Basic Concepts of Mathematical Modeling in Computational Psychiatry

Klaas Enno Stephan

Translational Neuromodeling Unit

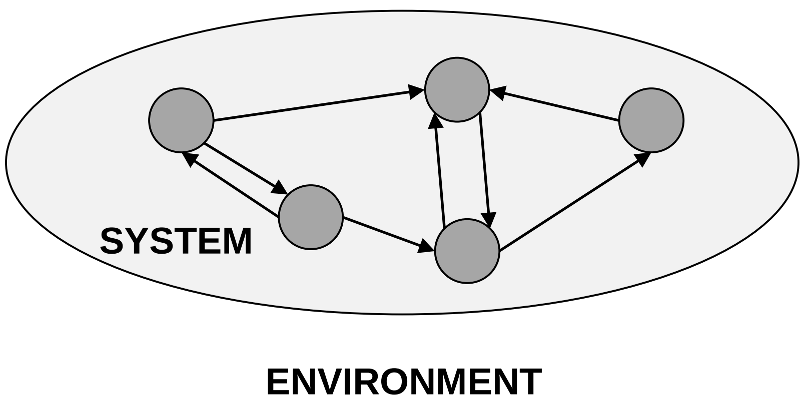
Universität Zürich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

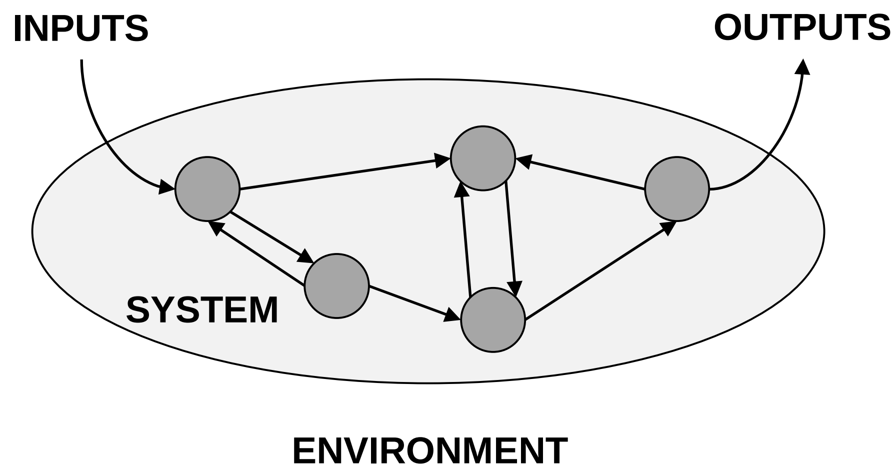
Systems

- 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)

closed system



open system

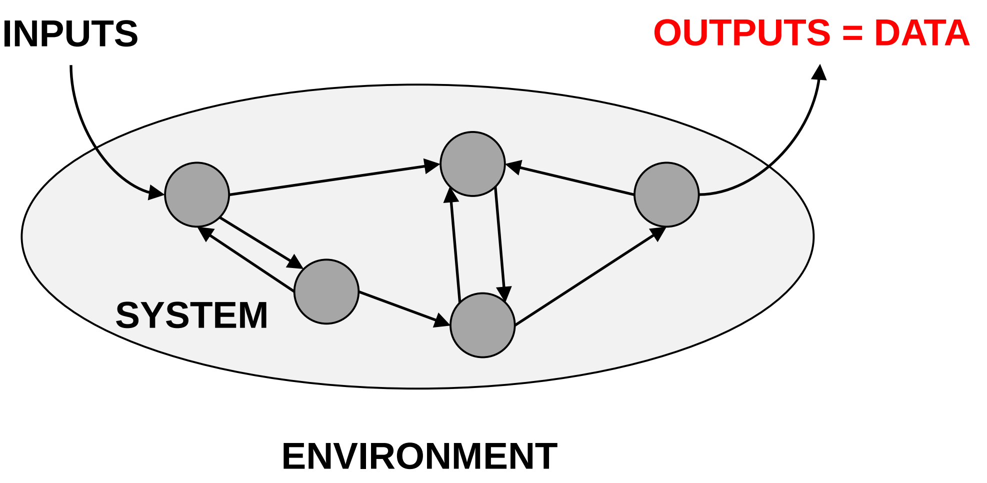


System models (state space models, latent process models)

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

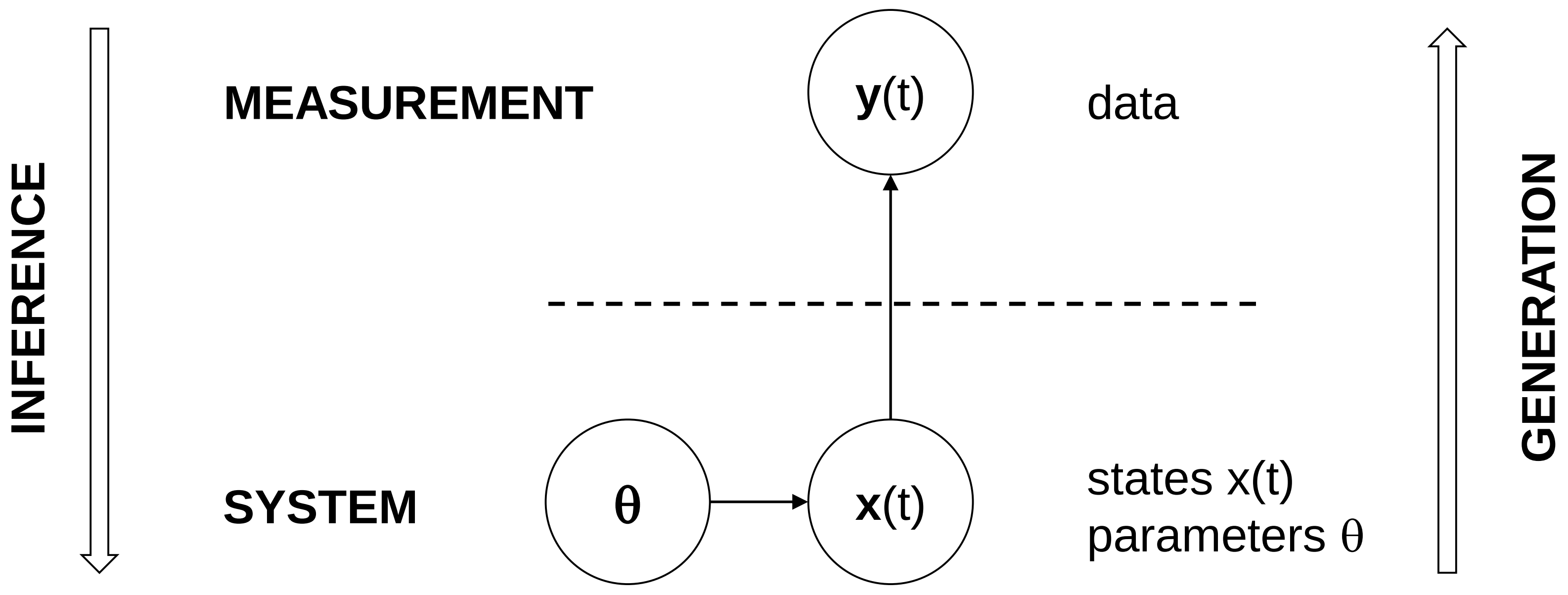
- central concept: hidden (latent) system states cause noisy measurements

