Bayesian inference and generative models

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, 22 November 2016

Slides with a yellow title were not covered in detail in the lecture and will not be part of the exam.

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
- contemporary **theories of brain function** (the "Bayesian brain")
  - predictive coding
  - free energy principle
  - active inference

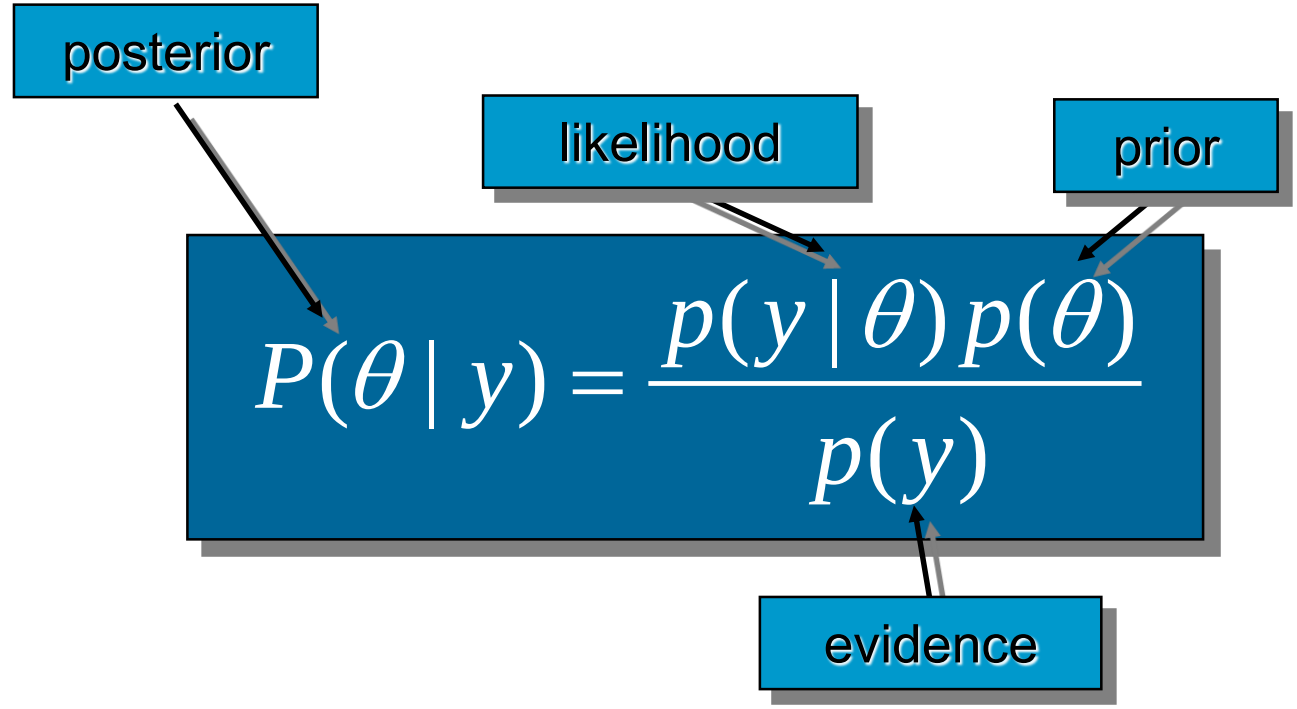
Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

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

Bayes‘ Theorem

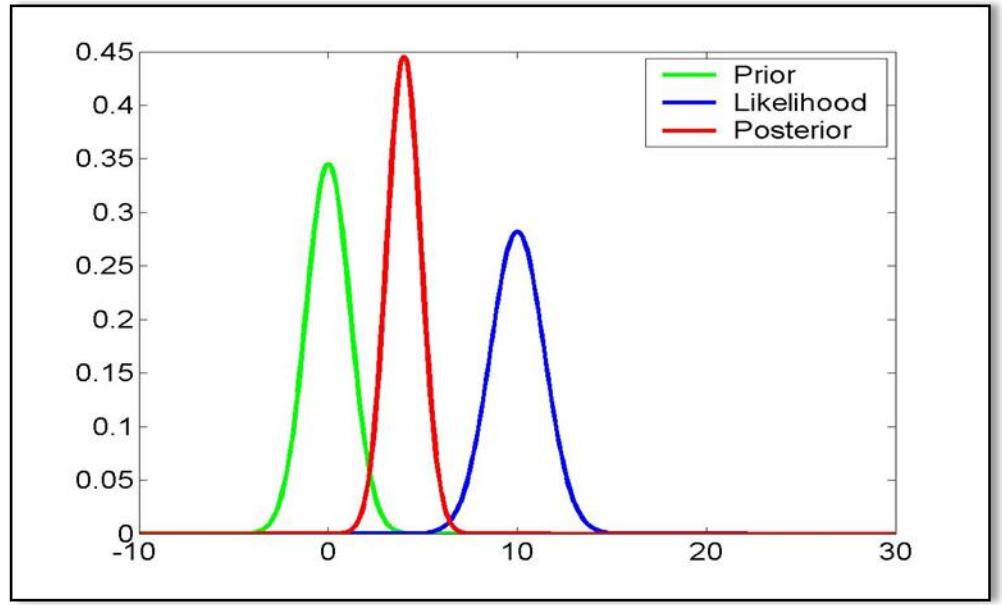
Reverend Thomas Bayes
1702 - 1761



“Bayes‘ Theorem describes, how an ideally rational person processes information."

Wikipedia

Bayesian inference: an animation



Generative models

- specify a joint probability distribution over all variables (observations and parameters)

- require a likelihood function and a prior:

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

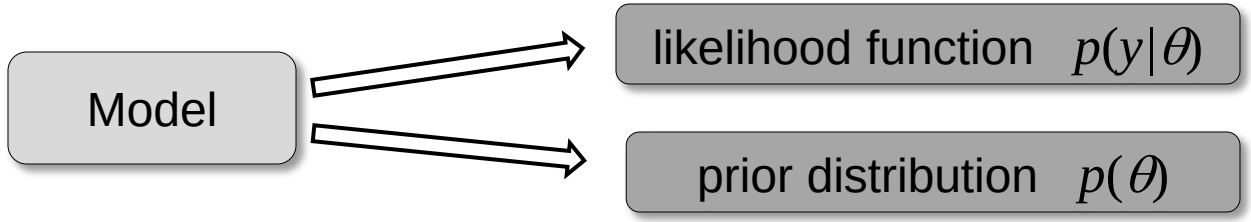
- can be used to randomly generate synthetic data (observations) by sampling from the prior
  - we can check in advance whether the model can explain certain phenomena at all

- model comparison based on the model evidence

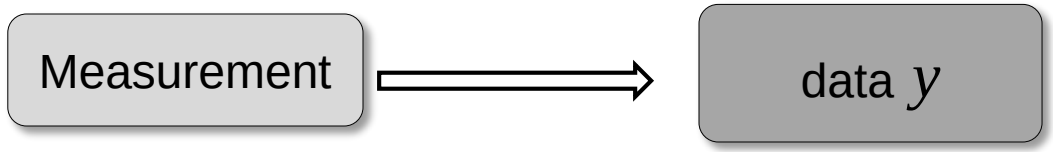
$$p(y \mid m) = \int p(y \mid \theta, m)\, p(\theta \mid m) d\theta$$

Principles of Bayesian inference

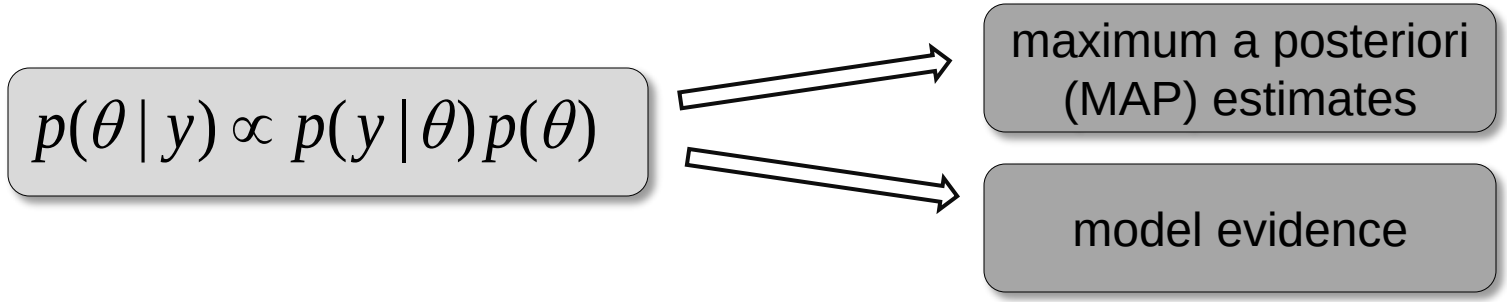
⇨ Formulation of a **generative model**



⇨ Observation of **data**



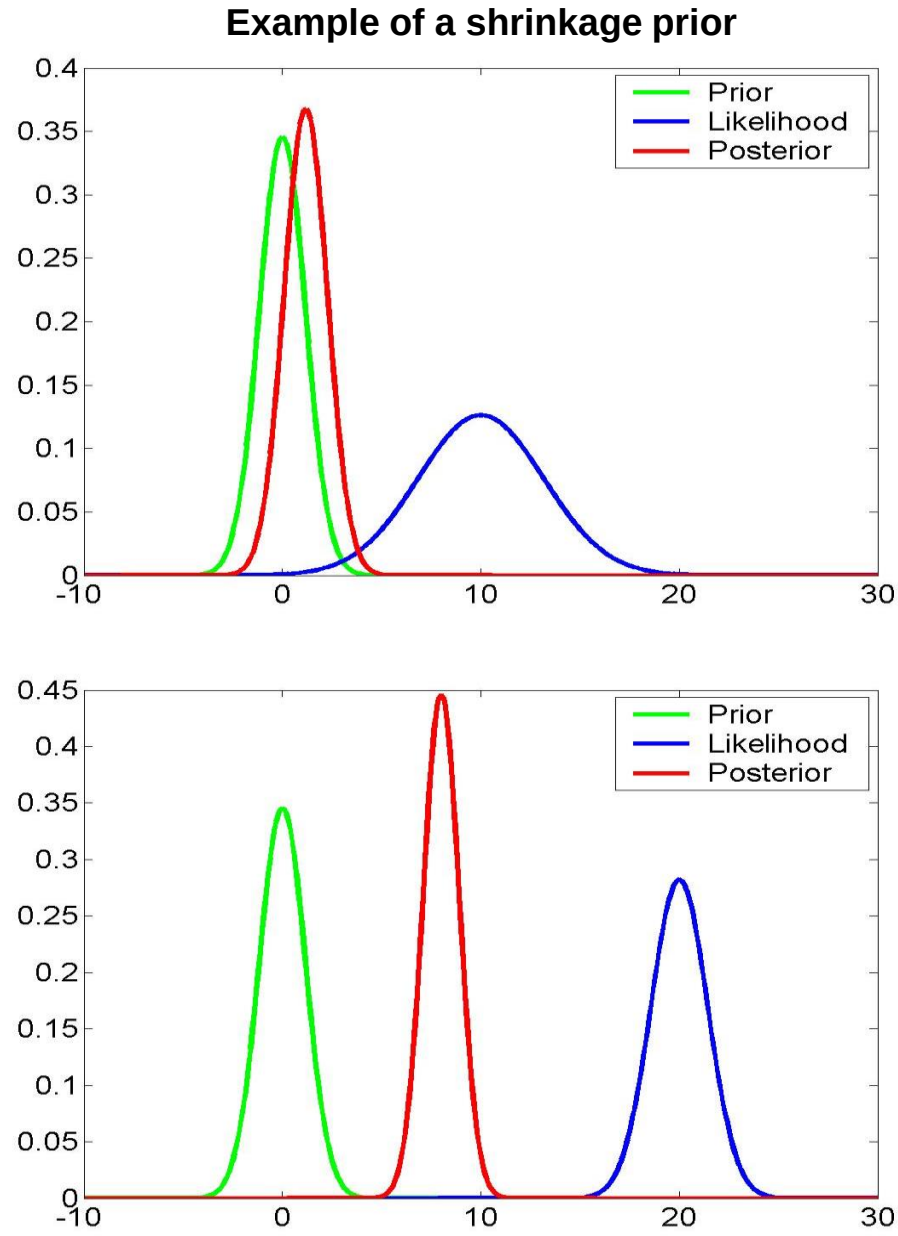
⇨ **Model inversion** – updating one's beliefs



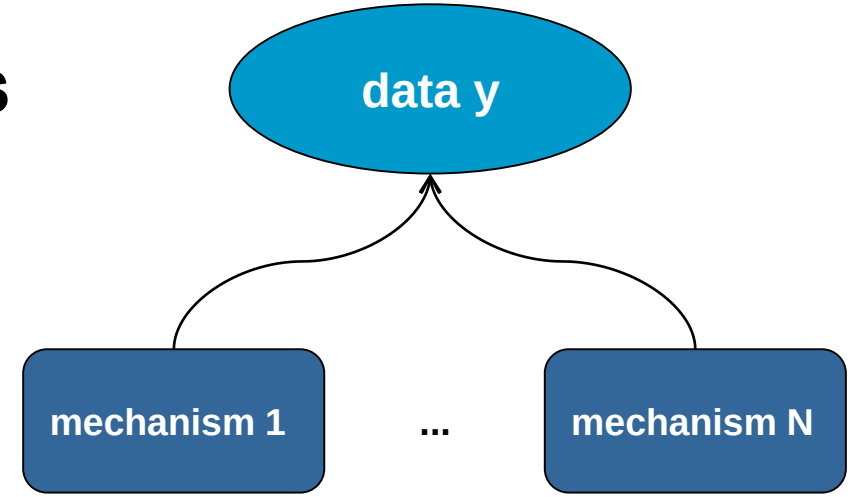
Priors

Priors can be of different sorts, e.g.

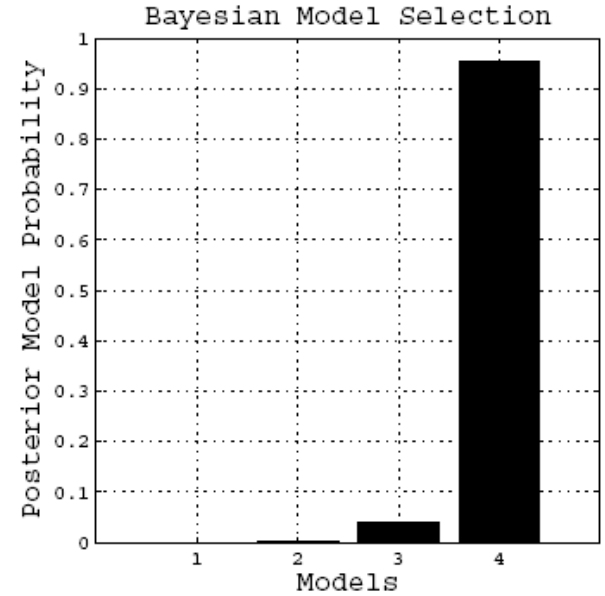
- empirical (previous data)
- empirical (estimated from current data using a hierarchical model → "empirical Bayes")
- uninformed
- principled (e.g., positivity constraints)
- shrinkage



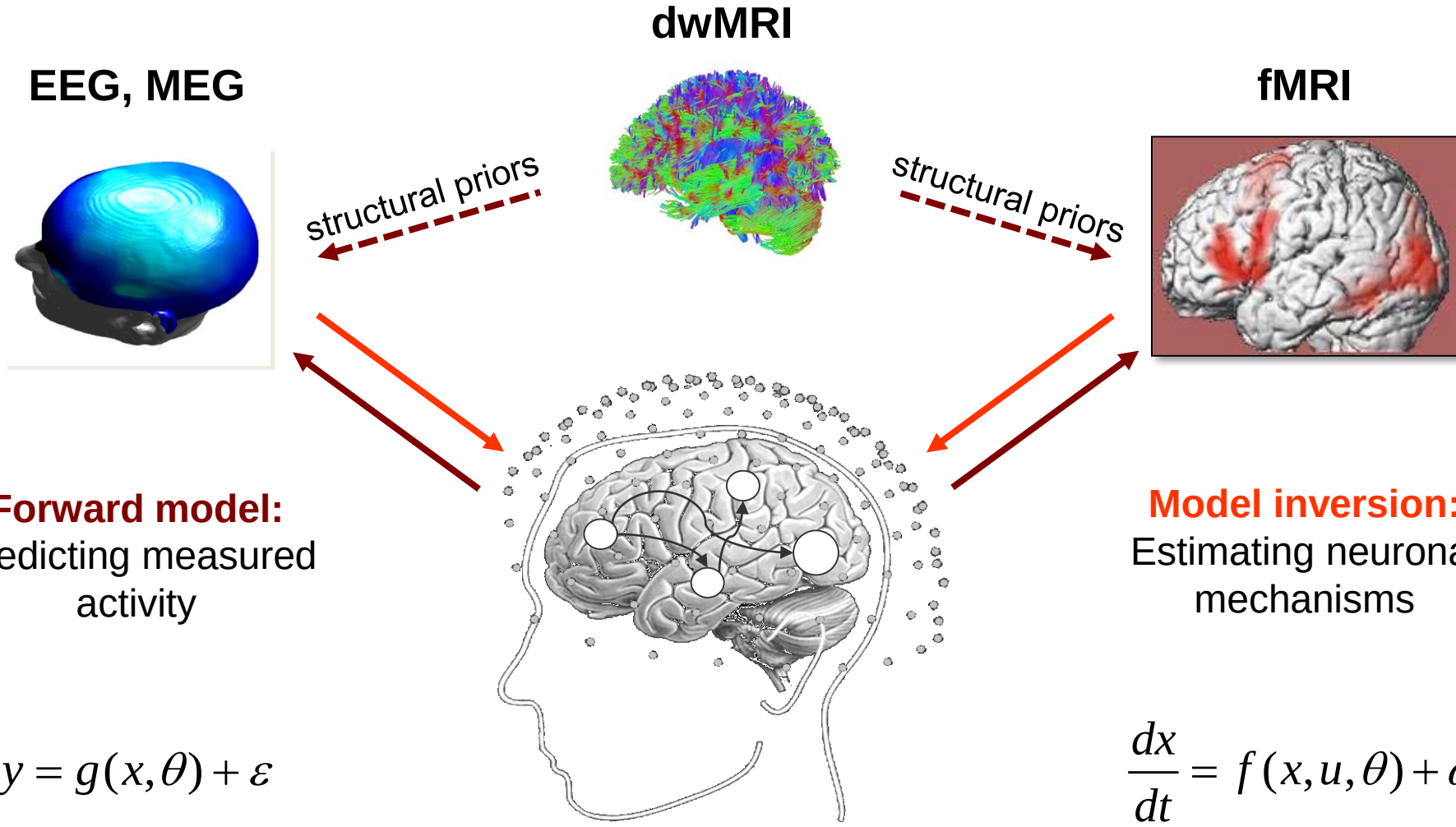
Advantages of generative models



- describe how observed data were generated by hidden **mechanisms**
- we can **check in advance** whether a model can explain certain phenomena at all
- force us to think mechanistically and be **explicit** about pathophysiological theories
- **formal framework for differential diagnosis:**
  statistical comparison of competing generative models, each of which provides a different explanation of measured brain activity or clinical symptoms



