UMBC

AN HONORS UNIVERSITY IN MARYLAND


# Mixed Membership Word Embeddings for Computational Social Science

**James Foulds (Jimmy)**

Department of Information Systems
University of Maryland, Baltimore County

*UMBC ACM Faculty Talk, April 5 2018*

*Paper to be presented at the International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*

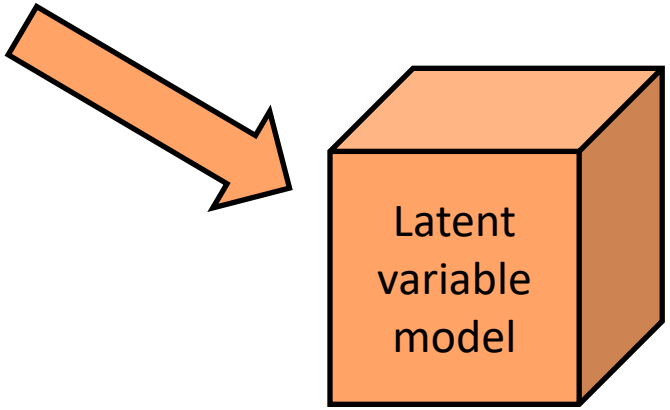
# Latent Variable Modeling


Complicated, noisy,
high-dimensional

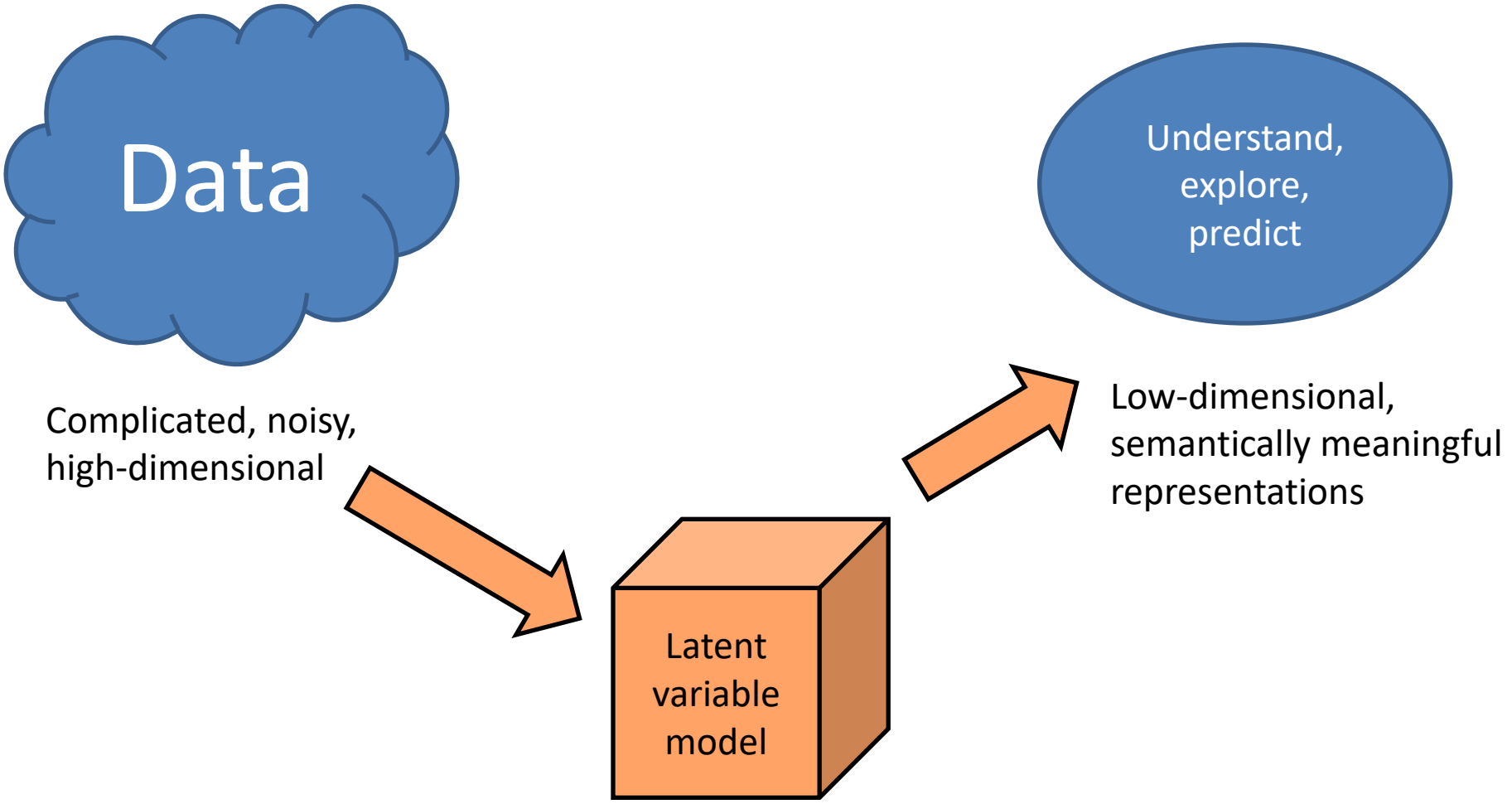
# Latent Variable Modeling


Understand,
explore,
predict

Complicated, noisy,
high-dimensional



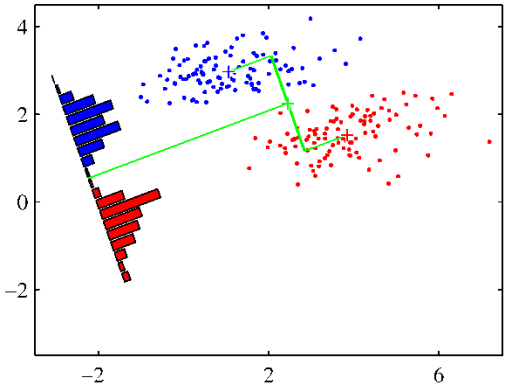
# Latent Variable Modeling



# Latent Variable Modeling

- Latent variable modeling is a general, principled approach for making sense of complex data sets

- Core principles:
  - Dimensionality reduction



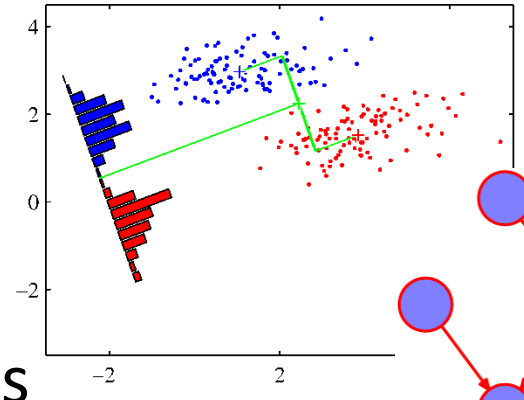


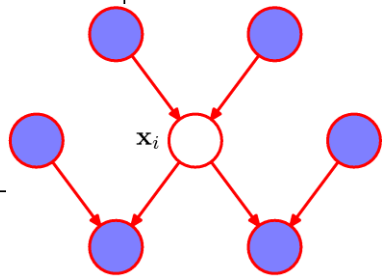

Images due to Chris Bishop, Pattern Recognition and Machine Learning book

# Latent Variable Modeling

- Latent variable modeling is a general, principled approach for making sense of complex data sets

- Core principles:
  - Dimensionality reduction
  - Probabilistic graphical models



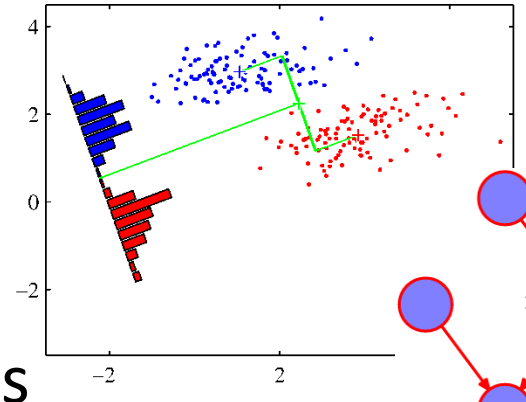



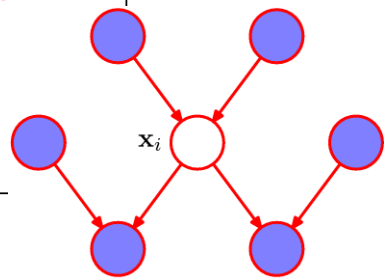



Images due to Chris Bishop, Pattern Recognition and Machine Learning book

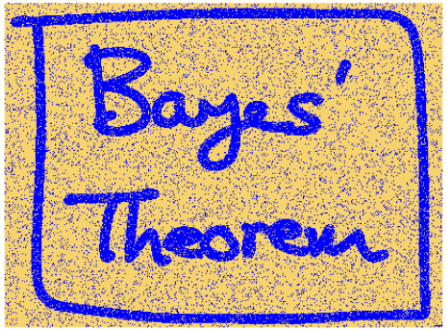
# Latent Variable Modeling

- Latent variable modeling is a general, principled approach for making sense of complex data sets



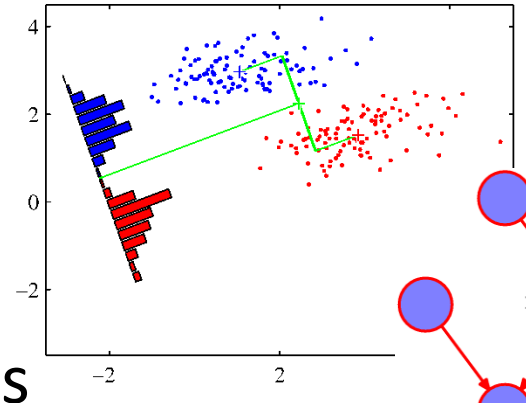

- Core principles:
  - Dimensionality reduction
  - Probabilistic graphical models
  - Statistical inference, especially Bayesian inference

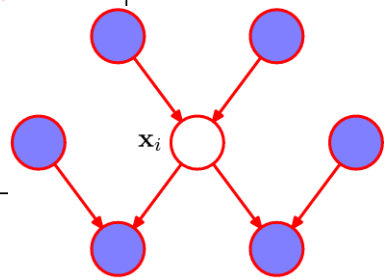



Images due to Chris Bishop, Pattern Recognition and Machine Learning book

# Latent Variable Modeling

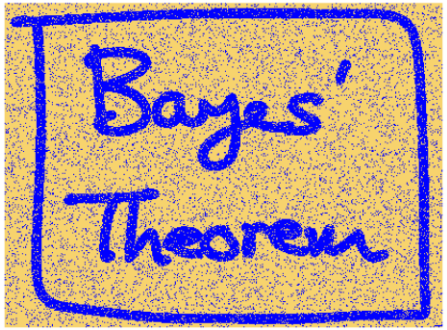
- Latent variable modeling is a general, principled approach for making sense of complex data sets




- Core principles:
  - Dimensionality reduction
  - Probabilistic graphical models
  - Statistical inference, especially Bayesian inference

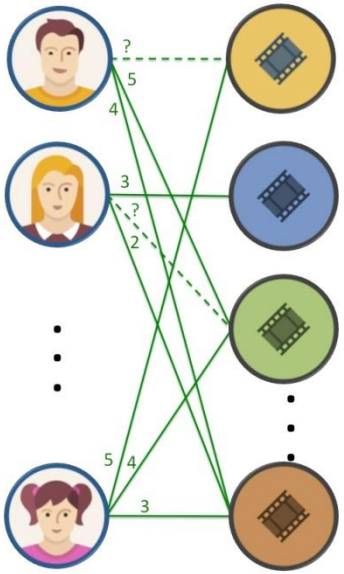
**Latent variable models are, basically, PCA on steroids!**




Images due to Chris Bishop, Pattern Recognition and Machine Learning book

# Motivating Applications

- Industry:
  - user modeling, recommender systems, and personalization, …




# Motivating Applications

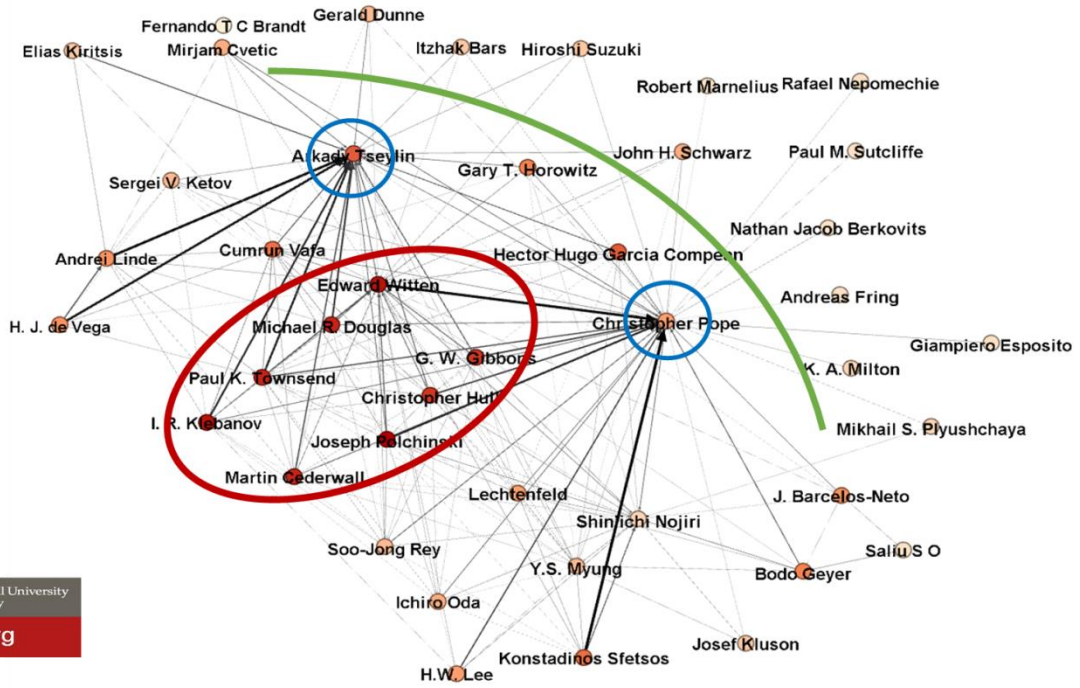
- Natural language processing
  - Machine translation
  - Document summarization
  - Parsing
  - Question answering
  - Named entity recognition
  - Sentiment analysis
  - Opinion mining

# Motivating Applications

- Furthering scientific understanding in:
  - Cognitive psychology (Griffiths and Tenenbaum, 2006)
  - Sociology (Hoff, 2008)
  - Political science (Gerrish and Blei, 2012)
  - The humanities (Mimno, 2012)
  - Genetics (Pritchard, 2000)
  - Climate science (Bain et al., 2011)
  - …

# Motivating Applications

- Social network analysis
  - Identify latent social groups/communities
  - Test sociological theories (homophily, stochastic equivalence, triadic closure, balance theory,…)




Cornell University Library
arXiv.org

# Motivating Applications

- Computational social science, digital humanities, …

**Neil deGrasse Tyson** @neiltyson · Feb 5

In science, when human behavior enters the equation, things go nonlinear. That's why Physics is easy and Sociology is hard.

