Bayesian inference and Bayesian model selection

Klaas Enno Stephan

Translational Neuromodeling Unit

Universität Zürich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lecture as part of "Methods & Models for fMRI data analysis",
University of Zurich & ETH Zurich, 27 November 2018

With slides from and many thanks to:

Kay Brodersen,

Will Penny,

Sudhir Shankar Raman

Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

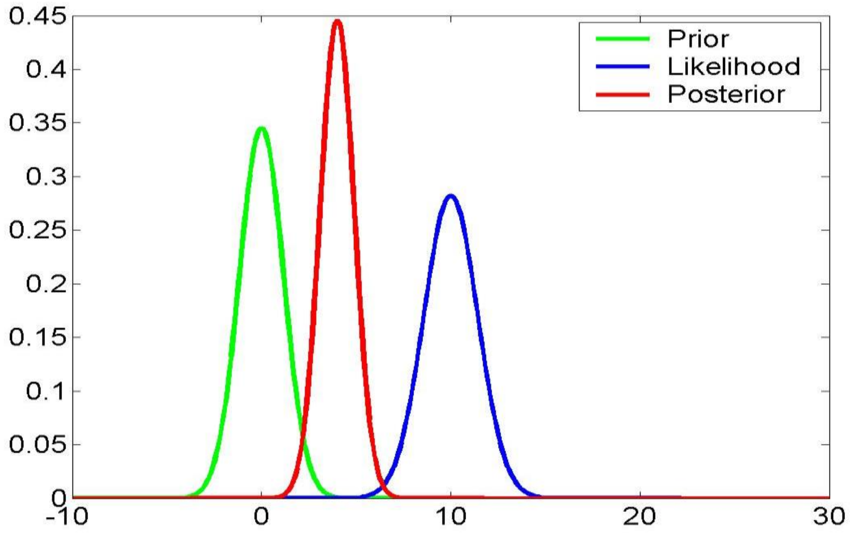
- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
  - computational psychosomatics
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
  - computational psychosomatics
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

Bayes' theorem



The Reverend Thomas Bayes
(1702-1761)

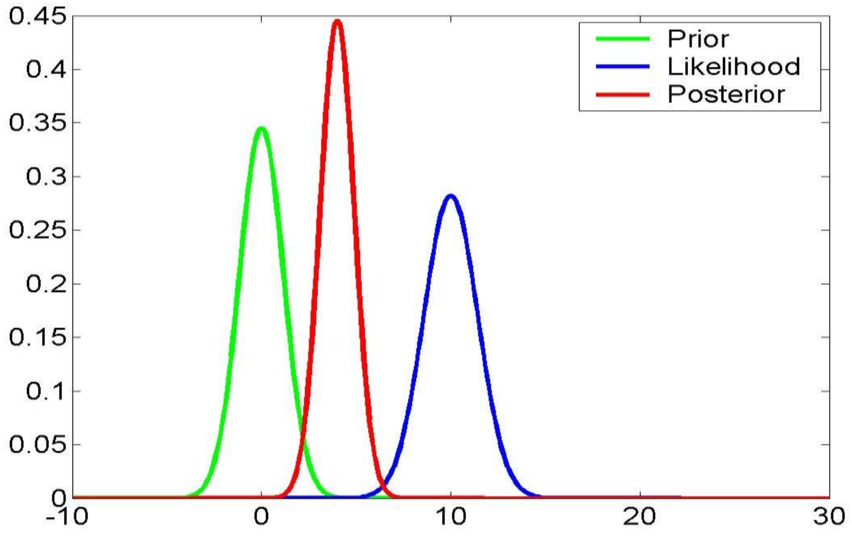
$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

posterior = likelihood • prior / evidence

“... the theorem expresses how a ... degree of belief should rationally change to account for availability of related evidence."

Wikipedia

Bayes' theorem



The Reverend Thomas Bayes
(1702-1761)

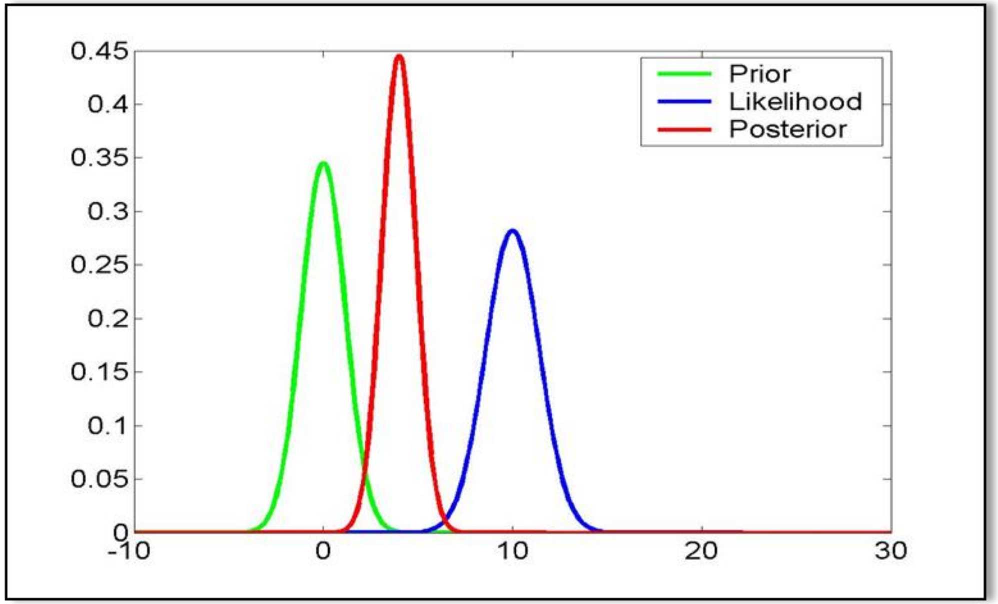
$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta)}$$

posterior = likelihood • prior / evidence

“... the theorem expresses how a ... degree of belief should rationally change to account for availability of related evidence."

Wikipedia

Bayesian inference: an animation



The evidence term

continuous θ

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta)}$$

discrete θ

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\sum_{\theta \in \Theta} p(y \mid \theta)\, p(\theta)}$$

Bayesian inference: An example (with fictitious probabilities)

- symptom:
  y=1: fever
  y=0: no fever

- disease:
  θ=1: Ebola
  θ=0: any other disease (AOD)

y \ θ	1: ebola	0: AOD
1: fever	99.99%	20%
0: no fever	0.01%	80%

p(y|θ)

- A priori:
  p(Ebola) $=10^{-6}$
  p(AOD) $= (1-10^{-6})$

- A patient presents with fever. What is the probability that he/she has ebola?

$$p(\theta=1 \mid y=1) = \frac{p(y=1 \mid \theta=1)\, p(\theta=1)}{\sum_{j\in\{0,1\}} p(y=1 \mid \theta=j)\, p(\theta=j)}$$

Bayesian inference: An example (with fictitious probabilities)

- symptom:
  y=1: fever
  y=0: no fever

- disease:
  θ=1: Ebola
  θ=0: any other disease (AOD)

y \ θ	1: ebola	0: AOD
1: fever	99.99%	20%
0: no fever	0.01%	80%

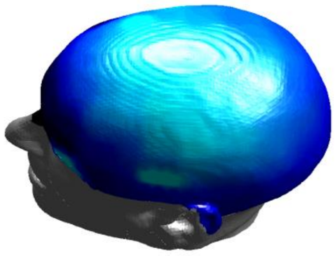
p(y|θ)

- A priori:
  $p(\text{Ebola}) = 10^{-6}$
  $p(\text{AOD}) = (1-10^{-6})$

- A patient presents with fever. What is the probability that he/she has ebola?

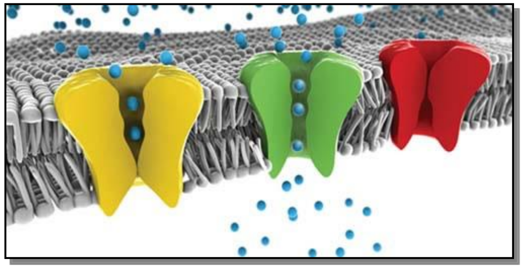
$$p(\theta = 1 \mid y = 1) = \frac{0.999 \cdot 10^{-6}}{0.999 \cdot 10^{-6} + 0.2 \cdot (1 - 10^{-6})} = 4.995 \cdot 10^{-6}$$

Generative models



$$p(y \mid \theta, m) \cdot p(\theta \mid m)$$

$$p(\theta \mid y, m)$$



1. specify the joint probability over data (observations) and parameters
2. enforce mechanistic thinking: how could the data have been caused?
3. generate synthetic data (observations) by sampling from the prior – can model explain certain phenomena at all?
4. inference about parameters → p(θ|y)
5. model evidence p(y|m): index of model quality

Bayesian inference in practice

⇨ Formulation of a **generative model**

Model ⇒ likelihood function $p(y|\theta)$

Model ⇒ prior distribution $p(\theta)$

⇨ Observation of **data**

Measurement ⇒ data y

⇨ **Model inversion** – updating one's beliefs

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

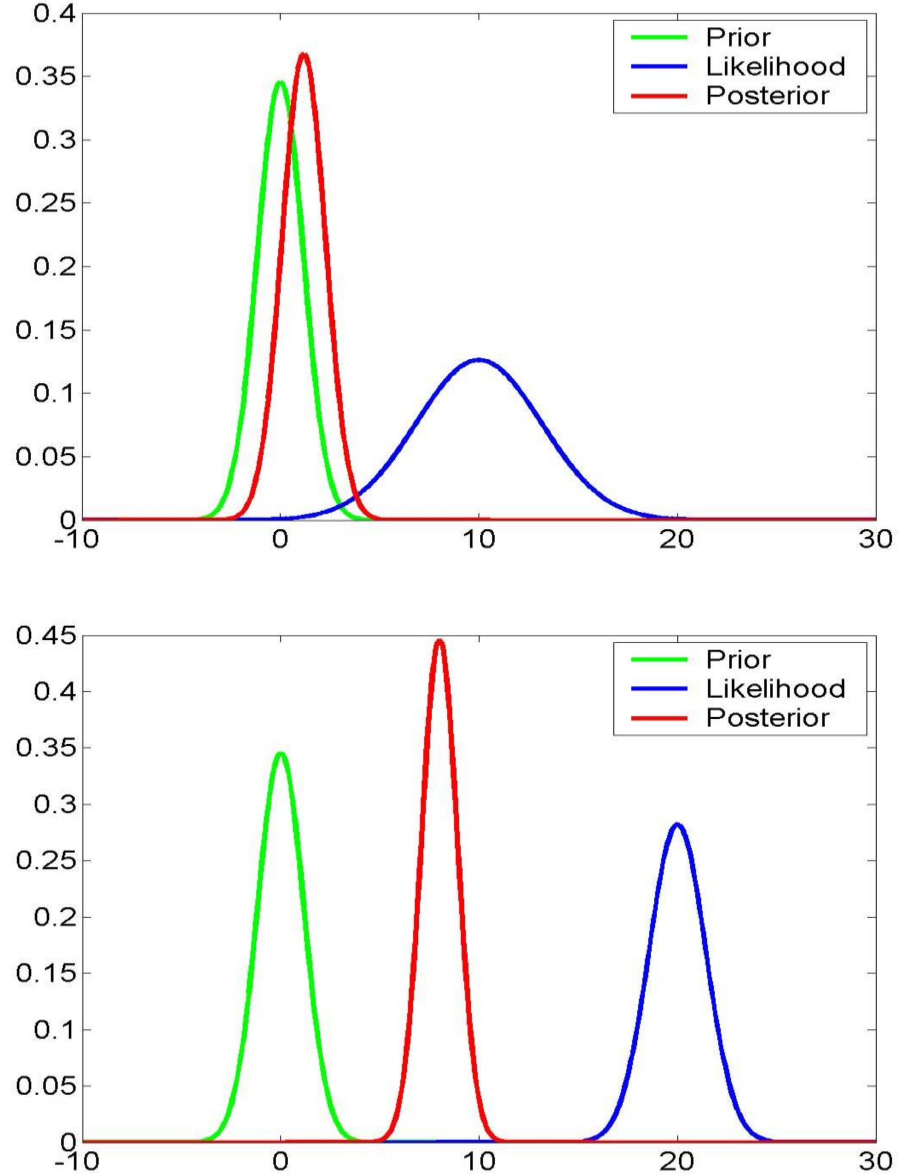
⇒ posterior

⇒ model evidence

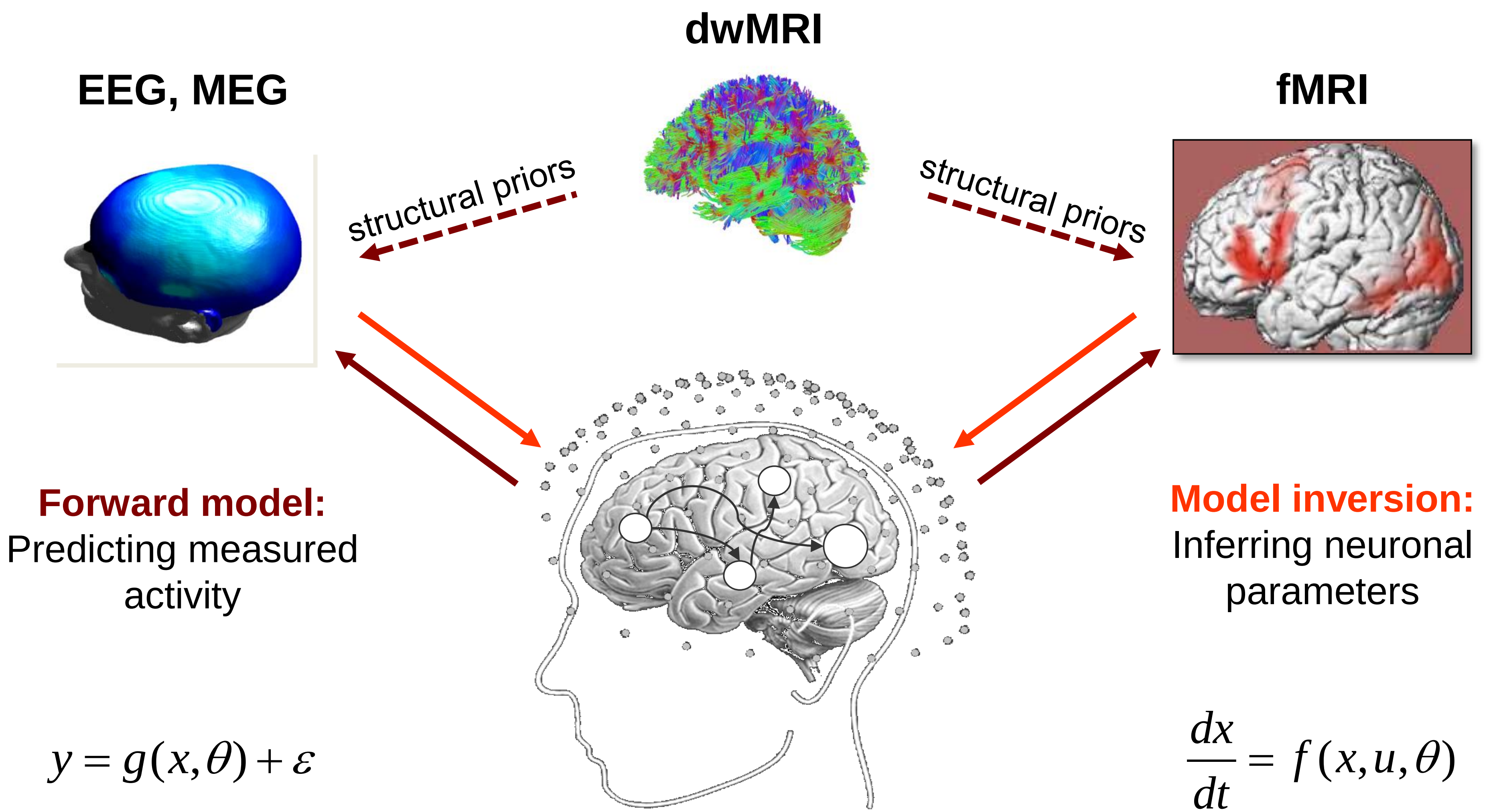
Priors

- Objective priors:
  - "non-informative" priors
  - objective constraints (e.g., non-negativity)
- Subjective priors:
  - subjective but not arbitrary
  - can express beliefs that result from understanding of the problem or system
  - can be result of previous empirical results
- Shrinkage priors:
  - emphasize regularization and sparsity
- Empirical priors:
  - learn parameters of prior distributions from the data ("empirical Bayes")
  - rest on a hierarchical model

