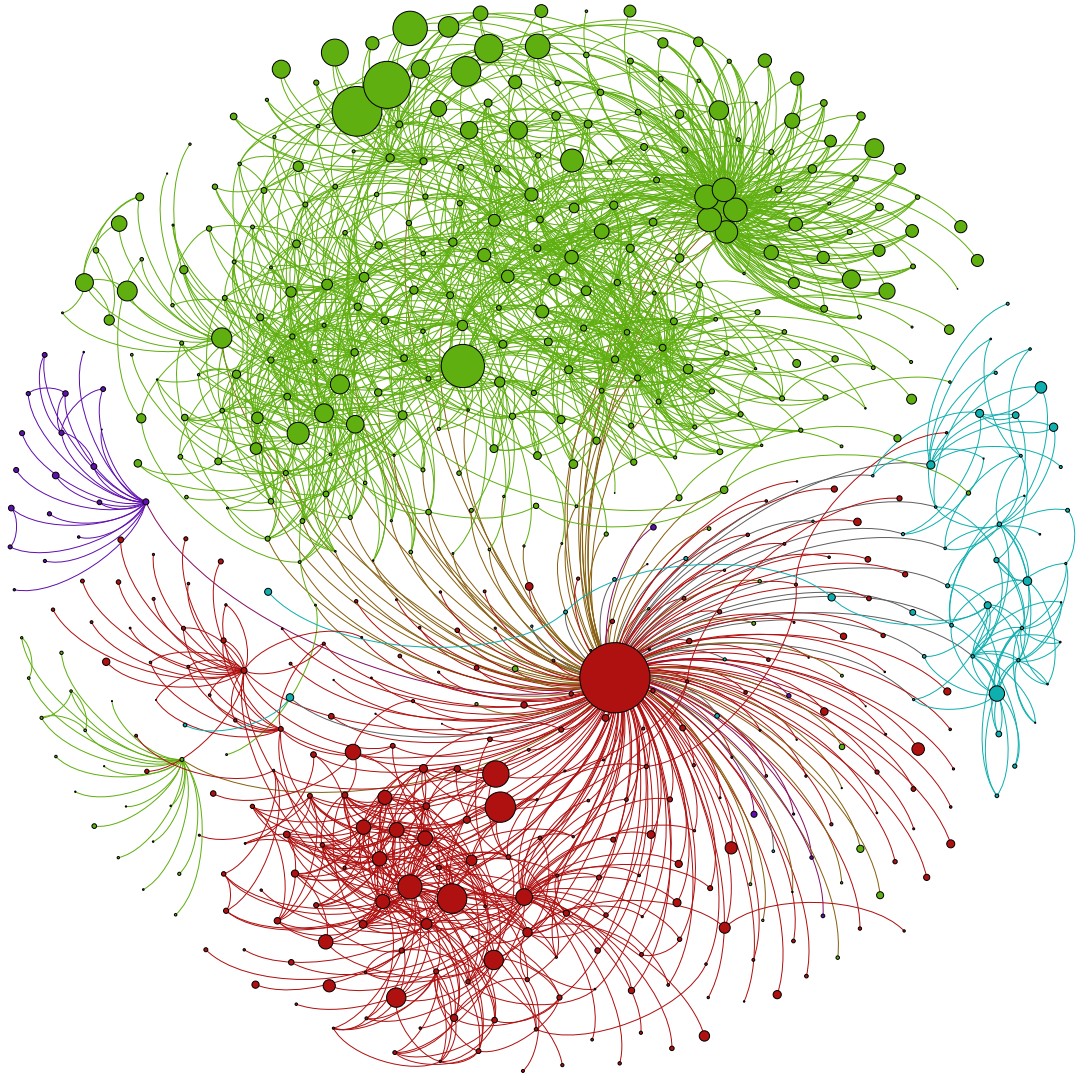
SCALING AND GENERALIZING APPROXIMATE BAYESIAN INFERENCE

David M. Blei
Departments of Computer Science and Statistics
Columbia University

We have ***complicated data***; we want to ***make sense*** of it.

PROBABILISTIC MACHINE LEARNING/BAYESIAN STATISTICS

- Statistical methods that ***connect domain knowledge to data***.
- Goal: A methodology that is ***expressive, scalable, easy to develop***

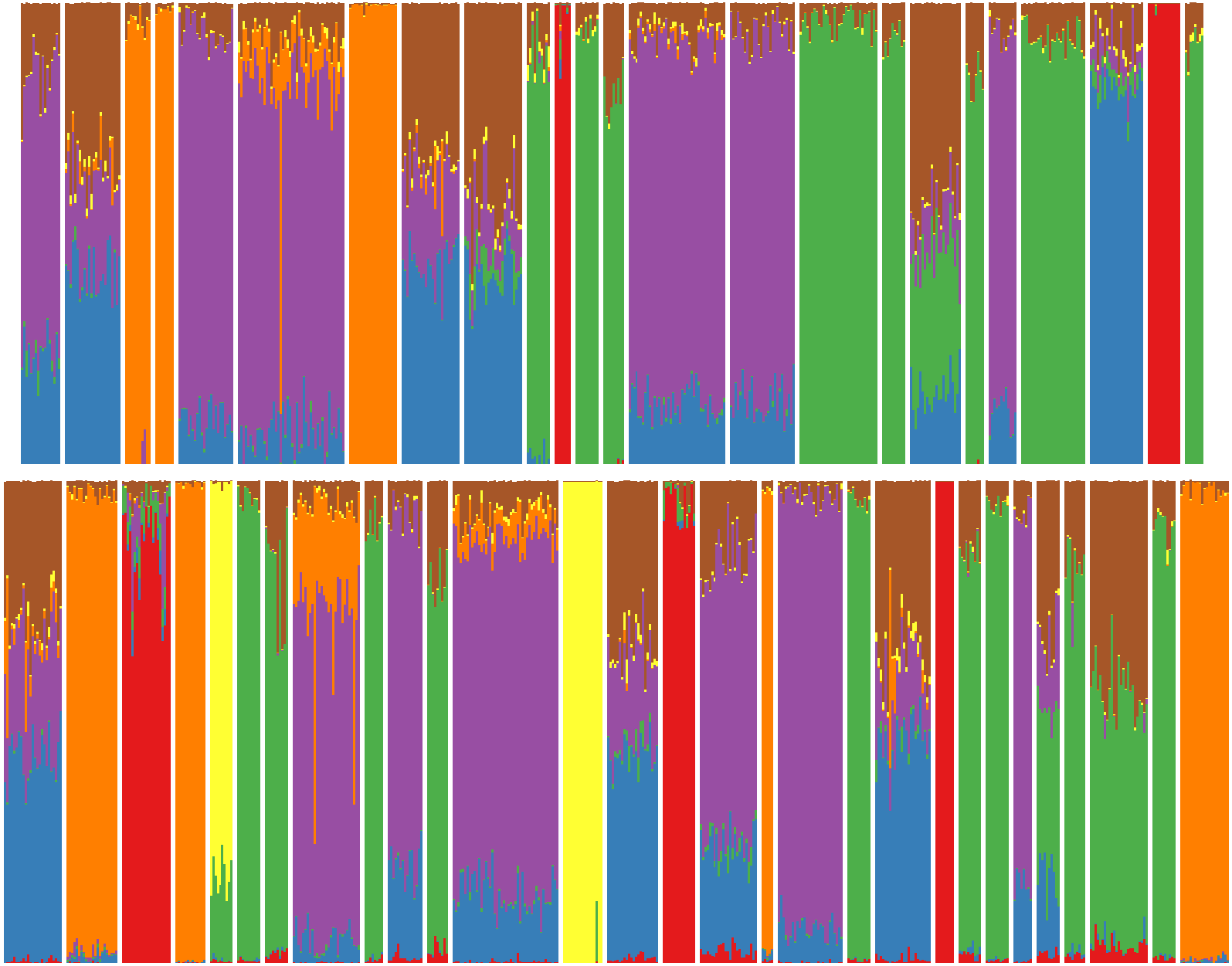


Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei 2013]

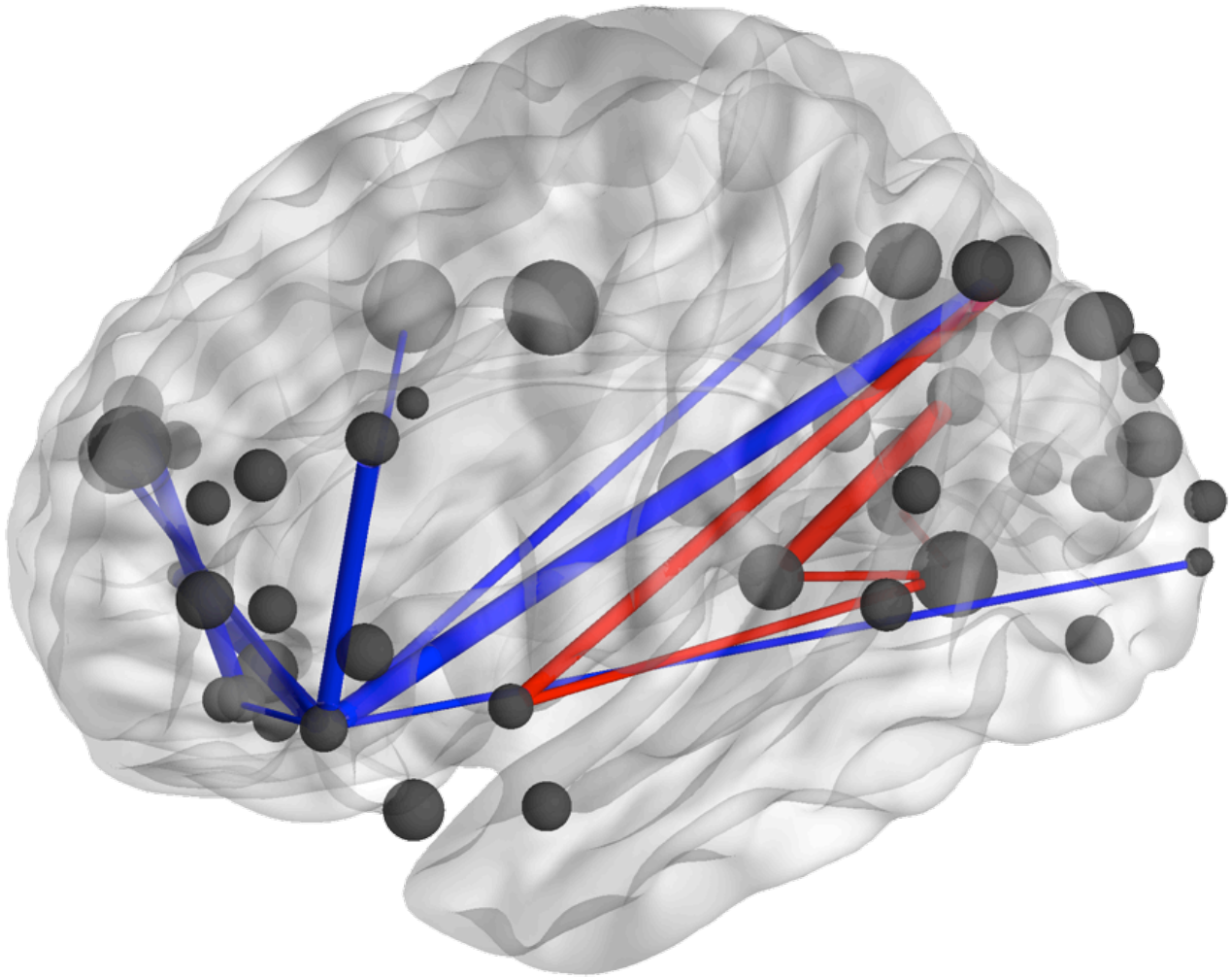
Topics found in 1.8M articles from the New York Times

[Hoffman+ 2013]



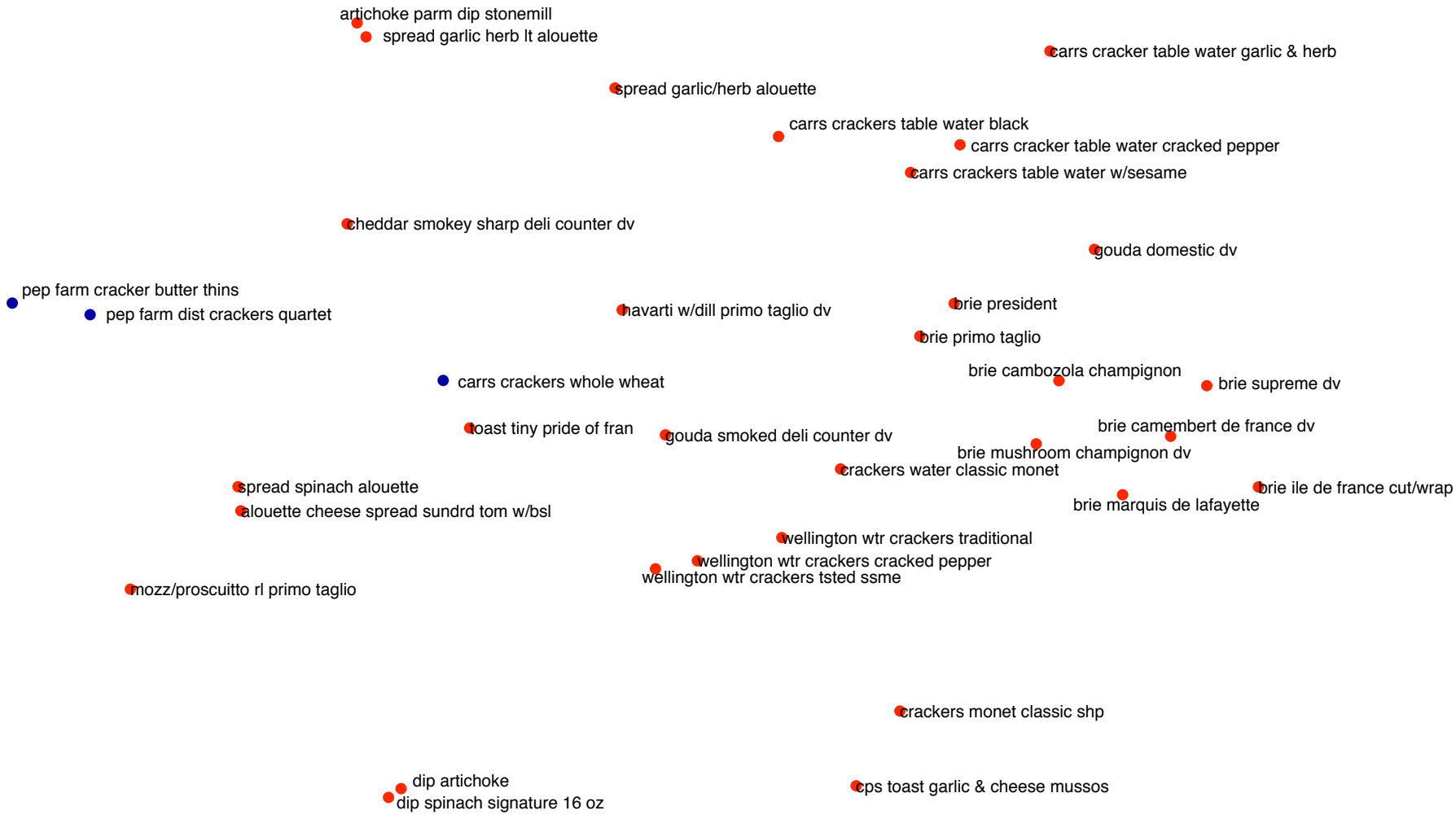
Population analysis of 2 billion genetic measurements

[Gopalan+ 2016]



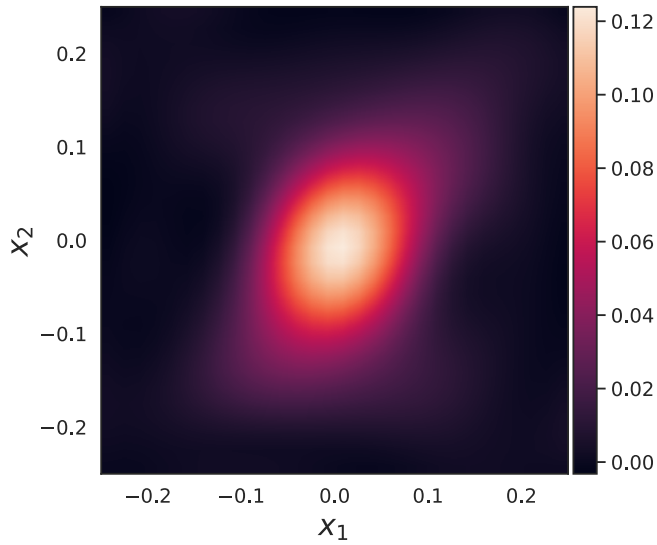
Neuroscience analysis of 220 million fMRI measurements

[Manning+ 2014]

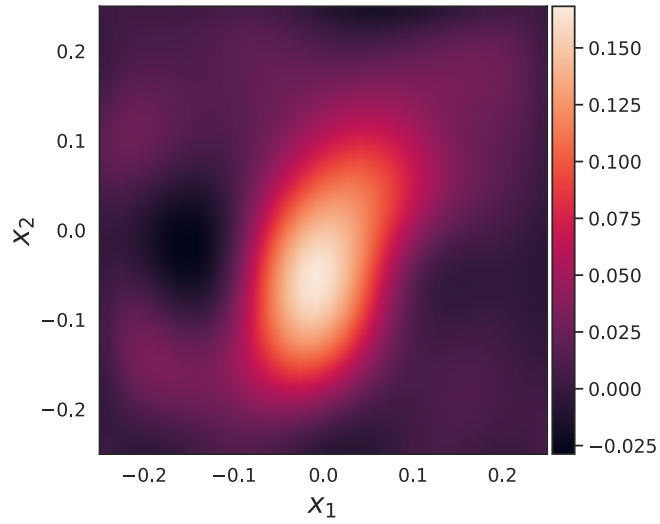


(Fancy) discrete choice analysis of 5.7M purchases

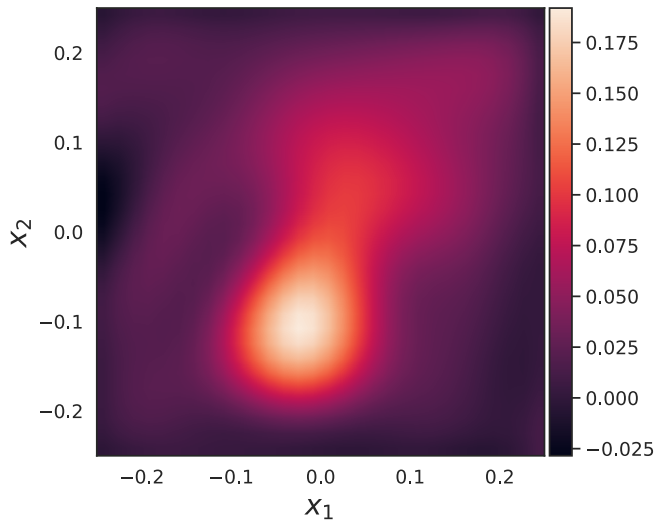
[Ruiz+ 2020]



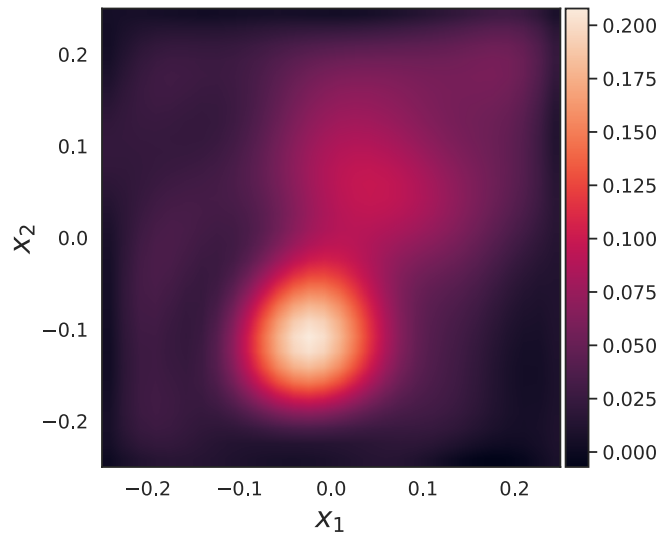
(a) $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 1{,}000$



