

YAHOO!

Graphical Models for the Internet

Alexander Smola & Amr Ahmed

Yahoo! Research & Australian National University

Santa Clara, CA

alex@smola.org blog.smola.org

Outline

- **Part 1 - Motivation**
 - Automatic information extraction
 - Application areas
- **Part 2 - Basic Tools**
 - Density estimation / conjugate distributions
 - Directed Graphical models and inference
- **Part 3 - Topic Models (our workhorse)**
 - Statistical model
 - Large scale inference (parallelization, particle filters)
- **Part 4 - Advanced Modeling**
 - Temporal dependence
 - Mixing clustering and topic models
 - Social Networks
 - Language models

Part 1 - Motivation

Data on the Internet

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

Finite resources

- **Editors are expensive**
- **Editors don't know users**
- **Barrier to i18n**
- **Abuse (intrusions are novel)**
- **Implicit feedback**
- **Data analysis (find interesting stuff rather than find x)**
- **Integrating many systems**
- **Modular design for data integration**
- **Integrate with given prediction tasks**

Invest in modeling and naming rather than data generation

Data on the Internet

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & c)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)

Finite resources

- Editors are expensive
- Editors / know users

unlimited amounts
of data

- New things are novel)
- Data analysis (find interesting stuff rather than find x)
- Integrating many systems
- Modular design for data integration
- Integrate with given prediction tasks

Invest in modeling and naming
rather than data generation

Clustering documents

Clustering documents

The screenshot shows the United Airlines website interface. At the top, there's the United logo and navigation links like 'My profile', 'Worldwide sites', and 'Customer service'. Below that are dropdown menus for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. A search bar is also present. The main content area features a 'BOOK FLIGHT' section with fields for 'From', 'To', 'Departing', and 'Returning'. There are also options for 'Roundtrip', 'One-way', and 'Multicity'. A large promotional banner in the center reads 'Use 30% fewer miles on your next United flight.' with an image of a large orange percentage sign. To the right, there's a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'. Below the flight booking section, there are links for 'United news and deals' and 'United-Continental merger'. At the bottom, there are links for 'Cars', 'Hotels', and 'Vacations'.

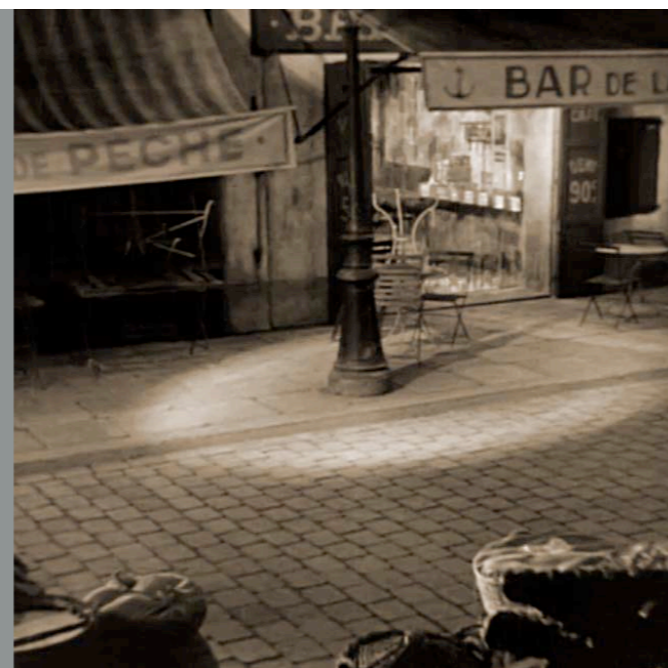
The screenshot shows the Australian National University (ANU) website. At the top, there's a 'Change Location' button and a search bar. Below that are navigation links for 'You Fly', 'Loyalty Programmes', and 'Promotions'. A blue navigation bar contains links for 'myEMAIL', 'IVLE', 'LIBRARY', 'MAPS', 'CALENDAR', 'SITEMAP', 'CONTACT', and 'e-CARDS'. Below this is another search bar with the text 'Search ANU...' and a 'GO' button. The main content area features a large banner with the text 'The Australian National University' and a navigation bar with links for 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. There are also several smaller banners and images, including one that says 'entred in Asia'.

Clustering documents

The screenshot shows the United Airlines website interface. At the top, there are navigation links for "My profile", "Worldwide sites", and "Customer service". Below this is a search bar and a menu with options like "Planning & booking", "Reservations & check-in", "Mileage Plus", and "Services & information". The main content area is divided into several sections: "Flights" with a "BOOK FLIGHT" and "REDEEM MILES" section, a "Log in" section with fields for Mileage Plus # or email address and password, and a "Travel information" section. There are also promotional banners, such as "Use 30% fewer miles on your next United flight" and "Earn up to 30,000 Bonus Miles".

The screenshot shows the Australian National University (ANU) website. At the top, there are navigation links for "EXPLORE ANU" and "A-Z INDEX", along with a search bar and utility links for "WEB", "CONTACTS", and "MAP". The main header features the ANU logo and the text "The Australian National University". Below this is a navigation menu with options like "HOME", "FUTURE STUDENTS", "CURRENT STUDENTS", "RESEARCH & EDUCATION", "ABOUT ANU", and "STAFF". The main content area features a large banner for "Ash forests rise and rise again" with a "read more" link. Below the banner are several featured articles: "Forests renew after Black Saturday fires", "School of Music at Floriade", "Undergraduate studies", and "Higher Degree Research". At the bottom, there are navigation buttons for "PROSPECTIVE STUDENTS", "CURRENT STUDENTS", "STAFF", "ALUMNI", and "VISITORS".

The screenshot shows the Chez Panisse website navigation menu. The menu items are: "RESERVATIONS", "RESTAURANT & CAFÉ", "MENUS", "RESTAURANT • CAFÉ", "MONDAY NIGHTS • WINE LIST", "ABOUT", "CHEZ PANISSE • ALICE WATERS", "OUR CHEFS • FRIENDS • PRESS", "FOUNDATION & MISSION", "SPECIAL EVENTS", "CALENDAR", "STORE", "BOOKS • POSTERS • GIFTS", "CONTACT", "INFORMATION", and "DIRECTIONS • MAILING LIST".



ng, Wining & Dining | Contact | Sitemap | About Suntec REIT



Clustering documents

The image shows a screenshot of the United Airlines website. The page features a navigation bar with 'UNITED' and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below the navigation, there are sections for 'Flights', 'Check-in', and 'Flight status'. A prominent red speech bubble with the word 'airline' is overlaid on the page. The website content includes a search form for flights, a 'Use 30% fewer miles on your next United flight' promotion, and a list of flight routes with prices, such as Singapore - Bangkok for SGD 395* and Singapore - Shanghai for SGD 824*.

The image shows a screenshot of the Australian National University (ANU) website. The page features a navigation bar with 'EXPLORE ANU', 'A-Z INDEX', and a search bar. Below the navigation, there are sections for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. A prominent red speech bubble with the word 'university' is overlaid on the page. The website content includes a news article titled 'Ash forests rise and rise again' and a list of navigation buttons for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

The image shows a screenshot of the Chez Panisse restaurant website. The page features a navigation bar with 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. Below the navigation, there are sections for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. A prominent red speech bubble with the word 'restaurant' is overlaid on the page. The website content includes a list of menu items and a background image of the restaurant's interior.

Today's mission

Find hidden structure in the data

Human understandable
Improved knowledge for estimation

Some applications

Hierarchical Clustering



NIPS 2010
Adams,
Ghahramani,
Jordan

Topics in text

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

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

Word segmentation

first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspeduponherknee,andthebrighteagereyeswerelookingupinto hers shecouldhearthevery tonesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderinghairthatwouldalwaysgetinto hereyesandstill as shelistened,orseemedtolisten,thewholeplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.thelonggrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashedhiswaythroughtheneighbouringpoolshcouldheartherattleoftheteacupsasthemarchhareandhisfriendssharedtheirneverendingmeal,andtheshrillvoiceofthequeen...



first, she dream ed of little alice herself ,and once again the tiny hand s were clasped upon her knee ,and the bright eager eyes were looking up into hers -- shecould hearthe very tone s of her voice , and see that queer little toss of herhead to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , thewhole place a round her became alive the strange creatures of her little sister 'sdream. thelong grass rustled ather feet as thewhitera bbit hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- shecould hearthe rattle ofthe tea cups as the marchhare and his friends shared their never -endingme a l ,and the ...

Mochihashi, Yamada, Ueda, ACL 2009

Language model

nevertheless ,
he was admired
by many of his immediate subordinates
for his long work hours
and dedication to building northwest
into what he called a “ mega carrier
.”

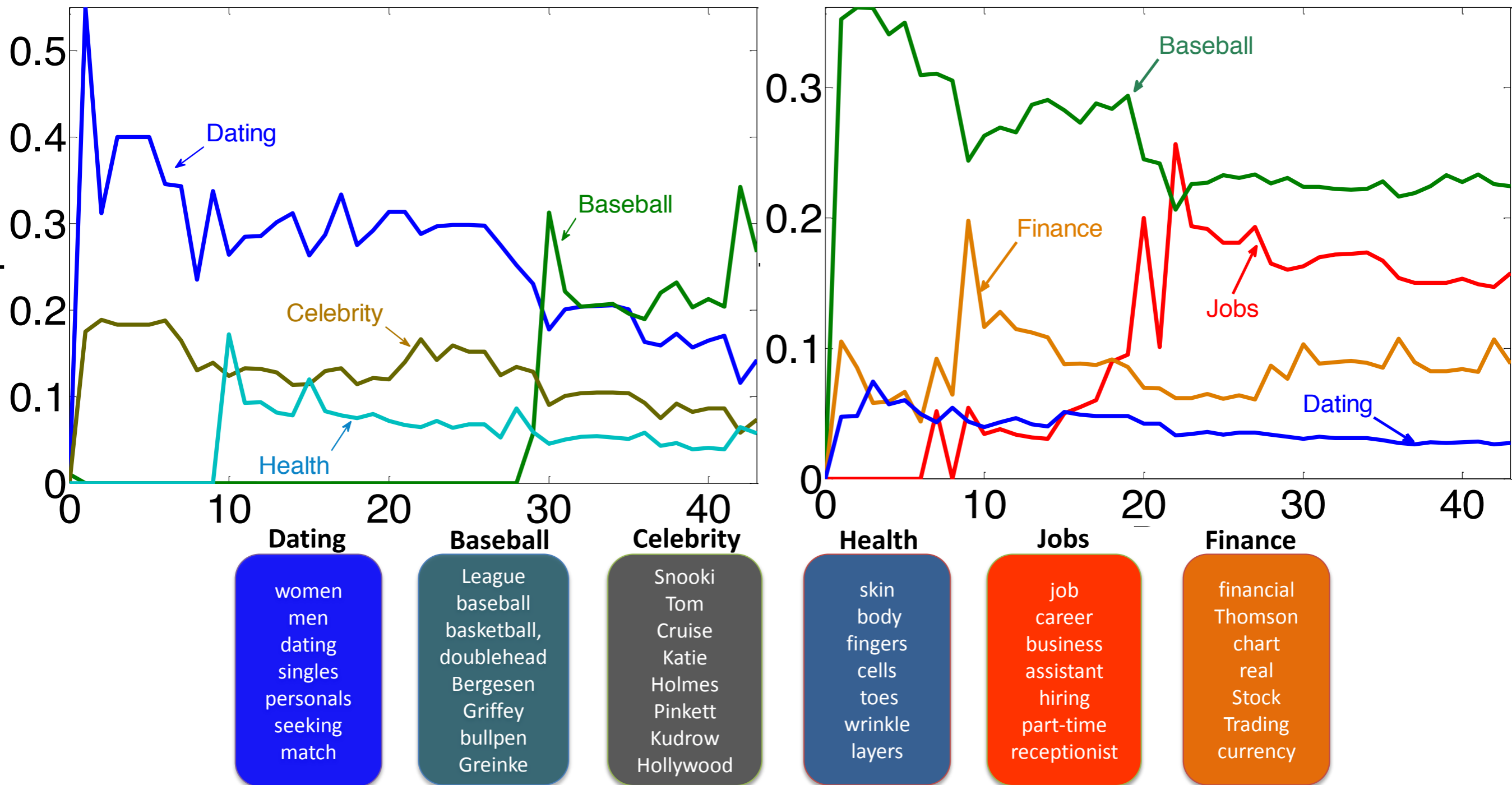
although
preliminary findings
were reported
more than a year ago ,
the latest results
appear
in today 's
new england journal of medicine ,
a forum
likely to bring new attention to the problem
.

south korea
registered a trade deficit of \$ 101 million
in october
, reflecting the country 's economic sluggishness
, according to government figures released wednesday

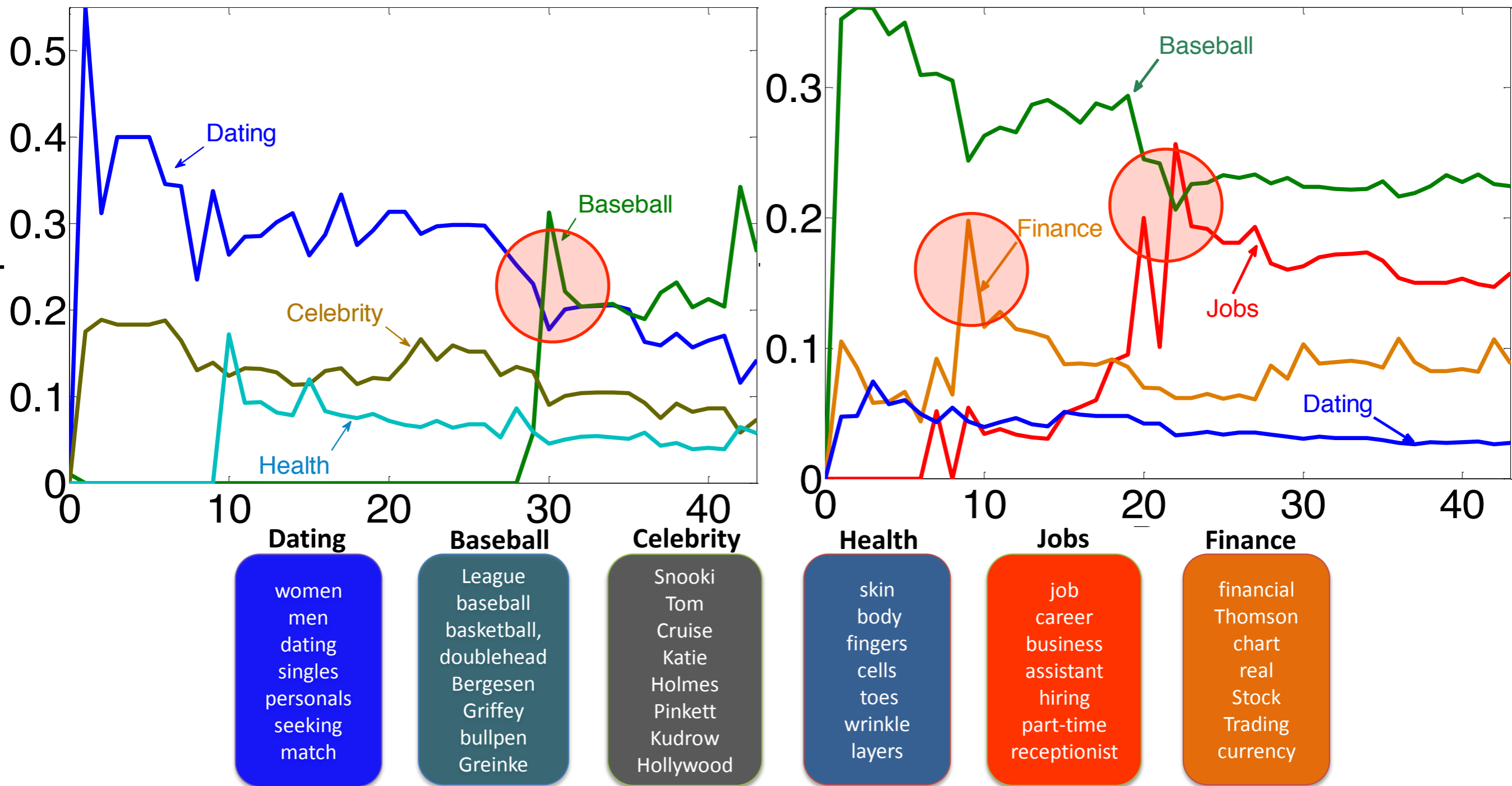
automatically synthesized
from Penn Treebank

Mochihashi, Yamada, Ueda
ACL 2009

User model over time



User model over time



Face recognition from captions



(a) Random samples from four clusters obtained using LDA on caption text [6].



(b) The corresponding clusters obtained by People-LDA.

Storylines from news

TOPICS

Sports

games
won
team
final
season
league
held

Politics

government
minister
authorities
opposition
officials
leaders
group

Unrest

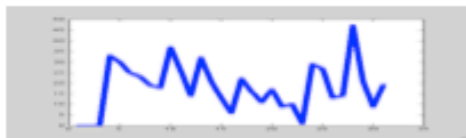
police
attack
run
man
group
arrested
move

Ahmed et al,
AISTATS 2011

STORYLINES

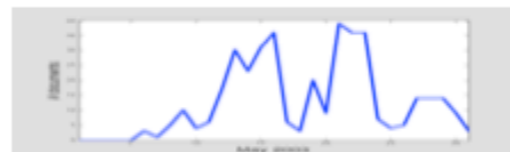
UEFA-soccer

champions	<i>Juventus</i>
goal	<i>AC Milan</i>
leg	<i>Real Madrid</i>
coach	<i>Milan</i>
striker	<i>Lazio</i>
midfield	<i>Ronaldo</i>
penalty	<i>Lyon</i>



Tax bills

tax	<i>Bush</i>
billion	<i>Senate</i>
cut	<i>US</i>
plan	<i>Congress</i>
budget	<i>Fleischer</i>
economy	<i>White House</i>
lawmakers	<i>Republican</i>

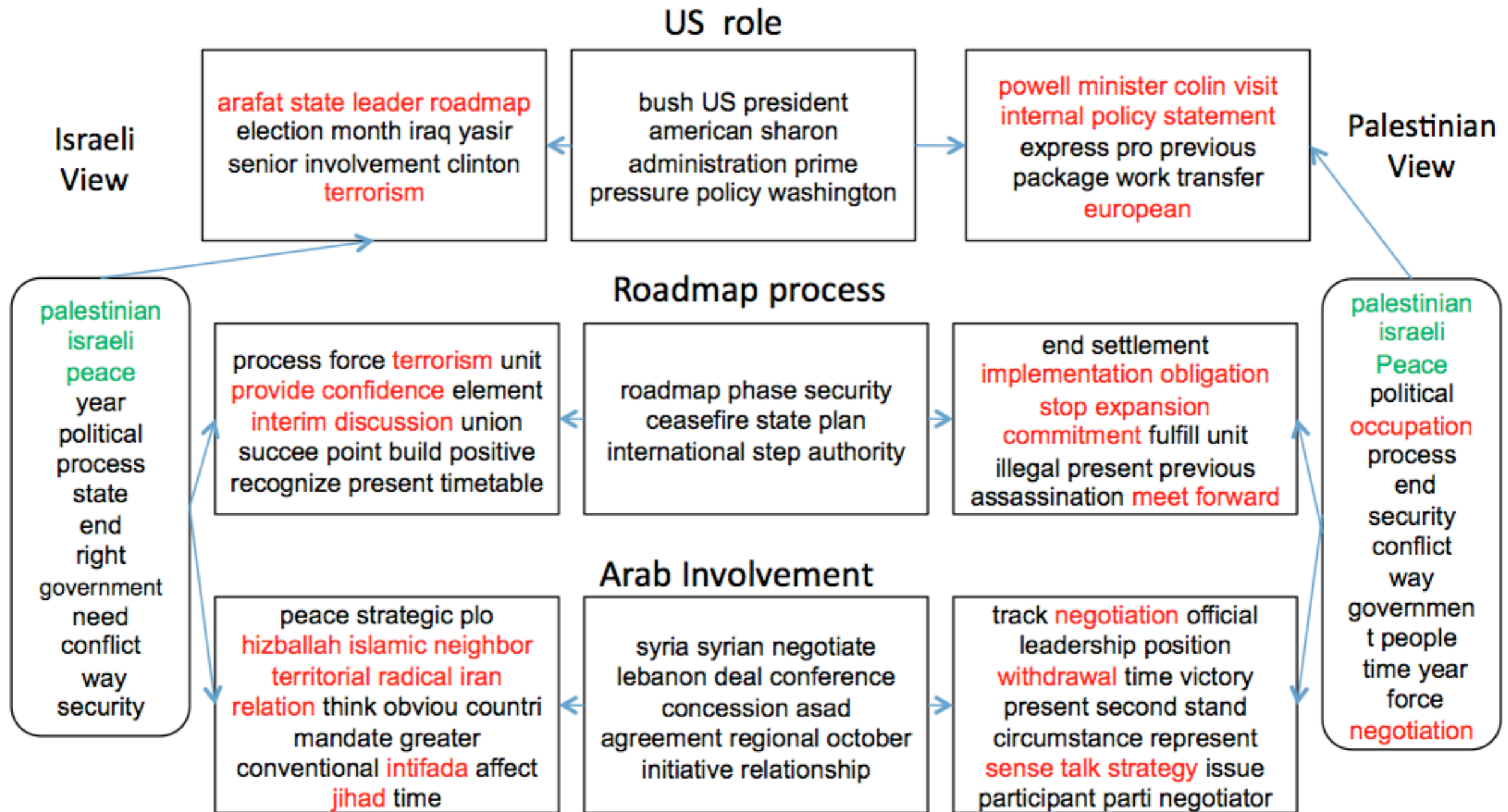


India-Pakistan tension

nuclear	<i>Pakistan</i>
border	<i>India</i>
dialogue	<i>Kashmir</i>
diplomatic	<i>New Delhi</i>
militant	<i>Islamabad</i>
insurgency	<i>Musharraf</i>
missile	<i>Vajpayee</i>





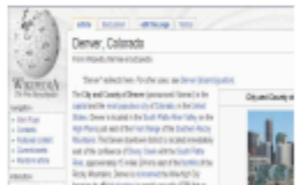


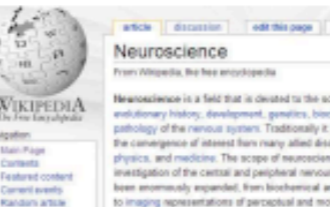


Ideology detection



Ahmed et al, 2010; Bitterlemons collection

Hypertext topic extraction

Topic 1			Topic 2		
neural	0.067	Artificial neural network  0.004	recognition	0.058	Speech recognition  0.004
network	0.047		speech	0.033	
networks	0.039	Neural network  0.003	language	0.015	Pattern recognition  0.004
learning	0.027		pattern	0.012	
artificial	0.017		handwriting	0.011	
data	0.015		evaluation	0.010	
models	0.014		robots	0.010	
function	0.014		systems	0.009	
Topic 3			Topic 4		
vancouver	0.051	Denver, Colorado  0.0008	brain	0.047	Cognitive science  0.003
denver	0.043		science	0.026	
city	0.041	Vancouver  0.0002	press	0.011	Neuroscience  0.002
retrieved	0.024		neurons	0.010	
colorado	0.011		mind	0.010	
area	0.009		systems	0.010	
population	0.009		human	0.010	
canada	0.008				

Gruber, Rosen-Zvi, Weiss; UAI 2008

Alternatives

Ontologies


dmoz open directory project In partnership with **Aol Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

[advanced](#)

<u>Arts</u> Movies , Television , Music ...	<u>Business</u> Jobs , Real Estate , Investing ...	<u>Computers</u> Internet , Software , Hardware ...
<u>Games</u> Video Games , RPGs , Gambling ...	<u>Health</u> Fitness , Medicine , Alternative ...	<u>Home</u> Family , Consumers , Cooking ...
<u>Kids and Teens</u> Arts , School Time , Teen Life ...	<u>News</u> Media , Newspapers , Weather ...	<u>Recreation</u> Travel , Food , Outdoors , Humor ...
<u>Reference</u> Maps , Education , Libraries ...	<u>Regional</u> US , Canada , UK , Europe ...	<u>Science</u> Biology , Psychology , Physics ...
<u>Shopping</u> Clothing , Food , Gifts ...	<u>Society</u> People , Religion , Issues ...	<u>Sports</u> Baseball , Soccer , Basketball ...
<u>World</u> Català , Dansk , Deutsch , Español , Français , Italiano , 日本語 , Nederlands , Polski , Русский , Svenska ...		


[Become an Editor](#) Help build the largest human-edited directory of the web



Copyright © 2011 Netscape

- continuous maintenance
- no guarantee of coverage
- difficult categories


Ontologies

 open directory project In partnership with **Aol Search.**

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

[advanced](#)

<u>Arts</u> Movies , Television , Music ...	<u>Business</u> Jobs , Real Estate , Investing ...	<u>Computers</u> Internet , Software , Hardware ...
<u>Games</u> Video Games , RPGs , Gambling ...	<u>Health</u> Fitness , Medicine , Alternative ...	<u>Home</u> Family , Consumers , Cooking ...
<u>Kids and Teens</u> Arts , School Time , Teen Life ...	<u>News</u> Media , Newspapers , Weather ...	<u>Recreation</u> Travel , Food , Outdoors , Humor ...
<u>Reference</u> Maps , Education , Libraries ...	<u>Regional</u> US , Canada , UK , Europe ...	<u>Science</u> Biology , Psychology , Physics ...
<u>Shopping</u> Clothing , Food , Gifts ...	<u>Society</u> People , Religion , Issues ...	<u>Sports</u> Baseball , Soccer , Basketball ...
<u>World</u> Català , Dansk , Deutsch , Español , Français , Italiano , 日本語 , Nederlands , Polski , Русский , Svenska ...		

[Become an Editor](#) Help build the largest human-edited directory of the web 

Copyright © 2011 Netscape

4,855,150 sites - 90,367 editors - over 1,005,887 categories

- continuous maintenance
- no guarantee of coverage
- difficult categories

Face Classification



Iranian Face Database
Pose, Age and Expression

|| HOME ||

Sampels

 Frontal, Pose Age: 4	 Frontal, Pose Age: 11	 Frontal, Pose, Expression Age: 16	 Frontal, Pose, Expression Age: 21
 Frontal, Pose, Expression Age: 27	 Frontal, Pose, Expression Age: 46	 Frontal, Pose, Expression Age: 65	 Frontal, Pose, Expression Age: 82

- 100-1000 people
- 10k faces
- curated (not realistic)
- expensive to generate

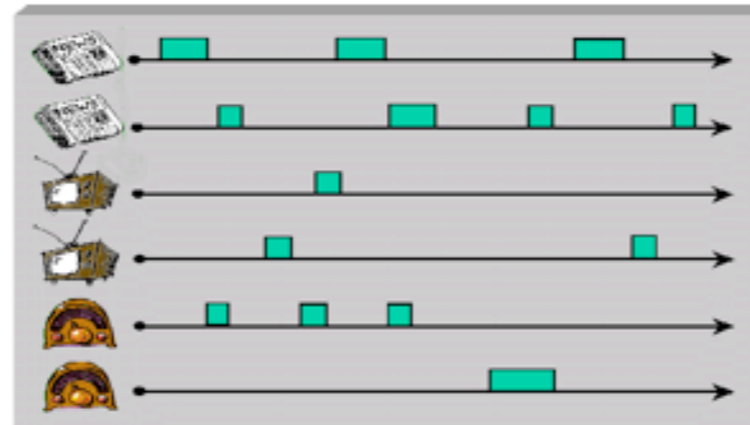
Topic Detection & Tracking

Information Technology Laboratory

Information Access Division (IAD)

NIST
National Institute of
Standards and Technology

Topic Detection and Tracking Evaluation



Topic Detection and Tracking research was pursued under the **DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program**:

Topic Detection and Tracking is an integral part of the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program. The goal of the TIDES program is to enable English-speaking users to access, correlate, and interpret multilingual sources of real-time information and to share the essence of this information with collaborators.

As a TIDES evaluation community, TDT provides a forum to discuss applications and techniques for detecting and tracking events that occur in real-time and the infrastructure to support common evaluations of component technologies. The TIDES project currently has one other evaluation community, The Text REtrieval Conference (TREC), and planning has begun for three new evaluations in the areas of Text Summarization, Question Answering and Quick Machine Translation.

- editorially curated training data
- expensive to generate
- subjective in selection of threads
- language specific

• [Multimodal Information Group Home](#)

• [Benchmark Tests](#)

• [Tools](#)

• [Test Beds](#)

• [Publications](#)

• [Links](#)

• [Contacts](#)

Advertising Targeting

Browse Ad Solutions

Media Spotlight

AUDIENCE

Affluents
Boomer Men
Boomer Women
Men 18-34
Men 18-49
Millennials
Online Dads
Online Moms
Women 18-34
Women 18-49

Your categories

Below you can edit the interests and inferred demographics that Google has associated with your cookie:

Category

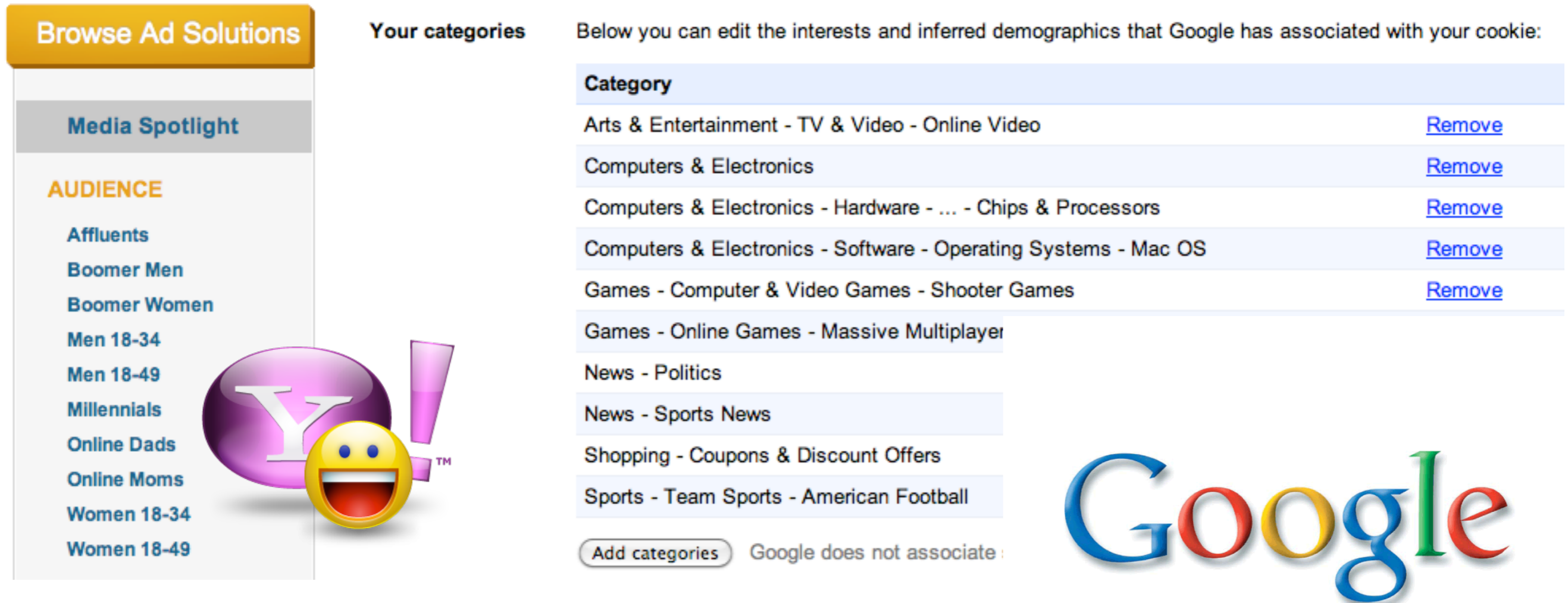
Arts & Entertainment - TV & Video - Online Video	Remove
Computers & Electronics	Remove
Computers & Electronics - Hardware - ... - Chips & Processors	Remove
Computers & Electronics - Software - Operating Systems - Mac OS	Remove
Games - Computer & Video Games - Shooter Games	Remove
Games - Online Games - Massive Multiplayer	Remove
News - Politics	Remove
News - Sports News	Remove
Shopping - Coupons & Discount Offers	Remove
Sports - Team Sports - American Football	Remove

[Add categories](#)

Google does not associate sensitive interest categories with your ads preferences.

- Needs training data **in every language**
- Is it really relevant for better ads?
- Does it **cover** relevant areas?

Advertising Targeting



The screenshot displays the 'Browse Ad Solutions' interface. On the left, under 'Media Spotlight', there is an 'AUDIENCE' section with a list of demographic and interest-based categories: Affluents, Boomer Men, Boomer Women, Men 18-34, Men 18-49, Millennials, Online Dads, Online Moms, Women 18-34, and Women 18-49. A cartoon character with a purple 'Y' and a yellow smiley face is positioned over the audience list. The main area, titled 'Your categories', contains a list of interests with 'Remove' links for each: Arts & Entertainment - TV & Video - Online Video, Computers & Electronics, Computers & Electronics - Hardware - ... - Chips & Processors, Computers & Electronics - Software - Operating Systems - Mac OS, Games - Computer & Video Games - Shooter Games, Games - Online Games - Massive Multiplayer, News - Politics, News - Sports News, Shopping - Coupons & Discount Offers, and Sports - Team Sports - American Football. At the bottom of this list is an 'Add categories' button and the text 'Google does not associate:'. The Google logo is visible in the bottom right corner of the interface.

- Needs training data **in every language**
- Is it really relevant for better ads?
- Does it **cover** relevant areas?

Challenges

- Scale
 - Millions to billions of instances (documents, clicks, users, messages, ads)
 - Rich structure of data (ontology, categories, tags)
 - Model description typically **larger than memory of single workstation**
- Modeling
 - Usually clustering or topic models **do not solve the problem**
 - Temporal structure of data
 - Side information for variables
 - **Solve problem. Don't simply apply a model!**
- Inference
 - 10k-100k clusters for hierarchical model
 - 1M-100M words
 - Communication is an issue for large state space

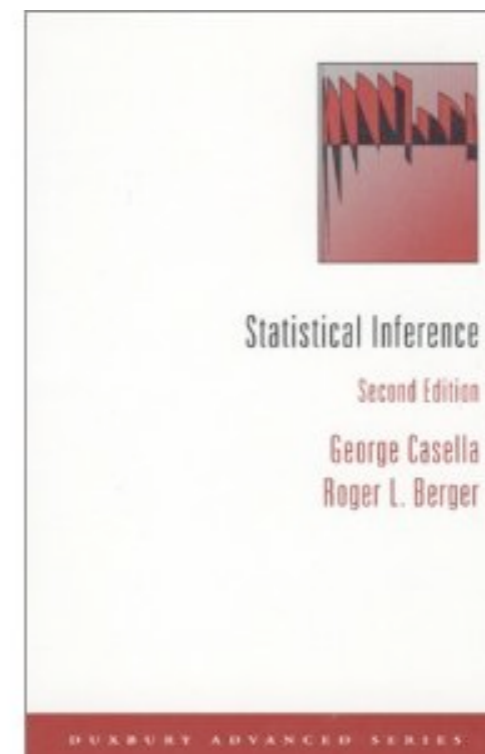
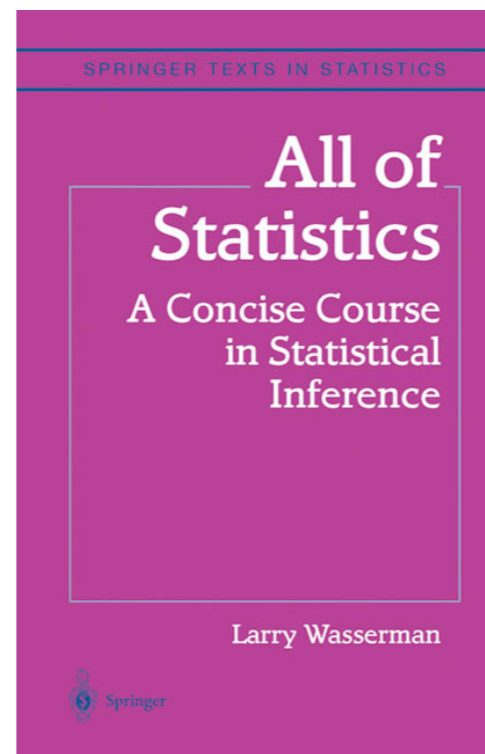
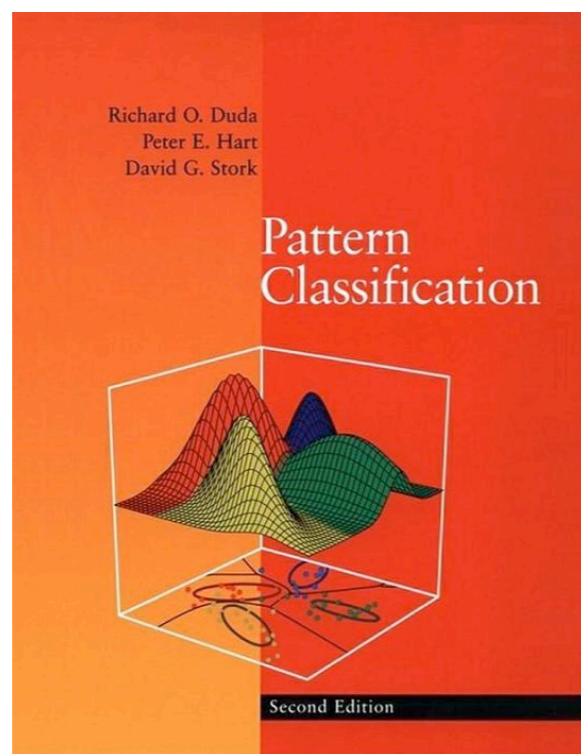
Summary - Part 1

- Essentially infinite amount of data
 - Labeling is prohibitively expensive
 - Not scalable for $i18n$
 - Even for *supervised* problems unlabeled data abounds. Use it.
 - User-understandable structure for representation purposes
 - Solutions are often customized to problem
- We can only cover building blocks in tutorial.**

Part 2 - Basic Tools



Statistics 101



Probability

- Space of events X
 - server status (working, slow, broken)
 - income of the user (e.g. \$95,000)
 - search queries (e.g. “graphical models”)
- Probability axioms (Kolmogorov)
$$\Pr(X) \in [0, 1], \Pr(\mathcal{X}) = 1$$
$$\Pr(\cup_i X_i) = \sum_i \Pr(X_i) \text{ if } X_i \cap X_j = \emptyset$$
- Example queries
 - $P(\text{server working}) = 0.999$
 - $P(90,000 < \text{income} < 100,000) = 0.1$

(In)dependence

- **Independence** $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
- Login behavior of two users (approximately)
- Disk crash in different colos (approximately)

(In)dependence

- **Independence** $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
 - Login behavior of two users (approximately)
 - Disk crash in different colos (approximately)
- **Dependent events**
 - Emails $\Pr(x, y) \neq \Pr(x) \cdot \Pr(y)$
 - Queries
 - News stream / Buzz / Tweets
 - IM communication
 - Russian Roulette

(In)dependence

- **Independence** $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
 - Login behavior of two users (approximately)
 - Disk crash in different colos (approximately)
- **Dependent events**
 - Emails $\Pr(x, y) \neq \Pr(x) \Pr(y)$
 - Queries
 - News stream / Buzz / Tweets
 - IM communication
 - Russian Roulette



Everywhere!

Independence



0.3

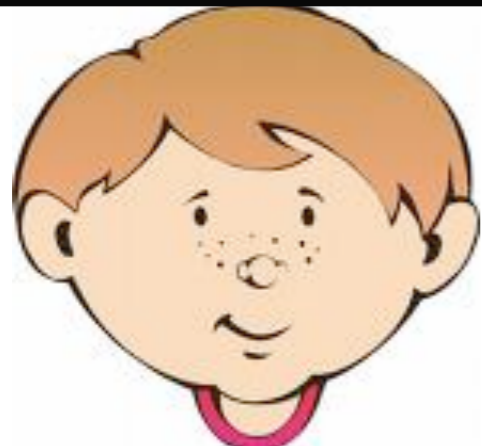
0.2



0.3

0.2

Dependence



0.45

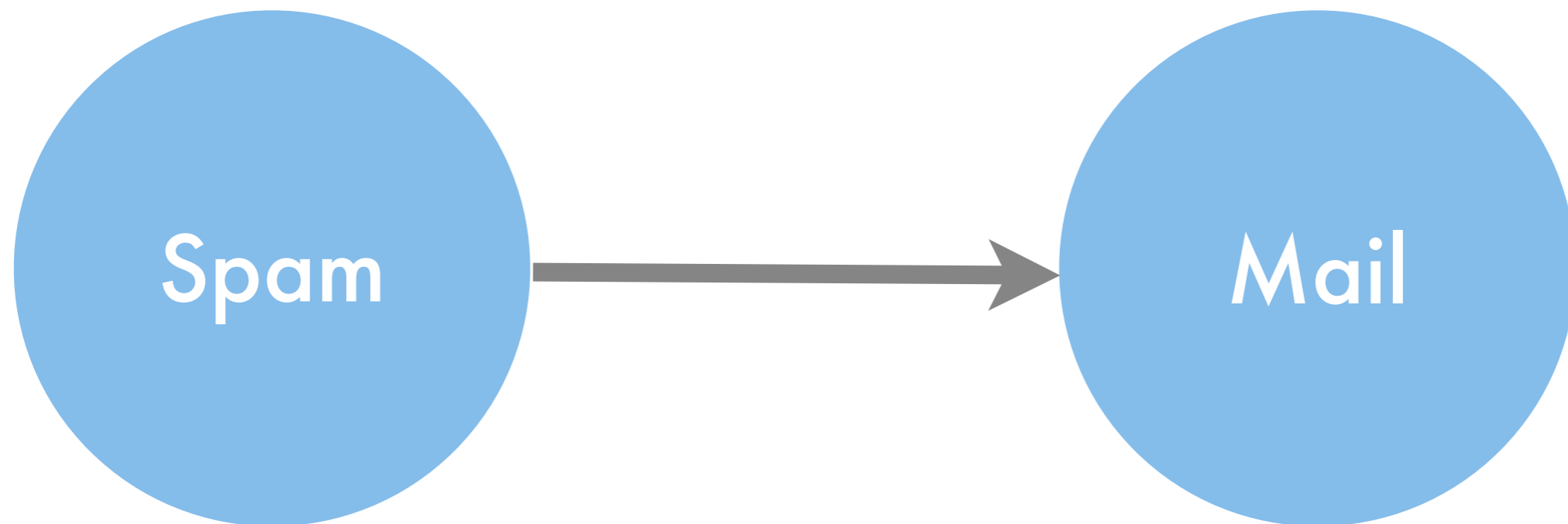
0.05



0.05

0.45

A Graphical Model



$$p(\text{spam}, \text{mail}) = p(\text{spam}) p(\text{mail} | \text{spam})$$

Bayes Rule

- **Joint Probability**

$$\Pr(X, Y) = \Pr(X|Y) \Pr(Y) = \Pr(Y|X) \Pr(X)$$

- **Bayes Rule**

$$\Pr(X|Y) = \frac{\Pr(Y|X) \cdot \Pr(X)}{\Pr(Y)}$$

- **Hypothesis testing**
- **Reverse conditioning**

AIDS test (Bayes rule)

- Data
 - Approximately **0.1%** are infected
 - Test detects **all** infections
 - Test reports positive for **1%** healthy people
- Probability of having AIDS if test is positive

AIDS test (Bayes rule)

- Data
 - Approximately **0.1%** are infected
 - Test detects **all** infections
 - Test reports positive for **1%** healthy people
- Probability of having AIDS if test is positive

$$\begin{aligned}\Pr(a = 1|t) &= \frac{\Pr(t|a = 1) \cdot \Pr(a = 1)}{\Pr(t)} \\ &= \frac{\Pr(t|a = 1) \cdot \Pr(a = 1)}{\Pr(t|a = 1) \cdot \Pr(a = 1) + \Pr(t|a = 0) \cdot \Pr(a = 0)} \\ &= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091\end{aligned}$$

Improving the diagnosis

Improving the diagnosis

- Use a follow-up test
 - Test 2 reports positive for 90% infections
 - Test 2 reports positive for 5% healthy people

$$\frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357$$

Improving the diagnosis

- Use a follow-up test
 - Test 2 reports positive for 90% infections
 - Test 2 reports positive for 5% healthy people

$$\frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357$$

- **Why can't we use Test 1 twice?**

Improving the diagnosis

- Use a follow-up test
 - Test 2 reports positive for 90% infections
 - Test 2 reports positive for 5% healthy people

$$\frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357$$

- **Why can't we use Test 1 twice?**

Outcomes are **not** independent but tests 1 and 2 are **conditionally independent**

$$p(t_1, t_2|a) = p(t_1|a) \cdot p(t_2|a)$$

Application: Naive Bayes



Naive Bayes Spam Filter

Naive Bayes Spam Filter

- **Key assumption**

Words occur independently of each other given the label of the document

$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

Naive Bayes Spam Filter

- **Key assumption**

Words occur independently of each other given the label of the document

$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

- **Spam classification via Bayes Rule**

$$p(\text{spam} | w_1, \dots, w_n) \propto p(\text{spam}) \prod_{i=1}^n p(w_i | \text{spam})$$

Naive Bayes Spam Filter

- **Key assumption**

Words occur independently of each other given the label of the document

$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

- **Spam classification via Bayes Rule**

$$p(\text{spam} | w_1, \dots, w_n) \propto p(\text{spam}) \prod_{i=1}^n p(w_i | \text{spam})$$

- **Parameter estimation**

Compute spam probability and word distributions for spam and ham

Naive Bayes Spam Filter

Equally likely phrases

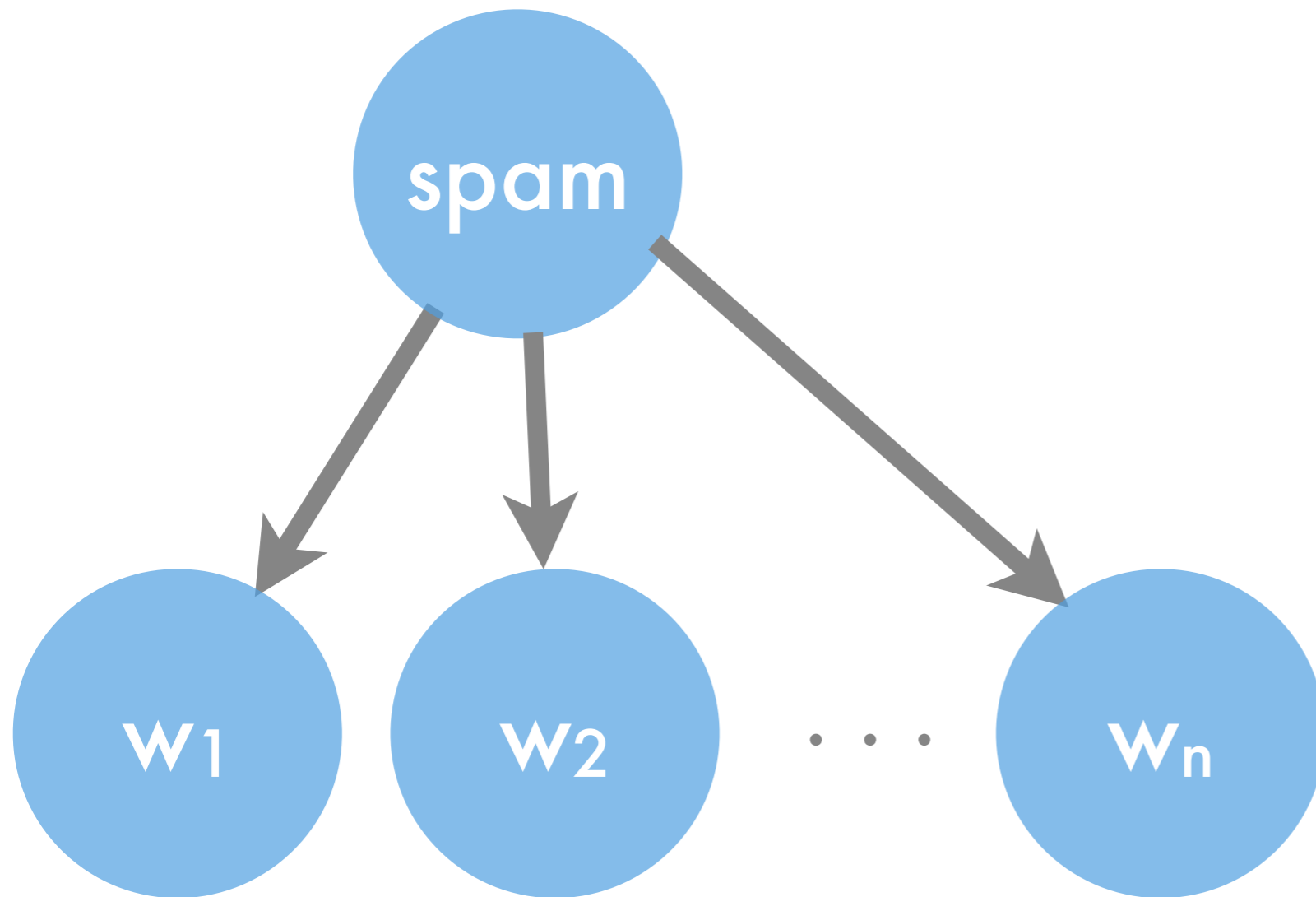
- Get rich quick. Buy WWW stock.
- Buy Viagra. Make your WWW experience last longer.
- You deserve a PhD from WWW University.
We recognize your expertise.

Naive Bayes Spam Filter

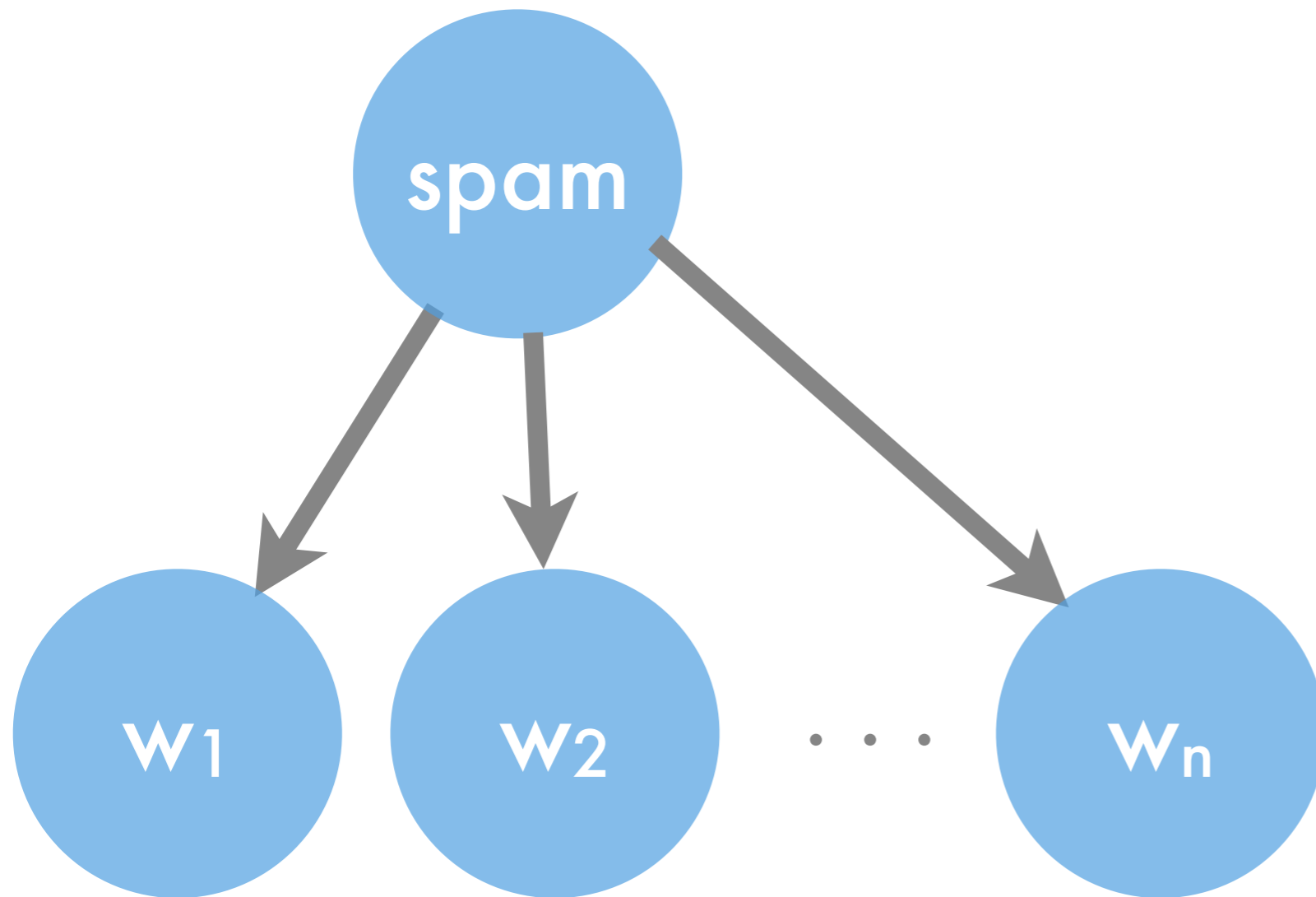
Equally likely phrases

- Get rich quick. Buy WWW stock.
- Buy Viagra. Make your WWW experience last longer.
- You deserve a PhD from WWW University.
We recognize your expertise.
- Make your rich WWW PhD experience last longer.

A Graphical Model

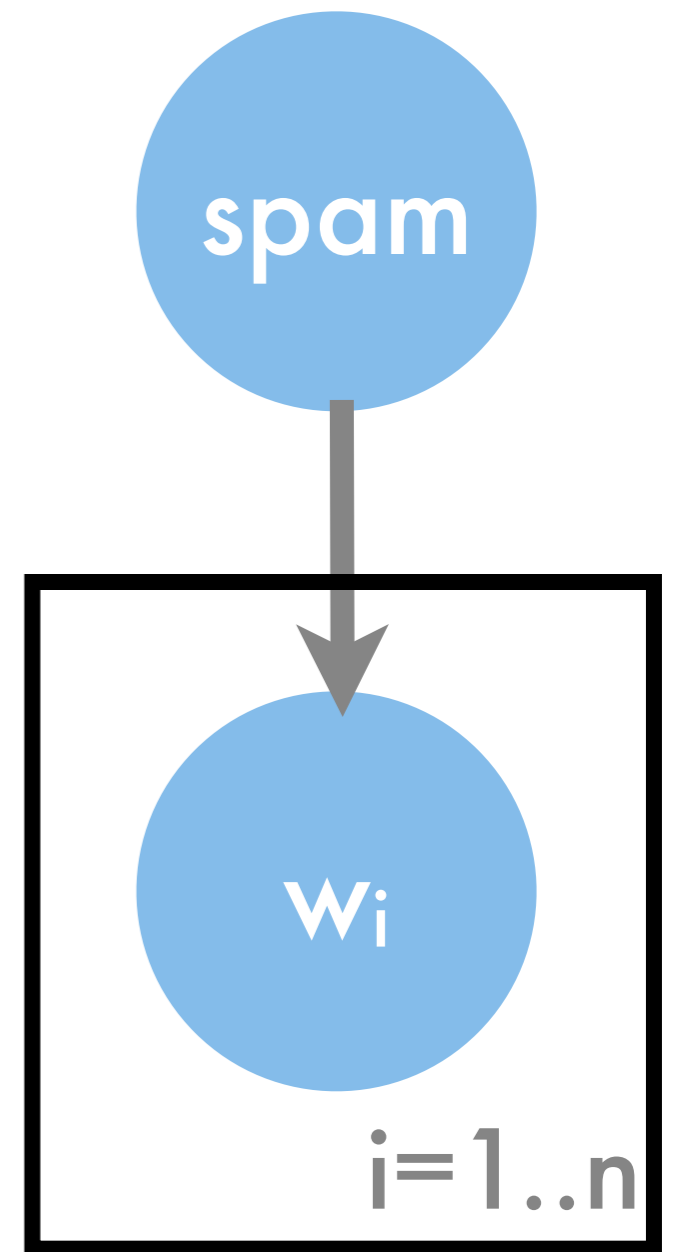
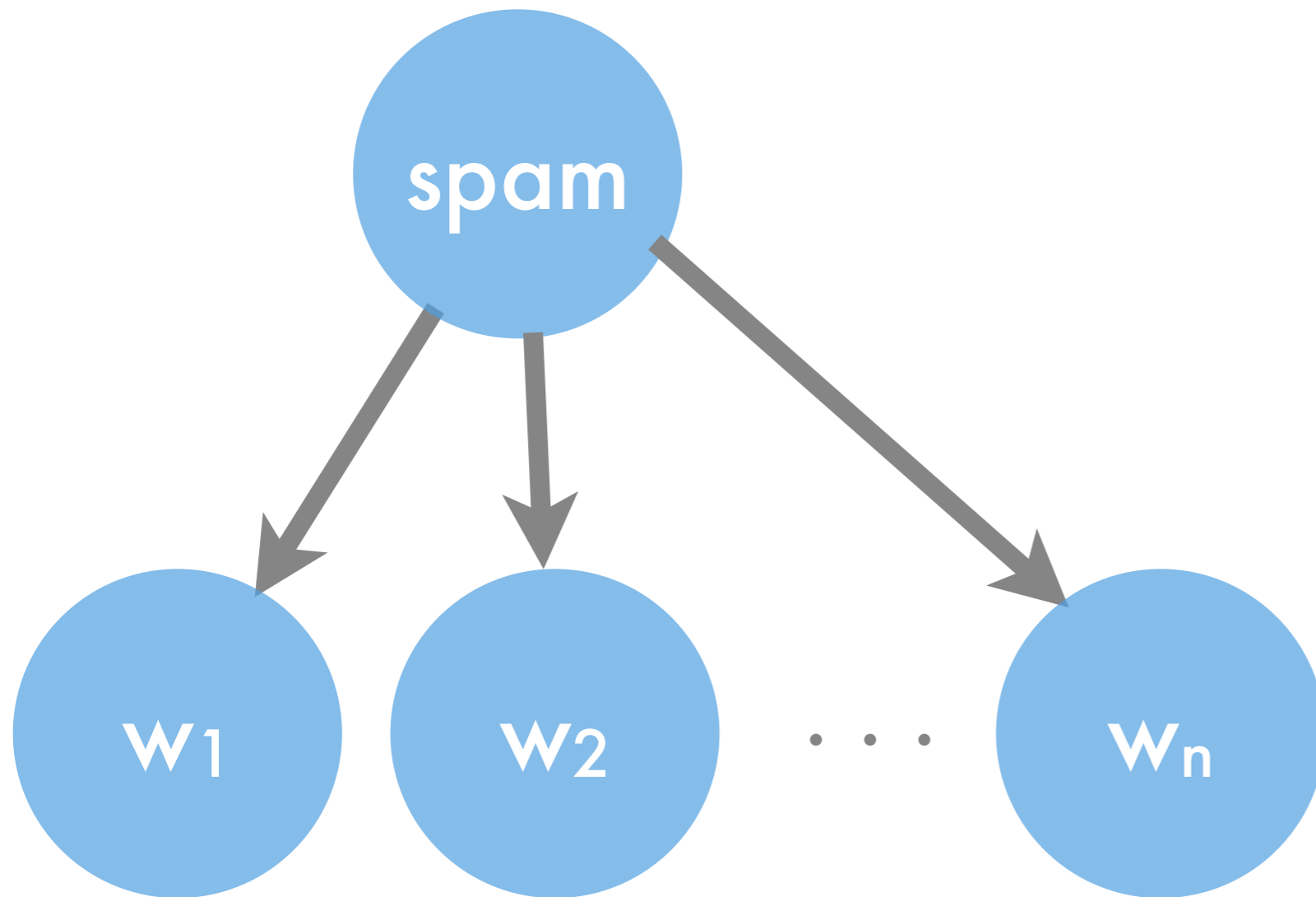


A Graphical Model



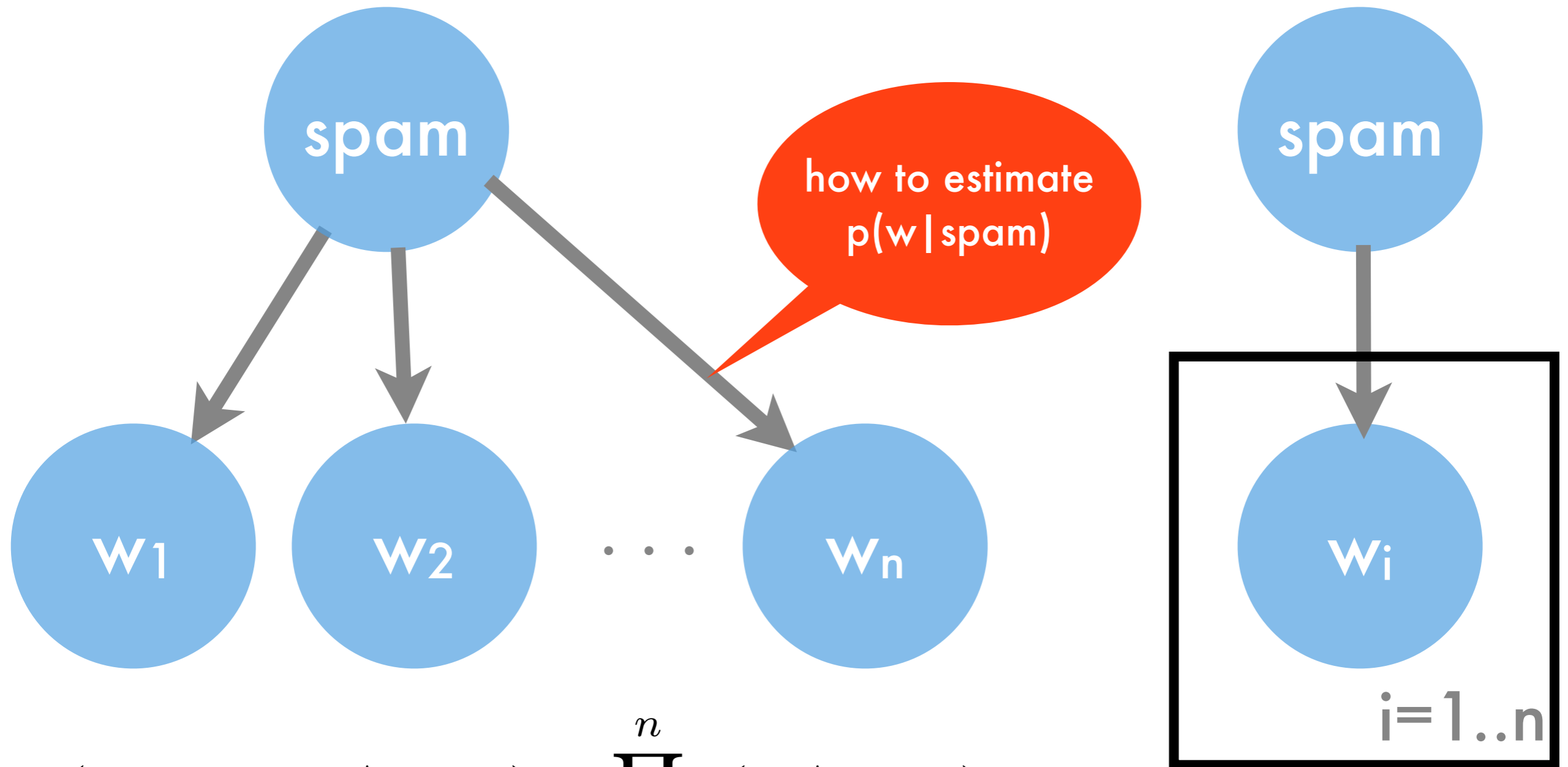
$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

A Graphical Model



$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

A Graphical Model



$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

Naive NaiveBayes Classifier

- Two classes (spam/ham)
- Binary features (e.g. presence of \$\$\$, viagra)
- Simplistic Algorithm
 - Count occurrences of feature for spam/ham
 - Count number of spam/ham mails

feature probability

$$p(x_i = \text{TRUE}|y) = \frac{n(i, y)}{n(y)} \text{ and } p(y) = \frac{n(y)}{n}$$

spam probability

$$p(y|x) \propto \frac{n(y)}{n} \prod_{i:x_i=\text{TRUE}} \frac{n(i, y)}{n(y)} \prod_{i:x_i=\text{FALSE}} \frac{n(y) - n(i, y)}{n(y)}$$

Naive NaiveBayes Classifier

what if $n(i,y)=0$?

what if $n(i,y)=n(y)$?

$$p(y|x) \propto \frac{n(y)}{n} \prod_{i:x_i=\text{TRUE}} \frac{n(i,y)}{n(y)} \prod_{i:x_i=\text{FALSE}} \frac{n(y) - n(i,y)}{n(y)}$$

Naive NaiveBayes Classifier

what if $n(i,y)=0$?

what if $n(i,y)=n(y)$?

$$p(y|x) \propto \frac{n(y)}{n} \prod_{i:x_i=\text{TRUE}} \frac{n(i,y)}{n(y)} \prod_{i:x_i=\text{FALSE}} \frac{n(y) - n(i,y)}{n(y)}$$

Estimating Probabilities



Two outcomes (binomial)

- Example: probability of 'viagra' in spam/ham
- Data likelihood

$$p(X; \pi) = \pi^{n_1} (1 - \pi)^{n_0}$$

- Maximum Likelihood Estimation
 - Constraint $\pi \in [0, 1]$
 - Taking derivatives yields

$$\pi = \frac{n_1}{n_0 + n_1}$$

n outcomes (multinomial)

- Example: USA, Canada, India, UK, NZ
- Data likelihood

$$p(X; \pi) = \prod_i \pi_i^{n_i}$$

- Maximum Likelihood Estimation

- Constrained optimization problem $\sum_i \pi_i = 1$
- Using log-transform yields

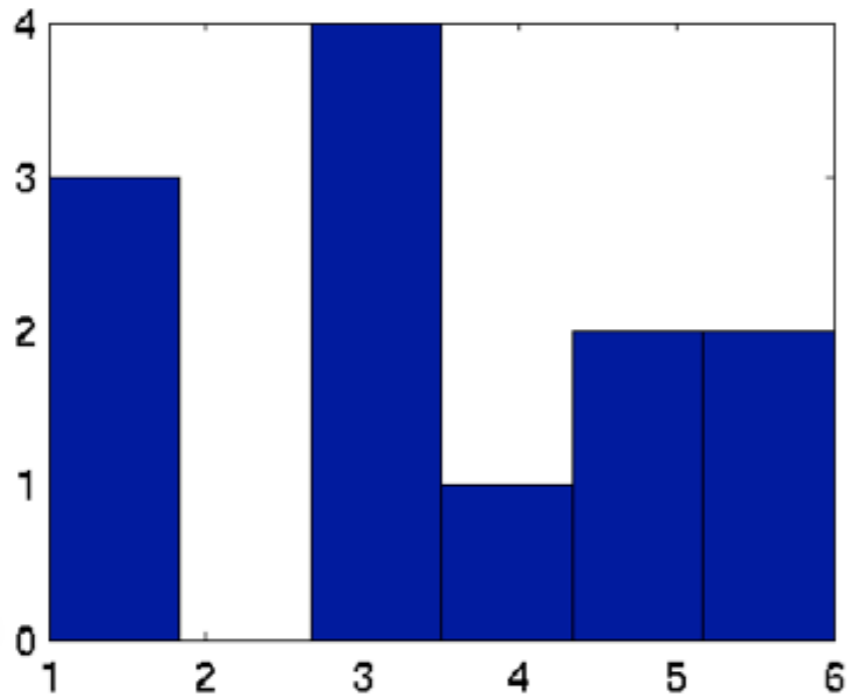
$$\pi_i = \frac{n_i}{\sum_j n_j}$$

Tossing a Dice

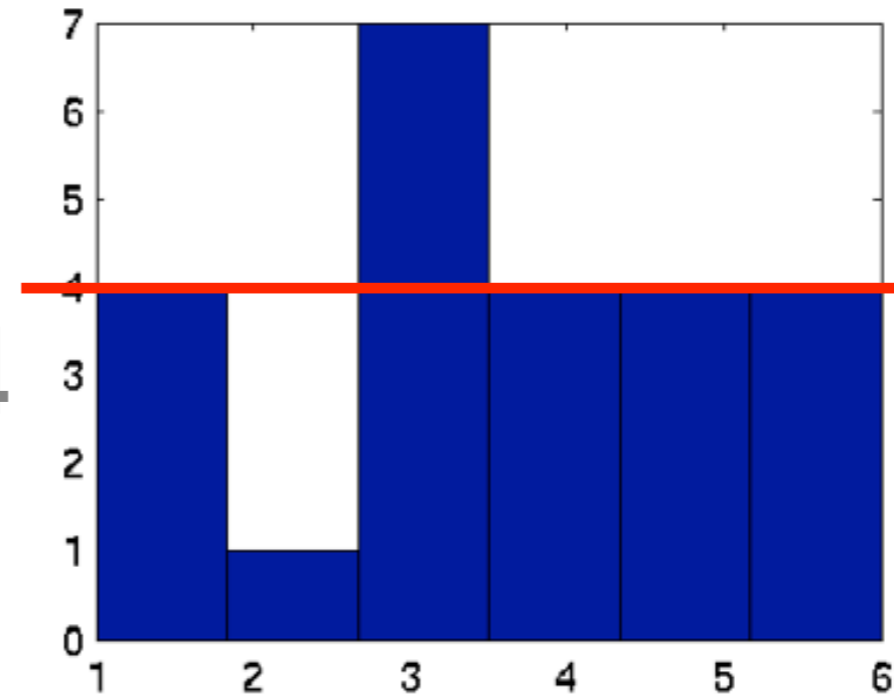
12



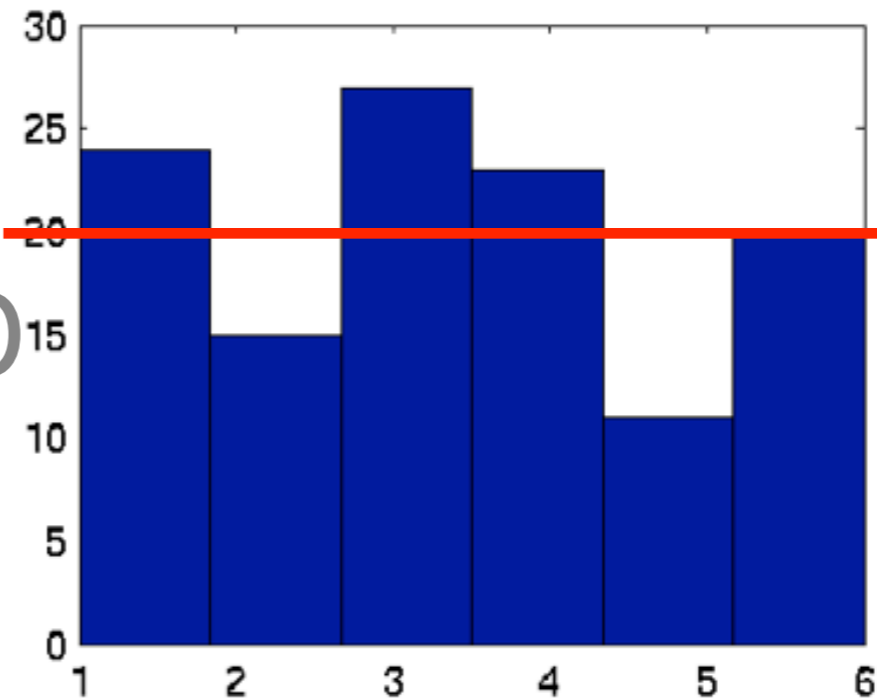
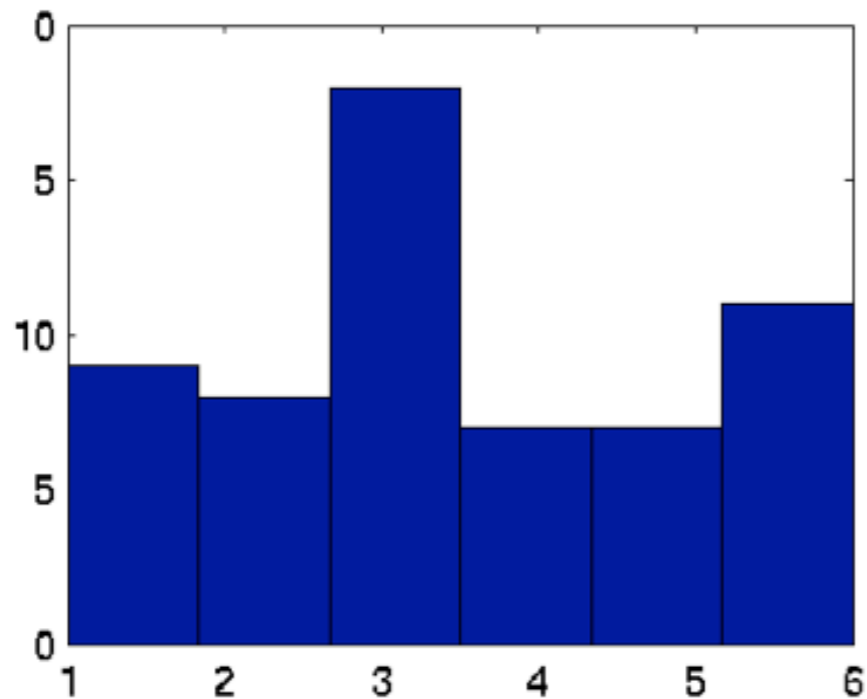
60



24



120

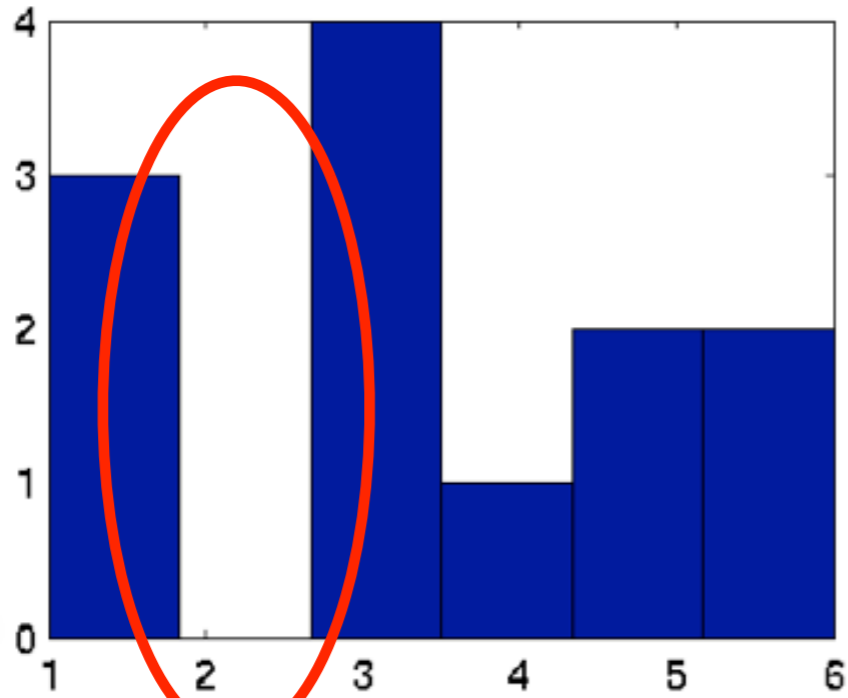


Tossing a Dice

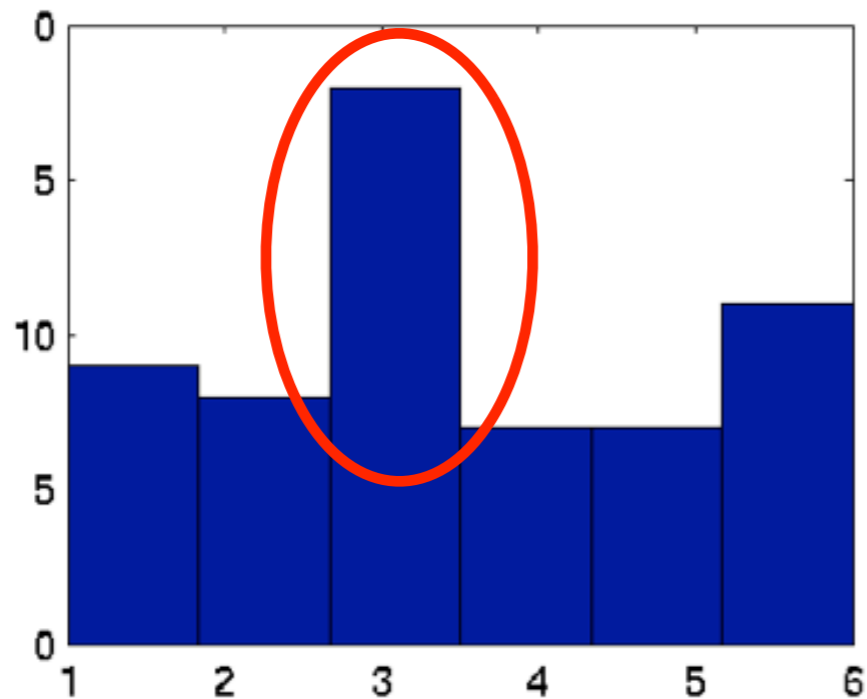
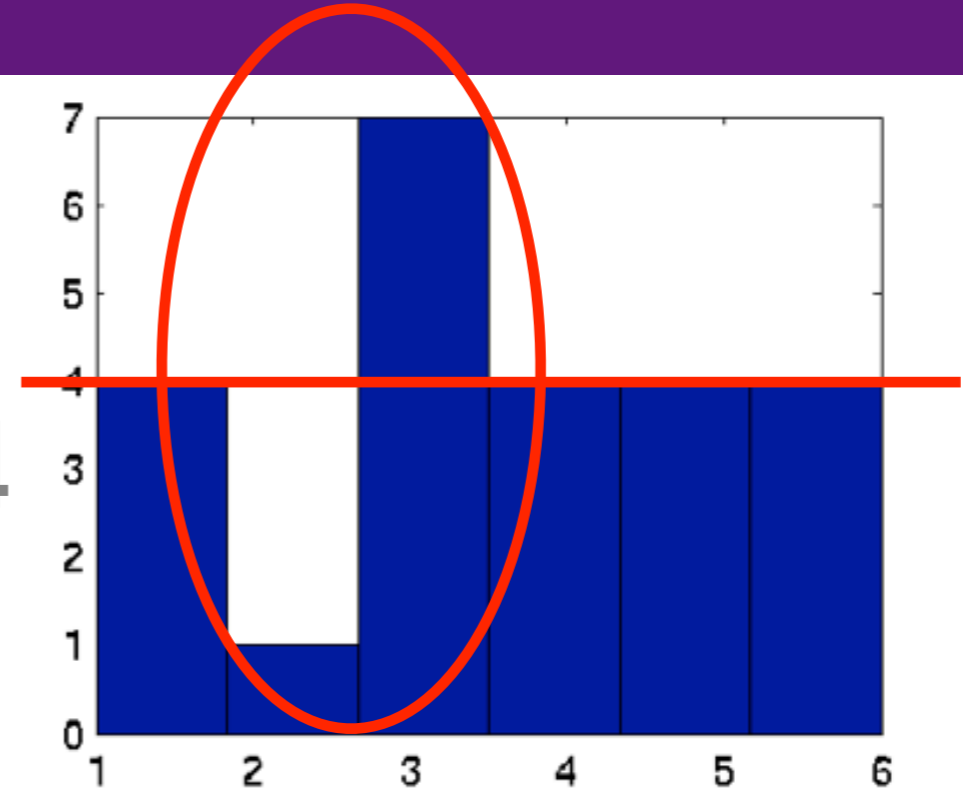
12



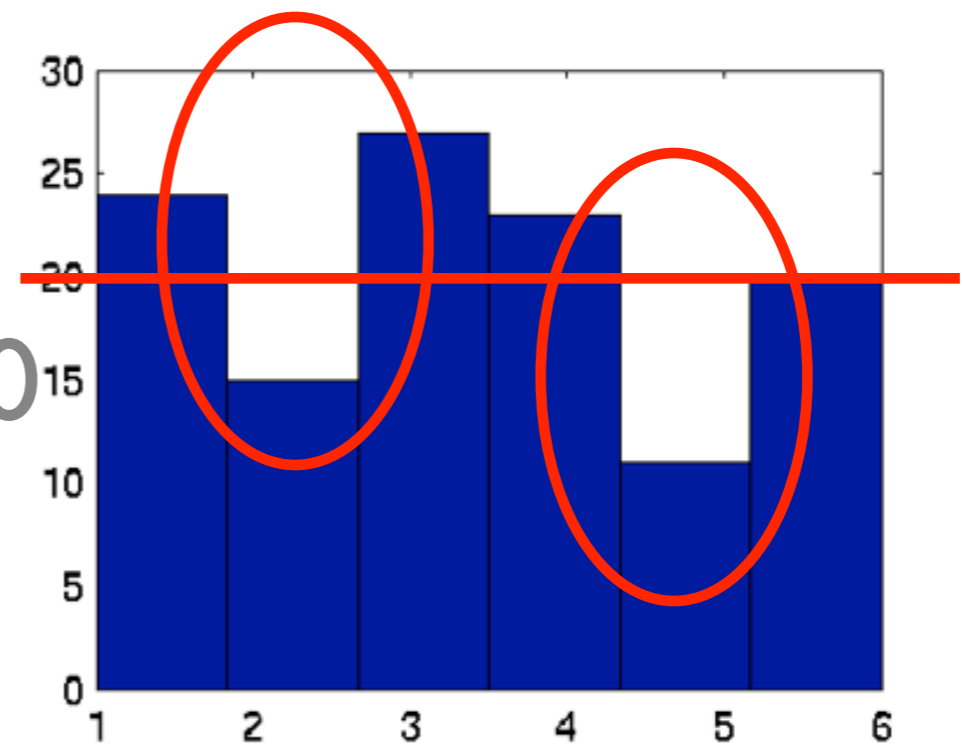
60



24



120



Conjugate Priors

- Unless we have lots of data estimates are weak
- Usually we have an idea of what to expect

$$p(\theta|X) \propto p(X|\theta) \cdot p(\theta)$$

we might even have 'seen' such data before

- Solution: add 'fake' observations

$$p(\theta) \propto p(X_{\text{fake}}|\theta) \text{ hence } p(\theta|X) \propto p(X|\theta)p(X_{\text{fake}}|\theta) = p(X \cup X_{\text{fake}}|\theta)$$

- Inference (generalized Laplace smoothing)

$$\frac{1}{n} \sum_{i=1}^n \phi(x_i) \longrightarrow \frac{1}{n+m} \sum_{i=1}^n \phi(x_i) + \frac{m}{n+m} \mu_0$$

fake count

fake mean

Conjugate Prior in action

$$m_i = m \cdot [\mu_0]_i$$

$$p(x = i) = \frac{n_i}{n} \longrightarrow p(x = i) = \frac{n_i + m_i}{n + m}$$

Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.15	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

Conjugate Prior in action

$$m_i = m \cdot [\mu_0]_i$$

- **Discrete Distribution**

$$p(x = i) = \frac{n_i}{n} \longrightarrow p(x = i) = \frac{n_i + m_i}{n + m}$$

- **Tossing a dice**

Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.15	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

