Graphical Models for the Internet

Amr Ahmed and Alexander Smola

Yahoo Research, Santa Clara, CA

Thus far ...

- Motivation
- Basic tools
  - Clustering
  - Topic Models
- Distributed batch inference
  - Local and global states
  - Star synchronization

Up next

- Inference
  - Online Distributed Sampling
  - Single machine multi-threaded inference
  - Online EM and Submodular Selection
- Applications
  - User tracking for behavioral Targeting
  - Content understanding
  - User modeling for content recommendation

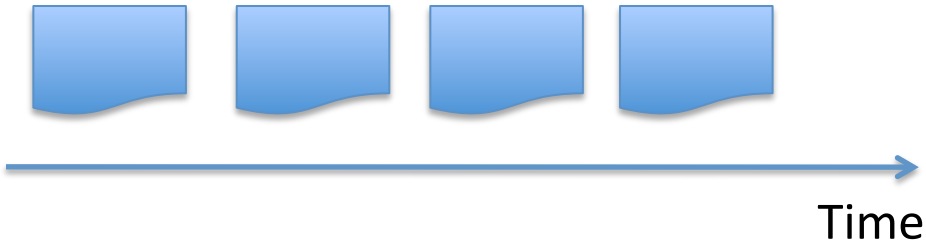
4. Online Model

Scenarios

- Batch Large-Scale
  - Covered in part 1

- Mini-batches
  - We already have a model
  - Data arrives in batches
  - We would like to keep model up-to-data

- Time-sensitive
  - Data arrives one item at a time
  - Model should be up-to-data

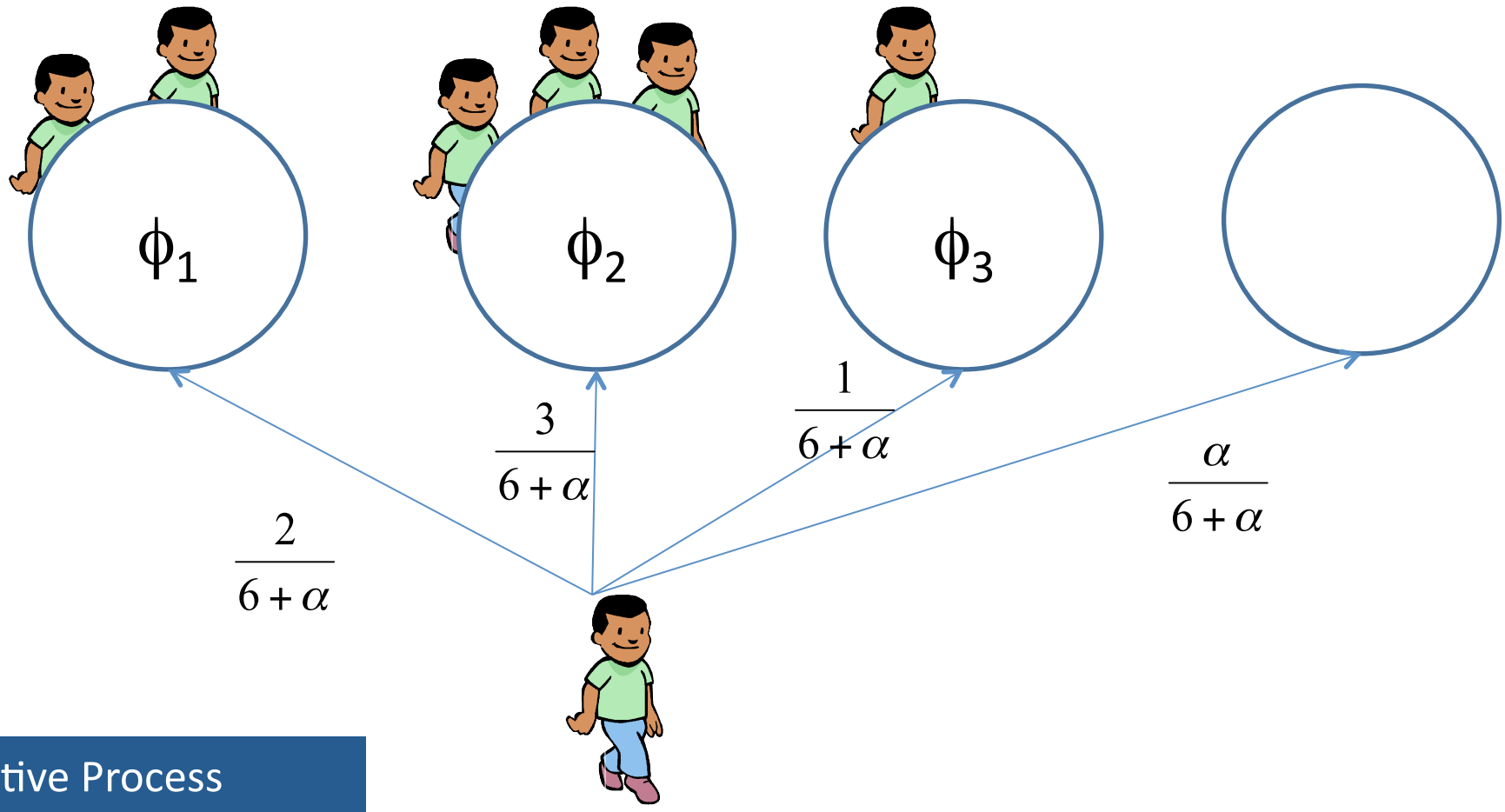


4.1 Dynamic Clustering

The Chinese Restaurant Process

- Allows the number of mixtures to grow with the data
- They are called non-parametric models
  - Means the number of effective parameters grow with data
  - Still have hyper-parameters that control the rate of growth
    - α: how fast a new cluster/mixture is born?
    - G_0: Prior over mixture component parameters

The Chinese Restaurant Process



Generative Process

-For data point x_i

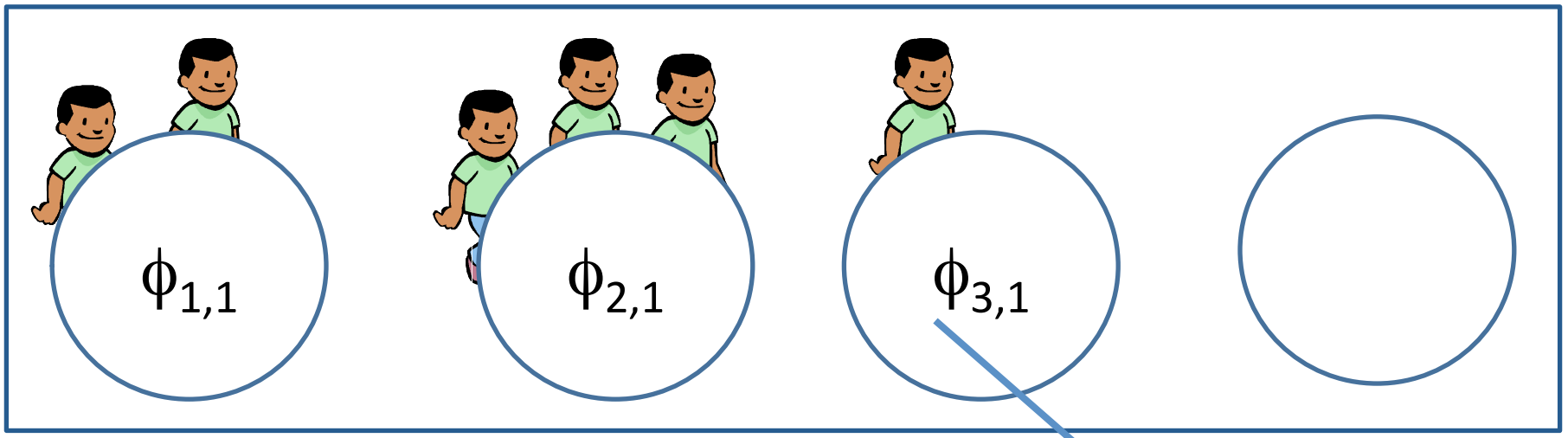
- Choose table $j \propto m_j$ and Sample $x_i \sim f(\phi_j)$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1})$

The rich gets richer effect
CANNOT handle sequential data

Recurrent CRP (RCRP) [Ahmed and Xing 2008]

- Adapts the number of mixture components over time
  - Mixture components can die out
  - New mixture components are born at any time
  - Retained mixture components parameters evolve according to a Markovian dynamics

The Recurrent Chinese Restaurant Process



T=1

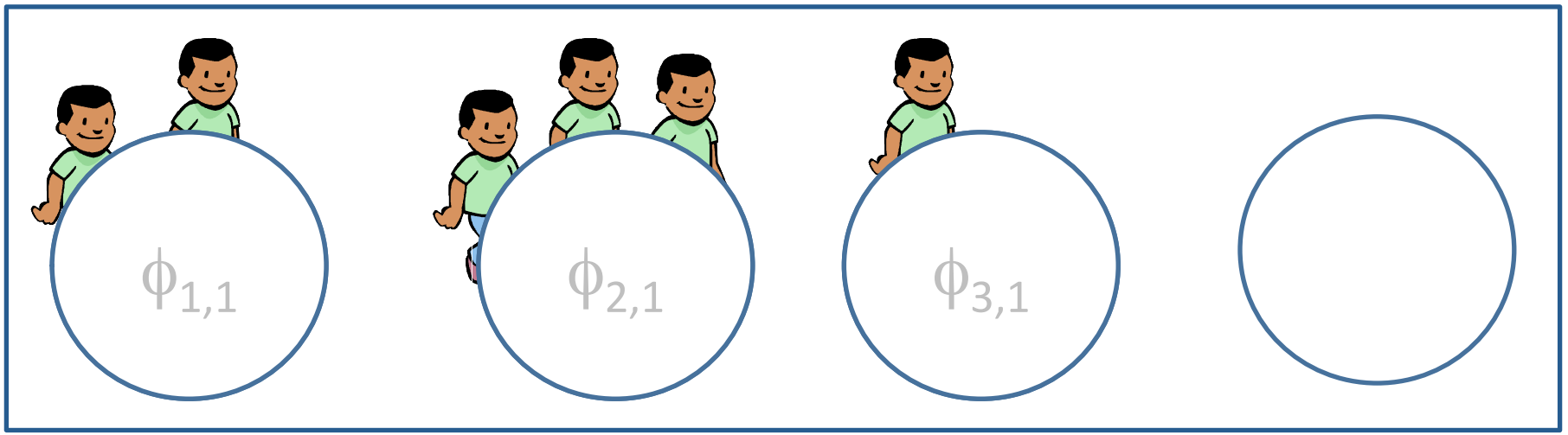
Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1

Generative Process

-Customers at time T=1 are seated as before:

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

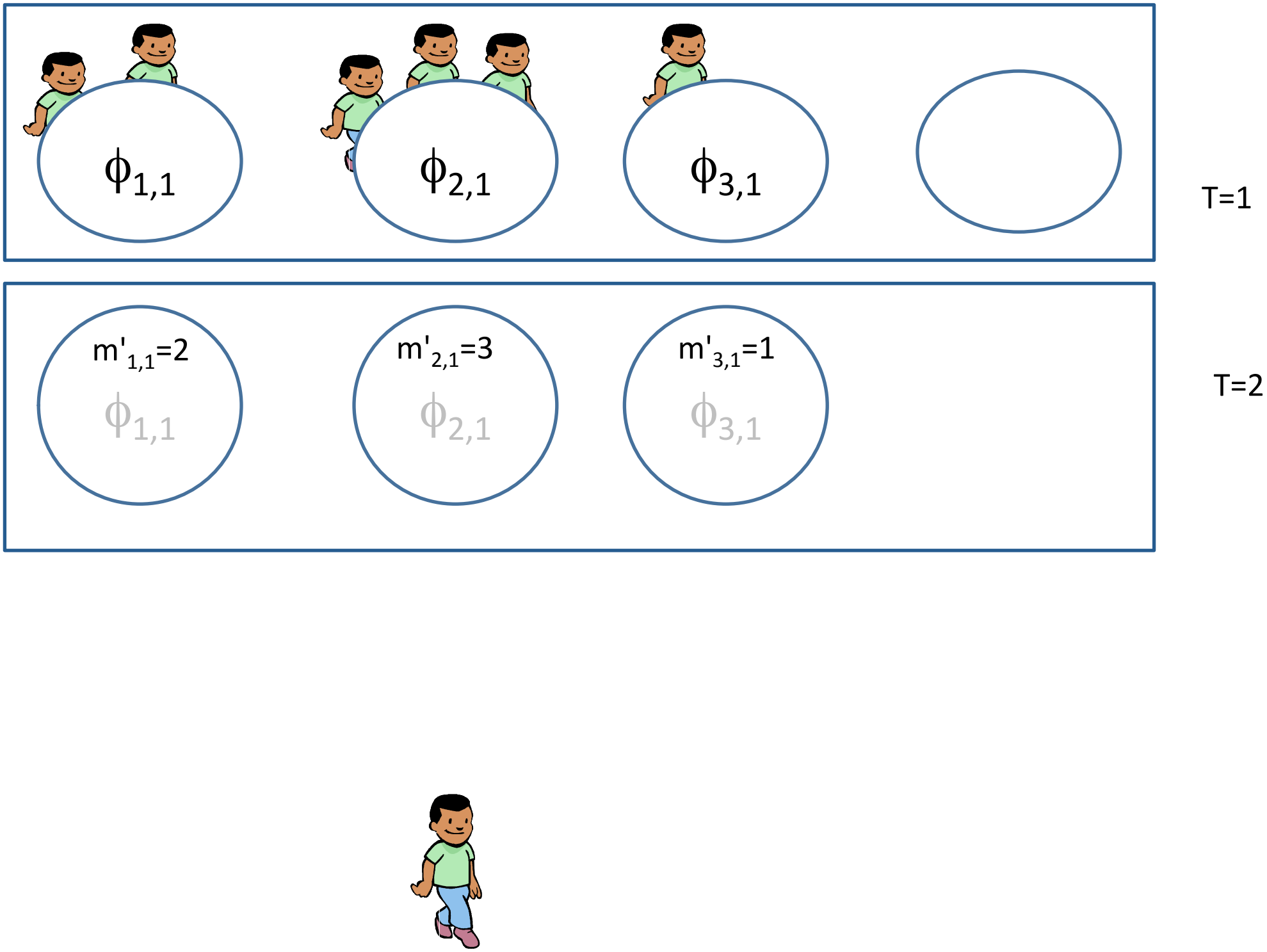
The Recurrent Chinese Restaurant Process

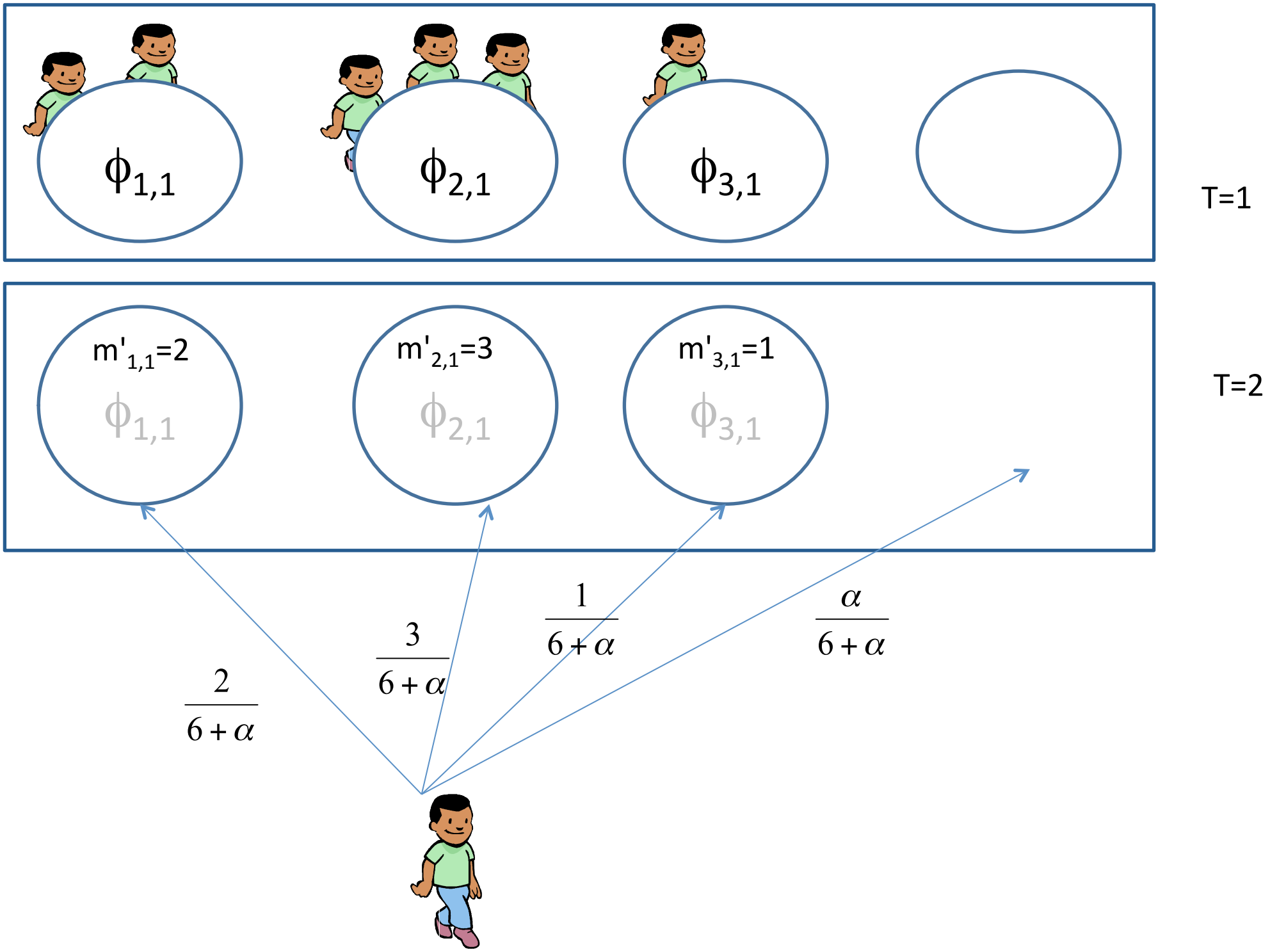


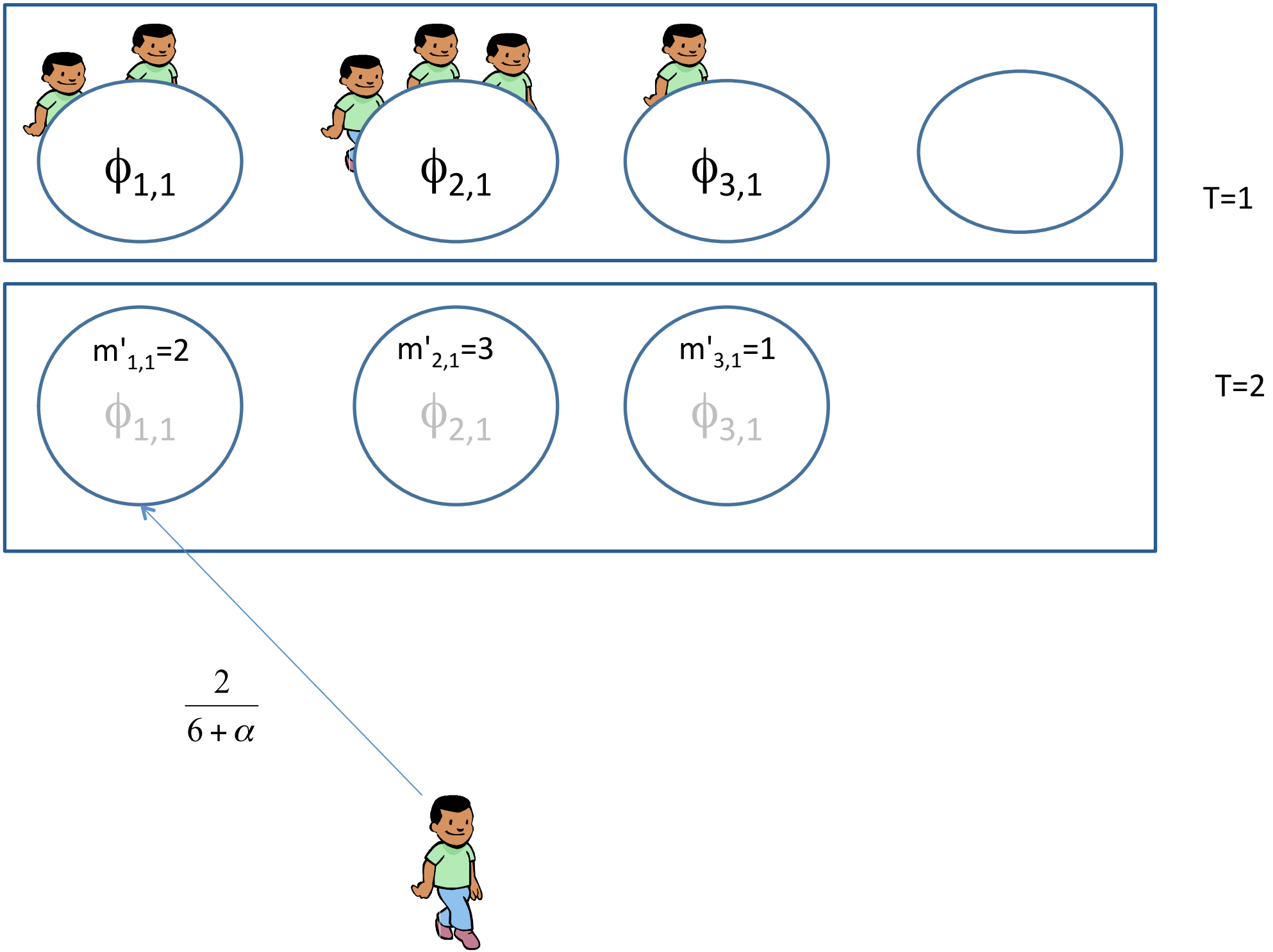
T=1

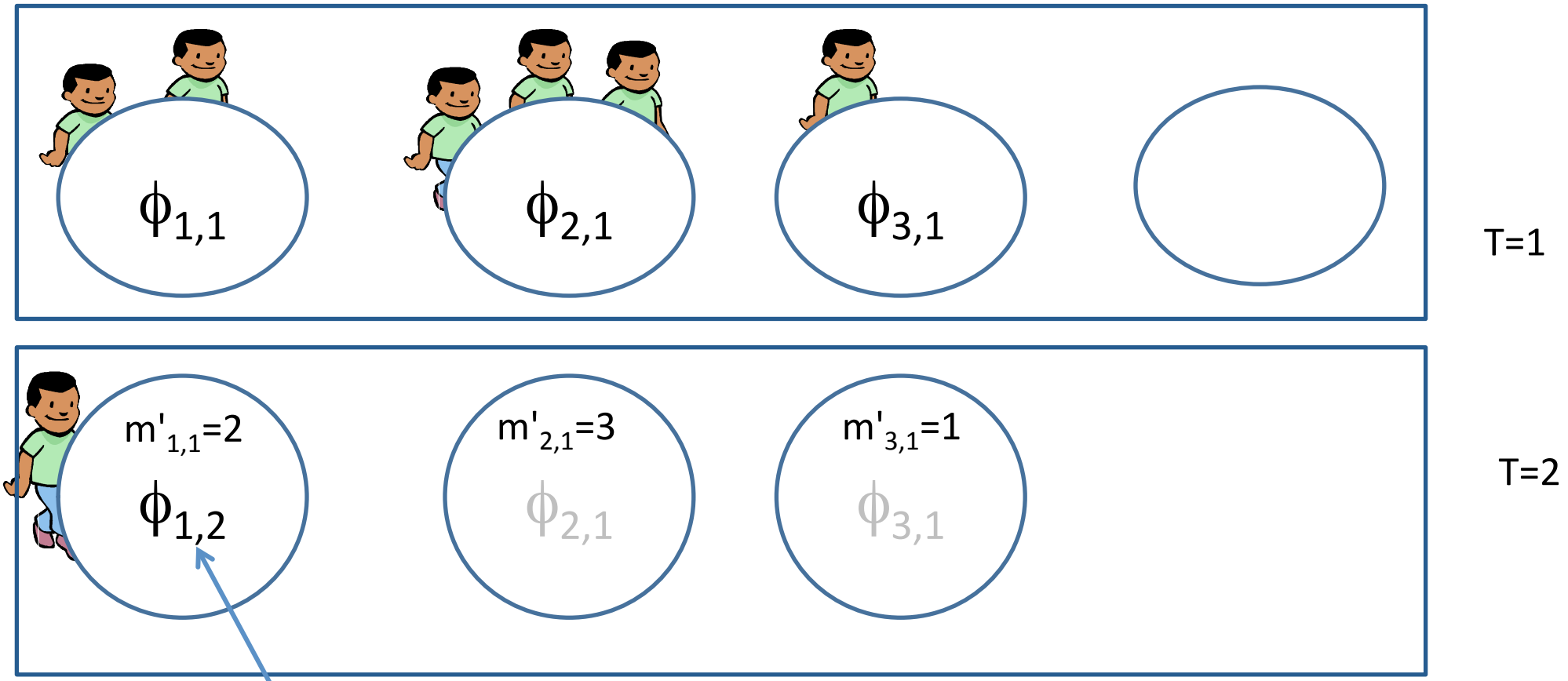
$m'_{1,1}=2$ $m'_{2,1}=3$ $m'_{3,1}=1$

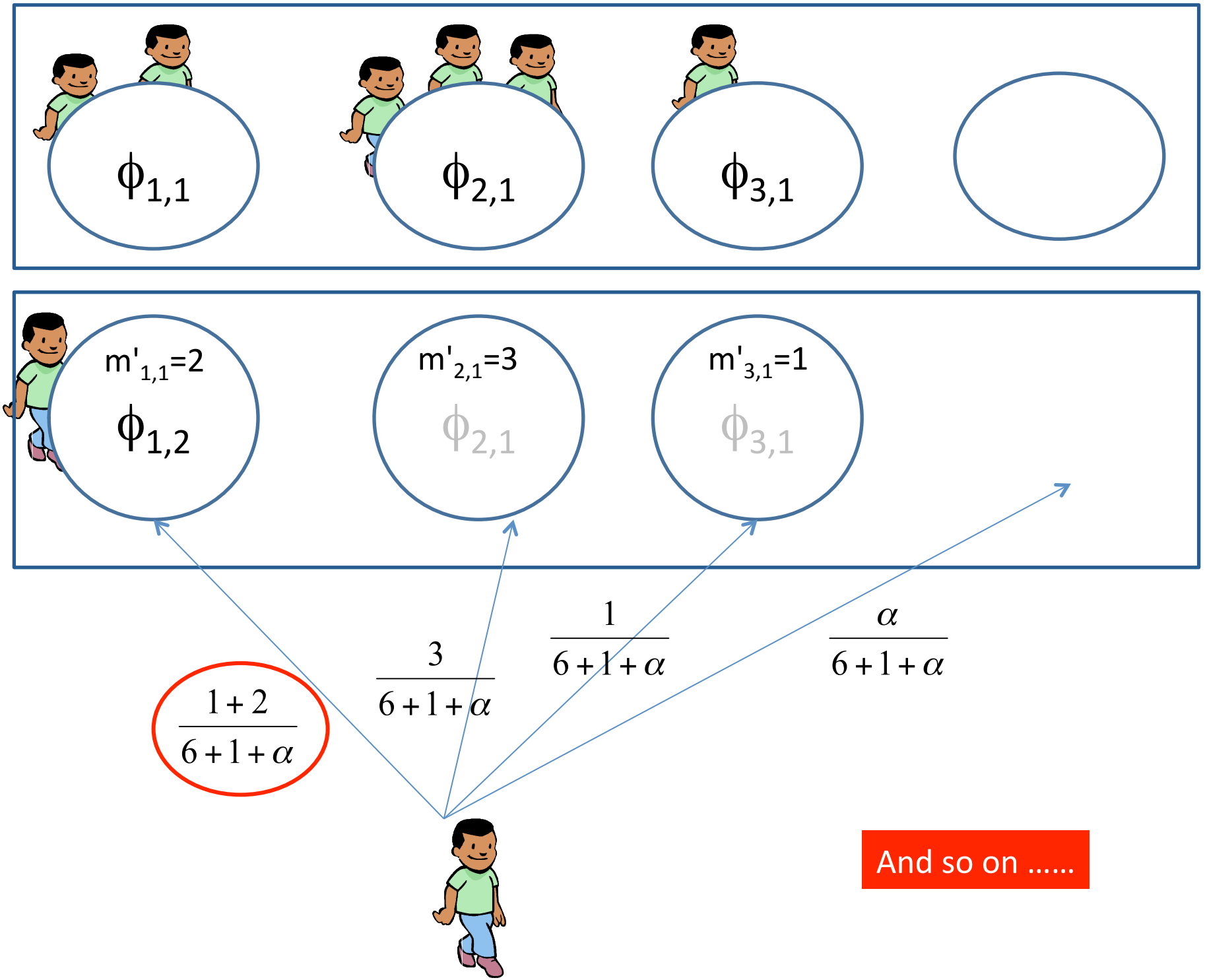
T=2

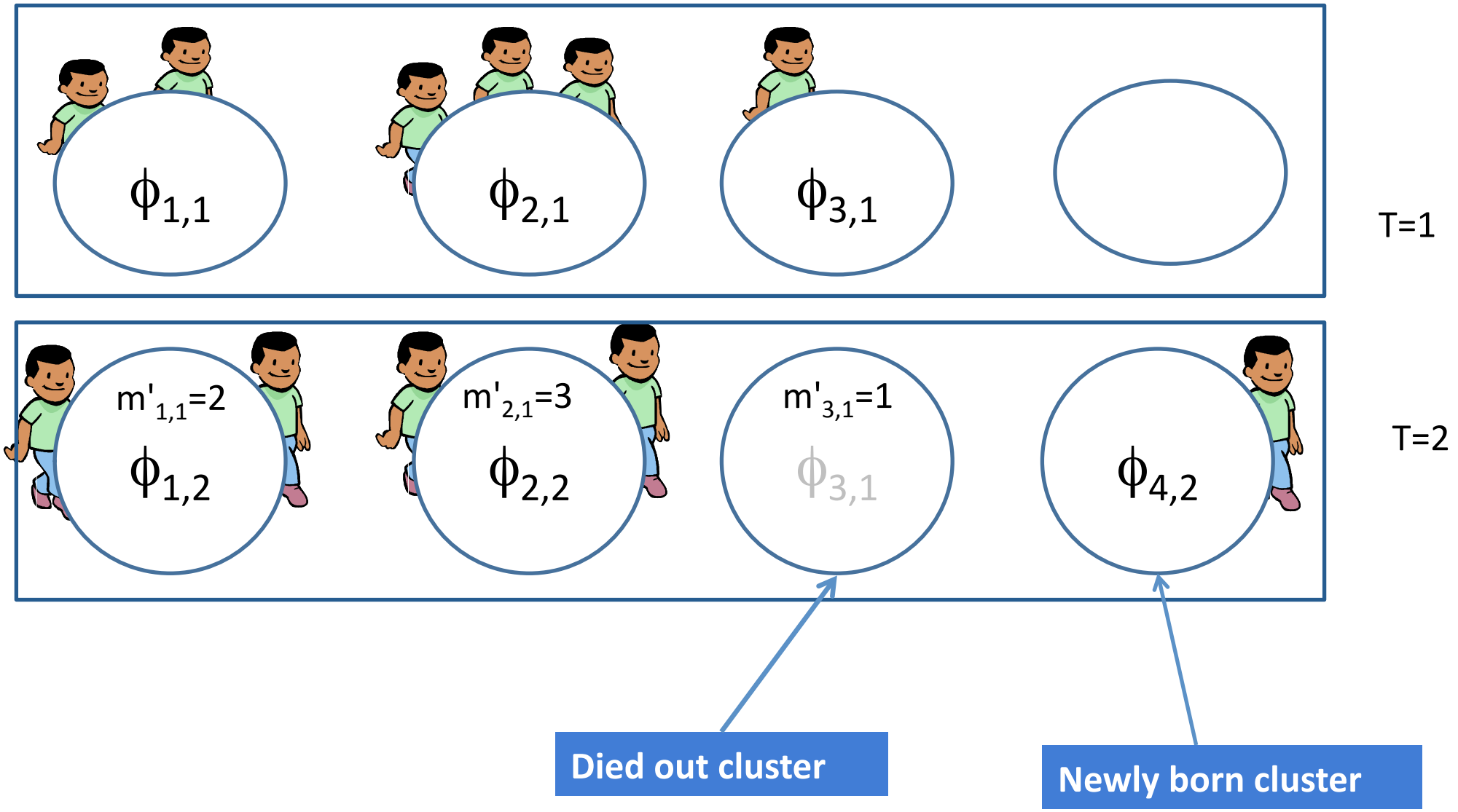

$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$m'_{1,1}=2$
$\phi_{1,1}$
$m'_{2,1}=3$
$\phi_{2,1}$
$m'_{3,1}=1$
$\phi_{3,1}$
T=2


