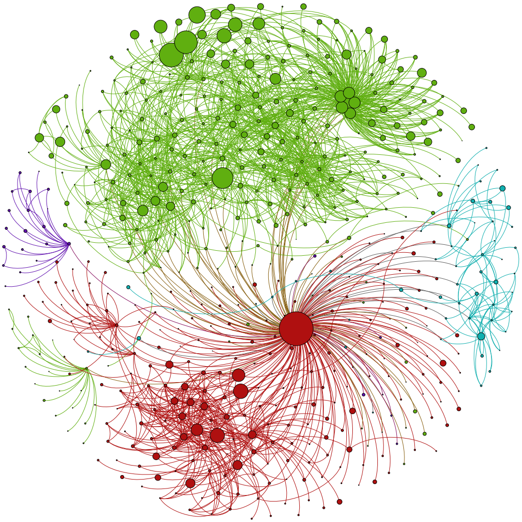
SCALING AND GENERALIZING APPROXIMATE BAYESIAN INFERENCE

David M. Blei
Columbia University

This talk is about how to discover

hidden patterns in large high-dimensional data sets.



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

[Gopalan and Blei, PNAS 2013]

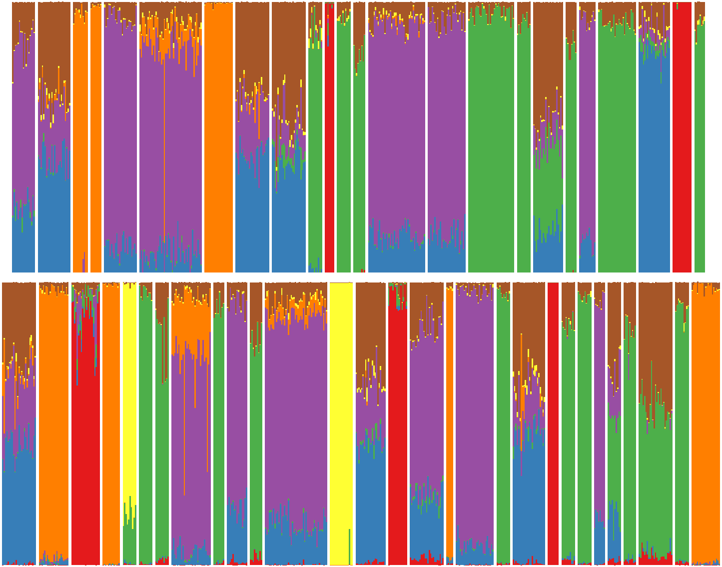
1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

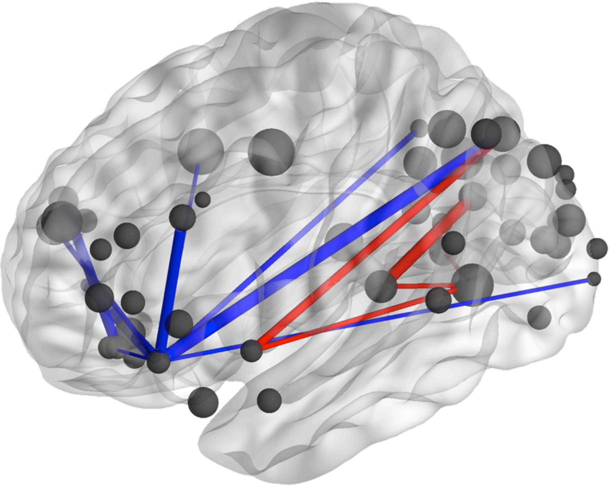
Topics found in 1.8M articles from the New York Times

[Hoffman, Blei, Wang, Paisley, JMLR 2013]



Population analysis of 2 billion genetic measurements

[Gopalan, Hao, Blei, Storey, to appear]



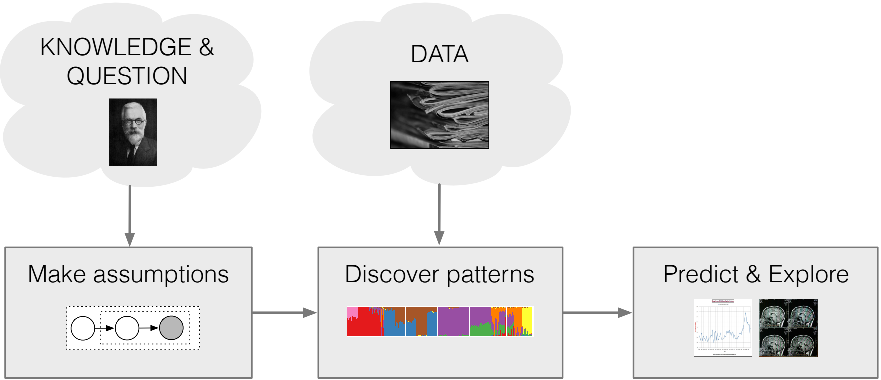
Neuroscience analysis of 220 million fMRI measurements

[Manning et al., PLOS ONE 2014]

Analysis of 1.7M taxi trajectories, in Stan

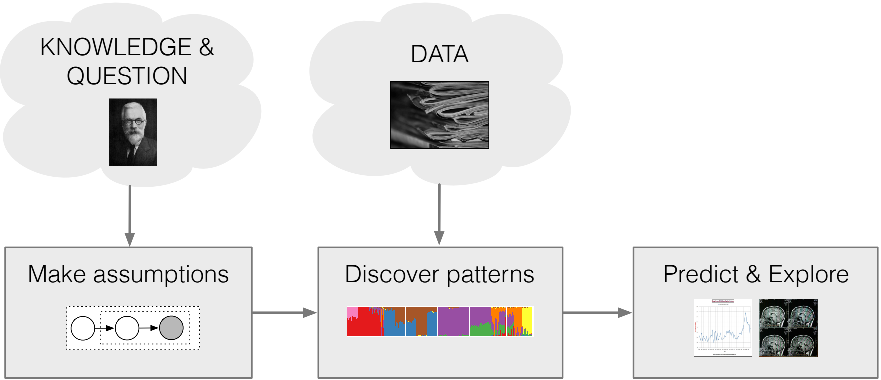
[Kucukelbir et al., 2016]

The probabilistic pipeline



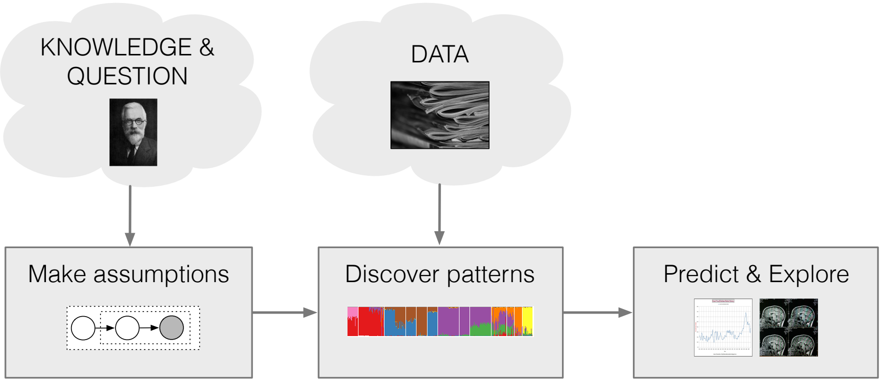
- Customized data analysis is important to many fields.
- Pipeline separates **assumptions**, **computation**, **application**
- Eases collaborative solutions to statistics problems

