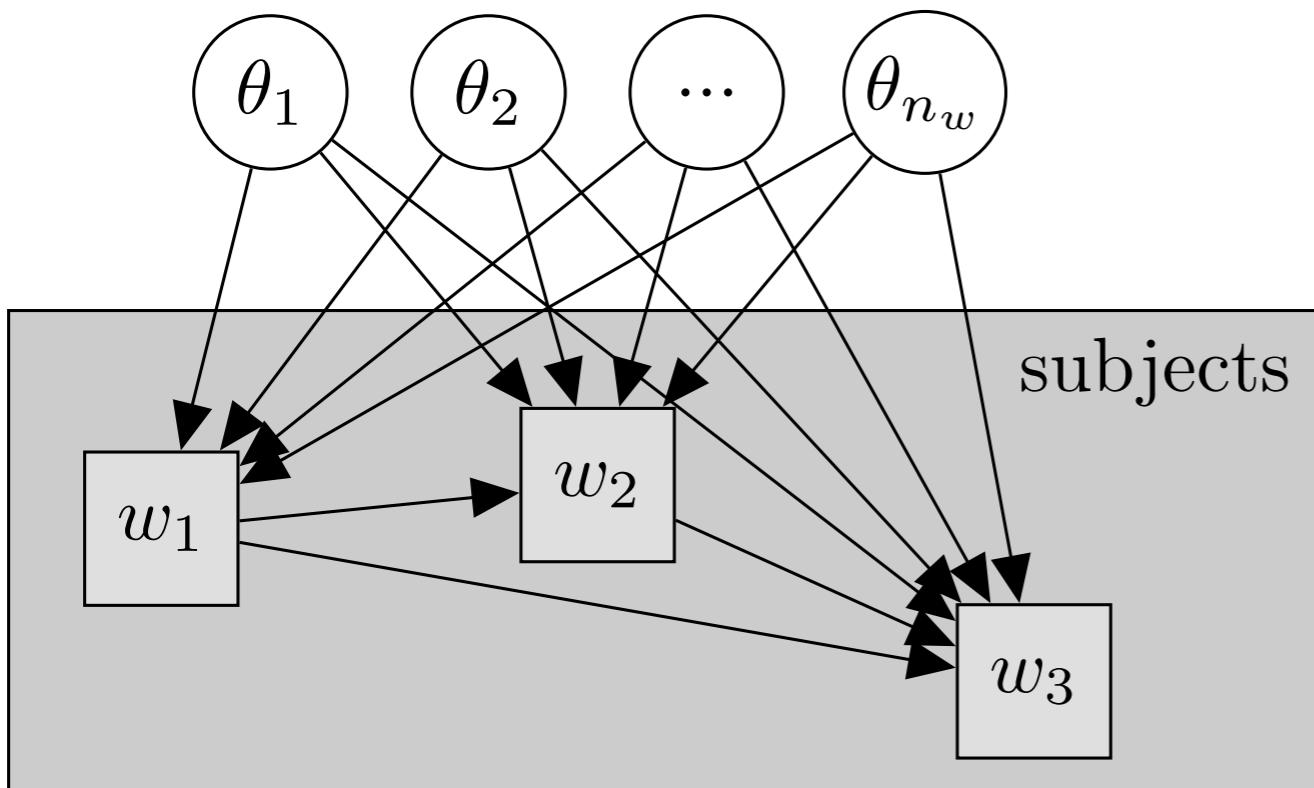


$$\theta \sim \text{Dirichlet}(1, 1, \dots, 1)$$

$$w \sim \text{Categorical}(\theta_1, \theta_2, \dots, \theta_{n_{words}})$$

- Noción de incertezza
- Predictibilidad del contexto ~ predictibilidad de palabra
- No hace uso de palabras extra

# Modelo de muestreo secuencial



$$\theta \sim \text{Dirichlet}(1, 1, \dots, 1)$$

$$w_1 \sim \text{Categorical}(\theta_1, \theta_2, \dots, \theta_{n_{words}})$$

$$w_2 \sim \text{Categorical}(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{n_{words}})$$

