Big Data and Large Scale Inference

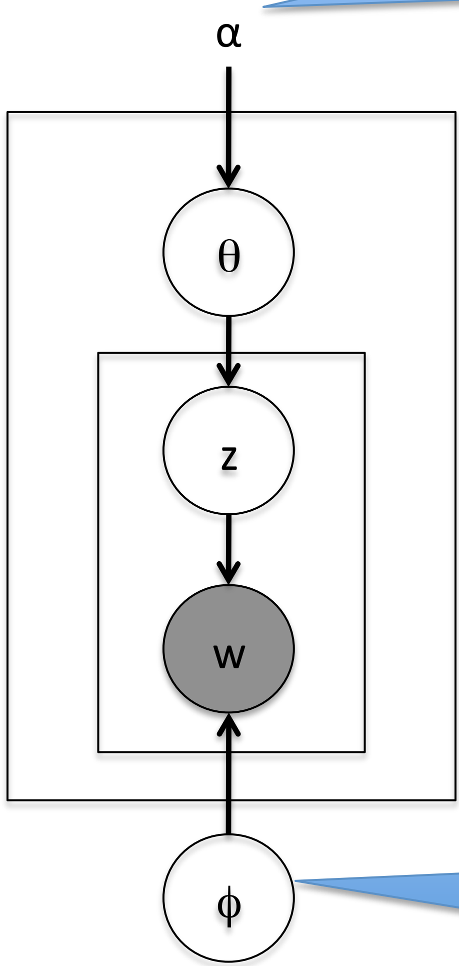
Amr Ahmed & Alex Smola

Research at Google

Wrapping up

- Distributed inference in latent variable models
  - Star Synchronization
  - Delta aggregation

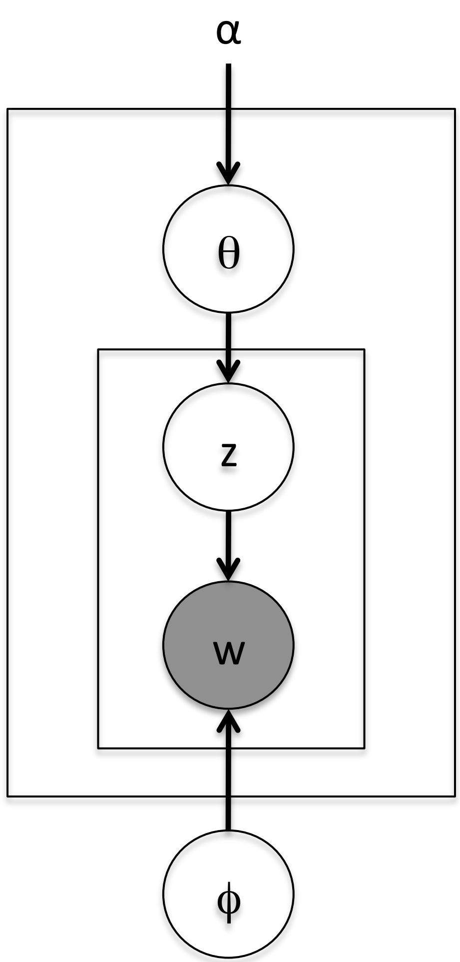
Wrapping up …



Prior over topic mixtures

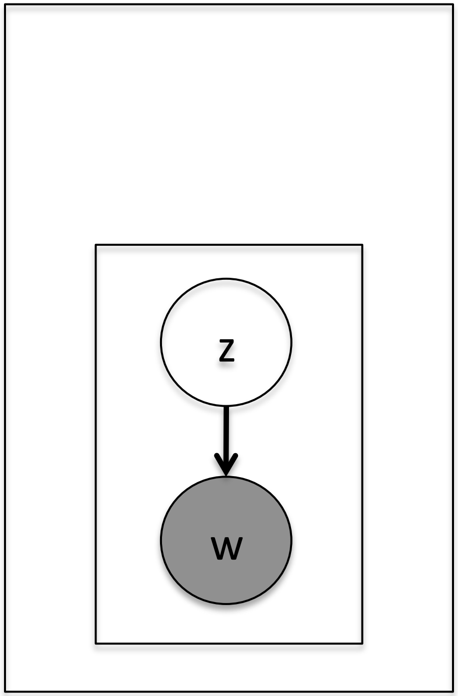
Topic-word distribution

Wrapping up …



- Global variables
  - Φ: Topic distribution over words
- Local variables
  - θ: topic mixing vector
  - Z: topic indicator

Wrapping up …

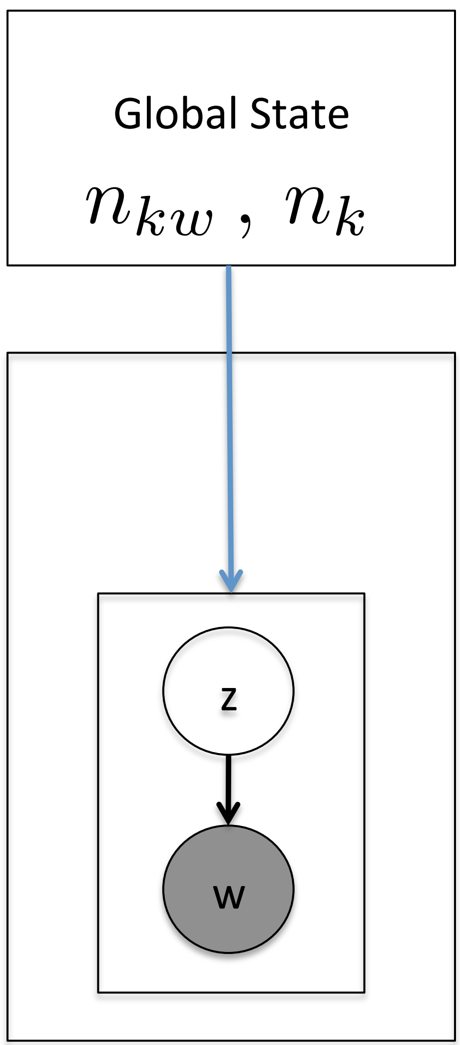


- Collapse global variables
  - Φ
- Collapse local variables
  - θ
- Couples all Zs
- Run collapsed sampler

$$P(z_{di} = k | w_{di} = w, z_{-di}) \propto$$

$$(n_{dk} + \alpha) \frac{n_{kw} + \beta}{n_k + W\beta}$$

Wrapping up …

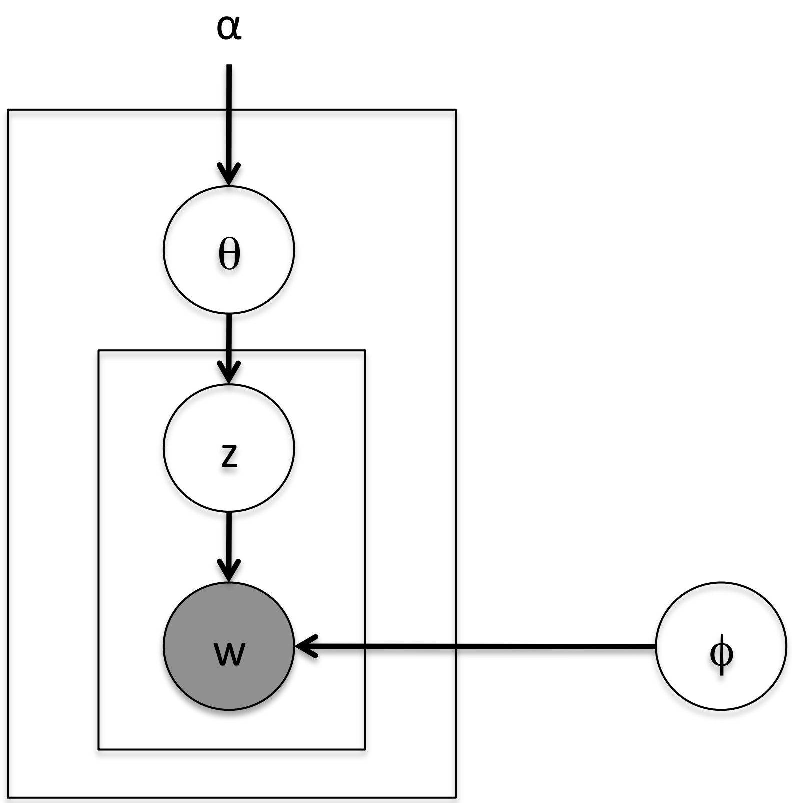


Local counts
(local state)

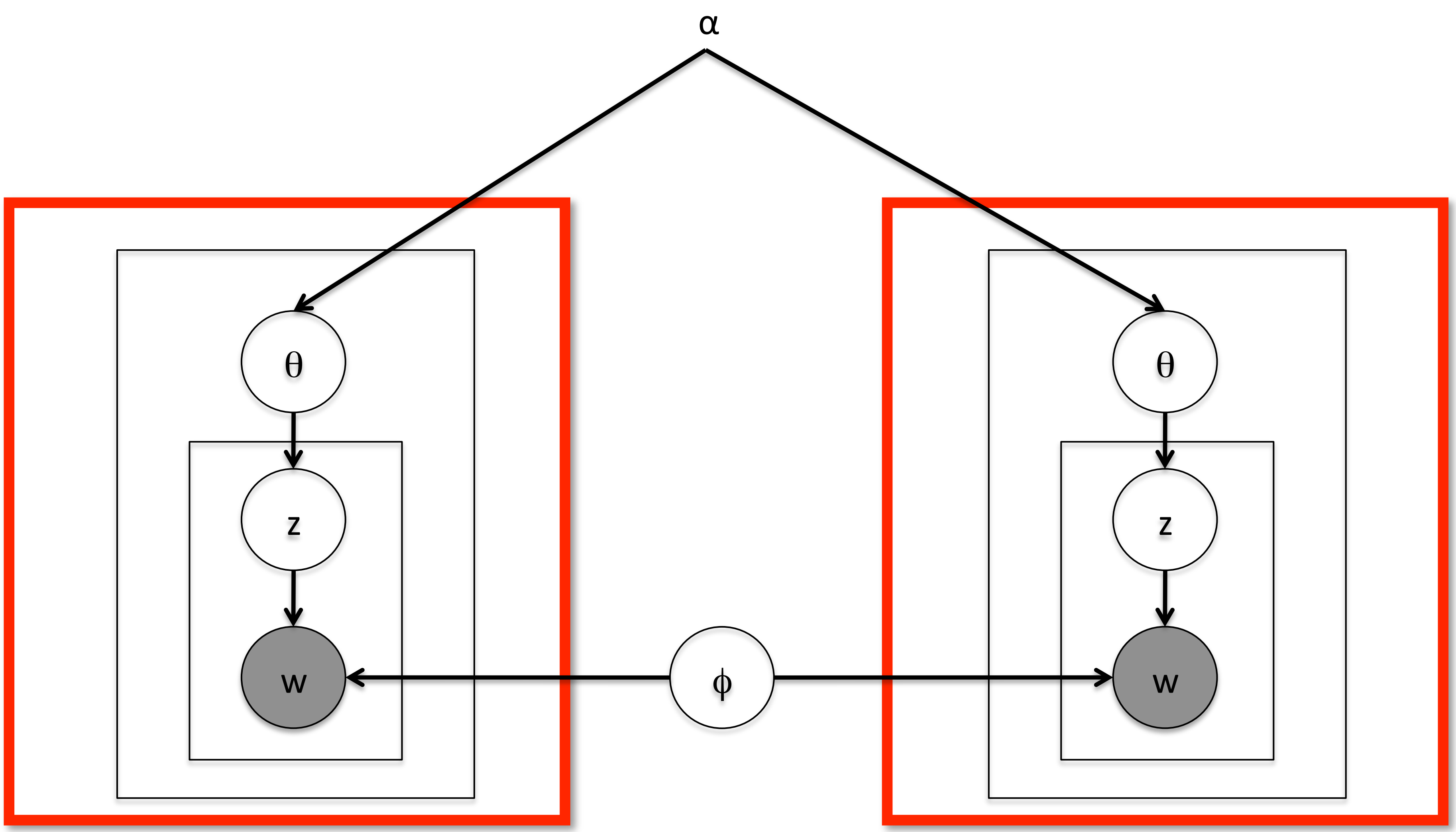
Global counts
(global state)

$$P(z_{di} = k | w_{di} = w, z_{-di}) \propto (n_{dk} + \alpha) \frac{n_{kw} + \beta}{n_k + W\beta}$$

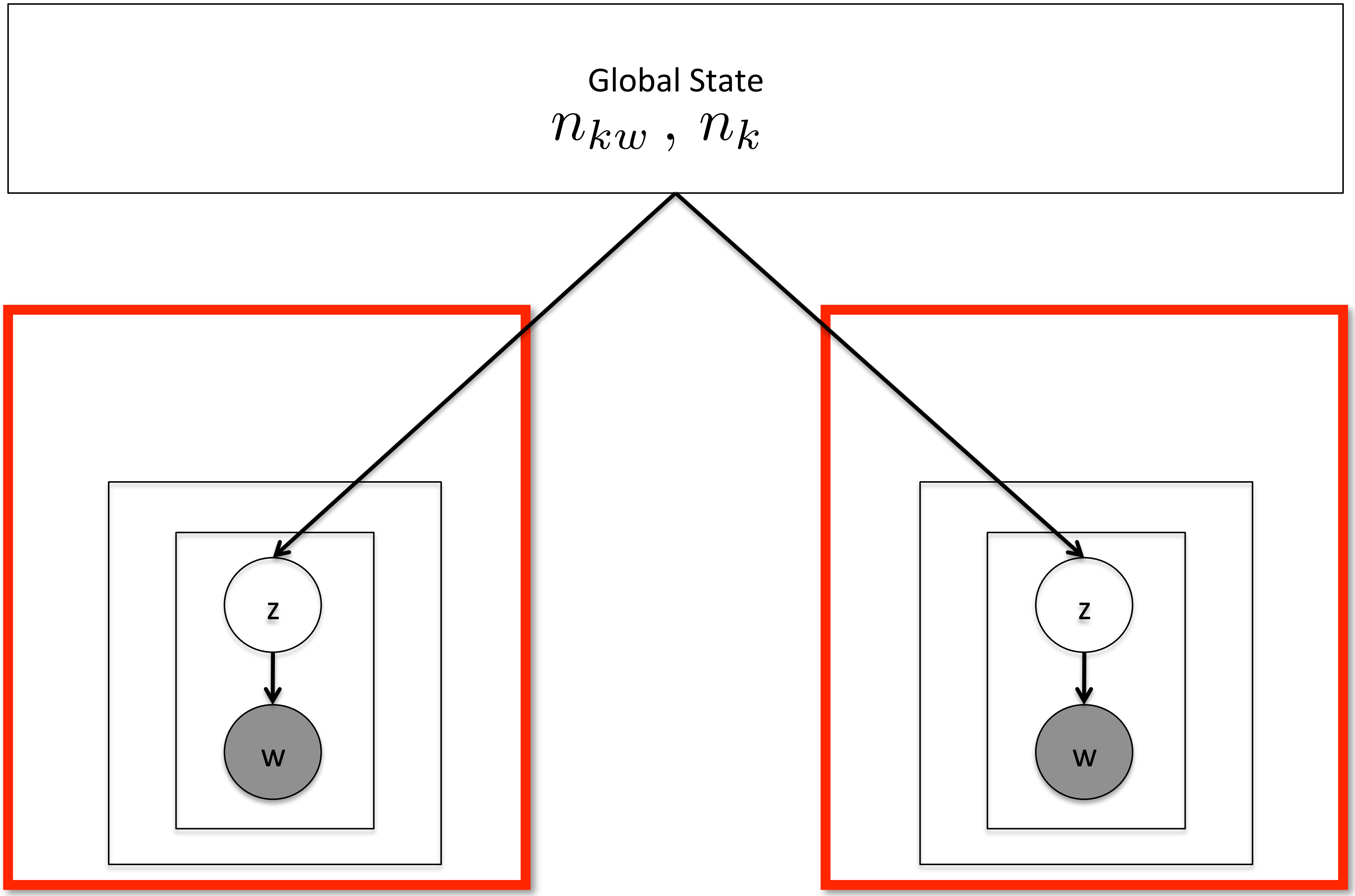
Distributed Inference: LDA



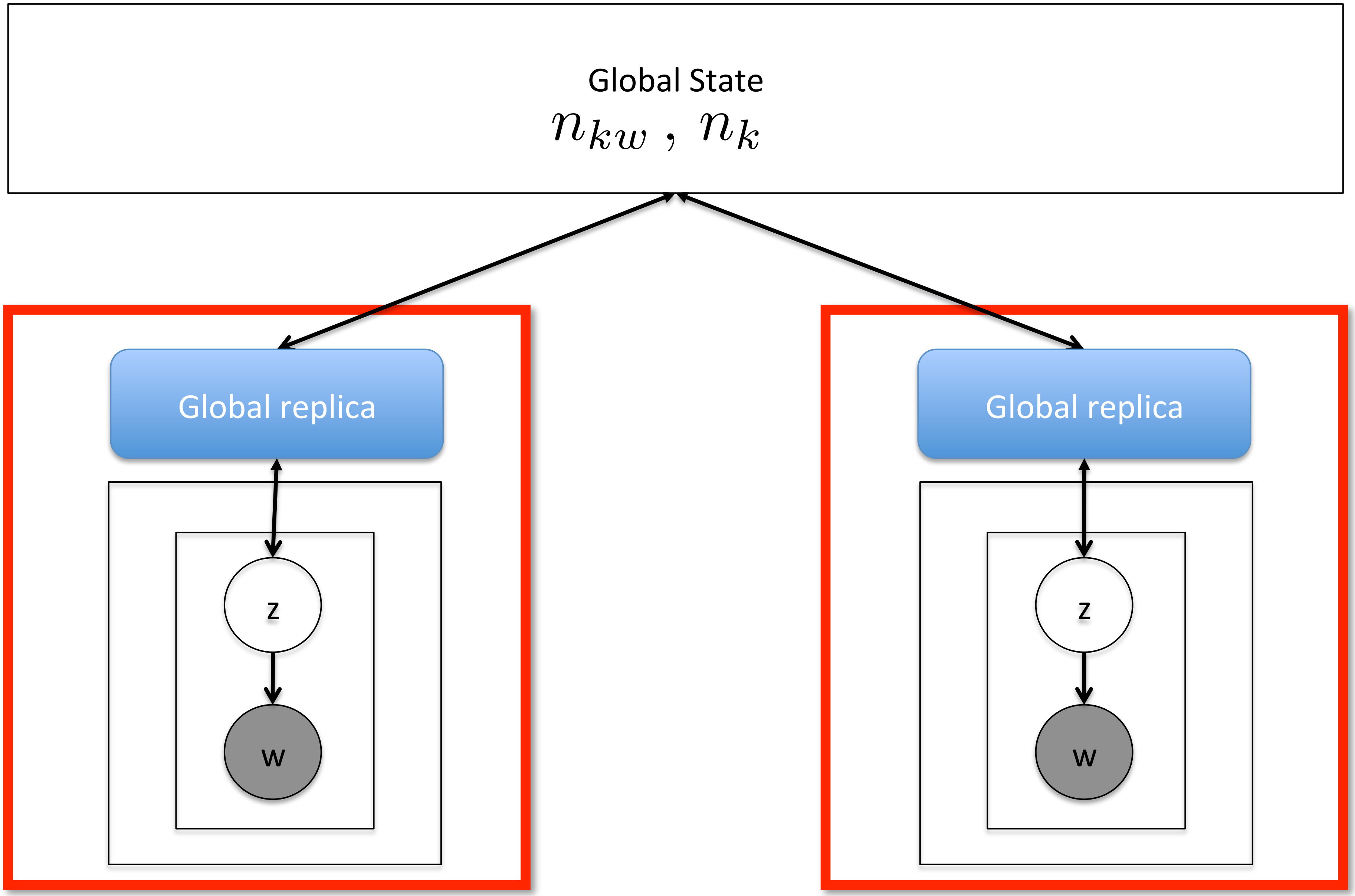
Distributed Inference: LDA



Distributed Inference: LDA

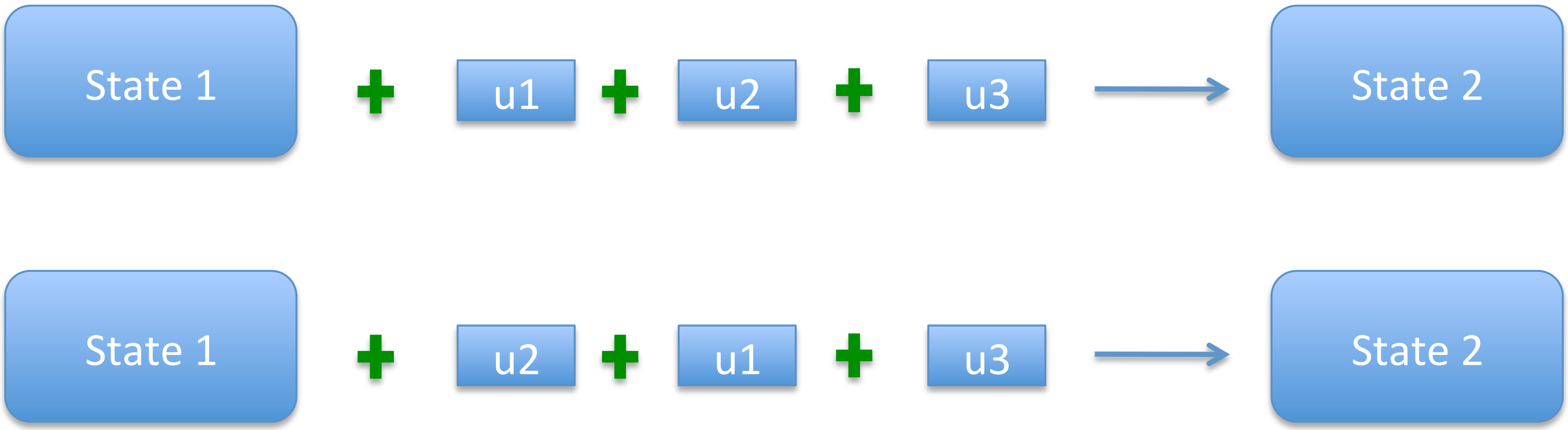


Distributed Inference: LDA



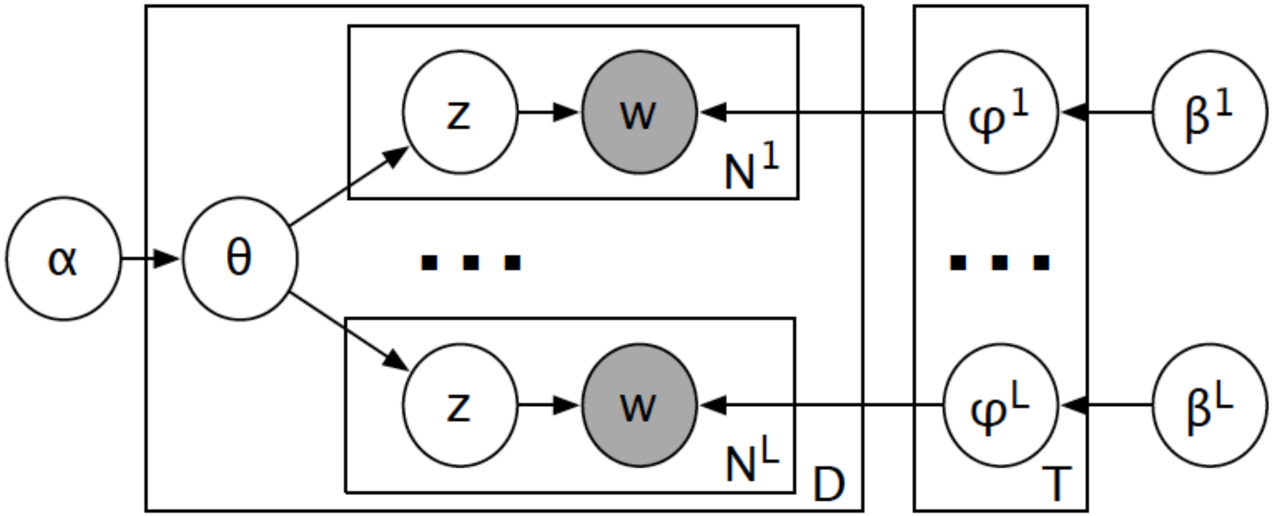
General Architecture

- Star synchronization
  - Works when variables depend on each other via aggregates
    - Counts, sums, etc.
  - When state objects form an Abelian group



Multilingual LDA

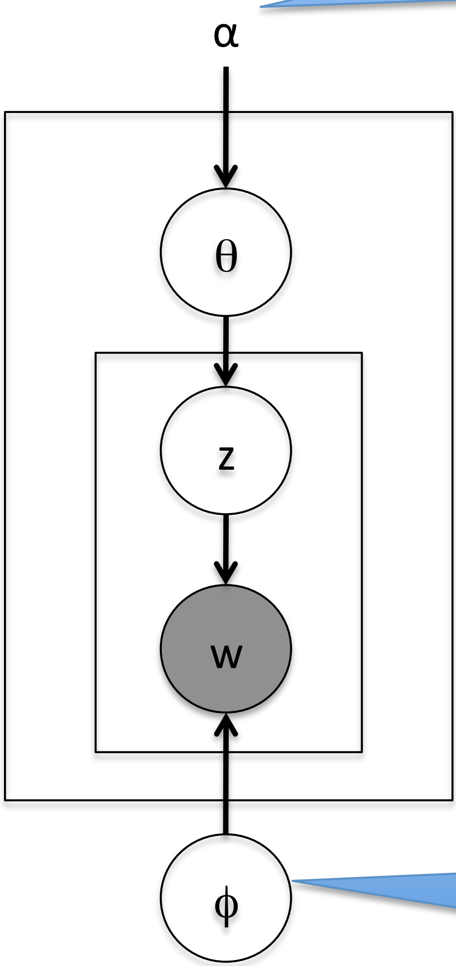
- Each topic has a distribution over words
- Fits parallel documents
  - Example: Wikipedia



What Is next?

- Can we fit any model only with those asynchronous primitives?
  - No
- We need synchronous operations
  - Parameter optimization
    - EM style algorithm
  - Non-collapsed global variables

The Need for Synchronous Processing

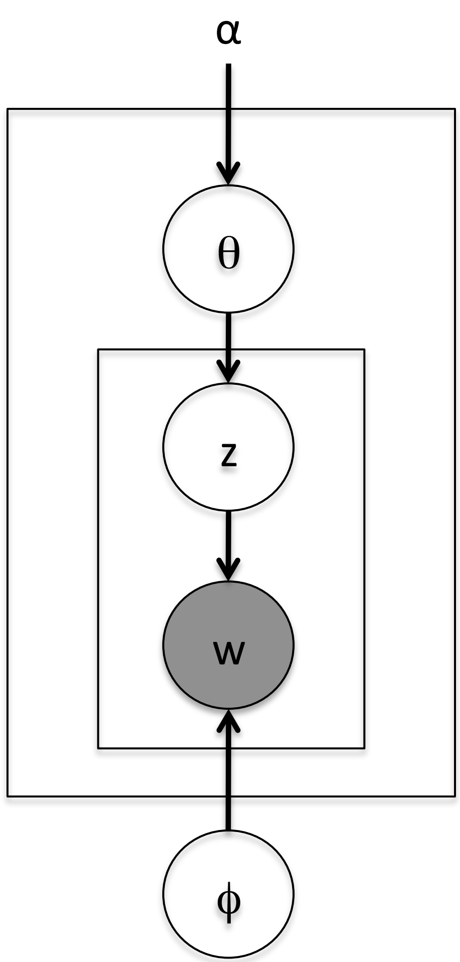


Prior over topic mixtures

What if we need to optimize over α?

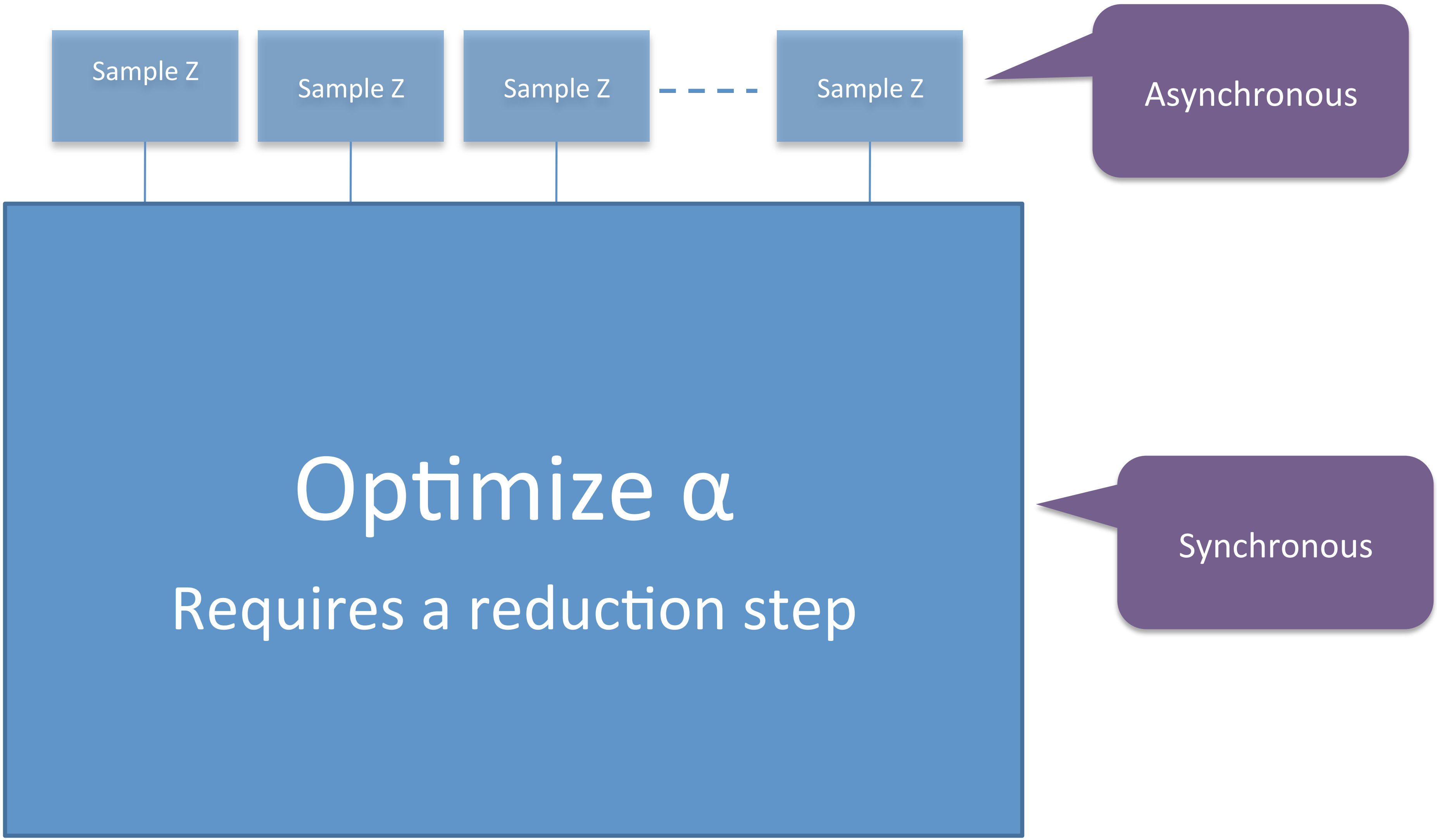
Topic-word distribution

The Need for Synchronous Processing

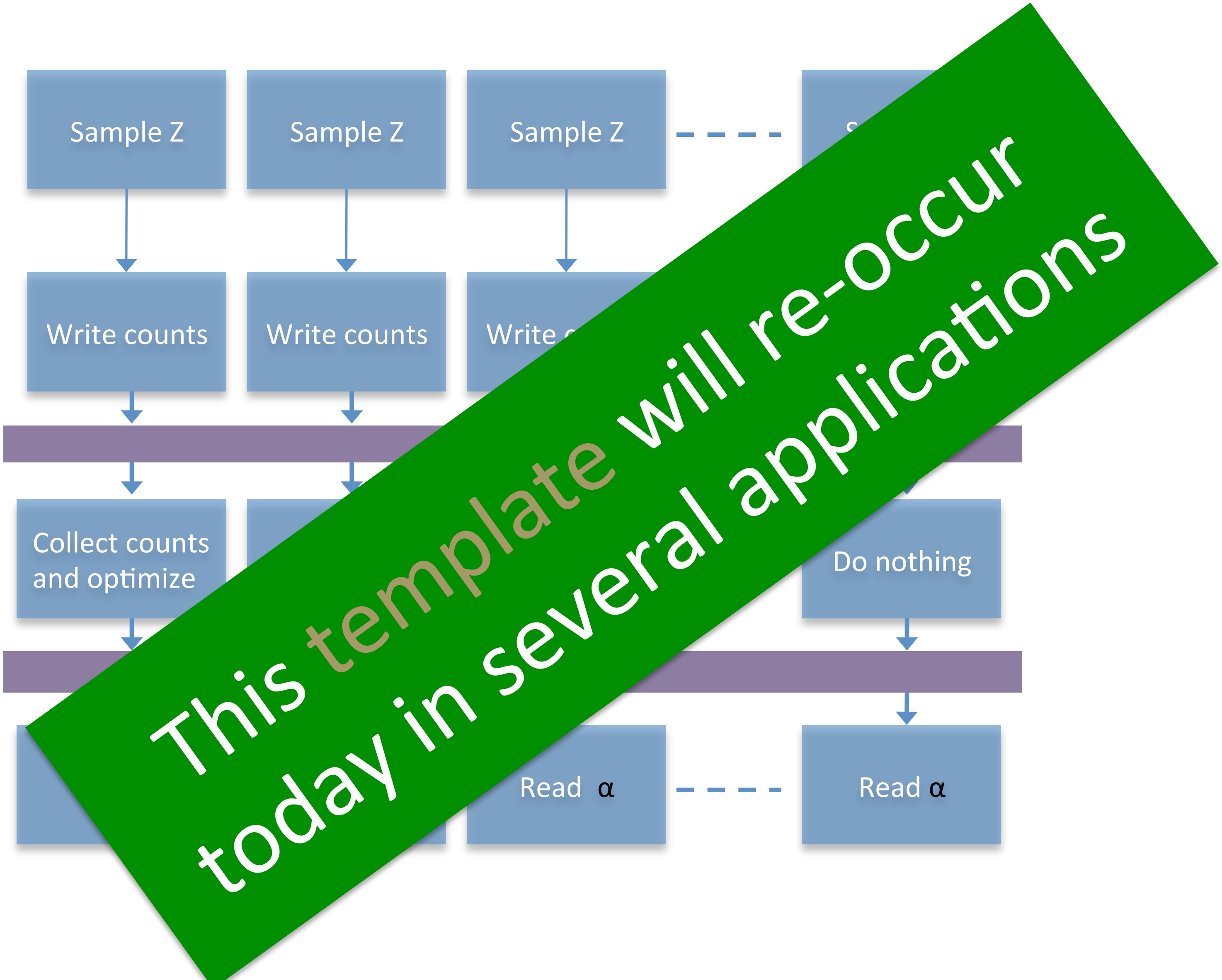


- E-Step
  - Run asynchronous collapsed sampler as before
- M-step
  - Reach a barrier
  - Collect values needed to optimize α
  - One machine optimizes α
  - Broadcast value back

Distributed Sampling Cycle



Distributed Sampling Cycle



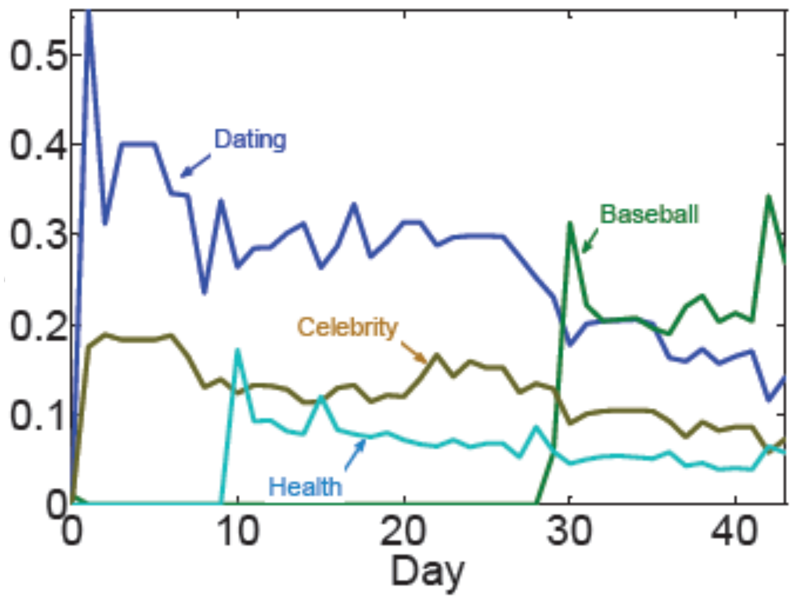
Up next

- Application
  - Temporal Modeling of user interests
  - Multi-domain user personalization
  - Graph factorization
  - Multi-task learning

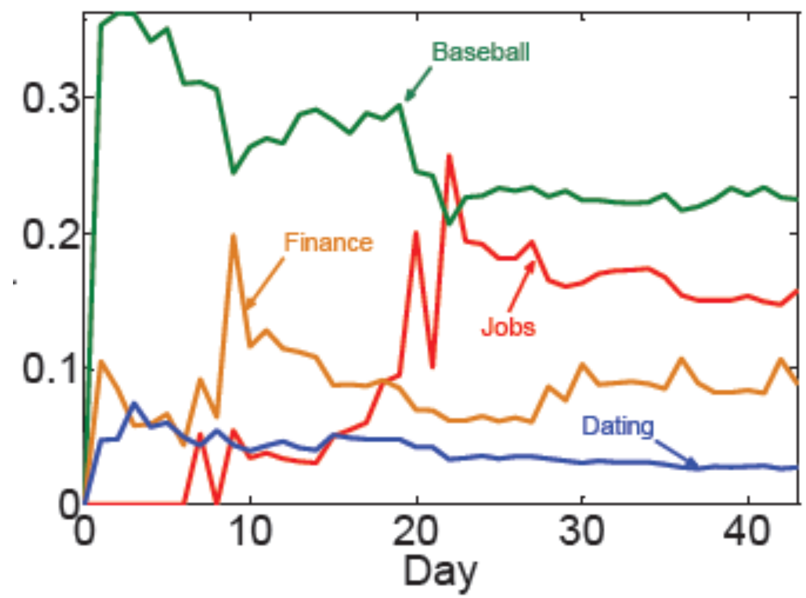
- Asynchronous Distributed Optimization
  - Can we get rid of the synchronous step?
  - Asynchronous consensus
  - Factorizing Y!M graph
    - 200 Million users and 10 Billion edges
    - The largest published work on graph factorization