The probabilistic pipeline



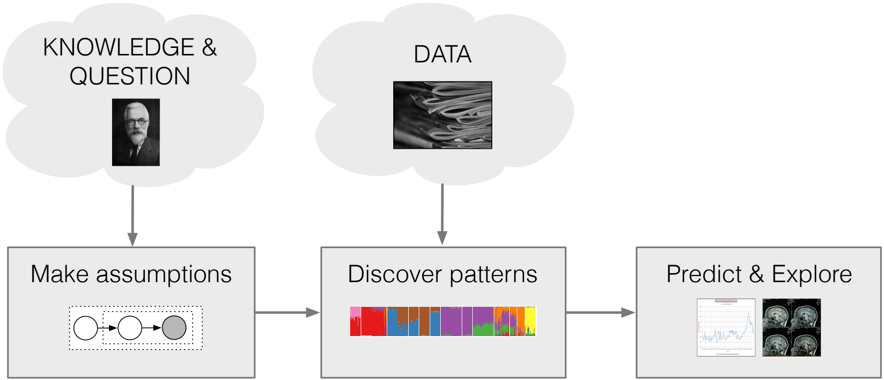
- **Inference** is the key algorithmic problem.
- Answers the question: What does this model say about this data?
- Our goal: **General** and **scalable** approaches to inference

The probabilistic pipeline



- Variational methods: inference as optimization [Jordan et al., 1999]
- Scale up with **stochastic variational inference** (SVI) [Hoffman et al., 2013]
- Generalize with **black box variational inference** (BBVI) [Ranganath et al., 2014]

The probabilistic pipeline

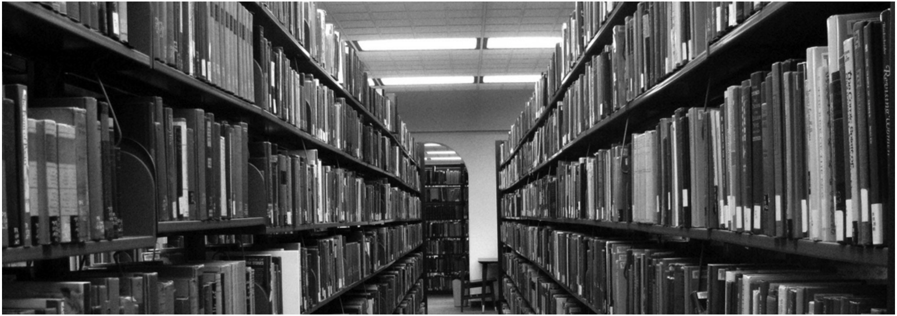


- Both approaches use **stochastic optimization**.
- SVI subsamples from a massive data set
- BBVI uses Monte Carlo to approximate difficult-to-compute expectations

STOCHASTIC VARIATIONAL INFERENCE

(with Matt Hoffman, Chong Wang, John Paisley)

Stochastic variational inference is an algorithm that
scales general Bayesian computation to massive data.



Motivation: Topic Modeling

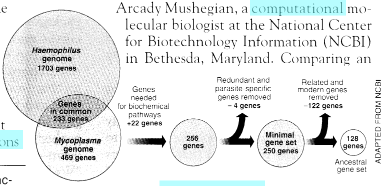
1. **Discover** the thematic structure in a large collection of documents
2. **Annotate** the documents
3. **Use** the annotations to visualize, organize, summarize, ...

Example: Latent Dirichlet allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



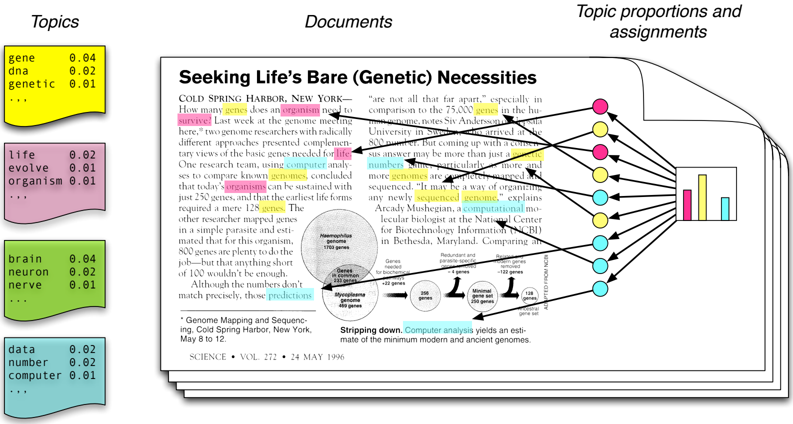
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

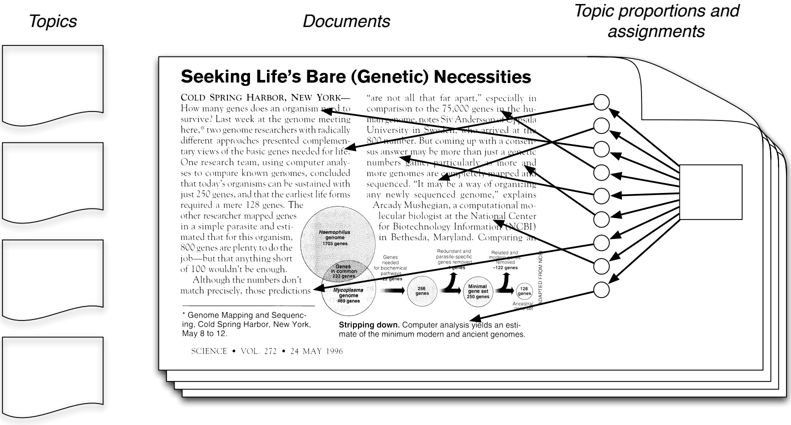
Documents exhibit multiple topics.

Example: Latent Dirichlet allocation (LDA)



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

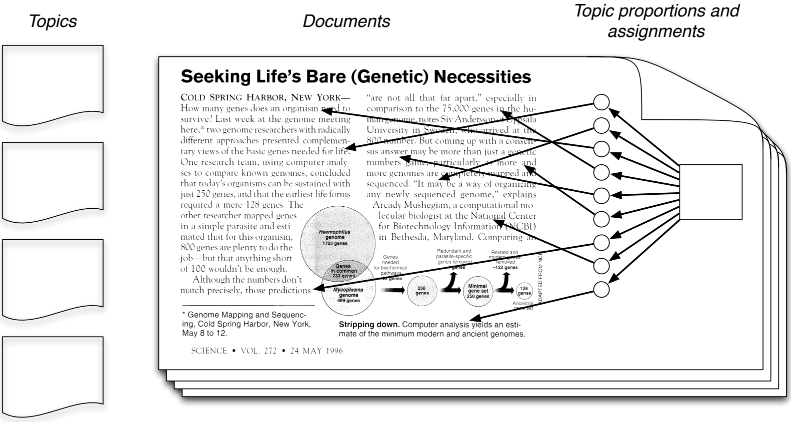
Example: Latent Dirichlet allocation (LDA)



- But we only observe the documents; the other structure is hidden.
- We compute the posterior