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

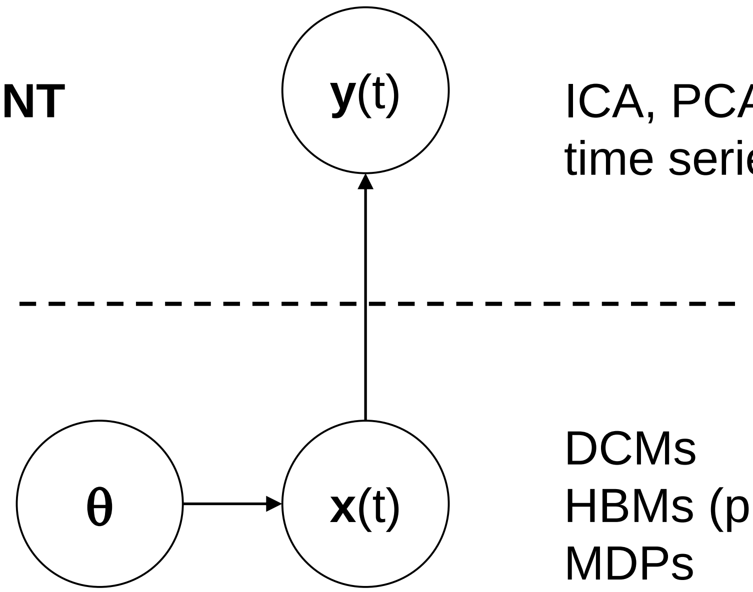


Forward modeling

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



Examples of approaches with/without forward modeling



MEASUREMENT

ICA, PCA
time series models

SYSTEM

DCMs
HBMs (predictive coding, HGF)
MDPs

States, parameters, inputs

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

- system states:

state vector

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

neurophysiological or algorithmic variables

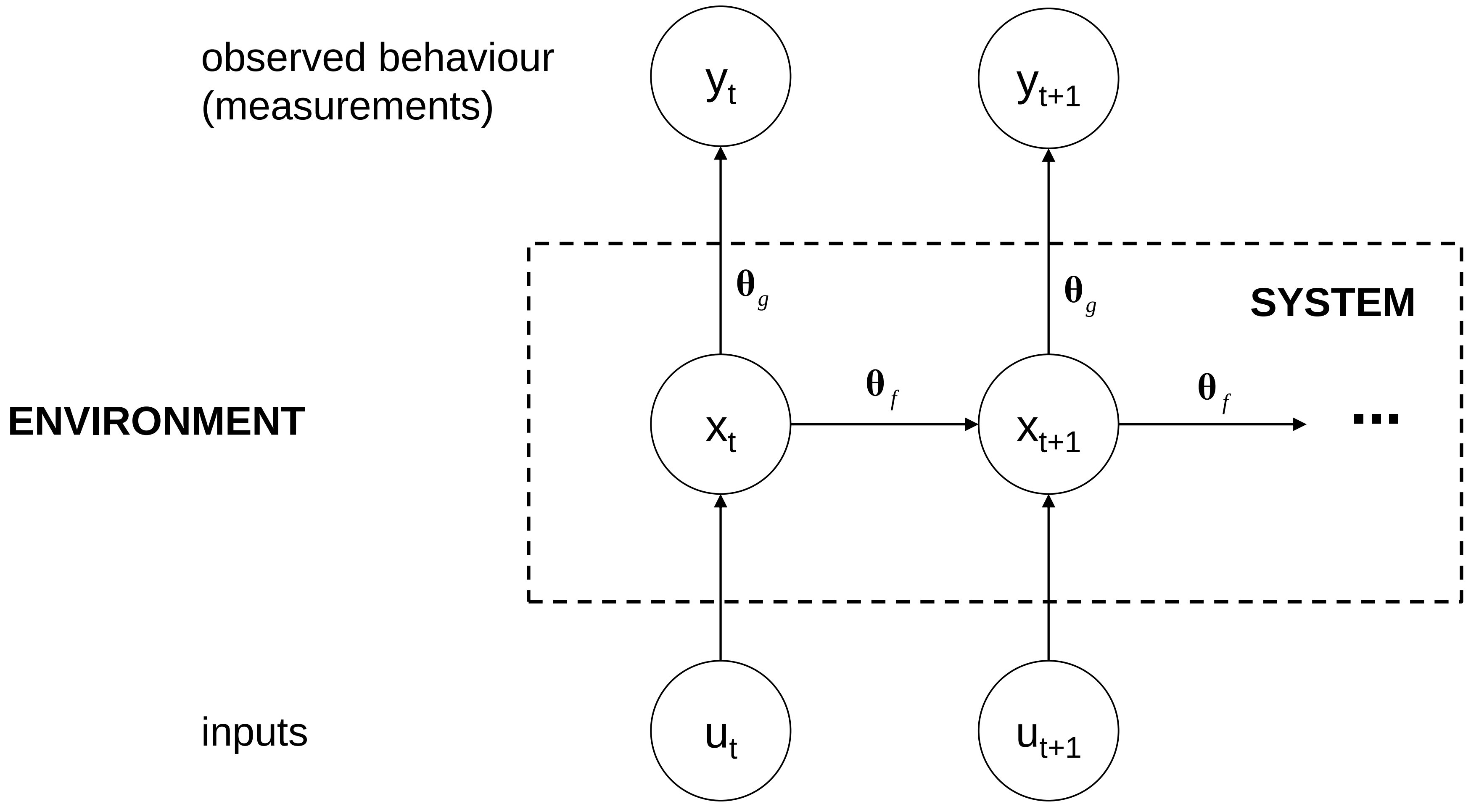
state (or evolution) equations, e.g.:

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

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

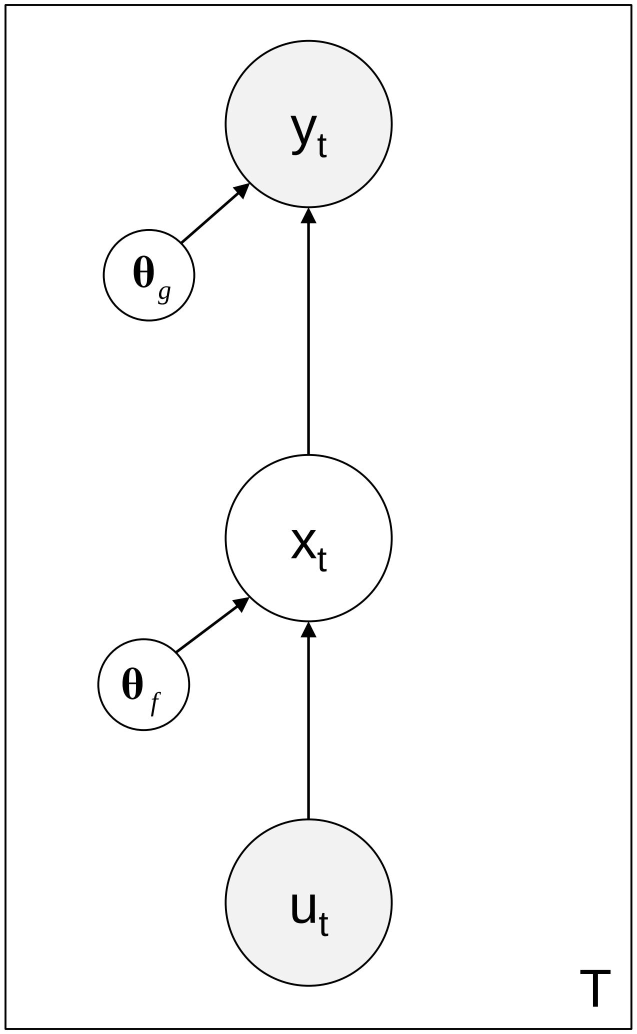
For discussion in the context of neuroimaging, see Stephan 2004, *J. Anat.*

