



# Big Data and Large Scale Inference

Amr Ahmed & Alex Smola

Research at Google

# Data on the Internet



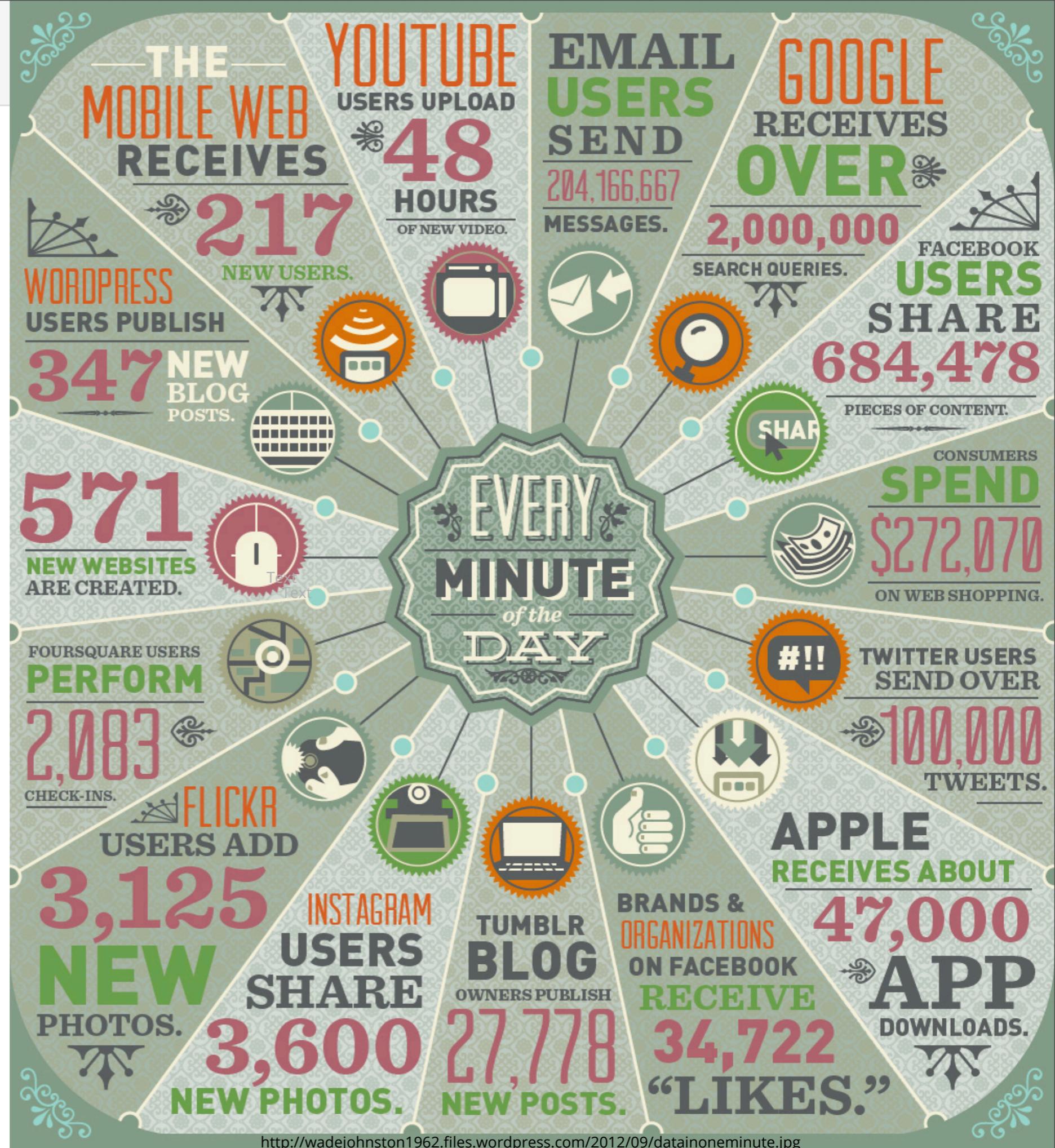
# Size calibration

- **Tiny** (2 cores)  
(512MB, 50MFlops, 1000 examples)
- **Small** (4 cores)  
(4GB, 10GFlops, 100k examples)
- **Medium** (16 cores)  
(32GB, 100GFlops, 1M examples)
- **Large** (256 cores)  
(512GB, 1TFlops, 100M examples)
- **Massive**  
... need to work hard to make it work



This  
is  
not  
a  
toy

dataset



# Google™ User generated content

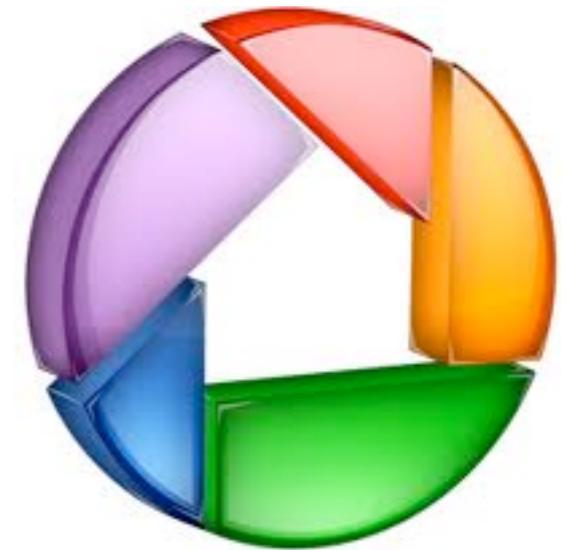
- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared, Google Now)
- 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)
- Link sharing (Facebook, Delicious, Buzz)
- Network traffic



flickr™



DISQUS



You Tube

yelp® 

Source: place source info here

>1B images, 40h video/minute

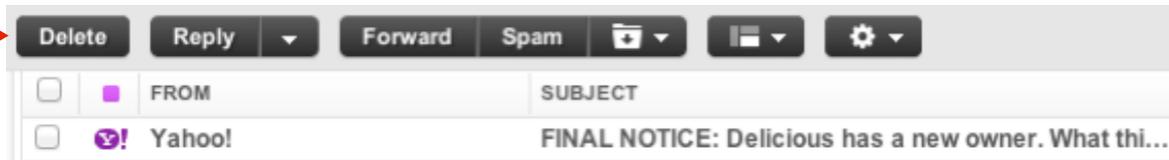
- Ads



- Click feedback



- Emails

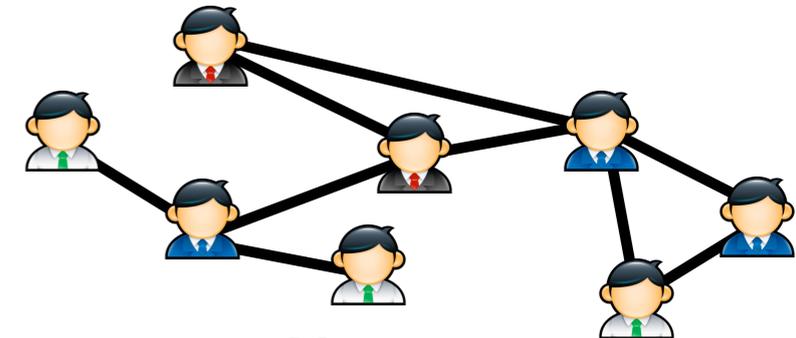


- Tags



- Location

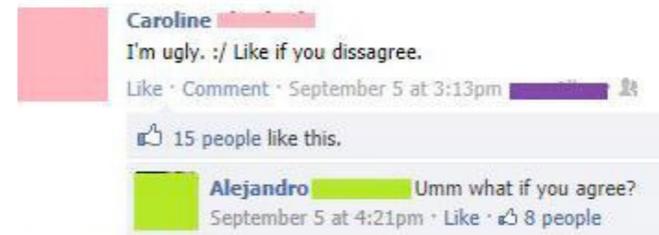
- Graphs



- Document collections



- Email/IM/Discussions



- Query stream



labeled

label free

# Summary

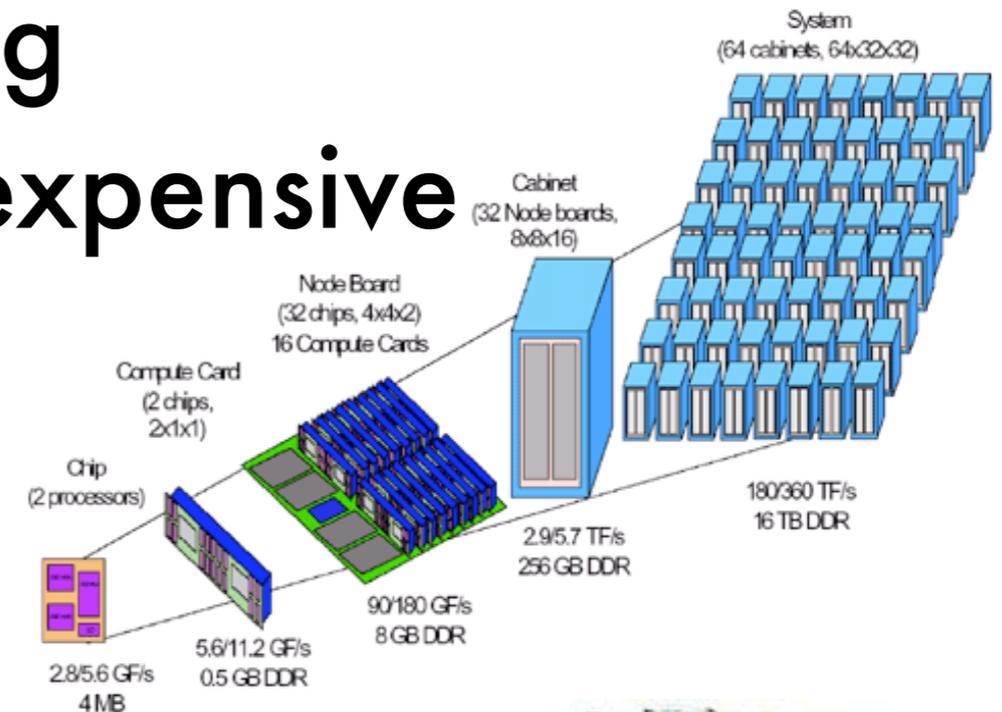
- Essentially infinite amount of data
  - Labeling is prohibitively expensive
  - Not scalable for  $i18n$
  - Even for *supervised* problems unlabeled data abounds. Use it.
  - User-understandable structure for representation purposes
  - Solutions are often customized to problem
- We can only cover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

# Hardware

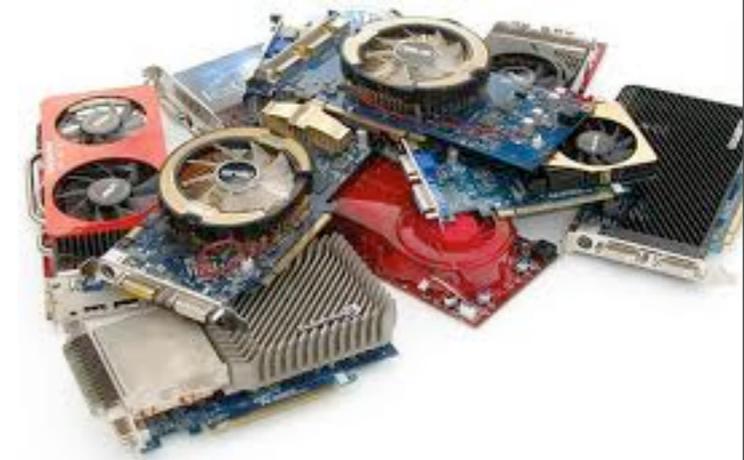


# Commodity Hardware

- High Performance Computing  
Very reliable, custom built, expensive



- Consumer hardware  
Cheap, efficient, easy to replicate,  
not very reliable, deal with it!



## Typical first year for a new cluster:

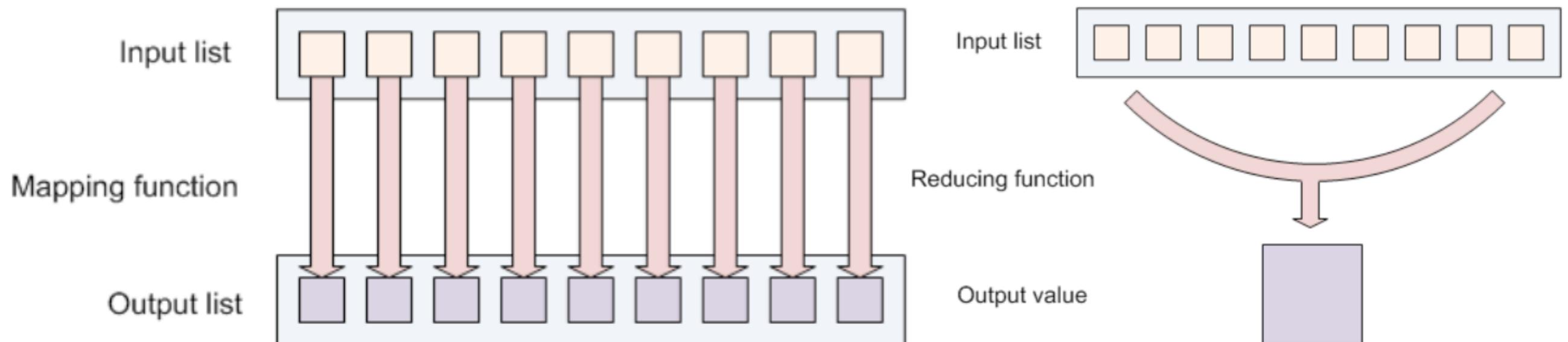
- ~0.5 **overheating** (power down most machines in <5 mins, ~1-2 days to recover)
- ~1 **PDU failure** (~500-1000 machines suddenly disappear, ~6 hours to come back)
- ~1 **rack-move** (plenty of warning, ~500-1000 machines powered down, ~6 hours)
- ~1 **network rewiring** (rolling ~5% of machines down over 2-day span)
- ~20 **rack failures** (40-80 machines instantly disappear, 1-6 hours to get back)
- ~5 **racks go wonky** (40-80 machines see 50% packetloss)
- ~8 **network maintenances** (4 might cause ~30-minute random connectivity losses)
- ~12 **router reloads** (takes out DNS and external vips for a couple minutes)
- ~3 **router failures** (have to immediately pull traffic for an hour)
- ~dozens of minor **30-second blips for dns**
- ~1000 **individual machine failures**
- ~thousands of **hard drive failures**

slow disks, bad memory, misconfigured machines, flaky machines, etc.

- Data (lower bounds)
  - 10 Billion documents (webpages, e-mails, ads, tweets)
  - 100 Million users on Google, Facebook, Twitter, Yahoo, Hotmail
  - 1 Million days of video on YouTube
  - 10 Billion images on Facebook
- Processing capability for single machine 1TB/hour  
**But we have much more data**
- Parameter space for models is too big for a single machine  
**Personalize content for many millions of users**
- Process on **many cores** and **many machines simultaneously**

# Map Reduce

- 1000s of (faulty) machines
- Lots of jobs are mostly embarrassingly parallel (except for a sorting/transpose phase)
- Functional programming origins
  - `Map(key,value)`  
processes each (key,value) pair and outputs a new (key,value) pair
  - `Reduce(key,value)`  
reduces all instances with same key to aggregate



from Ramakrishnan, Sakrejda, Canon, DoE 2011

# Map Reduce

- 1000s of (faulty) machines
- Lots of jobs are mostly embarrassingly parallel (except for a sorting/transpose phase)
- Functional programming origins
  - Map(key,value)  
processes each (key,value) pair and outputs a new (key,value) pair
  - Reduce(key,value)  
reduces all instances with same key to aggregate
- Example - **extremely naive** wordcount
  - Map(docID, document)  
for each document emit many (wordID, count) pairs
  - Reduce(wordID, count)  
sum over all counts for given wordID and emit (wordID, aggregate)

# Map Reduce

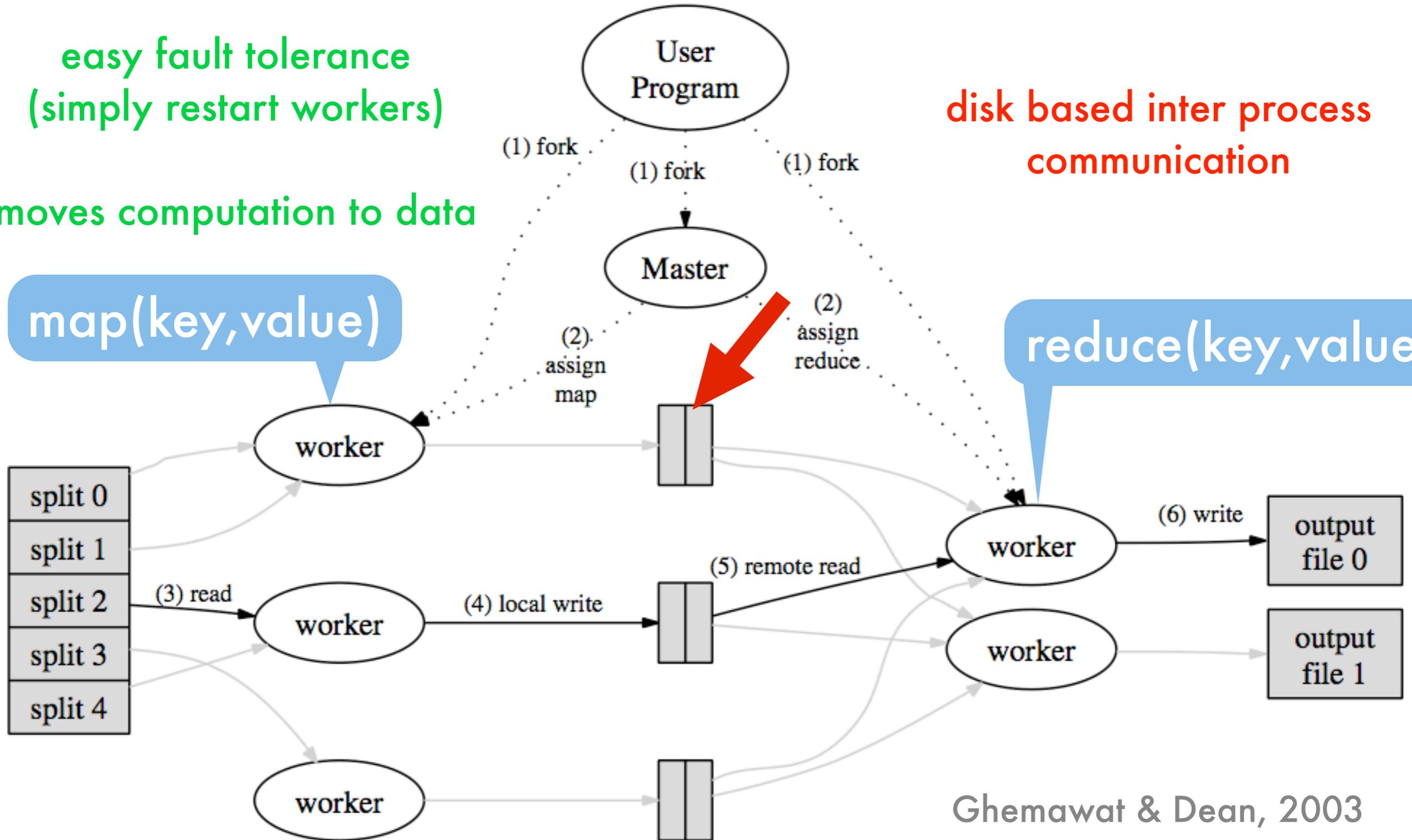
easy fault tolerance  
(simply restart workers)

moves computation to data

disk based inter process  
communication

map(key,value)

reduce(key,value)



Ghemawat & Dean, 2003

# Map Combine Reduce

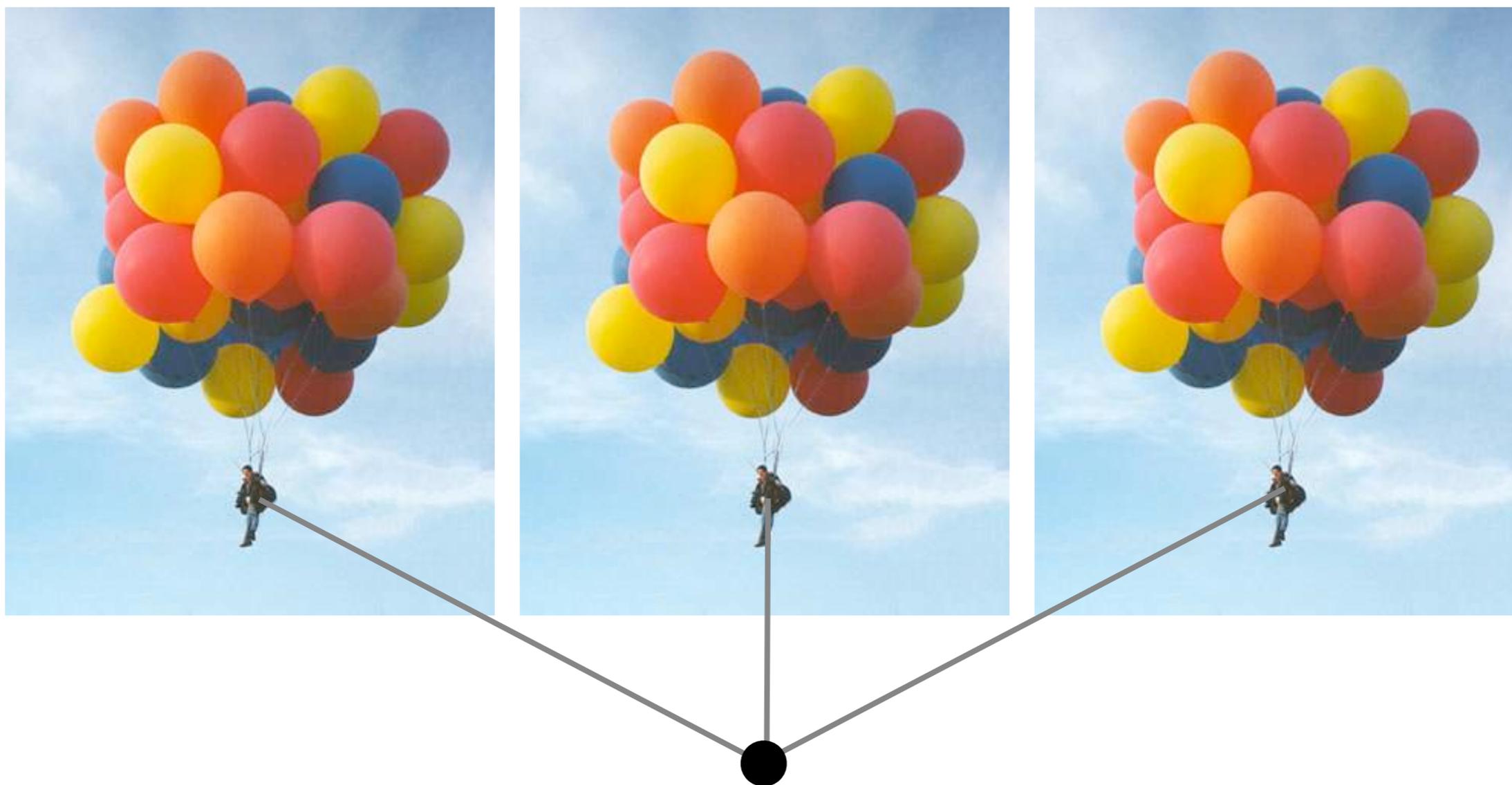
- Combine aggregates keys before sending to the reducer (saves bandwidth)
- Map must be stateless in blocks
- Reduce must be commutative in data
- Fault tolerance
  - Start jobs where the data is (move code not data - nodes run the file system, too)
  - Restart machines if maps fail (have replicas)
  - Restart reducers based on intermediate data
- Good fit for many algorithms
- Good if only a small number of MapReduce iterations needed
- Need to request machines at each iteration (time consuming)
- State lost in between maps
- Communication only via file I/O



# Motivation - Topic models



# Clustering



# Clustering

# Clustering

The screenshot shows the United Airlines website interface. At the top, there's the United logo and navigation links like 'My profile', 'Worldwide sites', and 'Customer service'. Below that are menu items for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. A search bar is also present. The main content area features a 'BOOK FLIGHT' section with fields for 'From' and 'To' airports, departure and return dates, and search options like 'Roundtrip', 'One-way', and 'Multicity'. A prominent banner offers 'Use 30% fewer miles on your next United flight.' with an image of a large orange percentage sign. To the right, there's a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password', and a 'Log in' button. Below the login section, there are links for 'Start earning miles today', 'united.com benefits and features', and 'Travel information'. At the bottom of the main content area, there's a 'United news and deals' section with various links and a 'United-Continental merger' announcement.

