YAHOO!®

Scalable Machine Learning

10. Distributed Inference and Applications

Alex Smola
Yahoo! Research and ANU

http://alex.smola.org/teaching/berkeley2012
Stat 260 SP 12

Outline

- Latent Dirichlet Allocation
  - Basic model
  - Sampling and efficient implementation
- Parallel Inference
  - Problem templates
  - Solution templates
- Applications

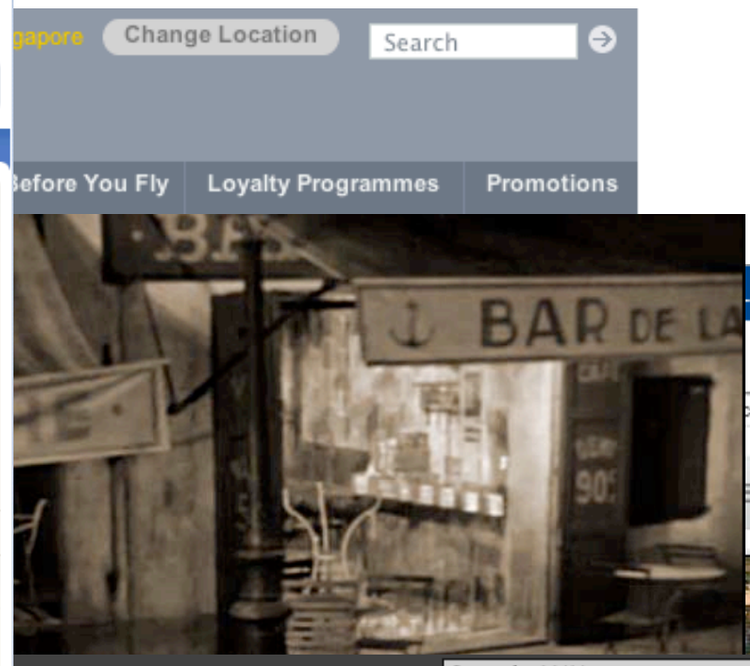

MAGIC Etch A Sketch® SCREEN
Latent Dirichlet
Allocation

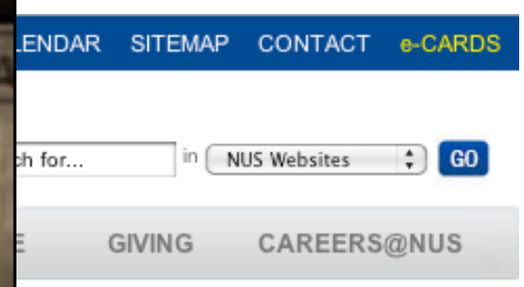
Grouping objects

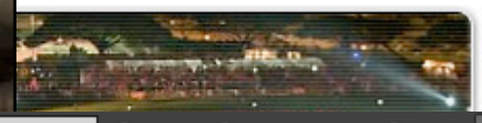
Grouping objects

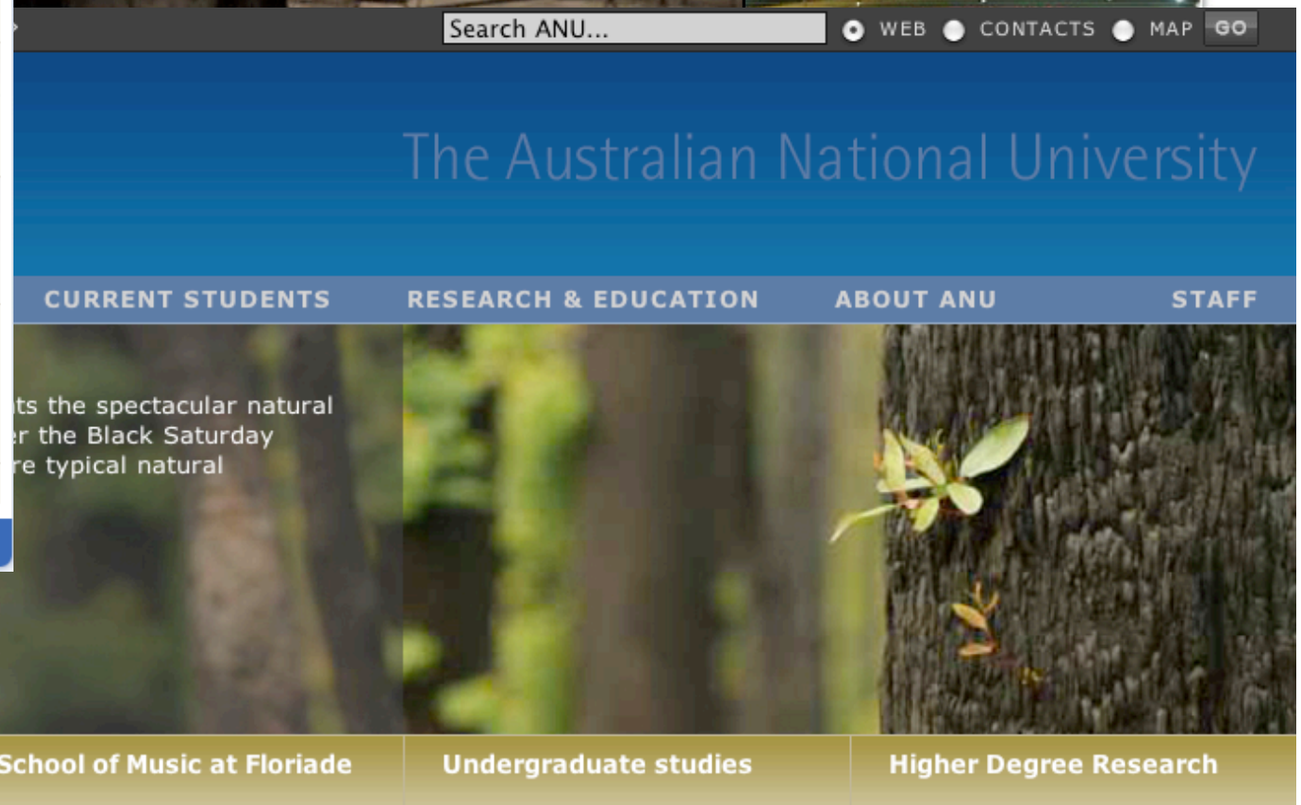
Singapore

Grouping objects

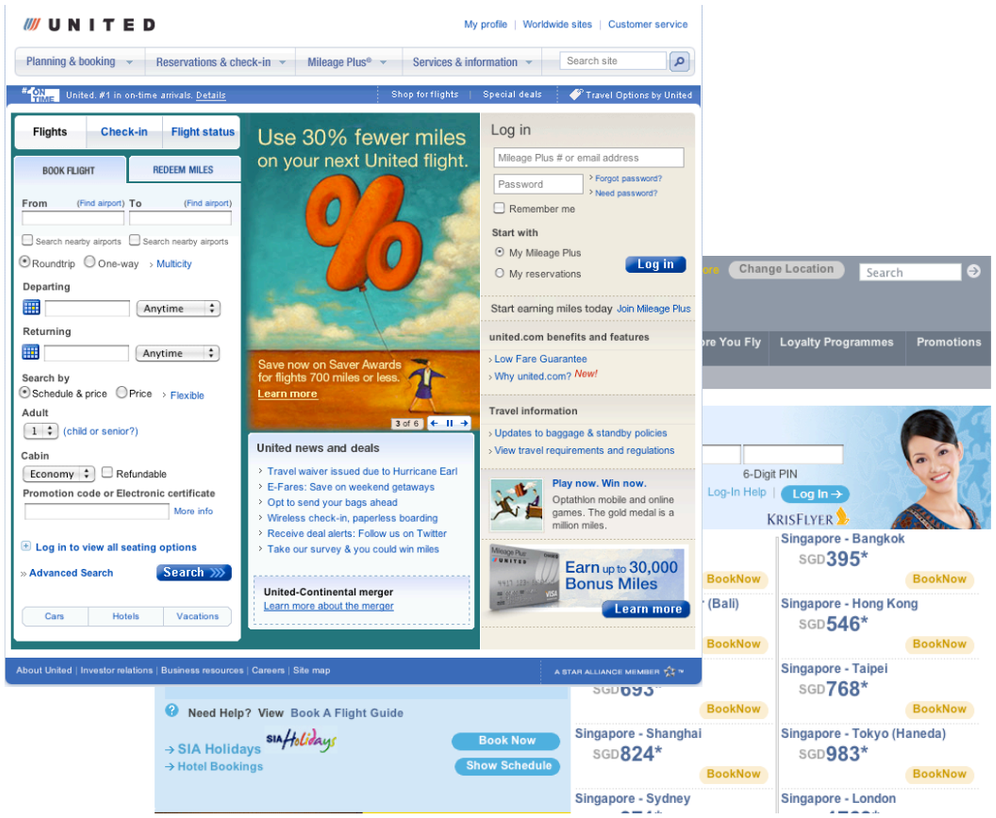






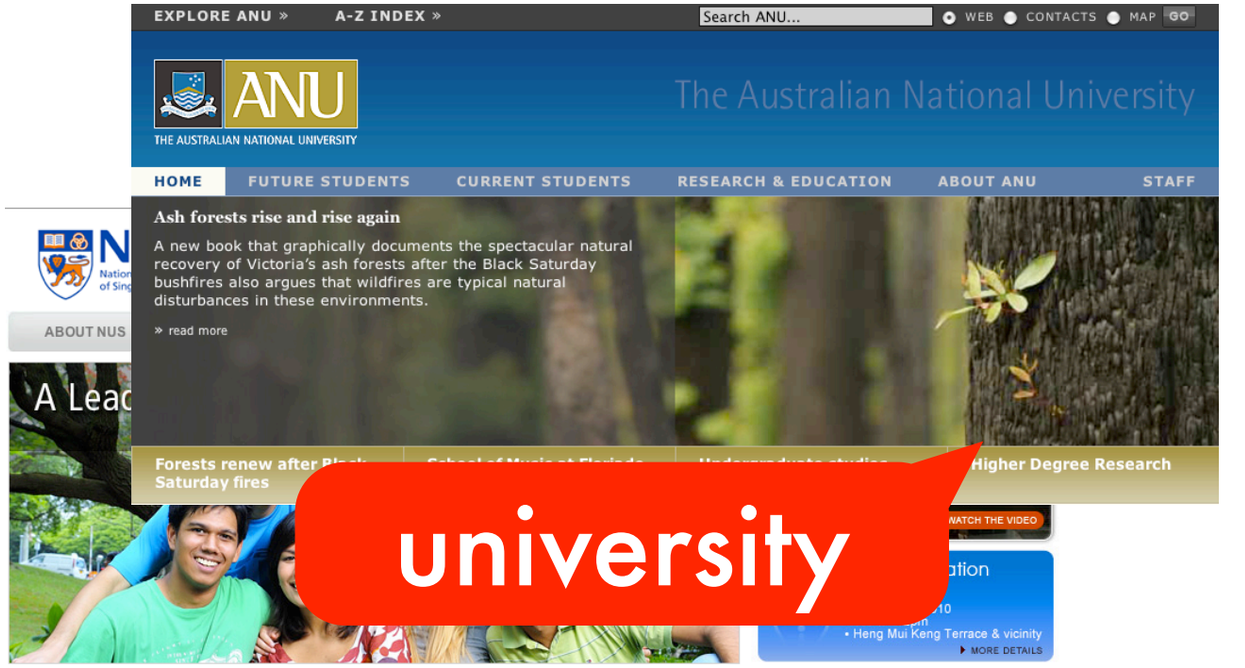


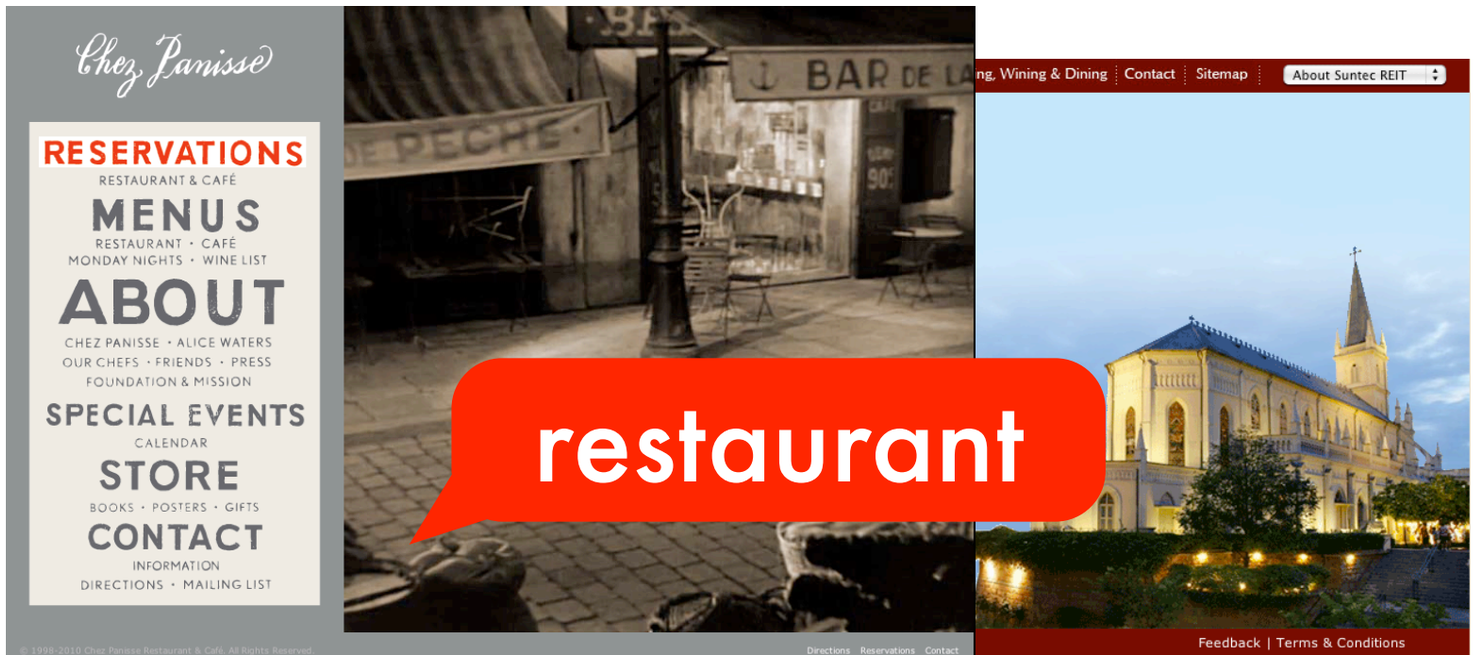
Grouping objects



YAHOO!

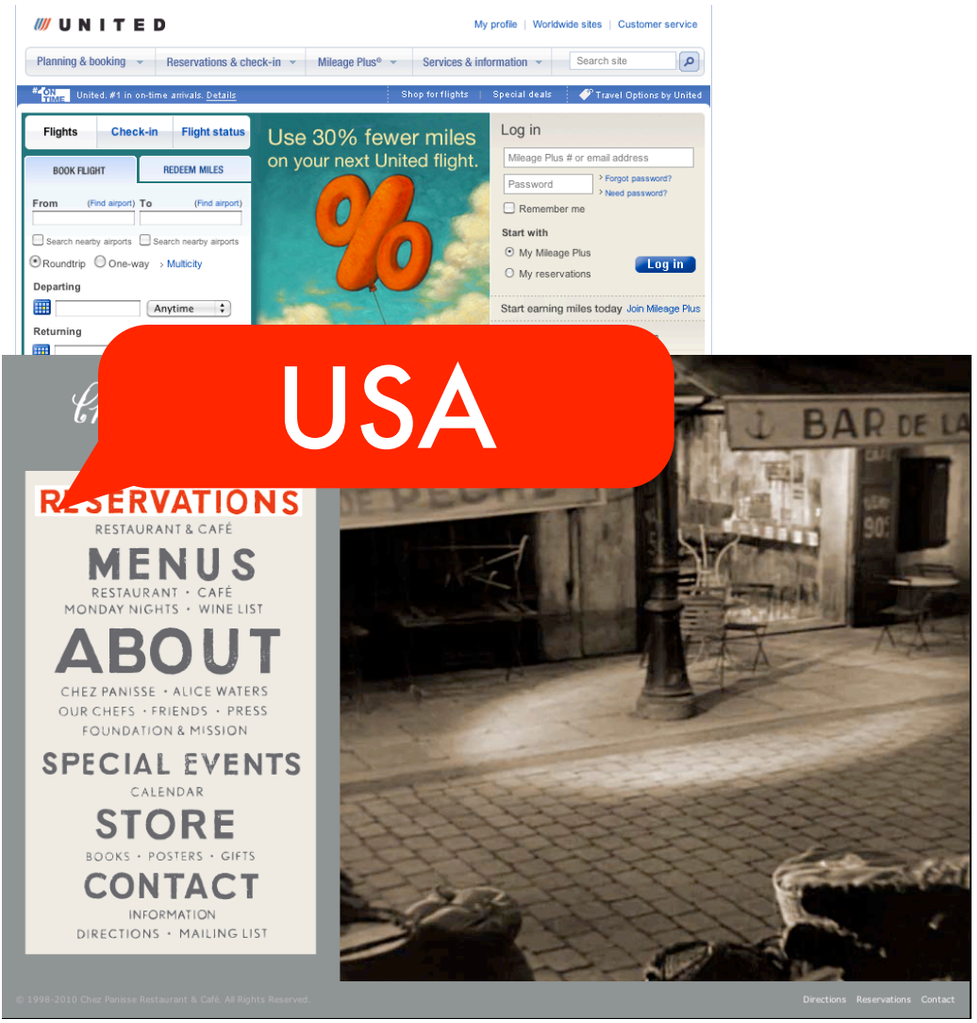
Grouping objects

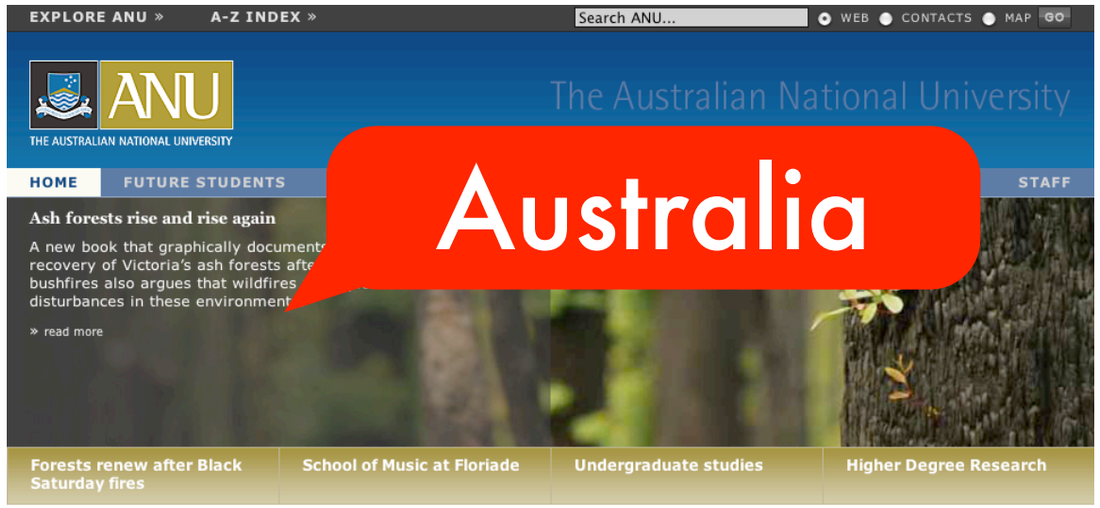




YAHOO!

Grouping objects

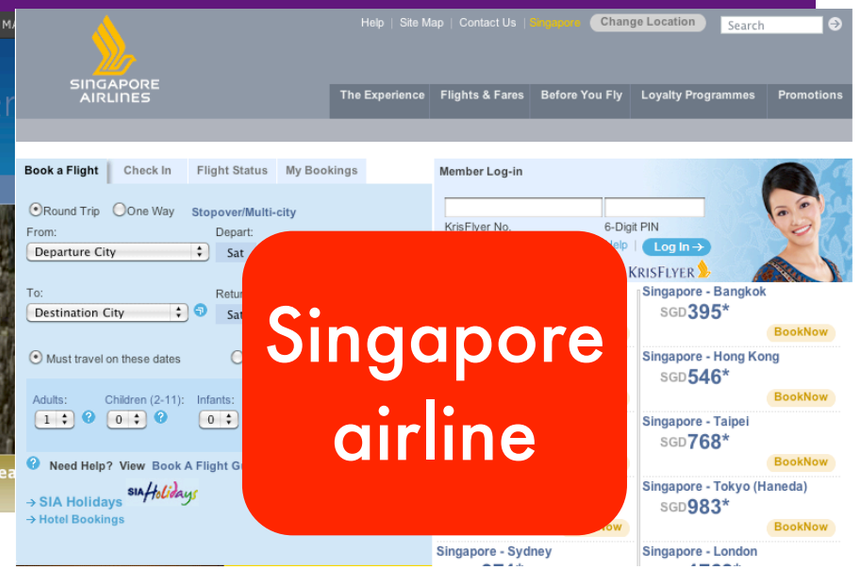




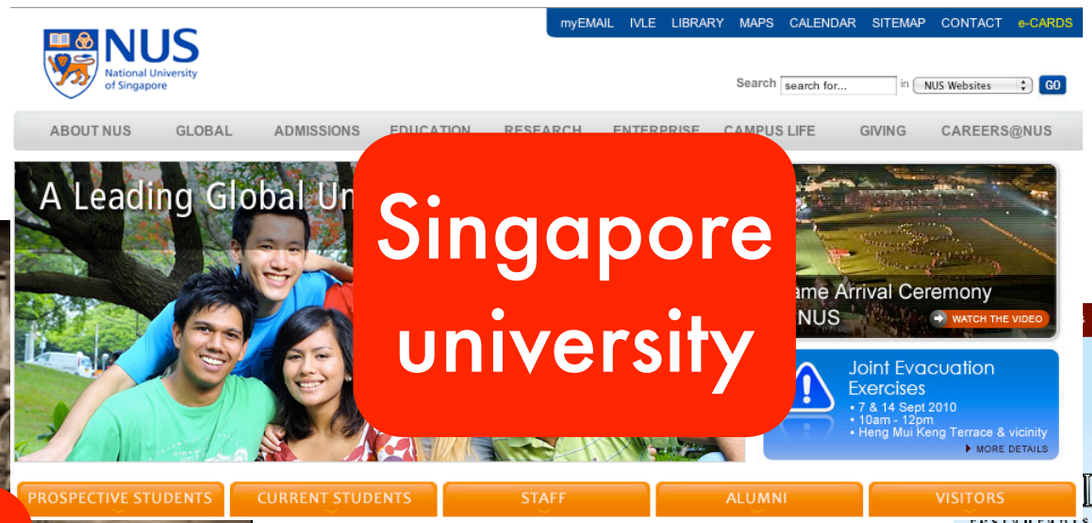
Topic Models

USA airline

Australia university



Singapore airline



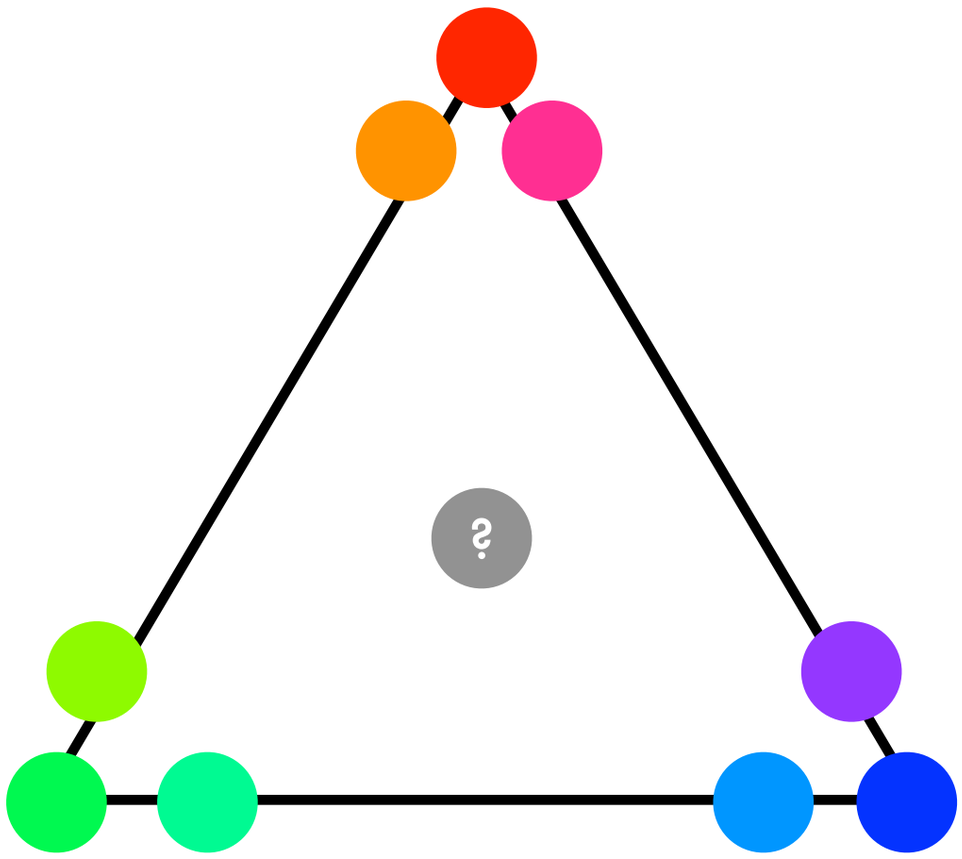
Singapore university

USA food

Singapore food

Clustering & Topic Models

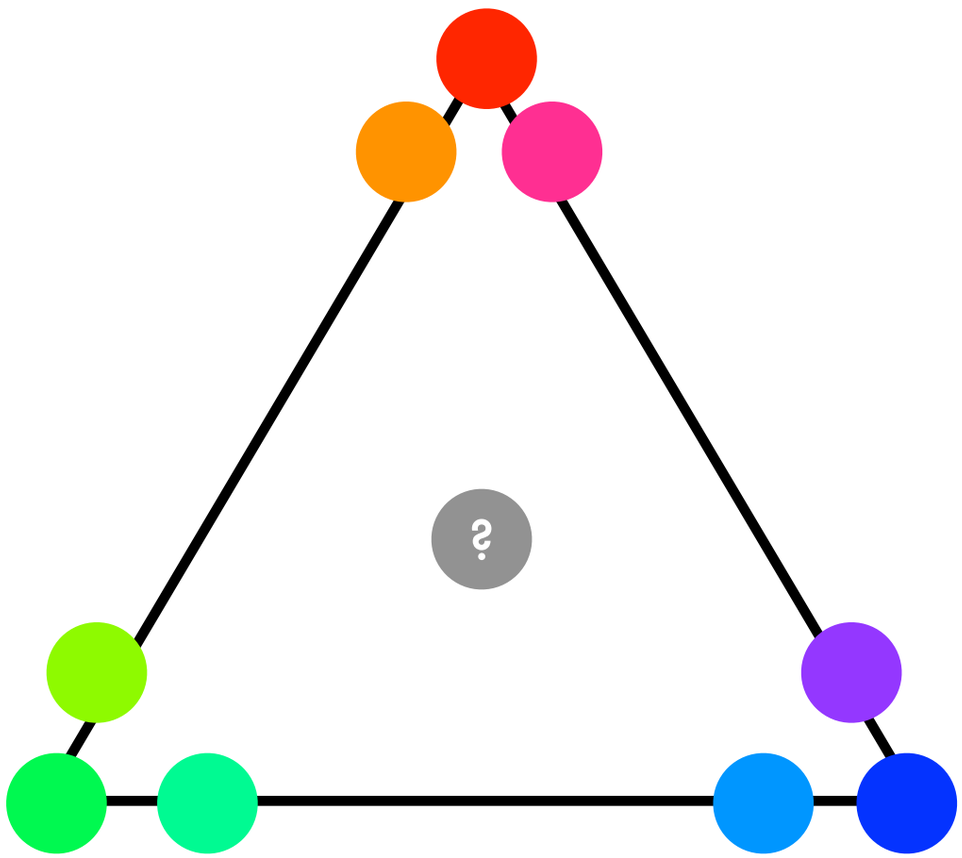
Clustering



group objects
by prototypes

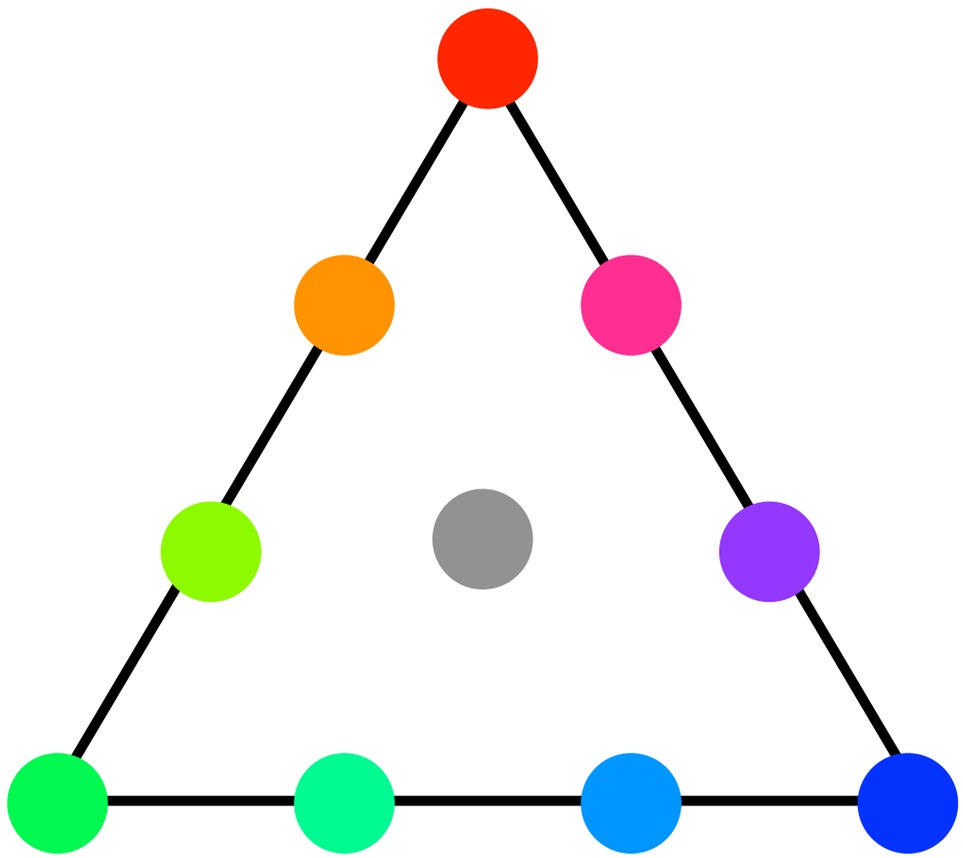
Clustering & Topic Models

Clustering



group objects
by prototypes

Topics

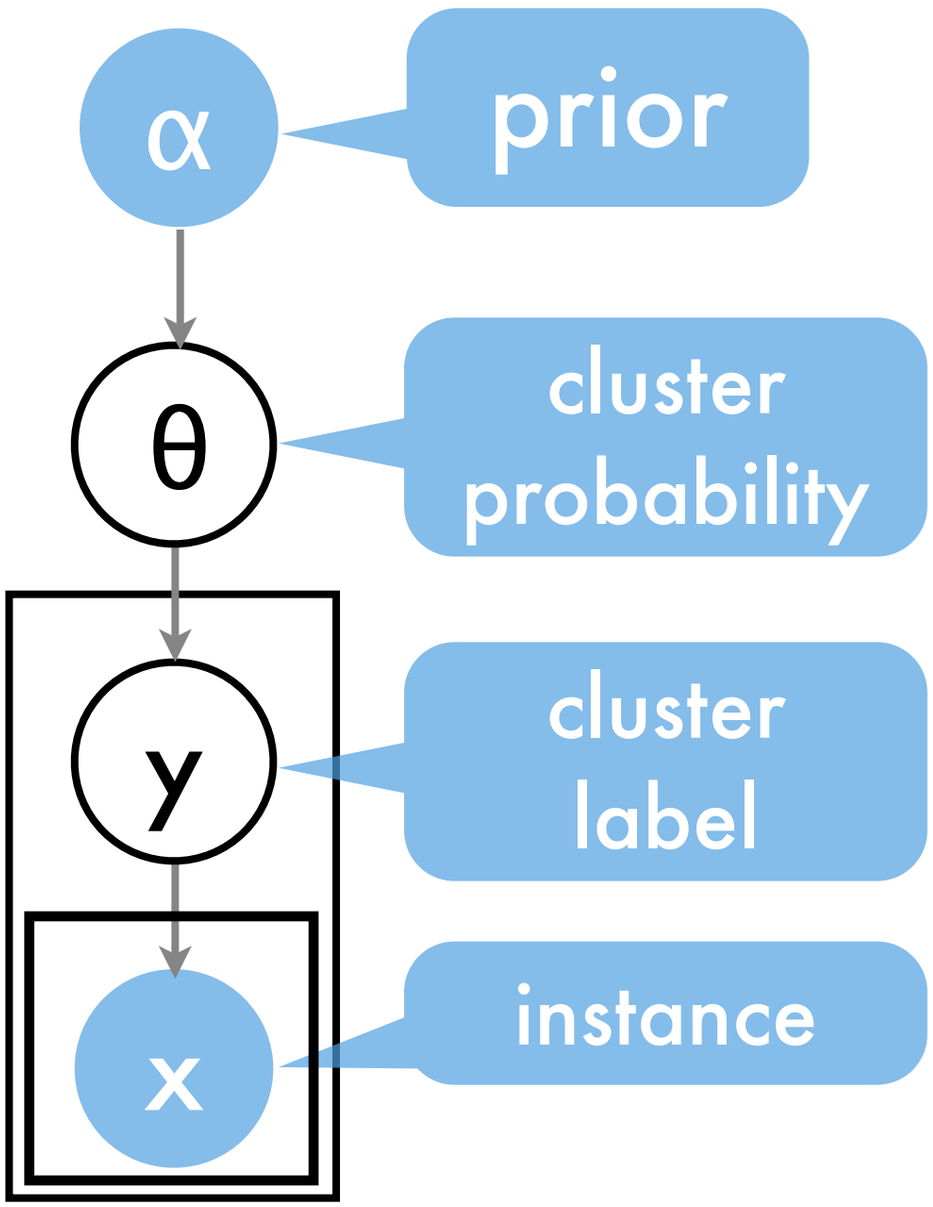


decompose objects
into prototypes

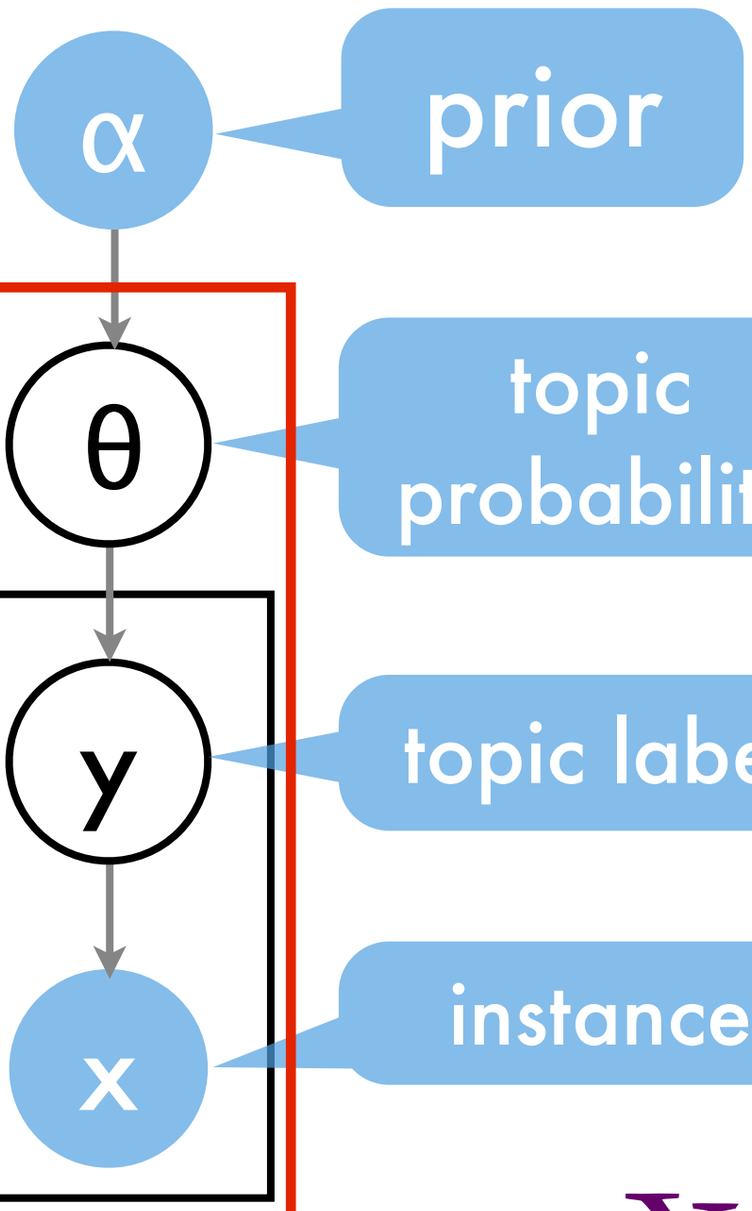
YAHOO!

Clustering & Topic Models

clustering

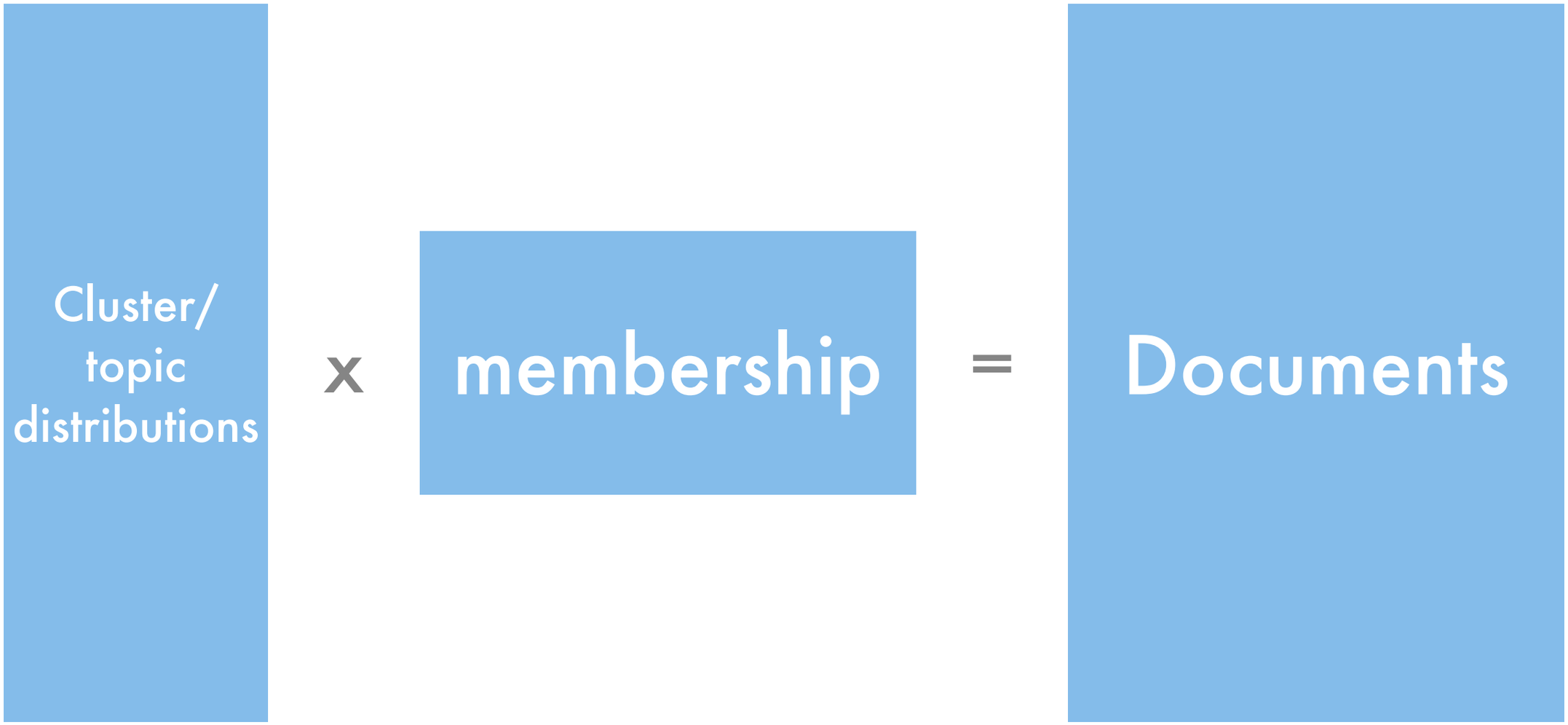


Latent Dirichlet Allocation



YAHOO!

Clustering & Topic Models



clustering: (0, 1) matrix
topic model: stochastic matrix
LSI: arbitrary matrices

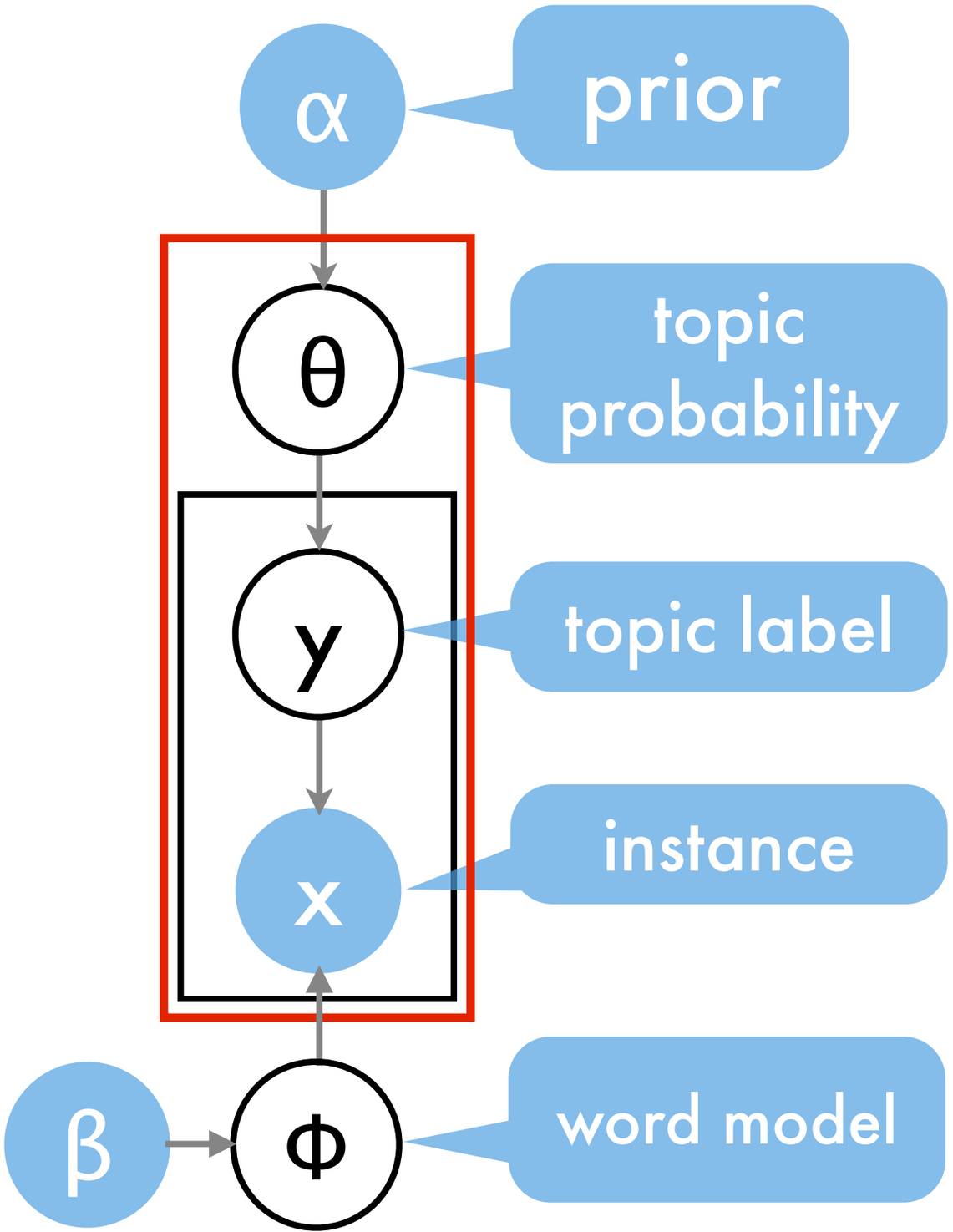
YAHOO!

Topics in text

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.

Latent Dirichlet Allocation; Blei, Ng, Jordan, JMLR 2003

Mathematical Details



- Dirichlet prior

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

- Topic label

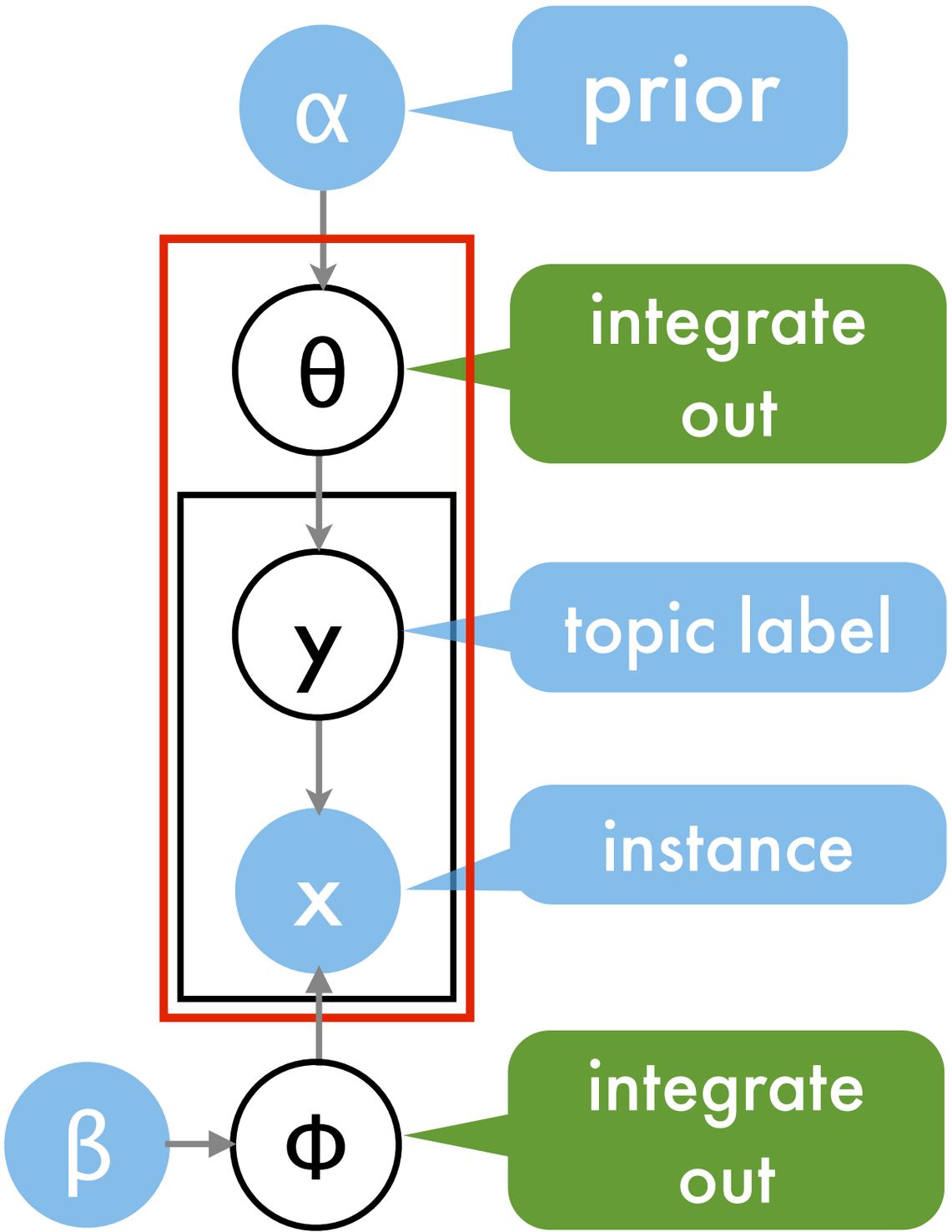
$$p(y|\theta) = \theta_y$$

- Word probability

$$p(x|\phi, y) = \phi_{x,y}$$

- Word model via Dirichlet prior

Collapsed Inference



- Dirichlet prior

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

- Topic label

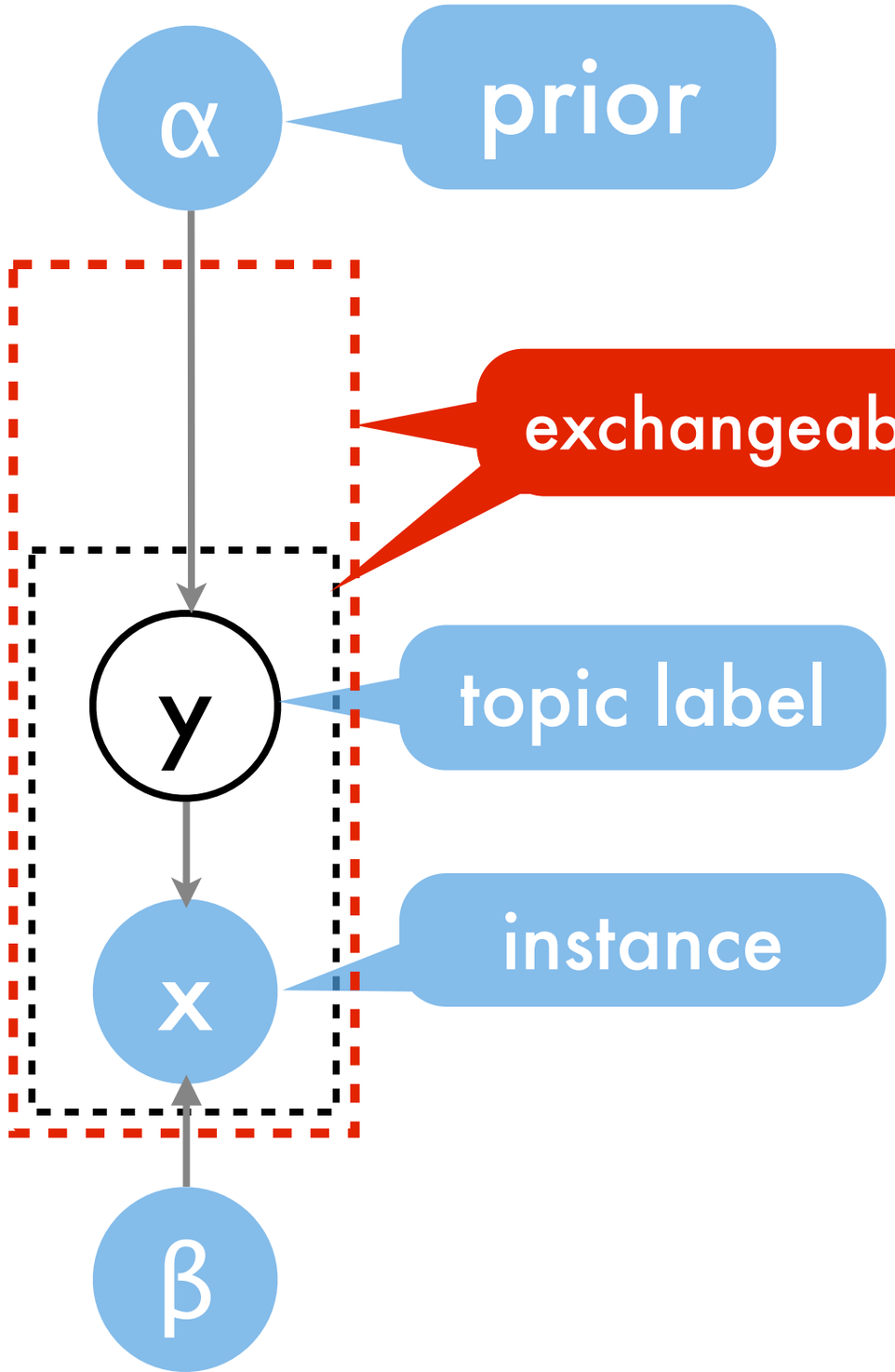
$$p(y|\theta) = \theta_y$$

- Word probability

$$p(x|\phi, y) = \phi_{x,y}$$

- Word model via Dirichlet prior

Collapsed Inference



- Collapsed Dirichlet for topics

$$p(y_{d1}, \ldots y_{dm}|\alpha)$$

- Collapsed Dirichlet for words

$$\prod_{j=1}^{k} p(W|y_{id} = j|\beta)$$

(need to partition between topics)

Recall - collapsing exponentials

- Conjugate priors

$$p(\theta) \propto p(X_{\text{fake}}|\theta)$$

Hence we know how to compute normalization

- **Prediction**

$$p(x|X) = \int p(x|\theta)p(\theta|X)d\theta$$

$$\propto \int p(x|\theta)p(X|\theta)p(X_{\text{fake}}|\theta)d\theta$$

$$= \int p(\{x\} \cup X \cup X_{\text{fake}}|\theta)d\theta$$

(Beta, binomial)
(Dirichlet, multinomial)
(Gamma, Poisson)
(Wishart, Gauss)

look up closed form expansions

http://en.wikipedia.org/wiki/Exponential_family

Collapsing exponentials

- Conjugate prior

$$p(\theta) = \exp\left(m_0 \left\langle \mu_0, \theta \right\rangle - m_0 g(\theta) - h(m_0\mu_0, m_0)\right)$$

- Posterior

$$\begin{aligned} p(\theta|X) &\propto \prod_{i=1}^{m} p(x_i|\theta)p(\theta) \\ &= \exp\left(\left\langle m_0\mu_0 + \sum_{i=1}^{m}\phi(x_i), \theta \right\rangle - (m_0+m)g(\theta) - h(m_0\mu_0, m_0)\right) \end{aligned}$$

- Computing the normalization yields

$$p(X|\mu_0, m_0) = \exp\left(h(m_0\mu_0 + m\mu[X], m_0 + m) - h(m_0\mu_0, m_0)\right)$$

Applyication to Dirichlet prior

- Normalization

$$p(X|\mu_0, m_0) = \exp\left(h(m_0\mu_0 + m\mu[X], m_0 + m) - h(m_0\mu_0, m_0)\right)$$

- In mean (not natural) parameters ...

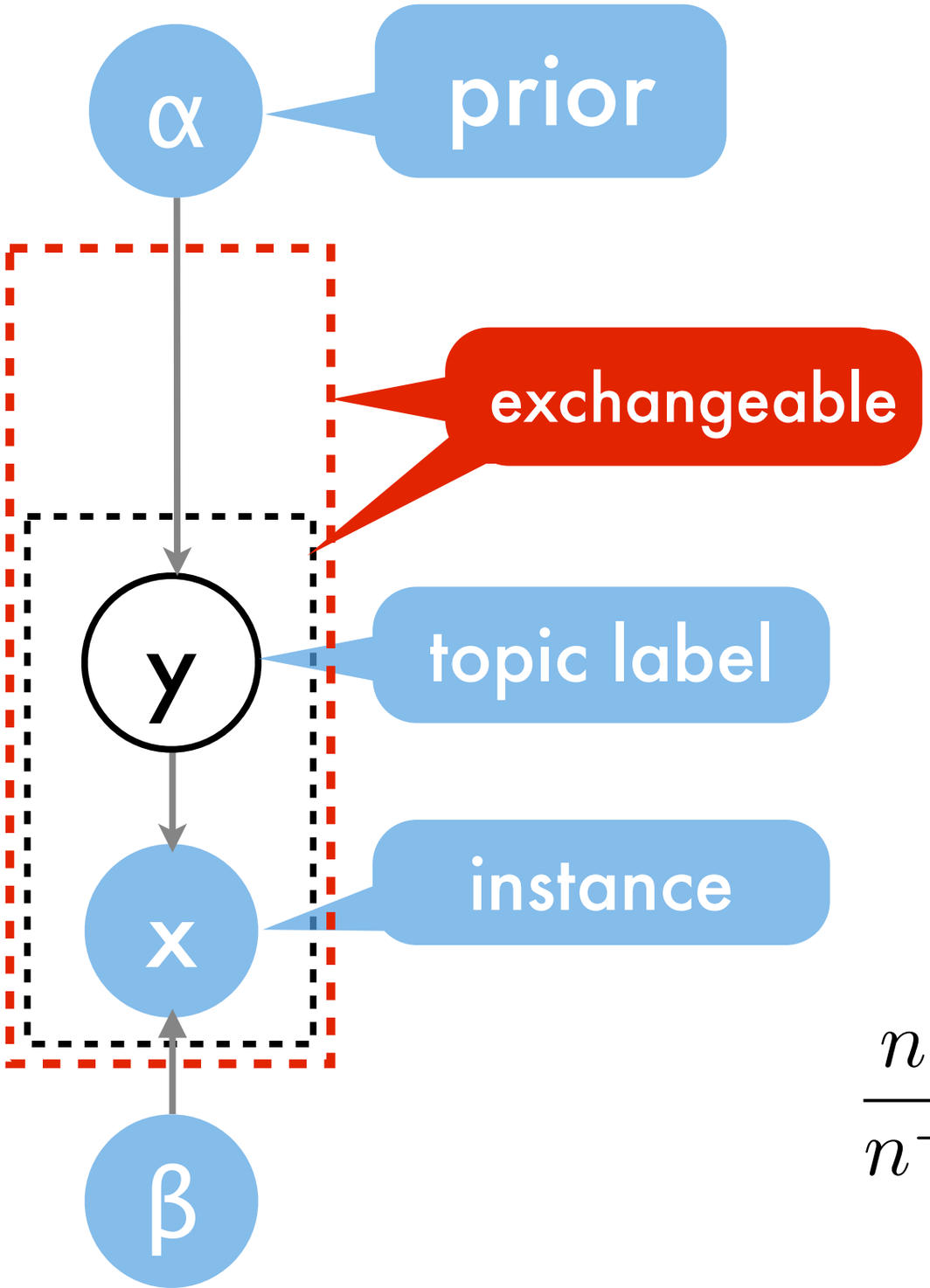
$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

$$h(\alpha) = \sum_i \log\Gamma(\alpha_i) - \log\Gamma\left(\sum_i \alpha_i\right)$$

- Change in normalization is Laplace smoother

$$\exp h(\alpha \cup X) - h(\alpha) = \frac{\alpha_i + n_i}{\sum_i \alpha_i + n_i}$$

Collapsed Inference

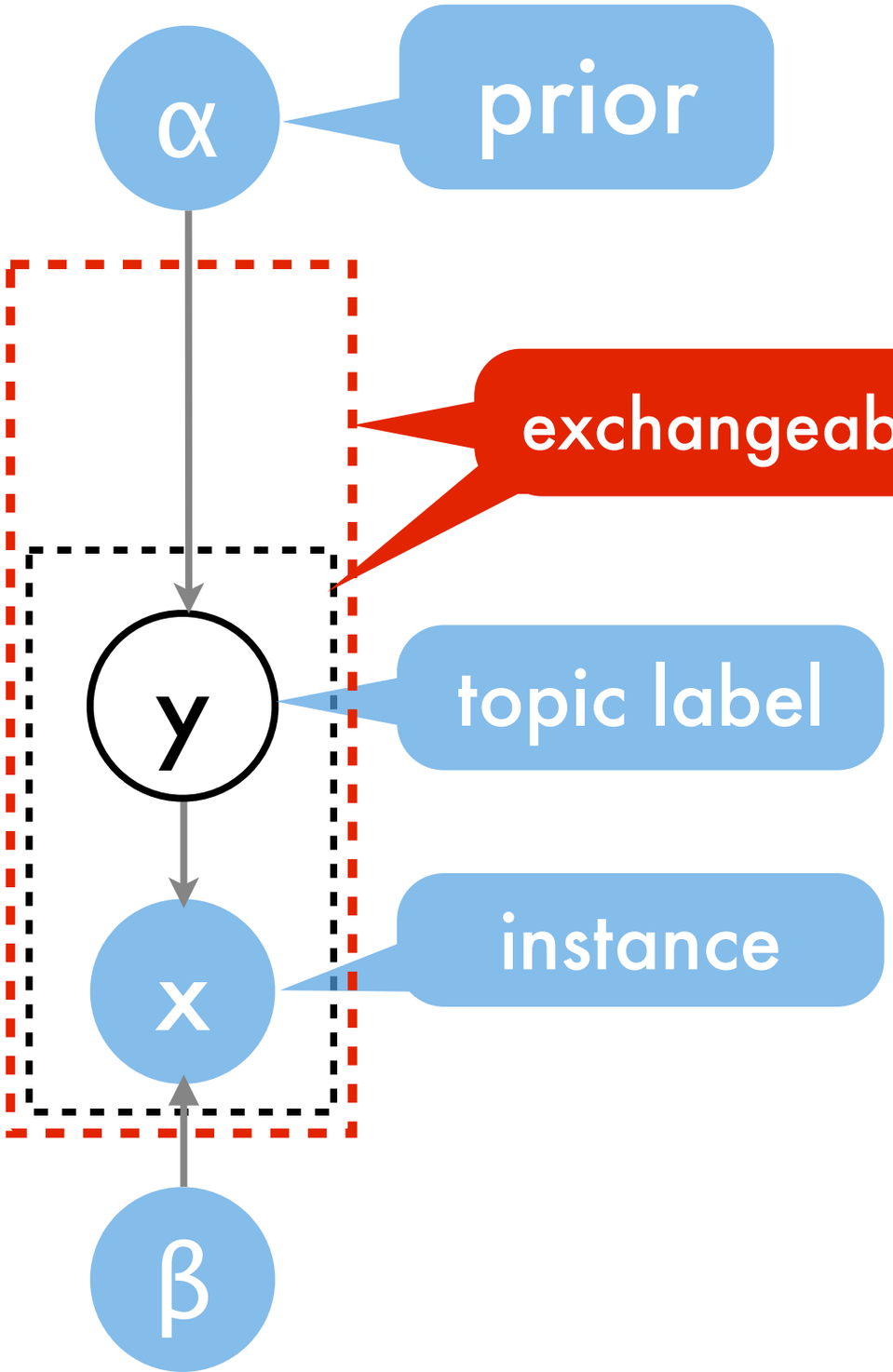


- Collapsed Dirichlet for topics
- Collapsed Dirichlet for words

- Unnormalized product (n are counters for y)

$$\frac{n^{-ij}(t,d)+\alpha_t}{n^{-i}(d)+\sum_t \alpha_t} \quad \frac{n^{-ij}(t,w)+\beta_t}{n^{-i}(t)+\sum_t \beta_t}$$

Acceleration



- For most words only few topics relevant
- Normalization requires sum over all topics regardless
- Exploit sparsity in n

$$\frac{n^{-ij}(t,d)+\alpha_t}{n^{-i}(d)+\sum_t \alpha_t} \quad \frac{n^{-ij}(t,w)+\beta_t}{n^{-i}(t)+\sum_t \beta_t}$$

Gibbs Sampler (Griffiths & Steyvers)

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
      - Update global (word, topic) table

YAHOO!

