Reducing the Sampling Complexity of Topic Models

Aaron Li

joint work with Amr Ahmed, Sujith Ravi, Alex Smola
CMU and Google

MACHINE LEARNING DEPARTMENT

Google

Carnegie Mellon University

Outline

- Topic Models
  - Inference algorithms
  - Losing sparsity at scale
- Inference algorithm
  - Metropolis Hastings proposal
  - Walker's Alias method for $O(k_d)$ draws
- Experiments
  - LDA, Pitman-Yor topic models, HPYM
  - Distributed inference

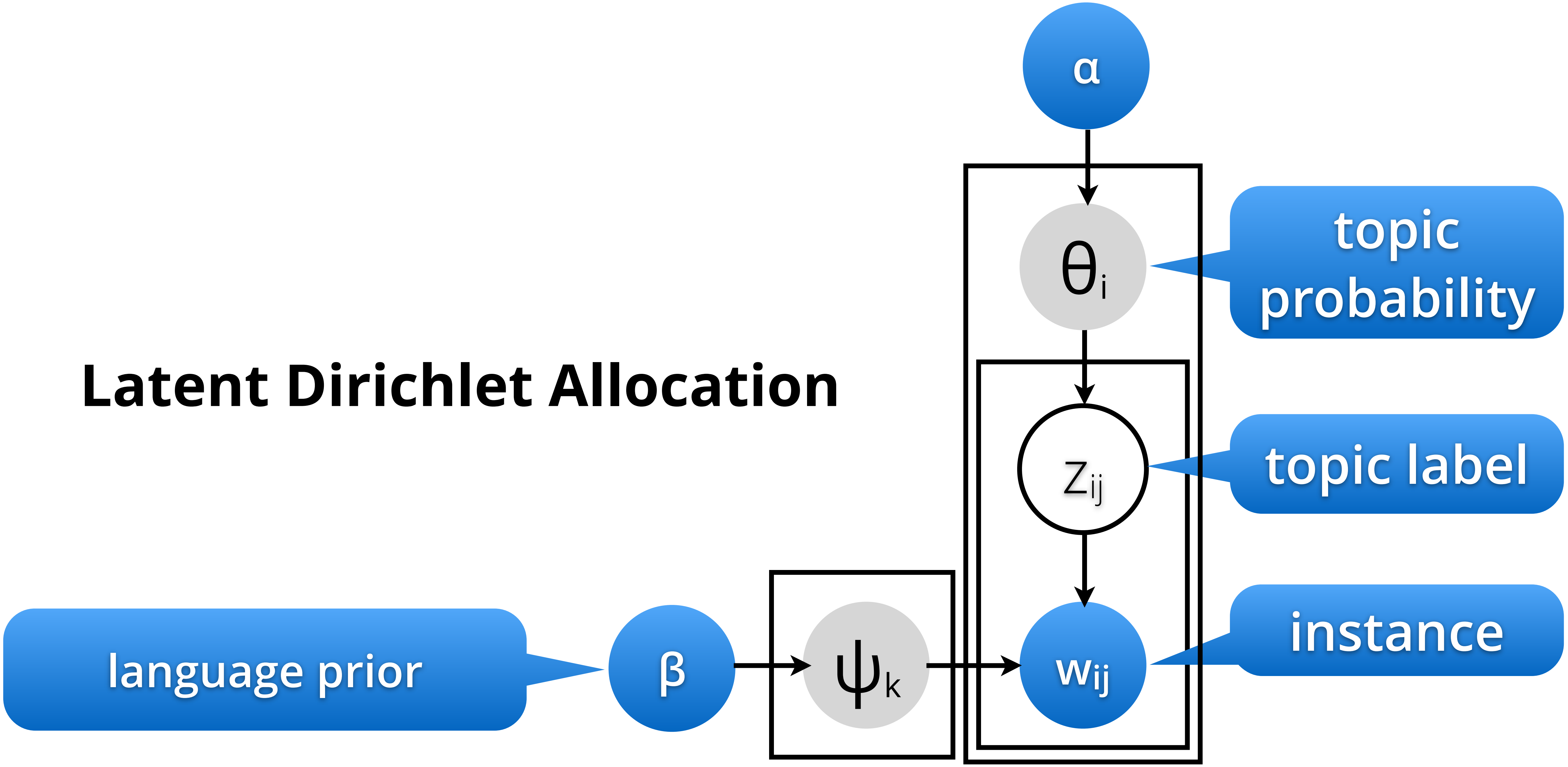
Models

Google

Carnegie Mellon University

Clustering & Topic Models

Latent Dirichlet Allocation



Google

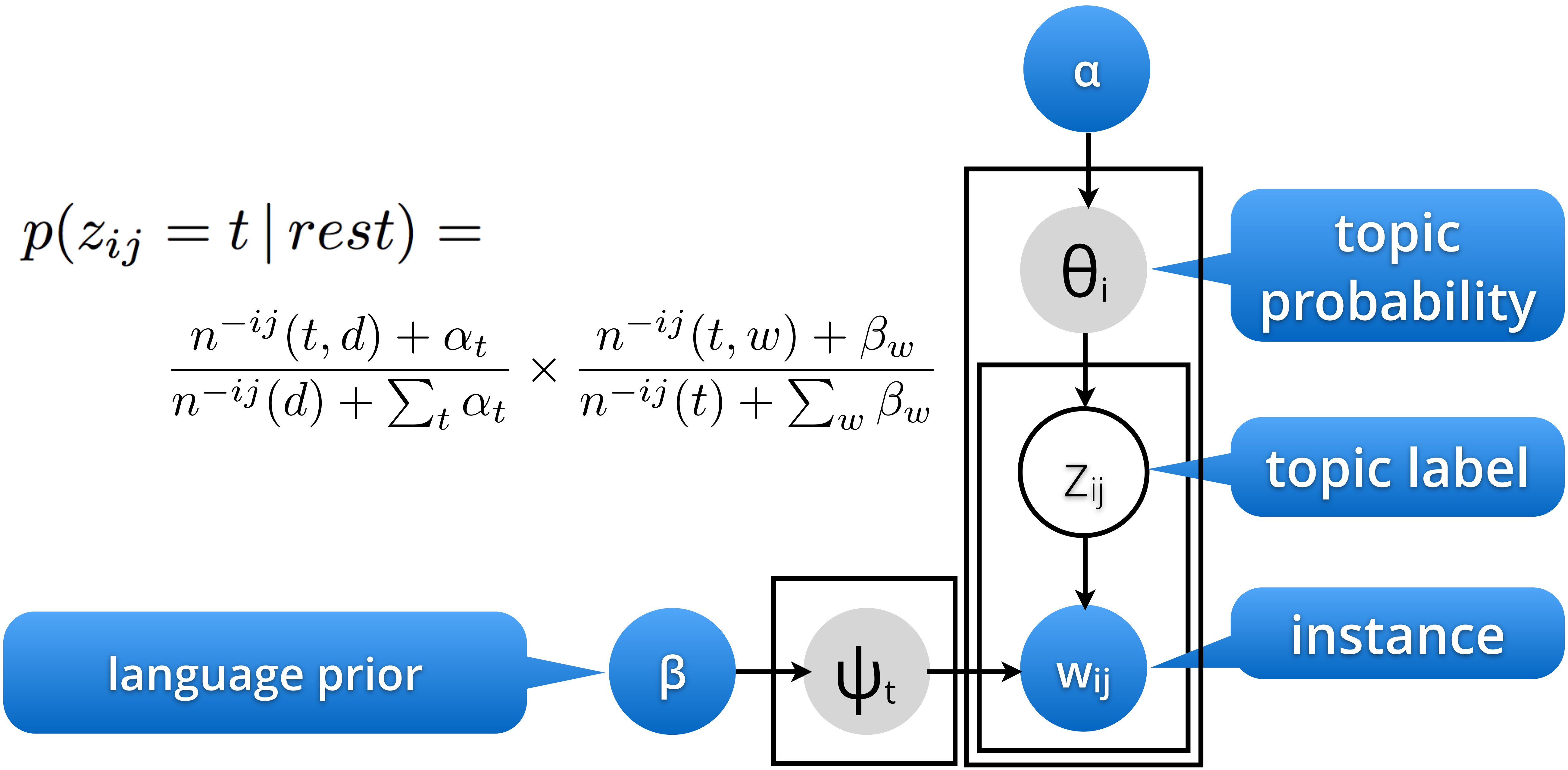
Carnegie Mellon University

Topics in text
(Blei, Ng, Jordan, 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Collapsed Gibbs Sampler (Griffiths & Steyvers, 2005)

$$p(z_{ij} = t \mid rest) = \frac{n^{-ij}(t,d) + \alpha_t}{n^{-ij}(d) + \sum_t \alpha_t} \times \frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \sum_w \beta_w}$$