The screenshot shows the Australian National University (ANU) website. At the top, there's a 'Change Location' button and a search bar. Below that are navigation links for 'You Fly', 'Loyalty Programmes', and 'Promotions'. A secondary navigation bar includes 'myEMAIL', 'IVLE', 'LIBRARY', 'MAPS', 'CALENDAR', 'SITEMAP', 'CONTACT', and 'e-CARDS'. A search bar with 'search for...' and 'GO' is also present. The main content area features a banner for 'The Australian National University' with a search bar for 'Search ANU...' and navigation links for 'RESEARCH', 'ENTERPRISE', 'CAMPUS LIFE', 'GIVING', and 'CAREERS@NUS'. Below the banner, there's a navigation bar with 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The main content area is partially obscured by a large image of a tree trunk.

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

Copyright © 2006 Chijmes. All rights reserved.

Forests renew after Black Saturday fires

School of Music at Floriade

Undergraduate studies

Higher Degree Research

# Clustering

The screenshot shows the United Airlines website interface. At the top, there are navigation links for 'My profile', 'Worldwide sites', and 'Customer service'. Below this is a search bar and a menu with categories like 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. The main content area is divided into several sections: 'BOOK FLIGHT' and 'REDEEM MILES' on the left; a large promotional banner for 'Use 30% fewer miles on your next United flight.' in the center; and a 'Log in' section on the right. Below the login section, there are links for 'More You Fly', 'Loyalty Programmes', and 'Promotions'. At the bottom, there is a 'Need Help?' section with links to 'SIA Holidays' and 'Hotel Bookings', and a 'Book Now' button.

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

The screenshot shows the website for Chez Patisse, a restaurant and café. The top navigation bar includes links for 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. The main content area is divided into several sections: 'RESERVATIONS RESTAURANT & CAFÉ', 'MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST', 'ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION', 'SPECIAL EVENTS CALENDAR', 'STORE BOOKS • POSTERS • GIFTS', and 'CONTACT INFORMATION DIRECTIONS • MAILING LIST'. The background of the website features a photograph of the restaurant's interior, showing a bar area with a sign that reads 'BAR DE LA PEPECHE'. At the bottom, there is a 'Feedback | Terms & Conditions' link.

# Clustering



The image shows a screenshot of the United Airlines website. A red speech bubble with the word "airline" is overlaid on the page. The website features a navigation menu with options like "Planning & booking", "Reservations & check-in", and "Mileage Plus". There are sections for "BOOK FLIGHT" and "REDEEM MILES", a "Log in" section, and a "United news and deals" section. A prominent offer states "Use 30% fewer miles on your next United flight." with a large percentage sign graphic.



The image shows a screenshot of the Australian National University (ANU) website. A red speech bubble with the word "university" is overlaid on the page. The website features a navigation menu with options like "HOME", "FUTURE STUDENTS", "CURRENT STUDENTS", "RESEARCH & EDUCATION", "ABOUT ANU", and "STAFF". There is a main content area with a featured article titled "Ash forests rise and rise again" and a "Higher Degree Research" section. The ANU logo and name are prominently displayed at the top.



The image shows a screenshot of the Chez Pannise restaurant website. The page is a navigation menu with the following items: "RESERVATIONS RESTAURANT & CAFÉ", "MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST", "ABOUT CHEZ PANISSE • ALICE WATERS OUR CHEFS • FRIENDS • PRESS FOUNDATION & MISSION", "SPECIAL EVENTS CALENDAR", "STORE BOOKS • POSTERS • GIFTS", and "CONTACT INFORMATION DIRECTIONS • MAILING LIST". The restaurant's name "Chez Pannise" is written in a cursive font at the top.



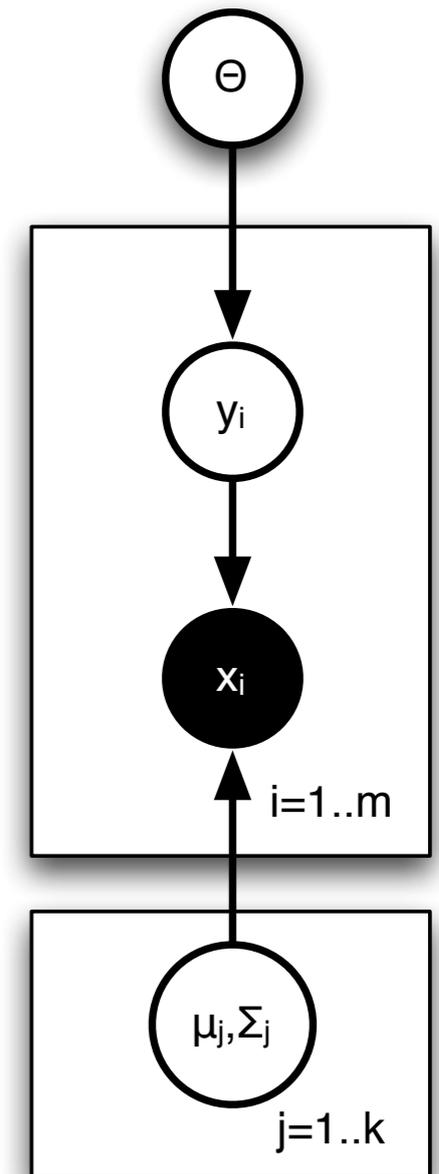
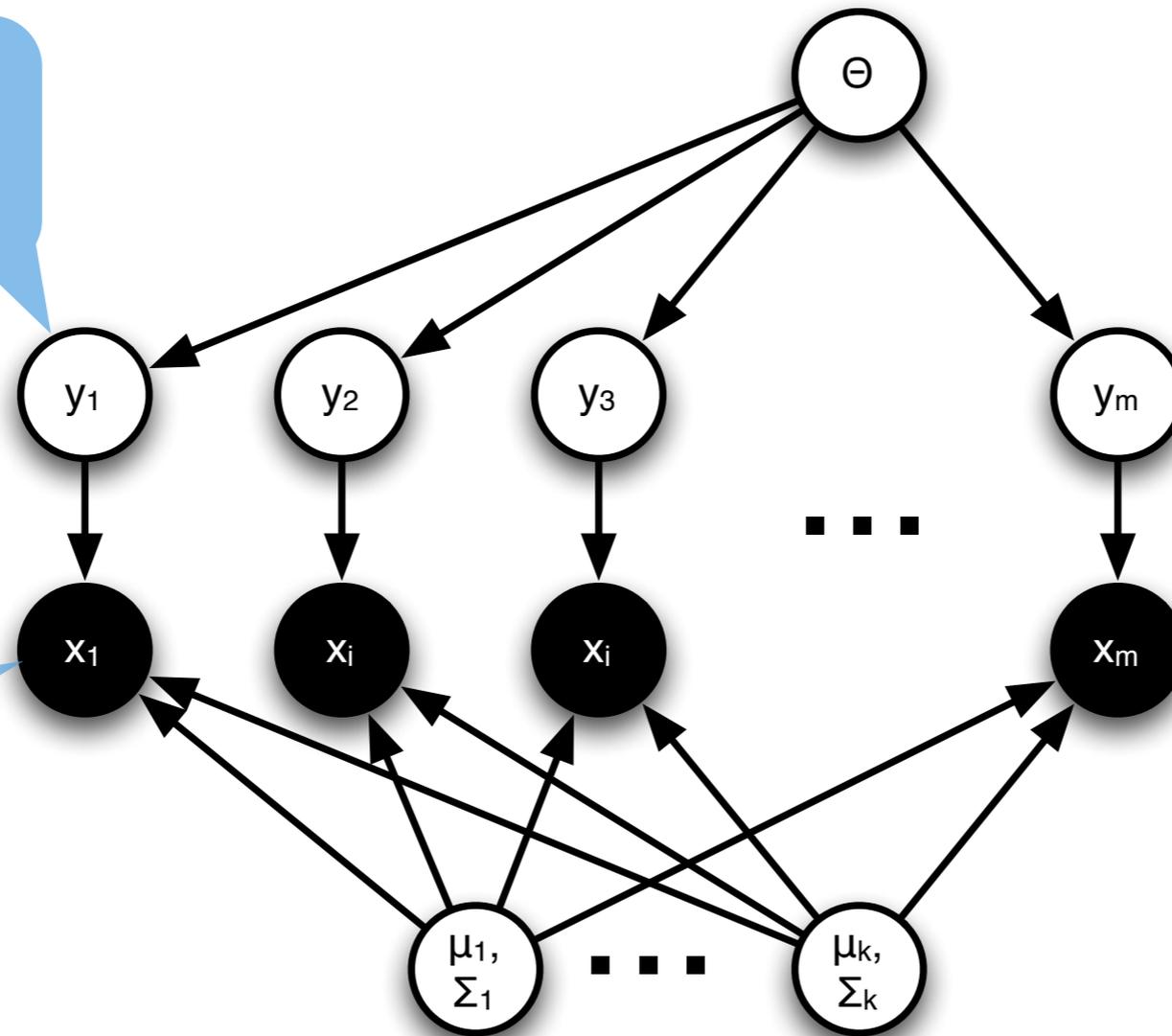
The image shows a screenshot of the Suntec REIT website. The page features a navigation menu with options like "Home", "Wining & Dining", "Contact", "Sitemap", and "About Suntec REIT". The main content area shows a large image of a church building at night, illuminated with lights. The website has a dark red header and footer.

restaurant

# Generative Model

cluster ID

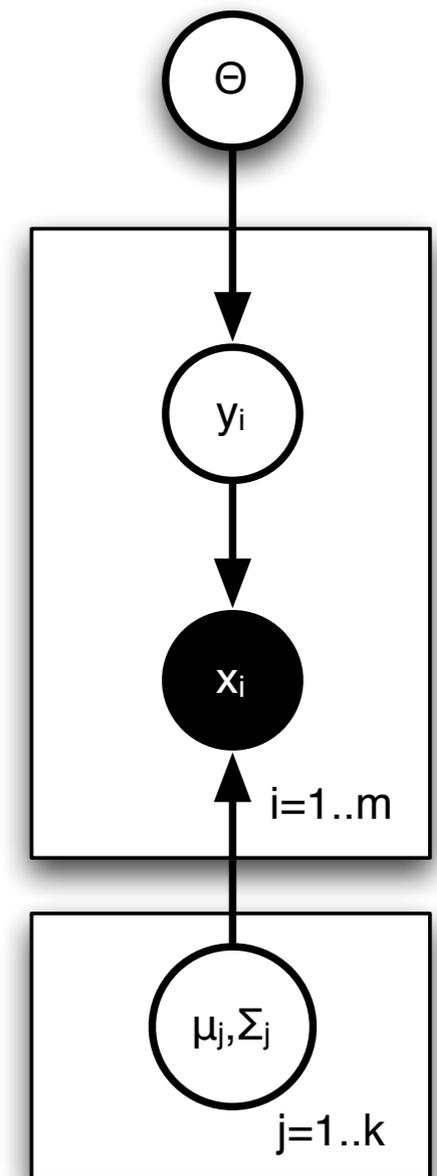
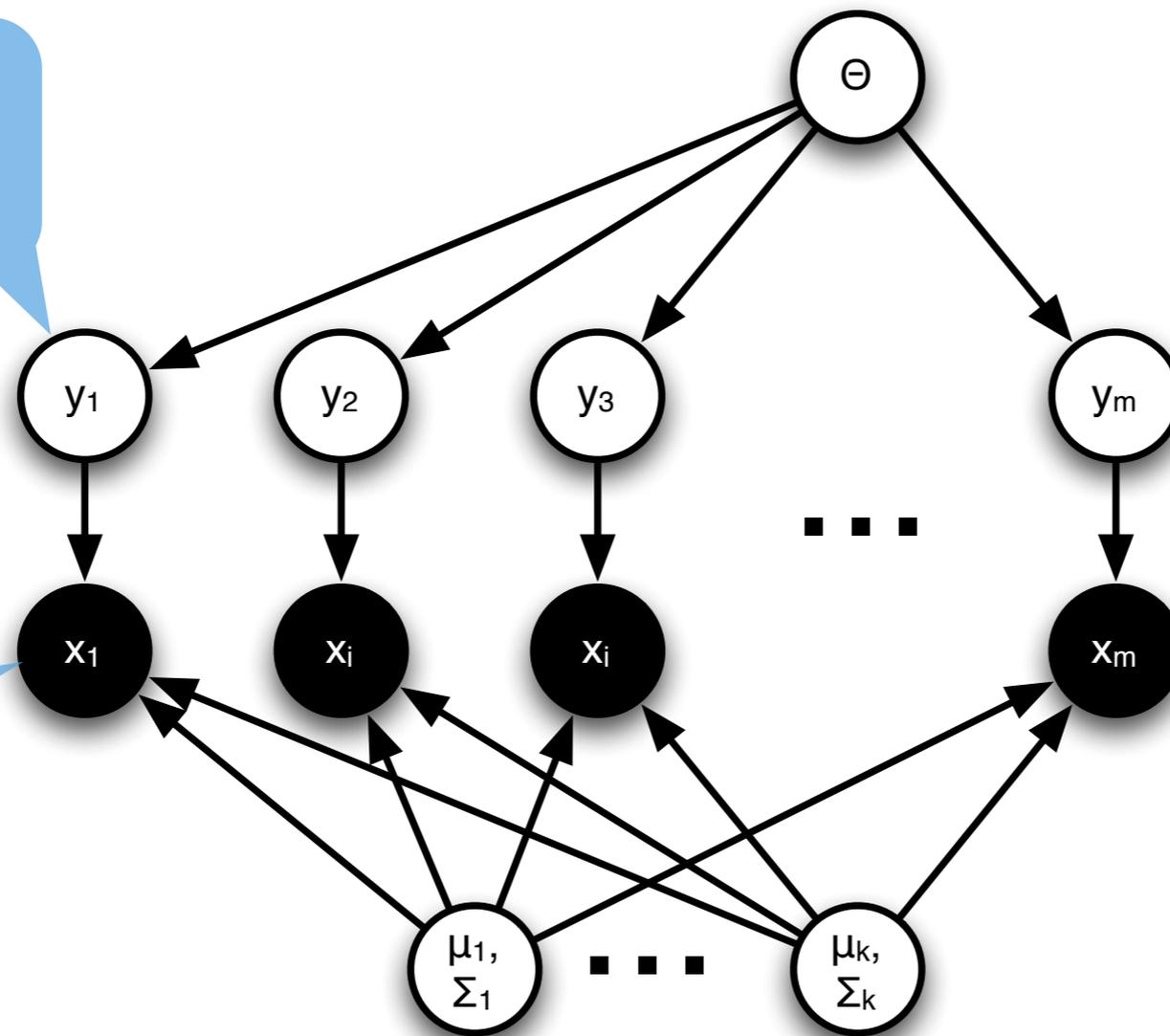
objects



# Generative Model

cluster ID

objects



$$p(X, Y | \theta, \sigma, \mu) = \prod_{i=1}^n p(x_i | y_i, \sigma, \mu) p(y_i | \theta)$$

# What can we cluster?

# What can we cluster?

mails      text      urls      products

news      queries      users

spammers      ads      locations

abuse      events

# Grouping objects

# Grouping objects

The image shows a composite of two website screenshots. The top screenshot is from Singapore Airlines, featuring the logo, navigation links like 'The Experience', 'Flights & Fares', and a search bar. The bottom screenshot is from the National University of Singapore (NUS), showing the university logo, navigation menus, and several content tiles. A red speech bubble with the word 'Singapore' is overlaid on the NUS page, pointing to the word 'Singapore' in the 'Asia' tile. The 'Asia' tile includes a photo of a couple and a link to 'CLICK HERE TO FIND OUT MORE'. Other tiles include 'Flame Arrival Ceremony at NUS' with a 'WATCH THE VIDEO' button, and 'Joint Evacuation Exercises' with a warning icon and event details. At the bottom of the NUS page are buttons for 'STAFF', 'ALUMNI', and 'VISITORS'. The footer of the NUS page contains copyright information for Chijmes (© 2006) and links for 'Feedback | Terms & Conditions'.

# Grouping objects

The screenshot shows the United Airlines website interface. At the top left is the United logo. Navigation links include "My profile", "Worldwide sites", and "Customer service". A main menu contains "Planning & booking", "Reservations & check-in", "Mileage Plus", and "Services & information". A search bar is present. Below the menu, there are tabs for "Flights", "Check-in", and "Flight status". The "BOOK FLIGHT" section includes fields for "From" and "To" airports, departure and return dates, and search options like "Roundtrip", "One-way", and "Multicity". A large promotional banner in the center features a large orange percentage sign and the text "Use 30% fewer miles on your next United flight." To the right, there is a "Log in" section for Mileage Plus members, with fields for "Mileage Plus # or email address" and "Password". Below this, there are links for "Start with My Mileage Plus" and "My reservations". The bottom of the page includes a footer with "About United", "Investor relations", "Business resources", "Careers", and "Site map".

The screenshot shows the Australian National University (ANU) website. At the top, there is a navigation menu with links for "CALENDAR", "SITEMAP", "CONTACT", and "e-CARDS". A search bar is located below the menu. The main banner features a photograph of a bar interior with the text "The Australian National University" in a large, blue font. Below the banner, there is a secondary navigation menu with links for "CURRENT STUDENTS", "RESEARCH & EDUCATION", "ABOUT ANU", and "STAFF". The bottom of the page contains a row of four featured articles: "Forests renew after Black Saturday fires", "School of Music at Floriade", "Undergraduate studies", and "Higher Degree Research".

The screenshot shows the footer of a website. It includes the text "Owned by:", "Managed by:", and "Property Manager:". Below this, there are logos for "SUNTEC Real Estate Investment Trust", "ARA", and "APC APAC Investment Management Pte Ltd". At the bottom, there is a copyright notice: "Copyright © 2006 Chijmes. All rights reserved."