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



Forward model:
Predicting measured activity

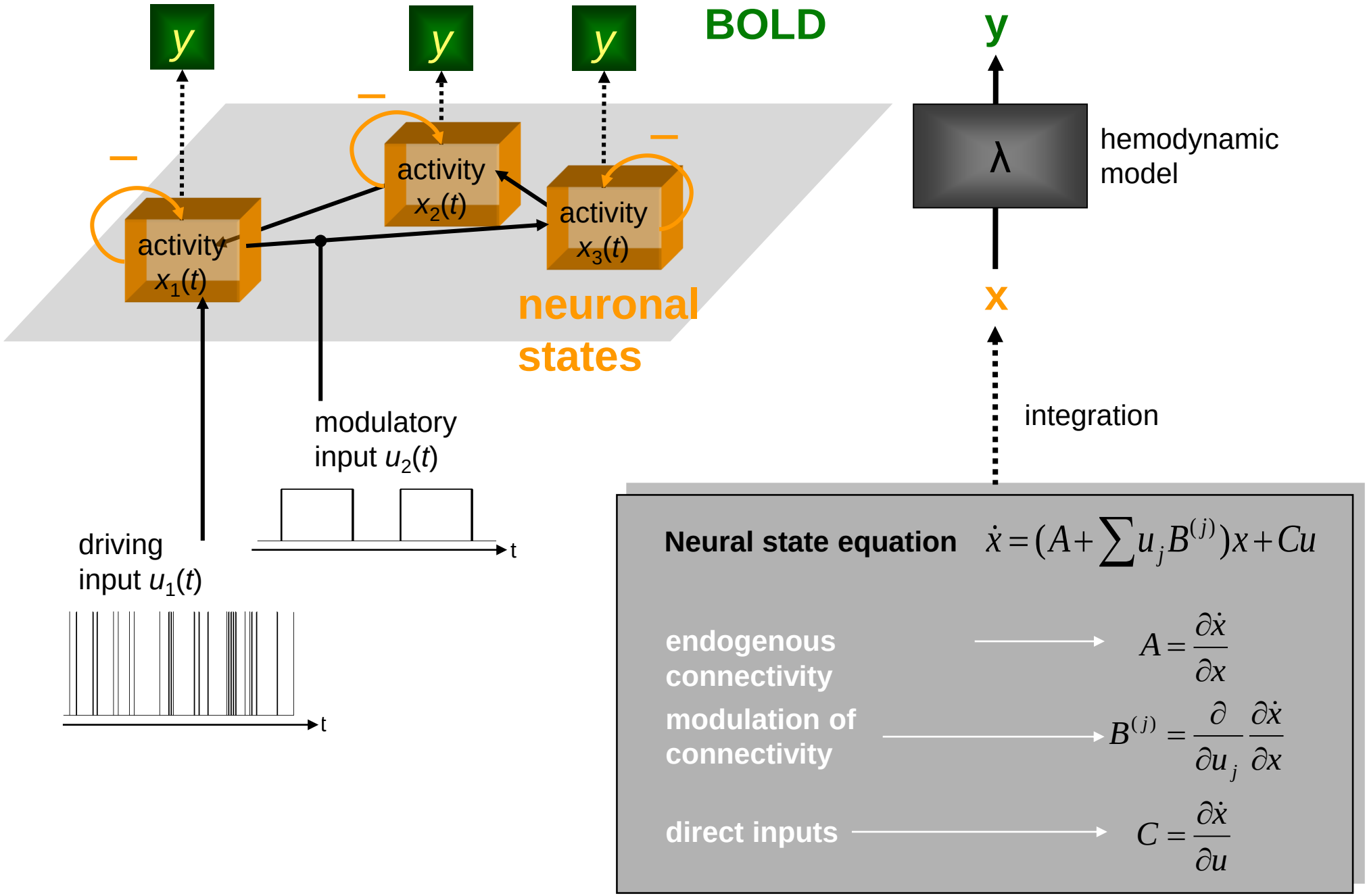
$$y = g(x,\theta) + \varepsilon$$

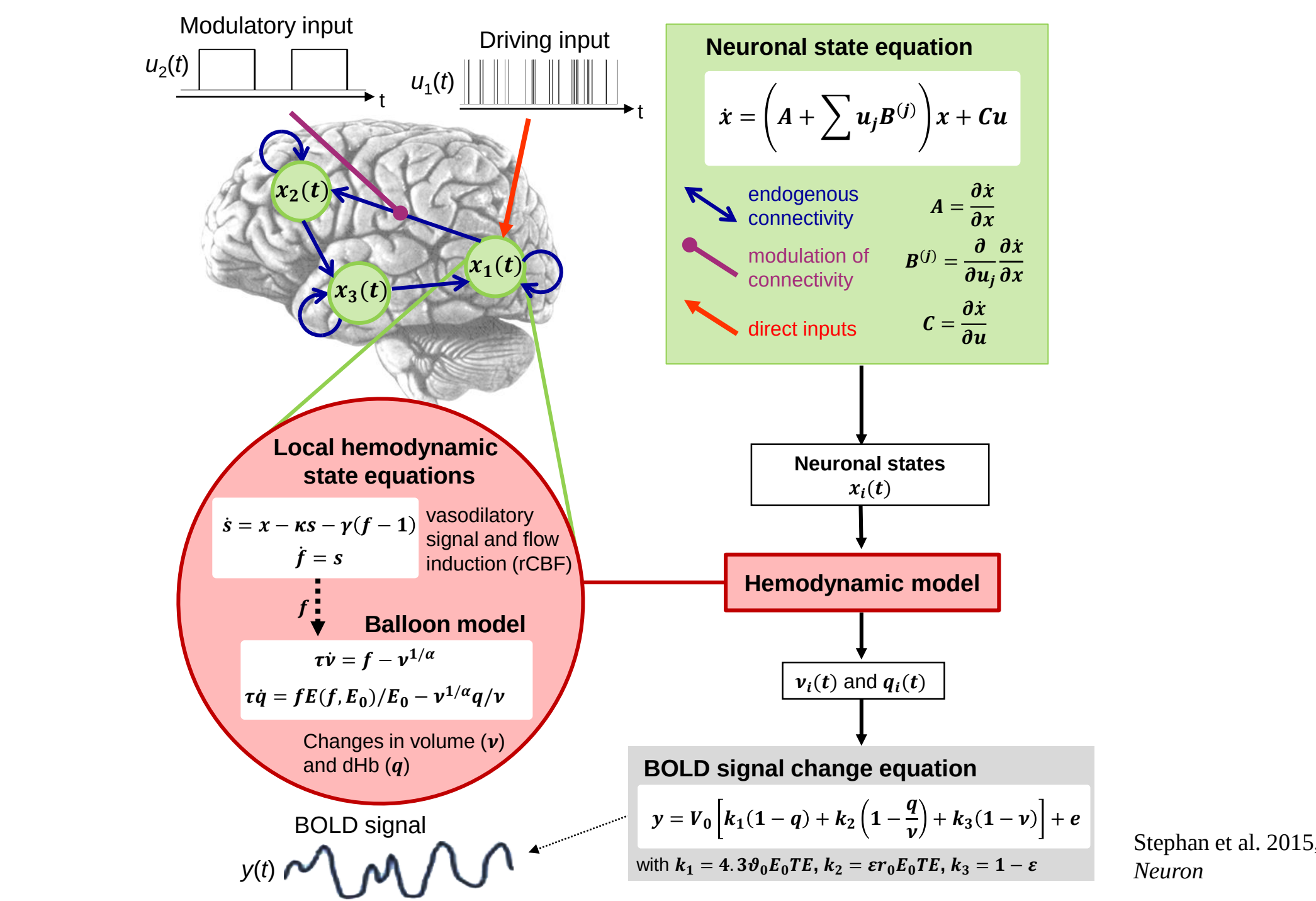
Model inversion:
Estimating neuronal mechanisms

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

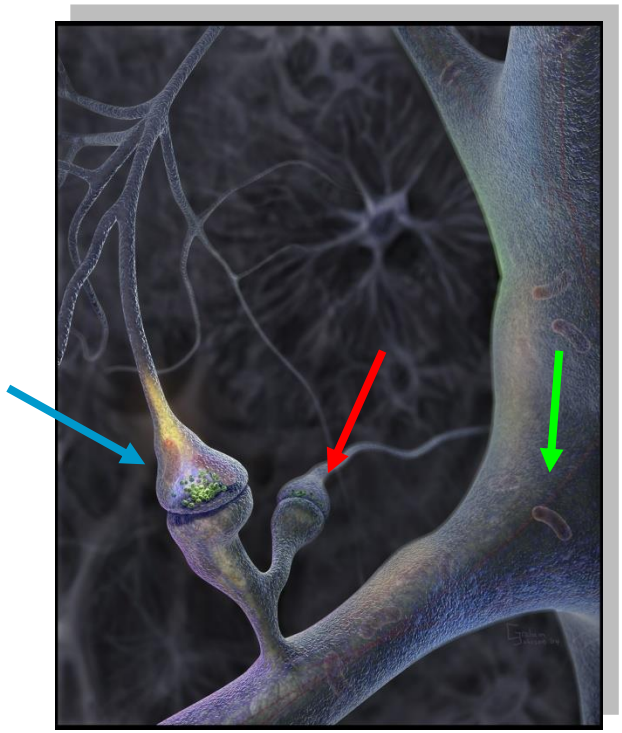
Friston et al. 2003, *NeuroImage*

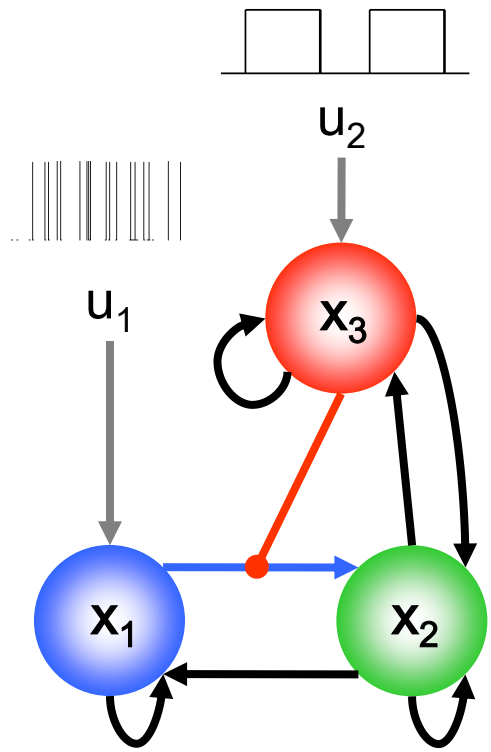
Stephan et al. 2009, *NeuroImage*

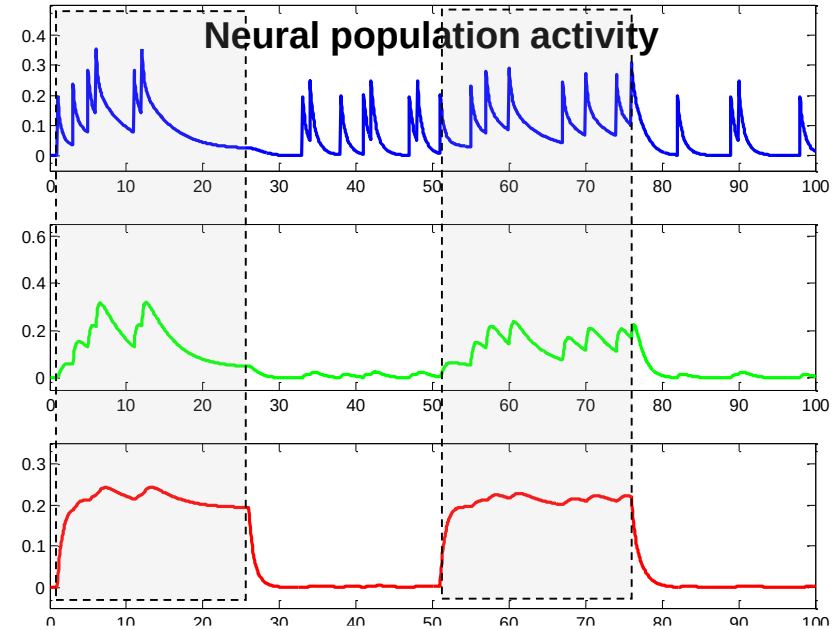

y
y
y
BOLD
activity
$x_1(t)$
activity
$x_2(t)$
activity
$x_3(t)$
neuronal
states
y
λ
hemodynamic
model
x
integration
modulatory
input $u_2(t)$
t
driving
input $u_1(t)$
t
Neural state equation $\dot{x} = (A + \sum u_j B^{(j)})x + Cu$
endogenous
connectivity
$A = \frac{\partial \dot{x}}{\partial x}$
modulation of
connectivity
$B^{(j)} = \frac{\partial}{\partial u_j}\frac{\partial \dot{x}}{\partial x}$
direct inputs
$C = \frac{\partial \dot{x}}{\partial u}$

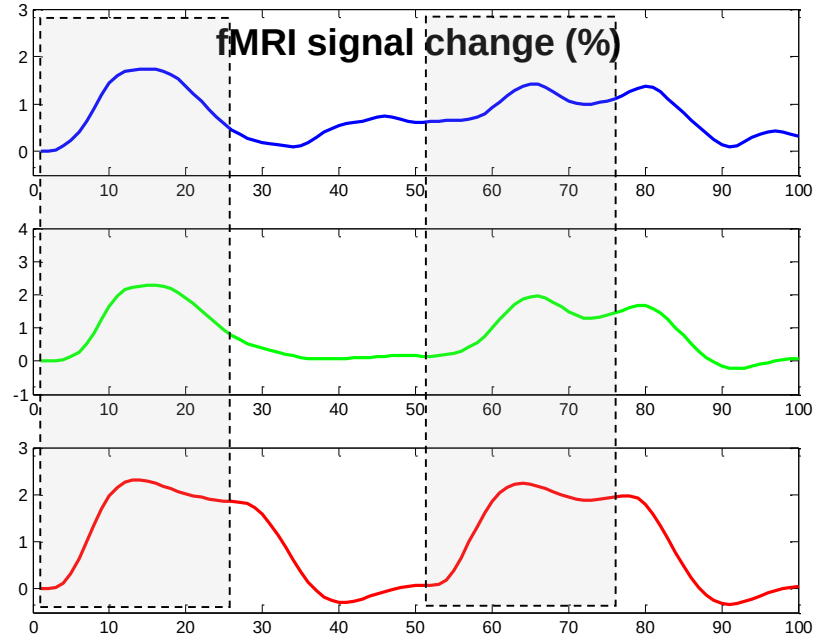


Stephan et al. 2015, *Neuron*









Nonlinear Dynamic Causal Model for fMRI

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

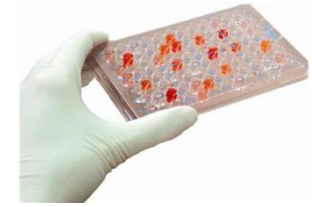
Stephan et al. 2008, *NeuroImage*

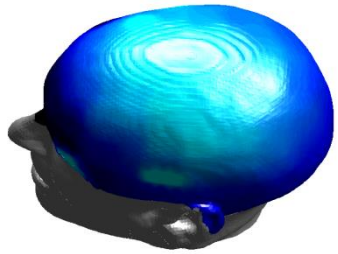
Why should I know about Bayesian inference?