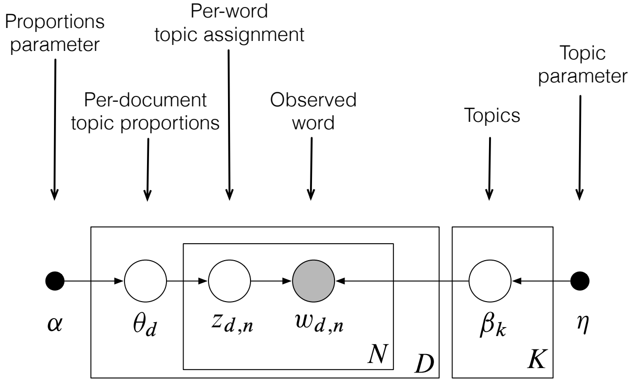
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Example: Latent Dirichlet allocation (LDA)



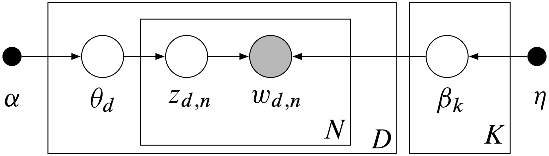
- Many data sets contain millions of documents.
- This requires inference about billions of variables.
- SVI can scale to these data.

LDA as a graphical model



- Encodes **assumptions** about data with a factorization of the joint
- Connects assumptions to **algorithms** for computing with data
- Defines the **posterior** (through the joint)

Posterior inference

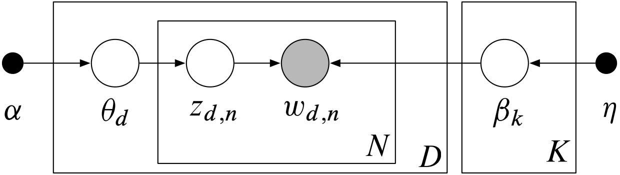


- The posterior of the latent variables given the documents is

$$p(\beta, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_{\boldsymbol{\beta}} \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}.$$

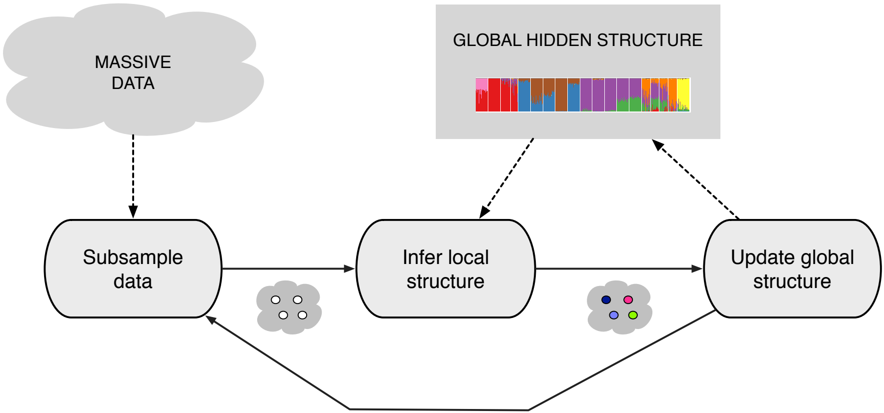
- We can't compute the denominator, the marginal $p(\mathbf{w})$.
- We use approximate inference.

Classical variational inference

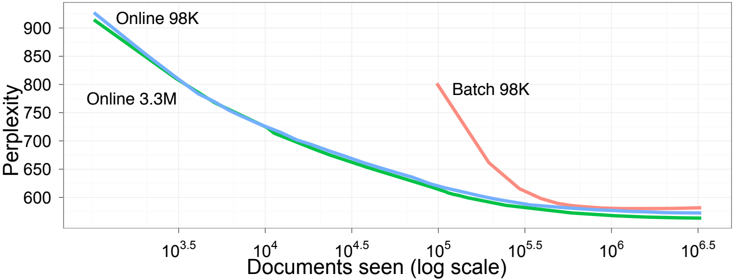


- Originally, we used variational methods to fit this model. [Blei et al., 2003]
- Classical variational inference is inefficient:
  - Do some local computation *for each data point*.
  - Aggregate these computations to re-estimate global structure.
  - Repeat.
- This cannot handle massive data.

Stochastic variational inference



Stochastic variational inference



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

[Hoffman et al., 2010]

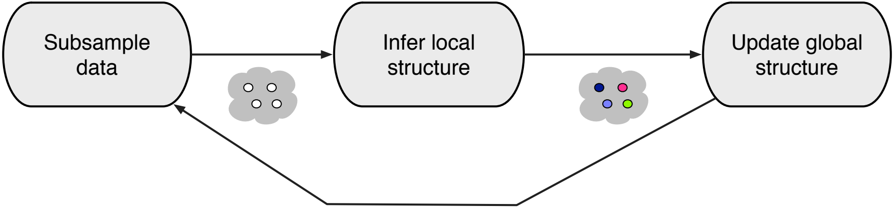
1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

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

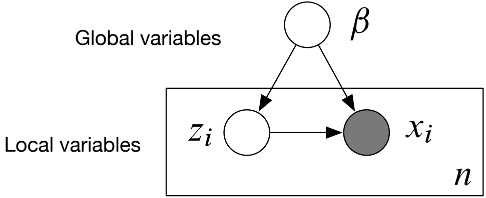
Stochastic variational inference



Next parts of this talk:

1. Define a generic class of models
2. Derive classical mean-field variational inference
3. Use the classical algorithm to derive stochastic variational inference

A generic class of models

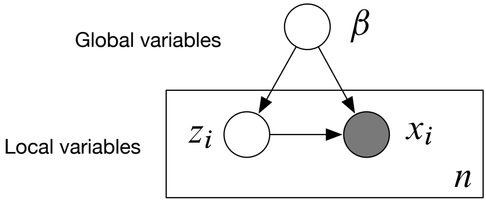


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- 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.

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

A generic class of models

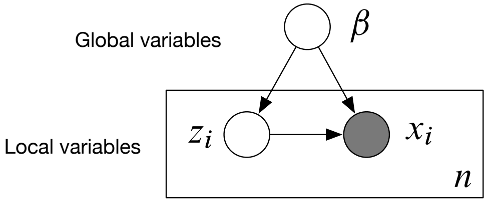


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variables.
- Assume each complete conditional is in the exponential family,

$$p(z_i \mid \beta, x_i) = h(z_i) \exp\{\eta_\ell(\beta, x_i)^\top z_i - a(\eta_\ell(\beta, x_i))\}$$
$$p(\beta \mid \mathbf{z}, \mathbf{x}) = h(\beta) \exp\{\eta_g(\mathbf{z}, \mathbf{x})^\top \beta - a(\eta_g(\mathbf{z}, \mathbf{x}))\}.$$

A generic class of models



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

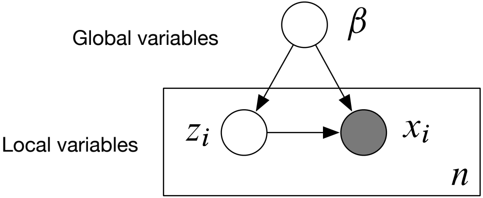
- A **complete conditional** is the conditional of a latent variable given the observations and other latent variable.