Modeling User Interests

User-1



User-2



Dating

women
men
dating
singles
personals
seeking
match

Baseball

League
baseball
basketball,
doublehead
Bergesen
Griffey
bullpen
Greinke

Celebrity

Snooki
Tom
Cruise
Katie
Holmes
Pinkett
Kudrow
Hollywood

Health

skin
body
fingers
cells
toes
wrinkle
layers

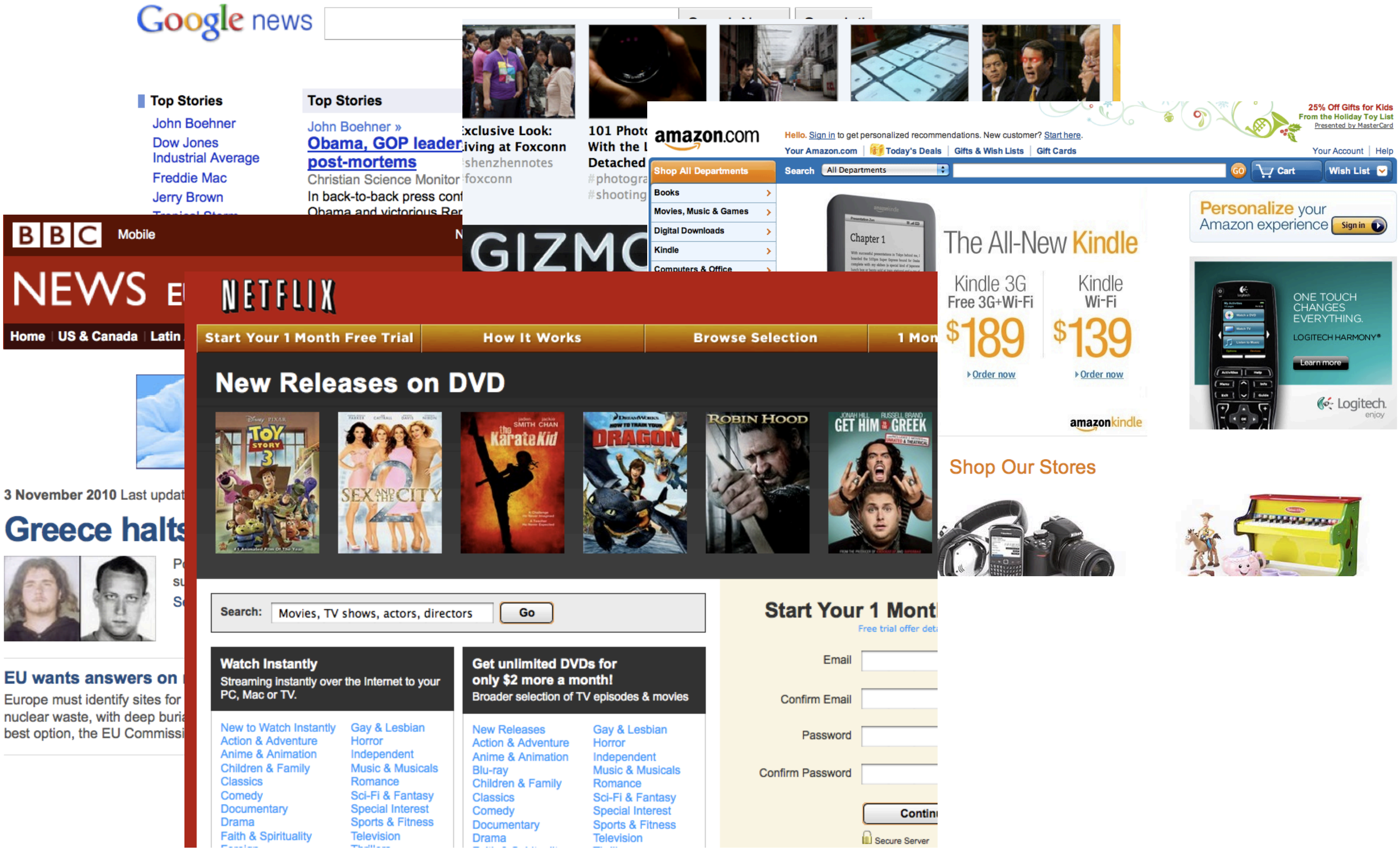
Jobs

job
career
business
assistant
hiring
part-time
receptionist

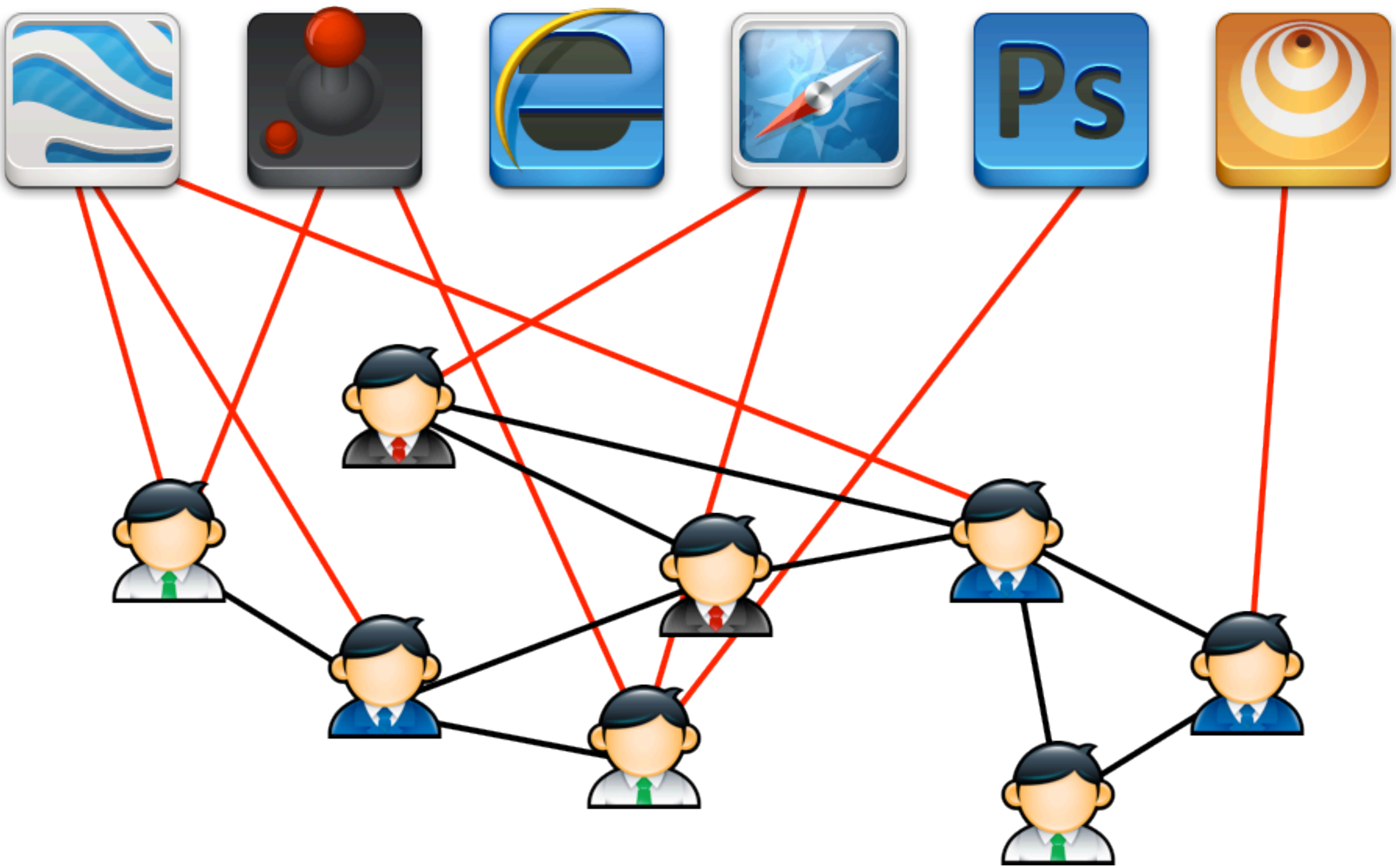
Finance

financial
Thomson
chart
real
Stock
Trading
currency

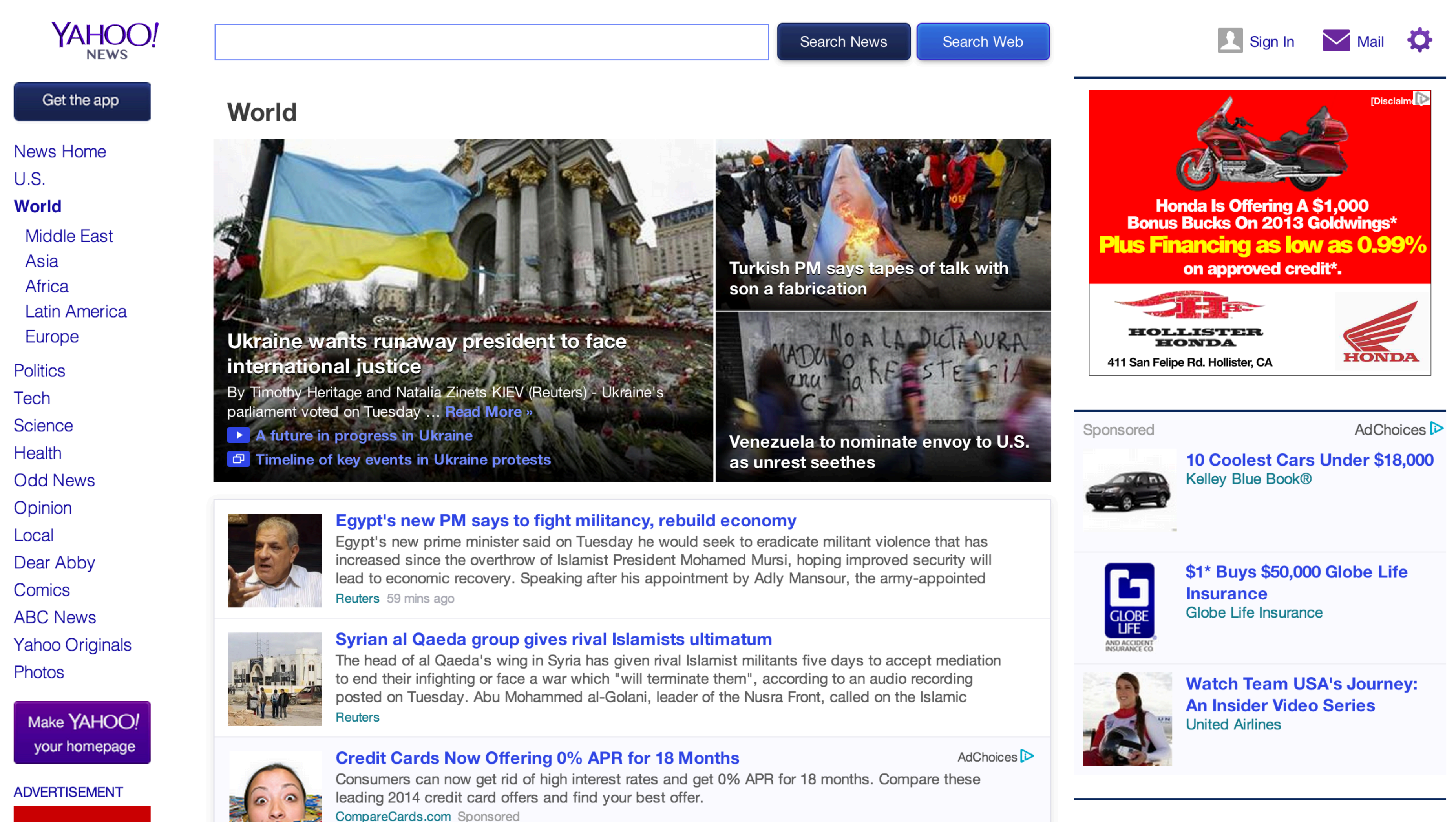
Multi-domain Personalization



Graph Factorization: Social Network

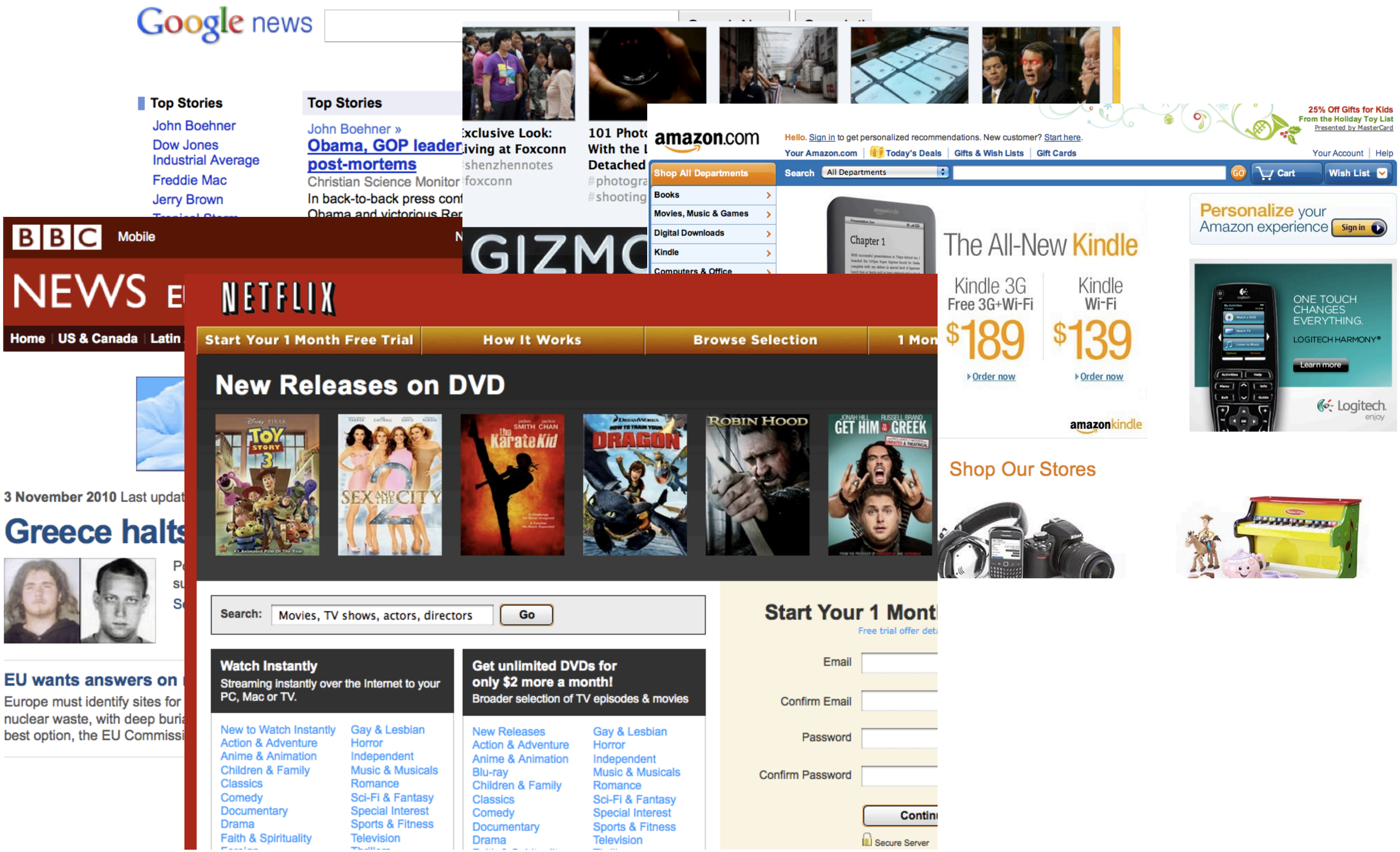


Computational Advertising: Multitask learning



Multi-Domain Personalization

Problem



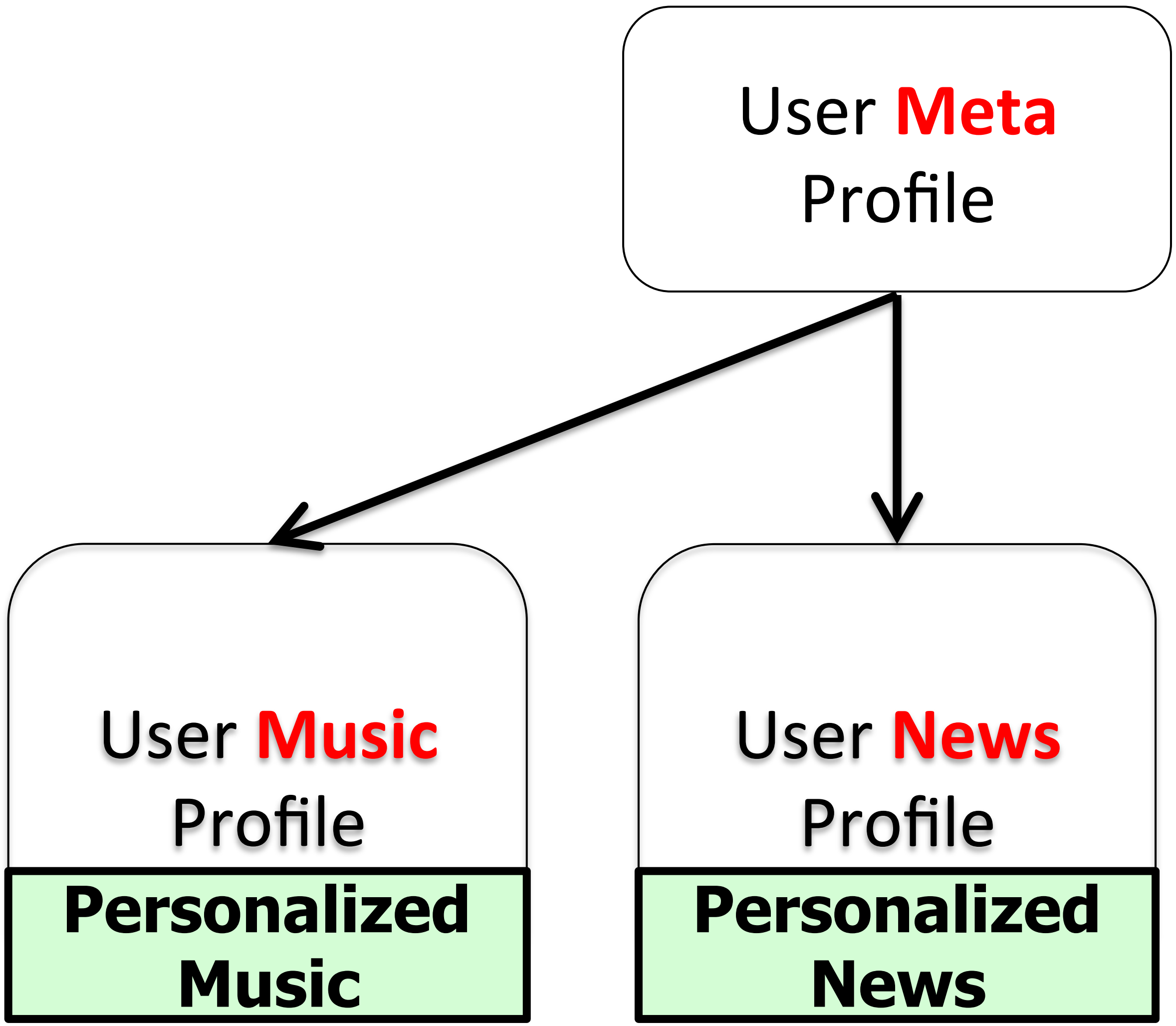
Google news
Top Stories
John Boehner
Dow Jones Industrial Average
Freddie Mac
Jerry Brown
Top Stories
John Boehner »
Obama, GOP leader
post-mortems
Christian Science Monitor
In back-to-back press conf
xclusive Look:
iving at Foxconn
shenzhennotes
foxconn
101 Photo
With the
Detached
photogra
shooting
GIZMO
BBC Mobile
NEWS
Home US & Canada Latin
3 November 2010 Last updat
Greece halts
EU wants answers on
Europe must identify sites for
nuclear waste, with deep buri
best option, the EU Commissi
amazon.com
Hello. Sign in to get personalized recommendations. New customer? Start here.
Your Amazon.com
Today's Deals
Gifts & Wish Lists
Gift Cards
Your Account
Help
25% Off Gifts for Kids
From the Holiday Toy List
Presented by MasterCard
Shop All Departments
Search
All Departments
GO
Cart
Wish List
Books
Movies, Music & Games
Digital Downloads
Kindle
Computers & Office
The All-New Kindle
Kindle 3G
Free 3G+Wi-Fi
$189
Order now
Kindle
Wi-Fi
$139
Order now
amazonkindle
Personalize your Amazon experience
Sign in
ONE TOUCH CHANGES EVERYTHING.
LOGITECH HARMONY®
Learn more
Logitech
Shop Our Stores
NETFLIX
Start Your 1 Month Free Trial
How It Works
Browse Selection
New Releases on DVD
Search:
Movies, TV shows, actors, directors
Go
Watch Instantly
Streaming instantly over the Internet to your PC, Mac or TV.
New to Watch Instantly
Action & Adventure
Anime & Animation
Children & Family
Classics
Comedy
Documentary
Drama
Faith & Spirituality
Gay & Lesbian
Horror
Independent
Music & Musicals
Romance
Sci-Fi & Fantasy
Special Interest
Sports & Fitness
Television
Get unlimited DVDs for only $2 more a month!
Broader selection of TV episodes & movies
New Releases
Action & Adventure
Anime & Animation
Blu-ray
Children & Family
Classics
Comedy
Documentary
Drama
Gay & Lesbian
Horror
Independent
Music & Musicals
Romance
Sci-Fi & Fantasy
Special Interest
Sports & Fitness
Television
Start Your 1 Mont
Free trial offer deta
Email
Confirm Email
Password
Confirm Password
Contin
Secure Server

Multi-domain Personalization

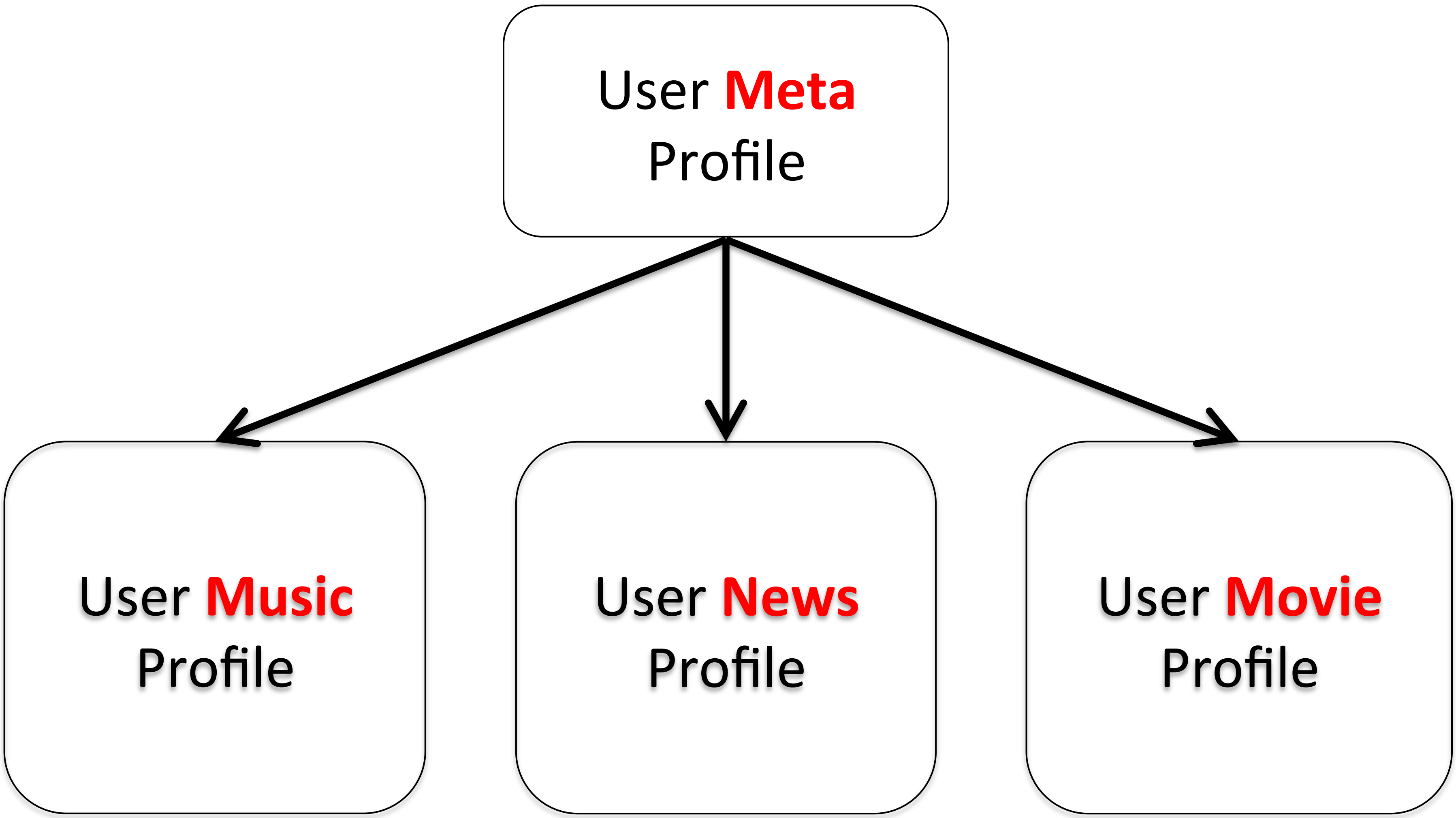
- Intuition
  - We observe user interaction with news and movies
  - Can we predict his music taste?

- Interaction definition
  - A bag of words describing objects user interacts with in a given domain

Example


User Meta Profile
User Music Profile
Personalized Music
User News Profile
Personalized News

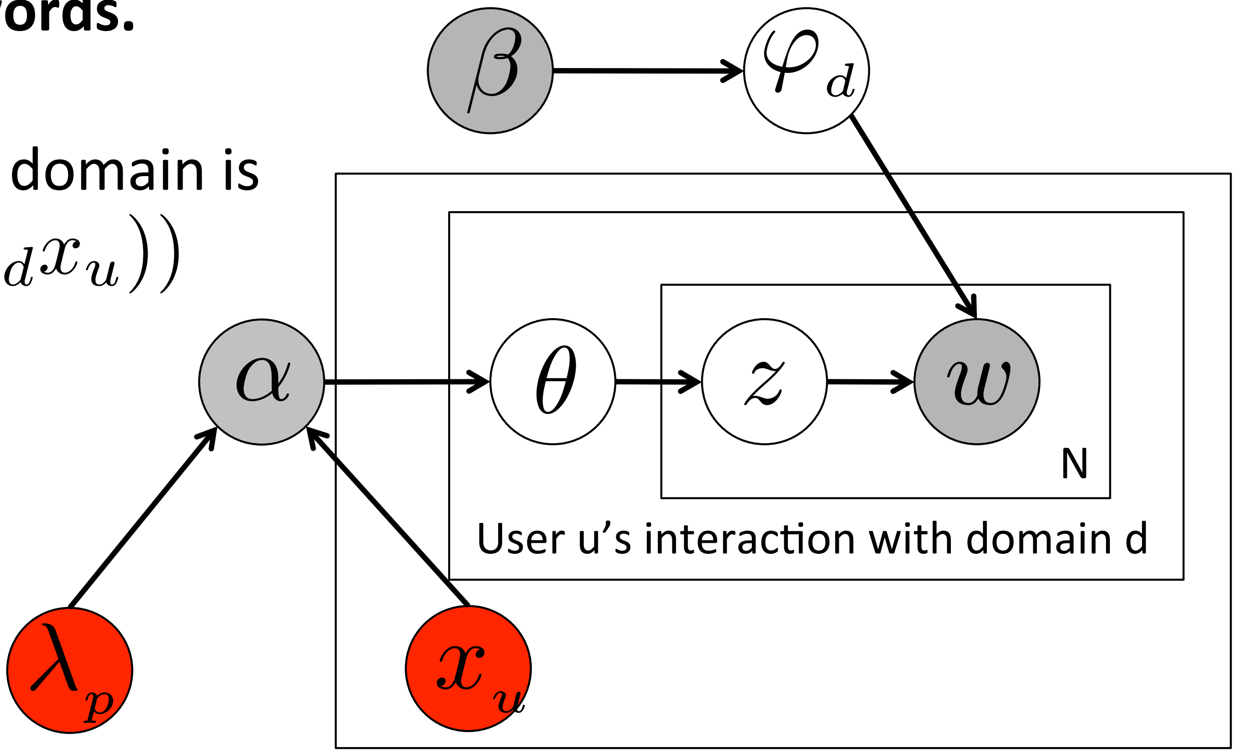
Example


User Meta Profile
User Music Profile
User News Profile
User Movie Profile

The Model

A **user's interaction with a domain** is a **bag of words.**
A **topic** is a **mixture of words.**

User's **prior** interest in a domain is
$$\alpha = \log(1 + \exp(\lambda_d x_u))$$



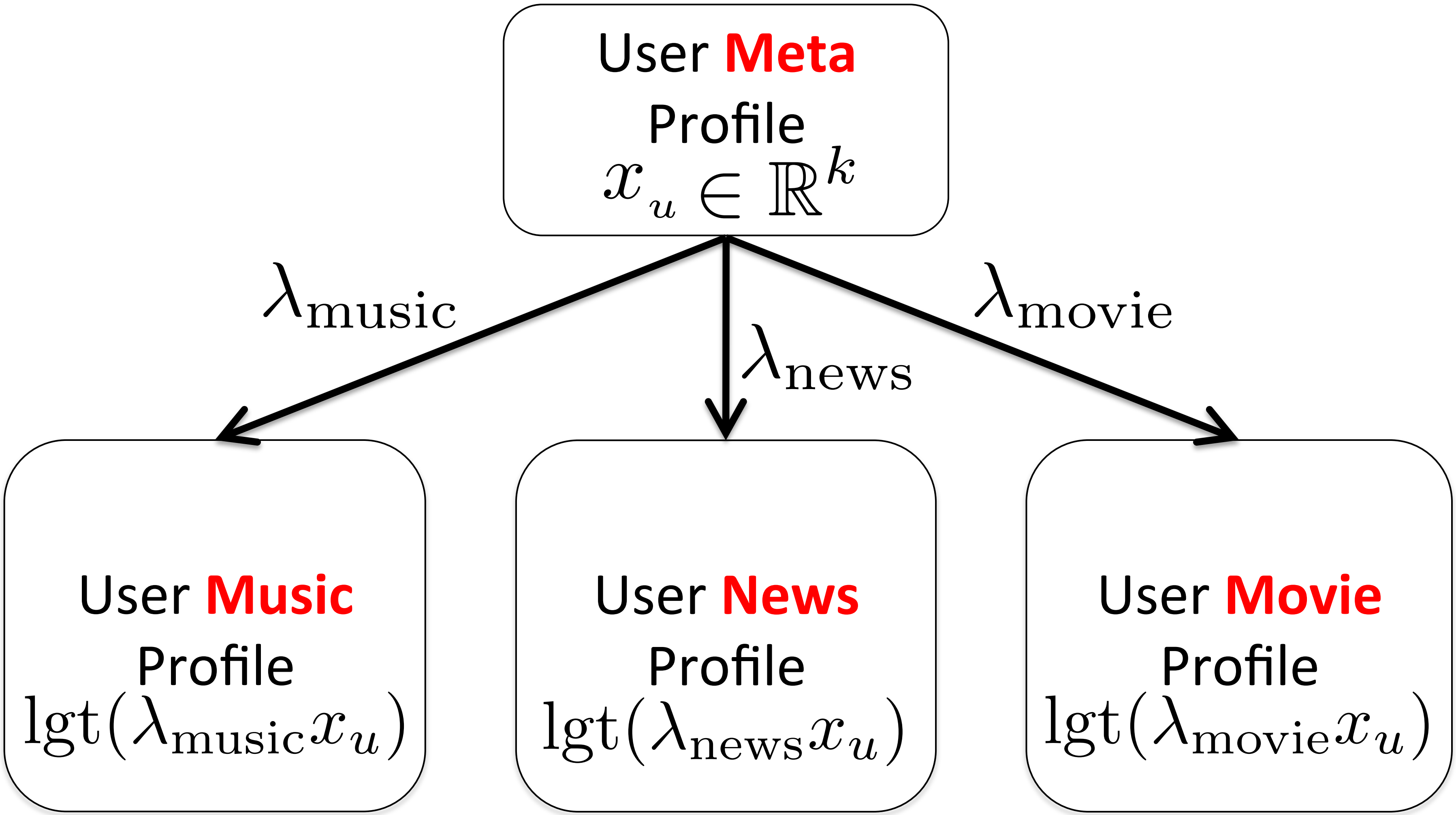
Each user has a meta-profile: $x_u \in \mathbb{R}^k$

Each domain has a latent matrix: $\lambda_d \in \mathbb{R}^{k \times t_d}$

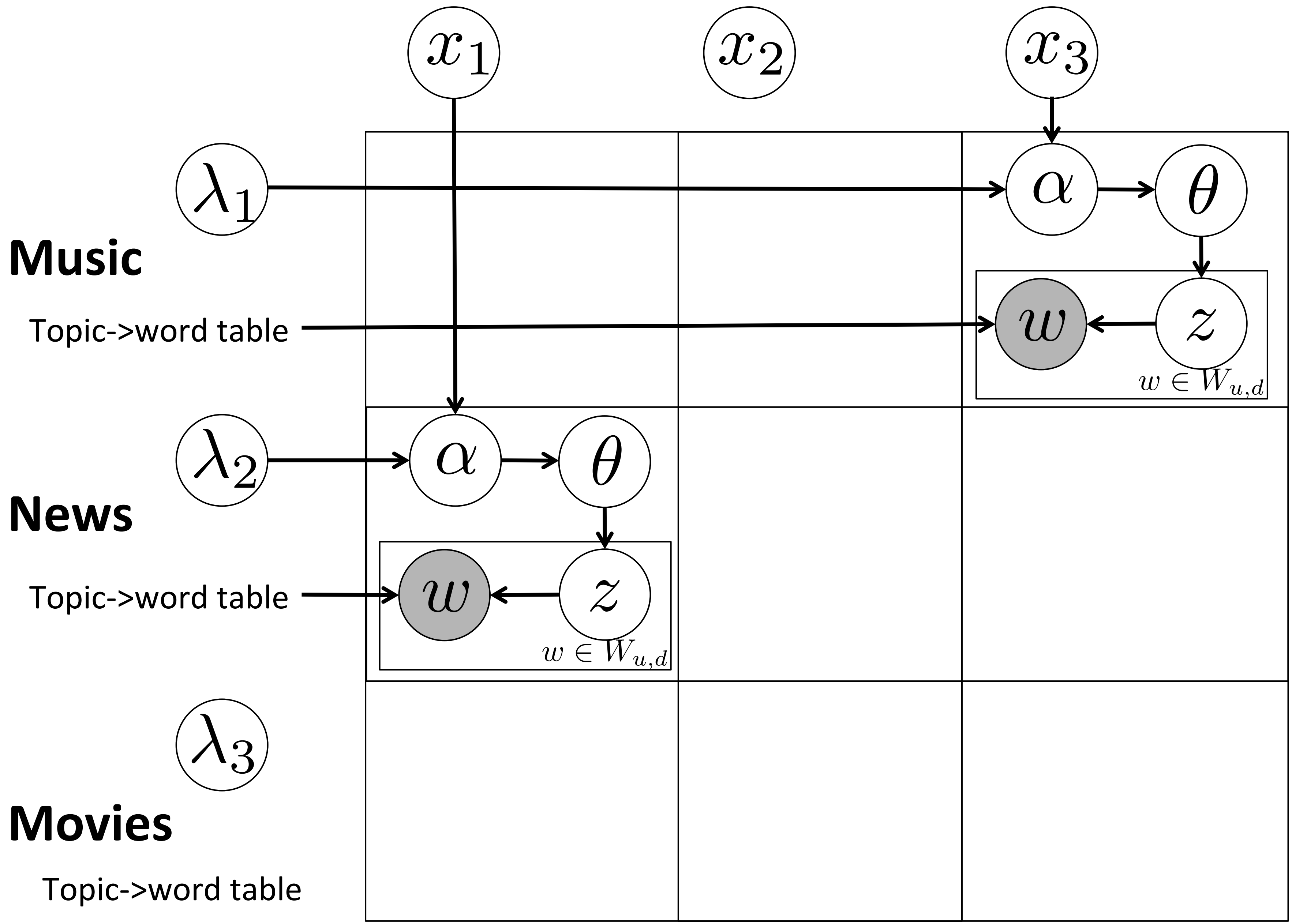
Slide credit Yucheng Low

The Model

$$\mathrm{lgt}(x) = \log(1 + \exp(x))$$

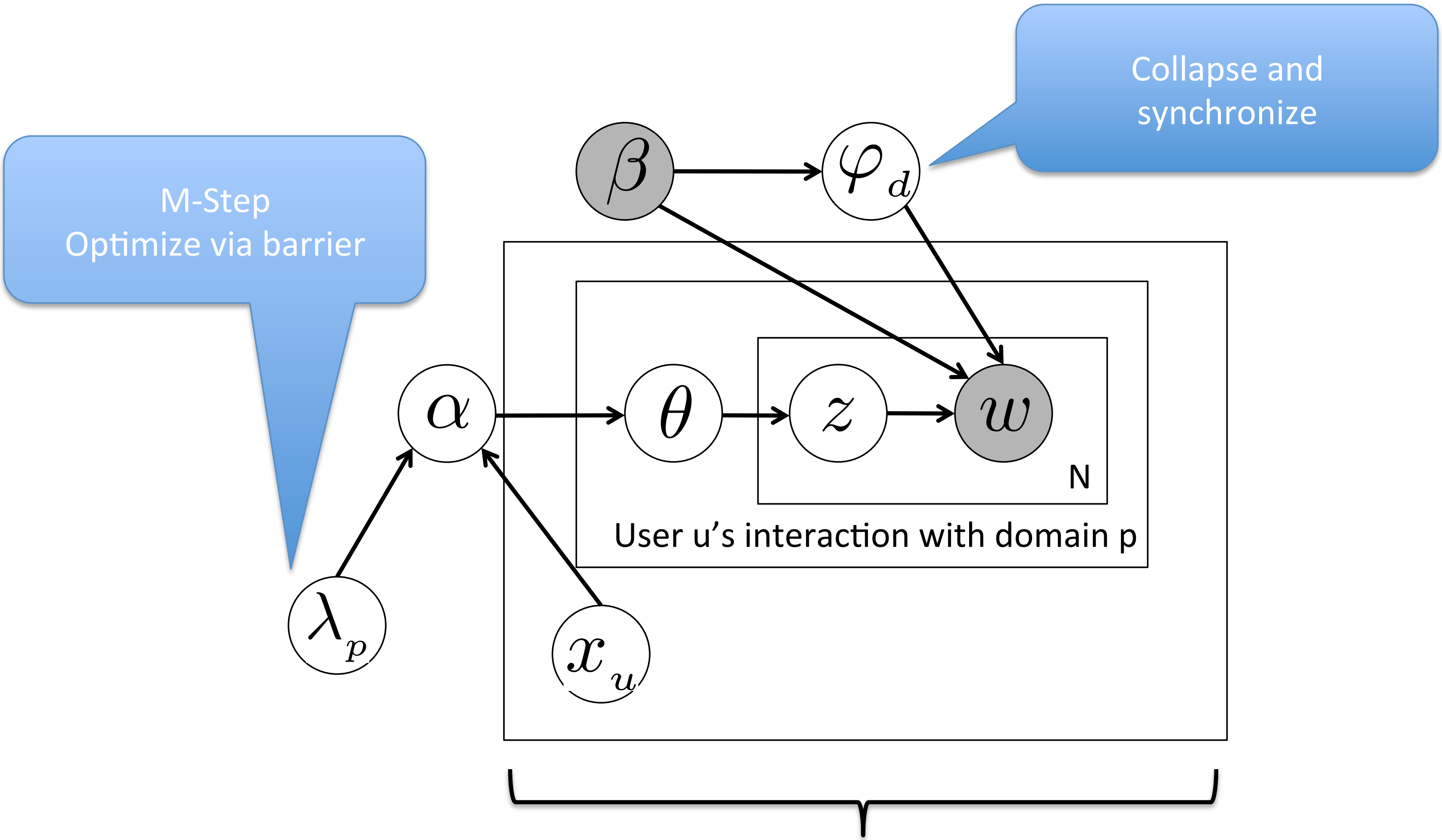


Slide credit Yucheng Low



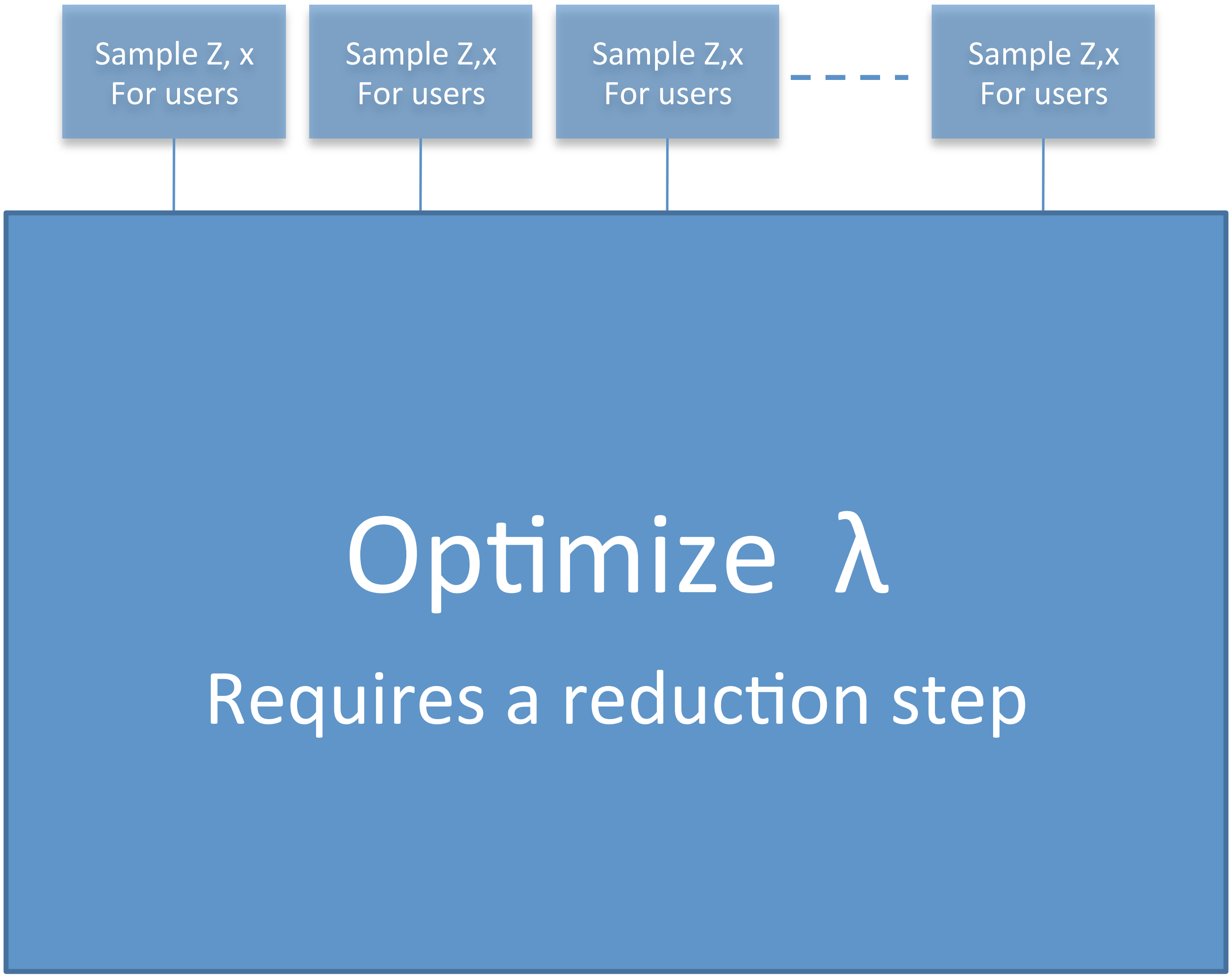
Slide credit Yucheng Low

Inference and Learning

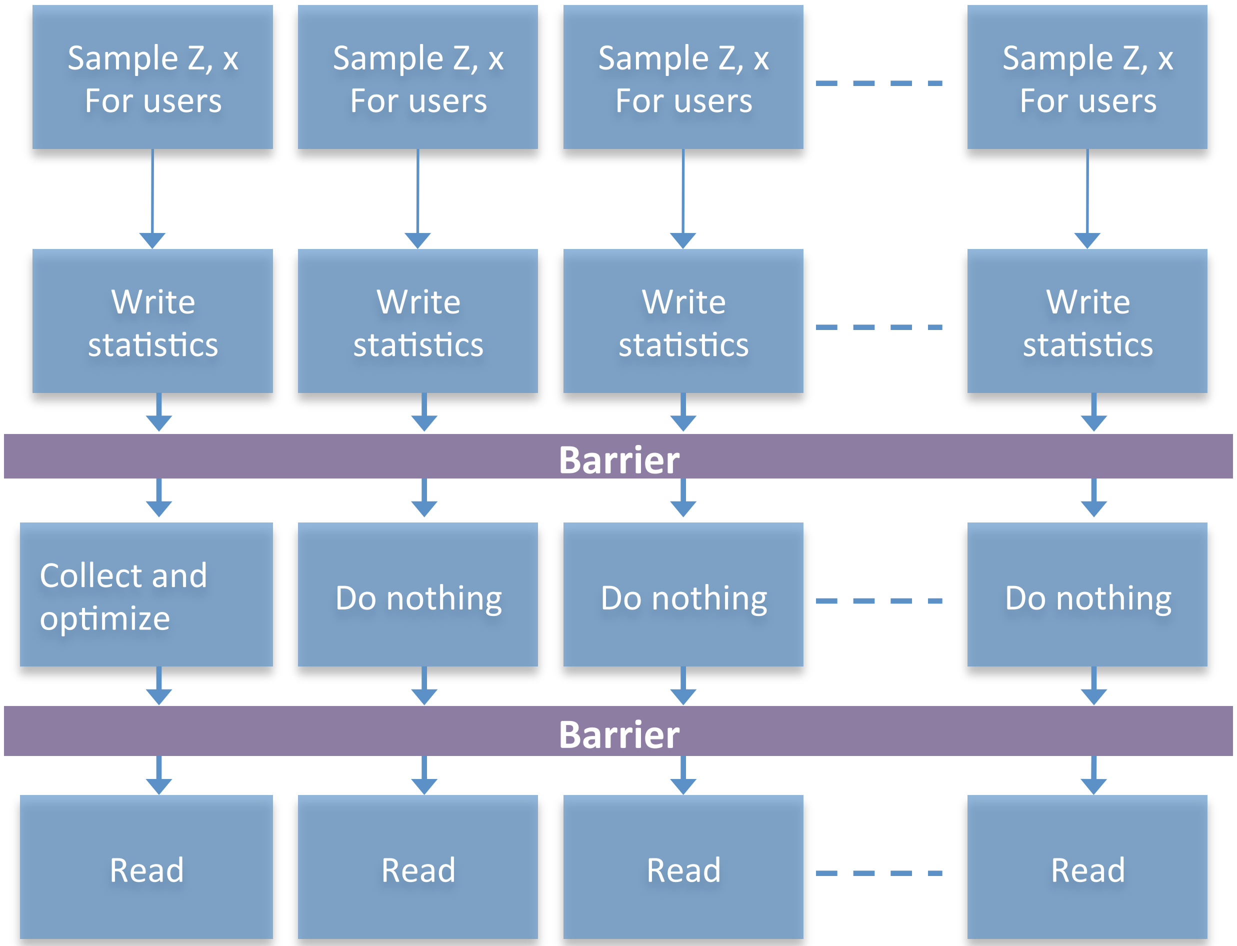


E-Step: sample local variables

Distributed Sampling Cycle



Distributed Sampling Cycle



Results

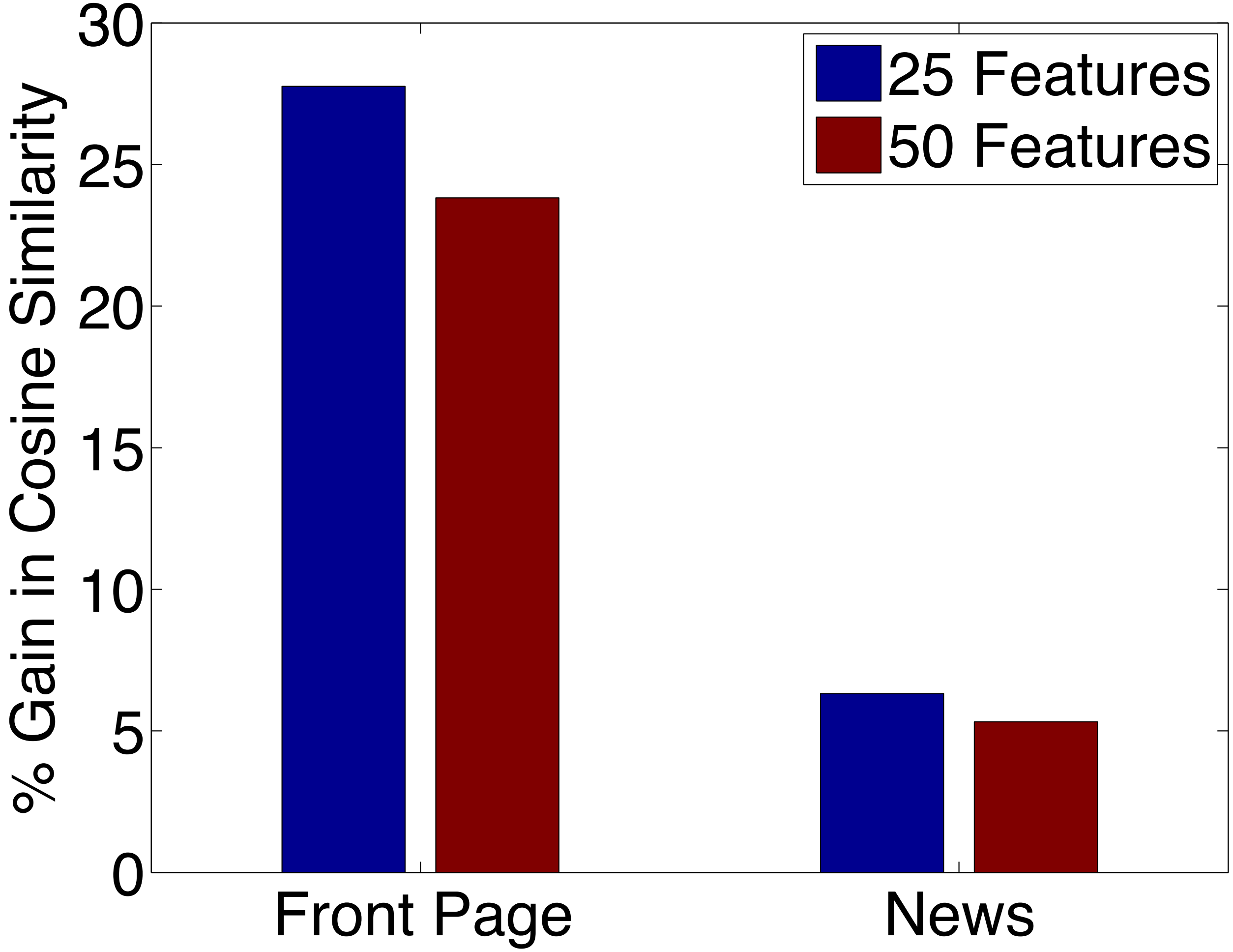
- **2 domain dataset.**

  Frontpage and News clicks of **5.6 million users.**

  Frontpage/News: Article text for each click.