Gibbs Sampler (Griffiths & Steyvers)

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
      - Update global (word, topic) table

this kills parallelism

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
  - Update global (word, topic) table

$$p(t|w_{ij}) \propto \beta_w \frac{\alpha_t}{n(t)+\bar{\beta}} + \beta_w \frac{n(t,d=i)}{n(t)+\bar{\beta}} + \frac{n(t,w=w_{ij})\,[n(t,d=i)+\alpha_t]}{n(t)+\bar{\beta}}$$

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
  - Update global (word, topic) table

$$p(t|w_{ij}) \propto \beta_w \frac{\alpha_t}{n(t)+\bar{\beta}} + \beta_w \frac{n(t,d=i)}{n(t)+\bar{\beta}} + \frac{n(t,w=w_{ij})\,[n(t,d=i)+\alpha_t]}{n(t)+\bar{\beta}}$$

slow

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
  - Update global (word, topic) table

changes rapidly

$$p(t|w_{ij}) \propto \beta_w \frac{\alpha_t}{n(t)+\bar{\beta}} + \beta_w \frac{n(t,d=i)}{n(t)+\bar{\beta}} + \frac{n(t,w=w_{ij})\,[n(t,d=i)+\alpha_t]}{n(t)+\bar{\beta}}$$

slow

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
  - Update global (word, topic) table

changes rapidly

$$p(t|w_{ij}) \propto \beta_w \frac{\alpha_t}{n(t)+\bar{\beta}} + \beta_w \frac{n(t,d=i)}{n(t)+\bar{\beta}} + \frac{n(t,w=w_{ij})\,[n(t,d=i)+\alpha_t]}{n(t)+\bar{\beta}}$$

slow

moderately fast

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

- For 1000 iterations do
  - For each document do
    - For each word in the document do
      - Resample topic for the word
      - Update local (document, topic) table
      - Update CPU local (word, topic) table
  - Update global (word, topic) table

table out of sync

memory inefficient

blocking

network bound

changes rapidly

$$p(t|w_{ij}) \propto \beta_w \frac{\alpha_t}{n(t)+\bar{\beta}} + \beta_w \frac{n(t,d=i)}{n(t)+\bar{\beta}} + \frac{n(t,w=w_{ij})\,[n(t,d=i)+\alpha_t]}{n(t)+\bar{\beta}}$$

slow

moderately fast

YAHOO!

Fully asynchronous sampler

- For 1000 iterations do (independently per computer)
  - For each thread/core do
    - For each document do
      - For each word in the document do
        - Resample topic for the word
        - Update local (document, topic) table
        - Generate computer local (word, topic) message
    - In parallel update local (word, topic) table
  - In parallel update global (word, topic) table

YAHOO!

Fully asynchronous sampler

- For 1000 iterations do (independently per computer)
  - For each thread/core do
    - For each document do
      - For each word in the document do
        - Resample topic for the word
        - Update local (document, topic) table
        - Generate computer local (word, topic) message
    - In parallel update local (word, topic) table
  - In parallel update global (word, topic) table

network bound

concurrent cpu hdd net

YAHOO!

Fully asynchronous sampler

- For 1000 iterations do (independently per computer)
  - For each thread/core do
    - For each document do
      - For each word in the document do
        - Resample topic for the word
        - Update local (document, topic) table
        - Generate computer local (word, topic) message
    - In parallel update local (word, topic) table
  - In parallel update global (word, topic) table

network bound

memory inefficient

concurrent cpu hdd net

minimal view

YAHOO!

Fully asynchronous sampler

- For 1000 iterations do (independently per computer)
  - For each thread/core do
    - For each document do
      - For each word in the document do
        - Resample topic for the word
        - Update local (document, topic) table
        - Generate computer local (word, topic) message
    - In parallel update local (word, topic) table
  - In parallel update global (word, topic) table

network bound

memory inefficient

table out of sync

concurrent cpu hdd net

minimal view

continuous sync

YAHOO!

Fully asynchronous sampler

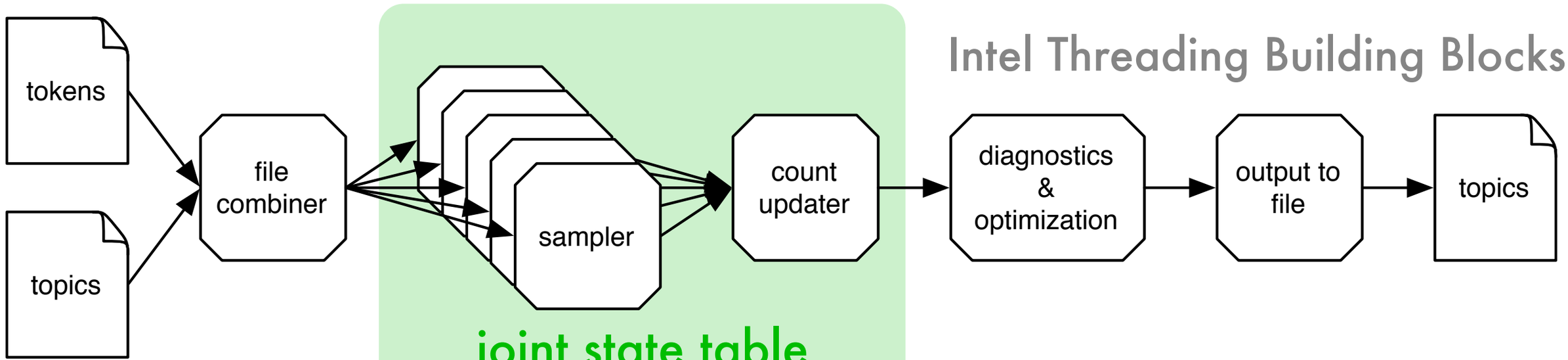
- For 1000 iterations do (independently per computer)
  - For each thread/core do
    - For each document do
      - For each word in the document do
        - Resample topic for the word
        - Update local (document, topic) table
        - Generate computer local (word, topic) message
    - In parallel update local (word, topic) table
  - In parallel update global (word, topic) table

network bound | memory inefficient | table out of sync | blocking

concurrent cpu hdd net | minimal view | continuous sync | barrier free

YAHOO!®

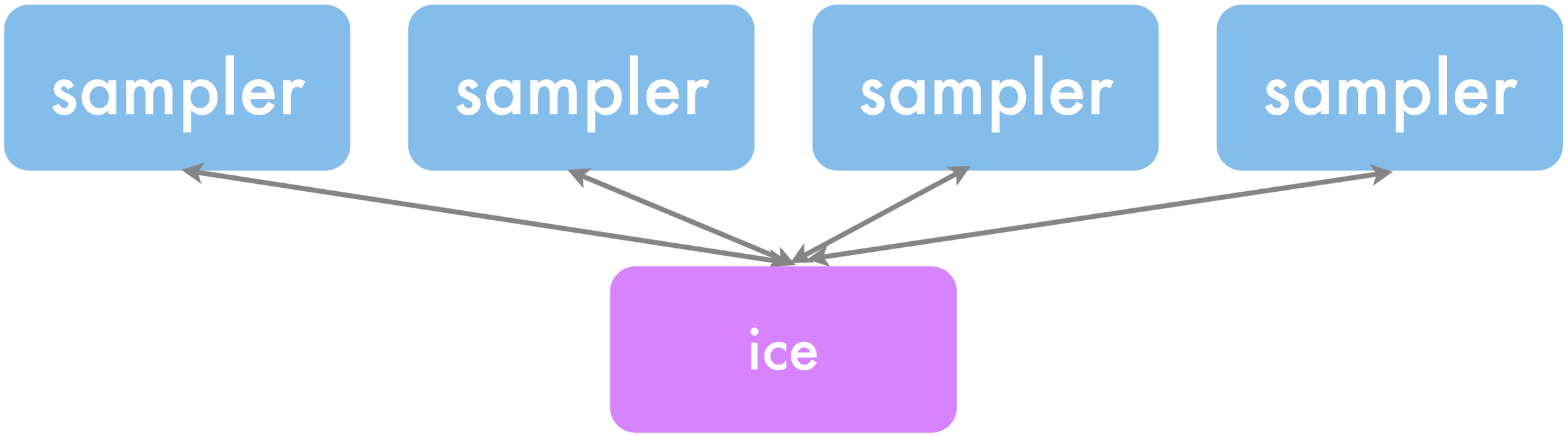
Multicore Architecture



- Decouple multithreaded sampling and updating (almost) avoids stalling for locks in the sampler
- Joint state table
  - much less memory required
  - samplers syncronized (10 docs vs. millions delay)
- Hyperparameter update via stochastic gradient descent
- No need to keep documents in memory (streaming)

YAHOO!

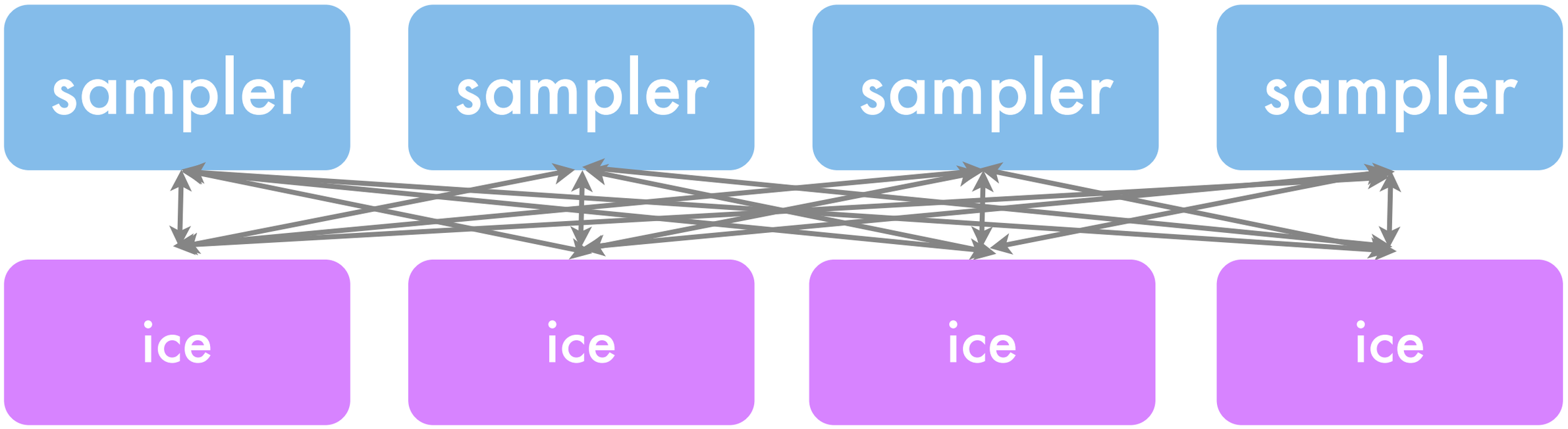
Cluster Architecture



- Distributed (key,value) storage via memcached
- Background asynchronous synchronization
  - single word at a time to avoid deadlocks
  - no need to have joint dictionary
  - uses disk, network, cpu simultaneously

YAHOO!

Cluster Architecture



- Distributed (key,value) storage via ICE
- Background asynchronous synchronization
  - single word at a time to avoid deadlocks
  - no need to have joint dictionary
  - uses disk, network, cpu simultaneously

YAHOO!

Making it work

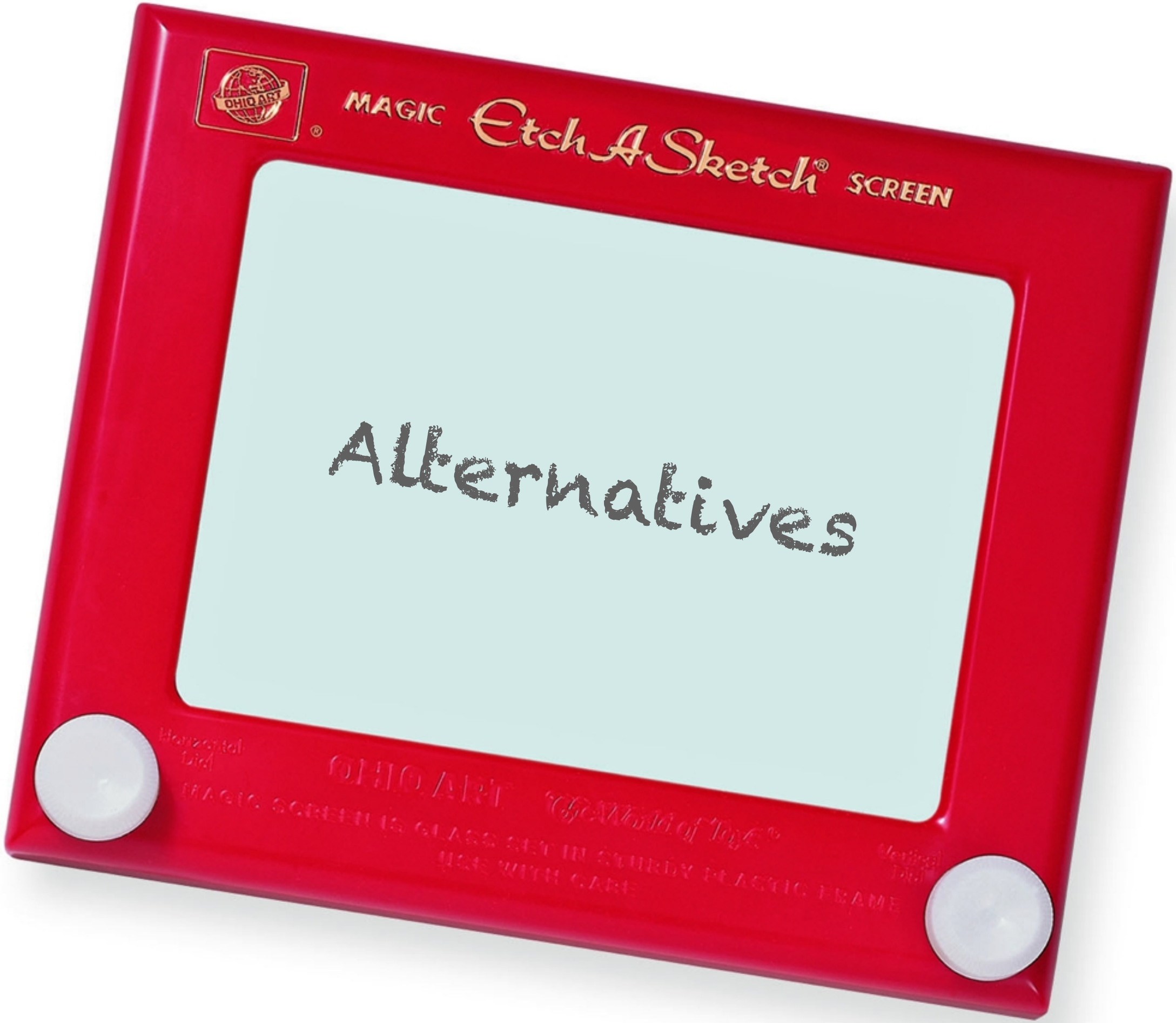
- Startup
  - Randomly initialize topics on each node (read from disk if already assigned - hotstart)
  - Sequential Monte Carlo for startup much faster
  - Aggregate changes on the fly
- Failover
  - State constantly being written to disk (worst case we lose 1 iteration out of 1000)
  - Restart via standard startup routine
- Achilles heel: need to restart from checkpoint if even a single machine dies.

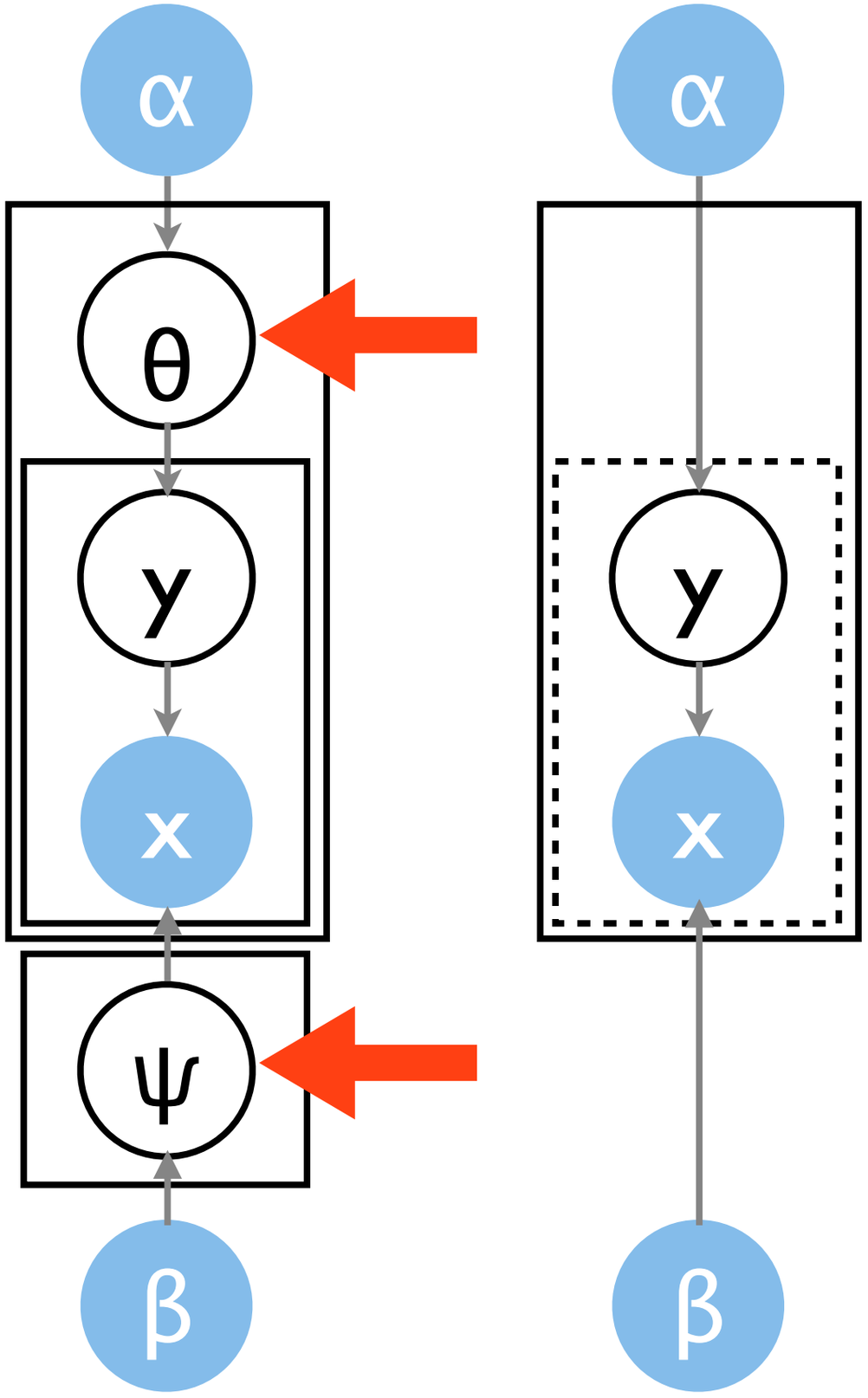
YAHOO!

Easily extensible

- Better language model (topical n-grams)
  can process millions of users (vs 1000s)
- Conditioning on side information (upstream)
  estimate topic based on authorship, source, joint user model ...
- Conditioning on dictionaries (downstream)
  integrate topics between different languages
- Time dependent sampler for user model
  approximate inference per episode

YAHOO!


MAGIC Etch A Sketch® SCREEN
Alternatives



Hal Daume; Joey Gonzalez

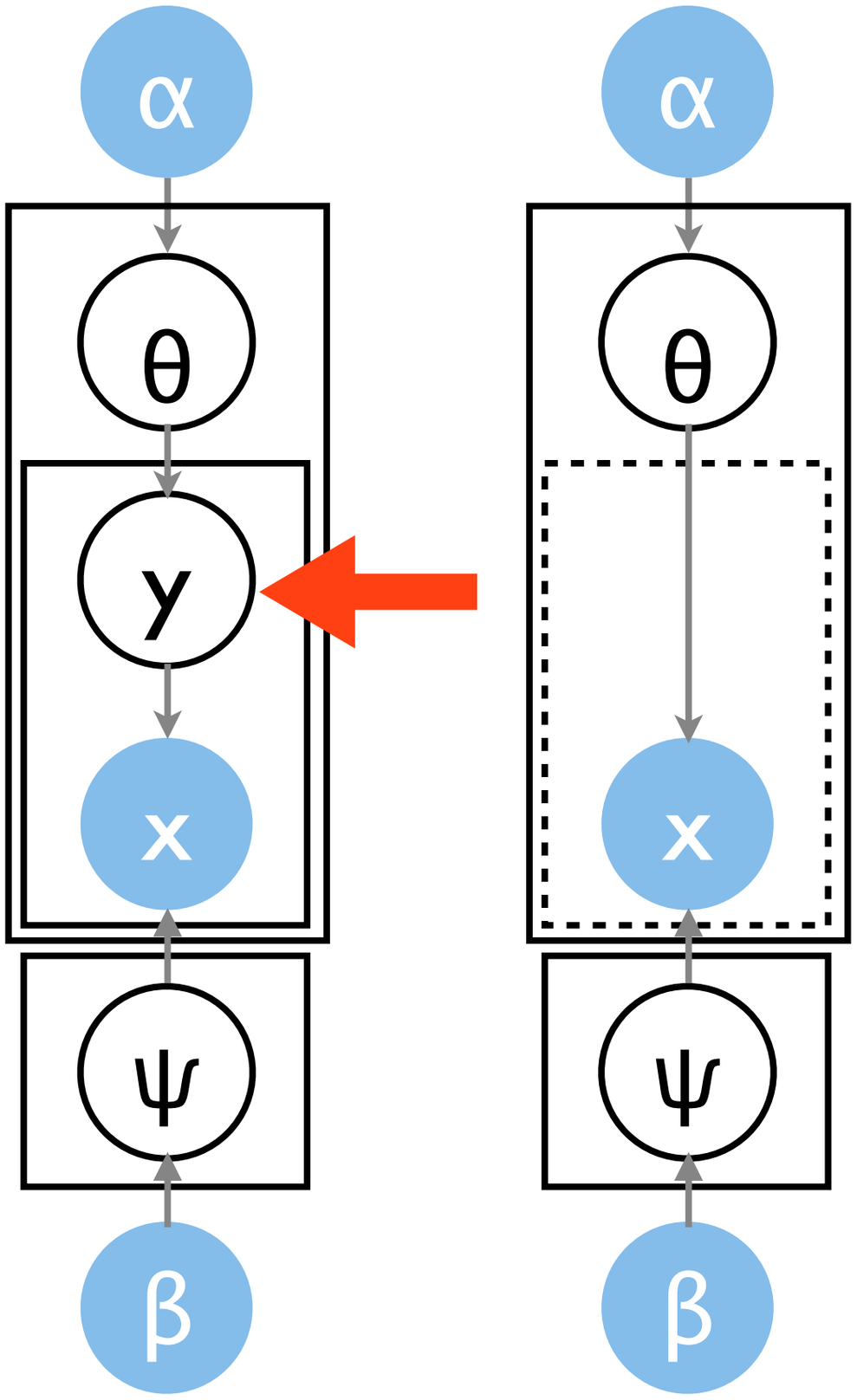
V1 - Brute force maximization

- Integrate out latent parameters θ and ψ

$$p(X, Y|\alpha, \beta)$$

- Discrete maximization problem in Y

- Hard to implement
- Overfits a lot (mode is not a typical sample)
- Parallelization infeasible



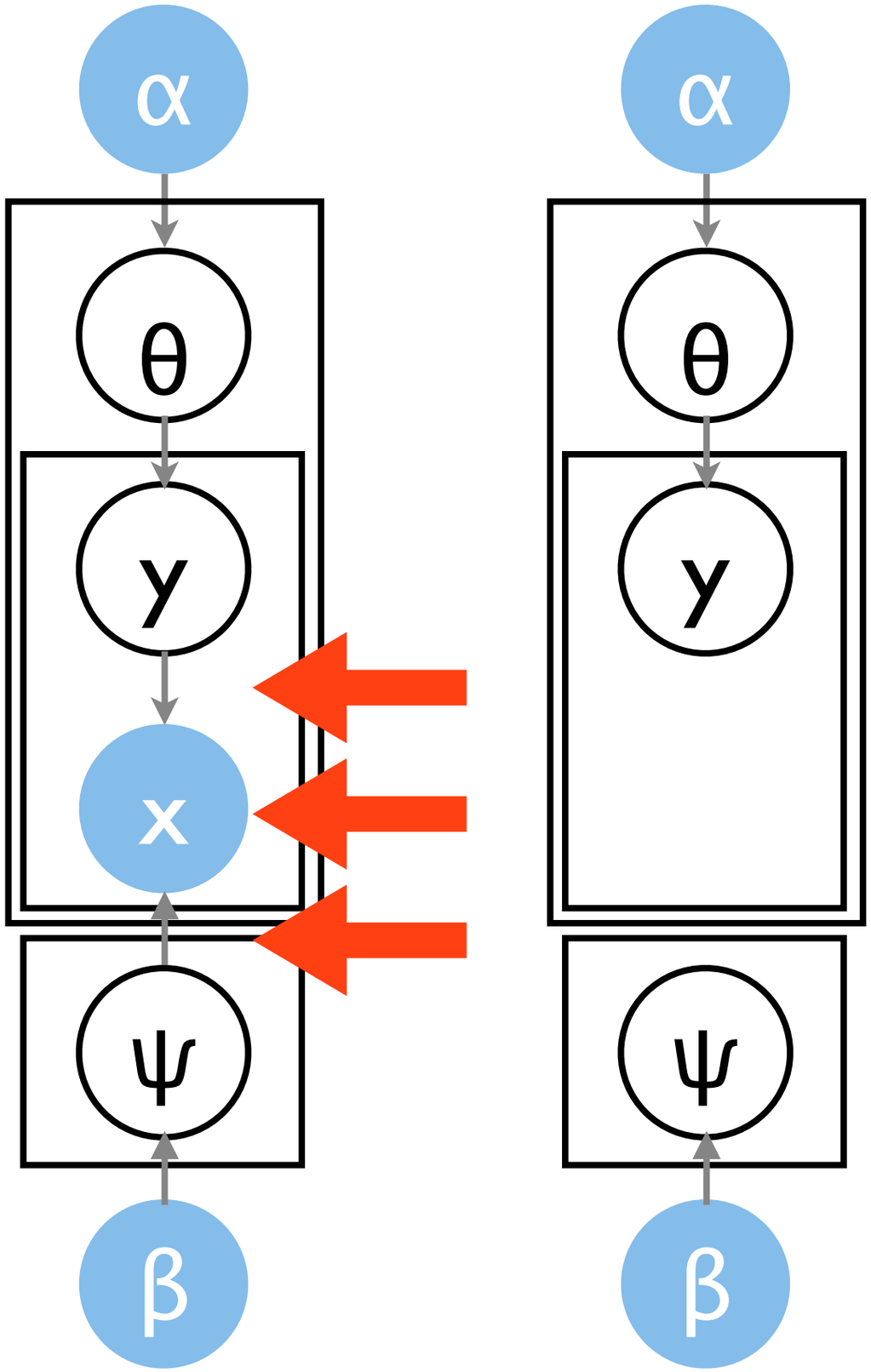
Hoffmann, Blei, Bach (in VW)

V2 - Brute force maximization

- Integrate out latent parameters y

$$p(X, \psi, \theta | \alpha, \beta)$$

- Continuous nonconvex optimization problem in θ and ψ
- Solve by stochastic gradient descent over documents
- Easy to implement
- Does not overfit much
- Great for small datasets
- Parallelization difficult/impossible
- Memory storage/access is O(T W) (this breaks for large models)
  - 1M words, 1000 topics = 4GB
  - Per document 1MFlops/iteration



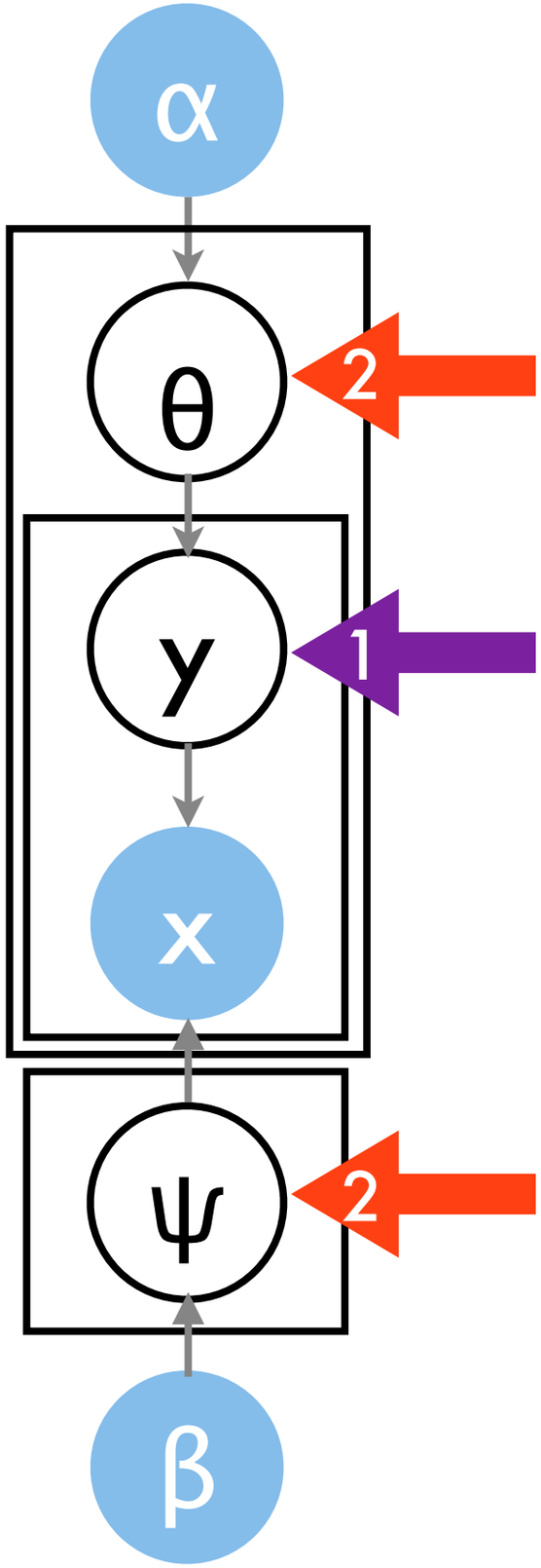
Blei, Ng, Jordan

V3 - Variational approximation

- Approximate intractable joint distribution by tractable factors

$$\log p(x) \geq \log p(x) - D(q(y)\|p(y|x))$$
$$= \int dq(y)\,[\log p(x) + \log p(y|x) - q(y)]$$
$$= \int dq(y) \log p(x,y) + H[q]$$

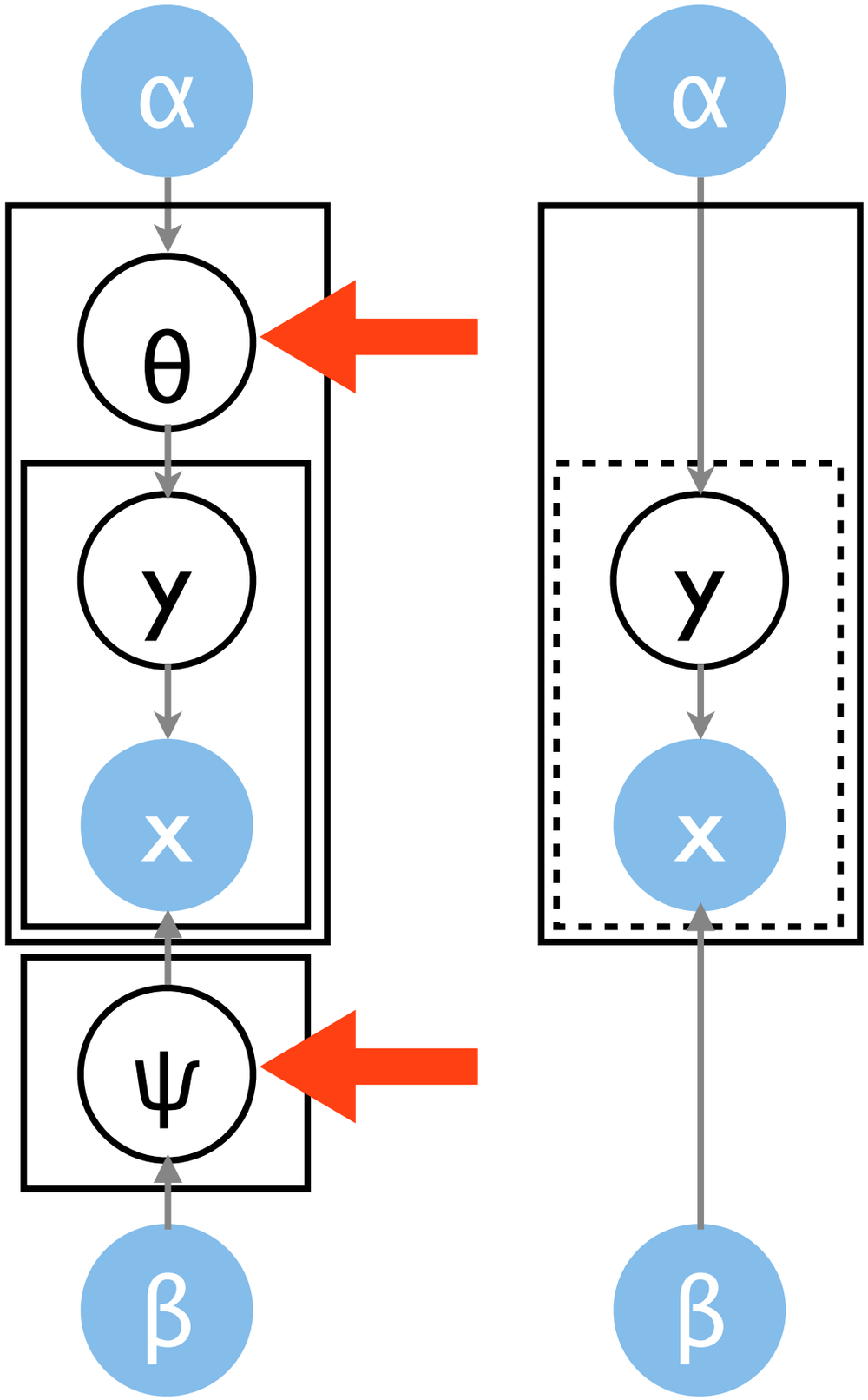
- Alternating convex optimization problem
- Dominant cost is matrix matrix multiply
- Easy to implement
- Great for small topics/vocabulary
- Parallelization easy (aggregate statistics)
- Memory storage is O(T W)
  (this breaks for large models)
- Model not quite as good as sampling



V4 - Uncollapsed Sampling

- Sample y_{ij}|rest
  Can be done in parallel
- Sample θ|rest and ψ|rest
  Can be done in parallel
- Compatible with MapReduce (only aggregate statistics)
- Easy to implement
- Children can be conditionally independent*
- Memory storage is O(T W) (this breaks for large models)
- Mixes slowly

*for the right model



Griffiths & Steyvers 2005

V5 - Collapsed Sampling

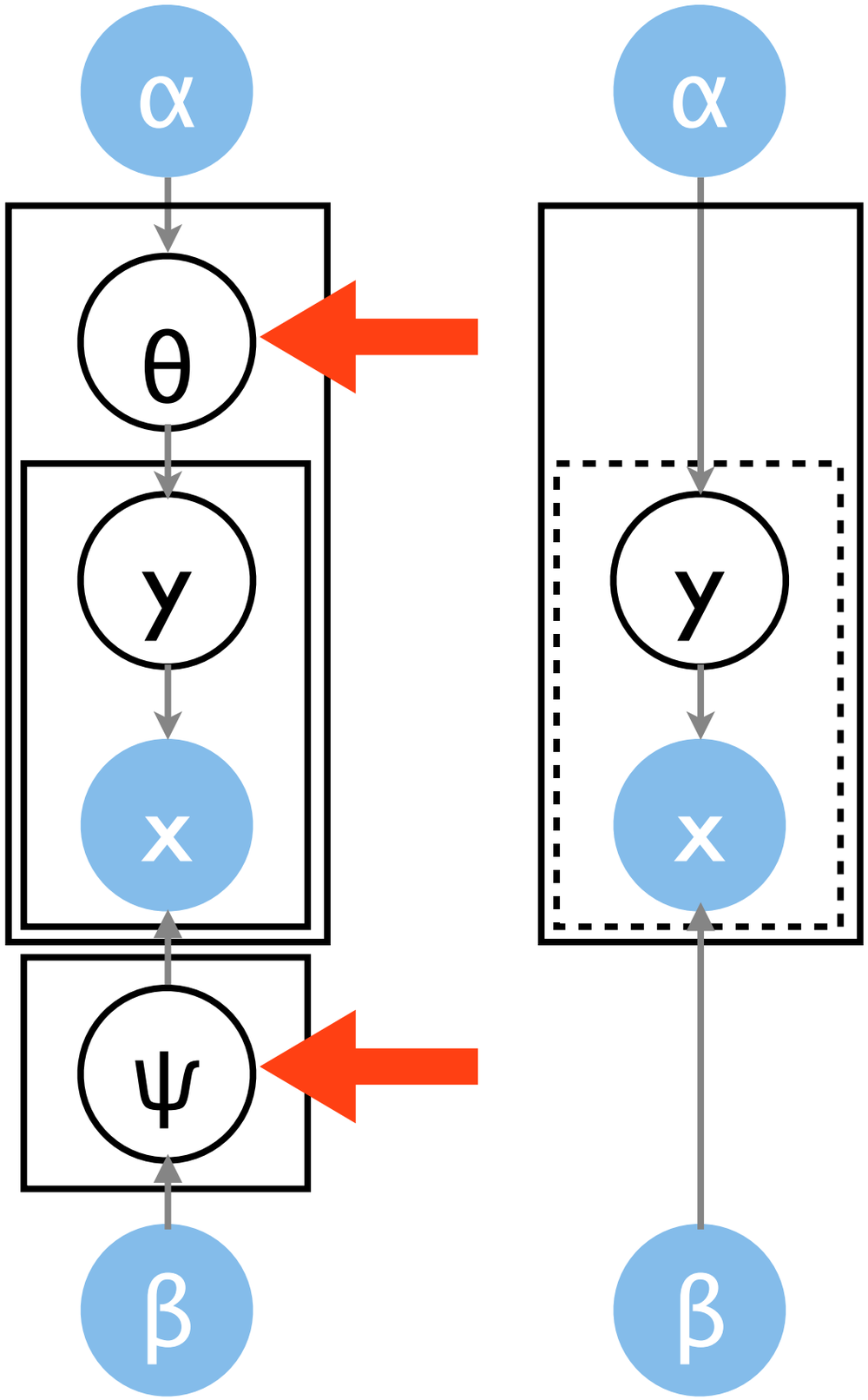
- Integrate out latent parameters θ and ψ

$$p(X, Y|\alpha, \beta)$$

- Sample one topic assignment $y_{ij}|X,Y^{-ij}$ at a time from

$$\frac{n^{-ij}(t,d)+\alpha_t}{n^{-i}(d)+\sum_t \alpha_t} \qquad \frac{n^{-ij}(t,w)+\beta_t}{n^{-i}(t)+\sum_t \beta_t}$$

- Fast mixing
- Easy to implement
- Memory efficient
- Parallelization infeasible (variables lock each other)



Griffiths & Steyvers 2005

V5 - Collapsed Sampling

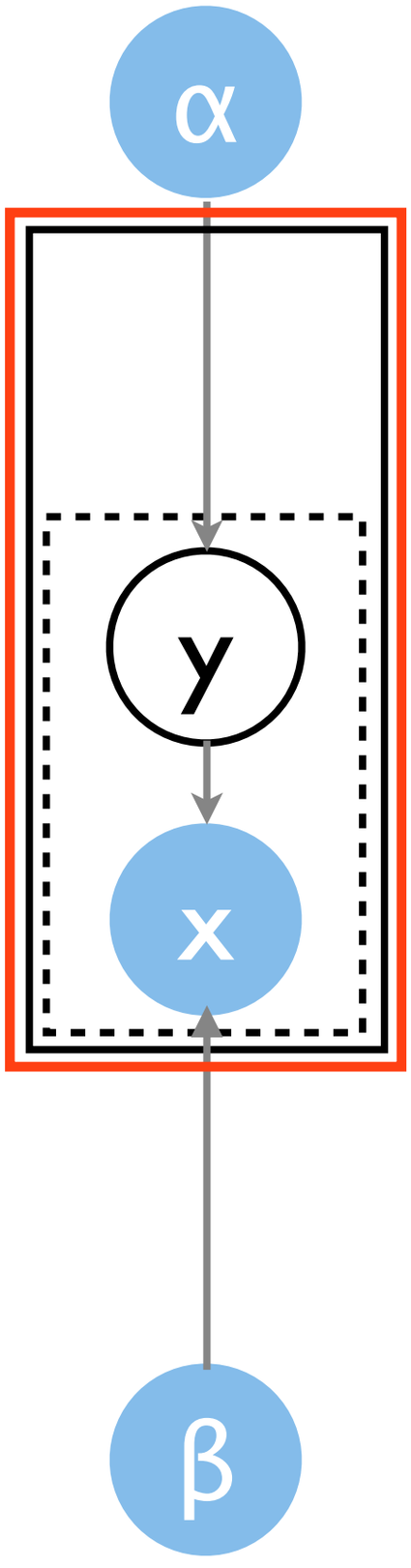
- Integrate out latent parameters θ and ψ

$$p(X, Y|\alpha, \beta)$$

- Sample one topic assignment $y_{ij}|X,Y^{-ij}$ at a time from

$$\frac{n^{-ij}(t,d)+\alpha_t}{n^{-i}(d)+\sum_t \alpha_t} \qquad \frac{n^{-ij}(t,w)+\beta_t}{n^{-i}(t)+\sum_t \beta_t}$$

- Fast mixing
- Easy to implement
- Memory efficient
- Parallelization infeasible (variables lock each other)



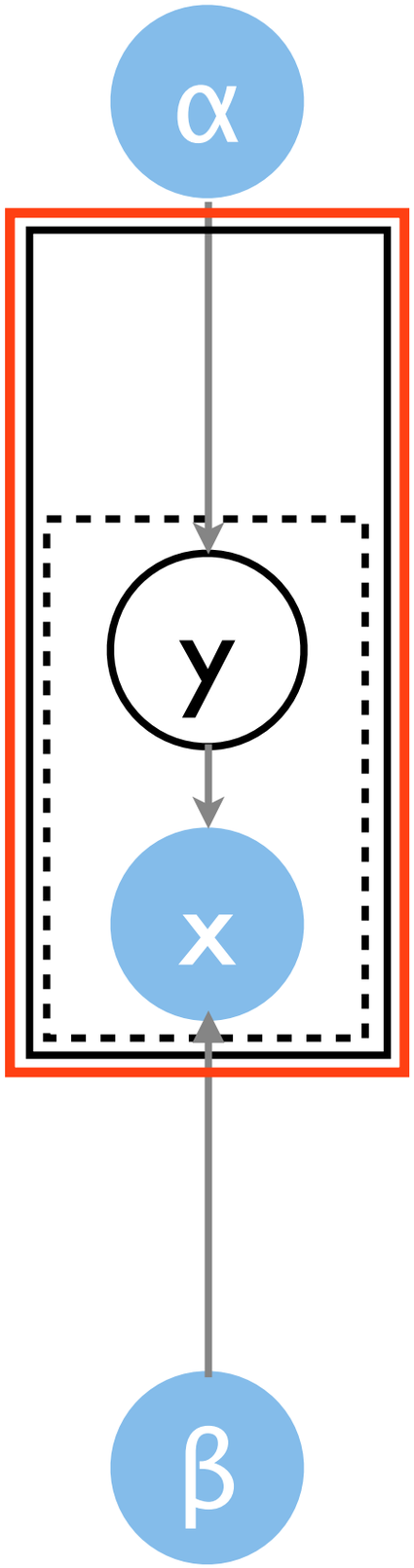
Asuncion, Smyth, Welling, … UCI
Mimno, McCallum, … UMass

V6 - Approximating the Distribution

- Collapsed sampler per machine

$$\frac{n^{-ij}(t,d)+\alpha_t}{n^{-i}(d)+\sum_t \alpha_t} \qquad \frac{n^{-ij}(t,w)+\beta_t}{n^{-i}(t)+\sum_t \beta_t}$$

- Defer synchronization between machines
  - no problem for n(t)
  - big problem for n(t,w)
- Easy to implement
- Can be memory efficient
- Easy parallelization
- Mixes slowly/worse likelihood



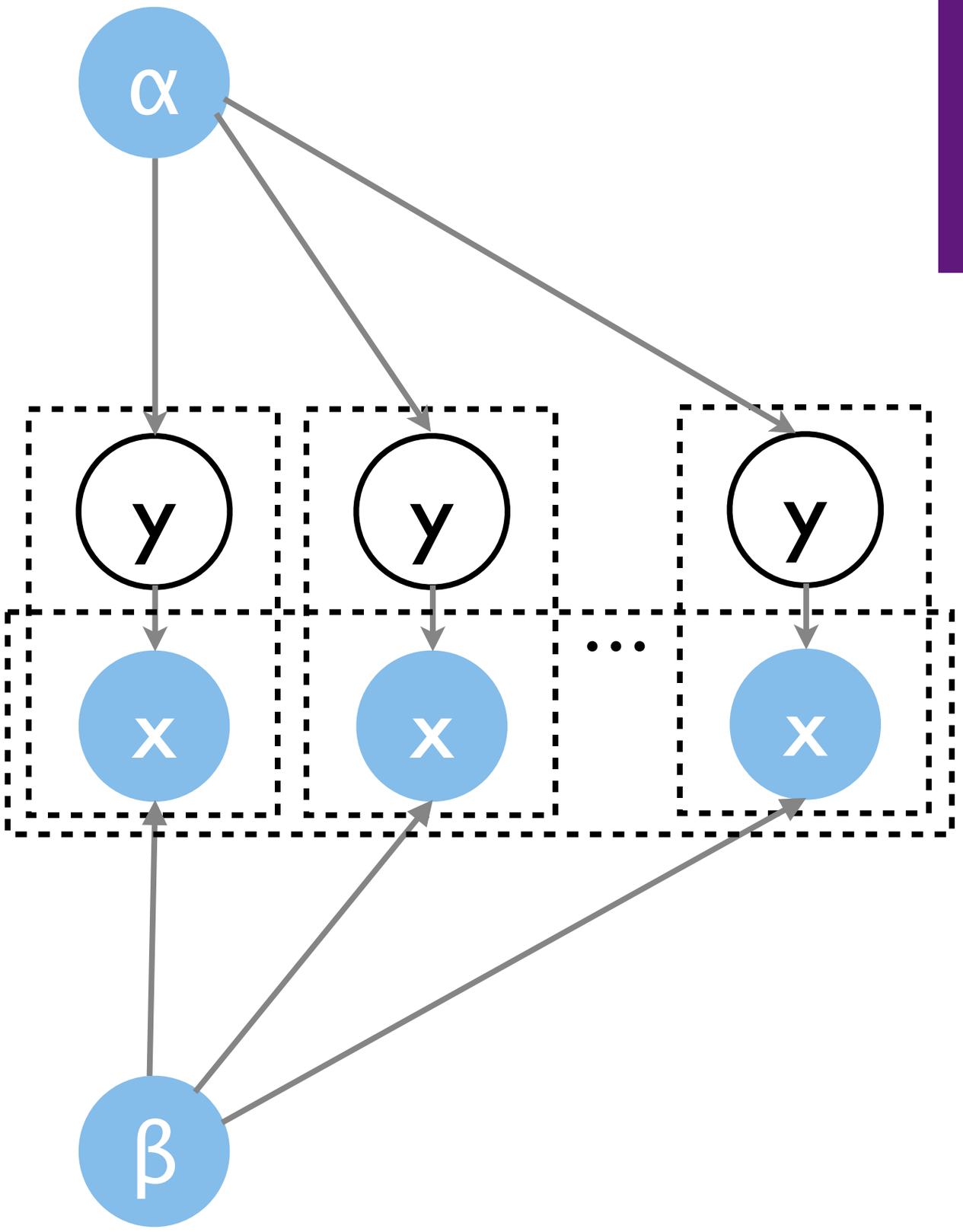
S. and Narayanamurthy, 2009
Ahmed, Gonzalez, et al., 2012

V7 - Better Approximations of the Distribution

- Collapsed sampler

$$\frac{n^{-ij}(t,d)+\alpha_t}{n^{-i}(d)+\sum_t \alpha_t} \quad \frac{n^{-ij}(t,w)+\beta_t}{n^{-i}(t)+\sum_t \beta_t}$$

- Make local copies of state
  - Implicit for multicore (delayed updates from samplers)
  - Explicit copies for multi-machine
- Not a hierarchical model (Welling, Asuncion, et al. 2008)
- Memory efficient (only need to view its own sufficient statistics)
- Multicore / Multi-machine
- Convergence speed depends on synchronizer quality



Canini, Shi, Griffiths, 2009
Ahmed et al., 2011

V8 - Sequential Monte Carlo

- Integrate out latent θ and ψ

$$p(X, Y|\alpha, \beta)$$

- Chain conditional probabilities

$$p(X, Y|\alpha, \beta) = \prod_{i=1}^{m} p(x_i, y_i|x_1, y_1, \ldots x_{i-1}, y_{i-1}, \alpha, \beta)$$

- For each particle sample

$$y_i \sim p(y_i|x_i, x_1, y_1, \ldots x_{i-1}, y_{i-1}, \alpha, \beta)$$

- Reweight particle by next step data likelihood

$$p(x_{i+1}|x_1, y_1, \ldots x_i, y_i, \alpha, \beta)$$

- Resample particles if weight distribution is too uneven

V8 - Sequential Monte Carlo

- One pass through data
- Data sequential parallelization is open problem
- Nontrivial to implement
  - Sampler is easy
  - Inheritance tree through particles is messy
- Need to estimate data likelihood (integration over y), e.g. as part of sampler
- This is multiplicative update algorithm with log loss …

Canini, Shi, Griffiths, 2009
Ahmed et al., 2011

- Integrate out latent θ and ψ

$$p(X, Y|\alpha, \beta)$$

- Chain conditional probabilities

$$p(X, Y|\alpha, \beta) = \prod_{i=1}^{m} p(x_i, y_i|x_1, y_1, \ldots x_{i-1}, y_{i-1}, \alpha, \beta)$$

- For each particle sample

$$y_i \sim p(y_i|x_i, x_1, y_1, \ldots x_{i-1}, y_{i-1}, \alpha, \beta)$$

- Reweight particle by next step data likelihood

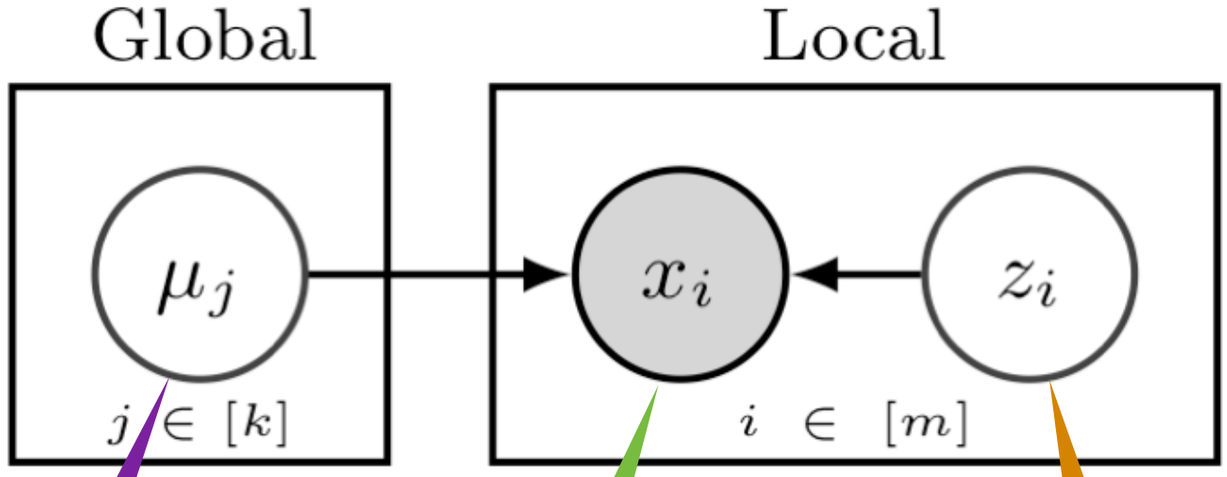
$$p(x_{i+1}|x_1, y_1, \ldots x_i, y_i, \alpha, \beta)$$

- Resample particles if weight distribution is too uneven

	Uncollapsed	Variational approximation	Collapsed natural parameters	Collapsed topic assignments
Optimization	overfits too costly	easy parallelization big memory footprint	overfits too costly	easy to optimize big memory footprint difficult parallelization
Sampling	slow mixing conditionally independent	n.a.	fast mixing difficult parallelization approximate inference by delayed updates particle filtering sequential	sampling difficult


MAGIC Etch A Sketch® SCREEN
Parallel
Inference

3 Problems

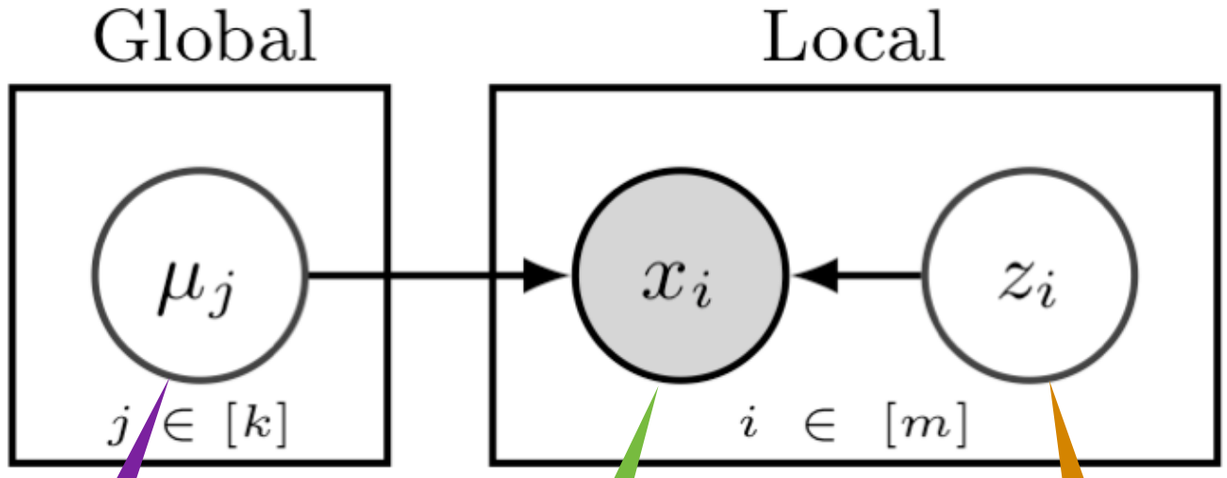


mean
variance
cluster weight

data

cluster ID

3 Problems



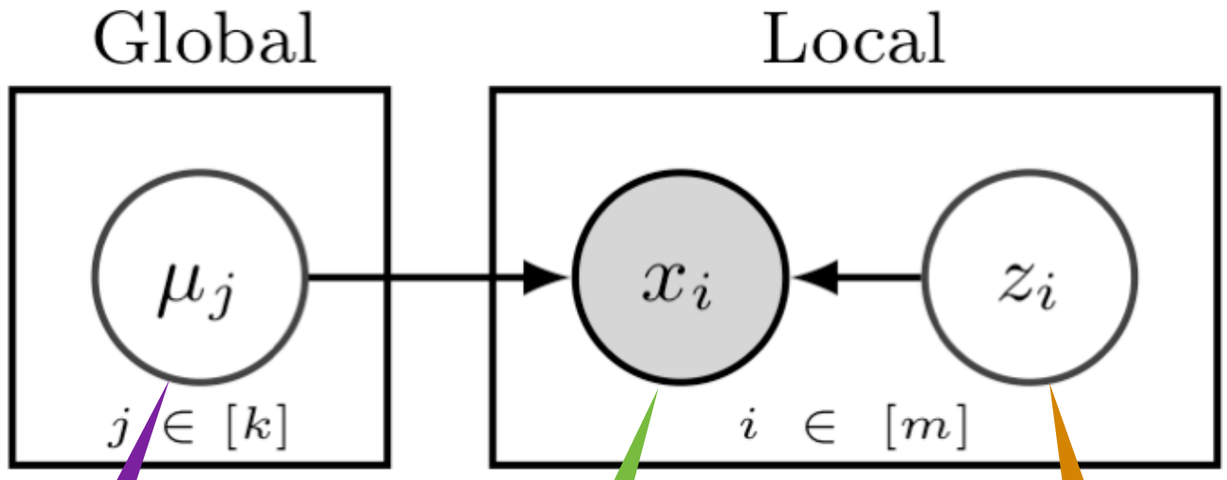
global state

data

local state

YAHOO!

3 Problems



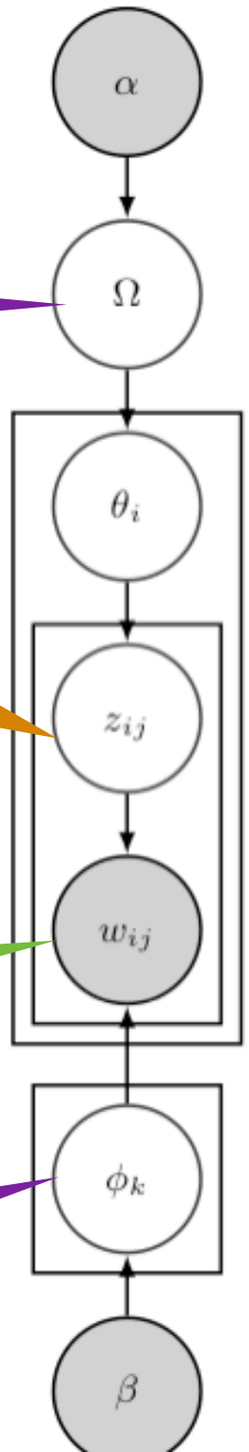
too big for single machine

huge

only local

3 Problems

Vanilla LDA



global state

local state

data

global state

3 Problems

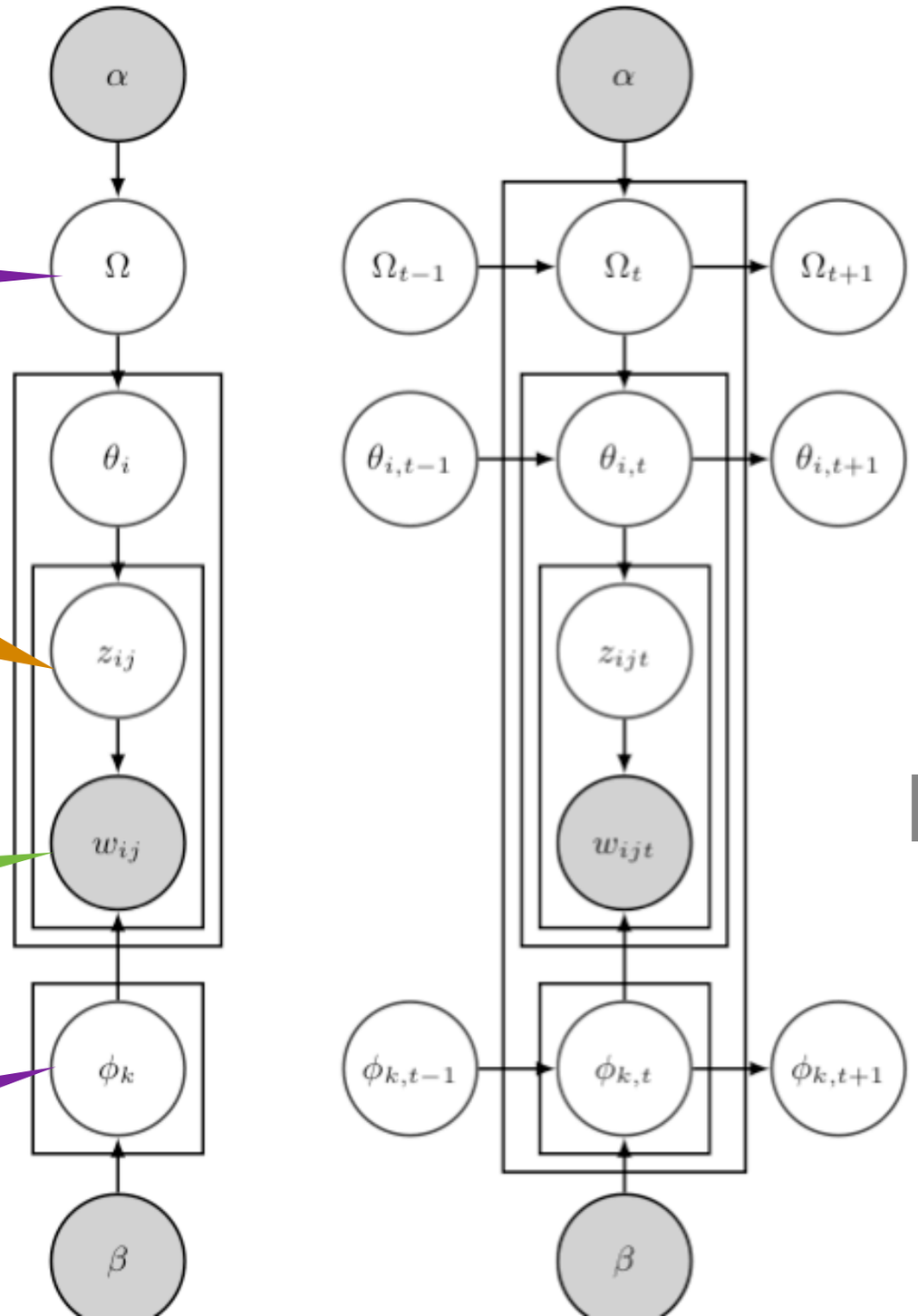
Vanilla LDA

global state

local state

data

global state

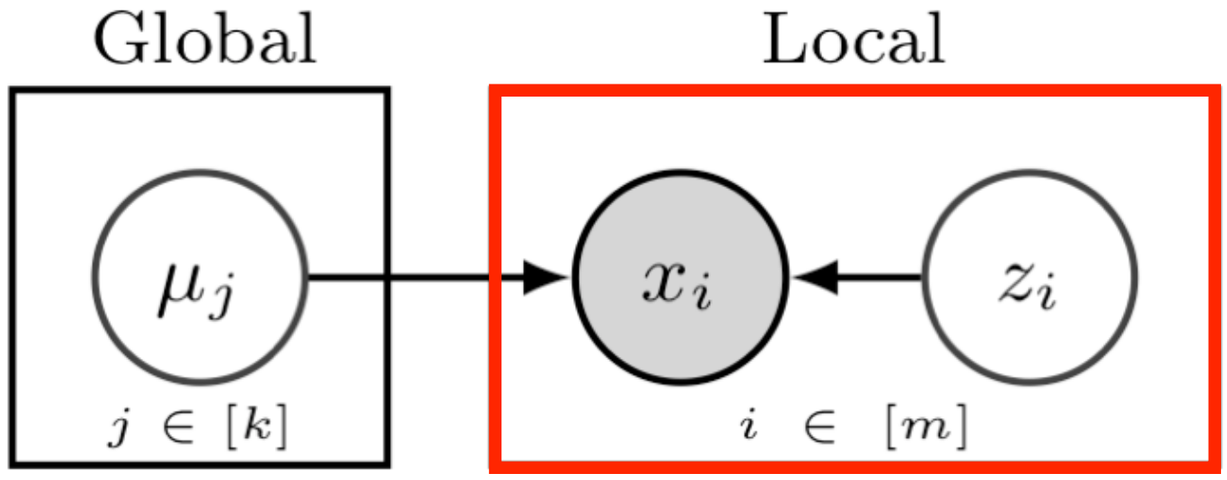


User profiling

YAHOO!

