

# Introduction to Computational Modeling: Generative Models

Klaas Enno Stephan



Translational Neuromodeling Unit

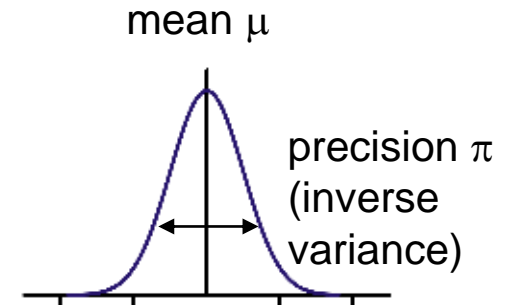


Universität  
Zürich<sup>UZH</sup>



Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

# A brief note on mathematical notations



- For example: Gaussian (Normal) distributions

- for scalars:  $p(x) = N(x; \mu, \sigma^2)$   $\mu = \text{mean}; \sigma^2 = \text{variance}$

- for vectors:  $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$   $\boldsymbol{\Sigma} = \text{covariance matrix}$   
 $= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$

- same thing, just expressed wrt. precision

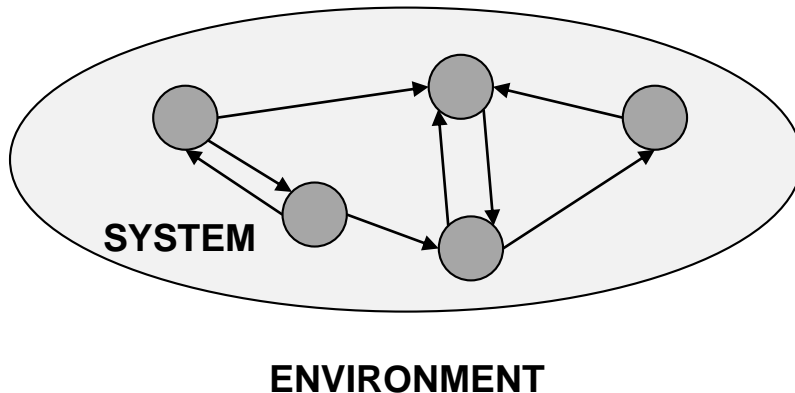
- for scalars:  $p(x) = N(x; \mu, \lambda^{-1})$   $\mu = \text{mean}; \lambda = 1/\sigma^2 = \text{precision}$

- for vectors:  $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$   $\boldsymbol{\Lambda} = \text{precision matrix}$

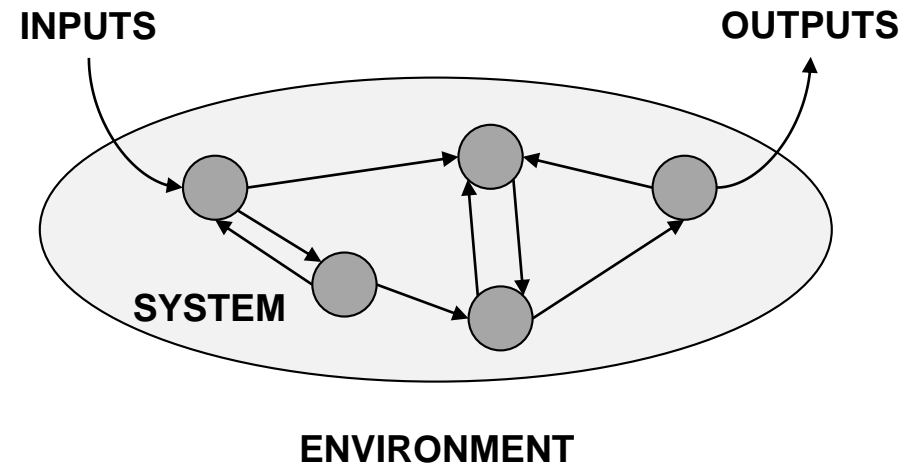
# Systems

- 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**

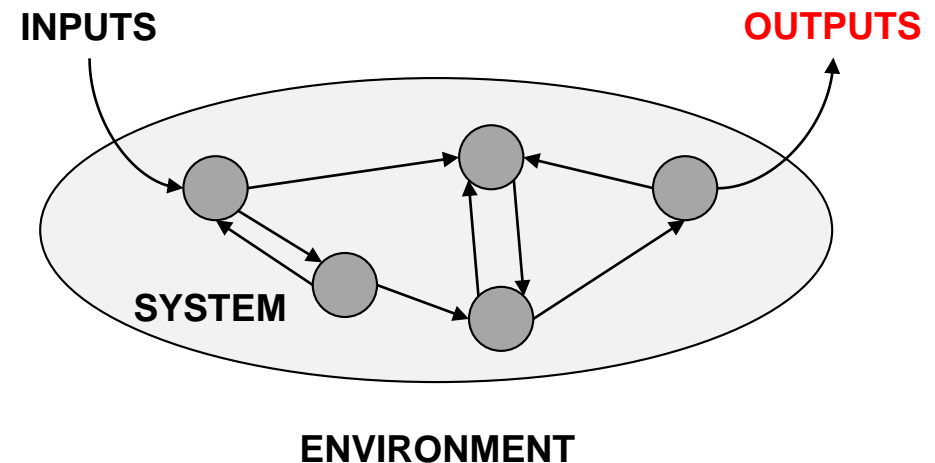


**open system**



# 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**  $\theta$
  - 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.:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \theta_f, \mathbf{u}(t)) \quad \text{as differential equation}$$

$$\mathbf{x}(t+1) = f(\mathbf{x}(t), \theta_f, \mathbf{u}(t)) \quad \text{as difference equation}$$

# State space representation

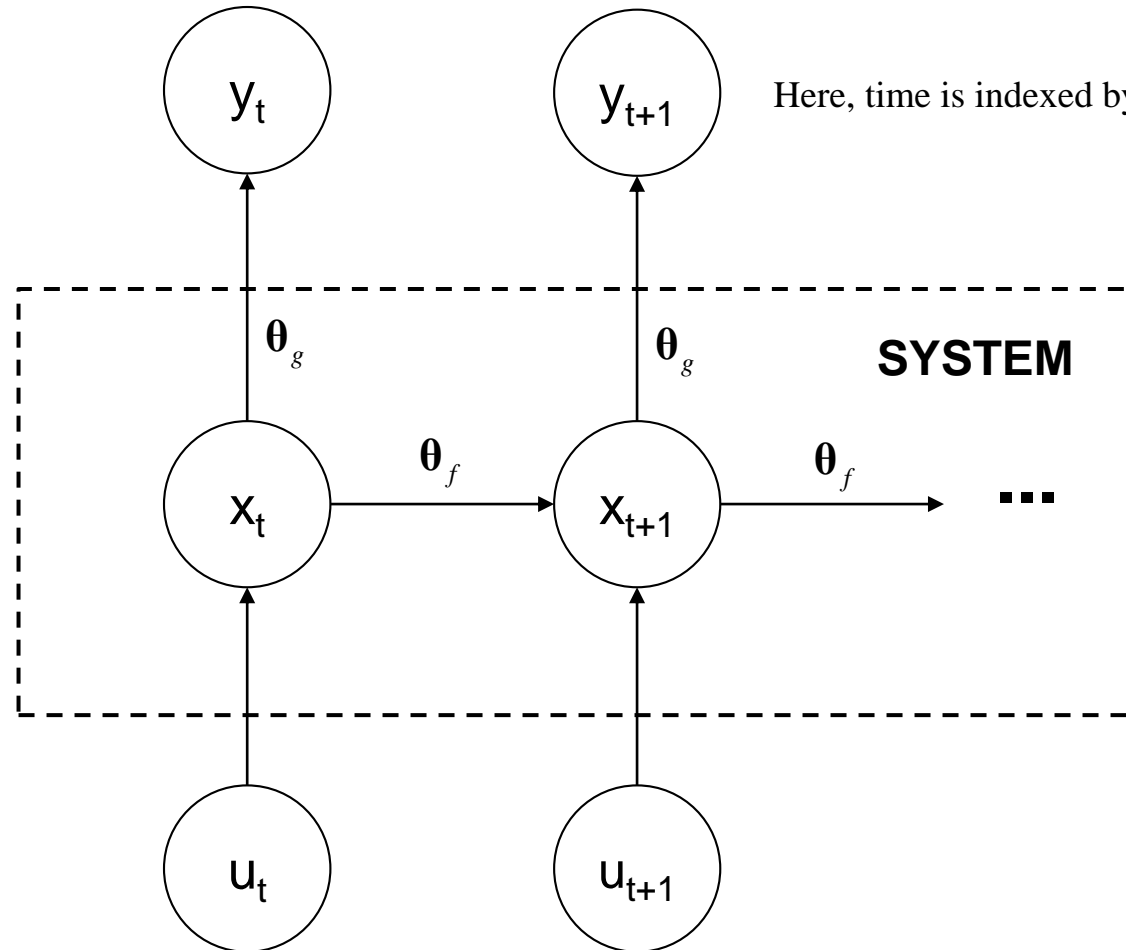
measurement  
(or observation, response)  
equation:

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

**ENVIRONMENT**

inputs

observed system behaviour



Here, time is indexed by subscripts.