Because Bayesian principles are fundamental for

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

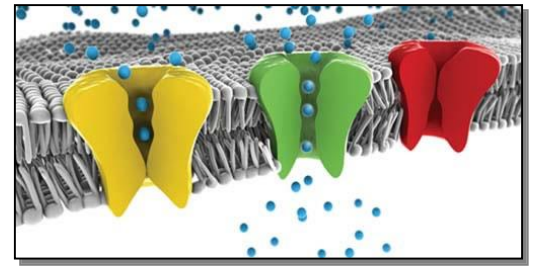
Generative models as "computational assays"

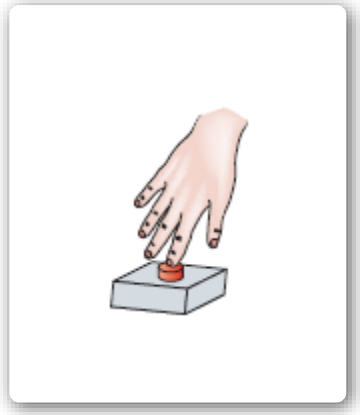




$$p(y \mid \theta, m) \cdot p(\theta \mid m)$$

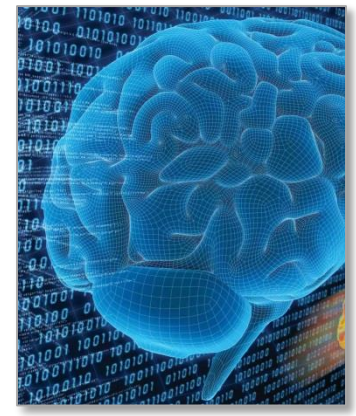
$$p(\theta \mid y, m)$$





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

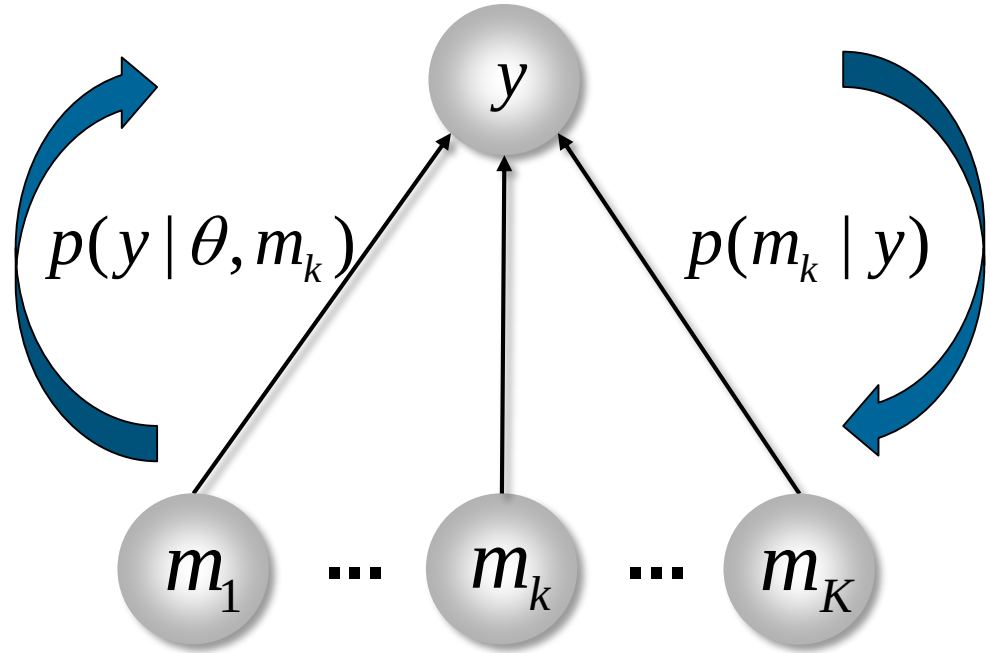
$$p(\theta \mid y, m)$$



Differential diagnosis based on generative models of disease symptoms

SYMPTOM
(behaviour
or physiology)

HYPOTHETICAL MECHANISM

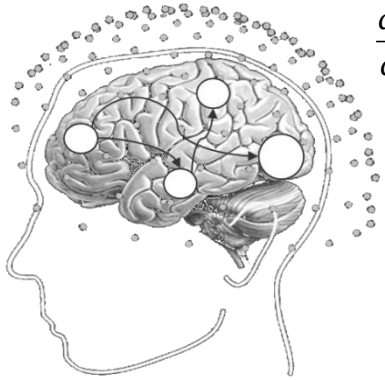


$$p(m_k \mid \mathrm{y}) = \frac{p(y \mid m_k)\, p(m_k)}{\sum_k p(y \mid m_k)\, p(m_k)}$$

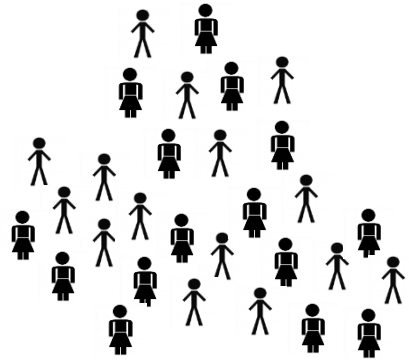
Stephan et al. 2016, *NeuroImage*

Translational Neuromodeling

❶ Computational assays: Models of disease mechanisms

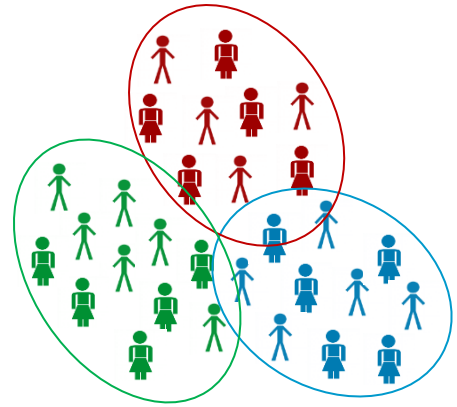


$$\frac{dx}{dt} = f(x,u,\theta) + \omega$$



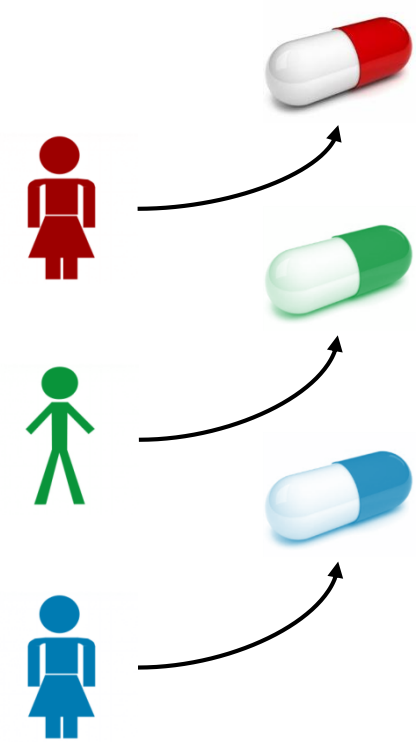
❷ Application to brain activity and behaviour of individual patients

❸ Detecting physiological subgroups (based on inferred mechanisms)



- disease mechanism A
- disease mechanism B
- disease mechanism C

❹ Individual treatment prediction

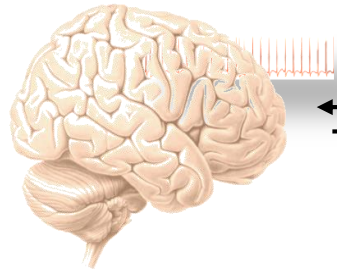


Stephan et al. 2015, *Neuron*

Perception = inversion of a hierarchical generative model

neuronal states

environm. states
others' mental states
bodily states



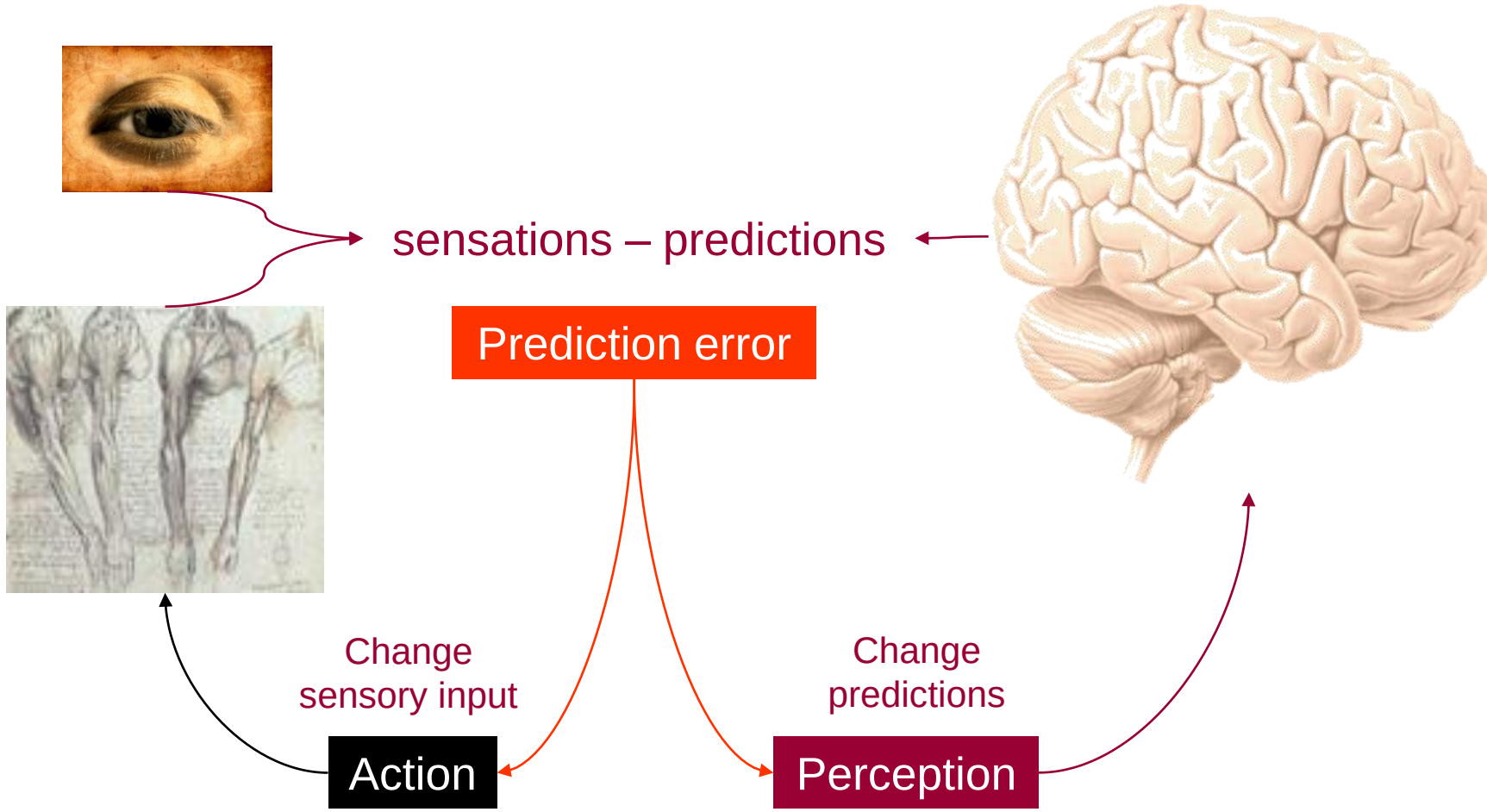
forward model

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

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

perception

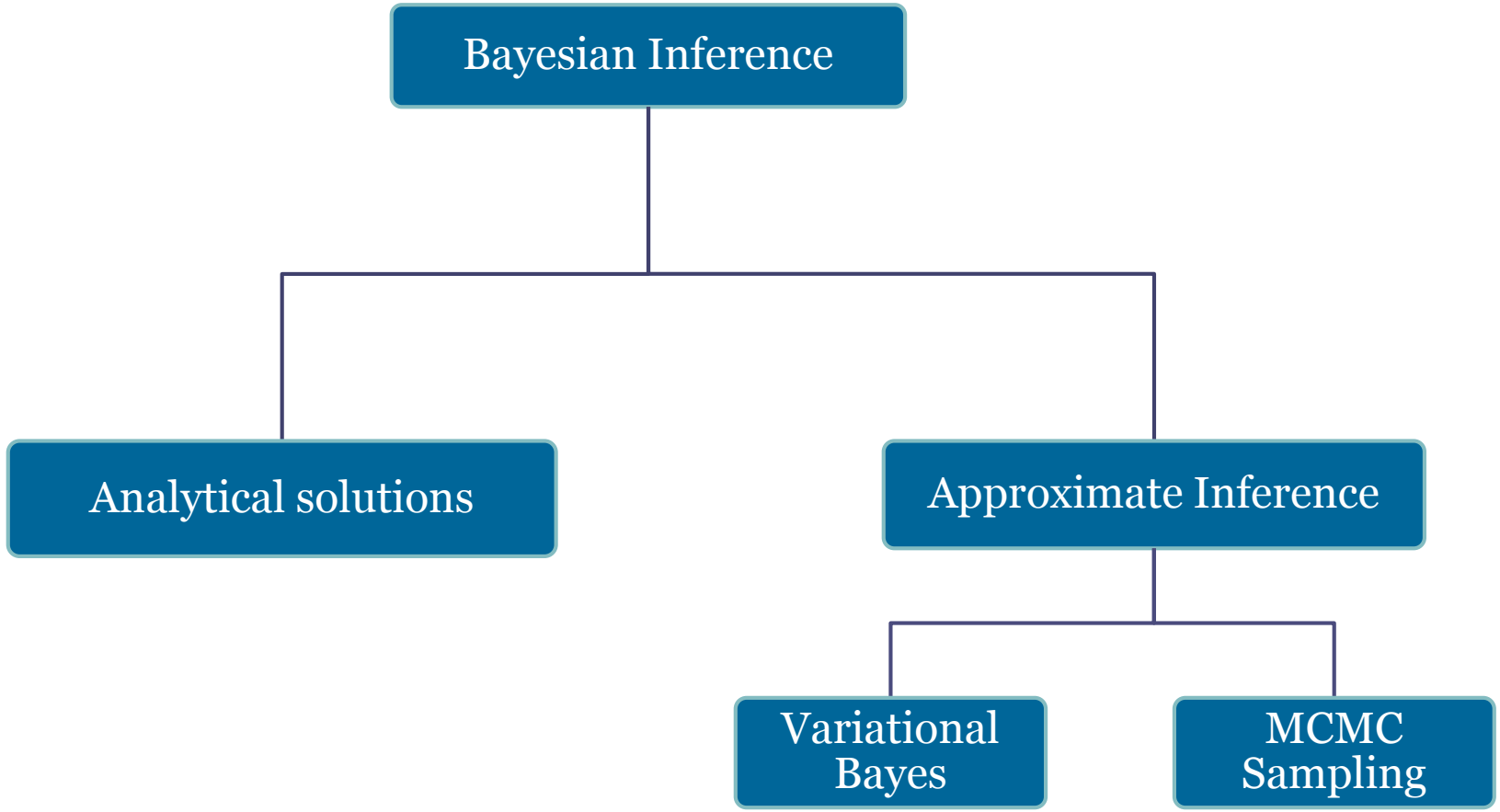
Example: free-energy principle and active inference



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?

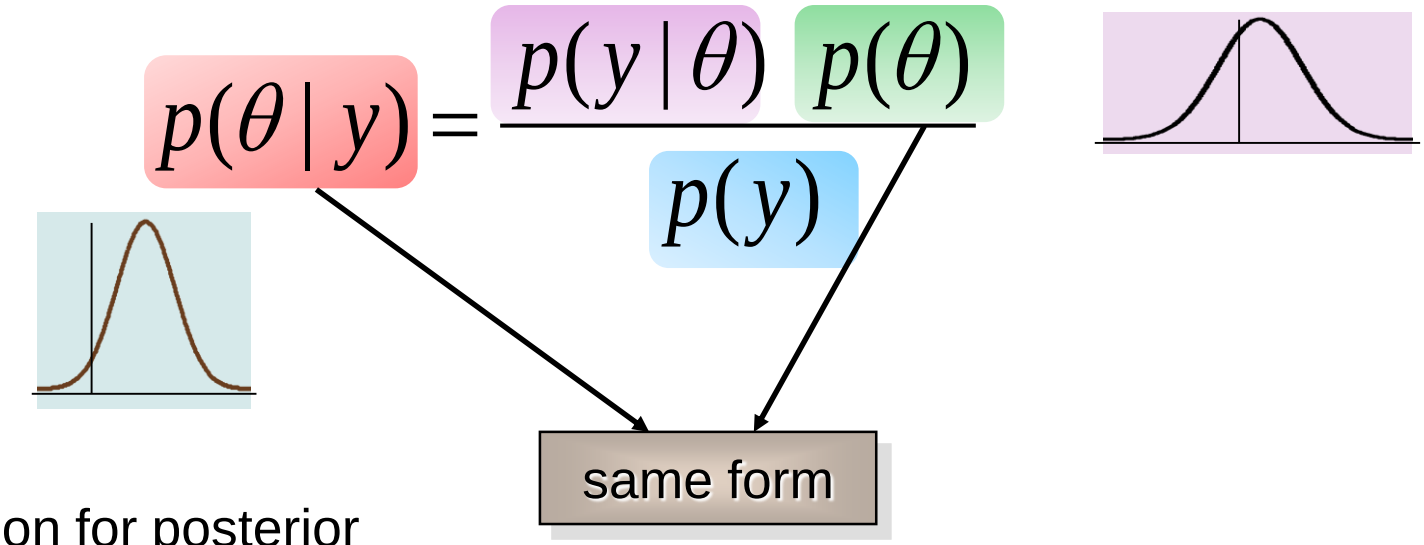
- **compute the posterior analytically**
  - requires conjugate priors
  - even then often difficult to derive an analytical solution

- **variational Bayes (VB)**
  - often hard work to derive, but fast to compute
  - cave: local minima, potentially inaccurate approximations

- **sampling methods (MCMC)**
  - guaranteed to be accurate in theory (for infinite computation time)
  - but may require very long run time in practice
  - convergence difficult to prove

Conjugate priors

If the posterior p(θ|x) is in the same family as the prior p(θ), the prior and posterior are called "conjugate distributions", and the prior is called a "conjugate prior" for the likelihood function.