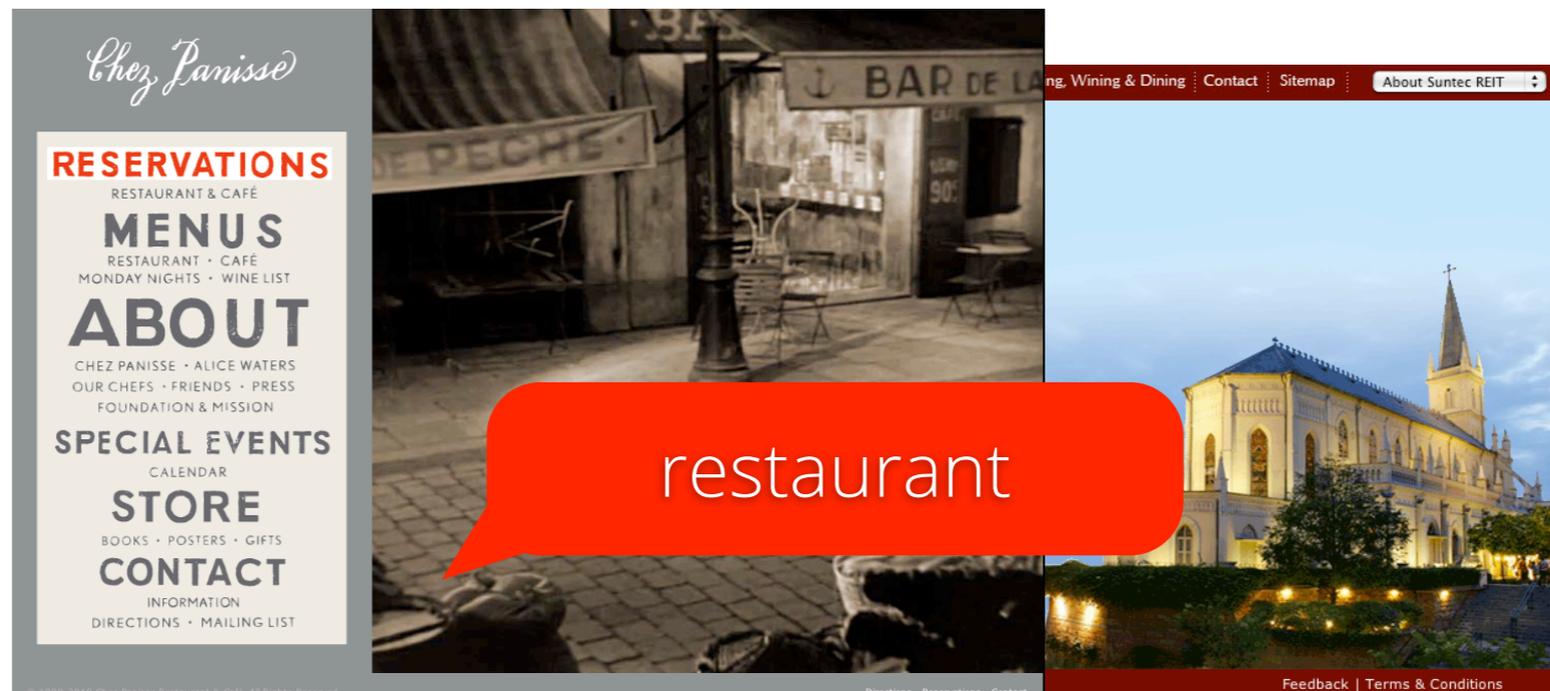
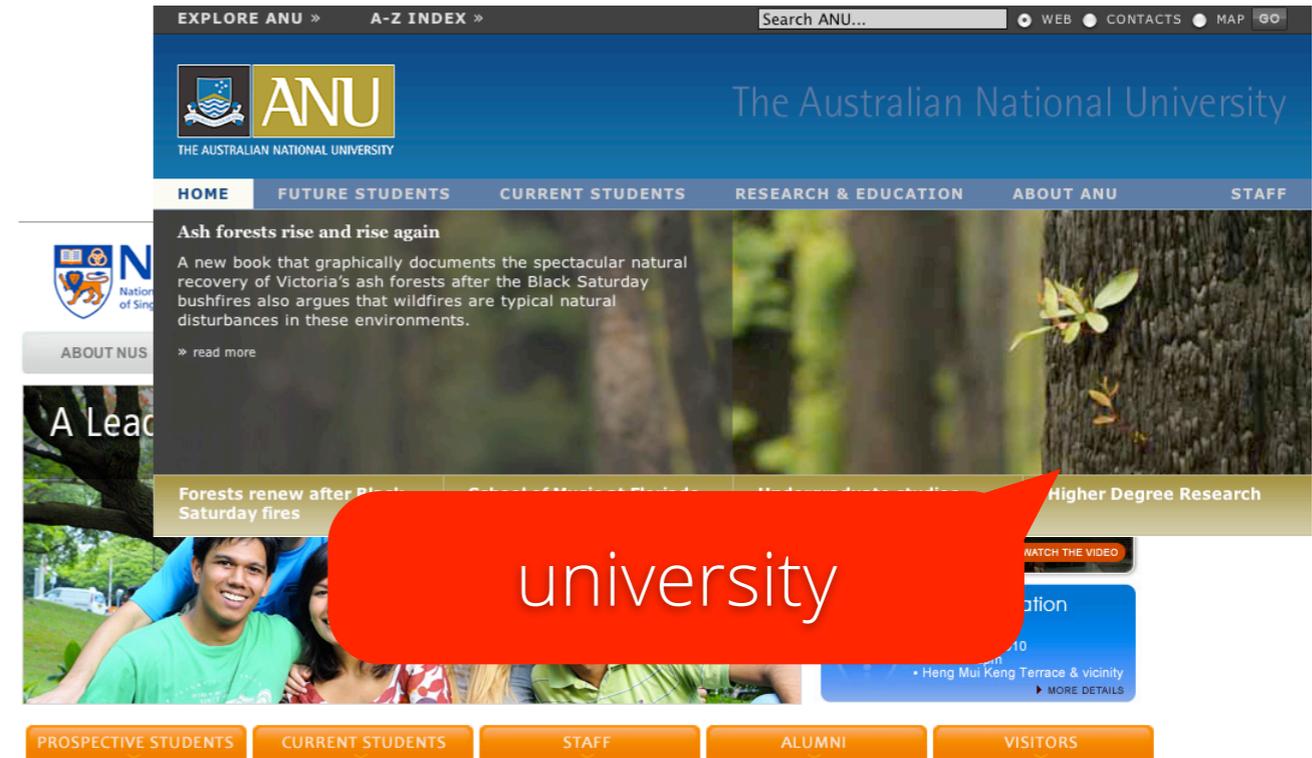
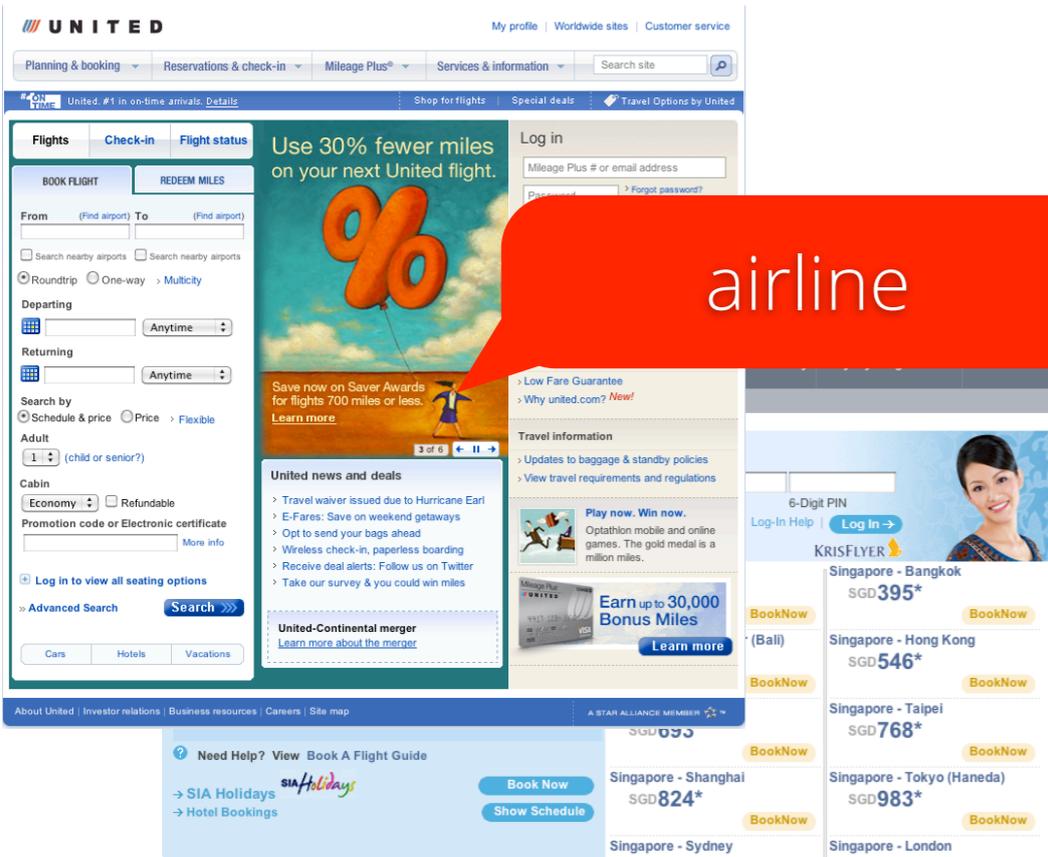
# Grouping objects

The screenshot shows the United.com website interface. It features a top navigation bar with 'UNITED' and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below this is a secondary navigation bar with categories like 'Planning & booking', 'Reservations & check-in', and 'Mileage Plus'. The main content area is divided into several sections: a 'BOOK FLIGHT' section with search filters for 'From', 'To', 'Departing', and 'Returning'; a 'REDEEM MILES' section; a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'; and a 'United news and deals' section with various promotional banners. A sidebar on the right contains a 'Change Location' search box and a 'KrisFlyer' section with flight routes and prices.

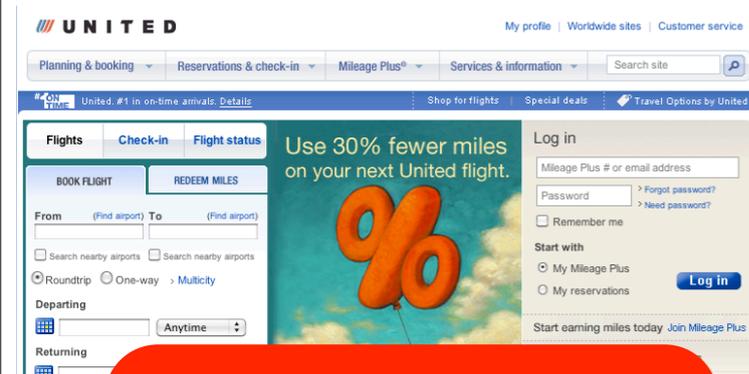
The screenshot shows the Australian National University (ANU) website. The top navigation bar includes 'EXPLORE ANU', 'A-Z INDEX', and a search bar. The main header features the ANU logo and the text 'The Australian National University'. Below the header is a secondary navigation bar with links for 'HOME',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

The screenshot shows the Chez Panisse website. The top navigation bar includes 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. The main content area features a navigation menu with links for 'RESERVATIONS', 'MENUS', 'ABOUT', 'SPECIAL EVENTS', 'STORE', and 'CONTACT'. The background image is a photograph of a restaurant interior with a sign that says 'BAR DE LA'. Below the navigation menu is a photograph of a church at night.

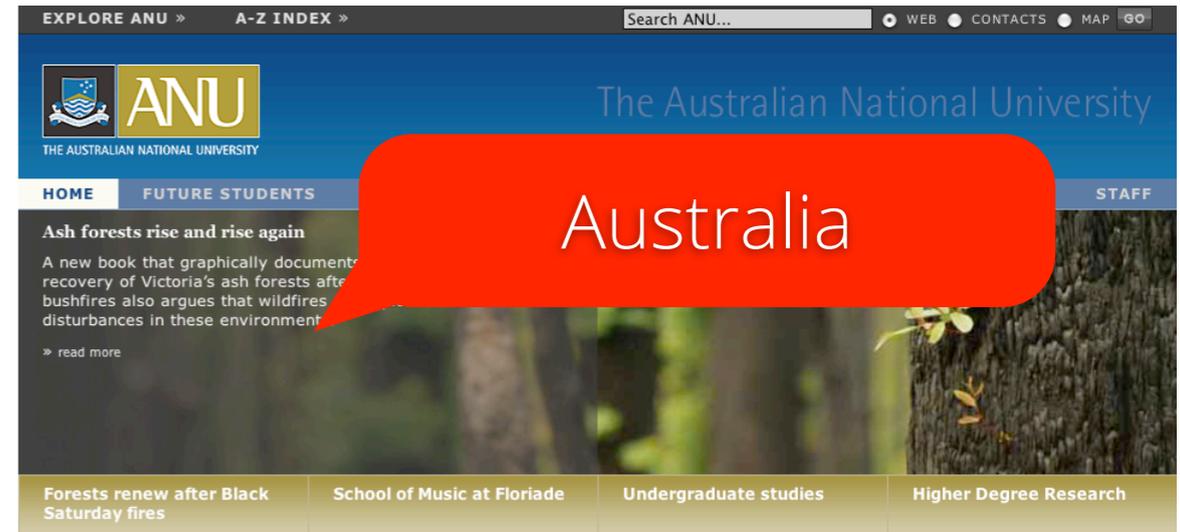
# Grouping objects



# Grouping objects



USA



Australia



Singapore

# Topic Models

UNITED  
My profile | Worldwide sites | Customer service  
Planning & booking | Reservations & check-in | Mileage Plus® | Services & information | Search site

Use 30% fewer miles on your next United flight.

BOOK FLIGHT | REDEEM MILES

From (Find airport) To (Find airport)

Departing: Anytime

Returning: Anytime

Log in

Mileage Plus # or email address

Password

Remember me

Start with

My Mileage Plus

My reservations

Log in

Start earning miles today Join Mileage Plus

Learn more

USA  
airline

EXPLORE ANU » A-Z INDEX » Search ANU... WEB CONTACTS M

ANU  
THE AUSTRALIAN NATIONAL UNIVERSITY

HOME FUTURE STUDENTS CUR ABOUT ANU

Ash forests rise and rise again

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

read more

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

Australia  
university

SINGAPORE AIRLINES

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

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

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

Round Trip | One Way | Stopover/Multi-city

From: Depart: Sat

Departure City

To: Return: Sat

Destination City

Must travel on these dates

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

Need Help? View Book A Flight

SIA Holidays | Hotel Bookings

Singapore - Bangkok SGD 395\* BookNow

Singapore - Hong Kong SGD 546\* BookNow

Singapore - Taipei SGD 768\* BookNow

Singapore - Tokyo (Haneda) SGD 983\* BookNow

Singapore - Sydney

Singapore - London

Singapore  
airline

NUS  
National University of Singapore

myEMAIL IVLE LIBRARY MAPS CALENDAR SITEMAP CONTACT CARDS

Search search for... in NUS Websites GO

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

A Leading Global Un

Game Arrival Ceremony

NUS WATCH THE VIDEO

Joint Evacuation Exercises

7 & 14 Sept 2010

10am - 12pm

Heng Mui Keng Terrace & vicinity

MORE DETAILS

PROSPECTIVE STUDENTS CURRENT STUDENTS STAFF ALUMNI VISITORS

Singapore  
university

Chez Panisse

RESERVATIONS  
RESTAURANT & CAFÉ

MENUS  
RESTAURANT • CAFÉ  
MONDAY NIGHTS • WINE LIST

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

SPECIAL EVENTS  
CALENDAR

STORE  
BOOKS • POSTERS • GIFTS

CONTACT  
INFORMATION  
DIRECTIONS • MAILING LIST

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

Directions Reservations Contact

USA  
food

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

IJMES  
restaurants • bars • shops

Discover a century of resplendent living history behind the cloisters

Chijmes, a premier lifestyle destination in Singapore

Owned by: Managed by: Property Manager:

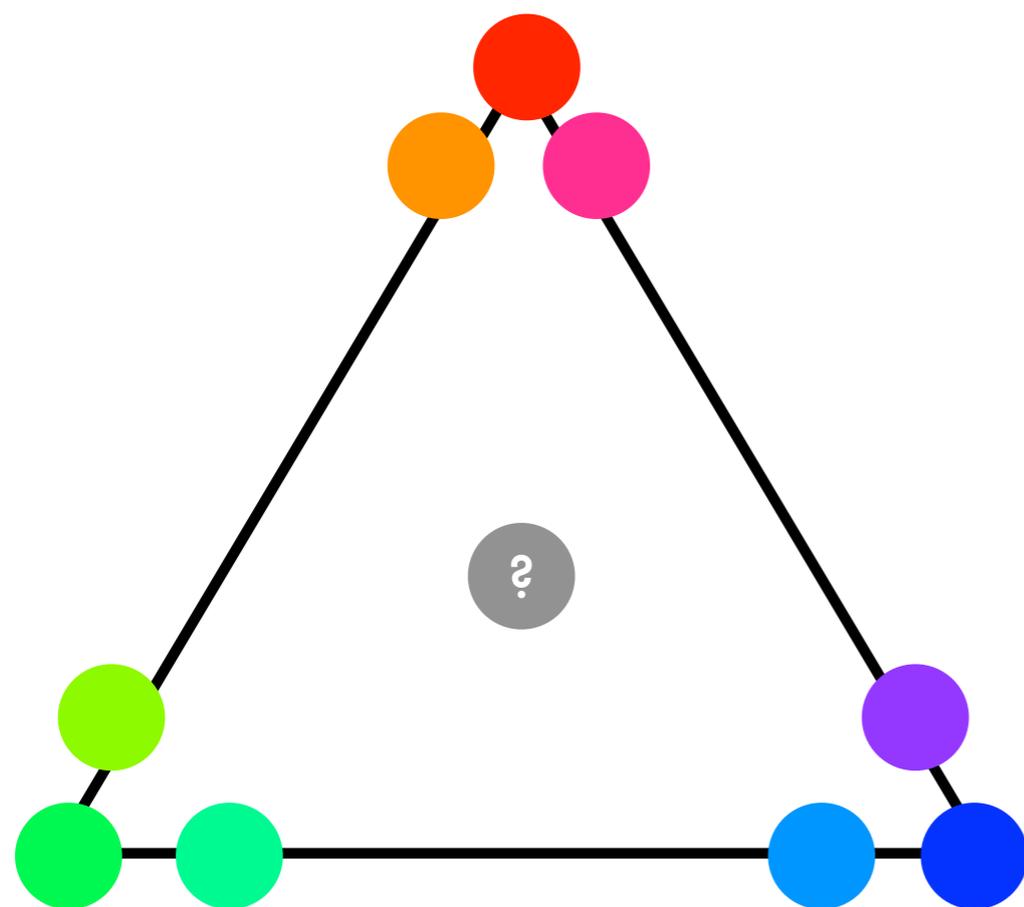
SUNTEC ARA IJMES

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

Singapore  
food

# Clustering & Topic Models

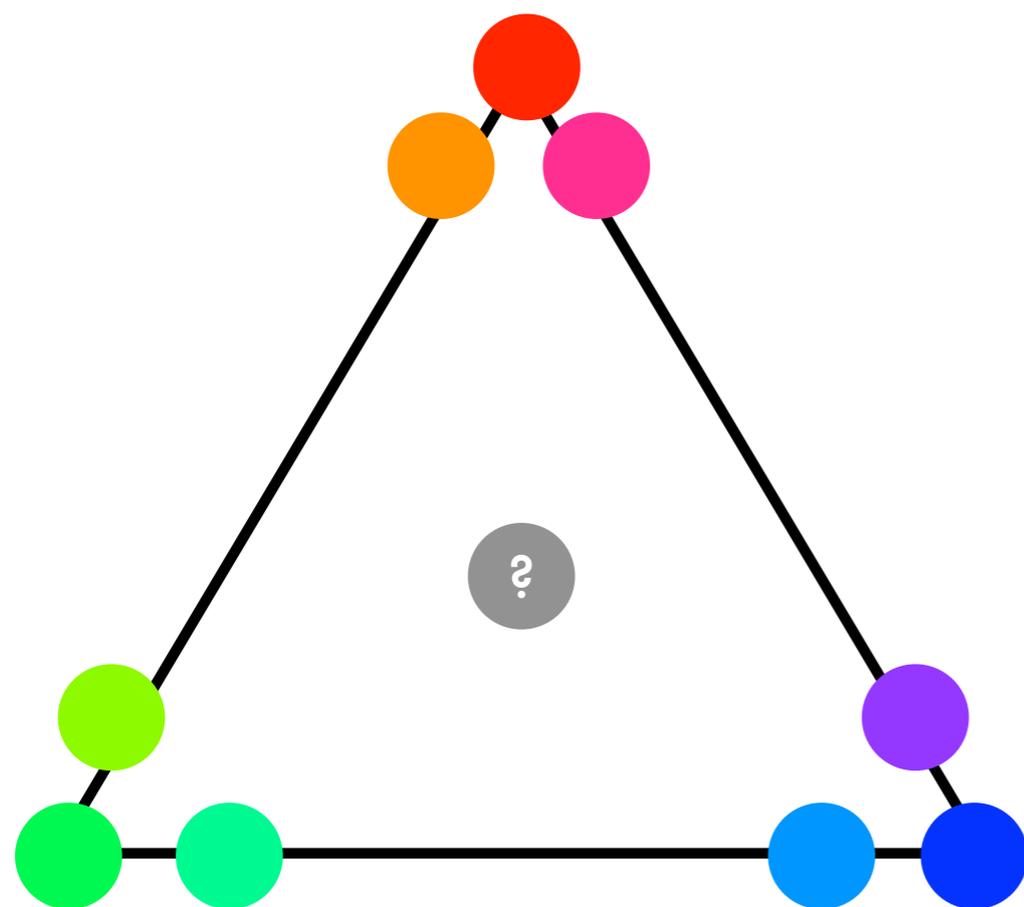
## Clustering



group objects  
by prototypes

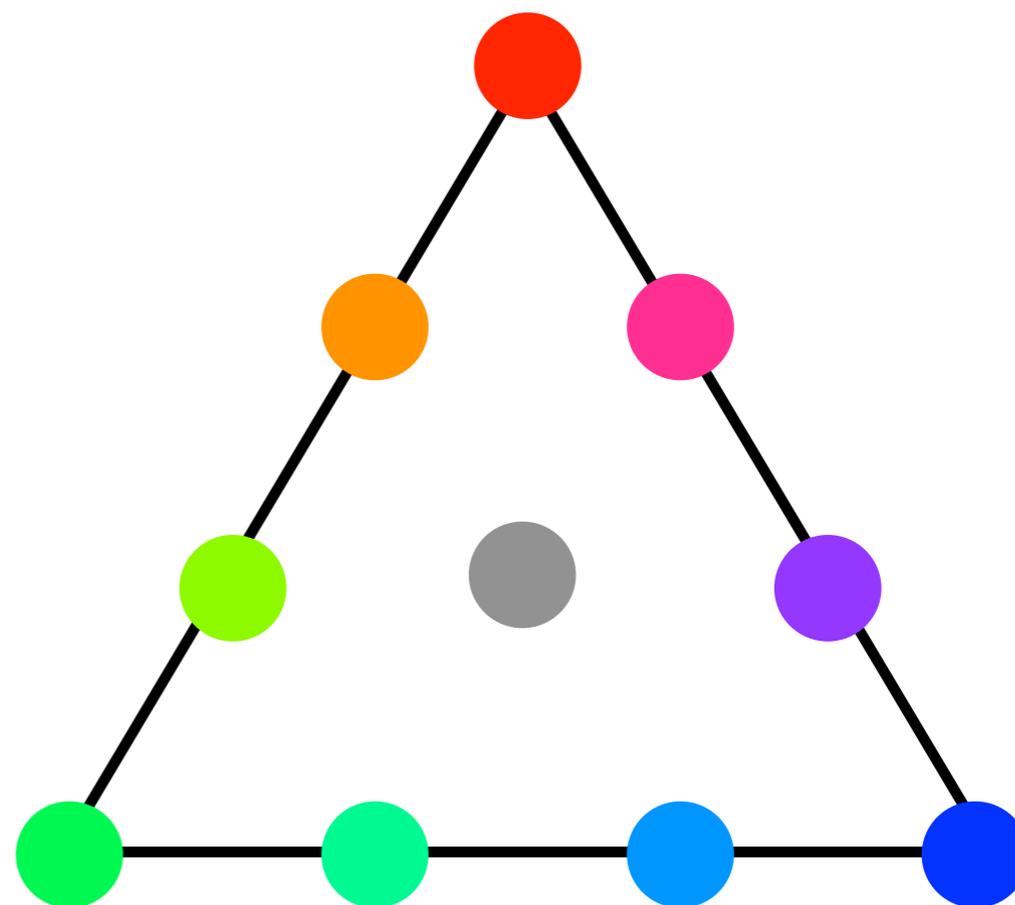
# Clustering & Topic Models

Clustering



group objects  
by prototypes

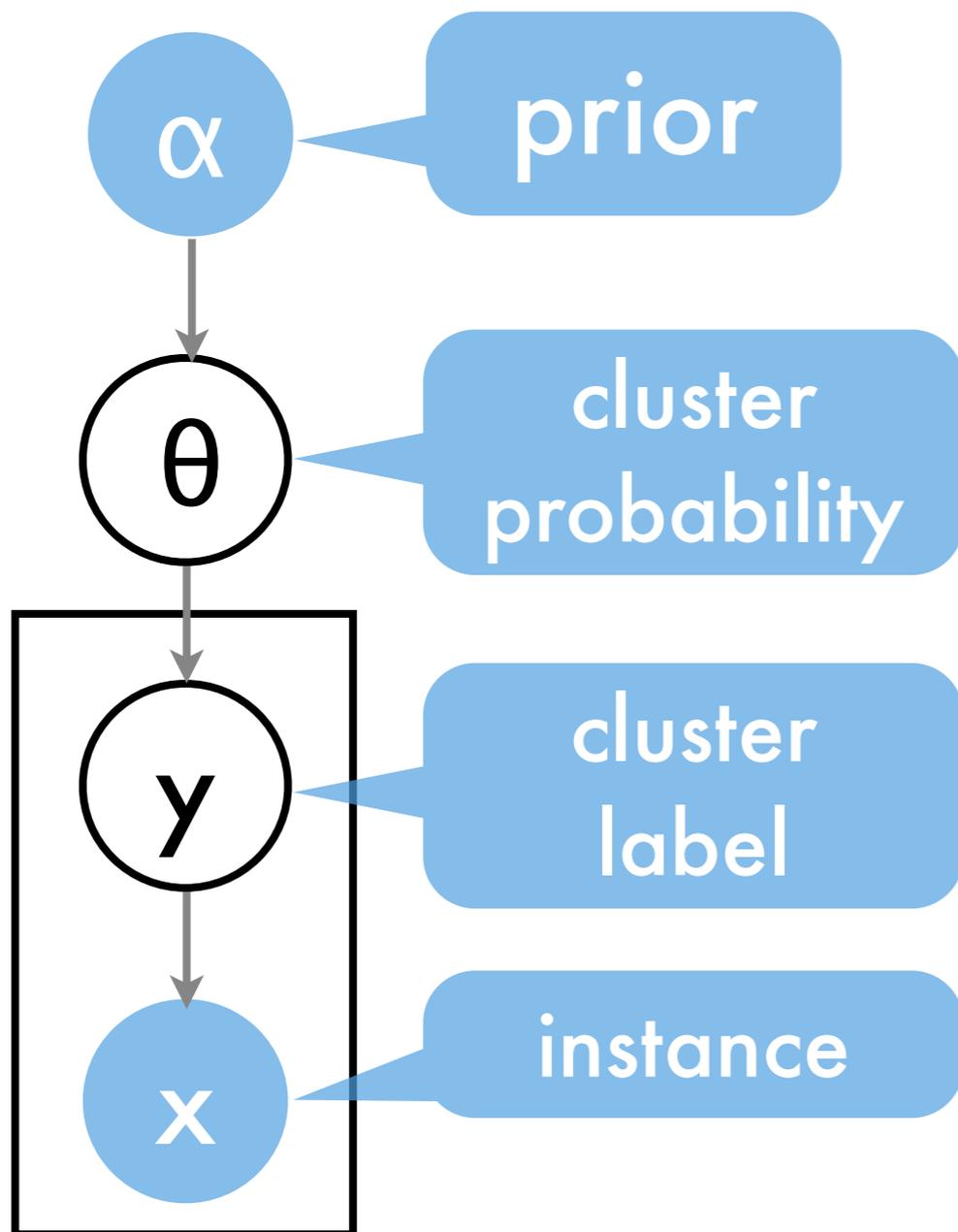
Topics



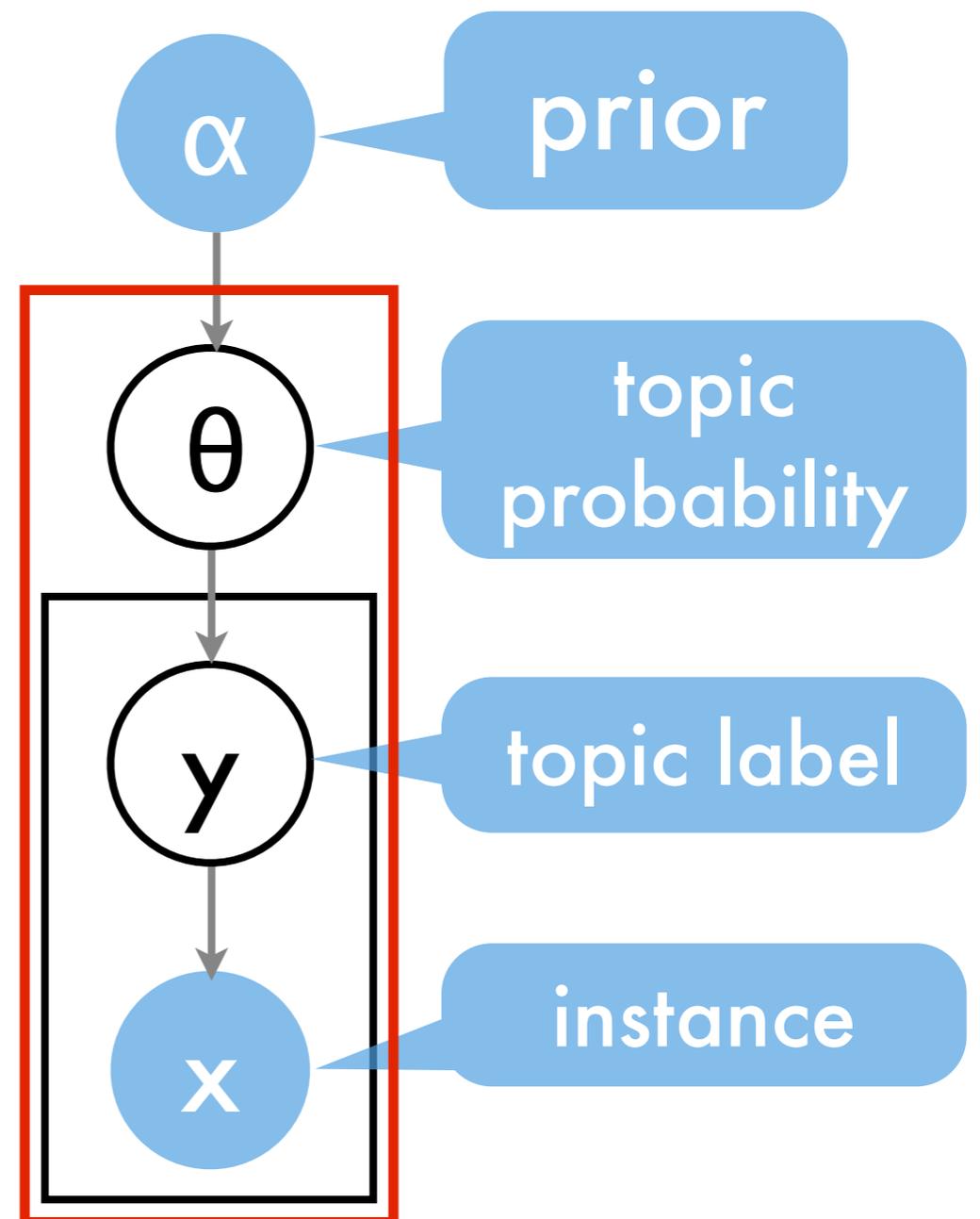
decompose objects  
into prototypes

# Clustering & Topic Models

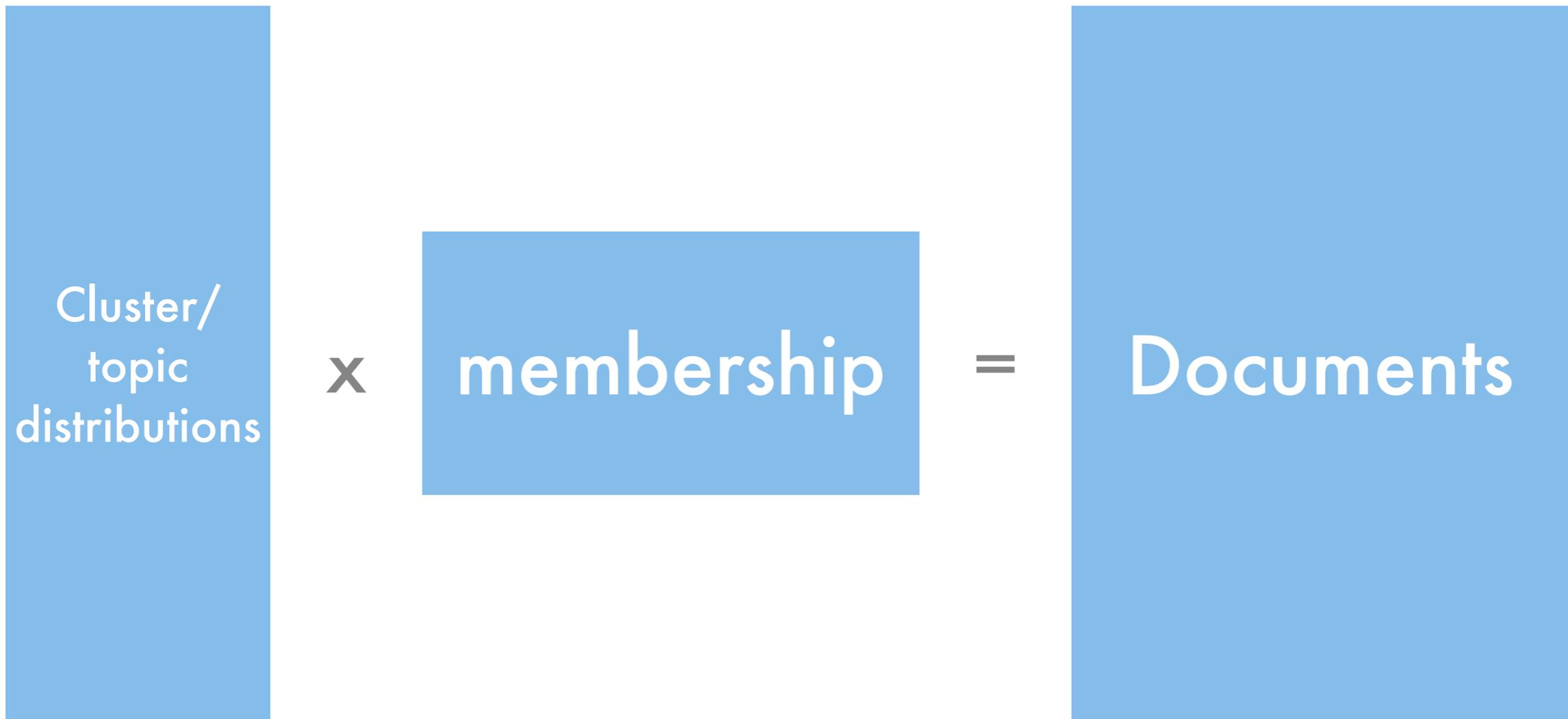
clustering



Latent Dirichlet Allocation



# Clustering & Topic Models



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

# Topics in text

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

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

# Example Topics

## “Arts”

## “Budgets”

## “Children”

## “Education”

---

NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

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

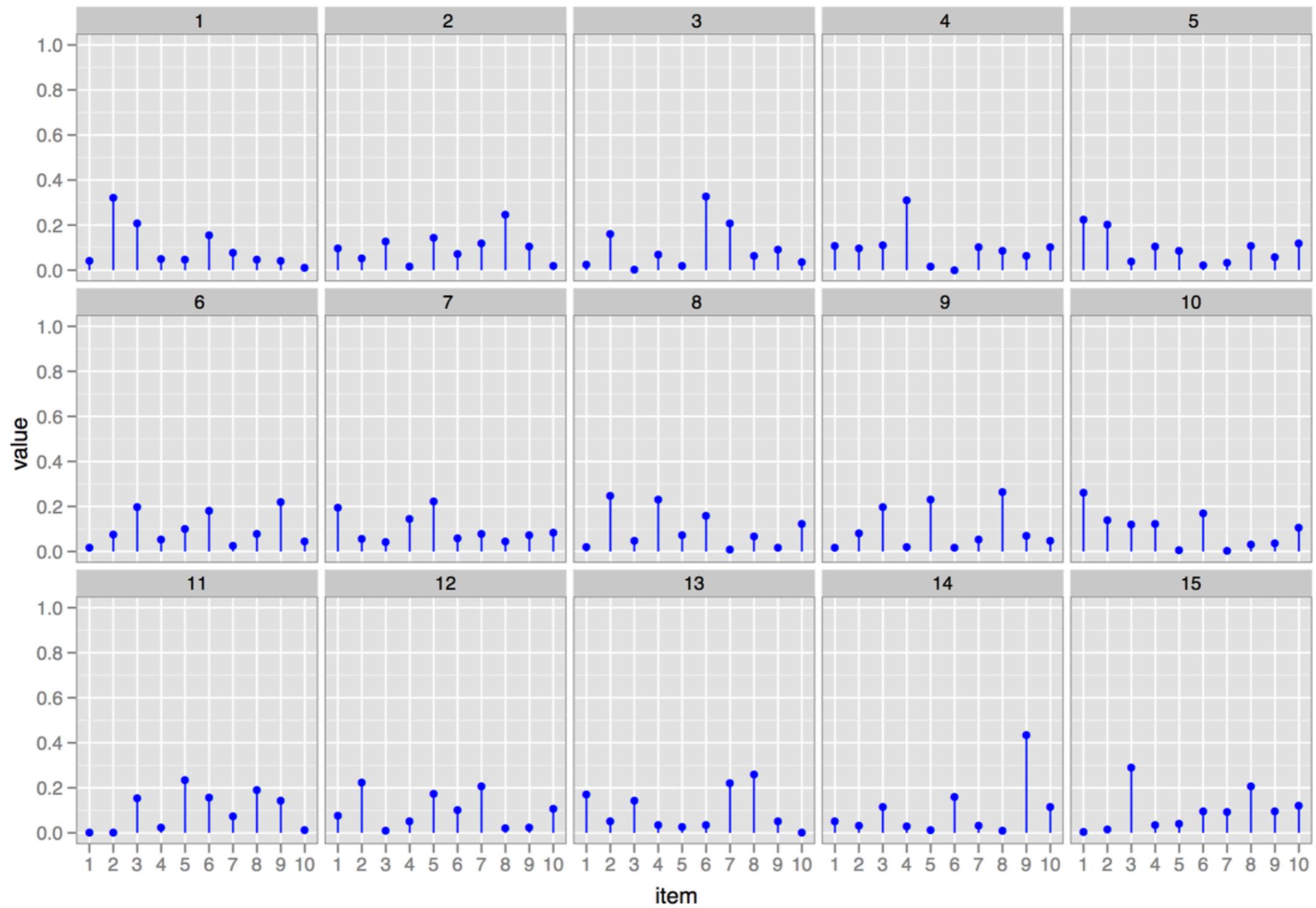
# Dirichlet Distribution

- Is a distribution over the simplex, i.e. positive vectors that sum to 1:

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

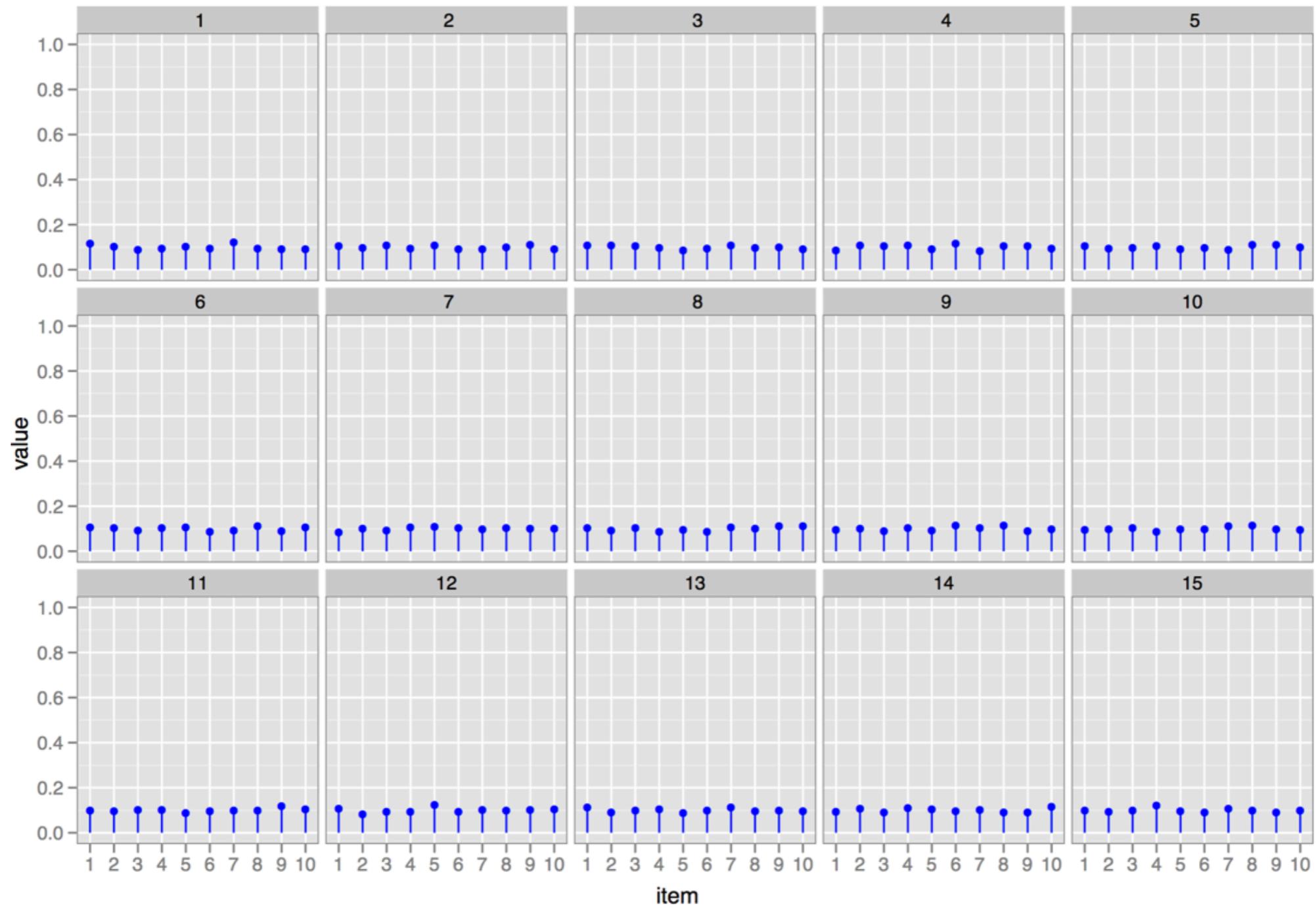
- $\alpha$  controls the shape of the distribution
- Expectations:  $E[\theta_i|\alpha] = \frac{\alpha_i}{\sum_i \alpha_i}$
- Conjugate to the multinomial distribution

$$\alpha = 1$$



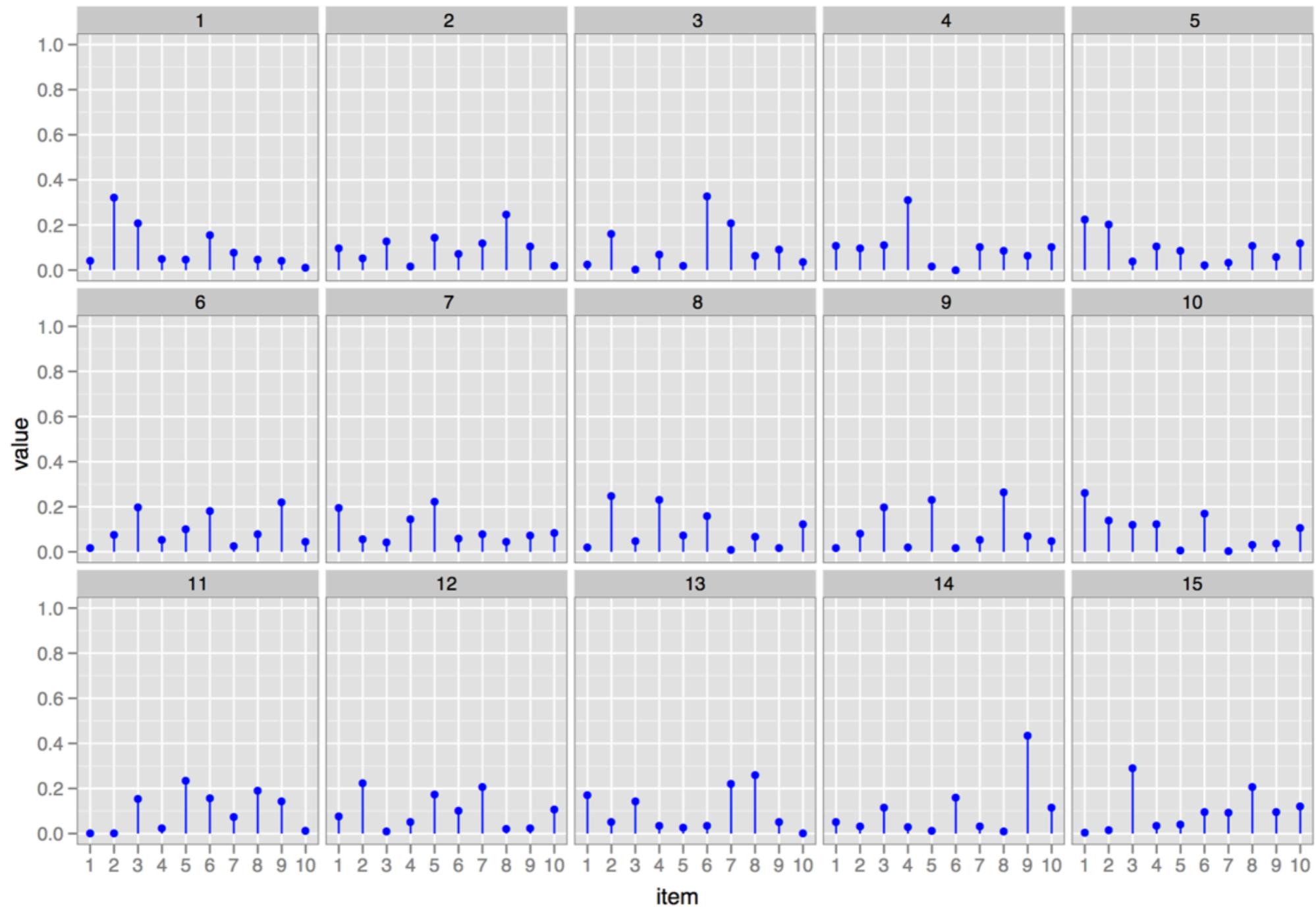
[Blei, LDA tutorial]

