Machine Learning + Statistics: Better Together Than Apart

Daniela Huppenkothen

DIRAC Institute, University of Washington

https://github.com/dhuppenkothen/cargese2018_tutorials

All models are wrong,
but some are useful.

— George Box

Iterative Understanding of Nature

Iterative Understanding of Nature

- Experimental Design

Iterative Understanding of Nature

- Experimental Design
- Data Collection

Iterative Understanding of Nature

- Experimental Design
- Data Collection
- Model Formulation

Iterative Understanding of Nature

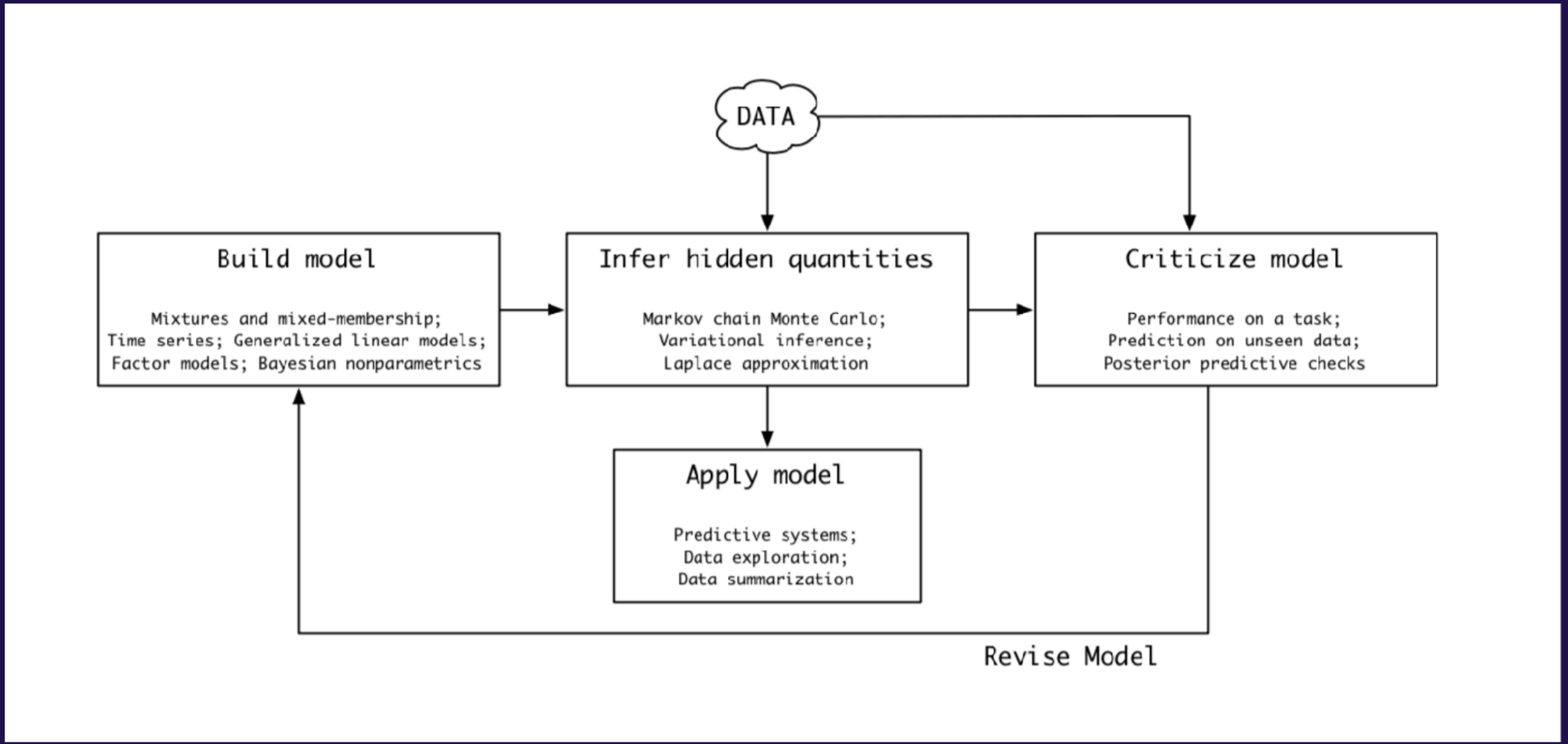
- Experimental Design
- Data Collection
- Model Formulation
- Model Criticism

Iterative Understanding of Nature

- Experimental Design
- Data Collection
- Model Formulation
- Model Criticism

Iterative Understanding of Nature

- Experimental Design
- Data Collection
- Model Formulation
- Model Criticism


DATA
Build model
Mixtures and mixed-membership;
Time series; Generalized linear models;
Factor models; Bayesian nonparametrics
Infer hidden quantities
Markov chain Monte Carlo;
Variational inference;
Laplace approximation
Criticize model
Performance on a task;
Prediction on unseen data;
Posterior predictive checks
Apply model
Predictive systems;
Data exploration;
Data summarization
Revise Model

Blei (2014)

Building + Revising Models

Building + Revising Models Part III

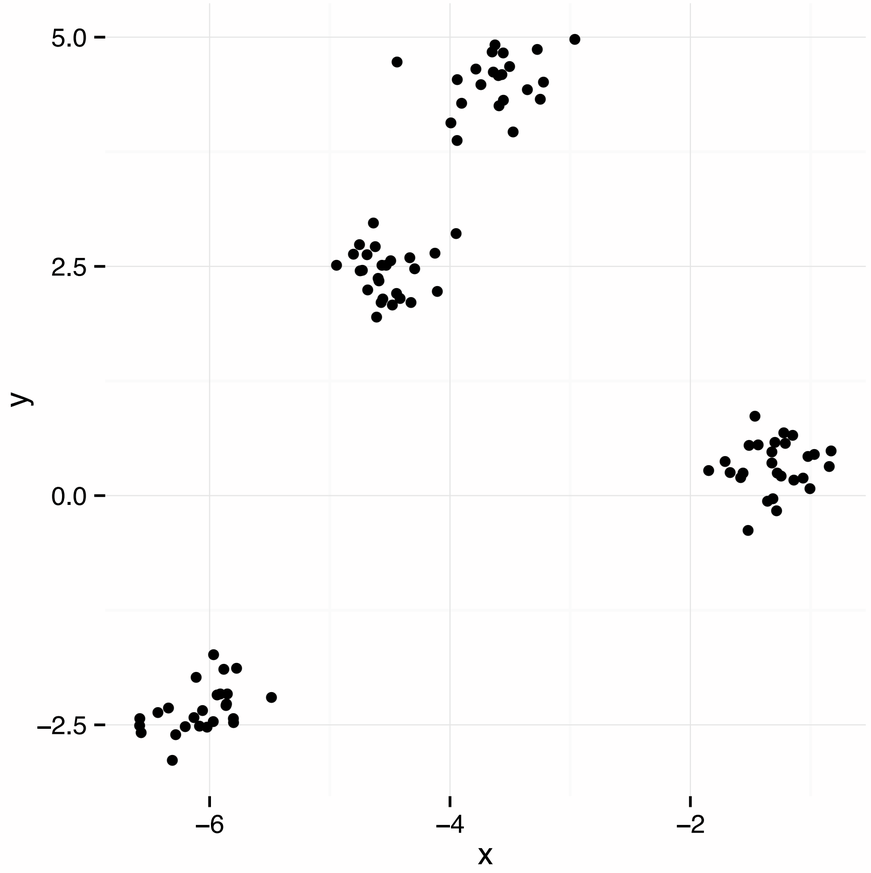
Building + Revising Models Part III (sort of)