Collapsed Gibbs Sampler

- For each document **i** do
  - For each word **j** in the document do
    - Resample topic for the word

sparse for most documents

sparse for small collections

$$\left(n^{-ij}(t,d) + \alpha_t\right) \times \frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \bar{\beta}}$$

dense

- Update (document, topic) table
- Update (word,topic) table

Exploiting Sparsity (Yao, Mimno, Mccallum, 2009)

- For each document **i** do
  - For each word **j** in the document do
    - Resample topic for the word

"constant" — sparse for most documents — sparse for small collections

$$\frac{\alpha_t \beta_w}{n^{-ij}(t) + \bar{\beta}} + n^{-ij}(t,d)\frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \bar{\beta}} + n^{-ij}(t,w)\frac{\alpha_t}{n^{-ij}(t) + \bar{\beta}}$$

    - Update (document, topic) table
    - Update (word,topic) table

amortized $O(k_d + k_w)$ time

Problem in Large Collections

For small datasets the assumption

$$k_d + k_w \ll k$$

is well satisfied.

For large datasets, assuming that the probability of occurrence for a given topic for a word is bounded from below by δ, Then the probability of the topic occurring at least once for a word in a collection of n documents is given by

$$1 - (1 - \delta)^n \geq 1 - e^{-n\delta} \rightarrow 1 \; for \; n \rightarrow \infty$$

Exploiting Sparsity (Yao, Mimno, Mccallum, 2009)

- For each document **i** do
  - For each word **j** in the document do
    - Resample topic for the word

"constant"

sparse for most documents

dense for large collections

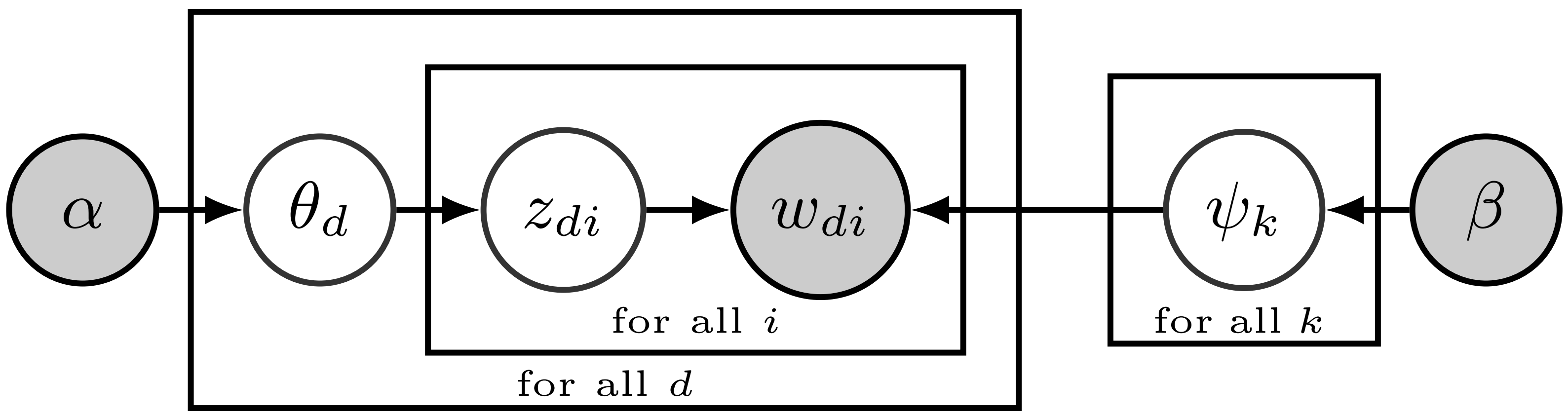
$$\frac{\alpha_t \beta_w}{n^{-ij}(t) + \bar{\beta}} + n^{-ij}(t,d)\frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \bar{\beta}} + n^{-ij}(t,w)\frac{\alpha_t}{n^{-ij}(t) + \bar{\beta}}$$

- Update (document, topic) table
- Update (word,topic) table

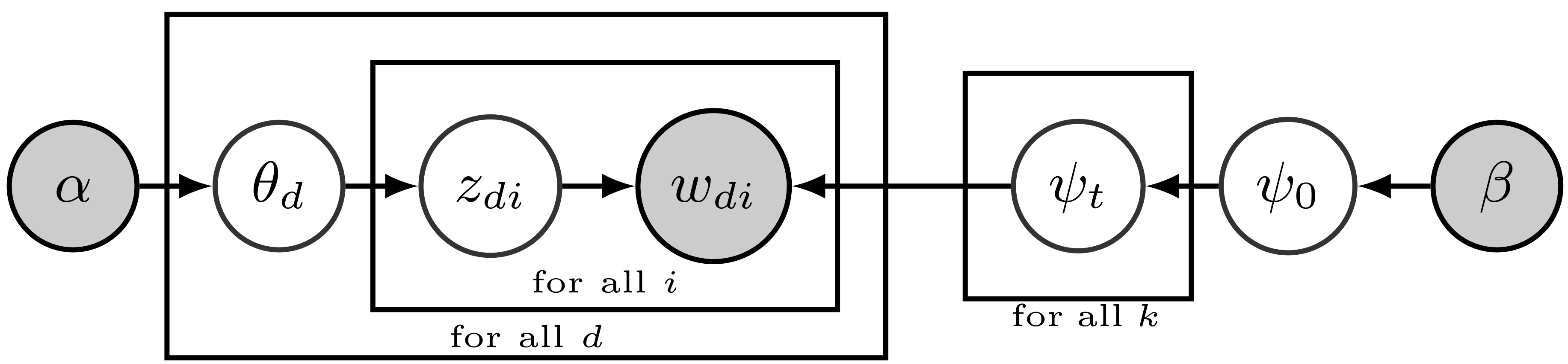
we solve this problem

More Models

- **LDA**



- **Poisson-Dirichlet Process**

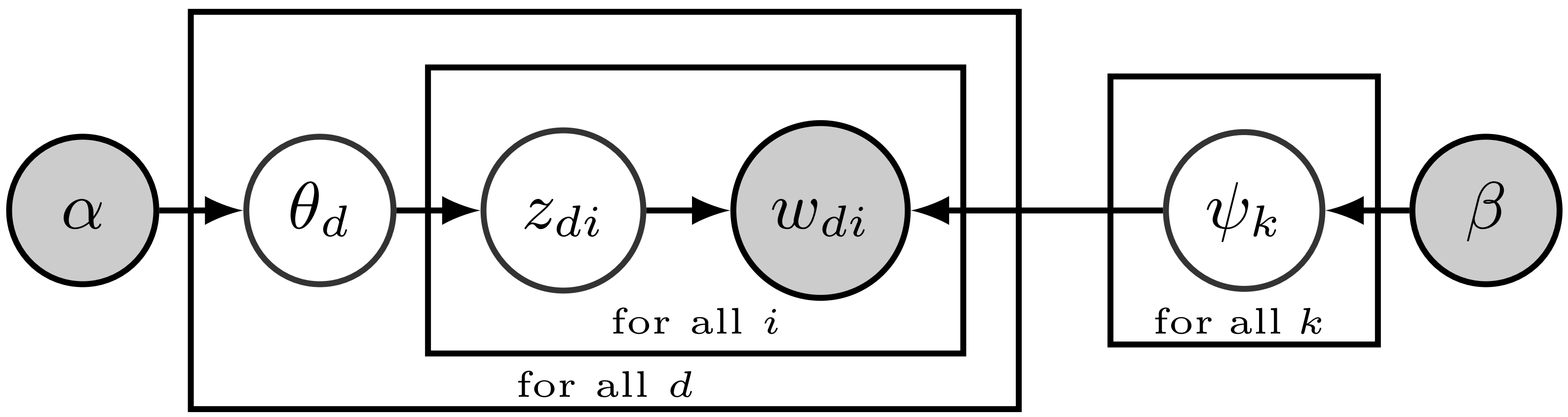


$$p(z_{di} = t, r_{di} = 0|\text{rest}) \propto \frac{\alpha_t + n_{dt}}{b_t + m_t} \frac{m_{tw} + 1 - s_{tw}}{m_{tw} + 1} \frac{S^{m_{tw}+1}_{s_{tw},a_t}}{S^{m_{tw}}_{s_{tw},a_t}}$$