$\alpha = 100$



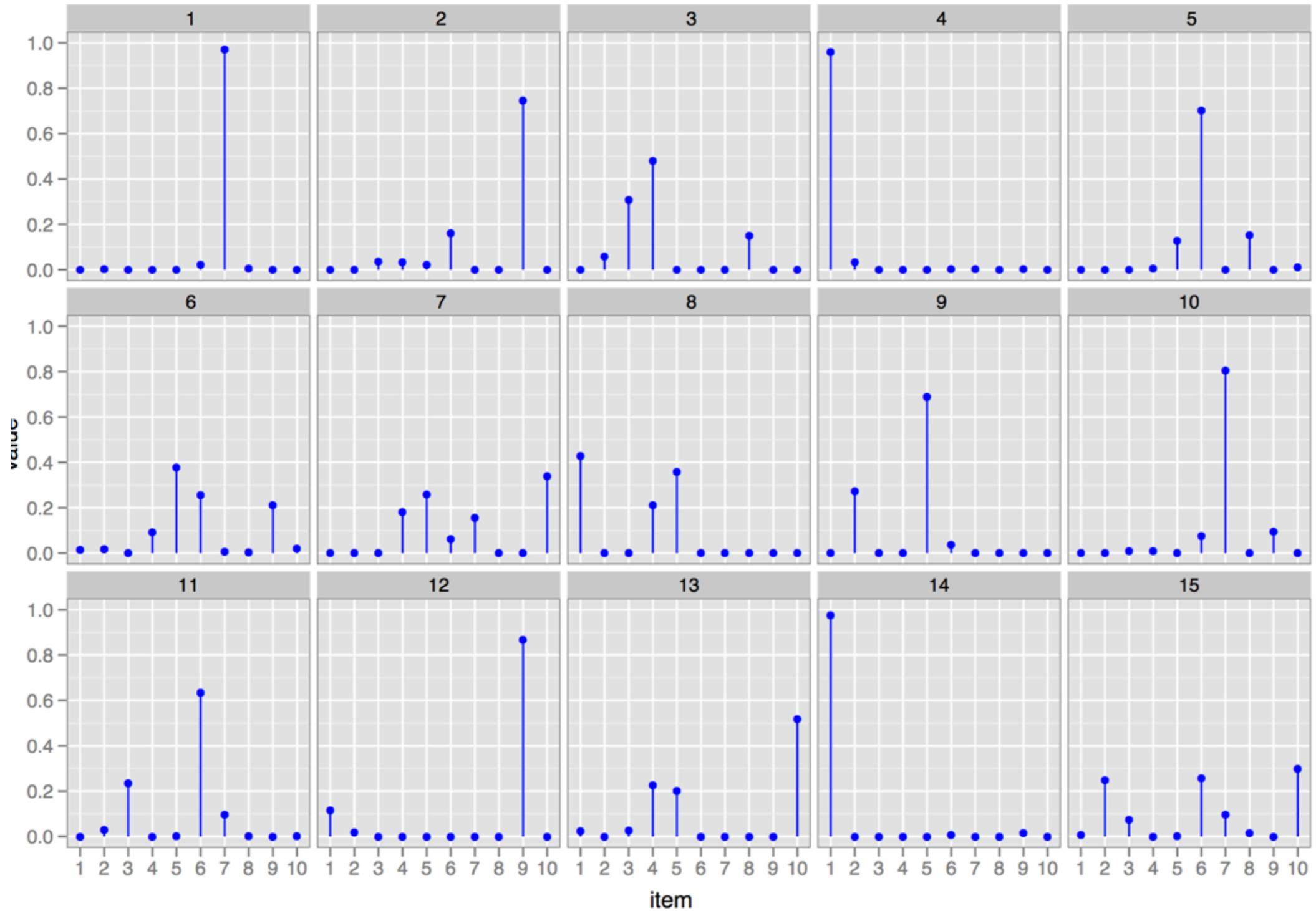
[Blei, LDA tutorial]

$$\alpha = 1$$



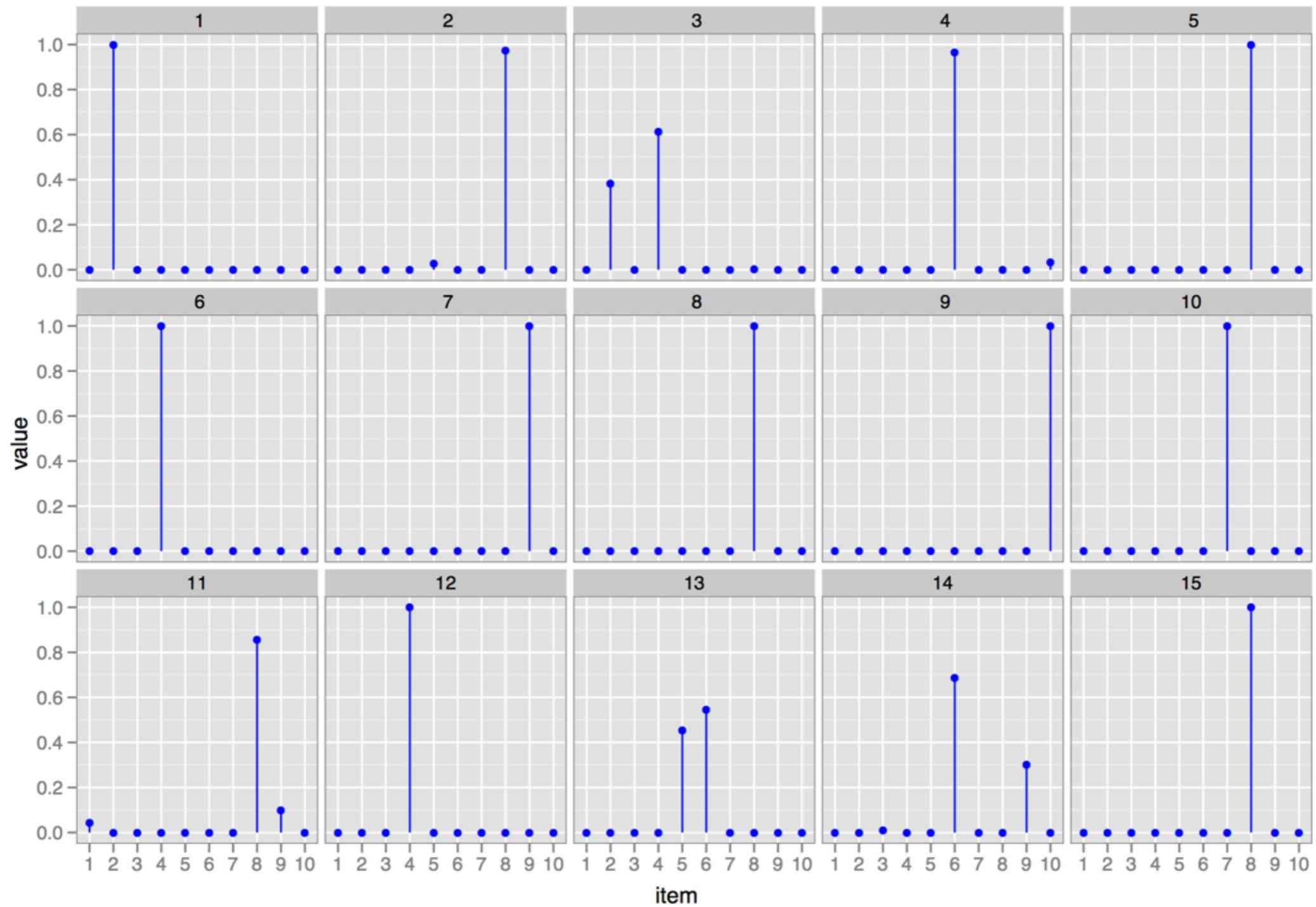
[Blei, LDA tutorial]

$\alpha = .1$



[Blei, LDA tutorial]

$\alpha = .01$



[Blei, LDA tutorial]

# Dirichlet Distribution

- Is a distribution over the simplex, i.e. positive vectors that sum to 1:

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

- Conjugate to the multinomial distribution

$$P(\theta|\alpha, x) = \frac{\Gamma(\sum_i x_i + \alpha_i)}{\prod_i \Gamma(x_i + \alpha_i)} \prod_i \theta_i^{x_i + \alpha_i - 1}$$

# Dirichlet Distribution

- **Prior**

$$P(\theta|\alpha) \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$$

$$E[\theta_i|\alpha] = \frac{\alpha_i}{\sum_i \alpha_i}$$

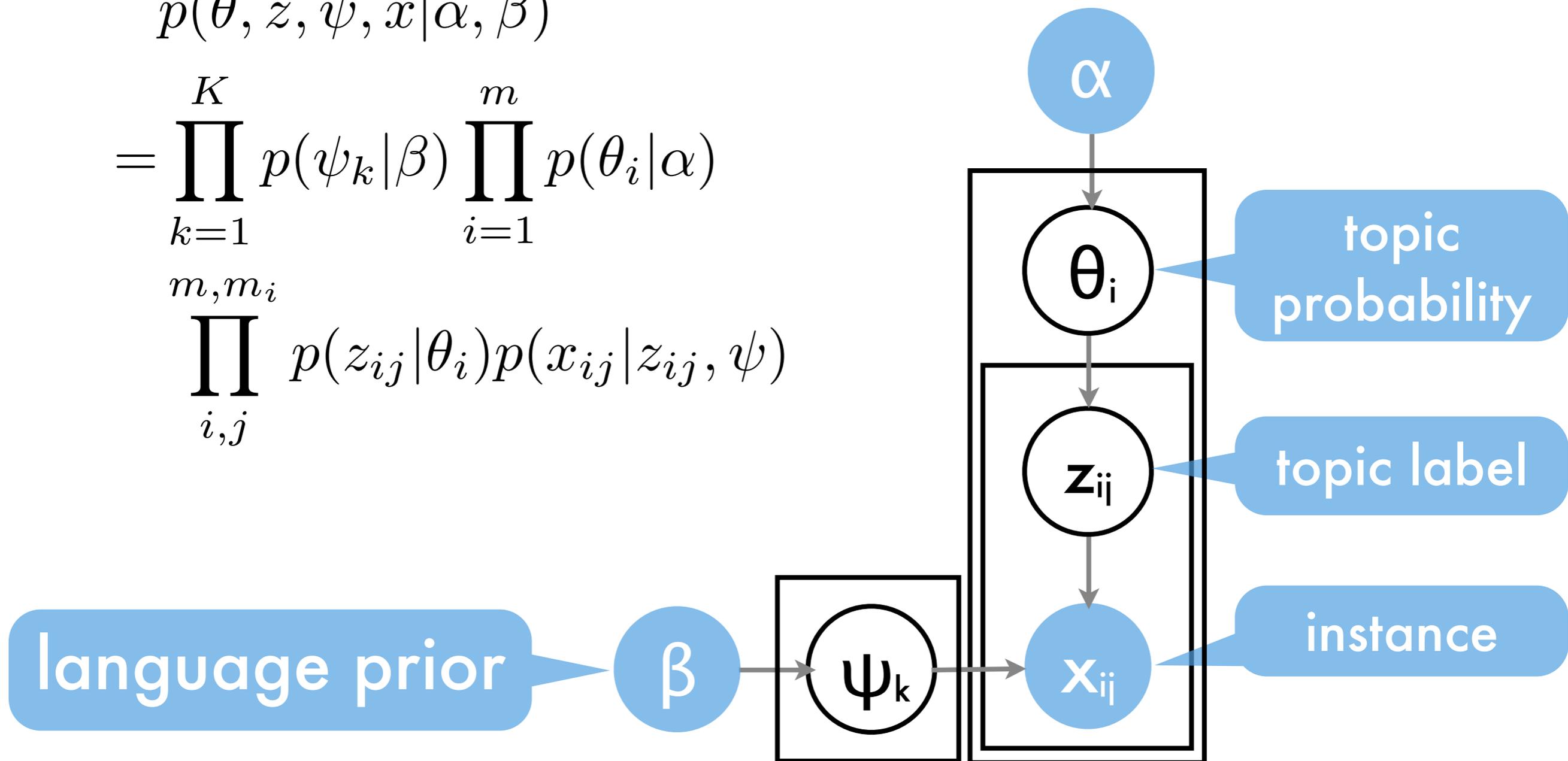
- **Posterior**

$$P(\theta|x, \alpha) \sim \text{Dir}(x_1 + \alpha_1, \dots, x_k + \alpha_k)$$

$$E[\theta_i|x, \alpha] = \frac{x_i + \alpha_i}{\sum_i x_i + \alpha_i}$$

# Joint Probability Distribution

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



# Joint Probability Distribution

sample  $\Psi$   
independently

sample  $\theta$   
independently

sample  $z$   
independently

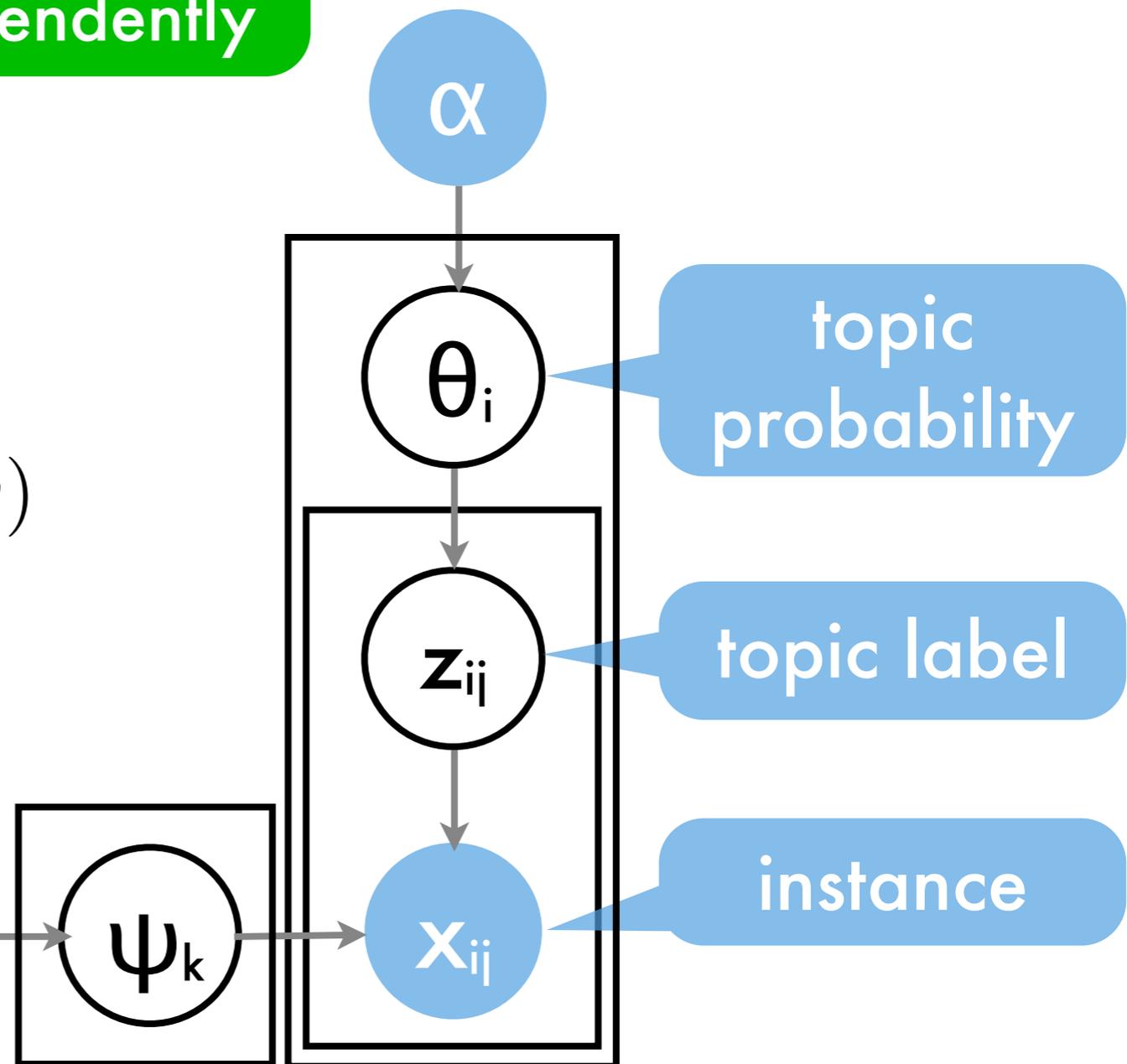
language prior

topic probability

topic label

instance

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



# Joint Probability Distribution

sample  $\Psi$   
independently

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

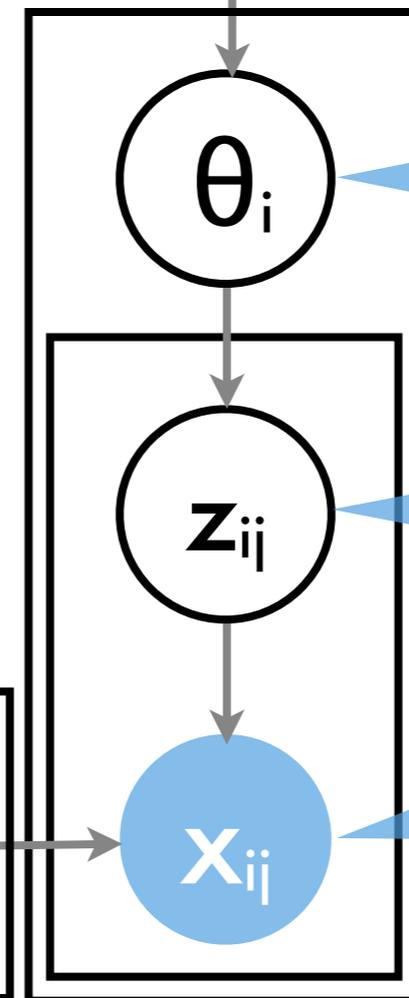
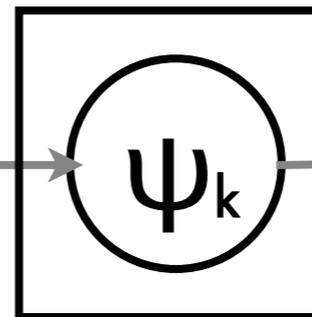
$$= \prod_{k=1}^K p(\psi_k | \beta) \prod_{i=1}^m p(\theta_i | \alpha)$$

sample  $\theta$   
independently

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

sample  $z$   
independently

language prior



slow

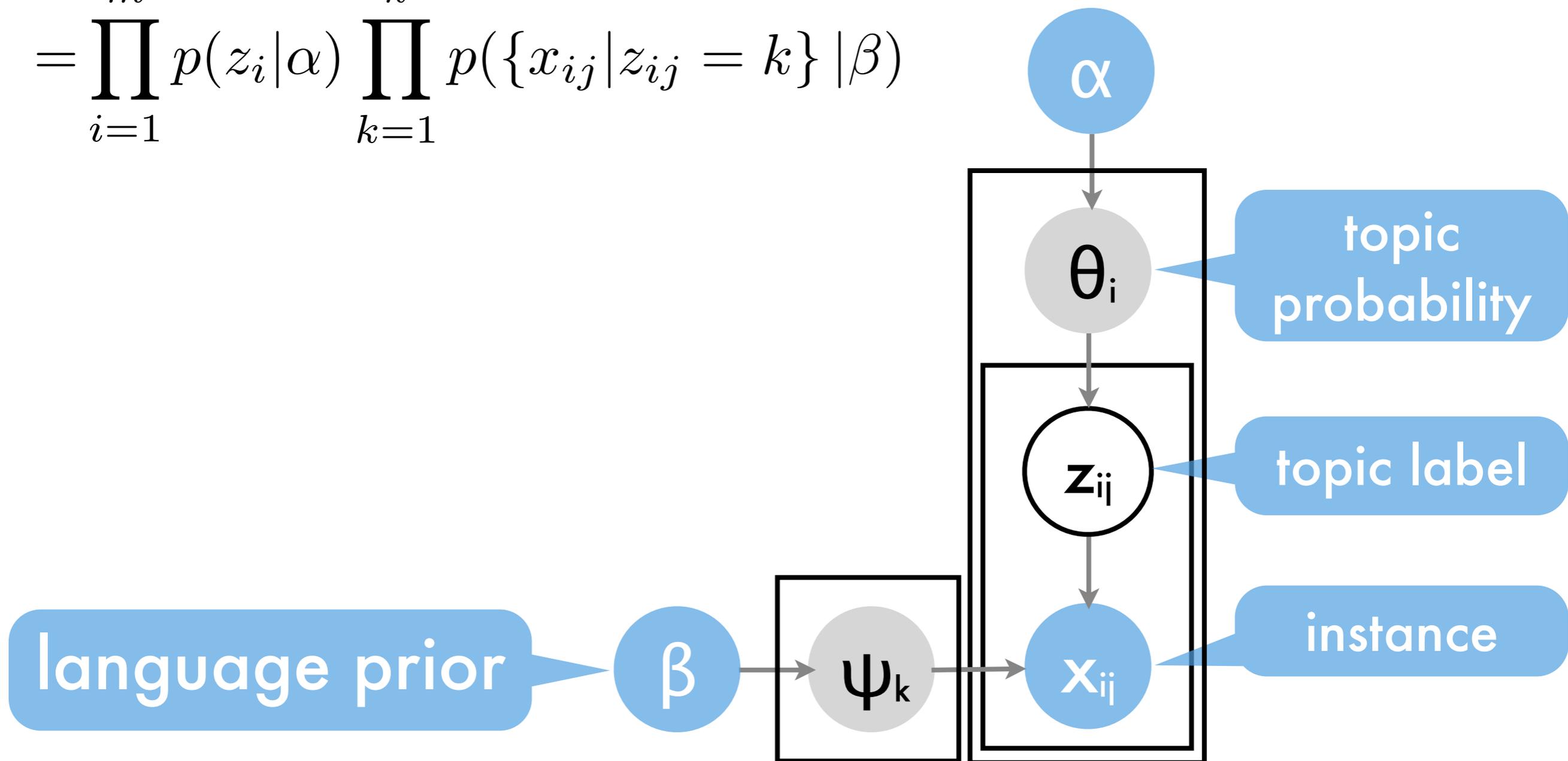
topic  
probability

topic label

instance

# Collapsed Sampler

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



# Collapsed Sampler

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

sample  $z$   
sequentially

language prior

$\beta$

$\psi_k$

$x_{ij}$

$\alpha$

$\theta_i$

$z_{ij}$

topic  
probability

topic label

instance

# Collapsed Sampler

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

sample  $z$   
sequentially

fast

language prior

$\beta$

$\psi_k$

$x_{ij}$

$z_{ij}$

$\theta_i$

$\alpha$

topic probability

topic label

instance

# Collapsed Sampler

Griffiths & Steyvers, 2005

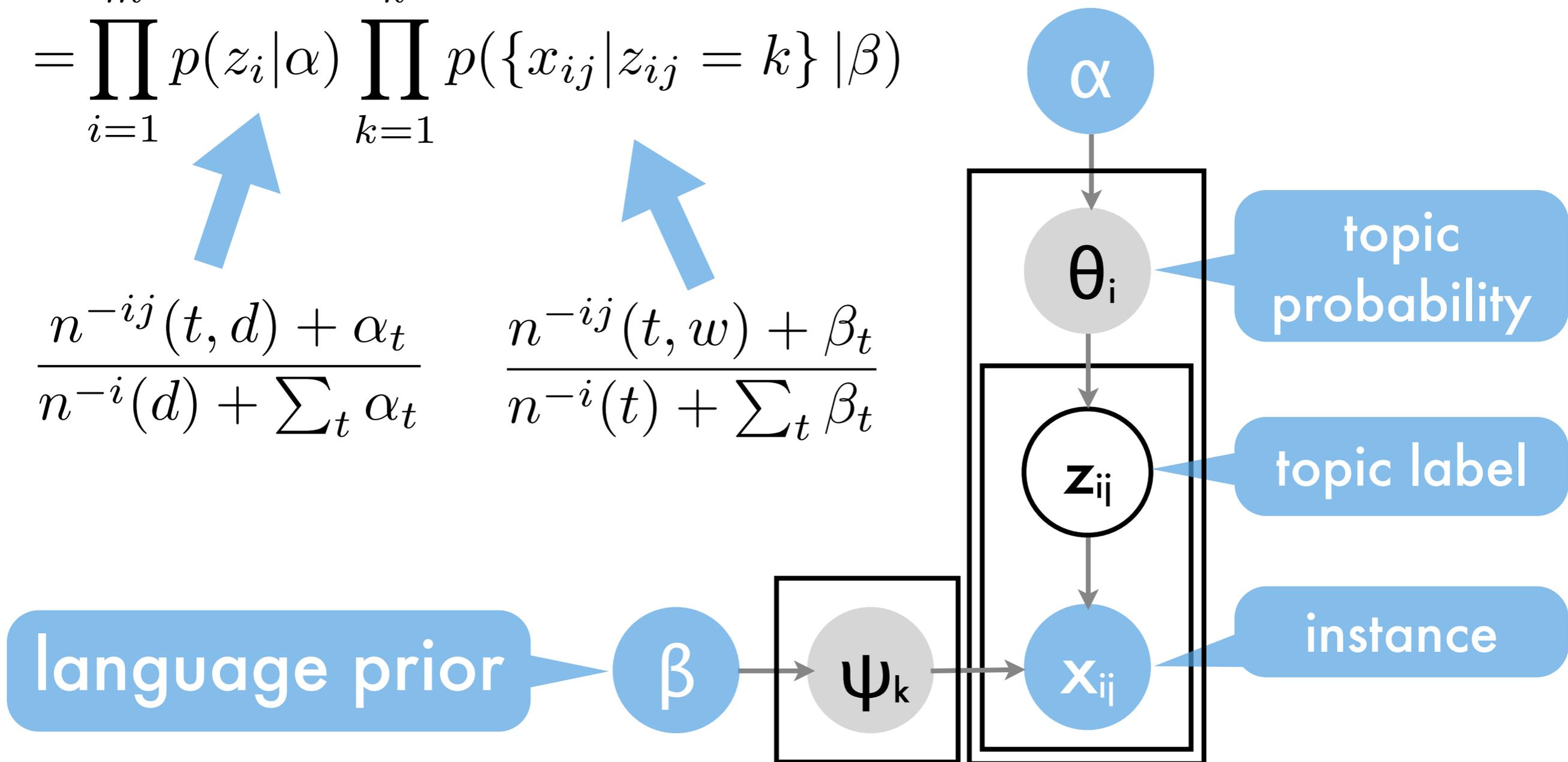
$$p(z, x | \alpha, \beta)$$

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

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

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

language prior



# Collapsed Sampler

Griffiths & Steyvers, 2005

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

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

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

language prior

$\beta$

$\psi_k$

$x_{ij}$

$z_{ij}$

$\theta_i$

$\alpha$

fast

topic probability

topic label

instance

# Derivations (was on the board)

$$p(z_{ij} = t | z^{-ij}, \alpha) =$$

total rule of probability

$$\int_{\theta} p(\theta, z_{ij} = t | z^{-ij}, \alpha) d\theta$$

Chain rule

=

$$\int_{\theta} p(\theta | z^{-ij}, \alpha) p(z_{ij} = t | \theta, z^{-ij}, \alpha) d\theta$$