- ⇨ analytical expression for posterior
- ⇨ examples (likelihood-prior):
  - Normal-Normal
  - Normal-inverse Gamma
  - Binomial-Beta
  - Multinomial-Dirichlet

Posterior mean & variance of univariate Gaussians

$$y = \theta + \varepsilon$$

Likelihood & Prior

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

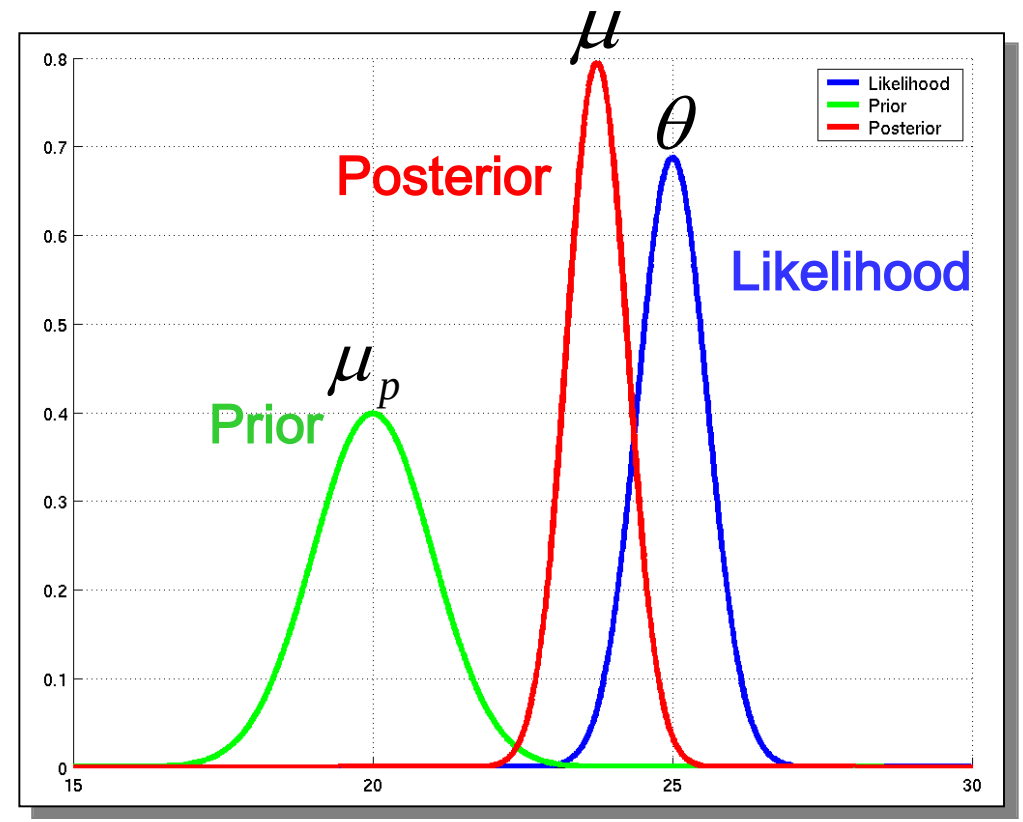
$$p(\theta) = N(\mu_p, \sigma_p^2)$$

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

$$y = \theta + \varepsilon$$

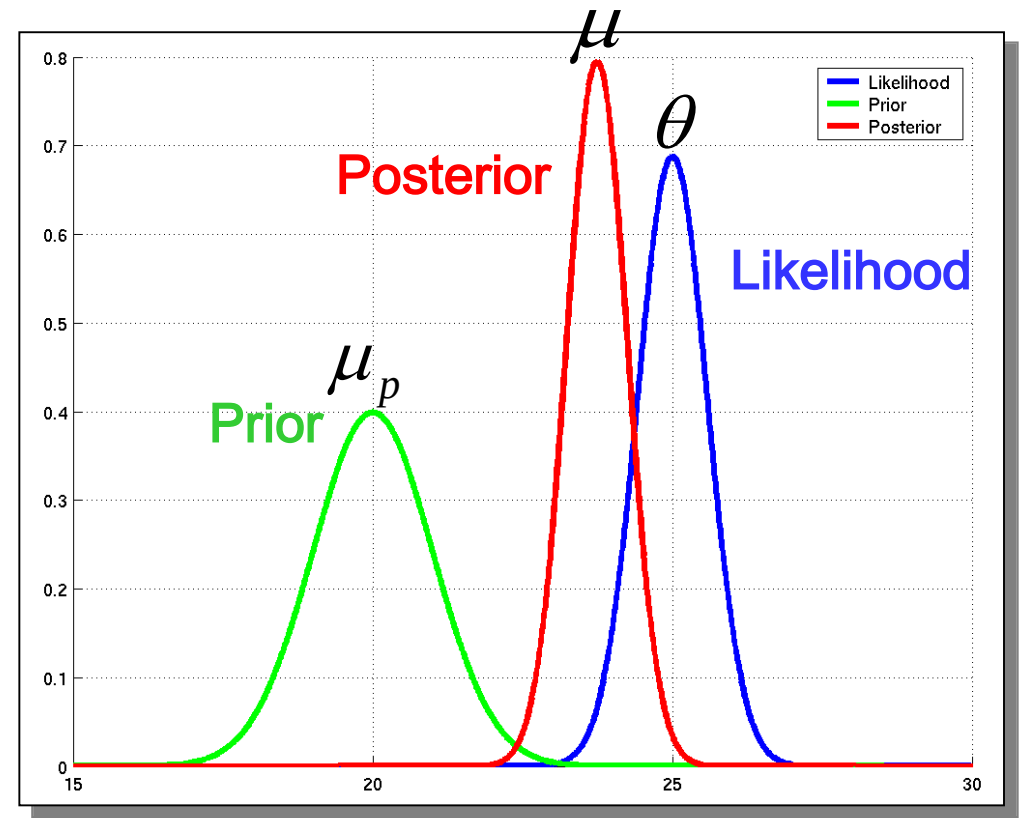
Likelihood & prior

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

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

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

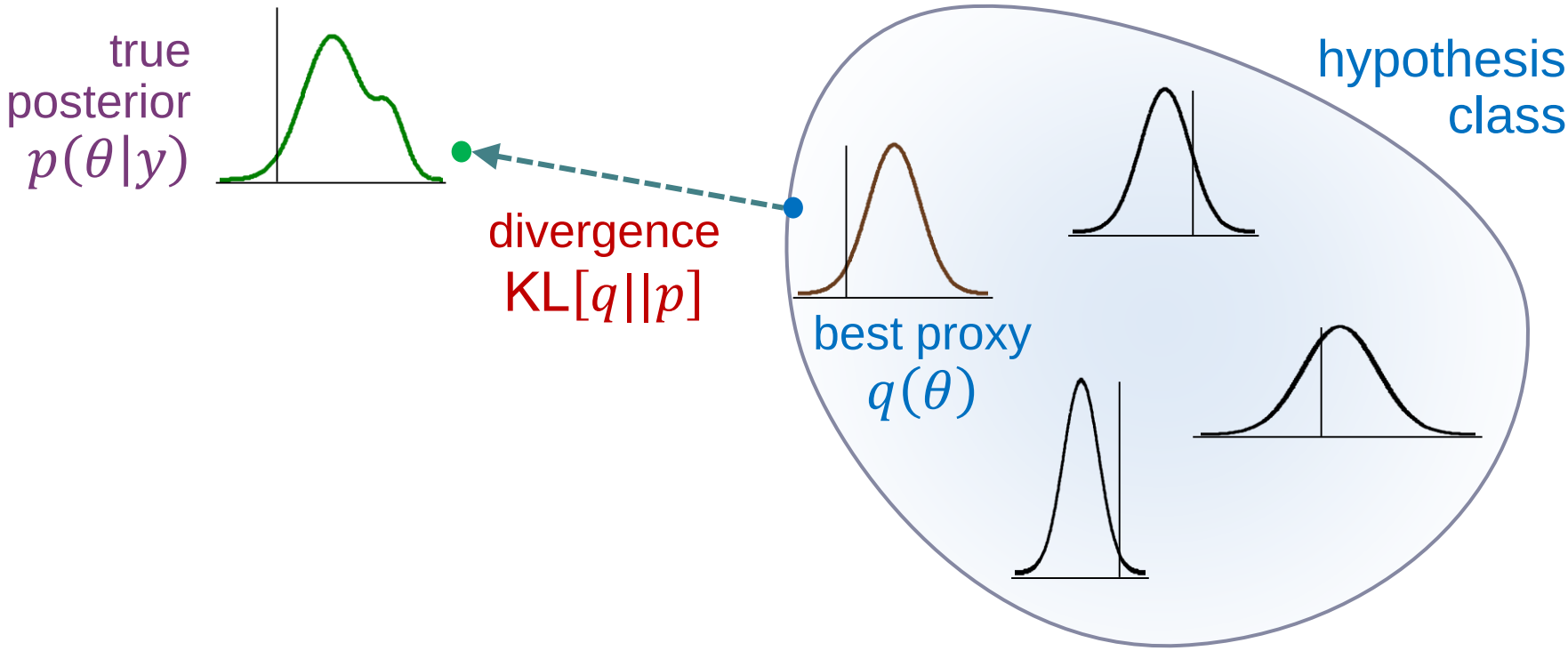
Relative precision weighting



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

- non-symmetric measure of the difference between two probability distributions P and Q
- $D_{KL}(P\|Q)$ = a measure of the information lost when Q is used to approximate P: the expected number of extra bits required to code samples from P when using a code optimized for Q, rather than using the true code optimized for P.

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

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

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= -\textstyle\sum_x p(x)\log q(x) + \sum_x p(x)\log p(x) \\ &= H(P,Q) - H(P) \end{aligned}$$

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

$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\text{divergence}} + \underbrace{F(q,y)}_{\text{neg. free energy}}$$

divergence ≥ 0 (unknown)

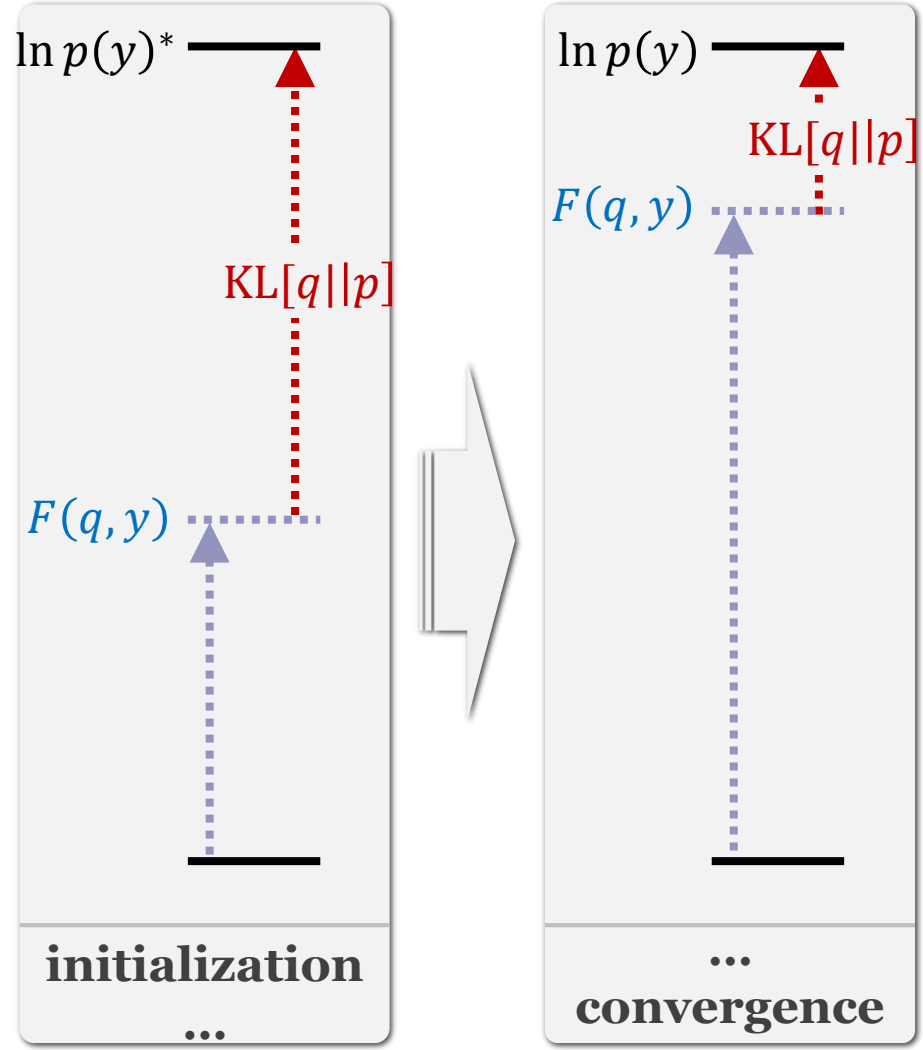
neg. free energy (easy to evaluate for a given q)

$F(q)$ 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

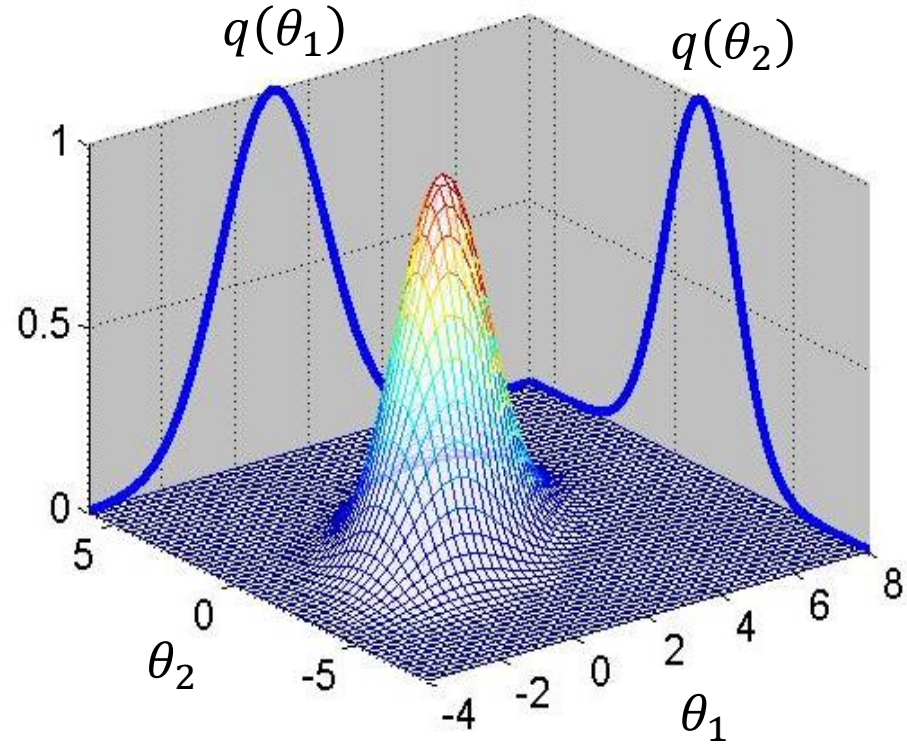
Factorize the approximate posterior $q(\theta)$ into independent partitions:

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

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.

VB (under mean-field assumption) in more detail

$$
\begin{aligned}
F(q,y) &= \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \mathrm{d}\theta \\
&= \int \prod_i q_i \times \left(\ln p(y,\theta) - \sum_i \ln q_i \right) \mathrm{d}\theta \qquad \text{mean-field assumption: } q(\theta) = \textstyle\prod_i q_i(\theta_i) \\
&= \int q_j \prod_{\setminus j} q_i \left(\ln p(y,\theta) - \ln q_j \right) \mathrm{d}\theta - \int q_j \prod_{\setminus j} q_i \sum_{\setminus j} \ln q_i \, \mathrm{d}\theta \\
&= \int q_j \left(\underbrace{\int \prod_{\setminus j} q_i \ln p(y,\theta) \, \mathrm{d}\theta_{\setminus j}}_{\langle \ln p(y,\theta) \rangle_{q_{\setminus j}}} - \ln q_j \right) \mathrm{d}\theta_j - \int q_j \int \prod_{\setminus j} q_i \ln \prod_{\setminus j} q_i \, \mathrm{d}\theta_{\setminus j} \, \mathrm{d}\theta_j \\
&= \int q_j \ln \frac{\exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}} \right)}{q_j} \mathrm{d}\theta_j + c \\
&= -\mathrm{KL}\left[q_j \,||\, \exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}} \right) \right] + c
\end{aligned}
$$

VB (under mean-field assumption) in more detail

In summary:

$$F(q,y) = -\mathrm{KL}\left[q_j \| \exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)\right] + c$$

Suppose the densities $q_{\setminus j} \equiv q(\theta_{\setminus j})$ are kept fixed. Then the approximate posterior $q(\theta_j)$ that maximizes $F(q,y)$ is given by:

$$\begin{aligned} q_j^* &= \arg\max_{q_j} F(q,y) \\ &= \frac{1}{Z}\exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right) \end{aligned}$$

Therefore:

$$\ln q_j^* = \underbrace{\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}}_{=:I(\theta_j)} - \ln Z$$

This implies a straightforward algorithm for variational inference:

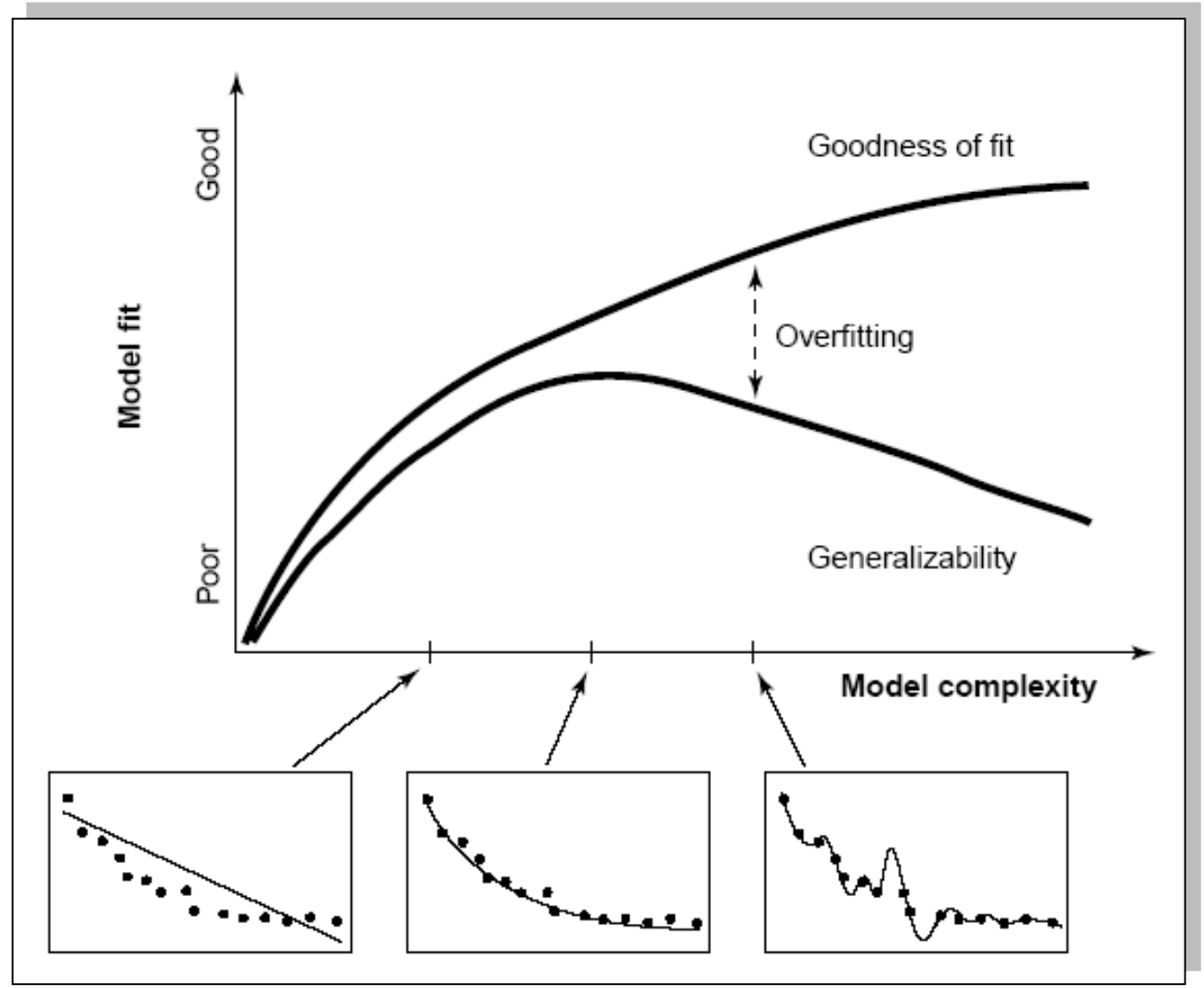
1. Initialize all approximate posteriors $q(\theta_i)$, e.g., by setting them to their priors.
2. Cycle over the parameters, revising each given the current estimates of the others.
3. Loop until convergence.