State space representation

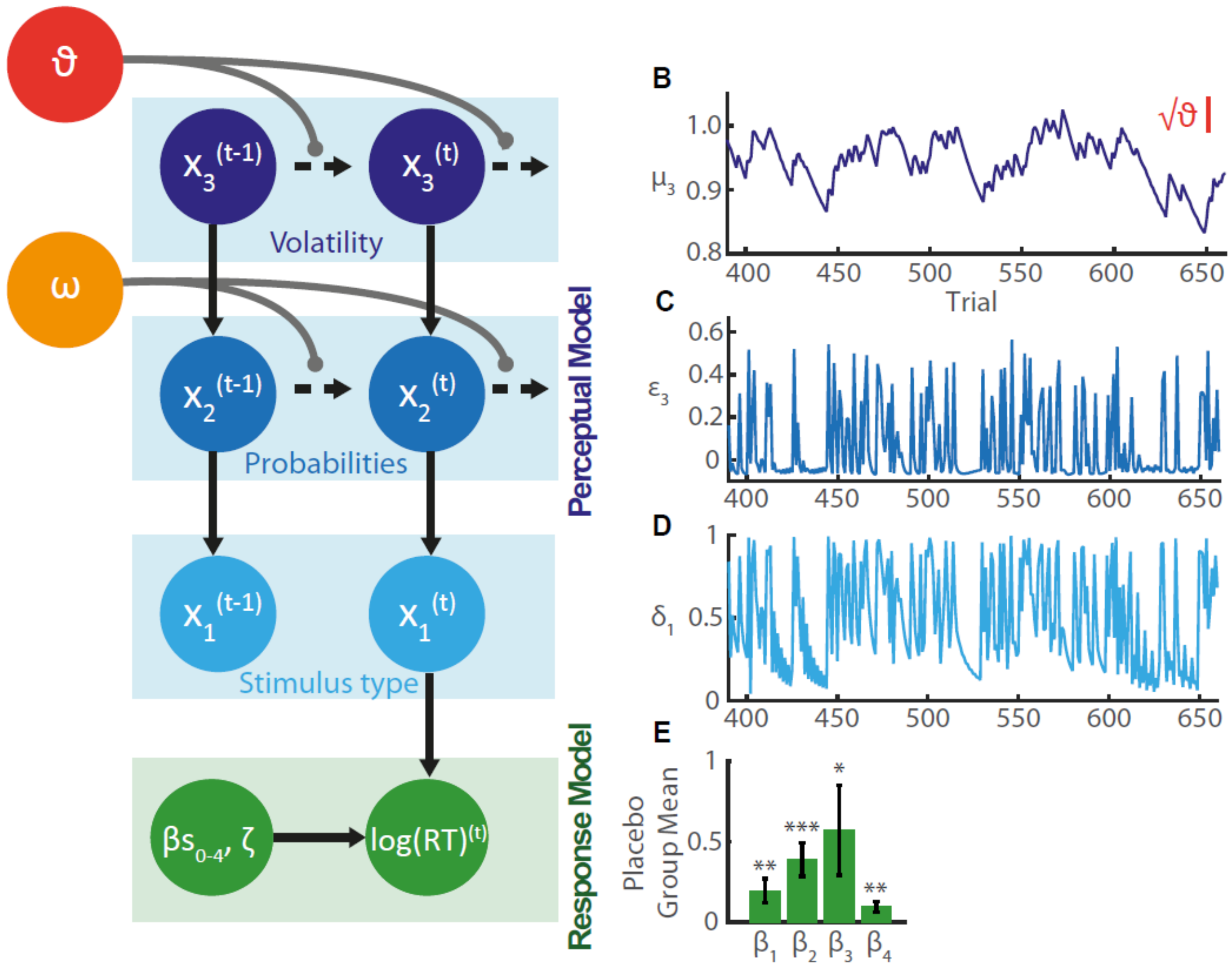


On this slide, time is indexed by subscripts.

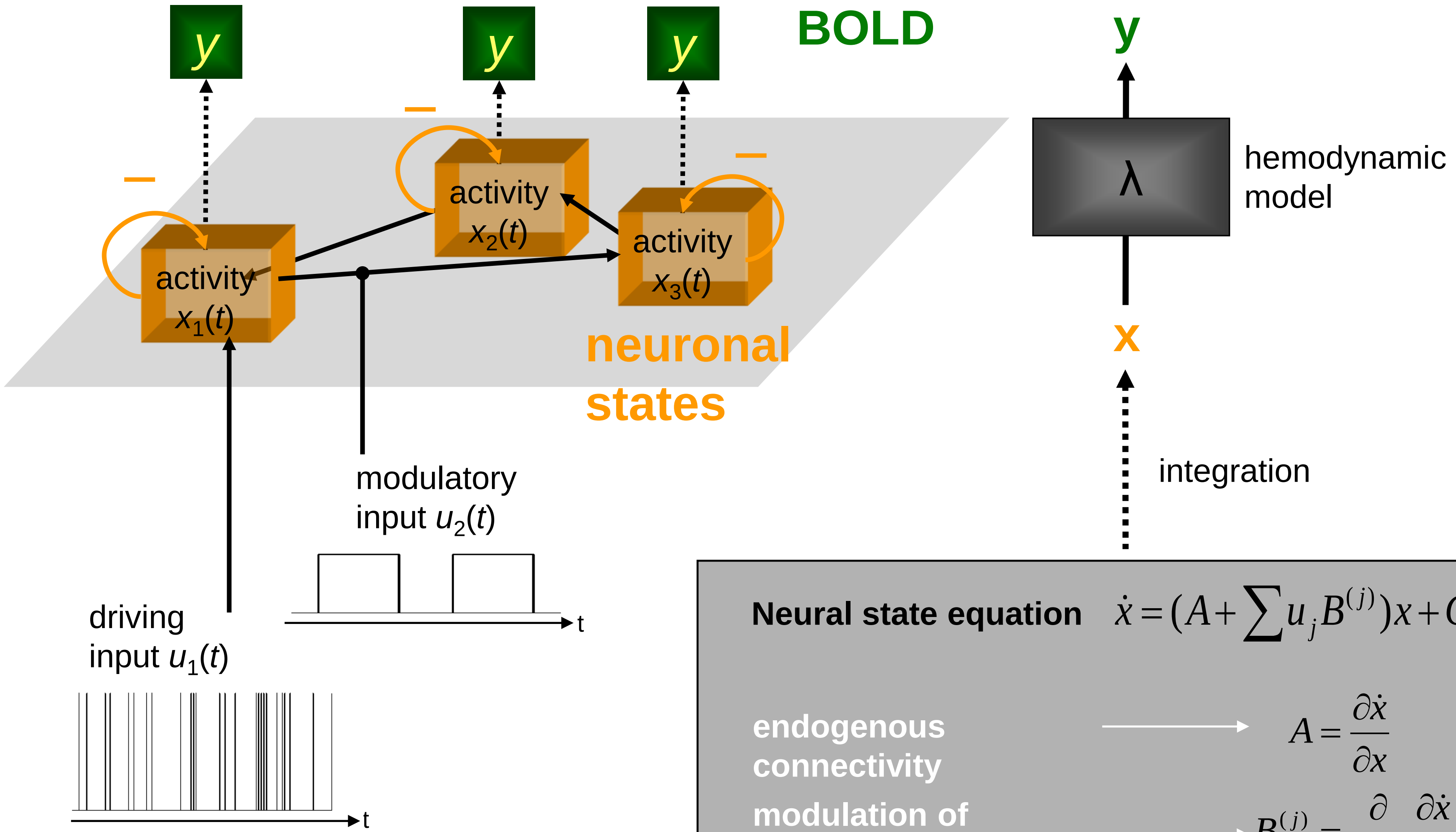
State space representation (graphical model)



