Introduction to Computational Modeling: Generative Models

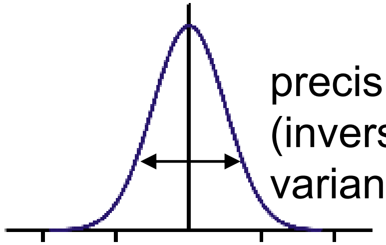
Klaas Enno Stephan

Translational Neuromodeling Unit

Universität Zürich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A brief note on mathematical notations



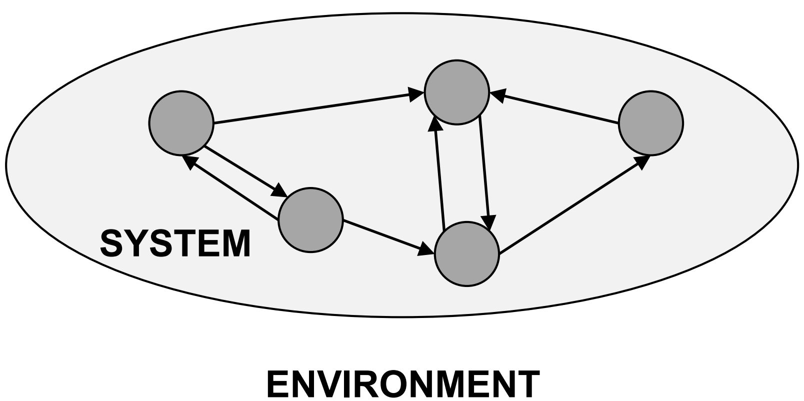
- For example: Gaussian (Normal) distributions
  - for scalars: $p(x) = N(x; \mu, \sigma^2)$ μ = mean; σ^2 = variance
  - for vectors: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ Σ = covariance matrix = $E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$

- same thing, just expressed wrt. precision
  - for scalars: $p(x) = N(x; \mu, \lambda^{-1})$ μ = mean; $\lambda = 1/\sigma^2$ = precision
  - for vectors: $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \Lambda^{-1})$ Λ = precision matrix

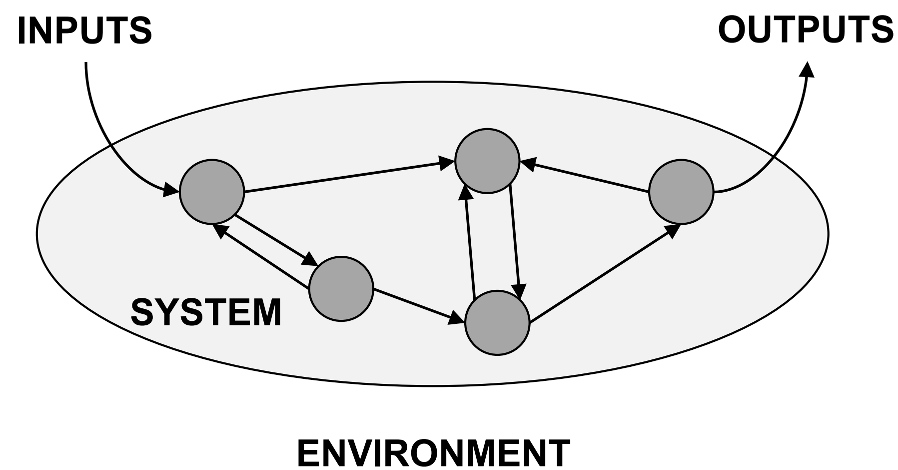
Systems

- 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



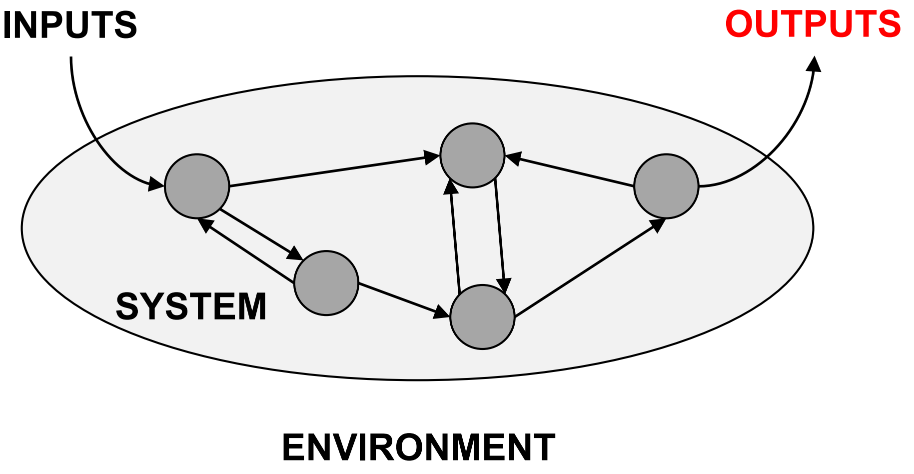
Stephan: *Translational Neuromodeling & Computational Psychiatry, in prep.*

System 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

- system models describe (at least) three things:
  – how system states evolve in time
  – how states determine system outputs
  – how observations of outputs are affected by noise



NB: Outputs can be
- actions (from the system's perspective)
- data (from an outside observer's view)

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

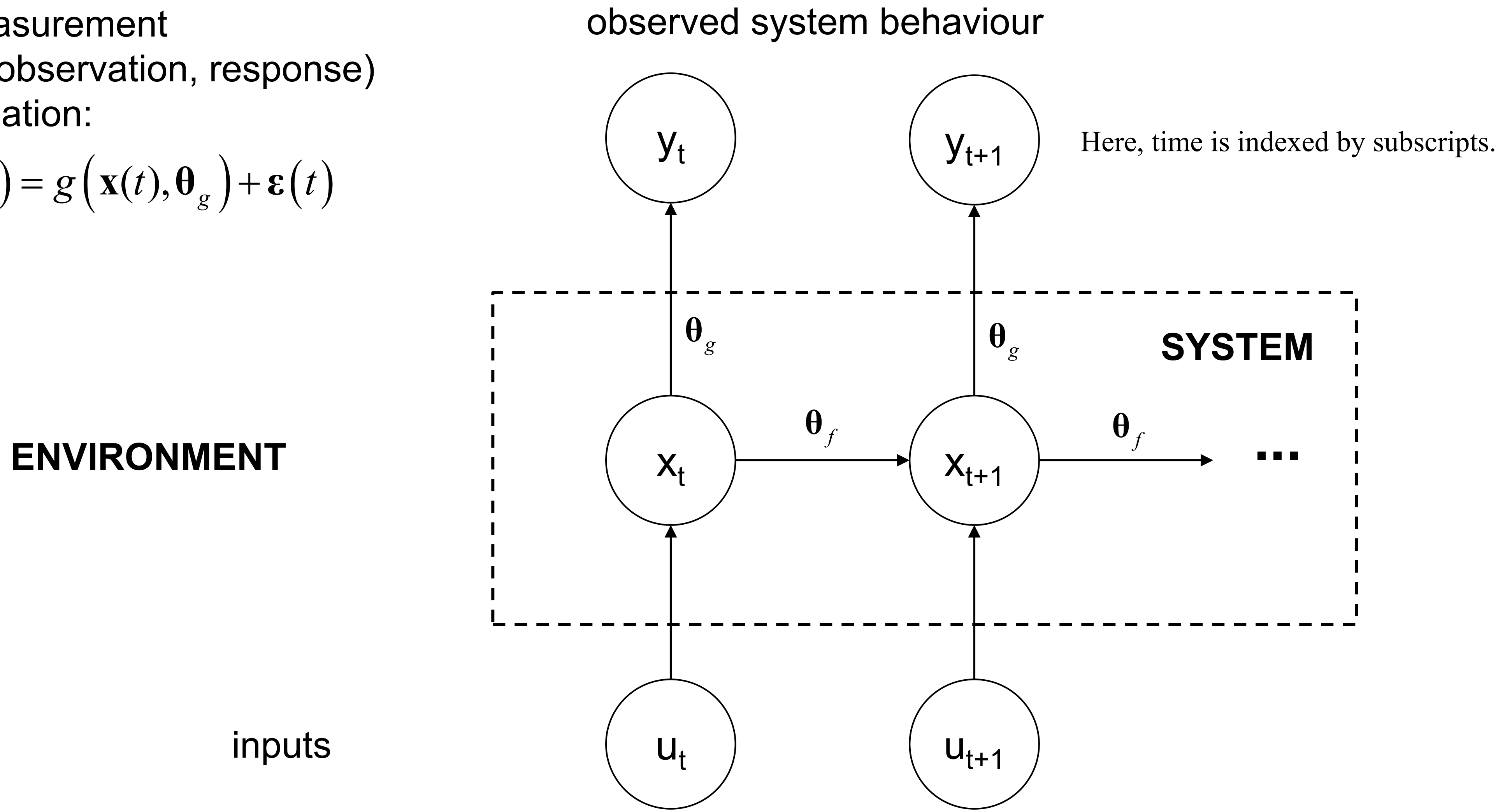
$$\mathbf{x}\left(t+1\right) = f\left(\mathbf{x}(t), \boldsymbol{\theta}_f, \mathbf{u}(t)\right)$$ as difference equation

