Mathematical Basics of Computational Psychiatry: Generative Modeling

Klaas Enno Stephan

Translational Neuromodeling Unit

Universität Zürich[UZH]

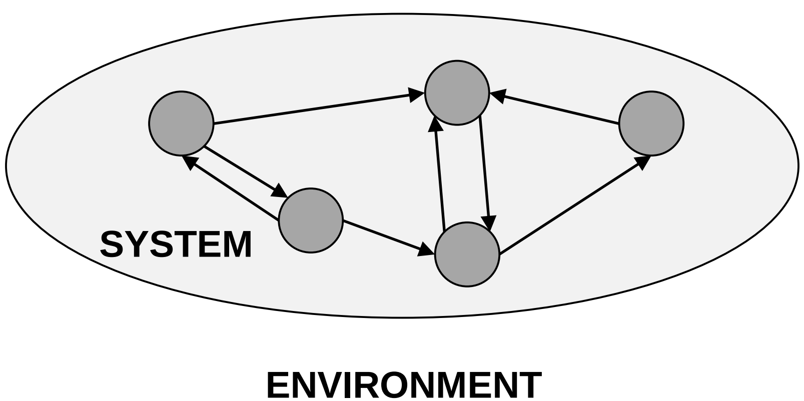
ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

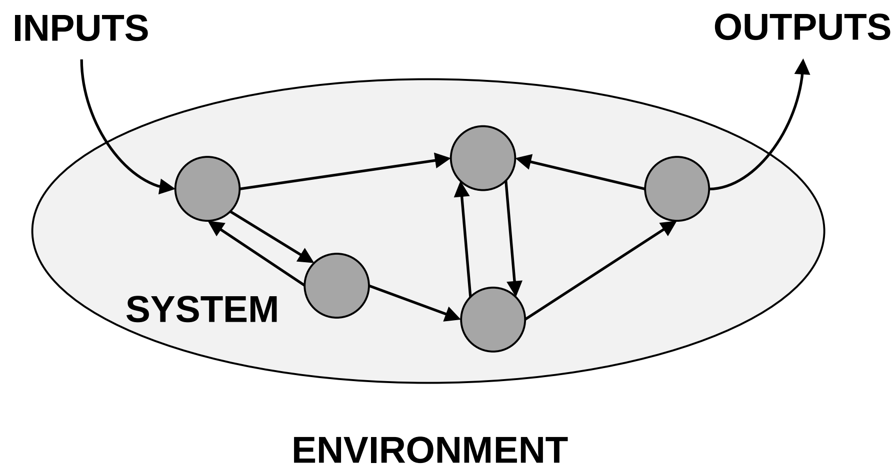
Systems

- system = a set of entities that interact to form a unified whole
- biological systems are open systems: they interact with their environment (exchange of energy, matter, information)

isolated system



open system

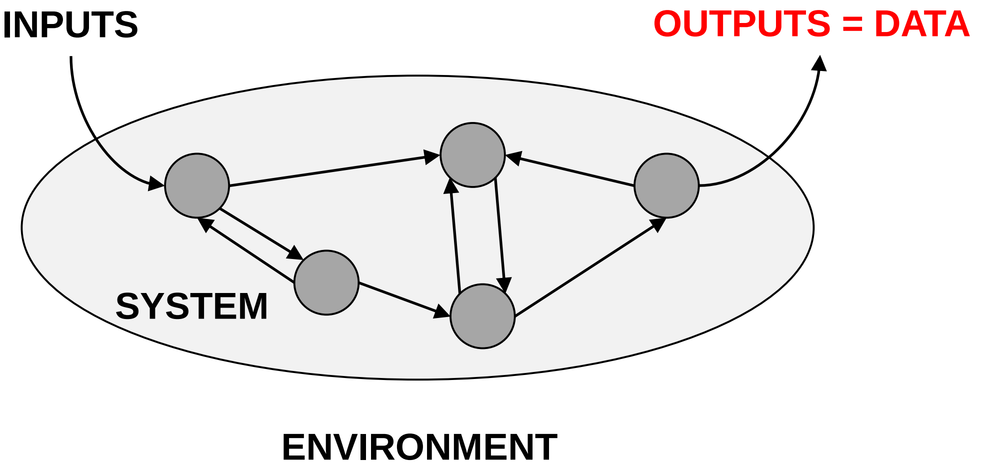


System models (state space models, latent process models)

- mathematically formal description of a system's behavior
(at an algorithmic or biophysical level that cannot be observed directly)

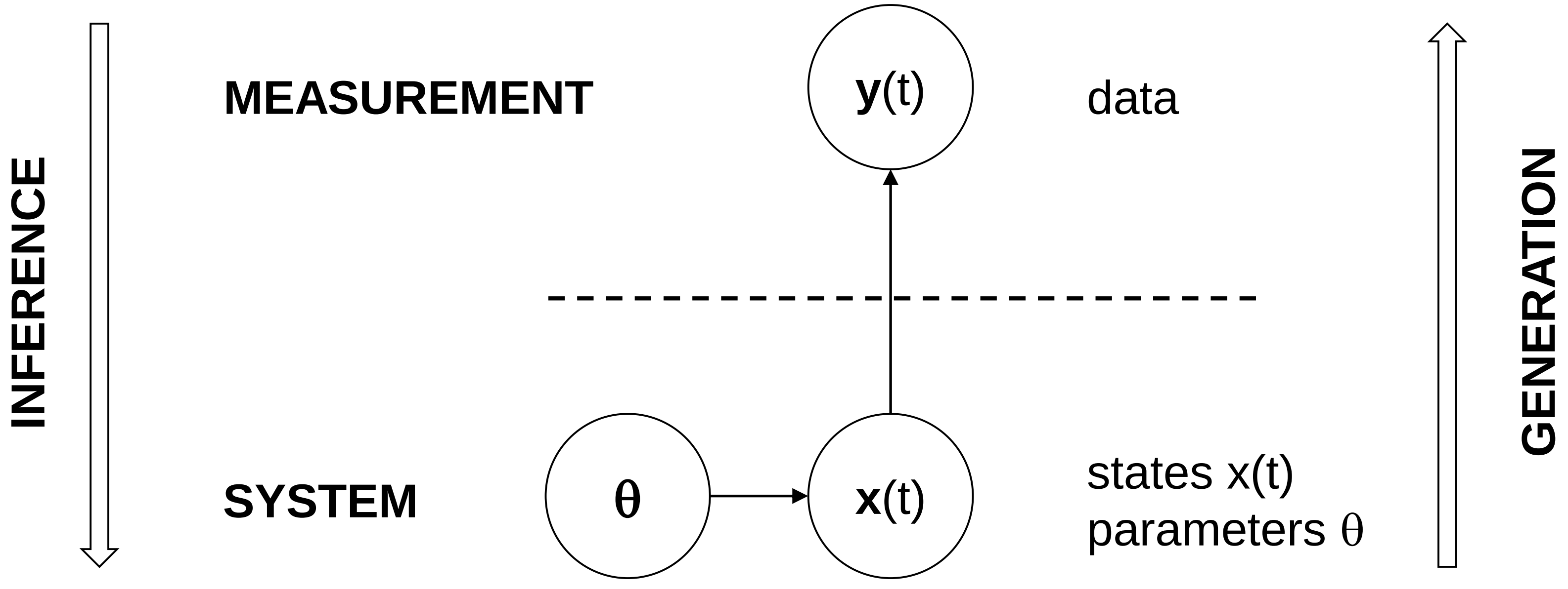
- central concept: hidden (latent) system states cause noisy measurements

- forward models that combine three things:
  - how system states evolve in time
  - how states determine system outputs
  - how outputs are corrupted by measurement noise



Forward modeling

- many ways to categorise modeling approaches
- one possibility: distinguish presence vs. absence of a forward model



States, parameters, inputs

- mandatory system components:
  - what are the relevant variables whose dynamics are of interest? → **states** $\mathbf{x}(t)$
  - what are structural determinants of their interactions? → **parameters** $\boldsymbol{\theta}$
  - what perturbations need to be considered? → **inputs** $\mathbf{u}(t)$

- system states:

state vector

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix}$$

neurophysiological or
algorithmic variables

state (or evolution) equations, e.g.:

$$\frac{d\mathbf{x}}{dt} = f\left(\mathbf{x}(t), \boldsymbol{\theta}_f, \mathbf{u}(t)\right)$$ as differential equation

$$\mathbf{x}(t+1) = f\left(\mathbf{x}(t), \boldsymbol{\theta}_f, \mathbf{u}(t)\right)$$ as difference equation

For a discussion of system theory in the context of neuroimaging, see Stephan 2004, *J. Anat.*