Conjugate Prior in action

$$m_i = m \cdot [\mu_0]_i$$

- **Discrete Distribution**

$$p(x = i) = \frac{n_i}{n} \longrightarrow p(x = i) = \frac{n_i + m_i}{n + m}$$

- **Tossing a dice**

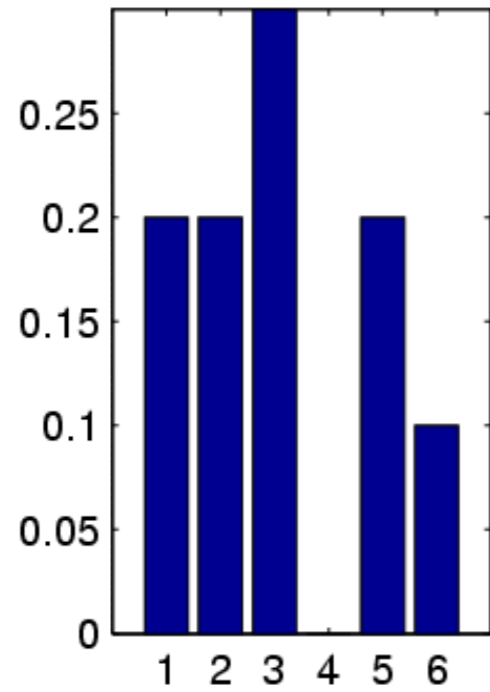
Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.15	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

- **Rule of thumb**

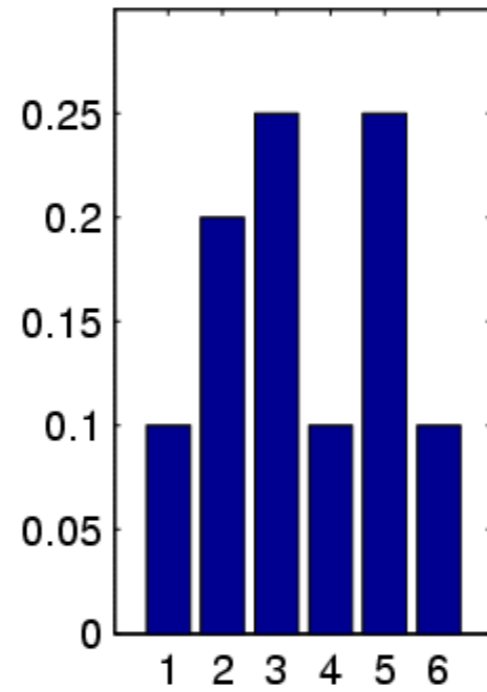
need 10 data points (or prior) per parameter

Honest dice

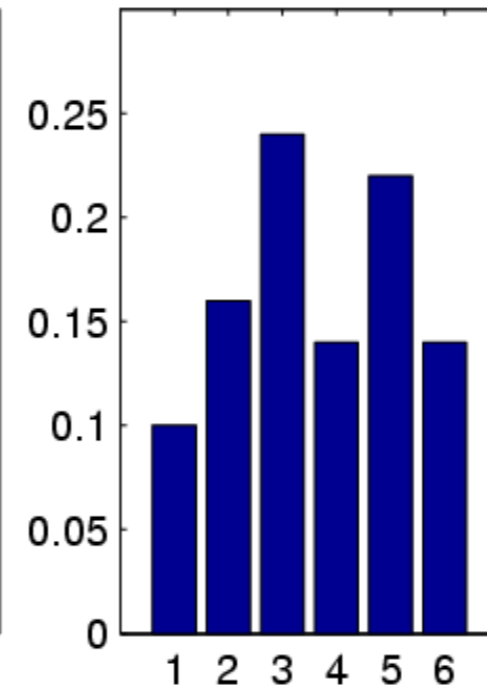
MLE



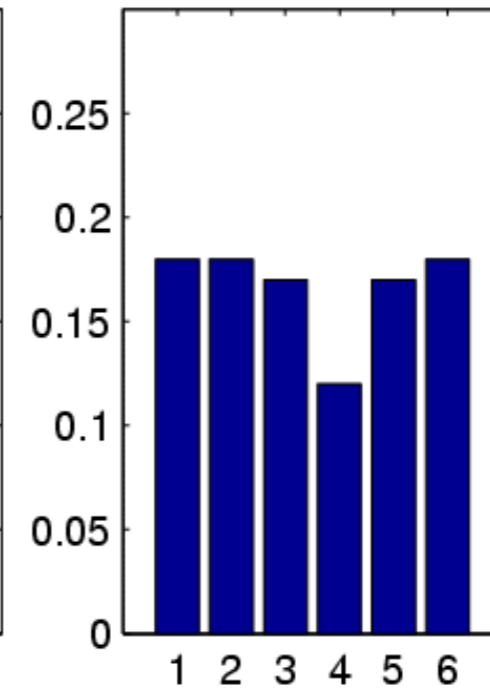
12



24

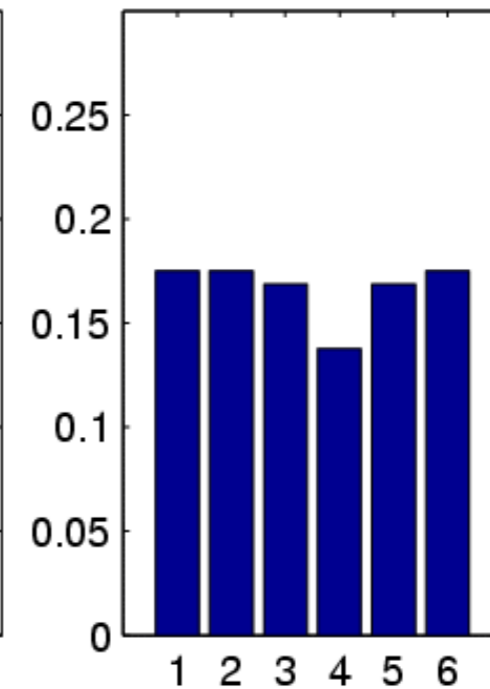
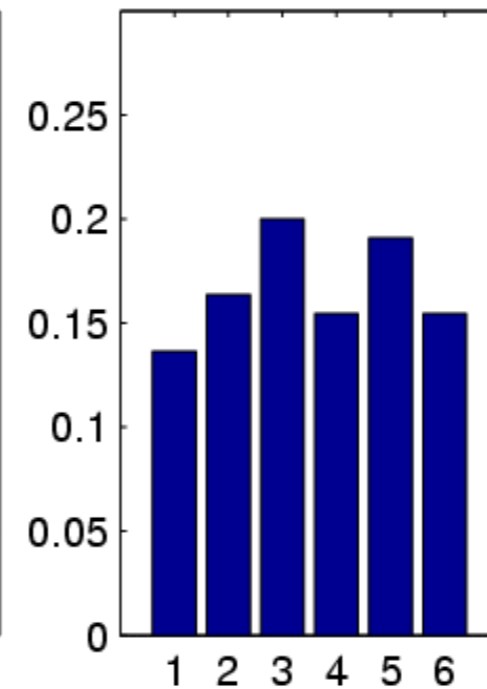
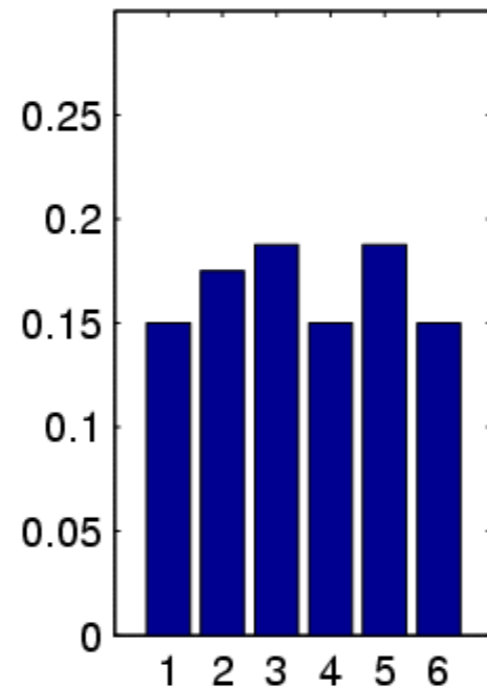
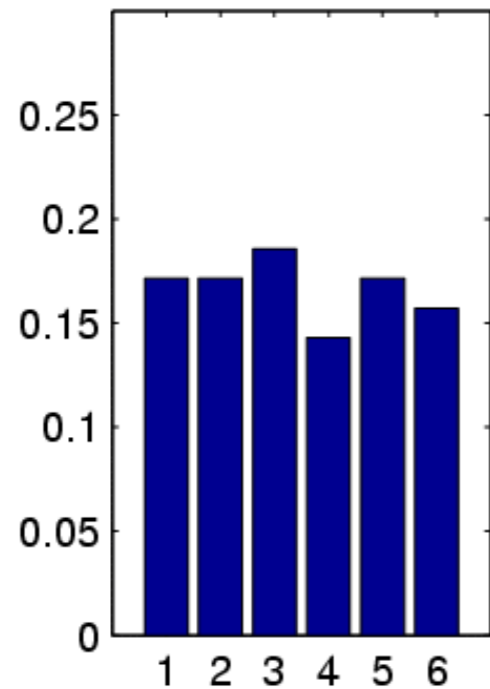


60



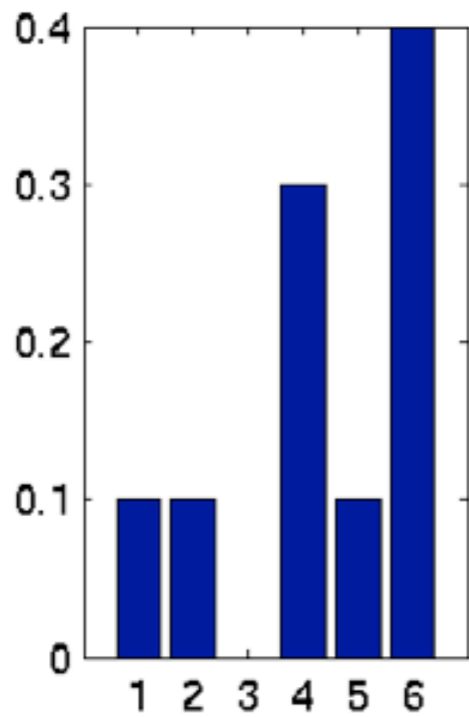
120

MAP

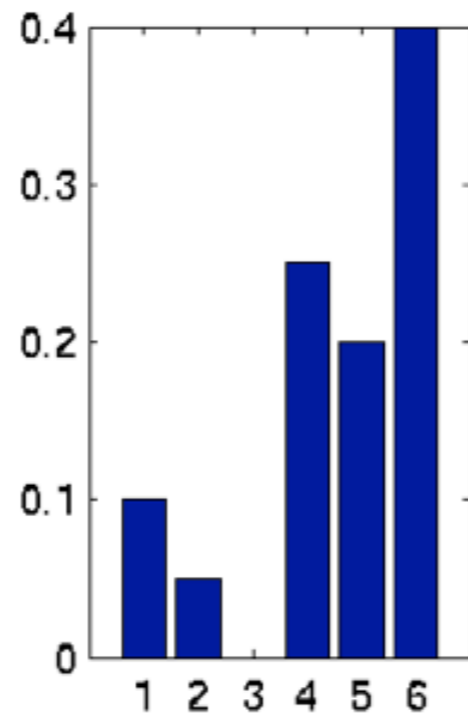


Tainted dice

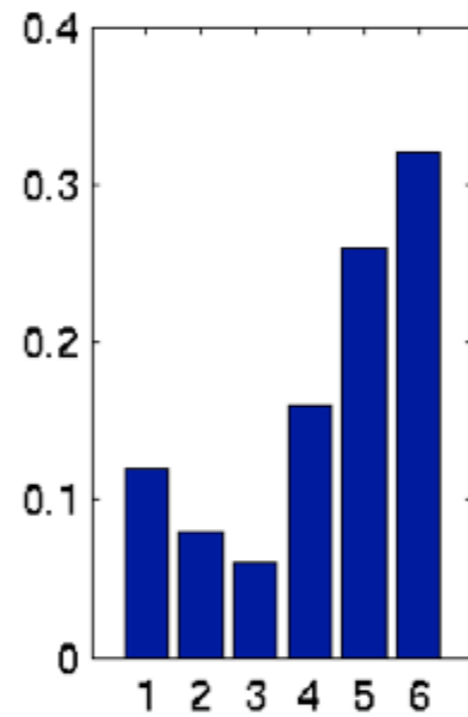
MLE



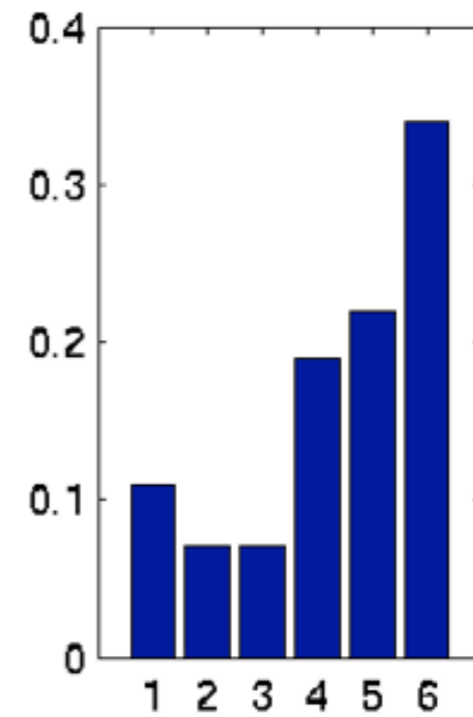
12



24

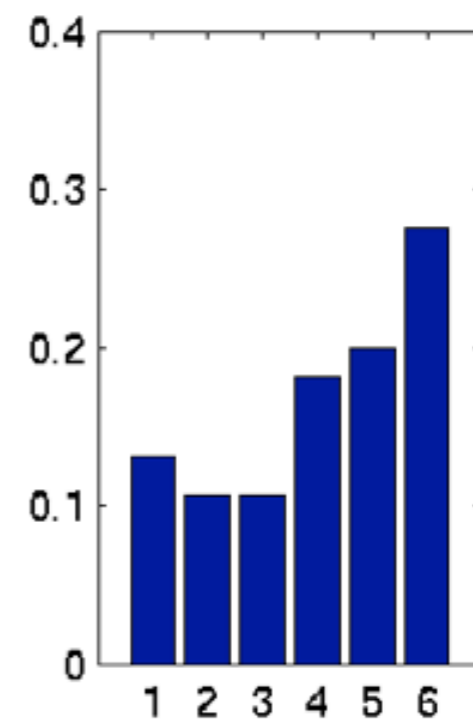
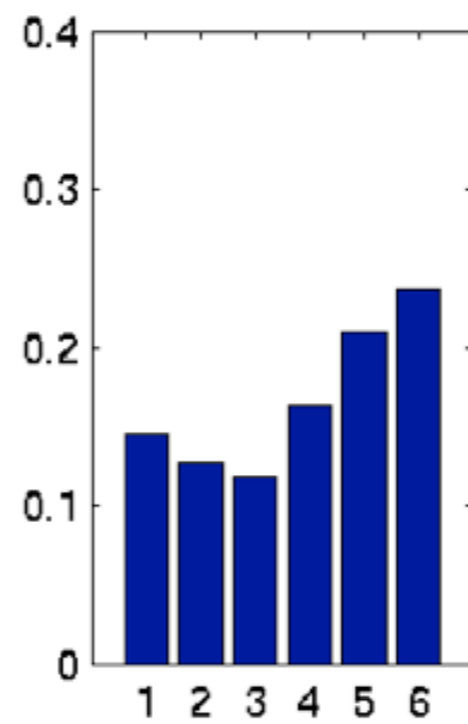
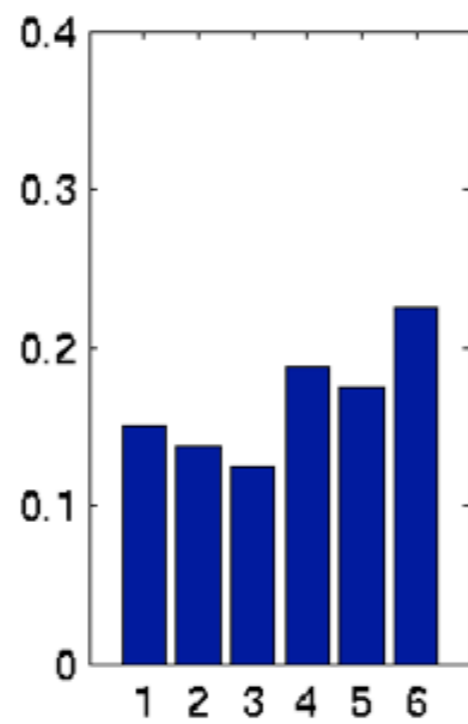
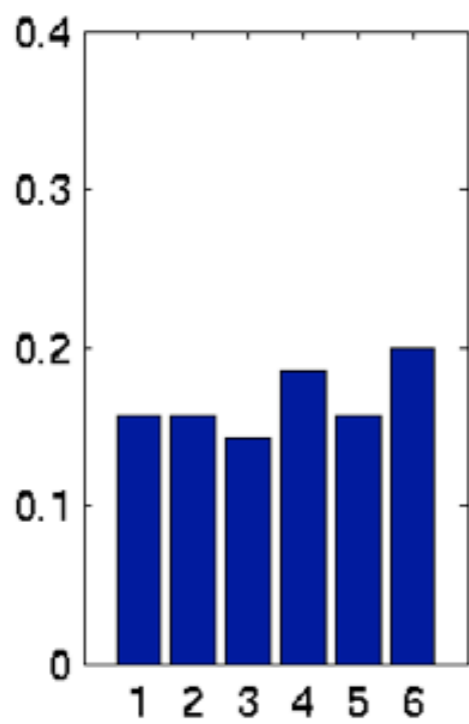


60



120

MAP



Exponential Families



Exponential Families

Exponential Families

- **Density function**

$$p(x; \theta) = \exp (\langle \phi(x), \theta \rangle - g(\theta))$$

where $g(\theta) = \log \sum_{x'} \exp (\langle \phi(x'), \theta \rangle)$

Exponential Families

- **Density function**

$$p(x; \theta) = \exp (\langle \phi(x), \theta \rangle - g(\theta))$$

$$\text{where } g(\theta) = \log \sum_{x'} \exp (\langle \phi(x'), \theta \rangle)$$

- **Log partition function generates cumulants**

$$\partial_{\theta} g(\theta) = \mathbf{E} [\phi(x)]$$

$$\partial_{\theta}^2 g(\theta) = \text{Var} [\phi(x)]$$

Exponential Families

- **Density function**

$$p(x; \theta) = \exp (\langle \phi(x), \theta \rangle - g(\theta))$$

$$\text{where } g(\theta) = \log \sum_{x'} \exp (\langle \phi(x'), \theta \rangle)$$

- **Log partition function generates cumulants**

$$\partial_{\theta} g(\theta) = \mathbf{E} [\phi(x)]$$

$$\partial_{\theta}^2 g(\theta) = \text{Var} [\phi(x)]$$

- **g is convex (second derivative is p.s.d.)**

Examples

- **Binomial Distribution**

$$\phi(x) = x$$

- **Discrete Distribution**

$$\phi(x) = e_x$$

(e_x is unit vector for x)

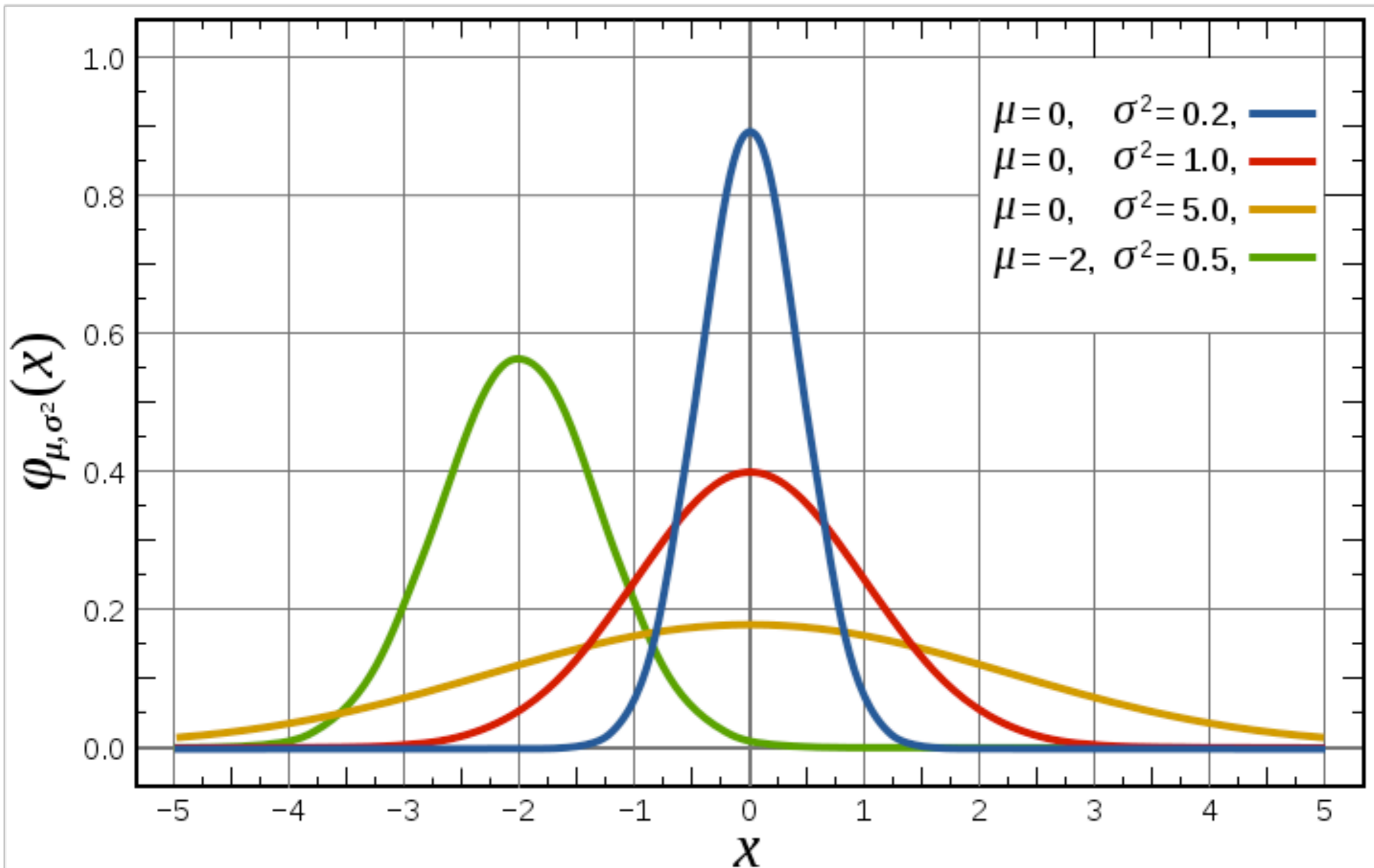
- **Gaussian**

$$\phi(x) = \left(x, \frac{1}{2} x x^\top \right)$$

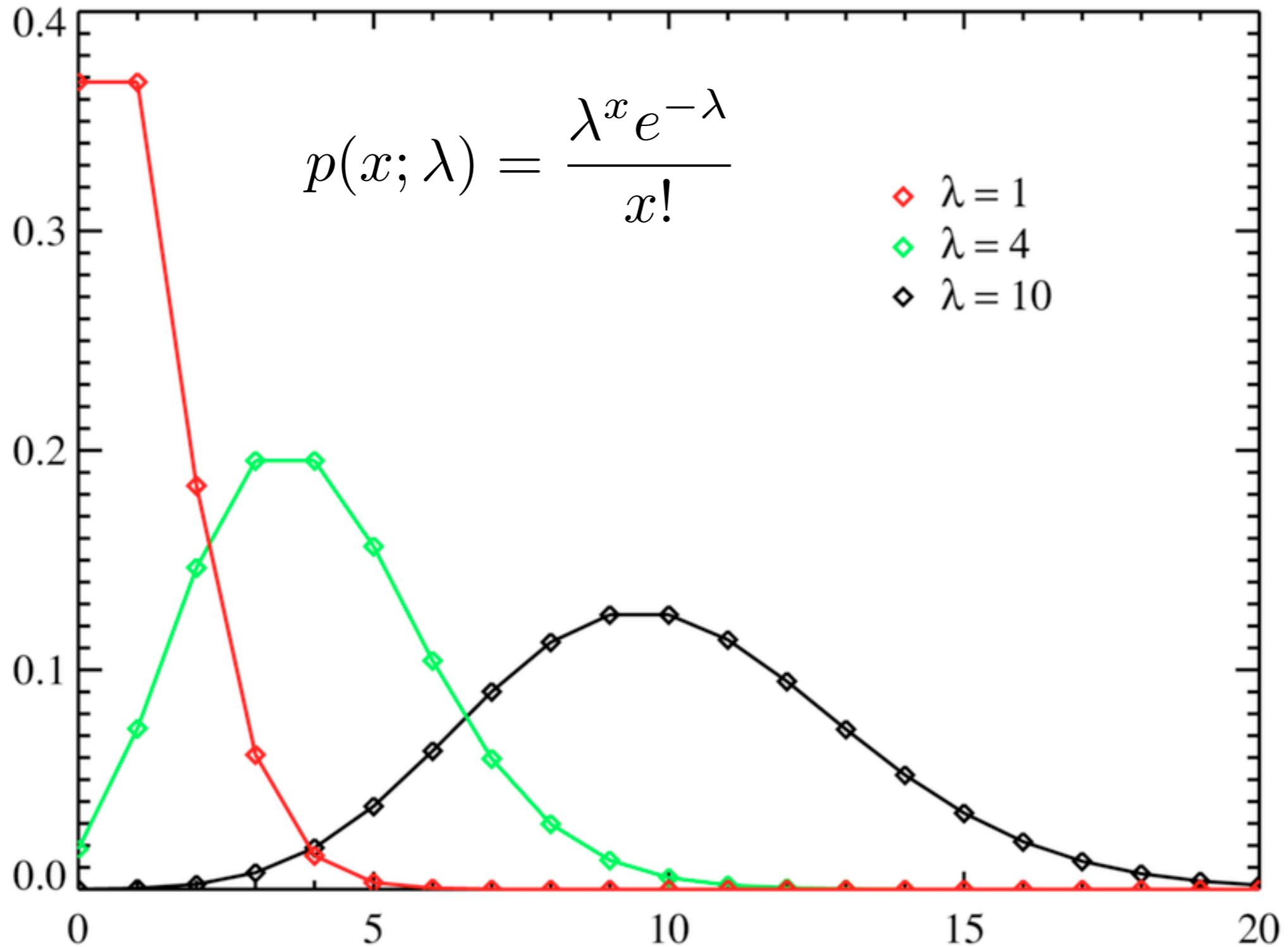
- **Poisson (counting measure $1/x!$)** $\phi(x) = x$

- **Dirichlet, Beta, Gamma, Wishart, ...**

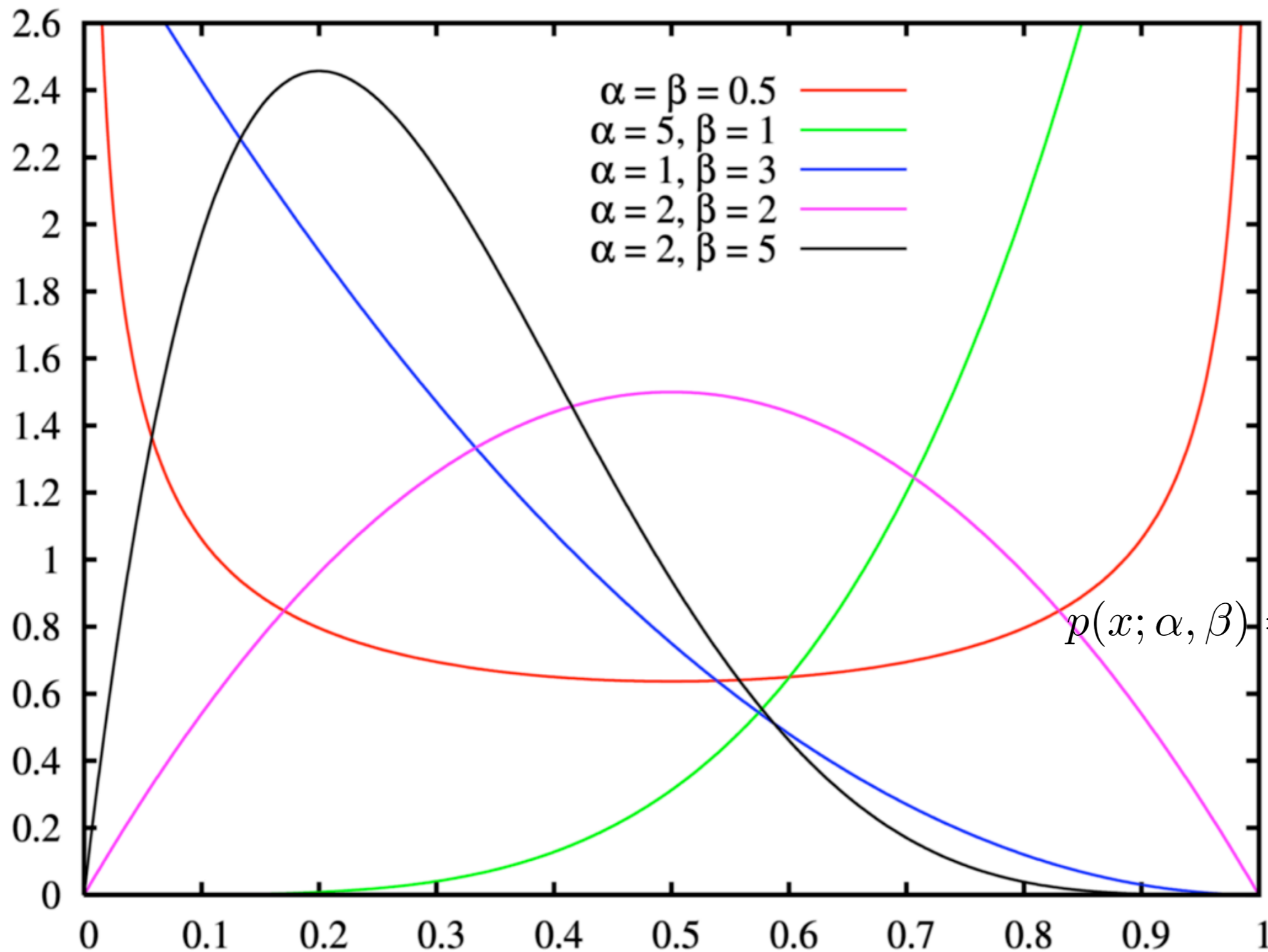
Normal Distribution



Poisson Distribution

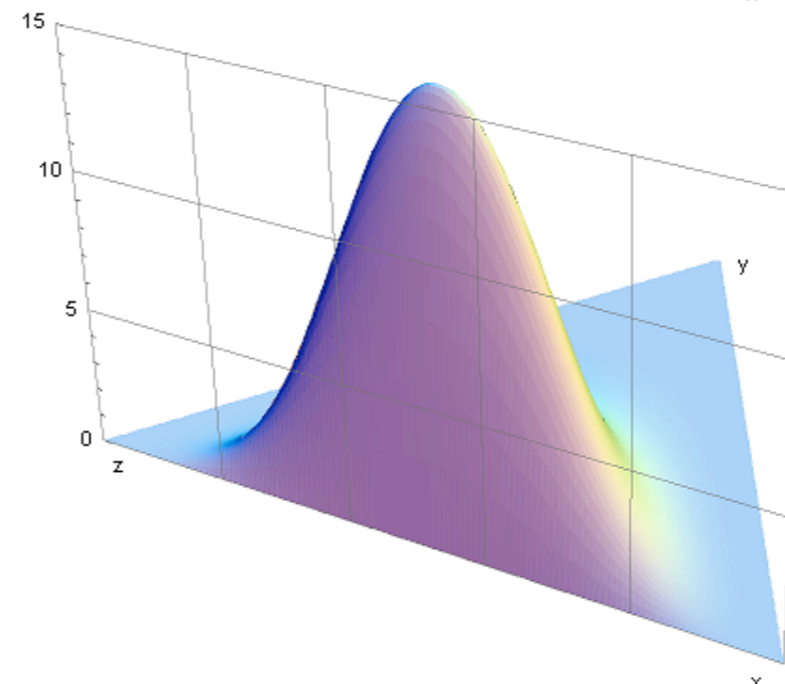
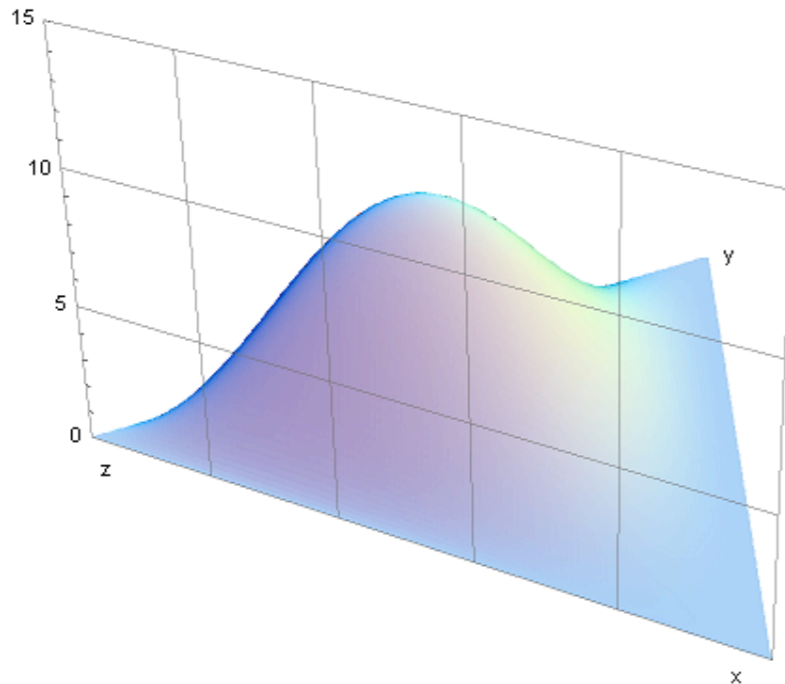
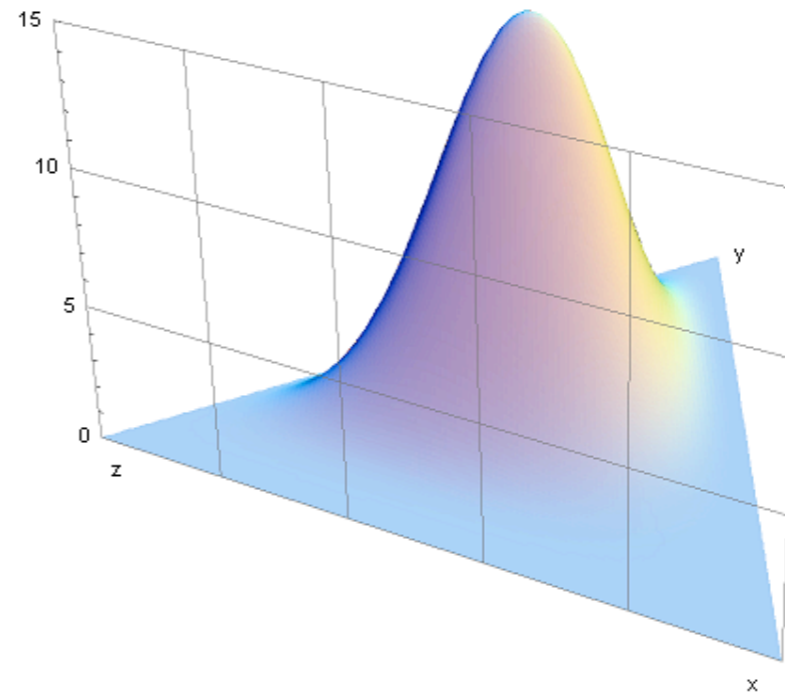
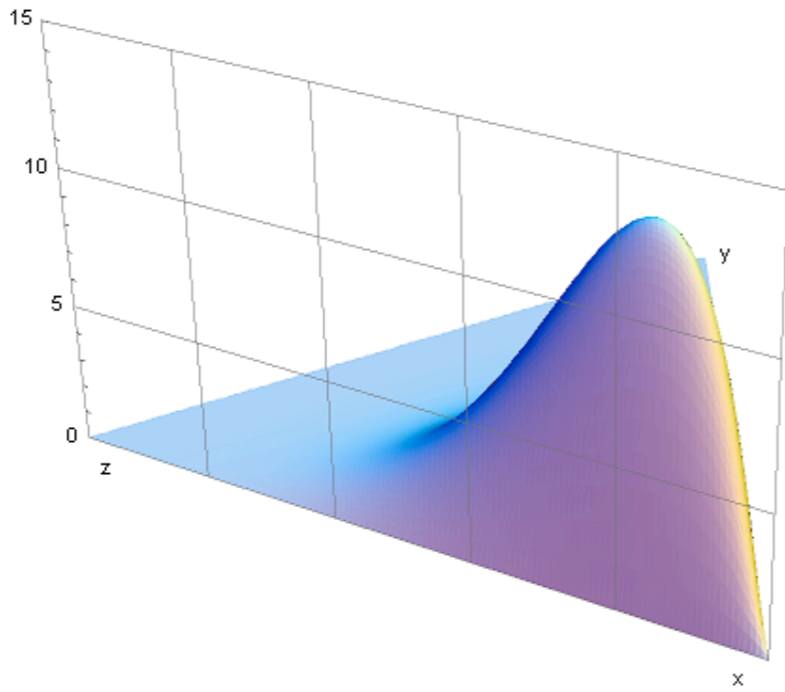


Beta Distribution



$$p(x; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}$$

Dirichlet Distribution



... this is a distribution over distributions ...

Maximum Likelihood

Maximum Likelihood

- **Negative log-likelihood**

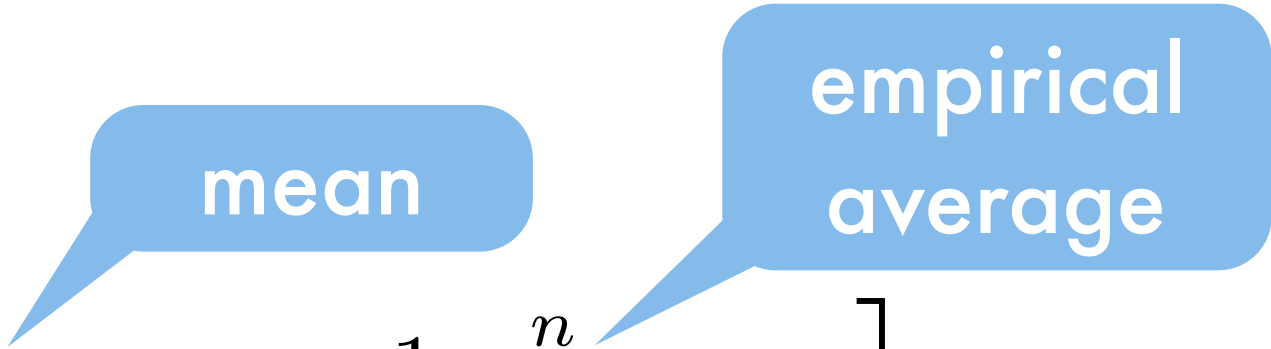
$$-\log p(X; \theta) = \sum_{i=1}^n g(\theta) - \langle \phi(x_i), \theta \rangle$$

Maximum Likelihood

- Negative log-likelihood

$$-\log p(X; \theta) = \sum_{i=1}^n g(\theta) - \langle \phi(x_i), \theta \rangle$$

- Taking derivatives

$$-\partial_{\theta} \log p(X; \theta) = m \left[\mathbf{E}[\phi(x)] - \frac{1}{m} \sum_{i=1}^n \phi(x_i) \right]$$


We pick the parameter such that the distribution matches the empirical average.

Example: Gaussian Estimation

- Sufficient statistics: x, x^2

- Mean and variance given by

$$\mu = \mathbf{E}_x[x] \text{ and } \sigma^2 = \mathbf{E}_x[x^2] - \mathbf{E}_x^2[x]$$

- Maximum Likelihood Estimate

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \hat{\mu}^2$$

- Maximum a Posteriori Estimate

$$\hat{\mu} = \frac{1}{n + n_0} \sum_{i=1}^n x_i \text{ and } \sigma^2 = \frac{1}{n + n_0} \sum_{i=1}^n x_i^2 + \frac{n_0}{n + n_0} \mathbf{1} - \hat{\mu}^2$$

smoother

smoother

Collapsing

- Conjugate priors

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

Hence we know how to compute normalization

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

(Beta, binomial)

(Dirichlet, multinomial)

(Gamma, Poisson)

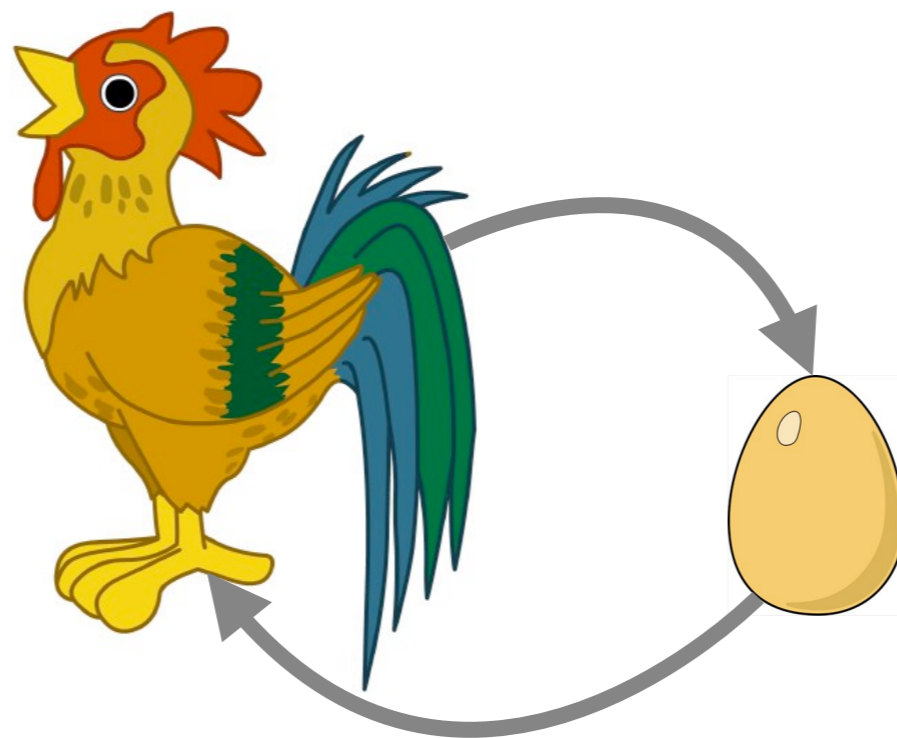
(Wishart, Gauss)

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

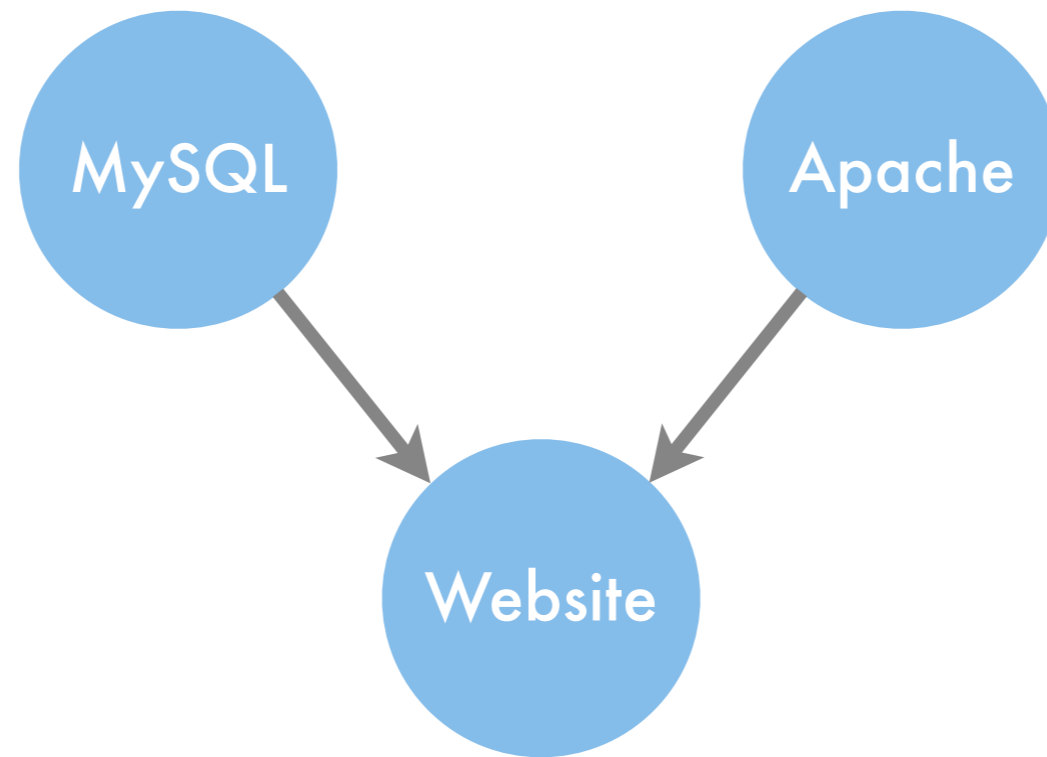
$$= \int p(\{x\} \cup X \cup X_{\text{fake}}|\theta)d\theta$$

look up closed
form expansions

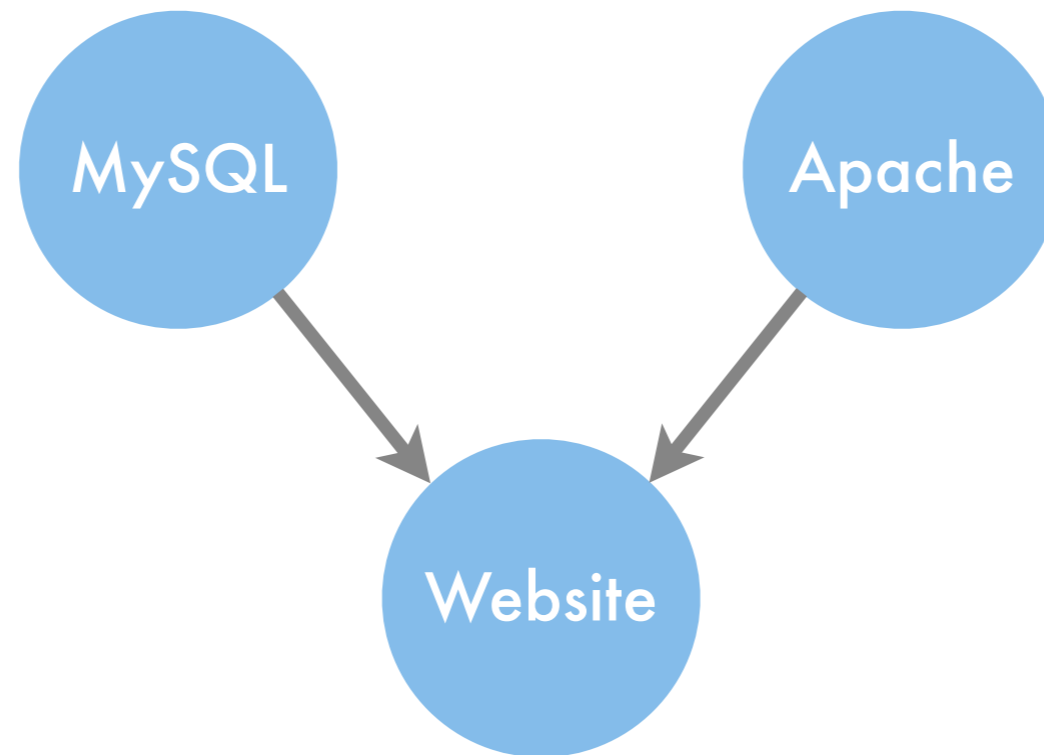
Directed Graphical Models



... some Web 2.0 service



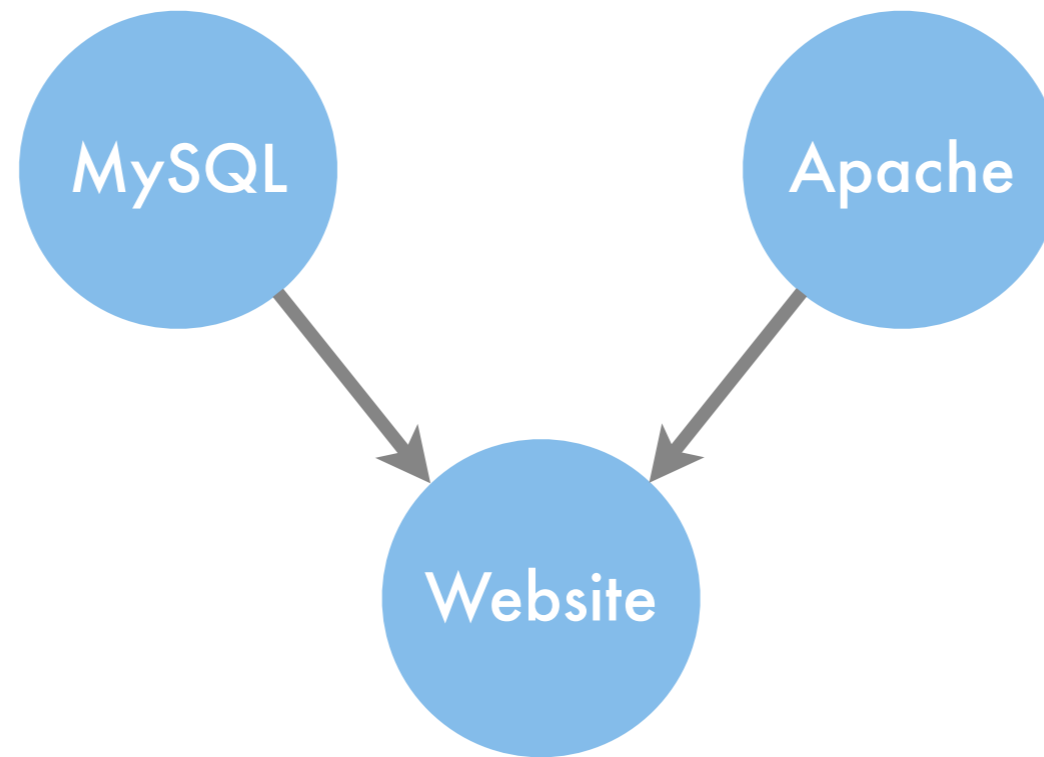
... some Web 2.0 service



- **Joint distribution (assume a and m are independent)**

$$p(m, a, w) = p(w|m, a)p(m)p(a)$$

... some Web 2.0 service



- **Joint distribution (assume a and m are independent)**

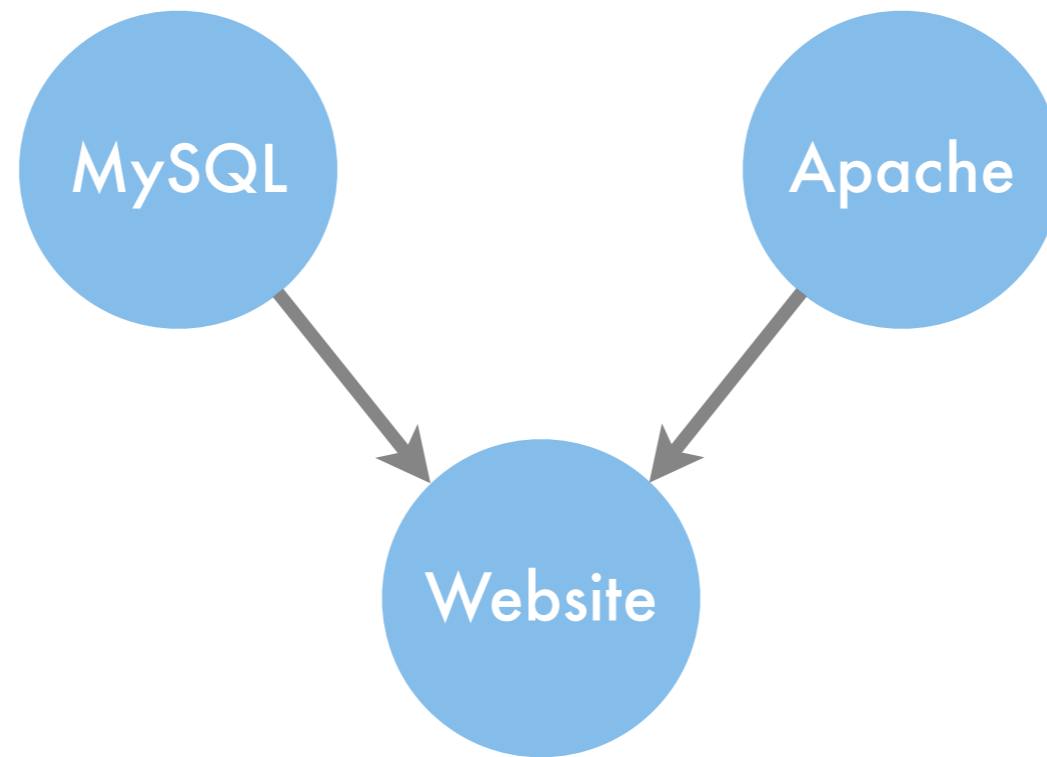
$$p(m, a, w) = p(w|m, a)p(m)p(a)$$

- **Explaining away**

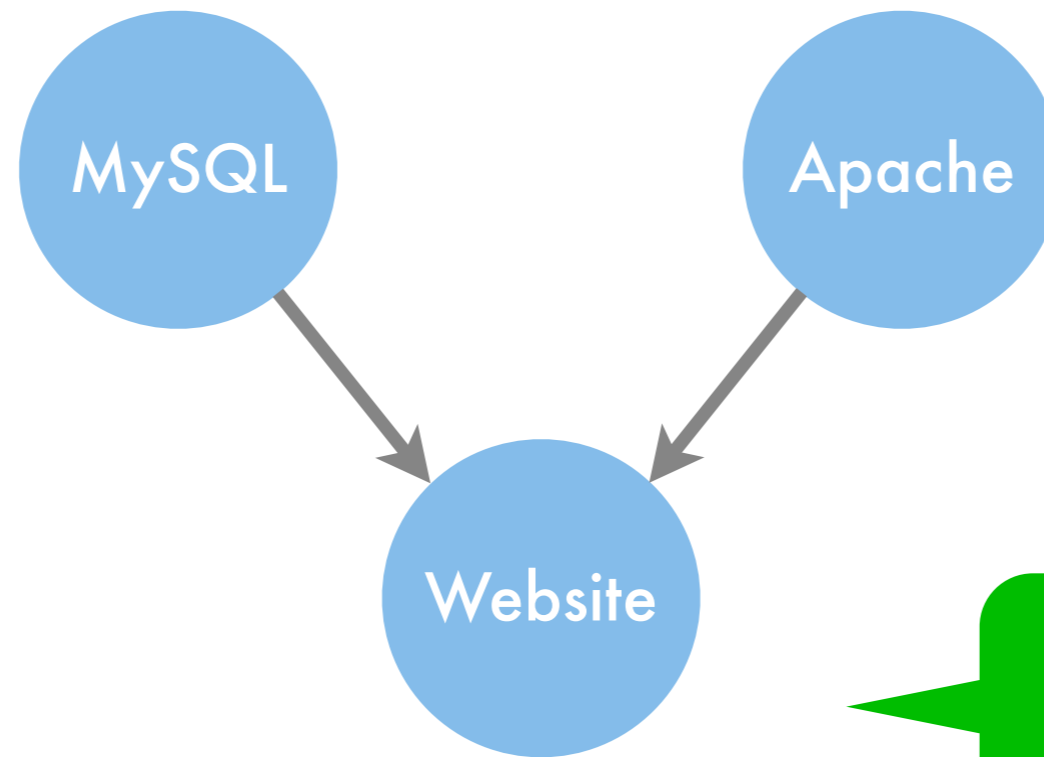
$$p(m, a|w) = \frac{p(w|m, a)p(m)p(a)}{\sum_{m', a'} p(w|m', a')p(m')p(a')}$$

a and m are dependent conditioned on w

... some Web 2.0 service



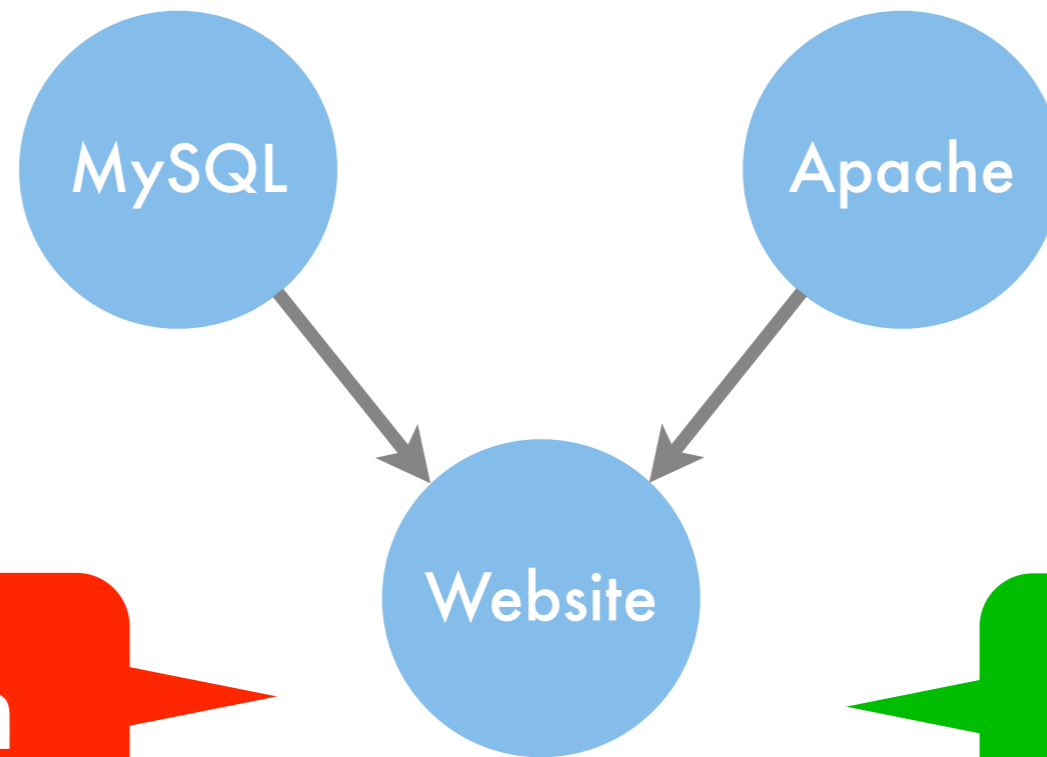
... some Web 2.0 service



is working

MySQL is working
Apache is working

... some Web 2.0 service



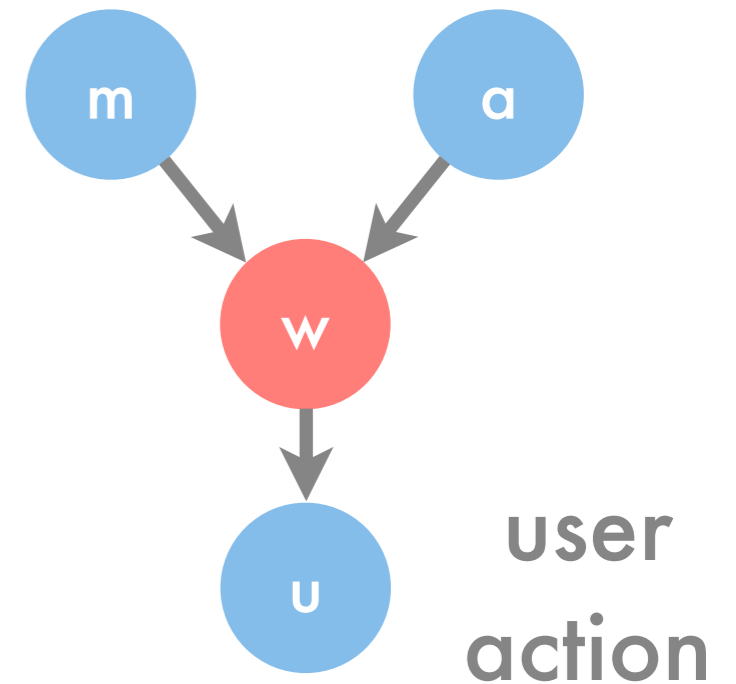
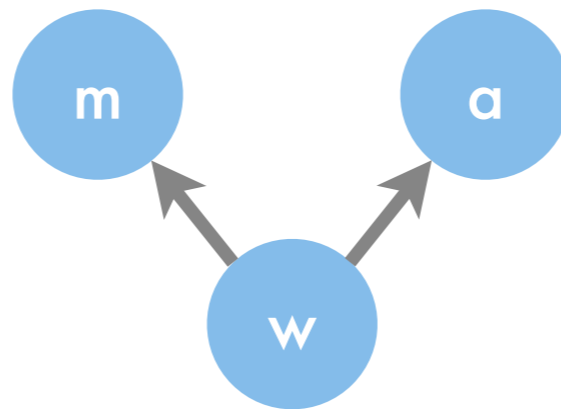
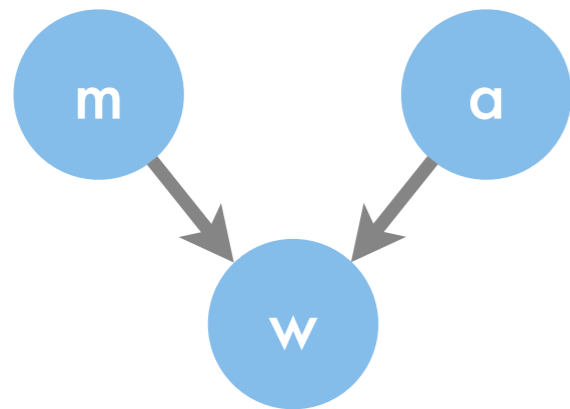
is broken

is working

At least one of the
two services is broken
(not independent)

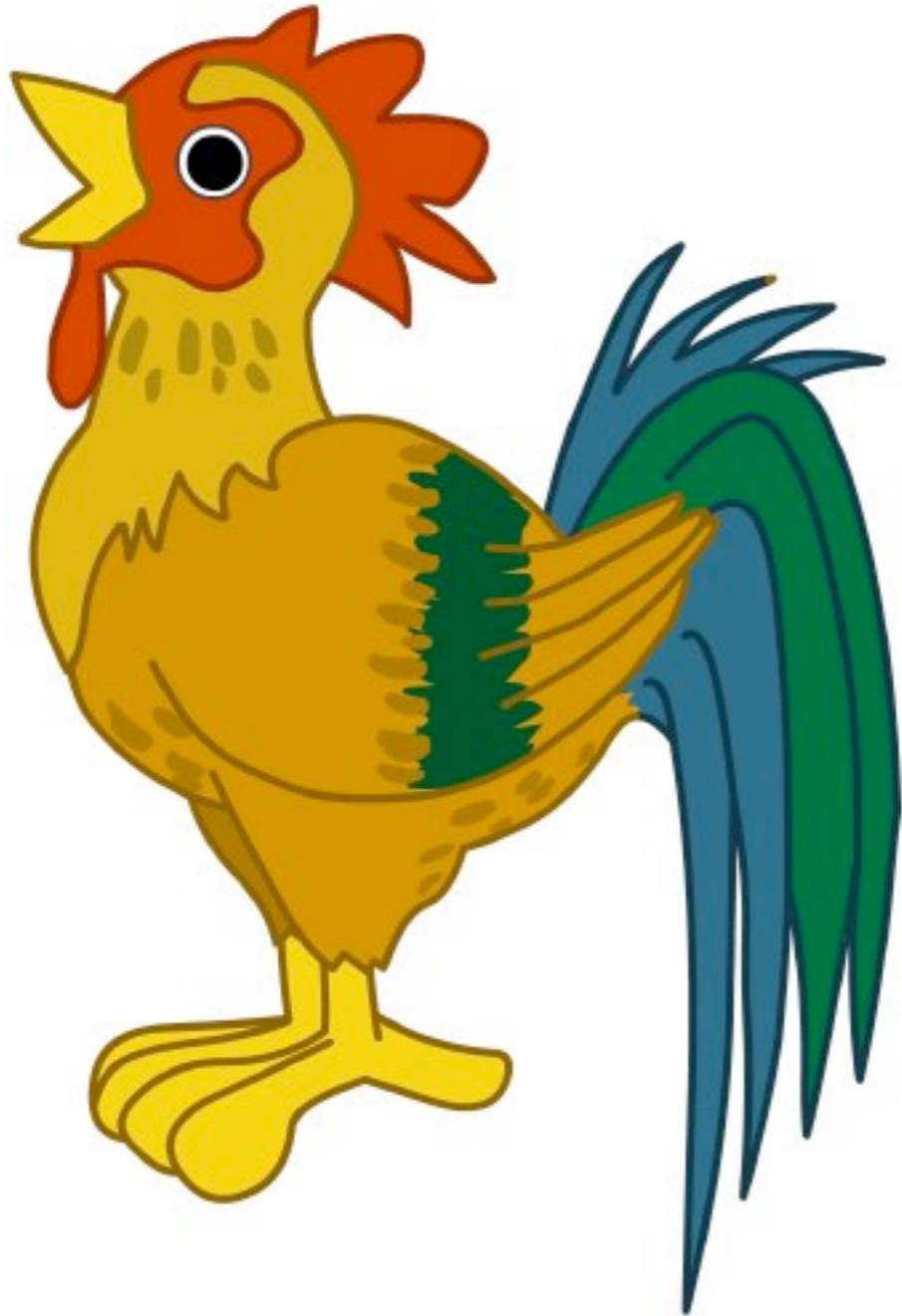
MySQL is working
Apache is working

Directed graphical model

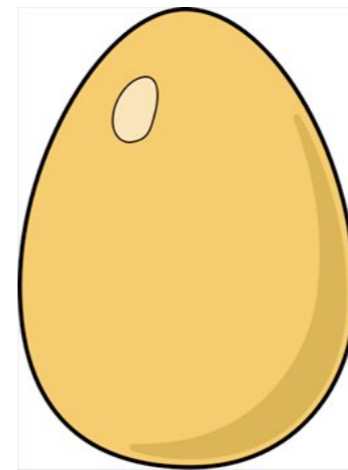
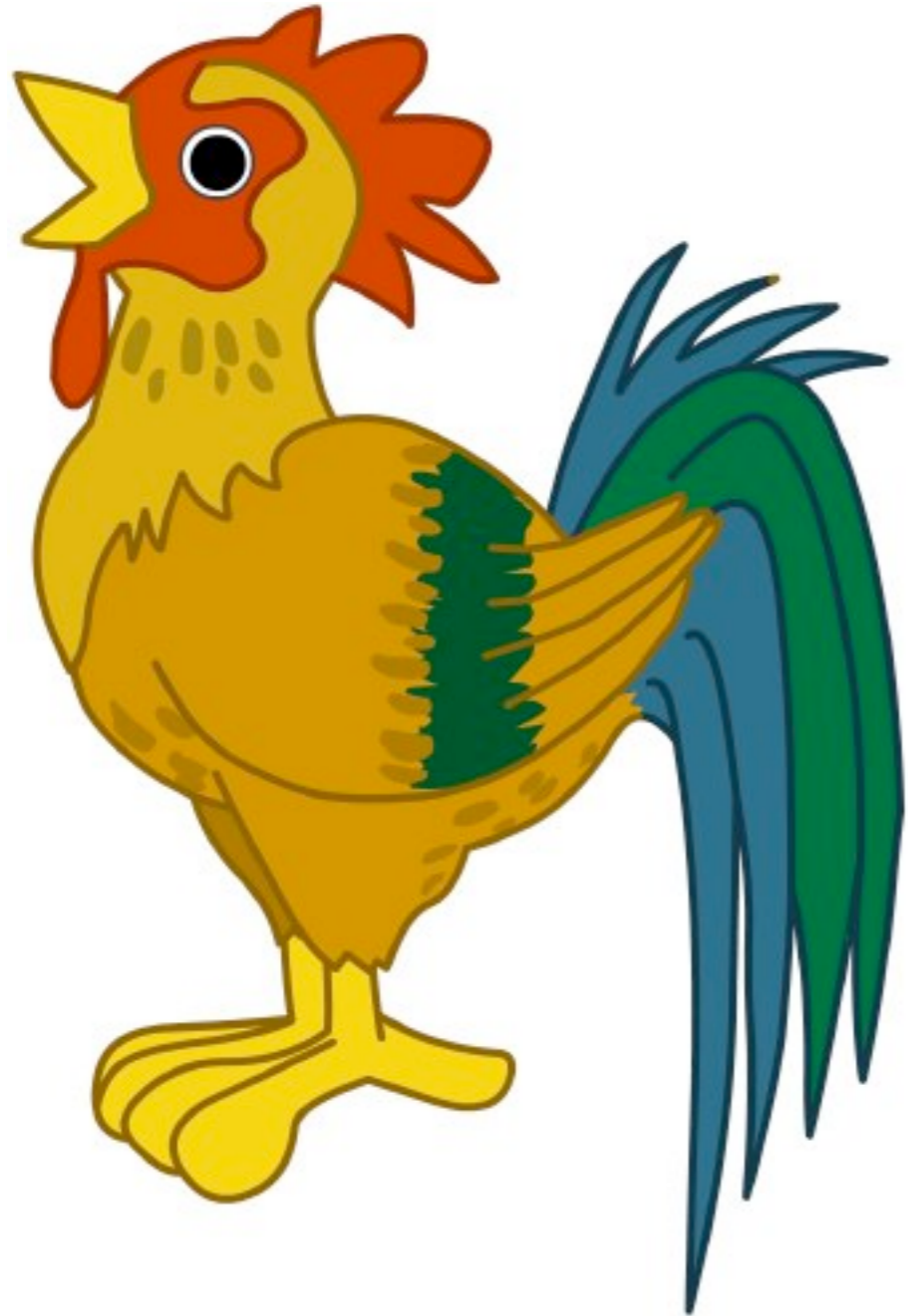


- Easier estimation
 - 15 parameters for full joint distribution
 - $1+1+3+1$ for factorizing distribution
- Causal relations
- Inference for unobserved variables

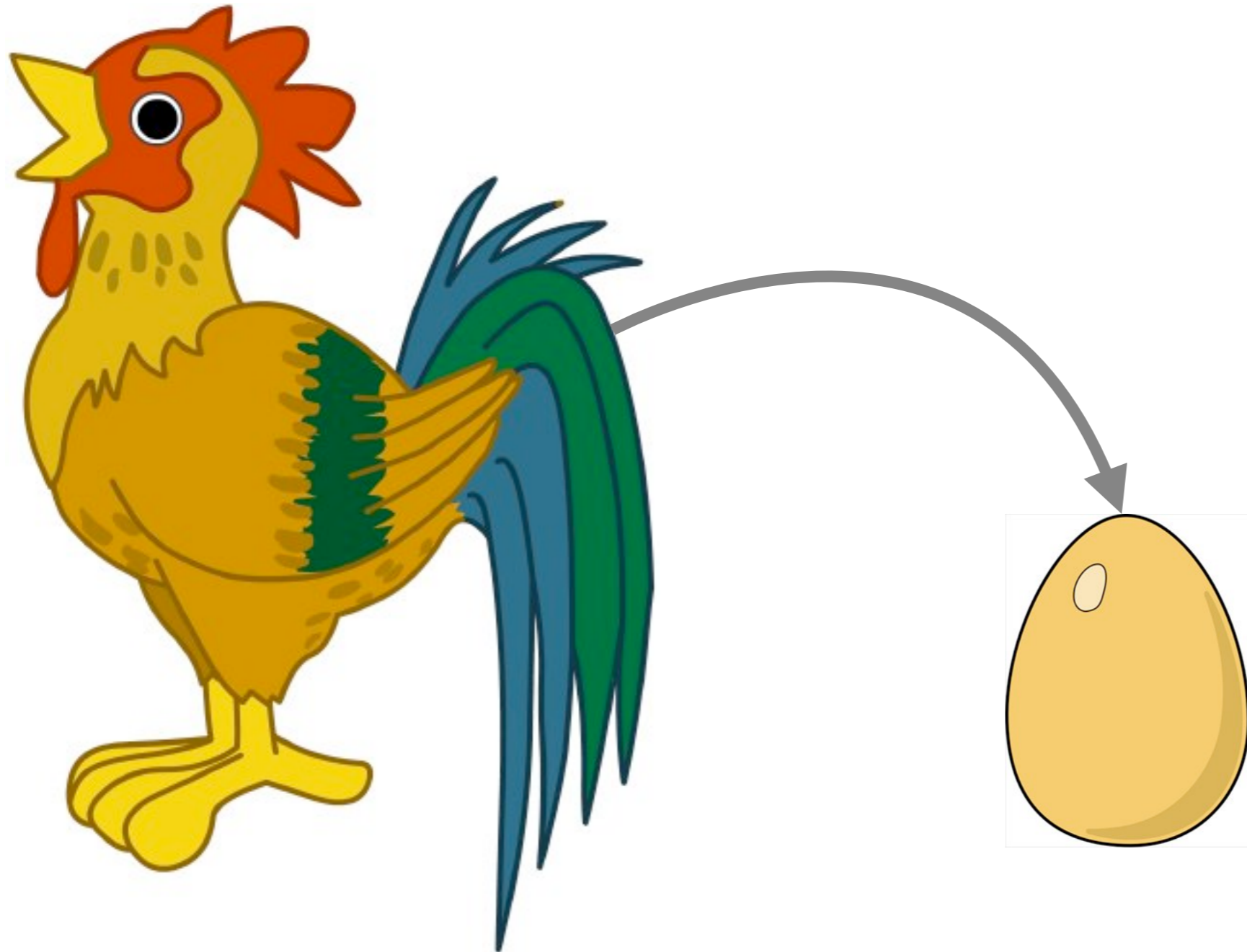
No loops allowed



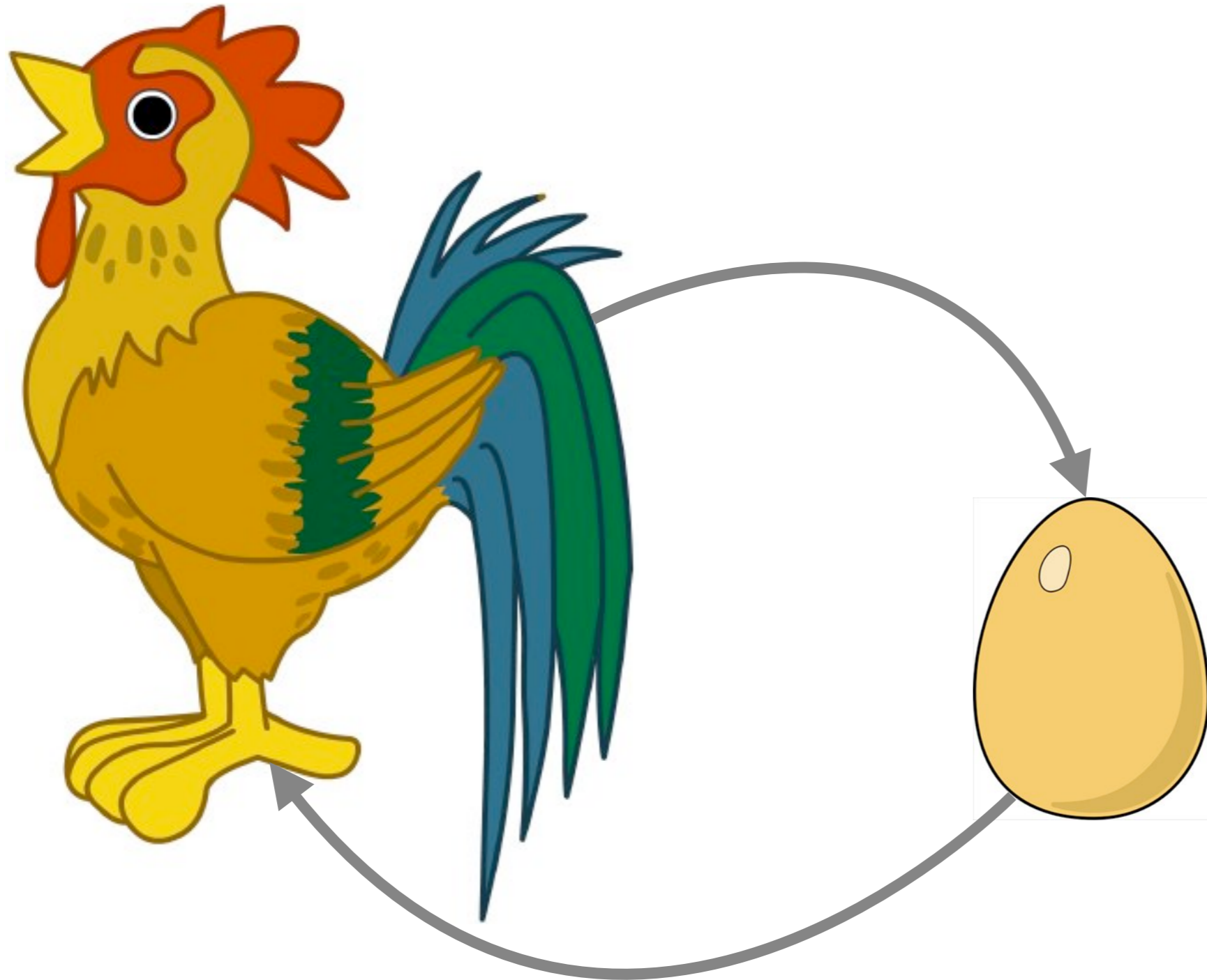
No loops allowed



No loops allowed

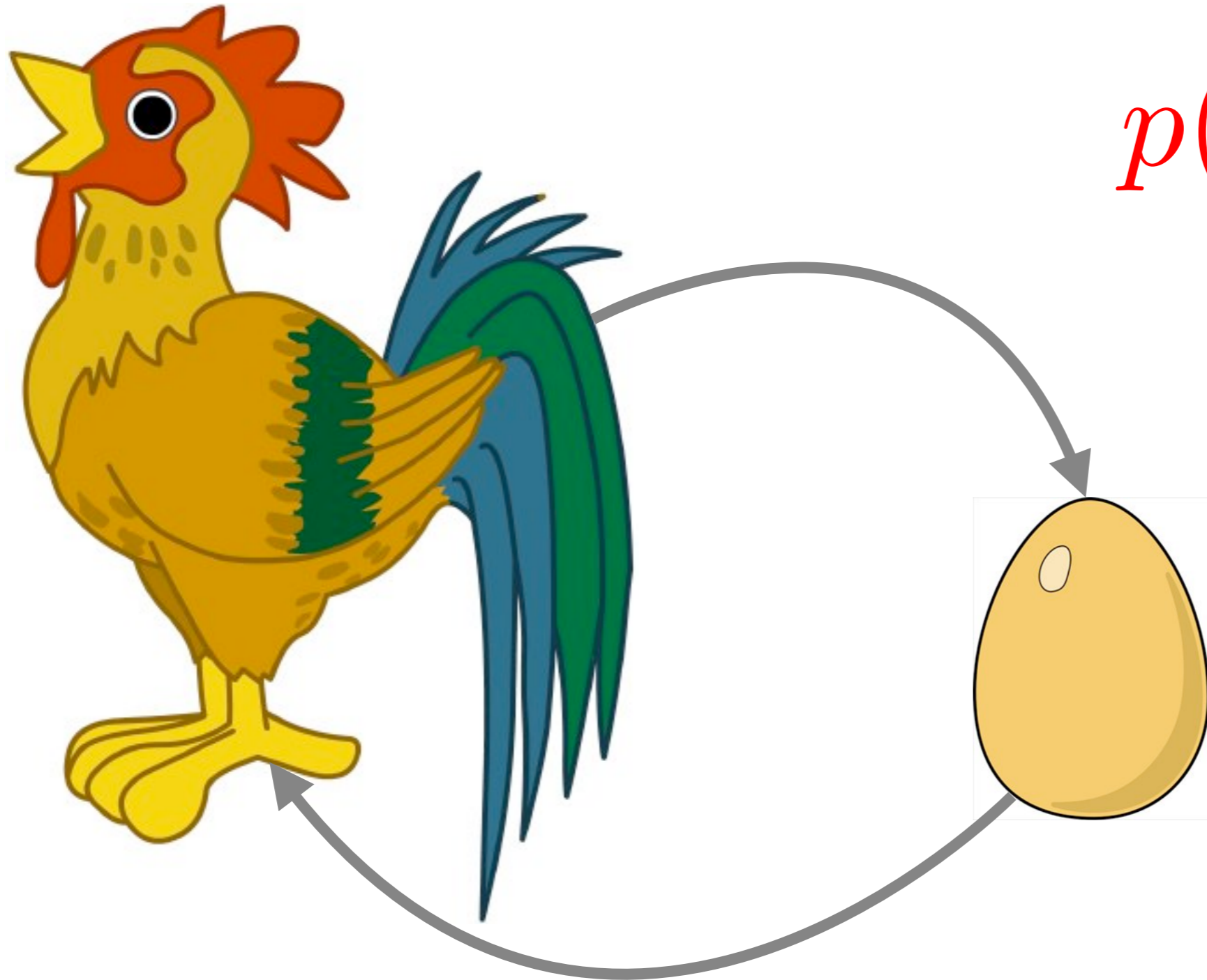


No loops allowed



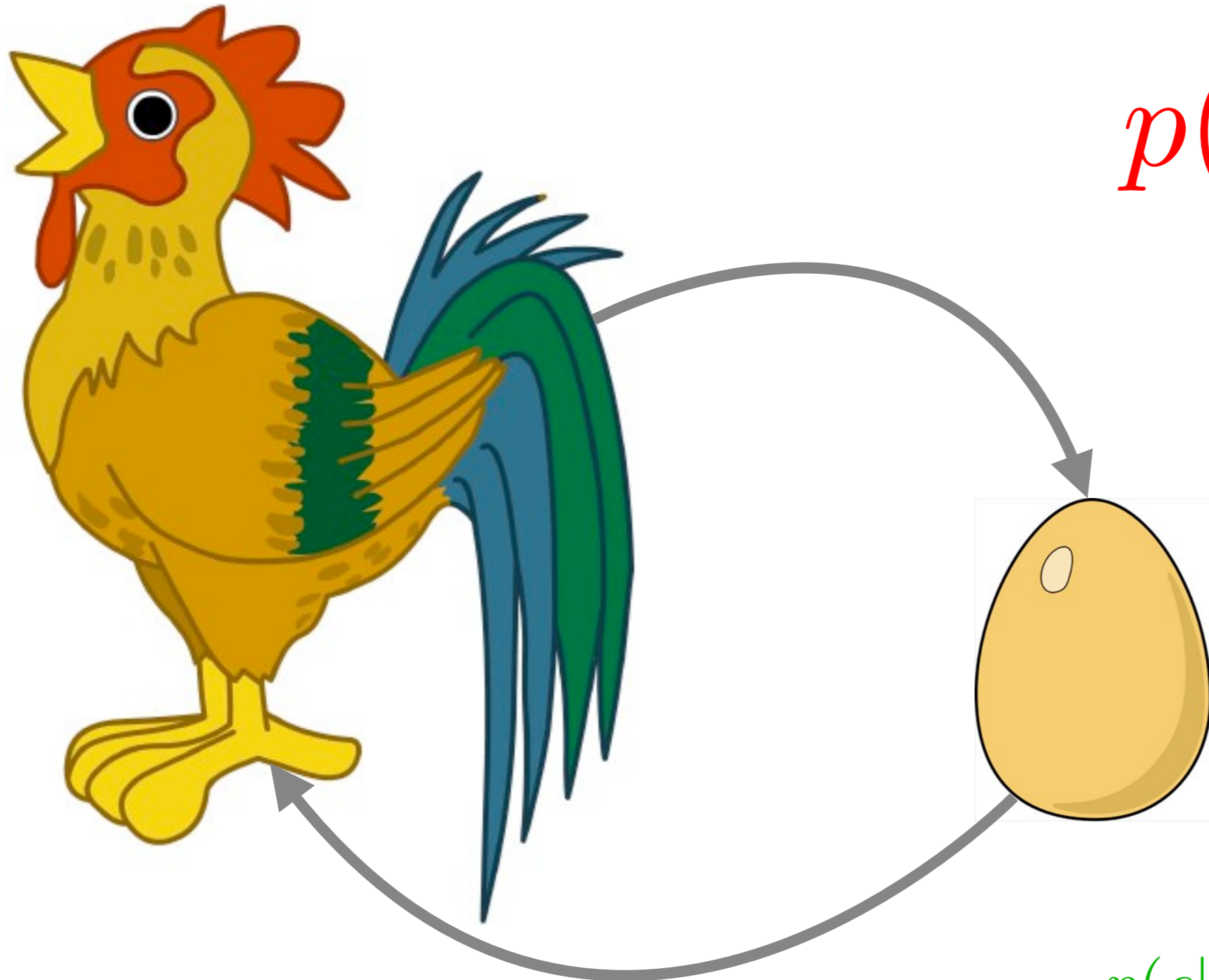
No loops allowed

$$p(c|e)p(e|c)$$



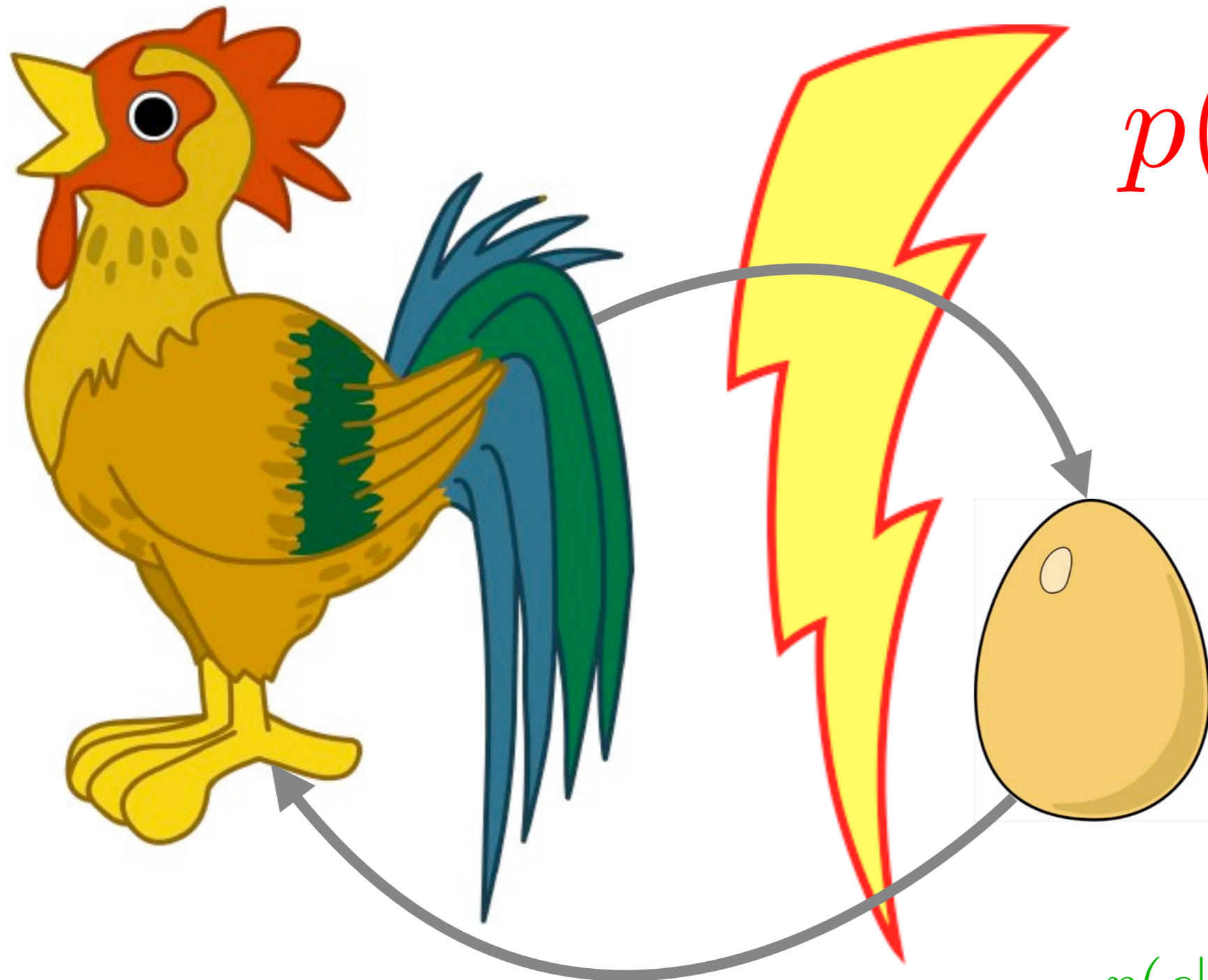
No loops allowed

$$p(c|e)p(e|c)$$



$$p(c|e)p(e) \text{ or } p(e|c)p(c)$$

No loops allowed



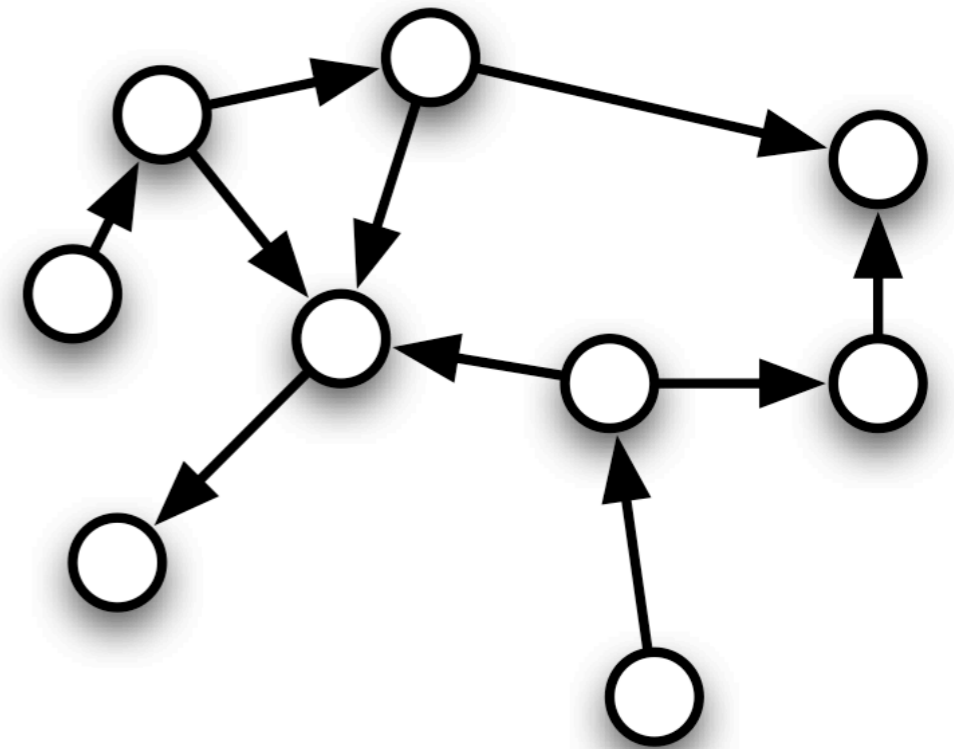
$$p(c|e)p(e|c)$$

$$p(c|e)p(e) \text{ or } p(e|c)p(c)$$

Directed Graphical Model

- Joint probability distribution

$$p(x) = \prod_i p(x_i | x_{\text{parents}(i)})$$



- Parameter estimation

- If x is fully observed the likelihood breaks up

$$\log p(x|\theta) = \sum_i \log p(x_i | x_{\text{parents}(i)}, \theta)$$

- If x is partially observed things get interesting
maximization, EM, variational, sampling ...