For a discussion of system theory in the context of neuroimaging, see Stephan 2004, *Journal of Anatomy*

State space representation

measurement
(or observation, response)
equation:

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



Stephan: *Translational Neuromodeling & Computational Psychiatry, in prep.*

Deterministic vs. stochastic state space models

- **deterministic models**
  - no state noise: $$\frac{d\mathbf{x}}{dt} = f\left(\mathbf{x}(t), \boldsymbol{\theta}_f, \mathbf{u}(t)\right)$$ ODEs

  $\rightarrow$ states $\mathbf{x}(t)$ fully determined by initial state $x(0)$, parameters $\boldsymbol{\theta}$ and inputs $\mathbf{u}(t)$

  $\rightarrow$ if inputs and initial state are known, inference on parameters sufficient to reconstruct state trajectories

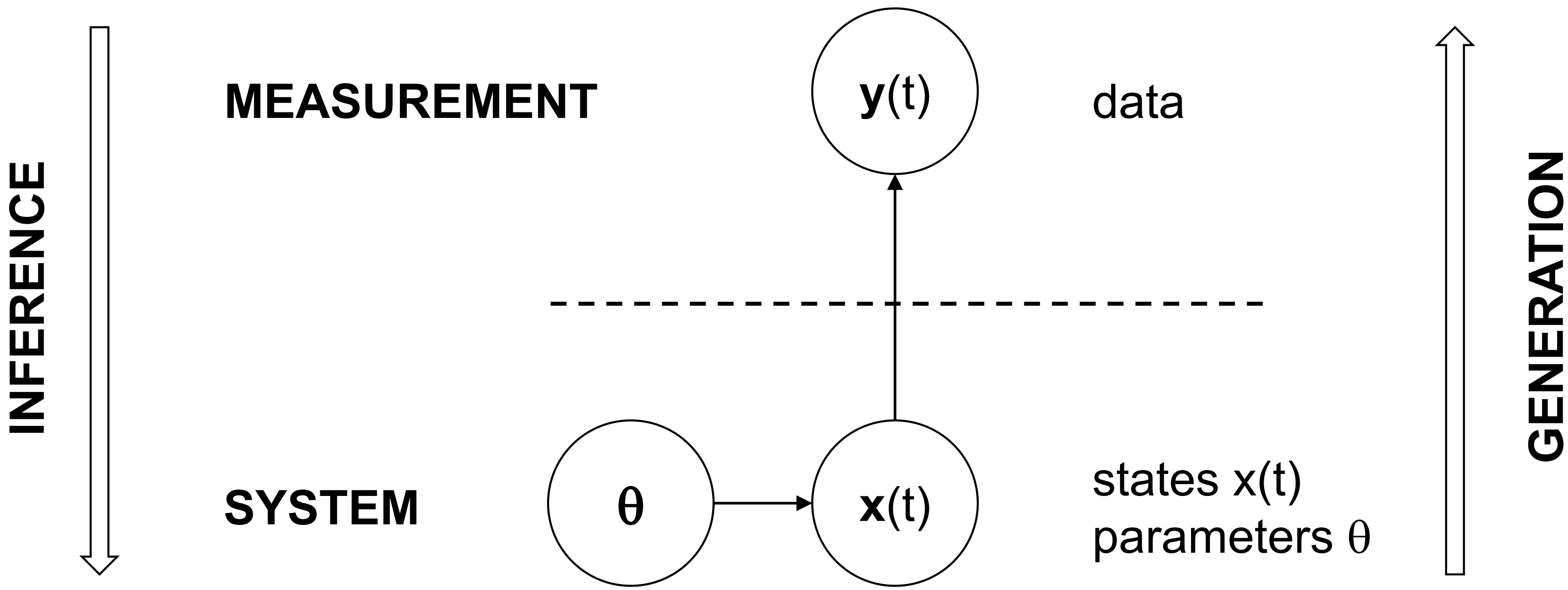
- **stochastic models**
  - state noise: $$\frac{d\mathbf{x}}{dt} = f\left(\mathbf{x}(t), \boldsymbol{\theta}_f, \mathbf{u}(t)\right) + \omega(t)$$ SDEs

  $\rightarrow$ states $\mathbf{x}(t)$ not fully determined by initial state, parameters and inputs

  $\rightarrow$ much tougher inference problem!

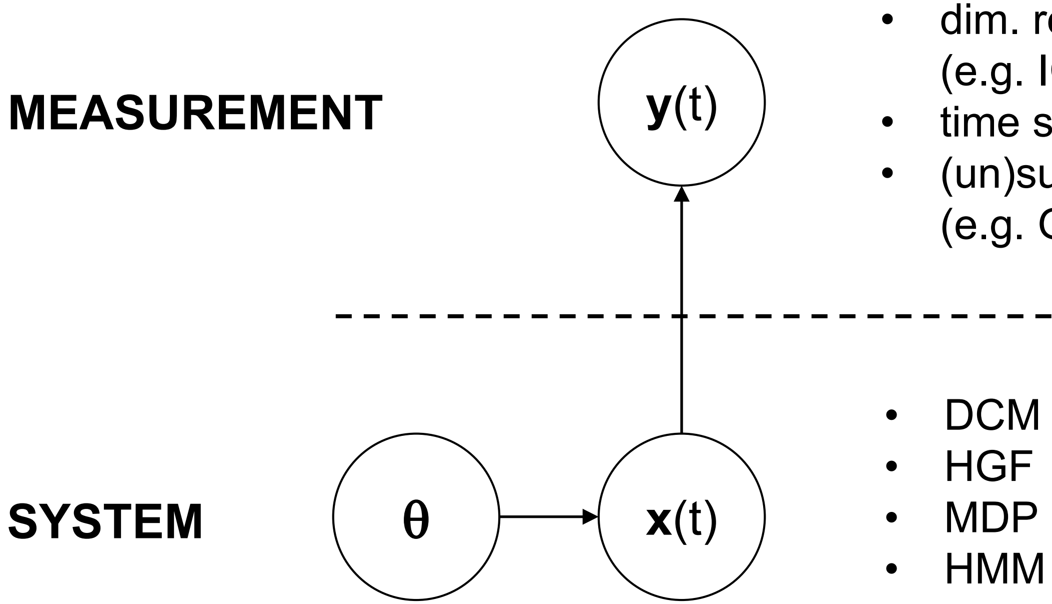
Models with/without latent states

- many ways to categorise modeling approaches
- one possibility: distinguish presence vs. absence of latent states



Stephan: *Translational Neuromodeling & Computational Psychiatry, in prep.*

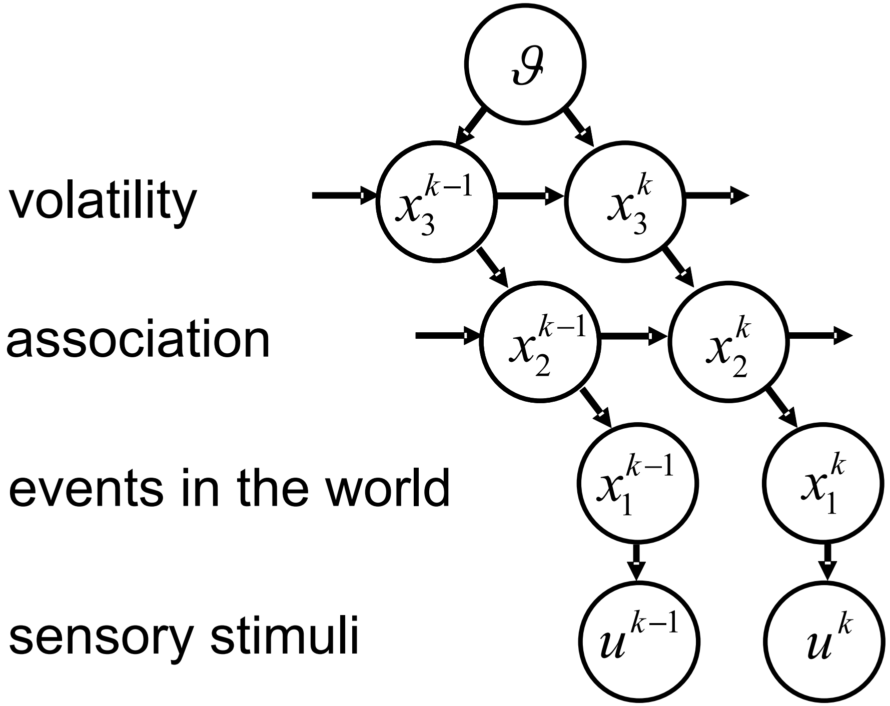
Examples of approaches with/without latent states

