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, Yah
- 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)

unlimited amounts of data

Finite resources

- Editors are expensive
- know users

ns are novel)

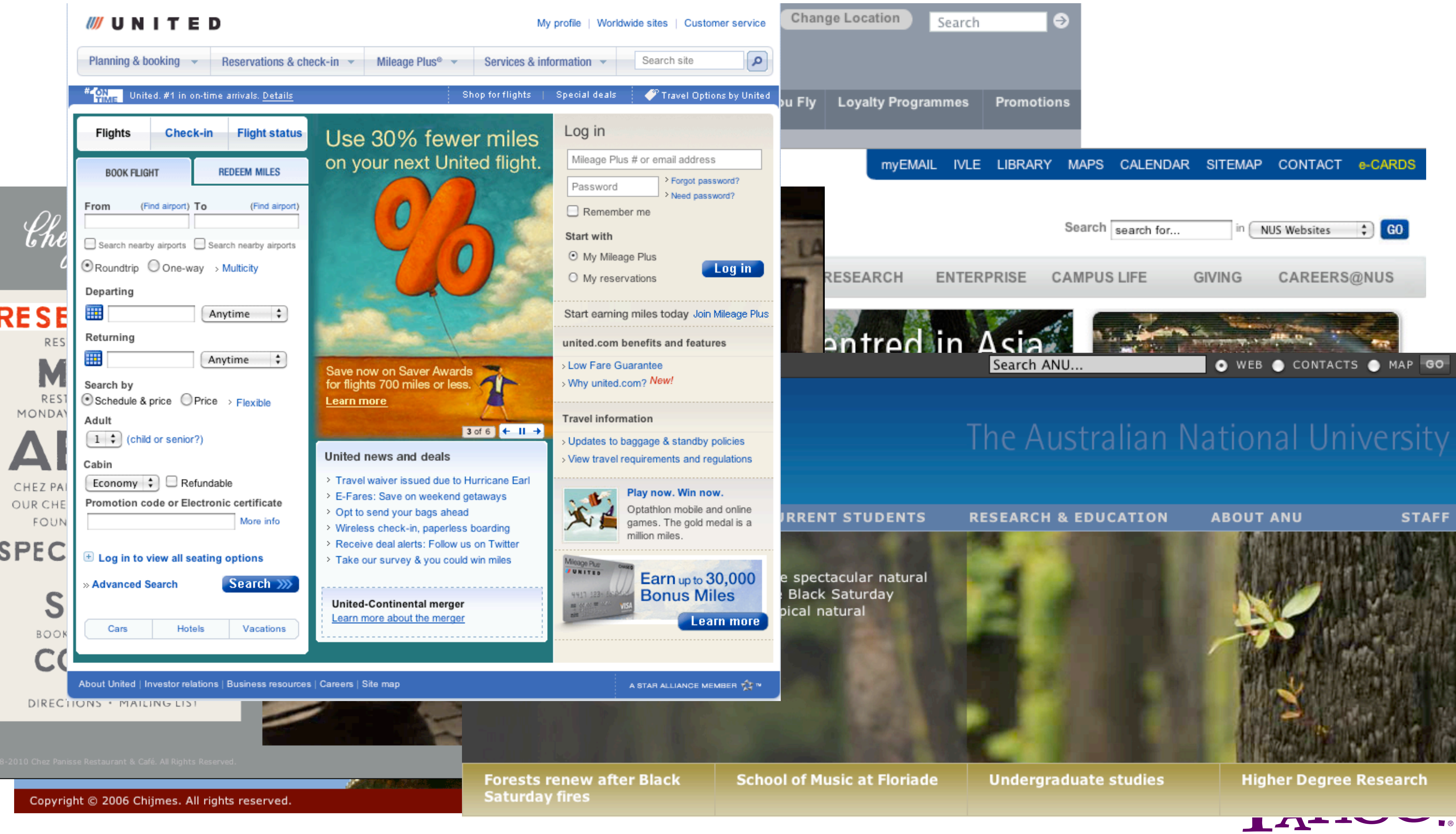
ck

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

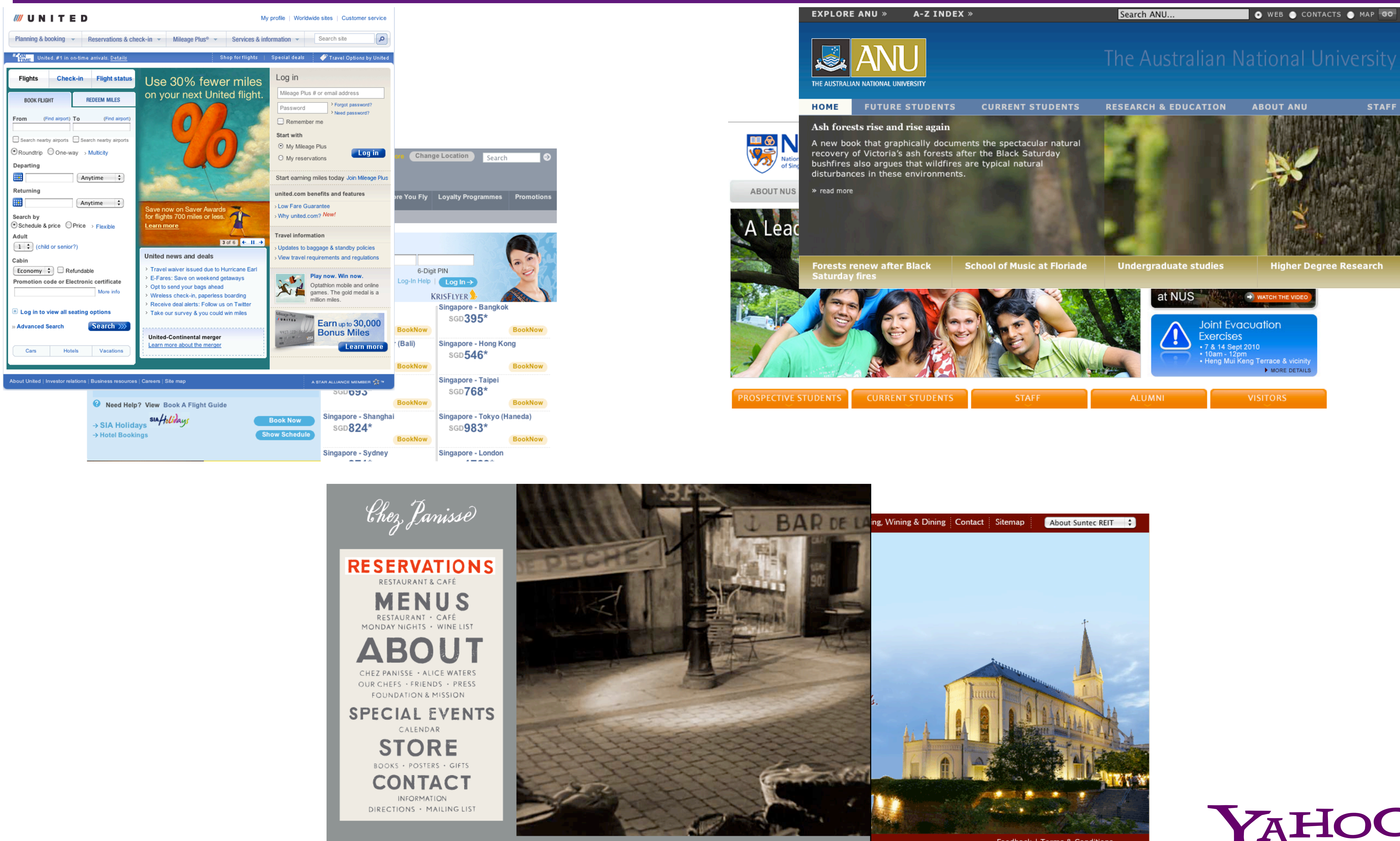
Invest in modeling and naming rather than data generation

Clustering documents

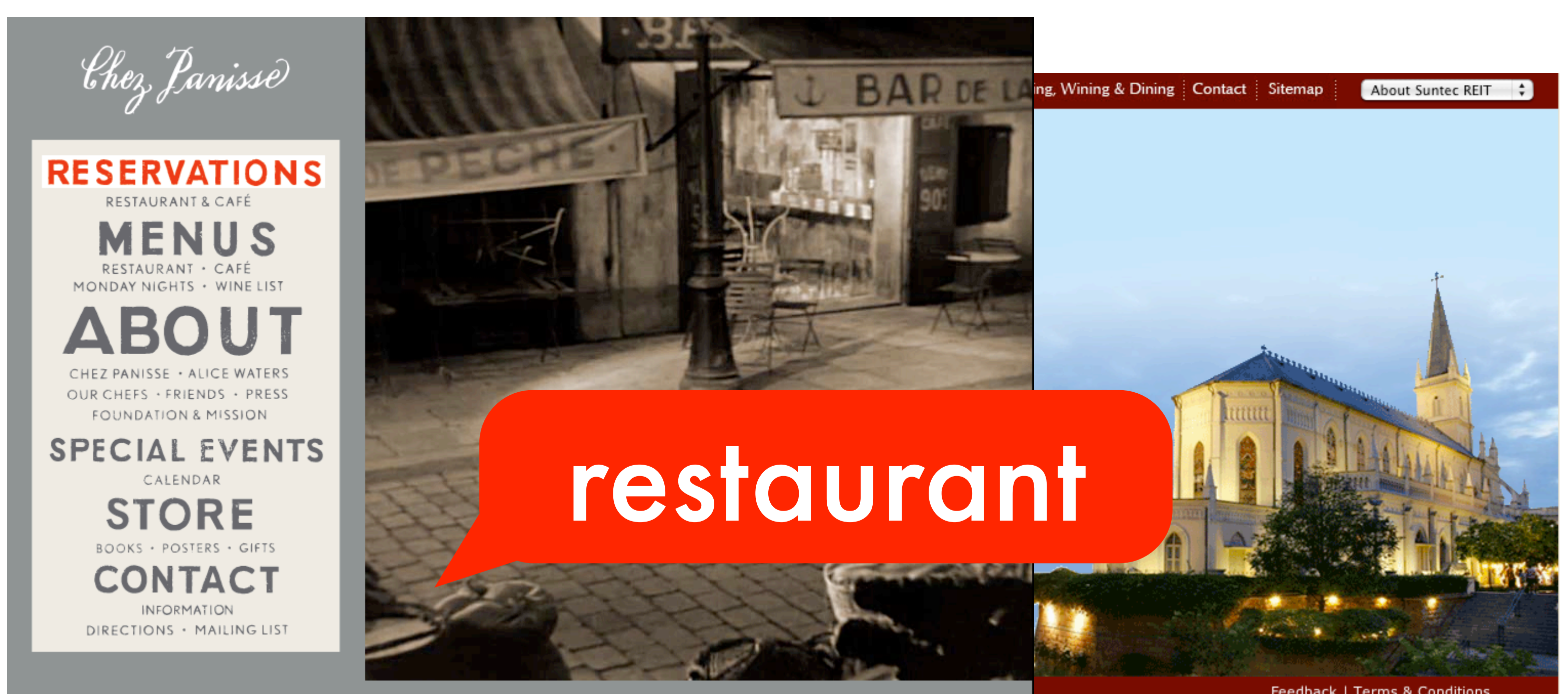
Clustering documents



Clustering documents



Clustering documents



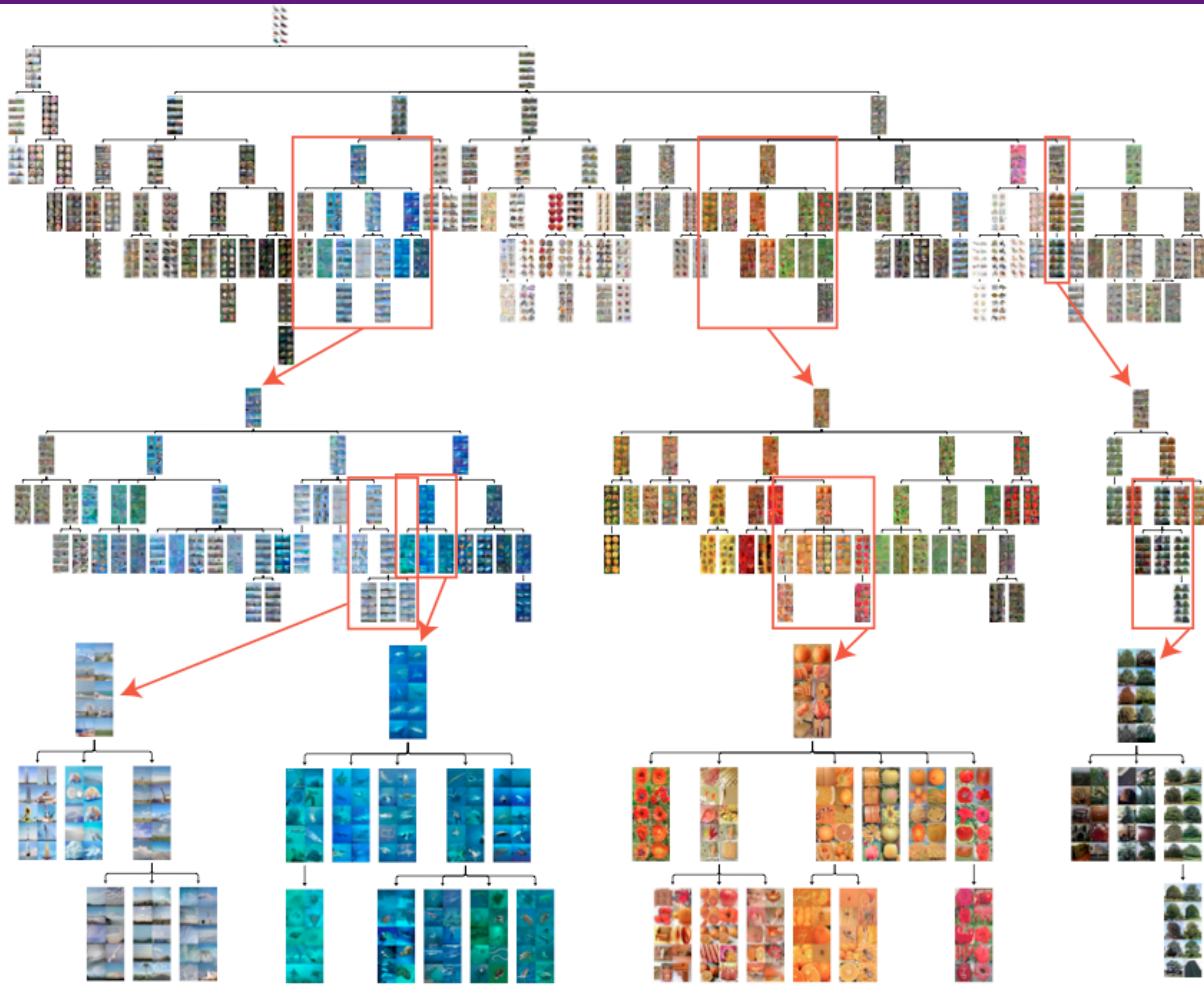
Today's mission

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

firstshedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspedupo
nherknee,andthebrighteagereyeswerelookingupintohersshecouldheartheveryto
nesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderinghai
rthatwouldalwaysgetintohereyesandstillasshelistened,orseemedtolisten,thewho
leplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.thelo
Nggrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashed
Hiswaythroughtheneighbouringpoolshecouldheartherattleoftheteacupsasthemar
chhareandhisfriendssharedtheirneverendingmeal,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
.

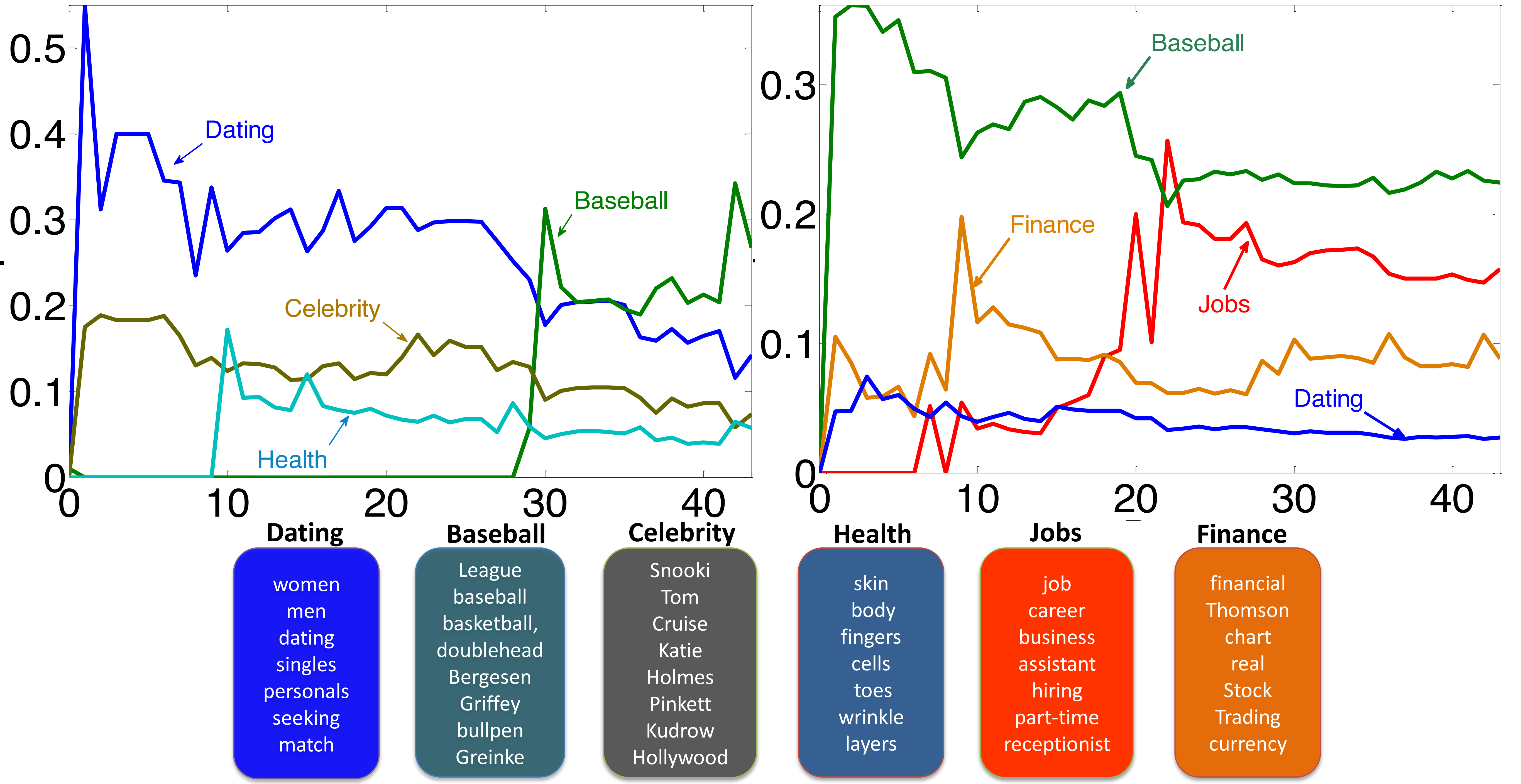
---

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

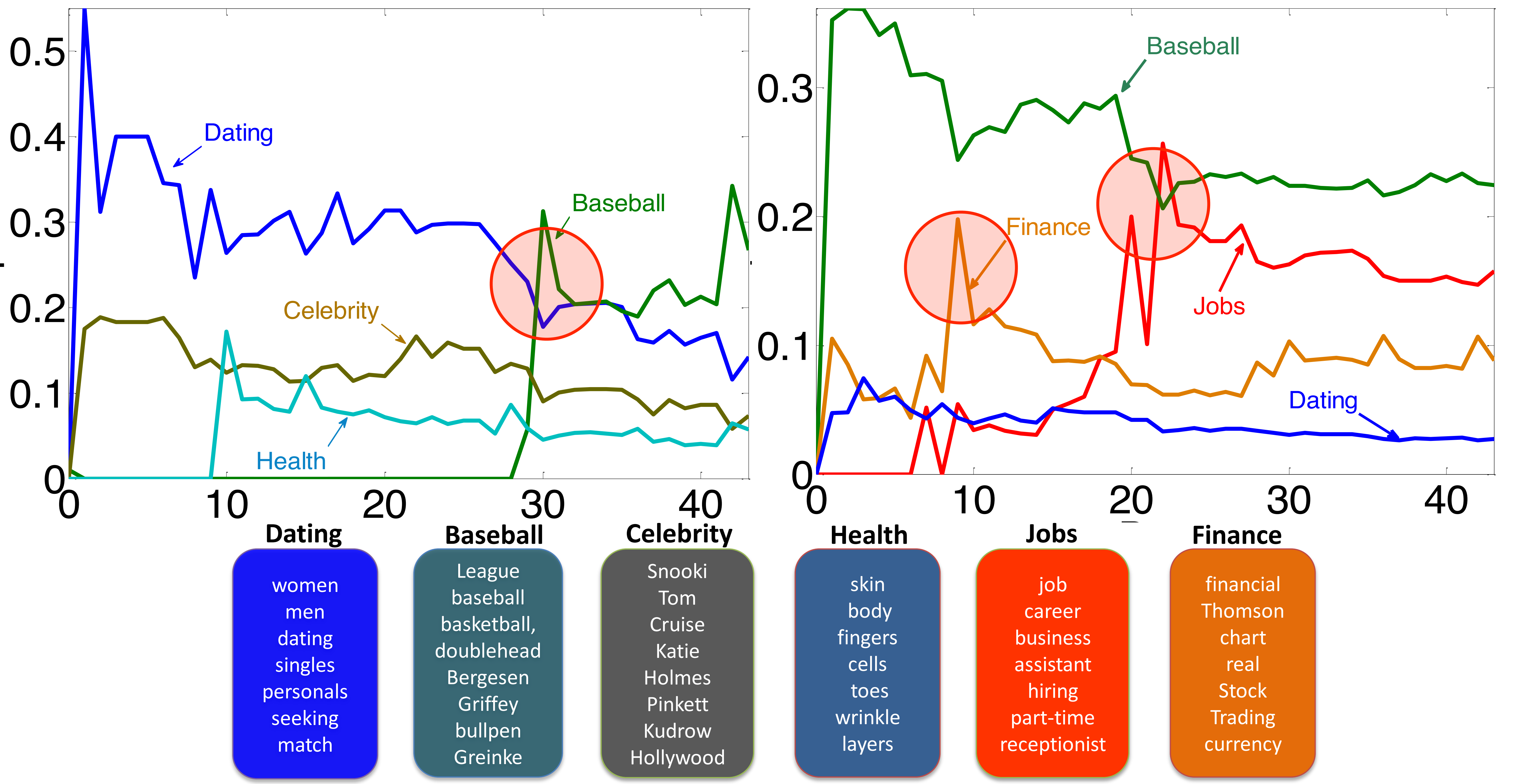
Mochihashi, Yamada, Ueda
ACL 2009

User model over time



Ahmed et al., KDD 2011

User model over time



Ahmed et al., KDD 2011

Face recognition from captions

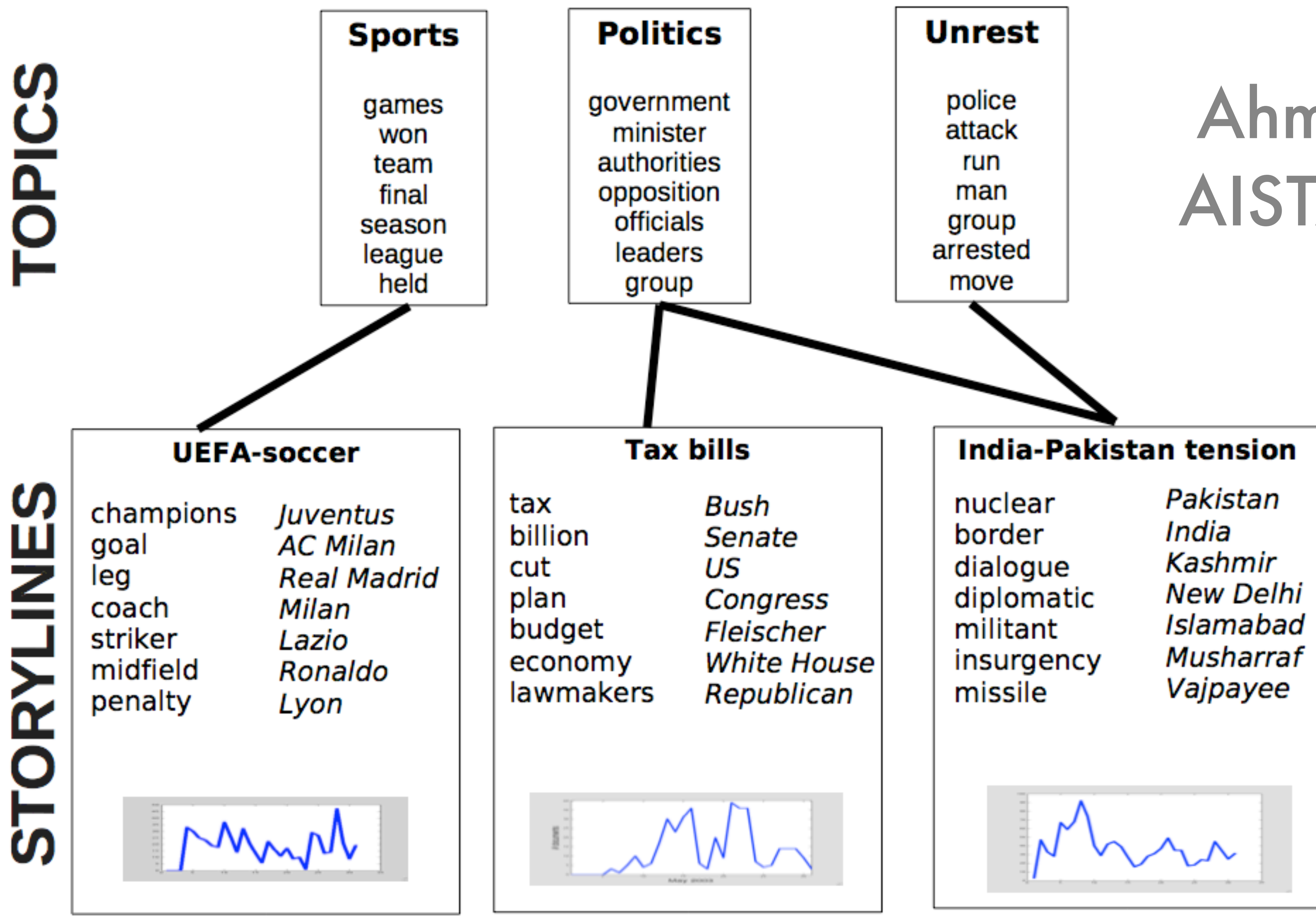
(a) Random samples from four clusters obtained using LDA on caption text [6].

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

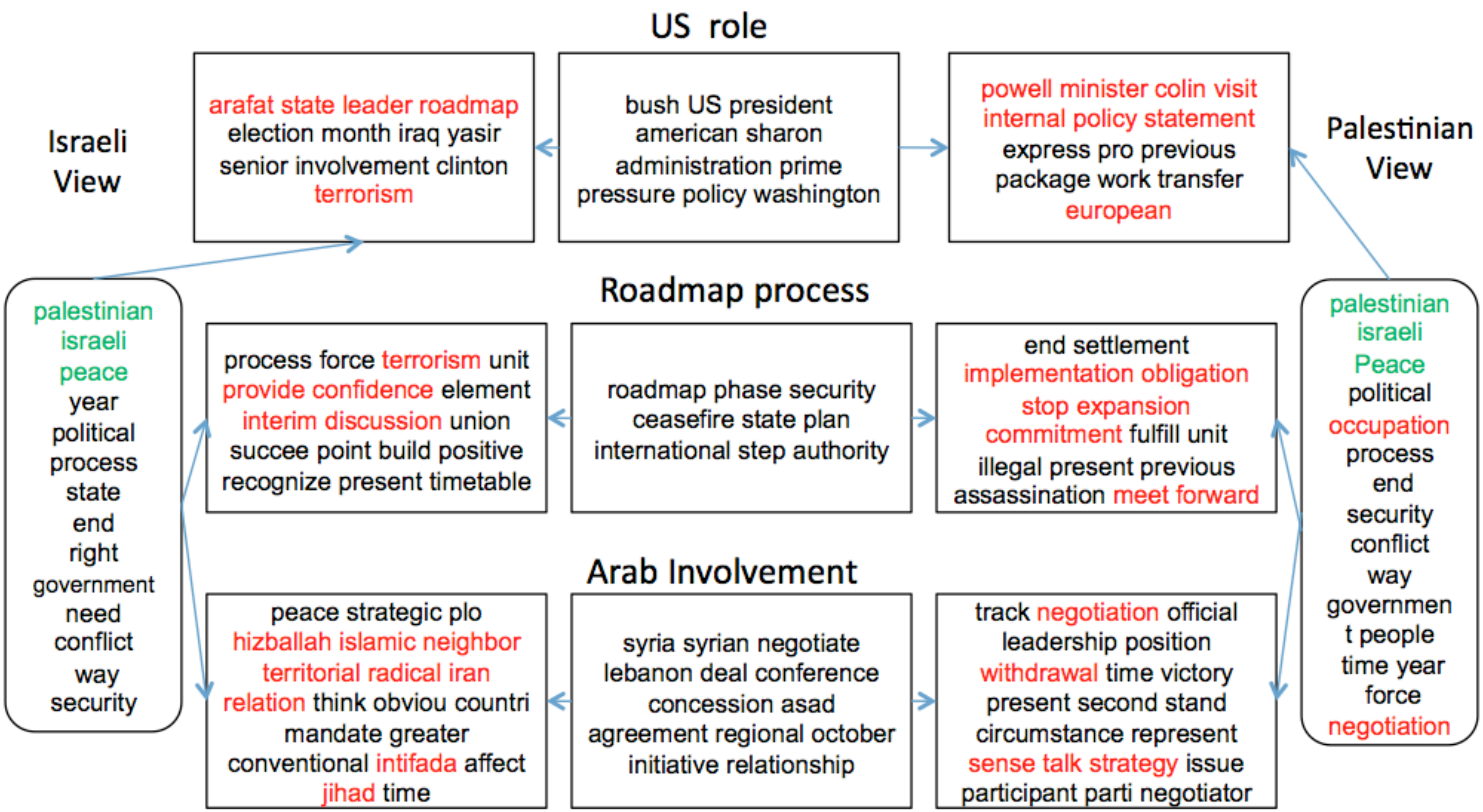
Jain, Learned-Miller, McCallum, ICCV 2007

Storylines from news

Ahmed et al, AISTATS 2011



Ideology detection



Ahmed et al, 2010; Bitterlemons collection

Hypertext topic extraction

Topic 1

neural	0.067
network	0.047
networks	0.039
learning	0.027
artificial	0.017
data	0.015
models	0.014
function	0.014

Artificial neural network 0.004

Neural network 0.003

Topic 2

recognition	0.058
speech	0.033
language	0.015
pattern	0.012
handwriting	0.011
evaluation	0.010
robots	0.010
systems	0.009

Speech recognition 0.004

Pattern recognition 0.004

Topic 3

vancouver	0.051
denver	0.043
city	0.041
retrieved	0.024
colorado	0.011
area	0.009
population	0.009
canada	0.008

Denver, Colorado 0.0008

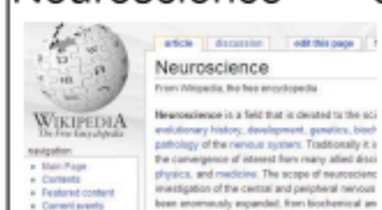
Vancouver 0.0002

Topic 4

brain	0.047
cognitive	0.026
science	0.016
press	0.011
neurons	0.010
mind	0.010
systems	0.010
human	0.010

Cognitive science 0.003

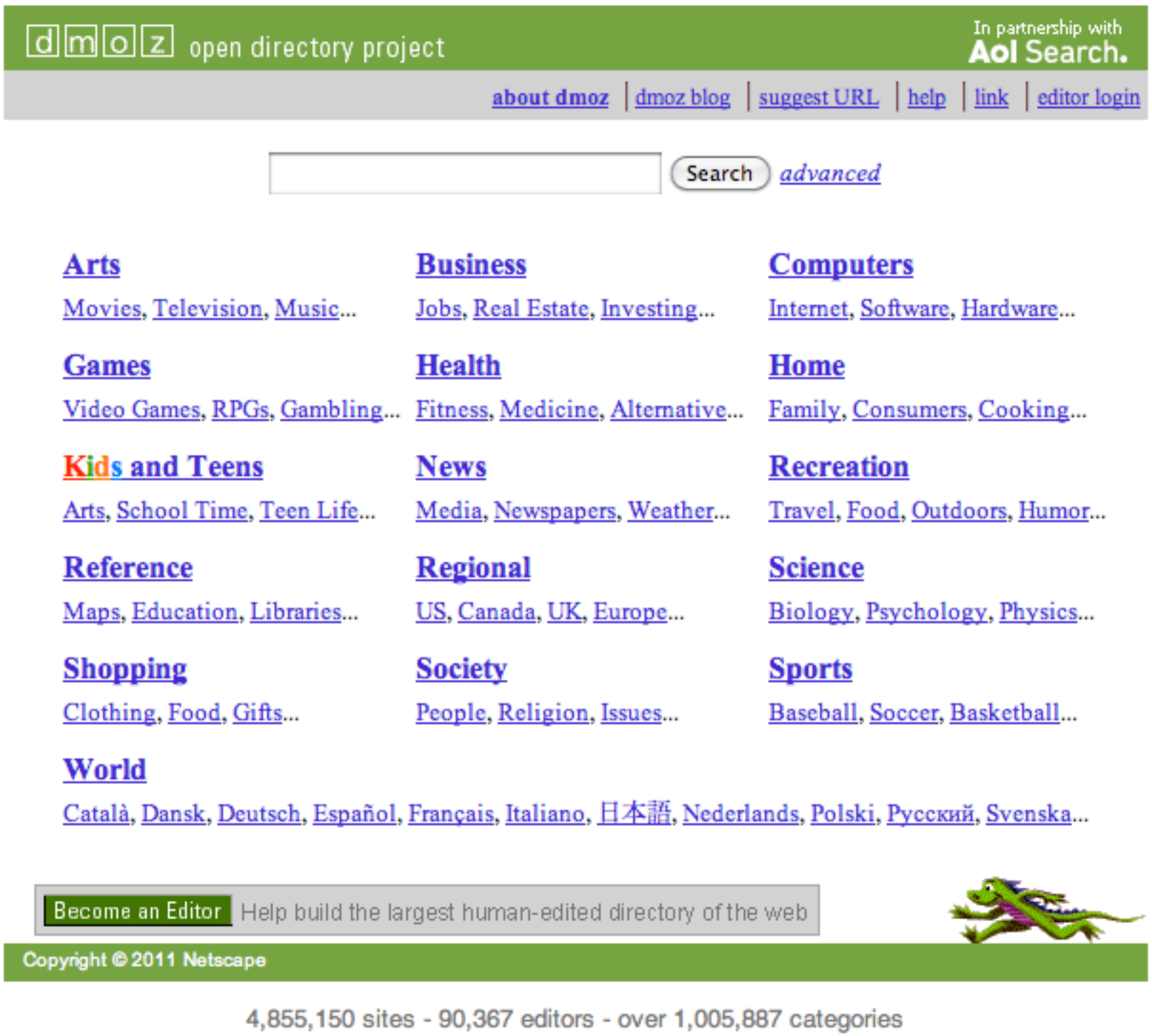
Neuroscience 0.002



Gruber, Rosen-Zvi, Weiss; UAI 2008

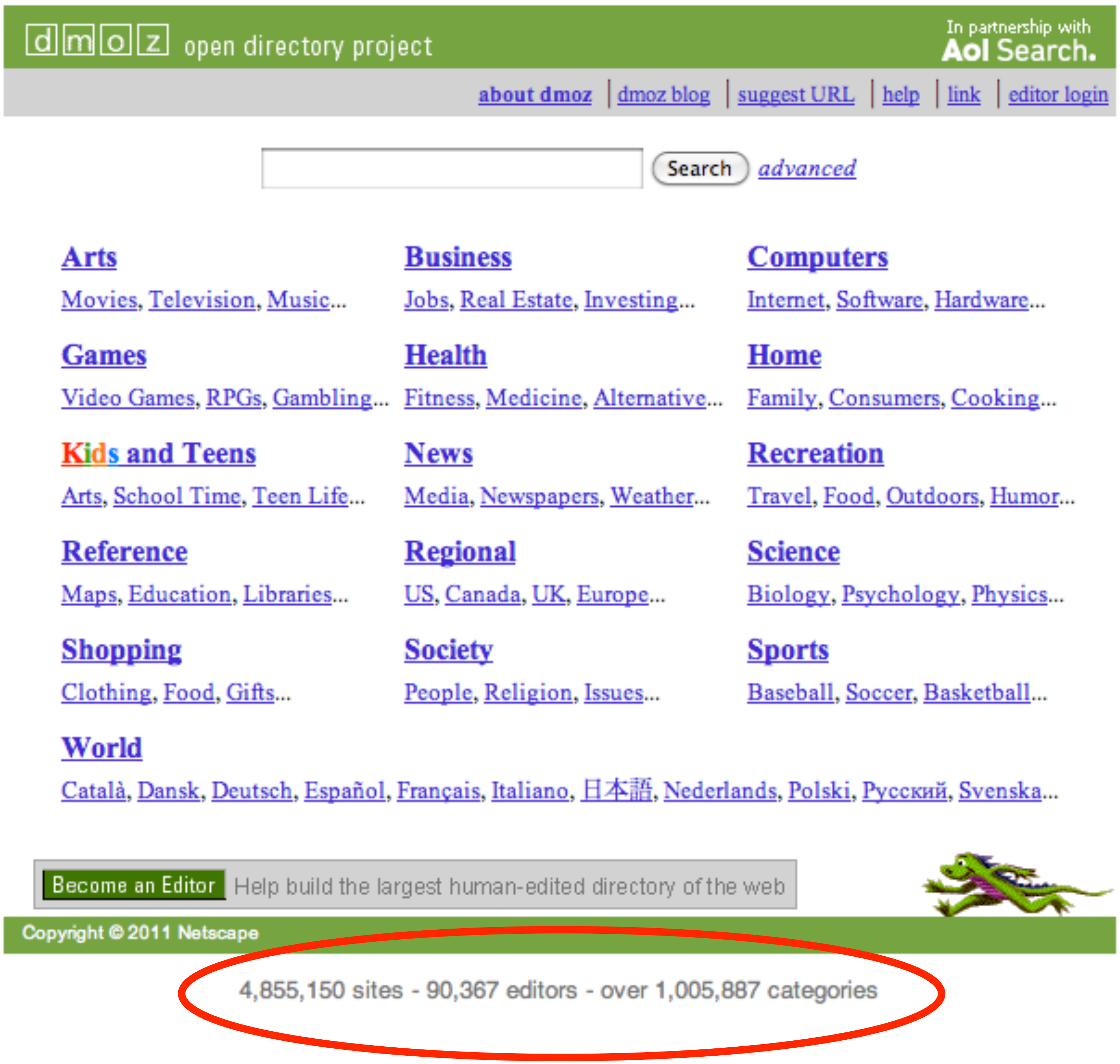
Alternatives

Ontologies



- continuous maintenance
- no guarantee of coverage
- difficult categories

Ontologies



- continuous maintenance
- no guarantee of coverage
- difficult categories

Face Classification

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

Topic Detection & Tracking

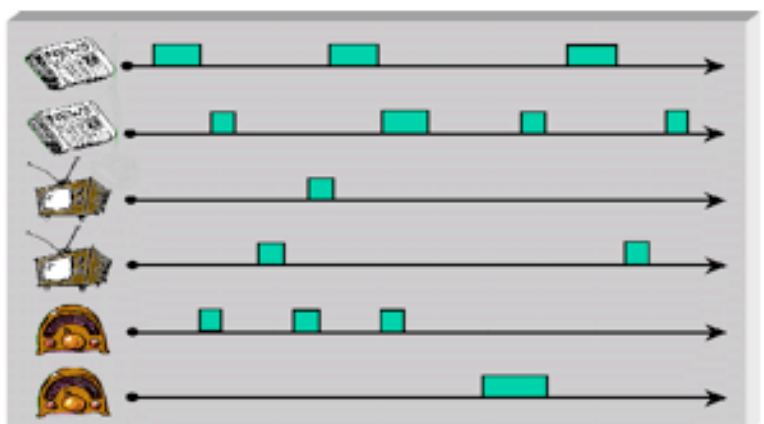
Information Technology Laboratory

Information Access Division (IAD)

NIST
National Institute of Standards and Technology

- Multimodal Information Group Home
- Benchmark Tests
- Tools
- Test Beds
- Publications
- Links
- Contacts

Topic Detection and Tracking Evaluation



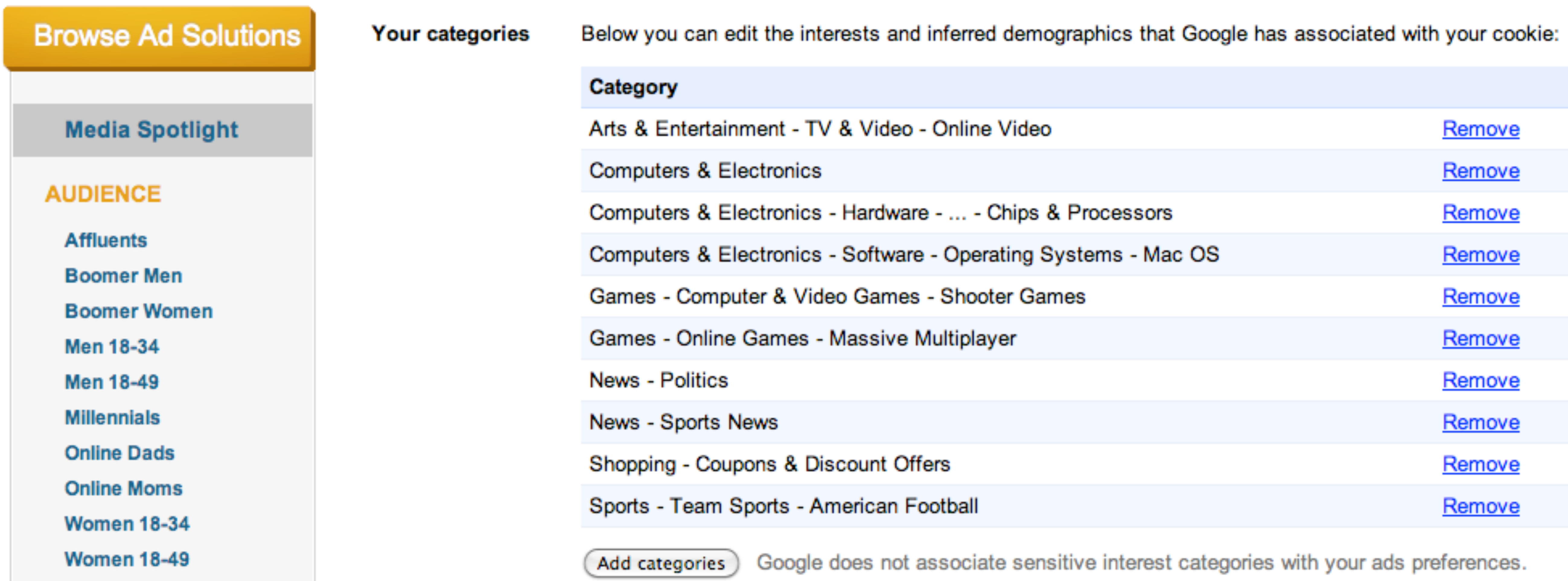
Topic Detection and Tracking research was pursued under the **DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program**:

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.

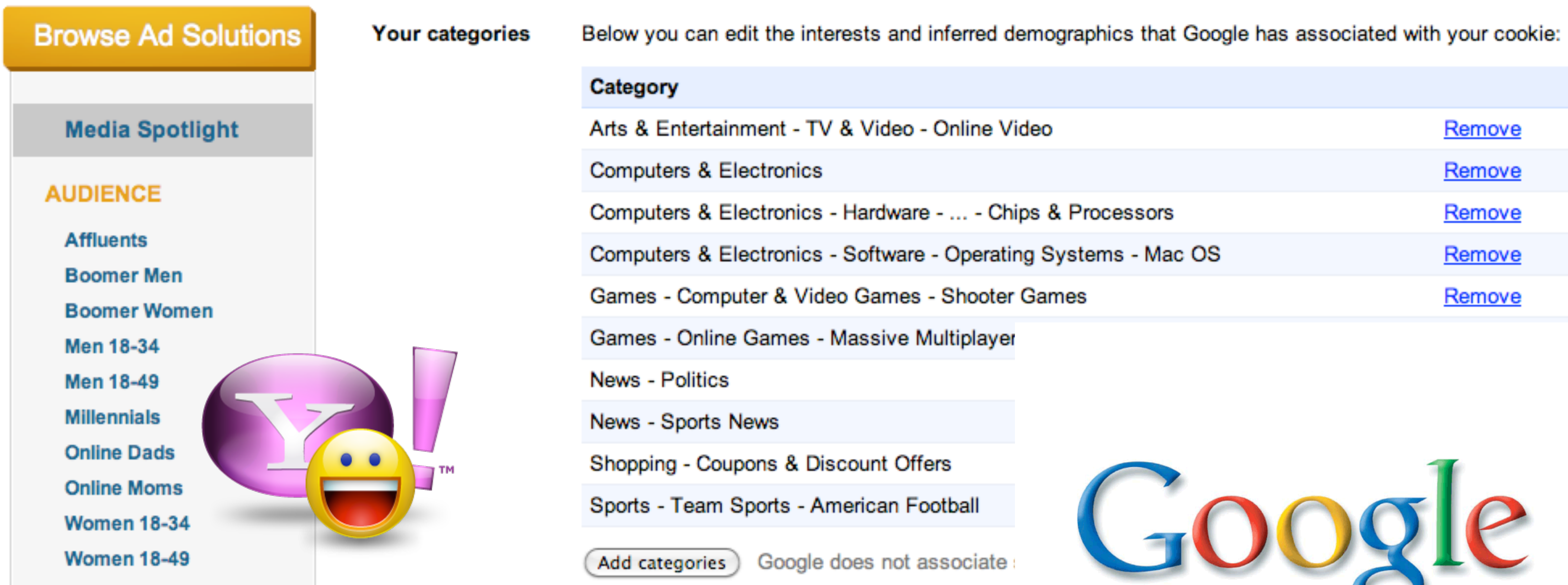
- editorially curated training data
- expensive to generate
- subjective in selection of threads
- language specific

Advertising Targeting



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

Advertising Targeting



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

Challenges

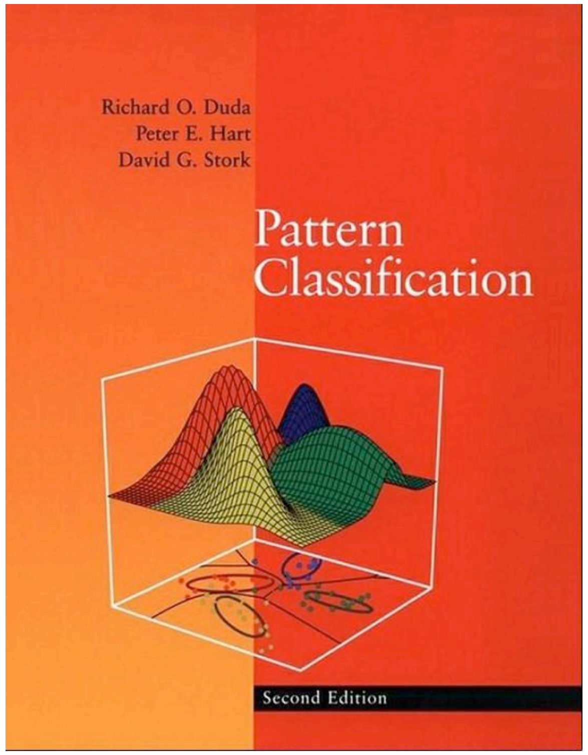
- Scale
  - Millions to billions of instances
    (documents, clicks, users, messages, ads)
  - Rich structure of data (ontology, categories, tags)
  - Model description typically larger than memory of single workstation
- Modeling
  - Usually clustering or topic models do not solve the problem
  - Temporal 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

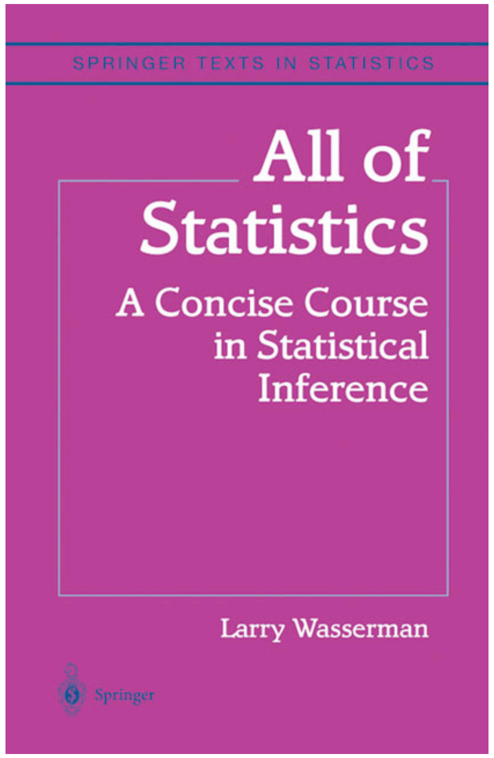
Summary - Part 1

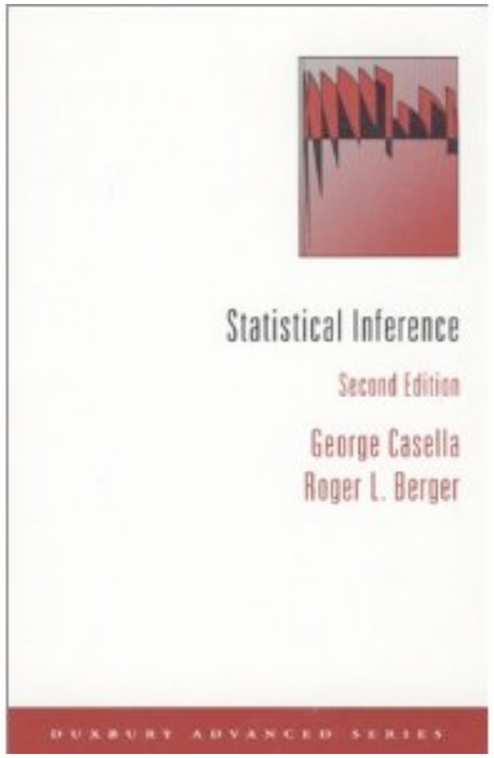
- Essentially infinite amount of data
- Labeling is prohibitively expensive
- Not scalable for i18n
- Even for *supervised* problems unlabeled data abounds. Use it.
- User-understandable structure for representation purposes
- Solutions are often customized to problem
  We can only cover building blocks 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(server working) = 0.999
  - P(90,000 < 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

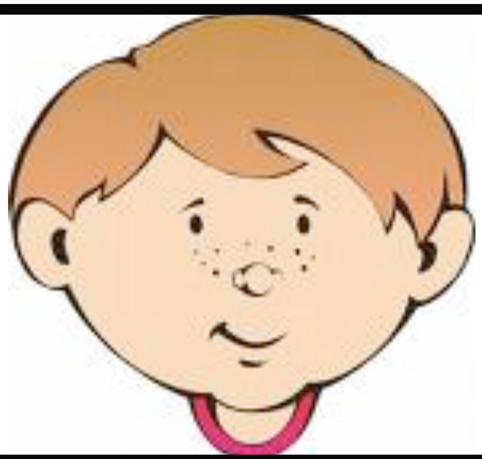
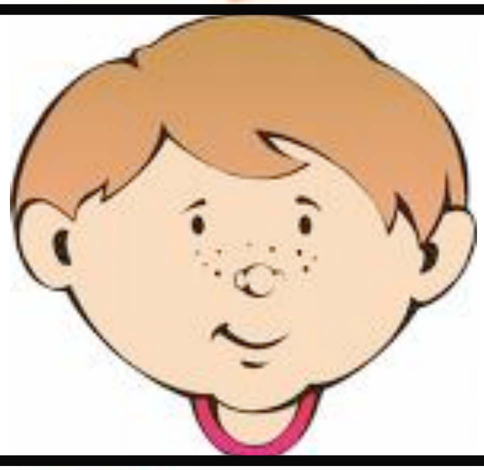
- Independence $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
  - Login behavior of two users (approximately)
  - Disk crash in different colos (approximately)
- Dependent events
  - Emails $\Pr(x, y) \neq \Pr(x) \cdot \Pr(y)$
  - Queries
  - News stream / Buzz / Tweets
  - IM communication
  - Russian Roulette

(In)dependence

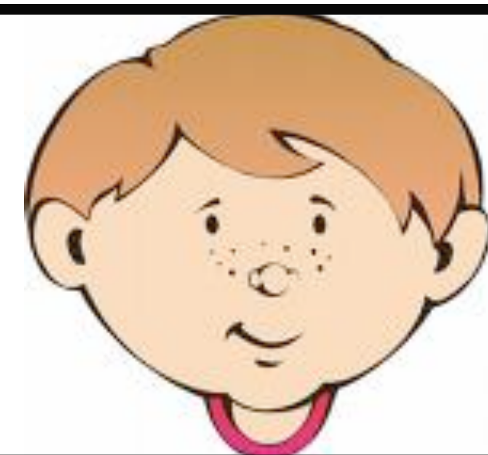
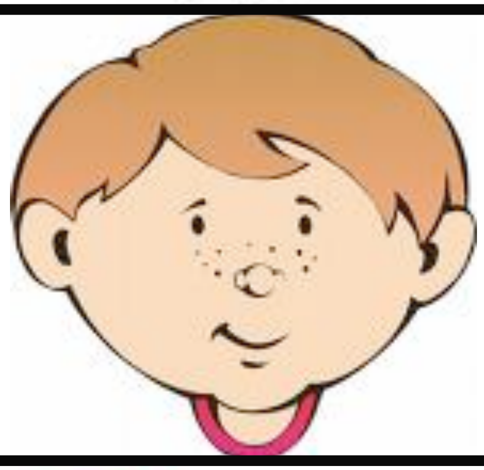
- Independence $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
  - Login behavior of two users (approximately)
  - Disk crash in different colos (approximately)
- Dependent events
  - Emails $\Pr(x, y) \neq \Pr(x)\ \Pr(y)$
  - Queries
  - News stream / Buzz / Tweets
  - IM communication
  - Russian Roulette

Everywhere!

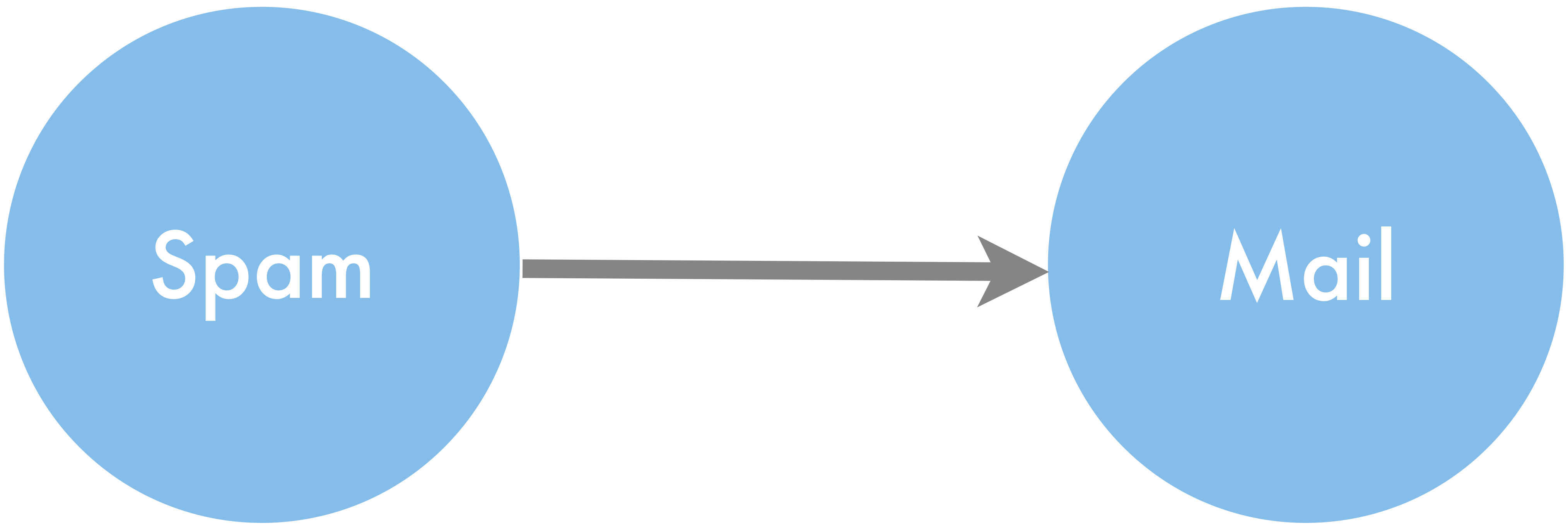
Independence

		
	0.3	0.2
	0.3	0.2

Dependence

		
	0.45	0.05
	0.05	0.45

A Graphical Model



p(spam, mail) = p(spam) p(mail|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

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

Naive Bayes Spam Filter

- Key assumption
  Words occur independently of each other given the label of the document

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

- Spam classification via Bayes Rule

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

Naive Bayes Spam Filter

- Key assumption
  Words occur independently of each other given the label of the document

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

- Spam classification via Bayes Rule

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

- 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.

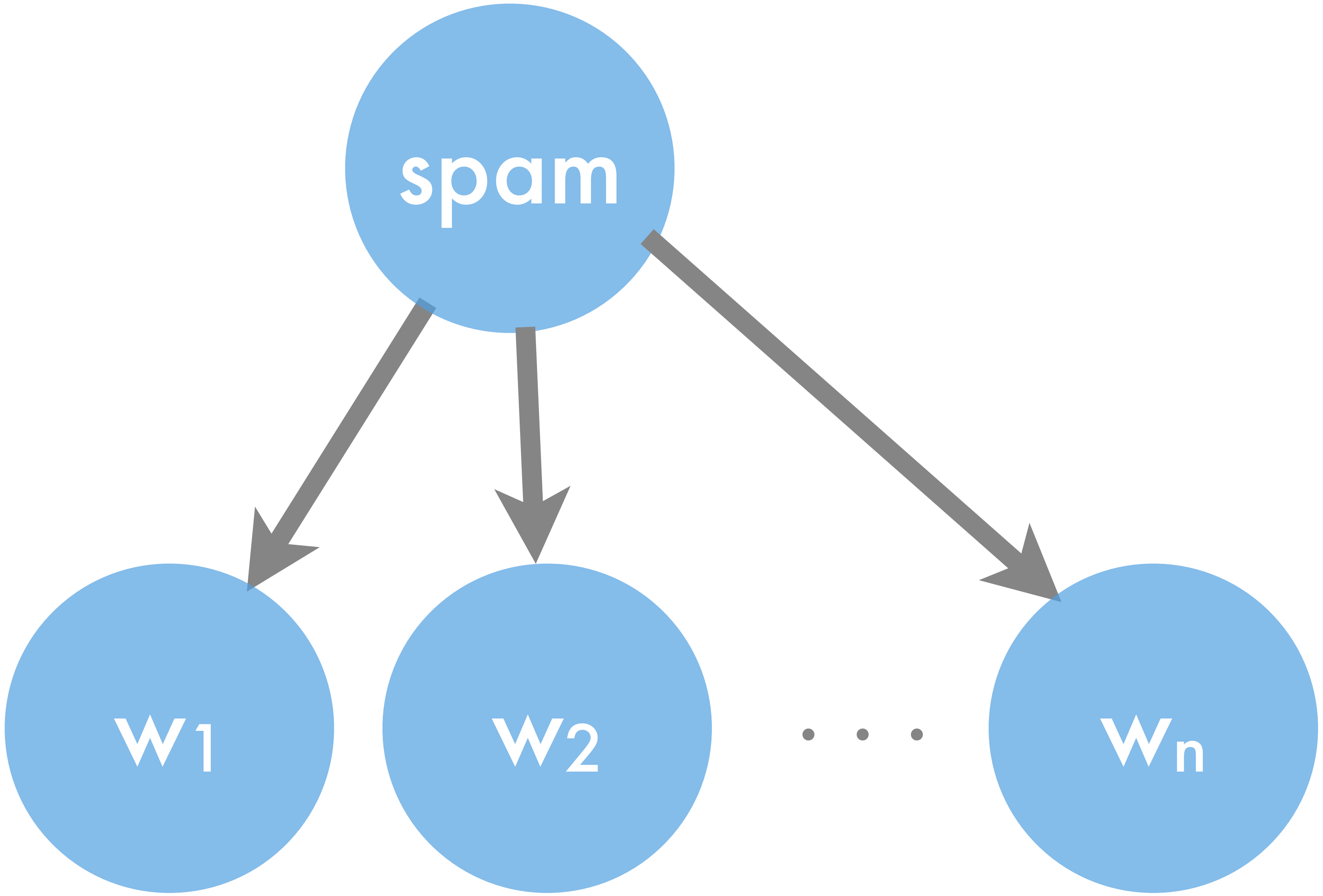
Naive Bayes Spam Filter

Equally likely phrases

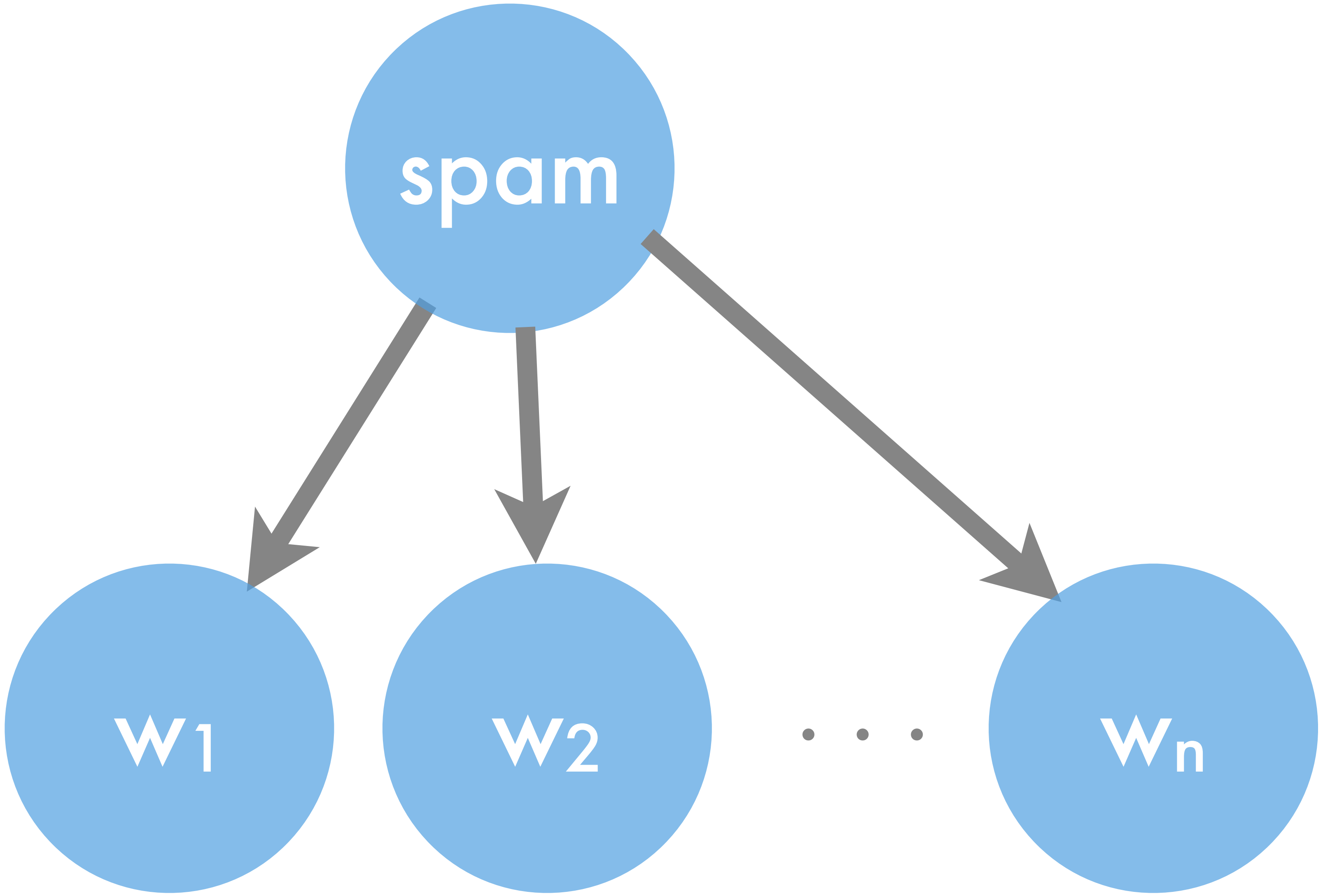
- Get rich quick. Buy WWW stock.
- Buy Viagra. Make your WWW experience last longer.
- You deserve a PhD from WWW University.
  We recognize your expertise.

- Make your rich WWW PhD experience last longer.

A Graphical Model

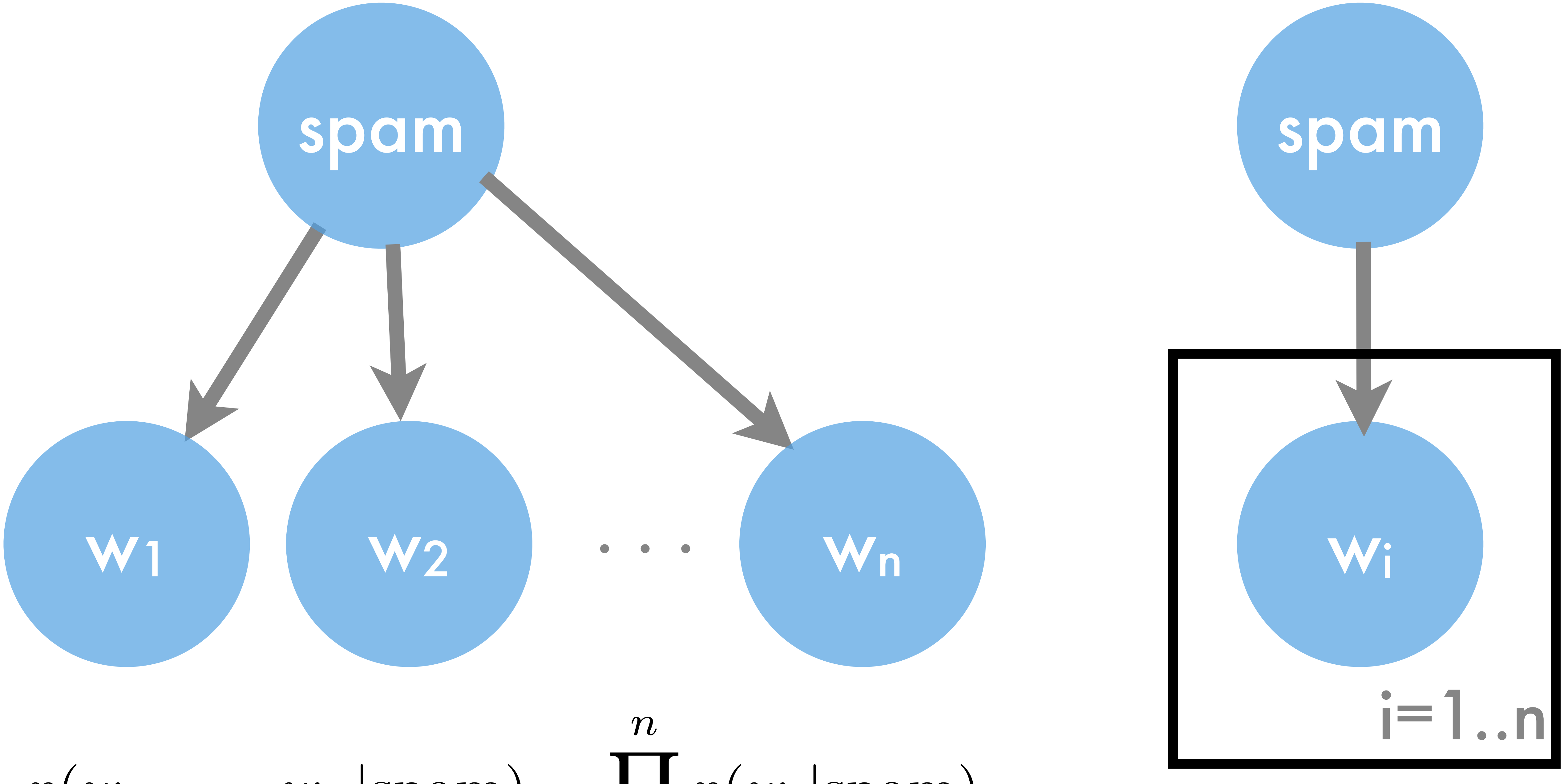


A Graphical Model



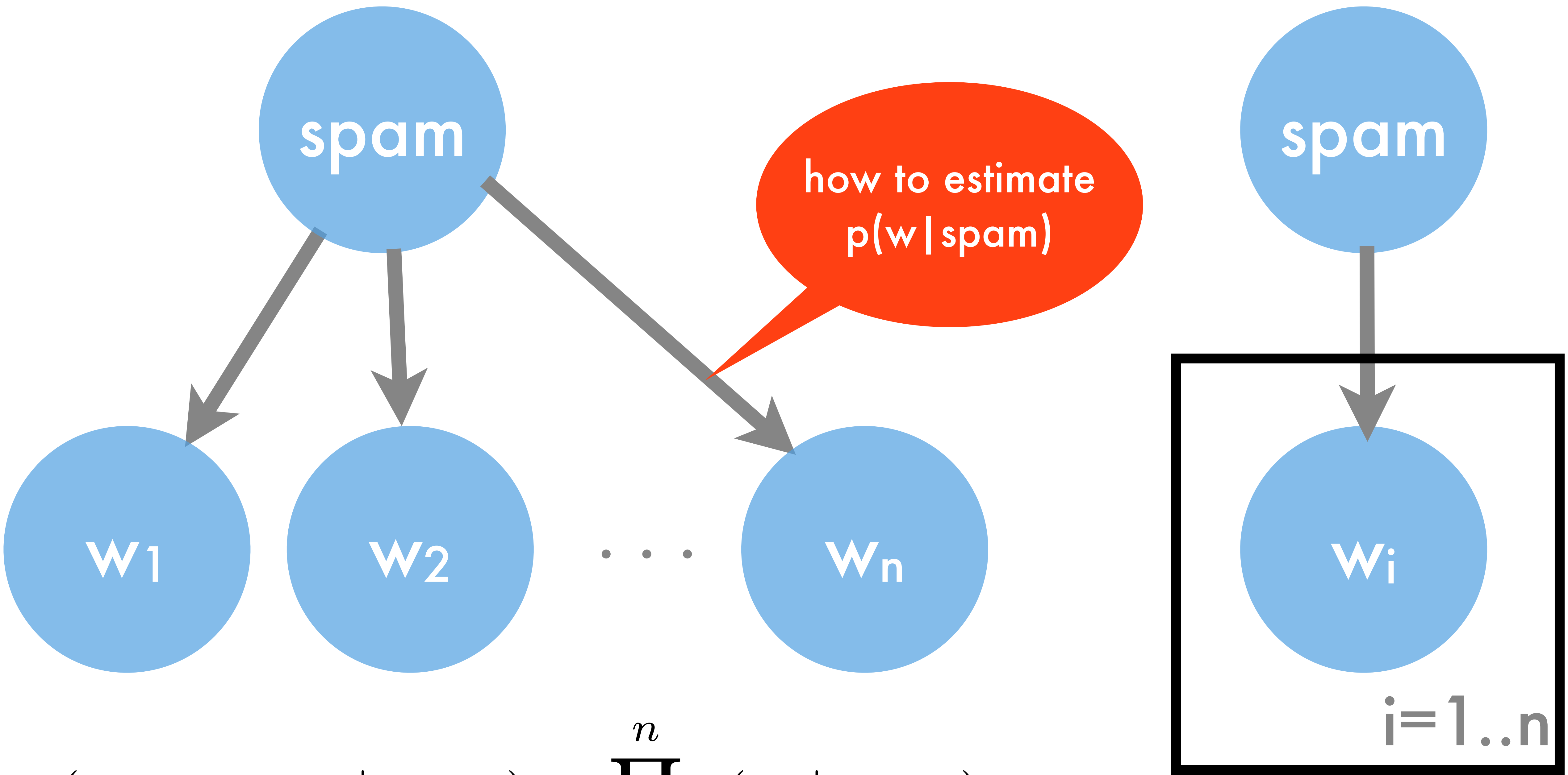
$$p(w_1, \ldots, w_n|\text{spam}) = \prod_{i=1}^{n} p(w_i|\text{spam})$$

A Graphical Model



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

A Graphical Model



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

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

spam probability

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

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

Naive NaiveBayes Classifier

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

what if n(i,y)=0?

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

Naive NaiveBayes Classifier

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

what if n(i,y)=0?