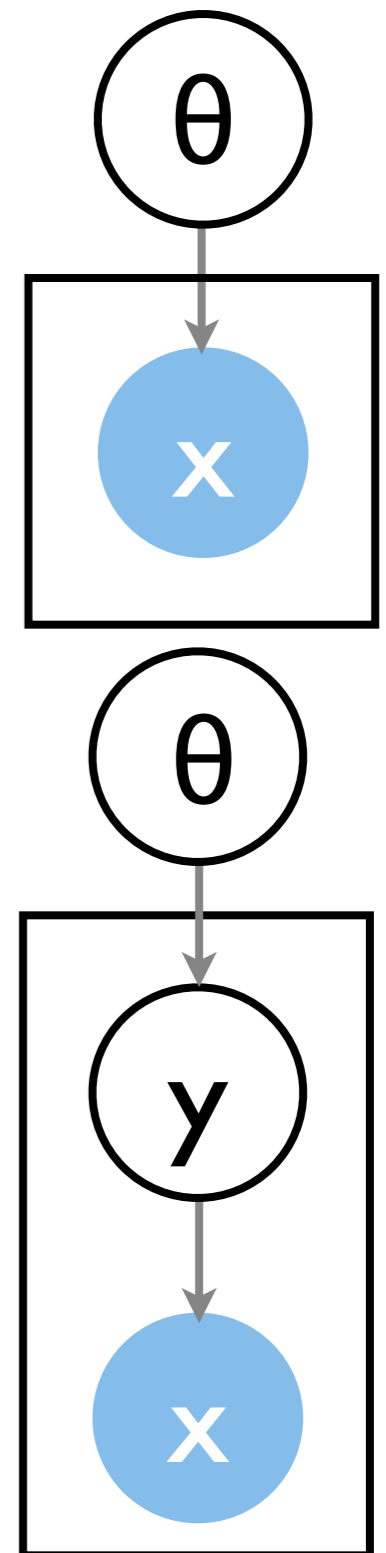
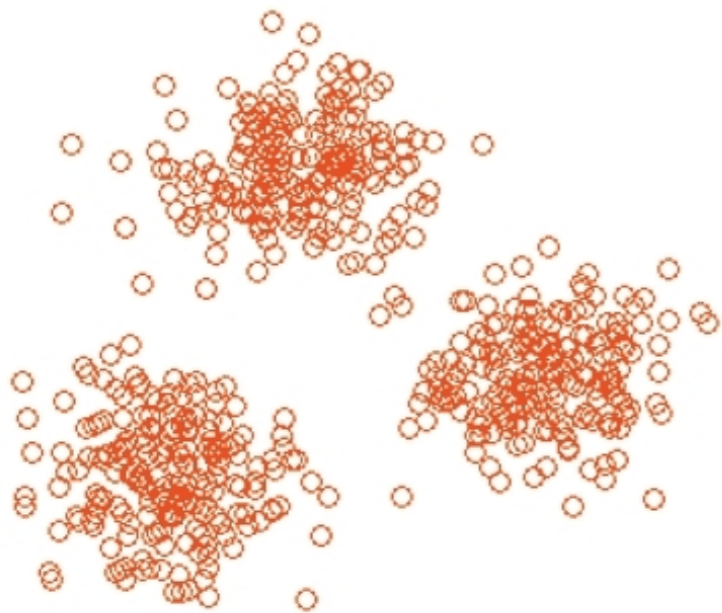
Clustering

Density Estimation

$$p(x, \theta) = p(\theta) \prod_{i=1}^n p(x_i | \theta)$$

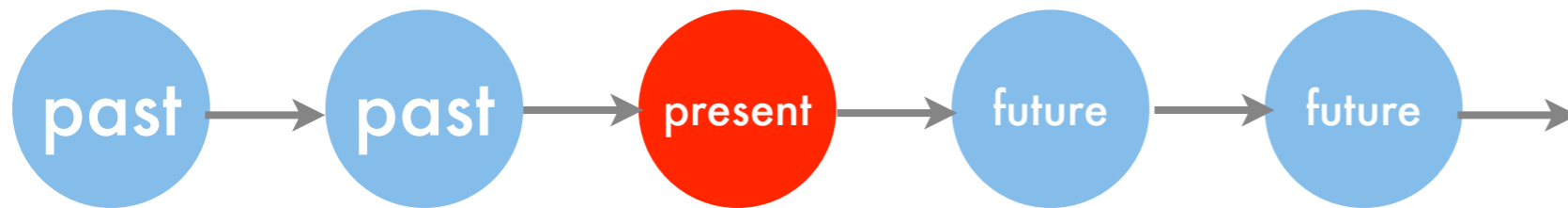
Clustering

$$p(x, y, \theta) = p(\pi) \prod_{k=1}^K p(\theta_k) \prod_{i=1}^n p(y_i | \pi) p(x_i | \theta, y_i)$$



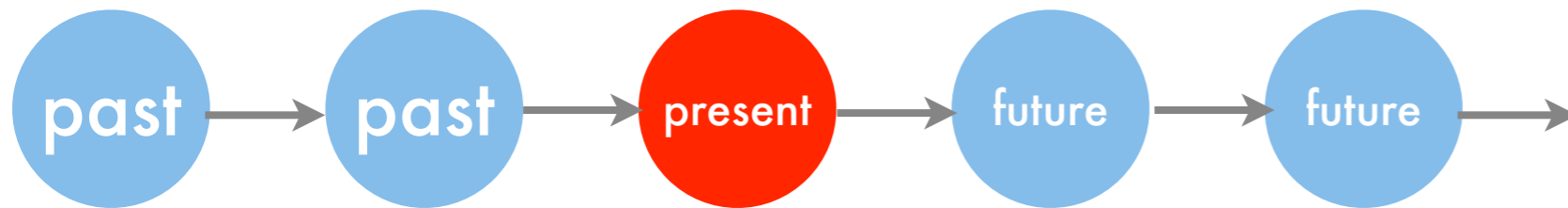
Chains

Markov Chain

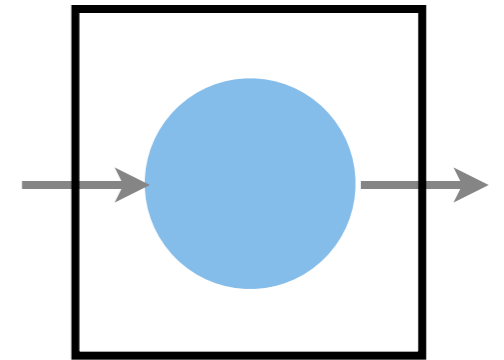


Chains

Markov Chain

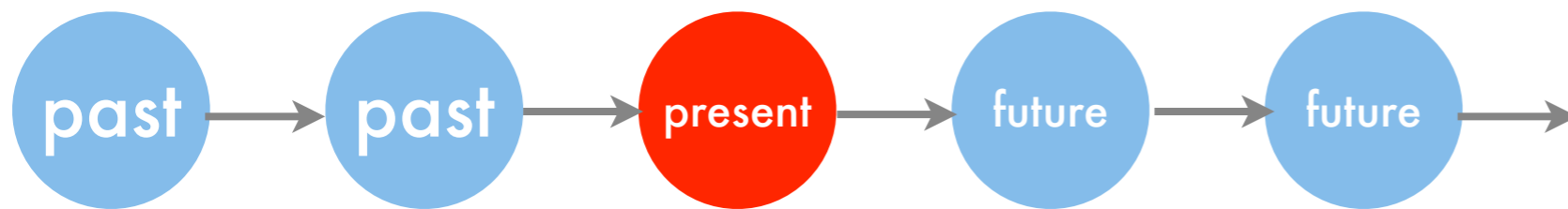


Plate

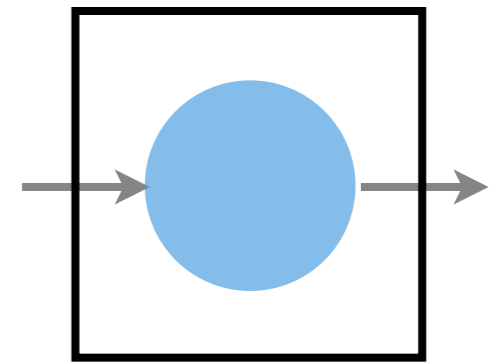


Chains

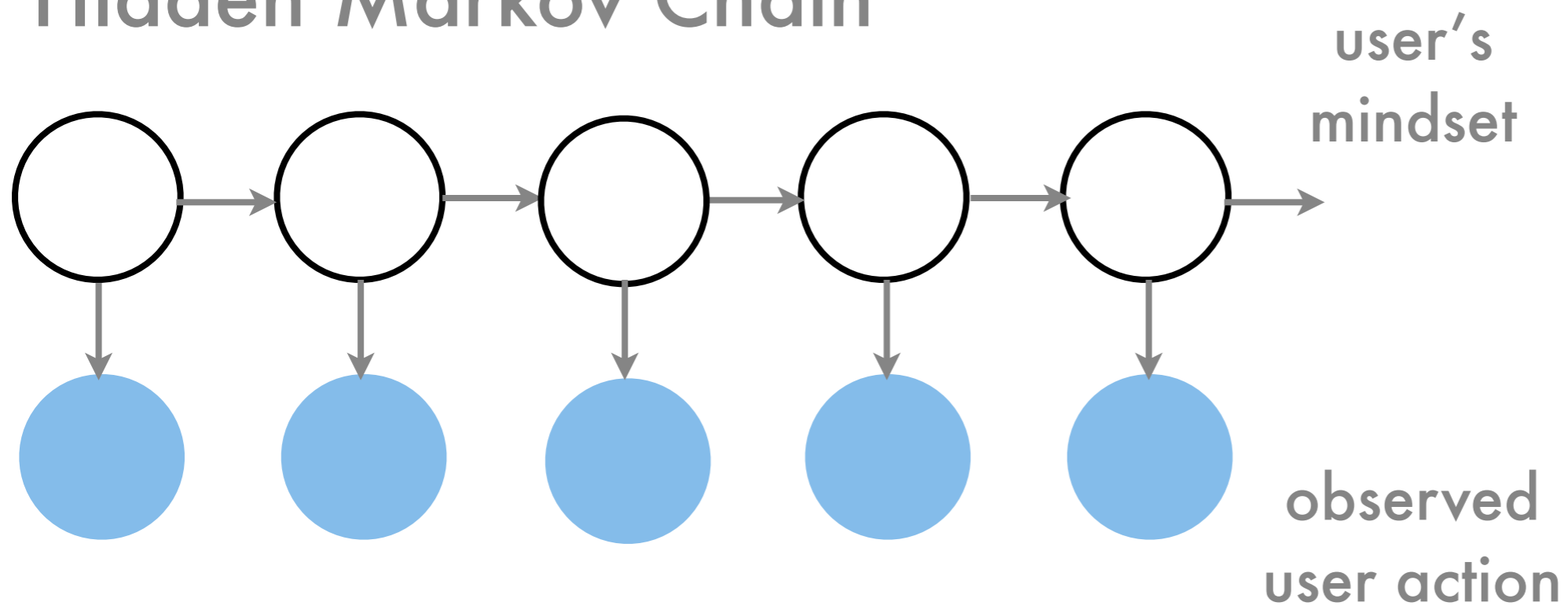
Markov Chain



Plate

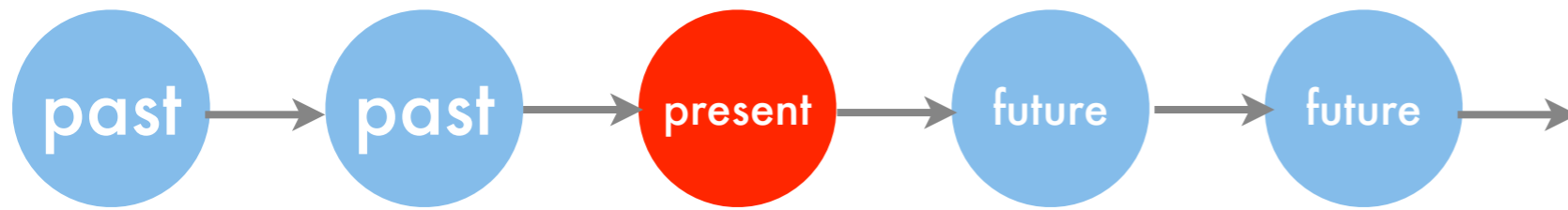


Hidden Markov Chain

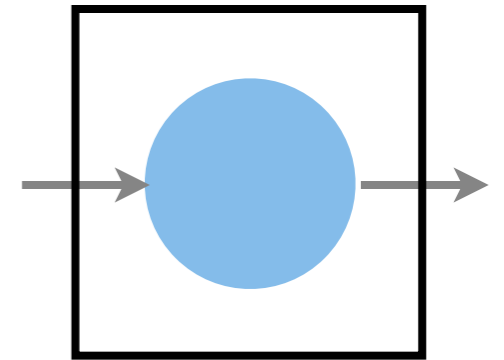


Chains

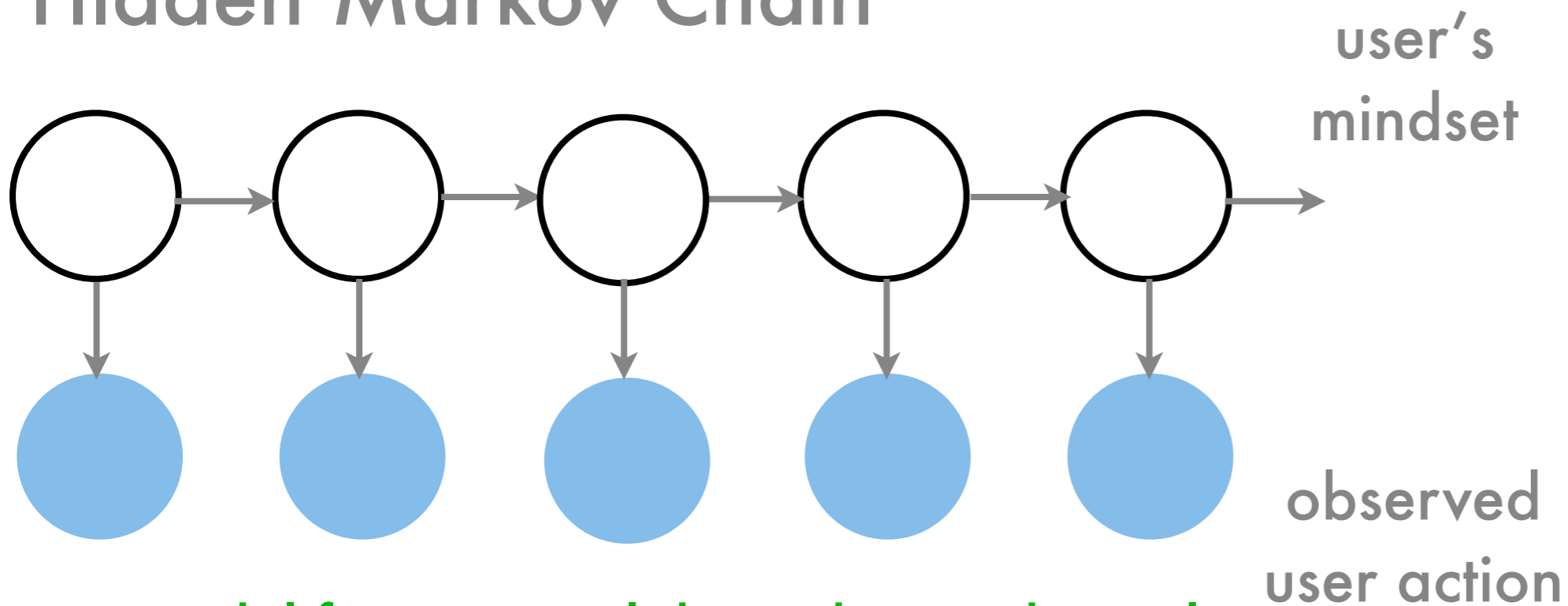
Markov Chain



Plate



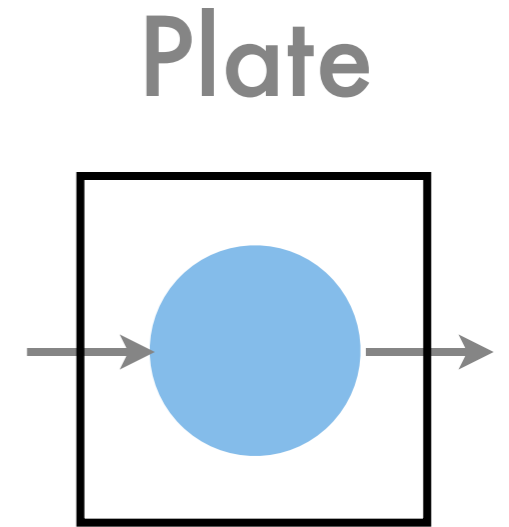
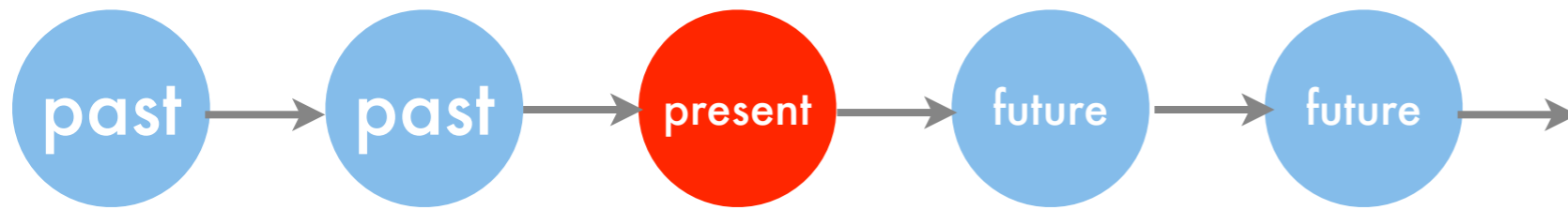
Hidden Markov Chain



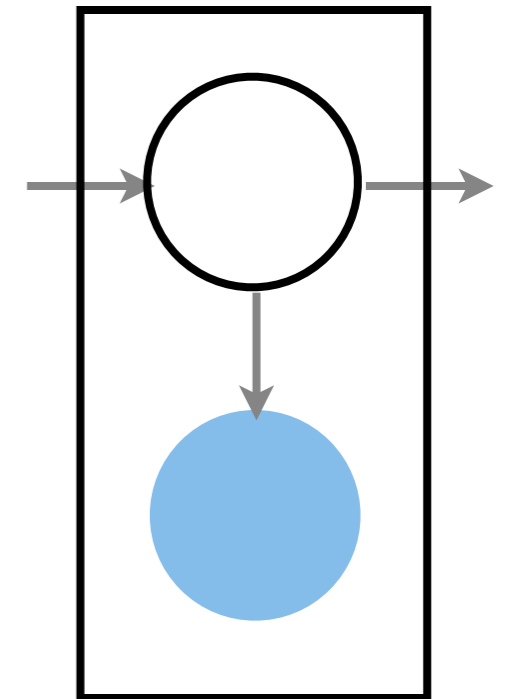
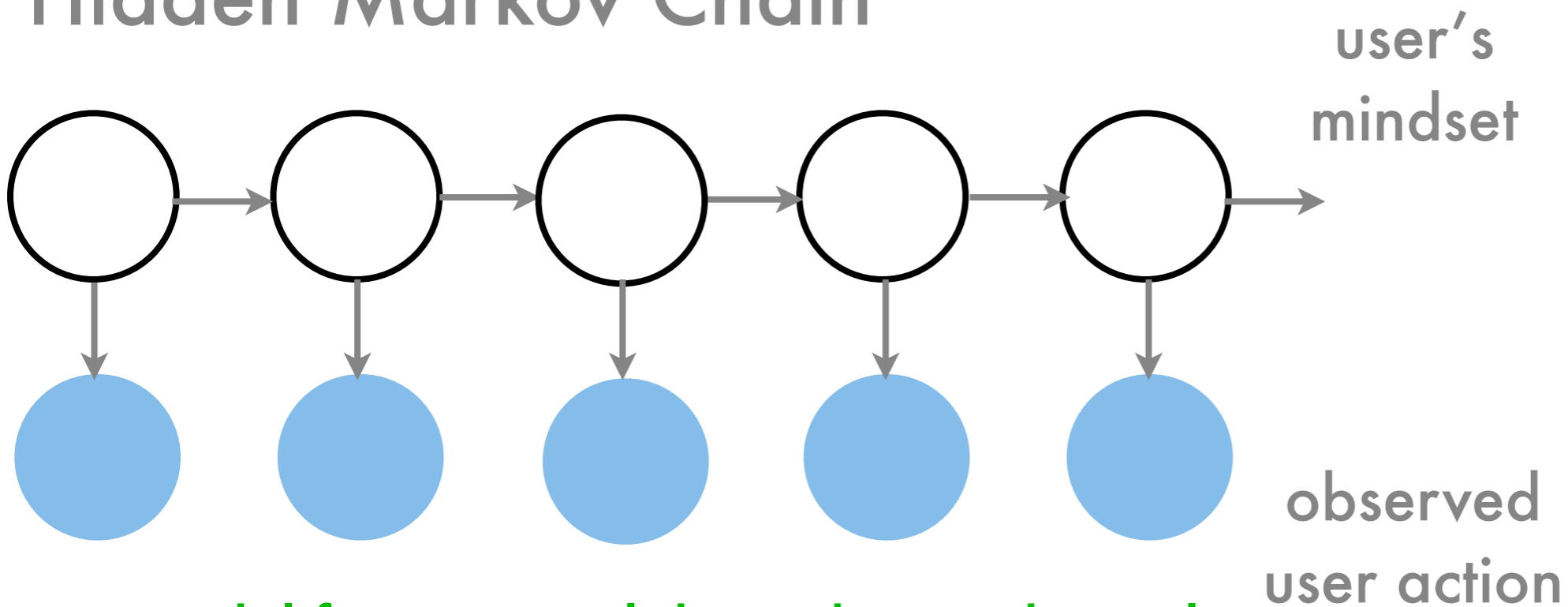
user model for traversal through search results

Chains

Markov Chain



Hidden Markov Chain



user model for traversal through search results

Chains

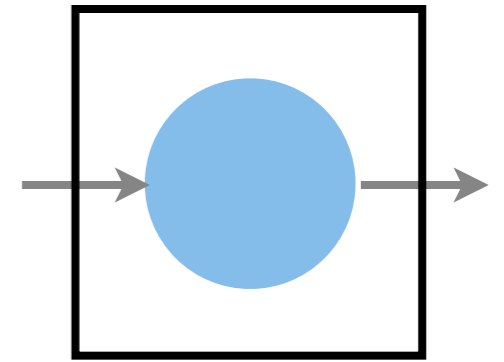
Markov Chain

$$p(x; \theta) = p(x_0; \theta) \prod_{i=1}^{n-1} p(x_{i+1} | x_i; \theta)$$

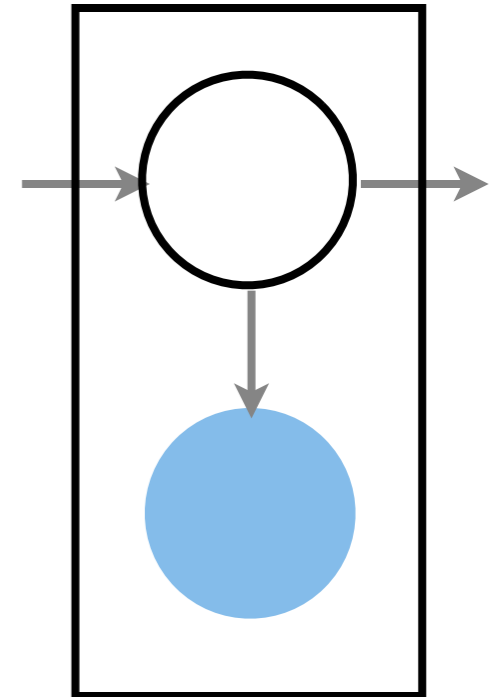
Hidden Markov Chain

$$p(x, y; \theta) = p(x_0; \theta) \prod_{i=1}^{n-1} p(x_{i+1} | x_i; \theta) \prod_{i=1}^n p(y_i | x_i)$$

Plate



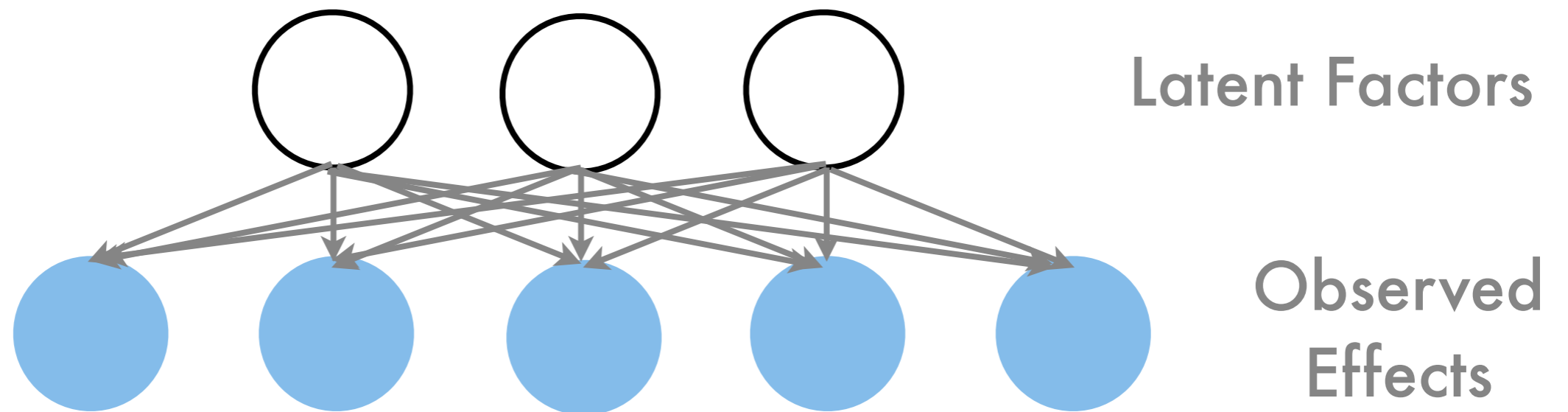
user's
mindset



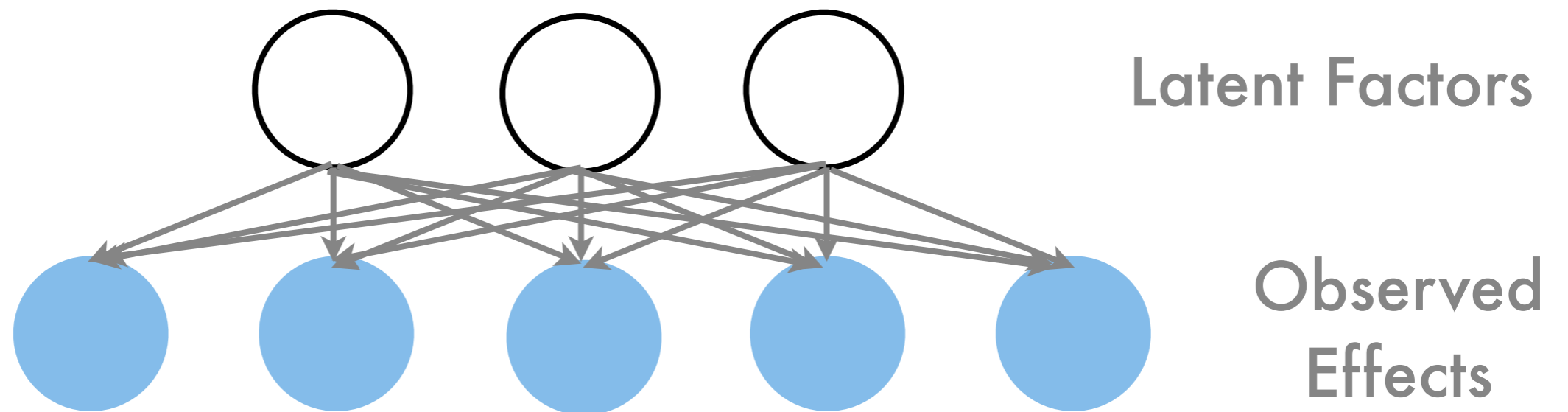
observed
user action

user model for traversal through search results

Factor Graphs

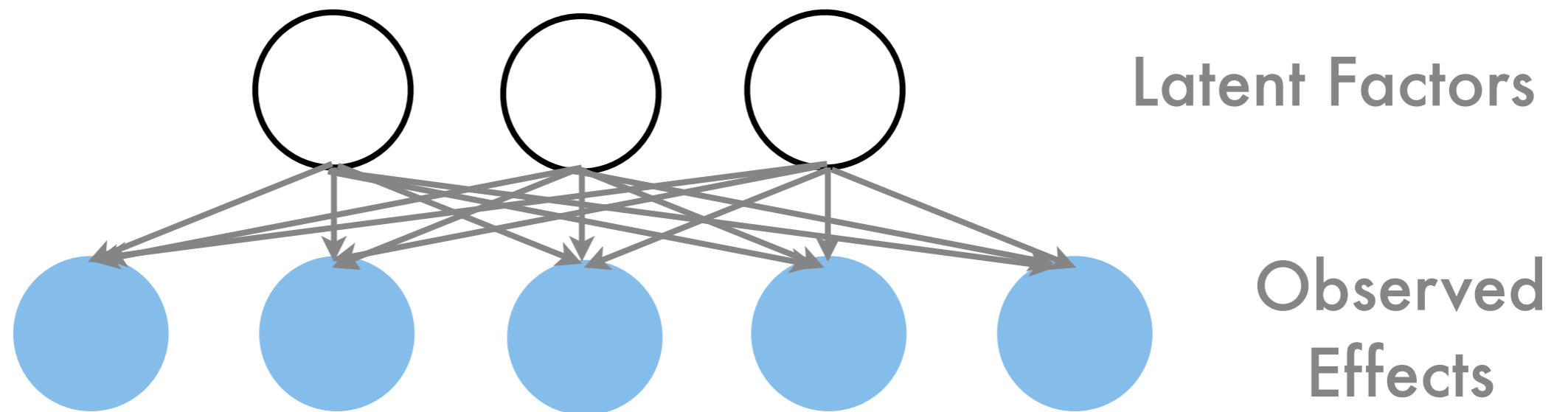


Factor Graphs



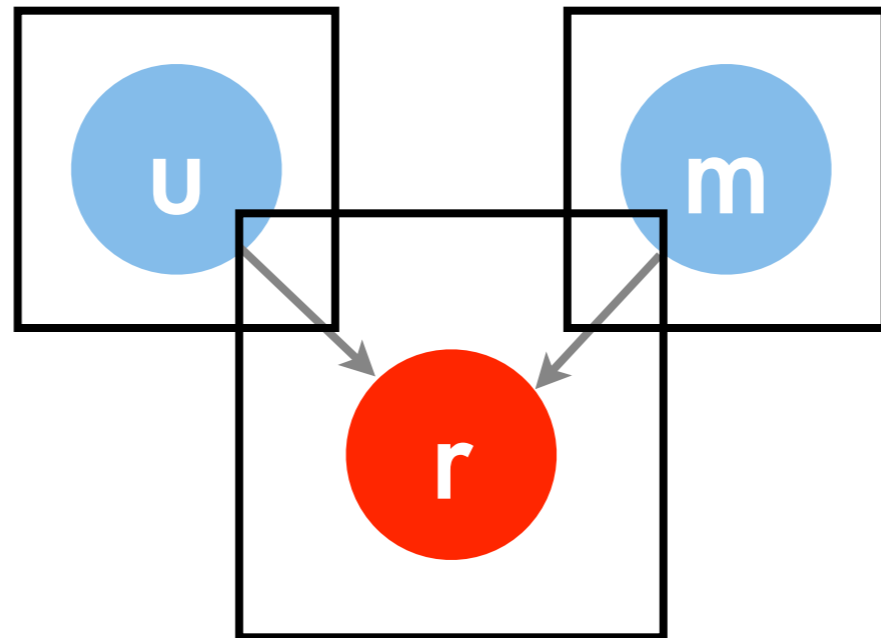
- **Observed effects**
Click behavior, queries, watched news, emails

Factor Graphs

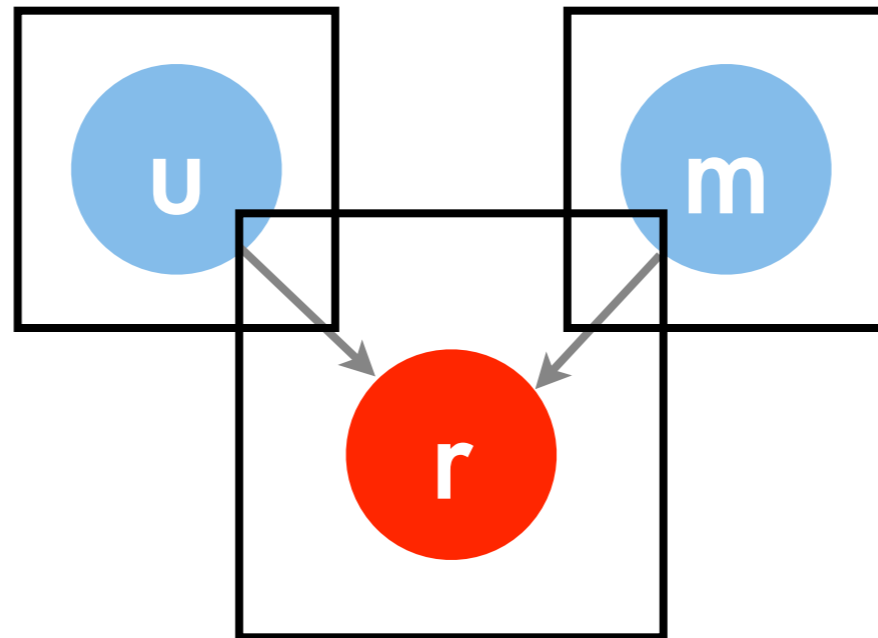


- **Observed effects**
Click behavior, queries, watched news, emails
- **Latent factors**
User profile, news content, hot keywords, social connectivity graph, events

Recommender Systems

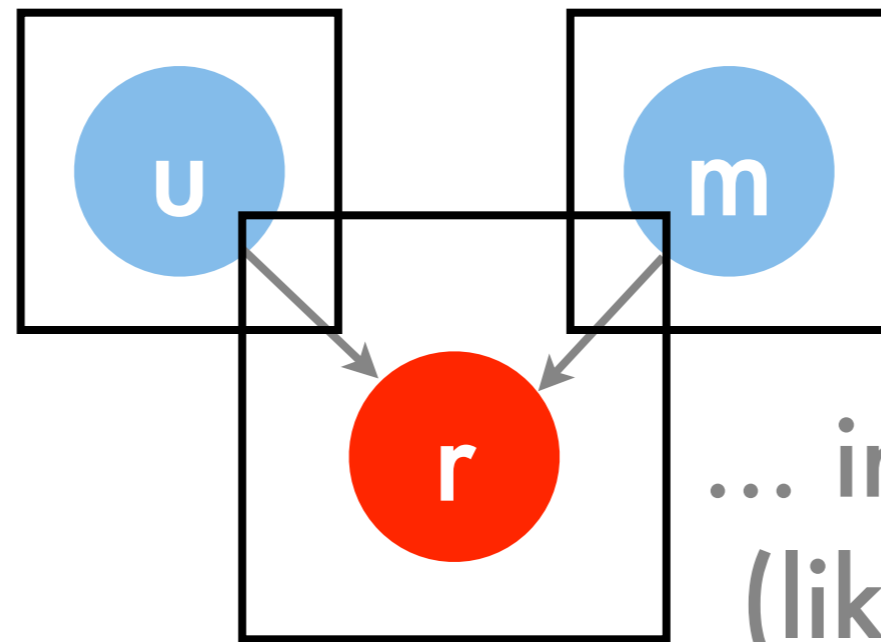


Recommender Systems



- **Users u**
- **Movies m**
- **Ratings r (but only for a subset of users)**

Recommender Systems

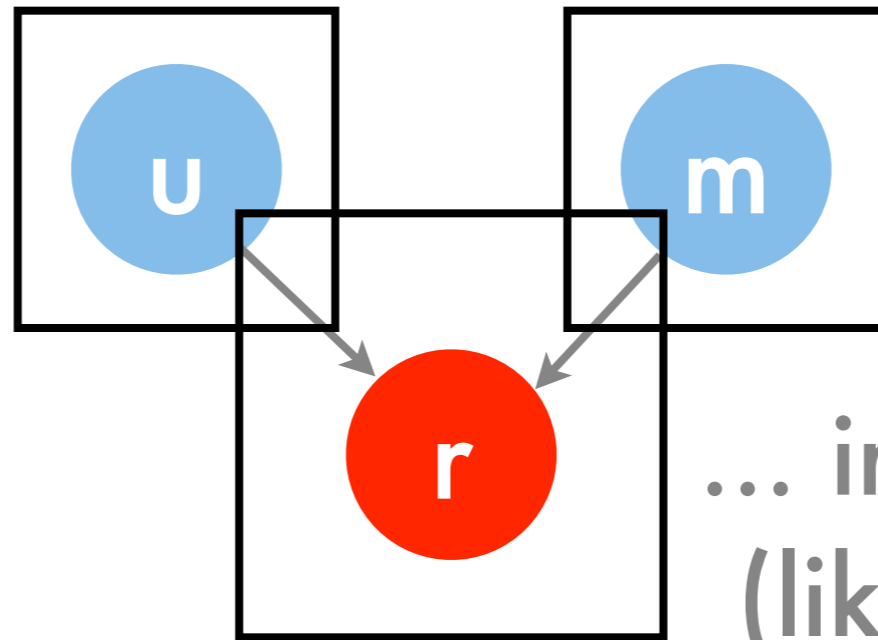


... intersecting plates ...
(like nested for loops)

- **Users u**
- **Movies m**
- **Ratings r (but only for a subset of users)**

Recommender Systems

news,
SearchMonkey
answers
social
ranking
OMG
personals



... intersecting plates ...
(like nested for loops)

- Users u
- Movies m
- Ratings r (but only for a subset of users)

Challenges



domain
expert



statistics

Challenges

- How to design models
 - Common (engineering) sense
 - Computational tractability



domain
expert



statistics

Challenges

- How to design models
 - Common (engineering) sense
 - Computational tractability
- Inference
 - Easy for fully observed situations
 - Many algorithms if not fully observed
 - Dynamic programming / message passing

domain
expert

statistics

Summary - Part 2

- Probability theory to estimate events
- Conjugate priors and Laplace smoothing
- Conjugate = phantasy data
- Collapsing
- Laplace smoothing
- Directed graphical models

Part 3 - Clustering & Topic Models

Inference Algorithms

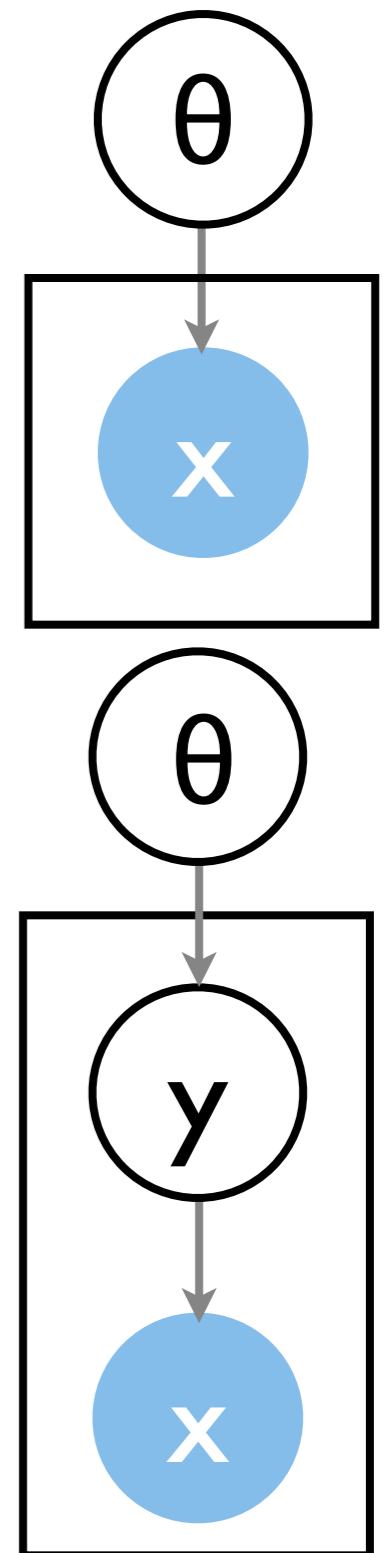
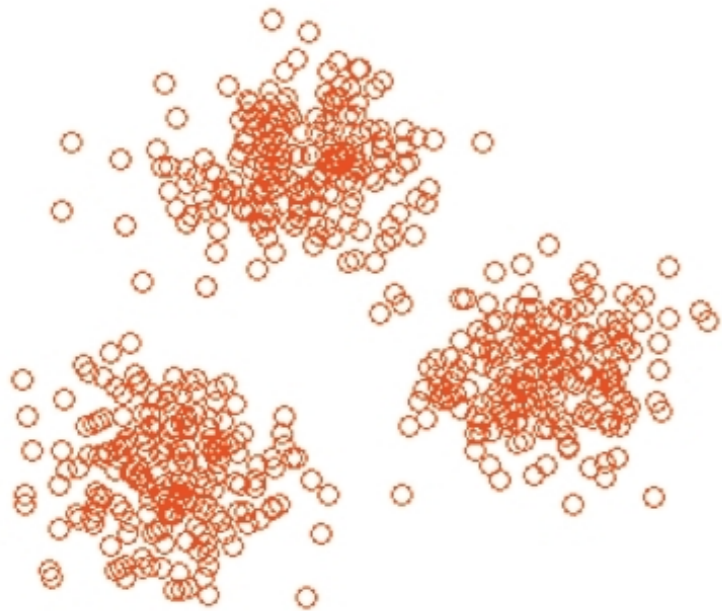
Clustering

Density Estimation

$$p(x, \theta) = p(\theta) \prod_{i=1}^n p(x_i | \theta)$$

Clustering

$$p(x, y, \theta) = p(\pi) \prod_{k=1}^K p(\theta_k) \prod_{i=1}^n p(y_i | \pi) p(x_i | \theta, y_i)$$



Clustering

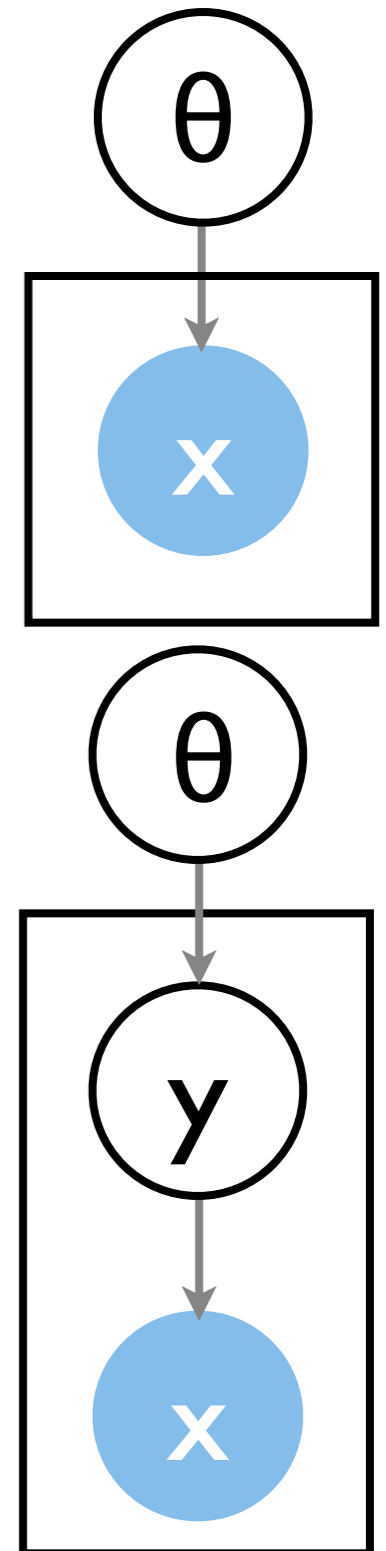
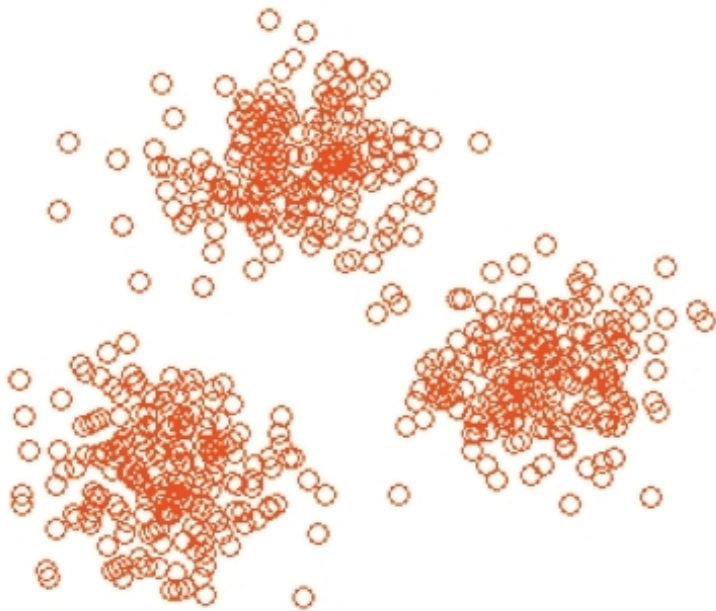
Density Estimation

$$p(x, \theta) = p(\theta) \prod_{i=1}^n p(x_i | \theta)$$

find θ

Clustering

$$p(x, y, \theta) = p(\pi) \prod_{k=1}^K p(\theta_k) \prod_{i=1}^n p(y_i | \pi) p(x_i | \theta, y_i)$$



Clustering

Density Estimation

log-concave

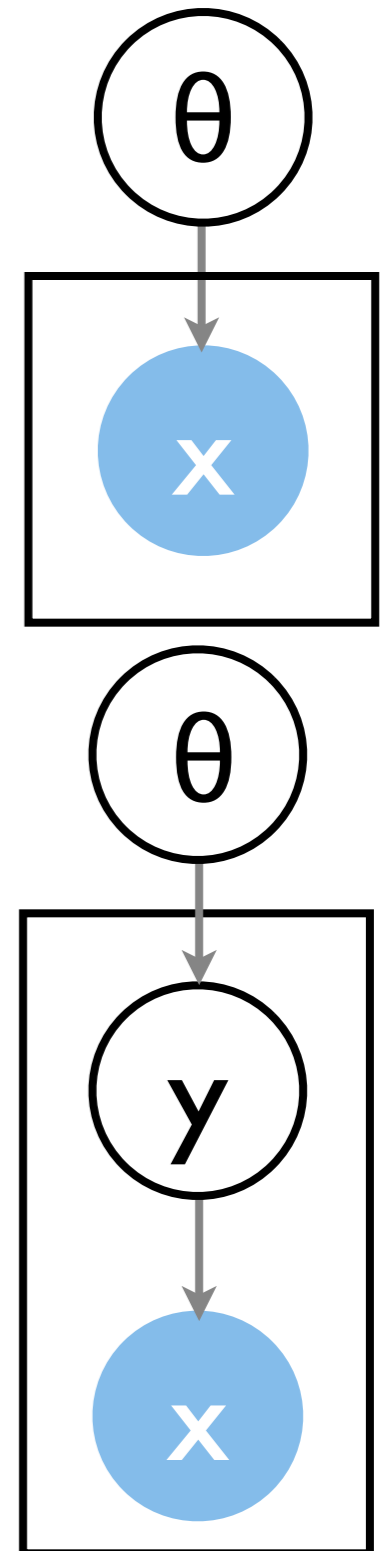
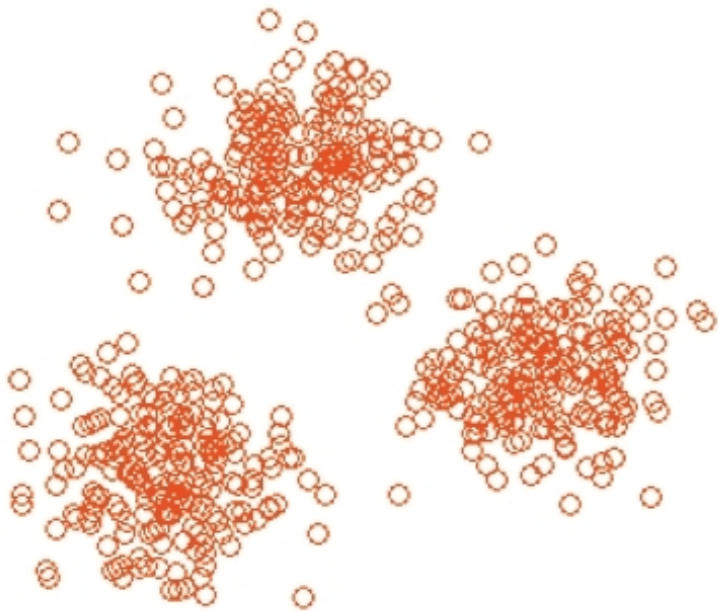
$$p(x, \theta) = p(\theta) \prod_{i=1}^n p(x_i | \theta)$$

find θ

Clustering

$$p(x, y, \theta) = p(\pi) \prod_{k=1}^K p(\theta_k) \prod_{i=1}^n p(y_i | \pi) p(x_i | \theta, y_i)$$

general nonlinear



Clustering

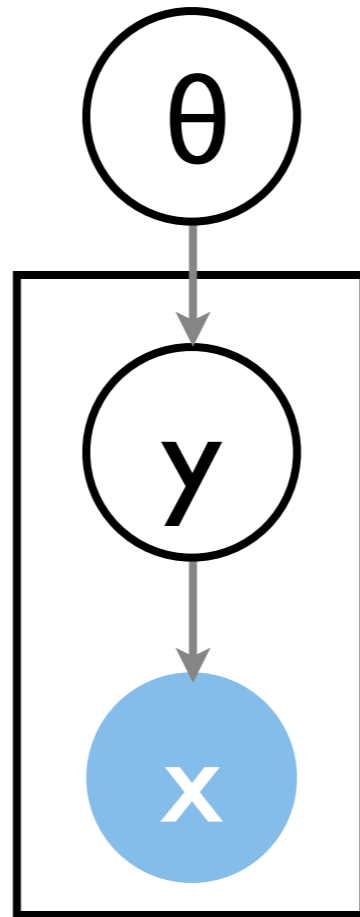
- **Optimization problem**

$$\text{maximize}_{\theta} \sum_y p(x, y, \theta)$$

$$\text{maximize}_{\theta} \log p(\pi) + \sum_{k=1}^K \log p(\theta_k) + \sum_{i=1}^n \log \sum_{y_i \in \mathcal{Y}} [p(y_i | \pi) p(x_i | \theta, y_i)]$$

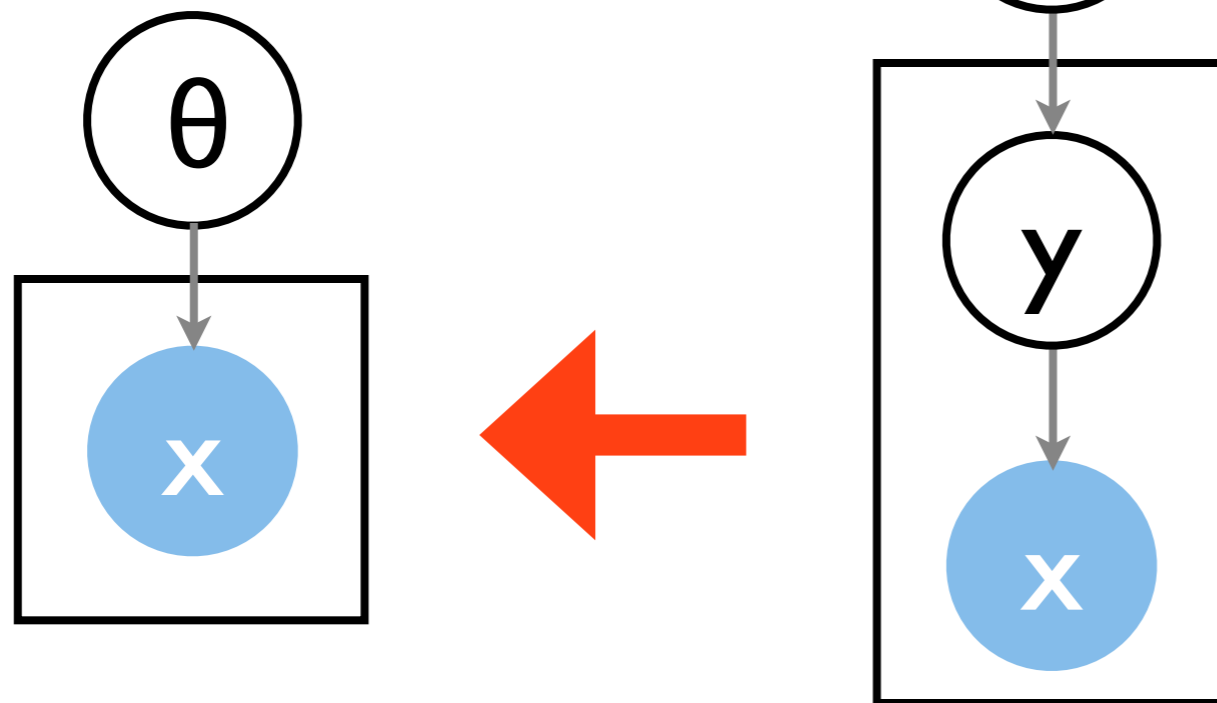
- **Options**
 - **Direct nonconvex optimization (e.g. BFGS)**
 - **Sampling (draw from the joint distribution)**
 - **Variational approximation**
(concave lower bounds aka EM algorithm)

Clustering



Clustering

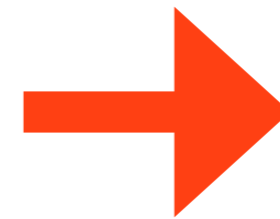
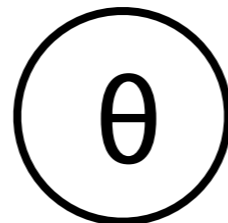
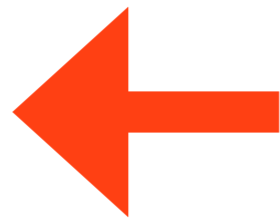
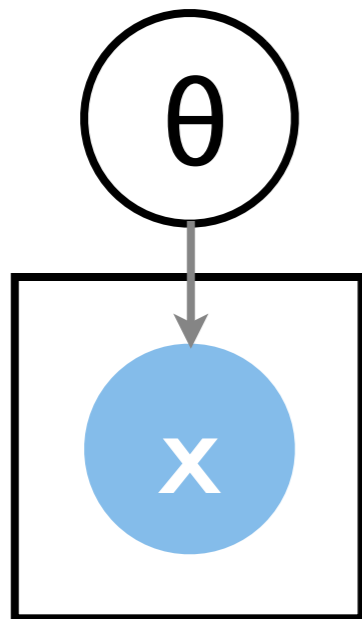
- Integrate out y



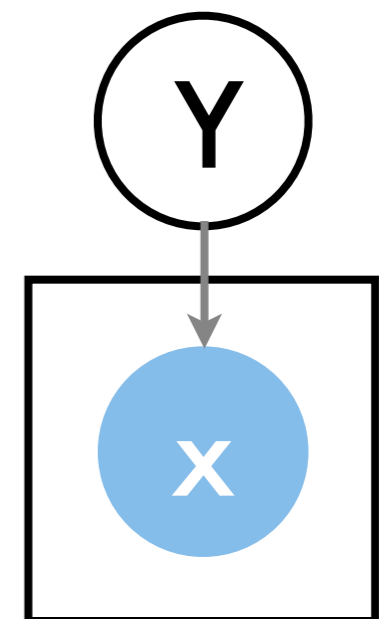
- Nonconvex optimization problem
- EM algorithm

Clustering

- Integrate out y



- Integrate out θ








- Nonconvex optimization problem
- EM algorithm

- Y is coupled
- Sampling
- Collapsed p

$$p(y|x) \propto p(\{x\} | \{x_i : y_i = y\} \cup X_{\text{fake}})p(y|Y \cup Y_{\text{fake}})$$






Gibbs sampling

- **Sampling:**
Draw an instance x from distribution $p(x)$
- **Gibbs sampling:**
 - In most cases direct sampling not possible
 - Draw one set of variables at a time

		
	0.45	0.05
	0.05	0.45

Gibbs sampling






- Sampling:
Draw an instance x from distribution $p(x)$
- Gibbs sampling:
 - In most cases direct sampling not possible
 - Draw one set of variables at a time

		
	0.45	0.05
	0.05	0.45

(b,g) - draw $p(.,g)$

Gibbs sampling






- Sampling:
Draw an instance x from distribution $p(x)$
- Gibbs sampling:
 - In most cases direct sampling not possible
 - Draw one set of variables at a time

		
	0.45	0.05
	0.05	0.45

(b, g) - draw $p(., g)$
 (g, g) - draw $p(g, .)$

Gibbs sampling






- Sampling:
Draw an instance x from distribution $p(x)$
- Gibbs sampling:
 - In most cases direct sampling not possible
 - Draw one set of variables at a time

		
	0.45	0.05
	0.05	0.45

(b, g) - draw $p(., g)$
 (g, g) - draw $p(g, .)$
 (g, g) - draw $p(., g)$

Gibbs sampling






- Sampling:
Draw an instance x from distribution $p(x)$
- Gibbs sampling:
 - In most cases direct sampling not possible
 - Draw one set of variables at a time

		
	0.45	0.05
	0.05	0.45

(b, g) - draw $p(., g)$
 (g, g) - draw $p(g, .)$
 (g, g) - draw $p(., g)$
 (b, g) - draw $p(b, .)$

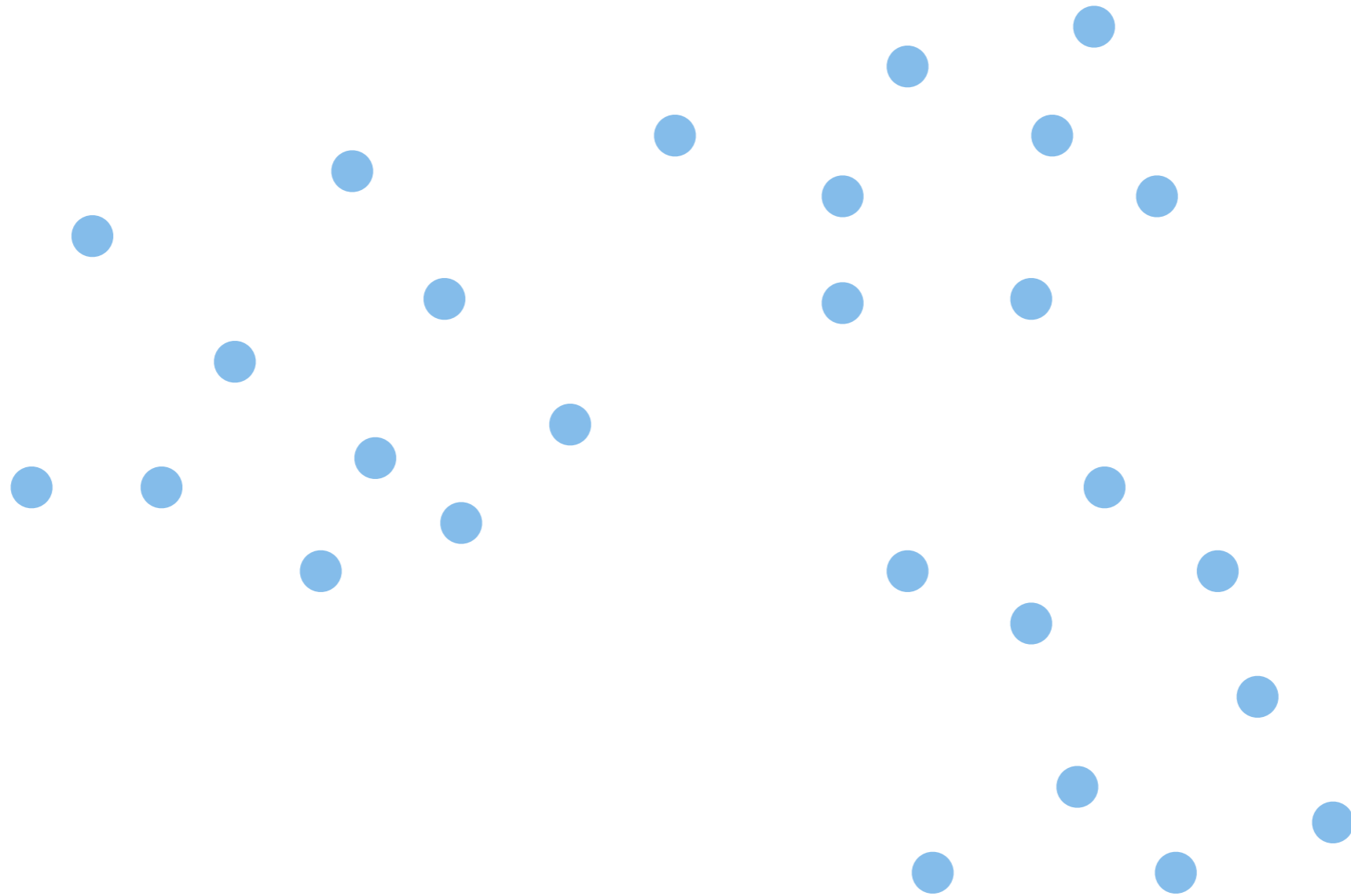
Gibbs sampling

- Sampling:
Draw an instance x from distribution $p(x)$
- Gibbs sampling:
 - In most cases direct sampling not possible
 - Draw one set of variables at a time

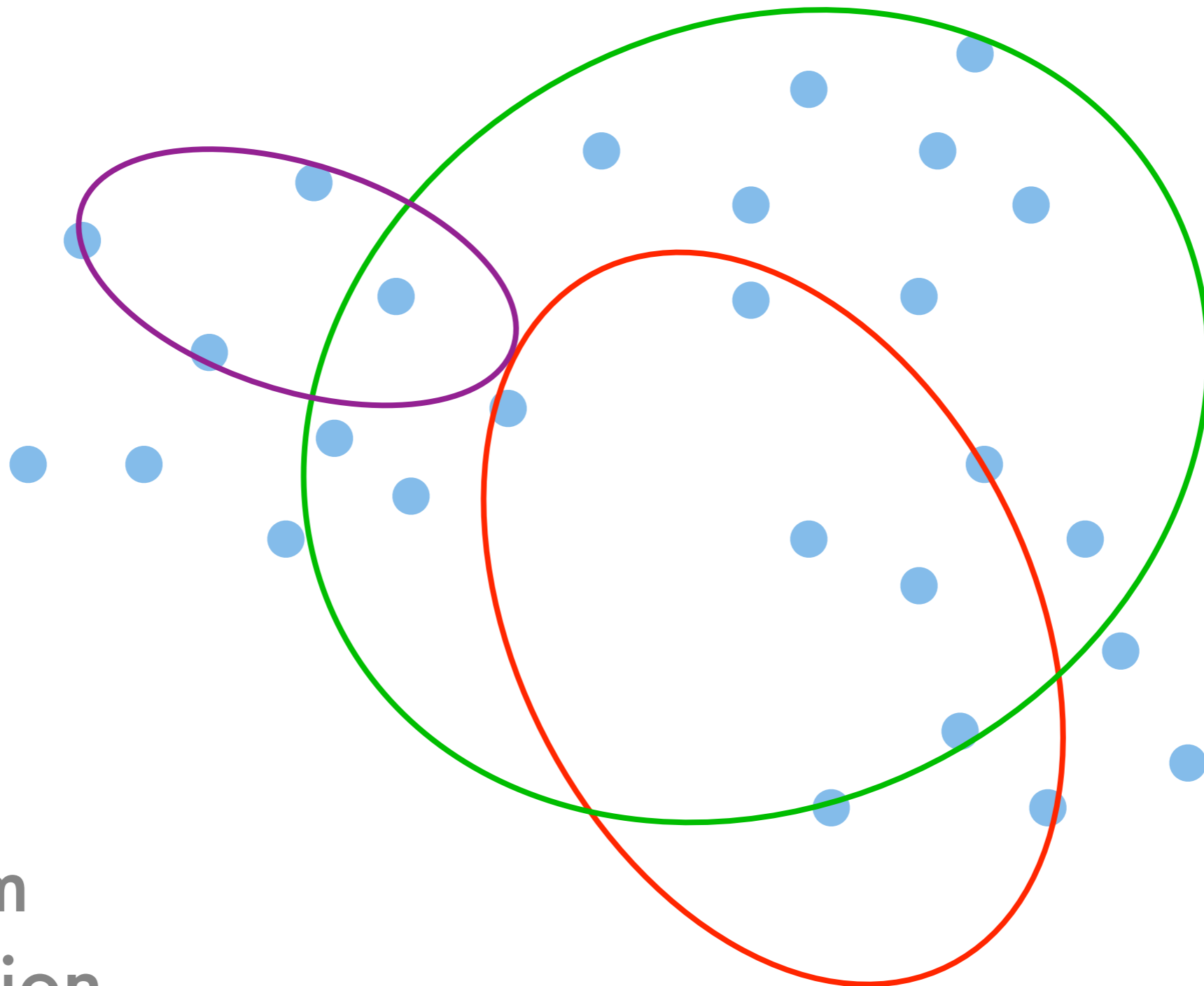
		
	0.45	0.05
	0.05	0.45

(b, g) - draw $p(., g)$
 (g, g) - draw $p(g, .)$
 (g, g) - draw $p(., g)$
 (b, g) - draw $p(b, .)$
 (b, b) ...