- The global parameter comes from conjugacy [Bernardo and Smith, 1994]

$$\eta_g(\mathbf{z}, \mathbf{x}) = \alpha + \textstyle\sum_{i=1}^{n} t(z_i, x_i),$$

where α is a hyperparameter and $t(\cdot)$ are sufficient statistics for $[z_i, x_i]$.

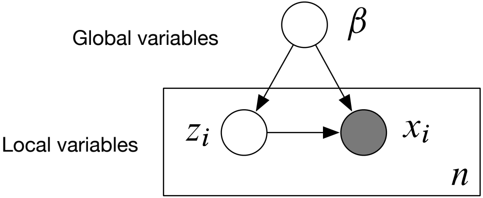
A generic class of models



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

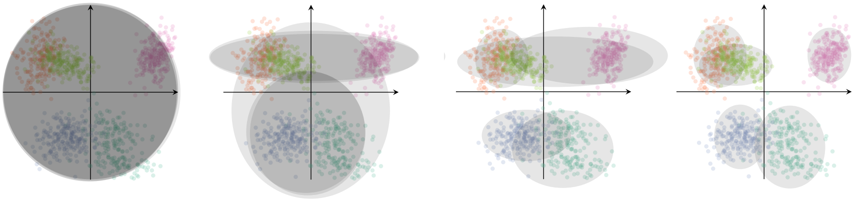
- **Dirichlet process mixture** with conjugate base measure [Sethuraman, 1994]
  - Local: mixture assignments for each observation (categorical).
  - Global: stick lengths (beta) and mixture components (e.g., Gaussian).
- Other BNP models, e.g.,
  - Beta-Bernoulli process
  - Hierarchical Dirichlet process

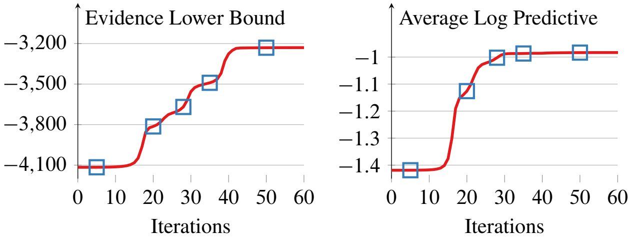
A generic class of models



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

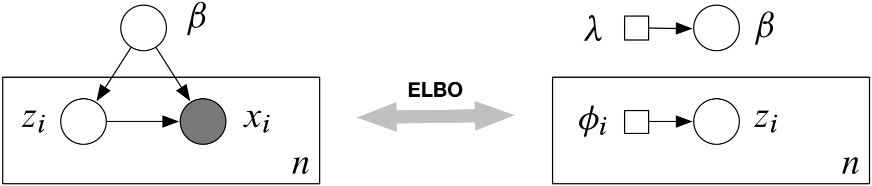
- Bayesian mixture models
- Time series models
  (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization
  (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression
  (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models
  (LDA and some variants)





- **Variational methods** turn *inference into optimization*
- Idea: Fit a simple distribution to be close (in KL) to the exact posterior
- Here: A simple mixture of Gaussians [image by Alp Kucukelbir]

Mean-field variational inference



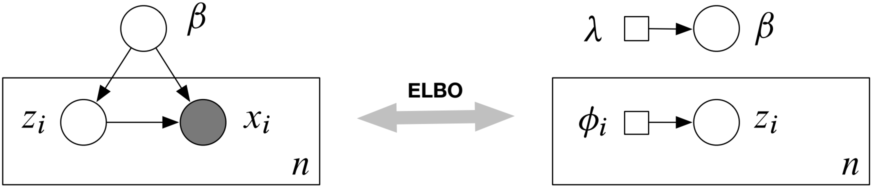
- Goal: Minimize KL divergence between a family q and the posterior p.
- Mean-field assumption: Set $q(\beta, \mathbf{z})$ to be fully factored,

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

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

$$\begin{aligned} p(\beta \mid \mathbf{z}, \mathbf{x}) &= h(\beta) \exp\{\eta_g(\mathbf{z}, \mathbf{x})^\top \beta - a(\eta_g(\mathbf{z}, \mathbf{x}))\} \\ q(\beta \mid \lambda) &= h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\}. \end{aligned}$$

Mean-field variational inference



- Optimize the evidence lower bound, equivalent to optimizing negative KL,

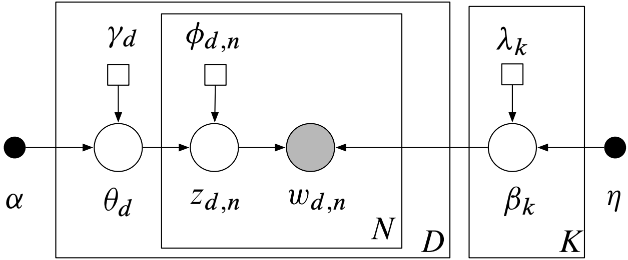
$$\mathscr{L}(\lambda, \phi_{1:n}) = \mathbb{E}_q[\log p(\beta, \mathbf{Z}, \mathbf{x})] - \mathbb{E}_q[\log q(\beta, \mathbf{Z})].$$

- Traditional VI uses coordinate ascent [Ghahramani and Beal, 2001]

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

- Iteratively update each parameter, holding others fixed.
  Notice the relationship to Gibbs sampling.

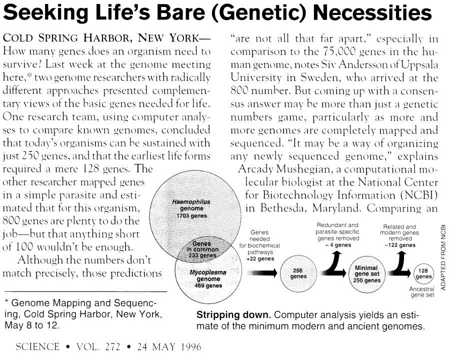
Mean-field variational inference for LDA



- The local variables are the per-document variables θ_d and $\mathbf{z}_d$.
- The global variables are the topics $\beta_1, \ldots, \beta_K$.
- The variational distribution is

$$q(\beta, \boldsymbol{\theta}, \mathbf{z}) = \prod_{k=1}^{K} q(\beta_k \mid \lambda_k) \prod_{d=1}^{D} q(\theta_d \mid \gamma_d) \prod_{n=1}^{N} q(z_{d,n} \mid \phi_{d,n})$$

Mean-field variational inference for LDA



Seeking Life's Bare (Genetic) Necessities

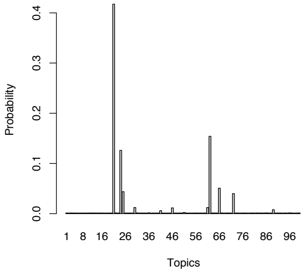
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



Mean-field variational inference for LDA

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

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

repeat

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

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

  end

  Set global parameter

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

  .

until *the ELBO has converged*

This is inefficient: We analyze all data before completing one iteration.