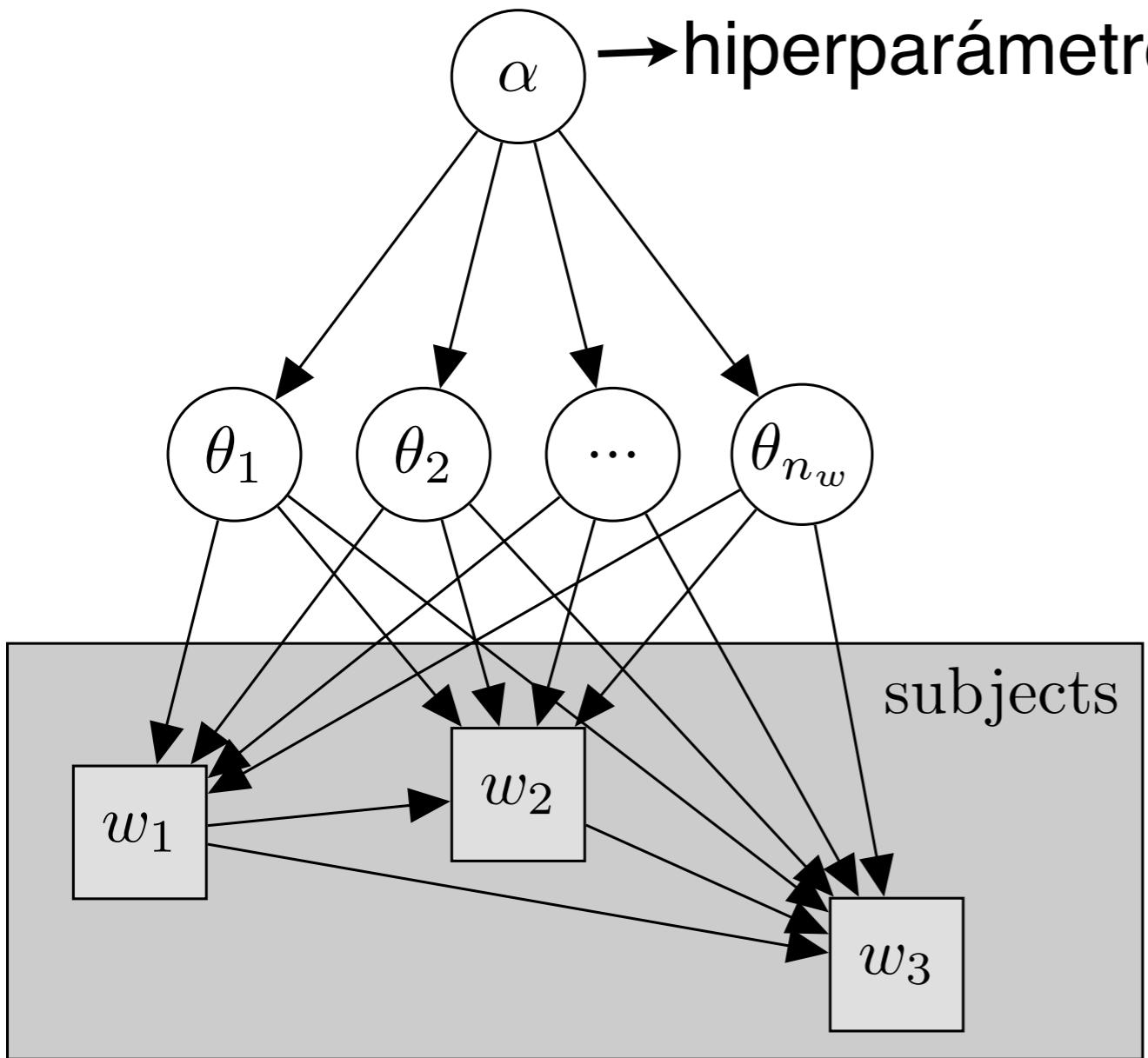
$$w_3 \sim \text{Categorical}(\tilde{\tilde{\theta}}_1, \tilde{\tilde{\theta}}_2, \dots, \tilde{\tilde{\theta}}_{n_{words}})$$

$$\tilde{\theta}_i = \begin{cases} \frac{\theta_i}{1-\theta_{w_1}} & \text{if } i \neq w_1 \\ 0 & \text{if } i = w_1 \end{cases}$$

$$\tilde{\tilde{\theta}}_i = \begin{cases} \frac{\tilde{\theta}_i}{1-\tilde{\theta}_{w_2}} & \text{if } i \neq w_2 \\ 0 & \text{if } i = w_2 \end{cases}$$

- Noción de incertezza
  - Predictibilidad del contexto ~ predictibilidad de palabra
  - Usa palabras extra
- Modelos jerárquicos

# Modelo Bayesiano Jerárquico



→ hiperparámetro de predictibilidad del *contexto*

$$\alpha \sim \text{Uniform}(0.05, 1)$$

$$\theta \sim \text{Dirichlet}(\alpha, \alpha, \dots, \alpha)$$

$$w_1 \sim \text{Categorical}(\theta_1, \theta_2, \dots, \theta_{n_{words}})$$

$$w_2 \sim \text{Categorical}(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{n_{words}})$$

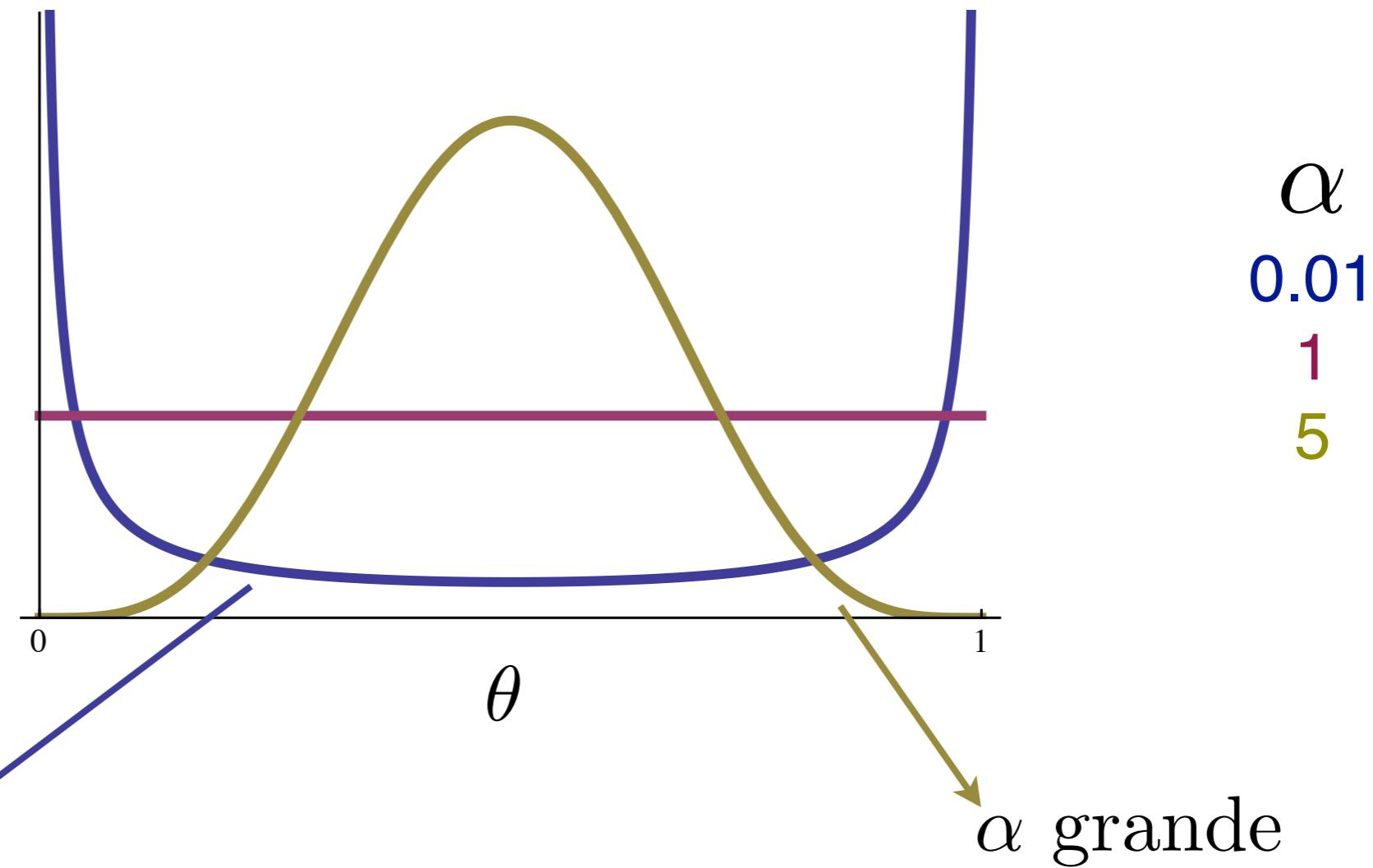
$$w_3 \sim \text{Categorical}(\tilde{\tilde{\theta}}_1, \tilde{\tilde{\theta}}_2, \dots, \tilde{\tilde{\theta}}_{n_{words}})$$