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?

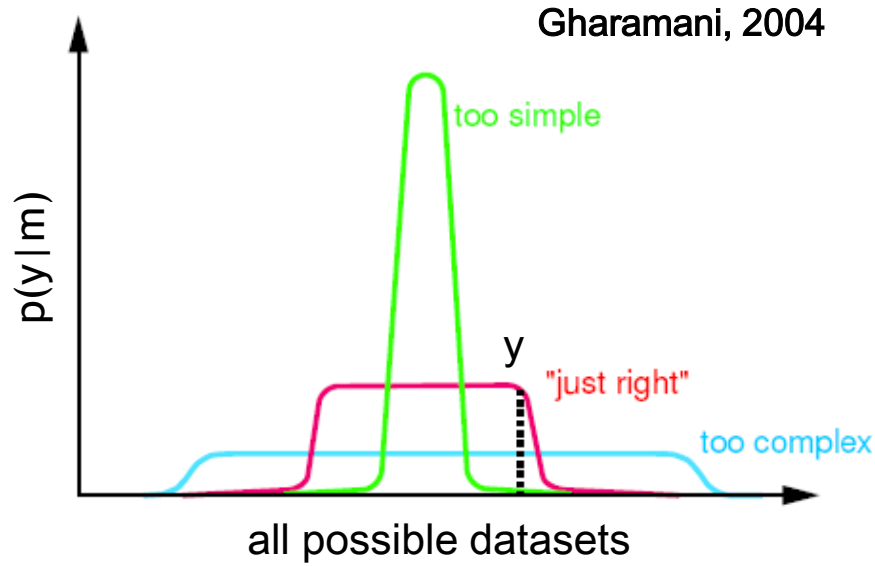


Pitt & Miyung (2002) *TICS*

Bayesian model selection (BMS)

Model evidence (marginal likelihood):

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



→ accounts for both accuracy and complexity of the model

→ "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?"

Various approximations, e.g.:

- negative free energy, AIC, BIC

McKay 1992, *Neural Comput.*
Penny et al. 2004a, *NeuroImage*

Model space (hypothesis set) M

Model space M is defined by prior on models.

Usual choice: flat prior over a small set of models.

$$p(m) = \begin{cases} 1/|M| \text{ if } m \in M \\ 0 \text{ if } m \notin M \end{cases}$$

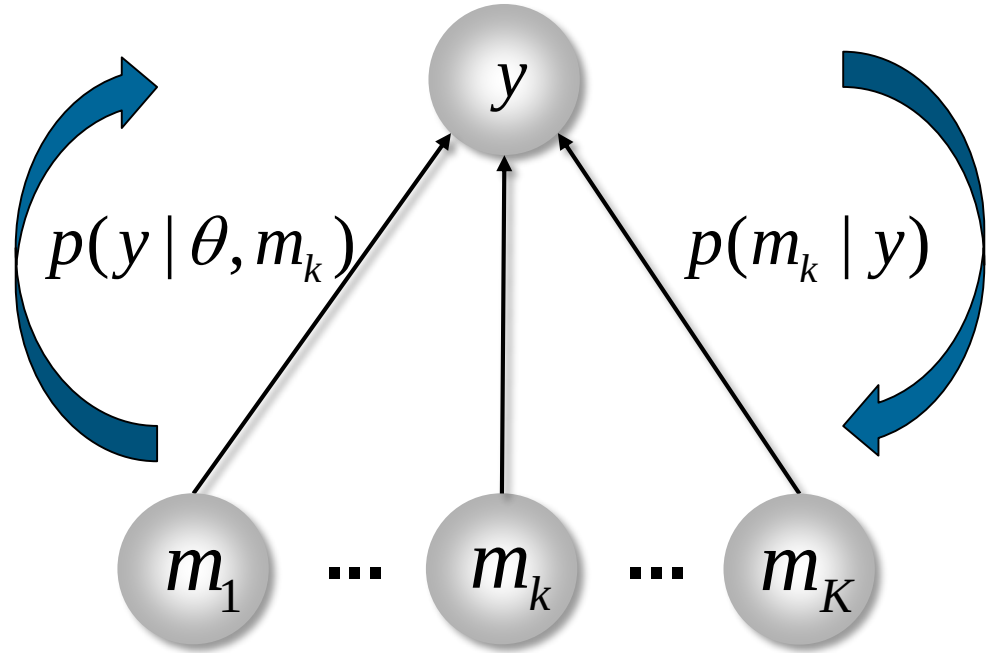
In this case, the posterior probability of model i is:

$$p(m_i \mid y) = \frac{p(y \mid m_i)\,p(m_i)}{\sum_{j=1}^{|M|} p(y \mid m_j)\,p(m_j)} = \frac{p(y \mid m_i)}{\sum_{j=1}^{|M|} p(y \mid m_j)}$$

Differential diagnosis based on generative models of disease symptoms

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. 2016, *NeuroImage*

Approximations to the model evidence

Logarithm is a monotonic function → Maximizing log model evidence = Maximizing model evidence

Log model evidence = balance between fit and complexity

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

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

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

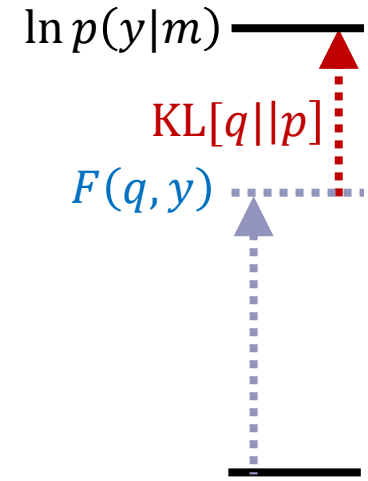
No. of parameters

No. of data points

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:

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

The (negative) free energy approximation

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

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

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

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):

$$KL\left[q(\theta), p(\theta \mid m)\right]$$

$$= \frac{1}{2}\ln\left|C_\theta\right| - \frac{1}{2}\ln\left|C_{\theta|y}\right| + \frac{1}{2}\left(\mu_{\theta|y} - \mu_\theta\right)^T C_\theta^{-1}\left(\mu_{\theta|y} - \mu_\theta\right)$$

- 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:

positive value, [0; ∞[

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

Kass & Raftery classification:

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

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

Fixed effects BMS at group level

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

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

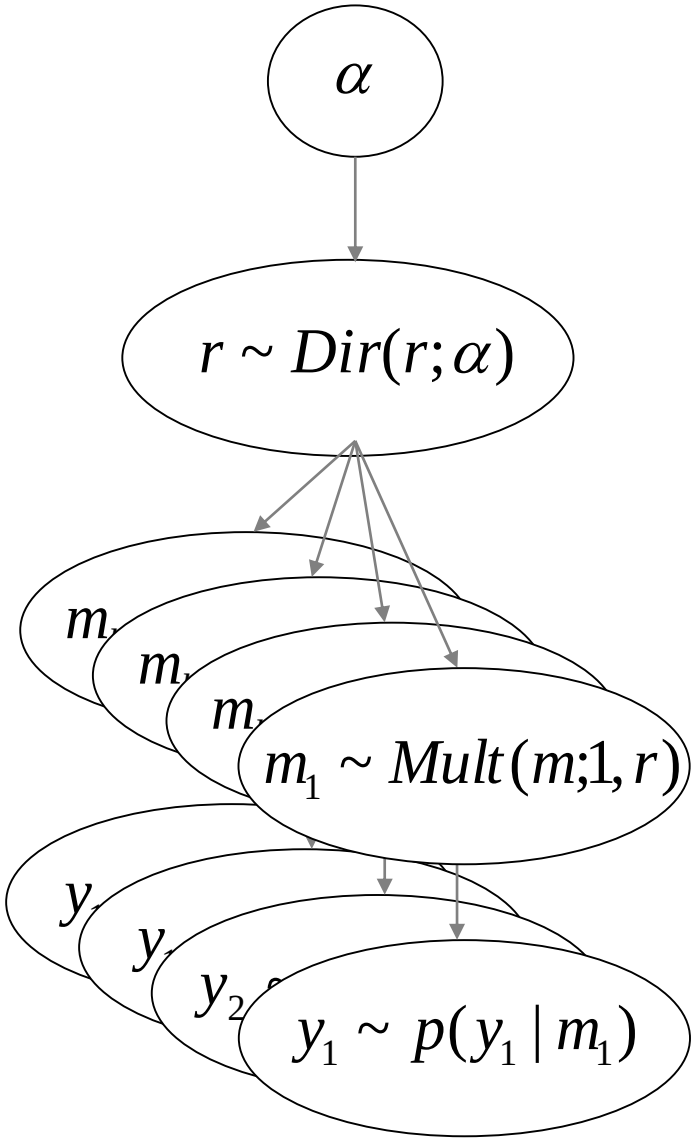
Average Bayes factor (ABF):

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

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

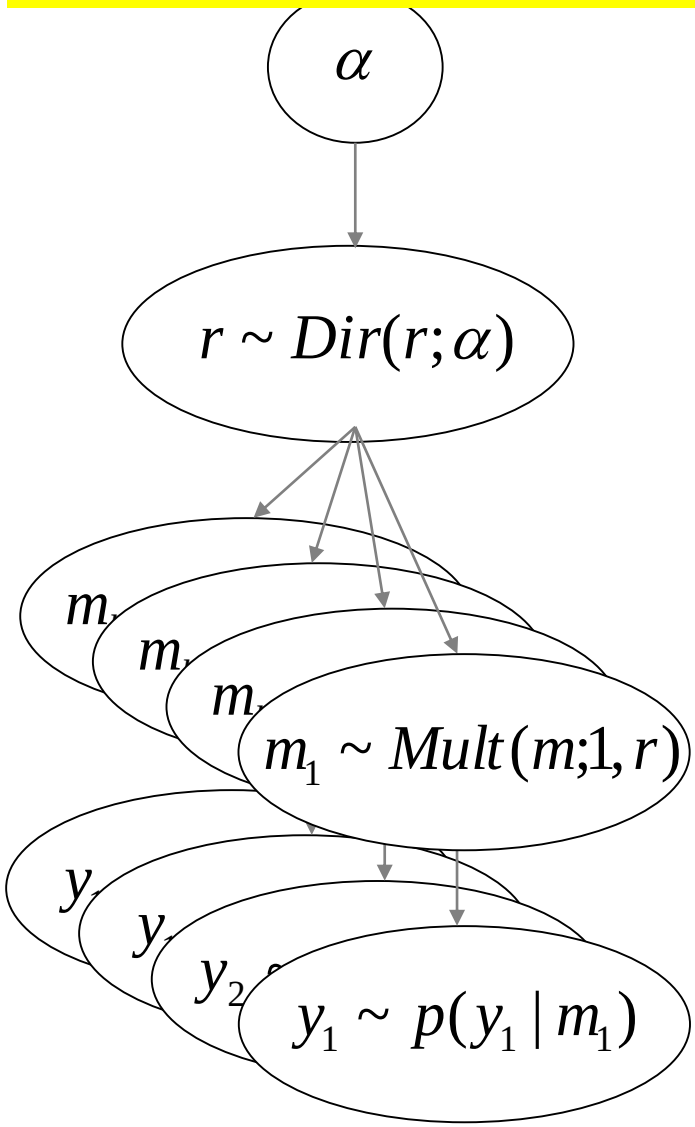
Multinomial distribution of model labels m

Model inversion by Variational Bayes or MCMC

Measured data y

Stephan et al. 2009, *NeuroImage*

Random effects BMS



$$p(r \mid \alpha) = Dir\left(r, \alpha\right) = \frac{1}{Z\left(\alpha\right)} \prod_k r_k^{\alpha_k - 1}$$

$$Z(\alpha) = \prod_k \Gamma\left(\alpha_k\right) \Big/ \Gamma\left(\sum_k \alpha_k\right)$$

$$p(m_n \mid r) = \prod_k r_k^{m_{nk}}$$

$$p(y_n \mid m_{nk}) = \int p(y \mid \vartheta)\, p(\vartheta \mid m_{nk})\, d\vartheta$$

Stephan et al. 2009, *NeuroImage*

❶ Write down joint probability and take the log

$$
\begin{aligned}
p(y,r,m) &= p(y\,|\,m)\,p(m\,|\,r)\,p(r\,|\,\alpha_0) \\
&= p(r\,|\,\alpha_0)\left[\prod_n p(y_n\,|\,m_n)\,p(m_n\,|\,r)\right] \\
&= \frac{1}{Z(\alpha_0)}\left[\prod_k r_k^{\alpha_{0k}-1}\right]\left[\prod_n p(y_n\,|\,m_n)\prod_k r_k^{m_{nk}}\right] \\
&= \frac{1}{Z(\alpha_0)}\prod_n\left[\prod_k\left[p(y_n\,|\,m_{nk})\,r_k\right]^{m_{nk}}\,r_k^{\alpha_{0k}-1}\right]
\end{aligned}
$$

$$
\ln p(y,r,m) = -\ln Z(\alpha_0) + \sum_n\sum_k\left((\alpha_{0k}-1)\ln r_k + m_{nk}\left(\log p(y_n\,|\,m_{nk}) + \ln r_k\right)\right)
$$

❷ Mean field approx.

$$q(r,m) = q(r)q(m)$$

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

$$q(r) \propto \exp\left(I\left(r\right)\right)$$

$$q(m) \propto \exp\left(I\left(m\right)\right)$$

$$I\left(r\right) = \left\langle \log p\left(y,r,m\right)\right\rangle_{q(m)}$$

$$I\left(m\right) = \left\langle \log p\left(y,r,m\right)\right\rangle_{q(r)}$$

❹ Iterative updating of sufficient statistics of approx. posteriors

$\alpha = \alpha_0$ $\qquad\qquad$ $\alpha_0 = [1,\ldots,1]$

Until convergence

$$u_{nk} = \exp\left(\ln p(y_n \mid m_{nk}) + \Psi(\alpha_k) - \Psi\left(\sum_k \alpha_k \right) \right)$$

$$g_{nk} = \frac{u_{nk}}{\sum_k u_{nk}}$$

$$\beta_k = \sum_n g_{nk}$$

$$\alpha = \alpha_0 + \beta$$

end

$g_{nk} = q(m_{nk} = 1)$ our (normalized) posterior belief that model *k* generated the data from subject *n*

$\beta_k = \sum_n g_{nk}$ expected number of subjects whose data we believe were generated by model *k*

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

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

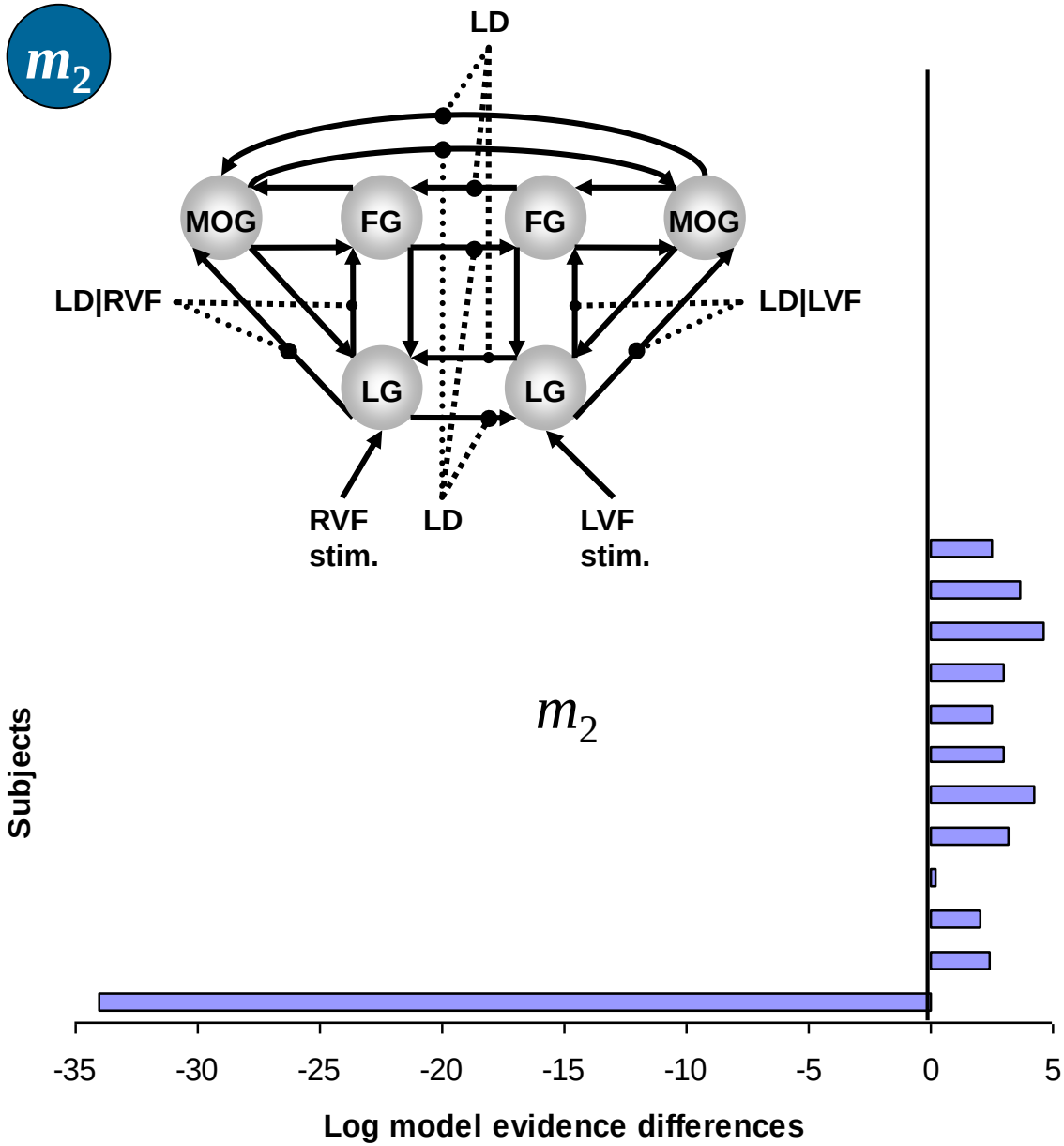
3. **exceedance probability** that a particular model *k* is more likely than any other model (of the *K* models tested), given the group data