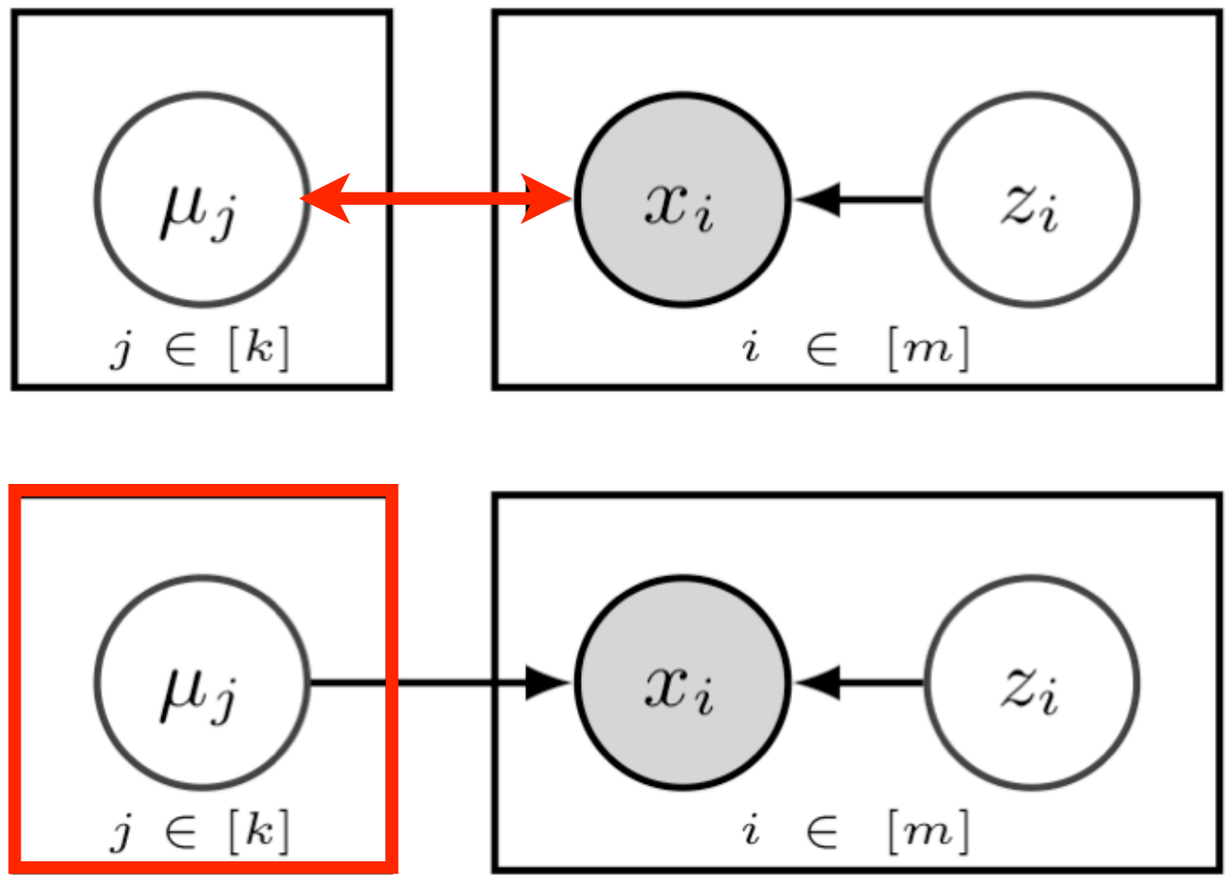
3 Problems

local state is too large



does not fit into memory

global state is too large



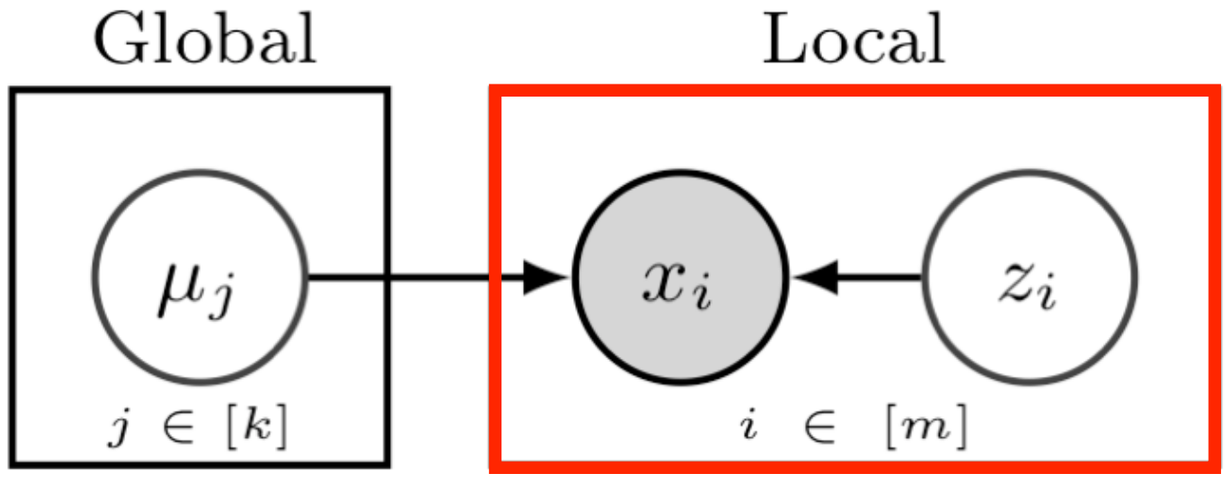
network load & barriers

does not fit into memory

YAHOO!

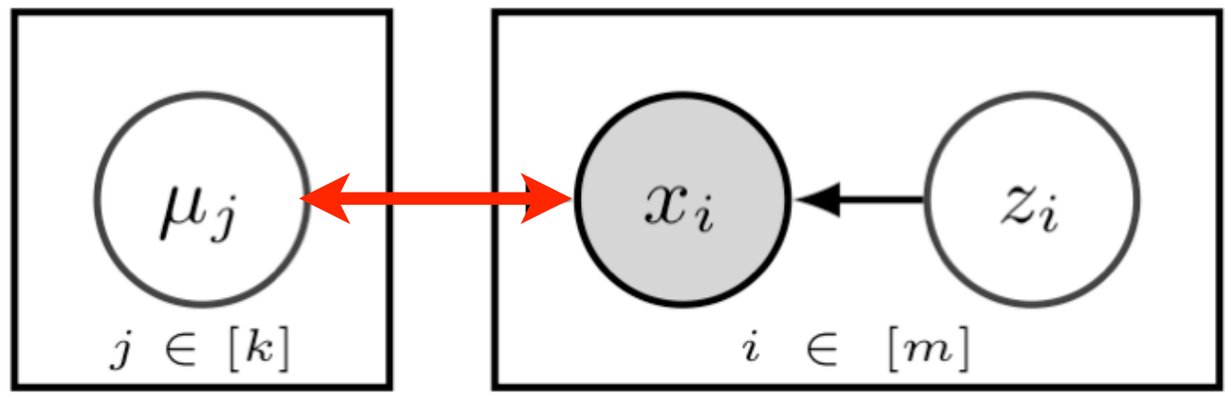
3 Problems

local state is too large

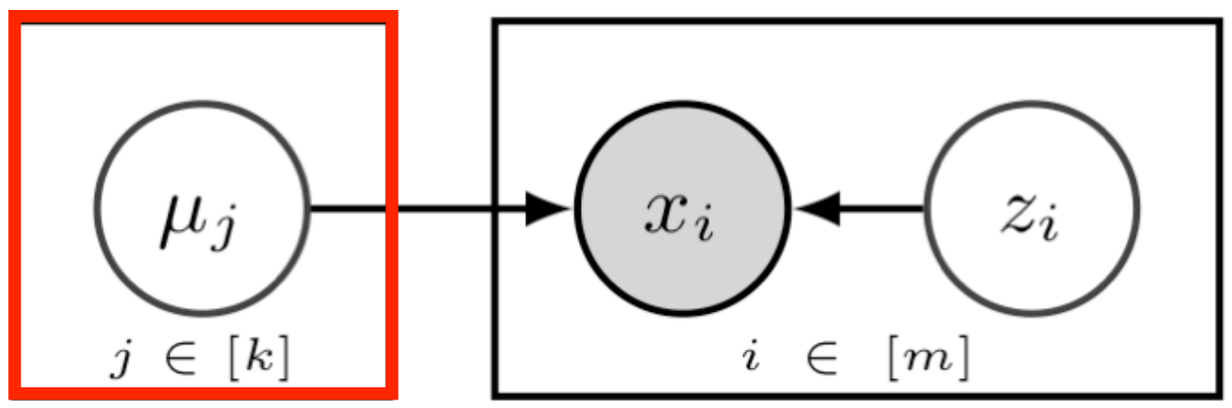


stream local data from disk

global state is too large



network load & barriers

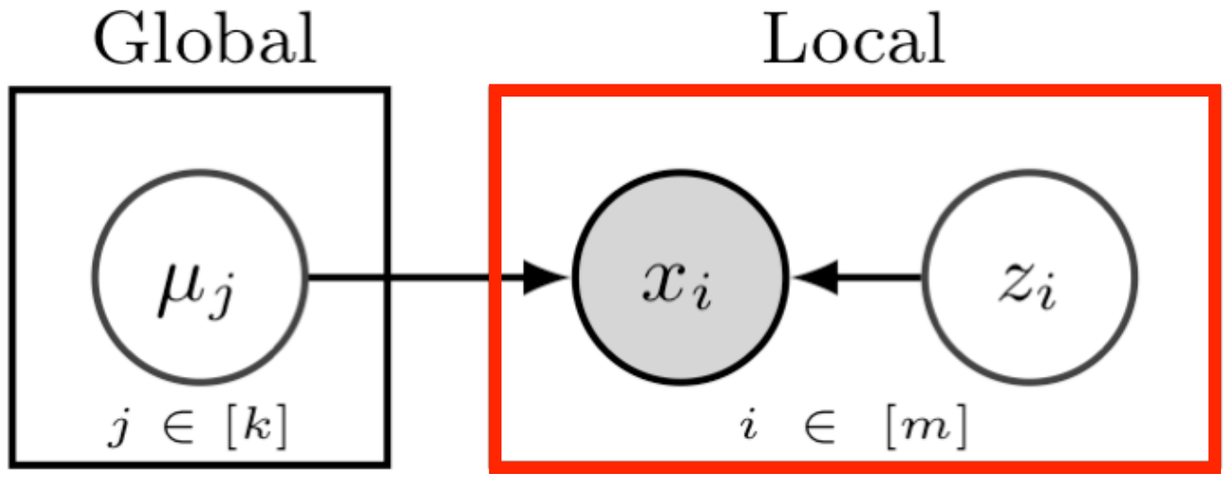


does not fit into memory

YAHOO!

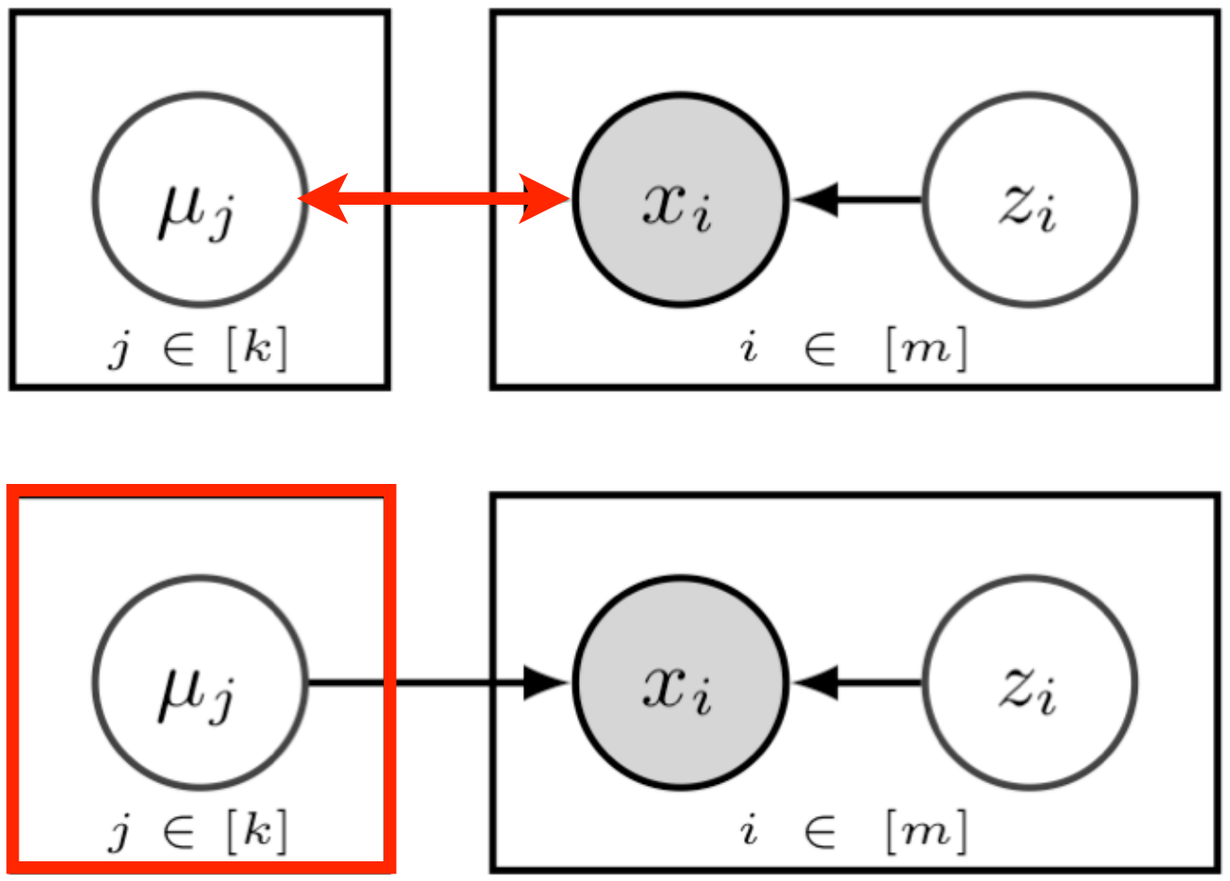
3 Problems

local state is too large



stream local data from disk

global state is too large



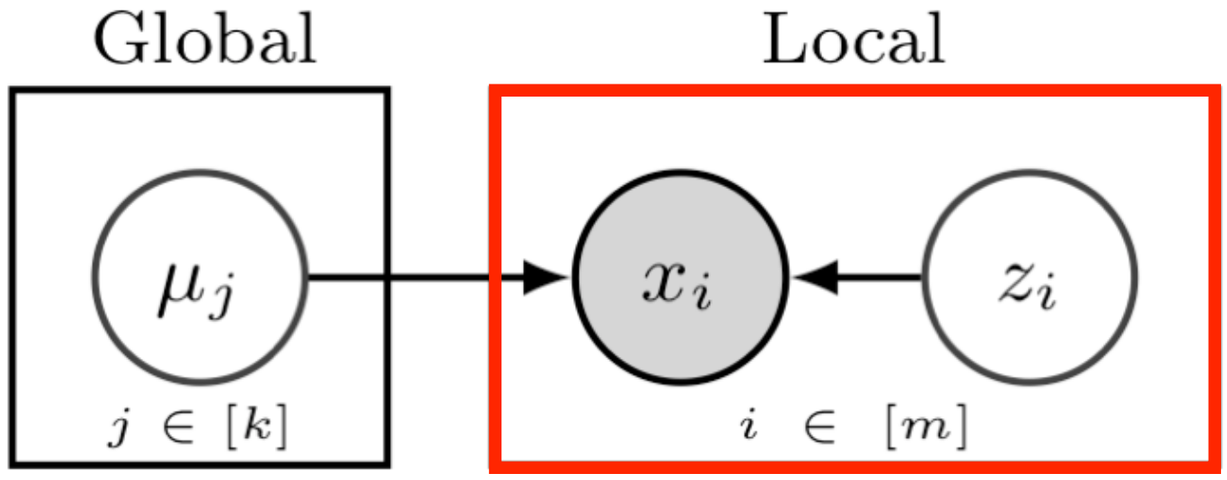
asynchronous synchronization

does not fit into memory

YAHOO!

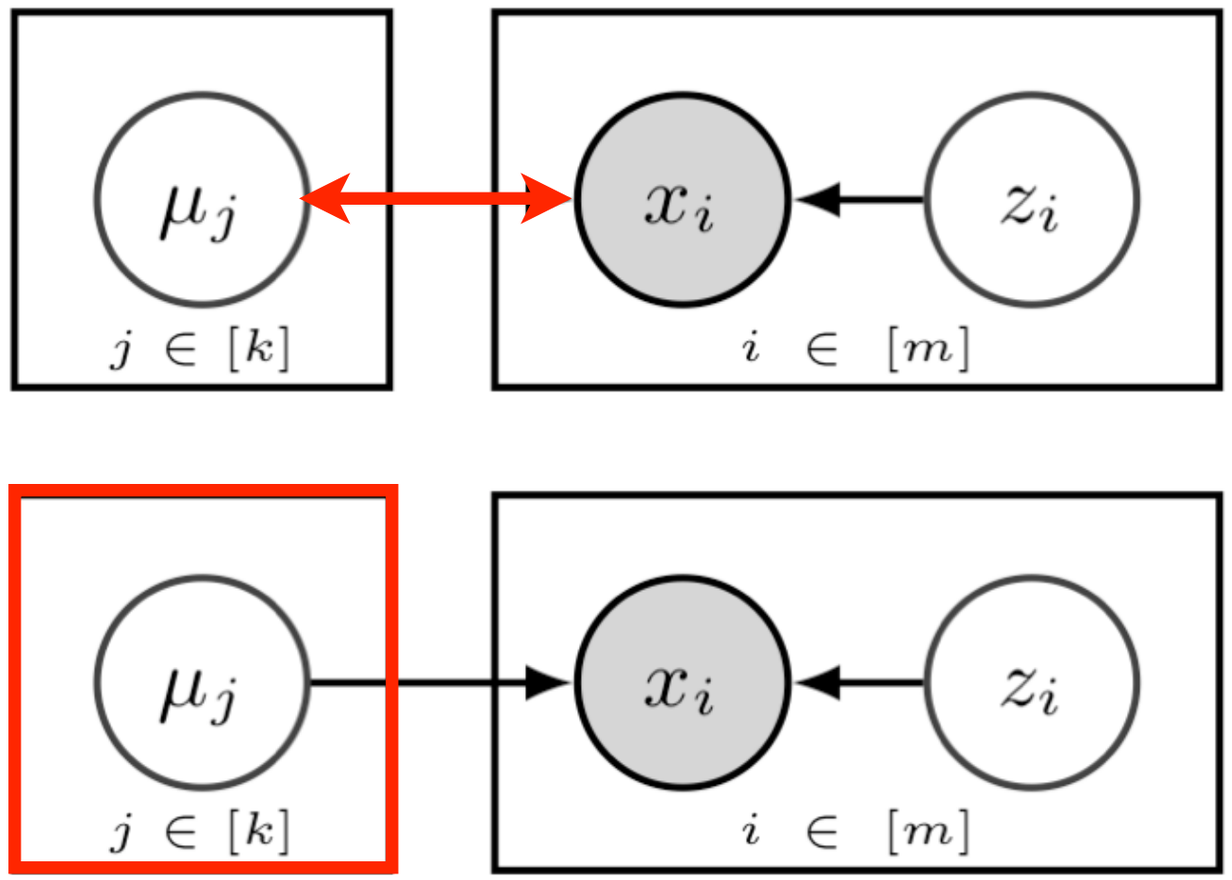
3 Problems

local state is too large



stream local data from disk

global state is too large

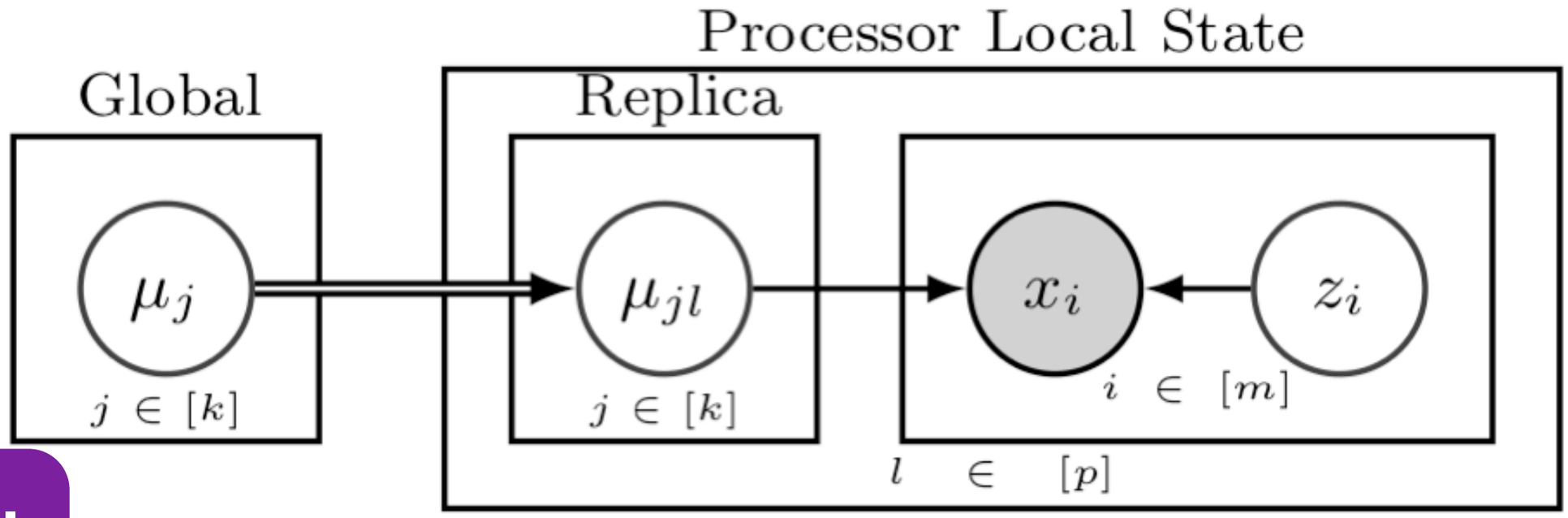


asynchronous synchronization

partial view

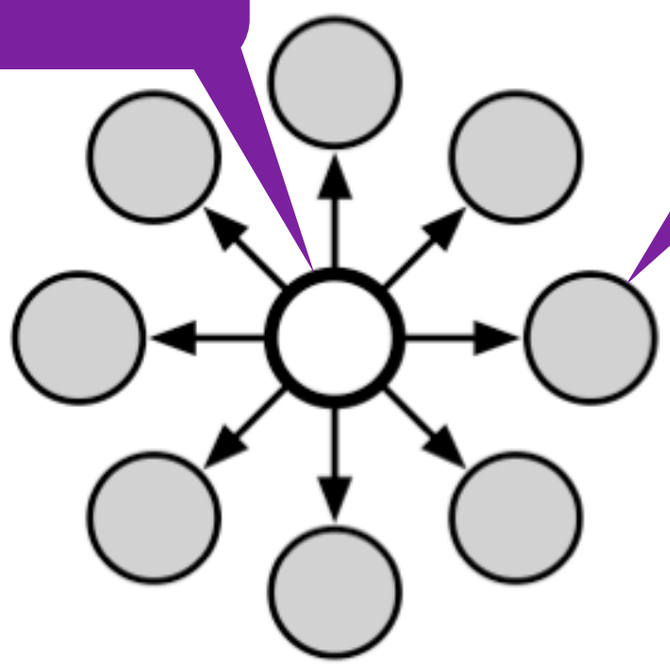
YAHOO!

Distribution

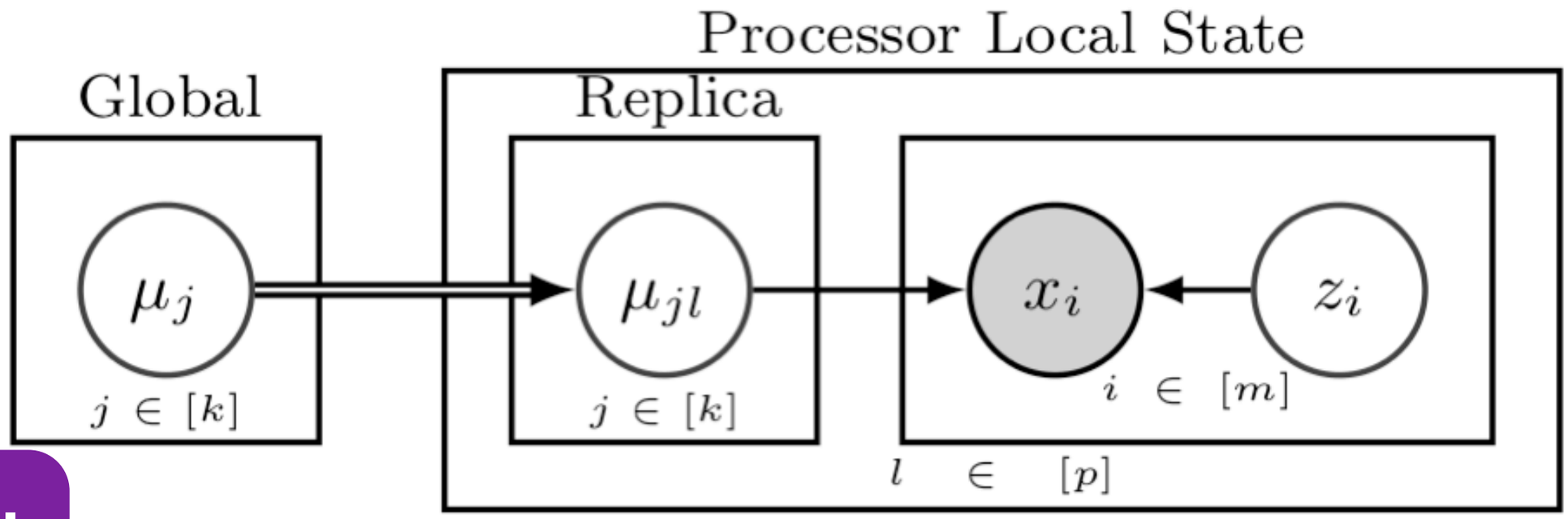


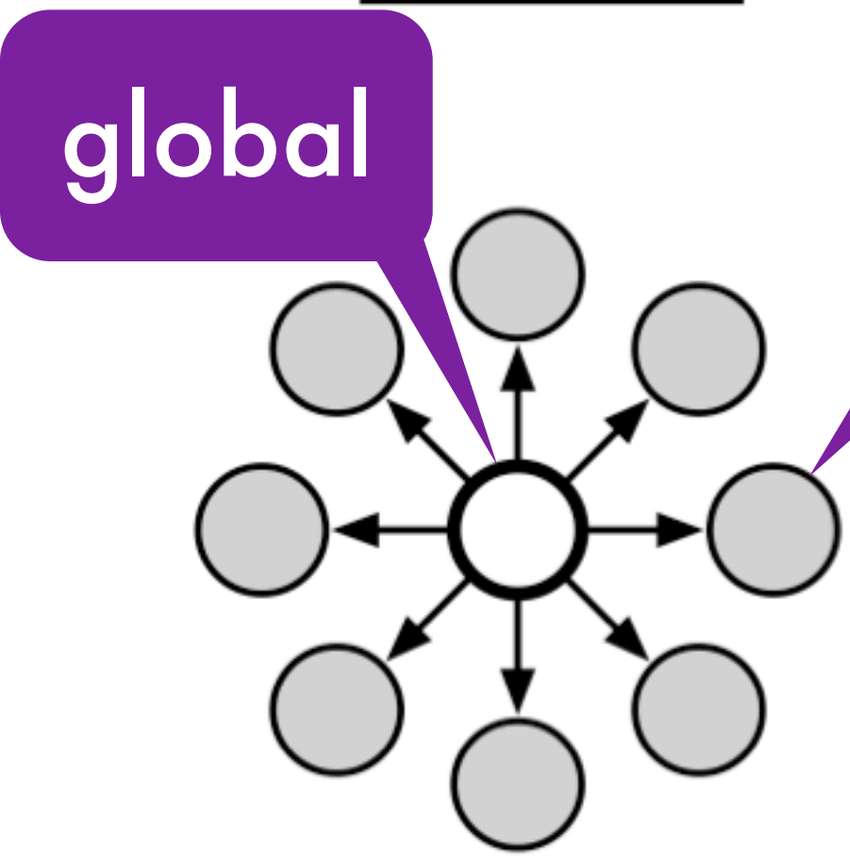
global

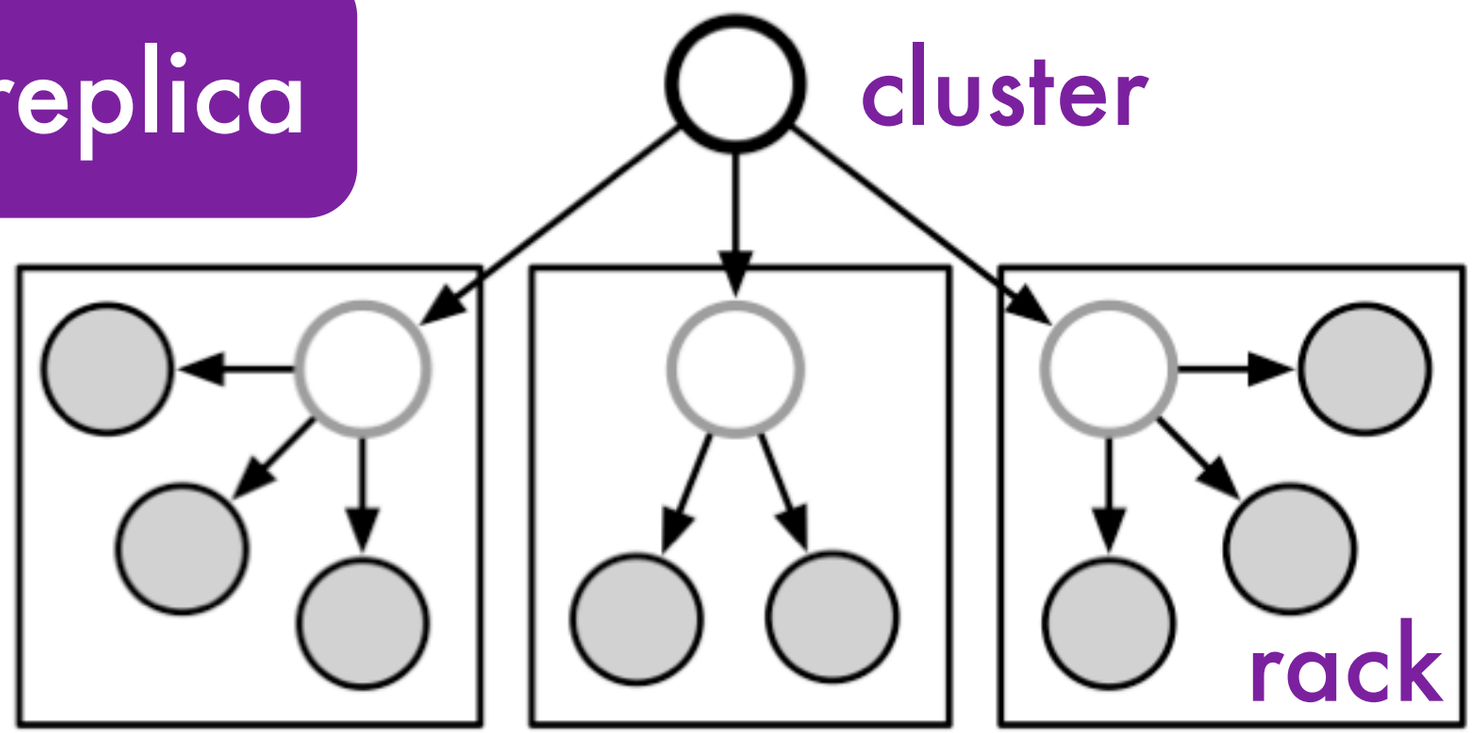
replica



Distribution







Synchronization

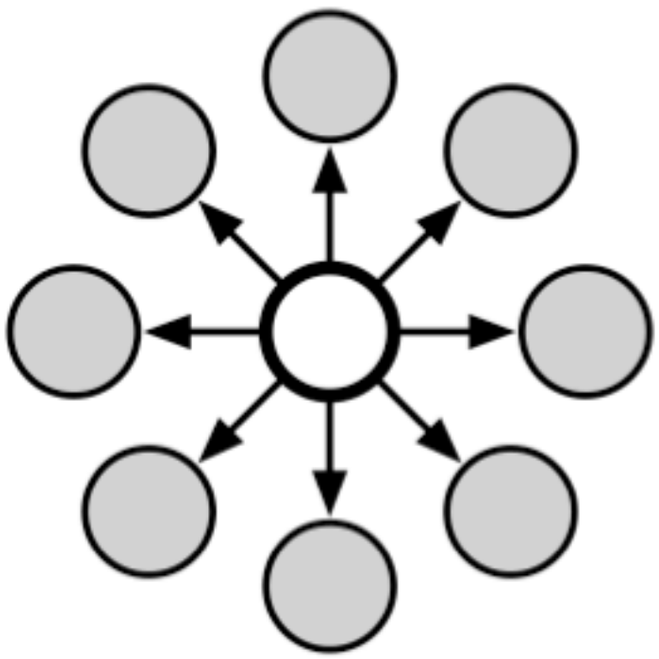
- Child updates local state
  - Start with common state
  - Child stores old and new state
  - Parent keeps global state
- Transmit differences asynchronously
  - Inverse element for difference
  - Abelian group for commutativity (sum, log-sum, cyclic group, exponential families)

local to global

$$\delta \leftarrow x \ominus x^{\text{old}}$$

$$x^{\text{old}} \leftarrow x$$

$$x^{\text{global}} \leftarrow x^{\text{global}} \oplus \delta$$



global to local

$$x \leftarrow x \oplus (x^{\text{global}} \ominus x^{\text{old}})$$

$$x^{\text{old}} \leftarrow x^{\text{global}}$$

Synchronization

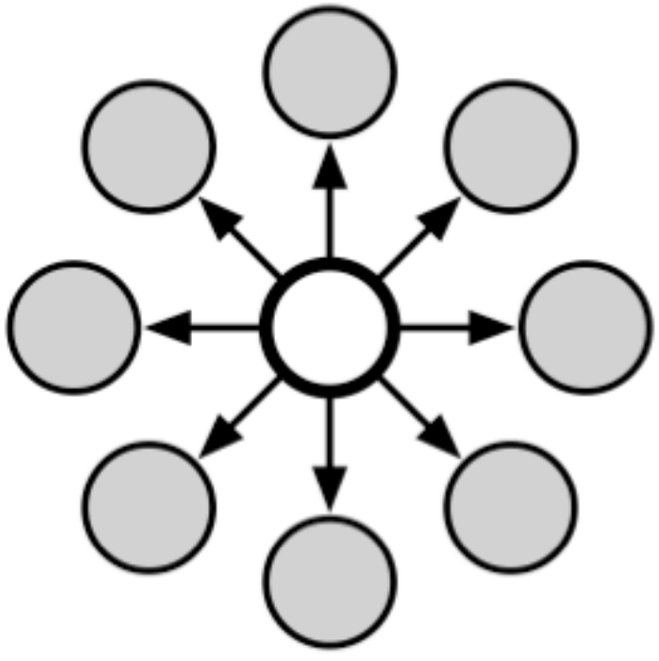
- Naive approach (dumb master)
  - Global is only (key,value) storage
  - Local node needs to lock/read/write/unlock master
  - Needs a 4 TCP/IP roundtrips - latency bound
- Better solution (smart master)
  - Client sends message to master / in queue / master incorporates it
  - Master sends message to client / in queue / client incorporates it
  - Bandwidth bound (>10x speedup in practice)

local to global

$$\delta \leftarrow x \ominus x^{\text{old}}$$

$$x^{\text{old}} \leftarrow x$$

$$x^{\text{global}} \leftarrow x^{\text{global}} \oplus \delta$$



global to local

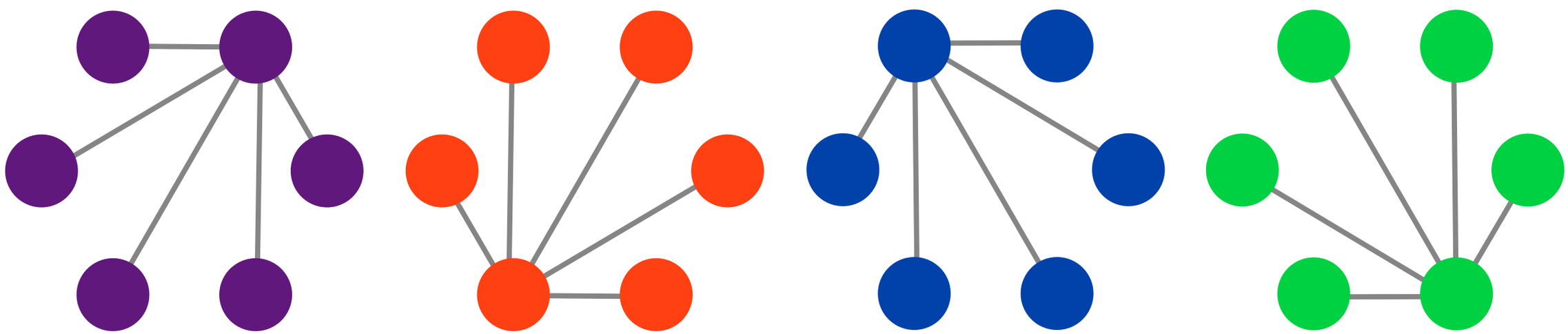
$$x \leftarrow x \oplus (x^{\text{global}} \ominus x^{\text{old}})$$

$$x^{\text{old}} \leftarrow x^{\text{global}}$$

Distribution

- Dedicated server for variables
  - Insufficient bandwidth (hotspots)
  - Insufficient memory
- Select server e.g. via consistent hashing

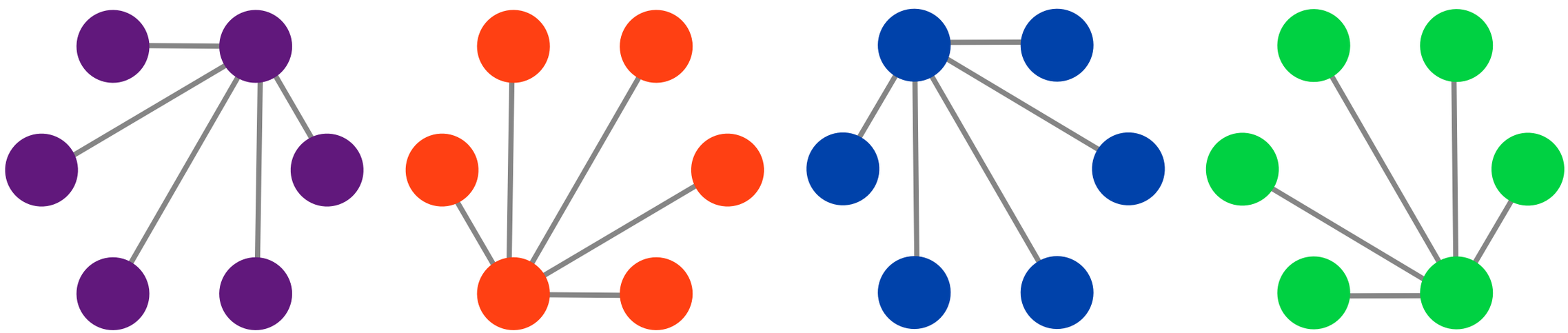
$$m(x) = \operatorname*{argmin}_{m \in M} h(x, m)$$



Distribution & fault tolerance

- Storage is O(1/k) per machine
- Communication is O(1) per machine
- Fast snapshots O(1/k) per machine (stop sync and dump state per vertex)

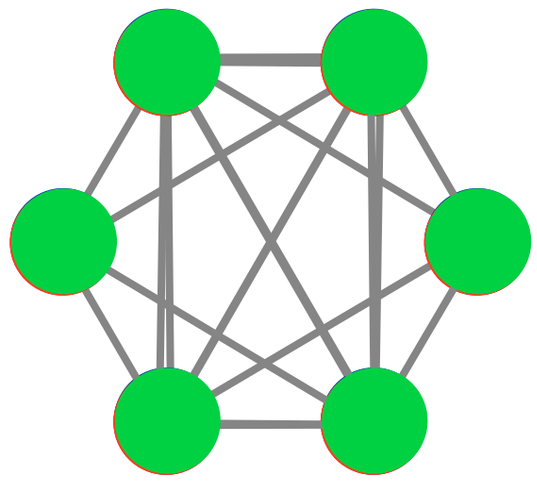
$$m(x) = \underset{m \in M}{\operatorname{argmin}}\, h(x, m)$$



Distribution & fault tolerance

- Storage is O(1/k) per machine
- Communication is O(1) per machine
- Fast snapshots O(1/k) per machine (stop sync and dump state per vertex)

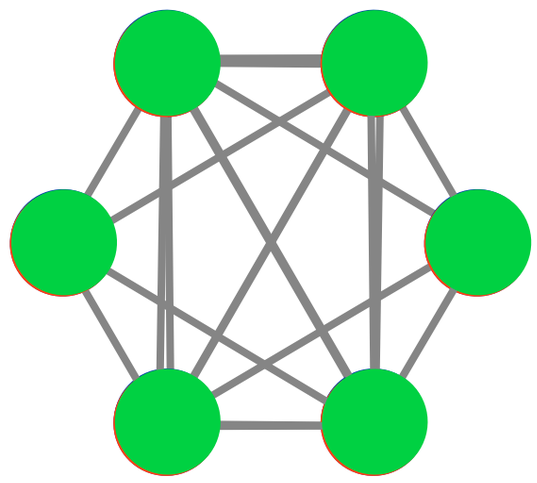
$$m(x) = \underset{m \in M}{\operatorname{argmin}}\, h(x, m)$$



Distribution & fault tolerance

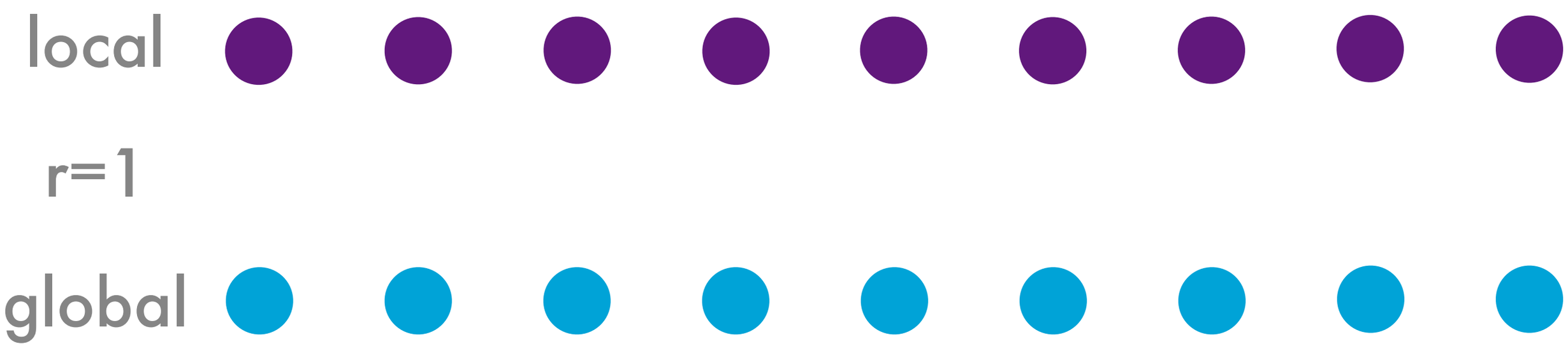
- Storage is O(1/k) per machine
- Communication is O(1) per machine
- Fast snapshots O(1/k) per machine (stop sync and dump state per vertex)
- O(k) open connections per machine
- O(1/k) throughput per machine

$$m(x) = \underset{m \in M}{\operatorname{argmin}}\, h(x, m)$$



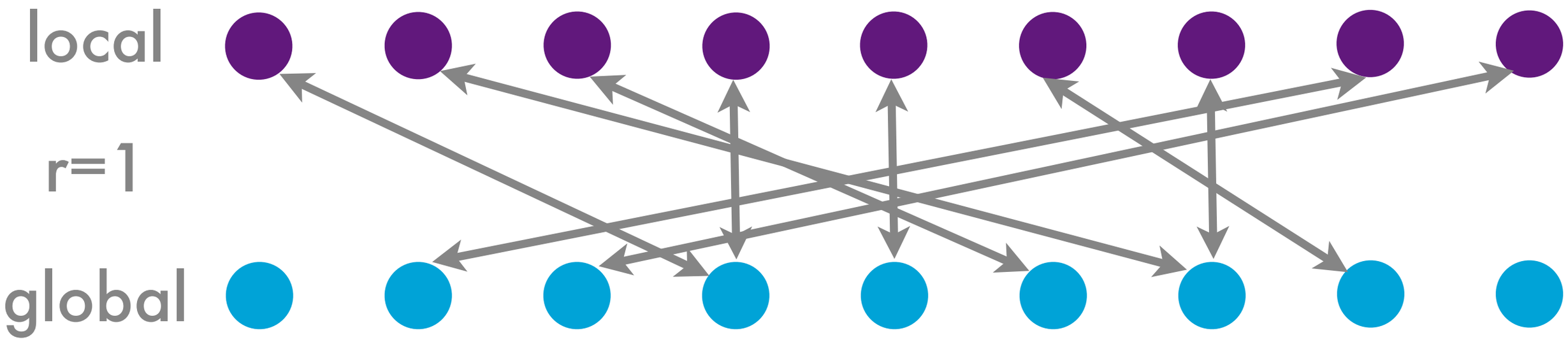
Synchronization

- Data rate between machines is O(1/k)
- Machines operate asynchronously (barrier free)
- Solution
  - Schedule message pairs
    - Communicate with r random machines simultaneously



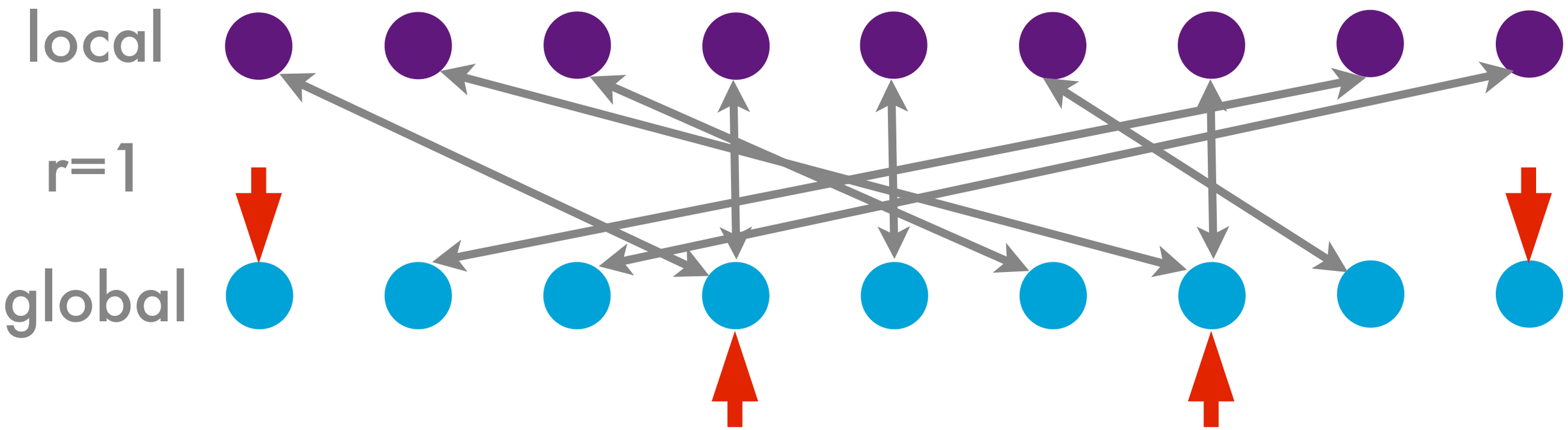
Synchronization

- Data rate between machines is O(1/k)
- Machines operate asynchronously (barrier free)
- Solution
  - Schedule message pairs
    - Communicate with r random machines simultaneously



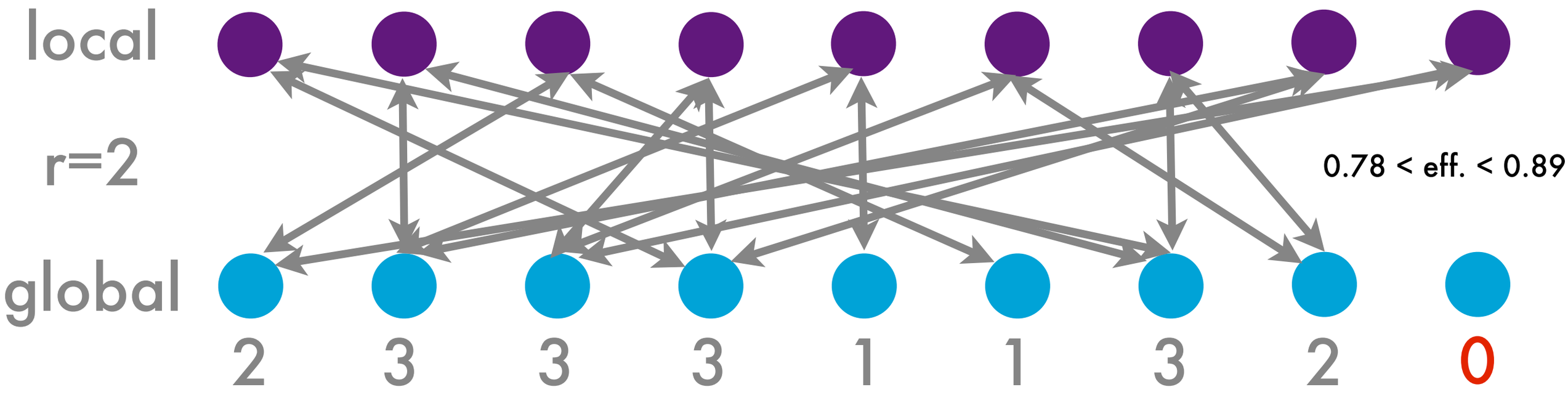
Synchronization

- Data rate between machines is O(1/k)
- Machines operate asynchronously (barrier free)
- Solution
  - Schedule message pairs
    - Communicate with r random machines simultaneously



Synchronization

- Data rate between machines is O(1/k)
- Machines operate asynchronously (barrier free)
- Solution
  - Schedule message pairs
  - Communicate with r random machines simultaneously



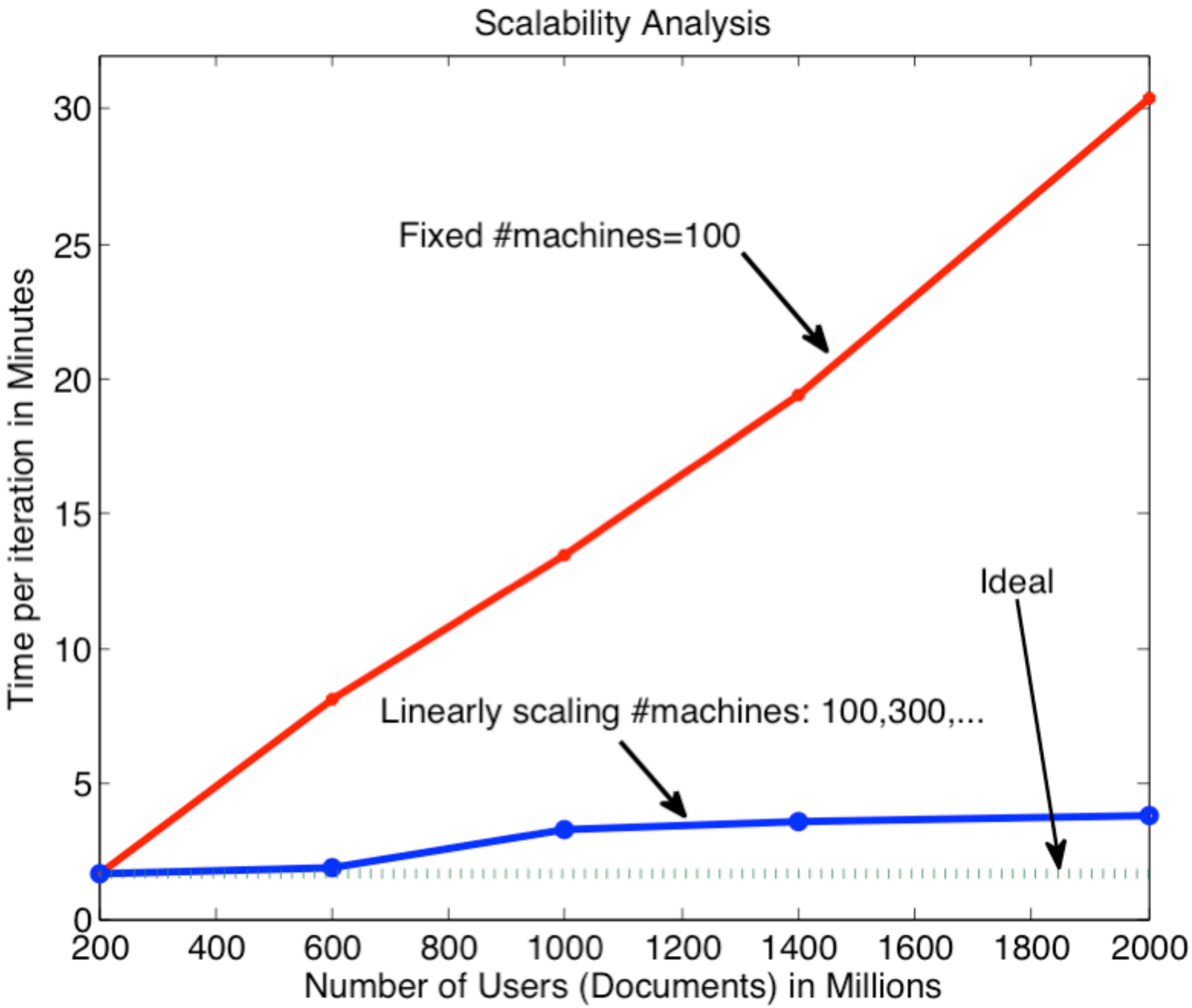
Synchronization

- Data rate between machines is O(1/k)
- Machines operate asynchronously (barrier free)
- Solution
  - Schedule message pairs
  - Communicate with r random machines simultaneously
  - Use Luby-Rackoff PRPG for load balancing
- Efficiency guarantee

$$1 - e^{-r} \sum_{i=0}^{r} \left[1 - \frac{i}{r}\right] \frac{r^i}{i!} \leq \text{Eff} \leq 1 - e^{-r}$$

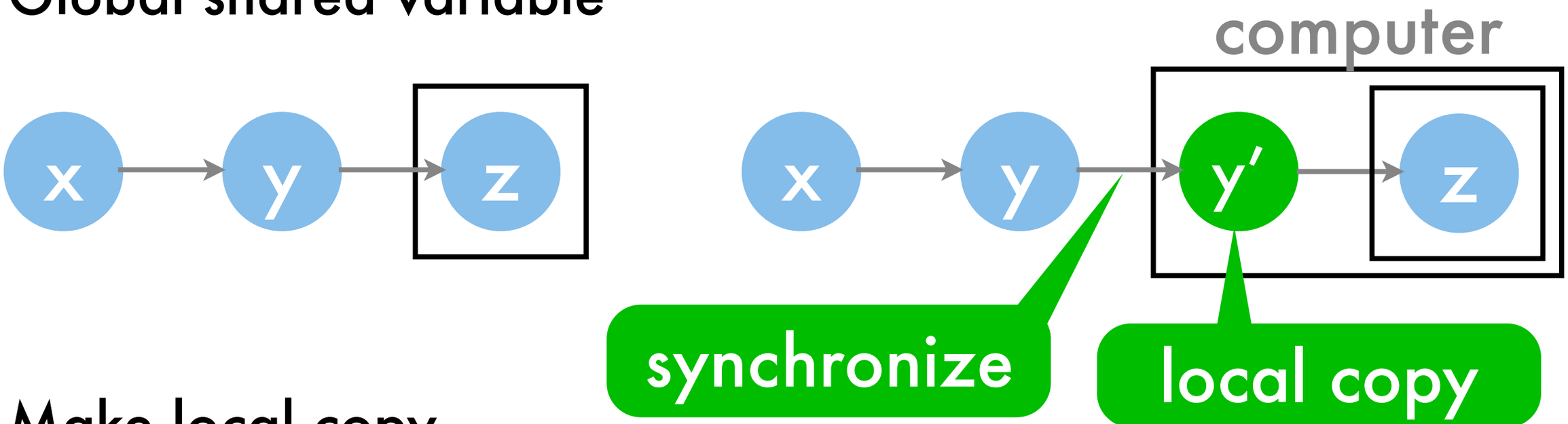
4 simultaneous connections are sufficient

Scalability



Summary
Variable Replication

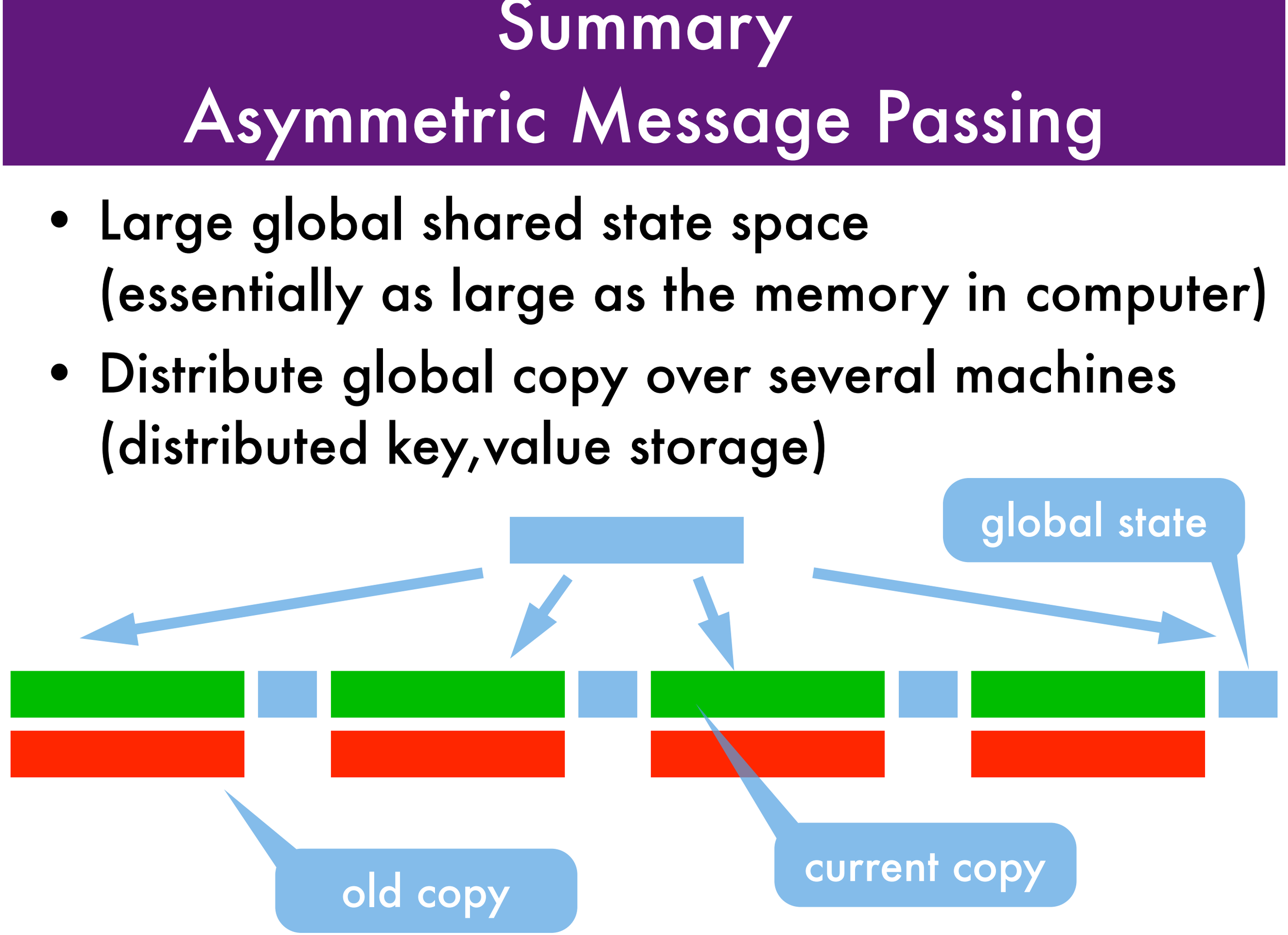
- Global shared variable



- Make local copy
  - Distributed (key,value) storage table for global copy
  - Do all bookkeeping locally (store old versions)
  - Sync local copies asynchronously using message passing (no global locks are needed)
- This is an approximation!