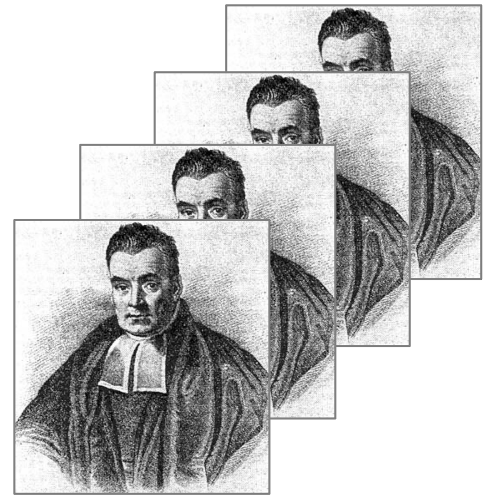


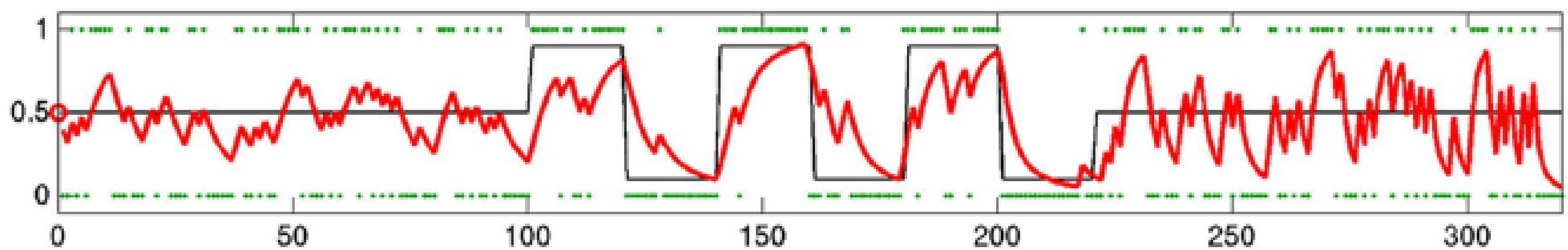
- dim. reduction/decomposition methods (e.g. ICA, PCA)
- time series models
- (un)supervised learning methods from ML (e.g. GP, SVM ...)

- DCM
- HGF
- MDP (RL, active inference)
- HMM

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

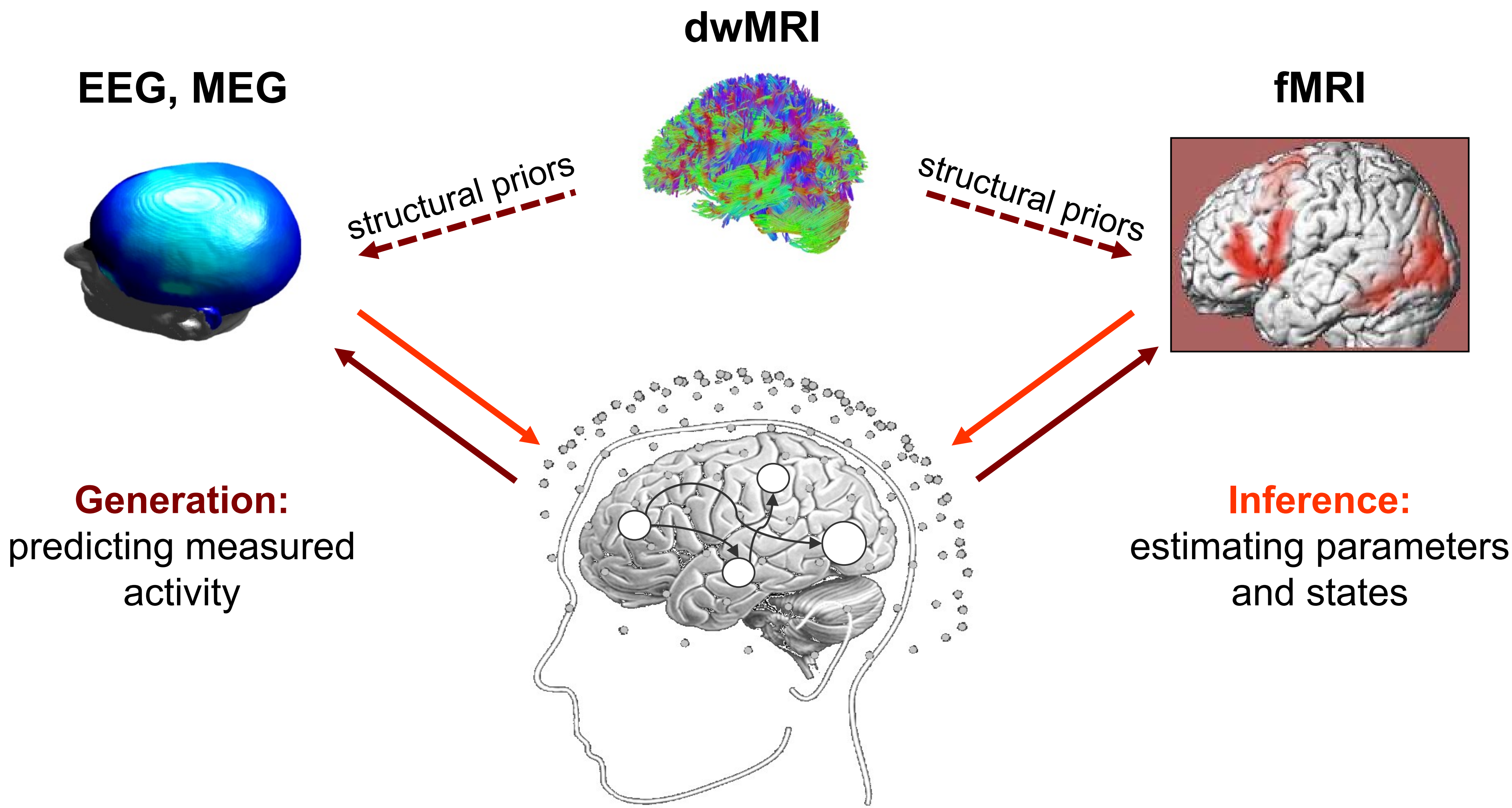






Mathys et al. 2011, 2014, *Front. Hum. Neurosci.*

... and DCMs of fMRI and EEG/MEG data



adapted from:
Stephan et al. 2017, *NeuroImage*

Maximum likelihood estimation (MLE)

- Given a system model and measured data, we would like to estimate the values of the model parameters.

- Once we have specified our assumptions about the nature of the observation noise (e.g. IID Gaussian), we can compute the **likelihood $\mathbf{p(y|\theta)}$**, i.e.: Given a particular value of $\boldsymbol{\theta}$**,** how likely are the observed data y under the chosen model?

- We could then search for the parameter value that maximises the (log) likelihood. This is the parameter value for which the model fits the data best.

- This is known as **maximum likelihood estimation (MLE)**:

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

Maximum likelihood estimation (MLE)

- Given a system model and measured data, we would like to estimate the values of the model parameters.

- Once we have specified our assumptions about the nature of the observation noise (e ... e.:
  Given a ... ler the
  chosen

More in the talk on maximum likelihood estimation.

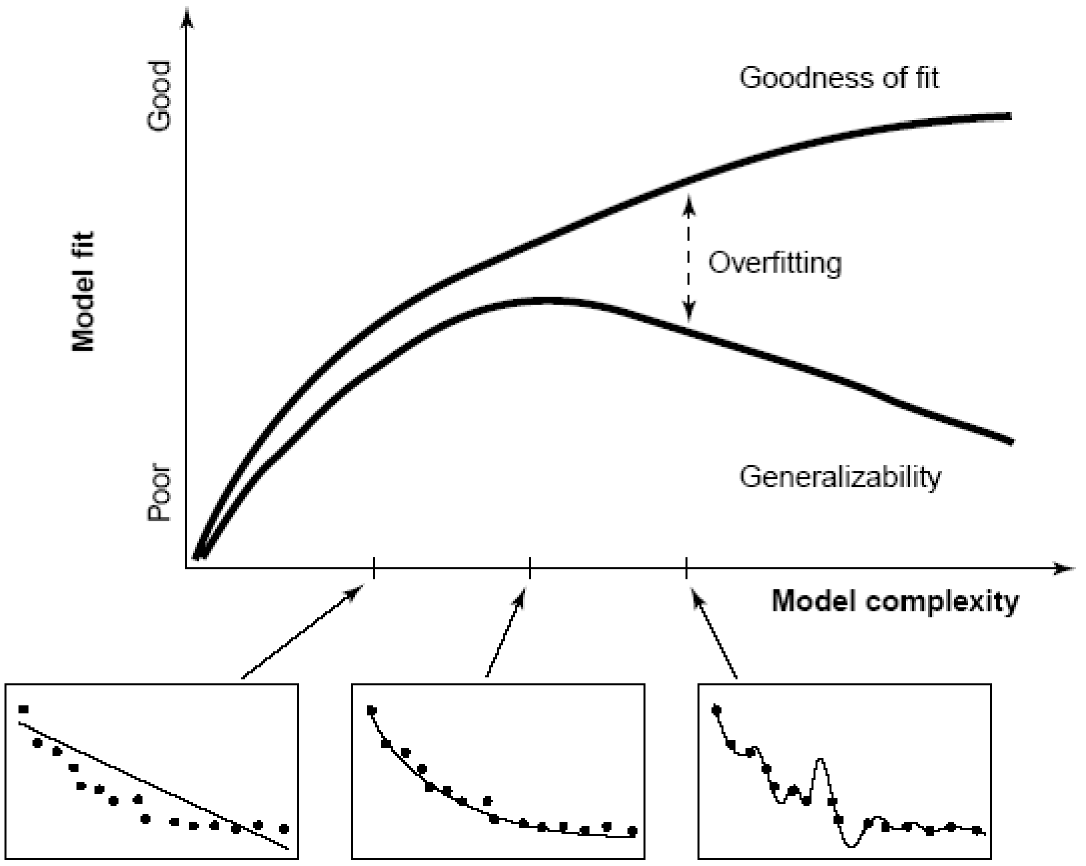
- We could then search for the parameter value that maximises the (log) likelihood. This is the parameter value for which the model fits the data best.