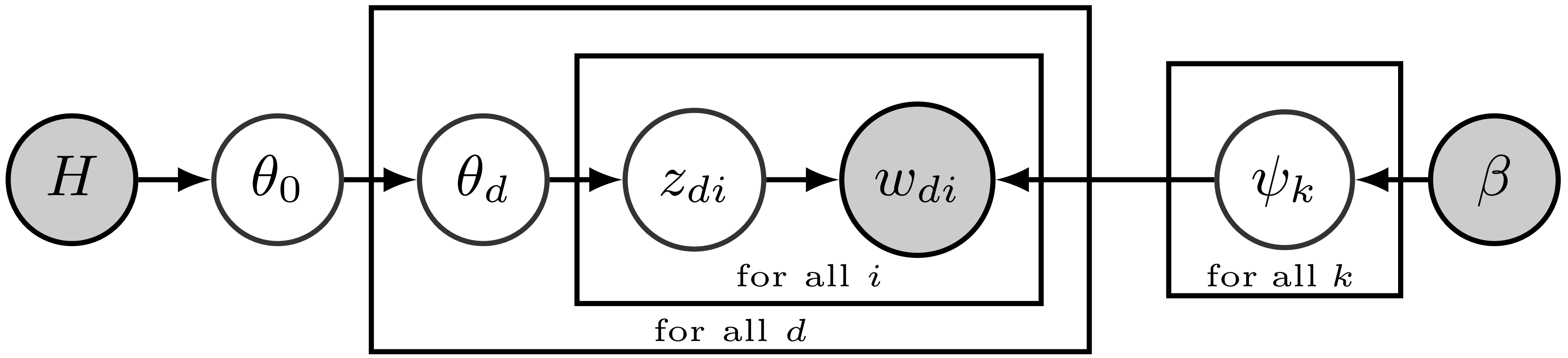
$$p(z_{di} = t, r_{di} = 1|\text{rest}) \propto (\alpha_t + n_{dt}) \frac{b_t + a_t s_t}{b_t + m_t} \frac{s_{tw} + 1}{m_{tw} + 1} \frac{\gamma + s_{tw}}{\bar{\gamma} + s_t} \frac{S^{m_{tw}+1}_{s_{tw}+1,a_t}}{S^{m_{tw}}_{s_{tw},a_t}}$$

More Models

- LDA



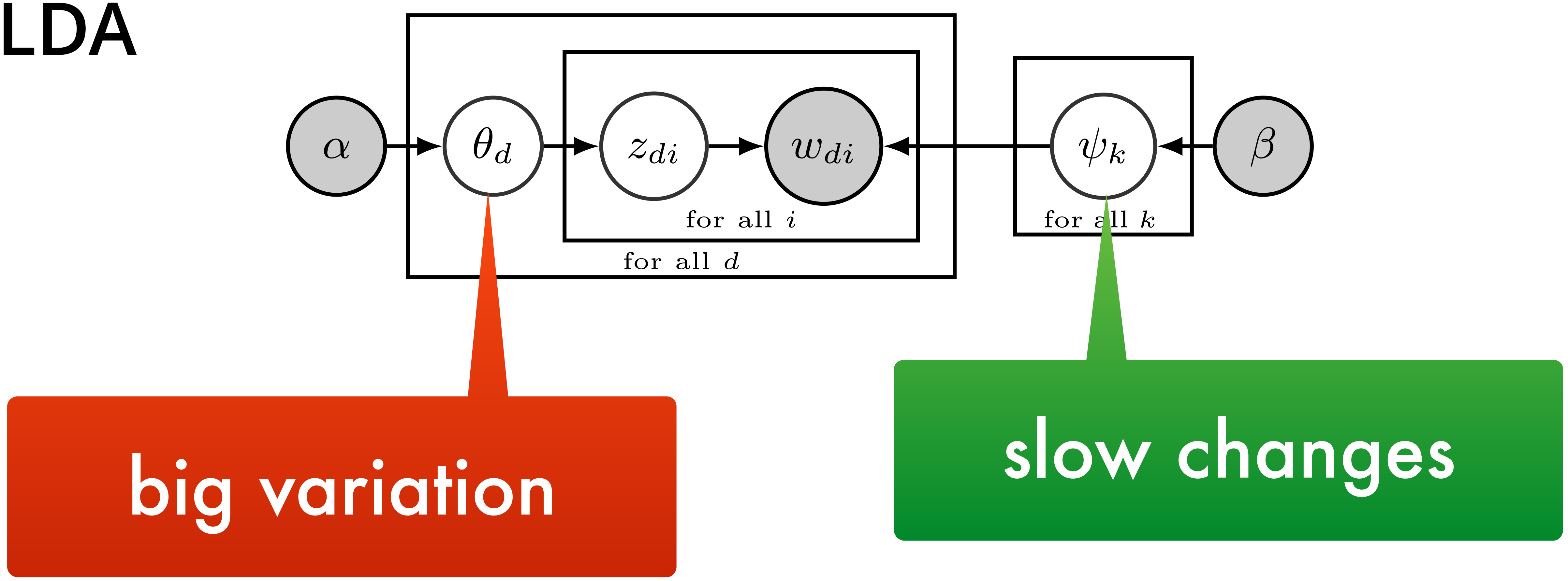
- Hierarchical-Dirichlet Process



… even more mess for topic distribution

Key Idea of the Paper

- LDA



- Approximate slowly changing distribution by fixed distribution. Use Metropolis Hastings
- Amortized O(1) time proposals

Metropolis Hastings Sampler

Lazy decomposition

- **Exploiting topic sparsity in documents**

$$\left(n^{-ij}(t,d)+\alpha_t\right)\frac{n^{-ij}(t,w)+\beta_w}{n^{-ij}(t)+\sum_w \beta_w}$$

$$=n^{-ij}(t,d)\frac{n^{-ij}(t,w)+\beta_w}{n^{-ij}(t)+\sum_w \beta_w}+\alpha_t\frac{n^{-ij}(t,w)+\beta_w}{n^{-ij}(t)+\sum_w \beta_w}$$

Sparse
O(k_d) time samples

Often dense but
slowly varying

- **Normalization costs O(k) operations!**

Google

Carnegie Mellon University

Lazy decomposition

- **Exploiting topic sparsity in documents**

$$\left(n^{-ij}(t,d) + \alpha_t\right) \frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \sum_w \beta_w}$$

$$= n^{-ij}(t,d) \frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \sum_w \beta_w} + \alpha_t \frac{n^{-ij}(t,w) + \beta_w}{n^{-ij}(t) + \sum_w \beta_w}$$

Sparse
$O(k_d)$ time samples

Approximate by
stale $q(t|w)$

- **Normalization costs $O(k_d + 1)$ operations!**

Lazy decomposition

- **Exploiting topic sparsity in documents**

$$\left(n^{-ij}(t,d)+\alpha_t\right)\frac{n^{-ij}(t,w)+\beta_w}{n^{-ij}(t)+\sum_w \beta_w}$$

$$=n^{-ij}(t,d)\frac{n^{-ij}(t,w)+\beta_w}{n^{-ij}(t)+\sum_w \beta_w}+\alpha_t\frac{n^{-ij}(t,w)+\beta_w}{n^{-ij}(t)+\sum_w \beta_w}$$

$$\approx q(t|d)+q(t|w)$$

Sparse

Static

- **Normalization costs O(k_d + 1) operations!**

Metropolis Hastings with stationary proposal distribution

- We want to sample from p but only have q
- Metropolis Hastings
  - Draw x from q(x) and accept move from x'

$$\min\left(1, \frac{p(x)}{p(x')}\frac{q(x')}{q(x)}\right)$$

  - We only need to evaluate ratios of p and q
  - This is a chain. It mixes rapidly in experiments.