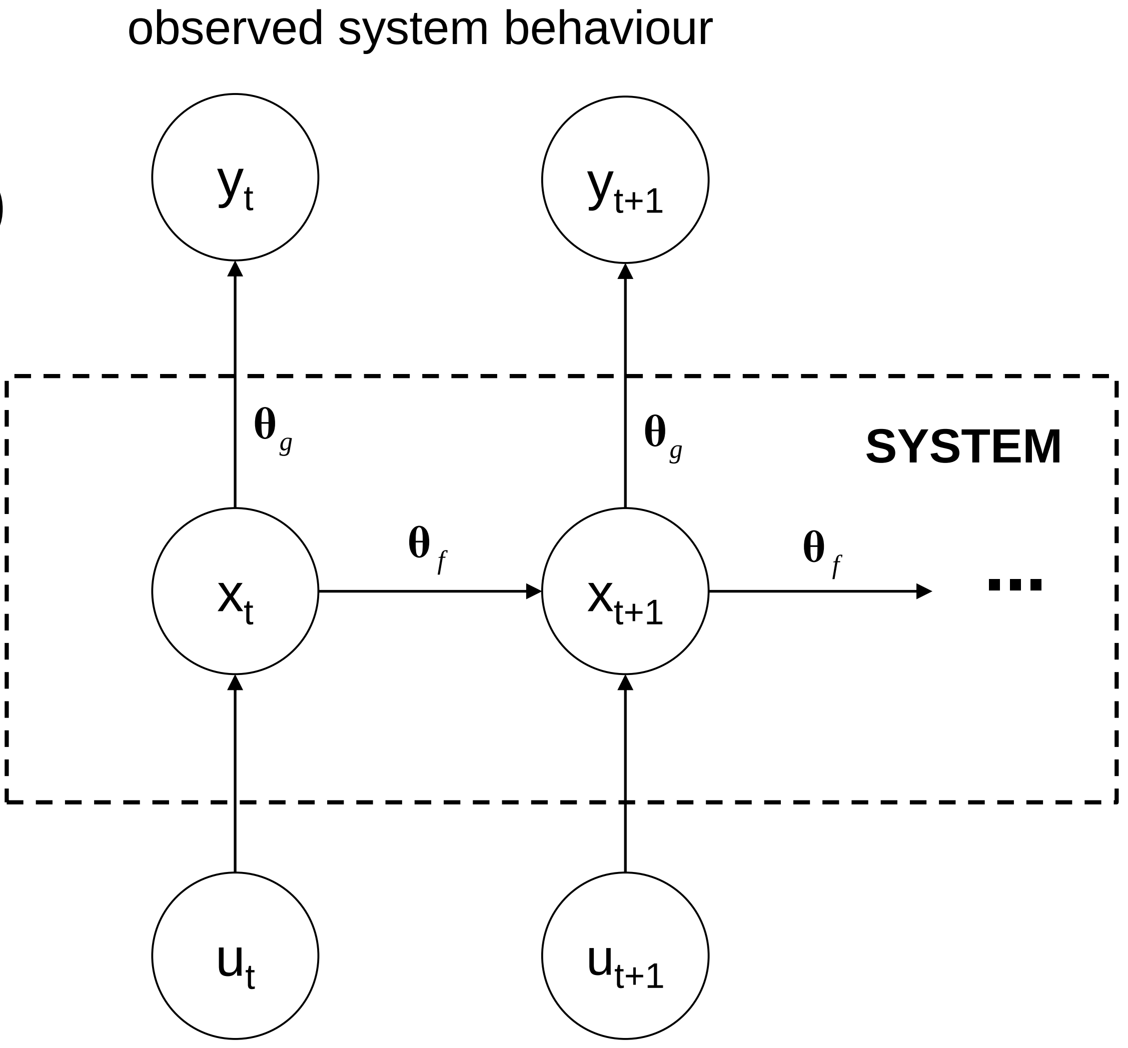
State space representation

measurement equation:

$$\mathbf{y}(t) = g\left(\mathbf{x}(t), \boldsymbol{\theta}_g\right) + \boldsymbol{\varepsilon}(t)$$