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

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

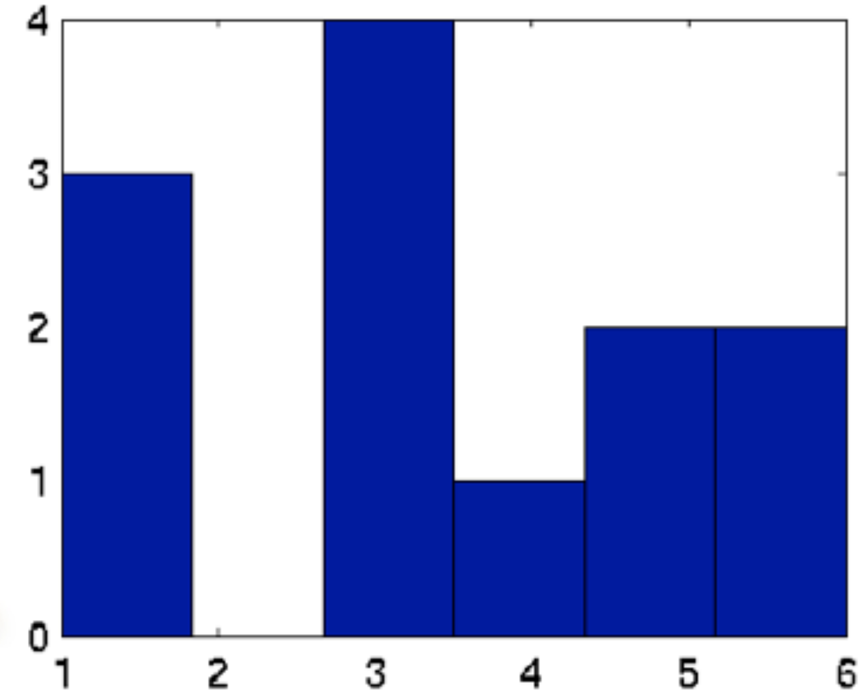
$$p(X;\pi) = \prod_i \pi_i^{n_i}$$

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

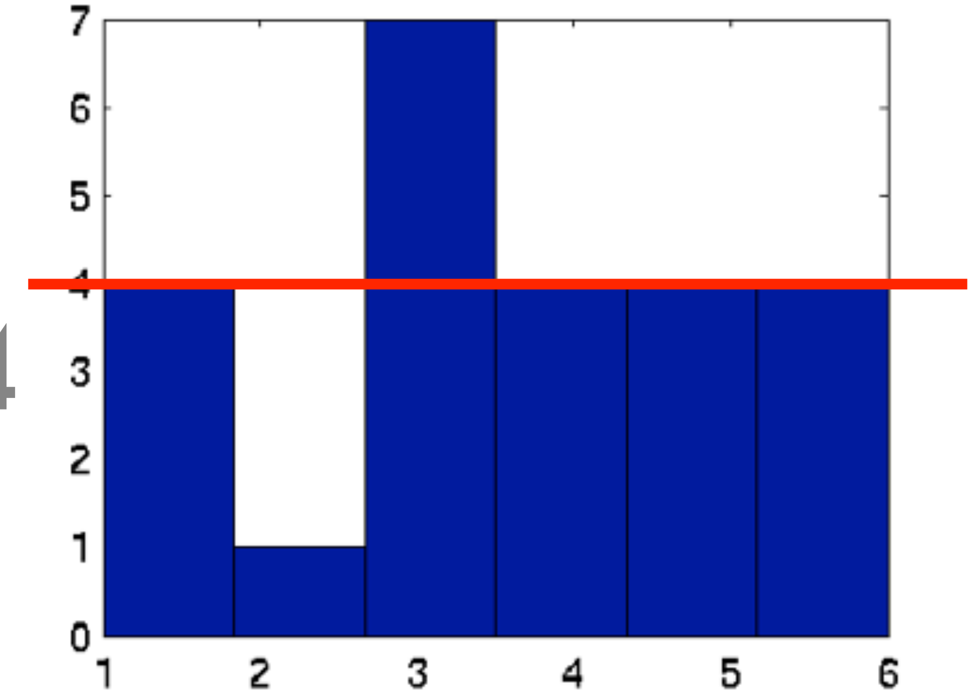
$$\pi_i = \frac{n_i}{\sum_j n_j}$$

Tossing a Dice

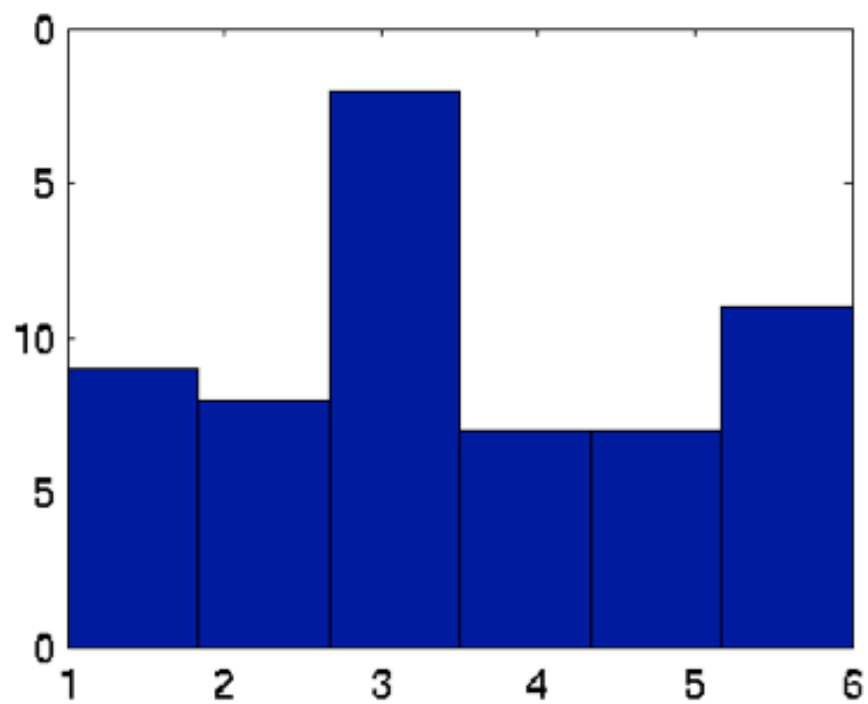
12



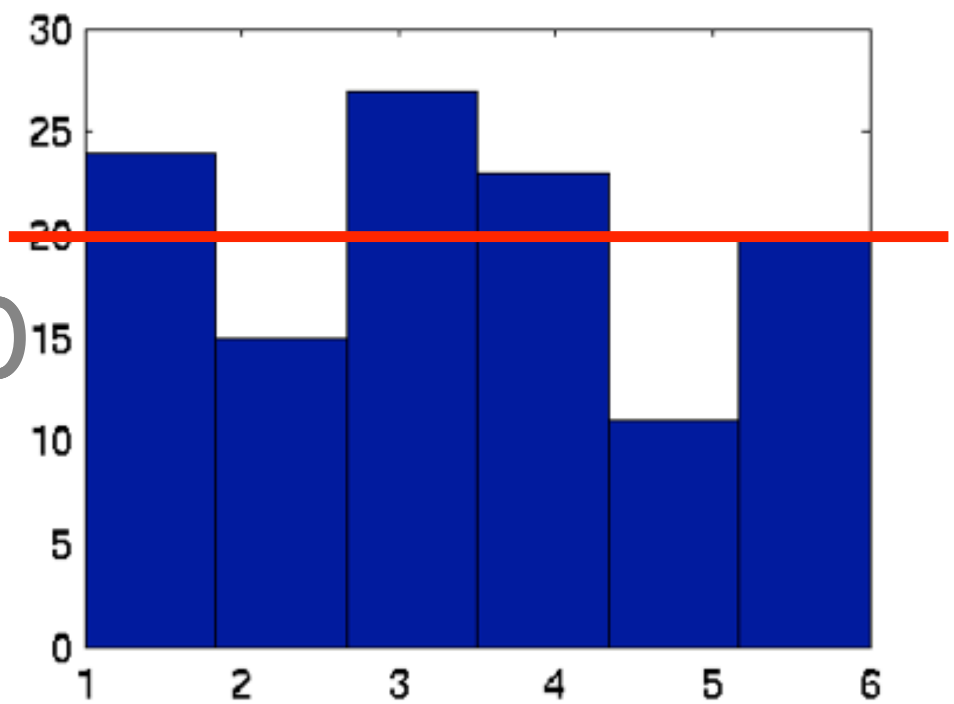
24



60

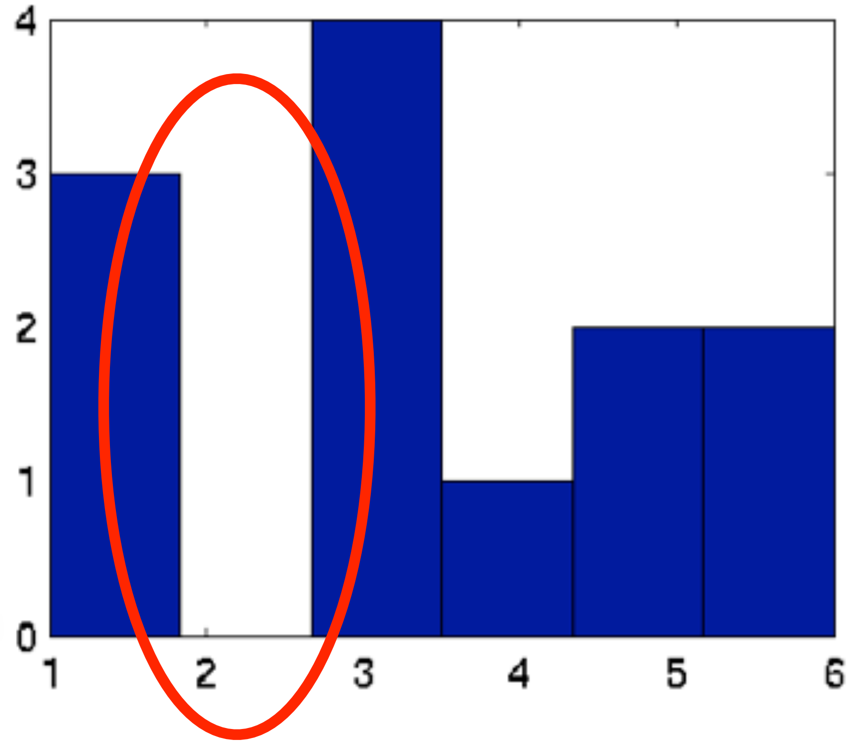


120

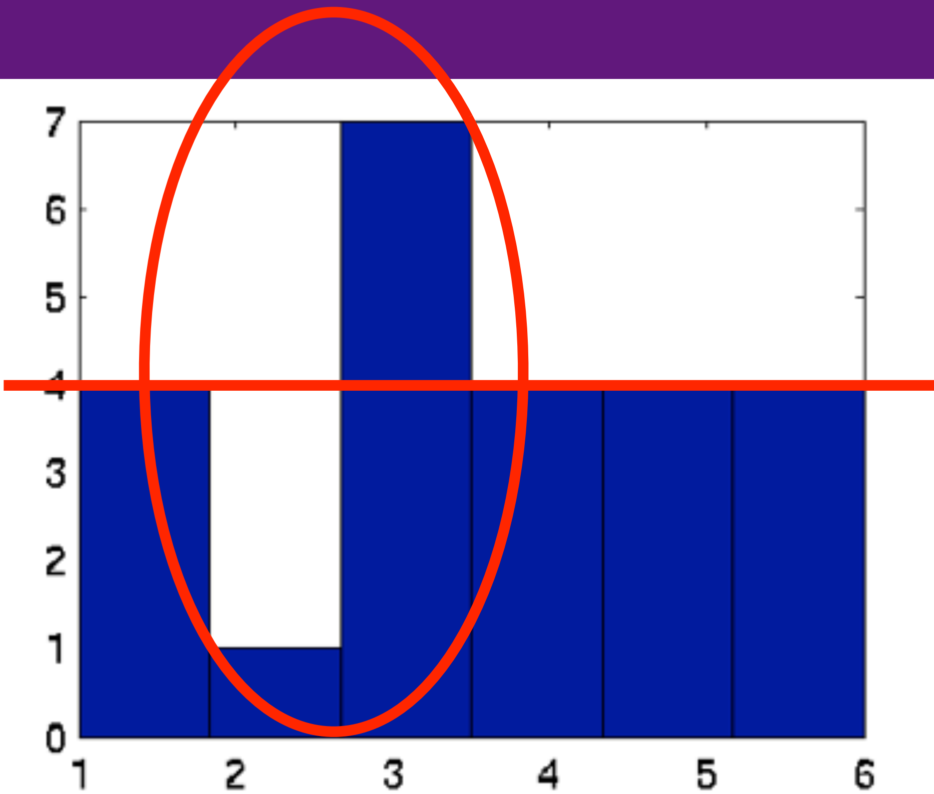


Tossing a Dice

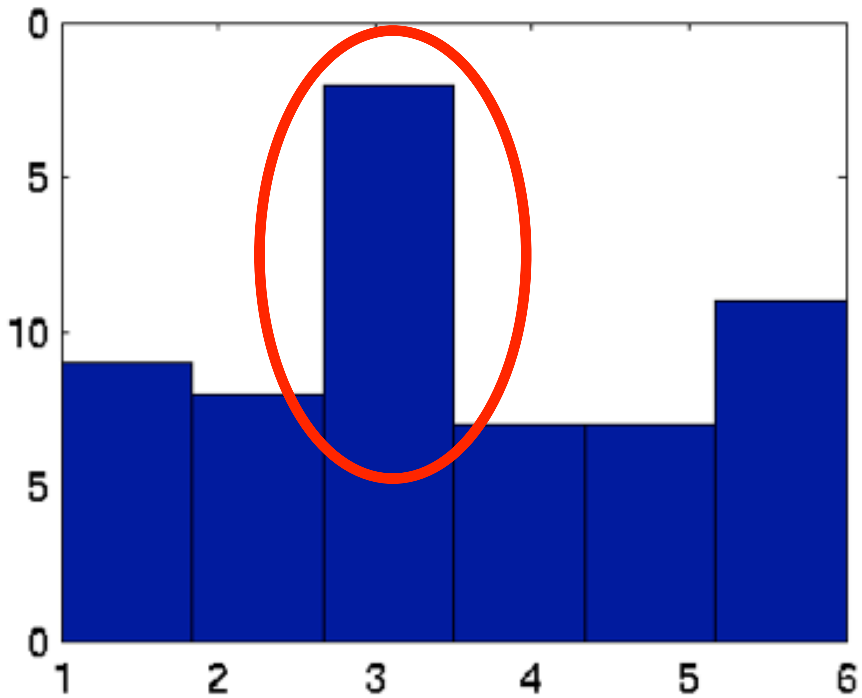
12



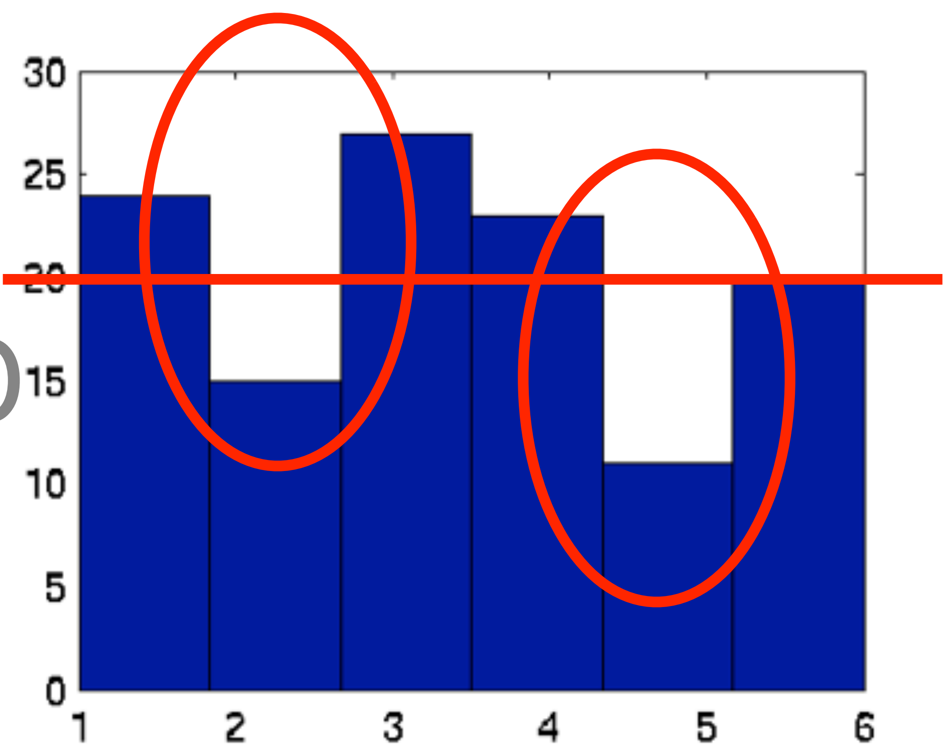
24



60



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**

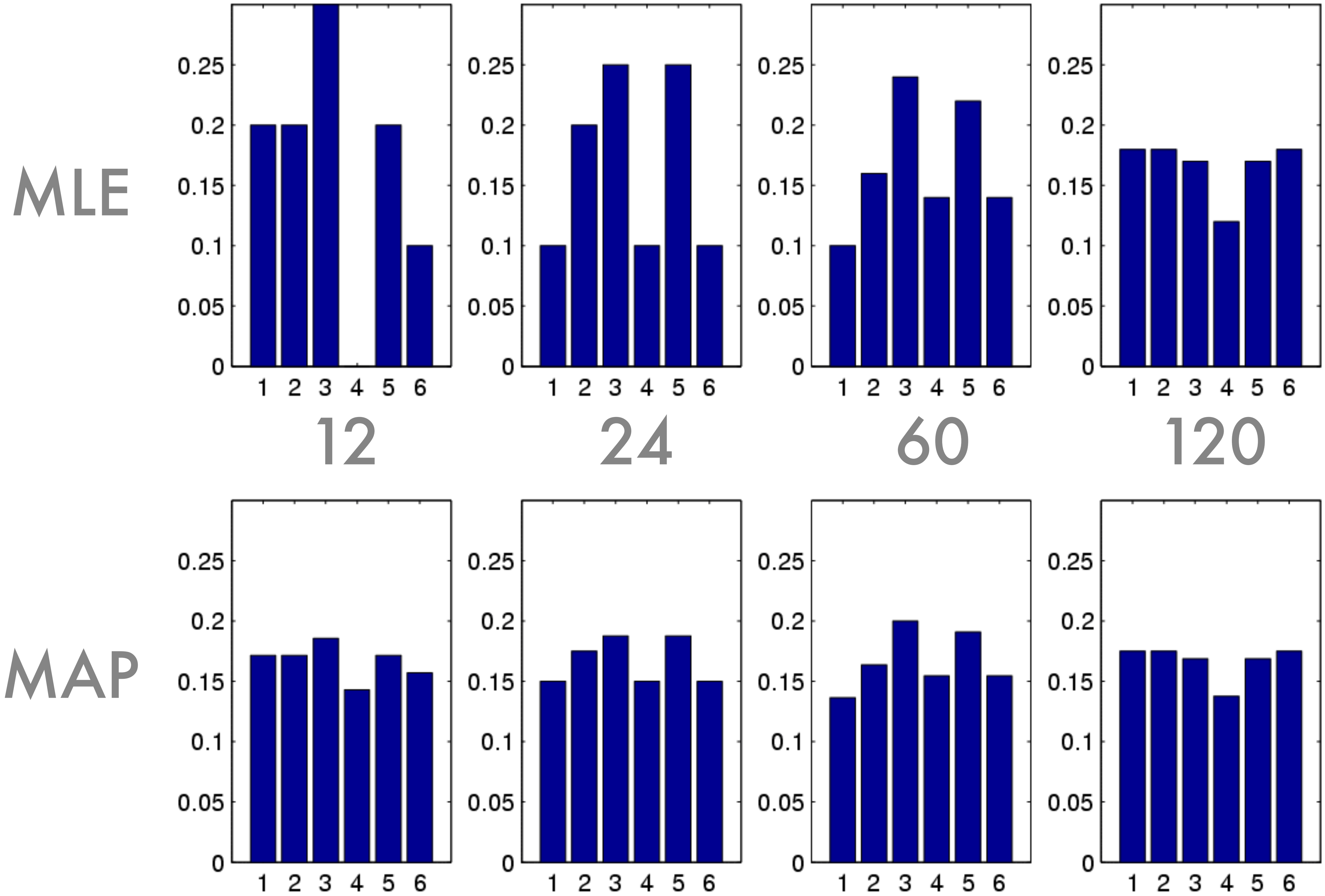
$$p(x = i) = \frac{n_i}{n} \longrightarrow p(x = i) = \frac{n_i + m_i}{n + m}$$

- **Tossing a dice**

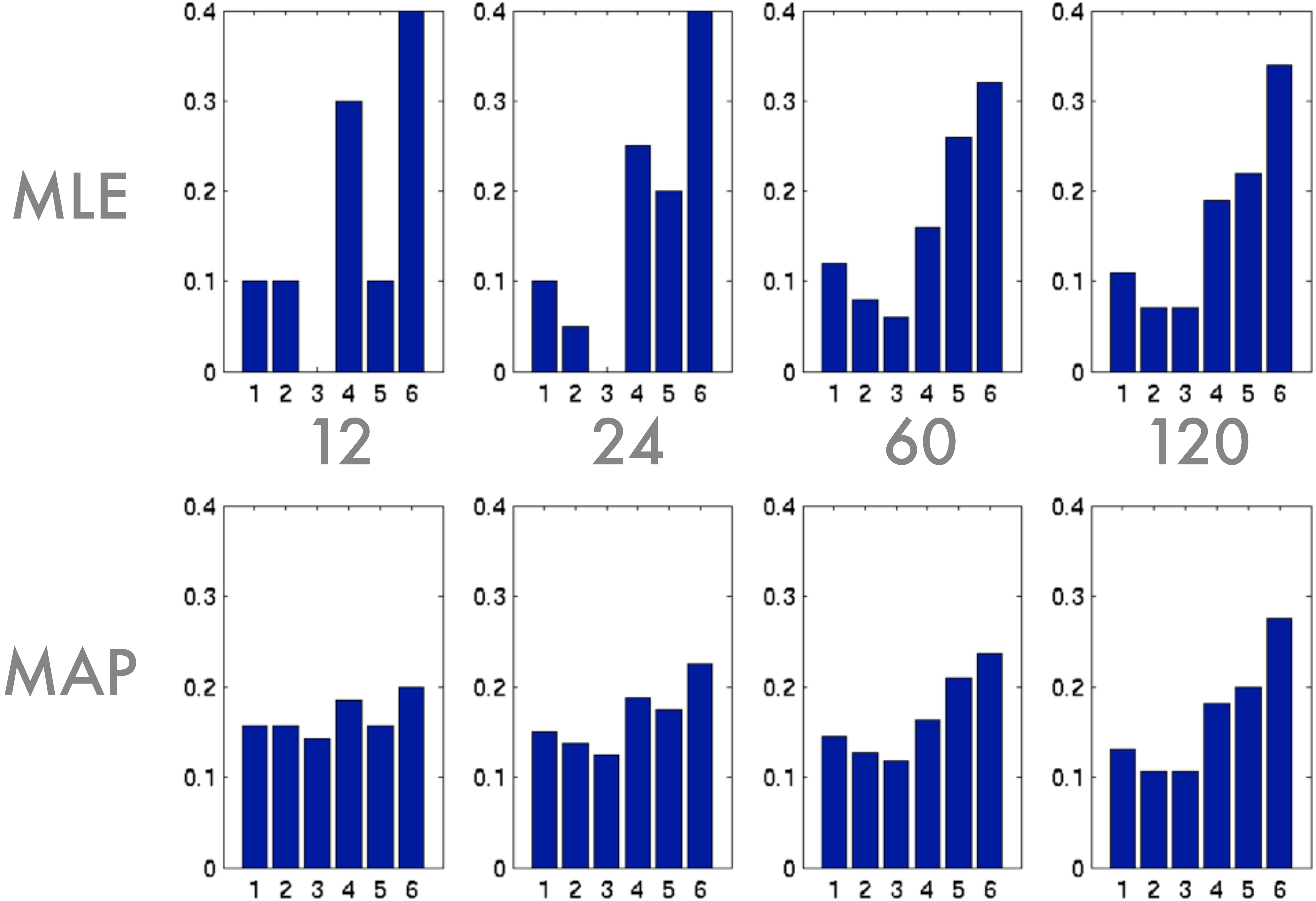
Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.15	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

- Rule of thumb
  need 10 data points (or prior) per parameter

Honest dice



Tainted dice



Exponential Families

Exponential Families

Exponential Families

- Density function

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

Exponential Families

- Density function

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

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

- Log partition function generates cumulants

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

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

Exponential Families

- Density function

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

- Log partition function generates cumulants

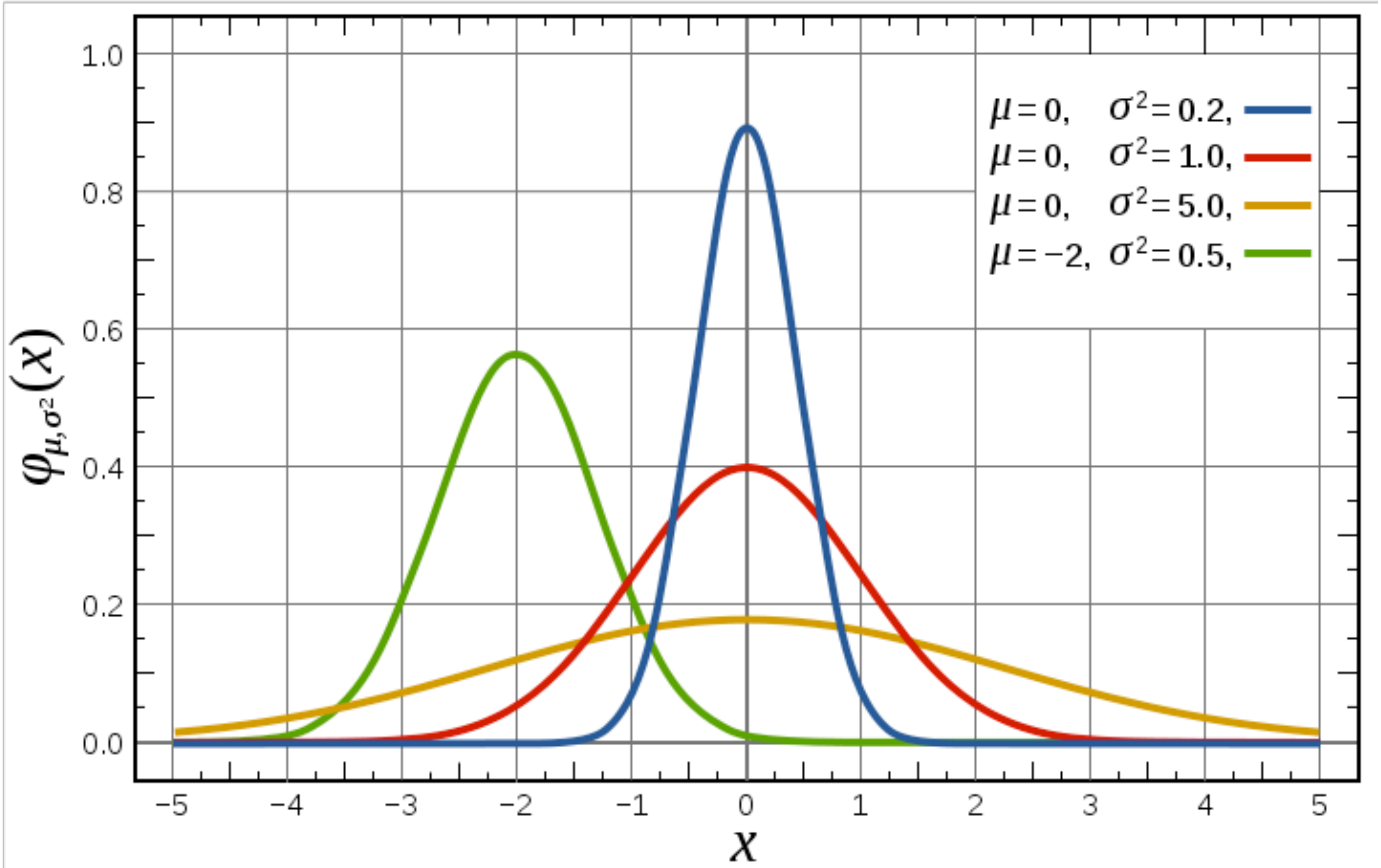
$$\partial_\theta g(\theta) = \mathbf{E}\left[\phi(x)\right]$$
$$\partial_\theta^2 g(\theta) = \mathrm{Var}\left[\phi(x)\right]$$

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

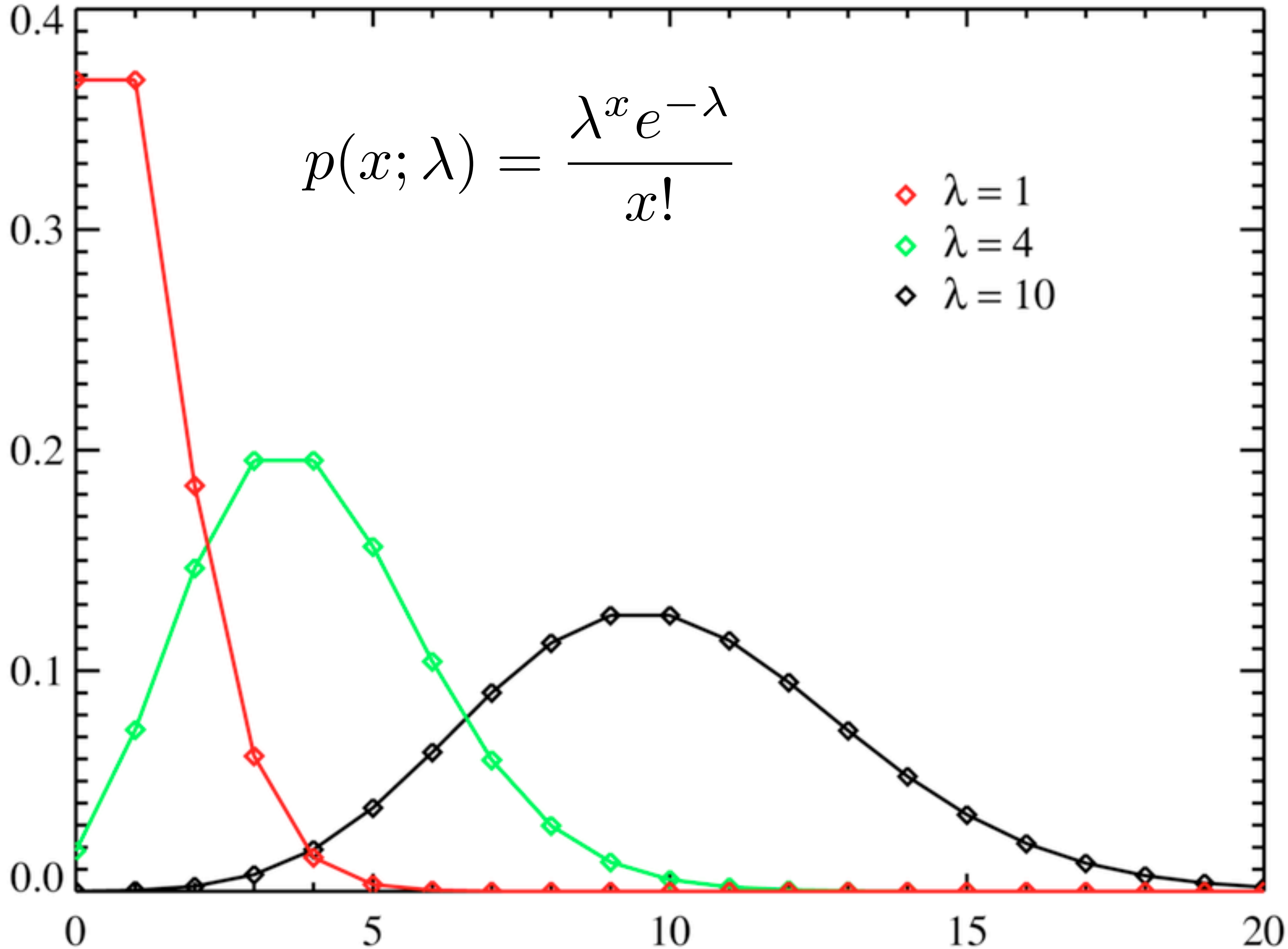
Examples

- 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, ...

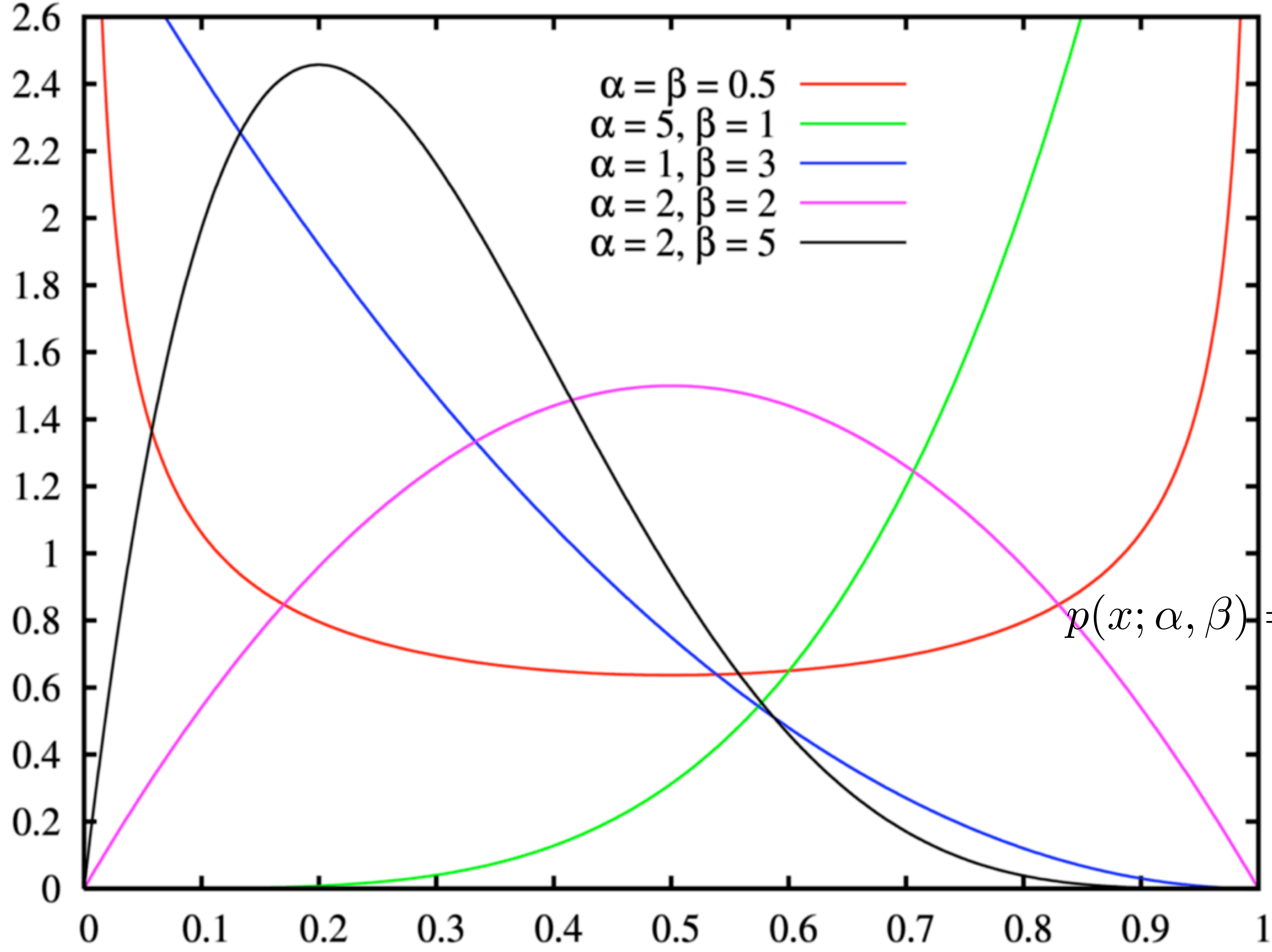
Normal Distribution



Poisson Distribution

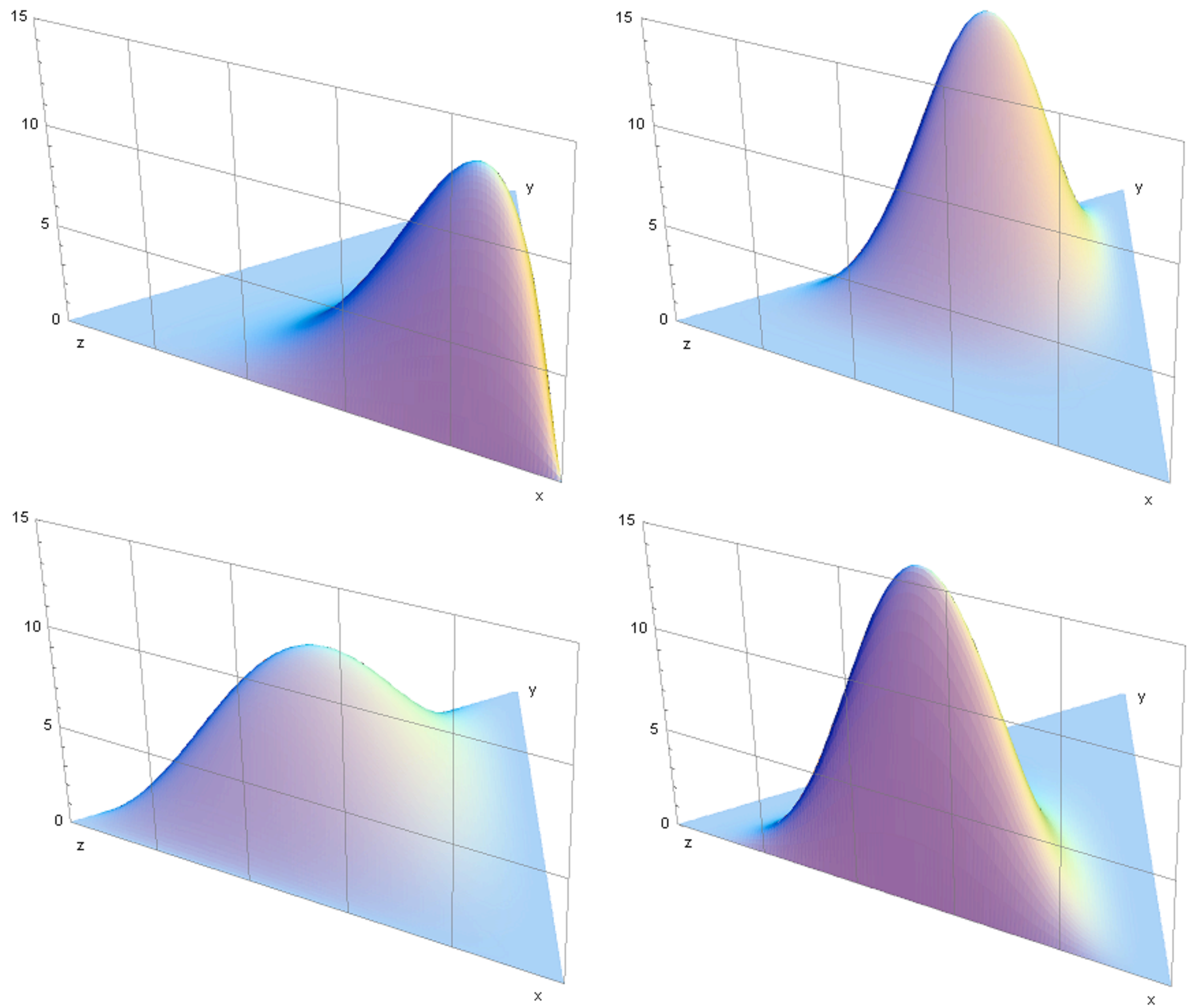


Beta Distribution



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

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**

mean

empirical average

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

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

smoother

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

smoother

Collapsing

- Conjugate priors

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

  Hence we know how to compute normalization

- Prediction

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

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

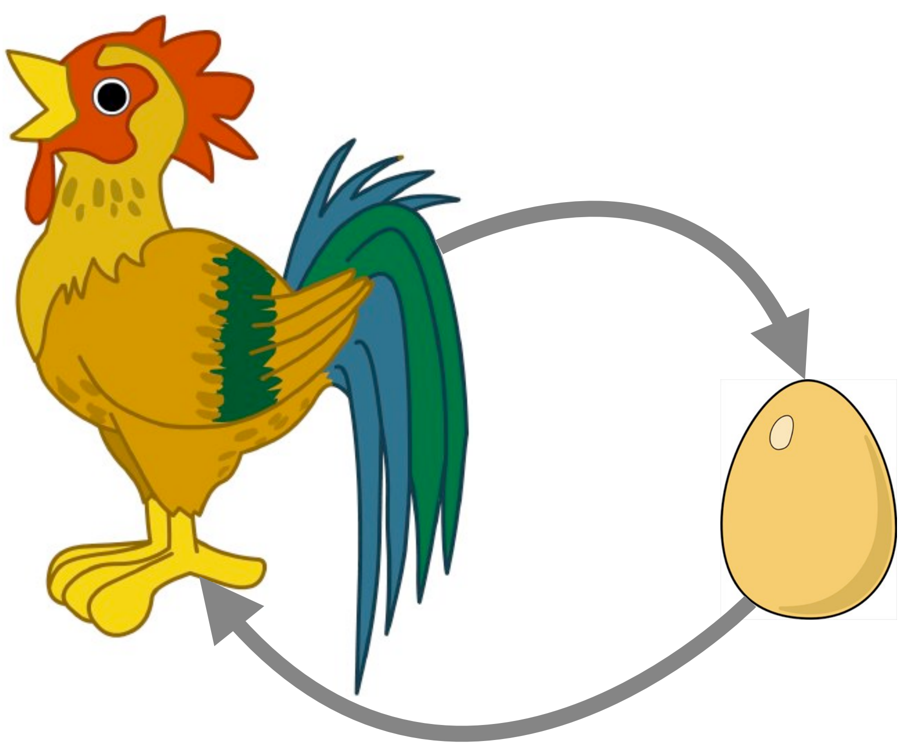
$$= \int p(\{x\} \cup X \cup X_{\text{fake}}|\theta)d\theta$$

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

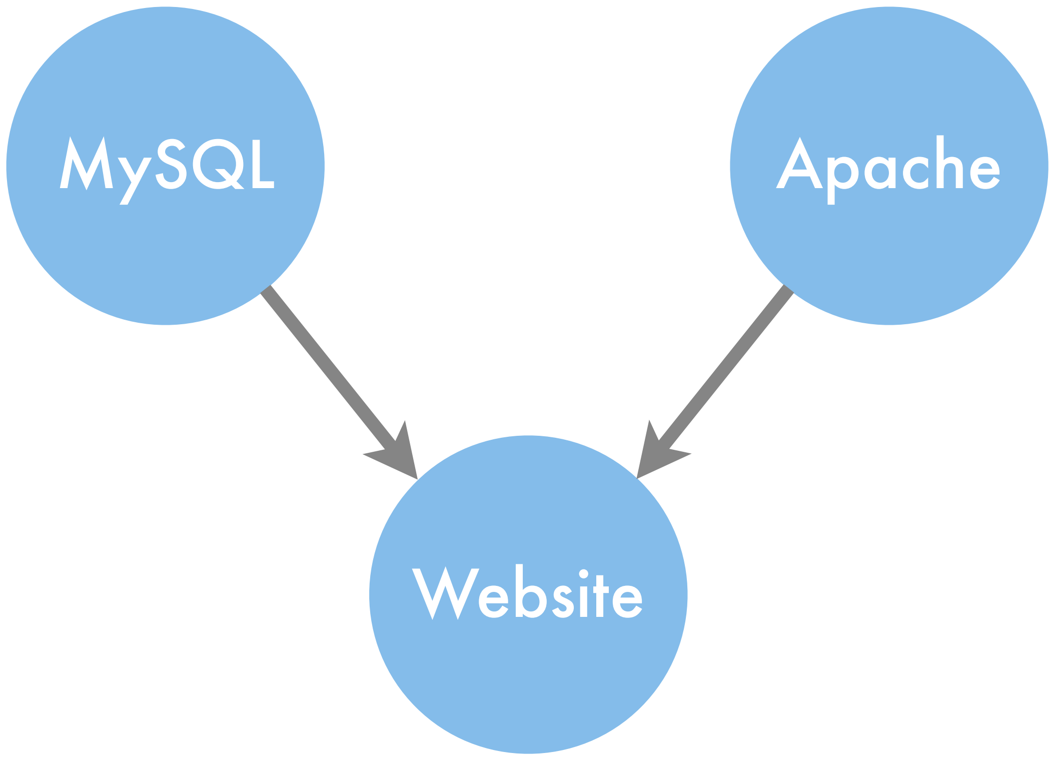
look up closed form expansions

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

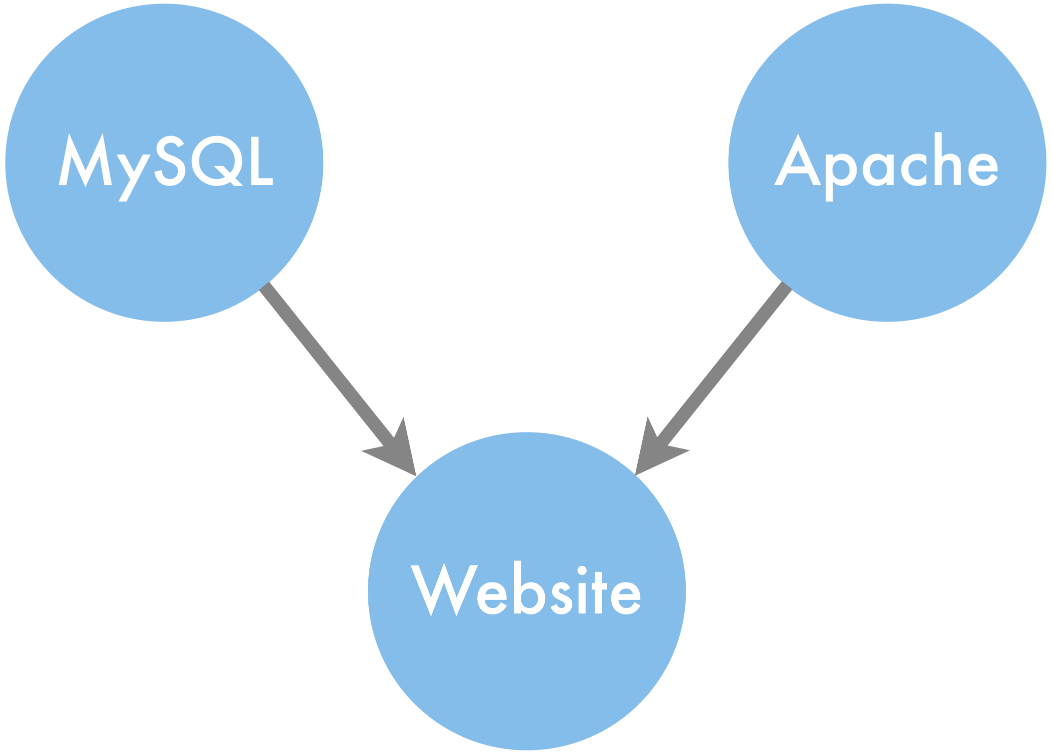
Directed Graphical Models



… some Web 2.0 service



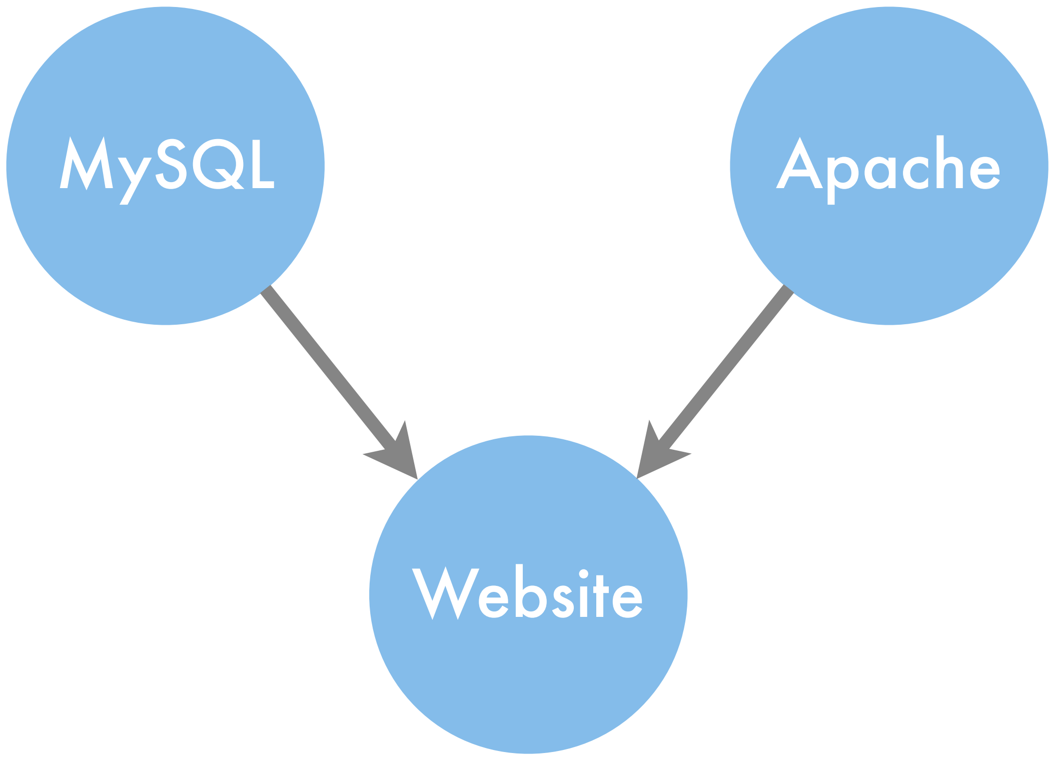
... some Web 2.0 service



- 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)

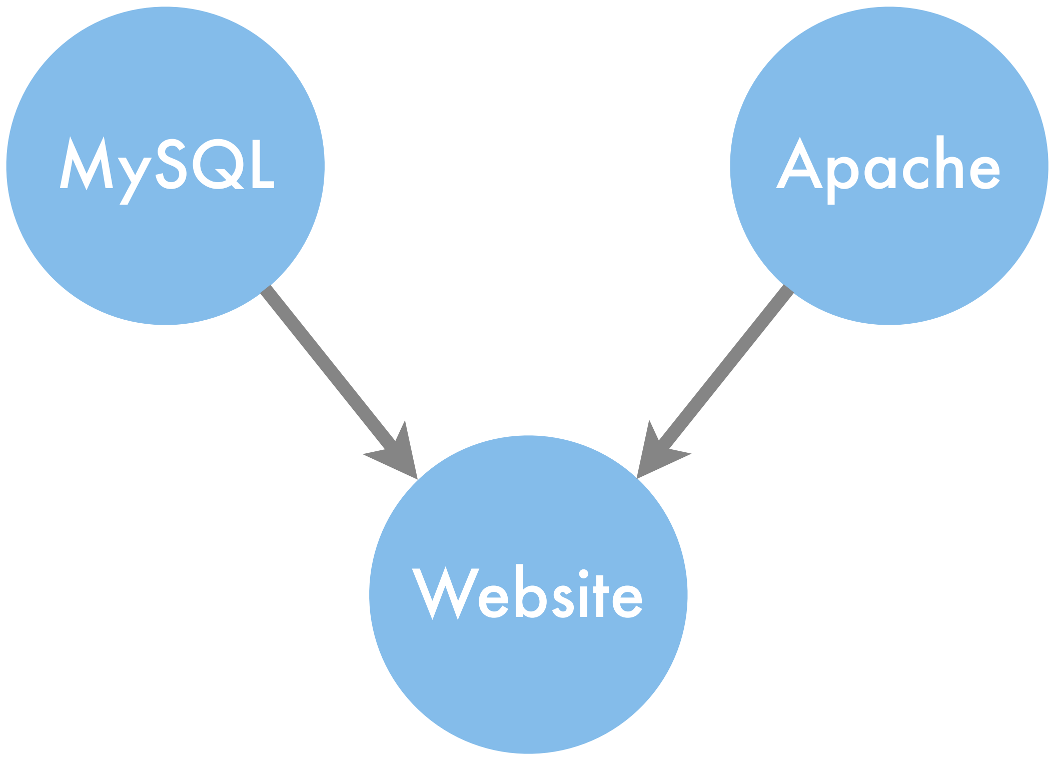
$$p(m, a, w) = p(w|m, a)p(m)p(a)$$

- Explaining away

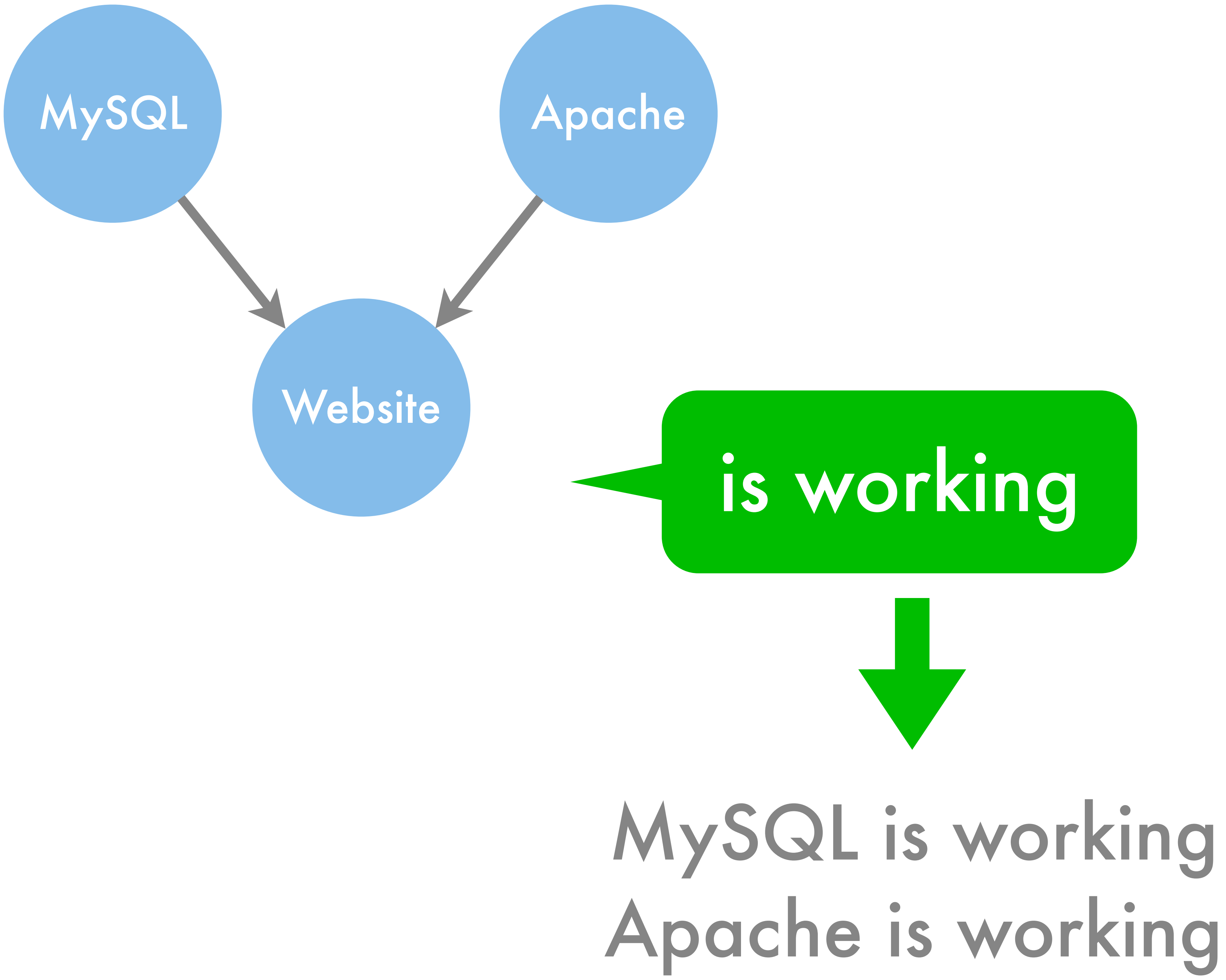
$$p(m, a|w) = \frac{p(w|m, a)p(m)p(a)}{\sum_{m', a'} p(w|m', a')p(m')p(a')}$$

a and m are dependent conditioned on w

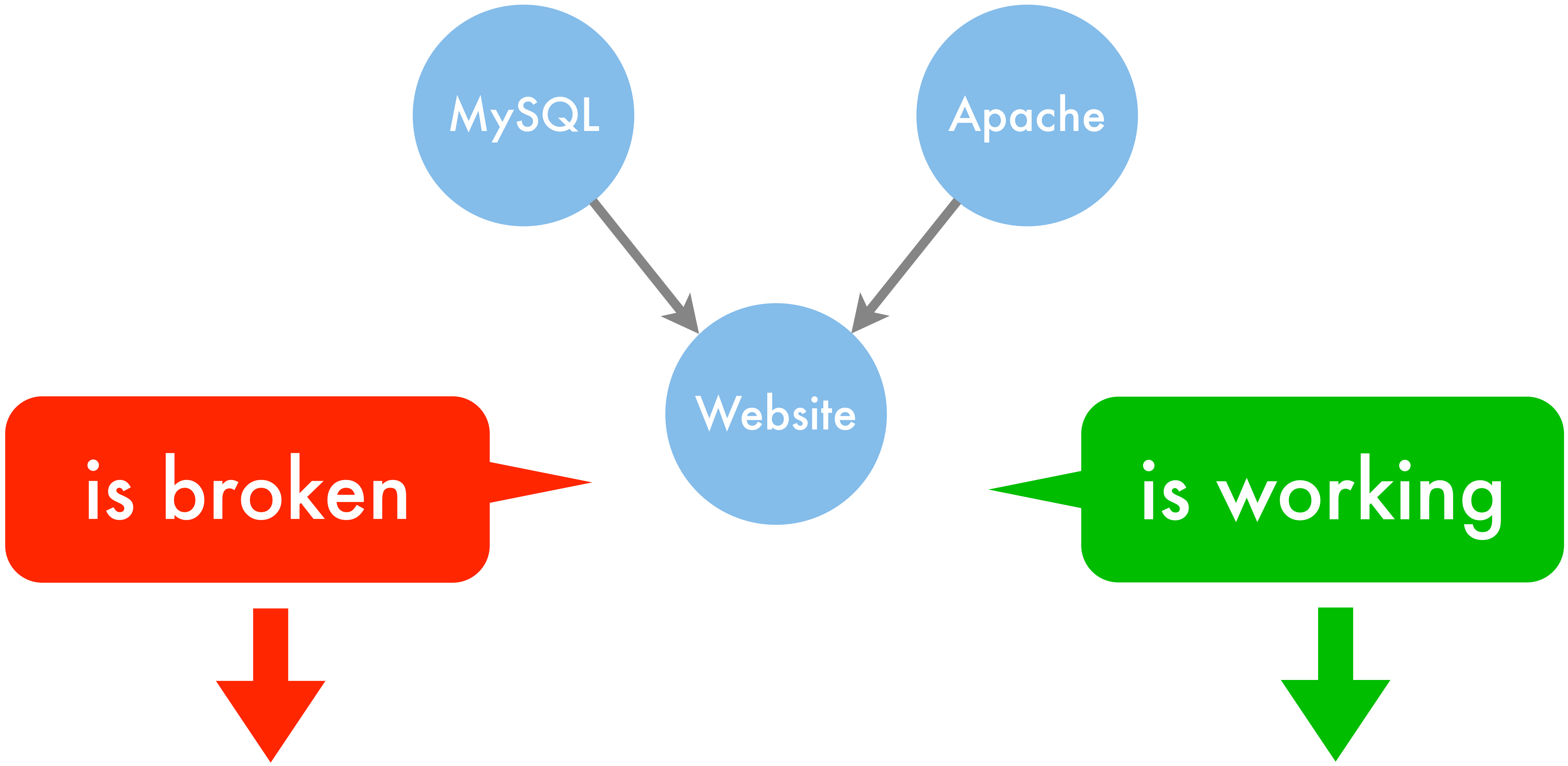
… some Web 2.0 service



... some Web 2.0 service



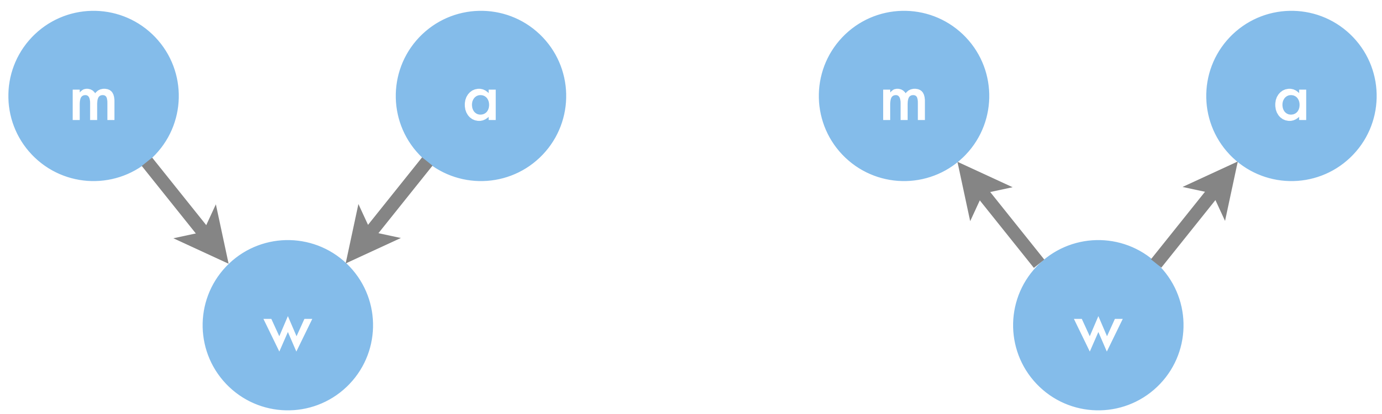
… some Web 2.0 service

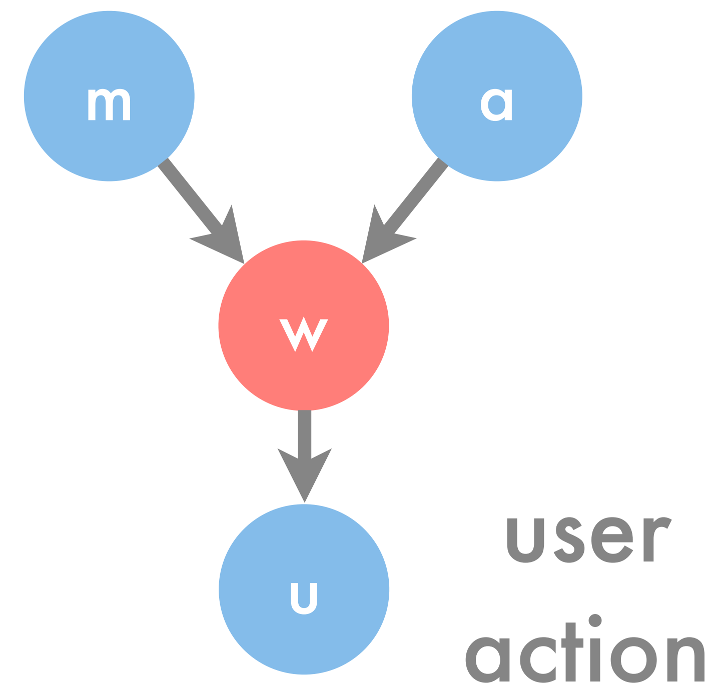


At least one of the two services is broken (not independent)

MySQL is working
Apache is working

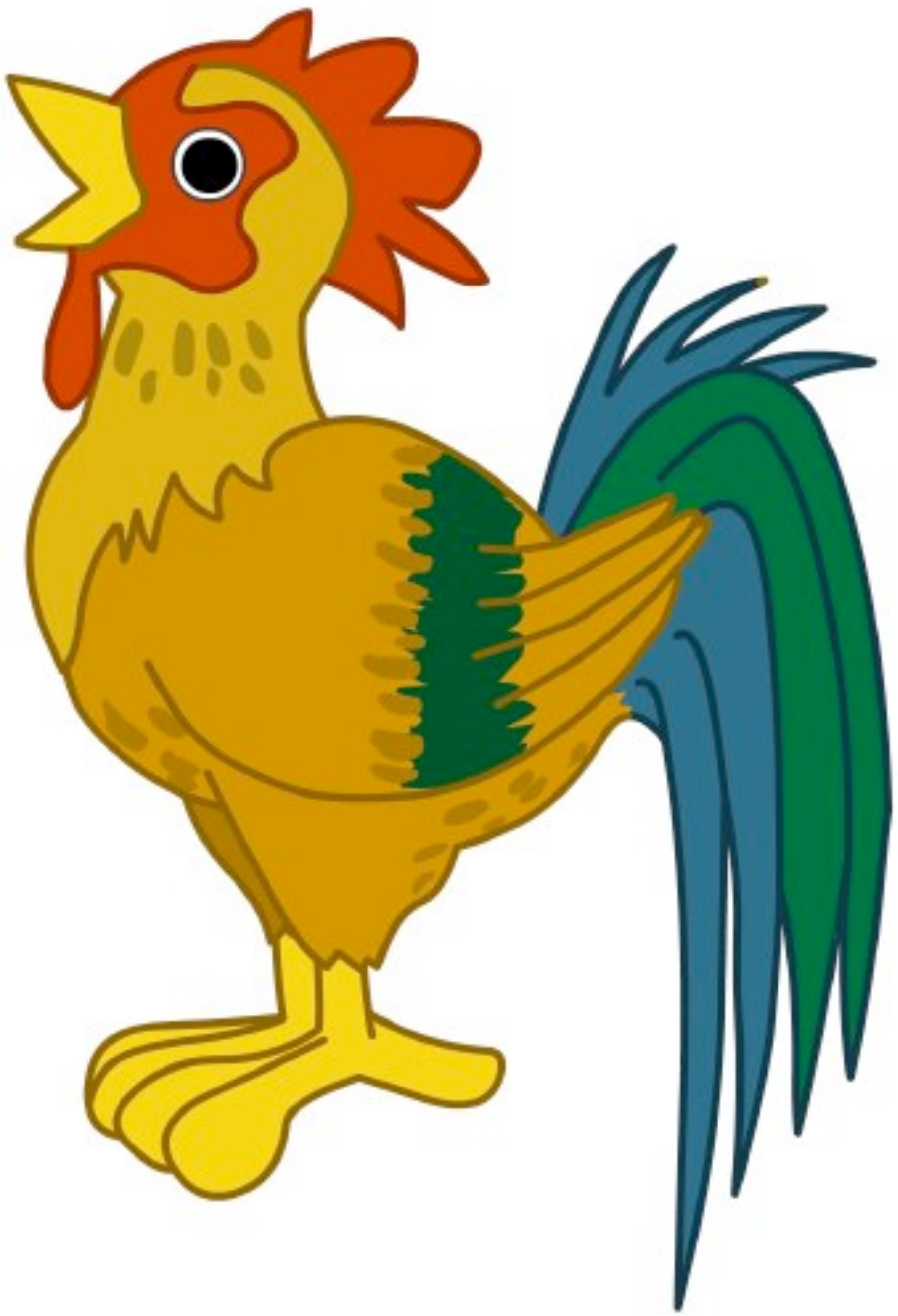
Directed graphical model



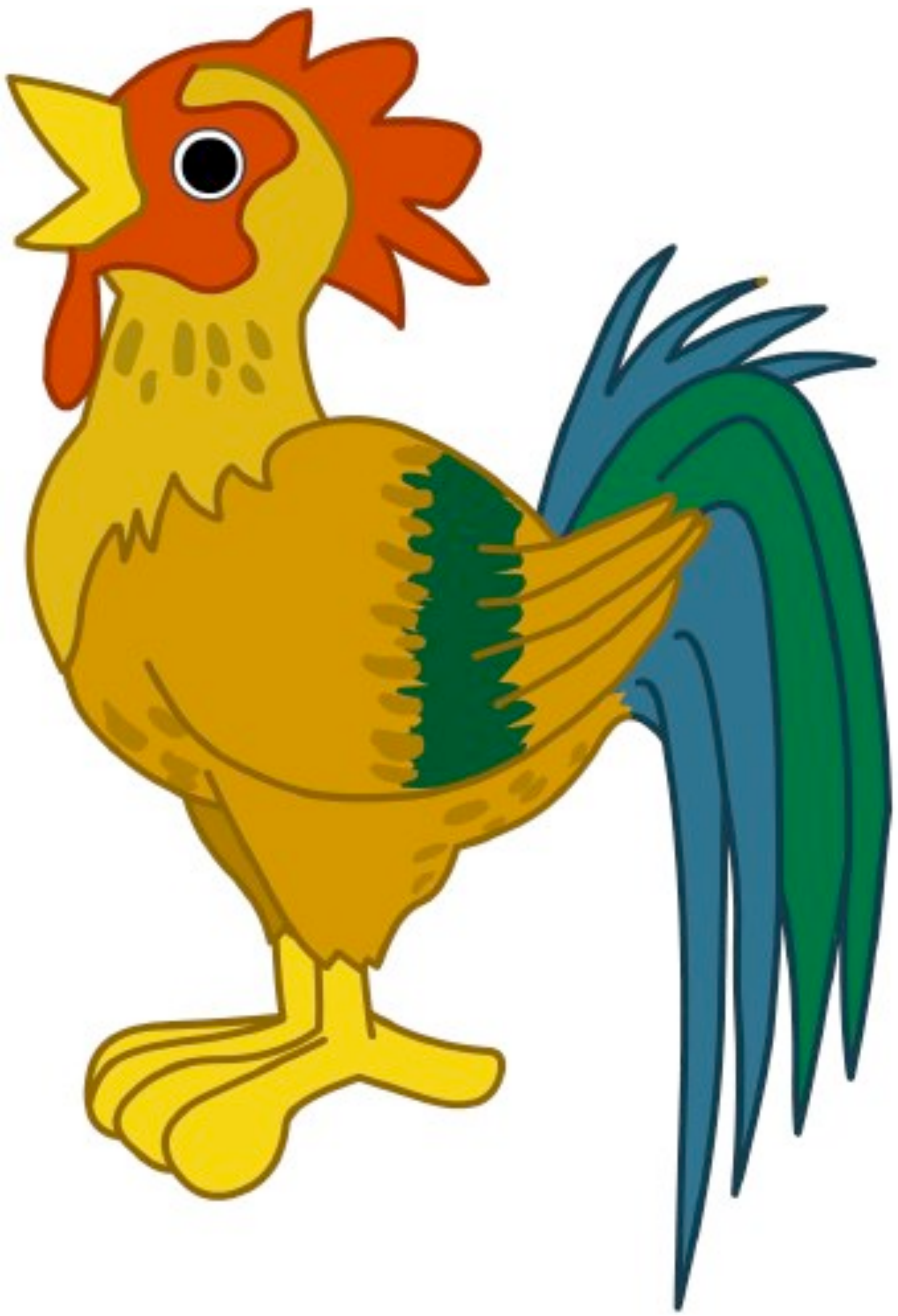


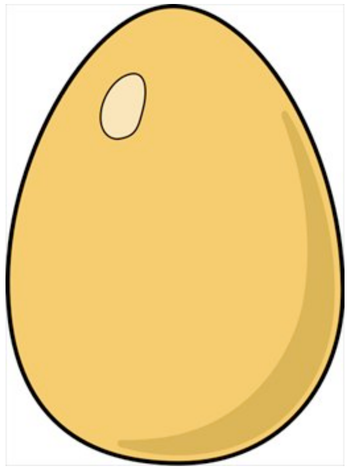
- Easier estimation
  - 15 parameters for full joint distribution
  - 1+1+3+1 for factorizing distribution
- Causal relations
- Inference for unobserved variables

No loops allowed

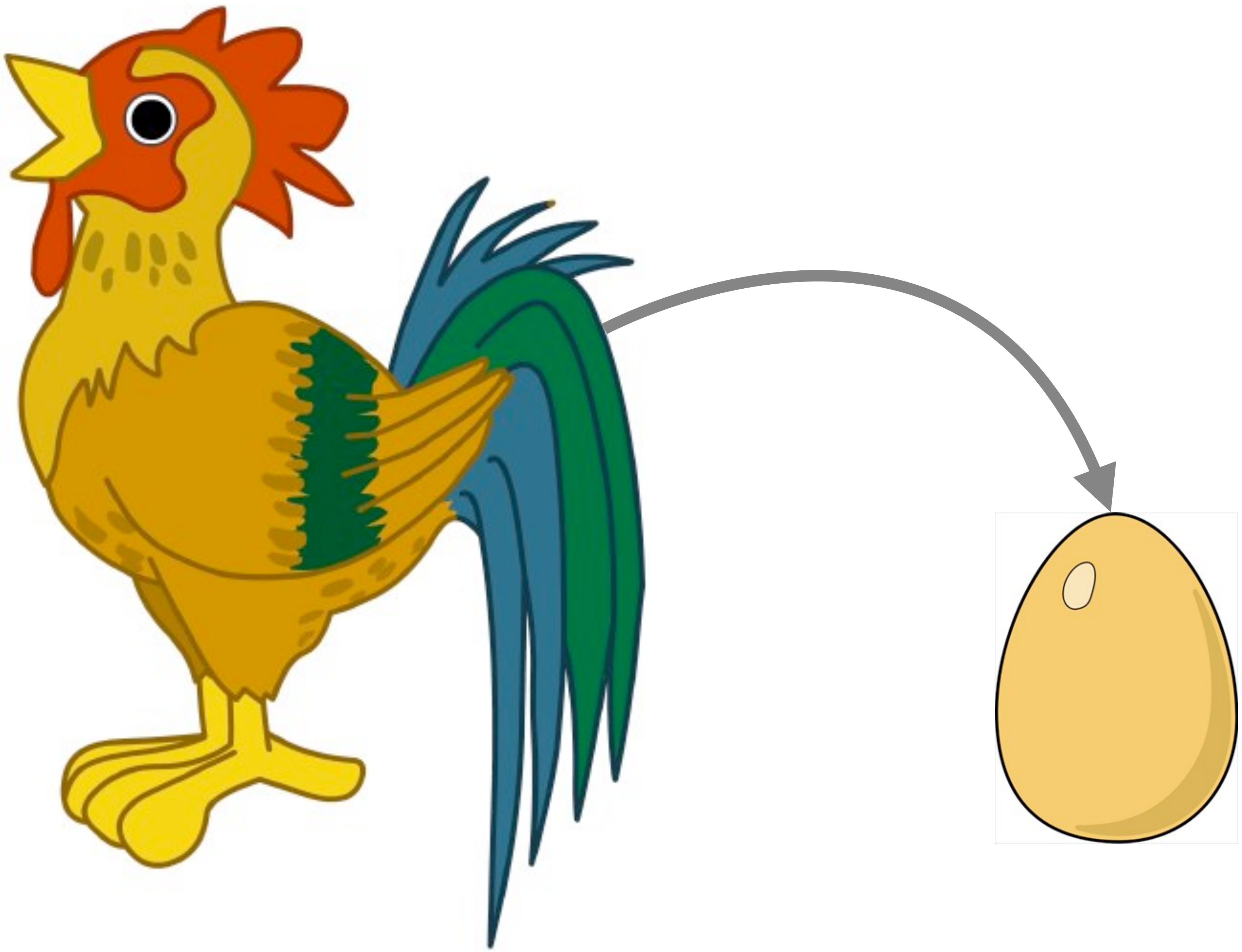


No loops allowed

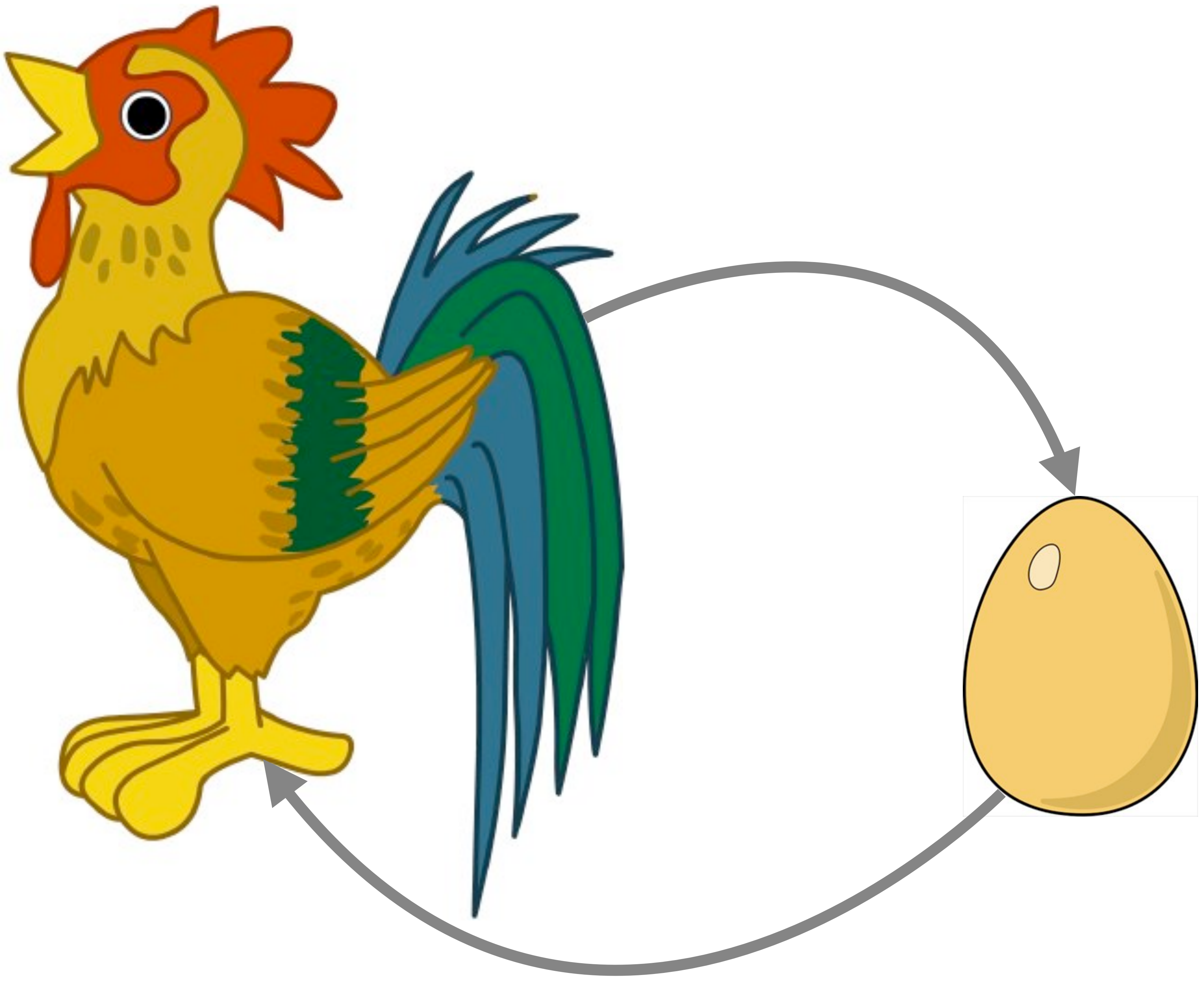




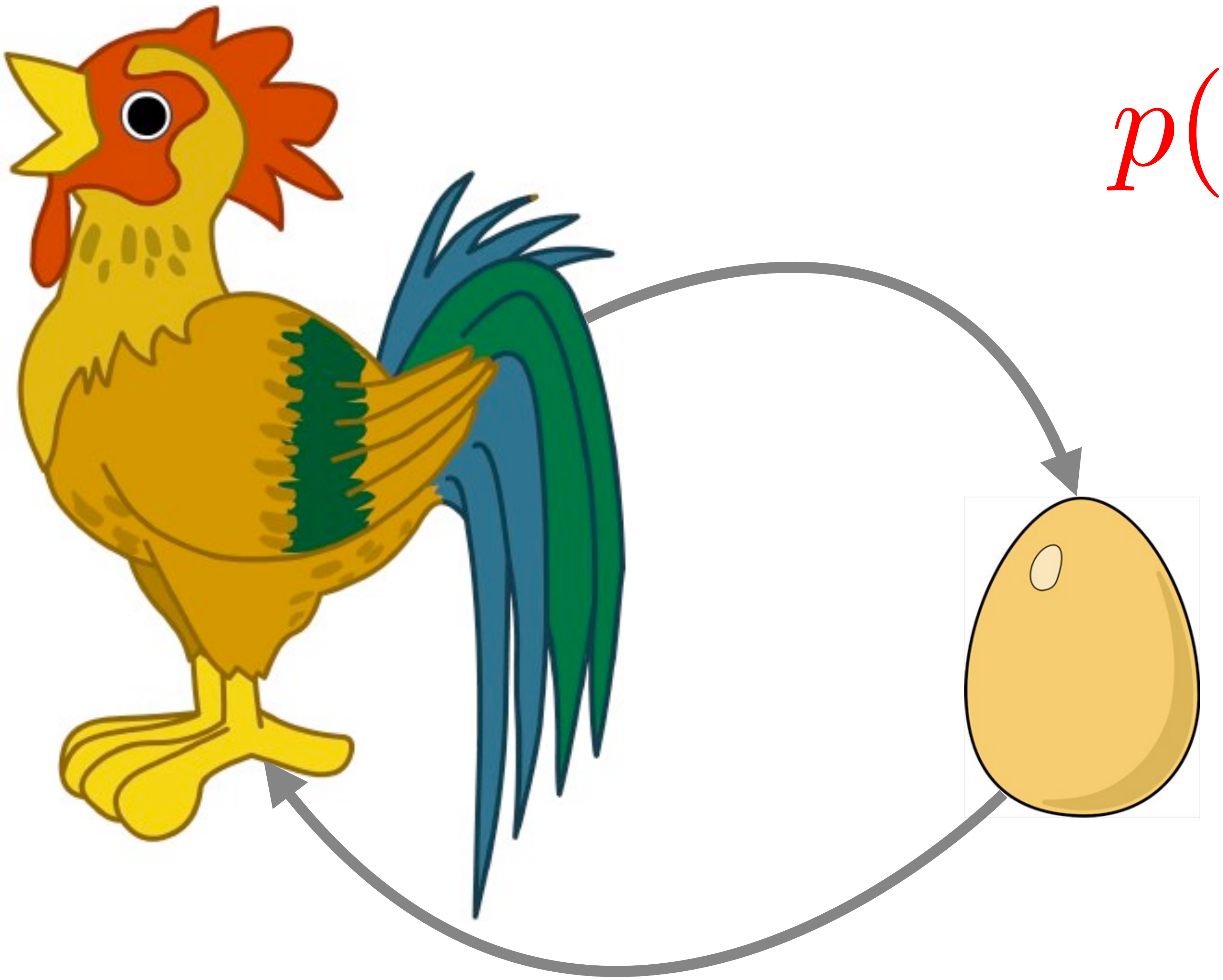
No loops allowed



No loops allowed

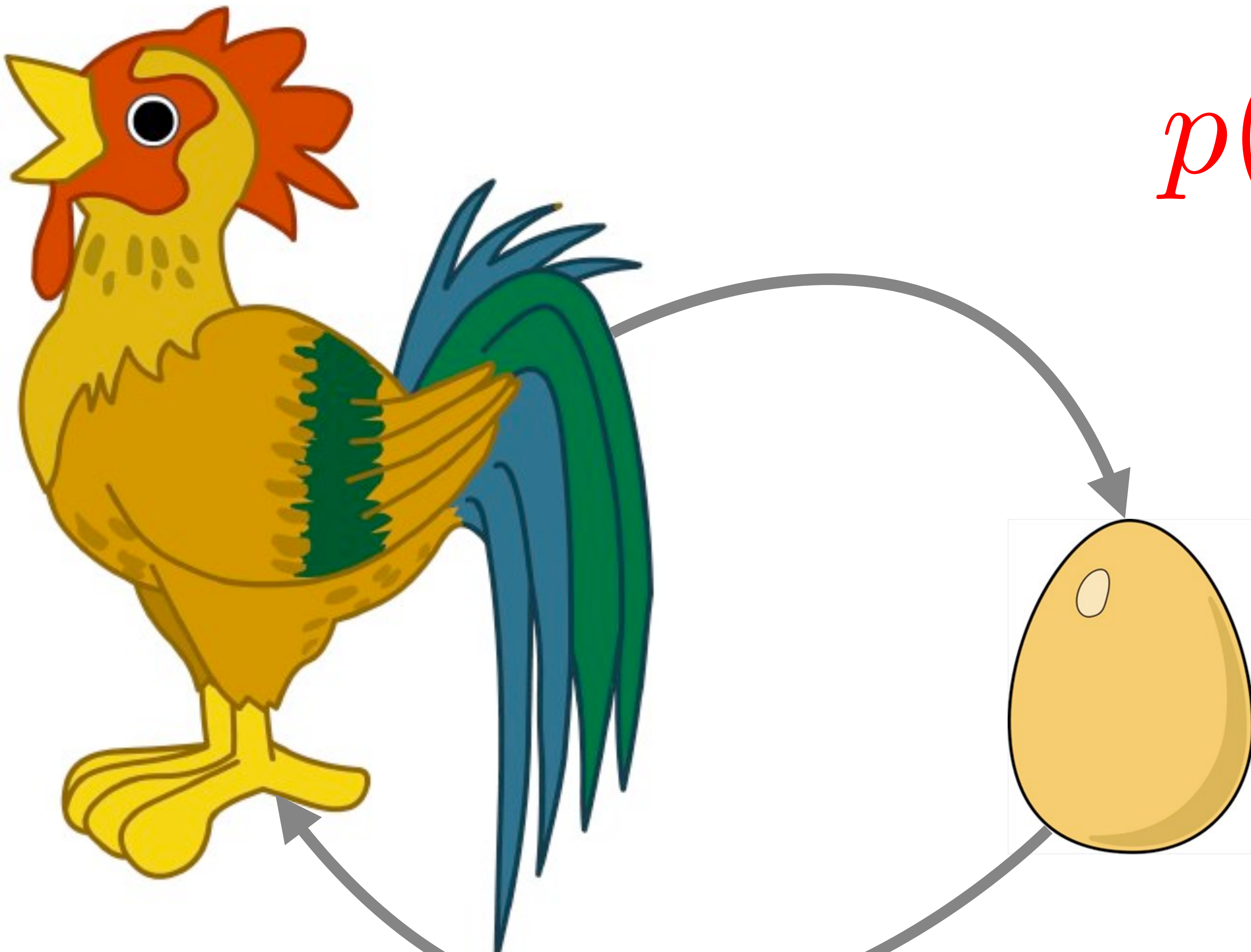


No loops allowed



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

No loops allowed

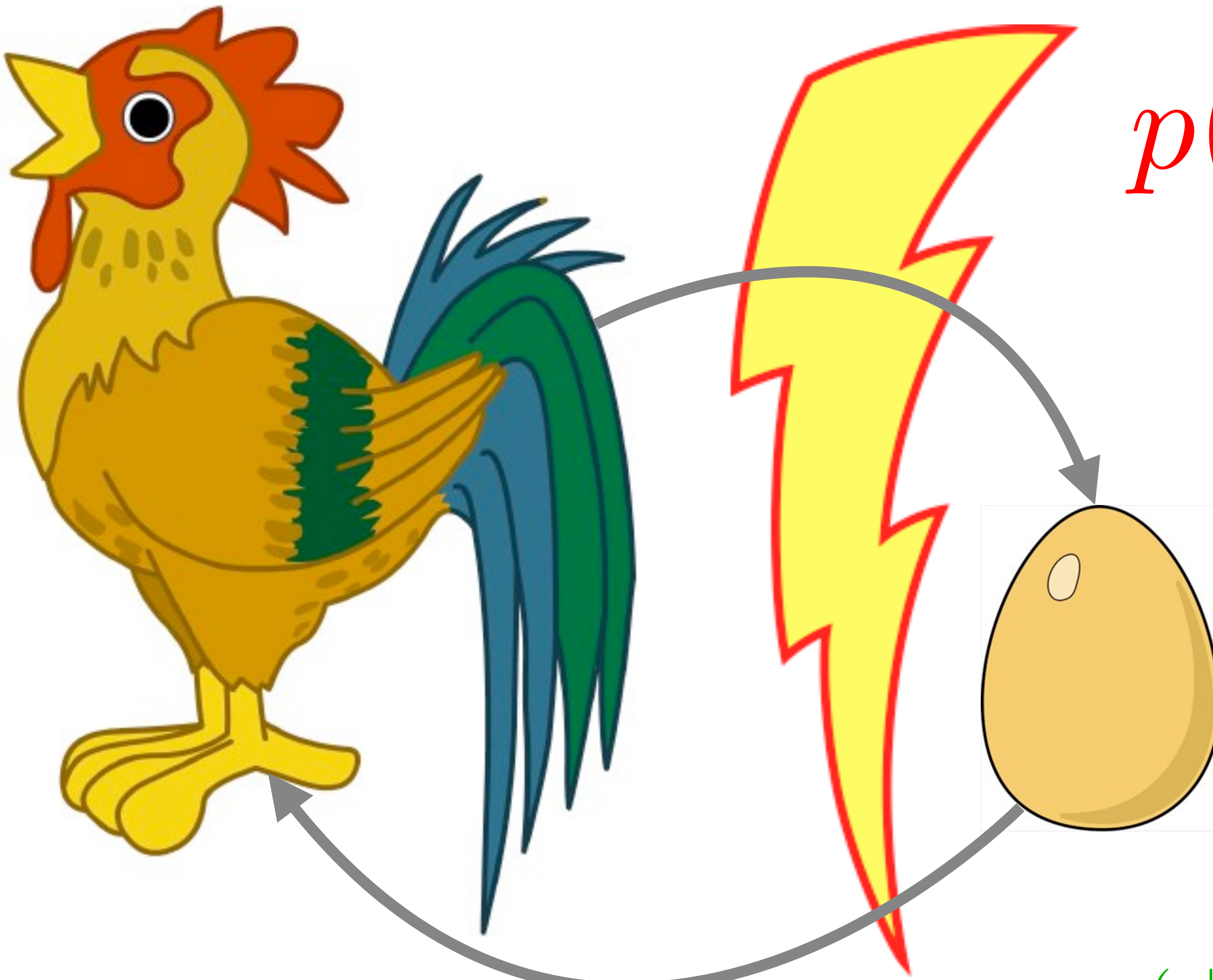


$$p(c|e)p(e|c)$$

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

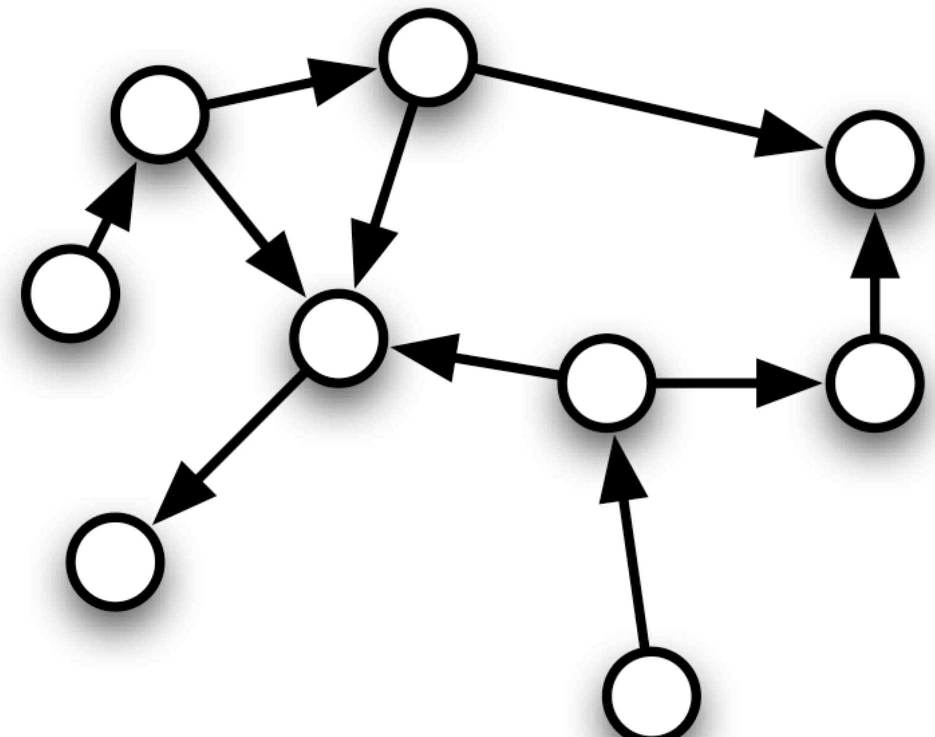
No loops allowed

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



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

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

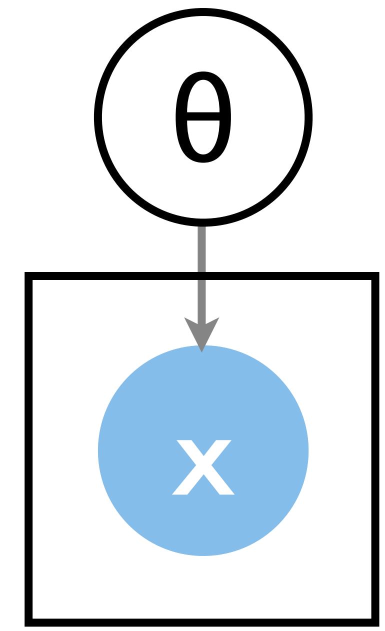
$$\log p(x|\theta) = \sum_i \log p(x_i | x_{\text{parents(i)}}, \theta)$$