Application to Topic Models

- Recall - we split topic probability

$$q(t) \propto q(t|d) + q(t|w)$$

k_d Sparse

Dense but static

- Dense part has normalization precomputed
- Sparse part can easily be normalized
- Sample from q(t) and
evaluate p(t|w,d) only for the draws

In a nutshell

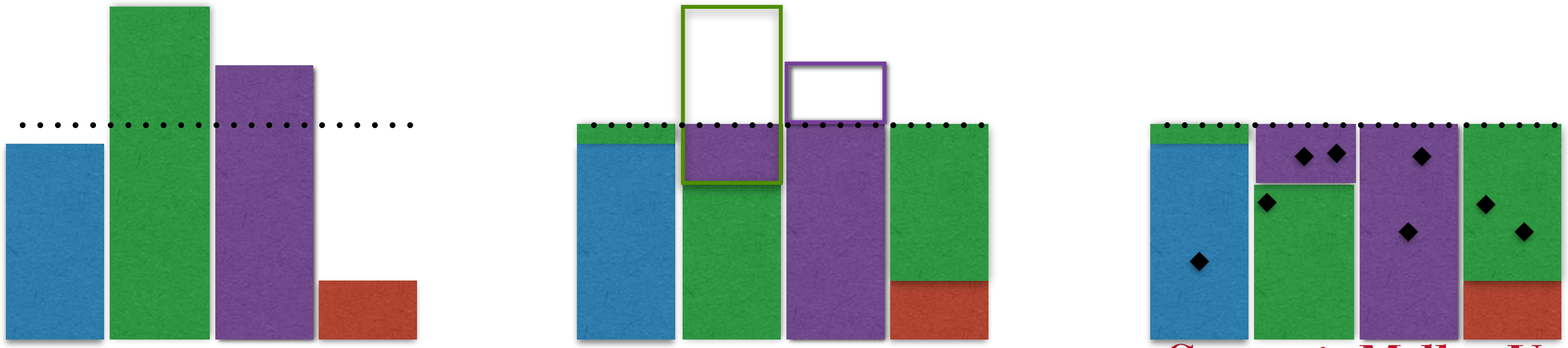
$$q(t) \propto q(t|d) + q(t|w)$$

- Sparse part for document (topics, topic hierarchy, etc.) Evaluate this exactly
- Dense part for generative model (language, images, …) Approximate this by stale model
- Metropolis Hastings sampler to correct
- Need fast way to draw from stale model

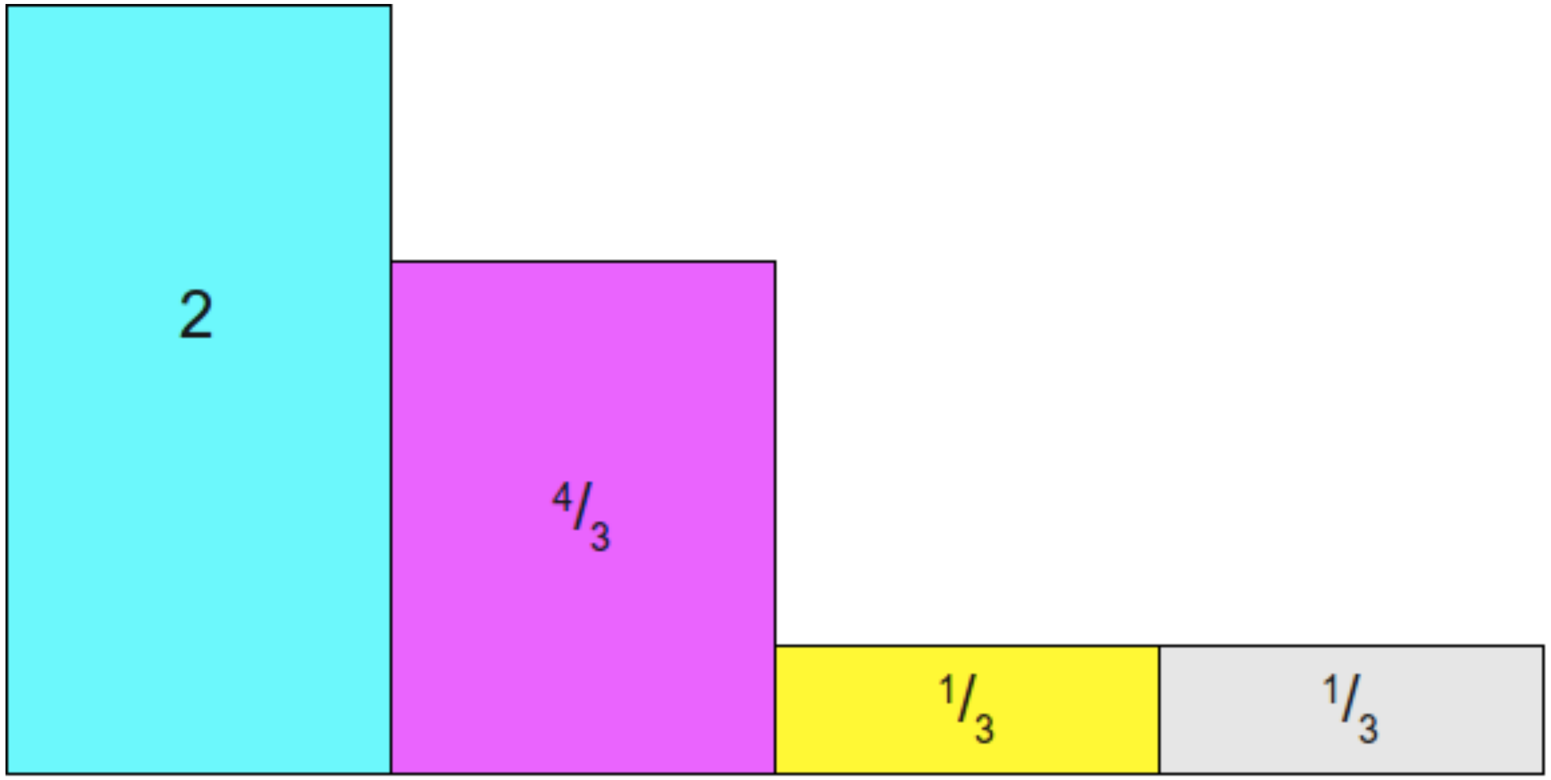
Alias Sampling

Walker's Alias Method

- Draw from discrete distribution in O(1) time
- Requires O(n) preprocessing
  - Group all x with n p(x) < 1 into L (rest in H)
  - Fill each of the small ones up by stealing from H. This yields (i,j, p(i)) triples.
  - Draw from uniform over n, then from p(i)