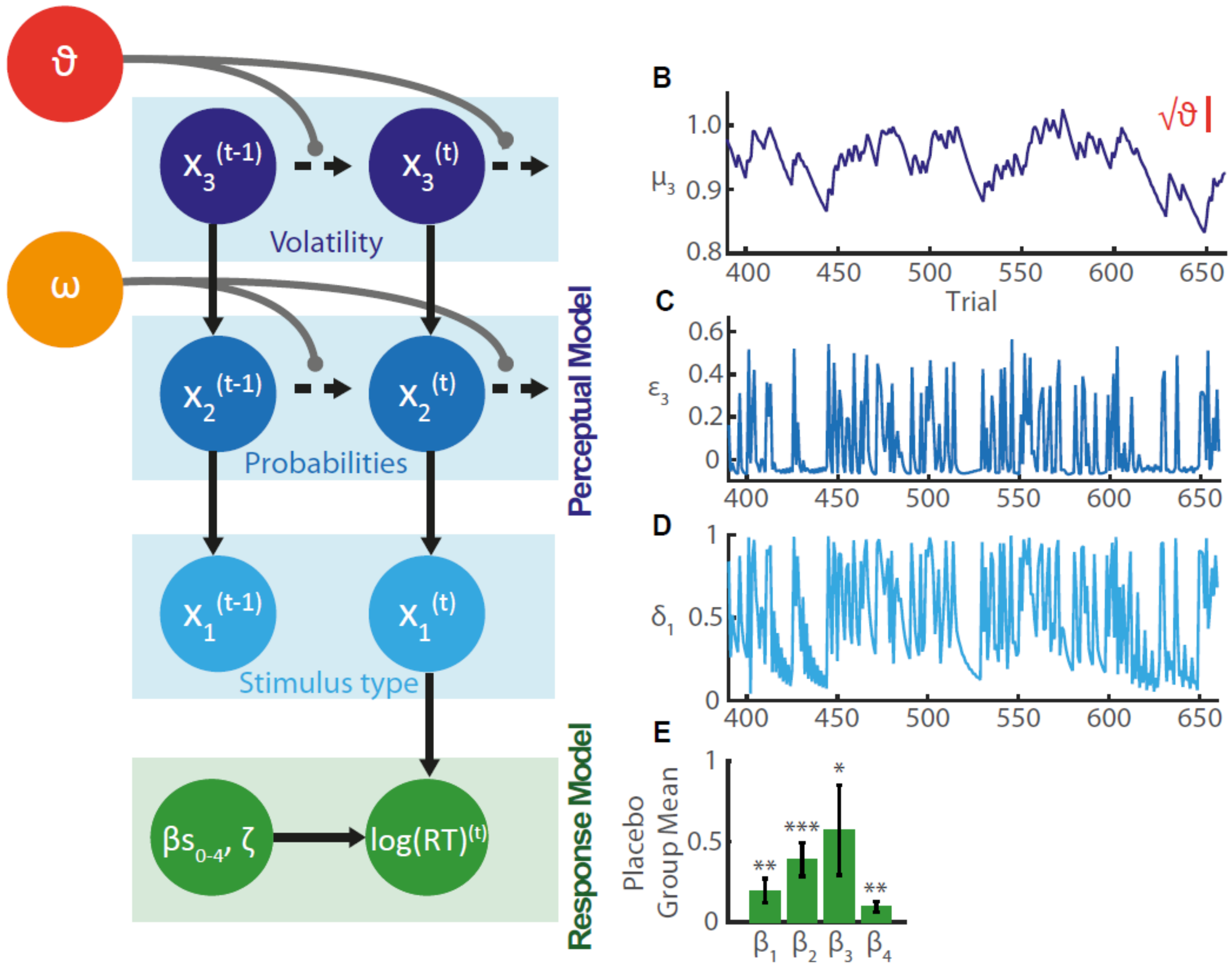
ENVIRONMENT

inputs

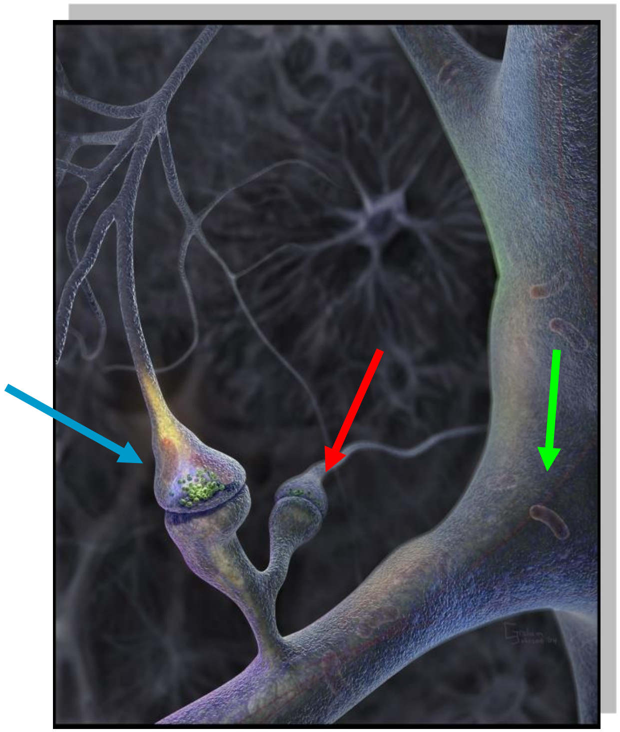


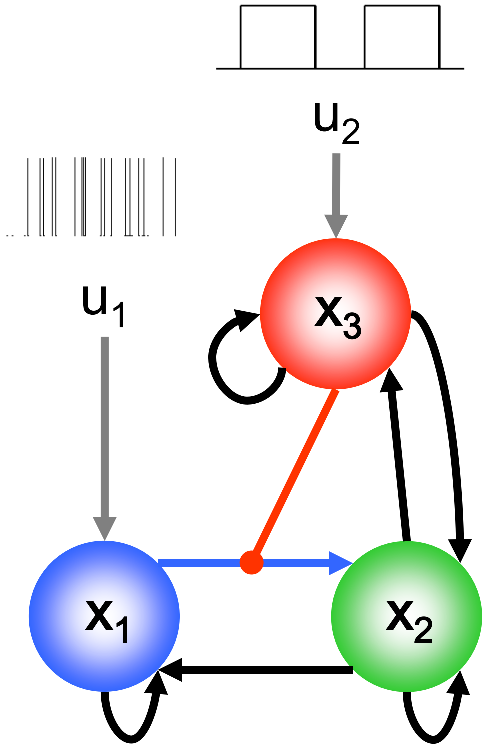
On this slide, time is indexed by subscripts.

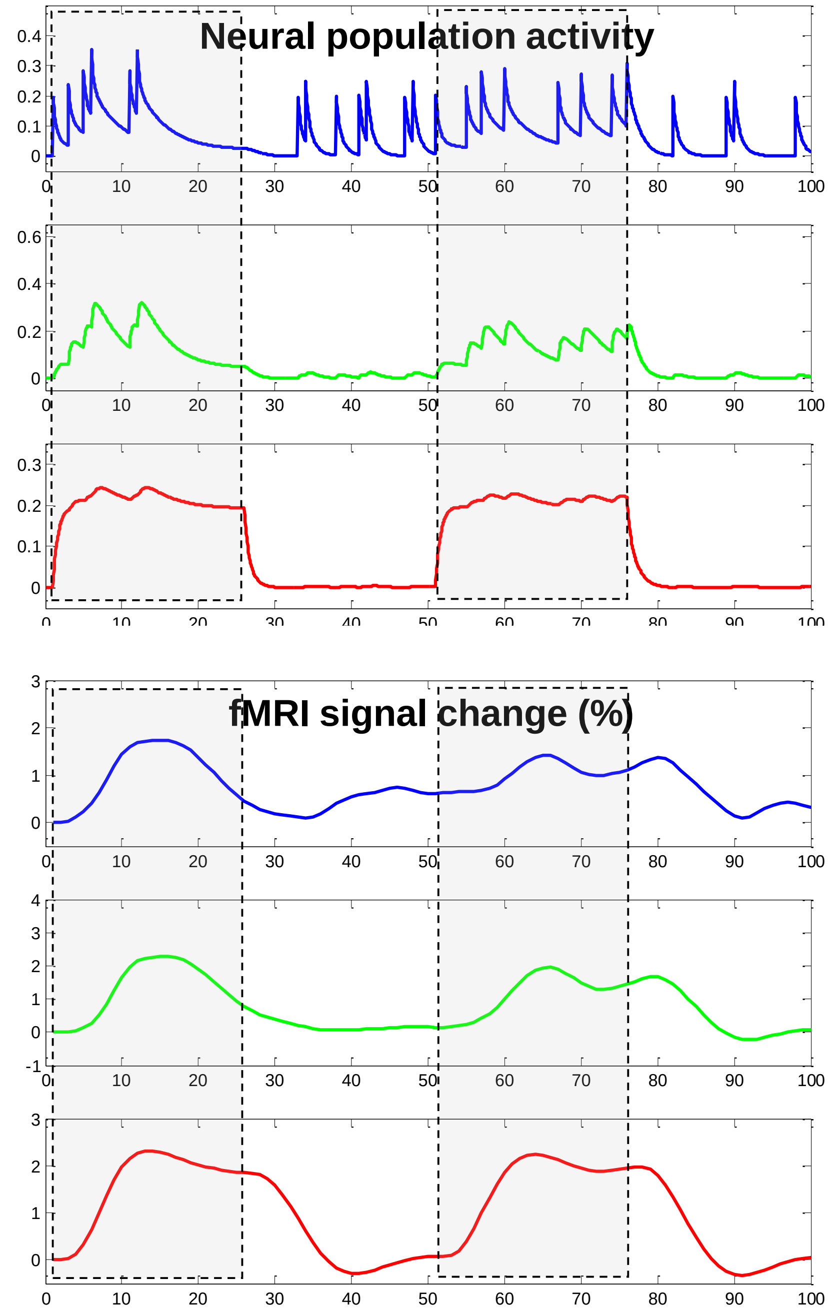
Examples of models discussed later in the course: HGF...



Marshall, Mathys et al. 2016, *PLoS Biology*







... and nonlinear DCM for fMRI

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

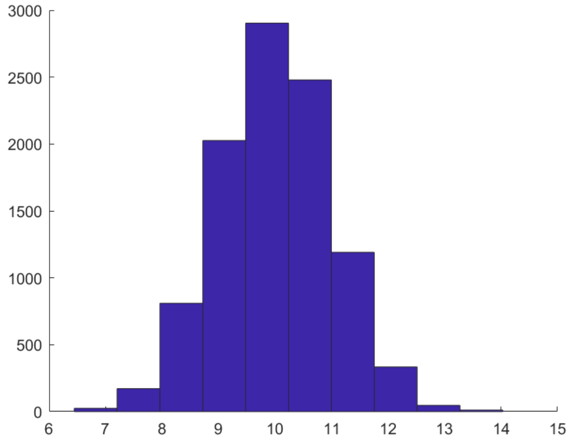
Stephan et al. 2008, *NeuroImage*

Statistical interlude: random variables/vectors

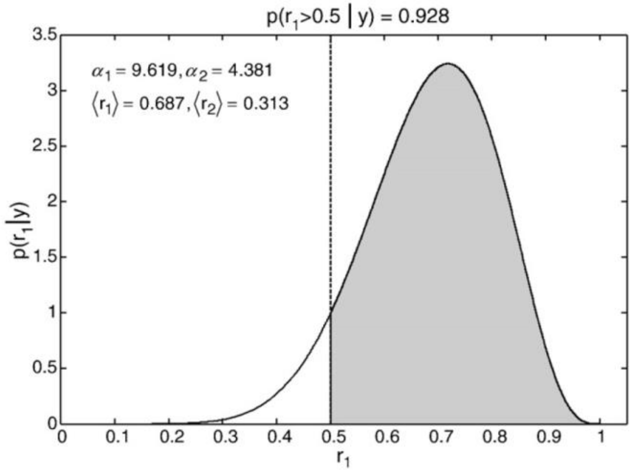
- **random variable**: a variable whose possible values are outcomes of a random phenomenon
- **random vector:** a vector of random variables

Statistical interlude: probability distributions and densities

- **probability distribution:**
  - describes the probability that a **discrete** random variable takes on a particular value



- **probability density:**
  - describes the probability of a **continuous** random variable falling within a particular range of values



Statistical interlude: probability distributions and densities

- notation example (Normal densities):
  - for scalars: $p(x) = N(x; \mu, \sigma^2)$ μ = mean; σ^2 = variance
  - for vectors: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ Σ = covariance matrix
    = E[$(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$]

Statistical interlude: probability distributions and densities

- notation example (Normal densities):
  - for scalars: $p(x) = N\left(x; \mu, \lambda^{-1}\right)$ μ = mean; $\lambda = 1/\sigma^2$ = precision
  - for vectors: $p(\mathbf{x}) = N\left(\mathbf{x}; \boldsymbol{\mu}, \Lambda^{-1}\right)$ Λ = precision matrix

same thing, just expressed wrt. precision

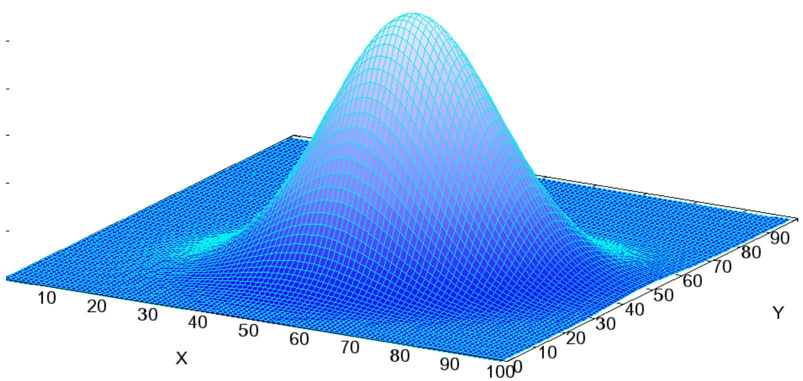
Statistical interlude: multivariate Gaussian/Normal