(b) $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 10{,}000$



(c) $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 100{,}000$

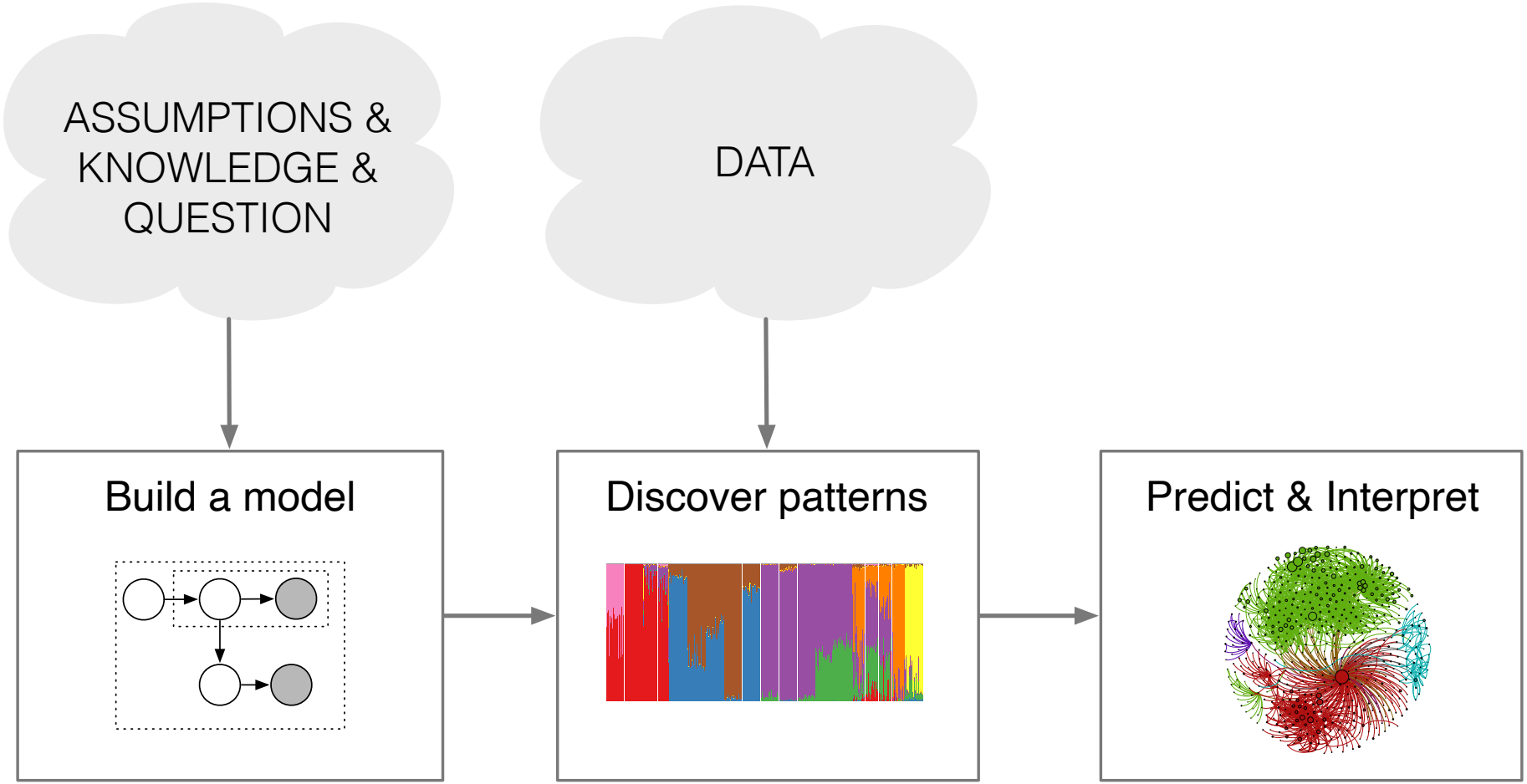


(d) $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 1{,}000{,}000$

Inferring the dust map from astronomical data

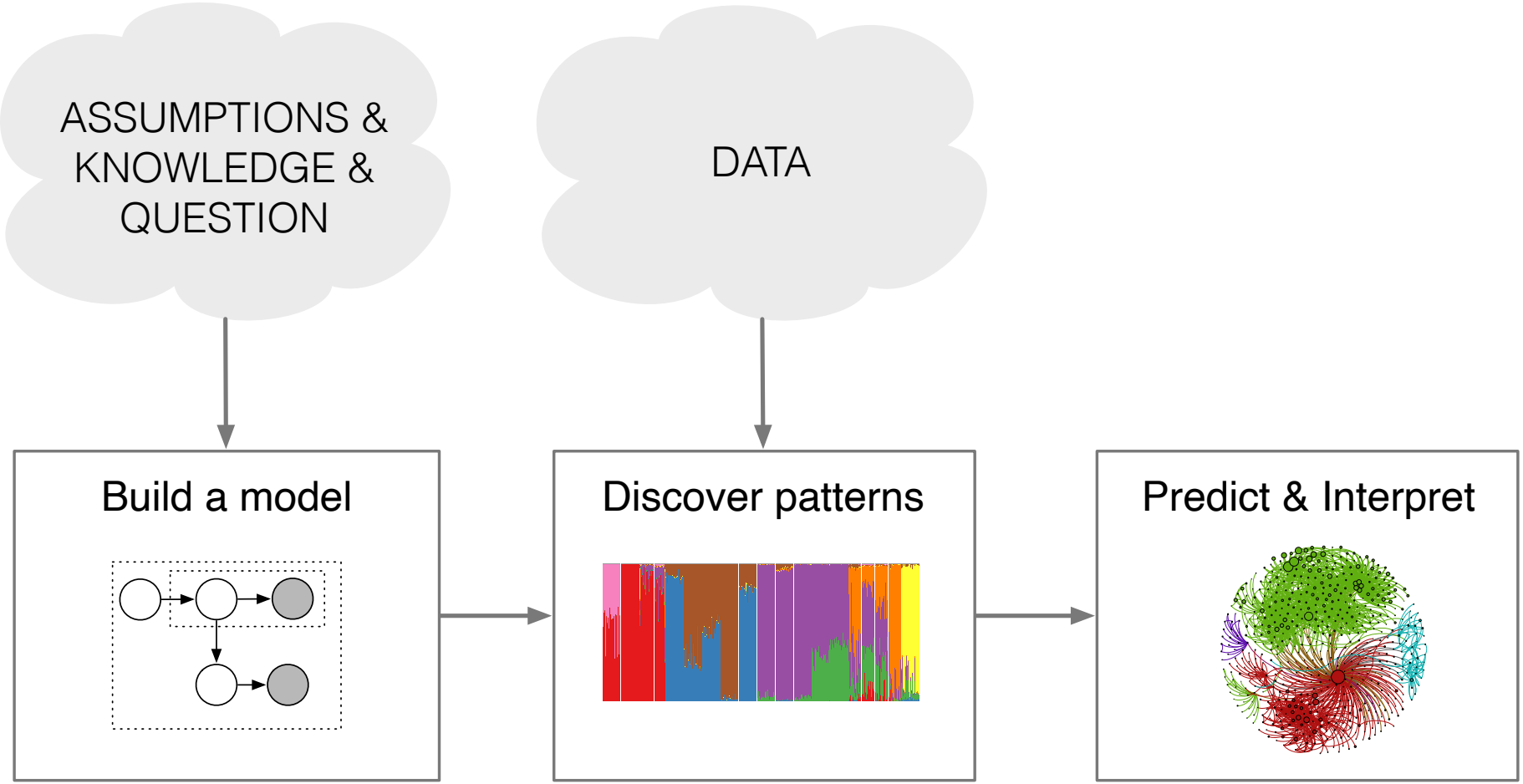
[L. Anderson+ 2022]

The probabilistic pipeline

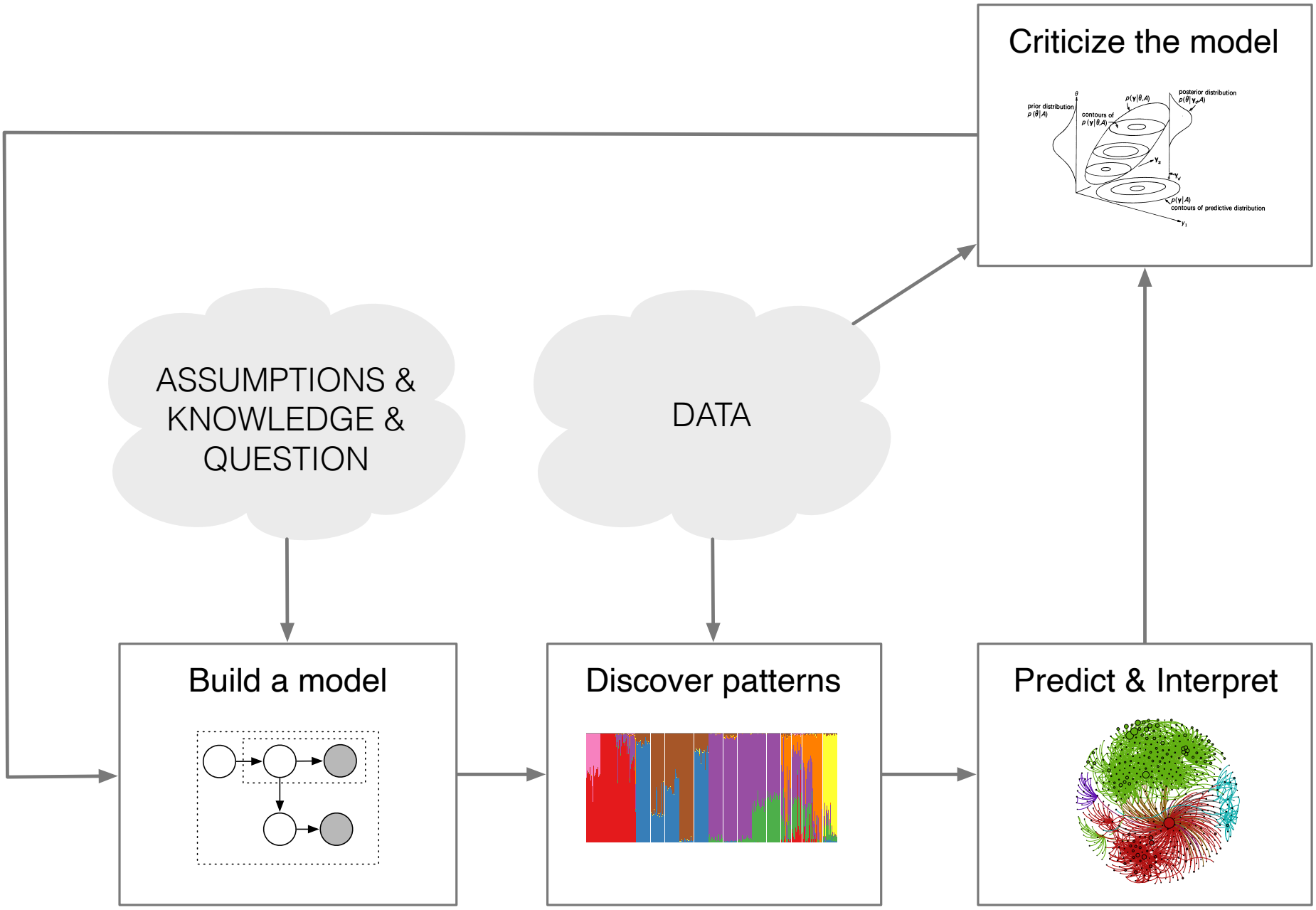


- Customized data analysis is important to many fields.
- Probabilistic ML separates **assumptions**, **computation**, **application**
- Eases collaborative solutions to ML/statistics problems

The probabilistic pipeline



- **Posterior inference** is the key algorithmic problem.
- Answers the question: What does this model say about this data?
- Today: **Scalable** and **general** approaches to posterior inference



[Box, 1980; Rubin, 1984; Gelman+ 1996; Blei, 2014]

Probabilistic machine learning / Bayesian statistics

- **Probabilistic model**: joint distribution of hidden variables $\mathbf{z}$ and observations $\mathbf{x}$,

$$p(\mathbf{z}, \mathbf{x})$$

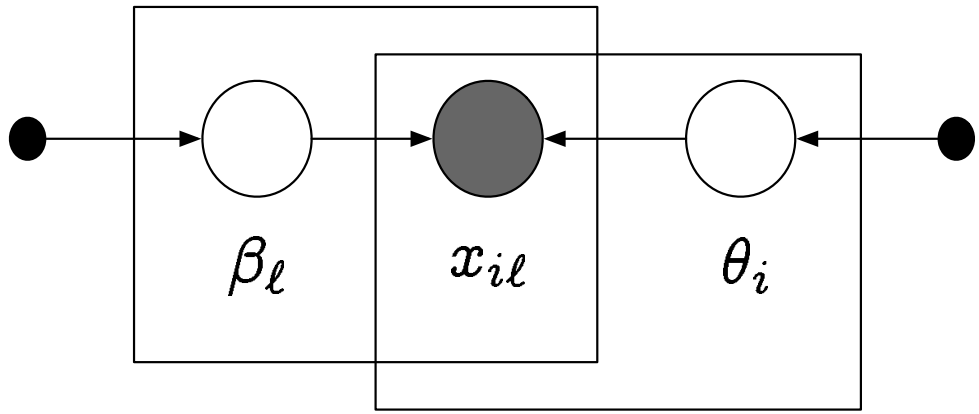
- Inference about the unknowns is through the **posterior**, the conditional distribution of the hidden variables given the observations

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

  (Note: There is no need to “be Bayesian” to calculate a posterior.)

- For most interesting models, the posterior is not tractable.
  We appeal to **approximate posterior inference**.

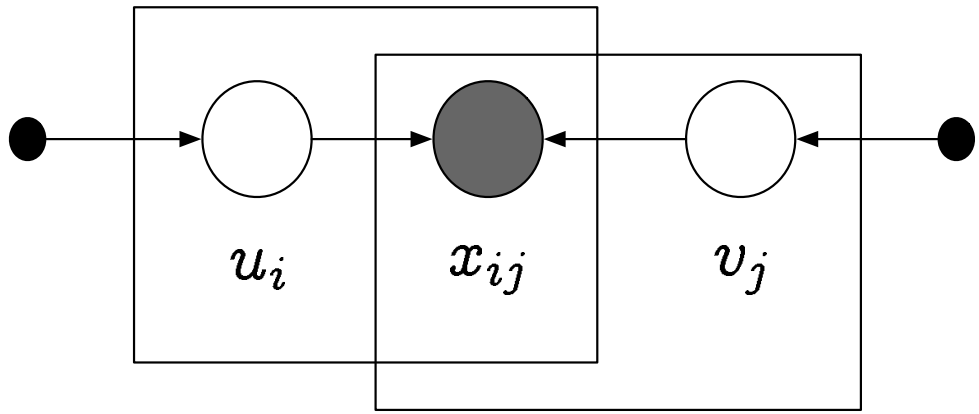
Structure [Pritchard+ 2000]



$$\begin{aligned} \beta_{\ell k} &\sim \text{beta}(a, b) & k &= 1 \ldots \text{K} \\ \theta_i &\sim \text{dirichlet}_\text{K}(\alpha) & i &= 1 \ldots m \\ x_{i\ell} &\sim \text{binomial}\left(2, \textstyle\sum_k \theta_{ik}\beta_{\ell k}\right) & \ell &= 1 \ldots \text{L} \end{aligned}$$

- A popular model for population genetics
- The data are (unphased) alleles at L locations.
- The posterior θ_i uncovers per-individual ancestry used, e.g., in causal adjustment.

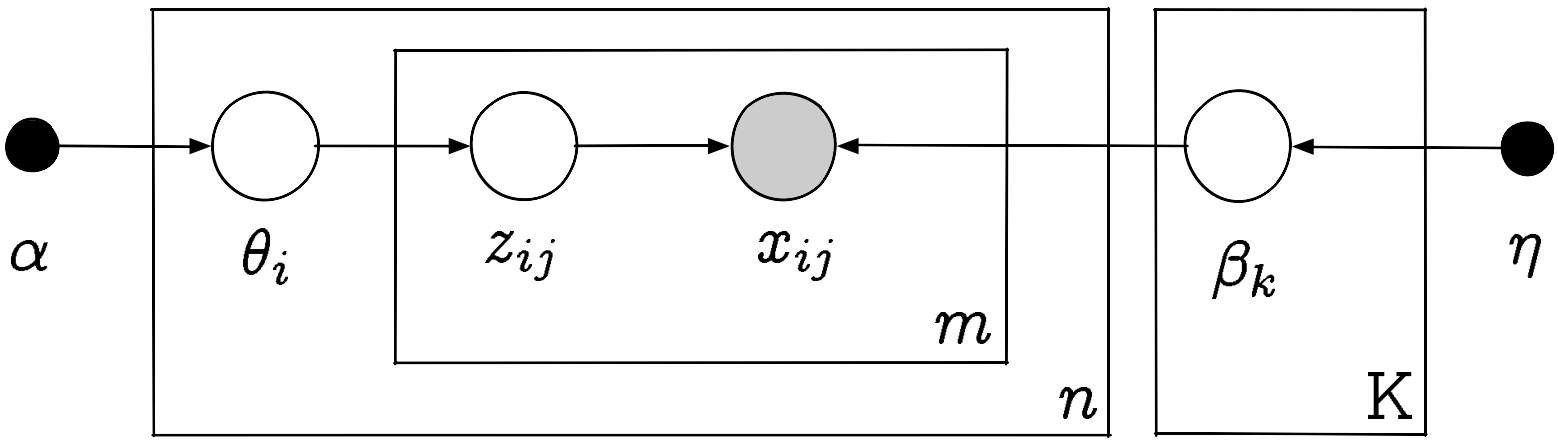
Poisson factorization [Gopalan+ 2015]



$$u_{ik} \sim \text{gamma}(a, b) \qquad k = 1 \ldots \text{K}$$
$$v_{jk} \sim \text{gamma}(c, d) \qquad i = 1 \ldots n$$
$$x_{ij} \sim \text{poisson}\left(\sum_k u_{ik} v_{jk}\right) \qquad j = 1 \ldots m$$

- A good model for recommendation systems
- Rows i are users ; columns j are items ; each x_{ij} is the number of clicks.
- Posterior per-row variables uncover user preferences.
  Posterior per-column variables uncover item attributes (like genre)

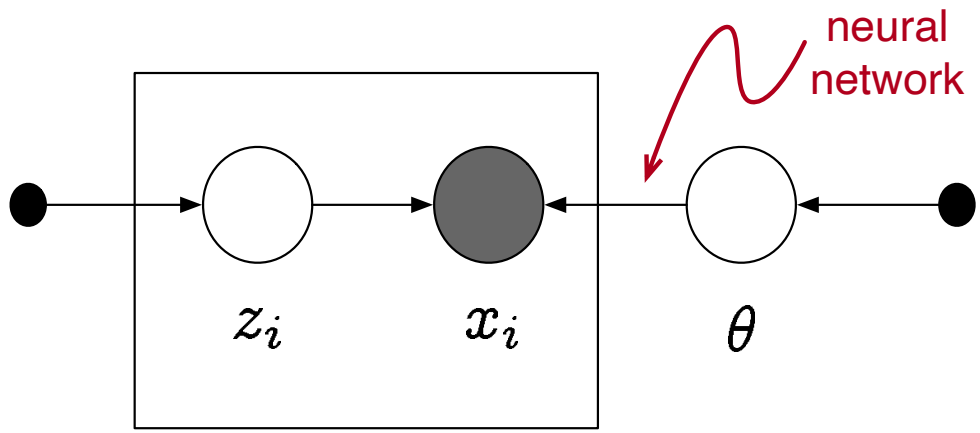
Latent Dirichlet allocation [Blei+ 2003]



$$\beta_k \sim \text{dirichlet}_V(\eta) \qquad k = 1 \ldots \text{K}$$
$$\theta_i \sim \text{dirichlet}_\text{K}(\alpha) \qquad i = 1 \ldots m$$
$$z_{ij} \sim \text{cat}(\theta_i) \qquad j = 1 \ldots n$$
$$x_{ij} \sim \text{cat}(\beta_{z_{ij}})$$

- A mixed-membership model of documents, a.k.a. a topic model.
- Posterior $\beta_{1:\text{K}}$ are *topics*, each a distribution over the vocabulary.
- The topics reflect themes that run through the collection.