$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
m'$_{1,1}$=2
m'$_{2,1}$=3
m'$_{3,1}$=1
$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=2
$\frac{2}{6+\alpha}$
$\frac{3}{6+\alpha}$
$\frac{1}{6+\alpha}$
$\frac{\alpha}{6+\alpha}$

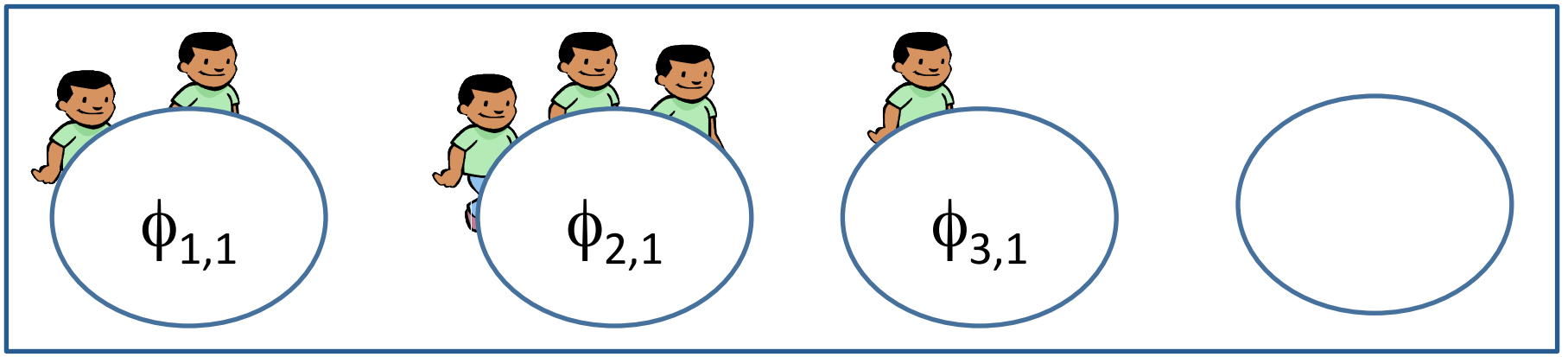

$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
m'$_{1,1}$=2
$\phi_{1,1}$
m'$_{2,1}$=3
$\phi_{2,1}$
m'$_{3,1}$=1
$\phi_{3,1}$
T=2
$\frac{2}{6+\alpha}$


$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
$m'_{1,1}=2$
$\phi_{1,2}$
$m'_{2,1}=3$
$\phi_{2,1}$
$m'_{3,1}=1$
$\phi_{3,1}$
T=2
Sample $\phi_{1,2} \sim P(.| \phi_{1,1})$

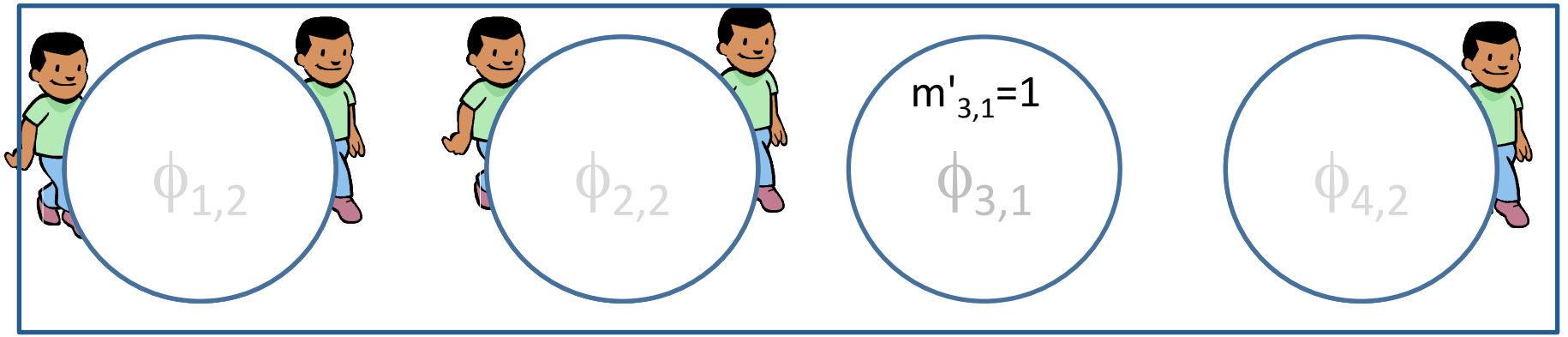

$\phi_{1,1}$
$\phi_{2,1}$
$\phi_{3,1}$
T=1
m'$_{1,1}$=2
$\phi_{1,2}$
m'$_{2,1}$=3
$\phi_{2,1}$
m'$_{3,1}$=1
$\phi_{3,1}$
T=2
$\frac{1+2}{6+1+\alpha}$
$\frac{3}{6+1+\alpha}$
$\frac{1}{6+1+\alpha}$
$\frac{\alpha}{6+1+\alpha}$
And so on ……



At the end of epoch 2



T=1



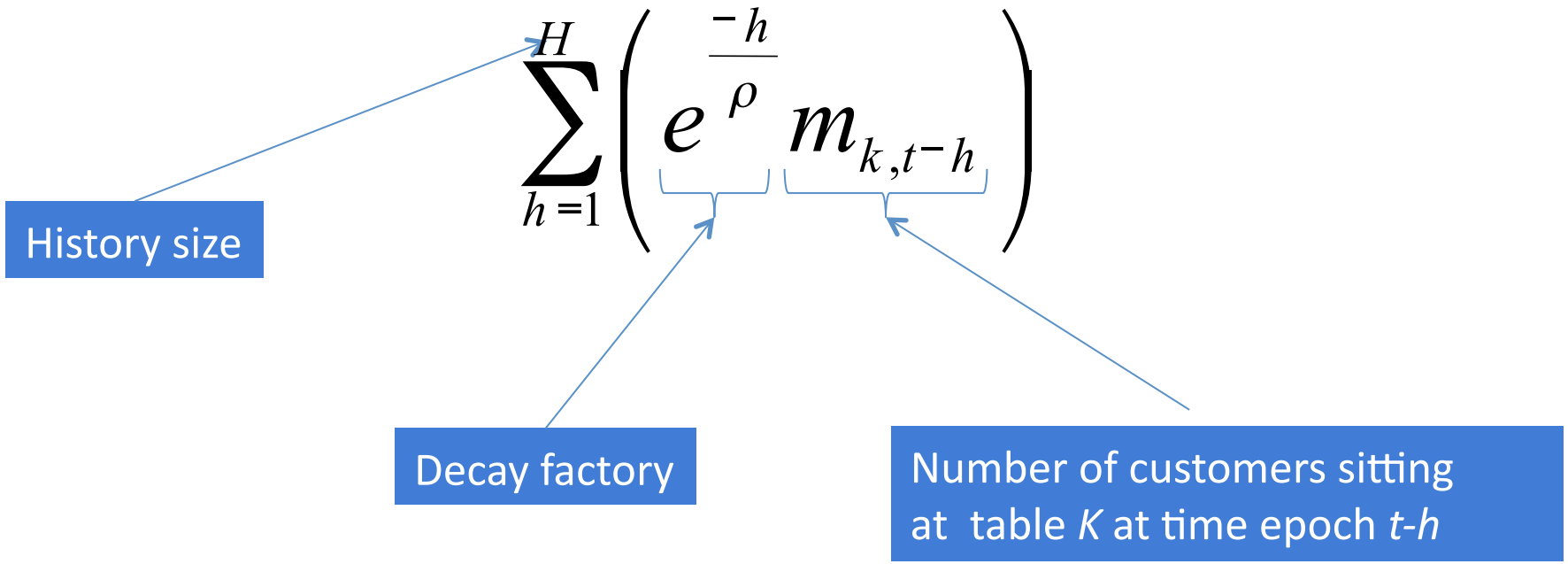
T=2

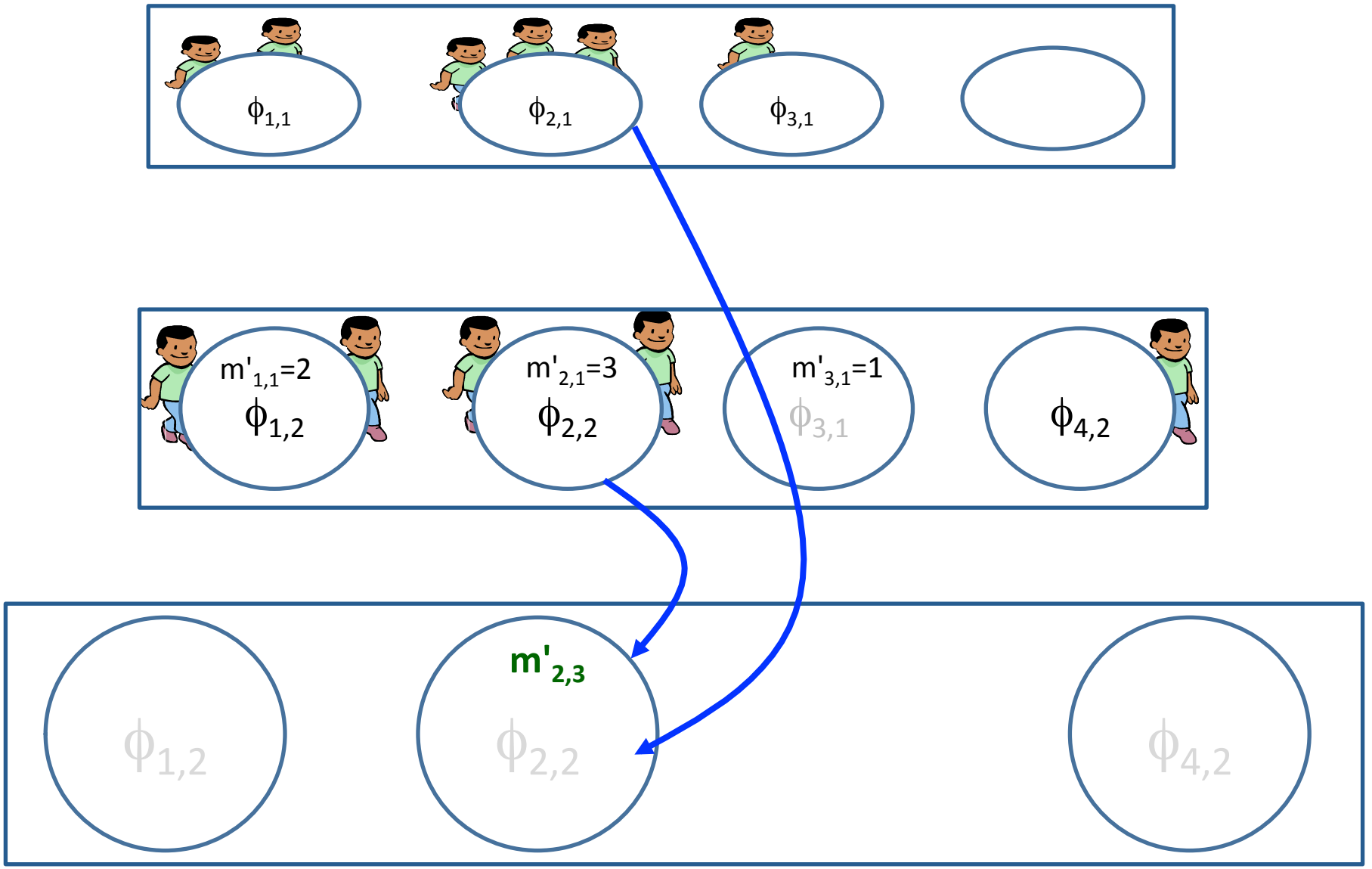
$m'_{1,2}=2$ $m'_{2,2}=2$ $m'_{4,2}=1$

T=3

Recurrent Chinese Restaurant Process

- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table *k* at time *t is*





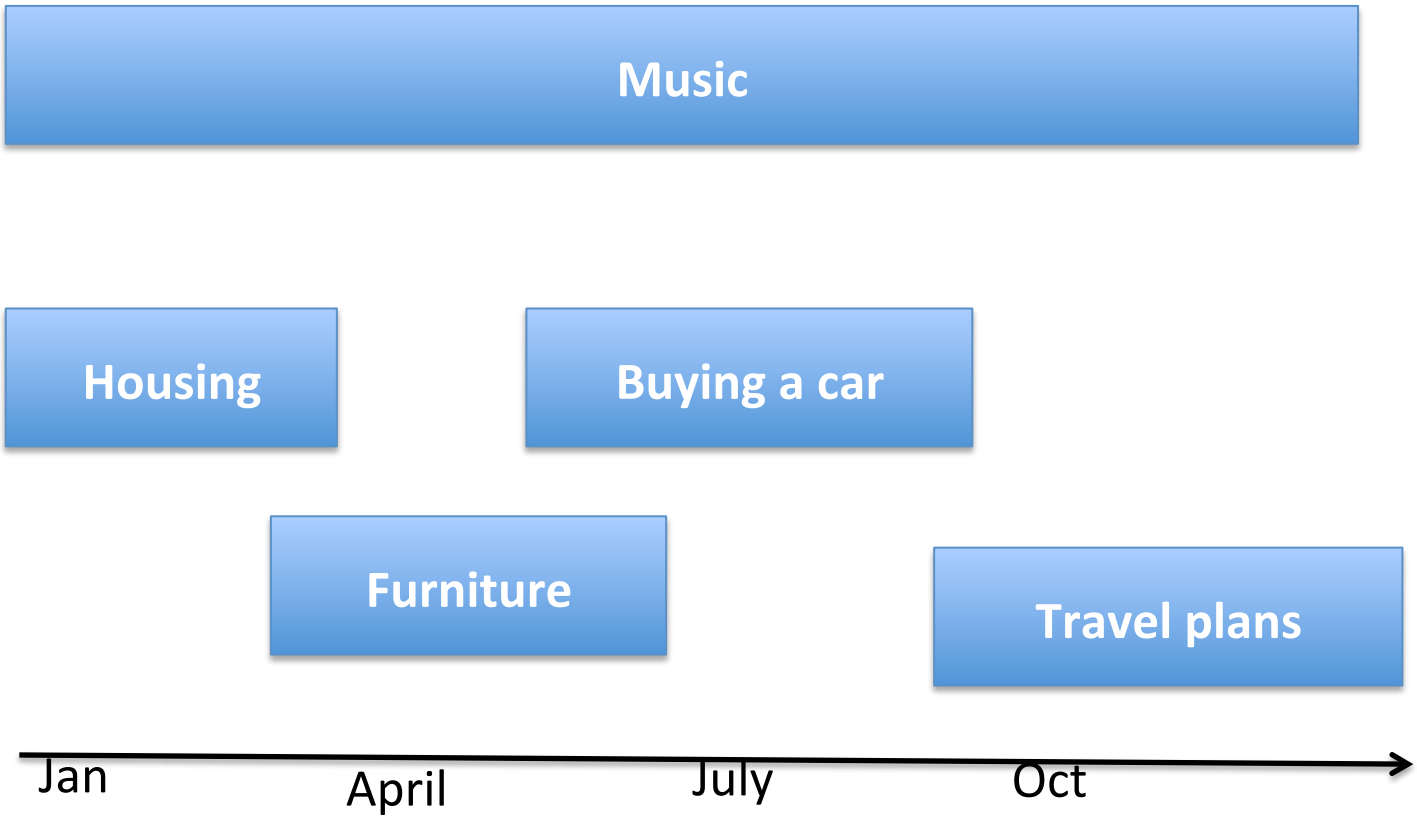
$$m'_{2,3} = \sum_{h=1}^{H} \left(e^{\frac{-h}{\rho}} m_{k,t-h} \right)$$

4.2 Online Distributed Inference

Tracking Users Interest

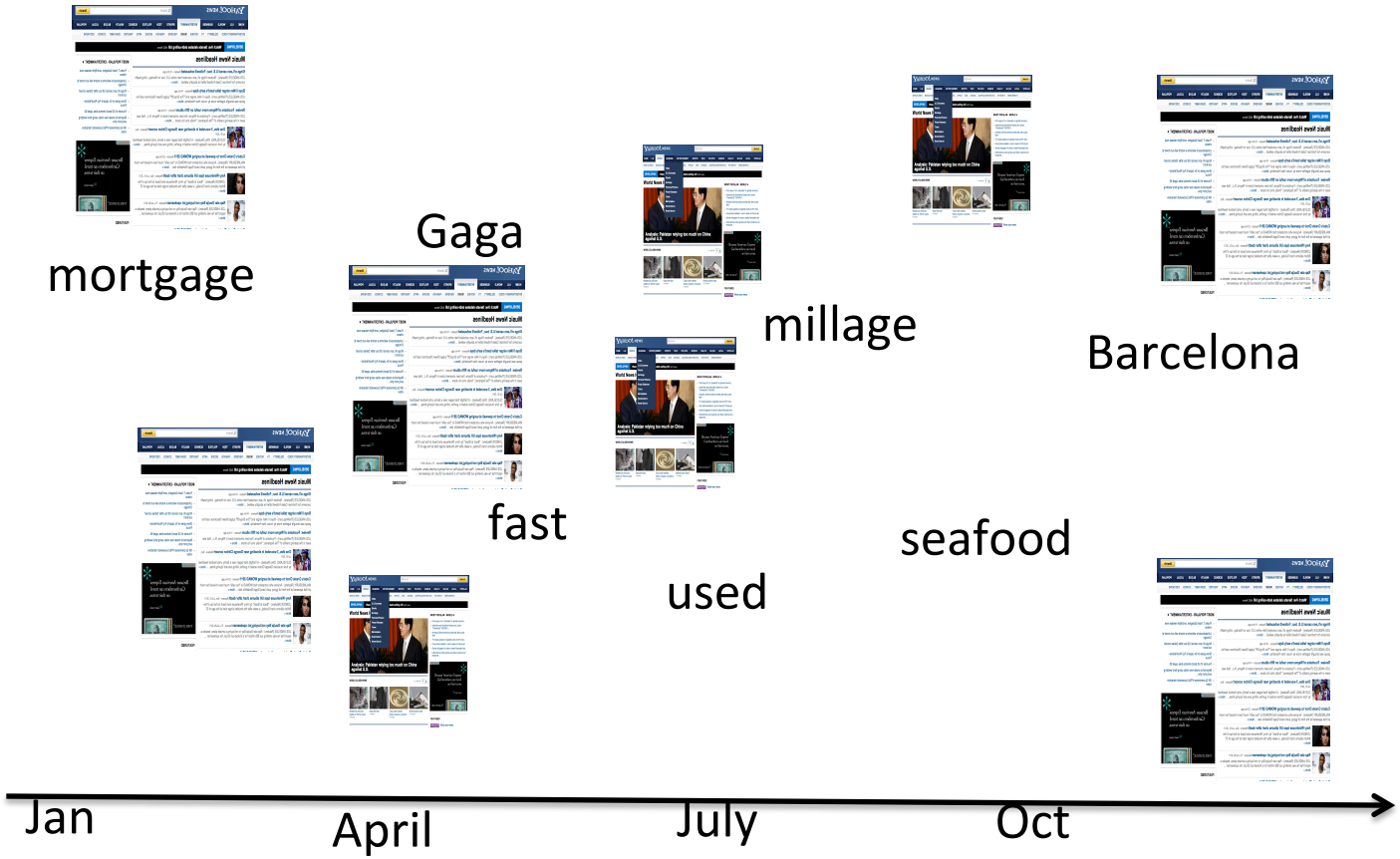
Characterizing User Interests

- 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

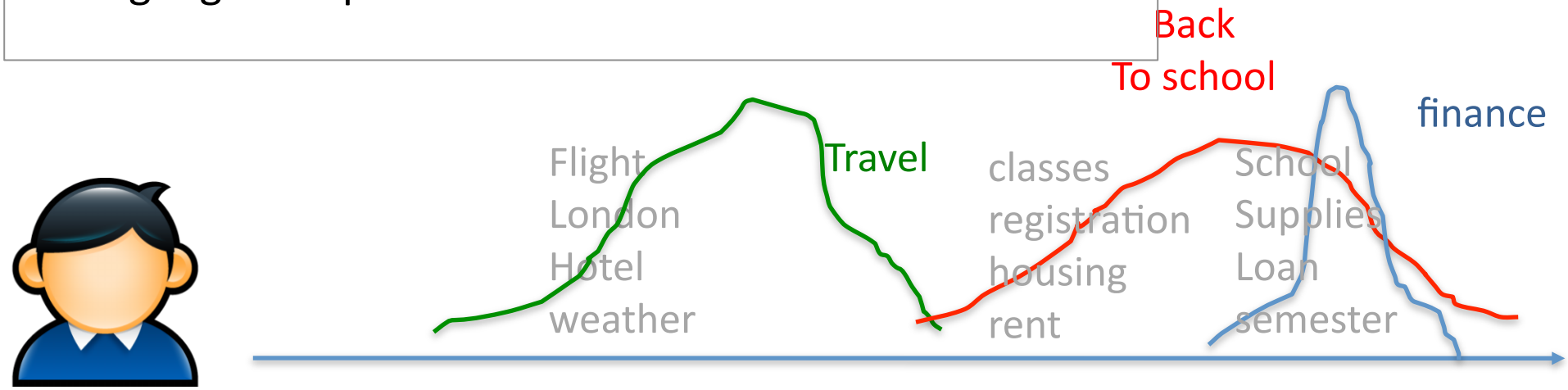
Problem formulation

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

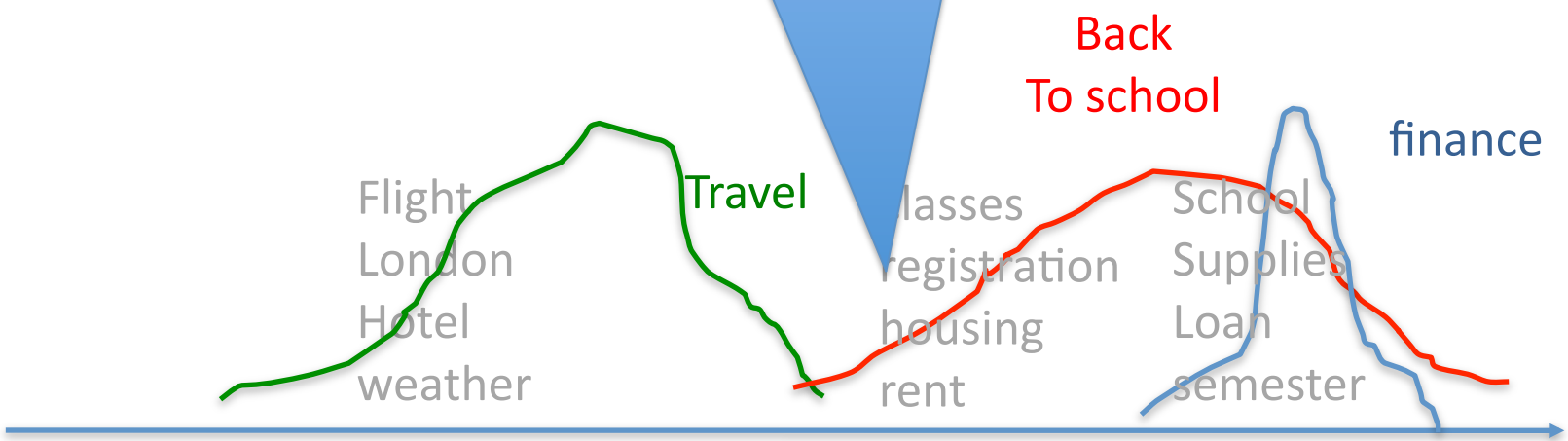


Problem formulation

When to show a financing ad?

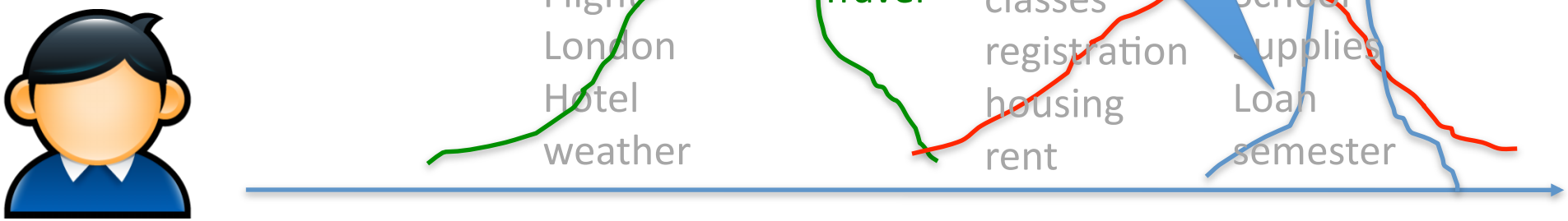
Problem formulation

When to show a financing ad?



Problem formulation

When to show a financing ad?



Problem formulation

When to show a hotel ad?

Problem formulation

When to show a hotel ad?

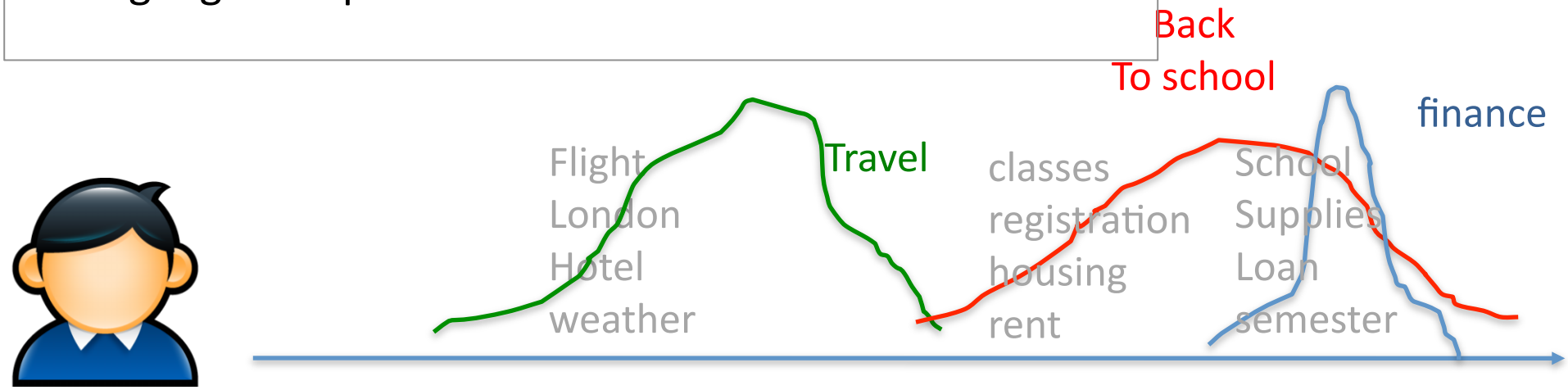
Problem formulation

Input

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

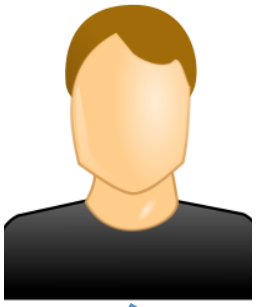
Output

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



Mixed-Membership Formulation

Objects



Job Hiring
speed price
part-time Camry
Career opening
bonus package

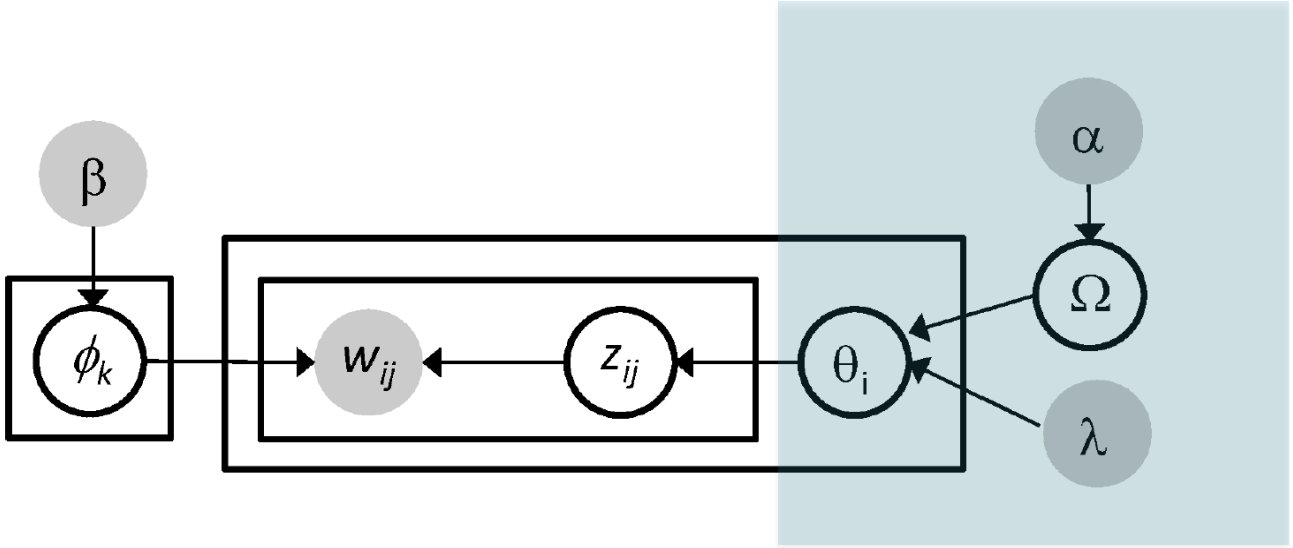
card diet calories
loan recipe milk
Weight lb kg

Degree of membership

Mixtures

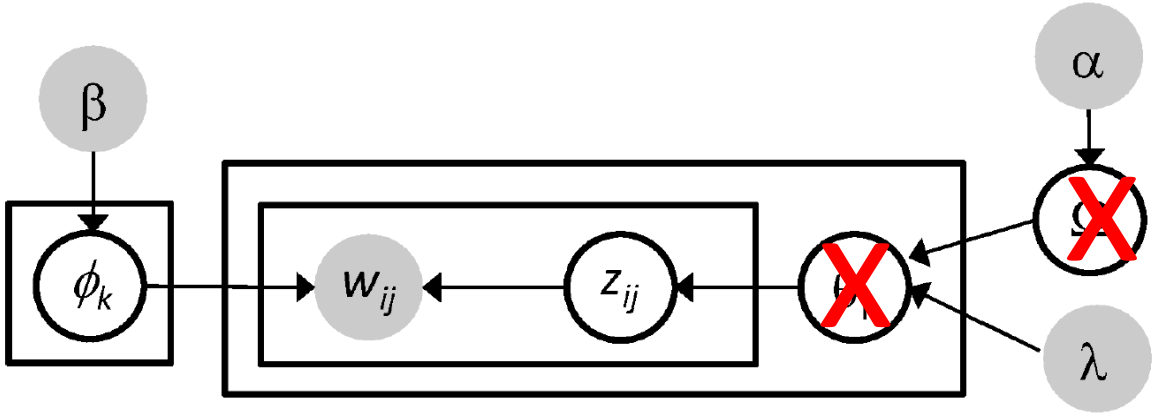
Diet	Cars	Job	Finance
Recipe	Car	job	Bank
Chocolate	Blue	Career	Online
Pizza	Book	Business	Credit
Food	Kelley	Assistant	Card
Chicken	Prices	Hiring	debt
Milk	Small	Part-time	portfolio
Butter	Speed	Receptionis	Finance
Powder	large	t	Chase

In Graphical Notation

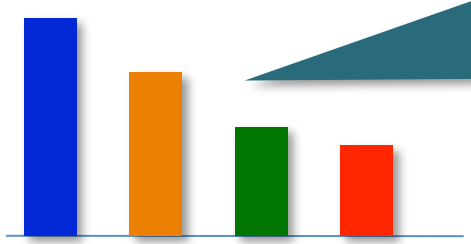


1. Draw once $\Omega|\alpha \sim \mathrm{Dir}(\alpha/K)$.
2. Draw each topic $\phi_k|\beta \sim \mathrm{Dir}(\beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i|\lambda, \Omega \sim \mathrm{Dir}(\lambda\Omega)$.
   (b) For each word
      (a) Draw a topic $z_{ij}|\theta_d \sim \mathrm{Mult}(\theta_i)$.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

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

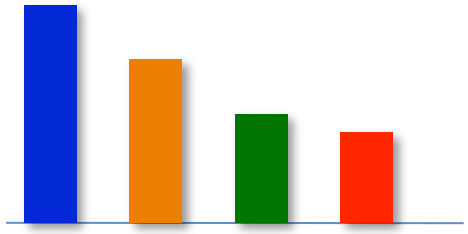
Food Chicken

Topic word-distributions

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

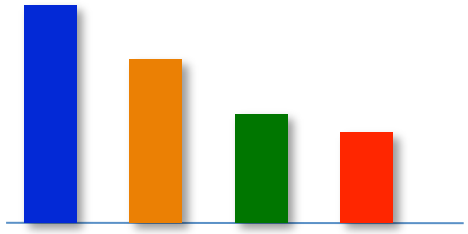
Food Chicken

.....