- This is known as **maximum likelihood estimation (MLE)**:

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

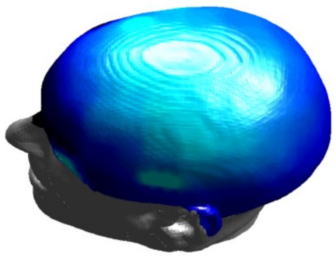
Overfitting

- MLE has various limitations. For example, for complex models and limited data, **overfitting** is a severe problem (see later talks in the course).
- 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*

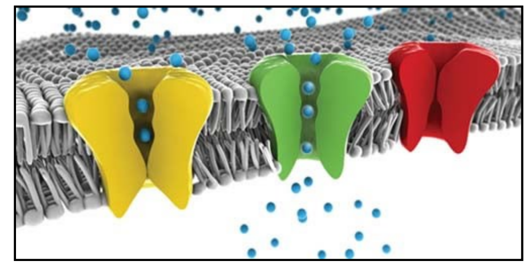
Generative models



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

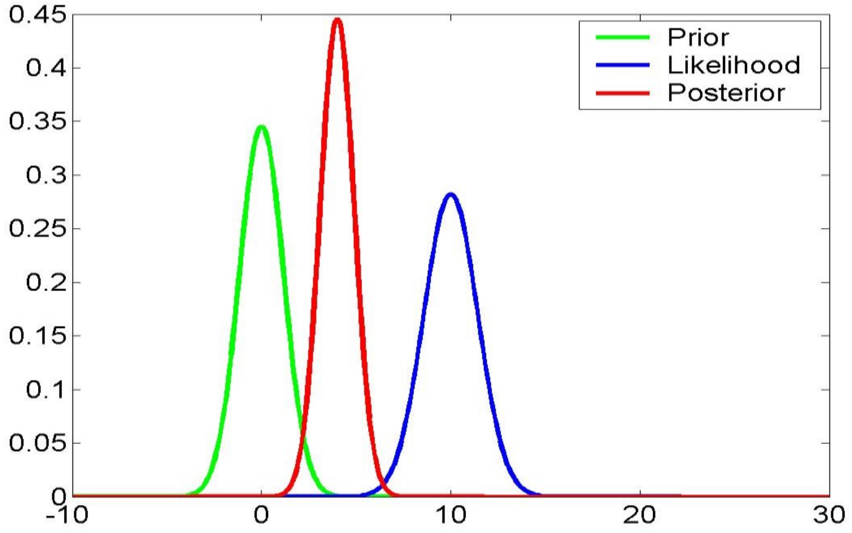
$$p(\boldsymbol{\theta} \mid \mathbf{y}, m)$$

y = data, θ = parameters, m = model



1. a probabilistic forward mapping from parameters to data, defined by likelihood and prior (joint probability)
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. model inversion = inference about parameters → posterior $p(\theta|y,m)$
5. natural basis for model comparison → model evidence $p(y|m)$

Bayes' rule



The Reverend Thomas Bayes
(1702-1761)

Likelihood × **prior**: generative model

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

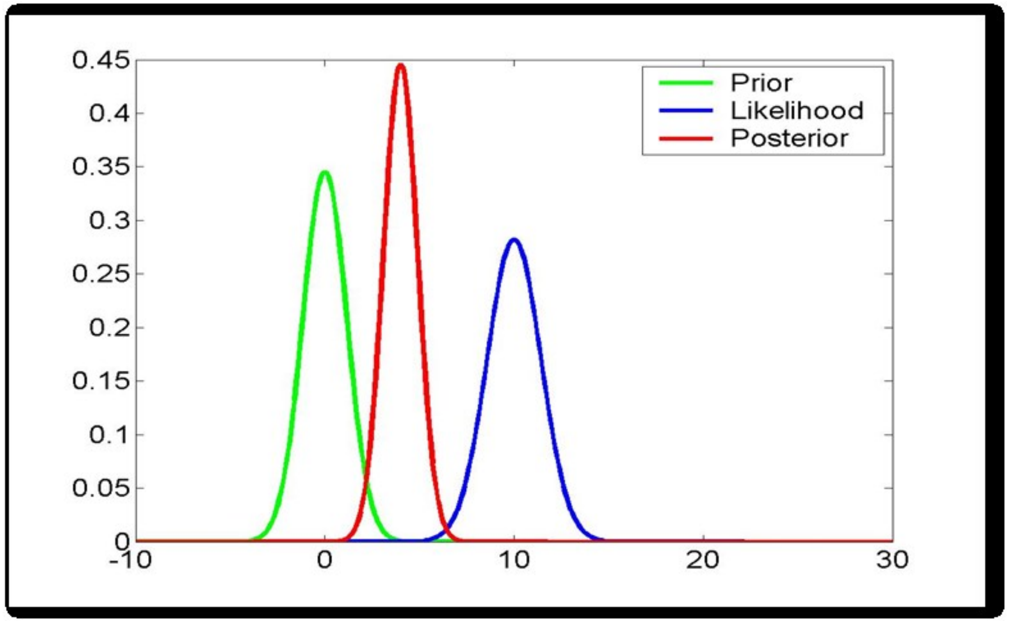
θ: parameters
y: data

Model evidence: normalisation term and index for model goodness

"... the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence."

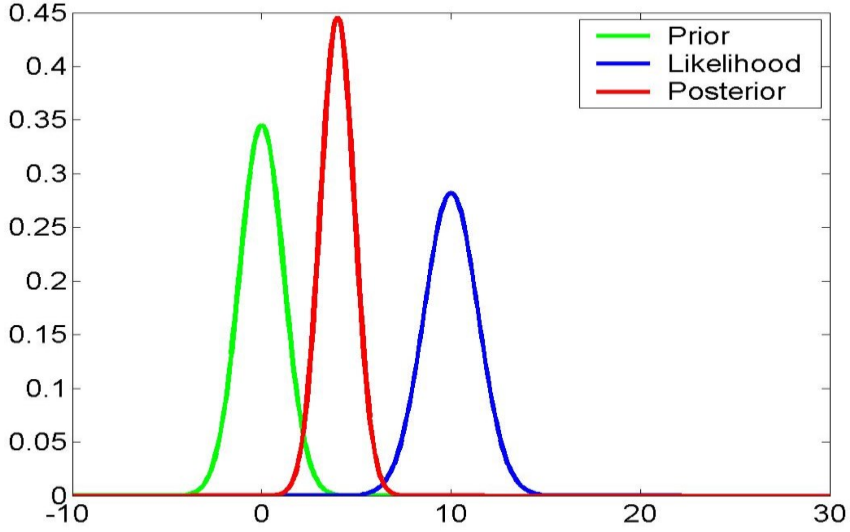
Wikipedia

Bayesian inference: an animation



Code courtesy by Guillaume Flandin

Bayes' rule

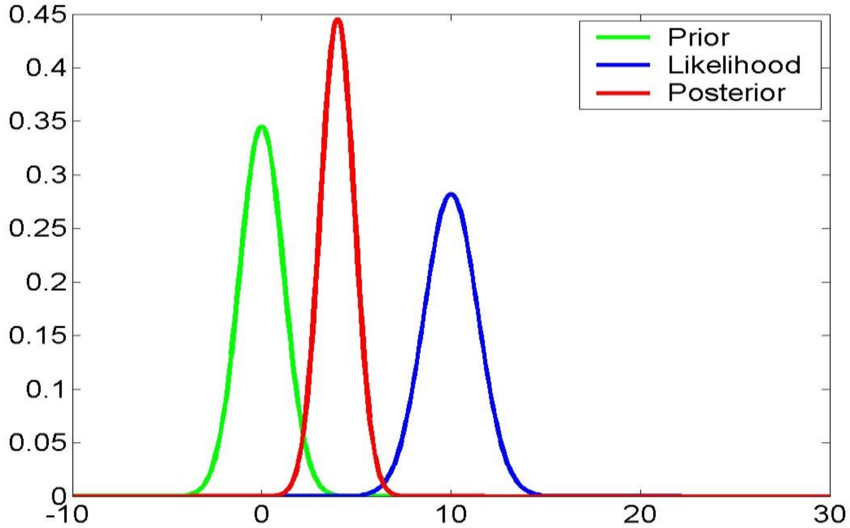


The Reverend Thomas Bayes
(1702-1761)

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

No change to previous equation – just making the choice of a particular model explicit.