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

Clustering

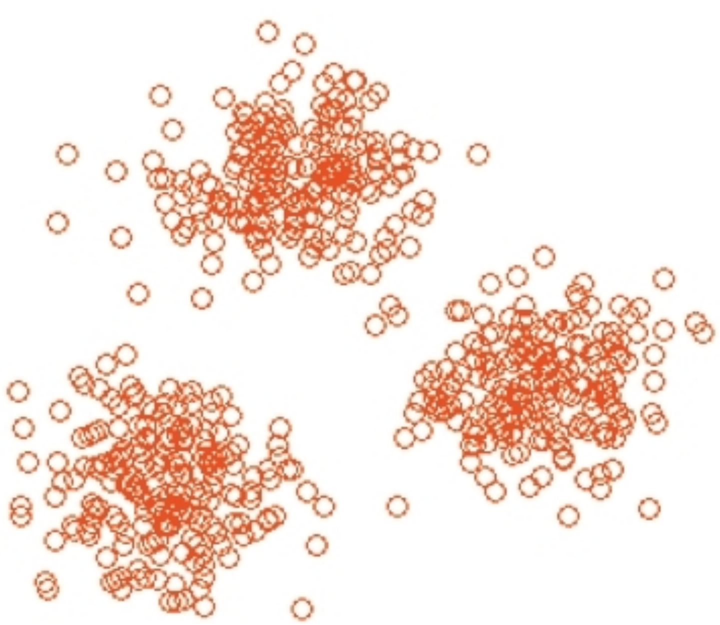
Density Estimation

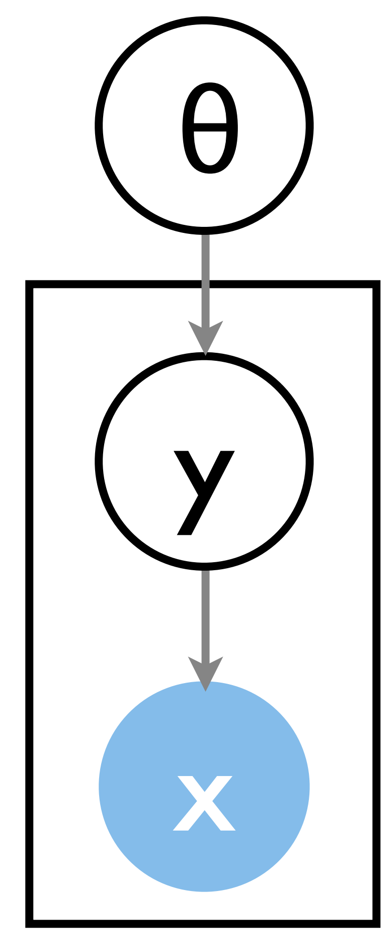
$$p(x,\theta) = p(\theta)\prod_{i=1}^{n} p(x_i|\theta)$$



Clustering

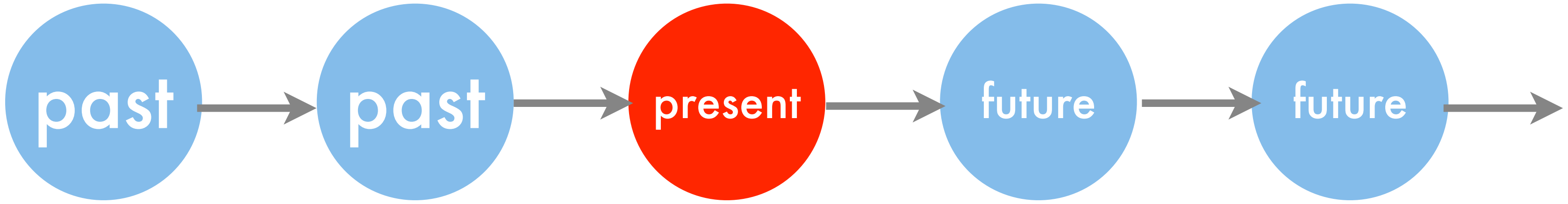
$$p(x,y,\theta) = p(\pi)\prod_{k=1}^{K} p(\theta_k)\prod_{i=1}^{n} p(y_i|\pi)p(x_i|\theta,y_i)$$





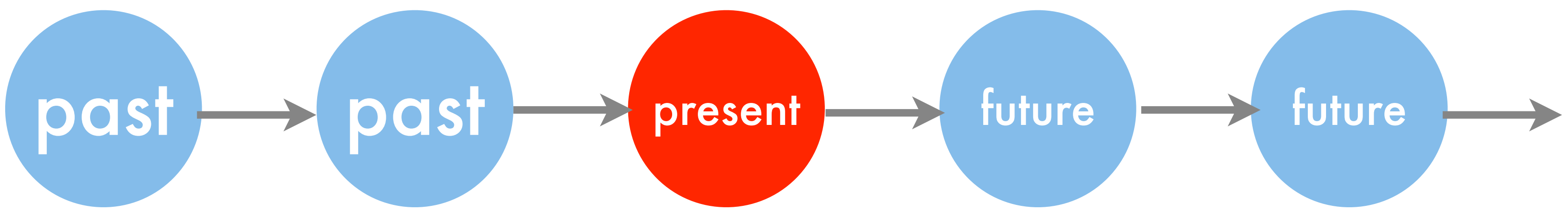
Chains

Markov Chain

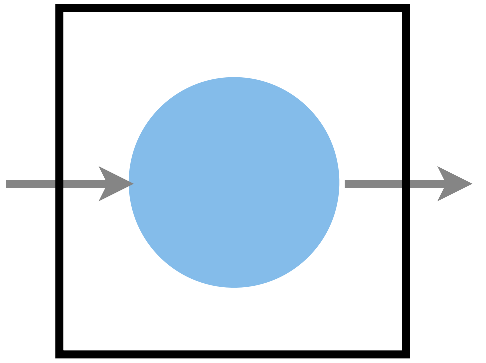


Chains

Markov Chain

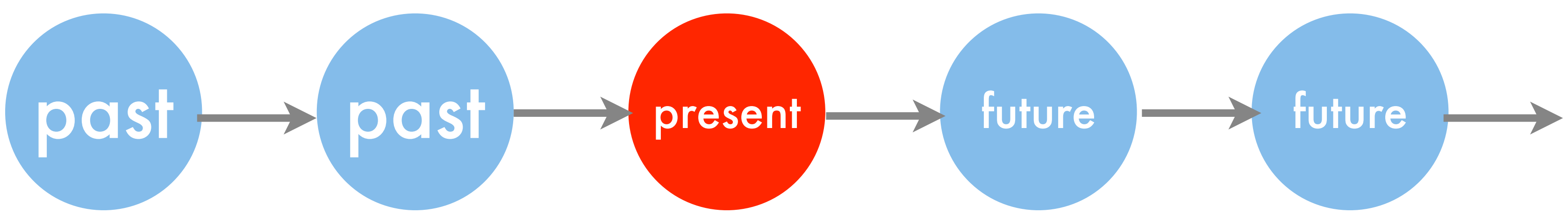


Plate

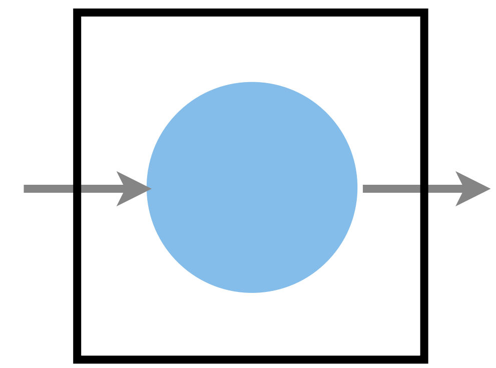


Chains

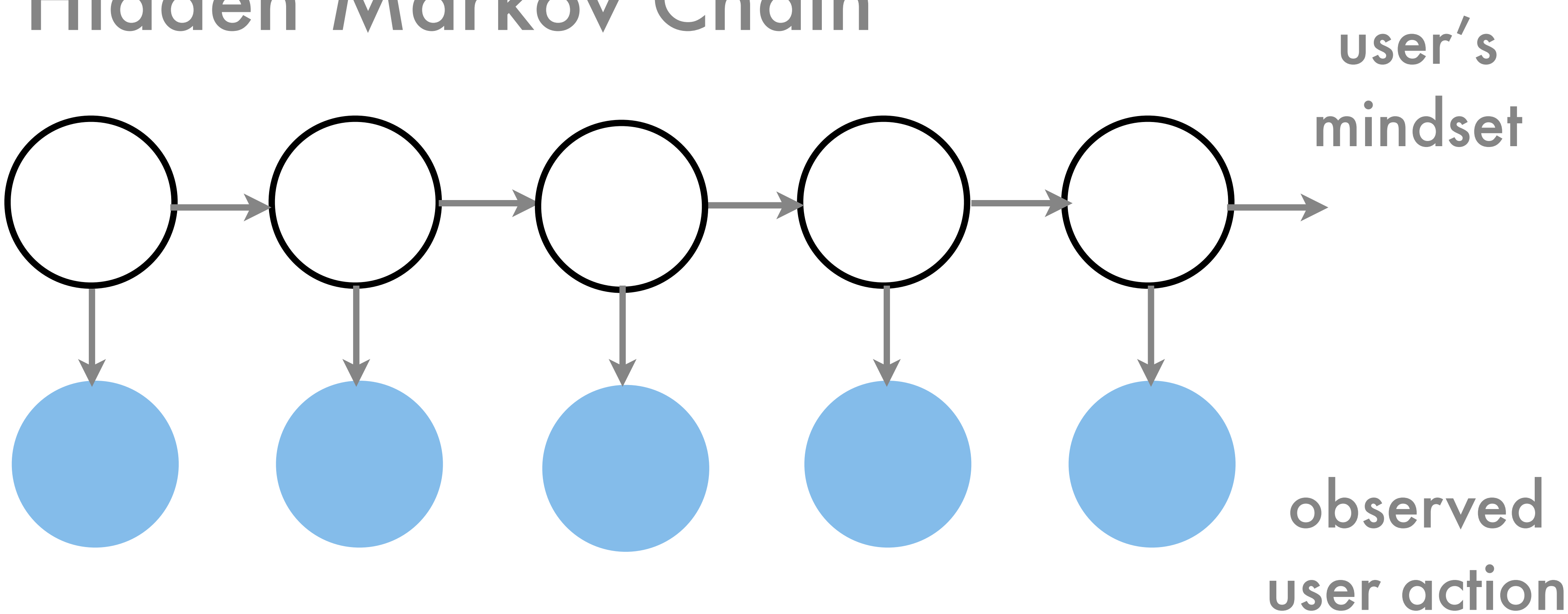
Markov Chain



Plate

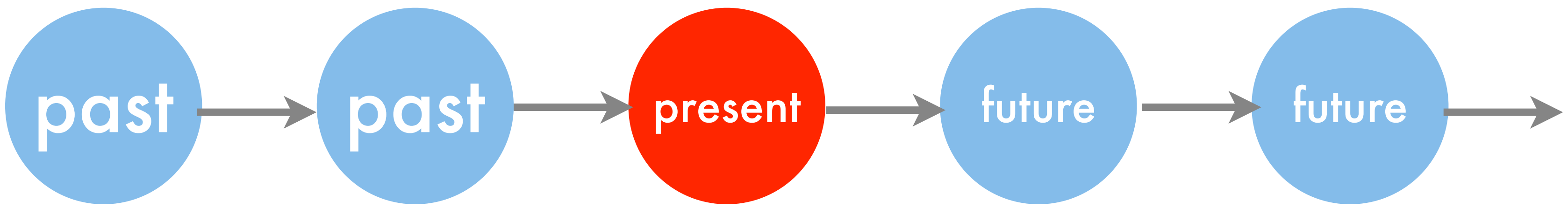


Hidden Markov Chain

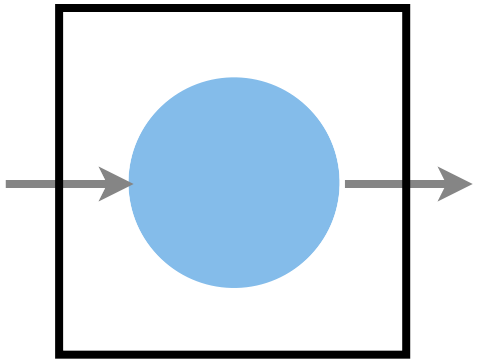


Chains

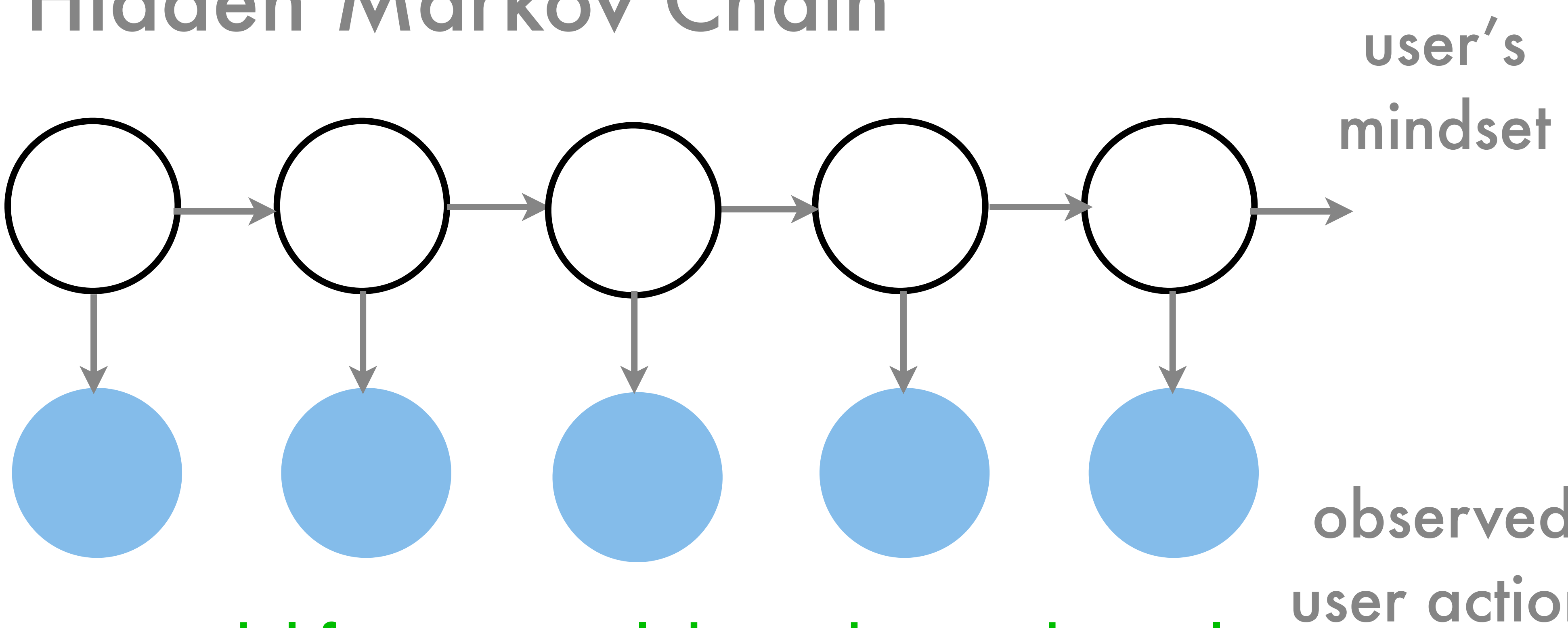
Markov Chain



Plate



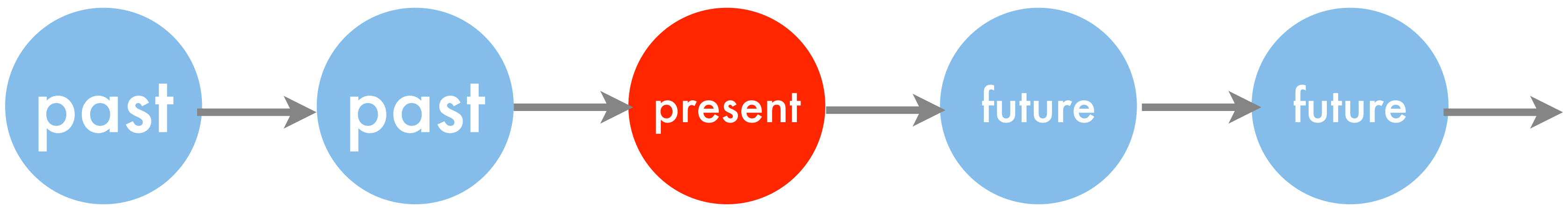
Hidden Markov Chain



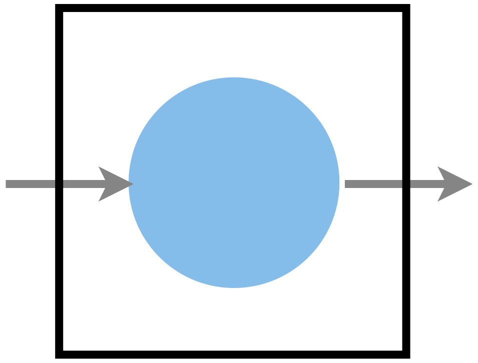
user model for traversal through search results

Chains

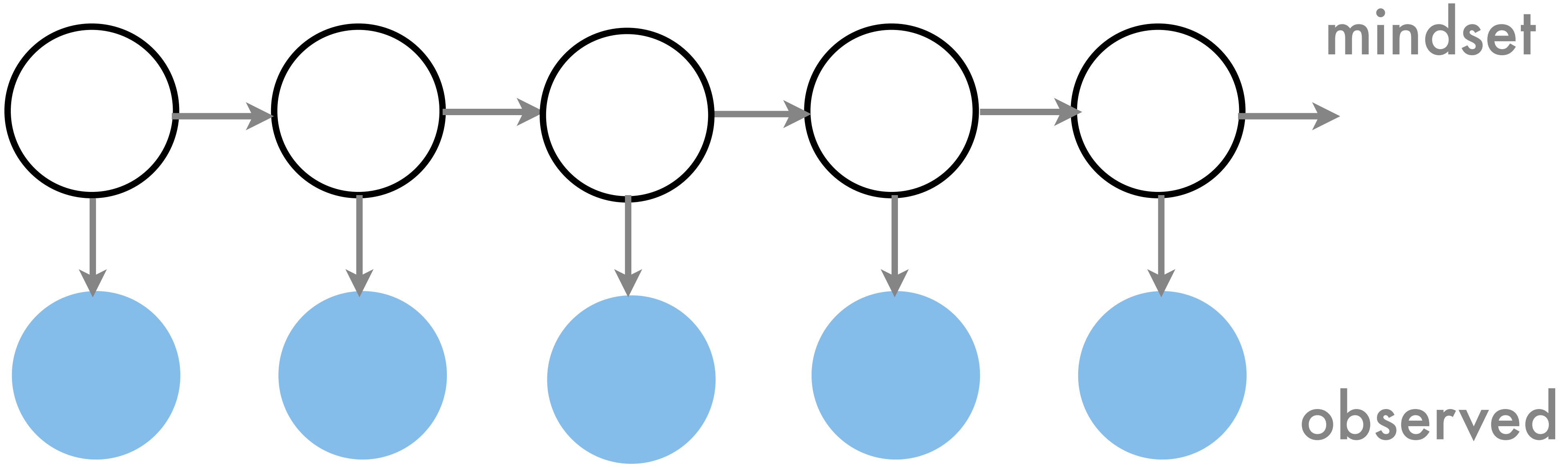
Markov Chain

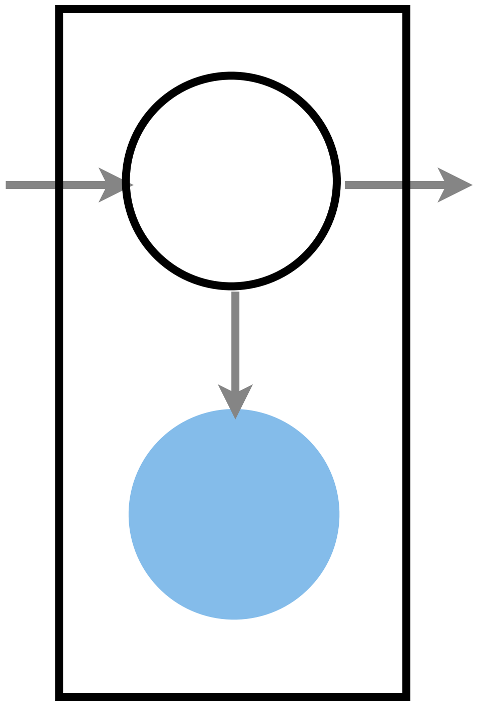


Plate



Hidden Markov Chain





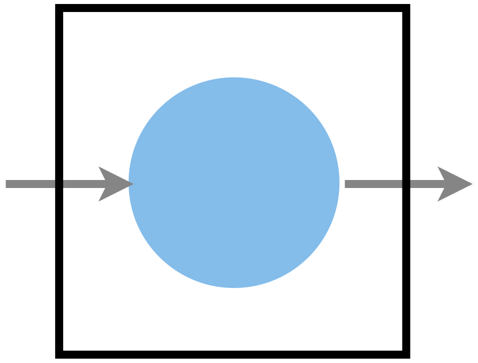
user model for traversal through search results

Chains

Markov Chain

Plate

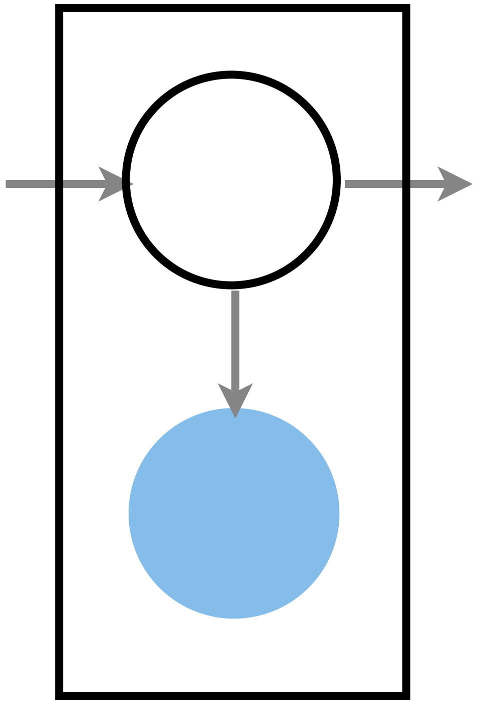
$$p(x;\theta) = p(x_0;\theta)\prod_{i=1}^{n-1} p(x_{i+1}|x_i;\theta)$$



Hidden Markov Chain

user's mindset

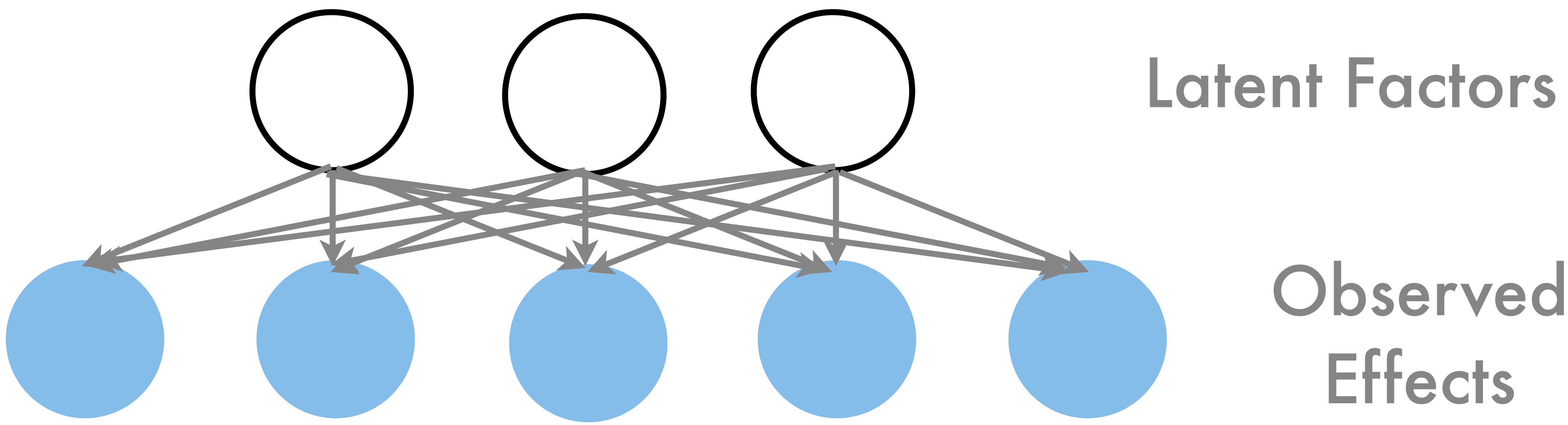
$$p(x,y;\theta) = p(x_0;\theta)\prod_{i=1}^{n-1} p(x_{i+1}|x_i;\theta)\prod_{i=1}^{n} p(y_i|x_i)$$



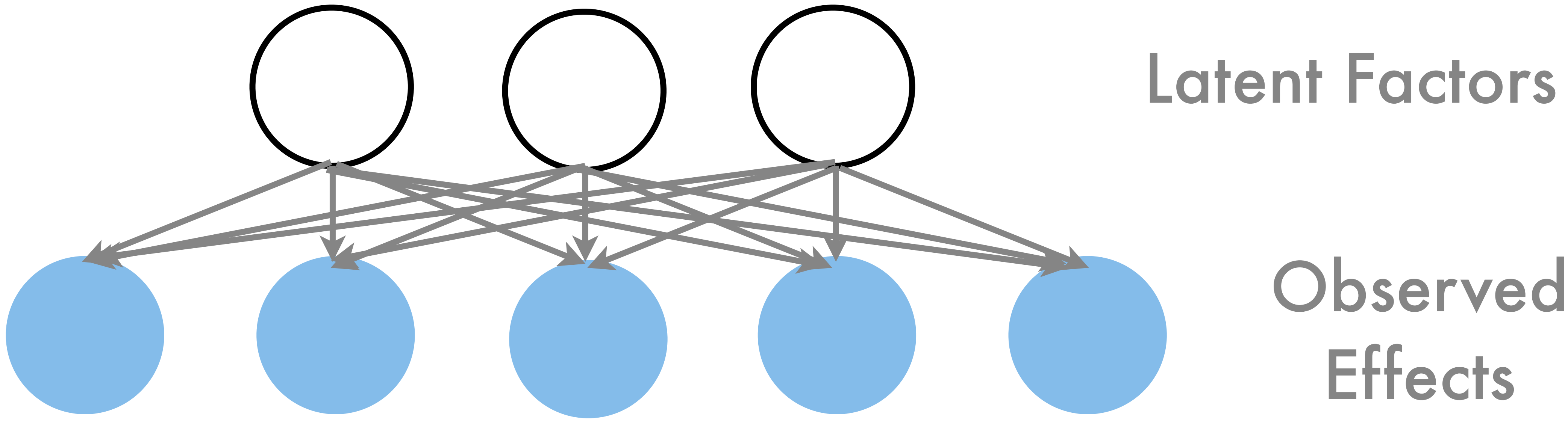
observed user action

user model for traversal through search results

Factor Graphs

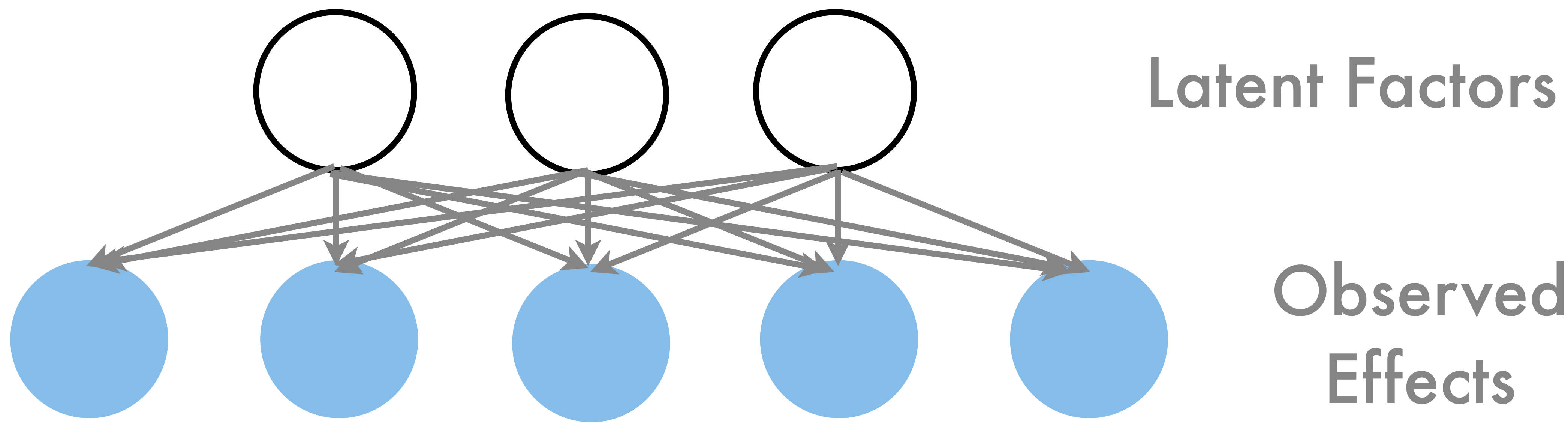


Factor Graphs



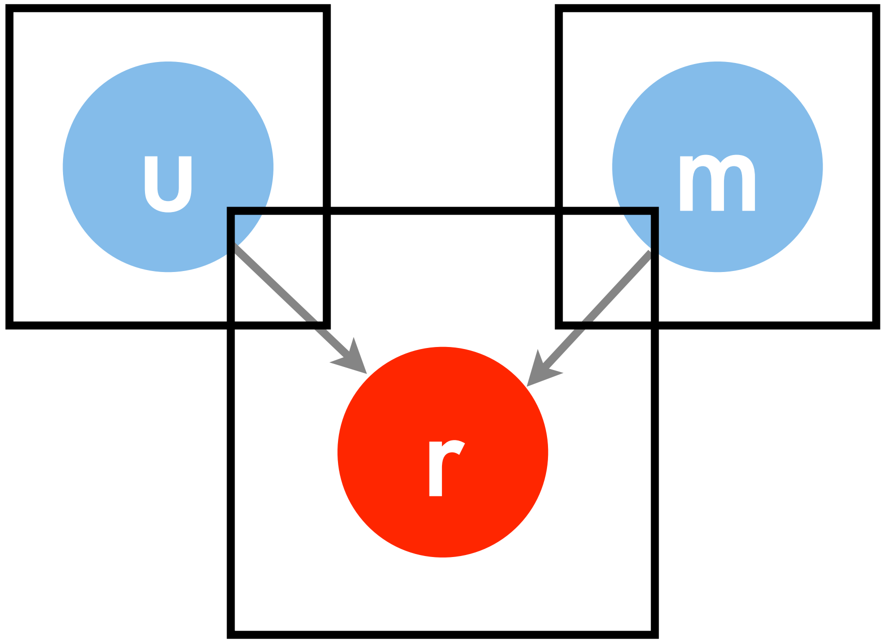
- Observed effects
  Click behavior, queries, watched news, emails

Factor Graphs

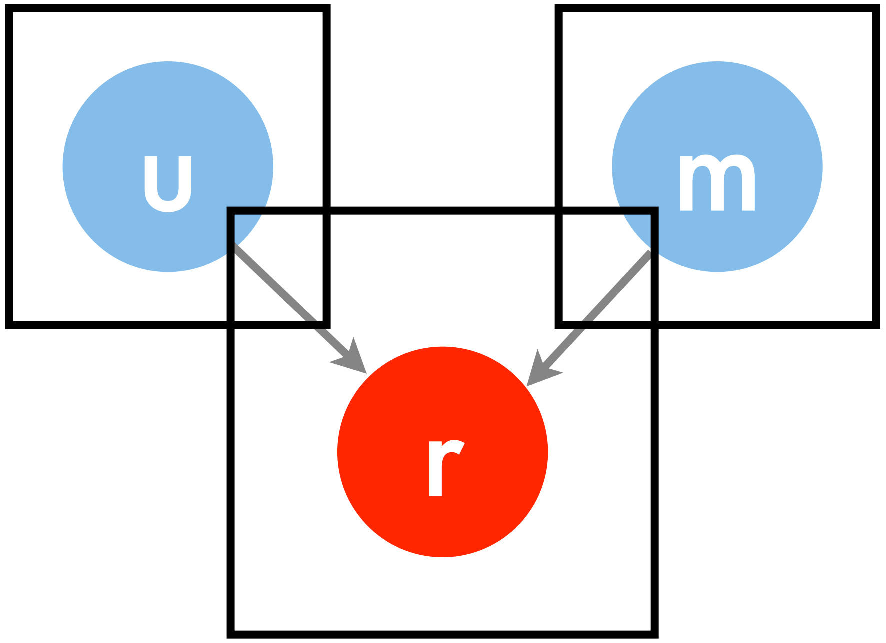


- Observed effects
  Click behavior, queries, watched news, emails
- Latent factors
  User profile, news content, hot keywords, social connectivity graph, events

Recommender Systems

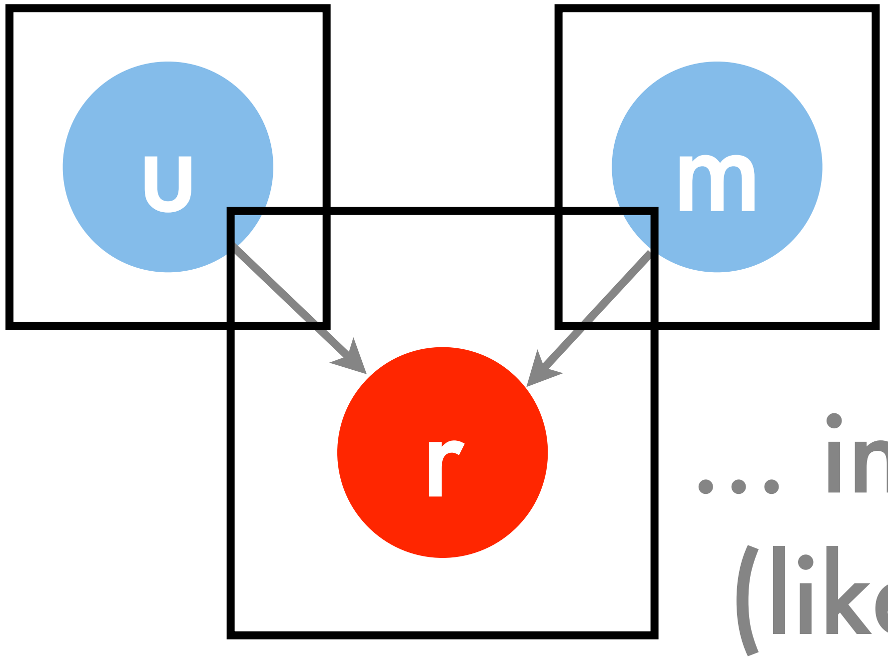


Recommender Systems



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

Recommender Systems

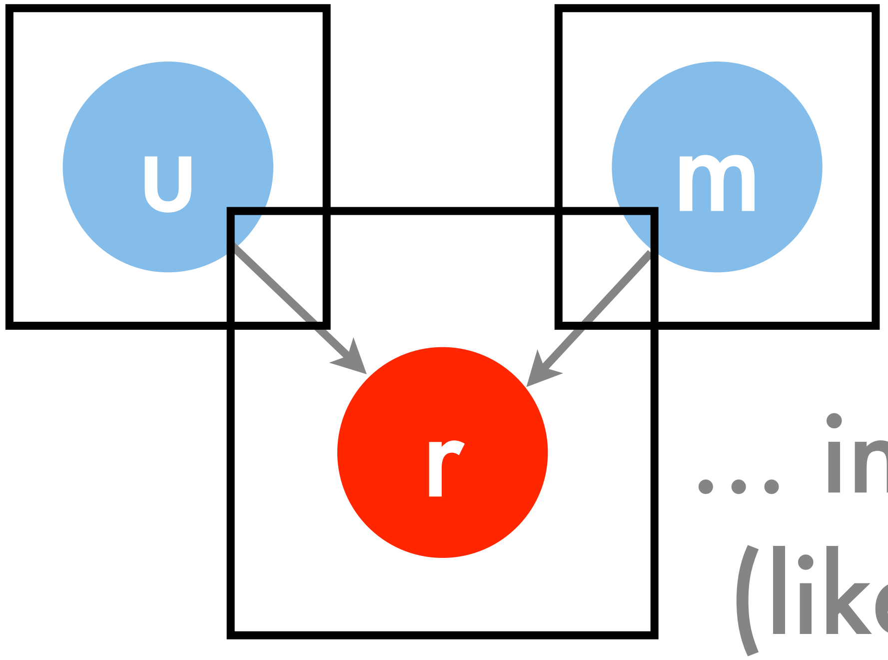


... intersecting plates ...
(like nested for loops)

- 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

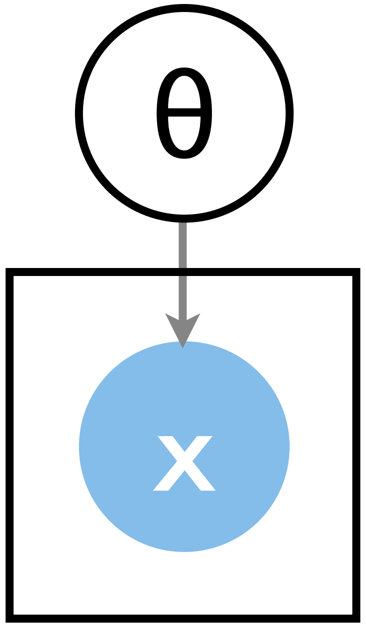
Part 3 - Clustering & Topic Models

Inference Algorithms

Clustering

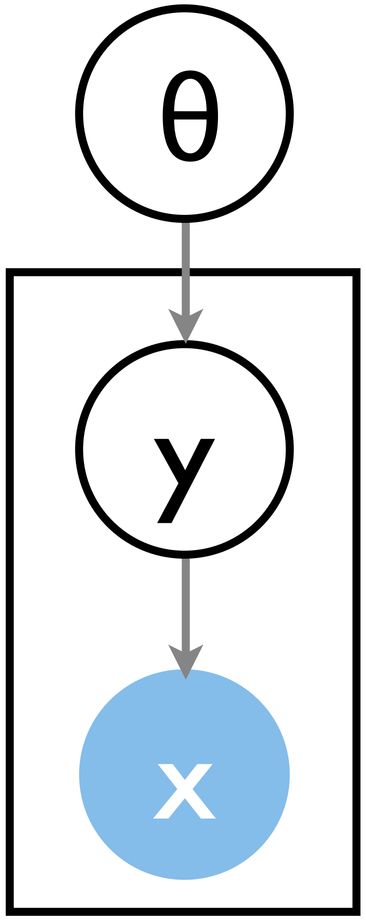
Density Estimation

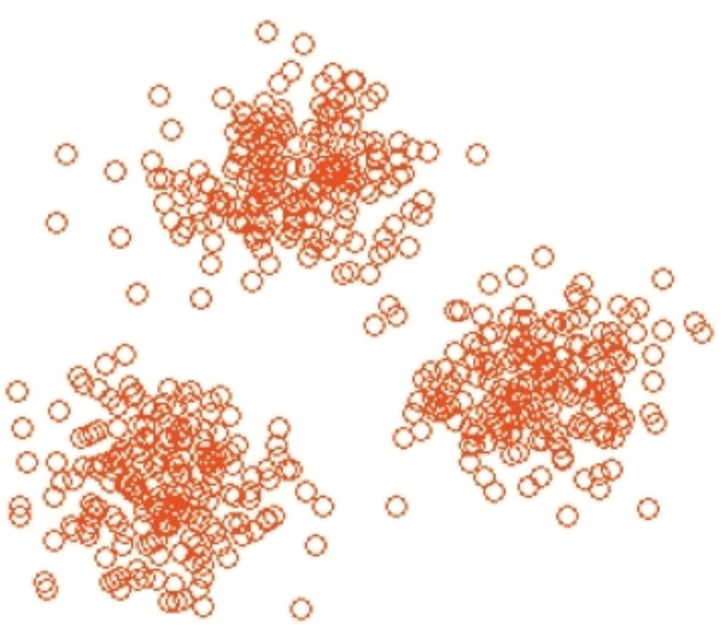
$$p(x, \theta) = p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$



Clustering

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



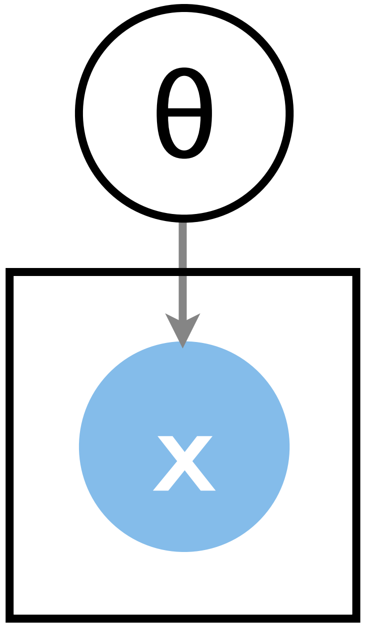


Clustering

Density Estimation

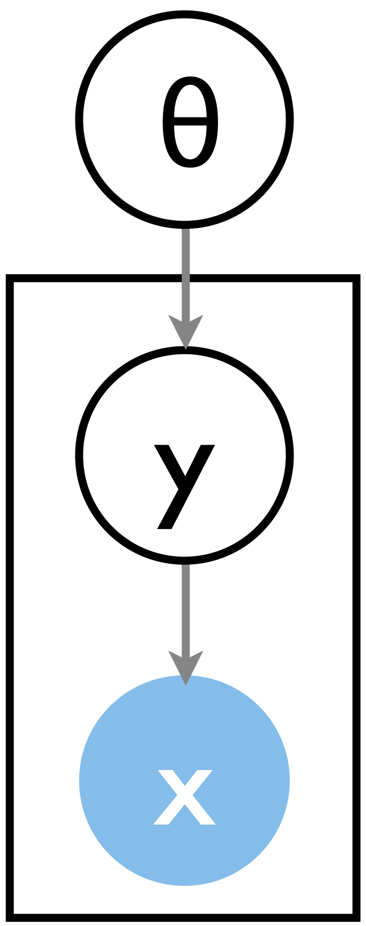
$$p(x, \theta) = p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

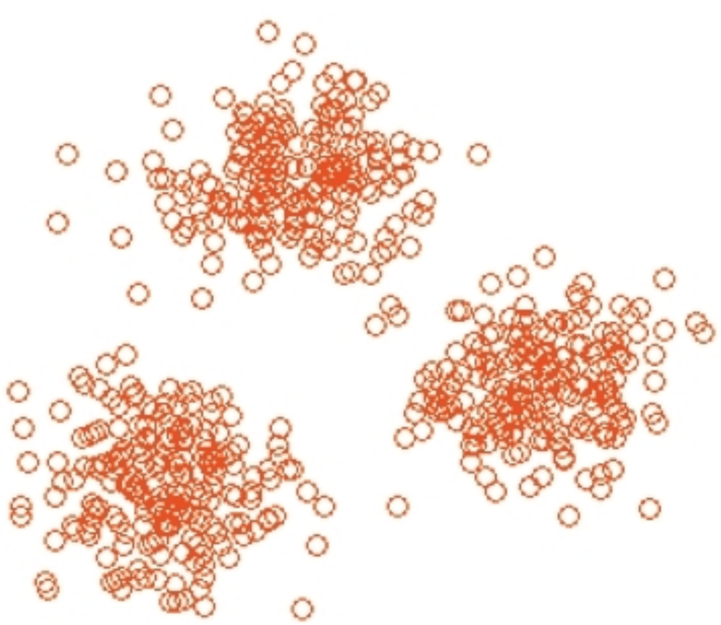
find θ



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

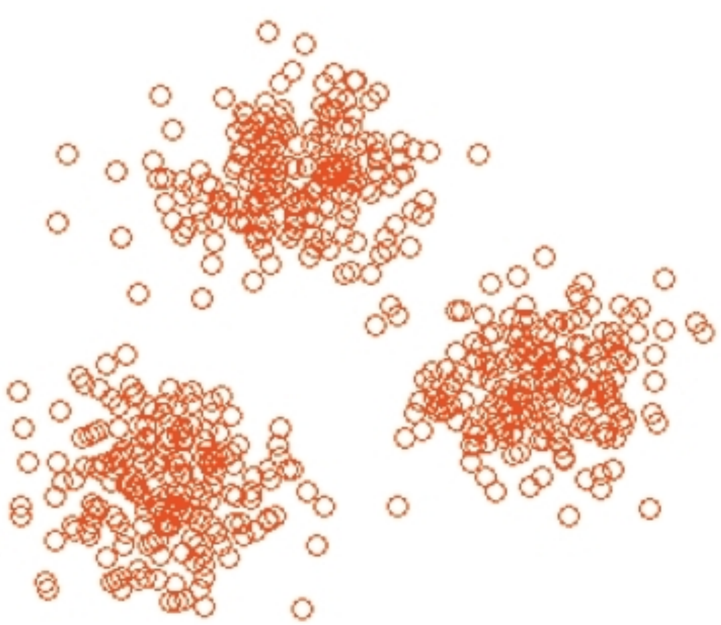
log-concave

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

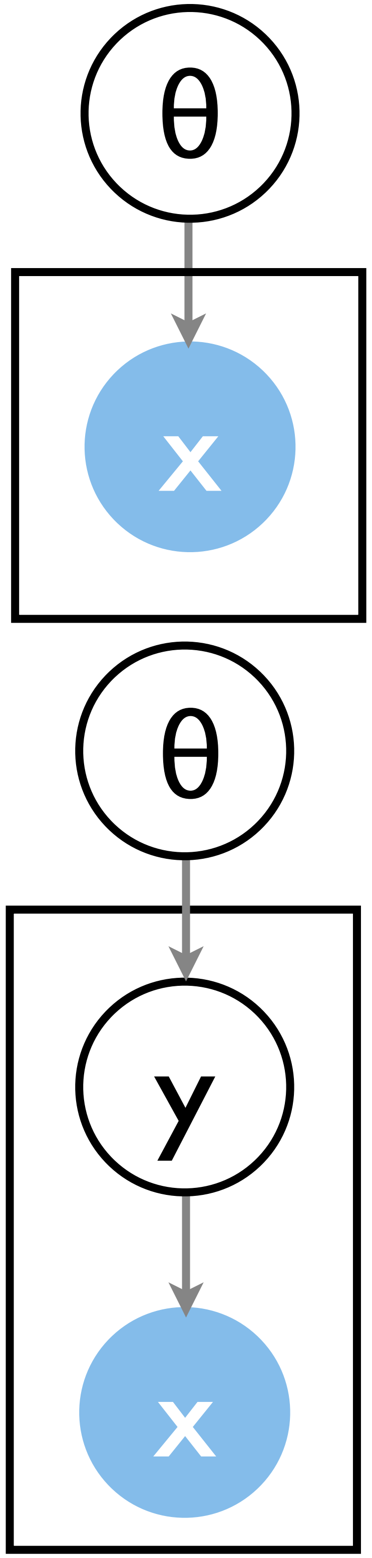
find θ

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

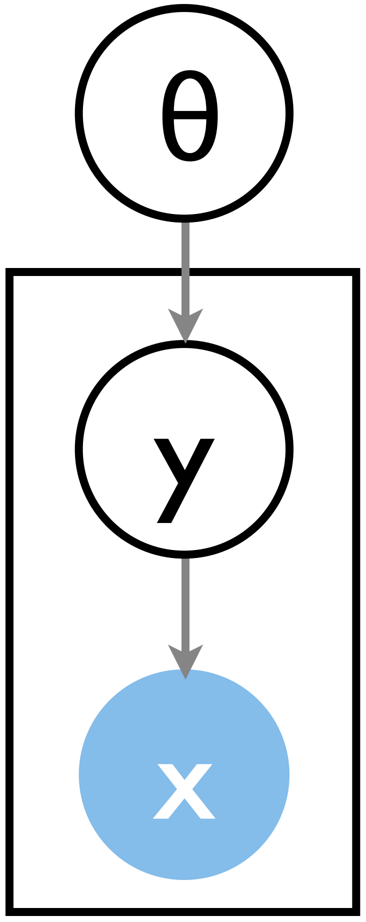
- Optimization problem

$$\underset{\theta}{\text{maximize}} \sum_{y} p(x, y, \theta)$$

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

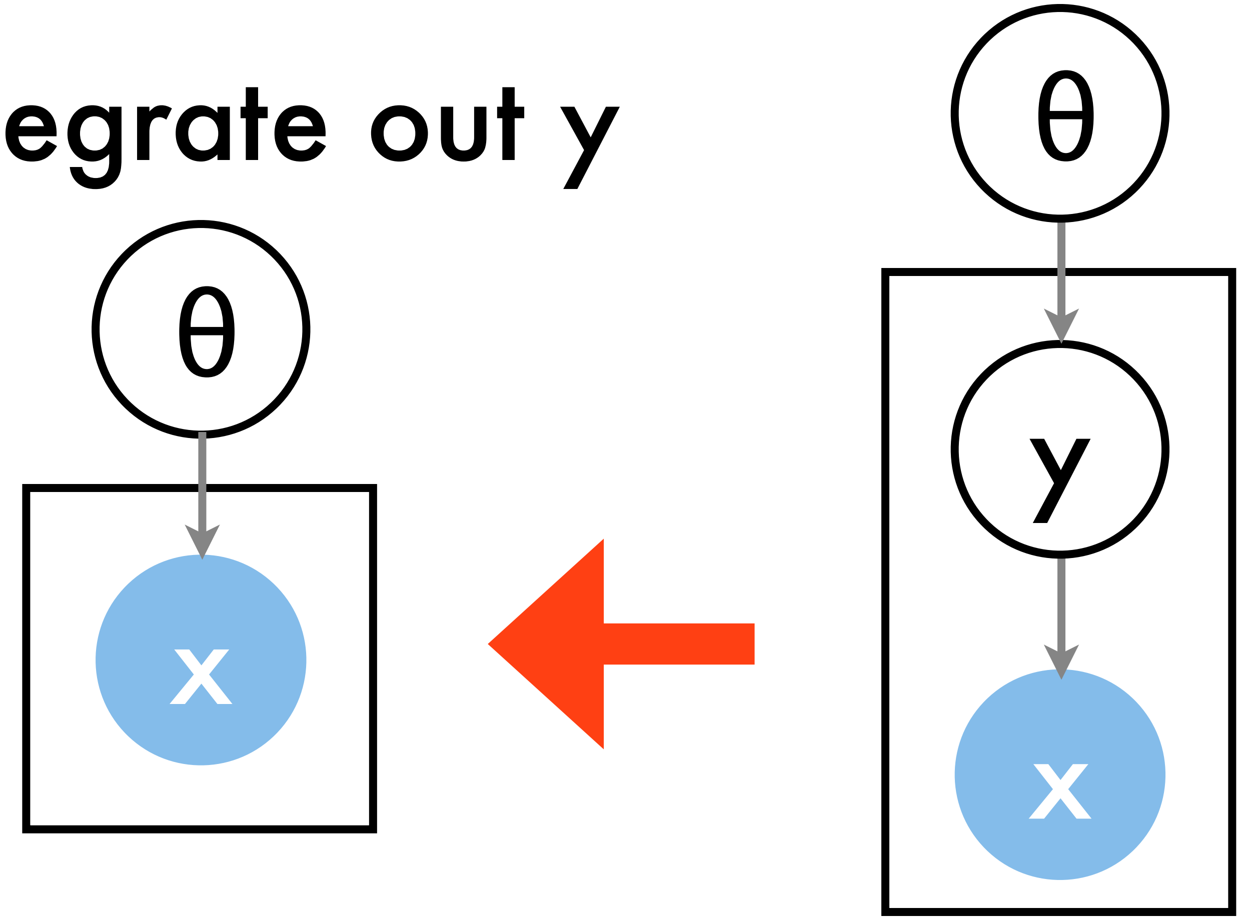
- Options
  - Direct nonconvex optimization (e.g. BFGS)
  - Sampling (draw from the joint distribution)
  - Variational approximation (concave lower bounds aka EM algorithm)

Clustering



Clustering

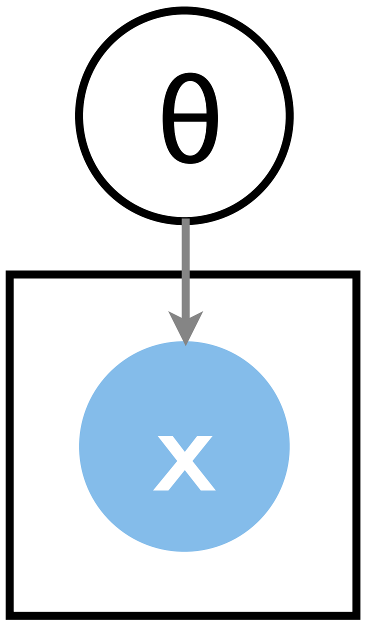
- Integrate out y



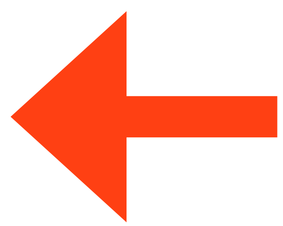
- Nonconvex optimization problem
- EM algorithm

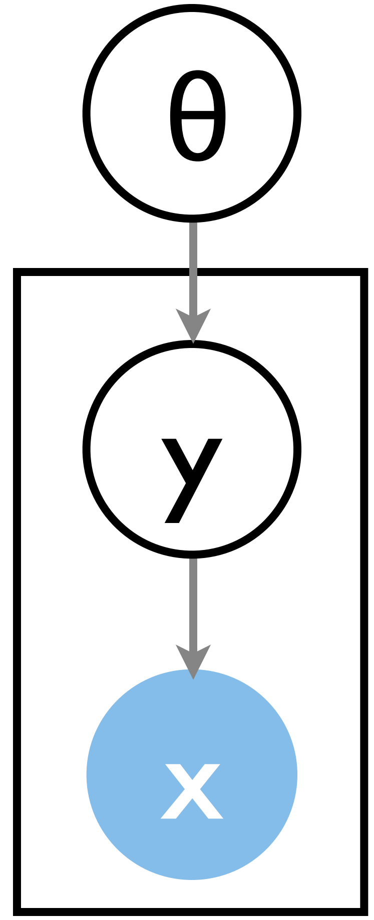
Clustering

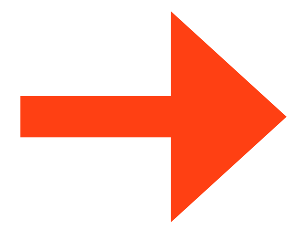
- Integrate out y



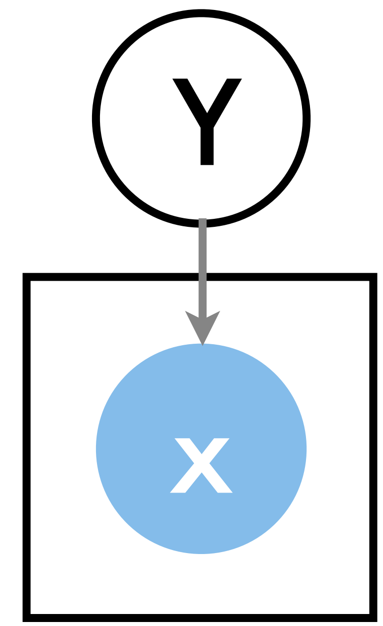
- Nonconvex optimization problem
- EM algorithm







- Integrate out θ



- Y is coupled
- Sampling
- Collapsed p

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

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,.)