Deep generative models [Kingma and Welling 2014, Rezende+ 2014]



$$\theta \sim p(\theta)$$
$$z_i \sim \text{normal}_{\text{K}}(0, 1) \qquad i = 1 \ldots n$$
$$x_i \sim \text{poisson}(\text{nn}(z_i\,;\,\theta)) \qquad \text{nn} : \mathbb{R}^{\text{K}} \to \mathbb{R}_+^{\text{m}}$$

- A neural network eats a latent variable to produce the observed data.
- This is a very flexible class of models of distributions $p(x) = \int p(z)p(x|z)\mathrm{d}z$.
- Inference is on neural network parameters and latent representations.

Probabilistic machine learning / Bayesian statistics

- **Probabilistic model**: joint distribution of hidden variables $\mathbf{z}$ and observations $\mathbf{x}$,

$$p(\mathbf{z}, \mathbf{x})$$

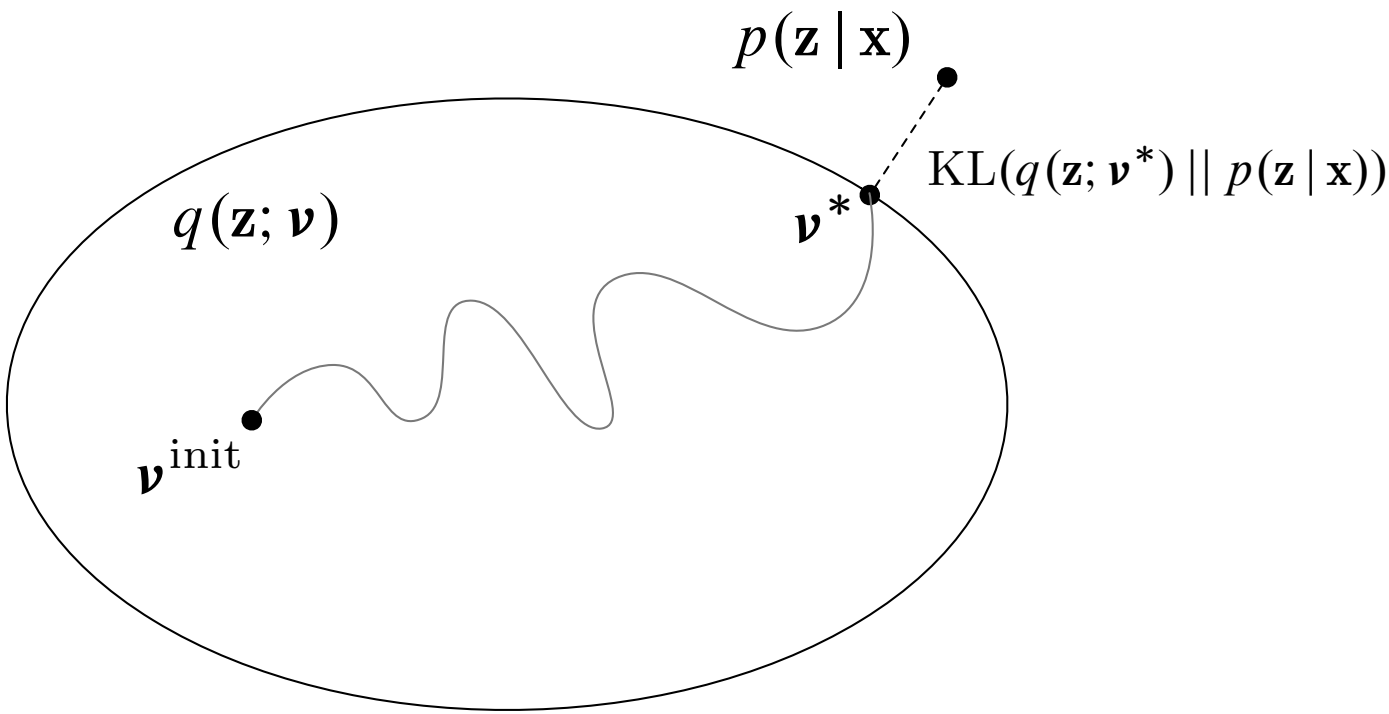
- Inference about the unknowns is through the **posterior**, the conditional distribution of the hidden variables given the observations

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

  (Note: There is no need to “be Bayesian” to calculate a posterior.)

- For most interesting models, the posterior is not tractable.
  We appeal to **approximate posterior inference**.

Variational inference



- VI solves **inference** with **optimization**.

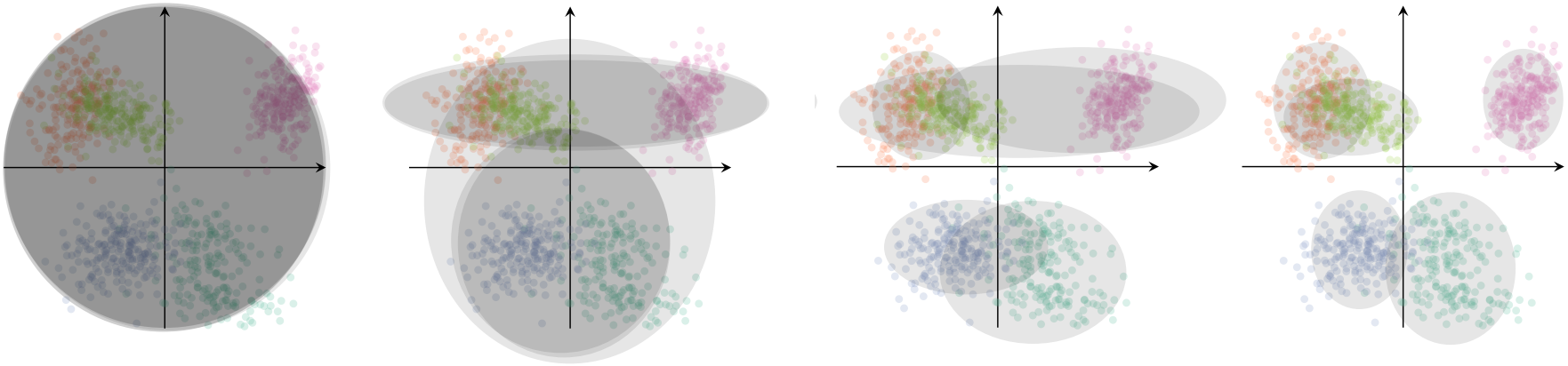
  (Contrast this with MCMC.)

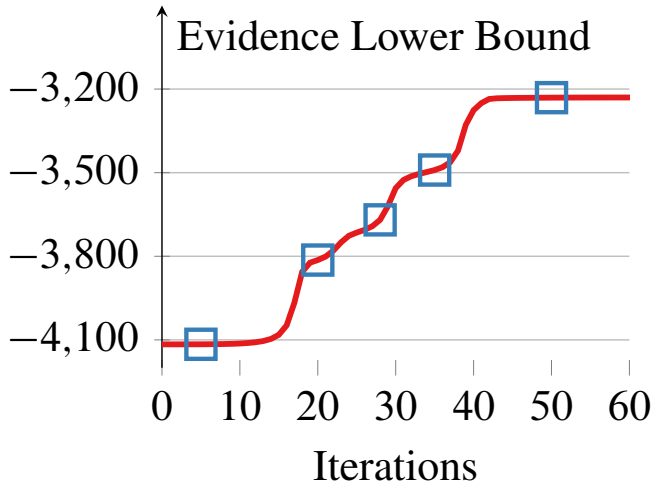
- Posit a **variational family** of distributions over the latent variables,

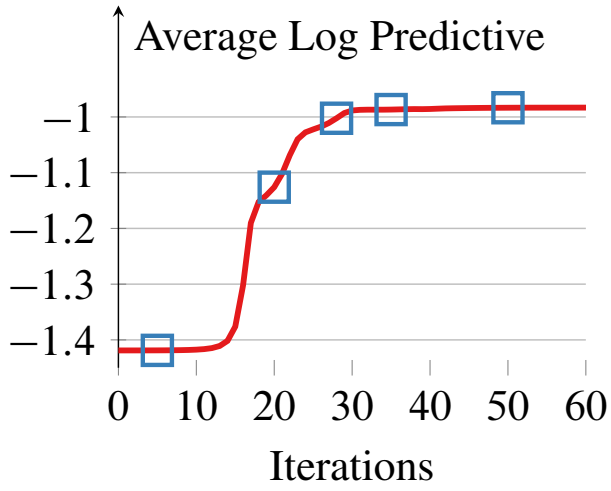
$$q(\mathbf{z}; \nu)$$

- Fit the **variational parameters** ν to be close (in KL) to the exact posterior.

Example: Mixture of Gaussians

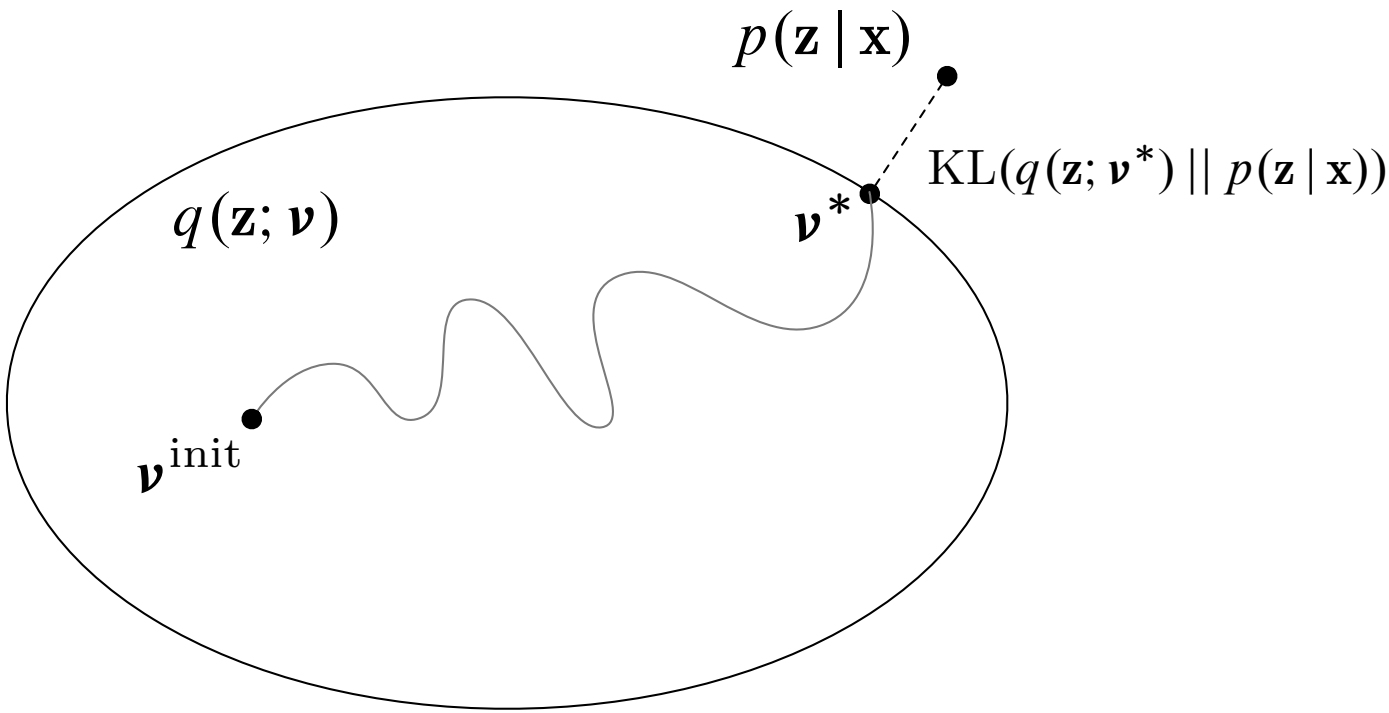






[images by Alp Kucukelbir; Blei+ 2017]

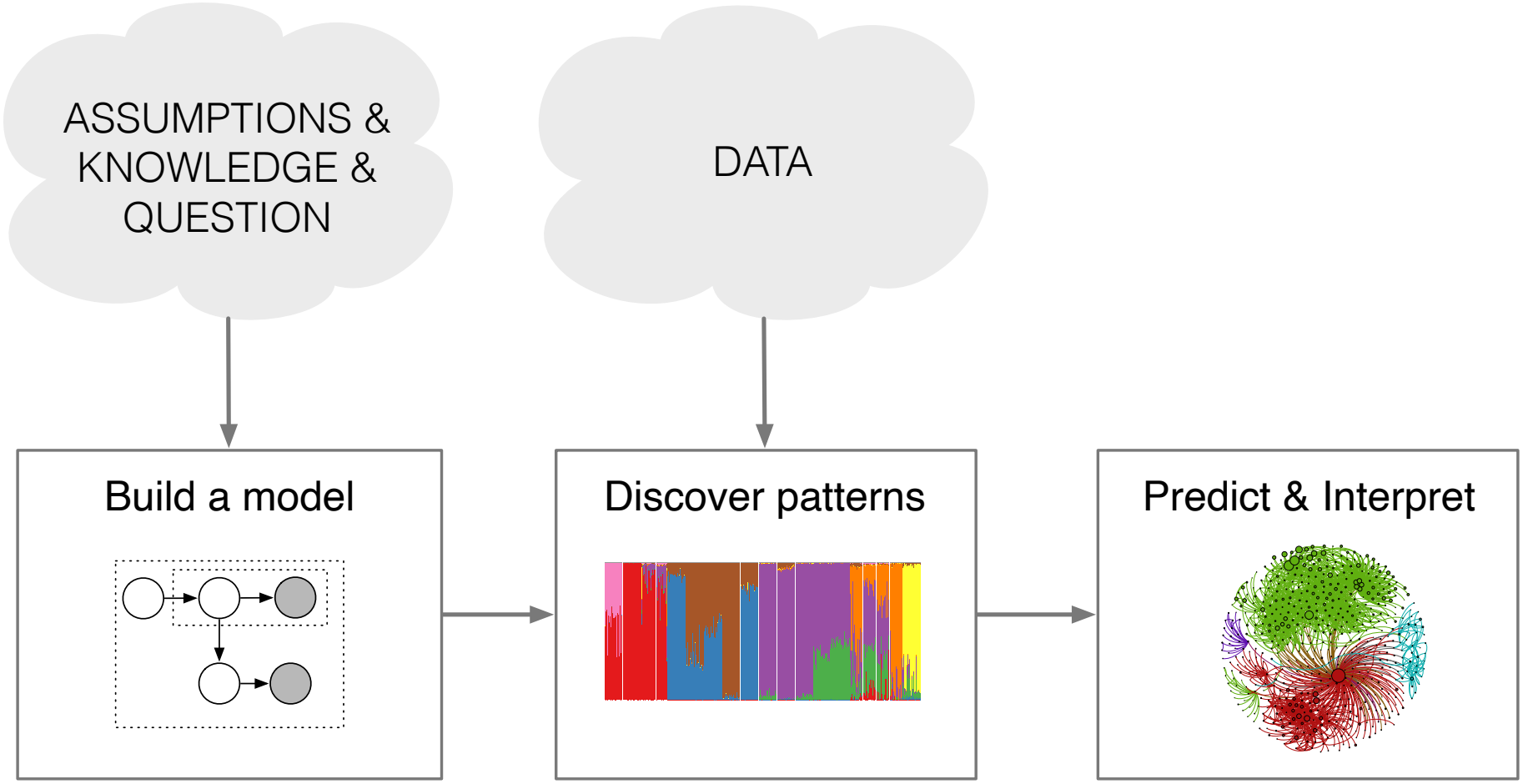
Today: Stochastic optimization makes VI better



- **Stochastic VI** scales up VI to massive data. [Hoffman+ 2013]
- **Black box VI** generalizes VI to a wide class of models. [Ranganath+ 2014]

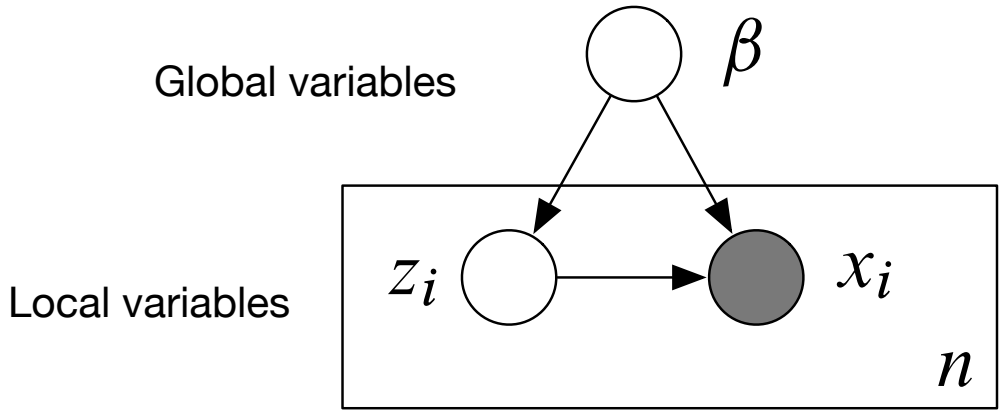
Stochastic Variational Inference

The probabilistic pipeline



How can we scale up variational inference to massive datasets?

Conditionally conjugate models

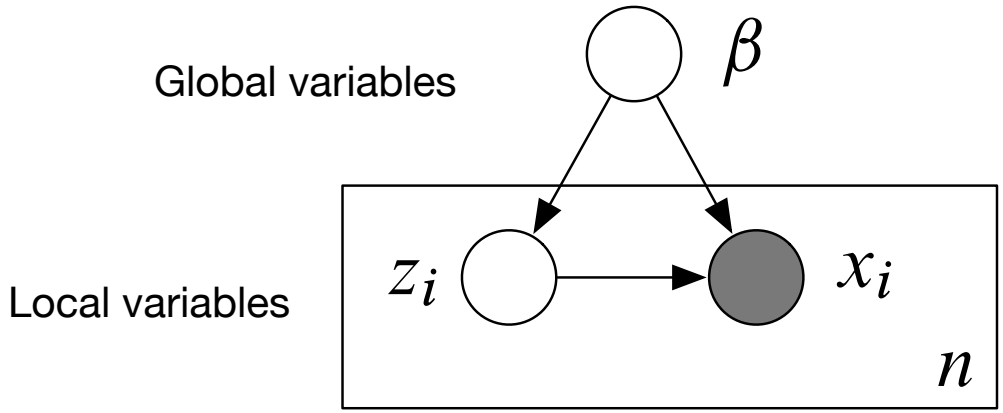


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- The observations are $\mathbf{x} = x_{1:n}$.
- The **local** variables are $\mathbf{z} = z_{1:n}$.
- The **global** variables are β.
- The ith data point x_i only depends on z_i and β.

Compute $p(\beta, \mathbf{z} \mid \mathbf{x})$.