# Deterministic vs. stochastic state space models

- **deterministic models**

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

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

- if inputs and initial state are known, inference on parameters sufficient to reconstruct state trajectories

- **stochastic models**

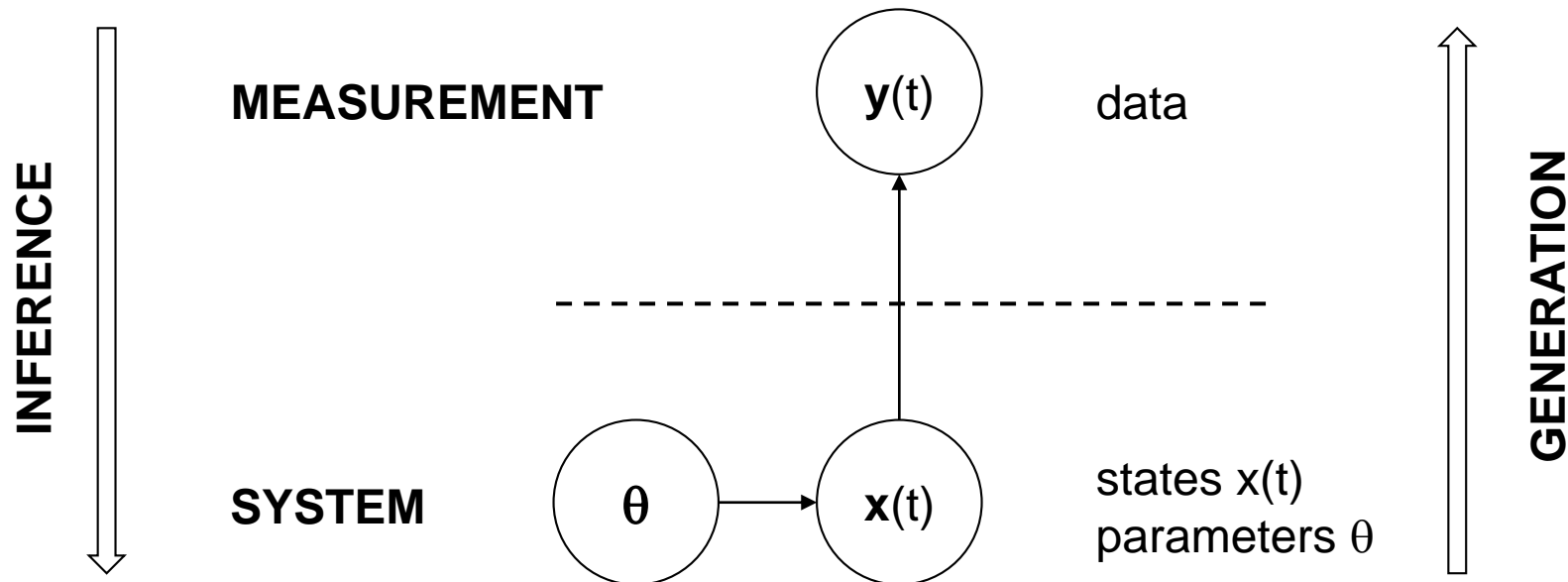
- state noise:  $\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \boldsymbol{\theta}_f, \mathbf{u}(t)) + \omega(t)$  SDEs

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

- much tougher inference problem!

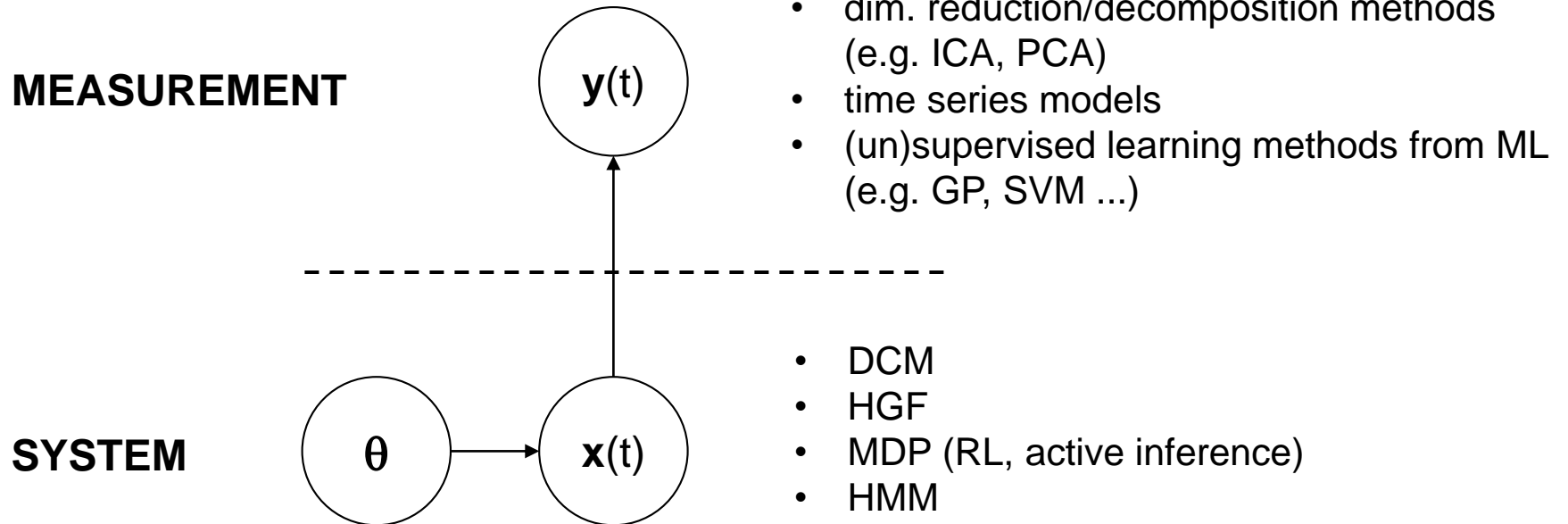
# Models with/without latent states

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





# Examples of approaches with/without latent states



# 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**  $p(\mathbf{y}|\boldsymbol{\theta})$ , i.e.:  
Given a particular value of  $\boldsymbol{\theta}$ , how likely are the observed data  $\mathbf{y}$  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)**:

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

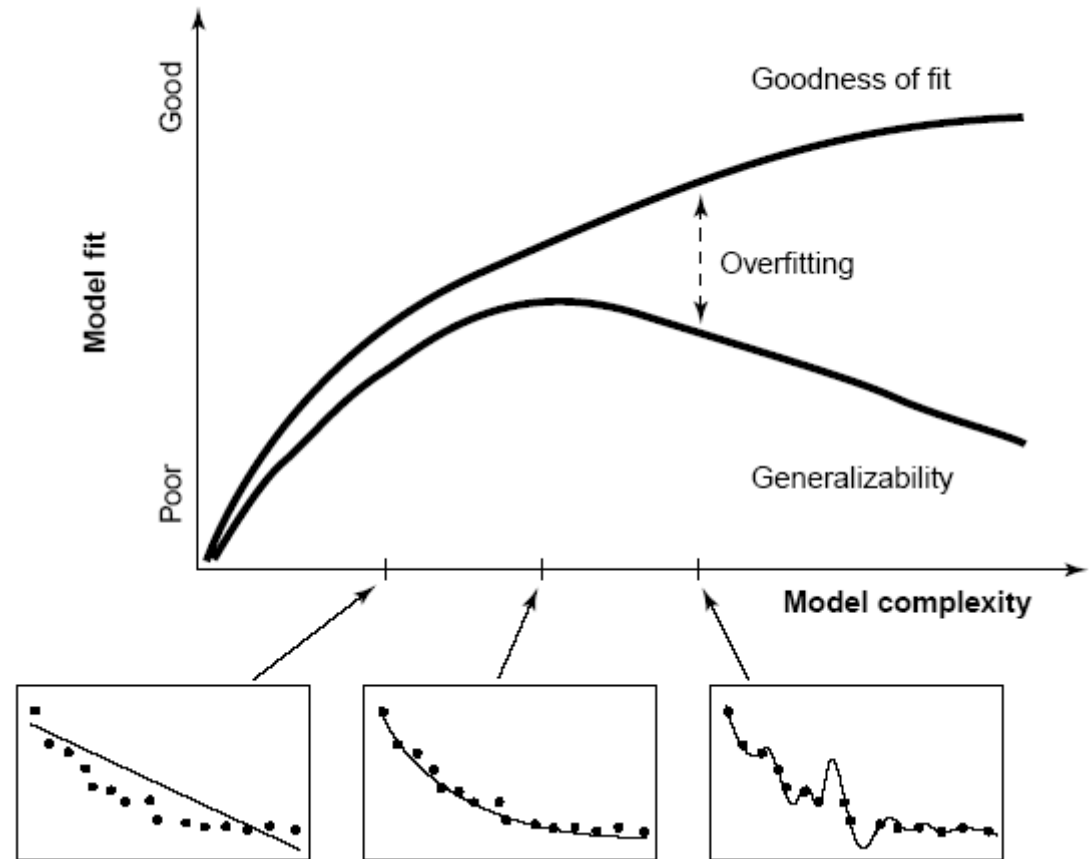
# Maximum likelihood estimation (MLE)

- 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., Gaussian), we can write down the likelihood function. Given a parameter value  $\theta$ , the likelihood is the probability of the chosen data given  $\theta$ .  
**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)**:

$$\hat{\theta}_{ML} = \arg \max_{\theta} \ln p(\mathbf{y} | \theta)$$

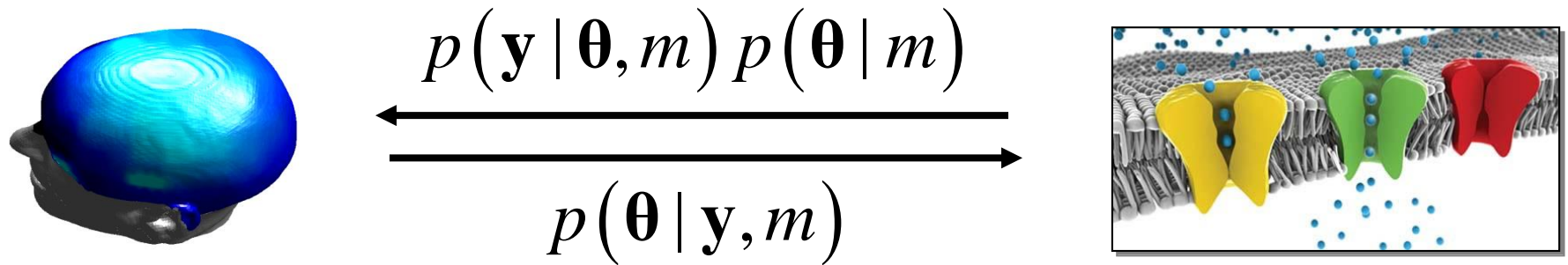
# Overfitting

- 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*

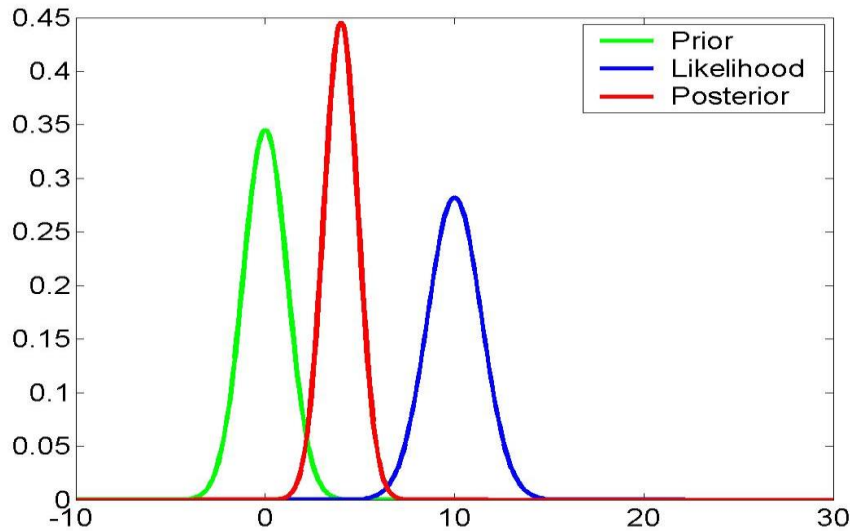
# Generative models



$y$  = data,  $\theta$  = 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  $\rightarrow$  posterior  $p(\theta|y,m)$
5. natural basis for model comparison  $\rightarrow$  model evidence  $p(y|m)$

# Bayes' rule

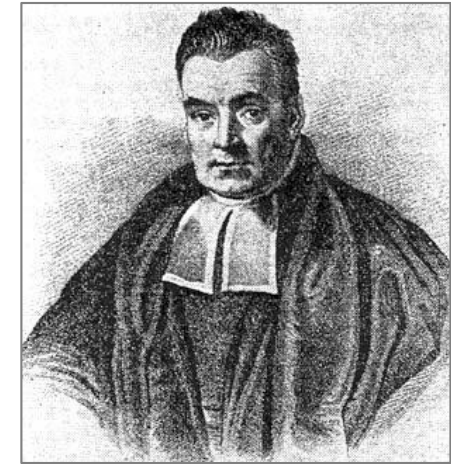


Likelihood  $\times$  prior: generative model

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)}$$

$\theta$ : parameters  
 $y$ : data

**Model evidence:** normalisation  
term and index for model goodness

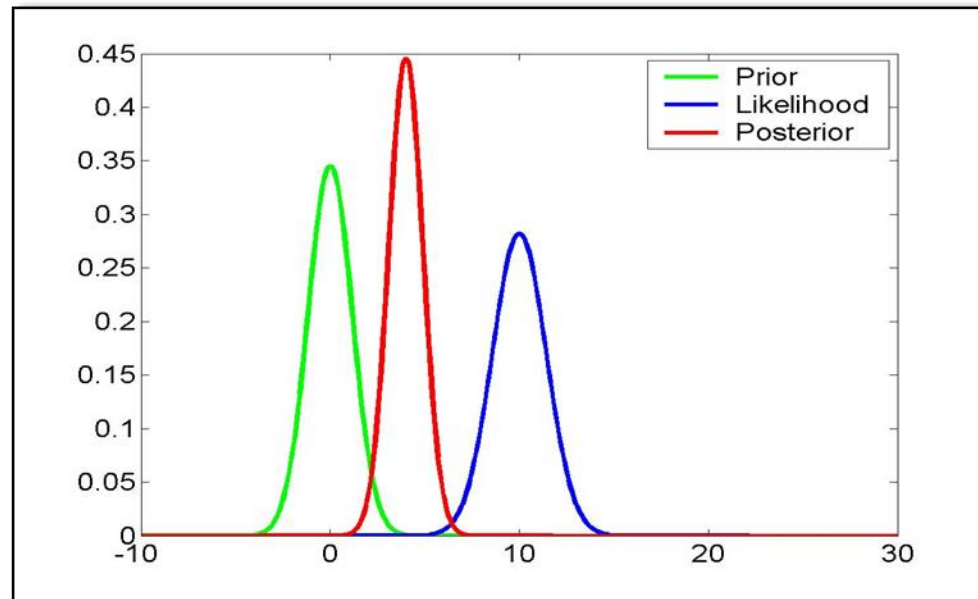


The Reverend Thomas Bayes  
(1702-1761)