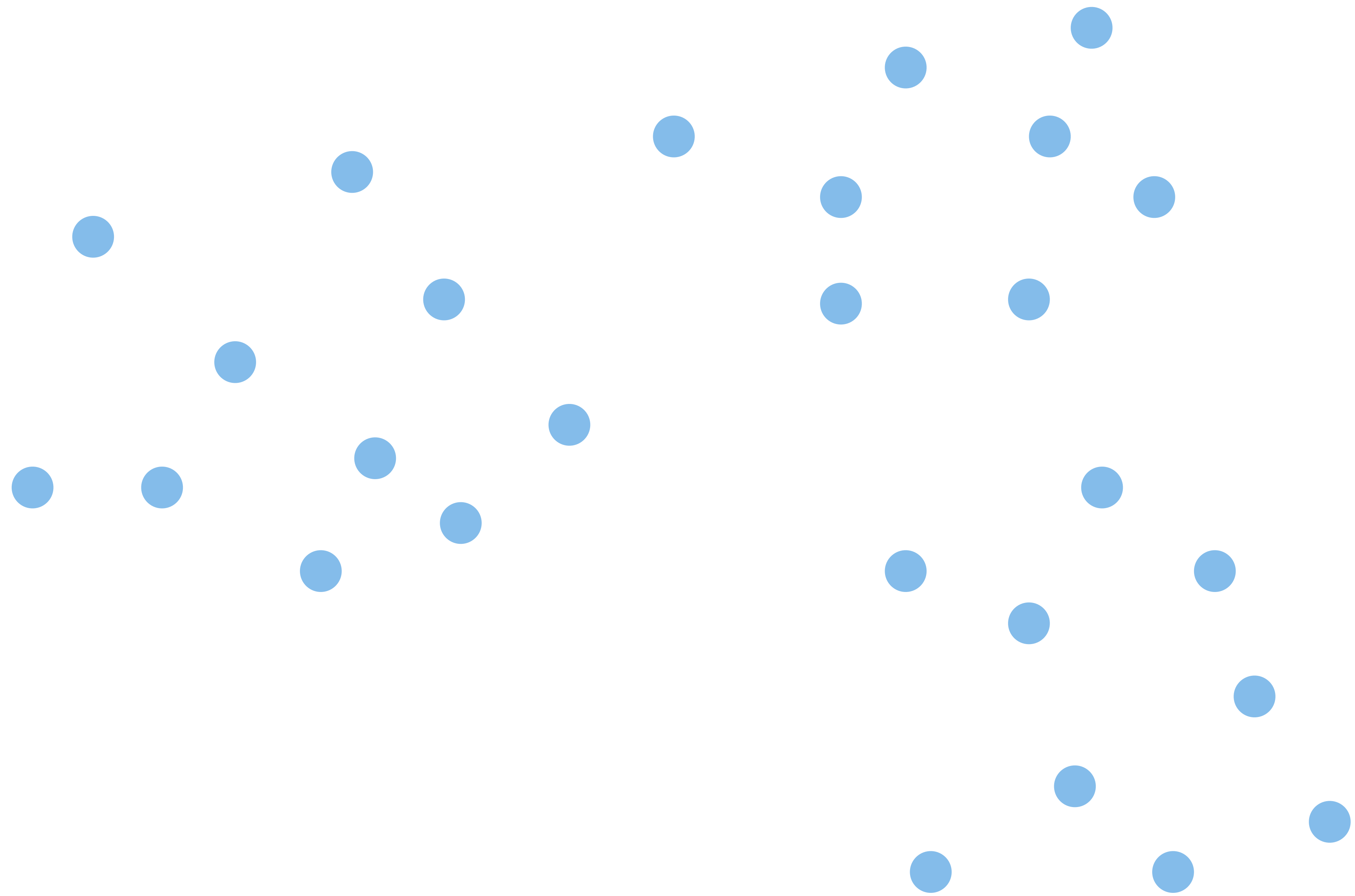
Gibbs sampling

- 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

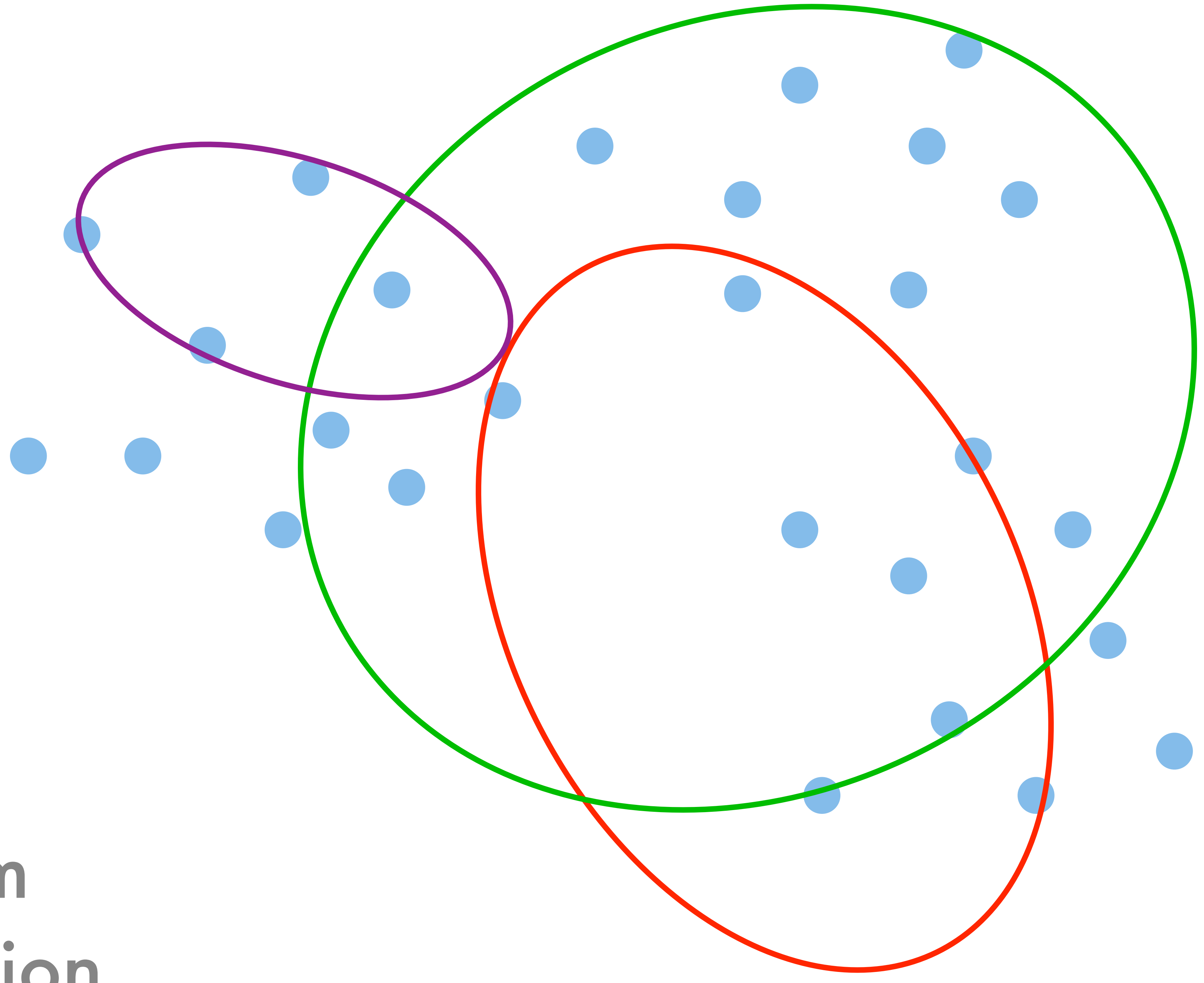
		
	0.45	0.05
	0.05	0.45

(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

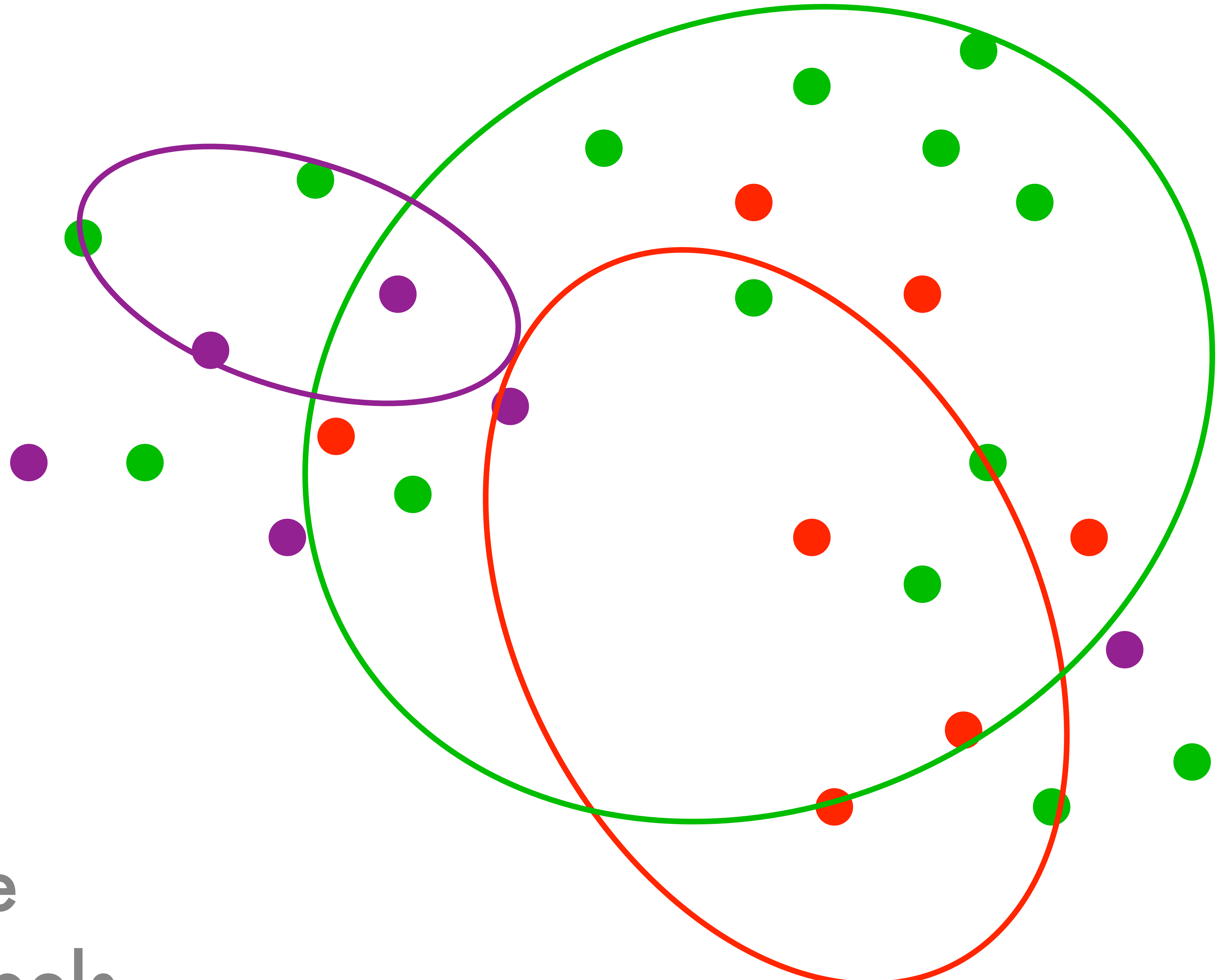


Gibbs sampling for clustering



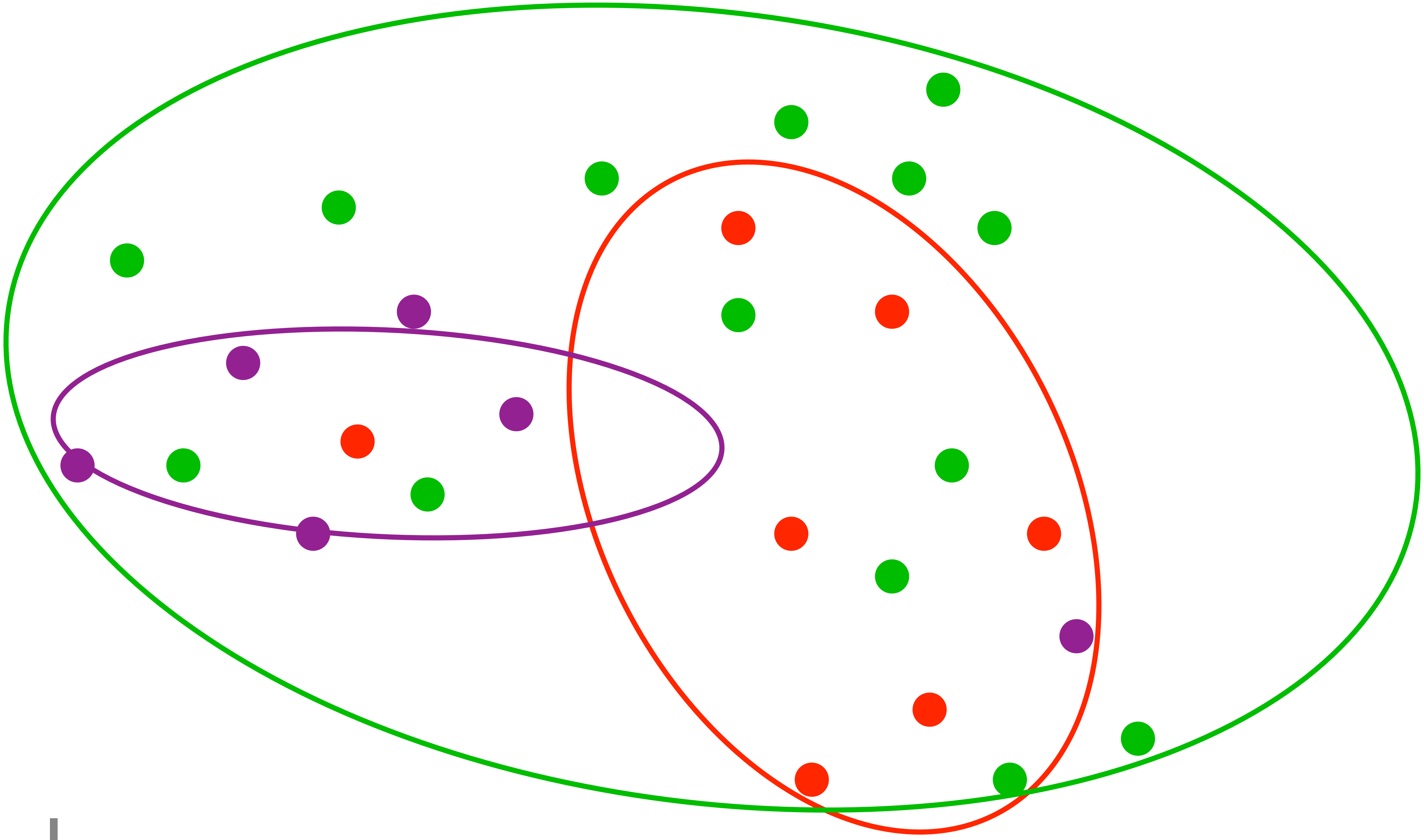
random
initialization

Gibbs sampling for clustering



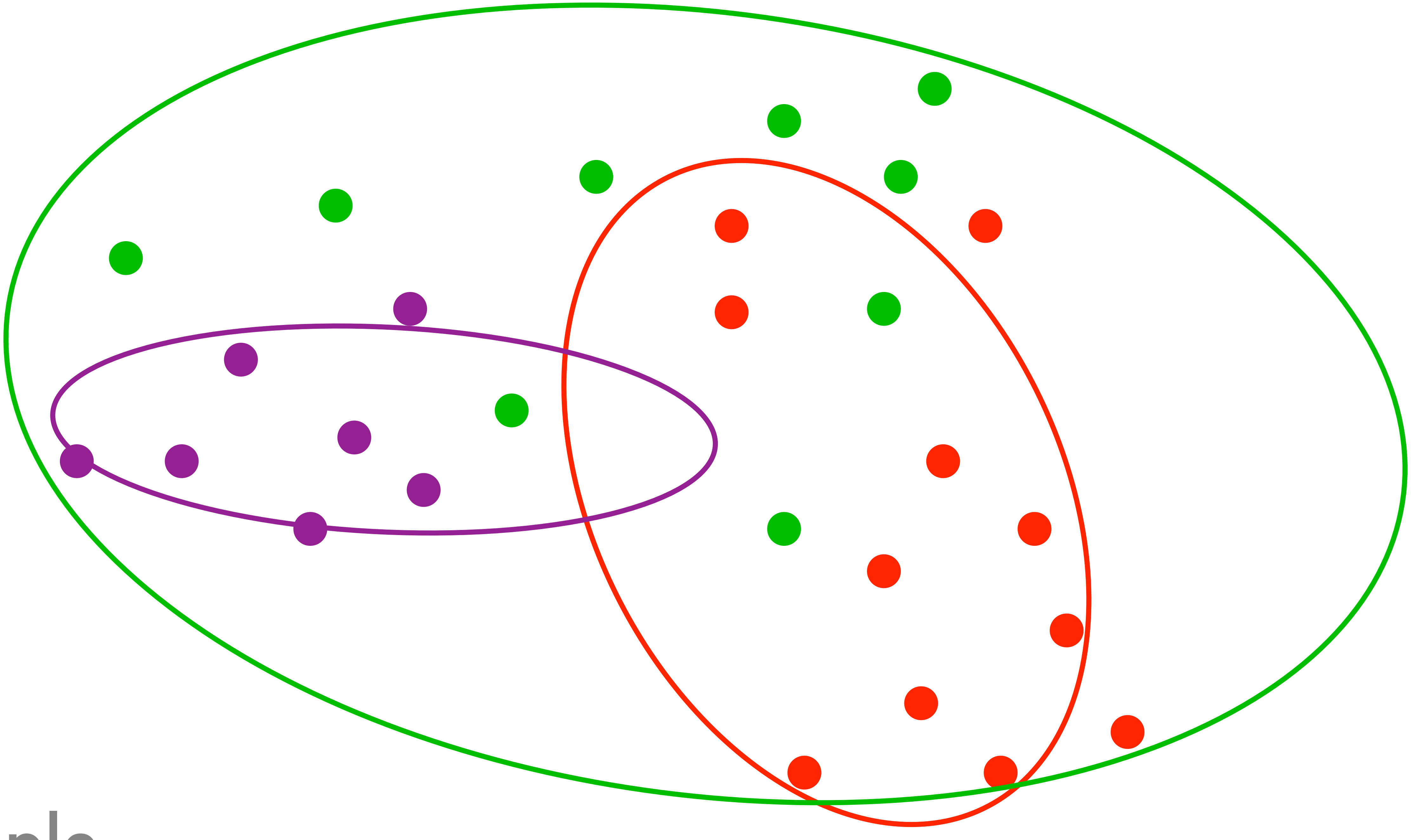
sample
cluster labels

Gibbs sampling for clustering



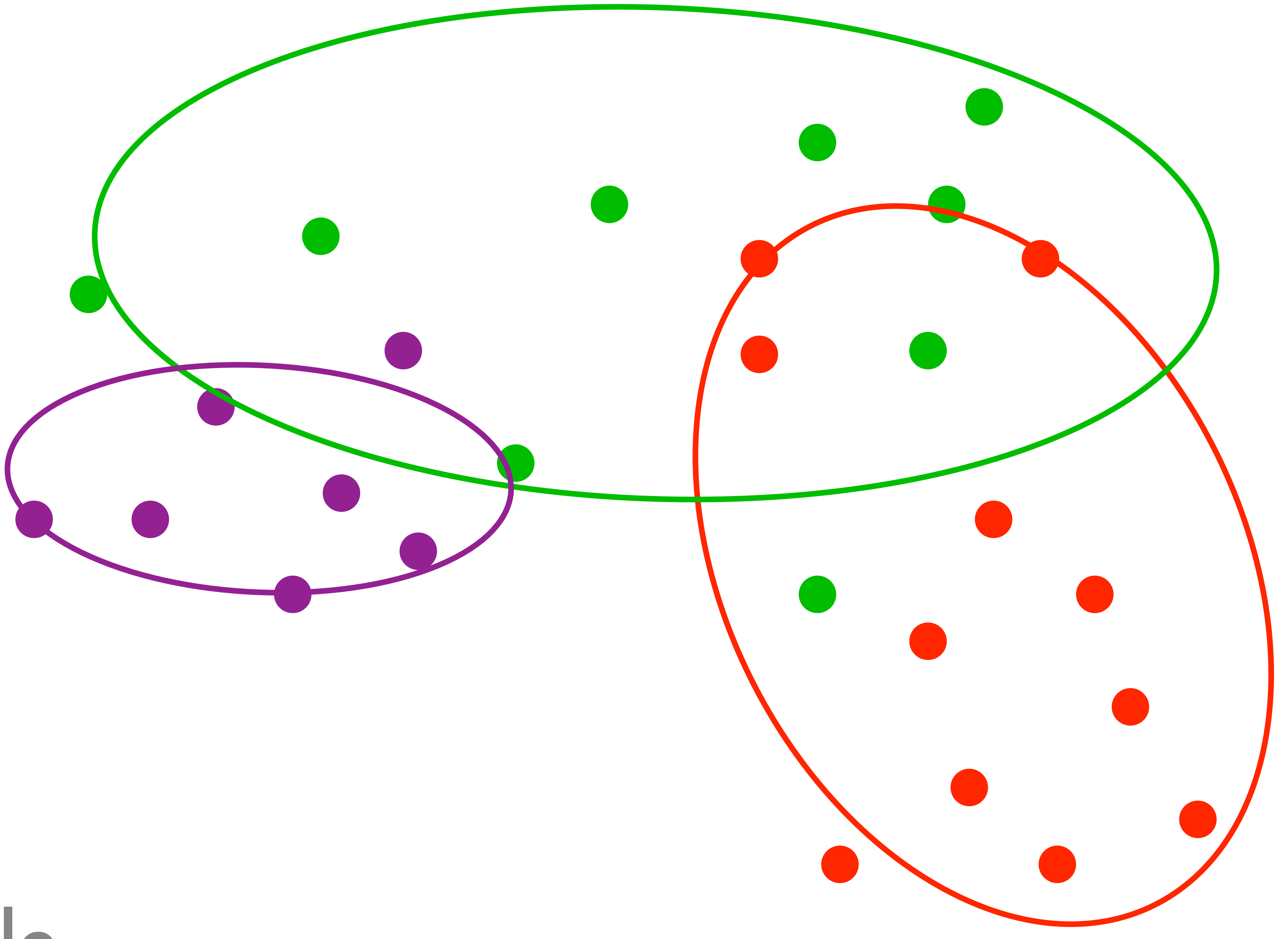
resample
cluster model

Gibbs sampling for clustering



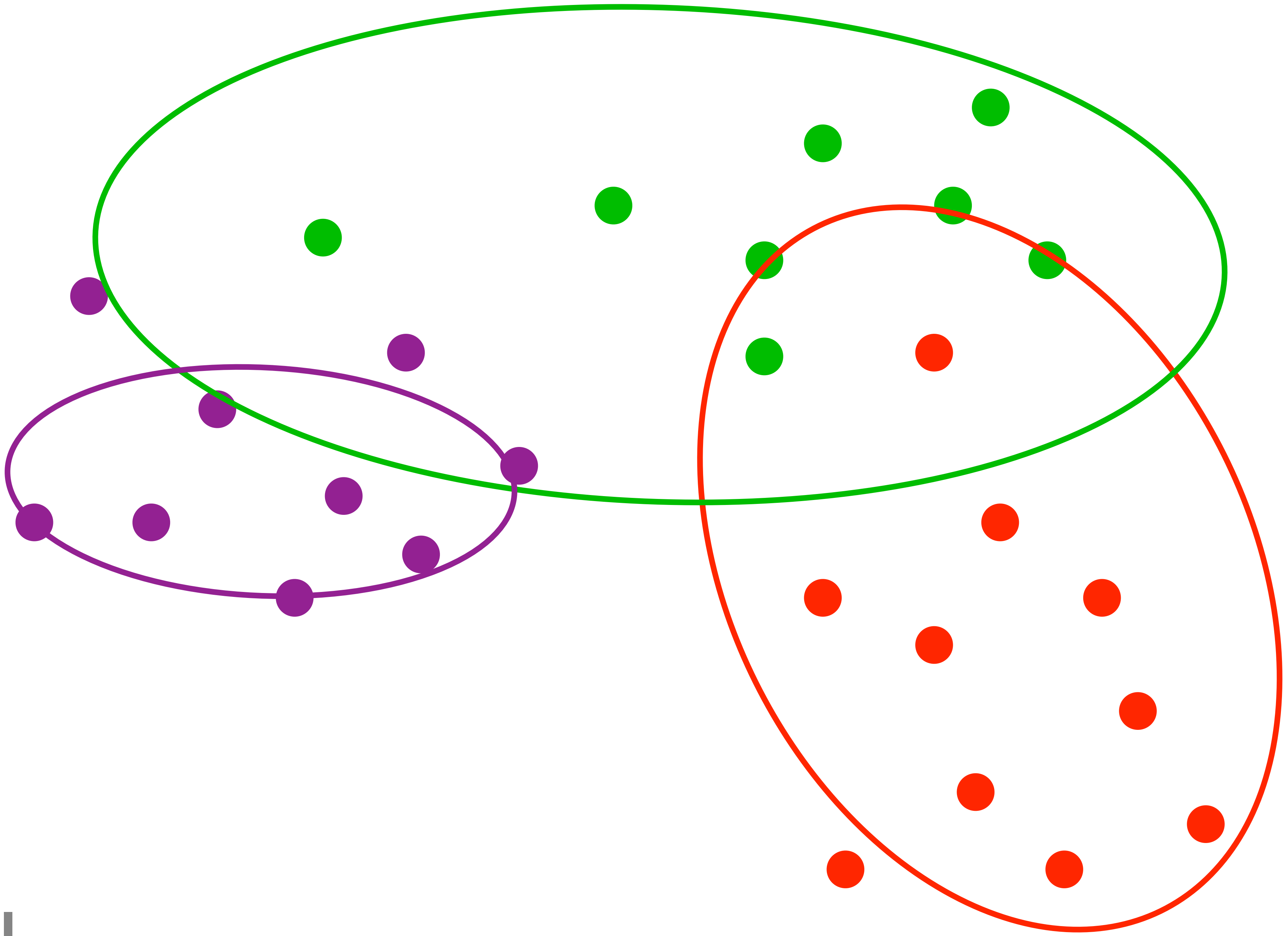
resample
cluster labels

Gibbs sampling for clustering



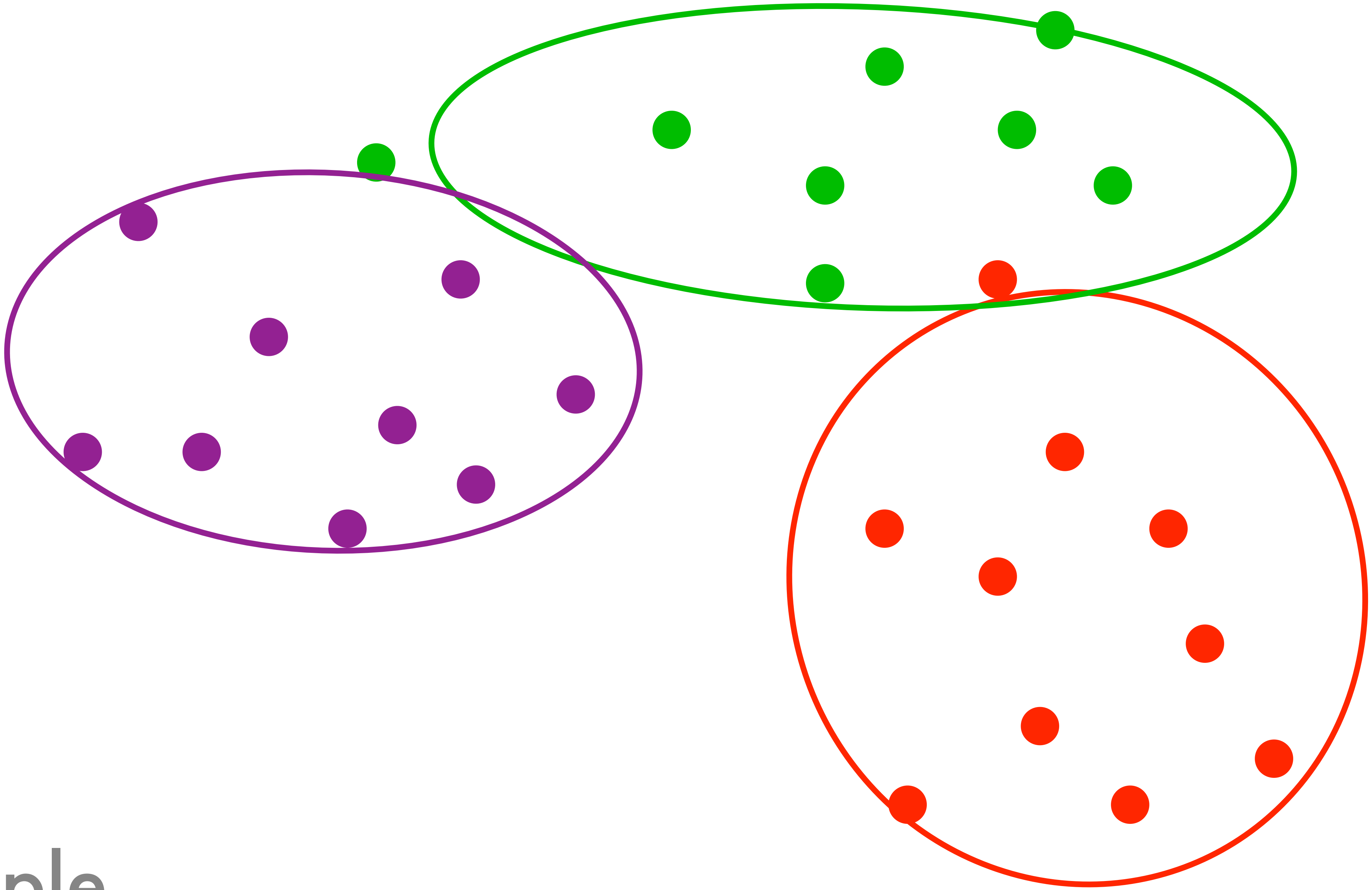
resample
cluster model

Gibbs sampling for clustering



resample
cluster labels

Gibbs sampling for clustering



resample
cluster model

e.g. Mahout Dirichlet Process Clustering

Inference Algorithm ≠ Model

Inference Algorithm ≠ Model

Corollary: EM ≠ Clustering

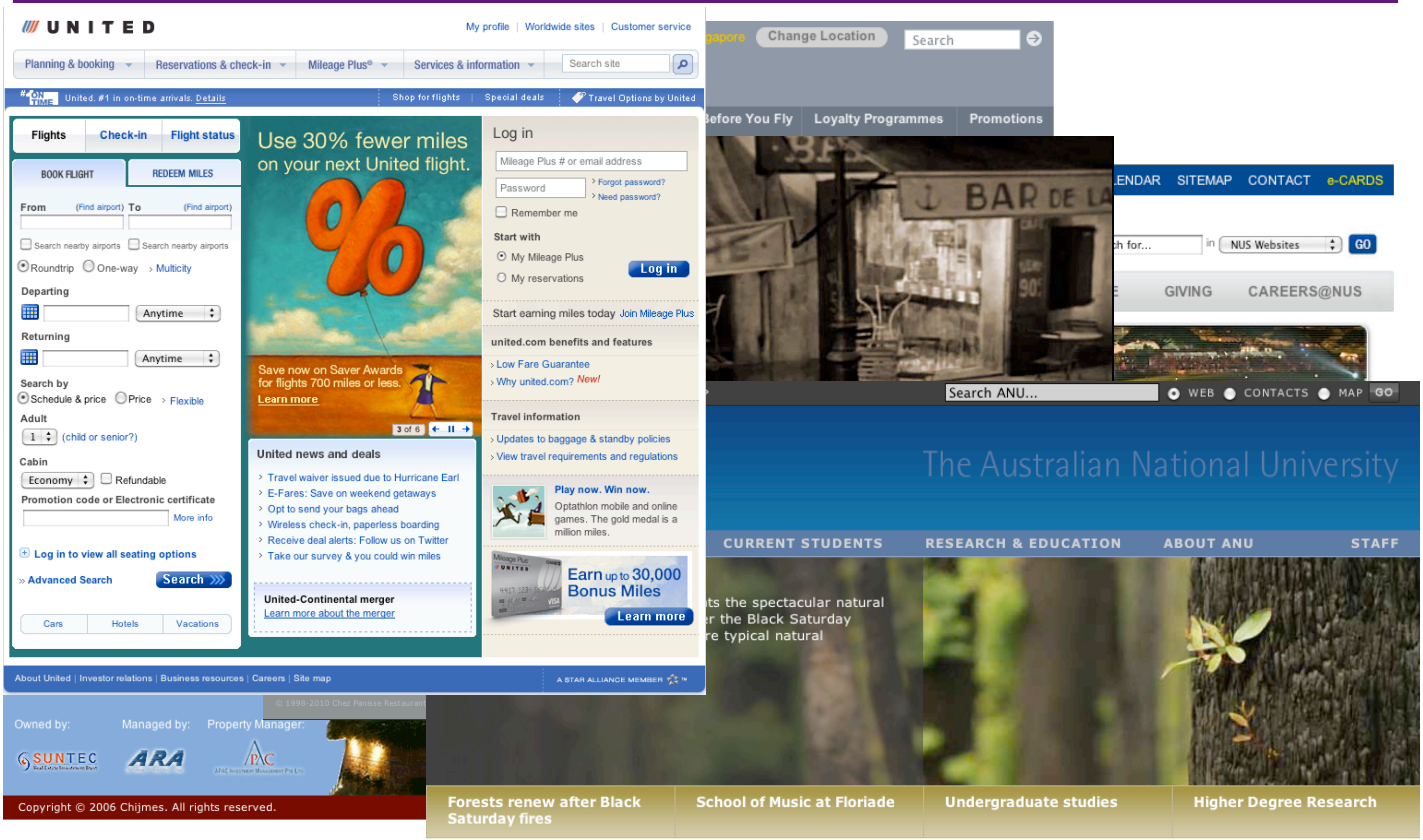
Topic models

Grouping objects

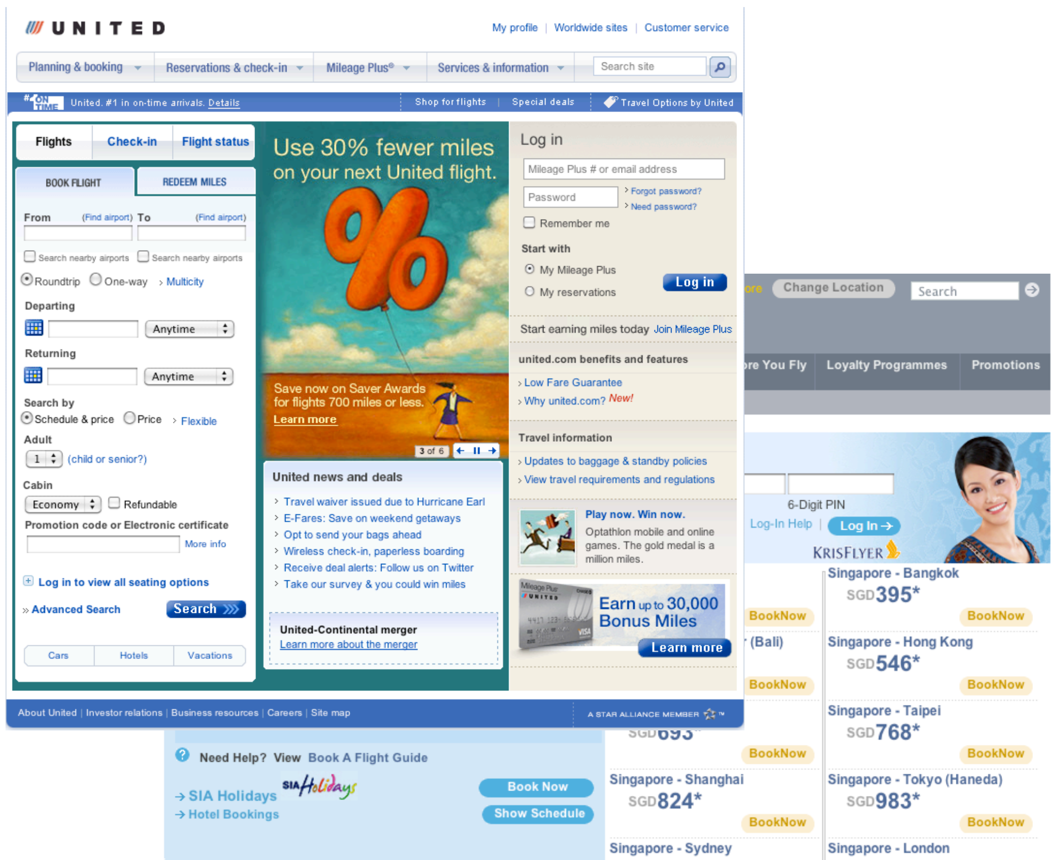
Grouping objects

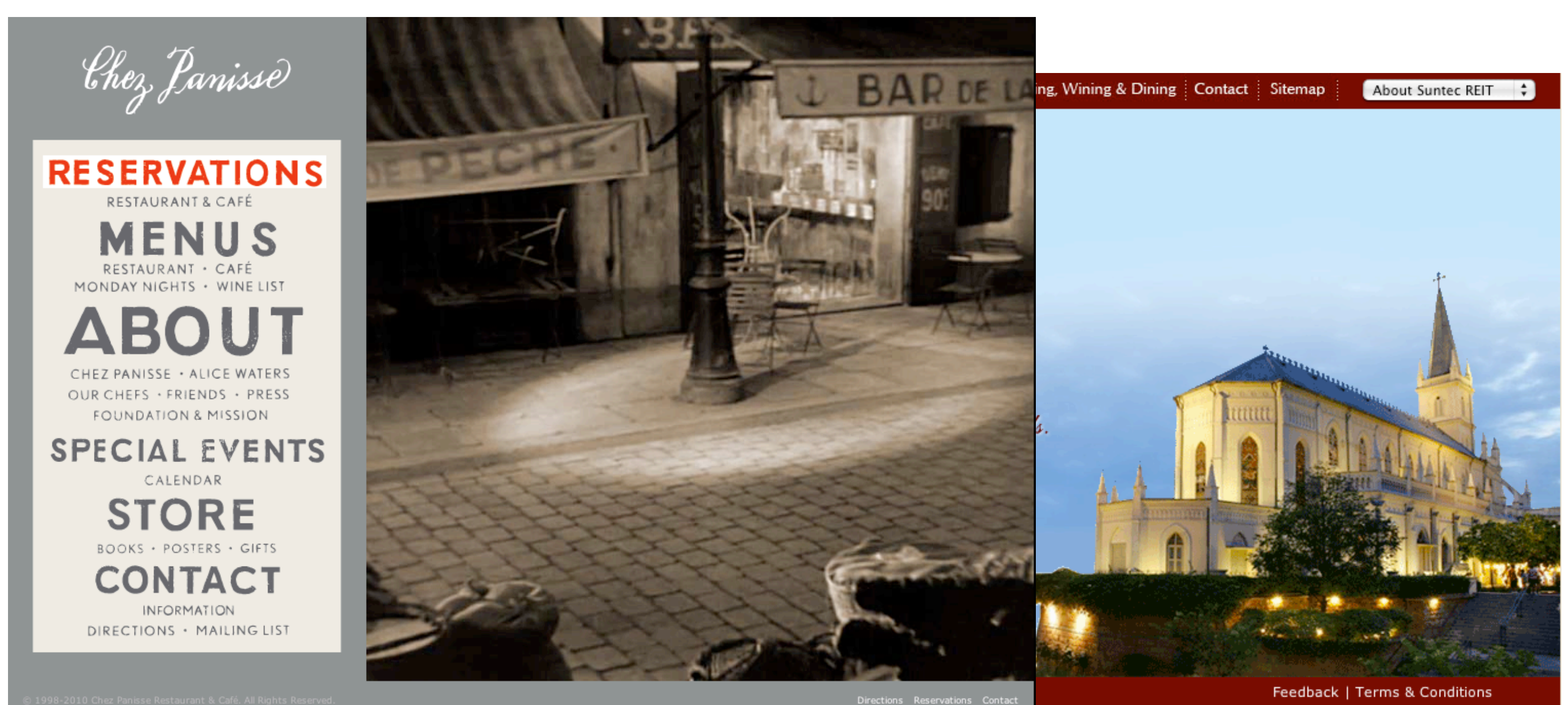
YAHOO!

Grouping objects



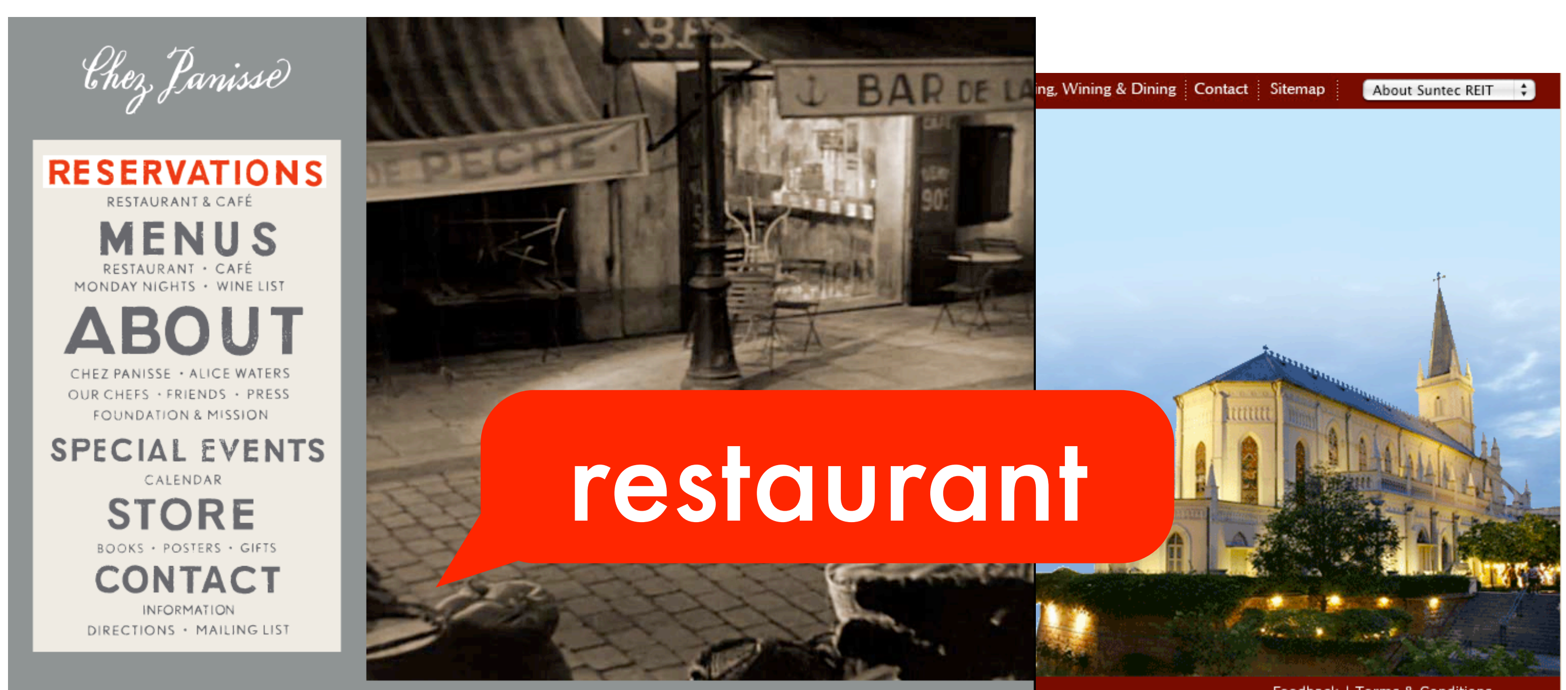
Grouping objects





YAHOO!

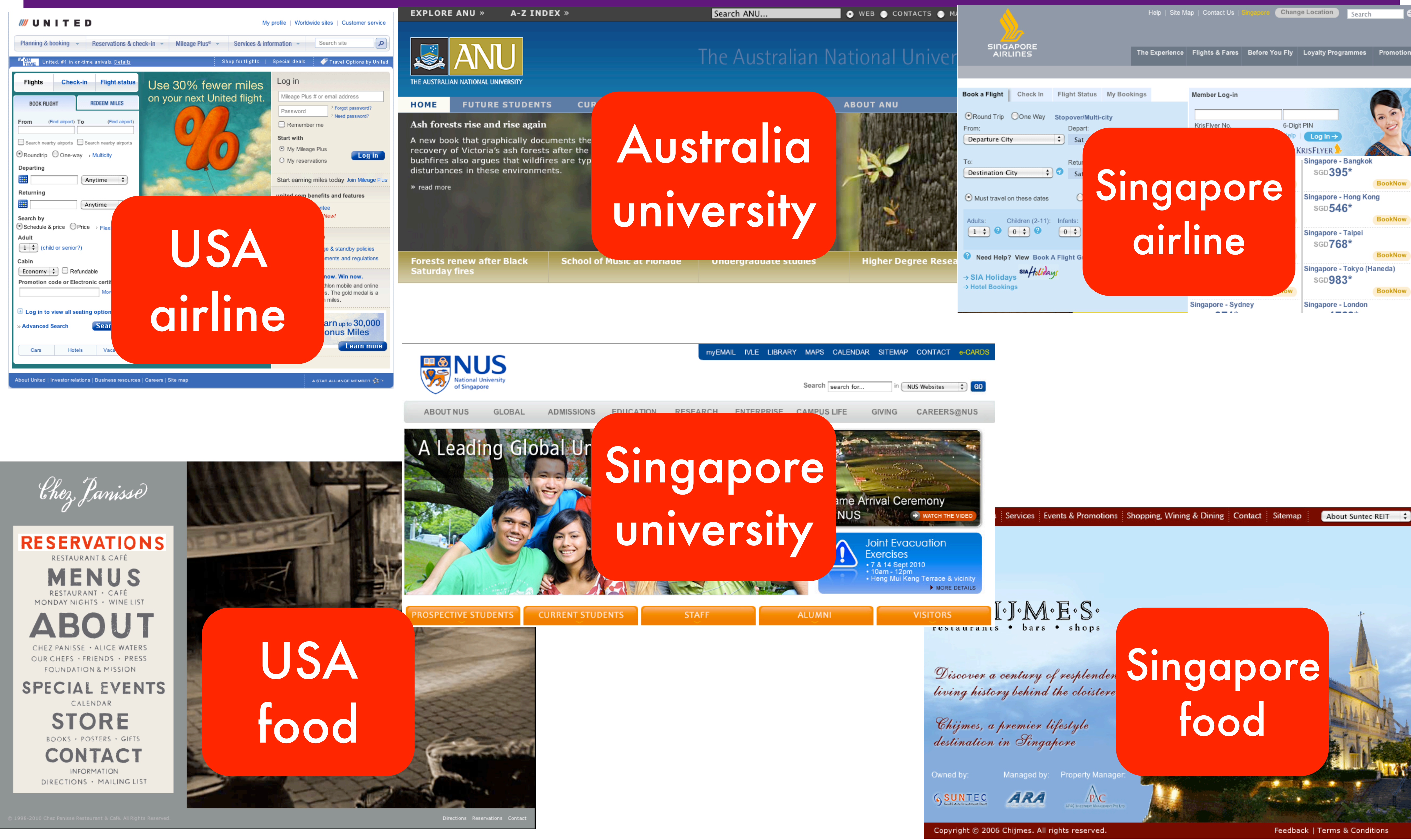
Grouping objects



YAHOO!

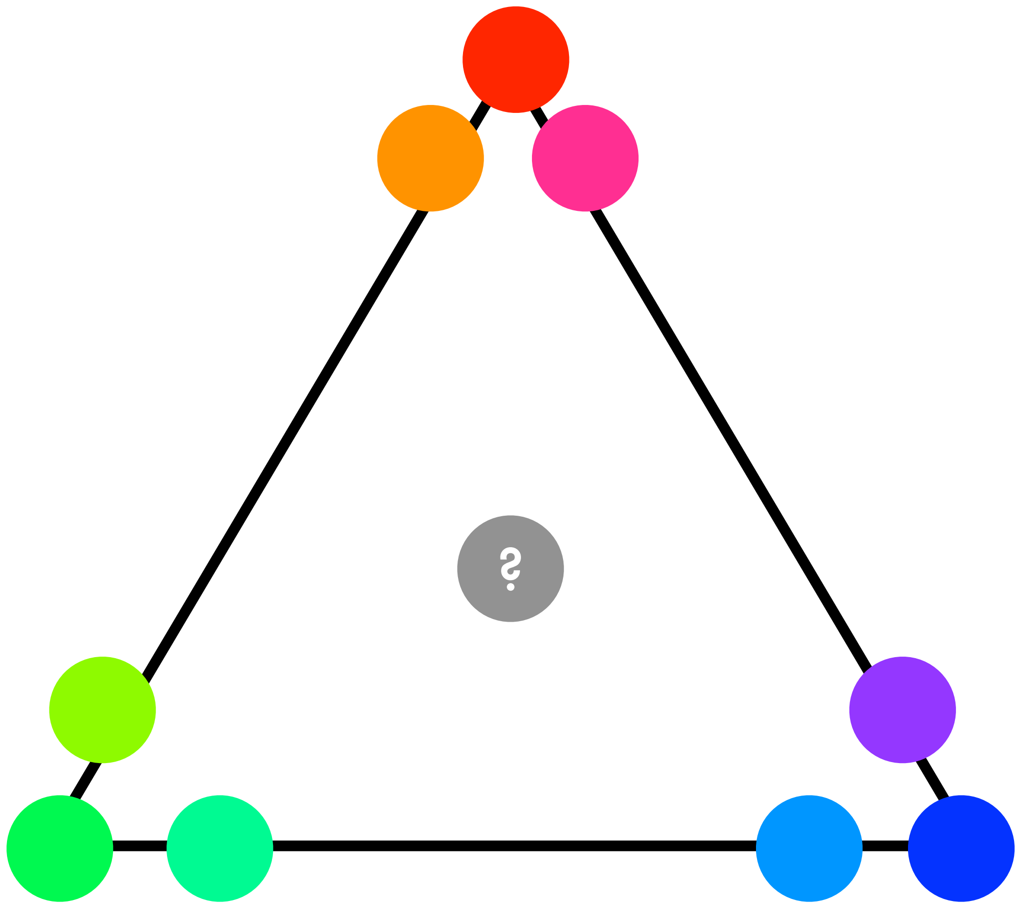
Grouping objects

Topic Models



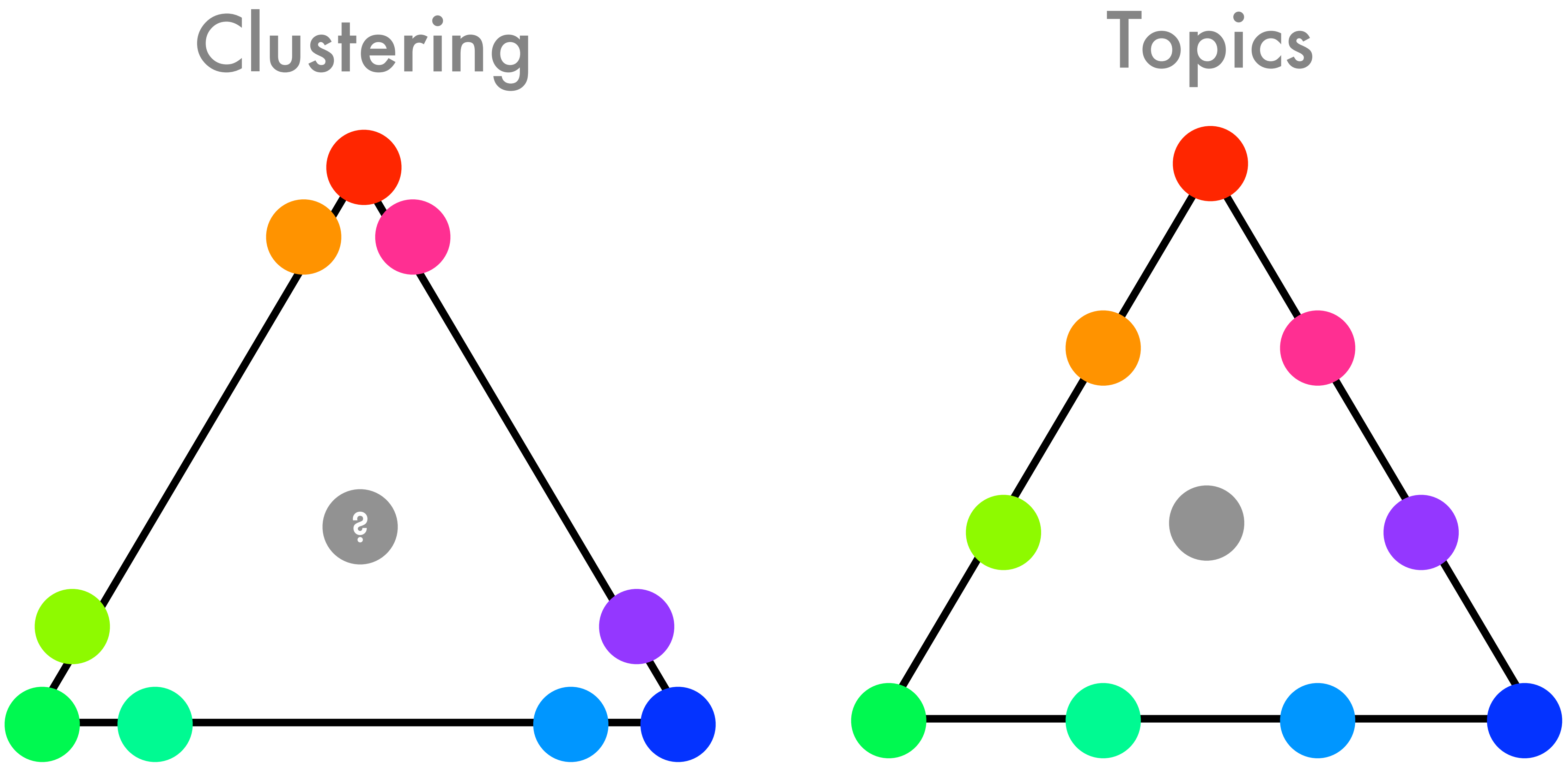
Clustering & Topic Models

Clustering



group objects
by prototypes

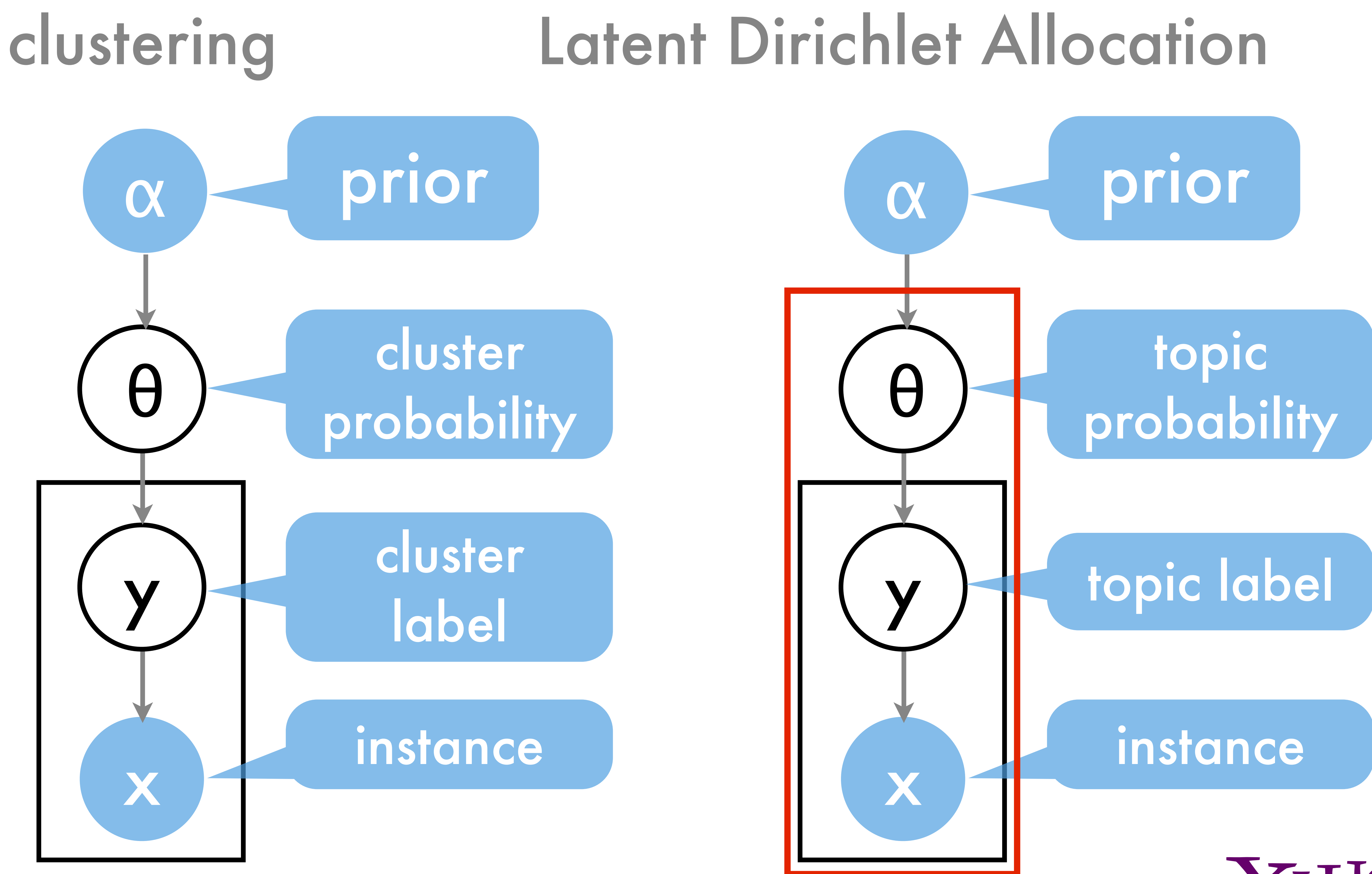
Clustering & Topic Models



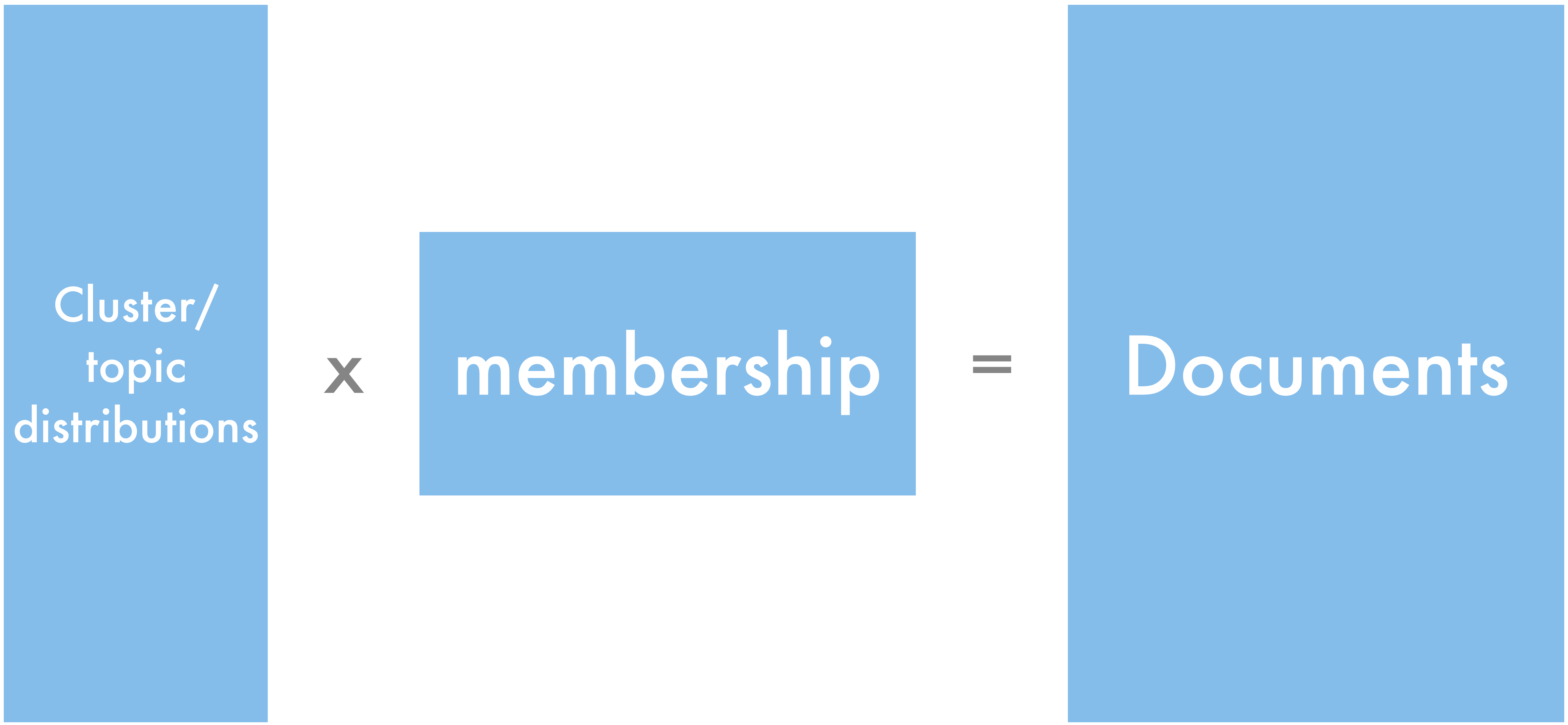
group objects by prototypes

decompose objects into prototypes

Clustering & Topic Models



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

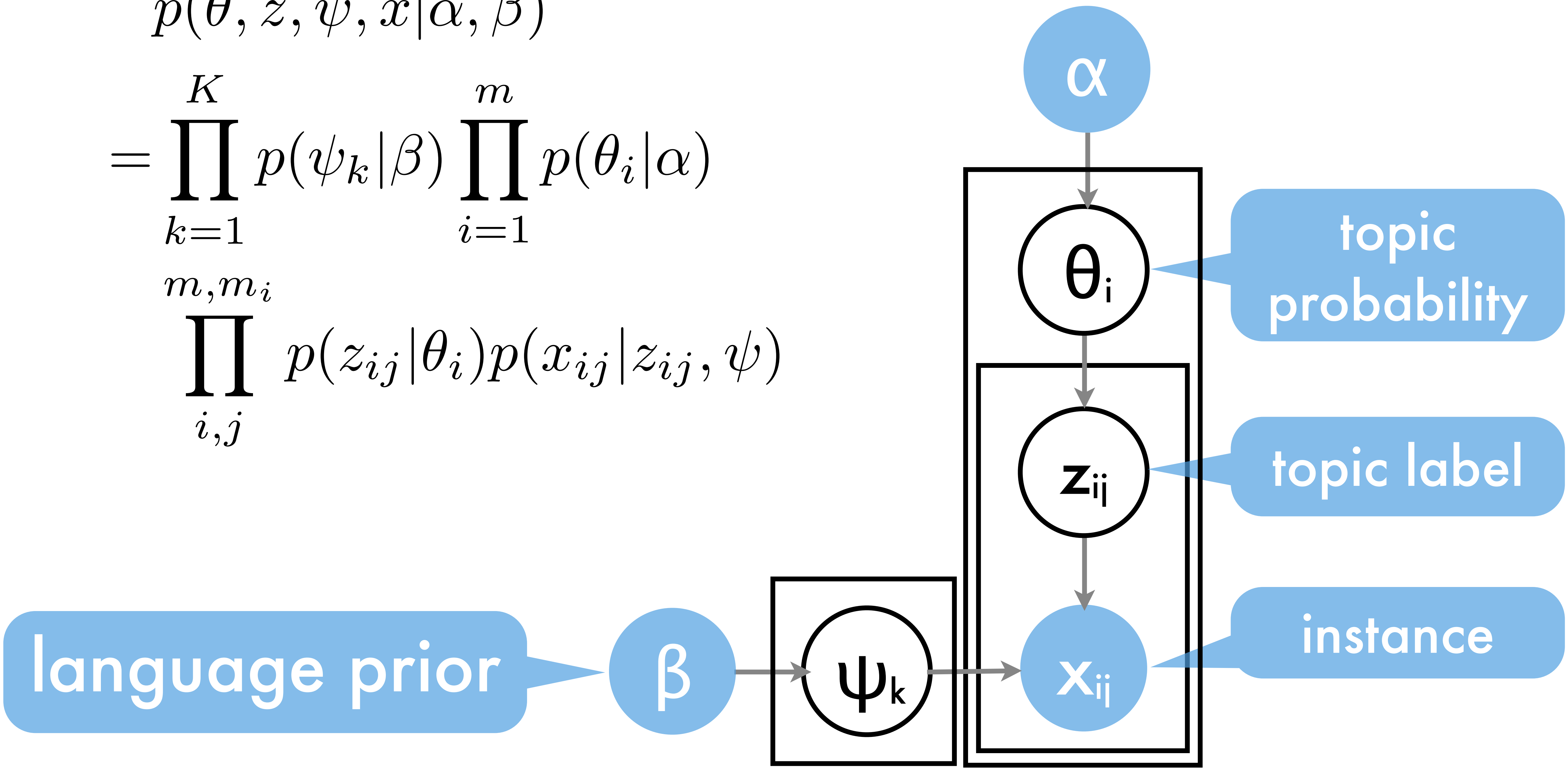
Collapsed Gibbs Sampler

Joint Probability Distribution

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

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

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



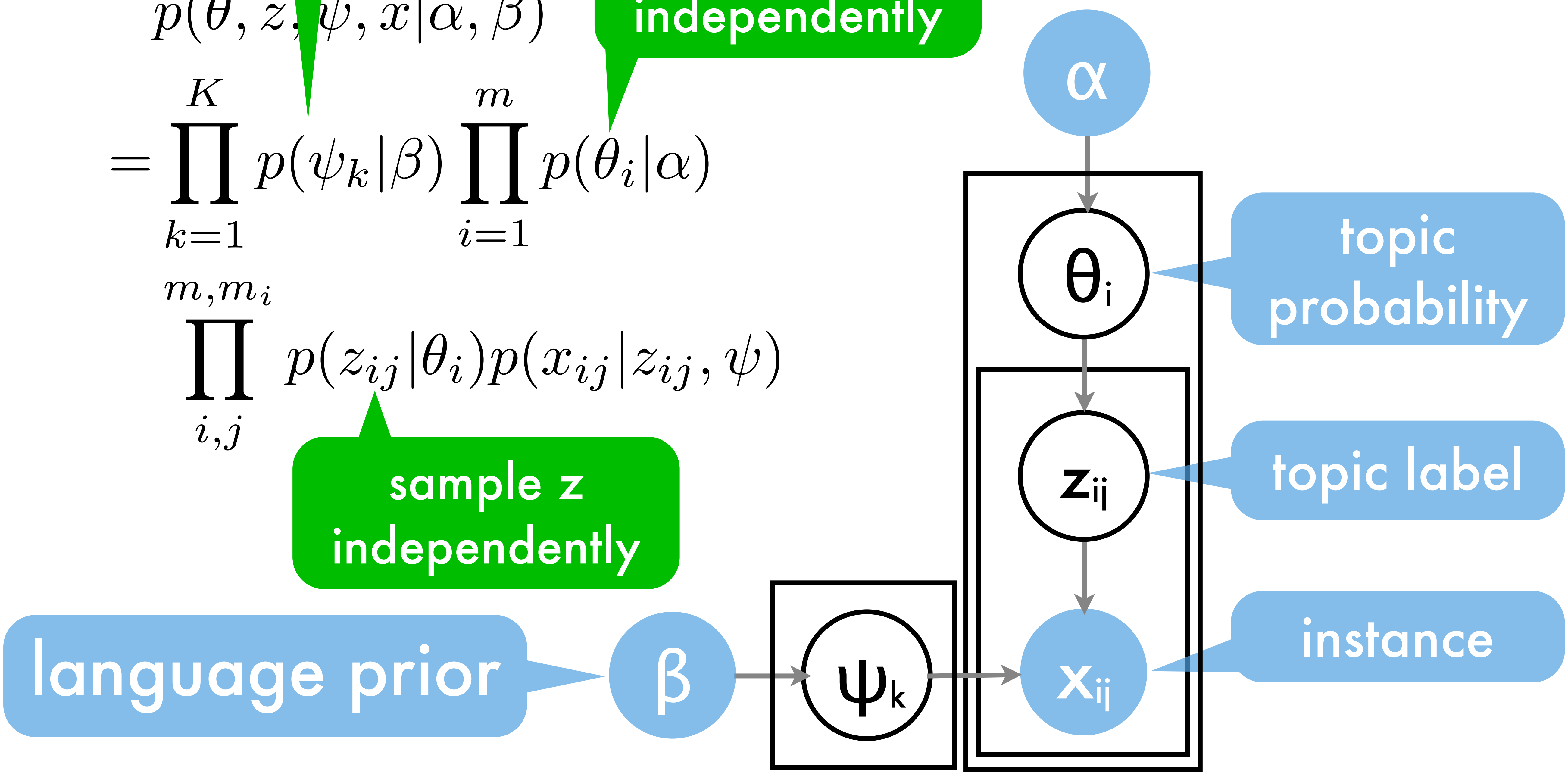
Joint Probability Distribution

sample Ψ independently

sample θ independently

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

sample z independently



Joint Probability Distribution

sample Ψ independently

sample θ independently

slow

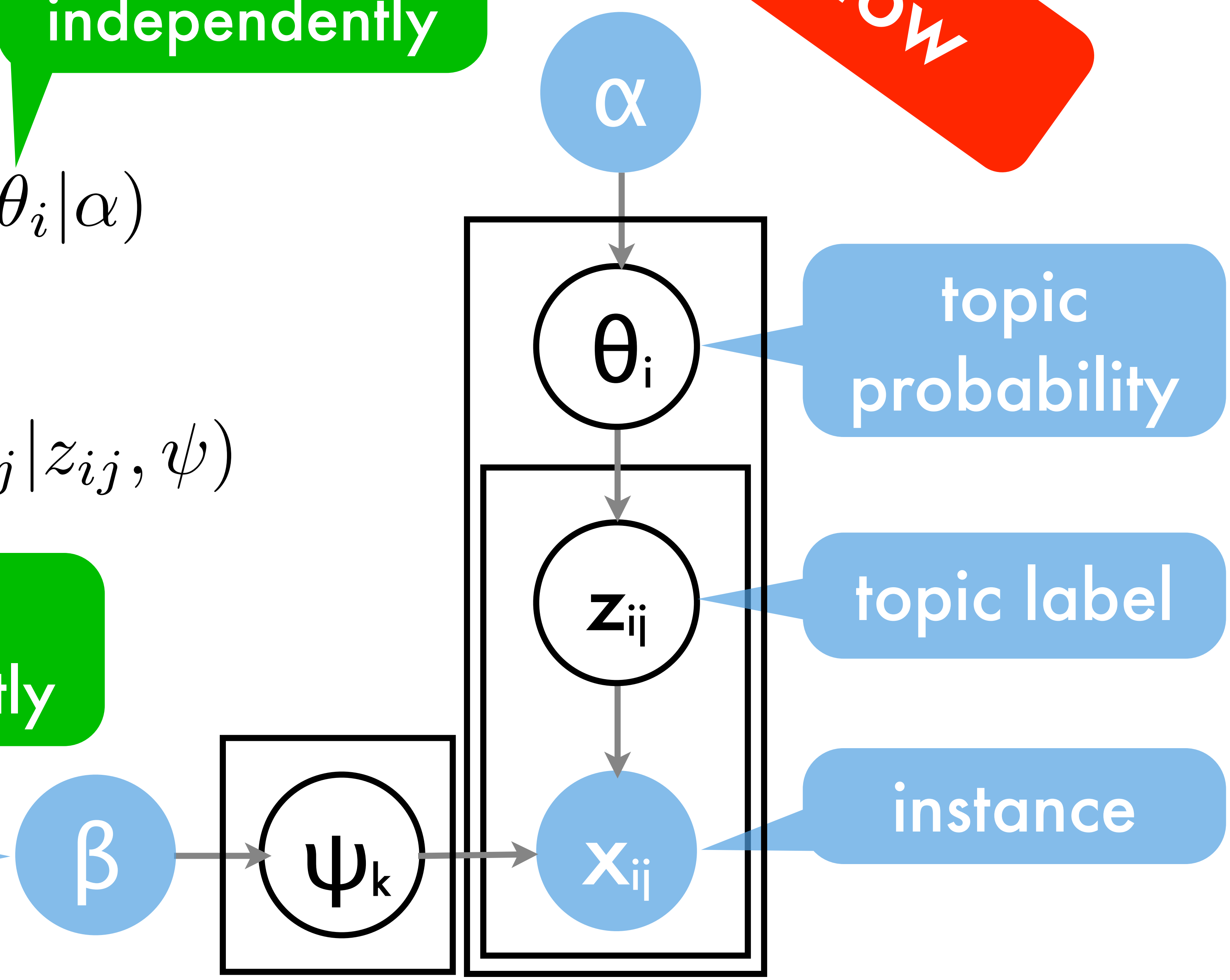
$$p(\theta, z, \psi, x|\alpha, \beta)$$

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

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

sample z independently

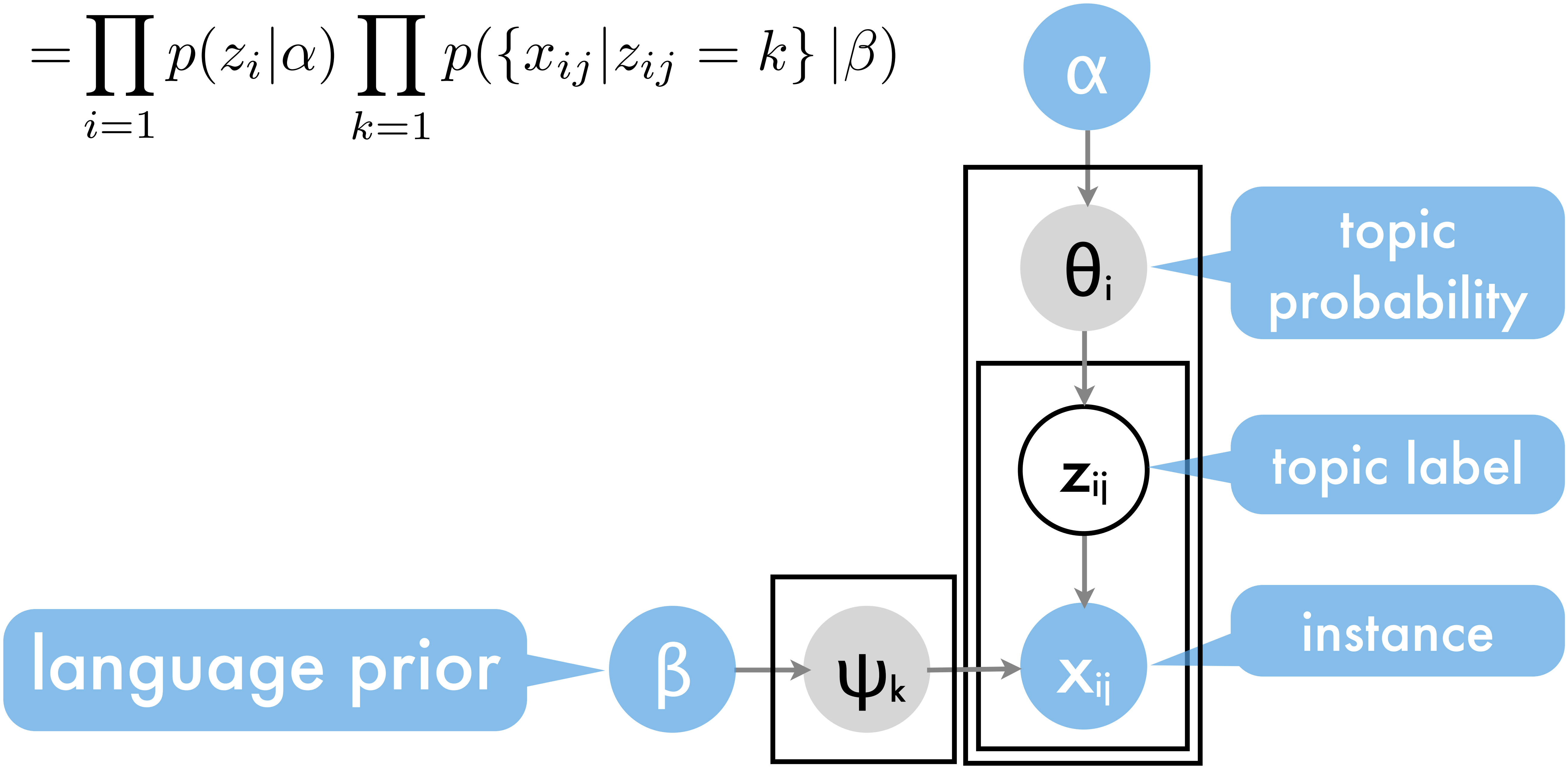
language prior



Collapsed Sampler

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

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

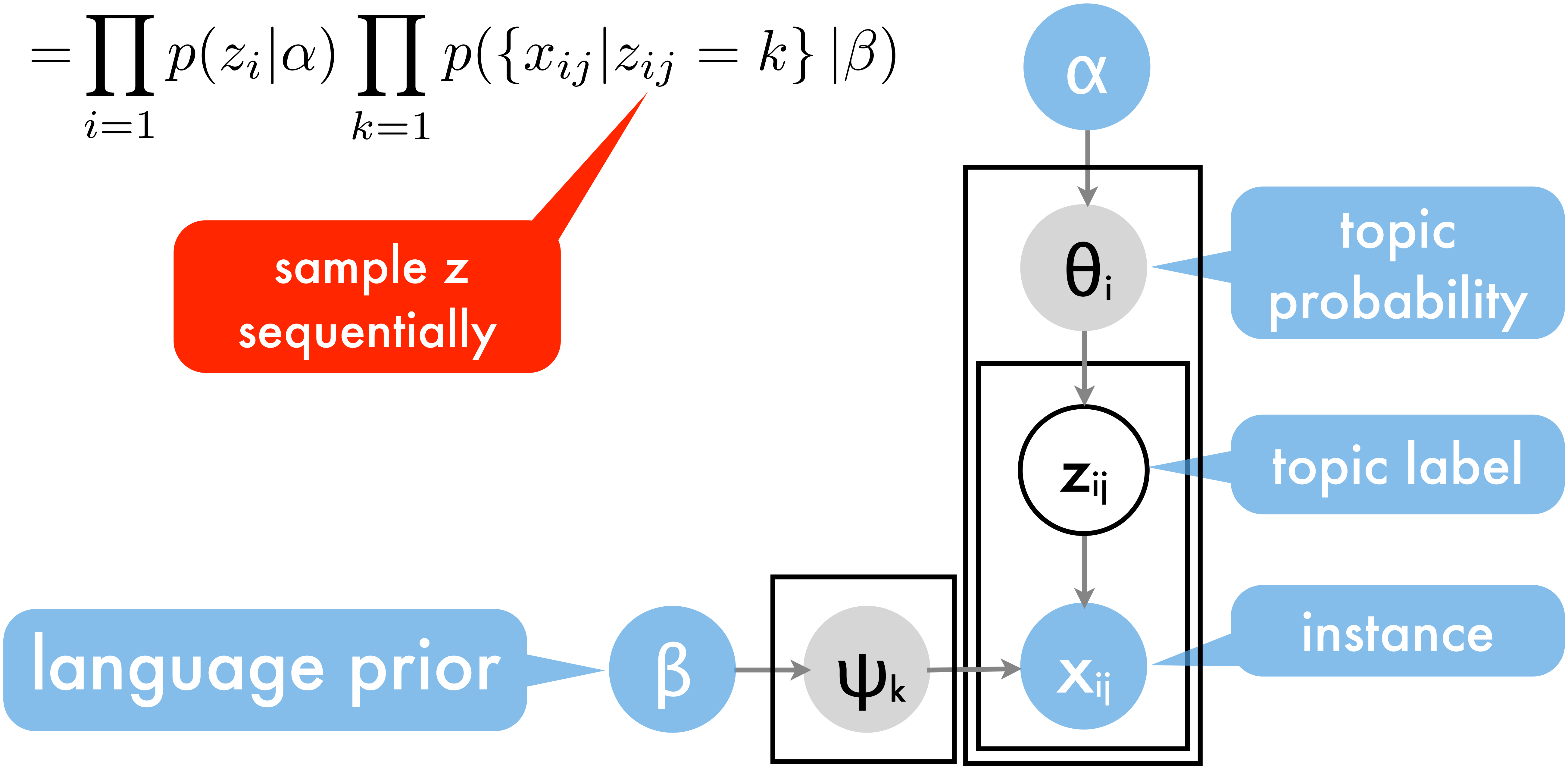


Collapsed Sampler

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

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

sample z sequentially



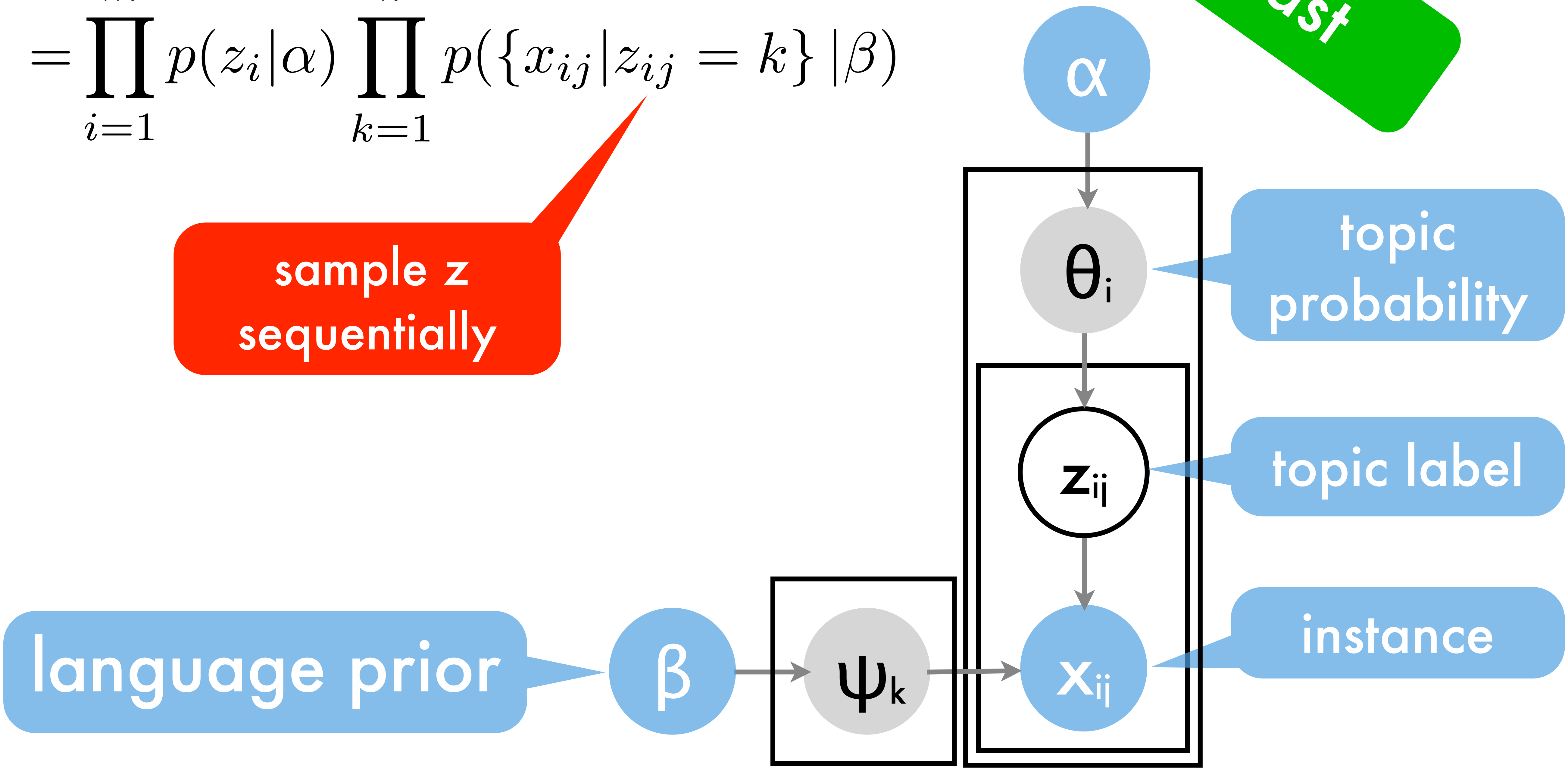
Collapsed Sampler

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

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

sample z sequentially

fast



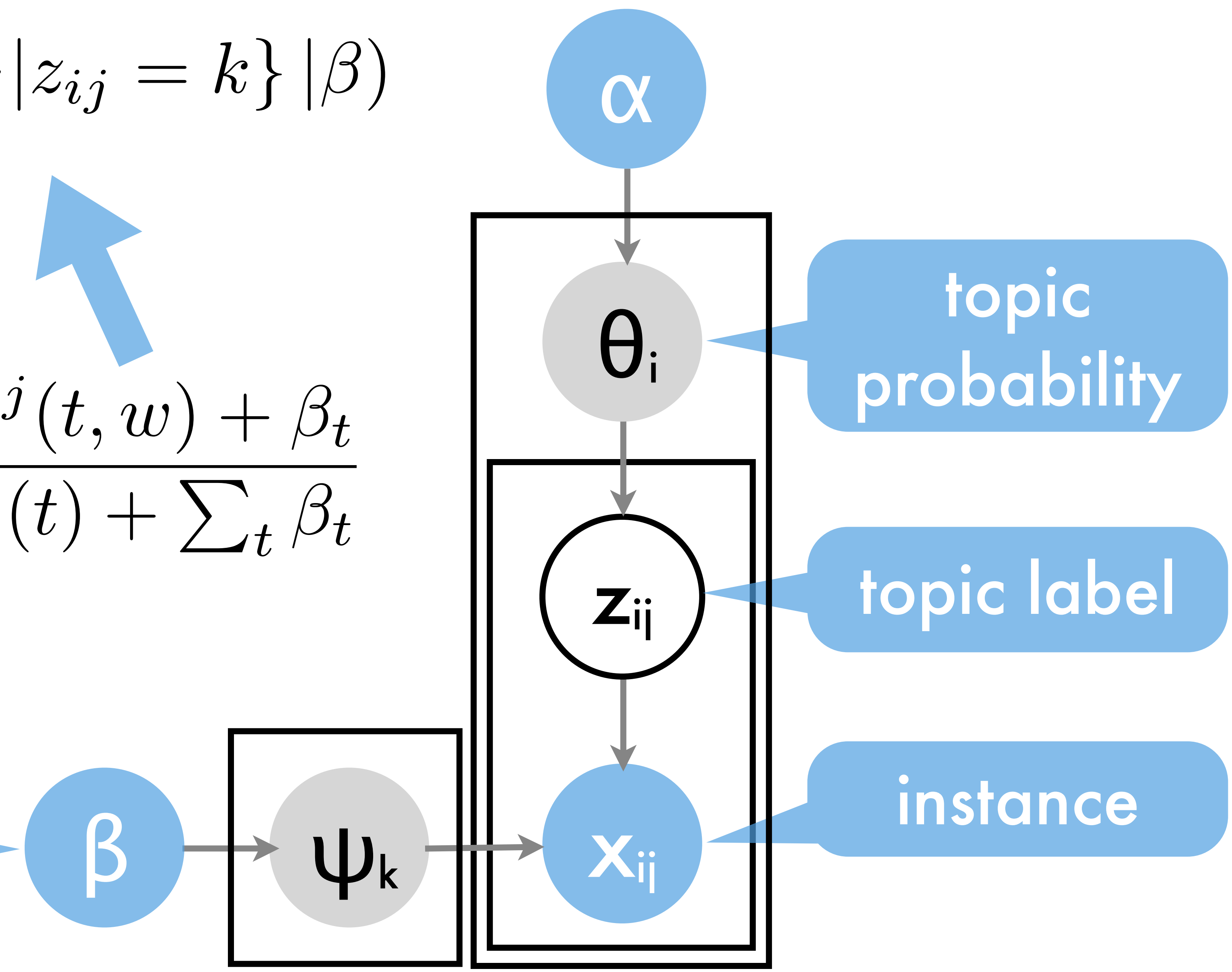
Collapsed Sampler

Griffiths & Steyvers, 2005

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

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

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



Collapsed Sampler

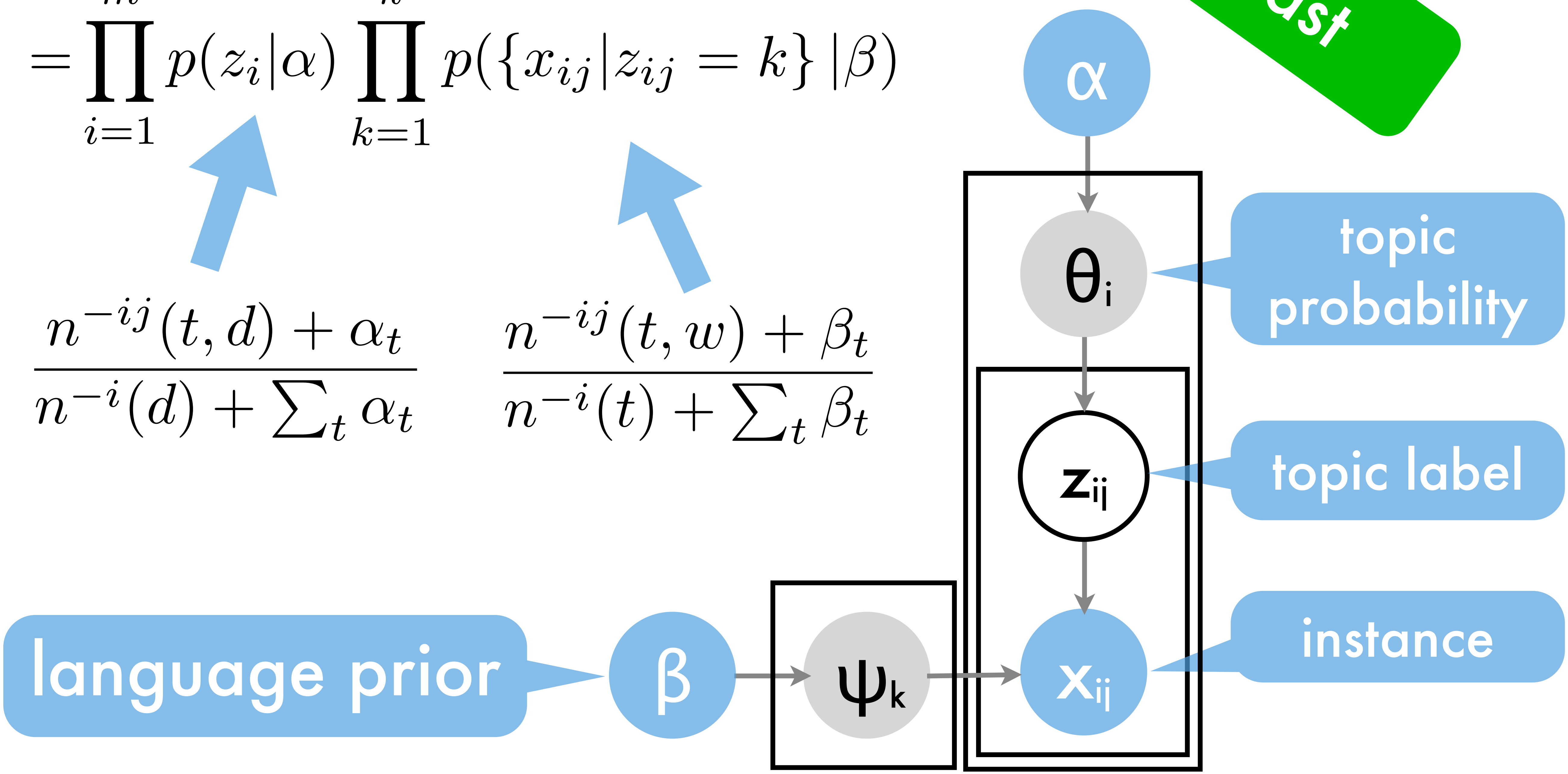
Griffiths & Steyvers, 2005

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

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

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

fast



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

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

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

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

slow

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

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

changes rapidly

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

slow

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

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

changes rapidly

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

slow

moderately fast

YAHOO!

State of the art
UMass Mallet, UC Irvine, Google

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

table out of sync

memory inefficient

blocking

network bound

changes rapidly

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

slow

moderately fast

YAHOO!

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

YAHOO!

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

YAHOO!

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

YAHOO!

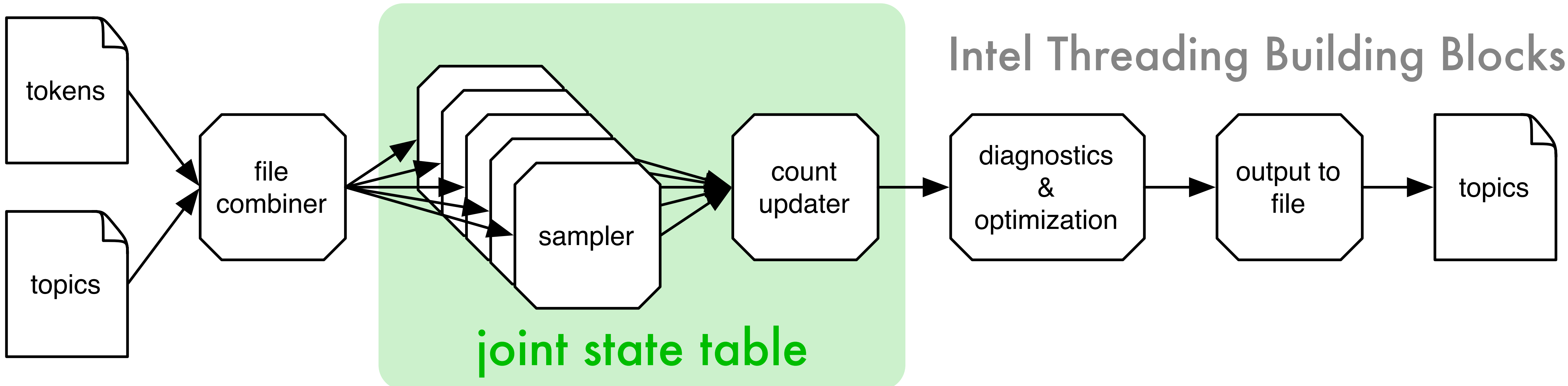
Our Approach

- 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

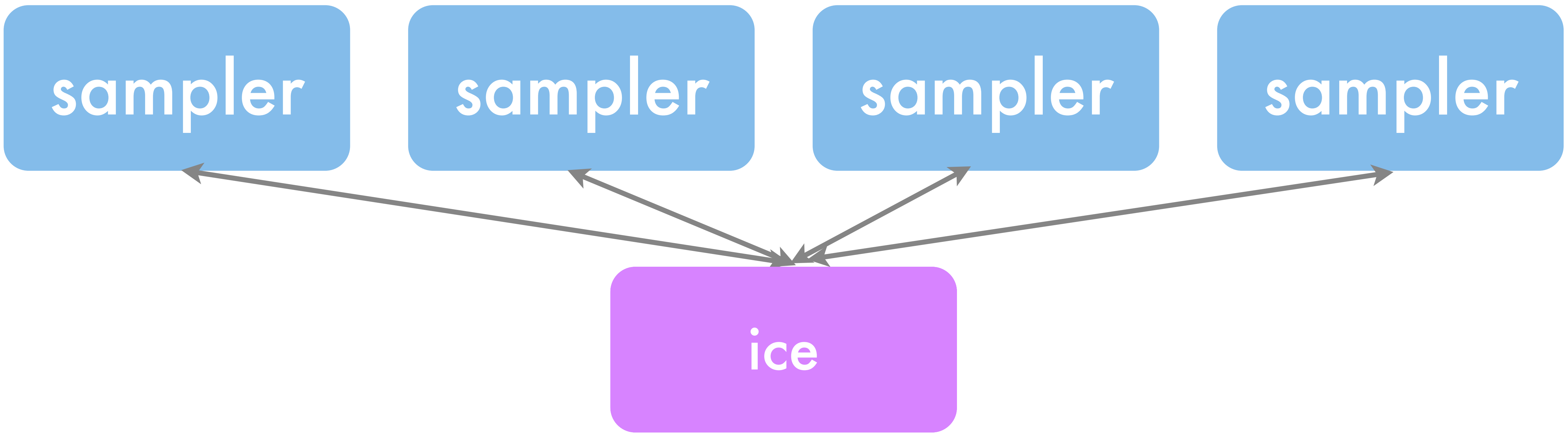
Architecture details

Multicore Architecture



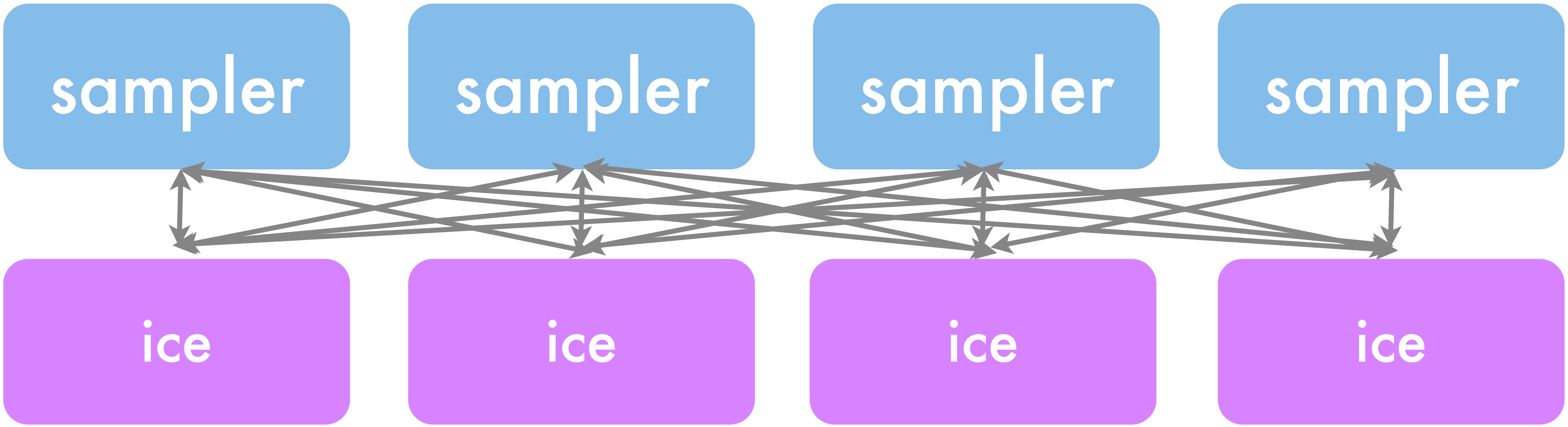
- Decouple multithreaded sampling and updating (almost) avoids stalling for locks in the sampler
- Joint state table
  - much less memory required
  - samplers syncronized (10 docs vs. millions delay)
- Hyperparameter update via stochastic gradient descent
- No need to keep documents in memory (streaming)

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

YAHOO!

Making it work

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

YAHOO!

Easily extensible

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

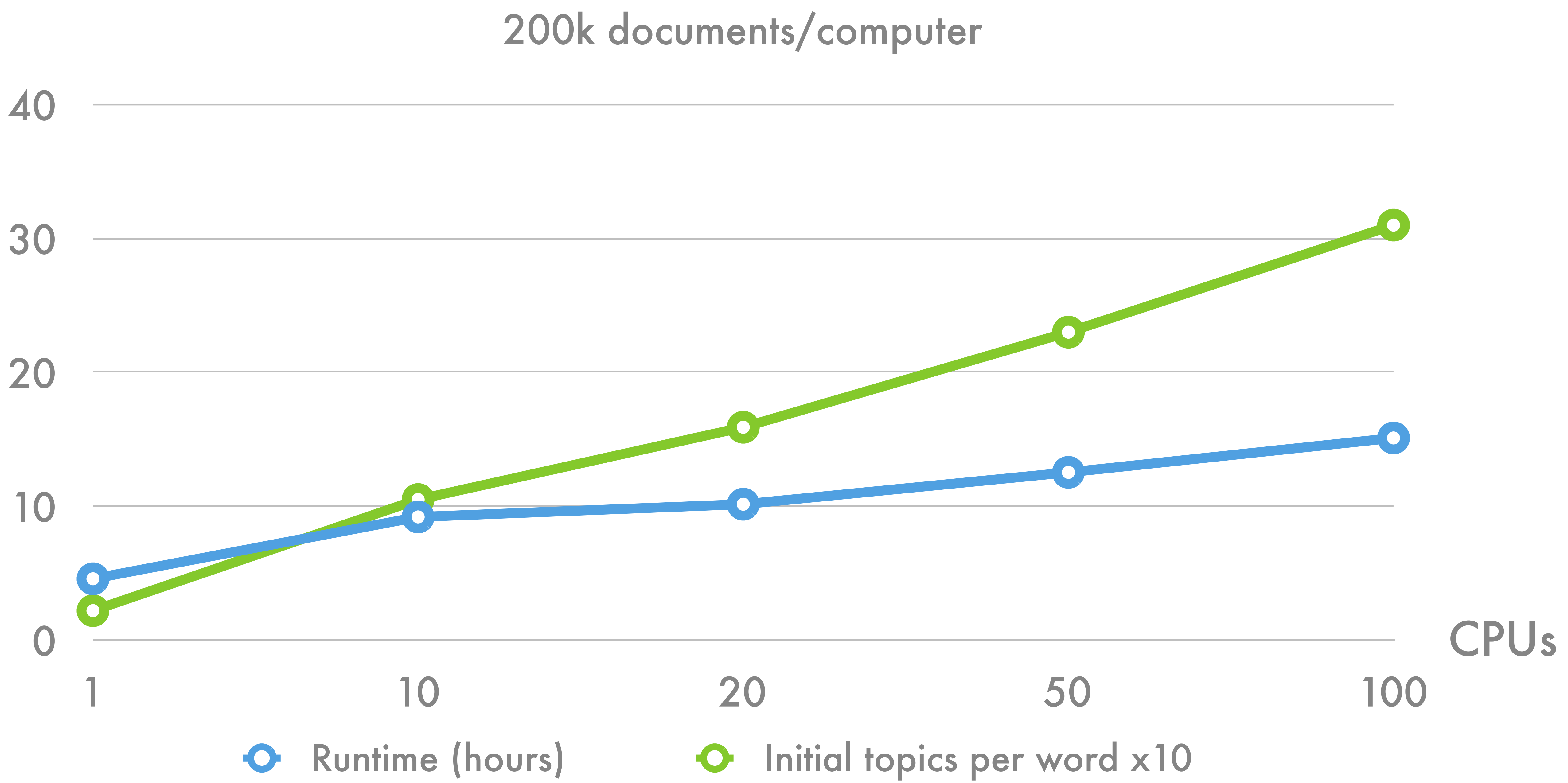
	Google LDA	Mallet	Irvine'08	Irvine'09	Yahoo LDA
Multicore	no	yes	yes	yes	yes
Cluster	MPI	no	MPI	point 2 point	memcached
State table	dictionary split	separate sparse	separate	separate	joint sparse
Schedule	synchronous exact	synchronous exact	synchronous exact	asynchronous approximate messages	asynchronous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

YAHOO!

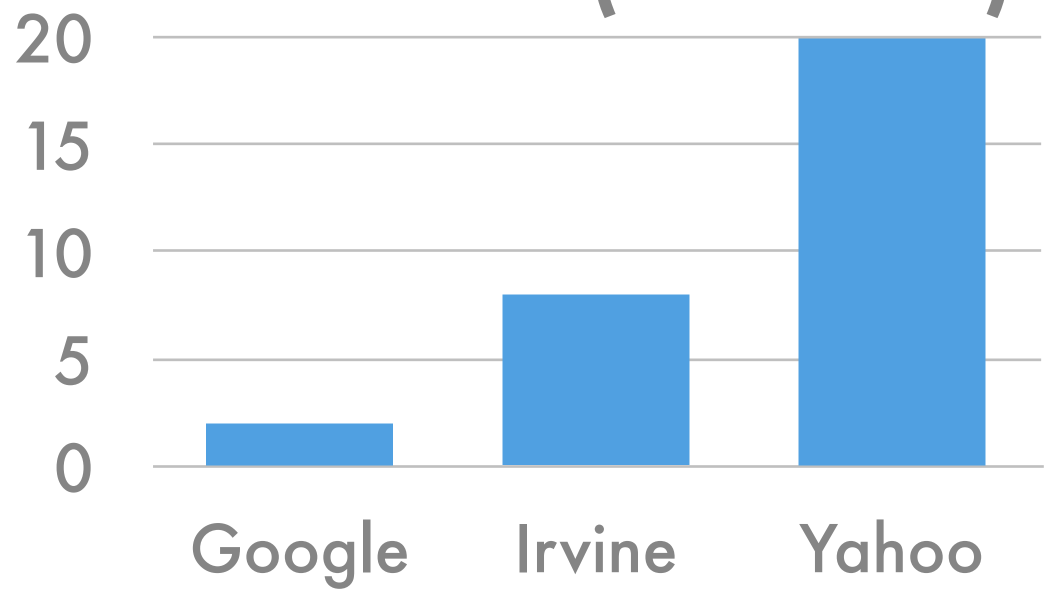
Scalability



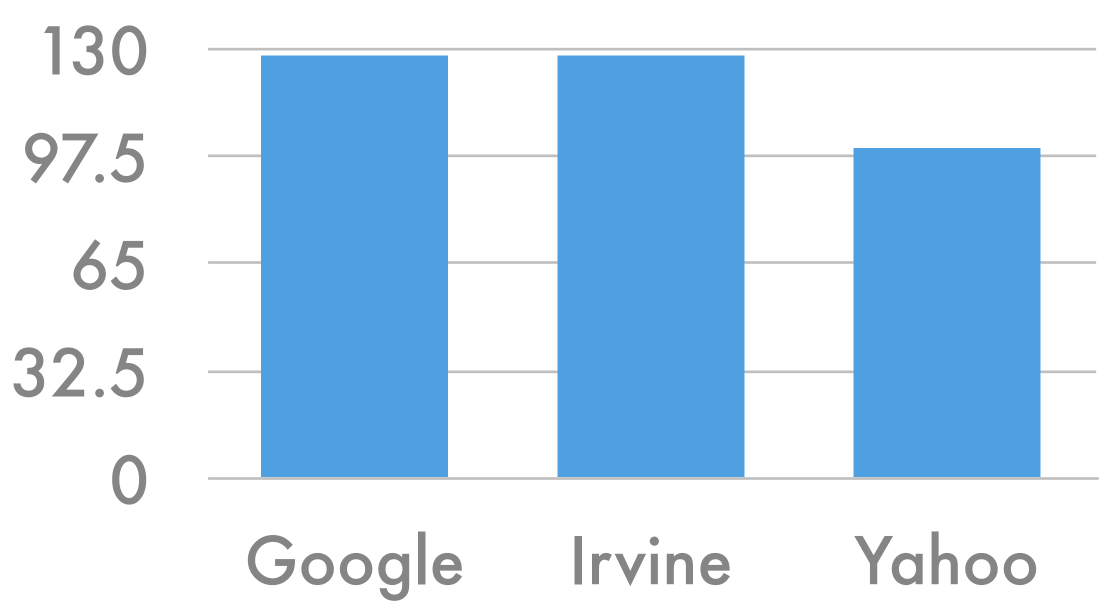
Likelihood even improves with parallelism!
-3.295 (1 node) -3.288 (10 nodes) -3.287 (20 nodes)

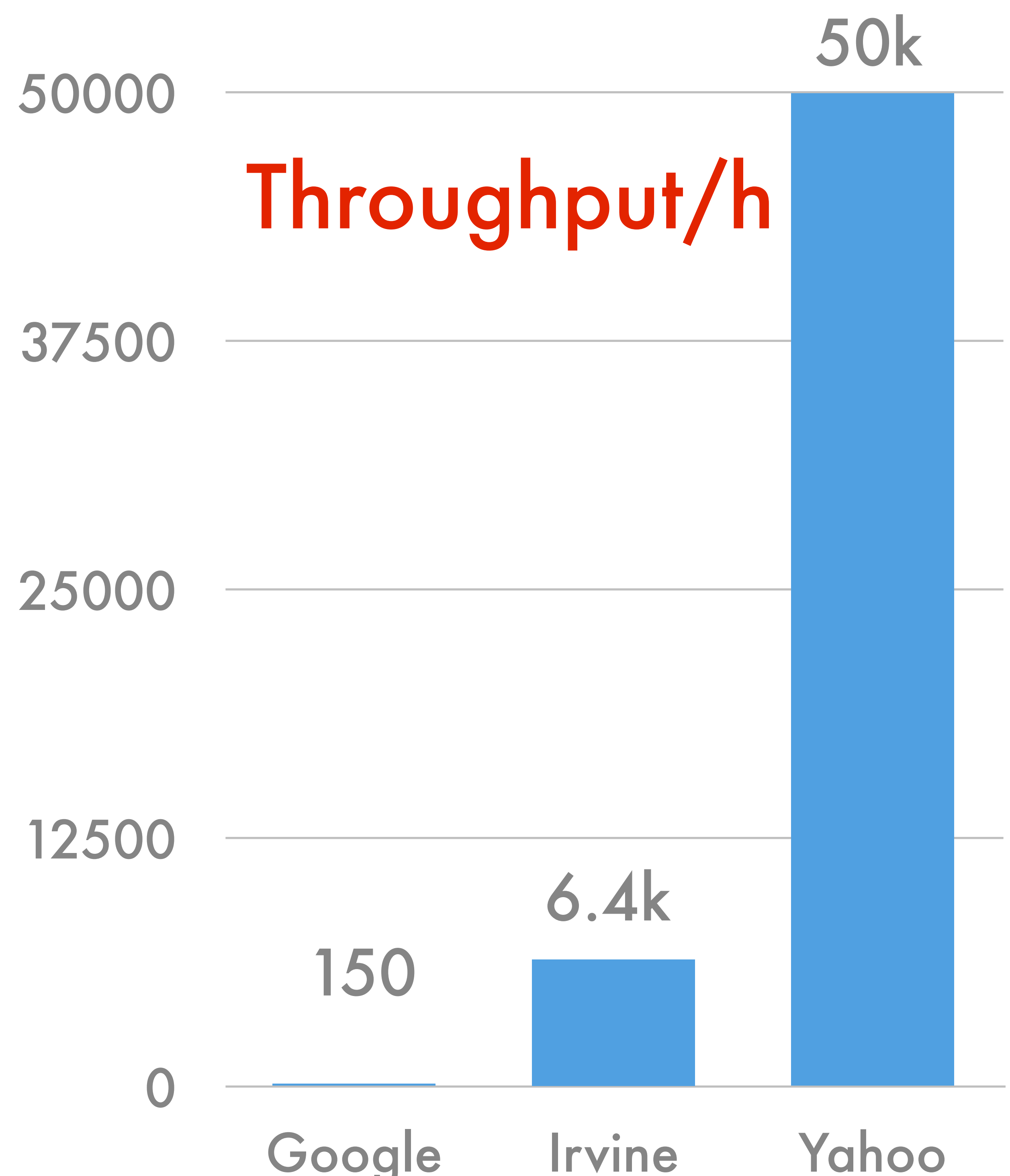
The Competition

Dataset size (millions)



Cluster size

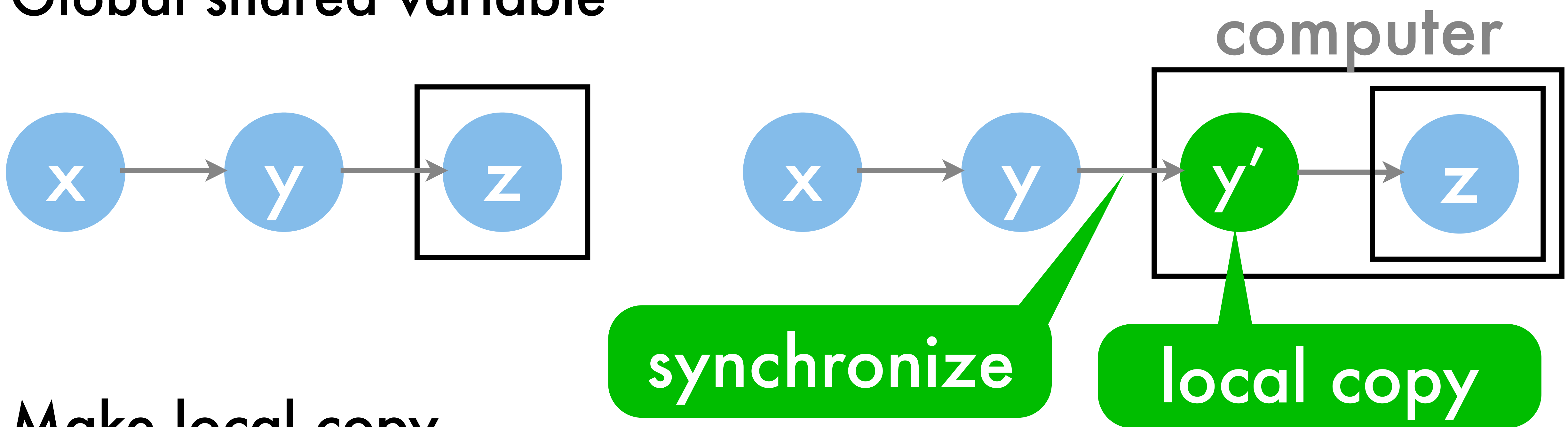




Design Principles

Variable Replication

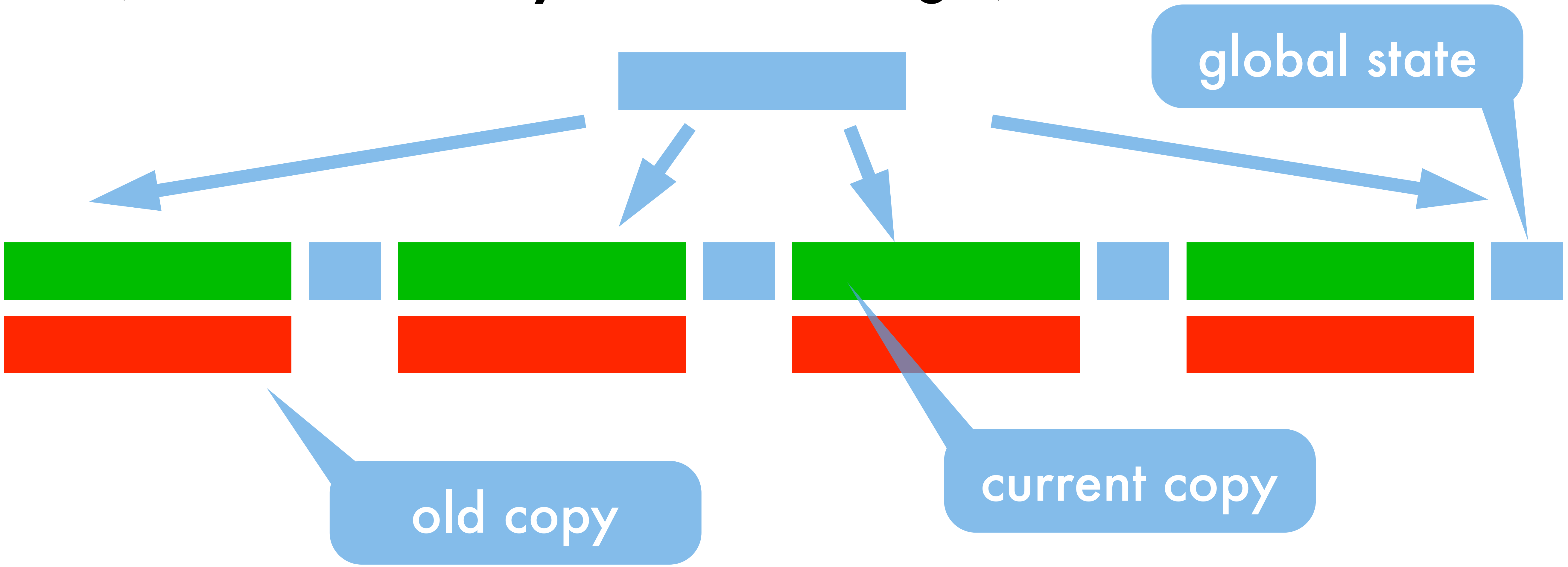
- Global shared variable



- 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!