Example of a shrinkage prior



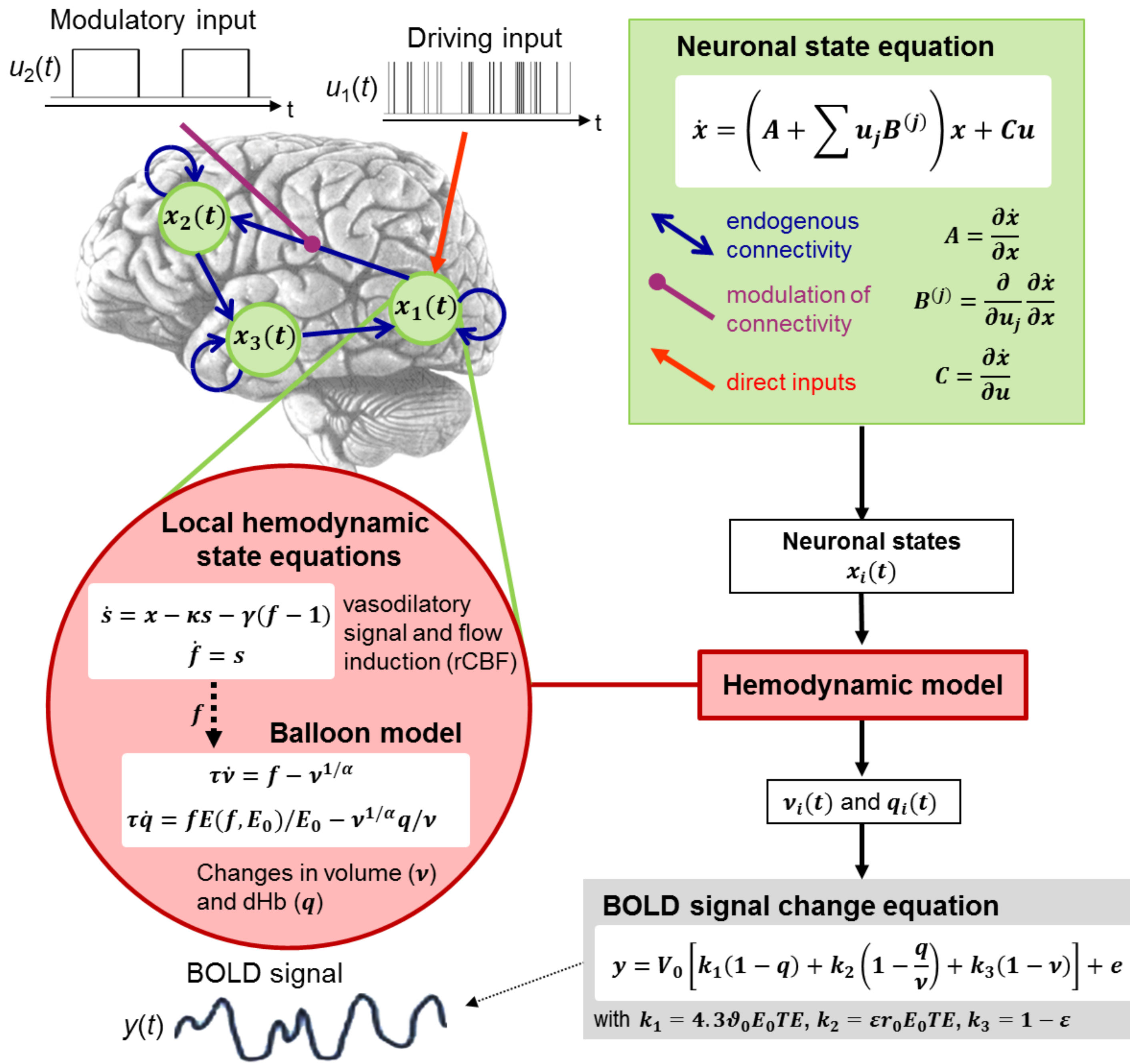
A generative modelling framework for fMRI & EEG: Dynamic causal modeling (DCM)



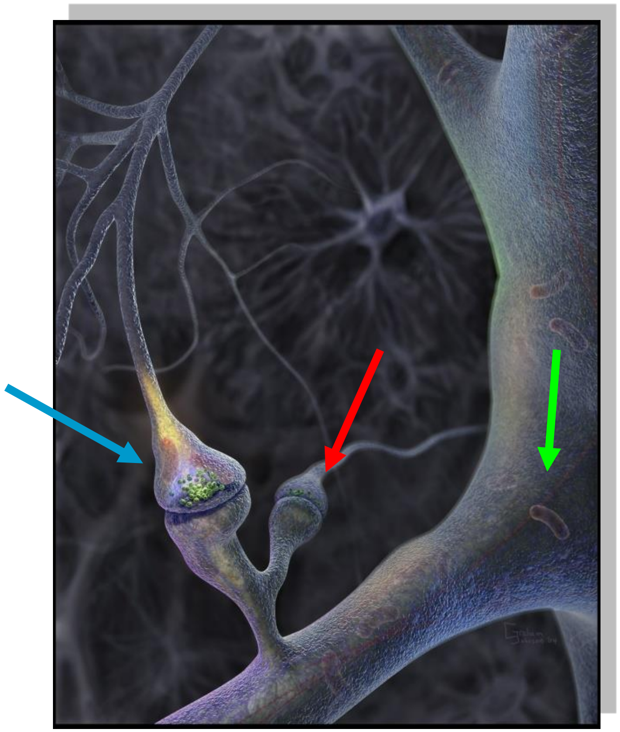
Friston et al. 2003, *NeuroImage*

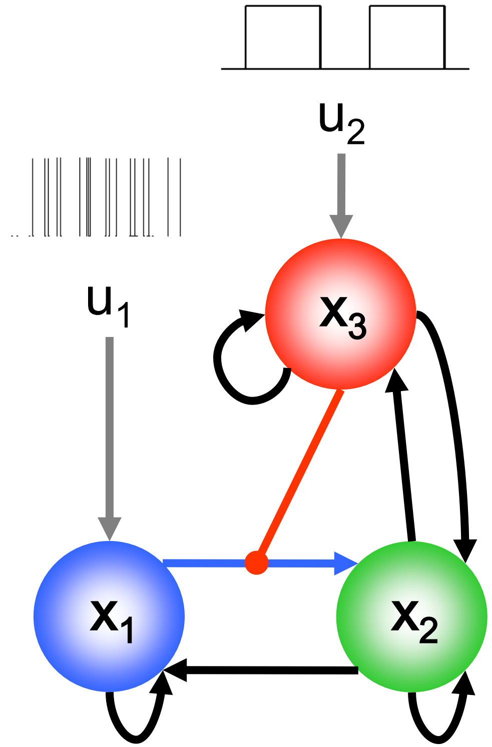
Stephan et al. 2009, *NeuroImage*

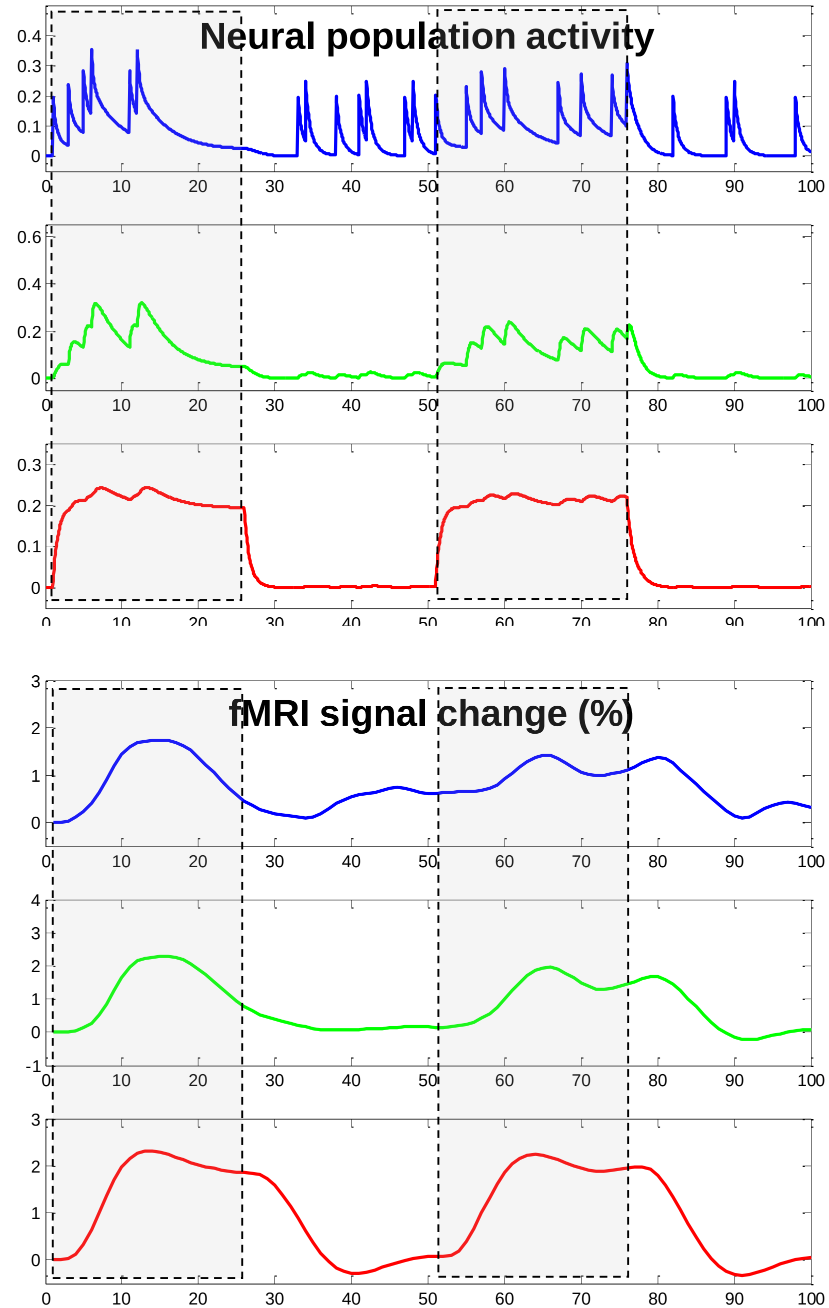
DCM for fMRI



Stephan et al. 2015, *Neuron*







Nonlinear Dynamic Causal Model for fMRI

$$\frac{dx}{dt} = \left(A + \sum_{i=1}^{m} u_i B^{(i)} + \sum_{j=1}^{n} x_j D^{(j)} \right) x + Cu$$

Stephan et al. 2008, *NeuroImage*

Bayesian system identification

$$u(t)$$

Neural dynamics

$$dx/dt = f(x,u,\theta)$$

Observer function

$$y = g(x,\theta) + \varepsilon$$

$$p(y \mid \theta, m) = N(g(\theta), \Sigma(\theta))$$
$$p(\theta, m) = N(\mu_\theta, \Sigma_\theta)$$

Inference on model structure

$$p(y \mid m) = \int p(y \mid \theta, m)\, p(\theta)\, d\theta$$

Inference on parameters

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)\, p(\theta, m)}{p(y \mid m)}$$

Design experimental inputs

Define likelihood model

Specify priors

Invert model

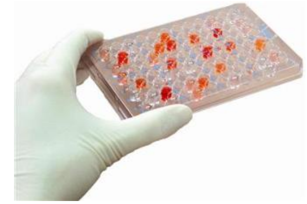
Make inferences

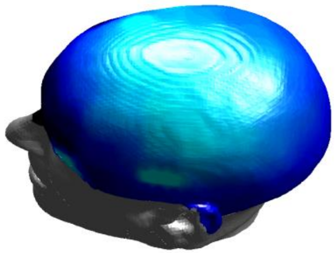
Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
  - computational psychosomatics
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

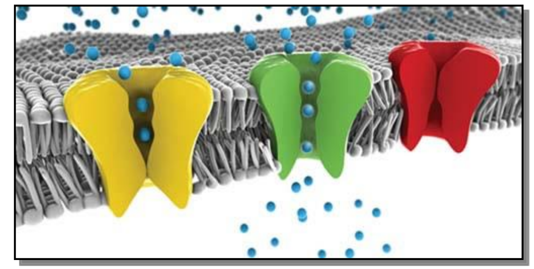
Generative models as "computational assays"

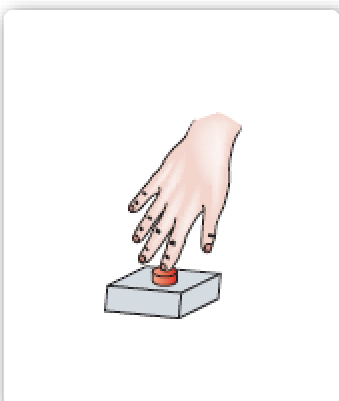




$$p(y\,|\,\theta,m)\cdot p(\theta\,|\,m)$$

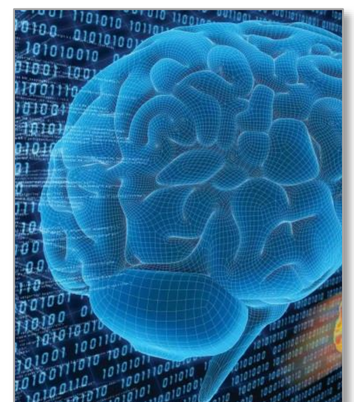
$$p(\theta\,|\,y,m)$$





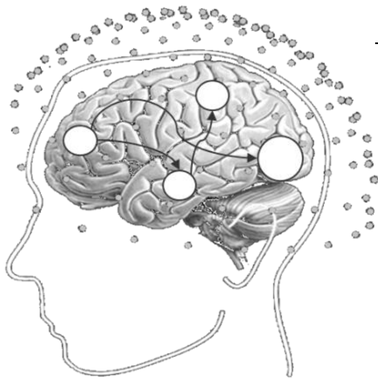
$$p(y\,|\,\theta,m)\cdot p(\theta\,|\,m)$$

$$p(\theta\,|\,y,m)$$

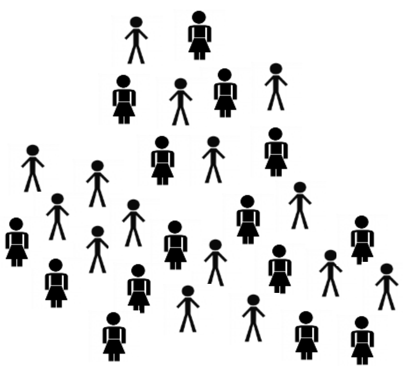


Translational Neuromodeling

❶ **Computational assays:
Models of disease mechanisms**

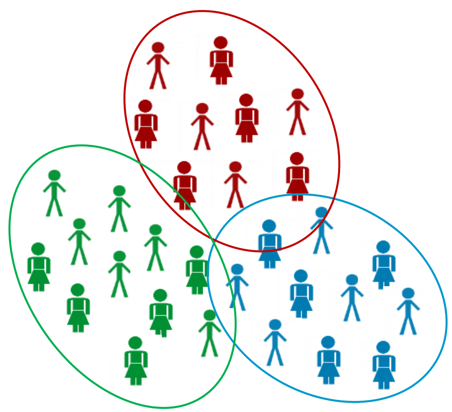


$$\frac{dx}{dt} = f(x,u,\theta) + \omega$$



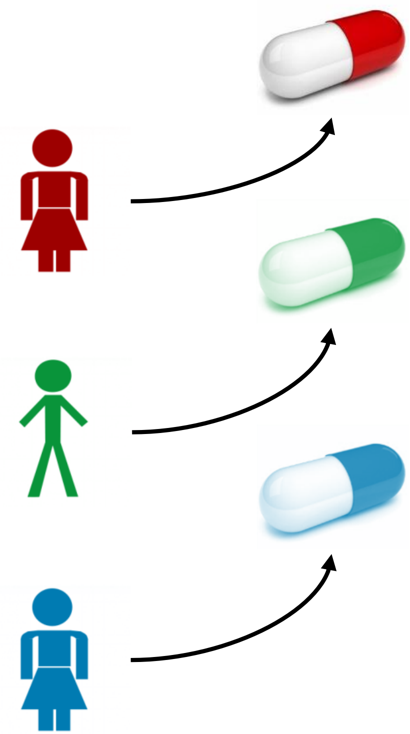
❷ **Application to brain activity and behaviour of individual patients**