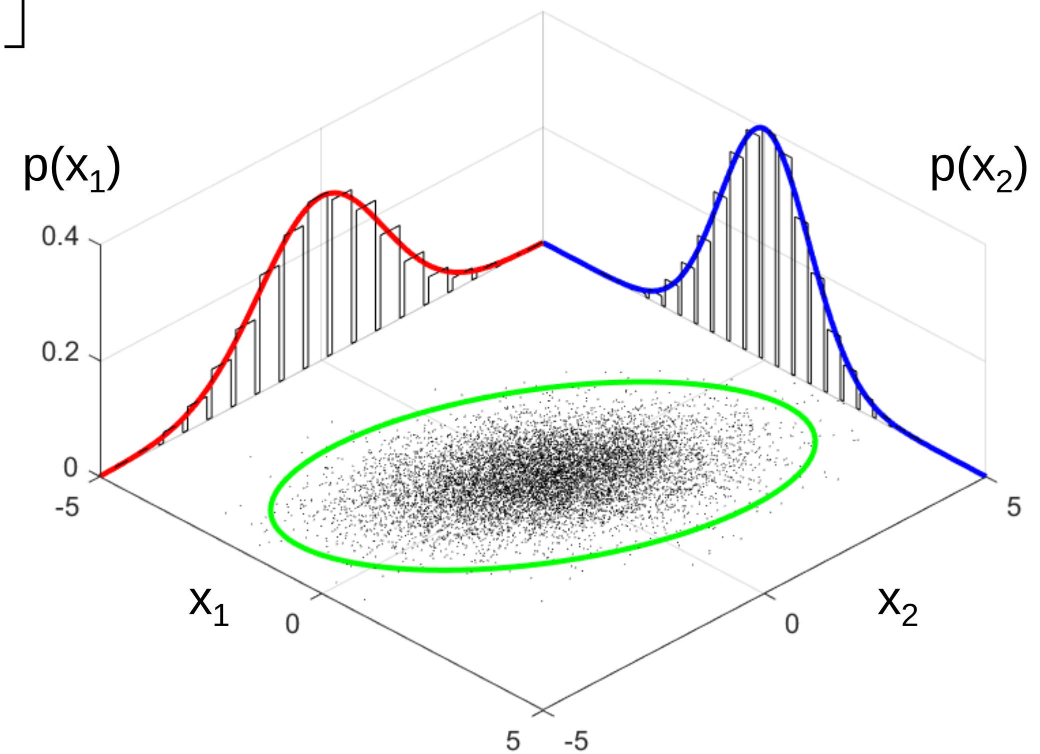
p-dimensional random vector: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

PDF: $$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

covariance matrix: $$\boldsymbol{\Sigma} = \mathbf{E}\left[\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right) \right]$$





Figures adapted from *Wikipedia*

Signal-generating equations (forward model) → likelihood

- State (evolution) equation *

$$\mathbf{x}(t+1) = f\left(\mathbf{x}(t), \boldsymbol{\theta}, \mathbf{u}(t)\right)$$

- Measurement (observation) equation

$$\mathbf{y}(t) = g\left(\mathbf{x}(t), \boldsymbol{\theta}\right) + \boldsymbol{\varepsilon}(t)$$

- Assuming IID Gaussian noise, write the (known) data as a probabilistic function of the (unknown) parameters:

$$\varepsilon = N\left(\varepsilon; 0, \sigma^2\right)$$

$$p\left(\mathbf{y} \mid \boldsymbol{\theta}\right) = N\left(\mathbf{y}; g\left(\mathbf{x}, \boldsymbol{\theta}\right), \sigma^2 \mathbf{I}\right)$$

- This turns our forward model into a probability statement:
  the **likelihood** of the observed data $\mathbf{y}$, given any particular value of $\boldsymbol{\theta}$.

* For simplicity, we assume deterministic state equations (no state noise) and absorb all parameters into a single vector $\theta=\{\theta_f,\theta_g\}$.

Maximum likelihood estimation (MLE)

- For any particular value of $\boldsymbol{\theta}$, we can refer to the definition of a multivariate Gaussian to compute the **likelihood of the entire dataset $\mathbf{Y}$** (all system nodes, all time points):

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} \left(\mathbf{y} - g(\mathbf{x}, \boldsymbol{\theta})\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - g(\mathbf{x}, \boldsymbol{\theta})\right) \right)$$

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = p\left(\mathbf{y}(1), \ldots, \mathbf{y}(T) \mid \boldsymbol{\theta}\right) = \prod_{t=1}^{T} p\left(\mathbf{y}(t) \mid \boldsymbol{\theta}\right)$$

- We could now search for the parameter value that maximises the likelihood (or, for numerical reasons, the log likelihood), or put simply: the parameter value for which the model fits the data best.
  This is known as **maximum likelihood estimation (MLE)**:

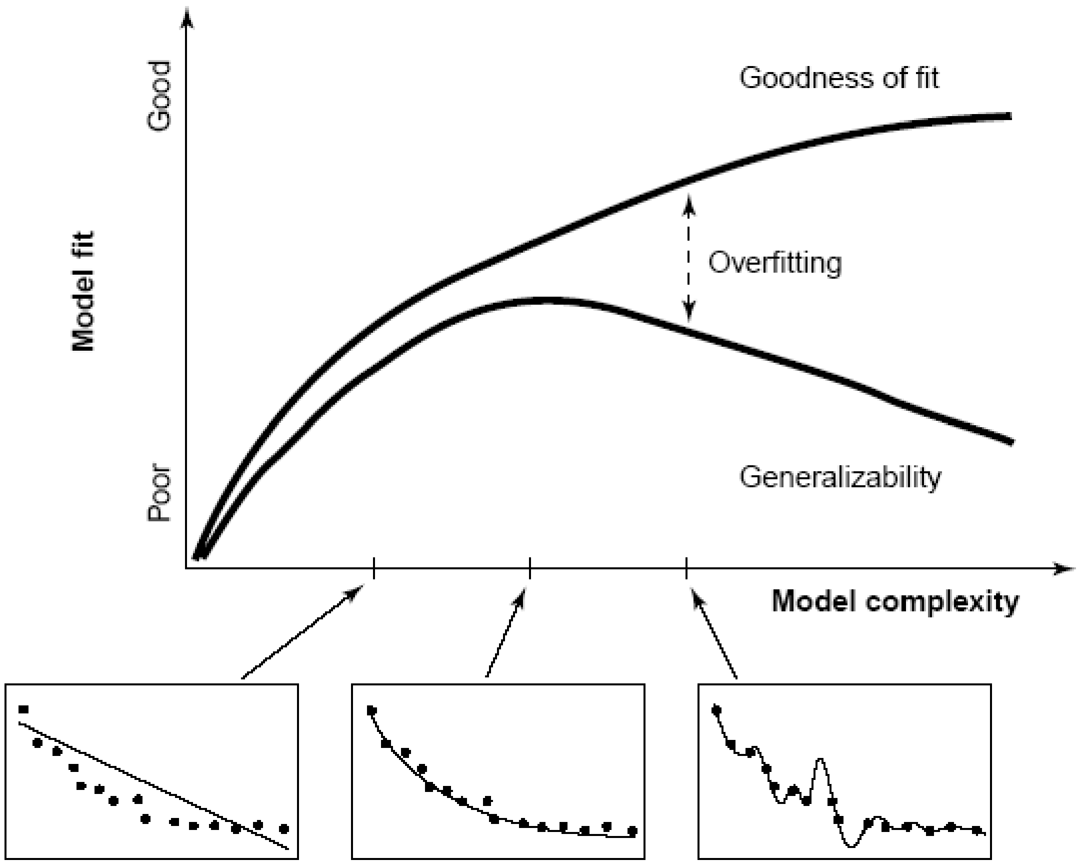
$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\theta} \ln p(\mathbf{Y} \mid \boldsymbol{\theta})$$

Overfitting

- MLE has various limitations. For example, for complex models and limited data, **overfitting** is a severe problem.

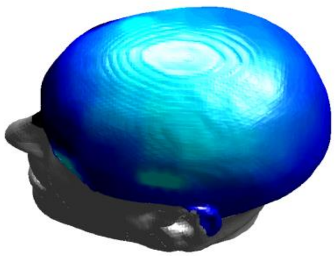
- For more robust inference, we turn to Bayesian methods
  → need to define a prior distribution of parameters

- Together, likelihood and prior define a **generative model**.