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

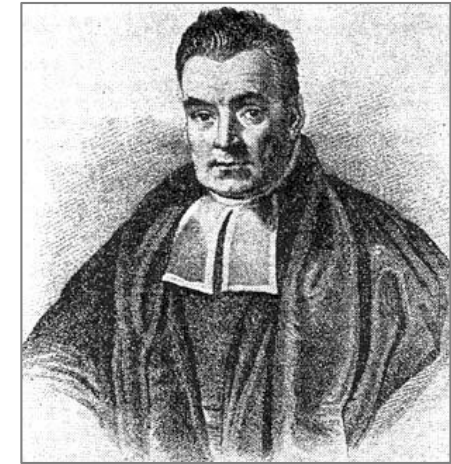
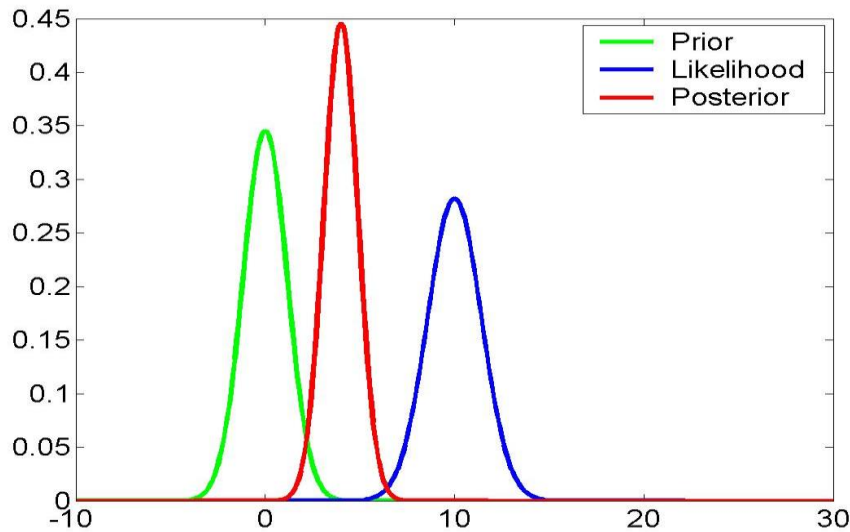
*Wikipedia*

# Bayesian inference: an animation



Code courtesy by Guillaume Flandin

# Bayes' rule



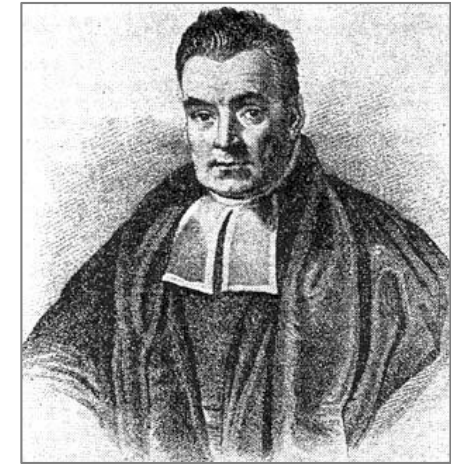
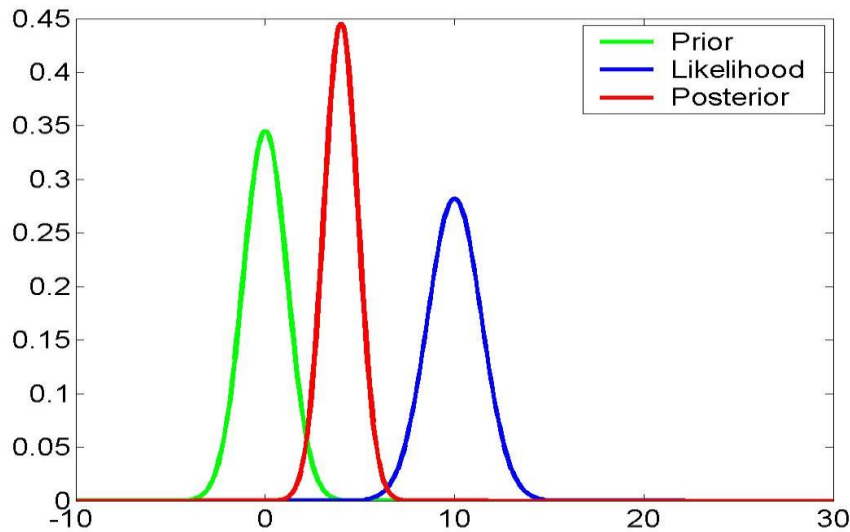
The Reverend Thomas Bayes  
(1702-1761)

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

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



# Bayes' rule



The Reverend Thomas Bayes  
(1702-1761)

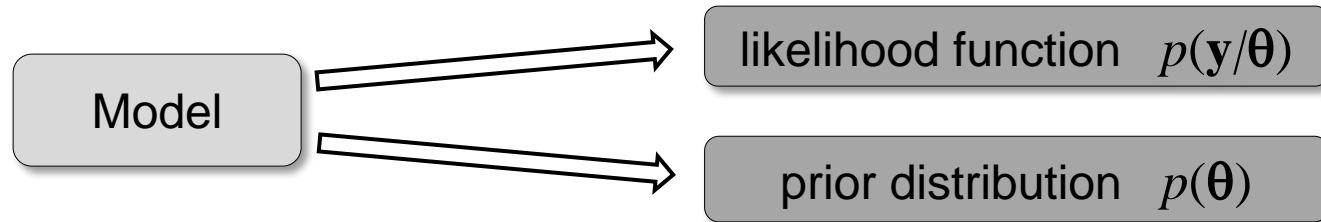
$$p(\boldsymbol{\theta} \mid \mathbf{y}, m) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}, m) p(\boldsymbol{\theta} \mid m)}{\int p(\mathbf{y} \mid \boldsymbol{\theta}, m) p(\boldsymbol{\theta} \mid m) d\boldsymbol{\theta}}$$

## **Evidence:**

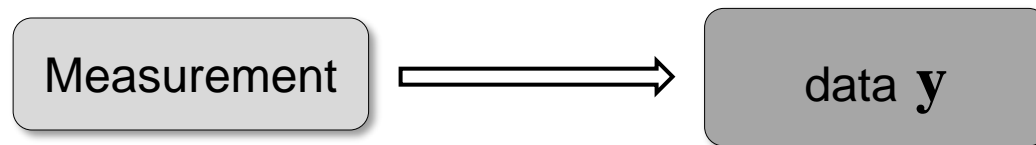
likelihood that data were generated by model  $m$ , averaging over all possible parameter values (as weighted by the prior).