Conditionally conjugate models



$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

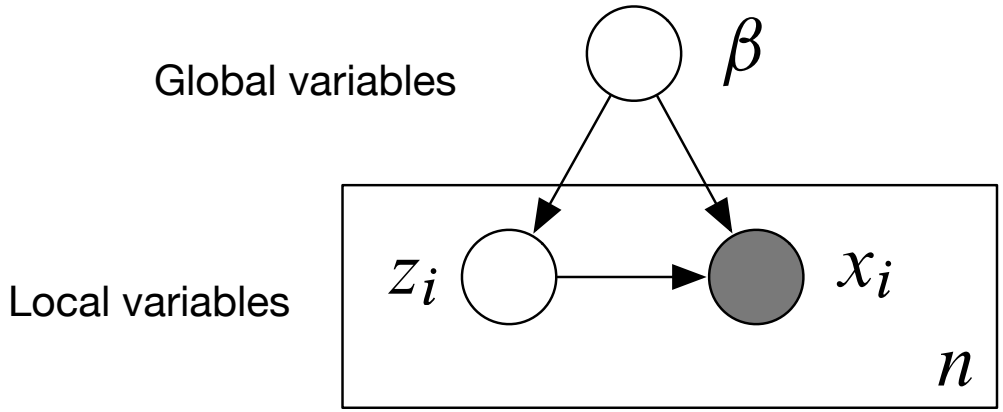
- **Complete conditional:**
  The distribution of a latent variable given the observations and other latent variables.

- Assume each complete conditional is in an exponential family [Brown 1986; Efron 2022] ,

$$p(z_i \mid \beta, x_i) = \text{expfam}(z_i \,;\, \eta_\ell(\beta, x_i))$$
$$p(\beta \mid \mathbf{z}, \mathbf{x}) = \text{expfam}(\beta \,;\, \eta_g(\mathbf{z}, \mathbf{x})),$$

where $\text{expfam}(z \,;\, \eta) = h(z) \exp\{\eta^\top t(z) - a(\eta)\}$.

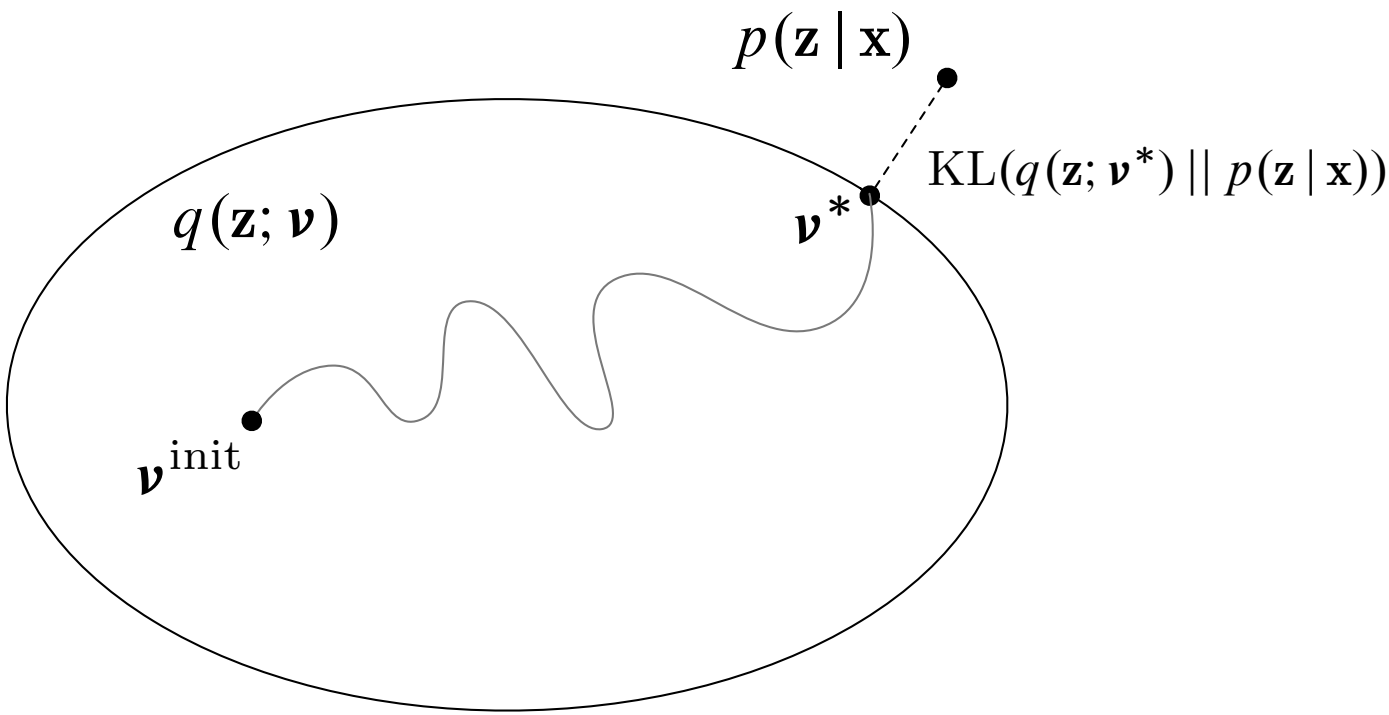
Conditionally conjugate models



$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- Bayesian mixture models
- Time series models
  (HMMs, linear dynamic systems)
- Factorial models
- Matrix factorization
  (factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression
  (linear, probit, Poisson)
- Stochastic block models
- Mixed-membership models
  (LDA and some variants)

Variational inference



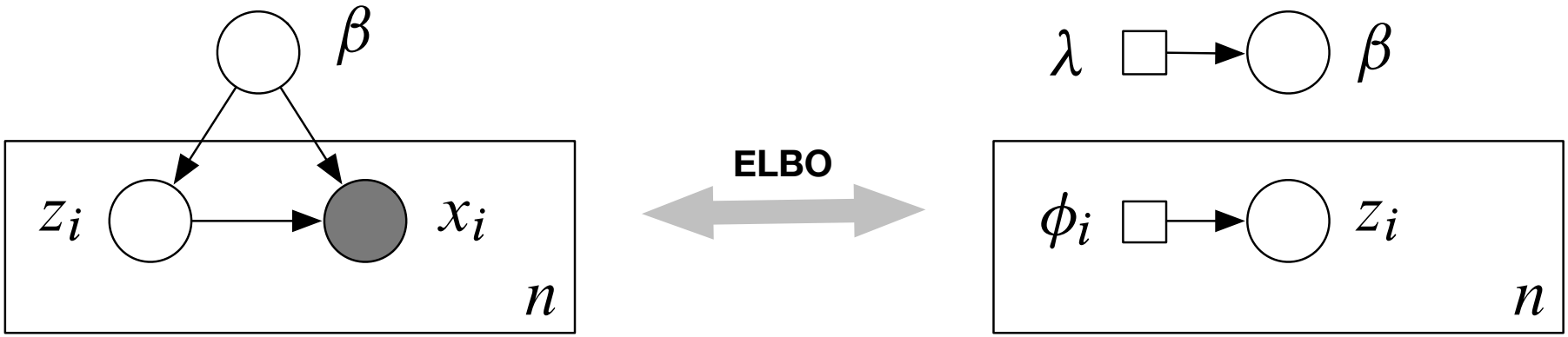
Minimize KL between $q(\beta, \mathbf{z}; \nu)$ and the posterior $p(\beta, \mathbf{z} \mid \mathbf{x})$.

The evidence lower bound

$$\mathcal{L}(\nu) = \underbrace{\mathbb{E}_q\left[\log p(\beta, \mathbf{z}, \mathbf{x})\right]}_{\text{Expected complete log likelihood}} - \underbrace{\mathbb{E}_q\left[\log q(\beta, \mathbf{z}; \nu)\right]}_{\text{Negative entropy}}$$

- KL is intractable; VI optimizes the **evidence lower bound** (ELBO) instead.
  - It is a lower bound on $\log p(\mathbf{x})$.
  - Maximizing the ELBO is equivalent to minimizing the KL.
- The ELBO trades off two terms.
  - The first term prefers $q(\cdot)$ to place its mass on the MAP estimate.
  - The second term encourages $q(\cdot)$ to be diffuse.
- Caveat: The ELBO is not convex.

Mean-field variational inference



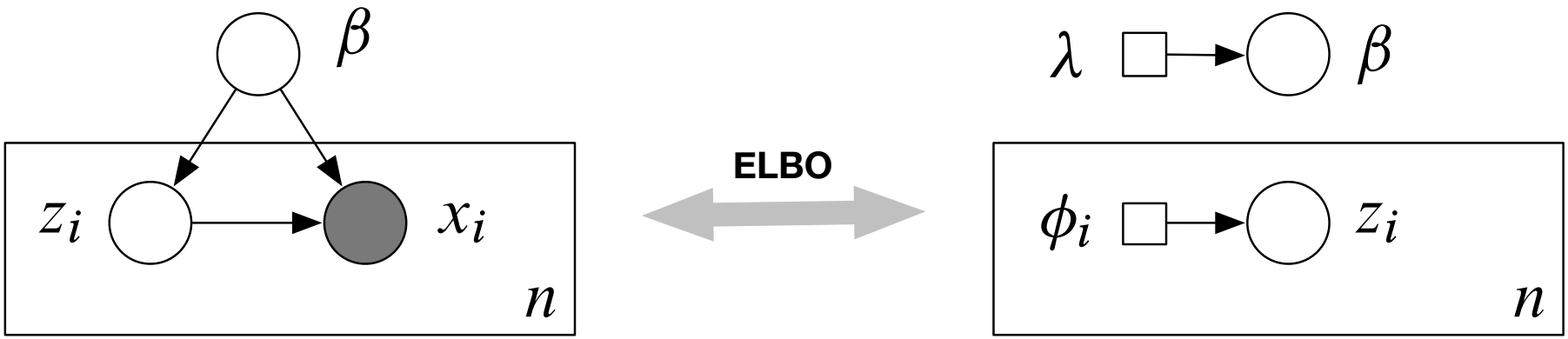
- The form of $q(\beta, \mathbf{z})$ defines the **variational family**.
- The **mean-field family** is fully factorized,

$$q(\beta, \mathbf{z}\,;\,\lambda, \phi) = q(\beta\,;\,\lambda)\prod_{i=1}^{n} q(z_i\,;\,\phi_i).$$

- Each factor is the same family as the model's complete conditional.

$$p(\beta \mid \mathbf{z}, \mathbf{x}) = \text{expfam}(\beta\,;\,\eta_g(\mathbf{z}, \mathbf{x}))$$
$$q(\beta\,;\,\lambda) = \text{expfam}(\beta\,;\,\lambda)$$

Mean-field variational inference



- Optimize the ELBO,

$$\mathcal{L}(\lambda, \phi) = \mathbb{E}_q \left[\log p(\beta, \mathbf{z}, \mathbf{x})\right] - \mathbb{E}_q \left[\log q(\beta, \mathbf{z})\right].$$

- Traditional VI uses coordinate ascent

$$\lambda^* = \mathbb{E}_\phi \left[\eta_g(\mathbf{z}, \mathbf{x})\right]; \; \phi_i^* = \mathbb{E}_\lambda \left[\eta_\ell(\beta, x_i)\right]$$

  It iteratively updates each parameter [Ghahramani and Beal, 2001] .

- Notice the relationship to Gibbs sampling [Gelfand and Smith, 1990] .

Coordinate ascent variational inference

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.

Initialize λ randomly.

while *not converged* **do**

  for *each data point* i **do**

    Set local parameter

$$\phi_i \leftarrow \mathbb{E}_\lambda \left[\eta_\ell(\beta, x_i) \right].$$

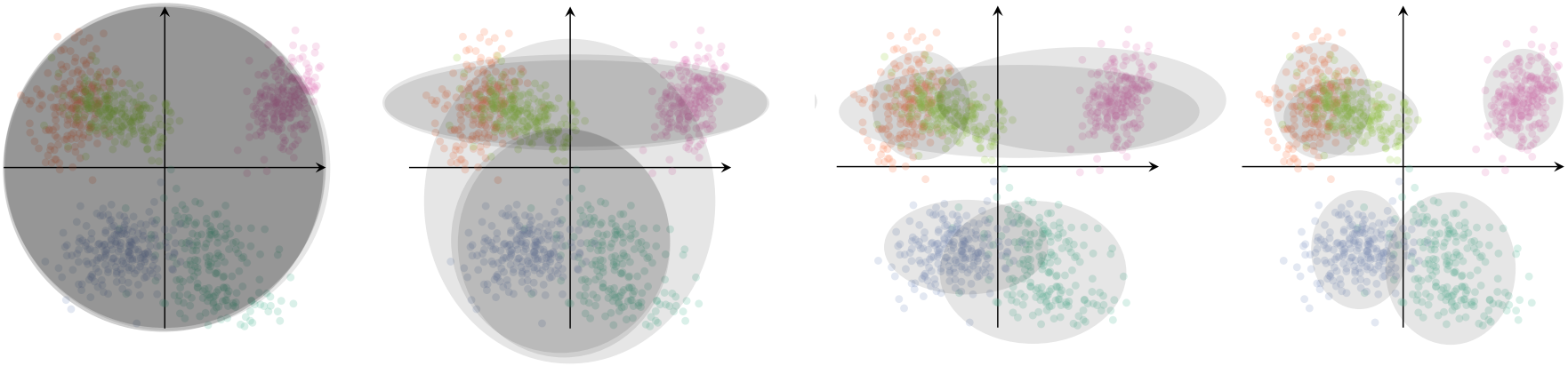
  end

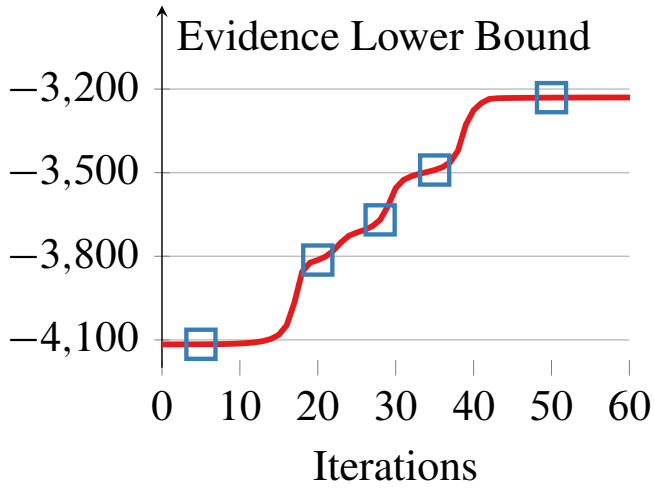
  Set global parameter

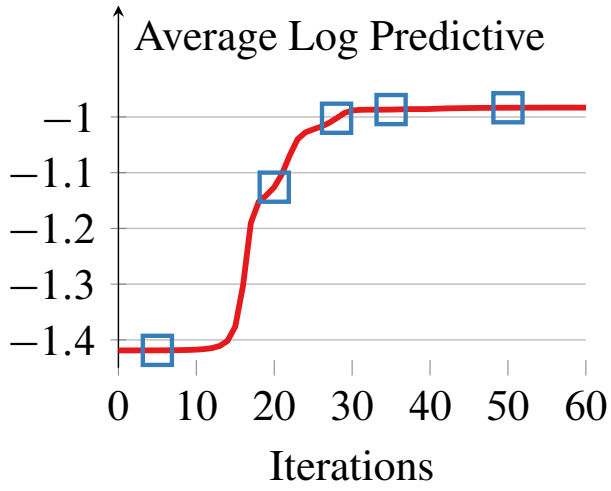
$$\lambda \leftarrow \alpha + \sum_{i=1}^{n} \mathbb{E}_{\phi_i} \left[t(Z_i, x_i) \right].$$

end

Example: Mixture of Gaussians

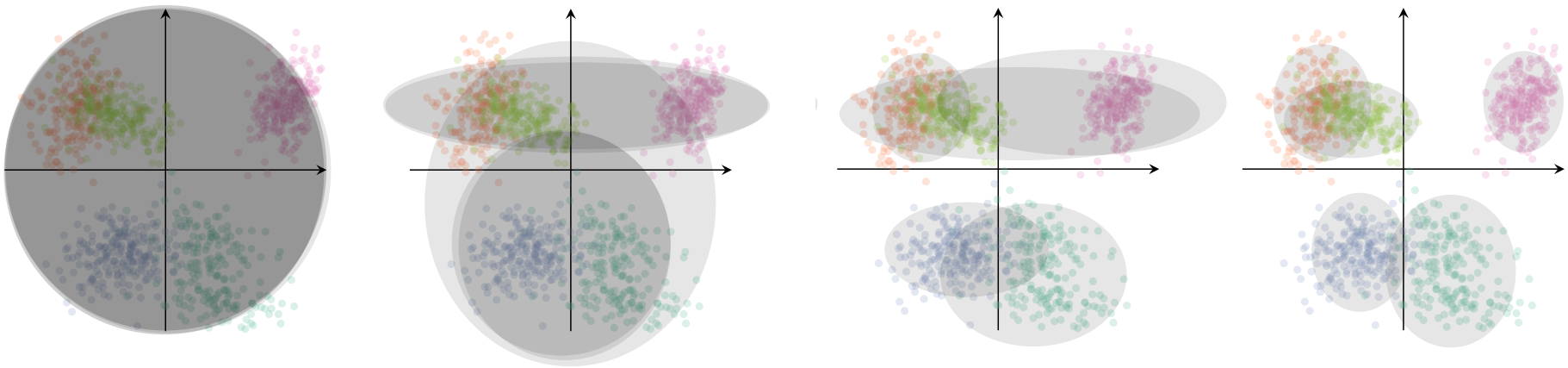






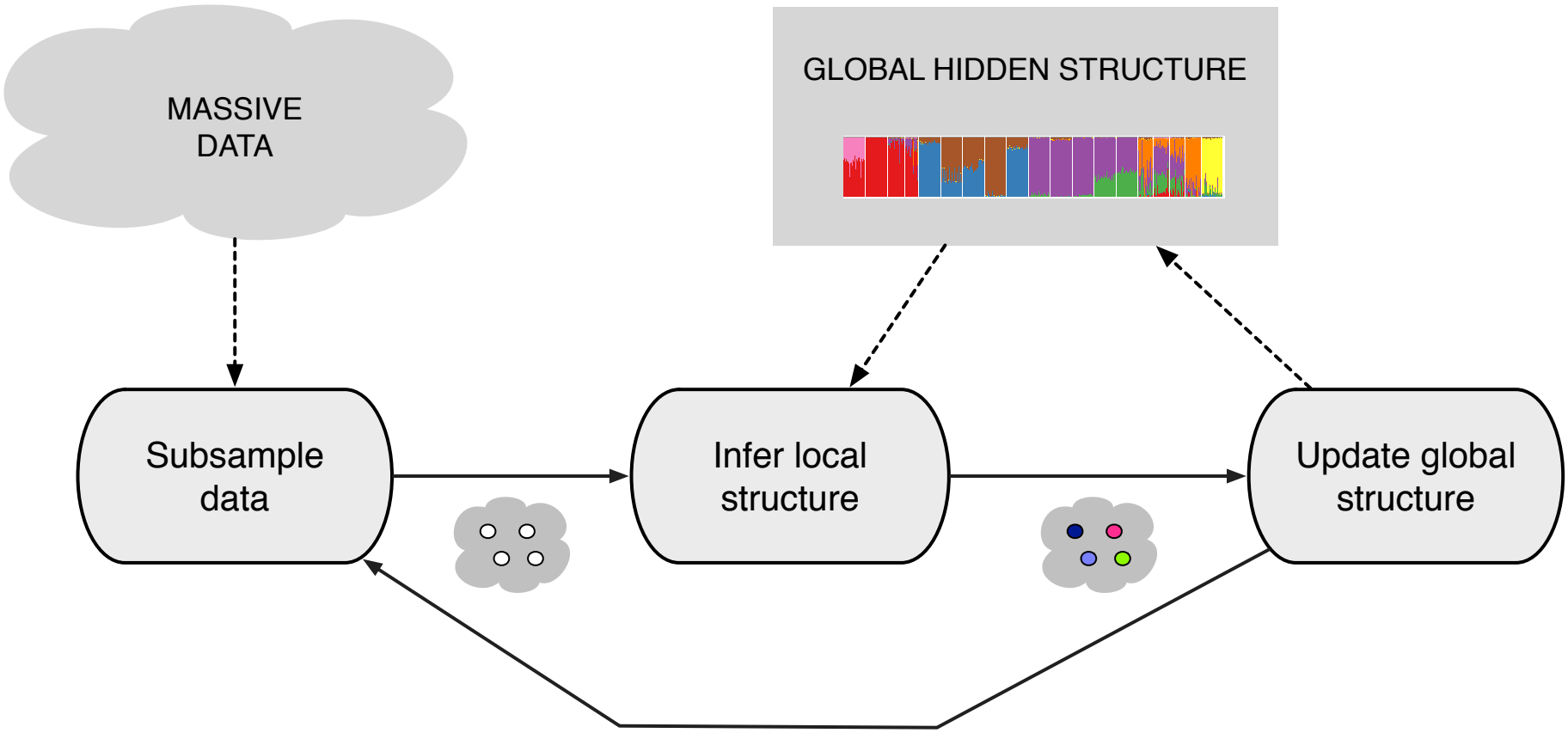
[images by Alp Kucukelbir; Blei+ 2017]

Stochastic variational inference



- Classical VI is inefficient:
  - Do some local computation *for each data point*.
  - Aggregate these computations to re-estimate global structure.
  - Repeat.
- **Stochastic variational inference** (SVI) scales VI to massive data.