Bayes' rule



The Reverend Thomas Bayes
(1702-1761)

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

Evidence:
likelihood that data were generated by model m, averaging over all possible parameter values (as weighted by the prior).

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

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

Maximum a posteriori (MAP) estimation

- A simple way to use a generative model (and go beyond MLE) is to compute MAP estimates.

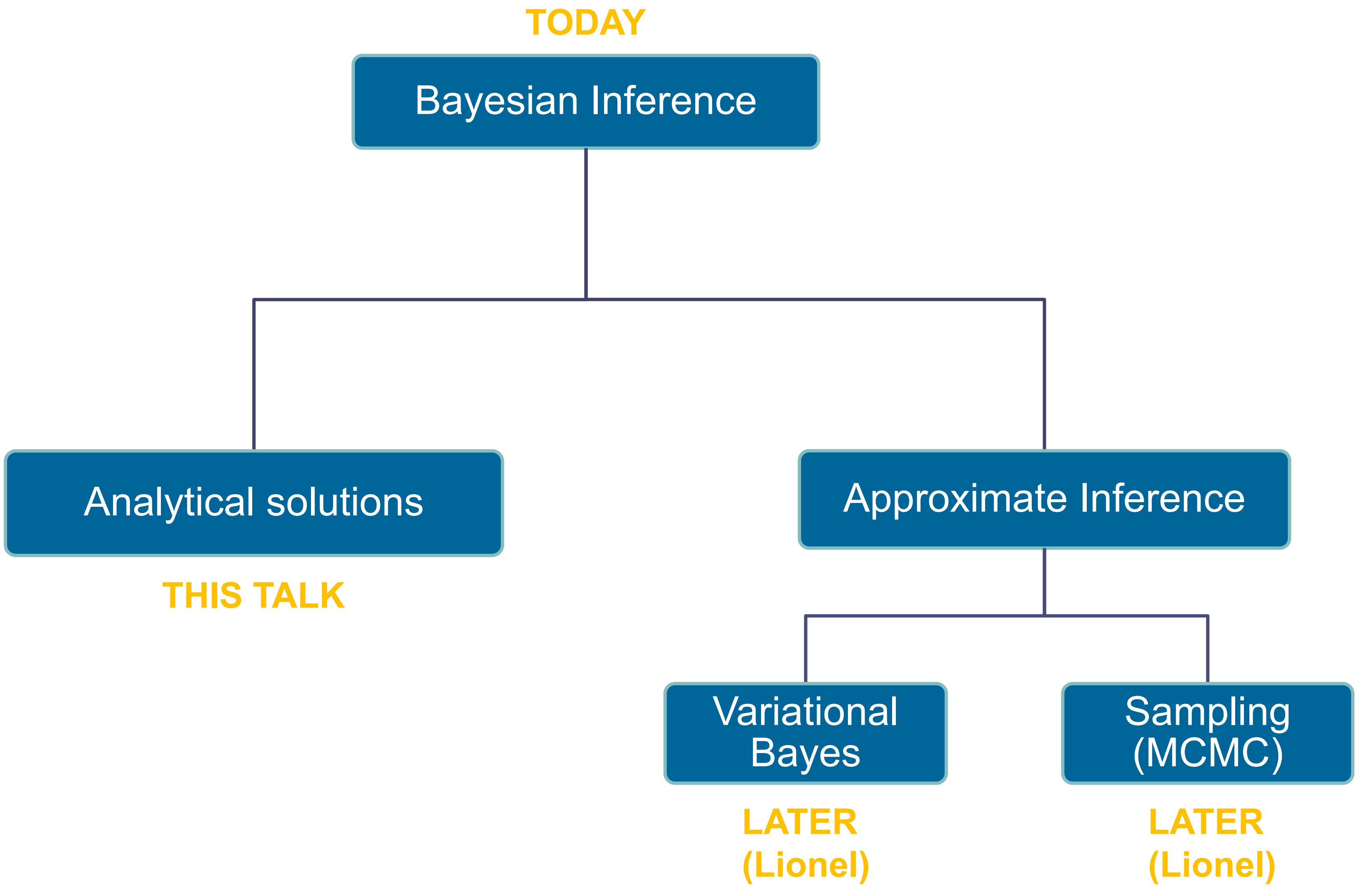
- This finds parameter values that maximize the numerator of Bayes' theorem:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{MAP} &= \arg\max_{\theta}\left[p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)\right] \\ &= \arg\max_{\theta}\left[\ln p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) + \ln p\left(\boldsymbol{\theta}\right)\right]\end{aligned}$$

- Advantages:
  - prior serves to regularize $\rightarrow$ can prevent overfitting
  - does not require one to evaluate the model evidence
  - simple to implement (e.g. numerical optimization methods)

- Disadvantages:
  - does not provide the full posterior, only a point estimate
  - no information about uncertainty

Methods for Bayesian inference



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 (e.g. Markov Chain Monte Carlo, MCMC)**
  - theoretically guaranteed to be accurate (for infinite computation time)
  - problems: may require very long run time in practice, only heuristics to decide about convergence in practice

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 $\propto$ Normal × Normal
  - Beta $\propto$ Binomial × Beta
  - Dirichlet $\propto$ Multinomial × Dirichlet

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

same form

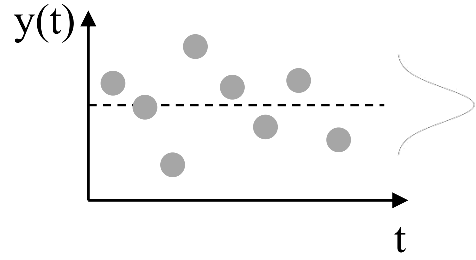
A simple example: univariate Gaussian belief update

Likelihood & prior

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

$$p(\theta) = N(\mu_{prior}, \sigma_{prior}^2)$$

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

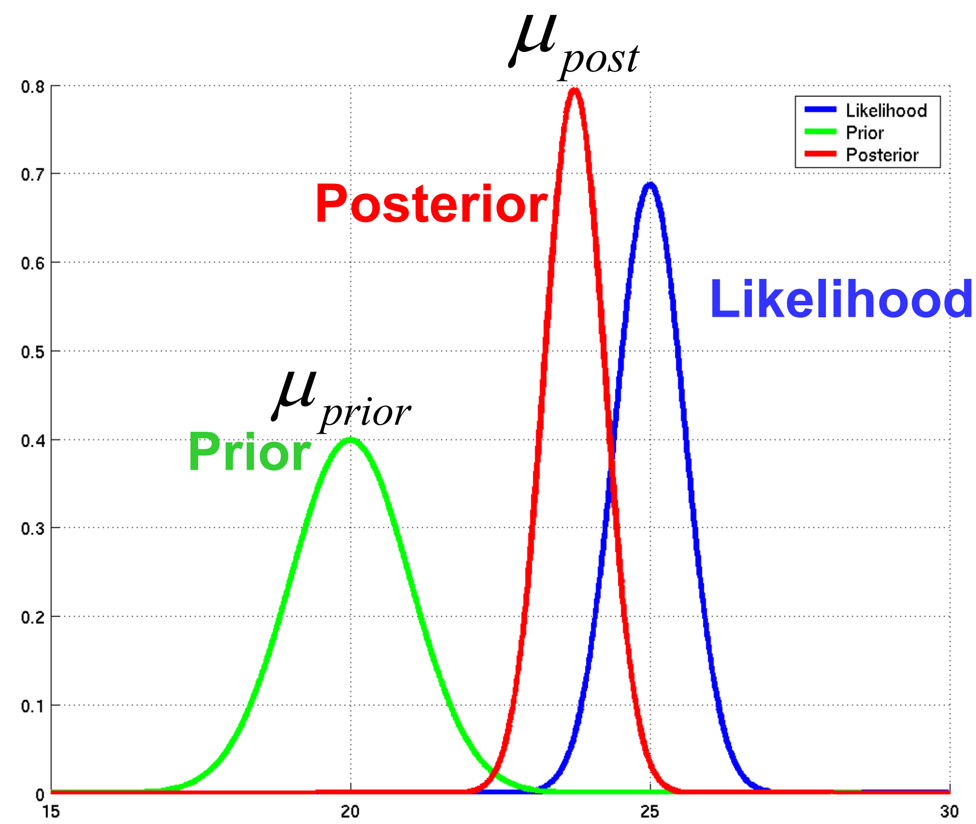


Posterior $p(\theta\,|\,y) = N\left(\mu_{post}, \lambda_{post}^{-1}\right)$
(for a single observation y)

$$\frac{1}{\sigma_{post}^2} = \frac{1}{\sigma_e^2} + \frac{1}{\sigma_{prior}^2}$$

$$\mu_{post} = \sigma_{post}^2 \left(\frac{1}{\sigma_e^2} y + \frac{1}{\sigma_{prior}^2} \mu_{prior} \right)$$

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



Adapted from a slide by Will Penny.