# Example: Mining Classics Journals

## Computational Historiography: Data Mining in a Century of Classics Journals

DAVID MIMNO, Princeton University

---

More than a century of modern Classical scholarship has created a vast archive of journal publications that is now becoming available online. Most of this work currently receives little, if any, attention. The collection is too large to be read by any single person and mostly not of sufficient interest to warrant traditional close reading. This article presents computational methods for identifying patterns and testing hypotheses about Classics as a field. Such tools can help organize large collections, introduce younger scholars to the history of the field, and act as a "survey," identifying anomalies that can be explored using more traditional methods.

Categories and Subject Descriptors: H.4.0 [**Information Systems Applications**]: General

General Terms: Experimentation

**ACM Reference Format:**

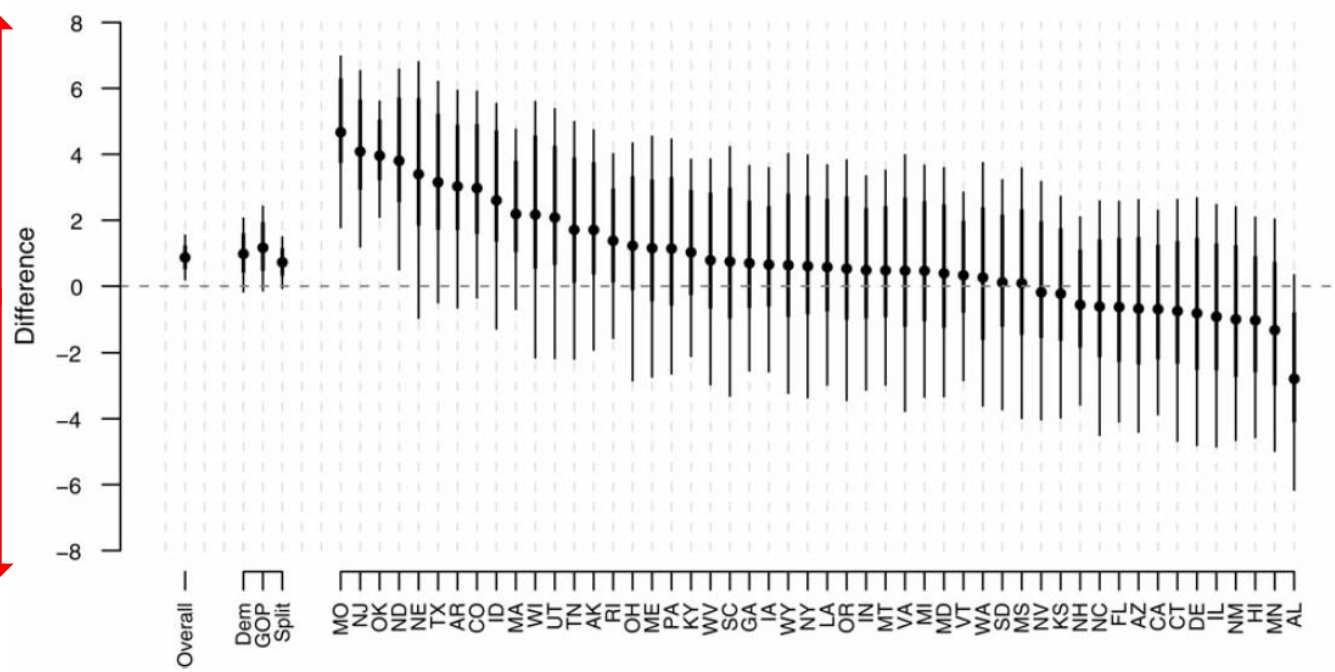
Mimno, D. 2012. Computational historiography: Data mining in a century of classics journals. ACM J. Comput. Cult. Herit. 5, 1, Article 3 (April 2012), 19 pages.
DOI = 10.1145/2160165.2160168 http://doi.acm.org/10.1145/2160165.2160168

---

# Example: Do U.S. Senators from the same state prioritize different issues? (Grimmer, 2010)

Schiller's theory is false

Schiller's theory is true




"This shows that contrary to the prediction's from Schiller's (2000) theory, senators from the same state emphasize a *more similar set of priorities than senators who represent different states*, in press releases from 2007."

Grimmer, J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. Political Analysis, 18(1):1–35, 2010.

# Example: Influence Relationships in the U.S. Supreme Court

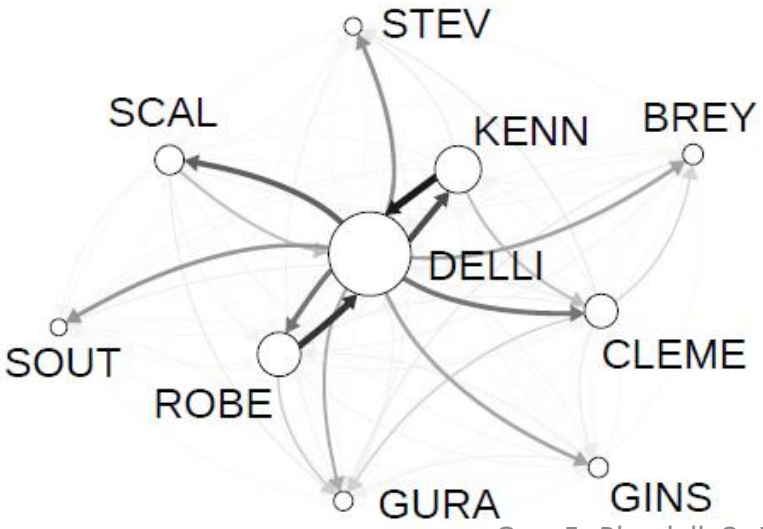
**The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation**

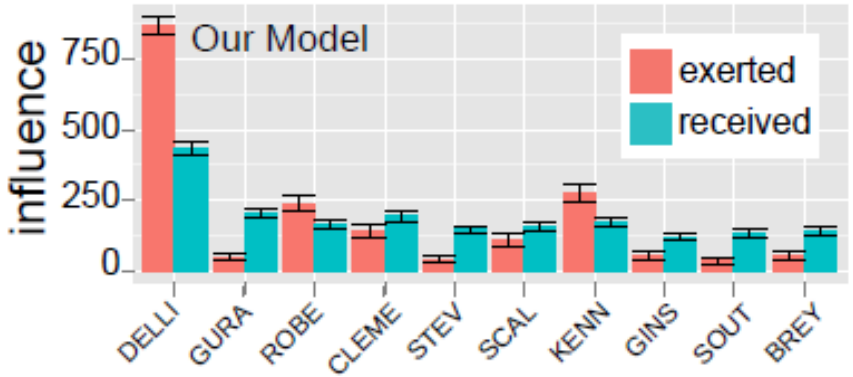

**Fangjian Guo**
Duke University
Durham, NC, USA
guo@cs.duke.edu

**Charles Blundell**
Gatsby Unit, UCL
London, UK
c.blundell@gatsby.ucl.ac.uk

**Hanna Wallach**
Microsoft Research
New York, NY, USA
wallach@microsoft.com

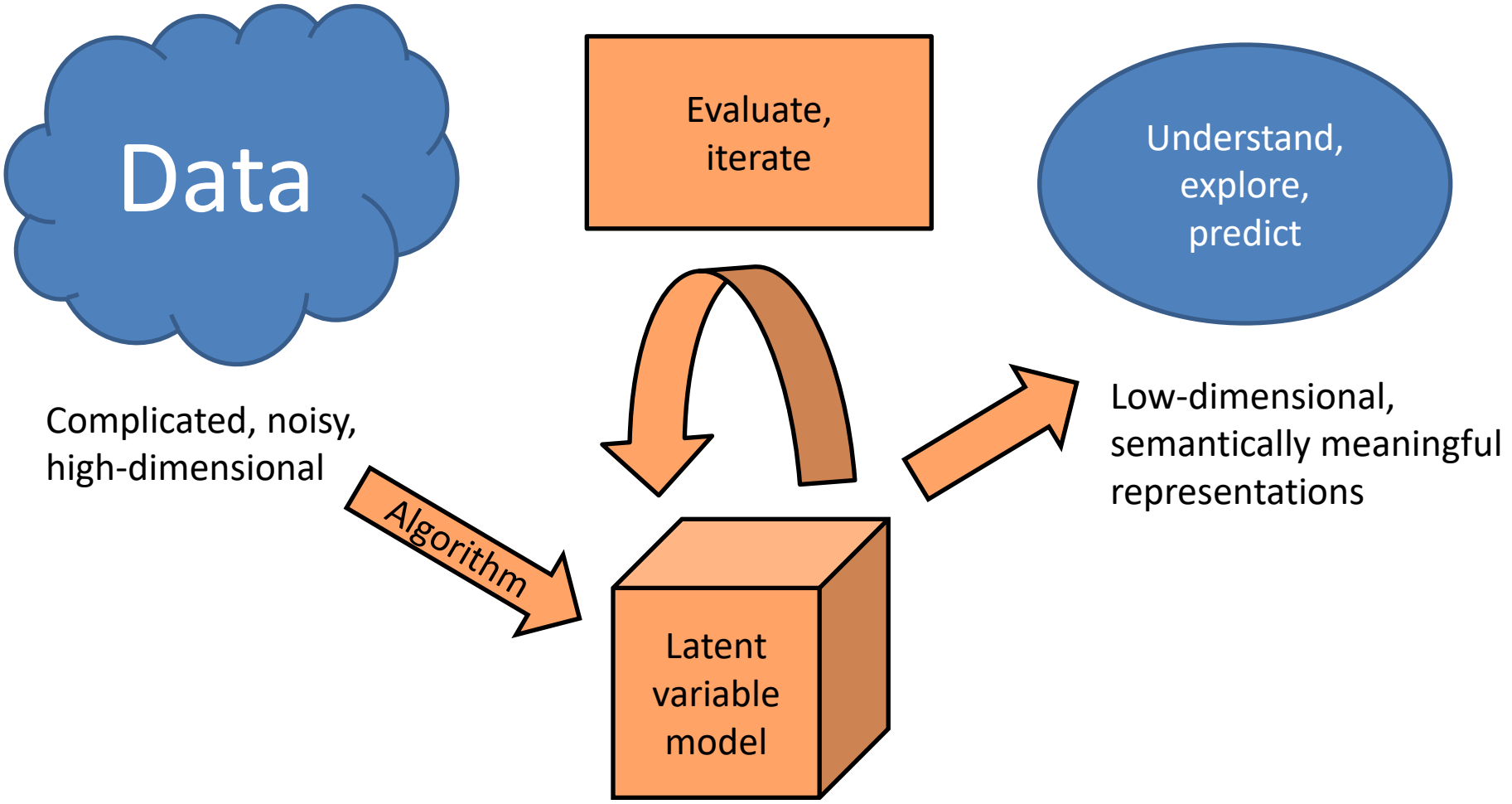
**Katherine Heller**
Duke University
Durham, NC, USA
kheller@stat.duke.edu