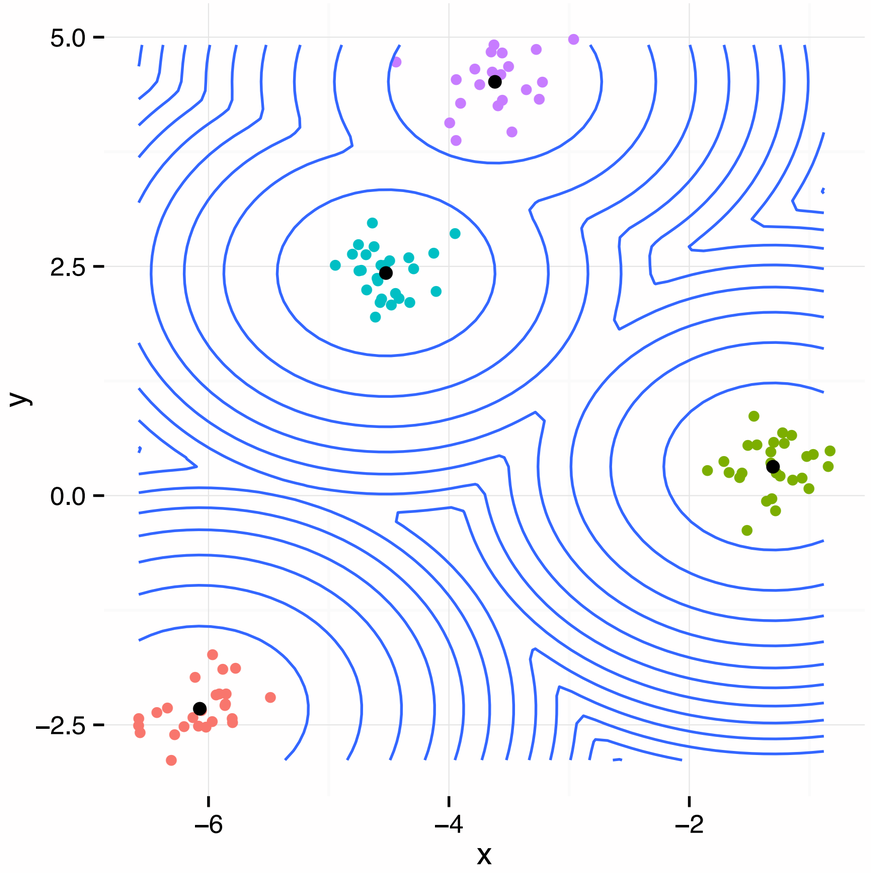
Wednesday: Bayesian Hierarchical Models

Yesterday: Machine Learning

Today: Both!

Gaussian Mixture Models





Blei (2014)

(1) The Generative Process

1. Draw mixture proportions $\theta \sim \text{Dirichlet}(\alpha)$
2. For each mixture component k, draw $\mu_k \sim \mathcal{N}(0, \sigma^2)$
3. For each data point i:
   a. Draw mixture assignment $z_i \mid \theta \sim \text{Discrete}(\theta)$
   b. Draw data point $x_i \mid z_i, \mu \sim \mathcal{N}(\mu_{z_i}, 1)$

(2) The Joint Probability Distribution

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

(2) The Joint Probability Distribution

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

cluster proportions

(2) The Joint Probability Distribution

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

cluster proportions

Gaussian means

(2) The Joint Probability Distribution

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

cluster proportions

Gaussian means

cluster assignments

(2) The Joint Probability Distribution

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

cluster proportions

Gaussian means

cluster assignments

data points

(2) The Joint Probability Distribution

Dirichlet

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

cluster proportions

Gaussian means

cluster assignments

data points

(2) The Joint Probability Distribution

Dirichlet Normal

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

cluster proportions

Gaussian means

cluster assignments

data points

(2) The Joint Probability Distribution

Dirichlet Normal Discrete

$$p(\theta, \mu, z, x \mid \sigma_0^2, \alpha) = p(\theta \mid \alpha) \prod_{k=1}^{K} p(\mu_k \mid \sigma_0^2) \prod_{i=1}^{N} (p(z_i \mid \theta) p(x_i \mid z_i, \mu, z_i))$$