$$\tilde{\theta}_i = \begin{cases} \frac{\theta_i}{1-\theta_{w_1}} & \text{if } i \neq w_1 \\ 0 & \text{if } i = w_1 \end{cases}$$

$$\tilde{\tilde{\theta}}_i = \begin{cases} \frac{\tilde{\theta}_i}{1-\tilde{\theta}_{w_2}} & \text{if } i \neq w_2 \\ 0 & \text{if } i = w_2 \end{cases}$$

- Noción de incerteza
- Predictibilidad de contexto *independiente*
- Usa palabras extra

# Predictibilidad del contexto



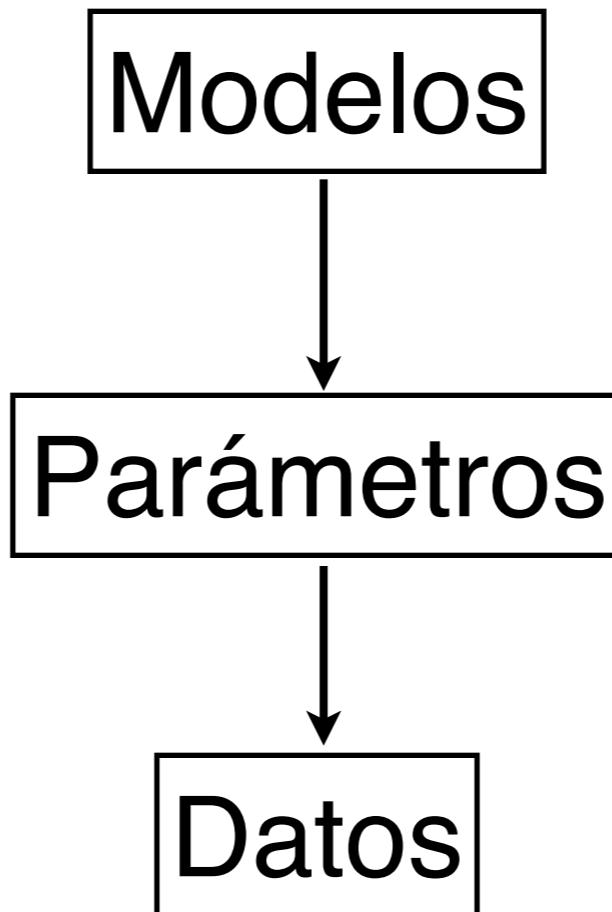
no sé qué palabra, pero alguna  
muy probable, las otras muy poco