- Measure gain relative to independent models on each domain

Results



% Gain in Cosine Similarity
30
25
20
15
10
5
0
25 Features
50 Features
Front Page
News

Analysis

Celebrity

sandra, oscar, oscars, red, carpet, bullock, golden, gown, bullocks, nominee, bestactress, sparkles, stunning,

vienna, bachelor, jake, pavelka, giraldi, finale, show, stars, dancing, love, season, time, abc,

Entertainment

Science

bacteria, fight, super, struggling, developed, doctors, resistant, lethal, virtually, drugs, antibiotic, competitors, chad,

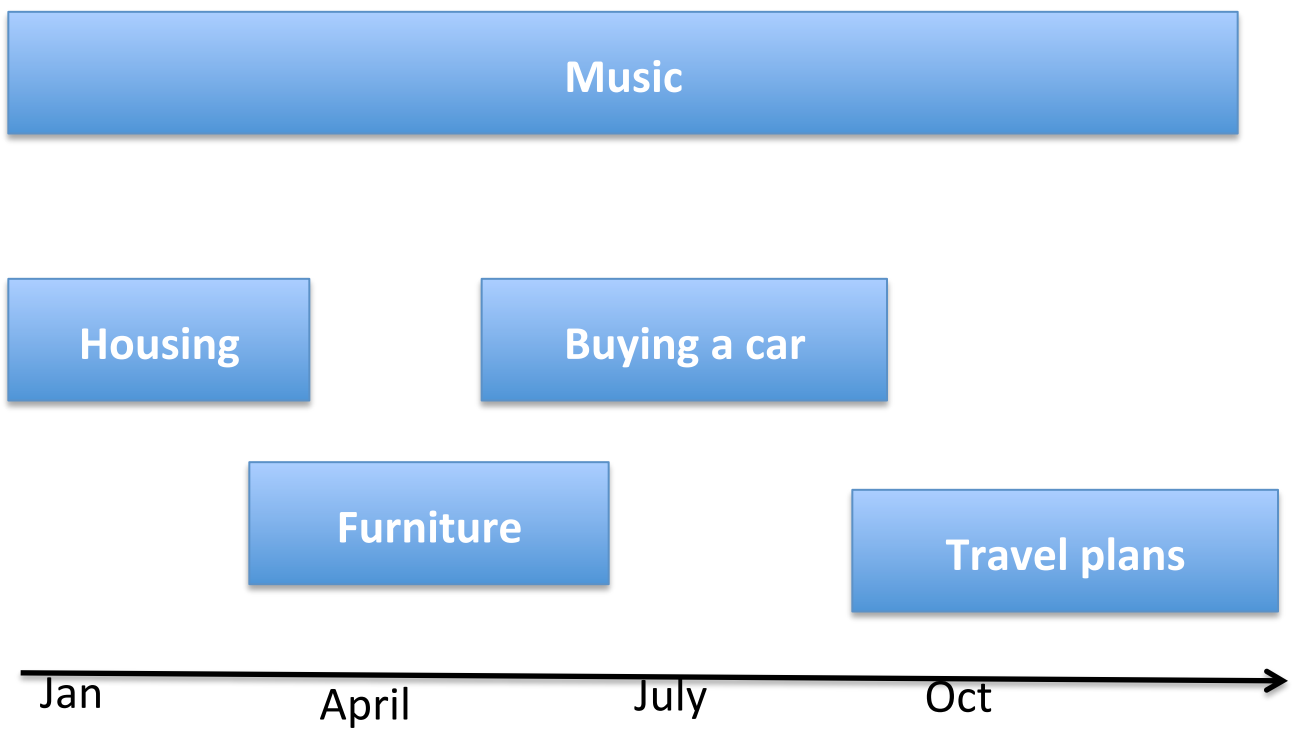
film, movie, movies, films, director, story, avatar, james, time, hollywood, big, make, hes, star,

Science Fiction

Tracking Users Interest

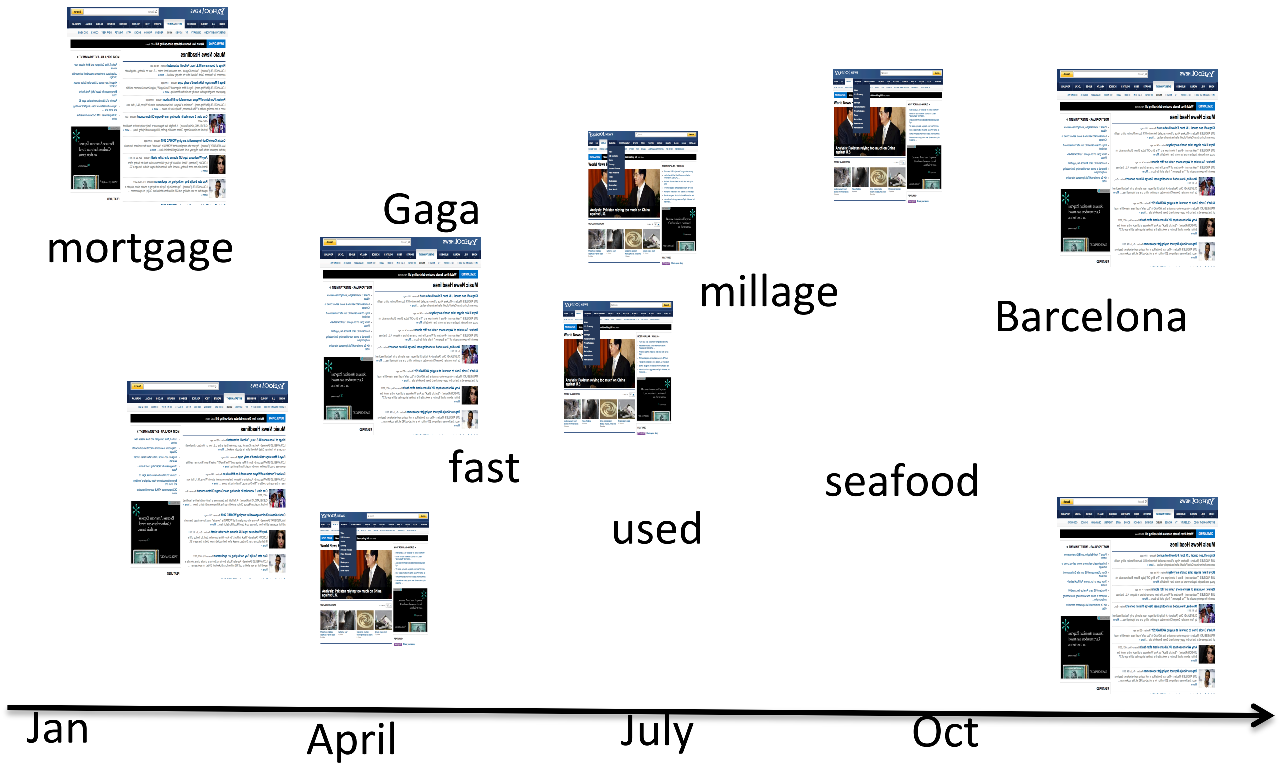
Characterizing User Interests

- Short term vs long-term



Characterizing User Interests

- Short term vs long-term
- Latent



Problem formulation

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 interests
- Dynamic interest representation
- Online and scalable inference
- Language independent

Flight
London
Hotel
weather

classes
registration
housing
rent

School
Supplies
Loan
semester

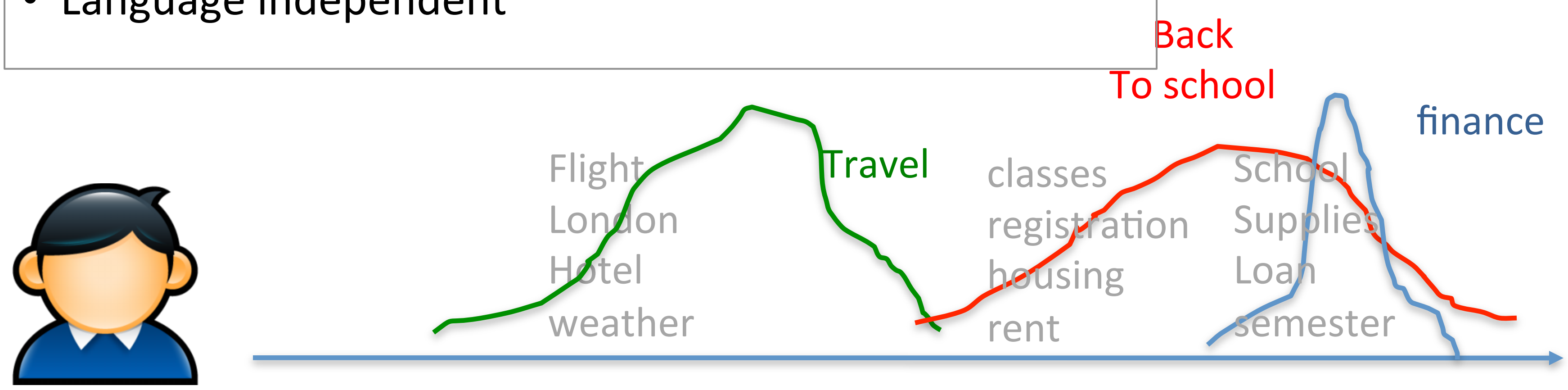
Problem formulation

Input

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

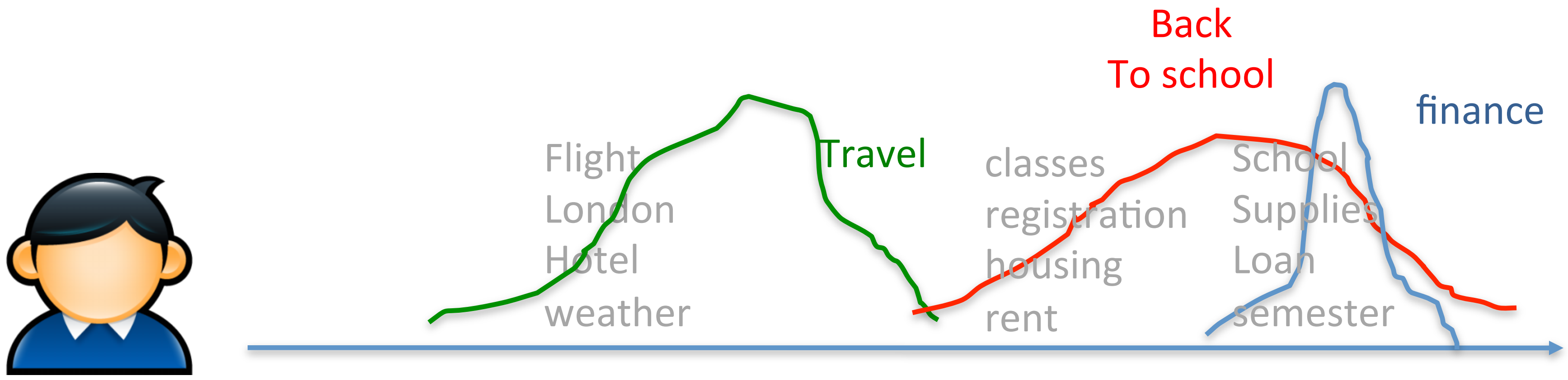
Output

- Users' daily distribution over interests
- Dynamic interest representation
- Online and scalable inference
- Language independent



Problem formulation

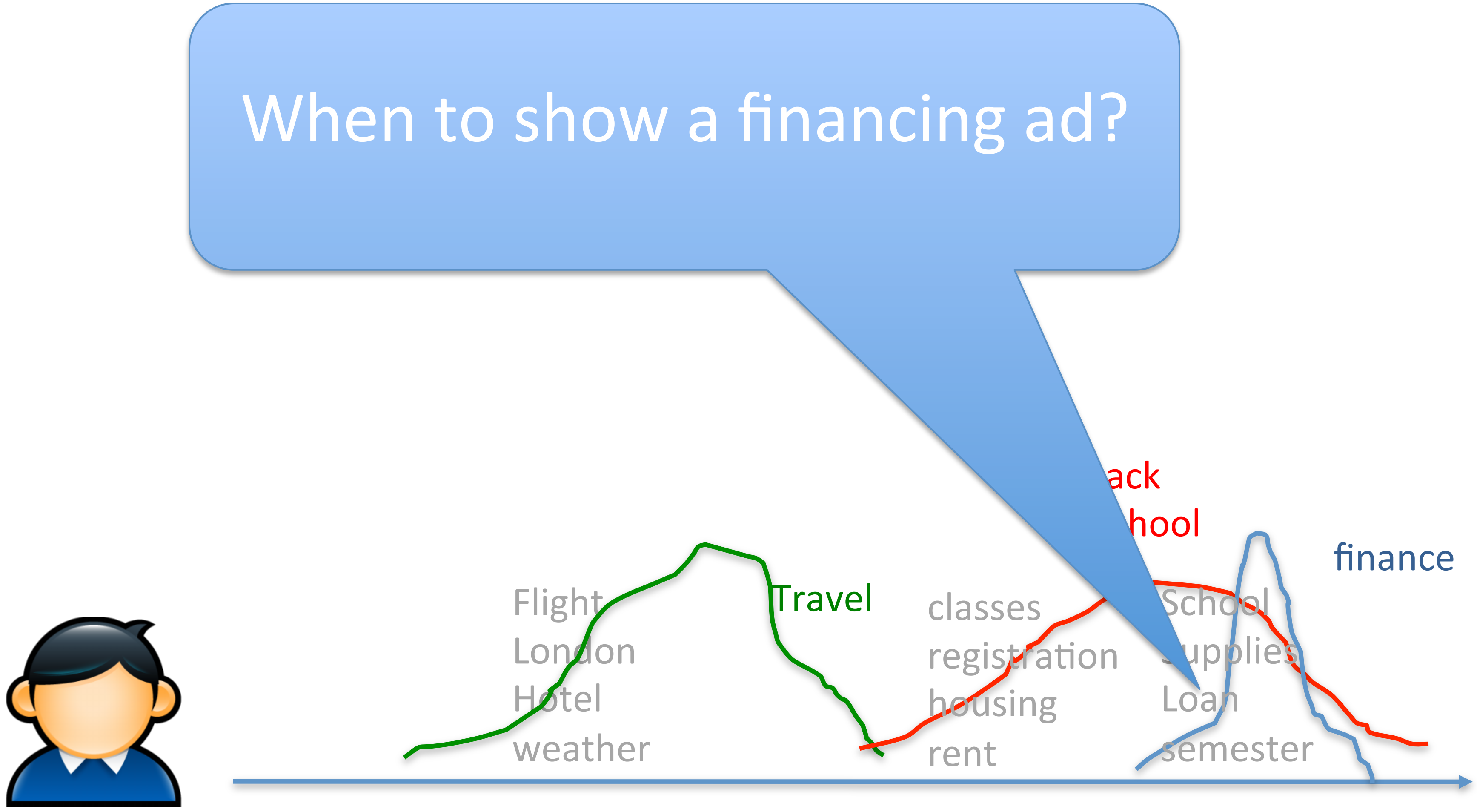
When to show a financing ad?



Problem formulation

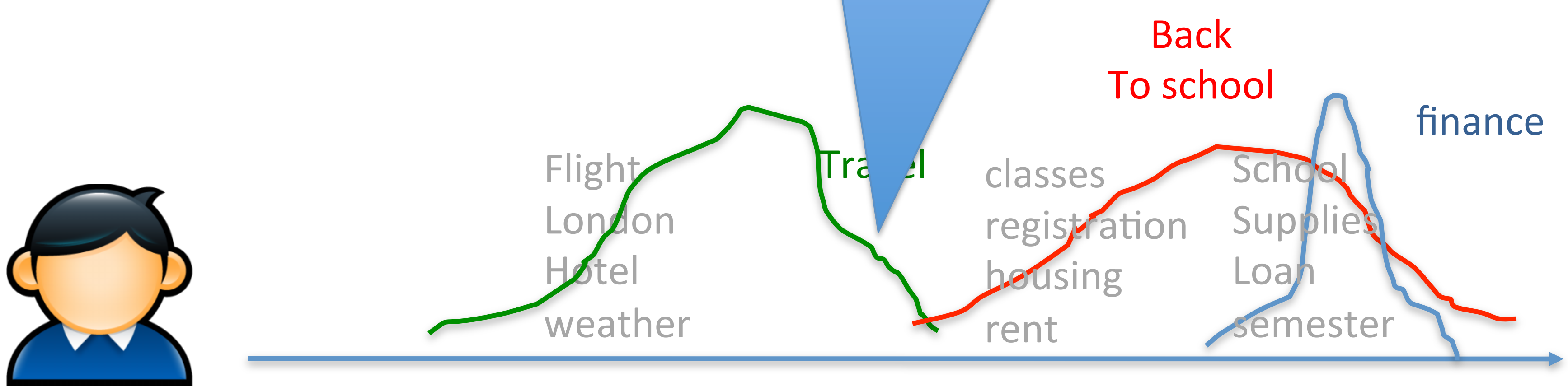
When to show a financing ad?

Problem formulation



Problem formulation

When to show a hotel ad?



Problem formulation

When to show a hotel ad?

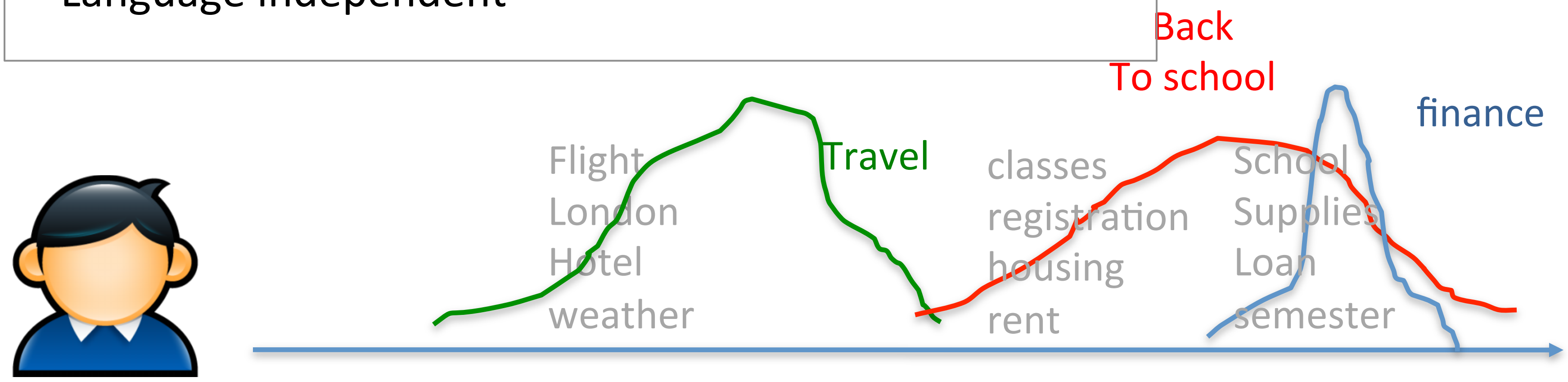
Problem formulation

Input

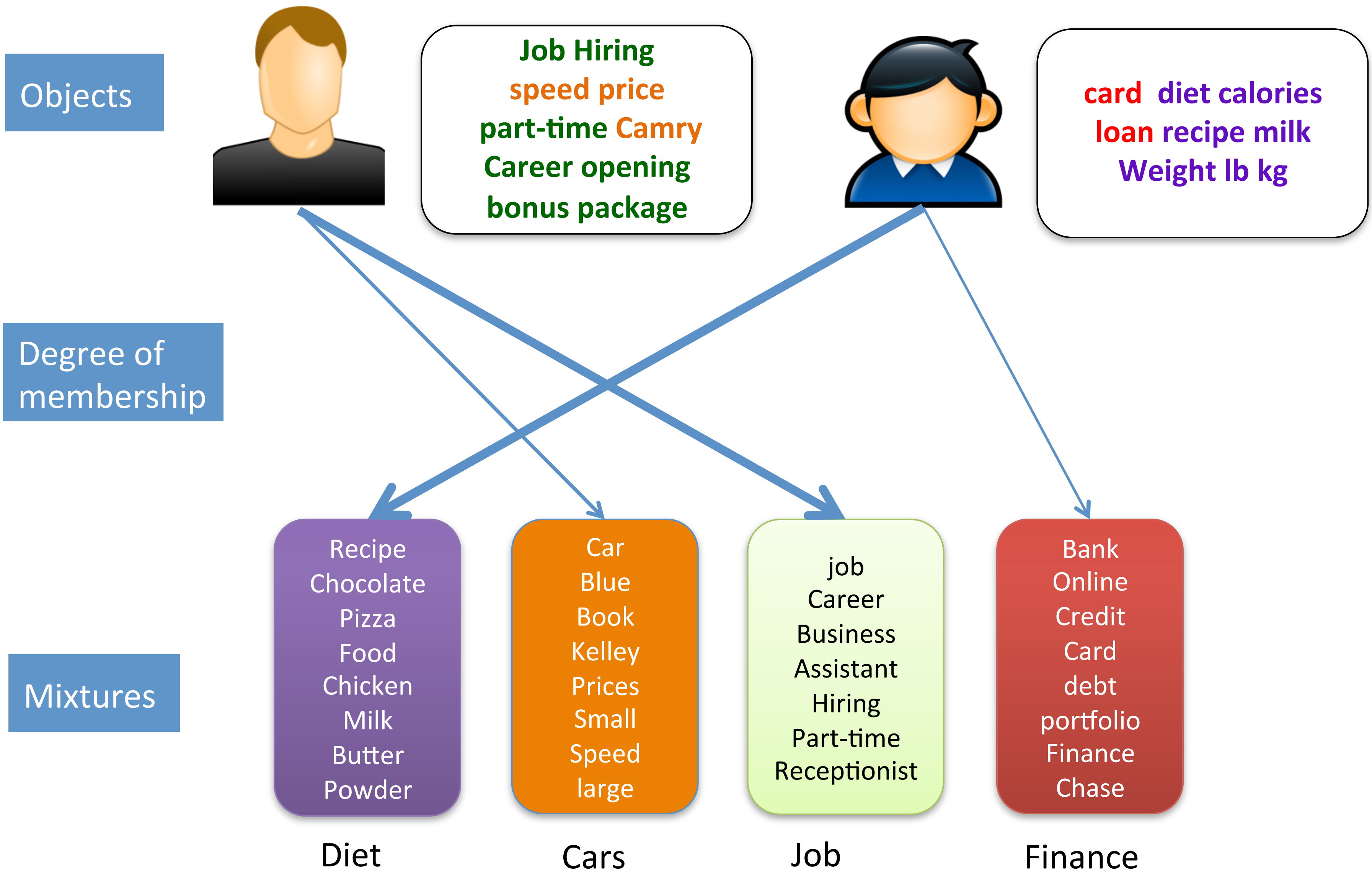
- Queries issued by the user or tags of watched content
- Snippet of page examined by user
- Time stamp of each action (day resolution)

Output

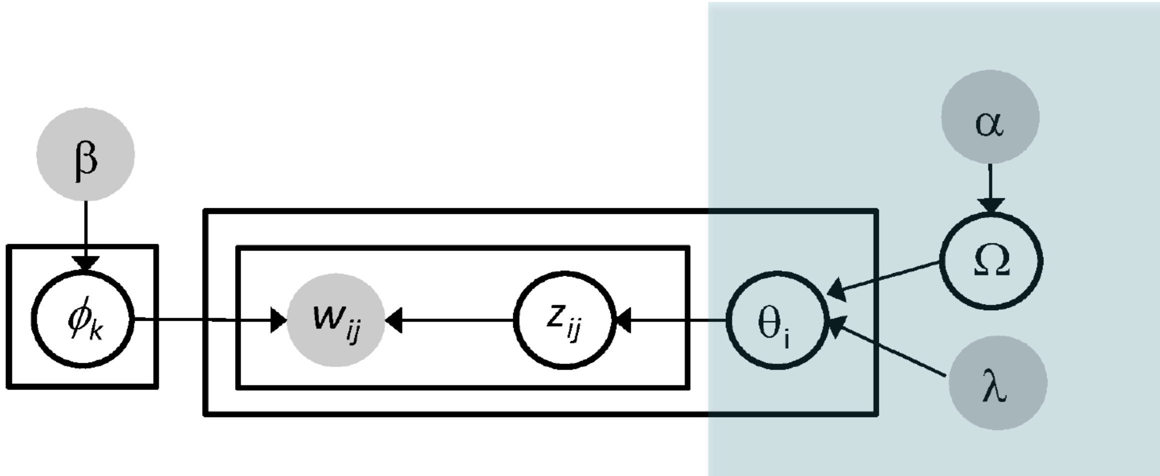
- Users' daily distribution over interests
- Dynamic interest representation
- Online and scalable inference
- Language independent



Mixed-Membership Formulation

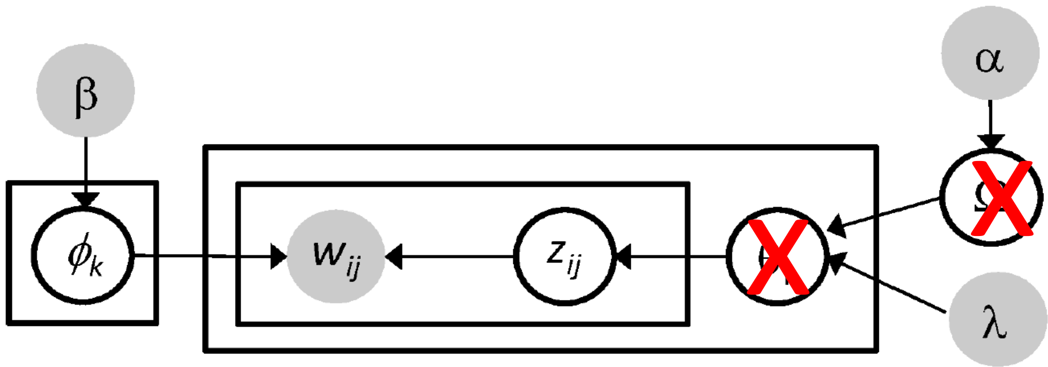


In Graphical Notation

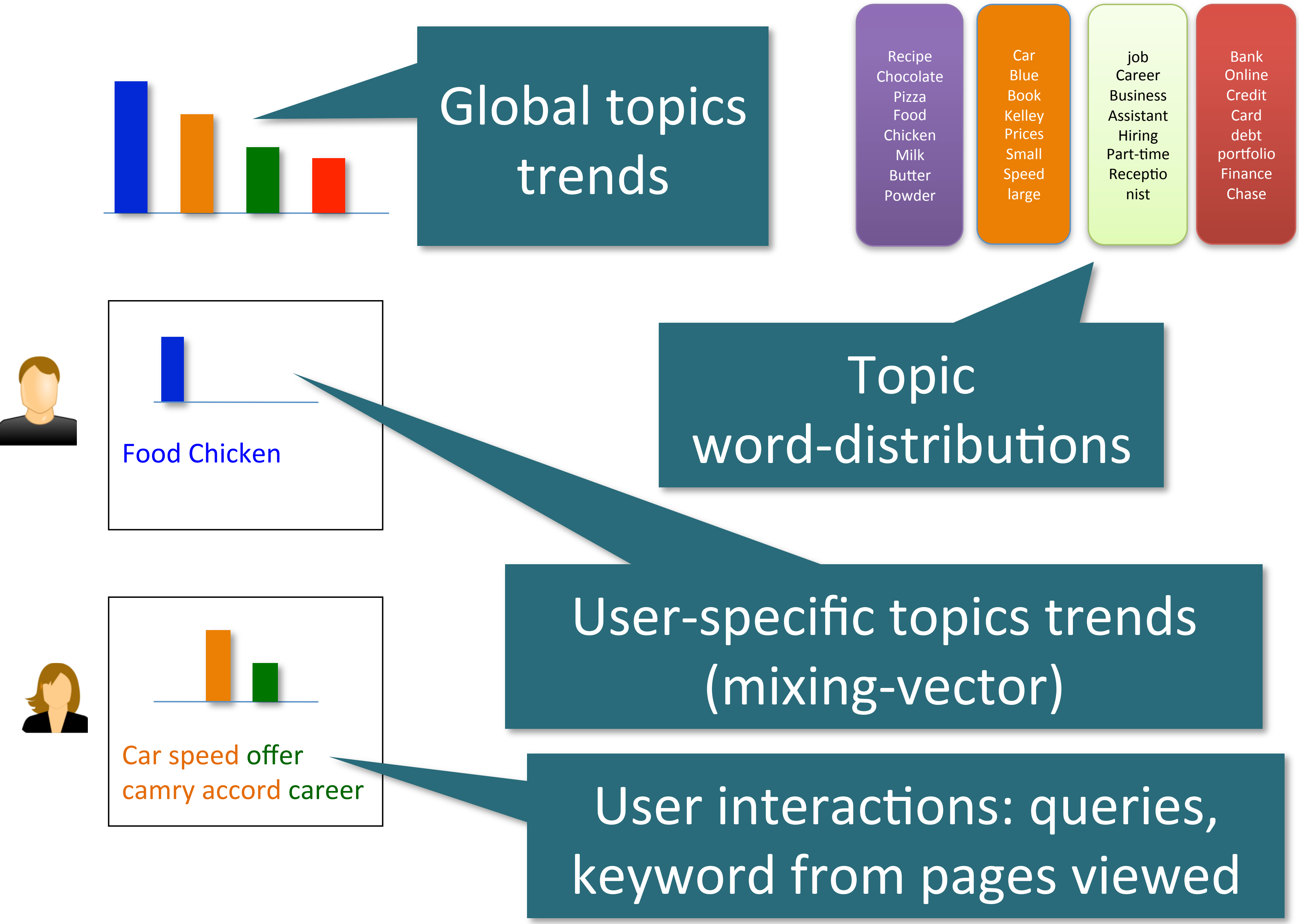


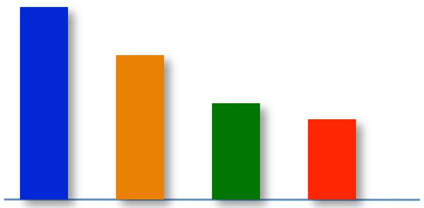
1. Draw once $\Omega|\alpha \sim \text{Dir}(\alpha/K)$.
2. Draw each topic $\phi_k|\beta \sim \text{Dir}(\beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i|\lambda, \Omega \sim \text{Dir}(\lambda\Omega)$.
   (b) For each word
      (a) Draw a topic $z_{ij}|\theta_d \sim \text{Mult}(\theta_i)$.
      (b) Draw a word $w_{ij}|z_{ij}, \phi \sim \text{Multi}(\phi_{z_{ij}})$.

In Polya-Urn Representation



- Collapse multinomial variables: θ, Ω
- Fixed-dimensional Hierarchal Polya-Urn representation
  - Chinese restaurant franchise


Global topics trends
Recipe Chocolate Pizza Food Chicken Milk Butter Powder
Car Blue Book Kelley Prices Small Speed large
job Career Business Assistant Hiring Part-time Receptio nist
Bank Online Credit Card debt portfolio Finance Chase
Topic word-distributions
Food Chicken
User-specific topics trends (mixing-vector)
Car speed offer camry accord career
User interactions: queries, keyword from pages viewed



Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

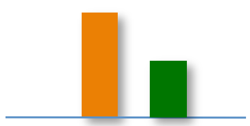
Car
Blue
Book
Kelley
Prices
Small
Speed
large

job
Career
Business
Assistant
Hiring
Part-time
Receptio
nist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

Food Chicken

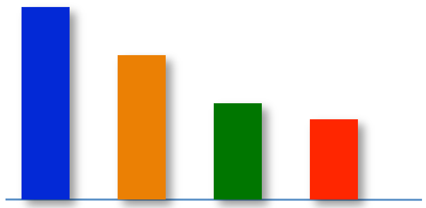
....



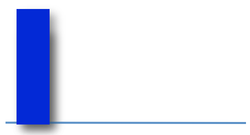
Car speed offer
camry accord career

Generative Process

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

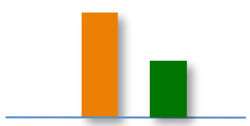


Recipe	Car	job	Bank
Chocolate	Blue	Career	Online
Pizza	Book	Business	Credit
Food	Kelley	Assistant	Card
Chicken	Prices	Hiring	debt
Milk	Small	Part-time	portfolio
Butter	Speed	Receptio	Finance
Powder	large	nist	Chase



Food Chicken

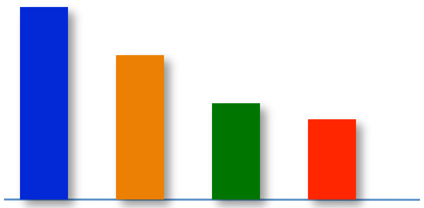
....

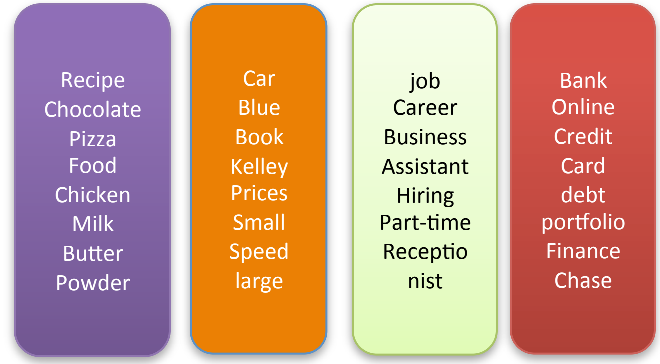


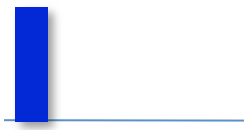
Car speed offer
camry accord career

Generative Process

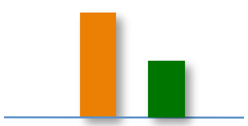
- For each user interaction
  - Choose an intent from local distribution
    - Sample word from the topic's word-distribution
  - Choose a new intent $\propto \lambda$
    - Sample a new intent from the global distribution
      - Sample word from the new topic word-distribution







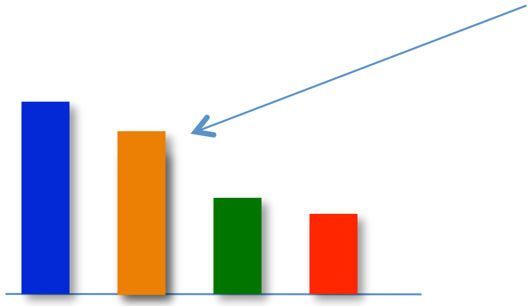
Food Chicken pizza ....