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]
- Has enabled modern machine learning

Natural gradients

Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program, Saitama 351-01, Japan

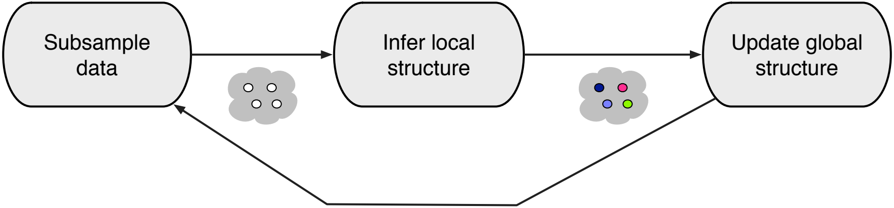
When a parameter space has a certain underlying structure, the ordinary gradient of a function does not represent its steepest direction, but the natural gradient does. Information geometry is used for calculating the natural gradients in the parameter space of perceptrons, the space of matrices (for blind source separation), and the space of linear dynamical systems (for blind source deconvolution). The dynamical behavior of natural gradient online learning is analyzed and is proved to be Fisher efficient, implying that it has asymptotically the same performance as the optimal batch estimation of parameters. This suggests that the plateau phenomenon, which appears in the backpropagation learning algorithm of multilayer perceptrons, might disappear or might not be so serious when the natural gradient is used. An adaptive method of updating the learning rate is proposed and analyzed.

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

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

- Computationally:
  - Compute coordinate updates.
  - Subtract the current variational parameters.

Stochastic variational inference



- Construct a noisy natural gradient
  - Sample a data point at random $j \sim \mathrm{Uniform}(1, \ldots, n)$.
  - Calculate
$$\tilde{\nabla}_\lambda \mathcal{L} = \alpha + n \mathbb{E}_{\phi_j^*}[t(Z_j, x_j)] - \lambda$$
- This is a good noisy gradient
  - The expectation (with respect to j) is the natural gradient.
  - Only requires the local parameters for one data point.

Stochastic variational inference

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.
Initialize λ randomly. Set ρ_t appropriately.

repeat

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

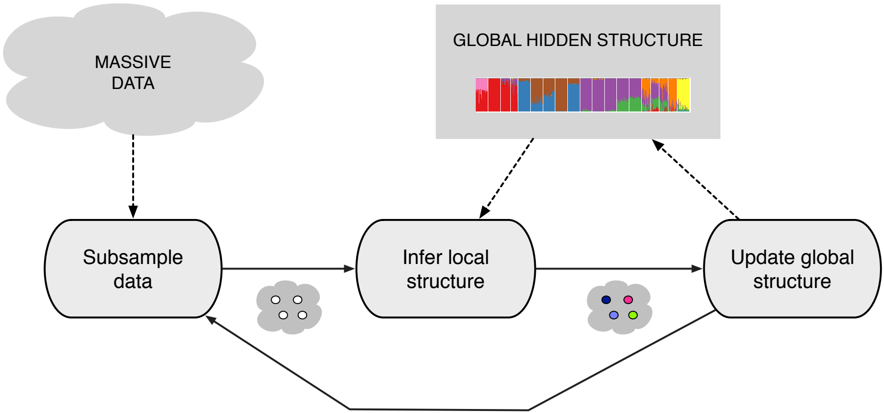
$$\hat{\lambda} = \alpha + n\mathbb{E}_\phi[t(Z_j, x_j)].$$

Set global parameter

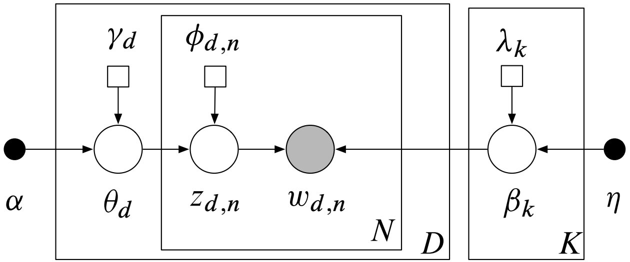
$$\lambda = (1 - \rho_t)\lambda + \rho_t\hat{\lambda}.$$

until *forever*

Stochastic variational inference

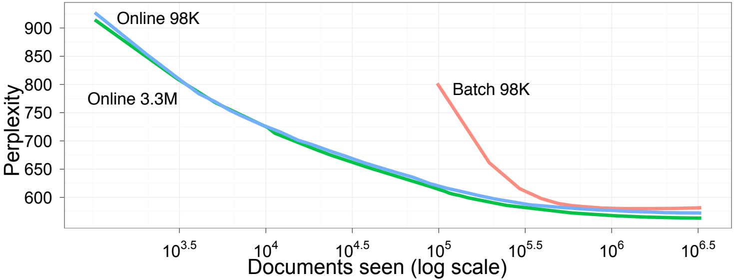


Stochastic variational inference in LDA



1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form intermediate topics from those local parameters
4. Update topics as a weighted average of intermediate and current topics

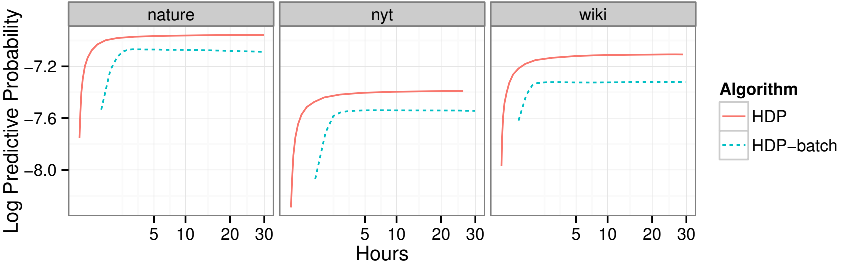
Stochastic variational inference in LDA



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 et al., 2010]

HDP topic models



- A study of large corpora with the HDP topic model [Teh et al., 2006]
- Details are in Hoffman et al., 2013.
- SVI is faster and lets us analyze more data.

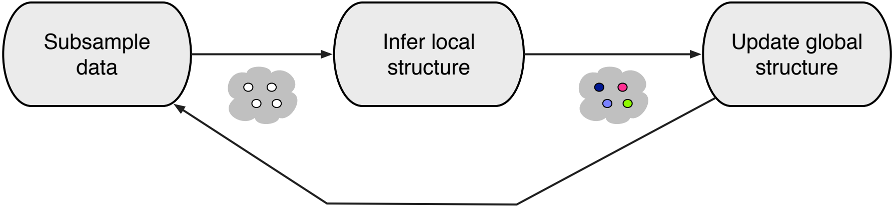
1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

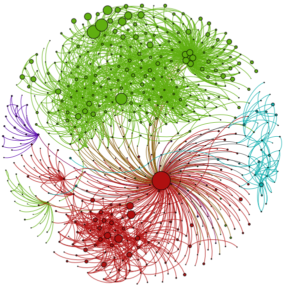
Topics using the HDP, found in 1.8M articles from the New York Times

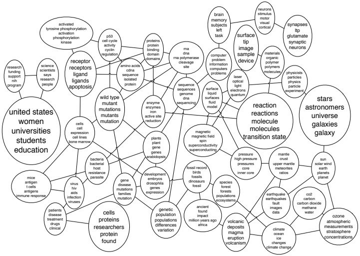
Stochastic variational inference



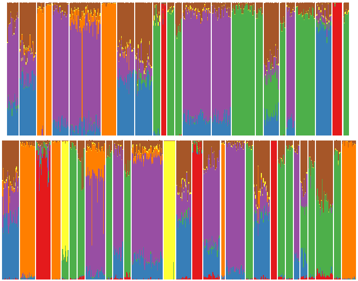
We derived an algorithm for **scalable variational inference**.