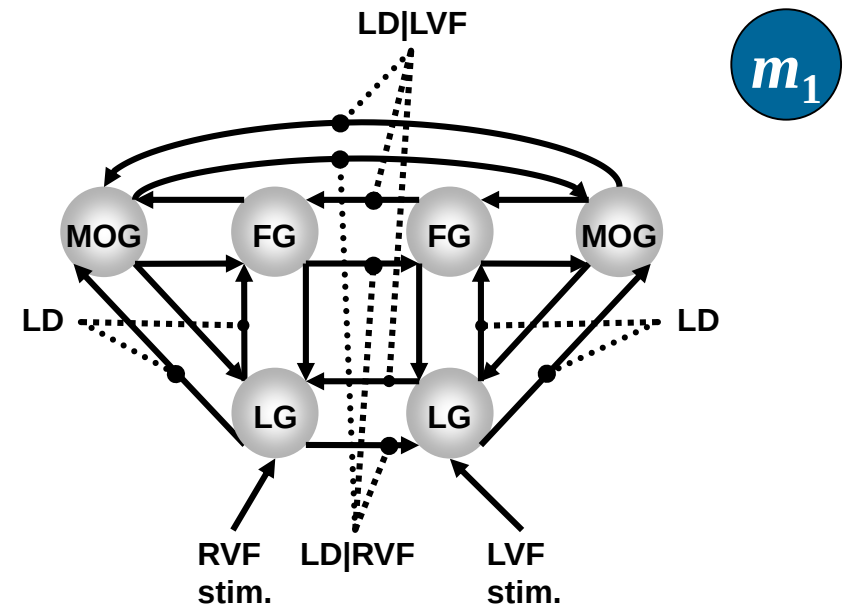
$$\exists k \in \{1...K\}, \forall j \in \{1...K \mid j \neq k\}:$$

$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

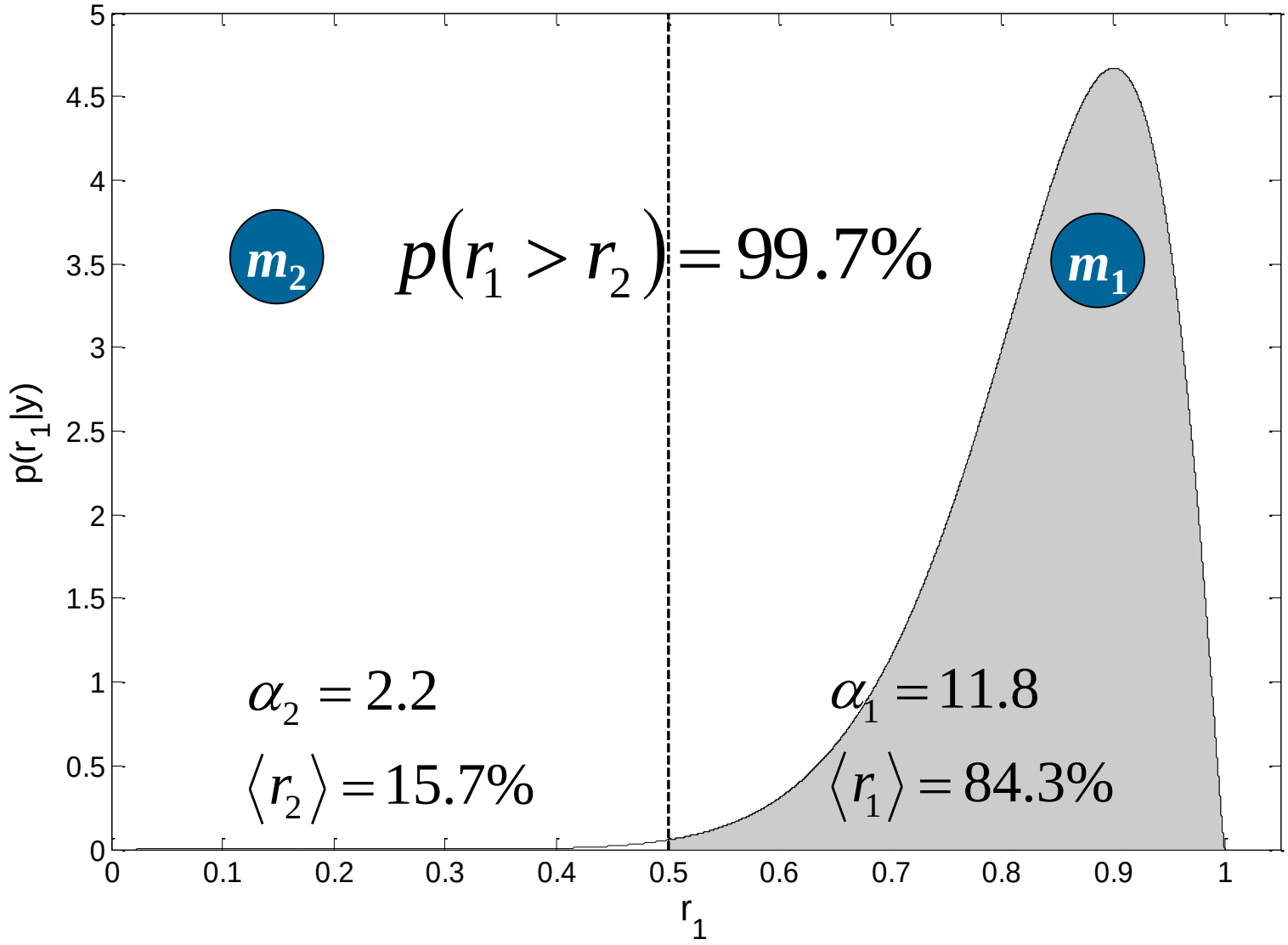
4. **protected exceedance probability**: see below

Example: Hemispheric interactions during vision





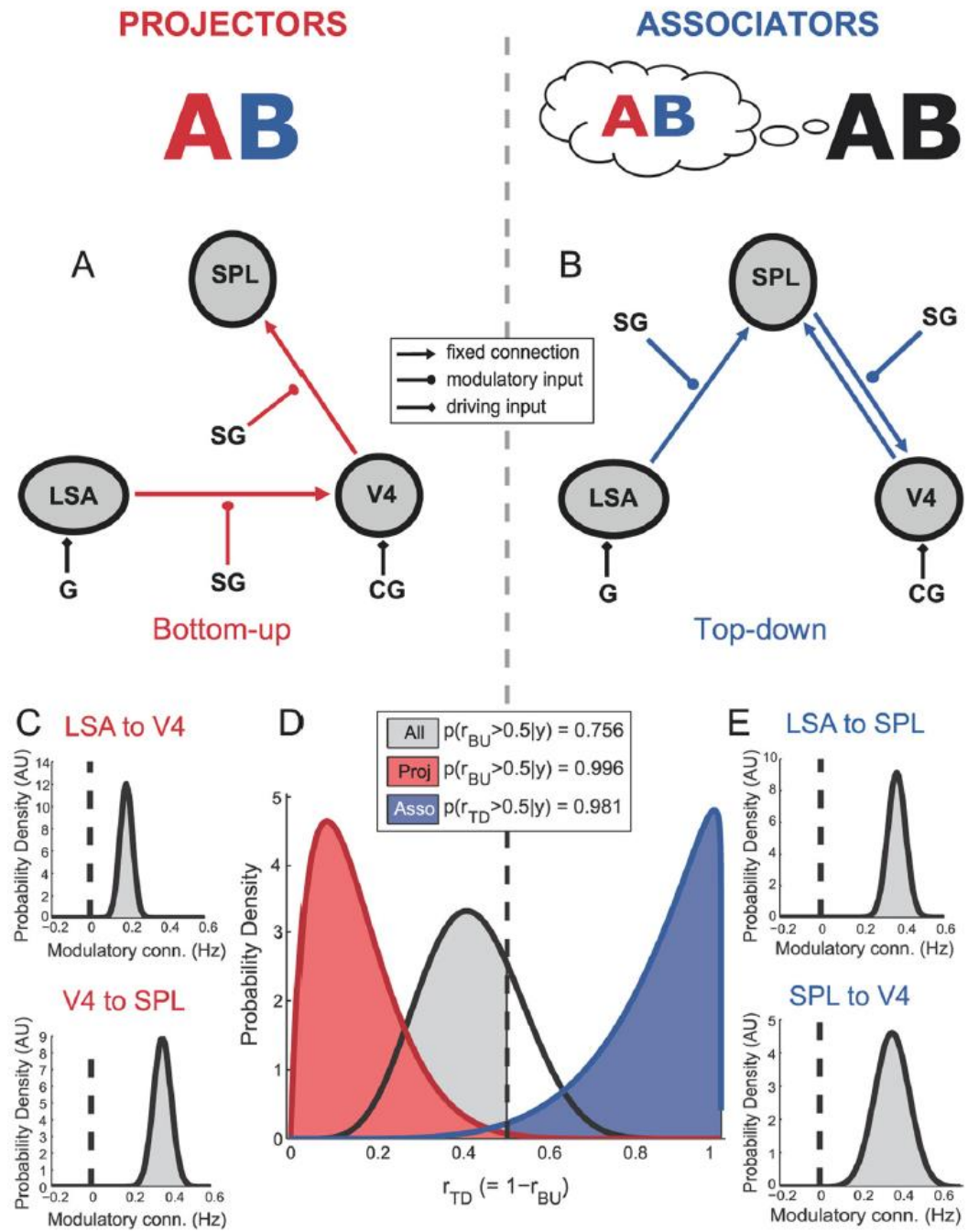
Data: Stephan et al. 2003, *Science*
Models: Stephan et al. 2007, *J. Neurosci.*



Stephan et al. 2009a, *NeuroImage*

Example: Synaesthesia

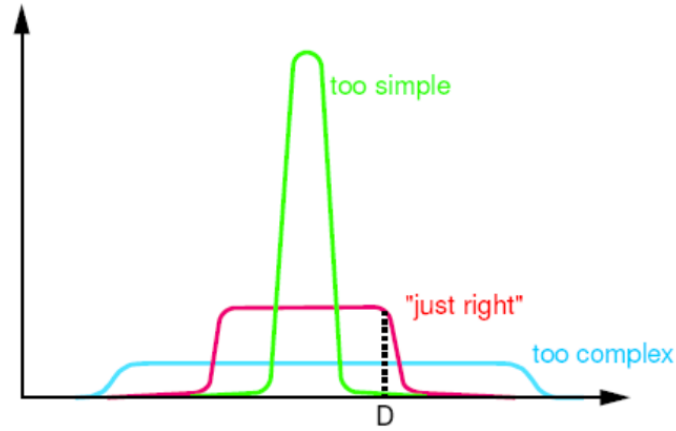
- “projectors” experience color externally colocalized with a presented grapheme
- “associators” report an internally evoked association
- across all subjects: no evidence for either model
- but BMS results map precisely onto projectors (bottom-up mechanisms) and associators (top-down)



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

Overfitting at the level of models

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



posterior model probability:

$$p(m \mid y)$$

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

BMA:

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

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

Model space partitioning: comparing model families

- partitioning model space into K subsets or families:

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

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

$$p(f_k)$$

- 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: fixed effects

- We wish to have a uniform prior at the family level:

$$p(f_k) = \frac{1}{K}$$

- This is related to the model level via the sum of the priors on models:

$$p(f_k) = \sum_{m \in f_k} p(m)$$

- Hence the uniform prior at the family level is:

$$\forall m \in f_k : p(m) = \frac{1}{K|f_k|}$$

- The probability of each family is then obtained by summing the posterior probabilities of the models it includes:

$$p(f_k \mid y_{1..N}) = \sum_{m \in f_k} p(m \mid y_{1..N})$$

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

Family-level inference: random effects

- The frequency of a family in the population is given by:

$$s_k = \sum_{m \in f_k} r_m$$

- In RFX-BMS, this follows a Dirichlet distribution, with a uniform prior on the parameters α (see above).

$$p(s) = Dir(\alpha)$$

- A uniform prior over family probabilities can be obtained by setting:

$$\forall m \in f_k : \alpha_{prior}(m) = \frac{1}{|f_k|}$$

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:

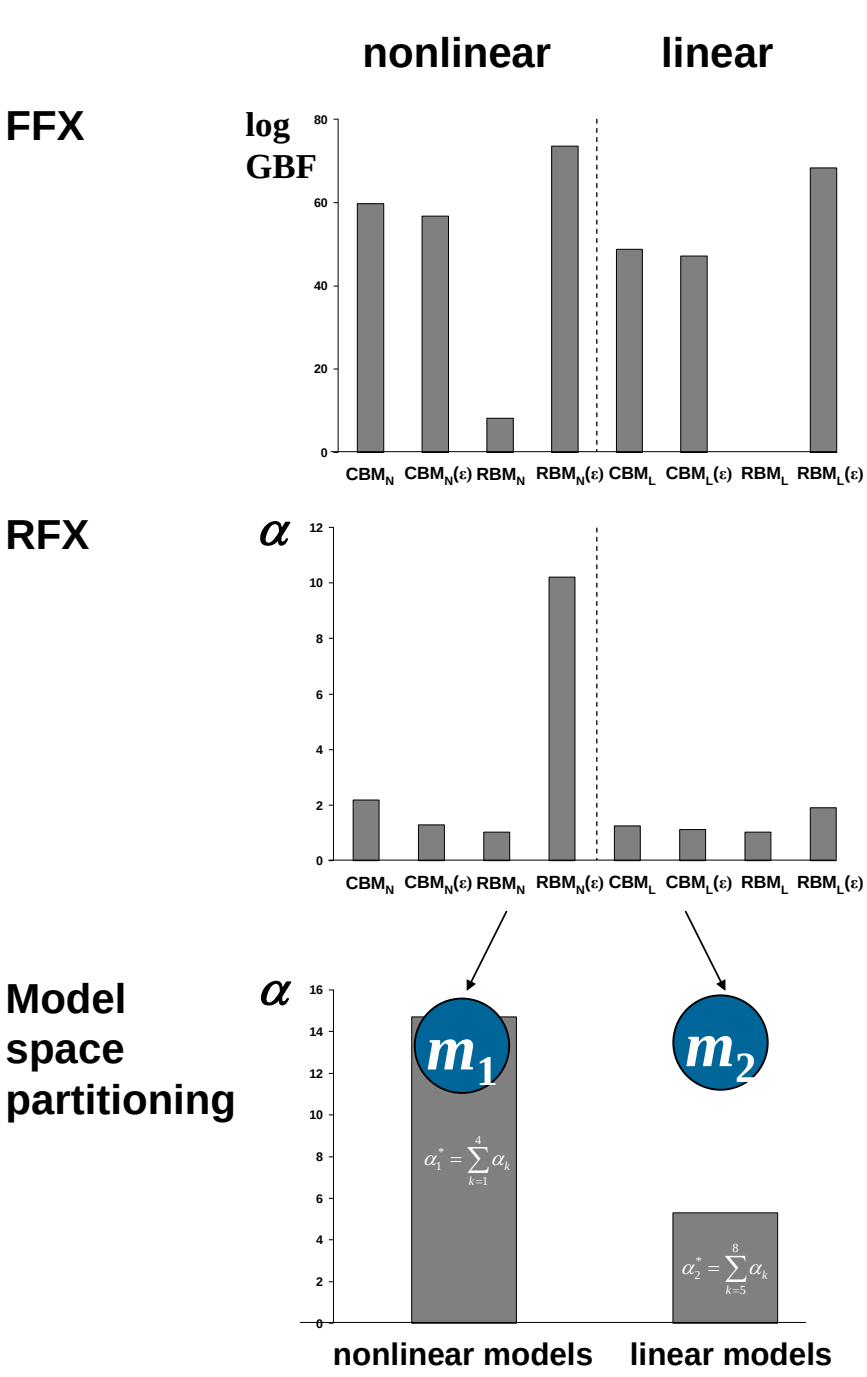
$$\left(r_1, r_2, ..., r_K\right) \sim Dir\left(\alpha_1, \alpha_2, ..., \alpha_K\right)$$