*migas migas migas migas  
hormigas aceitunas árboles*

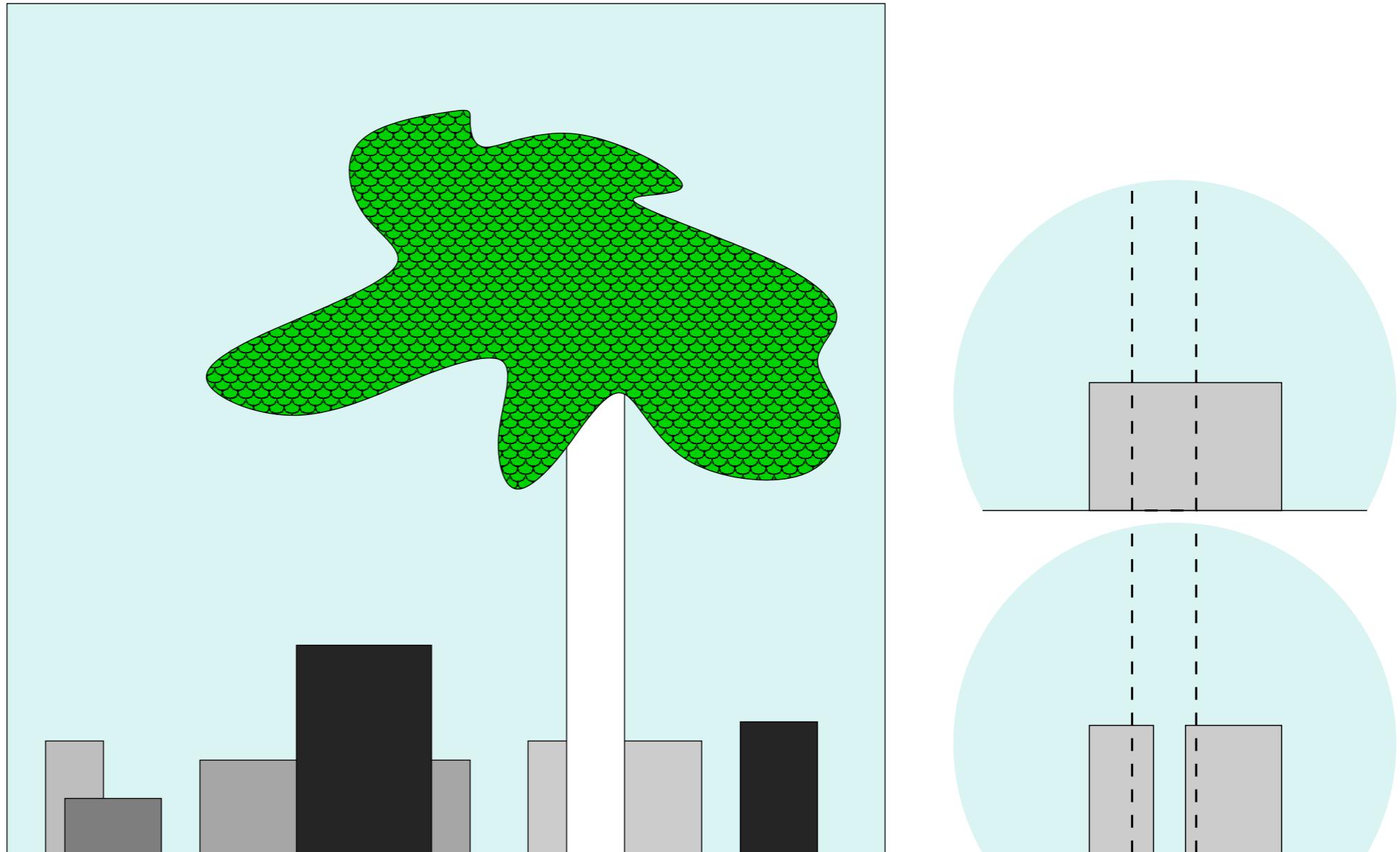
<

*migas migas migas migas  
hormigas hormigas hormigas*

# Comparación de Modelos y la “Navaja de Occam”



En principio, un problema jerárquico



¿Cuántas cajas hay atrás del árbol?

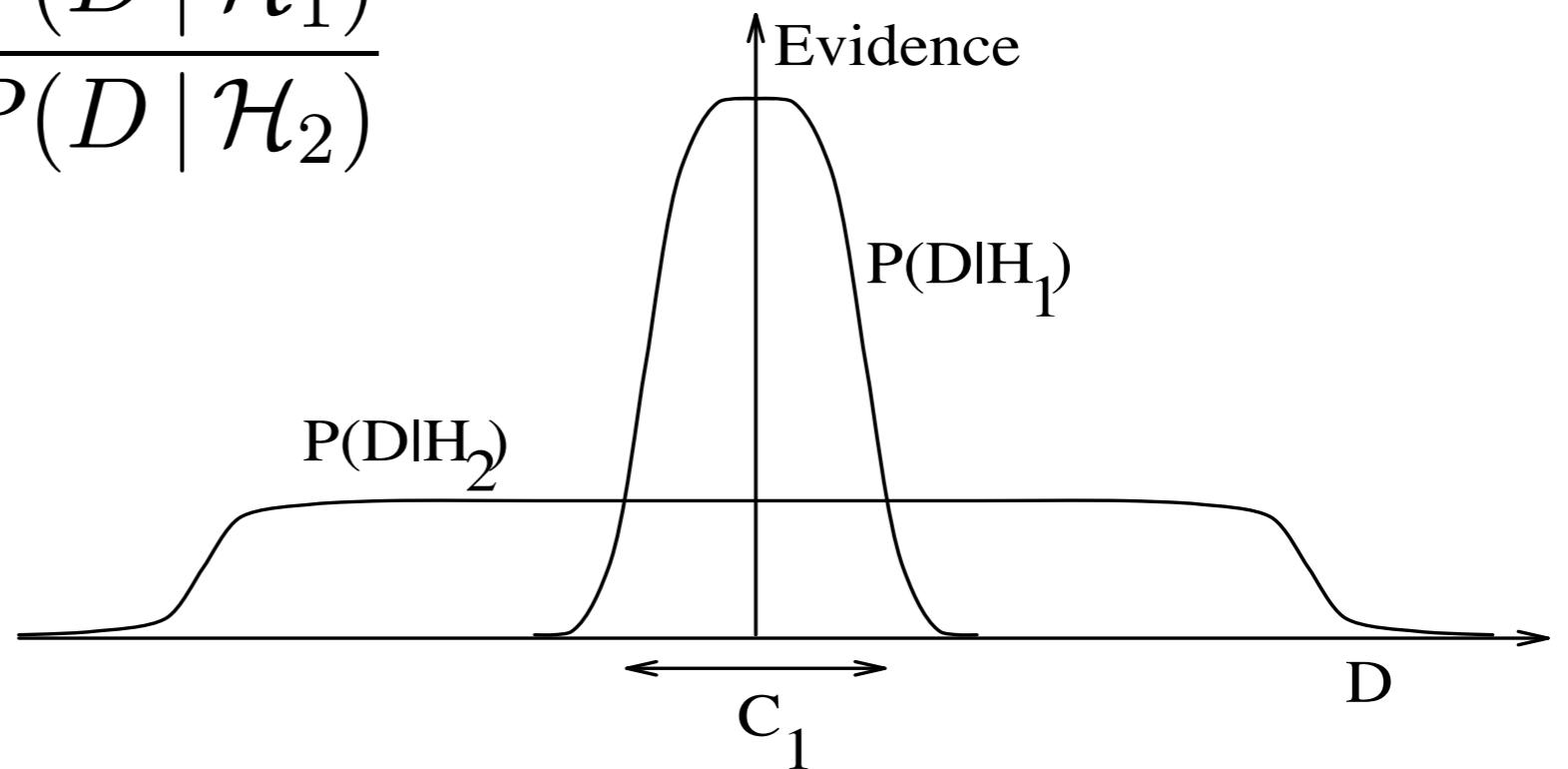
# Occam: prefiramos la explicación más simple

Dirac: porque es más *bella*  
o: ¡porque esta estrategia viene funcionando bien!