# Principles of generative modeling

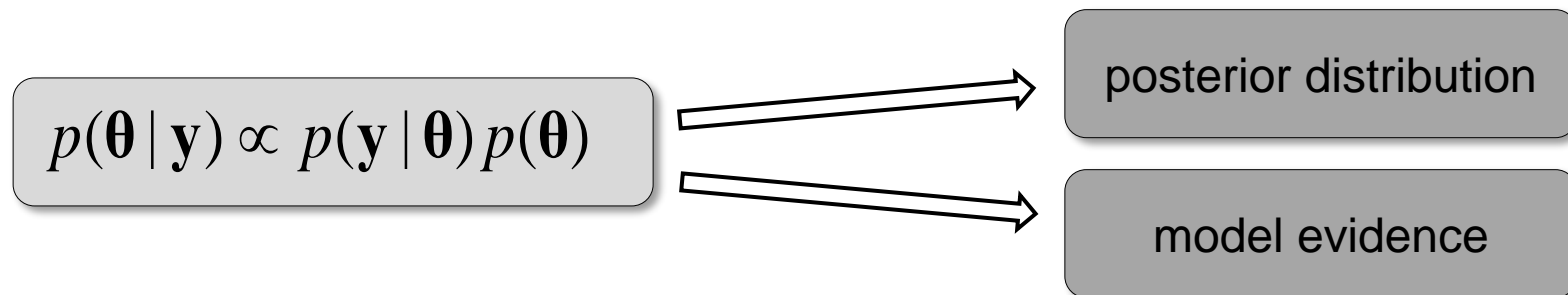
⇒ Specifying a **generative model**



⇒ Observation of **data**



⇒ **Model inversion**



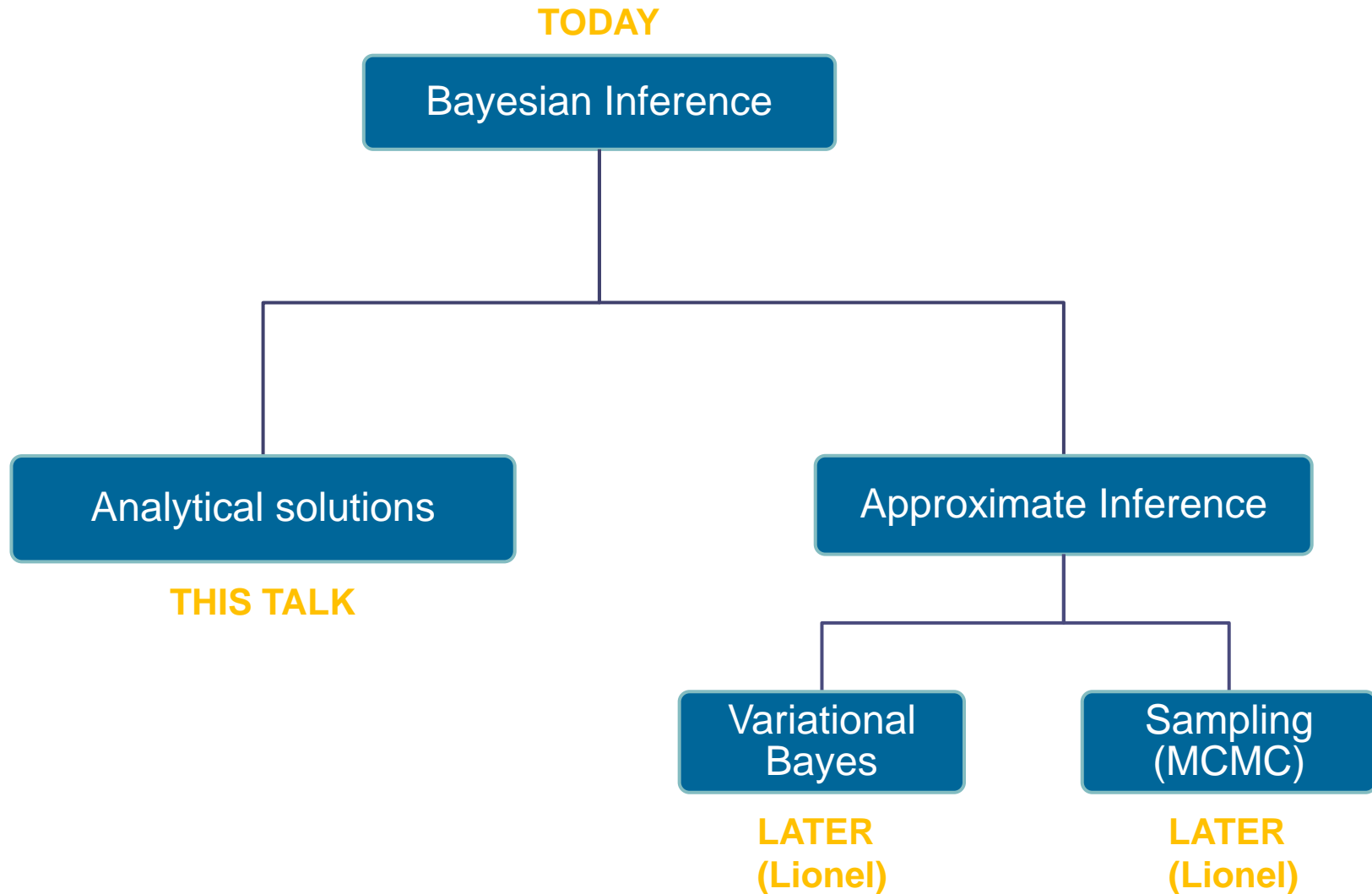
# Maximum a posteriori (MAP) estimation

- 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:

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

- Advantages:
  - prior serves to regularize → can prevent overfitting
  - does not require computing 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  $\times$  Normal
  - Beta  $\propto$  Binomial  $\times$  Beta
  - Dirichlet  $\propto$  Multinomial  $\times$  Dirichlet

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


same form

# A simple example: univariate Gaussian belief update

## Likelihood & prior

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

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

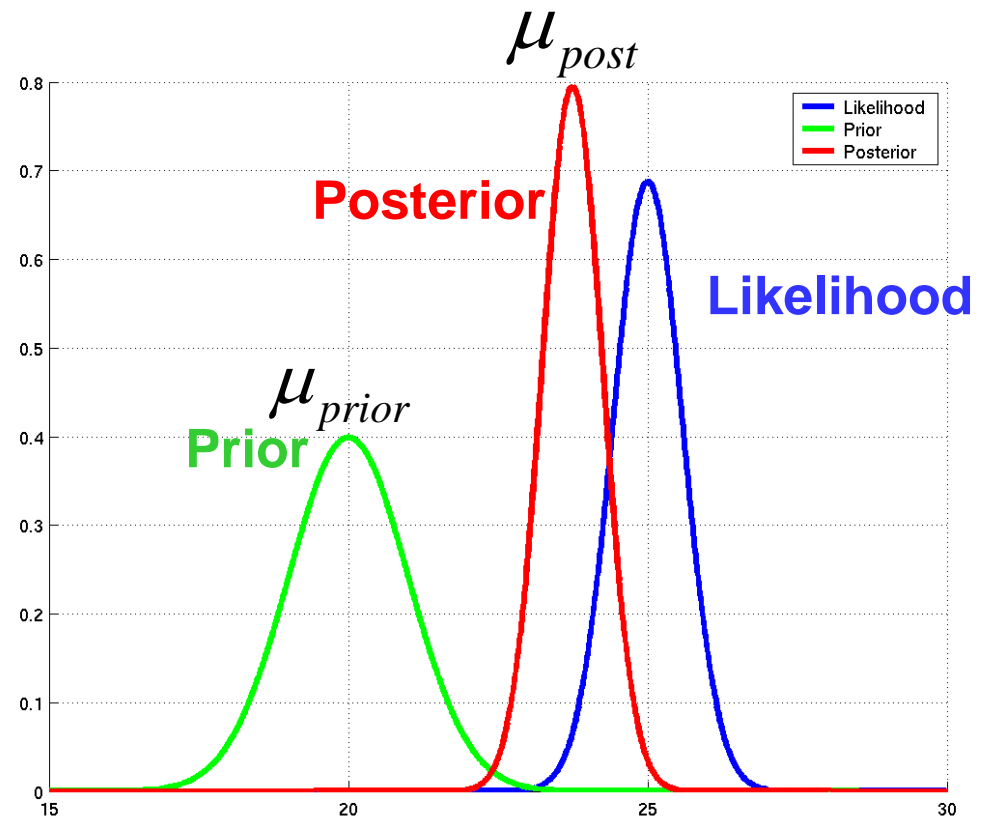
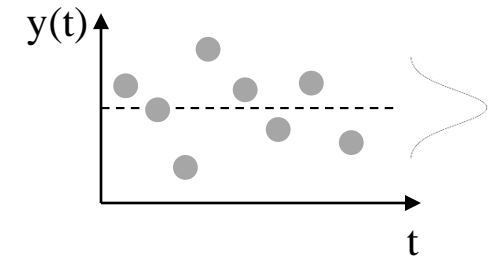
**Posterior**  $p(\theta | y) = N(\mu_{\text{post}}, \lambda_{\text{post}}^{-1})$   
(for a single observation  $y$ )