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 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

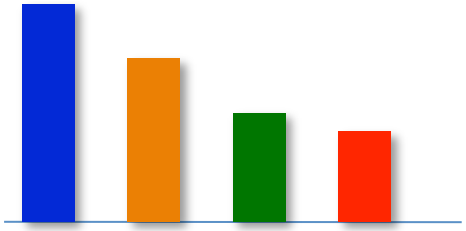
.....

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
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Blue
Book
Kelley
Prices
Small
Speed
large

job
Career
Business
Assistant
Hiring
Part-time
Receptio
nist

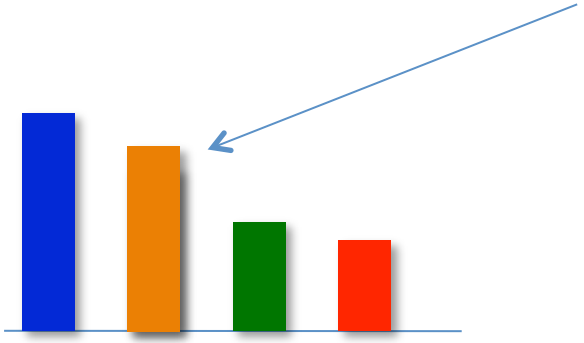
Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

Food Chicken
pizza .....

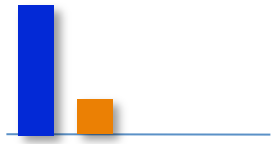
Car speed offer
camry accord career

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 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

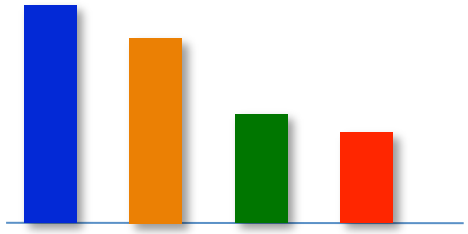


Food Chicken pizza .....

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

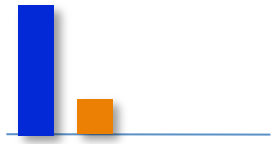


Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

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

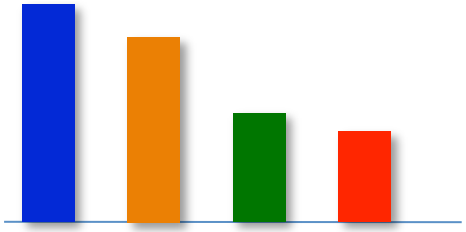


Food Chicken
pizza millage

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
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

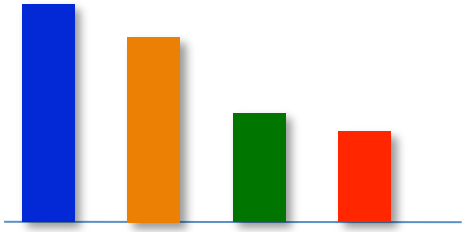
Food Chicken
pizza millage

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

Car speed offer
camry accord career

At time t



At time t+1

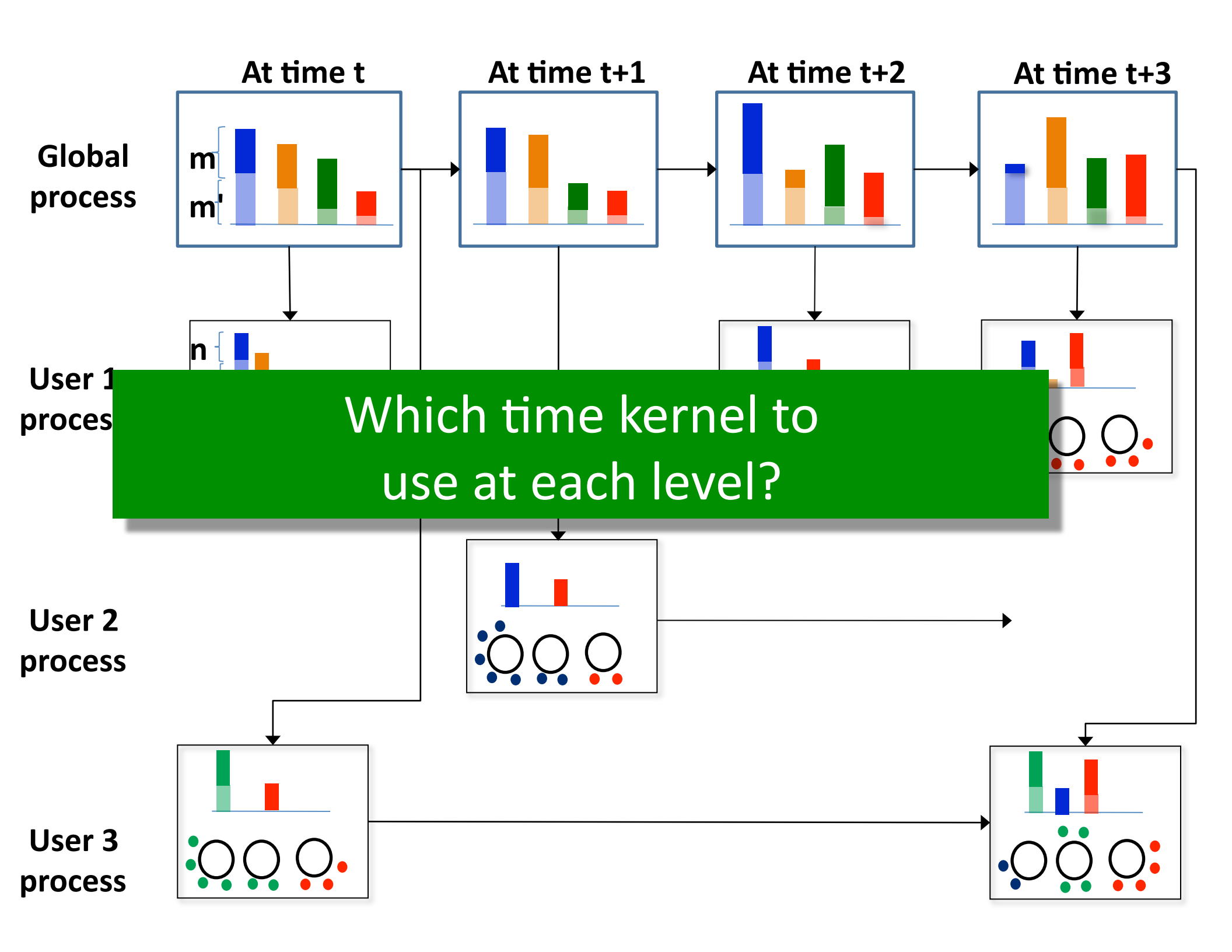
Recipe Chocolate Pizza Food Chicken Milk Butter Powder	Car Blue Book Kelley Prices Small Speed large	job Career Business Assistant Hiring Part-time Receptio nist	Bank Online Credit Card debt portfolio Finance Chase

Food Chicken
pizza millage

Car speed offer
camry accord career

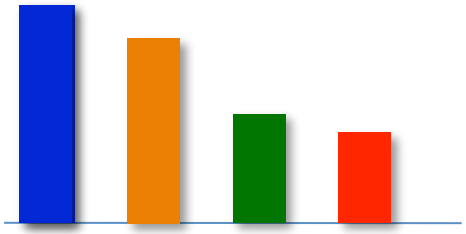
Build a dynamic model

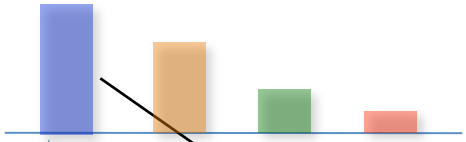
Connect each level using a RCRP


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

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

Pseudo counts

$= * \exp^{\frac{-1}{\lambda}}$

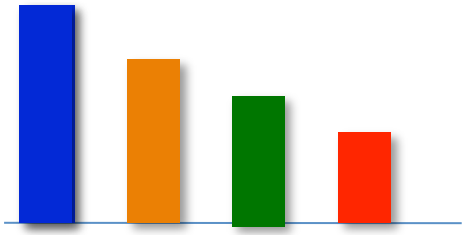
Decay factor

Car speed offer
camry accord career

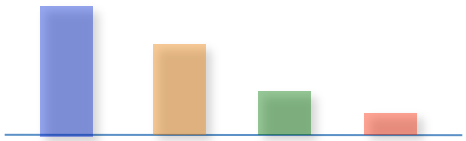
Observation 1

-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



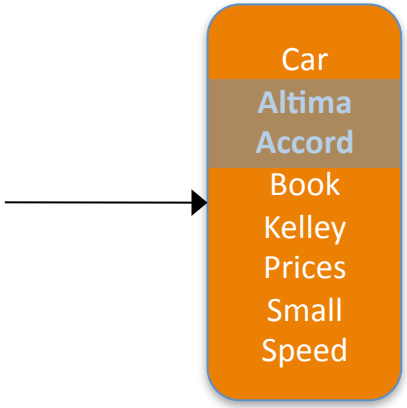
At time t+1



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



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

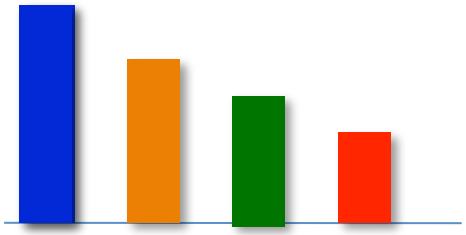
Intuition

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

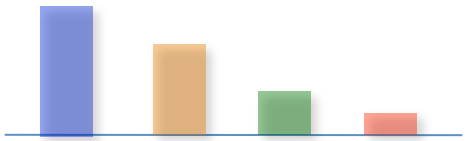
Observation 1

-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

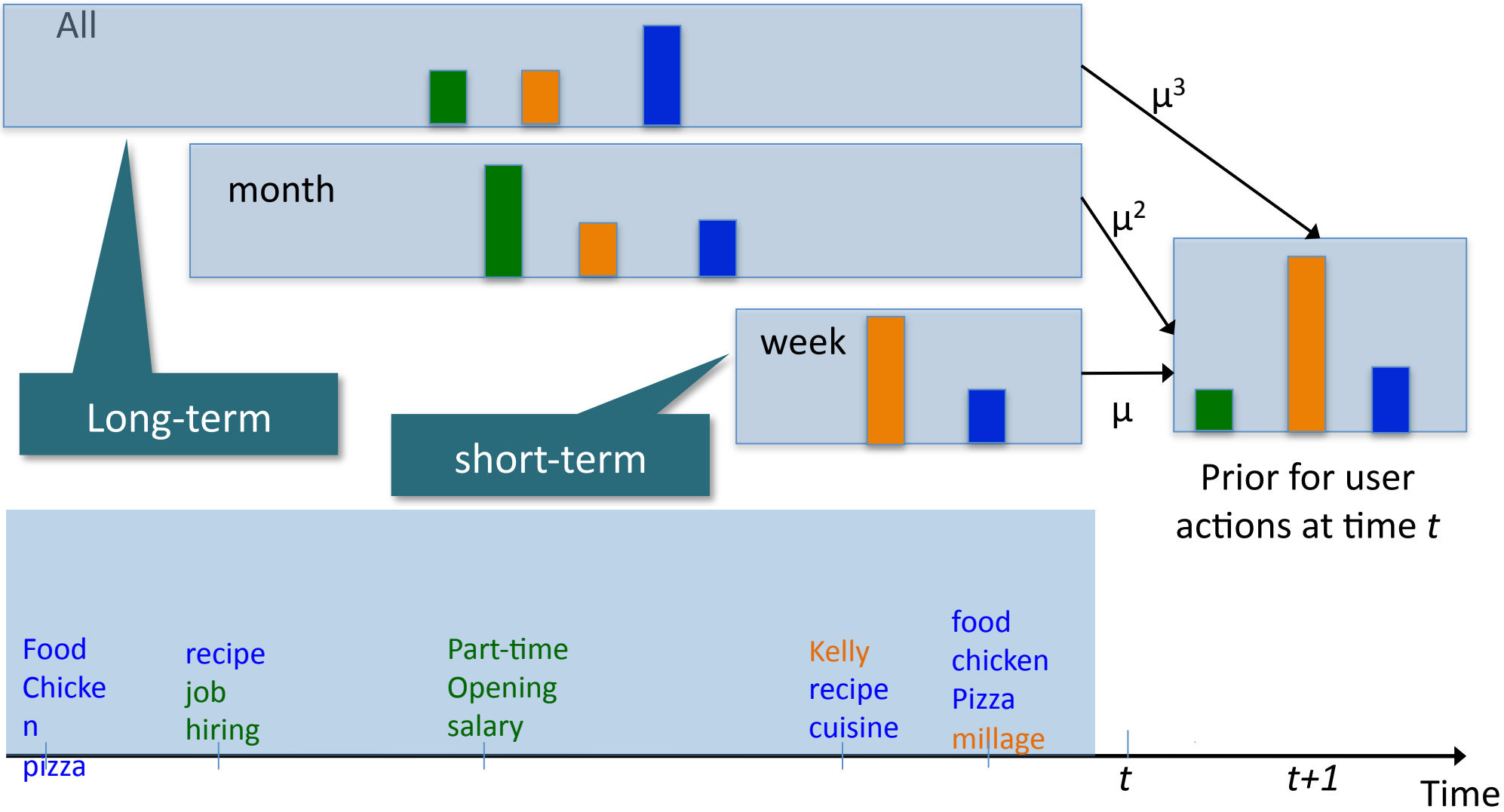
Food Chicken
pizza millage

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

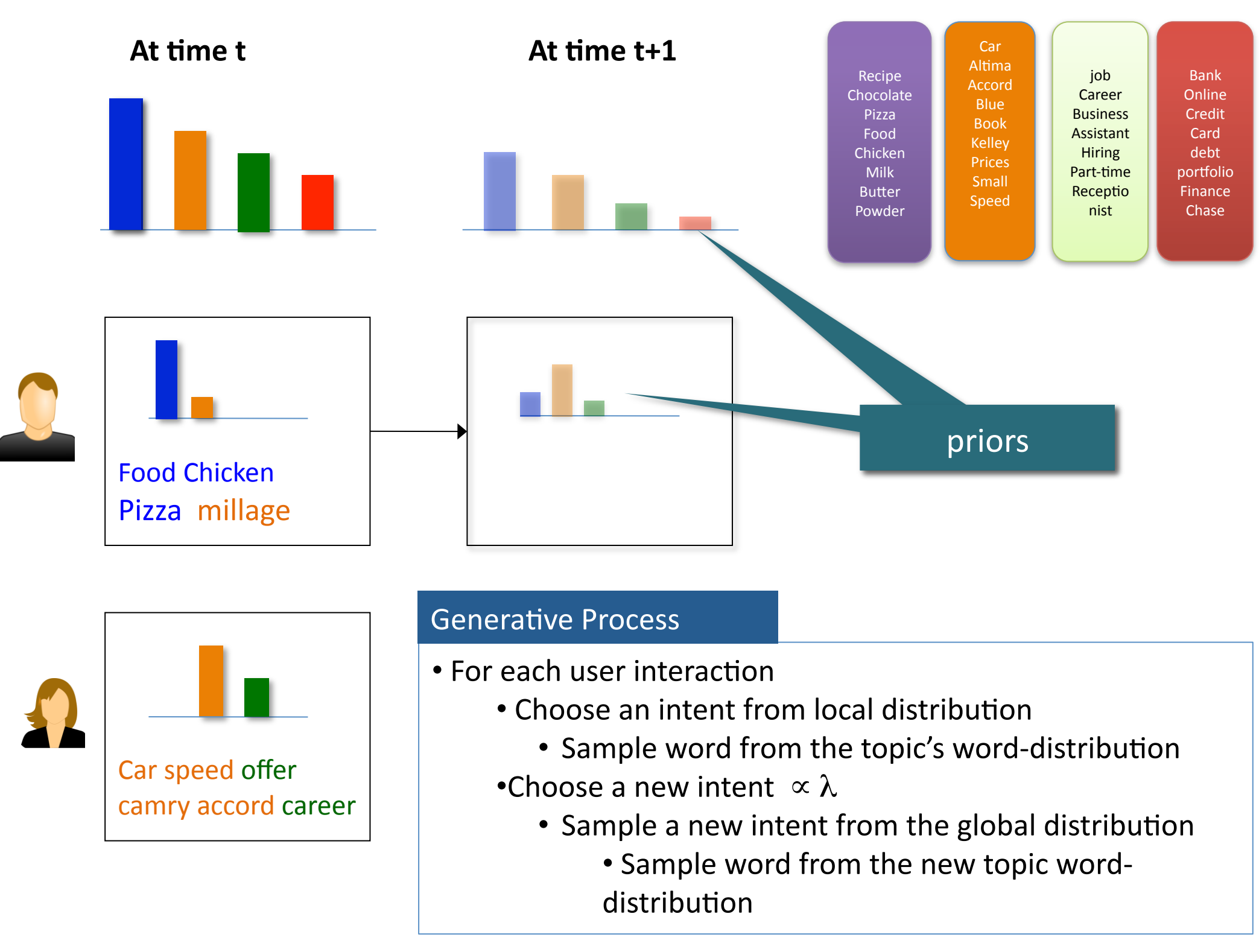
Car speed offer
camry accord career

Observation 2

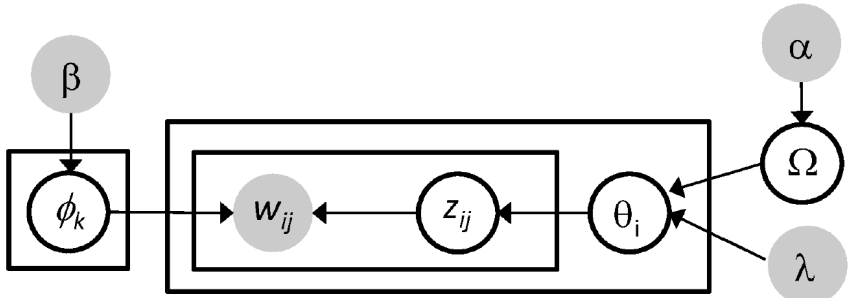
- User prior at time t+1 is a mixture of the user short and long term interest

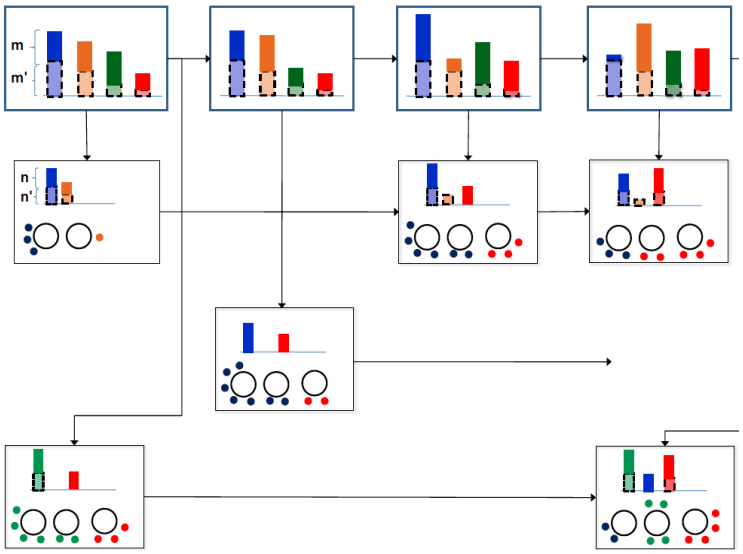


Diet	Cars	Job	Finance
Recipe	Car	job	Bank
Chocolate	Blue	Career	Online
Pizza	Book	Business	Credit
Food	Kelley	Assistant	Card
Chicken	Prices	Hiring	debt
Milk	Small	Part-time	portfolio
Butter	Speed	Receptionis	Finance
Powder	large	t	Chase


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
Receptio
nist
Bank
Online
Credit
Card
debt
portfolio
Finance
Chase
priors
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 the topic's word-distribution
•Choose a new intent ∝ λ
• Sample a new intent from the global distribution
• Sample word from the new topic word-distribution

Polya-Urn
RCRF Process

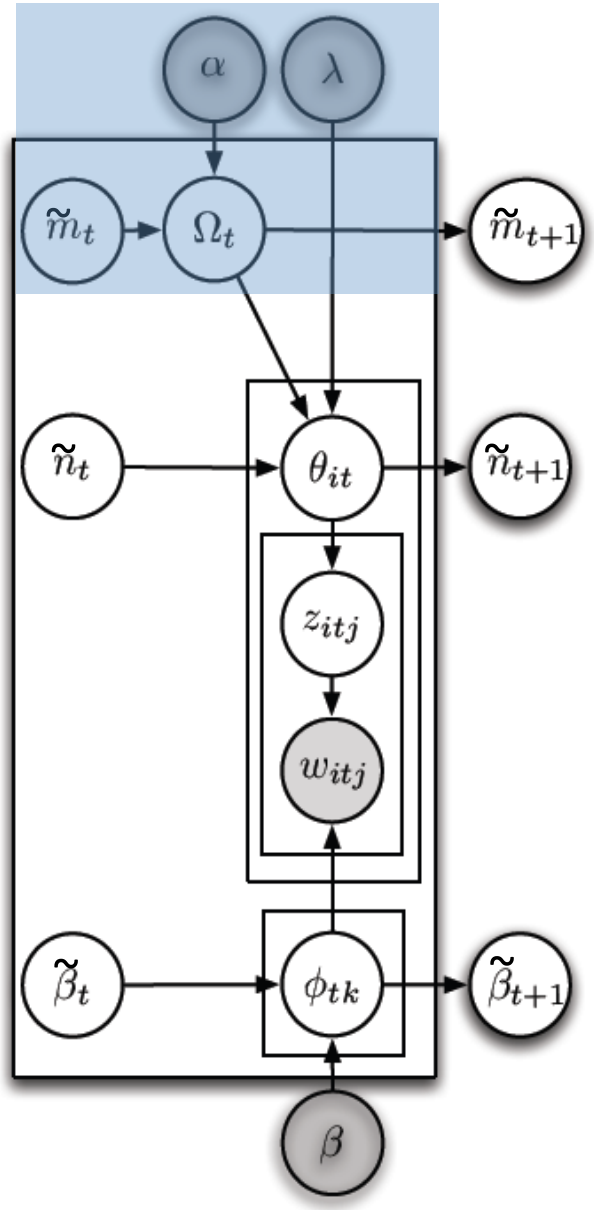




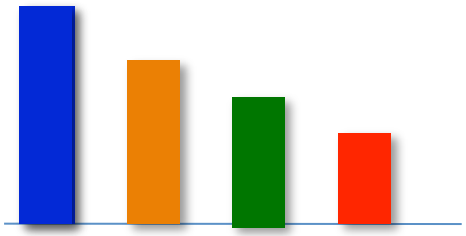
?

Simplified Graphical Model

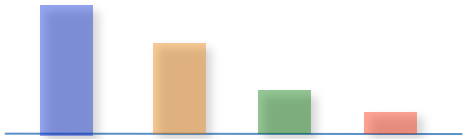
1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t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



At time t



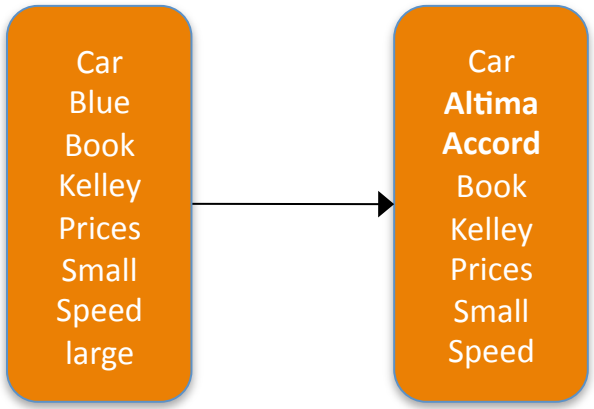
At time t+1

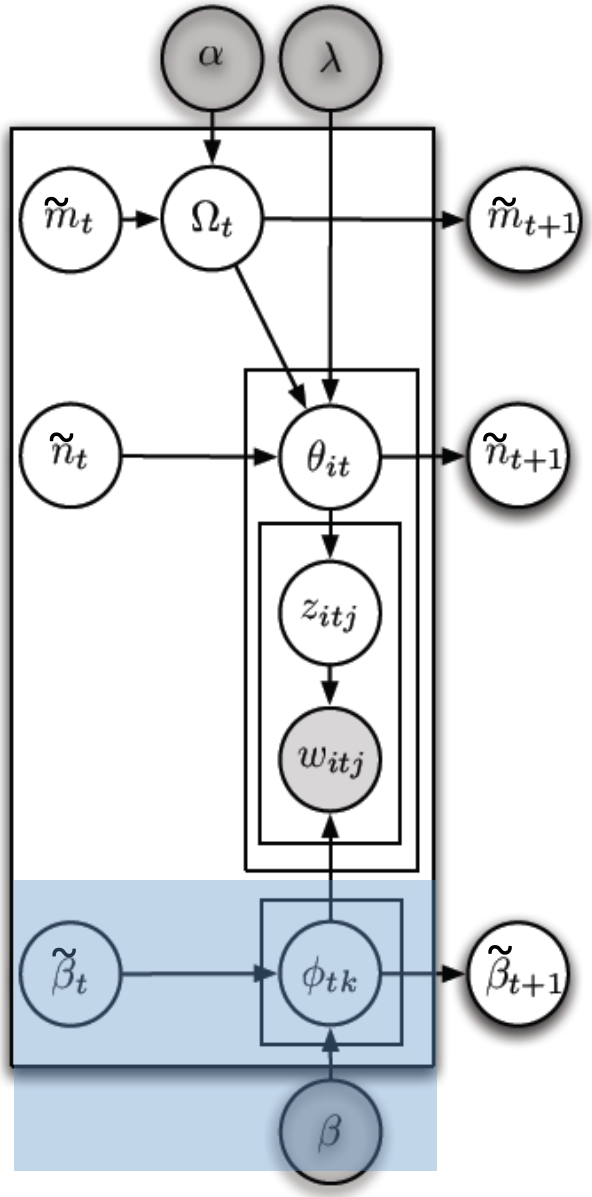


Simplified Graphical Model

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

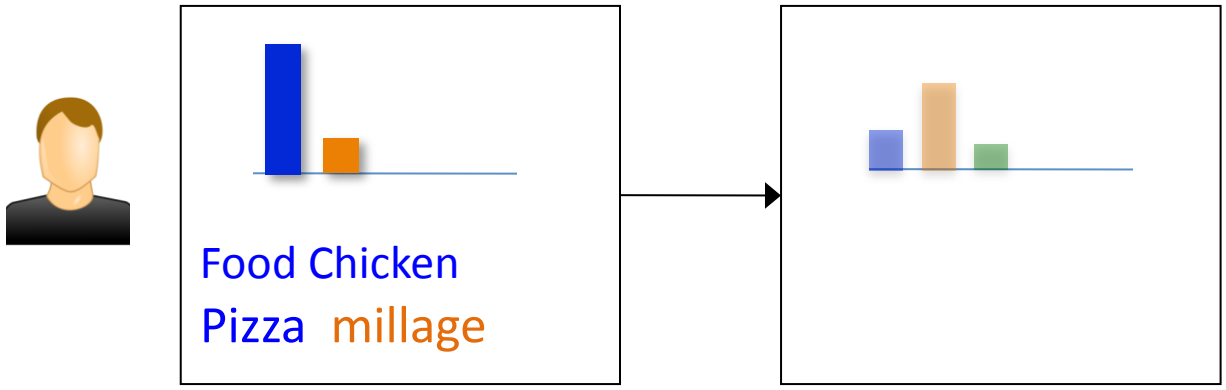
$$\tilde{\beta}_{kw}^t = \sum_{h=1}^{t-1} \exp^{\frac{h-t}{\kappa_0}} n_{kw}^h$$

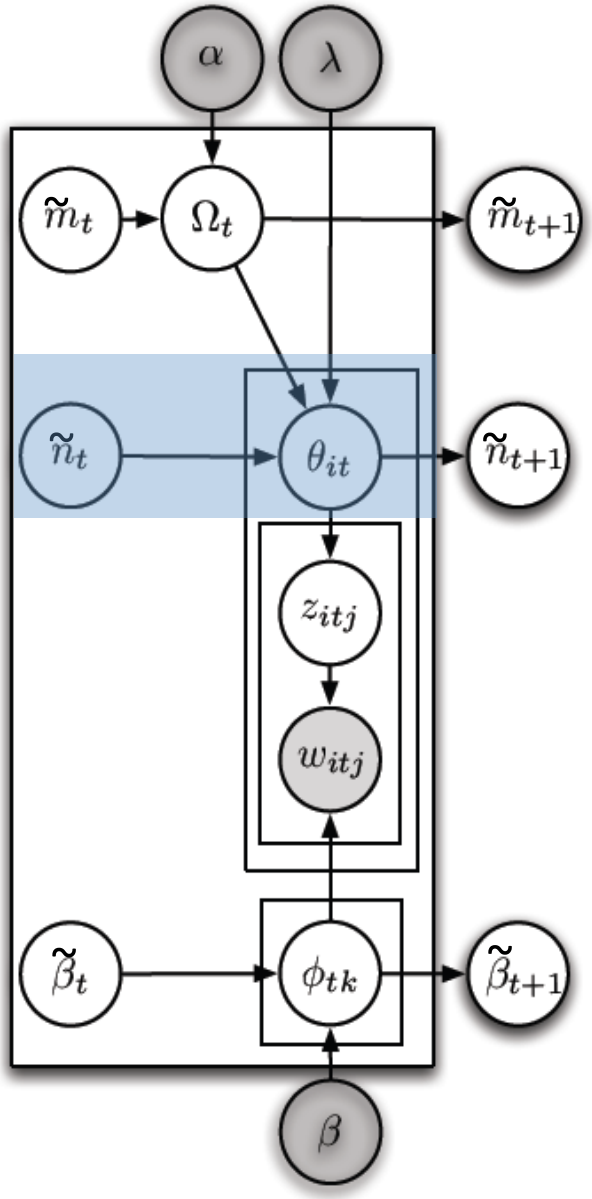




Simplified Graphical Model

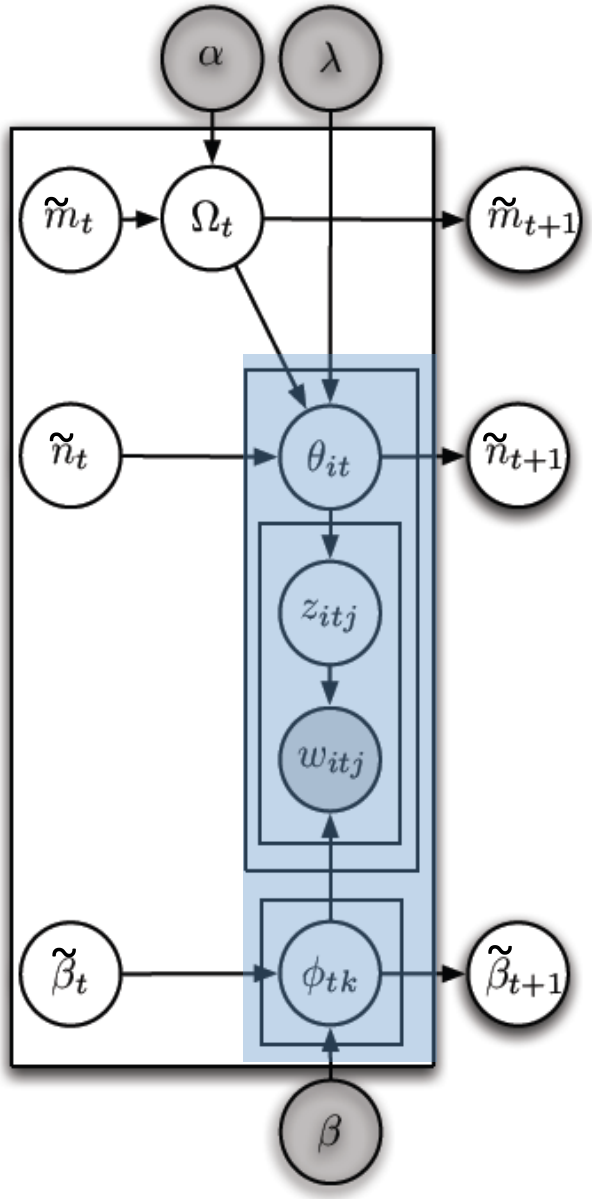
1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t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





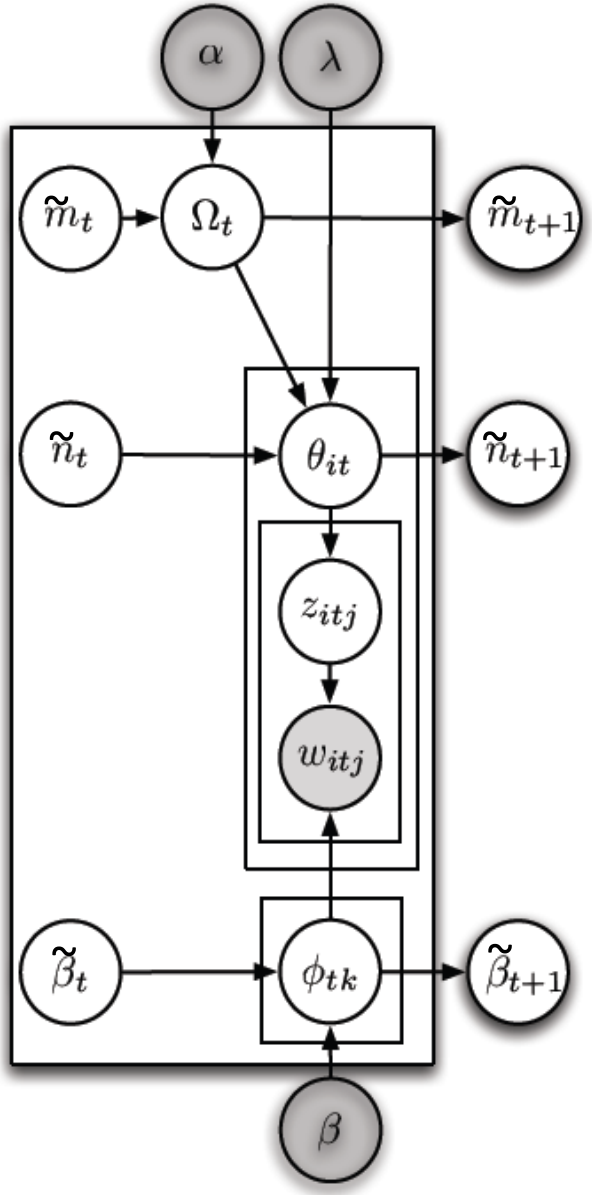
Simplified Graphical Model

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 $w_{in}^t|z_{ij}^t, \phi^t \sim \mathrm{Multi}(\phi_{z_{ij}^t}^t)$.