Gibbs sampling for clustering

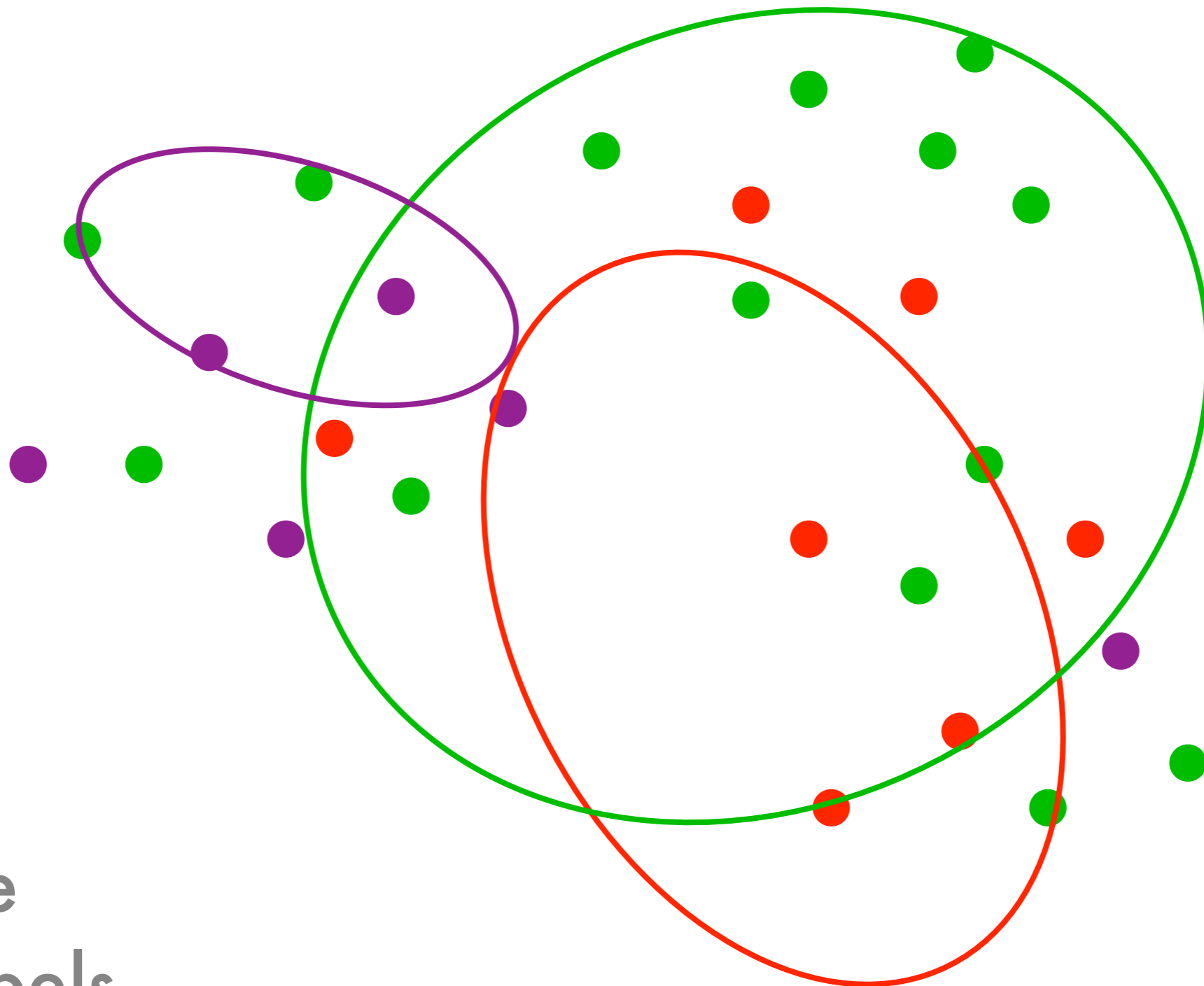


Gibbs sampling for clustering



random
initialization

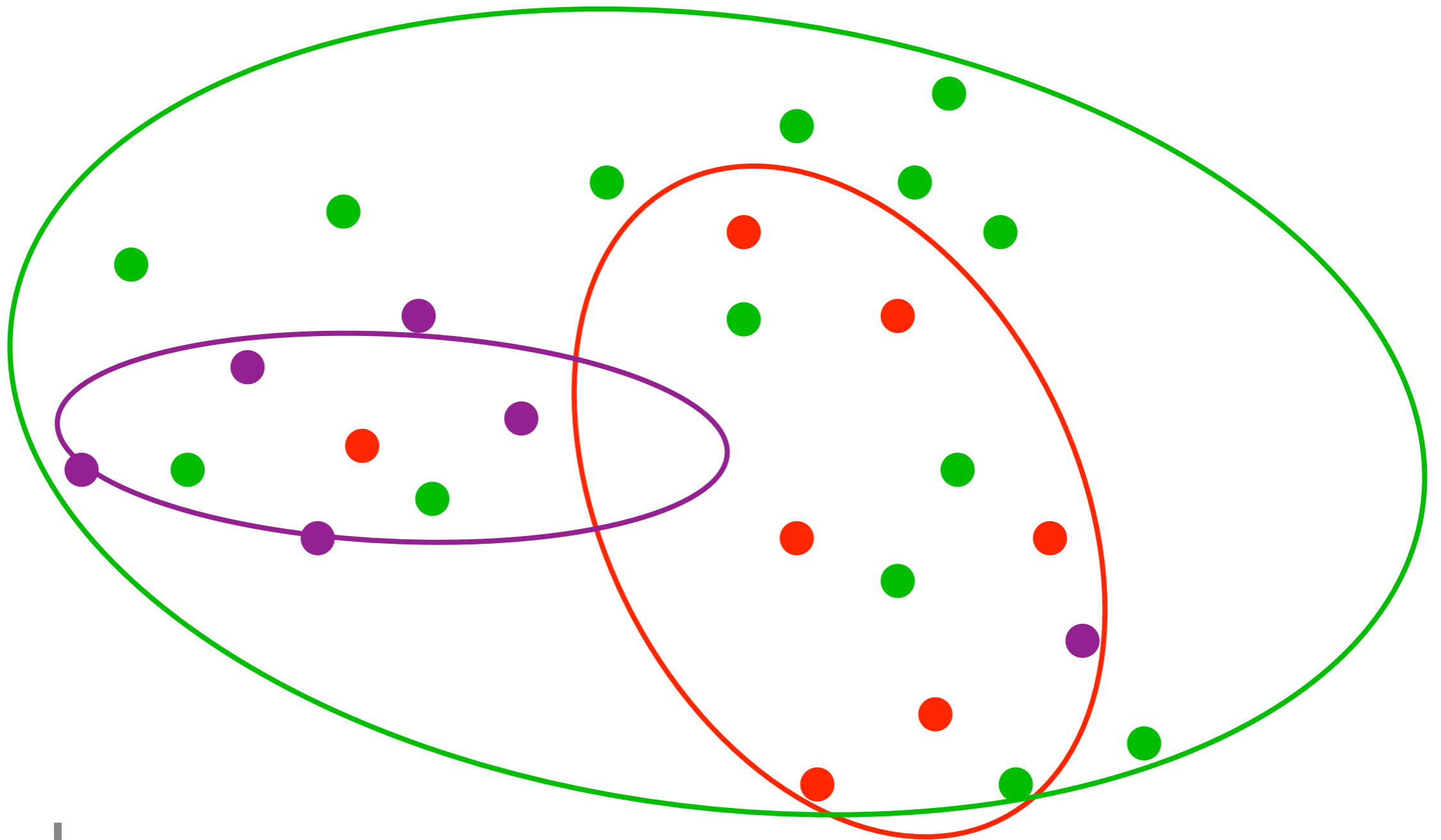
Gibbs sampling for clustering



sample

cluster labels

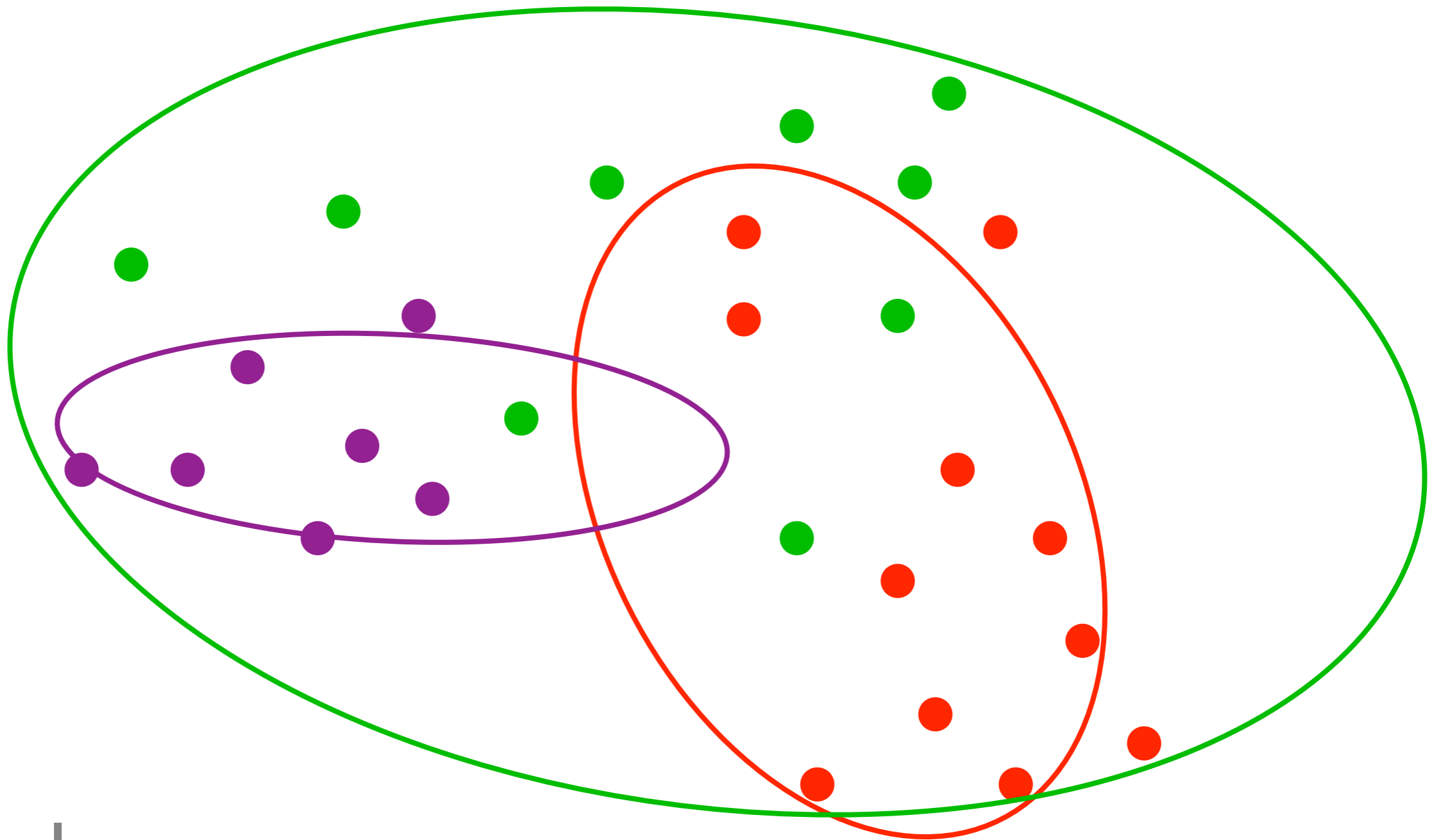
Gibbs sampling for clustering



resample

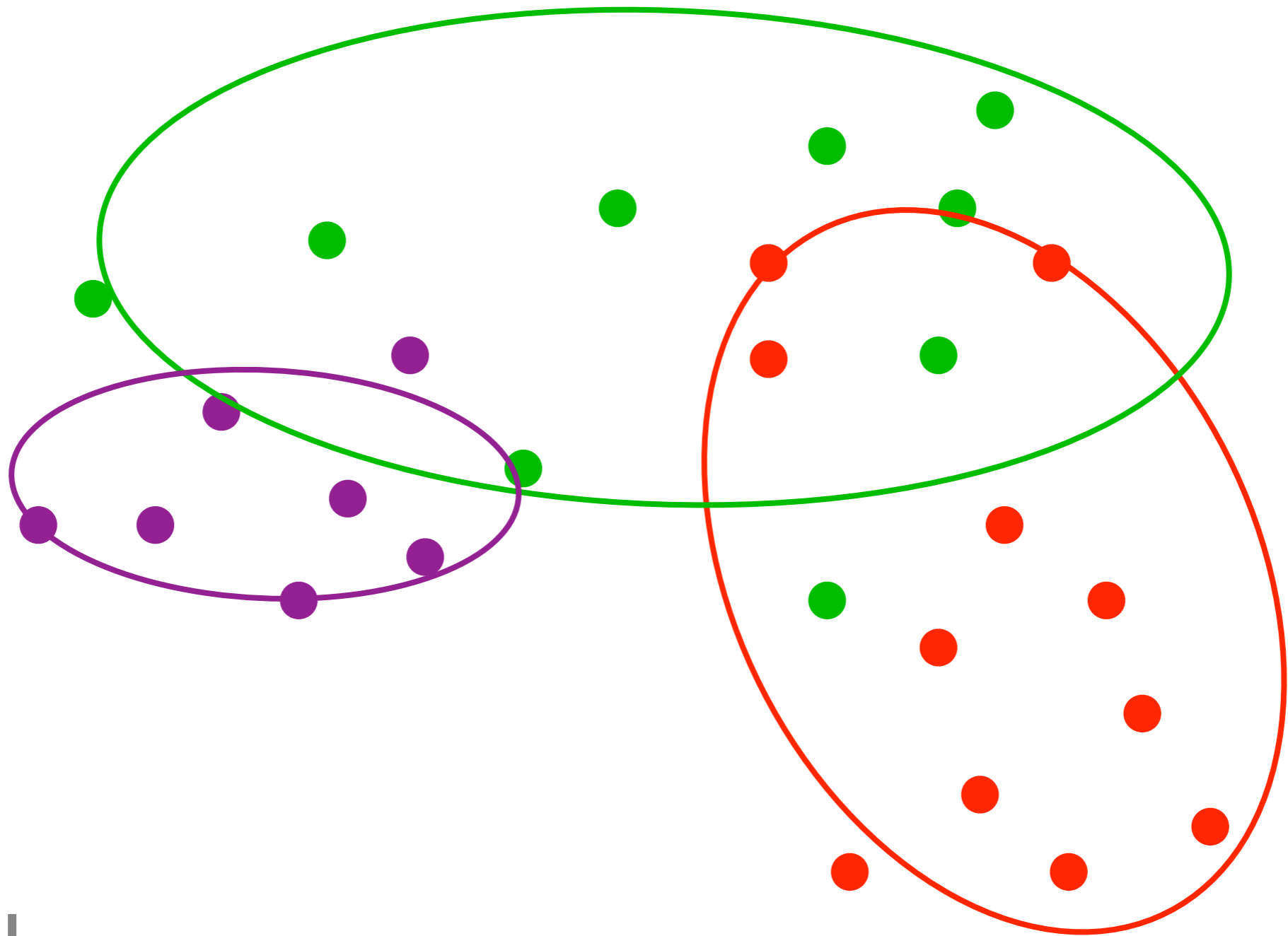
cluster model

Gibbs sampling for clustering



resample
cluster labels

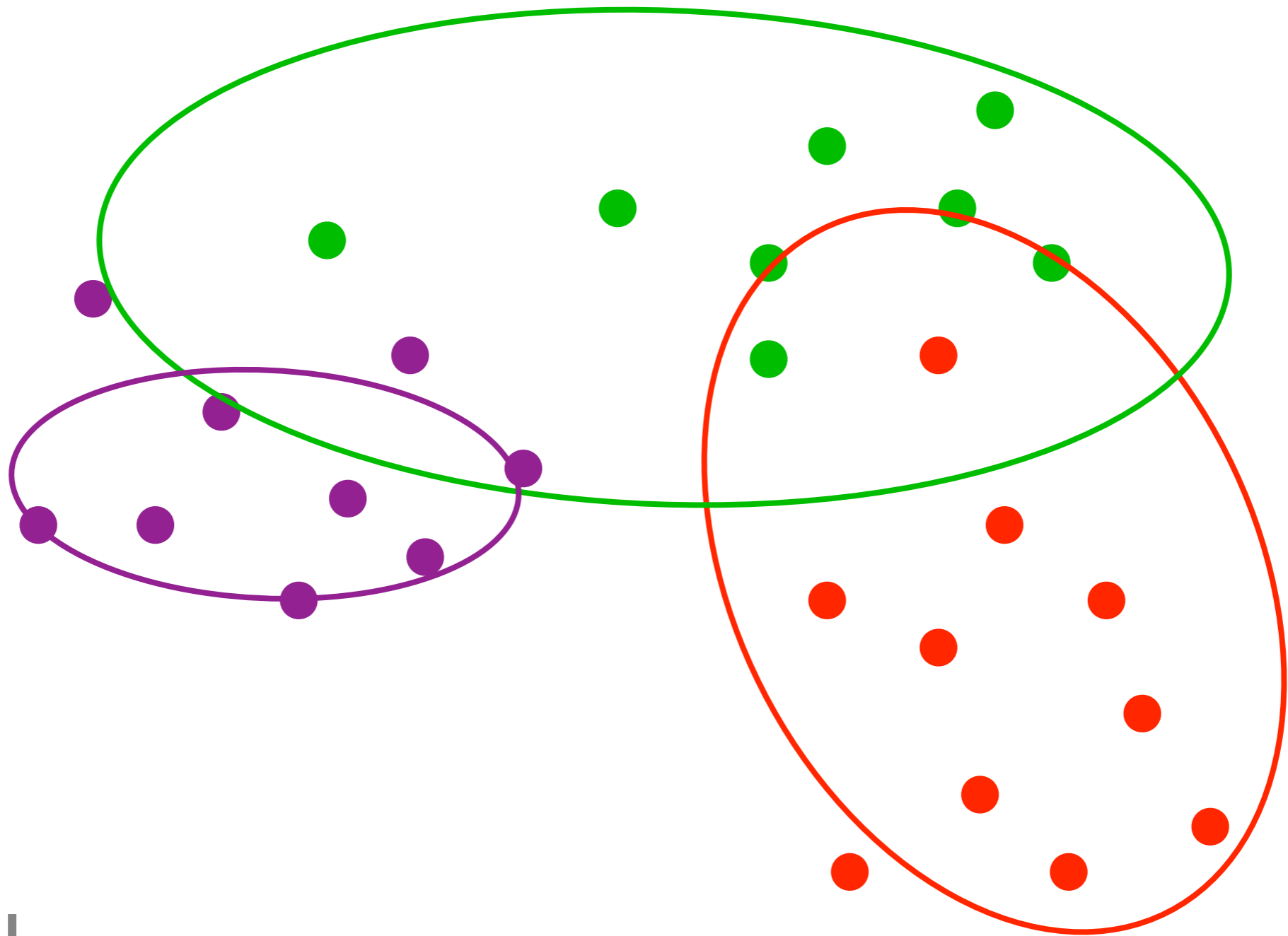
Gibbs sampling for clustering



resample

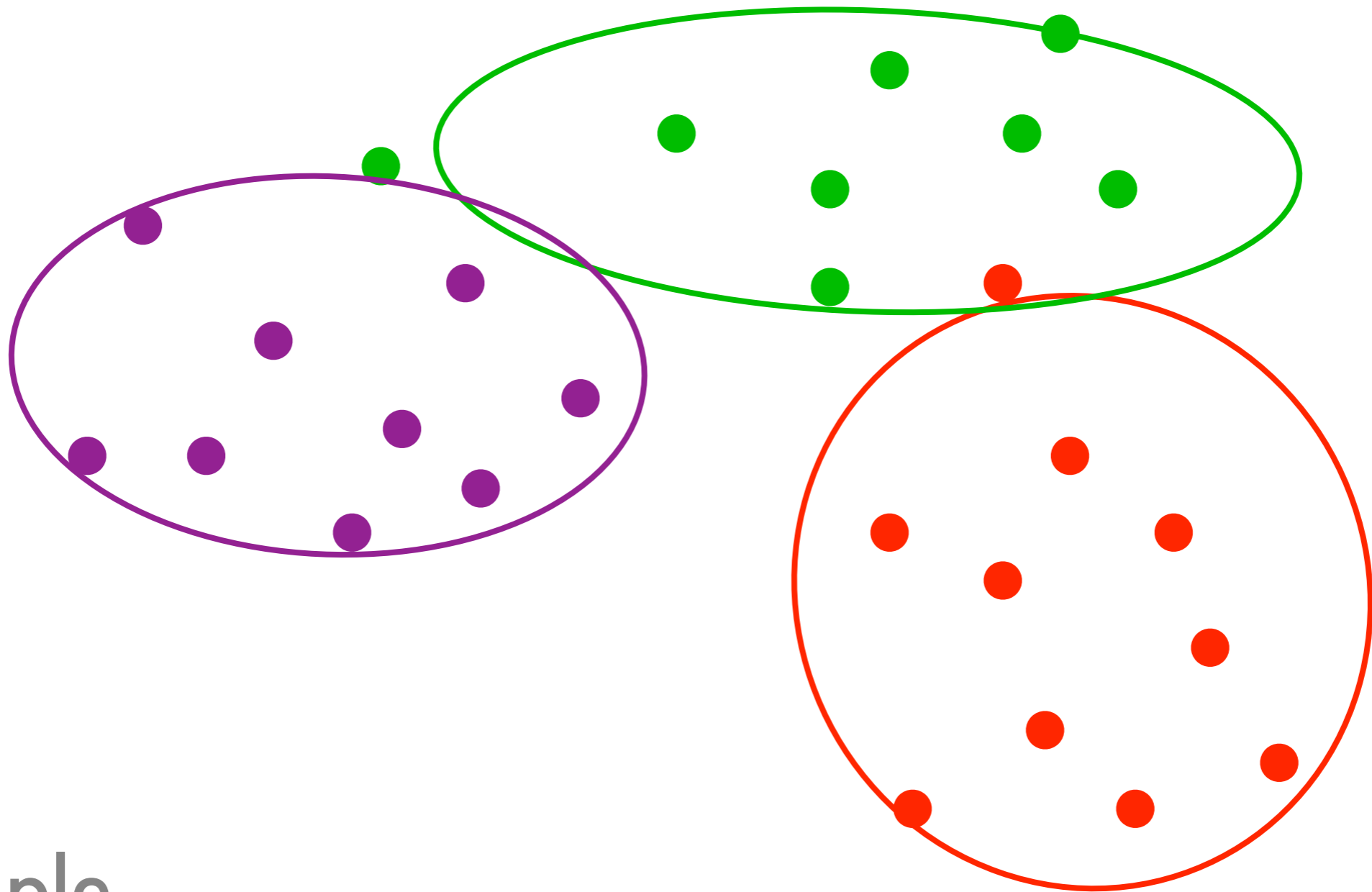
cluster model

Gibbs sampling for clustering



resample
cluster labels

Gibbs sampling for clustering



resample

cluster model

e.g. Mahout Dirichlet Process Clustering

Inference Algorithm \neq Model

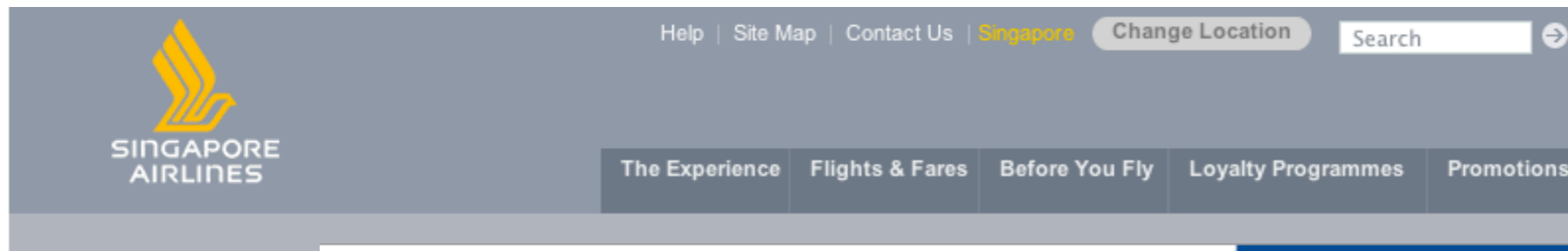
Inference Algorithm \neq Model

Corollary: EM \neq Clustering

Topic models

Grouping objects

Grouping objects



SINGAPORE AIRLINES

Help | Site Map | Contact Us | Singapore | Change Location | Search

The Experience | Flights & Fares | Before You Fly | Loyalty Programmes | Promotions

Book a Flight | Check In

Round Trip One Way

From:



myEMAIL | IVLE | LIBRARY | MAPS | CALENDAR | SITEMAP | CONTACT | e-CARDS

Search in GO

ABOUT NUS | GLOBAL | ADMISSIONS | ENTERPRISE | CAMPUS LIFE | GIVING | CAREERS@NUS

Home | About Us | Services | Events & Promotions | Shopping, Wining & Dining | Contact | Sitemap

Singapore

CHIJMES
restaurants • bars • shops

Discover a century of resplendent living history behind the cloistered walls.

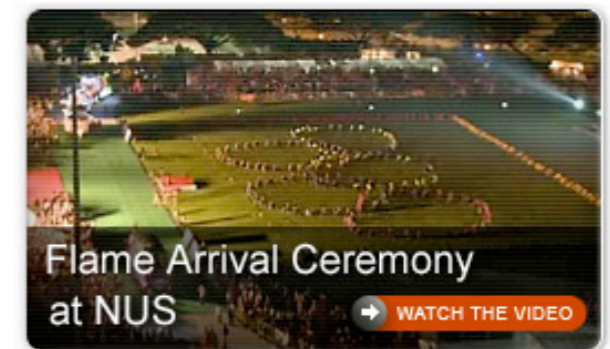
Chijmes, a premier lifestyle destination in Singapore

Owned by: Managed by: Property Manager:



Copyright © 2006 Chijmes. All rights reserved.

Feedback | Terms & Conditions



Flame Arrival Ceremony at NUS

WATCH THE VIDEO



Joint Evacuation Exercises

- 7 & 14 Sept 2010
- 10am - 12pm
- Heng Mui Keng Terrace & vicinity

MORE DETAILS

STAFF | ALUMNI | VISITORS

YAHOO!

Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's the United logo and navigation links like 'My profile', 'Worldwide sites', and 'Customer service'. Below that are menu items for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. A search bar is also present. The main content area features a 'BOOK FLIGHT' section with fields for 'From', 'To', 'Departing', and 'Returning'. There's also a 'REDEEM MILES' section. A large promotional banner for 'Use 30% fewer miles on your next United flight' is prominent. To the right, there's a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'. Below the login section, there are links for 'Start earning miles today' and 'united.com benefits and features'. At the bottom, there's a footer with 'About United', 'Investor relations', 'Business resources', 'Careers', and 'Site map'.

The screenshot shows the Australian National University (ANU) website. At the top, there's a navigation menu with links for 'Calendar', 'Sitemap', 'Contact', and 'e-CARDS'. Below that is a search bar with the text 'Search ANU...'. The main banner features the text 'The Australian National University' in a large, light blue font. Below the banner, there's another navigation menu with links for 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The background of the banner shows a close-up of a tree trunk with a small plant growing from it.

This section contains logos for three companies: Suntec, ARA, and APC. Below the logos, there's a small image of a tree trunk with a small plant growing from it, similar to the one in the ANU banner.

Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's a navigation bar with 'UNITED' logo and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below this is a search bar and a menu for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. The main content area features a large promotional banner for 'Use 30% fewer miles on your next United flight.' with a large orange percentage sign. To the right, there's a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'. Below the login section, there are links for 'Start with My Mileage Plus' and 'My reservations'. The bottom part of the page shows a list of flight routes with prices, such as 'Singapore - Bangkok SGD 395*', 'Singapore - Hong Kong SGD 546*', and 'Singapore - Taipei SGD 768*'. There are also buttons for 'Book Now' and 'Show Schedule'.

The screenshot shows the Australian National University (ANU) website. At the top, there's a navigation bar with 'EXPLORE ANU', 'A-Z INDEX', and a search bar. The main header features the ANU logo and the text 'The Australian National University'. Below this is a navigation menu with links for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The main content area features a large news article titled 'Ash forests rise and rise again' with a sub-headline 'A new book that graphically documents the spectacular natural recovery of Victoria's ash forests after the Black Saturday bushfires also argues that wildfires are typical natural disturbances in these environments.' Below the article, there are several smaller news items: 'Forests renew after Black Saturday fires', 'School of Music at Floriade', 'Undergraduate studies', and 'Higher Degree Research'. At the bottom, there's a navigation bar with links for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

The screenshot shows the Chez Panisse website. The main content area features a large menu with the following items: 'RESERVATIONS RESTAURANT & CAFÉ', 'MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST', 'ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION', 'SPECIAL EVENTS CALENDAR', 'STORE BOOKS • POSTERS • GIFTS', and 'CONTACT INFORMATION DIRECTIONS • MAILING LIST'. The background of the page shows a photograph of the restaurant's interior, which is a rustic, industrial-style space with brick walls and wooden tables.



ng, Wining & Dining | Contact | Sitemap | About Suntec REIT



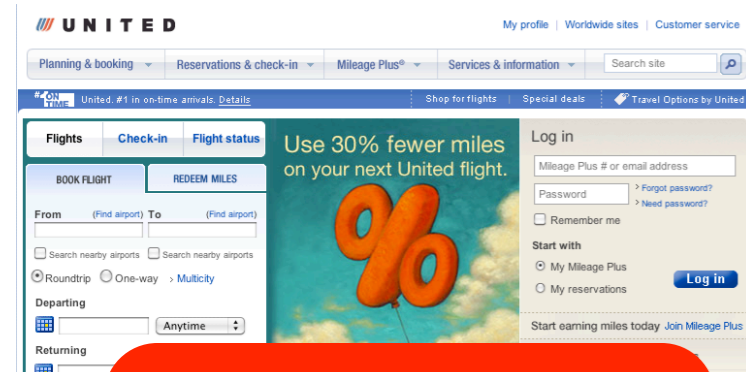
Grouping objects

The image shows a screenshot of the United Airlines website. The page features a navigation bar with 'UNITED' and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below the navigation, there are sections for 'Flights', 'Check-in', and 'Flight status'. A prominent red speech bubble with the word 'airline' is overlaid on the page. The website content includes a search form for flights, a 'Use 30% fewer miles on your next United flight' promotion, and a list of flight routes with prices, such as Singapore to Bangkok for SGD 395* and Singapore to London for SGD 768*.

The image shows a screenshot of the Australian National University (ANU) website. The page features a navigation bar with 'EXPLORE ANU', 'A-Z INDEX', and a search bar. Below the navigation, there is a header with the ANU logo and the text 'The Australian National University'. A prominent red speech bubble with the word 'university' is overlaid on the page. The website content includes a news article titled 'Ash forests rise and rise again' and a list of navigation buttons for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

The image shows a screenshot of the Chez Panisse restaurant website. The page features a navigation bar with 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. Below the navigation, there is a header with the text 'Chez Panisse' and a list of navigation buttons for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. A prominent red speech bubble with the word 'restaurant' is overlaid on the page. The website content includes a background image of a restaurant interior and a navigation bar at the bottom with 'Directions', 'Reservations', 'Contact', 'Feedback', 'Terms & Conditions'.

Grouping objects



UNITED My profile | Worldwide sites | Customer service

Planning & booking | Reservations & check-in | Mileage Plus® | Services & information | Search site

Use 30% fewer miles on your next United flight.

Log in

Mileage Plus # or email address

Password Forgot password? Need password?

Remember me

Start with

My Mileage Plus

My reservations

Start earning miles today Join Mileage Plus

USA



RESERVATIONS RESTAURANT & CAFÉ

MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST

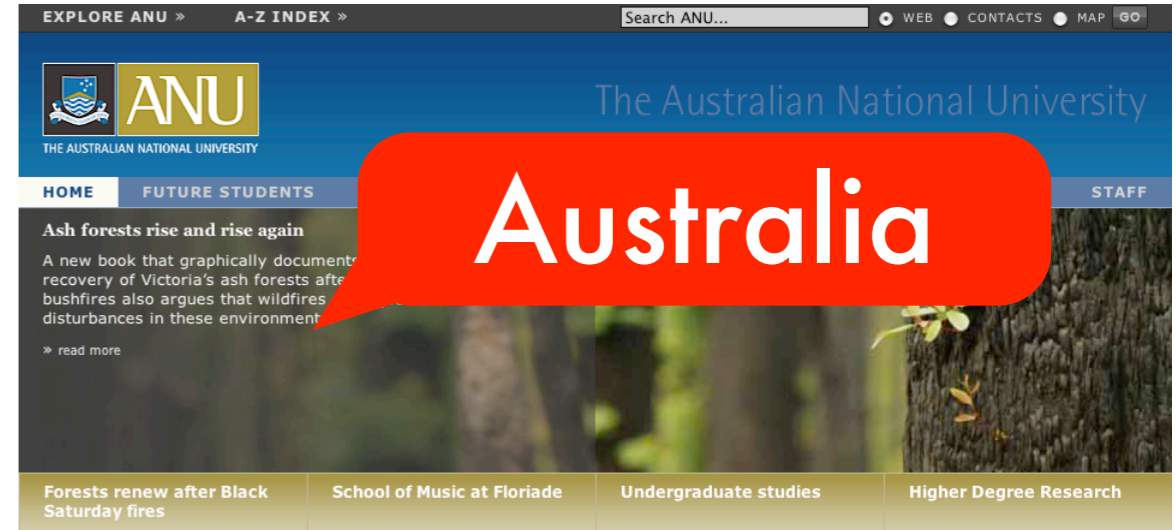
ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION

SPECIAL EVENTS CALENDAR

STORE BOOKS • POSTERS • GIFTS

CONTACT INFORMATION DIRECTIONS • MAILING LIST

© 1998-2010 Chez Panisse Restaurant & Café. All Rights Reserved.



EXPLORE ANU >> A-Z INDEX >> Search ANU... WEB CONTACTS MAP GO

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

The Australian National University

HOME FUTURE STUDENTS STAFF

Ash forests rise and rise again

A new book that graphically documents recovery of Victoria's ash forests after bushfires also argues that wildfires disturbances in these environment

read more

Forests renew after Black Saturday fires School of Music at Floriade Undergraduate studies Higher Degree Research

Australia



SINGAPORE AIRLINES

Book a Flight | Check in | Flight Status | My Bookings

Round Trip One Way Stopover/Multi-city

From: Departure City To: Destination City

Must travel on these dates

Adults: 1 Children (2-11): 0 Infants: 0

Need Help? View Book A Flight

SIA Holidays Hotel Bookings

NUS National University of Singapore

myEMAIL IVLE LIBRARY MAPS CALENDAR SITEMAP CONTACT e-CARDS

Search search for... in NUS Websites GO

ABOUT NUS GLOBAL ADMISSIONS EDUCATION RESEARCH ENTERPRISE CAMPUS LIFE GIVING CAREERS@NUS

A Leading Global University Centred in Asia

Home About Us Services Events & Promotions Shopping, Wining & Dining Contact Sitemap About Suntec REIT

Flame Arrival Ceremony at NUS WATCH THE VIDEO

Joint Evacuation Exercises 7 & 14 Sept 2010 10am - 12pm Heng Mui Keng Terrace & vicinity MORE DETAILS

ALUMNI VISITORS

Singapore



CHIJMES restaurant

Discover living in Singapore

Chijmes, a premier lifestyle destination in Singapore

Owned by: SUNTEC Managed by: ARA Property Manager: PC

Copyright © 2006 Chijmes. All rights reserved. Feedback | Terms & Conditions

YAHOO!

Topic Models

UNITED My profile | Worldwide sites | Customer service

Planning & booking | Reservations & check-in | Mileage Plus | Services & information | Search site

Use 30% fewer miles on your next United flight.

BOOK FLIGHT REDEEM MILES

From (Find airport) To (Find airport)

Roundtrip One-way Multicity

Departing Anytime

Returning Anytime

Search by Schedule & price Price & Flex

Adult (child or senior?)

Cabin Economy Refundable

Promotion code or Electronic certificate

Log in to view all seating options

Advanced Search

Cars Hotels Vacations

Learn more

USA
airline

EXPLORE ANU | A-Z INDEX | Search ANU... | WEB | CONTACTS

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

HOME | FUTURE STUDENTS | CURRICULUM | ABOUT ANU

Ash forests rise and rise again

A new book that graphically documents the recovery of Victoria's ash forests after the bushfires also argues that wildfires are typical disturbances in these environments.

Forests renew after Black Saturday fires | School of Music at Monash | Undergraduate studies | Higher Degree Research

Australia
university

SINGAPORE AIRLINES

The Experience | Flights & Fares | Before You Fly | Loyalty Programmes | Promotions

Book a Flight | Check In | Flight Status | My Bookings | Member Log-in

Round Trip One Way Stopover/Multi-city

From: Depart: Departure City

To: Return: Destination City

Must travel on these dates

Adults: Children (2-11): Infants:

Need Help? View Book A Flight | SIA Holidays | Hotel Bookings

Singapore - Bangkok SGD 395* | Singapore - Hong Kong SGD 546* | Singapore - Taipei SGD 768* | Singapore - Tokyo (Haneda) SGD 983* | Singapore - Sydney | Singapore - London

Singapore
airline

NUS National University of Singapore

myEMAIL | IVLE | LIBRARY | MAPS | CALENDAR | SITEMAP | CONTACT | CARDS

ABOUT NUS | GLOBAL | ADMISSIONS | EDUCATION | RESEARCH | ENTERPRISE | CAMPUS LIFE | GIVING | CAREERS@NUS

A Leading Global University

Game Arrival Ceremony

Joint Evacuation Exercises

7 & 14 Sept 2010

10am - 12pm

Heng Mui Keng Terrace & vicinity

PROSPECTIVE STUDENTS | CURRENT STUDENTS | STAFF | ALUMNI | VISITORS

Singapore
university

Chez Panisse

RESERVATIONS RESTAURANT & CAFÉ

MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST

ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION

SPECIAL EVENTS CALENDAR

STORE BOOKS • POSTERS • GIFTS

CONTACT INFORMATION DIRECTIONS • MAILING LIST

Directions Reservations Contact

USA
food

Services | Events & Promotions | Shopping, Wining & Dining | Contact | Sitemap | About Suntec REIT

Chijmes

Discover a century of resplendent living history behind the cloisters

Chijmes, a premier lifestyle destination in Singapore

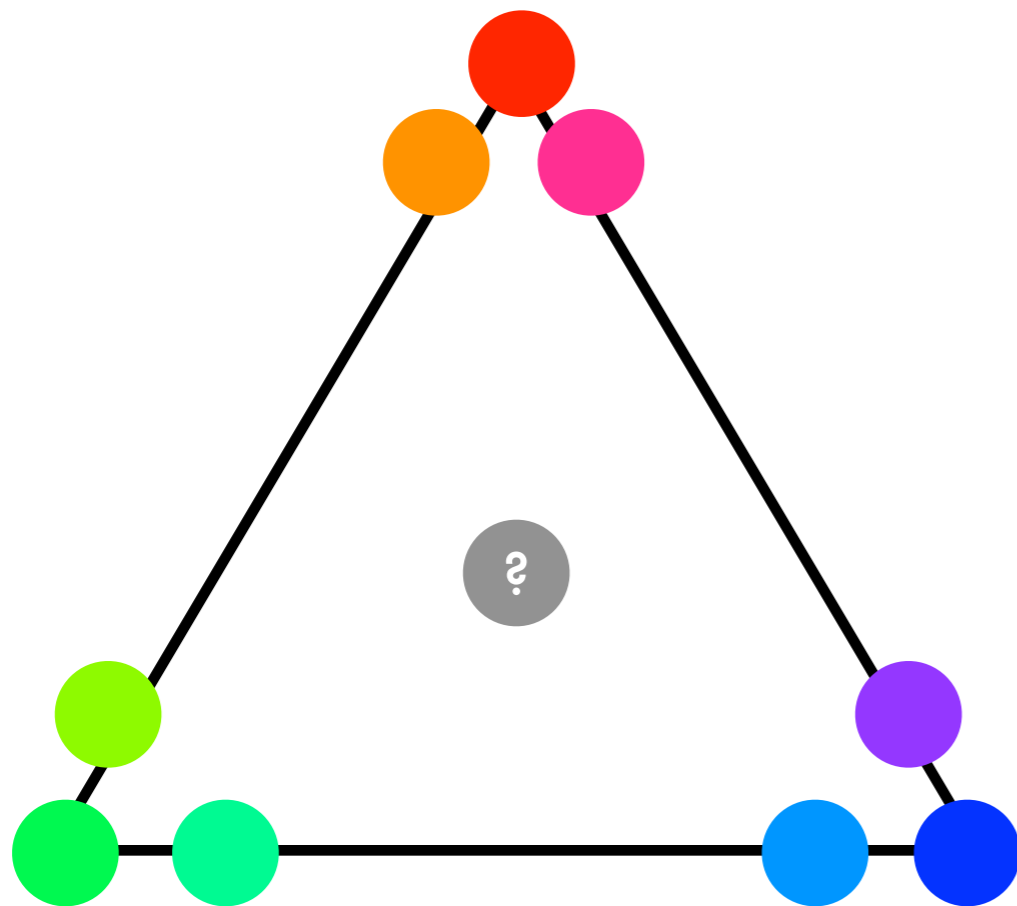
Owned by: SUNTEC | Managed by: ARA | Property Manager: APC

Copyright © 2006 Chijmes. All rights reserved. Feedback | Terms & Conditions

Singapore
food

Clustering & Topic Models

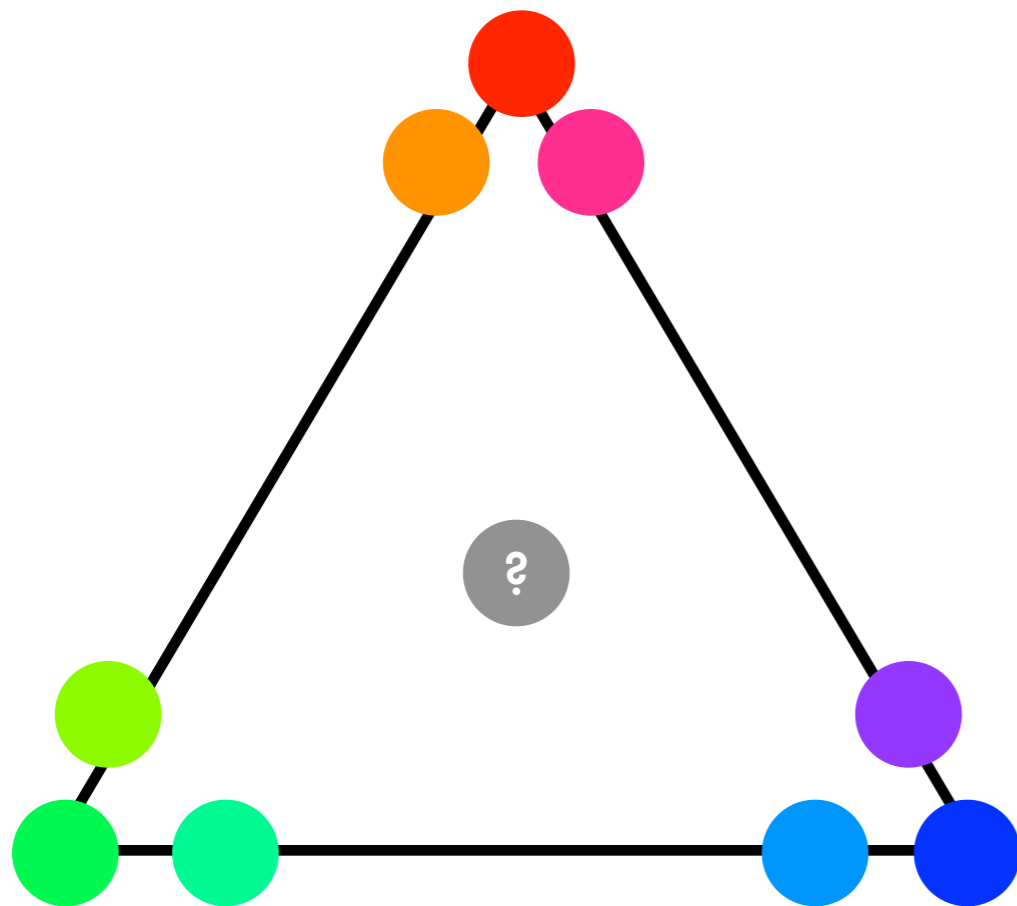
Clustering



group objects
by prototypes

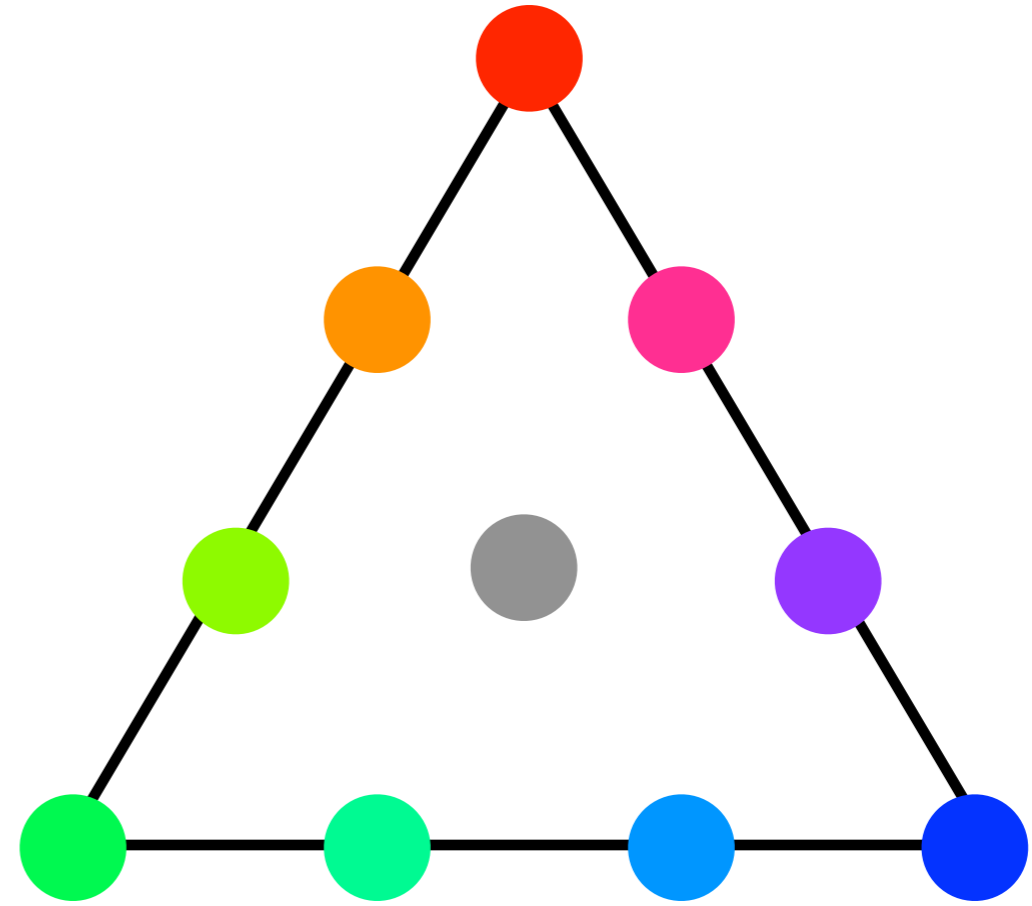
Clustering & Topic Models

Clustering



group objects
by prototypes

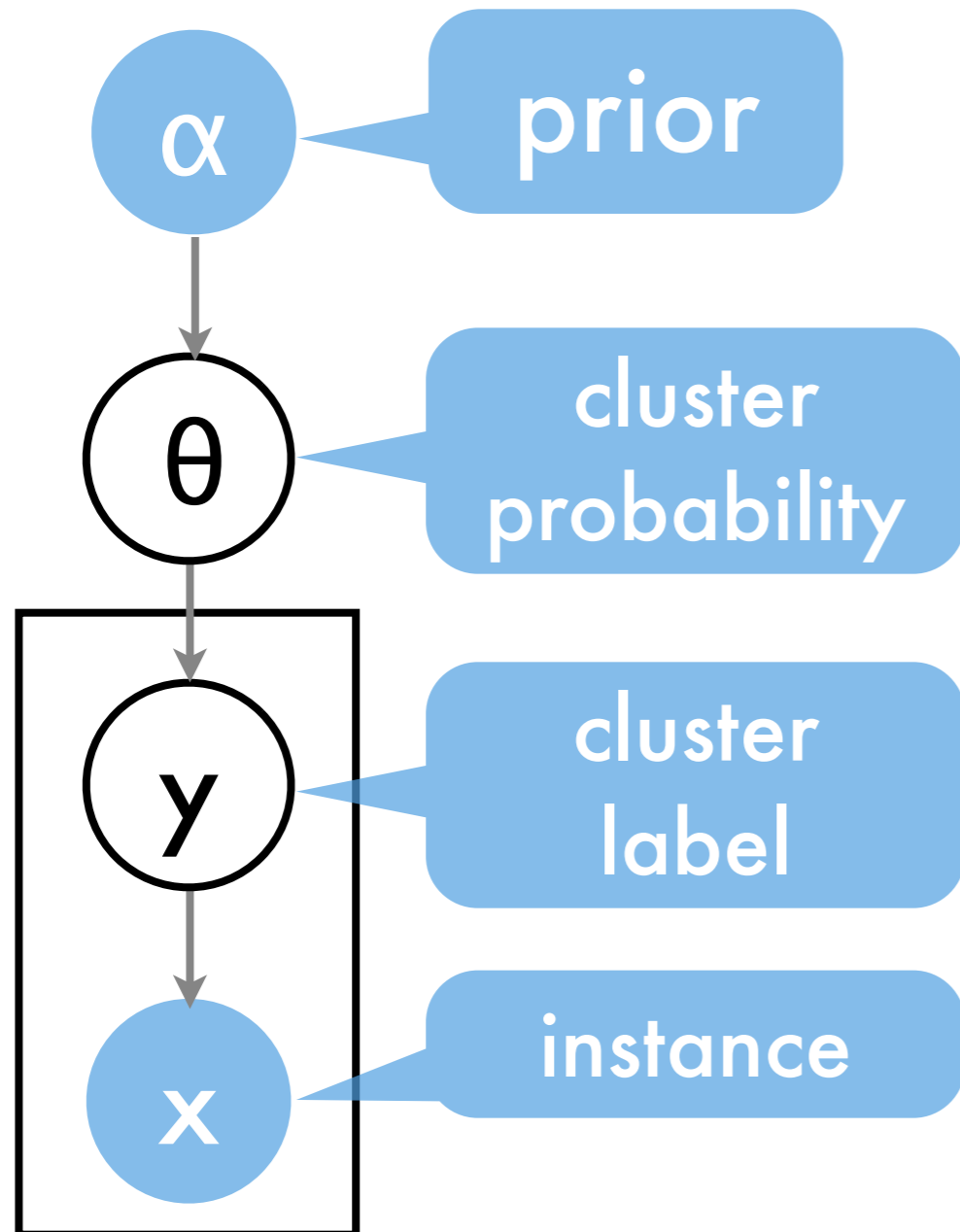
Topics



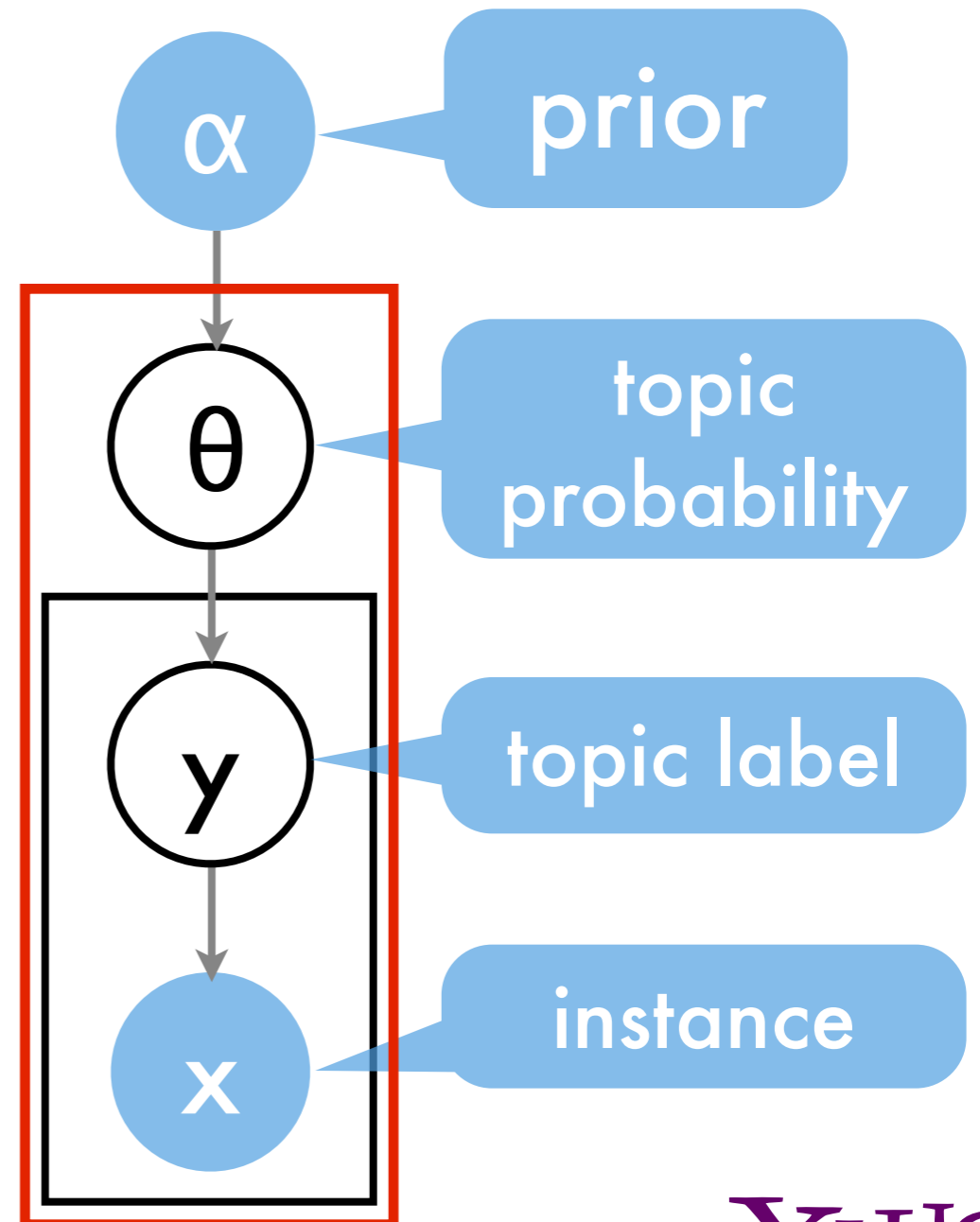
decompose objects
into prototypes

Clustering & Topic Models

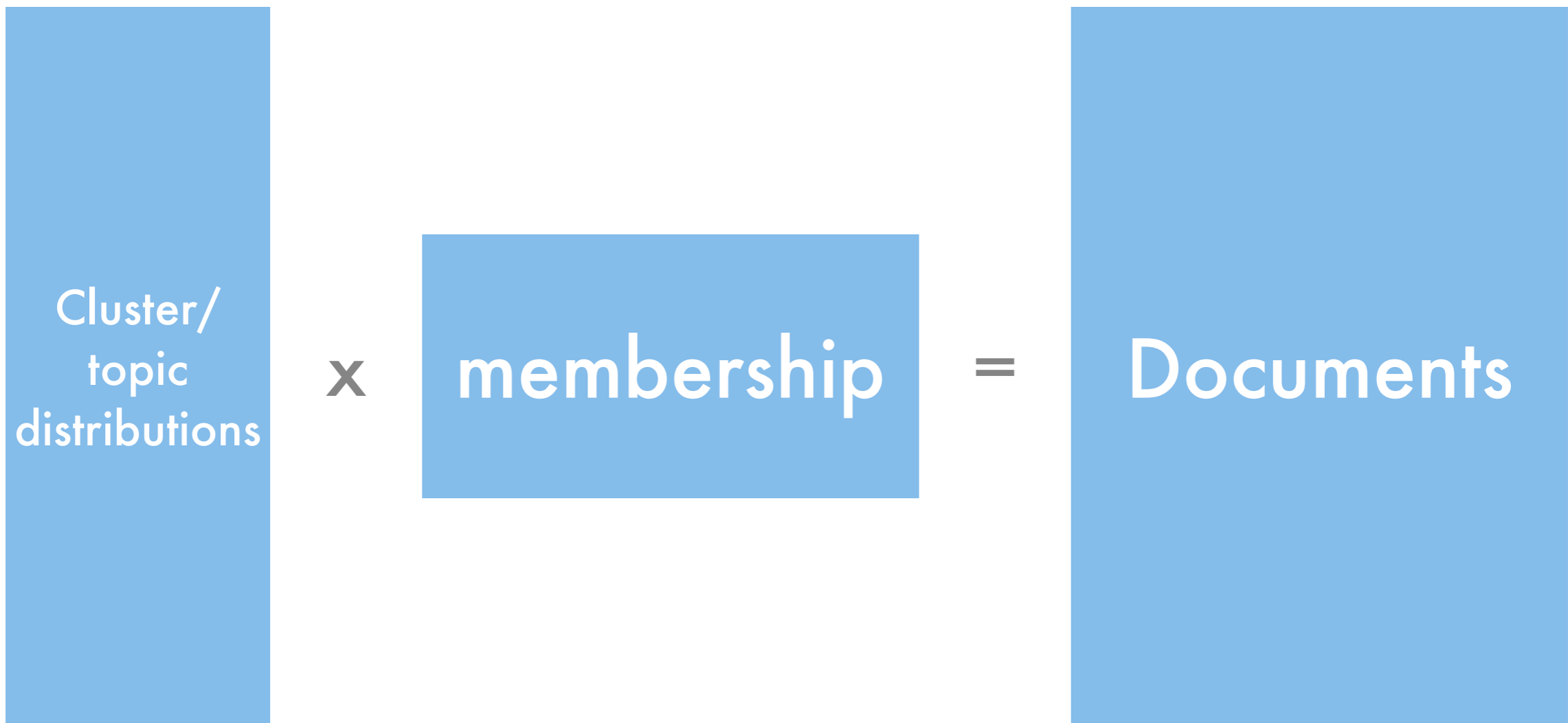
clustering



Latent Dirichlet Allocation



Clustering & Topic Models



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

Topics in text

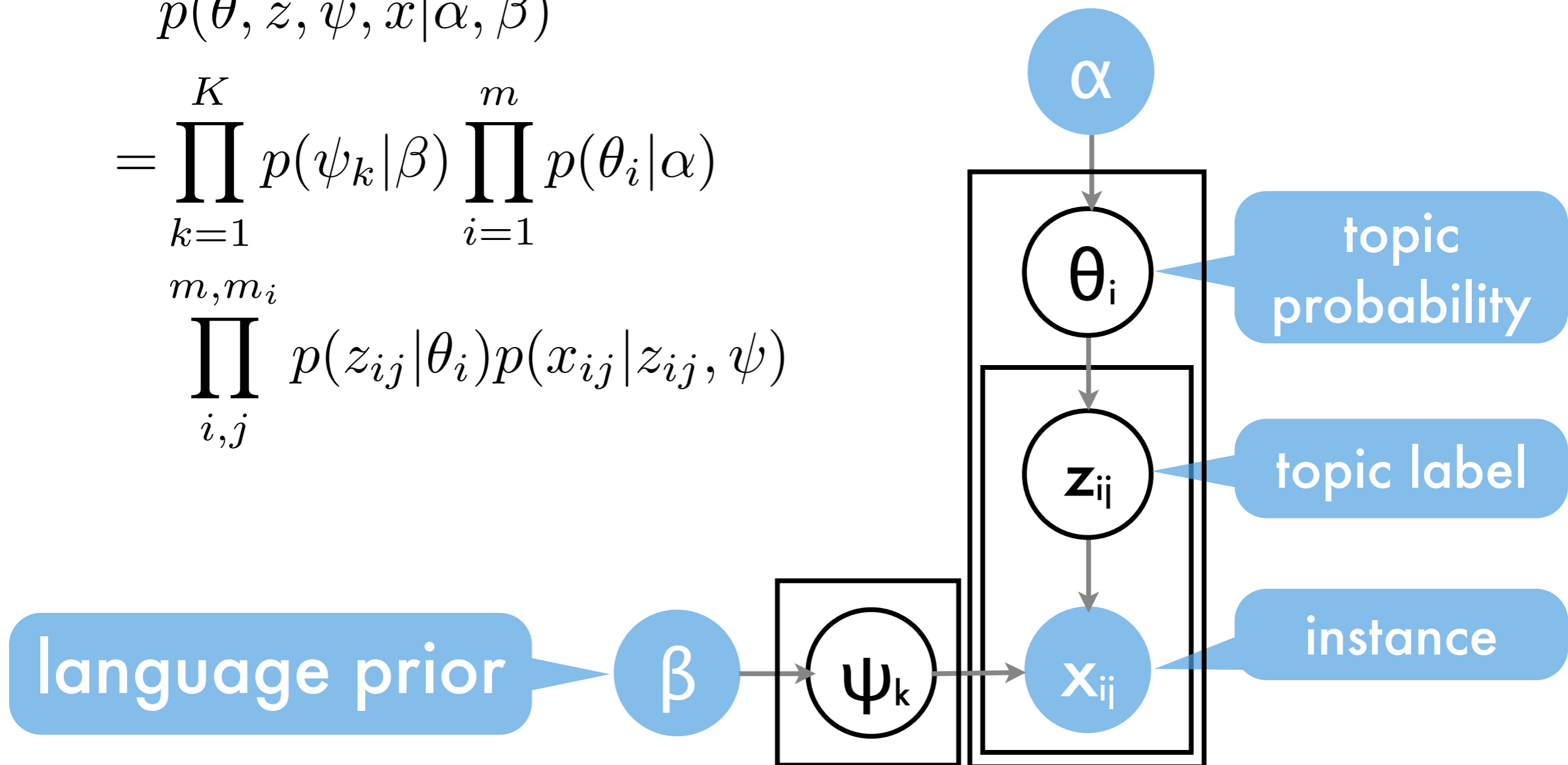
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as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

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

Collapsed Gibbs Sampler

Joint Probability Distribution

$$\begin{aligned} & p(\theta, z, \psi, x | \alpha, \beta) \\ &= \prod_{k=1}^K p(\psi_k | \beta) \prod_{i=1}^m p(\theta_i | \alpha) \\ & \quad \prod_{i,j} p(z_{ij} | \theta_i) p(x_{ij} | z_{ij}, \psi) \end{aligned}$$



Joint Probability Distribution

sample Ψ
independently

$$p(\theta, z, \psi, x | \alpha, \beta)$$

$$= \prod_{k=1}^K p(\psi_k | \beta) \prod_{i=1}^m p(\theta_i | \alpha)$$

sample θ
independently

$$\prod_{i,j} p(z_{ij} | \theta_i) p(x_{ij} | z_{ij}, \psi)$$

sample z
independently

language prior

β

Ψ_k

α

θ_i

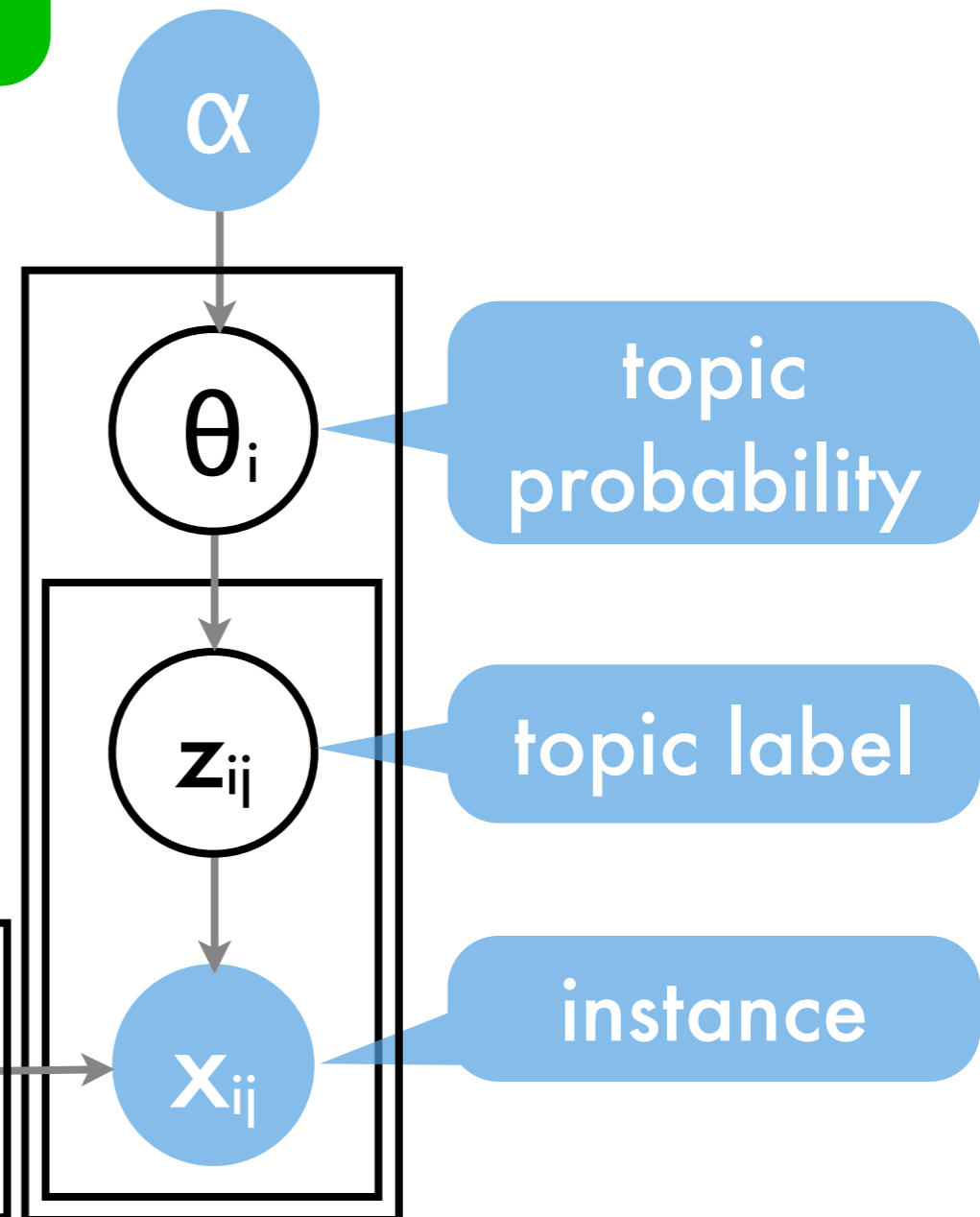
z_{ij}

x_{ij}

topic probability

topic label

instance



Joint Probability Distribution

sample Ψ
independently

$$p(\theta, z, \psi, x | \alpha, \beta)$$
$$= \prod_{k=1}^K p(\psi_k | \beta) \prod_{i=1}^m p(\theta_i | \alpha)$$
$$\prod_{i,j} p(z_{ij} | \theta_i) p(x_{ij} | z_{ij}, \psi)$$

sample θ
independently

sample z
independently

language prior

β

Ψ_k

α

θ_i

z_{ij}

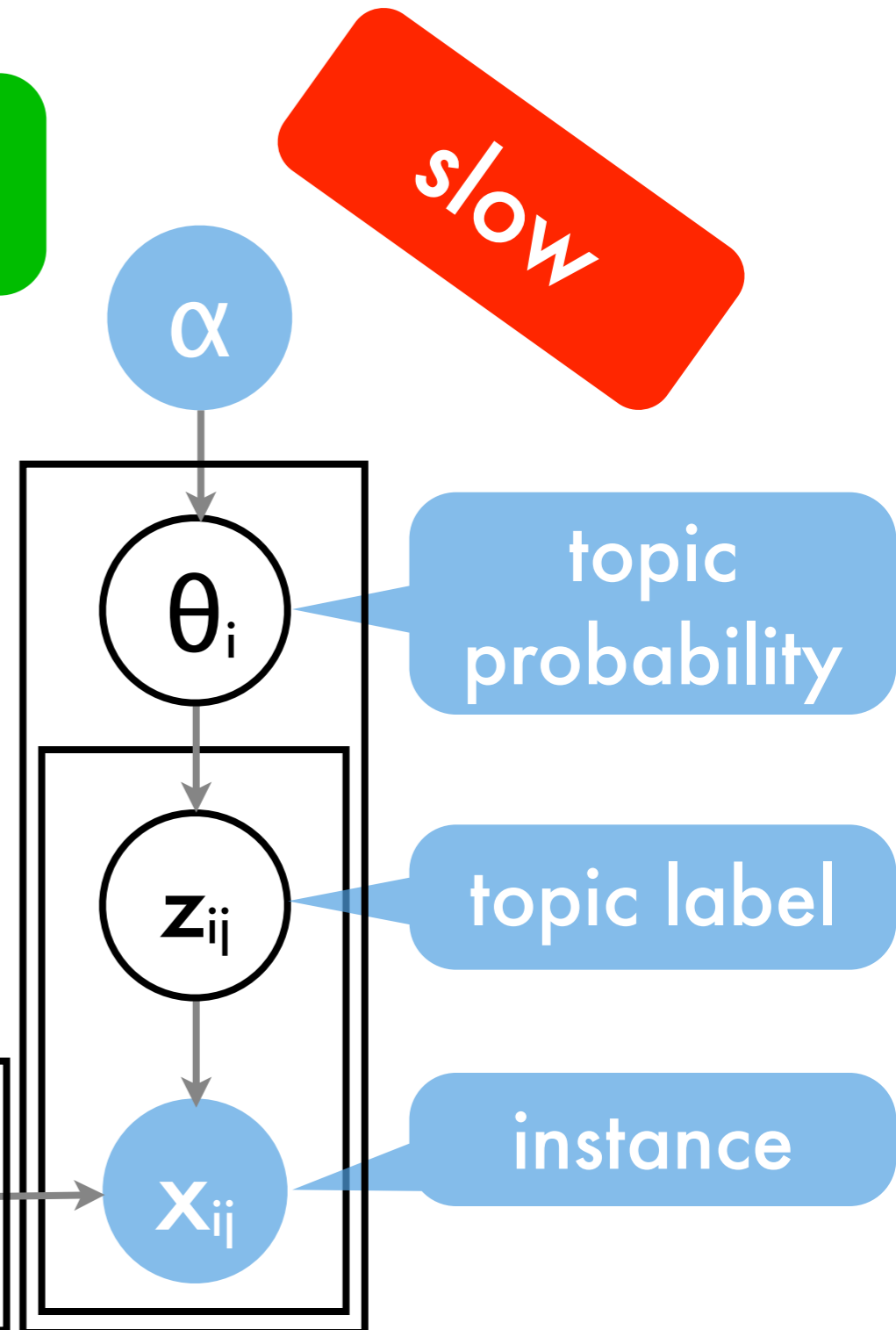
x_{ij}

slow

topic
probability

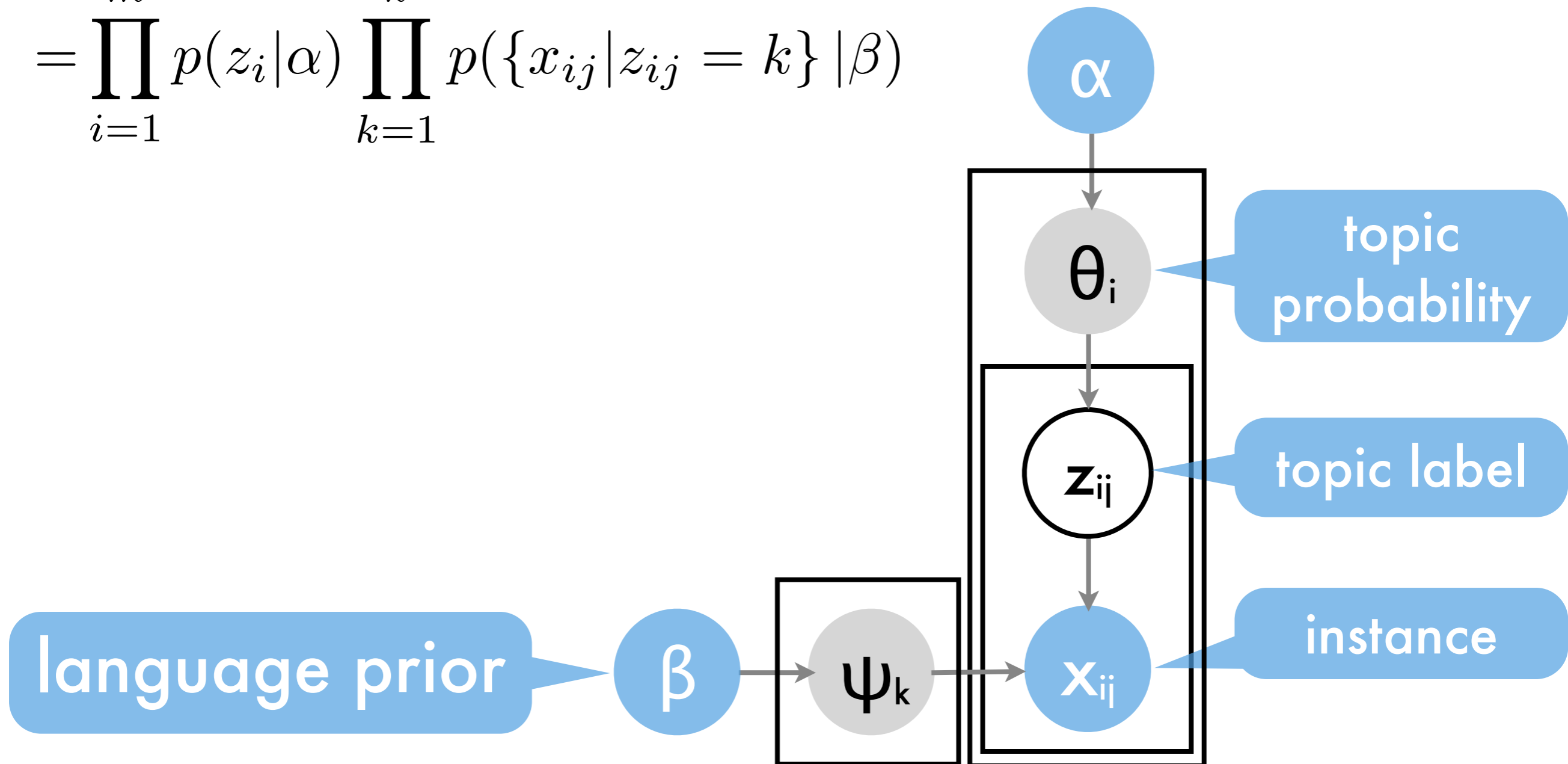
topic label

instance



Collapsed Sampler

$$p(z, x | \alpha, \beta)$$
$$= \prod_{i=1}^m p(z_i | \alpha) \prod_{k=1}^K p(\{x_{ij} | z_{ij} = k\} | \beta)$$



Collapsed Sampler

$$p(z, x | \alpha, \beta) = \prod_{i=1}^m p(z_i | \alpha) \prod_{k=1}^K p(\{x_{ij} | z_{ij} = k\} | \beta)$$

sample z
sequentially

language prior

β

ψ_k

x_{ij}

z_{ij}

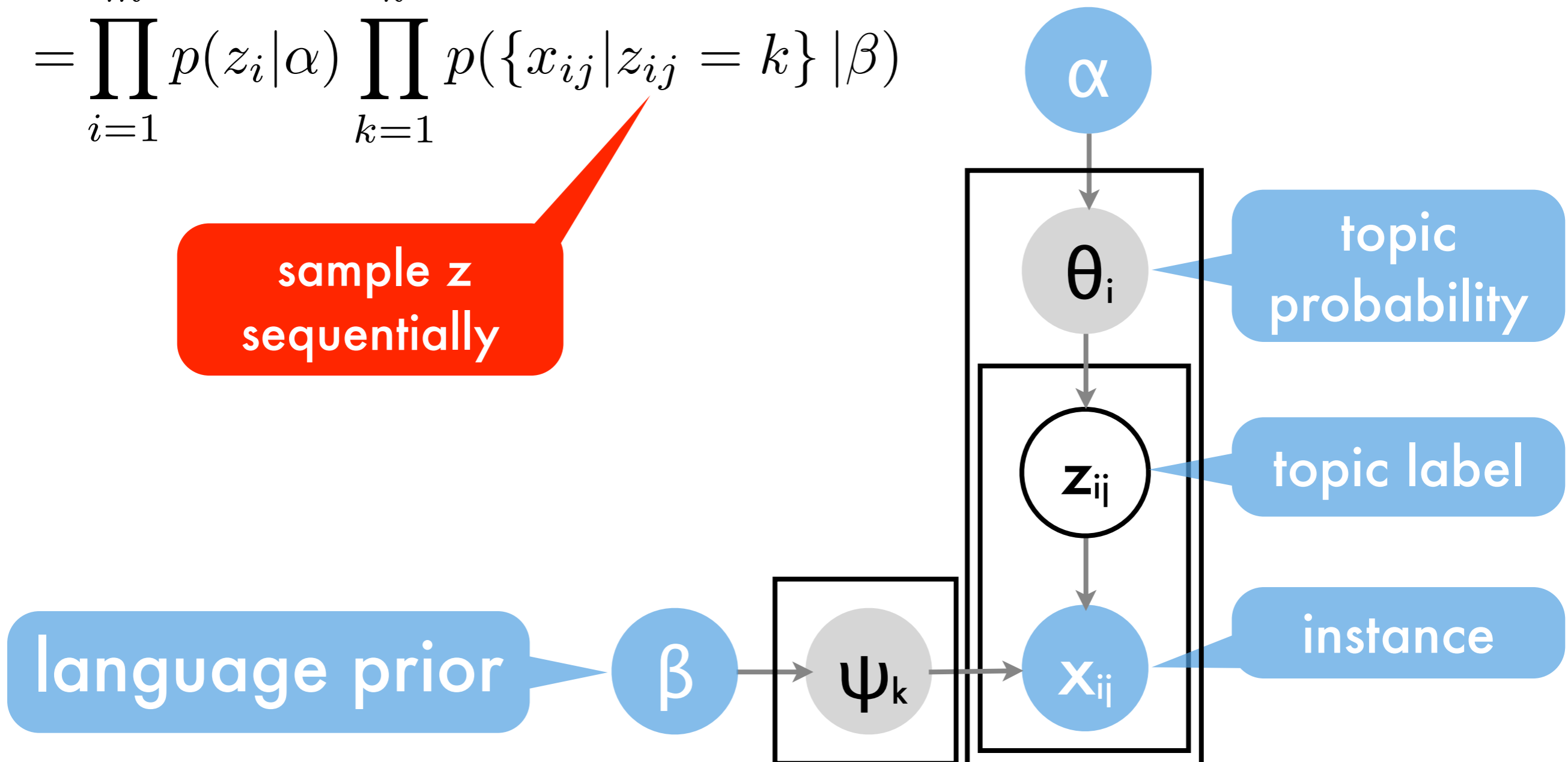
θ_i

α

topic probability

topic label

instance



Collapsed Sampler

$$p(z, x | \alpha, \beta) = \prod_{i=1}^m p(z_i | \alpha) \prod_{k=1}^K p(\{x_{ij} | z_{ij} = k\} | \beta)$$

sample z
sequentially

fast

language prior

β

ψ_k

x_{ij}

z_{ij}

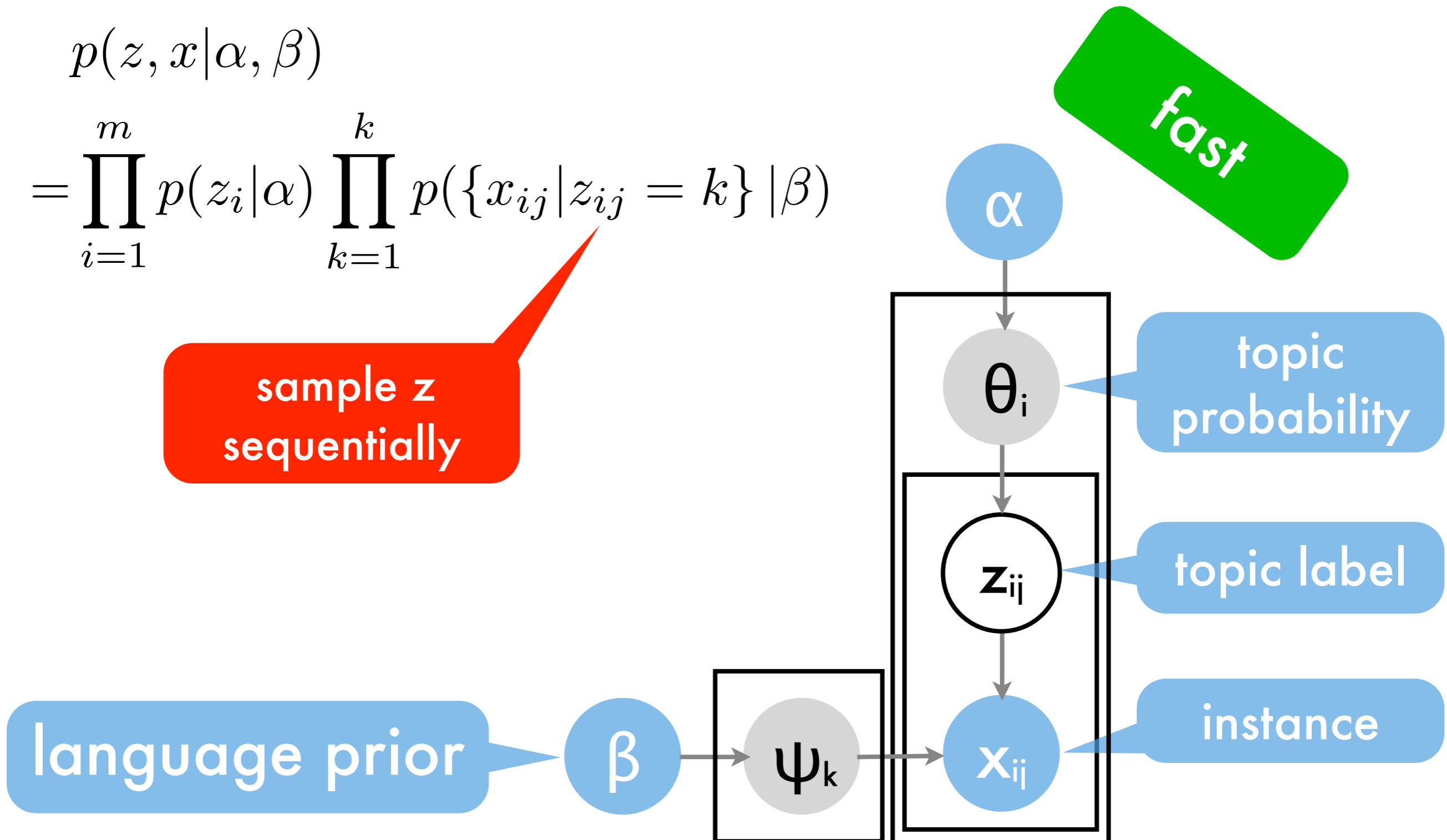
θ_i

α

topic probability

topic label

instance



Collapsed Sampler

Griffiths & Steyvers, 2005

$$p(z, x | \alpha, \beta)$$

$$= \prod_{i=1}^m p(z_i | \alpha) \prod_{k=1}^k p(\{x_{ij} | z_{ij} = k\} | \beta)$$

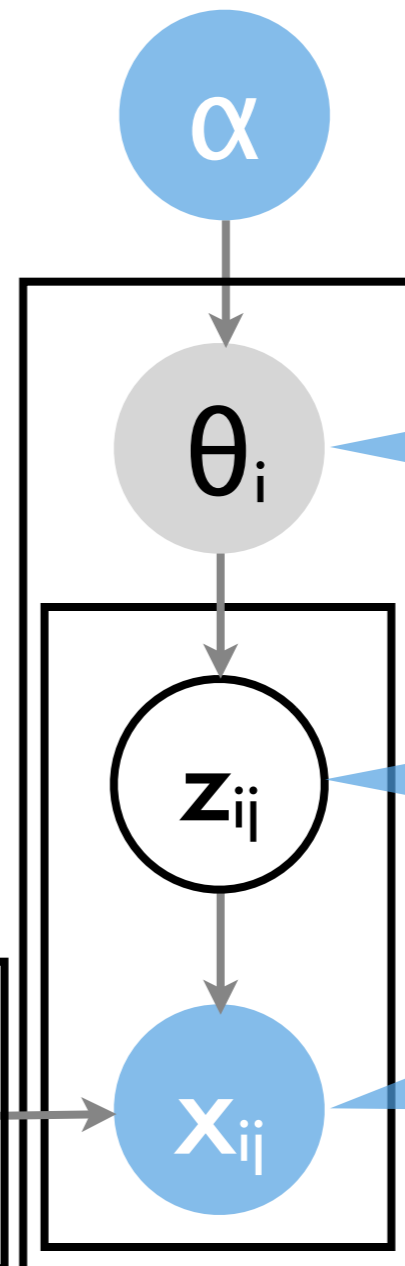
$$\frac{n^{-ij}(t, d) + \alpha_t}{n^{-i}(d) + \sum_t \alpha_t}$$

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

language prior

β

ψ_k



topic probability

topic label

instance

Collapsed Sampler

Griffiths & Steyvers, 2005

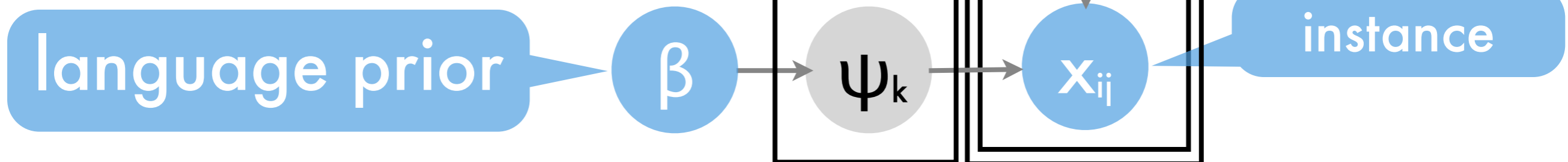
$$p(z, x | \alpha, \beta)$$

$$= \prod_{i=1}^m p(z_i | \alpha) \prod_{k=1}^k p(\{x_{ij} | z_{ij} = k\} | \beta)$$

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

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

language prior



Sequential Algorithm (Gibbs sampler)

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

Sequential Algorithm (Gibbs sampler)

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

this kills parallelism

State of the art

UMass Mallet, UC Irvine, Google

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

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

State of the art

UMass Mallet, UC Irvine, Google

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

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

slow

State of the art

UMass Mallet, UC Irvine, Google

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

changes rapidly

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

slow

State of the art

UMass Mallet, UC Irvine, Google

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

changes rapidly

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

slow

moderately fast

State of the art

UMass Mallet, UC Irvine, Google

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

table out of sync

memory inefficient

blocking

network bound

changes rapidly

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

slow

moderately fast

Our Approach

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

Our Approach

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

network
bound

concurrent
cpu hdd net

Our Approach

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

network
bound

memory
inefficient

concurrent
cpu hdd net

minimal
view

Our Approach

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

network
bound

memory
inefficient

table out
of sync

concurrent
cpu hdd net

minimal
view

continuous
sync

Our Approach

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

network
bound

memory
inefficient

table out
of sync

blocking

concurrent
cpu hdd net

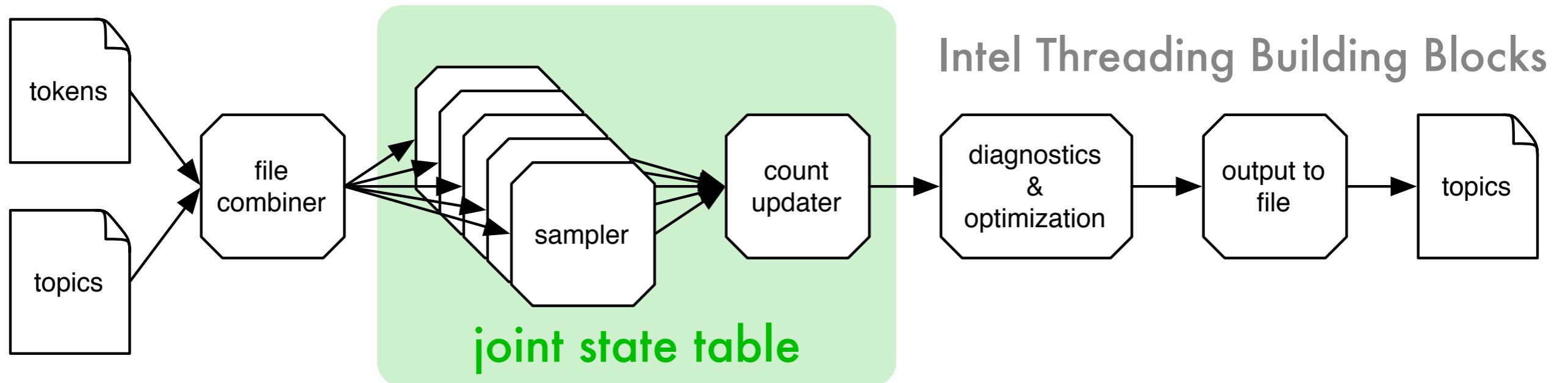
minimal
view

continuous
sync

barrier
free

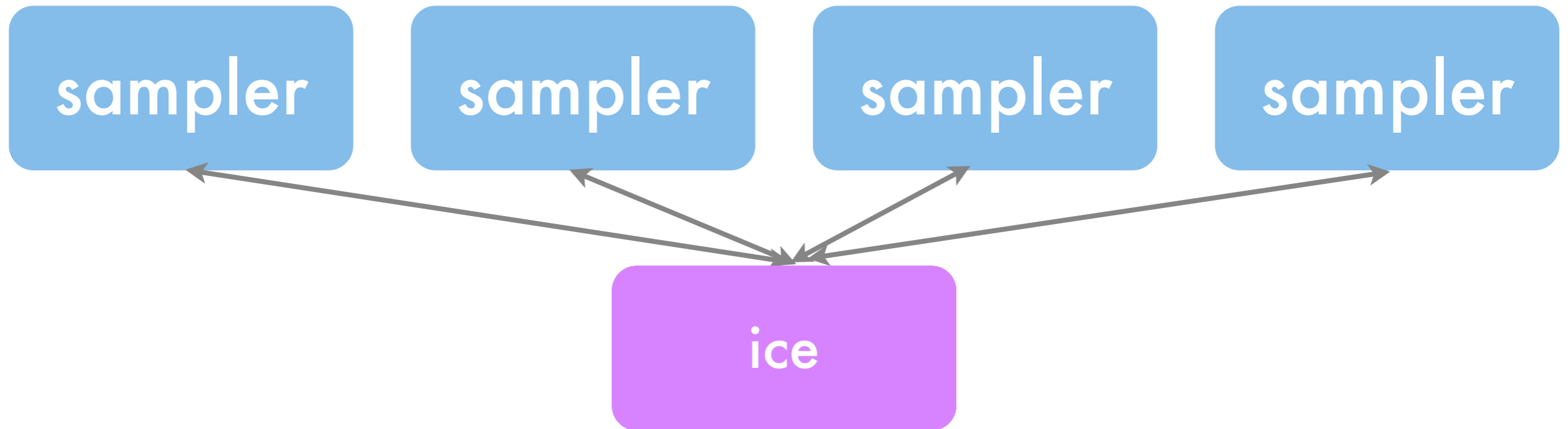
Architecture details

Multicore Architecture



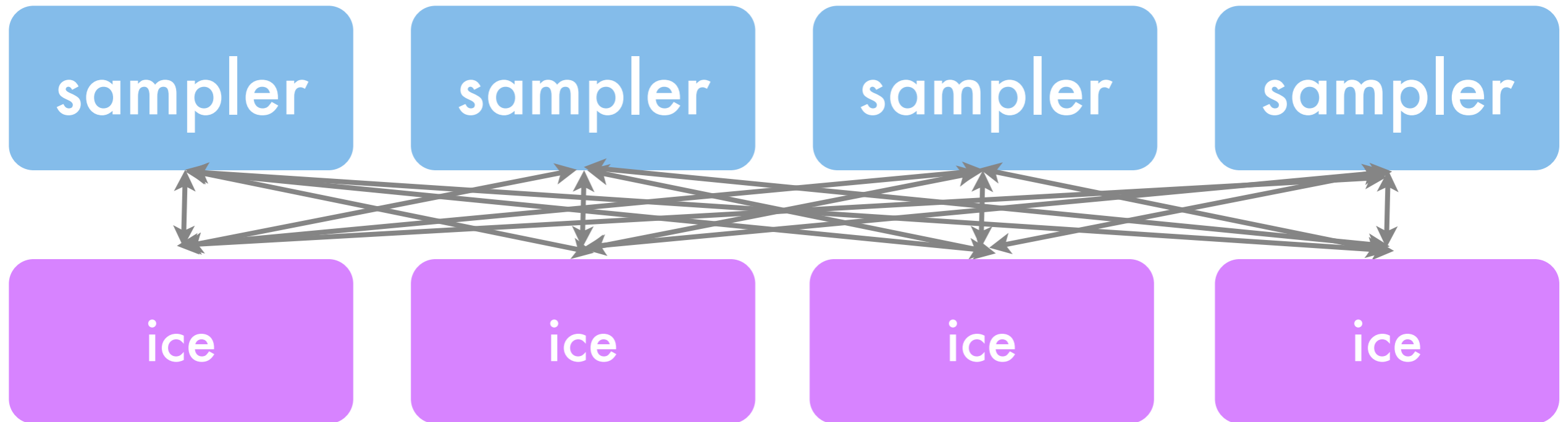
- Decouple multithreaded sampling and updating (almost) avoids stalling for locks in the sampler
- Joint state table
 - much less memory required
 - samplers synchronized (10 docs vs. millions delay)
- Hyperparameter update via stochastic gradient descent
- No need to keep documents in memory (streaming)