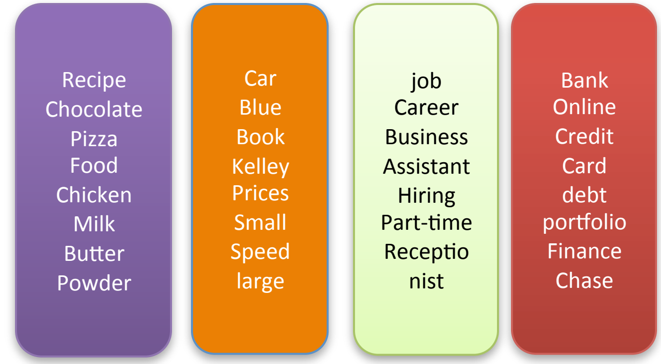


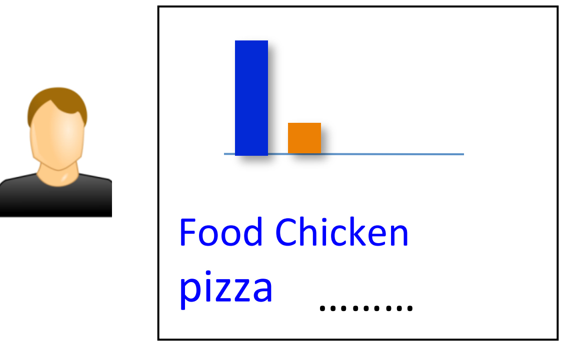
Car speed offer camry accord career

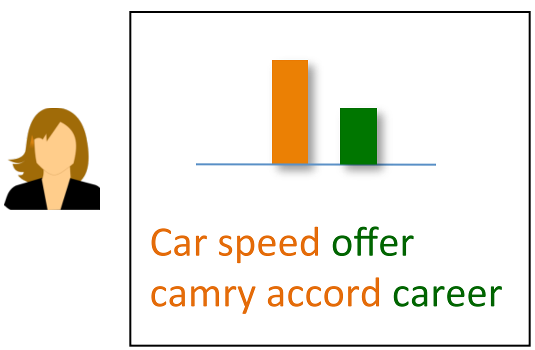
Generative Process

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



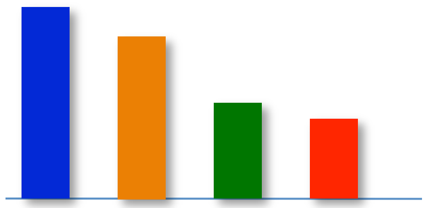






Generative Process

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

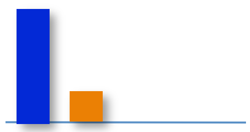


Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

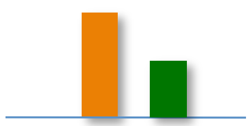
Car
Blue
Book
Kelley
Prices
Small
Speed
large

job
Career
Business
Assistant
Hiring
Part-time
Receptio
nist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



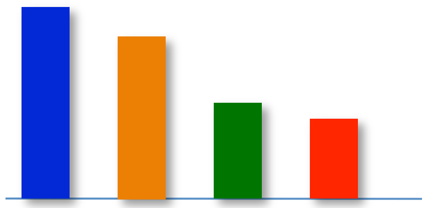
Food Chicken
pizza millage



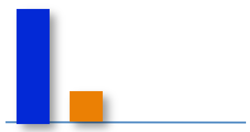
Car speed offer
camry accord career

Generative Process

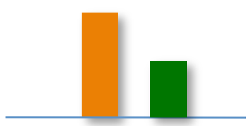
- For each user interaction
  - Choose an intent from local distribution
    - Sample word from topic's word-distribution
  - Choose a new intent $\propto \lambda$
    - Sample a new intent from the global distribution
      - Sample from word the new topic word-distribution



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



Food Chicken
pizza millage

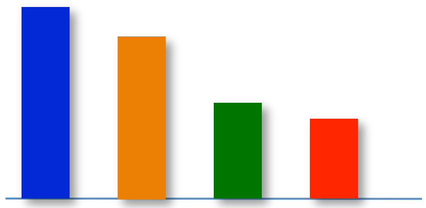


Car speed offer
camry accord career

Problems

- Static Model
- Does not evolve user’s interests
- Does not evolve the global trend of interests
- Does not evolve interest’s distribution over terms

At time t



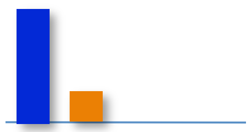
At time t+1

Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Blue
Book
Kelley
Prices
Small
Speed
large

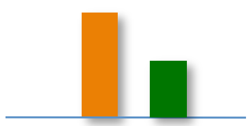
job
Career
Business
Assistant
Hiring
Part-time
Receptio
nist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



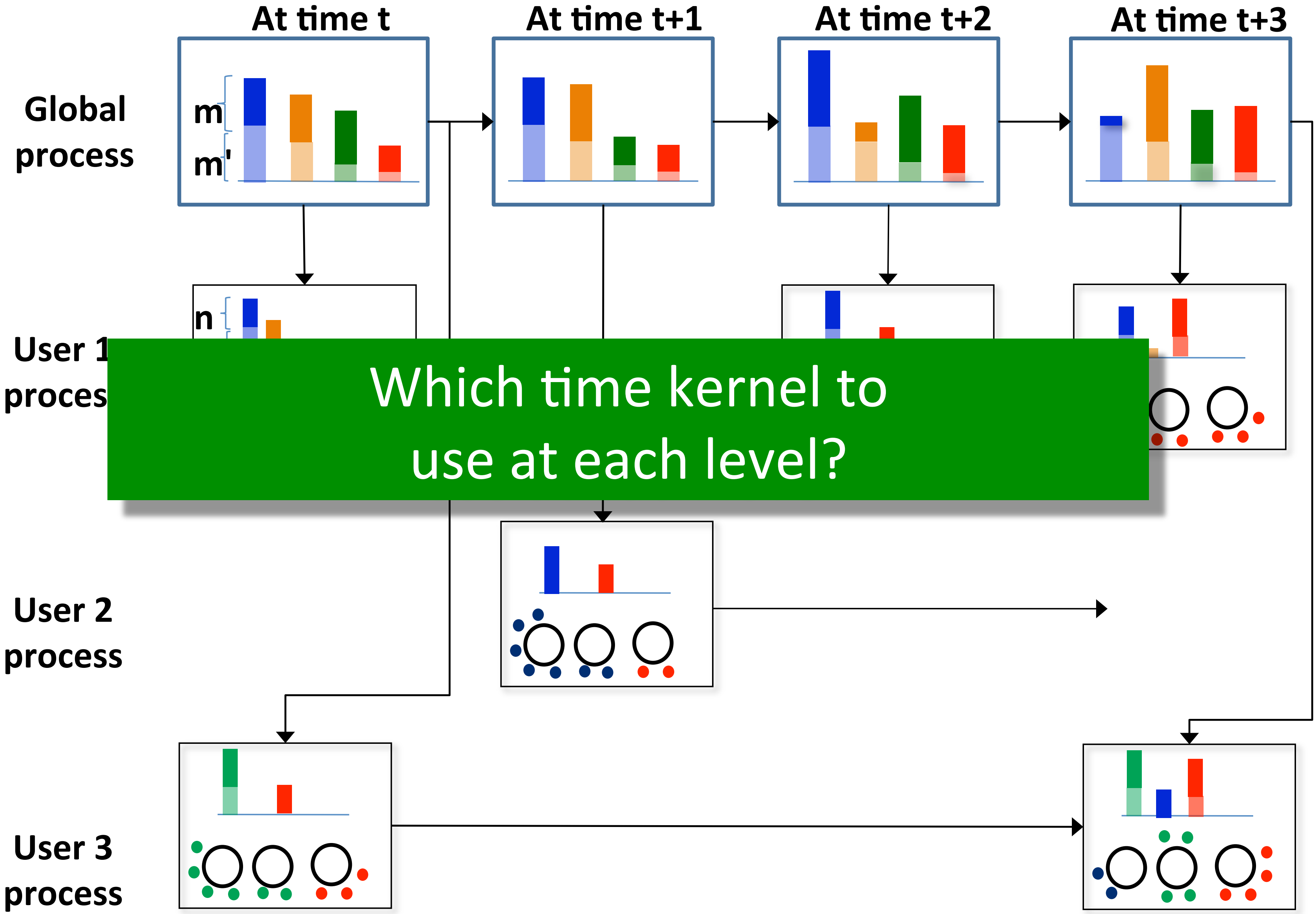
Food Chicken
pizza millage

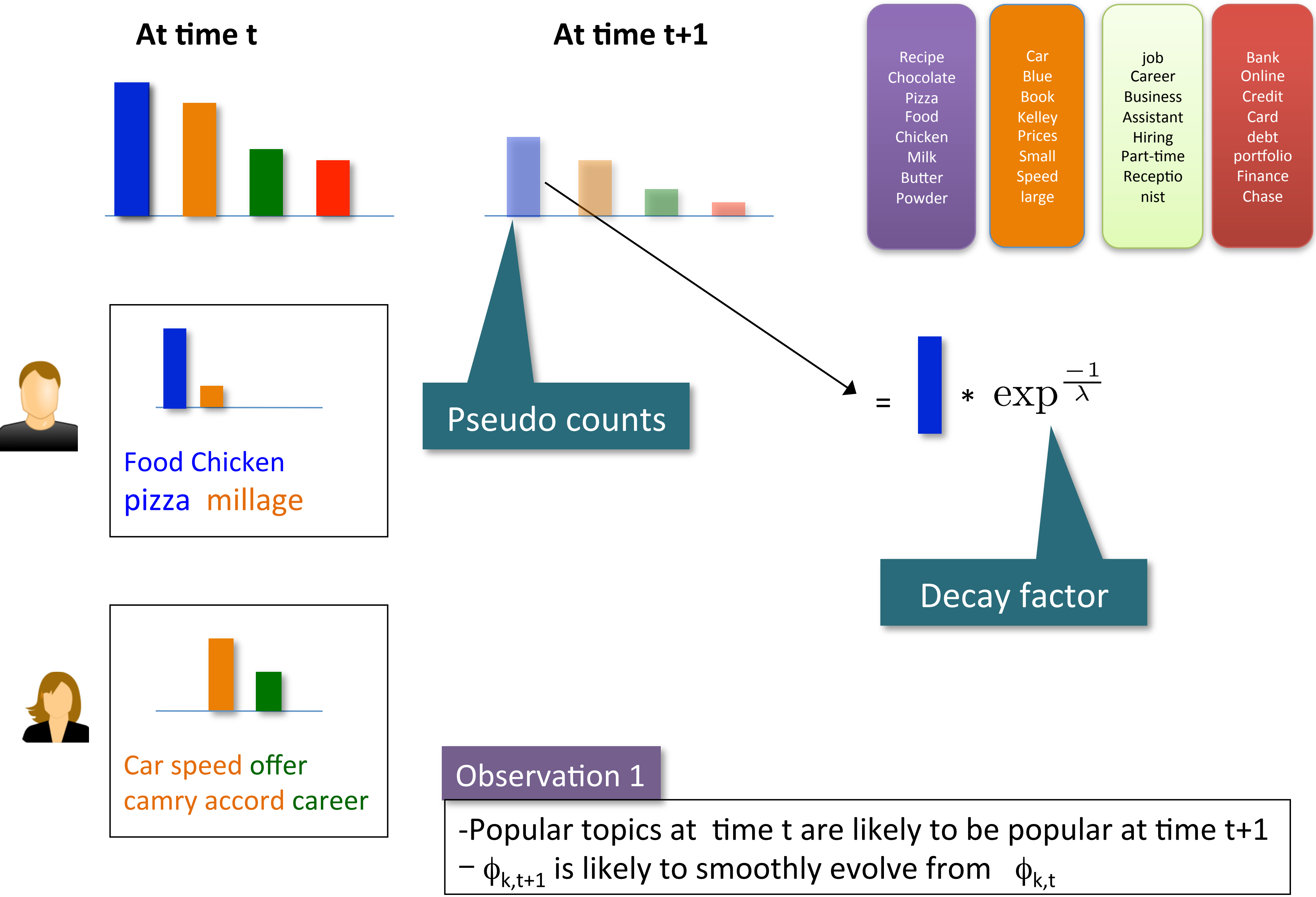
Build a dynamic model



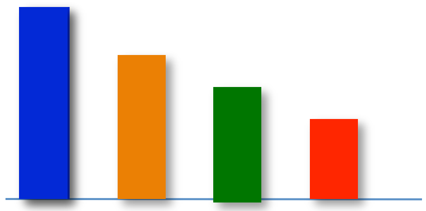
Car speed offer
camry accord career

Connect each level using a RCRP

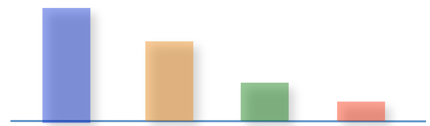

At time t
At time t+1
At time t+2
At time t+3
Global process
m
m'
n
User 1 proces
User 2 process
User 3 process
Which time kernel to use at each level?


At time t
At time t+1
Recipe Chocolate Pizza Food Chicken Milk Butter Powder
Car Blue Book Kelley Prices Small Speed large
job Career Business Assistant Hiring Part-time Receptio nist
Bank Online Credit Card debt portfolio Finance Chase
Pseudo counts
$= \; * \; \exp^{\frac{-1}{\lambda}}$
Decay factor
Food Chicken pizza millage
Car speed offer camry accord career
Observation 1
-Popular topics at time t are likely to be popular at time t+1
− $\phi_{k,t+1}$ is likely to smoothly evolve from $\phi_{k,t}$

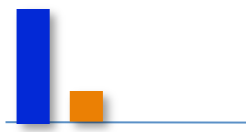
At time t



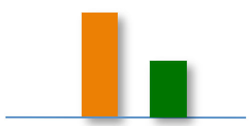
At time t+1



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



Food Chicken
pizza millage



Car speed offer
camry accord career

Car
Altima
Accord
Book
Kelley
Prices
Small
Speed

Intuition

Captures current trend of the car industry
(new release for e.g.)

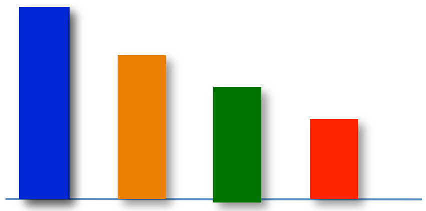
$\phi_{k,t}$ $\qquad$ $\phi_{k,t+1} \sim \mathrm{Dir}(\tilde{\beta}_{k,t+1})$

Observation 1

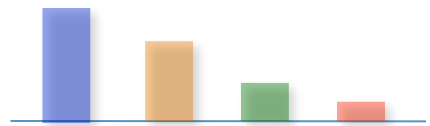
-Popular topics at time t are likely to be popular at time t+1

– $\phi_{k,t+1}$ **is likely to smoothly evolve** from $\phi_{k,t}$

At time t



At time t+1

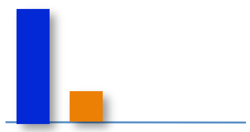


Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

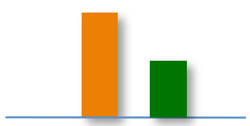
job
Career
Business
Assistant
Hiring
Part-time
Receptio
nist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



Food Chicken
pizza millage

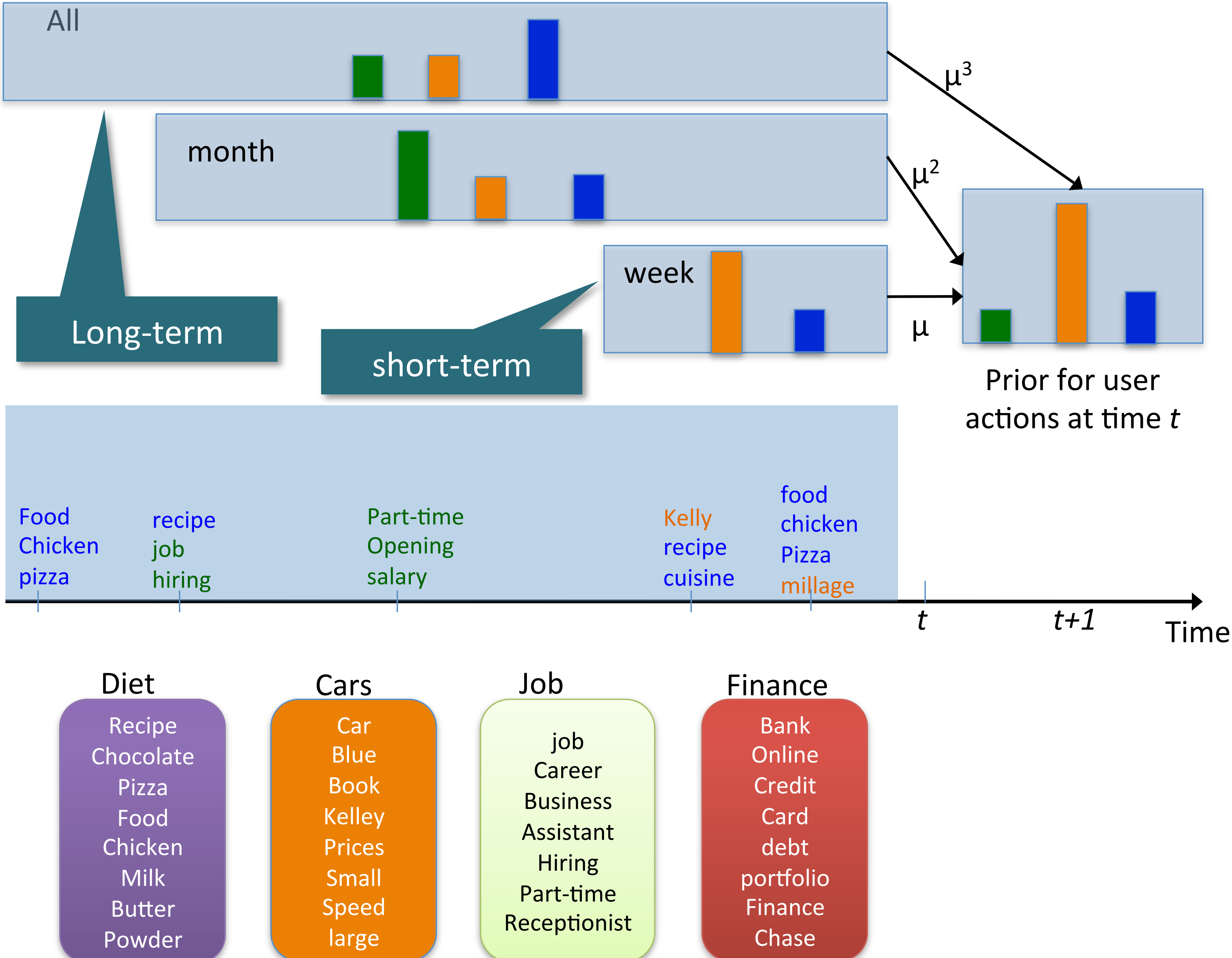
How do we get a prior that captures both long and short term interest?

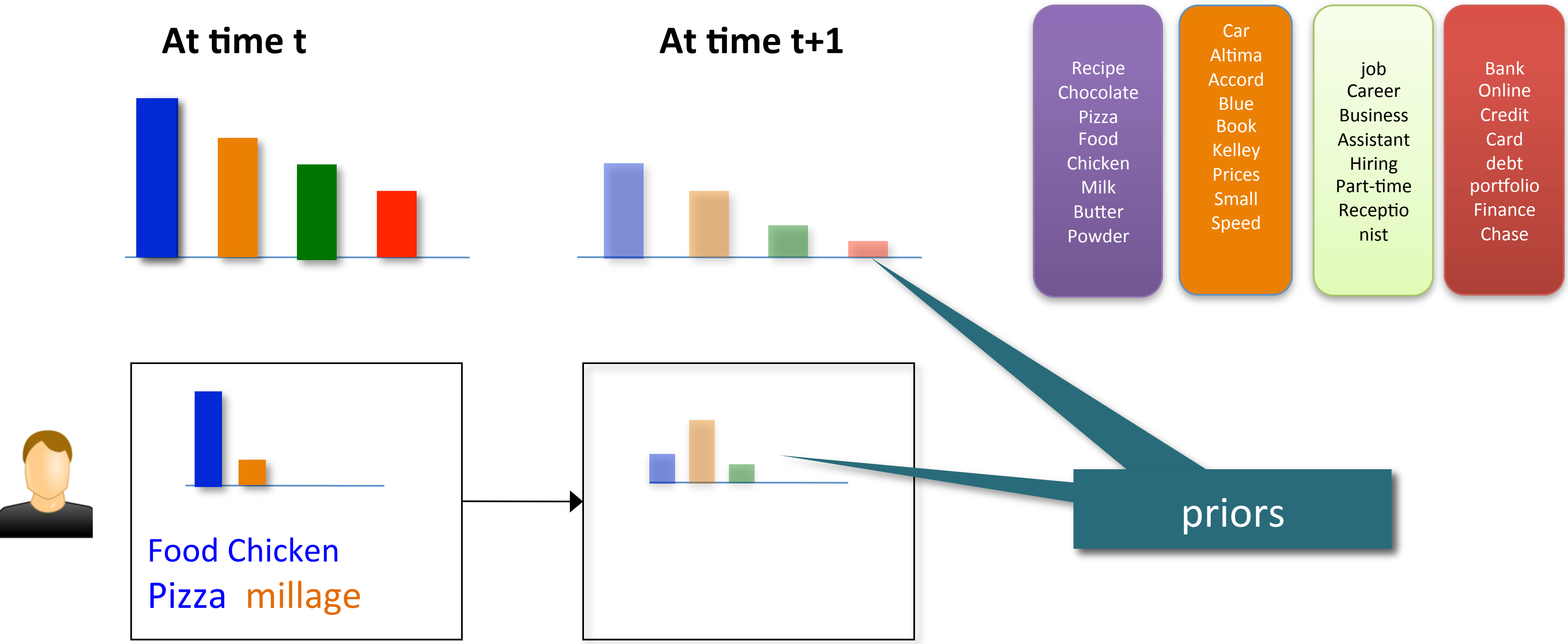


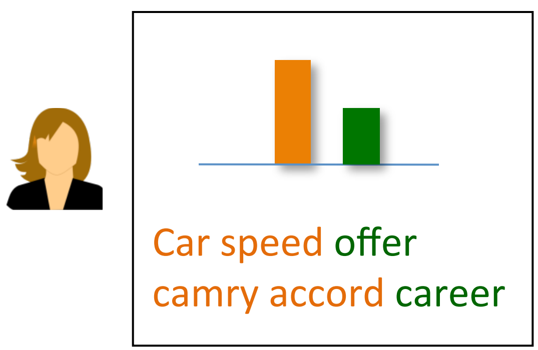
Car speed offer
camry accord career

Observation 2

- User prior at time t+1 is a mixture of the user short and long term interest

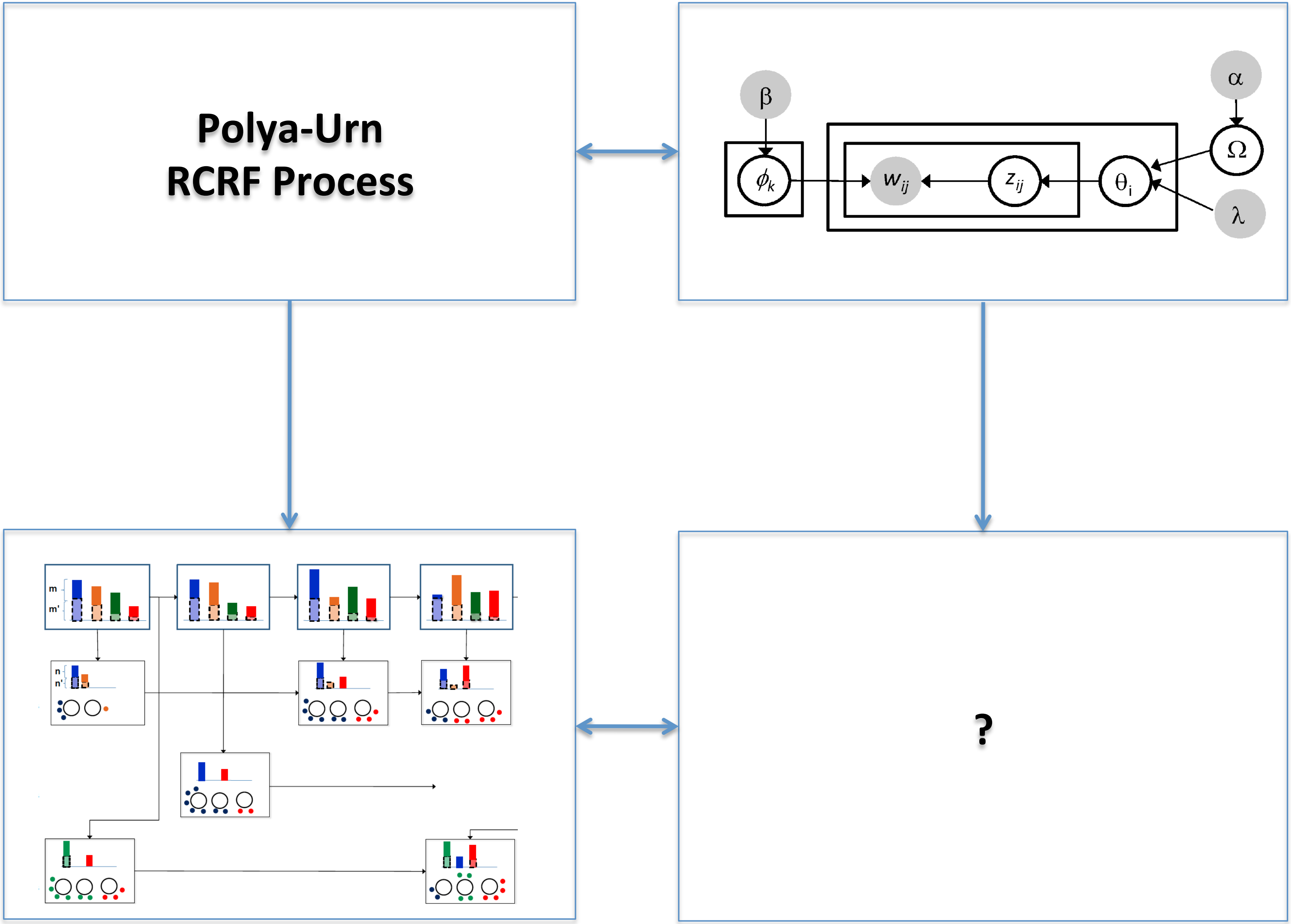

All
month
week
μ^3
μ^2
μ
Long-term
short-term
Prior for user actions at time t
Food Chicken pizza
recipe job hiring
Part-time Opening salary
Kelly recipe cuisine
food chicken Pizza millage
t
$t+1$
Time
Diet
Recipe Chocolate Pizza Food Chicken Milk Butter Powder
Cars
Car Blue Book Kelley Prices Small Speed large
Job
job Career Business Assistant Hiring Part-time Receptionist
Finance
Bank Online Credit Card debt portfolio Finance Chase





Generative Process

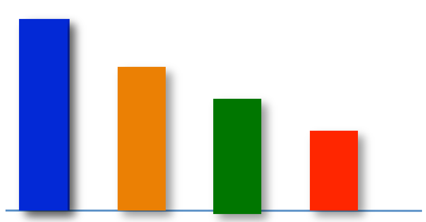
- For each user interaction
  - Choose an intent from local distribution
    - Sample word from the topic's word-distribution
  - Choose a new intent $\propto \lambda$
    - Sample a new intent from the global distribution
      - Sample word from the new topic word-distribution


Polya-Urn
RCRF Process
β
α
Ω
λ
ϕ_k
w_{ij}
z_{ij}
θ_i
m
m'
n
n'
?

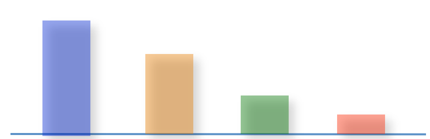
Simplified Graphical Model

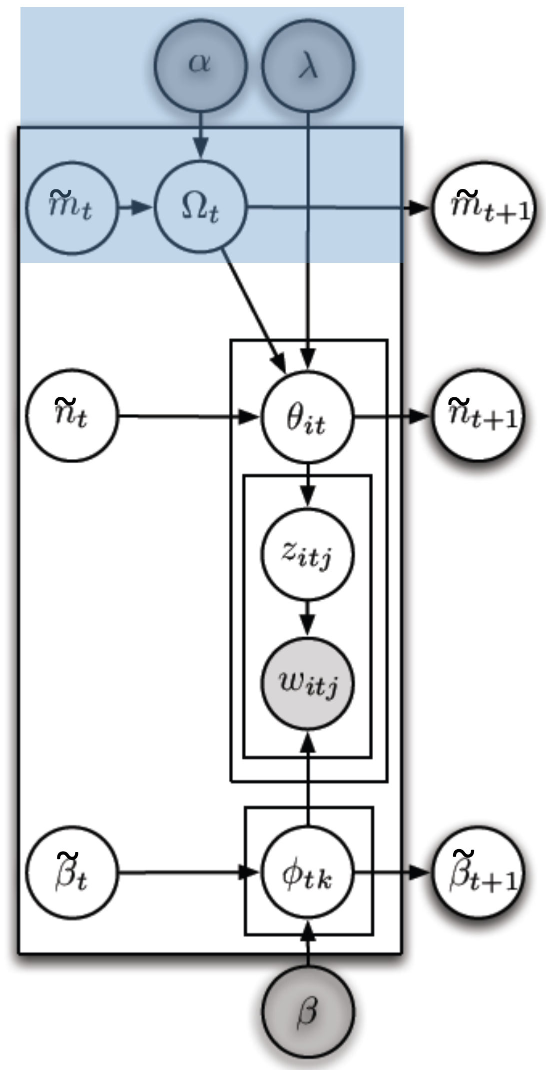
1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t \sim \mathrm{Dir}(\lambda\Omega^t + \tilde{\mathbf{n}}_i^t)$.
   (b) For each word
      (a) Draw a topic $z_{in}^t|\theta_i^t \sim \mathrm{Mult}(\theta_i^t)$.
      (b) Draw a word $w_{in}^t|z_{ij}^t, \phi^t \sim \mathrm{Multi}(\phi^t_{z_{ij}^t})$.

At time t



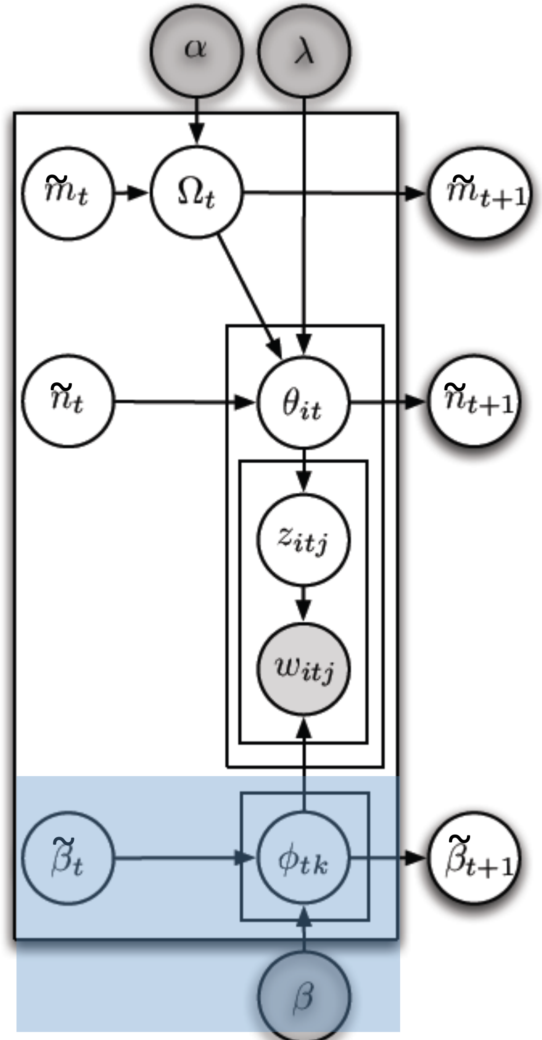
At time t+1





Simplified Graphical Model

1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t \sim \mathrm{Dir}(\lambda\Omega^t + \tilde{\mathbf{n}}_i^t)$.
   (b) For each word
      (a) Draw a topic $z_{in}^t|\theta_i^t \sim \mathrm{Mult}(\theta_i^t)$.
      (b) Draw a word $w_{in}^t|z_{ij}^t, \phi^t \sim \mathrm{Multi}(\phi^t_{z_{ij}^t})$.



$$\tilde{\beta}_{kw}^t = \sum_{h=1}^{t-1} \exp^{\frac{h-t}{\kappa 0}} n_{kw}^h$$

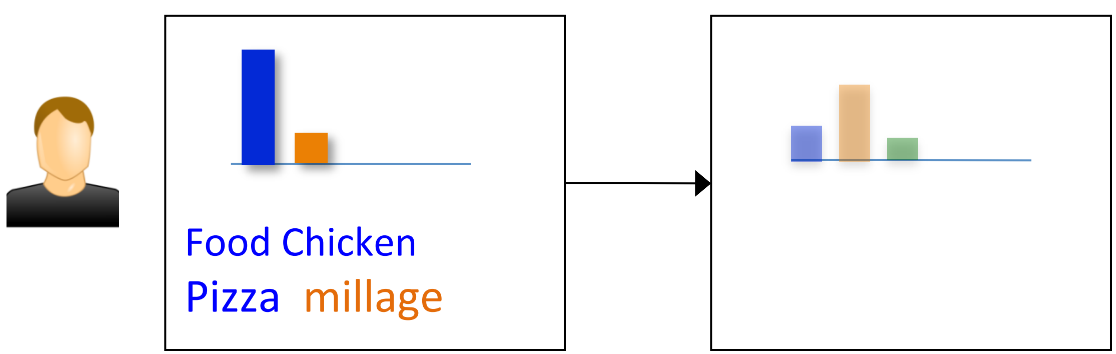
Car
Blue
Book
Kelley
Prices
Small
Speed
large

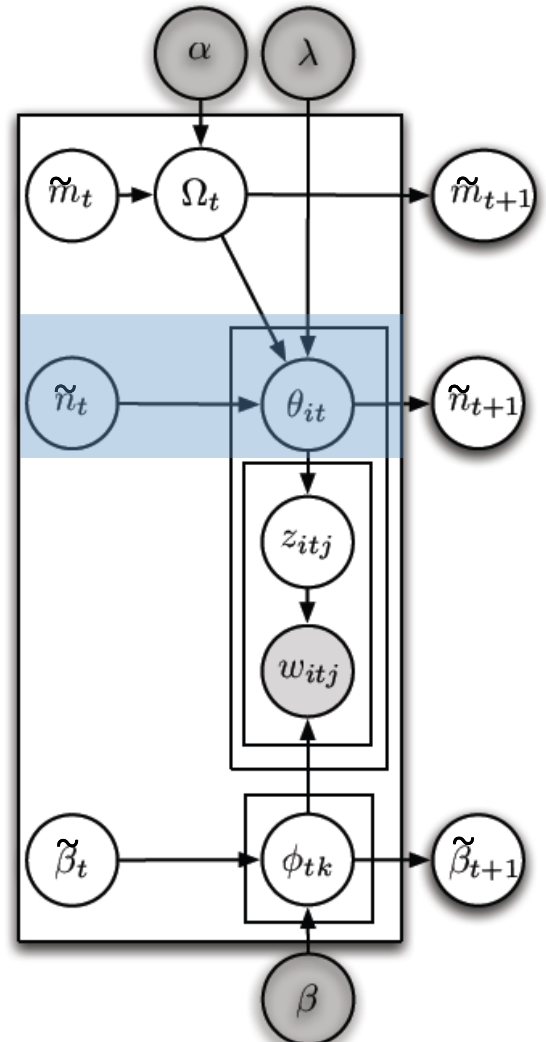
→

Car
Altima
Accord
Book
Kelley
Prices
Small
Speed

Simplified Graphical Model

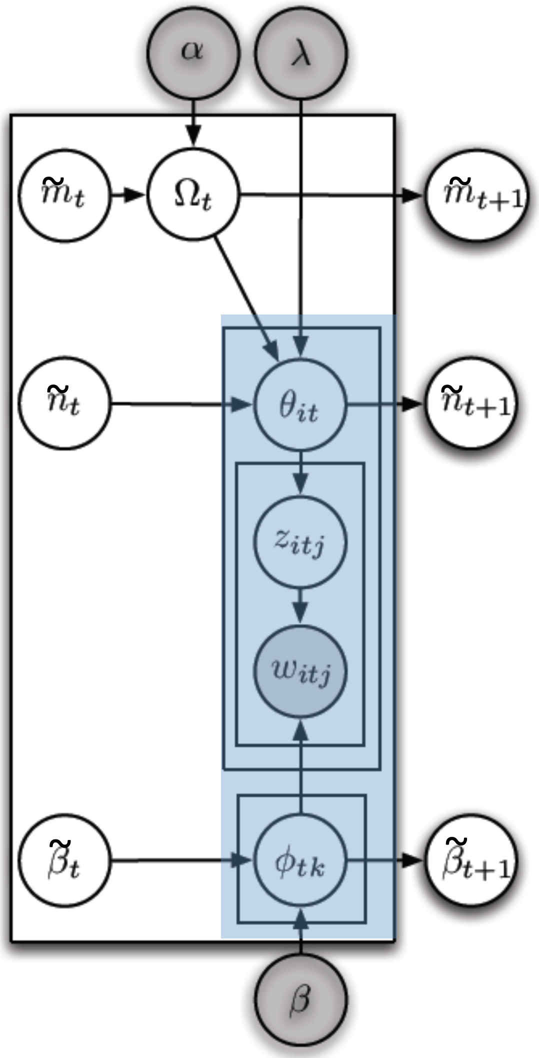
1. Draw once $\Omega^t | \alpha, \tilde{m}^t \sim \text{Dir}\Big(\tilde{\mathbf{m}}^t + \alpha/K\Big)$.
2. Draw each topic, $\phi^t_k | \beta, \tilde{\beta}^t_k \sim \text{Dir}(\tilde{\beta}^t_k + \beta)$.
3. For each user i:
   - (a) Draw topic proportions $\theta^t_i | \lambda, \Omega^t, \tilde{\mathbf{n}}^t_i \sim \text{Dir}(\lambda\Omega^t + \tilde{\mathbf{n}}^t_i)$.
   - (b) For each word
     - (a) Draw a topic $z^t_{in} | \theta^t_i \sim \text{Mult}(\theta^t_i)$.
     - (b) Draw a word $w^t_{in} | z^t_{ij}, \phi^t \sim \text{Multi}(\phi^t_{z^t_{ij}})$.





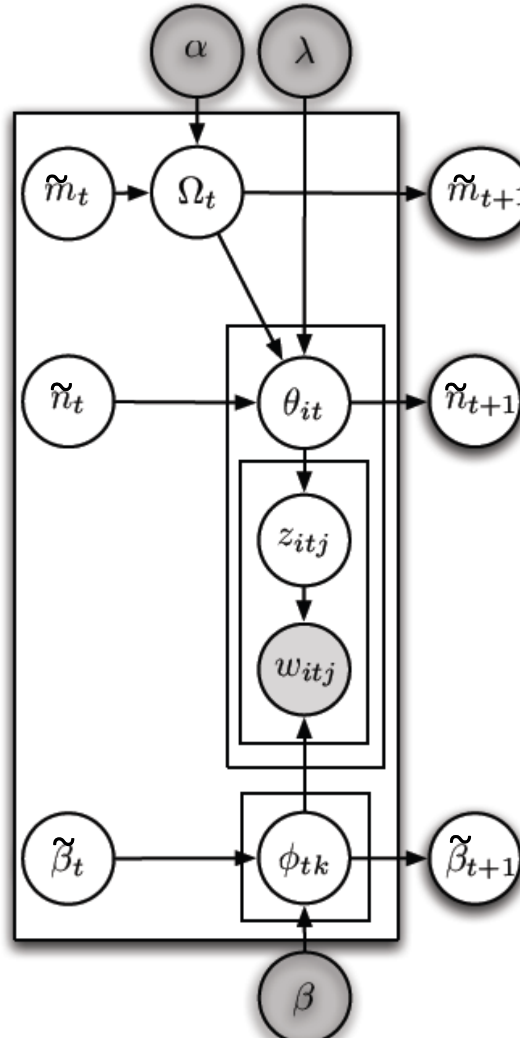
Simplified Graphical Model

1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t \sim \mathrm{Dir}(\lambda\Omega^t + \tilde{\mathbf{n}}_i^t)$.
   (b) For each word
      (a) Draw a topic $z_{in}^t|\theta_i^t \sim \mathrm{Mult}(\theta_i^t)$.
      (b) Draw a word $w_{in}^t|z_{ij}^t, \phi^t \sim \mathrm{Multi}(\phi^t_{z_{ij}^t})$.



Simplified Graphical Model

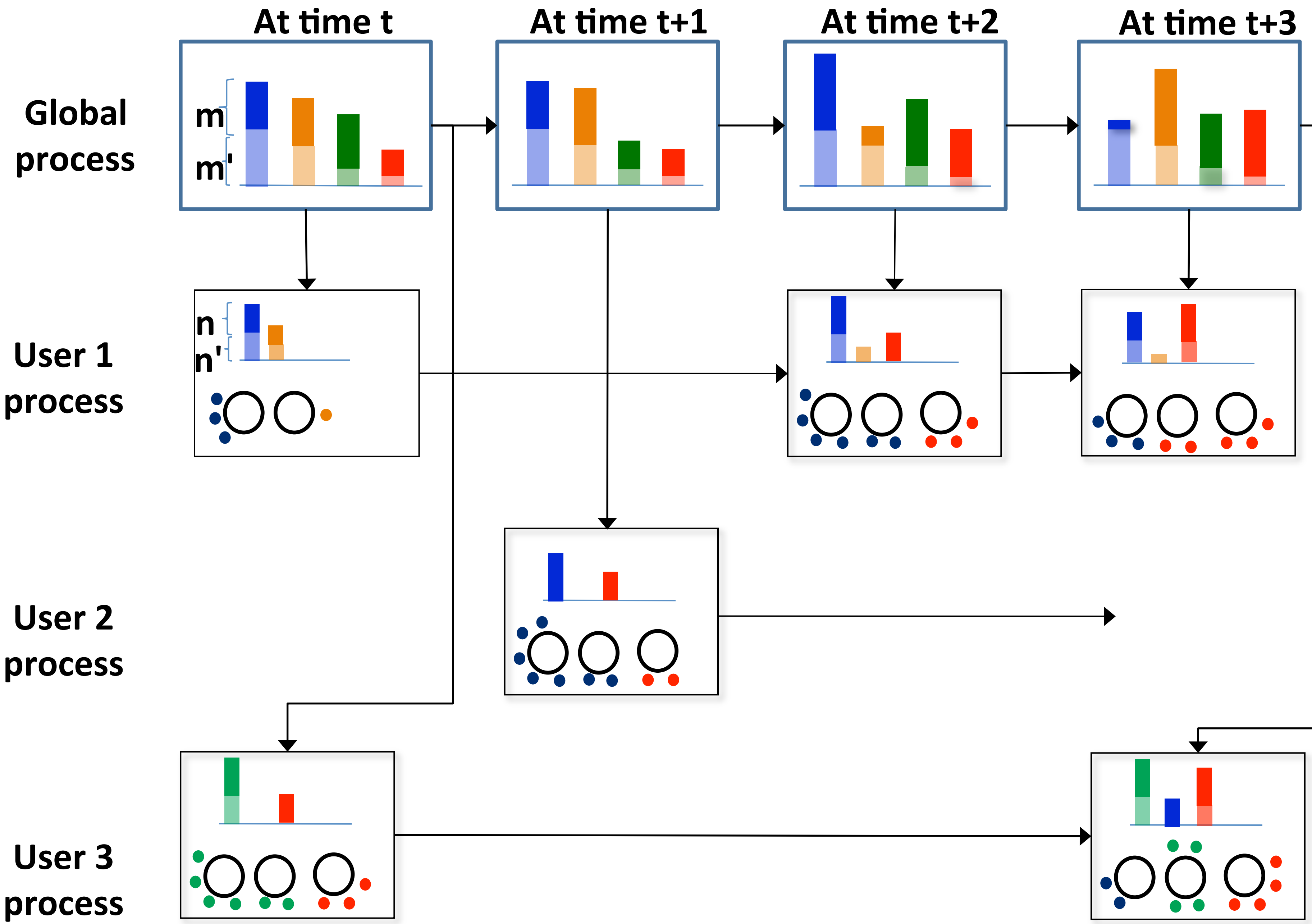
1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t \sim \mathrm{Dir}(\lambda\Omega^t + \tilde{\mathbf{n}}_i^t)$.
   (b) For each word
      (a) Draw a topic $z_{in}^t|\theta_i^t \sim \mathrm{Mult}(\theta_i^t)$.
      (b) Draw a word $w_{in}^t|z_{ij}^t, \phi^t \sim \mathrm{Multi}(\phi^t_{z_{ij}^t})$.