Stochastic variational inference



Stochastic optimization

A STOCHASTIC APPROXIMATION METHOD[1]

By Herbert Robbins and Sutton Monro

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

- Replace the gradient with cheaper noisy estimates [Robbins and Monro, 1951]
- Guaranteed to converge to a local optimum [Bottou, 1996]
- ***This algorithm has enabled modern machine learning.***

Stochastic optimization

A STOCHASTIC APPROXIMATION METHOD[1]

By Herbert Robbins and Sutton Monro

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

- Use noisy gradients to update

$$\nu_{t+1} = \nu_t + \rho_t \hat{\nabla}_\nu \mathcal{L}(\nu_t).$$

- Requires unbiased gradients $\mathbb{E}\left[\hat{\nabla}_\nu \mathcal{L}(\nu)\right] = \nabla_\nu \mathcal{L}(\nu)$
- Requires the step size sequence ρ_t follows Robbins-Monro conditions (Modern methods involve more sophisticated step-size schedules.)

The complete conditional of the global variable

- The complete conditional of the global variable is

$$p(\beta \mid \mathbf{z}, \mathbf{x}) = \text{expfam}(\beta\,;\, \eta_g(\mathbf{z}, \mathbf{x}))$$
$$\eta_g(\mathbf{z}, \mathbf{x}) = \alpha + \textstyle\sum_{i=1}^{n} t(z_i, x_i),$$

  where $t(\cdot, \cdot)$ is a function and α is the hyperparameter to the prior.

  (This is from classical theory of conjugate priors [Diaconis and Ylvisaker 1979].)

- The coordinate ascent update is

$$\lambda^* = \alpha + \textstyle\sum_{i=1}^{n} \mathbb{E}_{\phi_i}[t(Z_i, x_i)]$$

- For large datasets, this update is expensive.

Stochastic variational inference

- The **natural gradient** of the ELBO [Amari, 1998; Sato, 2001; Hoffman+ 2013] :

$$\nabla_\lambda^{\text{nat}} \mathcal{L}(\lambda) = \left(\alpha + \sum_{i=1}^n \mathbb{E}_{\phi_i^*}[t(Z_i, x_i)]\right) - \lambda.$$

- Construct a **noisy natural gradient**:

$$j \sim \text{Uniform}(1, \ldots, n)$$
$$\hat{\nabla}_\lambda^{\text{nat}} \mathcal{L}(\lambda) = \alpha + n\mathbb{E}_{\phi_j^*}[t(Z_j, x_j)] - \lambda.$$

- It is **good for stochastic optimization.**
  - Its expectation is the exact natural gradient (*unbiased*).
  - It only depends on optimized parameters of one data point (*cheap*).

Stochastic variational inference

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.

Initialize λ randomly.

Set ρ_t appropriately.

while *not converged* **do**

Sample $j \sim \mathrm{Unif}(1, \ldots, n)$.

Set local parameter

$$\phi \leftarrow \mathbb{E}_\lambda \left[\eta_\ell(\beta, x_j) \right].$$

Set intermediate global parameter

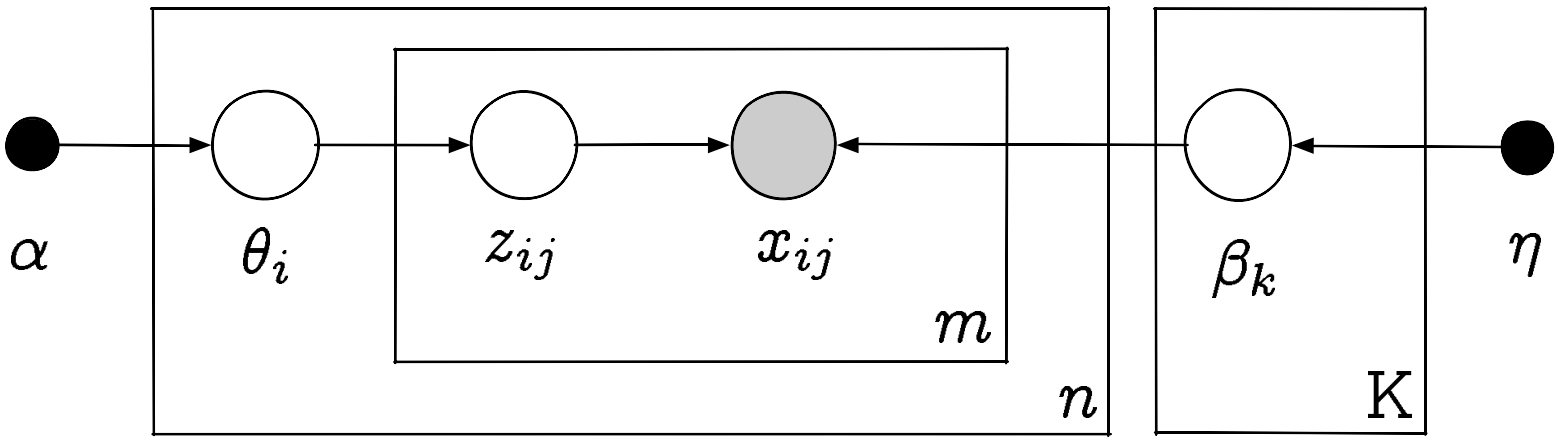
$$\hat{\lambda} = \alpha + n \mathbb{E}_\phi [t(Z_j, x_j)].$$

Set global parameter

$$\lambda = (1 - \rho_t)\, \lambda + \rho_t\, \hat{\lambda}.$$

end

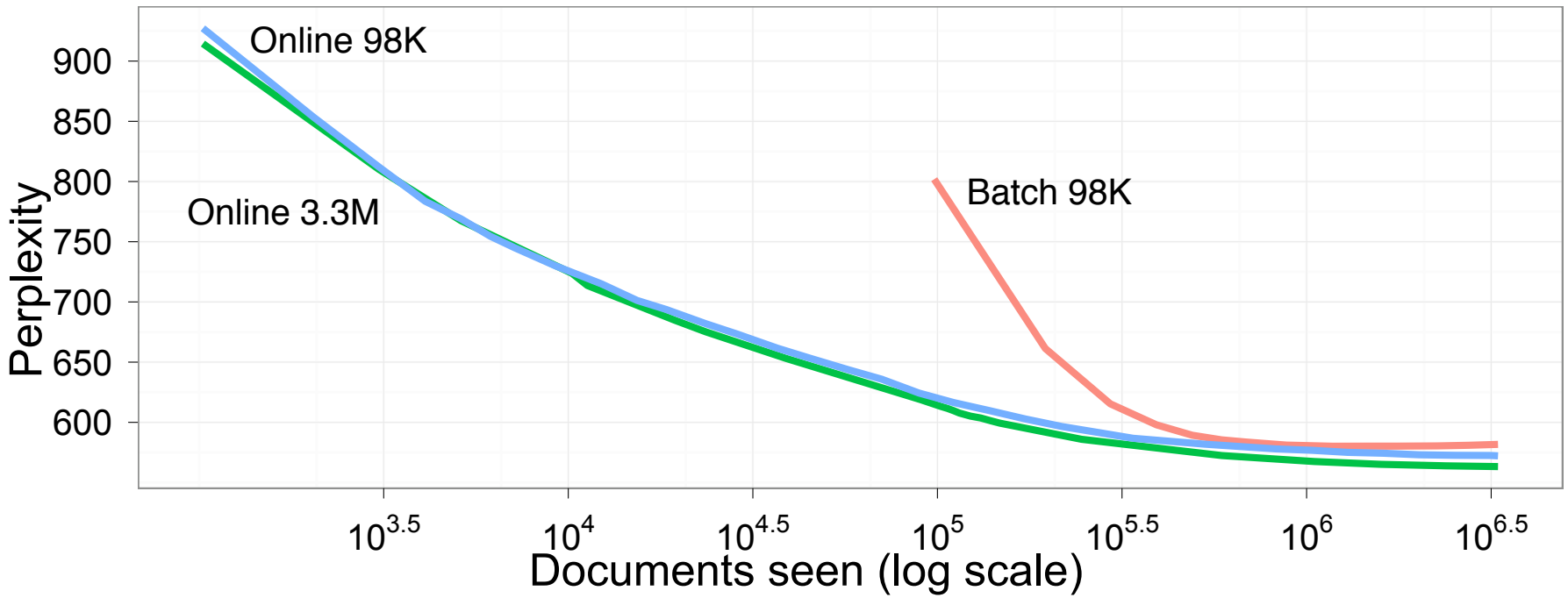
Latent Dirichlet allocation [Blei+ 2003]



$$\beta_k \sim \text{dirichlet}_V(\eta) \qquad k = 1 \ldots \text{K}$$
$$\theta_i \sim \text{dirichlet}_\text{K}(\alpha) \qquad i = 1 \ldots m$$
$$z_{ij} \sim \text{cat}(\theta_i) \qquad j = 1 \ldots n$$
$$x_{ij} \sim \text{cat}(\beta_{z_{ij}})$$

- A mixed-membership model of document collections, a.k.a. a topic model
- Posterior $\beta_{1:\text{K}}$ are *topics*, each a distribution over the vocabulary.
- The topics reflect themes that run through the collection.

Stochastic variational inference for LDA

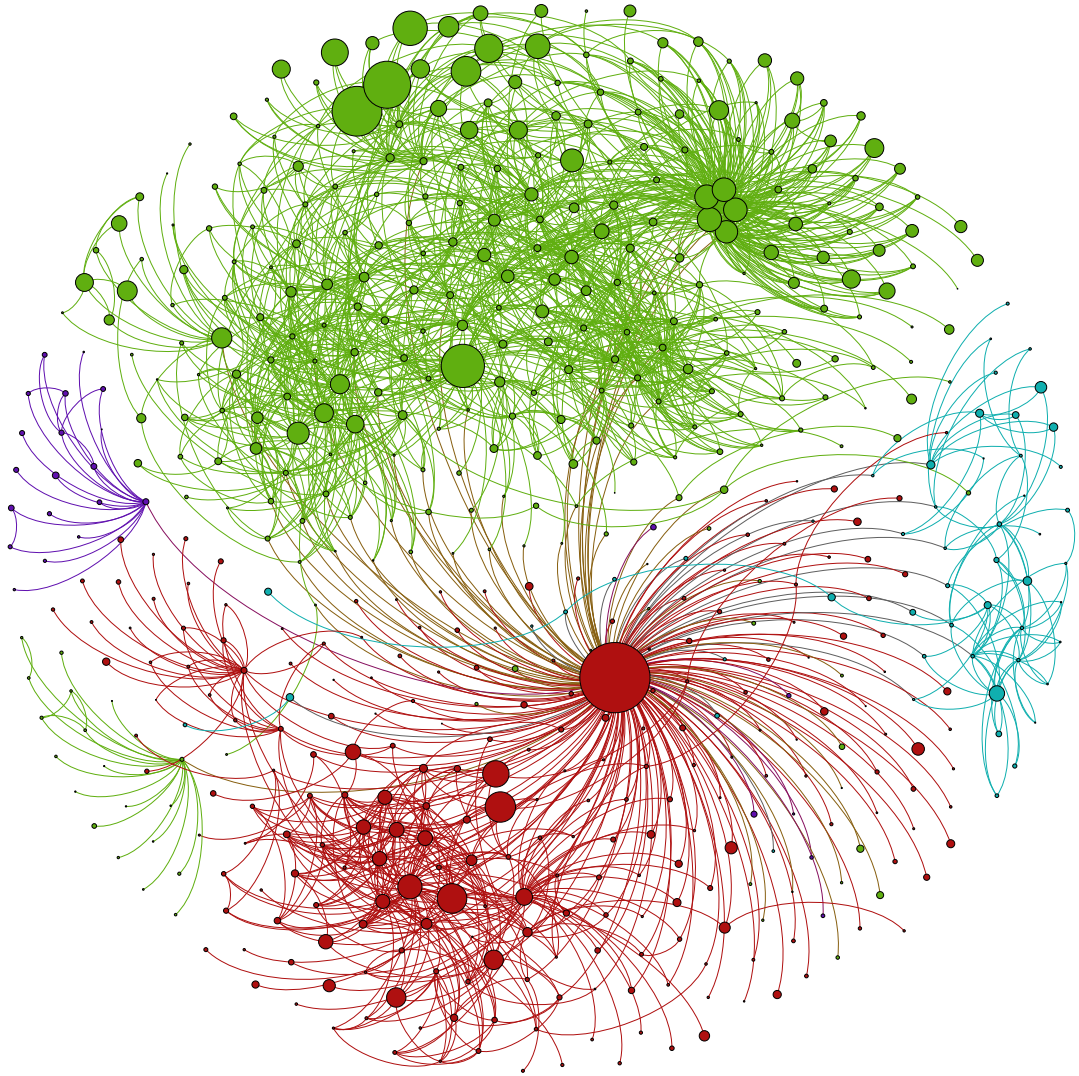


Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

[Hoffman+ 2010]

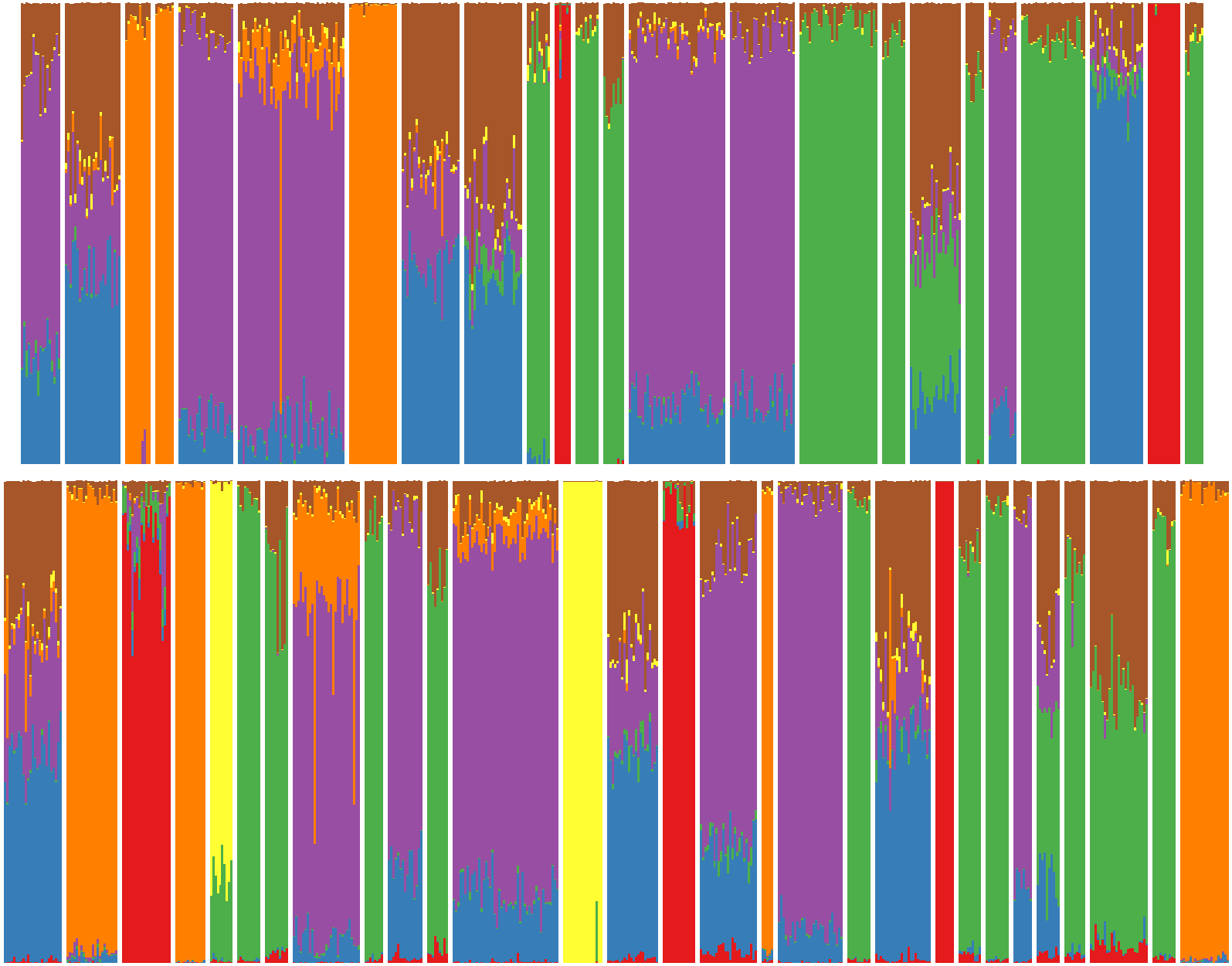
Topics using the HDP, found in 1.8M articles from the New York Times

[Hoffman+ 2013]



Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei 2013]



Population analysis of 2 billion genetic measurements

[Gopalan+ 2016]

Precursors and related work, especially about online EM

- A view of the EM algorithm that justifies incremental, sparse, and other variants

  [Neal and Hinton 1998]

- Convergence of a stochastic approximation version of the EM algorithm

  [Delyon+ 1999]

- Online model selection based on the variational Bayes

  [Sato 2001]

- Unsupervised variational Bayesian learning of nonlinear models

  [Honkela and Valpola 2003]

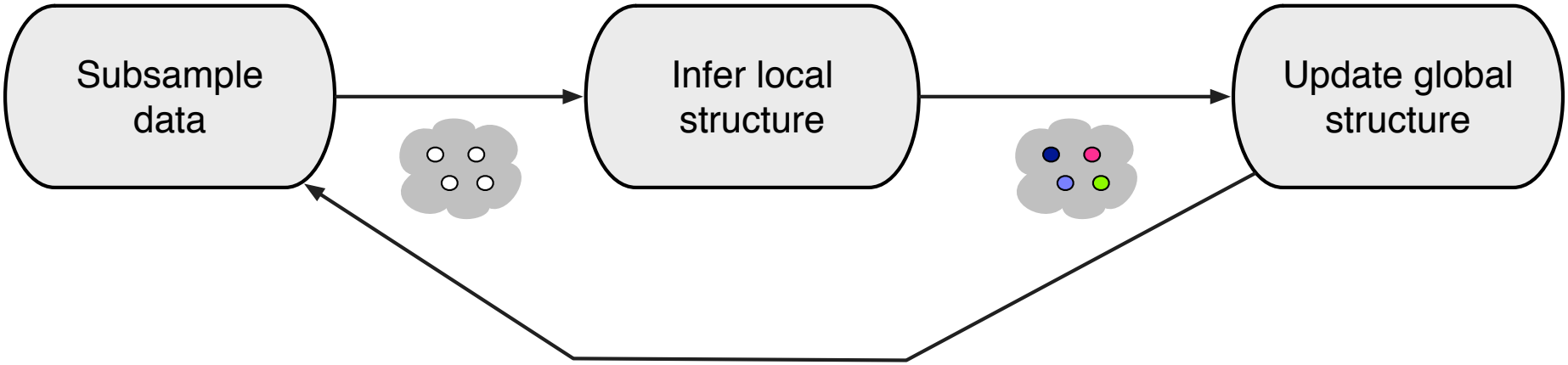
- On-line expectation-maximization algorithm for latent data models

  [Cappe and Moulines 2007]

- Online EM for unsupervised models

  [Liang and Klein 2009]]

SVI scales many models



- Bayesian mixture models
- Time series models (HMMs, linear dynamic systems)
- Factorial models
- Matrix factorization (factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic block models
- Mixed-membership models (LDA and some variants)

Black Box Variational Inference

A.1 Computing $\mathbf{E}[\log(\theta_i \mid \alpha)]$

The need to compute the expected value of the log of a single probability component under the Dirichlet arises repeatedly in deriving the inference and parameter estimation procedures for LDA. This value can be easily computed from the natural parameterization of the exponential family representation of the Dirichlet distribution.