❸ **Detecting physiological subgroups (based on inferred mechanisms)**



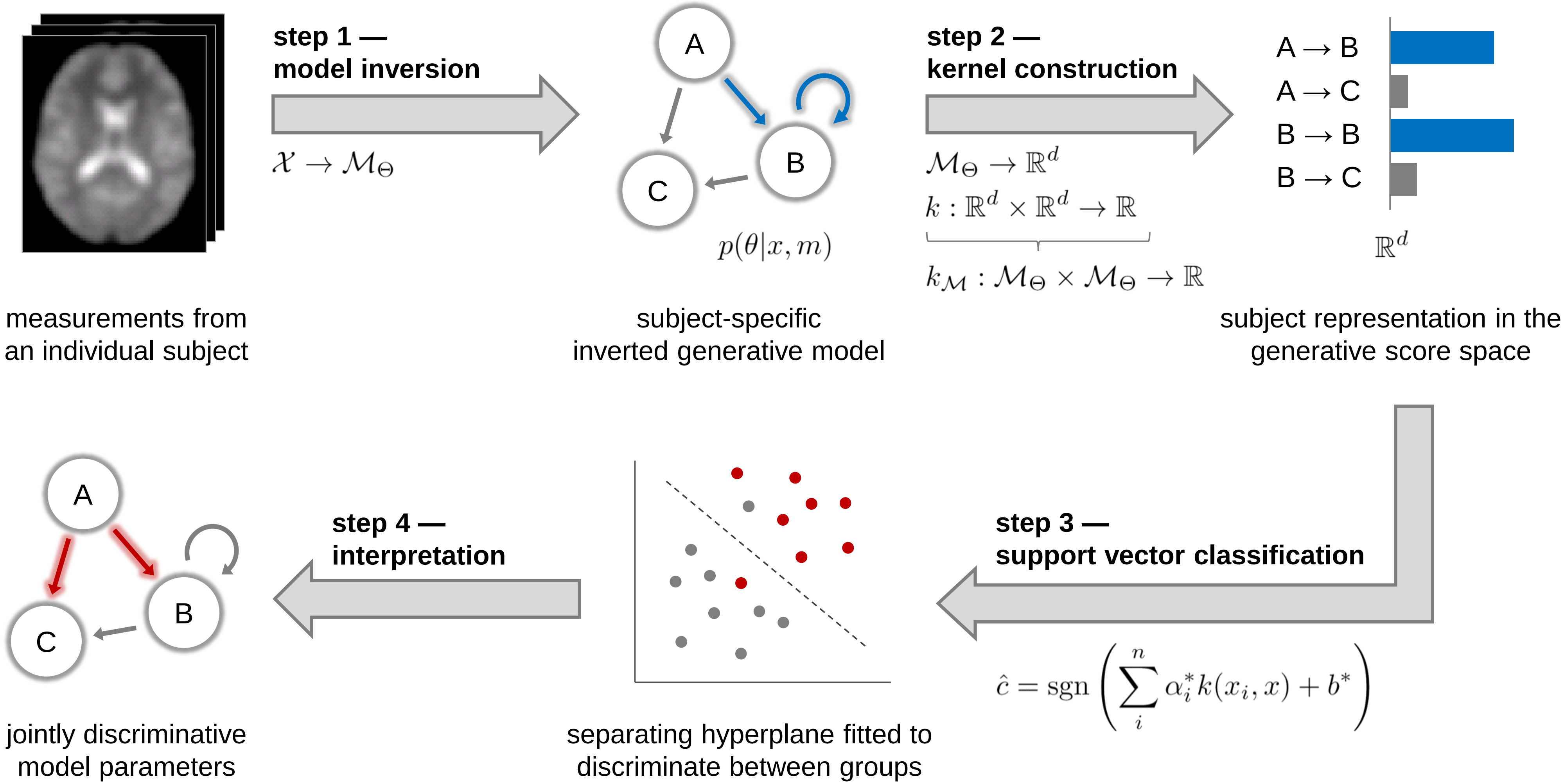
• disease mechanism A
• disease mechanism B
• disease mechanism C

❹ **Individual treatment prediction**



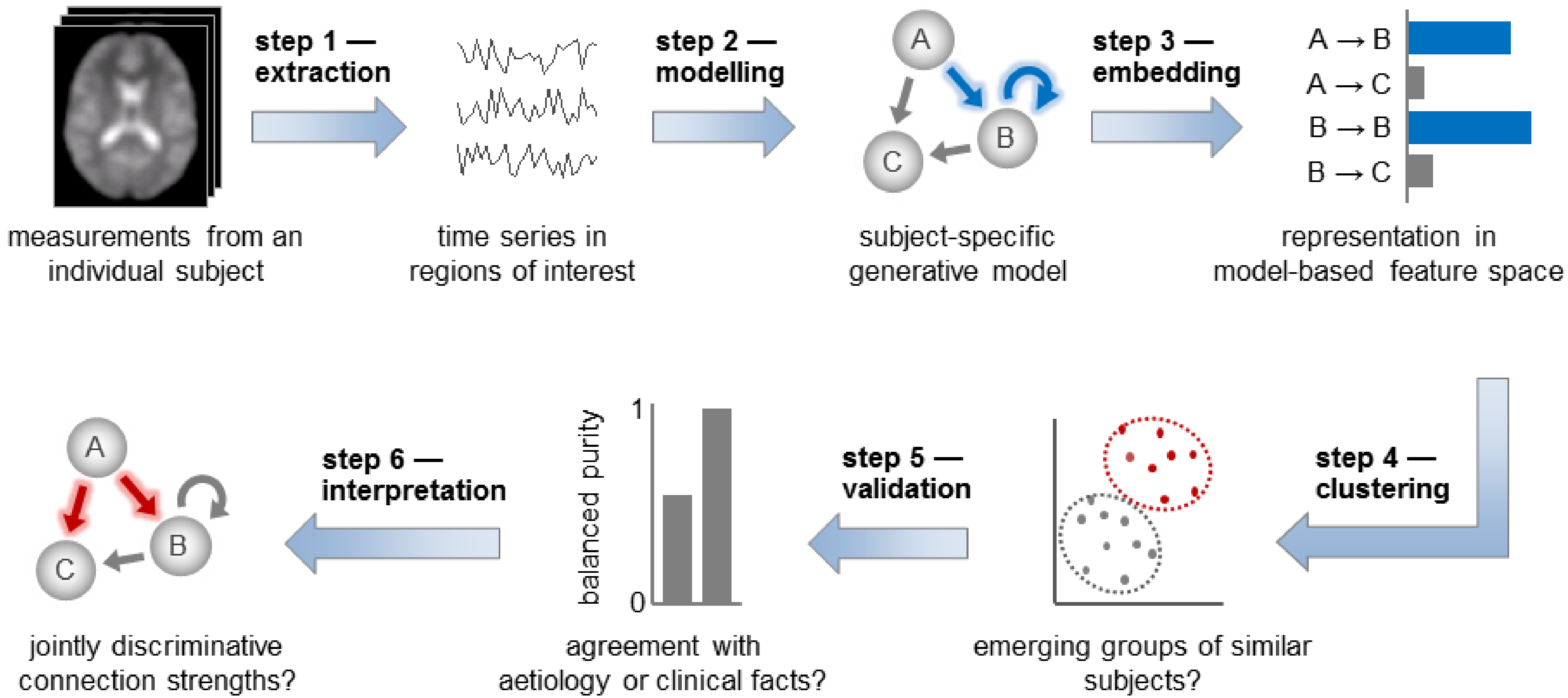
Stephan et al. 2015, *Neuron*

Generative embedding (supervised)



Brodersen et al. 2011, *PLoS Comput. Biol.*

Generative embedding (unsupervised)

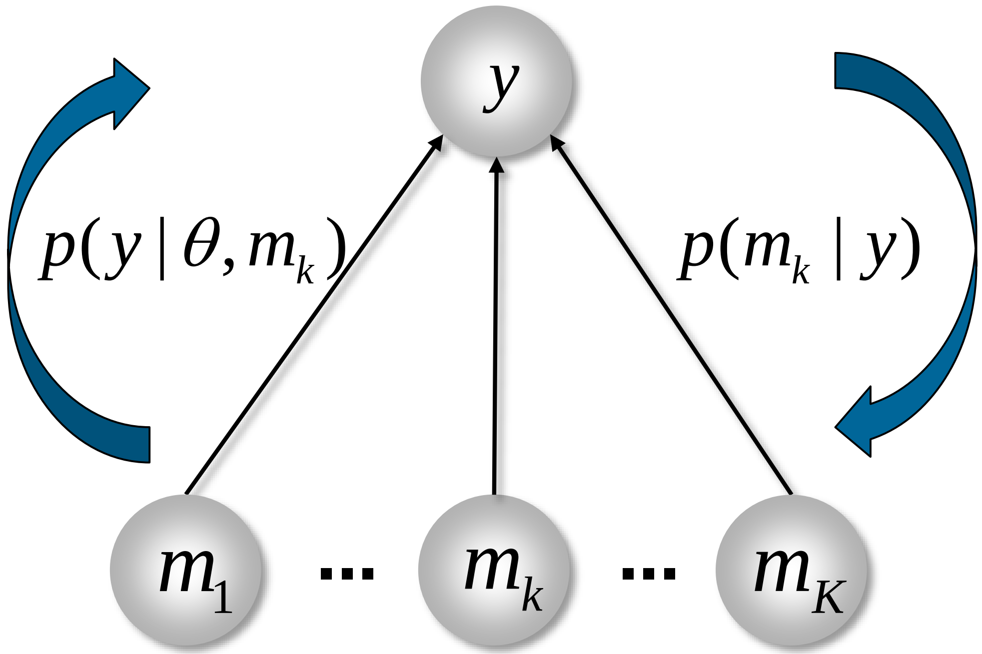


Brodersen et al. 2014, *NeuroImage Clinical*

Differential diagnosis by model selection

SYMPTOM
(behaviour
or physiology)

HYPOTHETICAL
MECHANISM



$$p(m_k \mid \mathrm{y}) = \frac{p(y \mid m_k)\, p(m_k)}{\sum_k p(y \mid m_k)\, p(m_k)}$$

Stephan et al. 2017, *NeuroImage*

Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

- **statistical inference** in general
- **system identification**
- **translational neuromodeling** ("computational assays")
  - computational psychiatry
  - computational neurology
  - computational psychosomatics
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