Topics evolve over time?

User's intent evolve over time?

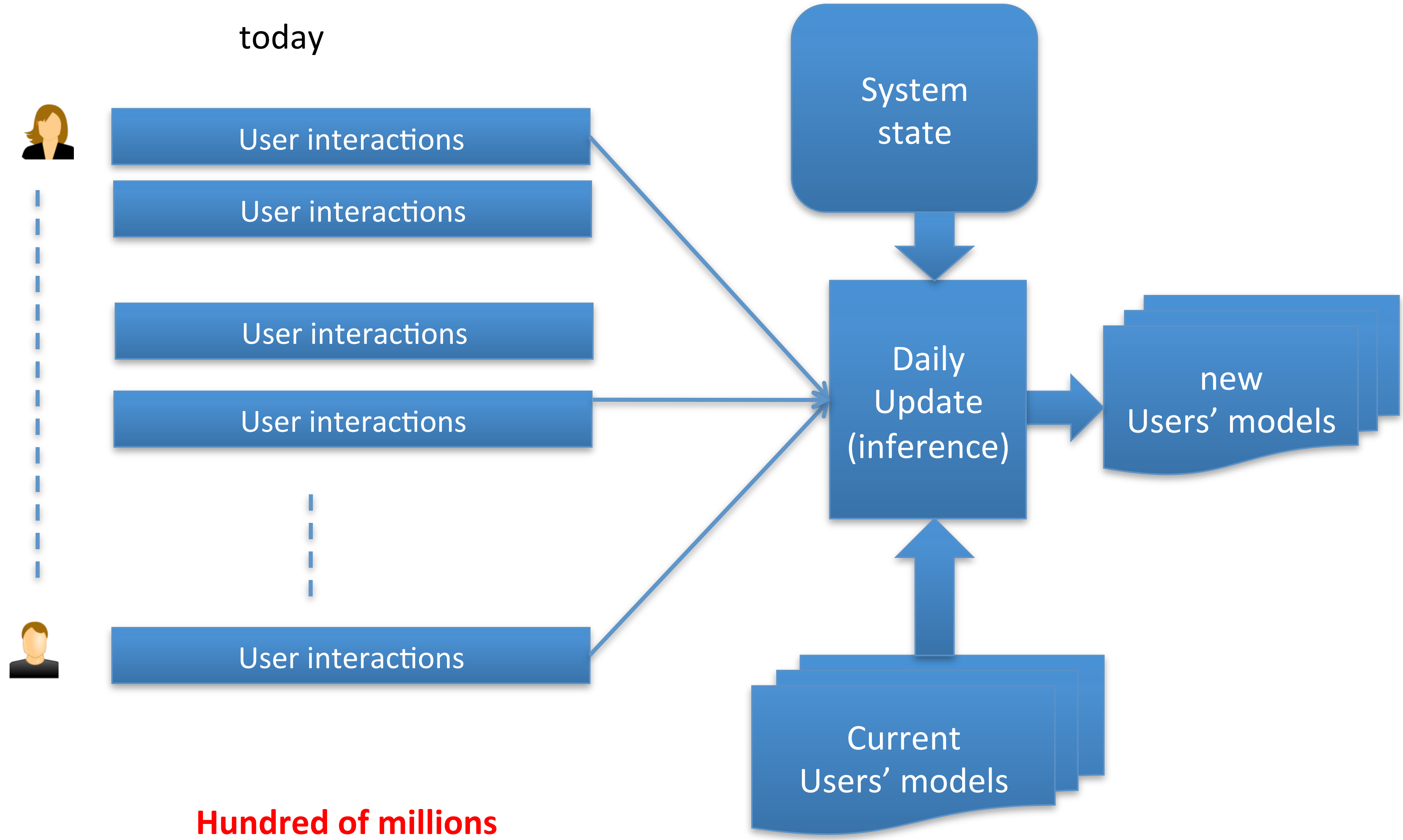
Capture long and term interests of users?


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

Online Distributed Inference

Work Flow

Work Flow

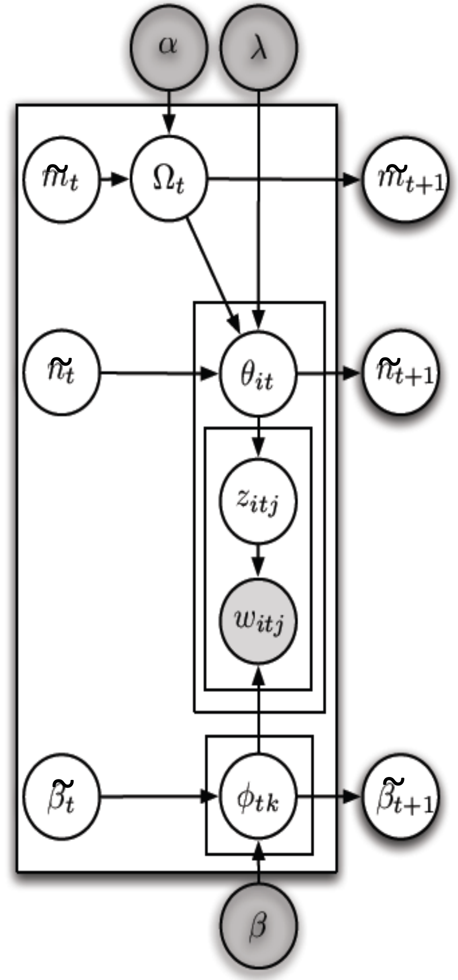

today
User interactions
User interactions
User interactions
User interactions
User interactions
System state
Daily Update (inference)
new Users' models
Current Users' models
Hundred of millions

Online Scalable Inference

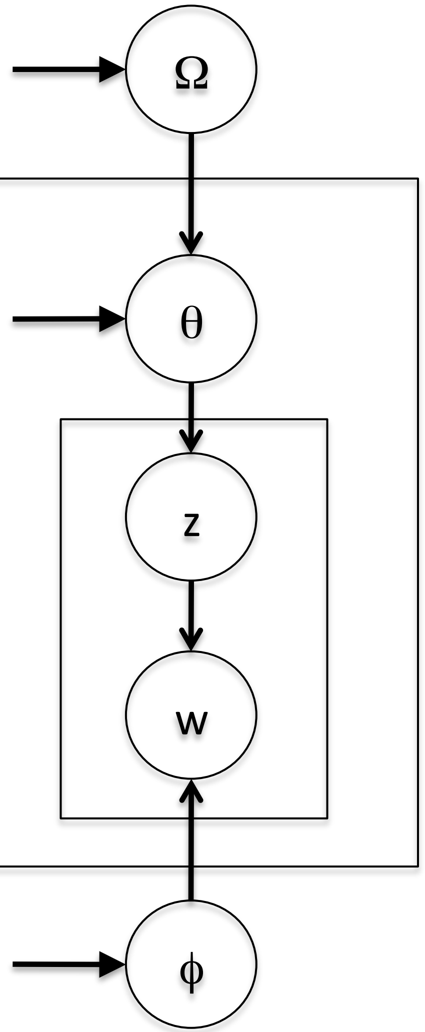
- Online algorithm
  - Greedy 1-particle filtering algorithm
  - Works well in practice
  - Collapse all multinomials except Ω_t
    - This makes distributed inference easier
  - At each time *t:*

$$P(\Omega^t, \mathbf{z}^t | \tilde{\mathbf{n}}^t, \tilde{\beta}^t, \tilde{\mathbf{m}}^t)$$

- Distributed scalable implementation
  - Used first part architecture as a subroutine
  - Added synchronous sampling capabilities

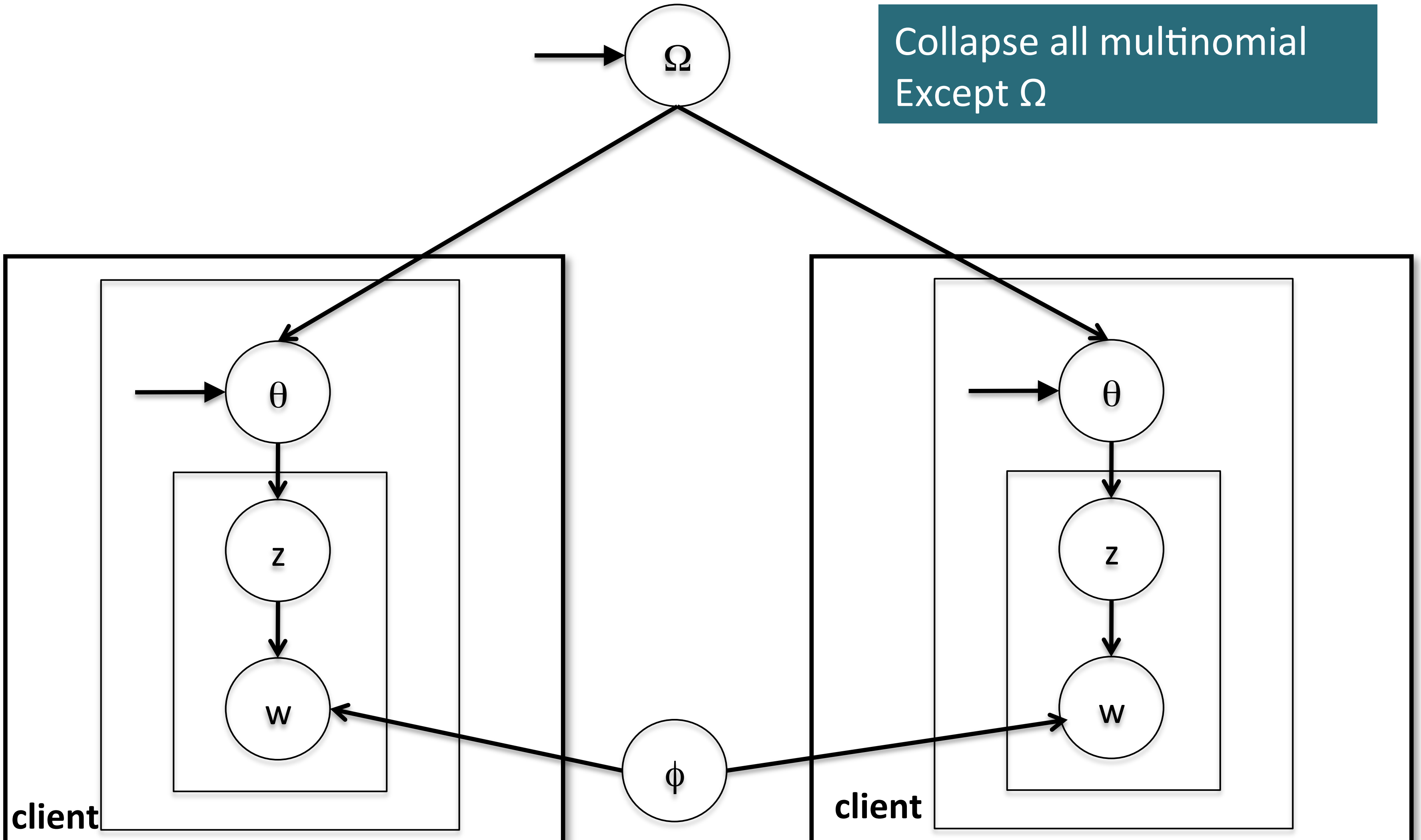


Distributed Inference (at time t)

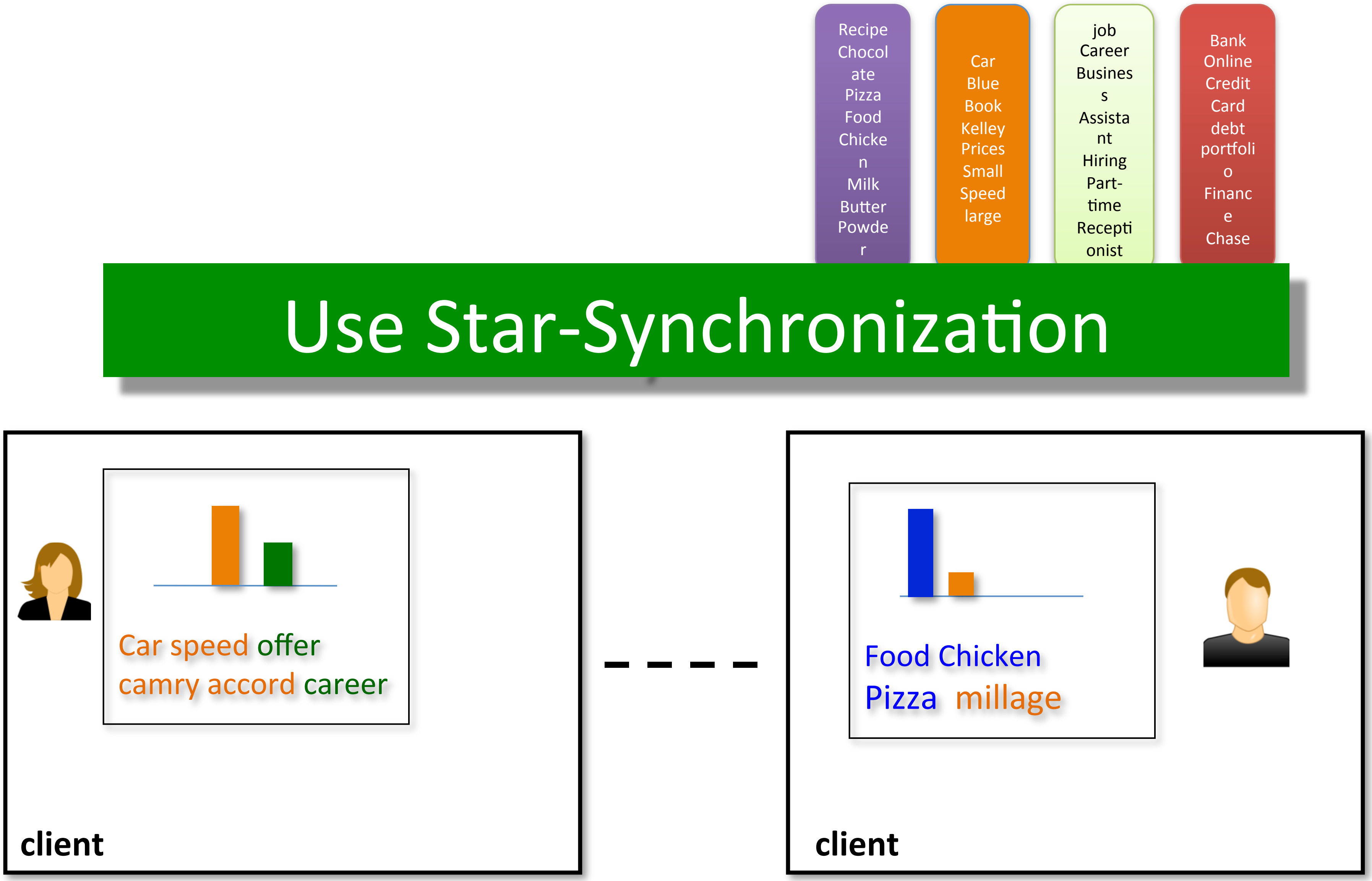


Distributed Inference (at time t)

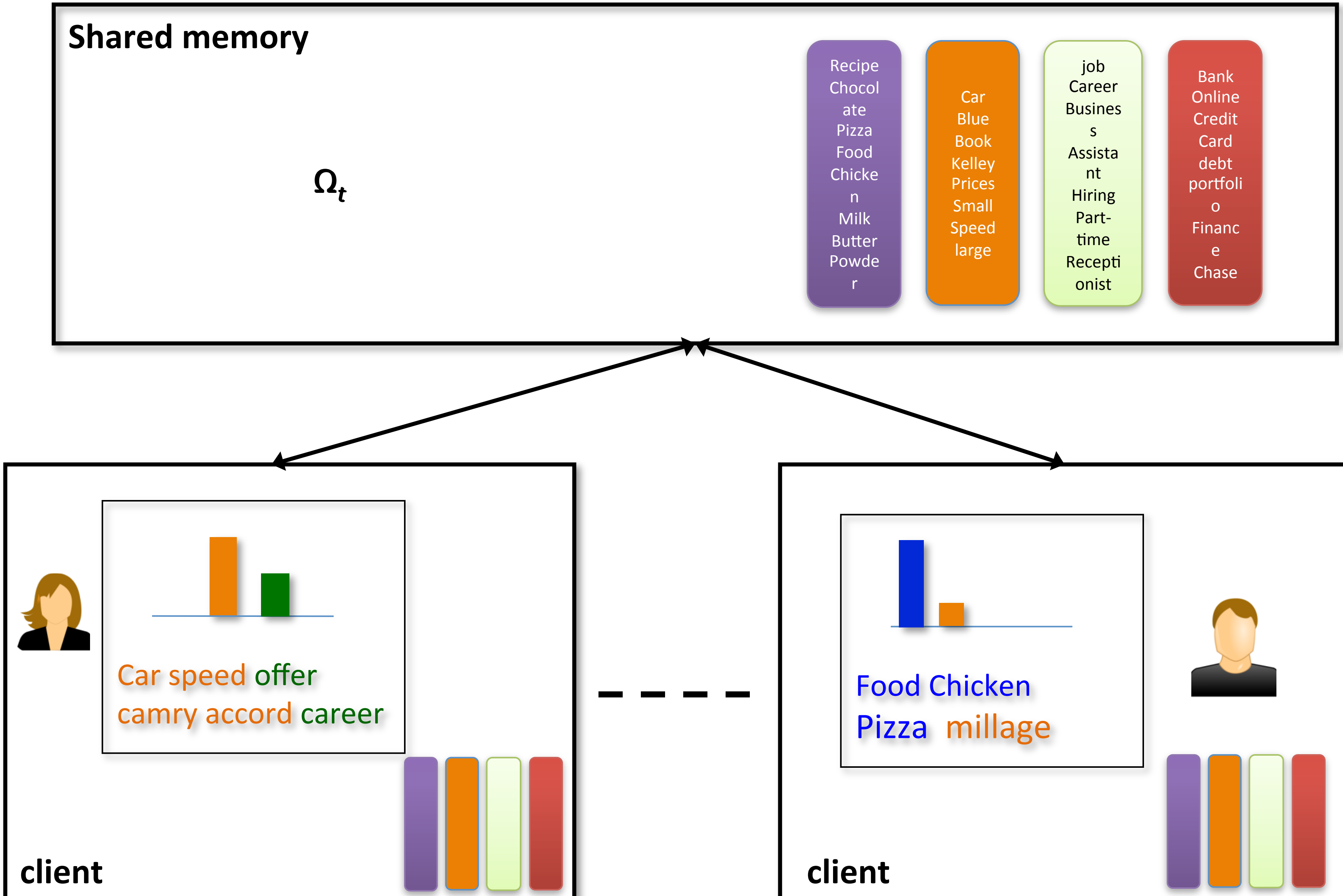
Collapse all multinomial
Except Ω



After collapsing



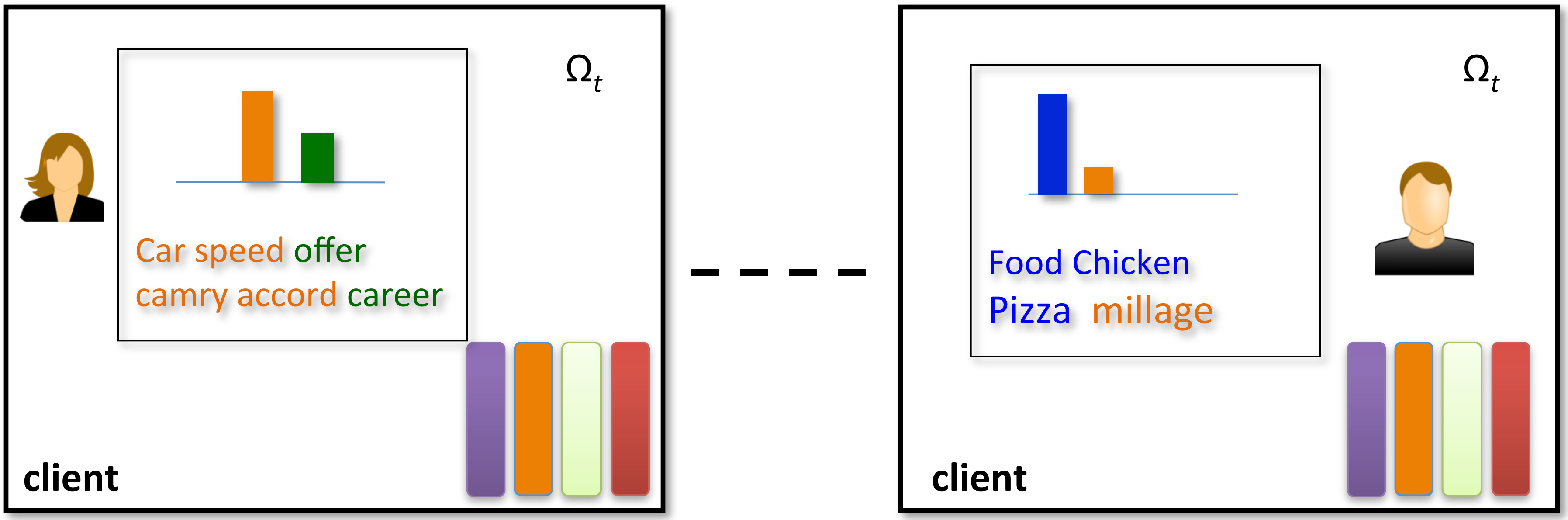
Fully Collapsed



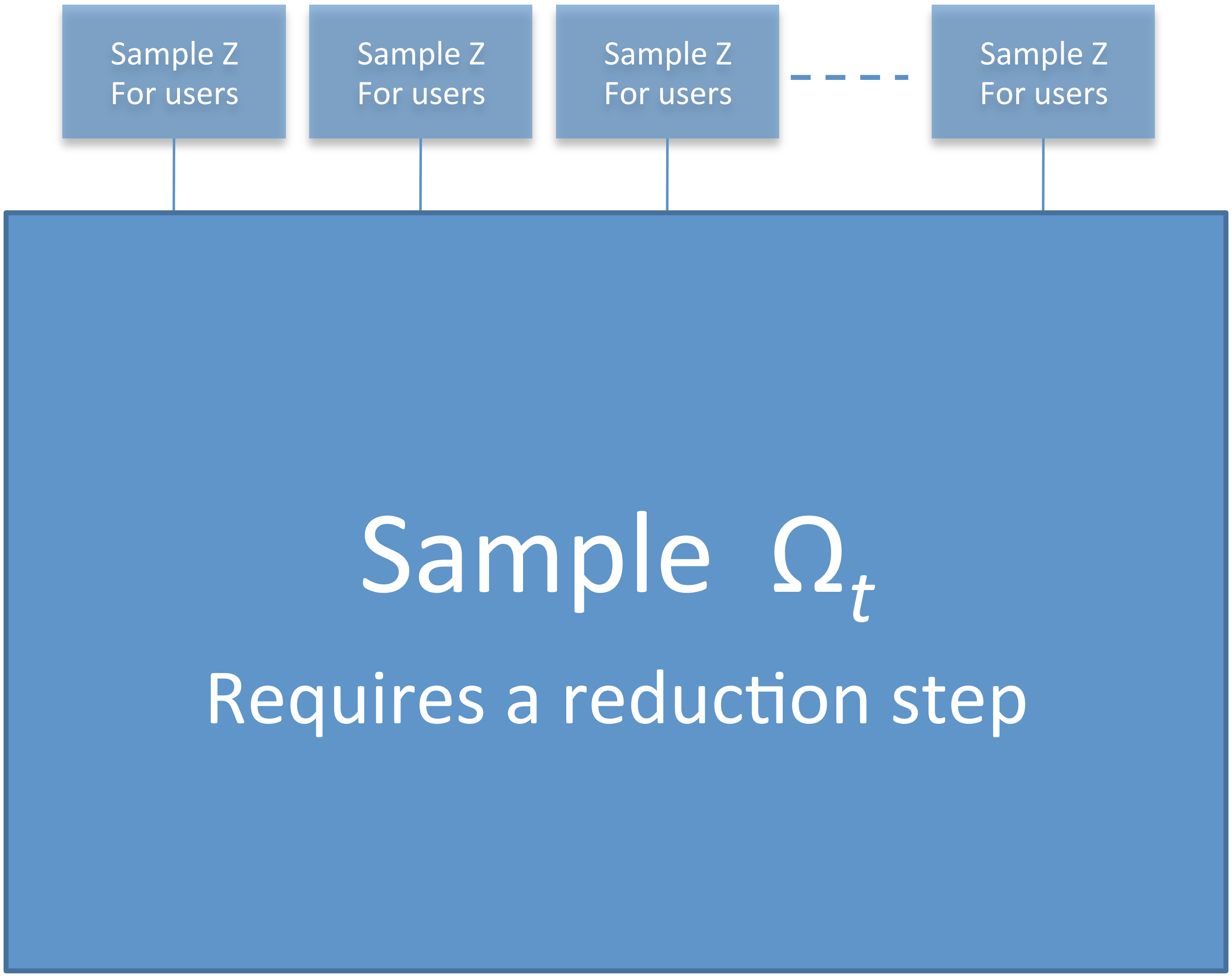
Semi-Collapsed

$$P(z_{ij}^t = k | w_{ij}^t = w, \Omega^t, \tilde{\mathbf{n}}_i^t)$$

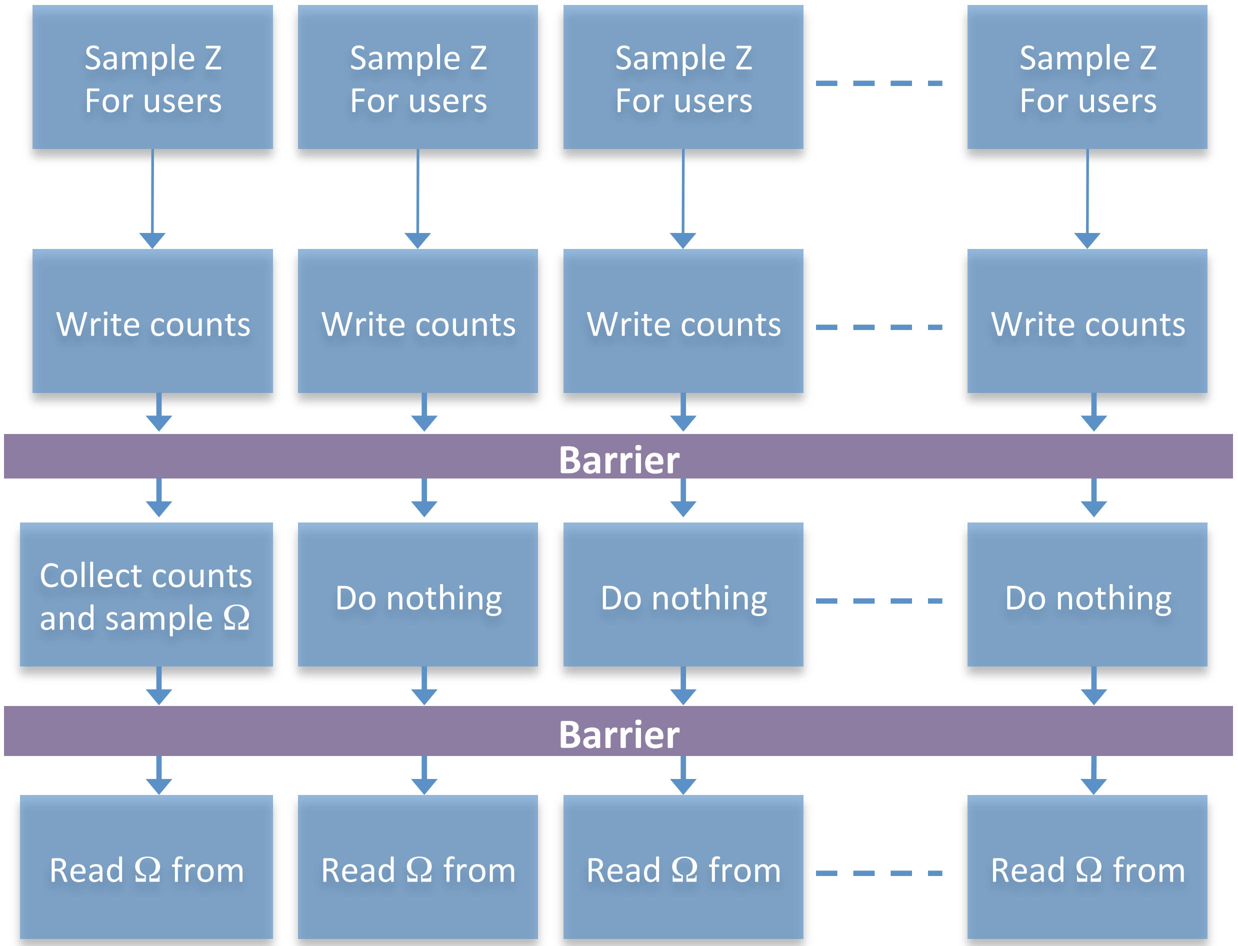
$$\propto \left(n_{ik}^{t,-j} + \tilde{n}_{ik}^t + \boxed{\lambda \Omega^t} \right) \frac{n_{kw}^{t,-j} + \tilde{\beta}_{kw}^t + \beta}{\sum_l n_{kl}^{t,-j} + \tilde{\beta}_{kl}^t + \beta}$$



Distributed Sampling Cycle



Distributed Sampling Cycle



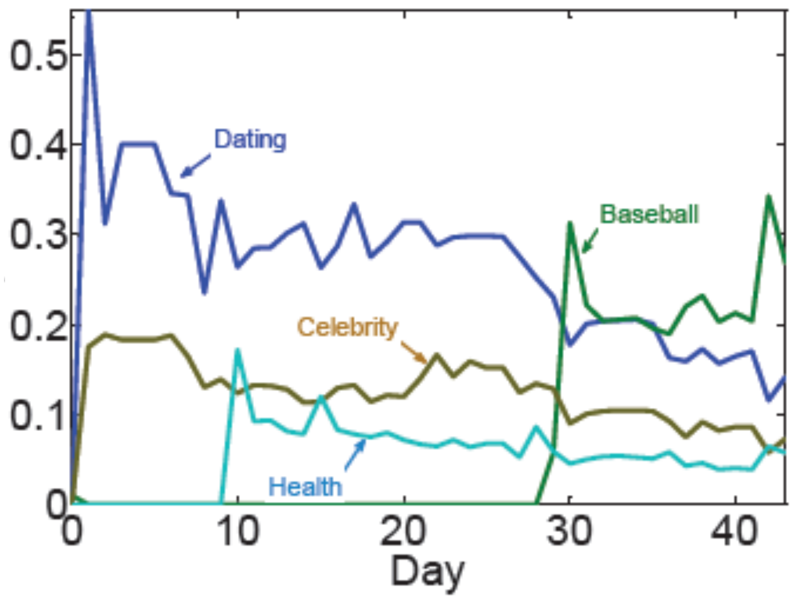
Experimental Results

- Tasks is predicting convergence in display advertising
- Use two datasets
  - 6 weeks of user history
  - Last week responses to Ads are used for testing
- Baseline:
  - User raw data as features
  - Static topic model

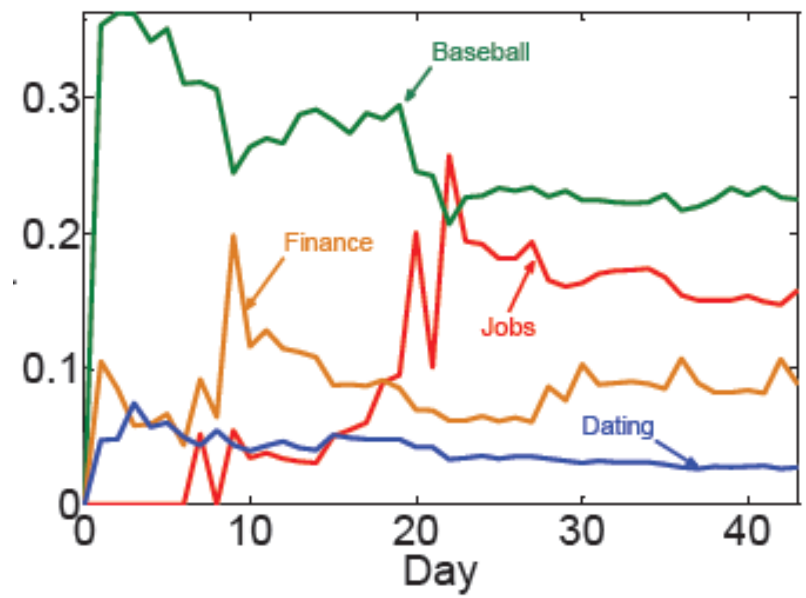
dataset	# days	# users	# campaigns	size
1	56	13.34M	241	242GB
2	44	33.5M	216	435GB

Interpretability

User-1



User-2



Dating

women
men
dating
singles
personals
seeking
match

Baseball

League
baseball
basketball,
doublehead
Bergesen
Griffey
bullpen
Greinke

Celebrity

Snooki
Tom
Cruise
Katie
Holmes
Pinkett
Kudrow
Hollywood

Health

skin
body
fingers
cells
toes
wrinkle
layers

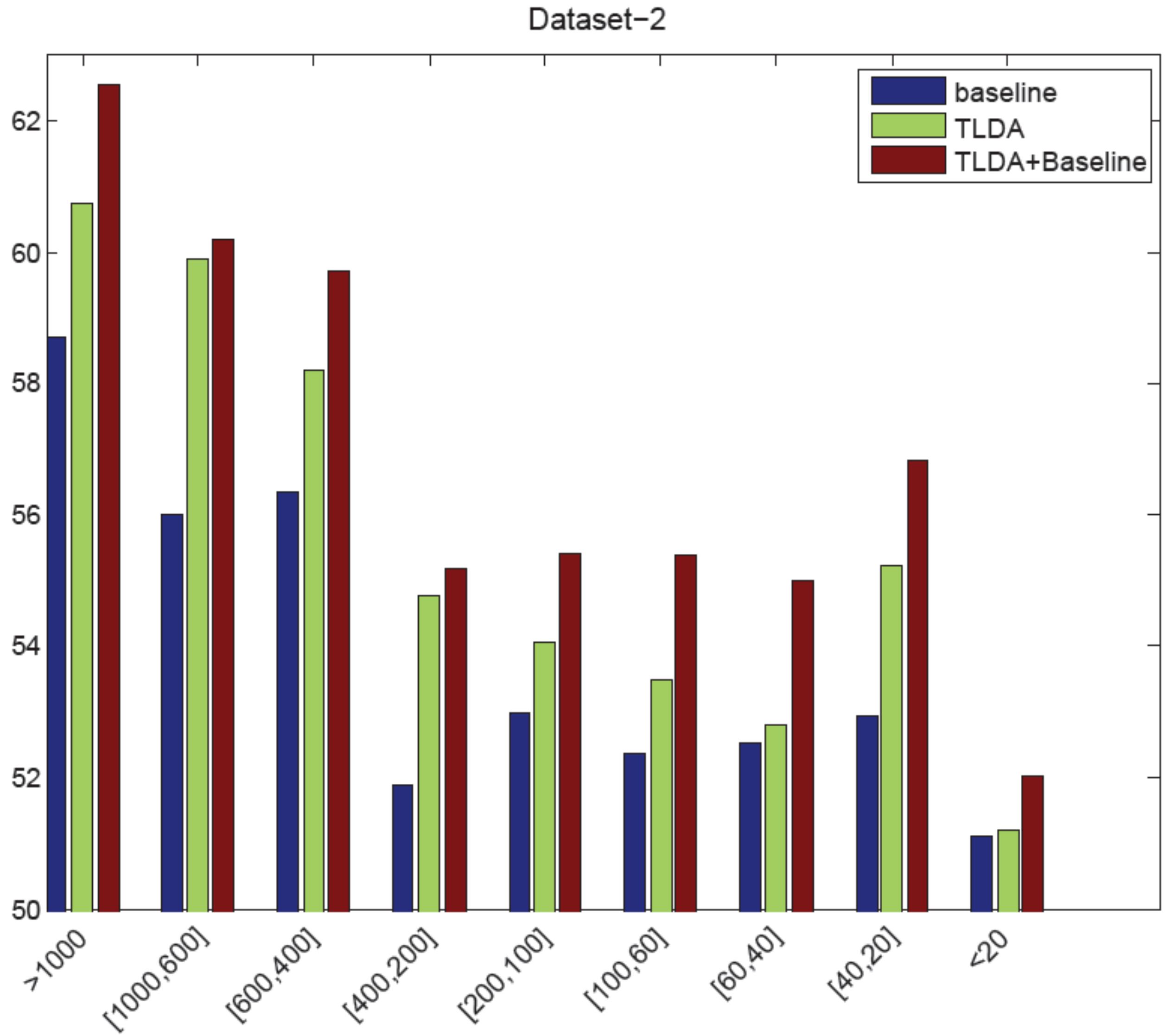
Jobs

job
career
business
assistant
hiring
part-time
receptionist

Finance

financial
Thomson
chart
real
Stock
Trading
currency

Performance in Display Advertising



ROC

Number of conversions

Performance in Display Advertising

Weighted ROC measure

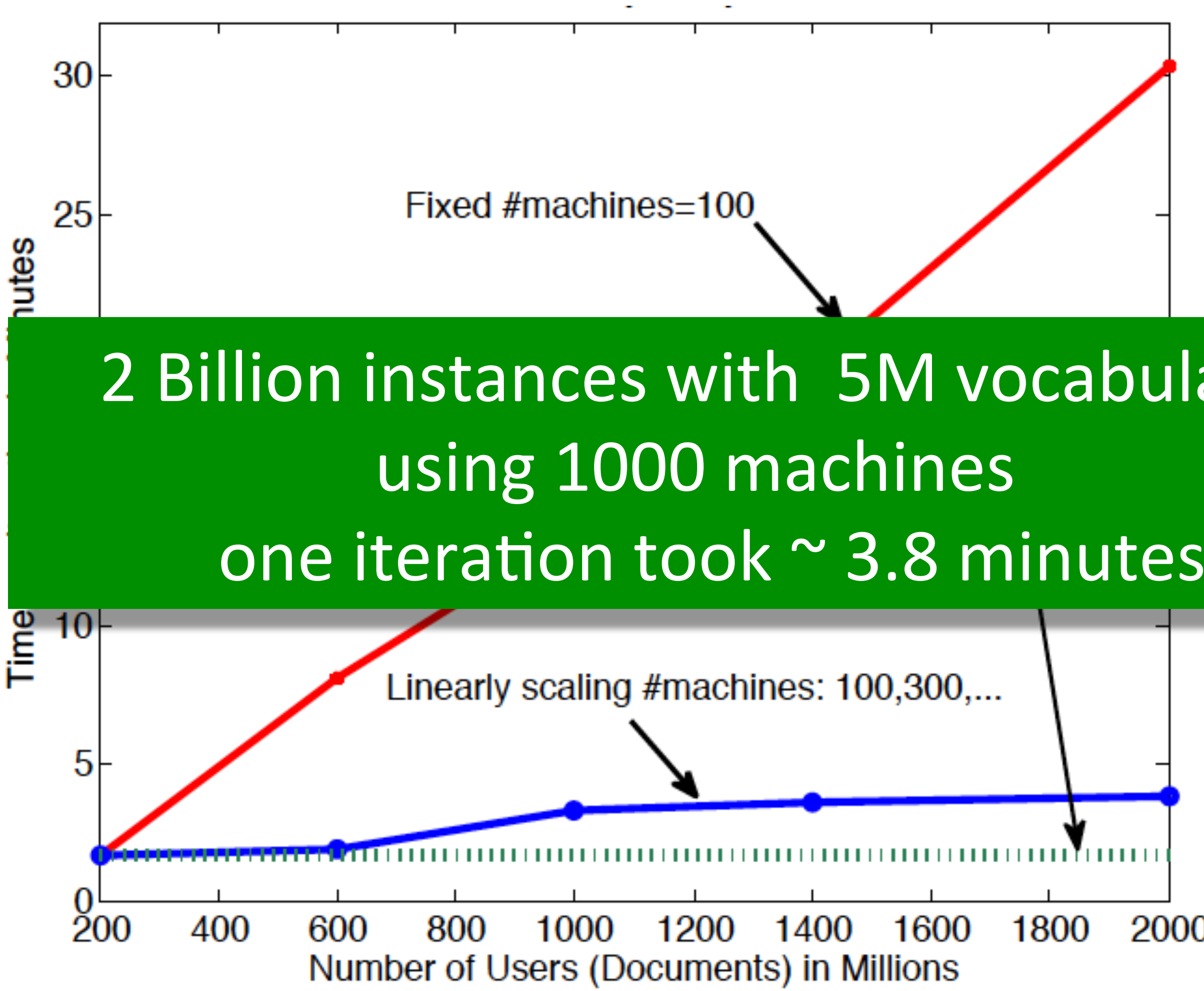
	base	TLDA	TLDA+base	LDA+base
dataset 1	54.40	55.78	**56.94**	55.80
dataset 2	57.03	57.70	**60.38**	58.54

Static
Batch models

Effect of number of topics

	topics	TLDA	TLDA + base
dataset 1	50	55.32	56.01
	100	55.5	56.56
	200	**55.8**	**56.94**
dataset 2	50	59.10	60.40
	100	**59.14**	**60.60**
	200	58.7	60.38

How Does It Scale?

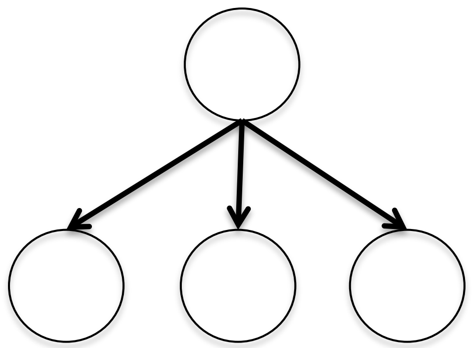


2 Billion instances with 5M vocabulary
using 1000 machines
one iteration took ~ 3.8 minutes

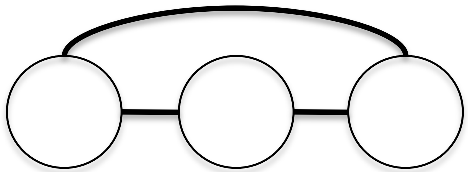
Distributed Inference Revisited

To collapse or not to collapse?