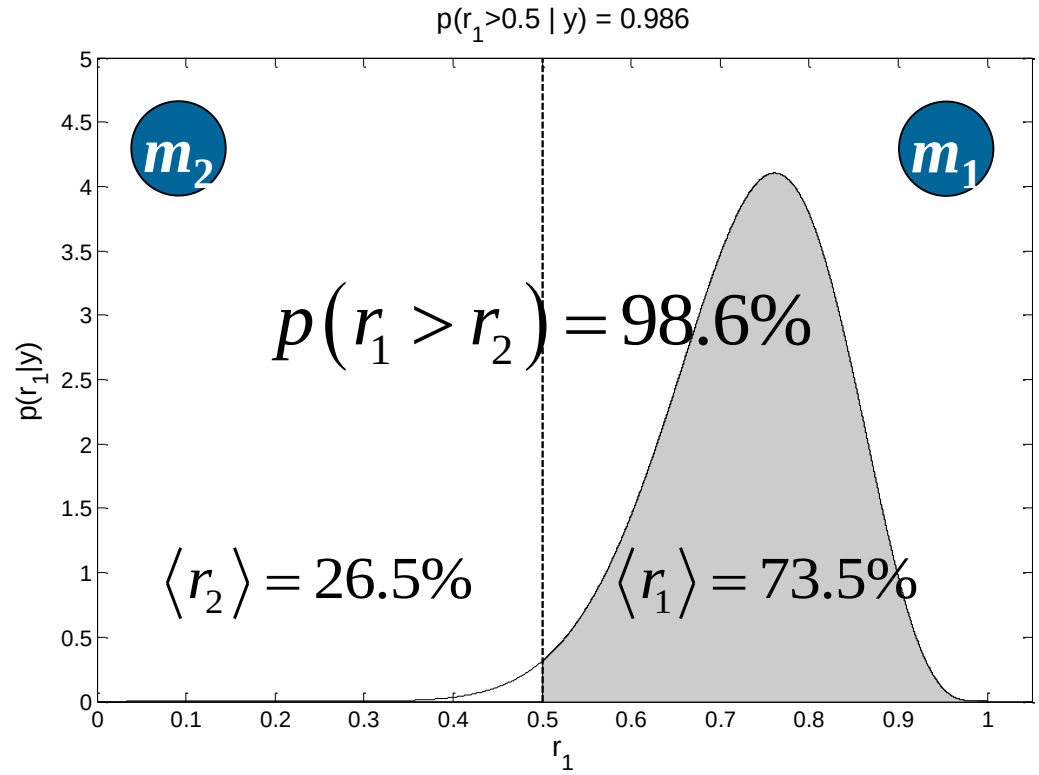
$$\Rightarrow r_1^* = \sum_{k \in N_1} r_k, r_2^* = \sum_{k \in N_2} r_k, ..., r_J^* = \sum_{k \in N_J} r_k$$

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

Stephan et al. 2009, *NeuroImage*

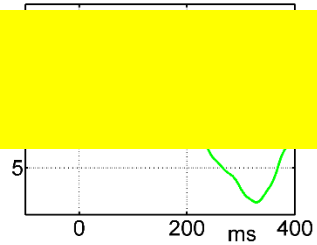
Model space partitioning: comparing model families



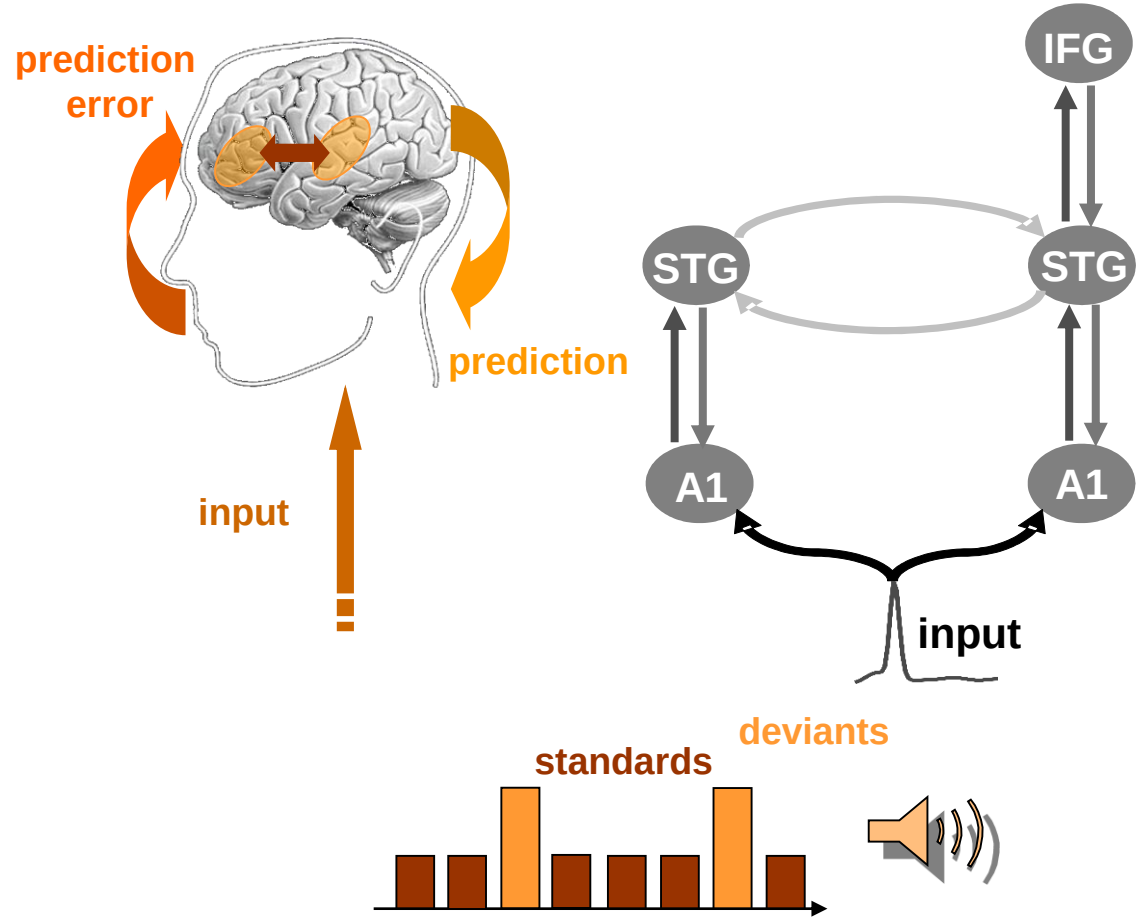


Stephan et al. 2009, *NeuroImage*

Mismatch negativity (MMN)



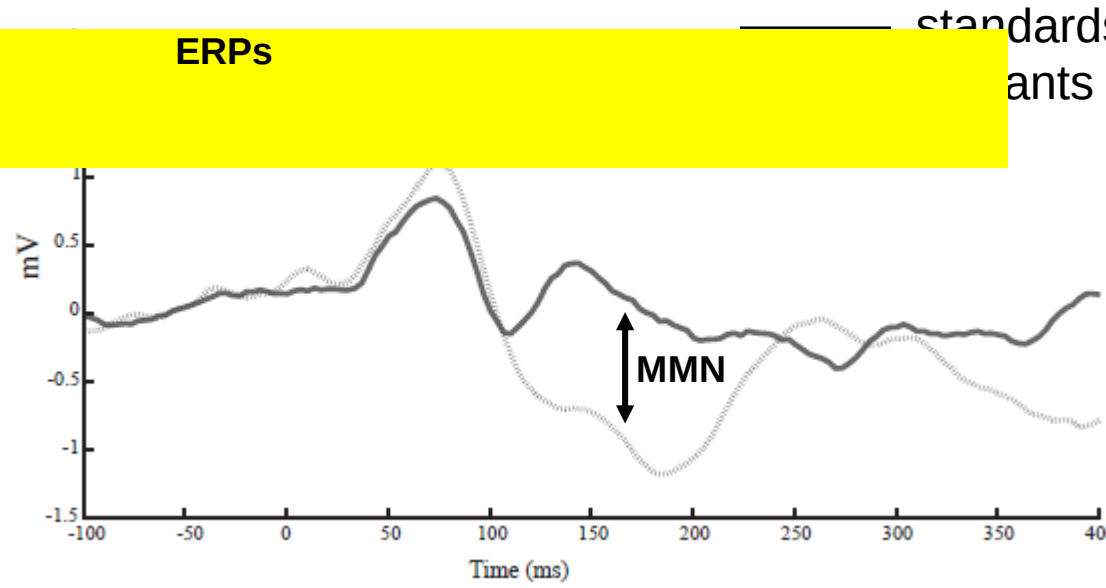
- elicited by surprising stimuli (scales with unpredictability)
- ↓ in schizophrenic patients
- classical interpretations:
  - pre-attentive change detection
  - neuronal adaptation
- current theories:
  - reflection of (hierarchical) Bayesian inference

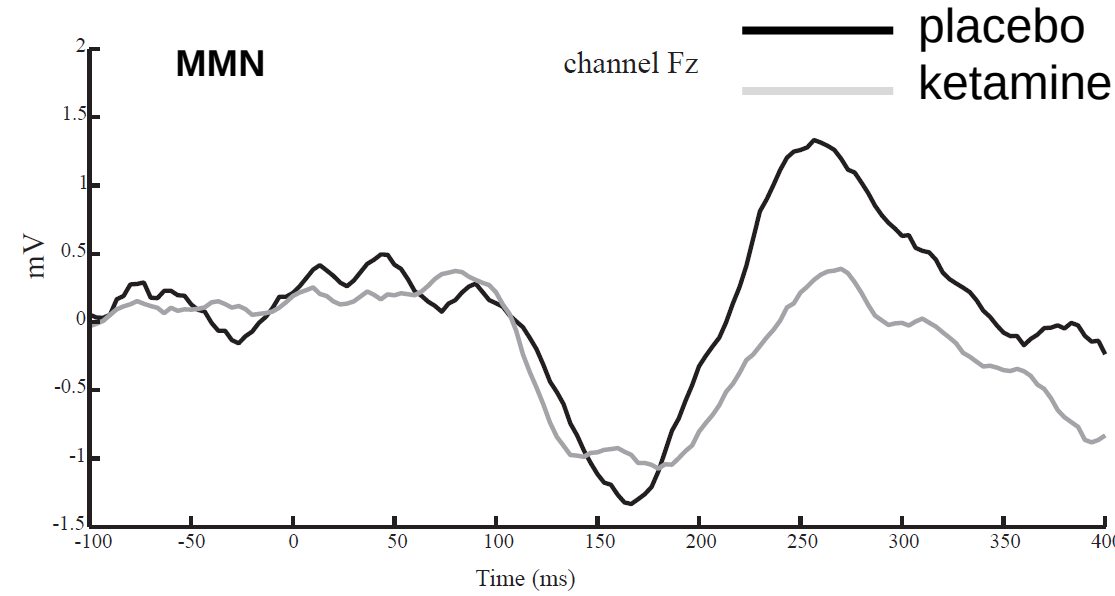


Garrido et al. 2009, *Clin. Neurophysiol.*

Mismatch negativity (MMN)

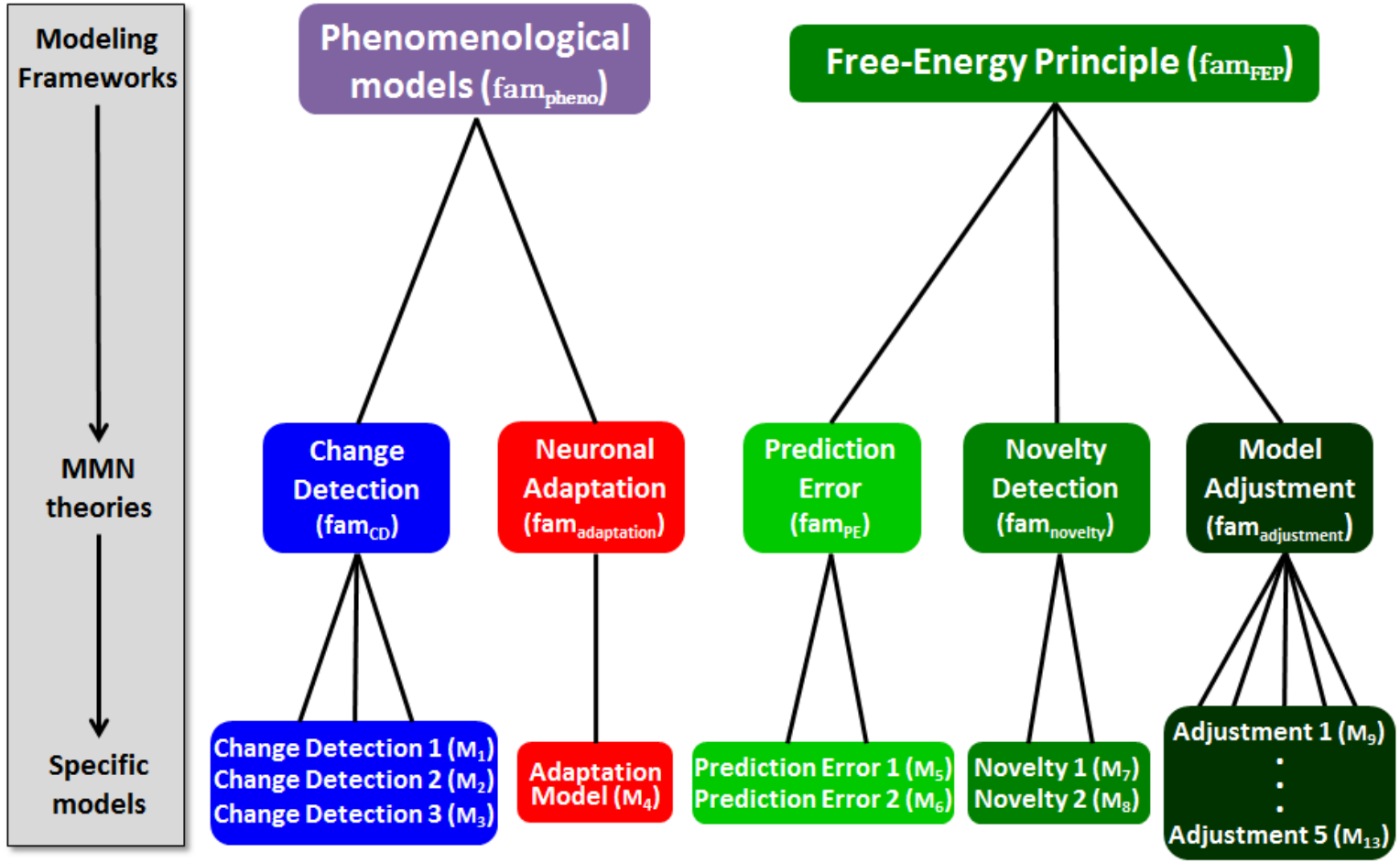
- ↓ in schizophrenic patients
- Highly relevant for computational assays of glutamatergic and cholinergic transmission:
  - + NMDAR
  - + ACh (nicotinic & muscarinic)
  - – 5HT
  - – DA





Schmidt, Diaconescu et al. 2013, *Cereb. Cortex*

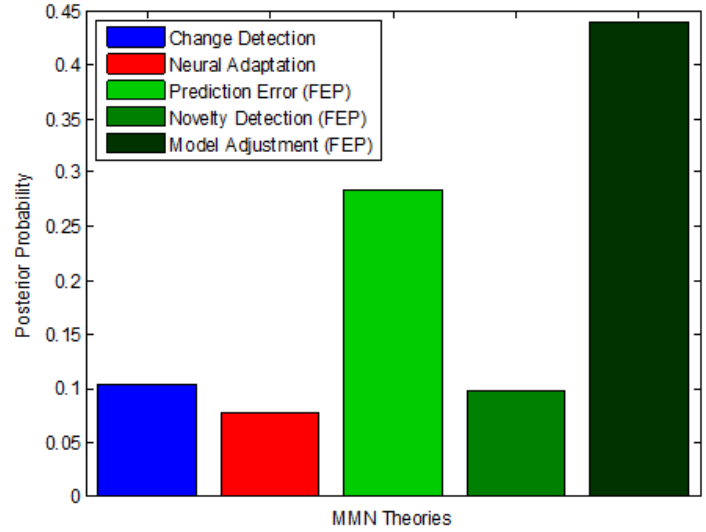
Modelling Trial-by-Trial Changes of the Mismatch Negativity (MMN)



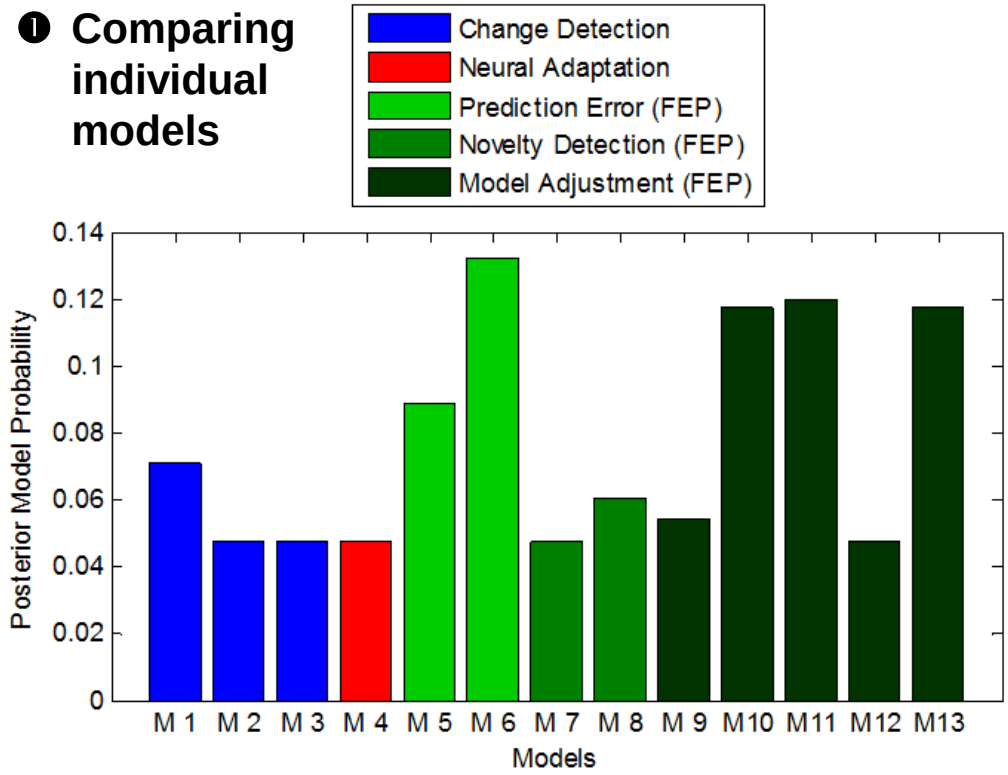
Lieder et al. 2013, *PLoS Comput. Biol.*