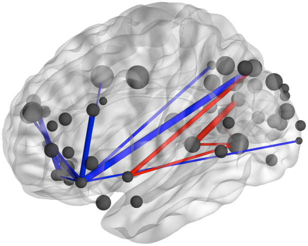
- Bayesian mixture models
- Time series models
  (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization
  (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression
  (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models
  (LDA and some variants)





activated
tyrosine phosphorylation
activation
phosphorylation
kinase
p53
cell cycle
activity
cyclin
regulation
proteins
protein
binding
domain
domains
rna
dna
rna polymerase
cleavage
site
research
funding
support
nih
program
science
scientists
says
research
people
receptor
receptors
ligand
ligands
apoptosis
amino acids
cdna
sequence
isolated
protein
sequence
sequences
genome
dna
sequencing
united states
women
universities
students
education
wild type
mutant
mutations
mutants
mutation
enzyme
enzymes
iron
active site
reduction
cells
cell
expression
cell lines
bone marrow
plants
plant
gene
genes
arabidopsis
bacteria
bacterial
host
resistance
parasite
mice
antigen
t cells
antigens
immune response
virus
hiv
aids
infection
viruses
gene
disease
mutations
families
mutation
development
embryos
drosophila
genes
expression
patients
disease
treatment
drugs
clinical
cells
proteins
researchers
protein
found
genetic
population
populations
differences
variation
brain
memory
subjects
left
task
neurons
stimulus
motor
visual
cortical
synapses
ltp
glutamate
synaptic
neurons
surface
tip
image
sample
device
computer
problem
information
computers
problems
materials
organic
polymer
polymers
molecules
laser
optical
light
electrons
quantum
physicists
particles
physics
particle
experiment
surface
liquid
surfaces
fluid
model
reaction
reactions
molecule
molecules
transition state
stars
astronomers
universe
galaxies
galaxy
magnetic
magnetic field
spin
superconductivity
superconducting
pressure
high pressure
pressures
core
inner core
mantle
crust
upper mantle
meteorites
ratios
sun
solar wind
earth
planets
planet
fossil record
birds
fossils
dinosaurs
fossil
species
forest
forests
populations
ecosystems
earthquake
earthquakes
fault
images
data
co2
carbon
carbon dioxide
methane
water
ancient
found
impact
million years ago
africa
volcanic
deposits
magma
eruption
volcanism
climate
ocean
ice
changes
climate change
ozone
atmospheric
measurements
stratosphere
concentrations



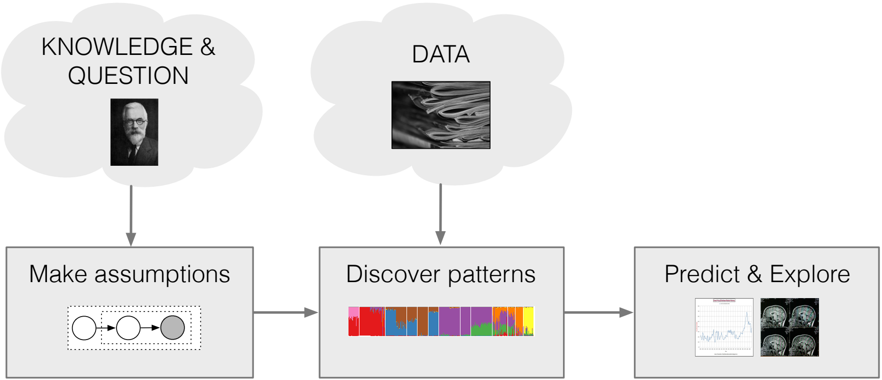


BLACK BOX VARIATIONAL INFERENCE

(with Rajesh Ranganath and Sean Gerrish)

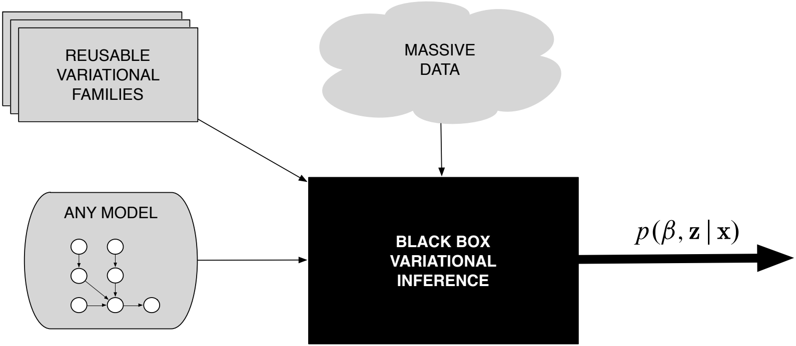
Black box variational inference is an algorithm that
efficiently performs Bayesian computation in any model.

Black box variational inference



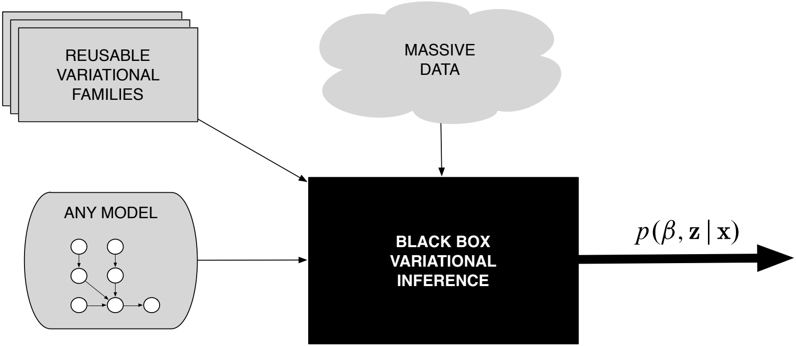
- Approximate inference can be difficult to derive.
- Especially true for models that are not conditionally conjugate
- E.g., discrete choice models, Bayesian generalized linear models, ...

Black box variational inference



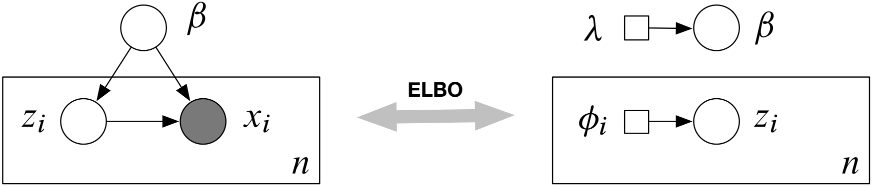
- Easily use variational inference with *any model*
- No exponential family requirements
- No mathematical work beyond specifying the model

Black box variational inference



- Sample from $q(\cdot)$
- Form noisy gradients without model-specific computation
- Use stochastic optimization

Black box variational inference



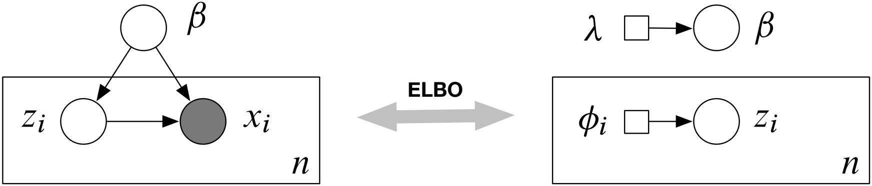
ELBO:

$$\mathcal{L}(\nu) = \mathbb{E}_q[\log p(\beta, \mathbf{Z}, \mathbf{x}) - \log q_\nu(\beta, \mathbf{Z})]$$