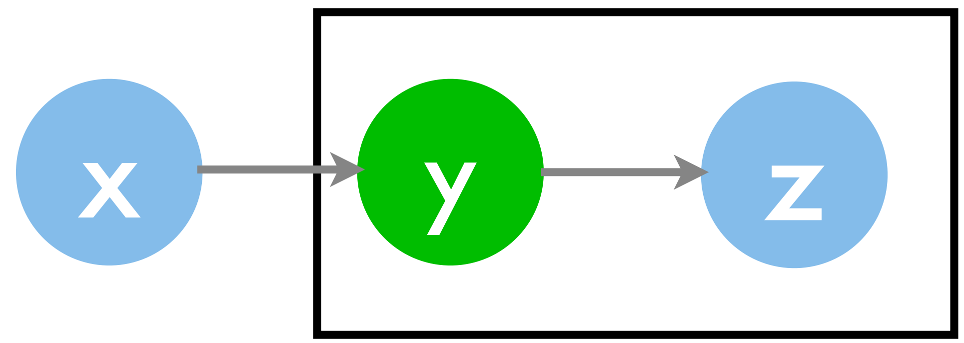
Asymmetric Message Passing

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

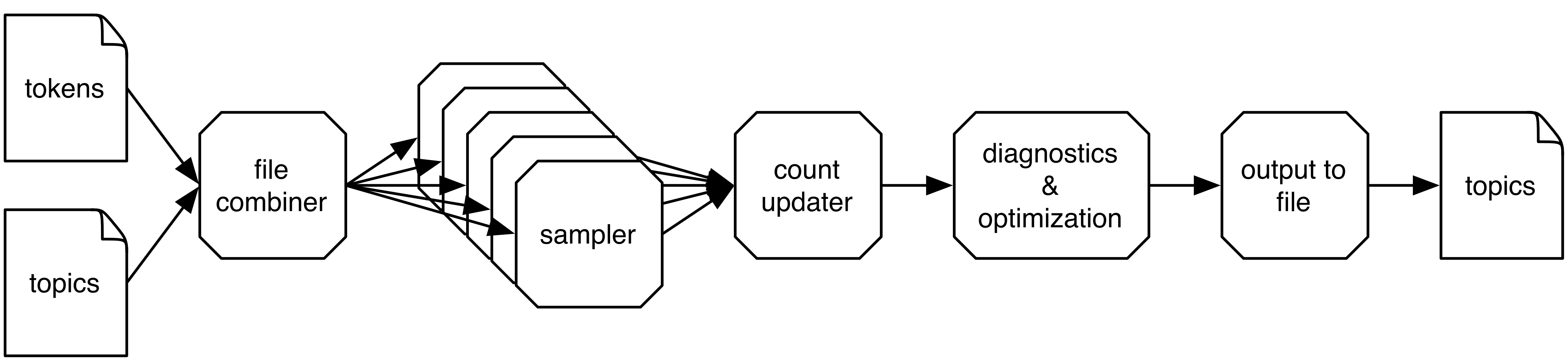


Out of core storage

- 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

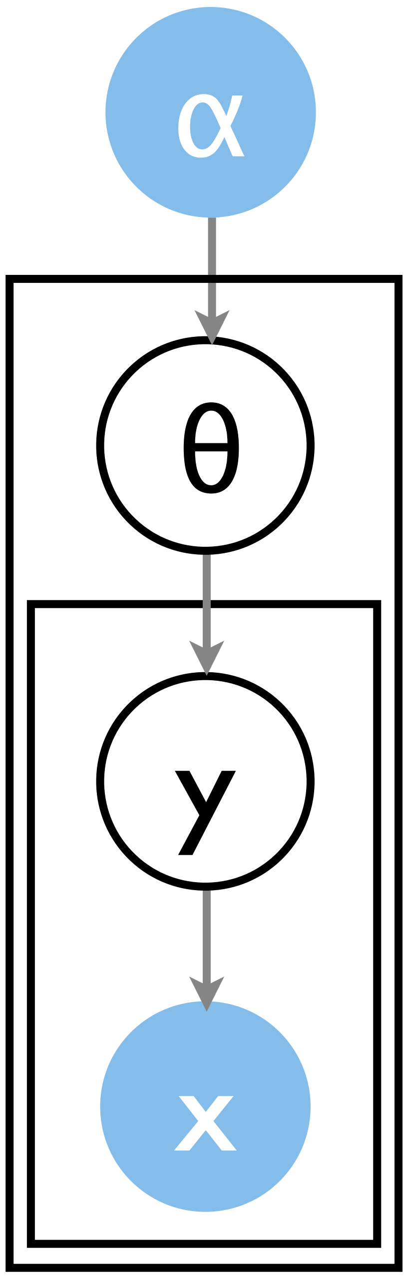
- Inference in graphical models
- Clustering
- Topic models
- Sampling
- Implementation details

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)
  - Simplical Mixture (Girolami & Kaban)
- Side information
  - Upstream conditioning (Mimno & McCallum)
  - Downstream conditioning (Petterson et al.)
  - Supervised LDA (Blei and McAulliffe 2007; Lacoste, Sha and Jordan 2008; Zhu, Ahmed and Xing 2009)



Correlated topic models

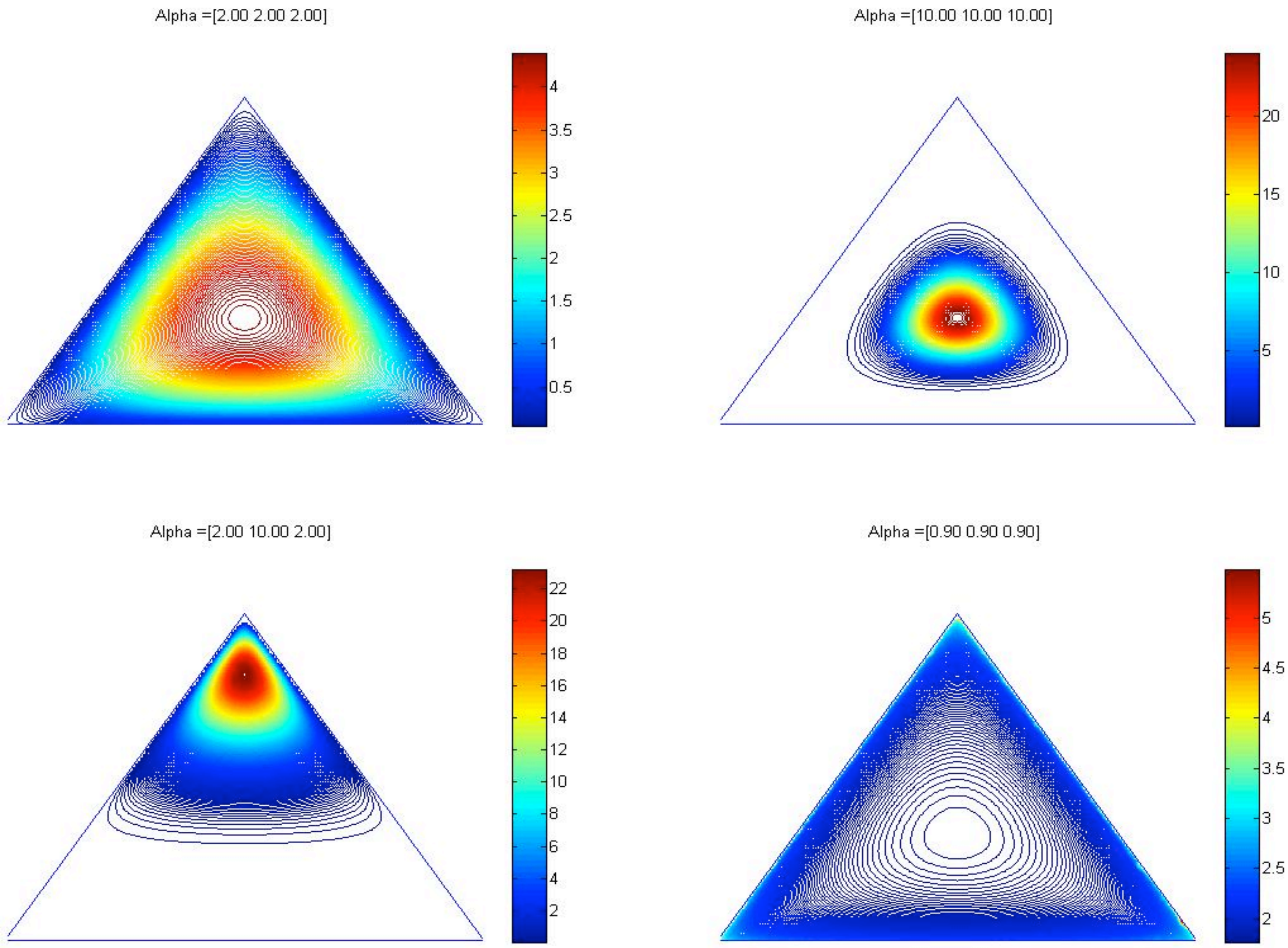
- Dirichlet distribution
  - Can only model which topics are hot
  - Does not model relationships between topics

Correlated topic models

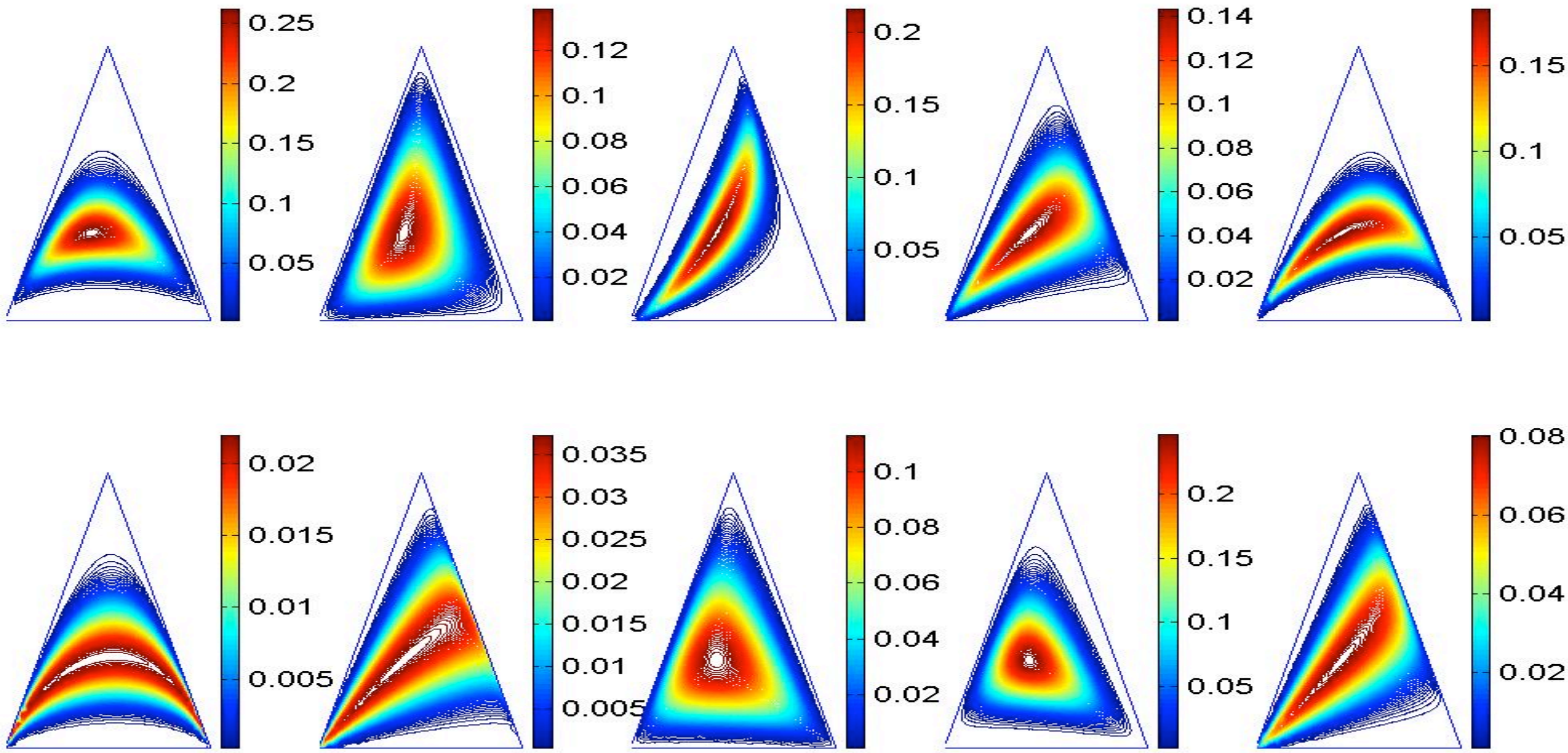
- Dirichlet distribution
  - Can only model which topics are hot
  - Does not model relationships between topics
- Key idea
  - We expect to see documents about sports and health but not about sports and politics
  - Uses a logistic normal distribution as a prior
- Conjugacy is no longer maintained
-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

Blei & Lafferty 2005; Ahmed & Xing 2007

Dirichlet prior on topics

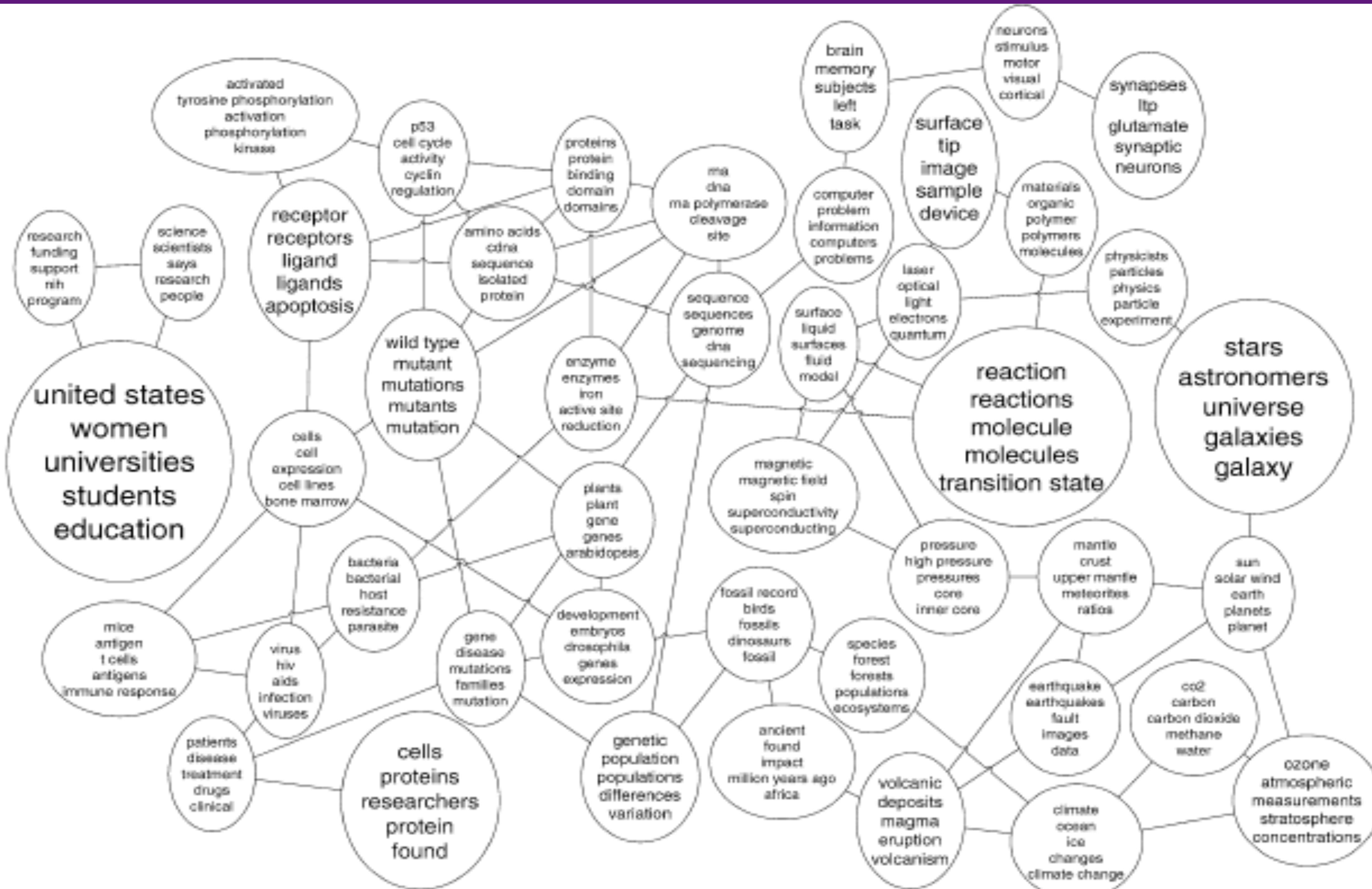


Log-normal prior on topics



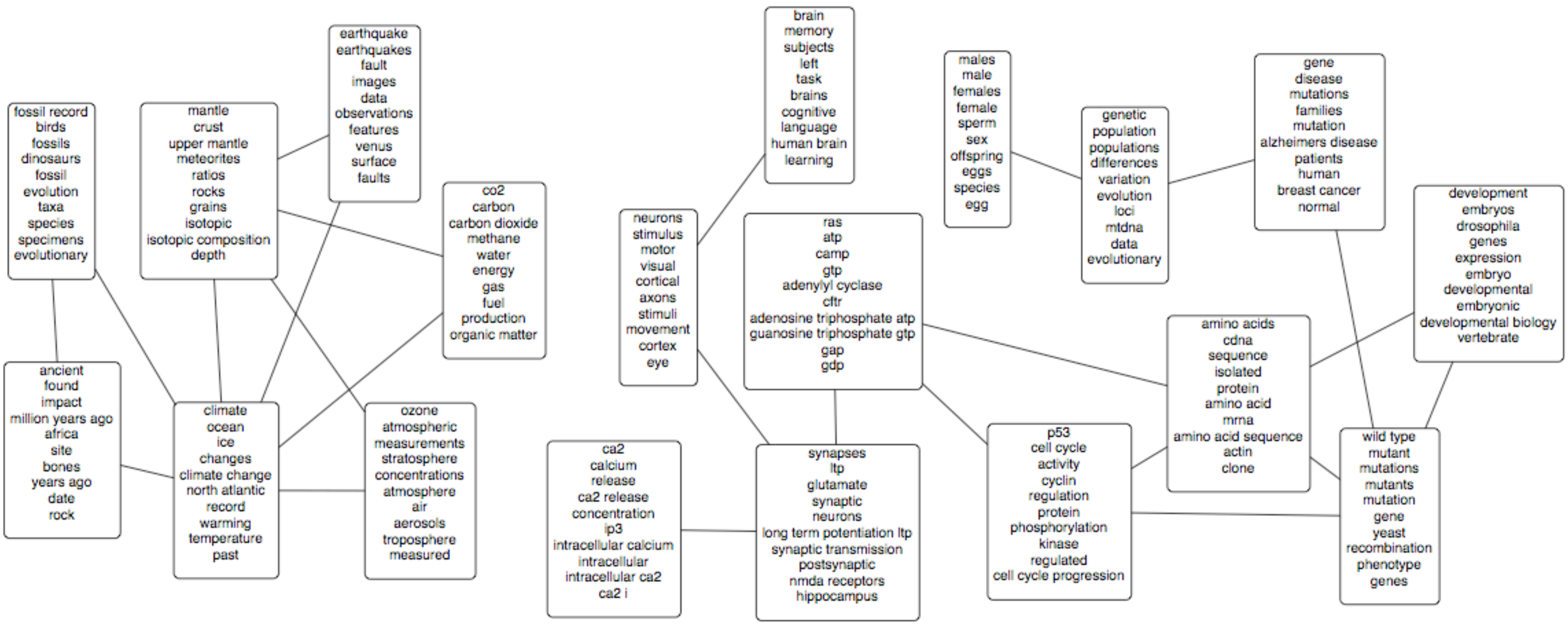
$\theta = e^{\eta - g(\eta)}$ **with** $\eta \sim \mathcal{N}(\mu, \Sigma)$

Correlated topics



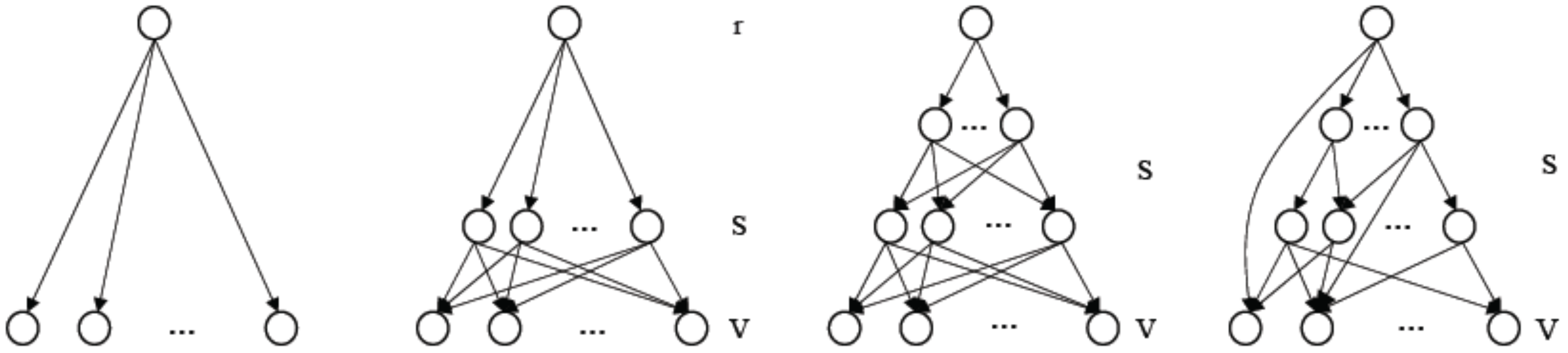
Blei and Lafferty 2005

Correlated topics



Pachinko Allocation

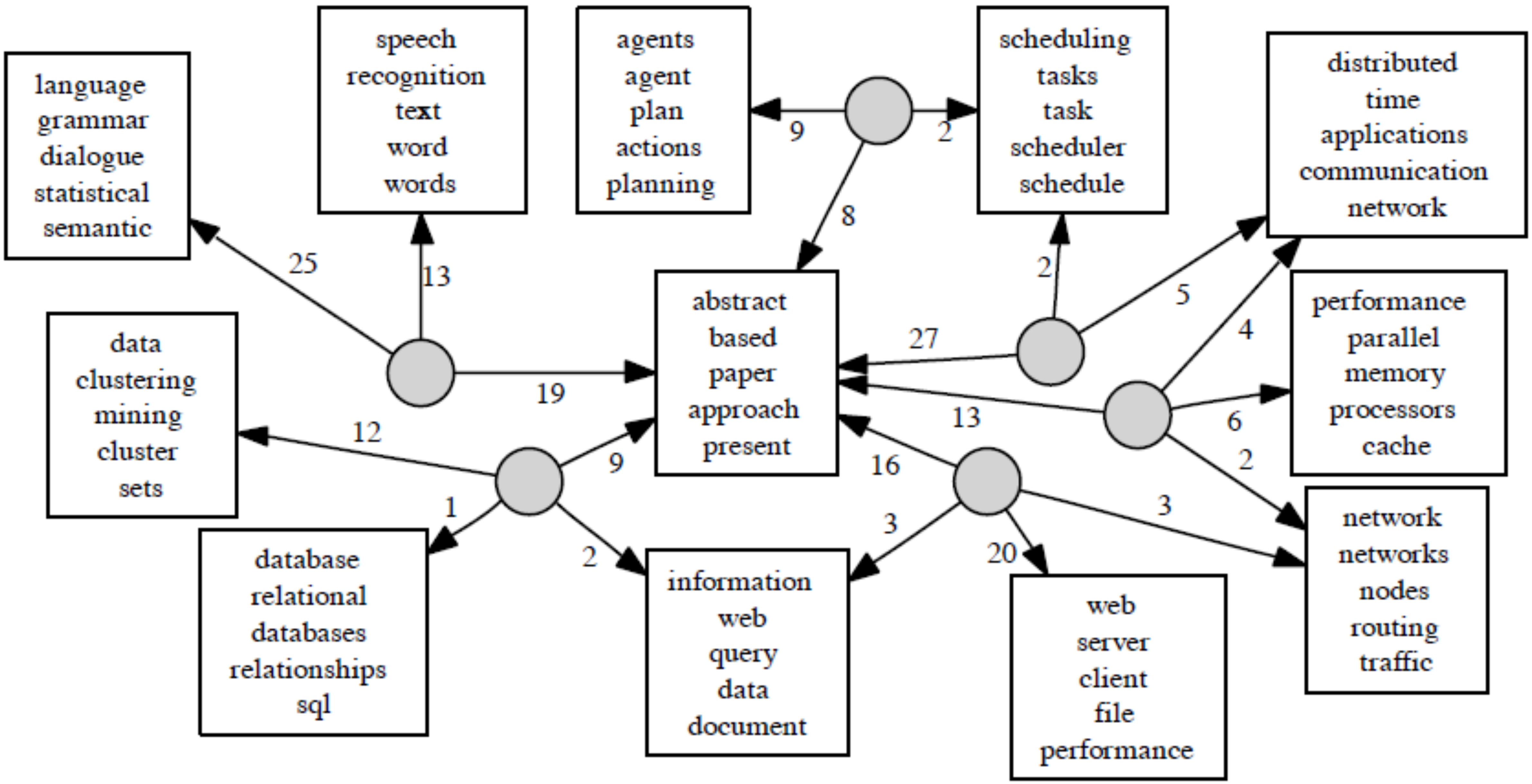
- Model the prior as a Directed Acyclic Graph
- Each document is modeled as multiple paths
- To sample a word, first select a path and then sample a word from the final topic
- The topics reside on the leaves of the tree



(a) Dirichlet Multinomial (b) LDA (c) Four-Level PAM (d) Arbitrary PAM

Li and McCallum 2006

Pachinko Allocation



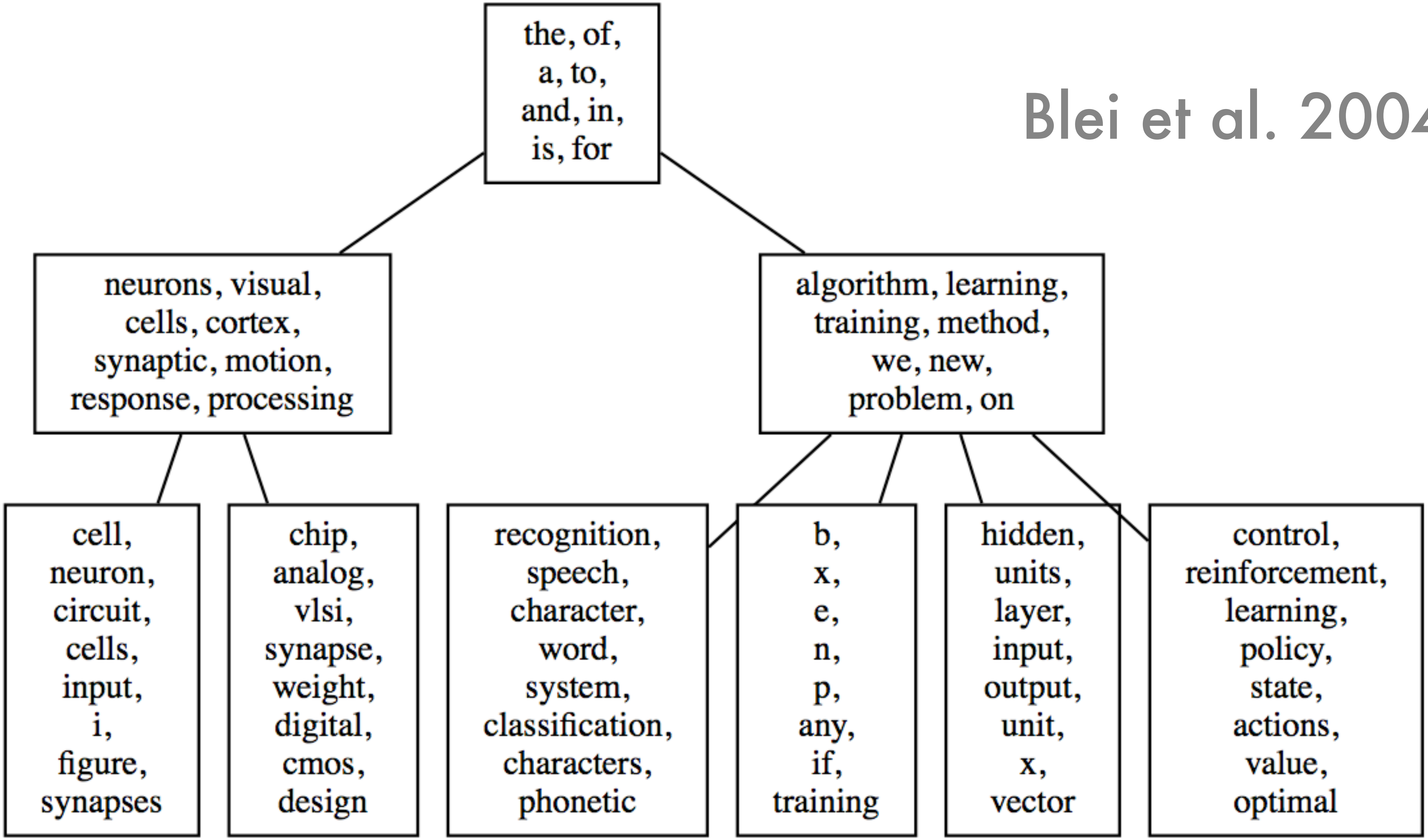
Li and McCallum 2006

Topic Hierarchies

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

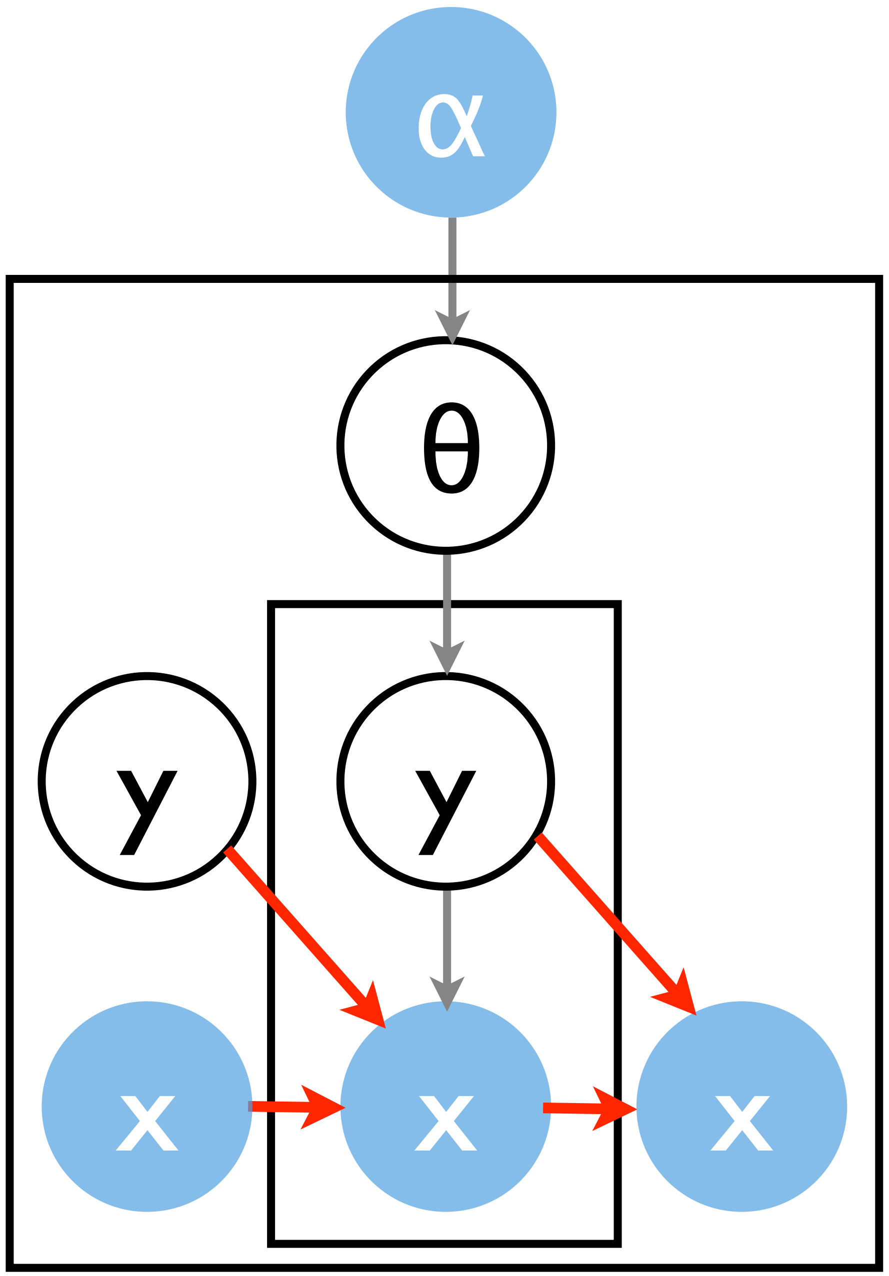
Topic Hierarchies

Blei et al. 2004



Topical n-grams

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



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

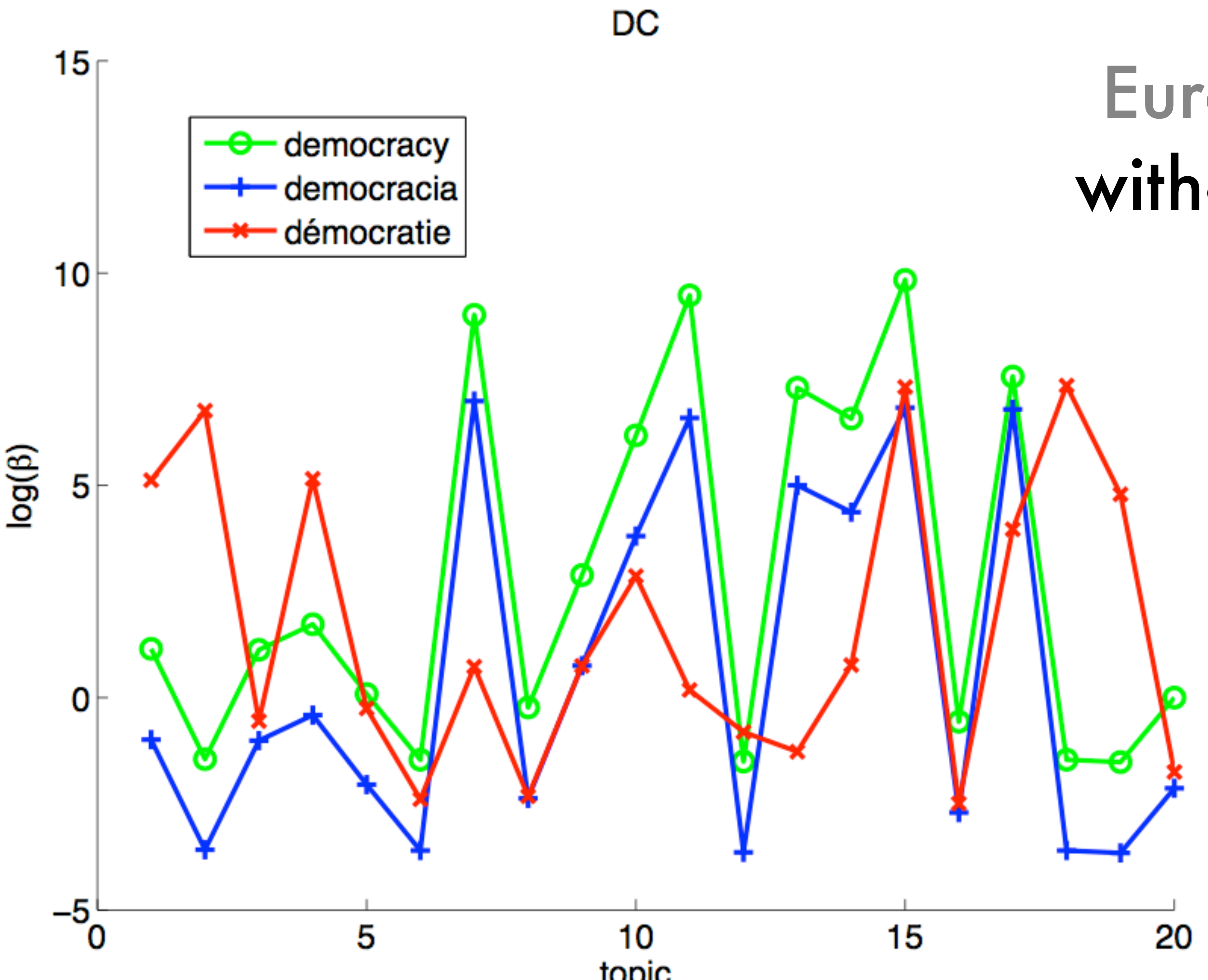
Topic n-grams

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

Side information

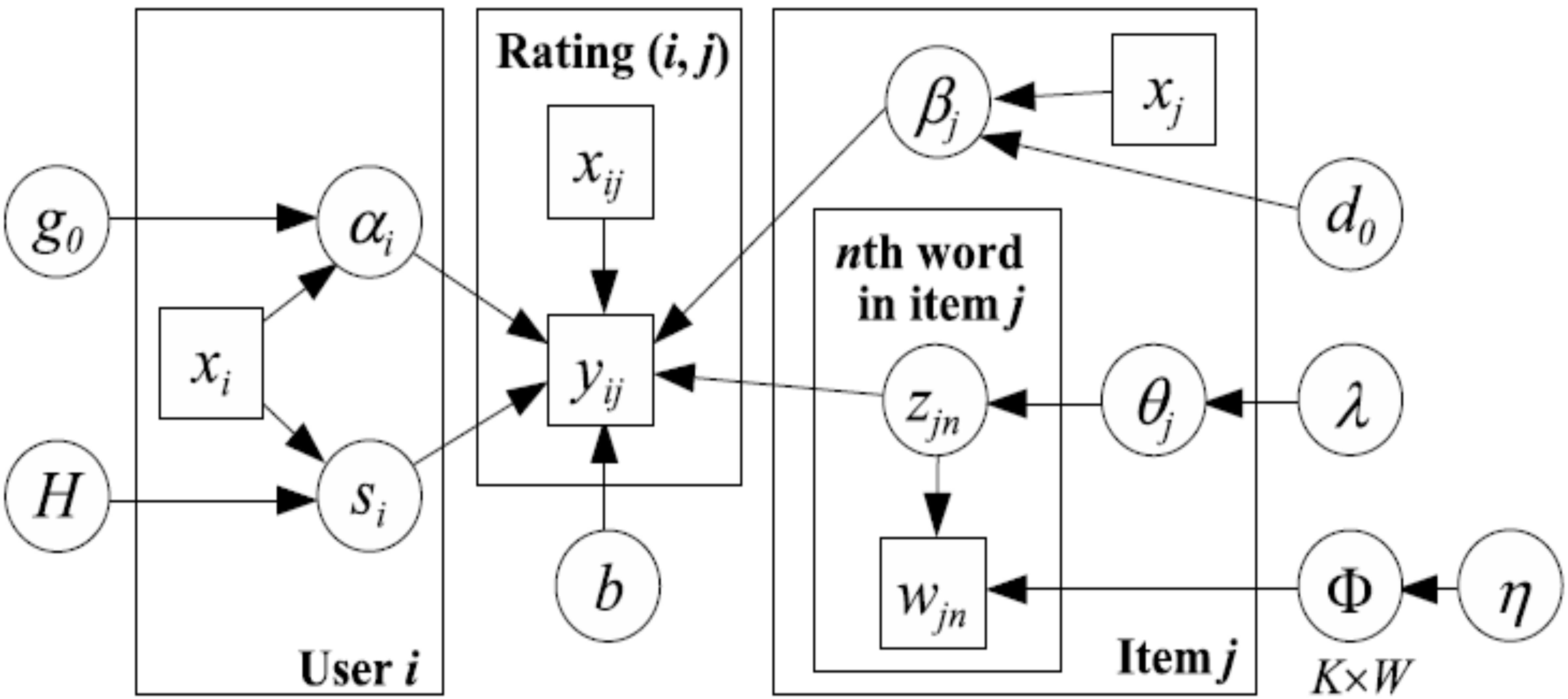
- Upstream conditioning (Mimno et al., 2008)
  - Document features are informative for topics
  - Estimate topic distribution e.g. based on authors, links, timestamp
- Downstream conditioning (Petterson et al., 2010)
  - Word features are informative on topics
  - Estimate topic distribution for words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

Downstream conditioning



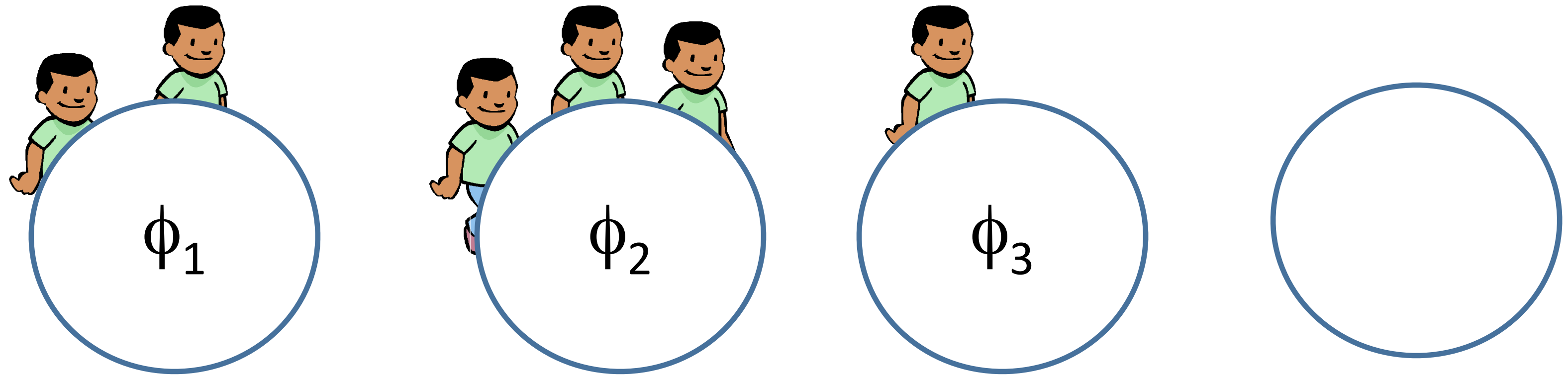
Europarl corpus
without alignment

Recommender Systems



Agarwal & Chen, 2010

Chinese Restaurant Process



Problem

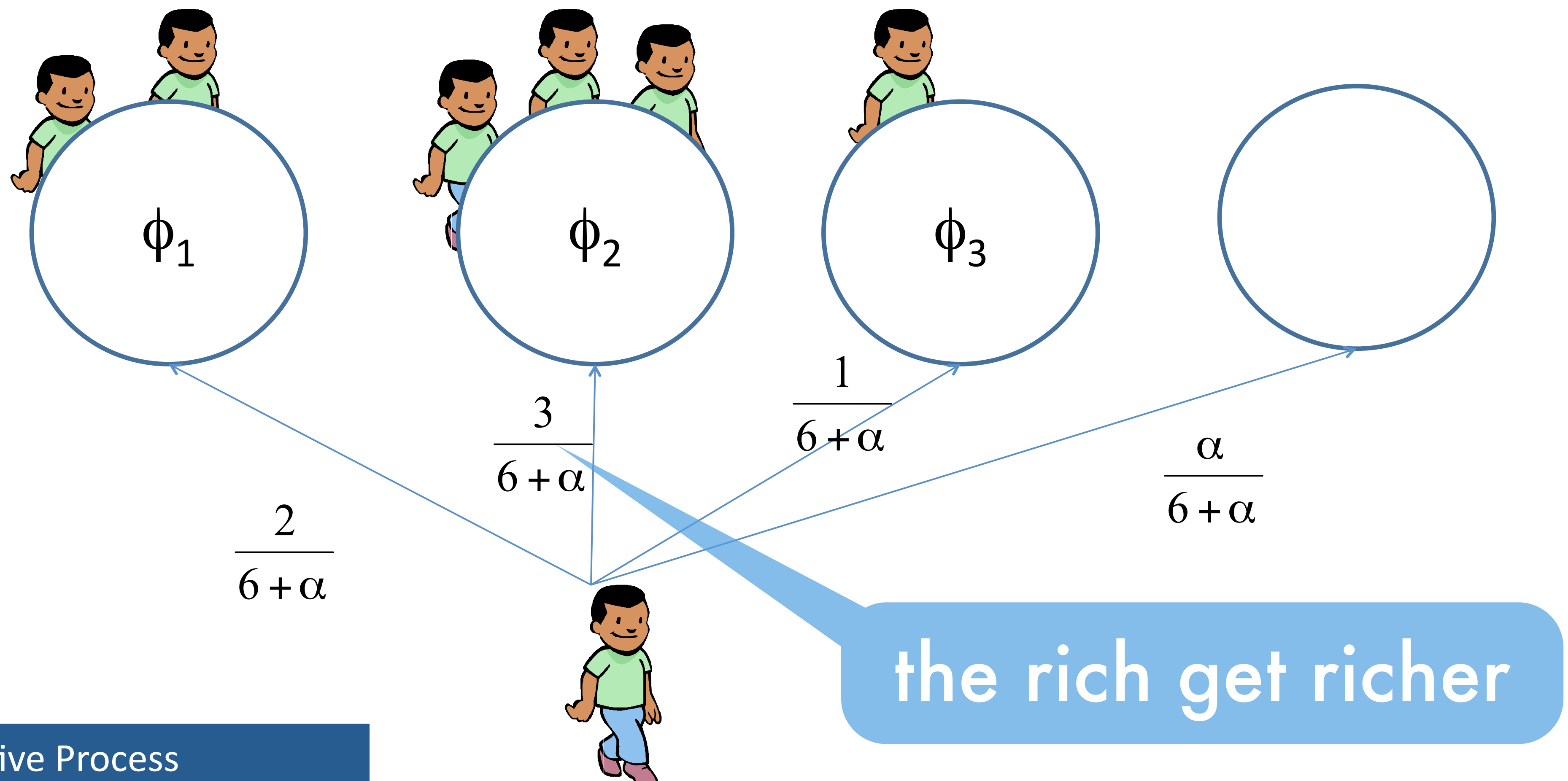
- How many clusters should we pick?
- How about a prior for infinitely many clusters?
- Finite model

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

- Infinite model
  Assume that the total smoother weight is constant

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

Chinese Restaurant Metaphor



Generative Process

-For data point x_i

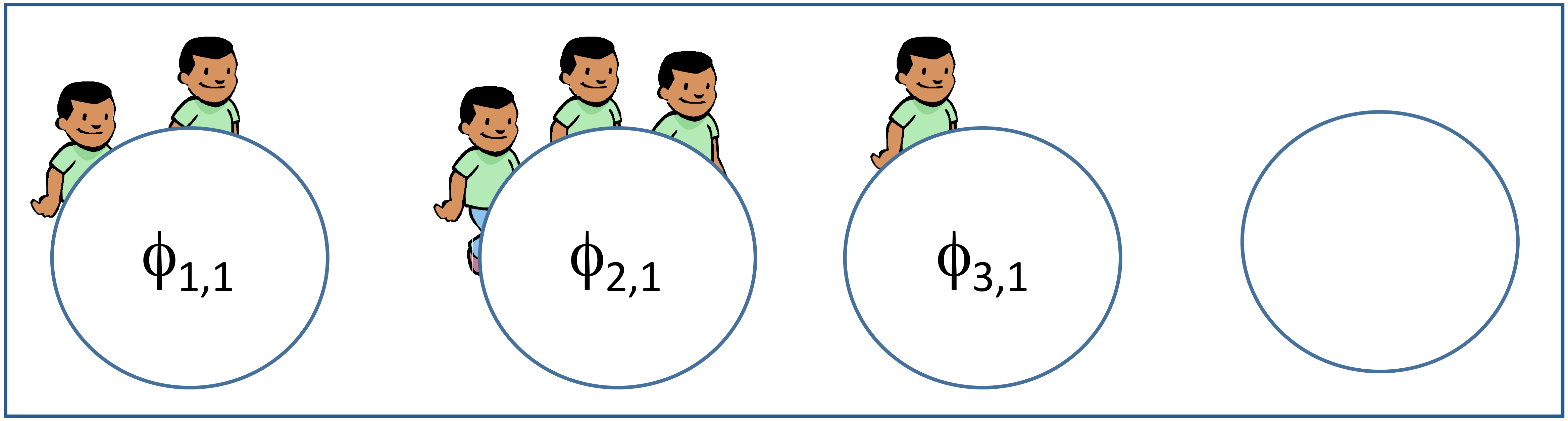
- Choose table $j \propto m_j$ and Sample $x_i \sim f(\phi_j)$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1})$

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

Evolutionary Clustering

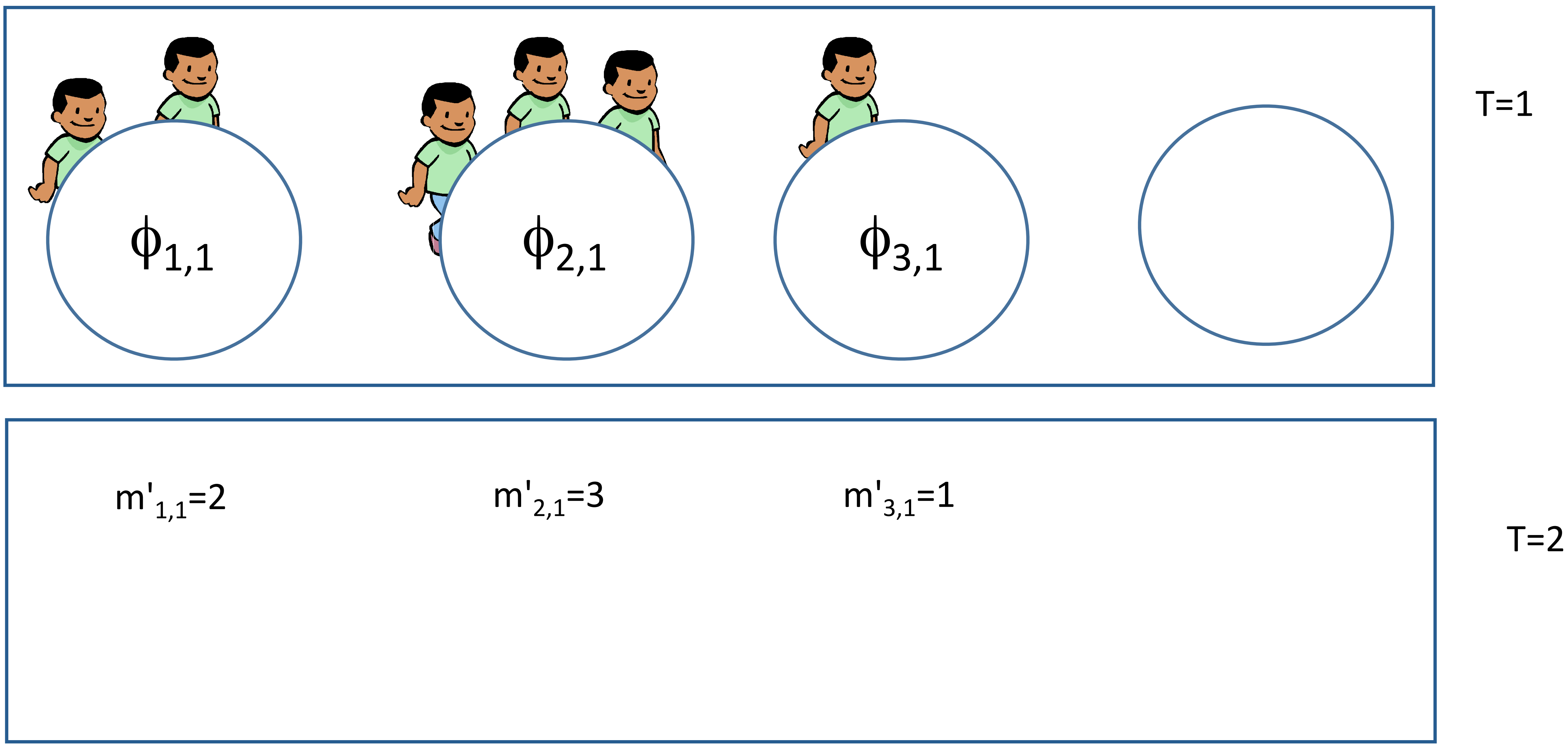
- Time series of objects, e.g. news stories
- Stories appear / disappear
- Want to keep track of clusters automatically

Recurrent Chinese Restaurant Process

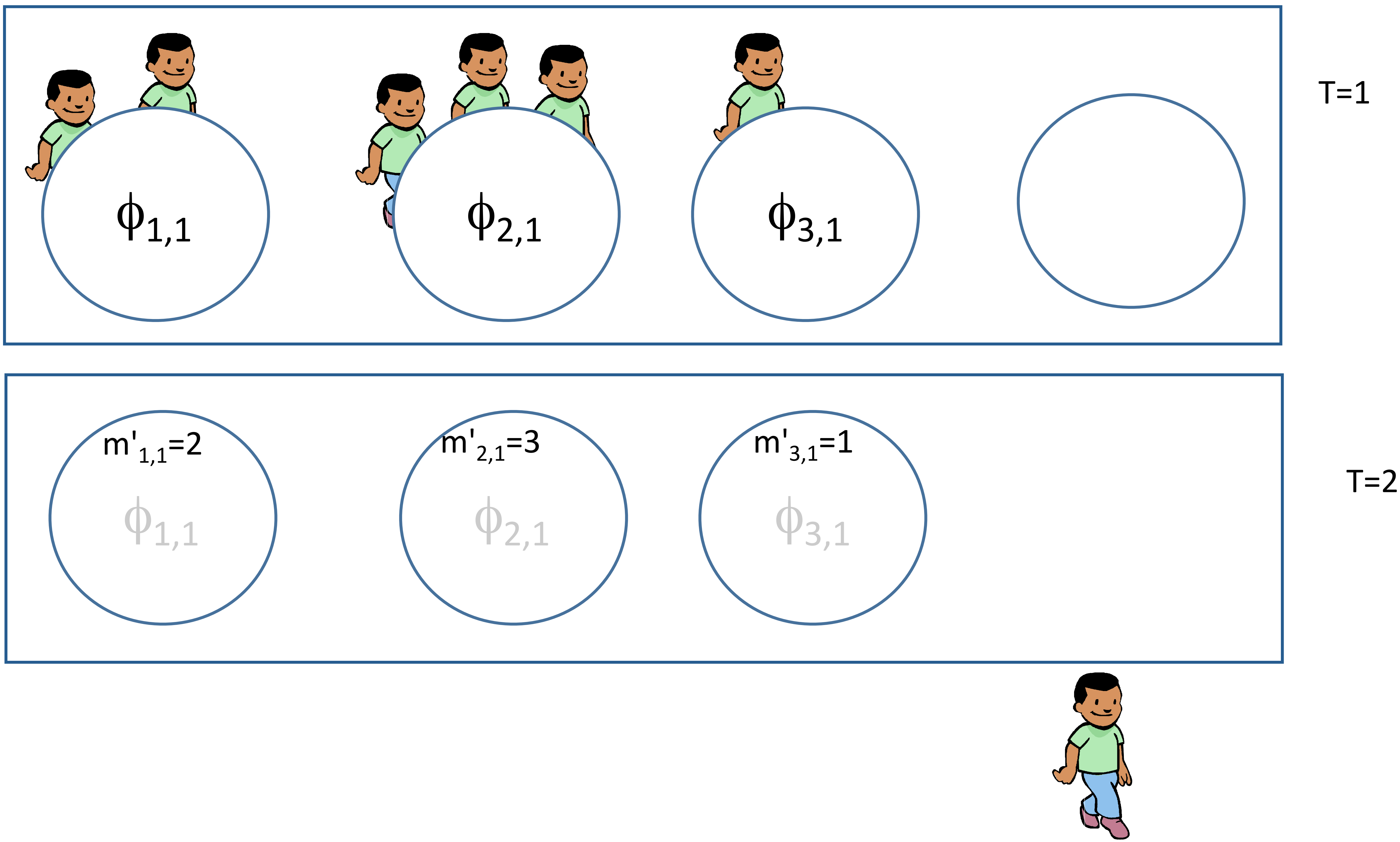


T=1

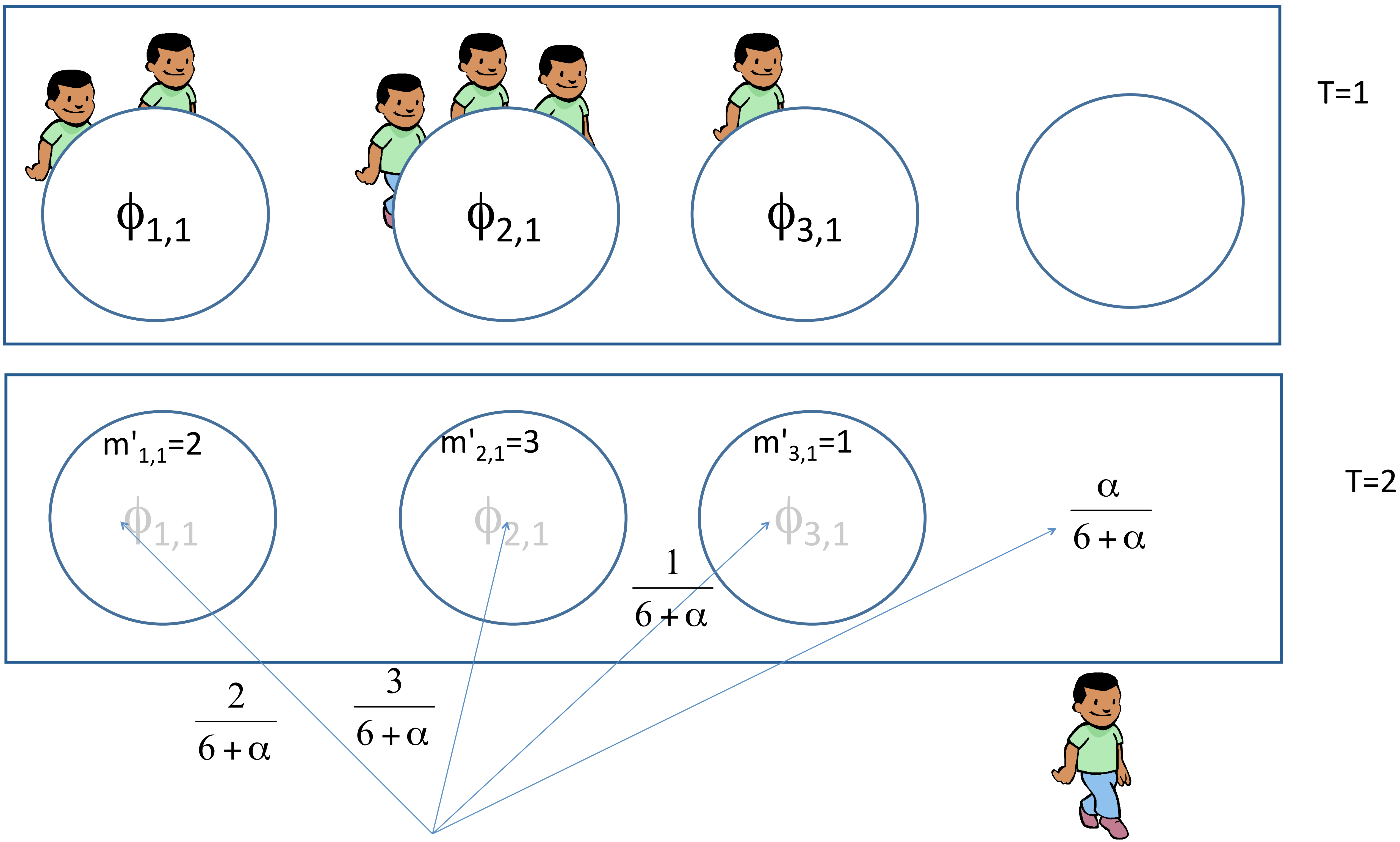
Recurrent Chinese Restaurant Process



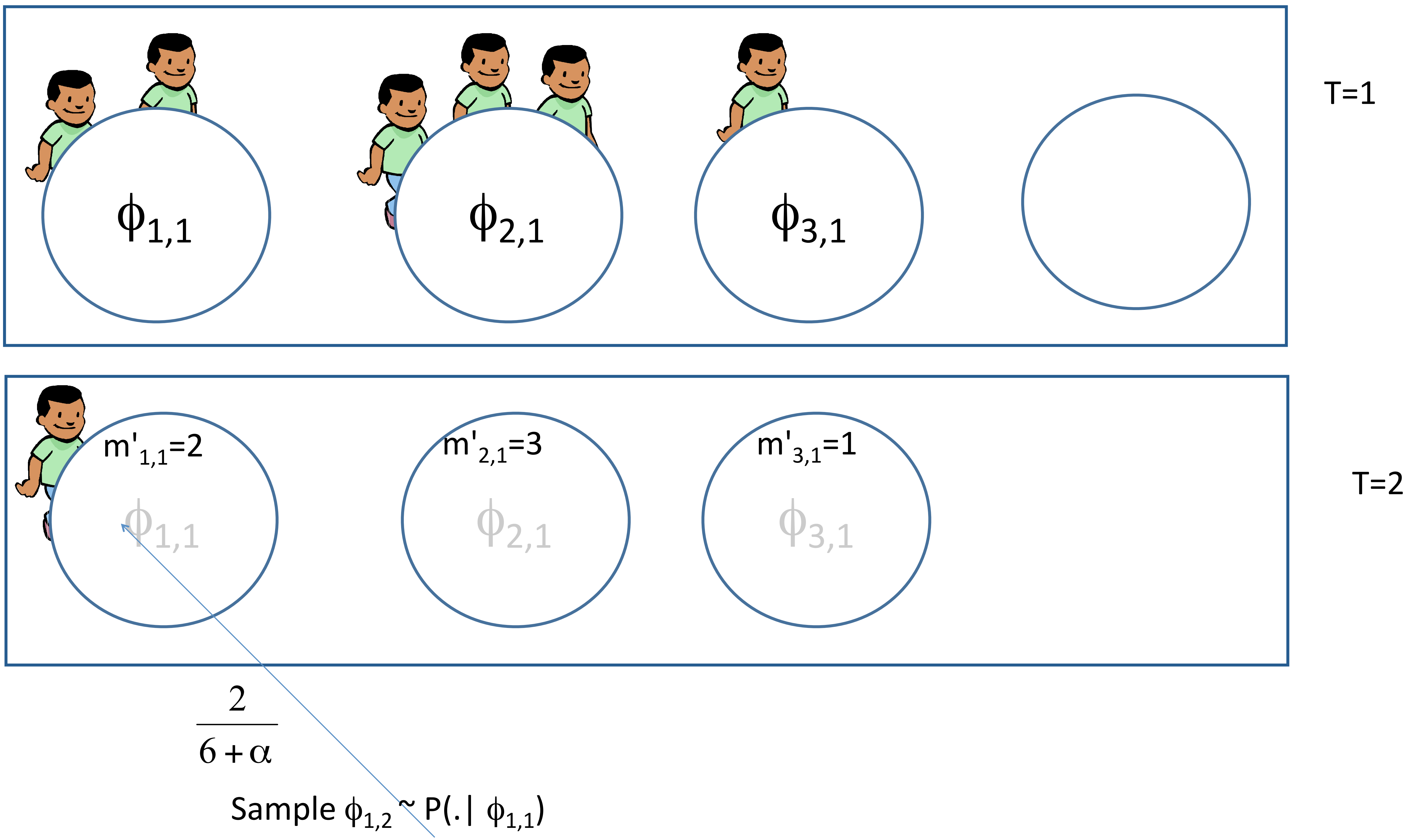
Recurrent Chinese Restaurant Process



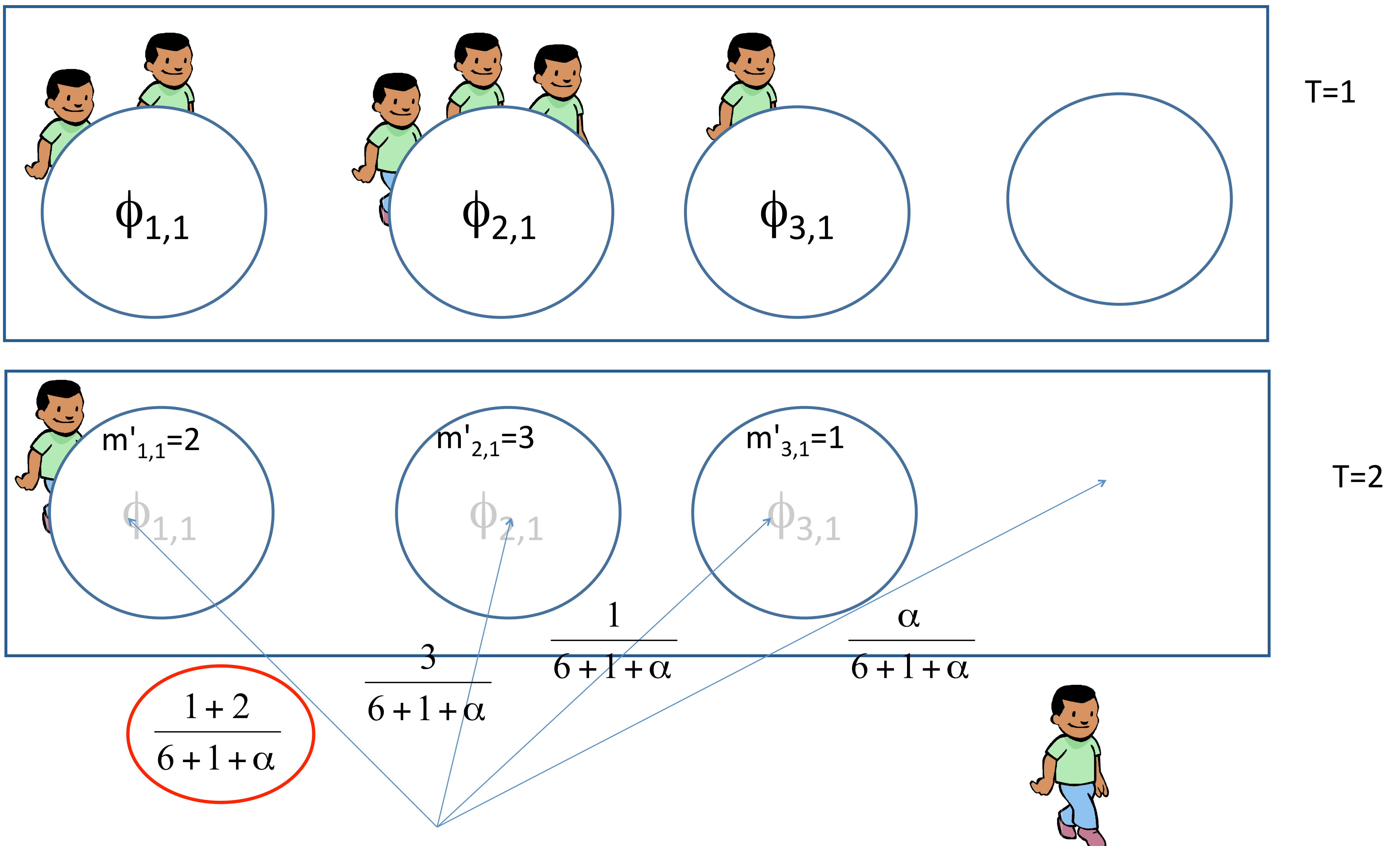
Recurrent Chinese Restaurant Process



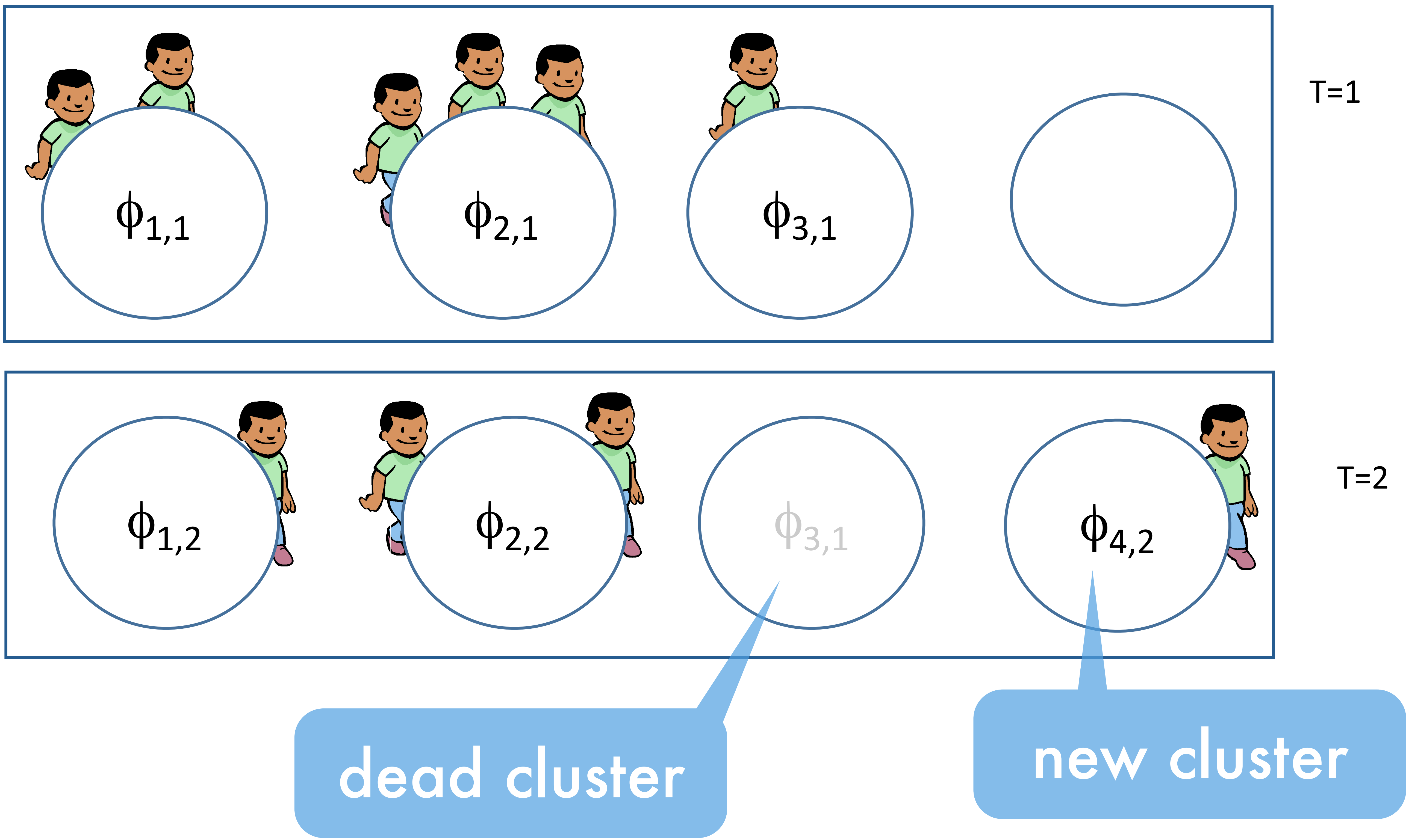
Recurrent Chinese Restaurant Process



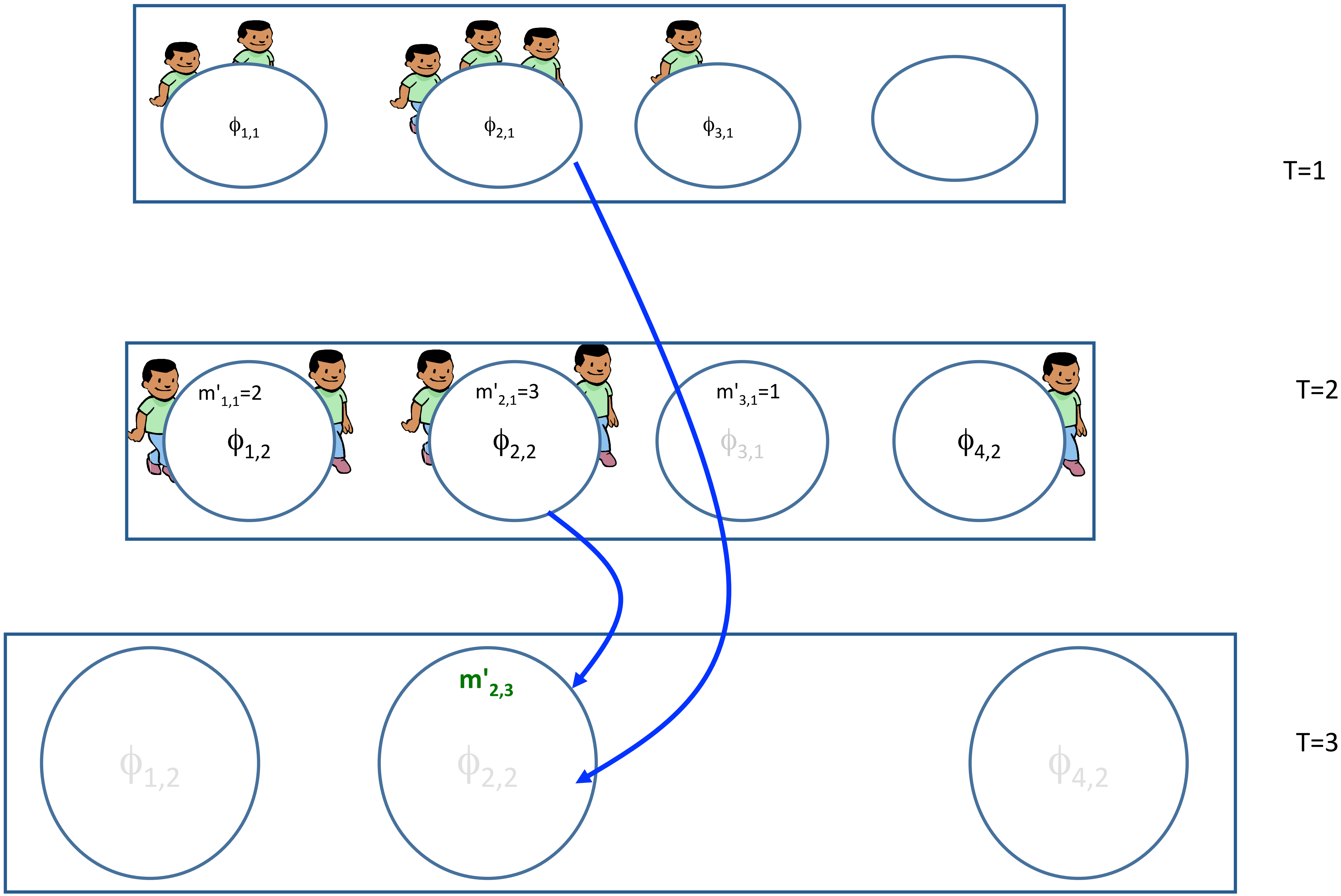
Recurrent Chinese Restaurant Process



Recurrent Chinese Restaurant Process



Longer History



TDPM Generative Power

DPM

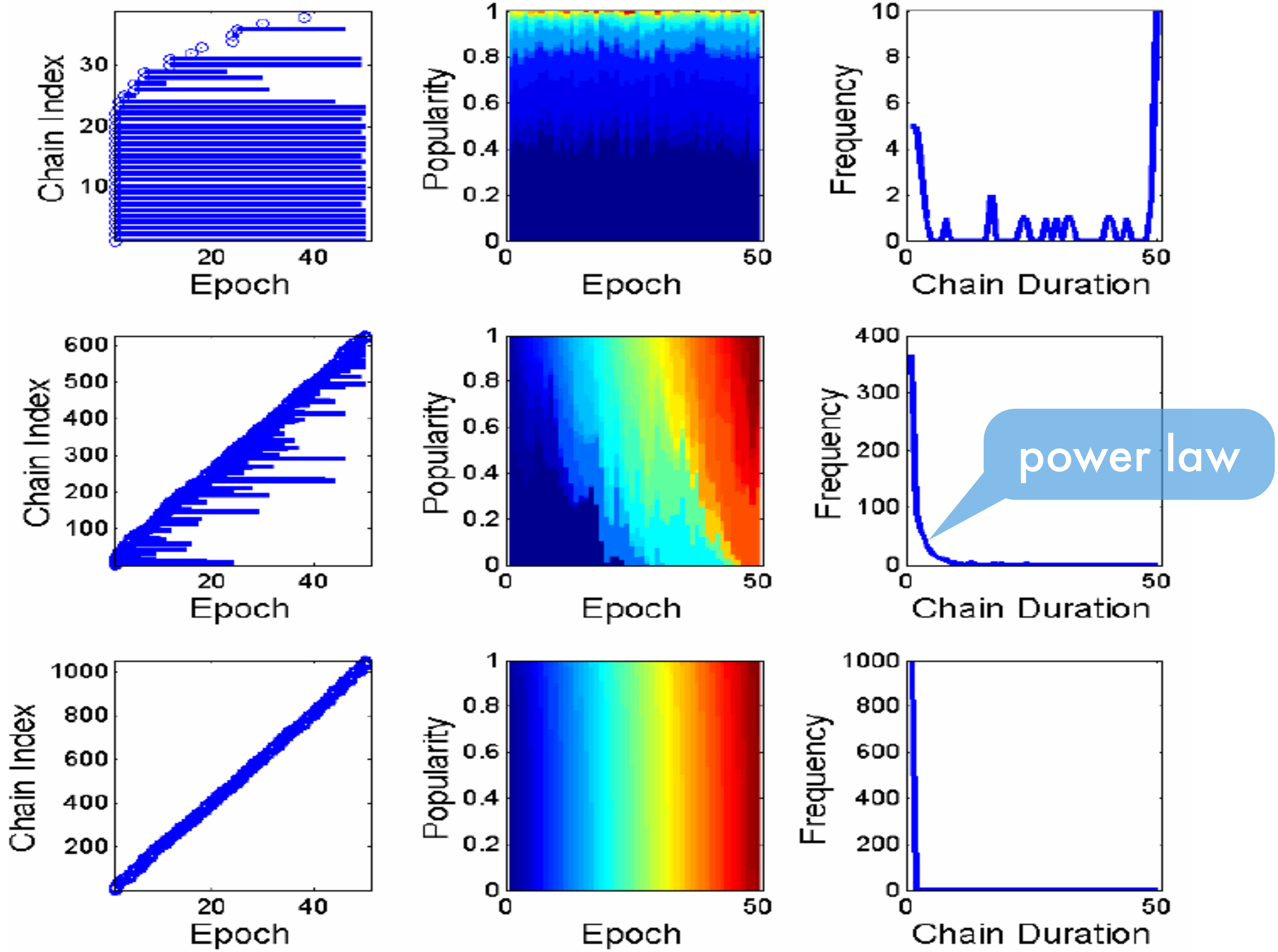
W= ∞
λ = ∞

TDPM

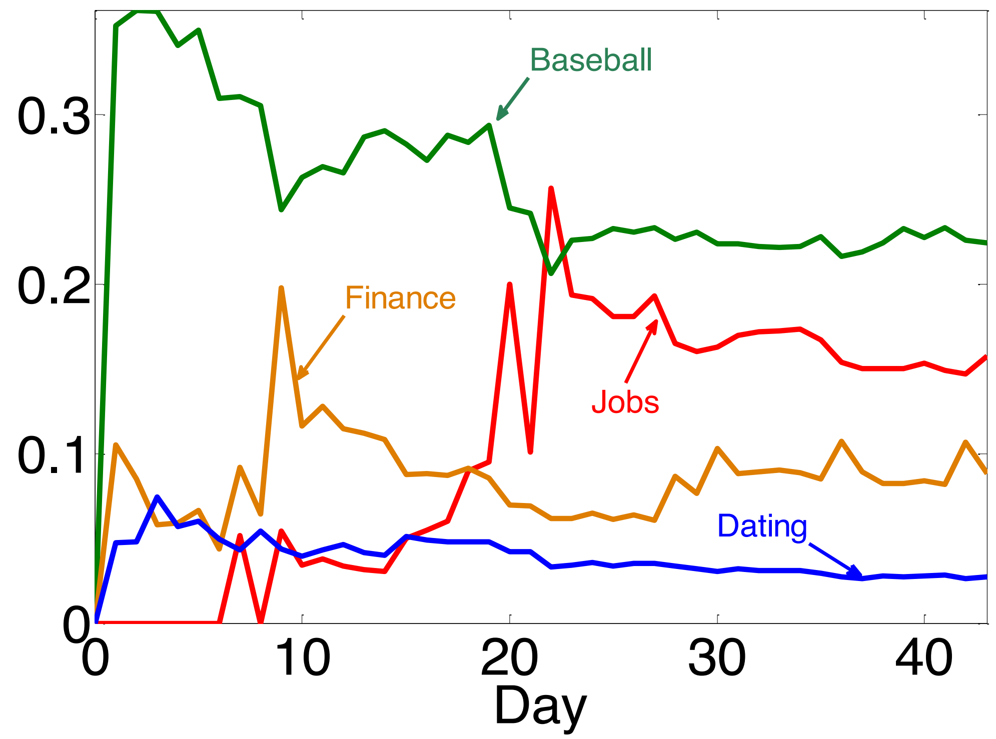
W=4
λ = .4

Independent DPMs

W= 0
λ = ? (any)