Cluster Architecture



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

Cluster Architecture



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

Making it work

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

Easily extensible

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

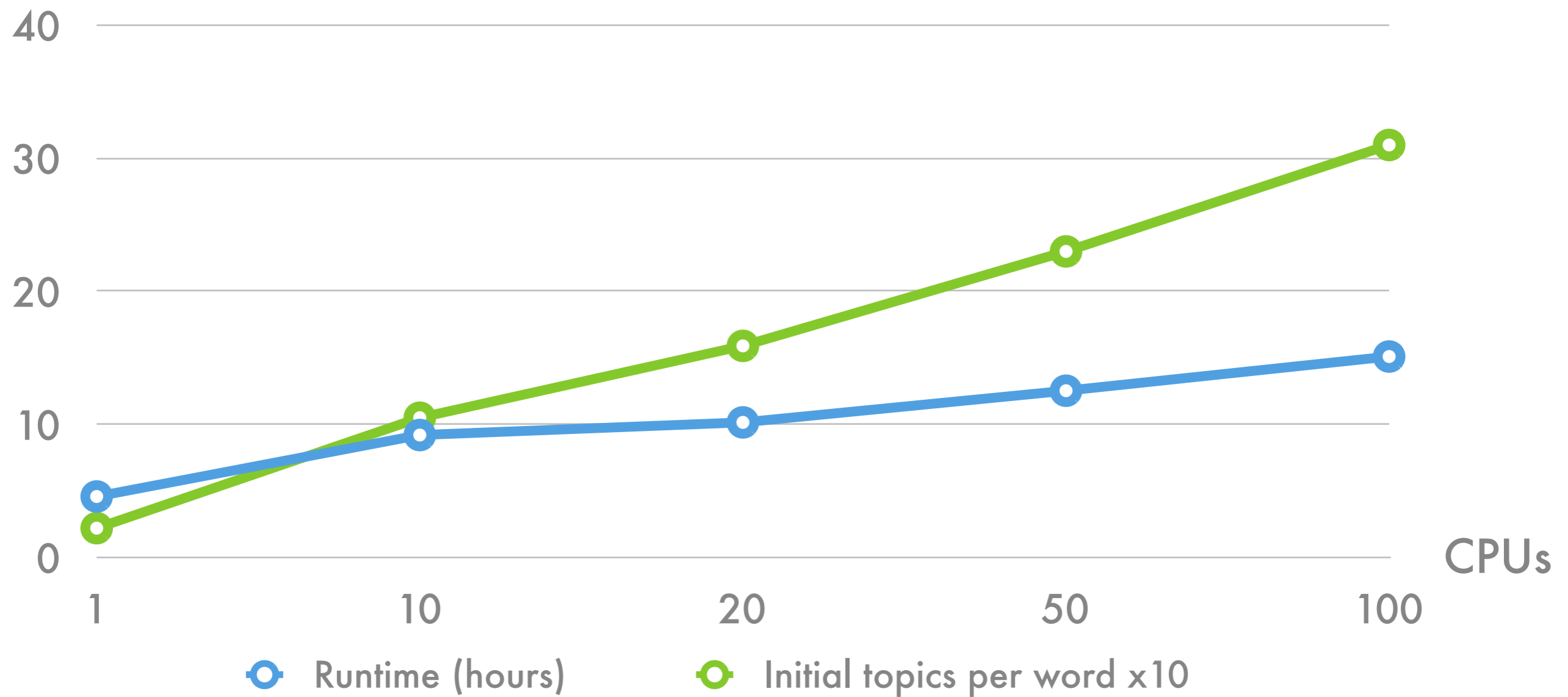
	Google LDA	Mallet	Irvine'08	Irvine'09	Yahoo LDA
Multicore	no	yes	yes	yes	yes
Cluster	MPI	no	MPI	point 2 point	memcached
State table	dictionary split	separate sparse	separate	separate	joint sparse
Schedule	synchronous exact	synchronous exact	synchronous exact	asynchronous approximate messages	asynchronous exact

Speed

- **1M documents per day** on 1 computer
(1000 topics per doc, 1000 words per doc)
- **350k documents per day** per node
(context switches & memcached & stray reducers)
- **8 Million docs** (Pubmed)
(sampler does not burn in well - too short doc)
 - Irvine: **128 machines, 10 hours**
 - Yahoo: **1 machine, 11 days**
 - Yahoo: **20 machines, 9 hours**
- **20 Million docs** (Yahoo! News Articles)
 - Yahoo: **100 machines, 12 hours**

Scalability

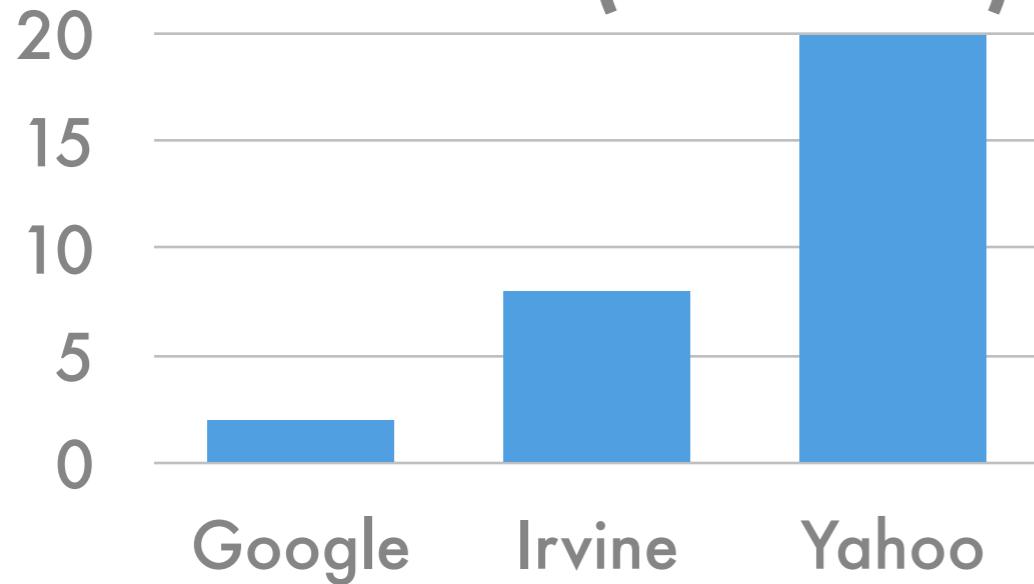
200k documents/computer



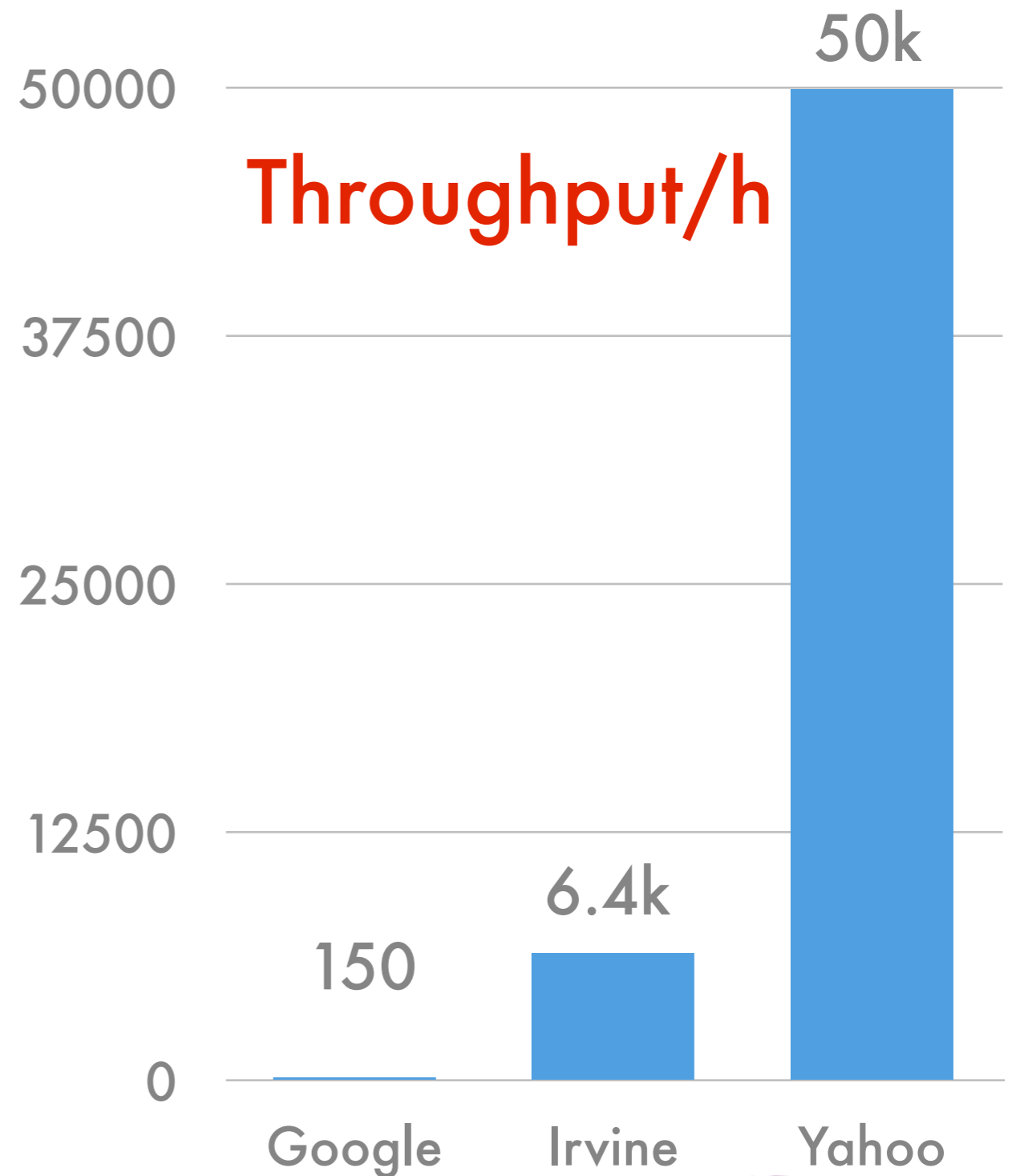
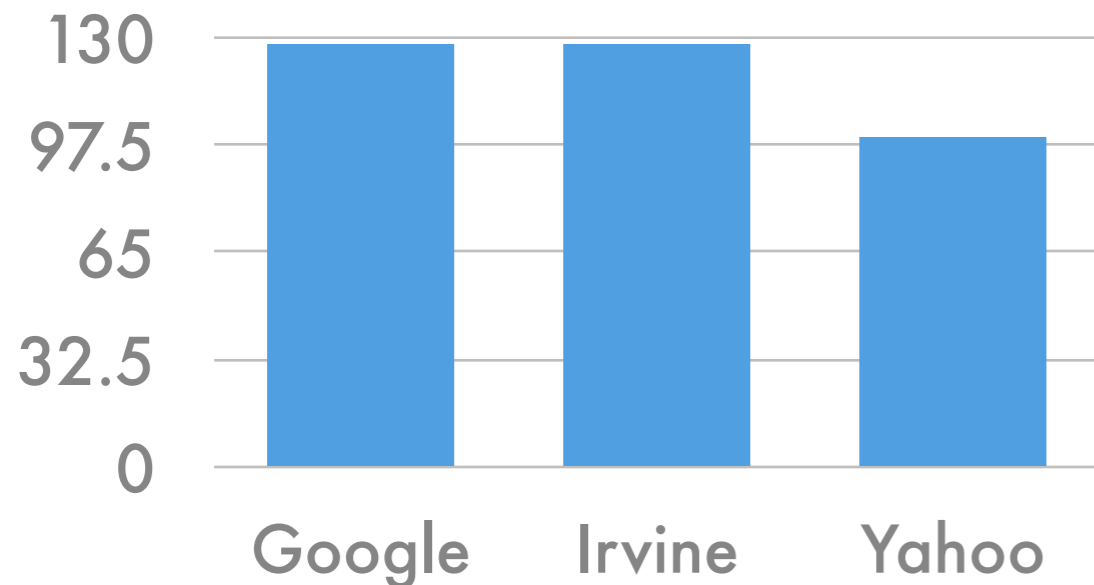
Likelihood even improves with parallelism!
-3.295 (1 node) -3.288 (10 nodes) -3.287 (20 nodes)

The Competition

Dataset size (millions)



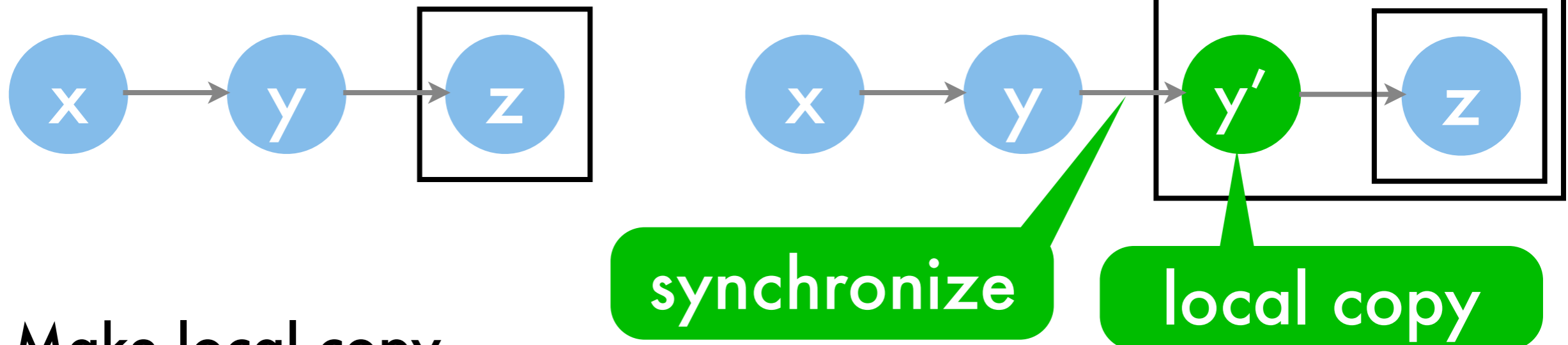
Cluster size



Design Principles

Variable Replication

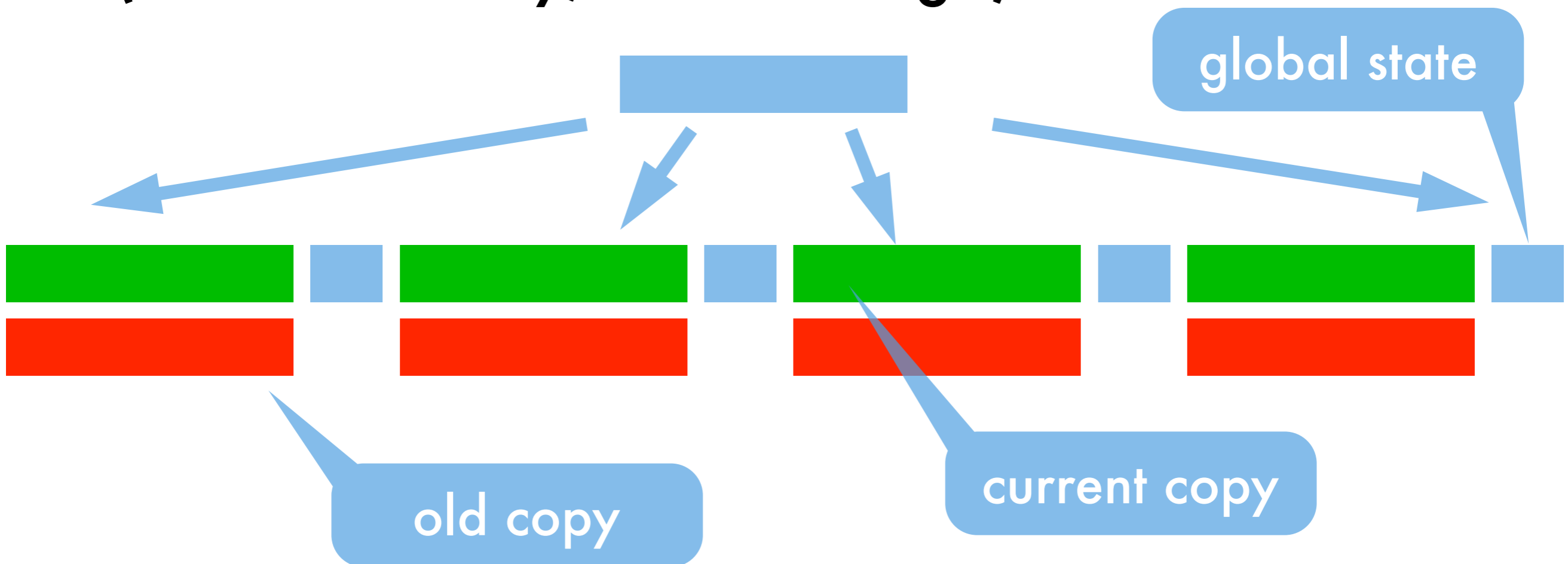
- Global shared variable



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

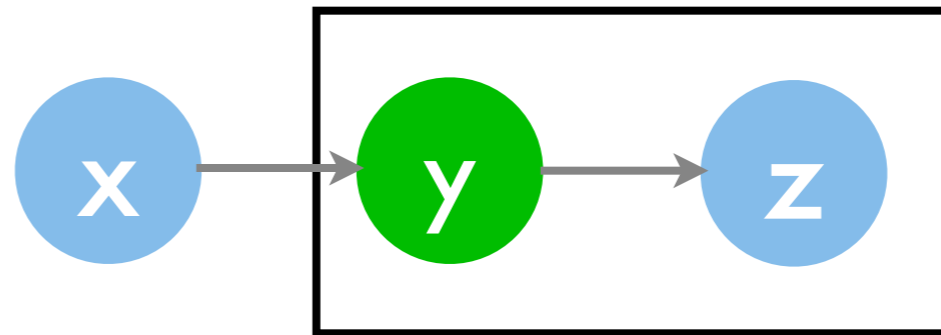
Asymmetric Message Passing

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

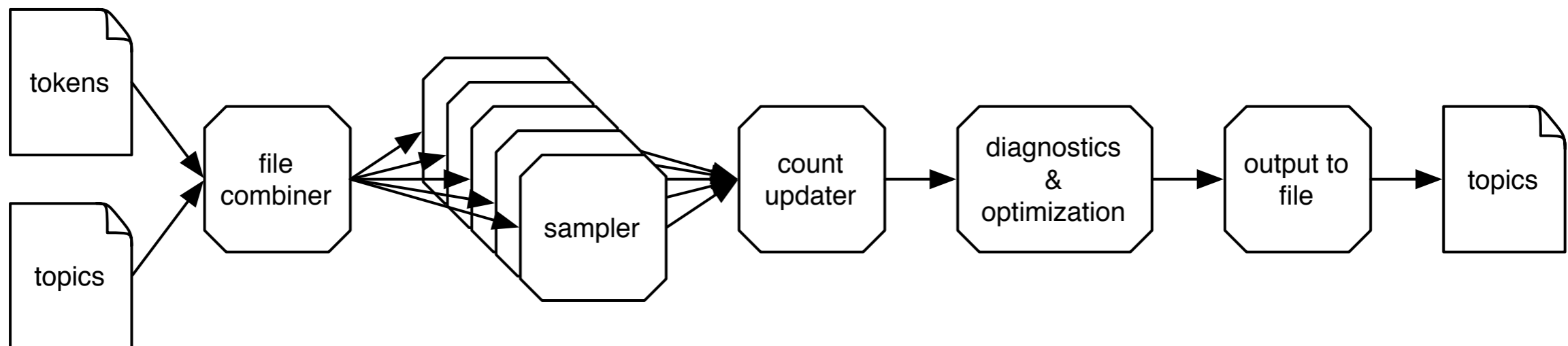


Out of core storage

- Very large state space



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



Summary - Part 3

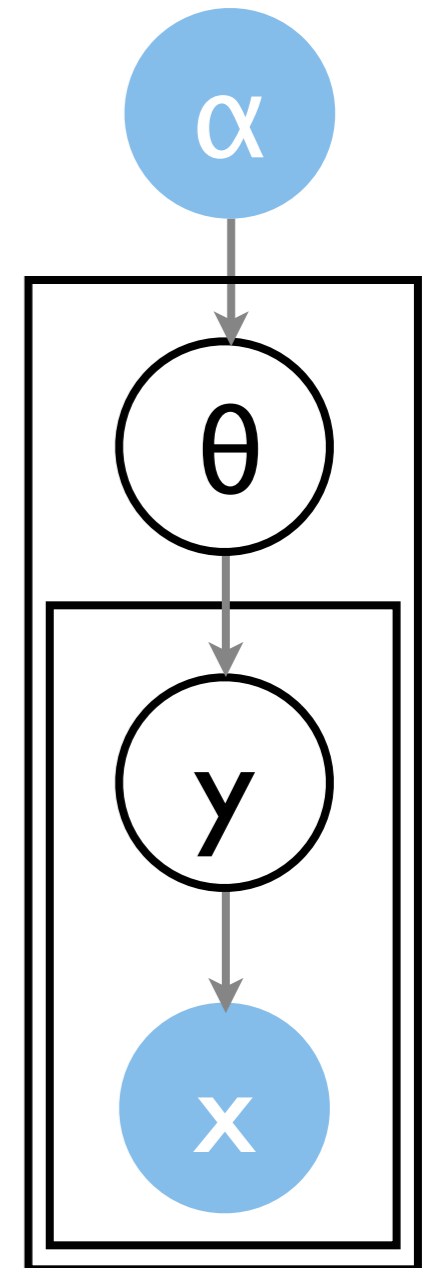
- Inference in graphical models
- Clustering
- Topic models
- Sampling
- Implementation details

Part 4 - Advanced Modeling

Advances in Representation

Extensions to topic models

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



Correlated topic models

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

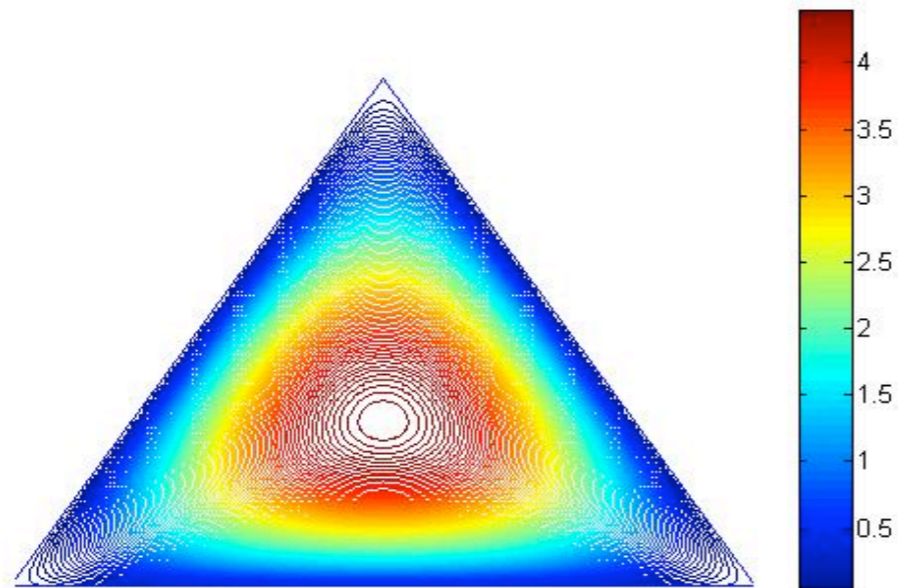
Correlated topic models

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

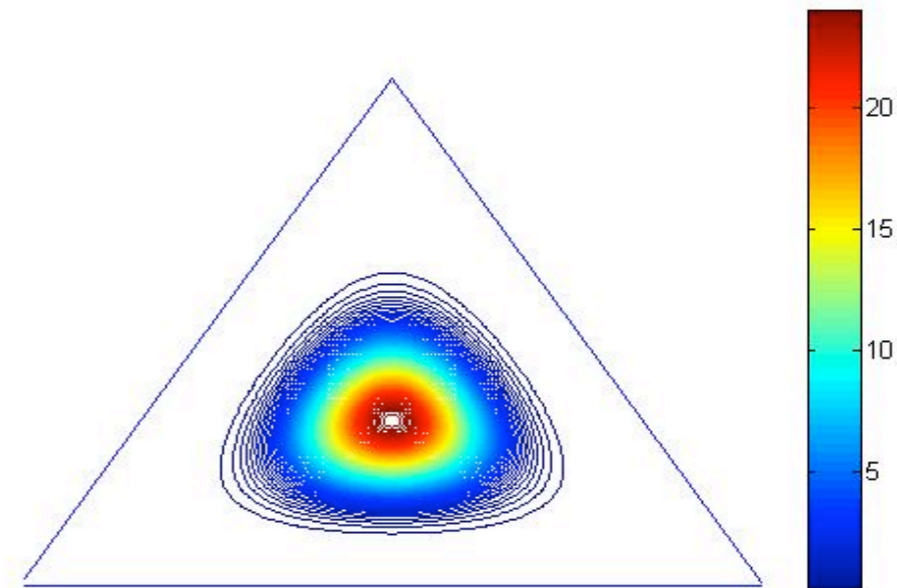
Blei & Lafferty 2005; Ahmed & Xing 2007

Dirichlet prior on topics

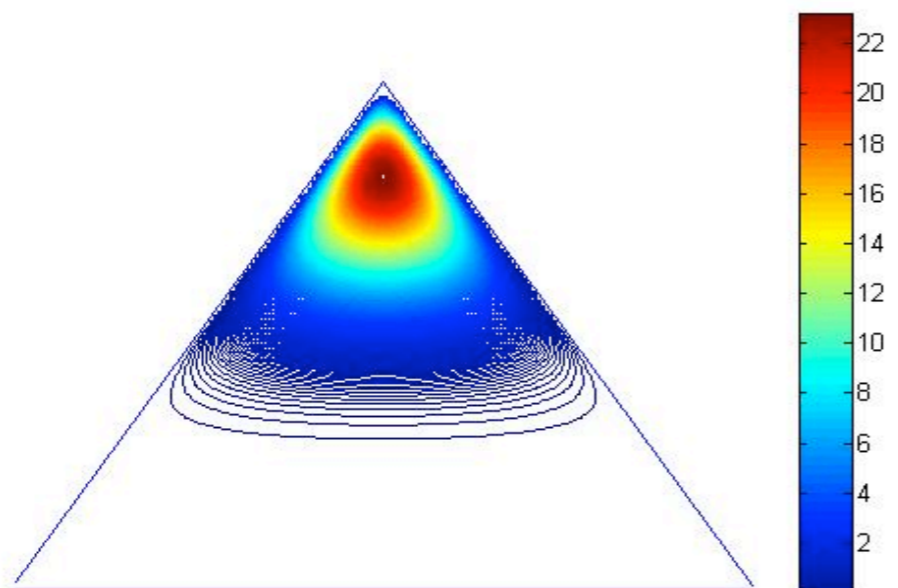
Alpha =[2.00 2.00 2.00]



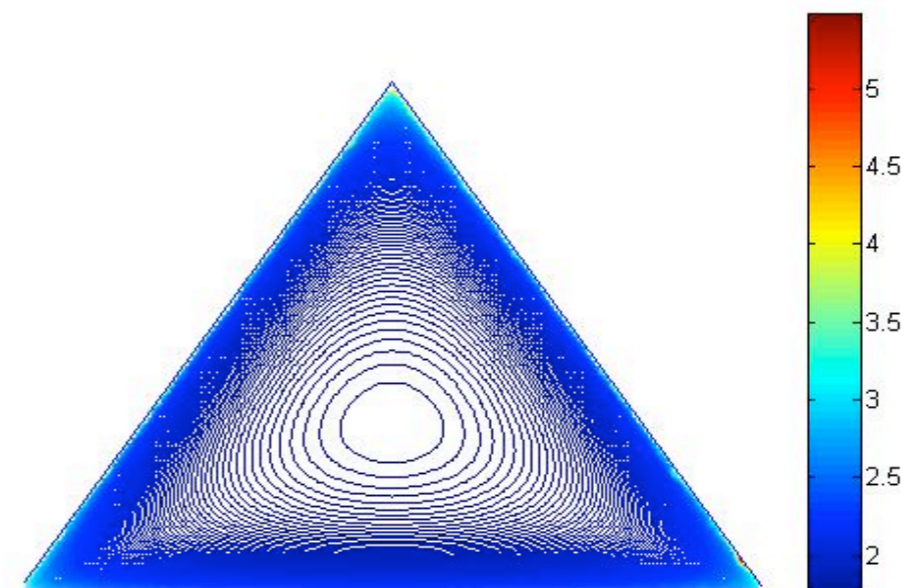
Alpha =[10.00 10.00 10.00]



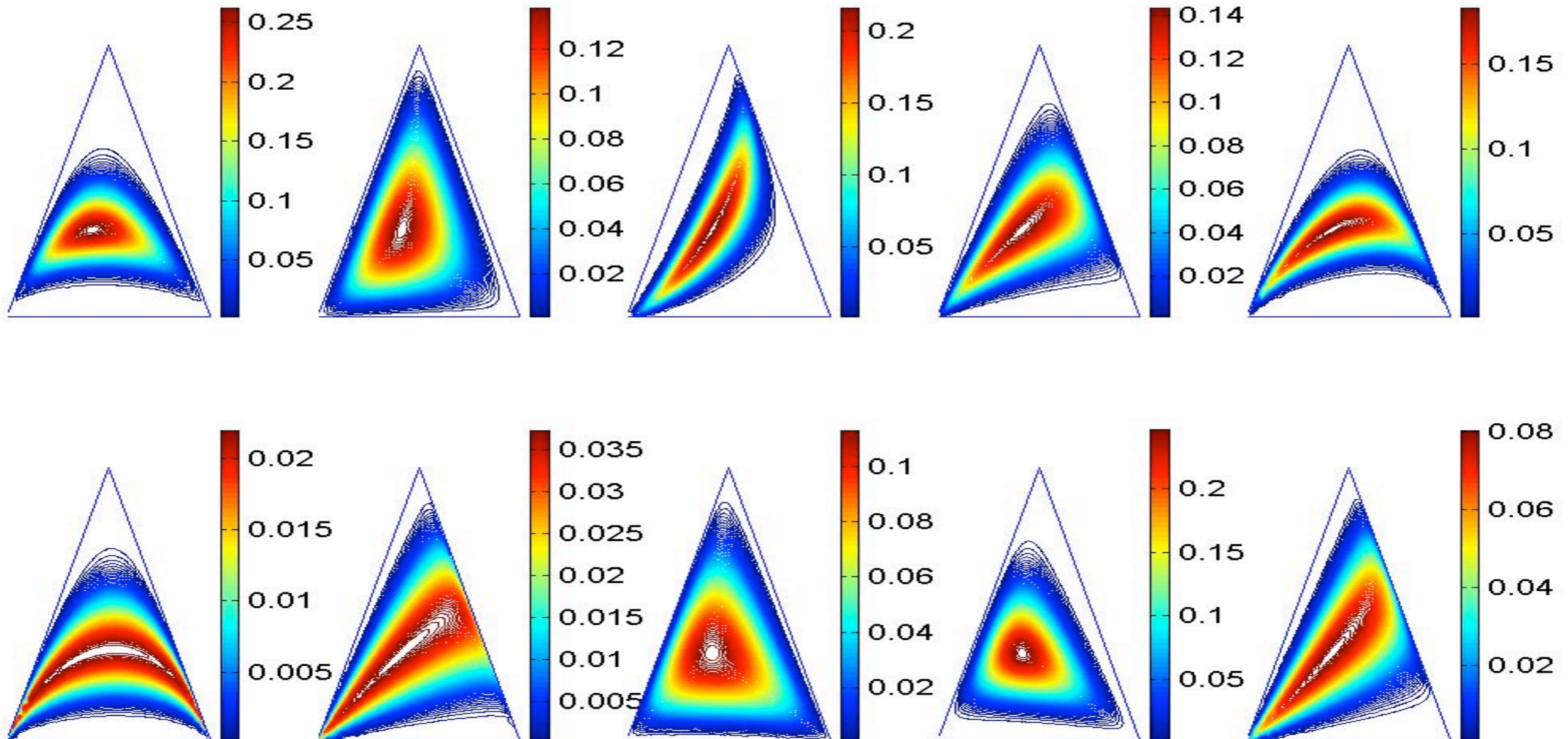
Alpha =[2.00 10.00 2.00]



Alpha =[0.90 0.90 0.90]

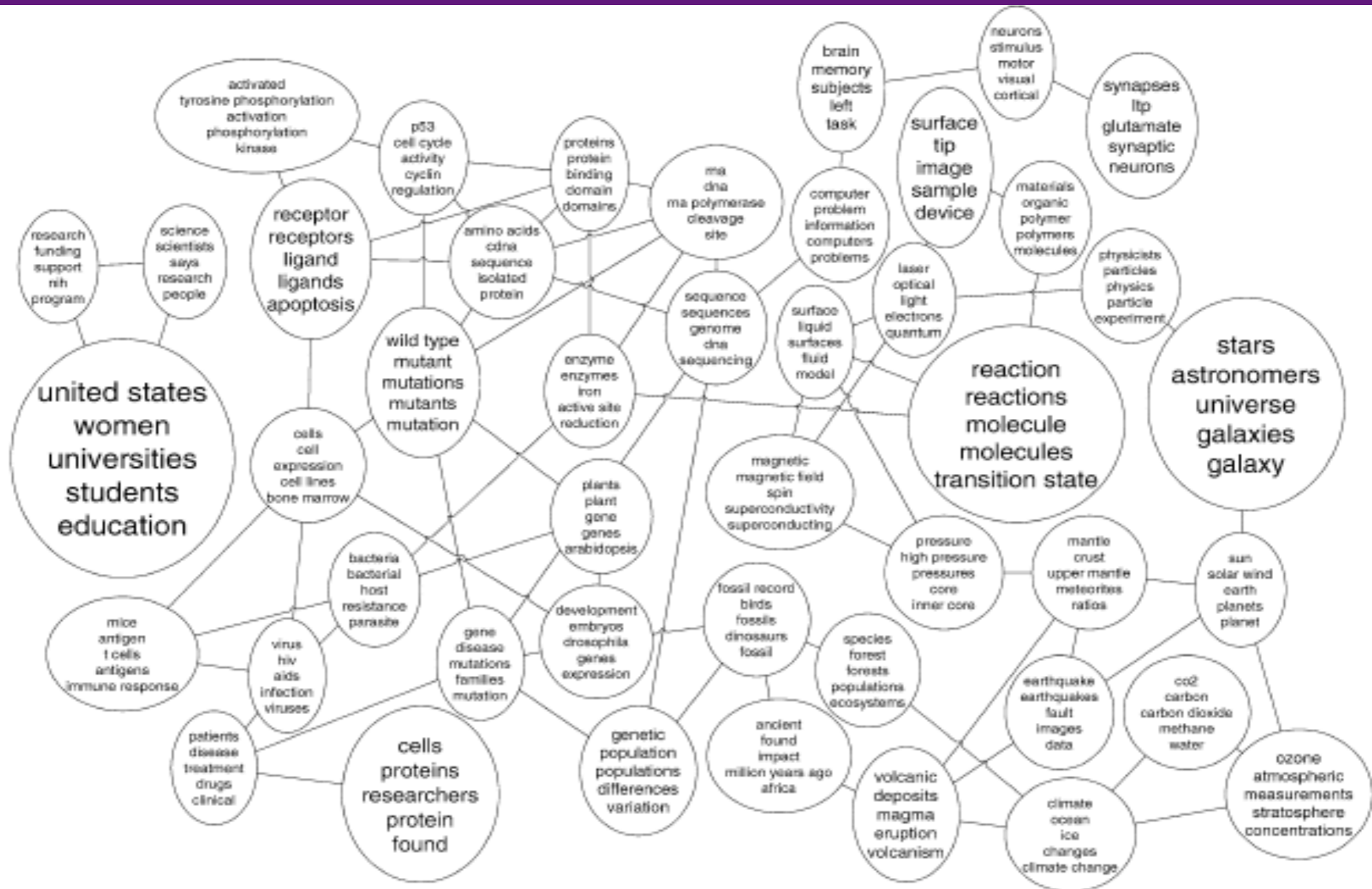


Log-normal prior on topics



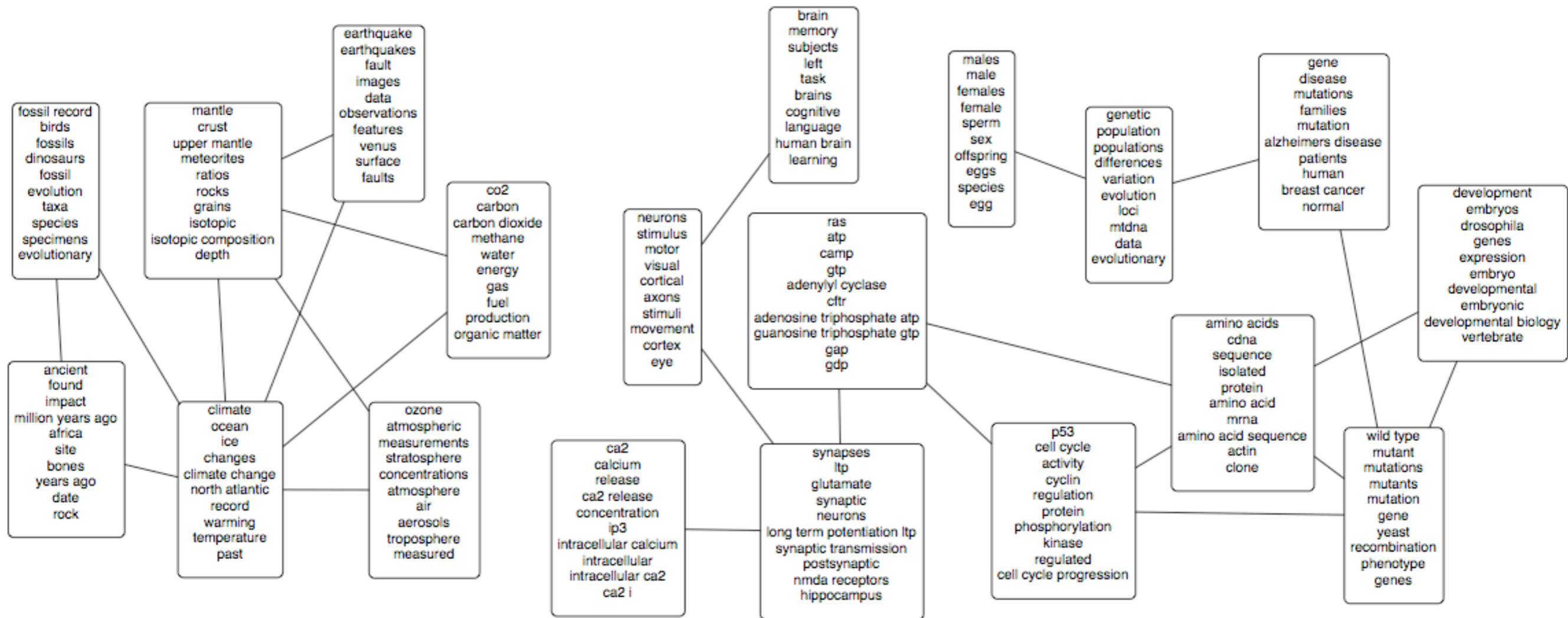
$$\theta = e^{\eta - g(\eta)} \quad \text{with} \quad \eta \sim \mathcal{N}(\mu, \Sigma)$$

Correlated topics



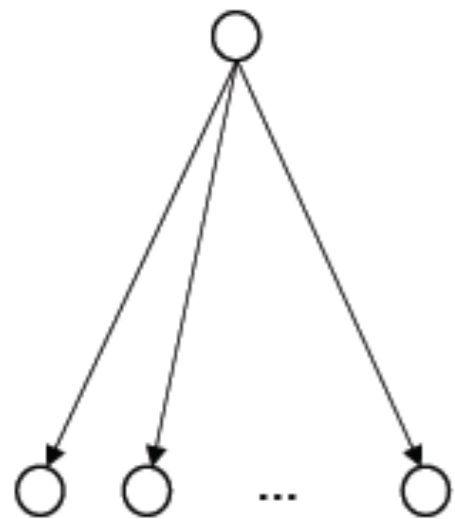
Blei and Lafferty 2005

Correlated topics

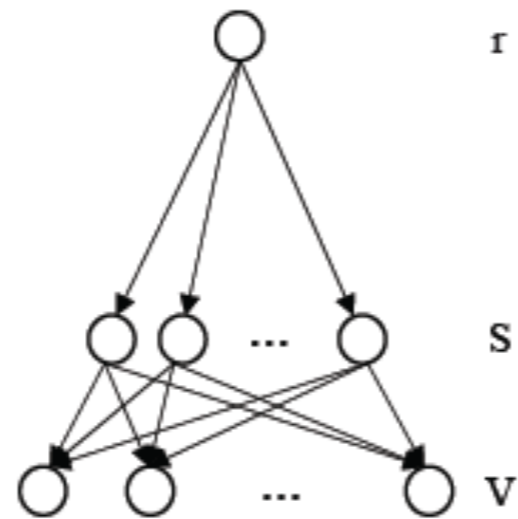


Pachinko Allocation

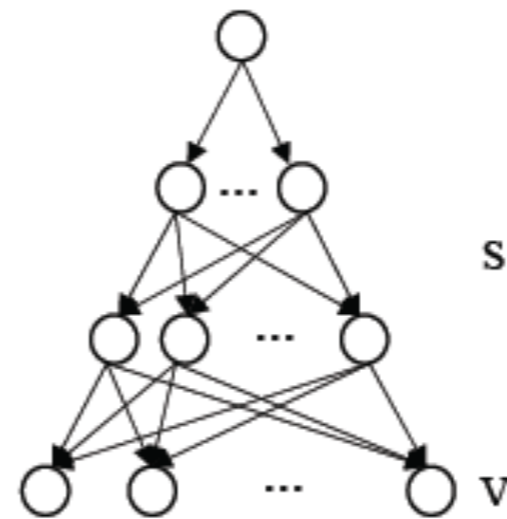
- Model the prior as a Directed Acyclic Graph
- Each document is modeled as multiple paths
- To sample a word, first select a path and then sample a word from the final topic
- The topics reside on the leaves of the tree



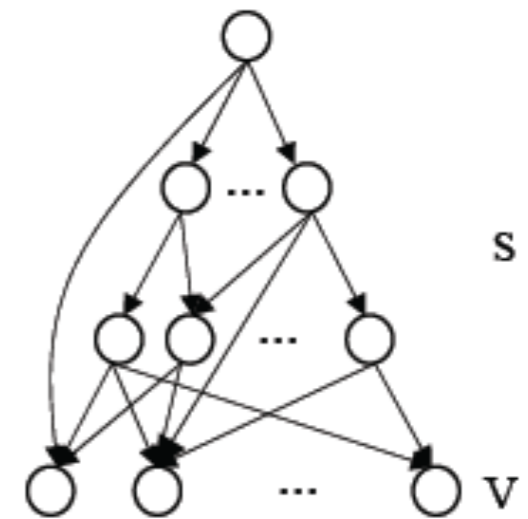
(a) Dirichlet Multinomial



(b) LDA

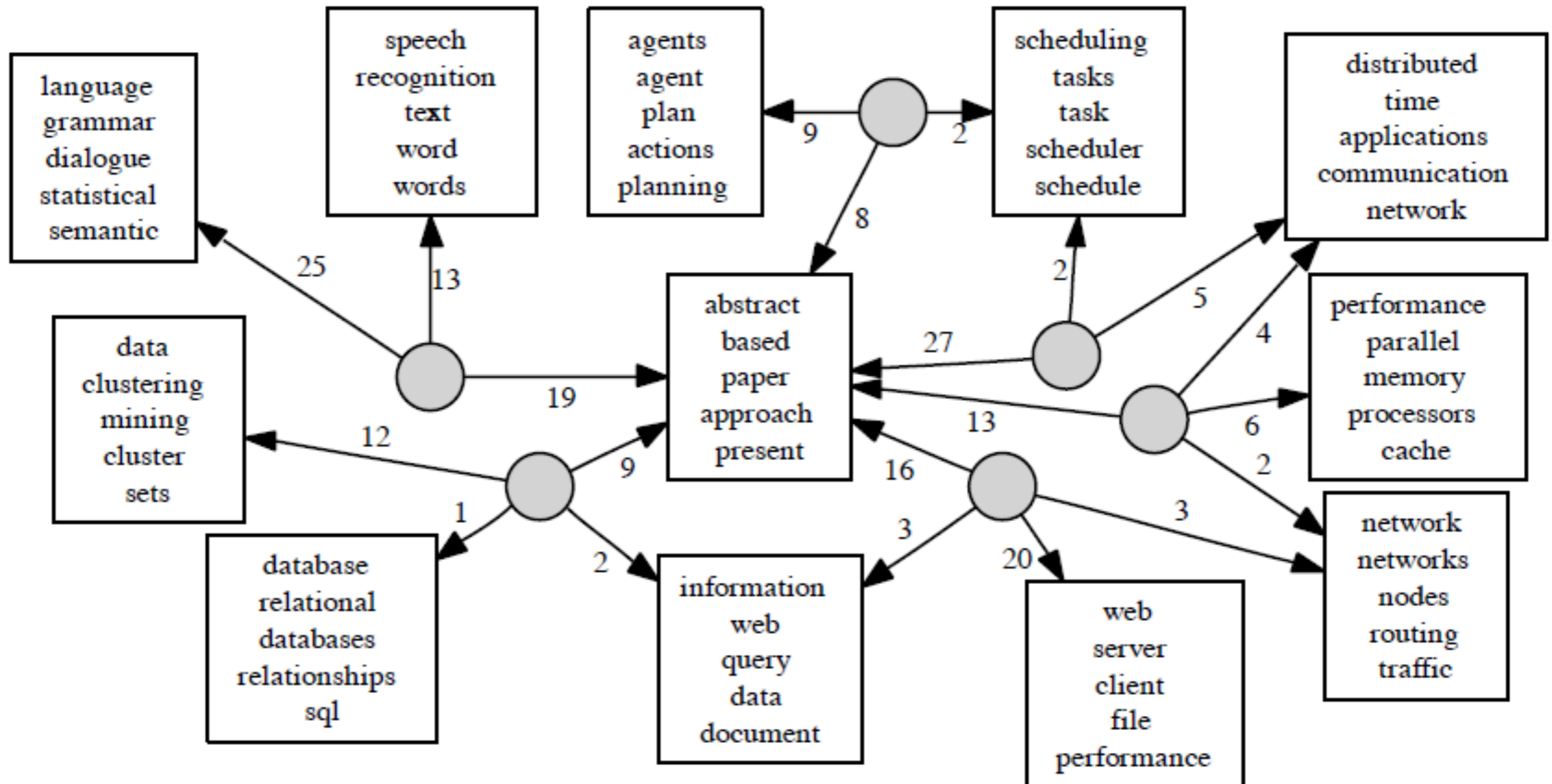


(c) Four-Level PAM



(d) Arbitrary PAM

Pachinko Allocation

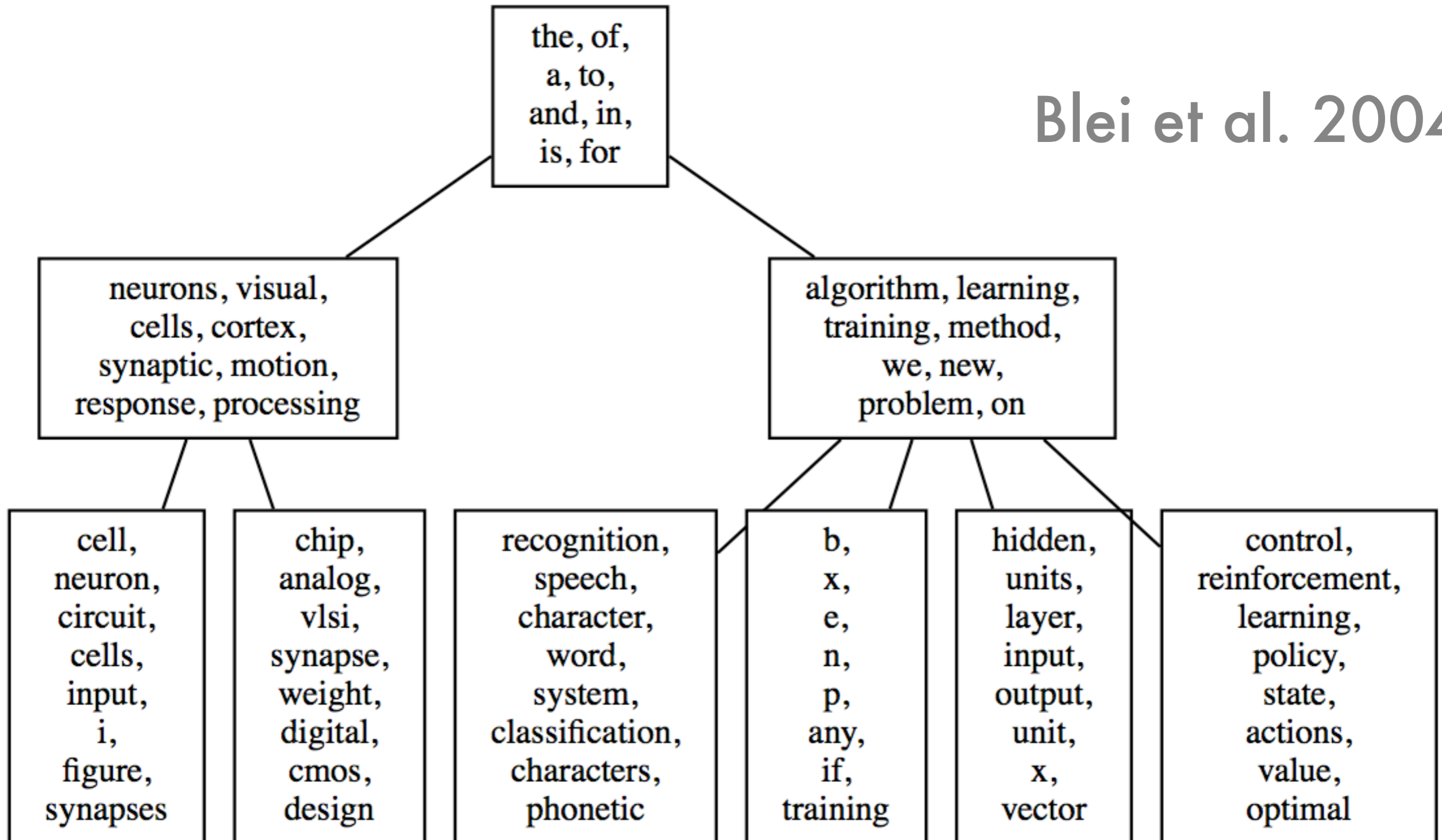


Topic Hierarchies

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

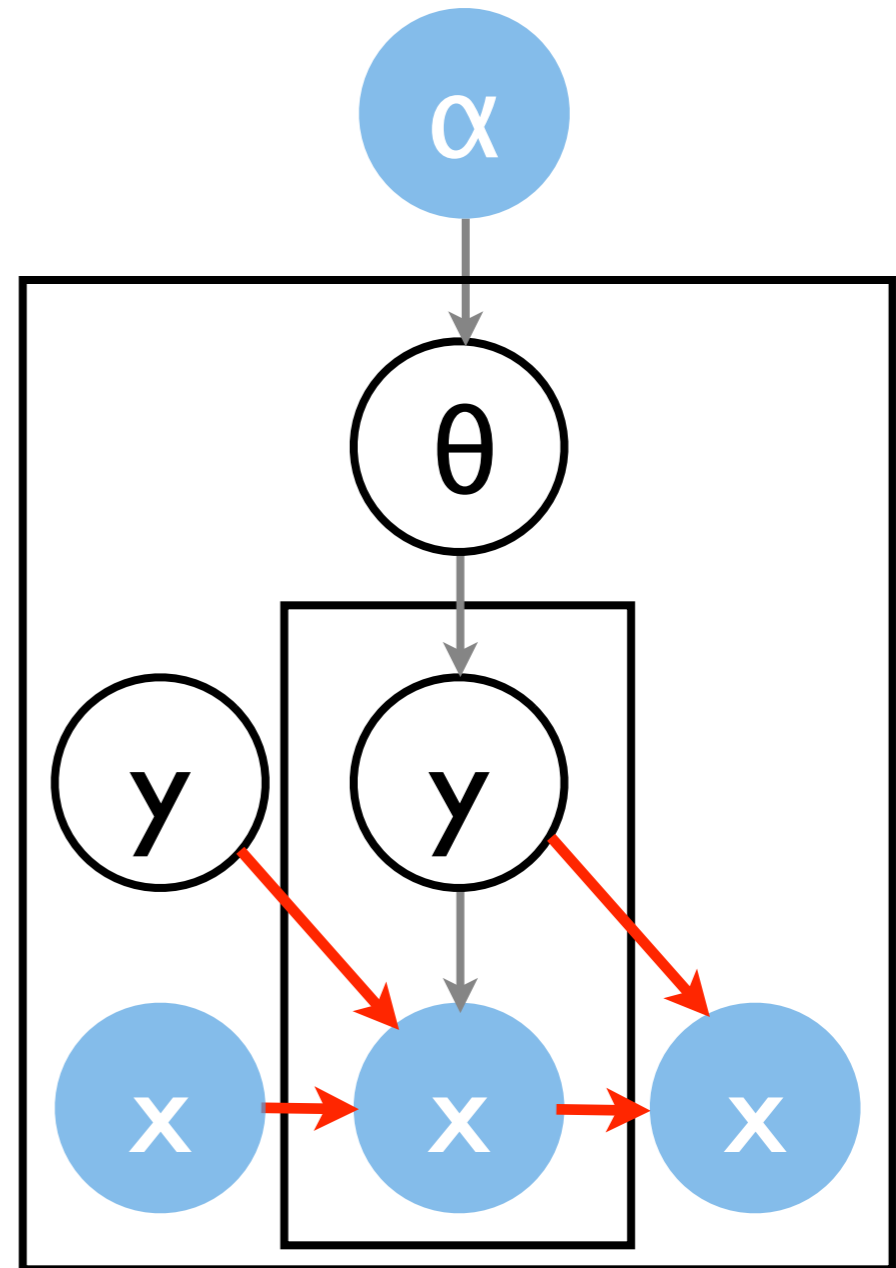
Topic Hierarchies

Blei et al. 2004



Topical n-grams

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



Topic n-grams

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

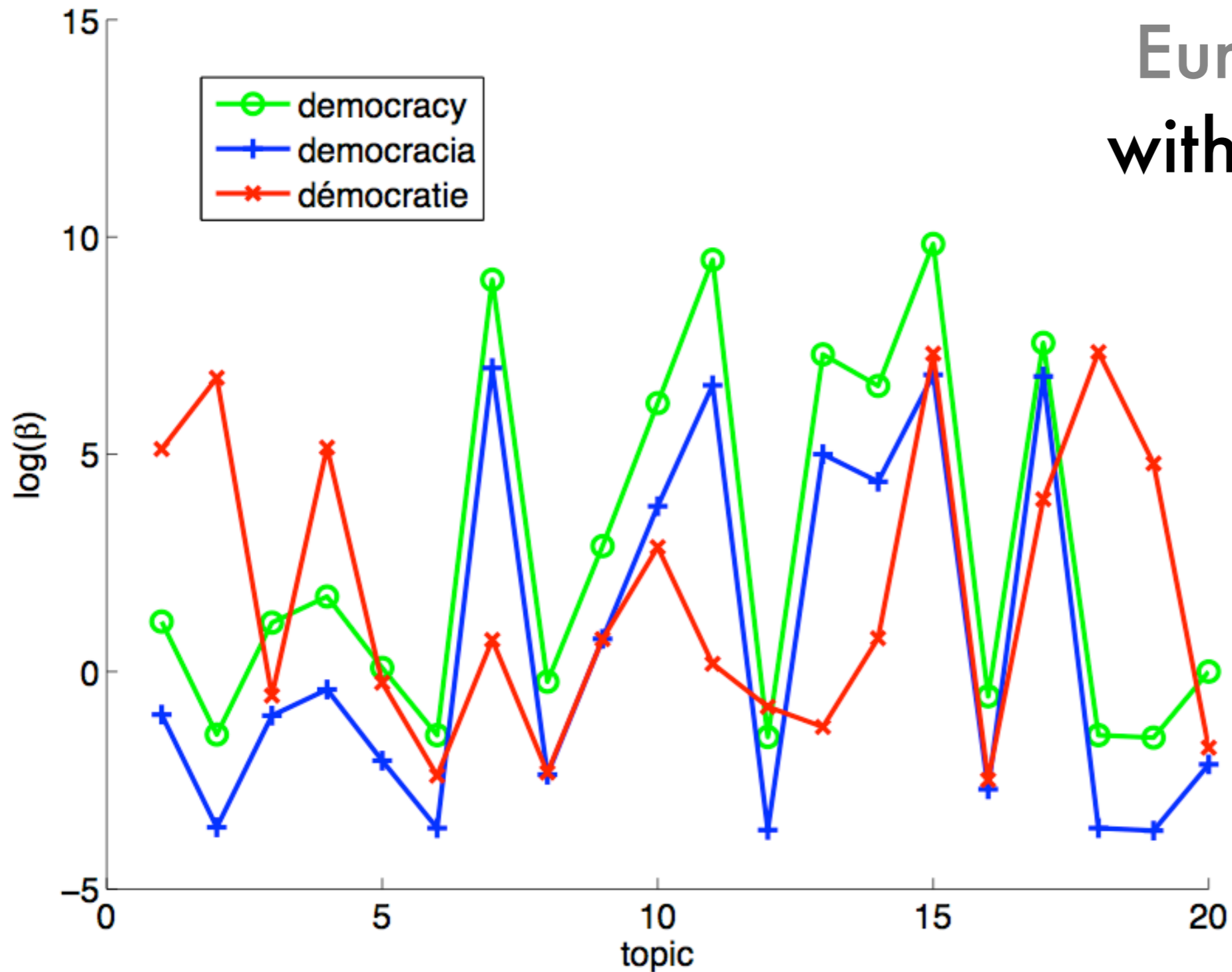
Side information

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

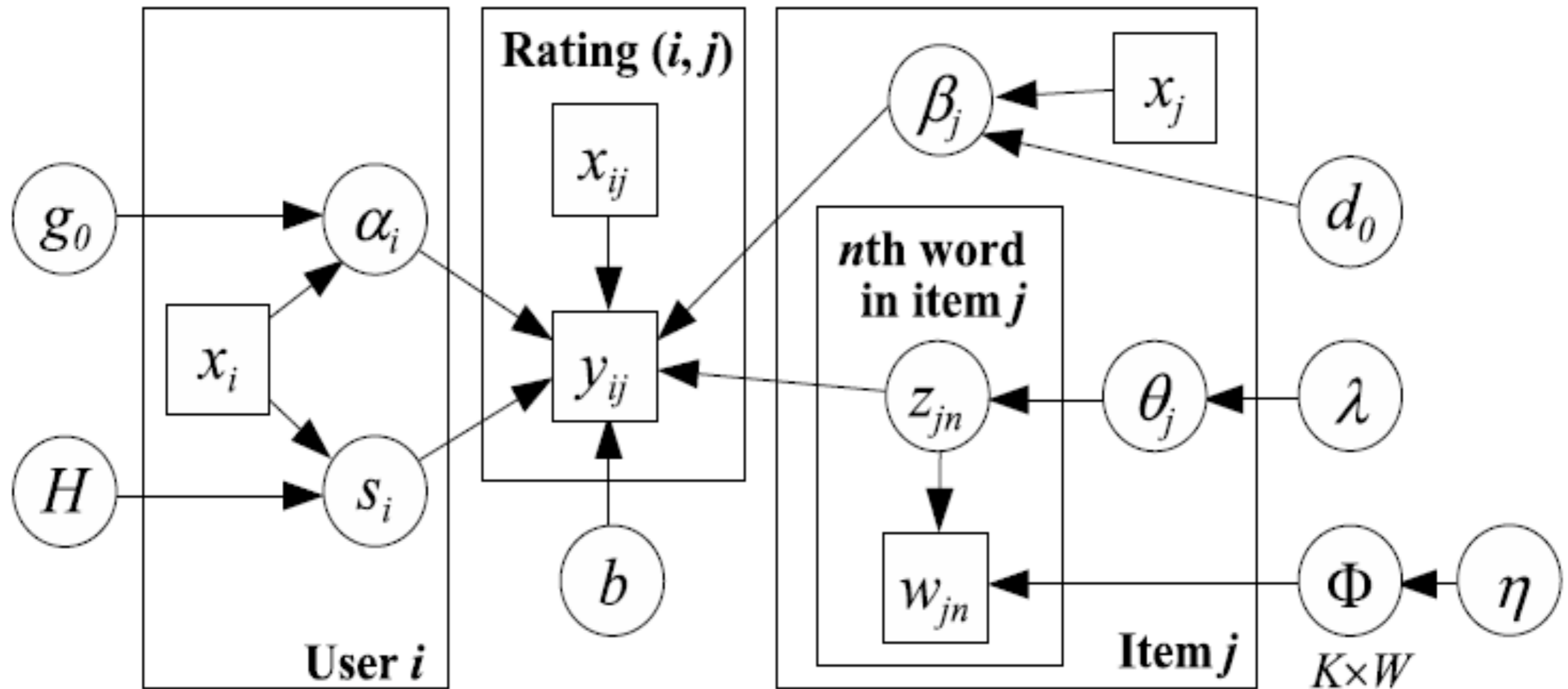
Downstream conditioning

DC

Europarl corpus
without alignment

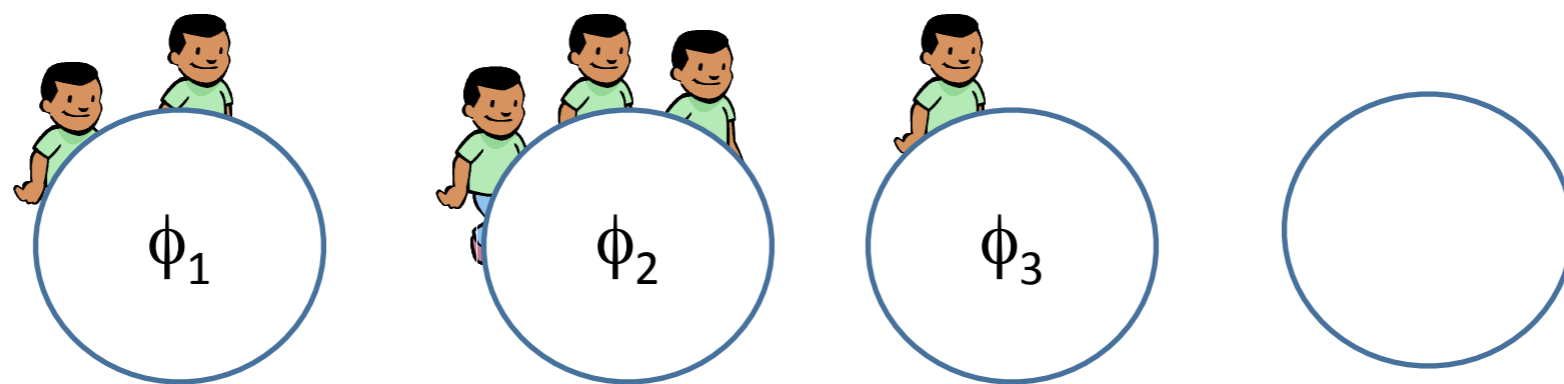


Recommender Systems



Agarwal & Chen, 2010

Chinese Restaurant Process



Problem

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

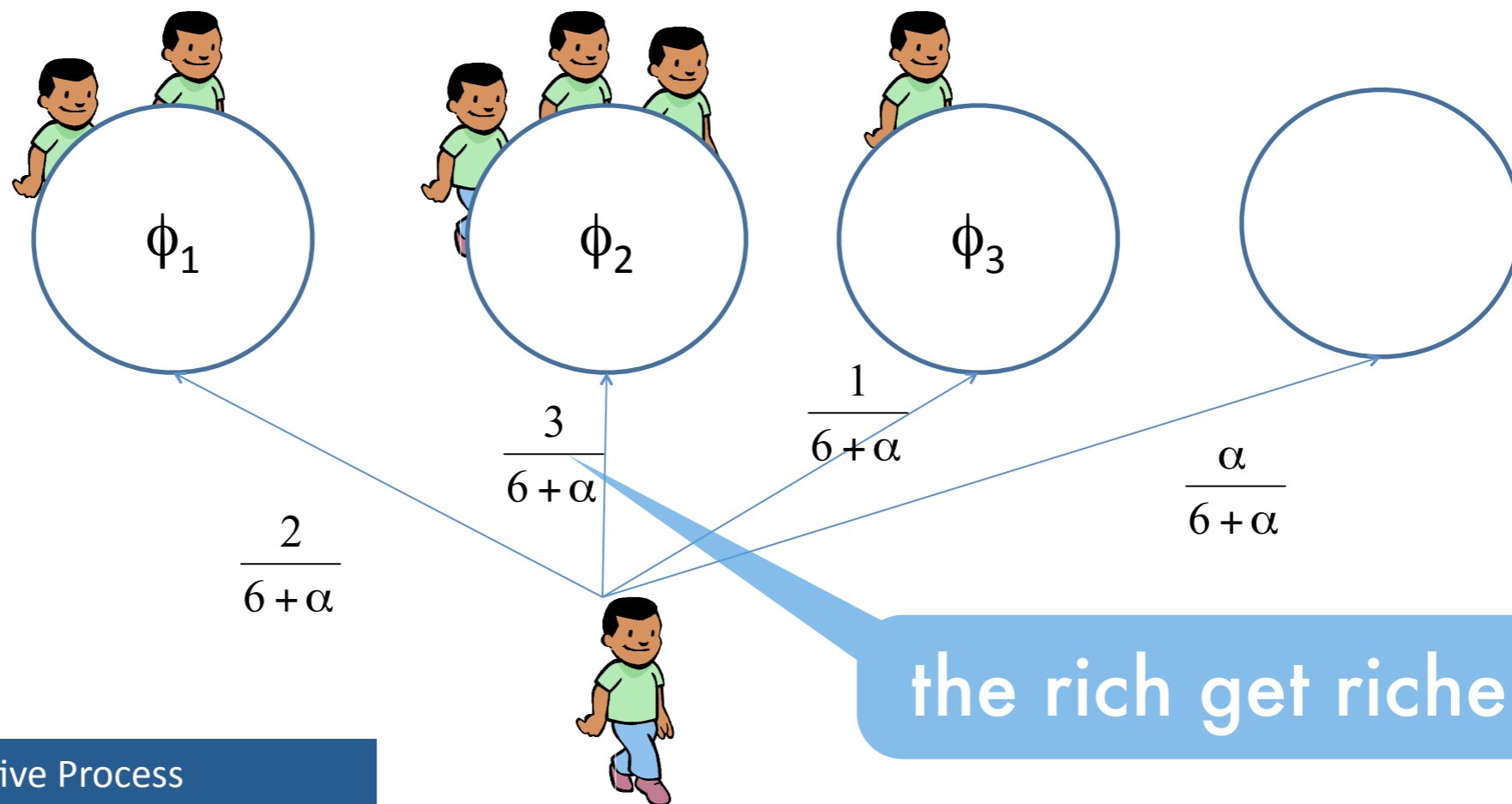
$$p(y|Y, \alpha) = \frac{n(y) + \alpha_y}{n + \sum_{y'} \alpha_{y'}}$$

- Infinite model

Assume that the total smoother weight is constant

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

Chinese Restaurant Metaphor



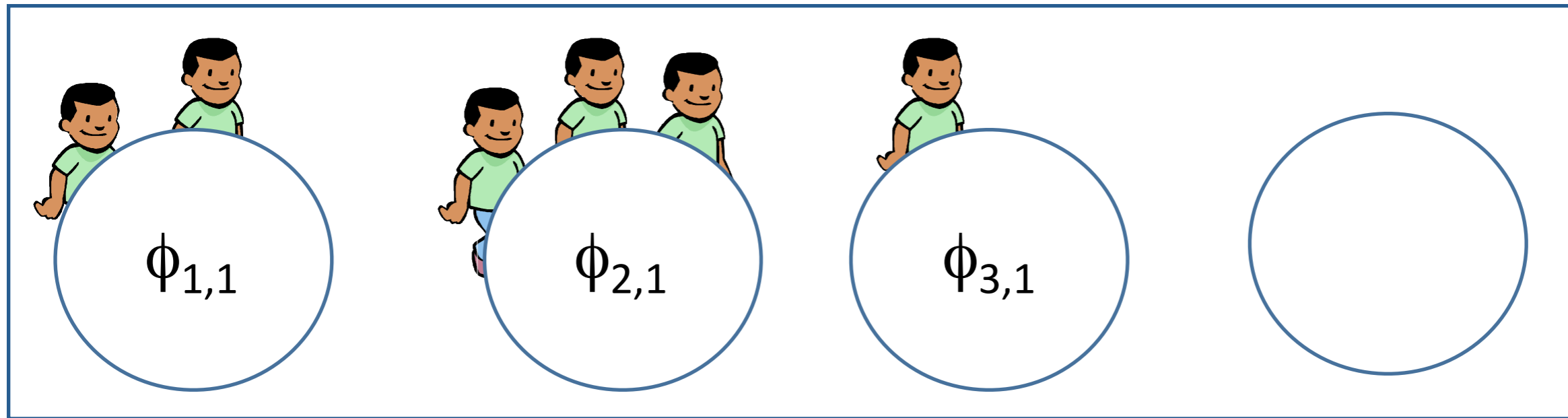
Generative Process

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

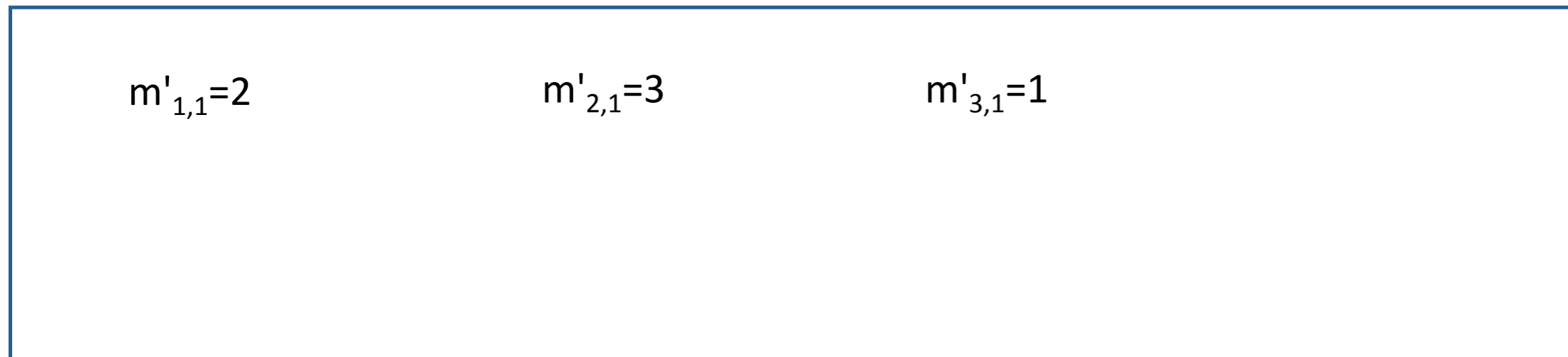
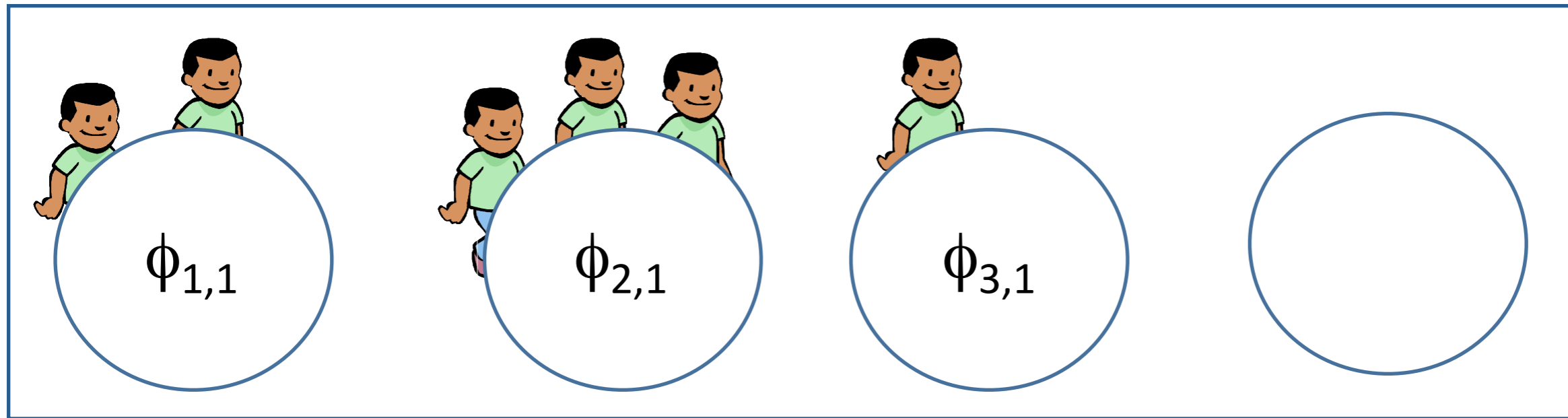
Evolutionary Clustering

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

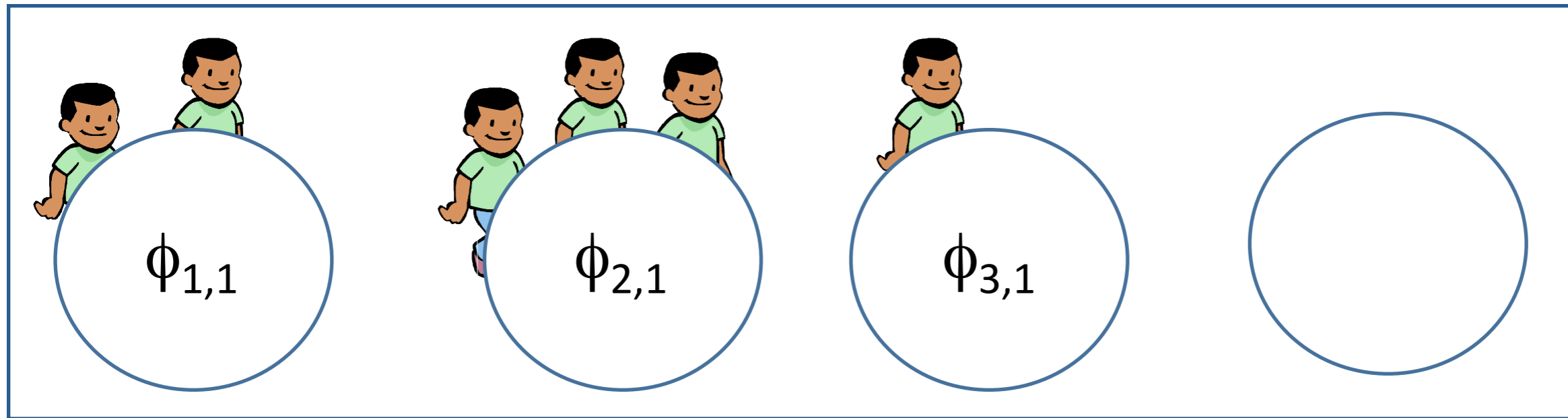
Recurrent Chinese Restaurant Process



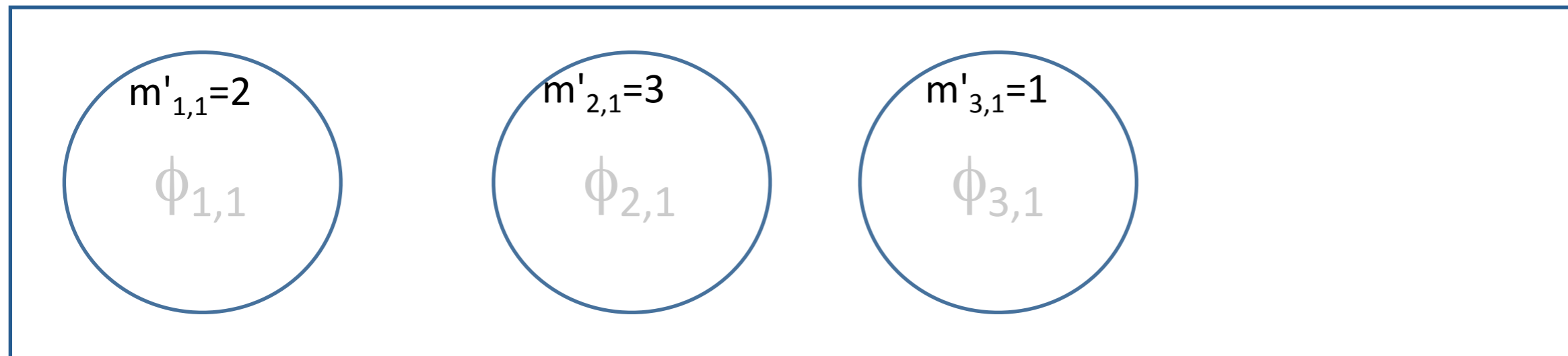
Recurrent Chinese Restaurant Process



Recurrent Chinese Restaurant Process



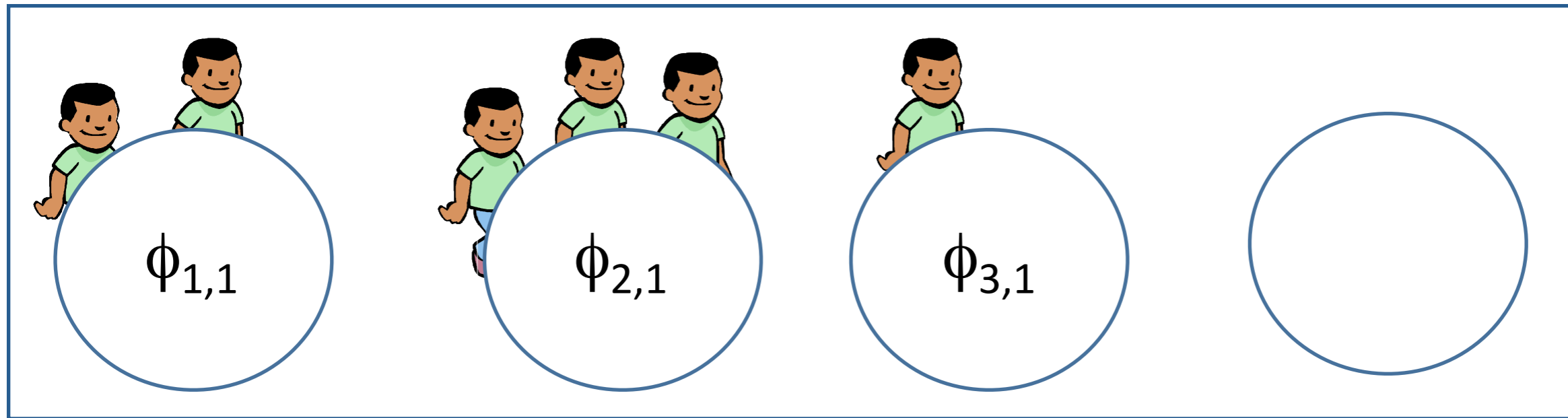
$T=1$



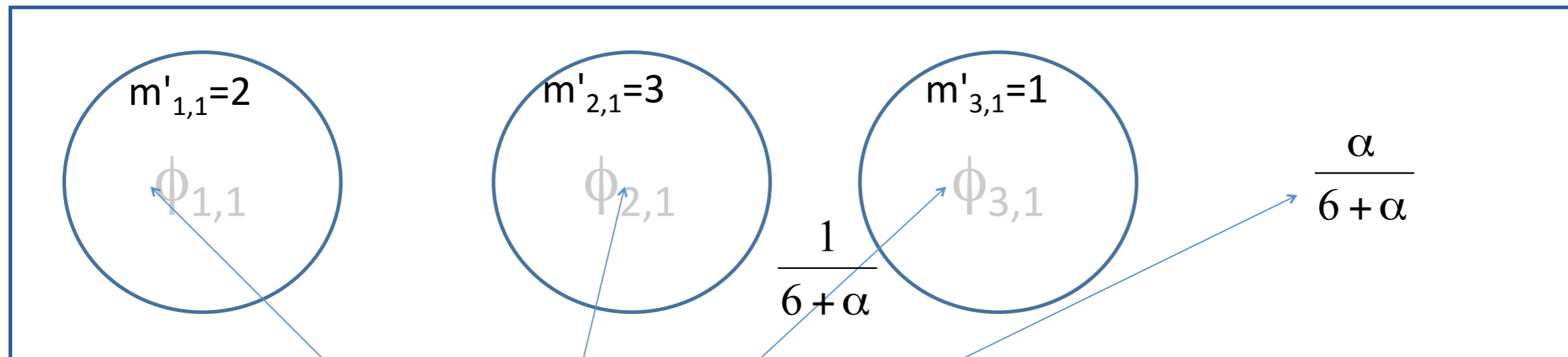
$T=2$



Recurrent Chinese Restaurant Process



$T=1$



$T=2$

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

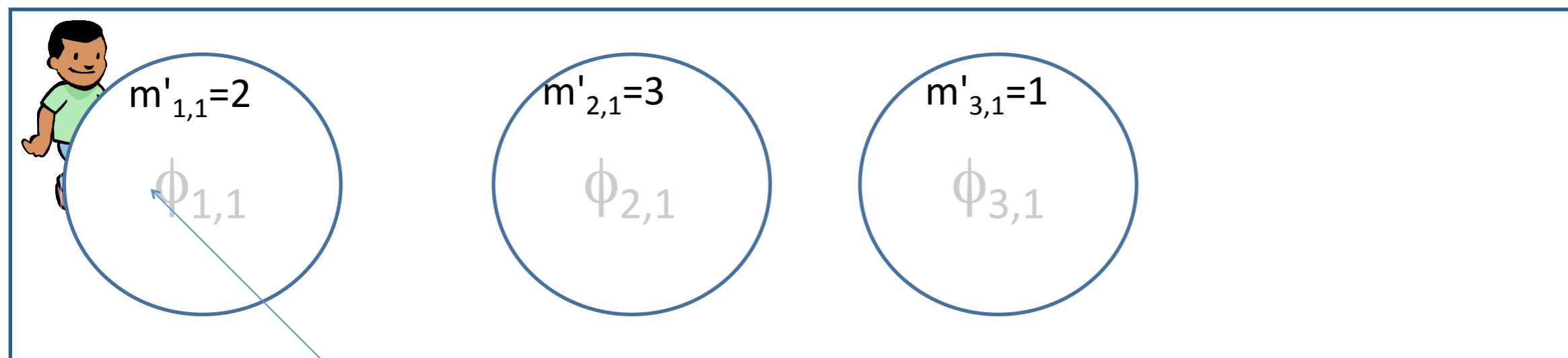
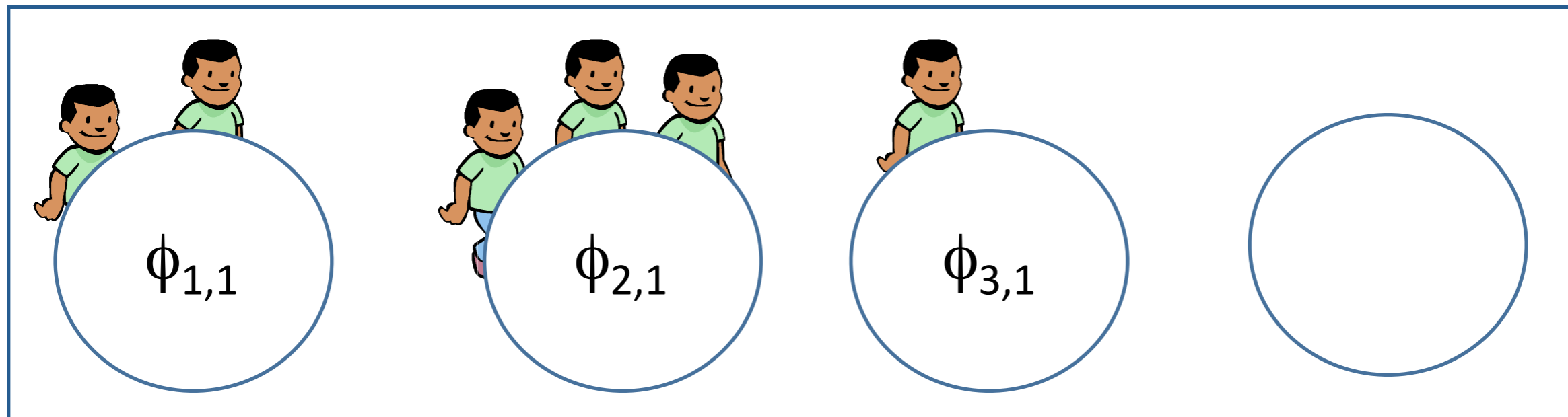
$$\frac{3}{6+\alpha}$$

$$\frac{1}{6+\alpha}$$

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



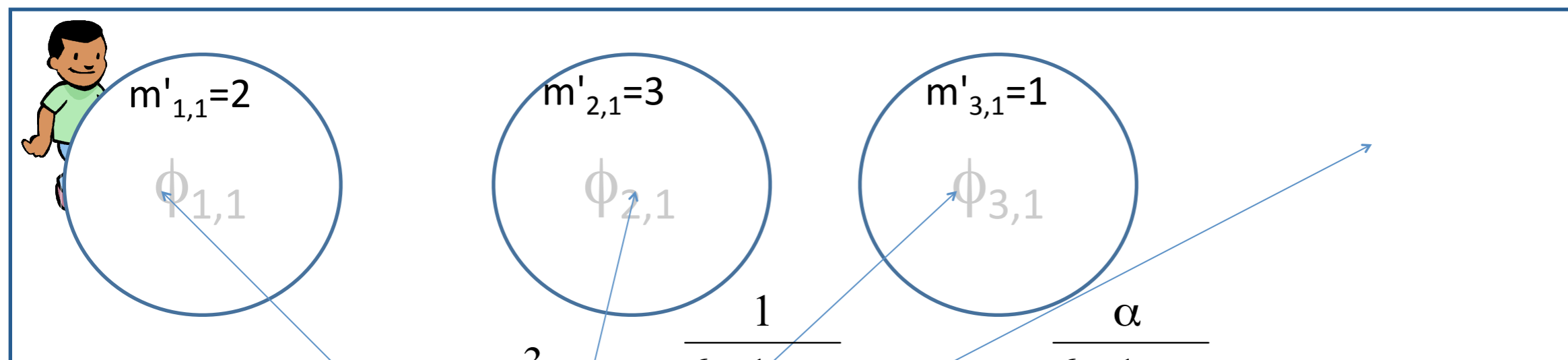
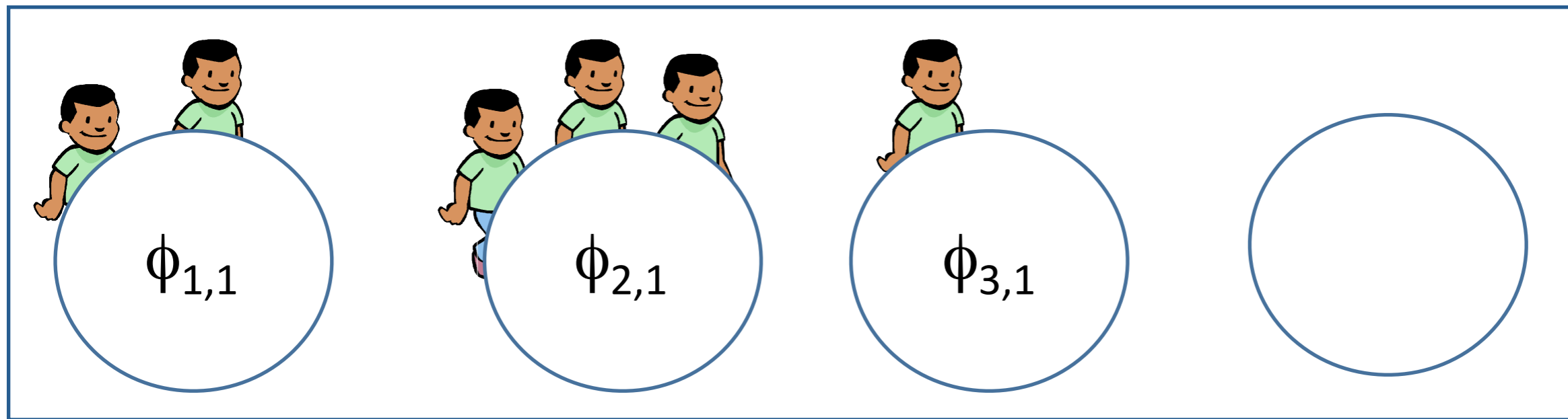
Recurrent Chinese Restaurant Process