A simple example: univariate Gaussian belief update

Likelihood & prior

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

$$p(\theta) = N(\mu_{prior}, \lambda_{prior}^{-1})$$

Posterior $p(\theta\,|\,y) = N\left(\mu_{post}, \lambda_{post}^{-1}\right)$

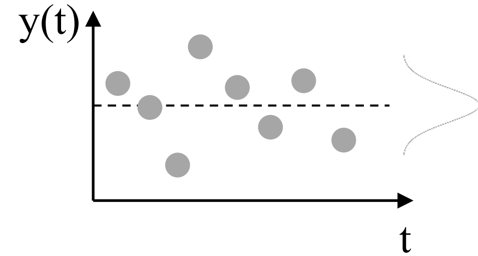
(for a single observation y)

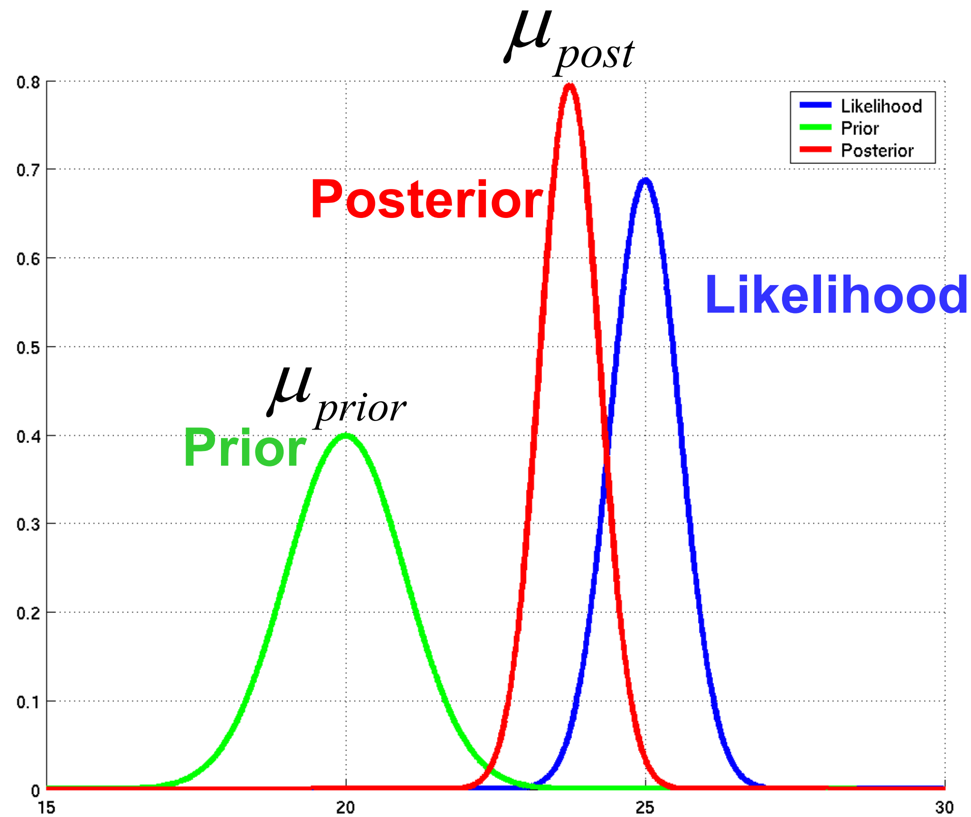
$$\lambda_{post} = \lambda_e + \lambda_{prior}$$

$$\mu_{post} = \frac{\lambda_e}{\lambda_{post}} y + \frac{\lambda_{prior}}{\lambda_{post}} \mu_{prior}$$

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

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



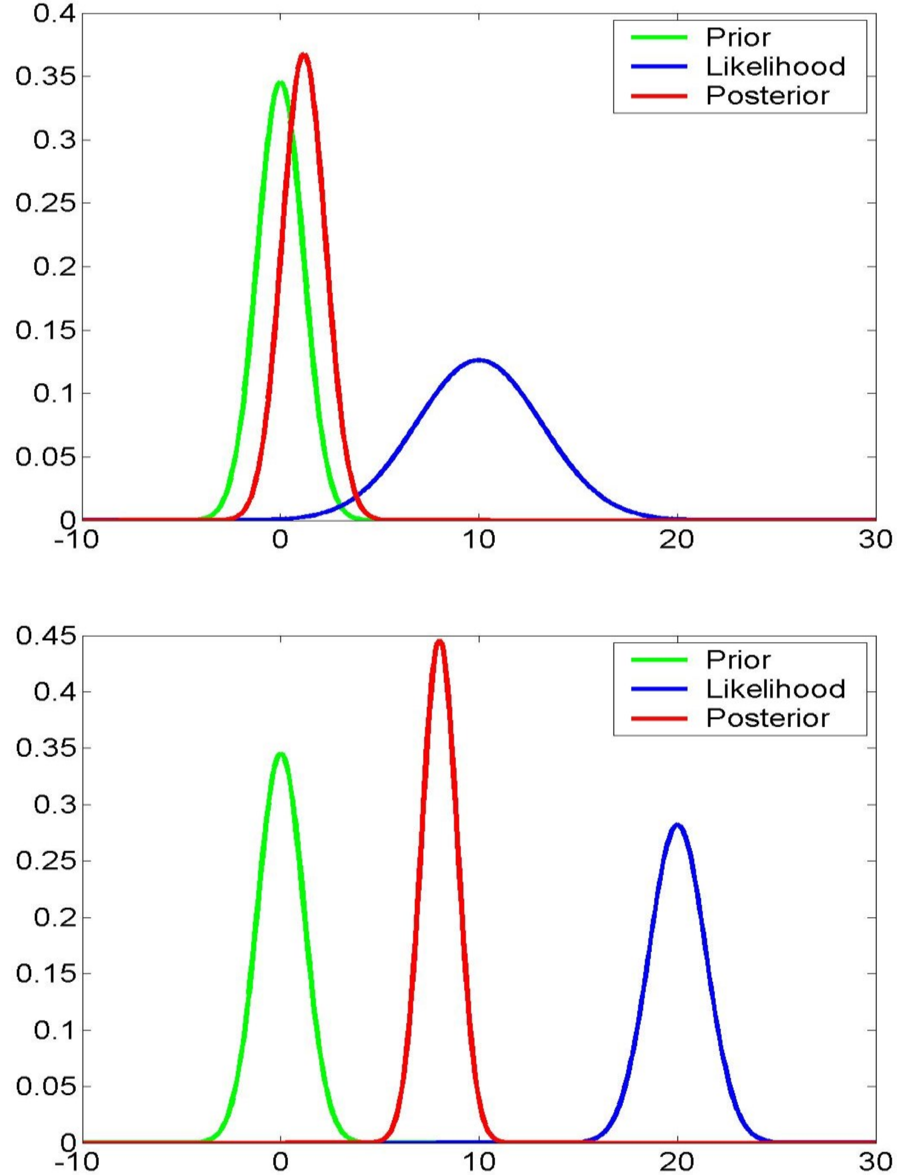


Adapted from a slide by Will Penny.

Choice of priors

- Objective priors:
  - "non-informative" priors
- Subjective priors:
  - subjective but not arbitrary
  - express beliefs that result from an understanding the problem or system
  - can be result of previous empirical results
  - can accommodate objective constraints (e.g., non-negativity)
- Shrinkage priors:
  - emphasise 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