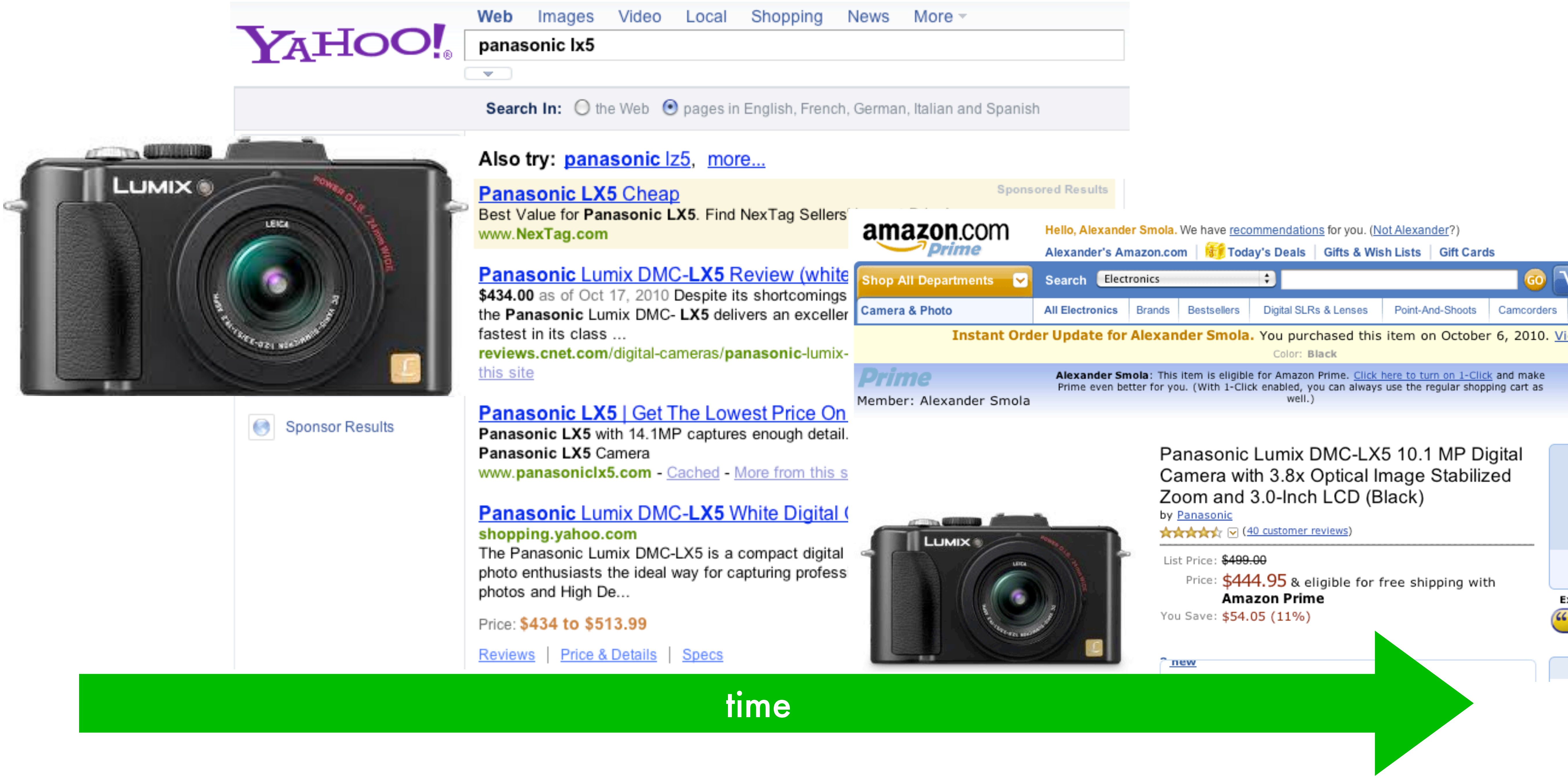


User modeling

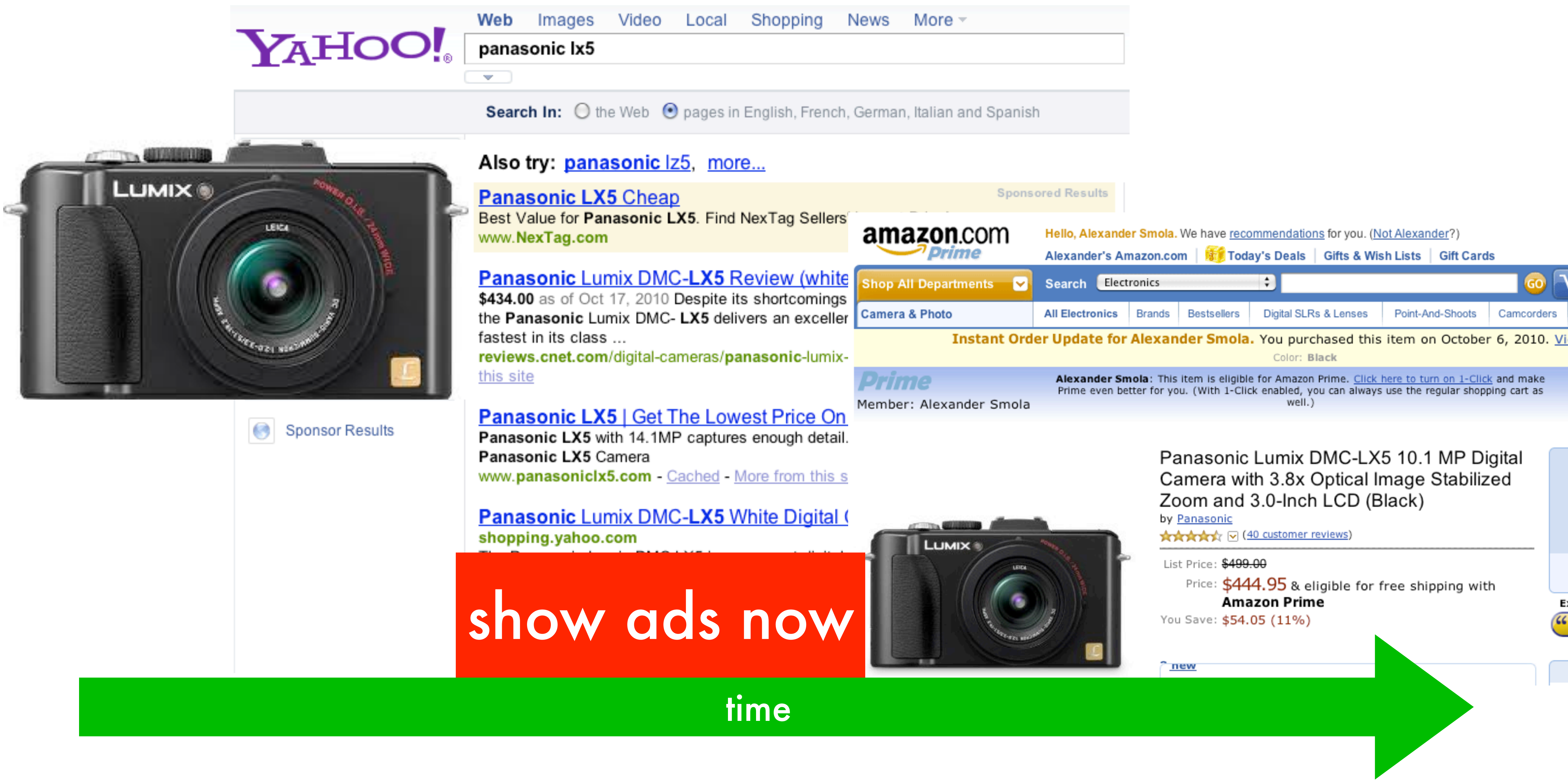


Buying a camera

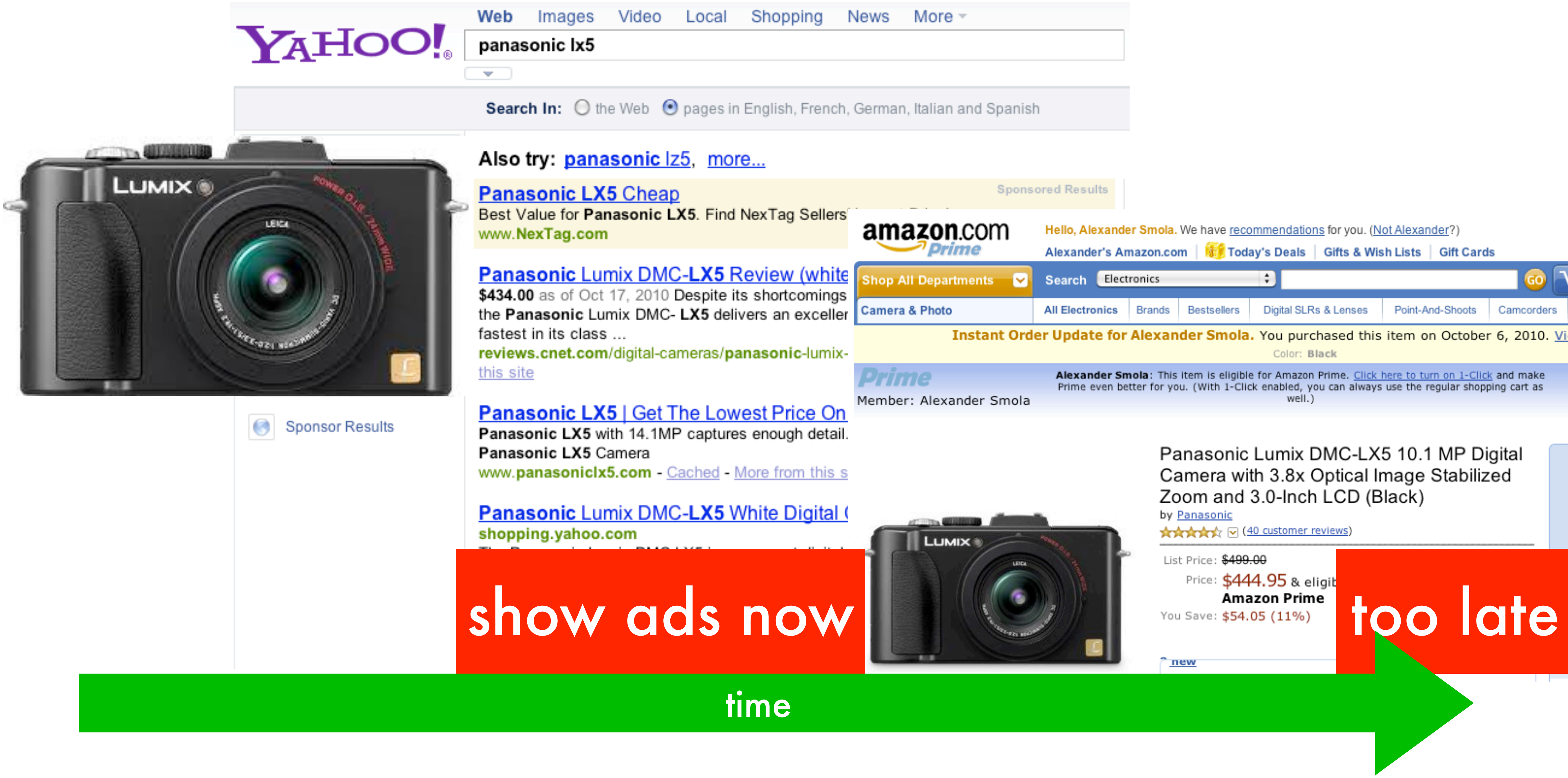
Buying a camera



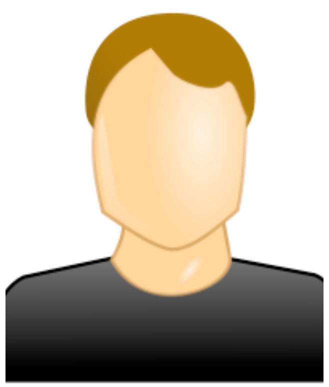
Buying a camera



Buying a camera



Car
Deals
van

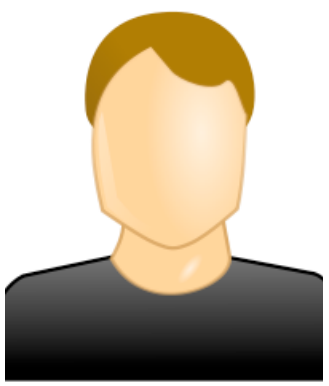


job
Hiring
diet

Car
Deals
van

Auto
Price
Used
inspection

Movies
Theatre
Art
gallery



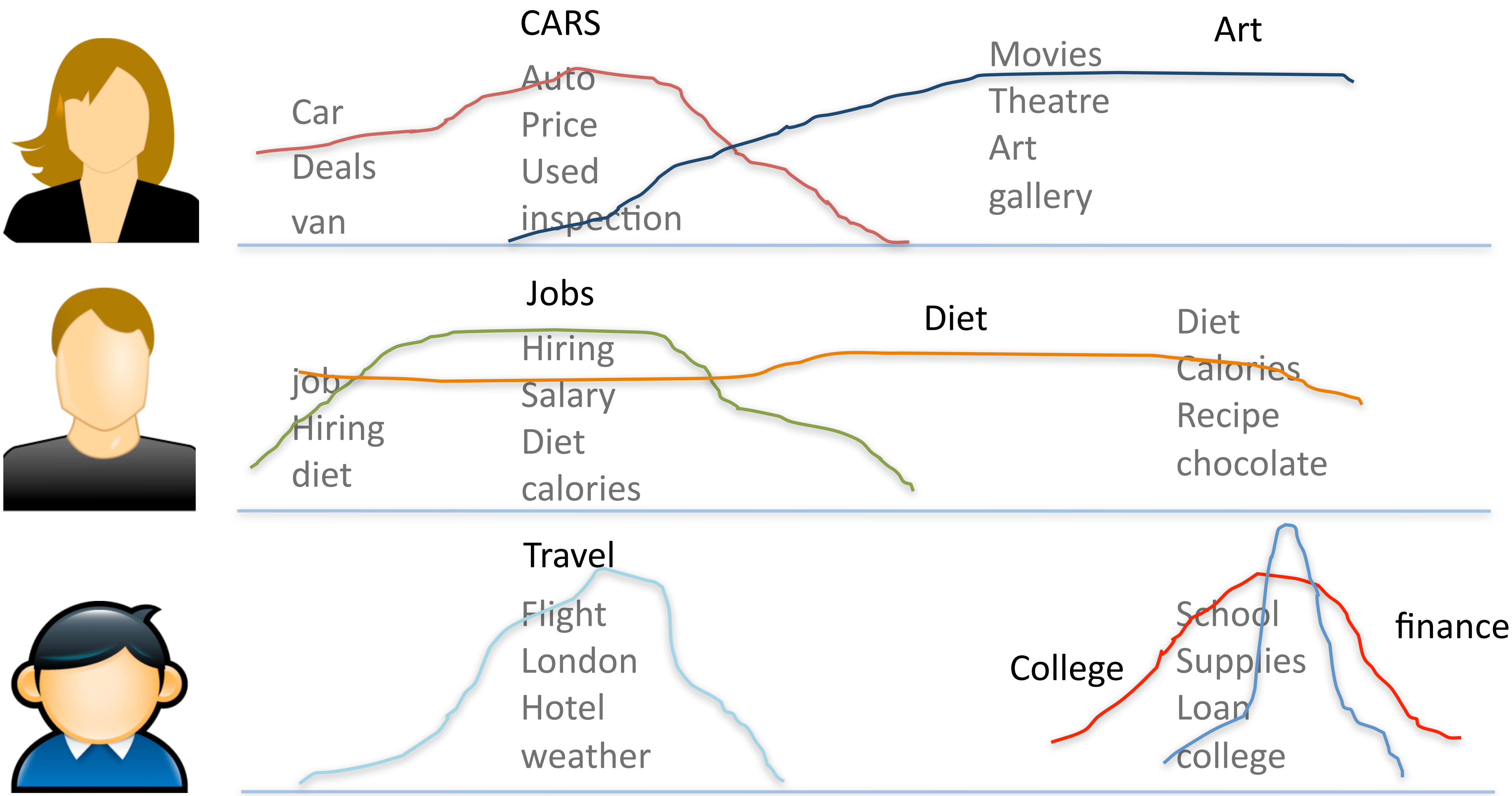
job
Hiring
diet

Hiring
Salary
Diet
calories

Diet
Calories
Recipe
chocolate

Flight
London
Hotel
weather

School
Supplies
Loan
college


CARS
Art
Car
Deals
van
Auto
Price
Used
inspection
Movies
Theatre
Art
gallery
Jobs
Diet
job
Hiring
diet
Hiring
Salary
Diet
calories
Diet
Calories
Recipe
chocolate
Travel
Flight
London
Hotel
weather
College
School
Supplies
Loan
college
finance

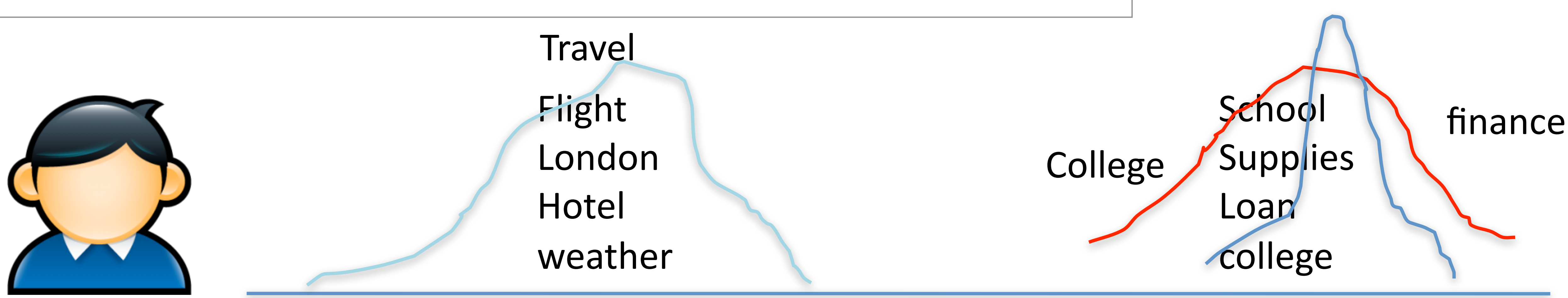
User modeling

Input

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

Output

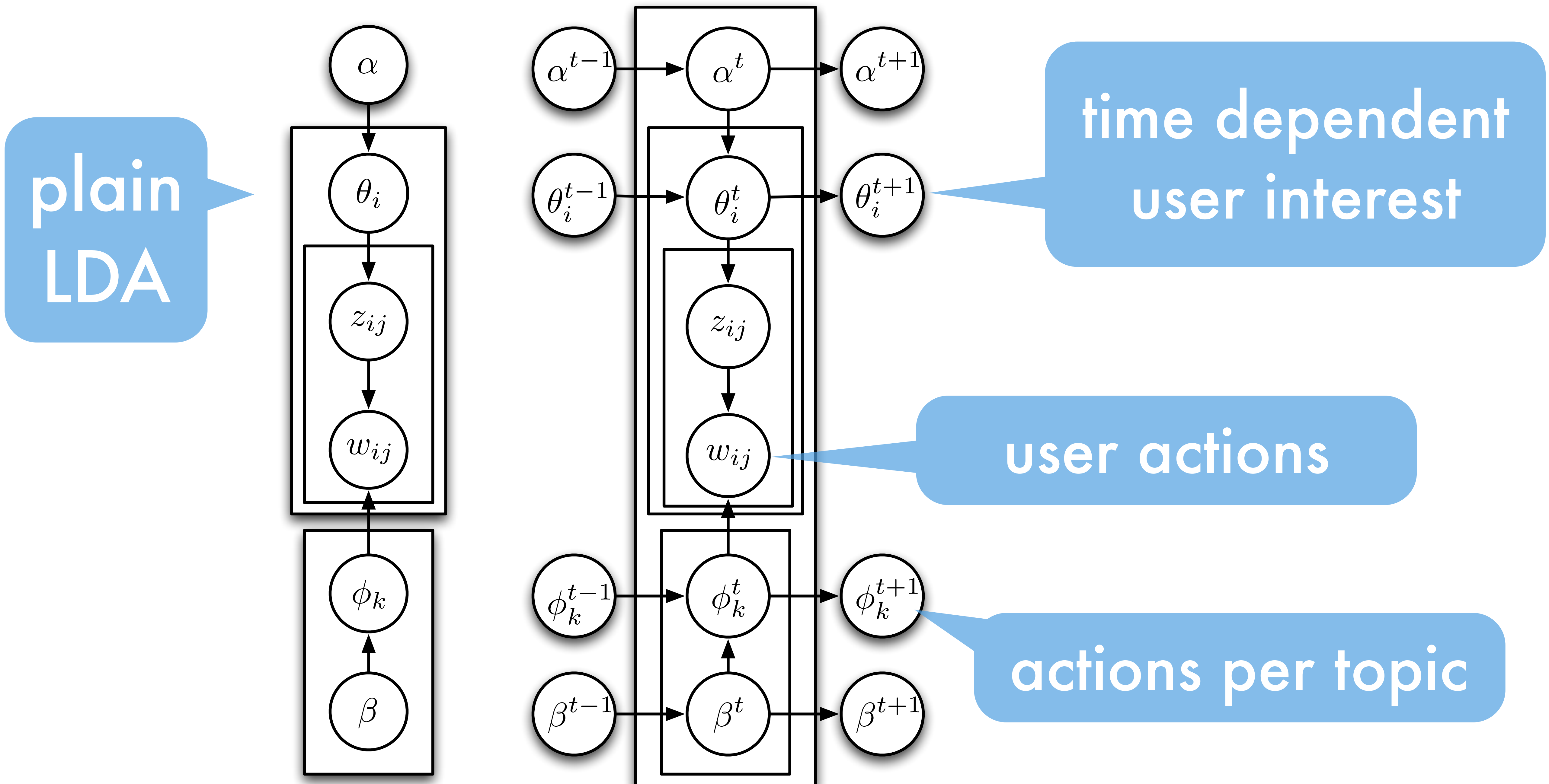
- Users' daily distribution over intents
- Dynamic intent representation

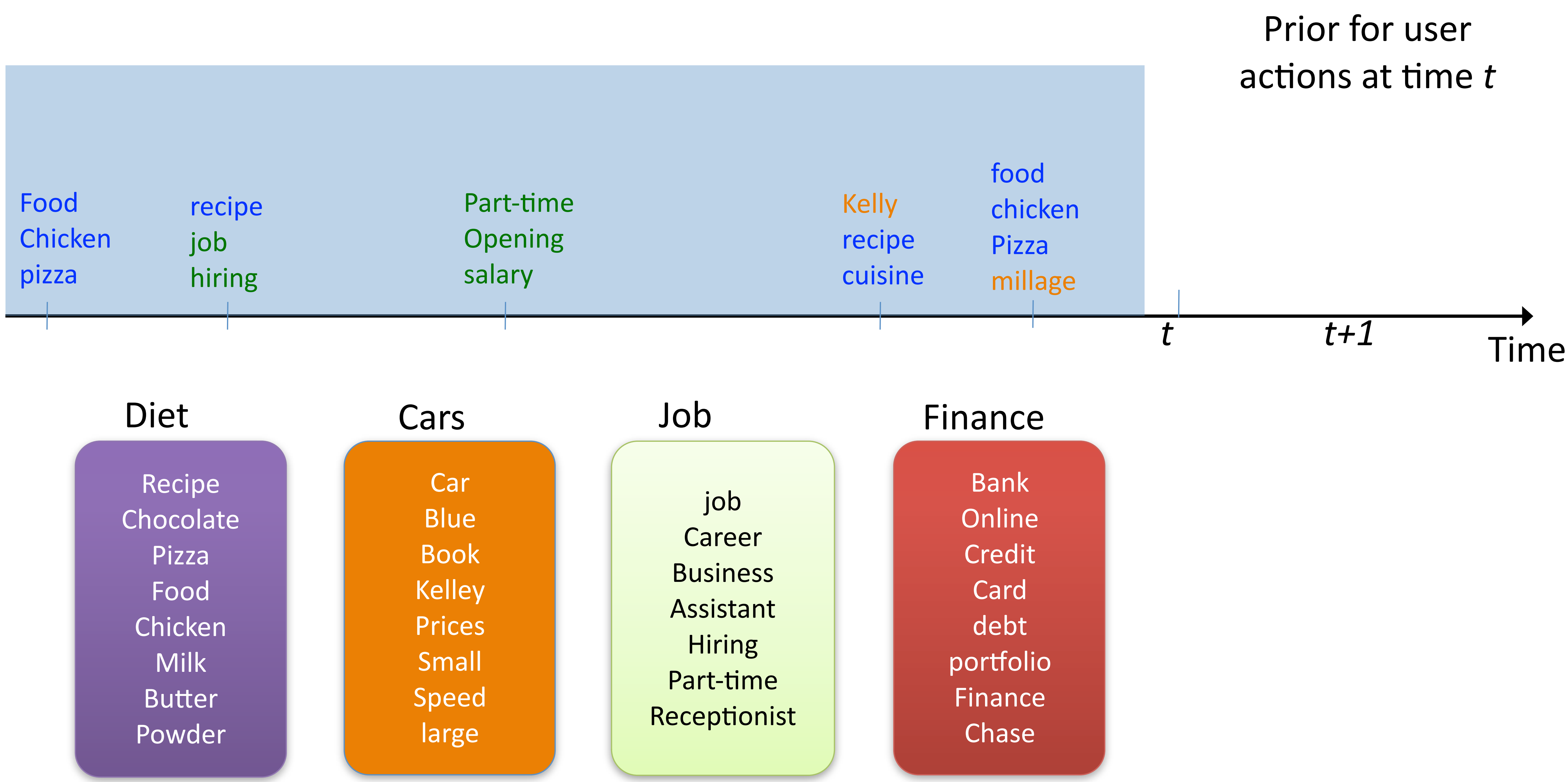


Time dependent models

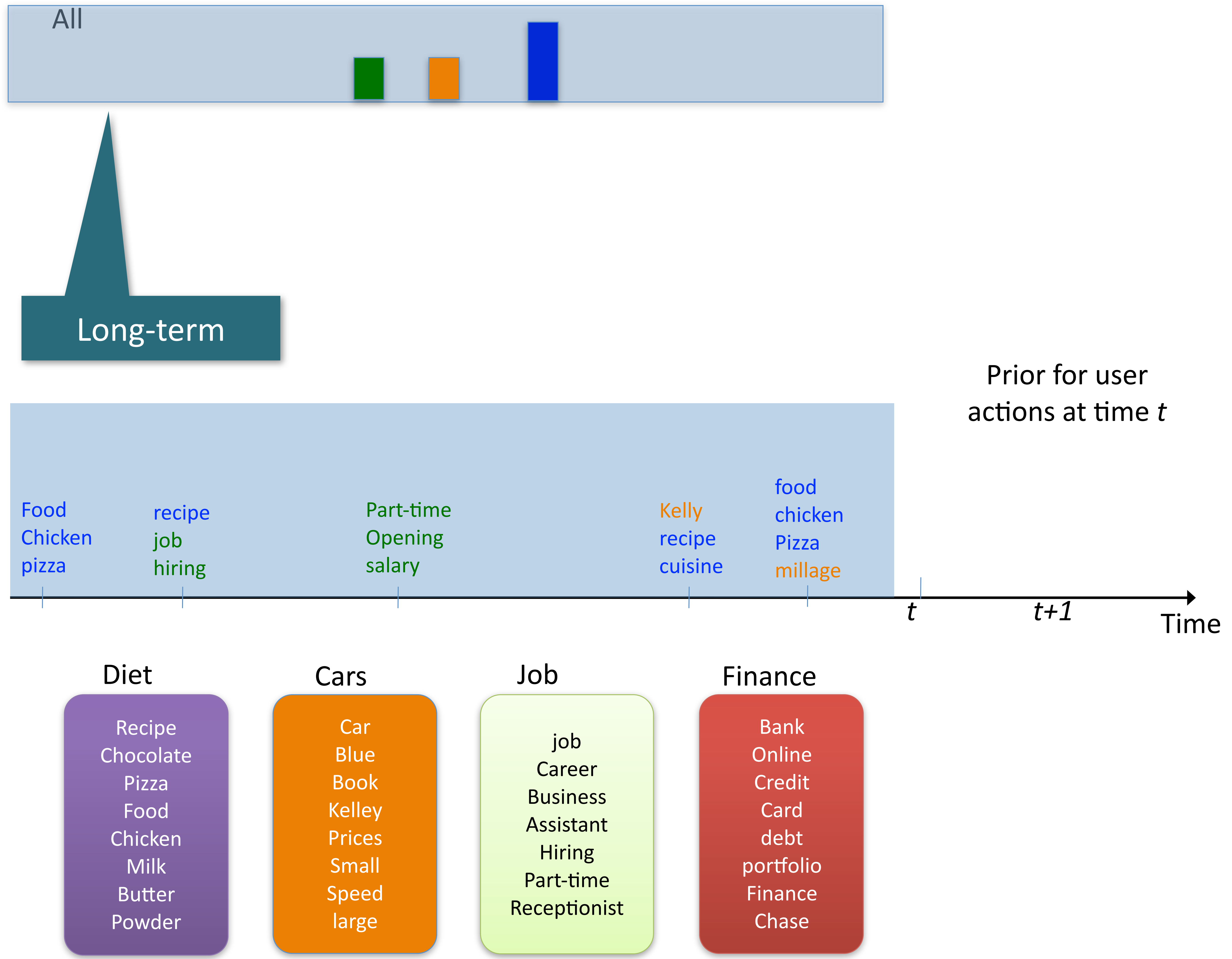
- LDA for topical model of users where
  - User interest distribution changes over time
  - Topics change over time
- This is like a Kalman filter except that
  - Don't know what to track (a priori)
  - Can't afford a Rauch-Tung-Striebel smoother
  - Much more messy than plain LDA

Graphical Model

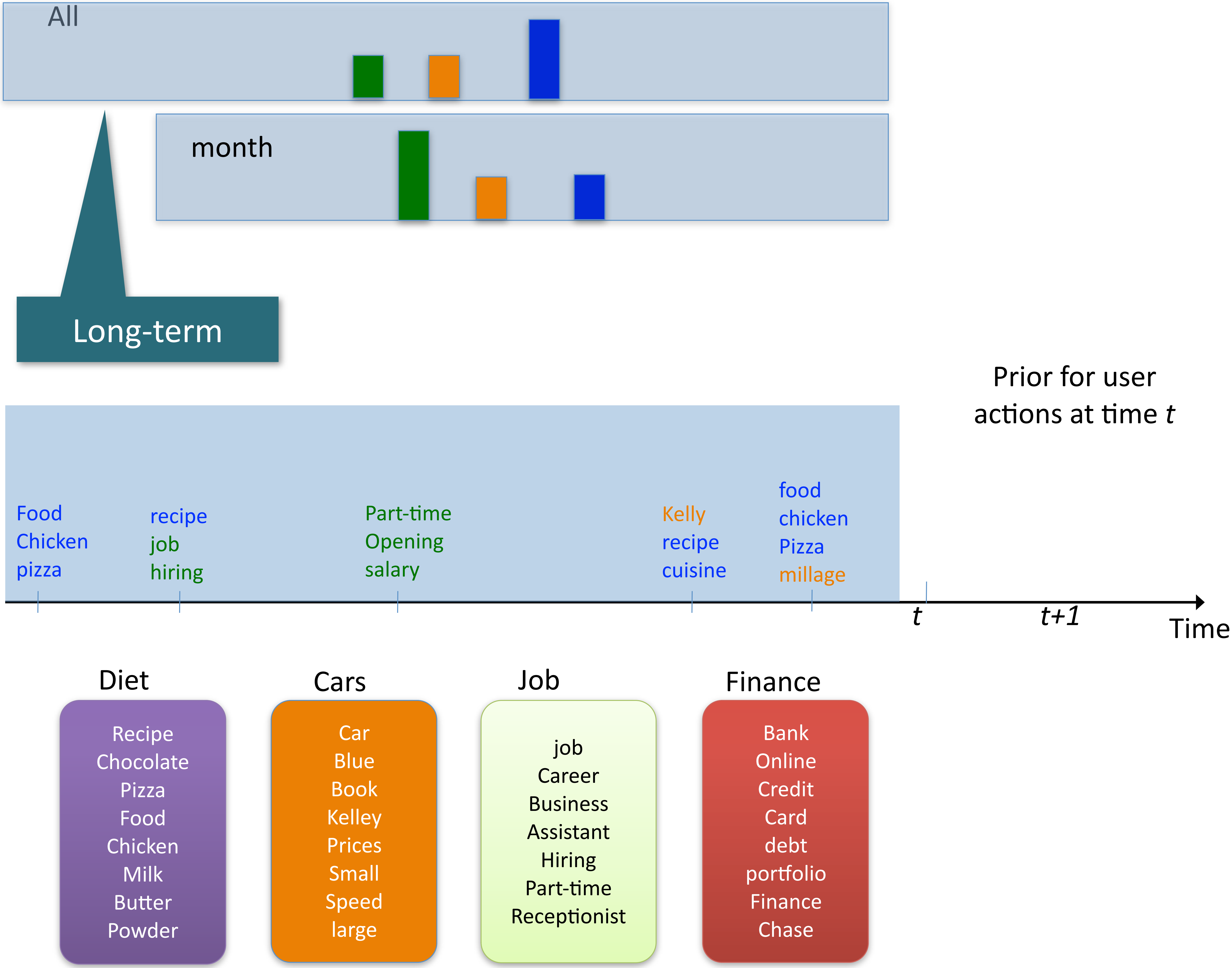




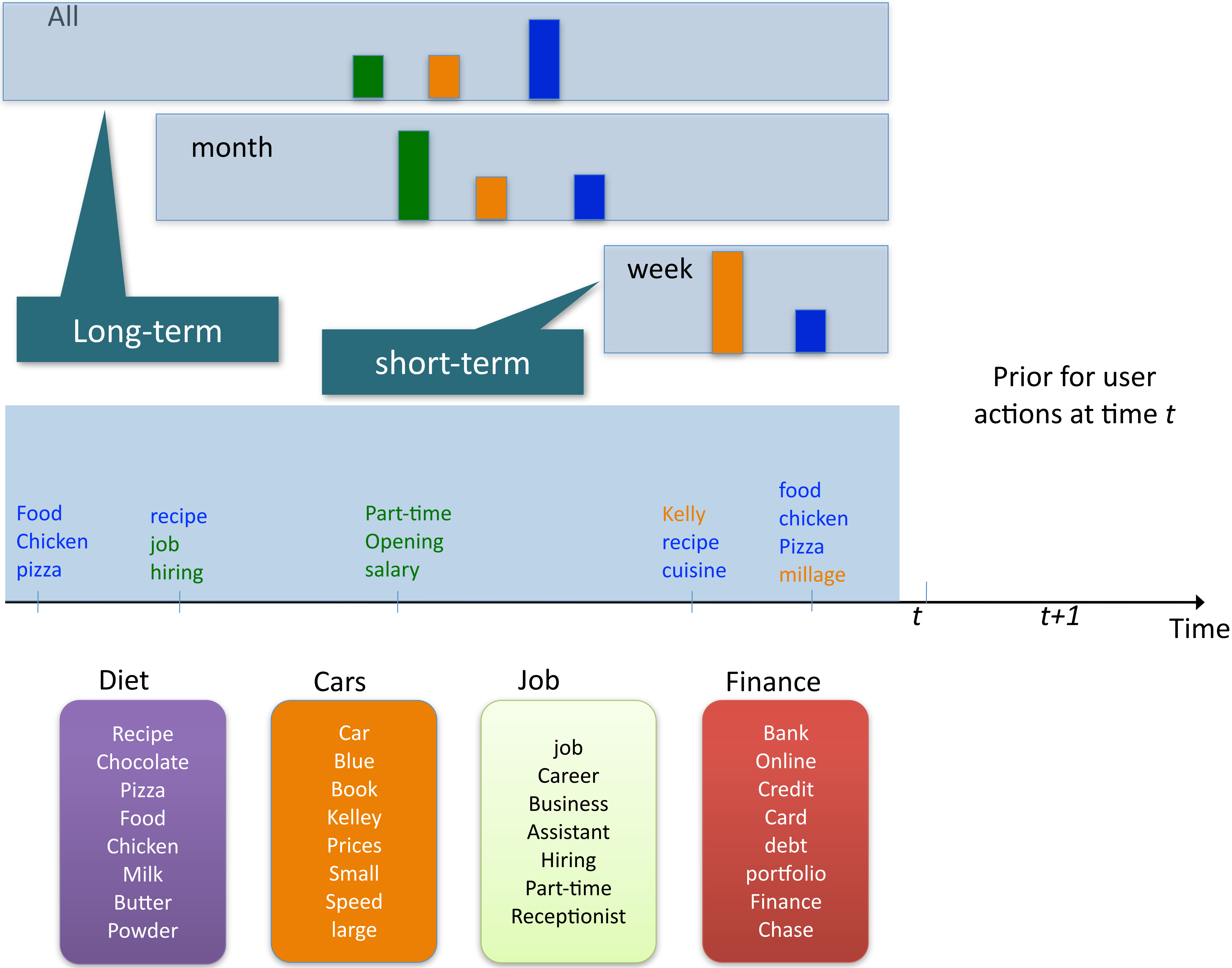
Prior for user actions at time t
Food
Chicken
pizza
recipe
job
hiring
Part-time
Opening
salary
Kelly
recipe
cuisine
food
chicken
Pizza
millage
t
t+1
Time
Diet
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder
Cars
Car
Blue
Book
Kelley
Prices
Small
Speed
large
Job
job
Career
Business
Assistant
Hiring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

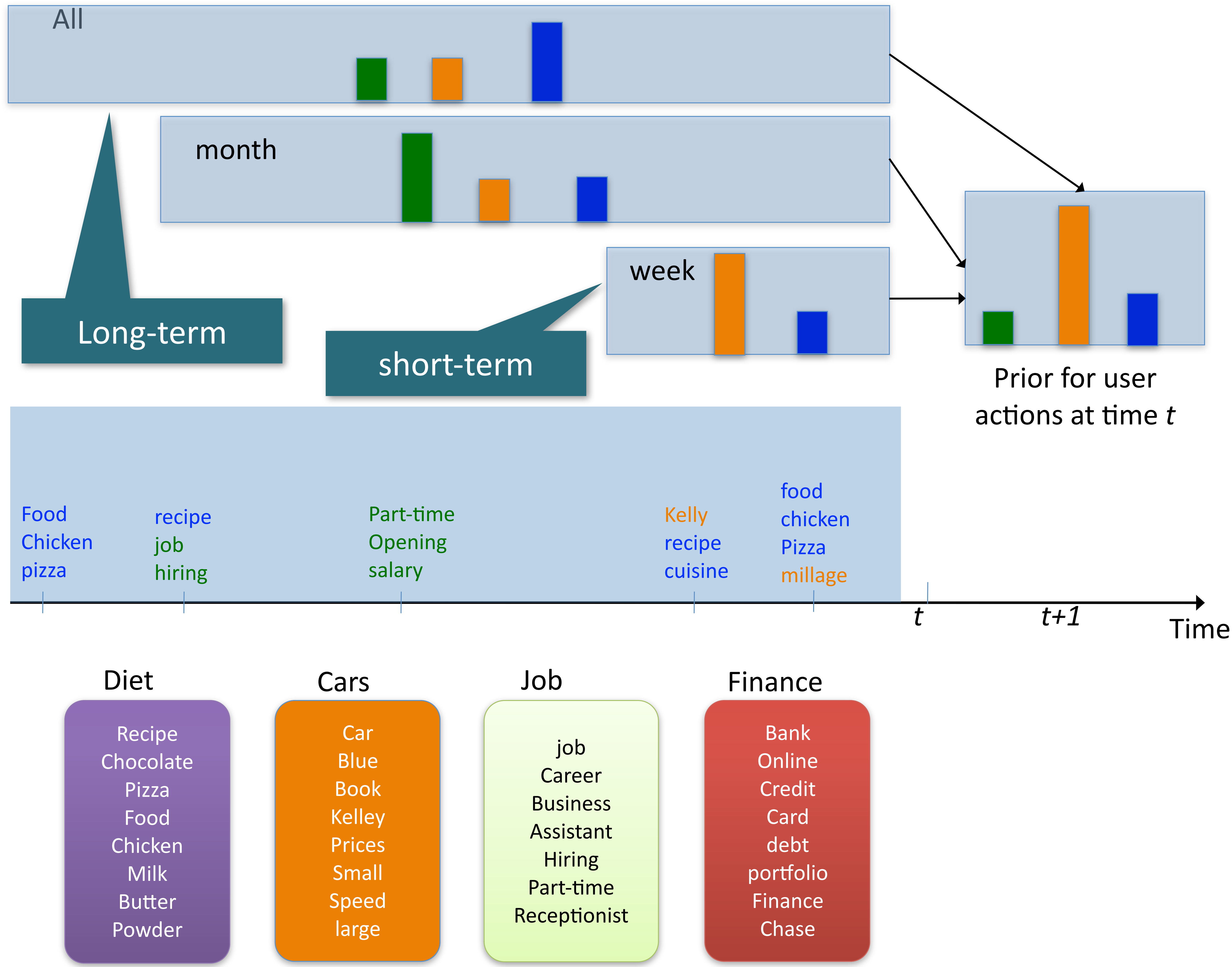


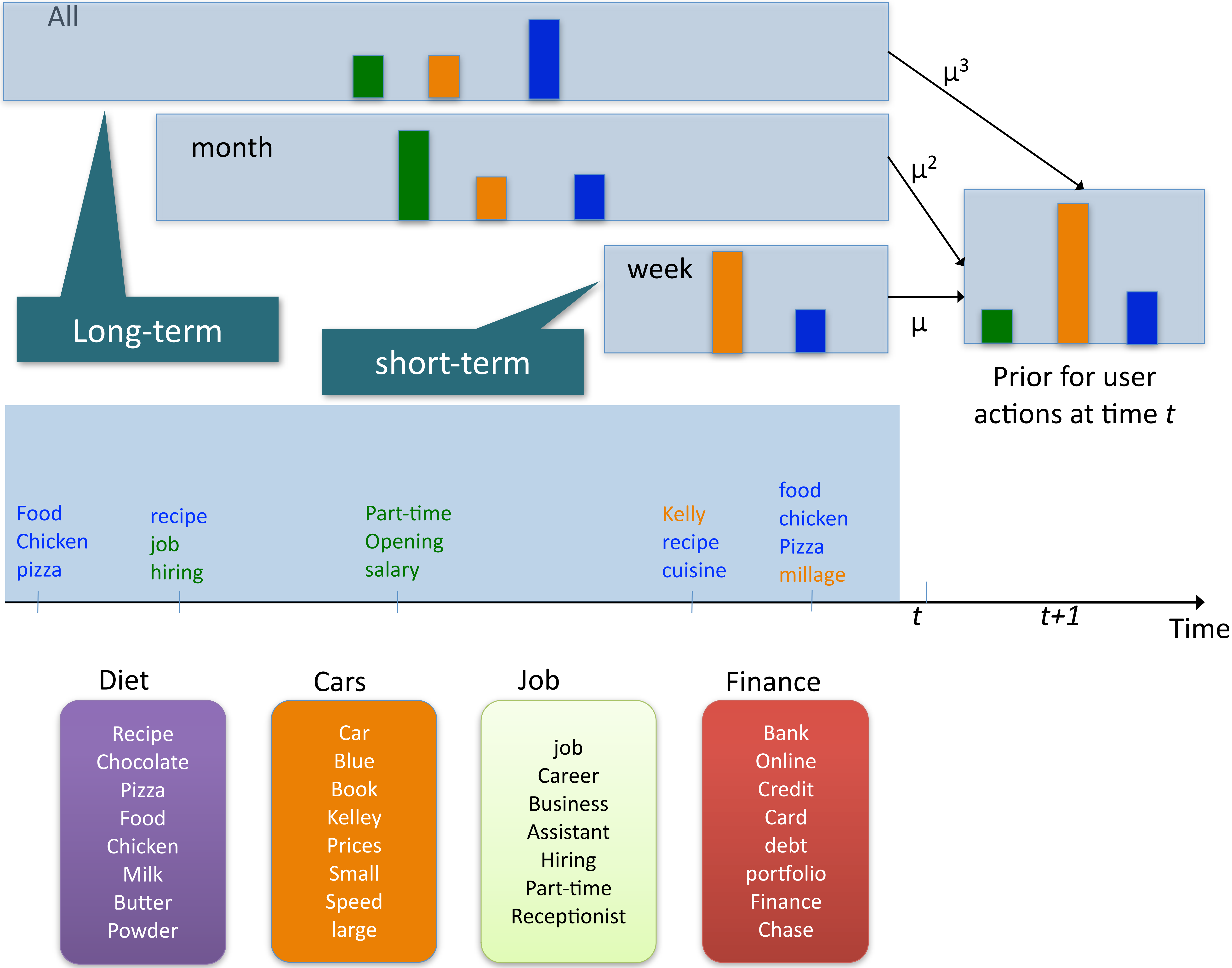
All
Long-term
Prior for user actions at time t
Food
Chicken
pizza
recipe
job
hiring
Part-time
Opening
salary
Kelly
recipe
cuisine
food
chicken
Pizza
millage
t
t+1
Time
Diet
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder
Cars
Car
Blue
Book
Kelley
Prices
Small
Speed
large
Job
job
Career
Business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



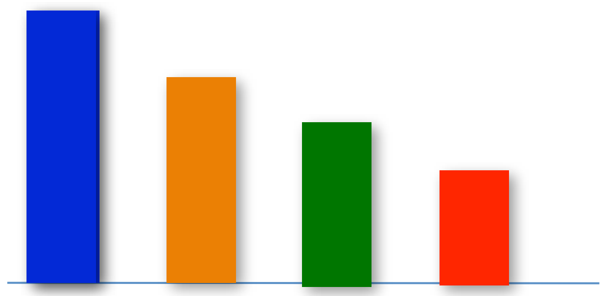
All
month
Long-term
Prior for user actions at time t
Food
Chicken
pizza
recipe
job
hiring
Part-time
Opening
salary
Kelly
recipe
cuisine
food
chicken
Pizza
millage
t
$t+1$
Time
Diet
Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder
Cars
Car
Blue
Book
Kelley
Prices
Small
Speed
large
Job
job
Career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


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

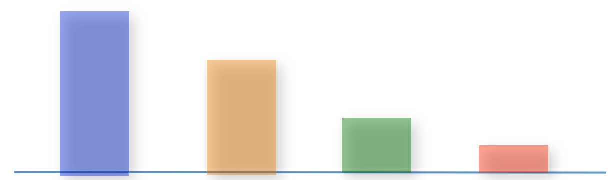

All
month
week
Long-term
short-term
Prior for user actions at time t
Food Chicken pizza
recipe job hiring
Part-time Opening salary
Kelly recipe cuisine
food chicken Pizza millage
t
t+1
Time
Diet
Recipe Chocolate Pizza Food Chicken Milk Butter Powder
Cars
Car Blue Book Kelley Prices Small Speed large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


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

At time t



At time t+1

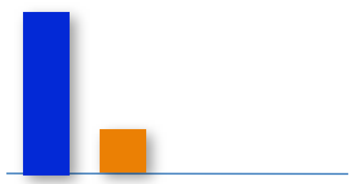


Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

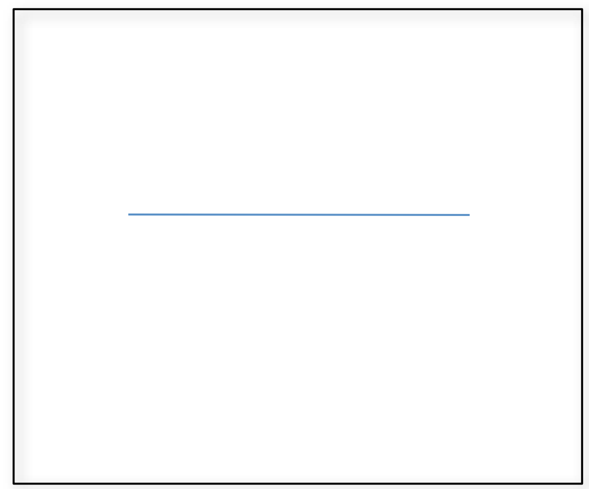
Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

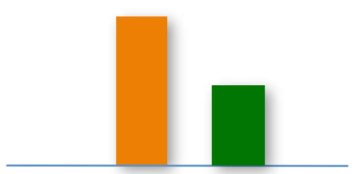
job
Career
Business
Assistant
Hiring
Part-time
Receptionist

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase



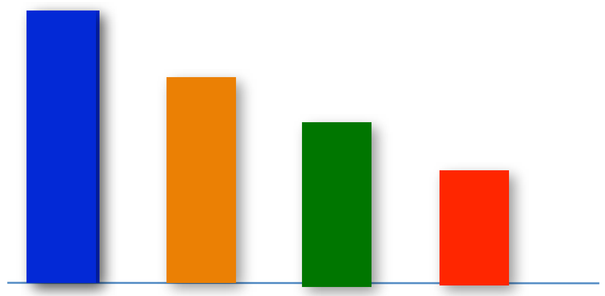
Food Chicken
Pizza mileage



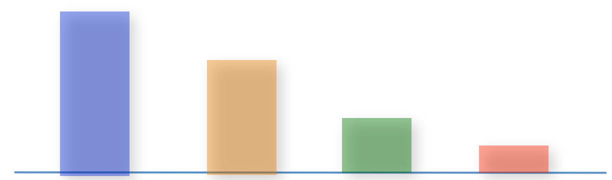


Car speed offer
Camry accord career

At time t



At time t+1

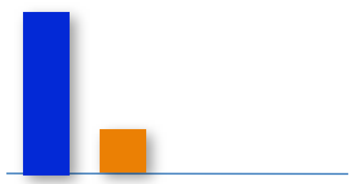


Recipe
Chocolate
Pizza
Food
Chicken
Milk
Butter
Powder

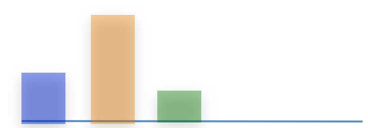
Car
Altima
Accord
Blue
Book
Kelley
Prices
Small
Speed

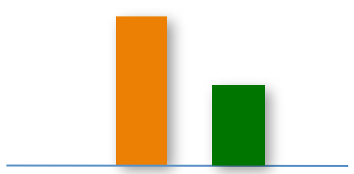
job
Career
Business
Assistant
Hiring
Part-time
Receptioni
st

Bank
Online
Credit
Card
debt
portfolio
Finance
Chase

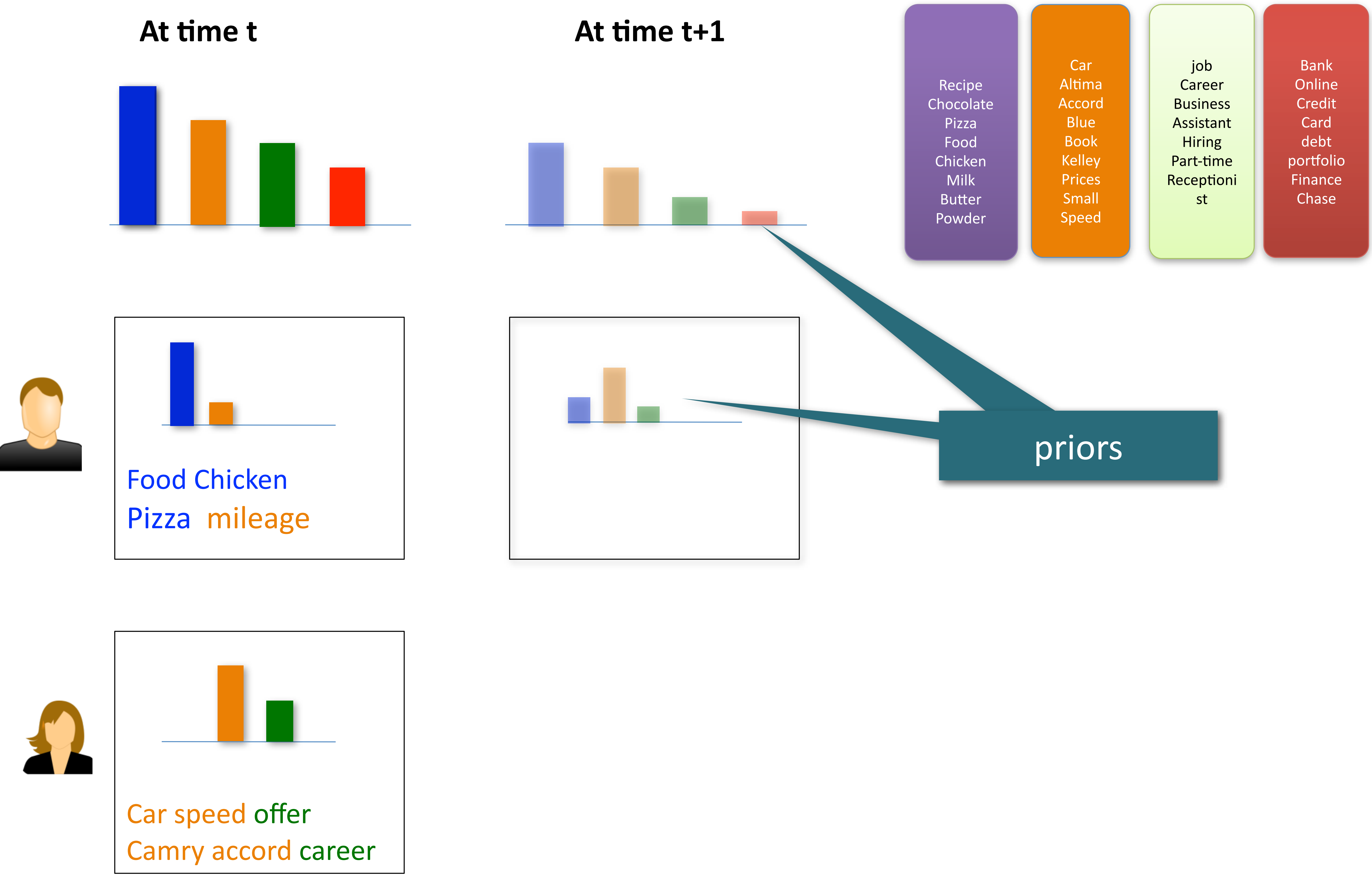


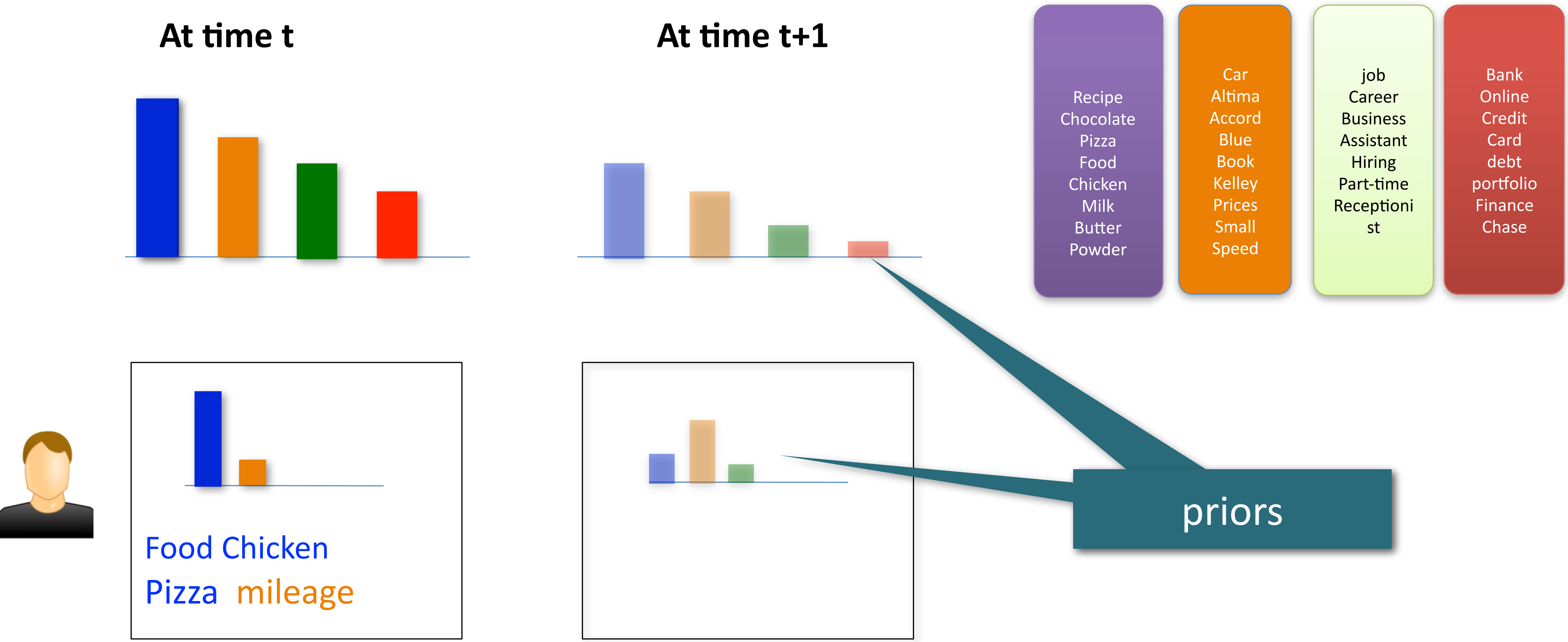
Food Chicken
Pizza mileage

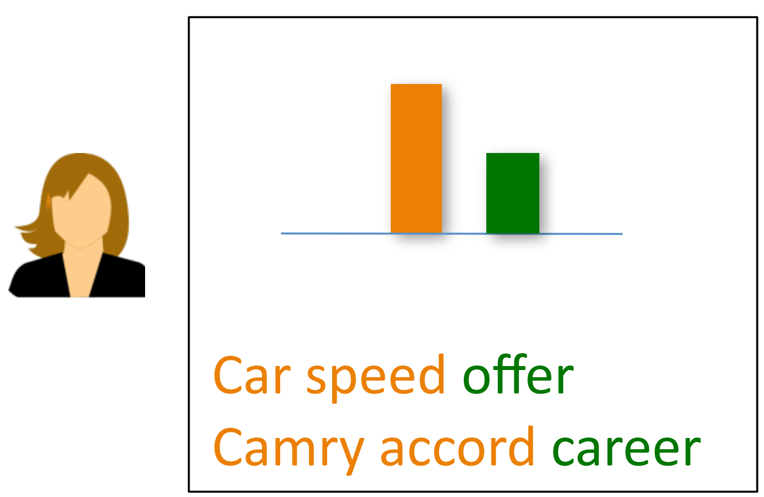




Car speed offer
Camry accord career

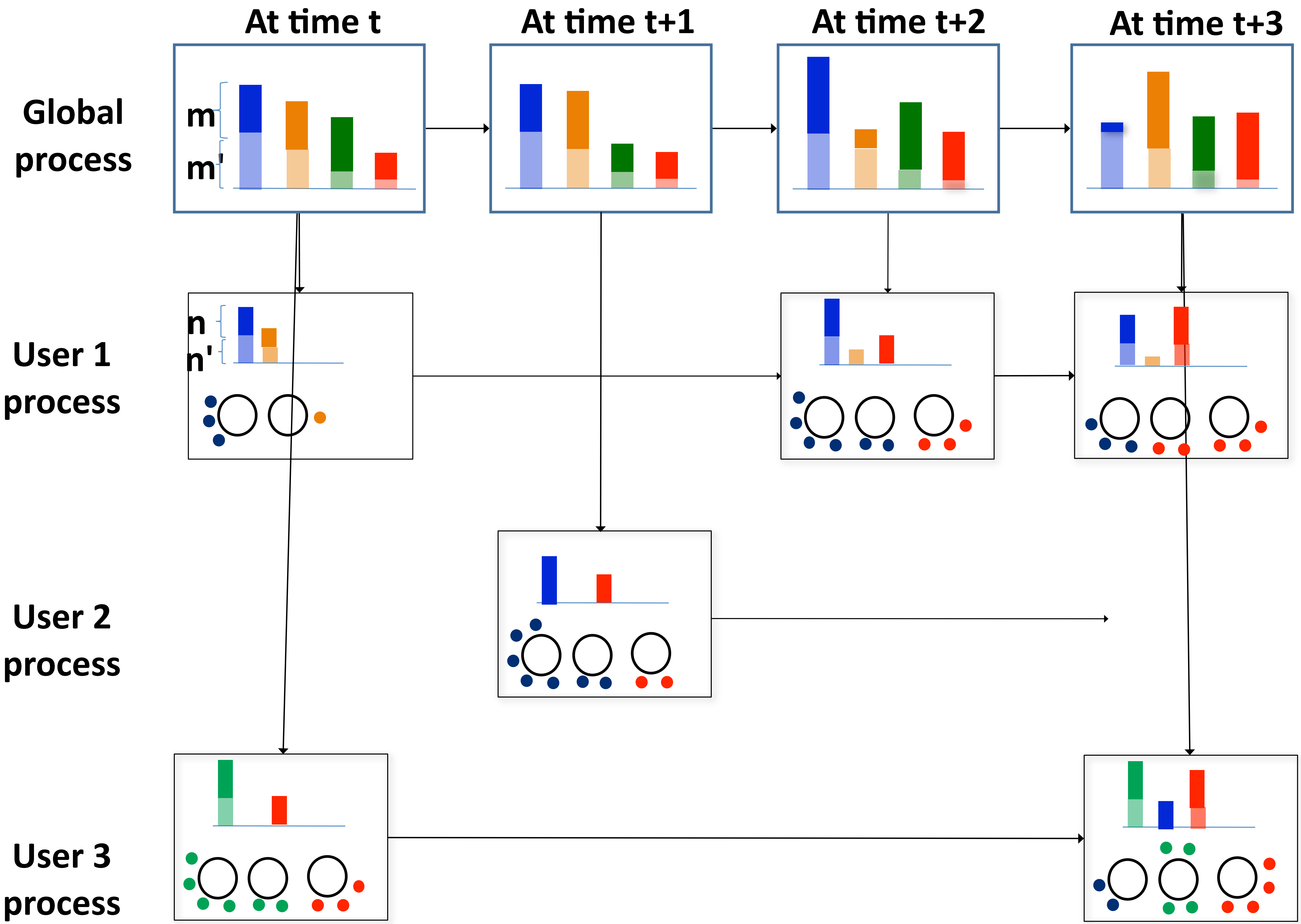

At time t
At time t+1
Recipe Chocolate Pizza Food Chicken Milk Butter Powder
Car Altima Accord Blue Book Kelley Prices Small Speed
job Career Business Assistant Hiring Part-time Receptionist
Bank Online Credit Card debt portfolio Finance Chase
priors
Food Chicken
Pizza mileage
Car speed offer
Camry accord career





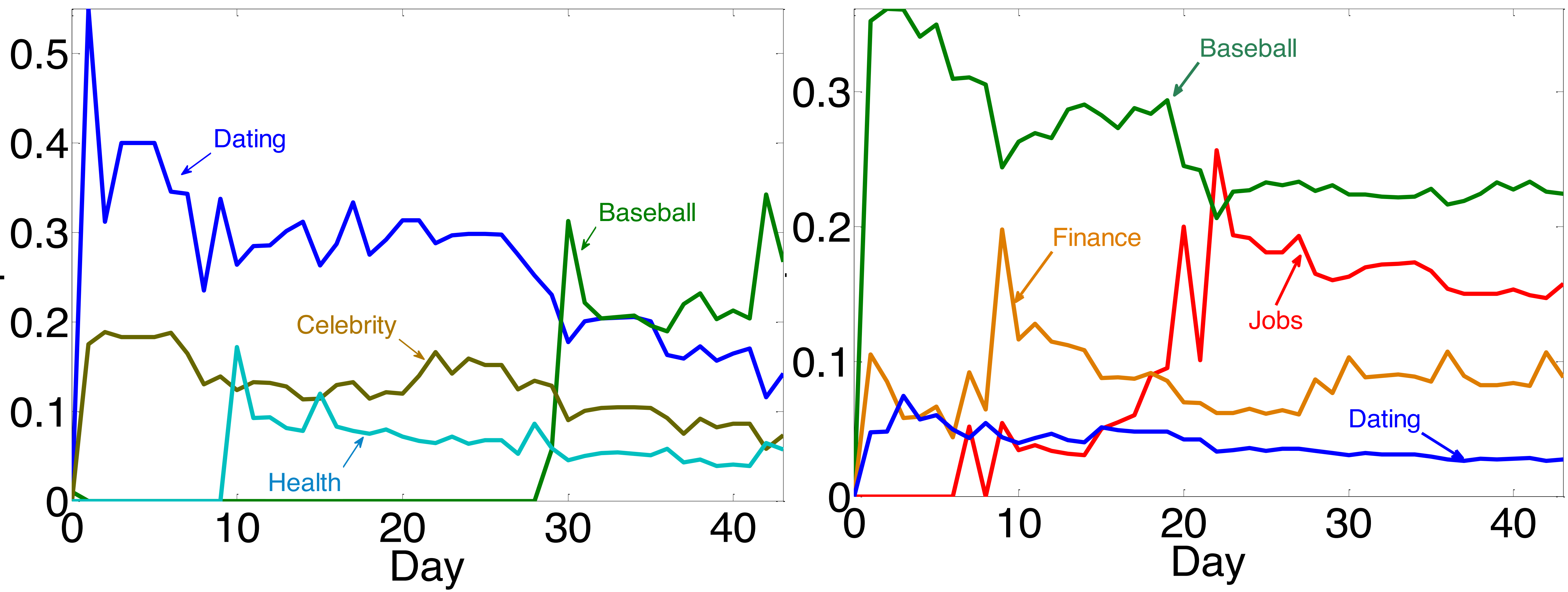
Generative Process

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



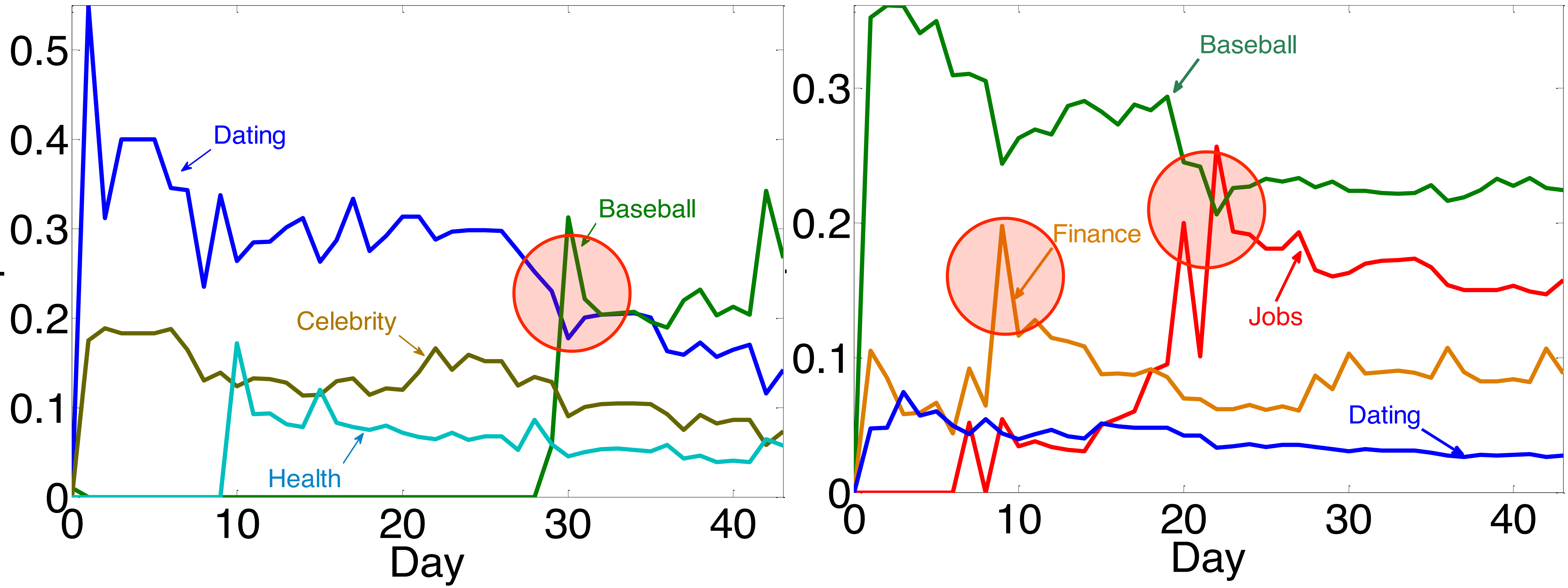
At time t
At time t+1
At time t+2
At time t+3
Global process
User 1 process
User 2 process
User 3 process
m
m'
n
n'

Sample users



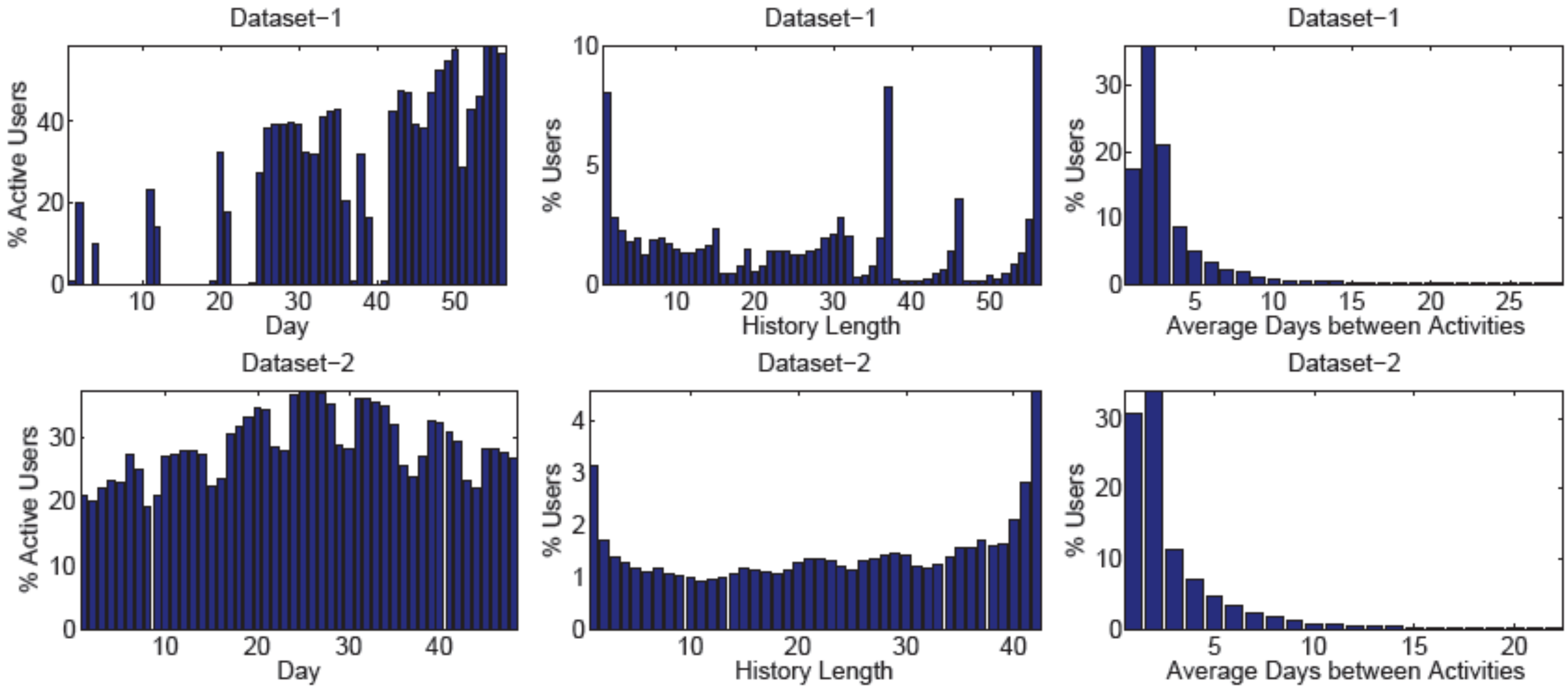
Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time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

Sample users



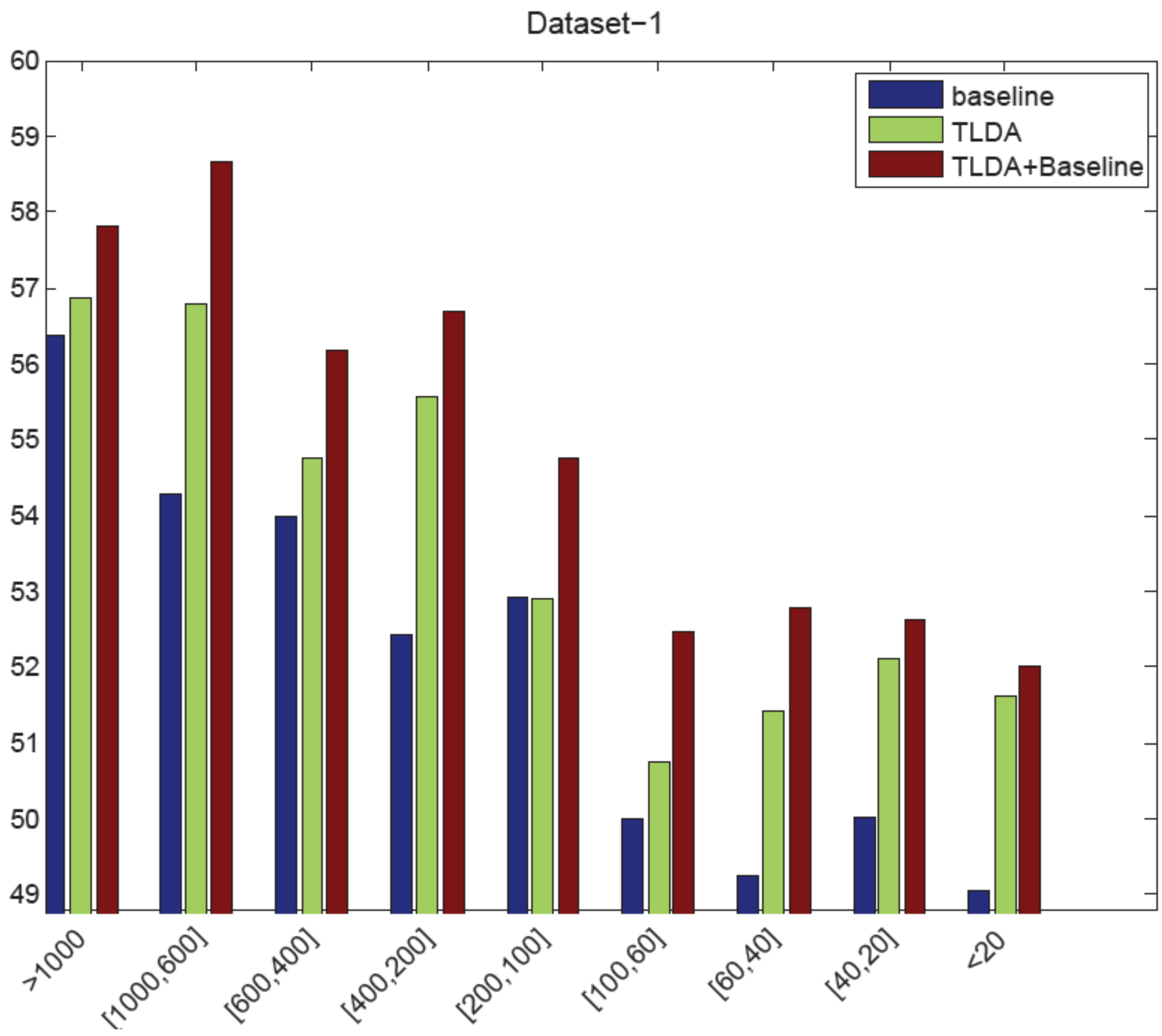
Dating	Baseball	Celebrity	Health	Jobs	Finance
women	League	Snooki	skin	job	financial
men	baseball	Tom	body	career	Thomson
dating	basketball,	Cruise	fingers	business	chart
singles	doublehead	Katie	cells	assistant	real
personals	Bergesen	Holmes	toes	hiring	Stock
seeking	Griffey	Pinkett	wrinkle	part-time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

Data

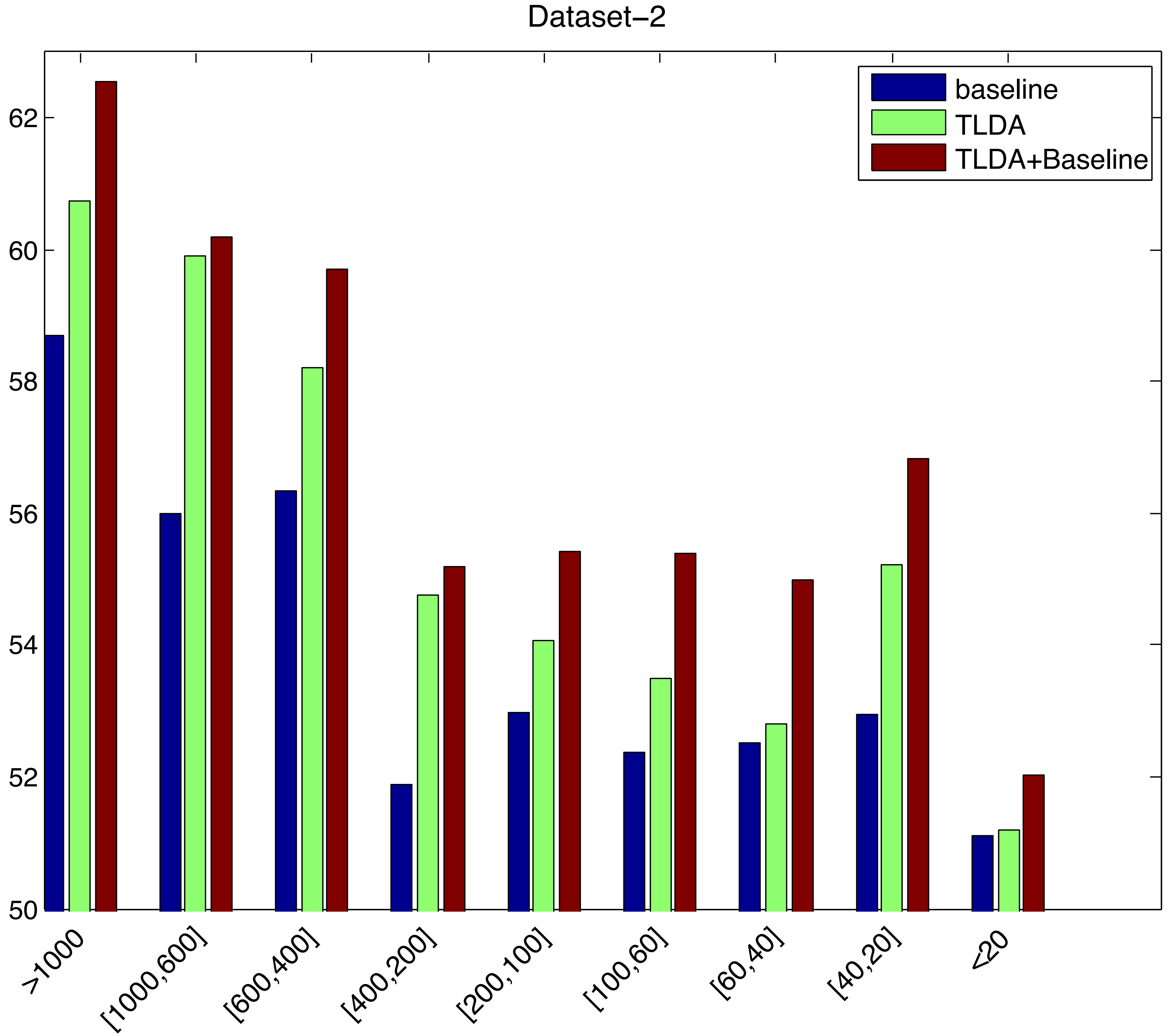


dataset	# days	# users	# campaigns	size
1	56	13.34M	241	242GB
2	44	33.5M	216	435GB

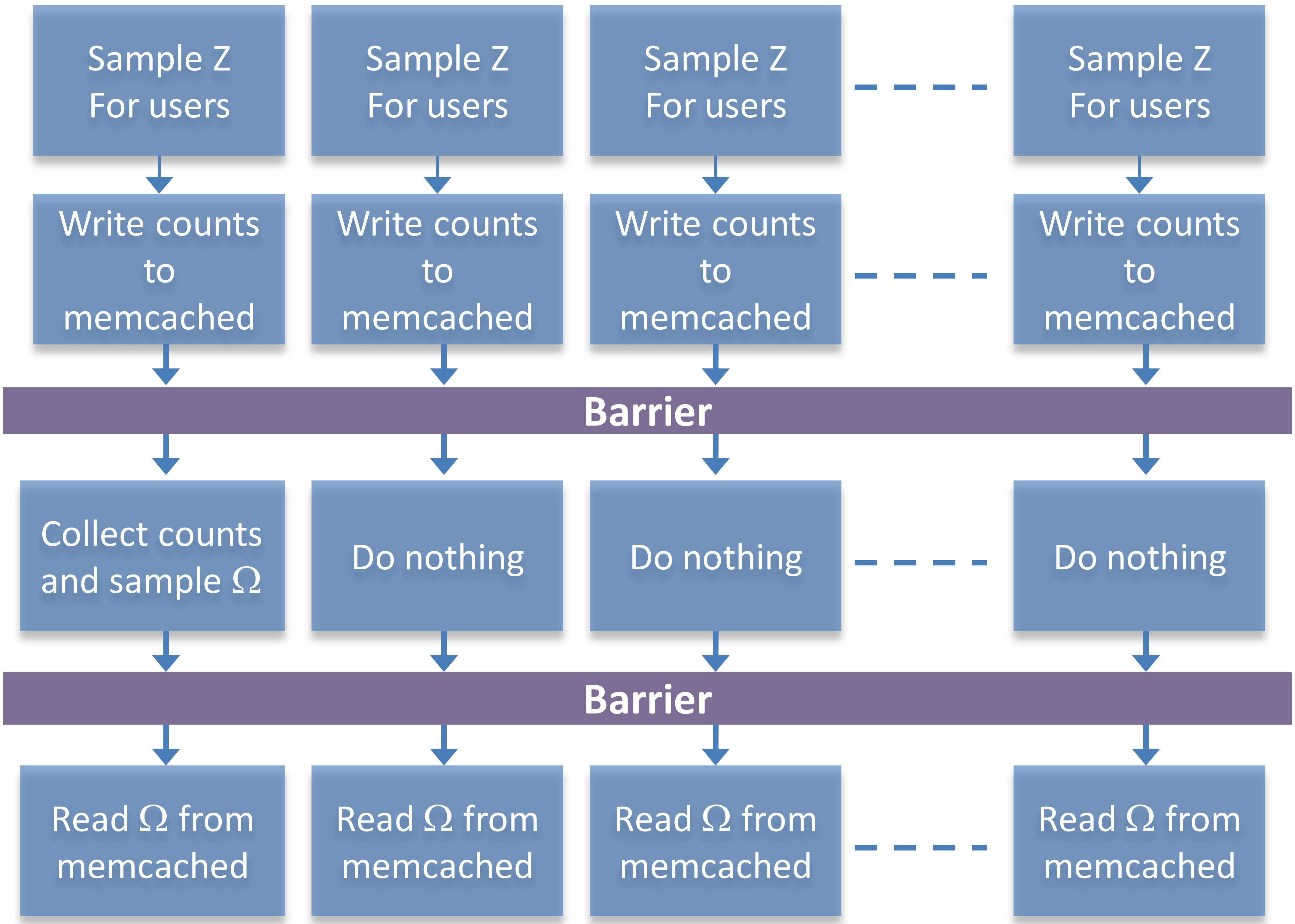
ROC score improvement



ROC score improvement



LDA for user profiling



News

News Stream

News Stream

Add-ons turn tax cut bill into 'Christmas t...'