- Not collapsing
  - Keeps conditional independence
    - Good for parallelization
    - Requires synchronous sampling
  - Might mix slowly



- Collapsing
  - Mixes faster
  - Hinder parallelism
  - Use star-synchronization
    - Works well if sibling depends on each others via aggregates
    - Requires asynchronous communication



Inference Primitive

- Collapse a variable
  - Star synchronization for the sufficient statistics
- Sampling a variable
  - Local
    - Sample it locally (possibly using the synchronized statistics)
  - Shared
    - Synchronous sampling using a barrier
- Optimizing a variable
  - Same as in the shared variable case
  - Ex. Conditional topic models

Asynchronous vs. Synchronous Optimization

Synchronous Processing

- Needed when
  - Ex: Optimizing a global variable
- Mostly requires a barrier
- Advantages
  - Easy to program
  - Well-understood reusable templates
- Disadvantages
  - The curse of the last reducer
  - You are as fast as the slowest machine!

Synchronous Processing

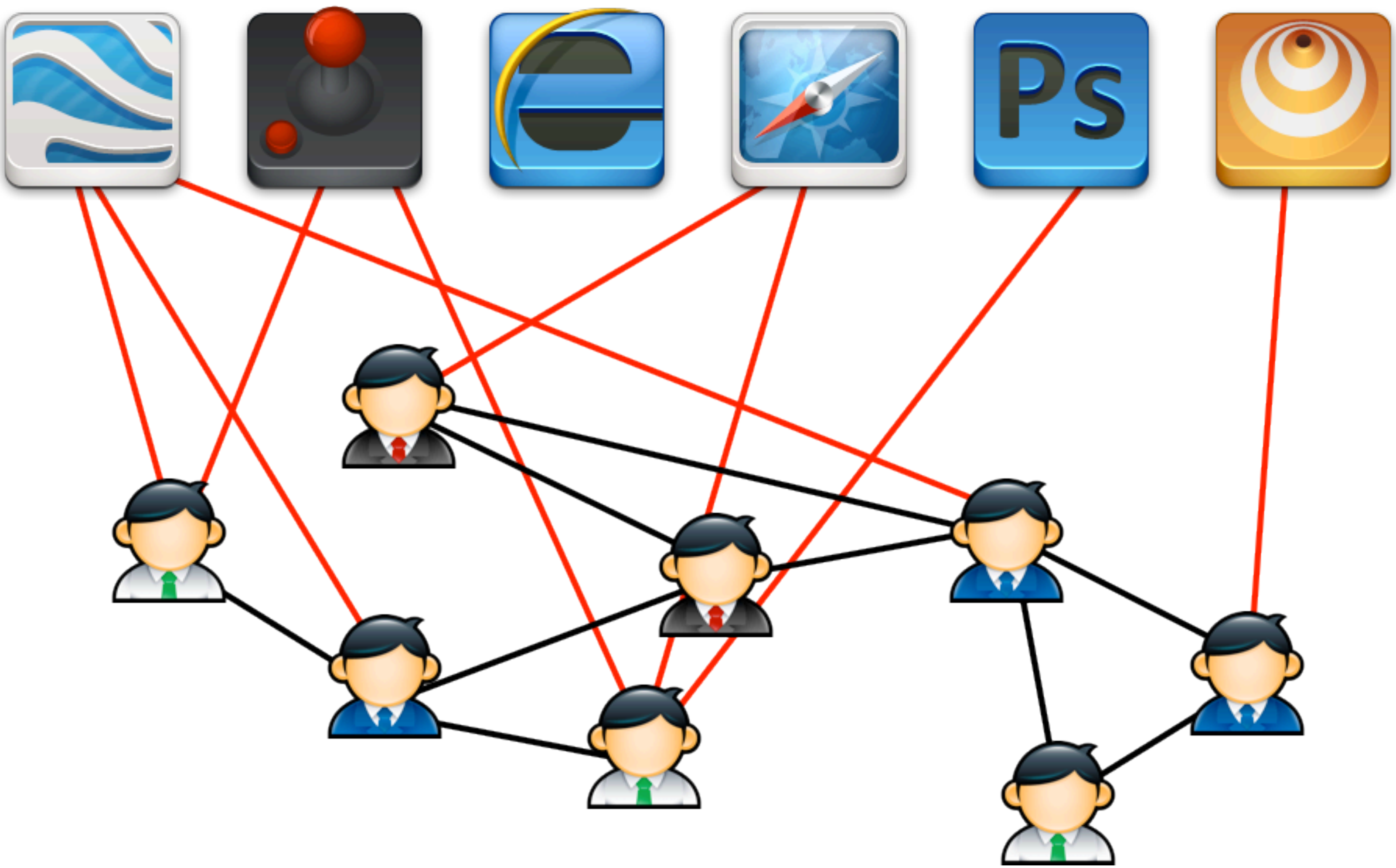
- Needed when
  - Ex: Optimize a global variab
- Mostly requires a barri
- Advantages
  - Easy to progr
  - Well-und sable template
- Disadv
  - The cu the last reducer
  - You are as fast as the slowest machine!

Can we do better?

Asynchronous Optimization

Graph Factorization

Graph Factorization: Social Network



Natural Graphs

- Social networks
  >1B vertices - Google+, Facebook, Twitter ...
- Mail graphs
  >200M vertices for slice of Yahoo Mail
- Language
  >1Mx10B vertices for (document,word) graph
- Computational advertising (ads, attributes)

Graph Factorization Problem

- Factor a graph into low rank components
- Assign a latent vector $Z_i \in \mathcal{R}^k$ with each node
- Optimize:

$$f(Y, Z, \lambda) = \frac{1}{2} \sum_{(i,j) \in E} \left(Y_{ij} - \langle Z_i, Z_j \rangle\right)^2 + \frac{\lambda}{2} \sum_i n_i \|Z_i\|^2$$

Observed value over edges

Predicted value

Regularization

Single-Machine Algorithm

- Just use stochastic gradient decent (SGD)

$$\frac{\partial f}{\partial Z_i} = - \sum_{j \in \mathcal{N}(i)} \big(Y_{ij} - \langle Z_i, Z_j \rangle\big) Z_j + \lambda n_i Z_i$$

- Cycle until convergence
  - Read a node, *i*
  - Update its latent factor

$$Z_i \leftarrow Z_i - \eta \big(\frac{\partial f}{\partial Z_i} \big)$$

Problem Scale

- Yahoo IM and Mail graphs
- Nodes are users
- Edges represent (log) number of messages
- 200 Million vertices
- 10 Billion edges

Challenges

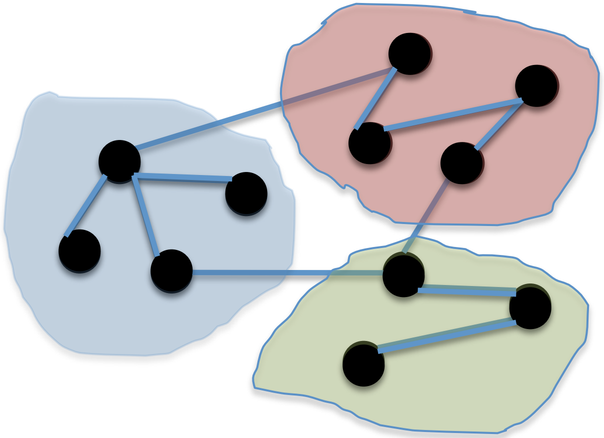
- Parameter storage
  - Too much for a single machine
- Approach
  - Distribute the graph over machines
    - How to partition the nodes?
  - Synchronization
    - How to synchronize replicated nodes
  - Communication
    - How to accommodate network topology

Challenges

Can we solve the problem with similar ideas to what we have covered?

Formulation as a Consensus Problem

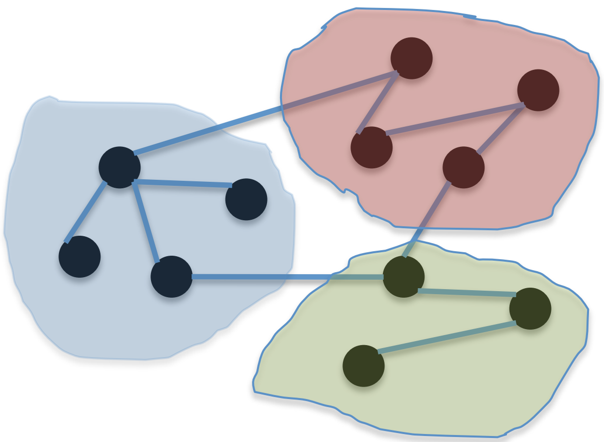
Partition and Replicate

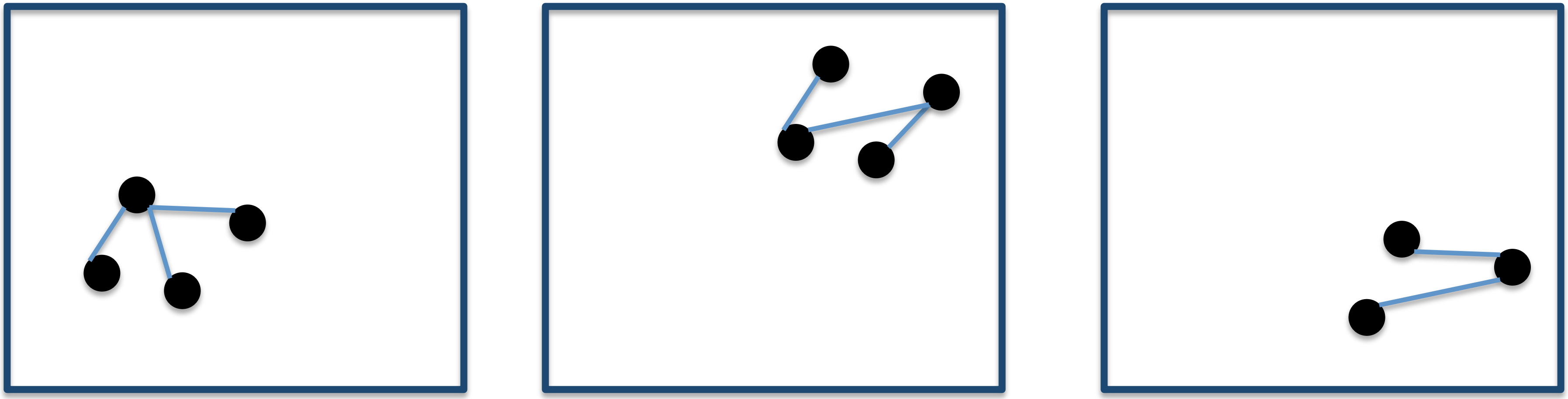


Partition and Replicate

- Cycle until convergence
  - Read a node, *i*
  - Update its latent factor

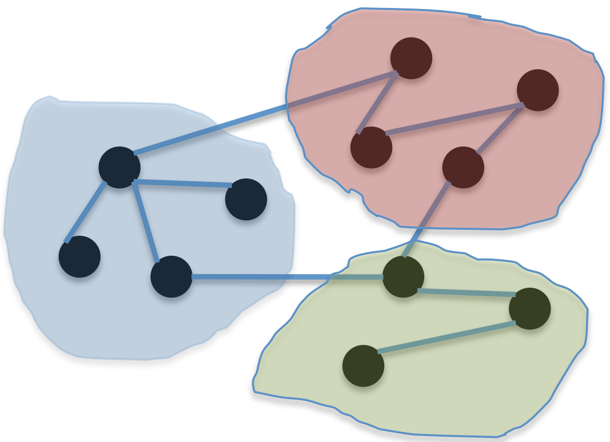
$$Z_i \leftarrow Z_i - \eta\left(\frac{\partial f}{\partial Z_i}\right)$$

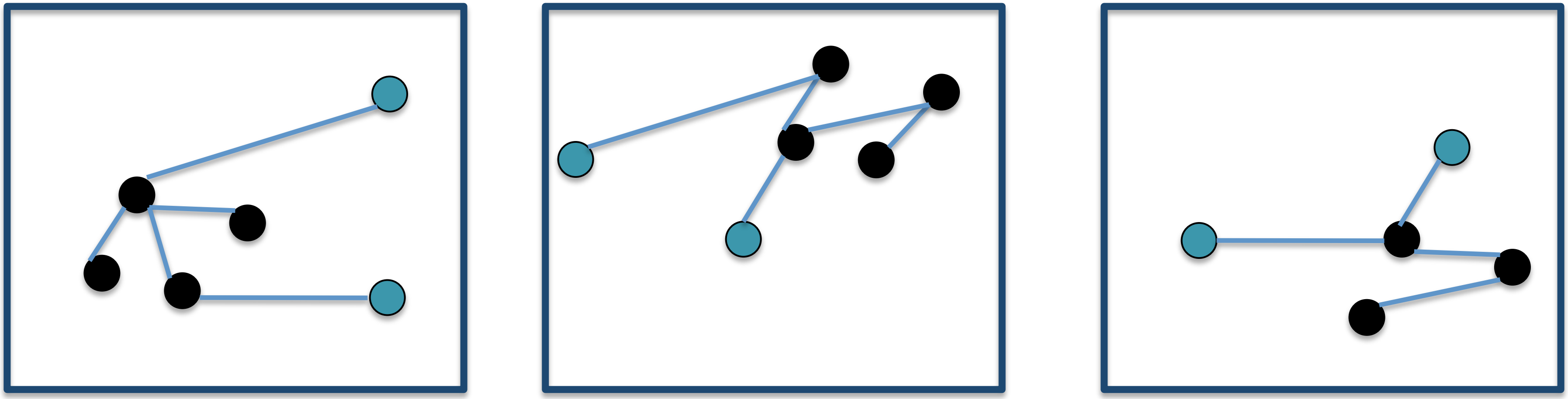




Partition and Replicate

- Problem
  - Some neighbors are missing
- Solution
  - Replicate and synchronize
  - **Borrowed** vs. owned nodes





Consensus Formulation

- Original problem

$$f(Y,Z,\lambda) = \frac{1}{2}\sum_{(i,j)\in E}\left(Y_{ij} - \langle Z_i, Z_j\rangle\right)^2 + \frac{\lambda}{2}\sum_i n_i\|Z_i\|^2$$

- Relaxed problem

$$\sum_{k=1}^{K} f_k(Y, X^{(k)}, \lambda) + \frac{1}{2}\sum_{k=1}^{K}\left[\mu\sum_{i\in V_k}\|Z_i - X_i^{(k)}\|^2\right]$$

Global factor

Local factors

Deviation

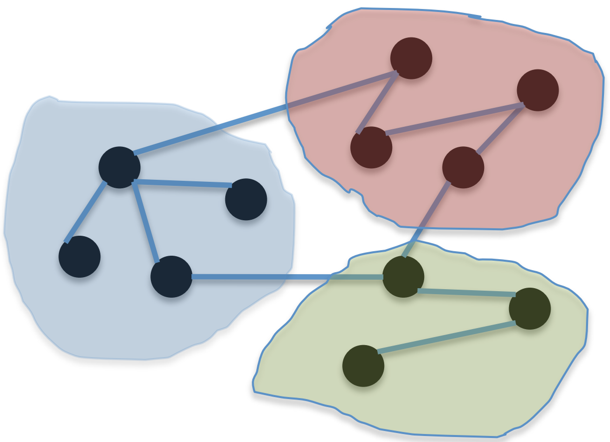
- Local problem

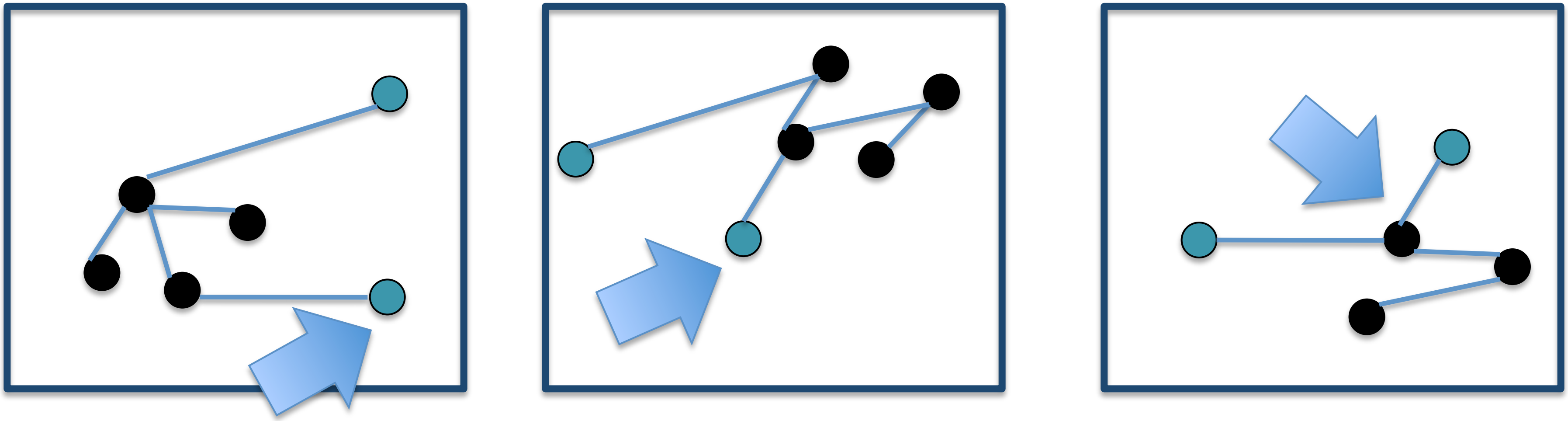
$$f_k(Y, X^{(k)}, \lambda)$$

$$= \frac{1}{2}\left[\sum_{\substack{(i,j)\in E,\\ i,j\in V_k}}\left(Y_{ij} - \langle X_i^{(k)}, X_j^{(k)}\rangle\right)^2 + \lambda\sum_{i\in V_k} n_i\|X_i^{(k)}\|^2\right]$$

Partition and Replicate

- Formulation
  - Introduce local copies
    - A factor per node X
  - Tie across machines
    - Introduce global factor Z
    - Penalizes deviations





Synchronous Optimization

Synchronous Algorithm

- Optimize joint objective over X,Z
- Local parameter updates
  - Run SGD until convergence

$$\text{minimize}_{X^{(k)}} f_k(Y, X^{(k)}, \lambda) + \frac{1}{2}\mu \sum_{i \in V_k} \|Z_i - X_i^{(k)}\|^2$$

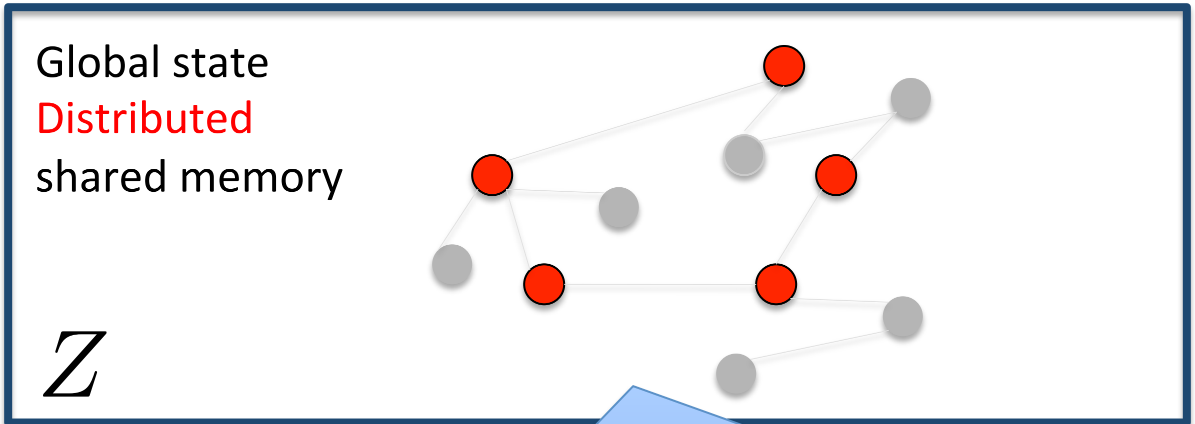
Fit the data

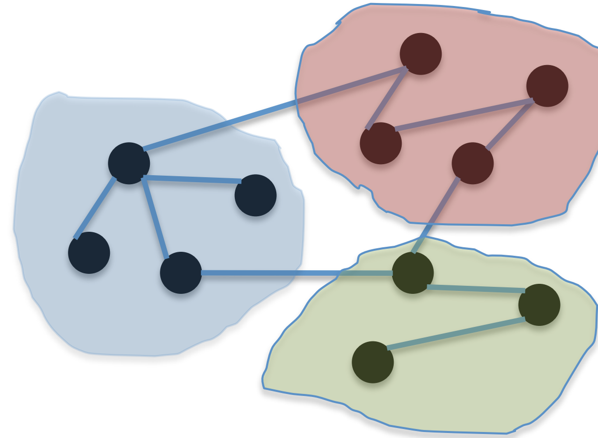
Minimize deviation

- Global parameter updates

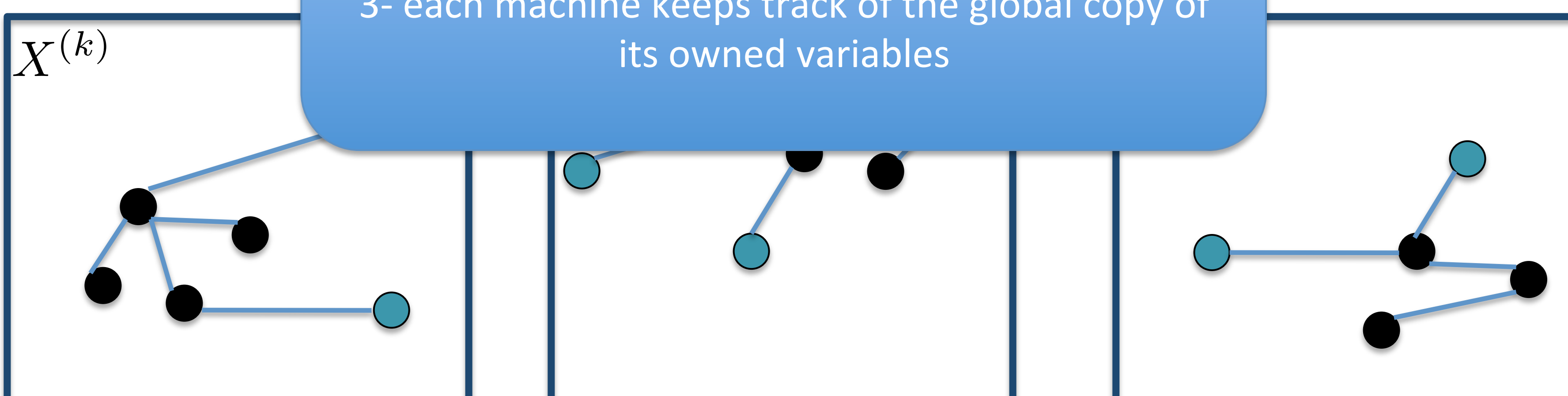
$$\text{minimize}_Z \quad \frac{1}{2} \sum_{k=1}^{K} \left[\mu \sum_{i \in V_k} \|Z_i - X_i^{(k)}\|^2 \right]$$

Synchronous Algorithms

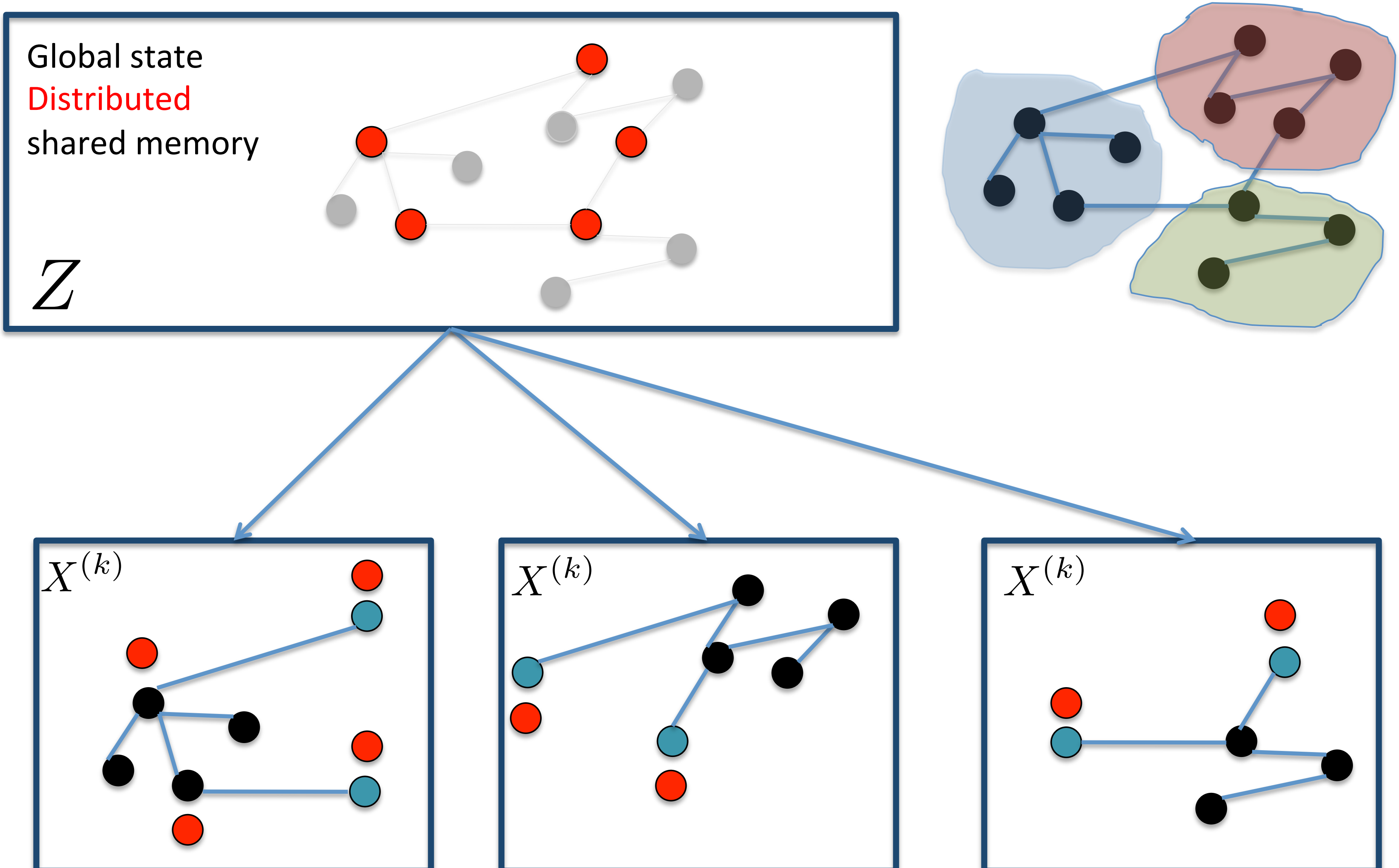




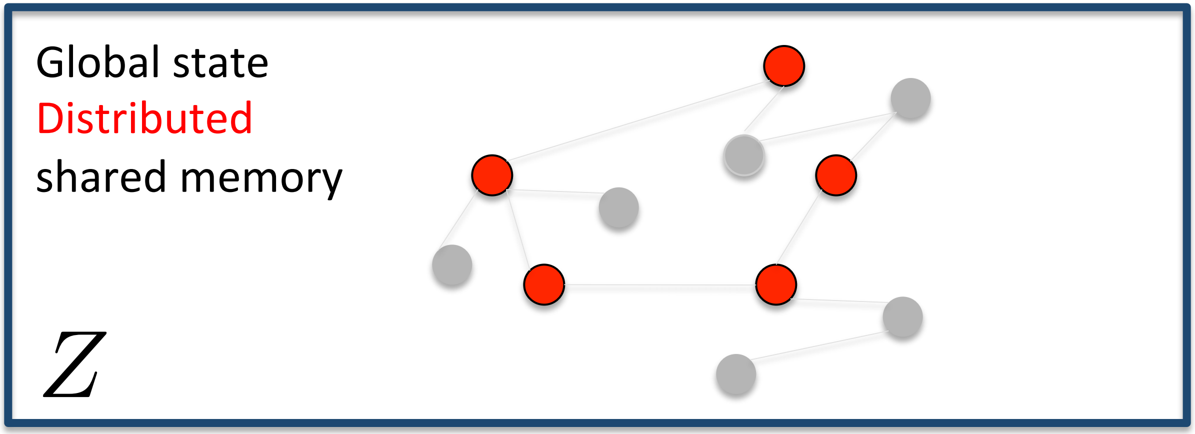
1- We only store replicated nodes
2- The global state is distributed across machines
3- each machine keeps track of the global copy of its owned variables

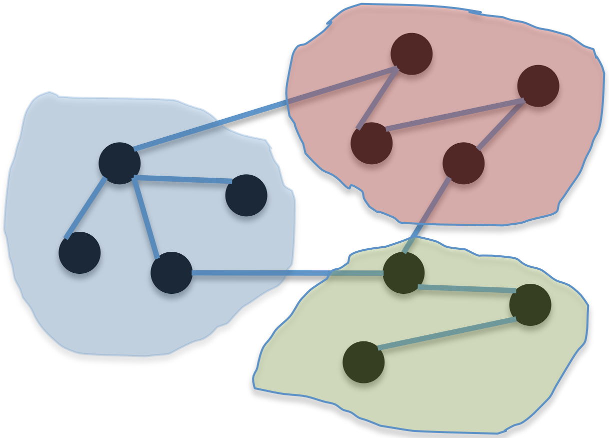


Step 1: Push global variables

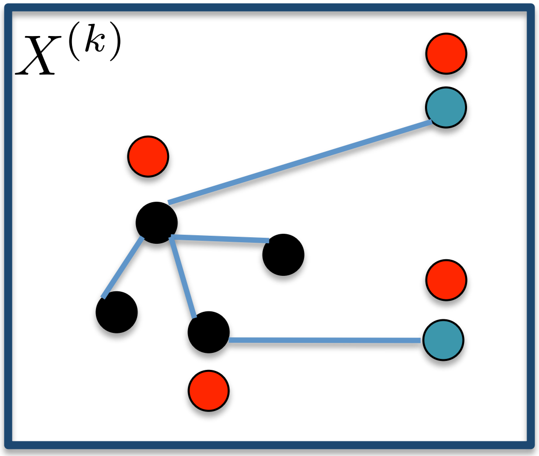


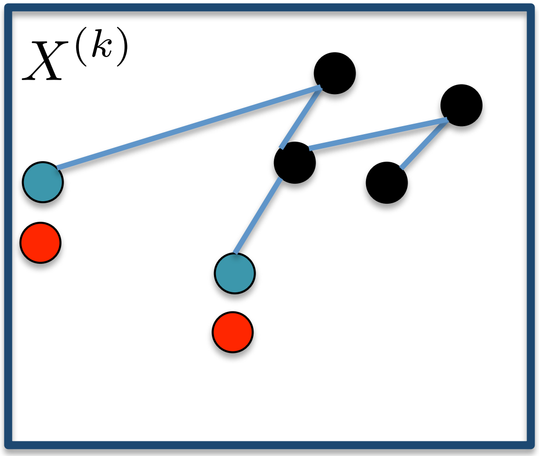
Step 2: Local Optimization

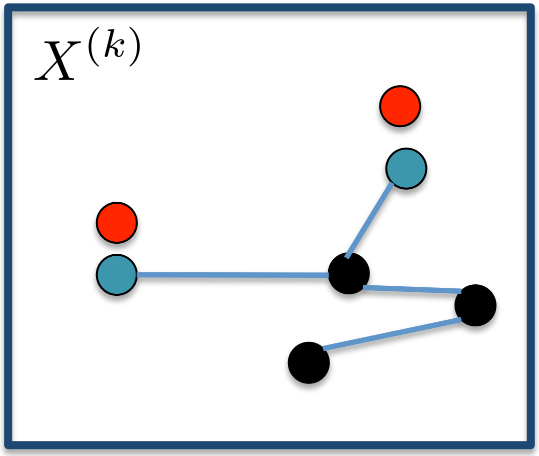




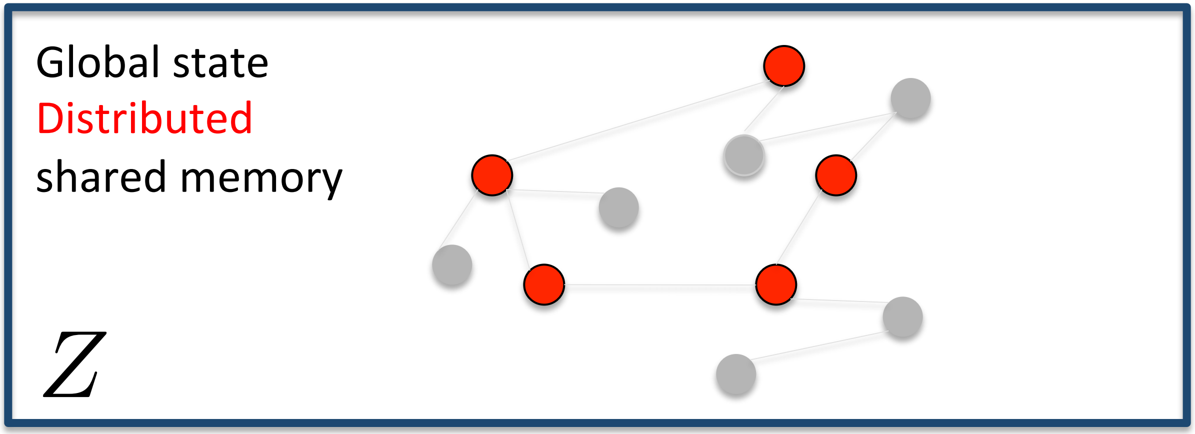
$$\text{minimize}_{X^{(k)}} f_k(Y, X^{(k)}, \lambda) + \frac{1}{2}\mu \sum_{i \in V_k} \|Z_i - X_i^{(k)}\|^2$$

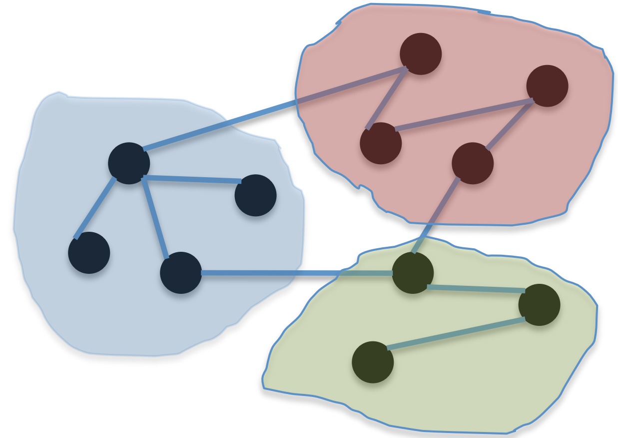




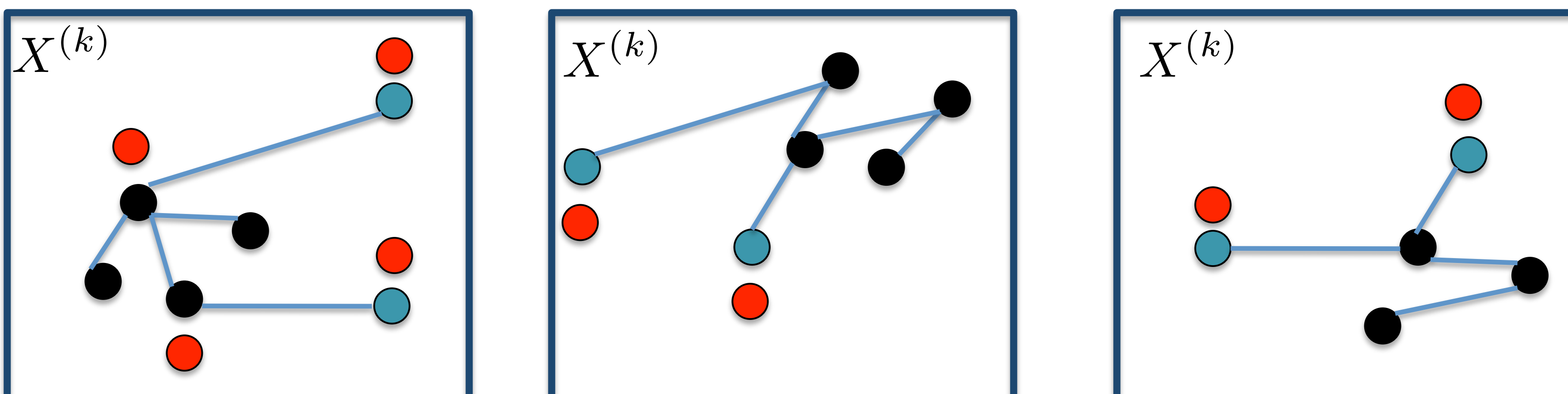


Step 3: Push and average

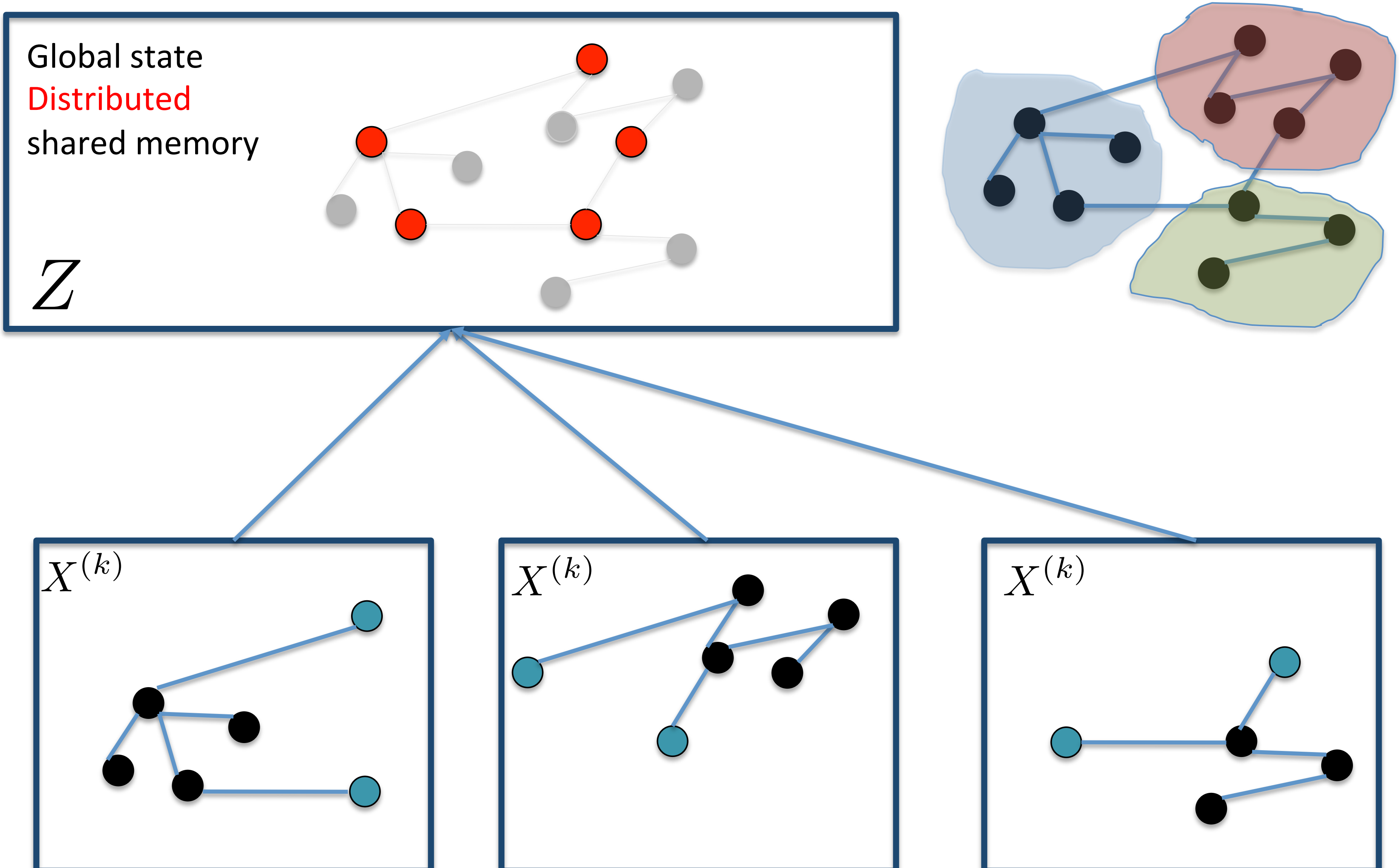




$$\text{minimize}_Z \quad \frac{1}{2}\sum_{k=1}^{K}\left[\mu \sum_{i \in V_k} \|Z_i - X_i^{(k)}\|^2\right]$$



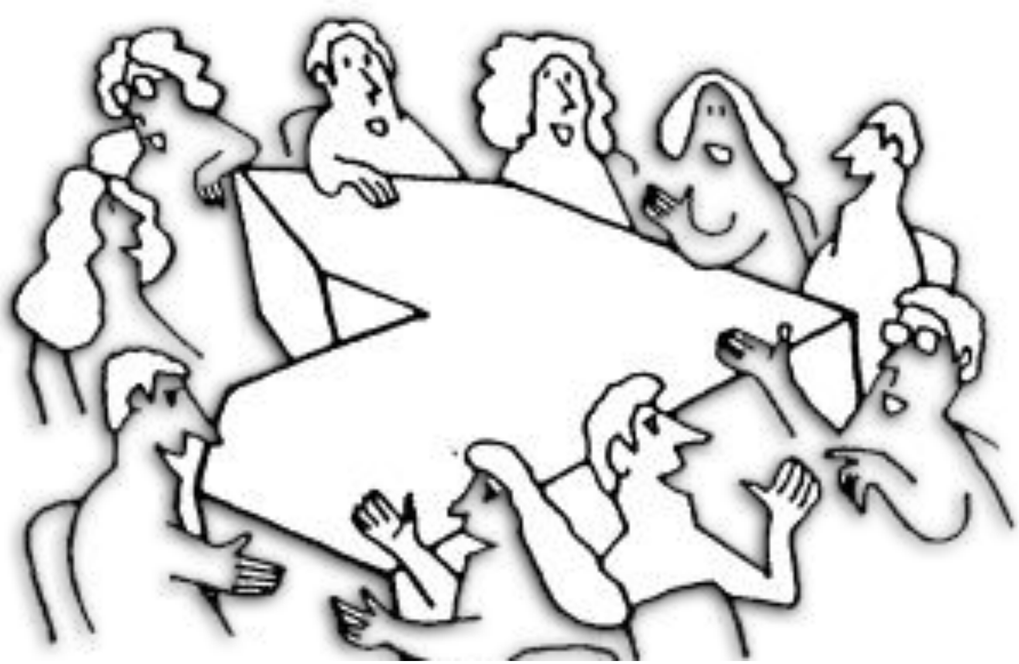
Step 3: Push and average



Summary of Synchronous Algorithm

- An **improvement** over standard Map-Reduce
- Curse of the last reducer
- You are as fast as the slowest machine
  - Optimize local variables
  - Barrier
  - Optimize global variables
  - Barrier
- **Can we do better?**

Asynchronous Optimization



An Asynchronous Algorithm

- Conceptual idea
  - Optimize X and Z jointly

$$\sum_{k=1}^{K} f_k(Y, X^{(k)}, \lambda) + \frac{1}{2}\sum_{k=1}^{K}\left[\mu \sum_{i \in V_k} \|Z_i - X_i^{(k)}\|^2\right]$$

- User SGD over (X,Z)
- Pick a local node
- Do a gradient step over corresponding X,Z!