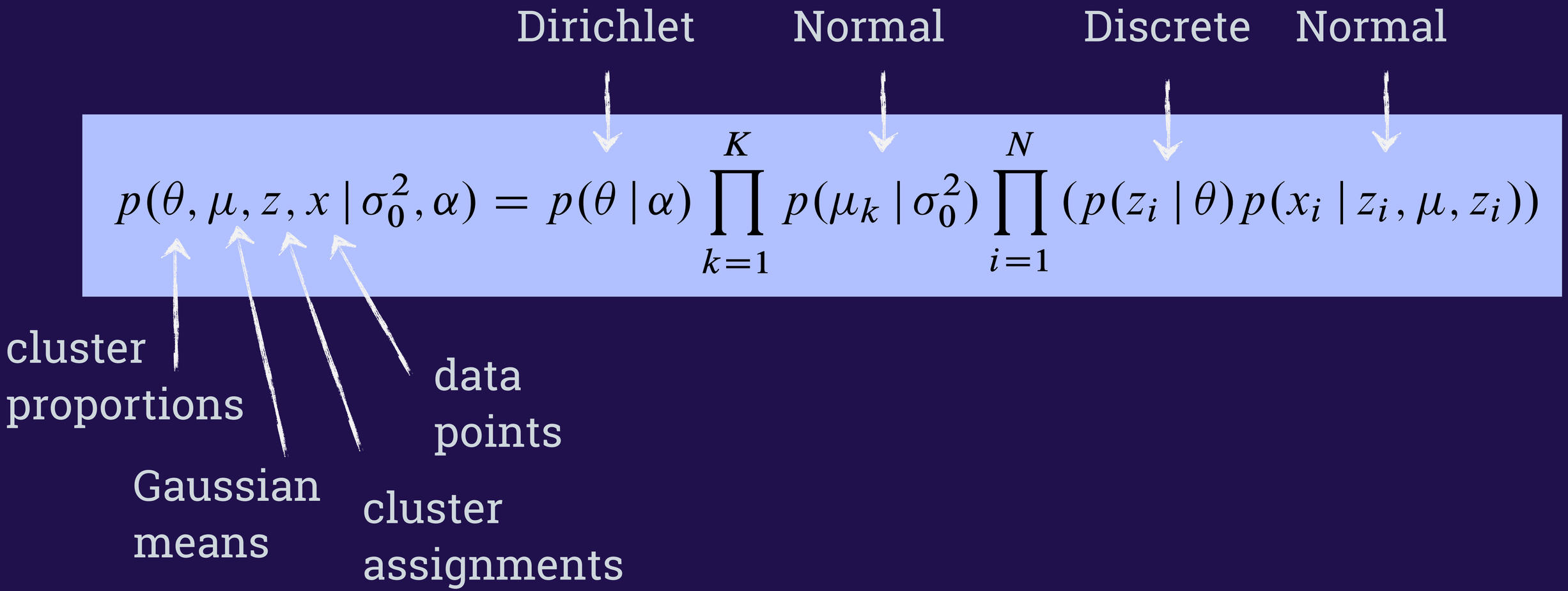
cluster proportions

Gaussian means

cluster assignments

data points

(2) The Joint Probability Distribution



(3) The Probabilistic Graphical Model

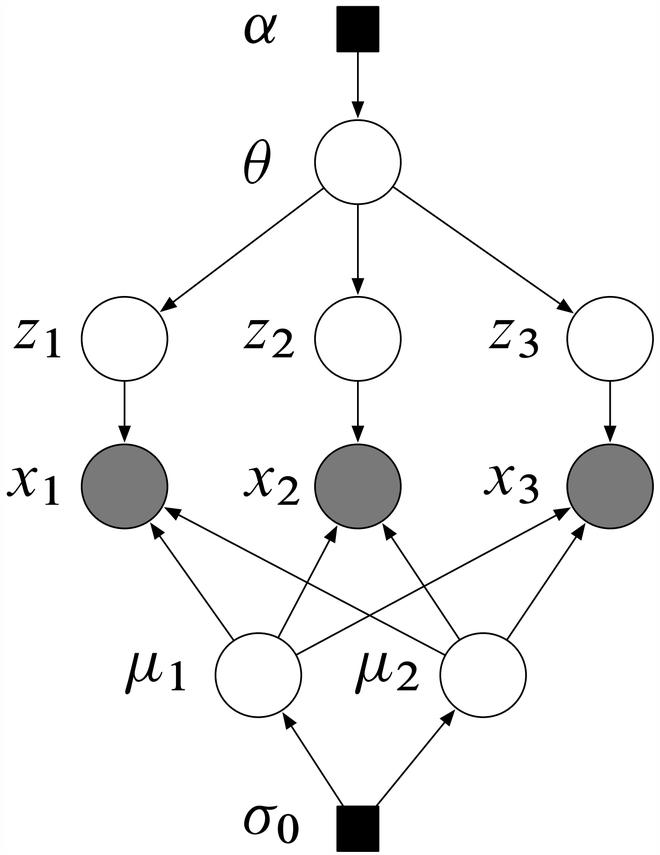
Exercise: Can you draw a graphical model for the Gaussian Mixture Model?

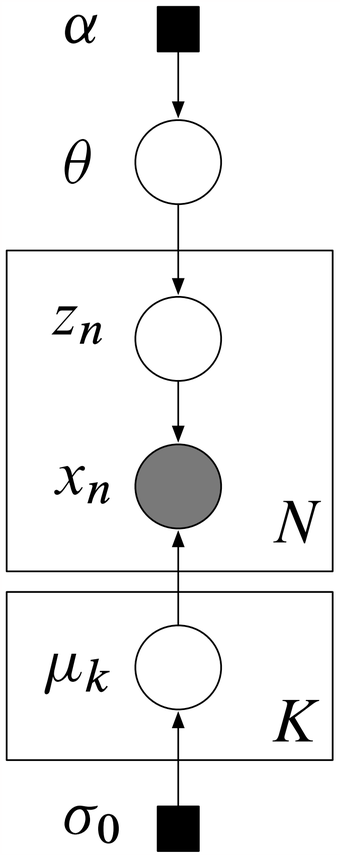
Blei (2014)

(3) The Probabilistic Graphical Model

Blei (2014)