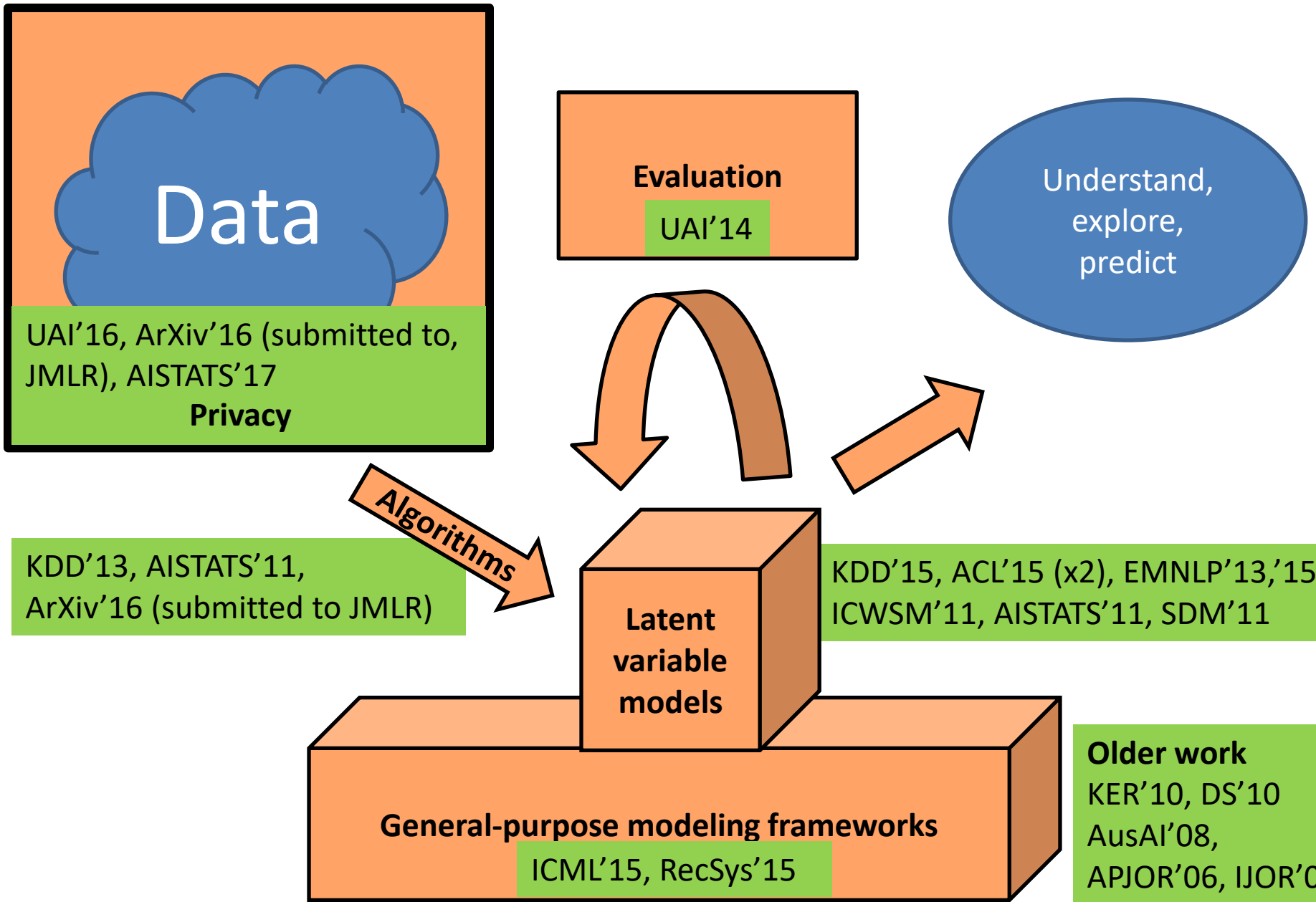




Guo, F., Blundell, C., Wallach, H., and Heller, K. (2015). AISTATS

# Box's Loop


Data
Evaluate, iterate
Understand, explore, predict
Complicated, noisy, high-dimensional
Algorithm
Latent variable model
Low-dimensional, semantically meaningful representations

# Overview of my Research



# Topic Models *(Blei et al., 2003)*

The quick brown fox jumps over the sly lazy dog

# Topic Models *(Blei et al., 2003)*

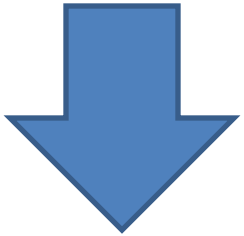
The quick brown fox jumps over the sly lazy dog

[5 6 37 1 4 30 5 22 570 12]

# Topic Models *(Blei et al., 2003)*

The quick brown fox jumps over the sly lazy dog

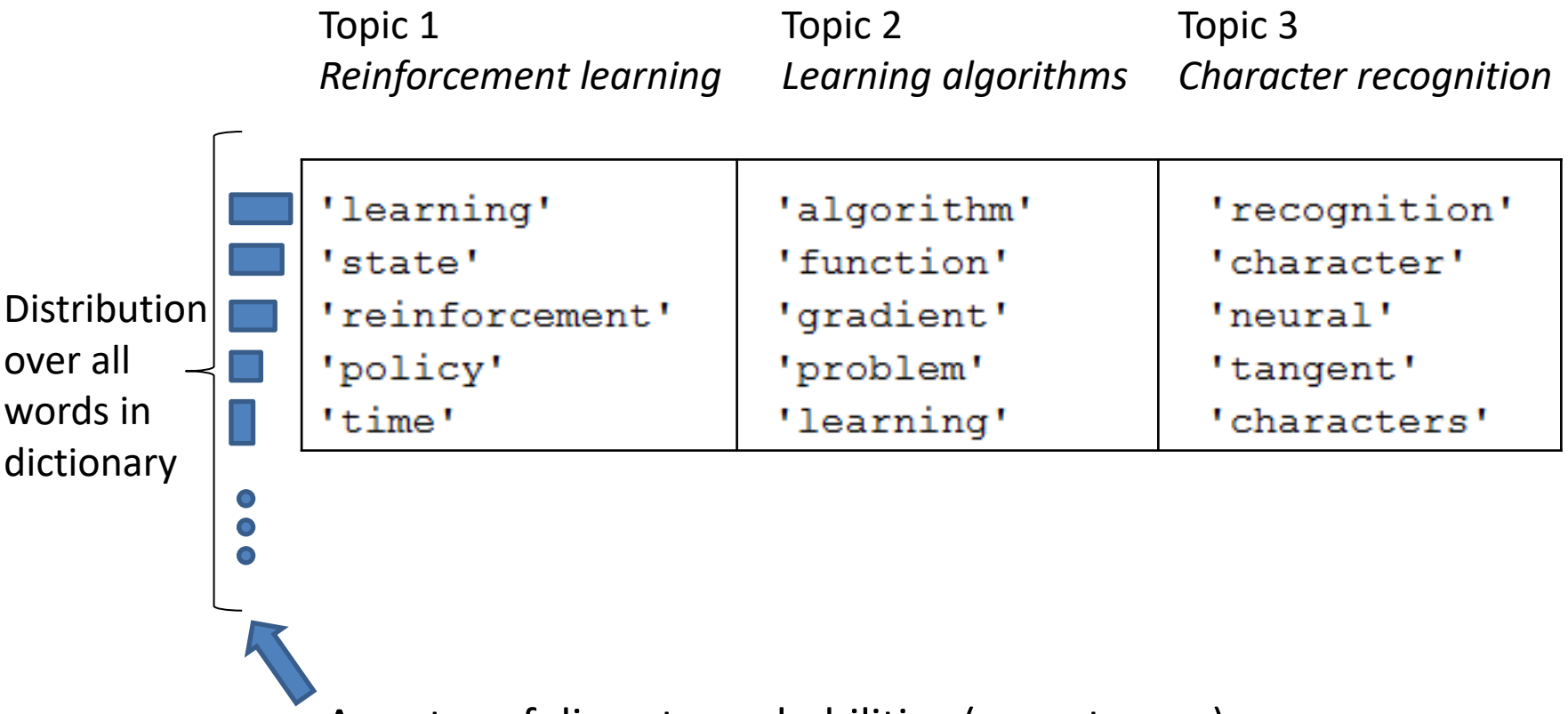
[5 6 37 1 4 30 5 22 570 12]



Foxes Dogs Jumping

[40% 40% 20% ]

# Topics


Topic 1
*Reinforcement learning*
Topic 2
*Learning algorithms*
Topic 3
*Character recognition*
'learning'
'state'
'reinforcement'
'policy'
'time'
'algorithm'
'function'
'gradient'
'problem'
'learning'
'recognition'
'character'
'neural'
'tangent'
'characters'
Distribution over all words in dictionary
A vector of discrete probabilities (sums to one)

# Topic Models for Computational Social Science

Computational Historiography: Data Mining in a Century of Classics Journals

Table IV. Selected Topics from JRS ($T = 150$), Ordered by Mean Publication Year (blue vertical line), 1911–2004

| | |
|---|---|
| fig tomb walls feet level wall tombs side plan room above house building blocks two floor small large remains |  |
| province asia roman antioch galatia minor cilicia name colonia governor cappadocia pamphylia strabo pisidia war probably studies part city |  |
| inscription stone inscriptions letters name published inscribed two above high dedication ramsay monument broken below plate monuments block |  |
| bronze museum now collection two pl glass fig british objects found head silver shape long pieces similar plate ashmolean |  |
| wall fort britain hadrian antonine roman forts stone scotland occupation work north vallum turf hill found building richmond milecastle |  |
| found roman pottery site mr coins near ware samian road small date occupation fragments hill objects iron gravel well |  |
| road river via miles valley bridge route through near modern ancient along roads map course plain line point bank |  |

time

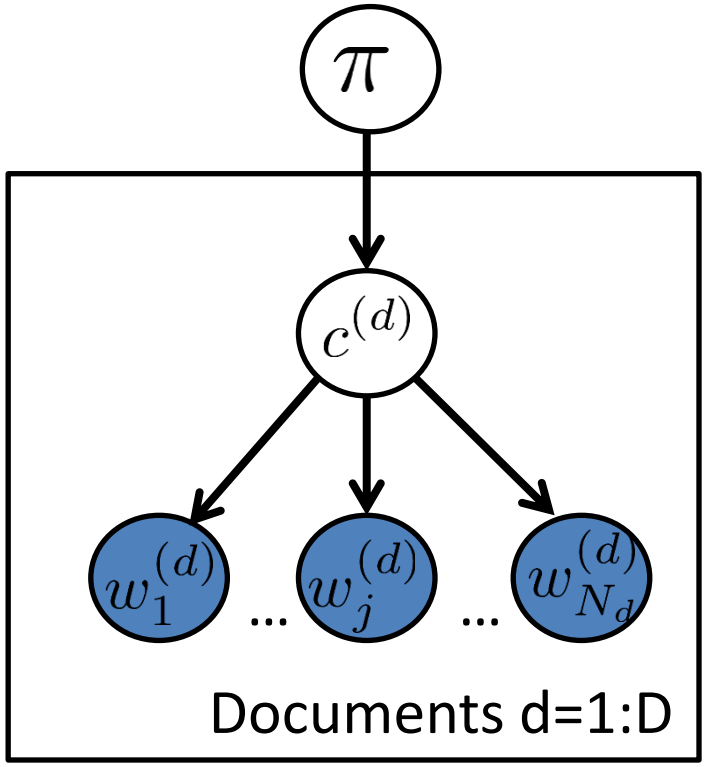
| | |
|---|---|
| cicero de philosophy philosophical porphyry stoic plato divination dreams philosopher life oracles work aristotle philosophers plutarch against divine |  |
| marriage women woman wife husband married yadin male bride girl daughter children sexual man law female mother augustine legal |  |
| horace poem poetry poet odes plancus poets poems catullus maecenas verse life literary lines ode nisbet old himself greek |  |
| per population cent total data figures age ooo high number figure rate average mortality italy model roman million empire |  |
| social political roman society public cultural within role status context power individual particular terms world traditional elite through culture |  |
| ovid propertius poet poetry poem love book cynthia elegy poems epic metamorphoses poetic elegiac literary prop gallus lover puella |  |

# Naïve Bayes Document Model

Assumed generative process:

- For each document $d$
  - Draw document's class, $c^{(d)} \sim \text{discrete}(\pi)$
  - For each word $i$ in document $d$
    - Sample the word, $w_i^{(d)} \sim \text{discrete}(\phi^{(c^{(d)})})$

Graphical model:



# Mixed Membership Modeling

- Naïve Bayes conditional independence assumption typically too strong, not realistic

- Mixed membership: relax “hard clustering” assumption to “soft clustering”
  - Membership distribution over clusters

  - E.g.:
    - Text documents belong to a distribution of topics
    - Social network individuals belong partly to multiple communities
    - Our genes come from multiple different ancestral populations

# Mixed Membership Modeling

- **Improves representational power**
  for a fixed number of topics/clusters
  - We can have a powerful model with **fewer clusters**

- **Parameter sharing**
  - Can learn on smaller datasets, especially with Bayesian approach to manage uncertainty in cluster assignments

# Topic Model Latent Representations

- **Unsupervised naïve Bayes** (latent class model)

| | Foxes | Dogs | Jumping |
|---|---|---|---|
| Doc 1 | 1 | | |
| Doc 2 | | | 1 |
| Doc 3 | | 1 | |

- **Topic model** (mixed membership model)

| | Foxes | Dogs | Jumping |
|---|---|---|---|
| Doc 1 | 0.4 | 0.4 | 0.2 |
| Doc 2 | 0.5 | 0.5 | |
| Doc 3 | 0.1 | | 0.9 |