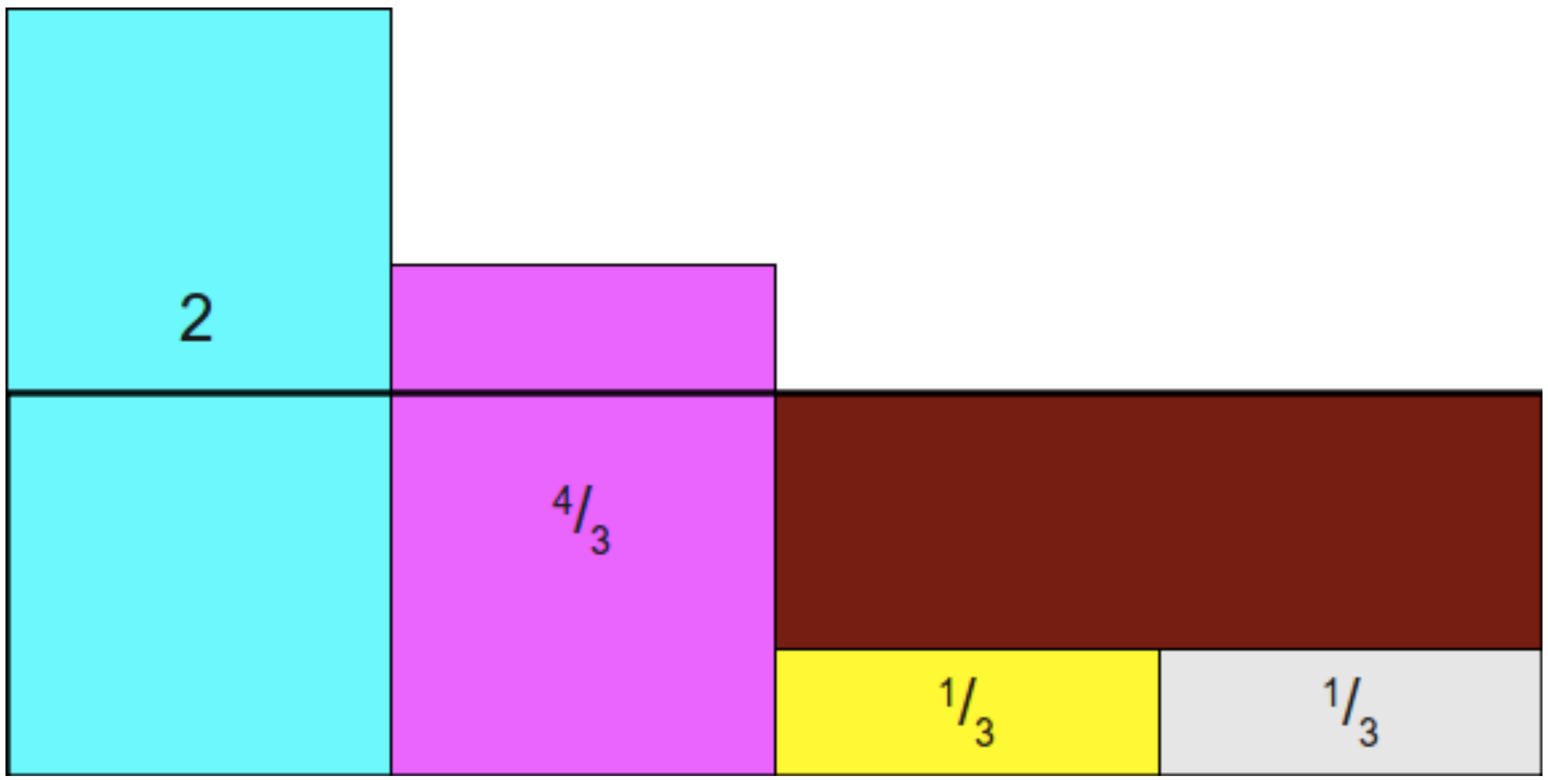
Probability distribution



Courtesy of keithschwartz.com

Google Carnegie Mellon University

Probability distribution

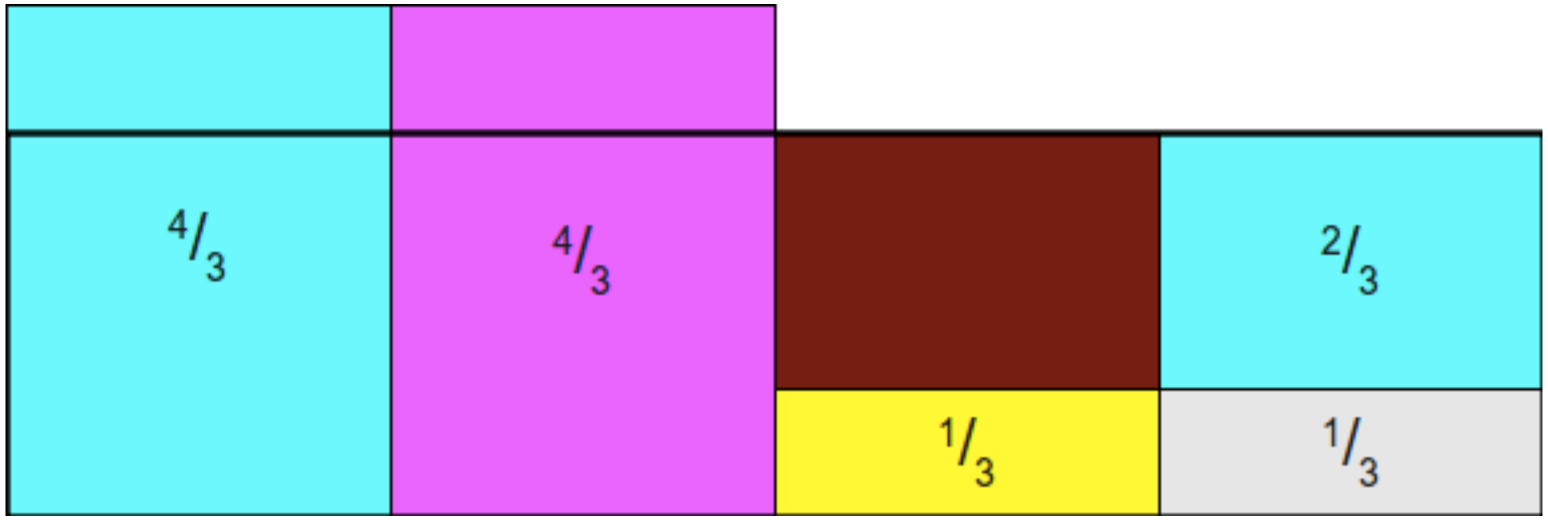


Splitting

Courtesy of keithschwartz.com

Google Carnegie Mellon University

Probability distribution



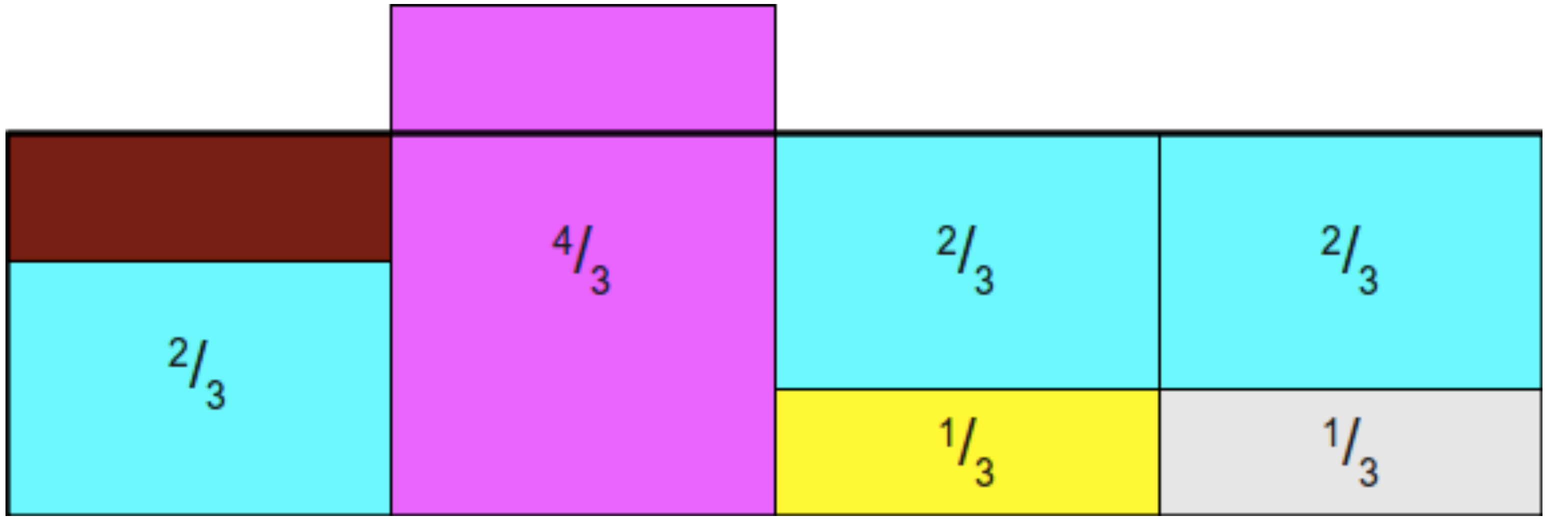
Filling up (4) with (1)