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



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



... and DCM

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 → $C = \frac{\partial \dot{x}}{\partial u}$

Signal-generating equations (forward model)

- 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):

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

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

- We could now search for the parameter value that maximises the likelihood (or, for numerical reasons, typically the log likelihood). This is known as **maximum likelihood estimation (MLE)**:

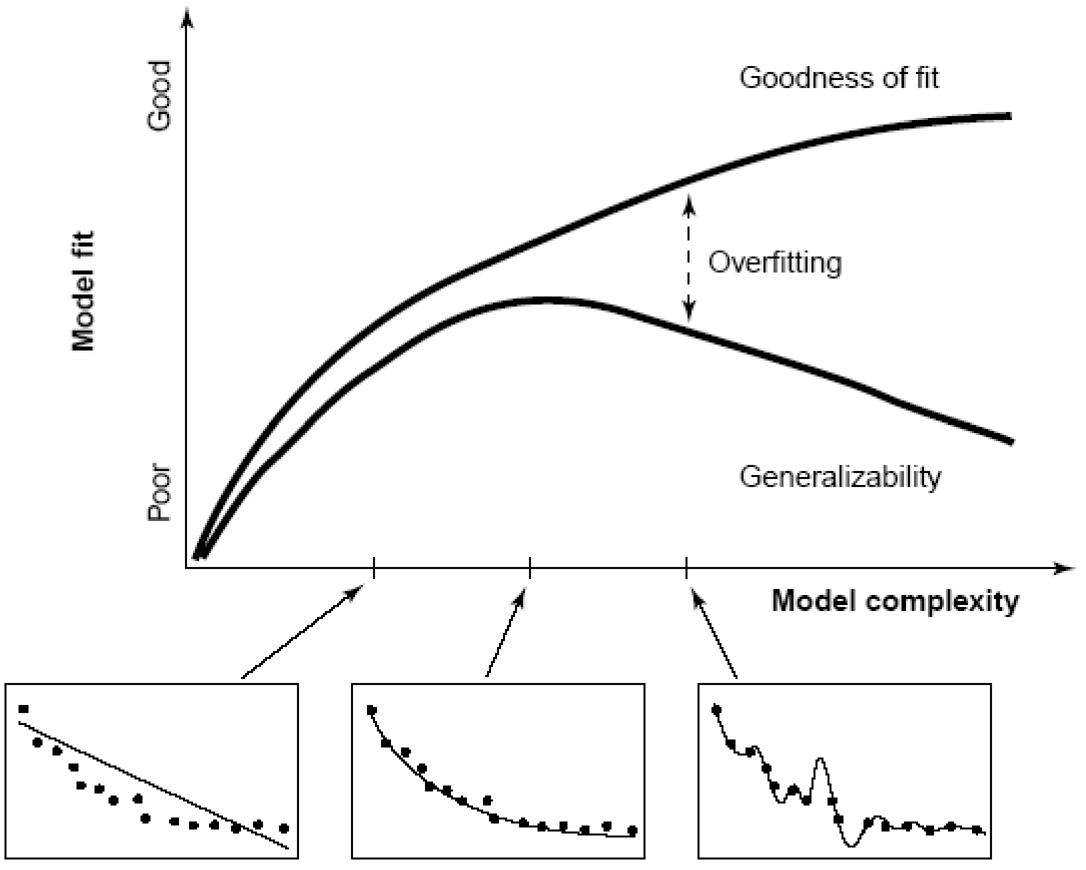
$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\theta} \ln p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right)$$

Overfitting

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

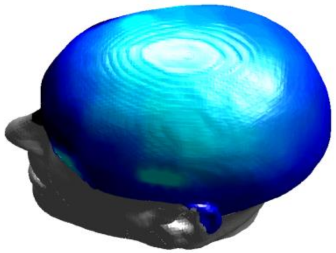
- For more robust inference,, we turn to Bayesian methods
  → need to define a prior distribution of parameters

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



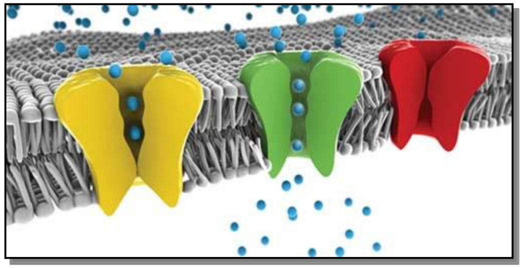
Pitt & Myung (2002) *TICS*

Generative models



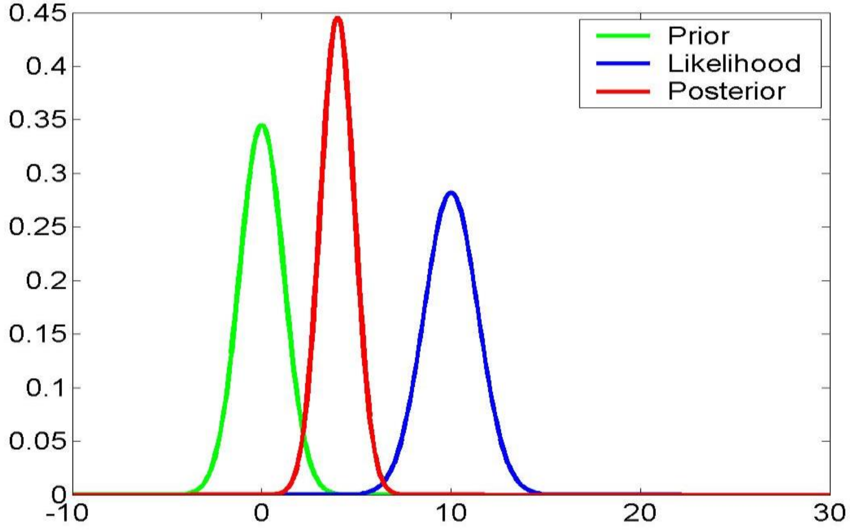
$$p(\mathbf{y} \mid \boldsymbol{\theta}, m)\, p(\boldsymbol{\theta} \mid m)$$