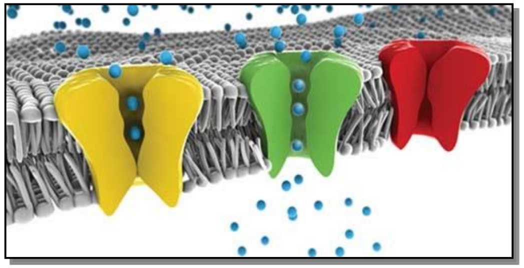
Pitt & Myung (2002) *TICS*

Generative models



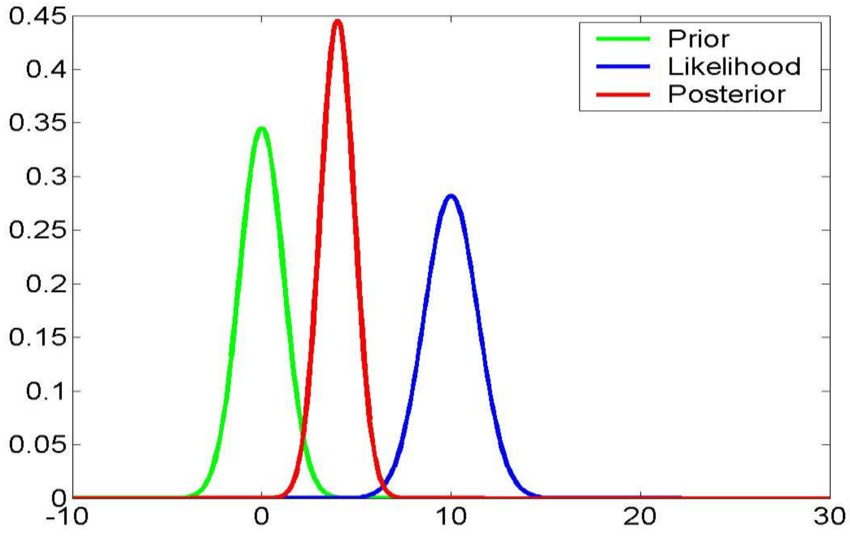
$$p(\mathbf{y} \mid \boldsymbol{\theta}, m)\, p(\boldsymbol{\theta} \mid m)$$

$$p(\boldsymbol{\theta} \mid \mathbf{y}, m)$$



1. a probabilistic forward mapping from parameters to data, defined by likelihood and prior
2. provide the joint probability of parameters and data
3. enforce mechanistic thinking: how could the data have been caused?
4. generate synthetic data (observations) by sampling from the prior – can model explain certain phenomena at all?
5. model inversion = inference about parameters → p(θ|y)

Bayes' theorem

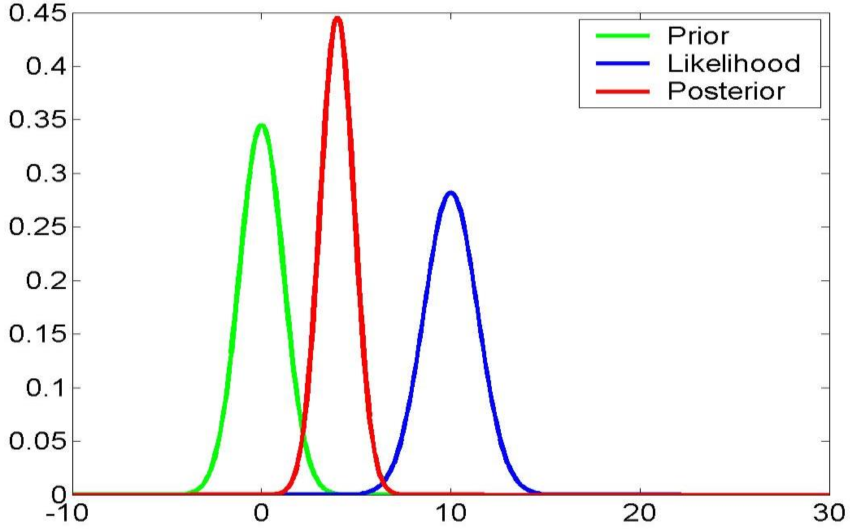


The Reverend Thomas Bayes
(1702-1761)

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

posterior = likelihood • prior / evidence

Bayes' theorem



The Reverend Thomas Bayes
(1702-1761)

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

posterior = likelihood • prior / evidence

Principles of generative modeling

⇨ Specifying a **generative model**

Model → likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$

Model → prior distribution $p(\boldsymbol{\theta})$

⇨ Observation of **data**

Measurement → data $\mathbf{y}$

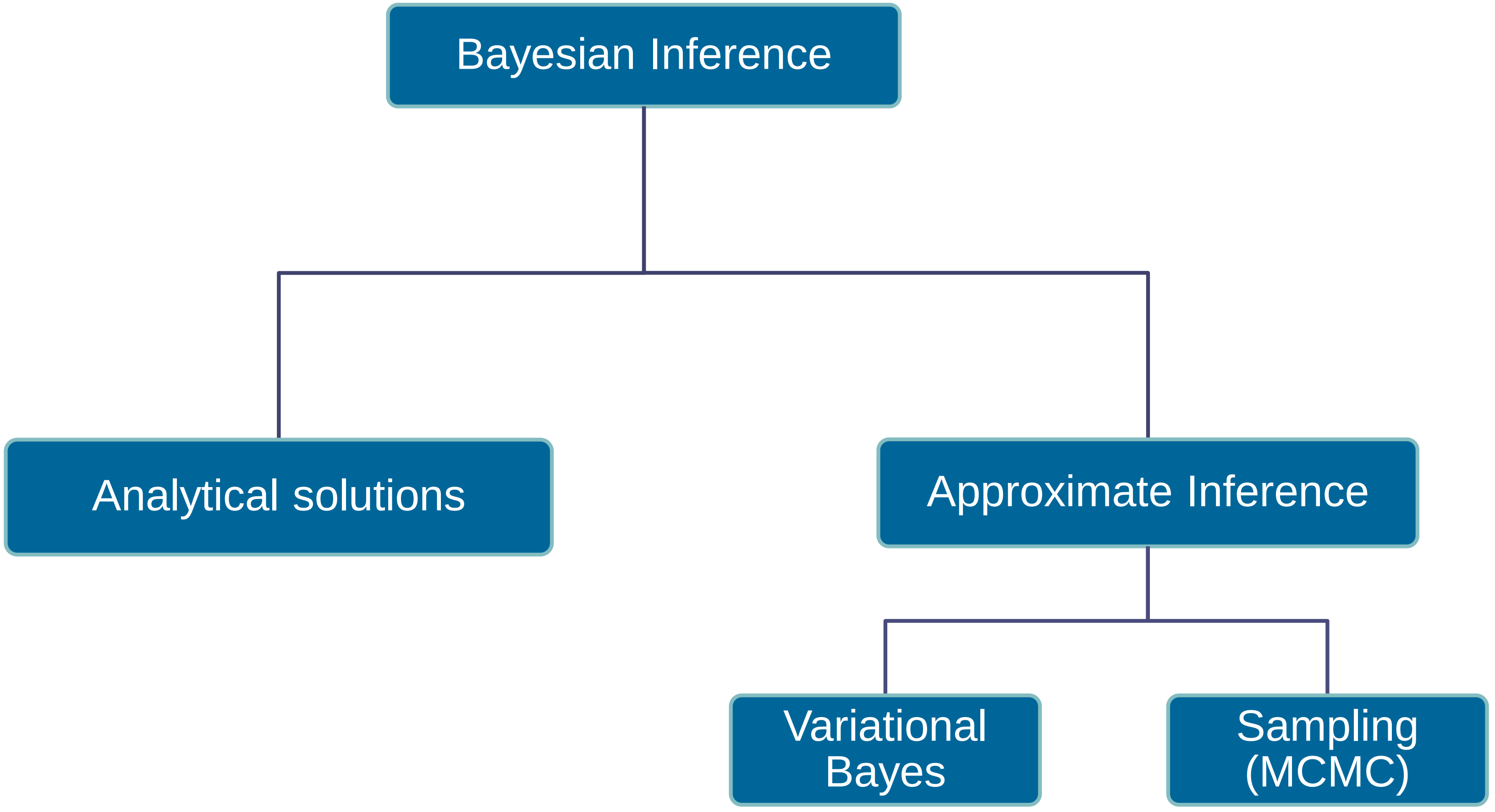
⇨ **Model inversion**

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto p(\mathbf{y}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})$$

→ posterior distribution

→ model evidence

Methods for model inversion



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

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

same form

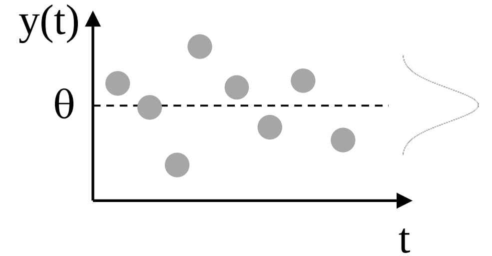
A simple example: univariate Gaussian belief update