Recall that a distribution is in the exponential family if it can be written in the form:

$$p(x \mid \eta) = h(x) \exp\left\{\eta^T T(x) - A(\eta)\right\},$$

where η is the natural parameter, $T(x)$ is the sufficient statistic, and $A(\eta)$ is the log of the normalization factor.

We can write the Dirichlet in this form by exponentiating the log of Eq. (1):

$$p(\theta \mid \alpha) = \exp\left\{\left(\sum_{i=1}^k (\alpha_i - 1)\log\theta_i\right) + \log\Gamma\left(\sum_{i=1}^k \alpha_i\right) - \sum_{i=1}^k \log\Gamma(\alpha_i)\right\}.$$

From this form, we immediately see that the natural parameter of the Dirichlet is $\eta_i = \alpha_i - 1$ and the sufficient statistic is $T(\theta_i) = \log\theta_i$. Furthermore, using the general fact that the derivative of the log normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic, we obtain:

$$\mathrm{E}[\log\theta_i \mid \alpha] = \Psi(\alpha_i) - \Psi\left(\sum_{j=1}^k \alpha_j\right)$$

where Ψ is the digamma function, the first derivative of the log Gamma function.

A.3.2 Variational Dirichlet

Next, we maximize Eq. (15) with respect to γ_i, the ith component of the posterior Dirichlet parameter. The terms containing γ_i are:

$$\begin{aligned} L_{[\gamma]} = &\sum_{i=1}^k (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right) + \sum_{n=1}^N \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right) \\ &- \log\Gamma\left(\sum_{j=1}^k \gamma_j\right) + \log\Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right). \end{aligned}$$

This simplifies to:

$$L_{[\gamma]} = \sum_{i=1}^k \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right)\left(\alpha_i + \sum_{n=1}^N \phi_{ni} - \gamma_i\right) - \log\Gamma\left(\sum_{j=1}^k \gamma_j\right) + \log\Gamma(\gamma_i).$$

We take the derivative with respect to γ_i:

$$\frac{\partial L}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \sum_{n=1}^N \phi_{ni} - \gamma_i\right) - \Psi'\left(\sum_{j=1}^k \gamma_j\right)\sum_{j=1}^k \left(\alpha_j + \sum_{n=1}^N \phi_{nj} - \gamma_j\right).$$

Setting this equation to zero yields a maximum at:

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni}. \tag{17}$$

Since Eq. (17) depends on the variational multinomial ϕ, full variational inference requires alternating between Eqs. (16) and (17) until the bound converges.

Finally, we expand Eq. (14) in terms of the model parameters (α, β) and the variational parameters (γ, ϕ). Each of the five lines below expands one of the five terms in the bound:

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) = &\log\Gamma\left(\sum_{j=1}^k \alpha_j\right) - \sum_{i=1}^k \log\Gamma(\alpha_i) + \sum_{i=1}^k (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right) \\ &+ \sum_{n=1}^N \sum_{i=1}^k \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right) \\ &+ \sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^V \phi_{ni} w_n^j \log\beta_{ij} \\ &- \log\Gamma\left(\sum_{j=1}^k \gamma_j\right) + \sum_{i=1}^k \log\Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right) \\ &- \sum_{n=1}^N \sum_{i=1}^k \phi_{ni}\log\phi_{ni}, \end{aligned} \tag{15}$$

where we have made use of Eq. (8).

In the following two sections, we show how to maximize this lower bound with respect to the variational parameters ϕ and γ.

A.3.1 Variational multinomial

We first maximize Eq. (15) with respect to ϕ_{ni}, the probability that the nth word is generated by latent topic i. Observe that this is a constrained maximization since $\sum_{i=1}^k \phi_{ni} = 1$.

We form the Lagrangian by isolating the terms which contain ϕ_{ni} and adding the appropriate Lagrange multipliers. Let β_{iv} be $p(w_n^v = 1 \mid z^i = 1)$ for the appropriate v. (Recall that each w_n is a vector of size V with exactly one component equal to one; we can select the unique v such that $w_n^v = 1$):

$$L_{[\phi_{ni}]} = \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right) + \phi_{ni}\log\beta_{iv} - \phi_{ni}\log\phi_{ni} + \lambda_n\left(\sum_{j=1}^k \phi_{ni} - 1\right),$$

where we have dropped the arguments of L for simplicity, and where the subscript ϕ_{ni} denotes that we have retained only those terms in L that are a function of ϕ_{ni}. Taking derivatives with respect to ϕ_{ni}, we obtain:

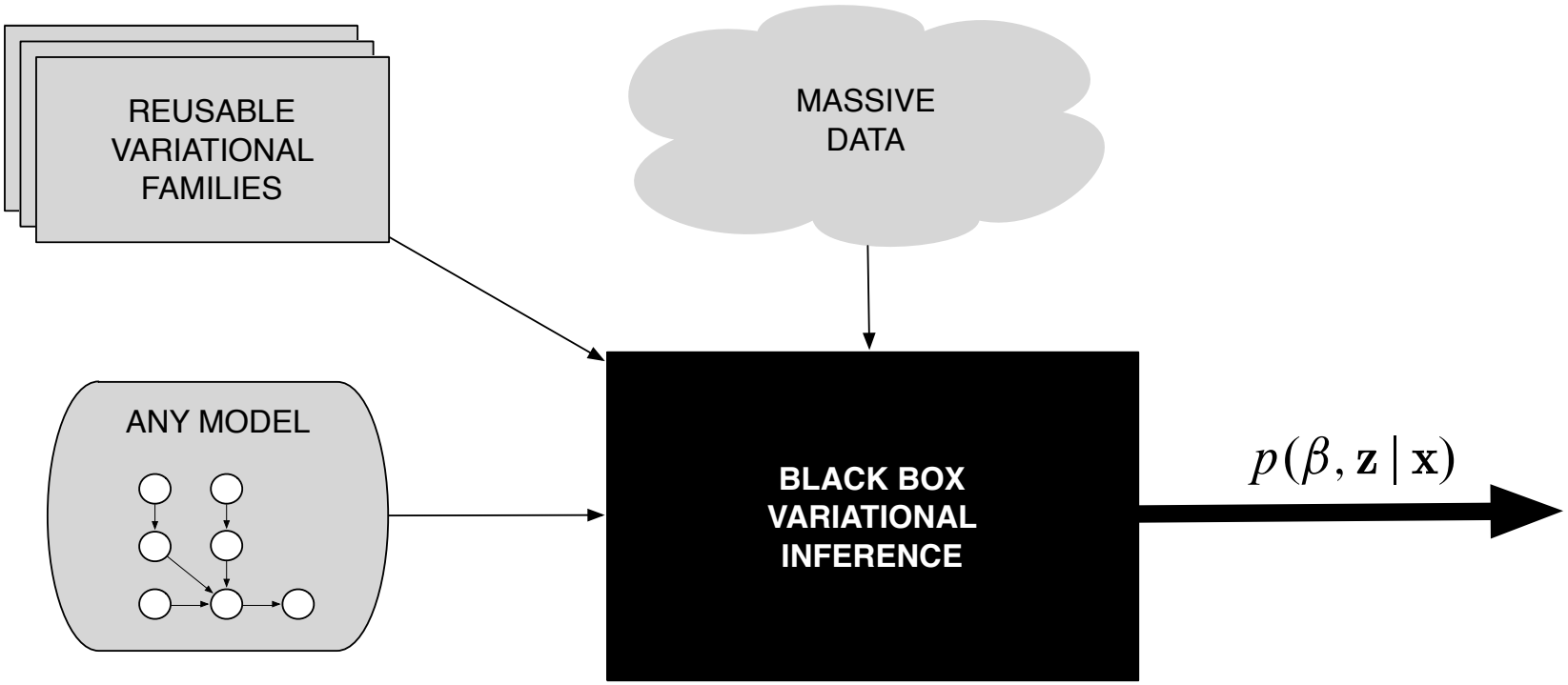
$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right) + \log\beta_{iv} - \log\phi_{ni} - 1 + \lambda.$$

Setting this derivative to zero yields the maximizing value of the variational parameter ϕ_{ni} (cf. Eq. 6):

$$\phi_{ni} \propto \beta_{iv}\exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^k \gamma_j\right)\right). \tag{16}$$

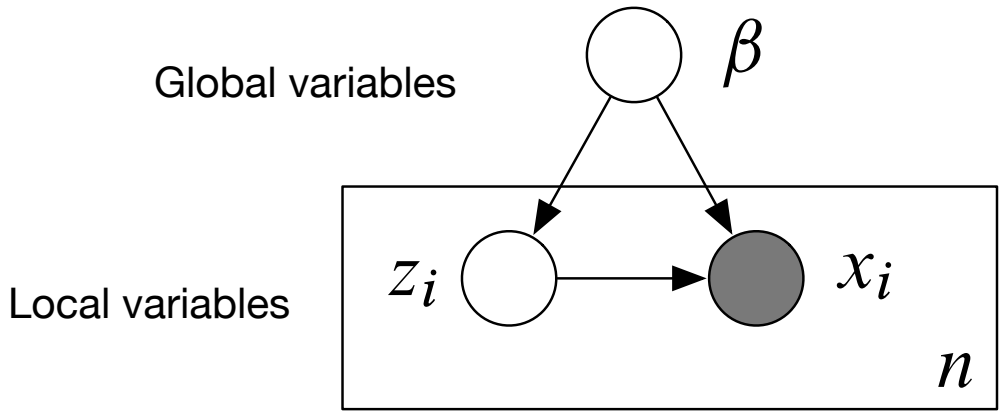
[from Blei+ 2003]

Black box variational inference



- Easily use variational inference with **any model**; no more appendices!
- Perform inference with **massive data**
- **No mathematical work** beyond specifying the model

Nonconjugate models



$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- Nonlinear time series models
- Deep latent Gaussian models
- Models with attention
- Generalized linear models
- Stochastic volatility models
- Discrete choice models
- Bayesian neural networks
- Deep exponential families
- Correlated topic models
- Sigmoid belief networks

Black box variational inference

$$\mathcal{L}(\nu) = \underbrace{\mathbb{E}_q\left[\log p(\beta, \mathbf{z}, \mathbf{x})\right]}_{\text{Expected complete log likelihood}} - \underbrace{\mathbb{E}_q\left[\log q(\beta, \mathbf{z}; \nu)\right]}_{\text{Negative entropy}}$$

The main idea behind BBVI:

- write the **gradient of the ELBO as an expectation**
- sample from $q(\cdot)$ to form a **Monte Carlo estimate of the gradient**
- use the MC estimate in a **stochastic optimization**

Black box variational inference

$$\mathcal{L}(\nu) = \underbrace{\mathbb{E}_q\left[\log p(\beta, \mathbf{z}, \mathbf{x})\right]}_{\text{Expected complete log likelihood}} - \underbrace{\mathbb{E}_q\left[\log q(\beta, \mathbf{z}; \nu)\right]}_{\text{Negative entropy}}$$

- Keep in mind the **black box criteria**.
- We should only need to:
  - sample from $q(\beta, \mathbf{z})$
  - evaluate things about $q(\beta, \mathbf{z})$
  - evaluate $\log p(\beta, \mathbf{z}, \mathbf{x})$
- These criteria let us perform approximate inference on many models.

BBVI # 1: The score gradient

$$\nabla_\nu \mathcal{L} = \mathbb{E}_{q(\mathbf{z};\nu)}[\underbrace{\nabla_\nu \log q(\mathbf{z};\nu)}_{\text{score function}} \underbrace{(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\nu))}_{\text{instantaneous ELBO}}]$$

- Use the score function to write the gradient as an expectation.

  [Ji+ 2010; Paisley+ 2012; Wingate+ 2013; Ranganath+ 2014; Mnih+ 2014]

- Also called the likelihood ratio or REINFORCE gradient

  [Glynn 1990; Williams 1992]

- Pushes ν to give high probability on $\mathbf{z}$ with large instantaneous ELBO.

BBVI # 1: The score gradient

$$\nabla_\nu \mathcal{L} = \mathbb{E}_{q(\mathbf{z};\nu)}[\underbrace{\nabla_\nu \log q(\mathbf{z};\nu)}_{\text{score function}} \underbrace{(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\nu))}_{\text{instantaneous ELBO}}]$$

Satisfies the **black box criteria** — no model-specific analysis needed.

- sample from $q(\mathbf{z};\nu)$
- evaluate $\nabla_\nu \log q(\mathbf{z};\nu)$
- evaluate $\log p(\mathbf{x},\mathbf{z})$ and $\log q(\mathbf{z})$

Score-gradient black box variational inference

Input: data $\mathbf{x}$, model $p(\mathbf{z}, \mathbf{x})$.

Initialize ν randomly.

Set ρ_j appropriately.

while *not converged* **do**

Take S samples from the variational distribution

$$\mathbf{z}[s] \sim q(\mathbf{z}; \nu) \quad s = 1 \ldots S$$

Calculate the noisy score gradient

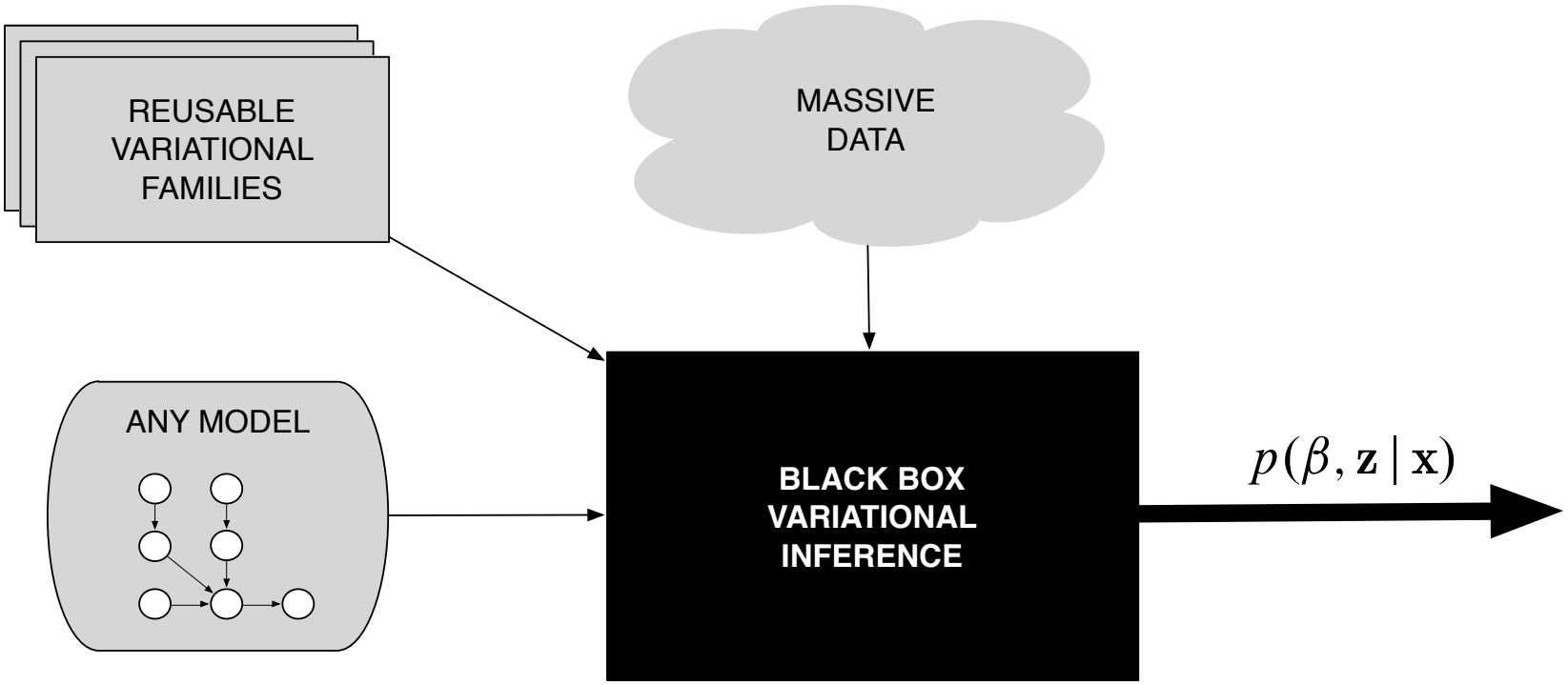
$$\tilde{g}_t = \frac{1}{S} \sum_{s=1}^{S} \nabla_\nu \log q(\mathbf{z}[s]; \nu_t)(\log p(\mathbf{x}, \mathbf{z}[s]) - \log q(\mathbf{z}[s]; \nu_t))$$

Update the variational parameters

$$\nu_{t+1} = \nu_t + \rho_t \, \tilde{g}_t$$

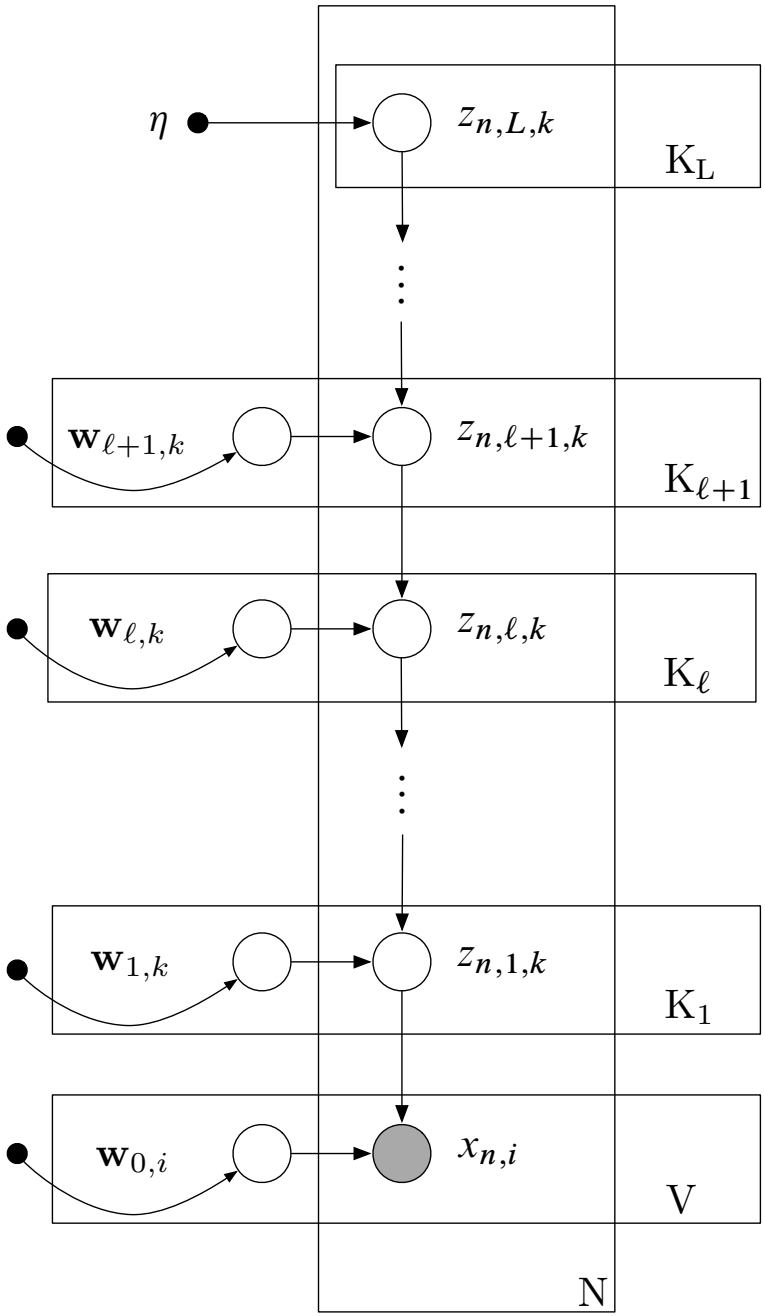
end

BBVI: Making it work



- Control the variance of the gradient [e.g., Paisley+ 2012; Ranganath+ 2014]
  - Rao-Blackwellization, control variates, importance sampling
- Adaptive step sizes [e.g., Duchi+ 2011; Kingma and Ba 2014; Kucukelbir+ 2016]
- SVI, for massive data [Hoffman+ 2013]