Simplified Graphical Model

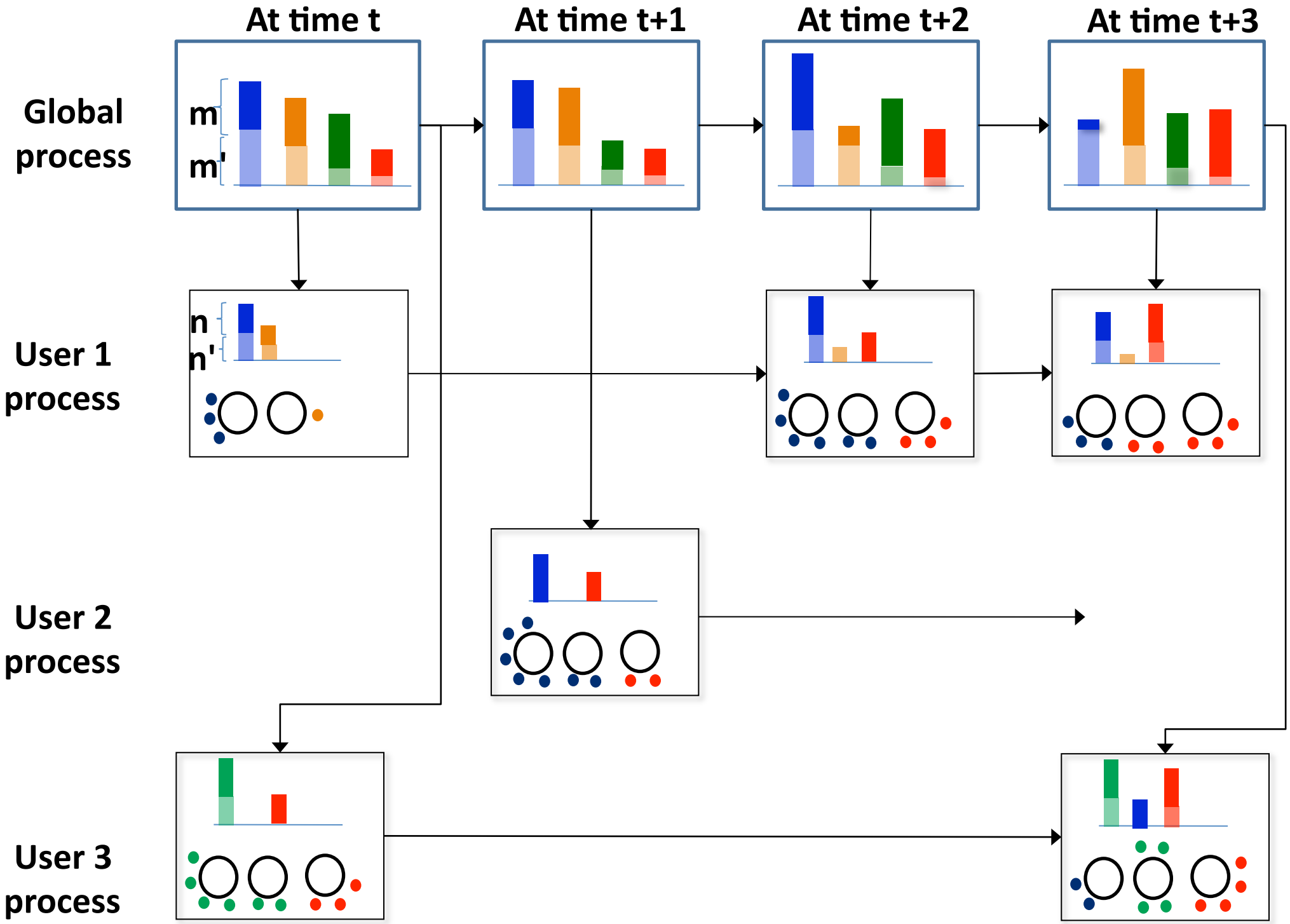
1. Draw once $\Omega^t|\alpha, \tilde{m}^t \sim \mathrm{Dir}\left(\tilde{\mathbf{m}}^t + \alpha/K\right)$.
2. Draw each topic, $\phi_k^t|\beta, \tilde{\beta}_k^t \sim \mathrm{Dir}(\tilde{\beta}_k^t + \beta)$.
3. For each user i:
   (a) Draw topic proportions $\theta_i^t|\lambda, \Omega^t, \tilde{\mathbf{n}}_i^t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



Topics evolve over time? ✔

User's intent evolve over time? ✔

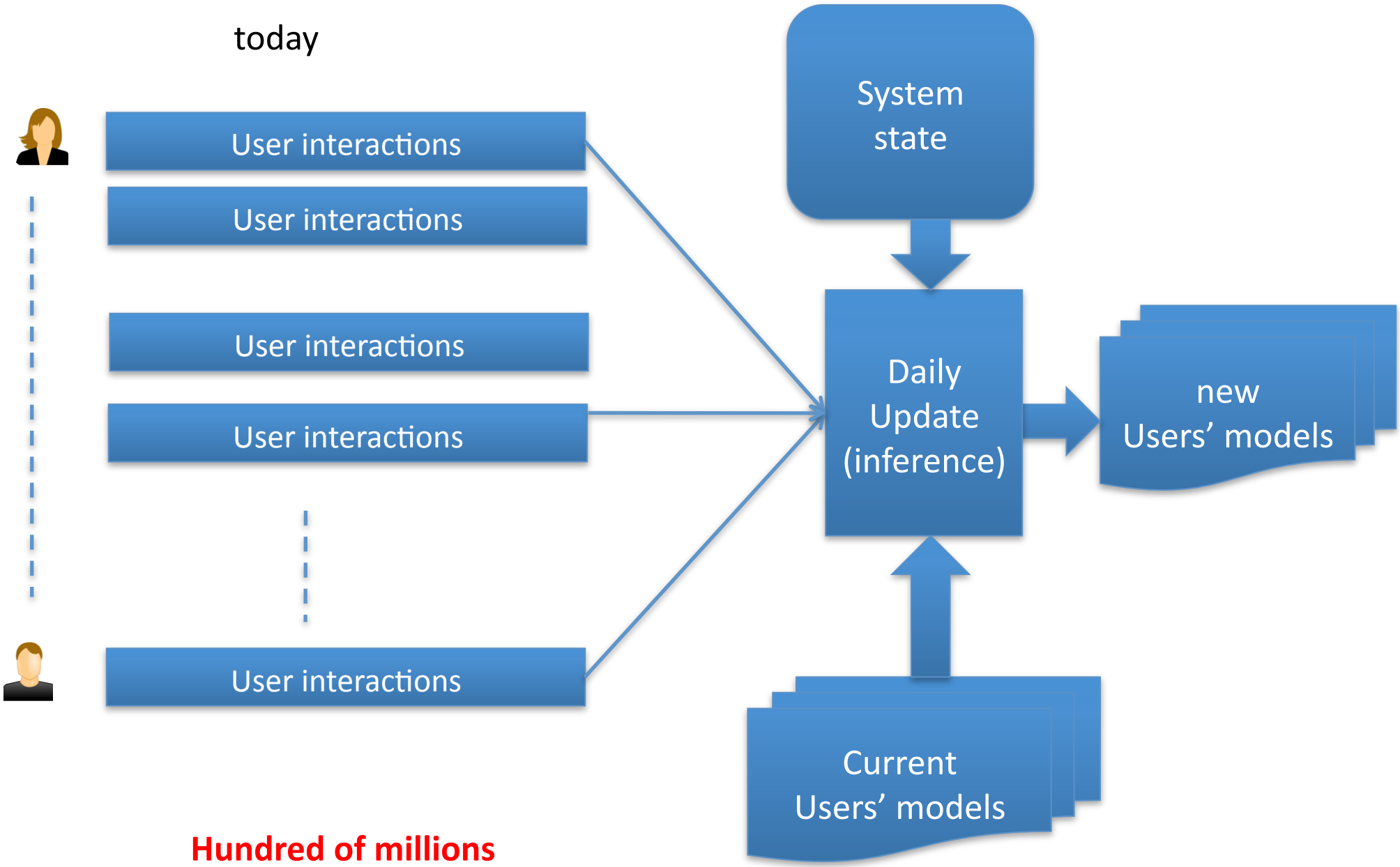
Capture long and term interests of users? ✔


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

4.2 Online Distributed Inference

Work Flow

Work Flow

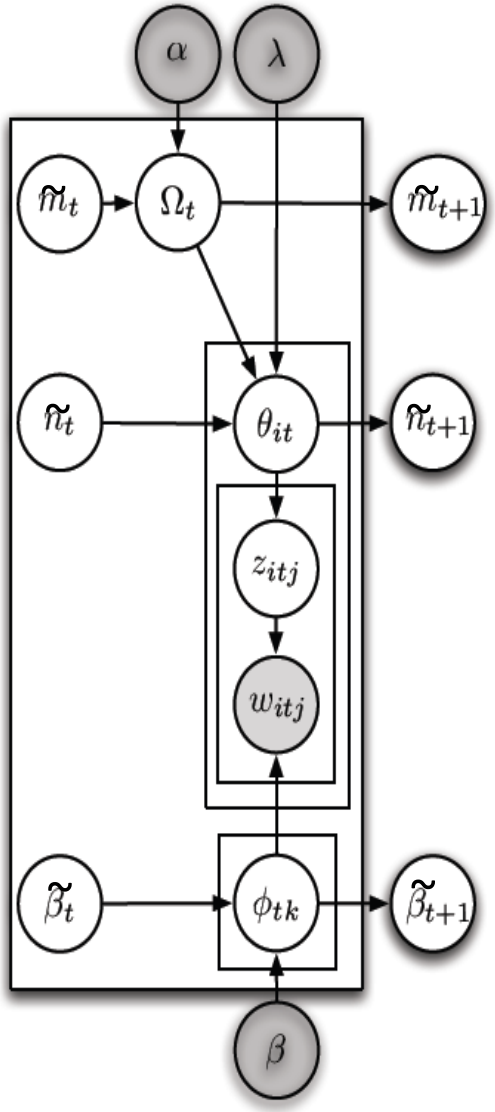

today
User interactions
User interactions
User interactions
User interactions
User interactions
System state
Daily Update (inference)
new Users' models
Current Users' models
Hundred of millions

Online Scalable Inference

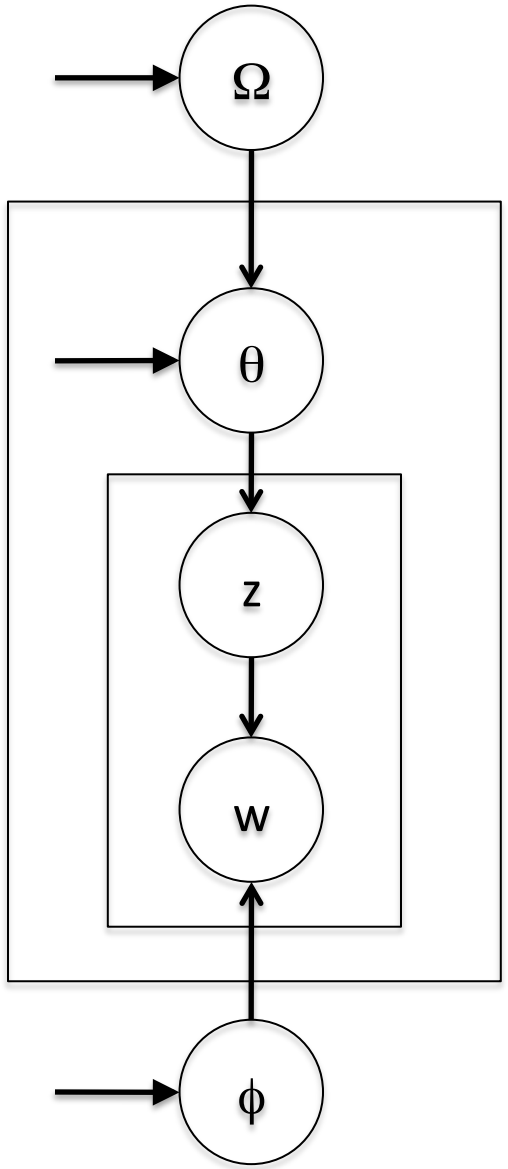
- Online algorithm
  - Greedy 1-particle filtering algorithm
  - Works well in practice
  - Collapse all multinomials except Ω_t
    - This makes distributed inference easier
  - At each time *t:*

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

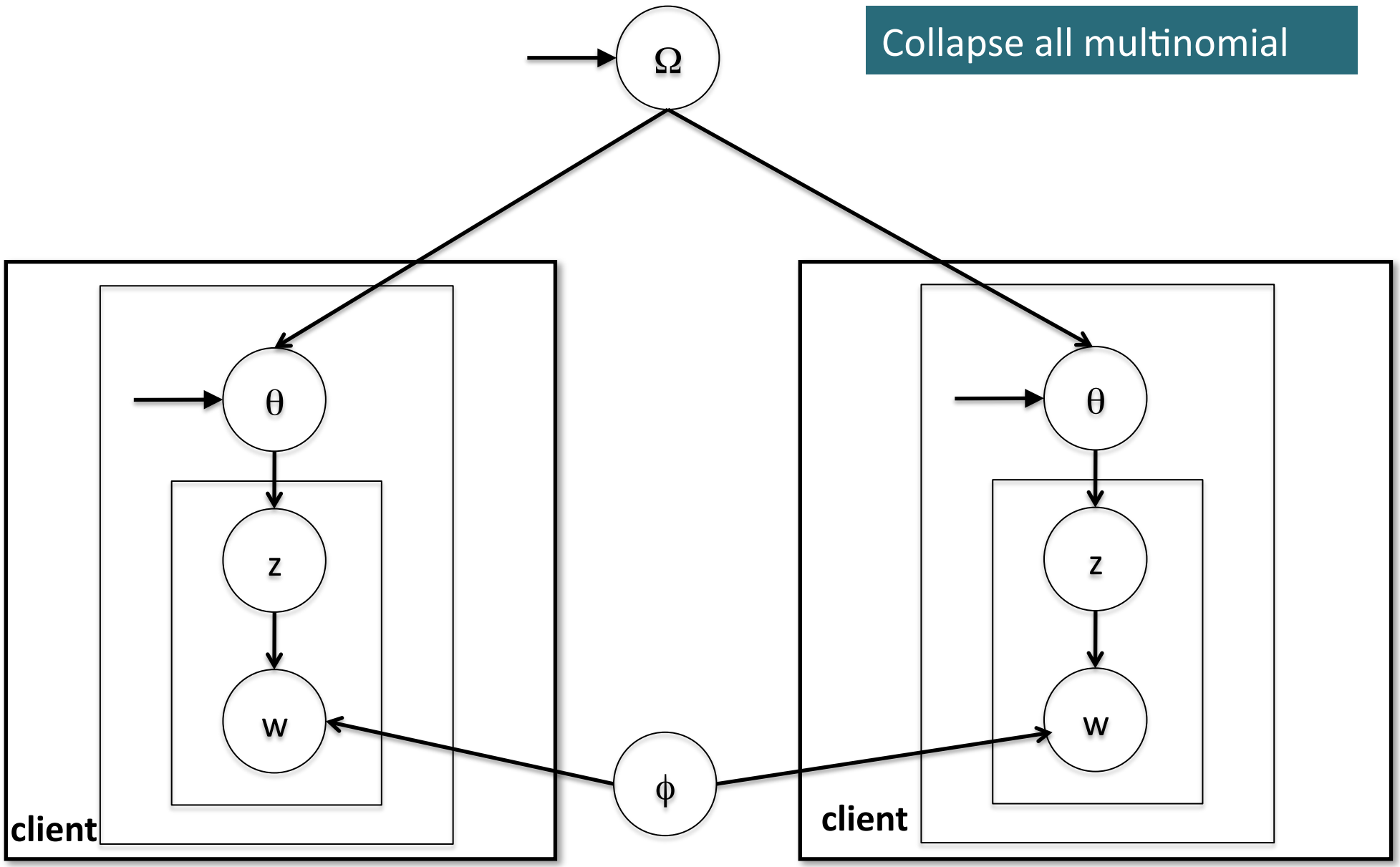
- Distributed scalable implementation
  - Used first part architecture as a subroutine
  - Added synchronous sampling capabilities



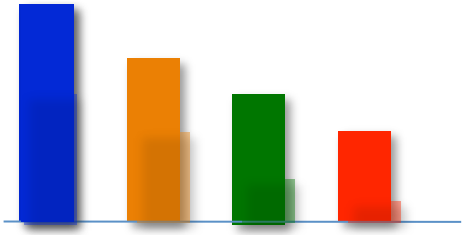
Distributed Inference (at time t)



Distributed Inference (at time t)



After collapsing



Recipe Chocolate Pizza Food Chicken Milk Butter Powder

Car Blue Book Kelley Prices Small Speed large

job Career Business Assistant Hiring Part-time Receptionist

Bank Online Credit Card debt portfolio Finance Chase

Use Star-Synchronization

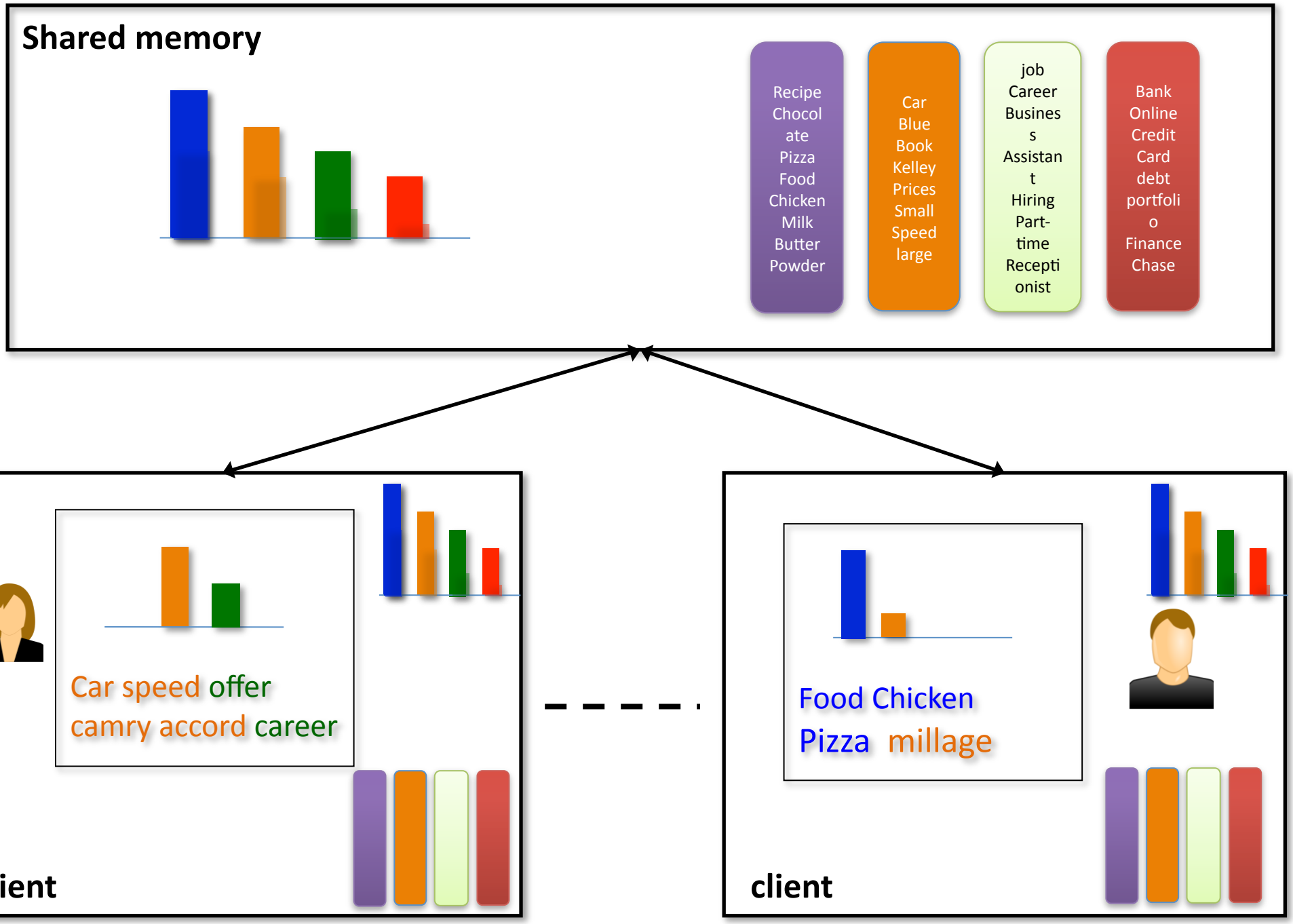
Car speed offer
camry accord career

client

Food Chicken
Pizza millage

client

Fully Collapsed



Distributed Inference (at time t)

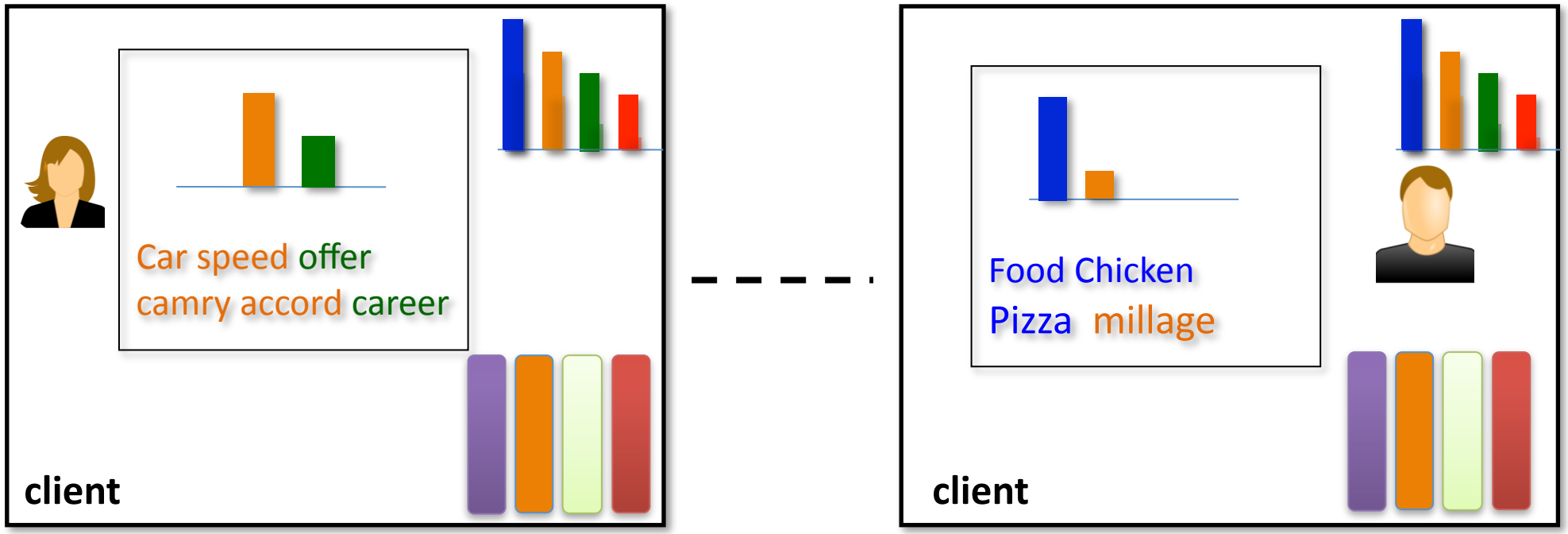
$$P(z_{ij}^t = k | w_{ij}^t = w, \Omega^t, \tilde{\mathbf{n}}_i^t) \propto$$

$$\left(n_{ik}^{t,-j} + \tilde{n}_{ik}^t + \lambda \frac{m_k^t + \tilde{m}_k^t + \frac{\alpha}{K}}{\sum_l m_l^t + \tilde{m}_l^t + \frac{\alpha}{K}} \right) \frac{n_{kw}^{t,-j} + \tilde{\beta}_{kw}^t + \beta}{\sum_l n_{kl}^{t,-j} + \tilde{\beta}_{kl}^t + \beta}$$