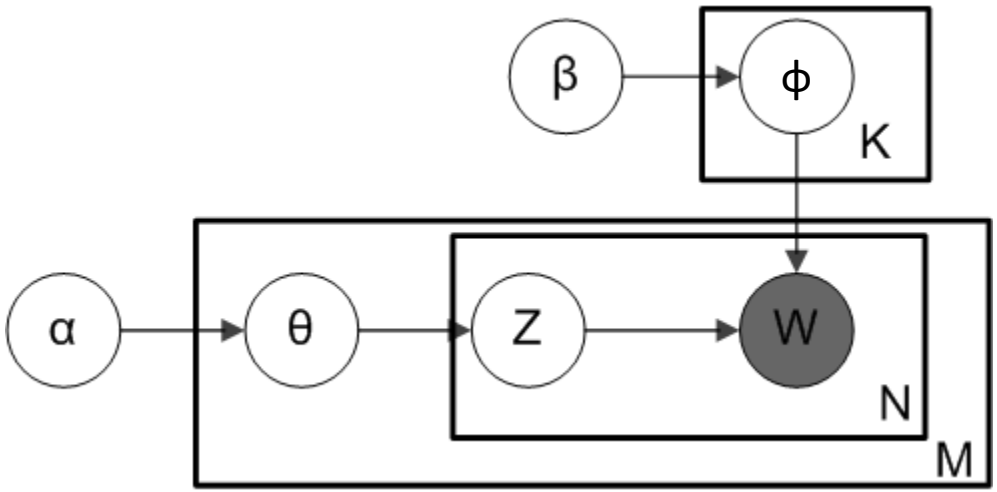
# Latent Dirichlet Allocation Topic Model (Blei et al., 2003)

Documents have **distributions over topics** $\theta^{(d)}$

Topics are **distributions over words** $\phi^{(k)}$

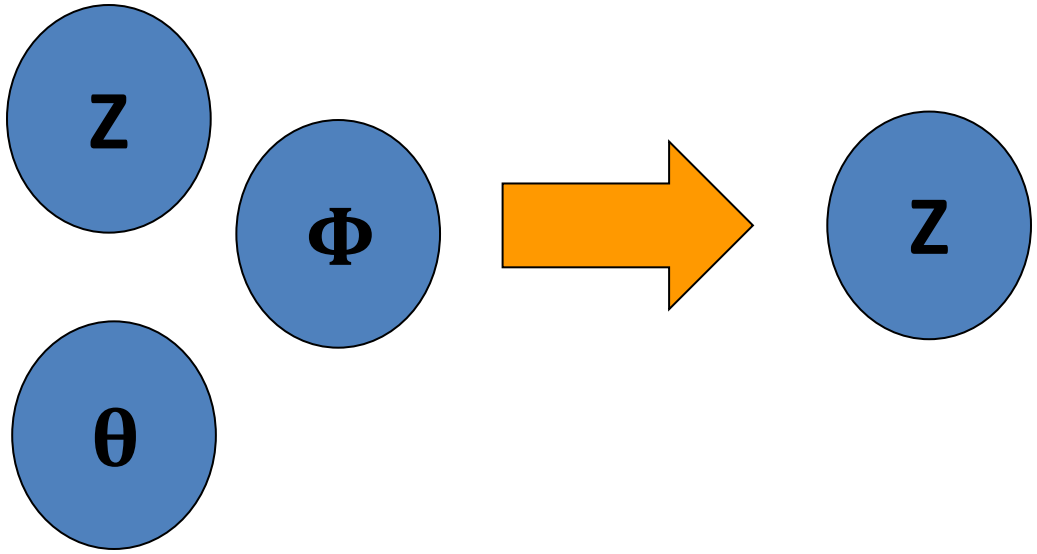
**Assumed generative process:** (full model includes priors on $\theta$, $\phi$)

- For each **document** d
  - For each **word** $w_{d,n}$
    - Draw a **topic assignment** $z_{d,n} \sim \text{Discrete}(\theta^{(d)})$
    - Draw a **word** from the chosen topic $w_{d,n} \sim \text{Discrete}(\phi^{(z_{d,n})})$



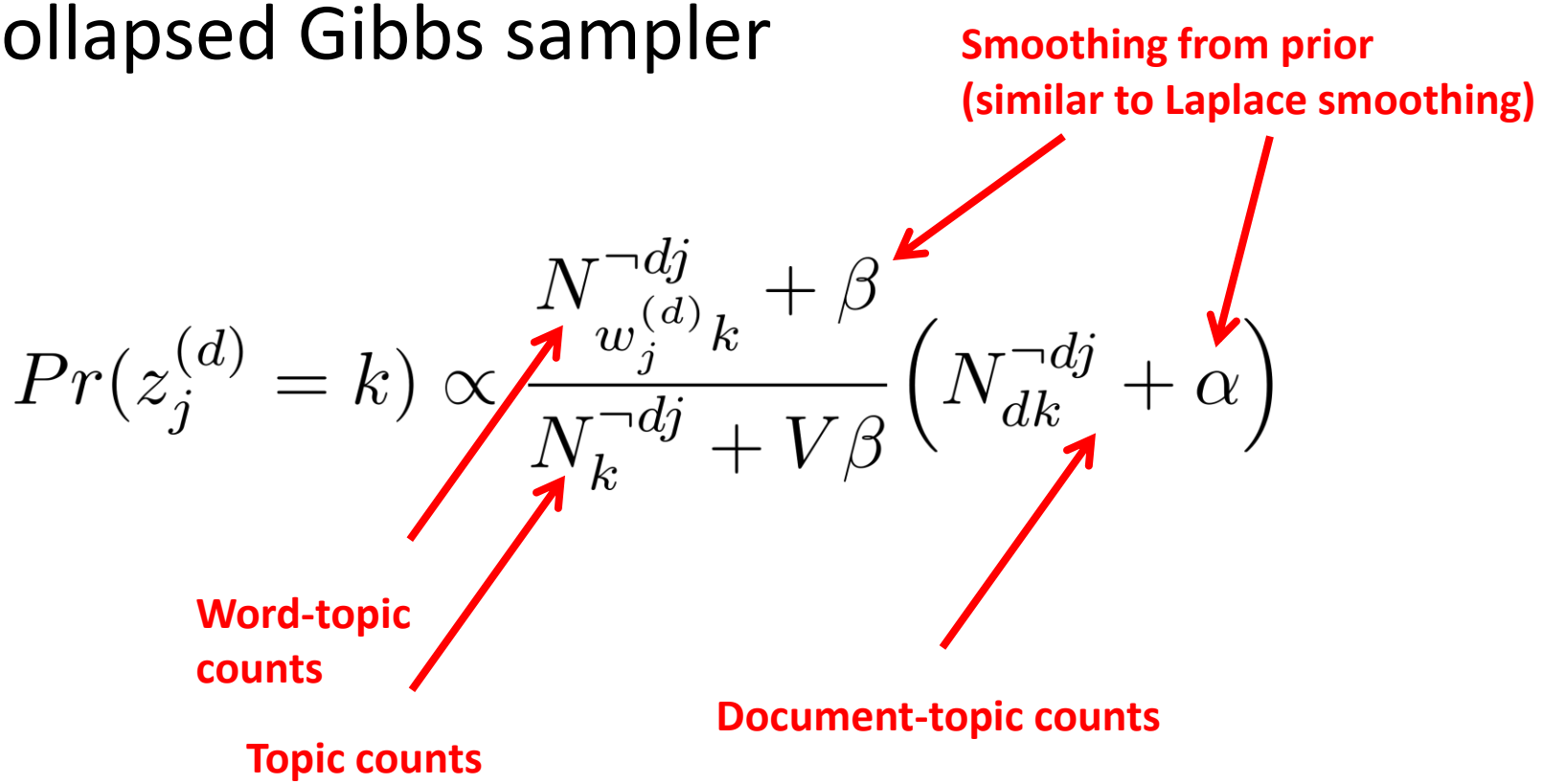
# Collapsed Gibbs sampler for LDA
## *Griffiths and Steyvers (2004)*

- Marginalize out the parameters, and perform inference on the **latent variables only**



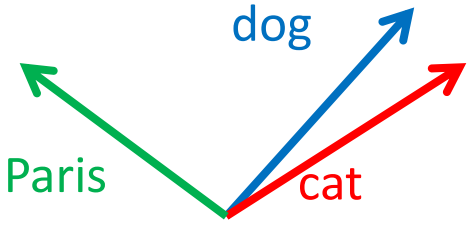
# Collapsed Gibbs sampler for LDA
## *Griffiths and Steyvers (2004)*

- Collapsed Gibbs sampler



# Word Embeddings

- Language models which learn to represent **dictionary words** with **vectors**




dog: (0.11, -1.5, 2.7, … )
cat: (0.15, -1.2, 3.2, … )
Paris: (4.5, 0.3, -2.1, …)

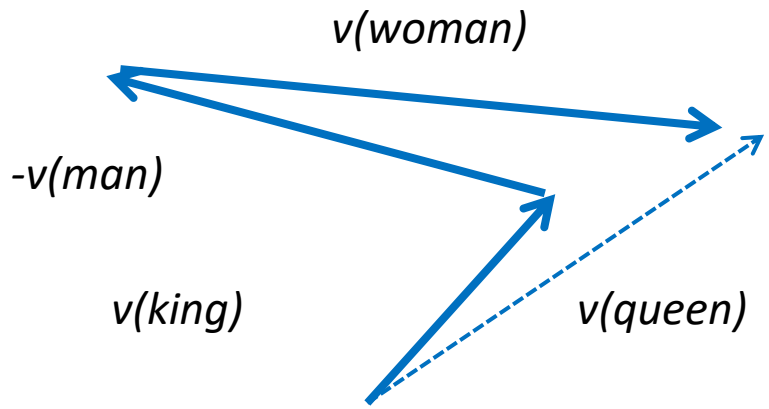
- Nuanced representations for words
- Improved performance for many NLP tasks
  - translation, part-of-speech tagging, chunking, NER, …
- NLP "from scratch"? (Collobert et al., 2011)

# Word Embeddings

- Vector arithmetic solves analogy tasks:

*man* is to *king* as *woman* is to _____?

*v(king) - v(man) + v(woman) ≈ v(queen)*



# The Distributional Hypothesis

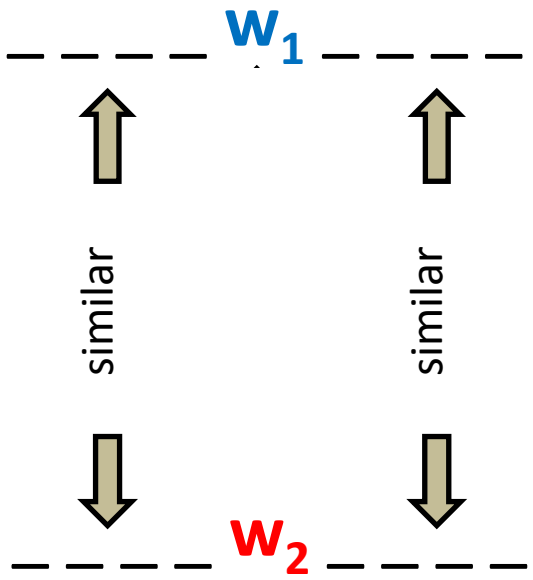
- "There is a correlation between **distributional similarity** and **meaning similarity**, which allows us to utilize the former in order to estimate the latter." (Sahlgren, 2008)

_ _ _ _ $w_1$ _ _ _ _

_ _ _ _ $w_2$ _ _ _ _

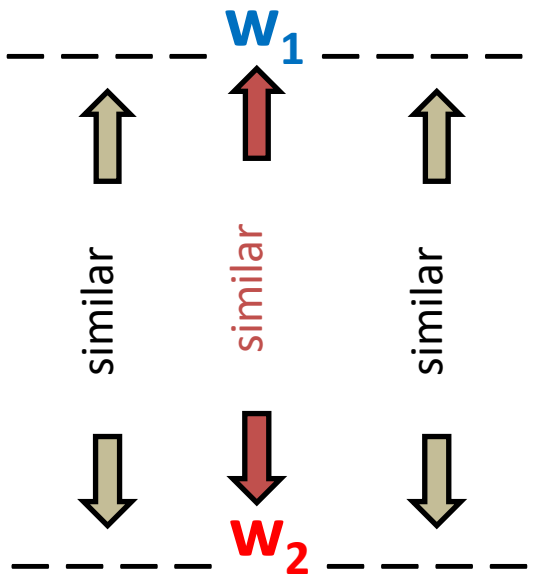
# The Distributional Hypothesis

- "There is a correlation between **distributional similarity** and **meaning similarity**, which allows us to utilize the former in order to estimate the latter." (Sahlgren, 2008)