Conceptual Idea

$$\sum_{k=1}^{K} f_k(Y, X^{(k)}, \lambda) + \frac{1}{2}\sum_{k=1}^{K}\left[\mu \sum_{i \in V_k} \|Z_i - X_i^{(k)}\|^2\right]$$

$$\frac{\partial f}{\partial Z_i}\left[X_i^{(k)}\right] = \mu(Z_i - X_i^{(k)}).$$

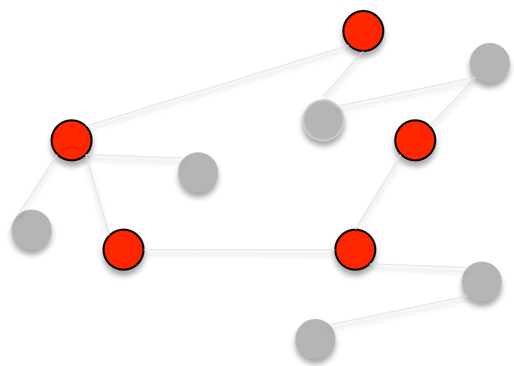
Cache the global variables
Locally (Asynchronous updates)

We don't have global copy locally

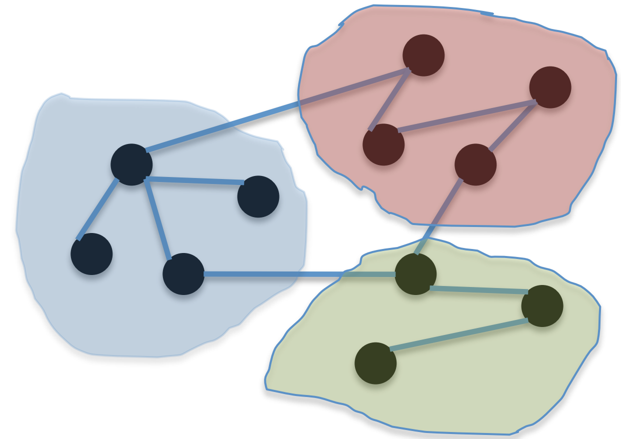
$$+ \lambda n_i X_i^{(k)} + \mu(X_i^{(k)} - Z_i).$$

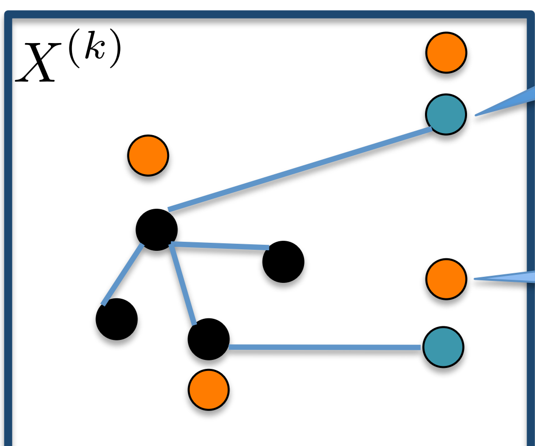
Parallel Updates

Global state
Distributed
shared memory



Z



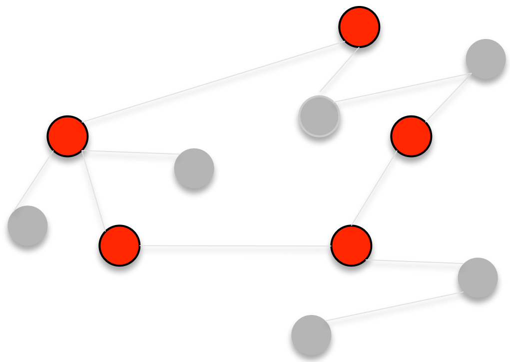


Indicate A borrowed node
Form other partitions

Last cached value of the
global variable

Parallel Asynchronous Updates

Global state
Distributed
shared memory



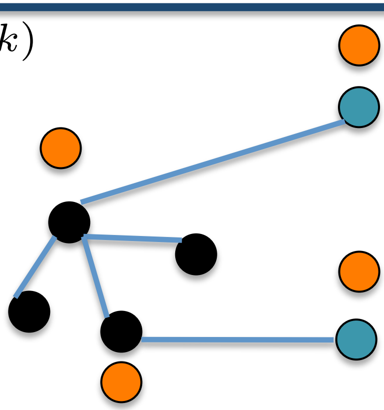
Z

-Receive local copy X_i from k
  -Update Z_i
  -Send back new Z_i to k

$$\frac{\partial f}{\partial Z_i}\left[X_i^{(k)}\right] = \mu(Z_i - X_i^{(k)}).$$

$$\frac{\partial f}{\partial X_i^{(k)}} = -\sum_{j \in N(i)} \left(Y_{ij} - \langle X_i^{(k)}, X_j^{(k)} \rangle\right) X_j^{(k)} + \lambda n_i X_i^{(k)} + \mu(X_i^{(k)} - Z_i^{(k)}).$$

$X^{(k)}$



-Cycle through nodes
-Update local copies

Computation thread

Synchronization thread Send

-Cycle through nodes
  - Send local copy to DSM

-Received Z_i from DSM
  - update cached copy

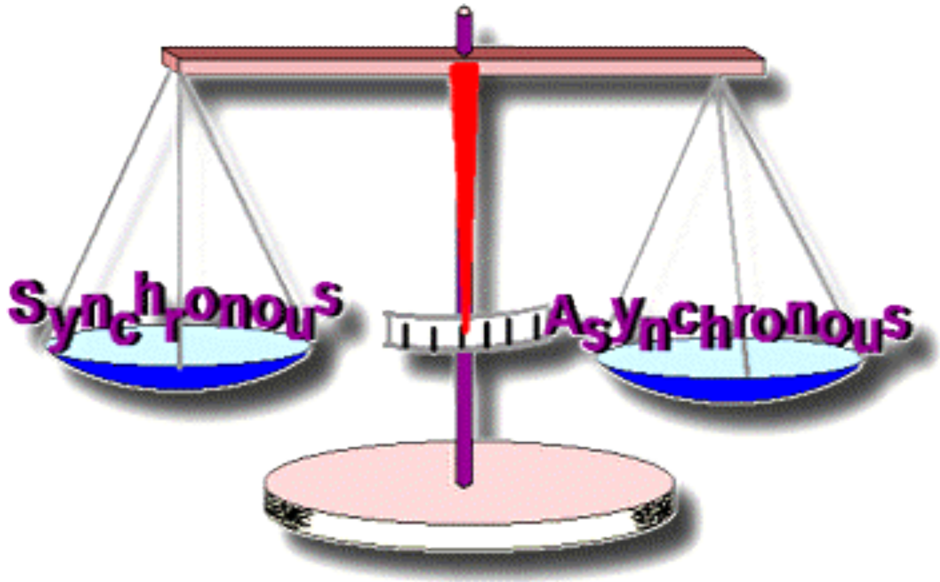
Synchronization thread receive

Summary of Asynchronous

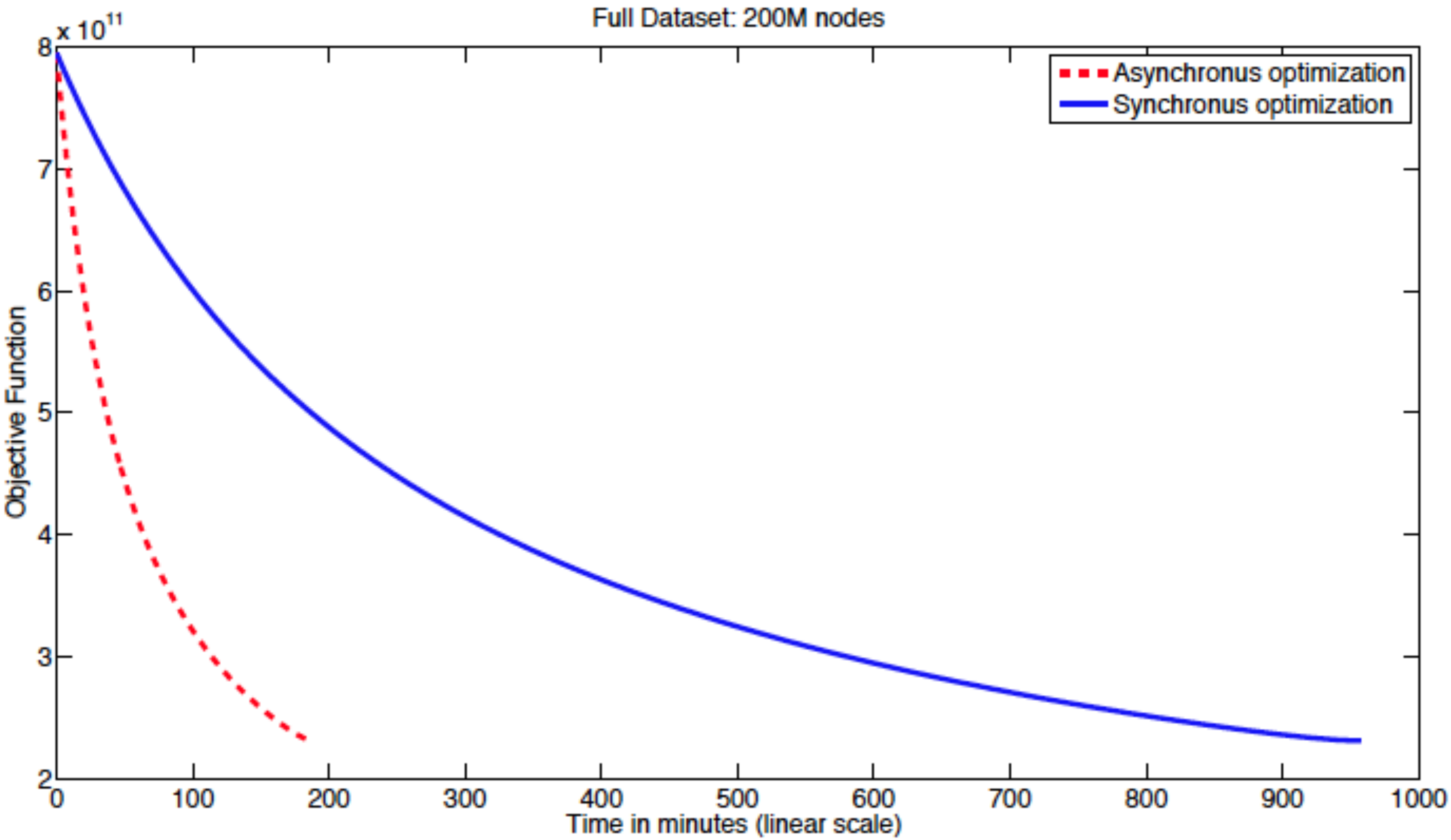
- Continuously update local variables X (via SGD)
- Continuously send local variables to global
- Continuously update global variable Z (via SGD)
- Continuously send & overwrite global variables to local

$$\sum_{k=1}^{K} f_k(Y, X^{(k)}, \lambda) + \frac{1}{2} \sum_{k=1}^{K} \Big[\mu \sum_{i \in V_k} \| Z_i - X_i^{(k)} \|^2 \Big]$$

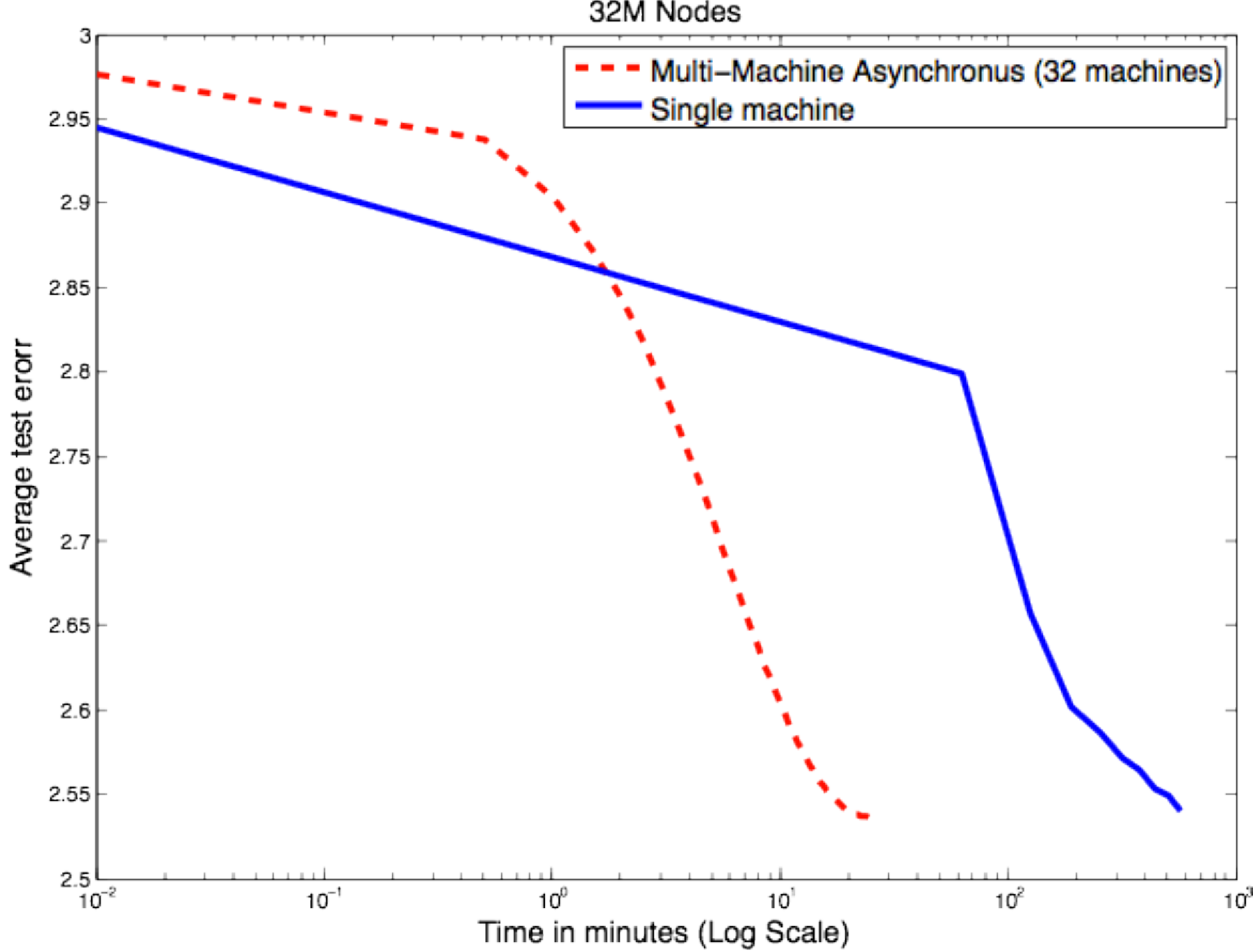
How Does it work?



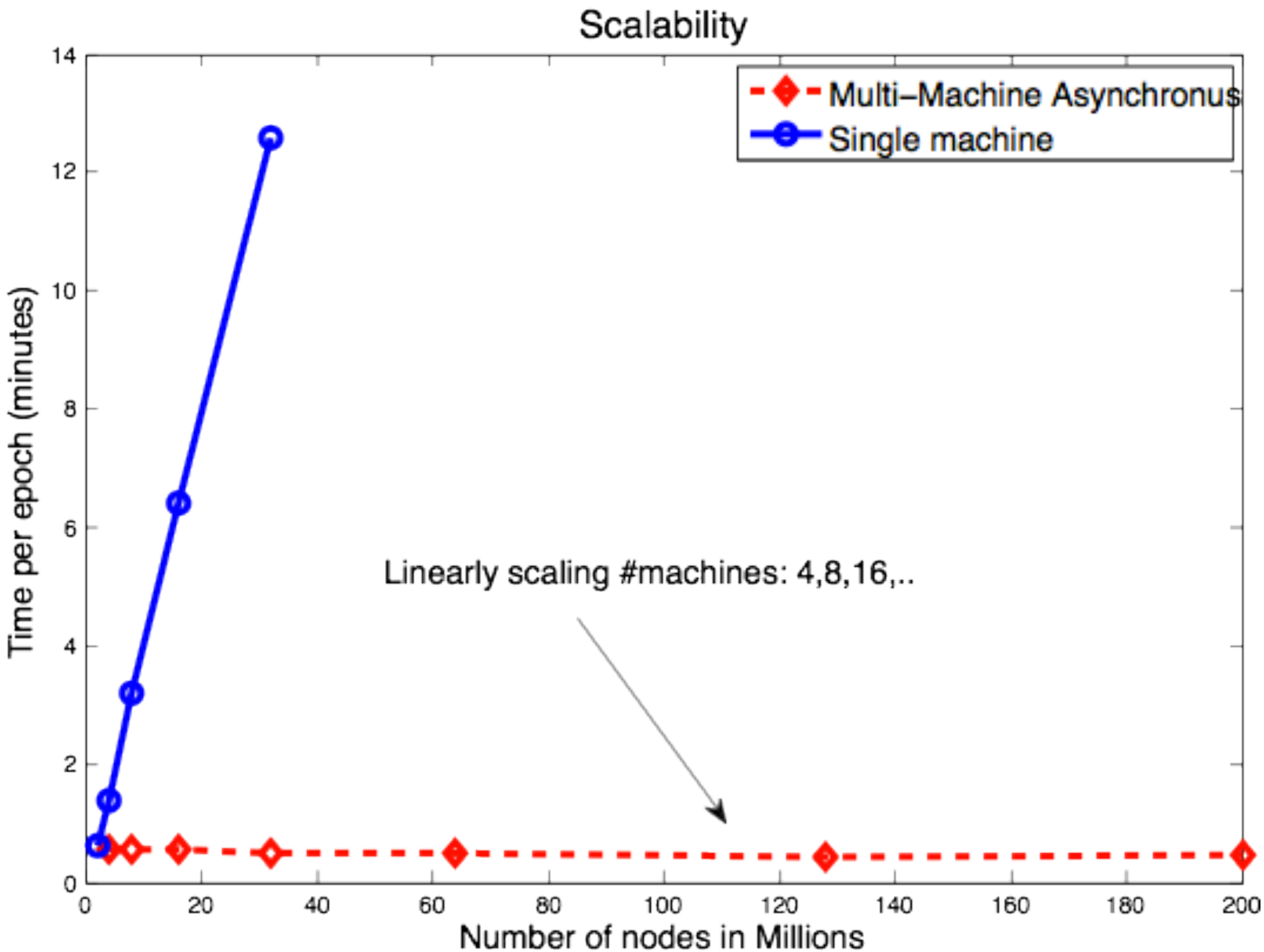
Sync Vs. Async.



Solution Quality



Scalability



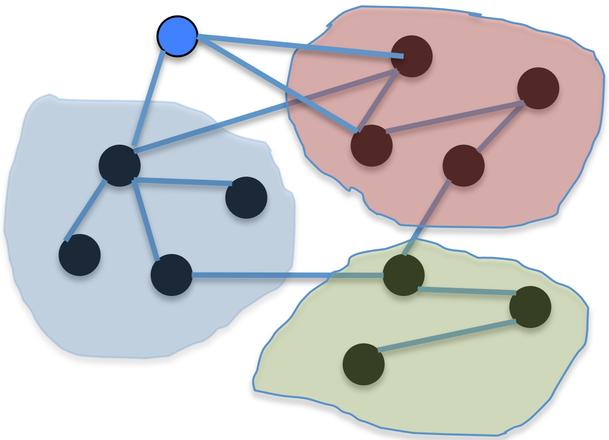
Practical Considerations

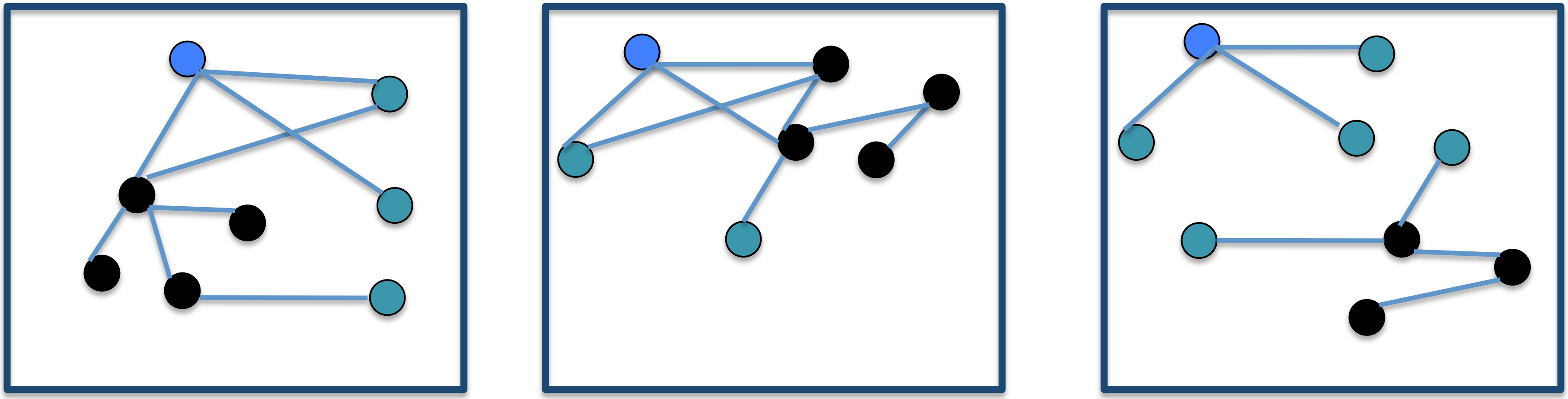
- How to partition the graph?
  - We want to minimize the number of borrowed nodes
    - Vertix cut vs. edge cut
    - Affect convergence

- Network Optimization
  - Take network topology into account

Single-pass greedy algorithm

- For each vertex *v*
  - For each partition *p*
    - We want to make sure that N(v) are in the same partition
    - Add N(v) / Nodes(p) to borrowed of *p*
  - Select p with minimum number of added borrowed nodes

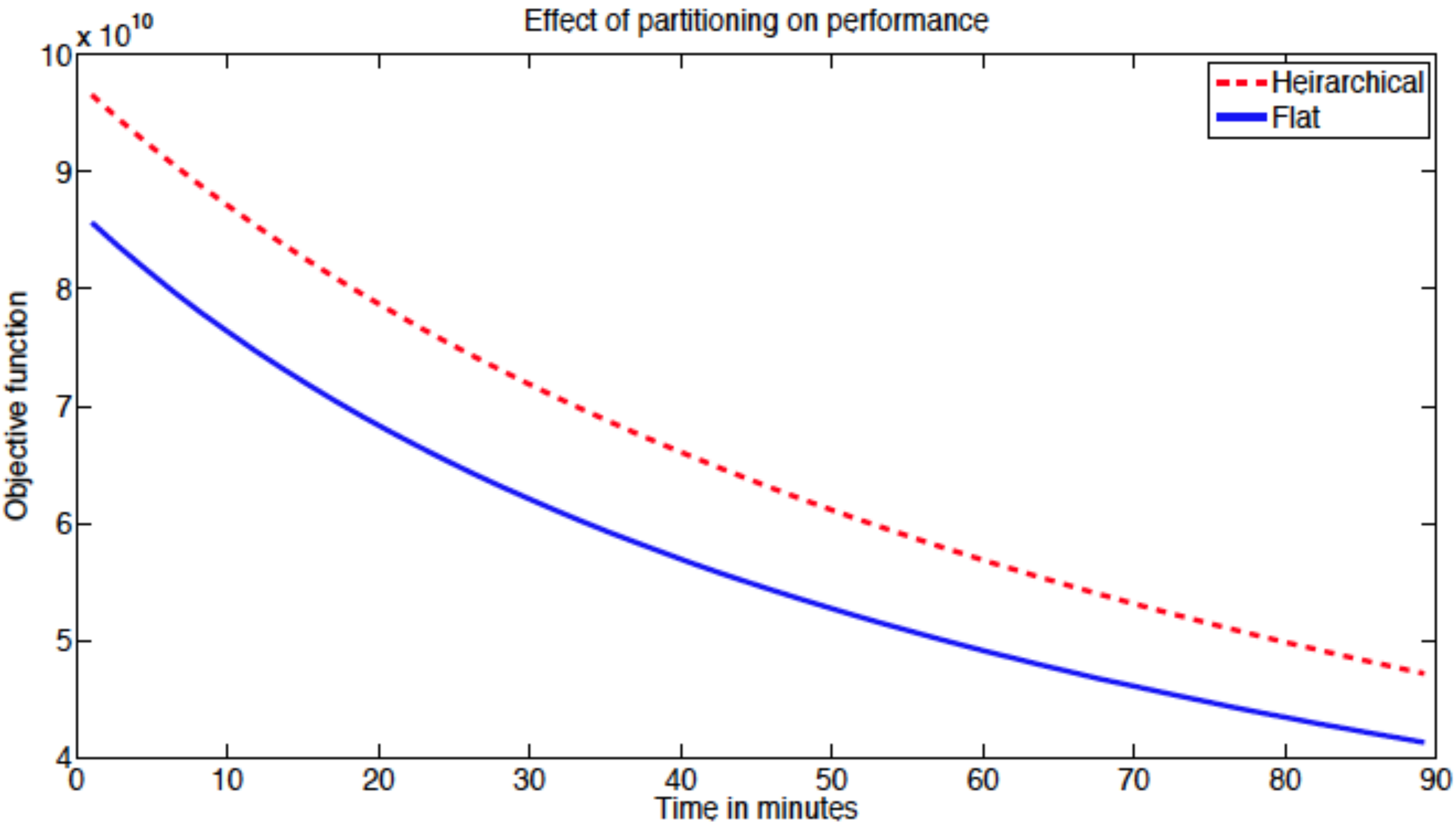




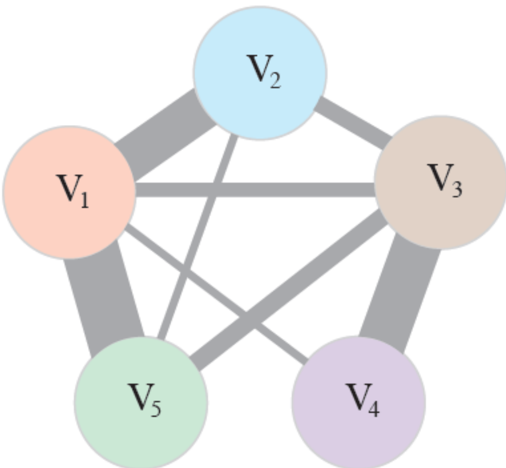
The Effect of Partitioning Quality

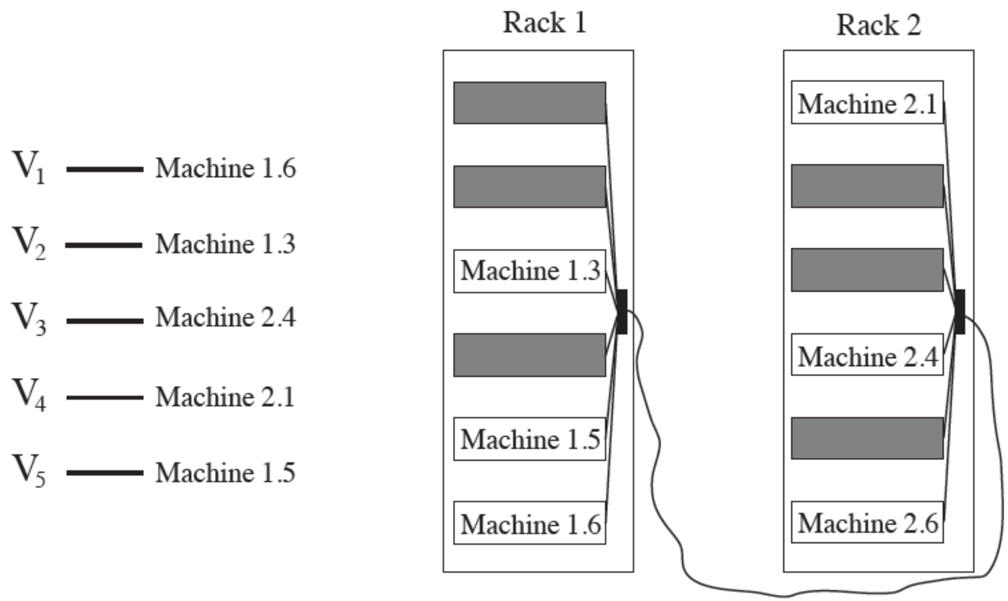
Method	Total borrowed nodes (millions)	Partitioning time (minutes)	Sync time (seconds)
Flat	252.31	166	71.5
Hierarchical	392.33	48.67	85.9
Hier-LSH	640.67	17.8	136.1
Hier-Random	720.88	11.6	145.2

The Effect of Partitioning Quality



Network Optimization

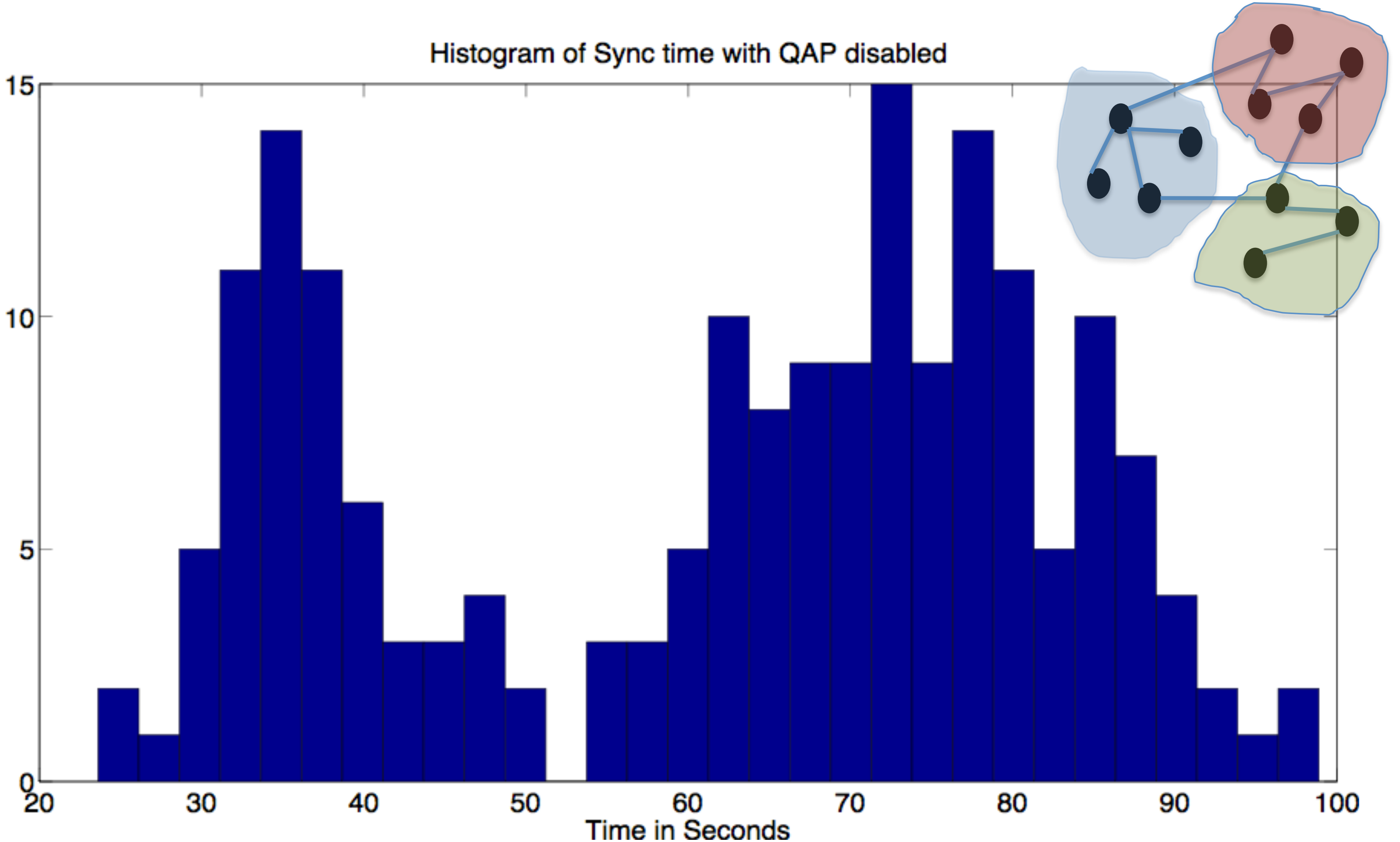




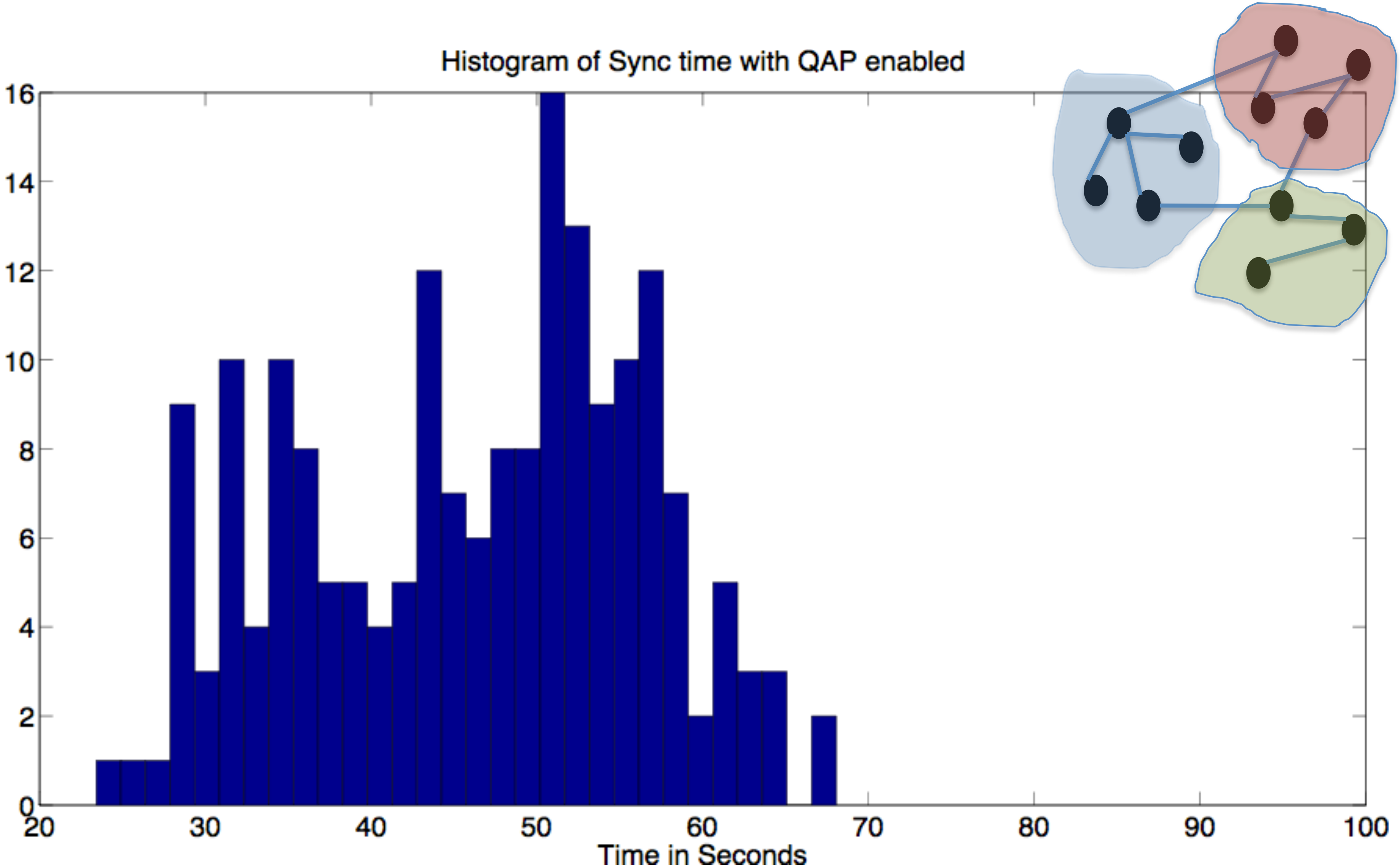
- We only know the layout at run time
- Solve a quadratic assignment problem

$$T(\pi) = \sum_{kl} C_{kl} D_{\pi(k)\pi(l)} = \sum_{kl} C_{kl} \sum_{uv} \pi_{ku} \pi_{lv} D_{uv} = \operatorname{tr} C \pi D \pi^\top$$

Sync time without QAP



Sync time with QAP



Summary

- Model as consensus problem
- Synchronous algorithms
  - Curse of the last reducer
- Asynchronous algorithms
  - Asynchronous parallel updates
  - Network topology optimization
  - Overlapping partitions
- Same idea applies to GMF models and collective graph factorization, matrix factorization, etc.

Hierarchical Multi-task Learning and Sparse Models

Computational Advertising

Display Advertising

- Behavioral targeting
- Given user feature vector
  - URL, queries, etc.
- Prediction problems for each campaign
  - Click prediction
  - Conversion prediction

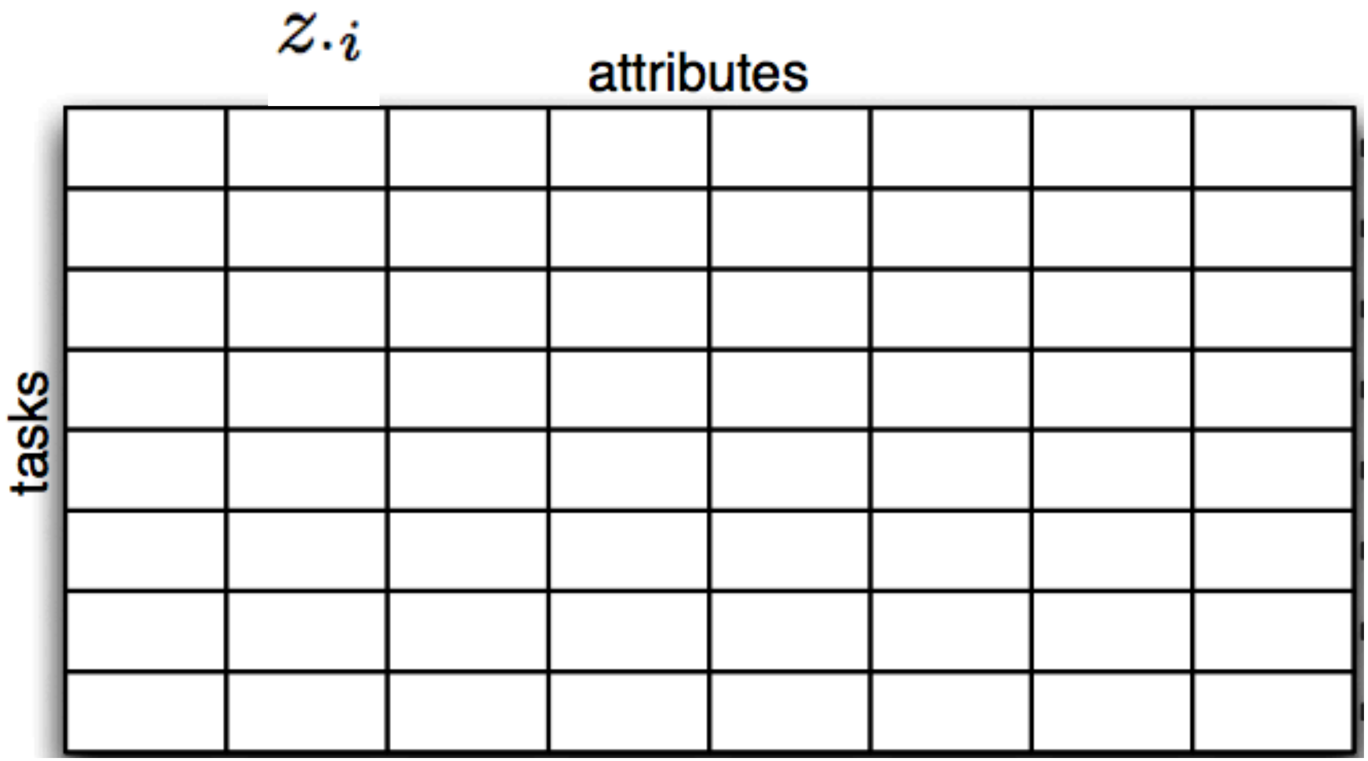
- Both are very sparse high-dimensional classification problems

Research Question

- Can we leverage data across tasks/sub-tasks?
  – Campaigns targeting sports lovers have similar clicking pattern
  – Can click data in one campaign help conversion?