Shorthand:

$$\mathcal{L}(\nu) = \mathbb{E}_q[D_\nu(\beta, \mathbf{Z})]$$

Black box variational inference



A noisy gradient:

$$\nabla_\nu \mathcal{L} \approx \frac{1}{B} \sum_{b=1}^{B} \nabla_\nu \log q_\nu\left(\beta_b, \mathbf{z}_b\right) D_\nu(\beta, \mathbf{Z})$$

where

$$(\beta_b, \mathbf{z}_b) \sim q_\nu(\beta, \mathbf{z})$$

The noisy gradient

$$\boxed{\nabla_\nu \mathcal{L}(\nu) \approx \frac{1}{B} \sum_{b=1}^{B} \nabla_\nu \log q_\nu(\beta_b, z_b) D_\nu(\beta_b, z_b)}$$

- We use these gradients in a stochastic optimization algorithm.
- Requirements:
  - Sampling from $q_\nu(\beta, \mathbf{z})$
  - Evaluating $\nabla_\nu \log q_\nu(\beta, \mathbf{z})$
  - Evaluating $\log p(\beta, \mathbf{z}, \mathbf{x})$
- A "black box": We can reuse $q(\cdot)$ across models

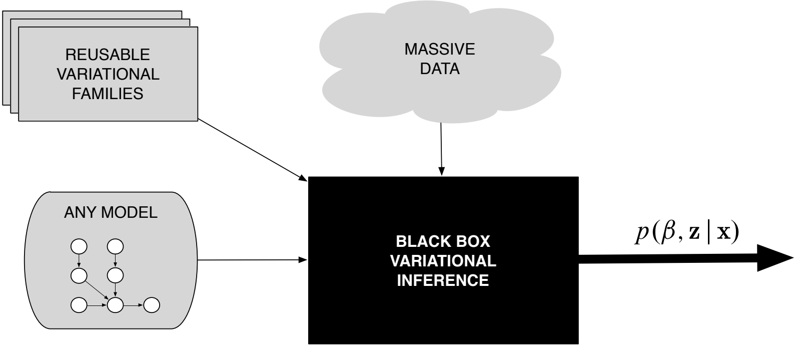
The noisy gradient

$$\boxed{\nabla_\nu \mathcal{L}(\nu) \approx \frac{1}{B}\sum_{b=1}^{B} \nabla_\nu \log q_\nu(\beta_b, z_b) D_\nu(\beta_b, z_b)}$$

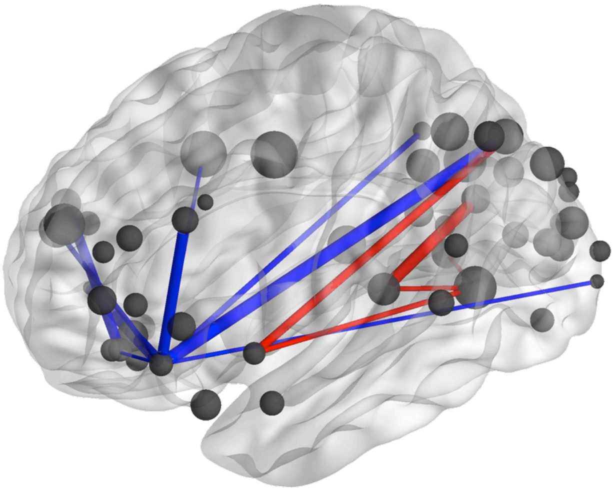
Making it work:

- Rao-Blackwellization for each component of the gradient
- Control variates, again using $\nabla_\nu \log q_\nu(\beta, z)$
- AdaGrad, for setting learning rates [Duchi, Hazan, Singer, 2011]
- Stochastic variational inference, for handling massive data

Monte Carlo Gradients of the ELBO

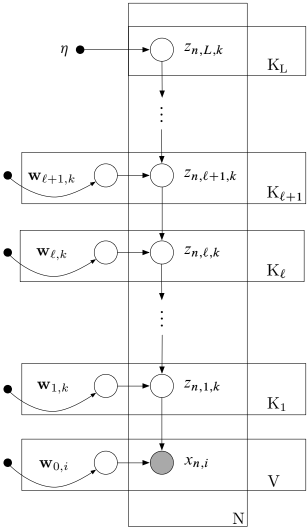


- MC gradient [Ji et al., 2010; Nott et al., 2012; Paisley et al., 2012; Ranganath et al. 2014]
- Autoencoders [Kingma and Welling, 2013/2014]
- Neural networks [Kingma et al., 2015; Mnih and Gregor, 2014; Rezende et al., 2014]
- A perspective from regression [Salimans and Knowles, 2014]
- Doubly stochastic VB [Titsias and Lazaro-Gredilla, 2014]



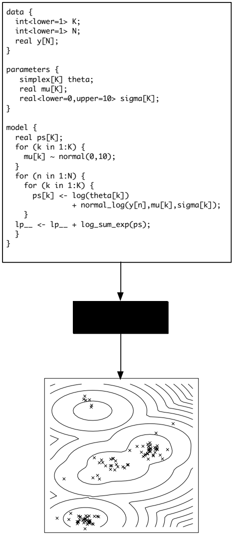
Neuroscience analysis of 220 million fMRI measurements

[Manning et al., PLOS ONE 2014]



Deep Exponential Families

[Ranganath et al., AISTATS 2015]



Probabilistic Programming

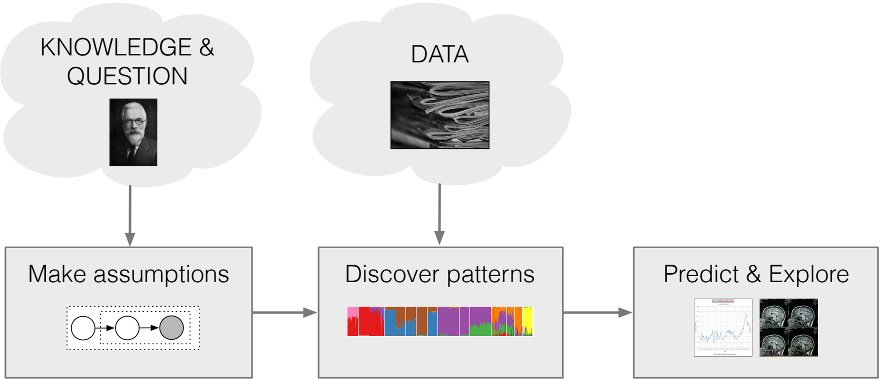
[Kucukelbir et al., 2015]

Edward: A library for probabilistic modeling, inference, and criticism

`github.com/blei-lab/edward`

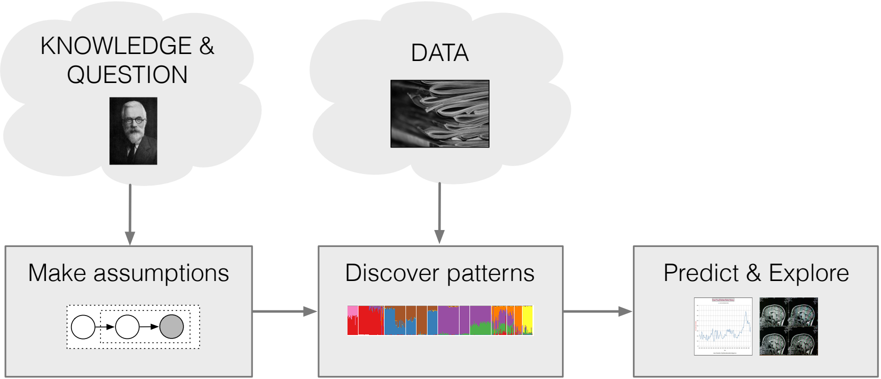
(lead by Dustin Tran)