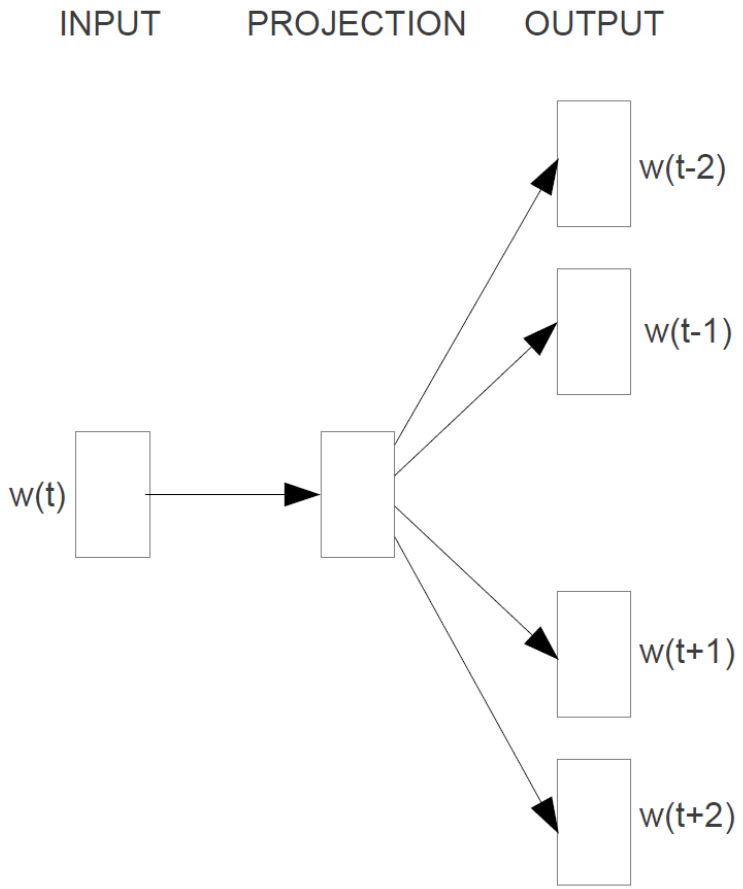
# The Distributional Hypothesis

- "There is a correlation between **distributional similarity** and **meaning similarity**, which allows us to utilize the former in order to estimate the latter." (Sahlgren, 2008)



# Word2vec (Mikolov et al., 2013)

## Skip-Gram




A log-bilinear classifier for the **context of a given word**

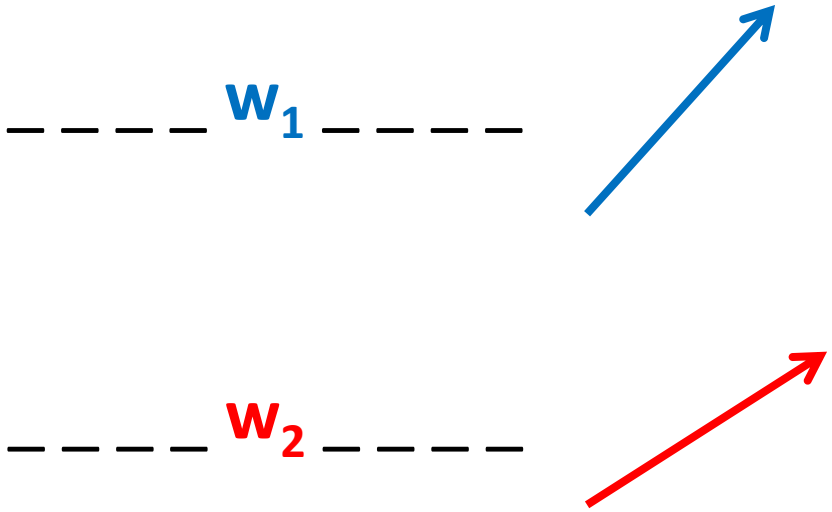
$$p(w_j|w_i) \propto \exp(v'^{\mathsf{T}}_{w_j} v_{w_i})$$

$v_w$ :"input" vectors

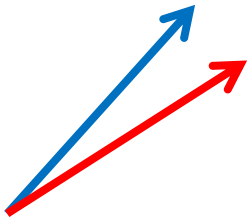
$v'_w$ :"output" vectors

Figure due to Mikolov et al. (2013)

# The Skip-Gram Encodes the Distributional Hypothesis




- Word vectors encode **distribution of context words**
- **Similar words** assumed to have **similar vectors**



# Word2vec (Mikolov et al., 2013)

- Key insights:
  - **Simple models** can be trained efficiently on **big data**
  - High-dimensional simple embedding models, trained on massive data sets, can **outperform sophisticated neural nets**

# Word Embeddings for Computational Social Science?

- Word embeddings have many advantages
  - Capture similarities between words
  - Often better classification performance than topic models

- Have not yet been widely adopted for **computational social science** research, perhaps due to the following limitations:
  - Target corpus of interest is often **not big data**
  - It is important for the model to be **interpretable**

# Contributions of this Work

- **Interpretable, statistically efficient embedding model**
- **Efficient training algorithm**, using recent advances from both topic models and word embeddings:
- Experimental results and **computational social science** case studies
- **Practical recommendations** and insights
  - use of *generic big data embeddings*, which is a very common practice in NLP

# The Skip-Gram as a Probabilistic Model

- Can view skip-gram as probabilistic model for ``generating'' context words

For each word in the corpus $w_i$

For each word $w_j \in context(i)$

Draw $w_j|w_i$ via $p(w_j|w_i) \propto exp({v'_{w_j}}^{\mathsf{T}} v_{w_i} + b_j)$

Conditional discrete distribution over words: can identify with a **topic**

# The Skip-Gram as a Probabilistic Model

**Naïve Bayes** conditional independence

For each word in the corpus $w_i$

For each word $w_j \in context(i)$

Draw $w_j|w_i$ via $p(w_j|w_i) \propto exp(v'_{w_j}{}^{\mathsf{T}} v_{w_i} + b_j)$

Observed "**cluster**" assignment

"**Topic**" distribution for input word $w_i$

# Analogous Topic Model Corresponding to Skip-Gram

**Naïve Bayes** conditional independence assumption

For each word in the corpus $w_i$

For each word $w_j \in context(i)$

Draw $w_j|w_i \sim \text{discrete}(\phi^{(w_i)})$

Observed "**cluster**" assignment

"**Topic**" for input word $w_i$

# Grid of Models' "Generative" Processes

| | Skip-gram | Skip-gram topic model |
|---|---|---|
| Naive Bayes | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) \propto exp({v'_{w_c}}^{\intercal} v_{w_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) = \text{Discrete}(\phi^{(w_i)})$ |
| Mixed membership | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) \propto exp({v'_{w_c}}^{\intercal} v_{z_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) = \text{Discrete}(\phi^{(z_i)})$ |

# Grid of Models' "Generative" Processes

| | Skip-gram | Skip-gram topic model |
|---|---|---|
| Naive Bayes | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) \propto exp({v'_{w_c}}^\intercal v_{w_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) = \text{Discrete}(\phi^{(w_i)})$ |
| Mixed membership | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) \propto exp({v'_{w_c}}^\intercal v_{z_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) = \text{Discrete}(\phi^{(z_i)})$ |

J. R. Foulds. Mixed Membership Word Embeddings for Computational Social Science. Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS), 2018.

# Grid of Models' "Generative" Processes

Identifying word distributions with topics leads to analogous topic model

| | Skip-gram | Skip-gram topic model |
|---|---|---|
| Naive Bayes | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) \propto exp({v'_{w_c}}^\intercal v_{w_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) = \text{Discrete}(\phi^{(w_i)})$ |
| Mixed membership | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) \propto exp({v'_{w_c}}^\intercal v_{z_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) = \text{Discrete}(\phi^{(z_i)})$ |

# Grid of Models' "Generative" Processes

Identifying word distributions with topics leads to analogous topic model

| | Skip-gram | Skip-gram topic model |
|---|---|---|
| Naive Bayes | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) \propto exp({v'_{w_c}}^\intercal v_{w_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) = \text{Discrete}(\phi^{(w_i)})$ |
| Mixed membership | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) \propto exp({v'_{w_c}}^\intercal v_{z_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) = \text{Discrete}(\phi^{(z_i)})$ |

Relax naïve Bayes assumption, replace with mixed membership model.
-flexible representation for words
-parameter sharing

# Grid of Models' "Generative" Processes

Identifying word distributions with topics leads to analogous topic model

| | Skip-gram | Skip-gram topic model |
|---|---|---|
| Naive Bayes | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) \propto exp({v'_{w_c}}^{\intercal} v_{w_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|w_i) = \text{Discrete}(\phi^{(w_i)})$ |
| Mixed membership | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) \propto exp({v'_{w_c}}^{\intercal} v_{z_i} + b_{w_c})$ | For each word in the corpus $w_i$<br>Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$<br>For each word $w_c \in context(i)$<br>Draw $w_c$ via $p(w_c\|z_i) = \text{Discrete}(\phi^{(z_i)})$ |

Reinstate word vector representation

Relax naïve Bayes assumption, replace with mixed membership model.
-flexible representation for words
-parameter sharing

# Mixed Membership Skip-Gram Topic Model

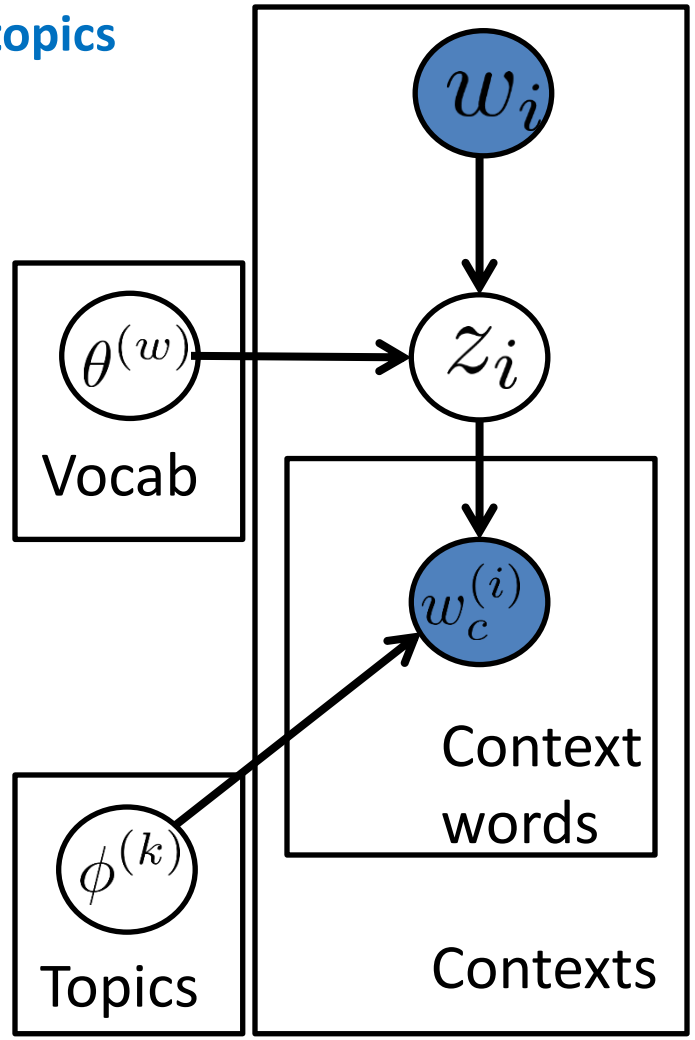
- Each input word has a **distribution over topics**
- Topics **shared** across all input words