Model comparison and selection

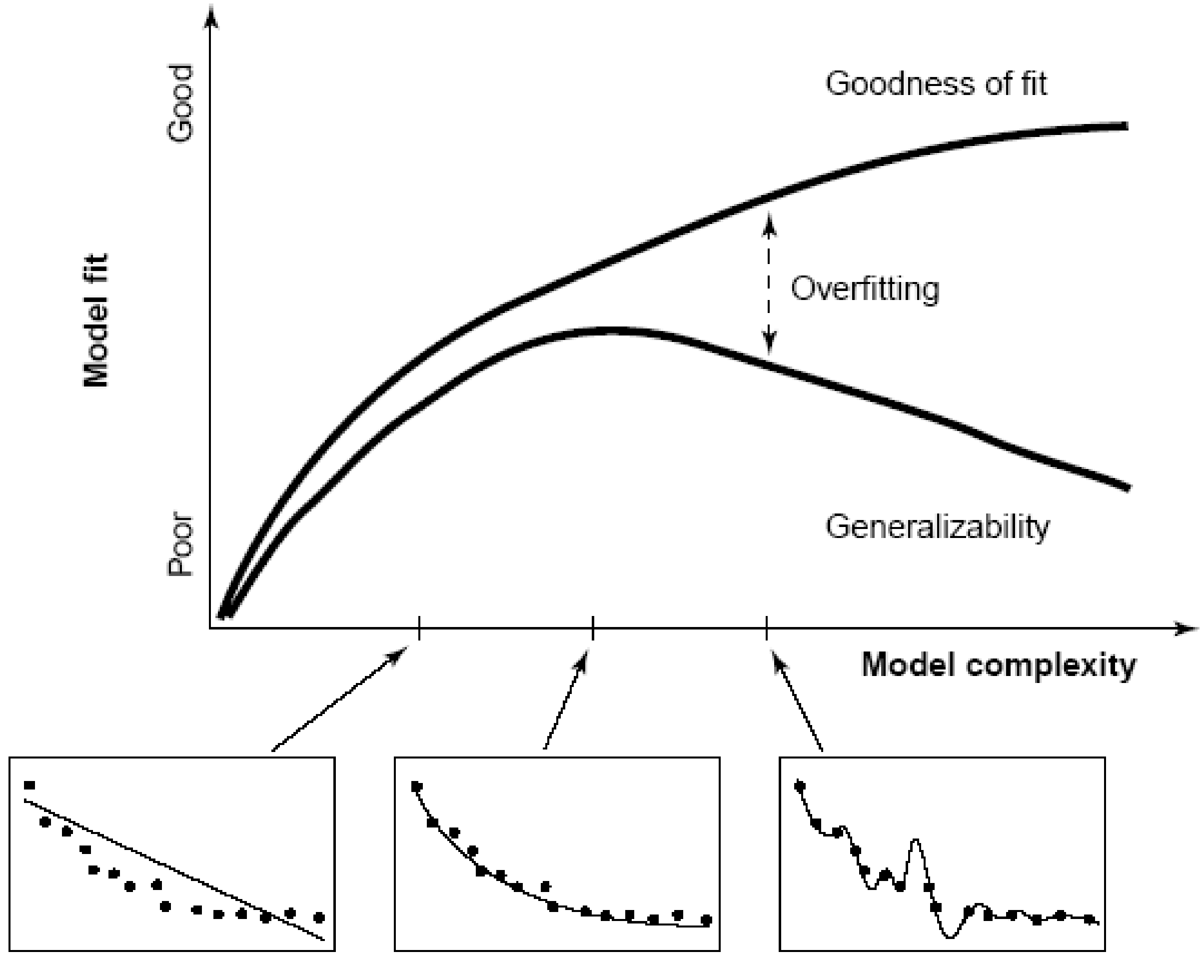
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

⇩

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*

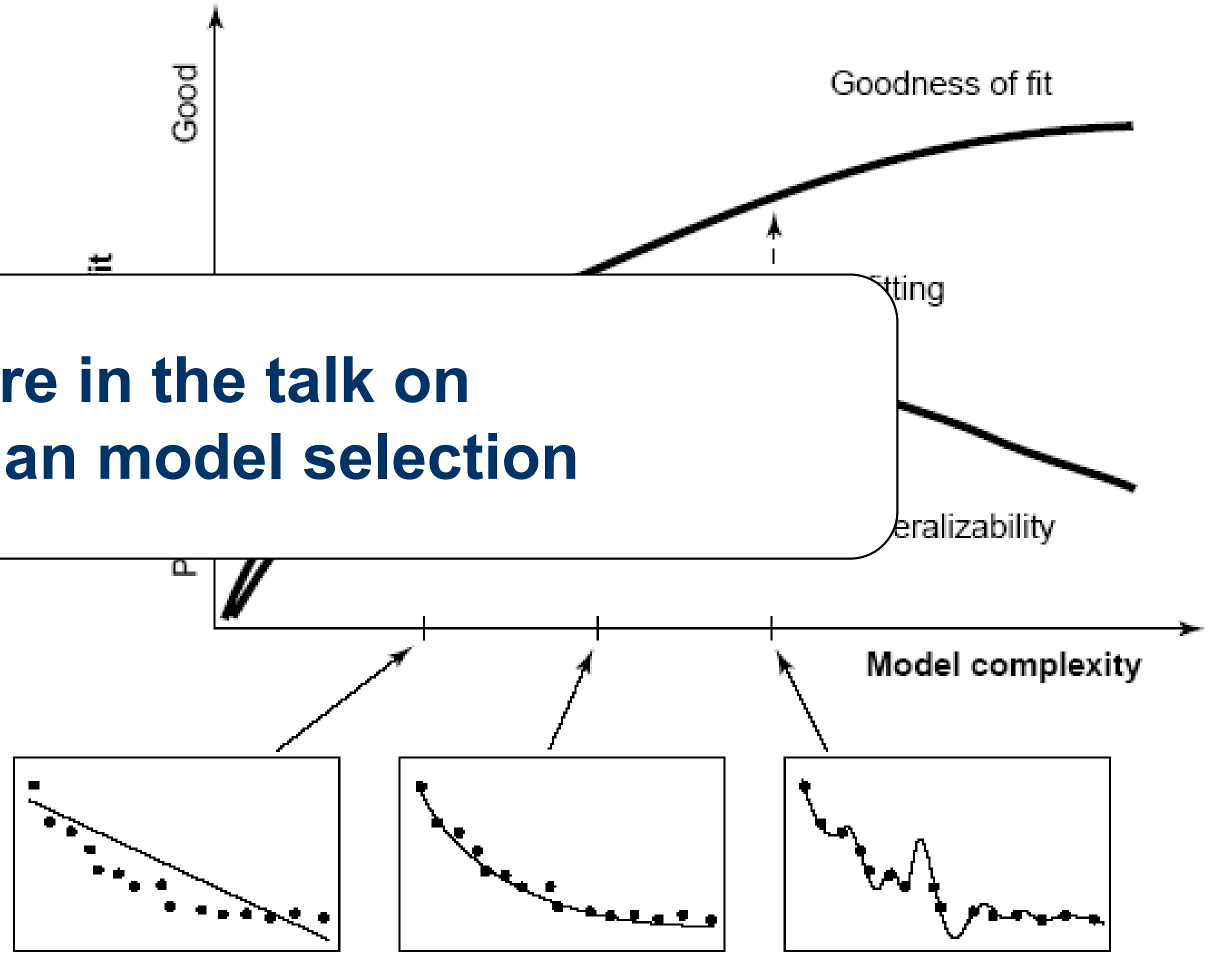
Model comparison and selection

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

More in the talk on
Bayesian model selection

Which mode
best balance
fit and model complexity?

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



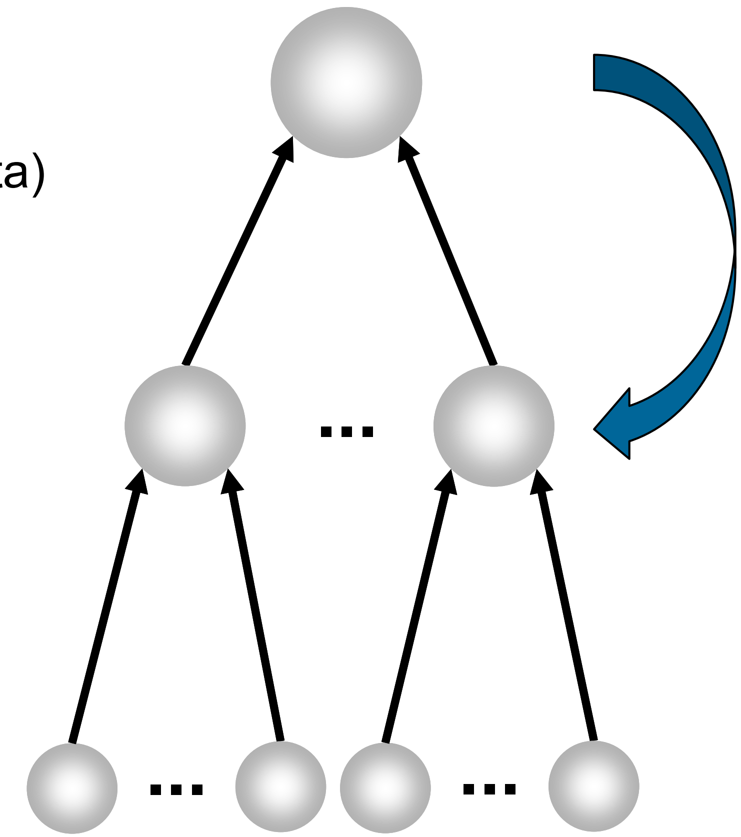
Pitt & Miyung (2002) *TICS*

Generative models as a basis for computational assays: key clinical questions

SYMPTOMS
(behavioural or physiological data)

MECHANISMS
(computational, physiological)

CAUSES
(aetiology)