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

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

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

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



# A simple example: univariate Gaussian belief update

## Likelihood & prior

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

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

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

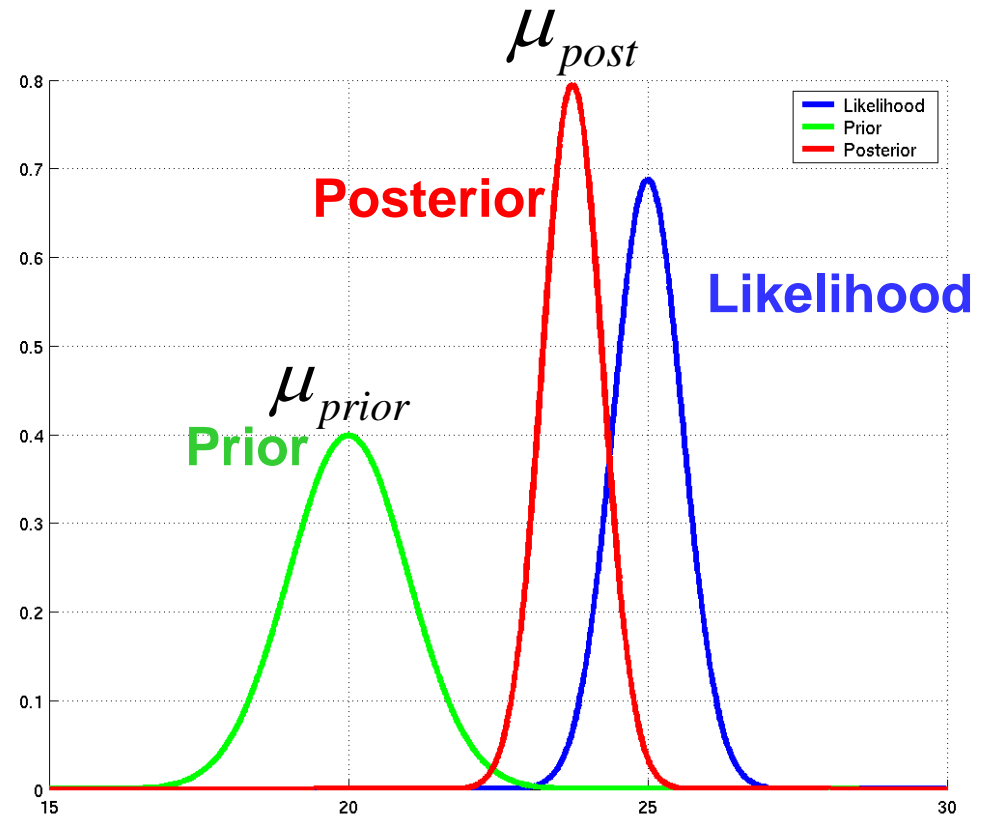
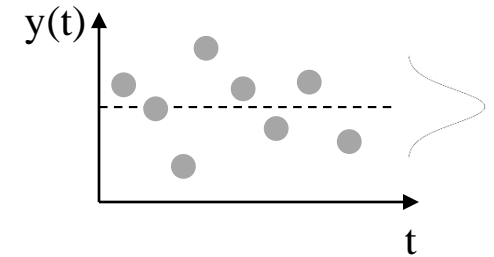
$$\lambda_{\text{post}} = \lambda_e + \lambda_{\text{prior}}$$

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

## relative precision weighting:

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

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

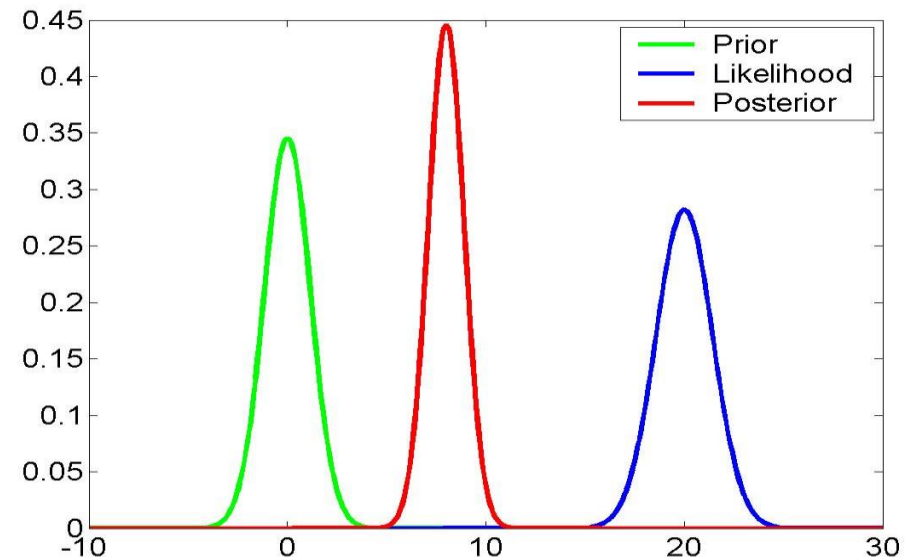
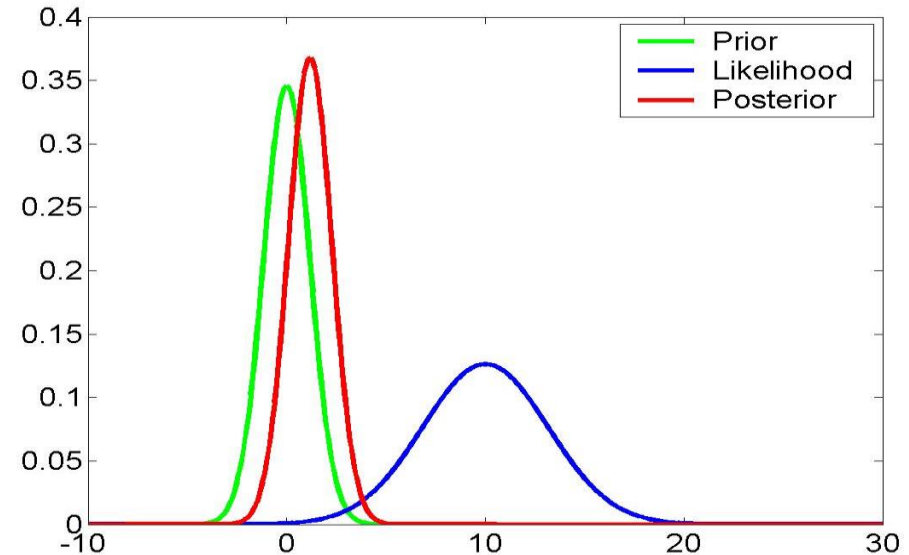




# Choice of priors

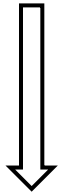
- Objective priors:
  - "non-informative" priors (e.g. Jeffreys' prior)
- Subjective priors:
  - subjective but not arbitrary
  - express beliefs that result from an understanding the problem or system
  - can be result of previous empirical results
  - can accommodate objective constraints (e.g., non-negativity)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

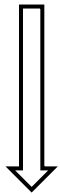


# Model comparison and selection

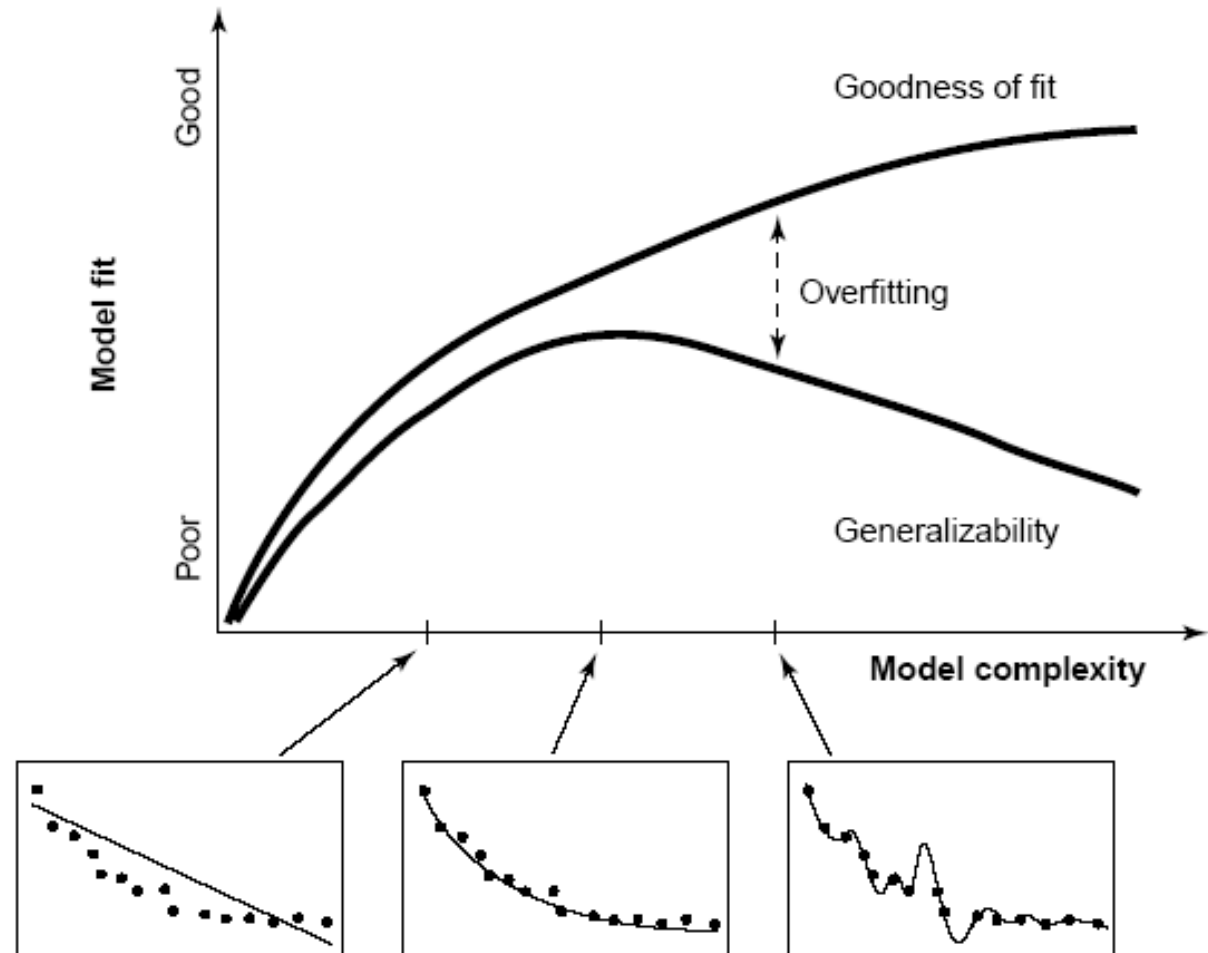
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?