For each word in the corpus $w_i$

Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$

For each word $w_j \in context(i)$

Draw $w_j | z_i \sim \text{Discrete}(\phi^{(z_i)})$



# Mixed Membership Skip-Gram

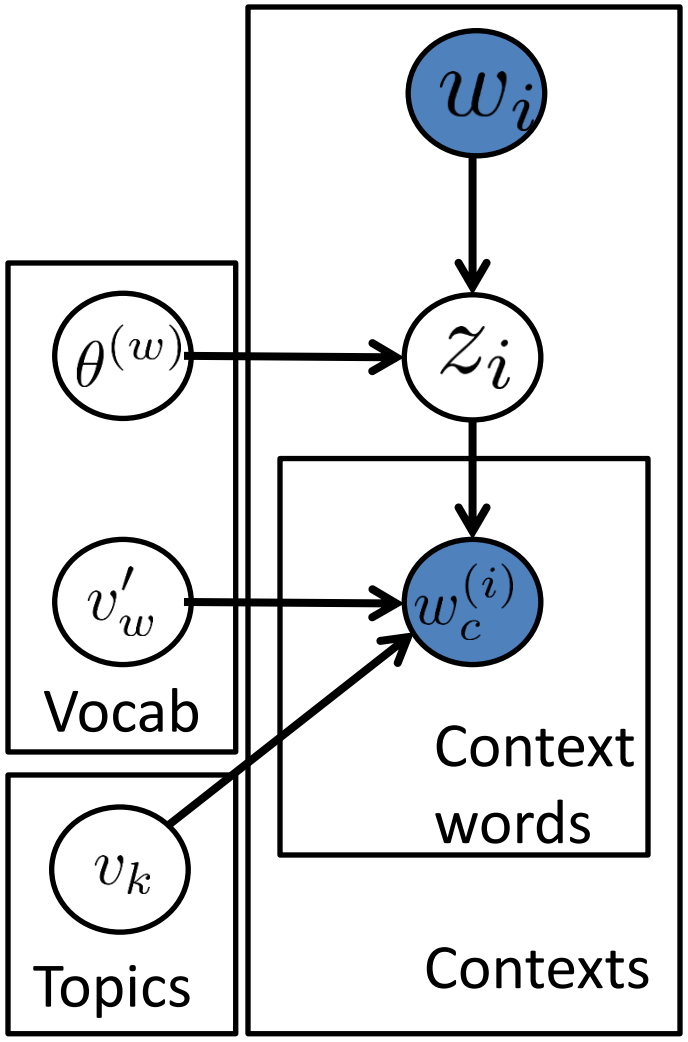
For each word in the corpus $w_i$

Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$

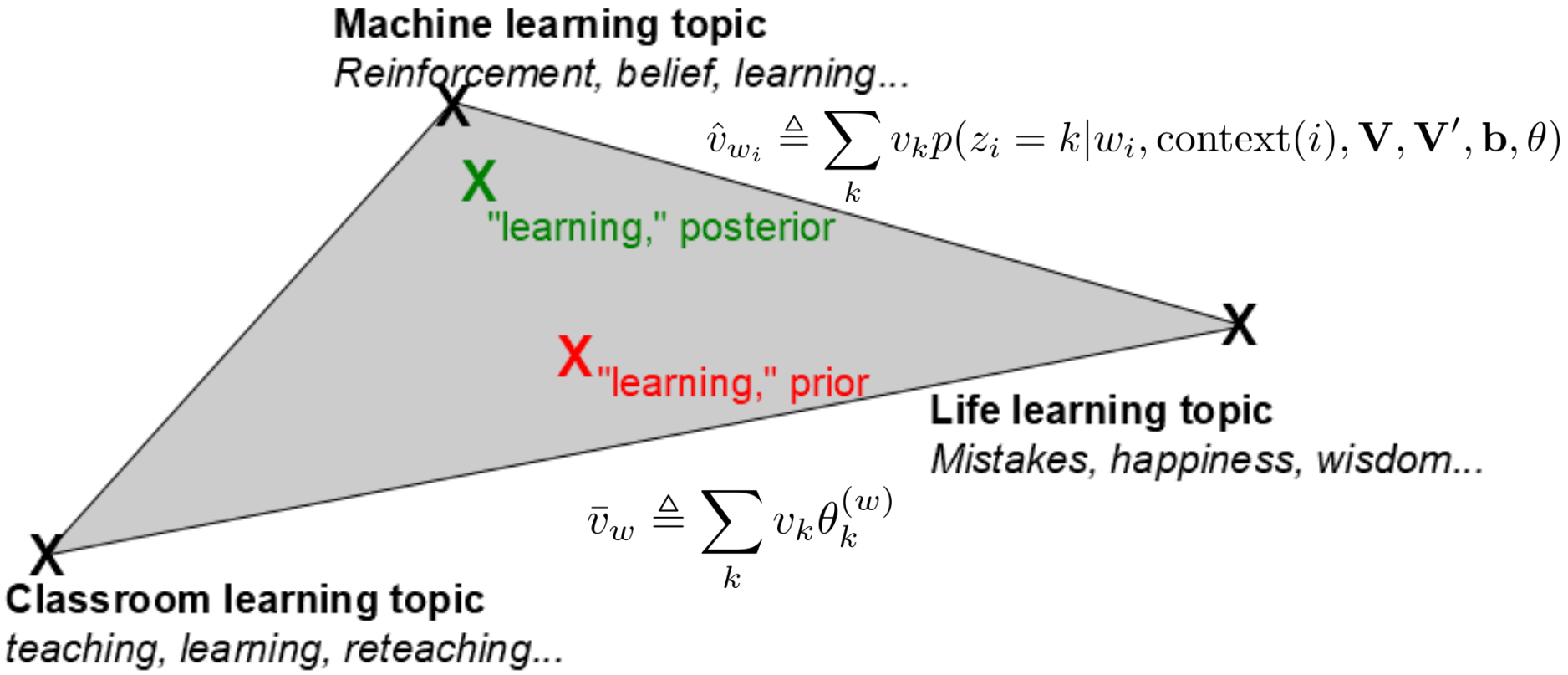
For each word $w_c \in context(i)$

Draw $w_c$ via

$$p(w_c|z_i) \propto exp({v'_{w_c}}^{\intercal} v_{z_i} + b_{w_c})$$



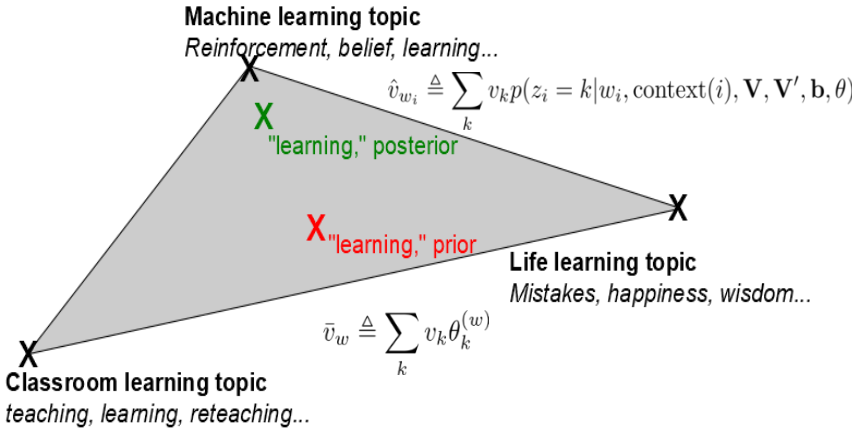
# Mixed Membership Word Embeddings




Context: *"We used an SVM when **learning** to predict the class labels."*

**Word embeddings are convex combinations of topic embeddings**

# Mixed Membership Word Embeddings




- Words have **mixed membership distributions** over topics $\theta^{(w)}$
- **Topics** have embeddings $\bar{v}_w$, words don't. Resolves **polysemy**
- **Fewer vectors than words**: **statistical efficiency** on small data
- Word embeddings recovered as prior mean $\bar{v}_w$ or posterior mean vectors $\hat{v}_{w_i}$ -**convex combinations of topic embeddings**
- **Interpretable**: topics can be interpreted via top words lists, word embeddings are defined in terms of topic embeddings

# Mixed Membership Skip-Gram Posterior Inference for Topic Vector

- **Context** can be leveraged for inferring the topic vector at test time, via **Bayes' rule**:

$$Pr(v_{w_i} = v_k | w_i, \text{context}(i), \mathbf{V}, \Theta) \propto Pr(z_i = k | w_i, \Theta) Pr(\text{context}(i) | z_i = k, \mathbf{V})$$

$$= \theta_k^{(w_i)} \prod_{c \in \text{context}(i)} \frac{exp(v'^{\intercal}_{w_c^{(i)}} v_k)}{\sum_{j'=1}^{V} exp(v'^{\intercal}_{j'} v_k)}$$

# Bayesian Inference for MMSG Topic Model

- Bayesian version of model with Dirichlet priors

- Collapsed Gibbs sampling

$$p(z_i = k|\cdot) \propto \left(n_k^{(w_i)\neg i} + \alpha_k\right) \prod_{c=1}^{|\text{context}(i)|} \frac{n_{w_c^{(i)}}^{(k)\neg_i} + \beta_{w_c^{(i)}} + n_{w_c^{(i)}}^{(i,c)}}{n^{(k)\neg i} + \sum_{w'} \beta_{w'} + c - 1}$$

# Bayesian Inference for MMSG Topic Model

- Challenge 1: want relatively large # topics

- Solution: **Metropolis-Hastings-Walker** algorithm (Li et al. 2014)
  - **Alias table** data structure, amortized O(1) sampling
  - **Sparse implementation**, sublinear in topics K
  - **Metropolis-Hastings correction** for sampling from stale distributions

# Metropolis-Hastings-Walker (Li et al. 2014)

Sparse

Dense, slow-changing

$$p(z_i = k|\cdot) \propto n_k^{(w_i)\neg i} A_{ik} + \alpha_k A_{ik}$$

$$A_{ik} = \prod_{c=1}^{|\text{context}(i)|} \frac{n_{w_c^{(i)}}^{(k)\neg_i} + \beta_{w_c^{(i)}} + n_{w_c^{(i)}}^{(i,c)}}{n^{(k)\neg i} + \sum_{w'} \beta_{w'} + c - 1}$$

- Approximate second term of the mixture, sample efficiently via alias tables, correct via Metropolis

# Metropolis-Hastings-Walker Proposal

- Dense part of Gibbs update is a "*product of experts*" (Hinton, 2004), expert for each context word

# Metropolis-Hastings-Walker Proposal

- Dense part of Gibbs update is a "*product of experts*" (Hinton, 2004), expert for each context word

- Use a "*mixture of experts*" proposal distribution

$$c \sim \text{Uniform}(|\text{context}(w_i)|) \ , \ q_{w_c}(k) \propto \frac{n_{w_c}^{(k)} + \beta_{w_c}}{n^{(k)} + \sum_{w'} \beta_{w'}}$$

# Metropolis-Hastings-Walker Proposal

- Dense part of Gibbs update is a "*product of experts*" (Hinton, 2004), expert for each context word

- Use a "*mixture of experts*" proposal distribution

$$c \sim \text{Uniform}(|\text{context}(w_i)|) \ , \ q_{w_c}(k) \propto \frac{n_{w_c}^{(k)} + \beta_{w_c}}{n^{(k)} + \sum_{w'} \beta_{w'}}$$

- Can sample efficiently from "experts" via alias tables

# Bayesian Inference for MMSG Topic Model

- Challenge 2: cluster assignment updates almost deterministic, vulnerable to local maxima

- Solution: **simulated annealing**
  - Anneal temperature of model
    - adjusting Metropolis-Hastings acceptance probabilities

# Approximate MLE for Mixed Membership Skip-Gram

- Online EM impractical
  - M-step is O(V)
  - E-step is O(KV)

# Approximate MLE for Mixed Membership Skip-Gram

- Online EM impractical
  - M-step is O(V)
  - E-step is O(KV)

- Approximate online EM
  - Key insight: MMSG topic model **equivalent** to word embedding model, up to Dirichlet prior
    - **Pre-solve E-step** via topic model CGS
    - Apply **Noise Contrastive Estimation** to solve M-step
  - Entire algorithm approximates maximum likelihood estimation via these two principled approximations

# Qualitative Results, NIPS Corpus

| | Input word = "Bayesian" |
|---|---|
| Model | Top words in topic for input word. Top 3 topics for word shown for mixed membership models. |
| SGTM | model networks learning neural bayesian data models approach network framework |
| SG | belief learning framework models methods markov function bayesian based inference |
| MMSGTM | bayesian model parameters posterior prior distribution approach likelihood variational inference |
| | neural networks computation bayesian learning mackay framework network functions practical |
| | carlo monte bayesian gaussian neural neal implementation methods models williams |
| MMSG | variational likelihood bayesian inference approach parameters marginal dirichlet posterior sampling |
| | neural bayesian learning networks computation framework regularization entropy press mackay |
| | neal rasmussen monte bayesian models http press neural barber carlo |

# Qualitative Results, NIPS Corpus

| Model | Input word = "Bayesian"<br>Top words in topic for input word. Top 3 topics for word shown for mixed membership models. |
|---|---|
| SGTM | model networks learning neural bayesian data models approach network framework |
| SG | belief learning framework models methods markov function bayesian based inference |
| MMSGTM | bayesian model parameters posterior prior distribution approach likelihood variational inference |
| | neural networks computation bayesian learning mackay framework network functions practical |
| | carlo monte bayesian gaussian neural neal implementation methods models williams |
| MMSG | variational likelihood bayesian inference approach parameters marginal dirichlet posterior sampling |
| | neural bayesian learning networks computation framework regularization entropy press mackay |
| | neal rasmussen monte bayesian models http press neural barber carlo |