(3) The Probabilistic Graphical Model





$\sim \text{Dirichlet}_K(\alpha)$

$\sim \text{Discrete}(\theta)$

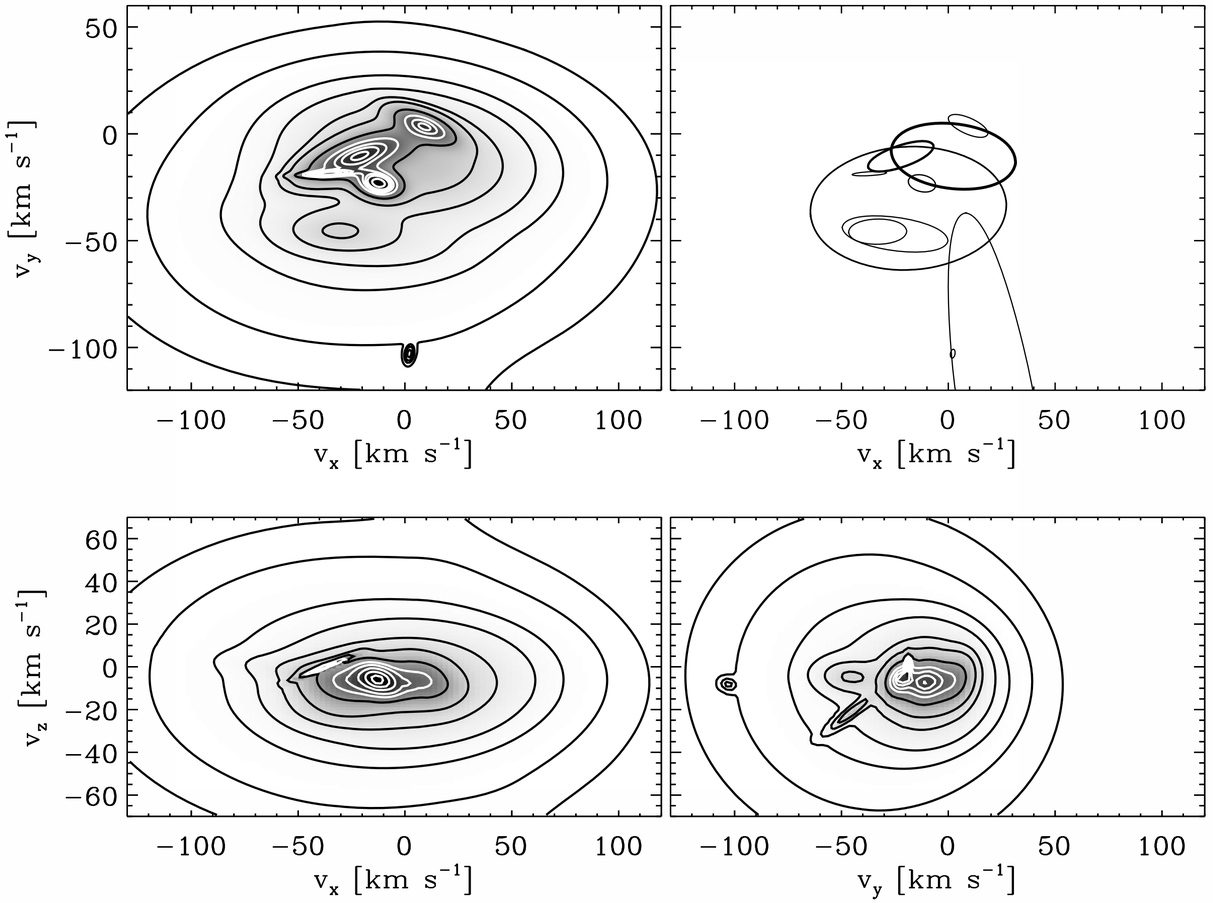
$\sim \mathcal{N}(\mu_{z_n}, \sigma^2)$

$\sim \mathcal{N}(0, \sigma_0^2)$

Blei (2014)

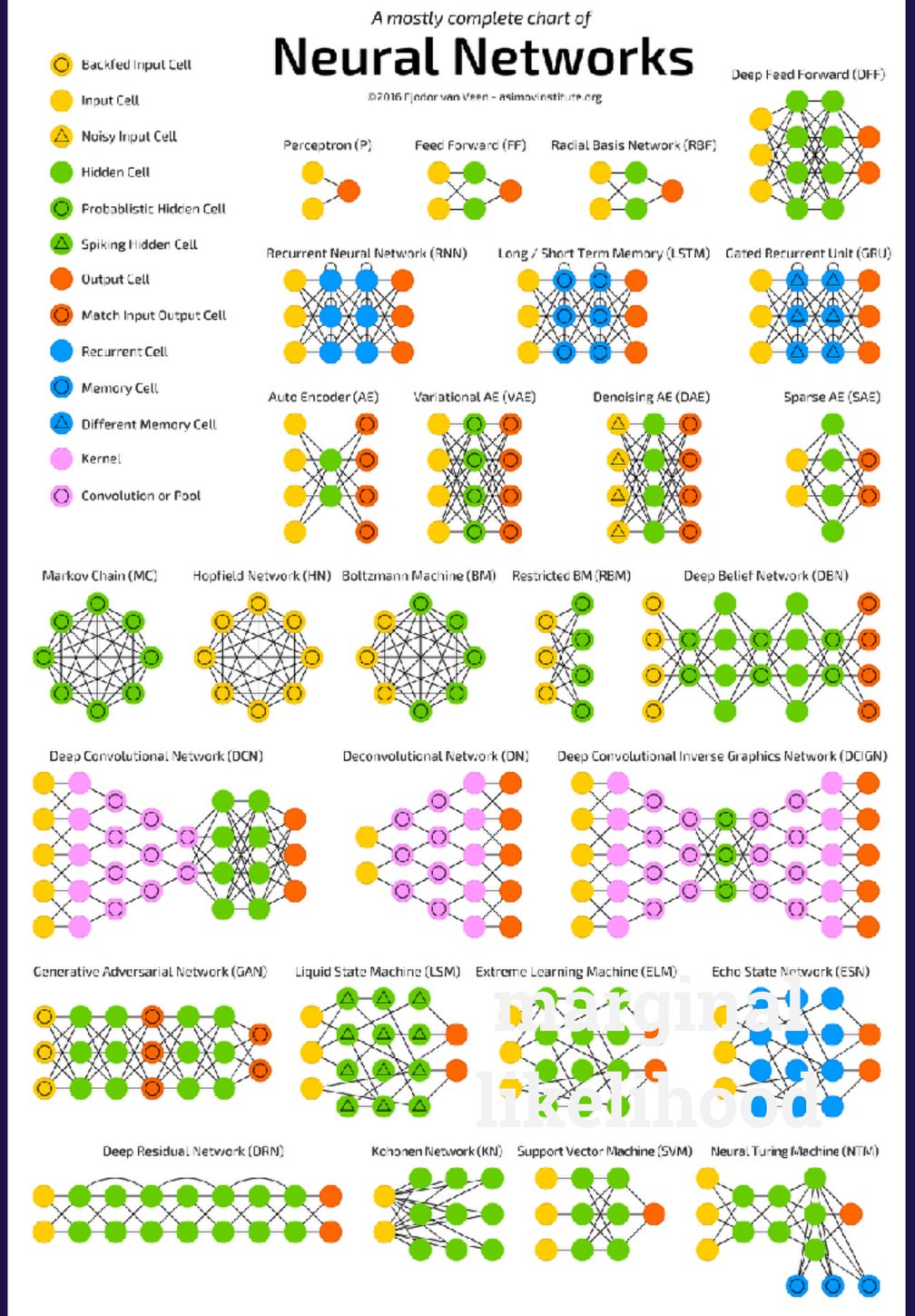
Measurement Uncertainties?
Missing Data?