Likelihood & prior

$$p(\mathbf{y} \mid \theta) = N\left(\theta, \lambda_e^{-1}\right)$$

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

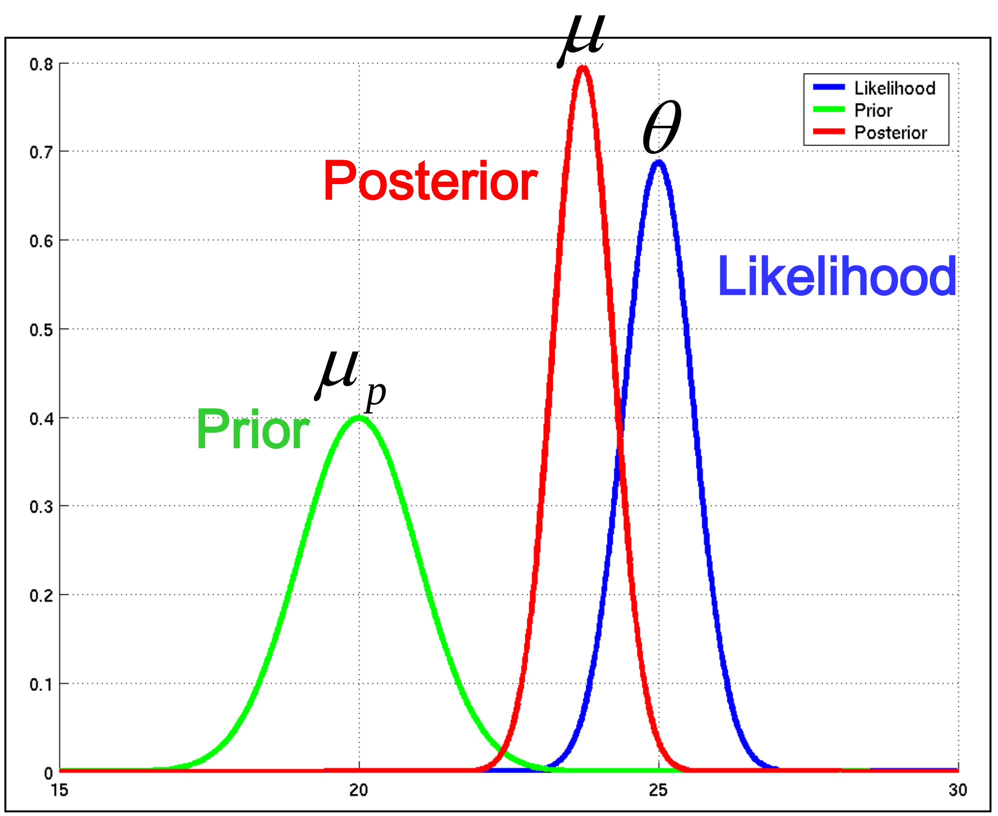
$$\mathbf{y} = \theta + \boldsymbol{\varepsilon}$$



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

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

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



relative precision weighting:
posterior mean = precision-weighted combination of prior mean and data mean

Model comparison and selection

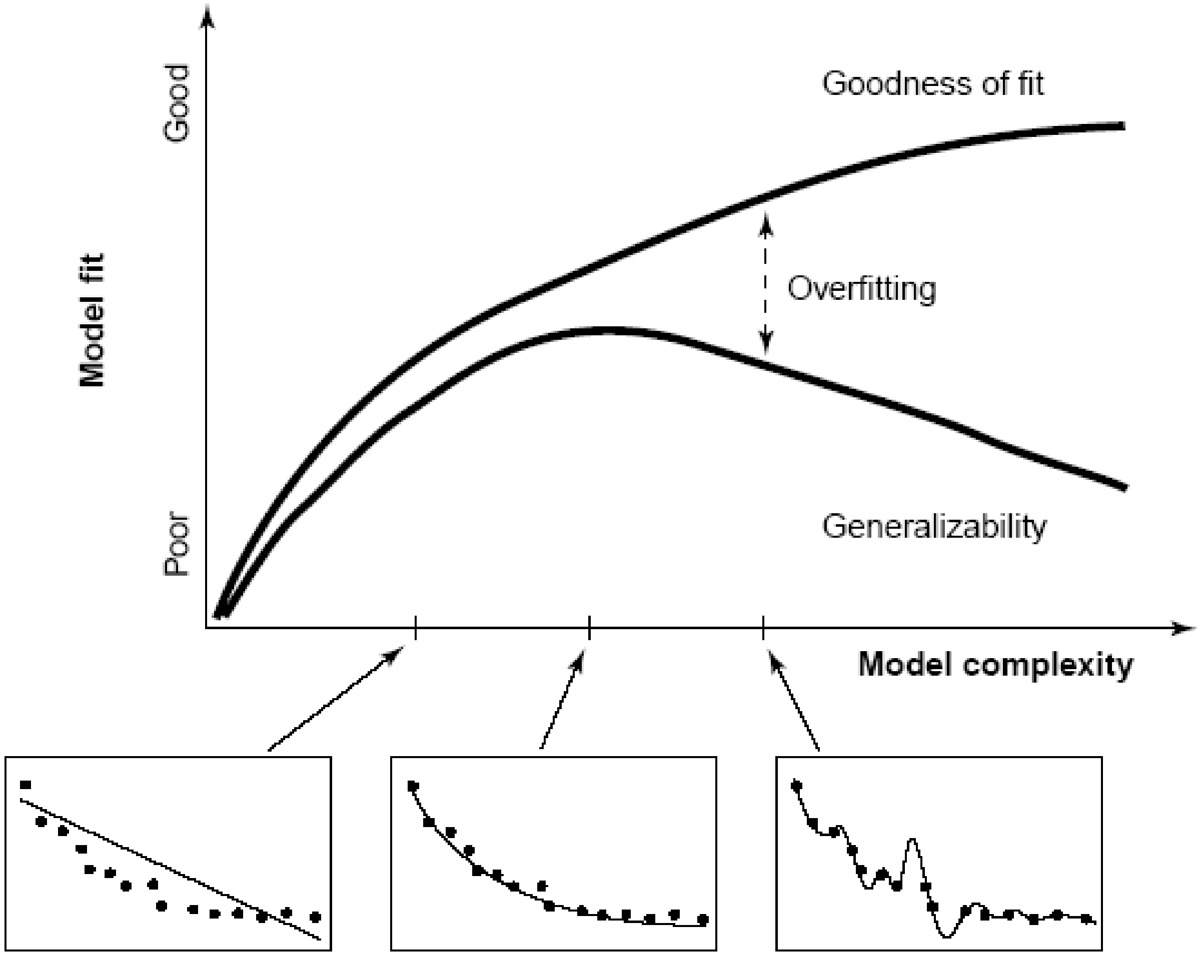
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

⇩

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)

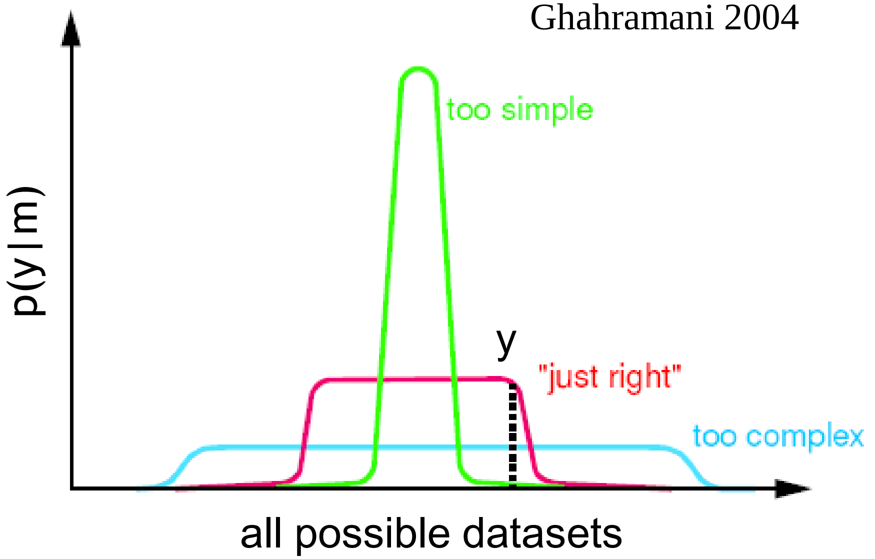
Model evidence:

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

⇨ 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



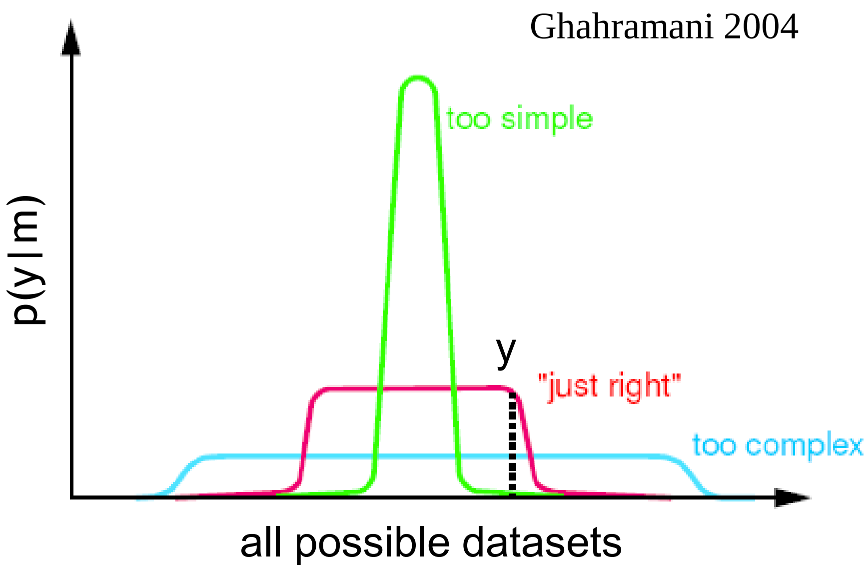
Various approximations:

- 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

Various approximations:

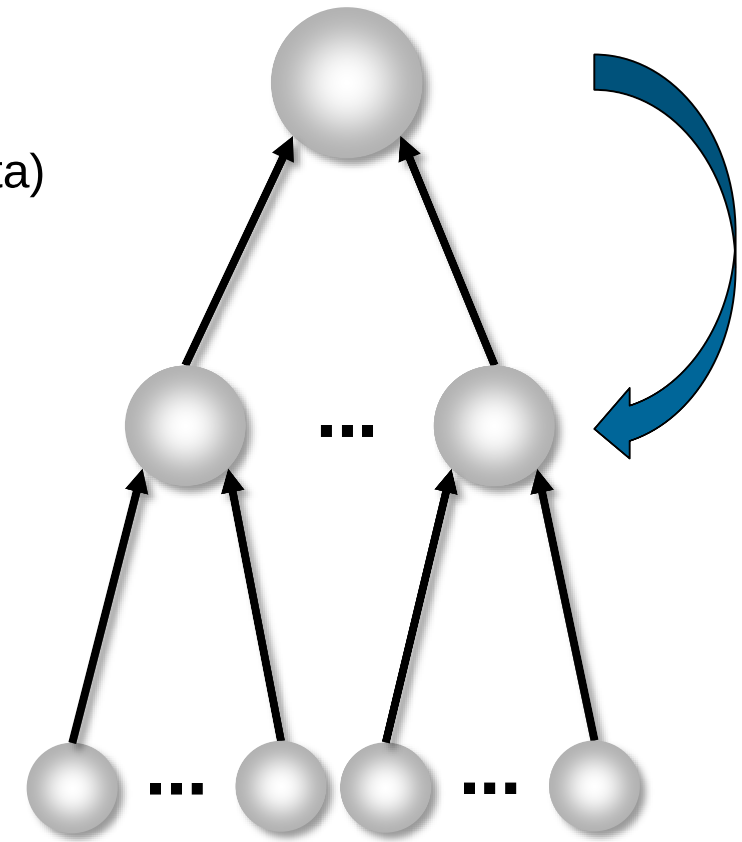
- negative free energy (F)
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

Generative models as computational assays for addressing key clinical questions

SYMPTOMS
(behavioural or physiological data)

MECHANISMS
(computational, physiological)

CAUSES
(aetiology)



❶ **differential diagnosis** of alternative disease mechanisms

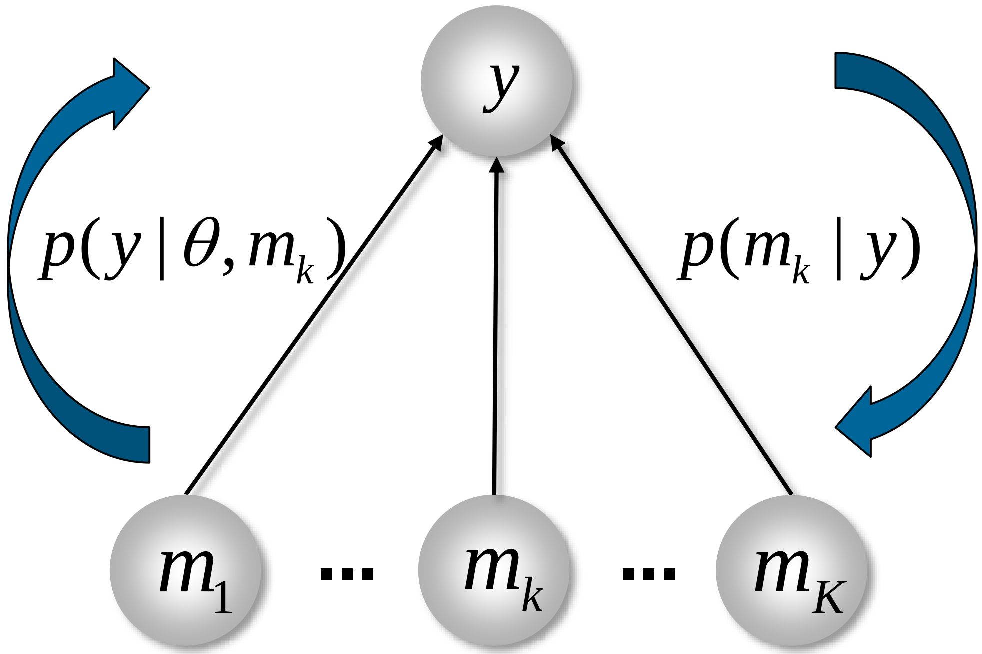
❷ **stratification / subgroup detection** into mechanistically distinct subgroups

❸ **prediction** of clinical trajectories and treatment response

➊ Differential diagnosis: model selection

SYMPTOM
(behaviour
or physiology)