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

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

Recurrent Chinese Restaurant Process



$$\frac{1+2}{6+1+\alpha}$$

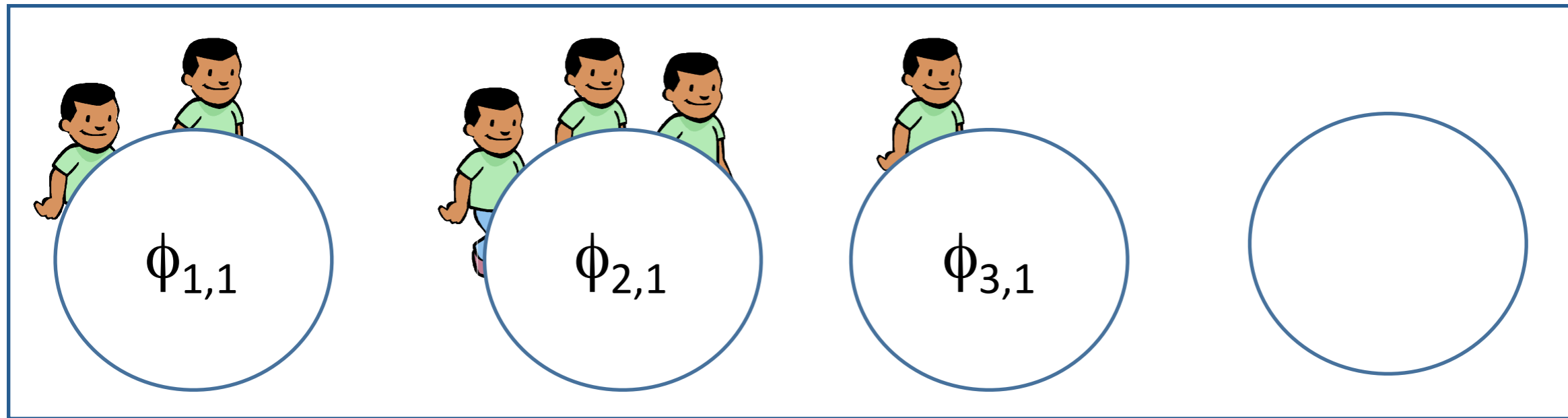
$$\frac{3}{6+1+\alpha}$$

$$\frac{1}{6+1+\alpha}$$

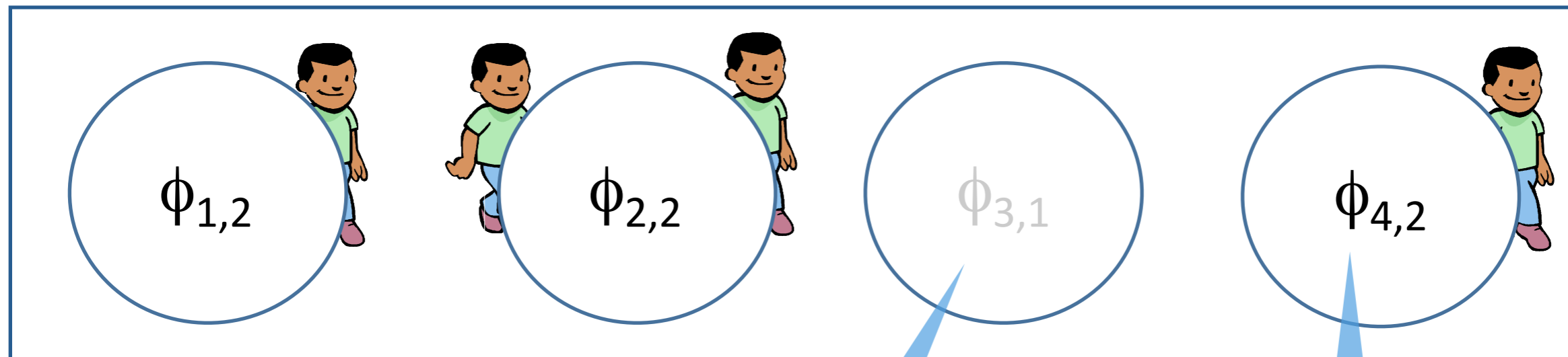
$$\frac{\alpha}{6+1+\alpha}$$



Recurrent Chinese Restaurant Process



$T=1$

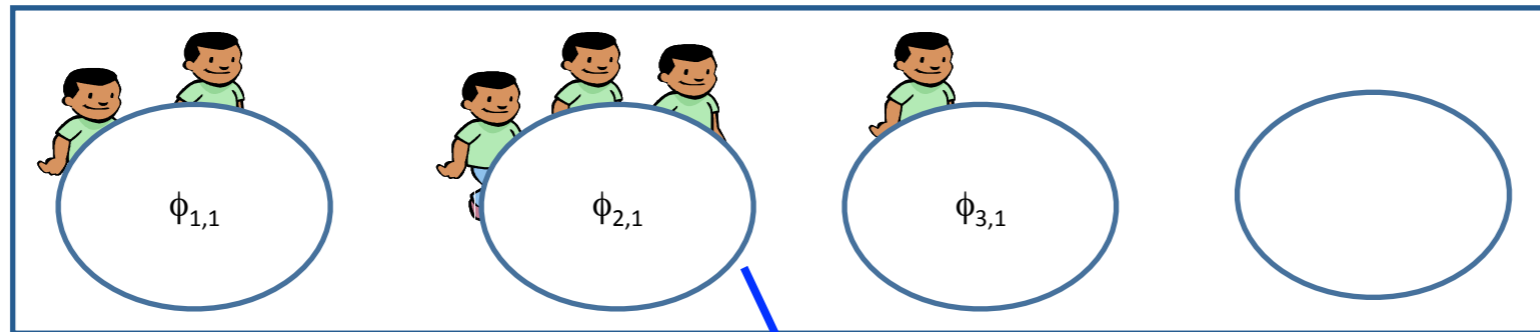


$T=2$

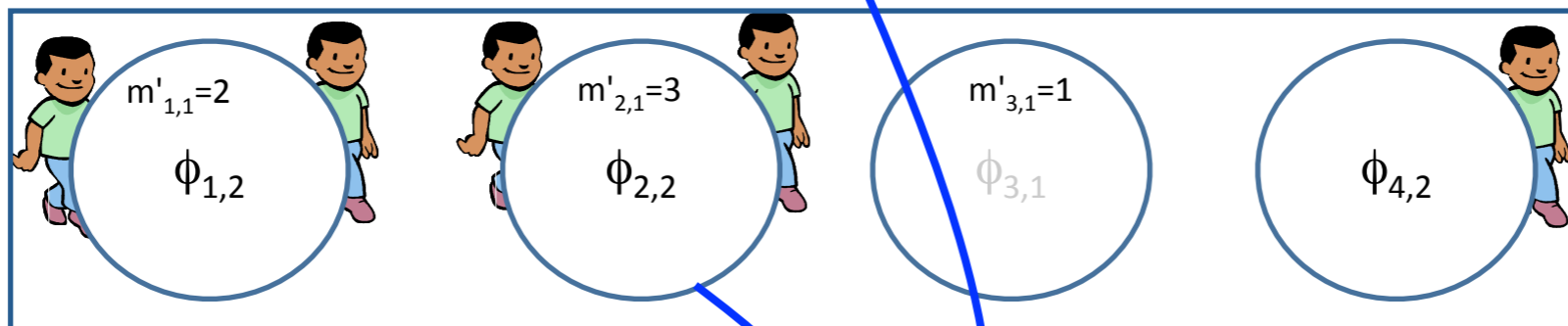
dead cluster

new cluster

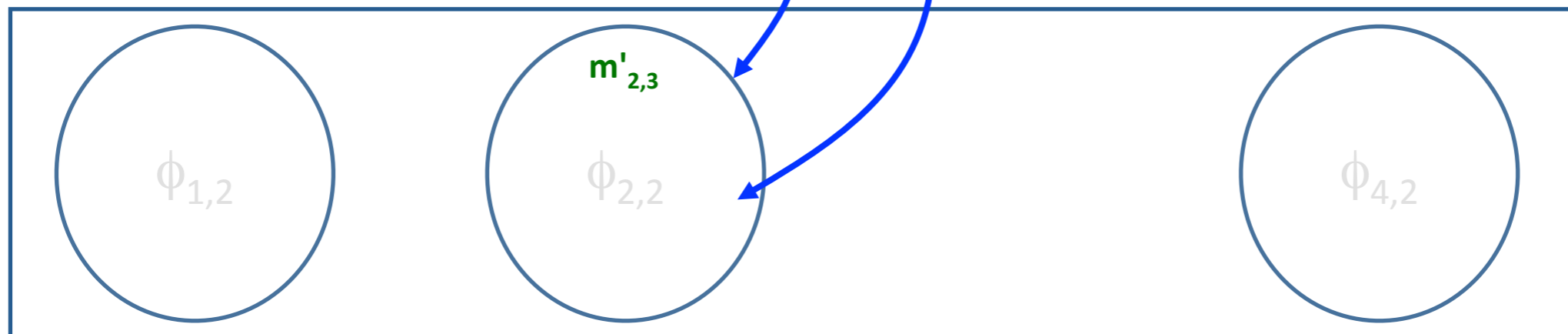
Longer History



T=1



T=2



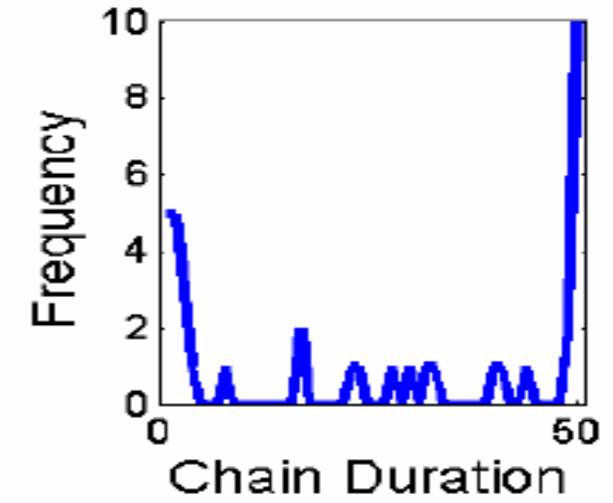
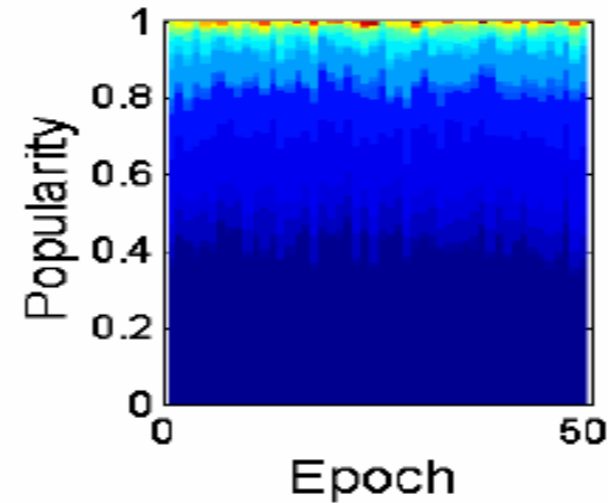
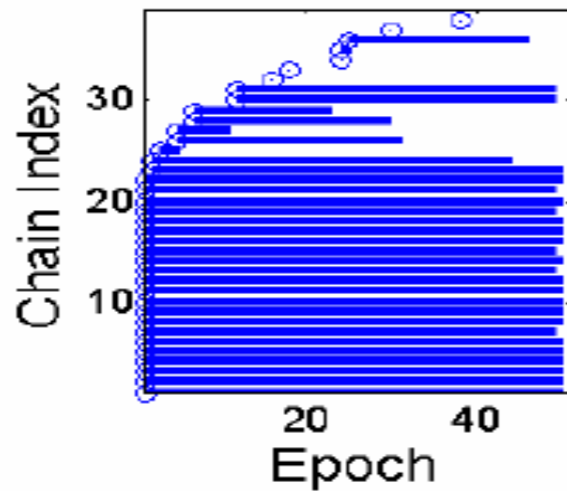
T=3

TDPM Generative Power

DPM

$$W = \infty$$

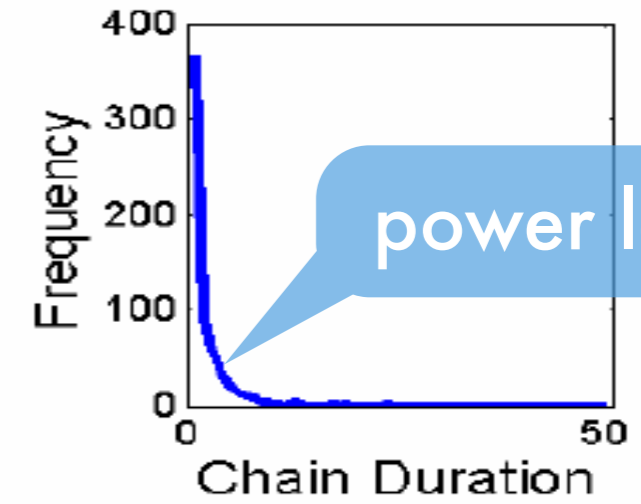
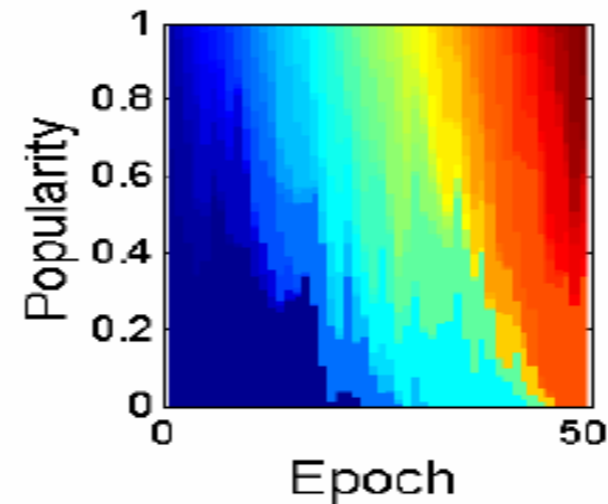
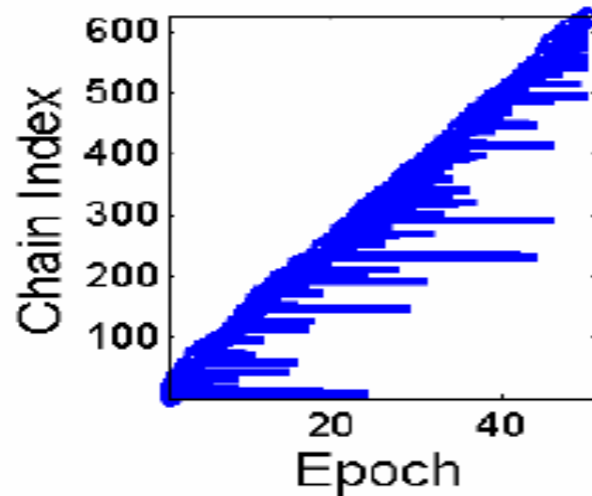
$$\lambda = \infty$$



TDPM

$$W = 4$$

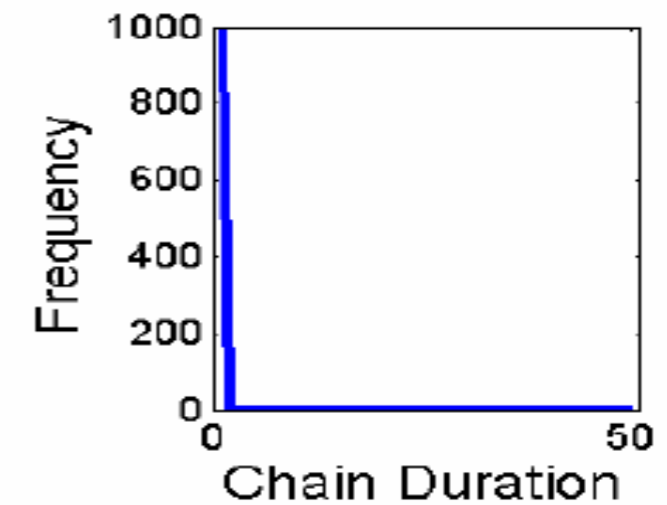
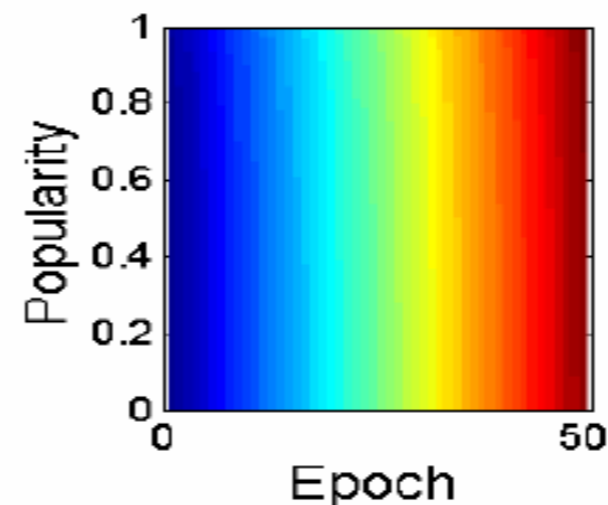
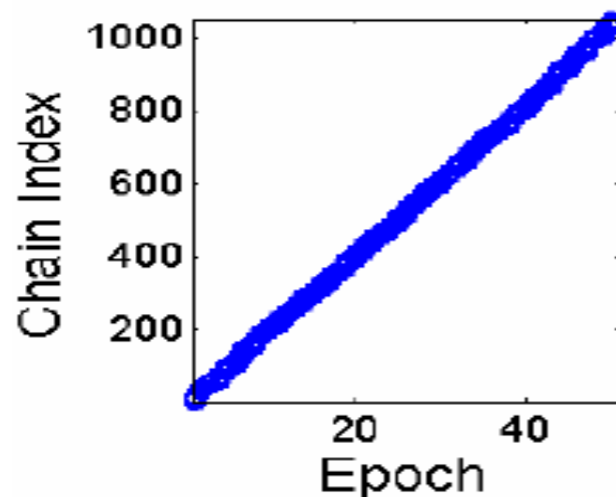
$$\lambda = .4$$



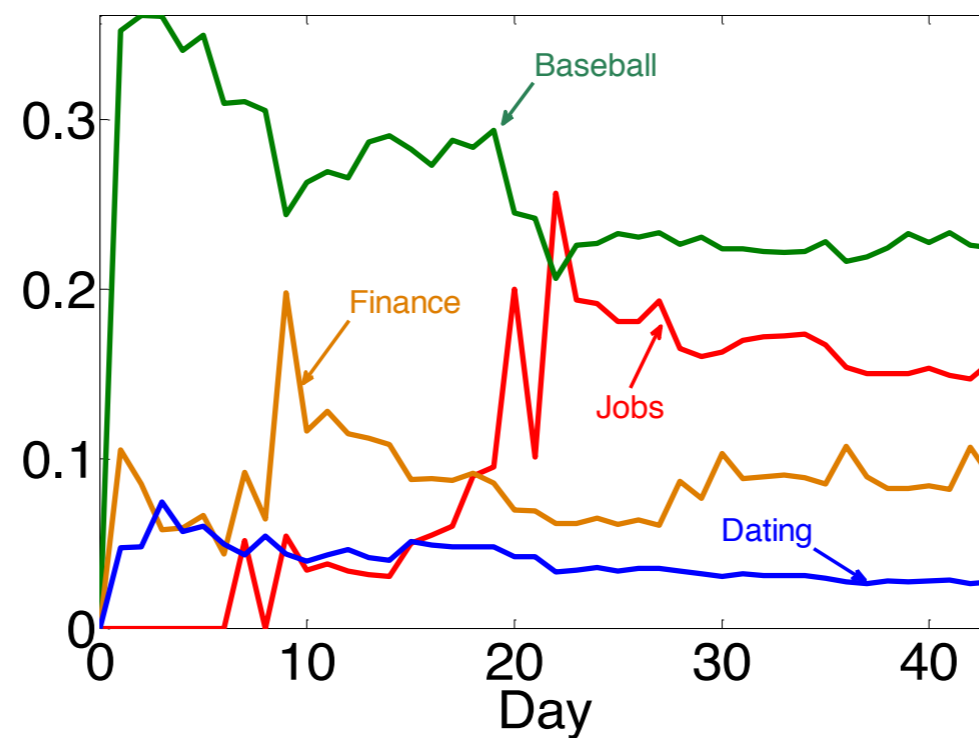
Independent DPMs

$$W = 0$$

$$\lambda = ? \text{ (any)}$$



User modeling



Buying a camera



time

Buying a camera

YAHOO! Web Images Video Local Shopping News More ▾

panasonic lx5

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



 Sponsor Results

Also try: [panasonic lx5](#), [more...](#)

Panasonic LX5 Cheap

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

Panasonic Lumix DMC-LX5 Review (white

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

Panasonic LX5 | Get The Lowest Price On

Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
www.panasoniclx5.com - [Cached](#) - [More from this site](#)

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

The Panasonic Lumix DMC-LX5 is a compact digital photo enthusiasts the ideal way for capturing professional photos and High De...

Price: **\$434 to \$513.99**

[Reviews](#) | [Price & Details](#) | [Specs](#)

Sponsored Results



Hello, **Alexander Smola**. We have [recommendations](#) for you. (Not Alexander?)

[Alexander's Amazon.com](#) |  [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments

Search

Camera & Photo

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

Instant Order Update for Alexander Smola. You purchased this item on October 6, 2010. [View](#)

Color: Black

Prime

Member: Alexander Smola

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



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

by [Panasonic](#)

★★★★☆ (40 customer reviews)

List Price: ~~\$499.00~~

Price: **\$444.95** & eligible for free shipping with

Amazon Prime

You Save: **\$54.05 (11%)**

[new](#)

time

Buying a camera

YAHOO! Web Images Video Local Shopping News More ▾

panasonic lx5

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



Also try: [panasonic lx5](#), [more...](#)

Panasonic LX5 Cheap

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

Panasonic Lumix DMC-LX5 Review (white

\$434.00 as of Oct 17, 2010 Despite its shortcomings the **Panasonic Lumix DMC-LX5** delivers an excellent fastest in its class ...

reviews.cnet.com/digital-cameras/panasonic-lumix-this-site

Panasonic LX5 | Get The Lowest Price On

Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
www.panasoniclx5.com - [Cached](#) - [More from this site](#)

Panasonic Lumix DMC-LX5 White Digital (

shopping.yahoo.com

Sponsor Results

Sponsored Results



Hello, **Alexander Smola**. We have [recommendations](#) for you. (Not Alexander?)

[Alexander's Amazon.com](#) | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments ▾

Search

Camera & Photo

All Electronics

Brands

Bestsellers

Digital SLRs & Lenses

Point-And-Shoots

Camcorders

Instant Order Update for Alexander Smola. You purchased this item on October 6, 2010. [View](#)

Color: **Black**

Prime

Member: Alexander Smola

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

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

by [Panasonic](#)

★★★★☆ (40 customer reviews)

List Price: ~~\$499.00~~

Price: **\$444.95** & eligible for free shipping with

Amazon Prime

You Save: **\$54.05 (11%)**



[new](#)

show ads now

time



Buying a camera

YAHOO! Web Images Video Local Shopping News More ▾

panasonic lx5

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



Also try: [panasonic lx5](#), [more...](#)

Panasonic LX5 Cheap

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

Panasonic Lumix DMC-LX5 Review (white

\$434.00 as of Oct 17, 2010 Despite its shortcomings the **Panasonic Lumix DMC-LX5** delivers an excellent performance in its class ...

reviews.cnet.com/digital-cameras/panasonic-lumix-lx5/
[this site](#)

Panasonic LX5 | Get The Lowest Price On

Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
www.panasoniclx5.com - [Cached](#) - [More from this site](#)

Panasonic Lumix DMC-LX5 White Digital (

shopping.yahoo.com

Sponsor Results

Sponsored Results



Hello, **Alexander Smola**. We have [recommendations](#) for you. (Not Alexander?)

[Alexander's Amazon.com](#) | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments ▾

Search

Camera & Photo

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

Instant Order Update for Alexander Smola. You purchased this item on October 6, 2010. [View](#)

Color: Black

Prime

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

Member: Alexander Smola

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

by [Panasonic](#)

★★★★☆ (40 customer reviews)

List Price: ~~\$499.00~~

Price: **\$444.95** & eligib

Amazon Prime

You Save: \$54.05 (11%)

[new](#)

show ads now

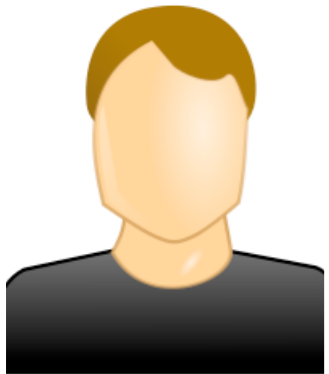
too late

time





Car
Deals
van



job
Hiring
diet

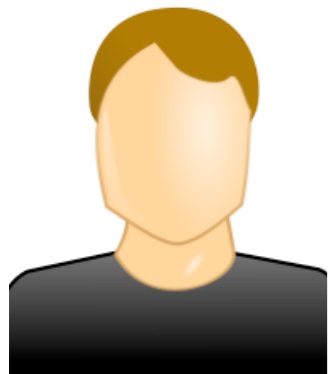




Car
Deals
van

Auto
Price
Used
inspection

Movies
Theatre
Art
gallery



job
Hiring
diet

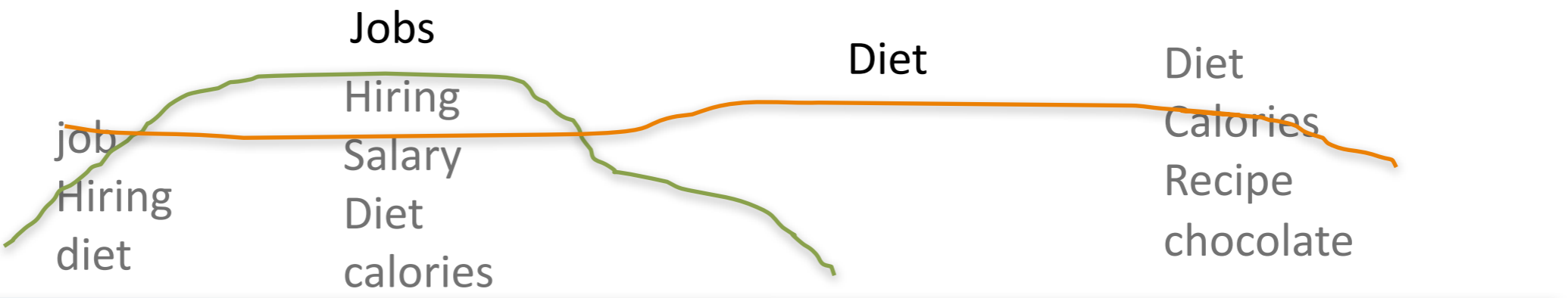
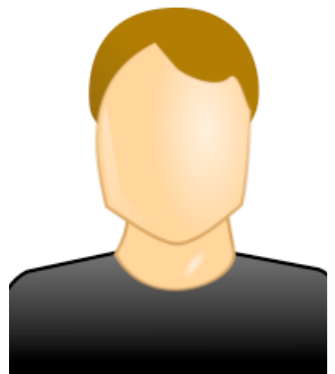
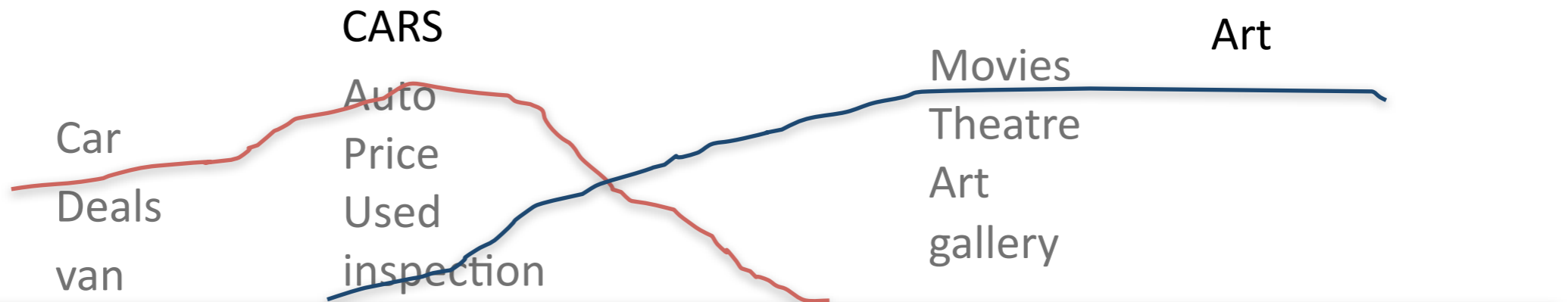
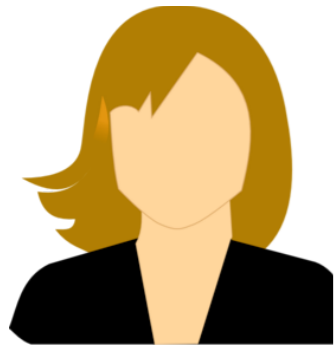
Hiring
Salary
Diet
calories

Diet
Calories
Recipe
chocolate



Flight
London
Hotel
weather

School
Supplies
Loan
college



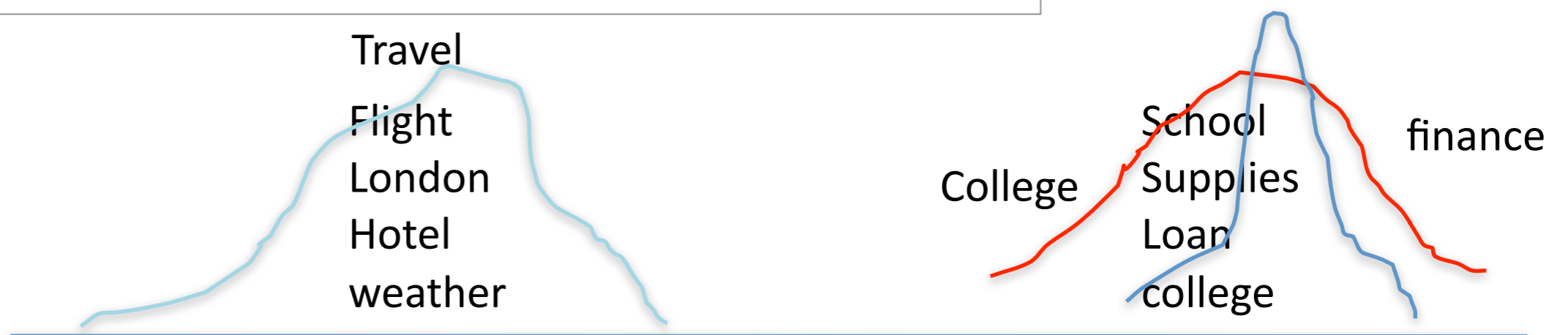
User modeling

Input

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

Output

- Users' daily distribution over intents
- Dynamic intent representation

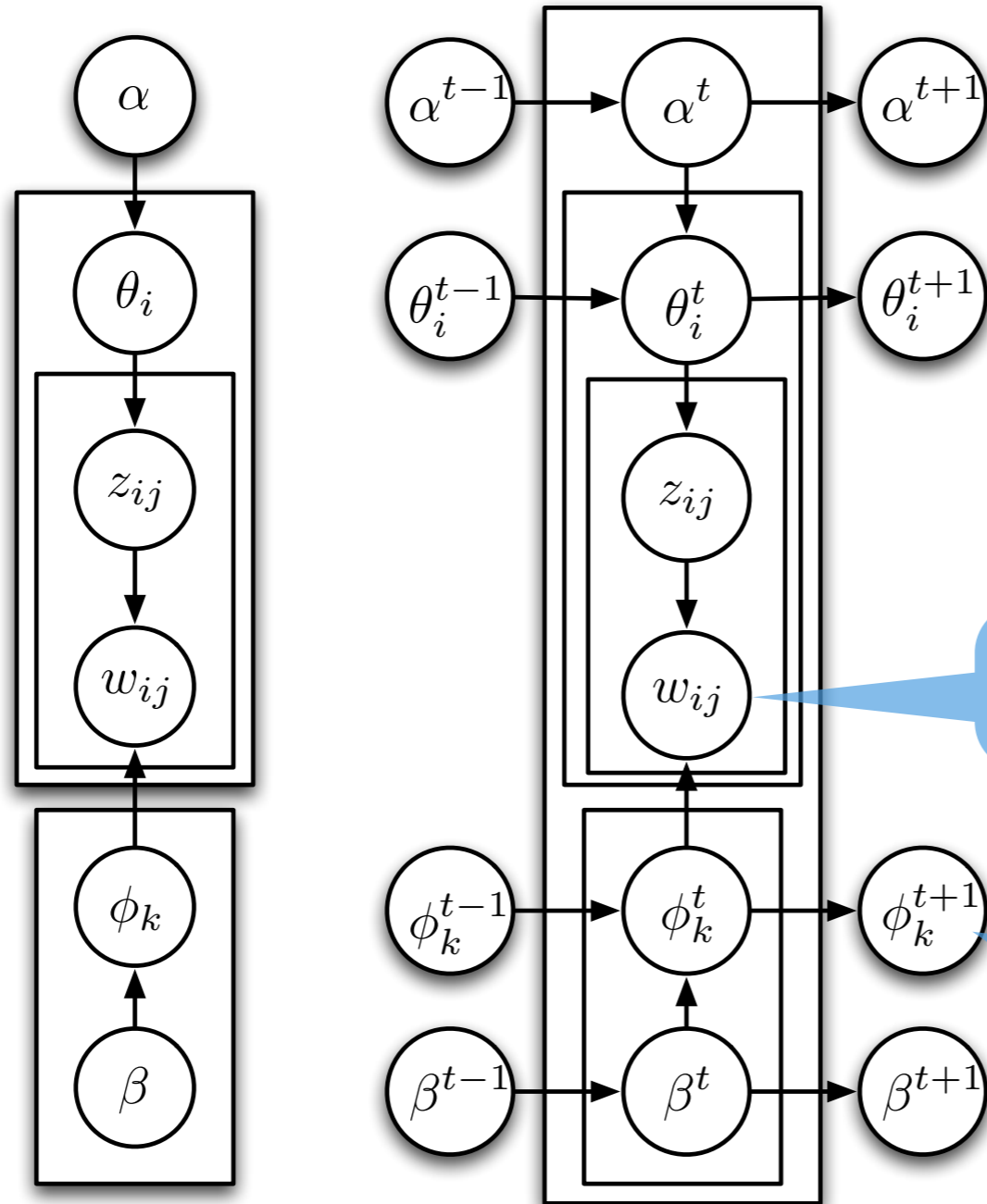


Time dependent models

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

Graphical Model

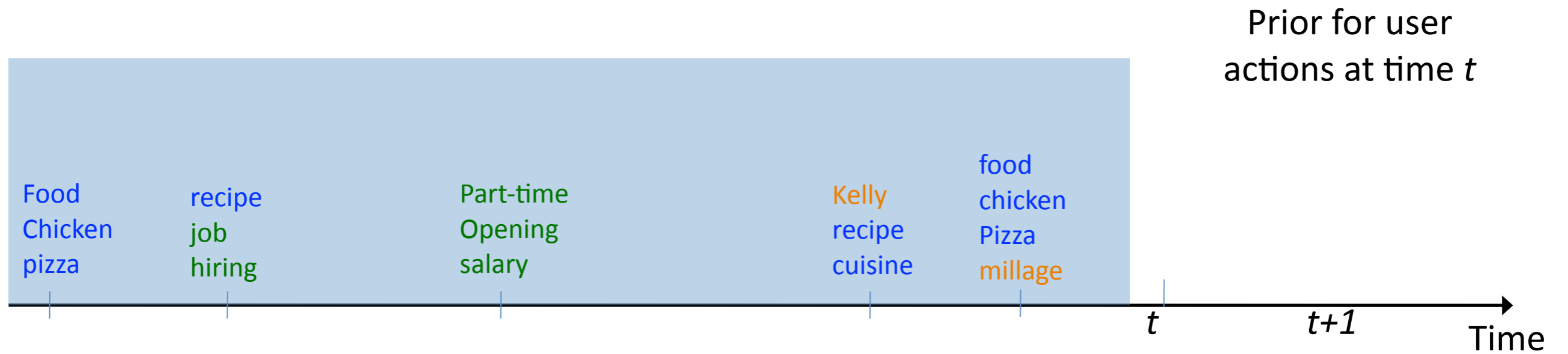
plain
LDA



time dependent
user interest

user actions

actions per topic



Diet

- Recipe
- Chocolate
- Pizza
- Food
- Chicken
- Milk
- Butter
- Powder

Cars

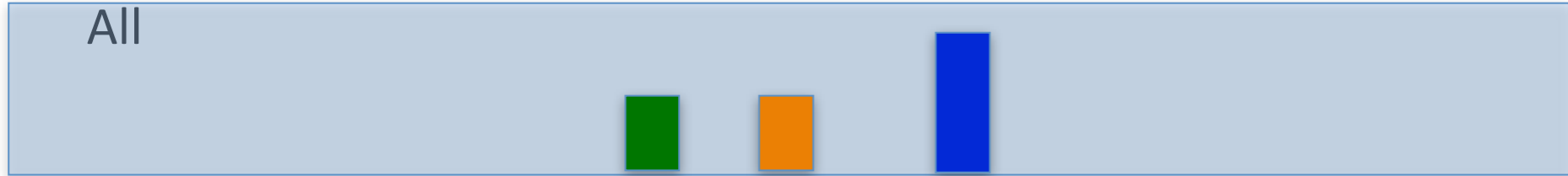
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

- job
- Career
- Business
- Assistant
- Hiring
- Part-time
- Receptionist

Finance

- Bank
- Online
- Credit
- Card
- debt
- portfolio
- Finance
- Chase



Long-term



Prior for user actions at time t

t $t+1$ Time

Diet

- Recipe
- Chocolate
- Pizza
- Food
- Chicken
- Milk
- Butter
- Powder

Cars

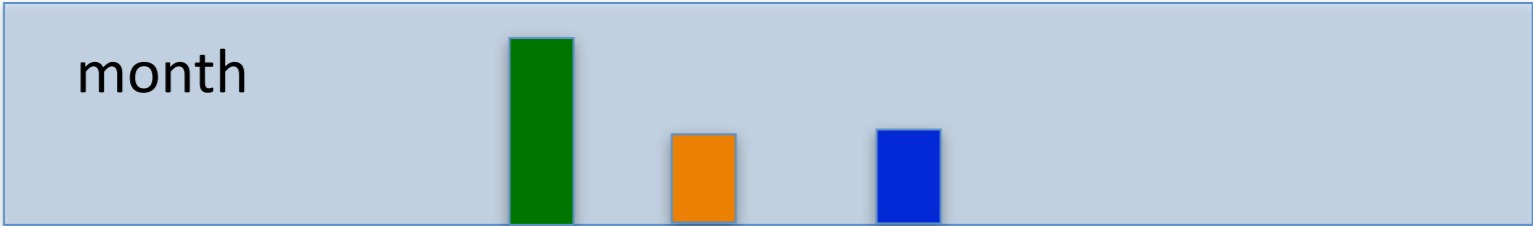
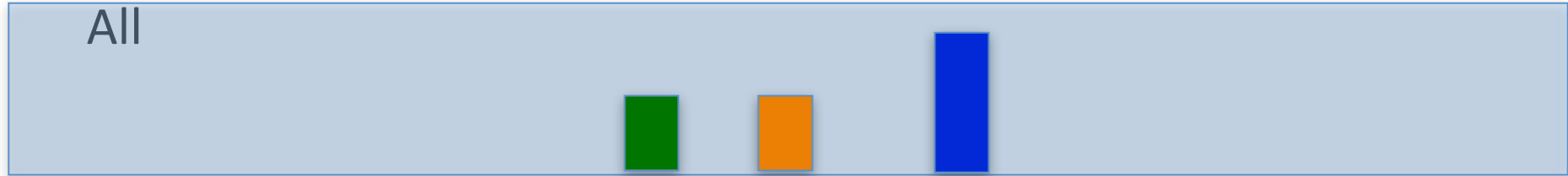
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

- job
- Career
- Business
- Assistant
- Hiring
- Part-time
- Receptionist

Finance

- Bank
- Online
- Credit
- Card
- debt
- portfolio
- Finance
- Chase



Long-term



Prior for user actions at time t

t $t+1$ Time

Diet

- Recipe
- Chocolate
- Pizza
- Food
- Chicken
- Milk
- Butter
- Powder

Cars

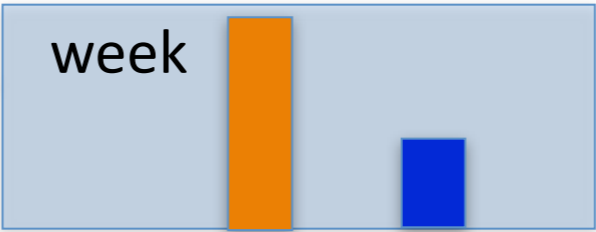
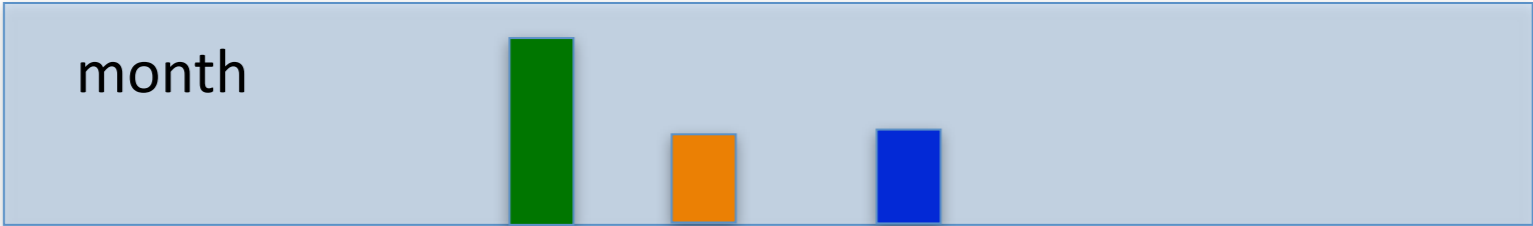
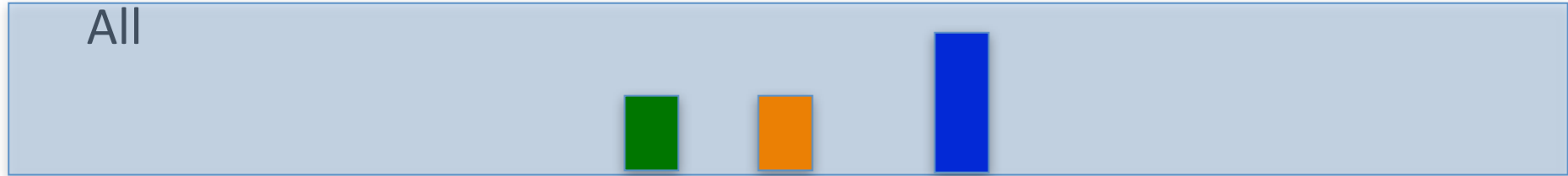
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

- job
- Career
- Business
- Assistant
- Hiring
- Part-time
- Receptionist

Finance

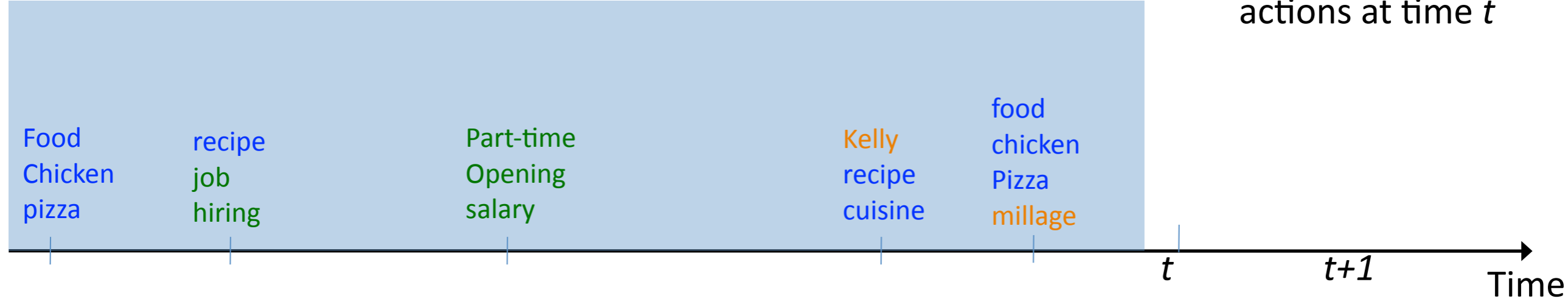
- Bank
- Online
- Credit
- Card
- debt
- portfolio
- Finance
- Chase



Long-term

short-term

Prior for user actions at time t



Diet

- Recipe
- Chocolate
- Pizza
- Food
- Chicken
- Milk
- Butter
- Powder

Cars

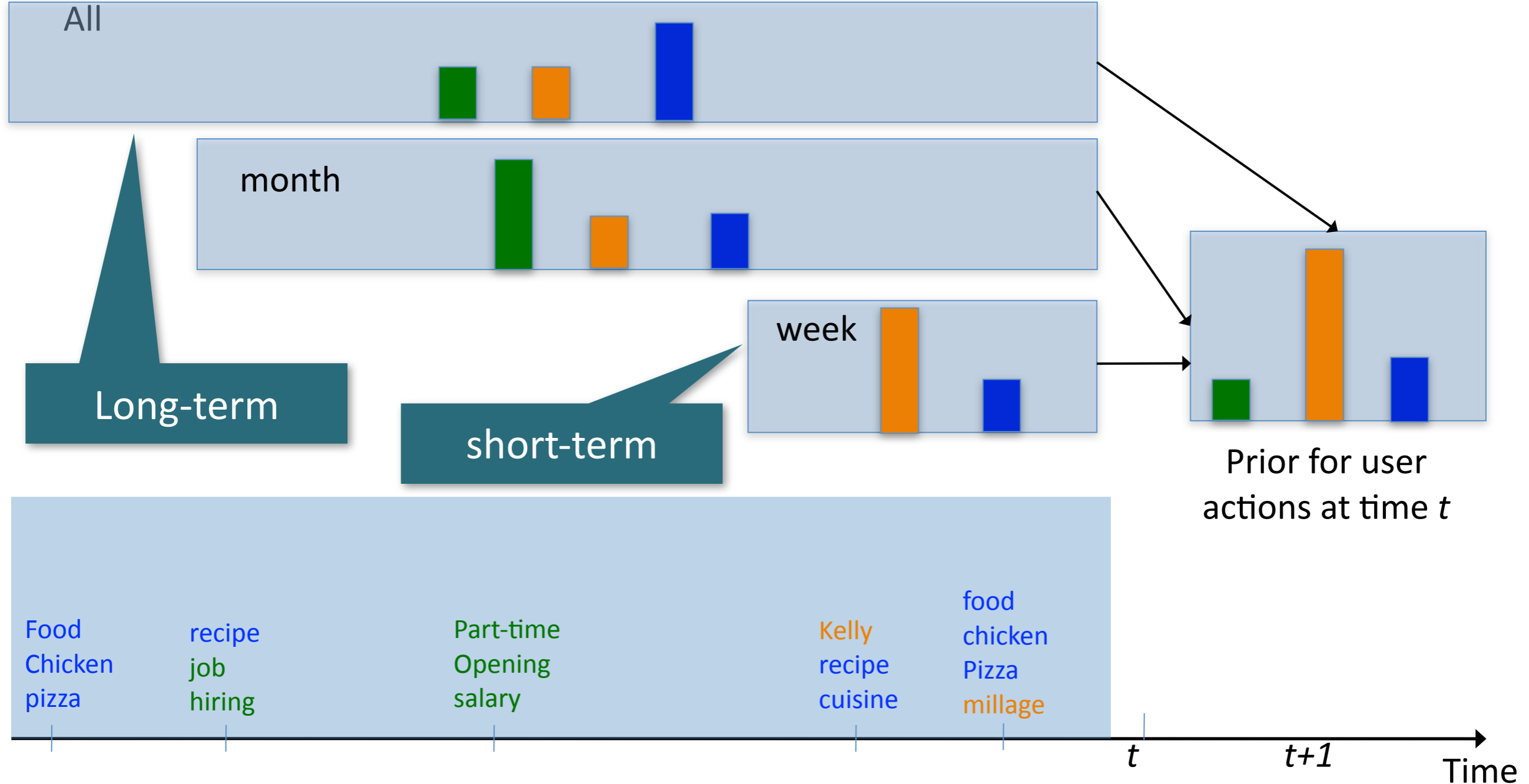
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

- job
- Career
- Business Assistant
- Hiring
- Part-time
- Receptionist

Finance

- Bank
- Online
- Credit Card
- debt portfolio
- Finance
- Chase



Diet

- Recipe
- Chocolate
- Pizza
- Food
- Chicken
- Milk
- Butter
- Powder

Cars

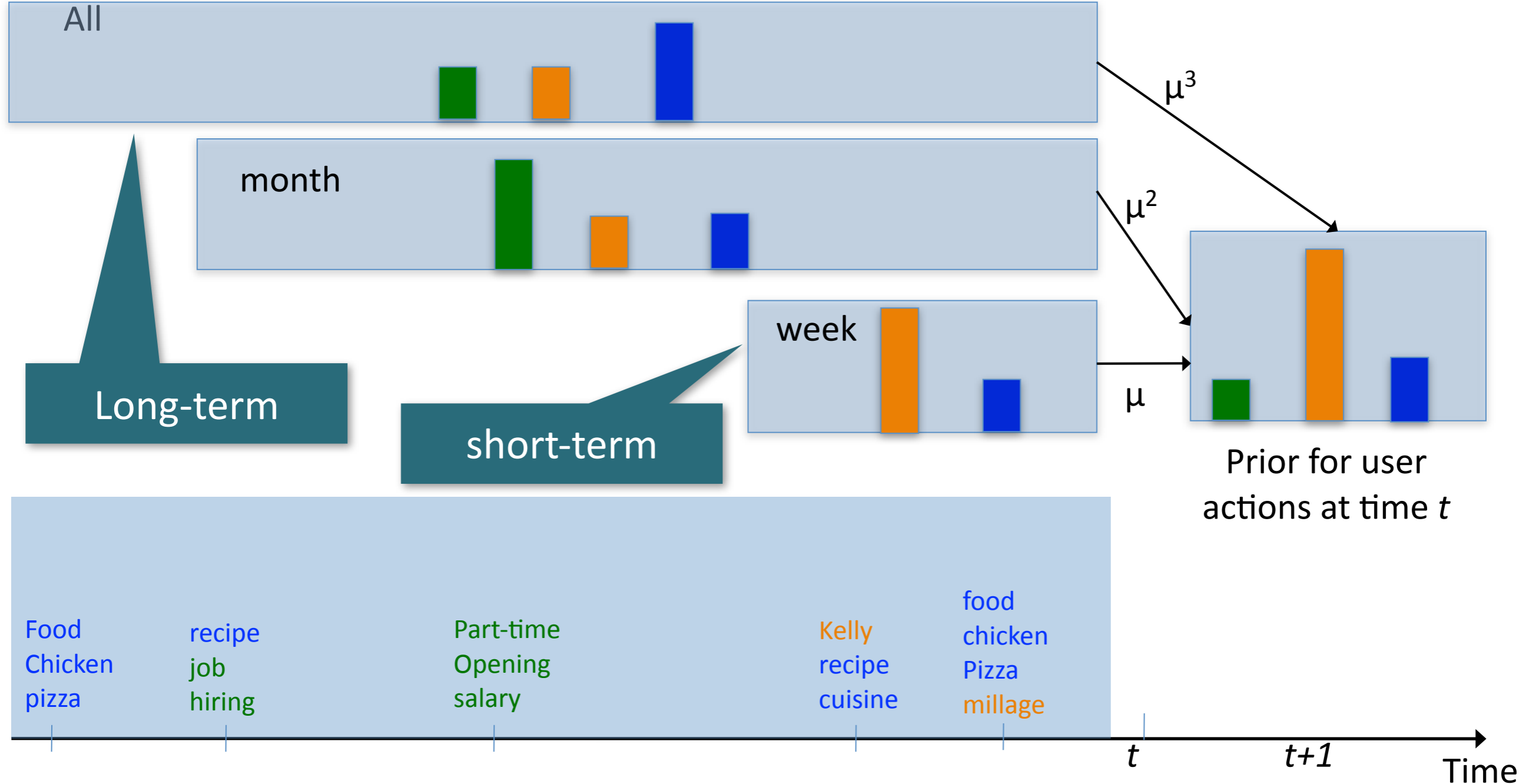
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

- job
- Career
- Business
- Assistant
- Hiring
- Part-time
- Receptionist

Finance

- Bank
- Online
- Credit
- Card
- debt
- portfolio
- Finance
- Chase



Diet

- Recipe
- Chocolate
- Pizza
- Food
- Chicken
- Milk
- Butter
- Powder

Cars

- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

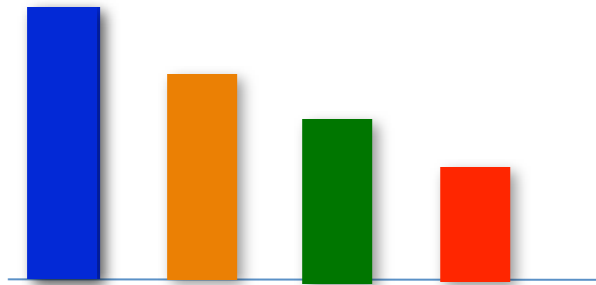
Job

- job
- Career
- Business
- Assistant
- Hiring
- Part-time
- Receptionist

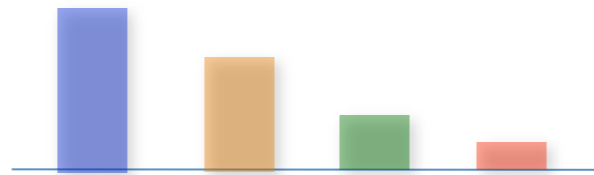
Finance

- Bank
- Online
- Credit
- Card
- debt
- portfolio
- Finance
- Chase

At time t



At time t+1



Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

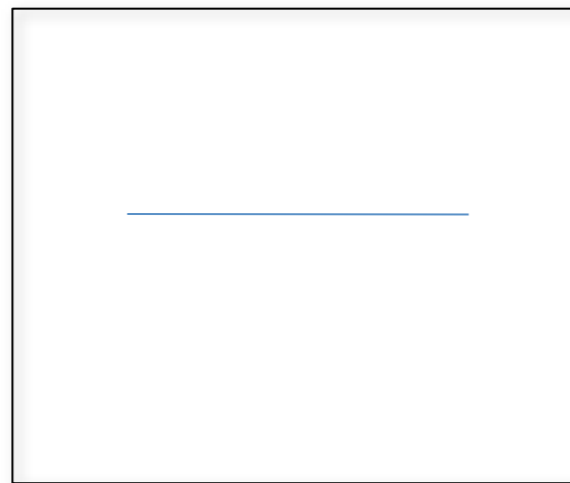
Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptioni
st

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

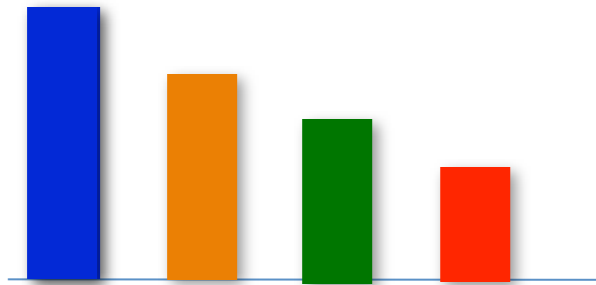


Food Chicken
Pizza mileage

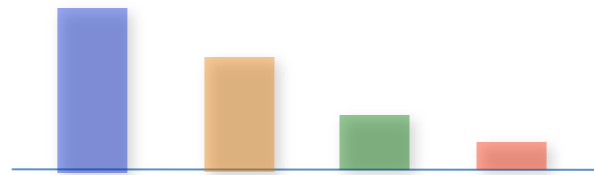


Car speed offer
Camry accord career

At time t



At time t+1



Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

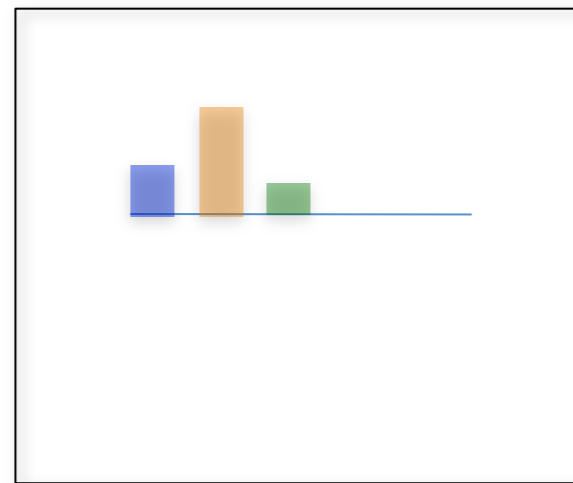
Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptioni
st

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

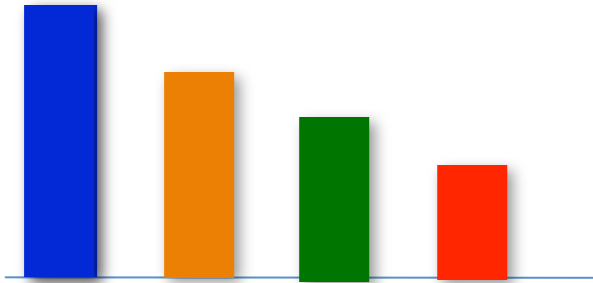


Food Chicken
Pizza mileage

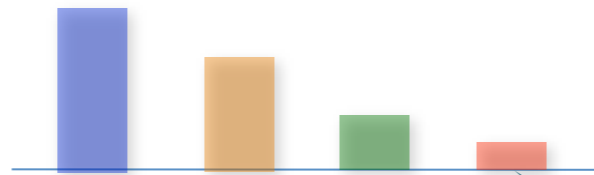


Car speed offer
Camry accord career

At time t



At time t+1



Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptioni
st

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



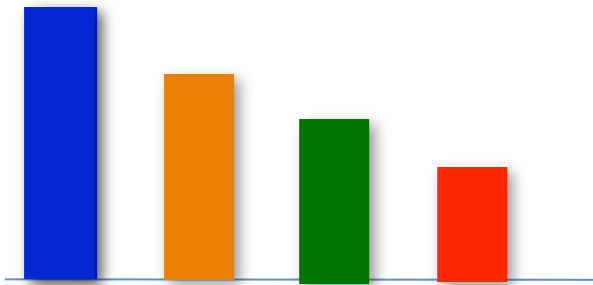
Food Chicken
Pizza mileage

priors

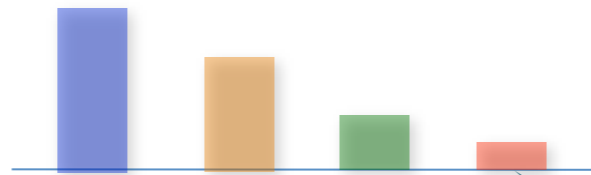


Car speed offer
Camry accord career

At time t



At time t+1



Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

job
Career
Business
Assistant
Hiring
Part-time
Receptioni
st

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



Food Chicken
Pizza mileage

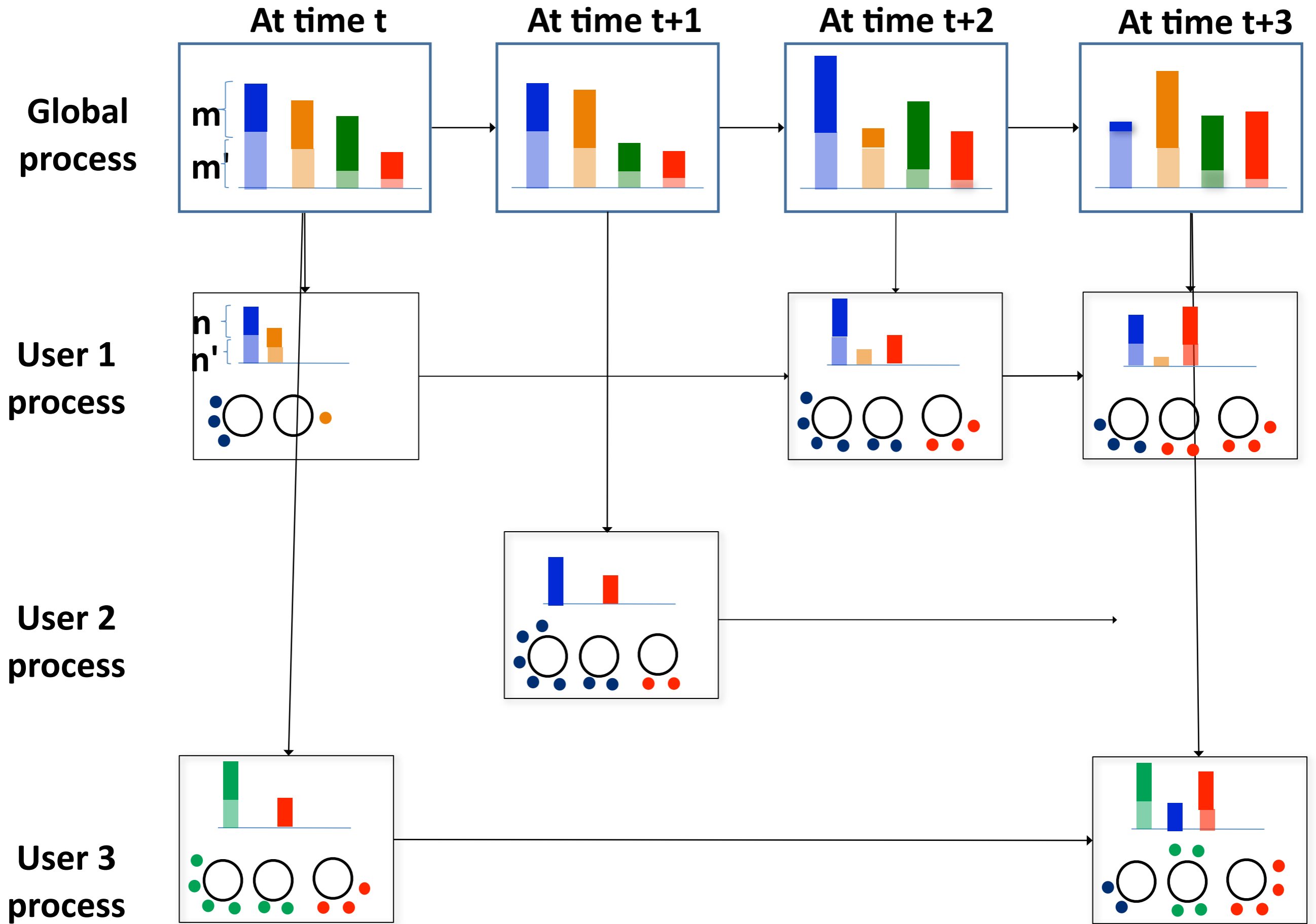
priors



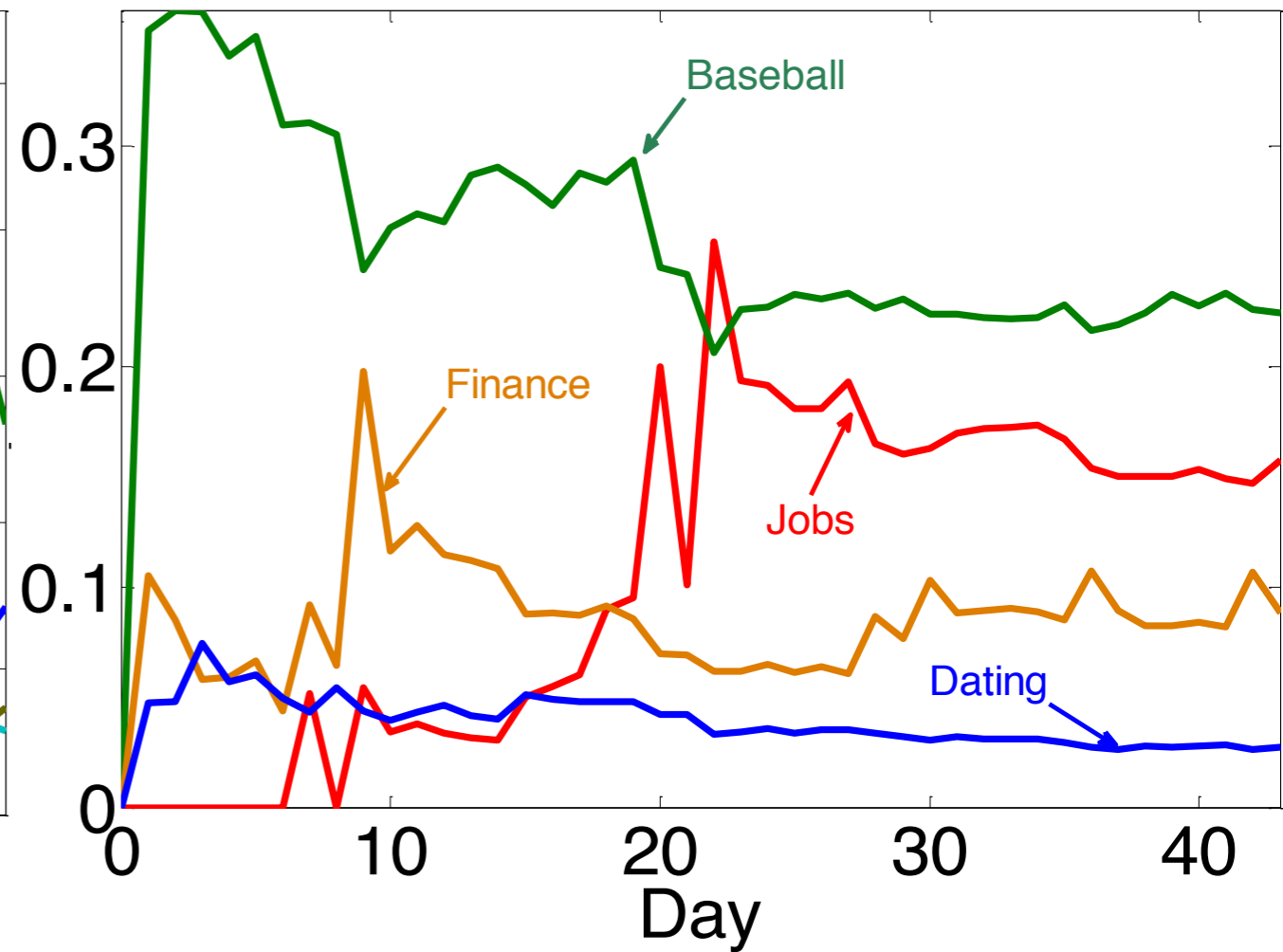
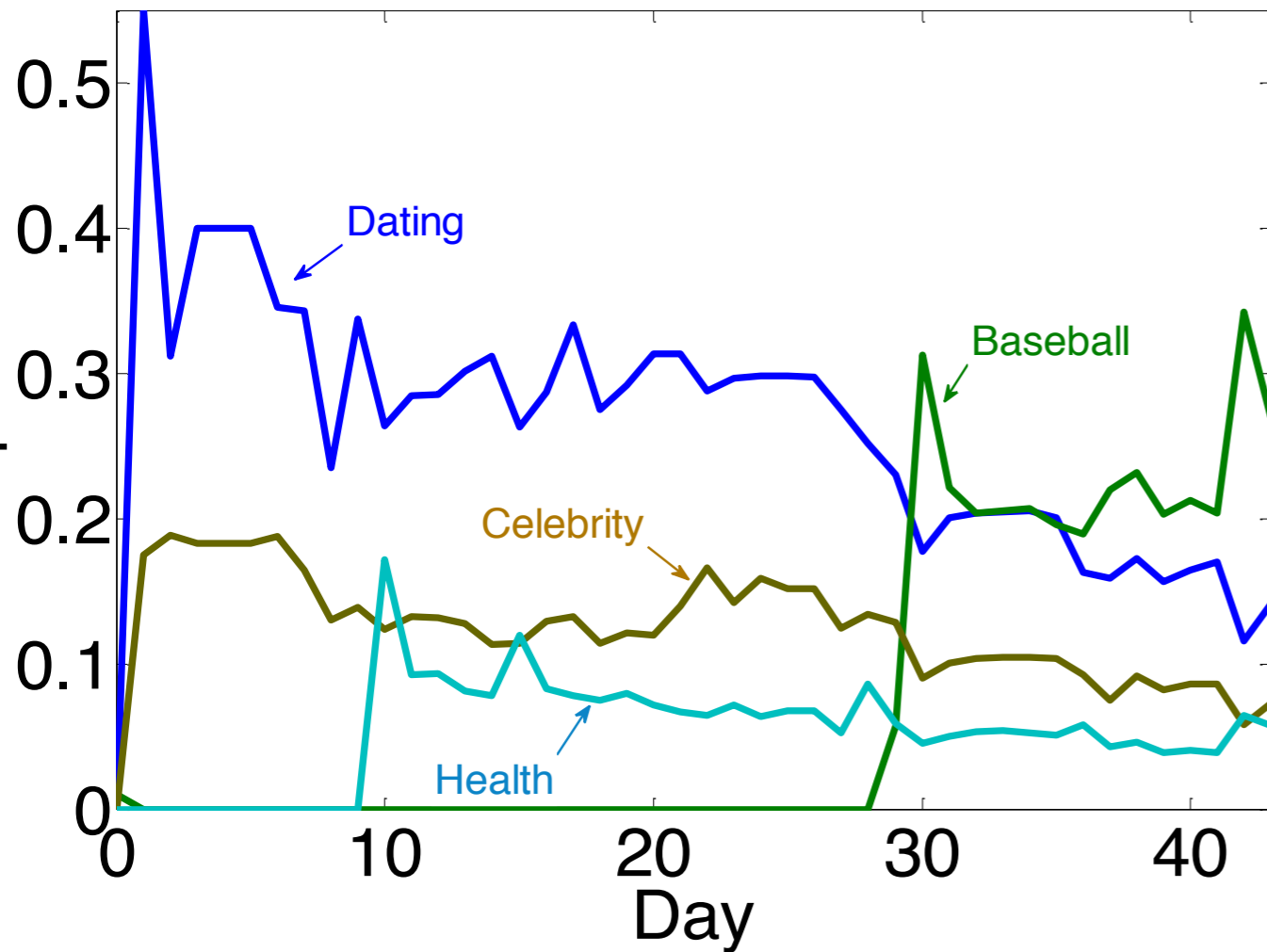
Car speed offer
Camry accord career

Generative Process

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



Sample users



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

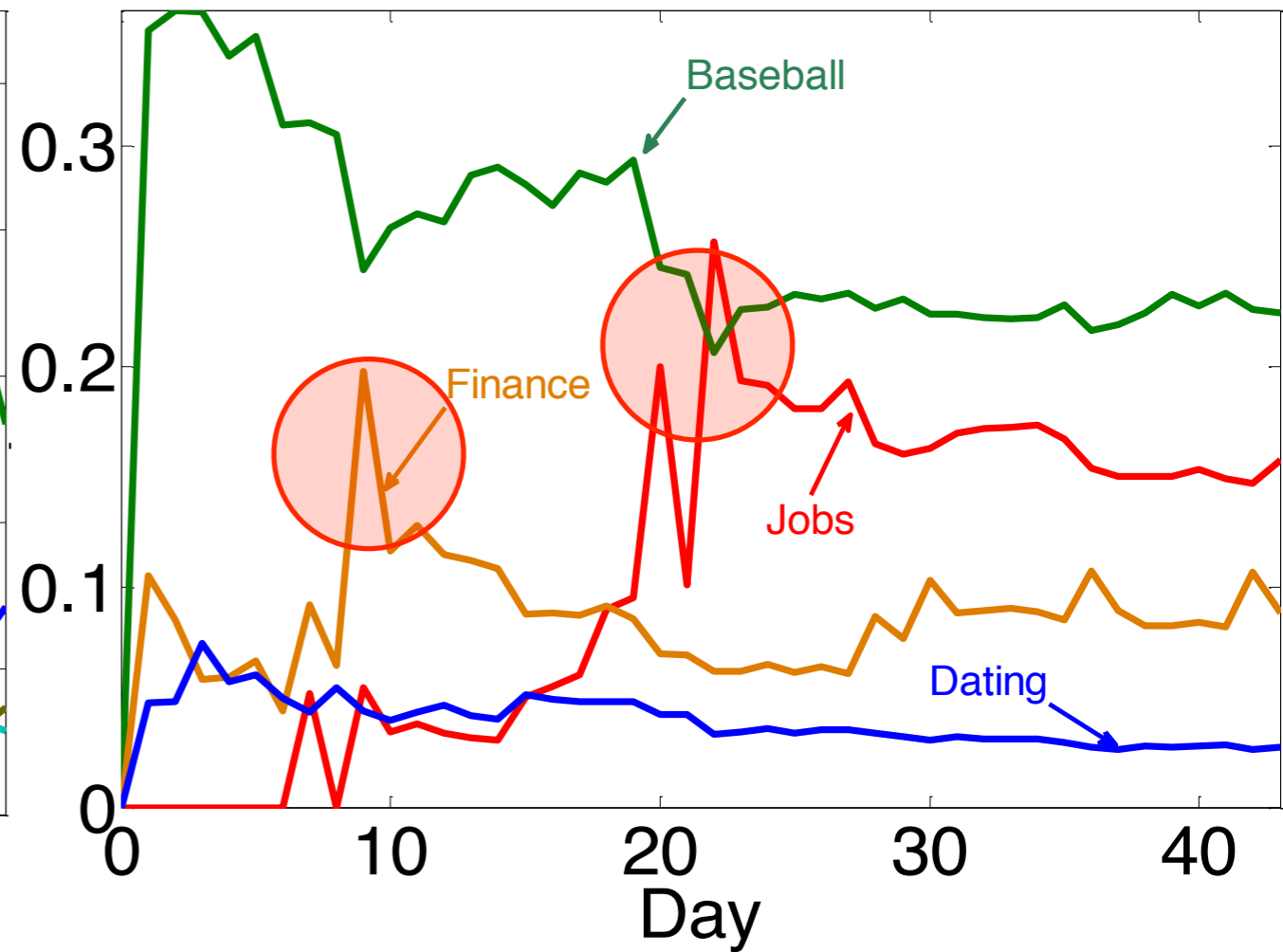
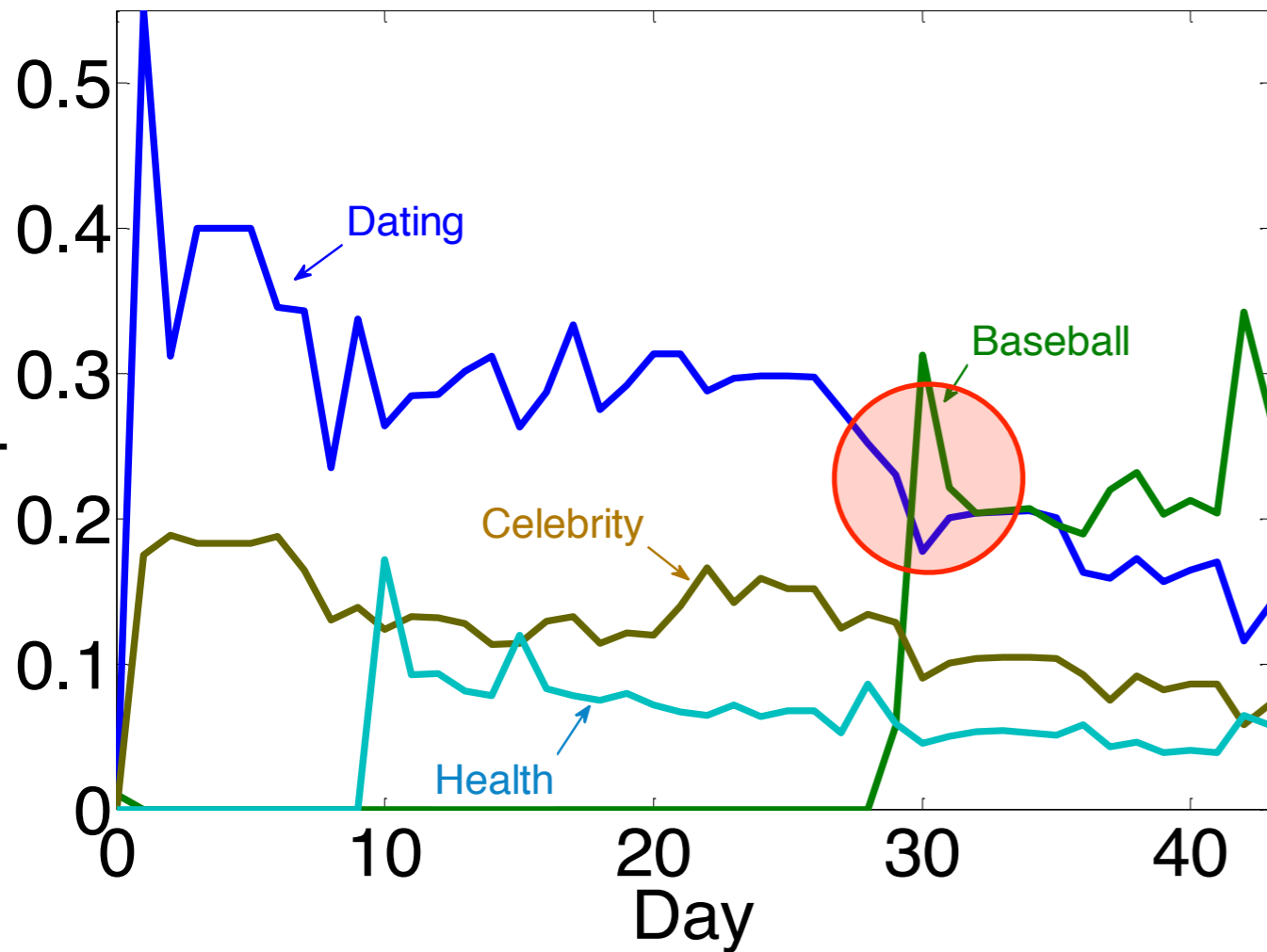
Jobs

job
career
business
assistant
hiring
part-time
receptionist

Finance

financial
Thomson
chart
real
Stock
Trading
currency

Sample users



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

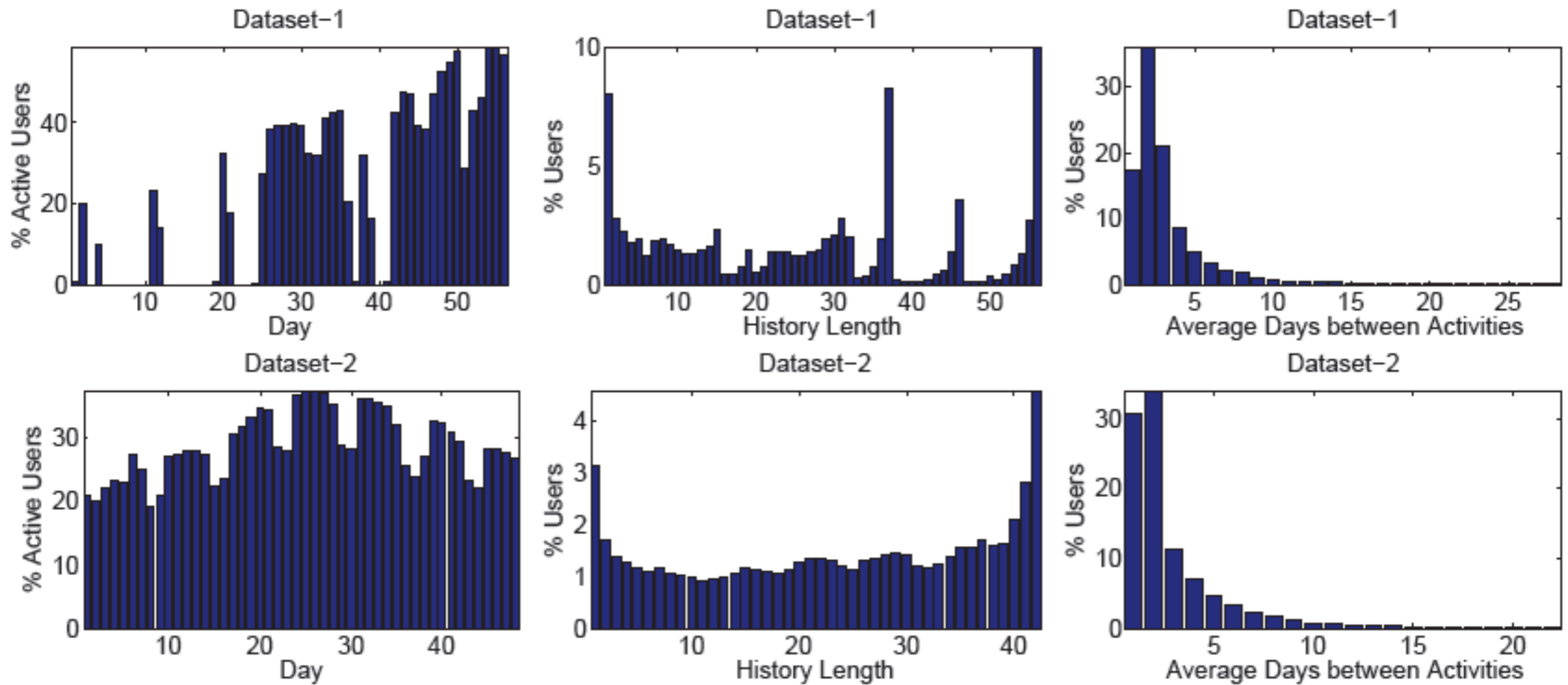
Jobs

job
career
business
assistant
hiring
part-time
receptionist

Finance

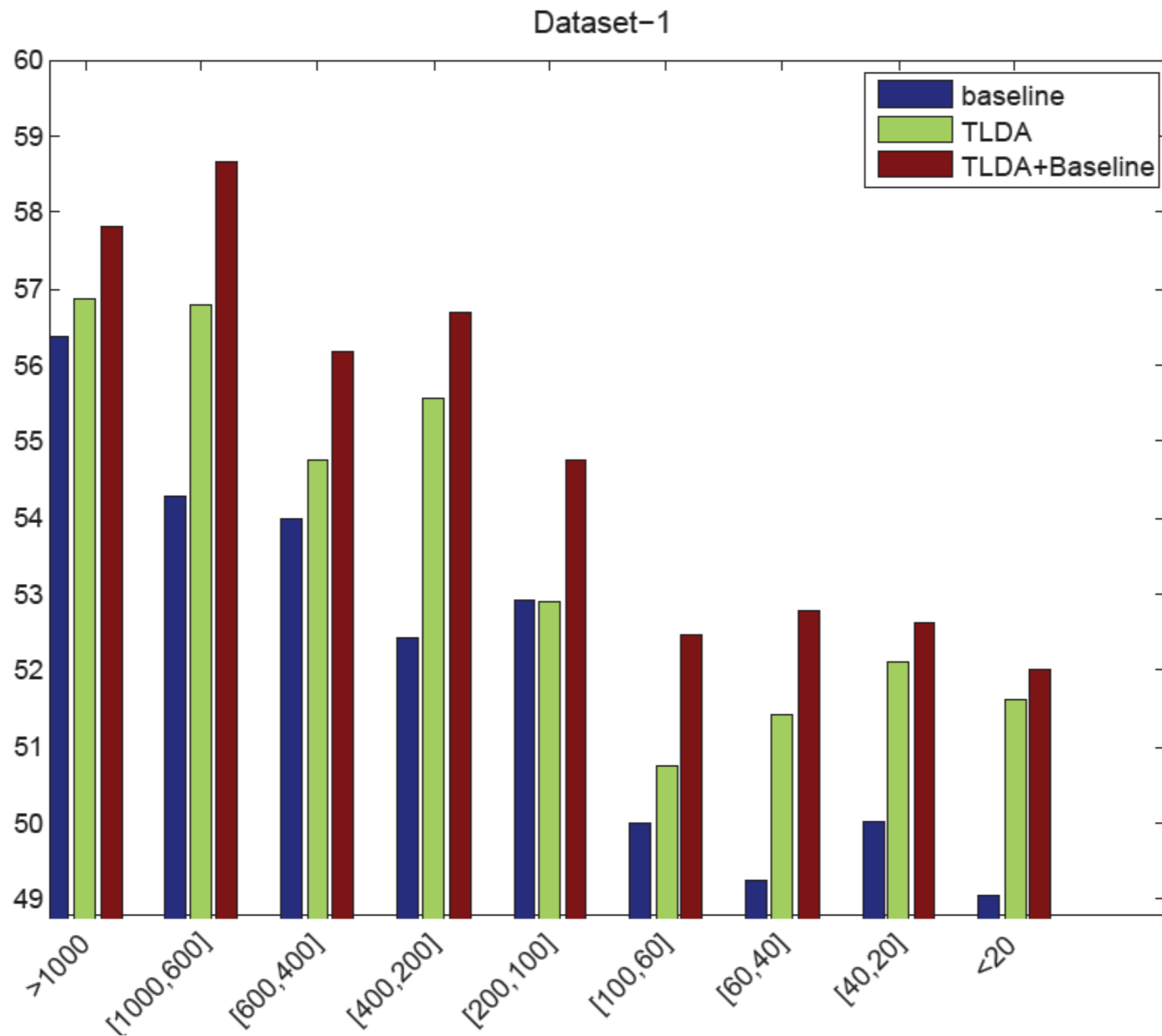
financial
Thomson
chart
real
Stock
Trading
currency