Courtesy of keithschwartz.com

Google

Carnegie Mellon University

Probability distribution

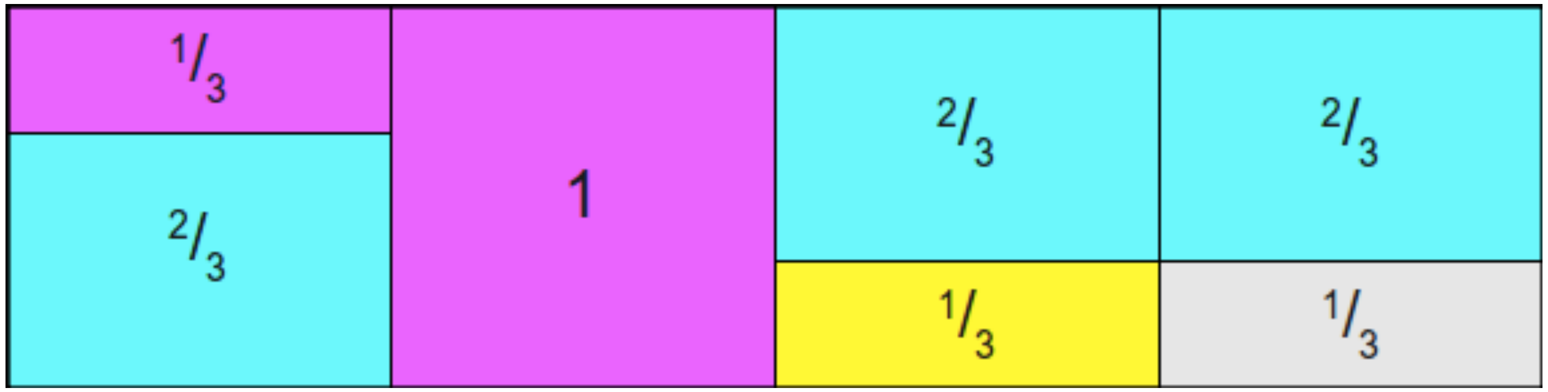


Filling up (3) with (1)

Courtesy of keithschwartz.com

Google
Carnegie Mellon University

Probability distribution



Filling up (1) with (2)

Courtesy of keithschwartz.com

Google
Carnegie Mellon University

Metropolis-Hastings-Walker

- Conditional topic probability

$$q(t) \propto q(t|d) + q(t|w)$$

k_d Sparse

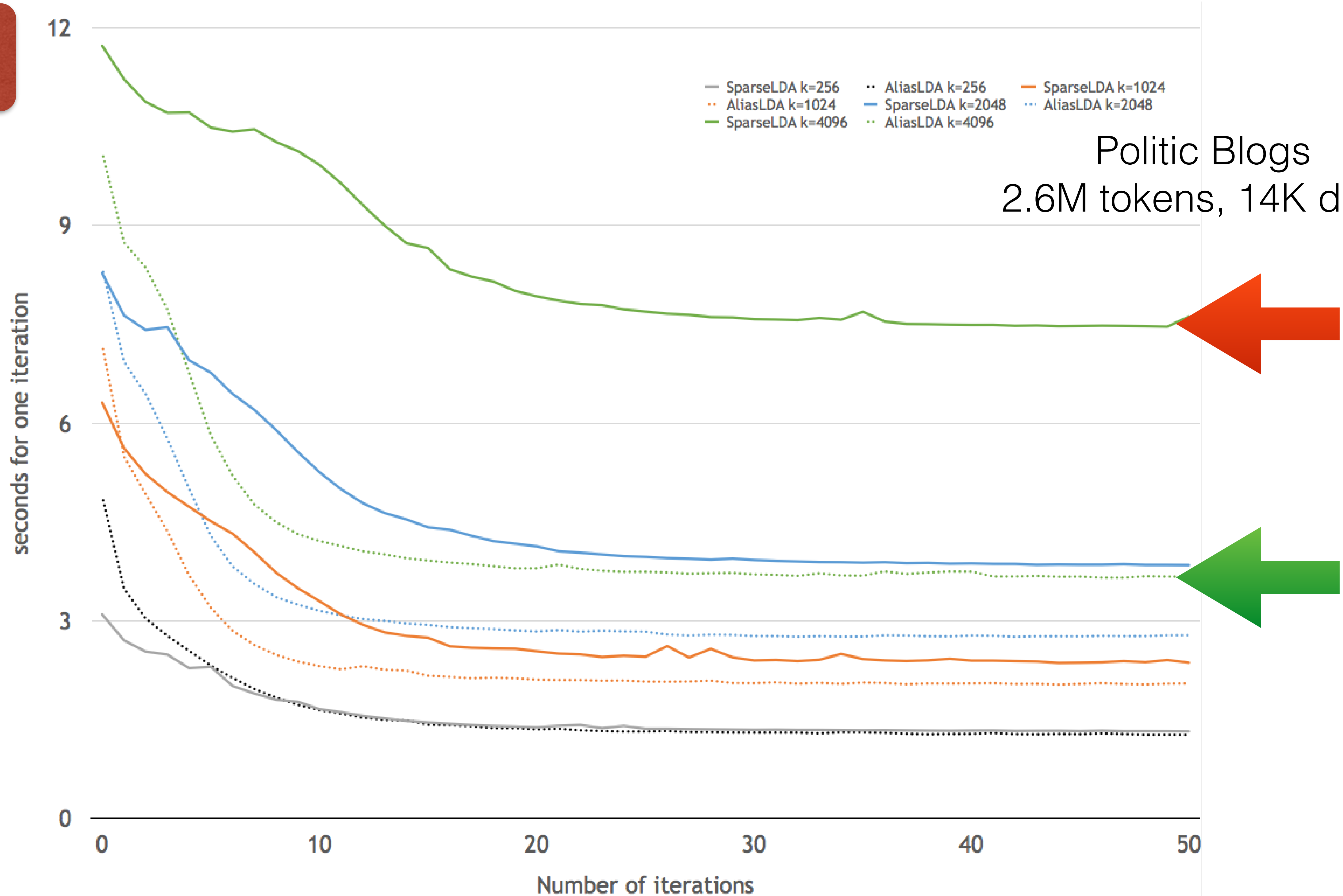
Dense but static

- Use Walker's method to draw from q(t|w)
- After k draws from q(t|w) recompute with current value
- Amortized $O(1 + k_d)$ sampler

Experiments

LDA: Varying the number of topics (4k)