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



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



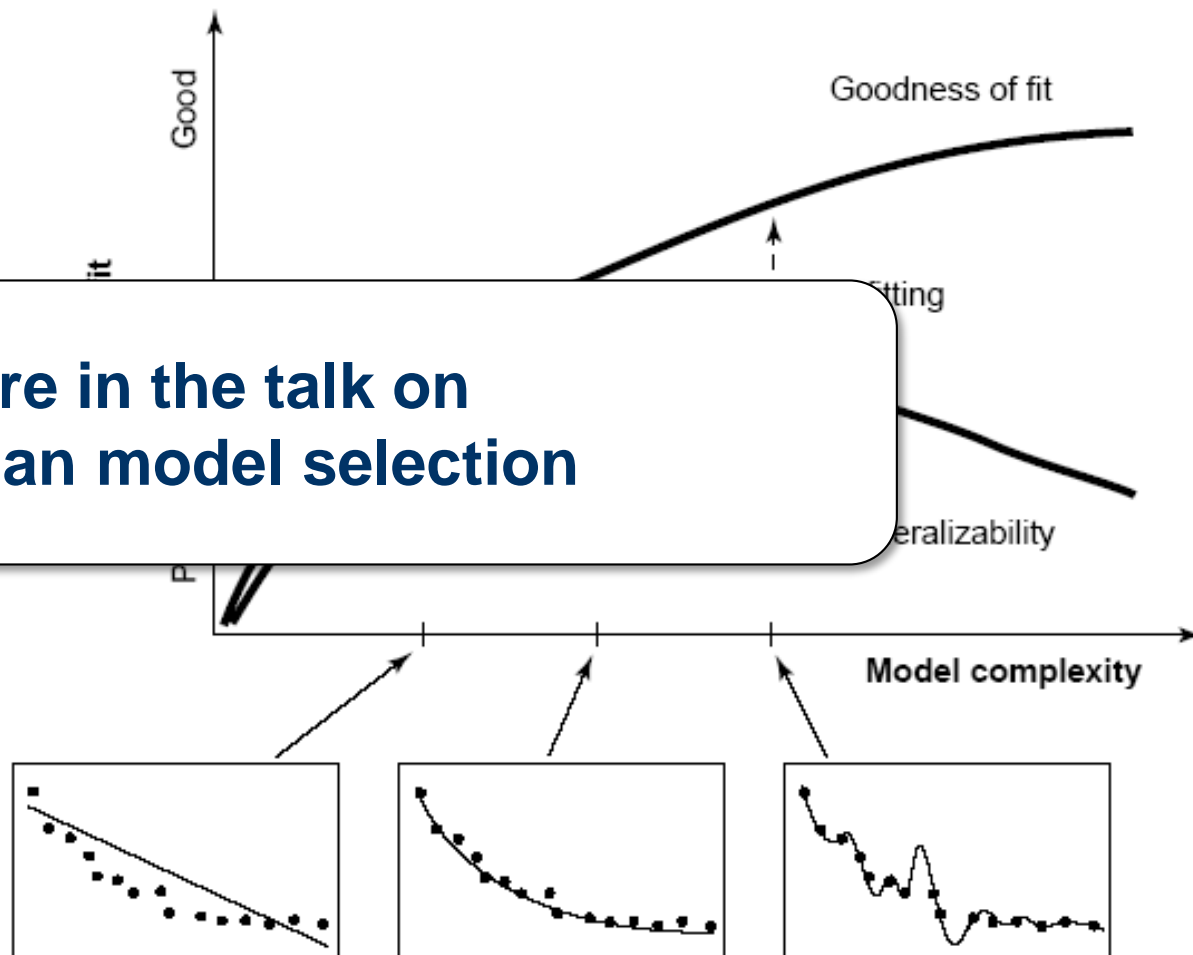
# Model comparison and selection

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

More in the talk on  
Bayesian model selection

Which model best balances fit and model complexity?

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



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

## **SYMPTOMS**

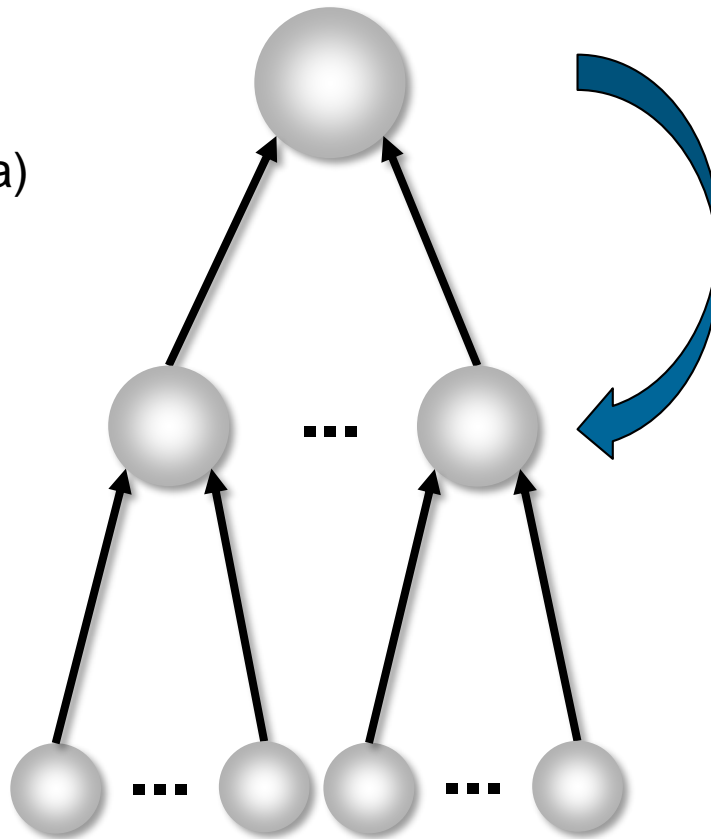
(behavioural or physiological data)

## **MECHANISMS**

(computational, physiological)

## **CAUSES**

(aetiology)