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



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

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 $\mathbf{y}$

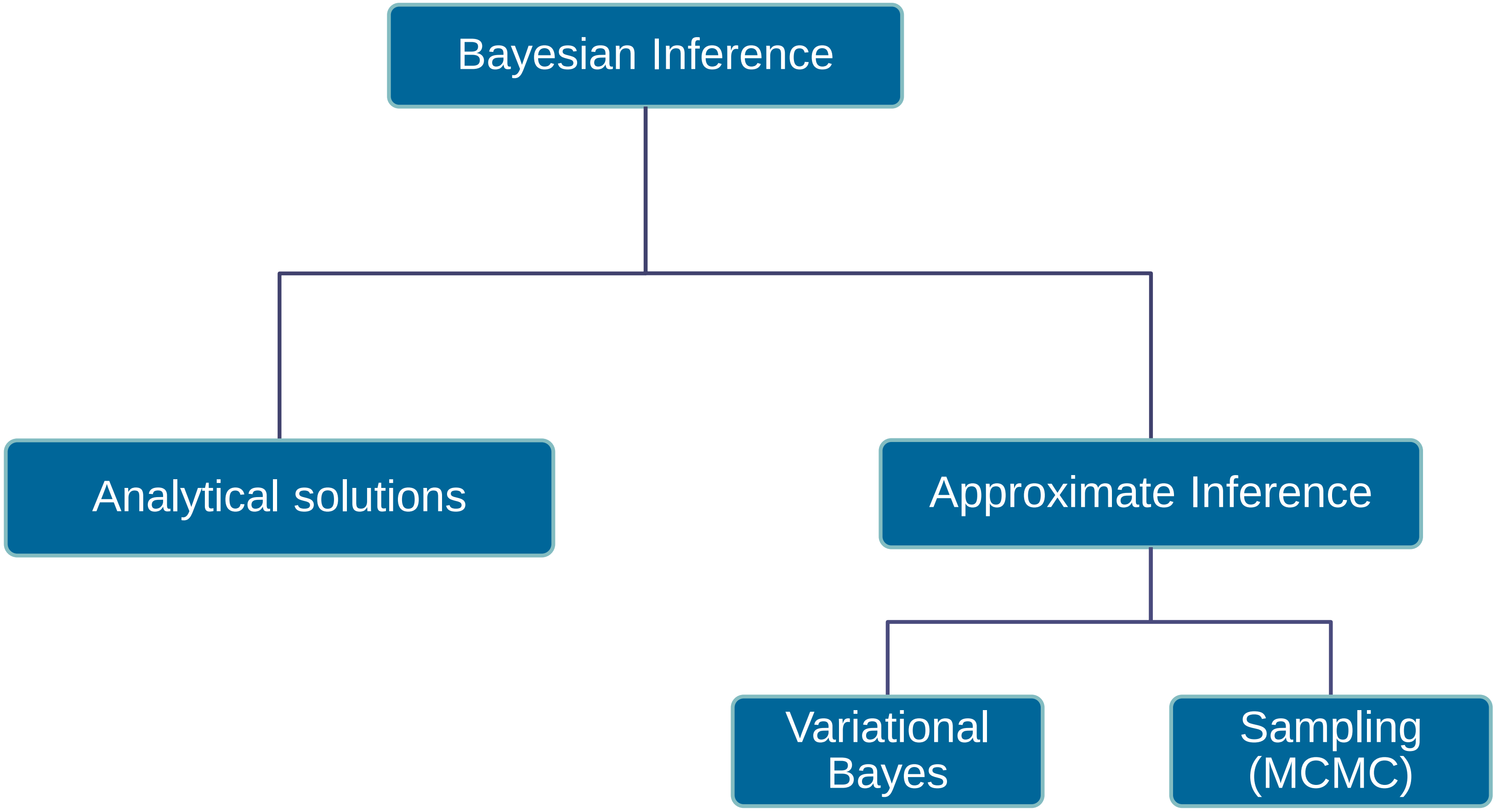
⇨ **Model inversion**

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

⇒ 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 methods (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

A simple example

$$y = \theta + \varepsilon$$

Likelihood & prior

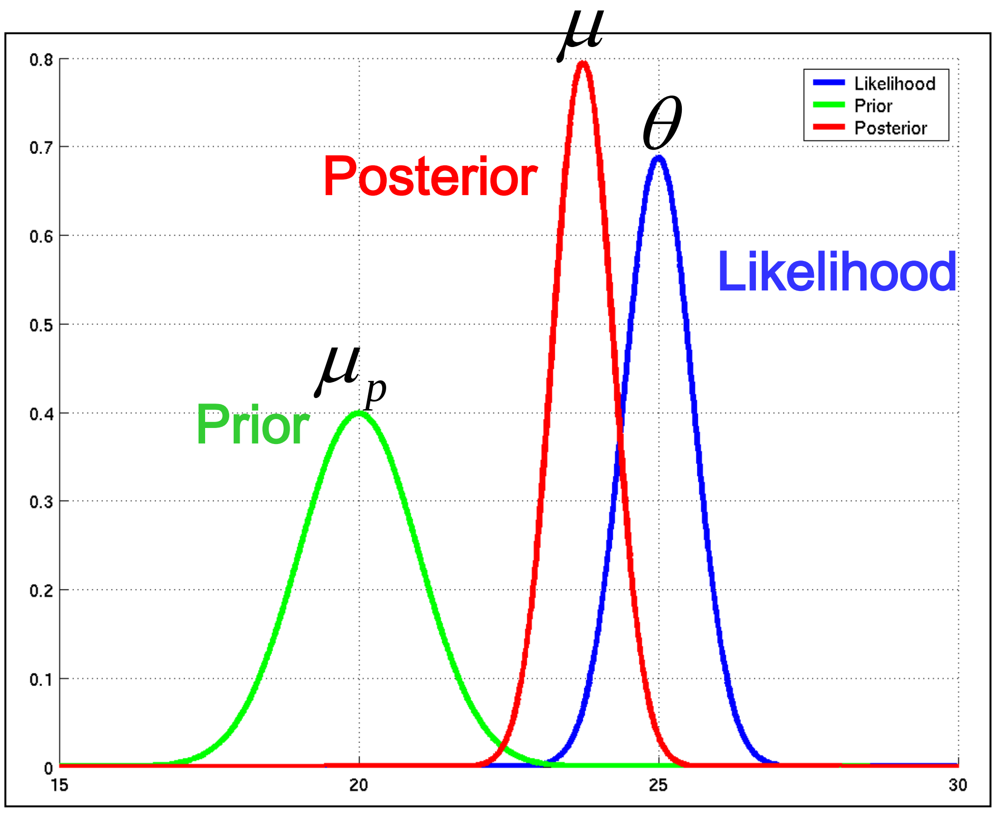
$$p(y|\theta) = N(\theta, \lambda_e^{-1})$$

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

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:
a principle we encounter throughout the course

Choice of priors

- Objective priors:
  - "non-informative" priors
  - but can still exert 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 the result of previous empirical results
- Shrinkage priors:
  - emphasize regularization and sparsity
- Empirical priors:
  - learn parameters of prior distributions from the data ("empirical Bayes")
  - rests on a hierarchical model