Perception = inversion of a hierarchical generative model

neuronal states

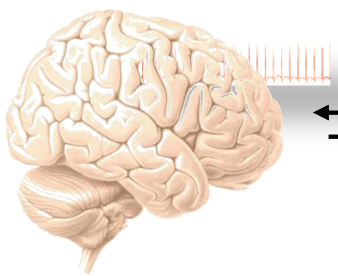
environm. states

others' mental states

bodily states

forward model

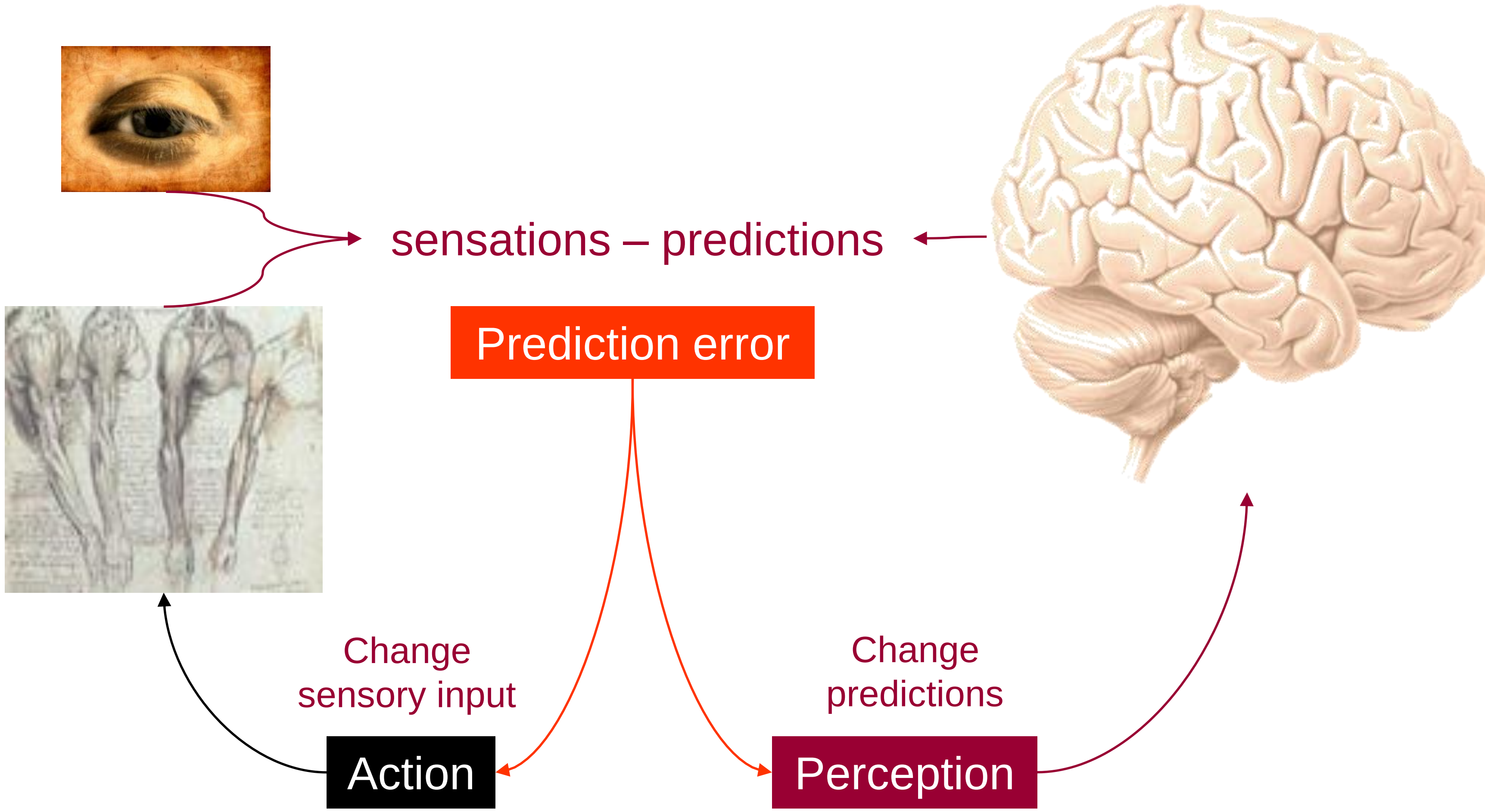
$$p(y \mid x, m) \cdot p(x \mid m)$$



$$p(x \mid y, m)$$

perception

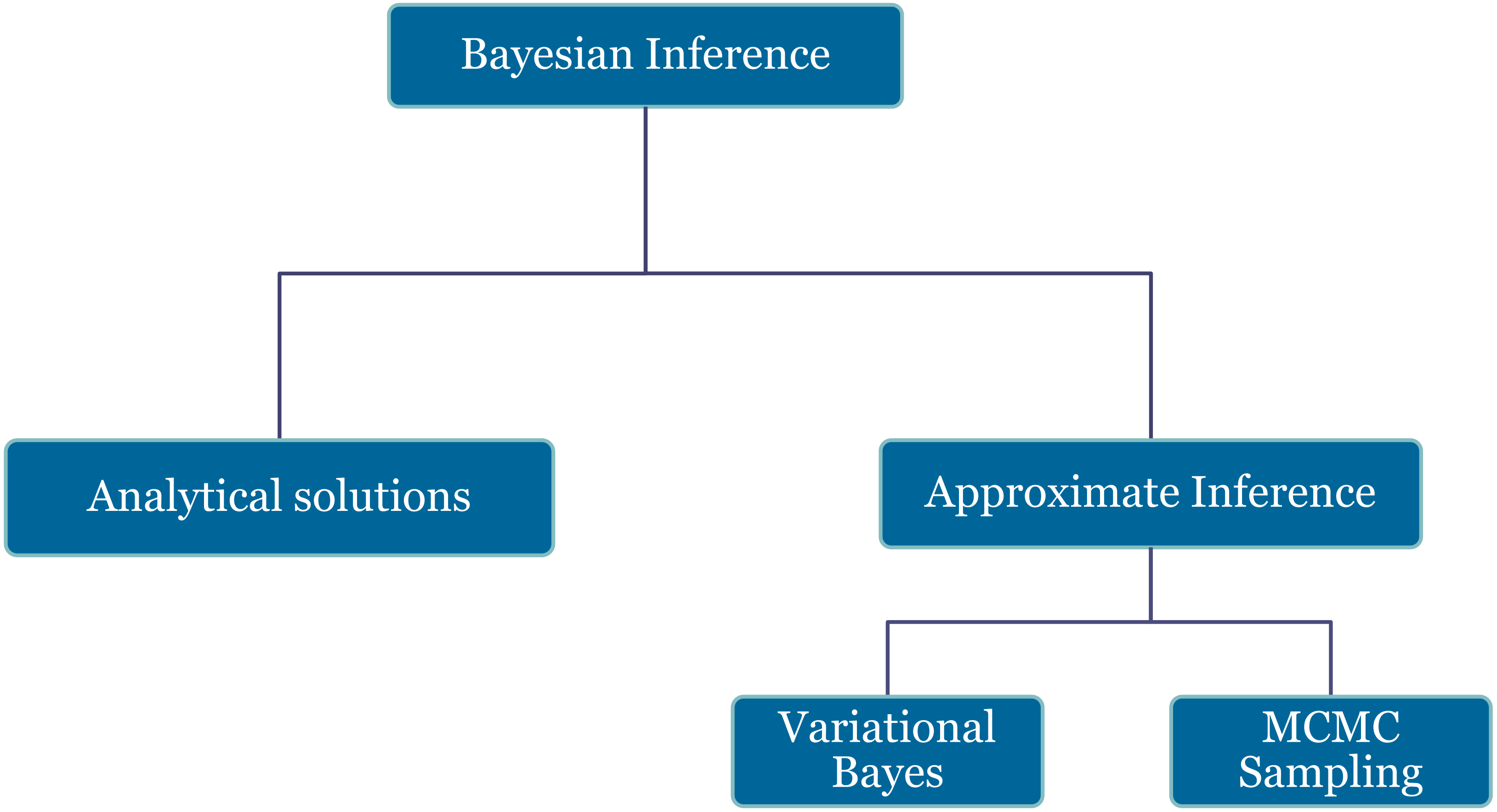
Free energy principle: predictive coding & active inference



Maximizing the evidence (of the brain's generative model)
= minimizing the surprise about the data (sensory inputs).

Friston et al. 2006,
J Physiol Paris

How is the posterior computed = how is a generative model inverted?



How is the posterior computed = how is a generative model inverted?

- **compute the posterior analytically**
  - requires conjugate priors
- **variational Bayes (VB)**
  - often hard work to derive, but fast to compute
  - uses approximations (approx. posterior, mean field)
  - problems: local minima, potentially inaccurate approximations
- **Sampling: Markov Chain Monte Carlo (MCMC)**
  - theoretically guaranteed to be accurate (for infinite computation time)
  - problems: may require very long run time in practice, convergence difficult to prove

Conjugate priors

- for a given likelihood function, the choice of prior determines the algebraic form of the posterior
- for some probability distributions a prior can be found such that the posterior has the same algebraic form as the prior
- such a prior is called "conjugate" to the likelihood
- examples:
  - Normal × Normal $\propto$ Normal
  - Beta × Binomial $\propto$ Beta
  - Dirichlet × Multinomial $\propto$ Dirichlet

$$\underbrace{p(\boldsymbol{\theta} \mid \mathbf{y})}_{\text{same form}} \propto p(\mathbf{y} \mid \boldsymbol{\theta})\, \underbrace{p(\boldsymbol{\theta})}_{\text{same form}}$$

same form

Posterior mean & variance of univariate Gaussians

$$y = \theta + \varepsilon$$

Likelihood & Prior

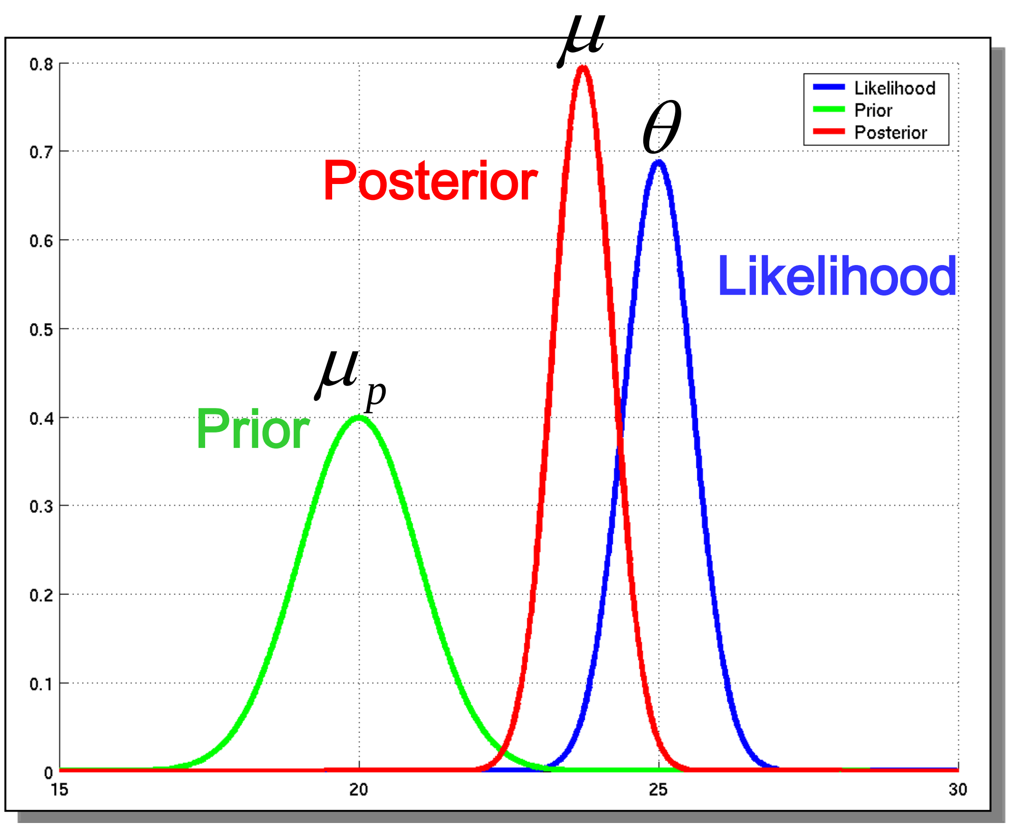
$$p(y \mid \theta) = N(\theta, \sigma_e^2)$$

$$p(\theta) = N(\mu_p, \sigma_p^2)$$

Posterior: $p(\theta \mid y) = N(\mu, \sigma^2)$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_e^2} + \frac{1}{\sigma_p^2}$$

$$\mu = \sigma^2 \left(\frac{1}{\sigma_e^2}\theta + \frac{1}{\sigma_p^2}\mu_p \right)$$



Posterior mean = variance-weighted combination of prior mean and data mean

Same thing – but expressed as precision weighting

Likelihood & prior

$$p(y\,|\,\theta) = N(\theta, \lambda_e^{-1})$$

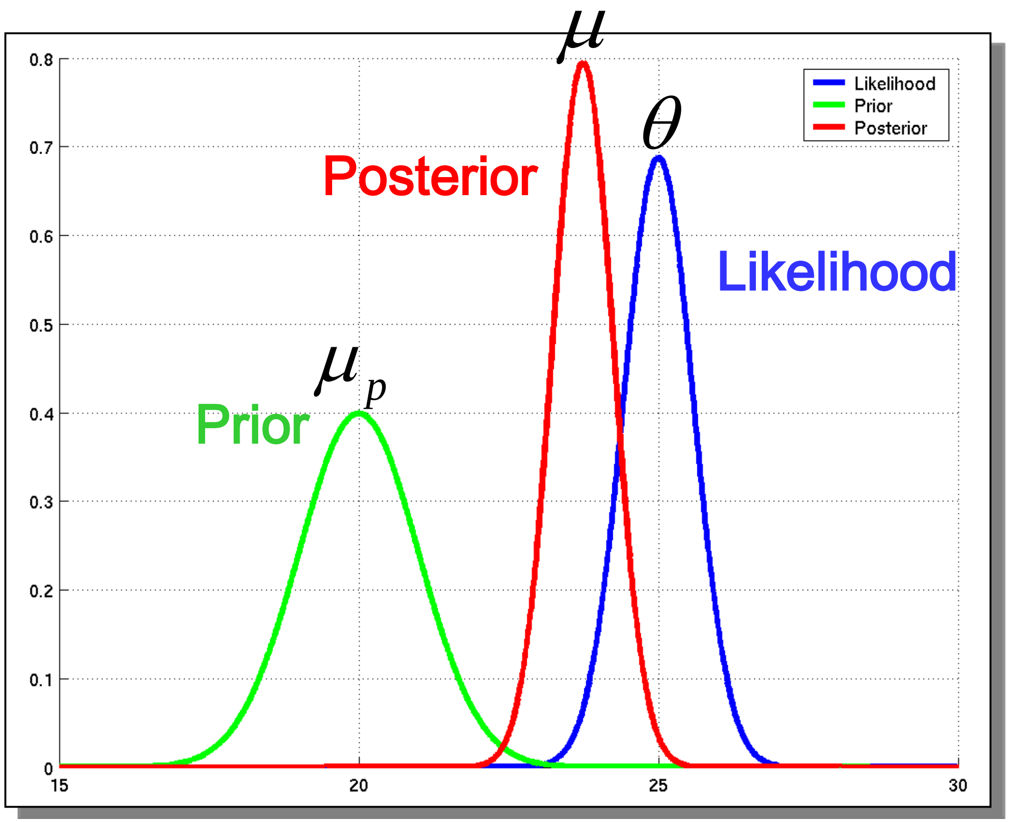
$$p(\theta) = N(\mu_p, \lambda_p^{-1})$$

$$y = \theta + \varepsilon$$

Posterior: $p(\theta\,|\,y) = N(\mu, \lambda^{-1})$

$$\lambda = \lambda_e + \lambda_p$$

$$\mu = \frac{\lambda_e}{\lambda}\theta + \frac{\lambda_p}{\lambda}\mu_p$$

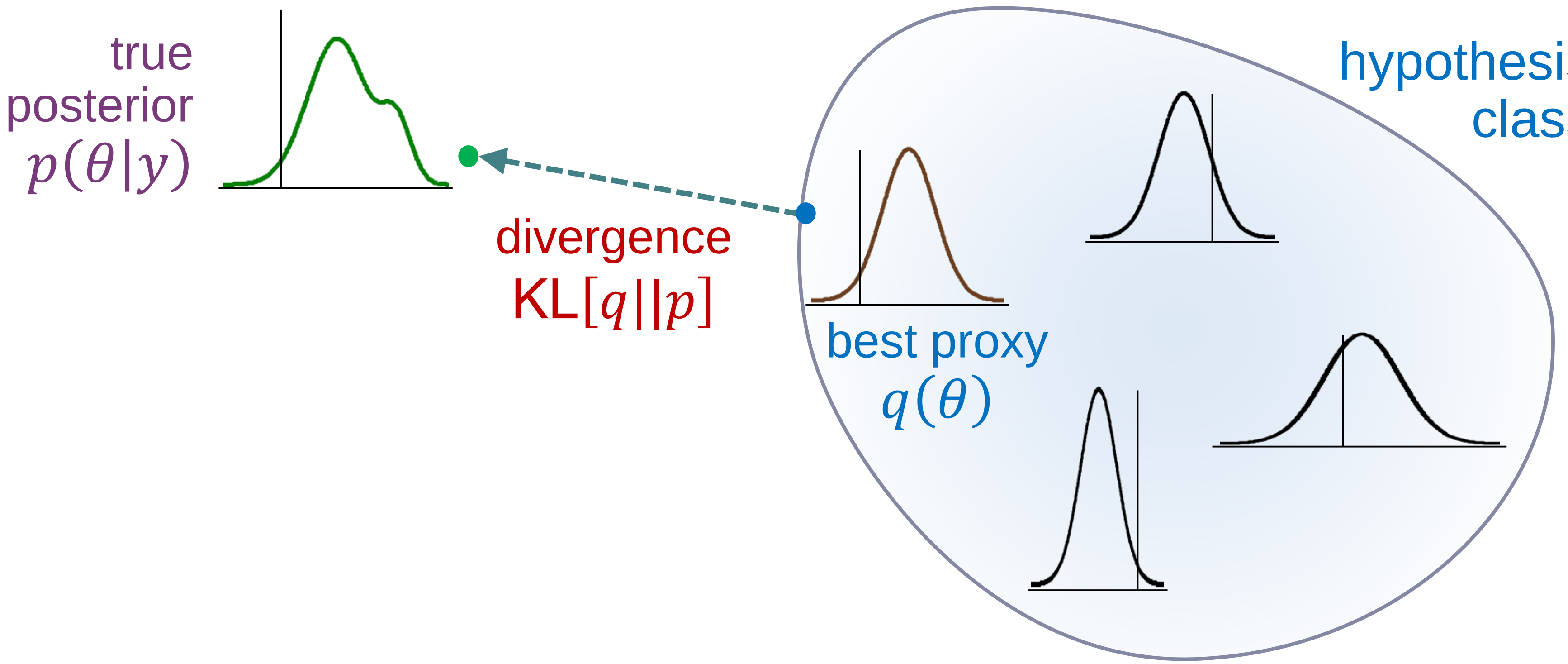


Relative precision weighting

Variational Bayes (VB)

Idea: find an approximate density $q(\theta)$ that is maximally similar to the true posterior $p(\theta|y)$.

This is often done by assuming a particular form for q (fixed form VB) and then optimizing its sufficient statistics.