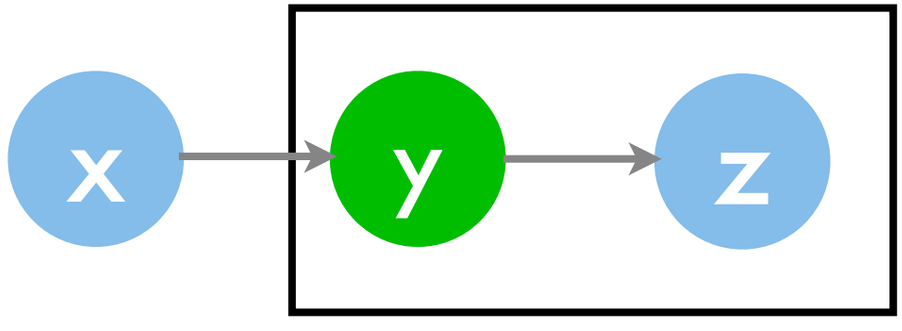
Summary
Asymmetric Message Passing

- Large global shared state space (essentially as large as the memory in computer)
- Distribute global copy over several machines (distributed key,value storage)

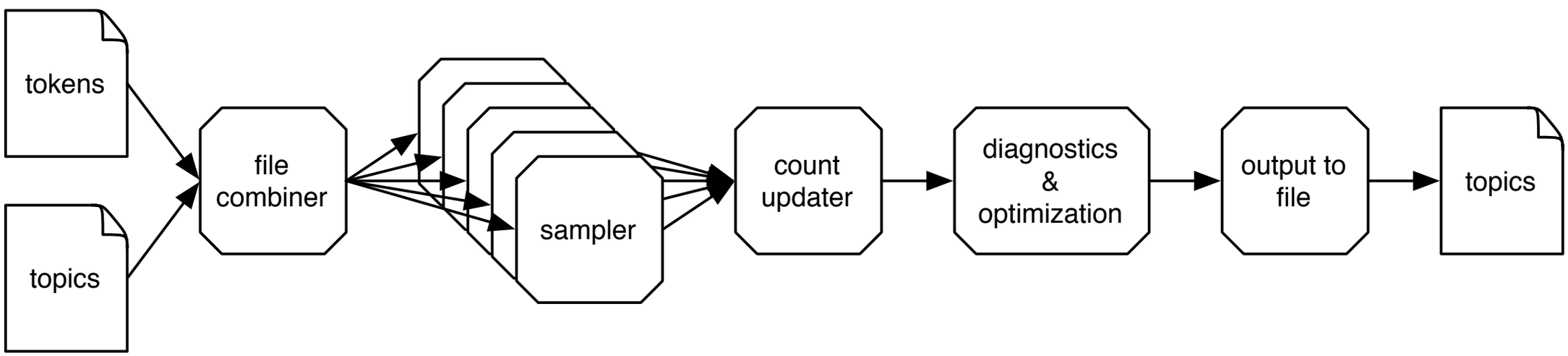


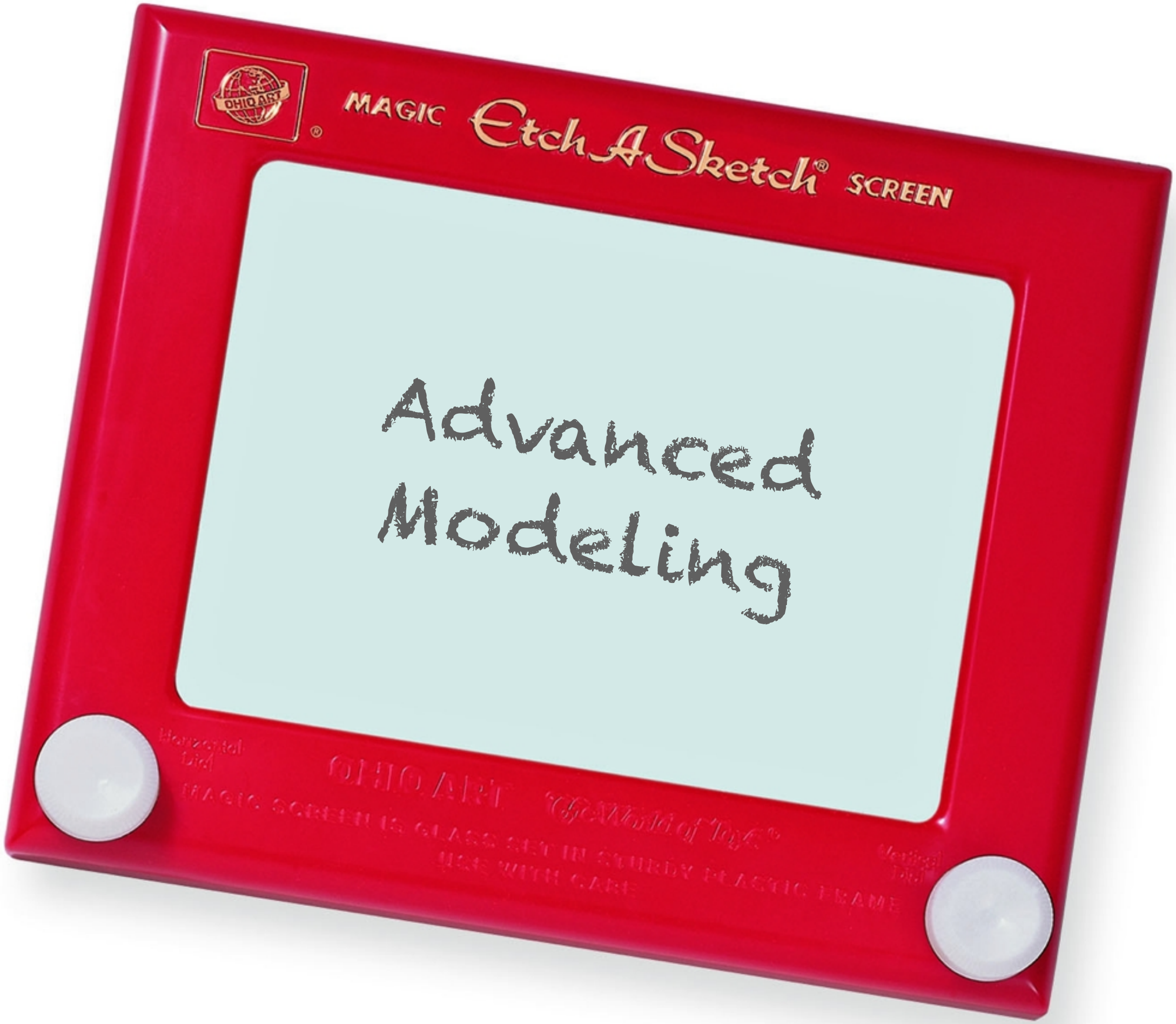
Summary
Out of core storage

- Very large state space



- Gibbs sampling requires us to traverse the data sequentially many times (think 1000x)
- Stream local data from disk and update coupling variable each time local data is accessed
- This is exact

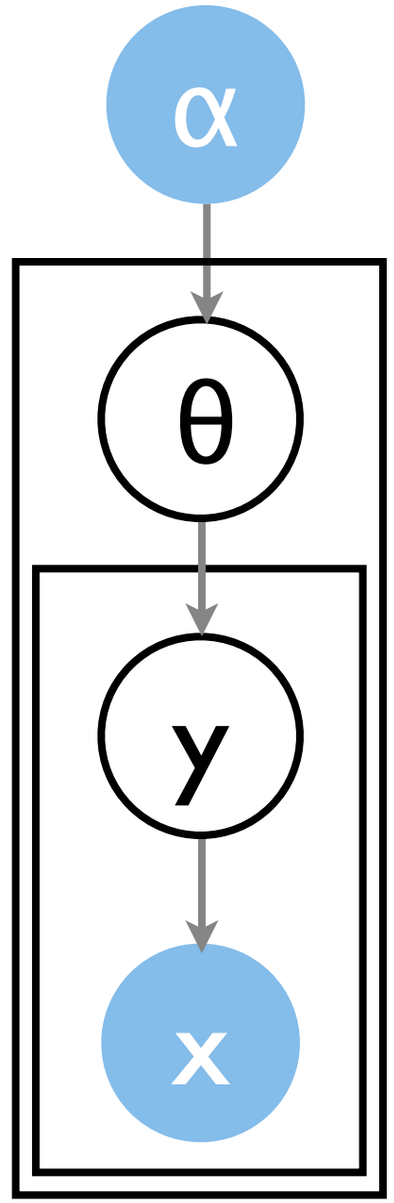



MAGIC Etch A Sketch® SCREEN
Advanced
Modeling

Advances in Representation

Extensions to topic models

- Prior over document topic vector
  - Usually as Dirichlet distribution
  - Use correlation between topics (CTM)
  - Hierarchical structure over topics
- Document structure
  - Bag of words
  - n-grams (Li & McCallum)
  - Simplical Mixture (Girolami & Kaban)
- Side information
  - Upstream conditioning (Mimno & McCallum)
  - Downstream conditioning (Petterson et al.)
  - Supervised LDA (Blei and McAulliffe 2007; Lacoste, Sha and Jordan 2008; Zhu, Ahmed and Xing 2009)



Correlated topic models

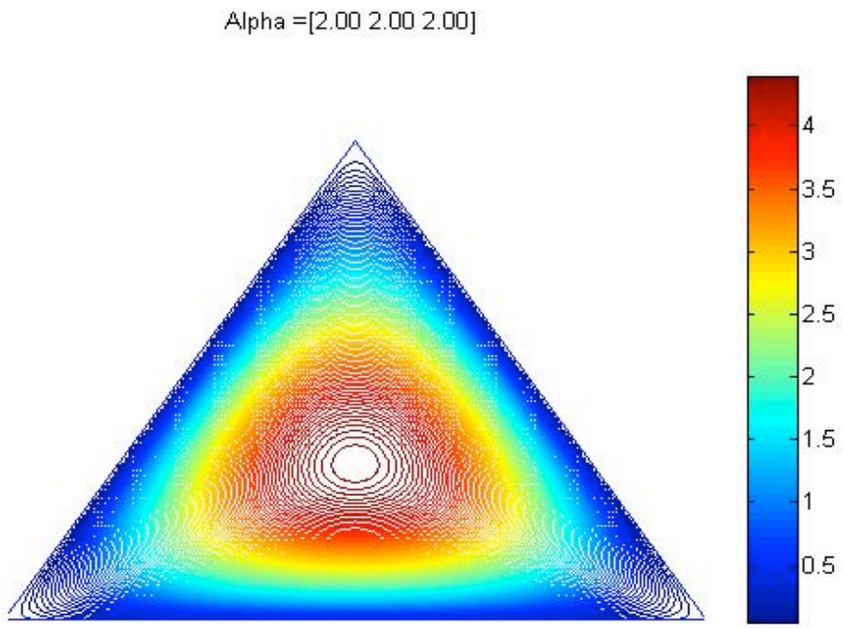
- Dirichlet distribution
  - Can only model which topics are hot
  - Does not model relationships between topics

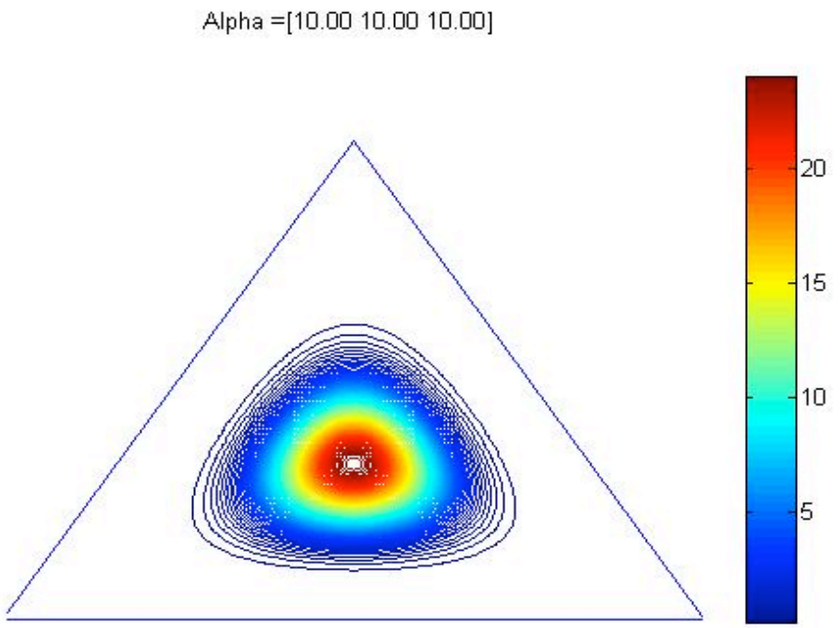
Correlated topic models

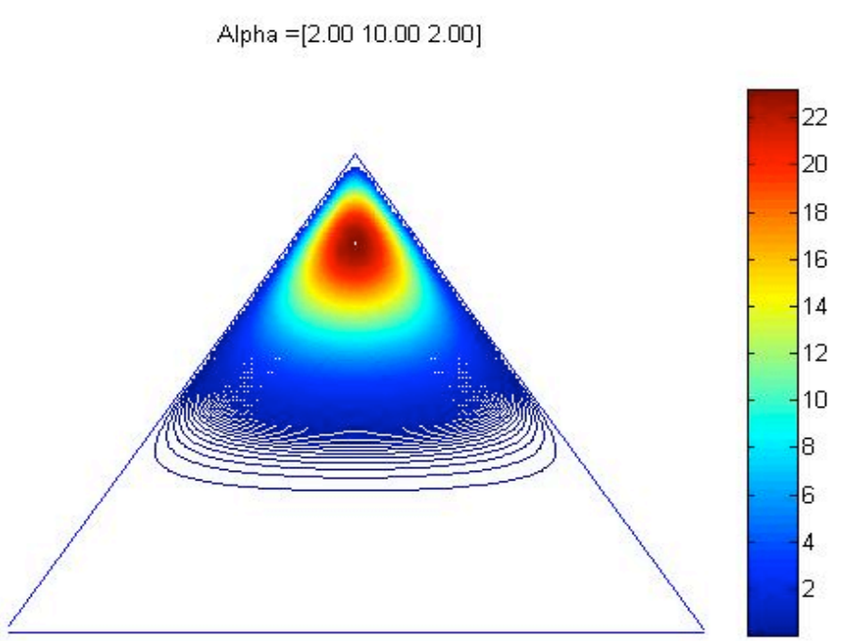
- Dirichlet distribution
  - Can only model which topics are hot
  - Does not model relationships between topics
- Key idea
  - We expect to see documents about sports and health but not about sports and politics
  - Uses a logistic normal distribution as a prior
- Conjugacy is no longer maintained
- Inference is harder than in LDA

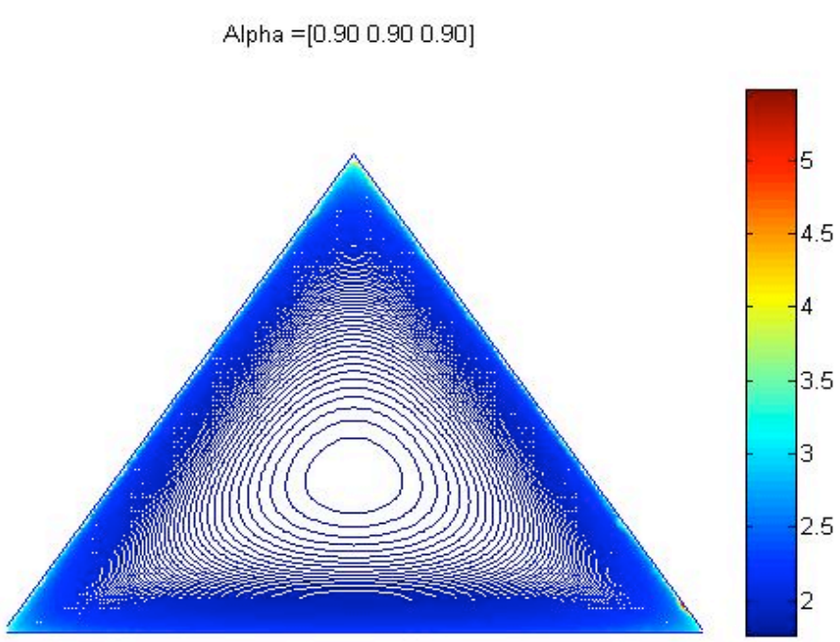
Blei & Lafferty 2005; Ahmed & Xing 2007

Dirichlet prior on topics

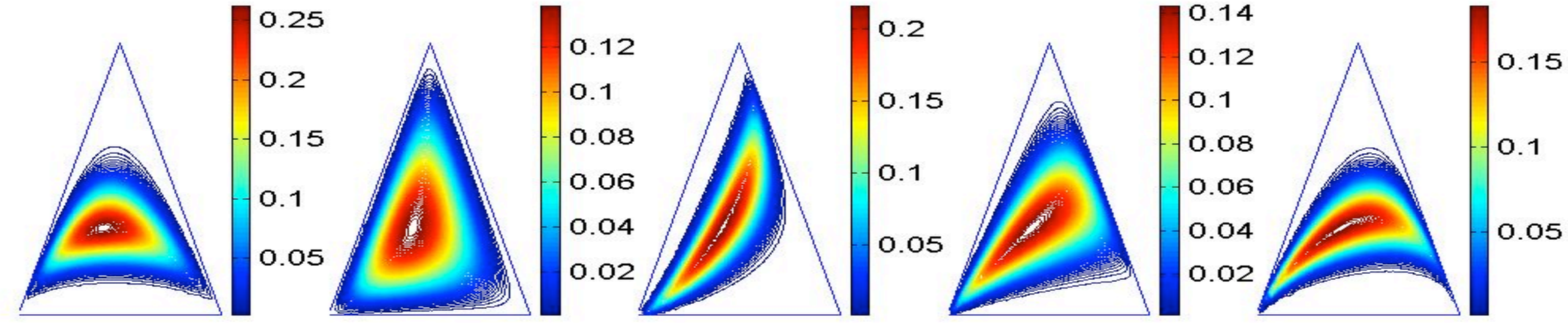


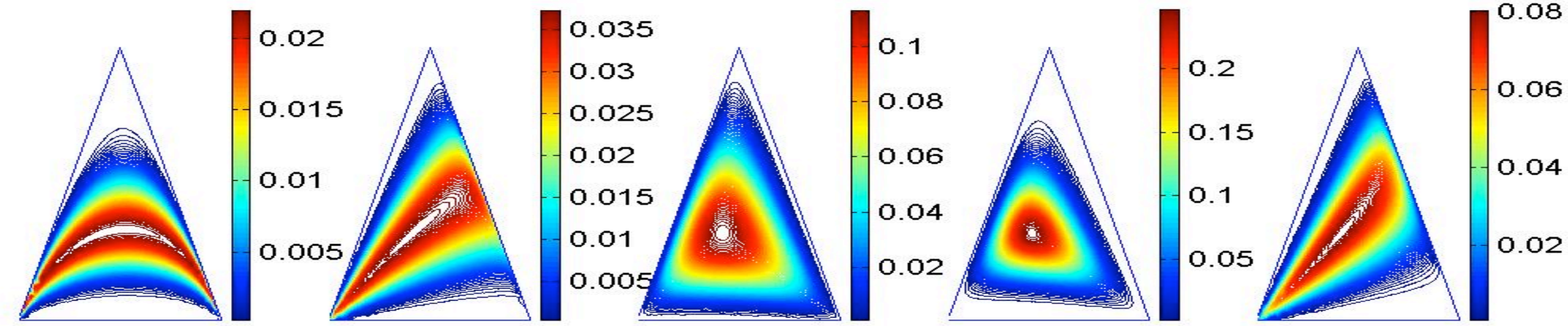






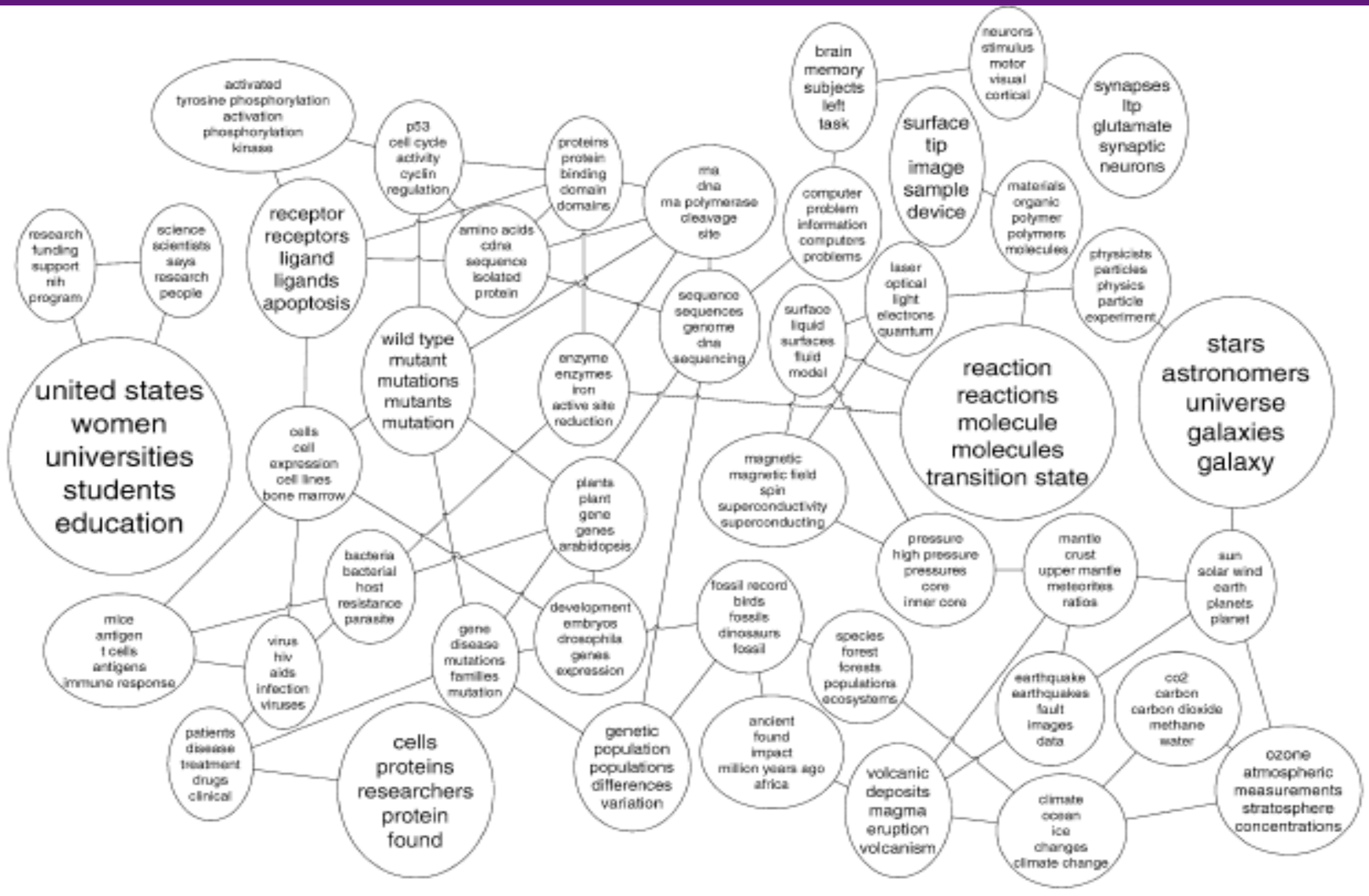
Log-normal prior on topics





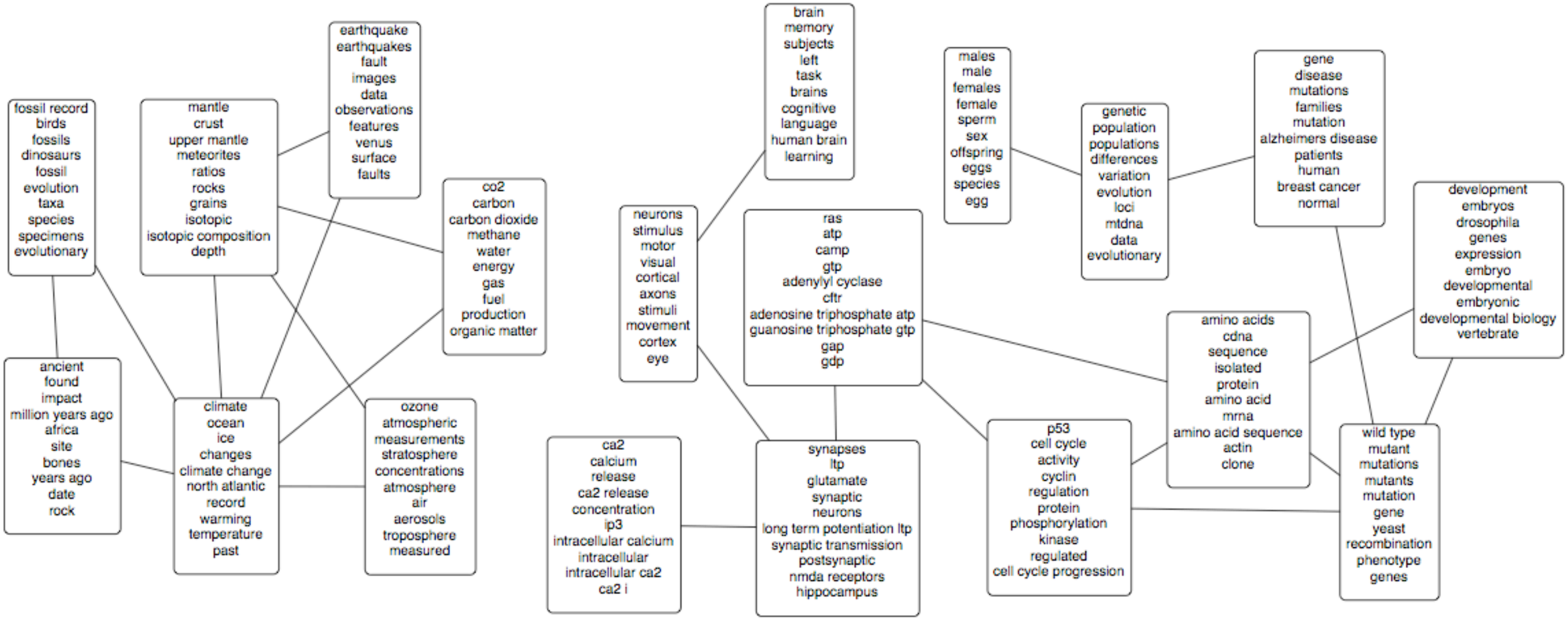
$$\theta = e^{\eta - g(\eta)} \quad \textbf{with} \quad \eta \sim \mathcal{N}(\mu, \Sigma)$$

Correlated topics



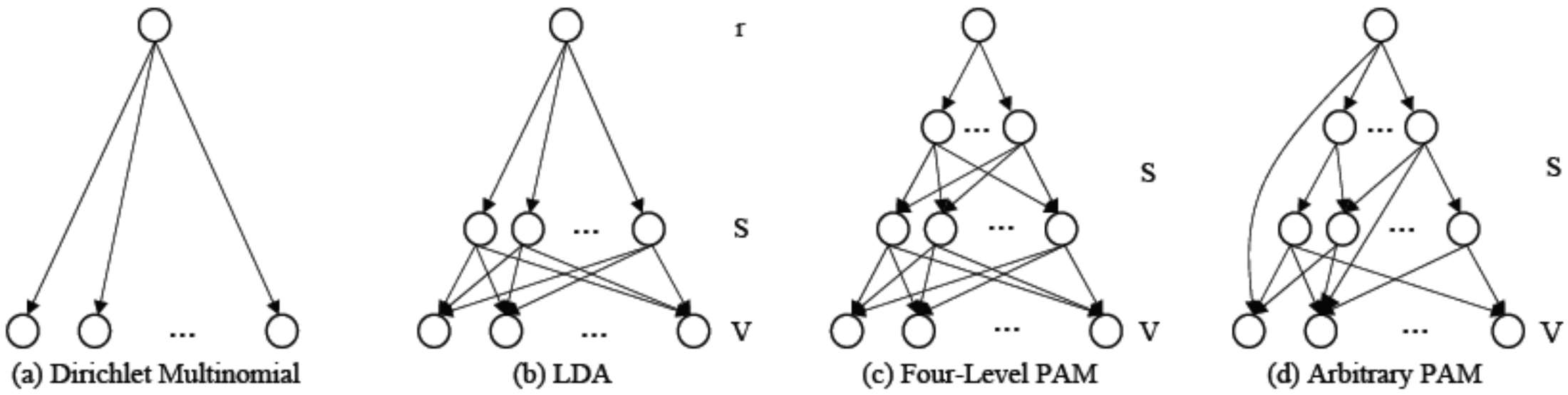
Blei and Lafferty 2005

Correlated topics



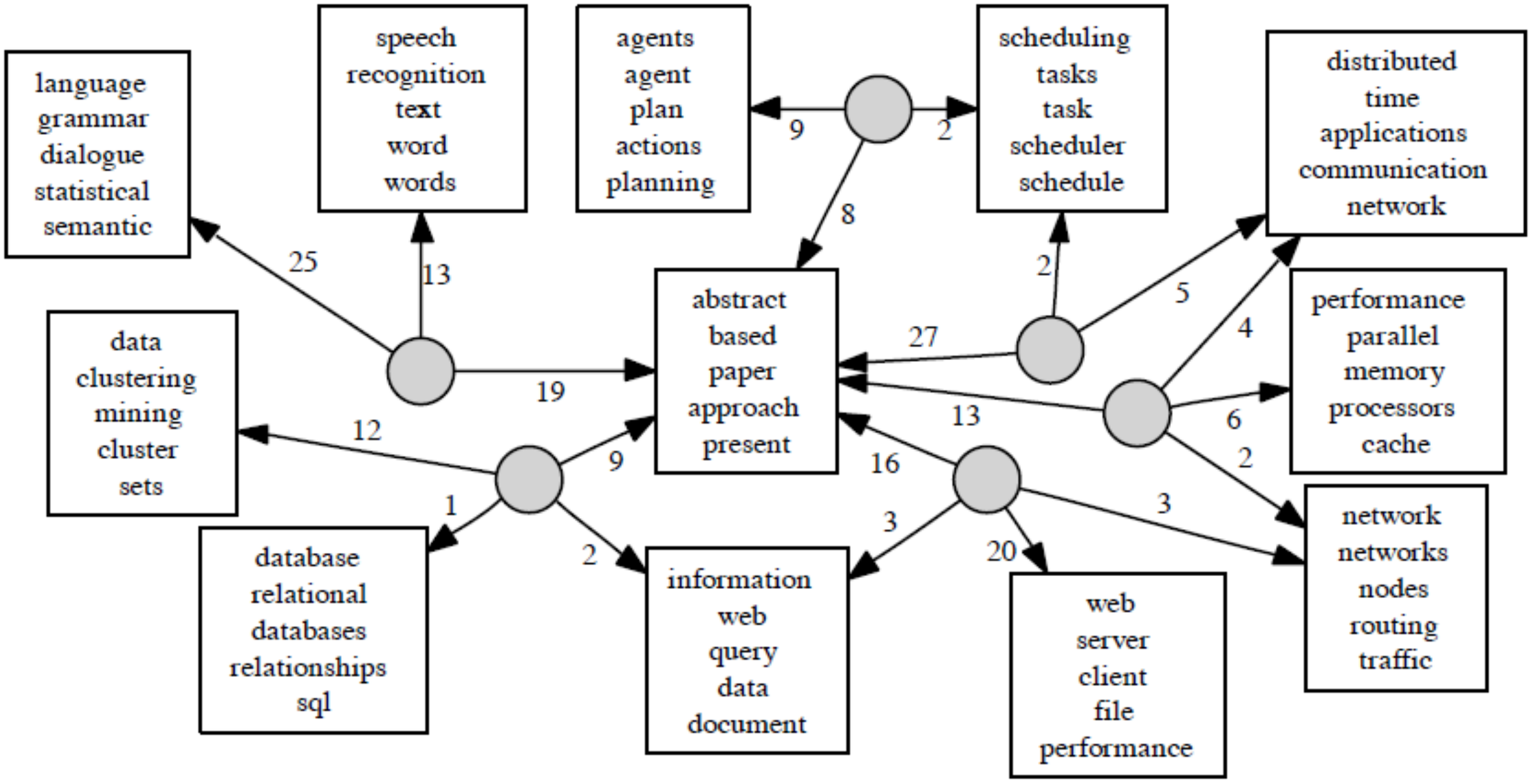
Pachinko Allocation

- Model the prior as a Directed Acyclic Graph
- Each document is modeled as multiple paths
- To sample a word, first select a path and then sample a word from the final topic
- The topics reside on the leaves of the tree



Li and McCallum 2006

Pachinko Allocation

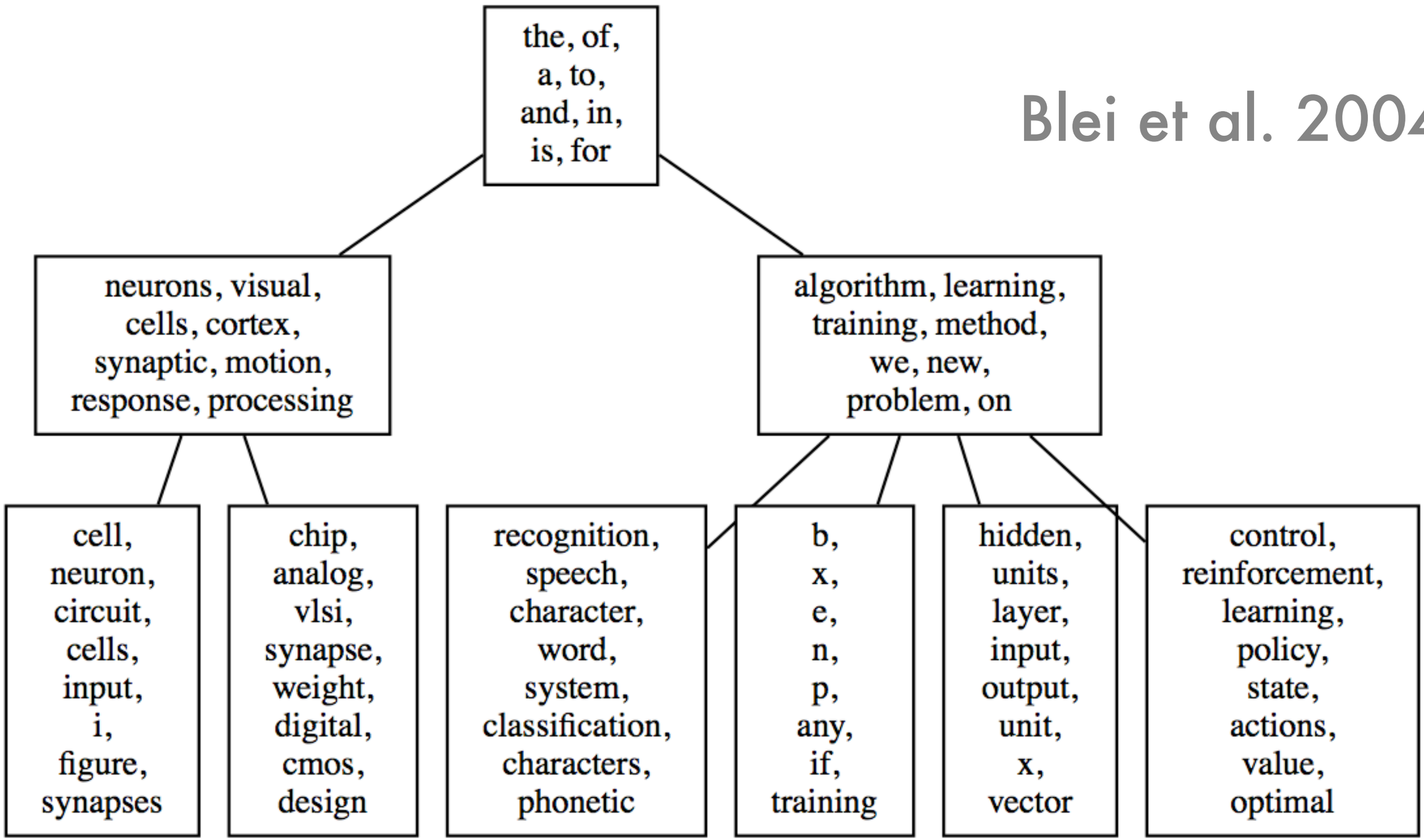


Li and McCallum 2006

Topic Hierarchies

- Topics can appear anywhere in the tree
- Each document is modeled as
  - Single path over the tree (Blei et al., 2004)
  - Multiple paths over the tree (Mimno et al.,2007)

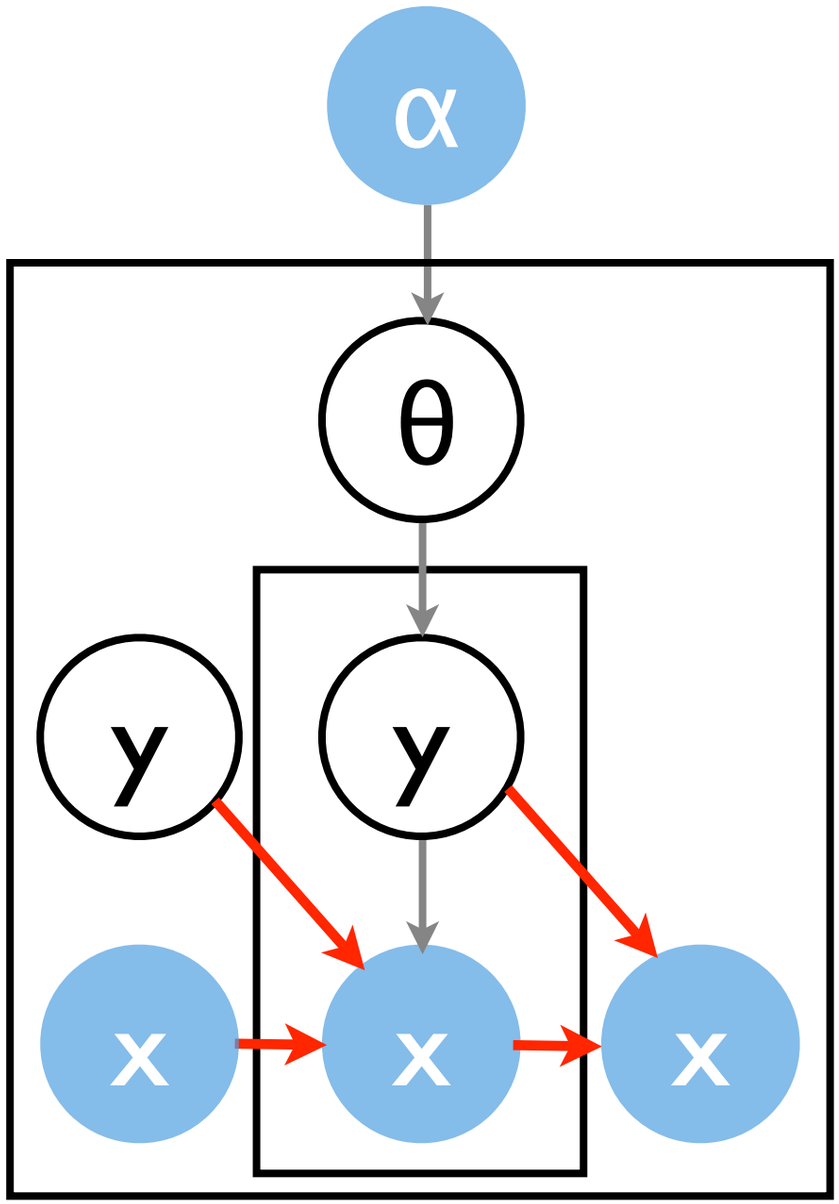
Topic Hierarchies



Blei et al. 2004

Topical n-grams

- Documents as bag of words
- Exploit sequential structure
- N-gram models
  - Capture longer phrases
  - Switch variables to determine segments
- Dynamic programming needed



Girolami & Kaban, 2003; Wallach, 2006; Wang & McCallum, 2007

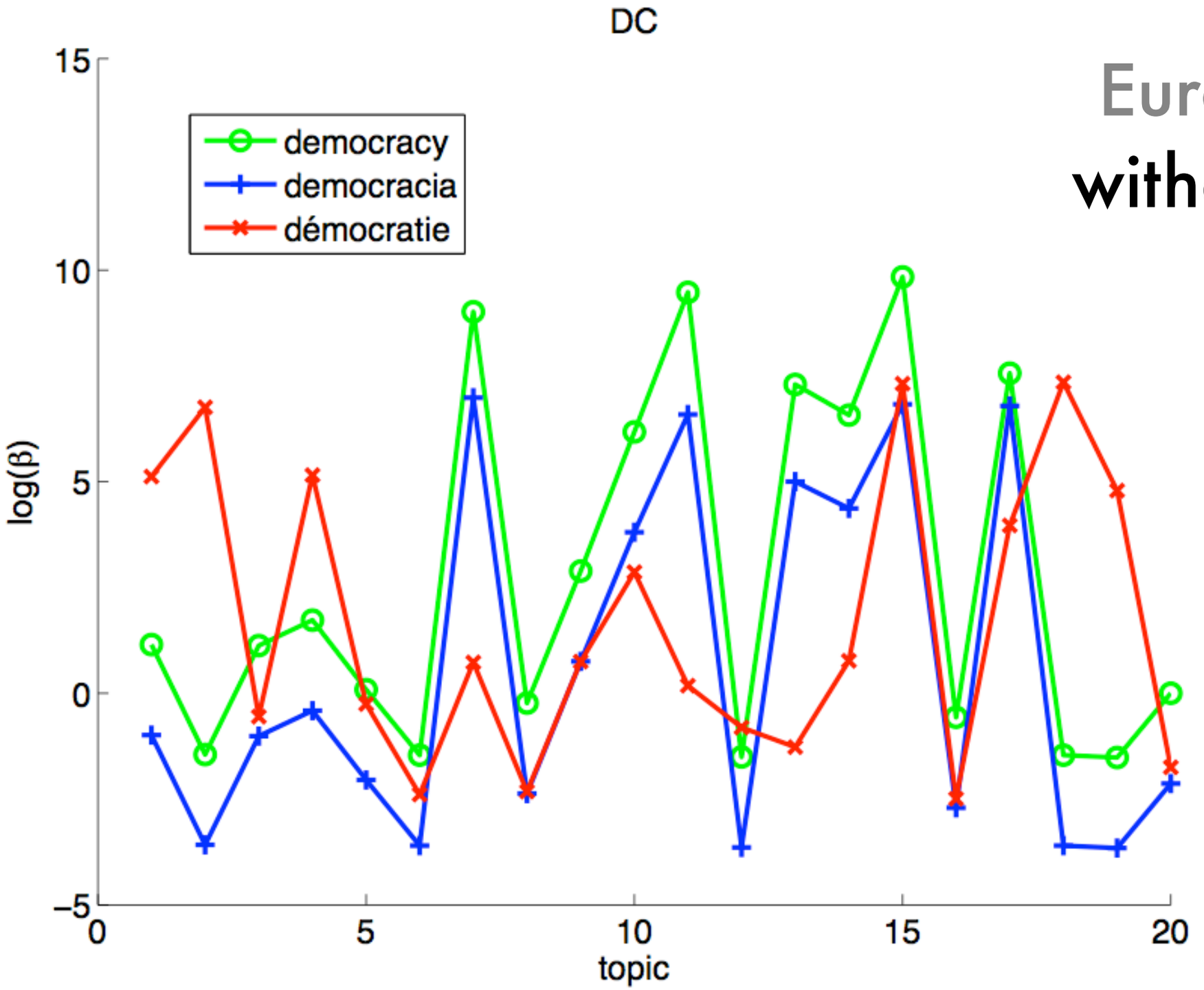
Topic n-grams

Speech Recognition			Support Vector Machines		
LDA	n-gram (2+)	n-gram (1)	LDA	n-gram (2+)	n-gram (1)
recognition	speech recognition	speech	kernel	support vectors	kernel
system	training data	word	linear	test error	training
word	neural network	training	vector	support vector machines	support
face	error rates	system	support	training error	margin
context	neural net	recognition	set	feature space	svm
character	hidden markov model	hmm	nonlinear	training examples	solution
hmm	feature vectors	speaker	data	decision function	kernels
based	continuous speech	performance	algorithm	cost functions	regularization
frame	training procedure	phoneme	space	test inputs	adaboost
segmentation	continuous speech recognition	acoustic	pca	kkt conditions	test
training	gamma filter	words	function	leave-one-out procedure	data
characters	hidden control	context	problem	soft margin	generalization
set	speech production	systems	margin	bayesian transduction	examples
probabilities	neural nets	frame	vectors	training patterns	cost
features	input representation	trained	solution	training points	convex
faces	output layers	sequence	training	maximum margin	algorithm
words	training algorithm	phonetic	svm	strictly convex	working
frames	test set	speakers	kernels	regularization operators	feature
database	speech frames	mlp	matrix	base classifiers	sv
mlp	speaker dependent	hybrid	machines	convex optimization	functions

Side information

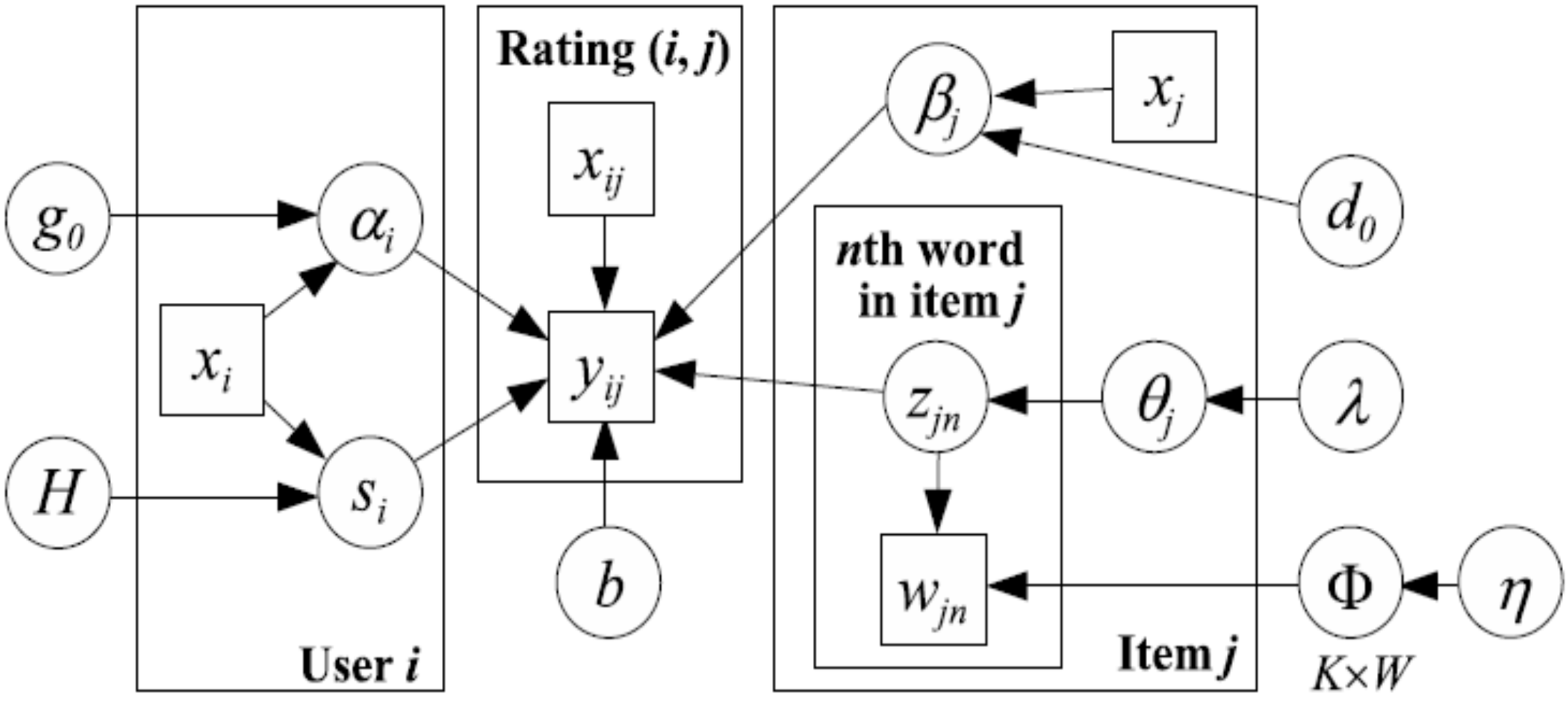
- Upstream conditioning (Mimno et al., 2008)
  - Document features are informative for topics
  - Estimate topic distribution e.g. based on authors, links, timestamp
- Downstream conditioning (Petterson et al., 2010)
  - Word features are informative on topics
  - Estimate topic distribution for words e.g. based on dictionary, lexical similarity, distributional similarity
- Class labels (Blei and McAulliffe 2007; Lacoste, Sha and Jordan 2008; Zhu, Ahmed and Xing 2009)
  - Joint model of unlabeled data and labels
  - Joint likelihood - semisupervised learning done right!

Downstream conditioning



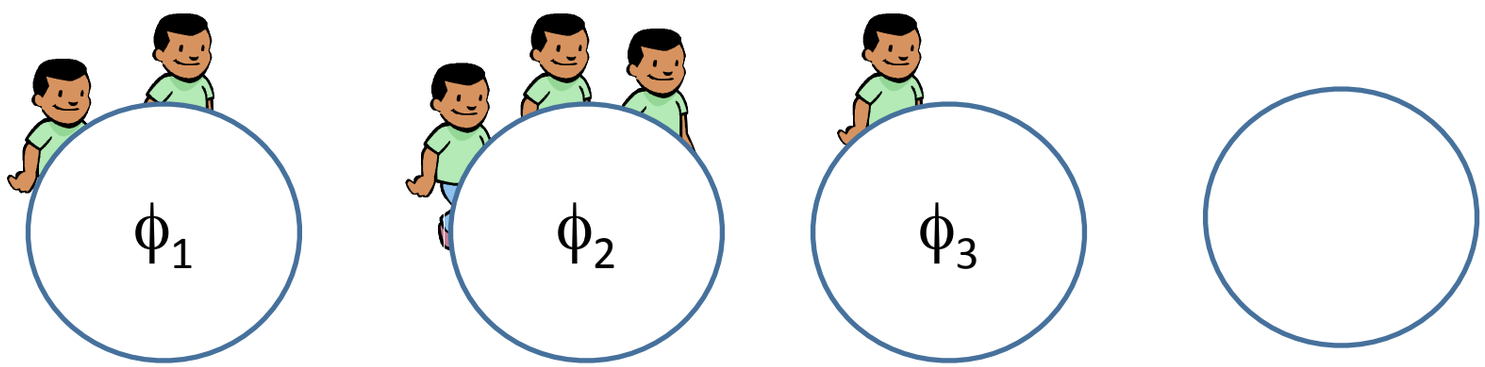
Europarl corpus
without alignment

Recommender Systems



Agarwal & Chen, 2010

Chinese Restaurant Process



Problem

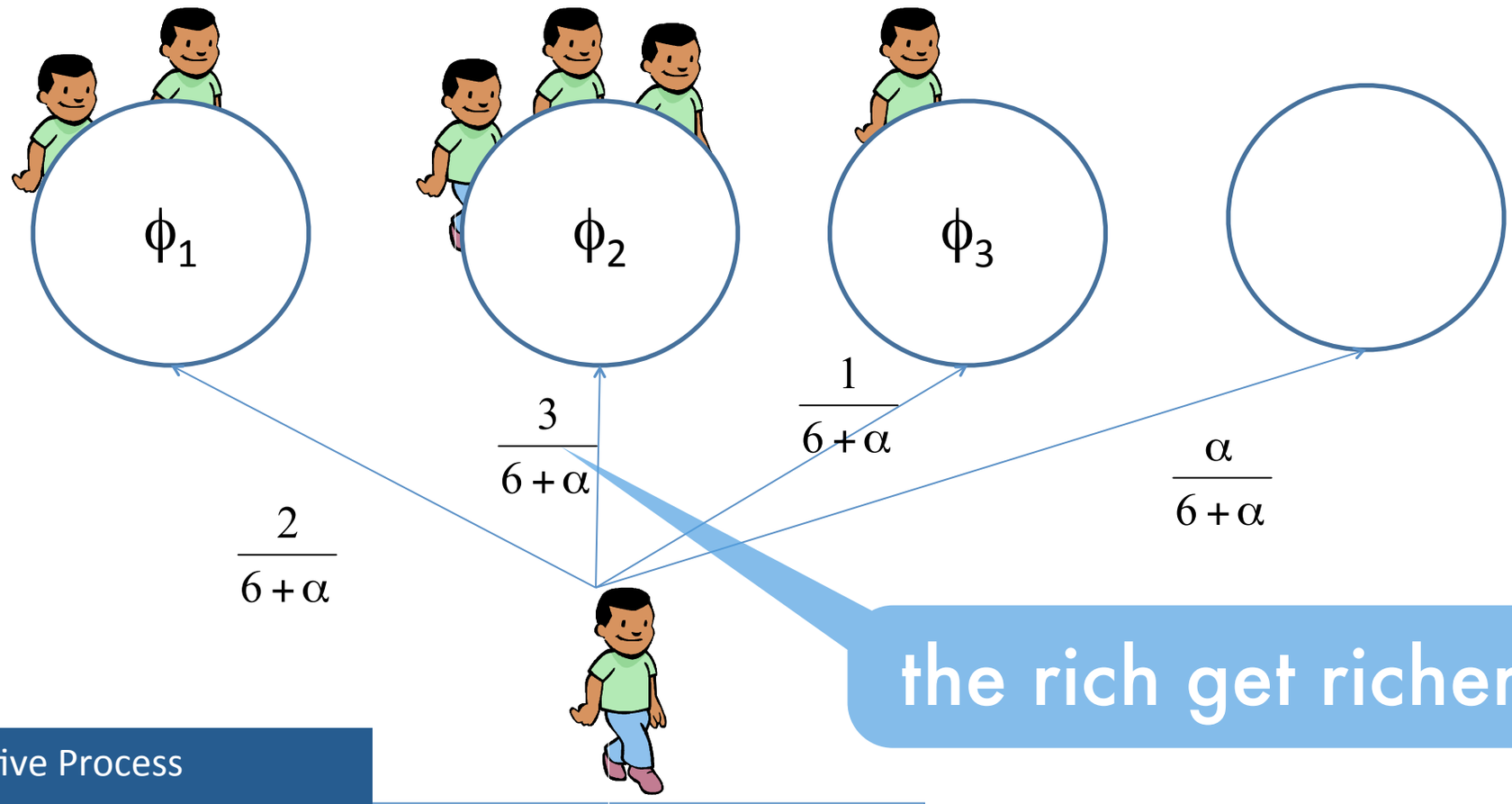
- How many clusters should we pick?
- How about a prior for infinitely many clusters?
- Finite model

$$p(y|Y,\alpha) = \frac{n(y) + \alpha_y}{n + \sum_{y'} \alpha_{y'}}$$

- **Infinite model**
  Assume that the total smoother weight is constant

$$p(y|Y,\alpha) = \frac{n(y)}{n + \sum_{y'} \alpha_{y'}} \textbf{ and } p(\text{new}|Y,\alpha) = \frac{\alpha}{n + \alpha}$$

Chinese Restaurant Metaphor



Generative Process

-For data point x_i

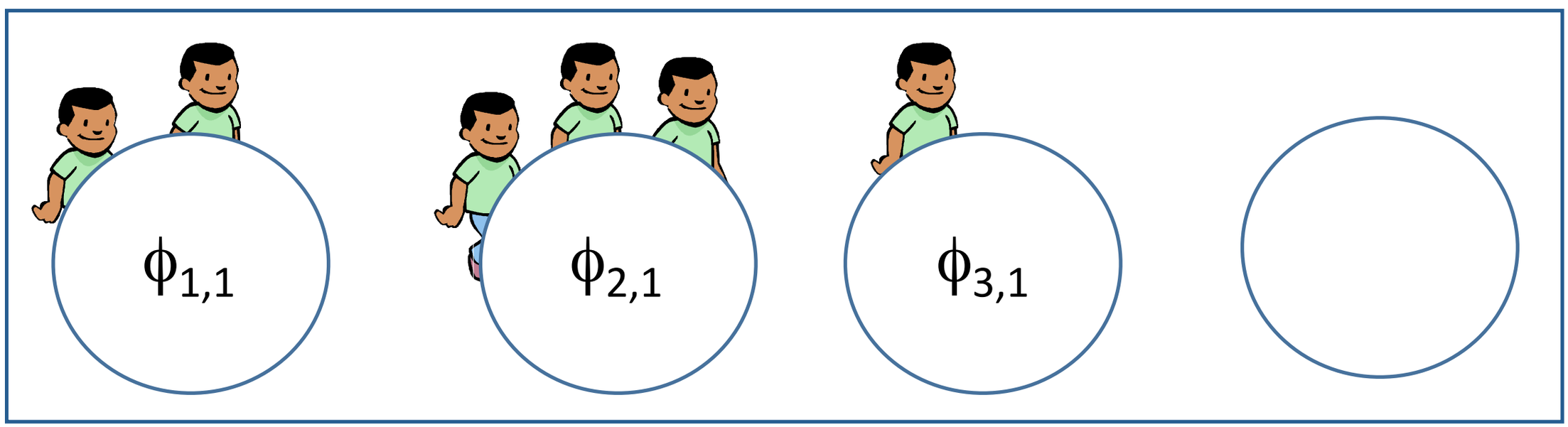
- Choose table $j \propto m_j$ and Sample $x_i \sim f(\phi_j)$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1})$

Pitman; Antoniak; Ishwaran; Jordan et al.; Teh et al.;

Evolutionary Clustering

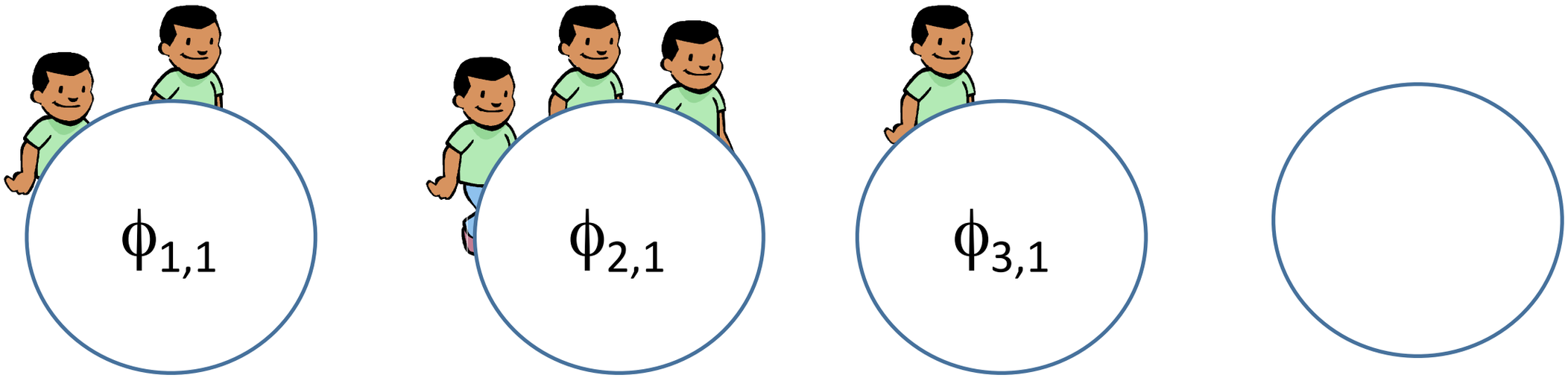
- Time series of objects, e.g. news stories
- Stories appear / disappear
- Want to keep track of clusters automatically