## Inferencia Bayesiana: lo dice la cuenta

$$\frac{P(\mathcal{H}_1 | D)}{P(\mathcal{H}_2 | D)} = \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)} \frac{P(D | \mathcal{H}_1)}{P(D | \mathcal{H}_2)}$$

Caricatura de modelos sencillos



-1, 3, 7, 11, ...

¿Cuáles son los próximos dos números de la secuencia?

15, 19

-19.9, 1043.8

sumar 4 al anterior

evaluar sobre el anterior  $x$ :

$$-x^3/11 + 9/11x^2 + 23/11$$

Formalizando:

$\mathcal{H}_a$  progresión aritmética, sumar  $n$

$\mathcal{H}_c$  función cúbica a partir del anterior  $x \rightarrow cx^3 + dx^2 + e$   
(con  $c, d, e$ : fracciones)

¡¡Tomemos *priors* iguales!!

# Calculamos la *Evidencia* para cada modelo

$\mathcal{H}_a$  progresión aritmética, sumar  $n$

$\mathcal{H}_c$  función cúbica a partir del anterior  $x \rightarrow cx^3 + dx^2 + e$   
(con  $c, d, e$ : fracciones)

(Tomamos intervalos -50 a 50)

$$P(D | \mathcal{H}_a) = \frac{1}{101} \frac{1}{101} = 0.00010$$

dónde empiezo y  
cuánto salto

$$\begin{aligned} P(D | \mathcal{H}_c) &= \left(\frac{1}{101}\right) \left(\frac{c}{101} \frac{d}{50}\right) \left(\frac{4}{101} \frac{1}{50}\right) \left(\frac{2}{101} \frac{1}{50}\right) \\ &= 0.000000000025 = 2.5 \times 10^{-12}. \end{aligned}$$

Los *odds* son de 40 millones a 1..

Lo mismo pasa en la ciencia: Copérnico vs. *epiciclos*

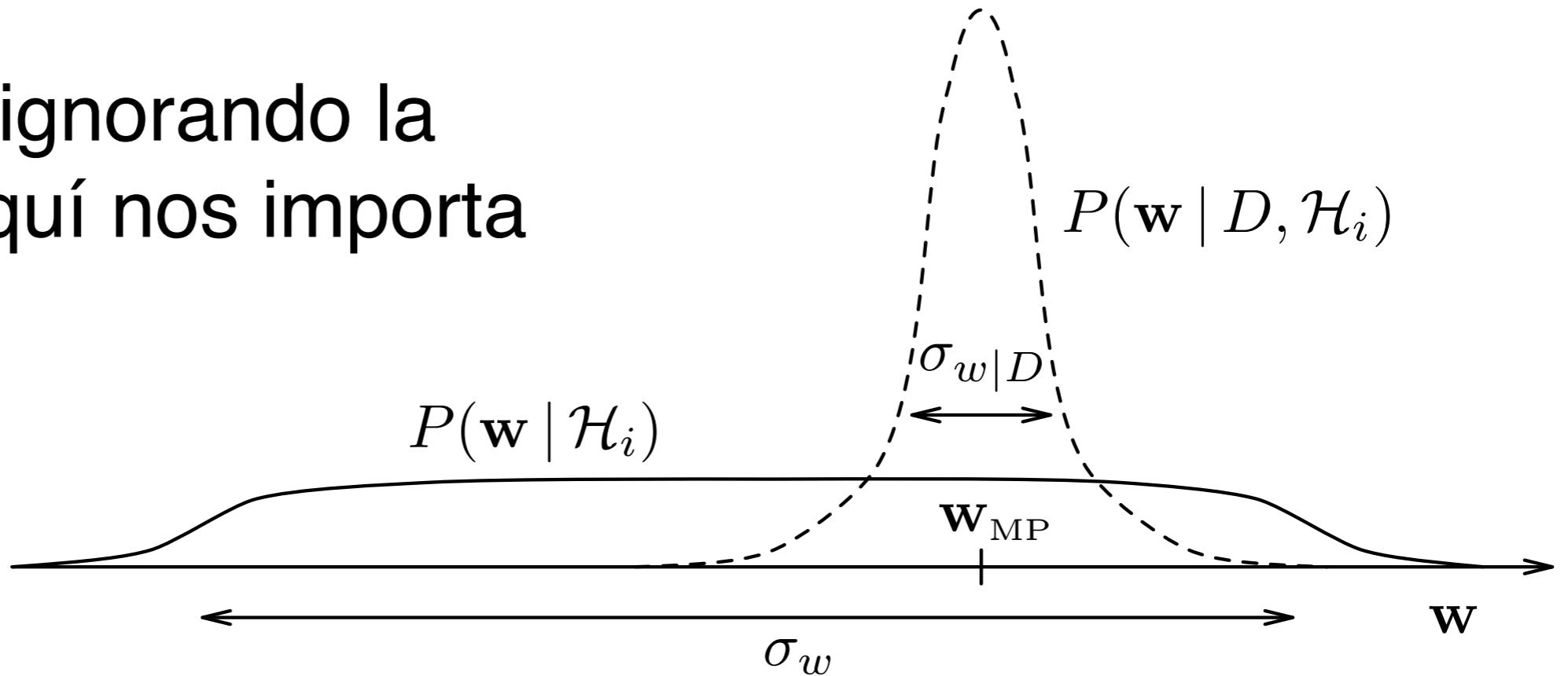
# Inferencia en dos niveles

## 1) Ajuste de Modelos

## 2) Comparación de Modelos

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Veníamos ignorando la evidencia.. aquí nos importa