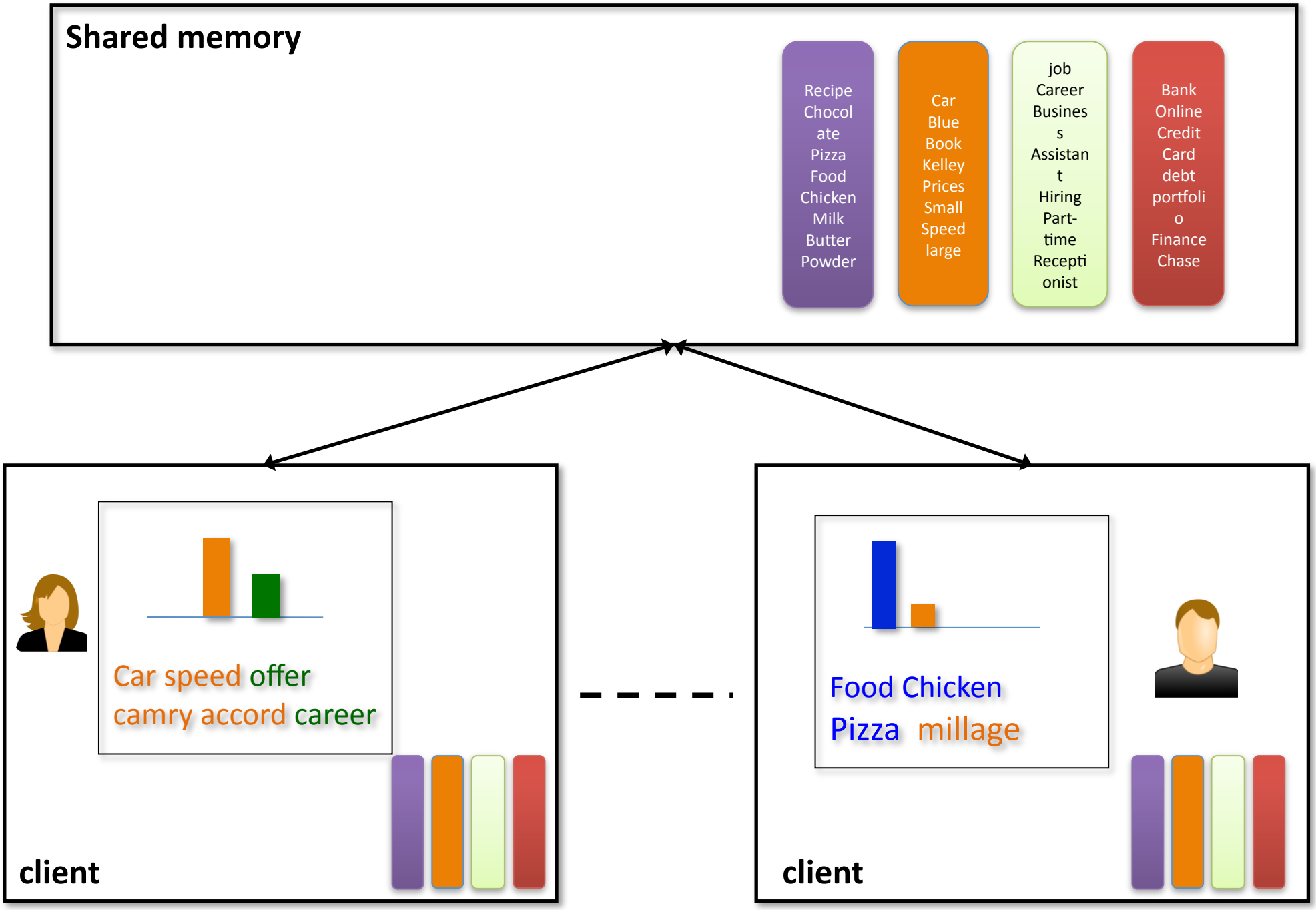
Local trend

Global trend

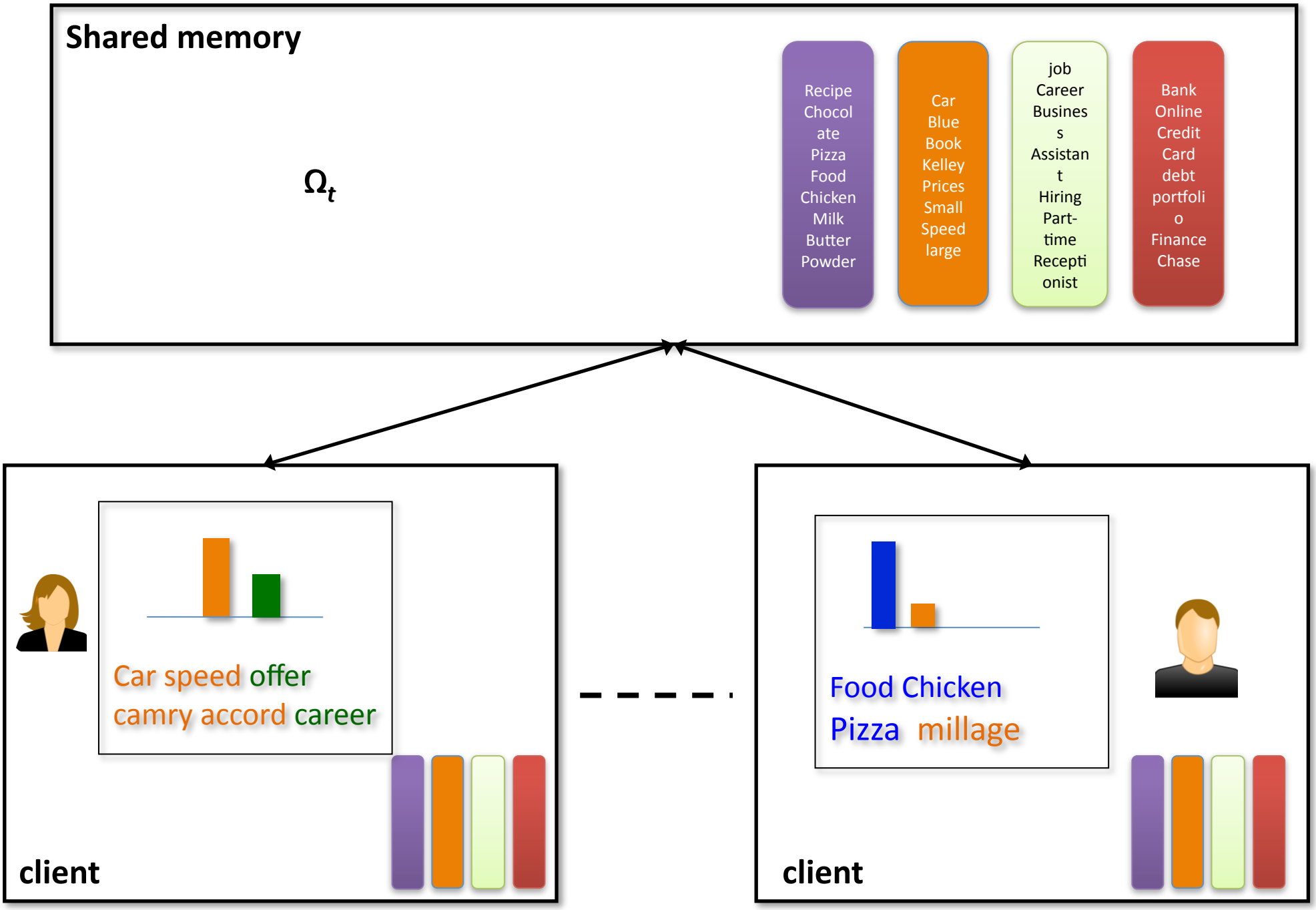
Topic factor



Semi-Collapsed



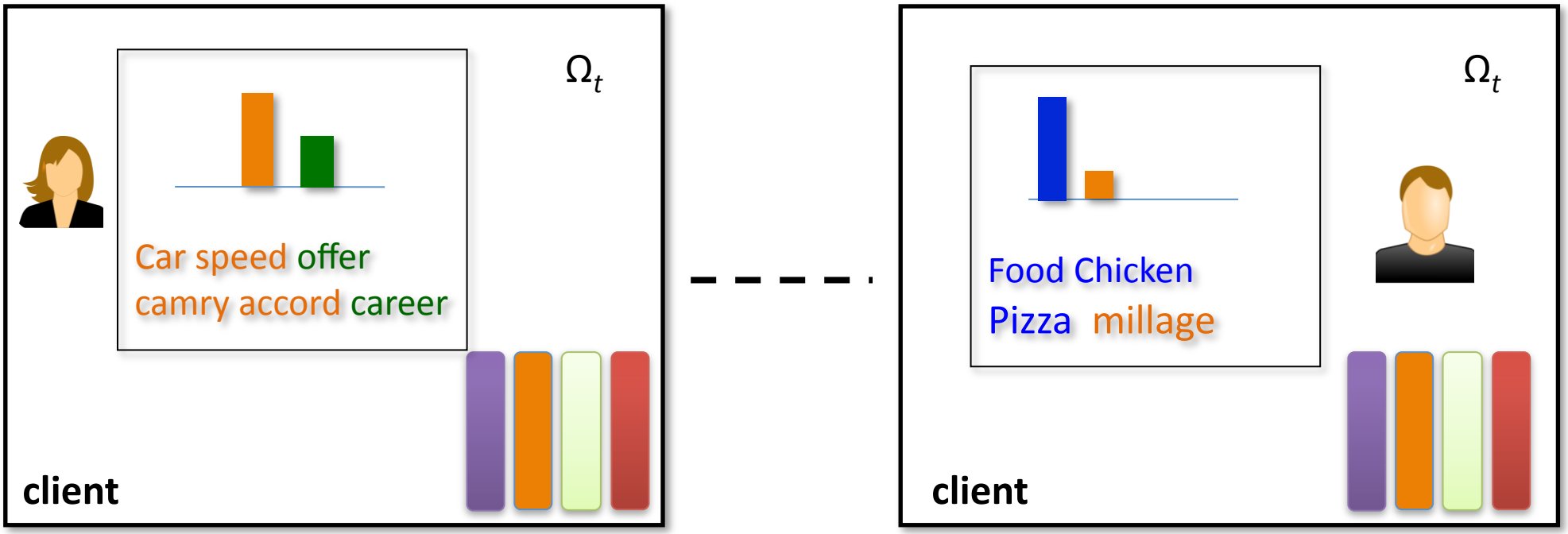
Semi-Collapsed



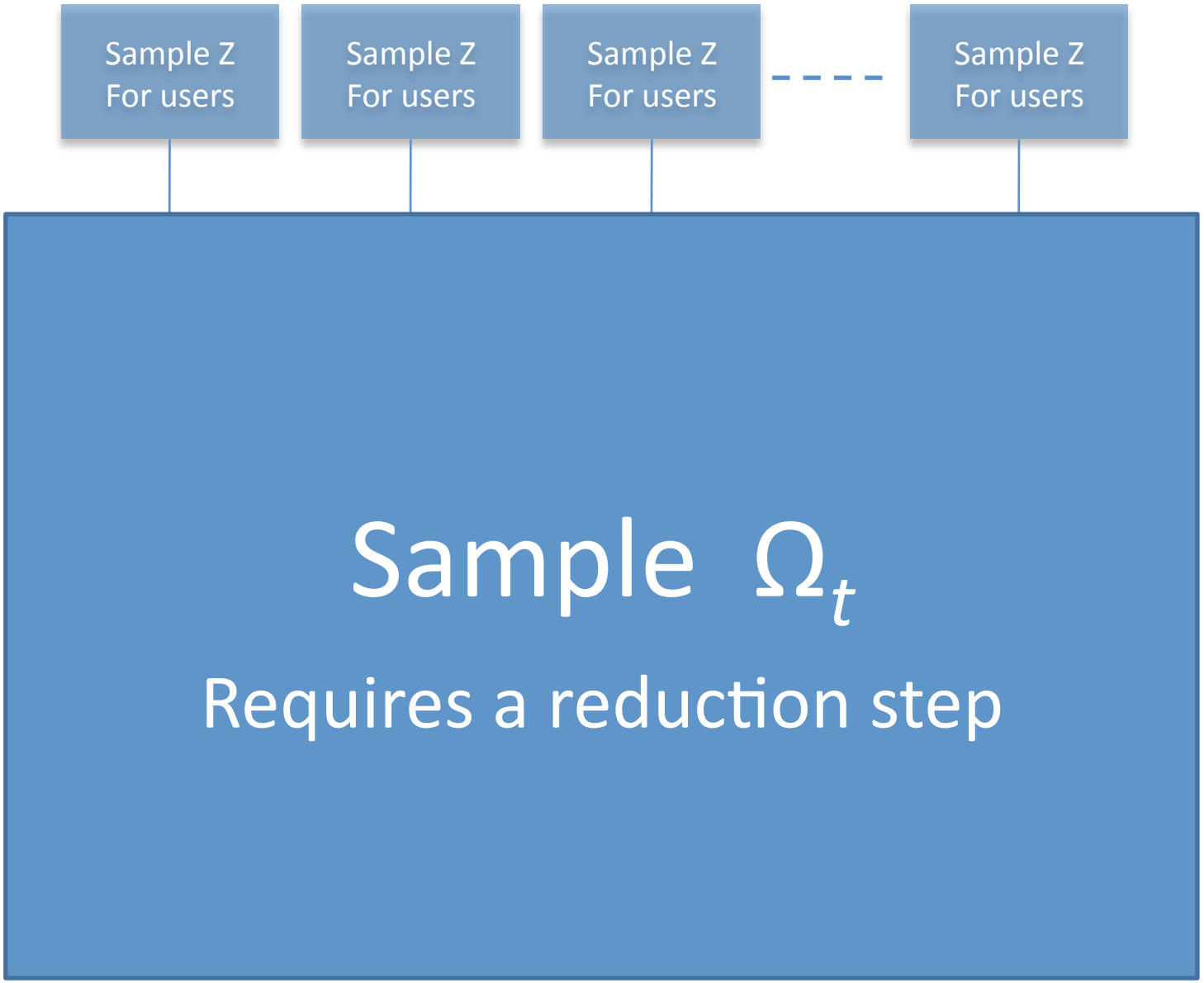
Semi-Collapsed

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

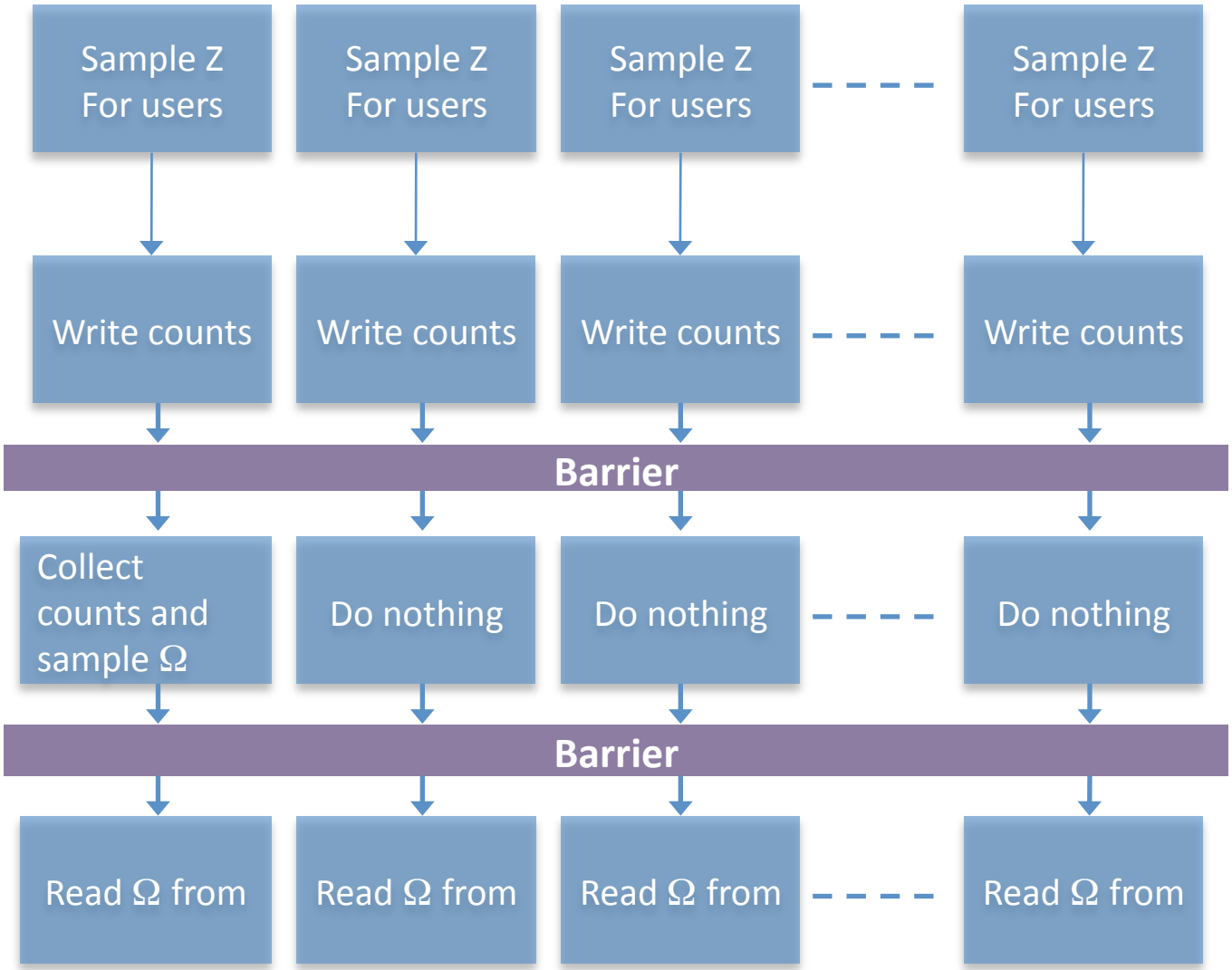
$$\propto \left(n_{ik}^{t,-j} + \tilde{n}_{ik}^t + \lambda \Omega^t \right) \frac{n_{kw}^{t,-j} + \tilde{\beta}_{kw}^t + \beta}{\sum_l n_{kl}^{t,-j} + \tilde{\beta}_{kl}^t + \beta}$$



Distributed Sampling Cycle



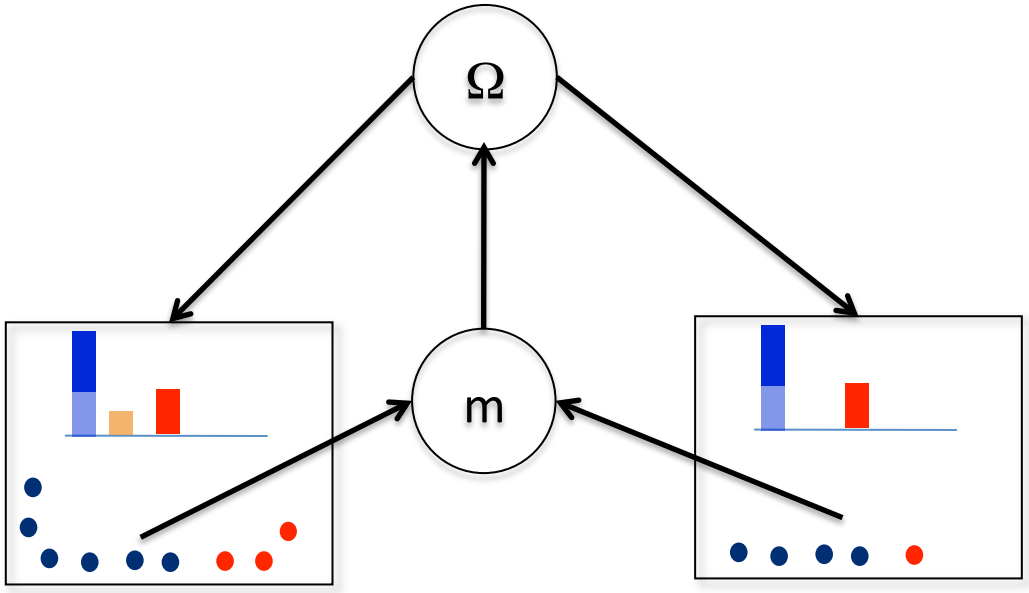
Distributed Sampling Cycle



Sampling Ω

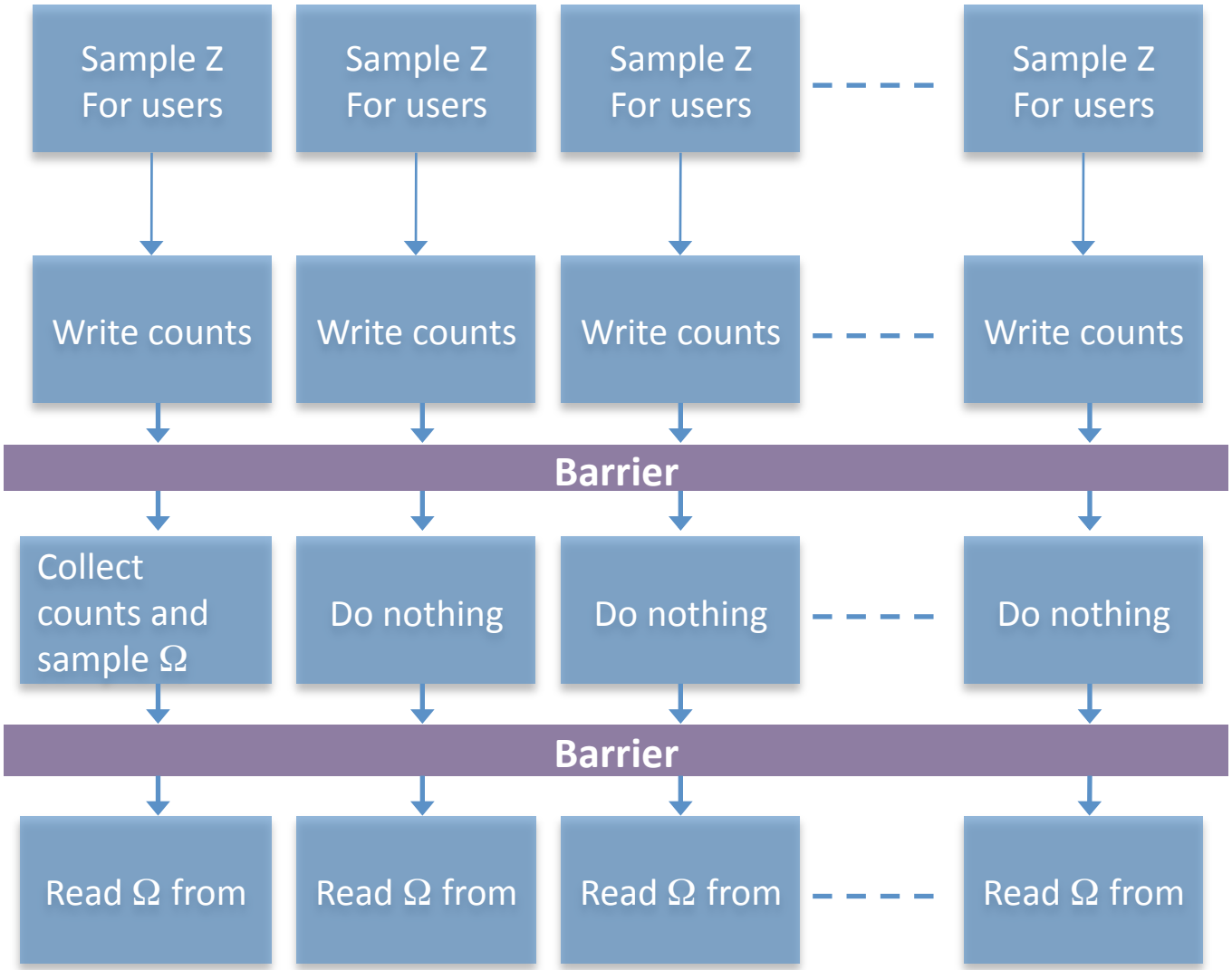
- Introduce auxiliary variable m_{kt}
  - How many times the global distribution was visited
  - $P(m_k^t | n_{1k}^t, \cdots, n_{ik}^t, \cdots)$ ~ AnotniaK

$$P(\Omega^t | \mathbf{m}^t, \tilde{\mathbf{m}}^t) \sim \mathrm{Dir}(\tilde{\mathbf{m}}^t + \mathbf{m}^t + \alpha/K)$$



$$\propto \left(n_{ik}^{t,-j} + \tilde{n}_{ik}^t + \lambda \Omega^t \right)$$

Distributed Sampling Cycle



4.2 Online Distributed Inference

Behavioral Targeting

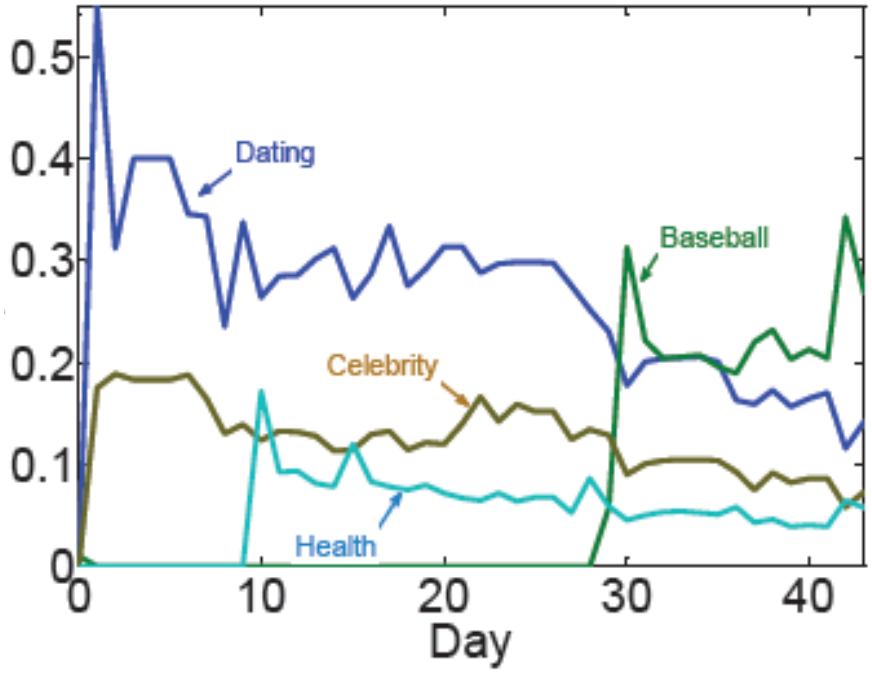
Experimental Results

- 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

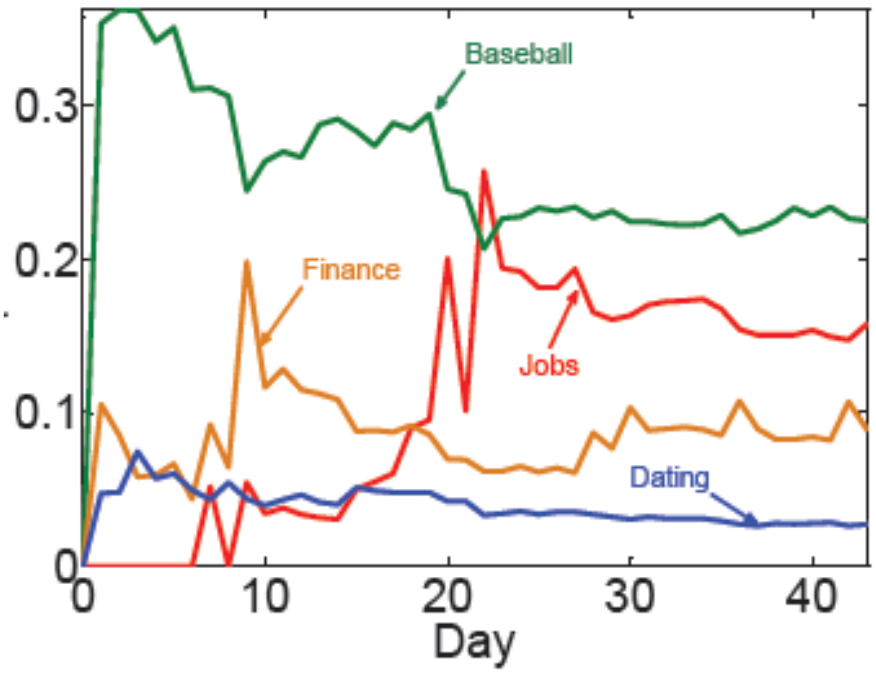
dataset	# days	# users	# campaigns	size
1	56	13.34M	241	242GB
2	44	33.5M	216	435GB

Interpretability

User-1

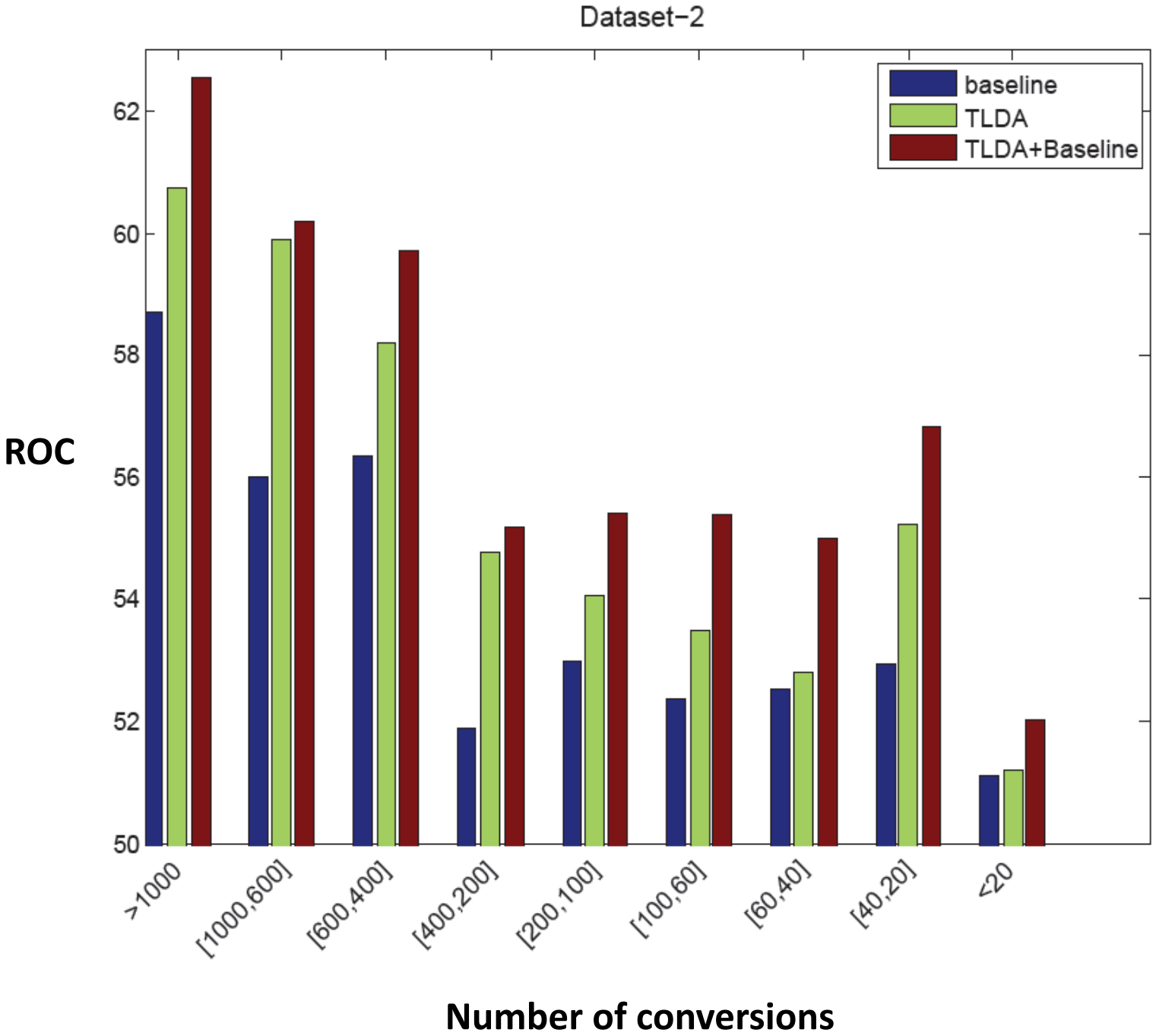


User-2



Dating	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			

Performance in Display Advertising



Performance in Display Advertising

Weighted ROC measure

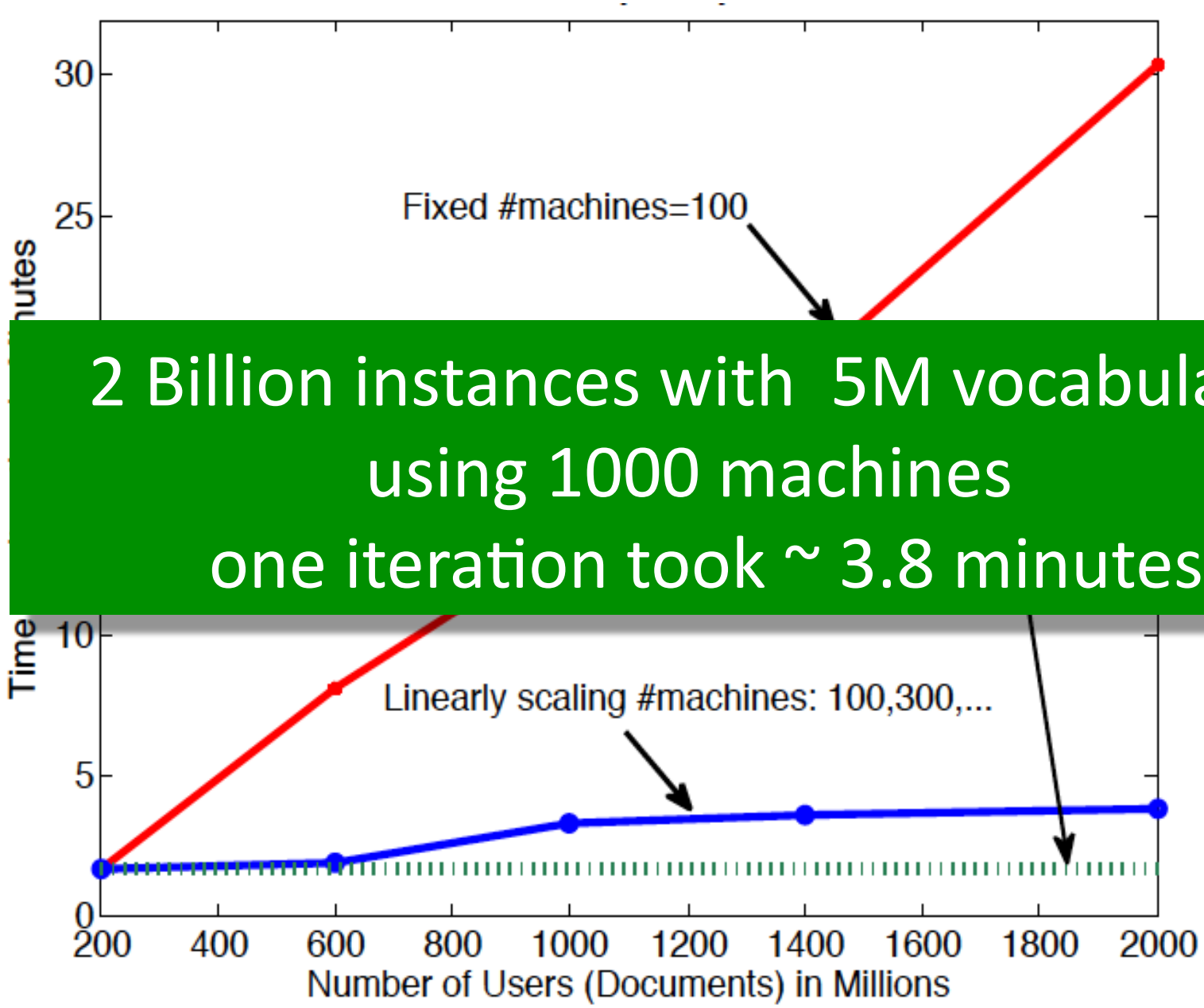
	base	TLDA	TLDA+base	LDA+base
dataset 1	54.40	55.78	**56.94**	55.80
dataset 2	57.03	57.70	**60.38**	58.54

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?

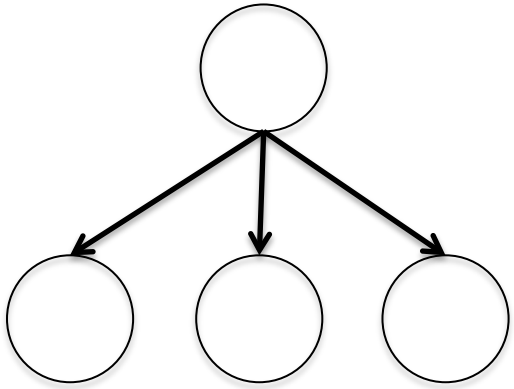


2 Billion instances with 5M vocabulary
using 1000 machines
one iteration took ~ 3.8 minutes

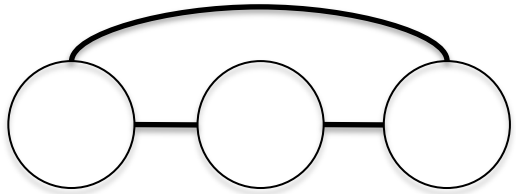
Distributed Inference Revisited

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

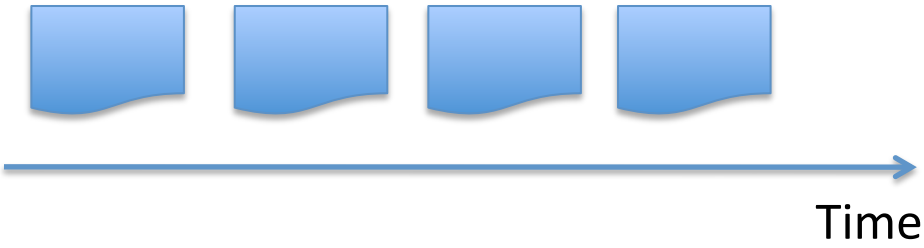
- Collapse a variable
  - Star synchronization for the sufficient statistics
- Sampling a variable
  - Local
    - Sample it locally (possibly using the synchronized statistics)
  - Shared
    - Synchronous sampling using a barrier
- Optimizing a variable
  - Same as in the shared variable case
  - Ex.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

Online Models

- Batch Large-Scale
  - Covered in part 1

- Mini-batches
  - We already have a model
  - Data arrives in batches
  - We would like to keep model up-to-data

- Time-sensitive
  - Data arrives one item at a time
  - Model should be up-to-data



What Is Coming?