The probabilistic pipeline

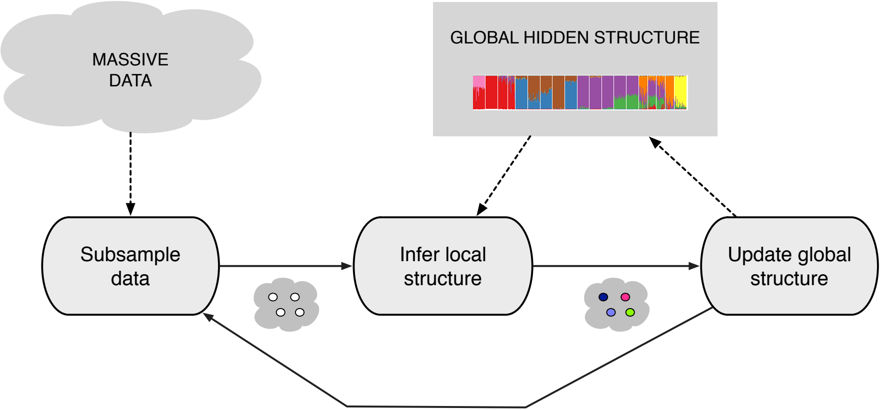


- Customized data analysis is important to many fields.
- Pipeline separates **assumptions**, **computation**, **application**
- Eases collaborative solutions to statistics problems

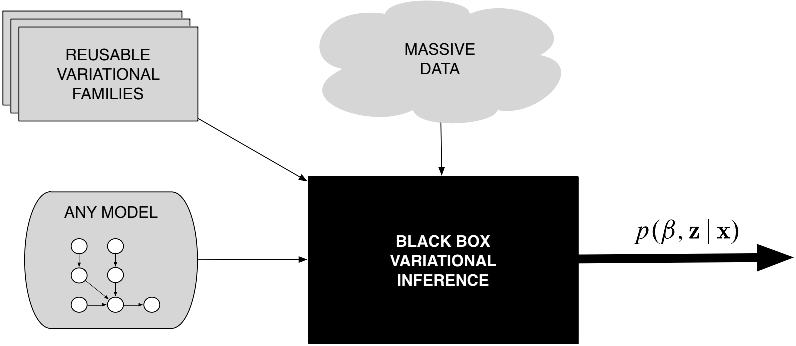
The probabilistic pipeline



- **Inference** is the key algorithmic problem.
- Answers the question: What does this model say about this data?
- Our goal: **General** and **scalable** approaches to inference



“Stochastic variational inference” [Hoffman et al., 2013, JMLR]



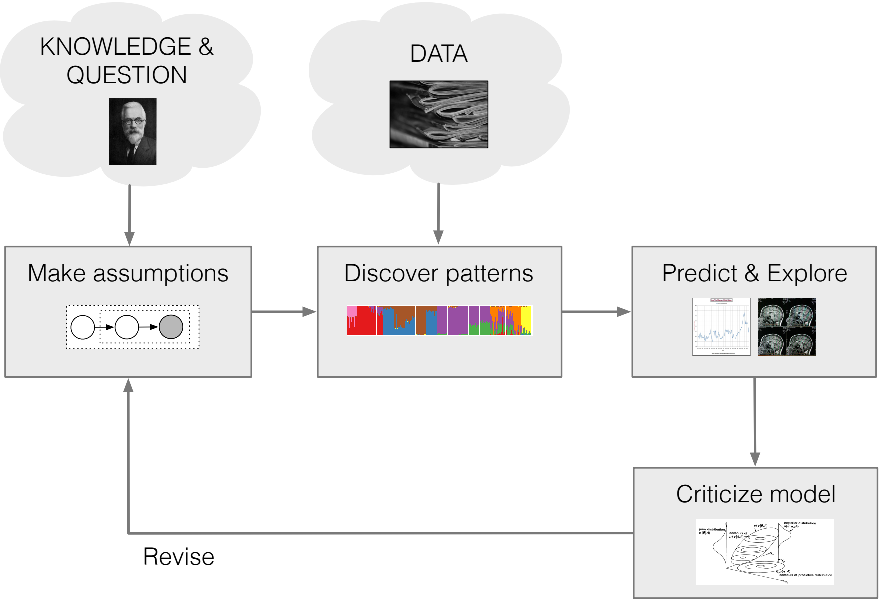
"Black box variational inference" [Ranganath et al., 2014, AISTATS]

Recent research (on the ArXiv):

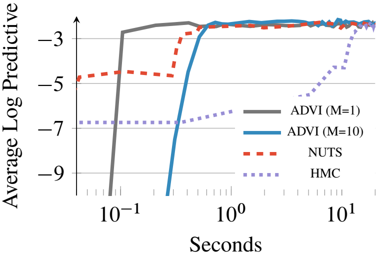
- "Variational Inference: A Review for Statisticians" (with J. McAuliffe and A. Kucukelbir)
- "Hierarchical Variational Models" (with R. Ranganath and D. Tran)
- "Automatic Differentiation Variational Inference" (with A. Kucukelbir, D. Tran, R. Ranganath, and A. Gelman)

Open and current research:

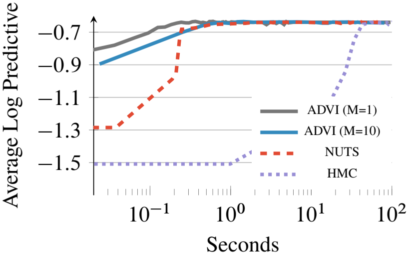
- How can we improve the gradient estimator in BBVI?
- Can we use alternative divergence measures? What are their properties?
- What are the statistical properties of VI? What is a good framework?
- How can we efficiently go beyond the mean field?


KNOWLEDGE &
QUESTION
DATA
Make assumptions
Discover patterns
Predict & Explore
Criticize model
Revise

Should I be skeptical about variational inference?



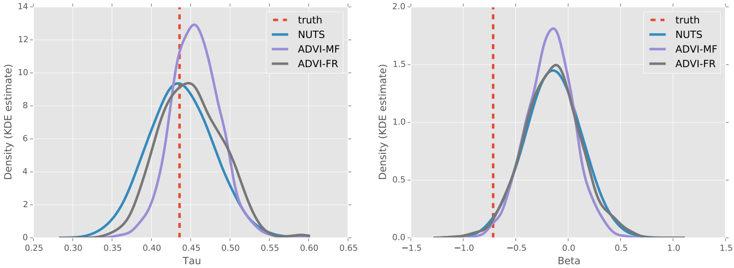
(a) Linear Regression with ARD



(b) Hierarchical Logistic Regression

- **MCMC enjoys theoretical guarantees.**
- But they usually get to the same place. [Kucukelbir et al., submitted]
- We need more theory about variational inference.

Should I be skeptical about variational inference?



- **Variational inference underestimates the variance of the posterior.**
- Relaxing the mean-field assumption can help.
- Here: A Poisson GLM [Giordano et al., 2015]