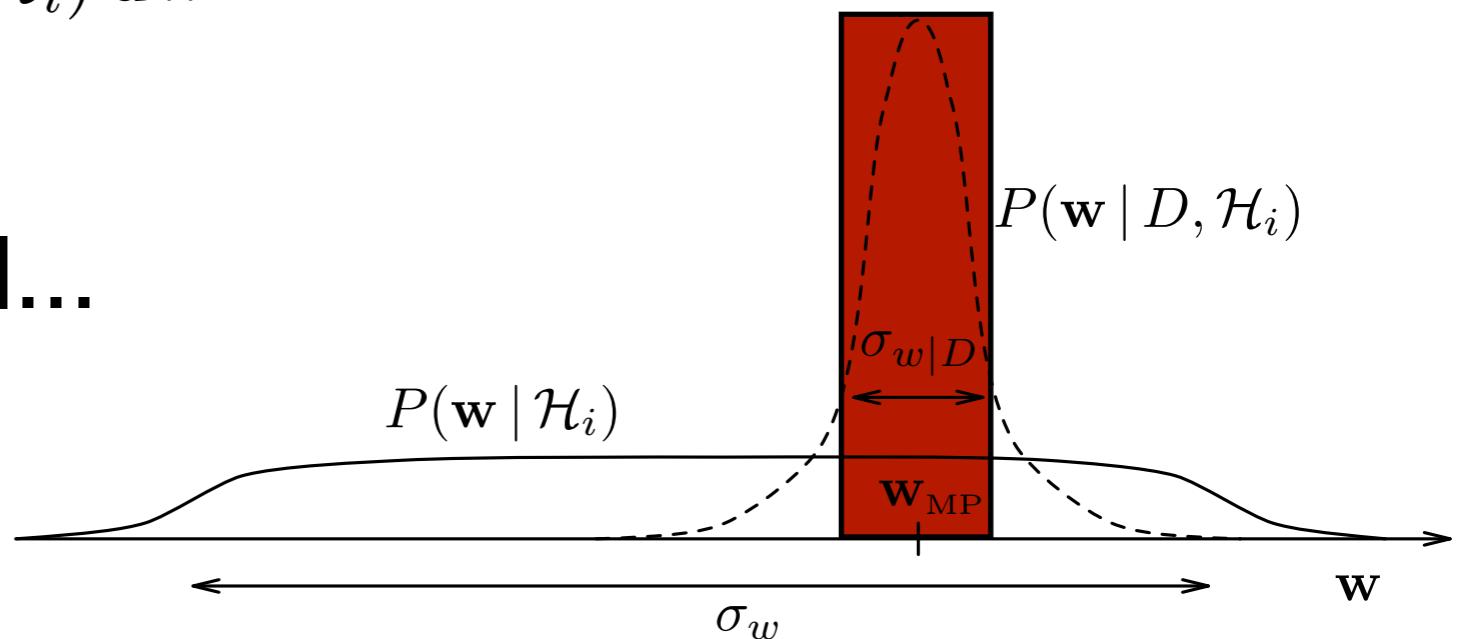
## 2) Comparación de Modelos

$$P(\mathcal{H}_i | D) \propto P(D | \mathcal{H}_i) P(\mathcal{H}_i)$$

Evidencia  
(era la normalización)

$$P(D | \mathcal{H}_i) = \int P(D | \mathbf{w}, \mathcal{H}_i) P(\mathbf{w} | \mathcal{H}_i) d\mathbf{w}$$

Aproximamos la integral...



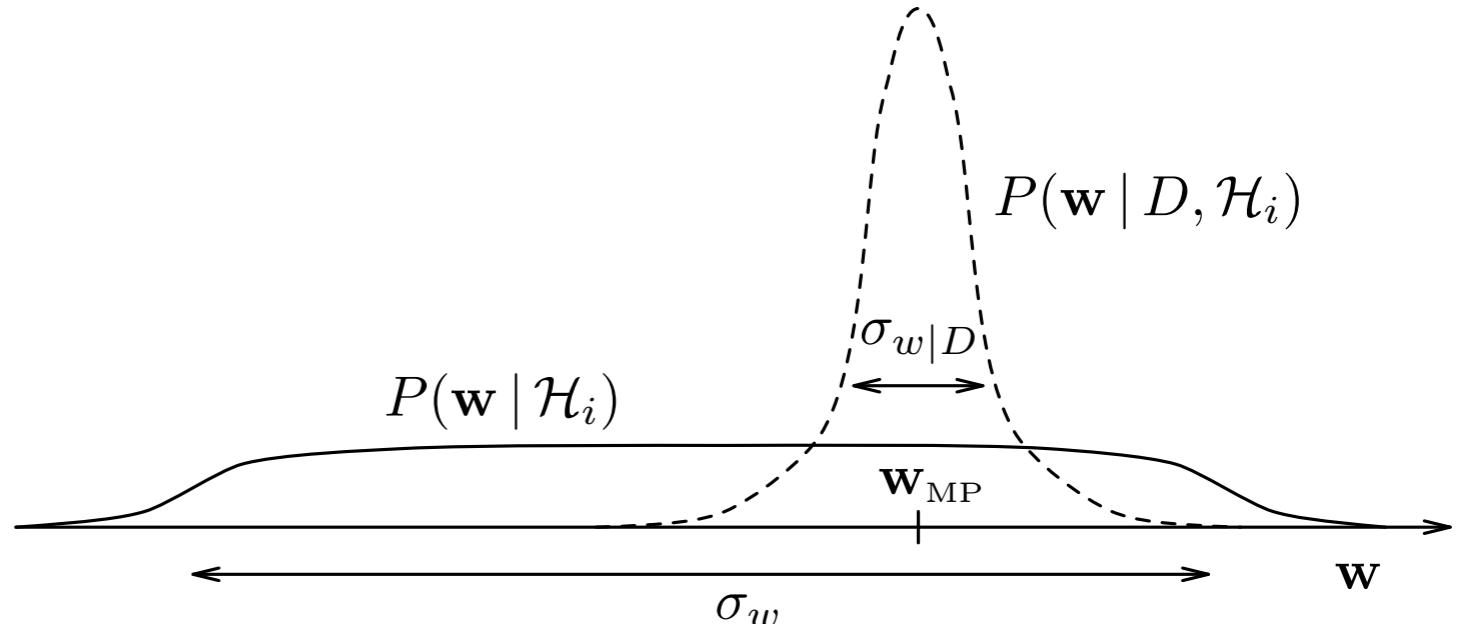
$$P(D | \mathcal{H}_i) \simeq \underbrace{P(D | \mathbf{w}_{\text{MP}}, \mathcal{H}_i)}_{\text{Best fit likelihood}} \times \underbrace{P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) \sigma_w|_D}_{\text{Occam factor}}$$