conditional independence

=

$$\int_{\theta} p(\theta | z^{-ij}, \alpha) p(z_{ij} = t | \theta) d\theta$$

=

$$\int_{\theta} p(\theta | z^{-ij}, \alpha) \theta_t d\theta$$

=

$$E_{p(\theta | z^{-ij}, \alpha)}[\theta_t]$$

=

mean of the posterior of  $\theta$  given other topic assignments

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

Derivation for the second factor follows similarly

# 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

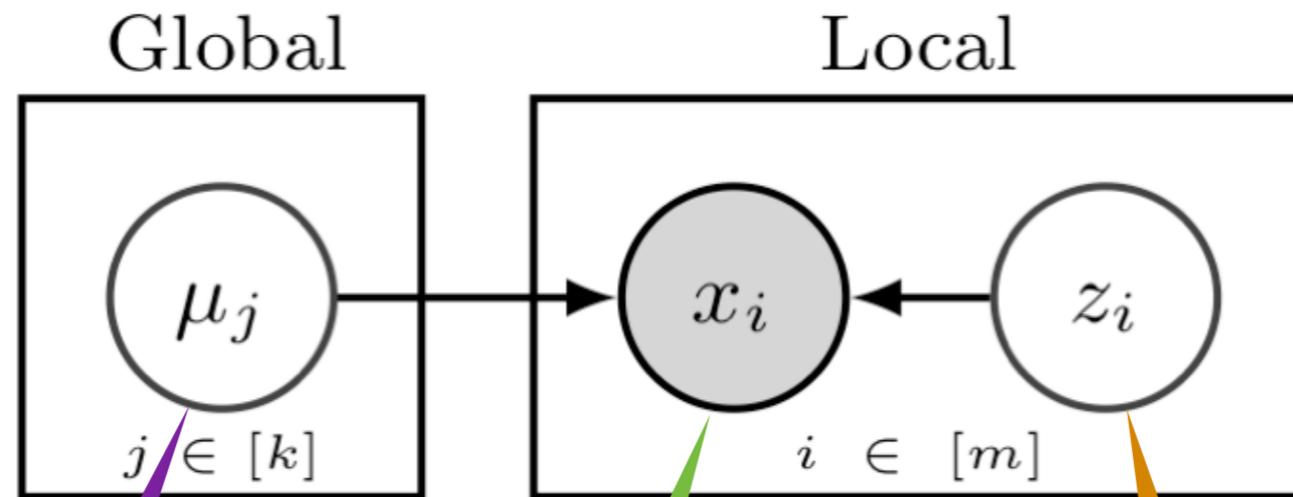
# Design Principles



# Scaling Problems



# 3 Problems

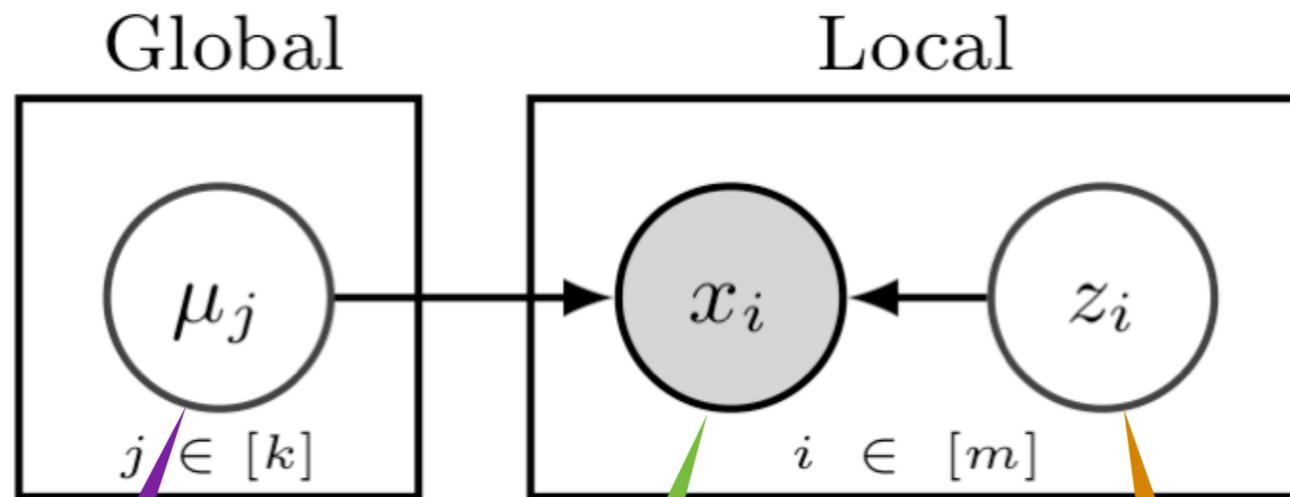


mean  
variance  
cluster weight

data

cluster ID

# 3 Problems

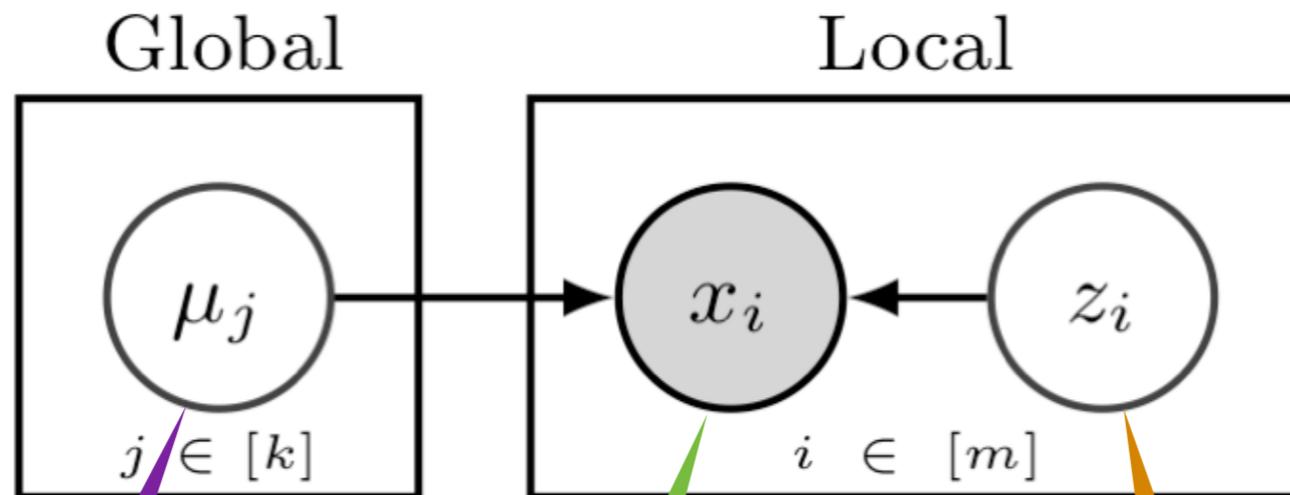


global state

data

local state

# 3 Problems



too big for single machine

huge

only local

# 3 Problems

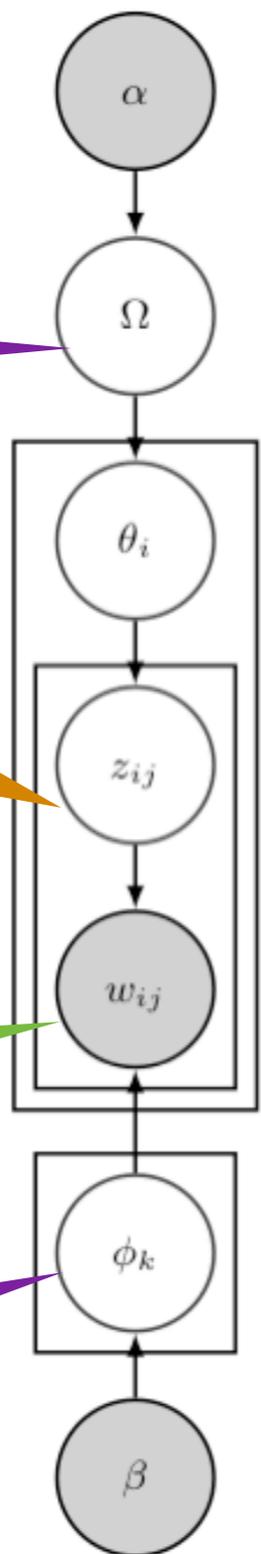
Vanilla LDA

global state

local state

data

global state



# 3 Problems

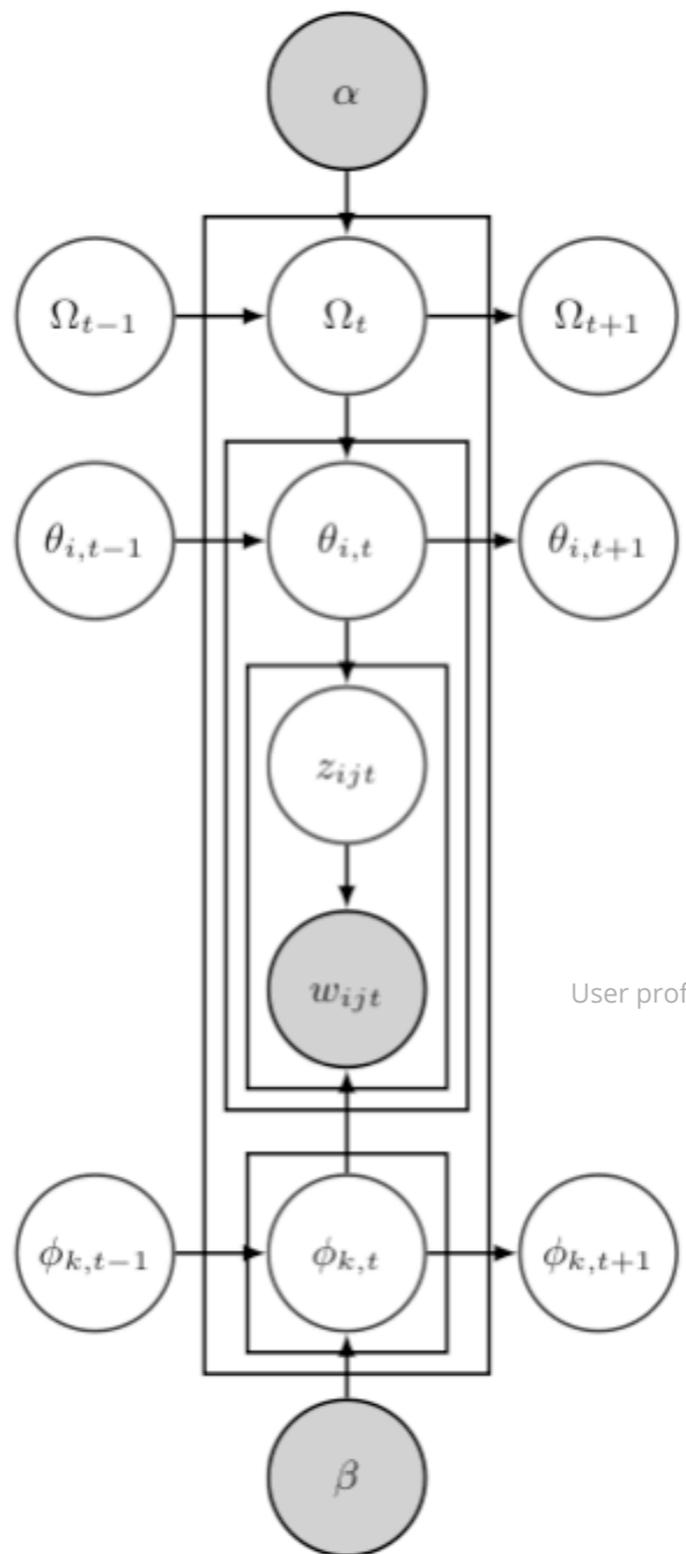
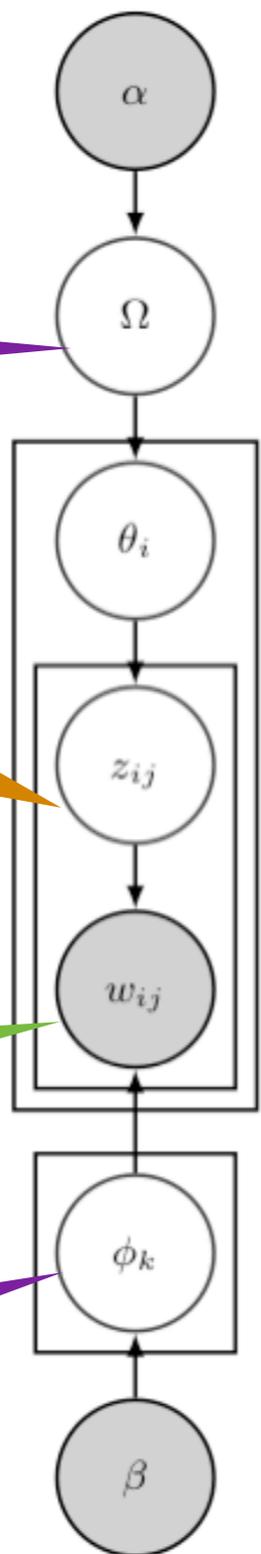
Vanilla LDA