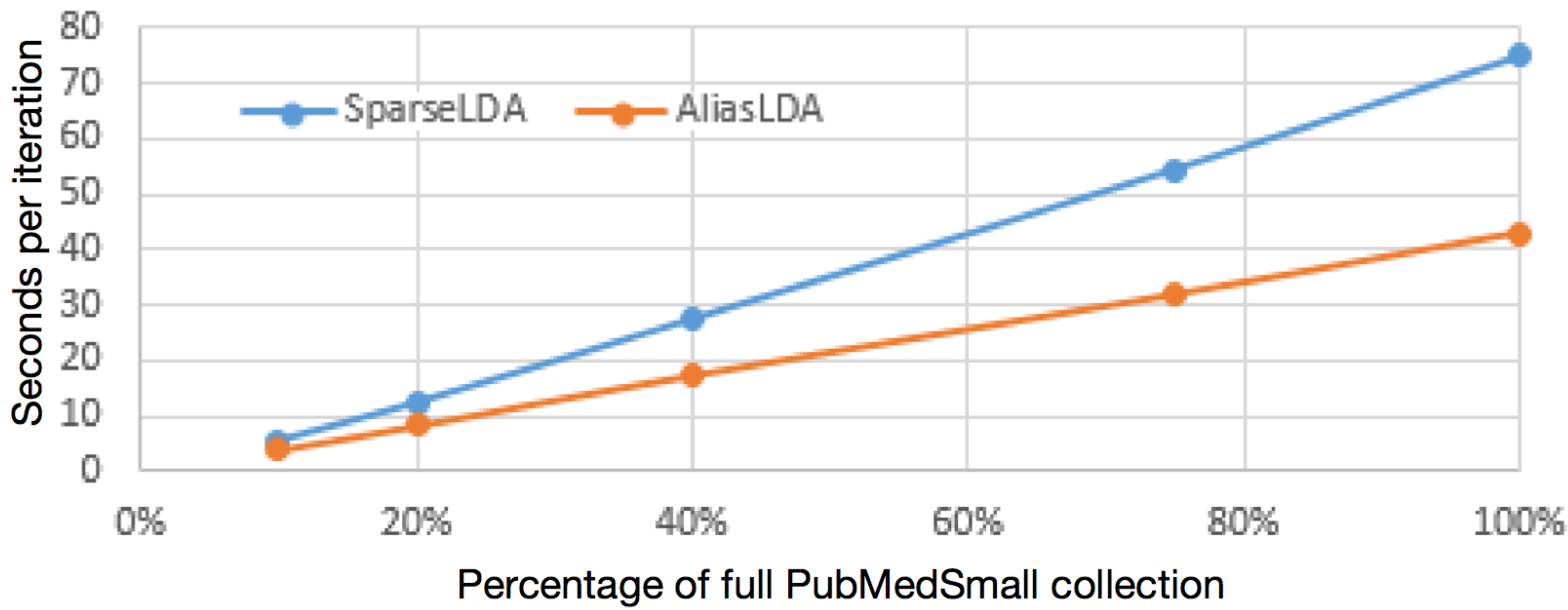
speed



Politic Blogs
2.6M tokens, 14K docs

LDA: Varying data size

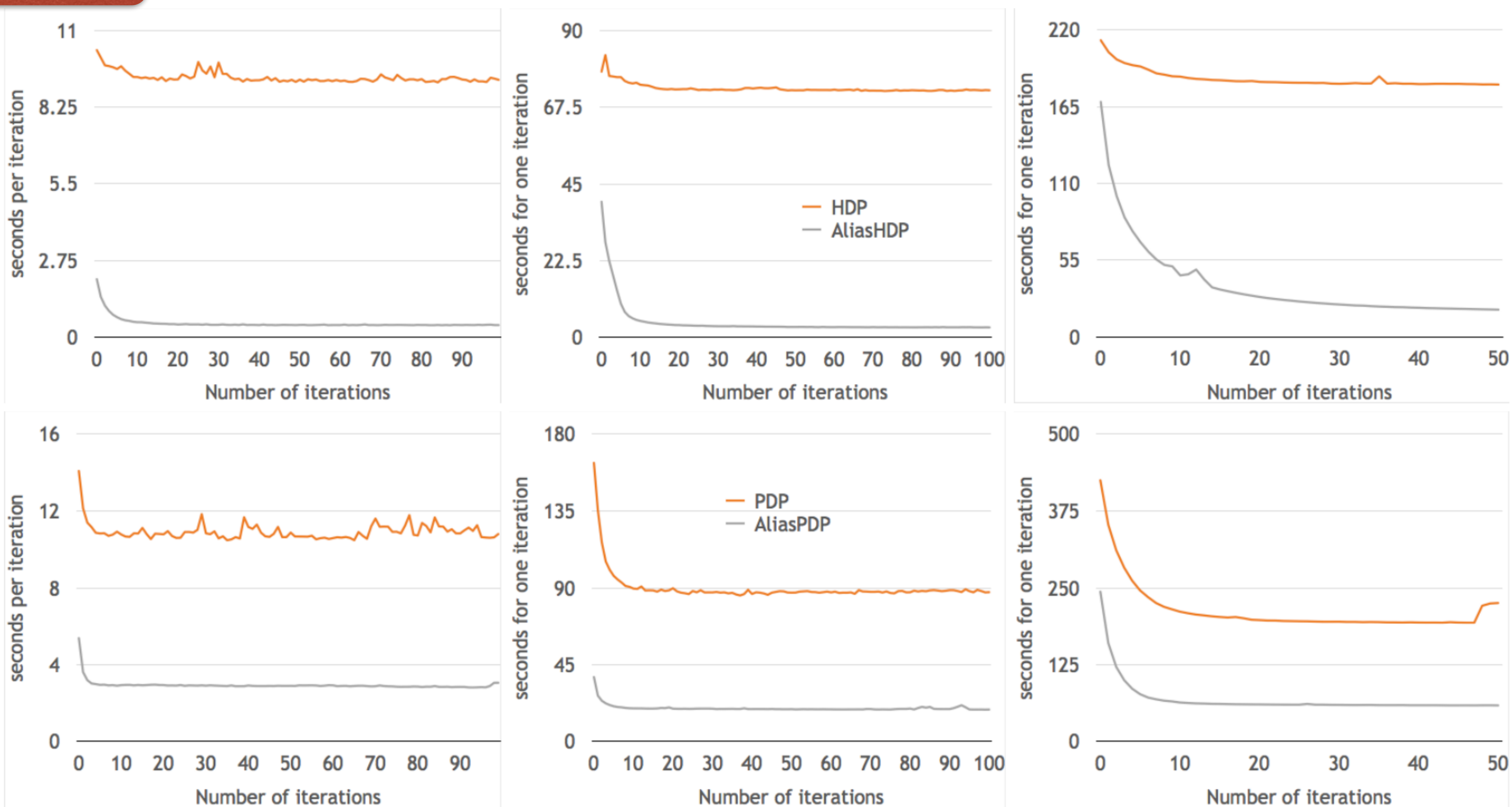
speed



HDP & PDP

speed

RS (321K tokens) | GPOL (2.6M tokens) | Enron (6M tokens)