Deep exponential families

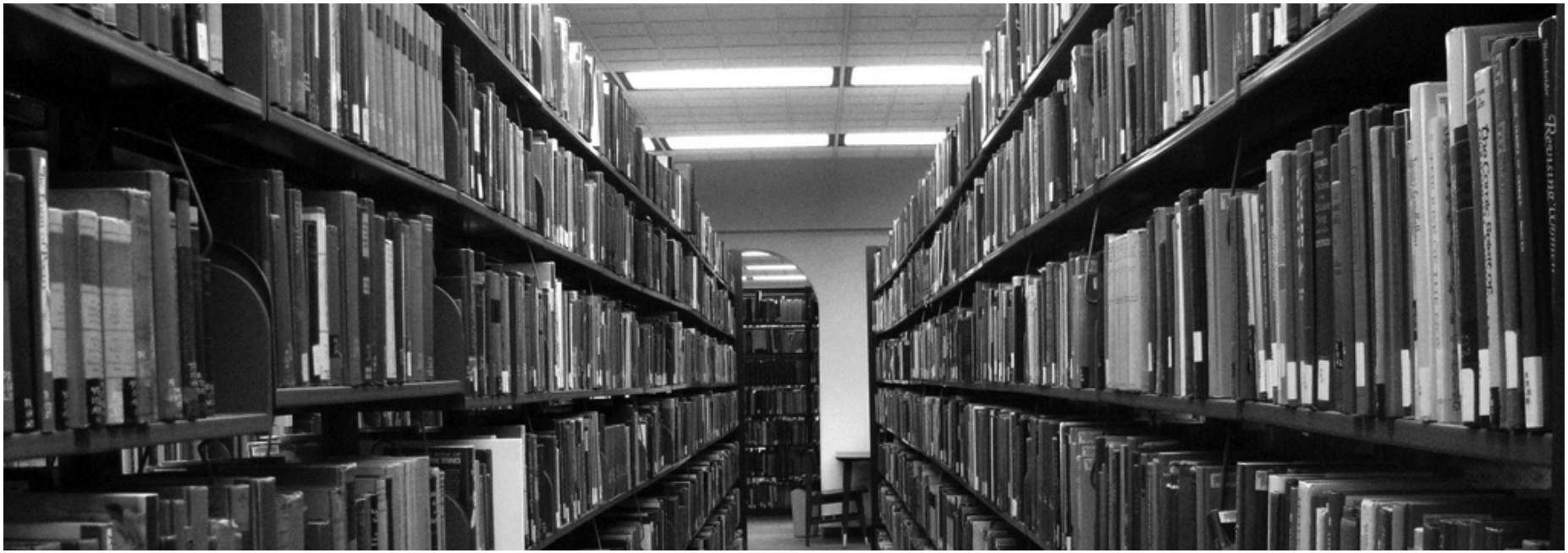


$$z_{n,L,k} \sim \text{Exp-Fam}(\eta)$$

$$z_{n,\ell,k} \sim \text{Exp-Fam}(g(\mathbf{w}_{\ell,k}^\top \mathbf{z}_{n,\ell+1}))$$

$$x_{n,i} \sim \text{Exp-Fam}(g(\mathbf{w}_{0,i}^\top \mathbf{z}_{n,1}))$$

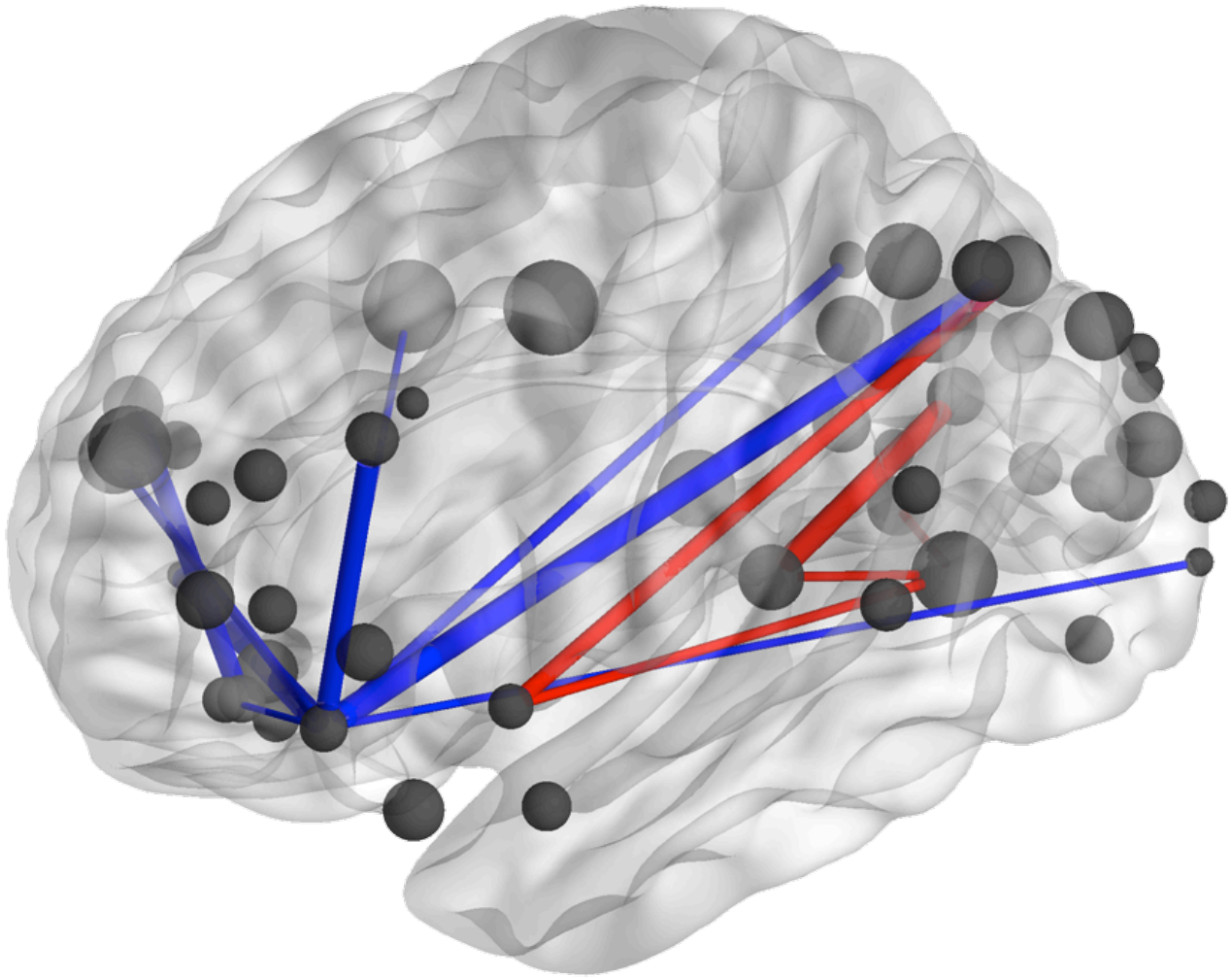
Empirical study of DEFs



- NYT and Science (about 150K documents in each, about 7K terms)
- Many models: adjusted depth, types of latents, priors, and link
- Held-out perplexity (lower is better) [Wallach+ 2009]

DEF evaluation

Model	$p(\mathbf{w})$	*NYT*	*Science*
LDA [Blei+ 2003]		2717	1711
DocNADE [Larochelle+ 2012]		2496	1725
Sparse Gamma 100	$\emptyset$	2525	1652
Sparse Gamma 100-30	Γ	2303	1539
Sparse Gamma 100-30-15	Γ	**2251**	1542
Sigmoid 100	$\emptyset$	2343	1633
Sigmoid 100-30	$\mathcal{N}$	2653	1665
Sigmoid 100-30-15	$\mathcal{N}$	2507	1653
Poisson 100	$\emptyset$	2590	1620
Poisson 100-30	$\mathcal{N}$	2423	1560
Poisson 100-30-15	$\mathcal{N}$	2416	1576
Poisson log-link 100-30	Γ	2288	**1523**
Poisson log-link 100-30-15	Γ	2366	1545



Neuroscience analysis of 220 million fMRI measurements

[Manning+ 2014]

BBVI #2: The reparameterization gradient

- Suppose $\log p(\mathbf{x}, \mathbf{z})$ and $\log q(\mathbf{z})$ are differentiable with respect to $\mathbf{z}$.
- Suppose the variational distribution can be written with a transformation,

$$\begin{aligned}\epsilon &\sim s(\epsilon) \\ \mathbf{z} &= t(\epsilon, \nu) \\ &\rightarrow \mathbf{z} \sim q(\mathbf{z}; \nu).\end{aligned}$$

  For example,

$$\begin{aligned}\epsilon &\sim \text{Normal}(0, 1) \\ z &= \epsilon\sigma + \mu \\ &\rightarrow z \sim \text{Normal}(\mu, \sigma^2).\end{aligned}$$

- The variational parameters are part of the transformation.
  But they are not involved in the “noise” distribution.

BBVI #2: The reparameterization gradient

$$\nabla_\nu \mathcal{L} = \mathbb{E}_{s(\epsilon)} \left[\underbrace{\nabla_{\mathbf{z}} [\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \nu)]}_{\text{gradient of instananeous ELBO}} \quad \underbrace{\nabla_\nu \, t(\epsilon, \nu)}_{\text{gradient of transformation}} \right]$$

- This is the reparameterization gradient, another tool for BBVI.

  [Glasserman 1991; Fu 2006; Kingma+ 2014; Rezende+ 2014; Titsias+ 2014]

- Can use autodifferentiation to take gradients (especially of the model)

- Can use and reuse different transformations [e.g., Naesseth+ 2017]

Black box variational inference

Input: data $\mathbf{x}$, model $p(\mathbf{z}, \mathbf{x})$.

Initialize ν randomly.

Set ρ_t appropriately.

while *not converged* **do**

Take S samples from the auxillary variable

$$\epsilon_s \sim s(\epsilon) \quad s = 1 \ldots S$$

Calculate the noisy gradient

$$\tilde{g}_t = \frac{1}{S} \sum_{s=1}^{S} \nabla_{\mathbf{z}} [\log p(\mathbf{x}, t(\epsilon_s, \nu_n)) - \log q(t(\epsilon_s, \nu_n); \nu_n)] \nabla_{\nu} t(\epsilon_s, \nu_n)$$

Update the variational parameters

$$\nu_{t+1} = \nu_t + \rho_t \tilde{g}_t$$

end

Shopper on 5.7M purchases.

[Ruiz+ 2017]

Analysis of 1.7M taxi trajectories, in Stan

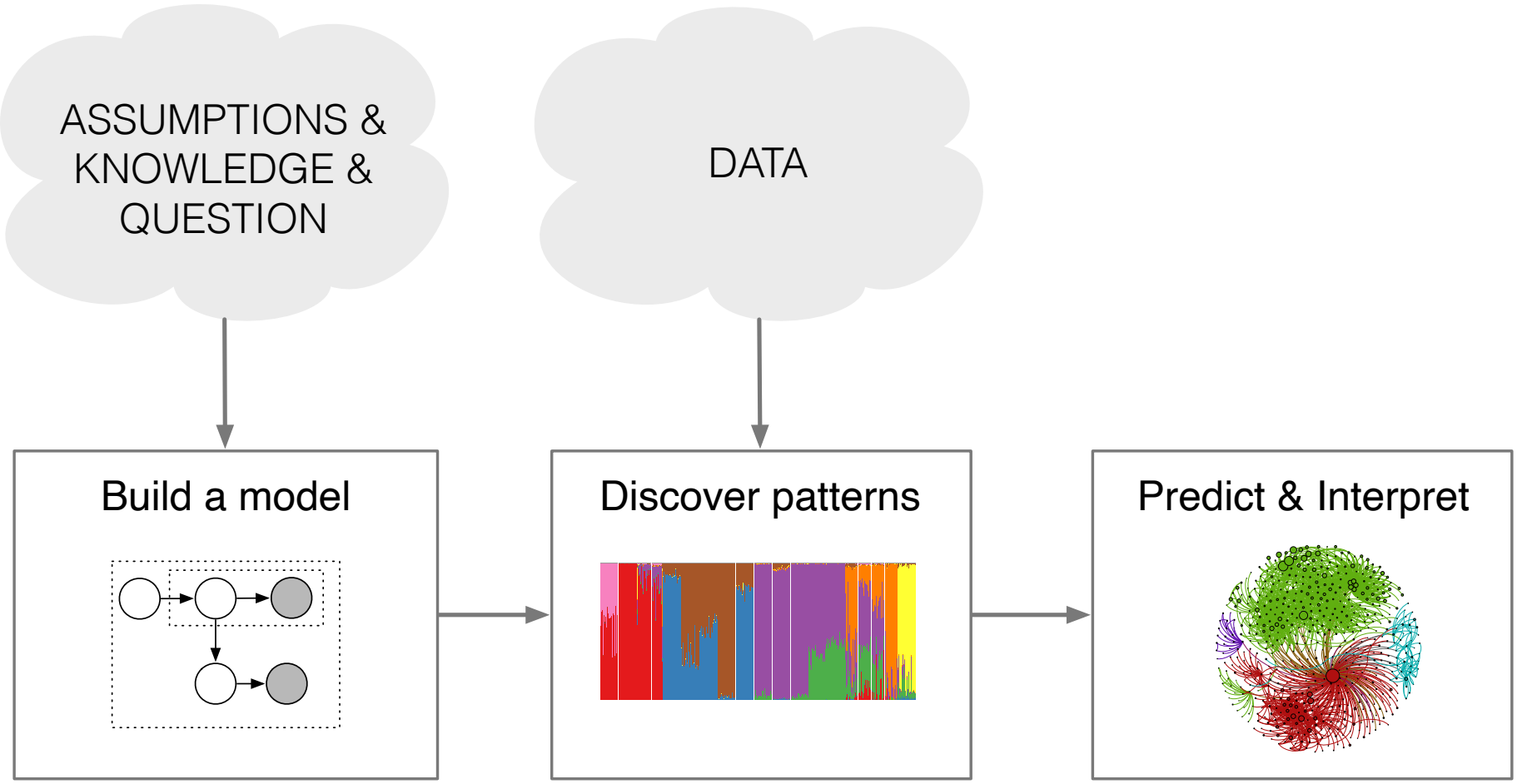
[Kucukelbir+ 2017]

Discussion

PROBABILISTIC MACHINE LEARNING

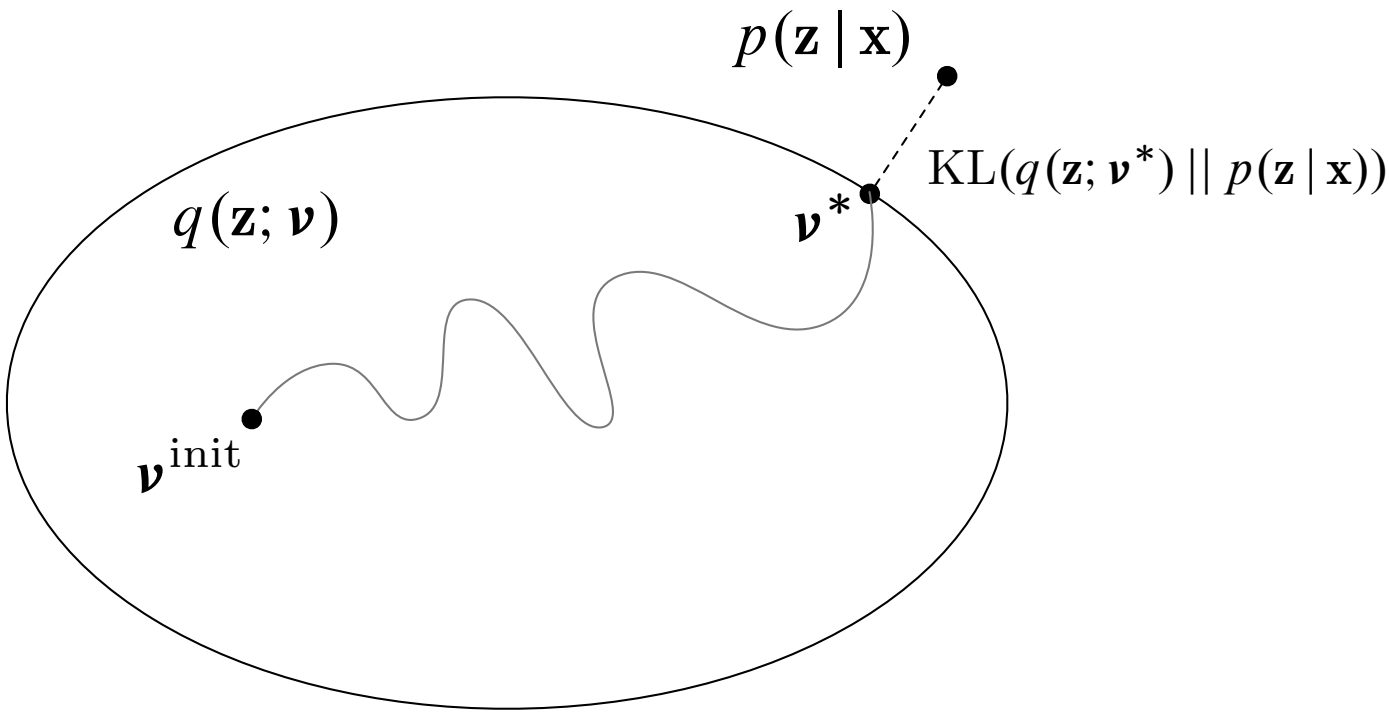
- ML methods that ***connect domain knowledge to data***.
- Provides a computational methodology for analyzing data
- Goal: A methodology that is ***expressive, scalable, easy to develop***

The probabilistic pipeline



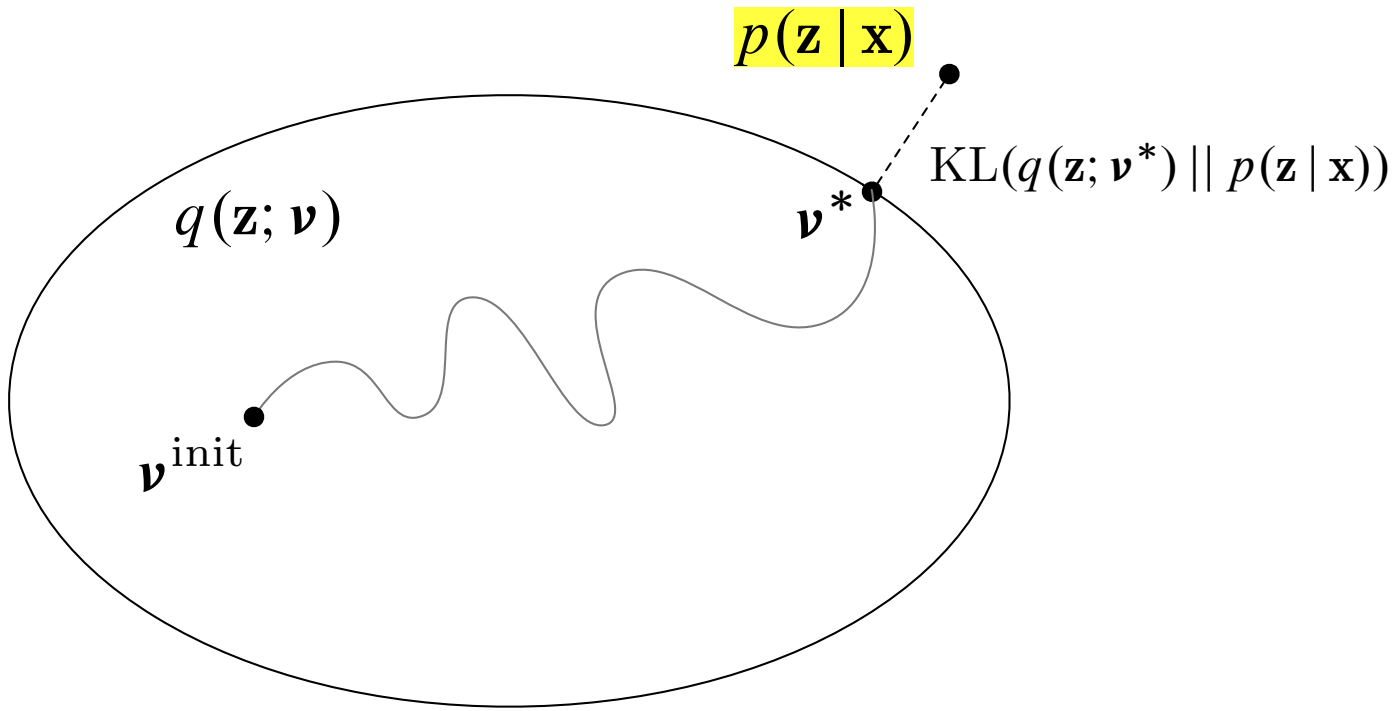
- **Posterior inference** is the key algorithmic problem.
- Answers the question: What does this model say about this data?
- VI provides **scalable** and **general** approaches to posterior inference