- Inference
  - Online Distributed Sampling
  - Single machine multi-threaded inference
  - Online EM and Submodular Selection
- Applications
  - User tracking for behavioral Targeting
  - Content understanding
  - User modeling for content recommendation

4.2 Scalable SMC Inference

Storylines

News Stream

AP

Republicans
and lawmake
Bill Clinton e
Full Story »

Video: Gibb
Slideshow:
Related: Ta

Add-ons turn tax

BEYOND FOSSIL FUELS

Using Waste, Swedish City Shrinks Its Fossil Fuel Use

As part of its
an undergrou

China says inflation up 5.1 percent in Nov

Associated Press | Buzz up! 19 votes | Share | tweet 0

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

By CARA ANNA, Associated Press – 1 hr 50 mins ago

BEIJING – China's inflation surged to a 28-month high in November, officials said Saturday, despite government efforts to increase food supplies and end diesel shortages.

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

News Stream

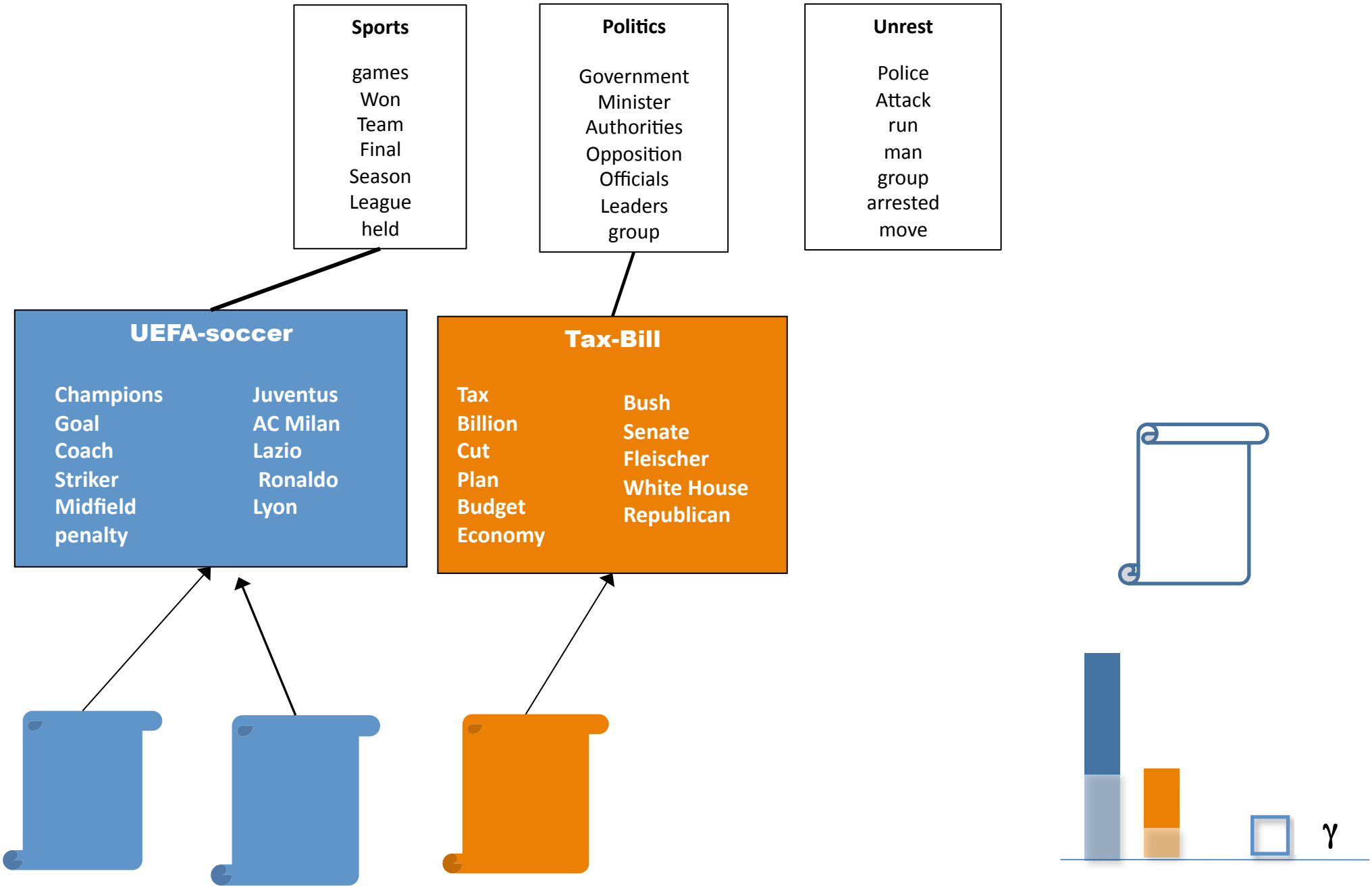
- Realtime news stream
  - Multiple sources (Reuters, AP, CNN, ...)
  - Same story from multiple sources
  - Stories are related

- Goals
  - Aggregate articles into a storyline
  - Analyze the storyline (topics, entities)
    - How does the story develop over time?
    - Who are the main entities?
    - What topics are addressed?

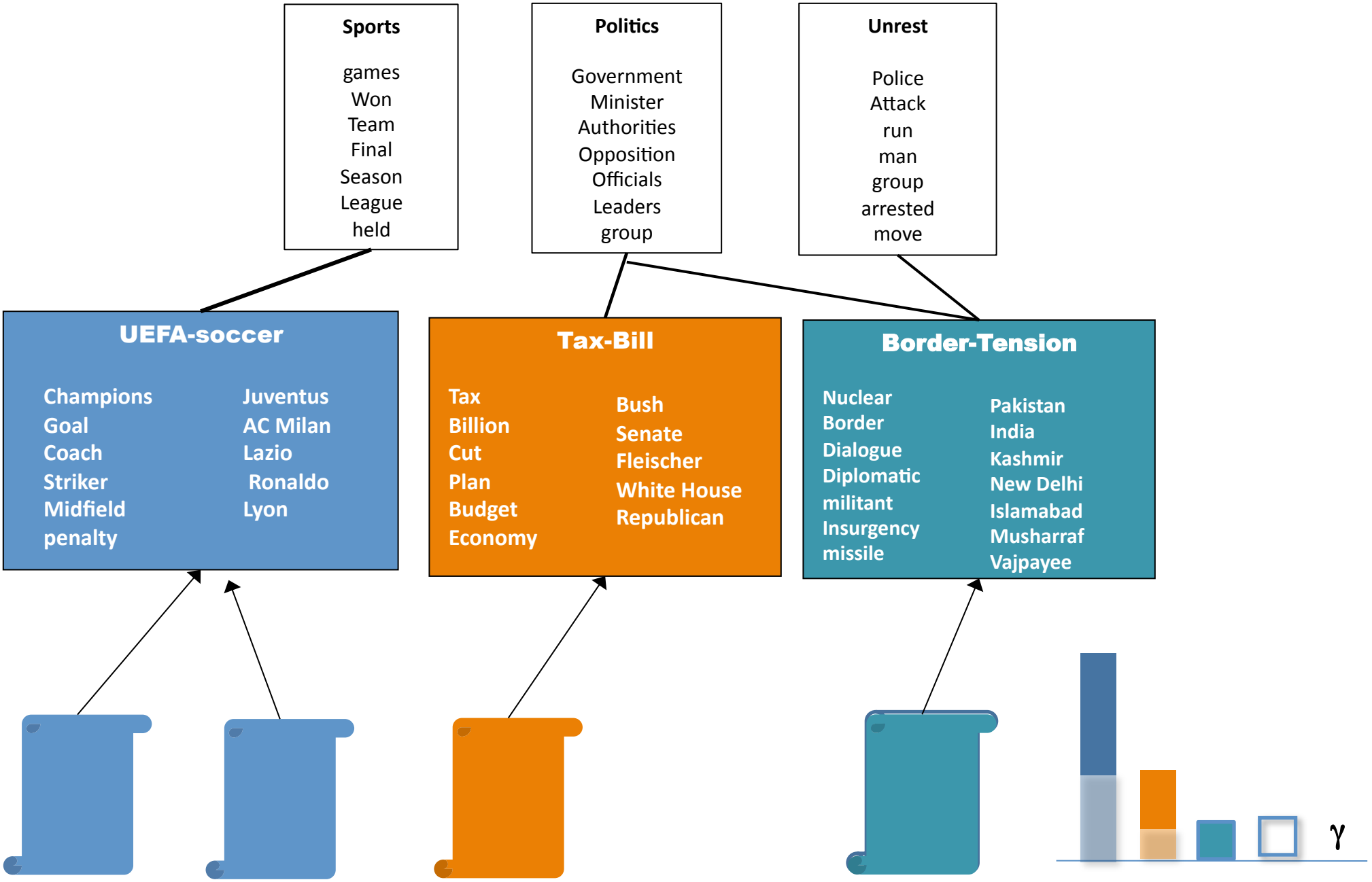
A Unified Model

- Jointly solves the three main tasks
  - Clustering,
  - Classification
  - Analysis
- Building blocks
  - A Topic model
    - High-level concepts (unsupervised classification)
  - Dynamic clustering (RCRP)
    - Discover tightly-focused concepts
      - Named entities
      - Story developments

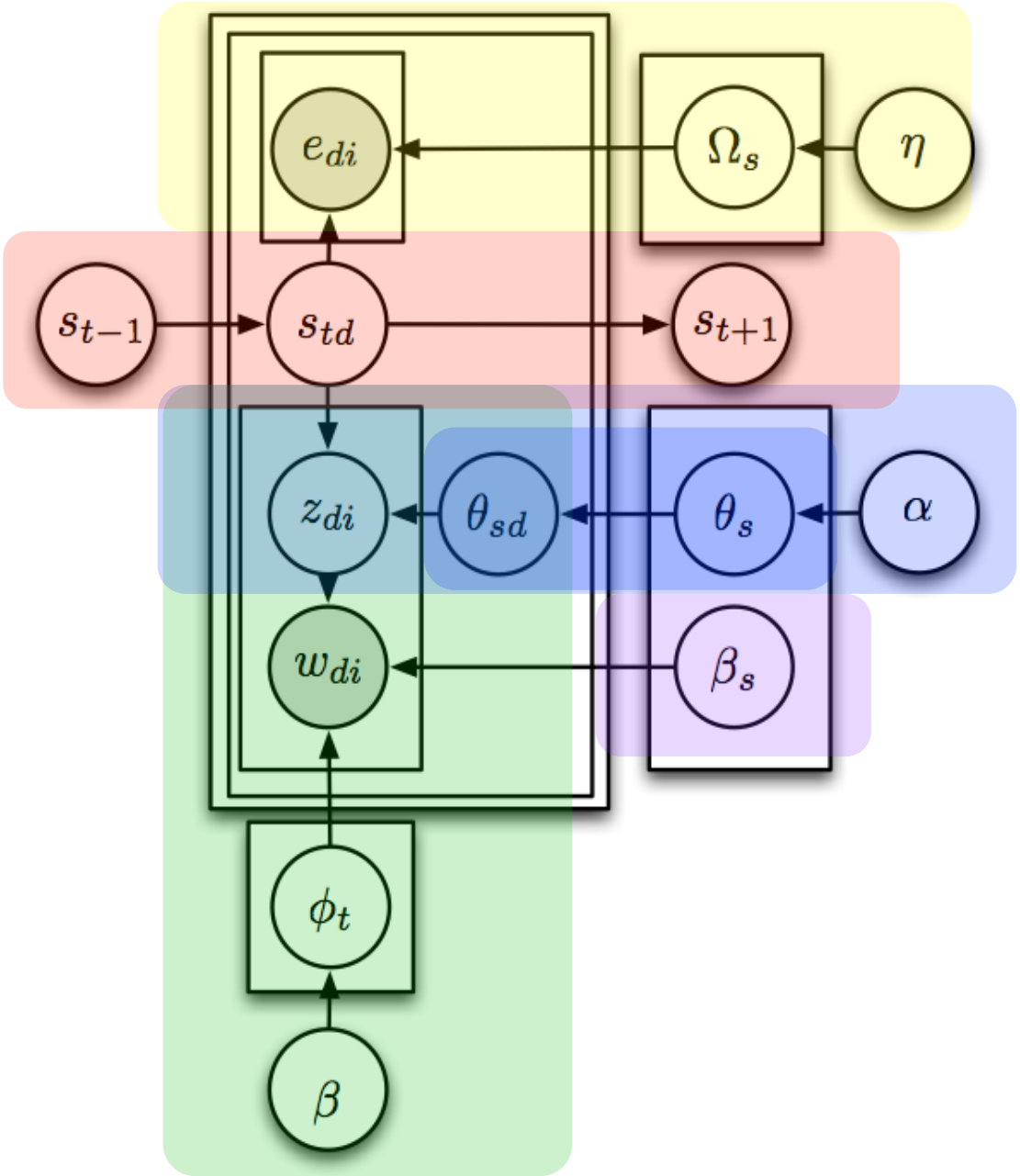
Infinite Dynamic Cluster-Topic Hybrid



Infinite Dynamic Cluster-Topic Hybrid



The Graphical Model



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

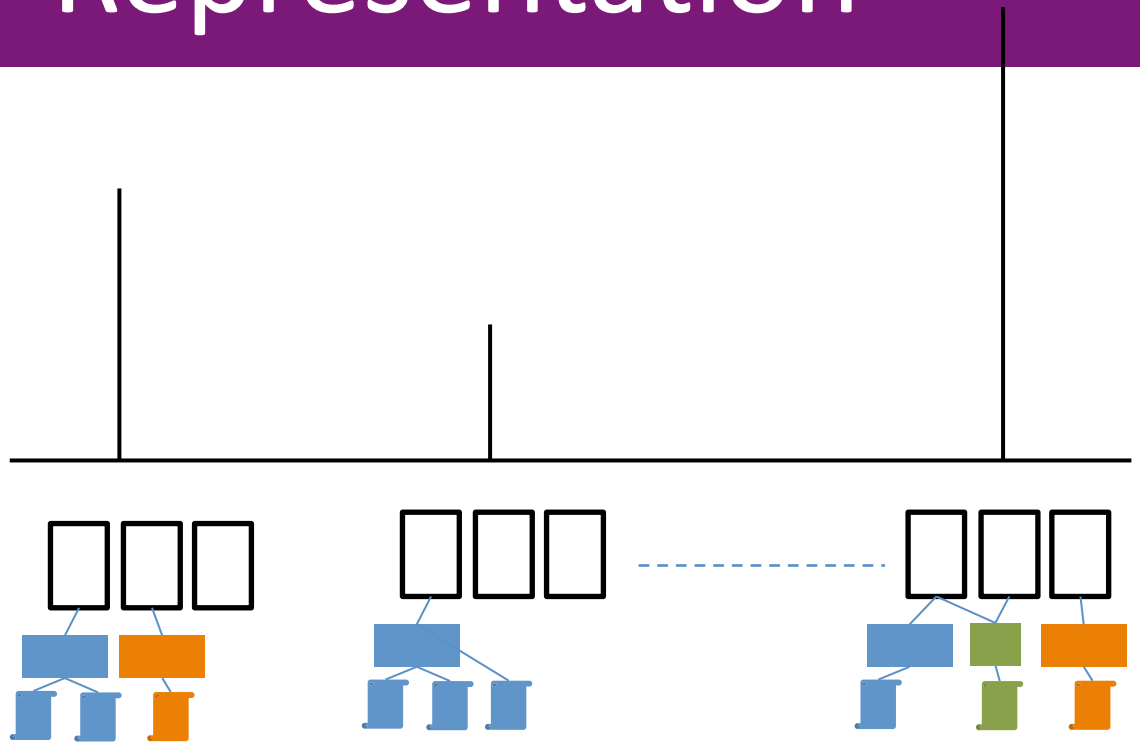
4.2 Fast SMC Inference

Inference via SMC

Online Inference Algorithm

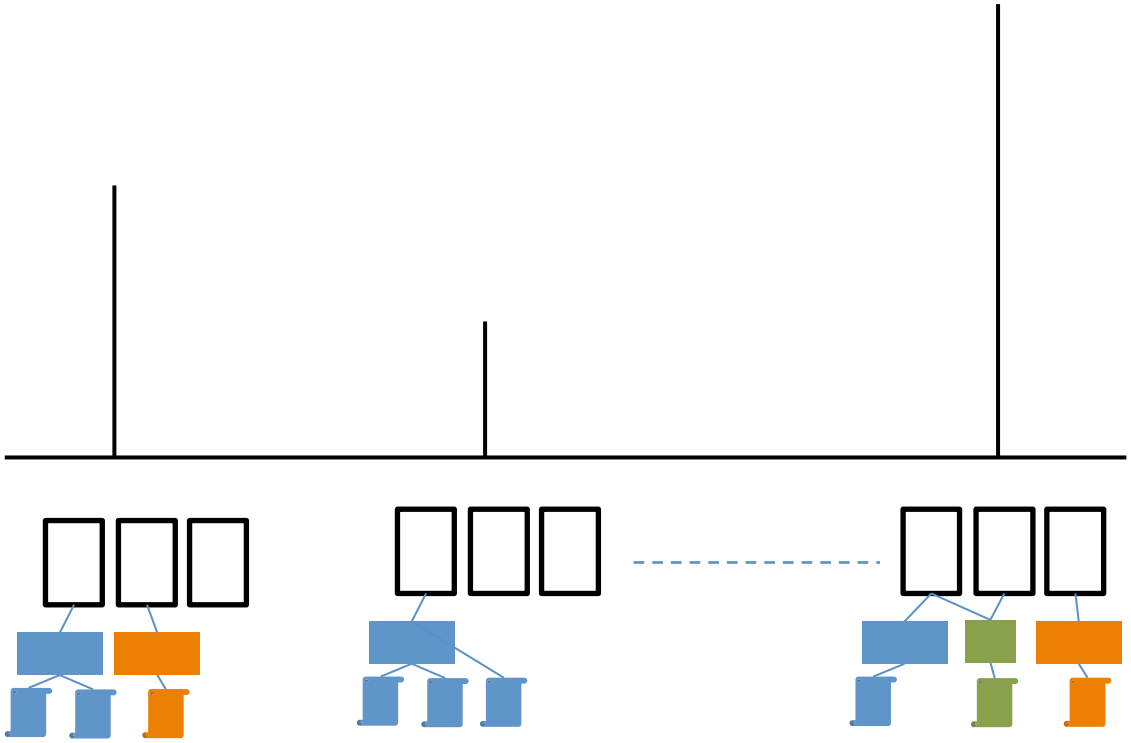
- A Particle filtering algorithm
- Each particle maintains a hypothesis
  - What are the stories
  - Document-story associations
  - Topic-word distributions
- Collapsed sampling
  - Sample ($\mathbf{z_d}$,s_d) only for each document

Particle Filter Representation



Algorithm 1 A Particle Filter Algorithm

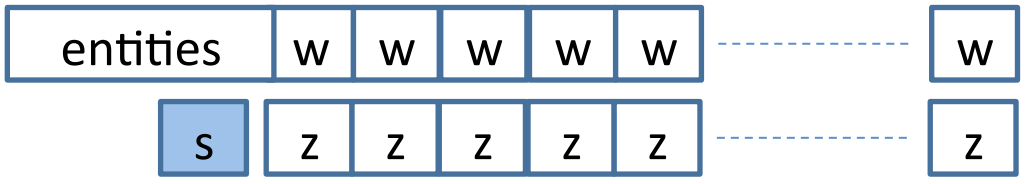
Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  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



Fold the document into the structure of each filter

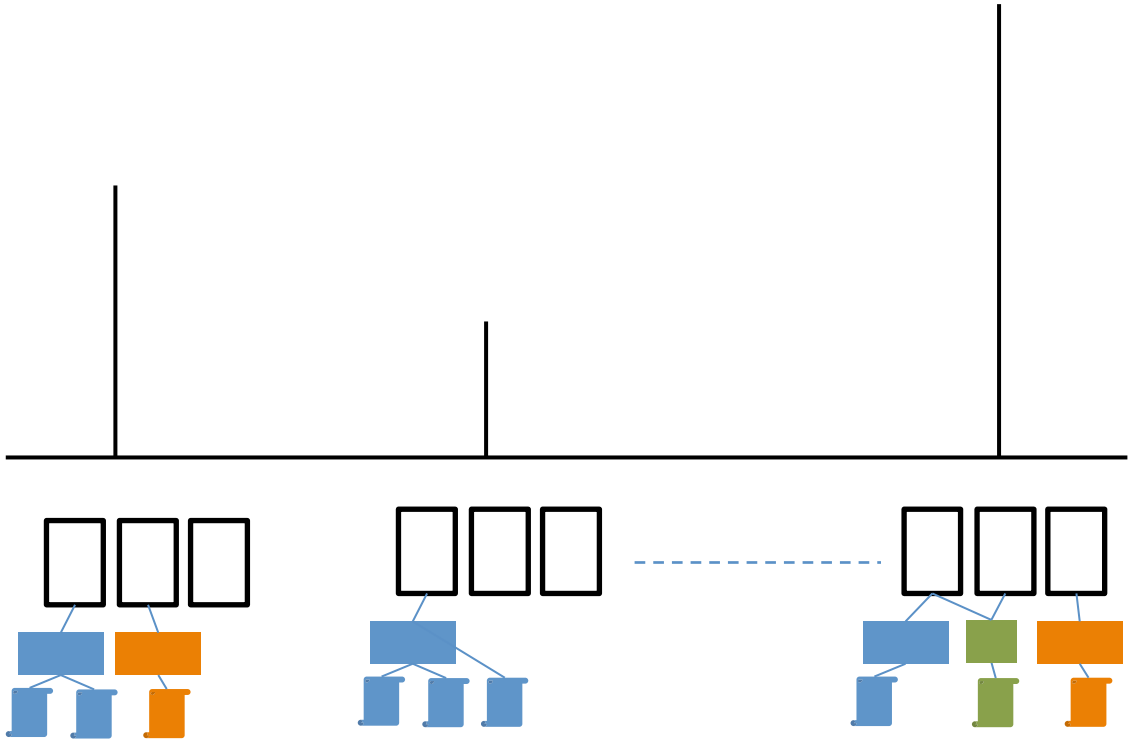
- **s** and **z** are tightly coupled
- Alternatives
  - Sample **s** then sample **z** (high variance)

Document td



Algorithm 1 A Particle Filter Algorithm

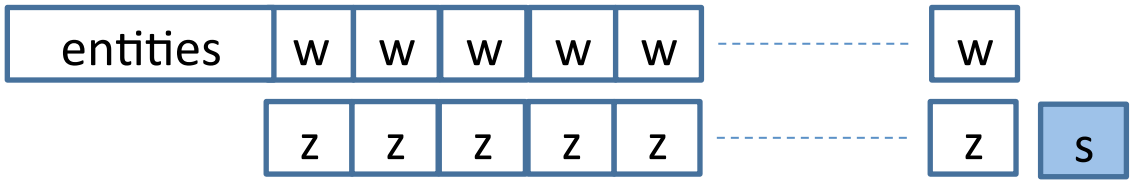
Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  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



Fold the document into the structure of each filter

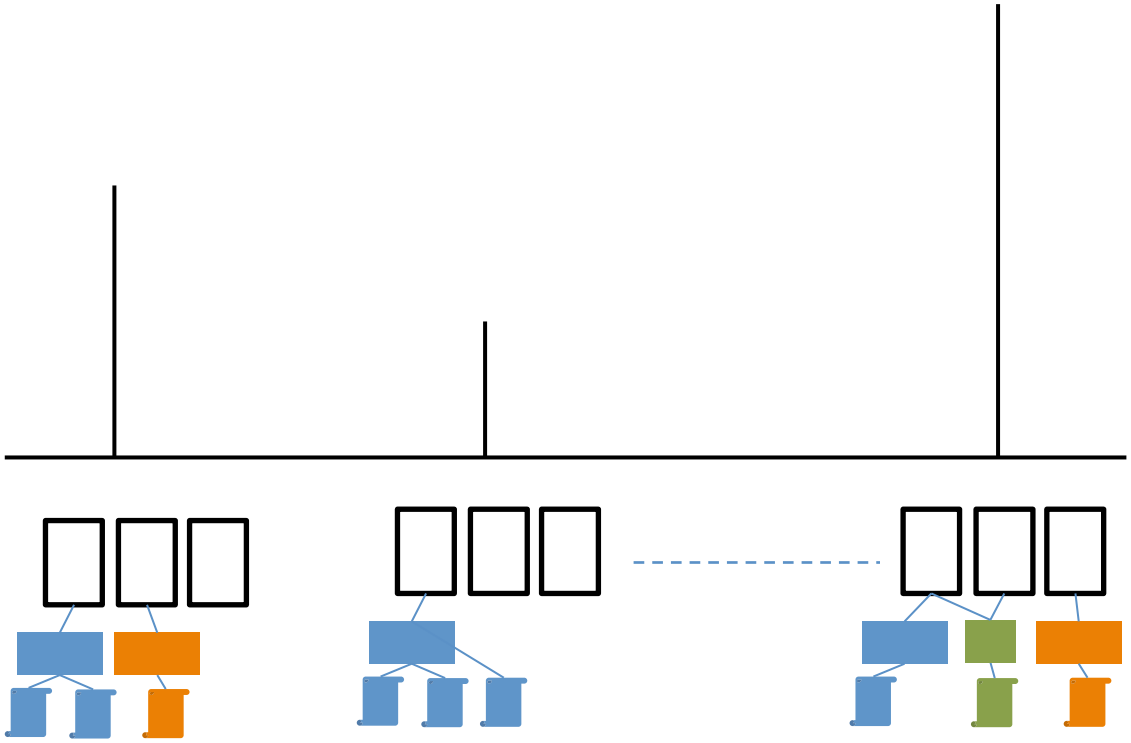
- **s** and **z** are tightly coupled
- Alternatives
  - Sample **s** then sample **z** (high variance)
  - Sample **z** then sample **s** (doesn't make sense)

Document td



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

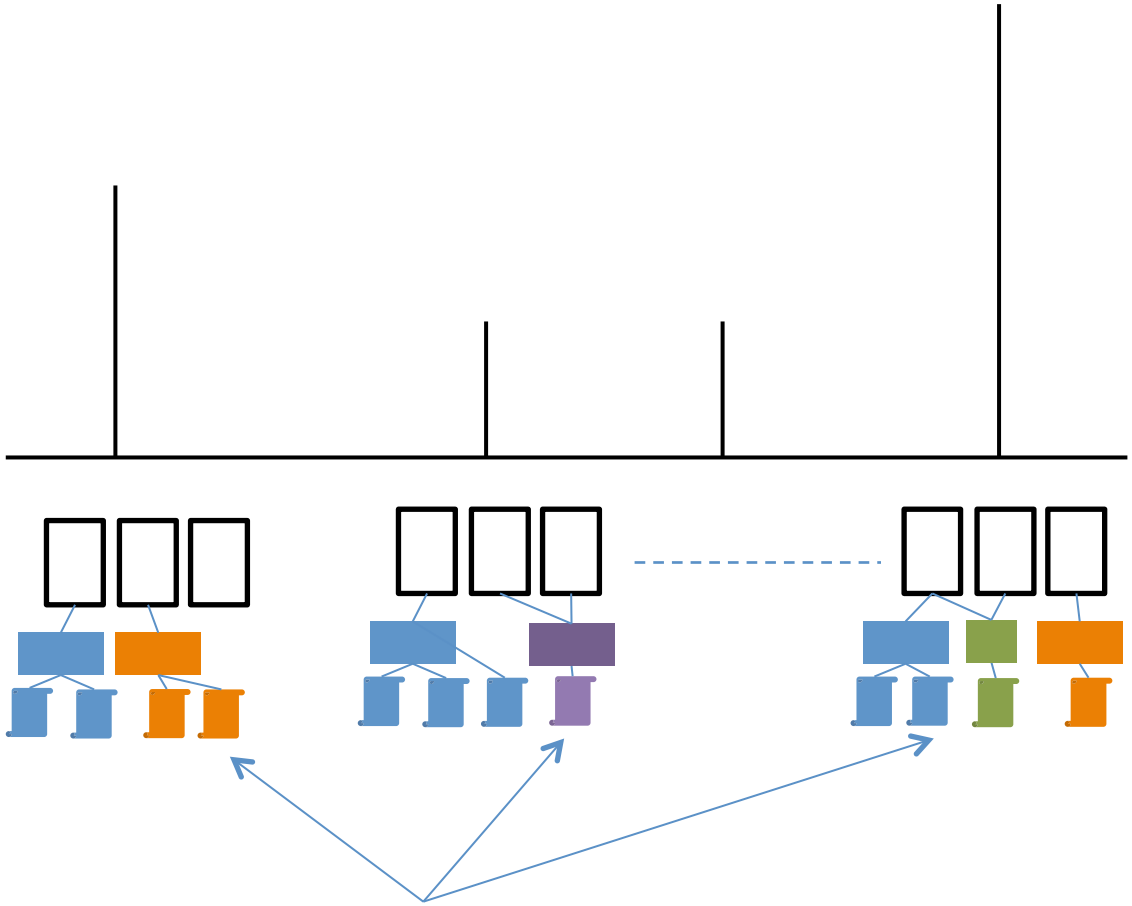


Fold the document into the structure of each filter

- **s** and **z** are tightly coupled
- Alternatives
  - Sample **s** then sample **z** (high variance)
  - Sample **z** then sample **s** (doesn't make sense)
- Idea
  - Run a few iterations of **MCMC over s and z**
  - Take last sample as the proposed value

Algorithm 1 A Particle Filter Algorithm

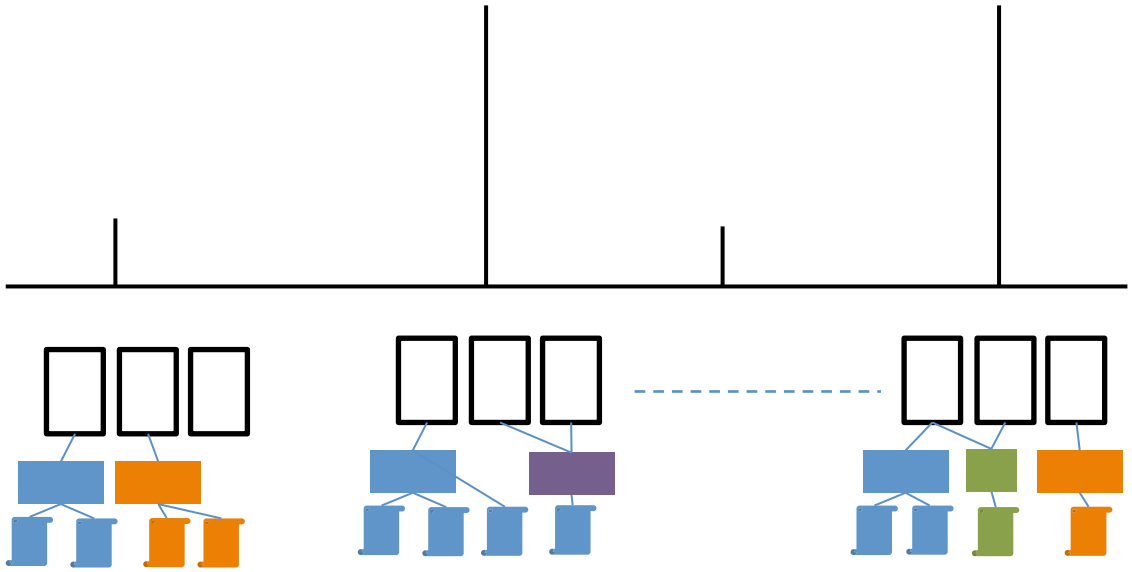
Initialize ω_1^f to $\frac{1}{F}$ for all $f \in \{1, \ldots F\}$
for each document d with time stamp t **do**
  for $f \in \{1, \ldots F\}$ **do**
    Sample $s_{td}^f, \mathbf{z}_{td}^f$ using MCMC
    $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$
  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



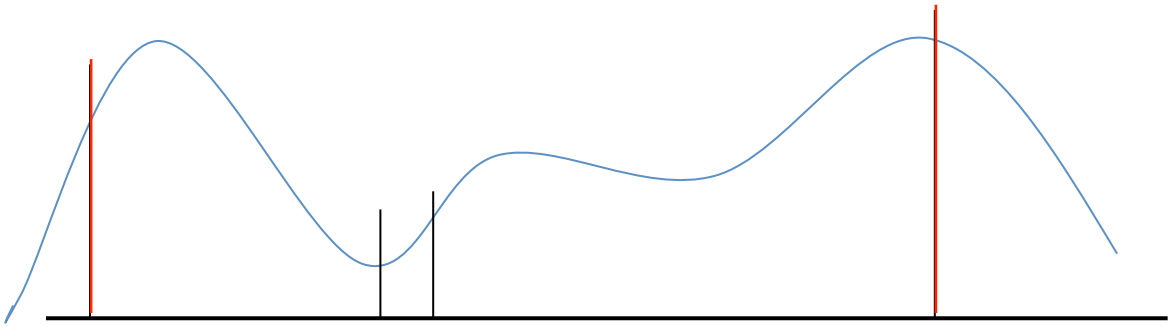
How good each filter look now?