AP – 1 hr 32 mins ago

WASHINGTON – In the

AP

Republicans is bec

and lawmakers. Bu

Bill Clinton even ba

Full Story »

Video: Gibbs: I Ha

Slideshow: Preside

Related: Tax fight

BEYOND FOSSIL FUELS

Using Waste, Swedish (

Johan Spanner for The New York Times

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

China says inflation up 5.1 perce...ber

Associated Press

Buzz up! 19 votes

Share

Print

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 su... ovember, officials said Saturday, despit... ase food supplies and end diesel shorta...

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

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

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

Suit to Recover Madoff's Money Calls Austrian an Accomplice

By DIANA B. HENRIQUES and PETER LATTMAN

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

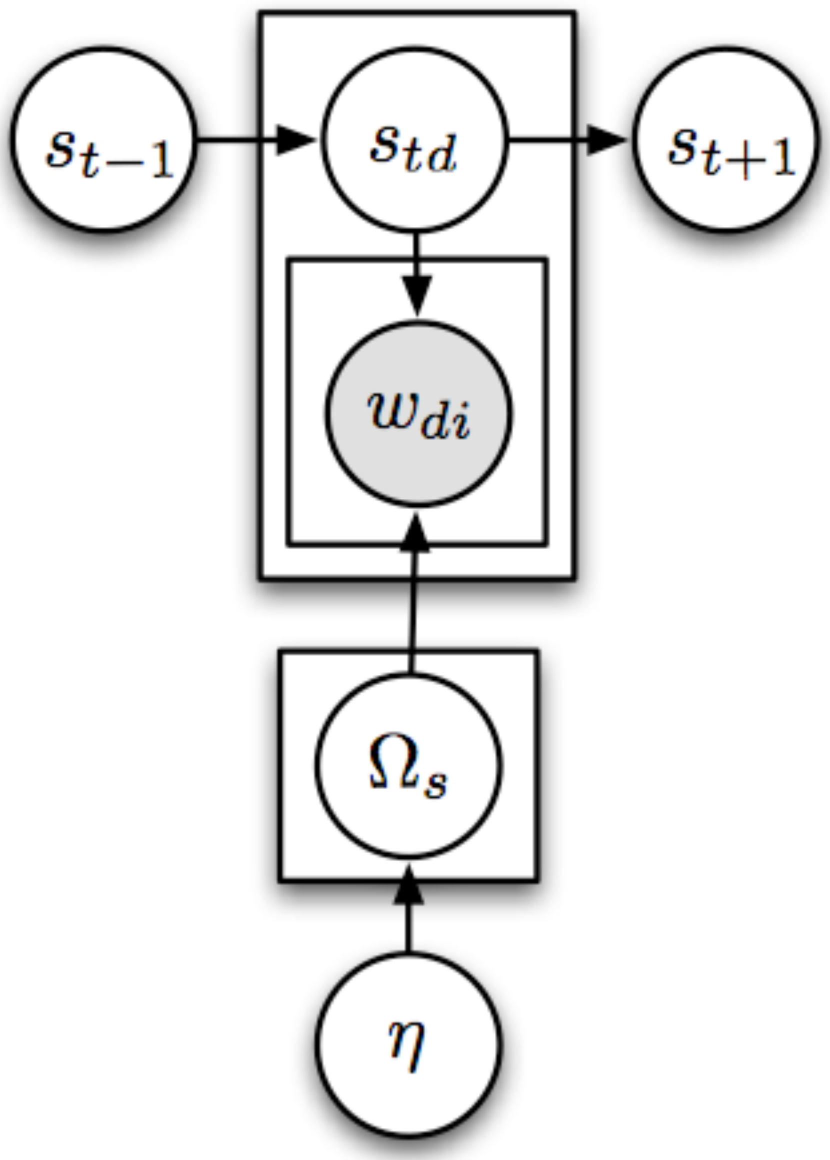
Post a Comment

News Stream

- 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



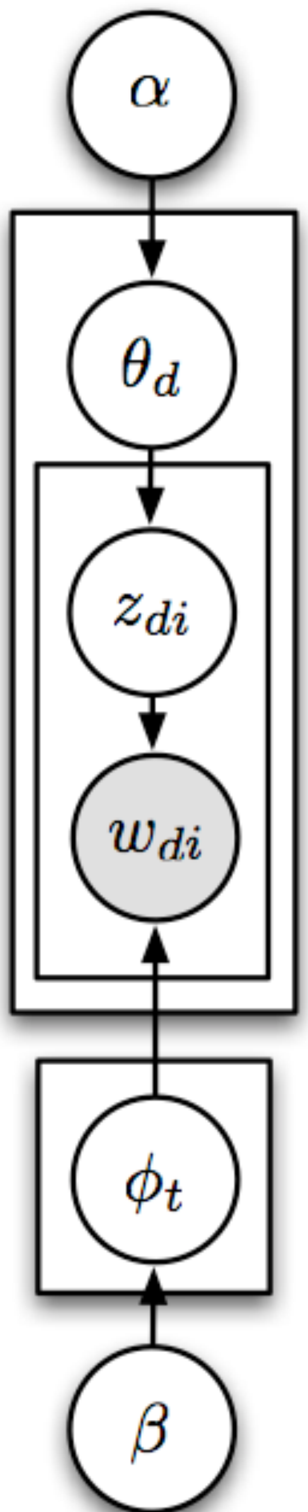
- Assume active story distribution at time t
- Draw story indicator
- Draw words from story distribution
- Down-weight story counts for next day

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

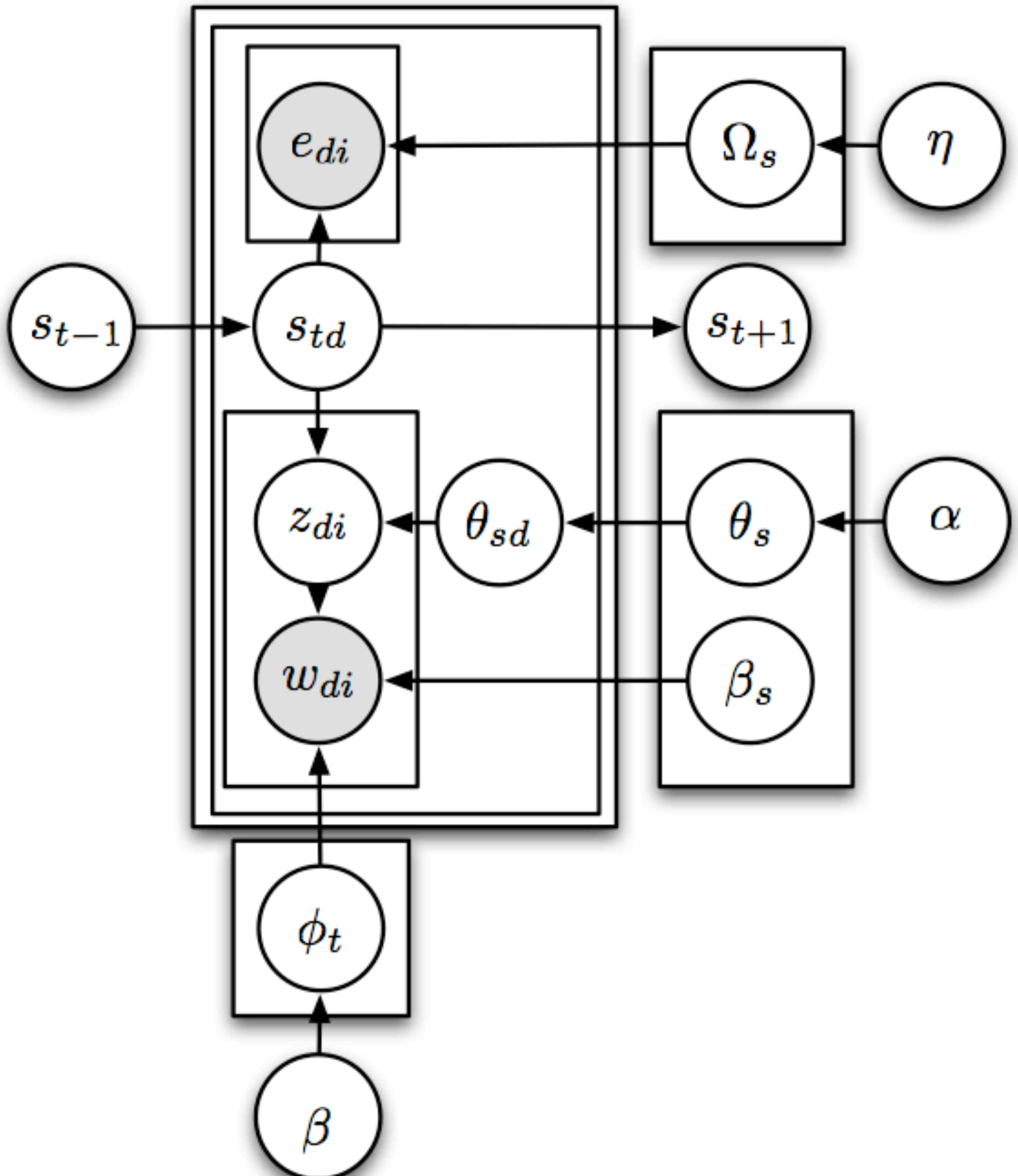
More Issues

- 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

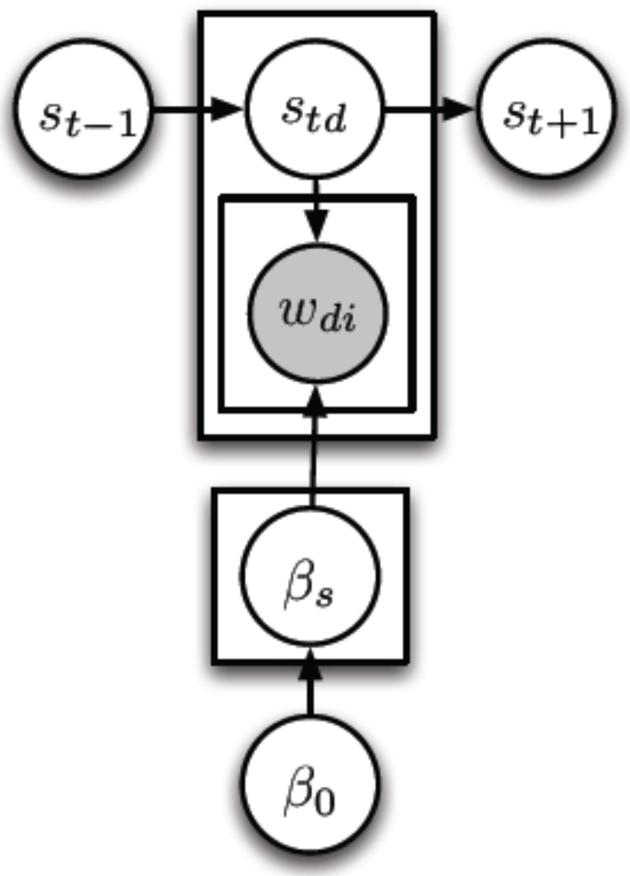
Storylines Model



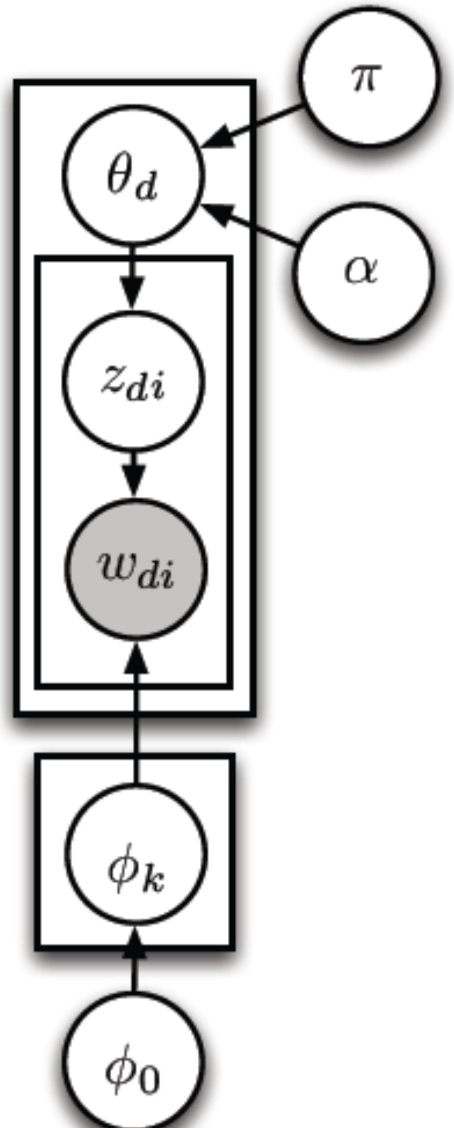
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

Storylines Model

Storylines Model

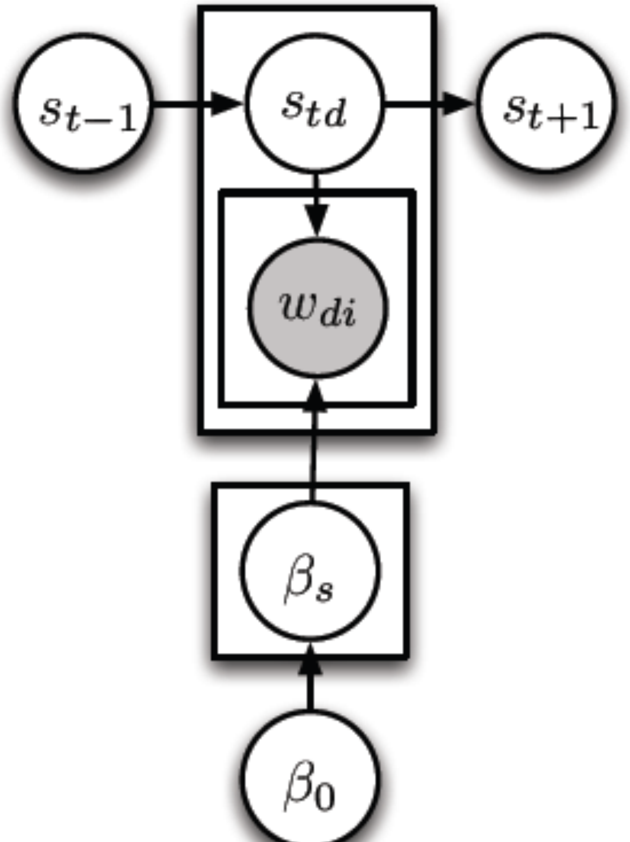


Tightly-focused

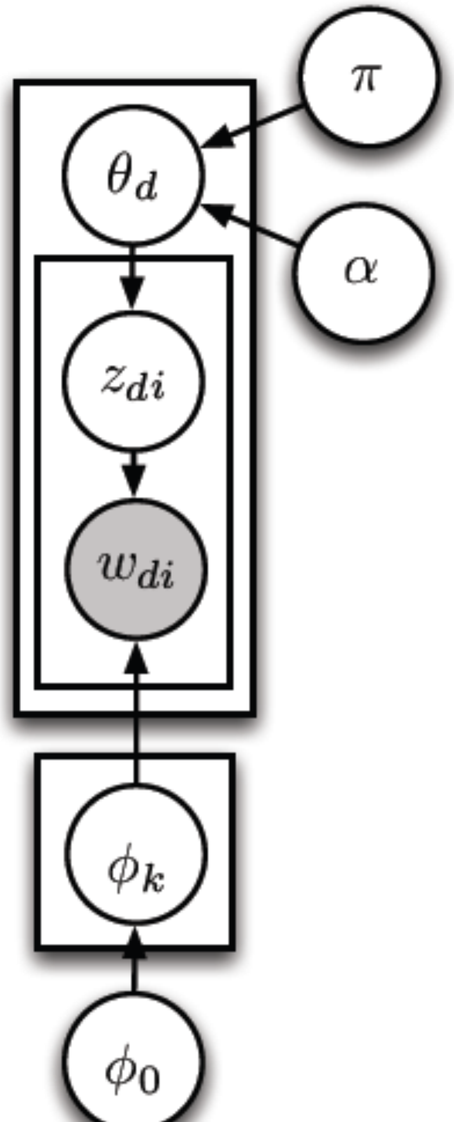


High-level concepts

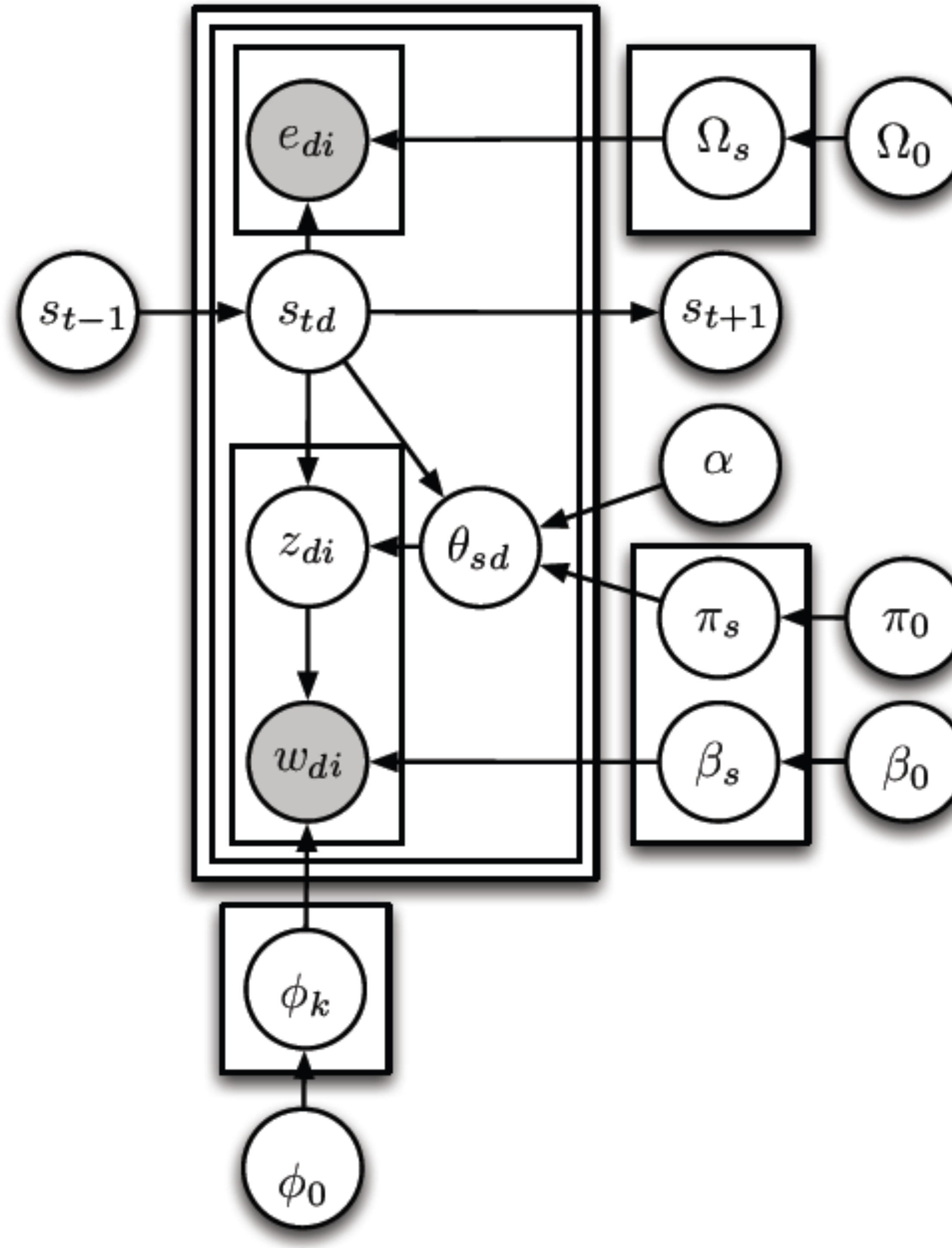
Storylines Model



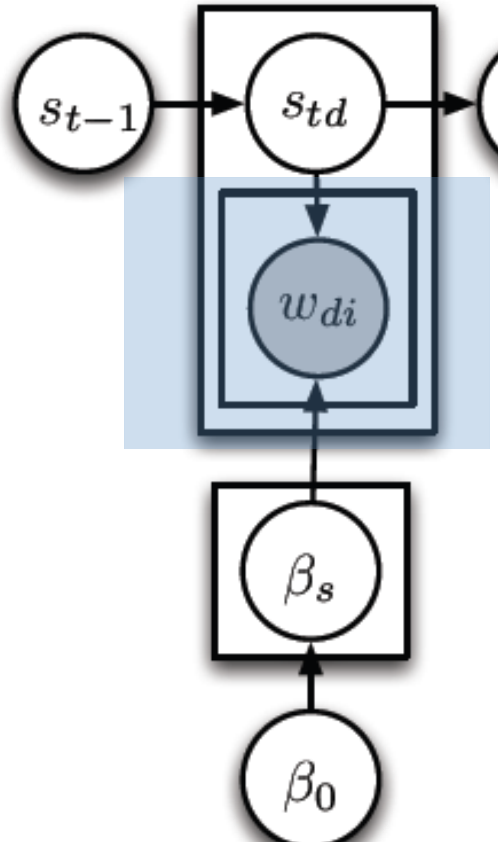
Tightly-focused

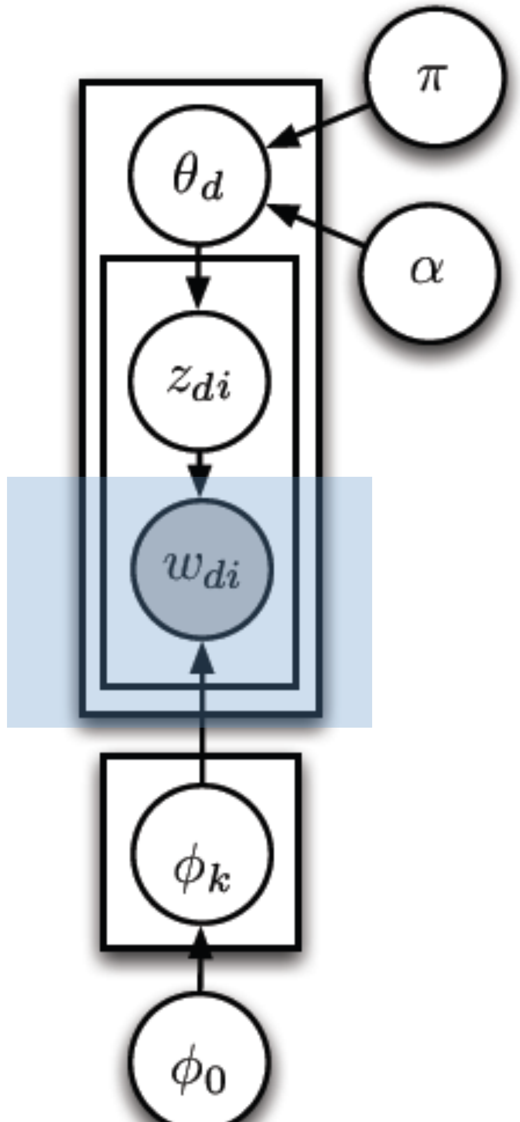


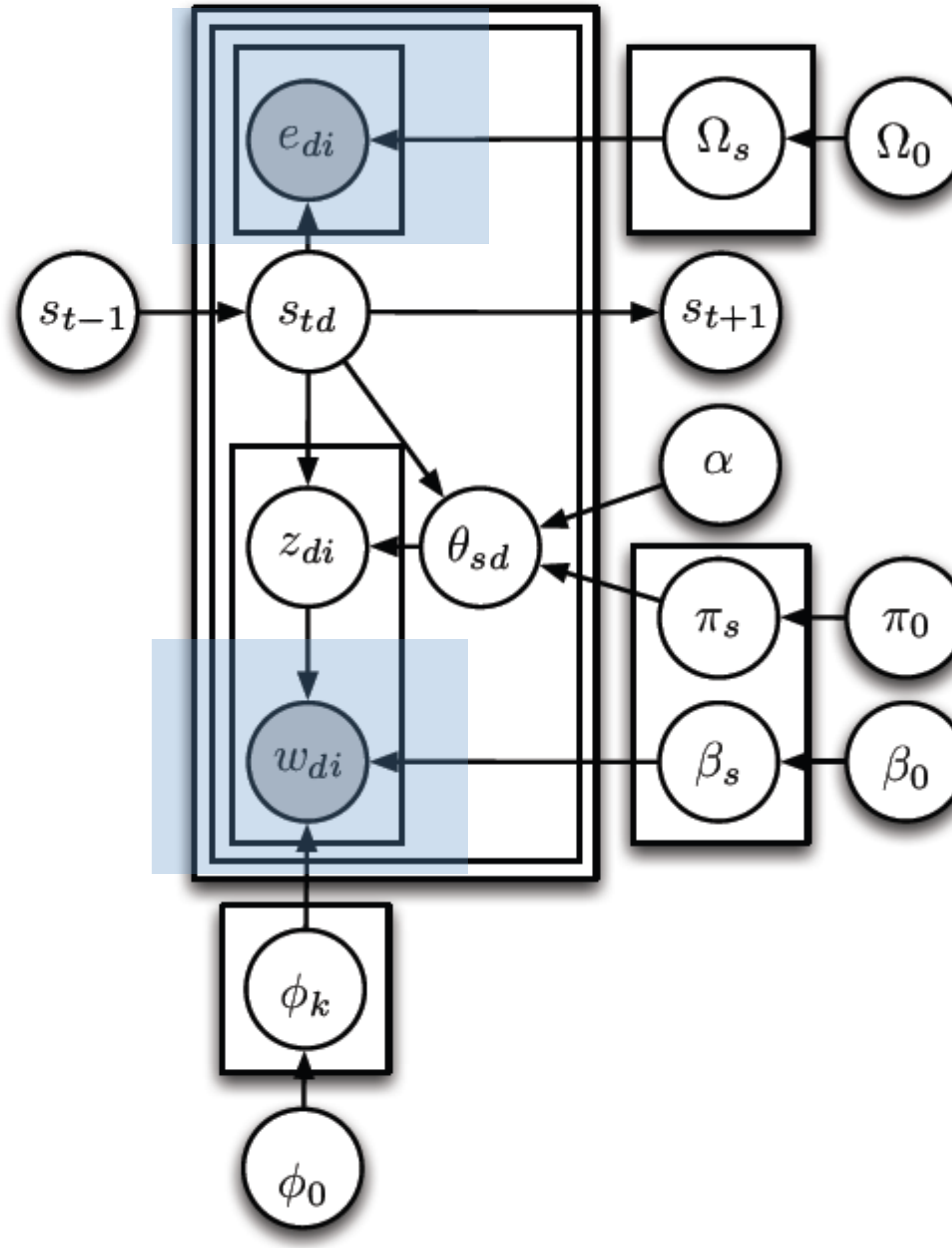
High-level concepts



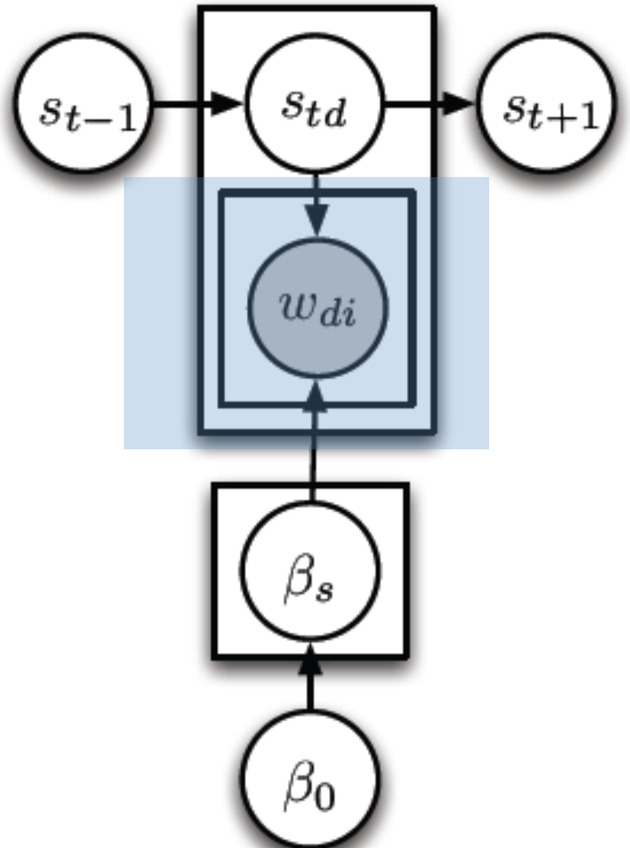
Storylines Model



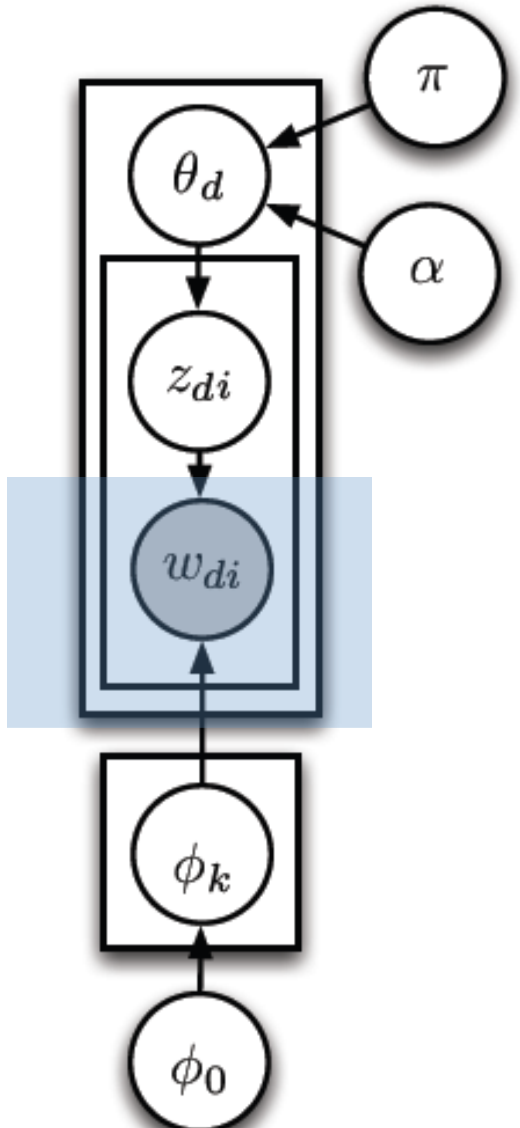




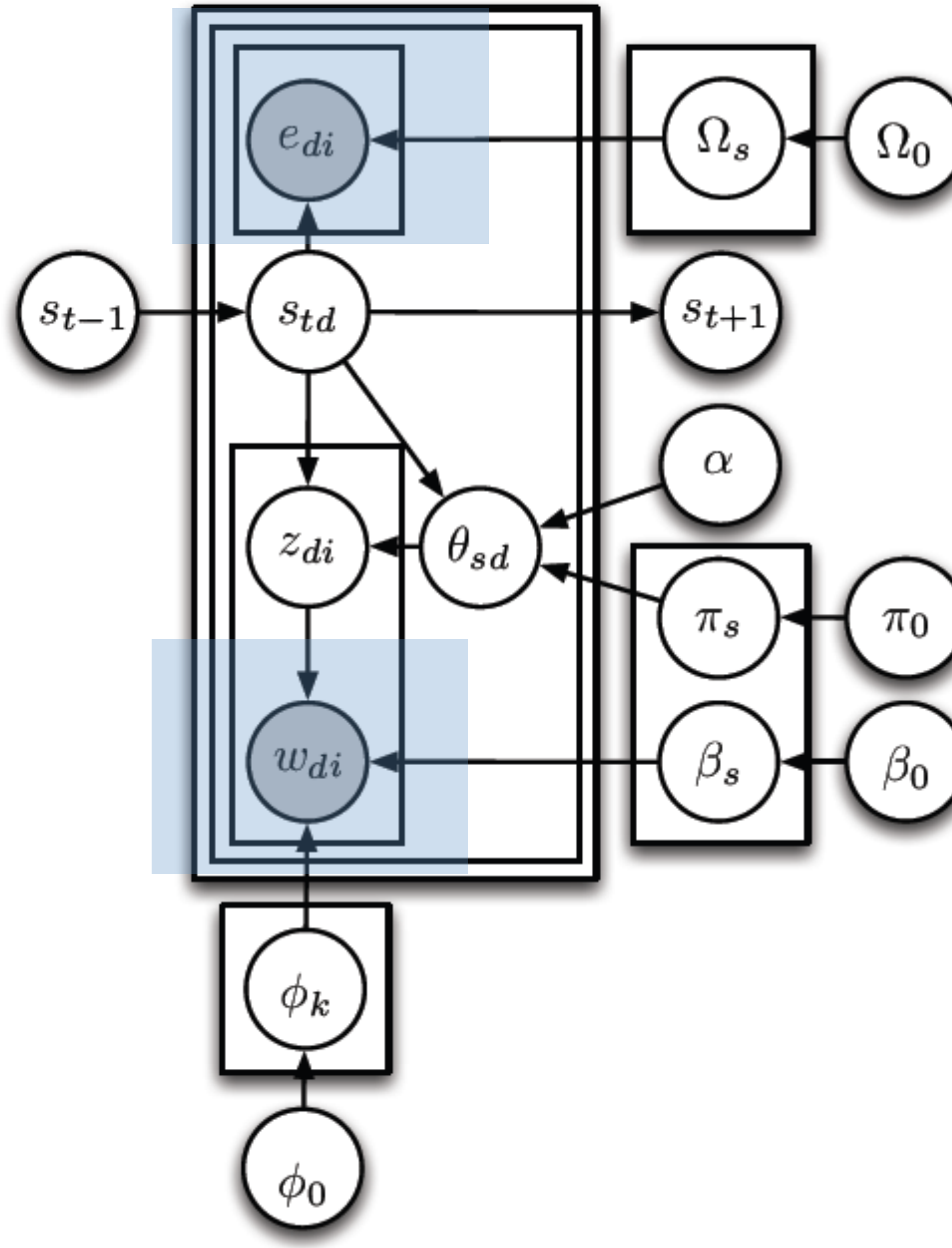
Storylines Model



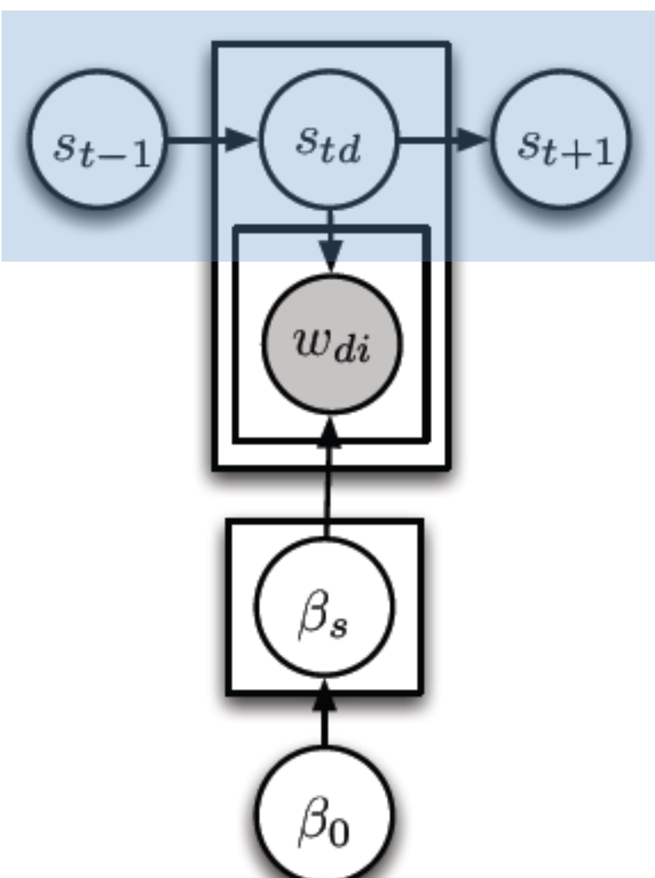
Tightly-focused

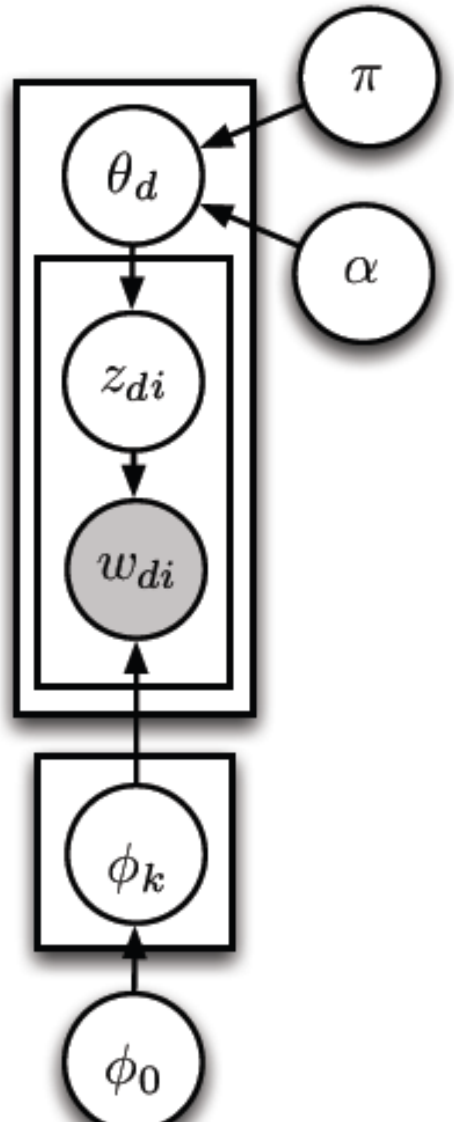


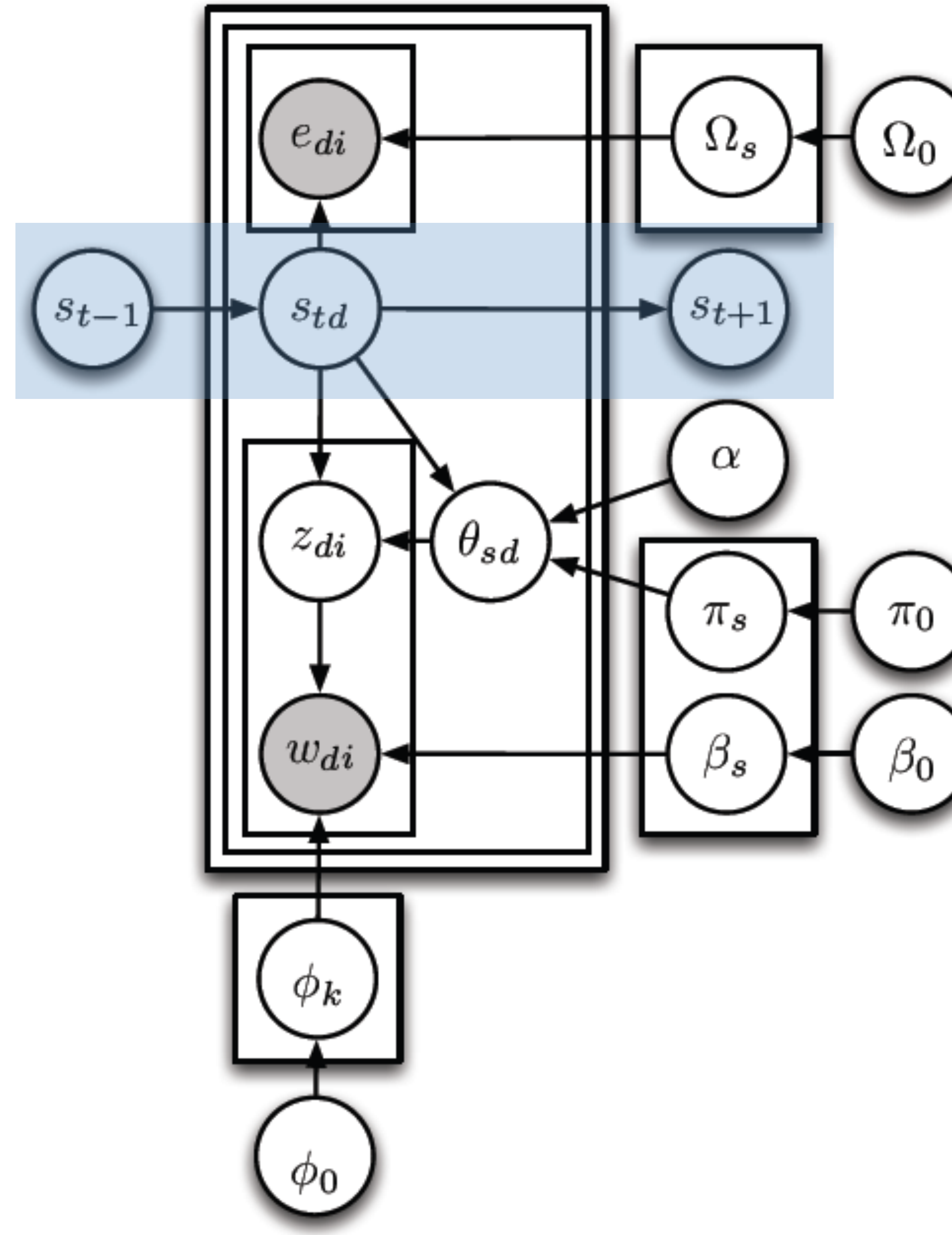
High-level concepts



Storylines Model







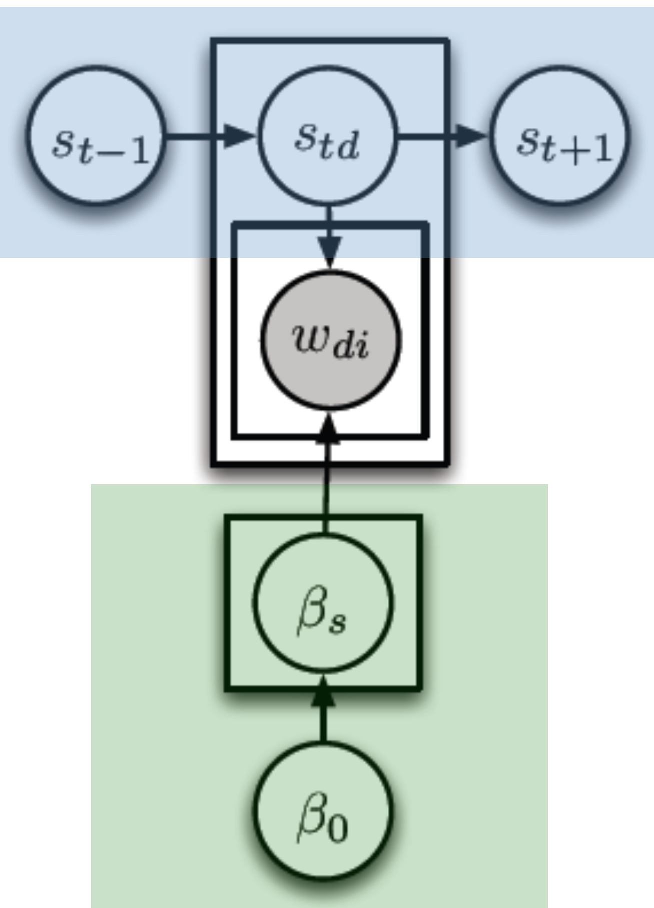
Each story has:

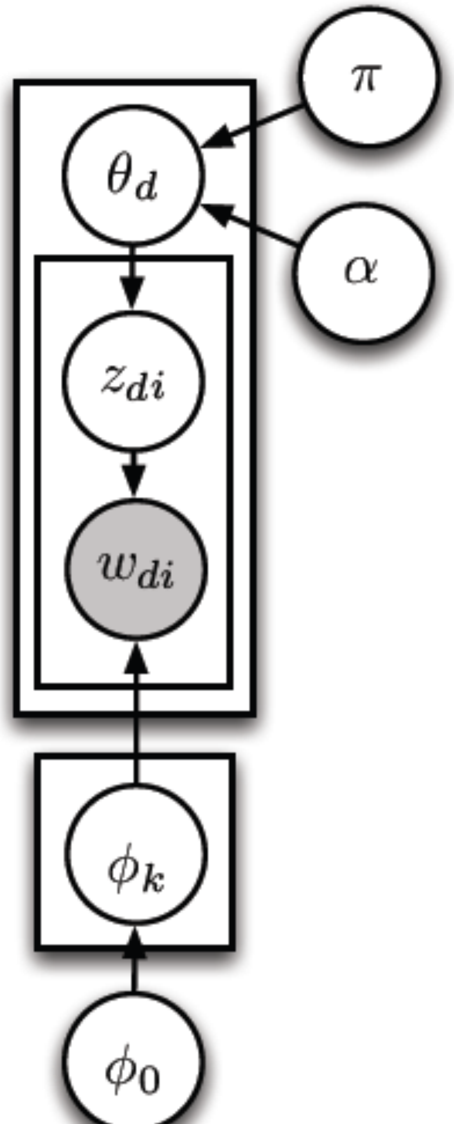
Distribution over words

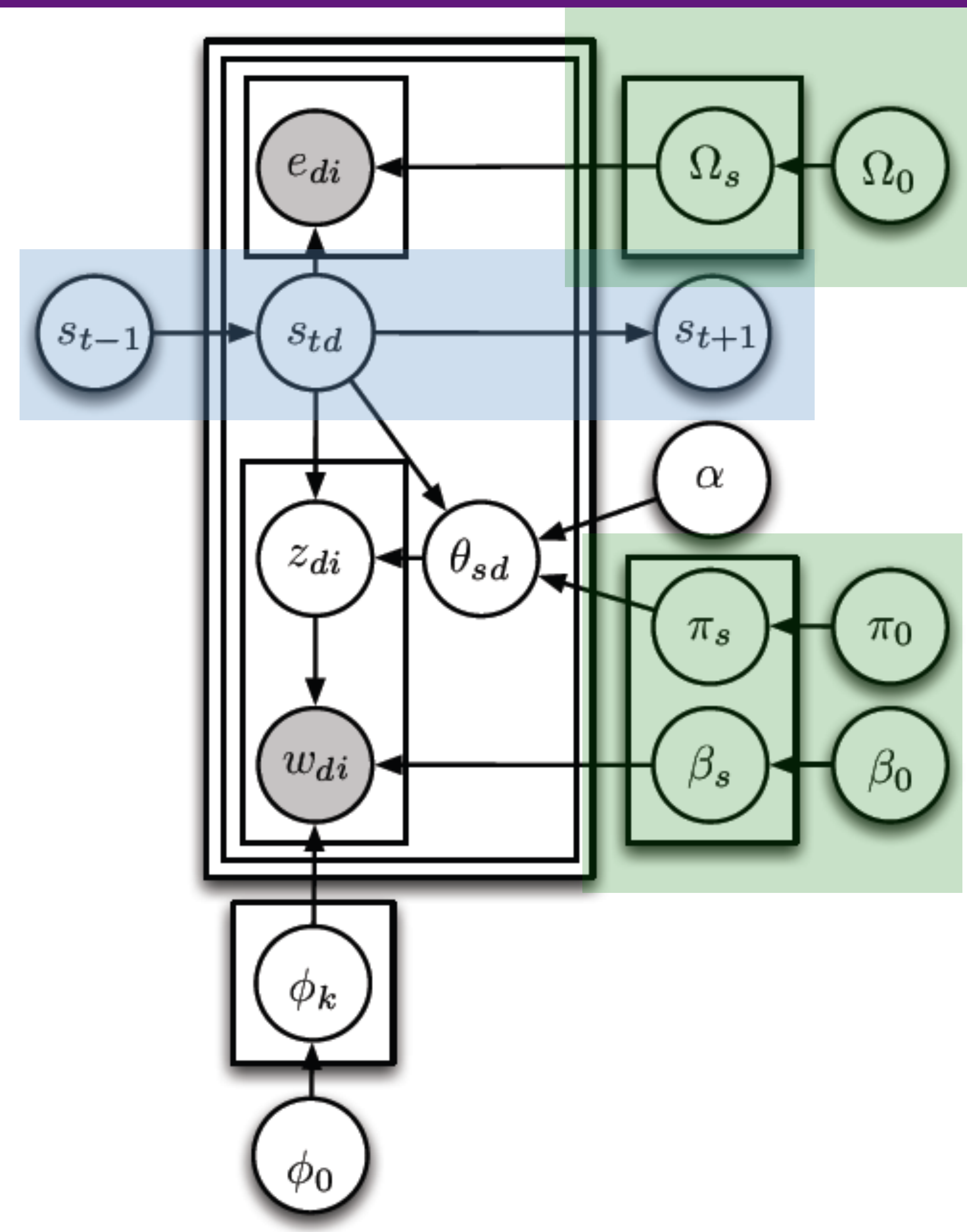
Distribution over topics

Distribution over named entites

Storylines Model







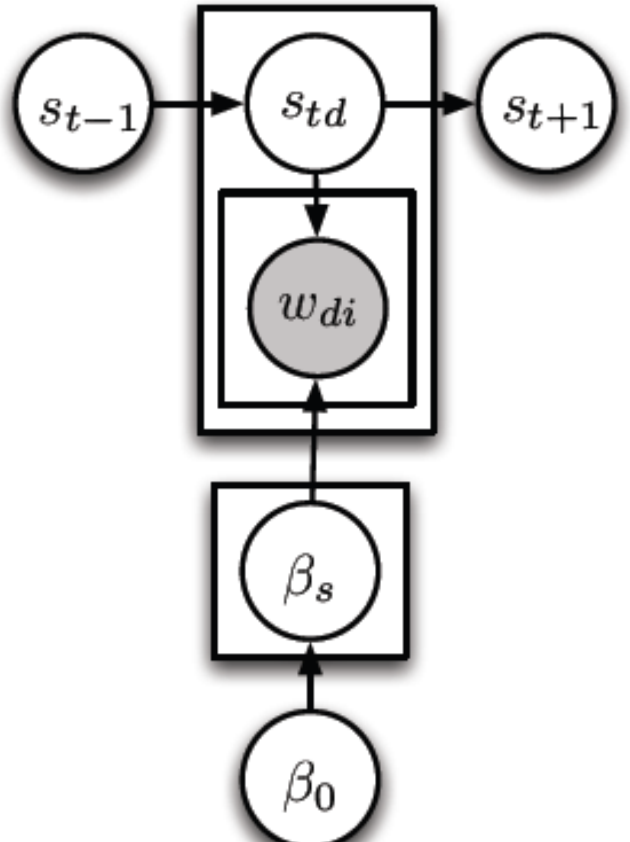
Each story has:

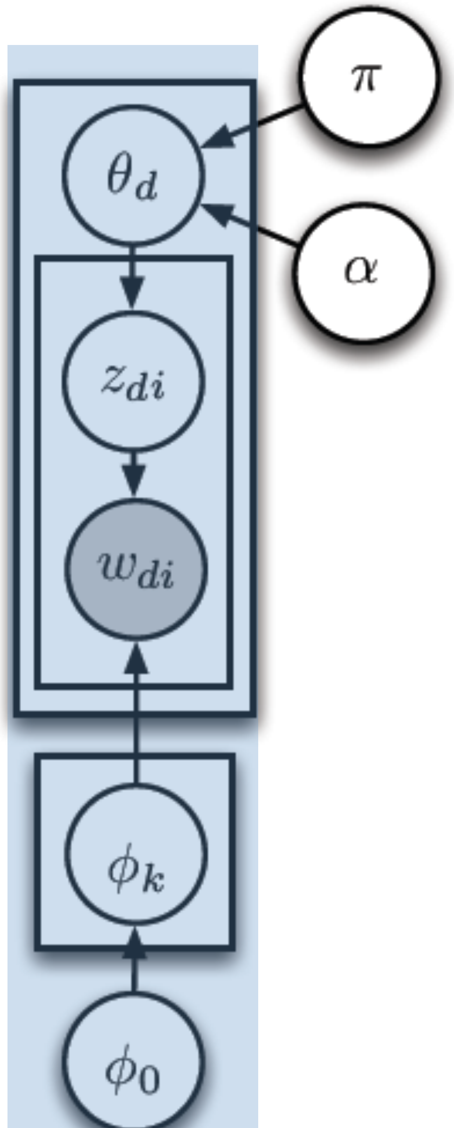
Distribution over words

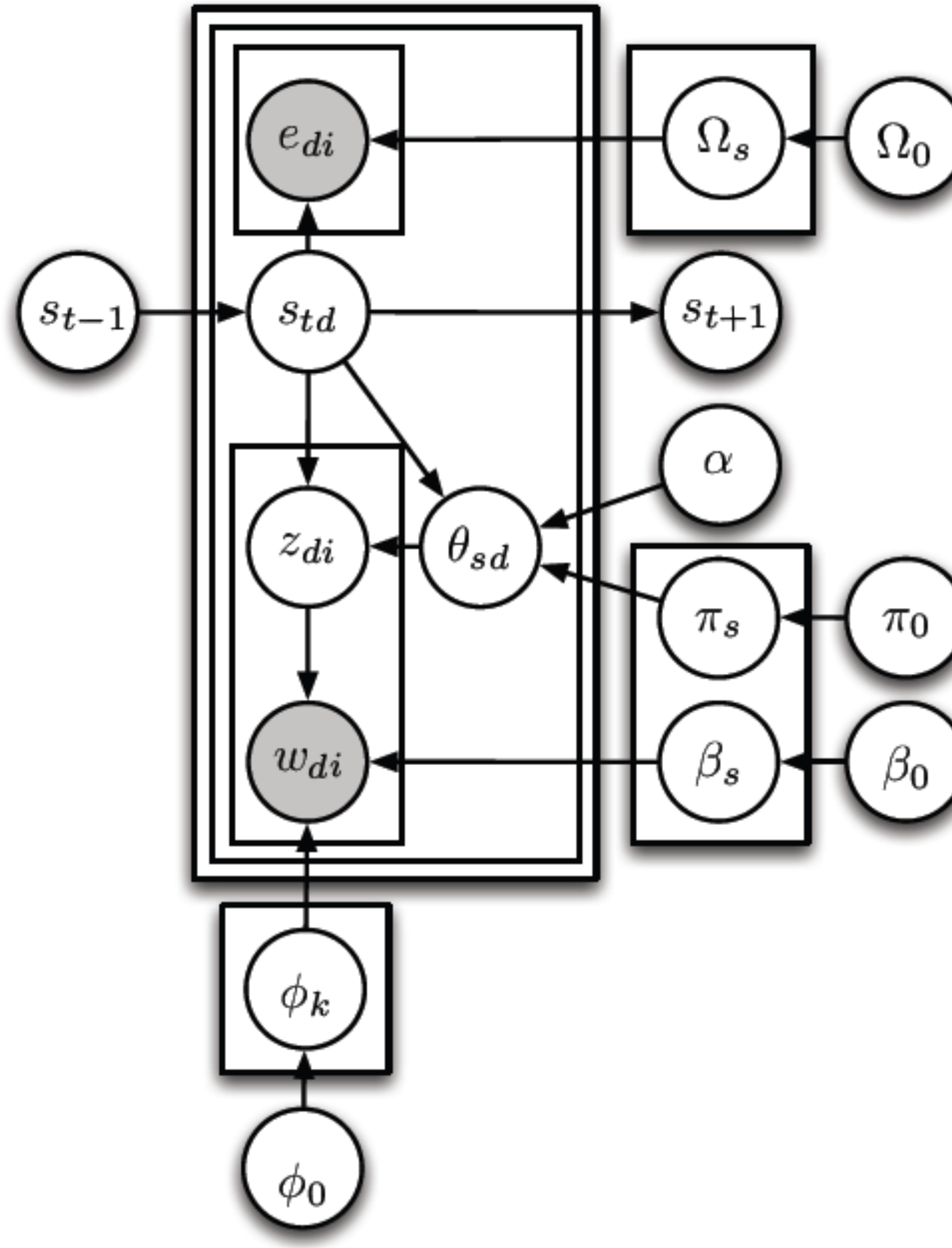
Distribution over topics

Distribution over named entites

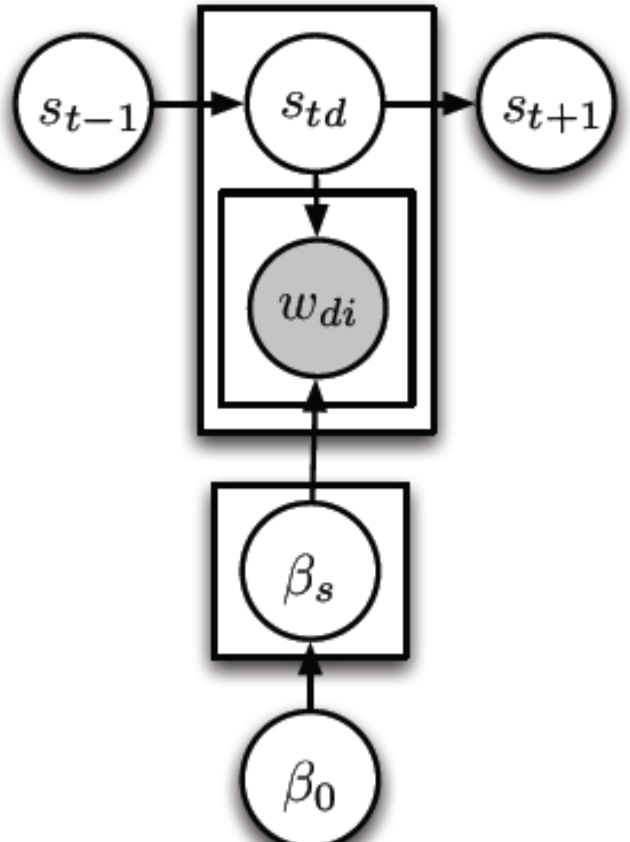
Storylines Model

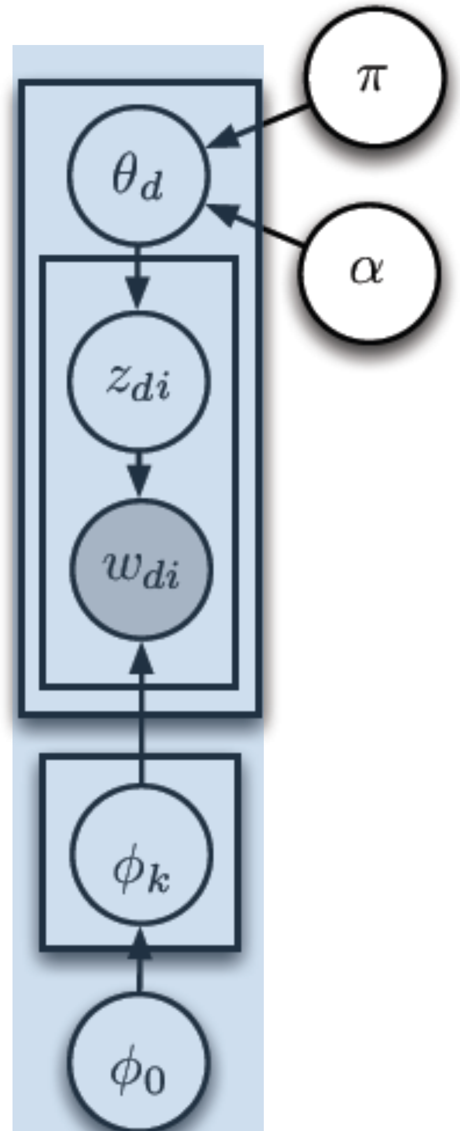


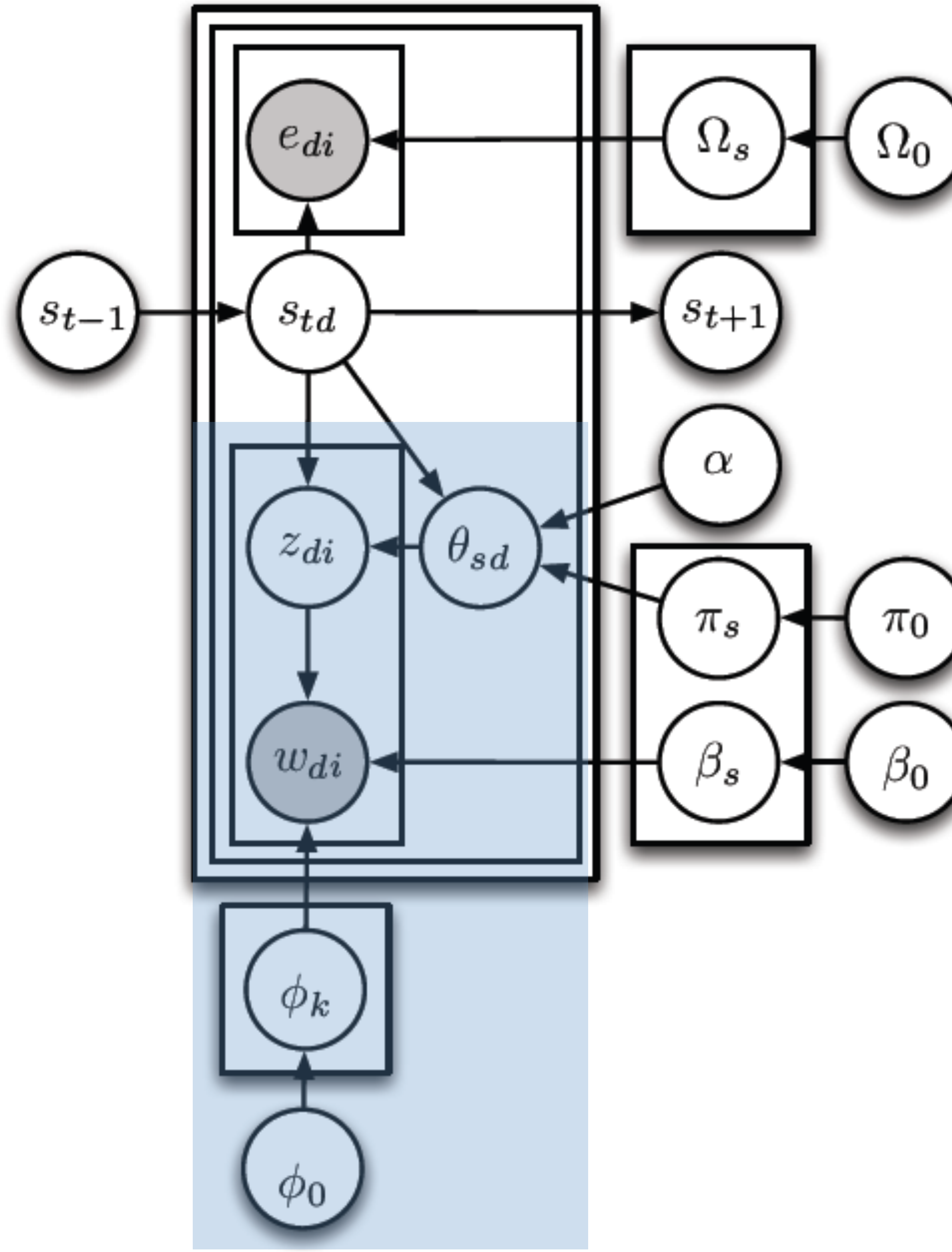




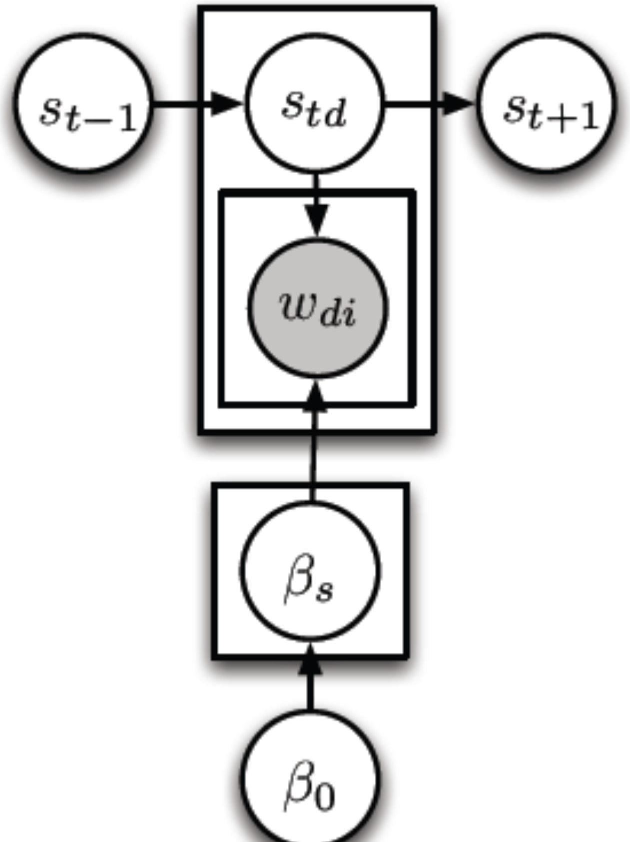
Storylines Model

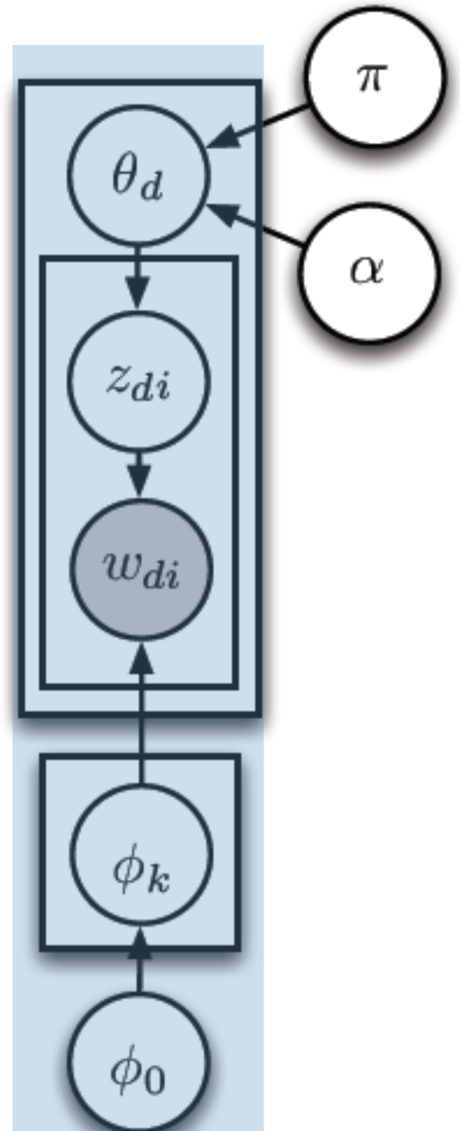


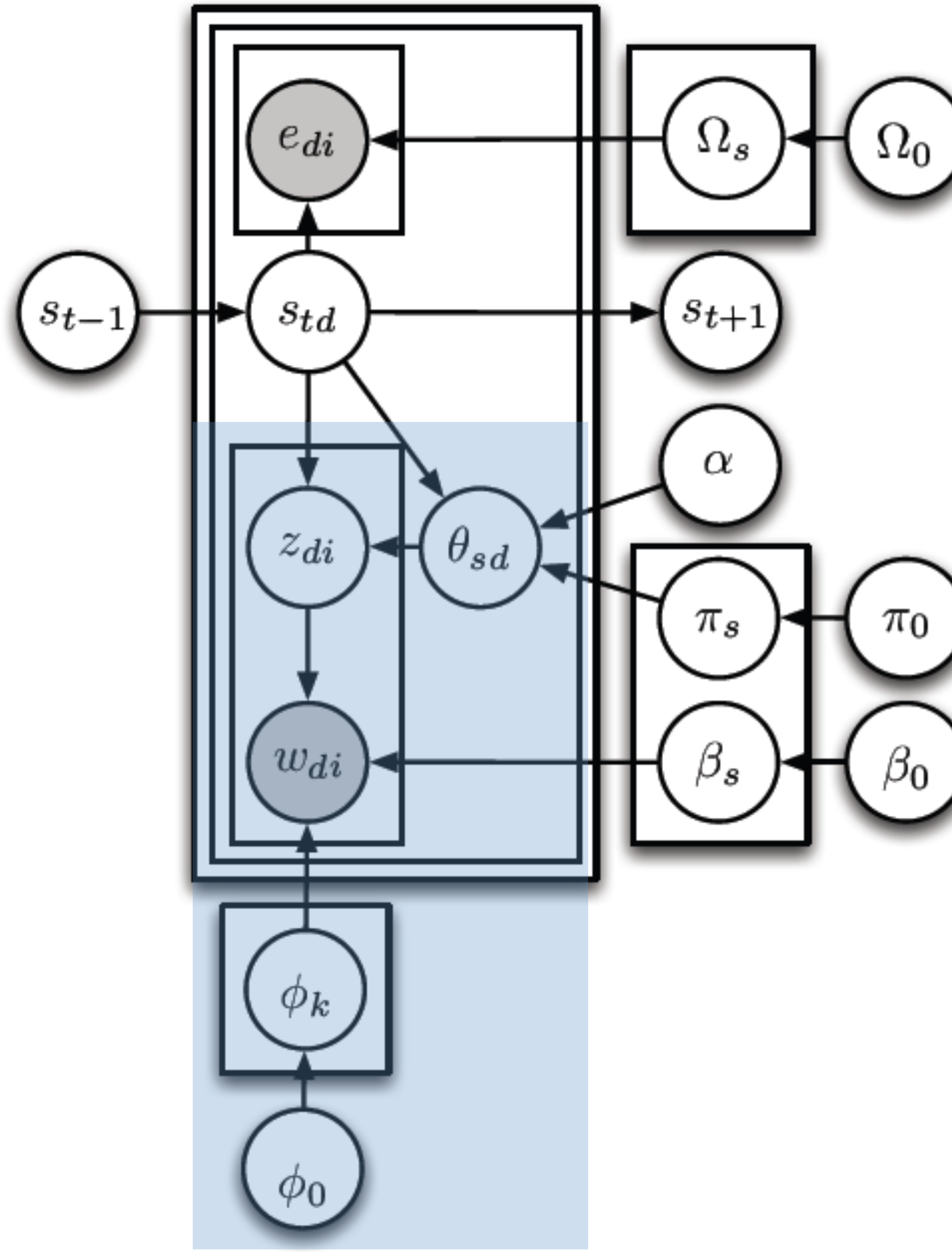




Storylines Model







Document’s topic mix is sampled from its story prior

Words inside a document either global or story specific

Generative process

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

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

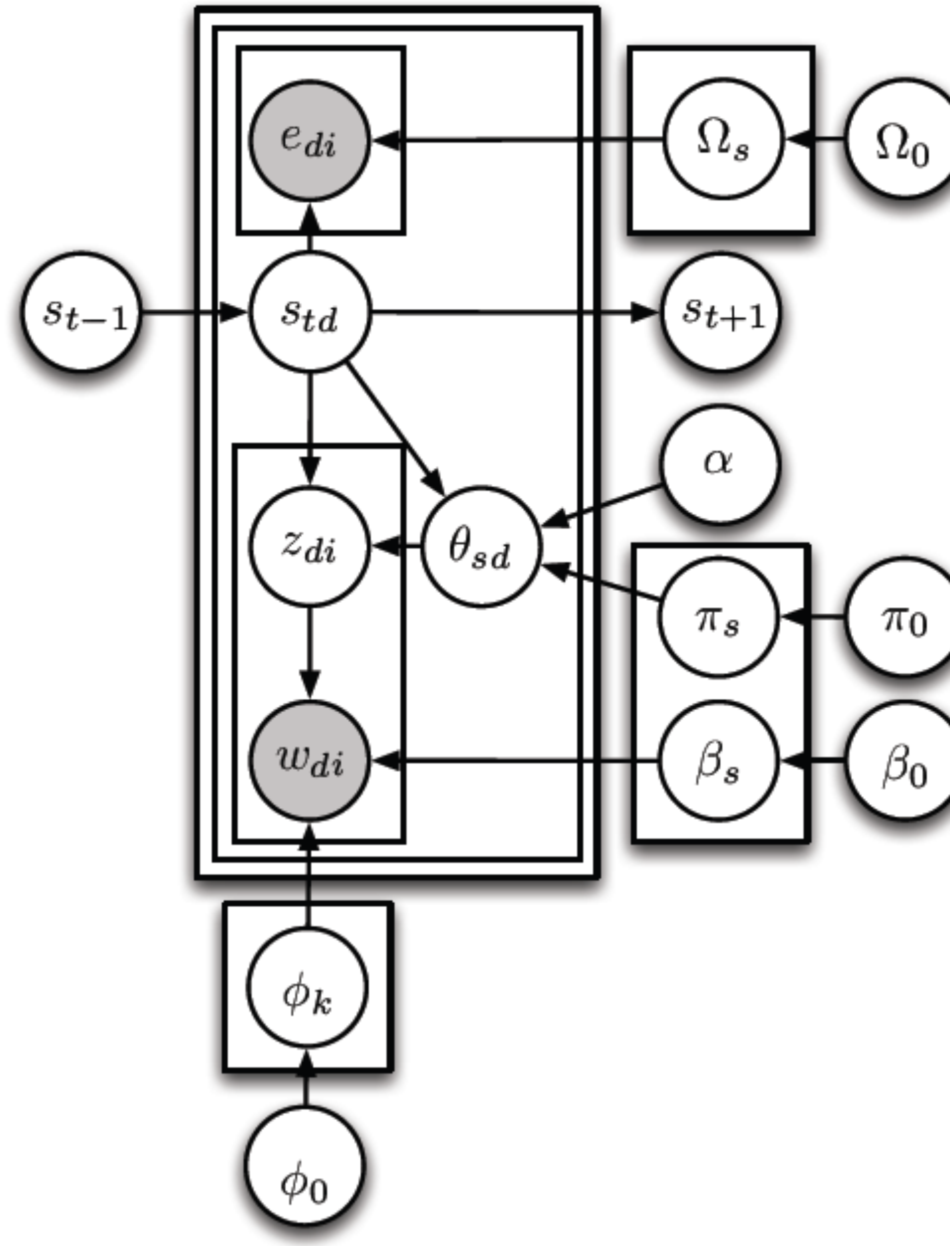
ii. Draw a distribution over named entities
$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$

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

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



Generative process

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

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

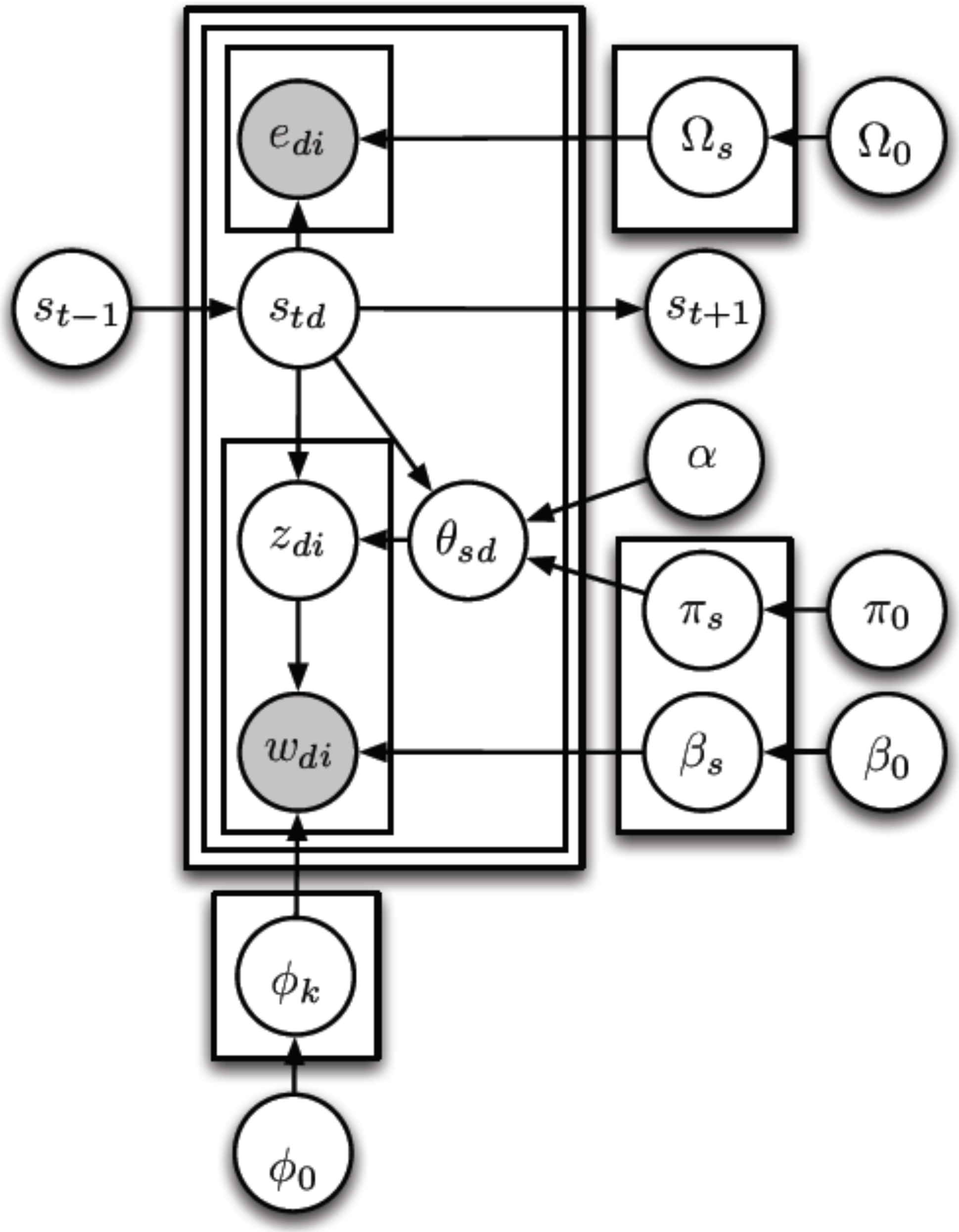
ii. Draw a distribution over named entities
$\Omega_{s_{new}} | G_0 \sim \mathrm{Dir}(\Omega_0)$

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

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

(d) Draw the words
$\mathbf{w}_{td} | s_{td} \sim \mathrm{LDA}\left(\theta_{s_{td}}, \{\phi_1, \cdots, \phi_K, \beta_{s_{td}}\}\right)$

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



Generative process

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

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

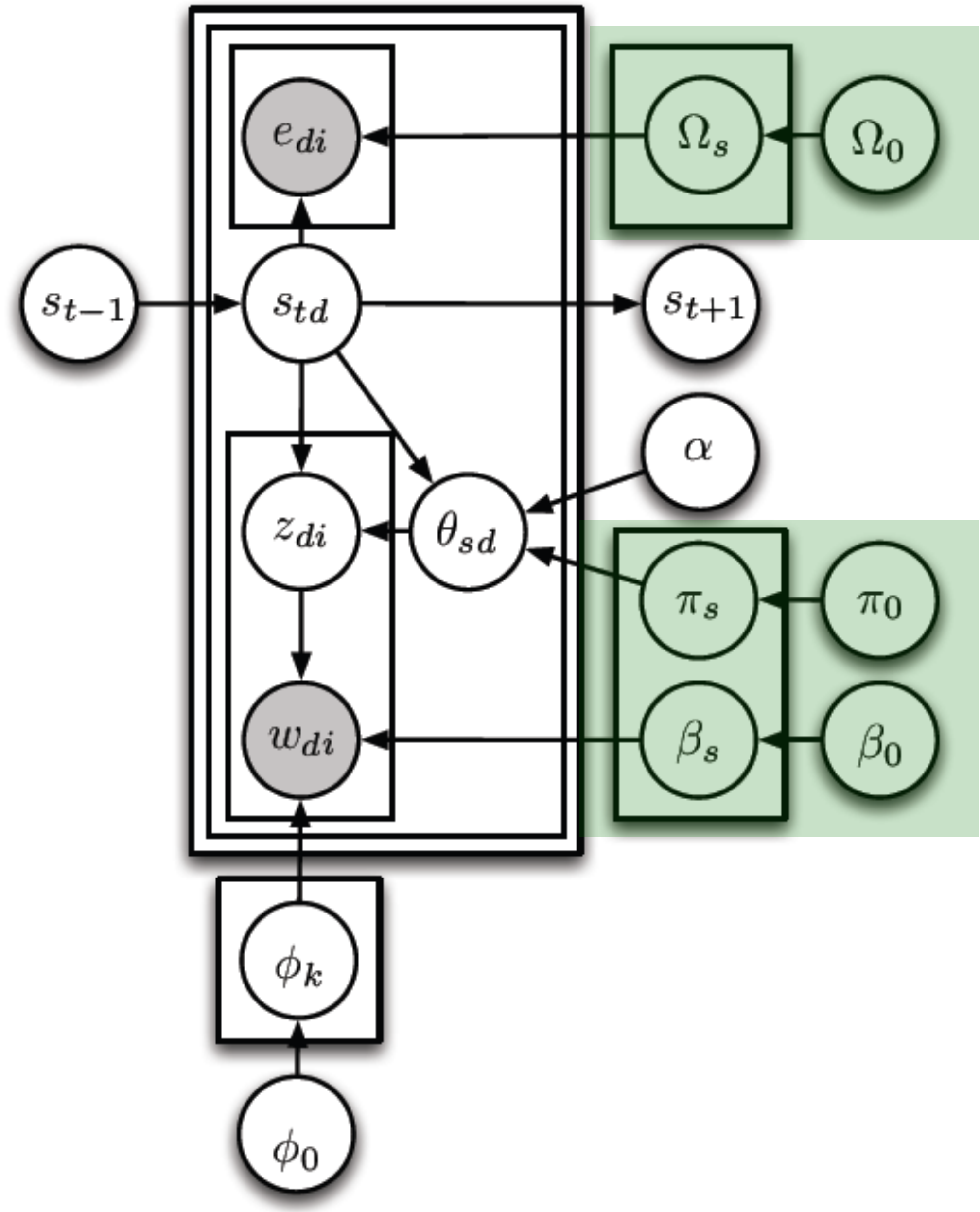
  ii. Draw a distribution over named entities $\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$