Model comparison and selection

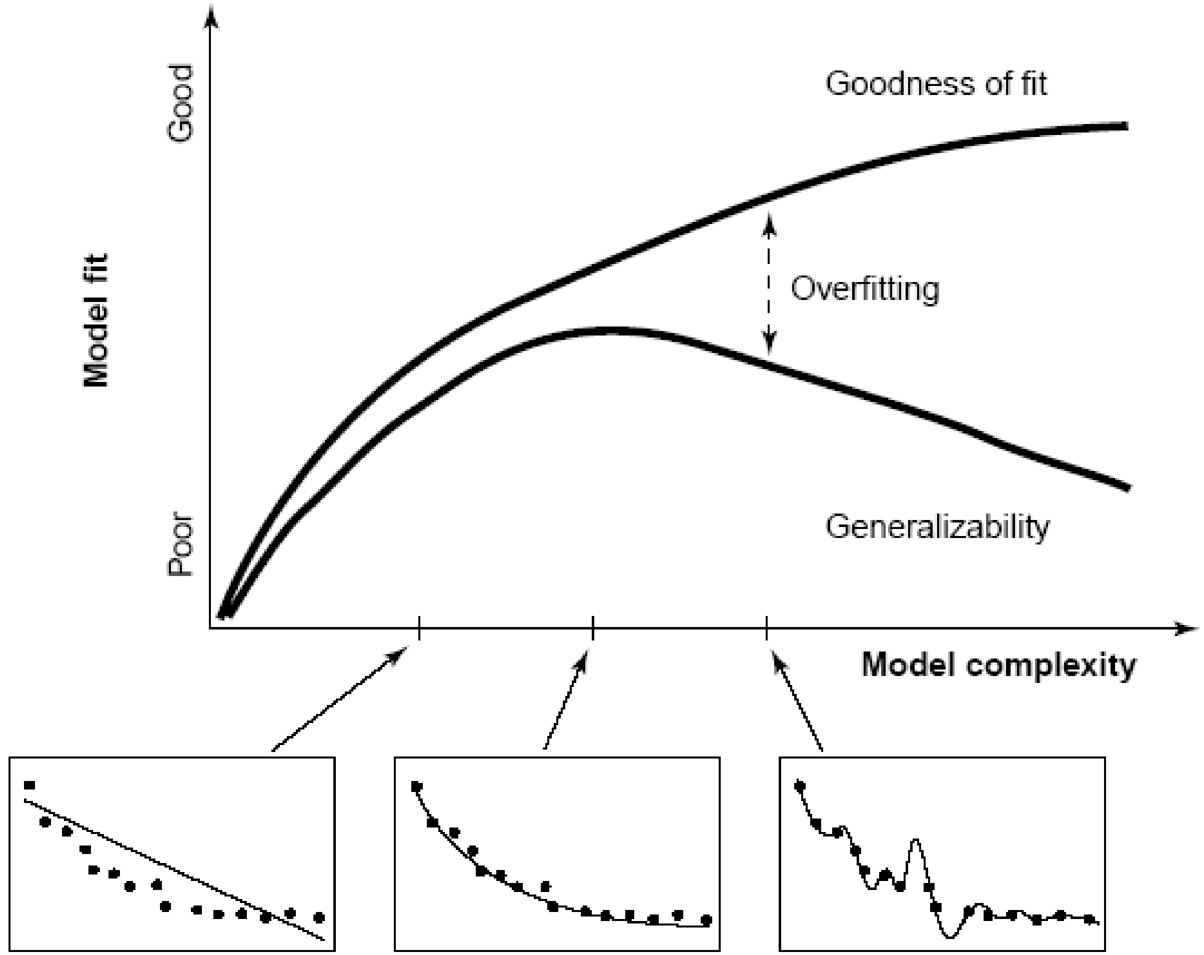
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

⇩

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

⇩

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



Pitt & Myung (2002) *TICS*

Bayesian model selection (BMS)

- First step of inference: define model space M

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

- Inference on model structure m:

Posterior model probability

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

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

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

Model evidence:

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

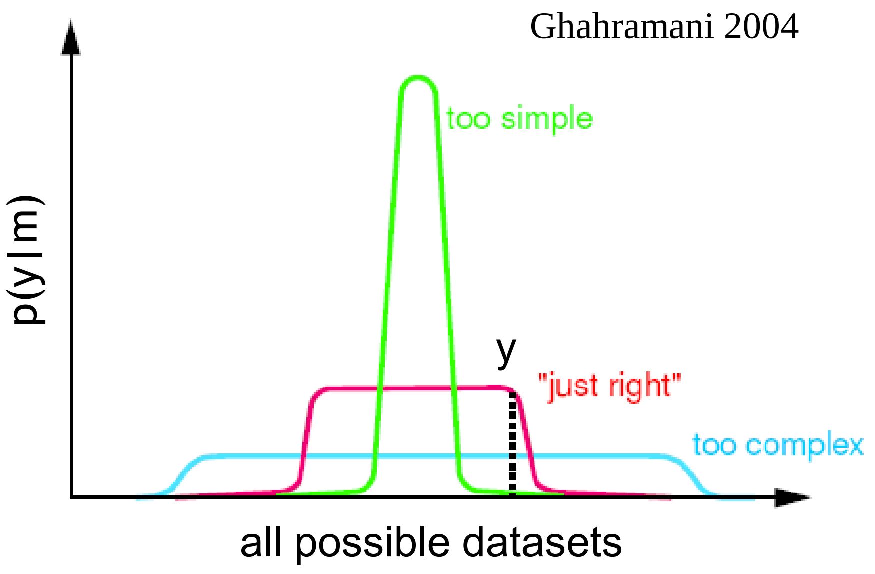
Bayesian model selection (BMS)

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

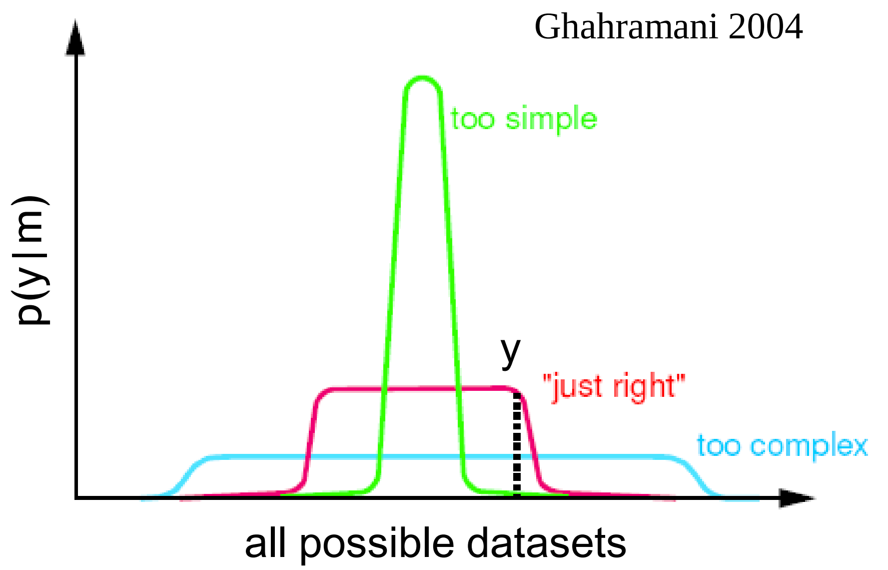


Various approximations:
- negative free energy (F)
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

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?"