Recurrent Chinese Restaurant Process



T=1

Recurrent Chinese Restaurant Process

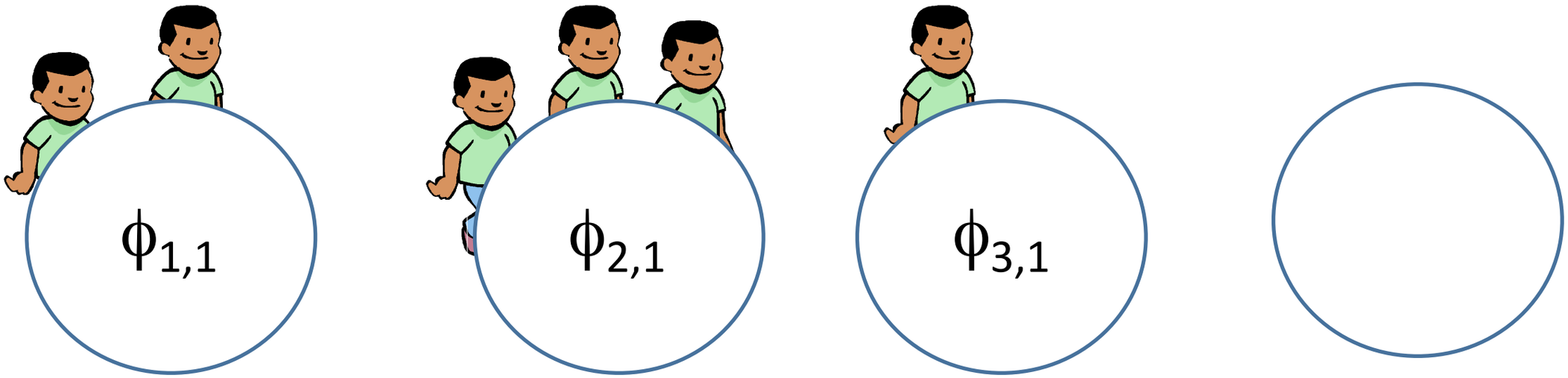


T=1

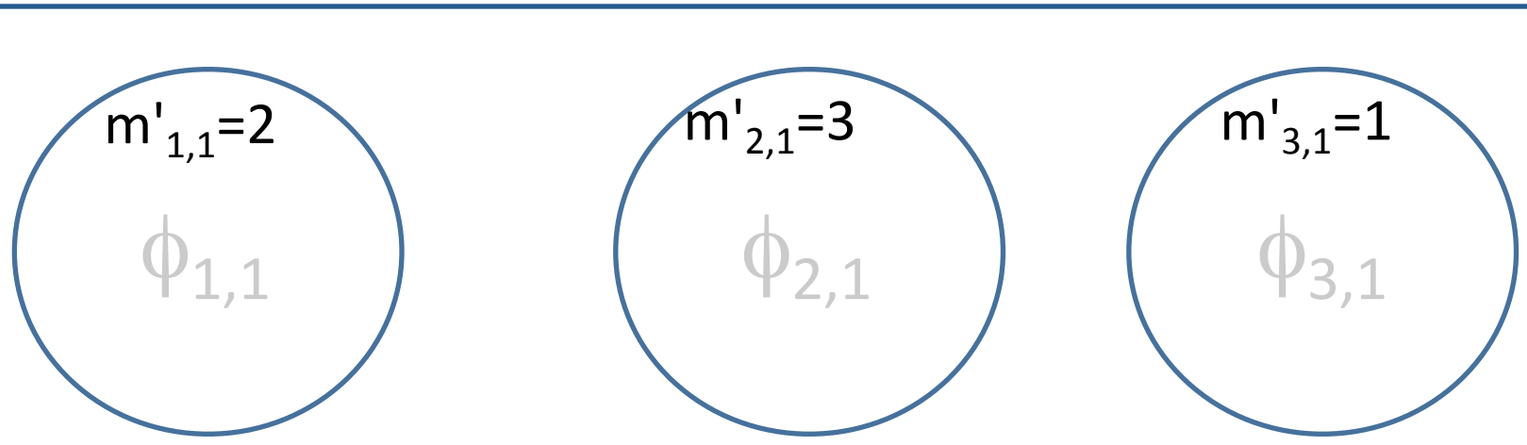
$m'_{1,1}=2$ $m'_{2,1}=3$ $m'_{3,1}=1$

T=2

Recurrent Chinese Restaurant Process

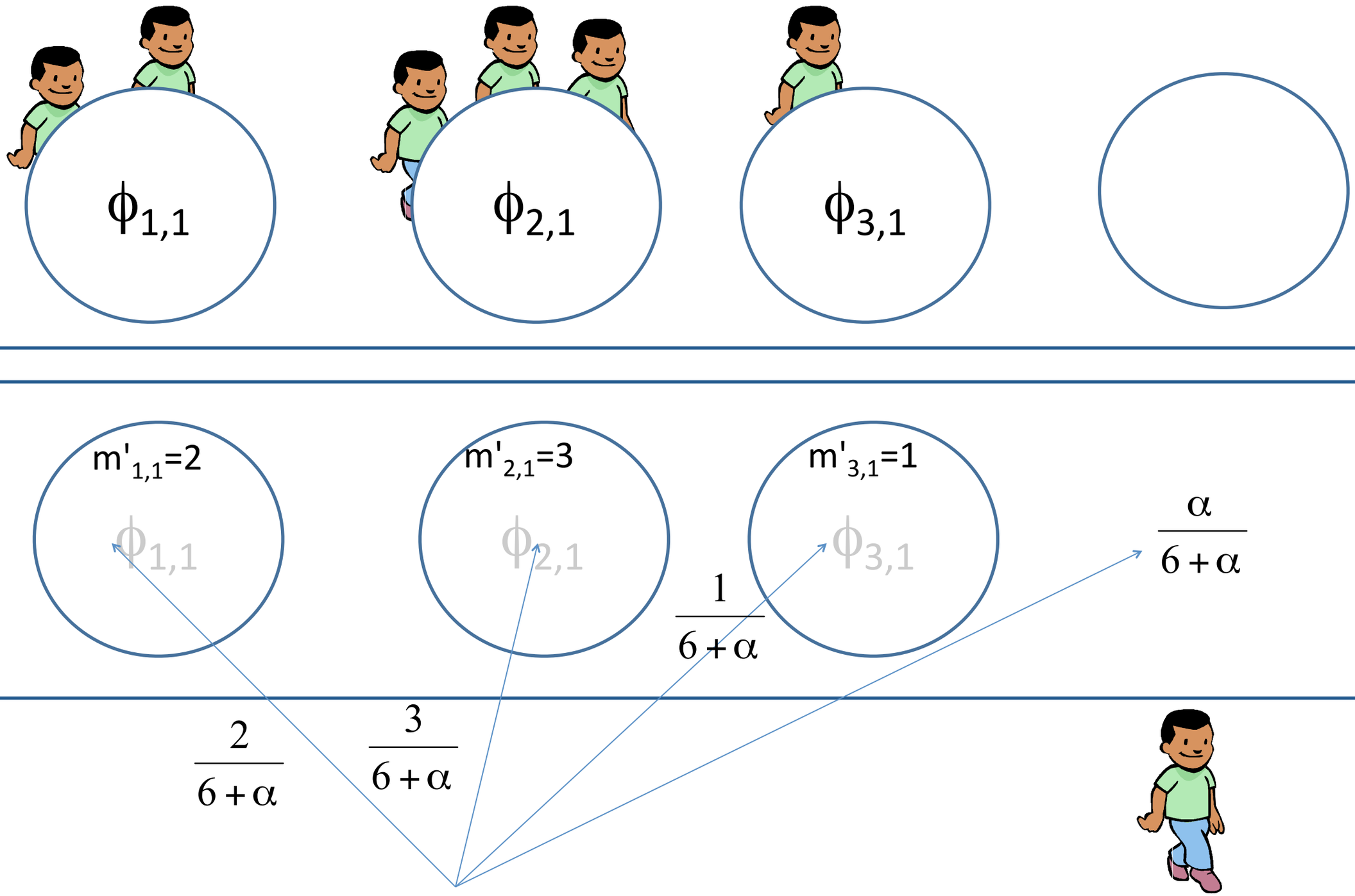


T=1

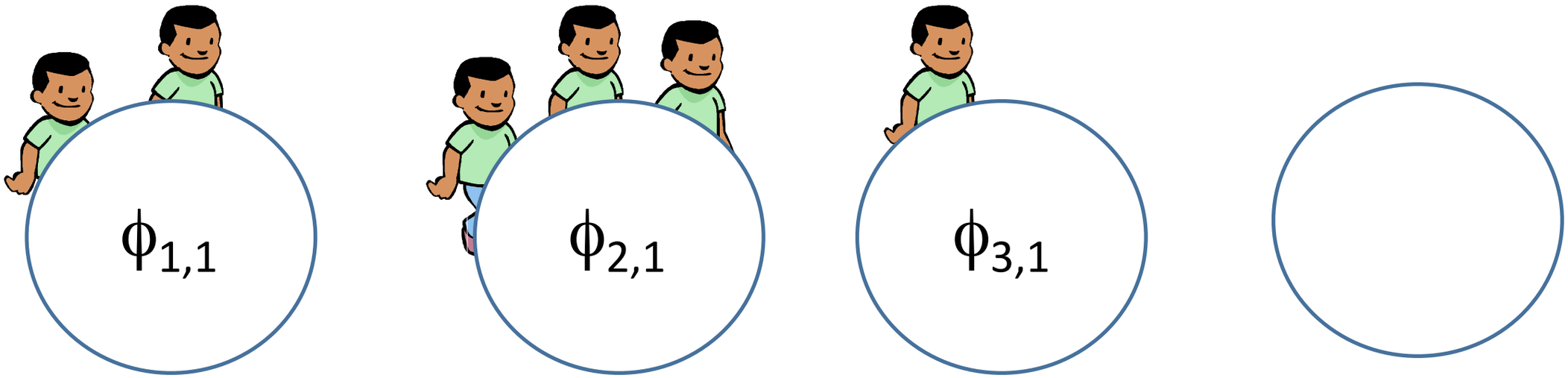


T=2

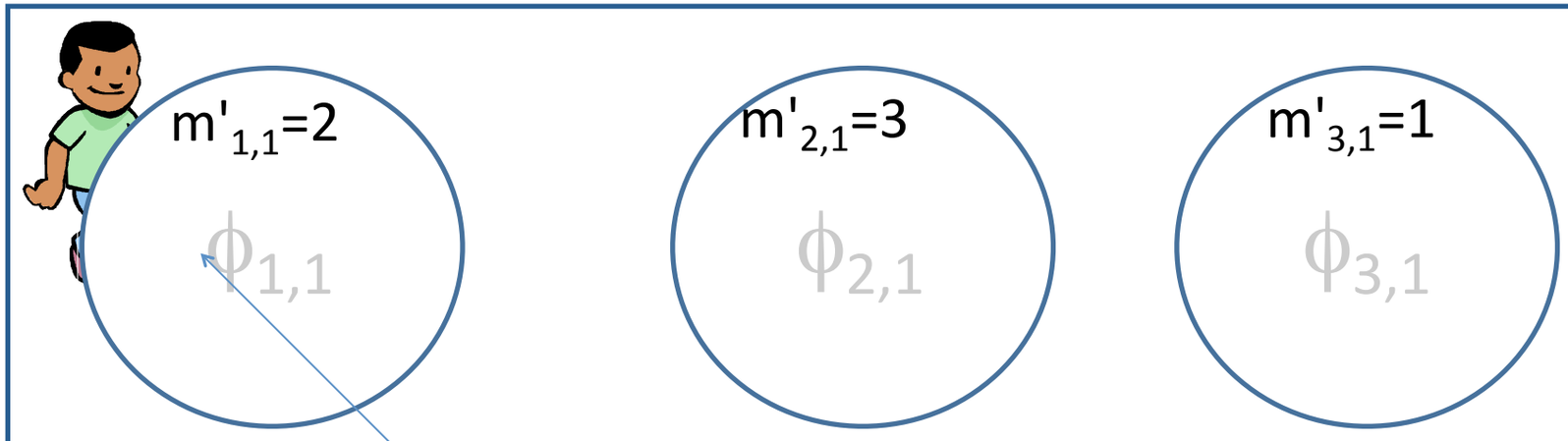
Recurrent Chinese Restaurant Process



Recurrent Chinese Restaurant Process



T=1

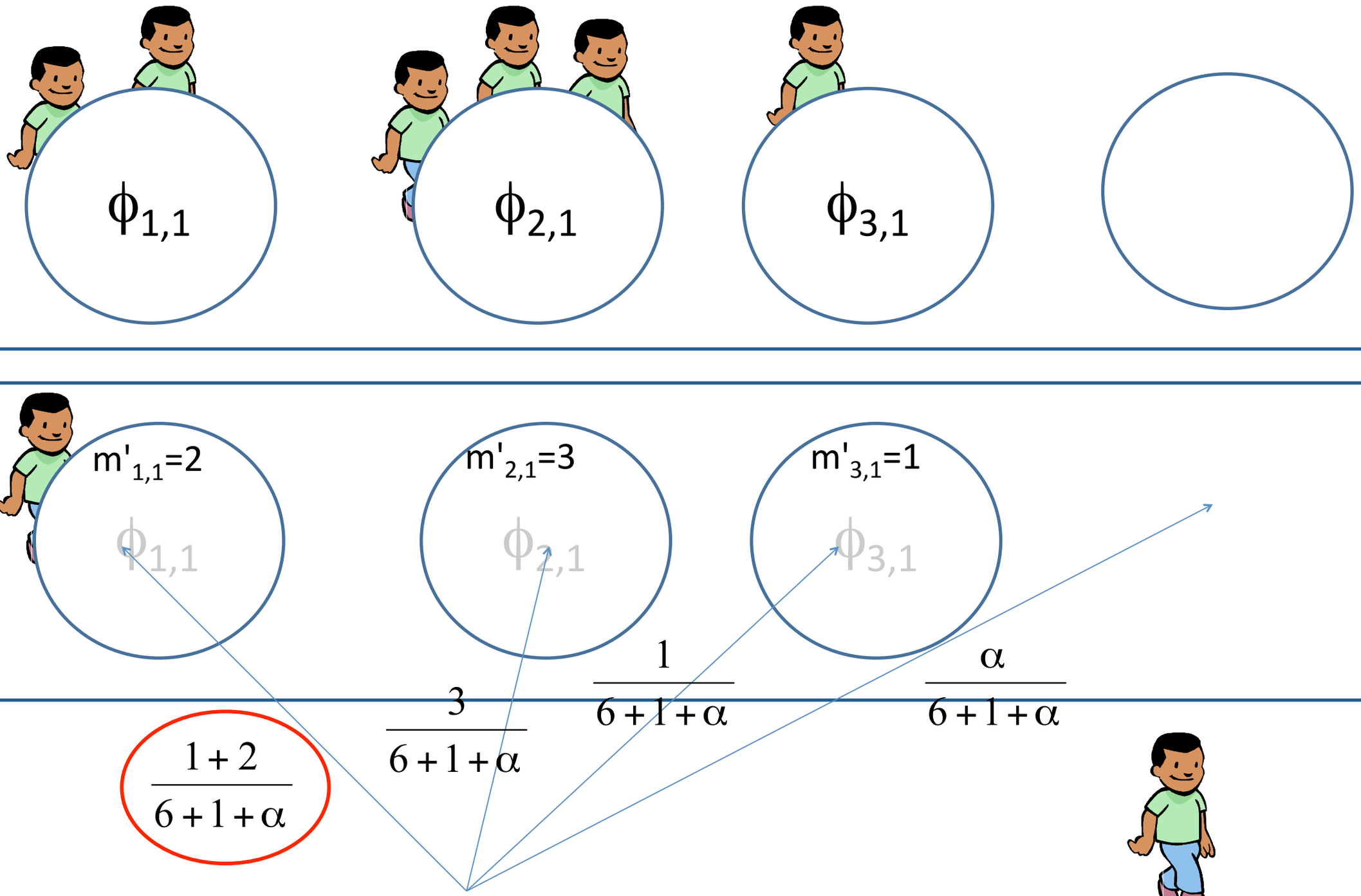


T=2

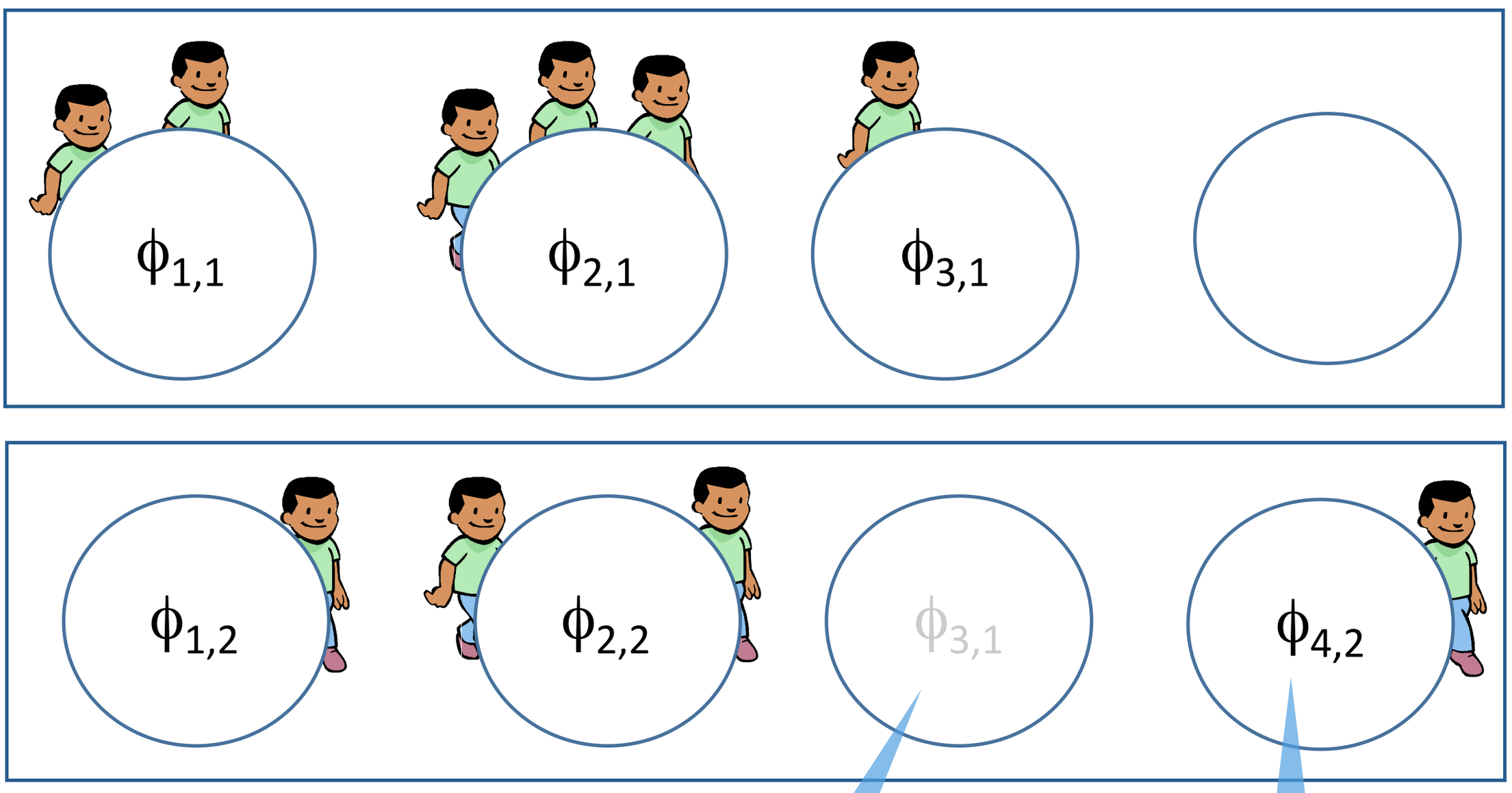
$$\frac{2}{6+\alpha}$$

Sample $\phi_{1,2} \sim P(.|\ \phi_{1,1})$

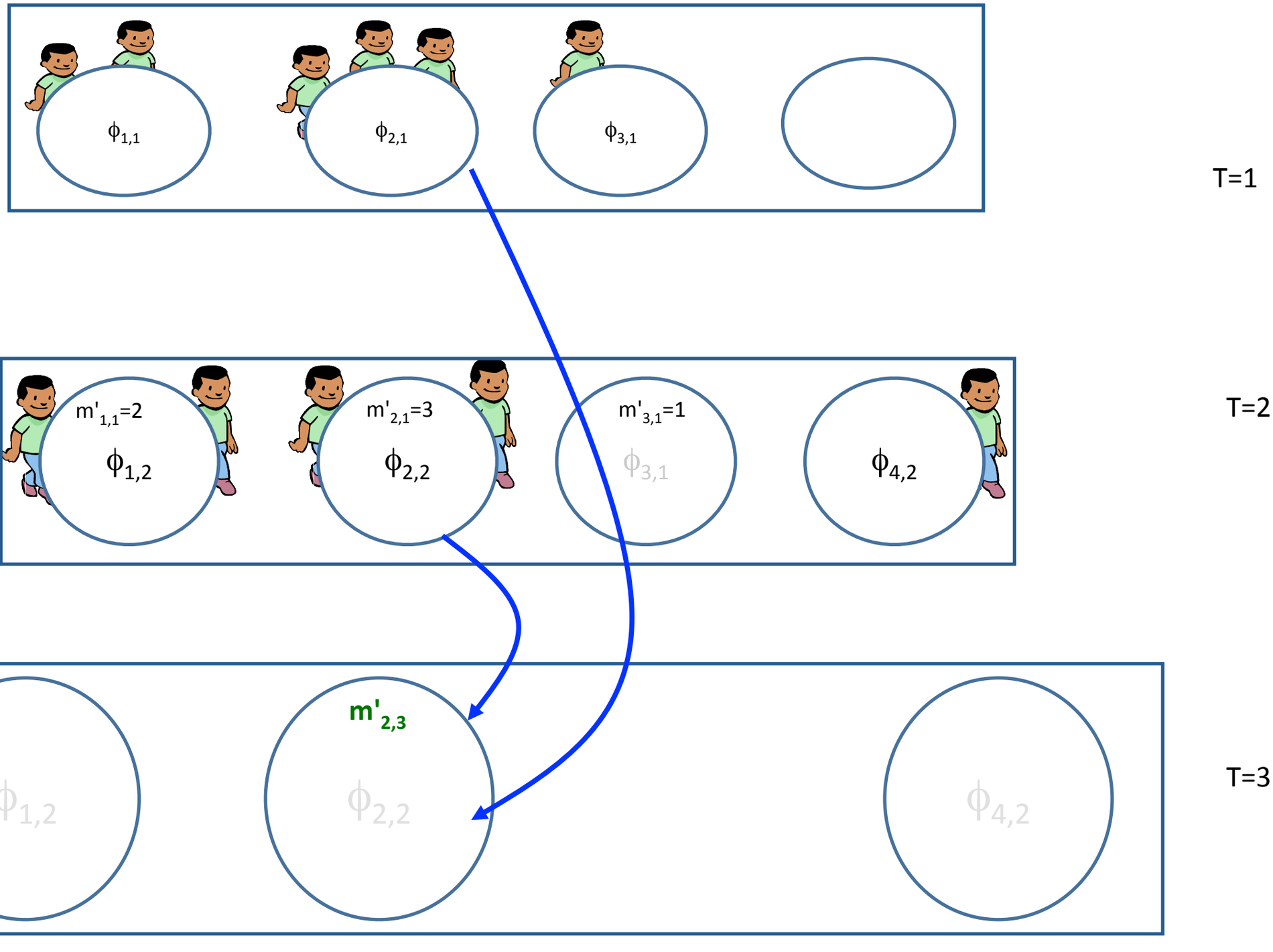
Recurrent Chinese Restaurant Process



Recurrent Chinese Restaurant Process



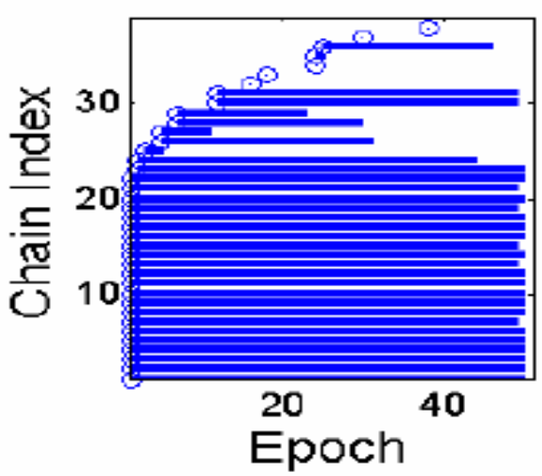
Longer History

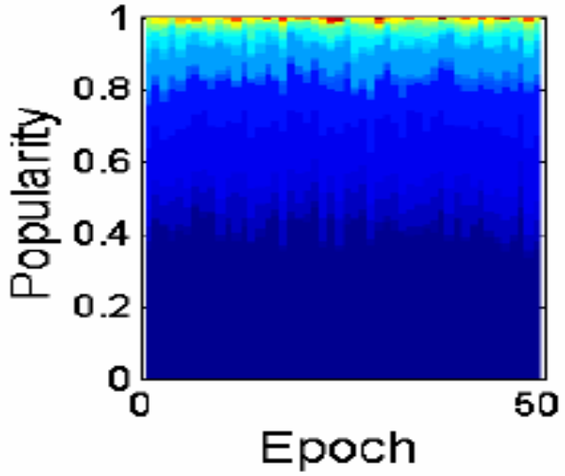


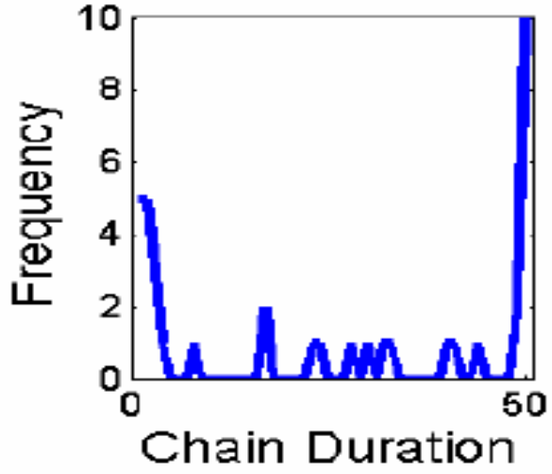
TDPM Generative Power

DPM

$W = \infty$
$\lambda = \infty$

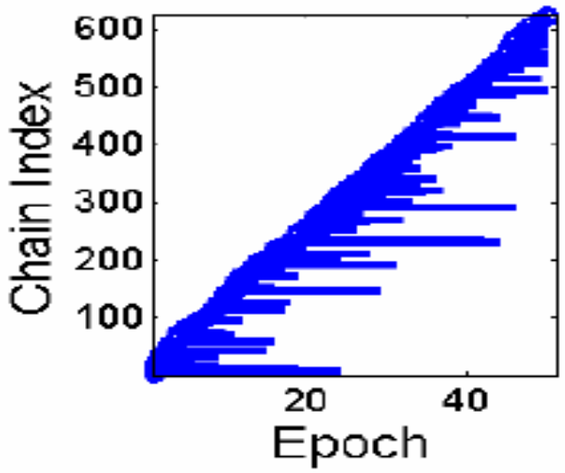


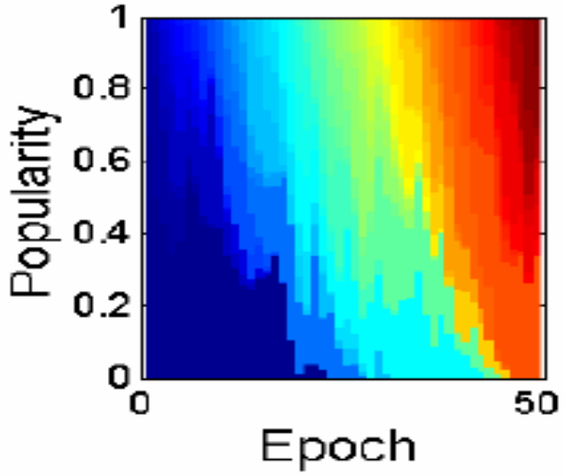


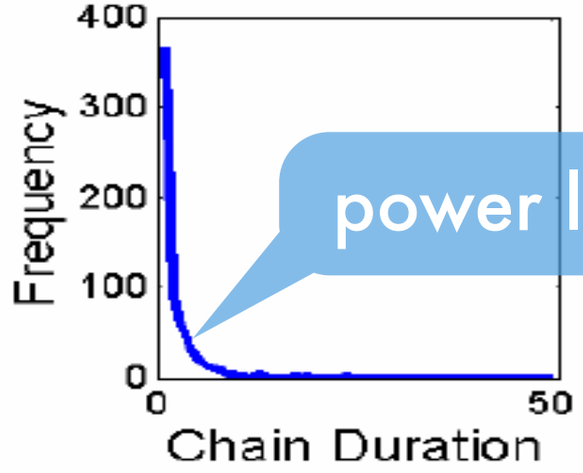


TDPM

$W = 4$
$\lambda = .4$

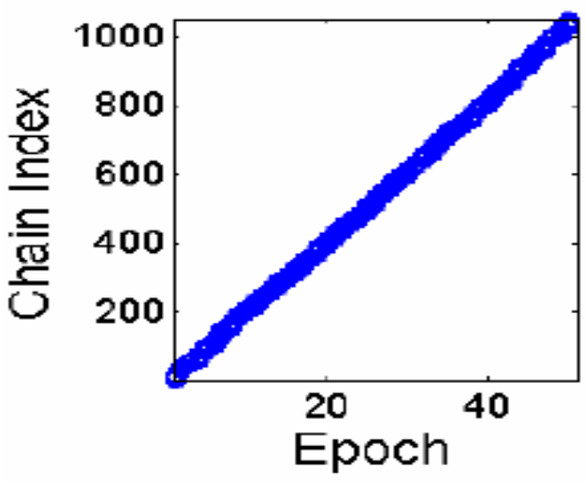


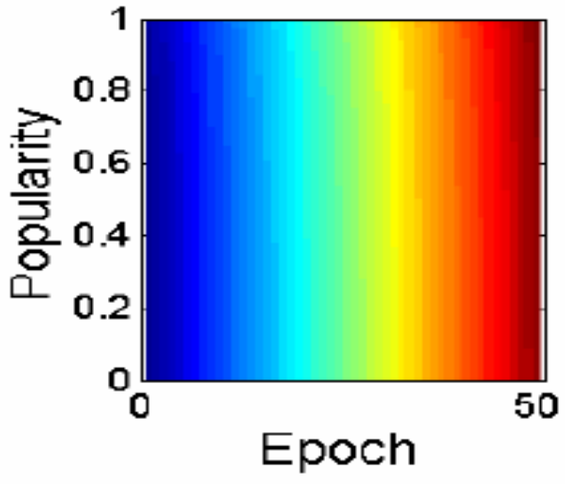


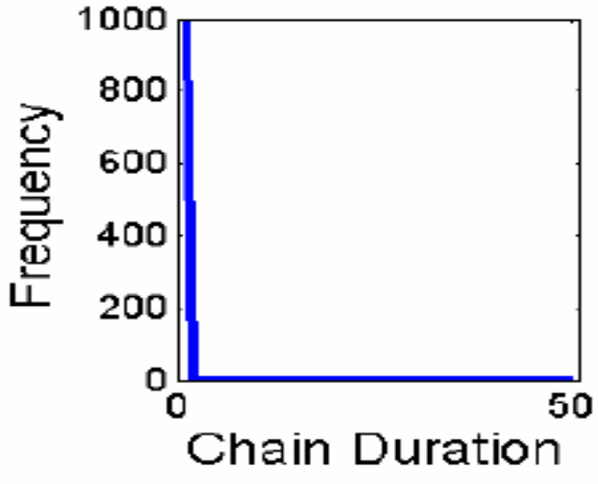


Independent DPMs

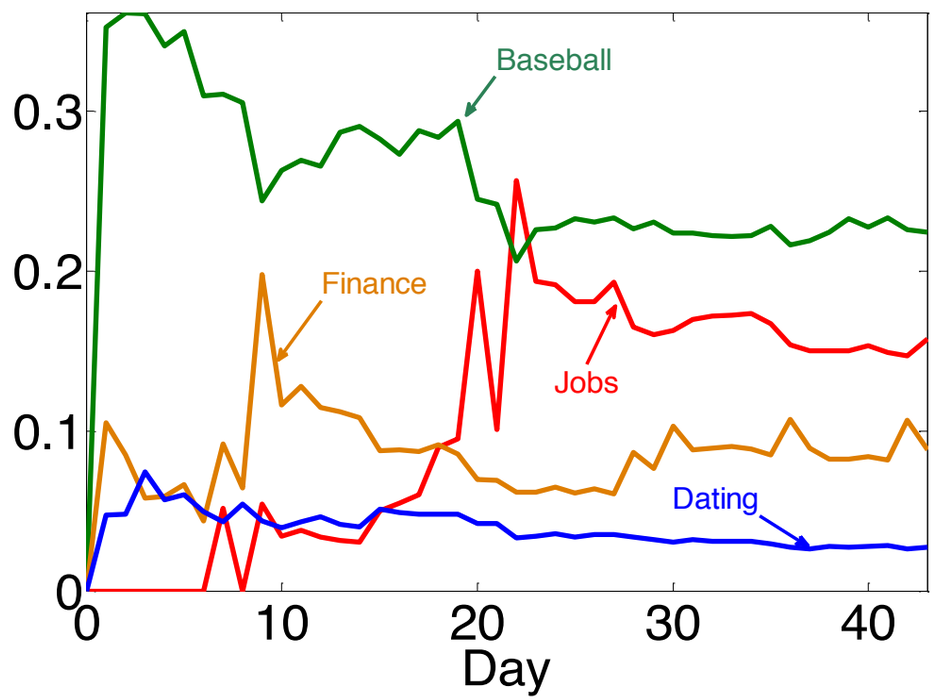
$W = 0$
$\lambda = ?$ (any)







User modeling



Buying a camera

Buying a camera

Sponsor Results

YAHOO!

Web Images Video Local Shopping News More

panasonic lx5

Search In: the Web pages in English, French, German, Italian and Spanish

Also try: panasonic lz5, more...

Sponsored Results

Panasonic LX5 Cheap
Best Value for **Panasonic LX5**. Find NexTag Sellers
www.**NexTag.com**

Panasonic Lumix DMC-**LX5** Review (white
$434.00 as of Oct 17, 2010 Despite its shortcomings the **Panasonic** Lumix DMC- **LX5** delivers an exceller fastest in its class ...
reviews.cnet.com/digital-cameras/**panasonic**-lumix-
this site

Panasonic LX5 | Get The Lowest Price On
Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
www.**panasoniclx5.com** - Cached - More from this s

Panasonic Lumix DMC-**LX5** White Digital (
shopping.yahoo.com
The Panasonic Lumix DMC-LX5 is a compact digital photo enthusiasts the ideal way for capturing profess photos and High De...

Price: **$434 to $513.99**

Reviews | Price & Details | Specs

amazon.com Prime

Hello, Alexander Smola. We have recommendations for you. (Not Alexander?)

Alexander's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop All Departments | Search Electronics | GO

Camera & Photo | All Electronics | Brands | Bestsellers | Digital SLRs & Lenses | Point-And-Shoots | Camcorders

Instant Order Update for Alexander Smola. You purchased this item on October 6, 2010. Vi
Color: **Black**

Prime
Member: Alexander Smola

Alexander Smola: This item is eligible for Amazon Prime. Click here to turn on 1-Click and make Prime even better for you. (With 1-Click enabled, you can always use the regular shopping cart as well.)

Panasonic Lumix DMC-LX5 10.1 MP Digital Camera with 3.8x Optical Image Stabilized Zoom and 3.0-Inch LCD (Black)
by Panasonic
(40 customer reviews)

List Price: $499.00
Price: $444.95 & eligible for free shipping with **Amazon Prime**
You Save: $54.05 (11%)

new

time

Buying a camera

Sponsor Results

YAHOO!

Web Images Video Local Shopping News More

panasonic lx5

Search In: the Web pages in English, French, German, Italian and Spanish

Also try: panasonic lz5, more...

Panasonic LX5 Cheap
Sponsored Results
Best Value for **Panasonic LX5**. Find NexTag Sellers
www.NexTag.com

Panasonic Lumix DMC-LX5 Review (white
$434.00 as of Oct 17, 2010 Despite its shortcomings the **Panasonic** Lumix DMC- **LX5** delivers an exceller fastest in its class ...
reviews.cnet.com/digital-cameras/**panasonic**-lumix-
this site

Panasonic LX5 | Get The Lowest Price On
Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
www.panasoniclx5.com - Cached - More from this s

Panasonic Lumix DMC-LX5 White Digital (
shopping.yahoo.com

show ads now

amazon.com Prime

Hello, Alexander Smola. We have recommendations for you. (Not Alexander?)
Alexander's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop All Departments | Search Electronics | GO

Camera & Photo | All Electronics | Brands | Bestsellers | Digital SLRs & Lenses | Point-And-Shoots | Camcorders

Instant Order Update for Alexander Smola. You purchased this item on October 6, 2010. Vi
Color: **Black**

Prime
Member: Alexander Smola

Alexander Smola: This item is eligible for Amazon Prime. Click here to turn on 1-Click and make Prime even better for you. (With 1-Click enabled, you can always use the regular shopping cart as well.)

Panasonic Lumix DMC-LX5 10.1 MP Digital Camera with 3.8x Optical Image Stabilized Zoom and 3.0-Inch LCD (Black)
by Panasonic
(40 customer reviews)

List Price: ~~$499.00~~
Price: $444.95 & eligible for free shipping with **Amazon Prime**
You Save: $54.05 (11%)

new

time

Buying a camera

Sponsor Results

YAHOO!

Web Images Video Local Shopping News More

panasonic lx5

Search In: the Web pages in English, French, German, Italian and Spanish

Also try: panasonic lz5, more...

Panasonic LX5 Cheap
Sponsored Results
Best Value for **Panasonic LX5**. Find NexTag Sellers
www.NexTag.com

Panasonic Lumix DMC-LX5 Review (white
$434.00 as of Oct 17, 2010 Despite its shortcomings the **Panasonic** Lumix DMC- **LX5** delivers an exceller fastest in its class ...
reviews.cnet.com/digital-cameras/**panasonic**-lumix-
this site

Panasonic LX5 | Get The Lowest Price On
Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
www.panasoniclx5.com - Cached - More from this s

Panasonic Lumix DMC-LX5 White Digital (
shopping.yahoo.com

amazon.com Prime

Hello, Alexander Smola. We have recommendations for you. (Not Alexander?)
Alexander's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop All Departments | Search Electronics | GO

Camera & Photo | All Electronics | Brands | Bestsellers | Digital SLRs & Lenses | Point-And-Shoots | Camcorders

Instant Order Update for Alexander Smola. You purchased this item on October 6, 2010. Vi
Color: **Black**

Prime
Member: Alexander Smola

Alexander Smola: This item is eligible for Amazon Prime. Click here to turn on 1-Click and make Prime even better for you. (With 1-Click enabled, you can always use the regular shopping cart as well.)

Panasonic Lumix DMC-LX5 10.1 MP Digital Camera with 3.8x Optical Image Stabilized Zoom and 3.0-Inch LCD (Black)
by Panasonic
(40 customer reviews)

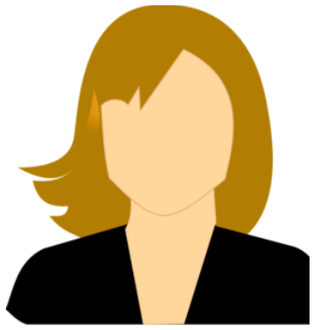
List Price: $499.00
Price: $444.95 & eligib
Amazon Prime
You Save: $54.05 (11%)

new

show ads now

too late

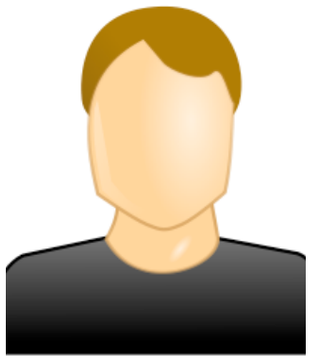
time



Car

Deals

van

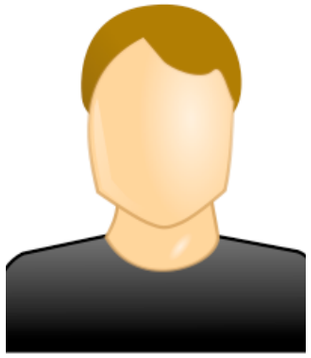


job

Hiring

diet

Car	Auto	Movies
Deals	Price	Theatre
van	Used	Art
	inspection	gallery



job	Hiring	Diet
Hiring	Salary	Calories
diet	Diet	Recipe
	calories	chocolate

	Flight	School
	London	Supplies
	Hotel	Loan
	weather	college

CARS

Car

Deals

van

Auto

Price

Used

inspection

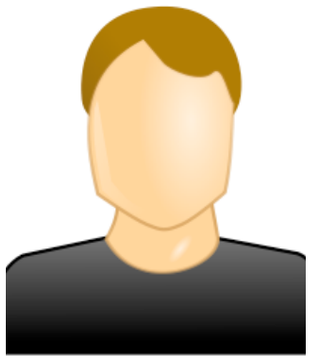
Art

Movies

Theatre

Art

gallery



Jobs

job

Hiring

diet

Hiring

Salary

Diet

calories

Diet

Diet

Calories

Recipe

chocolate

Travel

Flight

London

Hotel

weather

College

School

Supplies

Loan

college

finance

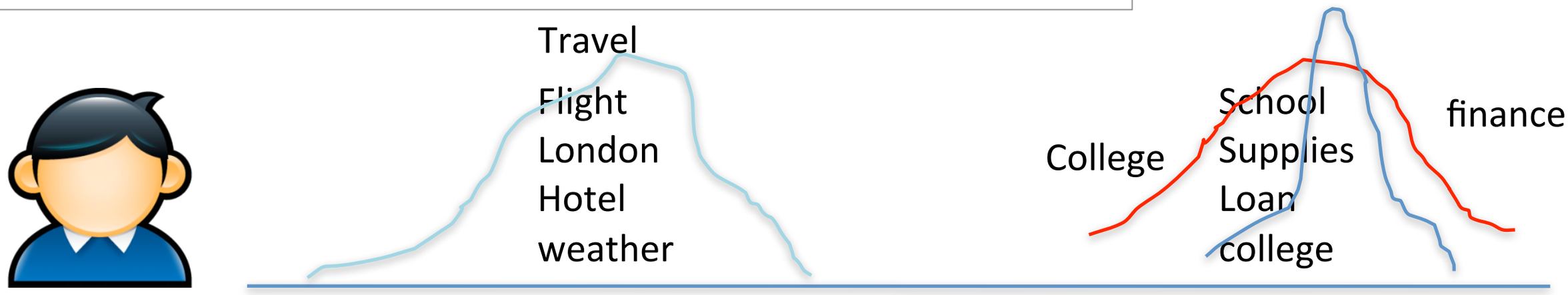
User modeling

Input

- Queries issued by the user or Tags of watched content
- Snippet of page examined by user
- Time stamp of each action (day resolution)

Output

- Users' daily distribution over intents
- Dynamic intent representation

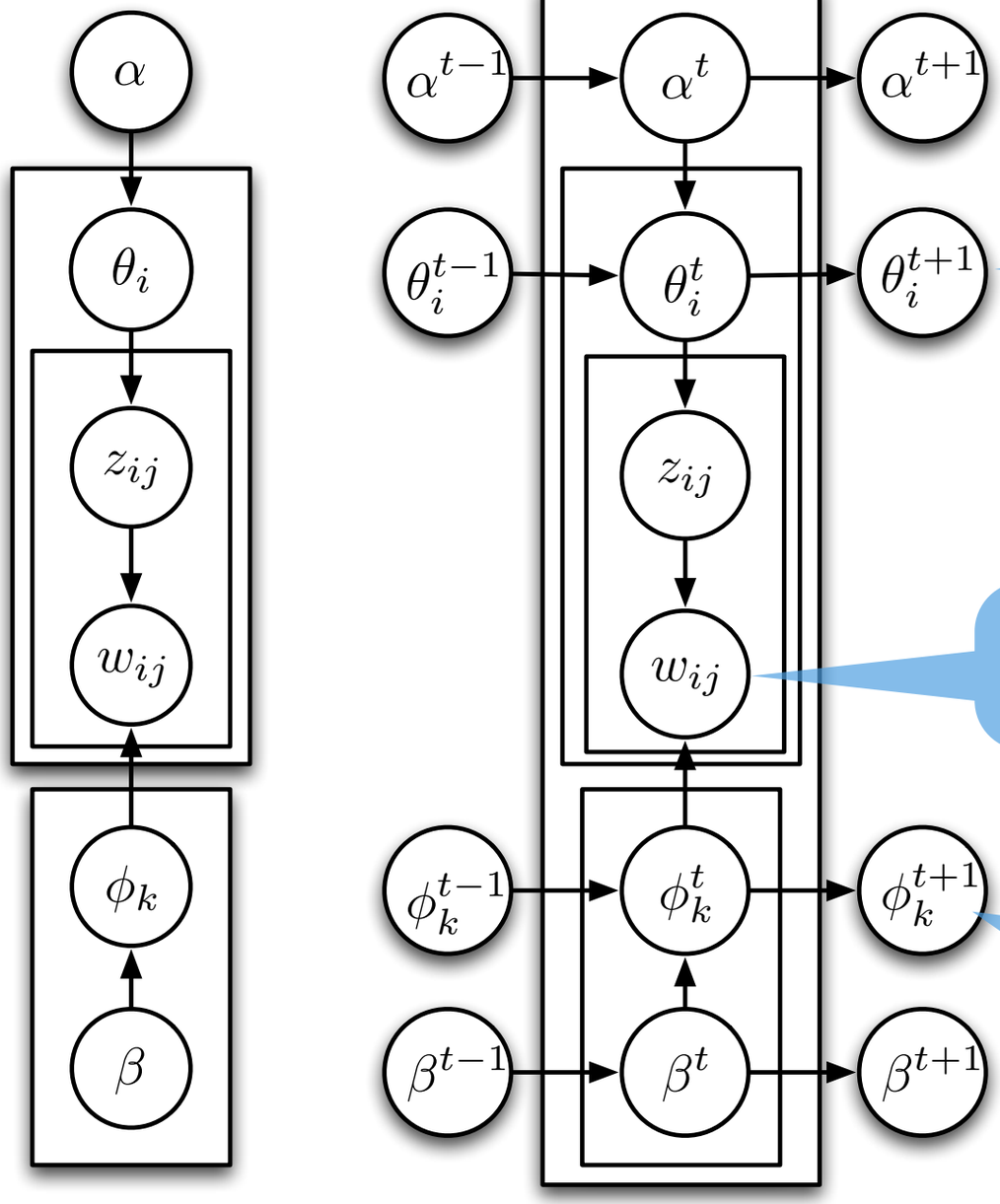


Time dependent models

- LDA for topical model of users where
  - User interest distribution changes over time
  - Topics change over time
- This is like a Kalman filter except that
  - Don’t know what to track (a priori)
  - Can’t afford a Rauch-Tung-Striebel smoother
  - Much more messy than plain LDA

Graphical Model

plain LDA



time dependent user interest

user actions

actions per topic

Prior for user actions at time *t*

Food
Chicken
pizza

recipe
job
hiring

Part-time
Opening
salary

Kelly
recipe
cuisine

food
chicken
Pizza
millage

t

t+1

Time

Diet

Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Cars

Car
Blue
Book
Kelley
Prices
Small
Speed
large

Job

job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Finance

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

All

Long-term

Prior for user actions at time *t*

Food
Chicken
pizza

recipe
job
hiring

Part-time
Opening
salary

Kelly
recipe
cuisine

food
chicken
Pizza
millage

t

t+1

Time

Diet

Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Cars

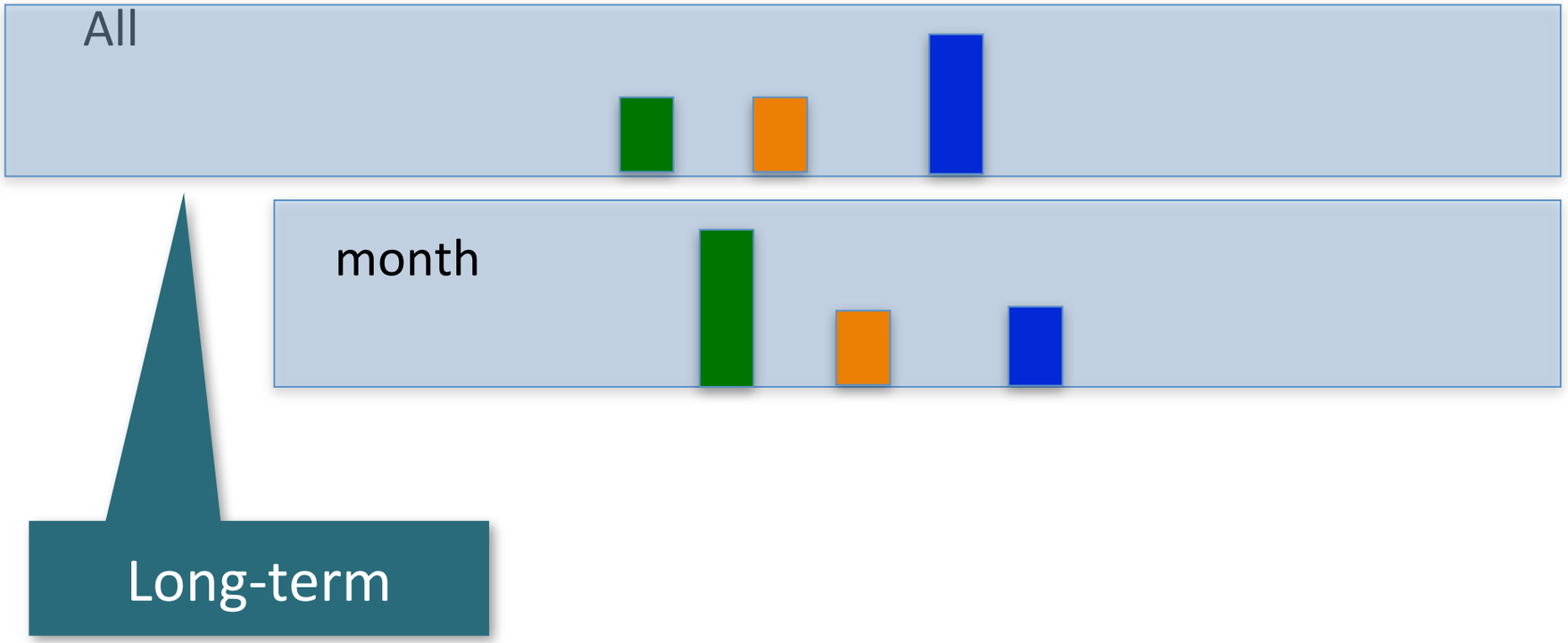
Car
Blue
Book
Kelley
Prices
Small
Speed
large

Job

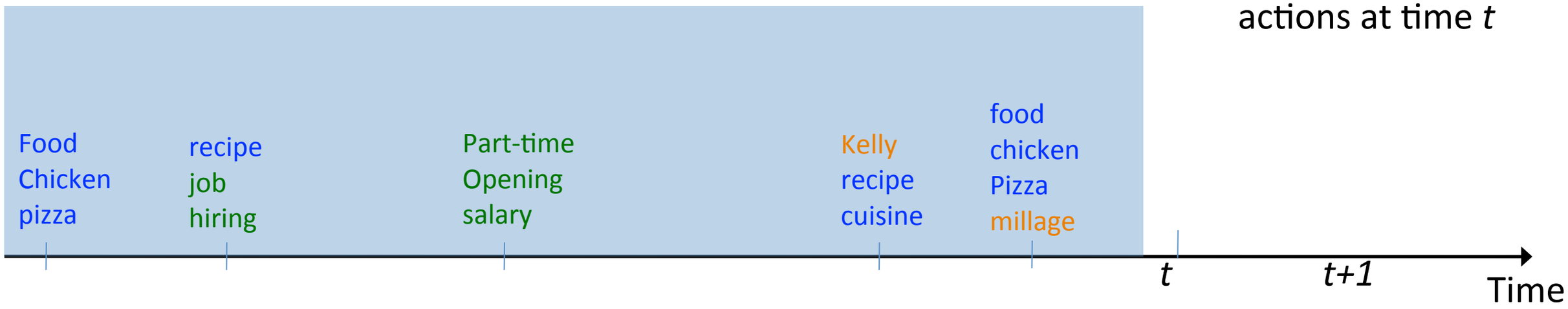
job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Finance

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



Prior for user actions at time *t*



Diet

Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Cars

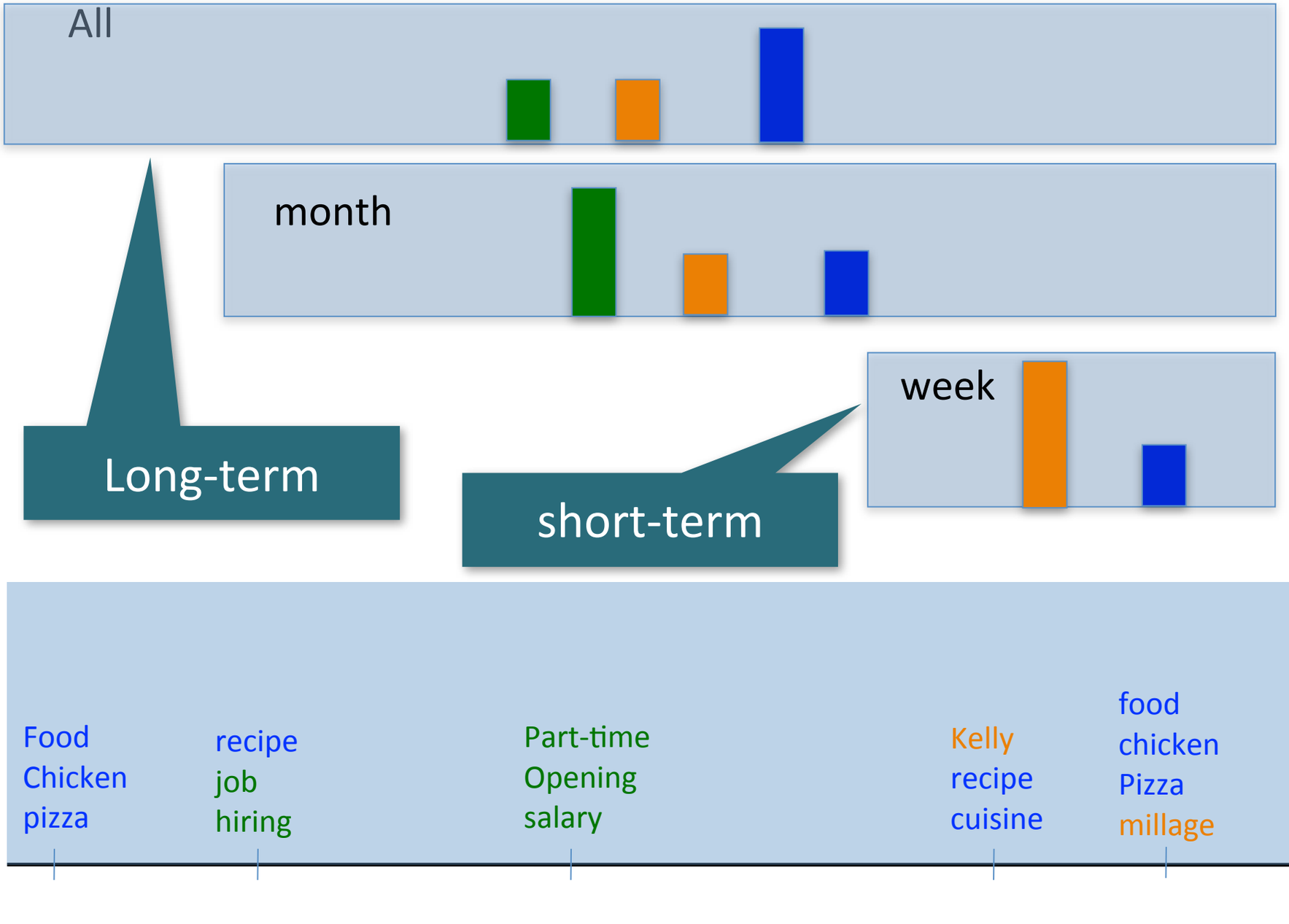
Car
Blue
Book
Kelley
Prices
Small
Speed
large

Job

job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Finance

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



Prior for user actions at time *t*

Diet

Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Cars

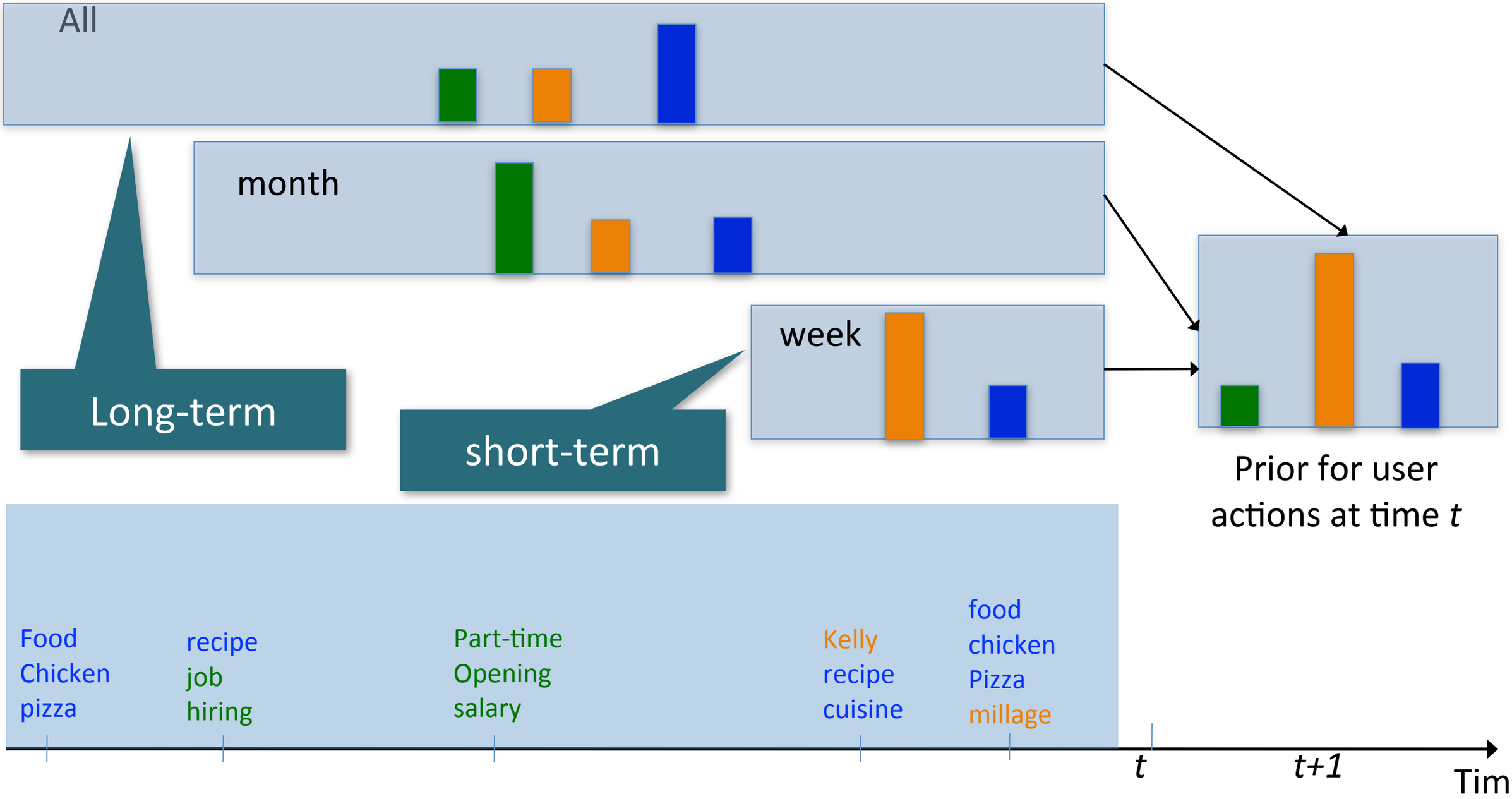
Car
Blue
Book
Kelley
Prices
Small
Speed
large

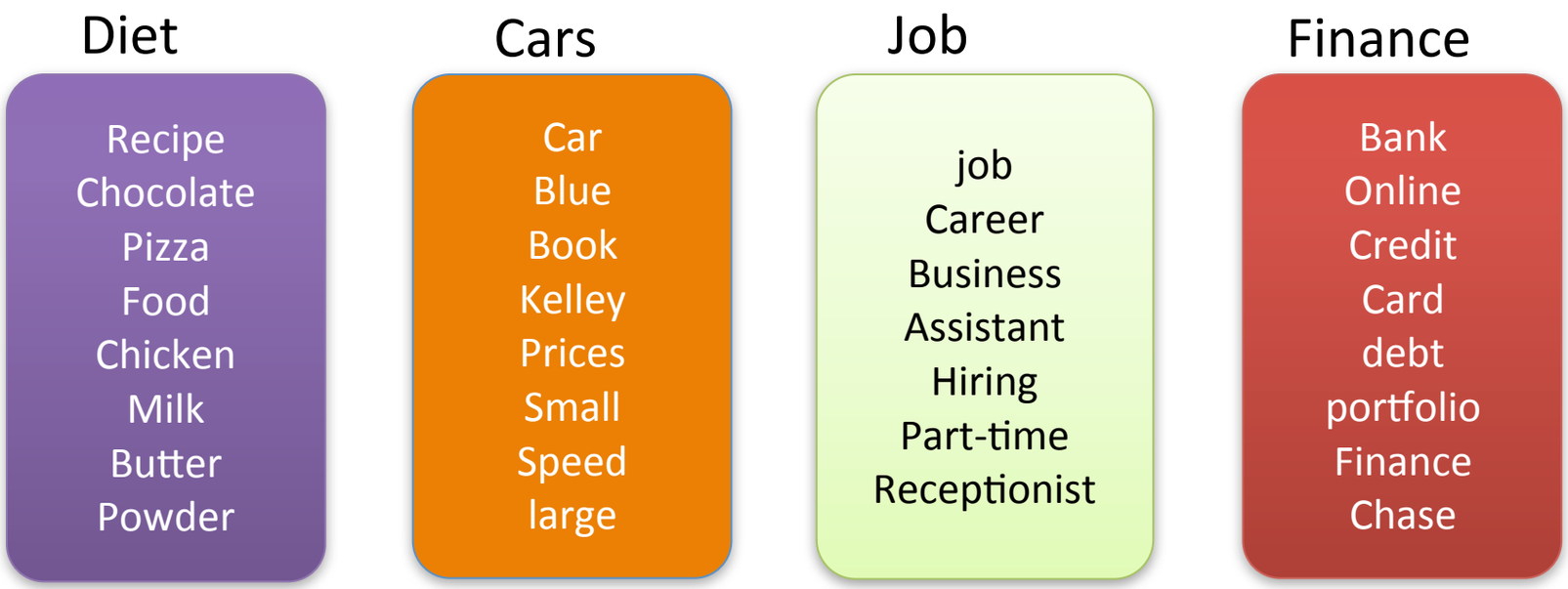
Job

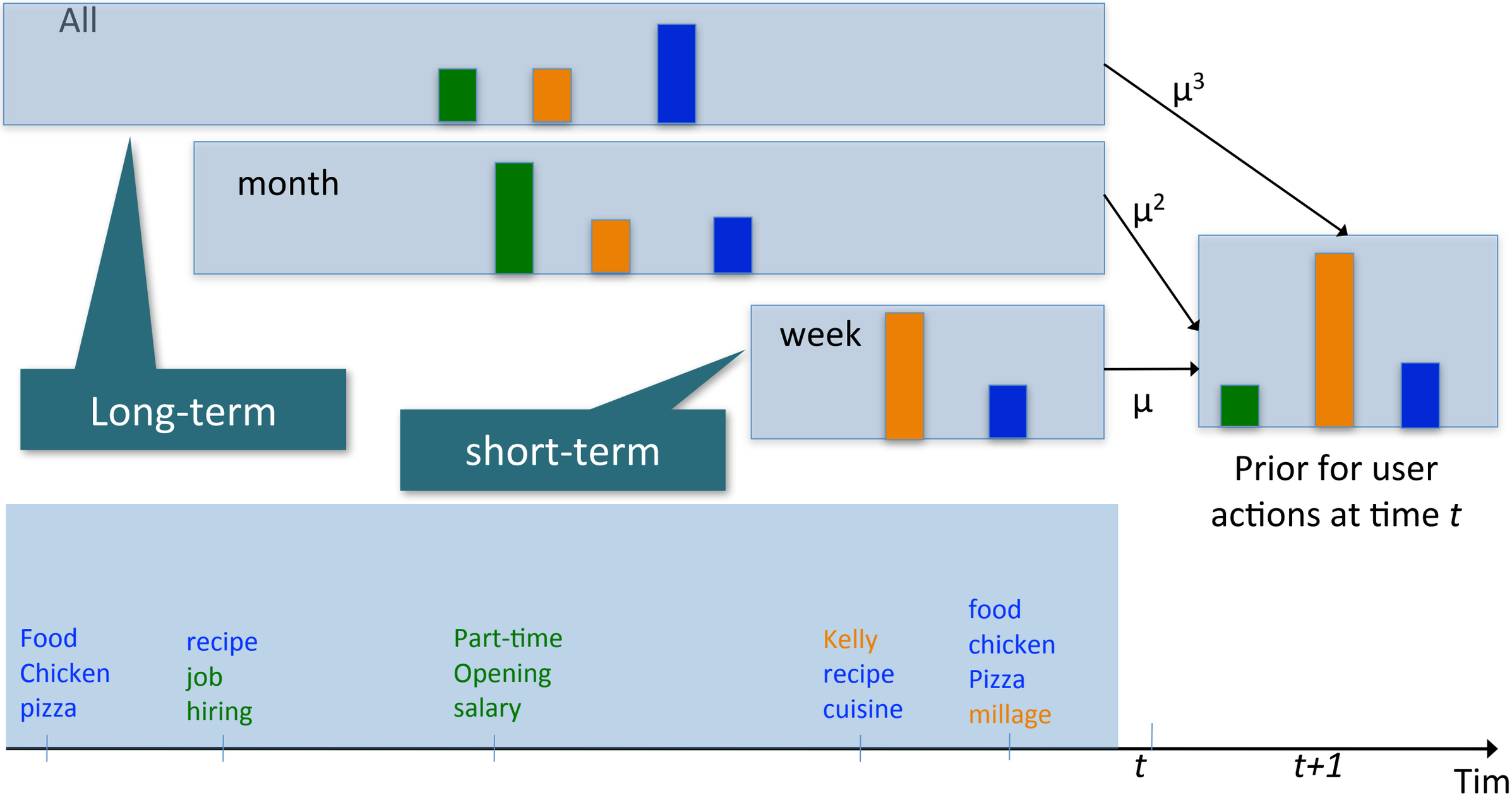
job
Career
Business
Assistant
Hiring
Part-time
Receptionist

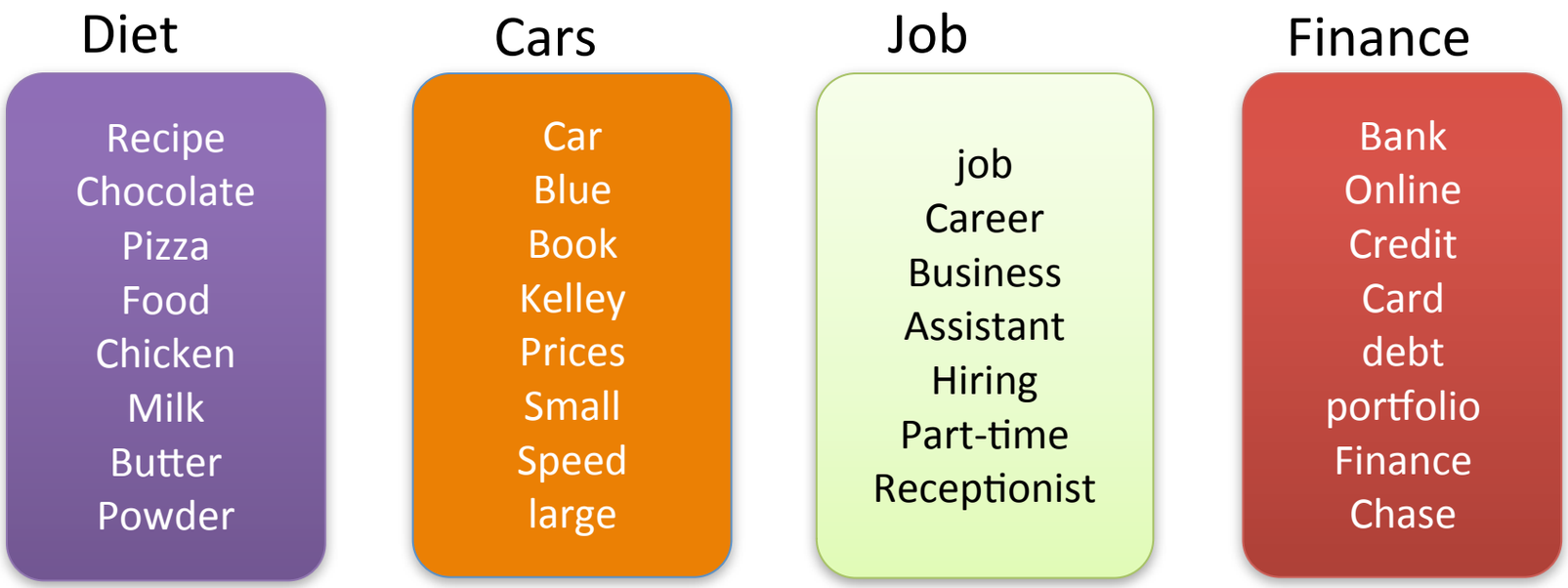
Finance

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

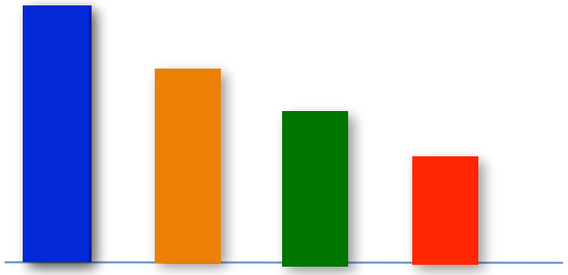

All
month
week
Long-term
short-term
Prior for user
actions at time t
Food
Chicken
pizza
recipe
job
hiring
Part-time
Opening
salary
Kelly
recipe
cuisine
food
chicken
Pizza
millage
t
t+1
Time


Diet
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder
Cars
Car
Blue
Book
Kelley
Prices
Small
Speed
large
Job
job
Career
Business
Assistant
Hiring
Part-time
Receptionist
Finance
Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