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

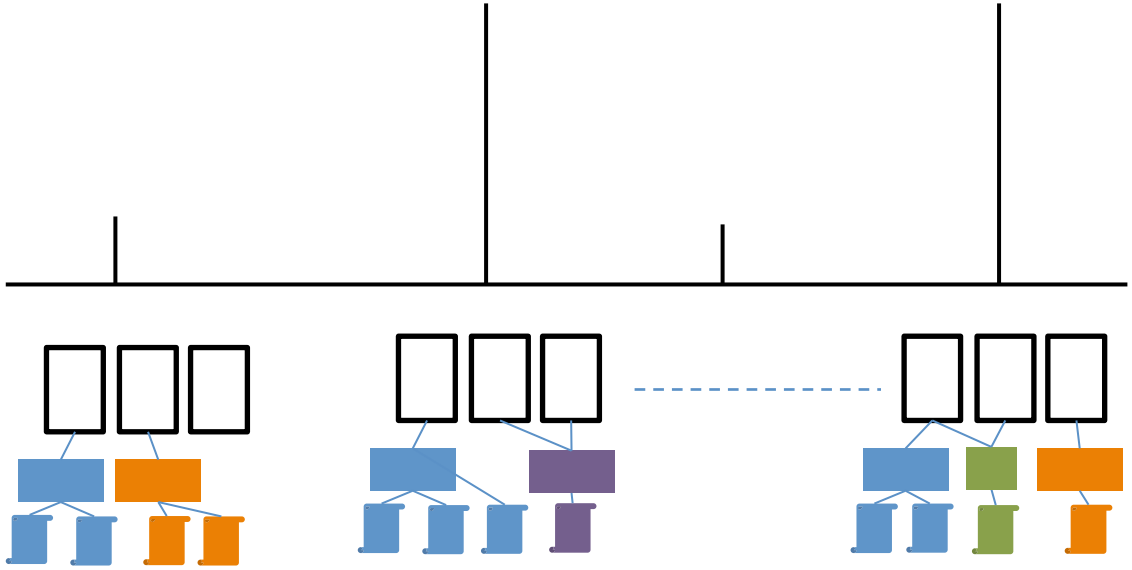


Get rid of bad filter
Replicate good one

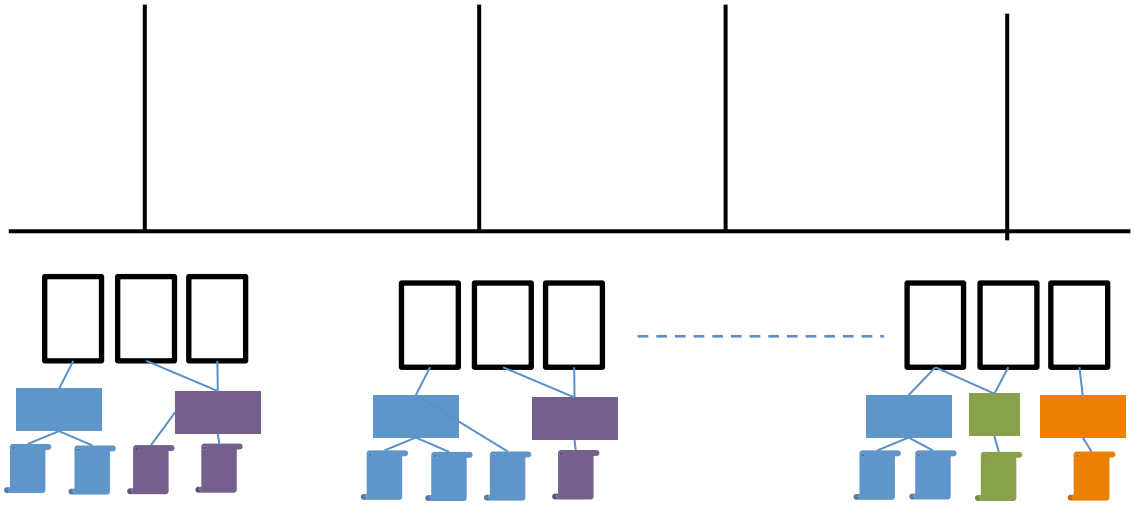


Algorithm 1 A Particle Filter Algorithm

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

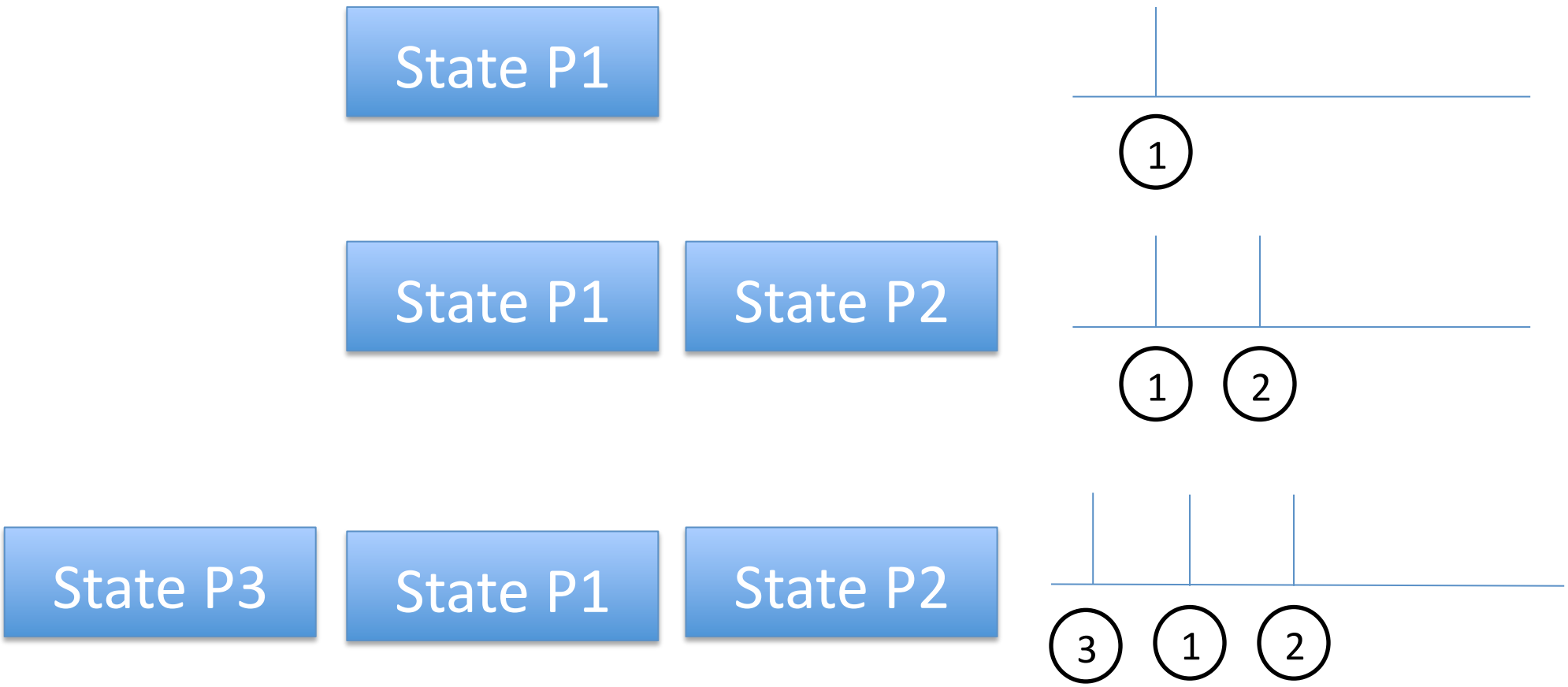


Get rid of bad filter
Replicate good one



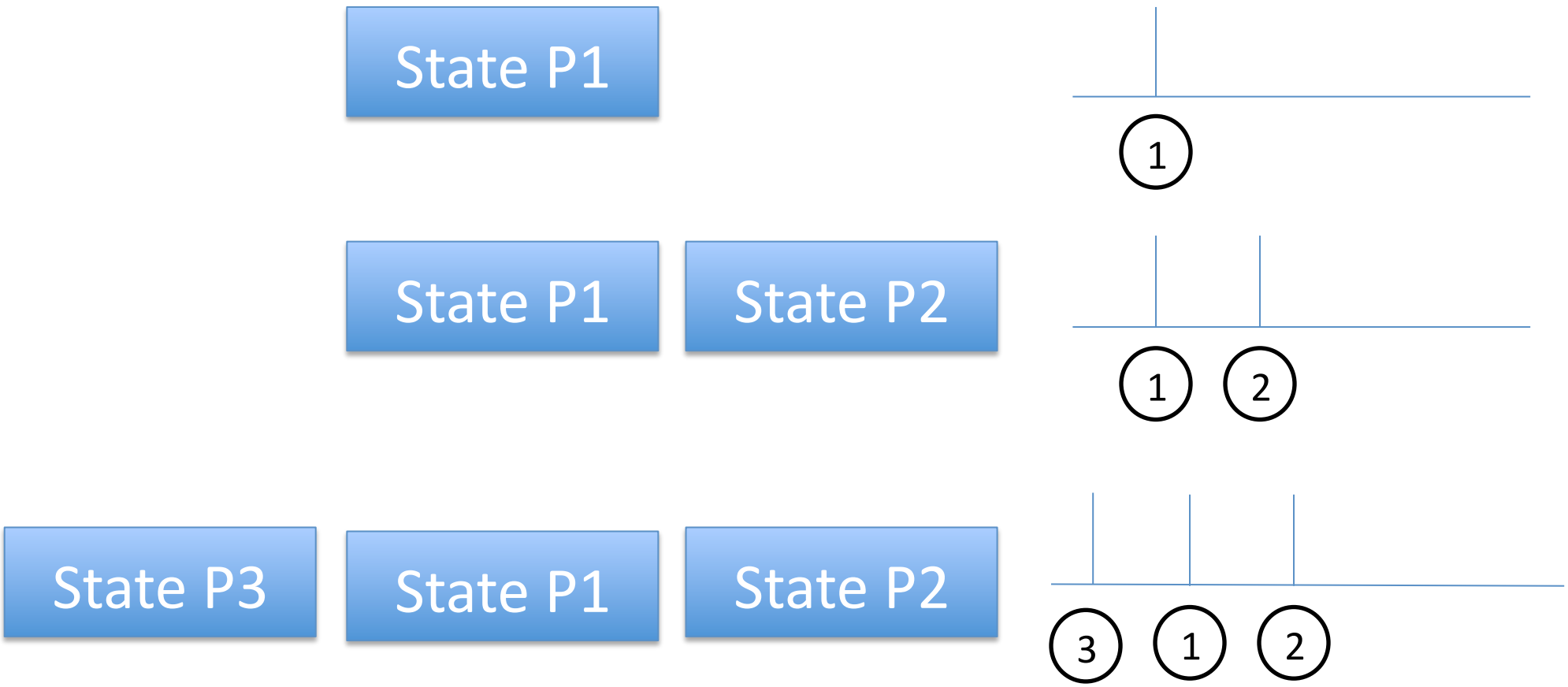
Efficient Computation and Storage

- Particles get replicated



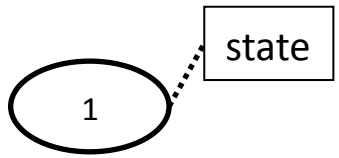
Efficient Computation and Storage

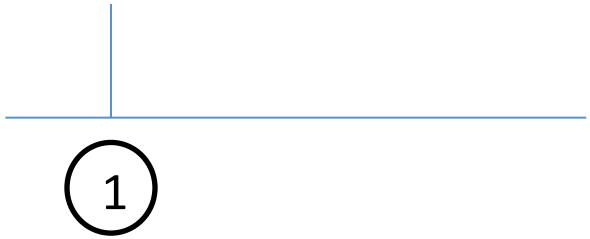
- Particles get replicated
  - Use thread-safe Inheritance tree



Efficient Computation and Storage

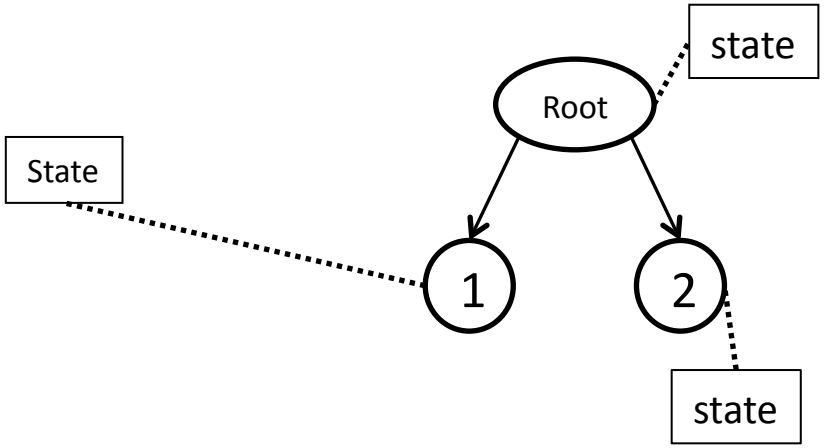
- Particles get replicated
  - Use thread-safe Inheritance tree [extends Canini et. Al 2009]

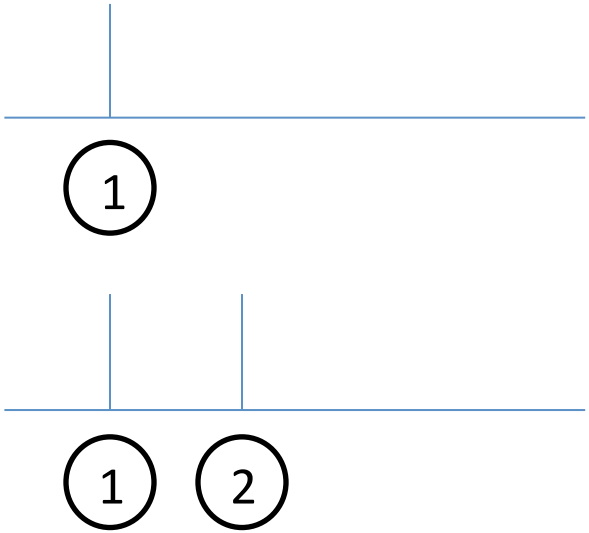




Efficient Computation and Storage

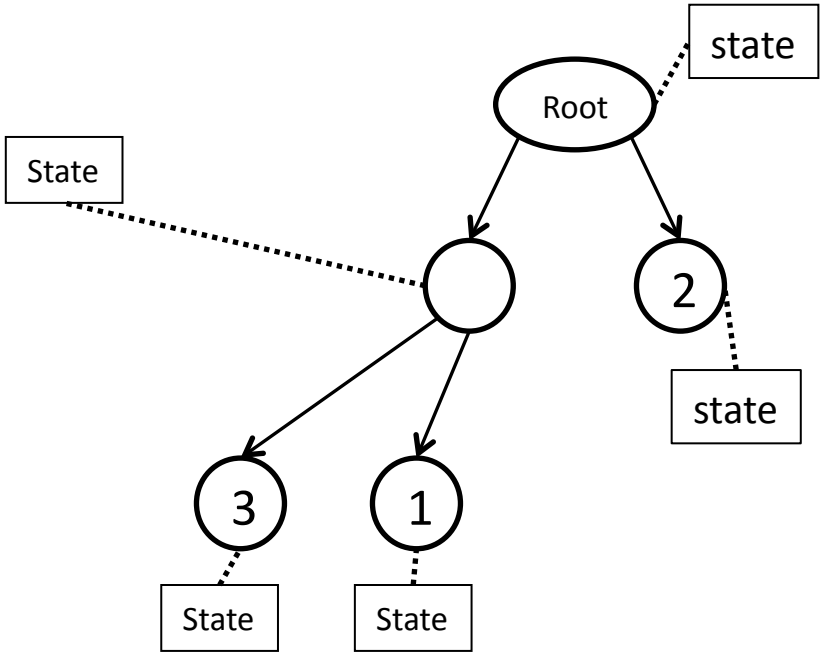
- Particles get replicated
  - Use thread-safe Inheritance tree [extends Canini et. Al 2009]

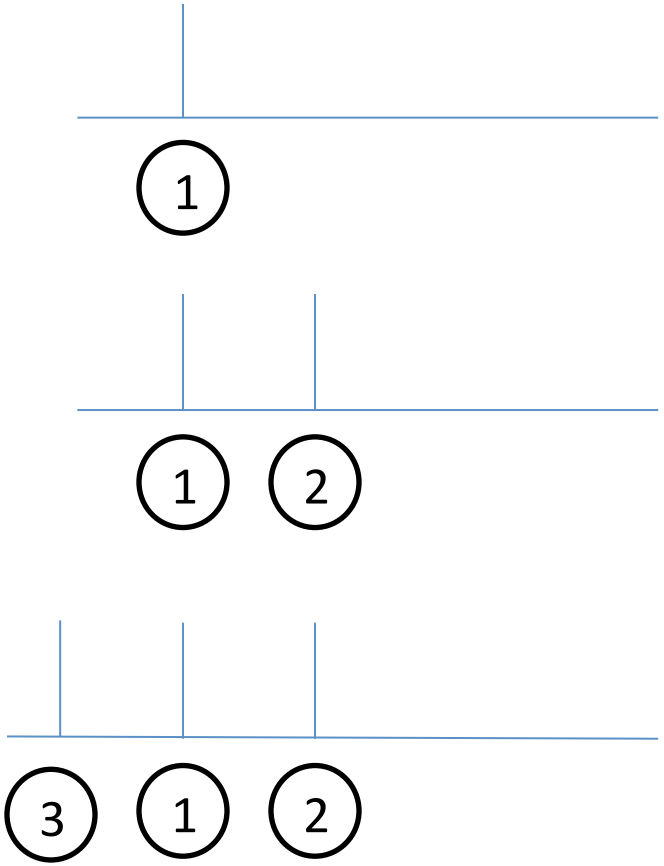




Efficient Computation and Storage

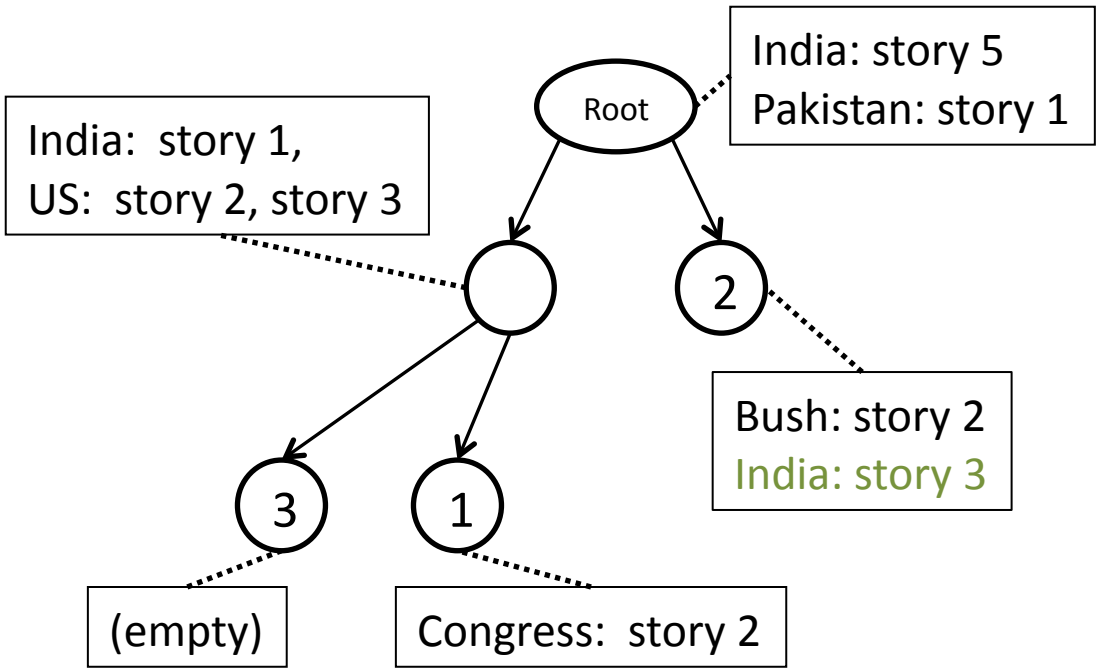
- Particles get replicated
  - Use thread-safe Inheritance tree [extends Canini et. Al 2009]

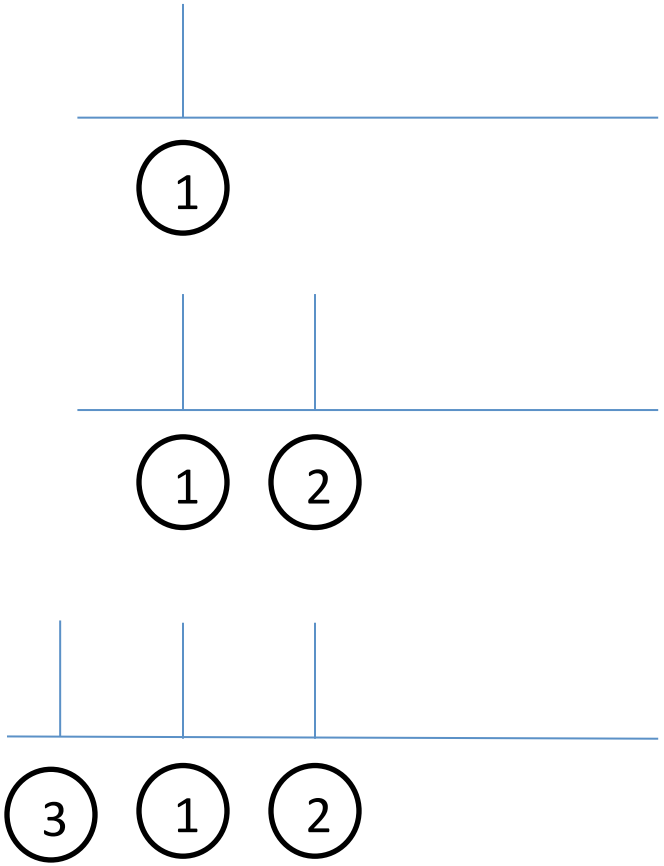




Efficient Computation and Storage

- Particles get replicated
  - Use thread-safe Inheritance tree [extends Canini et. Al 2009]
  - **Inverted representation** for fast lookup





Efficient Computation and Storage

- Why this is useful?

$$P(\mathbf{e}_{td}|s_{td} = s, \text{rest})$$

$$\frac{\Gamma\left(\sum_{e=1}^{E}\left[C_{se}^{-td} + \Omega_0\right]\right)}{\Gamma\left(\sum_{e=1}^{E}\left[C_{td,e} + C_{se}^{-td} + \Omega_0\right]\right)} \prod_{e=1}^{E} \frac{\Gamma\left(C_{td,e} + C_{se}^{-td} + \Omega_0\right)}{\Gamma\left(C_{se}^{-td} + \Omega_0\right)}$$

- Only focus on stories that mention at least one entity
  - Otherwise pre-compute and reuse
- We can use fast samplers for z as well [Yao et. Al. KDD09]

Experiments

- Yahoo! News datasets over two months
  - Three sub-sampled sets with different characteristics
- Editorially-labeled documents
  - Cannot-like and must-link pairs
- Performance measures using clustering accuracy
- Baseline
  - A strong *offline* Correlation *clustering* algorithm [WSDM 11]
    - Scaled with LSH to compute neighborhood graph (similar to Petrovic 2010)

Structured Browsing

Sports

games
Won
Team
Final
Season
League
held

Politics

Government
Minister
Authorities
Opposition
Officials
Leaders
group

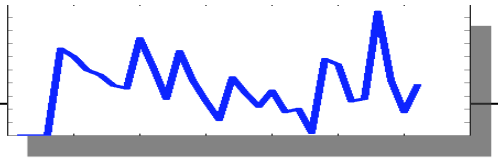
Unrest

Police
Attach
run
man
group
arrested
move

UEFA-soccer

Champions
Goal
Leg
Coach
Striker
Midfield
penalty

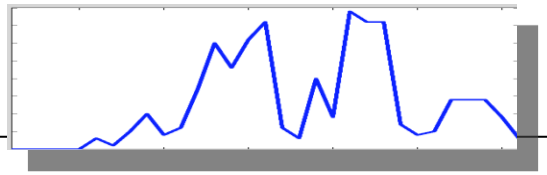
Juventus
AC Milan
Real Madrid
Milan
Lazio
Ronaldo
Lyon



Tax-bills

Tax
Billion
Cut
Plan
Budget
Economy
lawmakers

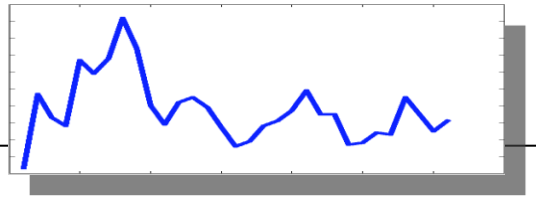
Bush
Senate
US
Congress
Fleischer
White House
Republican



Border-Tension

Nuclear
Border
Dialogue
Diplomatic
militant
Insurgency
missile

Pakistan
India
Kashmir
New Delhi
Islamabad
Musharraf
Vajpayee

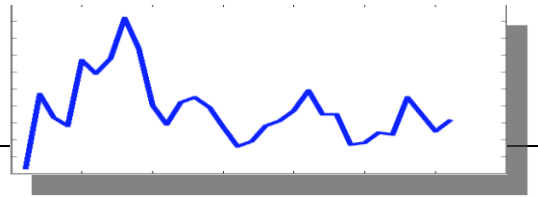


Structured Browsing

Border-Tension

Nuclear
Border
Dialogue
Diplomatic
militant
Insurgency
missile

Pakistan
India
Kashmir
New Delhi
Islamabad
Musharraf
Vajpayee



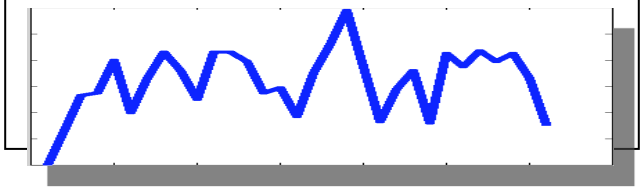
More Like India-Pakistan story

Based on topics

Middle-east-conflict

Peace
Roadmap
Suicide
Violence
Settlements
bombing

Israel
Palestinian
West bank
Sharon
Hamas
Arafat

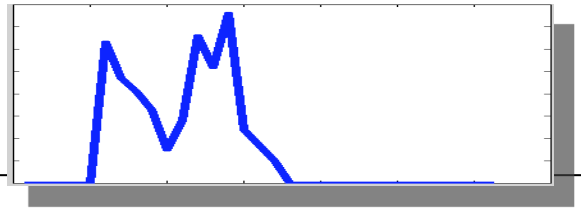


Nuclear+ topics [politics]

Nuclear programs

Nuclear
summit
warning
policy
missile
program

North Korea
South Korea
U.S
Bush
Pyongyang



Structured Browsing

Sports

games
Won
Team
Final
Season

Politics

Government
Minister
Authorities
Opposition
Officials

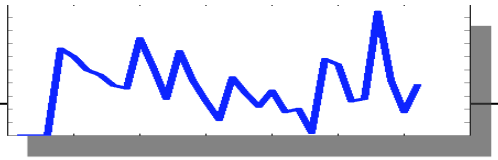
Unrest

Police
Attach
run
man
group

More on Personalization later on the talk

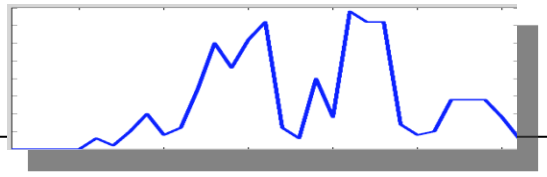
Champions
Goal
Leg
Coach
Striker
Midfield
penalty

Juventus
AC Milan
Real Madrid
Milan
Lazio
Ronaldo
Lyon



Tax
Billion
Cut
Plan
Budget
Economy
lawmakers

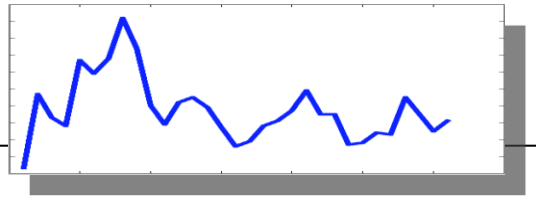
Bush
Senate
US
Congress
Fleischer
White House
Republican



on

Nuclear
Border
Dialogue
Diplomatic
militant
Insurgency
missile

Pakistan
India
Kashmir
New Delhi
Islamabad
Musharraf
Vajpayee



Quantitative Evaluation

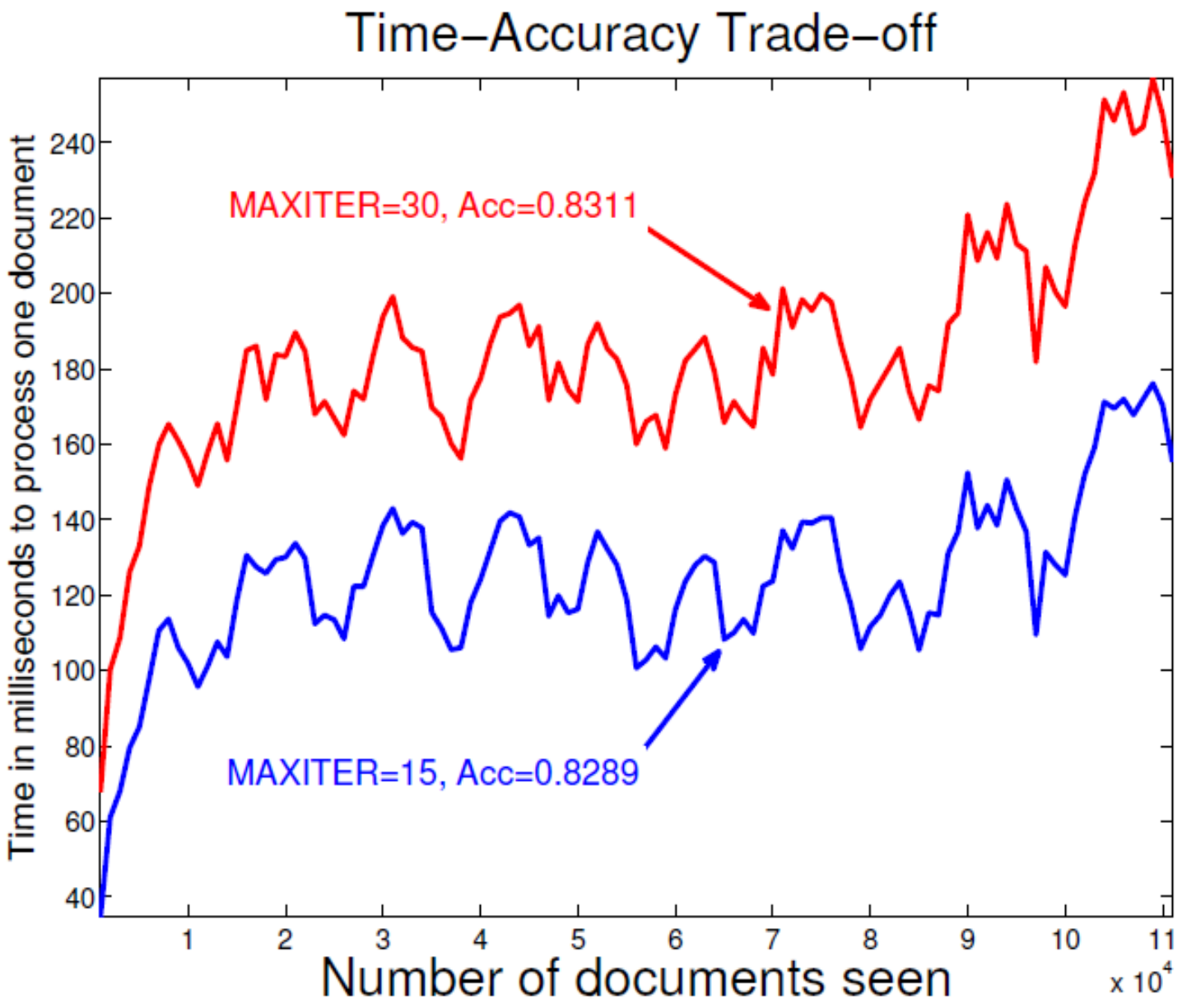
Number of topics = 100

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

Effect of number of topics

sample-No.	K=50	K=100	K=200	K=300
1	0.8261	**0.8289**	0.8186	0.8122
2	0.8293	**0.8388**	0.8344	0.8301
3	**0.8401**	0.8395	0.8373	0.8275

Scalability



Model Contribution

Removed Feature	Time	Names entites	Story words	Topics *(equiv. RCRP)*
Accuracy	0.8225	.6937	0.8114	0.7321

- Named entities are very important
- Removing time increase processing up to 2 seconds per document

Putting Things Together

Time vs. Machines

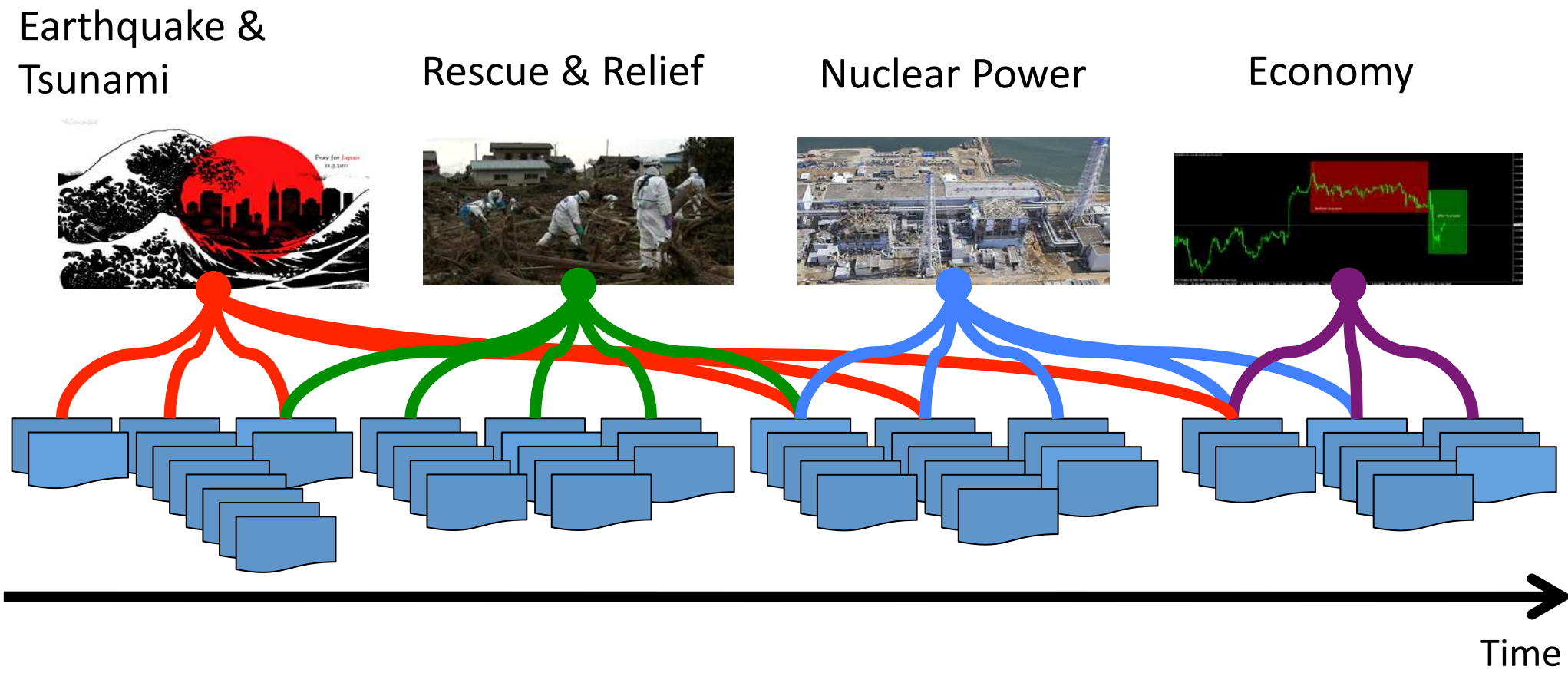
- Data arrives dynamically
- How to keep models up to date?