Evidence  $\simeq$  Best fit likelihood  $\times$  Occam factor

# El factor de Occam

*prior plana:*

$$P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) = 1/\sigma_w$$



$$P(D | \mathcal{H}_i) \simeq \underbrace{P(D | \mathbf{w}_{\text{MP}}, \mathcal{H}_i)}_{\text{Best fit likelihood}} \times \underbrace{P(\mathbf{w}_{\text{MP}} | \mathcal{H}_i) \sigma_{w|D}}_{\text{Occam factor}}$$

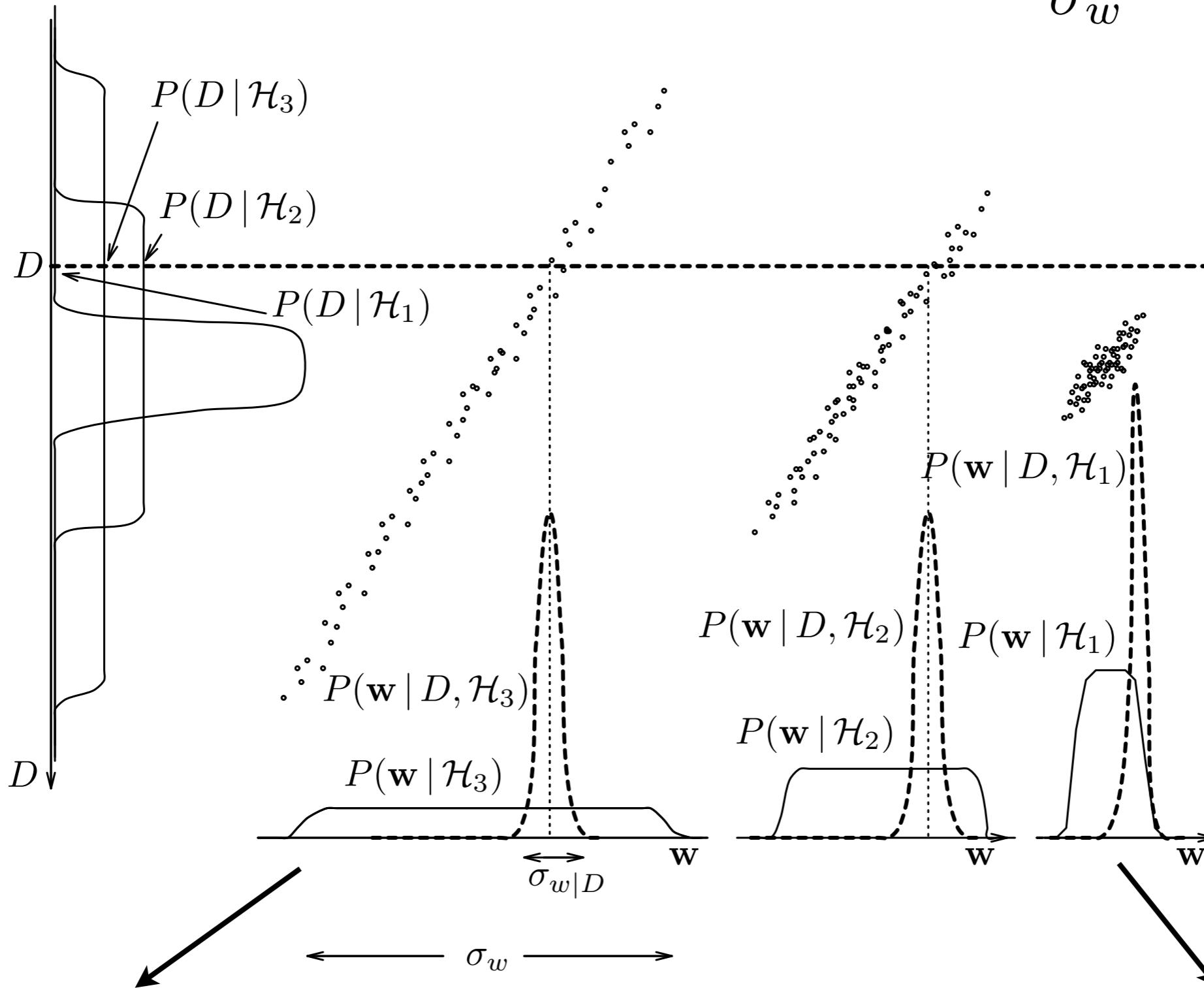
Evidence  $\simeq$  Best fit likelihood  $\times$  Occam factor

$$\text{Factor de Occam} = \frac{\sigma_{w|D}}{\sigma_w}$$

- penaliza modelos con rango grande de parámetros
- penaliza modelos cuyos parámetros requieren un ajuste fino a los datos

...¡y todo esto elevado al número de parámetros!