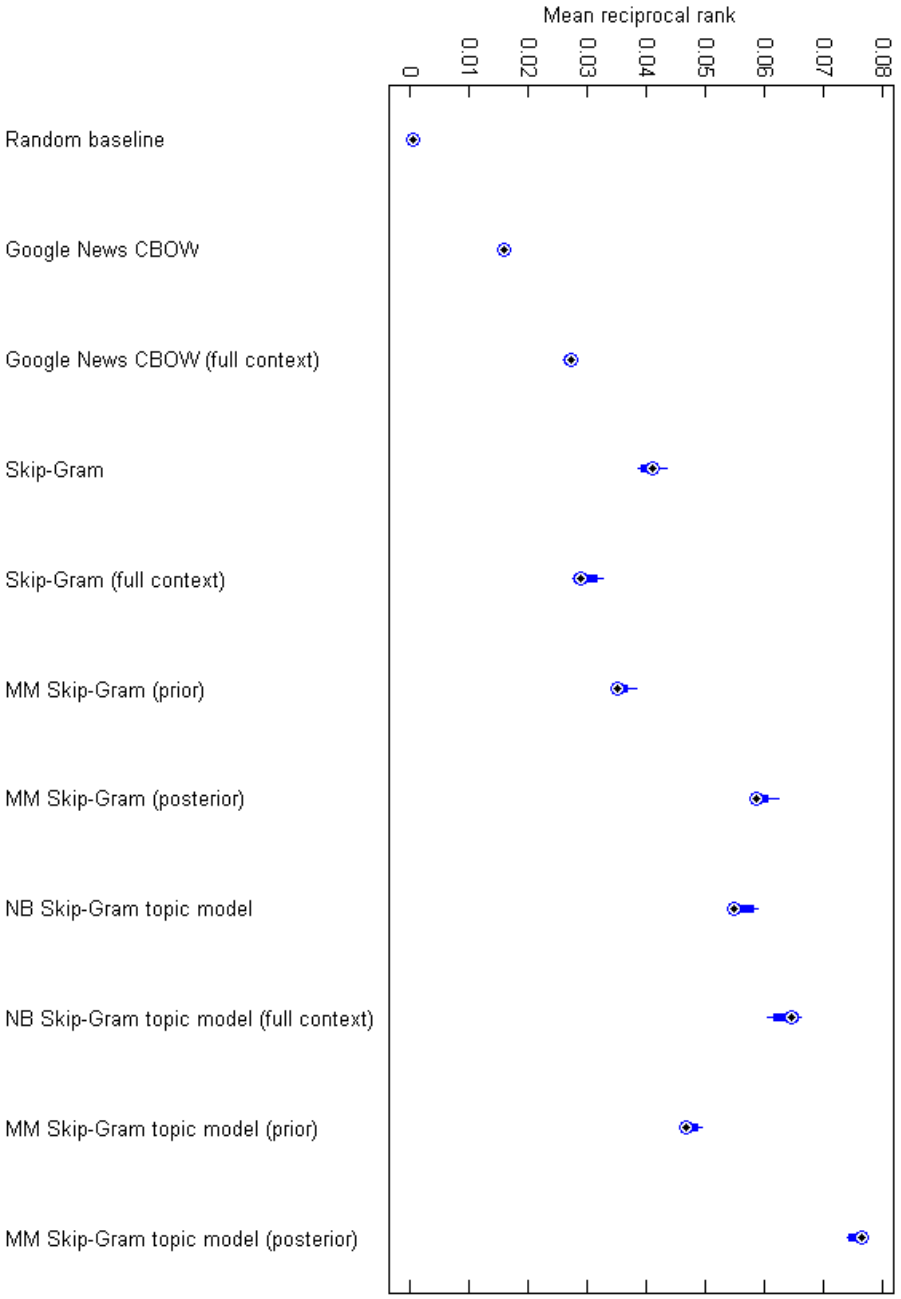
# Qualitative Results, NIPS Corpus

| | Input word = "Bayesian" |
|---|---|
| Model | Top words in topic for input word. Top 3 topics for word shown for mixed membership models. |
| SGTM | model networks learning neural bayesian data models approach network framework |
| SG | belief learning framework models methods markov function bayesian based inference |
| MMSGTM | bayesian model parameters posterior prior distribution approach likelihood variational inference |
| | neural networks computation bayesian learning mackay framework network functions practical |
| | carlo monte bayesian gaussian neural neal implementation methods models williams |
| MMSG | variational likelihood bayesian inference approach parameters marginal dirichlet posterior sampling |
| | neural bayesian learning networks computation framework regularization entropy press mackay |
| | neal rasmussen monte bayesian models http press neural barber carlo |

# Qualitative Results, NIPS Corpus

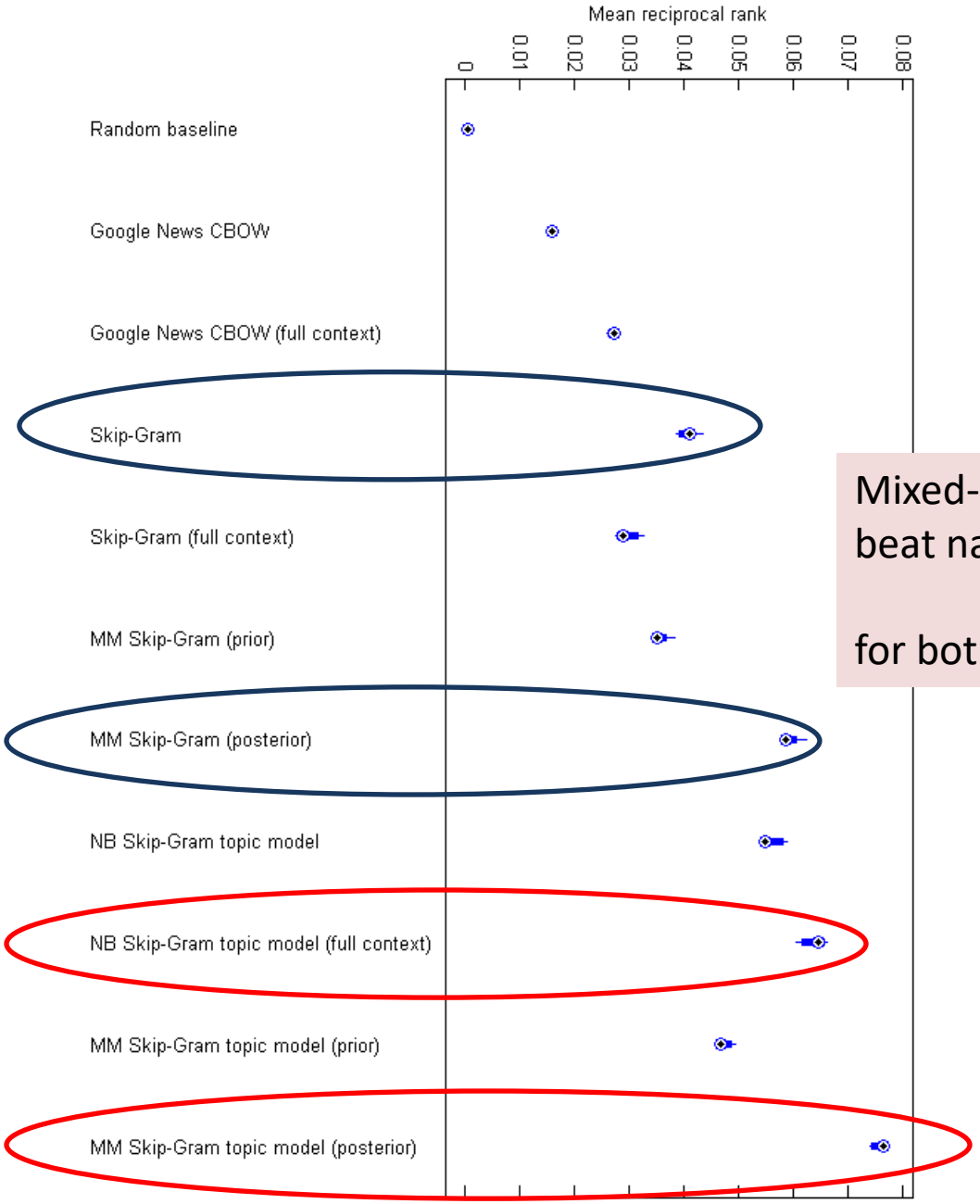
| | Input word = "Bayesian" |
|---|---|
| Model | Top words in topic for input word. Top 3 topics for word shown for mixed membership models. |
| SGTM | model networks learning neural bayesian data models approach network framework |
| SG | belief learning framework models methods markov function bayesian based inference |
| MMSGTM | bayesian model parameters posterior prior distribution approach likelihood variational inference |
| | neural networks computation bayesian learning mackay framework network functions practical |
| | carlo monte bayesian gaussian neural neal implementation methods models williams |
| MMSG | variational likelihood bayesian inference approach parameters marginal dirichlet posterior sampling |
| | neural bayesian learning networks computation framework regularization entropy press mackay |
| | neal rasmussen monte bayesian models http press neural barber carlo |

# Prediction Performance, NIPS Corpus




*(similar results on three other small datasets, see the paper)*

# Prediction Performance, NIPS Corpus

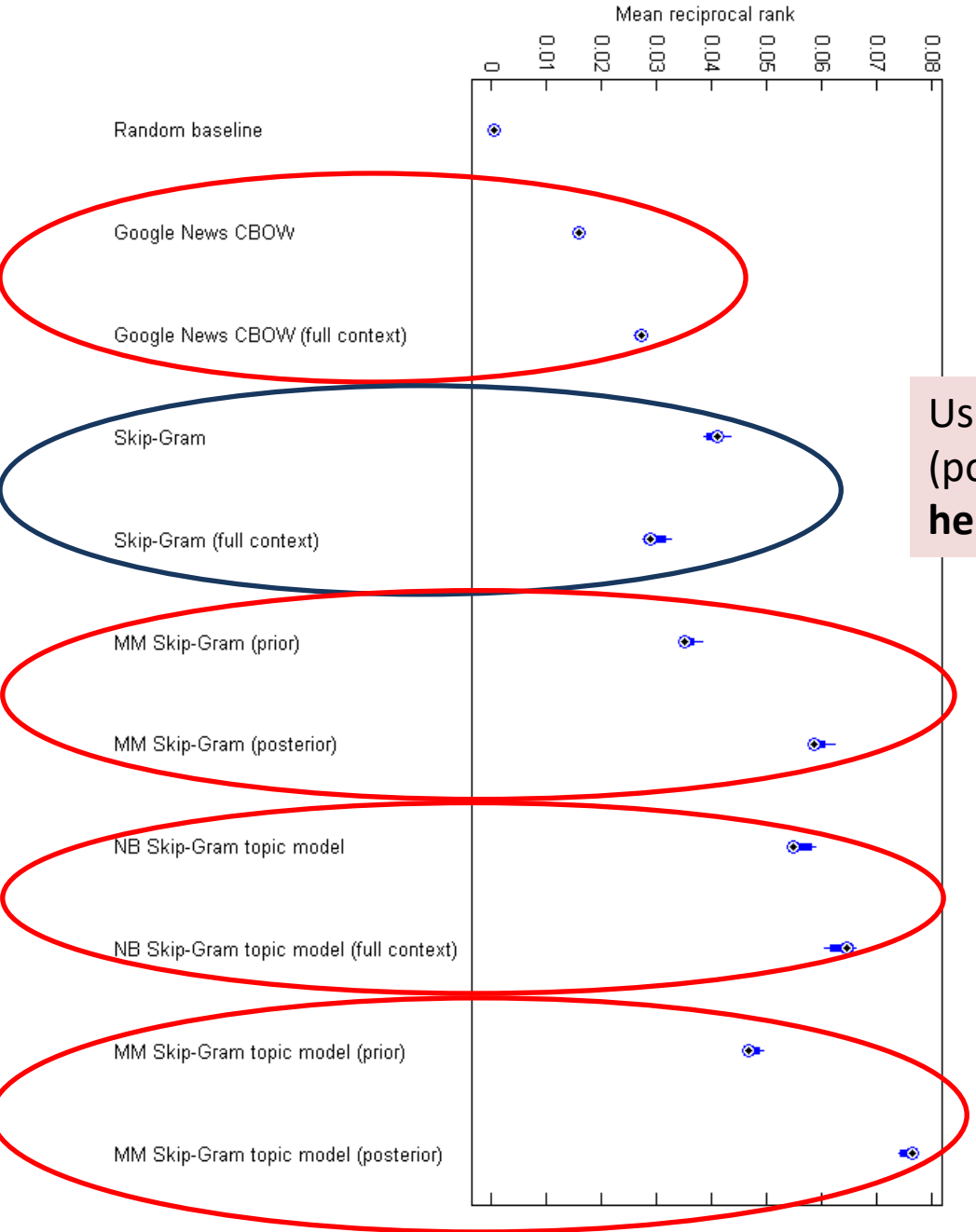



Mixed-membership models (w/ posterior) beat naïve Bayes models,

for both word embedding and topic models

*(similar results on three other small datasets, see the paper)*

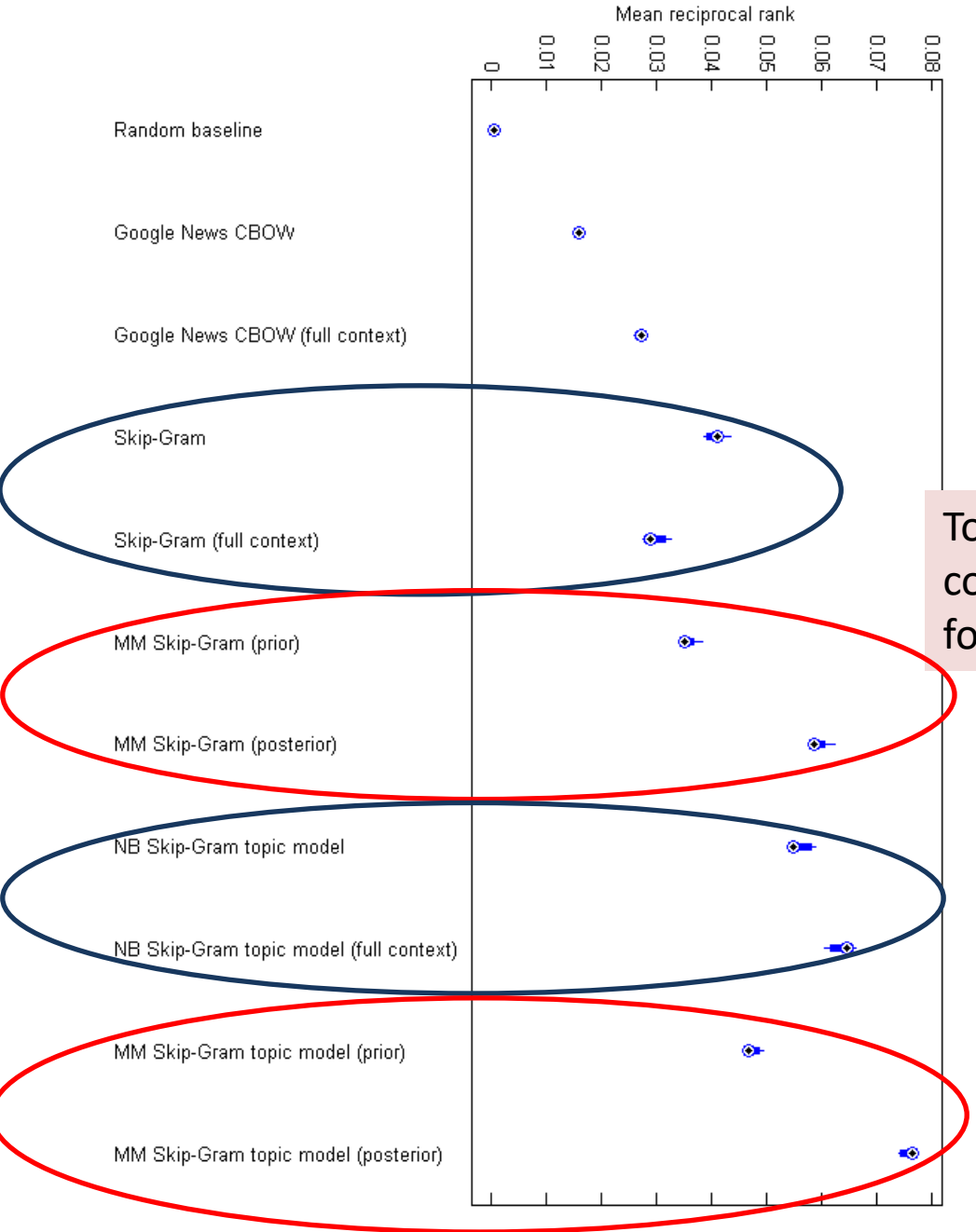
# Prediction Performance, NIPS Corpus




Using the **full context**
(posterior over topic or summing vectors)
**helps all models except the basic skip-gram**