Extreme Deconvolution



Bovy et al (2009)

Neural Networks + Probabilistic Models



Fjodor van Veen; https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

https://playground.tensorflow.org

Using Neural Networks as Emulators

posterior

likelihood

prior

$$P(\theta \mid x) = \frac{P(x \mid \theta) P(\theta)}{P(x)}$$

marginal likelihood

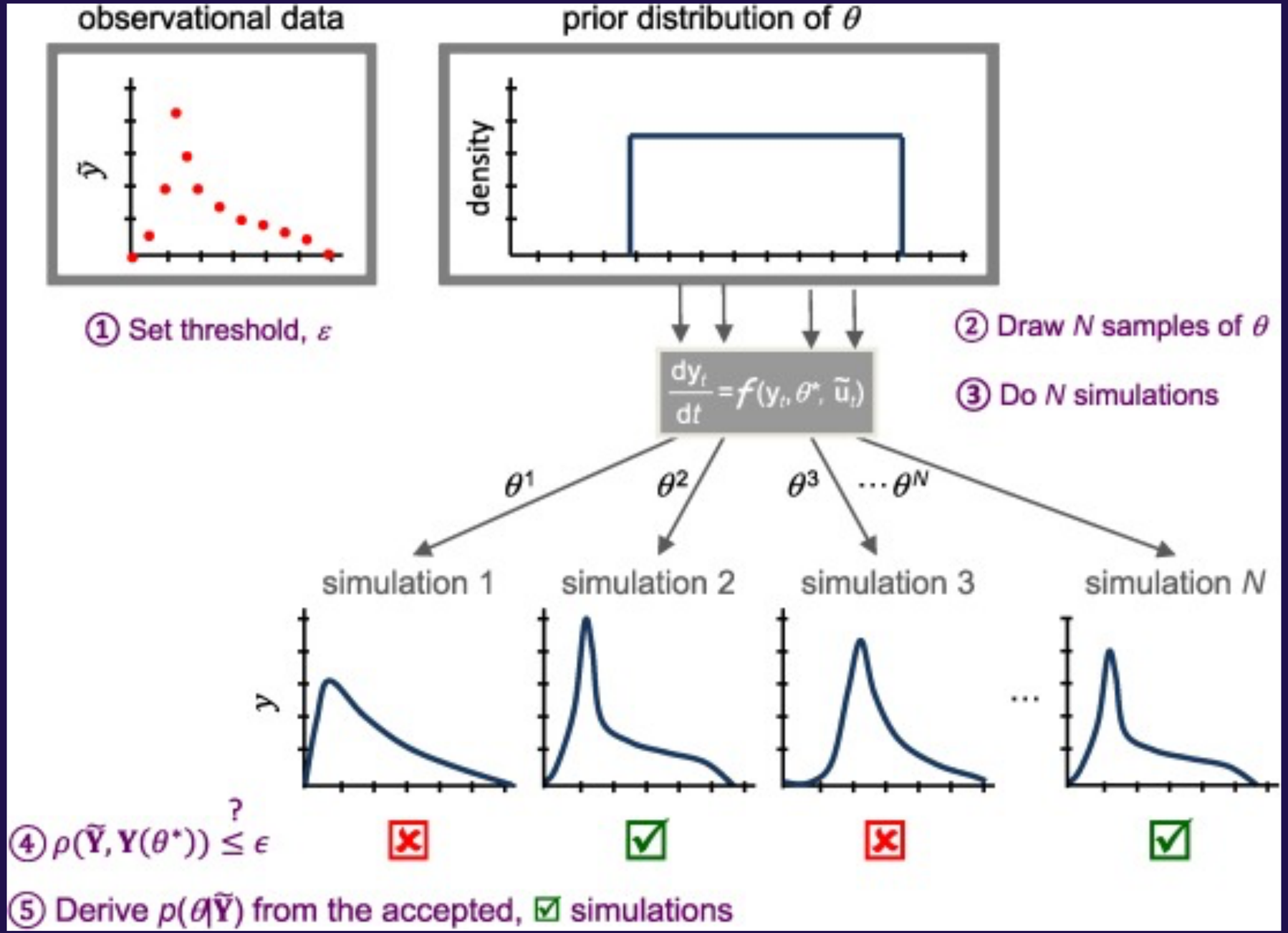
black box
simulator
~~likelihood~~