global state

local state

data

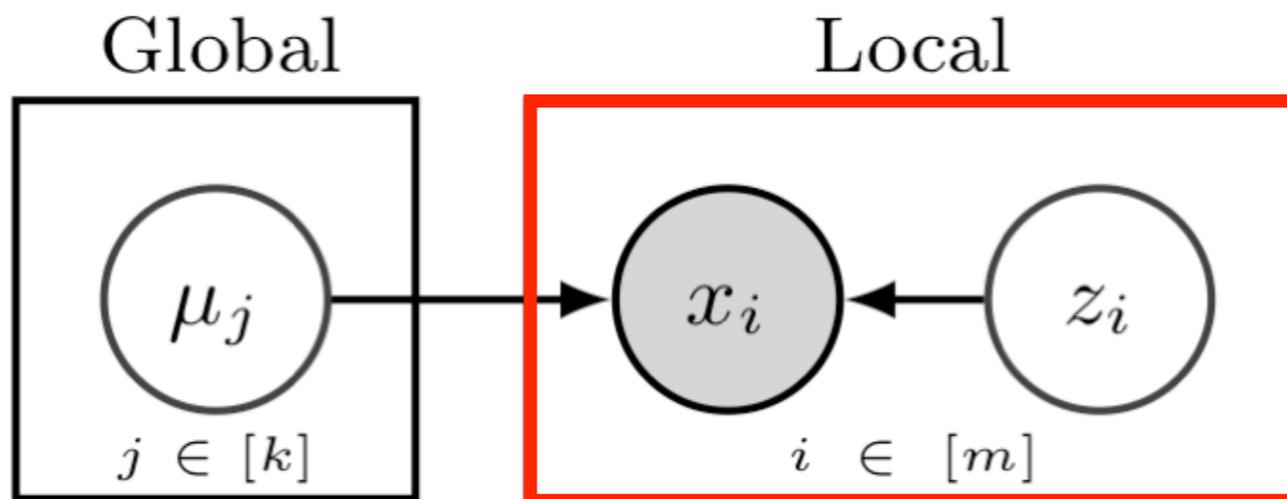
global state



User profiling

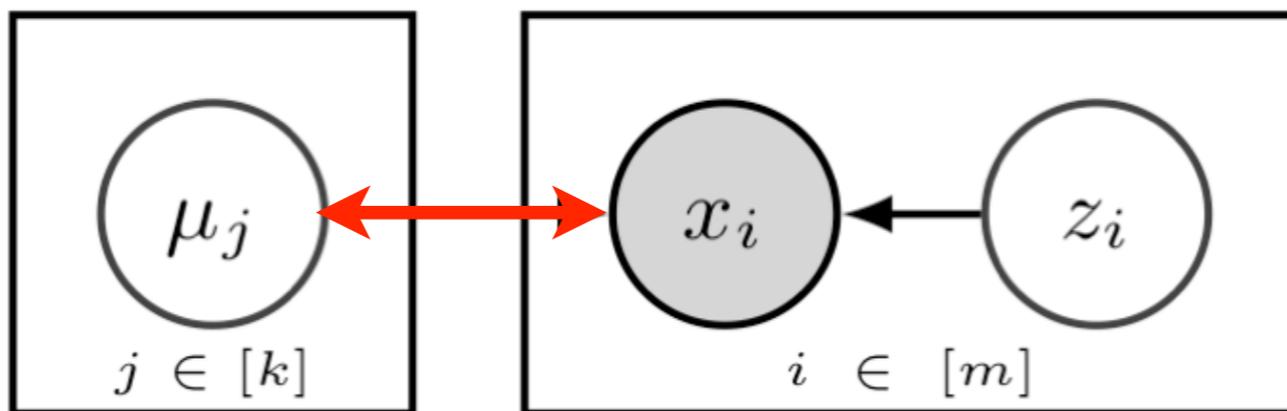
# 3 Problems

local state is too large

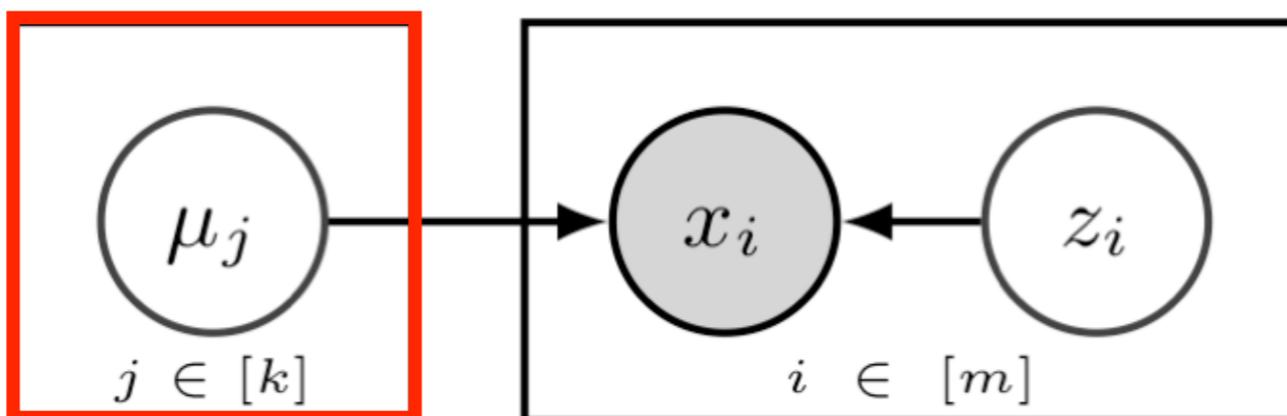


does not fit into memory

global state is too large



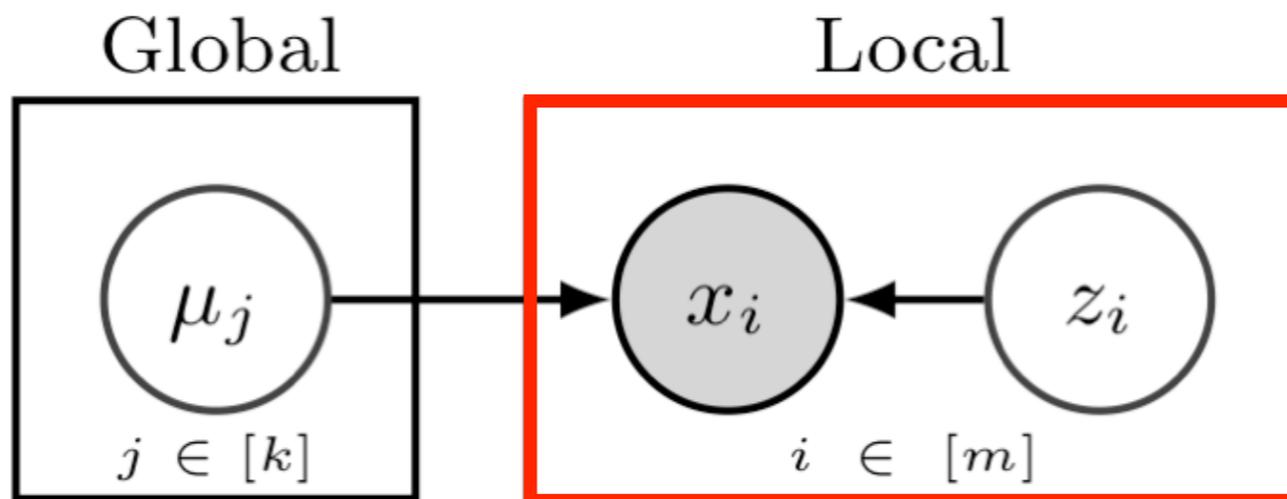
network load & barriers



does not fit into memory

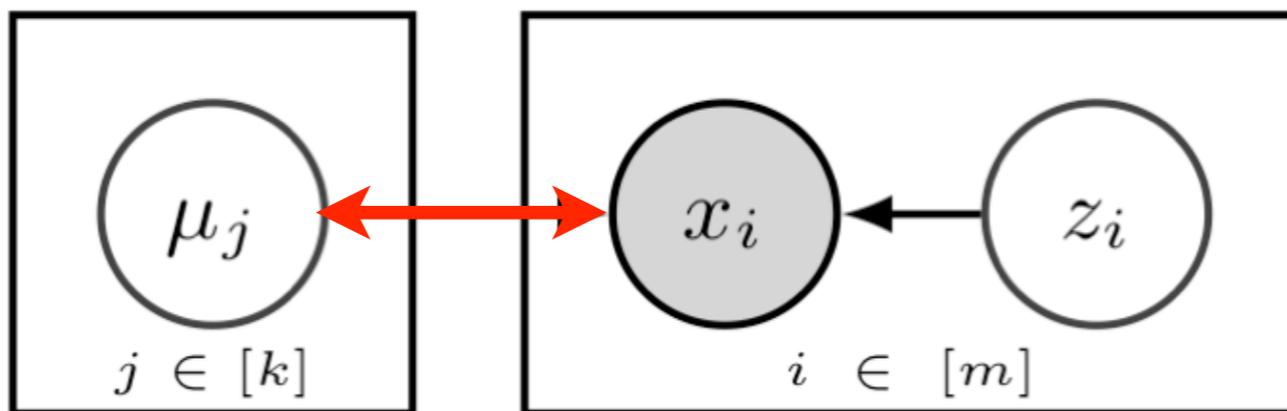
# 3 Problems

local state is too large

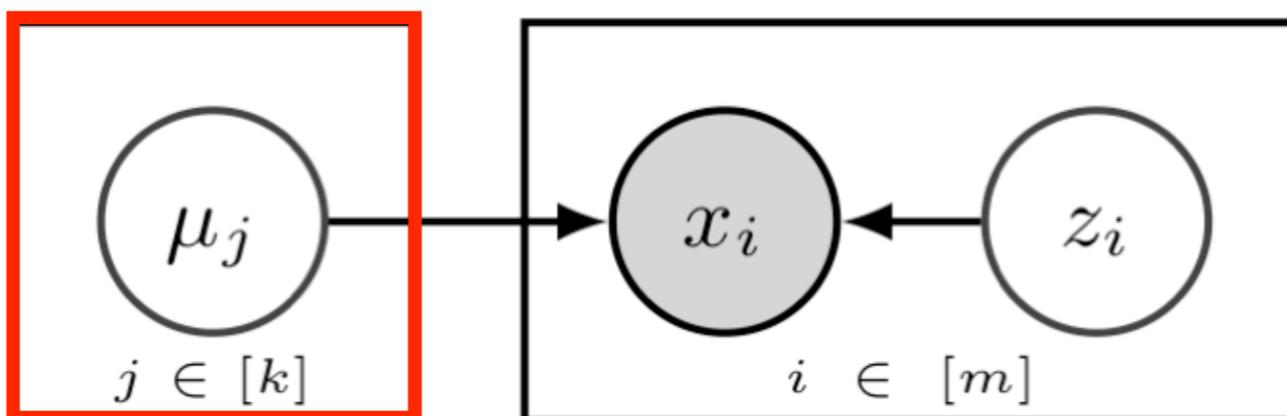


stream local data from disk

global state is too large



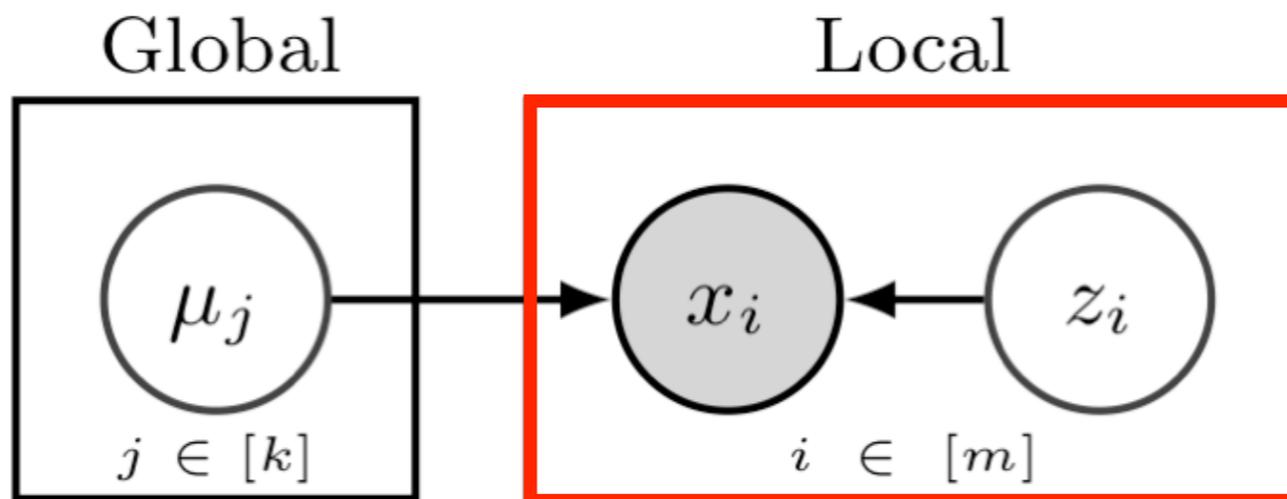
network load & barriers



does not fit into memory

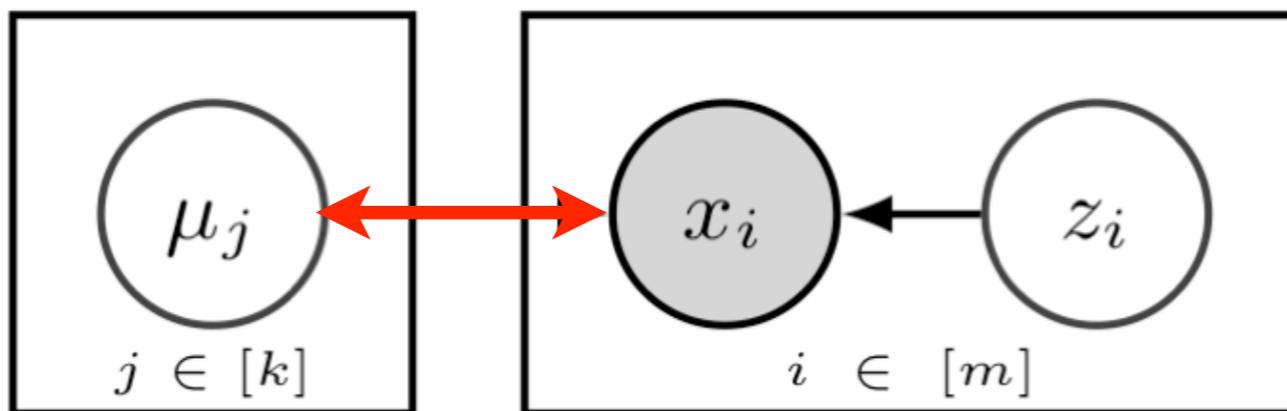
# 3 Problems

local state is too large

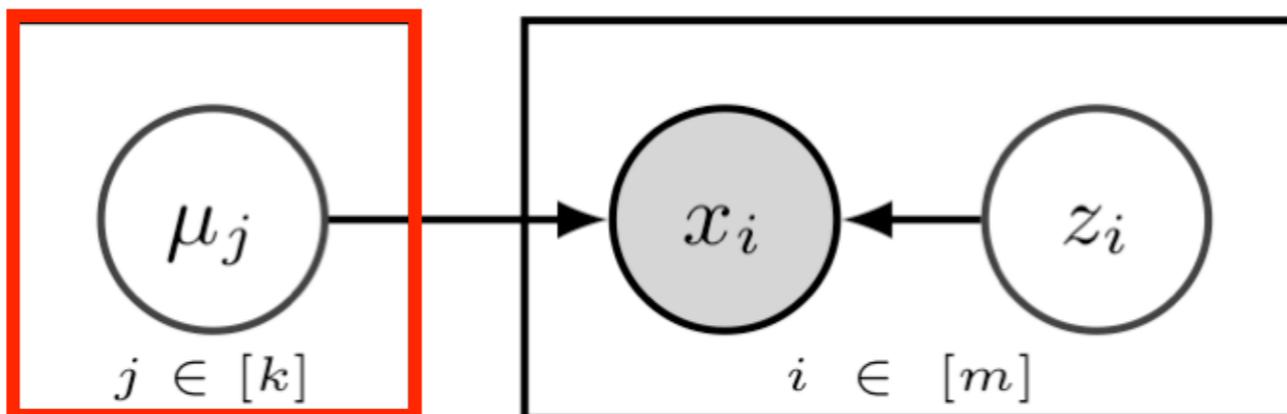


stream local data from disk

global state is too large



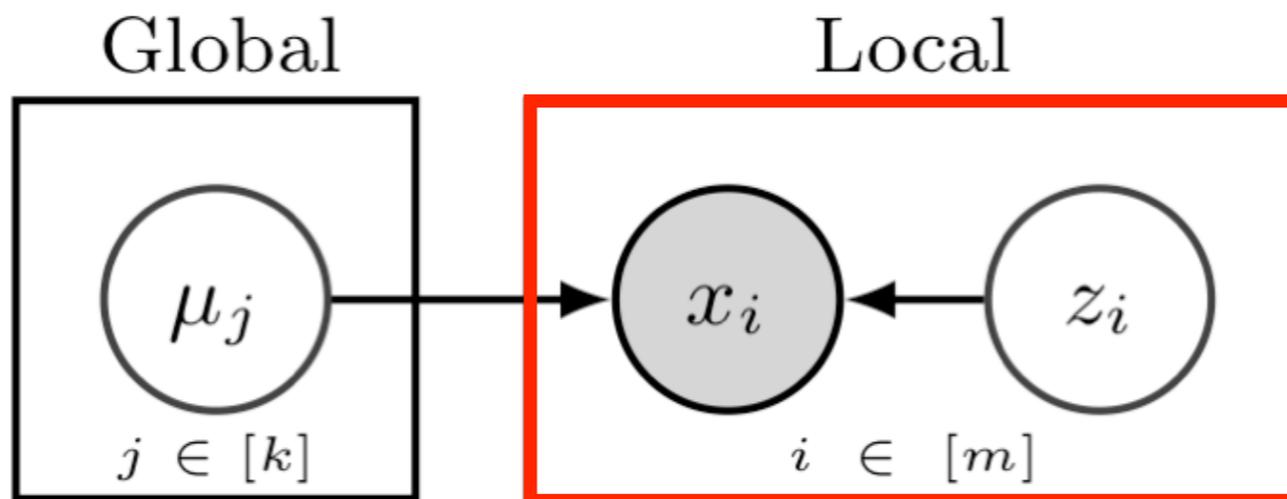
asynchronous synchronization



does not fit into memory

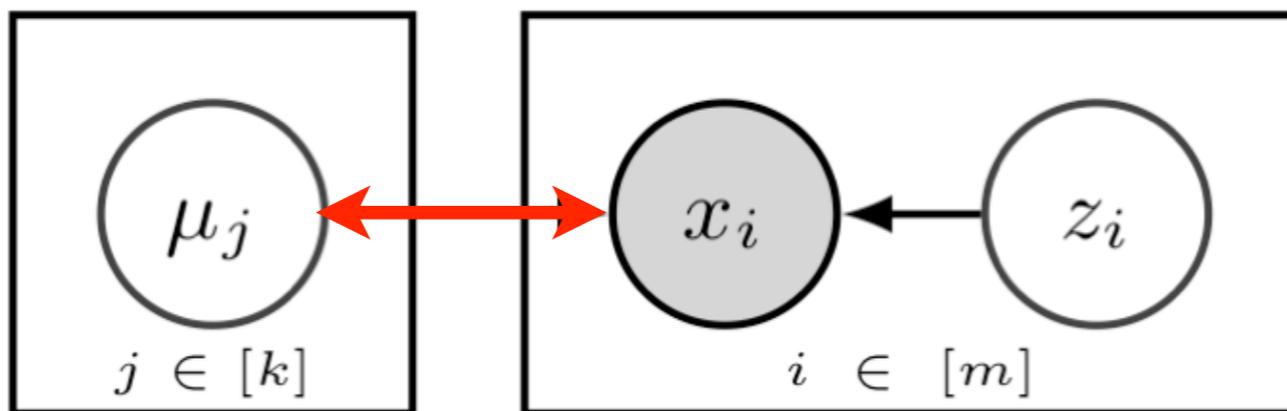
# 3 Problems

local state is too large

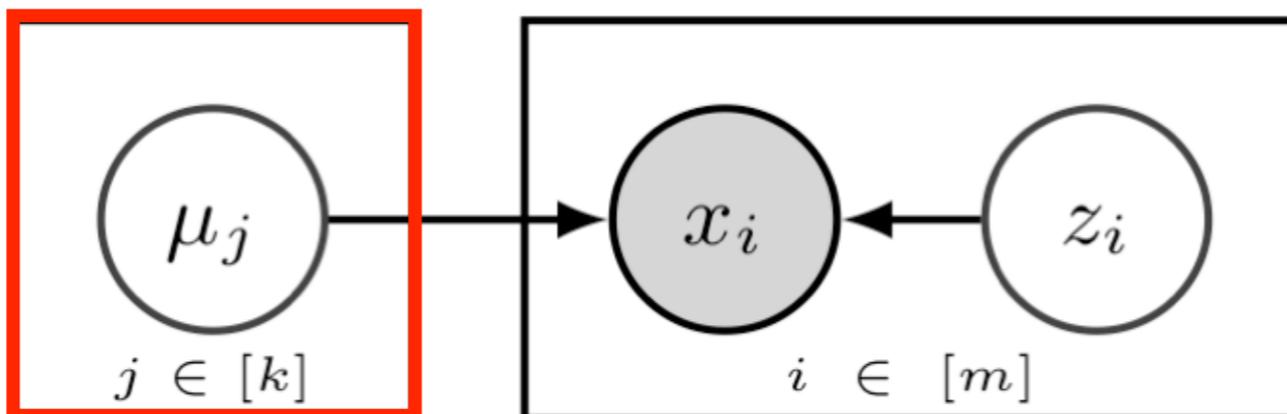


stream local data from disk

global state is too large



asynchronous synchronization

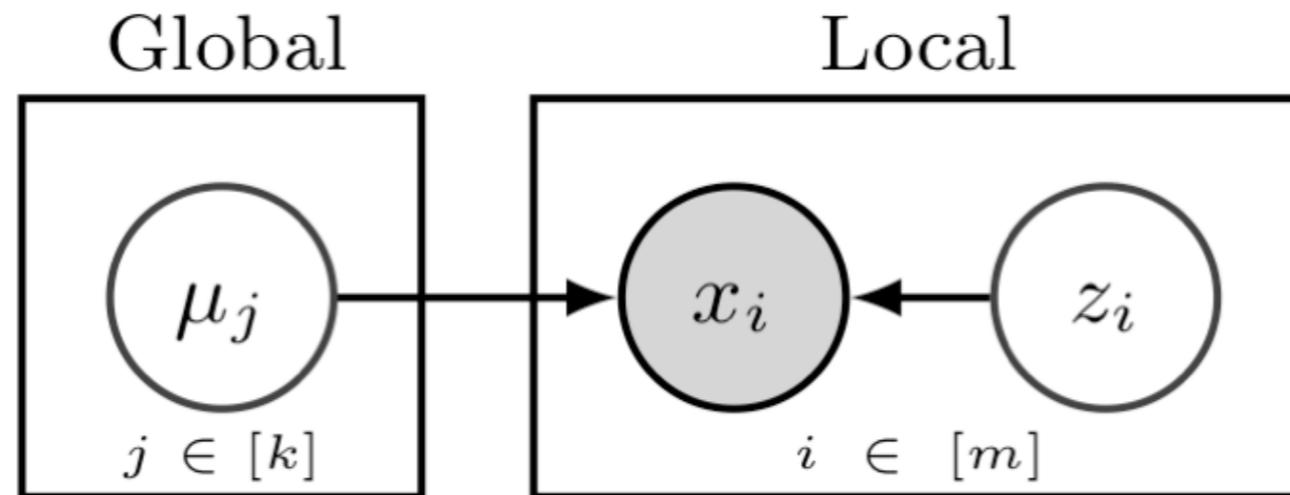


partial view

# Global state synchronization

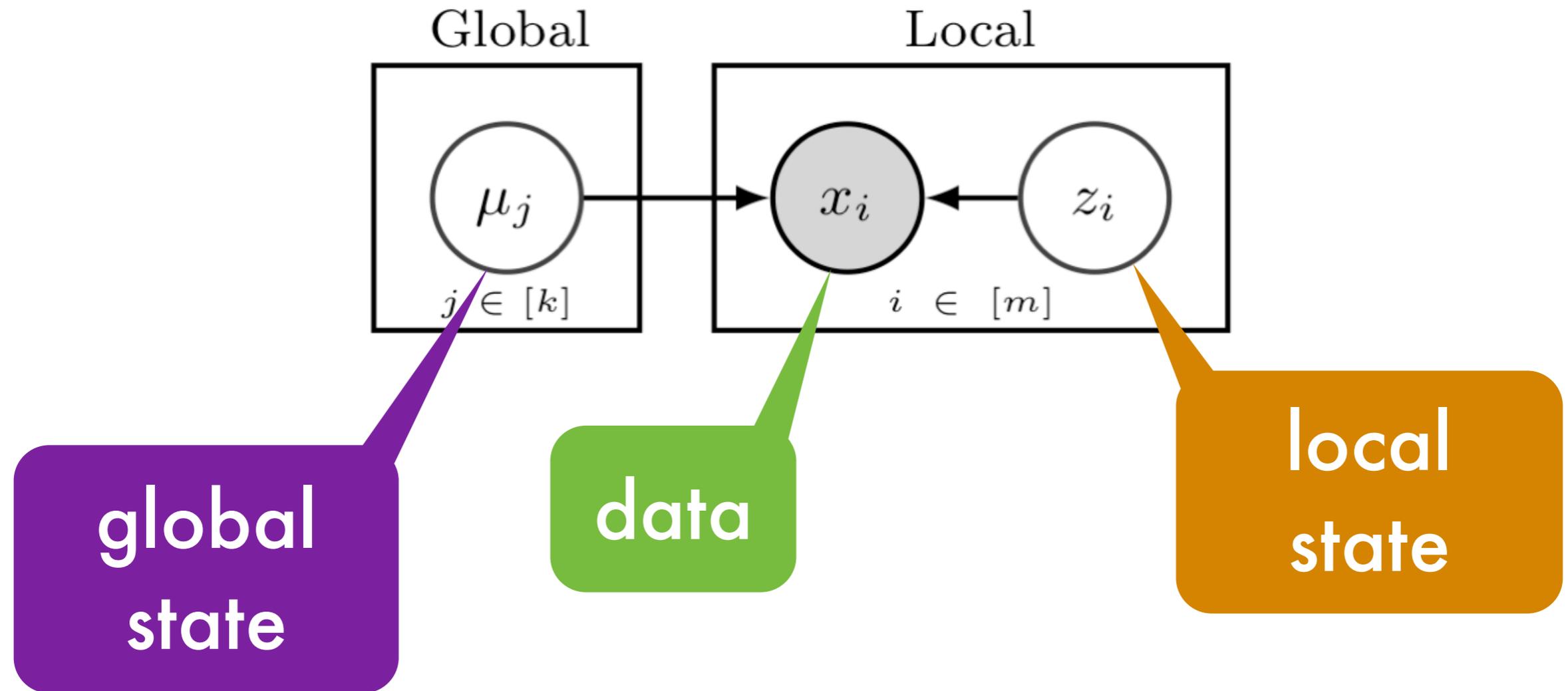


# Challenges



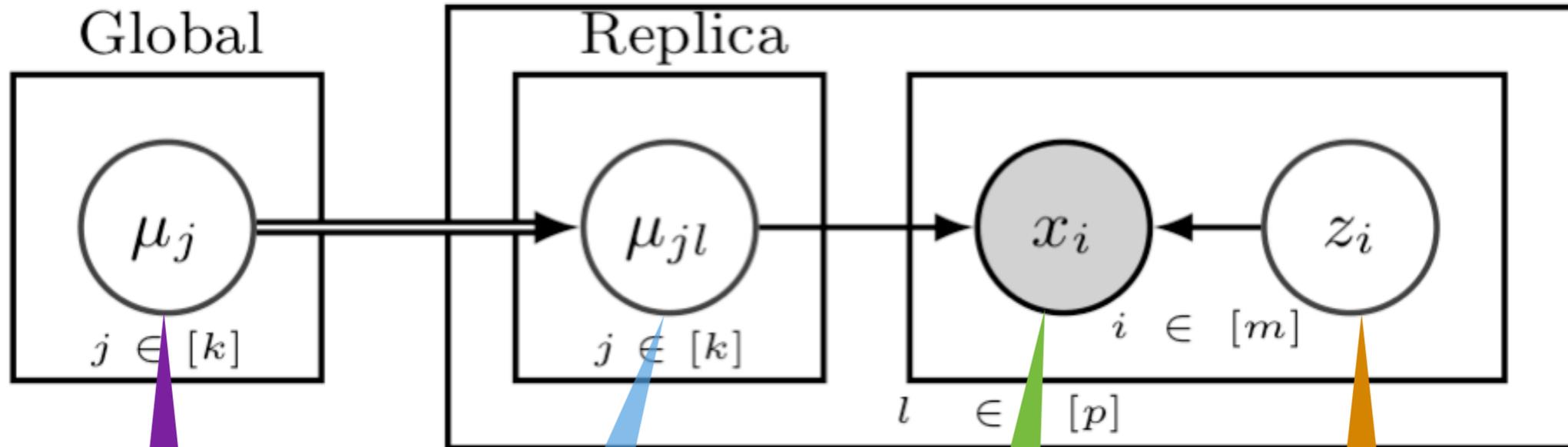
- Distribution (global)
- Synchronization (global)
- Fault tolerance
- Storage (local)