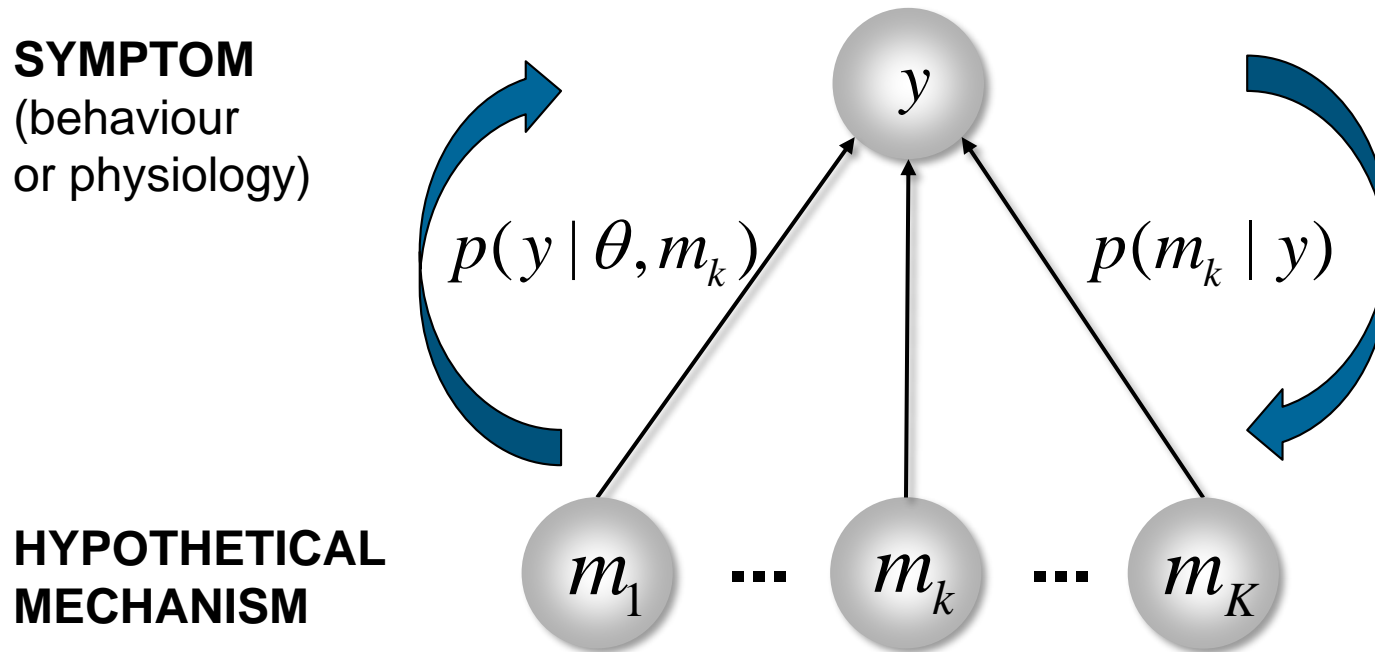


**① differential diagnosis** of alternative disease mechanisms

**② stratification / subgroup detection** into mechanistically distinct subgroups

**③ prediction** of clinical trajectories and treatment response

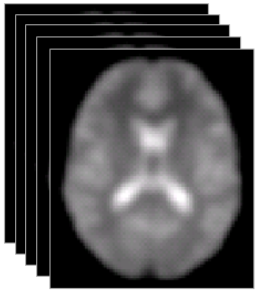
# Model selection for differential diagnosis



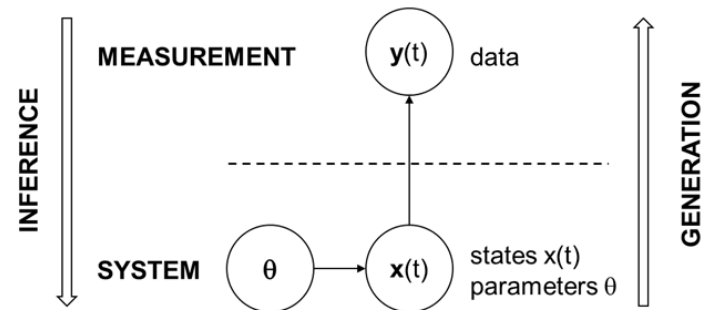
$$p(m_k | y) = \frac{p(y | m_k) p(m_k)}{\sum_k p(y | m_k) p(m_k)}$$

# Generative embedding

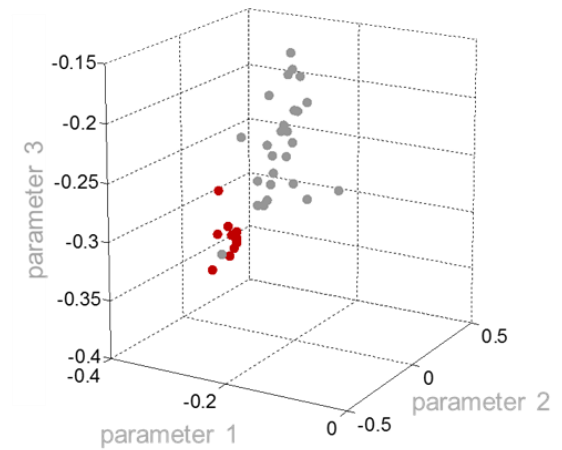
high-dimensional data



generative model



mechanistic interpretation

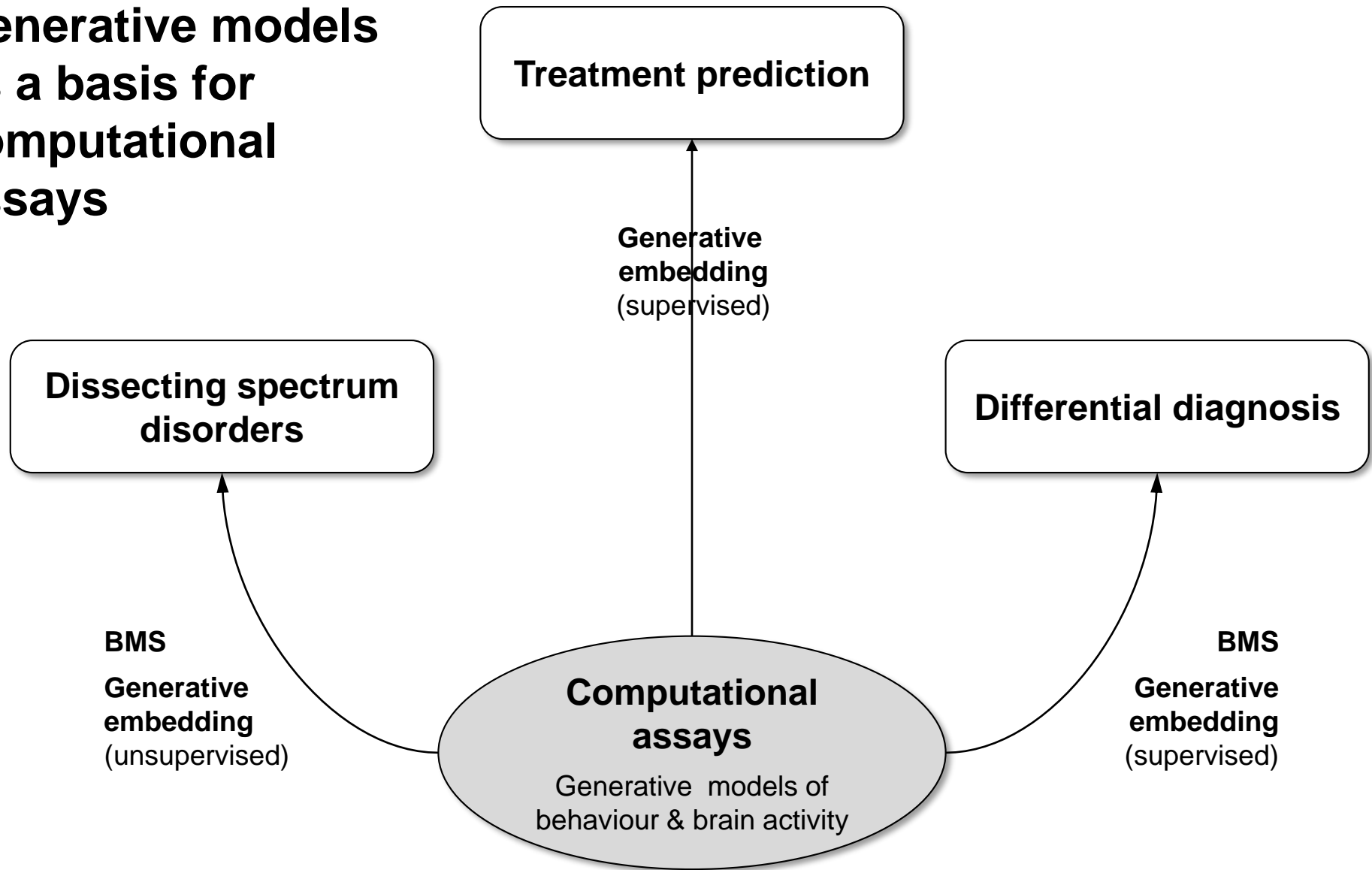


theory-driven  
dimensionality reduction



posterior densities  $\rightarrow$   
features for machine learning

# Generative models as a basis for computational assays



# 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**