Kullback–Leibler (KL) divergence

- asymmetric measure of the difference between two probability distributions P and Q

- Interpretations of $D_{KL}(P\|Q)$:
  - "Bayesian surprise" when Q=prior, P=posterior: measure of the information gained when one updates one's prior beliefs to the posterior P
  - a measure of the information lost when Q is used to approximate P

- non-negative: ≥0 (zero when P=Q)

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} \, \mathrm{d}x,$$

Variational calculus

Standard calculus

Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{\mathrm{d}f}{\mathrm{d}x}$

Example: maximize the likelihood expression $p(y|\theta)$ w.r.t. θ

Variational calculus

Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{\mathrm{d}F}{\mathrm{d}f}$

Example: maximize the entropy $H[p]$ w.r.t. a probability distribution $p(x)$

Leonhard Euler
(1707 – 1783)
Swiss mathematician, 'Elementa Calculi Variationum'

Variational Bayes

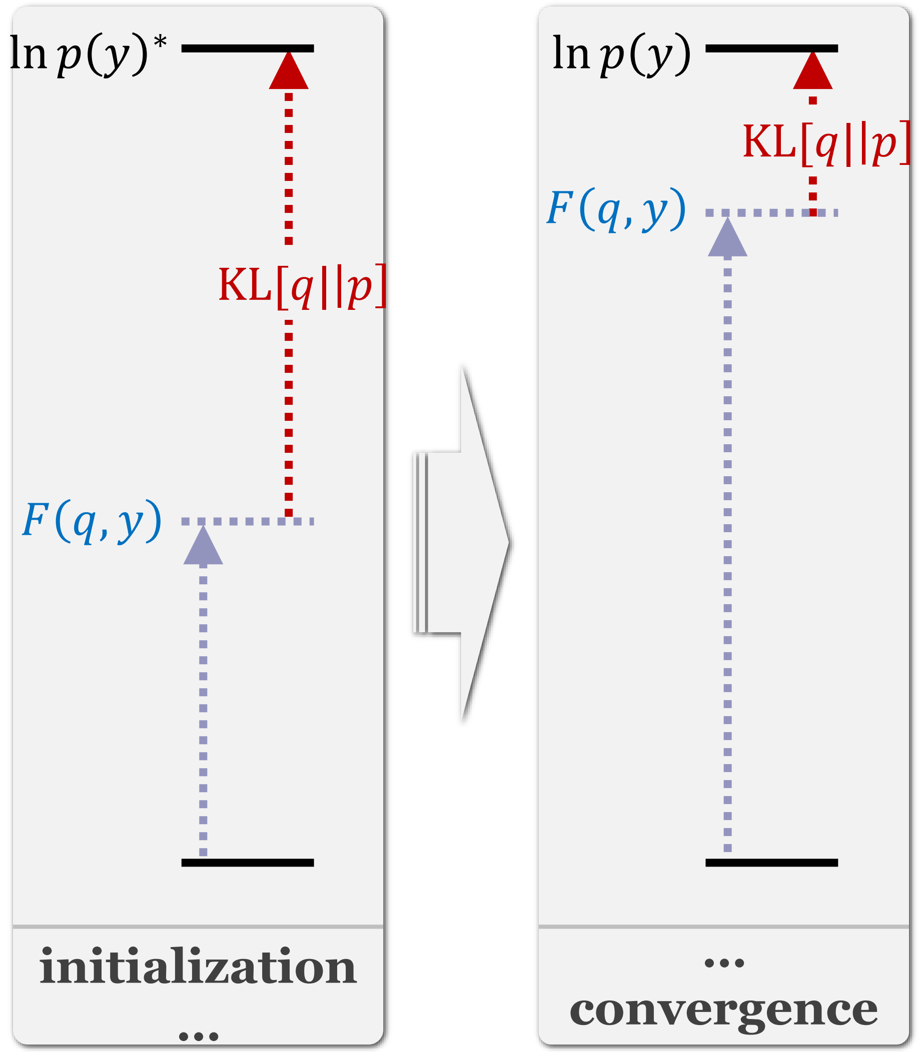
$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\substack{\textbf{divergence} \\ \geq\ 0 \\ \text{(unknown)}}} + \underbrace{F(q,y)}_{\substack{\textbf{neg. free} \\ \textbf{energy} \\ \text{(easy to evaluate} \\ \text{for a given } q)}}$$

$F(q)$ is a functional wrt. the approximate posterior $q(\theta)$.

Maximizing $F(q,y)$ is equivalent to:

- minimizing $\mathrm{KL}[q||p]$
- tightening $F(q,y)$ as a lower bound to the log model evidence

When $F(q,y)$ is maximized, $q(\theta)$ is our best estimate of the posterior.



Derivation of the (negative) free energy approximation

- See whiteboard!
- (or Appendix to Stephan et al. 2007, NeuroImage 38: 387-401)

Mean field assumption

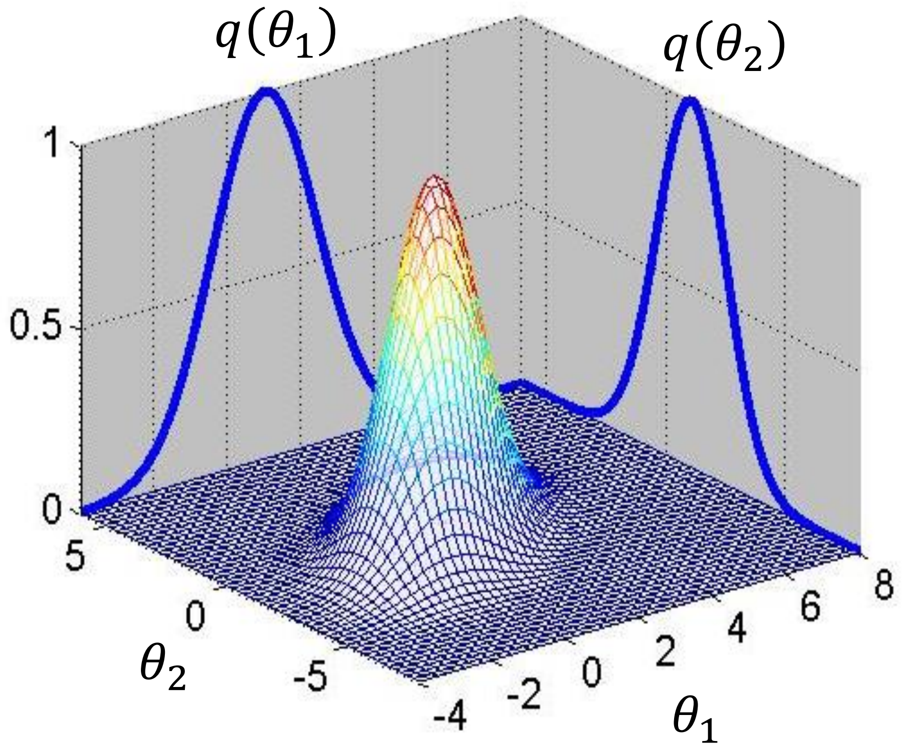
Factorize the approximate posterior $q(\theta)$ into independent partitions:

$$q(\theta) = \prod_i q_i(\theta_i)$$

where $q_i(\theta_i)$ is the approximate posterior for the i^{th} subset of parameters.

For example, split parameters and hyperparameters:

$$p(\theta, \lambda \mid y) \approx q(\theta, \lambda) = q(\theta)\, q(\lambda)$$



Jean Daunizeau, www.fil.ion.ucl.ac.uk/~jdaunize/presentations/Bayes2.pdf

VB in a nutshell (under mean-field approximation)

❶ Neg. free-energy approx. to model evidence.

$$\ln p(y|m) = F + KL\left[q(\theta,\lambda), p(\theta,\lambda|y)\right]$$

$$F = \left\langle \ln p(y|\theta,\lambda) \right\rangle_q - KL\left[q(\theta,\lambda), p(\theta,\lambda|m)\right]$$

❷ Mean field approx.

$$p(\theta,\lambda|y) \approx q(\theta,\lambda) = q(\theta)q(\lambda)$$

❸ Maximise neg. free energy wrt. q = minimise divergence, by maximising variational energies

$$q(\theta) \propto \exp(I_\theta) = \exp\left[\left\langle \ln p(y,\theta,\lambda) \right\rangle_{q(\lambda)}\right]$$

$$q(\lambda) \propto \exp(I_\lambda) = \exp\left[\left\langle \ln p(y,\theta,\lambda) \right\rangle_{q(\theta)}\right]$$

❹ Iterative updating of sufficient statistics of approx. posteriors by gradient ascent.

Model comparison and selection

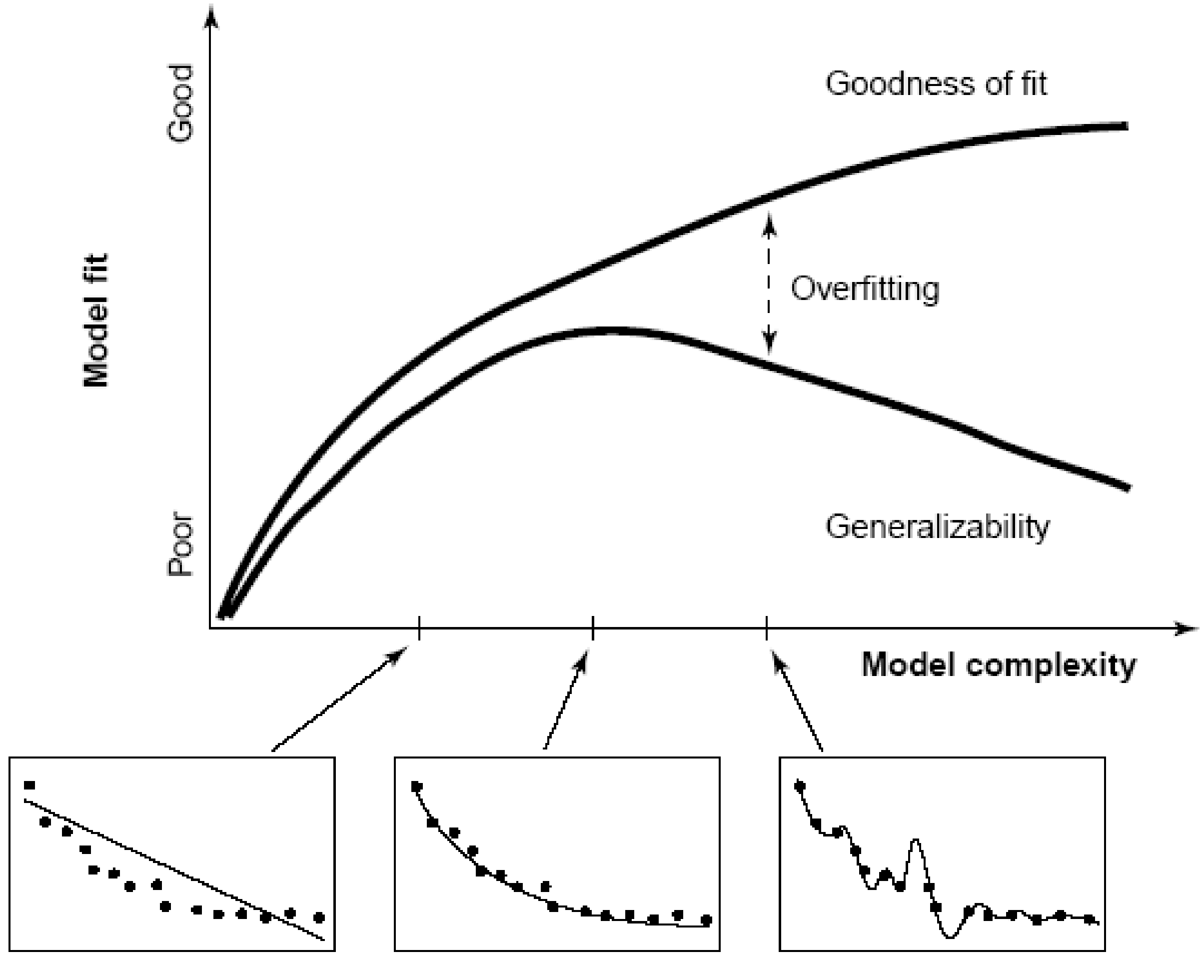
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

⇩

Which model represents the best balance between model fit and model complexity?

⇩

For which model m does $p(y|m)$ become maximal?



Pitt & Miyung (2002) *TICS*

Bayesian model selection (BMS)

- First step of inference: define model space M

$$|M| \in [1, \infty[$$

- Inference on model structure m:

Posterior model probability

$$p(m \mid y) = \frac{p(y \mid m)\, p(m)}{p(y)}$$

$$= \frac{p(y \mid m)\, p(m)}{\sum_m p(y \mid m)\, p(m)}$$

- For a uniform prior on m, model evidence sufficient for model selection

Model evidence:

$$p(y \mid m) = \int p(y \mid \theta, m)\, p(\theta \mid m)\; d\theta$$

Bayesian model selection (BMS)

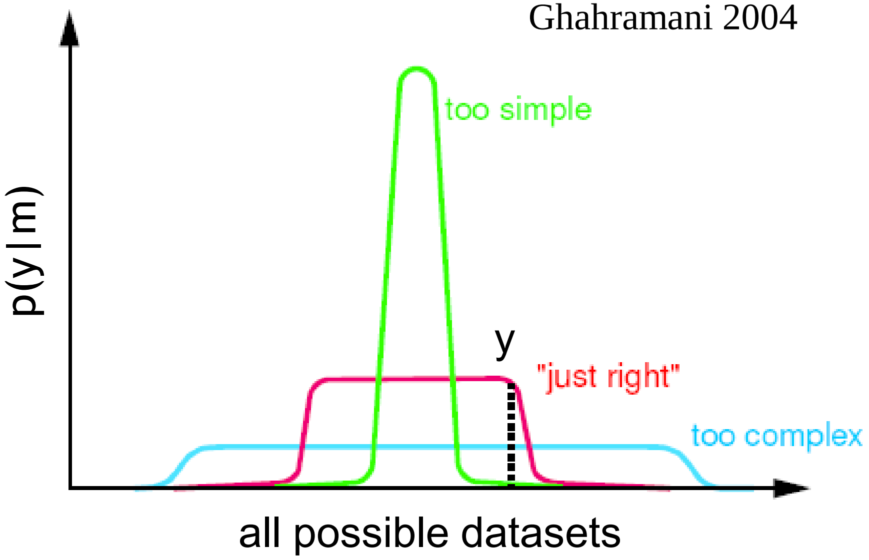
Model evidence:

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

⇨ probability that data were generated by model m, averaging over all possible parameter values (as specified by the prior)

⇨ accounts for both accuracy and complexity of the model

Ghahramani 2004



Various approximations:
- negative free energy (F)
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

Bayesian model selection (BMS)

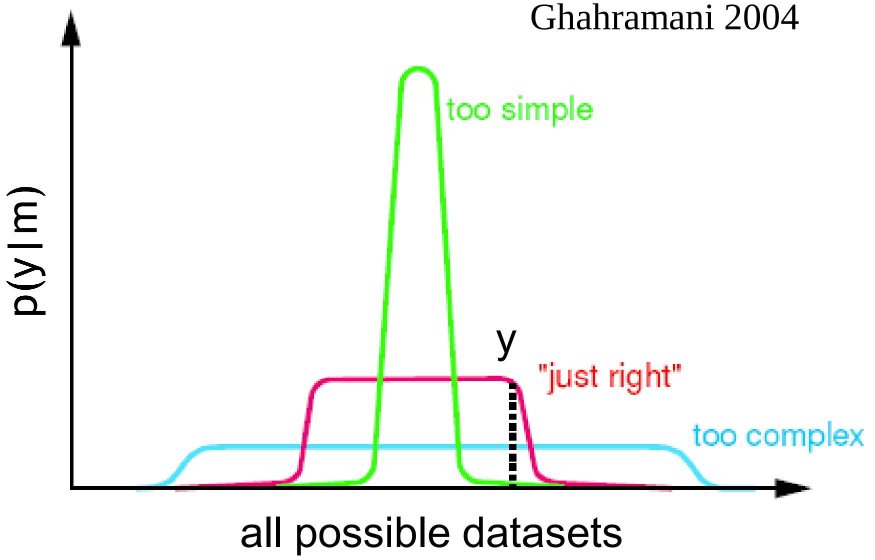
Model evidence:

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

⇨ "If I randomly sampled from my prior and plugged the resulting value into the likelihood function, how close would the predicted data be – on average – to my observed data?"