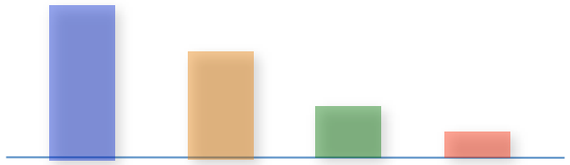

All
month
week
μ^3
μ^2
μ
Long-term
short-term
Prior for user actions at time *t*
Food
Chicken
pizza
recipe
job
hiring
Part-time
Opening
salary
Kelly
recipe
cuisine
food
chicken
Pizza
millage
t
t+1
Time


Diet
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder
Cars
Car
Blue
Book
Kelley
Prices
Small
Speed
large
Job
job
Career
Business
Assistant
Hiring
Part-time
Receptionist
Finance
Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

At time t



At time t+1

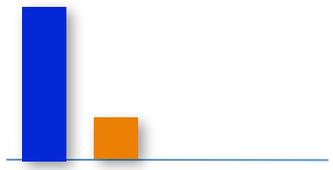


Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptioni
st

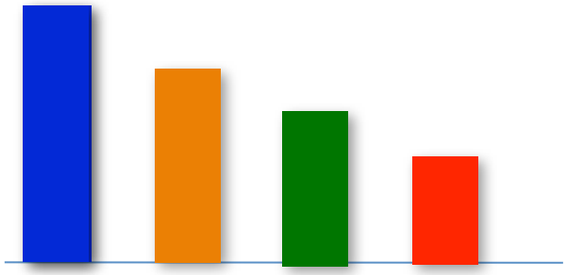
Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



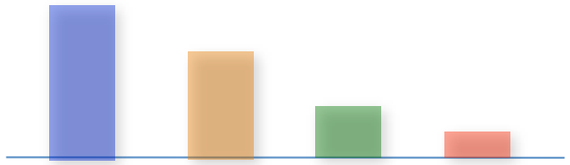
Food Chicken
Pizza mileage

Car speed offer
Camry accord career

At time t



At time t+1



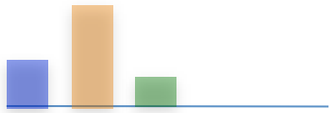
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

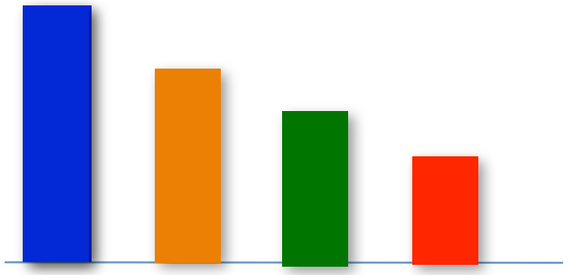
Food Chicken
Pizza mileage

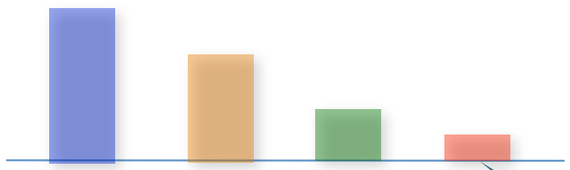


Car speed offer
Camry accord career

At time t

At time t+1





Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

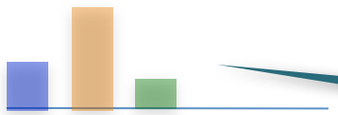
Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

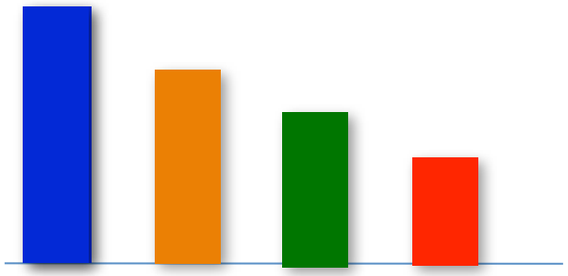
priors

Food Chicken
Pizza mileage

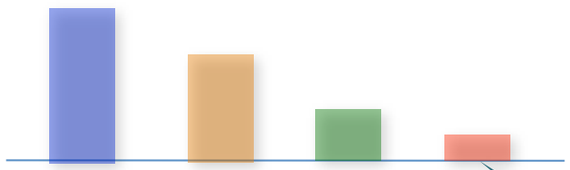


Car speed offer
Camry accord career

At time t



At time t+1



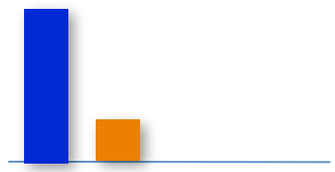
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

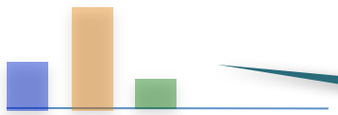
job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

priors



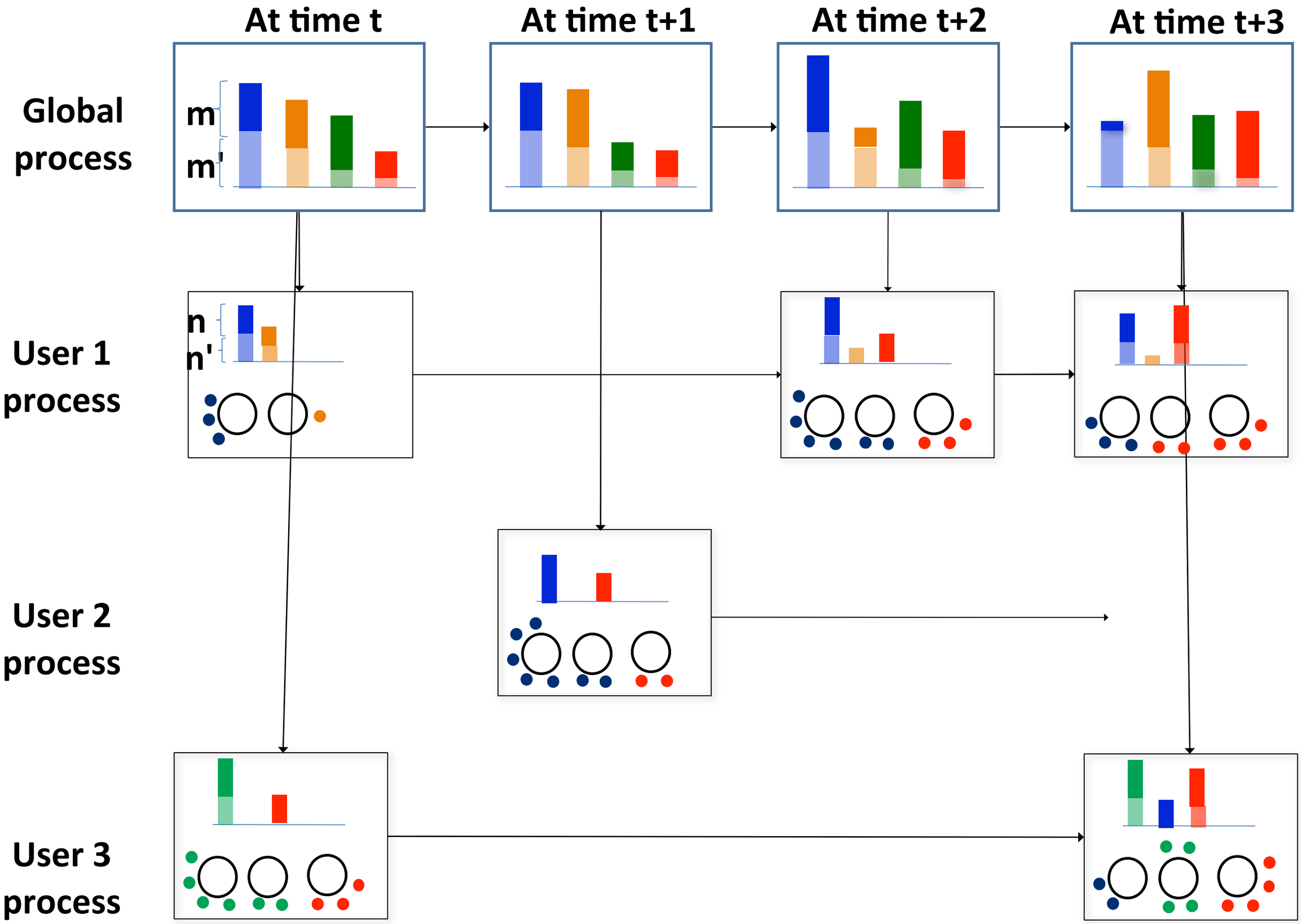
Food Chicken
Pizza mileage



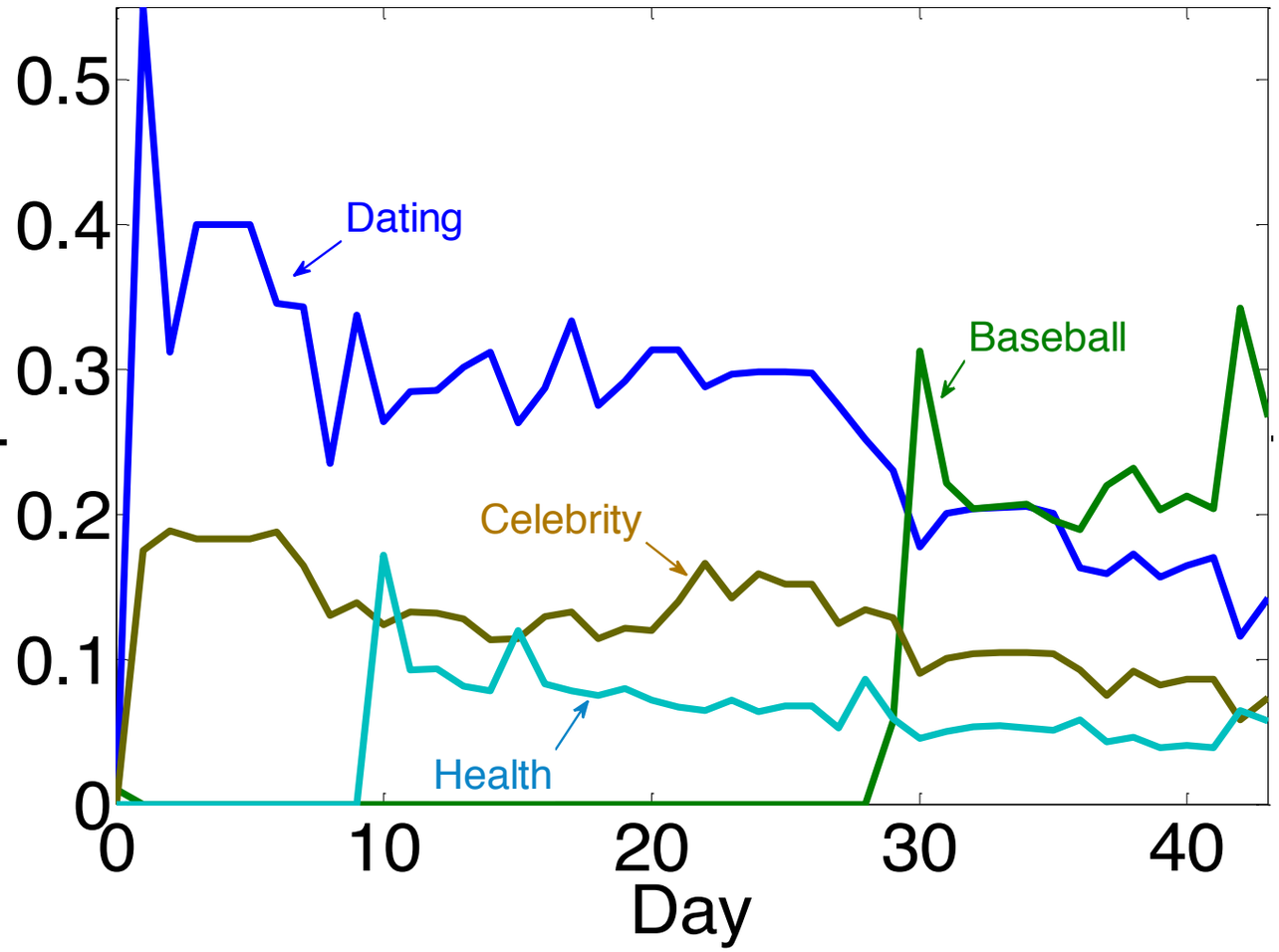
Car speed offer
Camry accord career

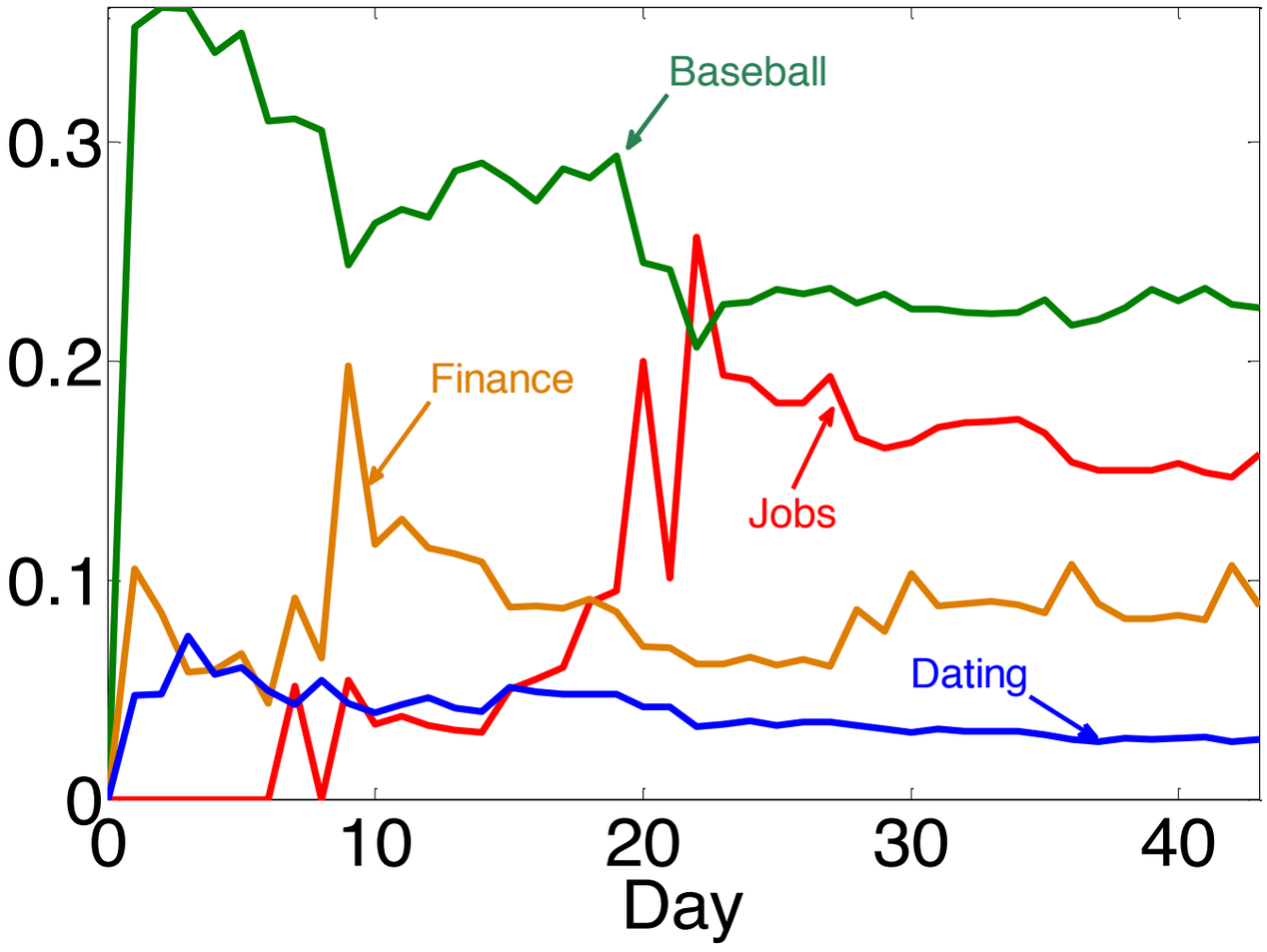
Generative Process

- For each user interaction
  - Choose an intent from local distribution
    - Sample word from the topic's word-distribution
  - Choose a new intent $\propto \alpha$
    - Sample a new intent from the global distribution
      - Sample word from the new topic word-distribution


At time t
At time t+1
At time t+2
At time t+3
Global process
m
m'
User 1 process
n
n'
User 2 process
User 3 process

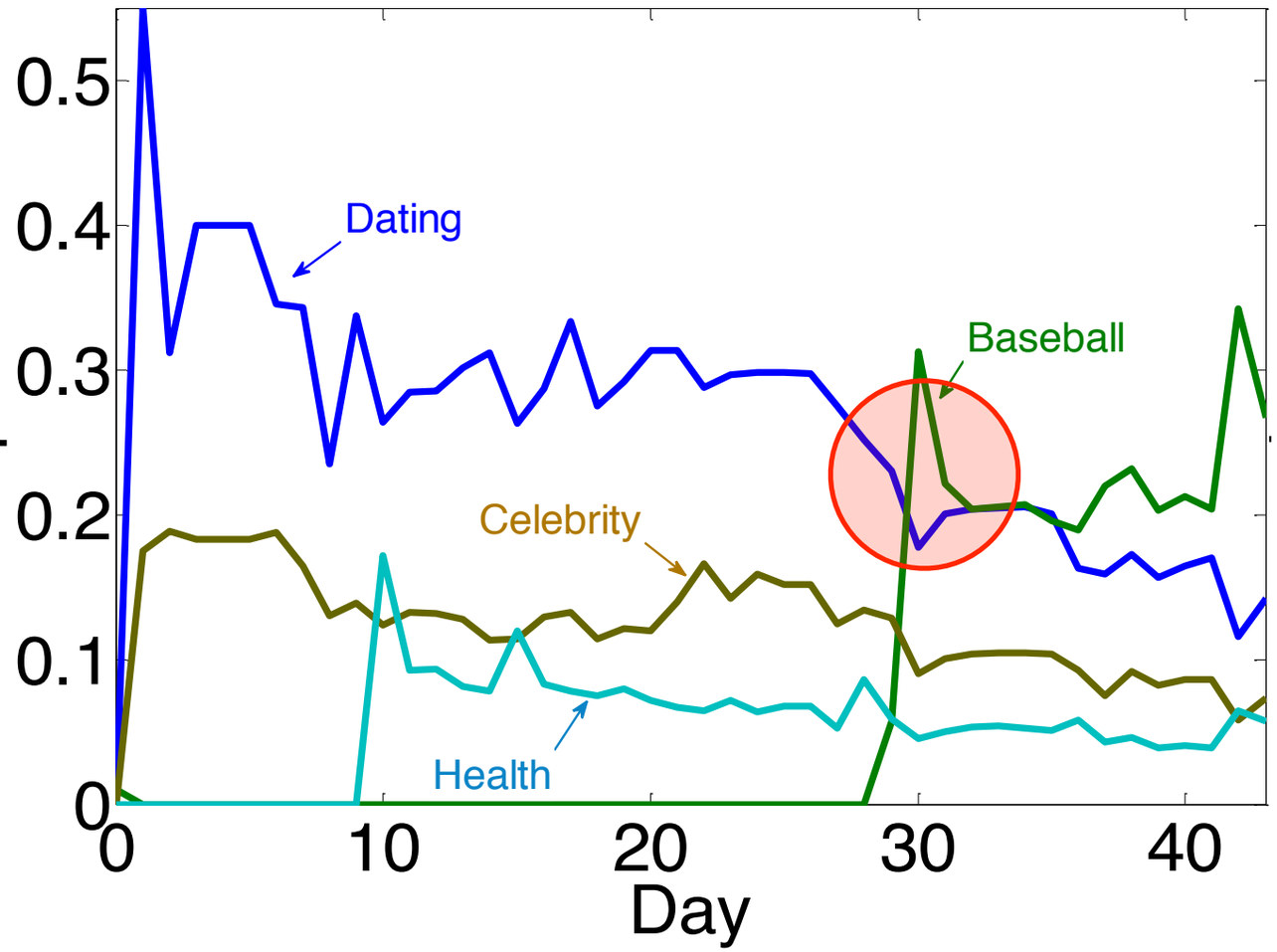
Sample users

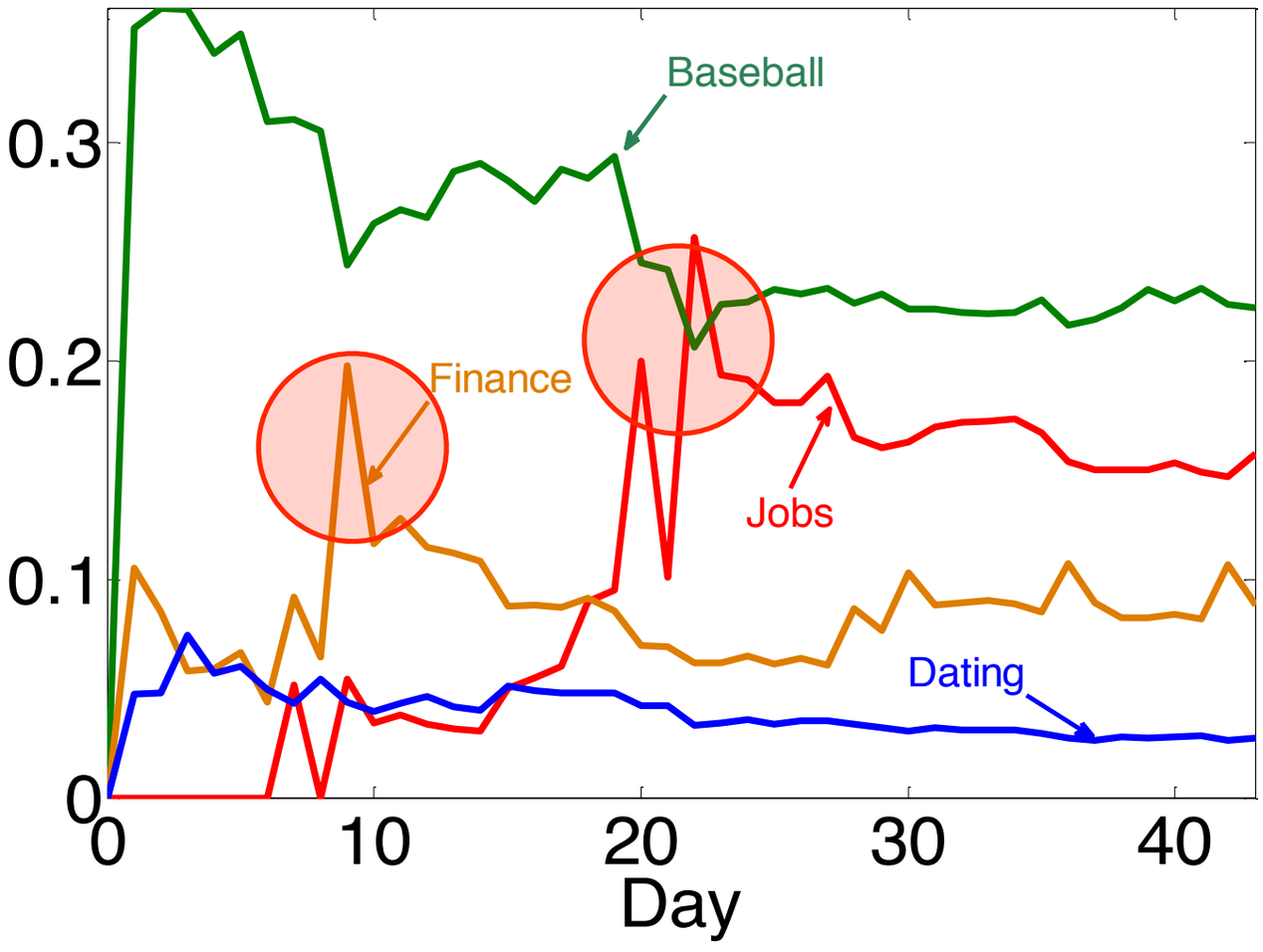




Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time	Trading
match	bullpen	Kudrow	layers	receptionist	currency
	Greinke	Hollywood			

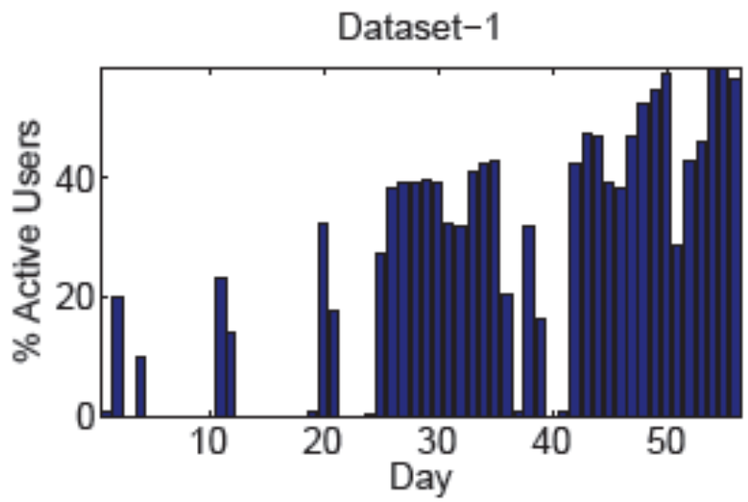
Sample users

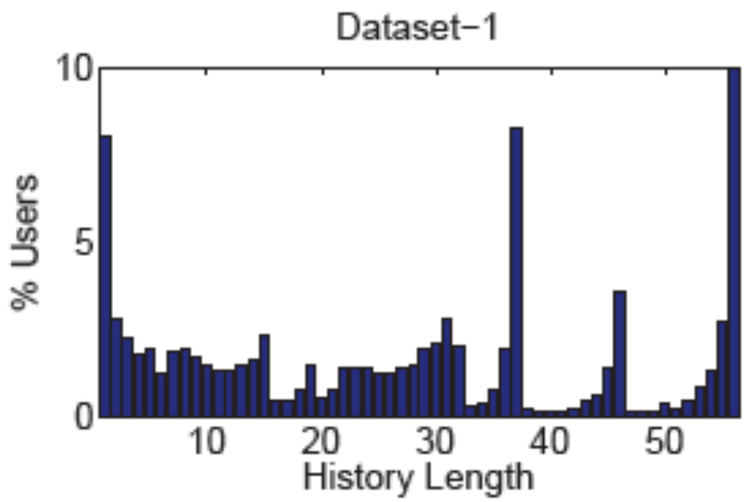


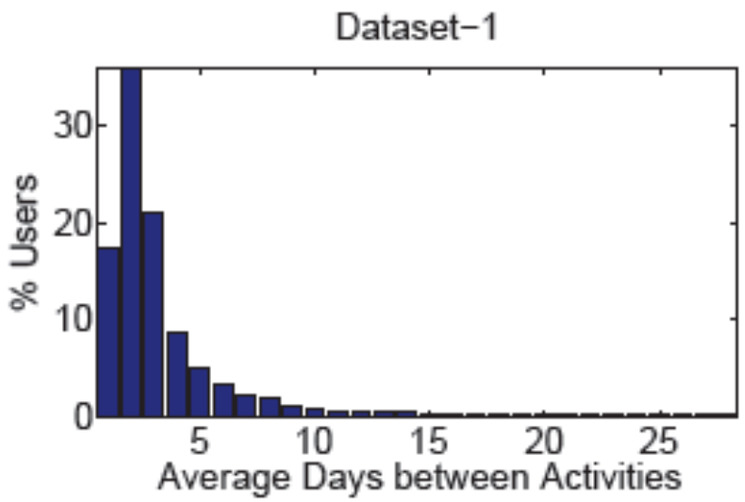


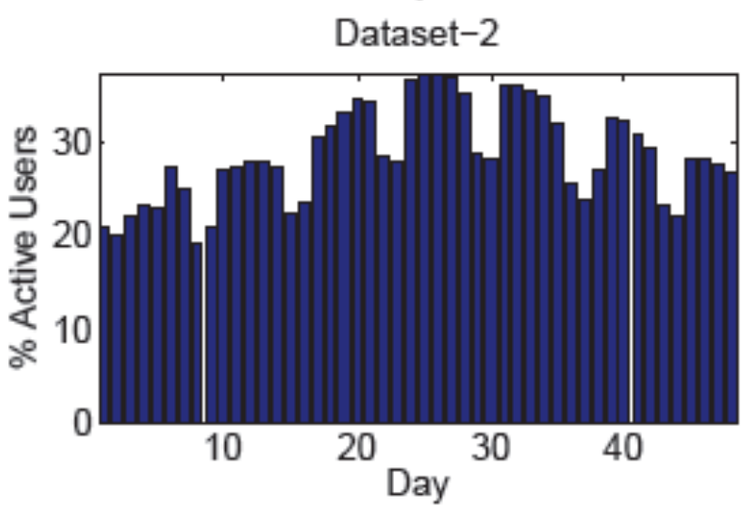
Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time	Trading
match	bullpen	Kudrow	layers	receptionist	currency
	Greinke	Hollywood			

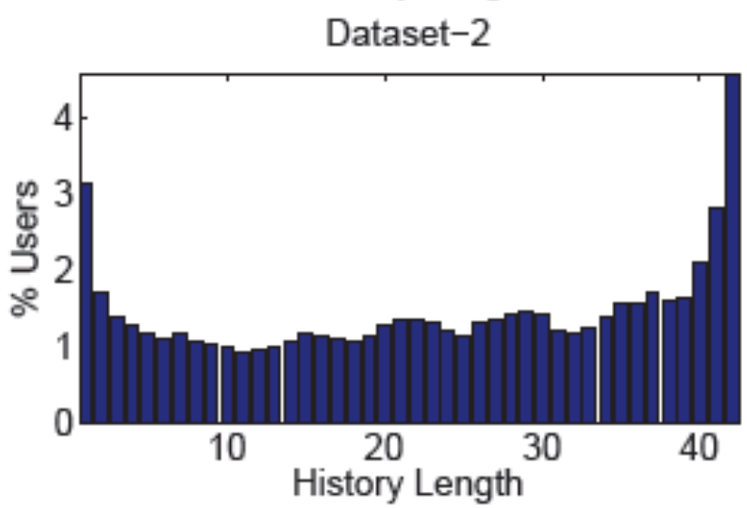
Data

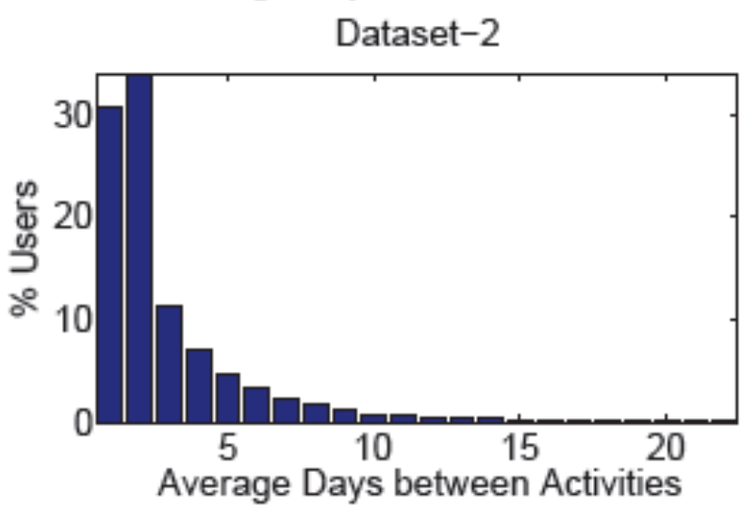






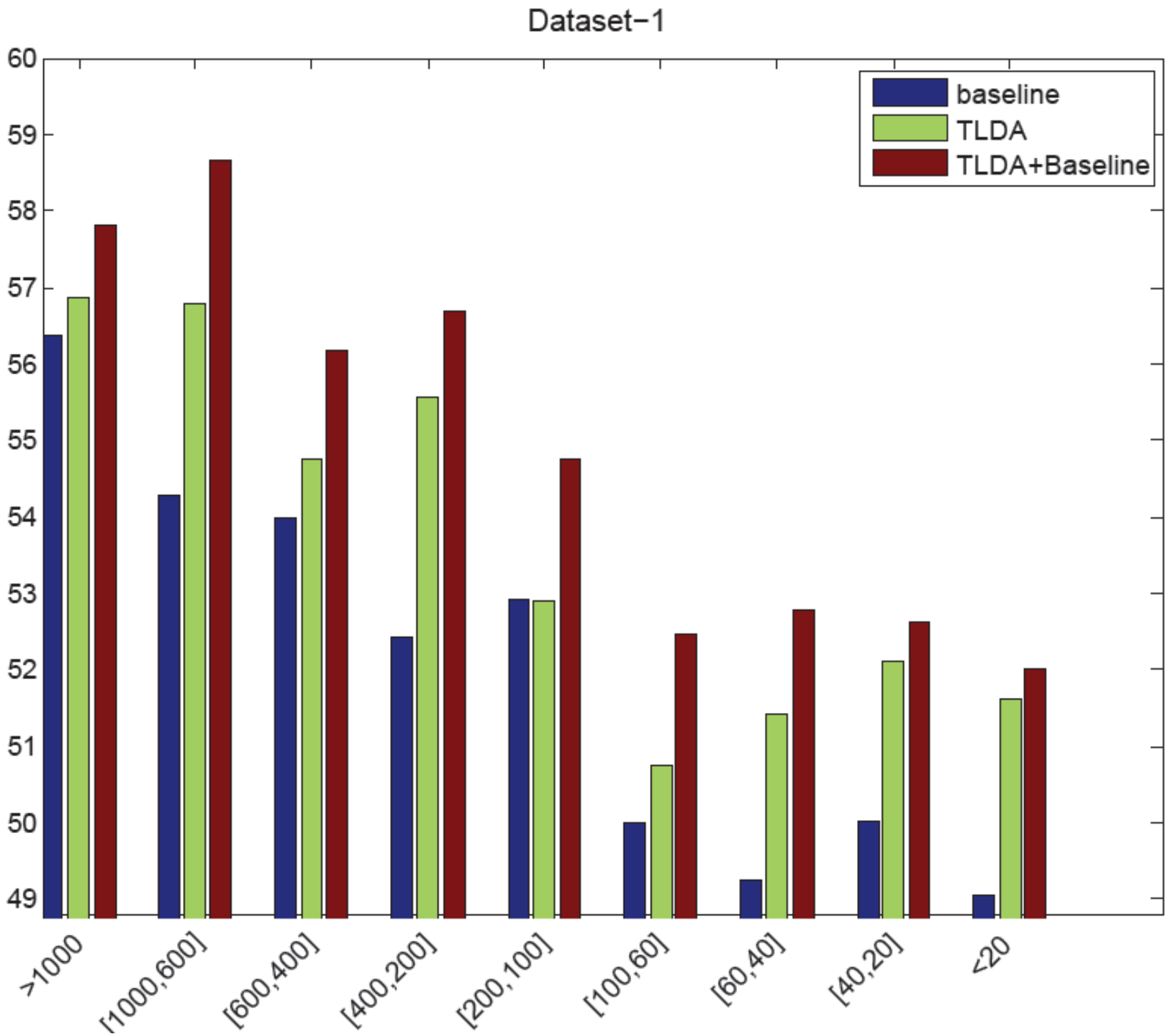




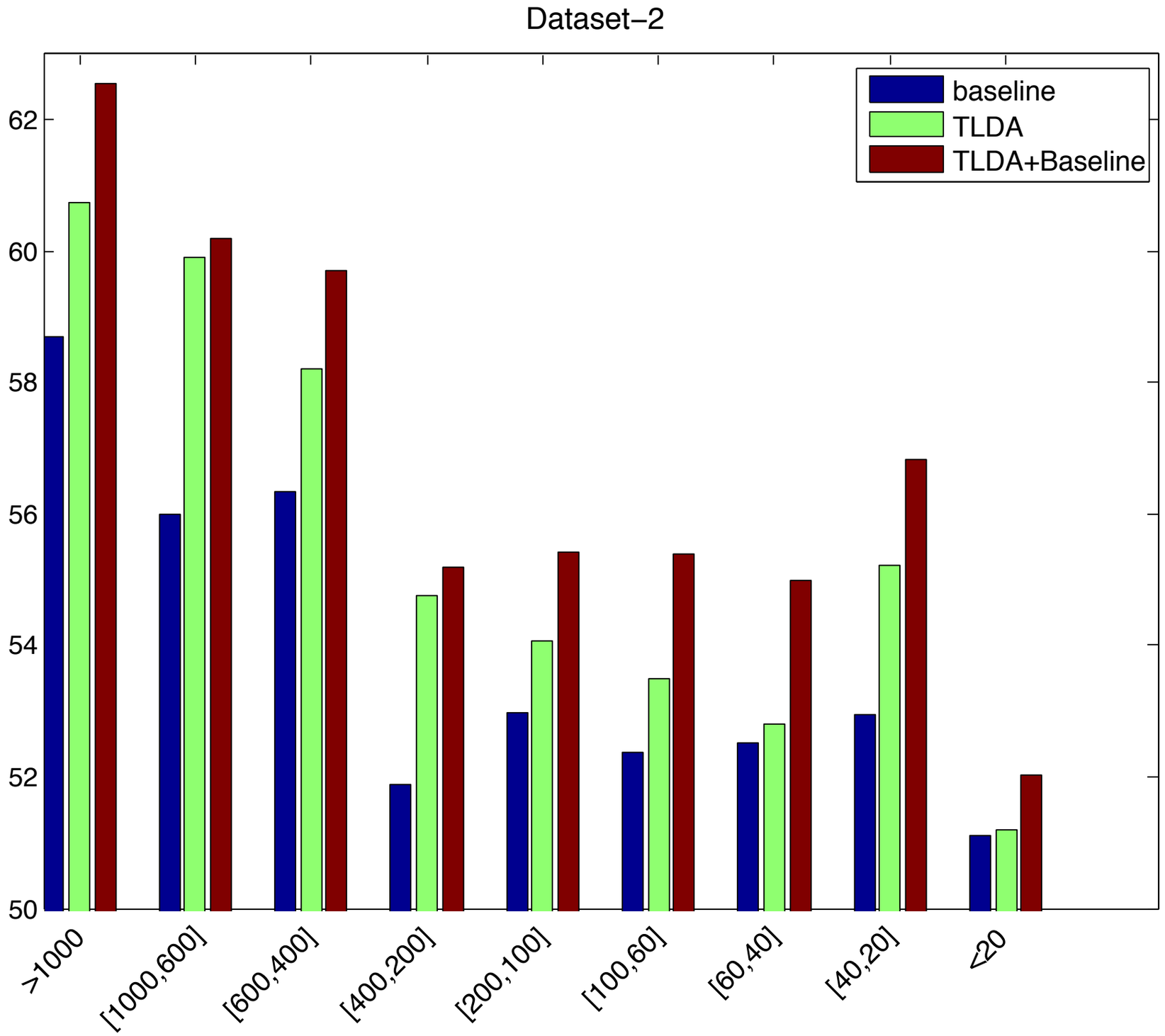


dataset	# days	# users	# campaigns	size
1	56	13.34M	241	242GB
2	44	33.5M	216	435GB

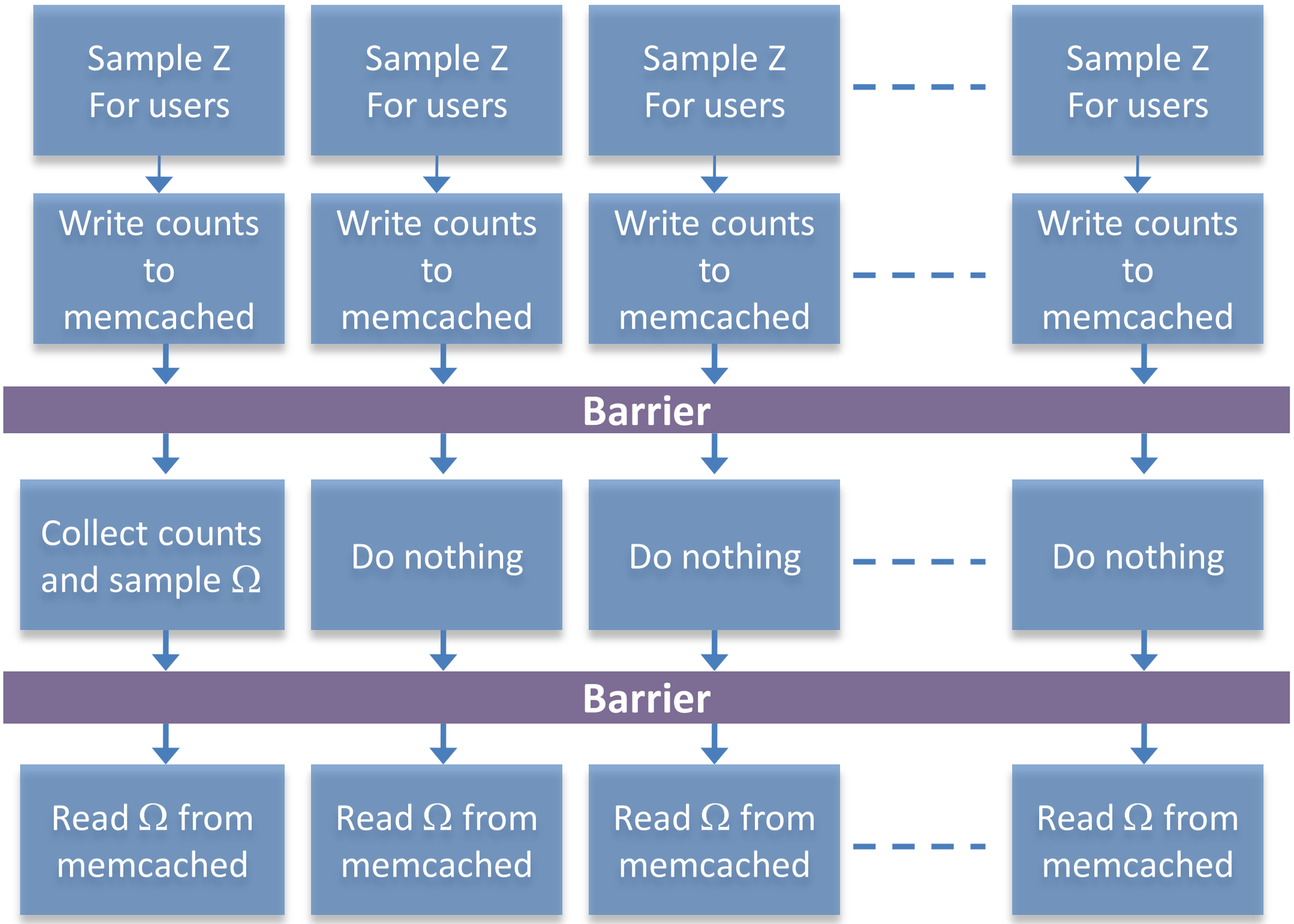
ROC score improvement



ROC score improvement



LDA for user profiling



News

News Stream

News Stream

AP

Republicans is bec
and lawmakers. Bu
Bill Clinton even ba
Full Story »

Video: Gibbs: I Hav
Slideshow: Preside
Related: Tax fight

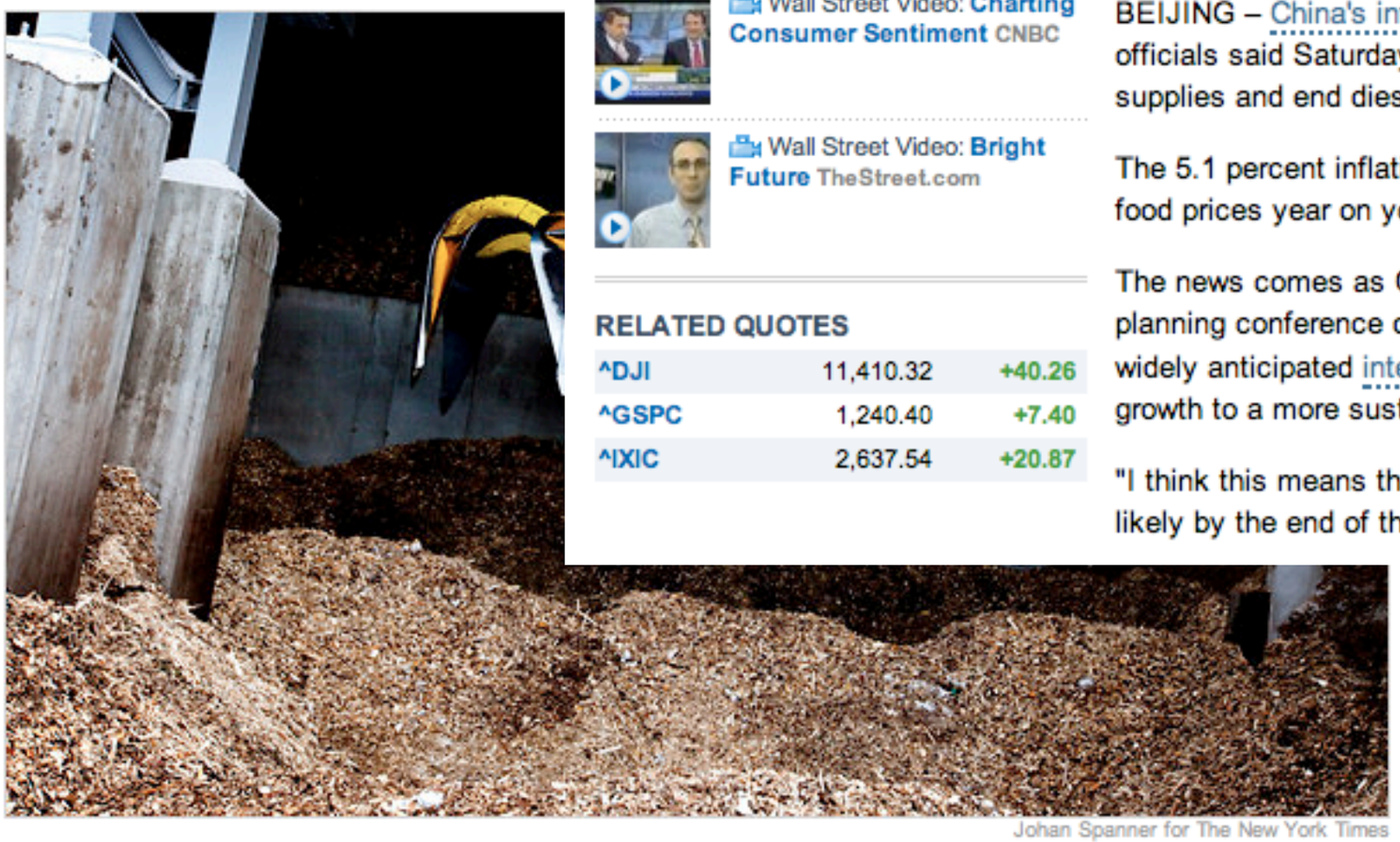
Add-ons turn tax cut bill into 'Christmas t

AP – 1 hr 32 mins ago

WASHINGTON – In the

BEYOND FOSSIL FUELS

Using Waste, Swedish C



Johan Spanner for The New York Times

As part of its citywide system, Kristianstad burns wood waste like tree prunings and scraps from flooring factories to power an underground district heating grid.

China says inflation up 5.1 perce

Associated Press

Buzz up! 19 votes | Share

By CARA ANNA, Associated Press

BEIJING – China's inflation su
officials said Saturday, despit
supplies and end diesel shorta

Wall Street Video: **Charting Consumer Sentiment** CNBC

Wall Street Video: **Bright Future** TheStreet.com

RELATED QUOTES

^DJI	11,410.32	+40.26
^GSPC	1,240.40	+7.40
^IXIC	2,637.54	+20.87

The 5.1 percent inflation rate was driven by a 11.7 percent jump in food prices year on year.

The news comes as China's leaders meet for the top economic planning conference of the year and as financial markets watch for a widely anticipated interest rate hike to help bring rapid economic growth to a more sustainable level.

"I think this means that an interest rate hike of 25 basis points is very likely by the end of the year," said CLSA analyst Andy Rothman.

Suit to Recover Madoff's Money Calls Austrian an Accomplice

By DIANA B. HENRIQUES and PETER LATTMAN

Sonja Kohn, an Austrian banker, is accused of masterminding a 23-year conspiracy that played a central role in financing the gigantic Ponzi scheme.

Post a Comment

Print

ber

ovember,

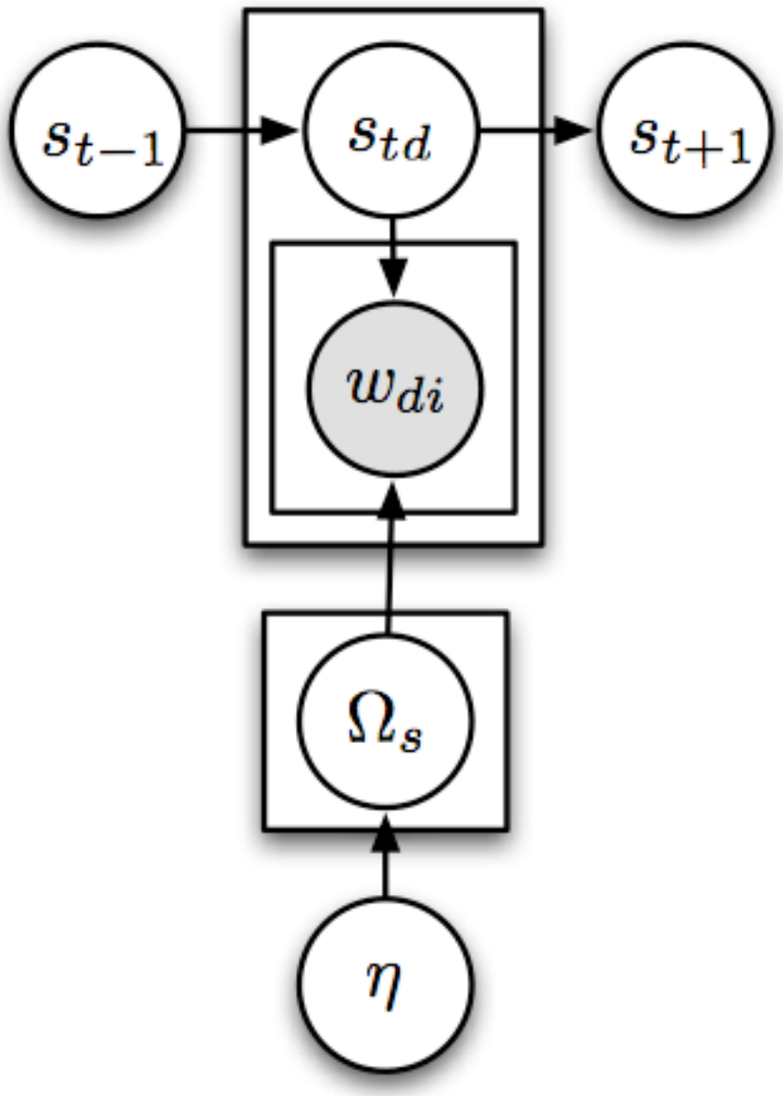
ase food

News Stream

- Over 1 high quality news article per second
- Multiple sources (Reuters, AP, CNN, …)
- Same story from multiple sources
- Stories are related

- Goals
  - Aggregate articles into a storyline
  - Analyze the storyline (topics, entities)

Clustering / RCRP



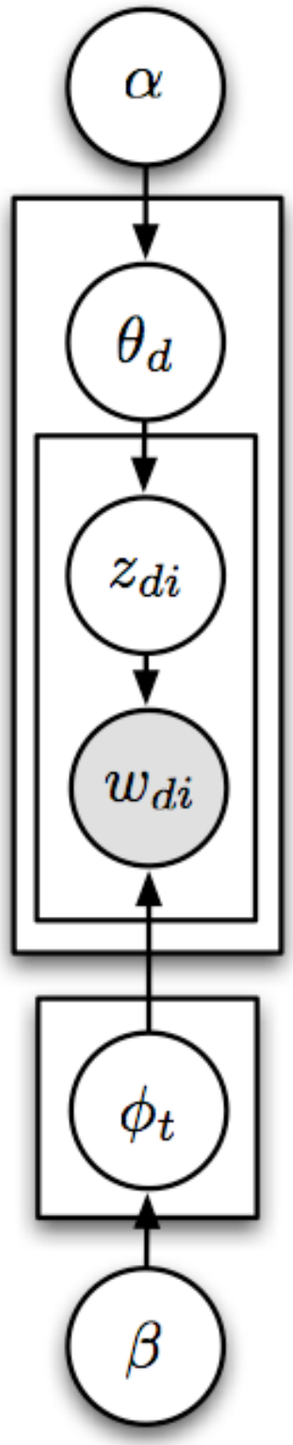
- Assume active story distribution at time t
- Draw story indicator
- Draw words from story distribution
- Down-weight story counts for next day

Ahmed & Xing, 2008

Clustering / RCRP

- Pro
  - Nonparametric model of story generation (no need to model frequency of stories)
  - No fixed number of stories
  - Efficient inference via collapsed sampler
- Con
  - We learn nothing!
  - No content analysis

Latent Dirichlet Allocation



- Generate topic distribution per article
- Draw topics per word from topic distribution
- Draw words from topic specific word distribution

Blei, Ng, Jordan, 2003

Latent Dirichlet Allocation

- Pro
  - Topical analysis of stories
  - Topical analysis of words (meaning, saliency)
  - More documents improve estimates
- Con
  - No clustering

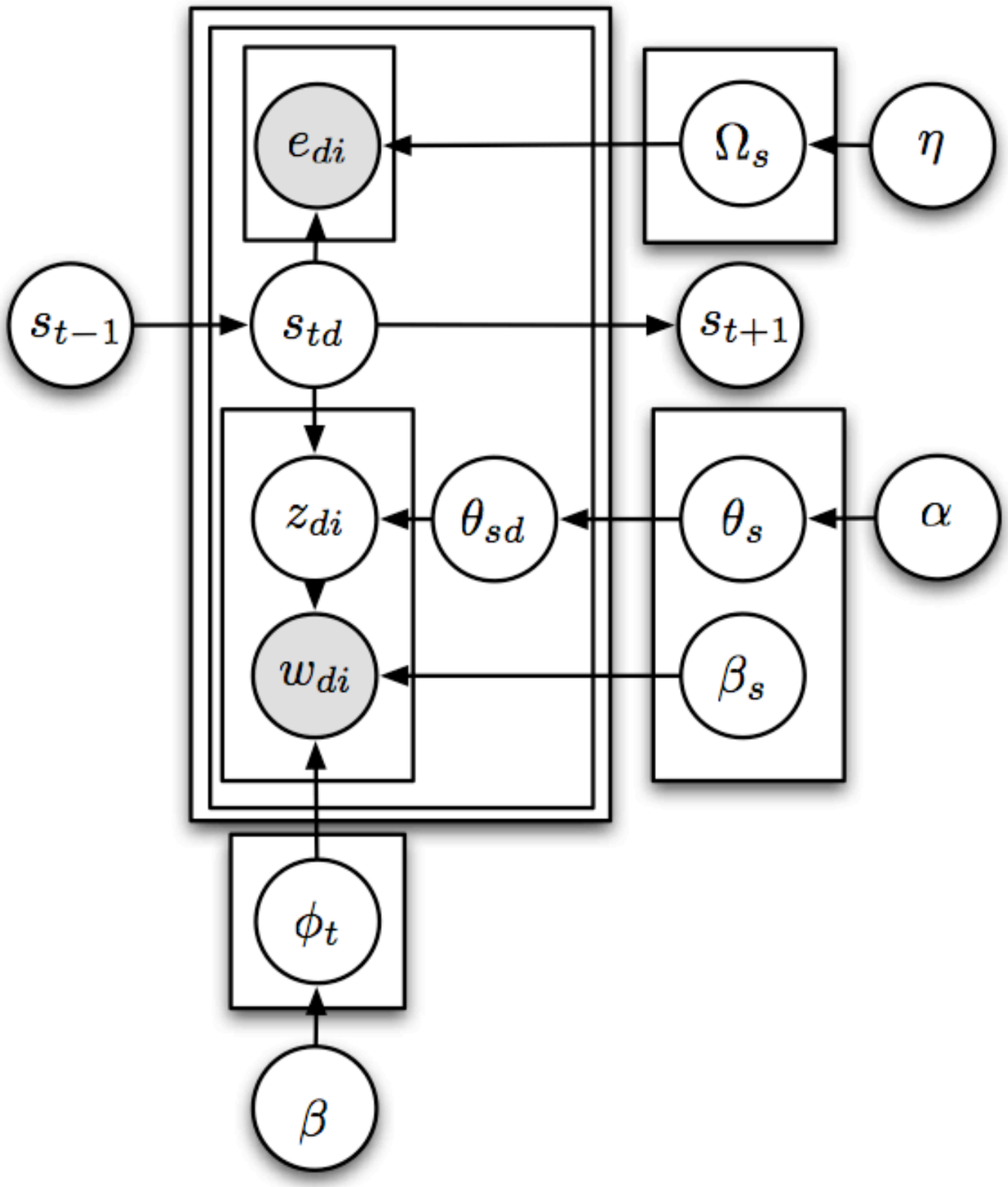
More Issues

More Issues

- Named entities are special, topics less
  (e.g. Tiger Woods and his mistresses)
- Some stories are strange
  (topical mixture is not enough - dirty models)
- Articles deviate from general story
  (Hierarchical DP)

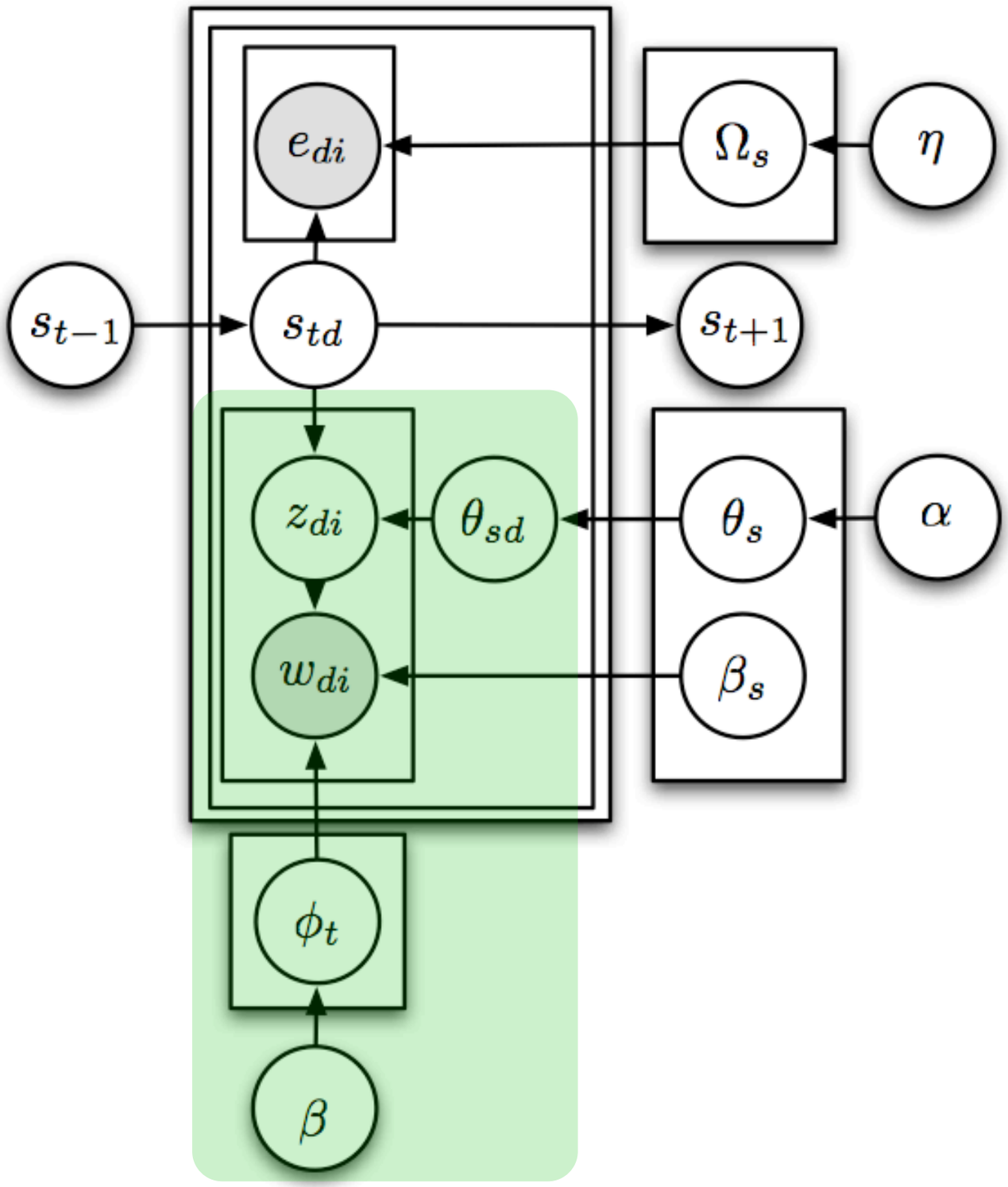
Storylines

Storylines Model



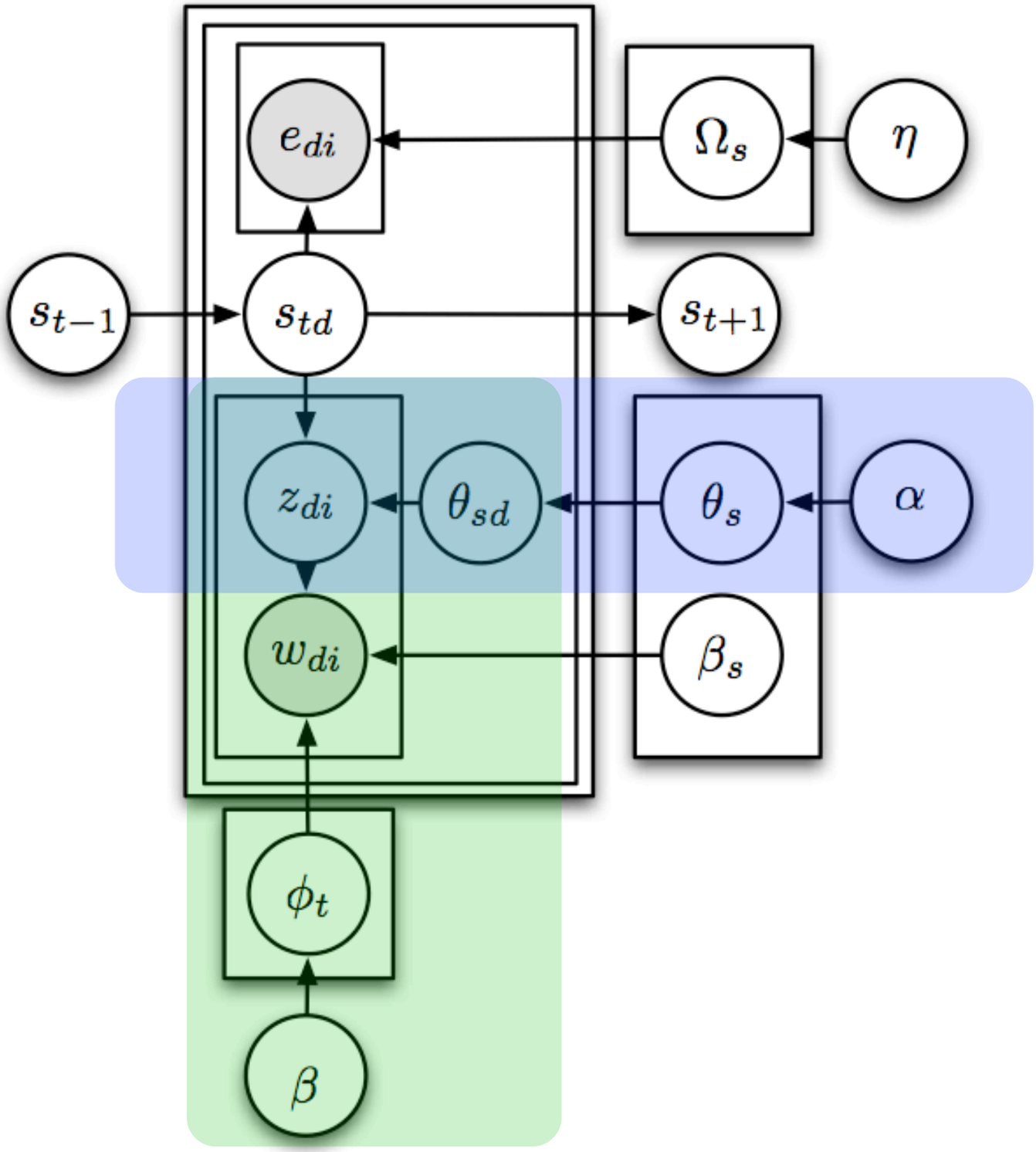
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model