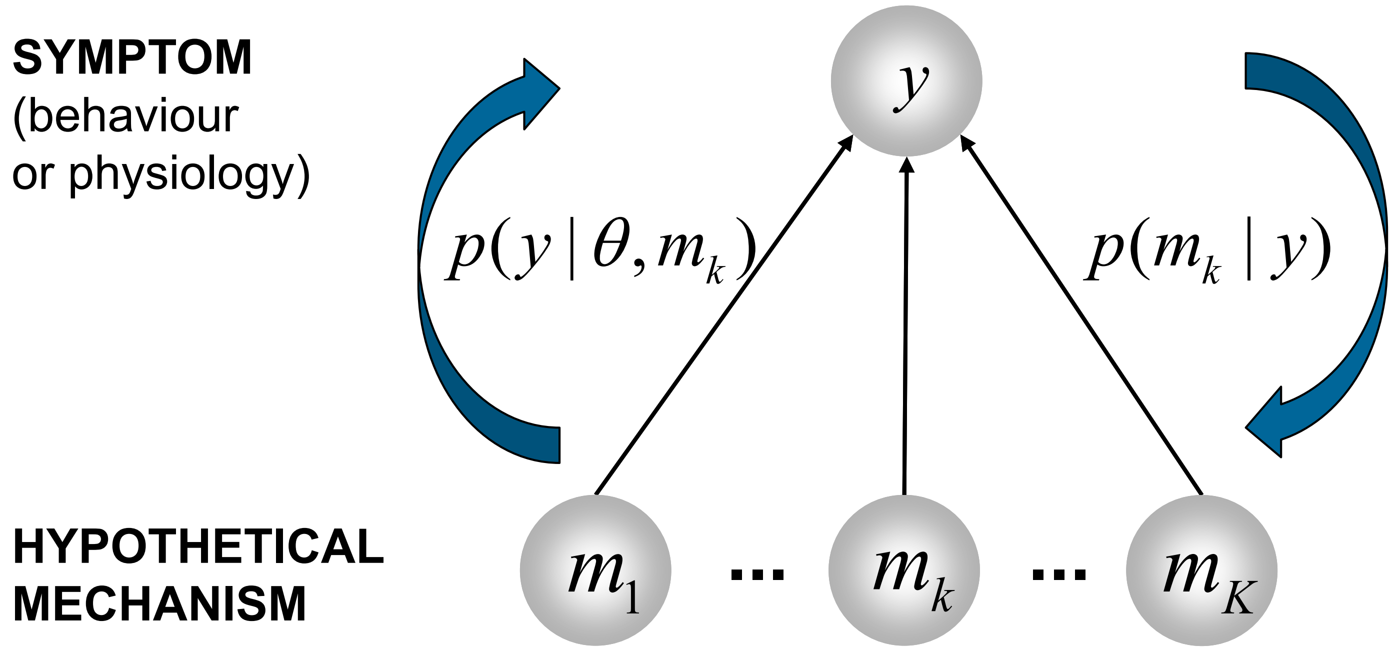
❶ **differential diagnosis** of alternative disease mechanisms

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

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

Stephan: *Translational Neuromodeling & Computational Psychiatry, in prep.*

Model selection for differential diagnosis

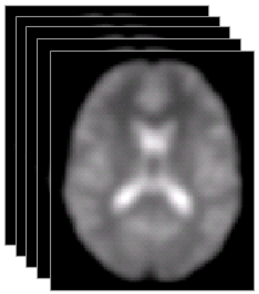


$$p(m_k \mid \mathrm{y}) = \frac{p(y \mid m_k)\, p(m_k)}{\sum_k p(y \mid m_k)\, p(m_k)}$$

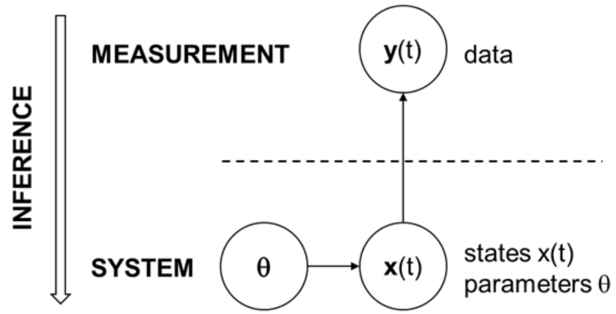
Stephan et al. 2017, *NeuroImage*

Generative embedding

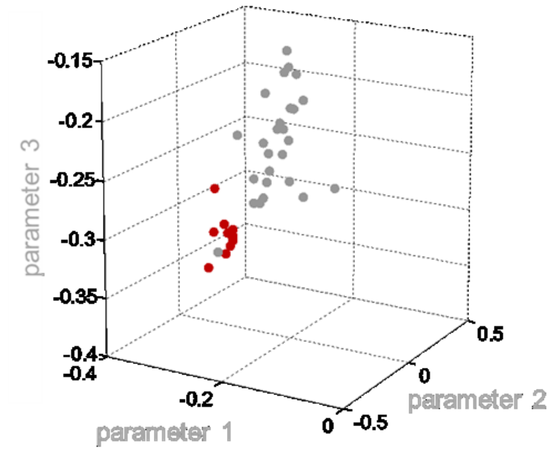
high-dimensional data



generative model



mechanistic interpretation

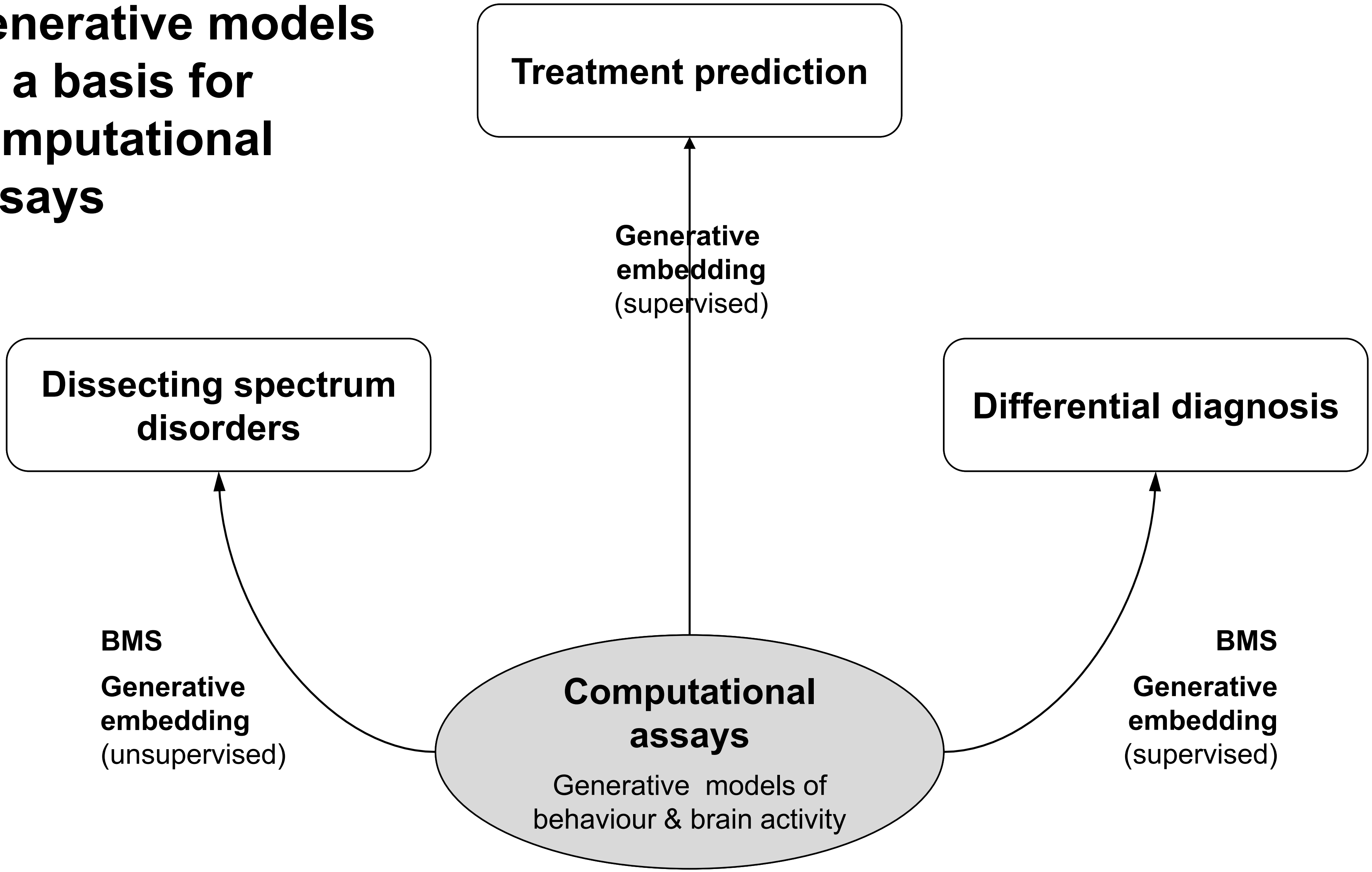


theory-driven
dimensionality reduction

posterior densities →
features for machine learning

Brodersen et al. 2011, *PLoS Comp. Biol.*
Brodersen et al. 2014, *NeuroImage Clinical*

Generative models as a basis for computational assays



adapted from:
Stephan & Mathys 2014, *Curr. Opin. Neurobiol.*

Further reading

Bayesian inference:

- Bishop CM (2006). Machine learning and pattern recognition. Springer, Heidelberg.

A simple introduction to General System Theory (in the context of neuroimaging):

- Stephan KE (2004) On the role of general system theory for functional neuroimaging. Journal of Anatomy 205: 443-470.

A generative modeling strategy for clinical applications:

- Stephan KE, Mathys C (2014) Computational Approaches to Psychiatry. Current Opinion in Neurobiology 25:85-92.
- 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