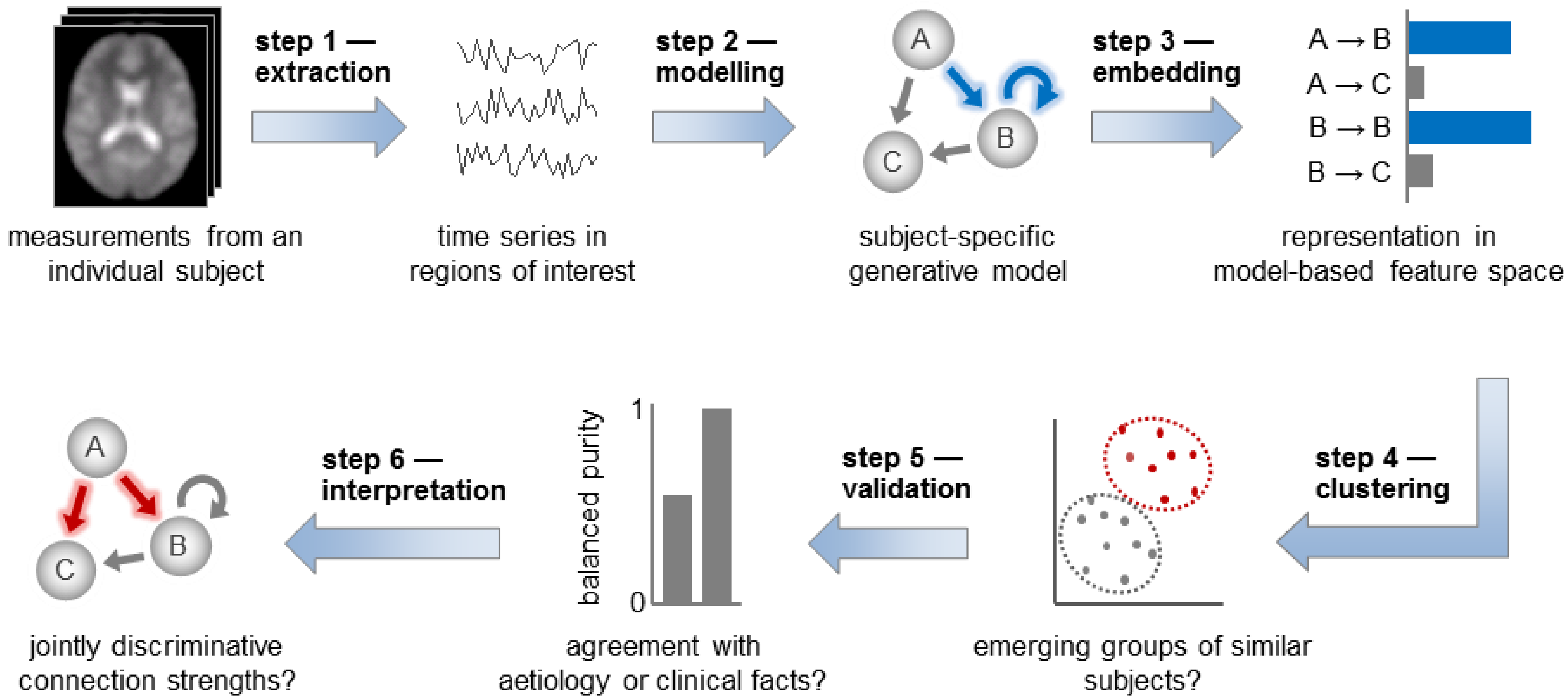
HYPOTHETICAL
MECHANISM



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

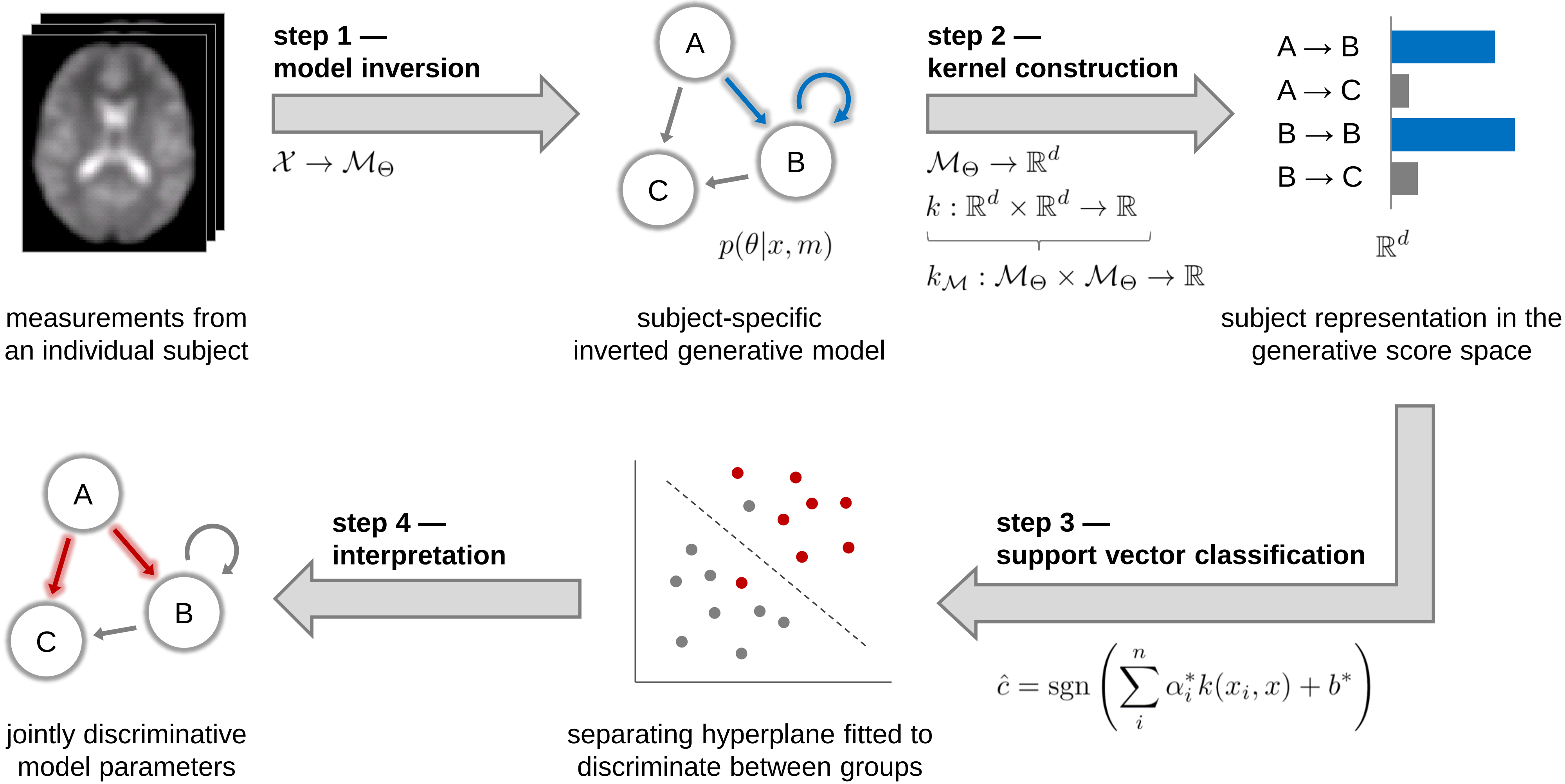
Stephan et al. 2017, *NeuroImage*

❷ Stratification / subgroup detection: Generative embedding (unsupervised)

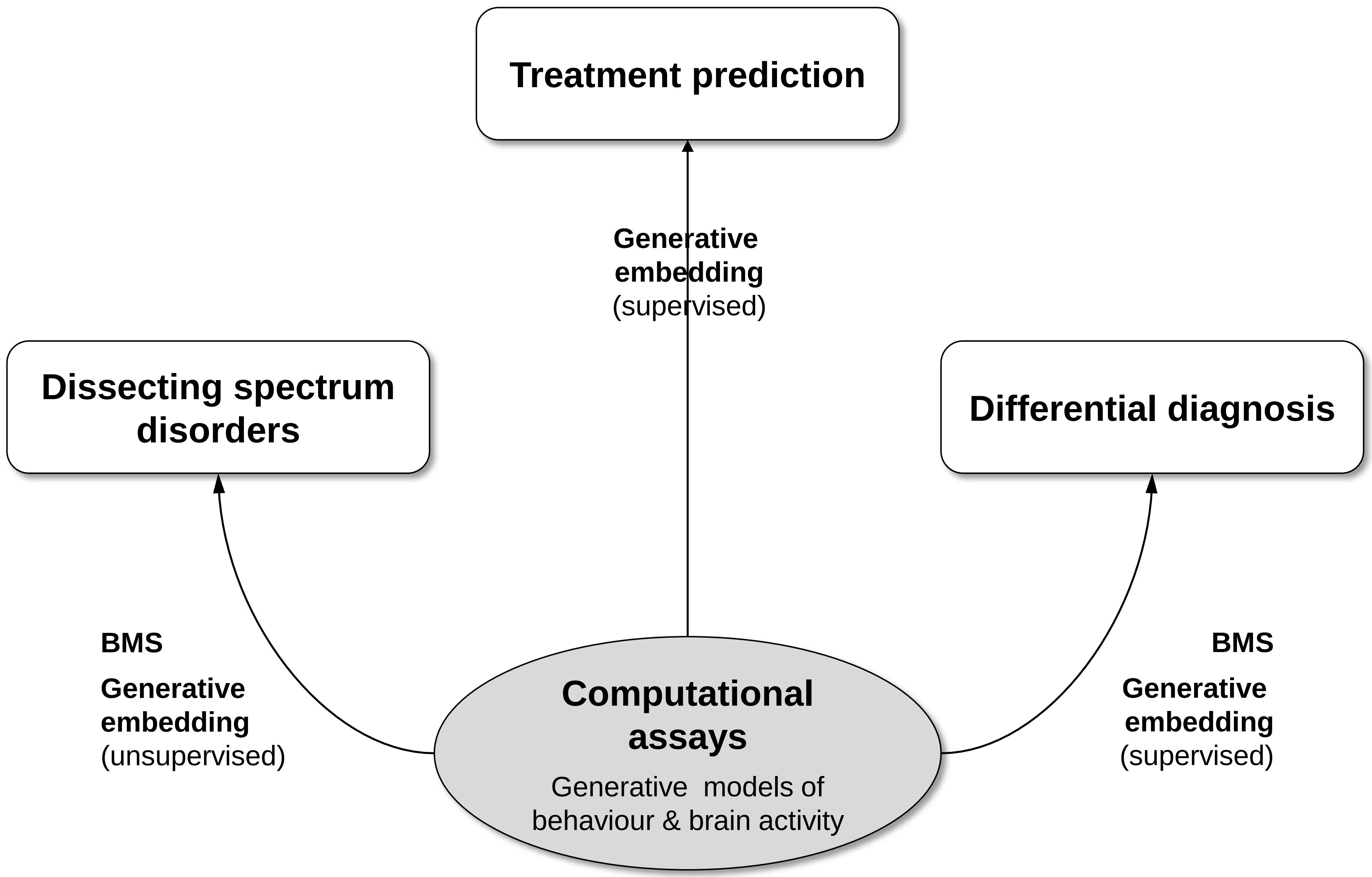


Brodersen et al. 2014, *NeuroImage Clinical*

❸ Prediction: Generative embedding (supervised)



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



Stephan & Mathys 2014, *Curr. Opin. Neurobiol.*

Further reading

- Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7, e1002079
- Brodersen, K.H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny,W.D., Buhmann, J.M., Stephan, K.E., 2014. Dissecting psychiatric spectrum disorders by generative embedding. Neuroimage Clin. 4, 98–111.
- Stephan KE (2004) On the role of general system theory for functional neuroimaging. Journal of Anatomy 205: 443-470.
- 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