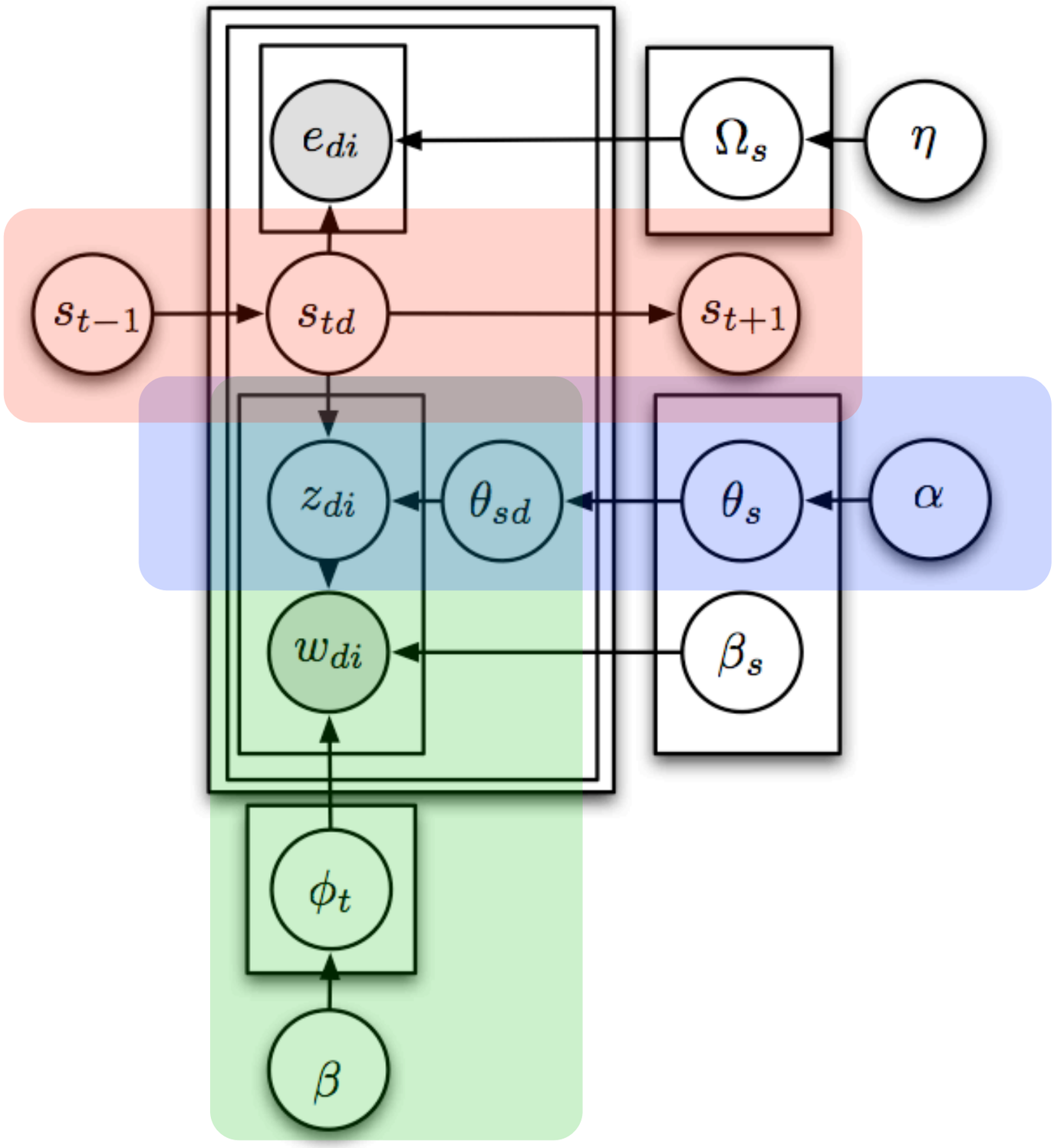
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model



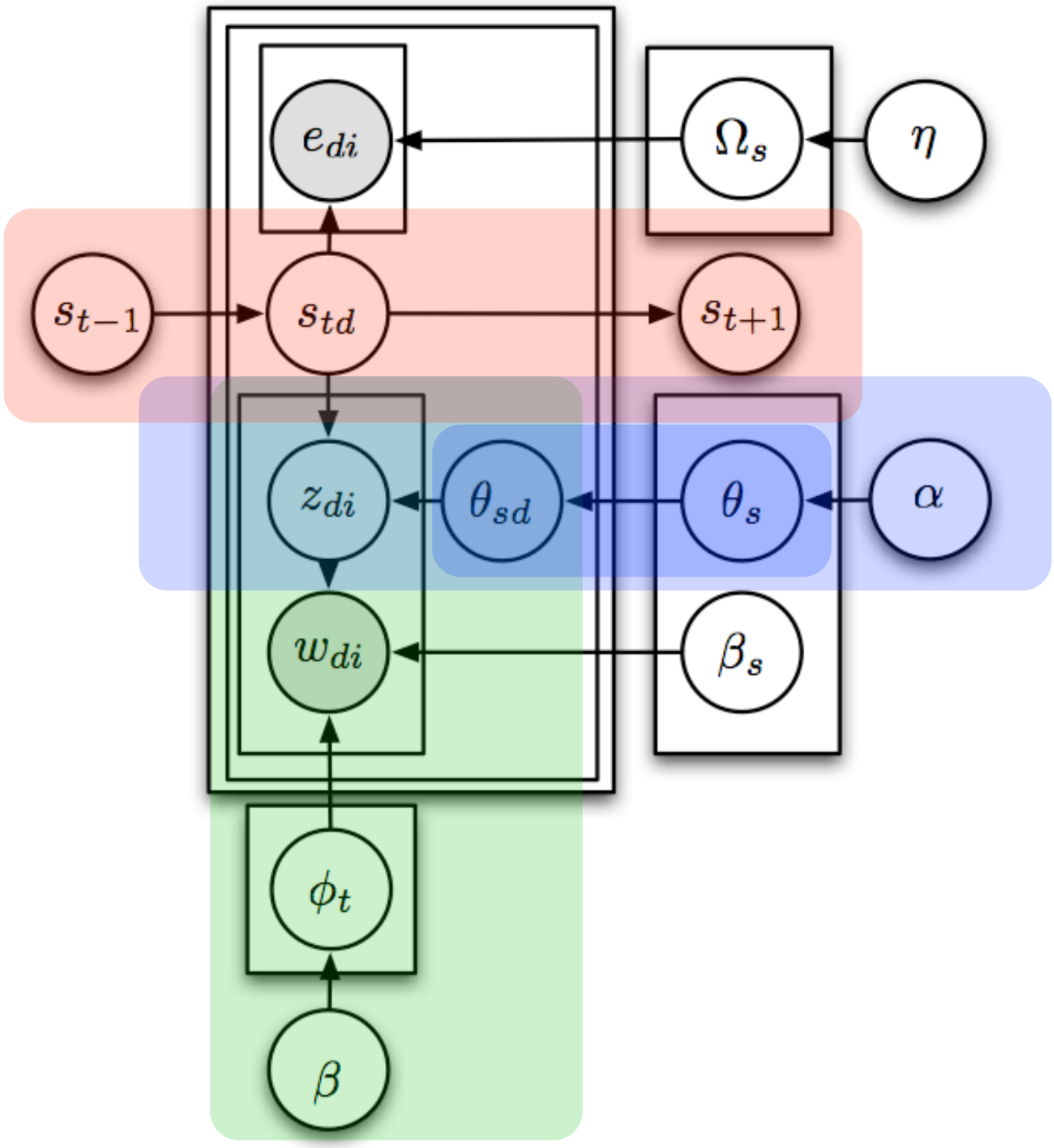
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model



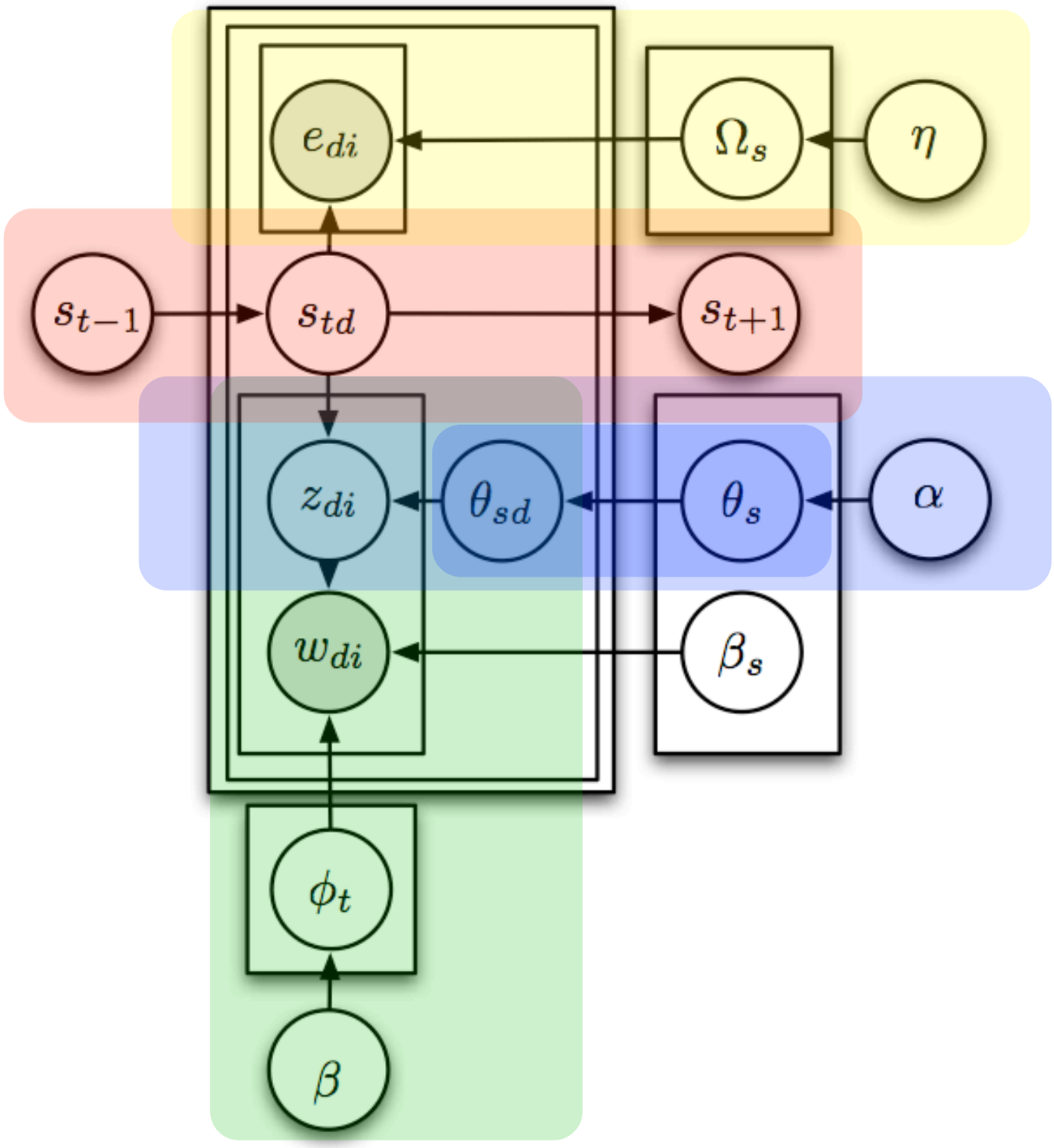
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model



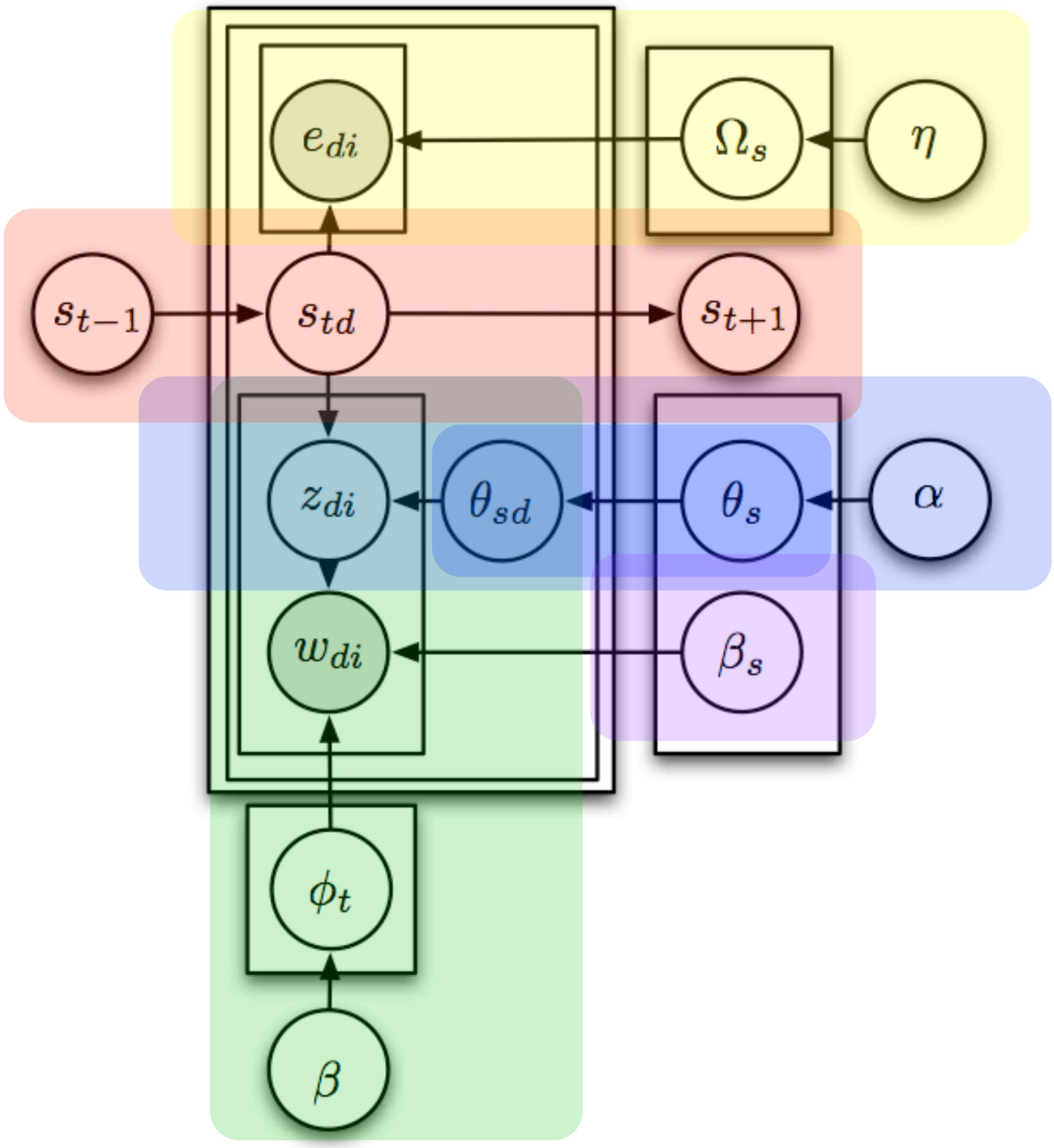
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model



- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model



- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Dynamic Cluster-Topic Hybrid

Sports
games
Won
Team
Final
Season
League
held

Politics
Government
Minister
Authorities
Opposition
Officials
Leaders
group

Accidents
Police
Attack
run
man
group
arrested
move

UEFA-soccer

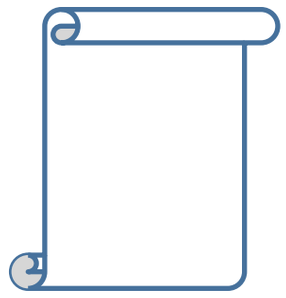
Champions
Goal
Coach
Striker
Midfield
penalty

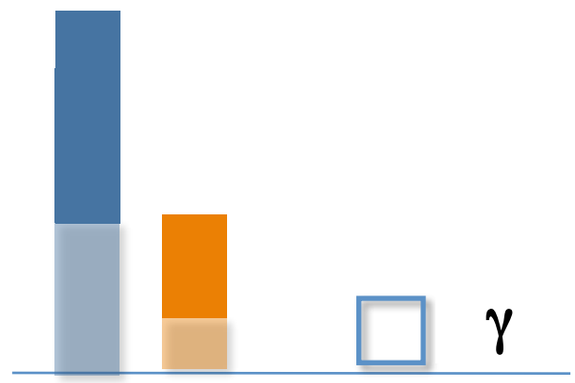
Juventus
AC Milan
Lazio
Ronaldo
Lyon

Tax-Bill

Tax
Billion
Cut
Plan
Budget
Economy

Bush
Senate
Fleischer
White House
Republican





Dynamic Cluster-Topic Hybrid

Sports
games
Won
Team
Final
Season
League
held

Politics
Government
Minister
Authorities
Opposition
Officials
Leaders
group

Accidents
Police
Attack
run
man
group
arrested
move

UEFA-soccer

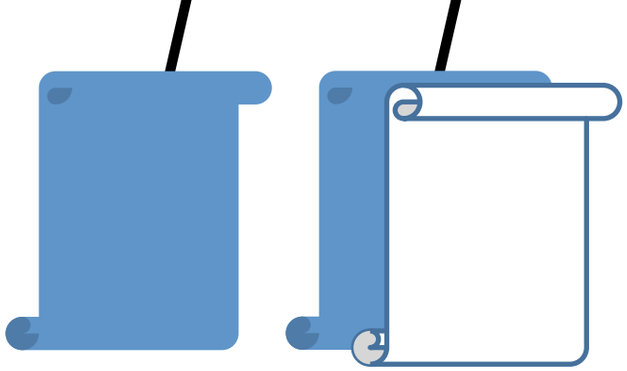
Champions
Goal
Coach
Striker
Midfield
penalty

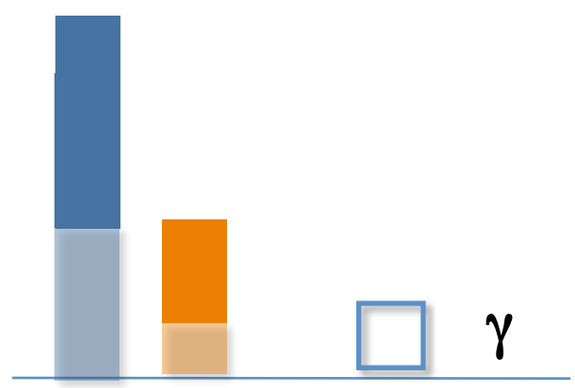
Juventus
AC Milan
Lazio
Ronaldo
Lyon

Tax-Bill

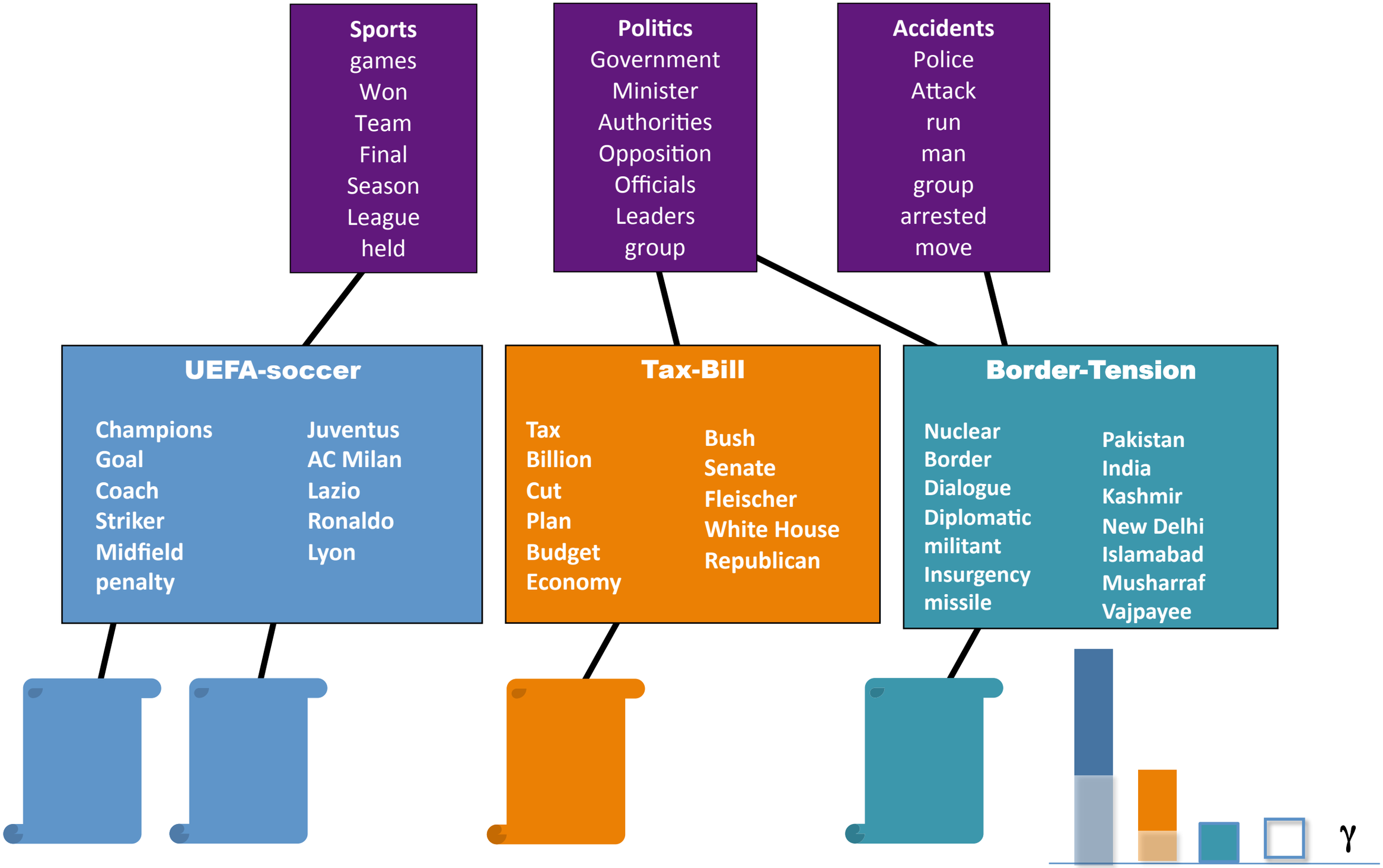
Tax
Billion
Cut
Plan
Budget
Economy

Bush
Senate
Fleischer
White House
Republican





Dynamic Cluster-Topic Hybrid



Inference

- We receive articles as a stream
  Want topics & stories now
- Variational inference infeasible
  (RCRP, sparse to dense, vocabulary size)
- We have a ‘tracking problem’
  - Sequential Monte Carlo
  - Use sampled variables of surviving particle
  - Use ideas from Cannini et al. 2009

Particle Filter

- Proposal distribution - draw stories s, topics z

$$p(s_{t+1}, z_{t+1} | x_{1\ldots t+1}, s_{1\ldots t}, z_{1\ldots t})$$

using Gibbs Sampling for each particle

- Reweight particle via

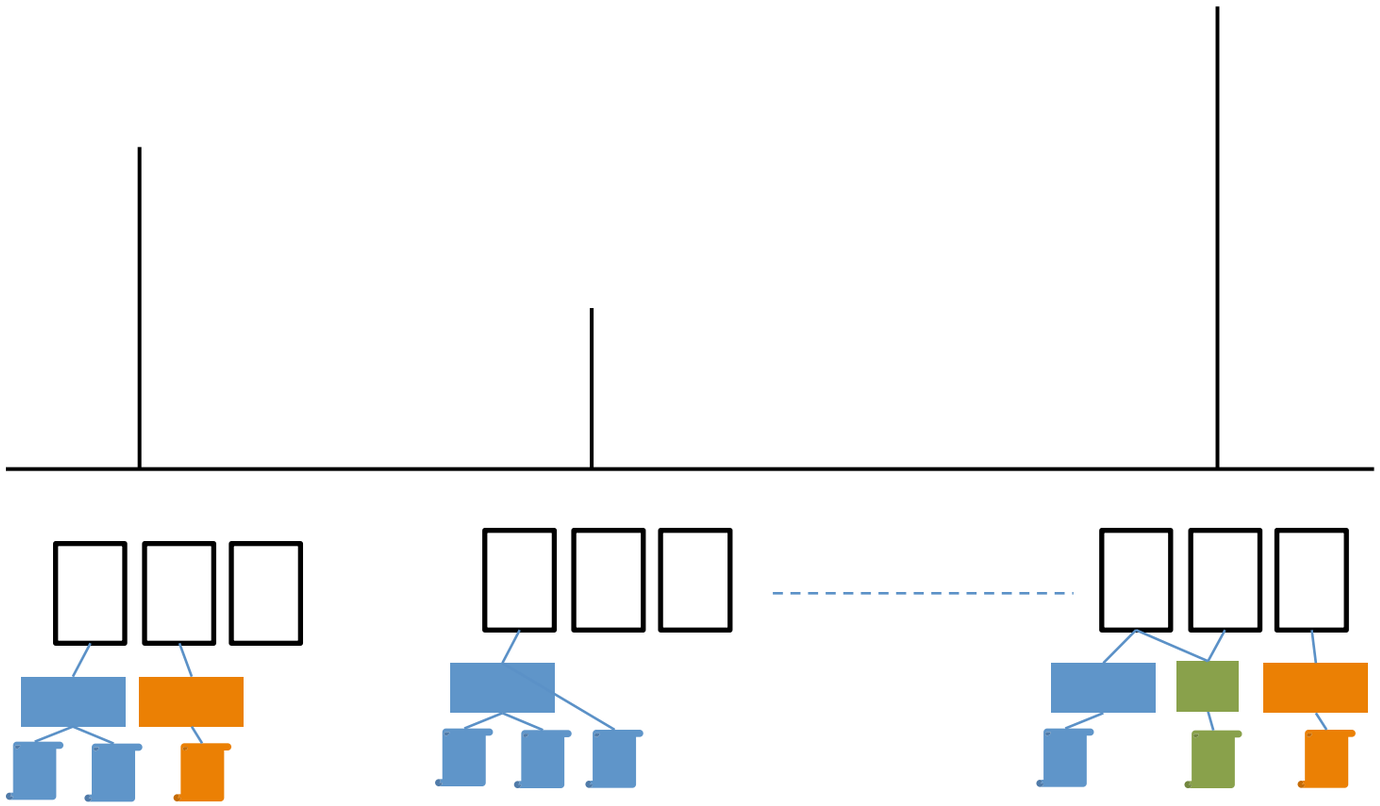
$$p(x_{t+1} | x_{1\ldots t}, s_{1\ldots t}, z_{1\ldots t})$$

new data

past state

- Resample particles if l_2 norm too large
  (resample some assignments for diversity, too)
- Compare to multiplicative updates algorithm
  In our case predictive likelihood yields weights

Particle Filter

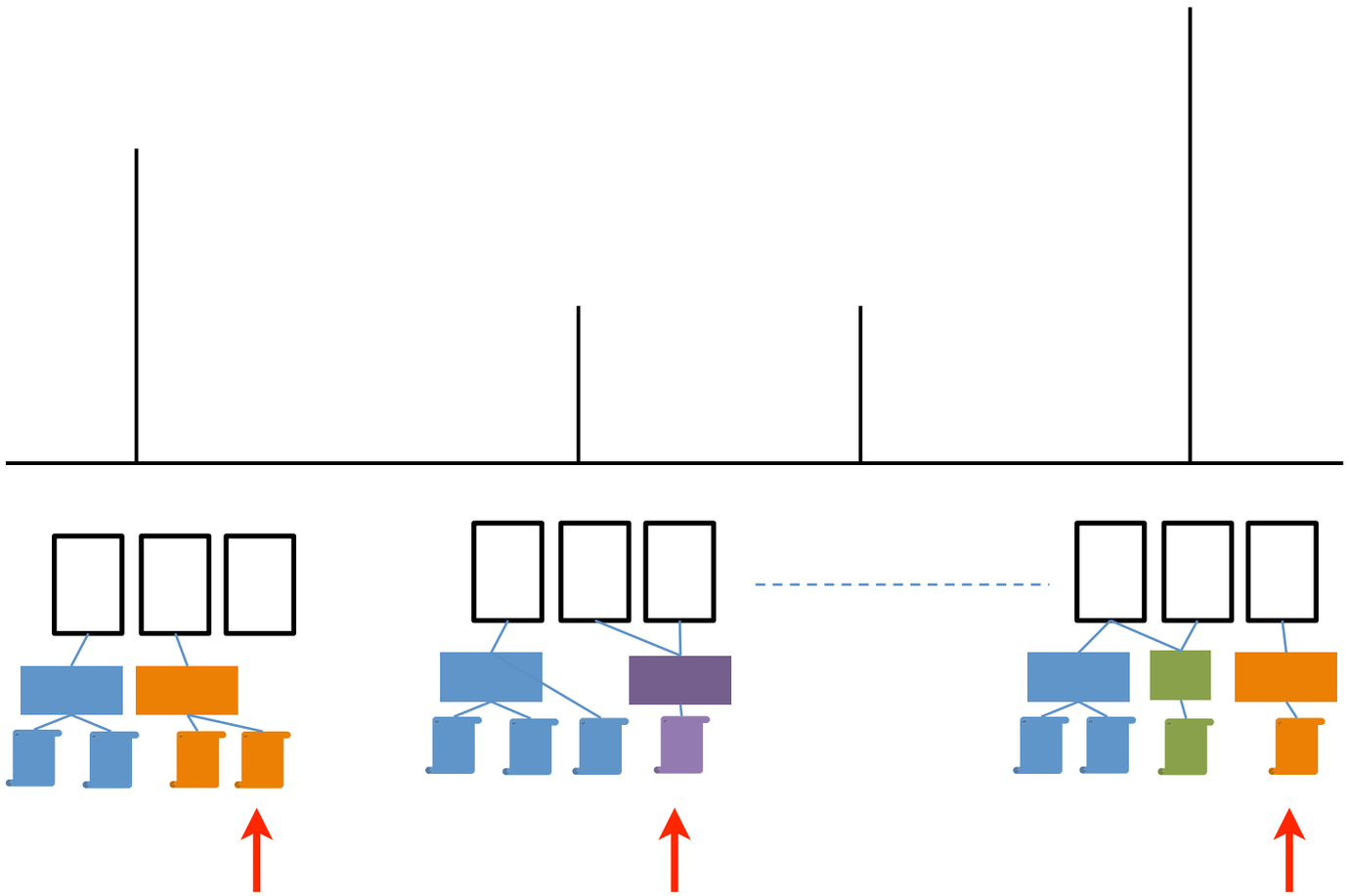


Algorithm 1 A Particle Filter Algorithm

Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  end for
  Normalize particle weights
  if $\|\omega_t\|_2^{-2} <$ threshold **then**
    resample particles
    for $f \in \{1, \ldots F\}$ **do**
      MCMC pass over 10 random past documents
    end for
  end if
end for

- **s** and **z** are tightly coupled
- Alternative to MCMC
  - Sample **s** then sample **z** (high variance)
  - Sample **z** then sample **s** (doesn't make sense)
- Idea (following a similar trick by Jain and Neal)
  - Run a few iterations of MCMC over s and z
  - Take last sample as the proposed value

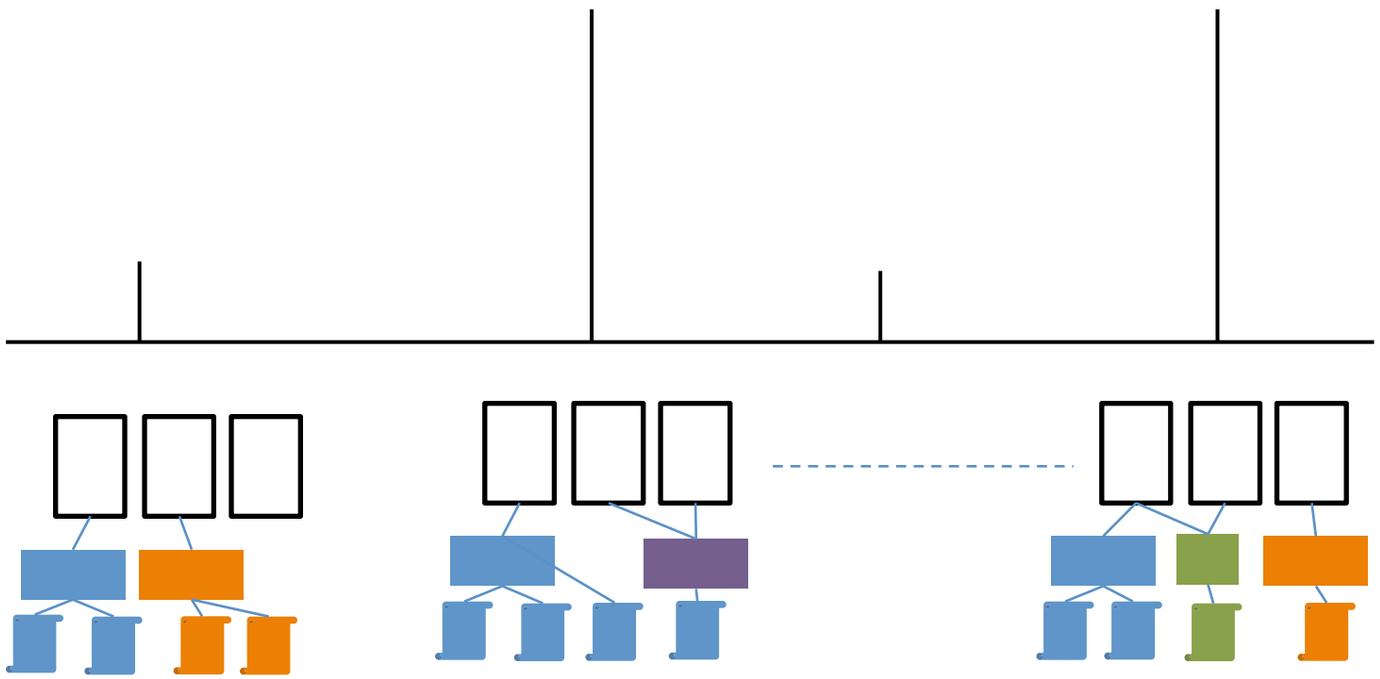
Particle Filter



Algorithm 1 A Particle Filter Algorithm

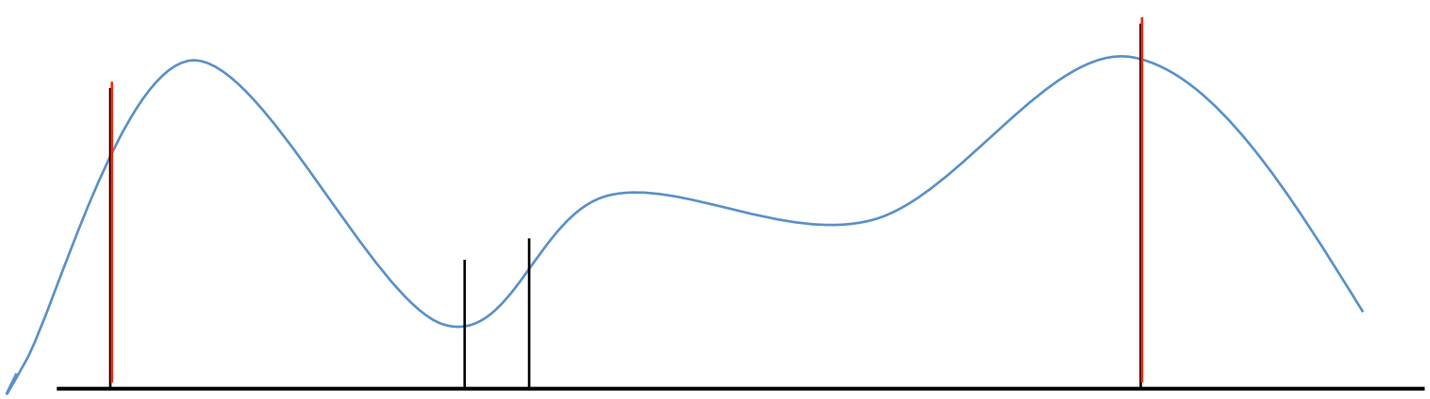
Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  end for
  Normalize particle weights
  if $\|\omega_t\|_2^{-2} <$ threshold **then**
    resample particles
    for $f \in \{1, \ldots F\}$ **do**
      MCMC pass over 10 random past documents
    end for
  end if
end for

Particle Filter

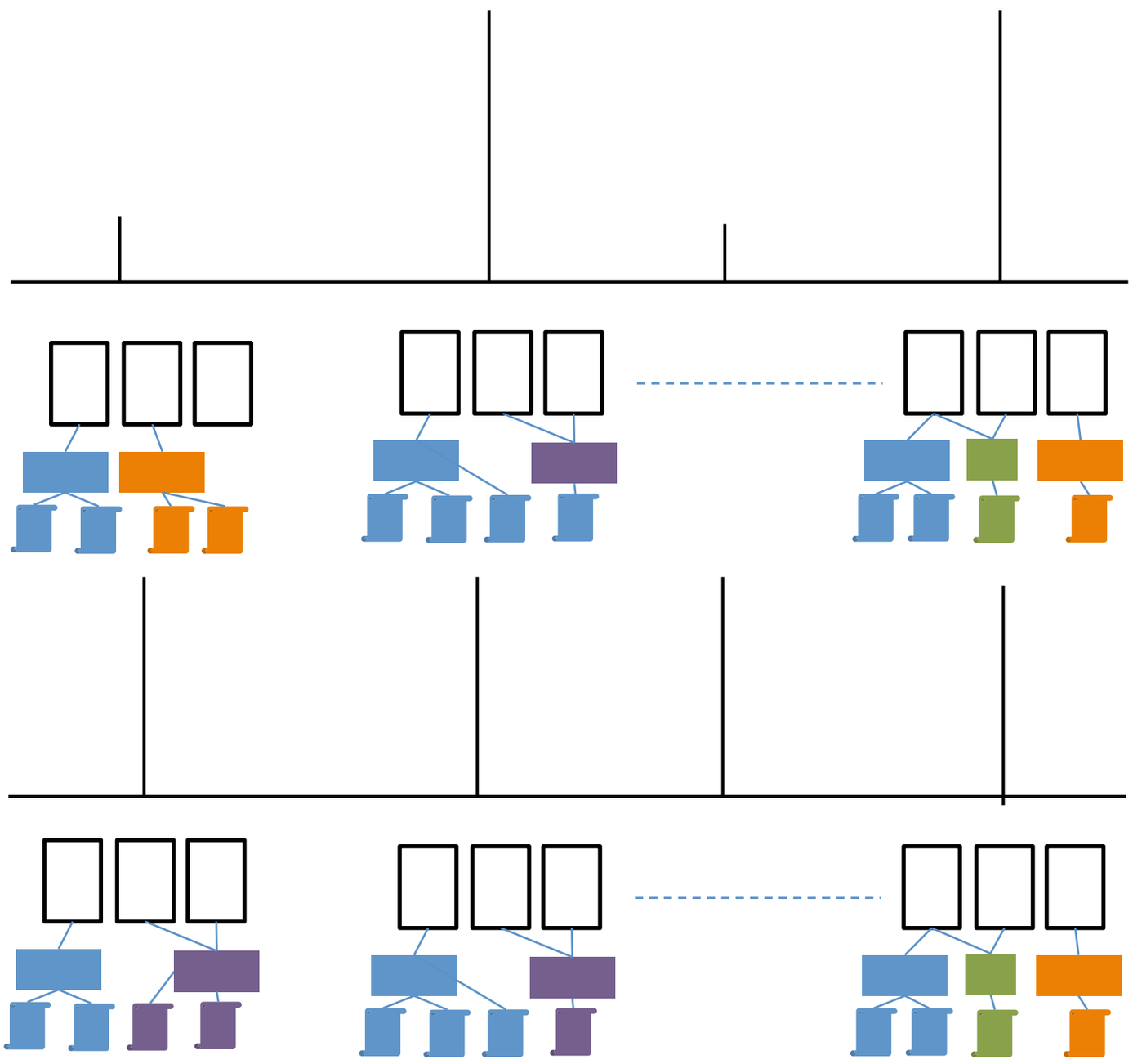


Algorithm 1 A Particle Filter Algorithm

Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  end for
  Normalize particle weights
  if $\|\omega_t\|_2^{-2} <$ threshold **then**
    resample particles
    for $f \in \{1, \ldots F\}$ **do**
      MCMC pass over 10 random past documents
    end for
  end if
end for



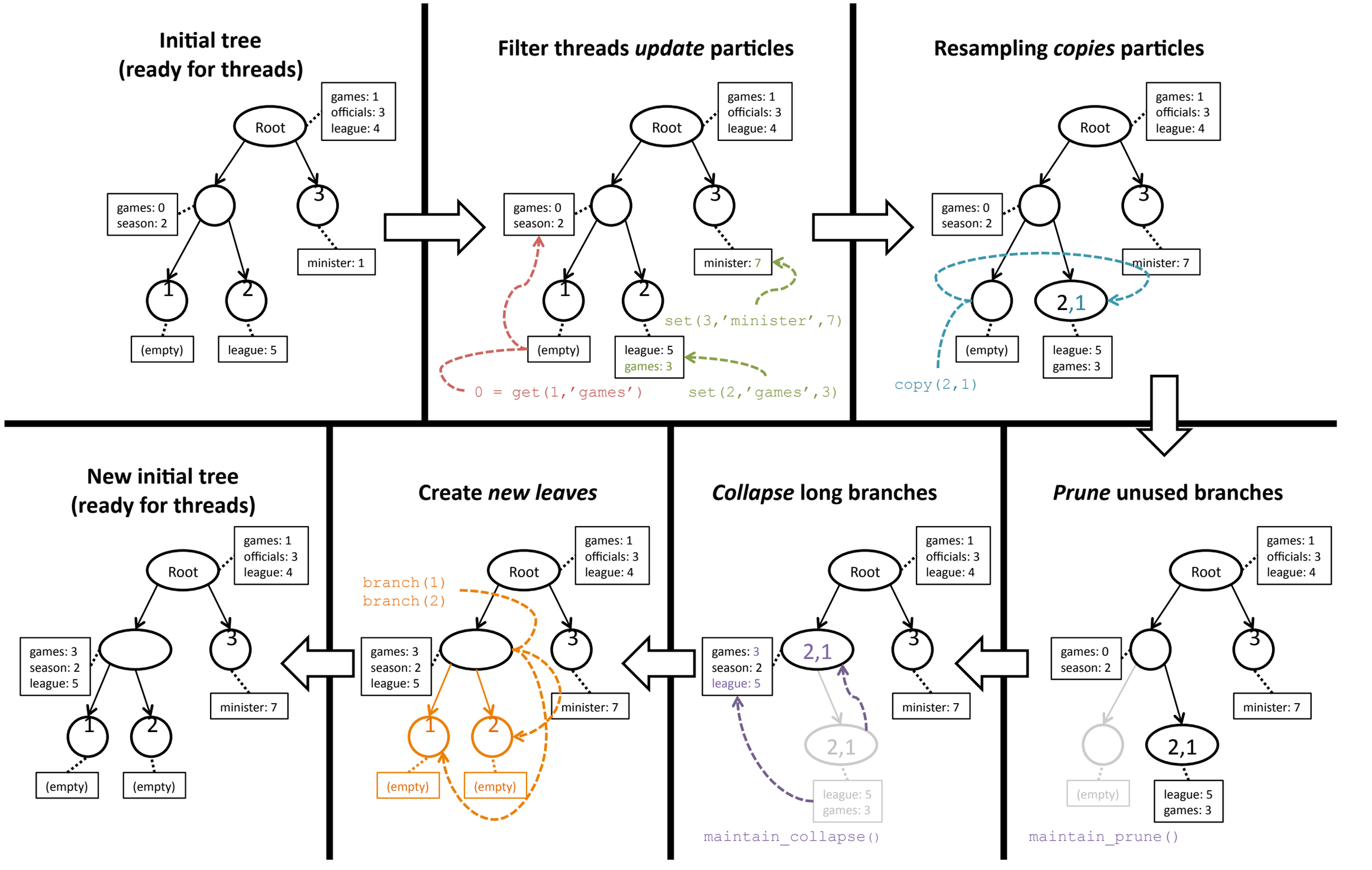
Particle Filter



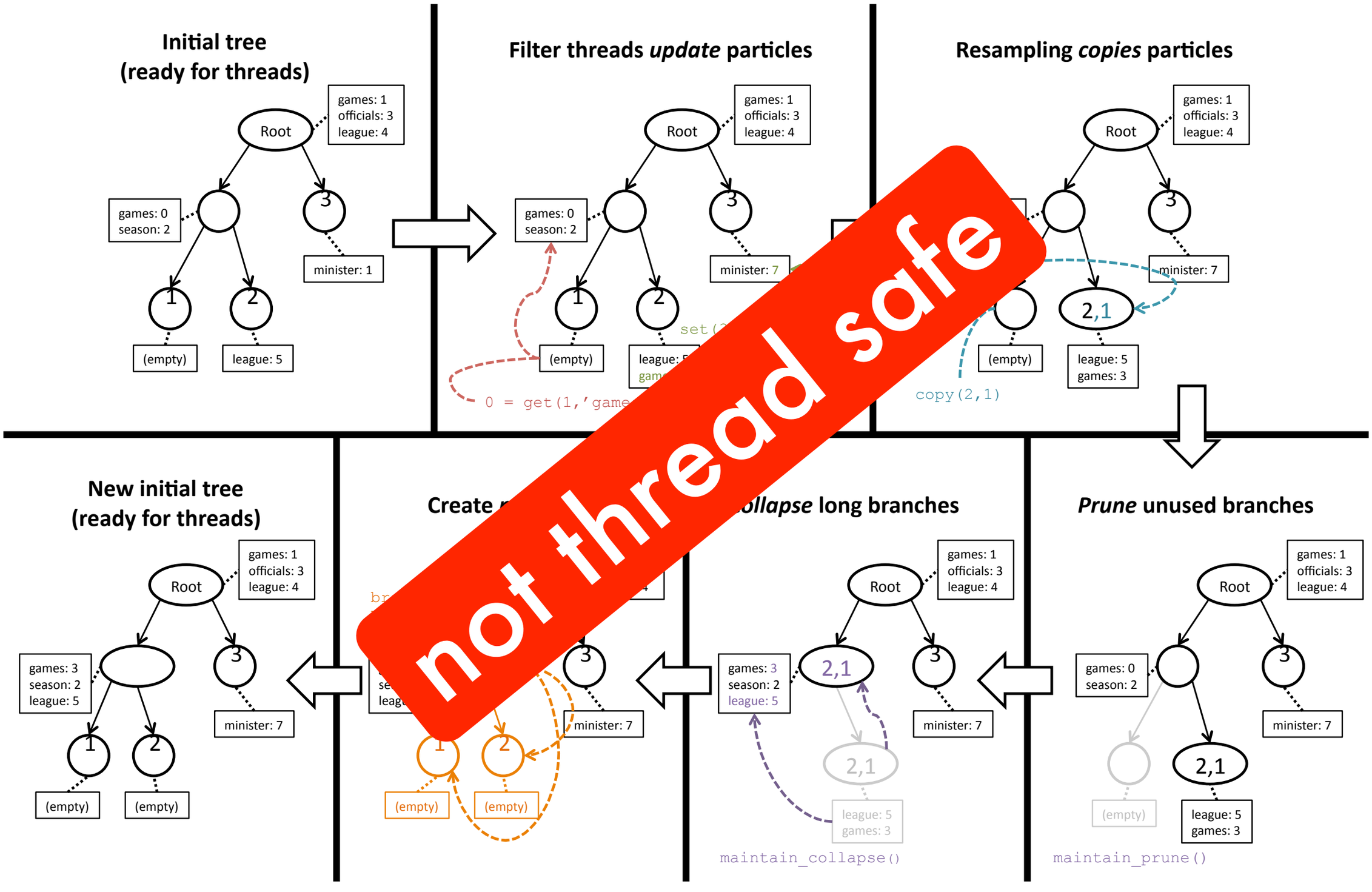
Algorithm 1 A Particle Filter Algorithm

Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  end for
  Normalize particle weights
  if $\|\omega_t\|_2^{-2} <$ threshold **then**
    resample particles
    for $f \in \{1, \ldots F\}$ **do**
      MCMC pass over 10 random past documents
    end for
  end if
end for

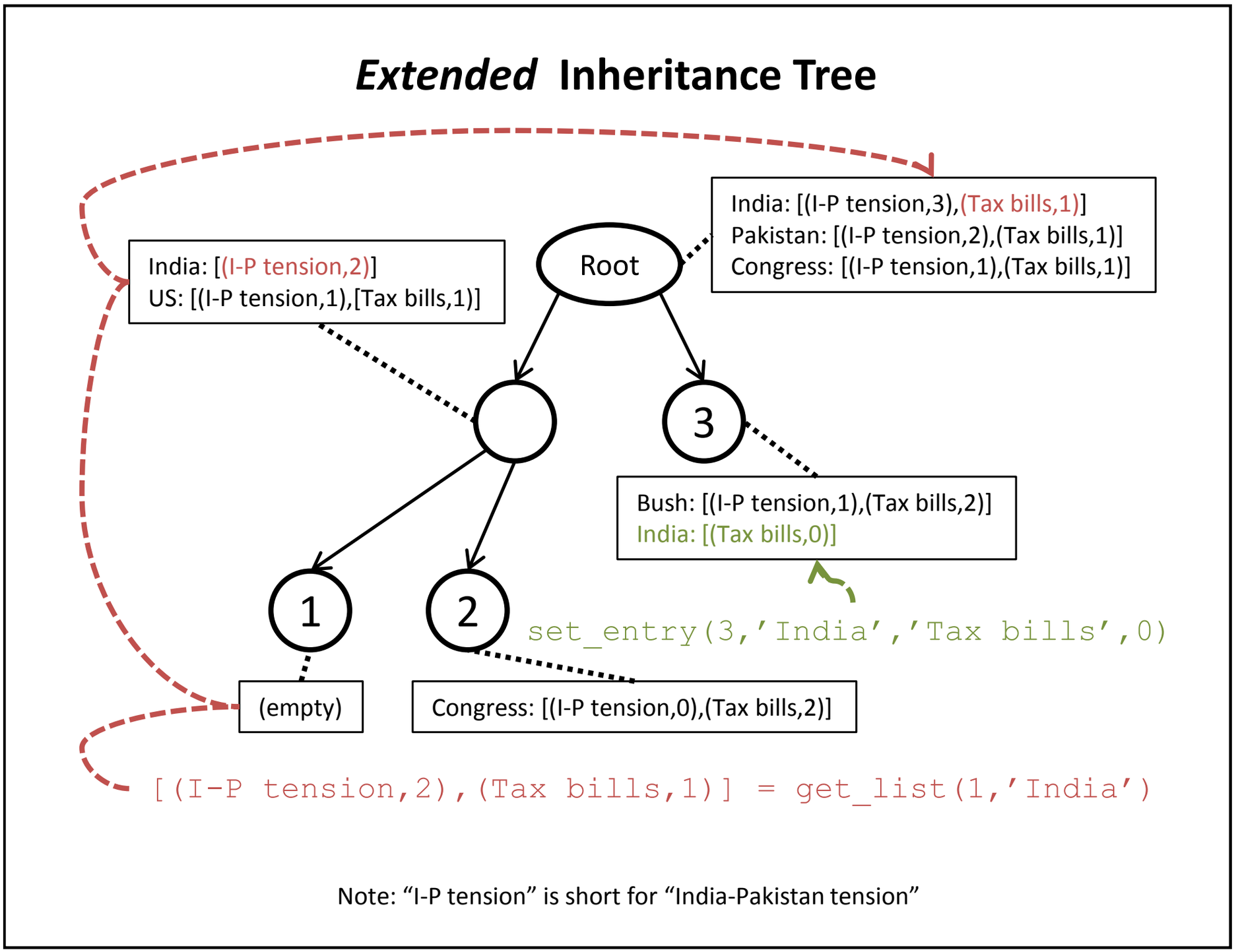
Inheritance Tree



Inheritance Tree



Extended Inheritance Tree



write only in the leaves (per thread)

Results

Ablation studies

- TDT5 (Topic Detection and Tracking)
  macro-averaged minimum detection cost: 0.714

- Removing features

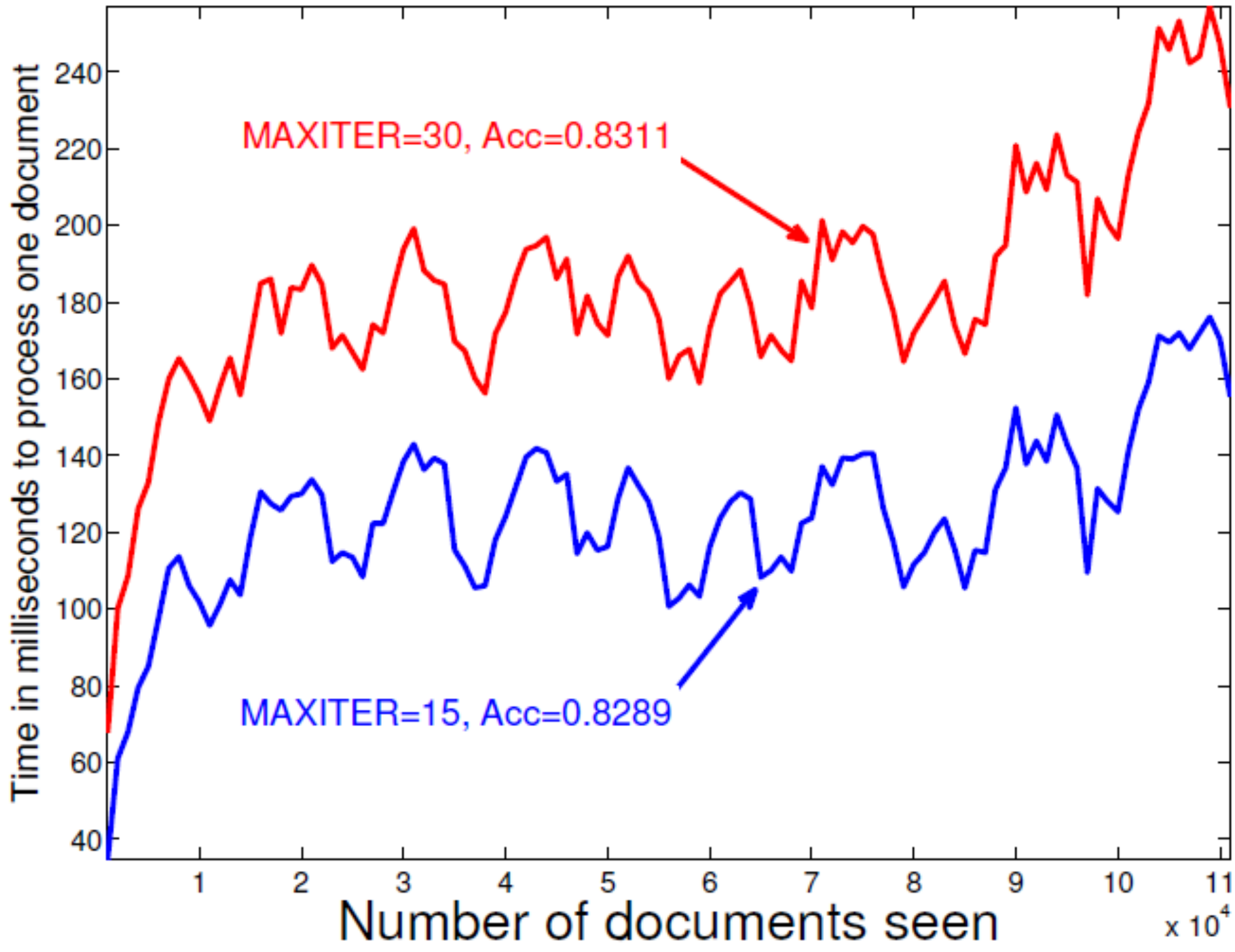
time	entities	topics	story words
0.84	0.90	0.86	0.75

Comparison

Sample No.	Sample size	Num Words	Num Entities	Story Acc.	LSHC Acc.
1	111,732	19,218	12,475	**0.8289**	0.738
2	274,969	29,604	21,797	**0.8388**	0.791
3	547,057	40,576	32,637	**0.8395**	0.800

Hashing & correlation clustering

Time-Accuracy trade off



Stories

TOPICS

Sports

games
won
team
final
season
league
held

Politics

government
minister
authorities
opposition
officials
leaders
group

Unrest

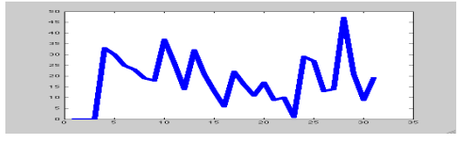
police
attack
run
man
group
arrested
move

STORYLINES

UEFA-soccer

champions
goal
leg
coach
striker
midfield
penalty

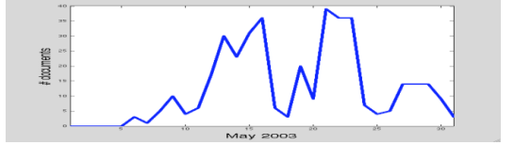
Juventus
AC Milan
Real Madrid
Milan
Lazio
Ronaldo
Lyon



Tax bills

tax
billion
cut
plan
budget
economy
lawmakers

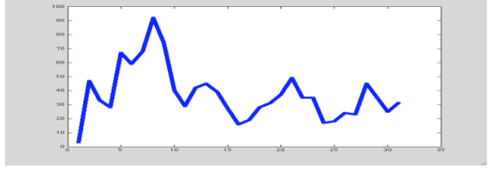
Bush
Senate
US
Congress
Fleischer
White House
Republican



India-Pakistan tension

nuclear
border
dialogue
diplomatic
militant
insurgency
missile

Pakistan
India
Kashmir
New Delhi
Islamabad
Musharraf
Vajpayee



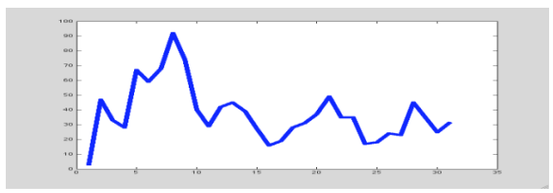
Related Stories

"Show similar stories by topic"

"Show similar stories, require the word nuclear*"*

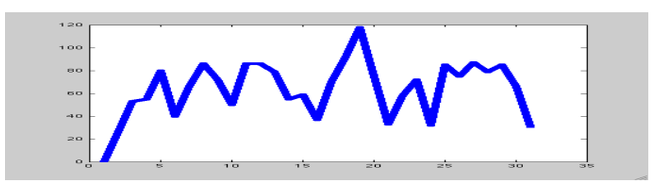
India-Pakistan tension

nuclear	*Pakistan*
border	*India*
dialogue	*Kashmir*
diplomatic	*New Delhi*
militant	*Islamabad*
insurgency	*Musharraf*
missile	*Vajpayee*



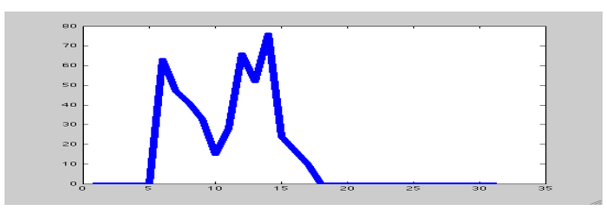
Middle-east conflict

Peace	*Israel*
Roadmap	*Palestinian*
Suicide	*West bank*
Violence	*Sharon*
Settlements	*Hamas*
bombing	*Arafat*



North Korea nuclear

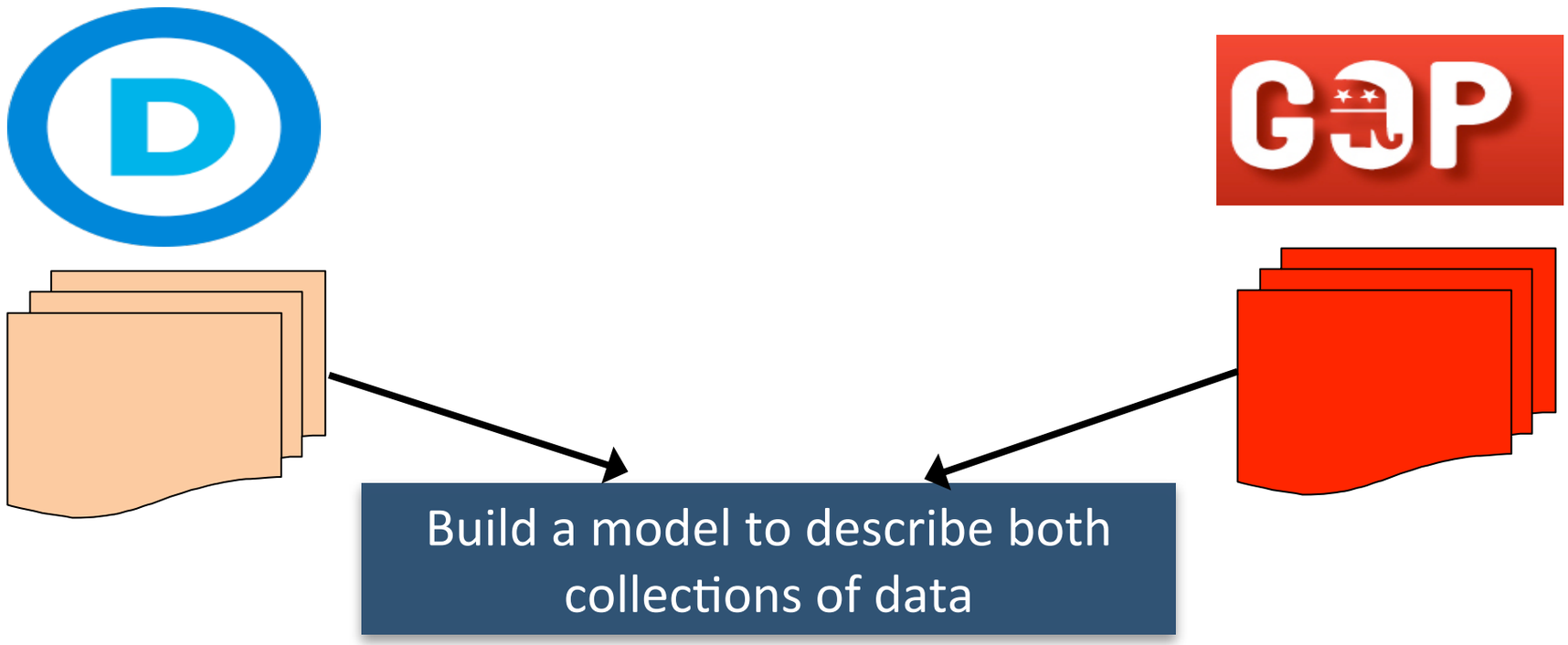
nuclear	*North Korea*
summit	*South Korea*
warning	*U.S*
policy	*Bush*
missile	*Pyongyang*
program	



Detecting Ideologies

Ahmed and Xing, 2010

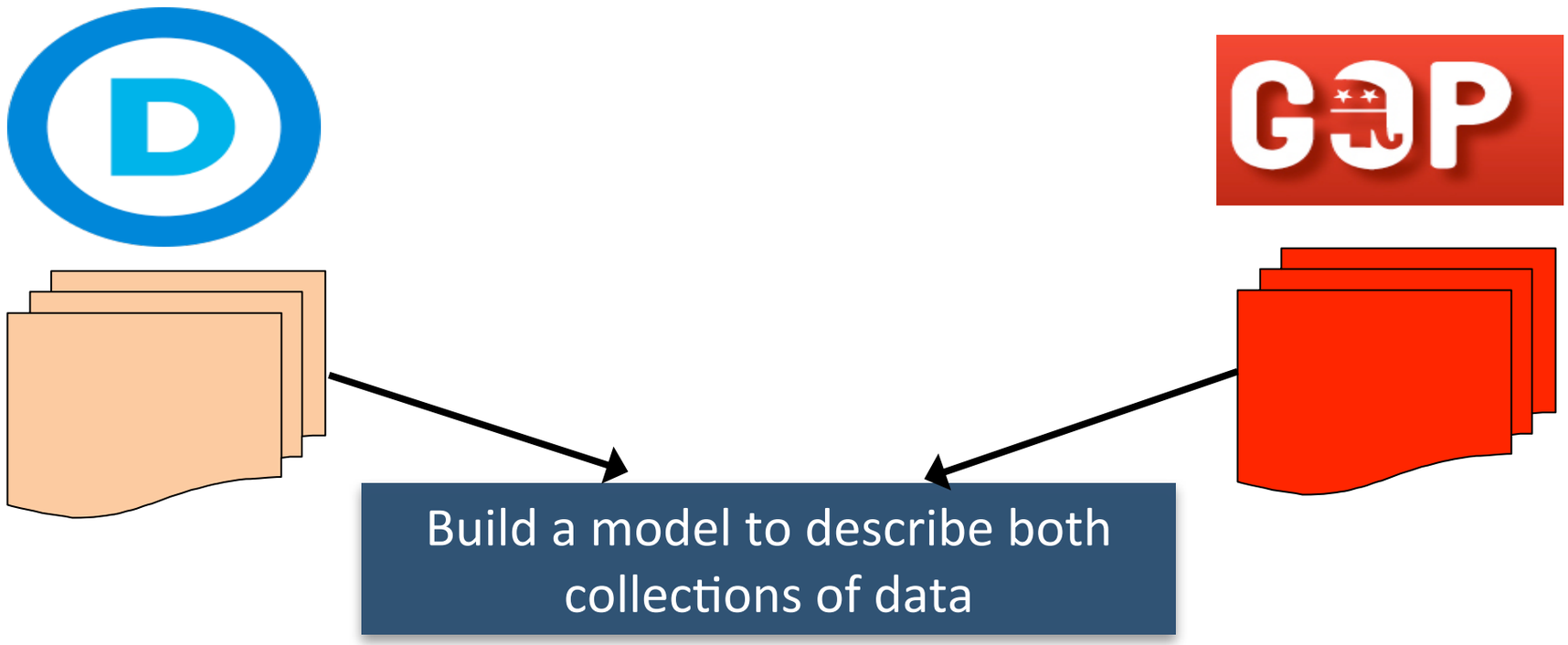
Ideologies



Visualization

- How does each ideology **view** mainstream events?
- On which topics do they **differ**?
- On which topics do they **agree**?