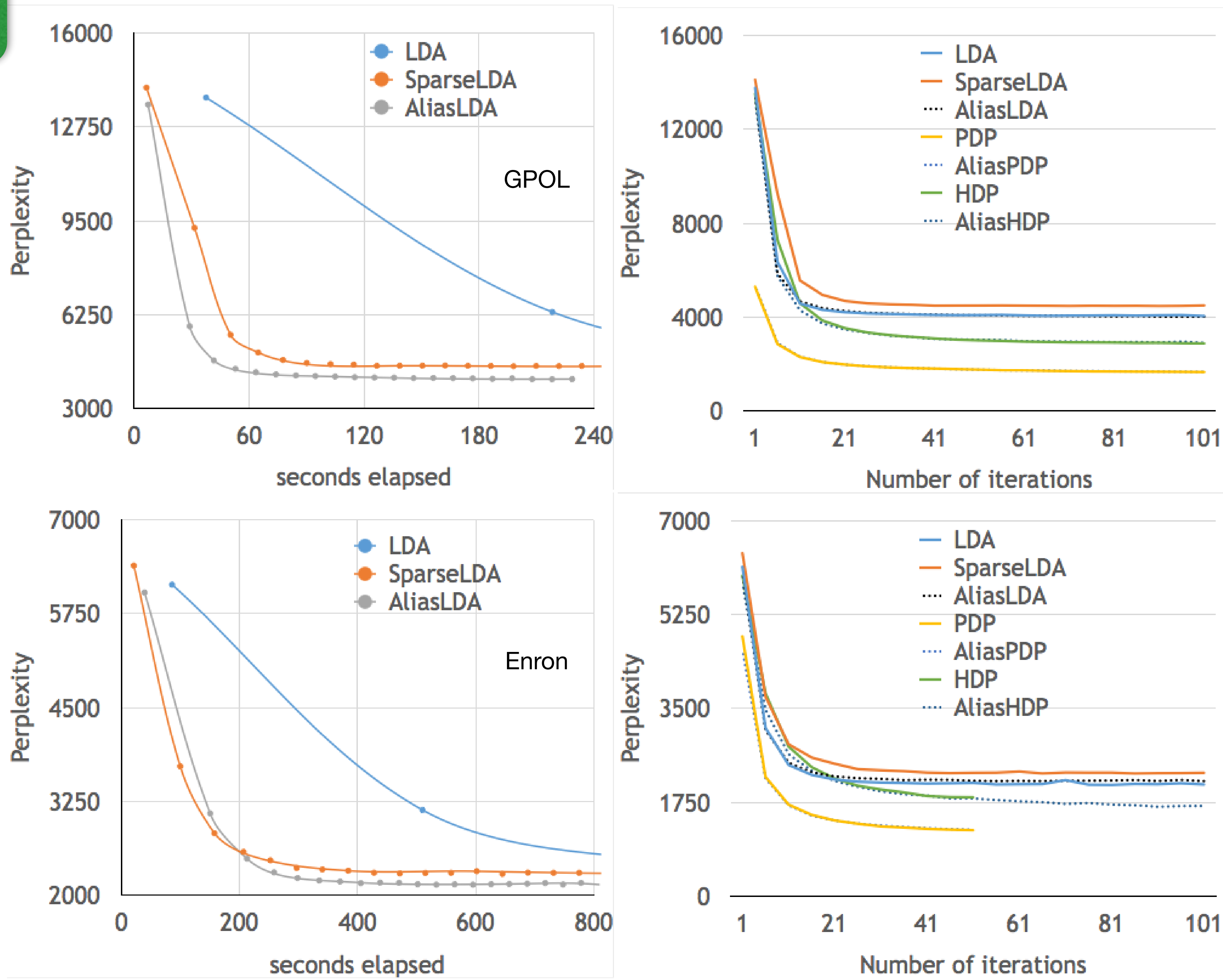
Google

Carnegie Mellon University

Perplexity

quality

Perplexity vs. Runtime | Perplexity vs. Iterations



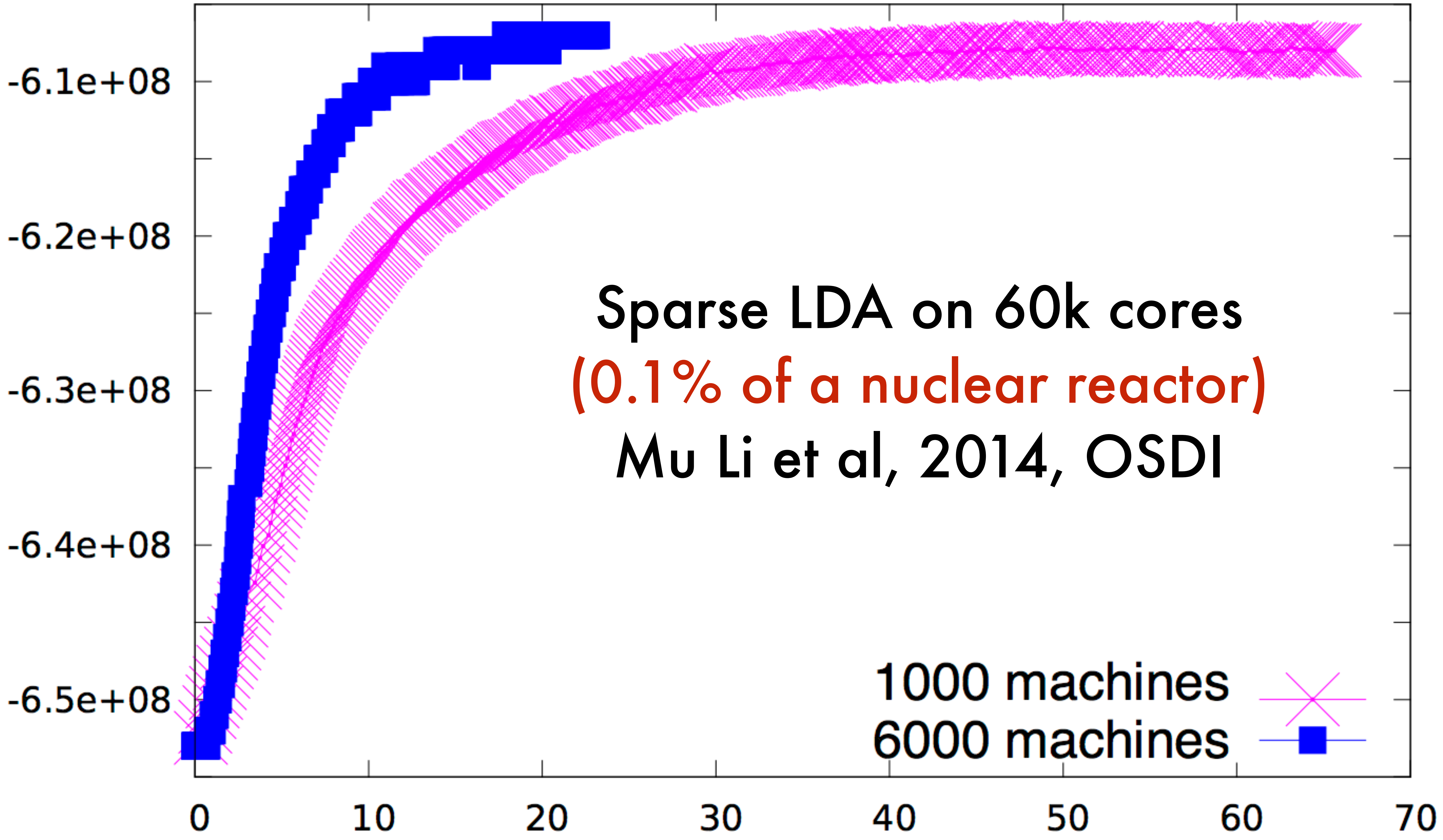
Google

Carnegie Mellon University

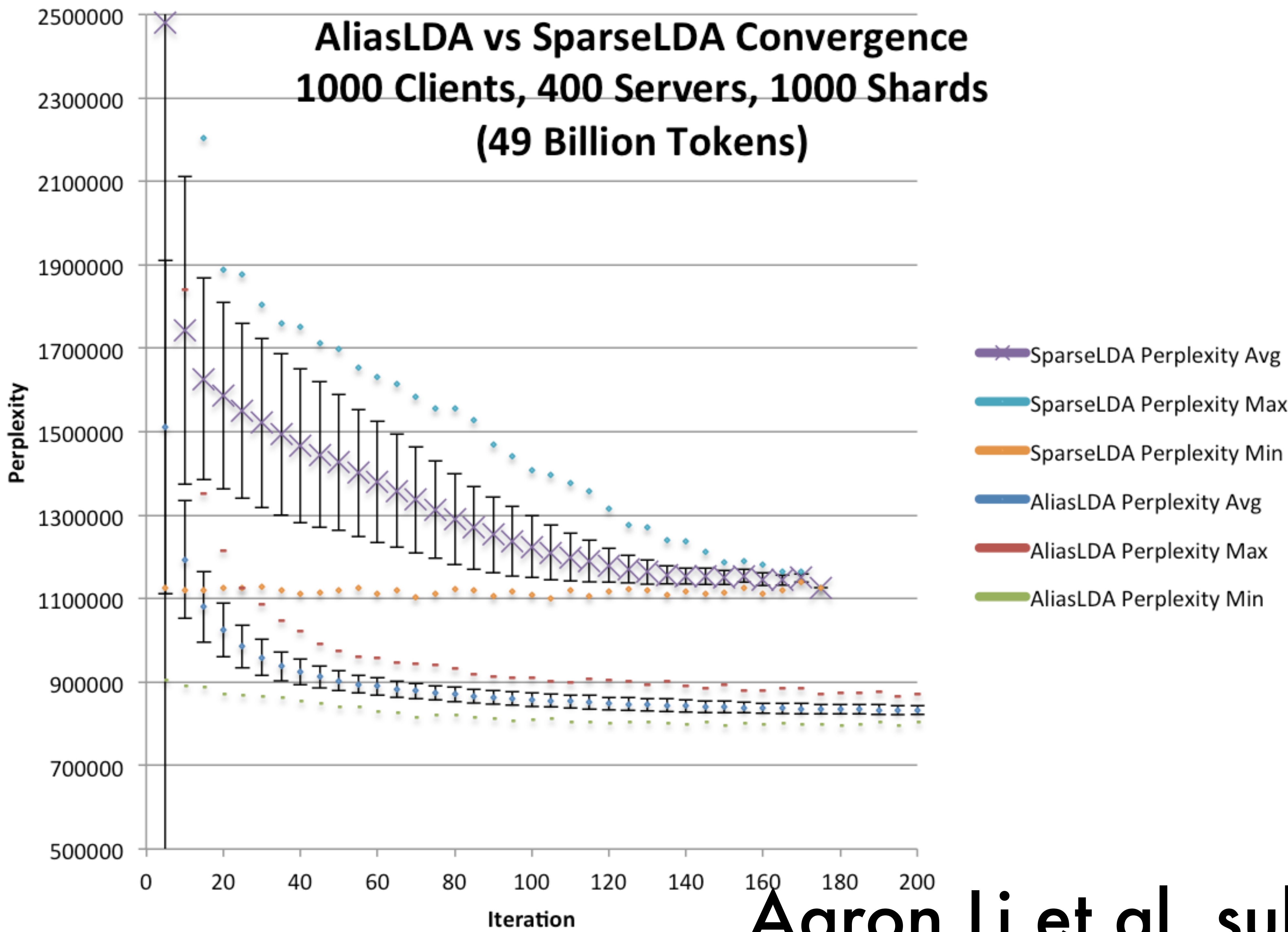
Summary

- Extends Sparse LDA concept of Yao et al.'09
  - Works for any sparse document model
  - Useful for many emissions models (Pitman Yor, Gaussians, etc.)
- Metropolis-Hastings-Walker
  - MH proposals on stale distribution
  - Recompute proposal after k draws for O(1)
- Fastest LDA sampler by a large margin

And now in parallel



Saving Nuclear Power Plants



Aaron Li et al, submitted

Google

Carnegie Mellon University