⇨ accounts for both accuracy and complexity of the model

Ghahramani 2004



Various approximations:

- negative free energy (F)
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

Approximations to the model evidence

Logarithm is a monotonic function $\Rightarrow$ Maximizing log model evidence = Maximizing model evidence

Log model evidence = balance between fit and complexity

$$\log p(y \mid m) = accuracy(m) - complexity(m)$$
$$= \log p(y \mid \theta, m) - complexity(m)$$

No. of parameters

No. of data points

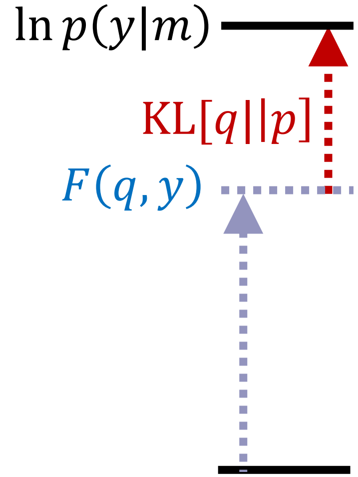
Akaike Information Criterion: $AIC = \log p(y \mid \theta, m) - p$

Bayesian Information Criterion: $BIC = \log p(y \mid \theta, m) - \frac{p}{2} \log N$

The (negative) free energy approximation F

F is a lower bound on the log model evidence:

$$\log p(y \mid m) = F + KL\left[q(\theta), p(\theta \mid y, m)\right]$$



Like AIC/BIC, F is an accuracy/complexity tradeoff:

$$F = \underbrace{\langle \log p(y \mid \theta, m) \rangle}_{accuracy} - \underbrace{KL\left[q(\theta), p(\theta \mid m)\right]}_{complexity}$$

The (negative) free energy approximation

- Log evidence is thus expected log likelihood (wrt. q) plus 2 KL's:

$$
\begin{aligned}
&\log p(y \mid m) \\
&= \langle \log p(y \mid \theta, m) \rangle - KL\left[q(\theta), p(\theta \mid m)\right] + KL\left[q(\theta), p(\theta \mid y, m)\right]
\end{aligned}
$$

$$
\boxed{
\begin{aligned}
F &= \log p(y \mid m) - KL\left[q(\theta), p(\theta \mid y, m)\right] \\
&= \underbrace{\langle \log p(y \mid \theta, m) \rangle}_{accuracy} - \underbrace{KL\left[q(\theta), p(\theta \mid m)\right]}_{complexity}
\end{aligned}
}
$$

The complexity term in *F*

- In contrast to AIC & BIC, the complexity term of the negative free energy *F* accounts for parameter interdependencies.
  Under Gaussian assumptions about the posterior (Laplace approximation):

$$
\begin{aligned}
&KL\left[q(\theta), p(\theta \mid m)\right] \\
&= \frac{1}{2}\ln\left|C_\theta\right| - \frac{1}{2}\ln\left|C_{\theta|y}\right| + \frac{1}{2}\left(\mu_{\theta|y} - \mu_\theta\right)^T C_\theta^{-1}\left(\mu_{\theta|y} - \mu_\theta\right)
\end{aligned}
$$

- The complexity term of *F* is higher
  - the more independent the prior parameters (↑ effective DFs)
  - the more dependent the posterior parameters
  - the more the posterior mean deviates from the prior mean

Bayes factors

To compare two models, we could just compare their log evidences.

But: the log evidence is just some number – not very intuitive!

A more intuitive interpretation of model comparisons is made possible by Bayes factors:

$$B_{12} = \frac{p(y \mid m_1)}{p(y \mid m_2)}$$

positive value, $[0; \infty[$

Kass & Raftery classification:

B_{12}	$p(m_1 \mid y)$	Evidence
1 to 3	50-75%	weak
3 to 20	75-95%	positive
20 to 150	95-99%	strong
$\geq$ 150	$\geq$ 99%	Very strong

Kass & Raftery 1995, *J. Am. Stat. Assoc.*

Fixed effects BMS at group level

Group Bayes factor (GBF) for 1...*K* subjects:

$$GBF_{ij} = \prod_k BF_{ij}^{(k)}$$

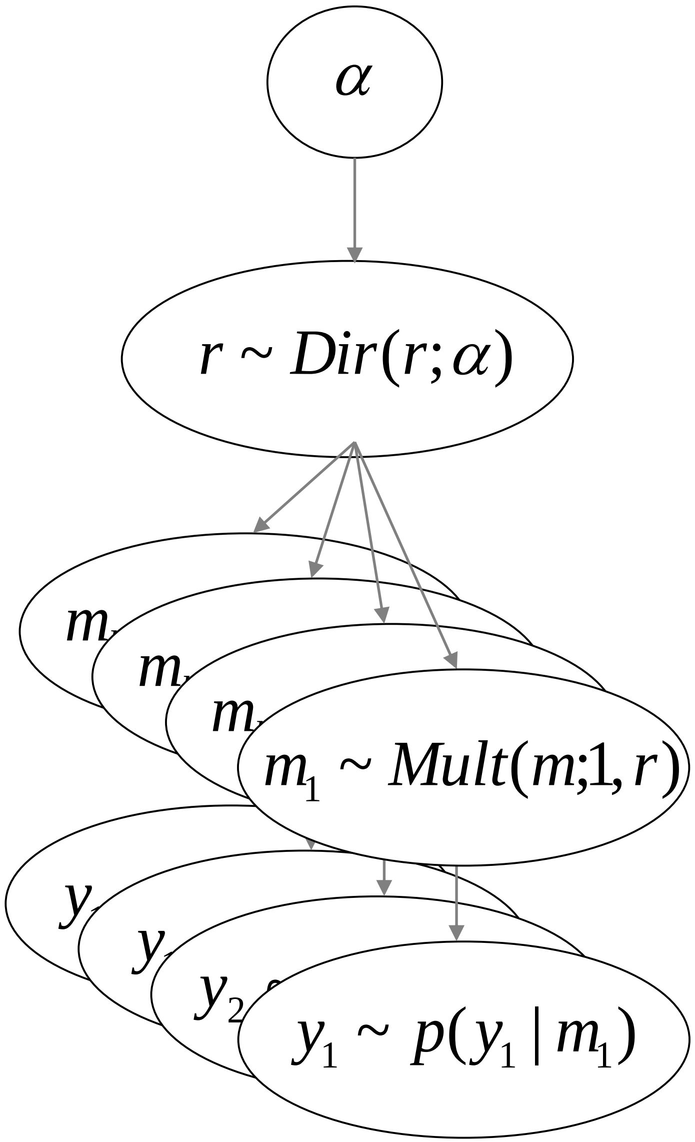
Average Bayes factor (ABF):

$$ABF_{ij} = \sqrt[K]{\prod_k BF_{ij}^{(k)}}$$

Problems:

- blind with regard to group heterogeneity
- sensitive to outliers

Random effects BMS for heterogeneous groups



Dirichlet parameters α
= “occurrences” of models in the population

Dirichlet distribution of model probabilities r

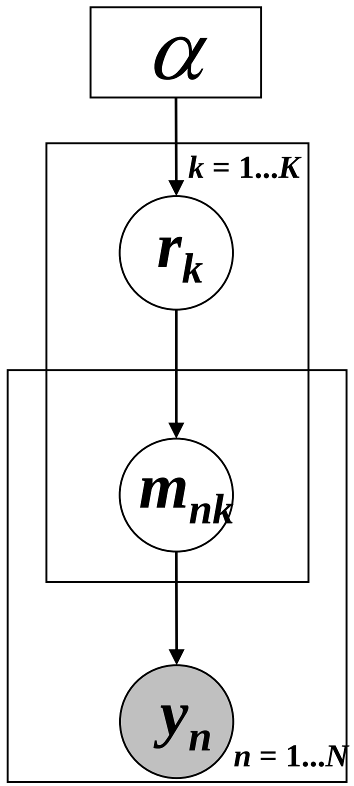
Multinomial distribution of model labels m

Model inversion by Variational Bayes or MCMC

Measured data y

Stephan et al. 2009, *NeuroImage*

Random effects BMS for heterogeneous groups



Dirichlet parameters α
= “occurrences” of models in the population

Dirichlet distribution of model probabilities r

Multinomial distribution of model labels m

Model inversion by Variational Bayes or MCMC

Measured data y

Stephan et al. 2009, *NeuroImage*

Four equivalent options for reporting model ranking by random effects BMS

1. **Dirichlet parameter estimates**

$$\alpha$$

2. **expected posterior probability** of obtaining the k-th model for any randomly selected subject

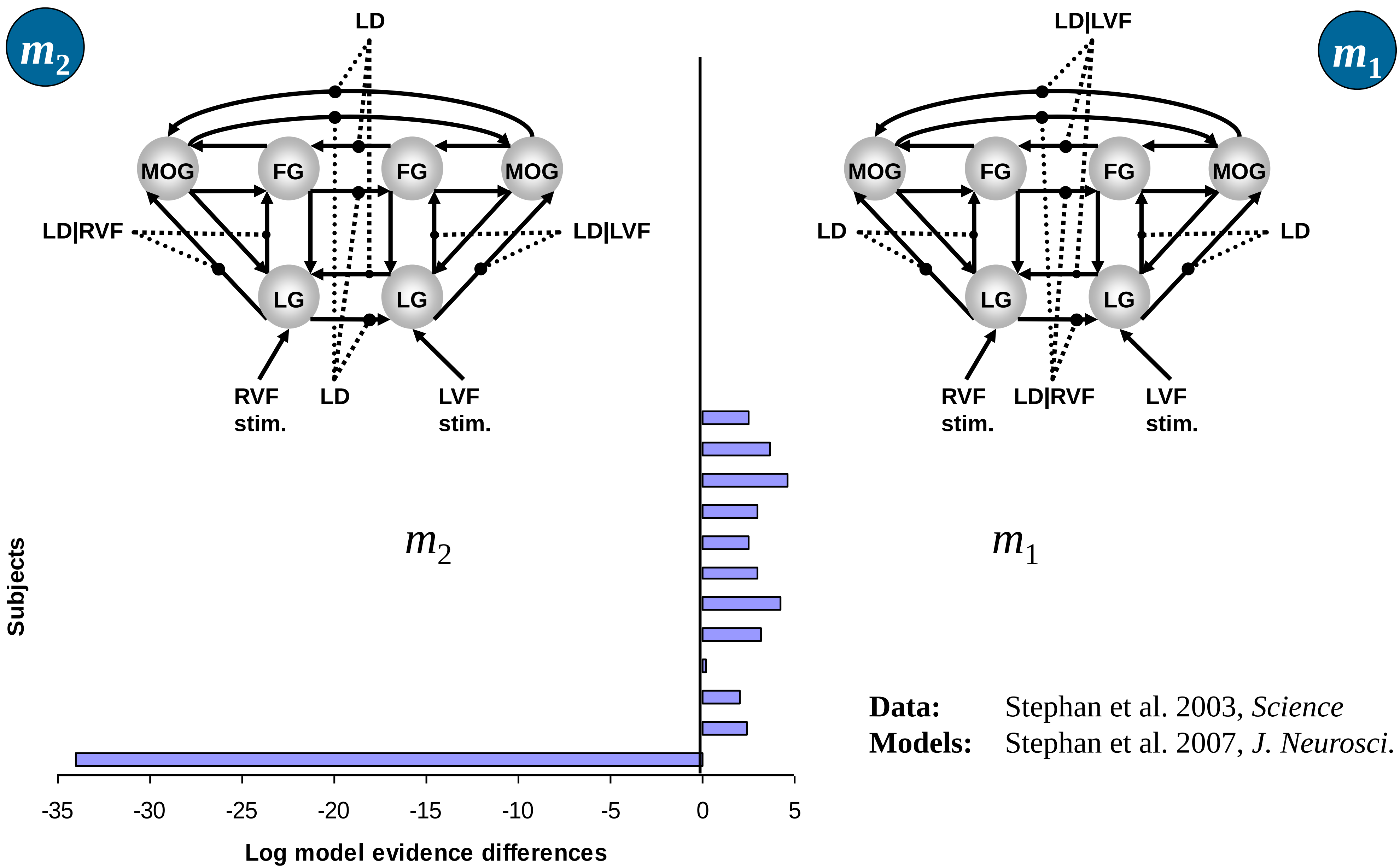
$$\langle r_k \rangle_q = \alpha_k / (\alpha_1 + \ldots + \alpha_K)$$

3. **exceedance probability** that a particular model *k* is more likely than any other model (of the *K* models tested), given the group data

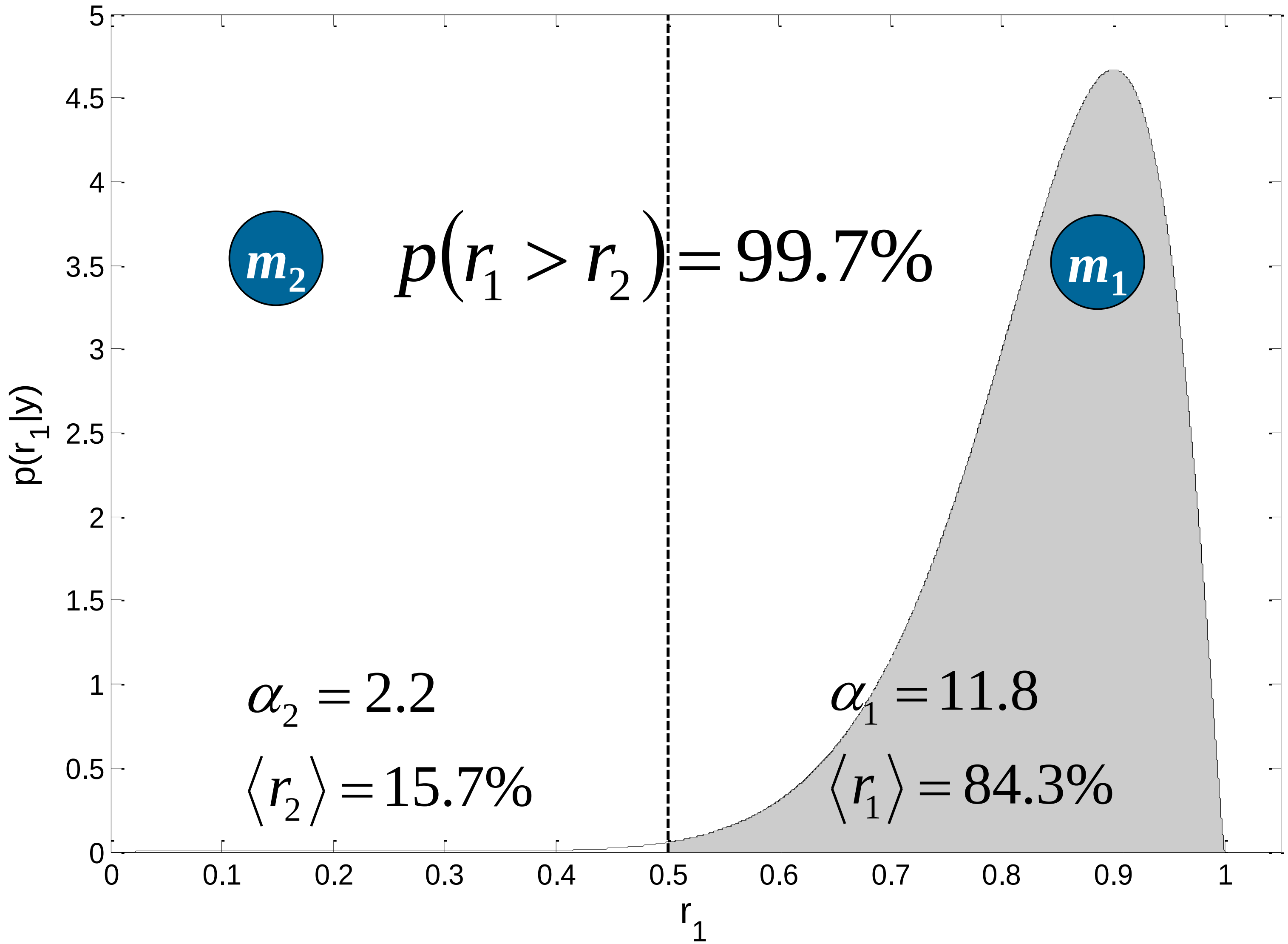
$$\exists k \in \{1...K\}, \forall j \in \{1...K \mid j \neq k\}:$$
$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