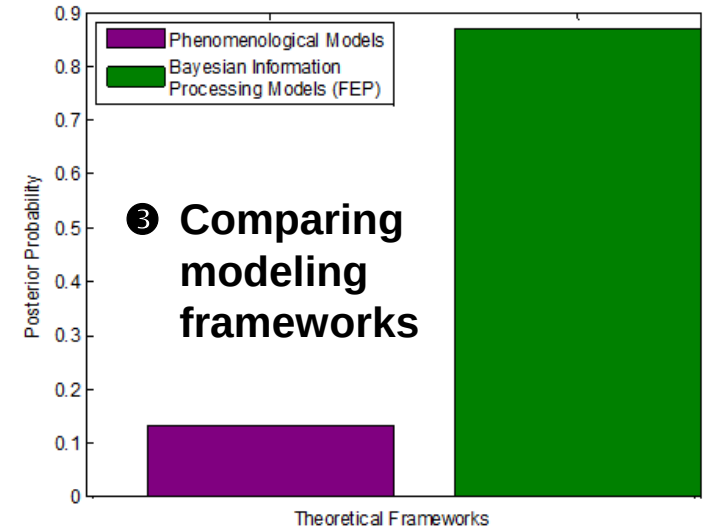
MMN model comparison at multiple levels

❷ **Comparing MMN theories**



❶ **Comparing individual models**





❸ **Comparing modeling frameworks**

Lieder et al. 2013, *PLoS Comput. Biol.*

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:

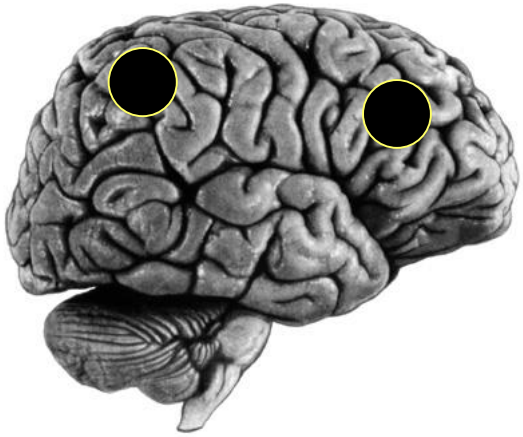
$$p(\theta \mid y)$$
$$= \sum_m p(\theta \mid y, m) p(m \mid y)$$

group-level BMA:

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

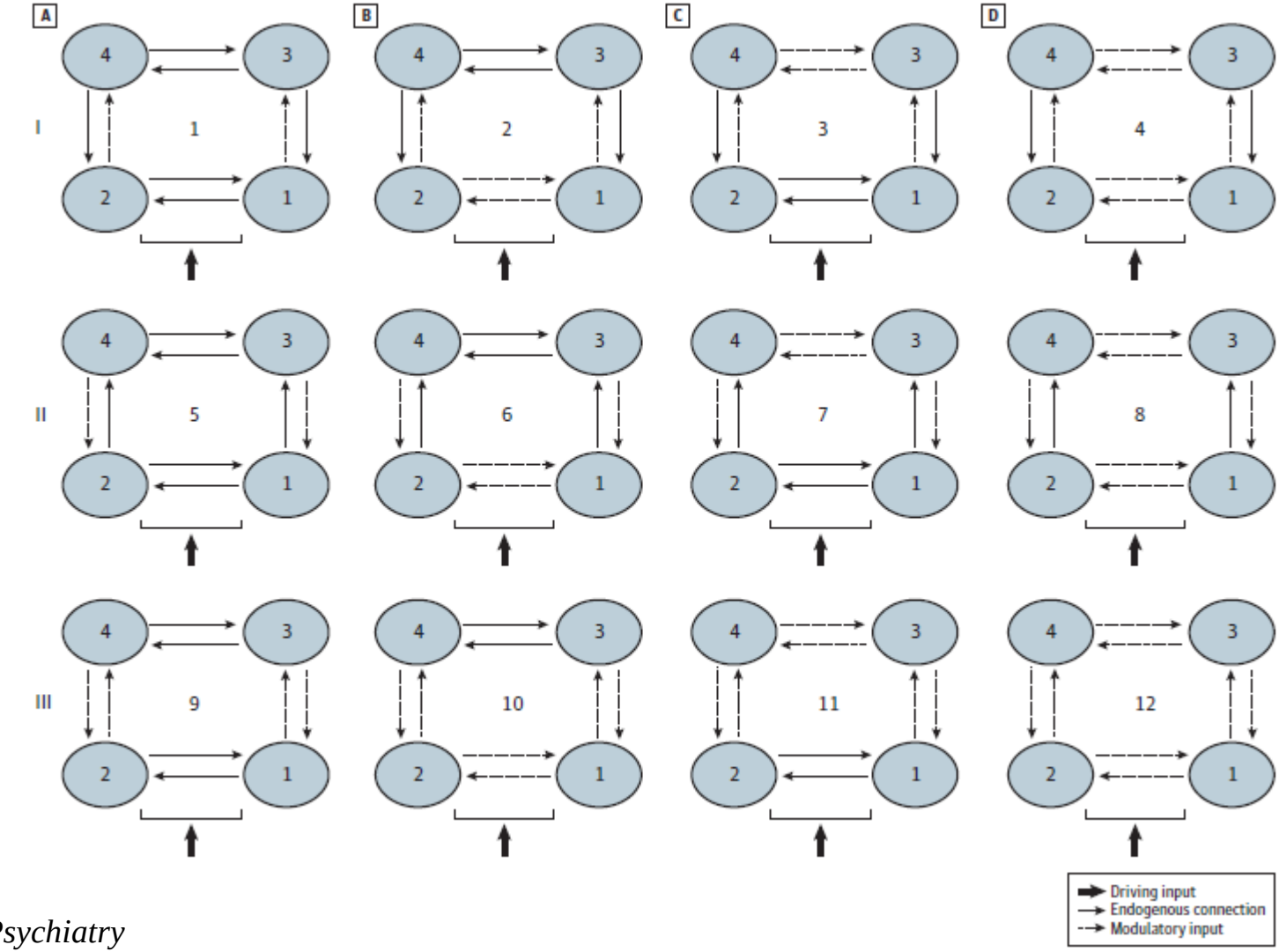
NB: $p(m|y_{1..N})$ can be obtained by either FFX or RFX BMS

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



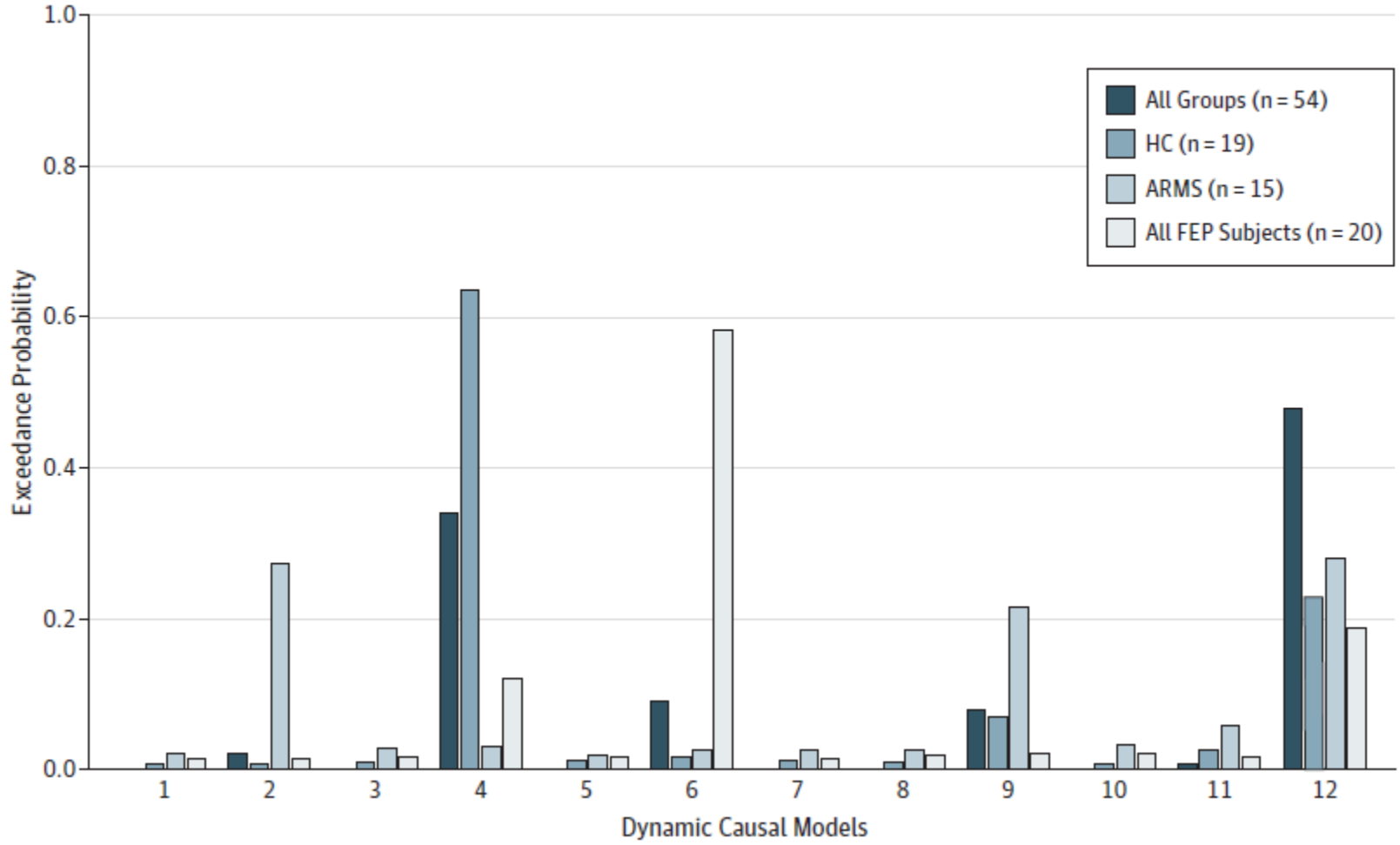
Prefrontal-parietal connectivity during working memory in schizophrenia

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



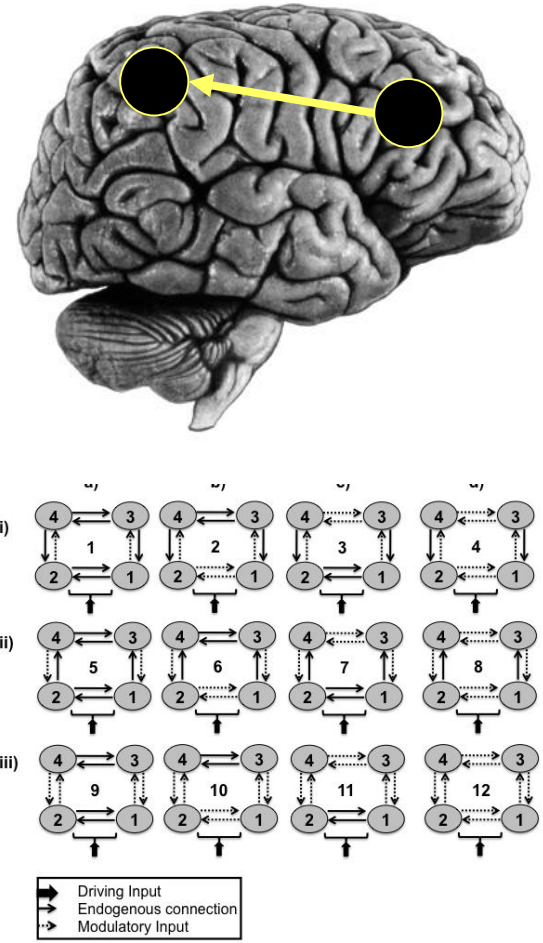
Schmidt et al. 2013, *JAMA Psychiatry*

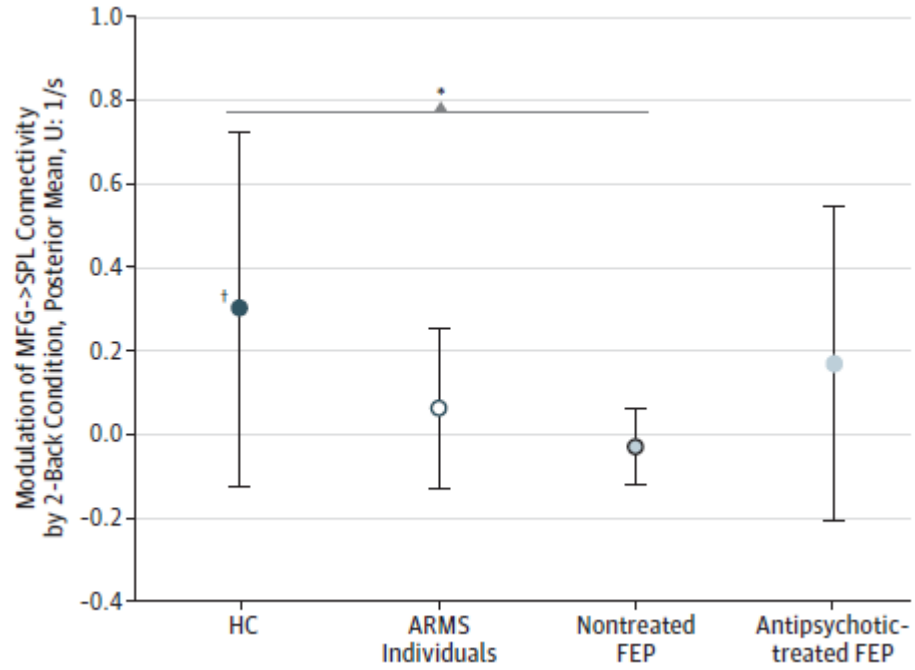
BMS results for all groups



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

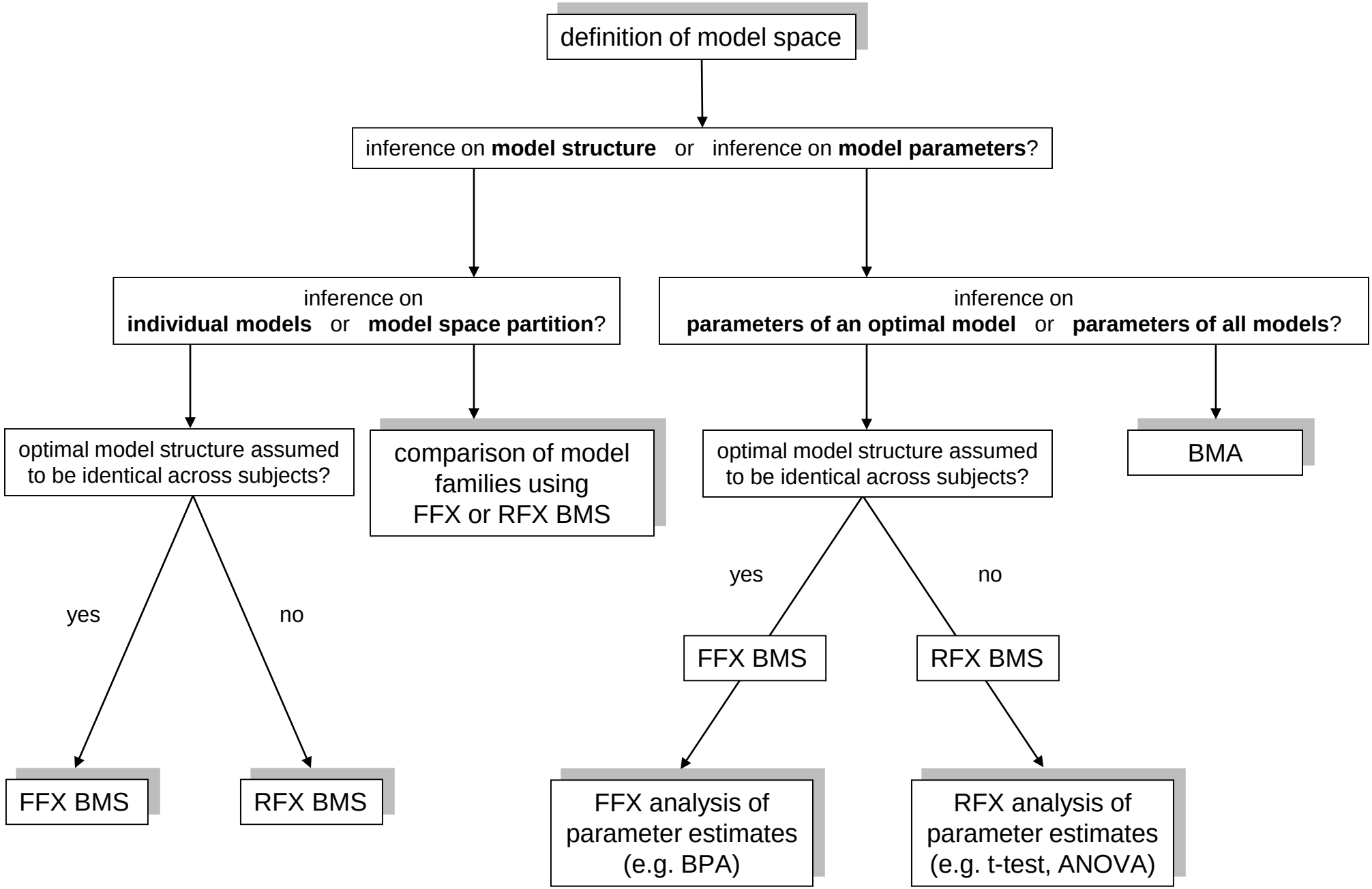
- EPs express our confidence that the posterior probabilities of models are different – under the hypothesis H_1 that models differ in probability: $r_k \neq 1/K$
- does not account for possibility "null hypothesis" H_0: $r_k = 1/K$
- **Bayesian omnibus risk (BOR)** of wrongly accepting H_1 over H_0:

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

- **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 (2016) Computational Neuroimaging Strategies for Single Patient Predictions. NeuroImage, in press. DOI: 10.1016/j.neuroimage.2016.06.038

Thank you