Ideologies

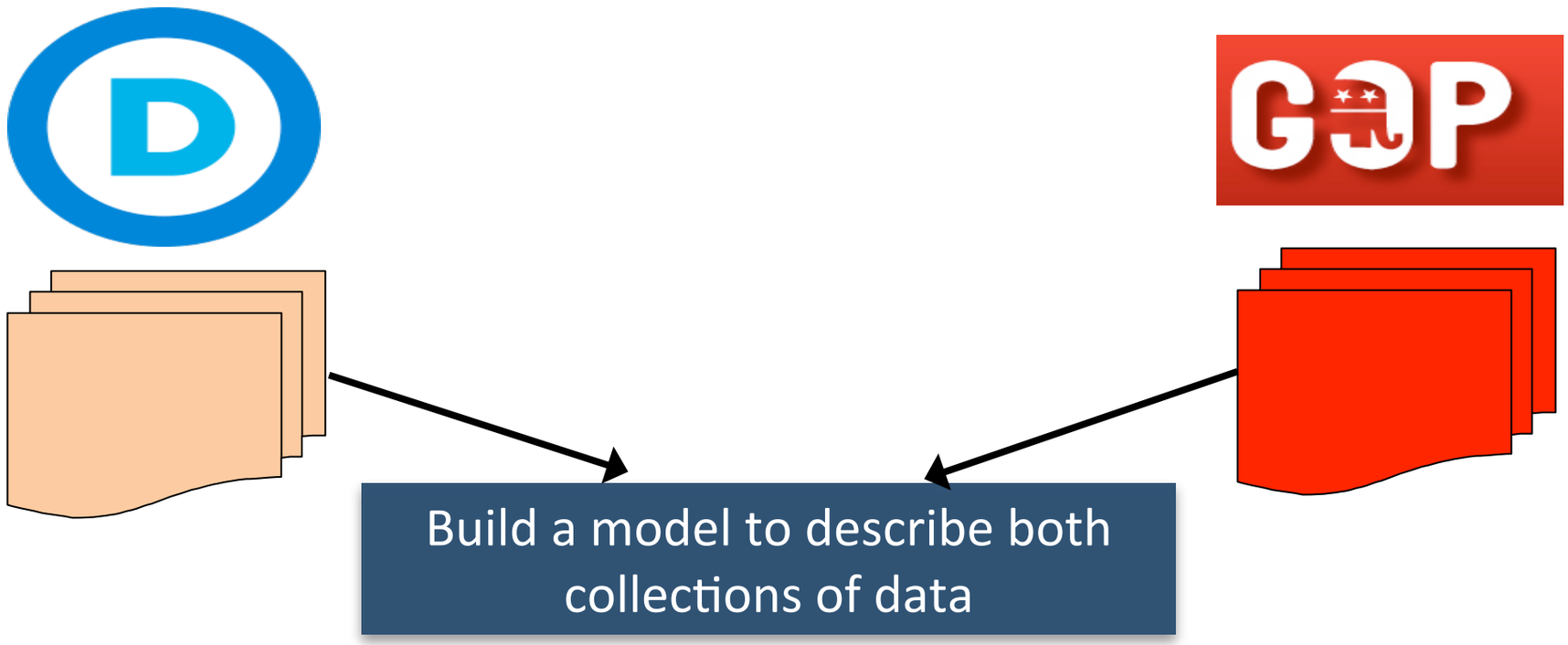


Visualization

Classification

- Given a new news article or a blog post, the system should infer
  - From which side it was written
  - Justify its answer on a topical level (view on abortion, taxes, health care)

Ideologies



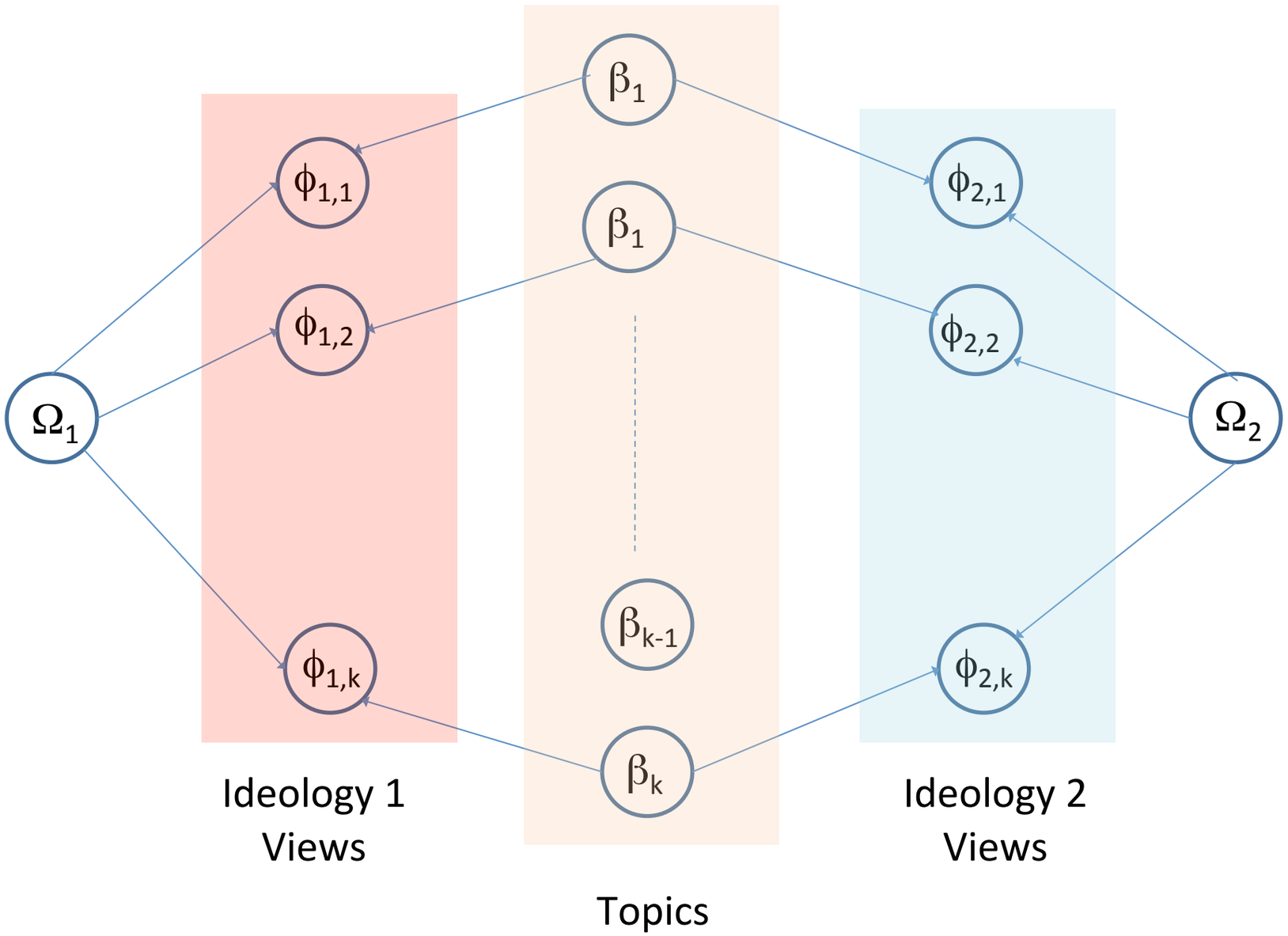
Visualization

Classification

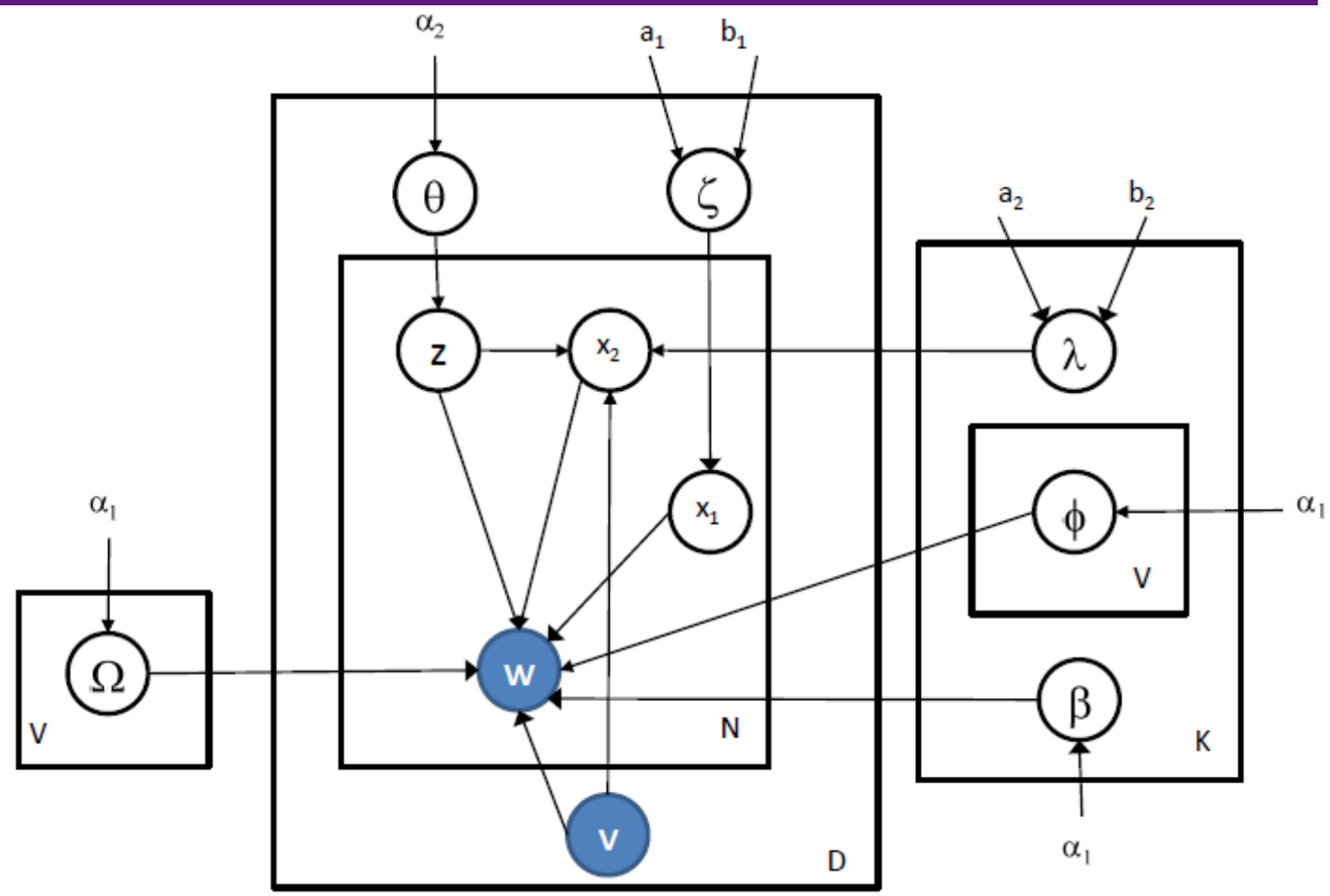
Structured browsing

- Given a **new** news article or a blog post, the user can ask for :
  - Examples of other articles from the same ideology about the same topic
  - Documents that could exemplify **alternative** views from **other ideologies**

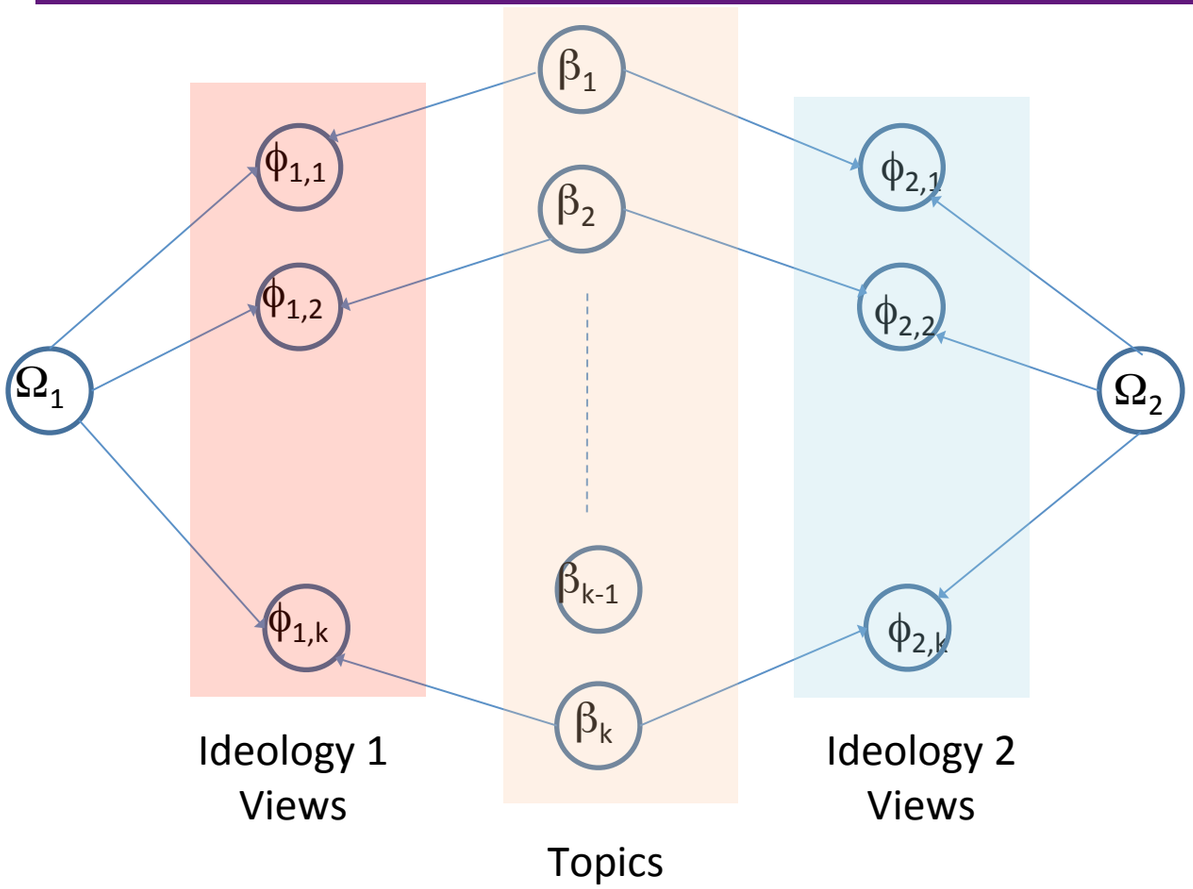
Building a factored model

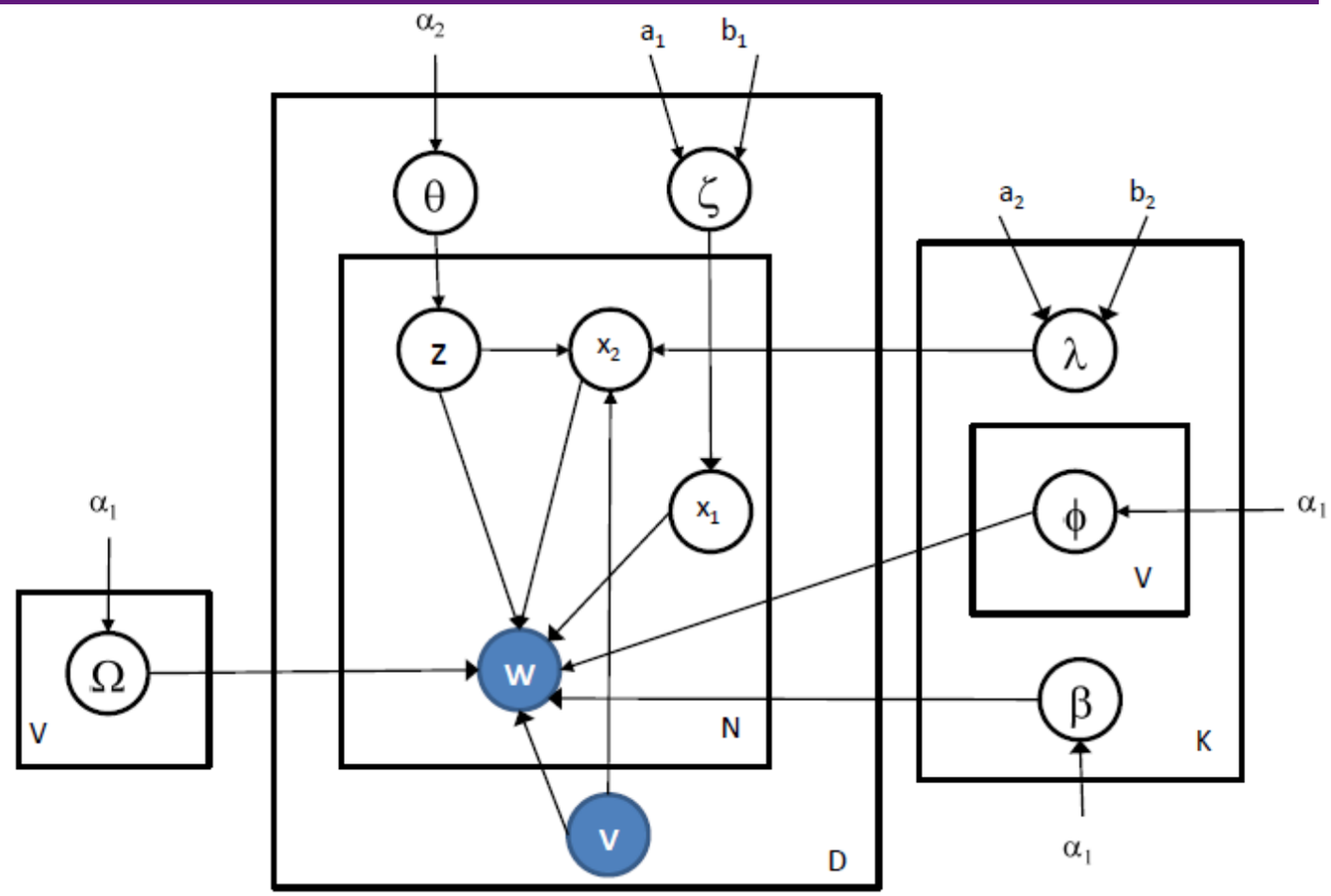


Building a factored model

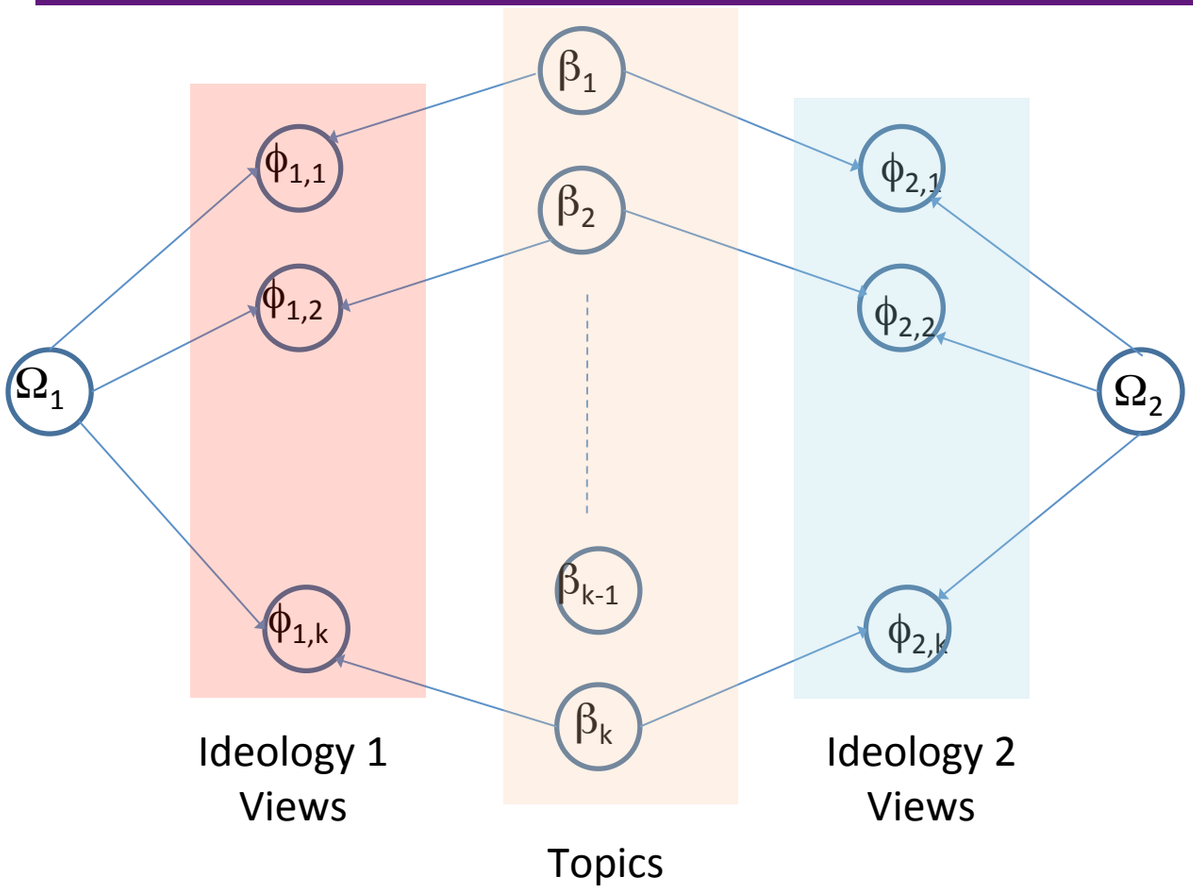


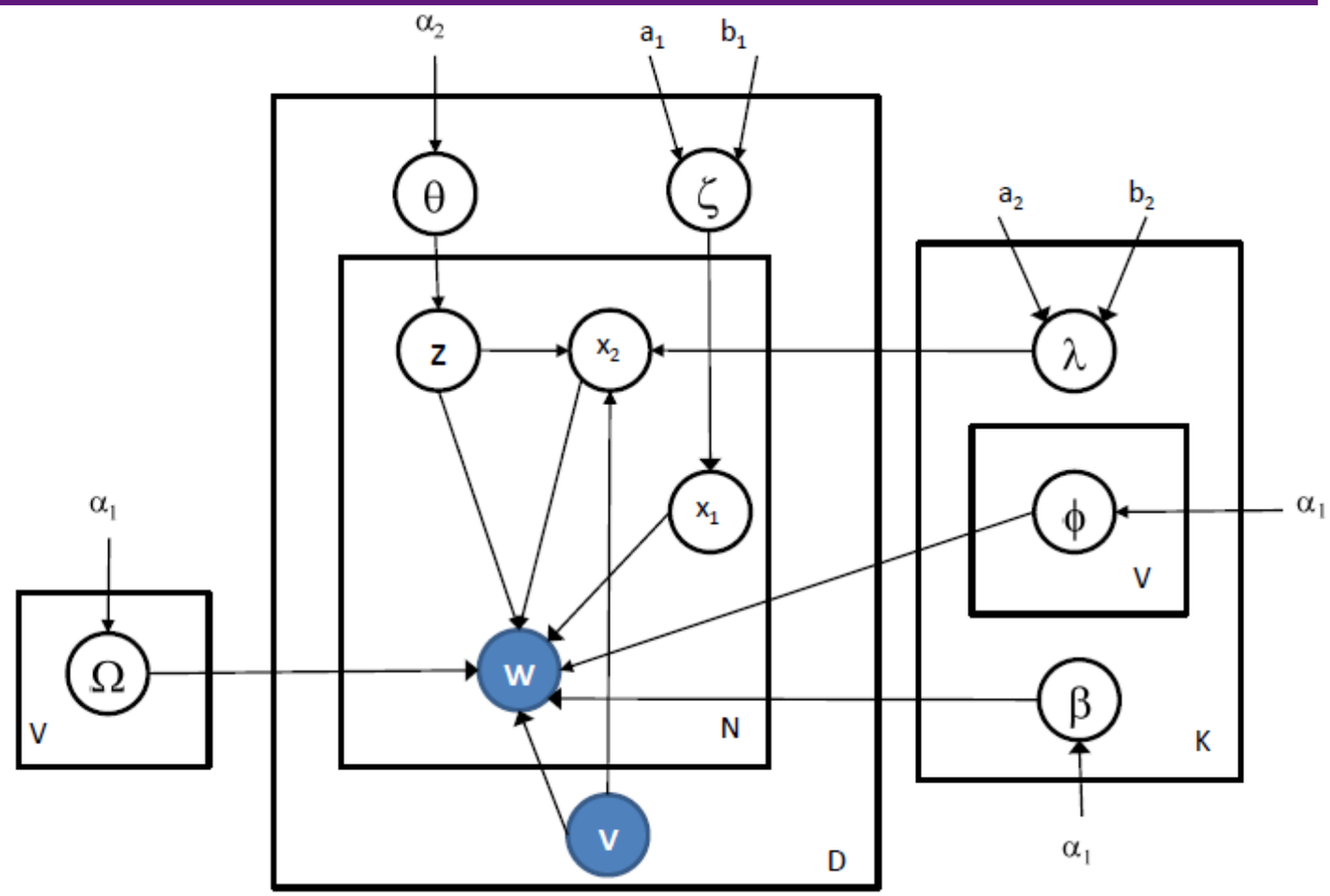
Building a factored model

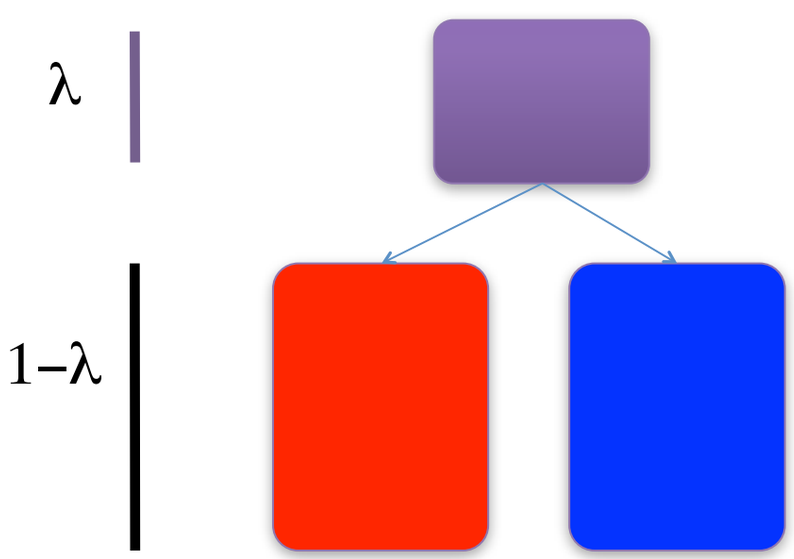


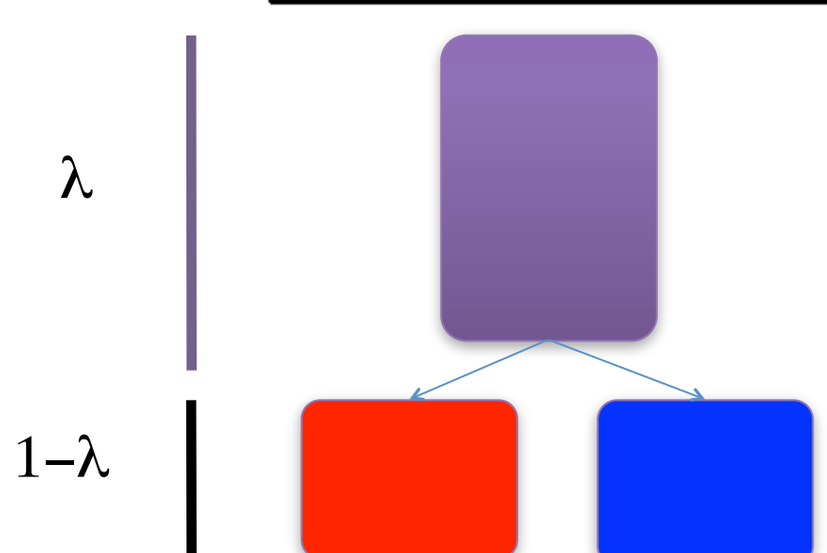


Building a factored model





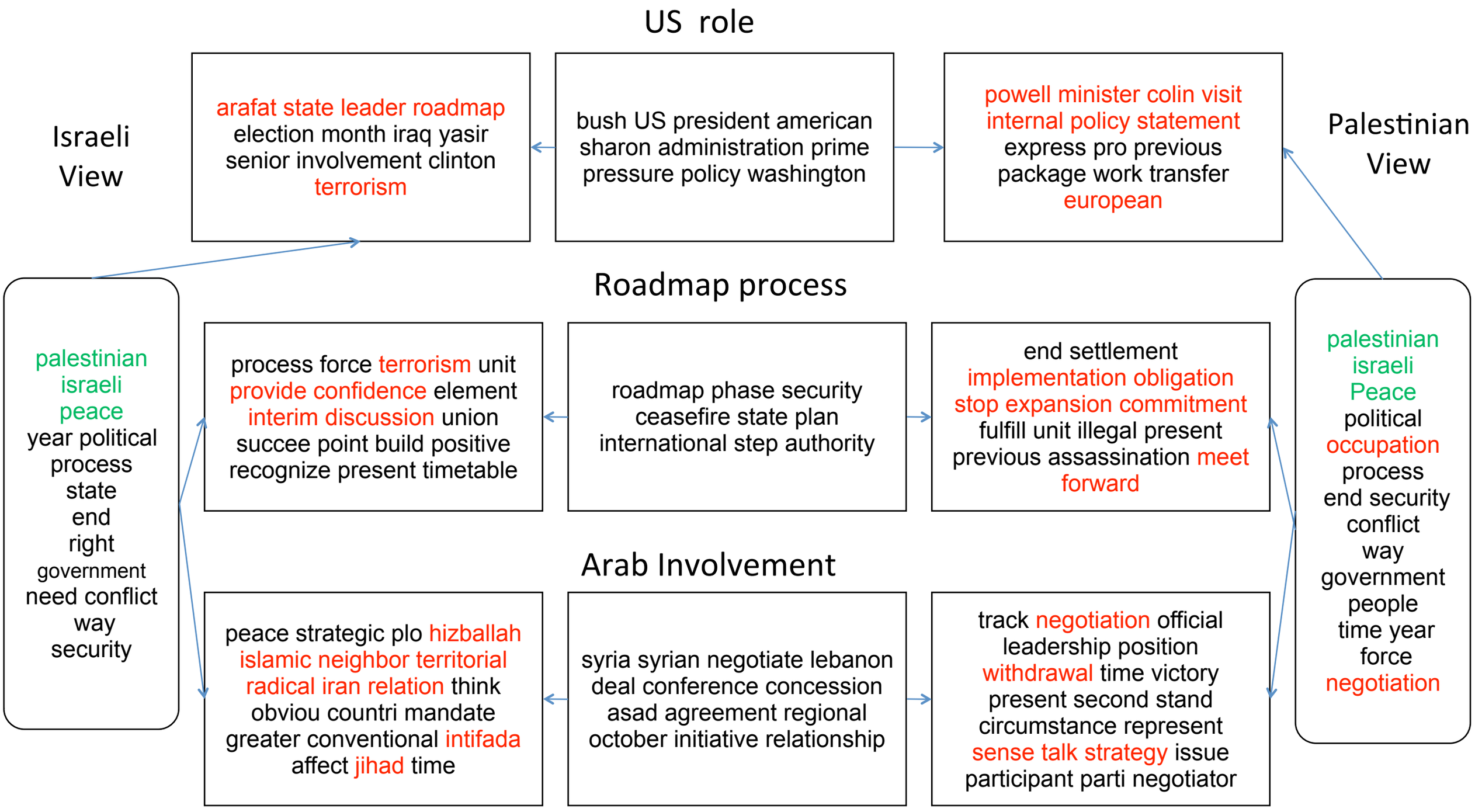




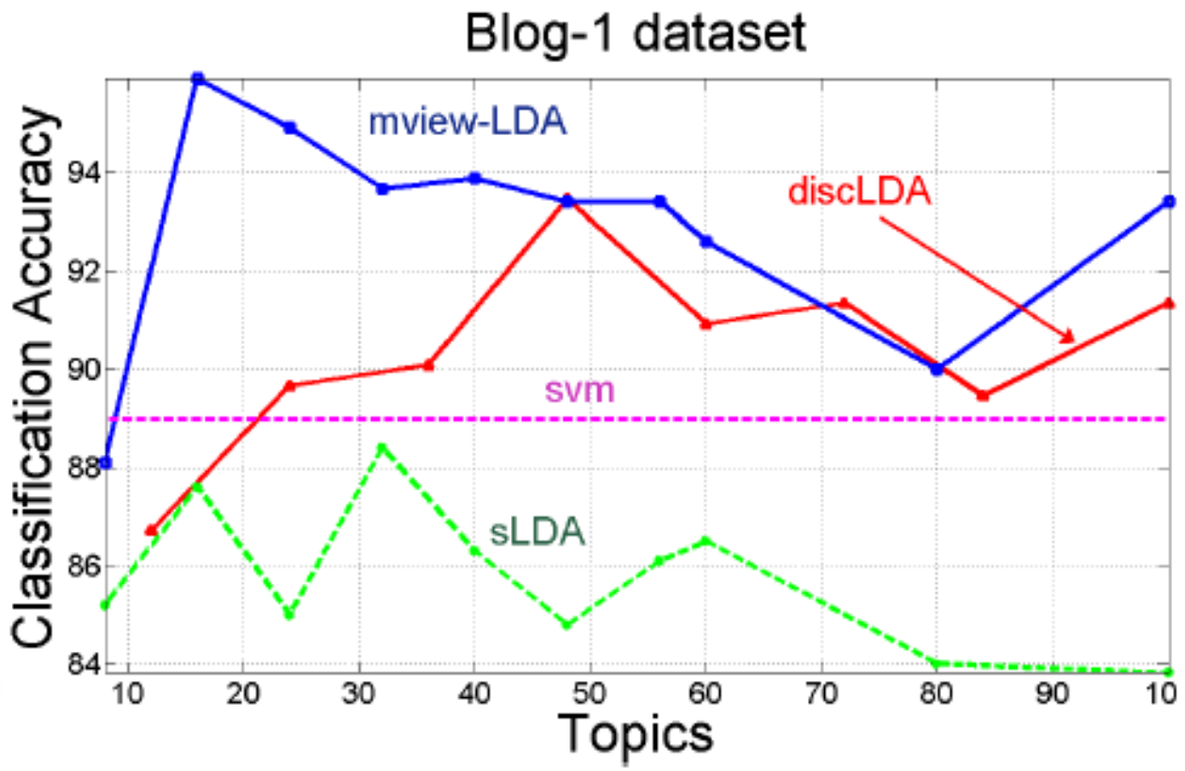
Data

- **Bitterlemons**:
  - Middle-east conflict, document written by Israeli and Palestinian authors.
  - ~300 documents form each view with average length 740
  - Multi author collection
  - 80-20 split for test and train
- **Political Blog-1:**
  - American political blogs (Democrat and Republican)
  - 2040 posts with average post length = 100 words
  - Follow test and train split as in (Yano et al., 2009)
- **Political Blog-**2 (test generalization to a new writing style)
  - Same as 1 but 6 blogs, 3 from each side
  - ~14k posts with ~200 words per post
  - 4 blogs for training and 2 blogs for test

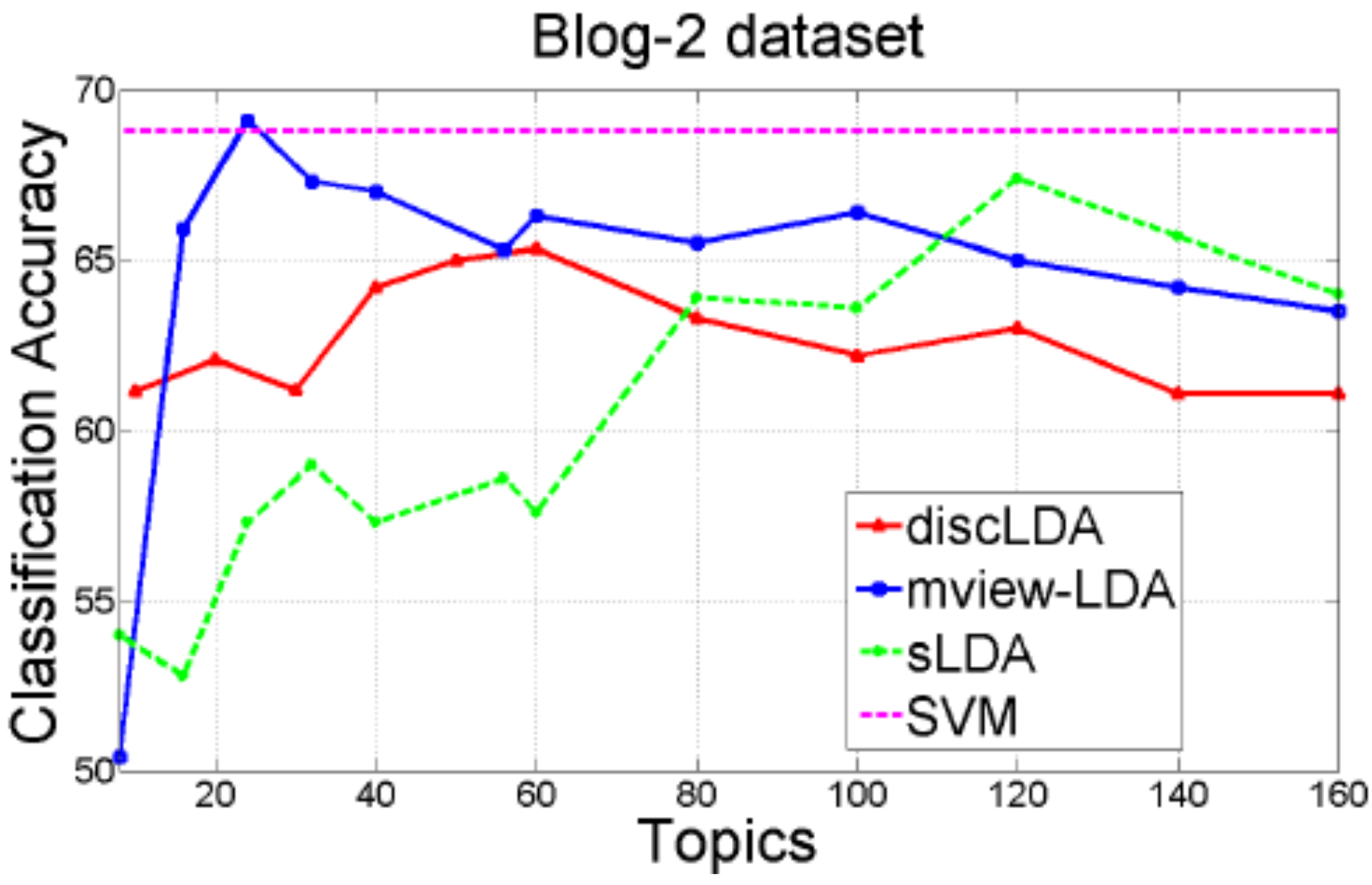
Bitterlemons dataset



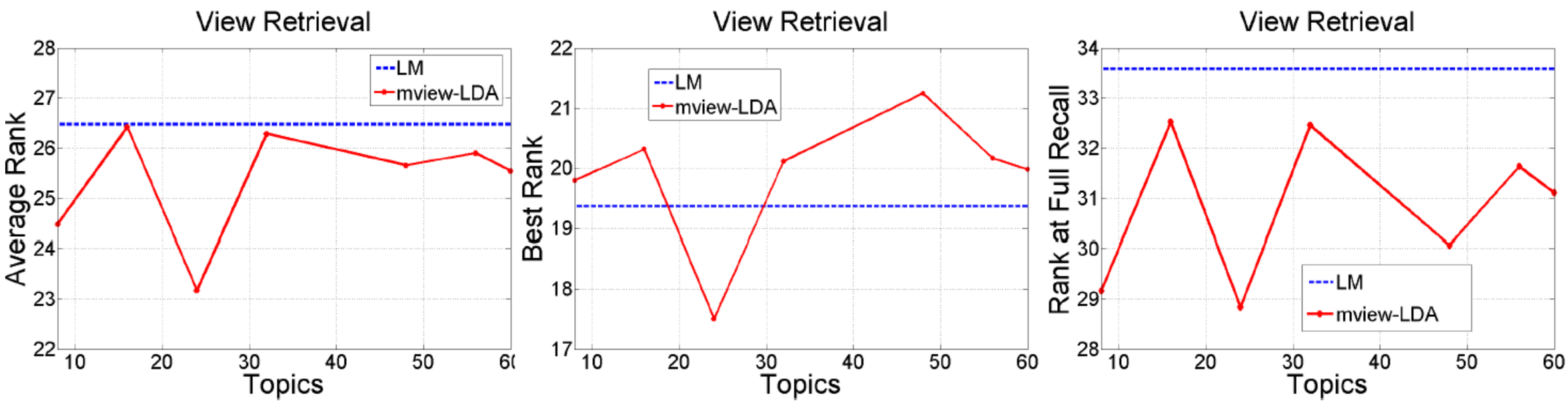
Classification accuracy



Generalization to new blogs

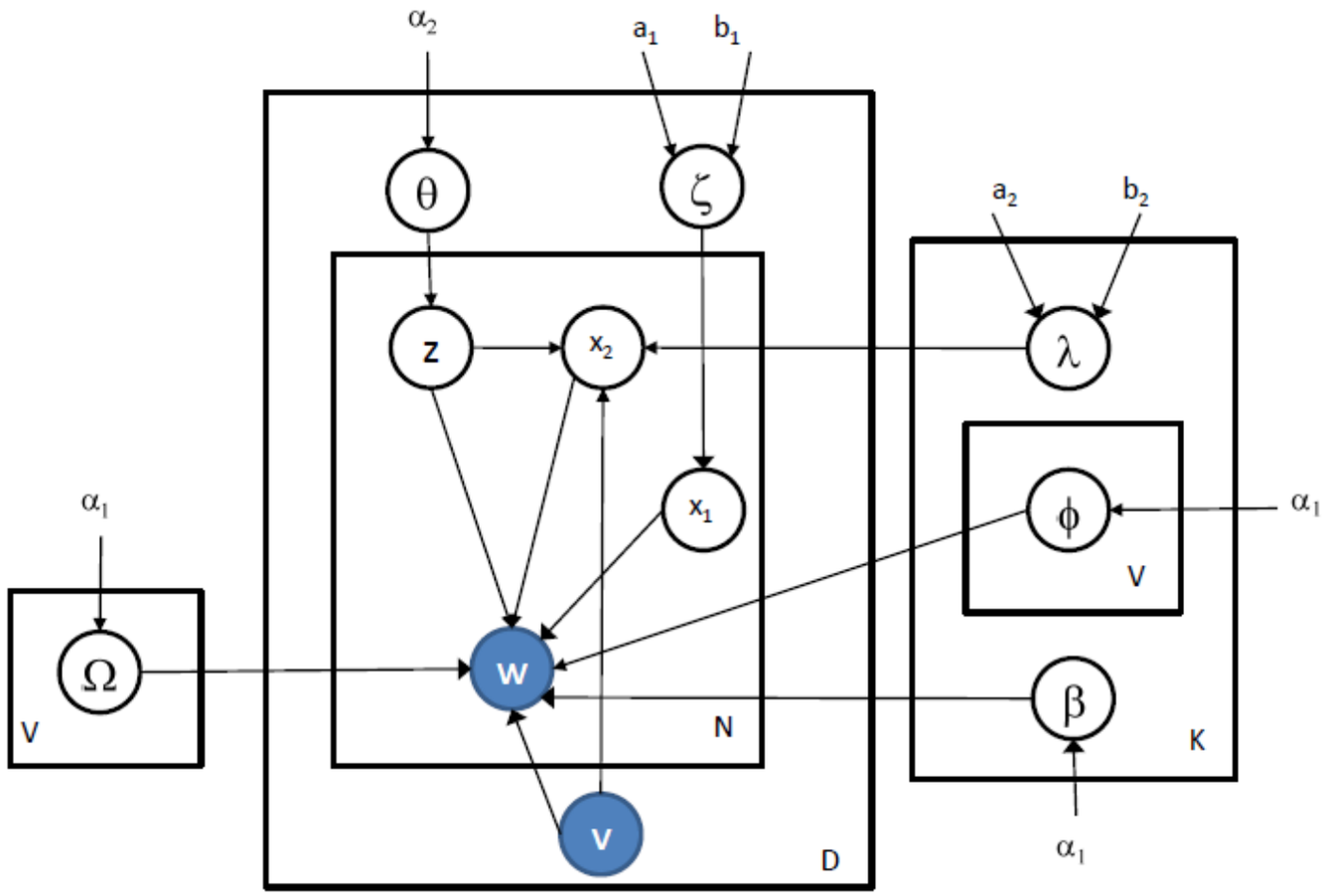


Finding alternate views

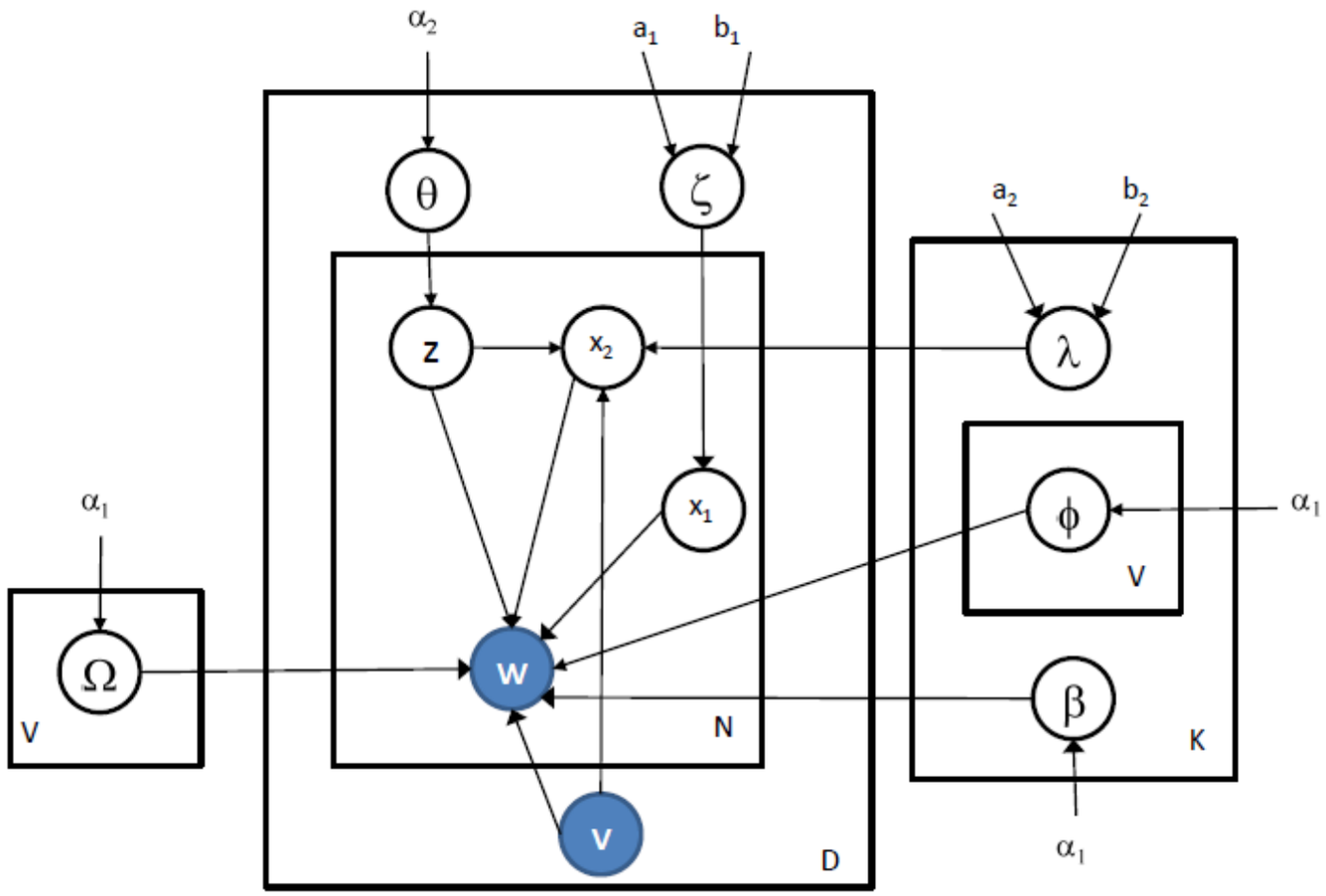


- Given a document written in one ideology, retrieve the equivalent
- Baseline: SVM + cosine similarity

Unlabeled data

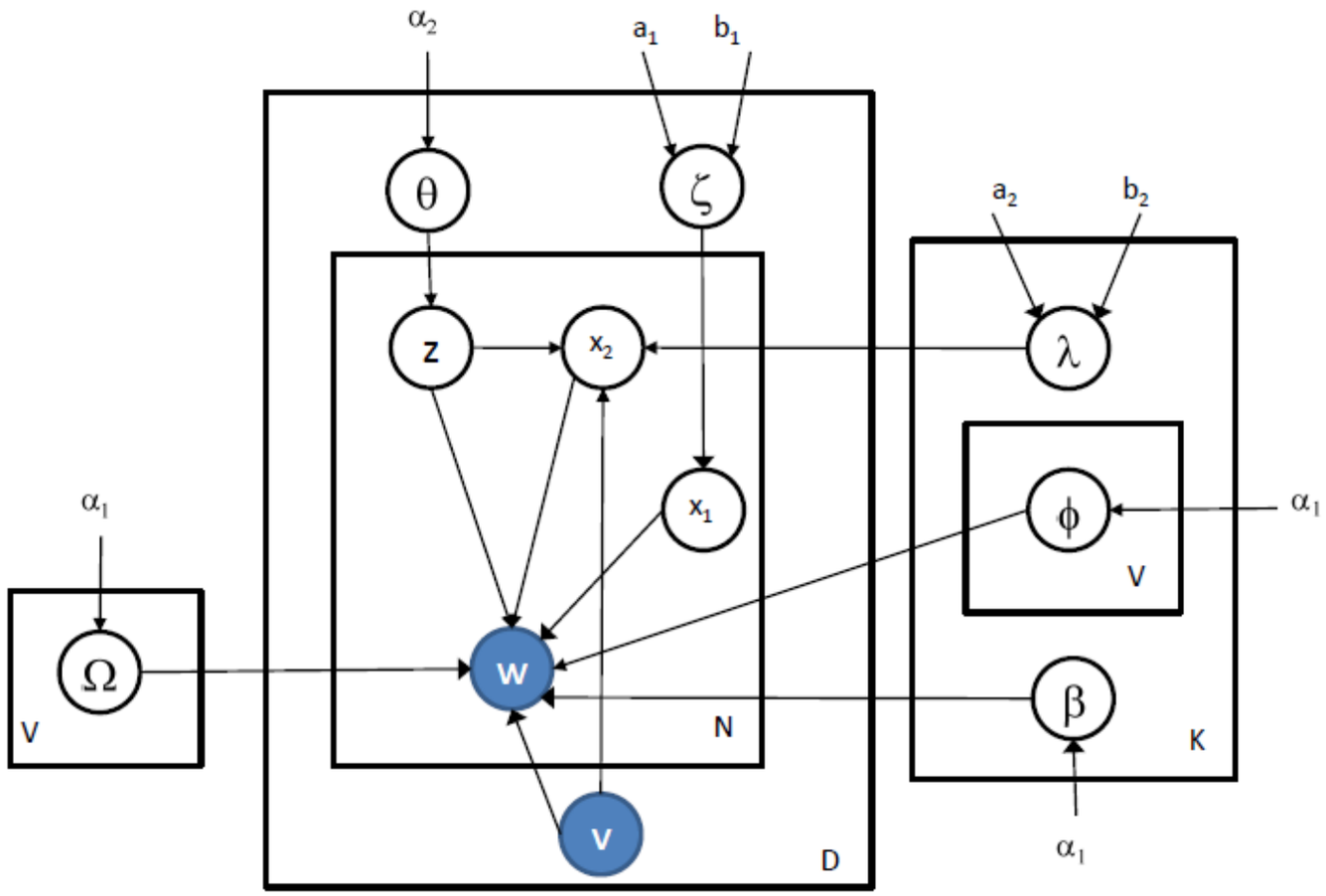


Unlabeled data



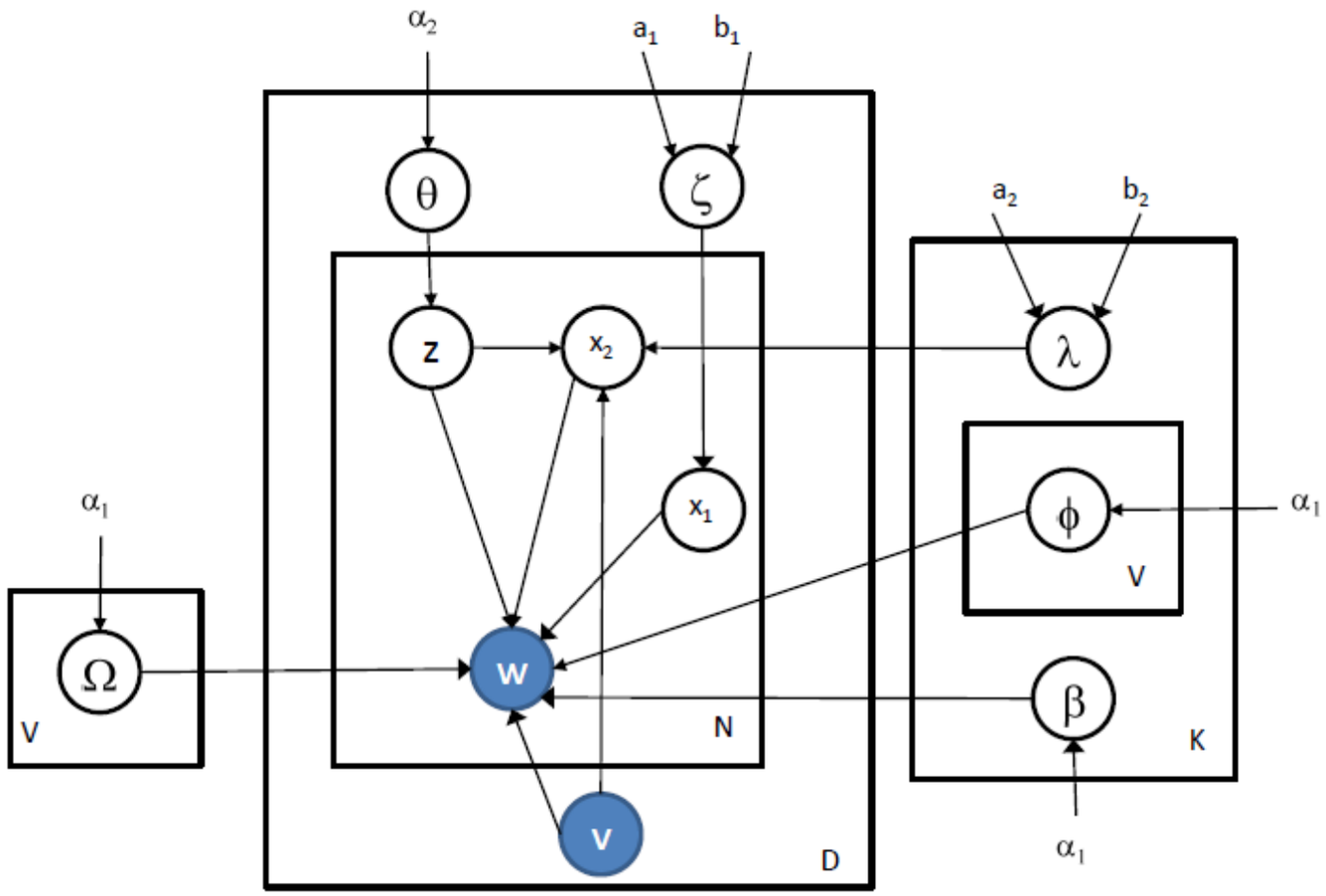
- In theory this is simple
  - Add a step that samples the document view (v)
  - **Doesn't mix** in practice because tight coupling between v and ($\mathbf{x}_1$,$\mathbf{x}_2$,**z**)

Unlabeled data



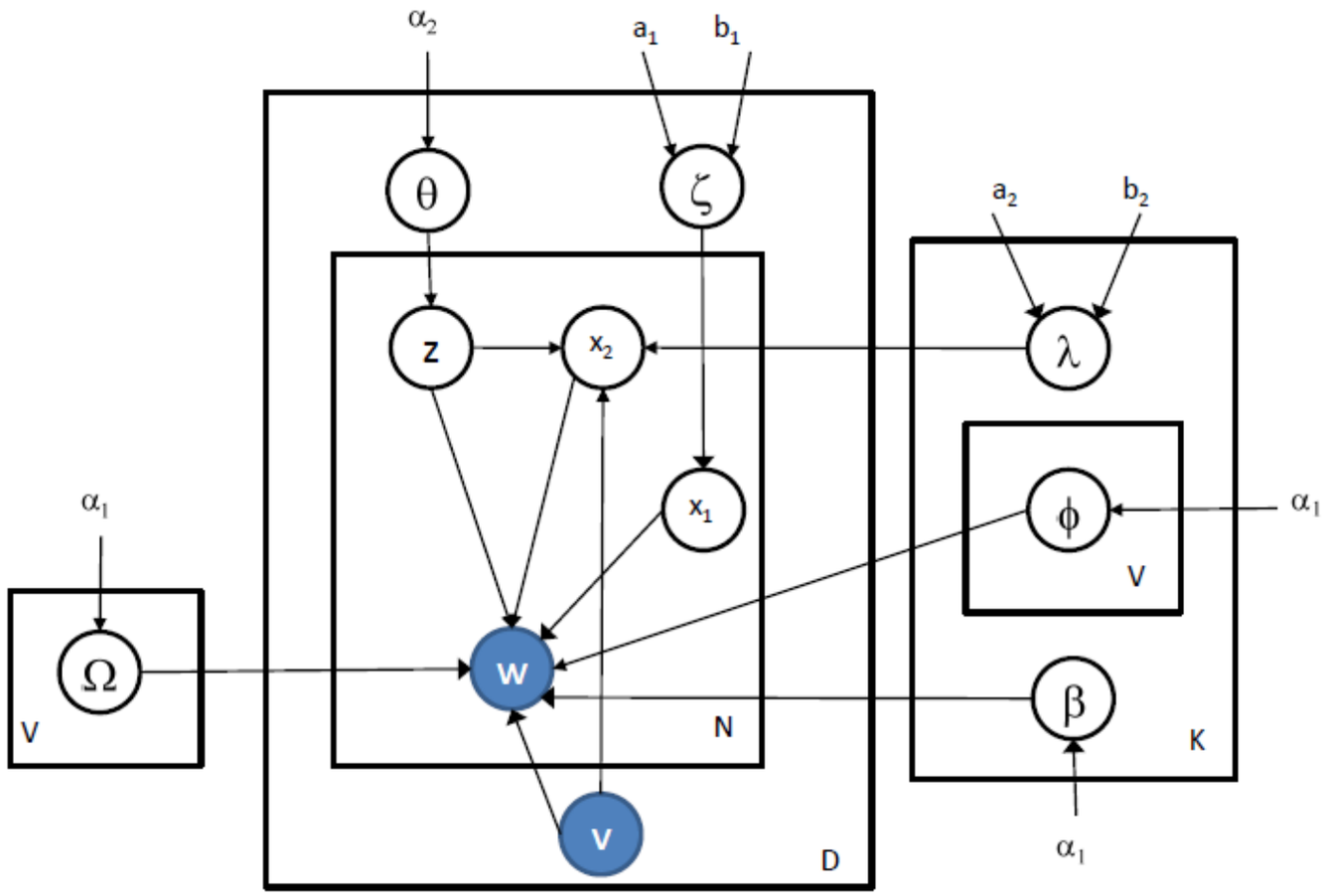
- In theory this is simple
  - Add a step that samples the document view (v)
  - **Doesn't mix** in practice because tight coupling between v and ($\mathbf{x}_1$,$\mathbf{x}_2$,**z**)
- Solution

Unlabeled data



- In theory this is simple
  - Add a step that samples the document view (v)
  - **Doesn't mix** in practice because tight coupling between v and ($\mathbf{x}_1$,$\mathbf{x}_2$,**z**)
- Solution
  - Sample v and ($\mathbf{x}_1$,$\mathbf{x}_2$,**z**) as a block using a Metropolis-Hasting step

Unlabeled data



- In theory this is simple
  - Add a step that samples the document view (v)
  - **Doesn't mix** in practice because tight coupling between v and ($\mathbf{x}_1$,$\mathbf{x}_2$,**z**)
- Solution
  - Sample v and ($\mathbf{x}_1$,$\mathbf{x}_2$,**z**) as a block using a Metropolis-Hasting step
  - This is a **huge proposal**!