4. **protected exceedance probability**: see below

Example: Hemispheric interactions during vision



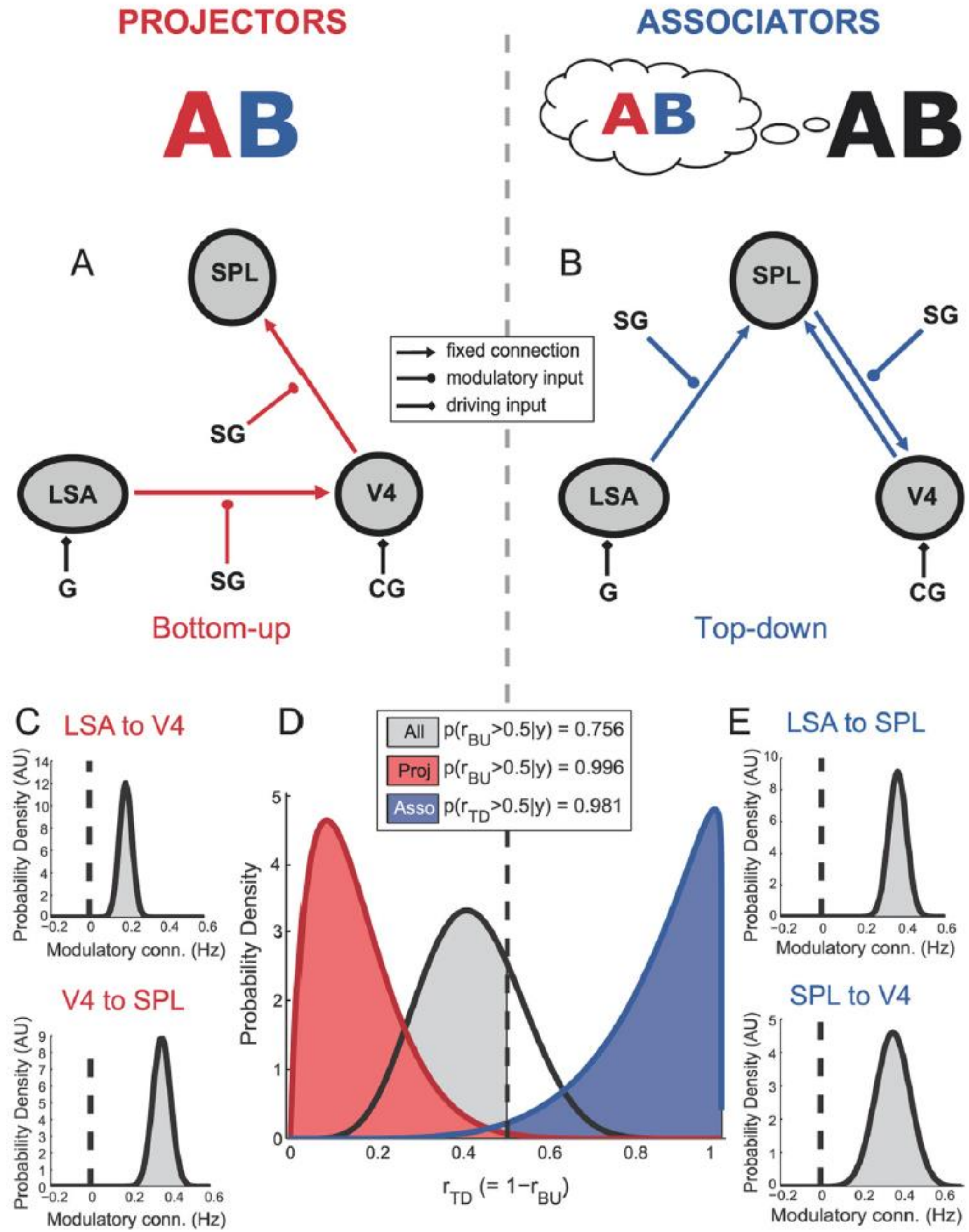
Data: Stephan et al. 2003, *Science*
Models: Stephan et al. 2007, *J. Neurosci.*



Stephan et al. 2009a, *NeuroImage*

Example: Synaesthesia

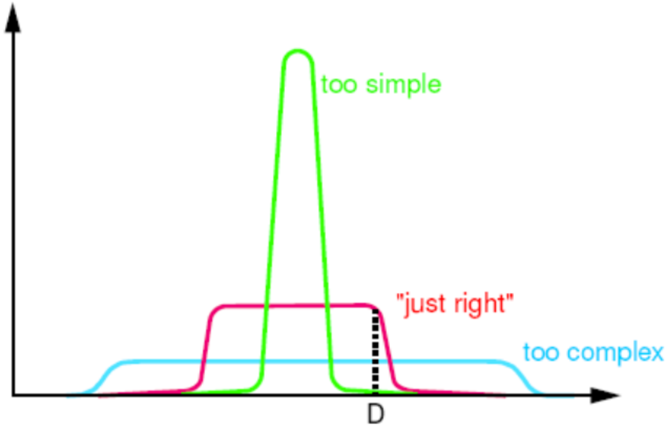
- "projectors" experience color externally colocalized with a presented grapheme
- "associators" report an internally evoked association
- across all subjects: no evidence for either model
- but BMS results map precisely onto projectors (bottom-up mechanisms) and associators (top-down)



van Leeuwen et al. 2011, *J. Neurosci.*

Overfitting at the level of models

- $\uparrow$ #models $\Rightarrow \uparrow$ risk of overfitting
- solutions:
  - regularisation: definition of model space = choosing priors p(m)
  - family-level BMS
  - Bayesian model averaging (BMA)



posterior model probability:

$$p\left(m \mid y\right) = \frac{p\left(y \mid m\right) p(m)}{\sum_m p\left(y \mid m\right) p(m)}$$

BMA:

$$p\left(\theta \mid y\right) = \sum_m p\left(\theta \mid y, m\right) p\left(m \mid y\right)$$

Model space partitioning: comparing model families

- partitioning model space into K subsets or families:

$$M = \left\{ f_1, \ldots, f_K \right\}$$

- pooling information over all models in these subsets allows one to compute the probability of a model family, given the data

$$p\left(f_k \right)$$

- effectively removes uncertainty about any aspect of model structure, other than the attribute of interest (which defines the partition)

Stephan et al. 2009, *NeuroImage*
Penny et al. 2010, *PLoS Comput. Biol.*

Family-level inference: random effects – a special case

- When the families are of equal size, one can simply sum the posterior model probabilities within families by exploiting the agglomerative property of the Dirichlet distribution:

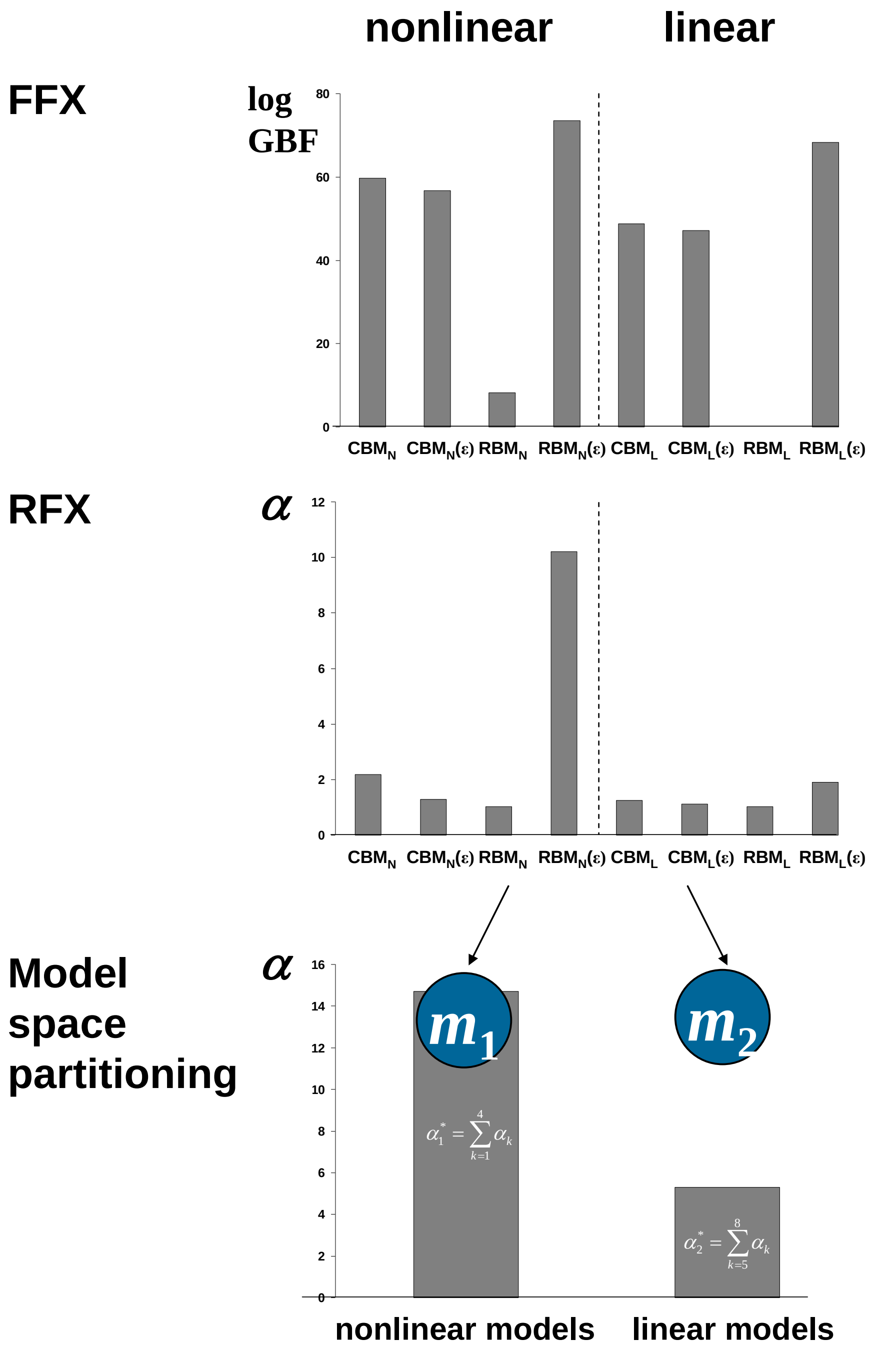
$$\left(r_1, r_2, \ldots, r_K\right) \sim Dir\left(\alpha_1, \alpha_2, \ldots, \alpha_K\right)$$

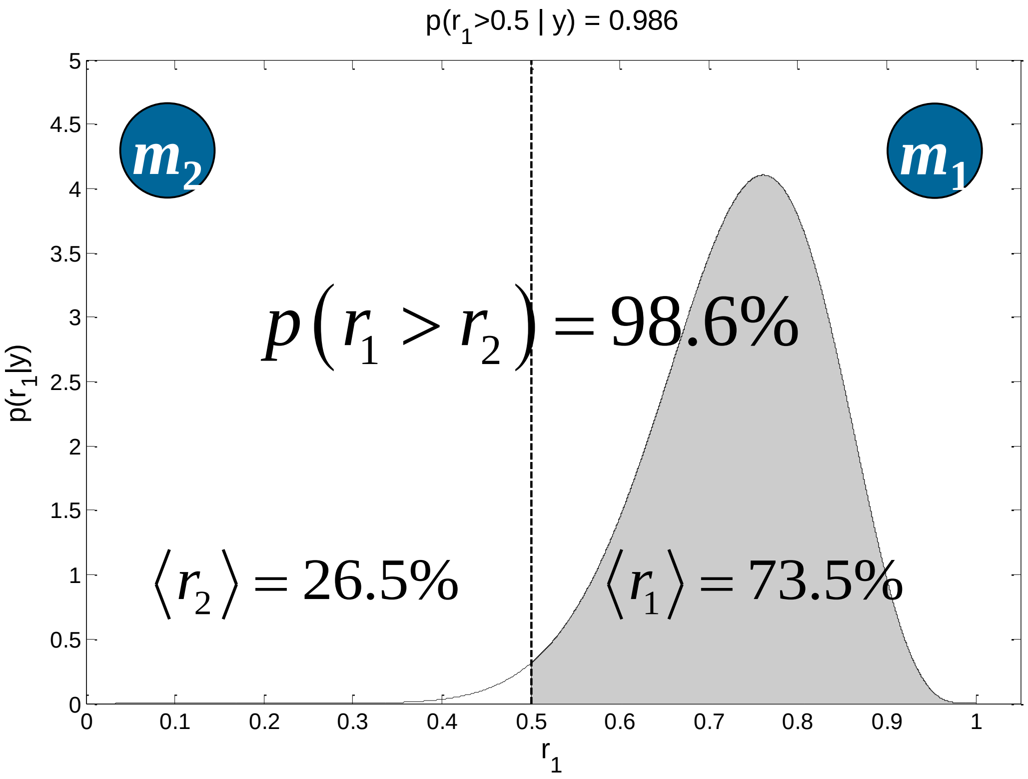
$$\Rightarrow r_1^* = \sum_{k \in N_1} r_k, r_2^* = \sum_{k \in N_2} r_k, \ldots, r_J^* = \sum_{k \in N_J} r_k$$

$$\sim Dir\left(\alpha_1^* = \sum_{k \in N_1} \alpha_k, \alpha_2^* = \sum_{k \in N_2} \alpha_k, \ldots, \alpha_J^* = \sum_{k \in N_J} \alpha_k\right)$$

Stephan et al. 2009, *NeuroImage*

Model space partitioning: comparing model families





Stephan et al. 2009, *NeuroImage*

Bayesian Model Averaging (BMA)

- abandons dependence of parameter inference on a single model and takes into account model uncertainty
- uses the entire model space considered (or an optimal family of models)
- averages parameter estimates, weighted by posterior model probabilities
- represents a particularly useful alternative
  - when none of the models (or model subspaces) considered clearly outperforms all others
  - when comparing groups for which the optimal model differs

single-subject BMA:

$$p(\theta \mid y) = \sum_m p(\theta \mid y, m) p(m \mid y)$$

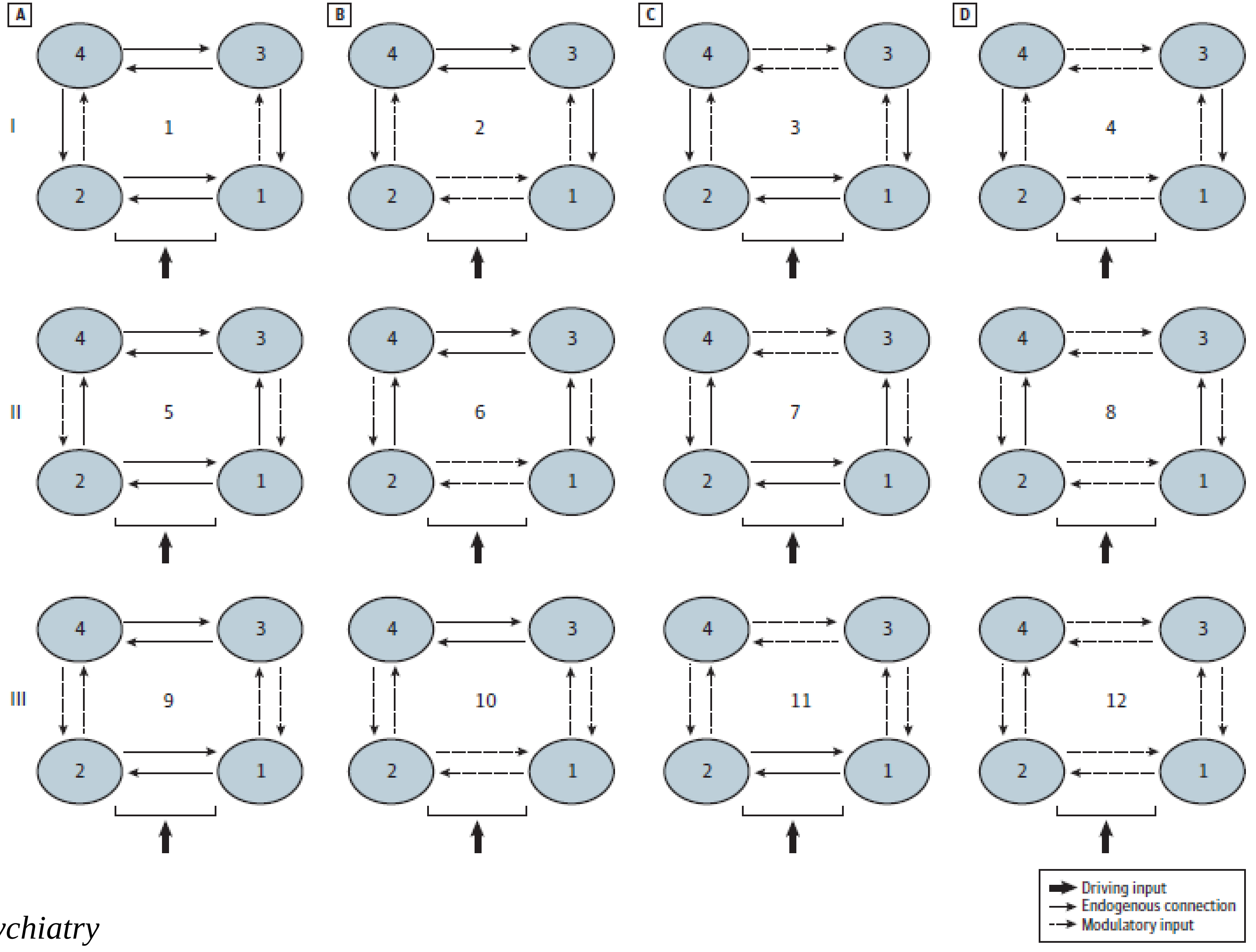
group-level BMA:

$$p(\theta_n \mid y_{1..N}) = \sum_m p(\theta_n \mid y_n, m) p(m \mid y_{1..N})$$