Data



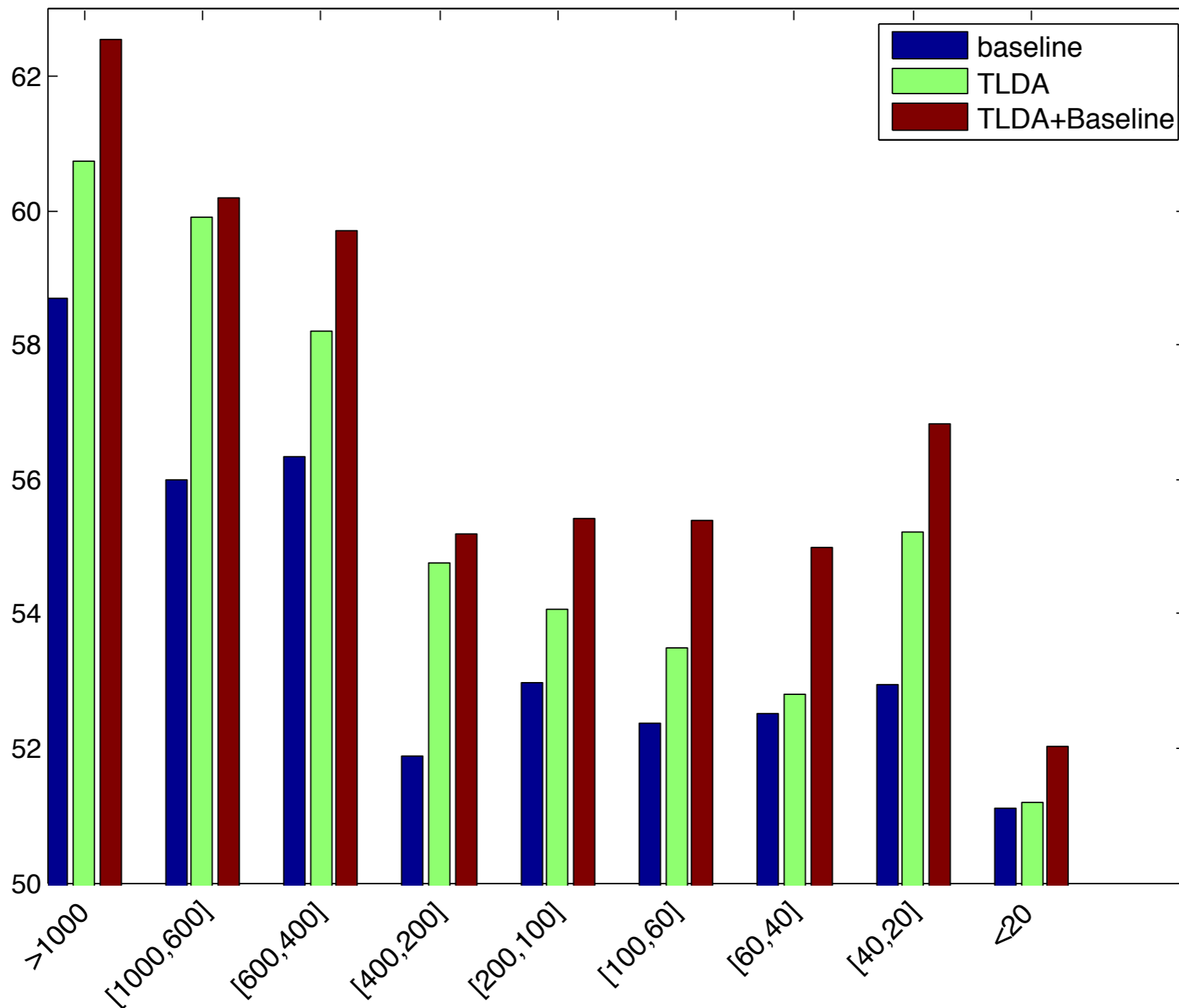
dataset	# days	# users	# campaigns	size
1	56	13.34M	241	242GB
2	44	33.5M	216	435GB

ROC score improvement

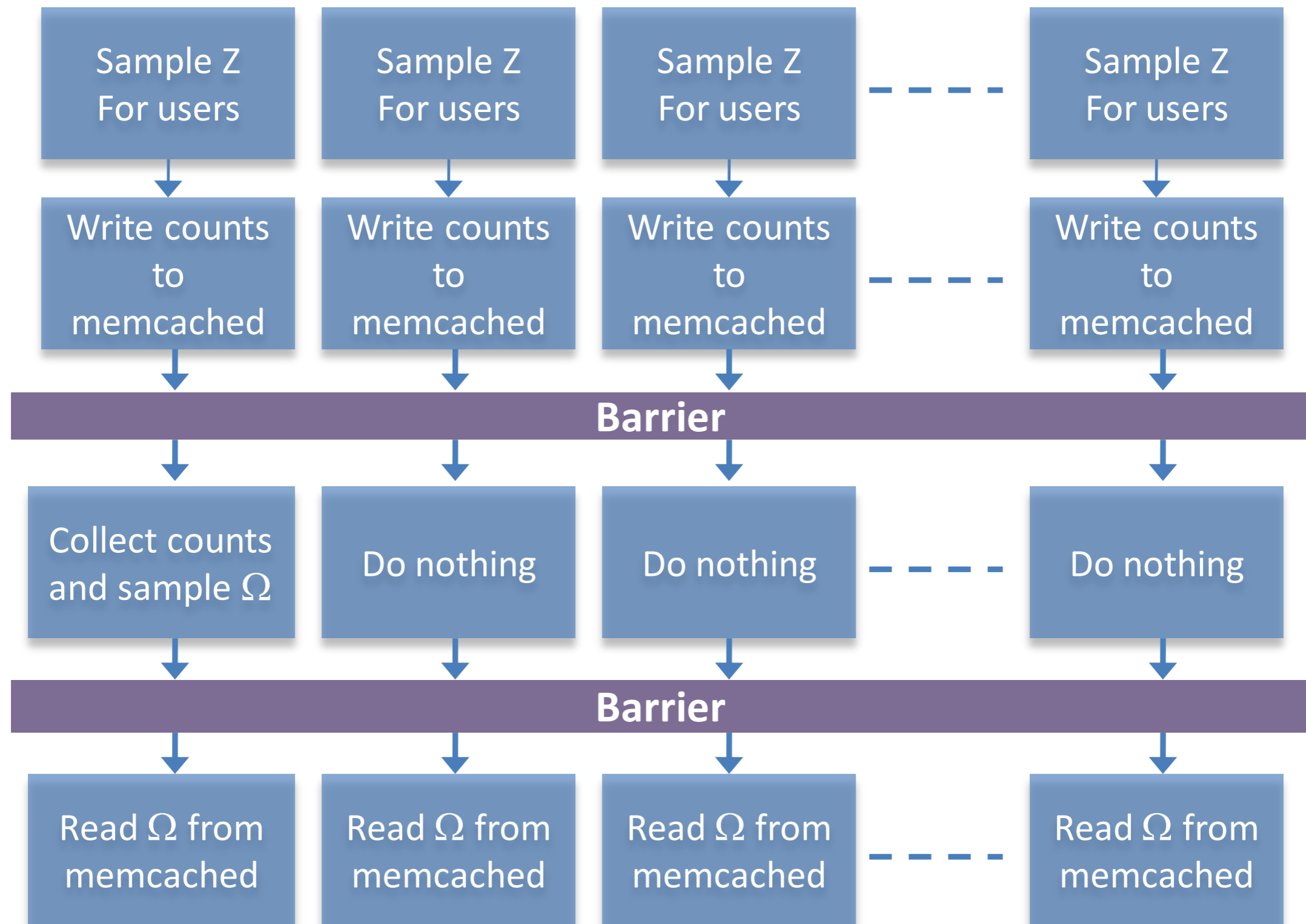


ROC score improvement

Dataset-2



LDA for user profiling



News

News Stream

News Stream



Add-ons turn tax cut bill into 'Christmas tree'

AP - 1 hr 32 mins ago
WASHINGTON - In the

BEYOND FOSSIL FUELS

Using Waste, Swedish



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

China says inflation up 5.1 per cent

Associated Press

Buzz up! 19 votes | Share



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

BEIJING - China's inflation surged Saturday, despite 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

er

Print

November, base food

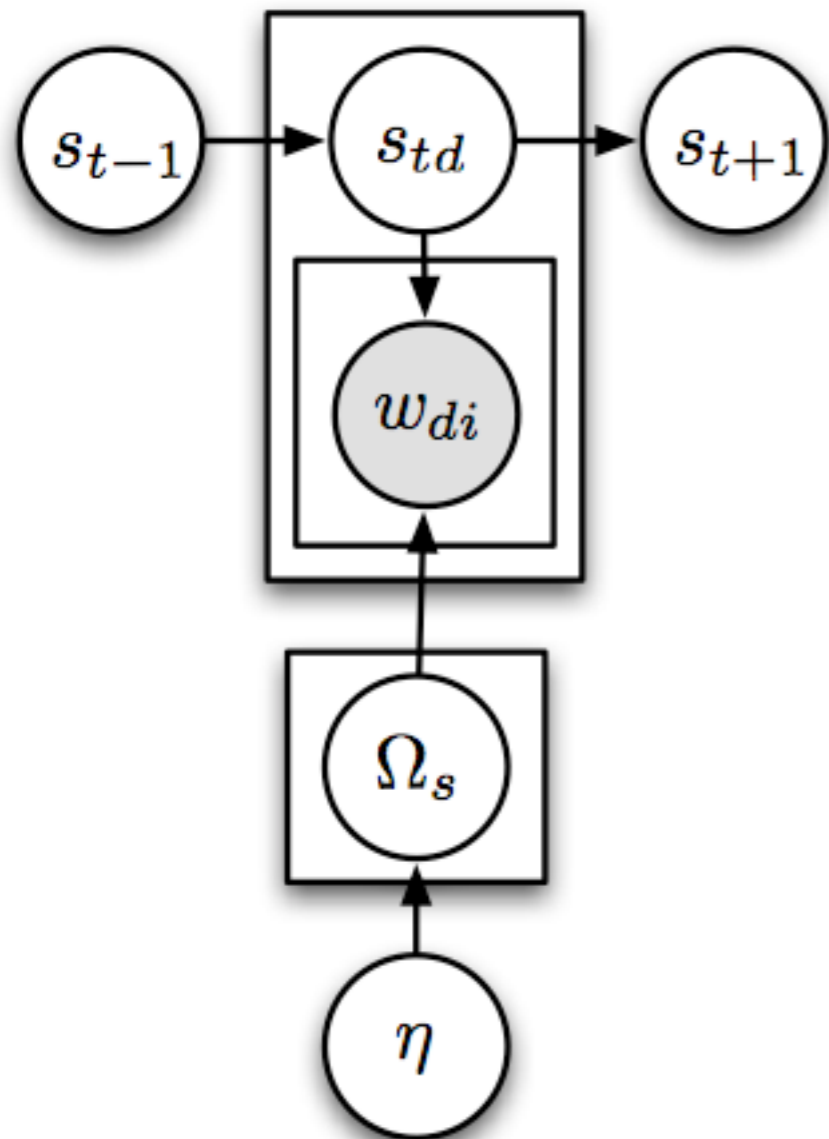
Johan Spanner for The New York Times

News Stream

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

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

Clustering / RCRP



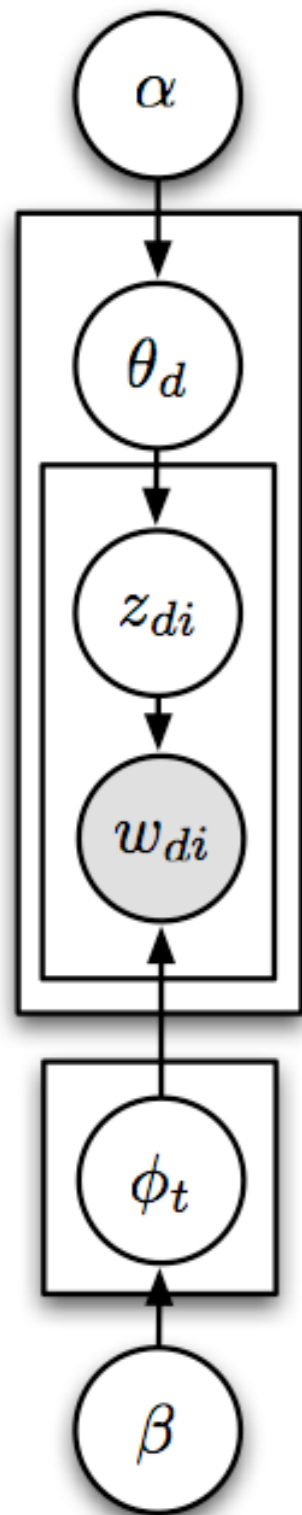
- Assume active story distribution at time t
- Draw story indicator
- Draw words from story distribution
- Down-weight story counts for next day

Ahmed & Xing, 2008

Clustering / RCRP

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

Latent Dirichlet Allocation



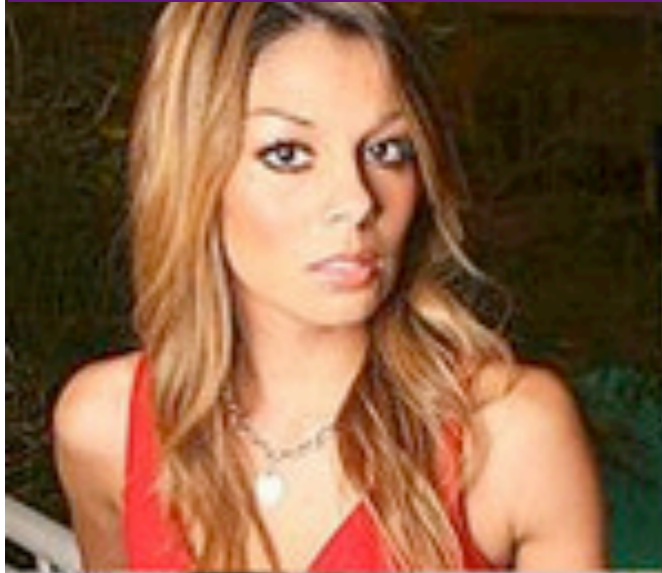
- Generate topic distribution per article
- Draw topics per word from topic distribution
- Draw words from topic specific word distribution

Blei, Ng, Jordan, 2003

Latent Dirichlet Allocation

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

More Issues



?



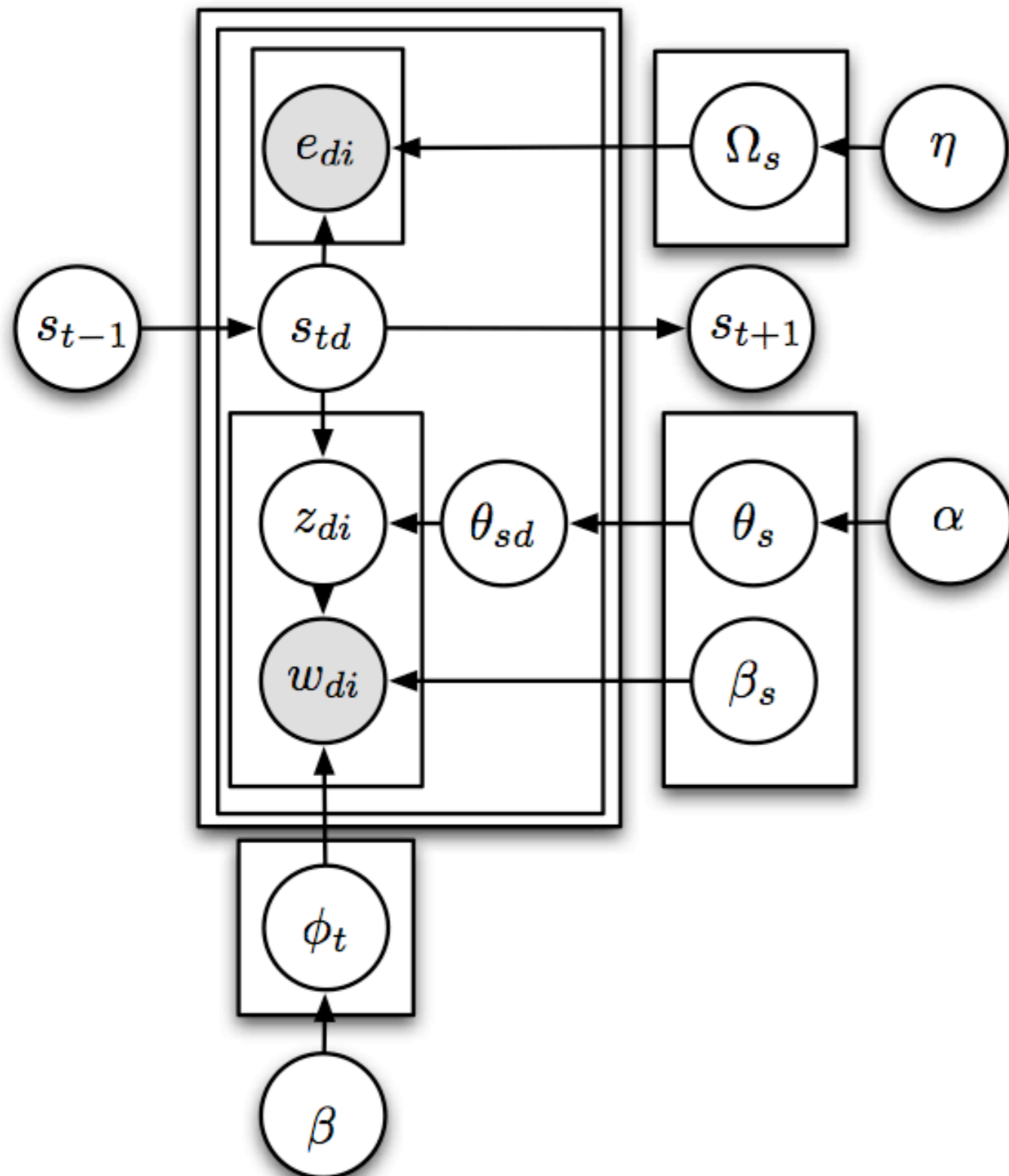
More Issues

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

Storylines

Amr Ahmed, Qirong Ho, Jake Eisenstein,
Alex Smola, Choon Hui Teo, 2011

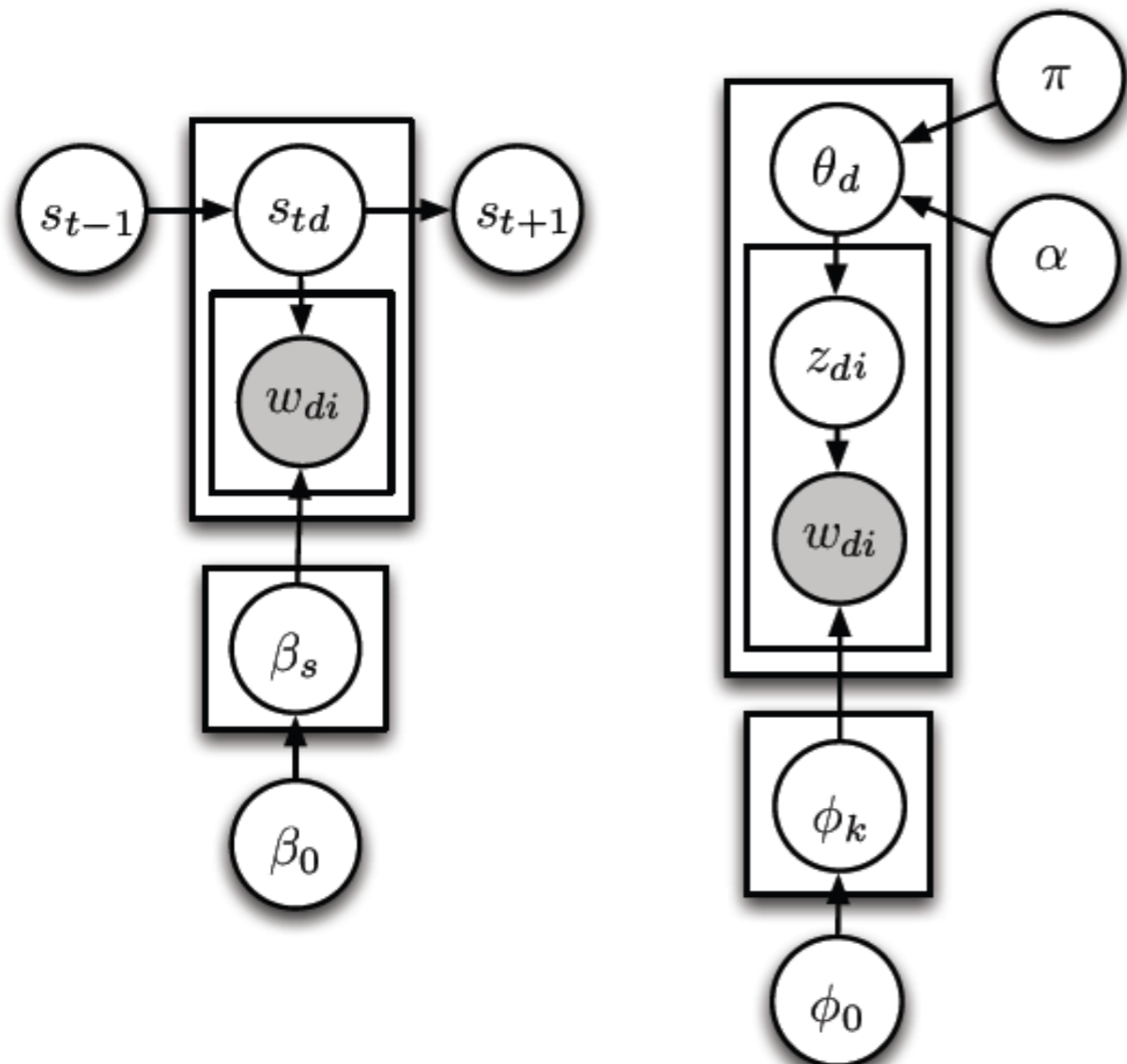
Storylines Model



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

Storylines Model

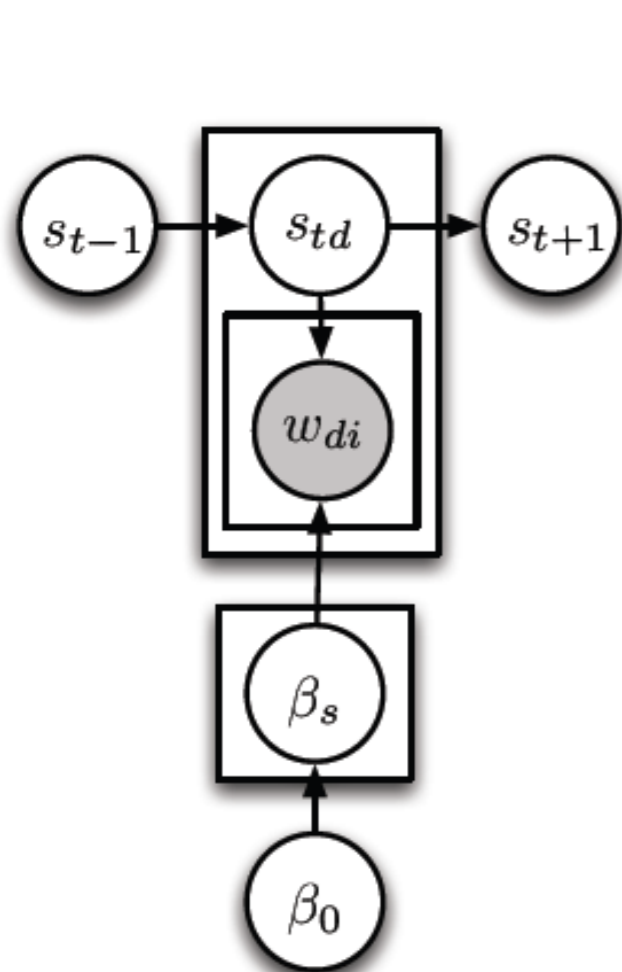
Storylines Model



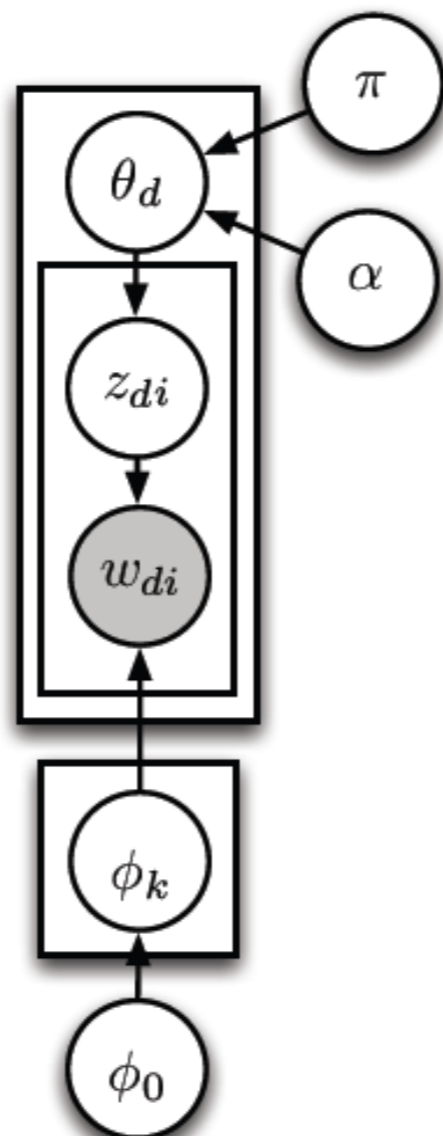
Tightly-focused

High-level
concepts

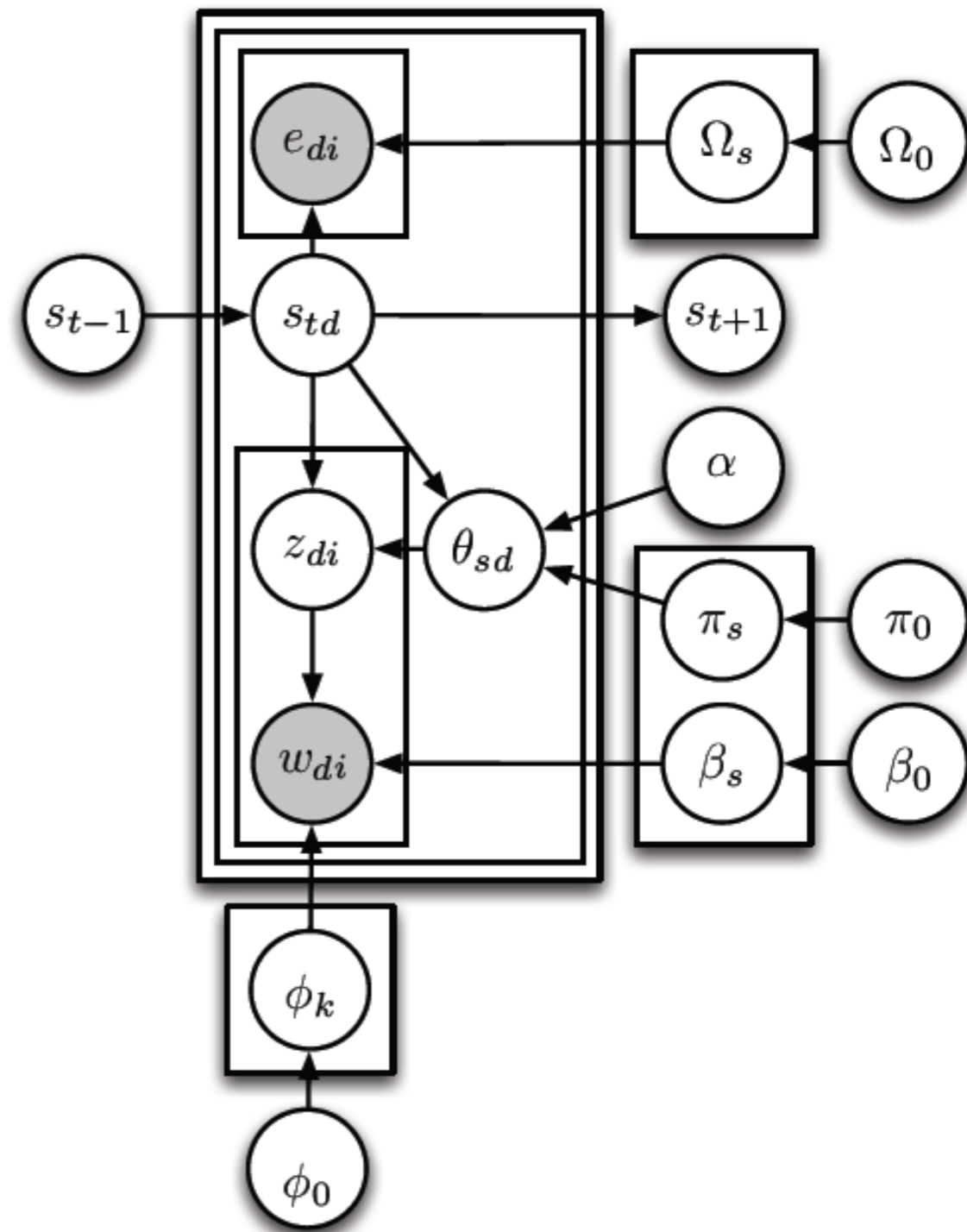
Storylines Model



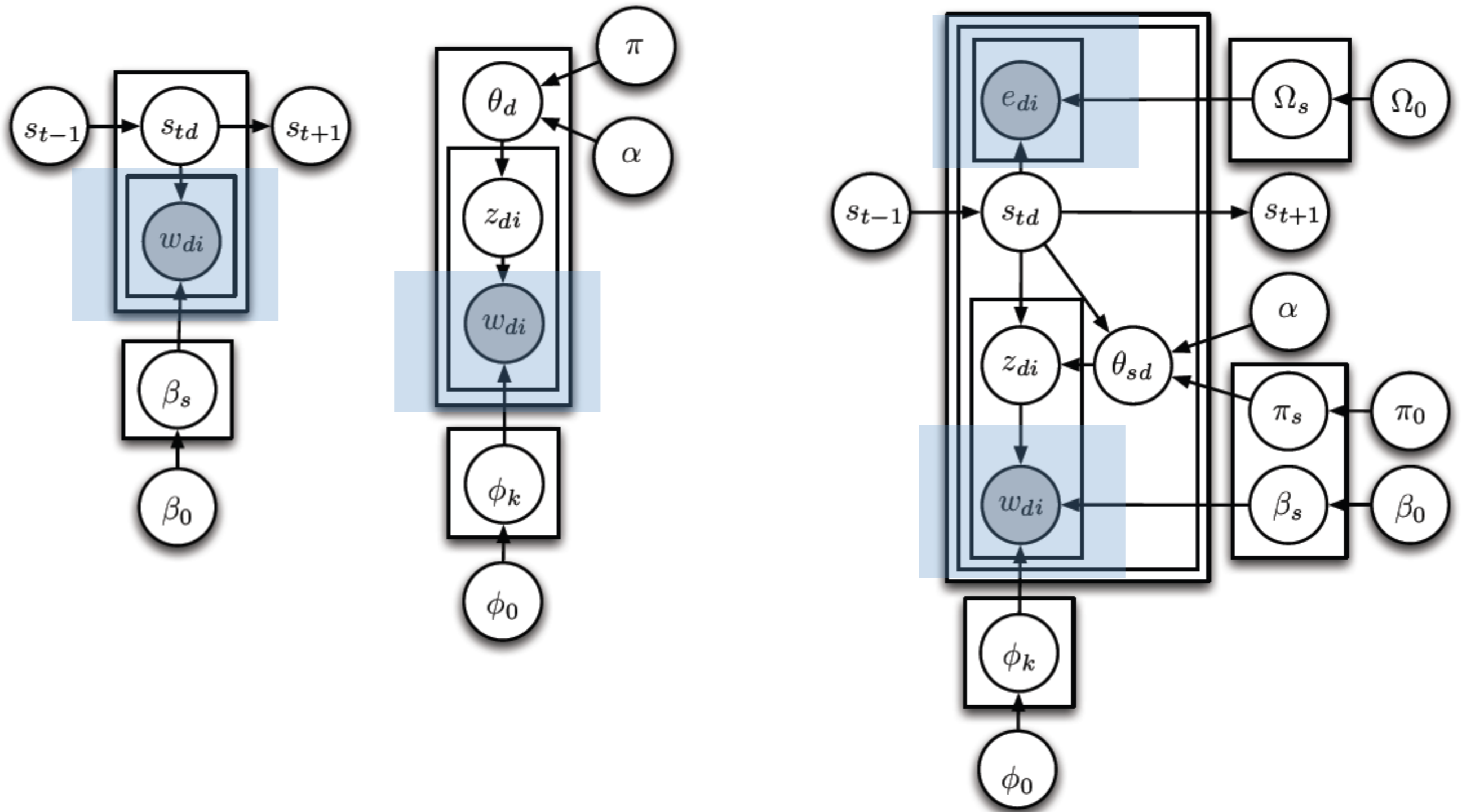
Tightly-focused



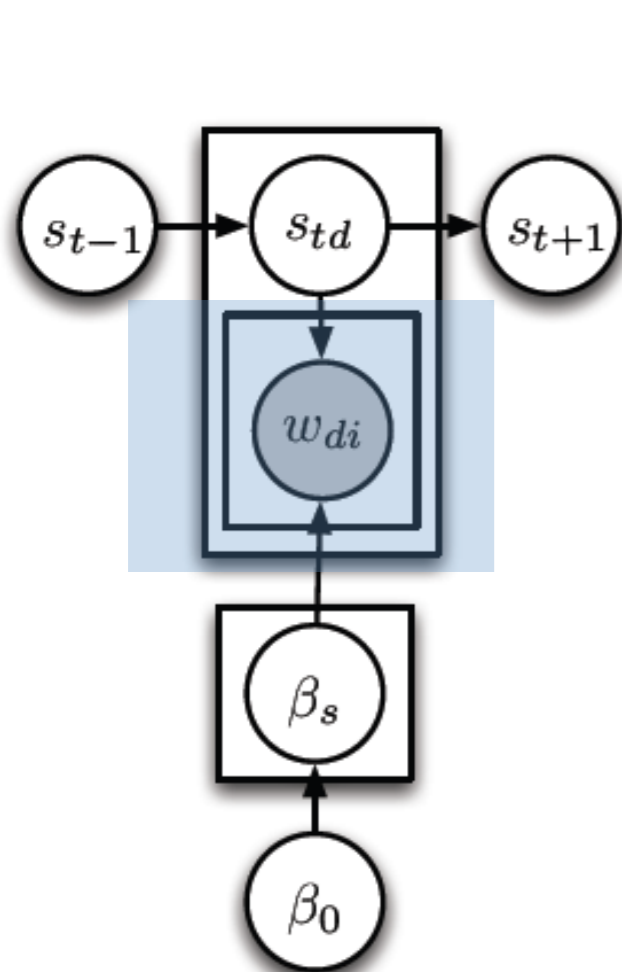
High-level concepts



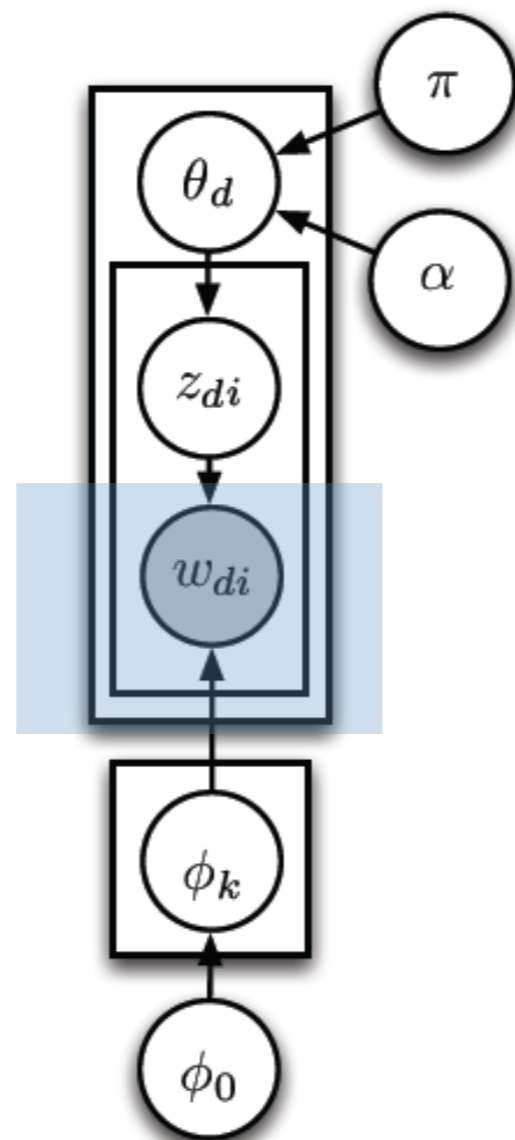
Storylines Model



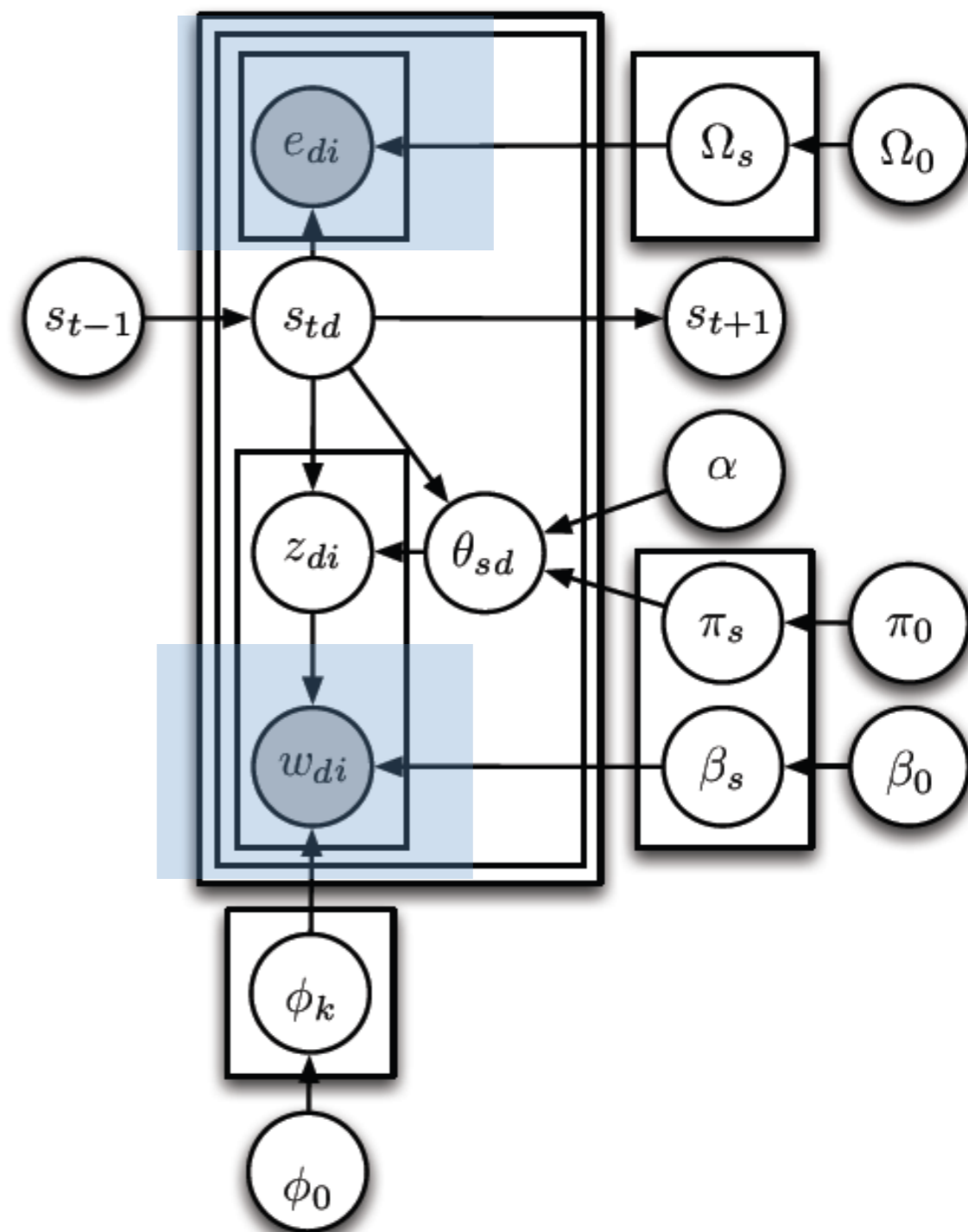
Storylines Model



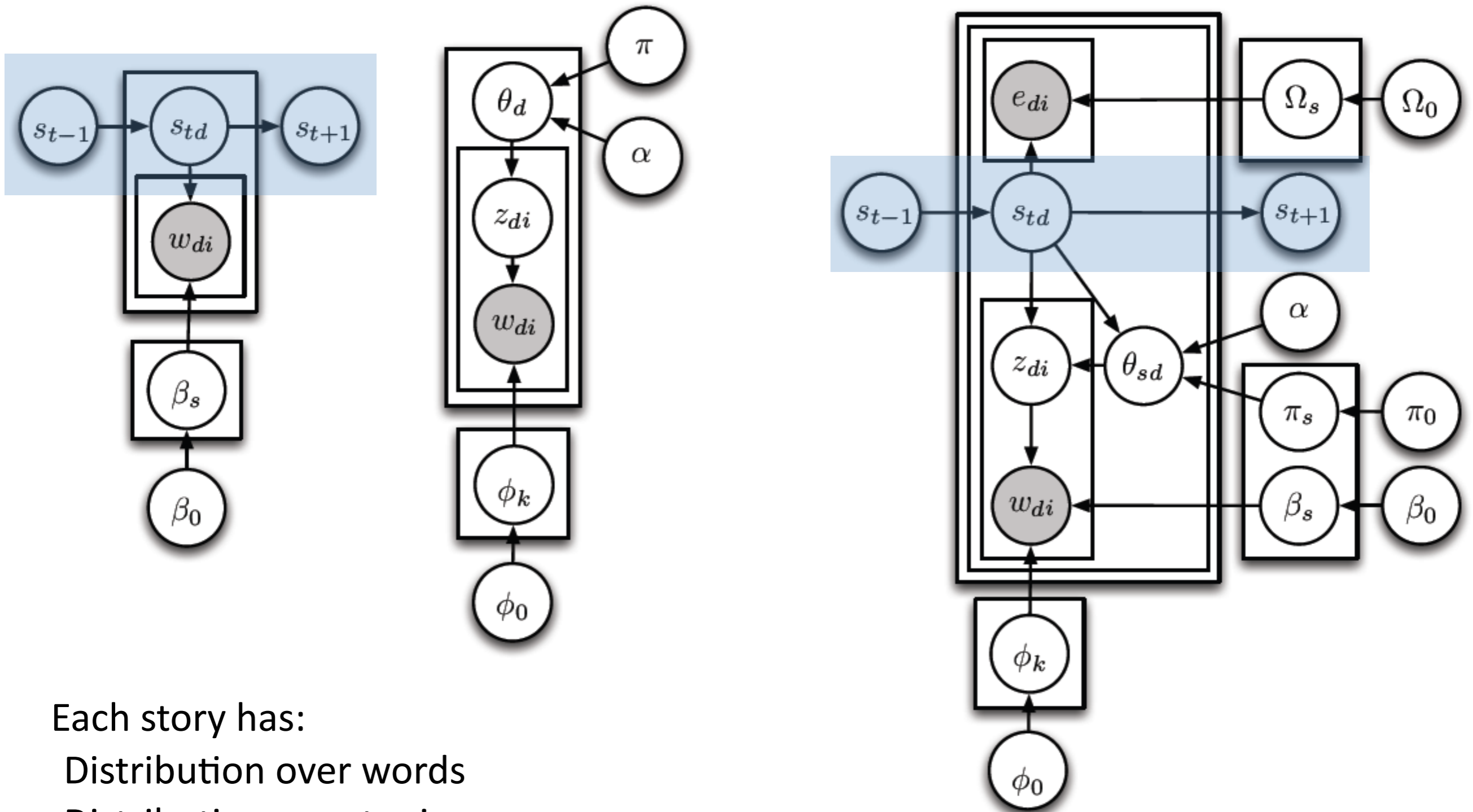
Tightly-focused



High-level concepts



Storylines Model



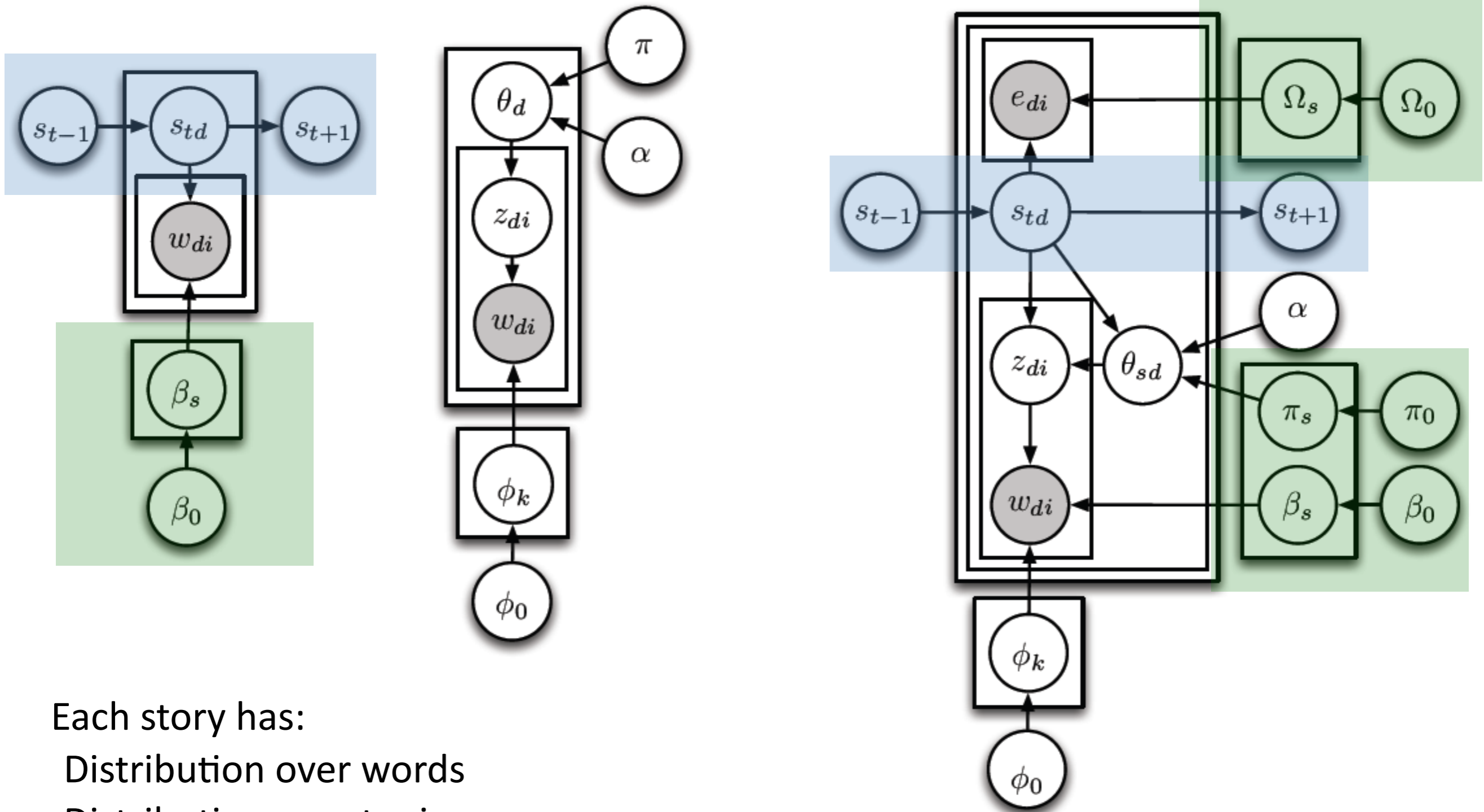
Each story has:

Distribution over words

Distribution over topics

Distribution over named entites

Storylines Model



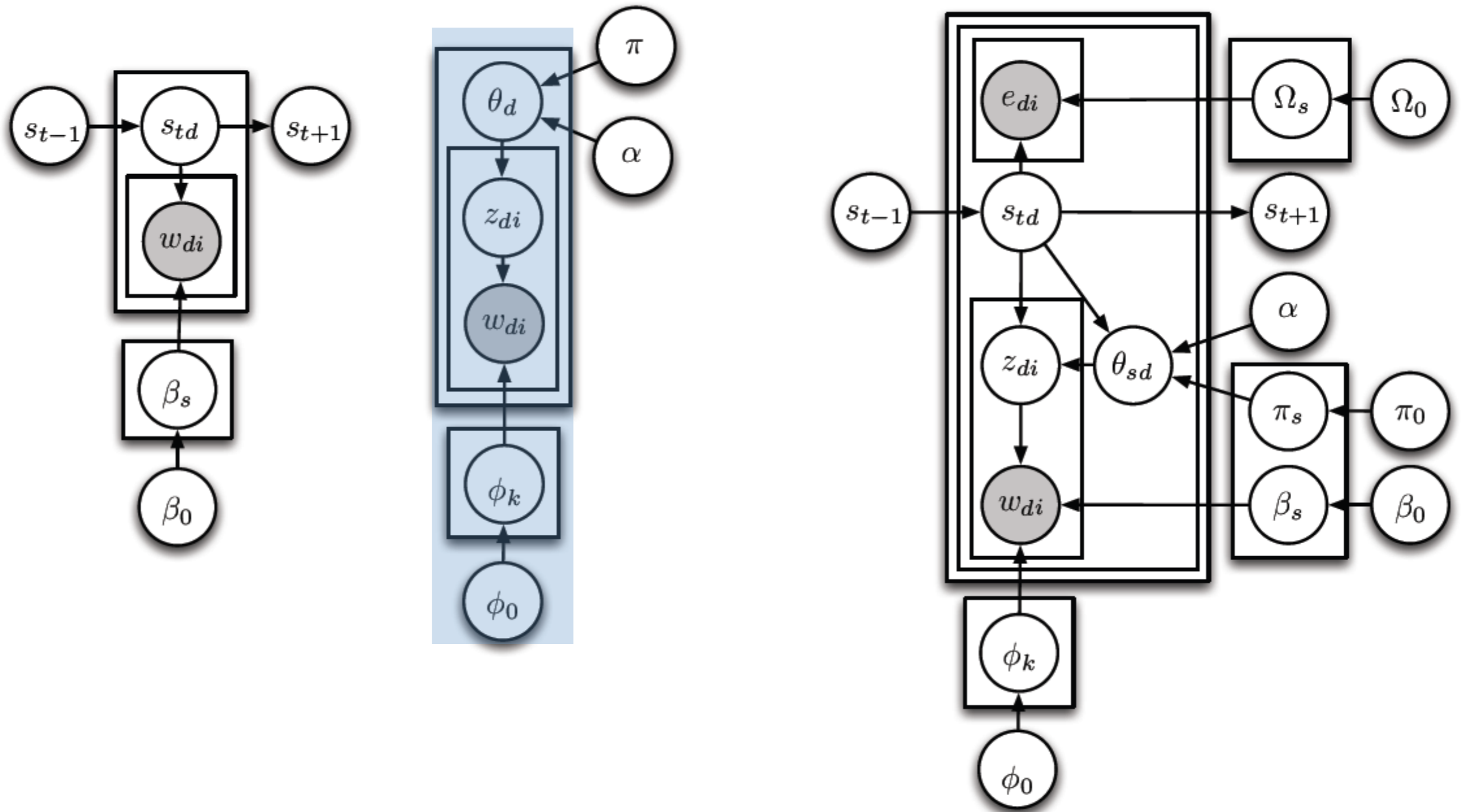
Each story has:

- Distribution over words

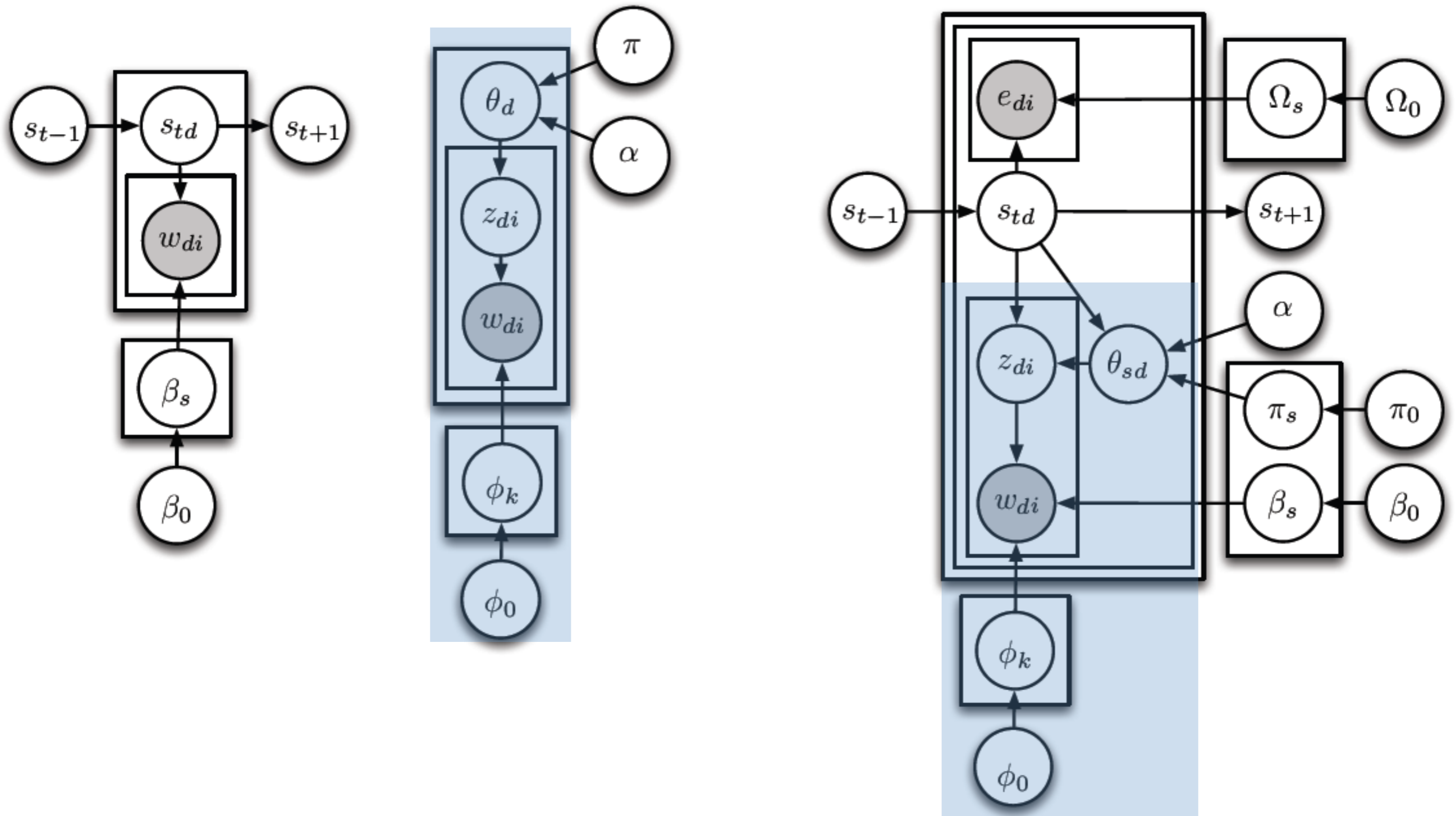
- Distribution over topics

- Distribution over named entites

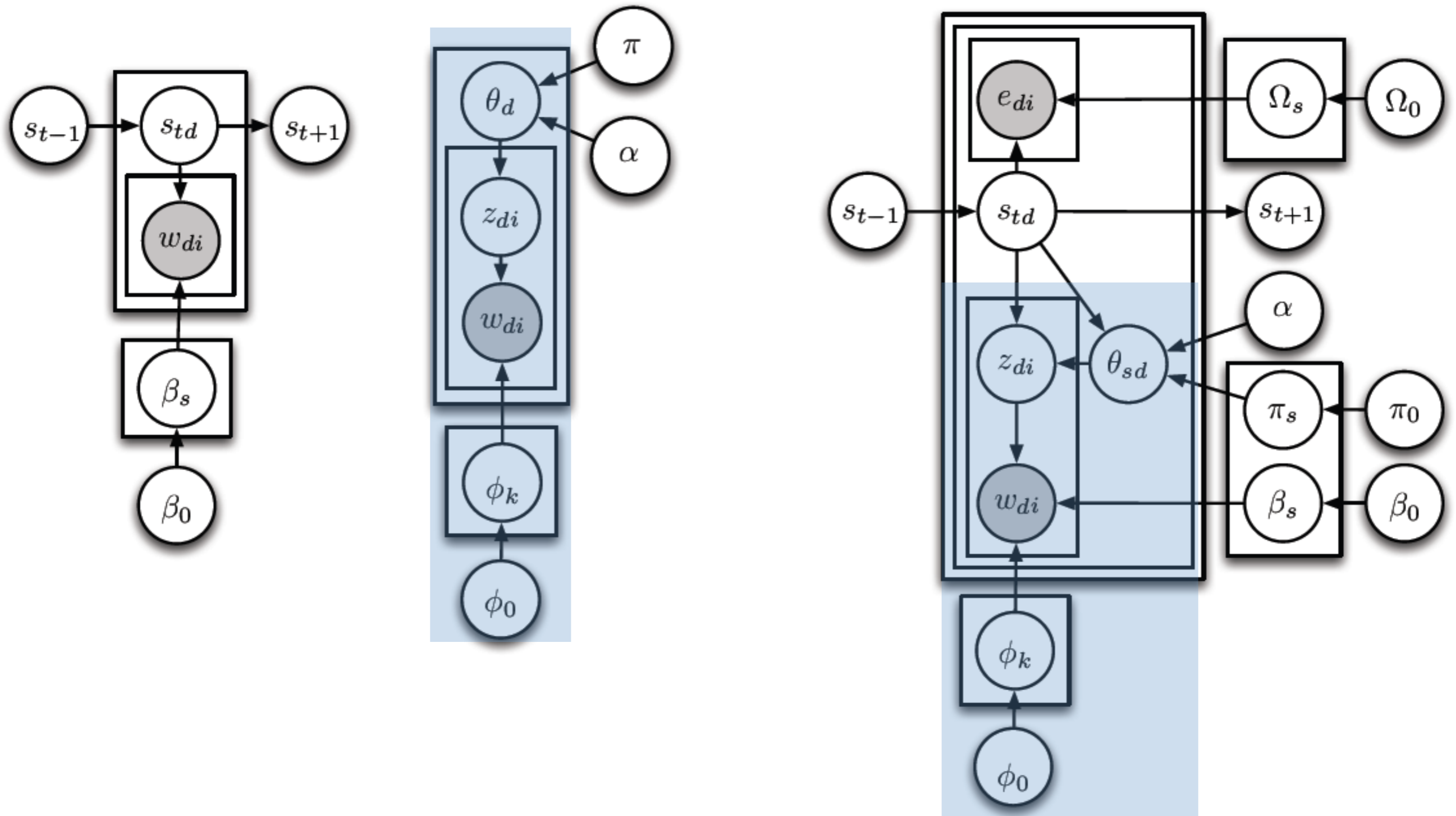
Storylines Model



Storylines Model



Storylines Model



Document's topic mix is sampled from its story prior
 Words inside a document either global or story specific

Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

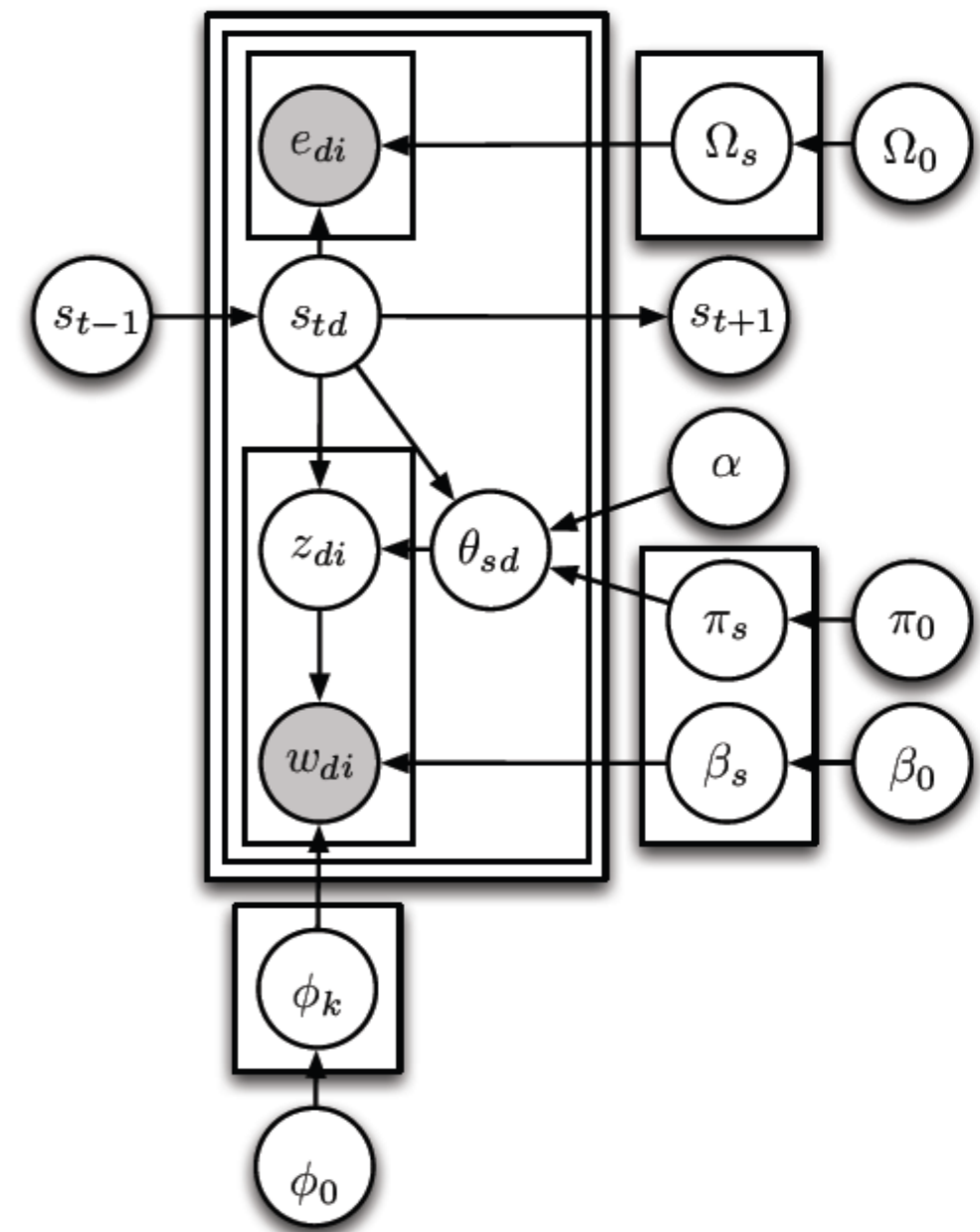
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

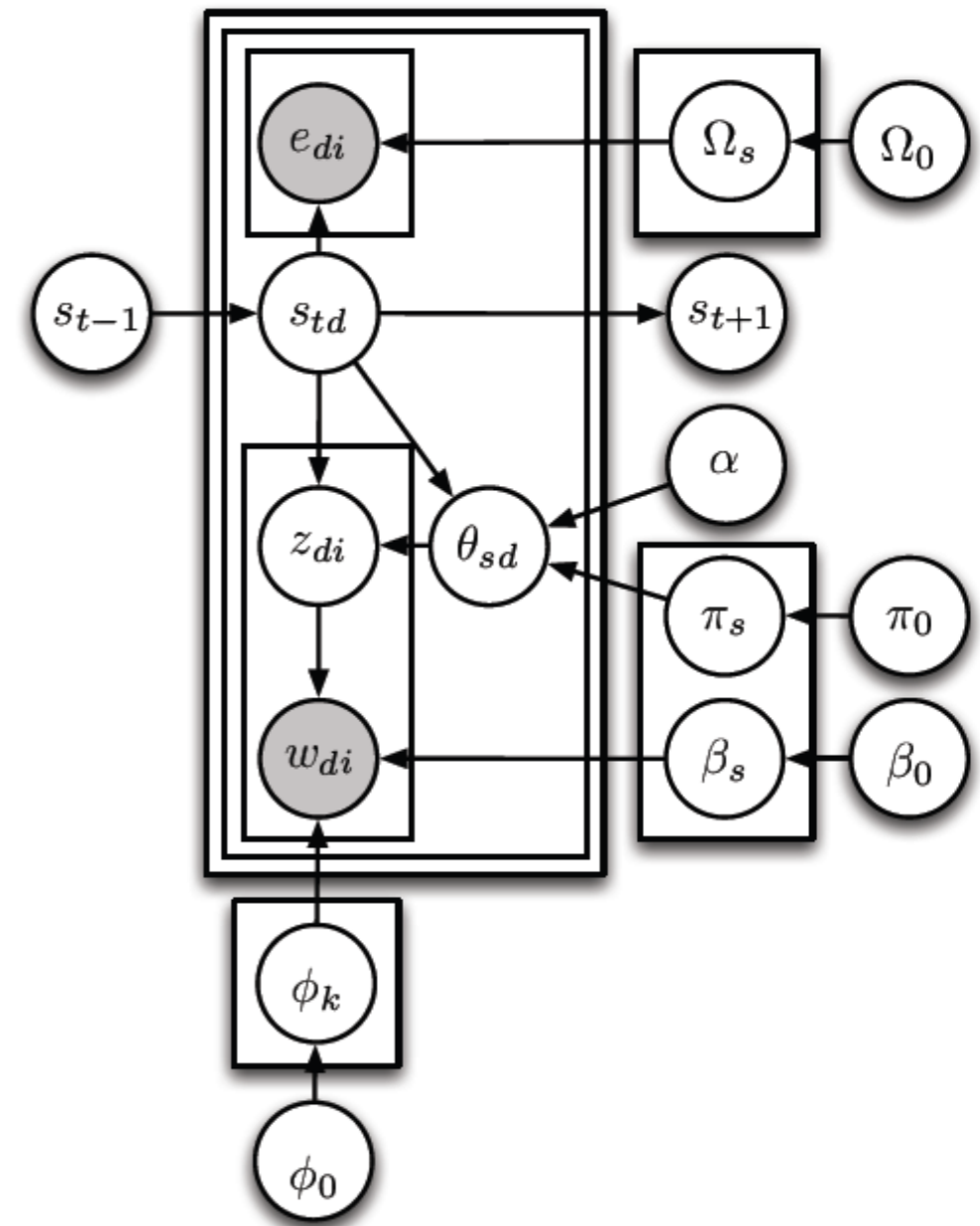
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

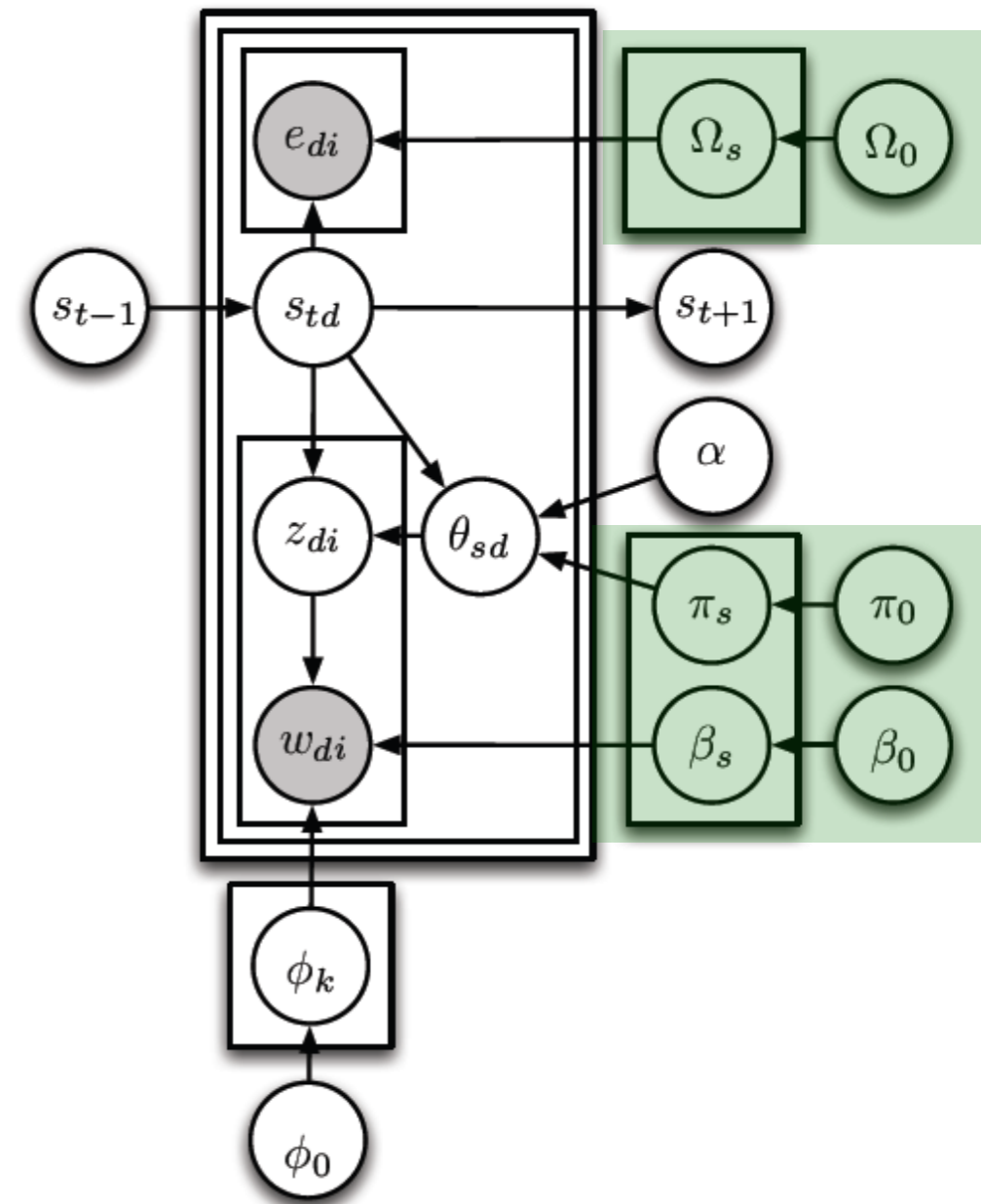
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

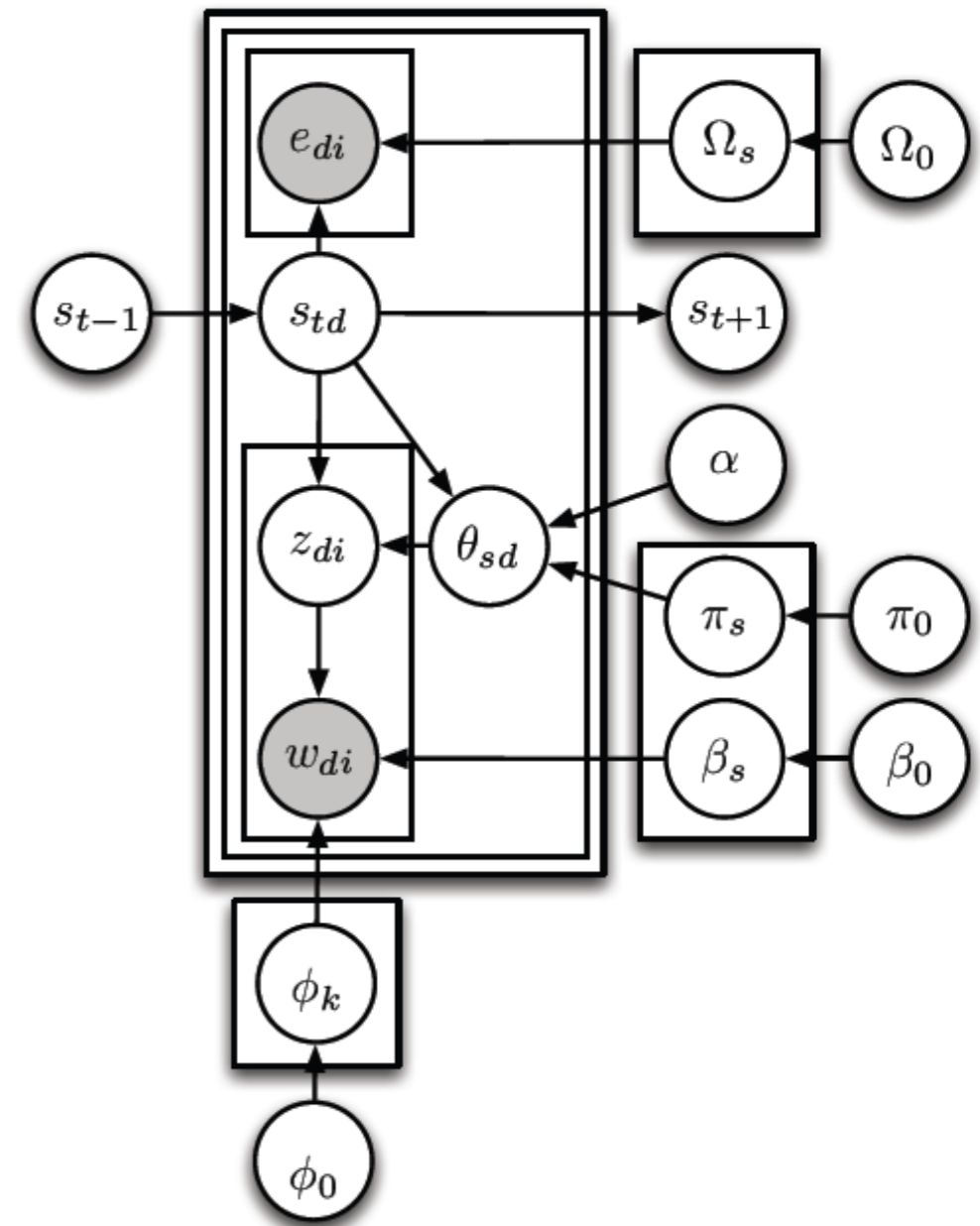
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

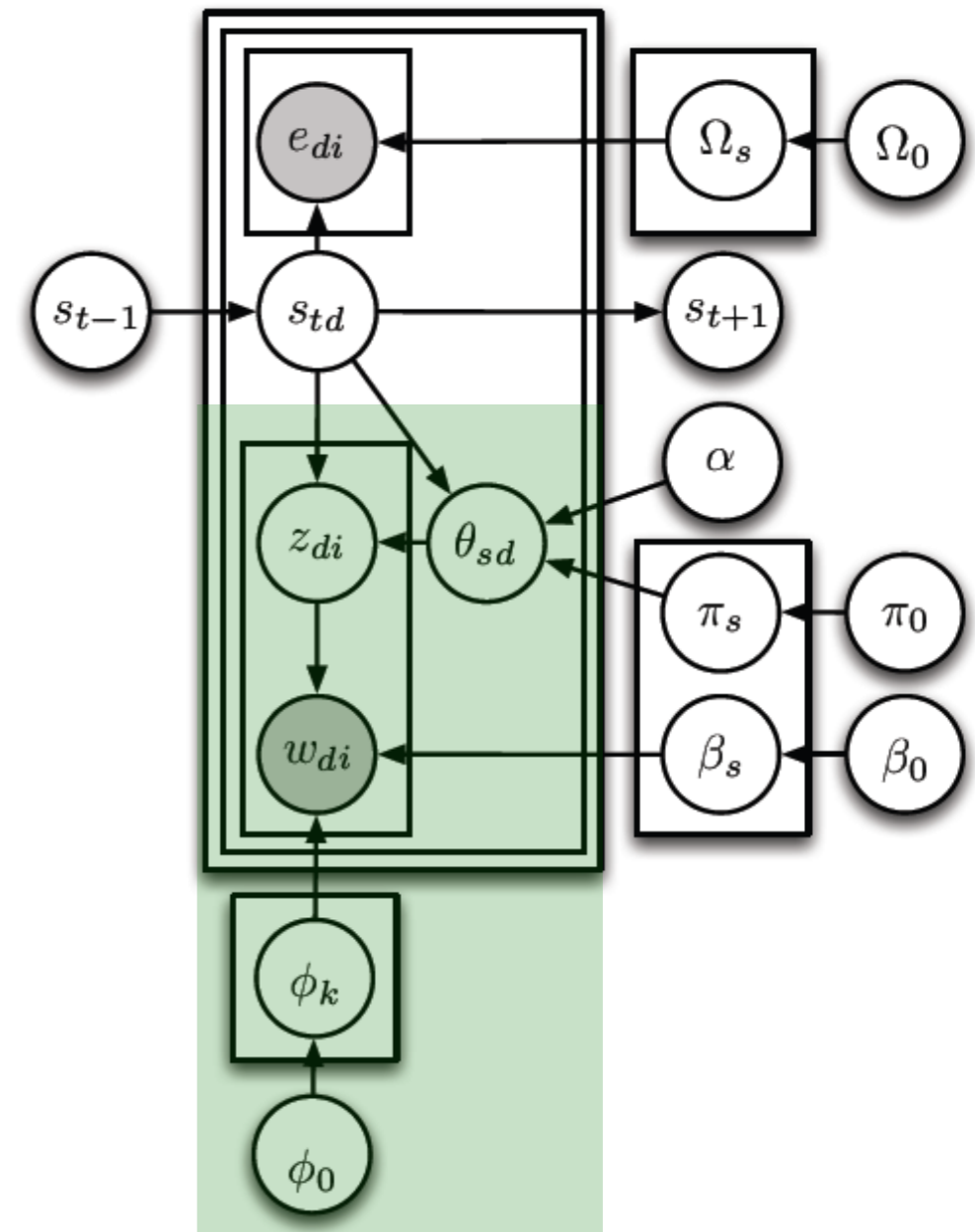
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

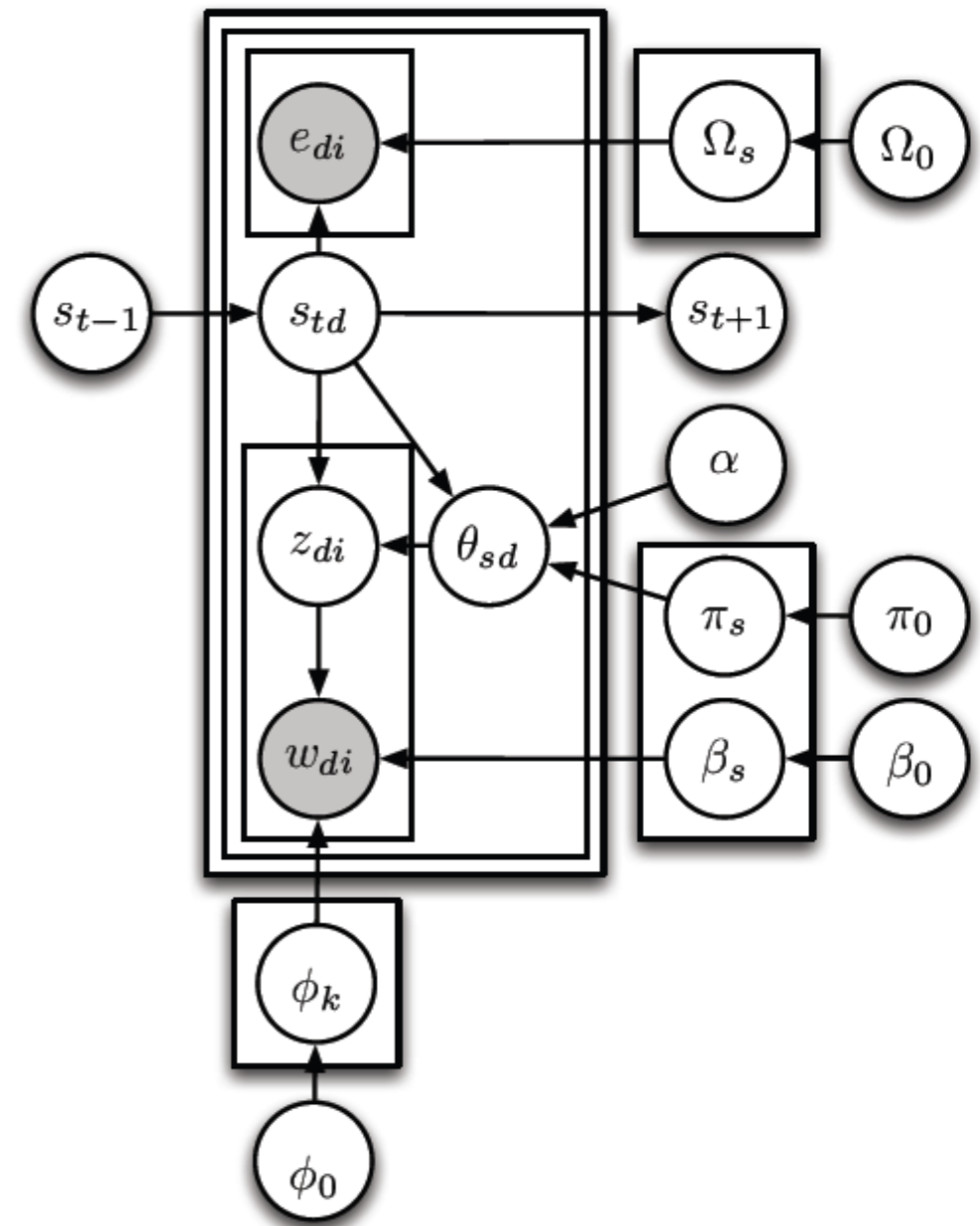
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Generative process

For each document $d \in \{1, \dots, D_t\}$:

(a) Draw the storyline indicator

$$s_{td} | \mathbf{s}_{1:t-1}, \mathbf{s}_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$$

(b) If s_{td} is a new storyline,

i. Draw a distribution over words

$$\beta_{s_{\text{new}}} | G_0 \sim \text{Dir}(\beta_0)$$

ii. Draw a distribution over named entities

$$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$$

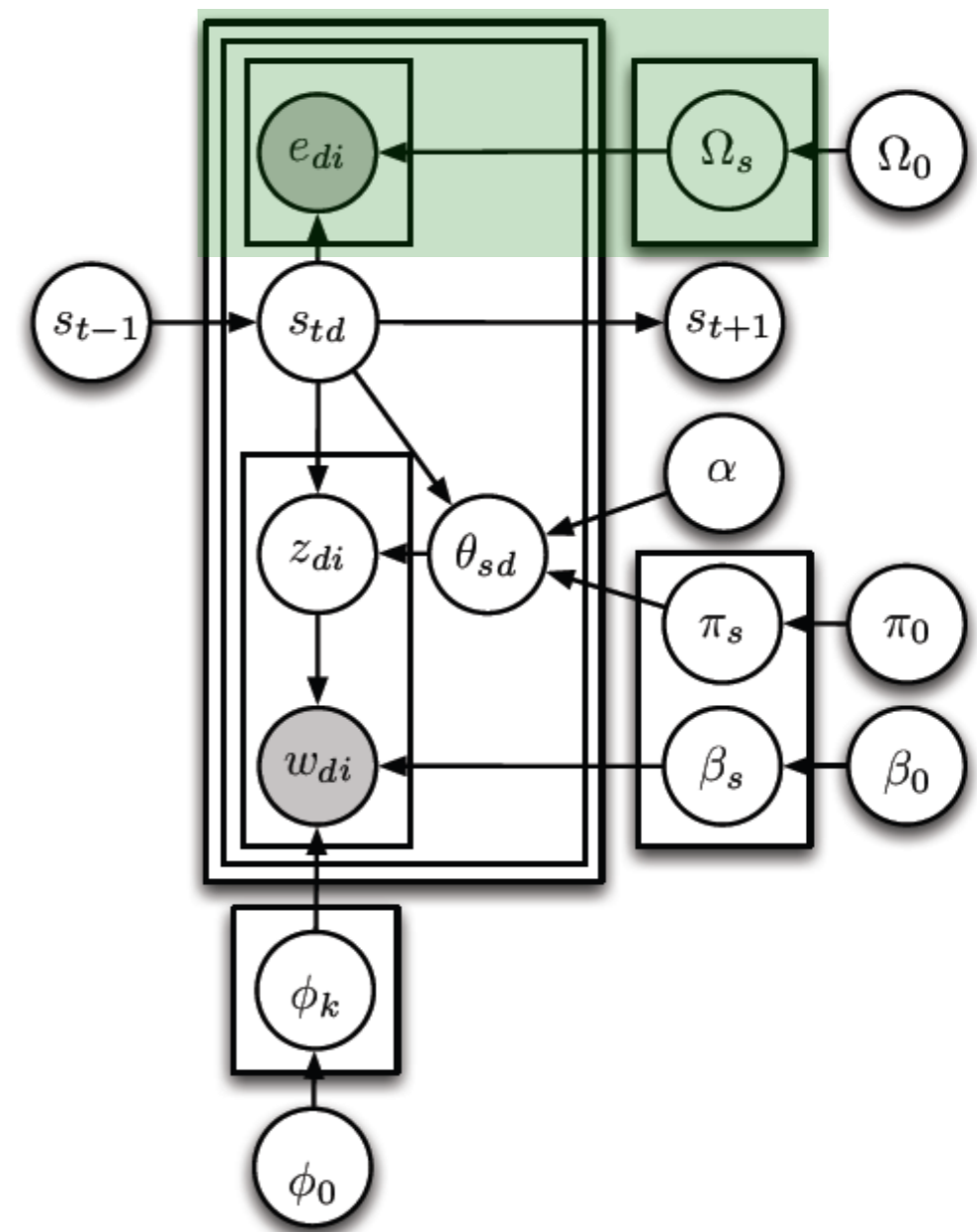
iii. Draw a Dirichlet distribution over topic proportions $\pi_{s_{\text{new}}} | G_0 \sim \text{Dir}(\pi_0)$

(c) Draw the topic proportions $\theta_{td} | s_{td} \sim \text{Dir}(\alpha \pi_{s_{td}})$

(d) Draw the words

$$\mathbf{w}_{td} | s_{td} \sim \text{LDA}(\theta_{s_{td}}, \{\phi_1, \dots, \phi_K, \beta_{s_{td}}\})$$

(e) Draw the named entities $\mathbf{e}_{td} | s_{td} \sim \text{Mult}(\Omega_{s_{td}})$



Estimation

- Sequential Monte Carlo (Particle Filter)
- For new time period draw stories s , topics z

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

using Gibbs Sampling for each particle

- Reweight particle via

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

- Regenerate particles if l2 norm too heavy

Numbers ...