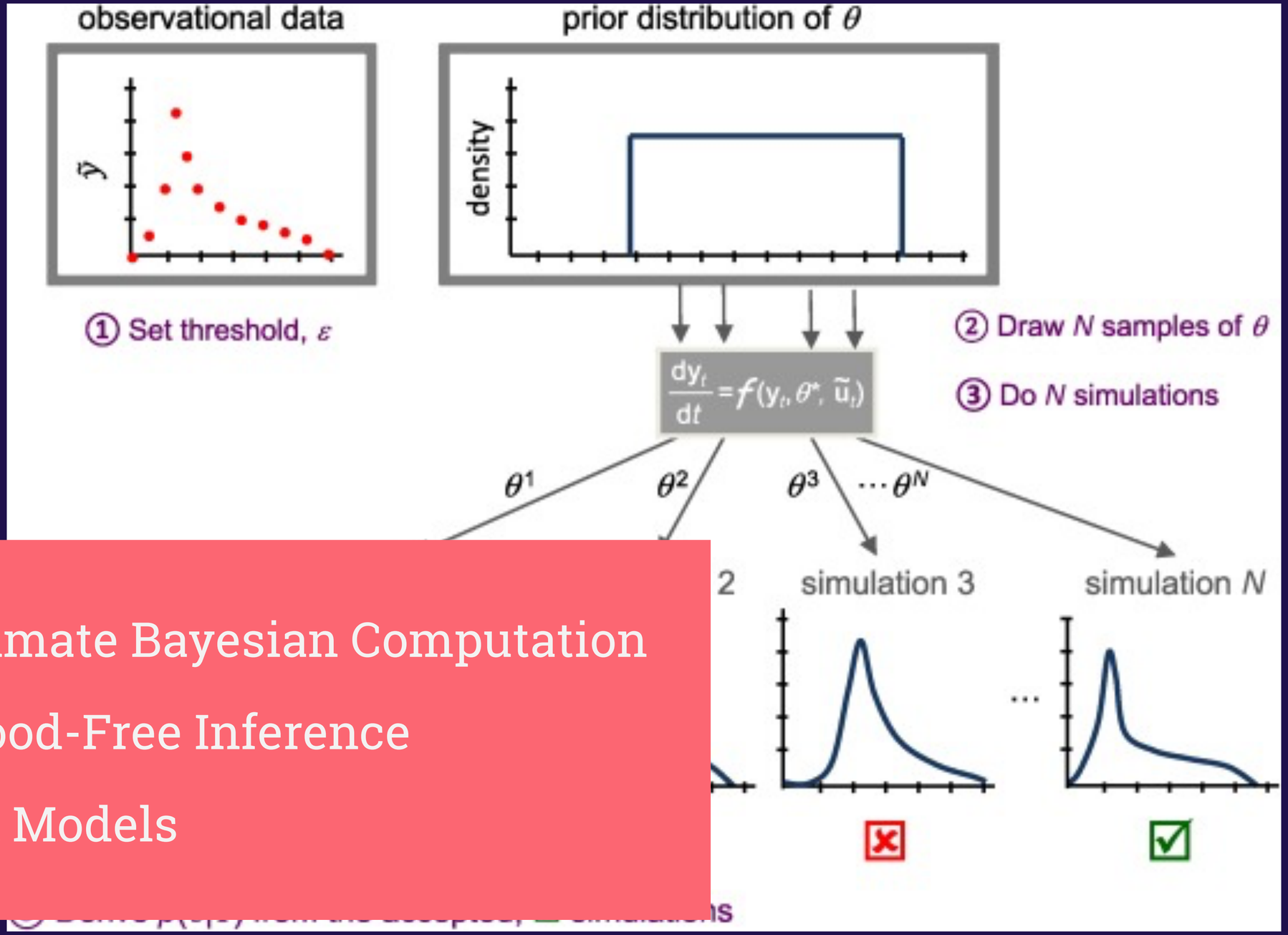
posterior

prior

$$P(\theta \mid x) = \frac{P(x \mid \theta) P(\theta)}{P(x)}$$

marginal
likelihood

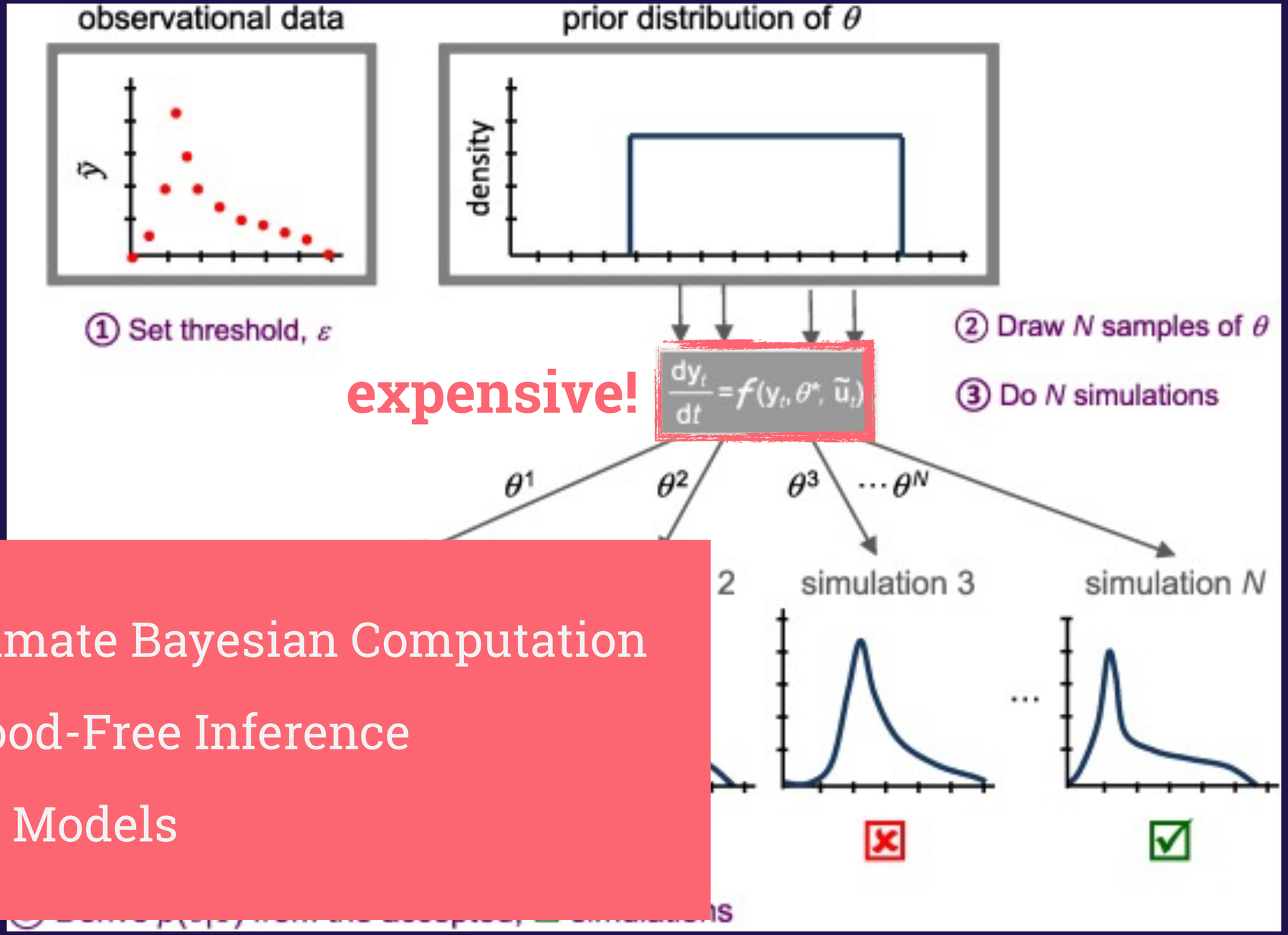

observational data
prior distribution of θ
density
① Set threshold, ε
② Draw N samples of θ
③ Do N simulations
dy_t/dt = f(y_t, θ*, ũ_t)
θ1
θ2
θ3
⋯ θN
simulation 1
simulation 2
simulation 3
simulation N
④ ρ(Ỹ, Y(θ*)) ≤ ε ?
⑤ Derive p(θ|Ỹ) from the accepted, ☑ simulations