NB: $p(m|y_{1..N})$ can be obtained by either FFX or RFX BMS

Penny et al. 2010, *PLoS Comput. Biol.*

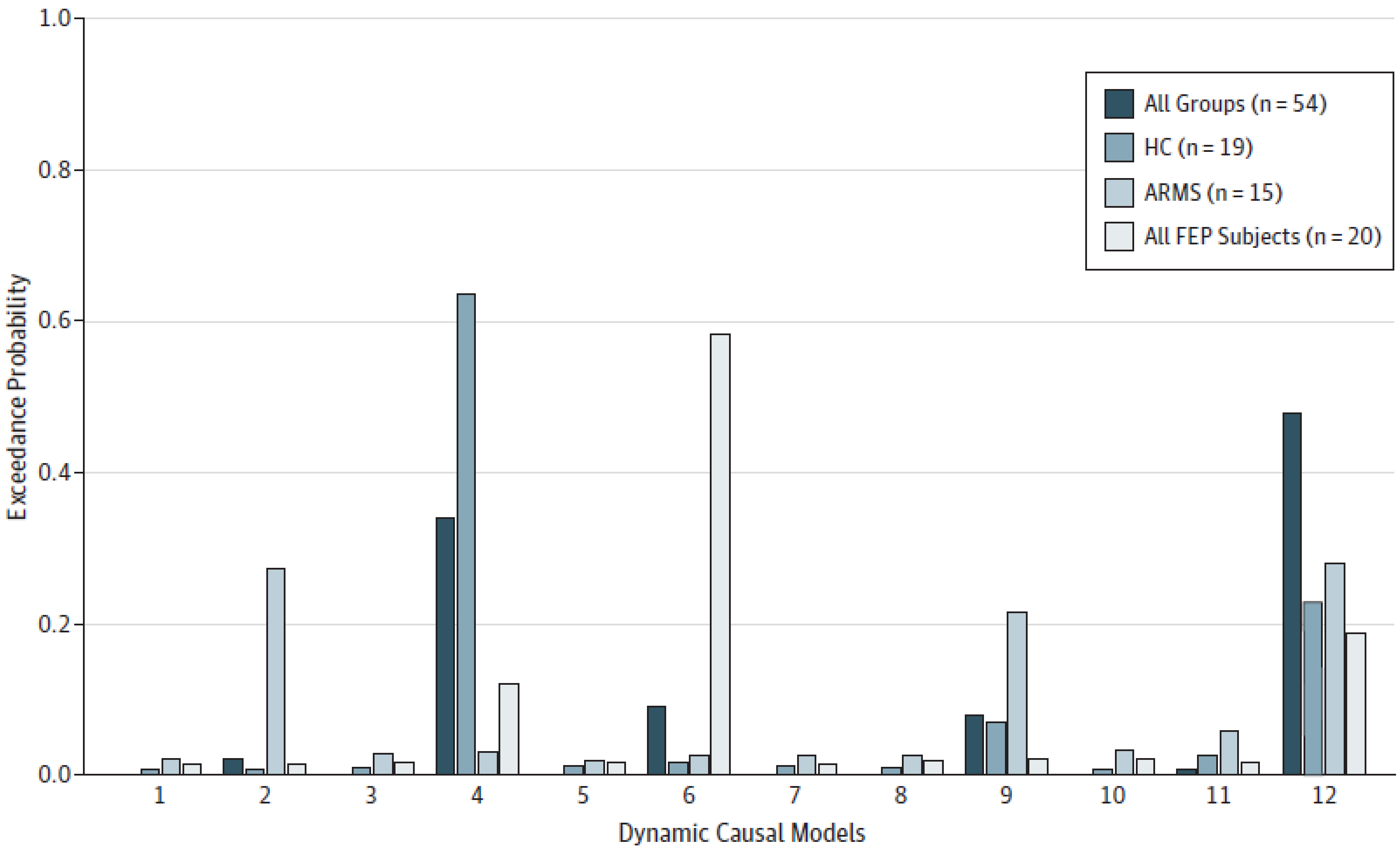
Prefrontal-parietal connectivity during working memory in schizophrenia



- 17 at-risk mental state (ARMS) individuals
- 21 first-episode patients (13 non-treated)
- 20 controls

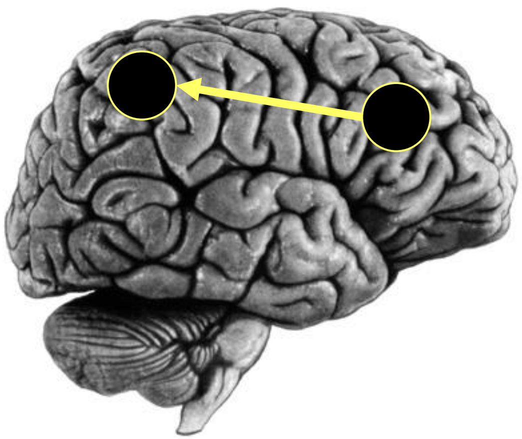
Schmidt et al. 2013, *JAMA Psychiatry*

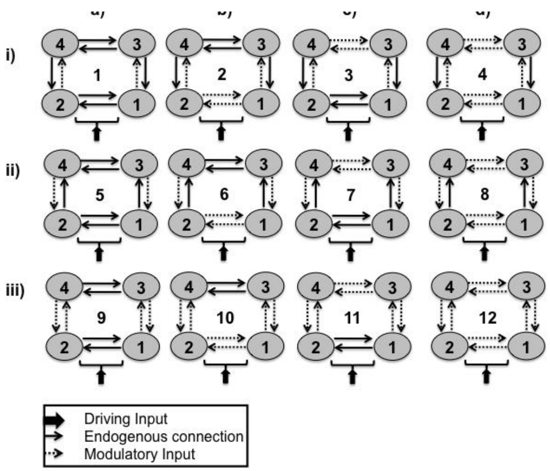
BMS results for all groups

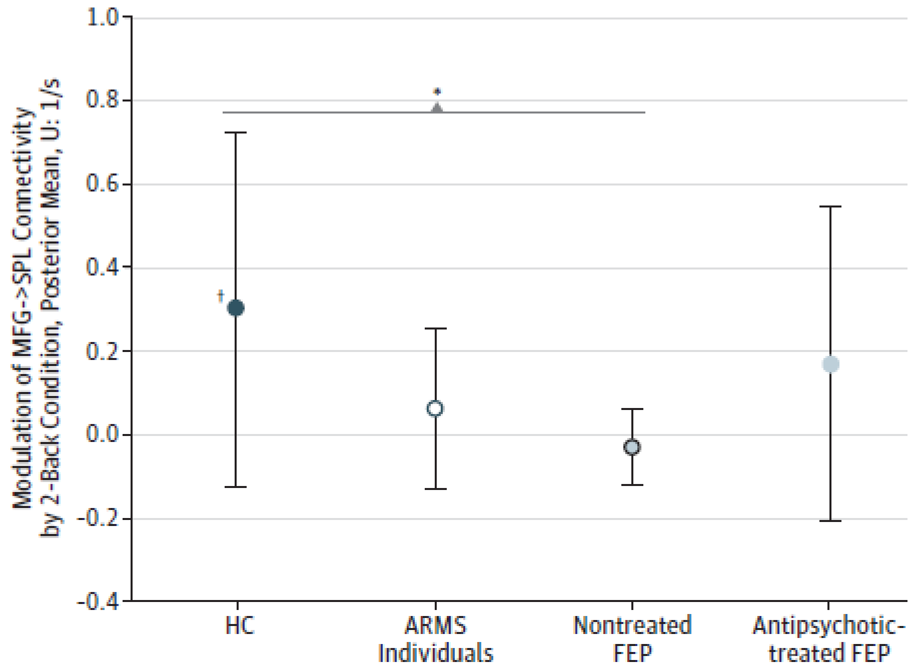


Schmidt et al. 2013, *JAMA Psychiatry*

BMA results: PFC → PPC connectivity







17 ARMS, 21 first-episode (13 non-treated), 20 controls

Schmidt et al. 2013, *JAMA Psychiatry*

Protected exceedance probability: Using BMA to protect against chance findings

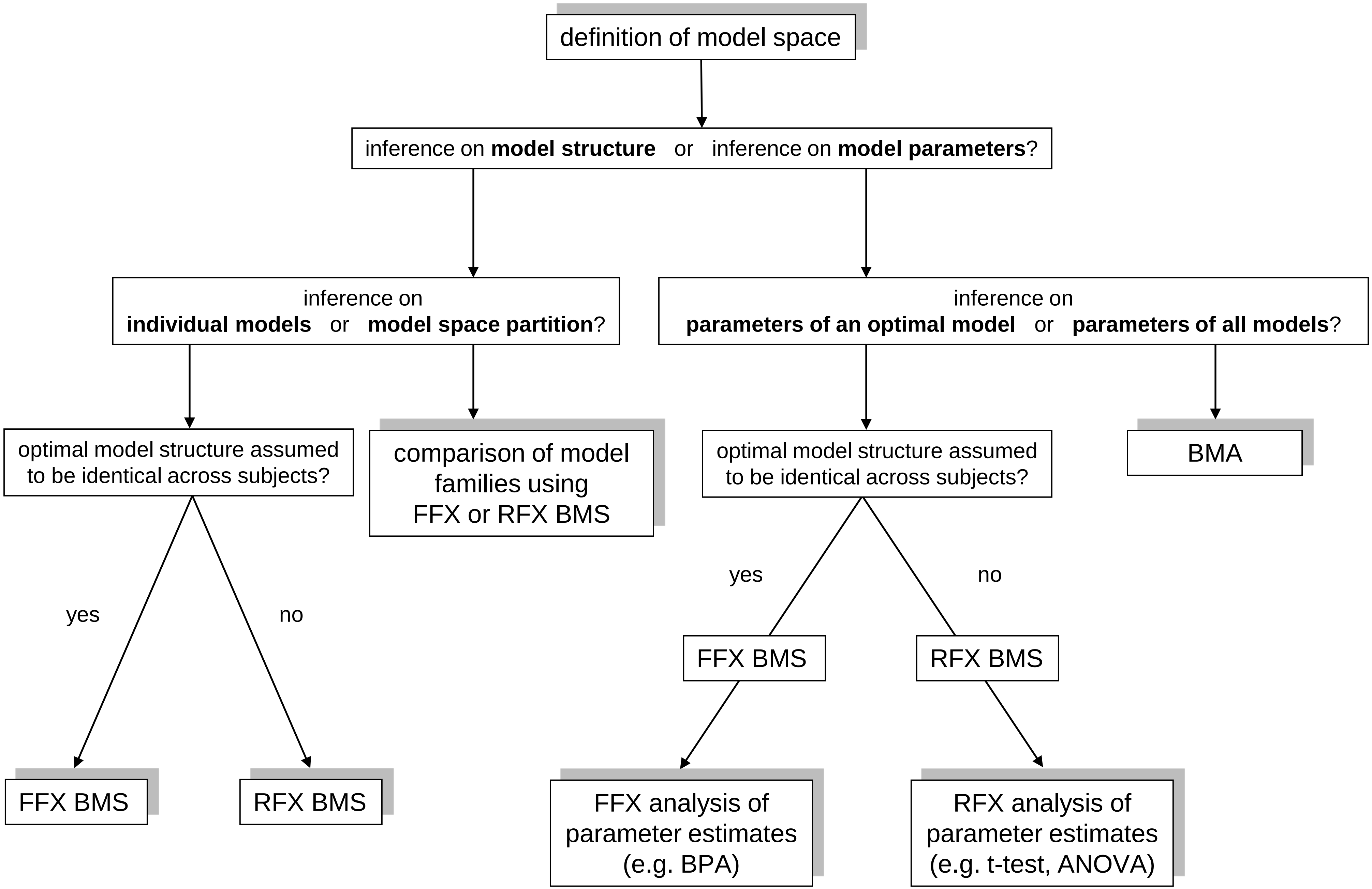
- EPs express our confidence that the posterior probabilities of models are different – under the hypothesis H_1 that models differ in probability: $r_k \neq 1/K$
- does not account for possibility "null hypothesis" H_0: $r_k = 1/K$
- **Bayesian omnibus risk (BOR)** of wrongly accepting H_1 over H_0:

$$P_o = \frac{1}{1 + \frac{p(m|H_1)}{p(m|H_0)}}.$$

- **protected EP**: Bayesian model averaging over H_0 and H_1:

$$\begin{aligned} \tilde{\varphi}_k &= P(r_k \geq r_{k' \neq k}|y) \\ &= P(r_k \geq r_{k' \neq k}|y, H_1)P(H_1|y) + P(r_k \geq r_{k' \neq k}|y, H_0)P(H_0|y) \\ &= \varphi_k(1 - P_0) + \frac{1}{K}P_0 \end{aligned}$$

Rigoux et al. 2014, *NeuroImage*



Stephan et al. 2010, *NeuroImage*

Further reading

- Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. NeuroImage 22:1157-1172.
- Penny WD, Stephan KE, Daunizeau J, Joao M, Friston K, Schofield T, Leff AP (2010) Comparing Families of Dynamic Causal Models. PLoS Computational Biology 6: e1000709.
- Penny WD (2012) Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59: 319-330.
- Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies – revisited. NeuroImage 84: 971-985.
- Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007) Comparing hemodynamic models with DCM. NeuroImage 38:387-401.
- Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. NeuroImage 46:1004-1017.
- Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ (2010) Ten simple rules for Dynamic Causal Modelling. NeuroImage 49: 3099-3109.
- Stephan KE, Iglesias S, Heinzle J, Diaconescu AO (2015) Translational Perspectives for Computational Neuroimaging. Neuron 87: 716-732.
- Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, Rigoux L, Moran RJ, Daunizeau J, Dolan RJ, Friston KJ, Heinz A (2017) Computational Neuroimaging Strategies for Single Patient Predictions. NeuroImage 145: 180-199.

Thank you