Stochastic optimization makes VI better



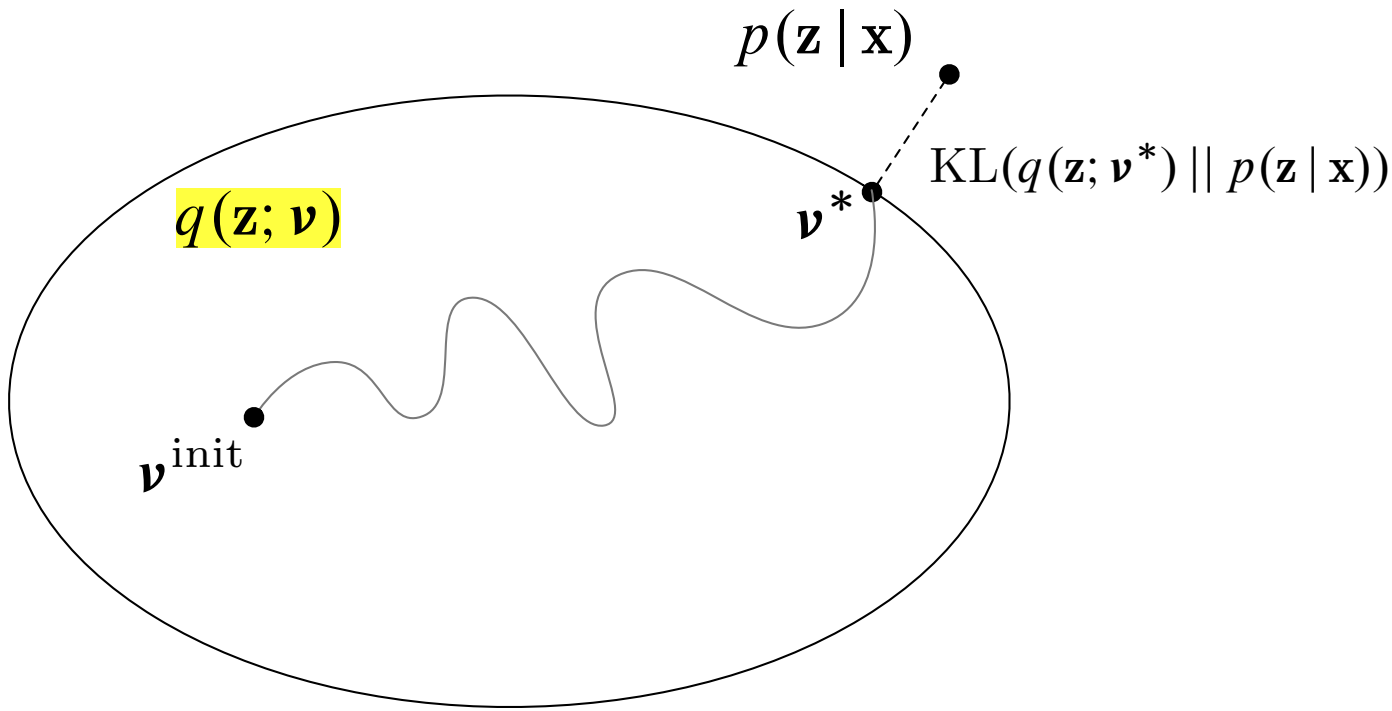
- **Stochastic VI** scales up VI to massive data.
- **Black box VI** generalizes VI to a wide class of models.

What classes of models can VI handle?



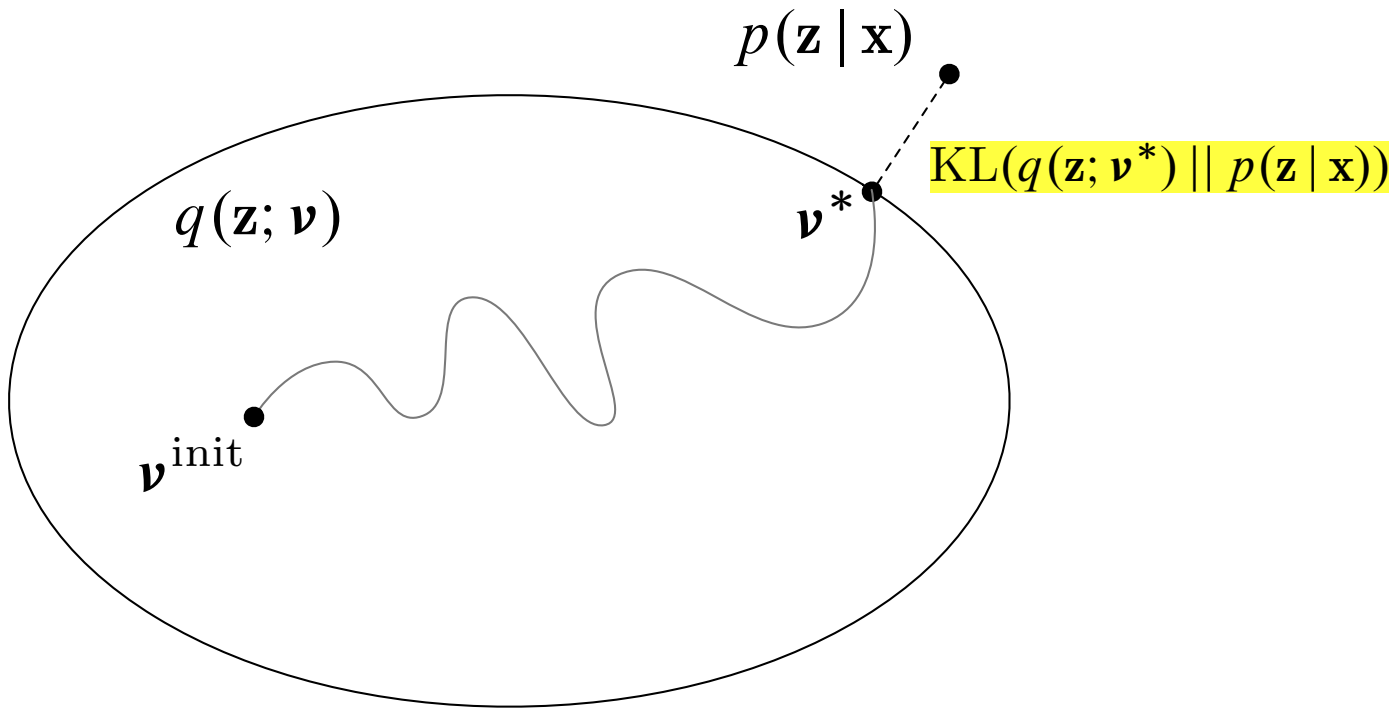
- Conditionally conjugate [Gharamani and Beal 2001; Hoffman+ 2013]
- Not ↑, but can differentiate the log likelihood [Kucukelbir+ 2015]
- Not ↑, but can calculate the log likelihood [Ranganath+ 2014]
- Not ↑, but can sample from the model [Ranganath+ 2017]

How can we expand the variational family?



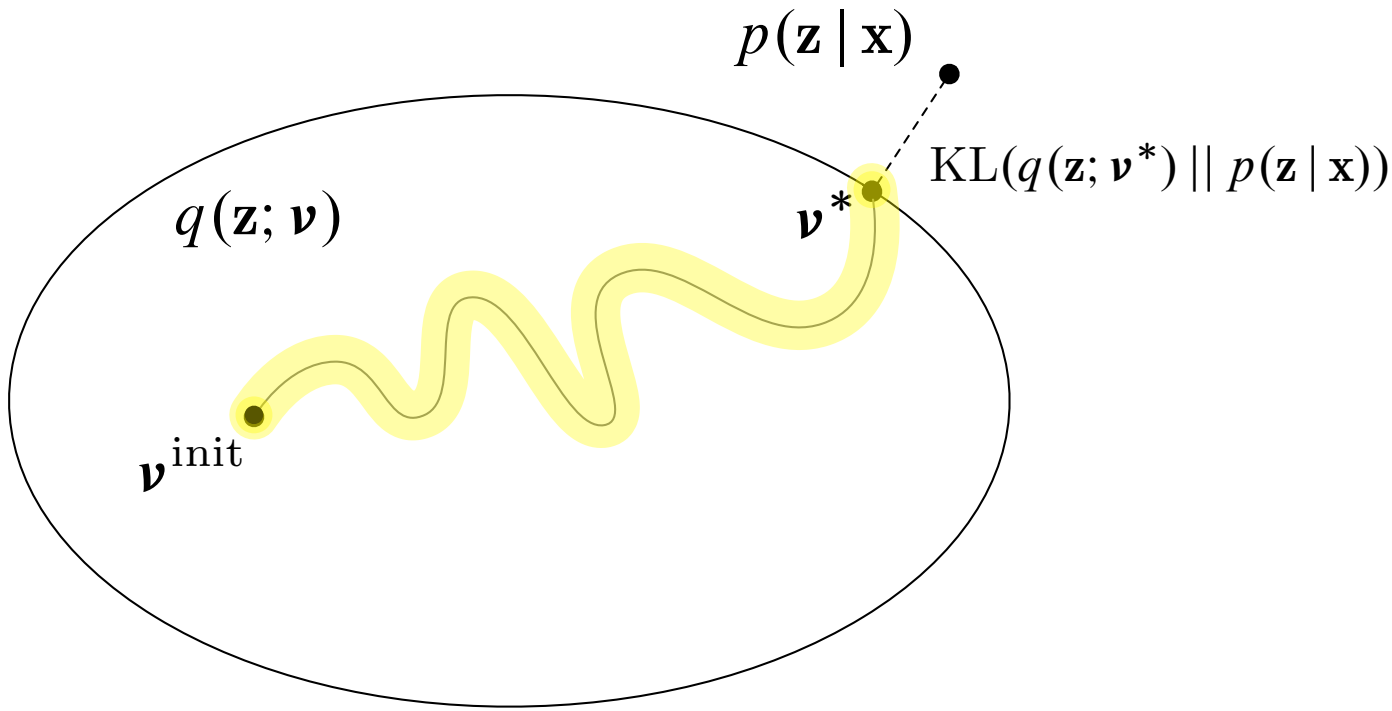
- Structured variational inference [Saul and Jordan 1996; Hoffman and Blei 2015]
- Variational models [Lawrence 2001; Ranganath+ 2015; Tran+ 2015]
- Amortized inference [Kingma and Welling 2014; Rezende+ 2014]
- Sequential Monte Carlo [Naesseth+ 2018; Maddison+ 2017; Le+ 2017]

Which distance should we use? How good is it?



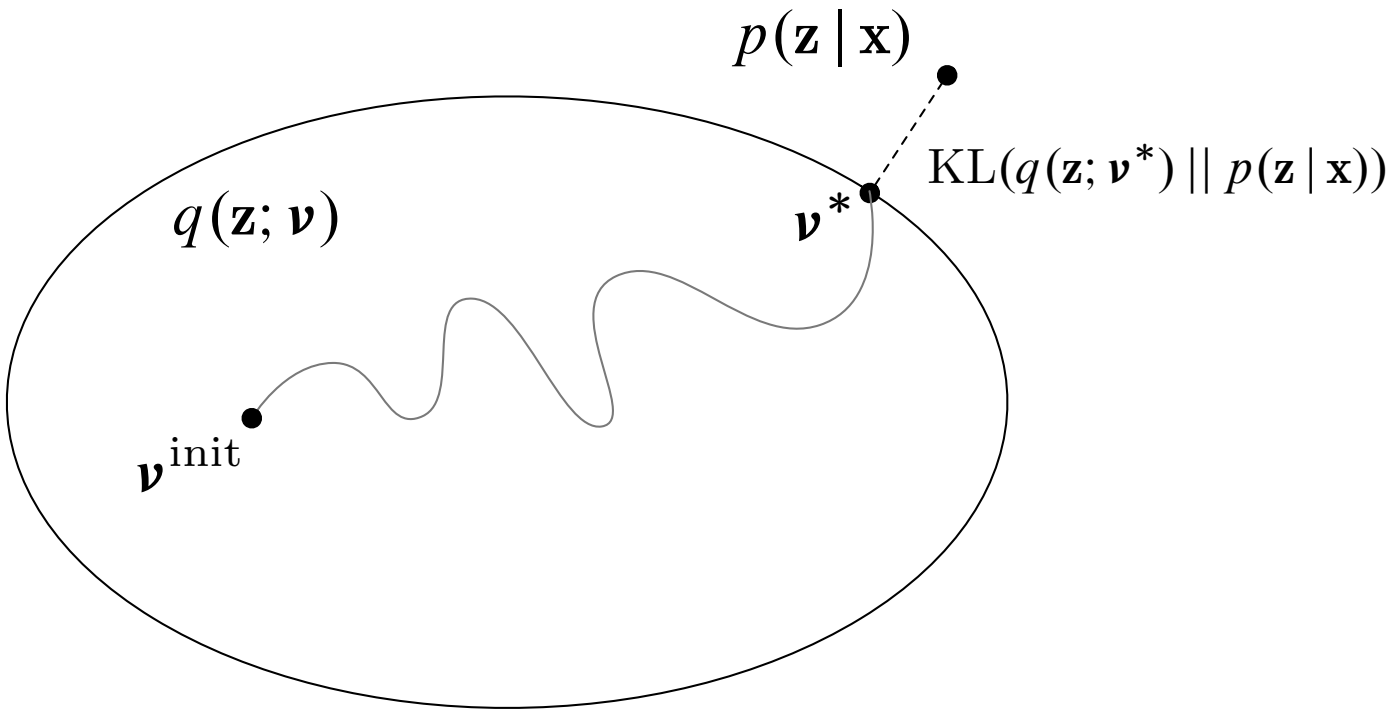
- The "inclusive" $\mathrm{KL}(p||q)$ [Minka 2001; Naesseth+ 2020]
- Generalized variational inference [Knoblauch+ 2019]
- Operator variational inference [Ranganath+ 2016]
- χ-variational inference [Dieng+ 2017]

Can we make the algorithm better?



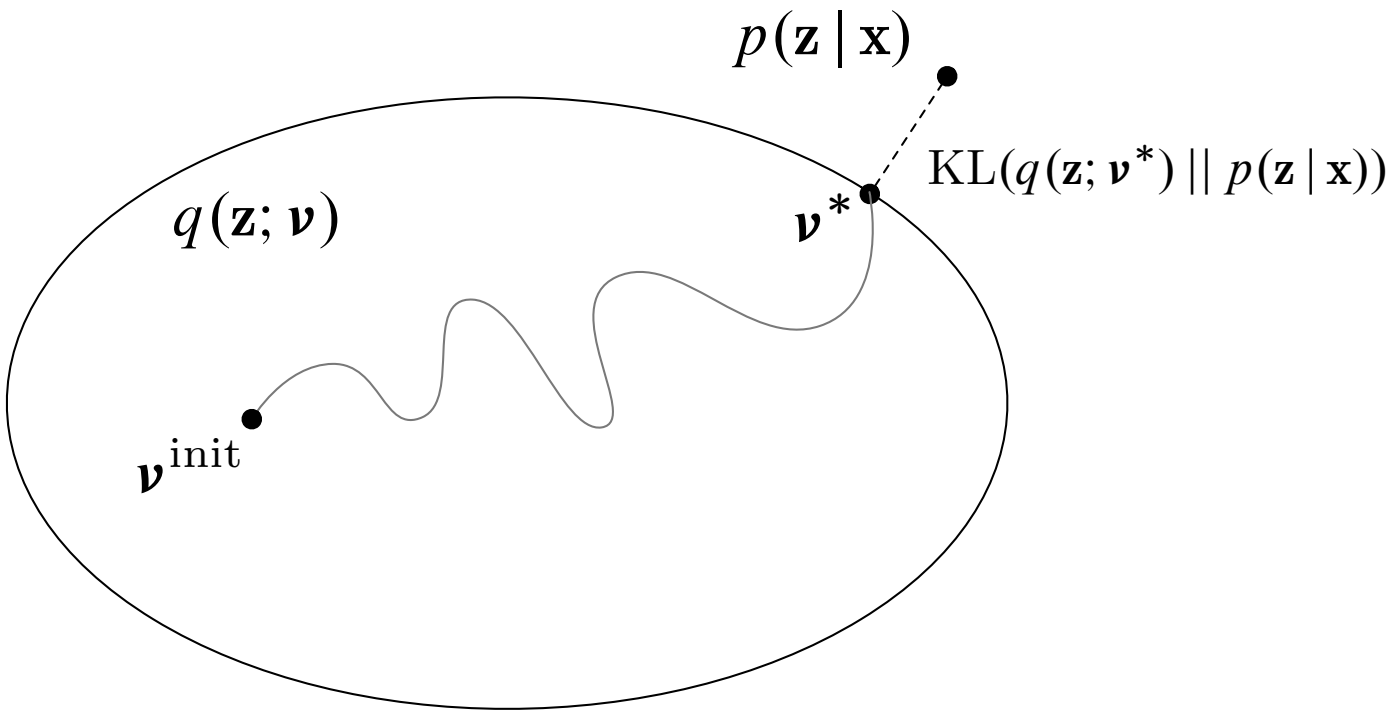
- SVI and structured SVI [Hoffman+ 2013; Hoffman and Blei 2015]
- Stochastic gradient descent as variational inference [Mandt+ 2017]
- Adaptive rates, averaged gradients, control variates, ... [Many papers]

What is guaranteed about VI?



- Asymptotic normality of Gaussian approximations [Hall+ 2011]
- Risk bounds for VI [Pati+ 2017]
- Bernstein Von-Mises, model misspecification [Wang and Blei 2019, 2020]
- Convergence rates for VI [Alquier+ 2016, Zhang and Gao 2019]

How can we use VI in practice?



- Correct for VI's underestimates of the posterior variance [Giordano+ 2015]
- Probabilistic programming [Minka 2014, Kucukelbir+ 2016, Bingham+ 2018, others]
- Best practices for running VI robustly across many models
- How to check variational inferences

References (from our group)

- D. Blei, A. Kucukelbir, J. McAuliffe. **Variational inference: A review for statisticians**. Journal of American Statistical Association, 2017.
- M. Hoffman, D. Blei, C. Wang, J. Paisley. **Stochastic variational inference**. Journal of Machine Learning Research, 2013.
- R. Ranganath, S. Gerrish, D. Blei. **Black box variational inference**. Artificial Intelligence and Statistics, 2014.
- A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, D. Blei. **Automatic differentiation variational inference**. Journal of Machine Learning Research, 2017.
- Y. Wang and D. Blei. **Frequentist consistency of variational Bayes.** Journal of the American Statistical Association, 2019.