Approximate Bayesian Computation

Likelihood-Free Inference

Implicit Models



Approximate Bayesian Computation

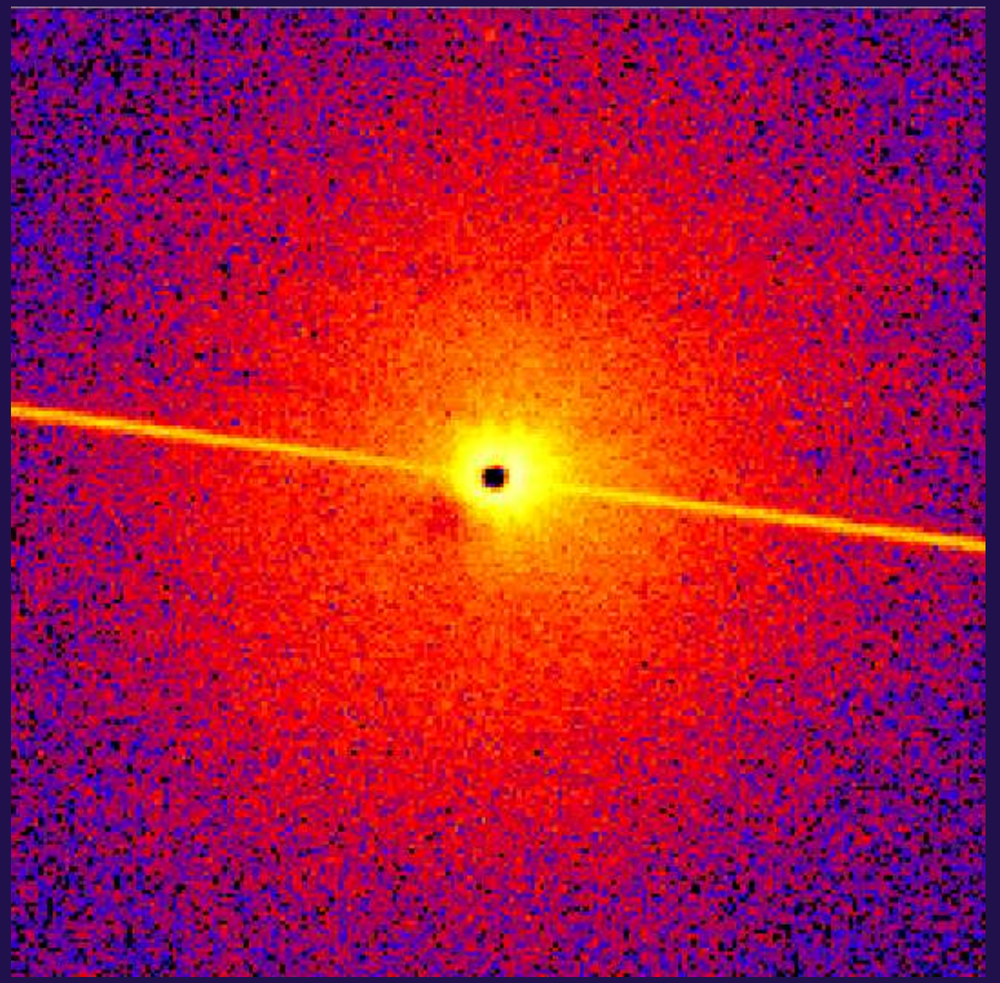
Likelihood-Free Inference

Implicit Models

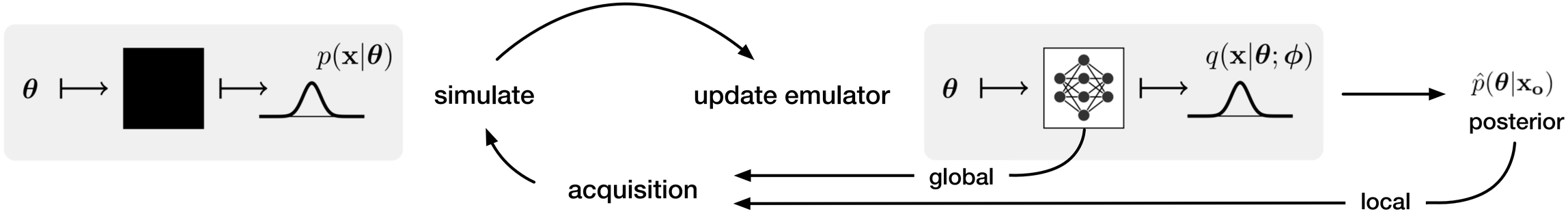
data analysis

data collection

physics



Lueckmann et al (2018)



also see: Brehmer et al (2018): Mining gold from implicit models to improve likelihood-free inference, Brehmer et al (2018): A Guide to Constraining Effective Field Theories with Machine Learning

Probabilistic Programming

pymc3

```
# setup model
model = pm.Model()
with model:
    # cluster sizes
    p = pm.Dirichlet('p', a=np.array([1., 1., 1.]), shape=k)
    # ensure all clusters have some points
    p_min_potential = pm.Potential('p_min_potential', tt.switch


    # cluster centers
    means = pm.Normal('means', mu=[0, 0, 0], sd=15, shape=k)
    # break symmetry
    order_means_potential = pm.Potential('order_means_potential
                                         tt.switch(means[1]-mea
                                         + tt.switch(means[2]-m

    # measurement error
    sd = pm.Uniform('sd', lower=0, upper=20)

    # latent cluster of each observation
    category = pm.Categorical('category',
                              p=p,
                              shape=ndata)

    # likelihood for each observed value
    points = pm.Normal('obs',
                       mu=means[category],
                       sd=sd,
                       observed=data)
```

```
# fit model
with model:
    step1 = pm.Metropolis(vars=[p, sd, means])
    step2 = pm.ElemwiseCategorical(vars=[category], values=[0,
    tr = pm.sample(10000, step=[step1, step2])
```

also:

- STAN
- Edward
- Tensorflow Probability

Exercise (if time): Write the chocolate productivity hierarchical model from the tutorial in a PyMC3 model

Model Interpretation + Critiquing

(see also Blei (2014))

Assess the fitness of the model

- How good is your model in general?
- Where does it fail in particular?