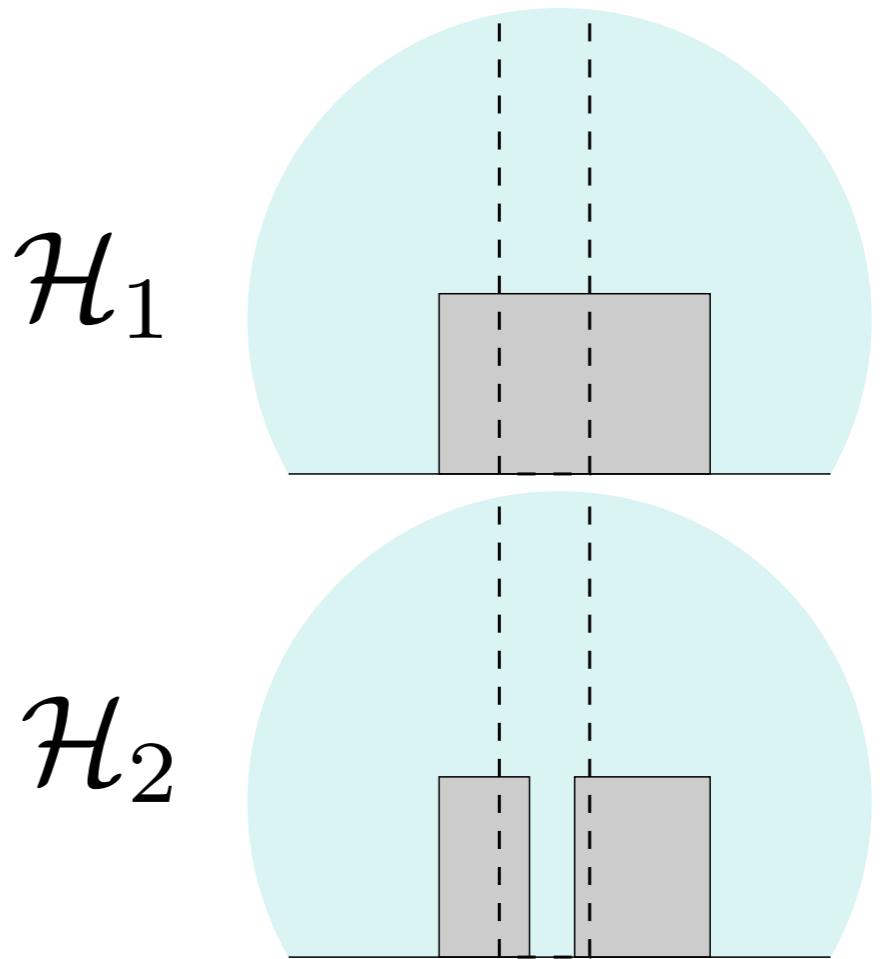
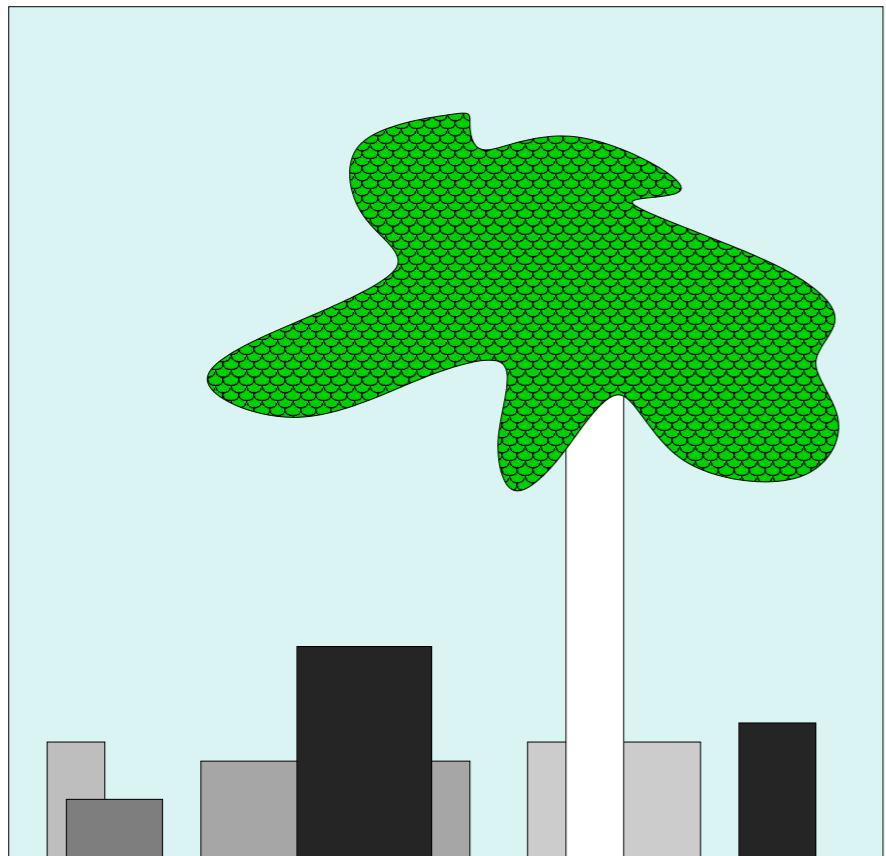
$$P(D|\mathcal{H}_i) \simeq P(D|\mathbf{w}_{MP}, \mathcal{H}_i) \times \frac{\sigma_w|D}{\sigma_w}$$



Buen *likelihood*,  
mal *Occam*

Buen *Occam*,  
mal *likelihood*

# De vuelta al árbol...



$$P(D | \mathcal{H}_1) = \frac{1}{20} \frac{1}{20} \frac{1}{20} \frac{1}{16}$$

$$P(D | \mathcal{H}_2) \simeq \frac{1}{20} \frac{1}{20} \frac{10}{20} \frac{1}{16} \frac{1}{20} \frac{1}{20} \frac{10}{20} \frac{1}{16} \frac{2}{2}$$

(aproximando restricciones en los parámetros)

$$\frac{P(D | \mathcal{H}_1)P(\mathcal{H}_1)}{P(D | \mathcal{H}_2)P(\mathcal{H}_2)} = \frac{1}{\frac{1}{20} \frac{10}{20} \frac{10}{20} \frac{1}{16}} \simeq 1000/1$$

factores de Occam