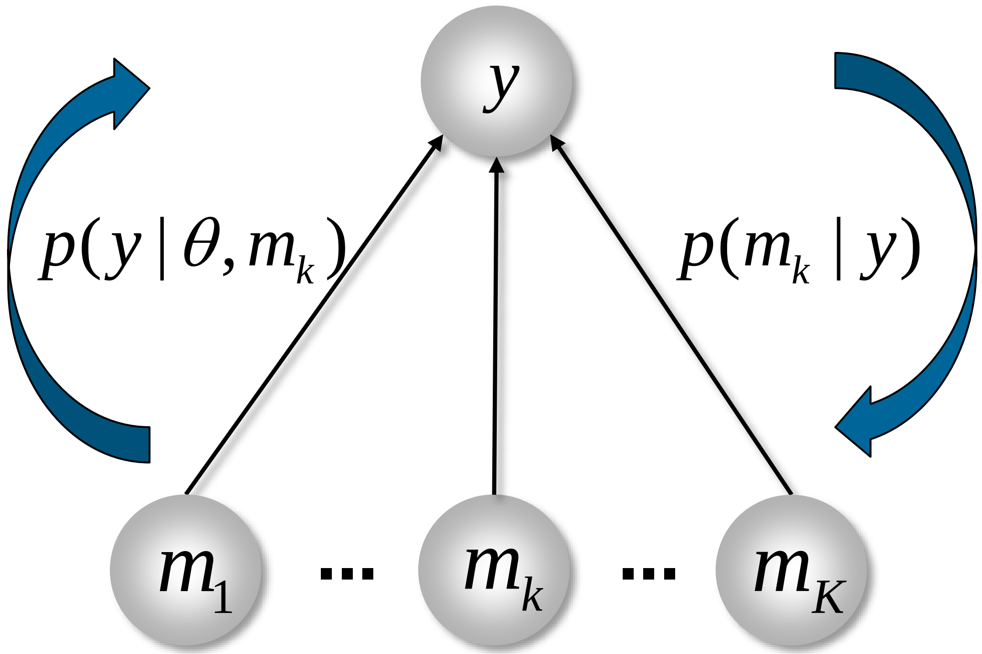
⇨ accounts for both accuracy and complexity of the model

Various approximations:

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

Differential diagnosis by model selection

SYMPTOM
(behaviour
or physiology)