	Batch	Mini-batches	Truly online
Single-Machine	Gibbs Variational	Online-LDA	SMC
Multi-Machine	Star-Synch.	Star-Synch + Synchronous step	?

4.3 User Preference

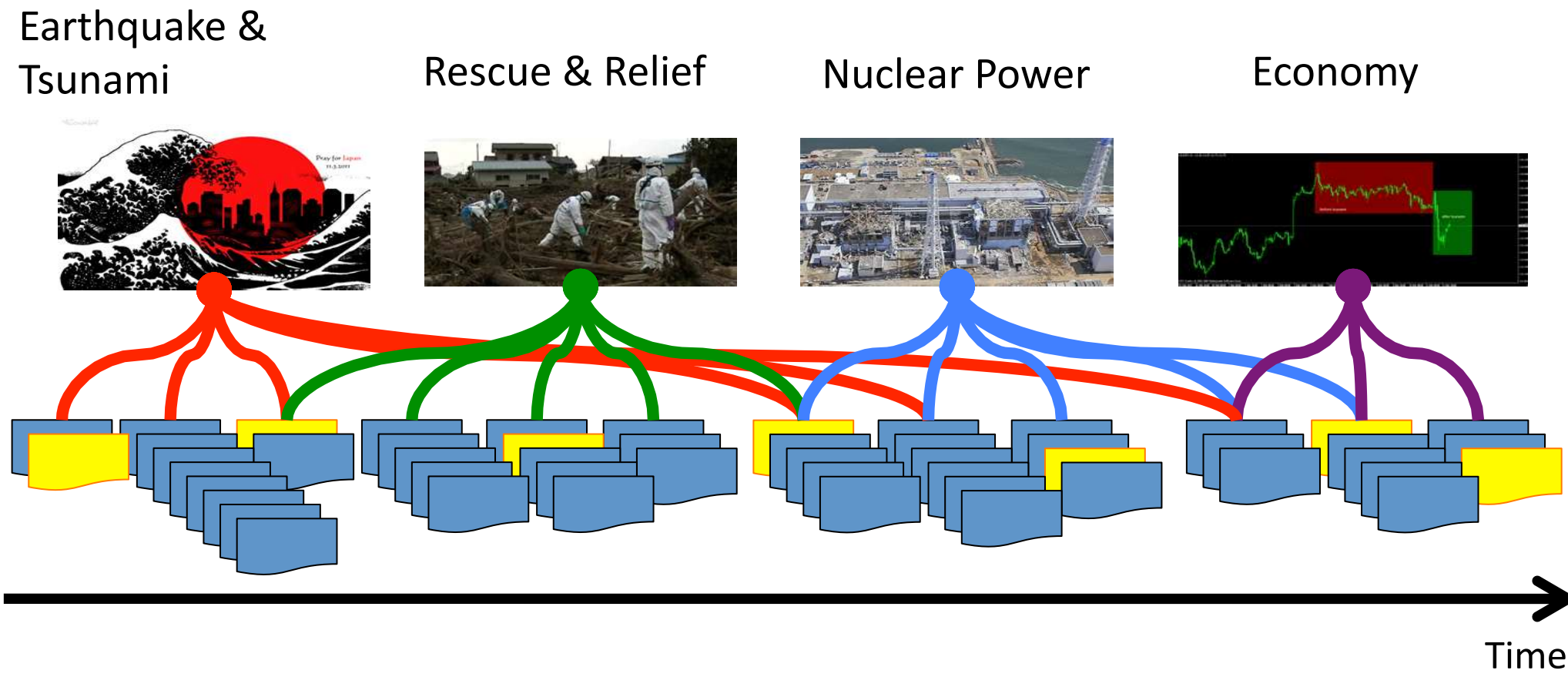
Online EM and Submodularity

Storyline Summarization



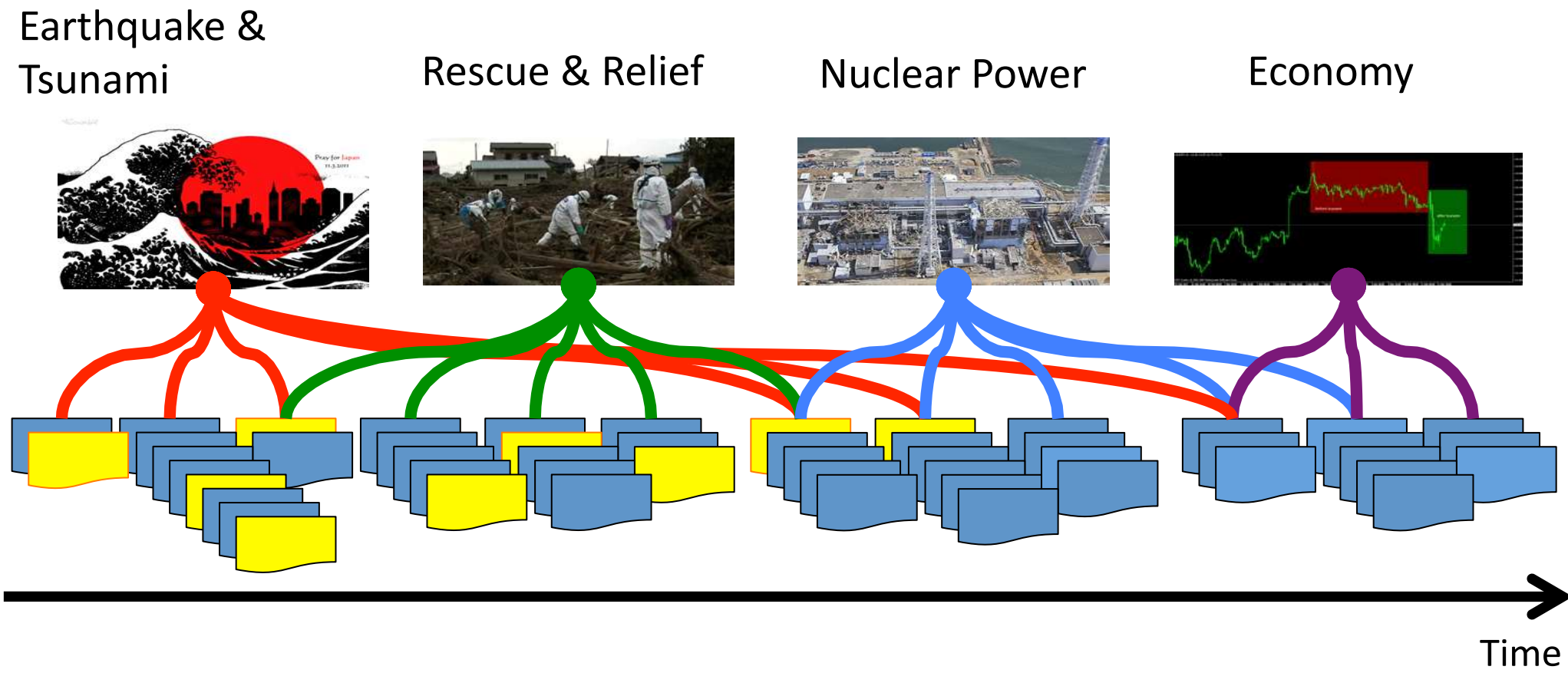
- How to summarize a storyline with few articles?

Storyline Summarization



- How to summarize a storyline with few articles?
- How to personalize the summary?

Storyline Summarization



- How to summarize a storyline with few articles?
- How to personalize the summary?

User Interaction

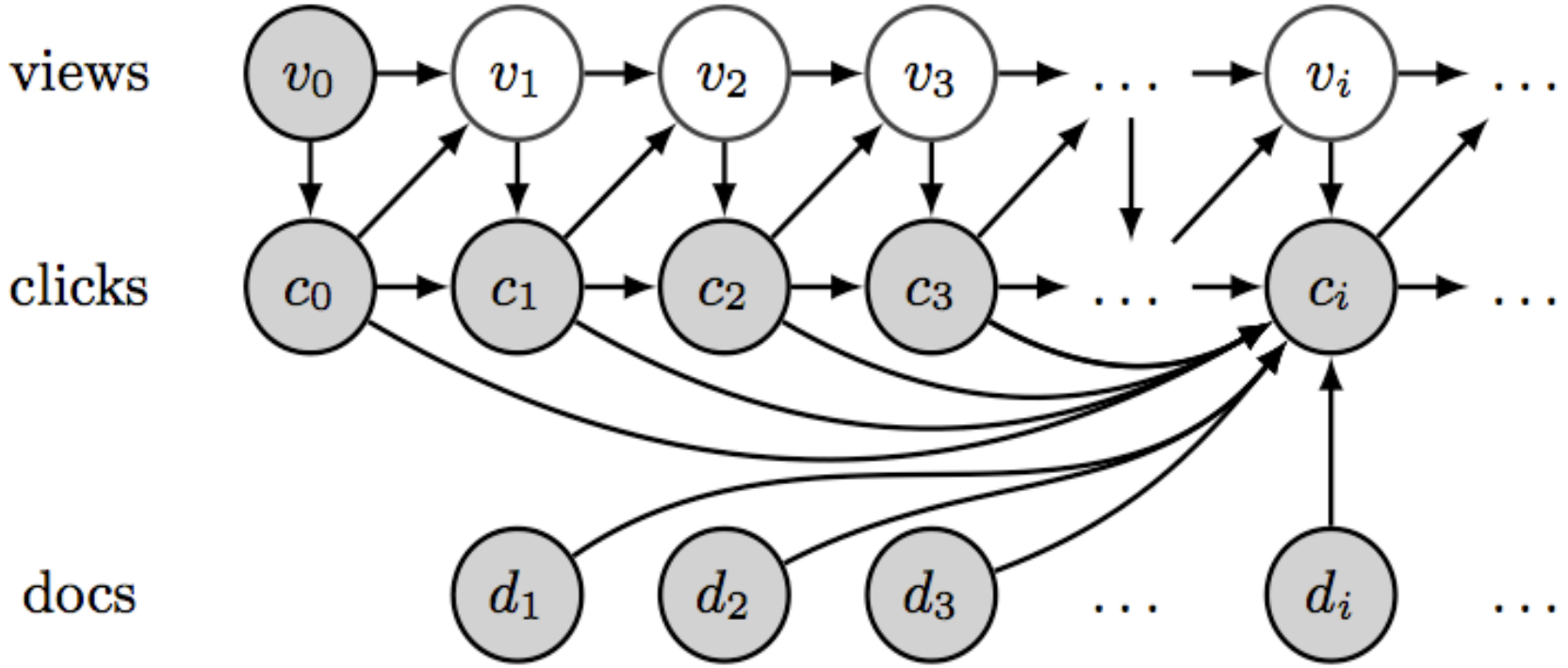
- Passive
  - We observe the user generated contents
  - Model user based on those content using unsupervised techniques
- Explicit
  - We present users with content
  - User give explicit feedback
    - Like/dislike
  - Learn user preference using supervised techniques
- Implicit
  - Mixture between the two
  - Present the user with items
  - Observe which items the user interact with
  - Learning user preference using semi-supervised models

User Satisfaction

- Modular
  - Present users with items she prefers
    - Regardless of the context
  - Targets relevance
  - Ex: vector space models

- Submodular
  - More of the same thing is not always better
    - Dimensioning return
  - Targets diversity
  - Ex: TDN [ElArini et. Al. KDD 09]

Sequential Click-View Model

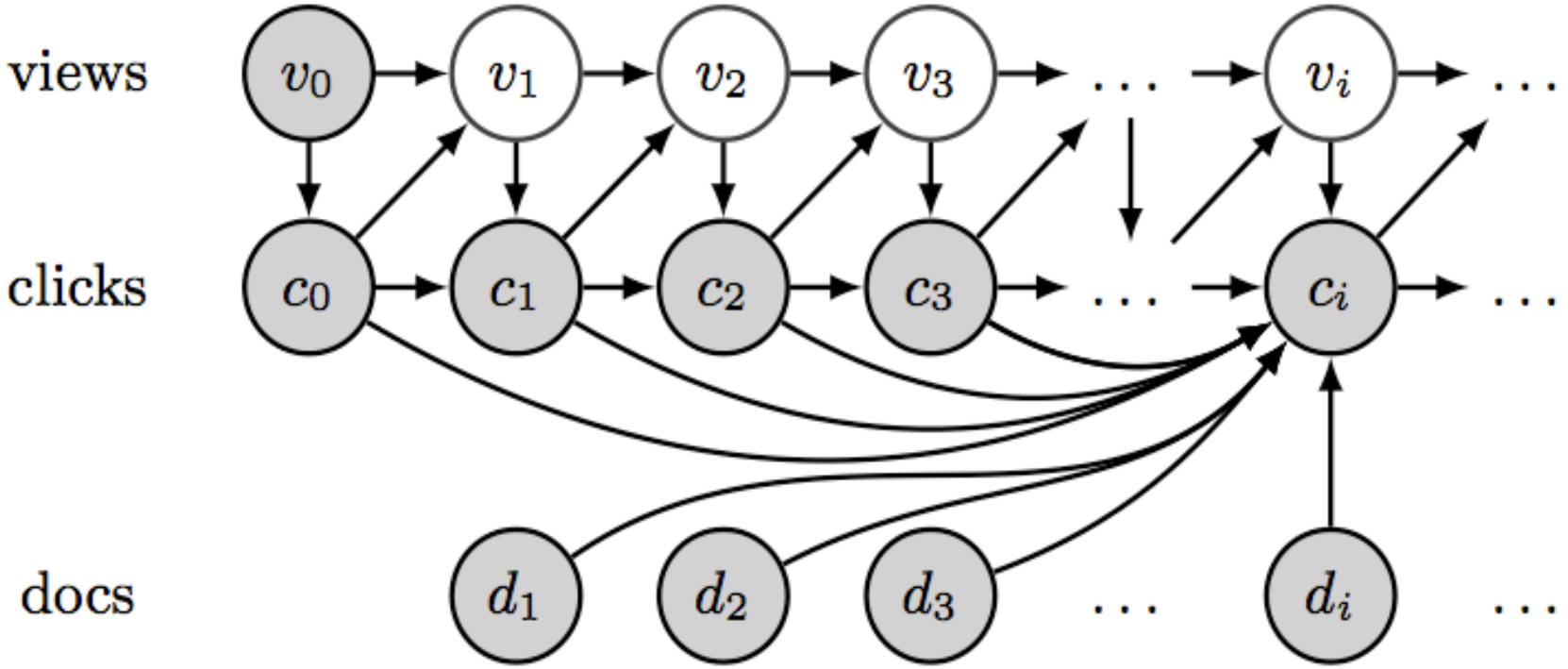


Modeling Views based on position

$$p(v_i = 1 \mid v_{i-1} = 1, c_{i-1} = 1) = \frac{1}{(1 + \exp(-\alpha_i))}$$

$$p(v_i = 1 \mid v_{i-1} = 1, c_{i-1} = 0) = \frac{1}{(1 + \exp(-\beta_i))}$$

Sequential Click-View Model



Modeling clicks using position and information gain

$$p(c_i = 1 \mid v_i = 1, c_{1,\ldots,i-1}) = \frac{1}{(1 + \exp(-\gamma_i - \sum_{j=1}^{i-1} c_j - \rho(A_i) + \rho(A_{i-1})))}$$

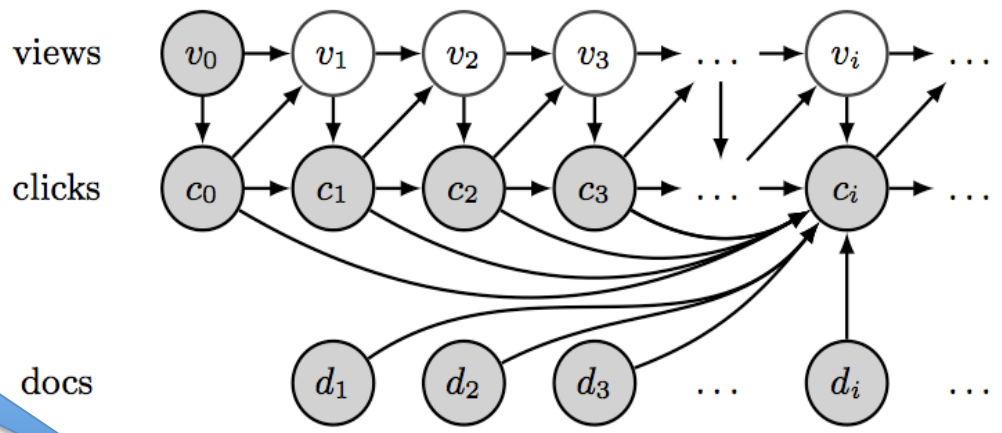
Coverage function's weights are learnt

Threshold

#clicks

Information gain

Sequential Click-View Model



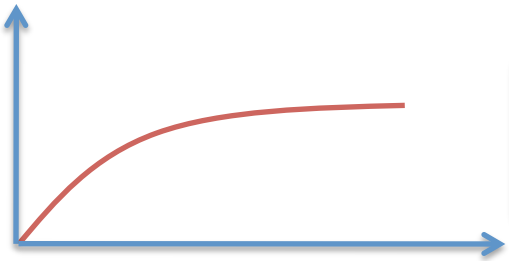
Selected Summary

Modular

$$\rho(D|S) := \sum_{s \in S} \sum_{j} [s]_j \left(a_j \sum_{d \in D} [d]_j + b_j \rho_j(D) \right).$$

Story

Features

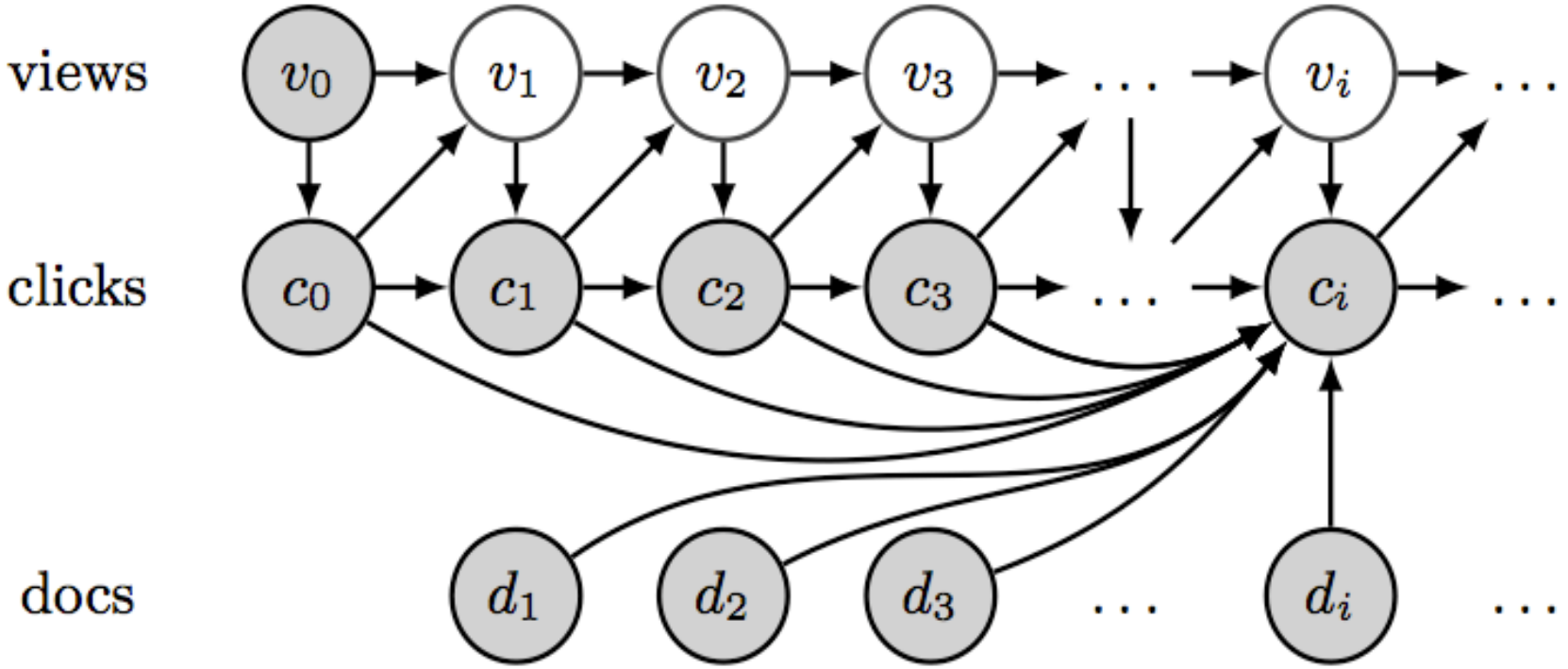


Submodular

Online Inference

- Treat missing views as hidden variables
  - Realistic interaction model
- Use the online EM algorithm
  - Infer the value of hidden variables
- Optimize parameters using SGD
  - Use additive weights
    - Background + story + category + user

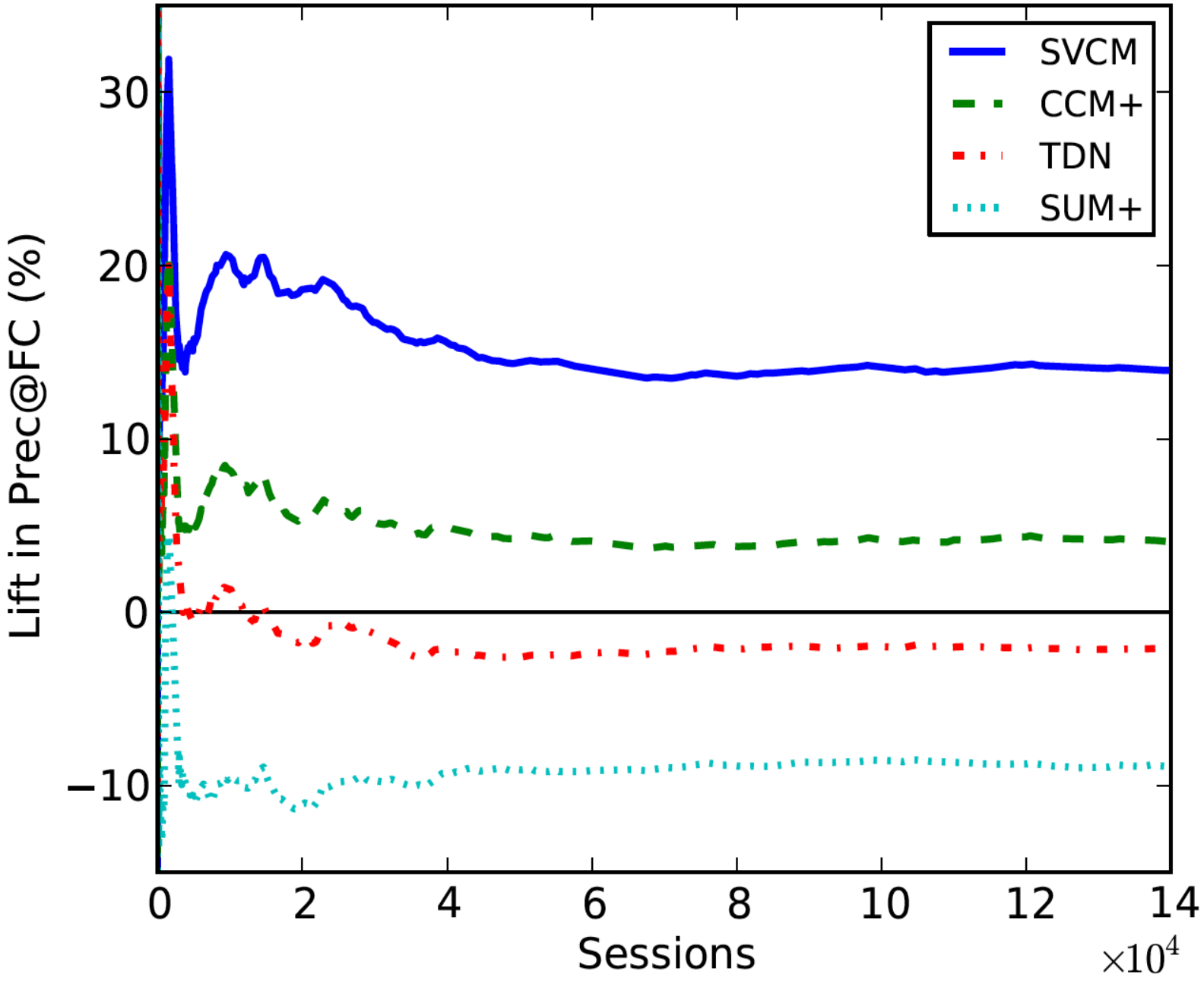
Online Inference



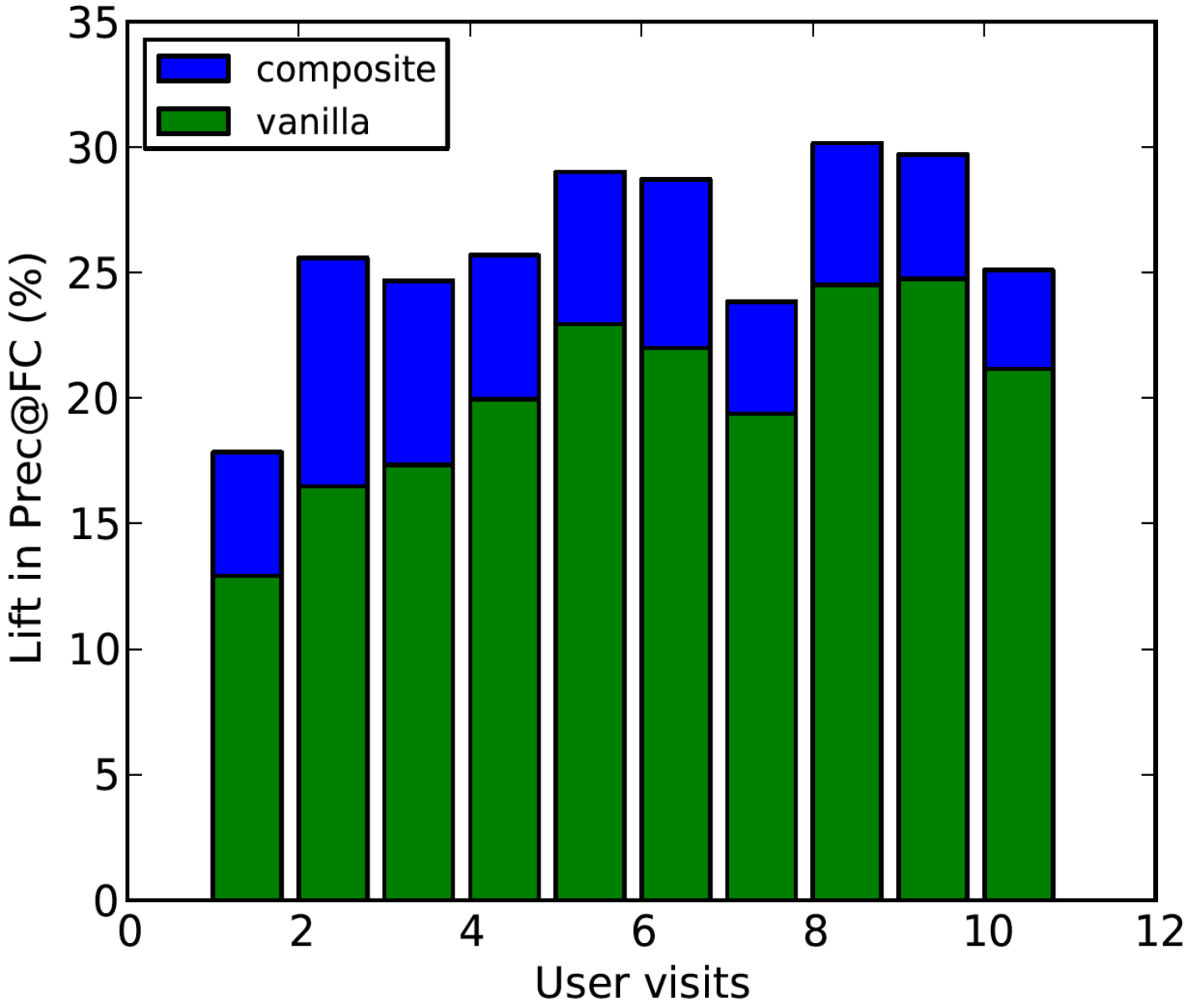
$$\Psi^* = \arg\min_{\Psi} \sum_{(c,d)} -\log p(c|\Psi, d) + \lambda\Omega(\Psi)$$

$$\Psi = \Psi_0 + \Psi_u + \Psi_s + \Psi_c.$$

How Does it Work?



How Does It Work?



5. Summary Future Directions

Summary

- Tools
  - Load distribution, balancing and synchronization
  - Clustering, Topic Models
- Models
  - Dynamic non-parametric models
  - Sequential latent variable models
- Inference Algorithms
  - Distributed batch
  - Sequential Monte Carlo
- Applications
  - User profiling
  - News content analysis & recommendation

Future Directions

- Theoretical bounds and guarantees
- Network data
  - Graph partitioning
- Non-parametric models
  - Learning structure from data
- Working under communication constraints
- Data distribution for particle filters