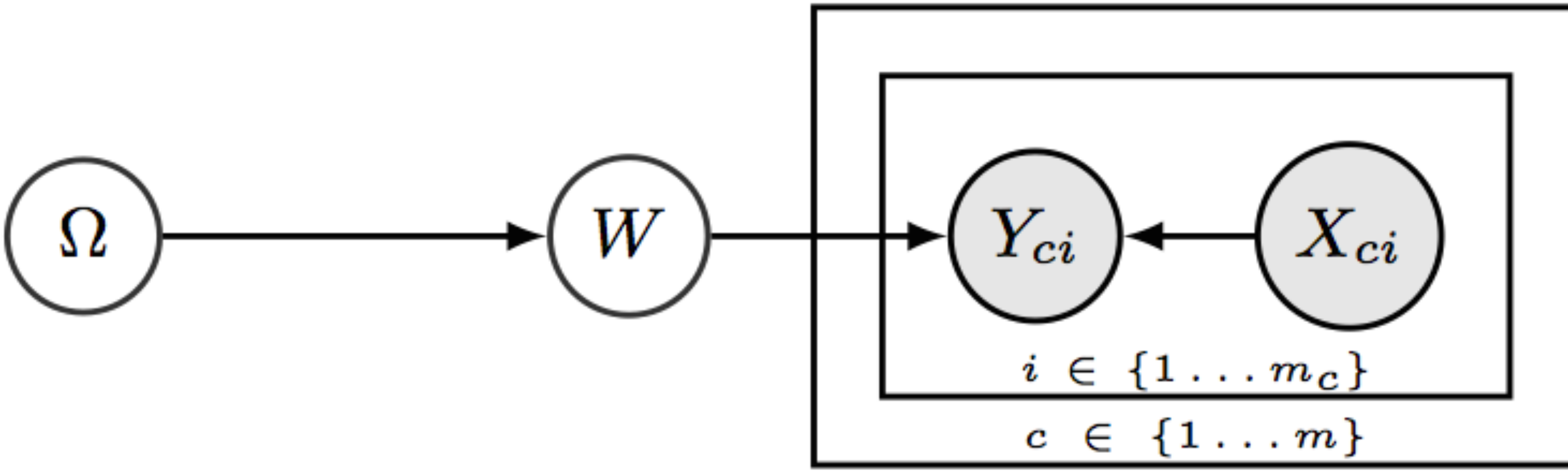
- **Challenges**
  - Hundred of millions of features
  - Thousands of campaigns
  - Billion of users
  - We want to learn sparse models for serving

Matrix-vitiate distribution



$$Z \sim \mathcal{N}(0, \mathbf{1}_d \otimes \Omega) \text{ or equivalently } z_{\cdot i} \sim \mathcal{N}(0, \Omega)$$

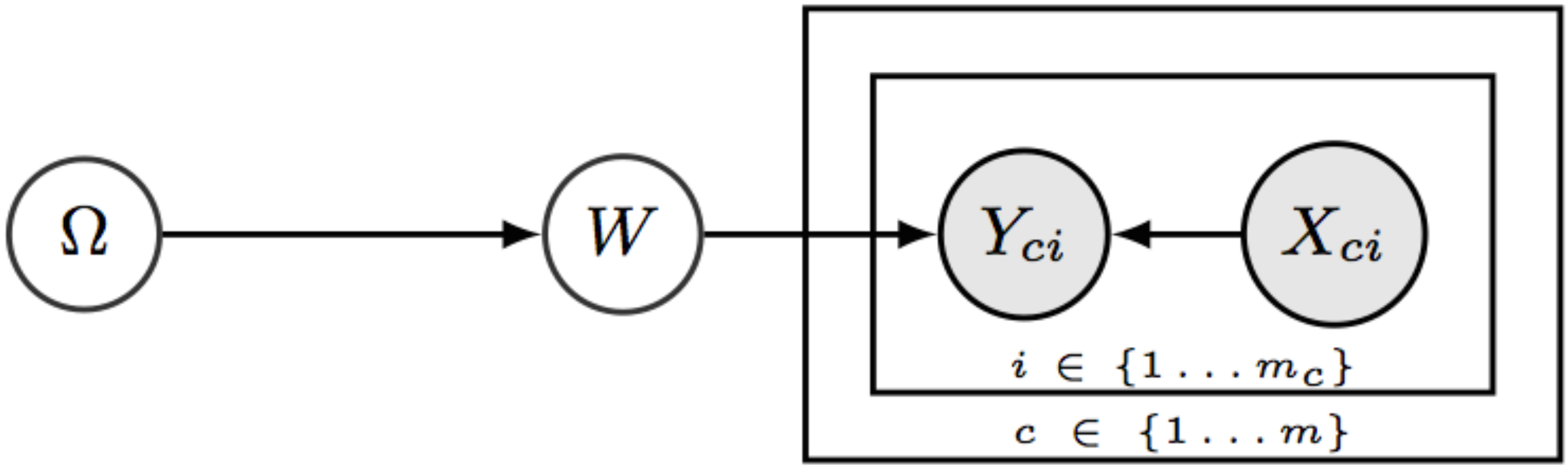
Multi-Task Learning



$$W \sim \mathcal{N}(0, \mathbf{1}_d \otimes \Omega) \text{ or equivalently } w_{\cdot i} \sim \mathcal{N}(0, \Omega)$$

$$-\log p(W|\Omega) = \operatorname{tr} W \Omega^{-1} W^\top + d \log|\Omega| + c$$

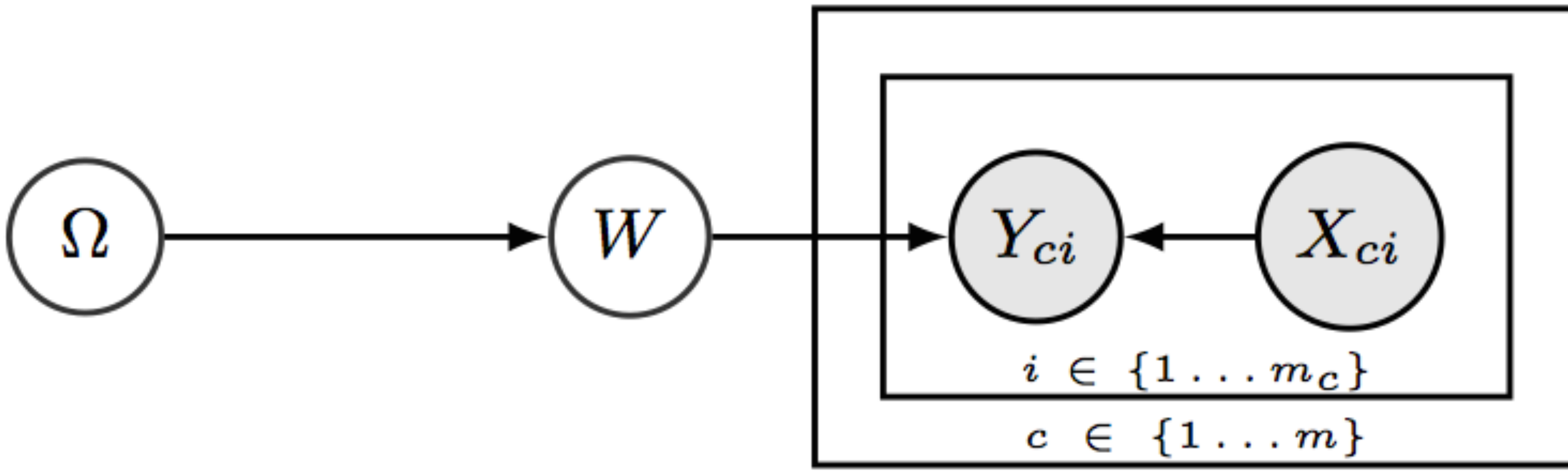
Multi-Task Learning



$$\underset{W,\Omega}{\text{minimize}} \sum_c -\log p(Y_c|X_c, w_c) + \lambda \operatorname{tr} W\Omega^{-1}W^\top$$

subject to $\Omega \succeq 0$ and $\operatorname{tr}\Omega = 1$

Multi-Task Learning

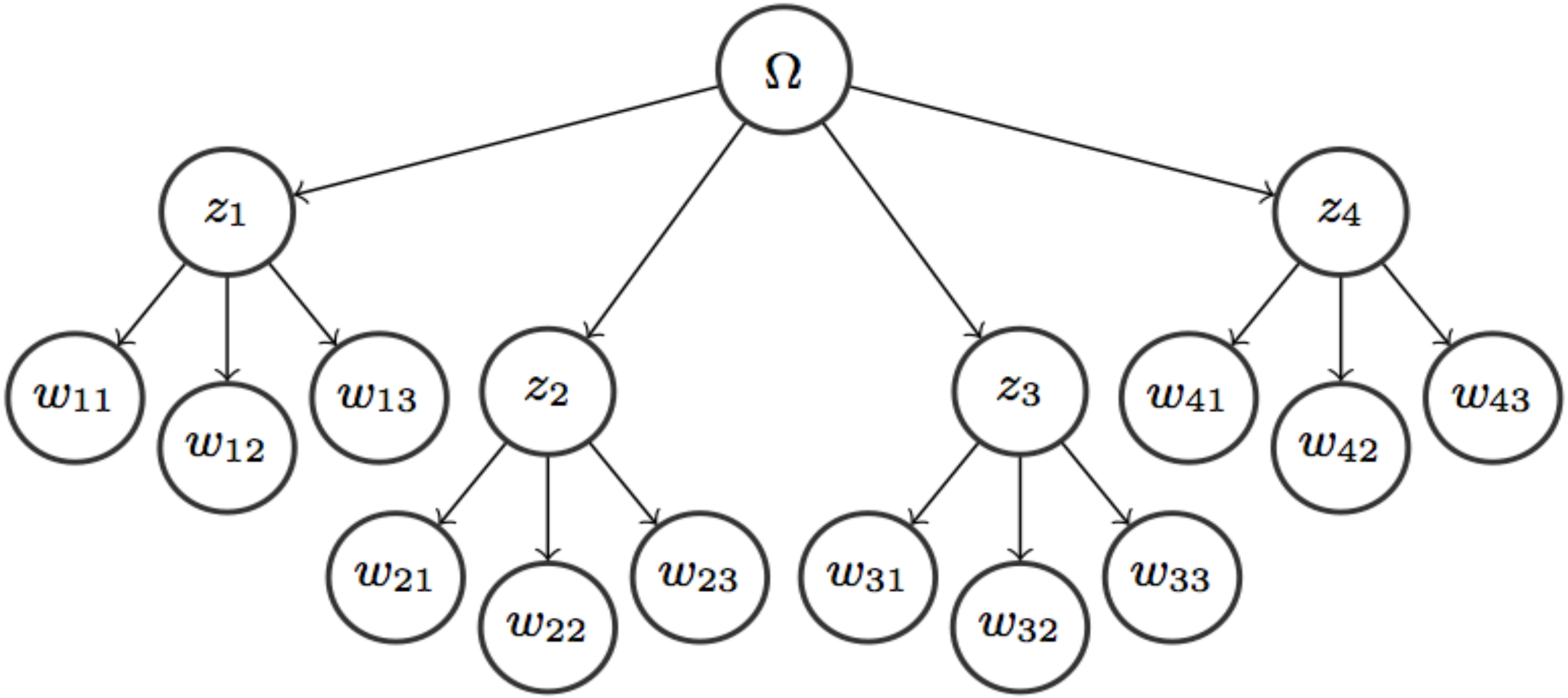


$$\underset{W,\Omega}{\text{minimize}} \sum_c -\log p(Y_c | X_c, w_c) + \lambda \operatorname{tr} W \Omega^{-1} W^\top$$

subject to $\Omega \succeq 0$ and $\operatorname{tr} \Omega =$

$$\hat{\Omega} = \frac{[W^\top W]^{-\frac{1}{2}}}{\operatorname{tr}[W^\top W]^{-\frac{1}{2}}}$$

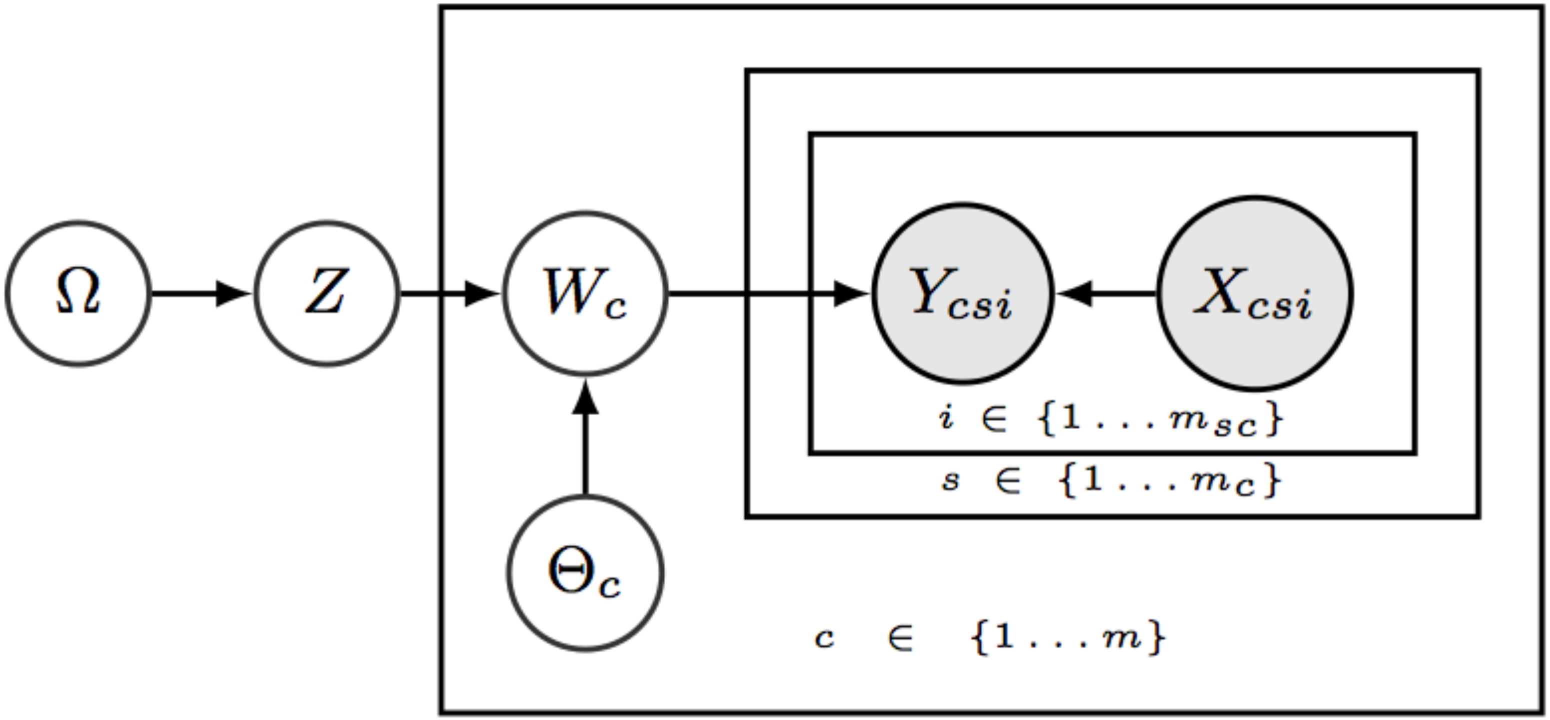
Hierarchical Multi-task learning



$Z \sim \mathcal{N}(0, \mathbf{1}_d \otimes \Omega)$ or equivalently $z_{\cdot i} \sim \mathcal{N}(0, \Omega)$

$w_{c \cdot i} \sim \mathcal{N}(1 \cdot z_{ci}, \Theta_c).$

In graphical Model



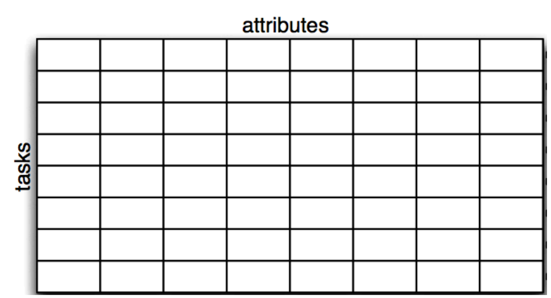
$Z \sim \mathcal{N}(0, \mathbf{1}_d \otimes \Omega)$ or equivalently $z_{\cdot i} \sim \mathcal{N}(0, \Omega)$

$w_{c \cdot i} \sim \mathcal{N}(1 \cdot z_{ci}, \Theta_c)$.

Optimization Problem

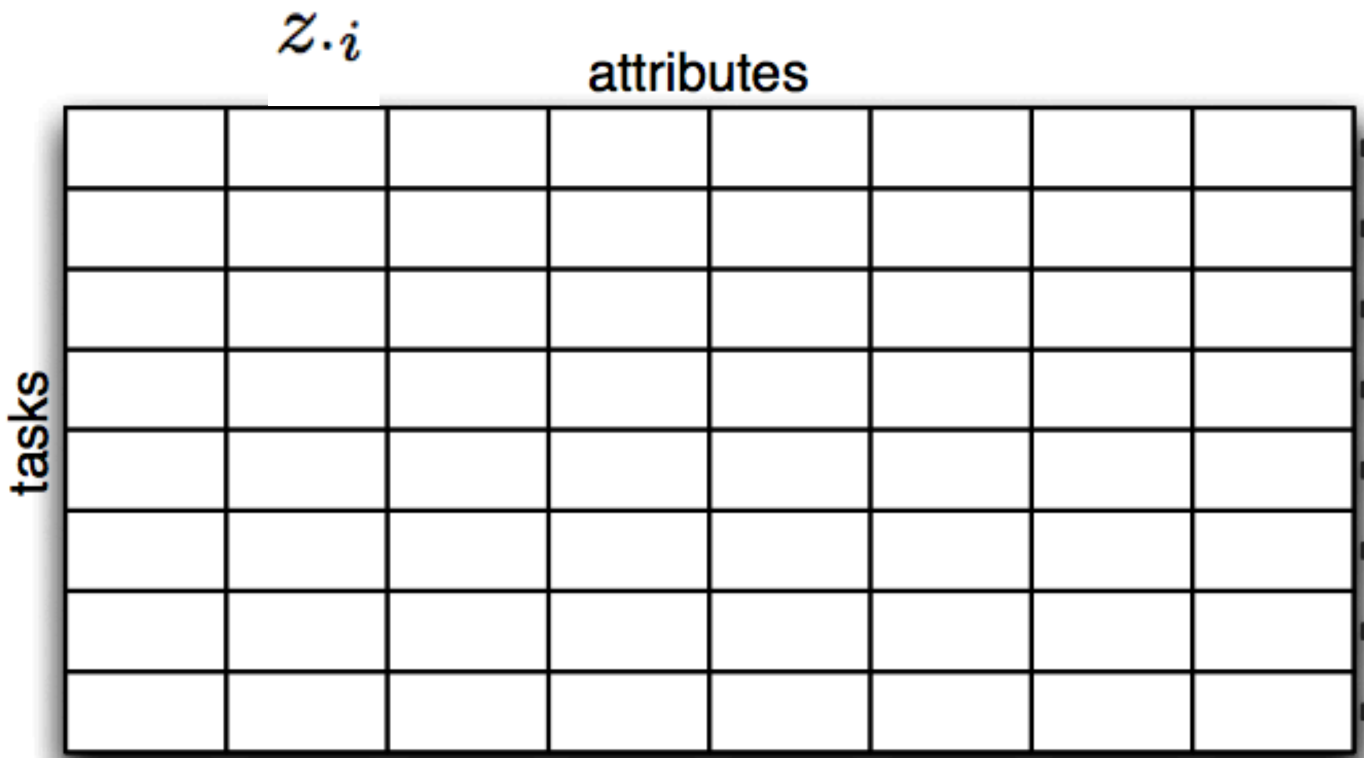
$$
\begin{aligned}
\underset{W,Z,\Omega,\Theta}{\text{minimize}} \quad & \sum_{csj} -\log p(y_{csj} \,|\, x_{csj}, w_{cs}) + \frac{1}{2}\operatorname{tr} Z^\top \Omega^{-1} Z \\
& + \sum_{c} \frac{1}{2} \operatorname{tr}(w_{c\cdot} - 1 \cdot z_c)^\top (w_{c\cdot} - 1 \cdot z_c)\Theta_c^{-1} \\
& + \lambda_1 \|Z\|_1 + \lambda_2 \|Z\|_{2,1} \qquad \text{(17a)} \\
& + \lambda_1 \|W\|_1 + \sum_{c} \lambda_2 \|W_c\|_{2,1}
\end{aligned}
$$

$$
\text{subject to } \Omega, \Theta_c \succeq 0 \text{ and } \operatorname{tr}\Omega = \operatorname{tr}\Theta_c = 1. \qquad \text{(17b)}
$$



Sparsity

$$|Z\|_1 + \lambda_2 \|Z\|_{2,1}$$



$$\|X\|_{p,q} := \Big\| \|X_{1\cdot}\|_p , \ldots \|X_{d\cdot}\|_p \Big\|_q$$

Optimization Problem

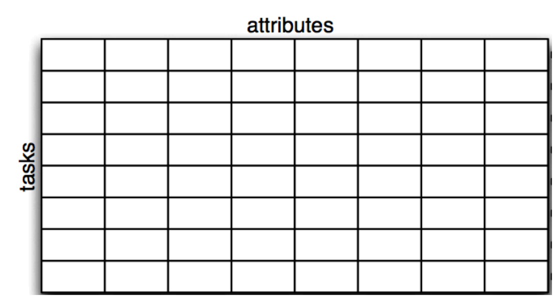
$$\underset{W,Z,\Omega,\Theta}{\text{minimize}} \quad \sum_{csj} -\log p(y_{csj}|x_{csj}, w_{cs}) + \frac{1}{2}\operatorname{tr} Z^\top \Omega^{-1} Z$$

$$+ \sum_{c} \frac{1}{2} \operatorname{tr}(w_{c\cdot} - 1 \cdot z_c)^\top (w_{c\cdot} - 1 \cdot z_c)\Theta_c^{-1}$$

$$+ \lambda_1 \|Z\|_1 + \lambda_2 \|Z\|_{2,1} \tag{17a}$$

$$+ \lambda_1 \|W\|_1 + \sum_c \lambda_2 \|W_c\|_{2,1}$$

$$\text{subject to } \Omega, \Theta_c \succeq 0 \text{ and } \operatorname{tr}\Omega = \operatorname{tr}\Theta_c = 1. \tag{17b}$$



Proximal Methods

$$\underset{a}{\text{minimize}}\, f(a) + \lambda\Omega[a]$$

$$b_{t+1} := a_t - \eta_t \partial_a f(a_t) \text{ and}$$

$$a_{t+1} = \underset{a}{\text{argmin}}\, \frac{1}{2t_t} \|a - b_{t+1}\|^2 + \lambda\Omega[a]$$

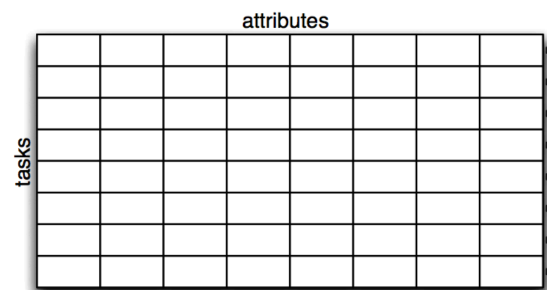
Example: L1

$$a_{t+1} \leftarrow sgn(b_{t+1}) max(0, |b_{t+1}| - t_i \lambda)$$

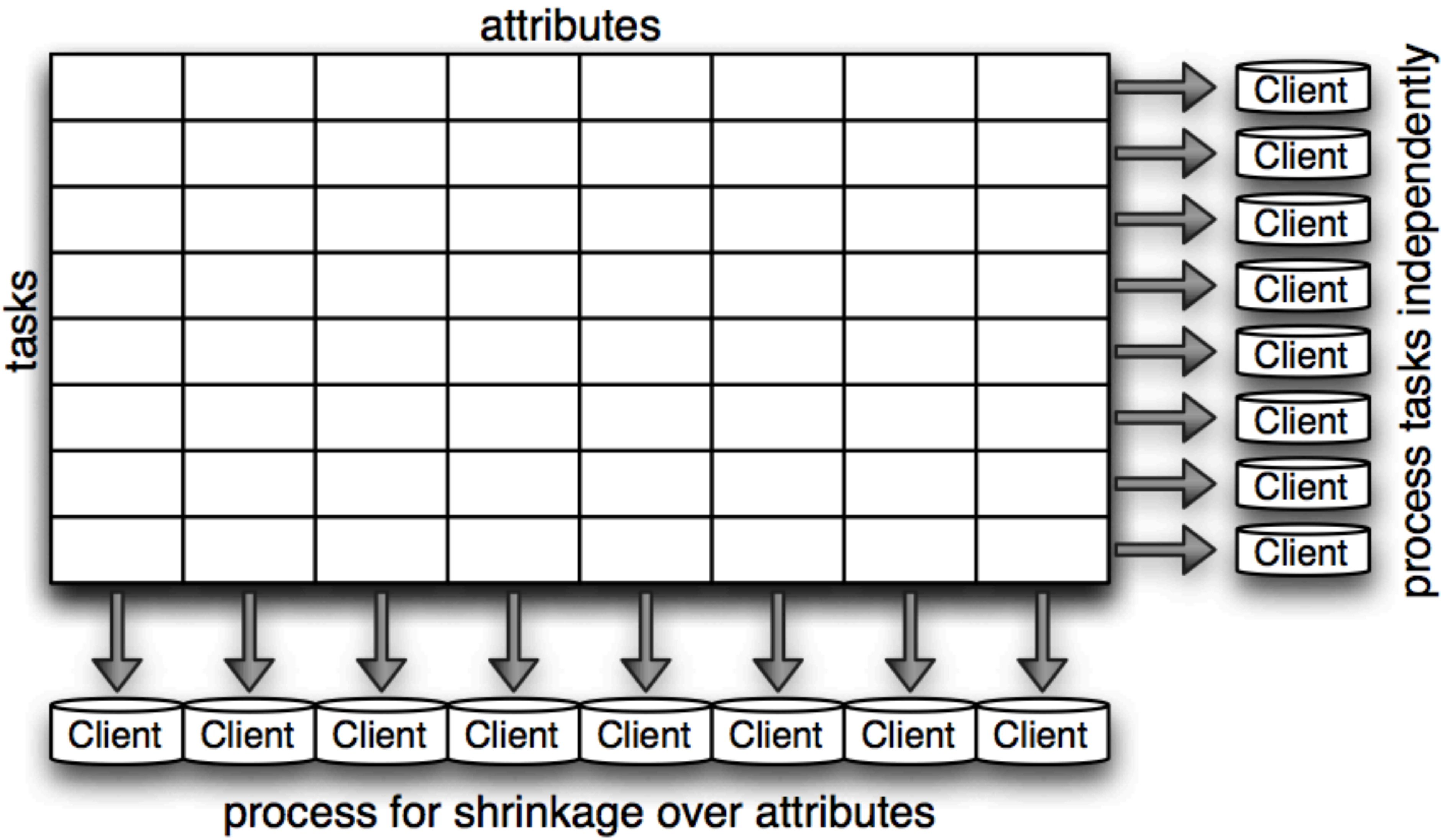
Optimization Problem

$$
\begin{aligned}
\underset{W,Z,\Omega,\Theta}{\text{minimize}} \quad & \sum_{csj} -\log p(y_{csj}|x_{csj}, w_{cs}) + \frac{1}{2} \operatorname{tr} Z^\top \Omega^{-1} Z \\
& + \sum_c \frac{1}{2} \operatorname{tr}(w_{c\cdot} - 1 \cdot z_c)^\top (w_{c\cdot} - 1 \cdot z_c)\Theta_c^{-1} \\
& + \lambda_1 \|Z\|_1 + \lambda_2 \|Z\|_{2,1} \qquad \text{(17a)} \\
& + \lambda_1 \|W\|_1 + \sum_c \lambda_2 \|W_c\|_{2,1}
\end{aligned}
$$

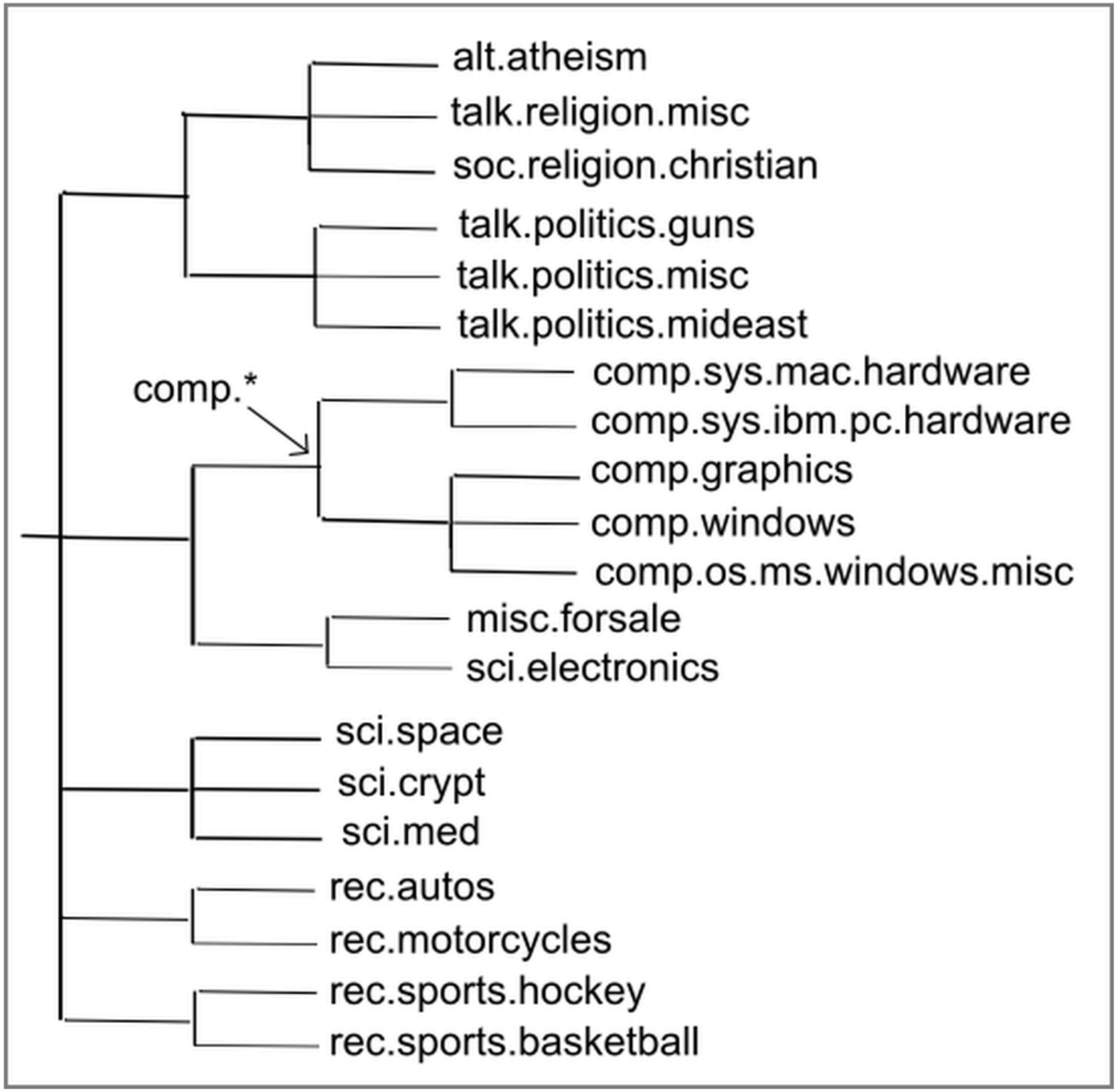
$$
\text{subject to } \Omega, \Theta_c \succeq 0 \text{ and } \operatorname{tr}\Omega = \operatorname{tr}\Theta_c = 1. \qquad \text{(17b)}
$$



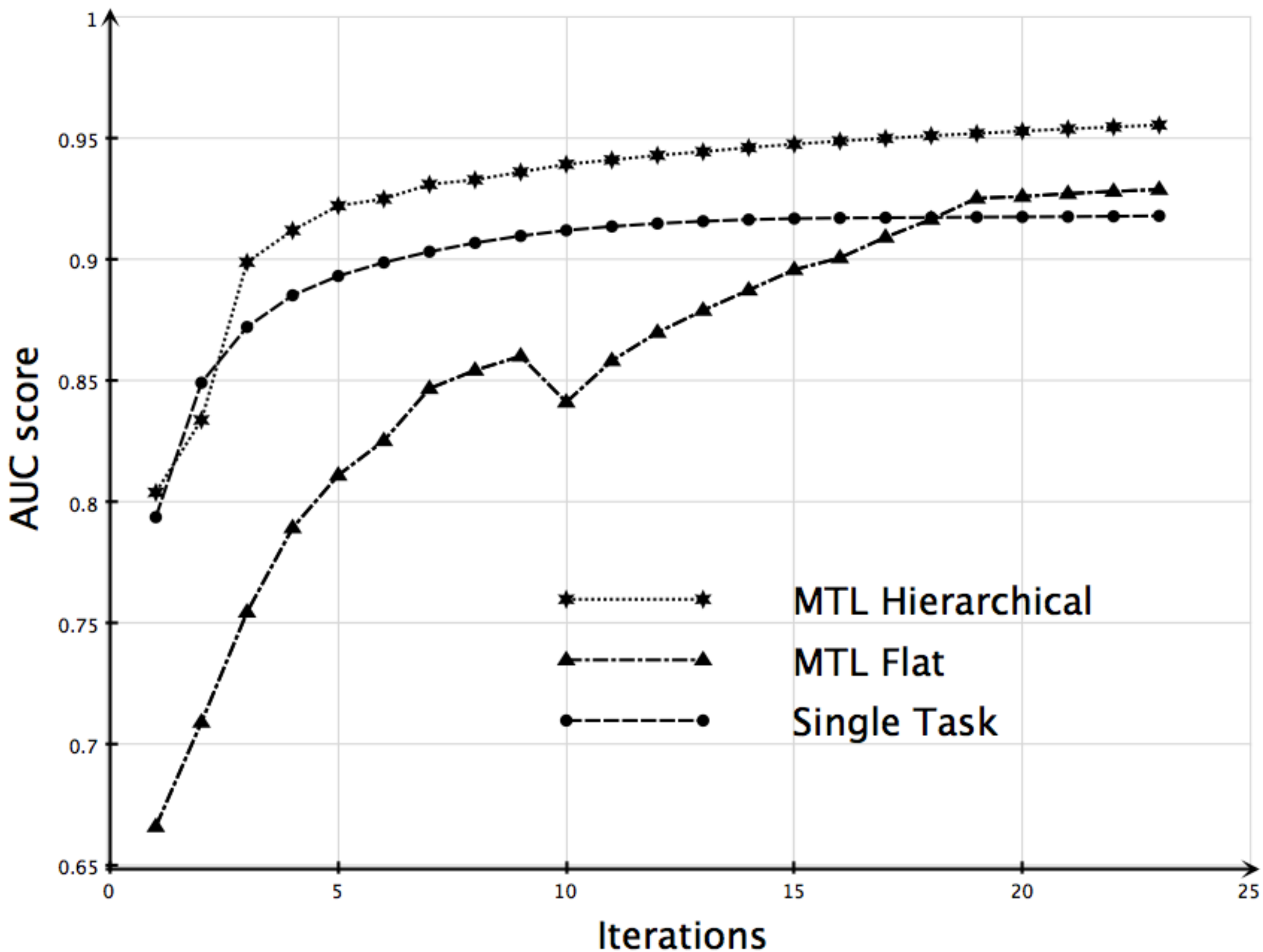
Distributed Implementation



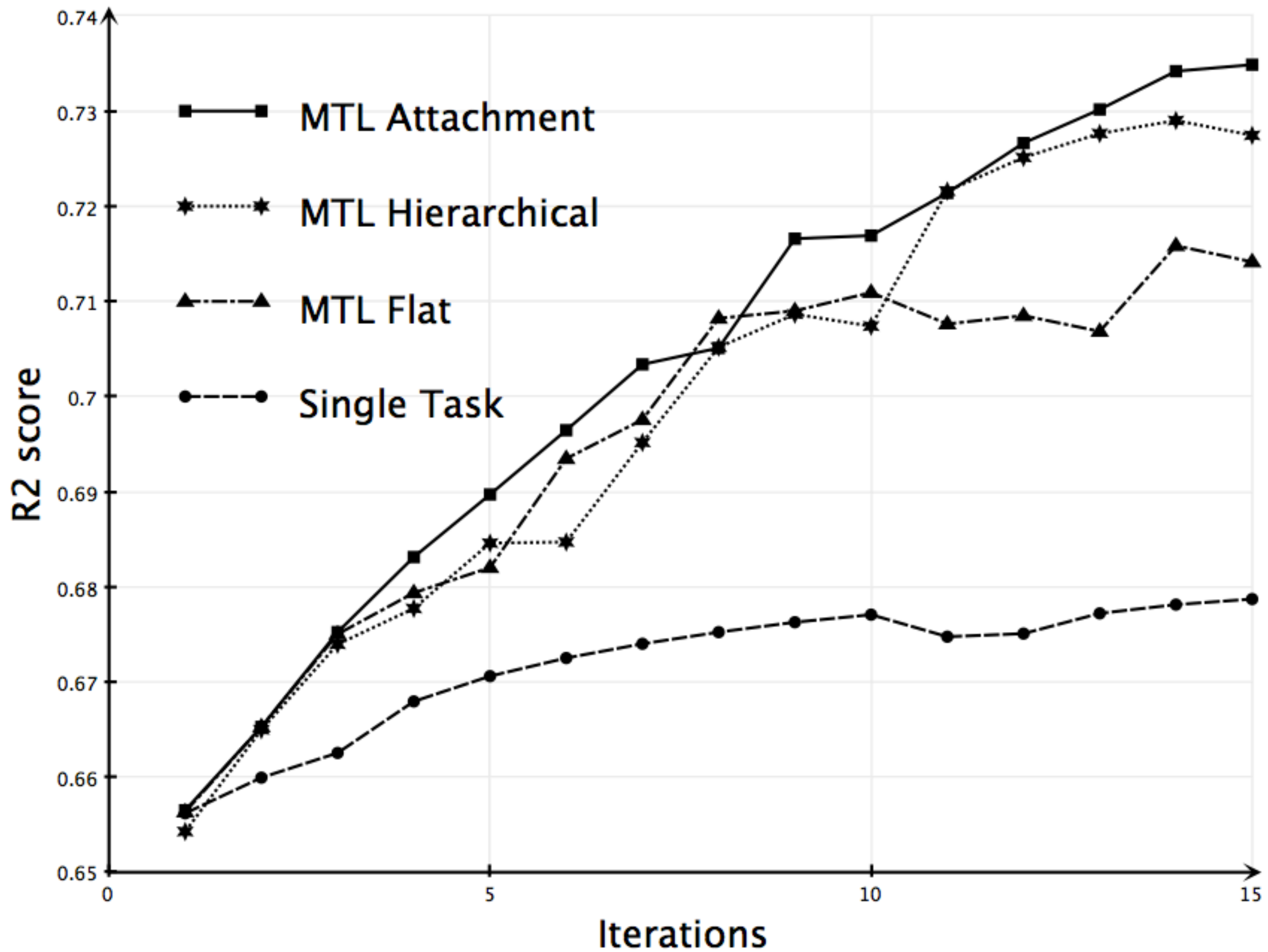
Public Dataset: 20-news group



Public Dataset: 20-news group



Public data: School dataset



Yahoo Advertising Dataset

days	users	features	campaigns	dataset size
56	10^9	934,000	630	1.4TB

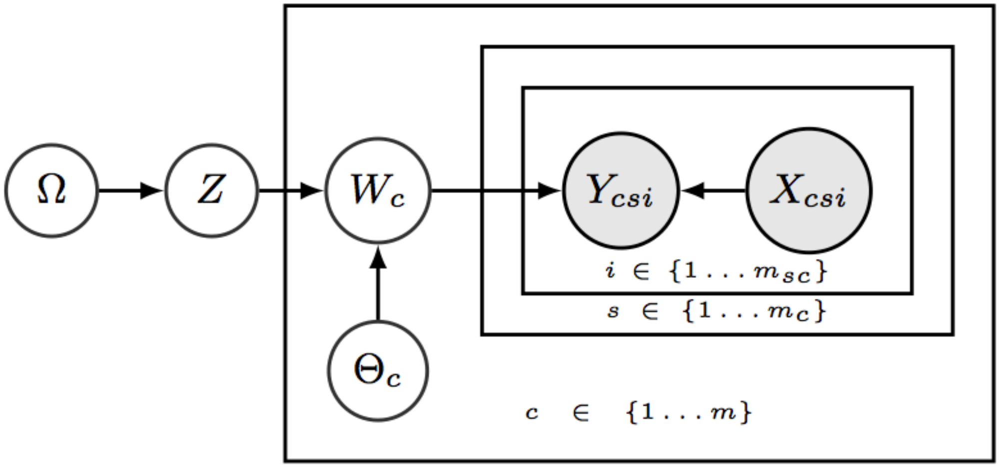
Table 2: Attachment multitask performance.

AUC	STL	ATT-MTRL
all subtasks	0.658	**0.674**
conversions	0.629	**0.653**
auxiliary (unattributed)	0.677	**0.714**
clicks	0.662	**0.671**

Ablation study

Table 4: Ablation study for ATT-MTRL.

AUC	conversions	all sub-tasks
L1	0.621	0.642
L1+L12	0.629	0.658
L1+L12+Θ	0.641	0.663
L1+L12+Θ+Ω	**0.653**	**0.674**



How sparse is the model?

Table 3: Feature selection effectiveness:

	Conversion AUC	features
STL + ℓ_2 + top features	0.606	10,000
STL + ℓ_2 + top features	0.609	30,000
STL + ℓ_2 + top features	0.607	50,000
ATT-MTRL (aggressive)	0.631	3,992
ATT-MTRL (conservative)	**0.653**	17,789

Summary

- Two Hierarchical multi-task learning formulation
- Distributed client-server optimization
- Sparse models
- Application in display advertising
- Can be extended to arbitrary levels

Advanced Directions

Advanced Directions

- Theoretical bounds and guarantees
- Non-parametric models
  - Learning structure from data
- Working under communication constraints
- More applications
  - Citation analysis
    - Graph factorization + LDA
- Semi-asynchronous algorithms

Selected References covered

- "probabilistic topic models", David Blei, review article.
- "Scalable Inference in Latent Variable Models", Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, Alex Smola, WSDM 2012.
- "Scalable Distributed Inference of Dynamic User Interests for Behavioral Targeting", Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, Alex Smola, KDD 2011
- "Multiple Domain User Personalization" , Yucheng Low, Deepak Agarwal and Alex Smola, KDD 2011.
- "The Dataminer Guide to Scalable Mixed-Membership and Nonparametric Bayesian Models", Amr Ahmed and Alexander J Smola, KDD 2013.
- "Distributed Large-scale Natural Graph Factorization" Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, Alexander J Smola, www 2013.
- "Hierarchical multitask learning: scalable algorithms and an application to conversion optimization in display advertising", Amr Ahmed Abhimanyu Das Alexander J. Smola, WSDM 2014.