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

time	entities	topics	story words
0.84	0.90	0.86	0.75

This is the best performance on TDT5!

- **Yahoo News data**
... beats all other clustering algorithms

Stories

TOPICS

Sports

games
won
team
final
season
league
held

Politics

government
minister
authorities
opposition
officials
leaders
group

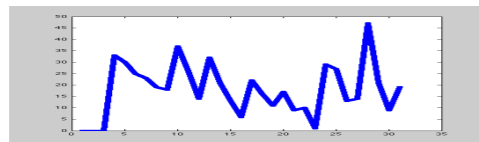
Unrest

police
attack
run
man
group
arrested
move

STORYLINES

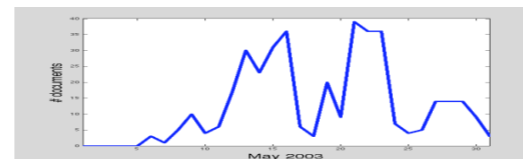
UEFA-soccer

champions	<i>Juventus</i>
goal	<i>AC Milan</i>
leg	<i>Real Madrid</i>
coach	<i>Milan</i>
striker	<i>Lazio</i>
midfield	<i>Ronaldo</i>
penalty	<i>Lyon</i>



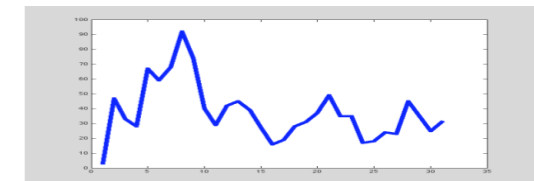
Tax bills

tax	<i>Bush</i>
billion	<i>Senate</i>
cut	<i>US</i>
plan	<i>Congress</i>
budget	<i>Fleischer</i>
economy	<i>White House</i>
lawmakers	<i>Republican</i>



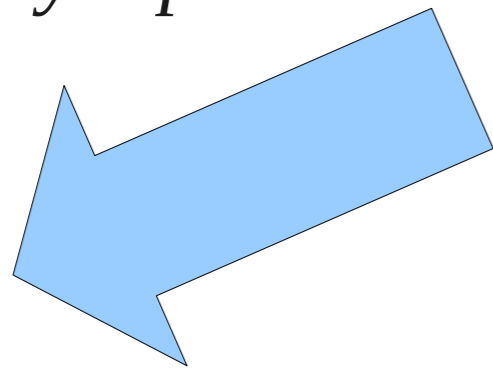
India-Pakistan tension

nuclear	<i>Pakistan</i>
border	<i>India</i>
dialogue	<i>Kashmir</i>
diplomatic	<i>New Delhi</i>
militant	<i>Islamabad</i>
insurgency	<i>Musharraf</i>
missile	<i>Vajpayee</i>



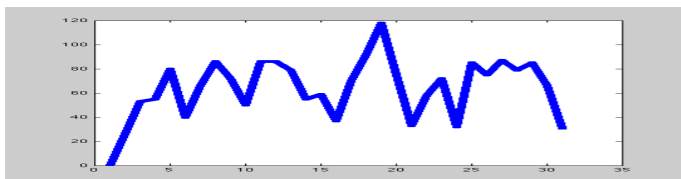
Related Stories

“Show similar stories by topic”



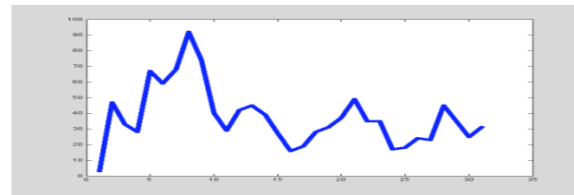
Middle-east conflict

Peace	<i>Israel</i>
Roadmap	<i>Palestinian</i>
Suicide	<i>West bank</i>
Violence	<i>Sharon</i>
Settlements	<i>Hamas</i>
bombing	<i>Arafat</i>

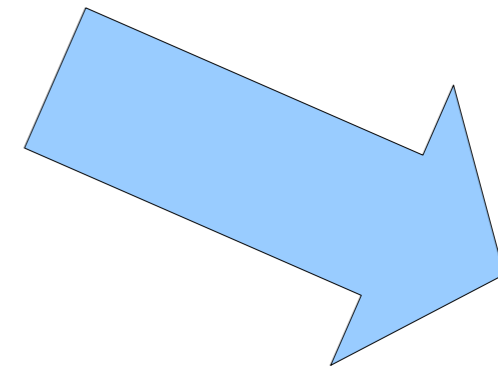


India-Pakistan tension

nuclear	<i>Pakistan</i>
border	<i>India</i>
dialogue	<i>Kashmir</i>
diplomatic	<i>New Delhi</i>
militant	<i>Islamabad</i>
insurgency	<i>Musharraf</i>
missile	<i>Vajpayee</i>

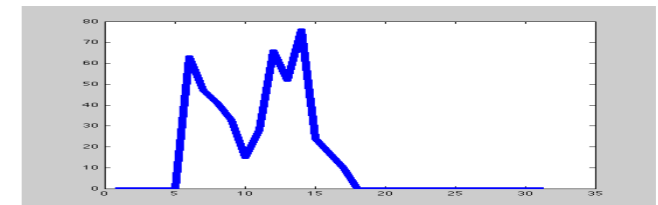


“Show similar stories, require the word nuclear”



North Korea nuclear

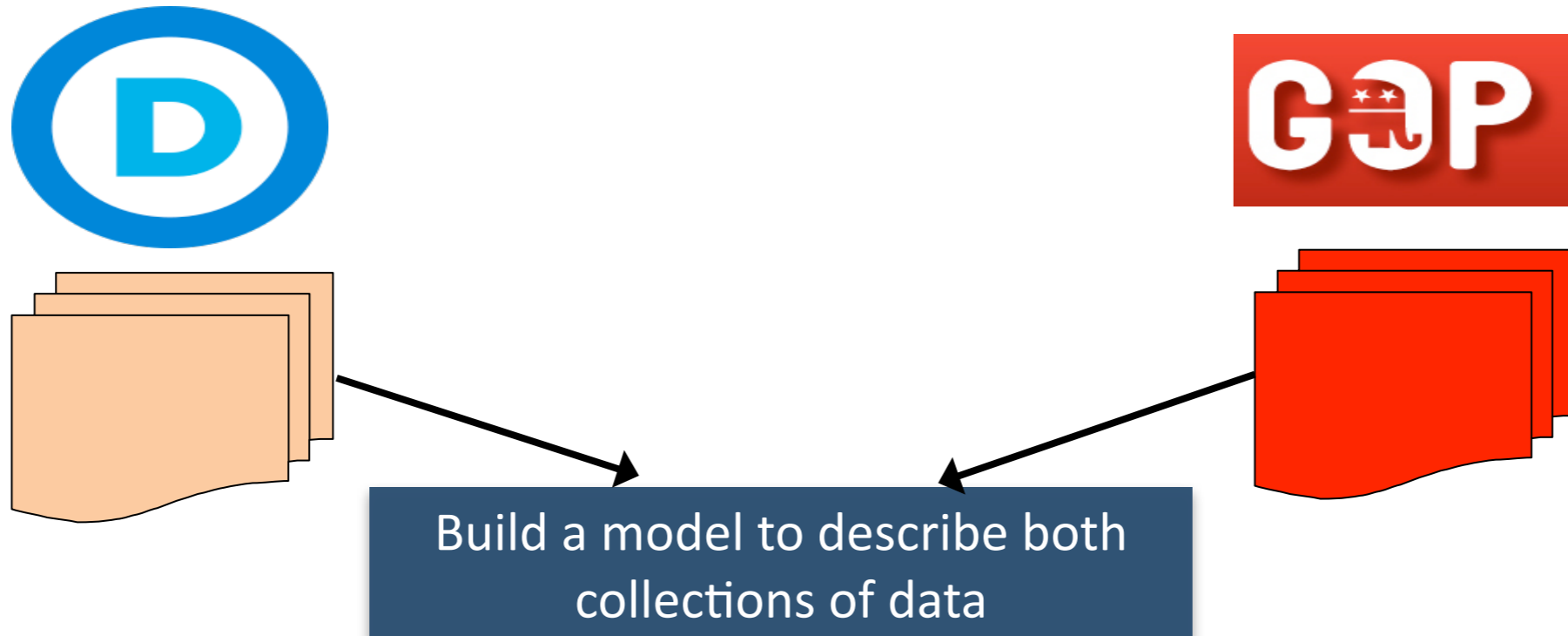
nuclear	<i>North Korea</i>
summit	<i>South Korea</i>
warning	<i>U.S</i>
policy	<i>Bush</i>
missile	<i>Pyongyang</i>
program	



Detecting Ideologies

Ahmed and Xing, 2010

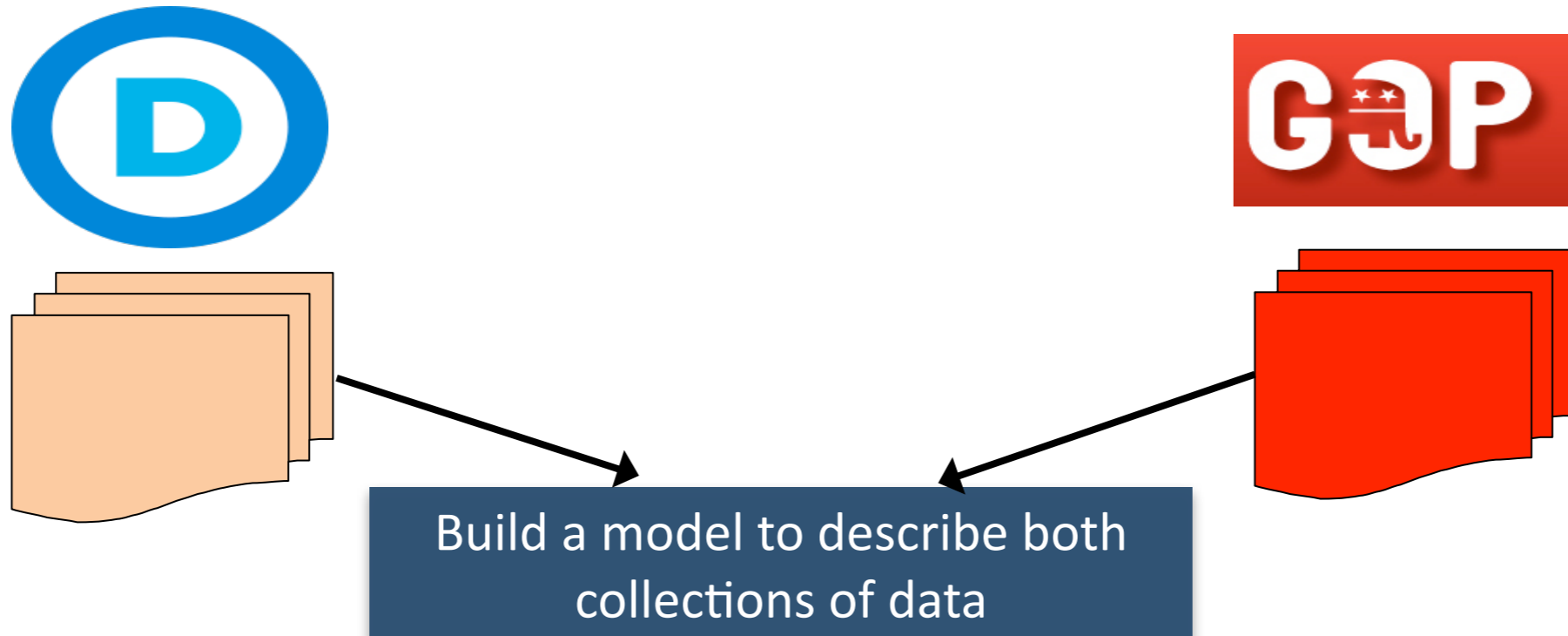
Ideologies



Visualization

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

Ideologies

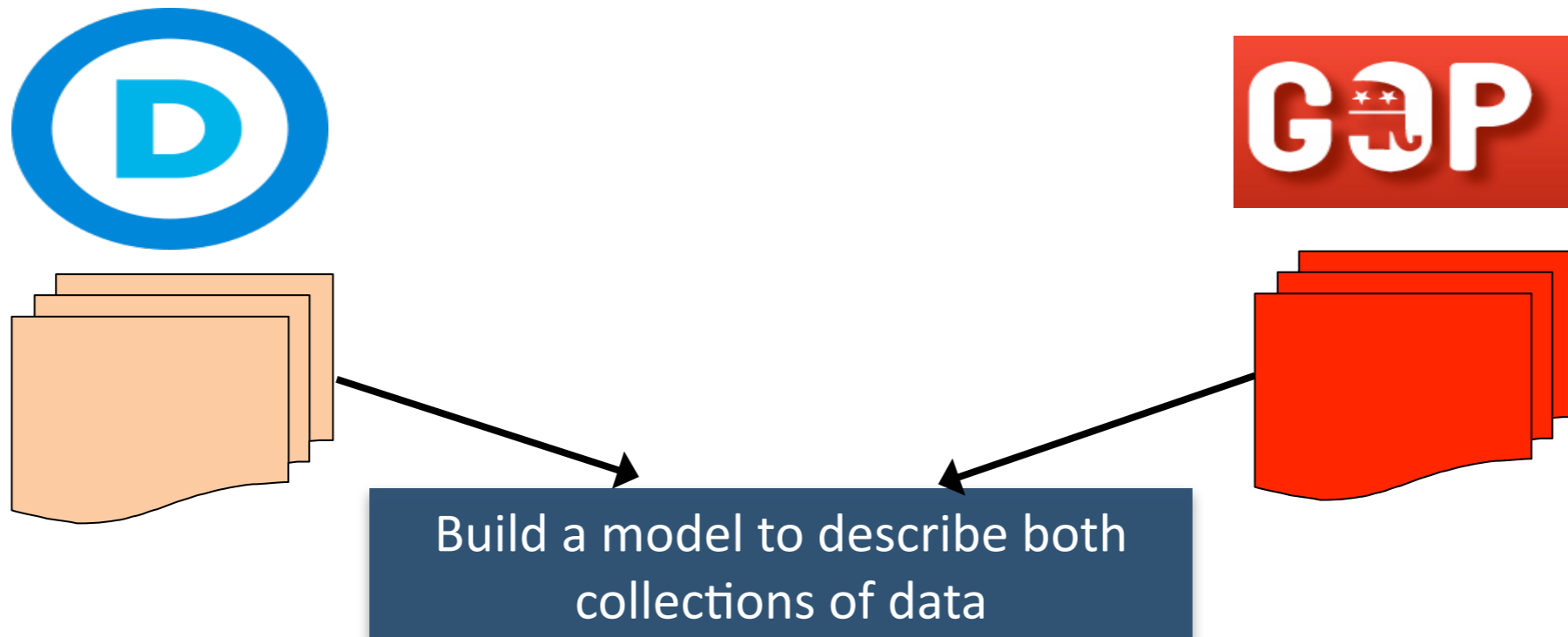


Visualization

Classification

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

Ideologies



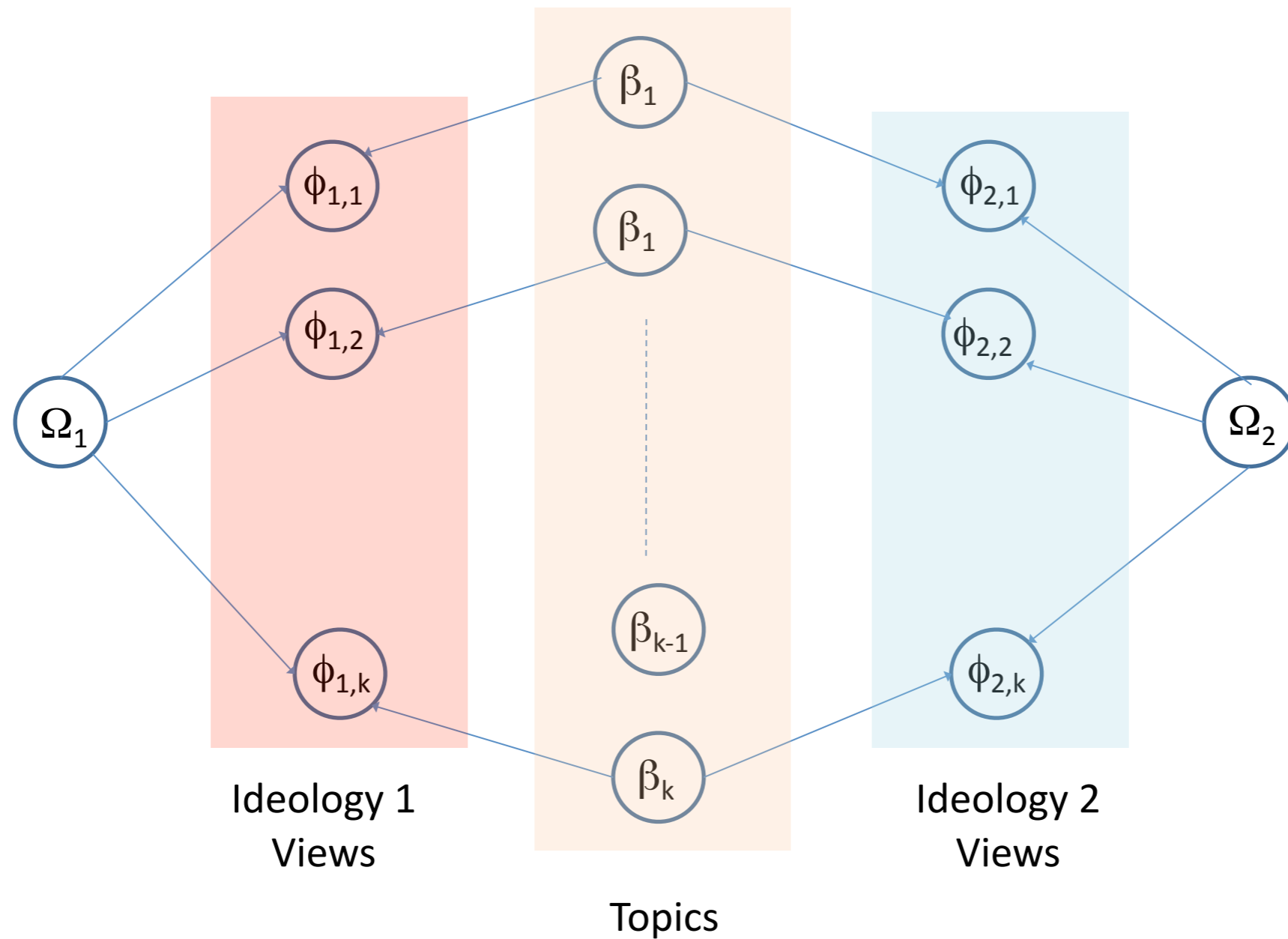
Visualization

Classification

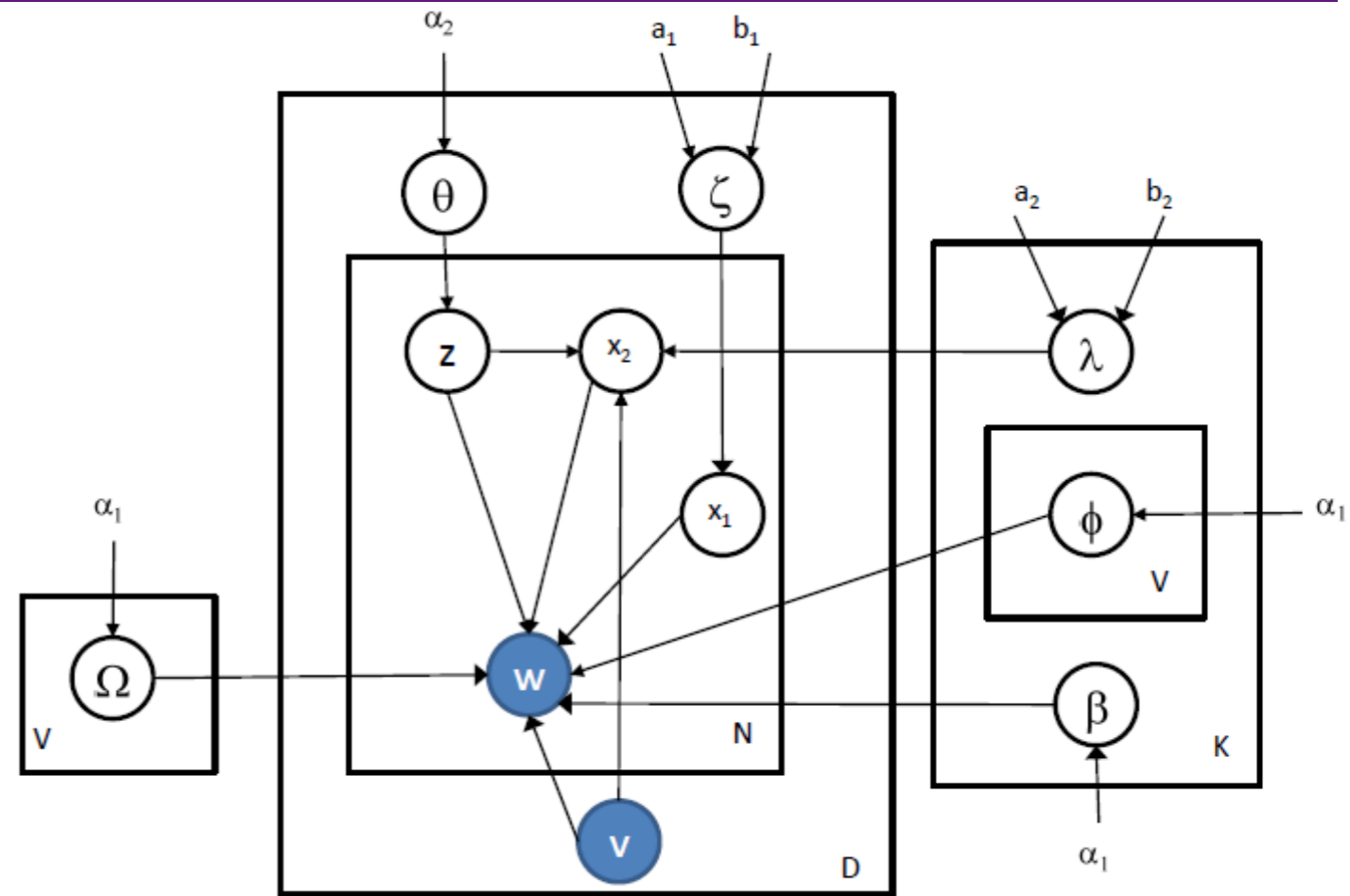
Structured browsing

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

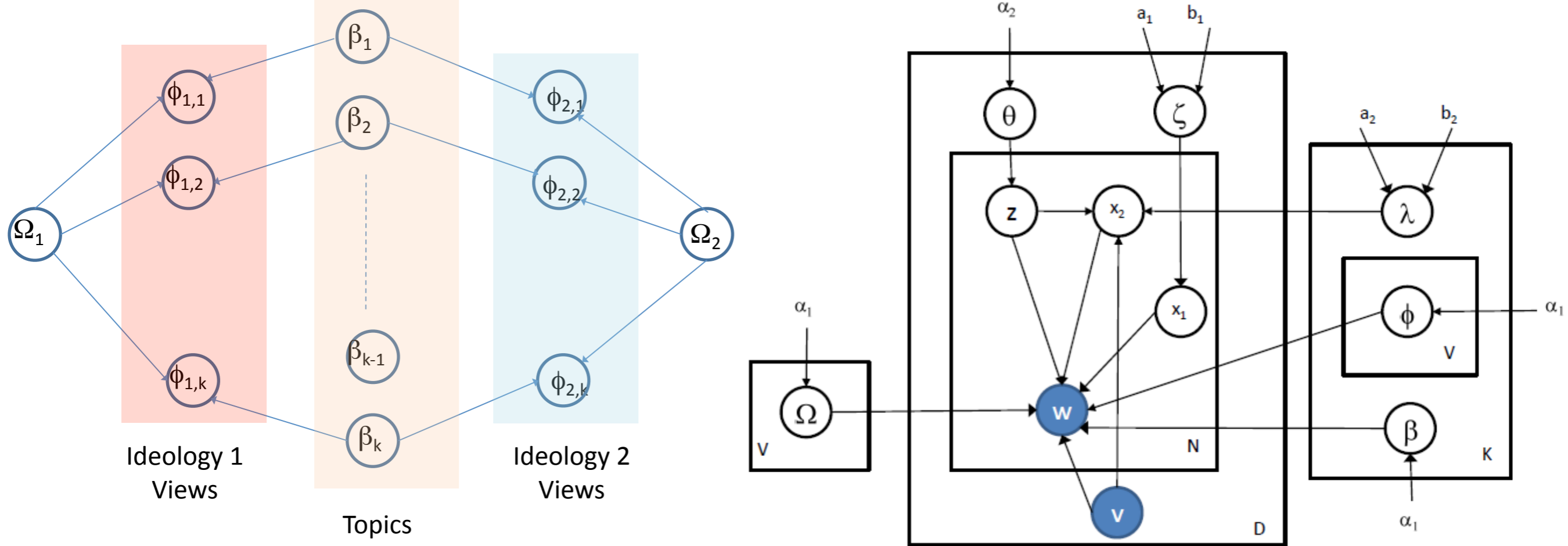
Building a factored model



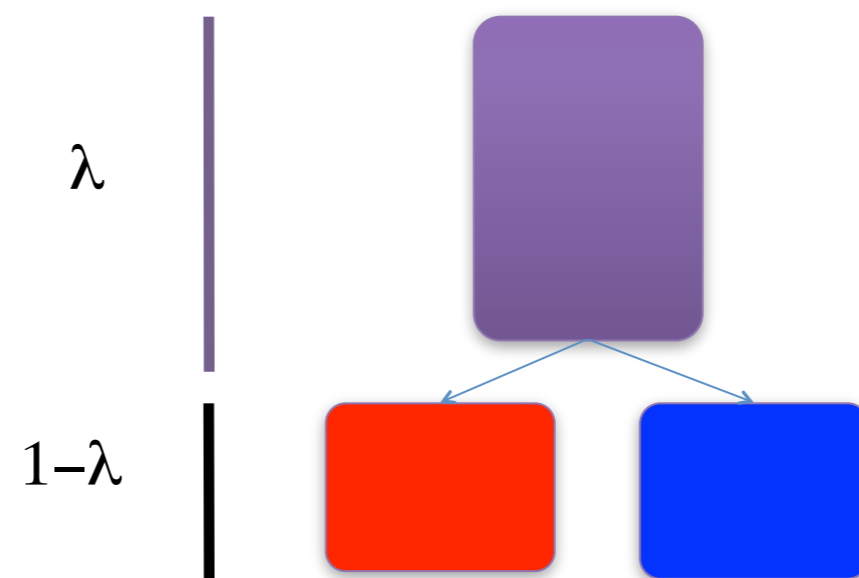
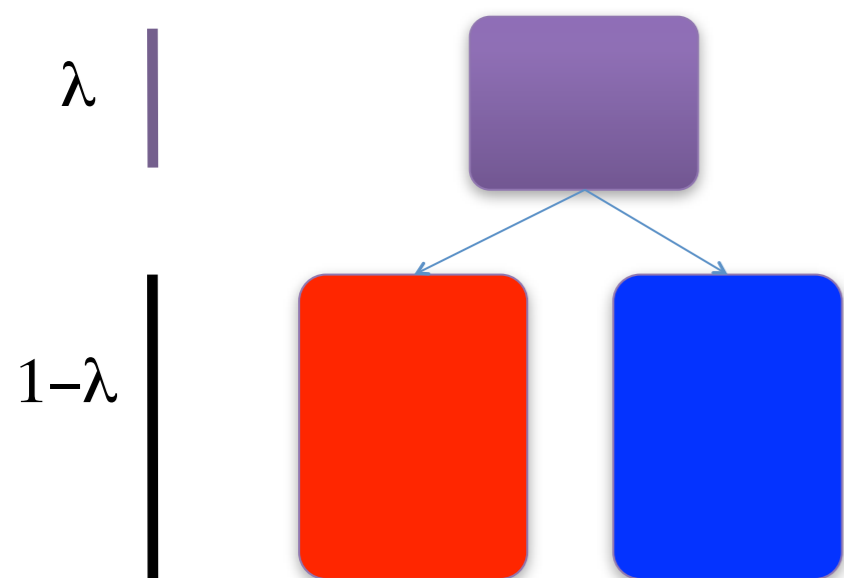
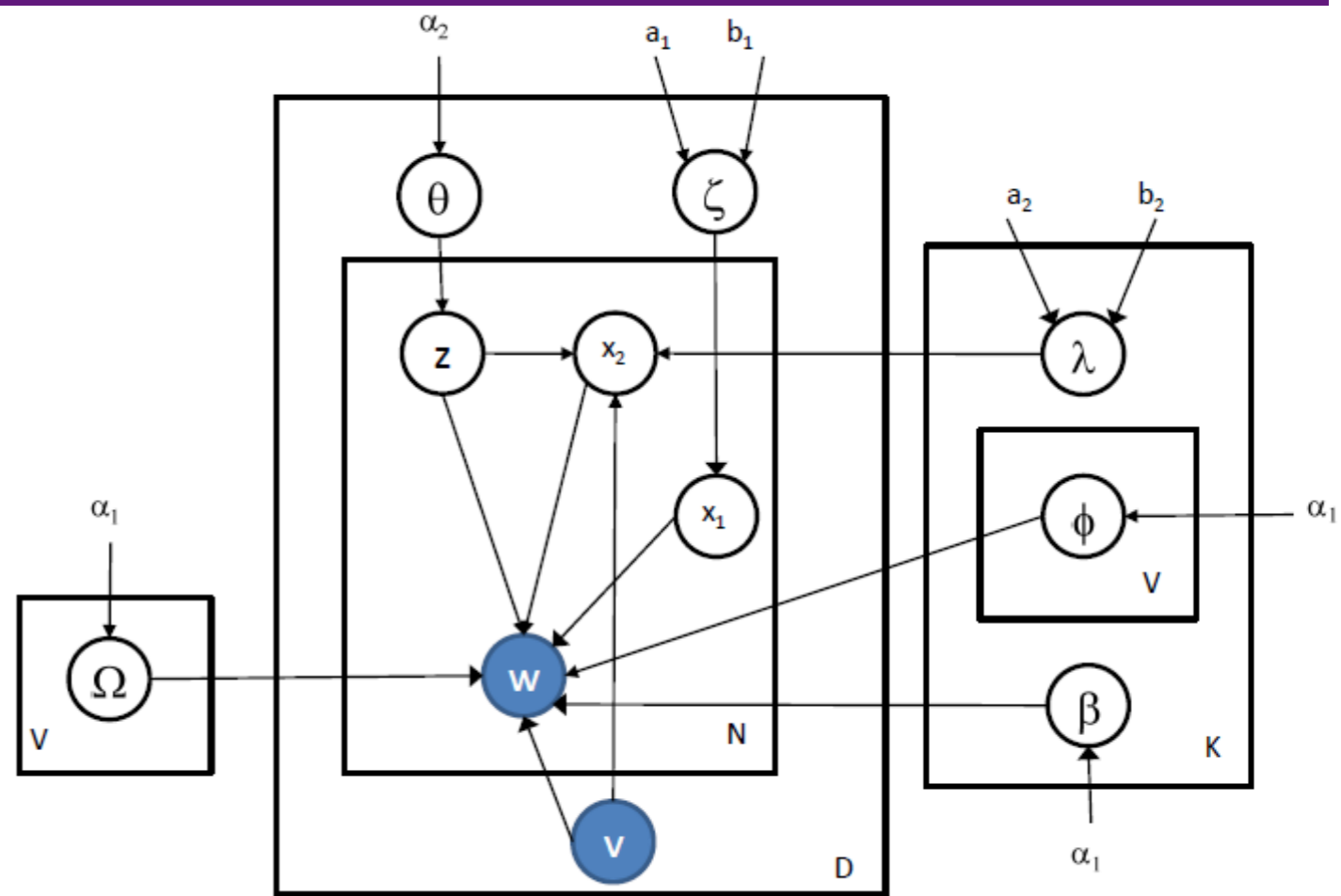
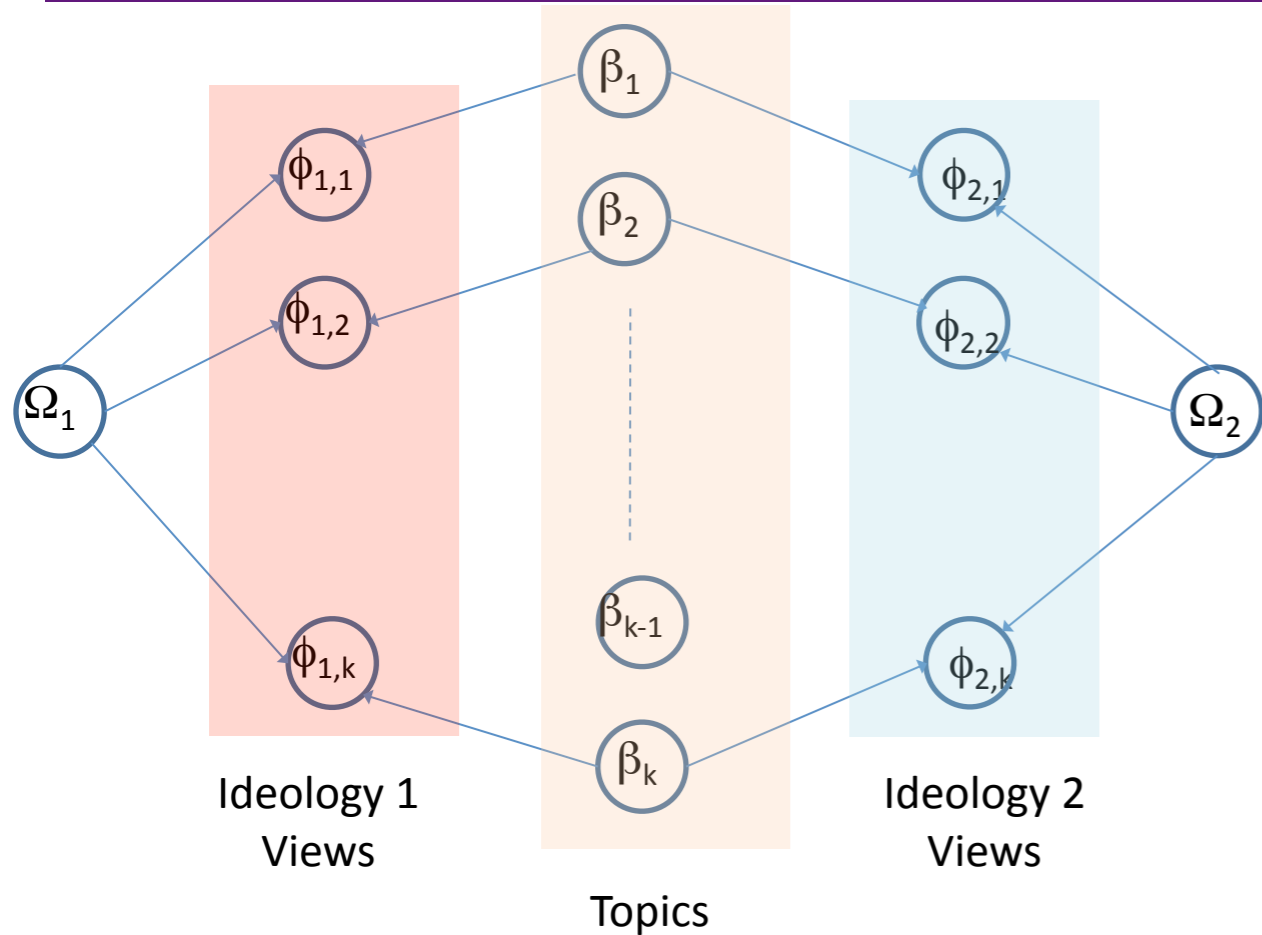
Building a factored model



Building a factored model



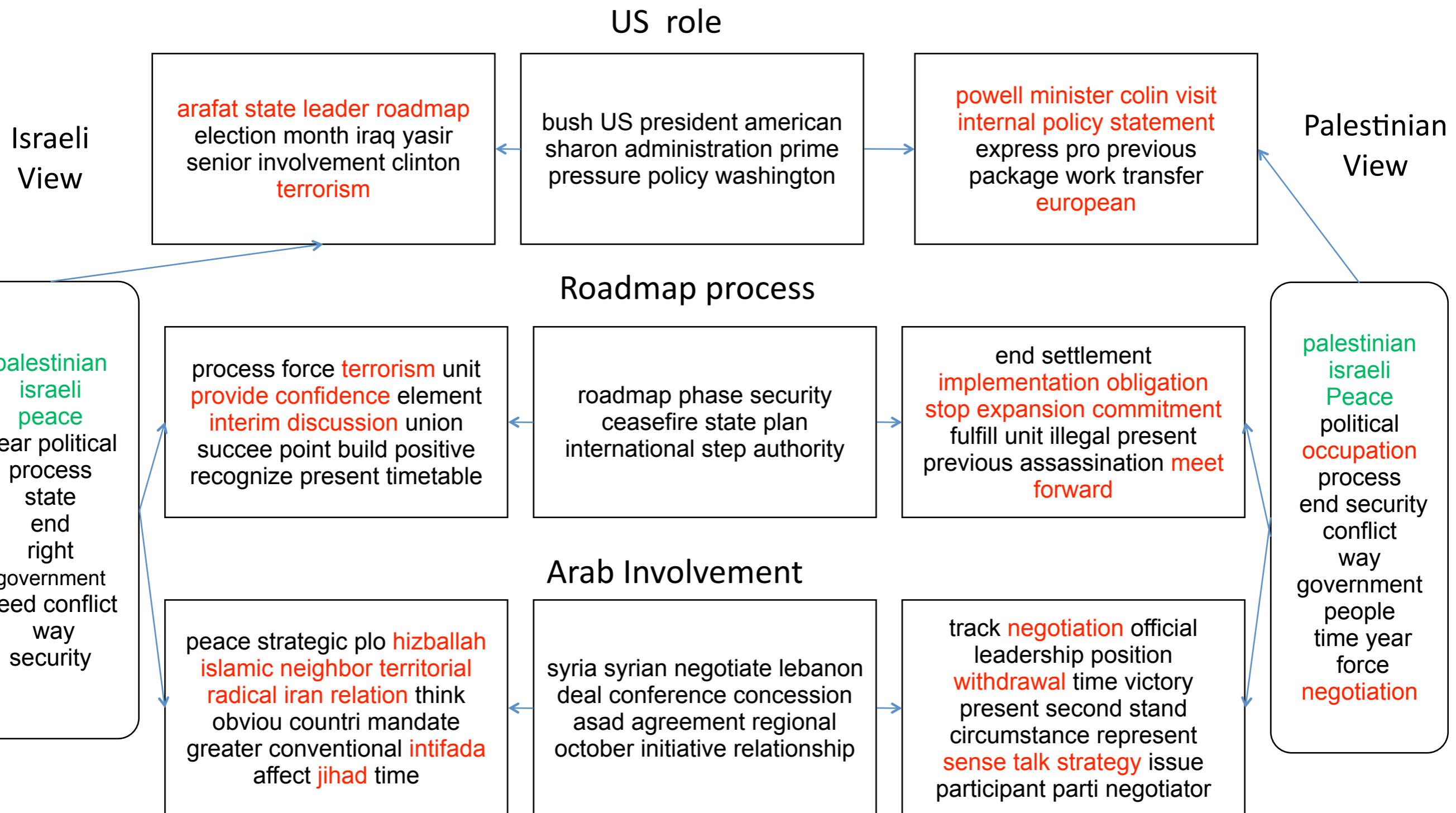
Building a factored model



Data

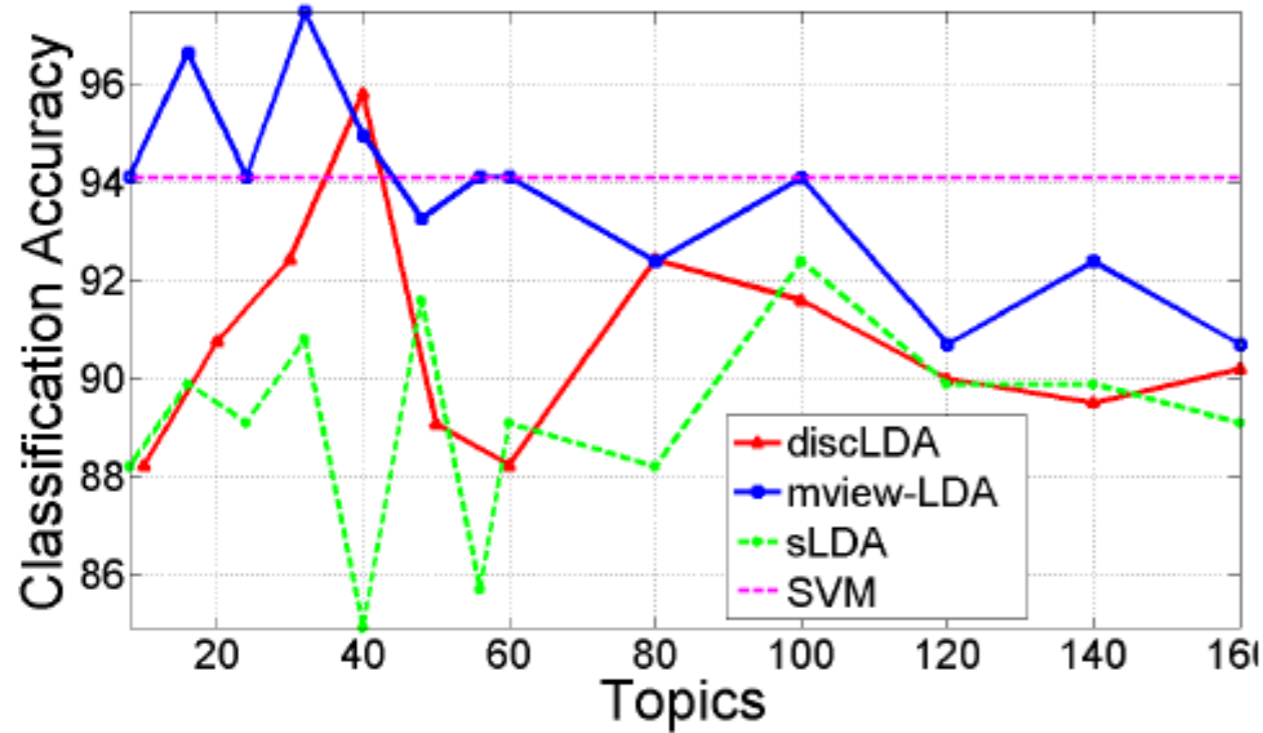
- **Bitterlemons:**
 - Middle-east conflict, document written by Israeli and Palestinian authors.
 - ~300 documents from each view with average length 740
 - Multi author collection
 - 80-20 split for test and train
- **Political Blog-1:**
 - American political blogs (Democrat and Republican)
 - 2040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

Bitterlemons dataset

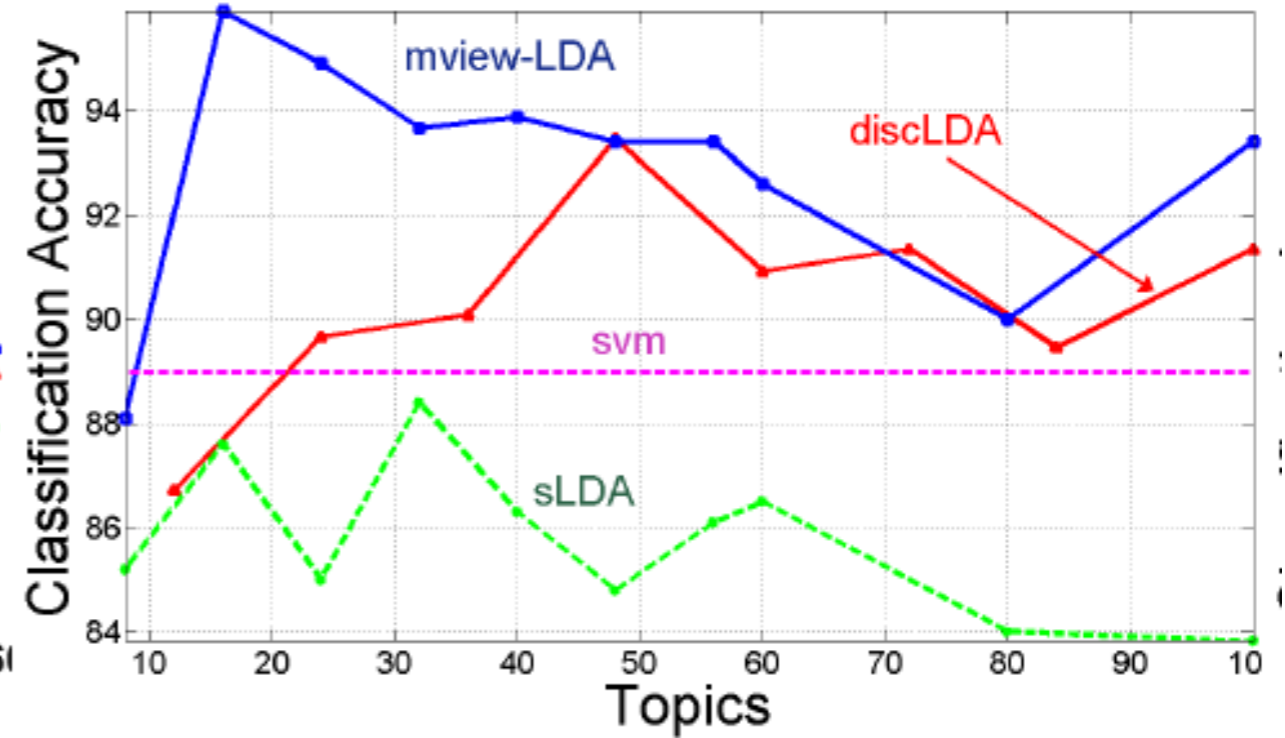


Classification accuracy

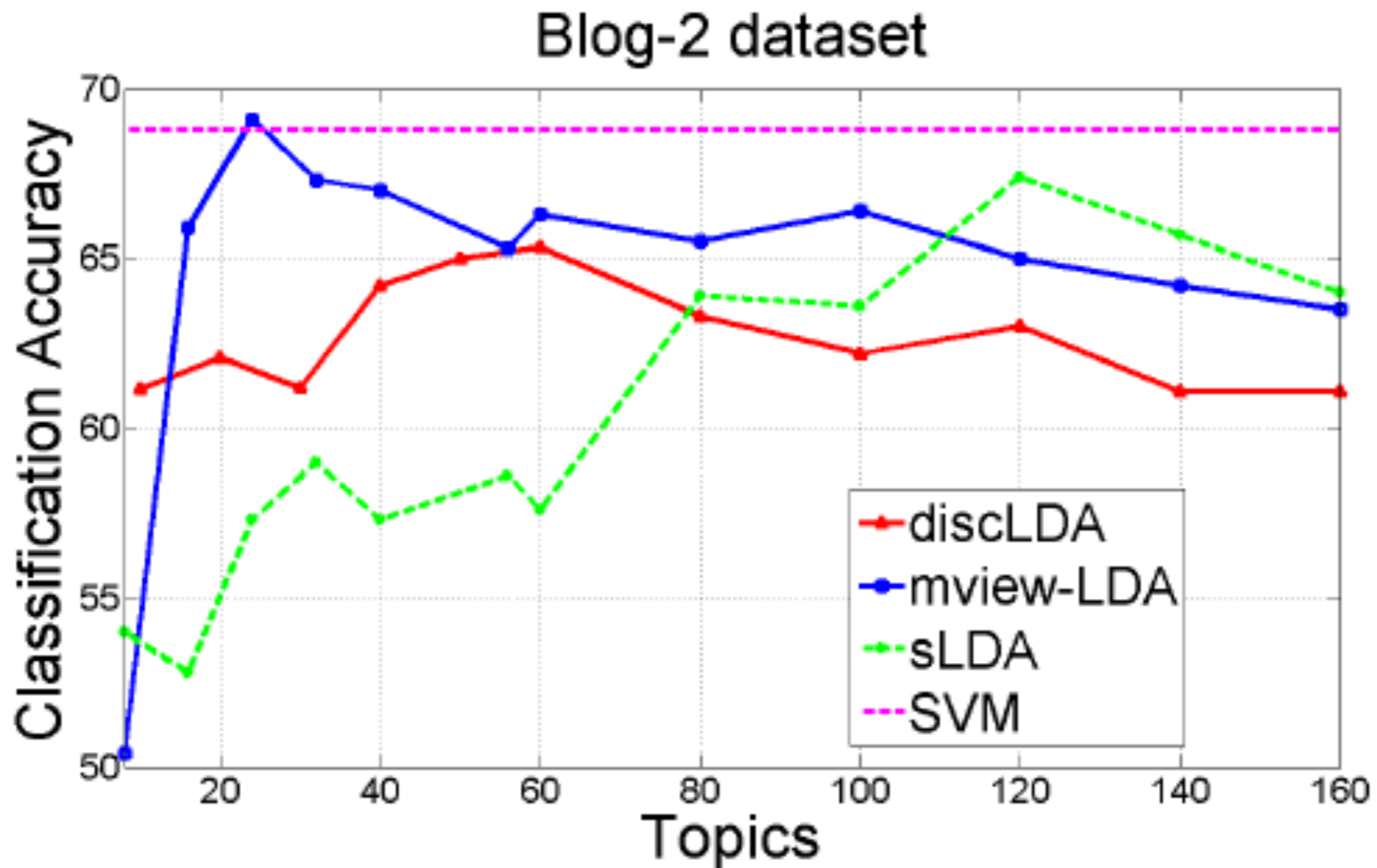
Bitterlemons dataset



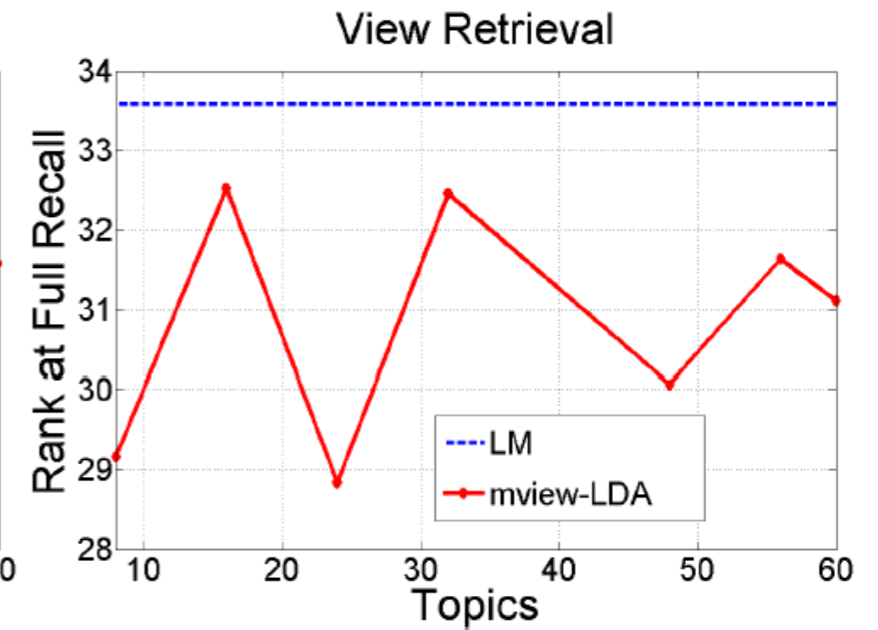
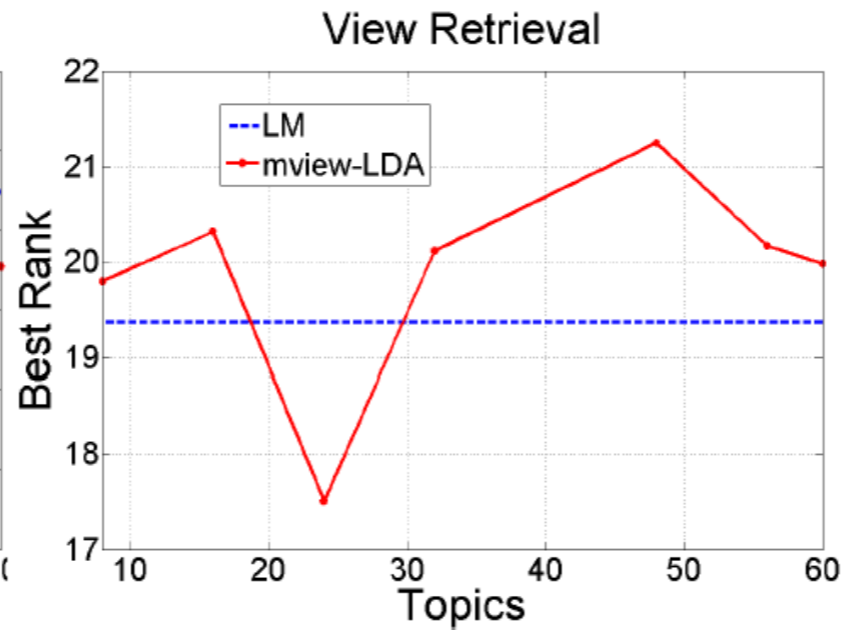
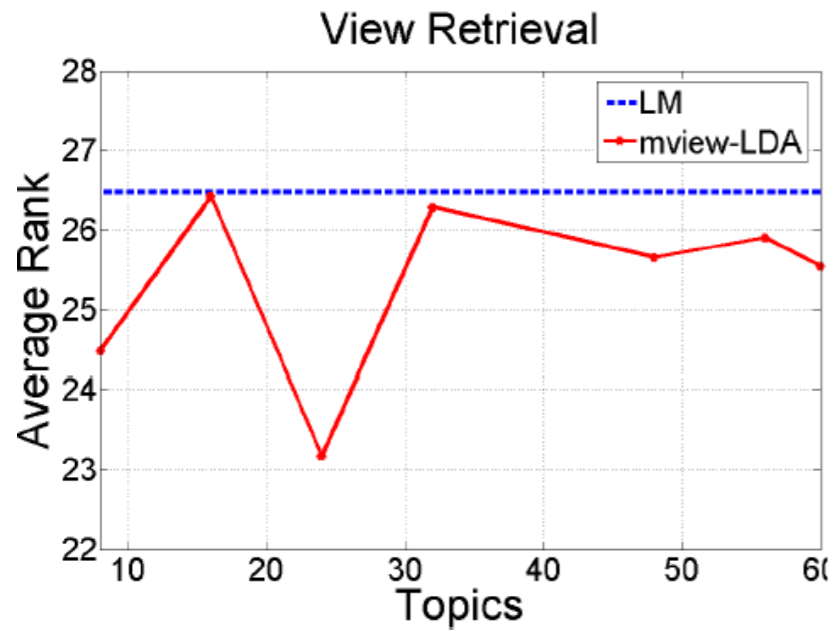
Blog-1 dataset



Generalization to new blogs

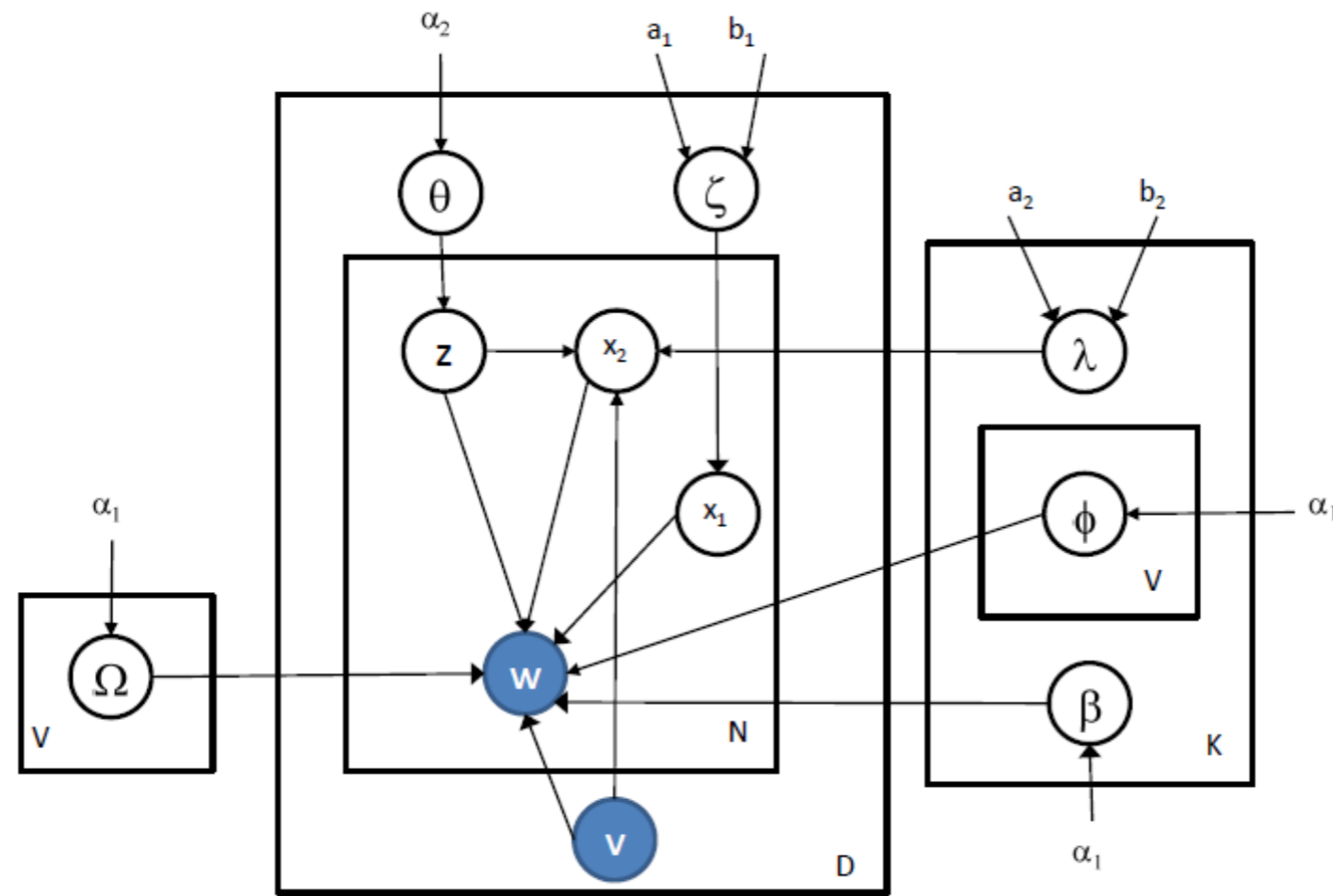


Finding alternate views

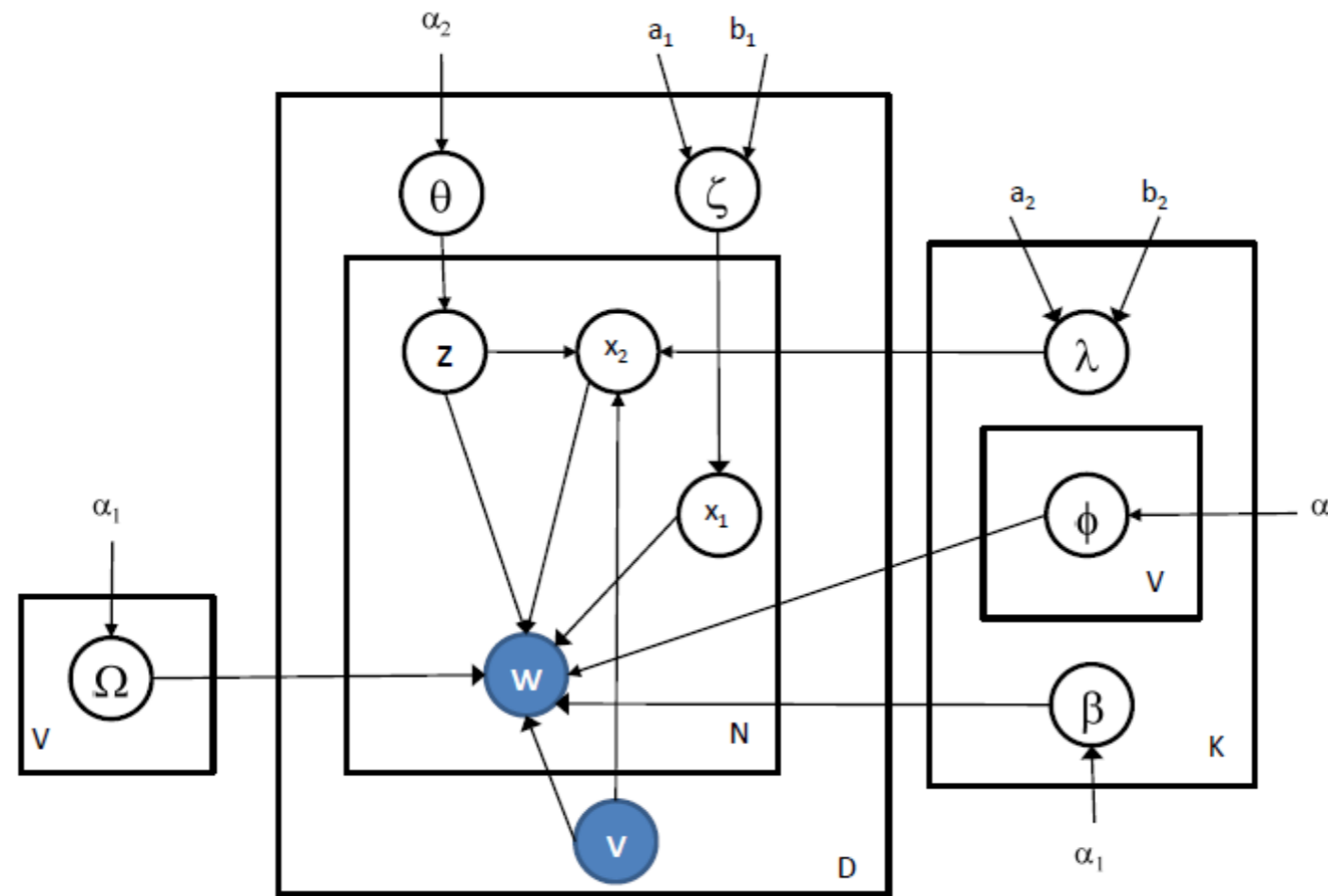


- Given a document written in one ideology, retrieve the equivalent
- Baseline: SVM + cosine similarity

Unlabeled data

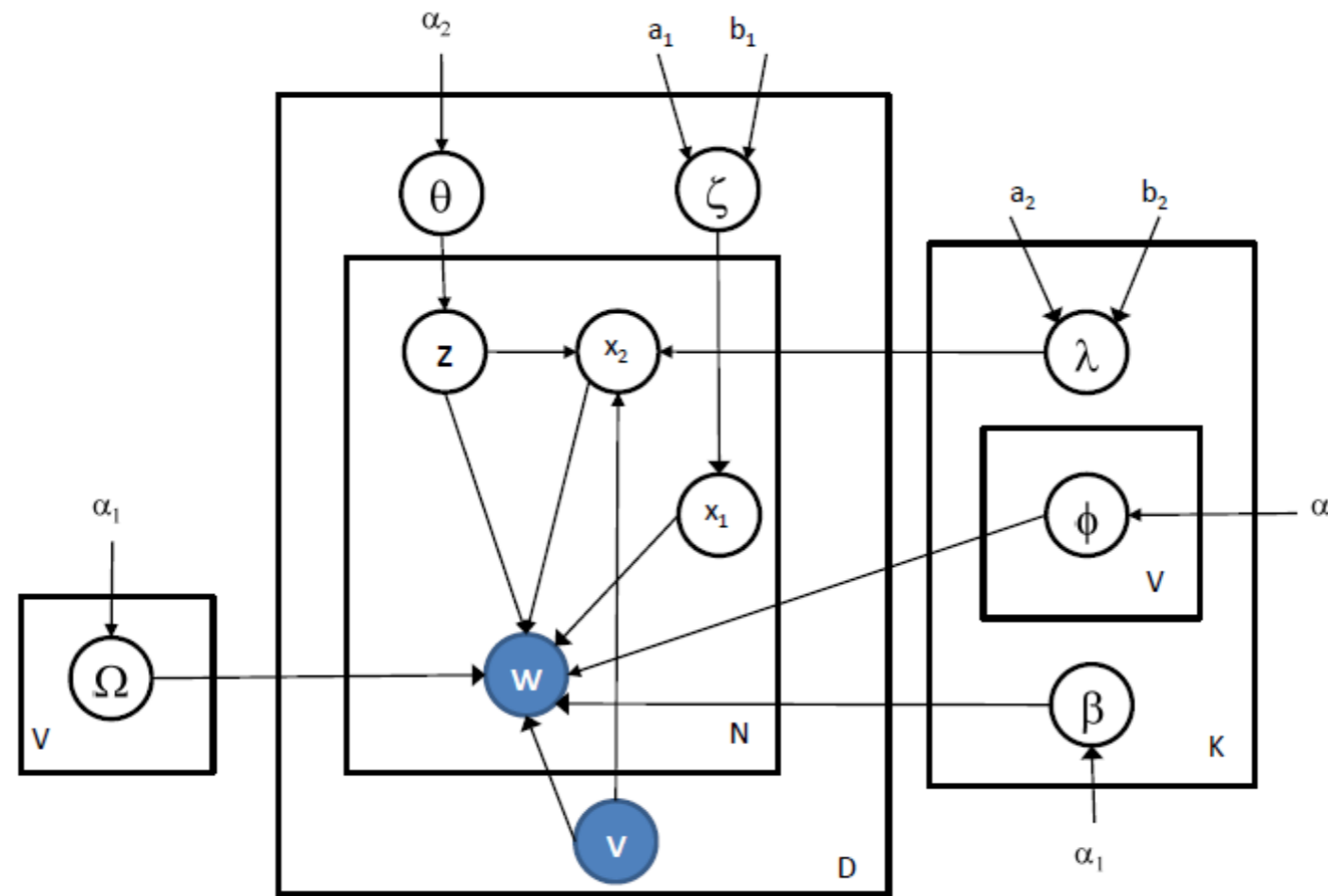


Unlabeled data



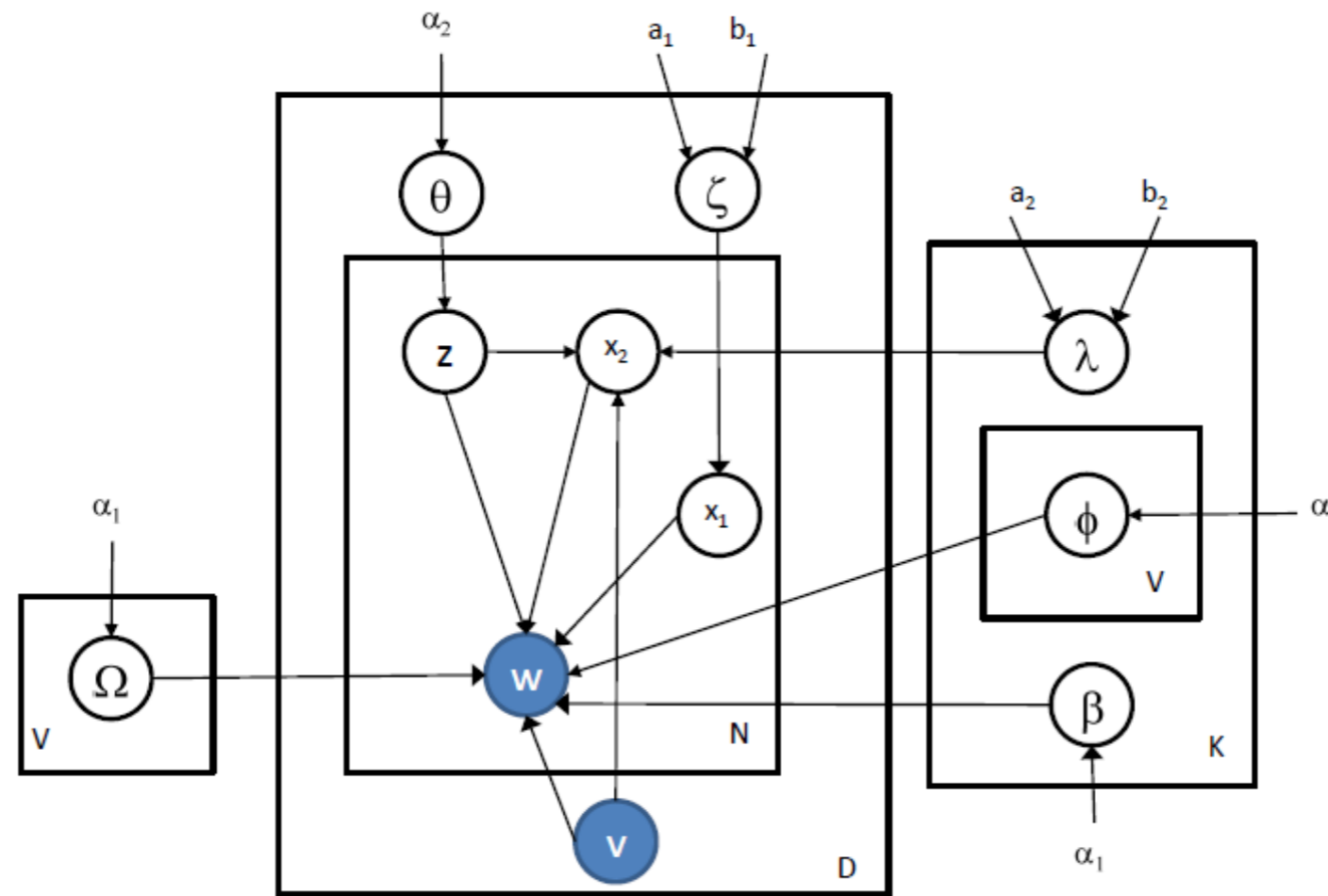
- In theory this is **simple**
 - Add a step that samples the document view (v)
 - **Doesn't mix** in practice because tight coupling between v and (x_1, x_2, z)

Unlabeled data



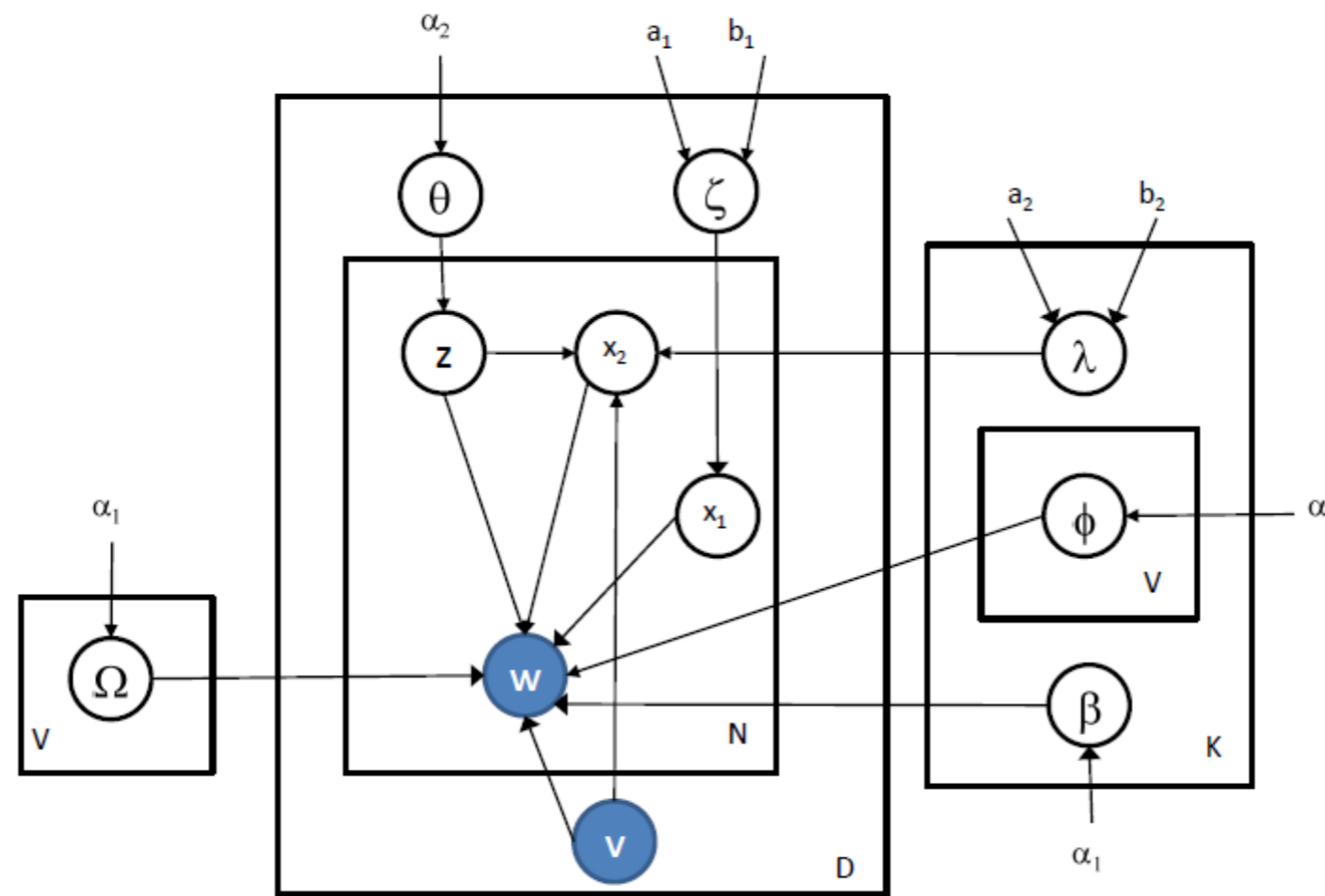
- In theory this is **simple**
 - Add a step that samples the document view (v)
 - **Doesn't mix** in practice because tight coupling between v and (x_1, x_2, z)
- Solution

Unlabeled data



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

Unlabeled data



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

Summary - Part 4

- Extensions to basic topic model (correlated topics, beyond bag of words, features)
- Chinese Restaurant Process
- Recurrent CRP
- User modeling
- Storylines
- Ideology detection

Related work

- **Tools**
 - **GraphLab (CMU - Guestrin, Low, Gonzalez ...)**
 - **Factorie (UMass - McCallum & coworkers)**
 - **HBC (Hal Daume)**
 - **Variational Bayes .NET (MSR Cambridge)**
- **See more on alex.smola.org / blog.smola.org**