*(similar results on three other small datasets, see the paper)*

# Prediction Performance, NIPS Corpus




Topic models beat their corresponding embedding models, for both **naïve Bayes** and **Mixed Membership**

*(similar results on three other small datasets, see the paper)*

# Downstream Tasks: Classification and Regression

| Dataset | #Classes | #Topics | Tf-idf | Google | MMSG | SG | MMSGTM | SG+MMSG | SG+MMSG+Google |
|---|---|---|---|---|---|---|---|---|---|
| 20 Newsgroups | 20 | 200 | 83.33 | 52.50 | 55.58 | 59.50 | 64.08 | 66.55 | 72.53 |
| Reuters-150 | 150 | 500 | 73.04 | 53.65 | 65.26 | 69.53 | 66.97 | 70.63 | 71.20 |
| Ohsumed | 23 | 500 | 43.07 | 20.56 | 31.82 | 37.57 | 32.41 | 39.53 | 40.27 |
| SOTU (RMSE) | Regression | 500 | 19.57 | 8.64 | 12.73 | 10.57 | 21.88 | 9.94 | 8.15 |

- Document categorization (classification accuracy, larger is better), and predicting the year of SOTU addresses (RMSE, smaller is better).
- **Target corpus beats generic big-data vectors** (except for SOTU, which is very small)
- **Skip-gram beats MMSG** for classification/regression – loss of granularity
- But, **concatenating the different vectors improves performance** over individual embeddings
    - MMSG, SG, generic Google vectors learn **complementary information**

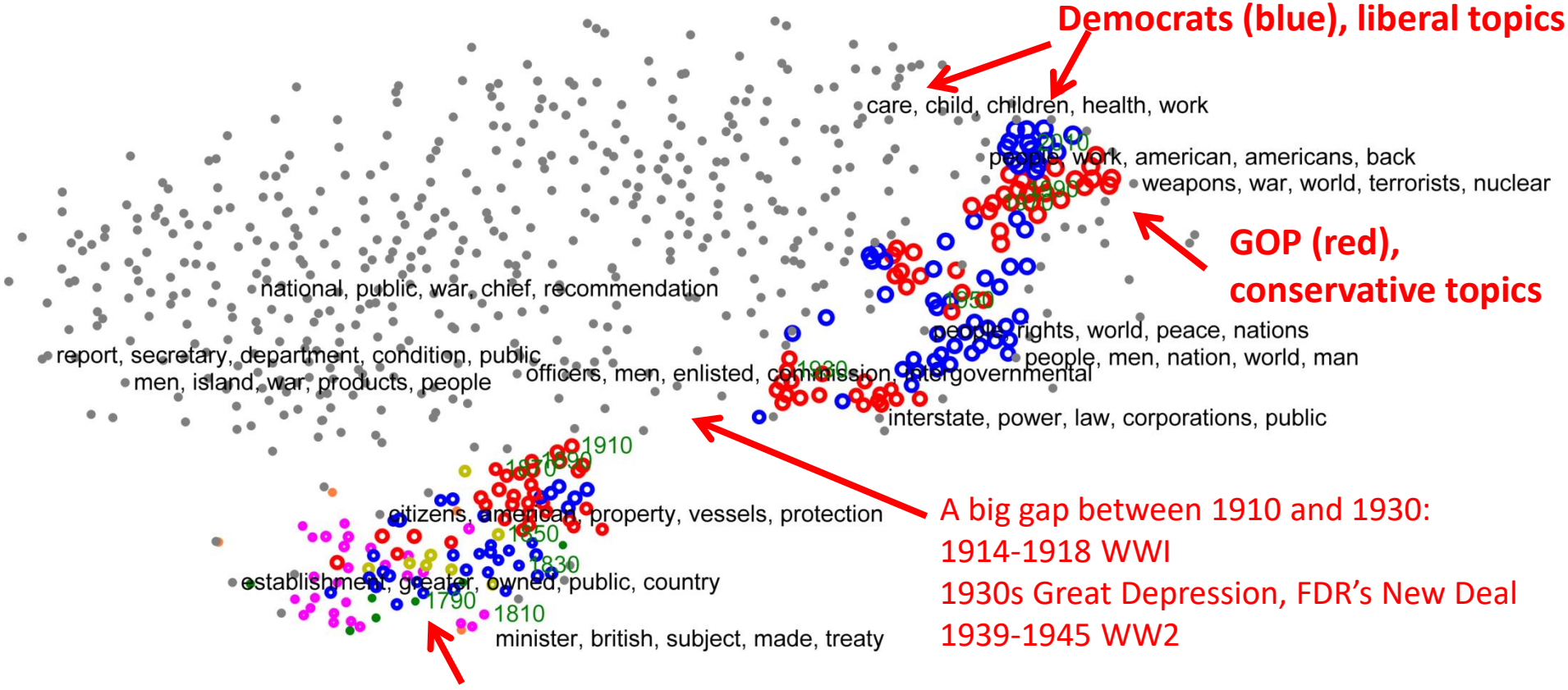
J. R. Foulds. Mixed Membership Word Embeddings for Computational Social Science. Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS), 2018.

# Vector Composition in Topic Space

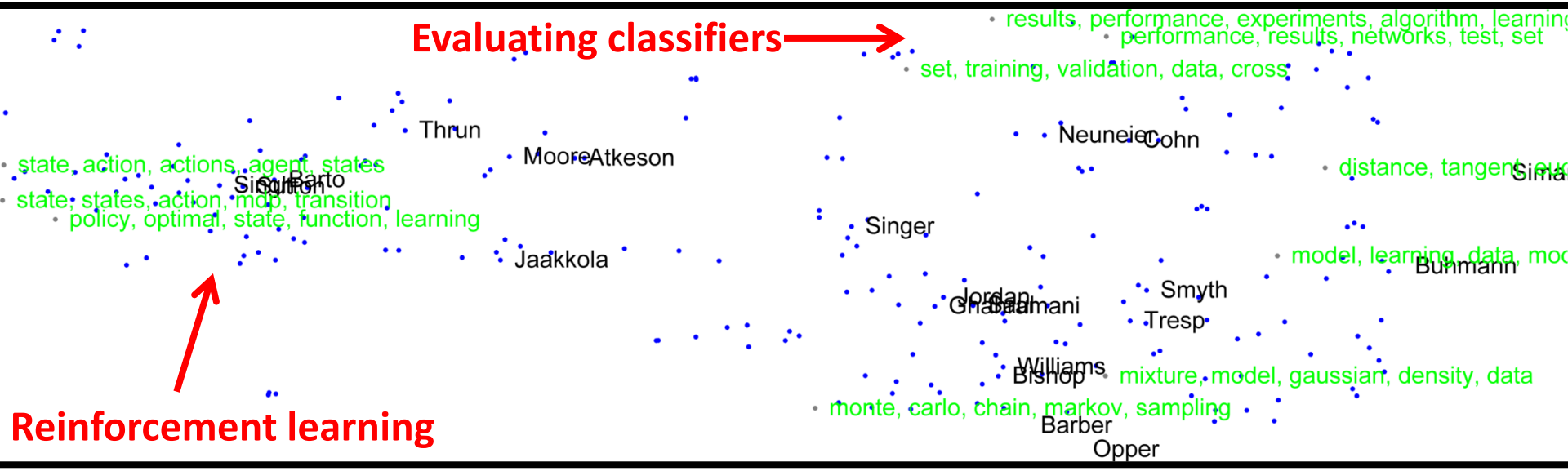
| Nearest topic after composition of mean vectors for words | |
|---|---|
| object + recognition | objects visual object recognition model |
| character + recognition | recognition segmentation character |
| speech + recognition | speech recognition hmm system hybrid |
| computer + vision | computer vision ieee image pattern |
| computer + science | university science colorado department |
| bias + variance | error training set data performance |
| covariance + variance | gaussian distribution model matrix |

J. R. Foulds. Mixed Membership Word Embeddings for Computational Social Science. Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS), 2018.

# State of the Union Addresses (t-SNE Projection)



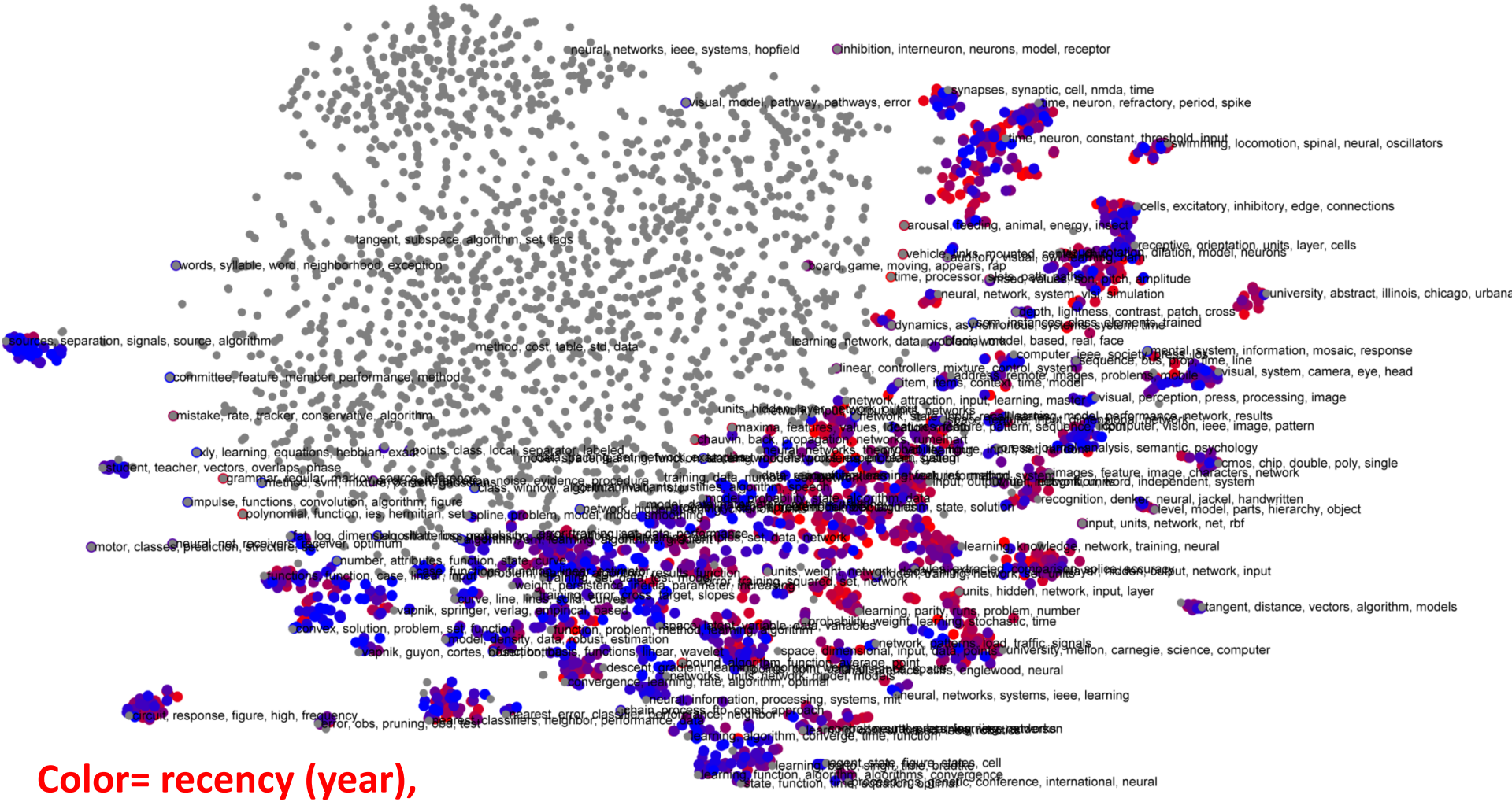
# NIPS Authors




**Blue = authors**

**Gray = topics**

# NIPS Documents




**Color= recency (year),**

**Red = older, blue = newer**

**Gray = topics**

# Conclusion

- Proposed **mixed membership**, **topic model** versions of skip-gram word embedding models
  - Statistically efficient, interpretable
- **Efficient training** via MHW collapsed Gibbs + NCE
- Proposed models **improve prediction**, useful for **computational social science**
- Ongoing/future work:
  - Scale to **big data** setting
  - Document embeddings

*Source code: https://github.com/jrfoulds/MixedMembershipWordEmbeddings*

# My Research Group: The Latent Lab