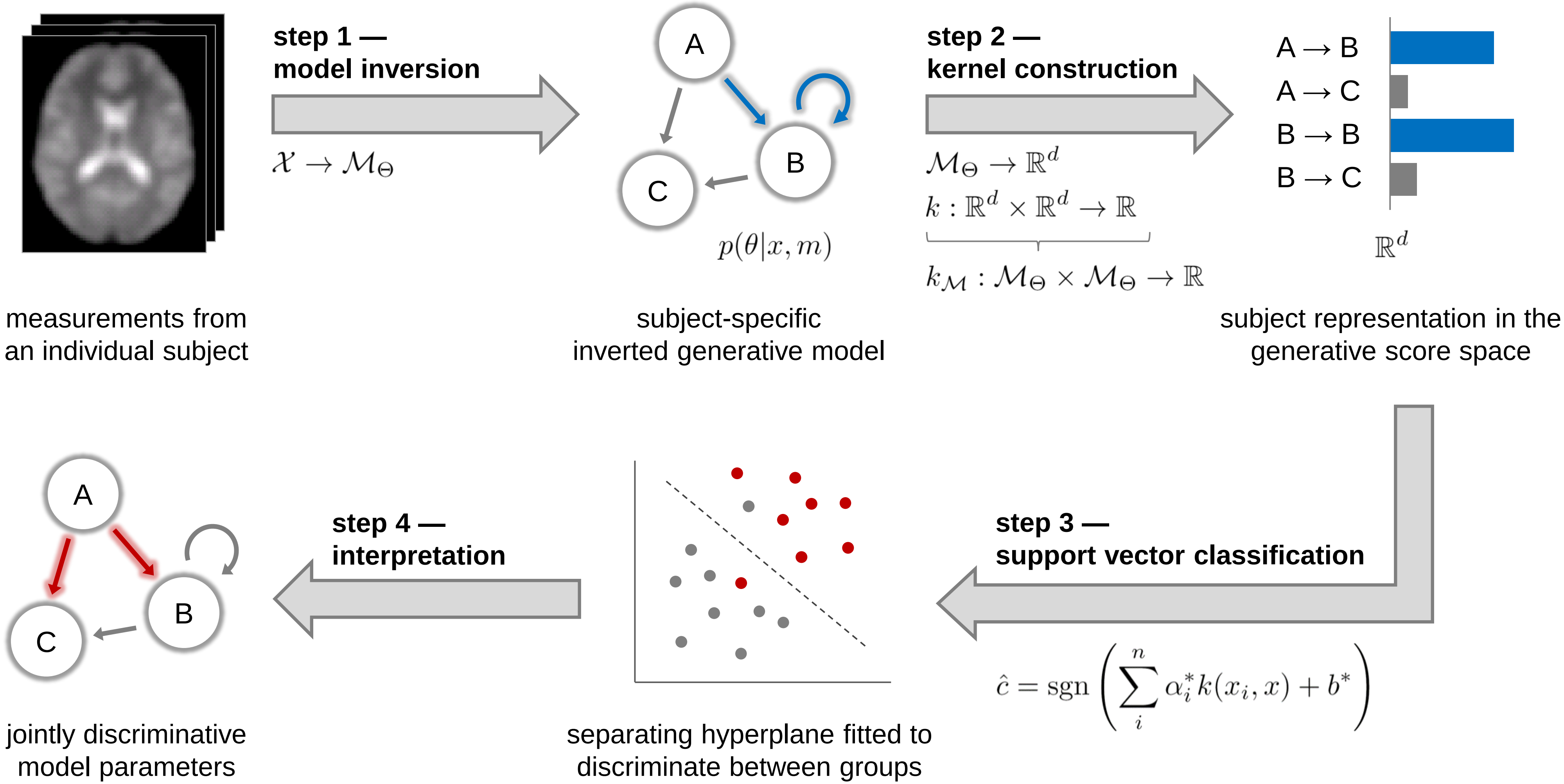


HYPOTHETICAL
MECHANISM

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

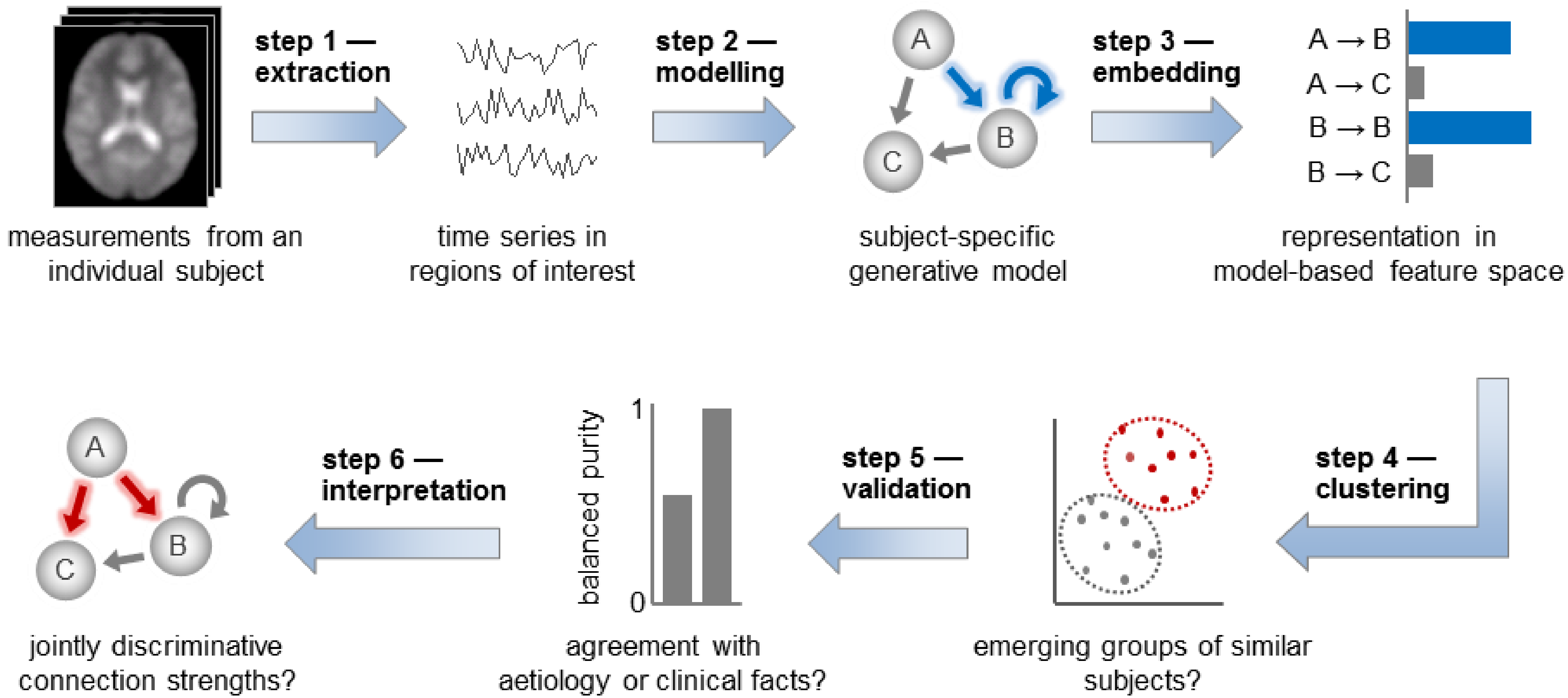
Stephan et al. 2017, *NeuroImage*

Generative embedding (supervised)

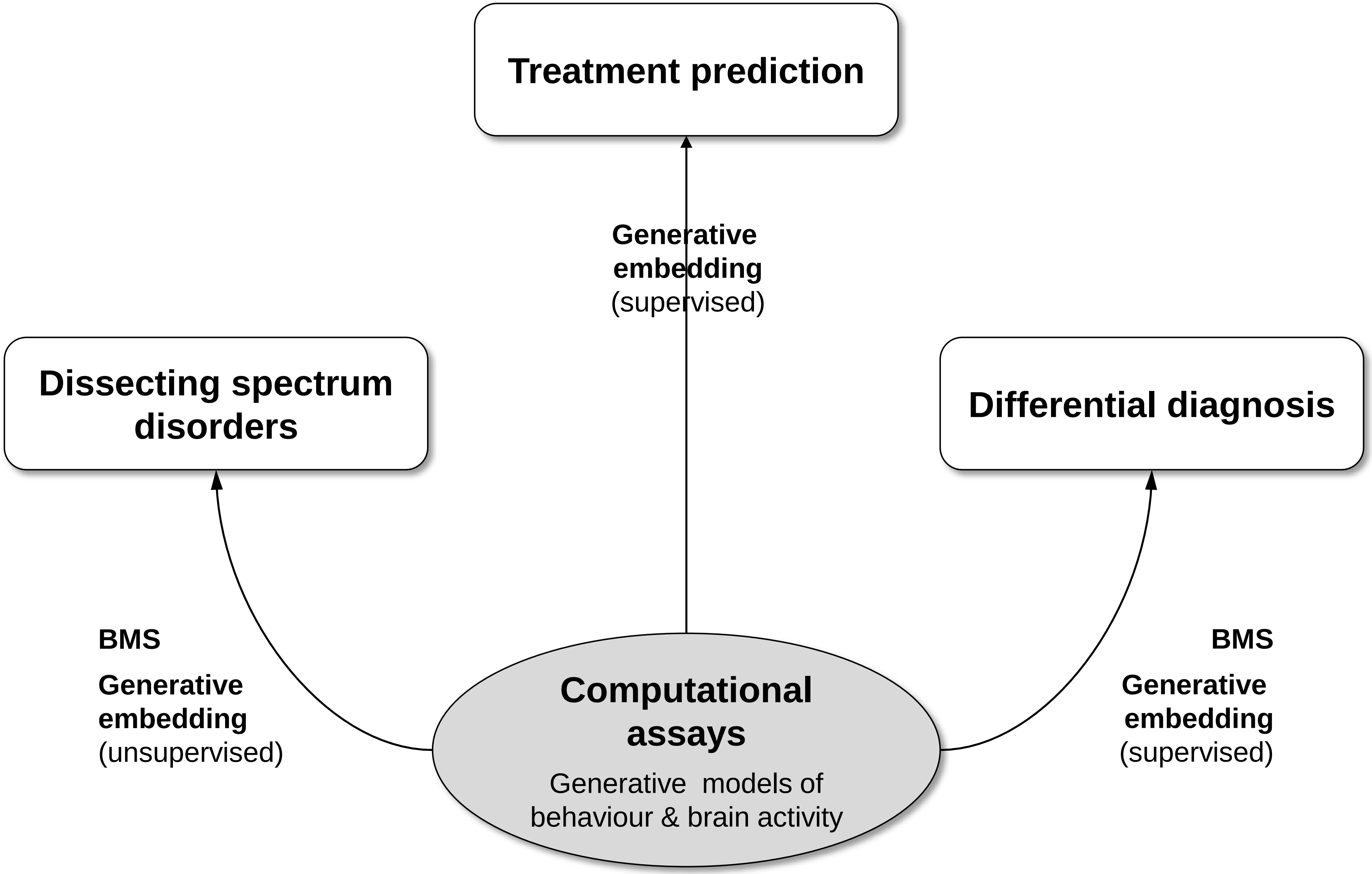


Brodersen et al. 2011, *PLoS Comput. Biol.*

Generative embedding (unsupervised)



Brodersen et al. 2014, *NeuroImage Clinical*


Treatment prediction
Generative embedding (supervised)
Dissecting spectrum disorders
Differential diagnosis
BMS
Generative embedding (unsupervised)
BMS
Generative embedding (supervised)
Computational assays
Generative models of behaviour & brain activity

Summary

- system models:
  - separating hidden system states from observed system behavior (measurements)
  - do not operate on raw data, but try to infer mechanisms that generated data
- this rests on a probabilistic forward model (from parameters to data)
  - state equations
  - observation equations
- inference:
  - MLE: straightforward but problematic
  - instead: Bayesian inversion of a generative model
  - inference on parameters and model structure

Thank you