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

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



Generative process

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

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

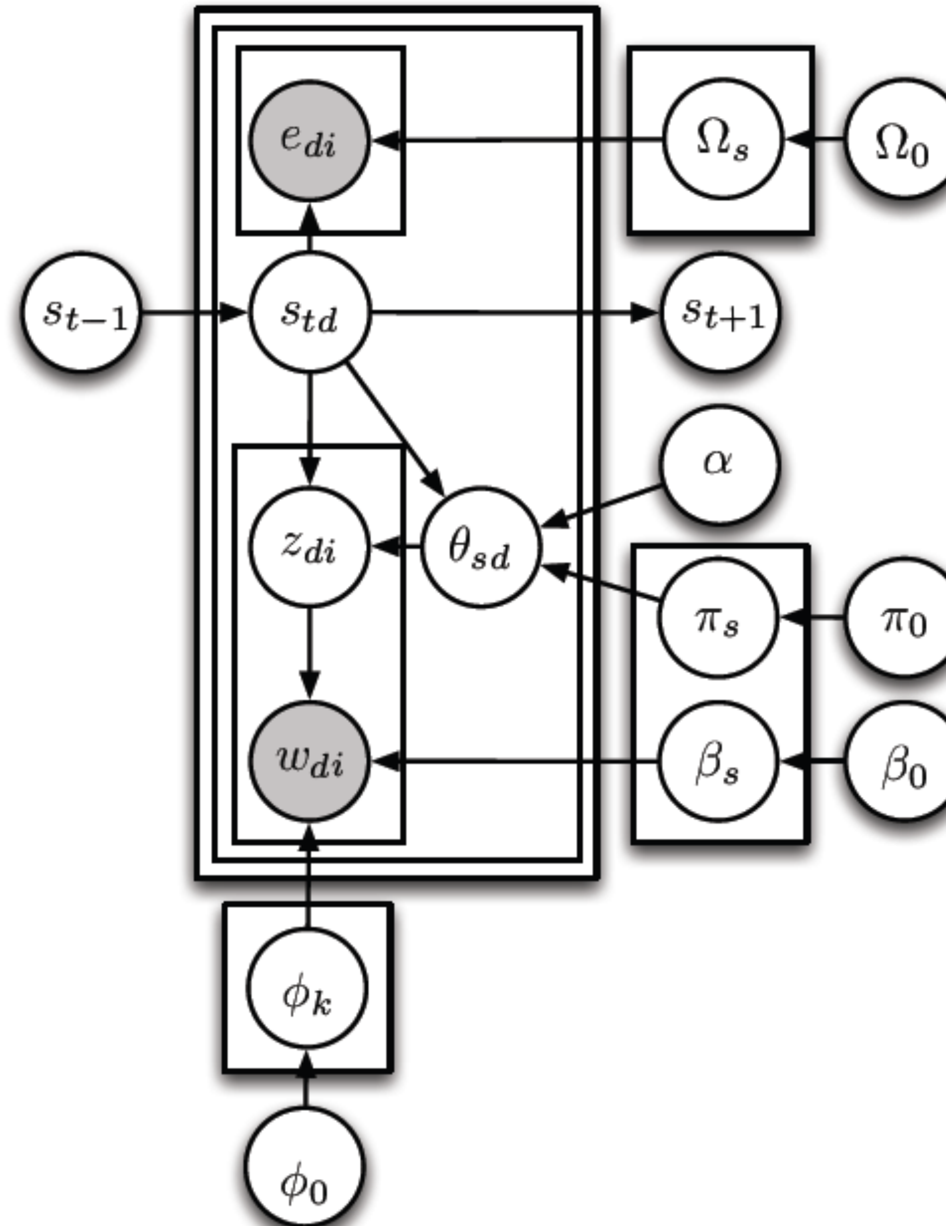
ii. Draw a distribution over named entities
$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$

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

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



Generative process

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

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

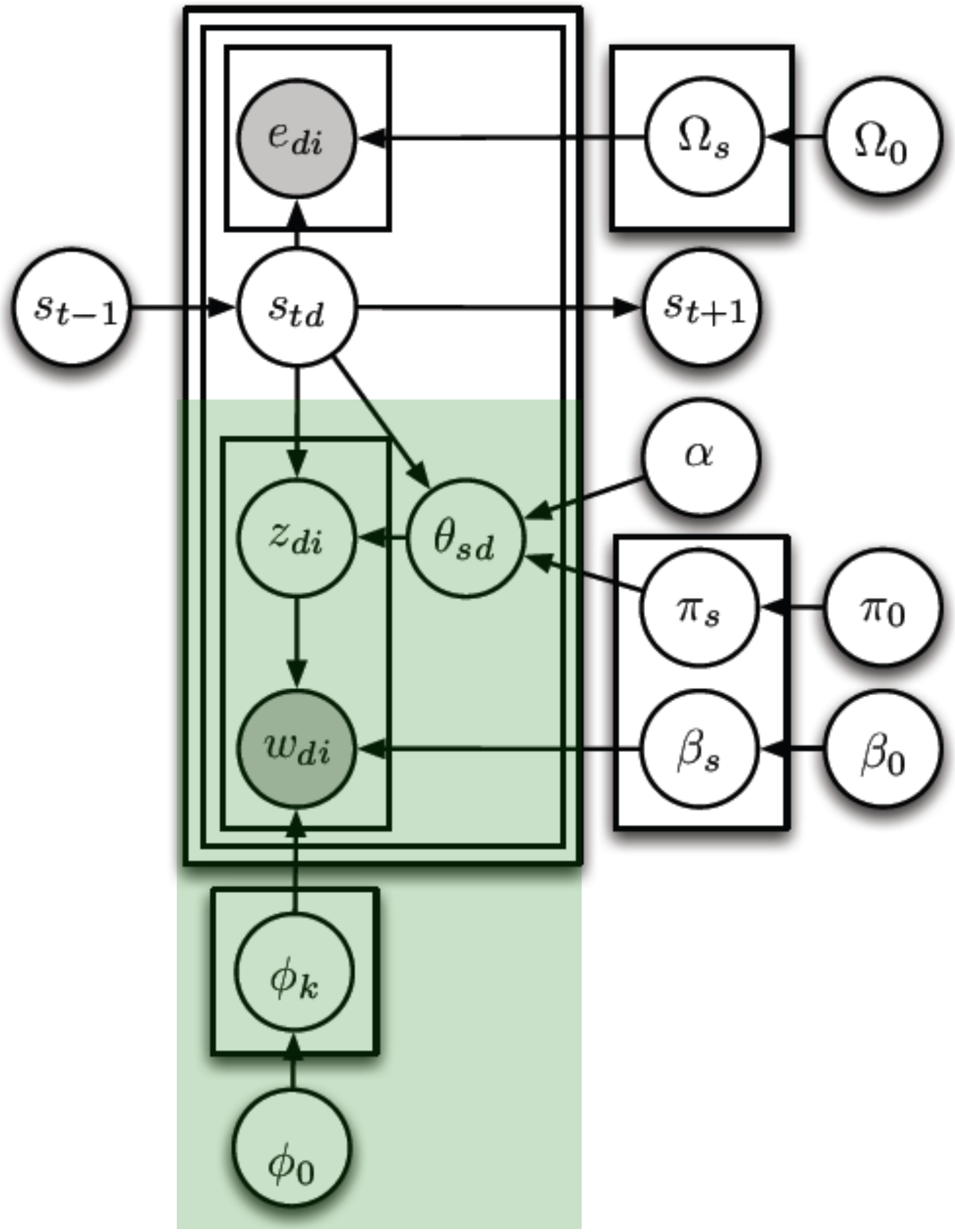
ii. Draw a distribution over named entities
$\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$

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

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



Generative process

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

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

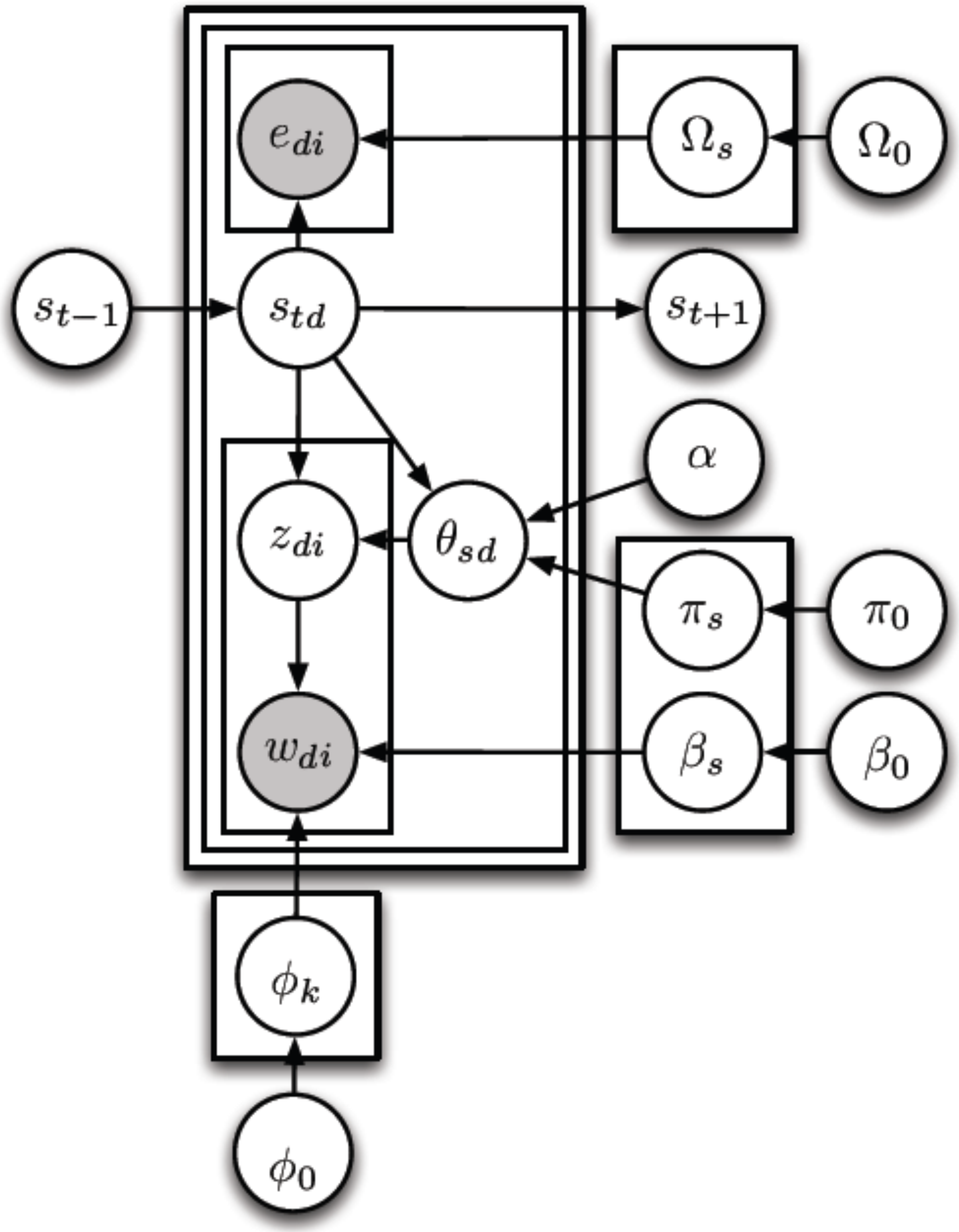
ii. Draw a distribution over named entities $\Omega_{s_{\text{new}}} | G_0 \sim \text{Dir}(\Omega_0)$

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

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



Generative process

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

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

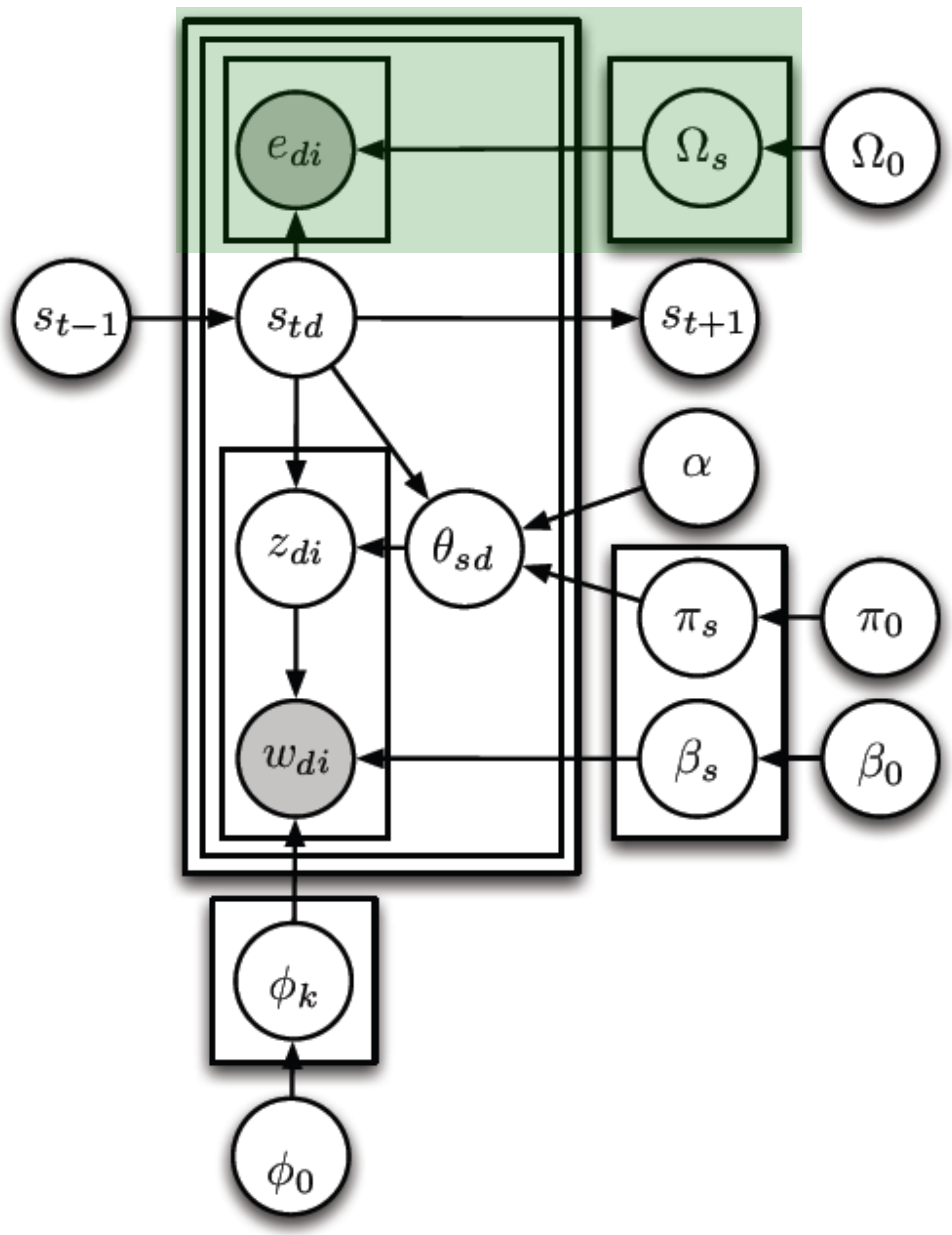
  ii. Draw a distribution over named entities $\Omega_{s_{\text{new}}}|G_0 \sim \text{Dir}(\Omega_0)$

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

(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

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

    using Gibbs Sampling for each particle
  - Reweight particle via

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

  - 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

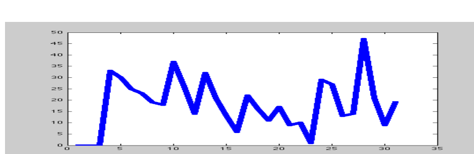
Unrest

police
attack
run
man
group
arrested
move

STORYLINES

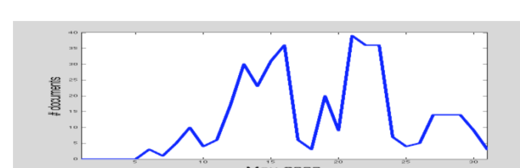
UEFA-soccer

champions	*Juventus*
goal	*AC Milan*
leg	*Real Madrid*
coach	*Milan*
striker	*Lazio*
midfield	*Ronaldo*
penalty	*Lyon*



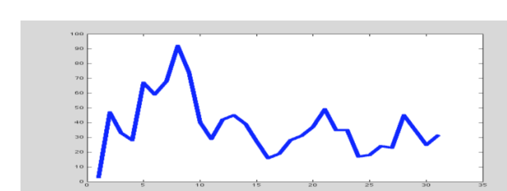
Tax bills

tax	*Bush*
billion	*Senate*
cut	*US*
plan	*Congress*
budget	*Fleischer*
economy	*White House*
lawmakers	*Republican*



India-Pakistan tension

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



Related Stories

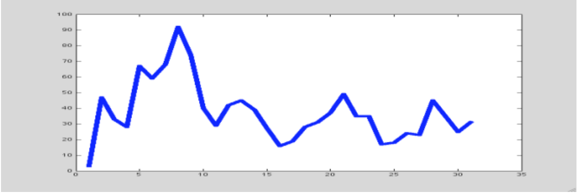
"Show similar stories by topic"

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

India-Pakistan tension

nuclear
border
dialogue
diplomatic
militant
insurgency
missile

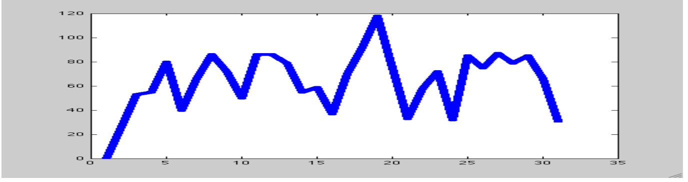
Pakistan
India
Kashmir
New Delhi
Islamabad
Musharraf
Vajpayee



Middle-east conflict

Peace
Roadmap
Suicide
Violence
Settlements
bombing

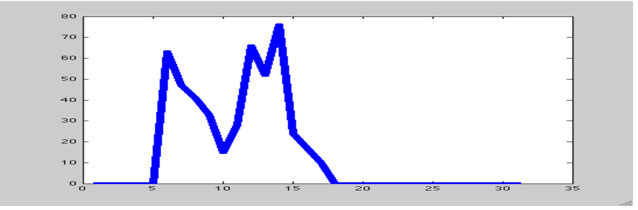
Israel
Palestinian
West bank
Sharon
Hamas
Arafat



North Korea nuclear

nuclear
summit
warning
policy
missile
program

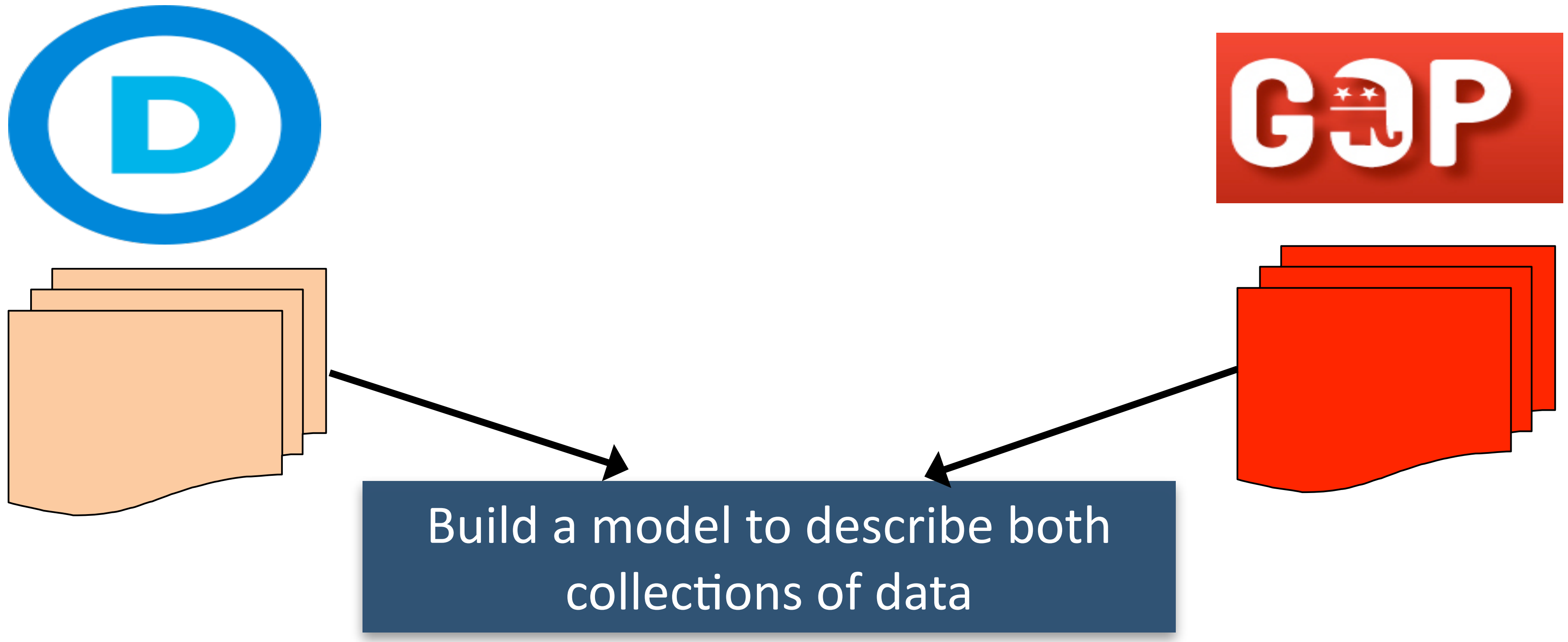
North Korea
South Korea
U.S
Bush
Pyongyang



Detecting Ideologies

Ahmed and Xing, 2010

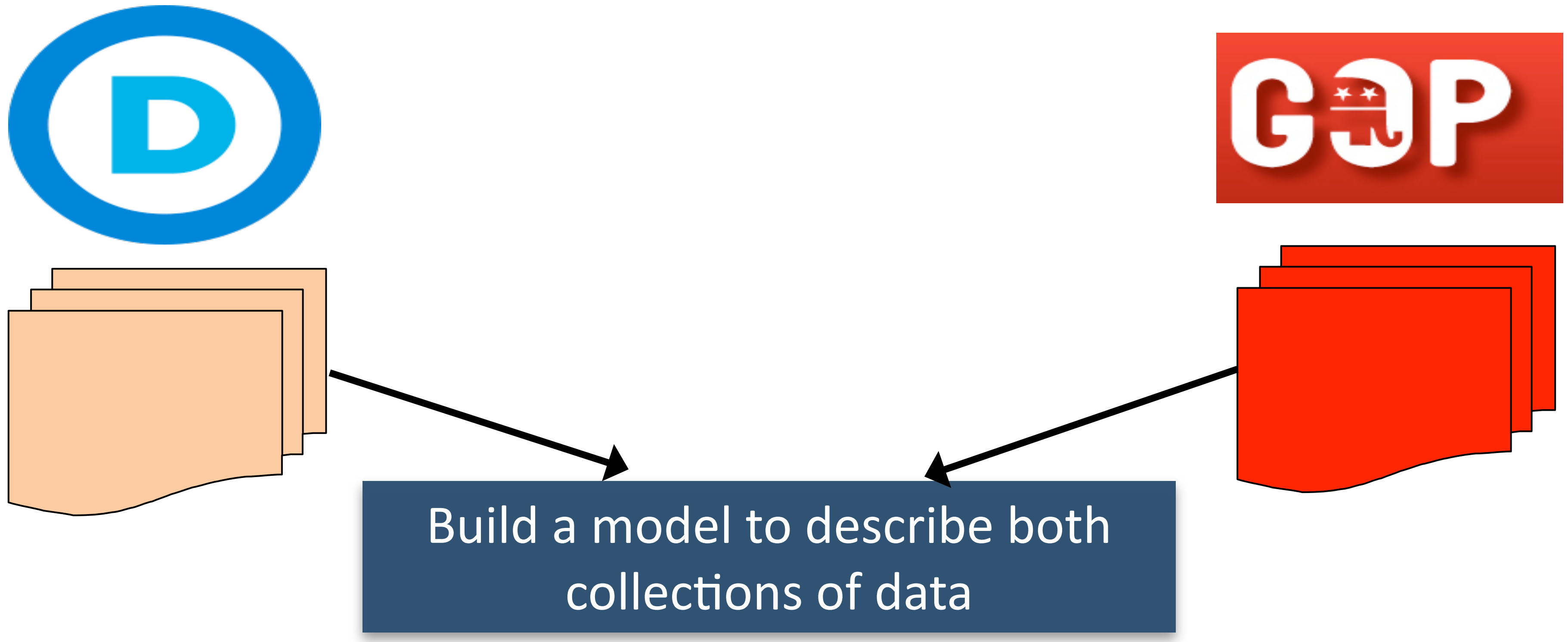
Ideologies



Visualization

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

Ideologies

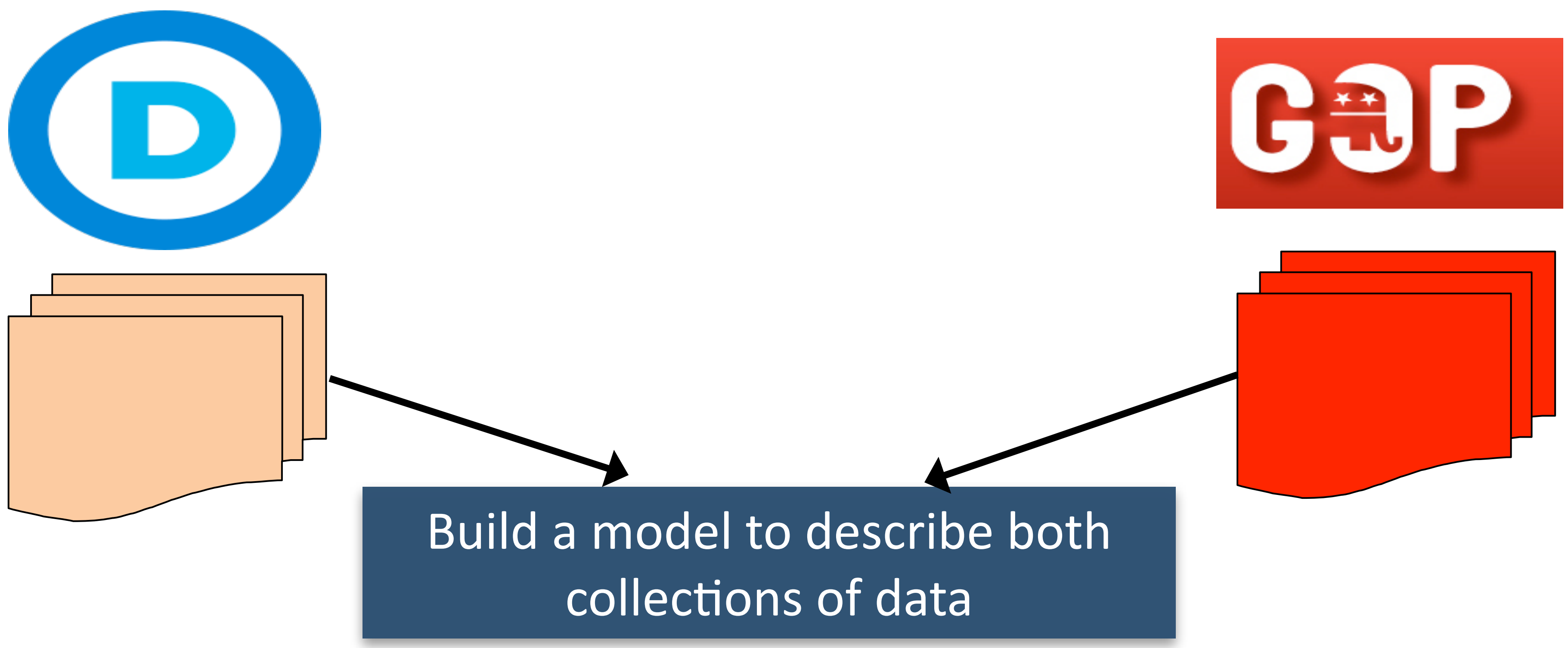


Visualization

Classification

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

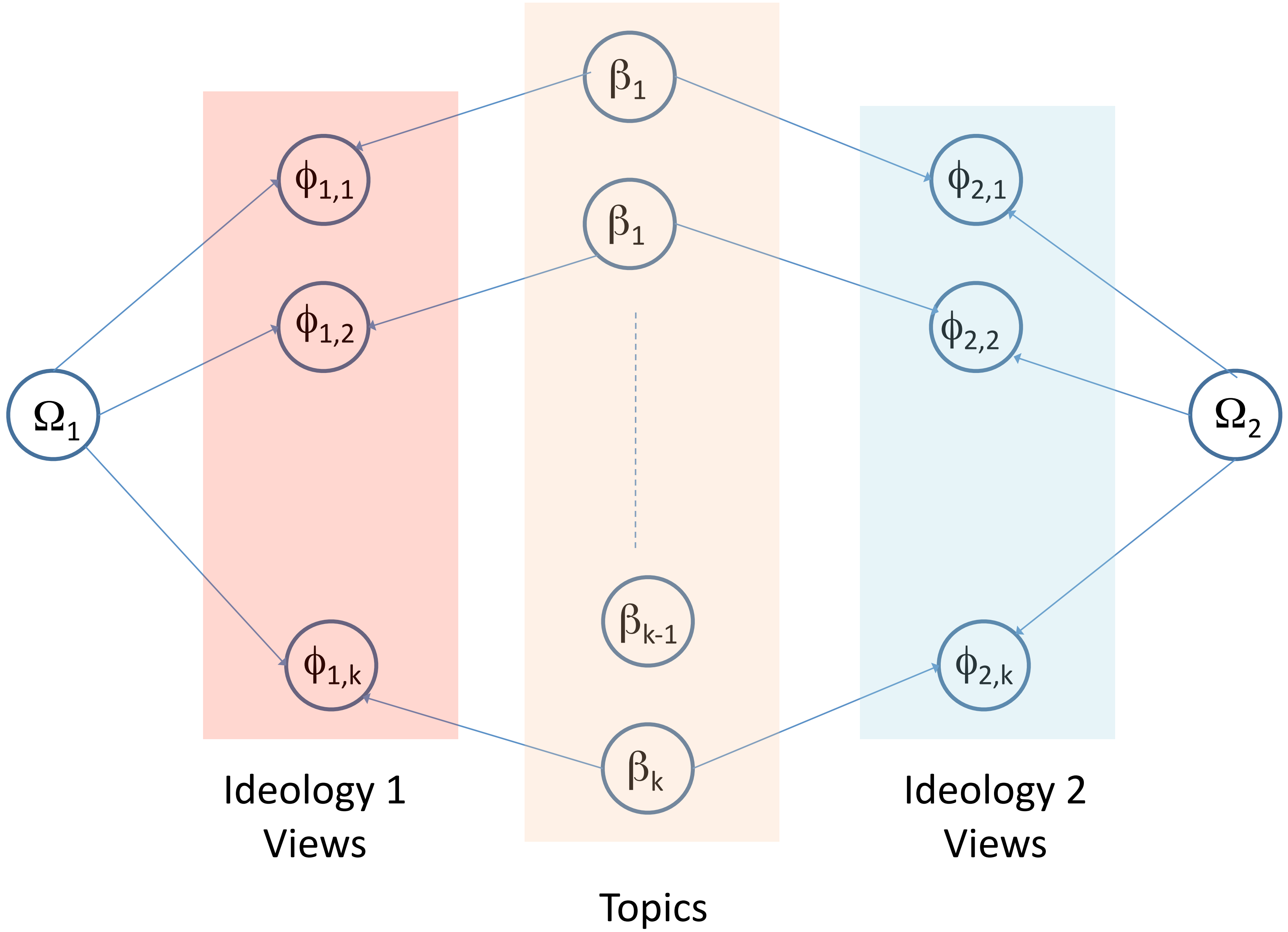
Ideologies



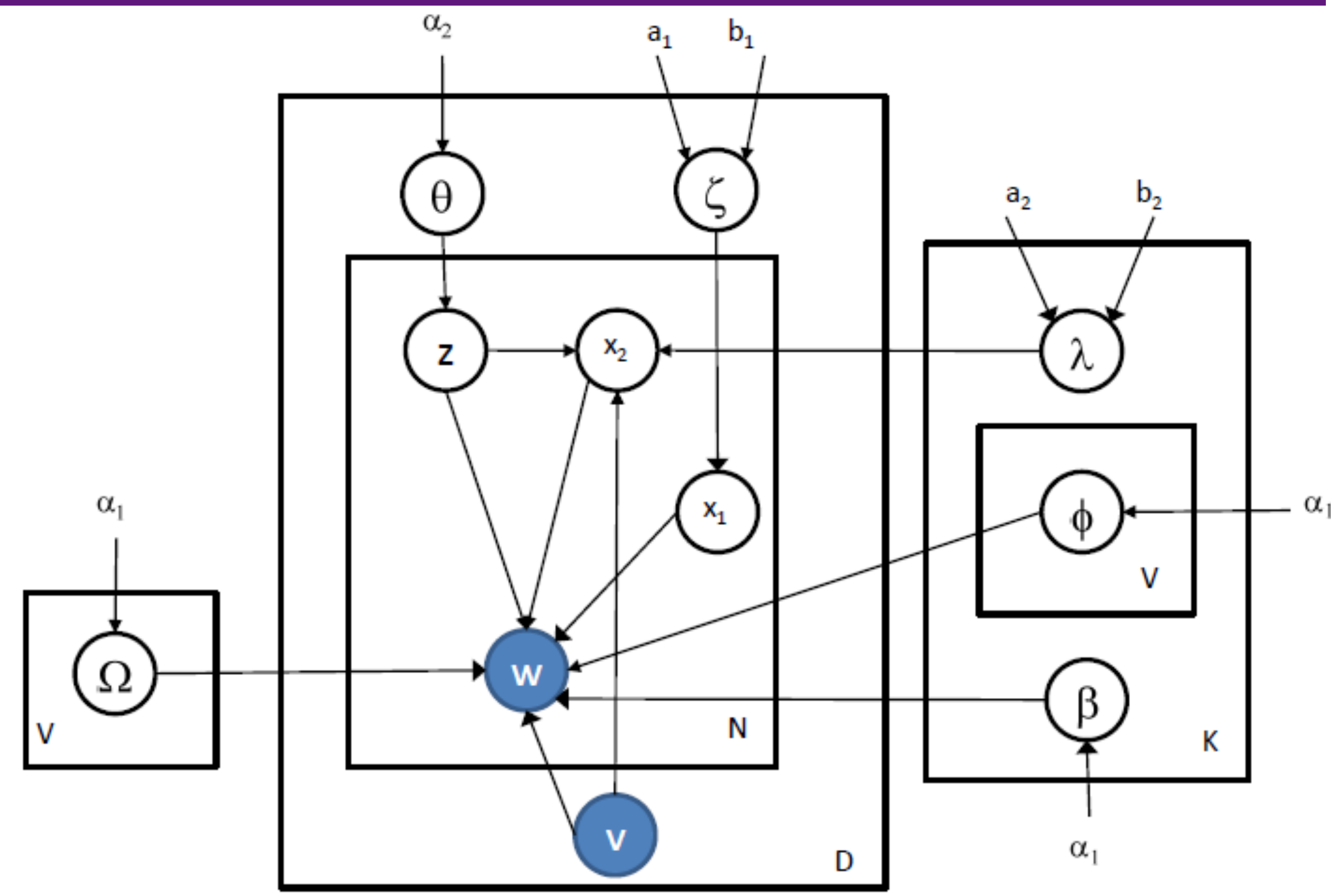
- Visualization
- Classification
- Structured browsing

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

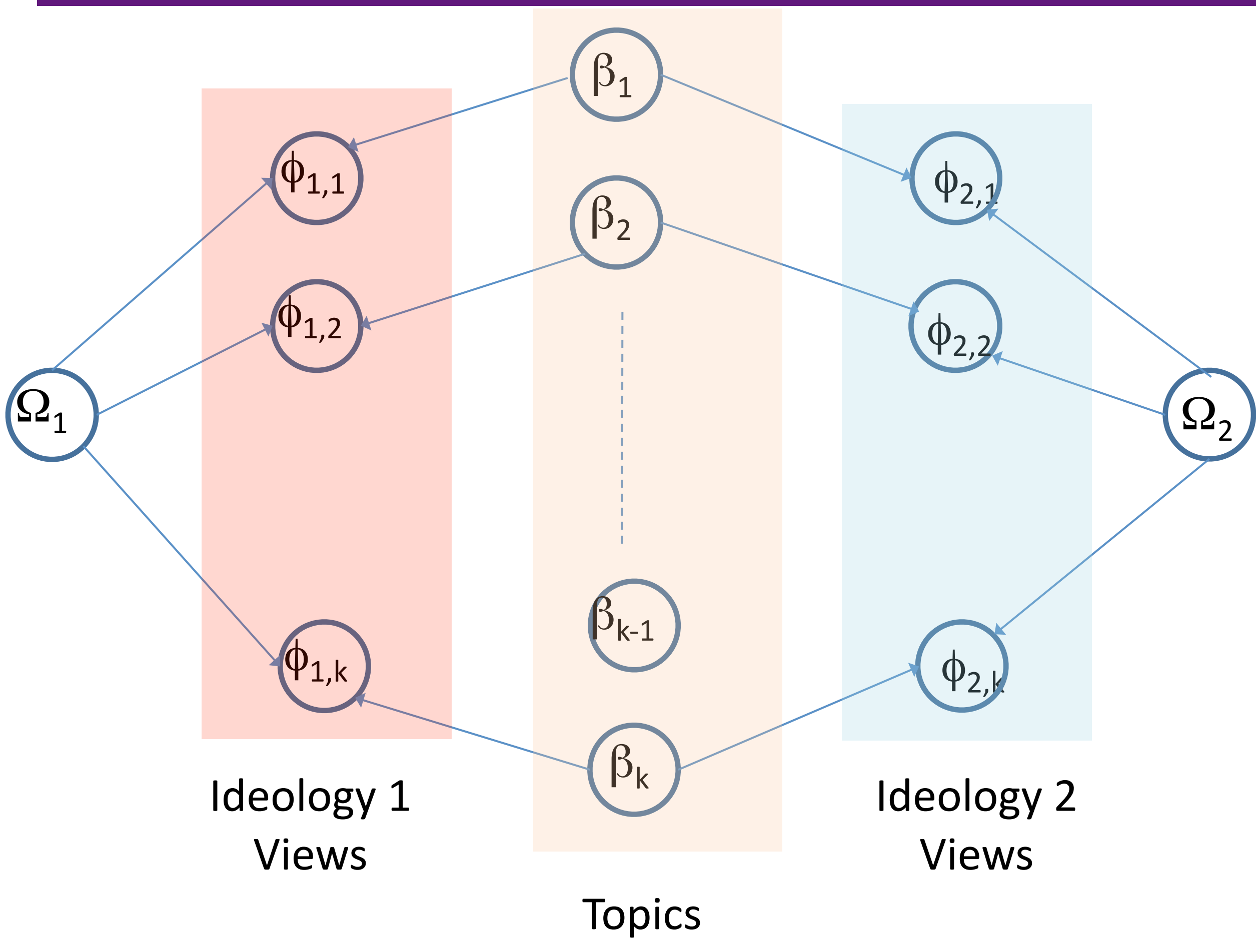
Building a factored model

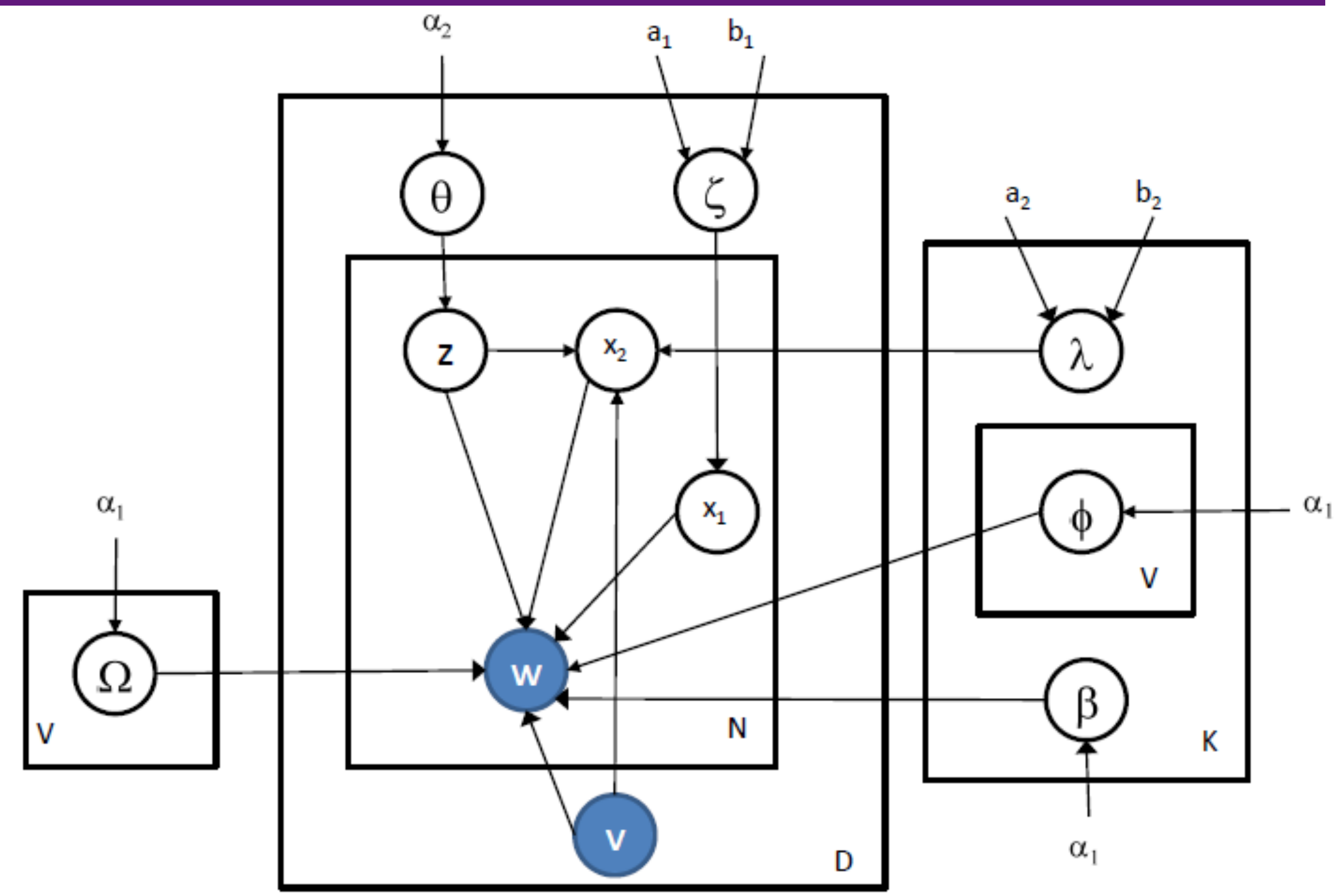


Building a factored model

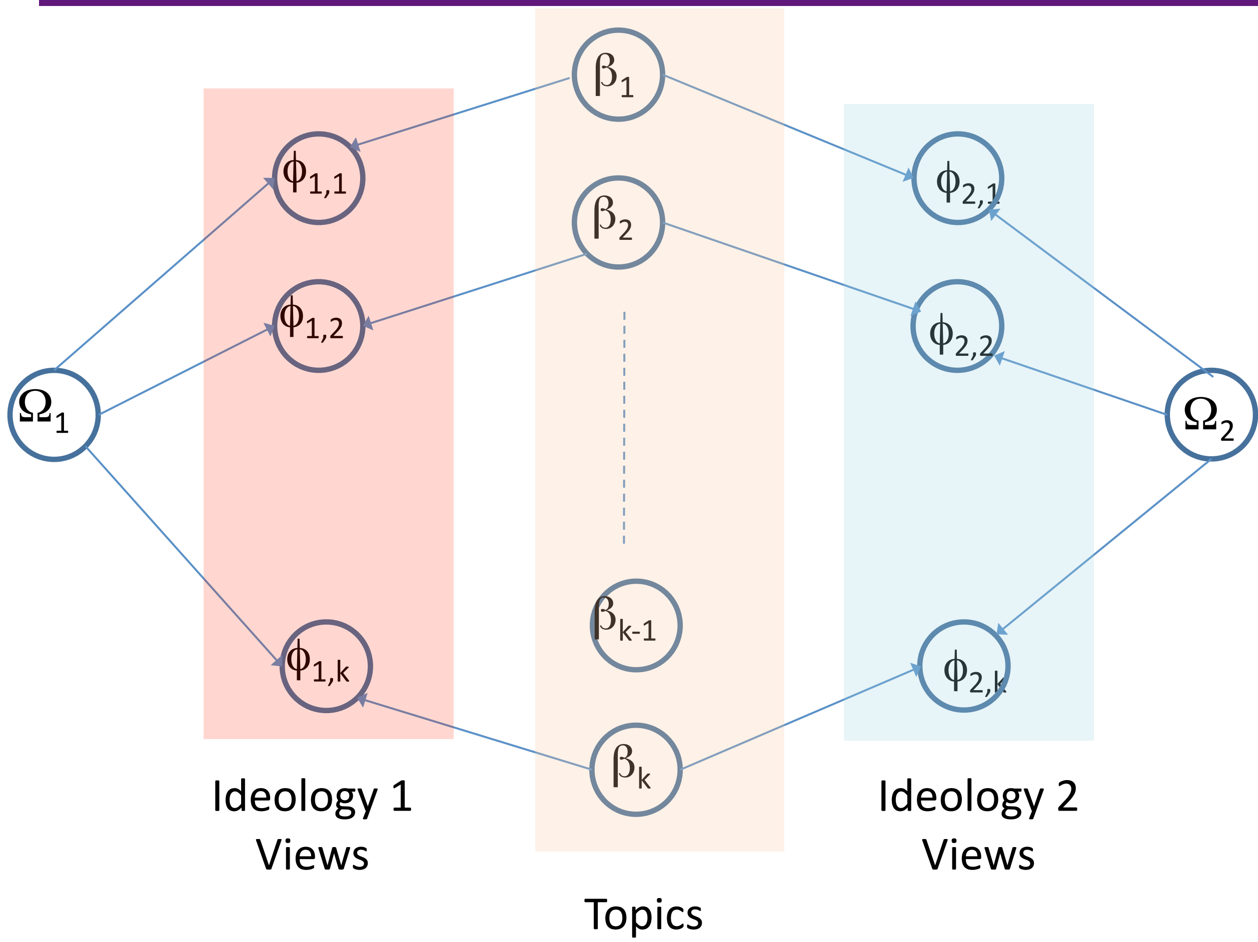


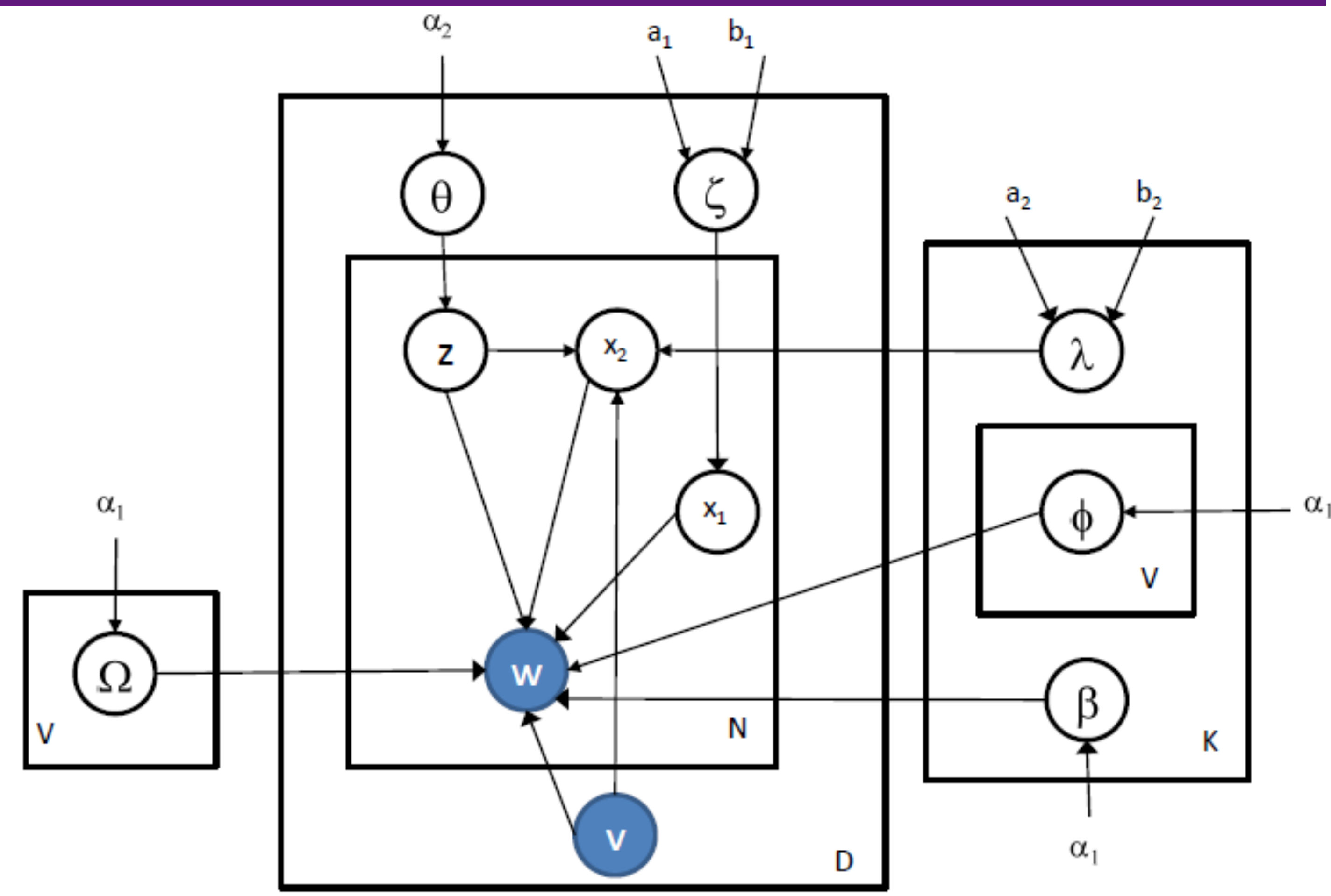
Building a factored model

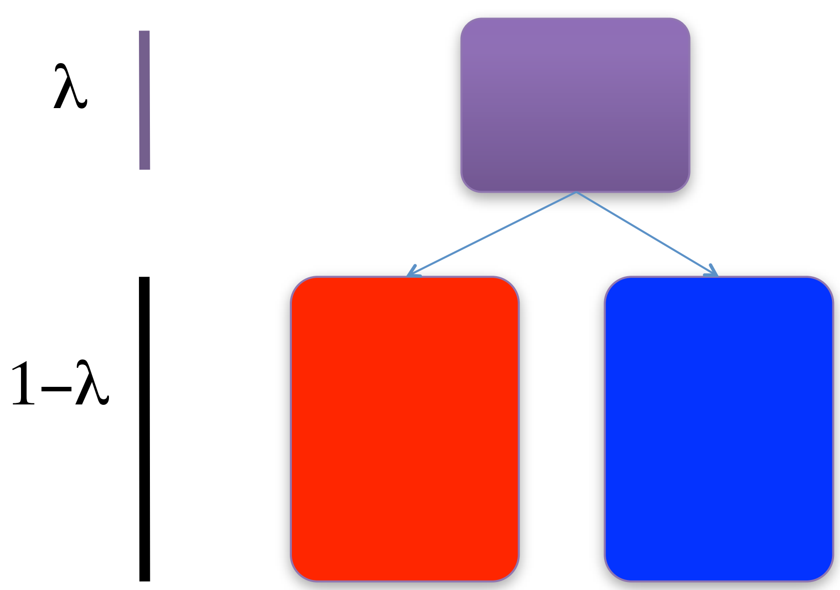


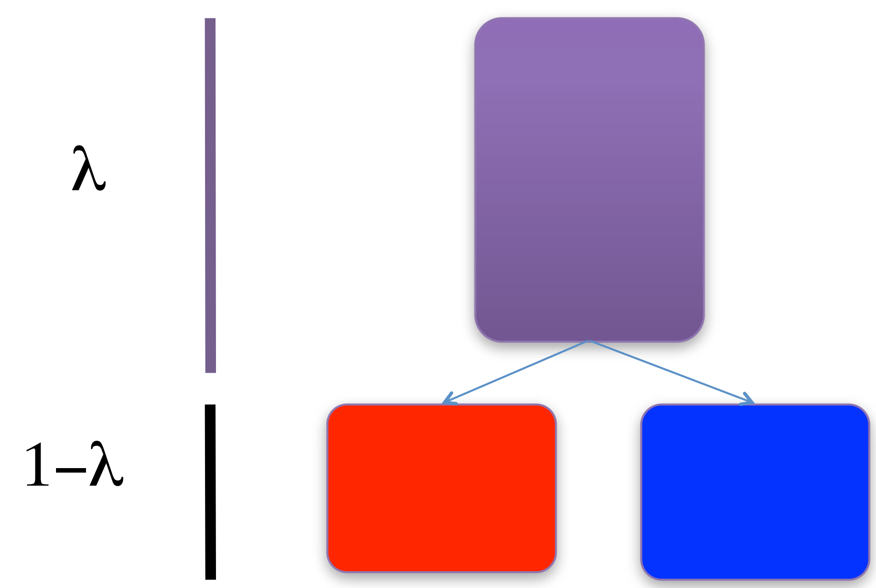


Building a factored model





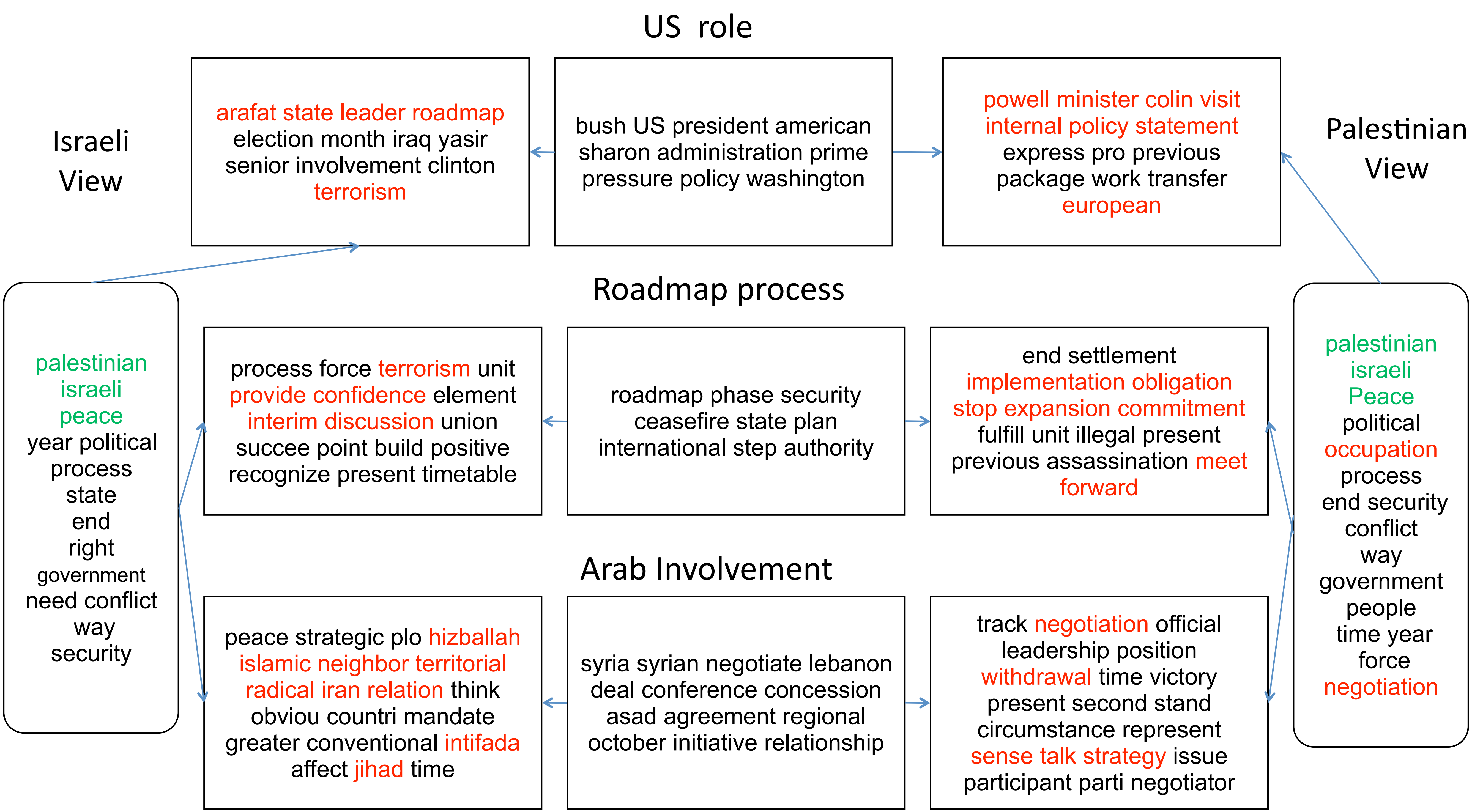




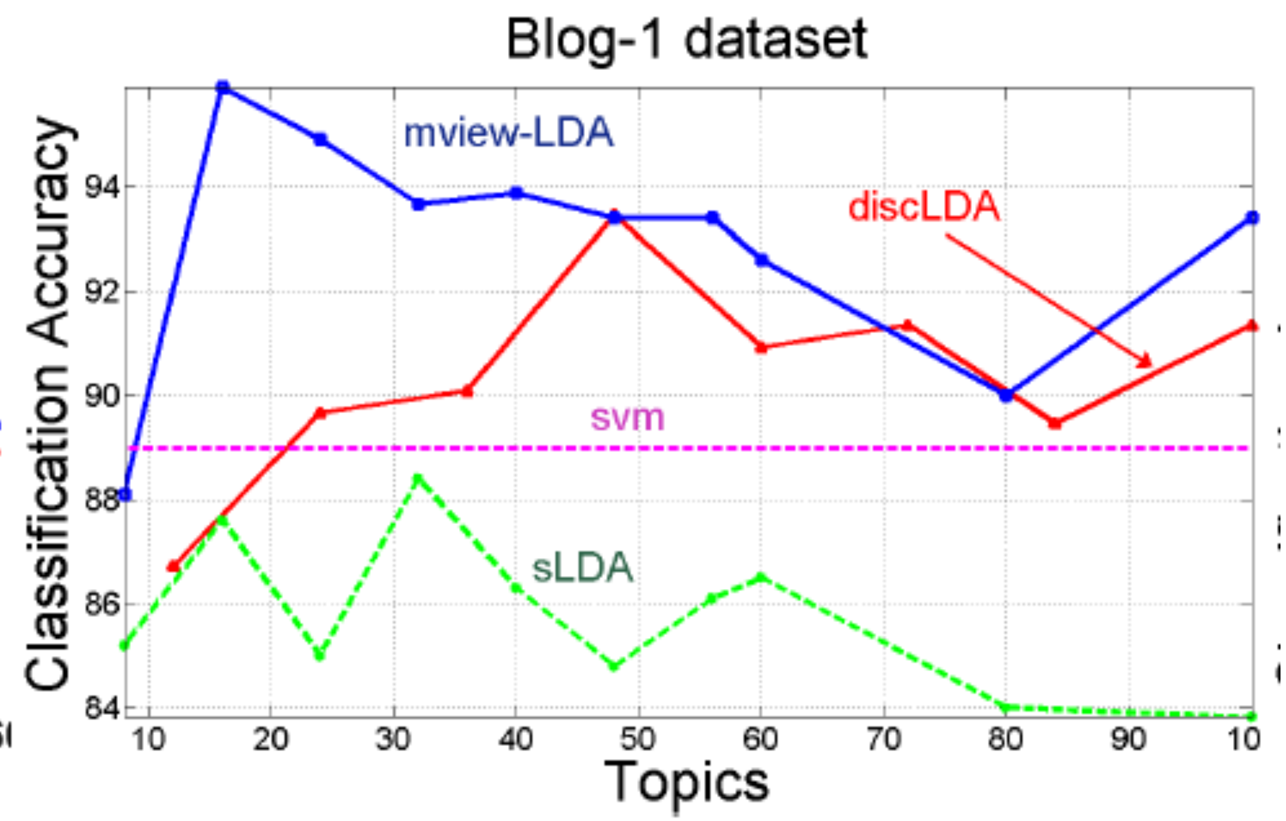
Data

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

Bitterlemons dataset

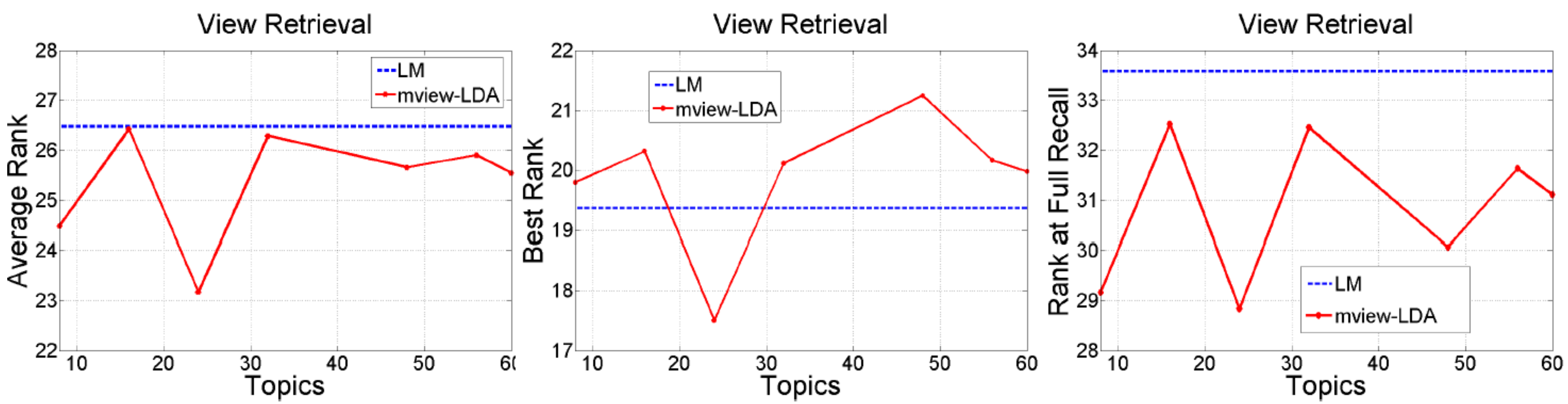


Classification accuracy



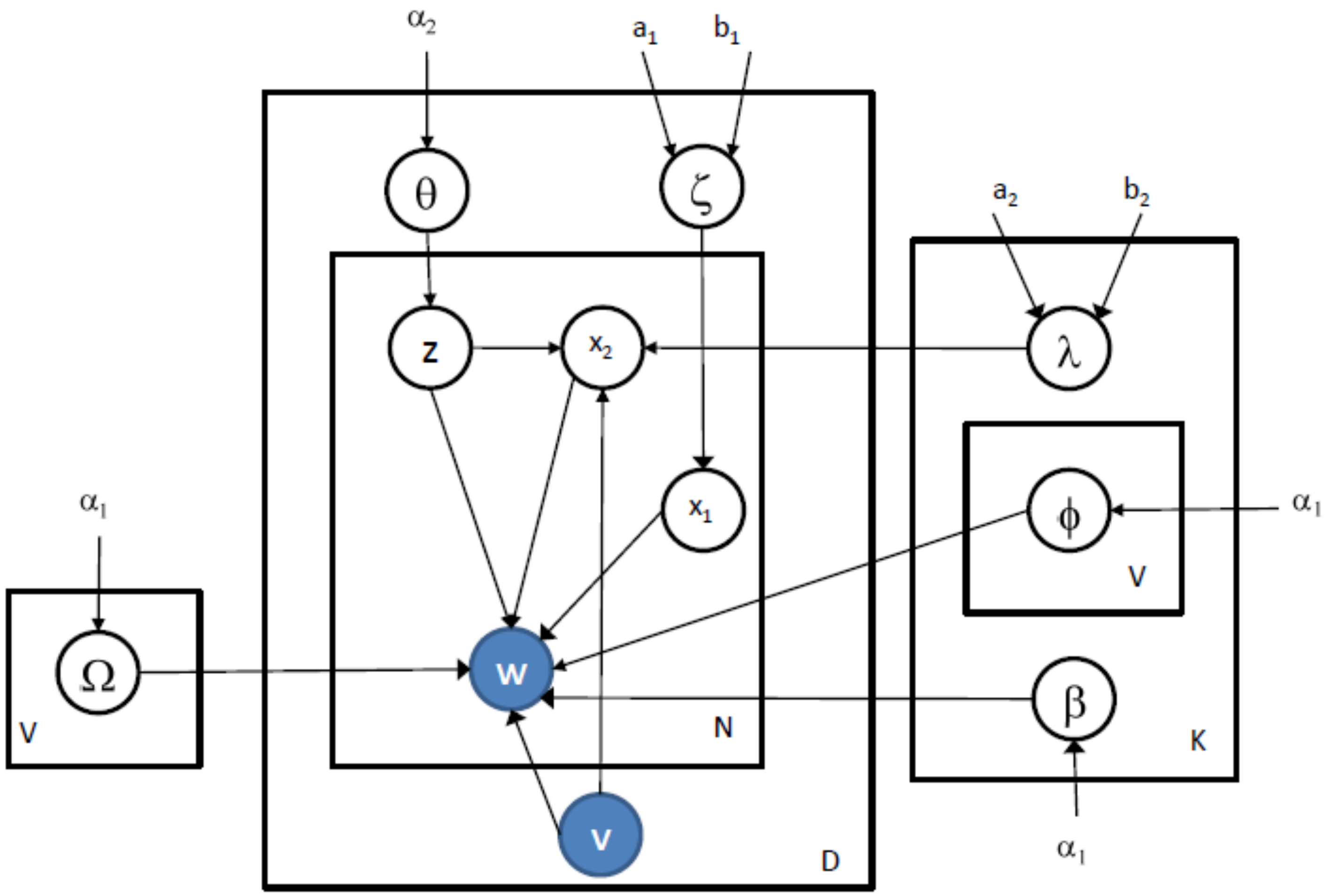
Generalization to new blogs

Finding alternate views

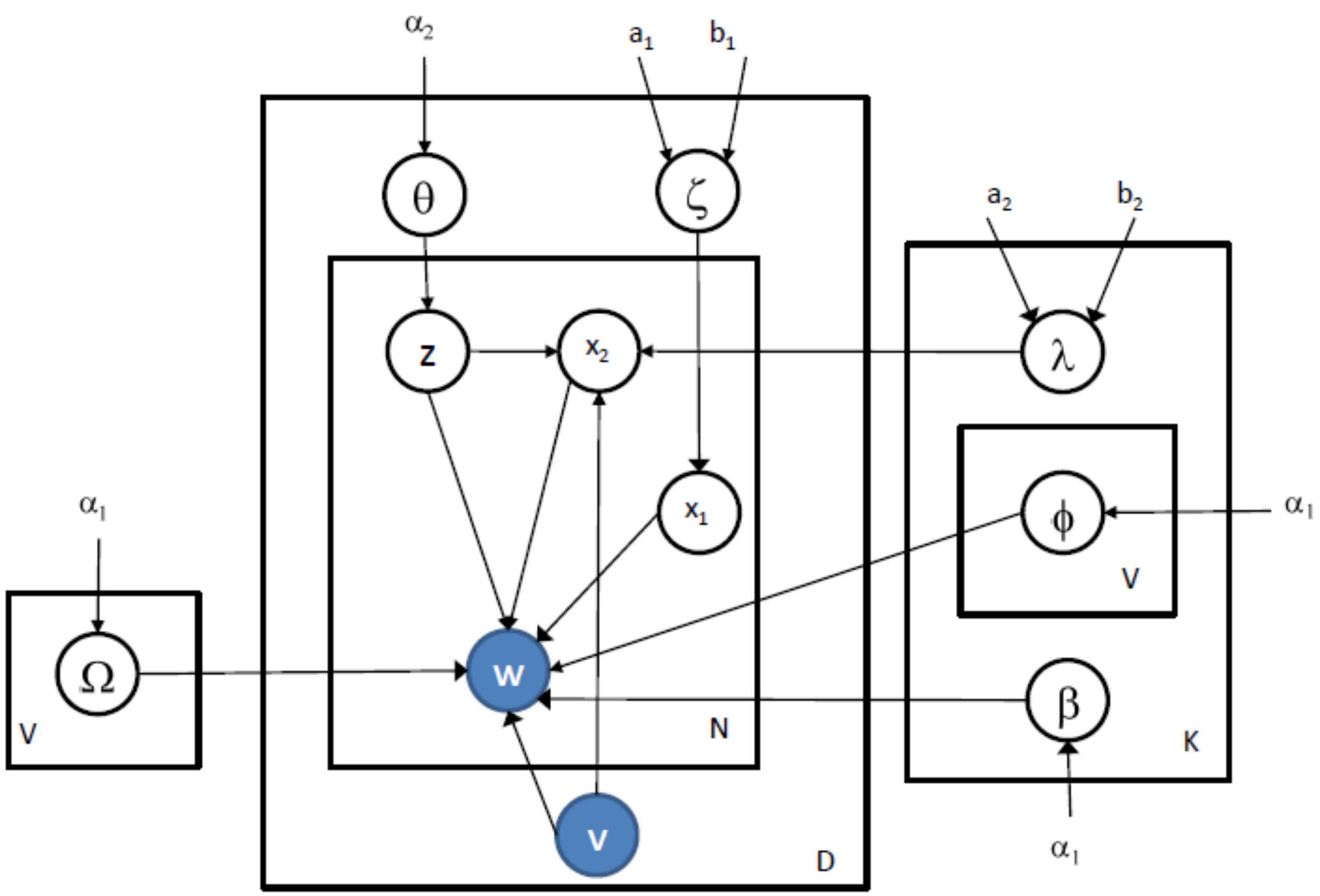


- Given a document written in one ideology, retrieve the equivalent
- Baseline: SVM + cosine similarity

Unlabeled data

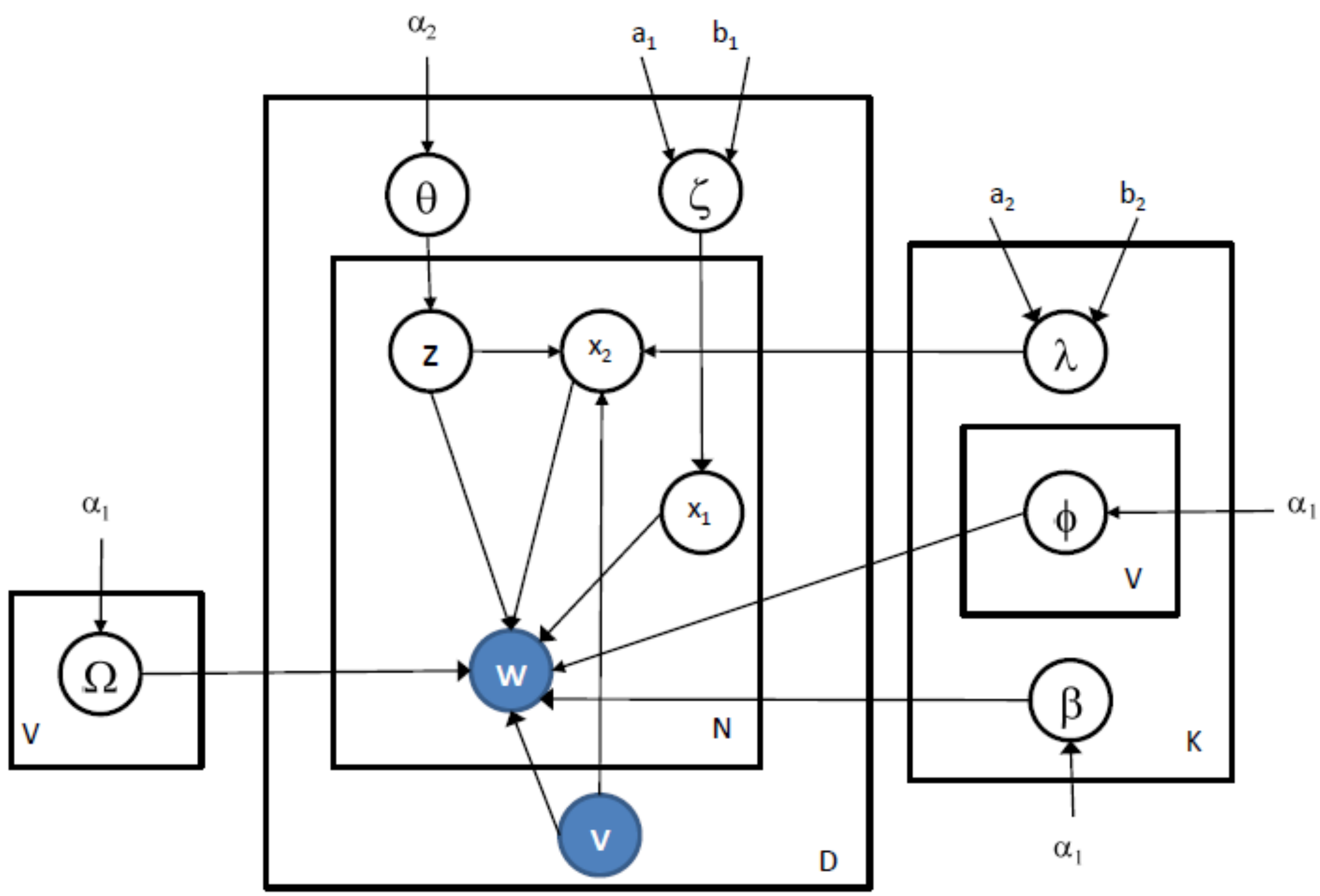


Unlabeled data



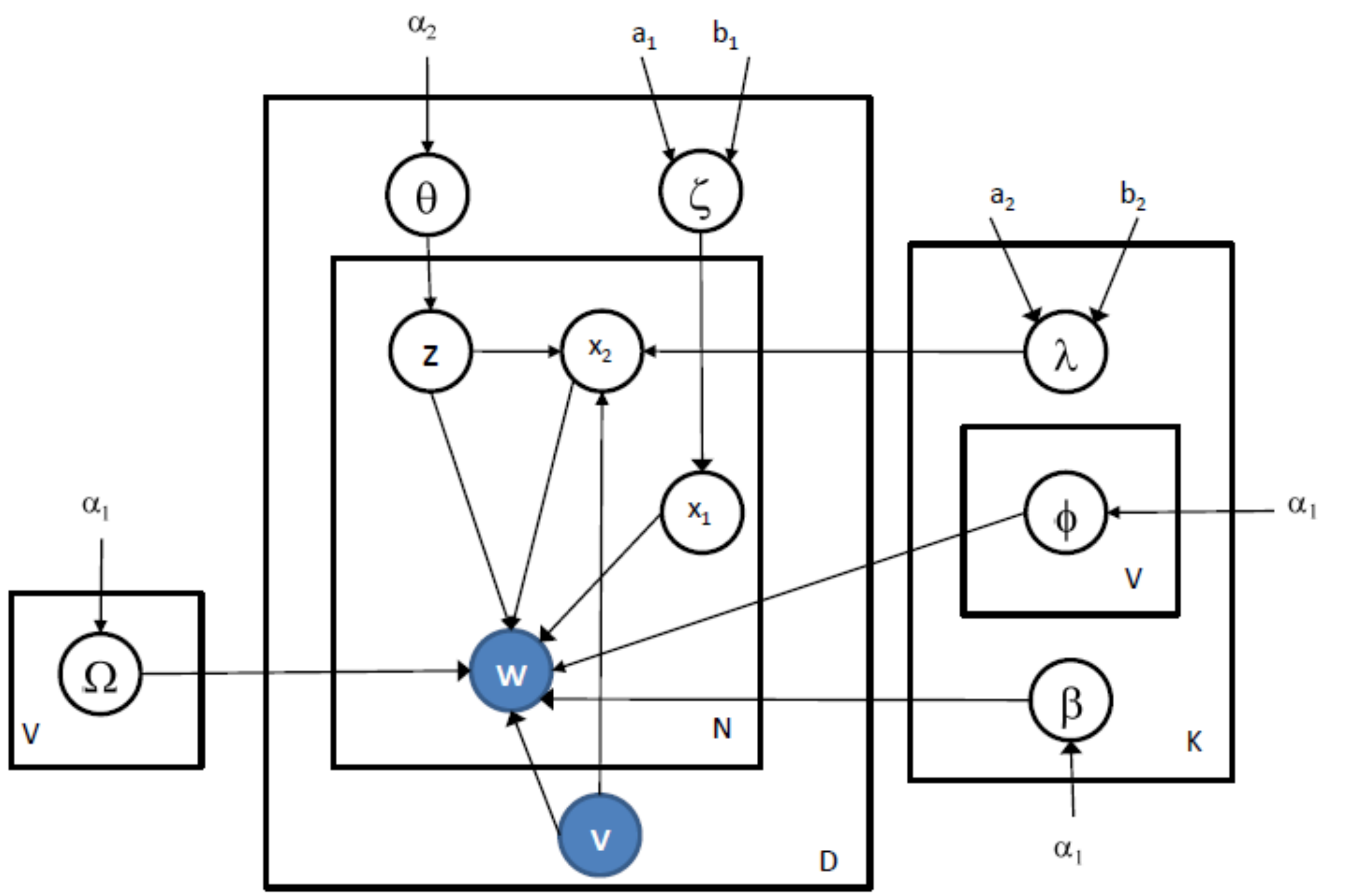
- In theory this is simple
    - Add a step that samples the document view (v)
    - **Doesn't mix** in practice because tight coupling between v and ($\mathbf{x}_1$,$\mathbf{x}_2$,$\mathbf{z}$)

Unlabeled data



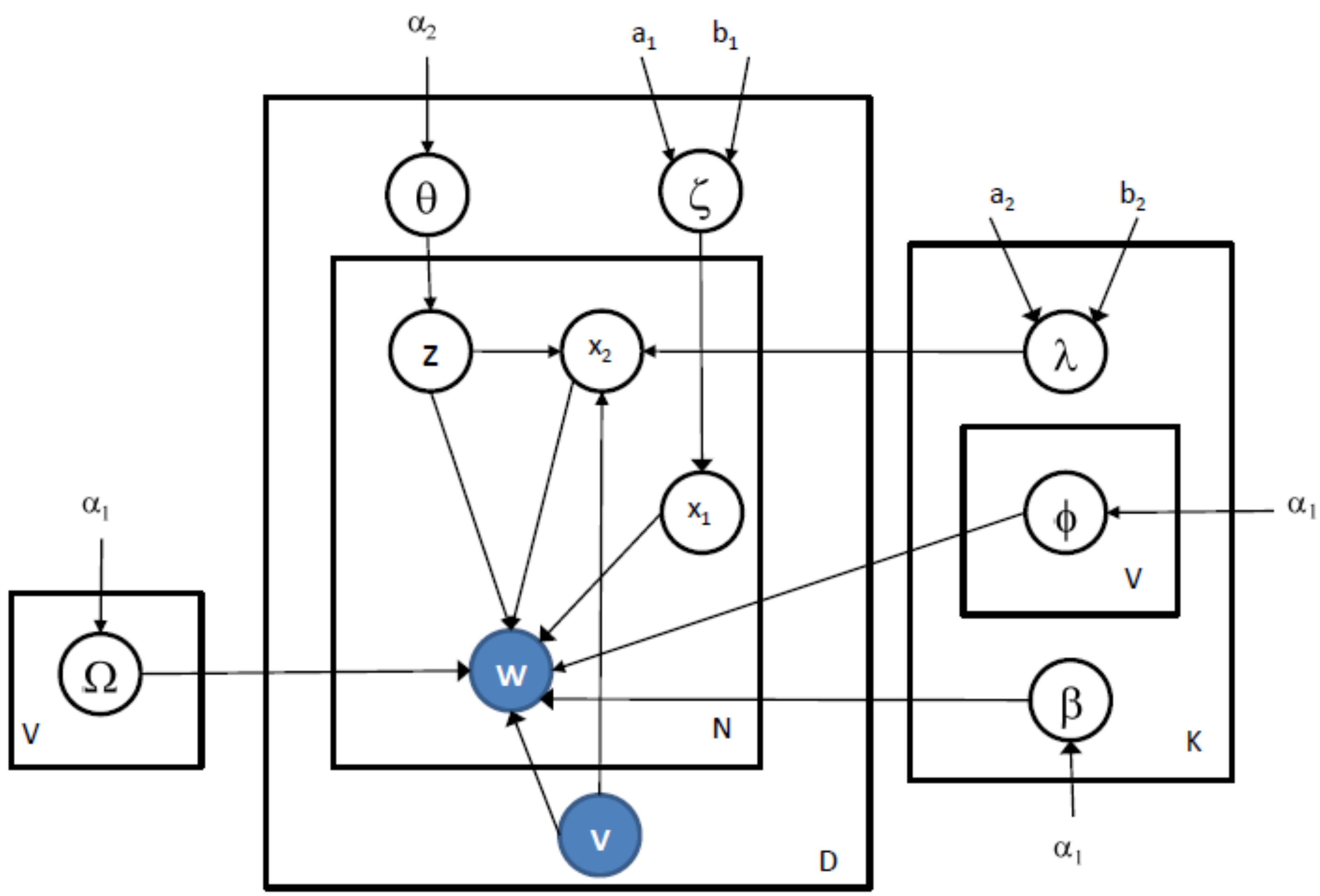
- In theory this is simple
  - Add a step that samples the document view (v)
  - **Doesn't mix** in practice because tight coupling between v and (**x_1**,**x_2**,**z**)
- Solution

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