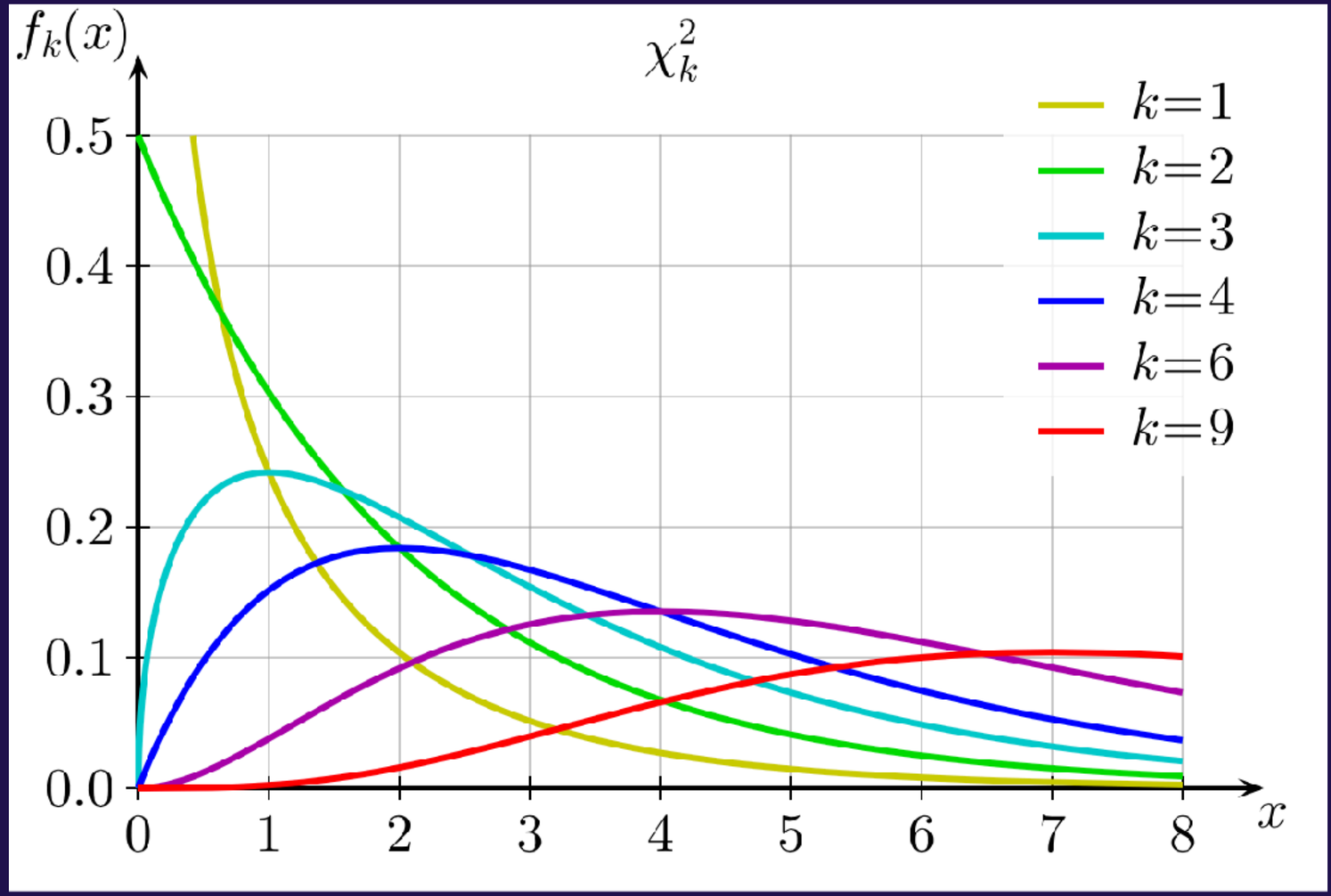
Be pragmatic about your models + inference

Goodness of Fit

$$\mathscr{L}(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m(x_i, \theta))^2$$

distributed as χ^2 with *n-k* degrees of freedom

Goodness of Fit



key concept: confront the model's posterior predictive distribution with the observed data

Posterior Predictive Distribution

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

Posterior Predictive Distribution

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

$$p(y' \mid \mathbf{y}, I) = \int_{\Theta} p(y' \mid \theta, \mathbf{y}, I) p(\theta \mid \mathbf{y}, I) d\theta$$

Posterior Predictive Distribution

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

$$p(y' \mid \mathbf{y}, I) = \int_{\Theta} p(y' \mid \theta, \mathbf{y}, I) p(\theta \mid \mathbf{y}, I) d\theta$$

Predictive Sample Re-use (cross-validation)

- Evaluate model assumption
- can also compare different sampling algorithms!

Posterior Predictive Checks

Is my model good enough in ways that matter?

Posterior Predictive Checks

Posterior Predictive Checks

1. Draw data points from the posterior predictive distribution:

Posterior Predictive Checks

1. Draw data points from the posterior predictive distribution:
    a. Draw parameters θ from $p(\theta \mid y)$

Posterior Predictive Checks

1. Draw data points from the posterior predictive distribution:
    a. Draw parameters θ from $p(\theta \mid y)$
    b. Draw replicated data y^{rep} from the sampled θ

Posterior Predictive Checks

1. Draw data points from the posterior predictive distribution:
    a. Draw parameters θ from $p(\theta \mid y)$
    b. Draw replicated data y^{rep} from the sampled θ
2. compute a summary statistic T(y) that encapsulates some structure in the data:

Posterior Predictive Checks

1. Draw data points from the posterior predictive distribution:
   a. Draw parameters θ from p(θ | *y*)
   b. Draw replicated data y^{rep} from the sampled θ
2. compute a summary statistic T(y) that encapsulates some structure in the data: $\mathrm{PPC} = P(T(y^{\mathrm{rep}} > T(y) \mid y))$

Posterior Predictive Checks

1. Draw data points from the posterior predictive distribution:
   a. Draw parameters θ from p(θ | *y*)
   b. Draw replicated data y^{rep} from the sampled θ
2. compute a summary statistic T(y) that encapsulates some structure in the data: $\mathrm{PPC} = P(T(y^{\mathrm{rep}} > T(y) \mid y))$

T(y) is adaptable to the needs of the practicioner!

Summary

Summary

- Be principled, but pragmatic about your models

Summary

- Be principled, but pragmatic about your models
- Start simple!

Summary

- Be principled, but pragmatic about your models
- Start simple!
- Explore combinations of statistical inference and machine learning (especially when you don’t care about the parameters)

Summary

- Be principled, but pragmatic about your models
- Start simple!
- Explore combinations of statistical inference and machine learning (especially when you don't care about the parameters)
- Critique your models!