# Distribution



# Distribution

Processor Local State



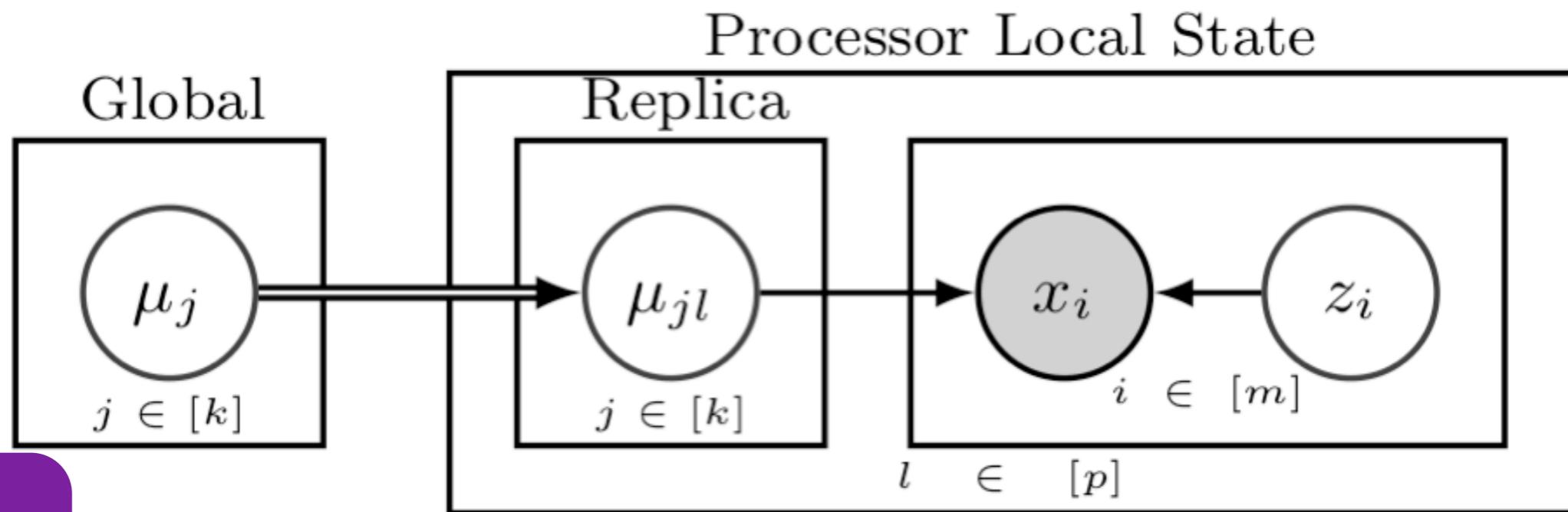
global  
state

copy

data

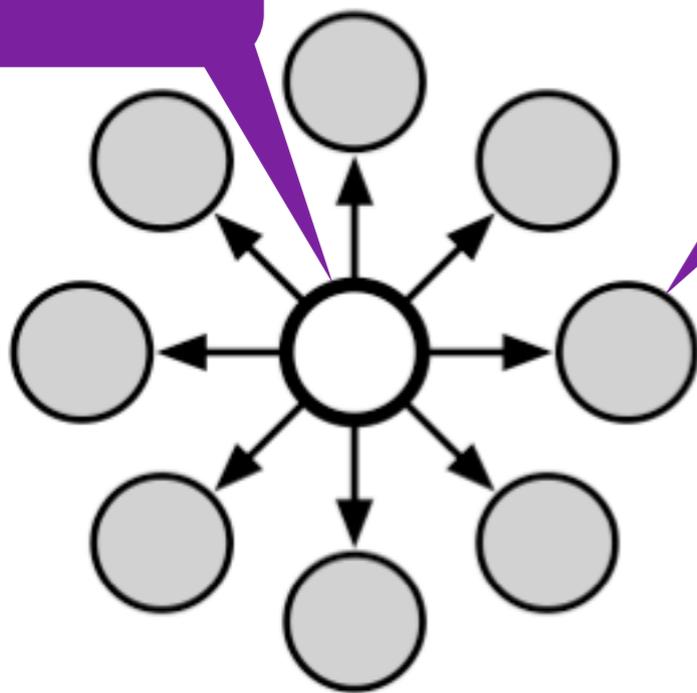
local  
state

# Distribution



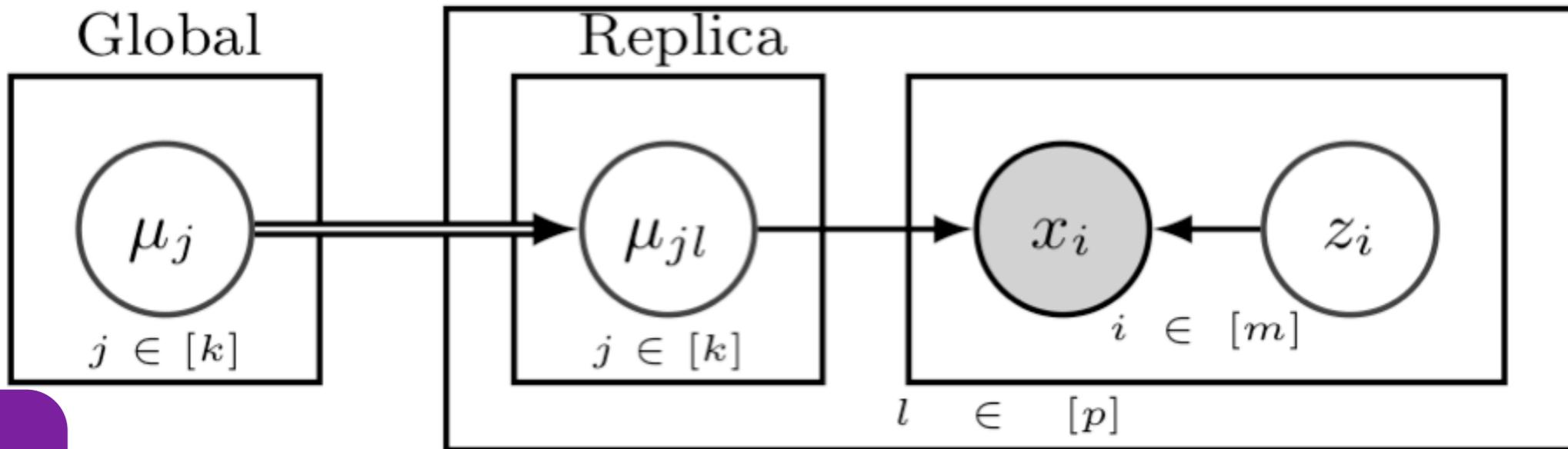
global

replica



# Distribution

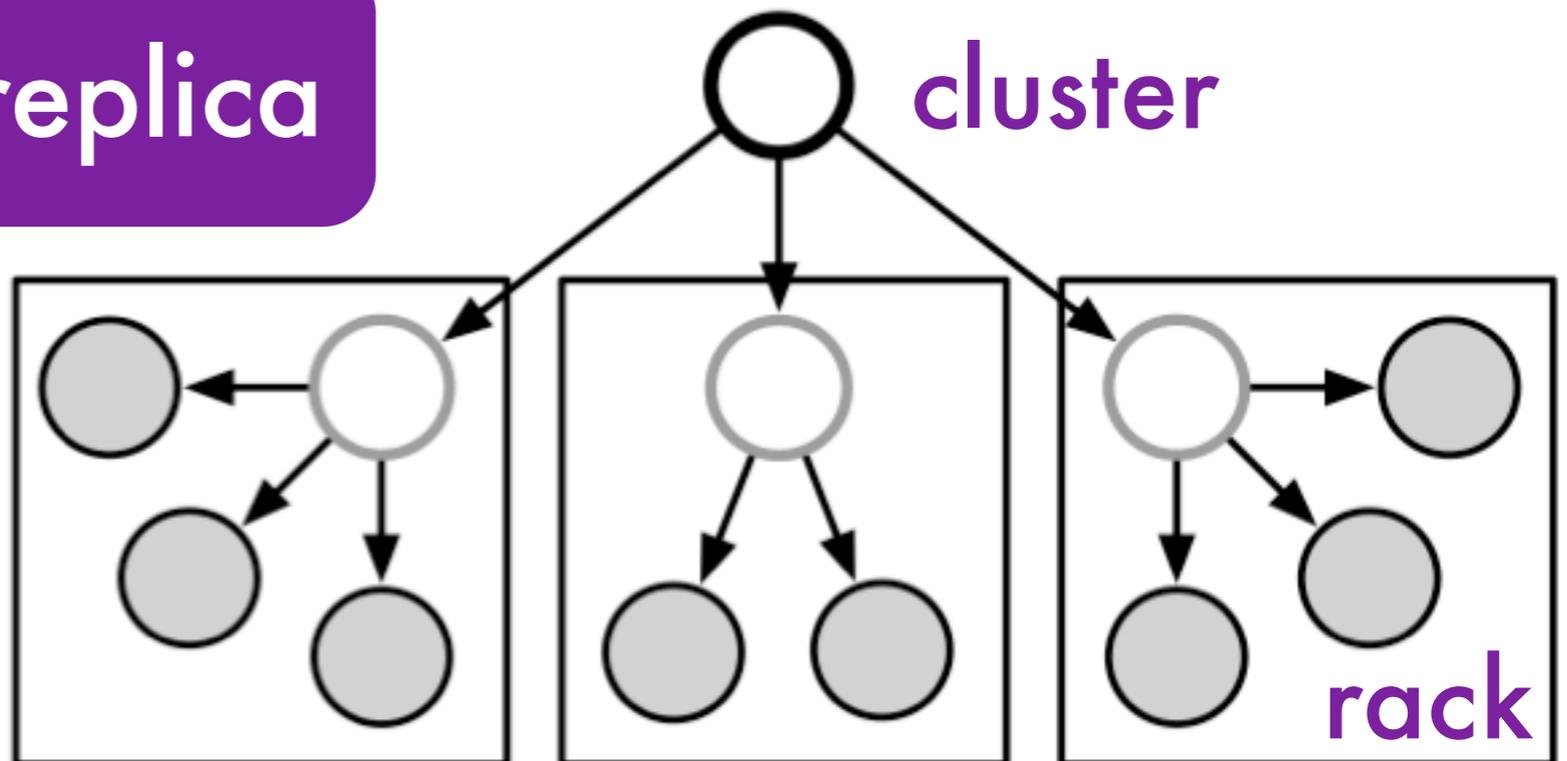
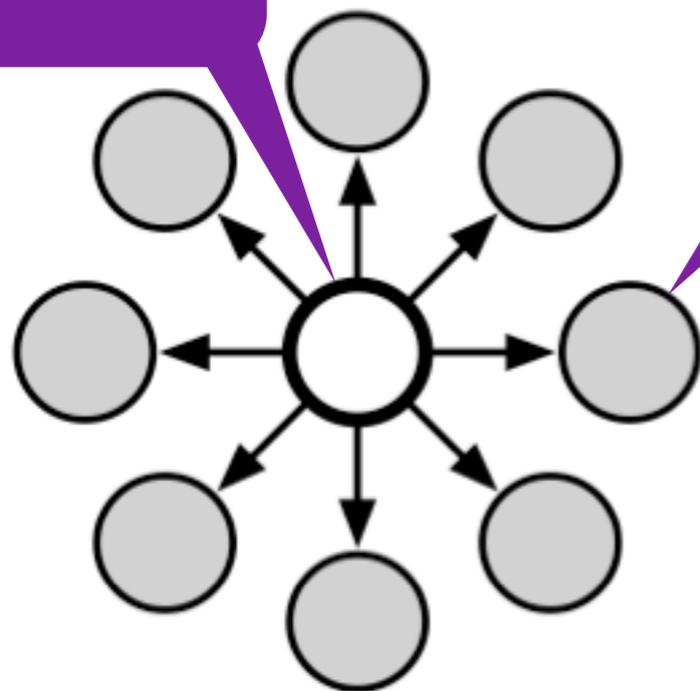
Processor Local State



global

replica

cluster



rack

# Synchronization

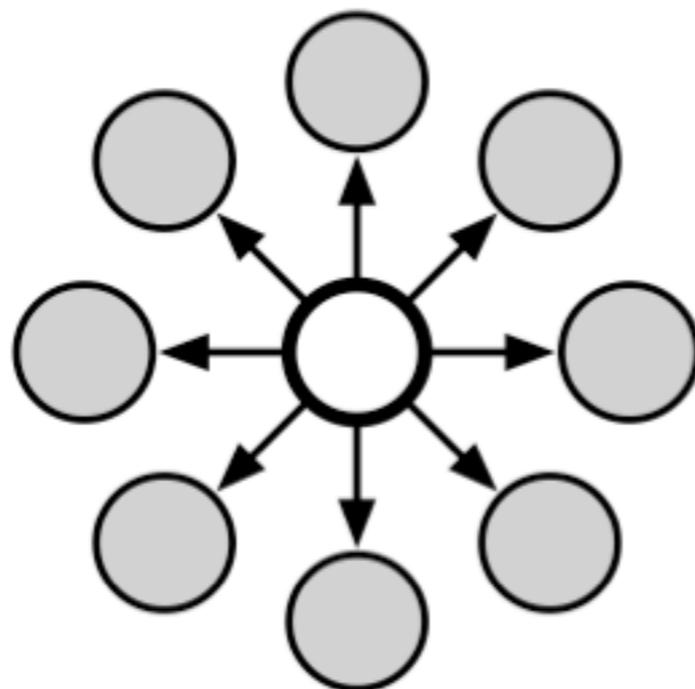
- Child updates local state
  - Start with common state
  - Child stores old and new state
  - Parent keeps global state
- Transmit differences asynchronously
  - Inverse element for difference
  - Abelian group for commutativity (sum, log-sum, cyclic group, exponential families)

local to global

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

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

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



global to local

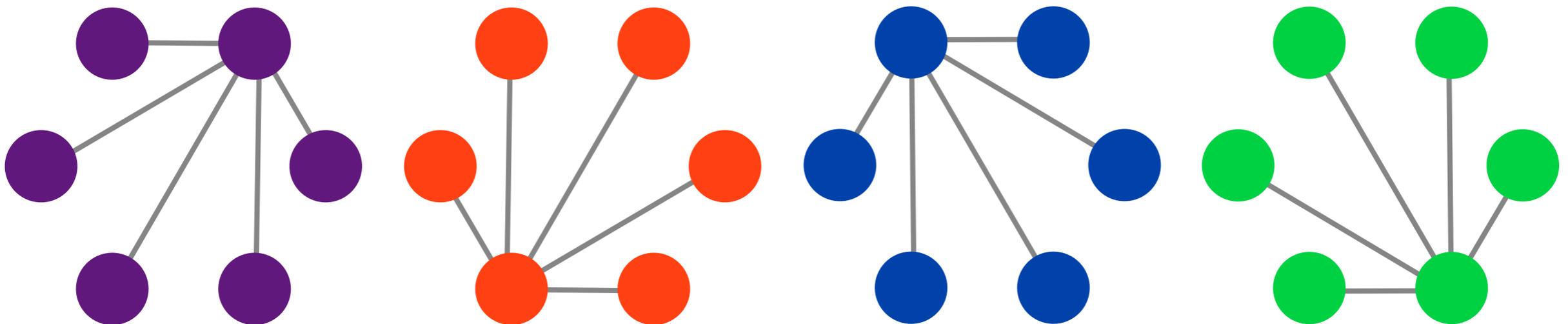
$$x \leftarrow x \oplus (x^{\text{global}} \ominus x^{\text{old}})$$

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

# Distribution

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

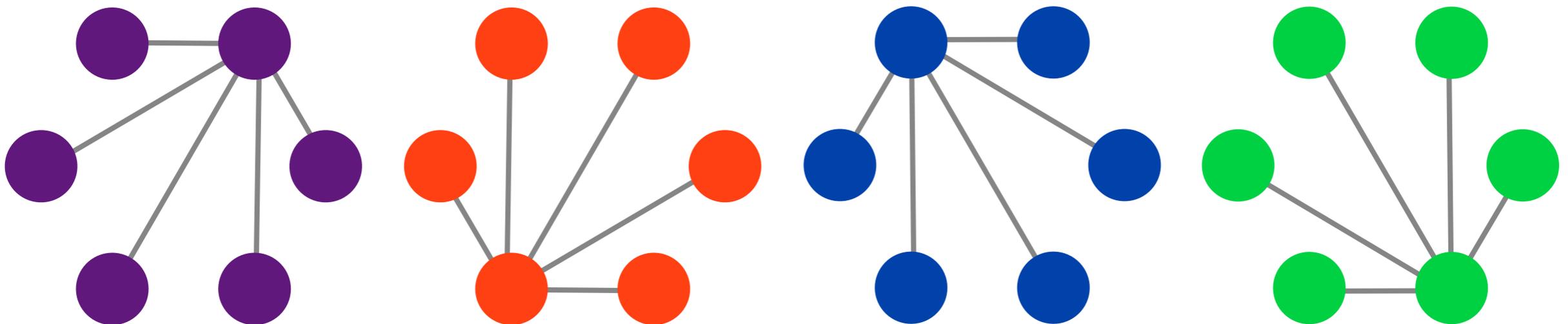
$$m(x) = \operatorname{argmin}_{m \in M} h(x, m)$$



# Distribution & fault tolerance

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

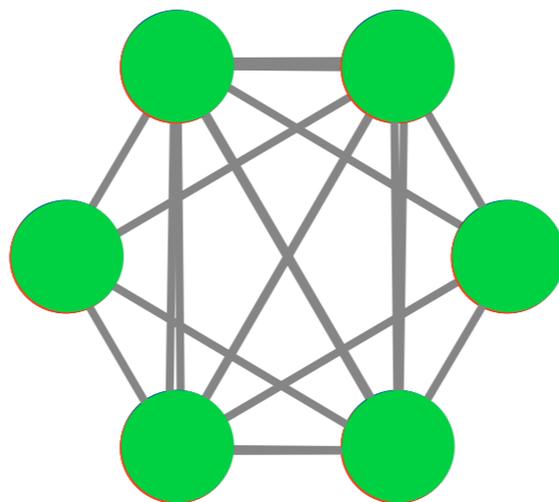
$$m(x) = \operatorname{argmin}_{m \in M} h(x, m)$$



# Distribution & fault tolerance

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

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



# Synchronization

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

client



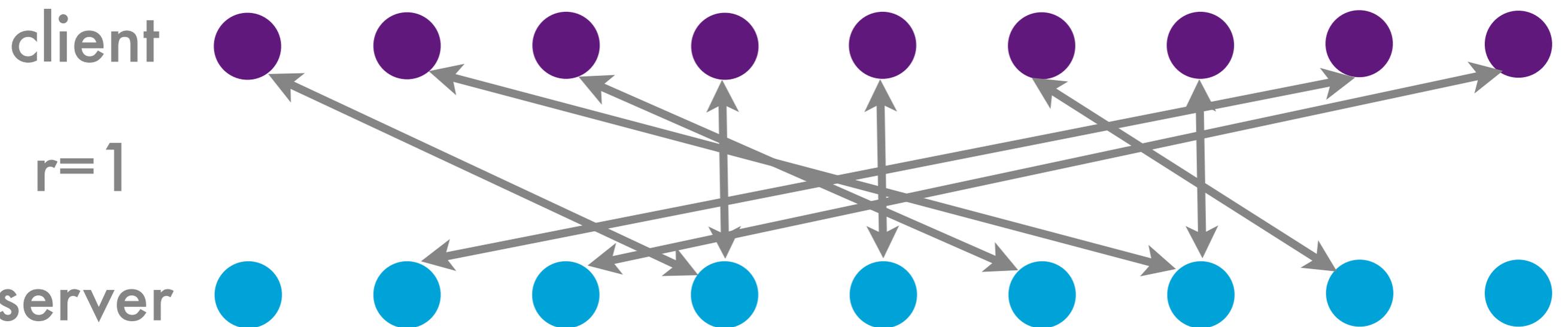
$r=1$

server



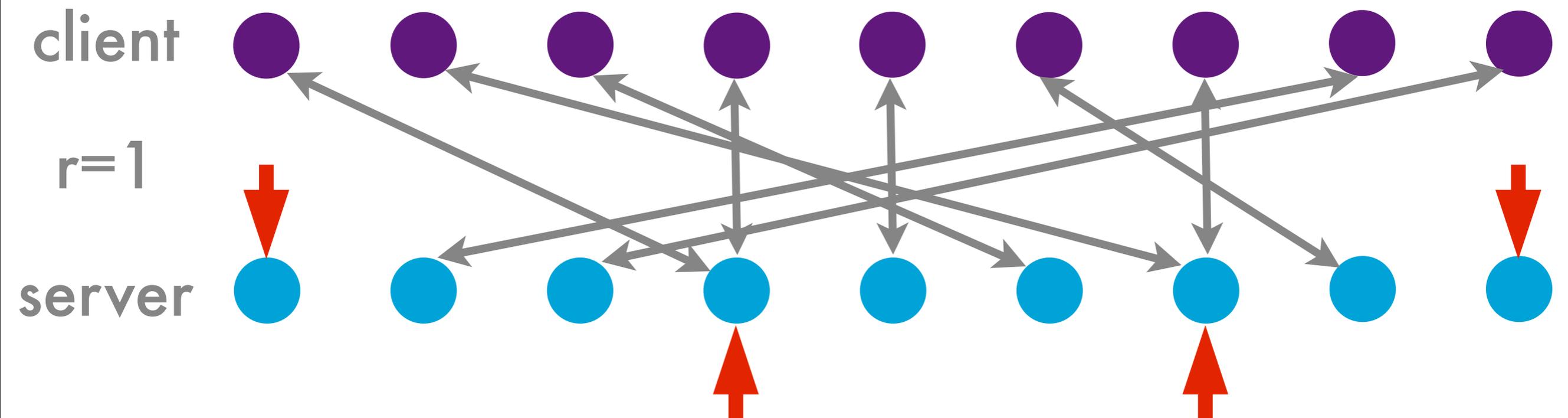
# Synchronization

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



# Synchronization

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



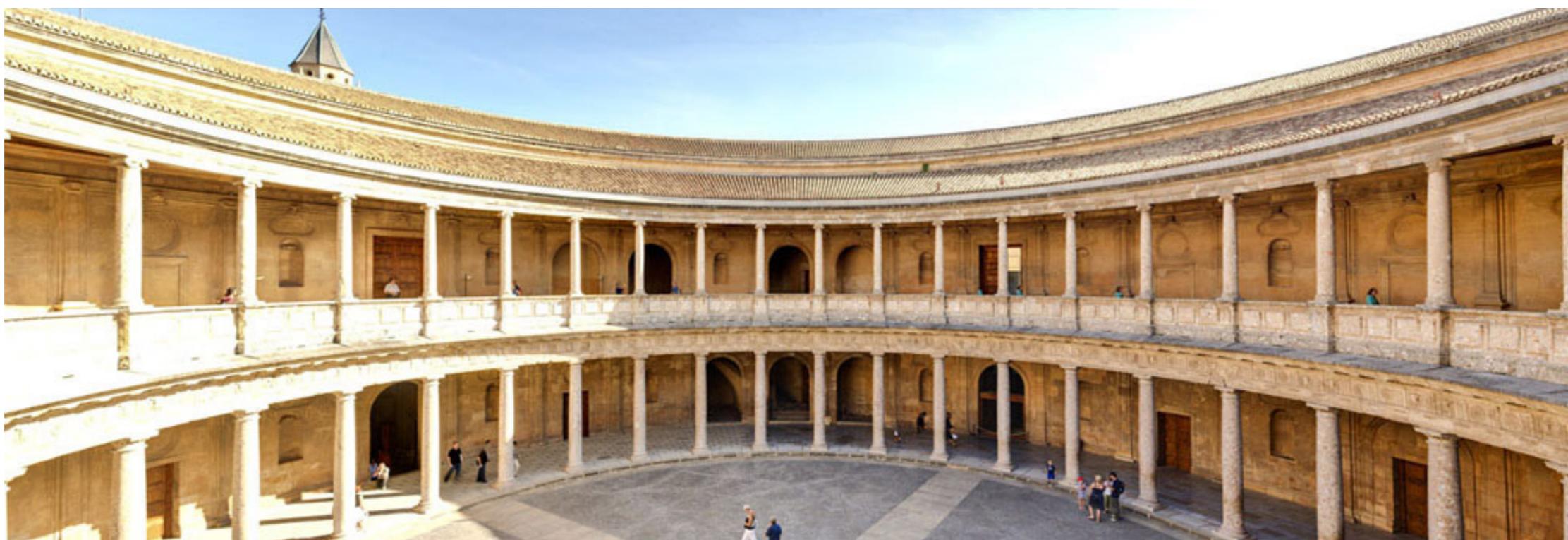
# Synchronization

- Data rate between machines is  $O(1/k)$
- Machines operate asynchronously (barrier free)
- Solution
  - Schedule message pairs
  - Communicate with  $r$  random machines simultaneously
- Efficiency guarantee [Ahmed et. al WSDM 2012]

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

4 simultaneous connections are sufficient

# Architecture



# 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

# Distributed asynchronous sampler

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

# Distributed asynchronous sampler

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

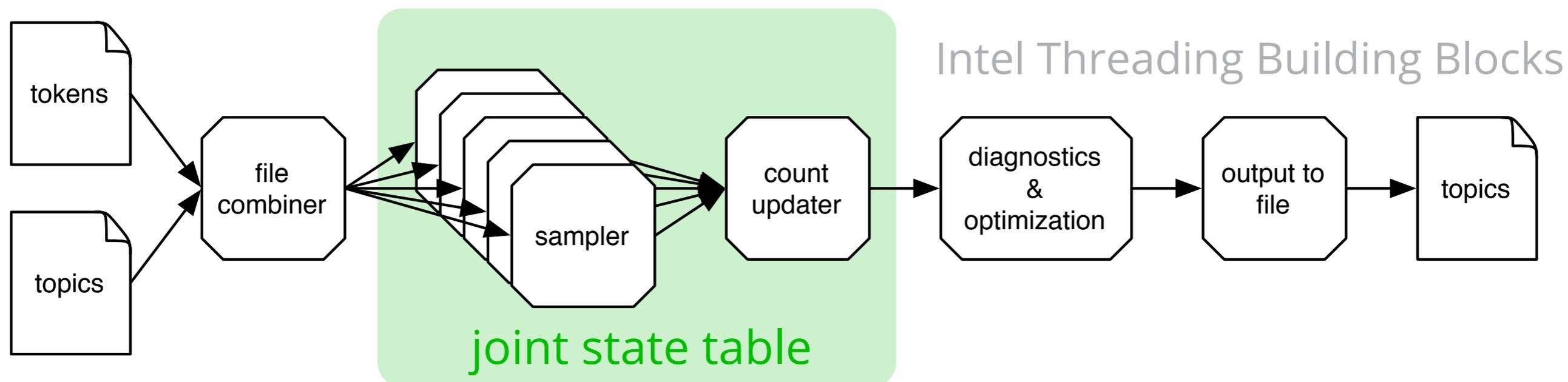
concurrent  
cpu hdd net

minimal  
view

continuous  
sync

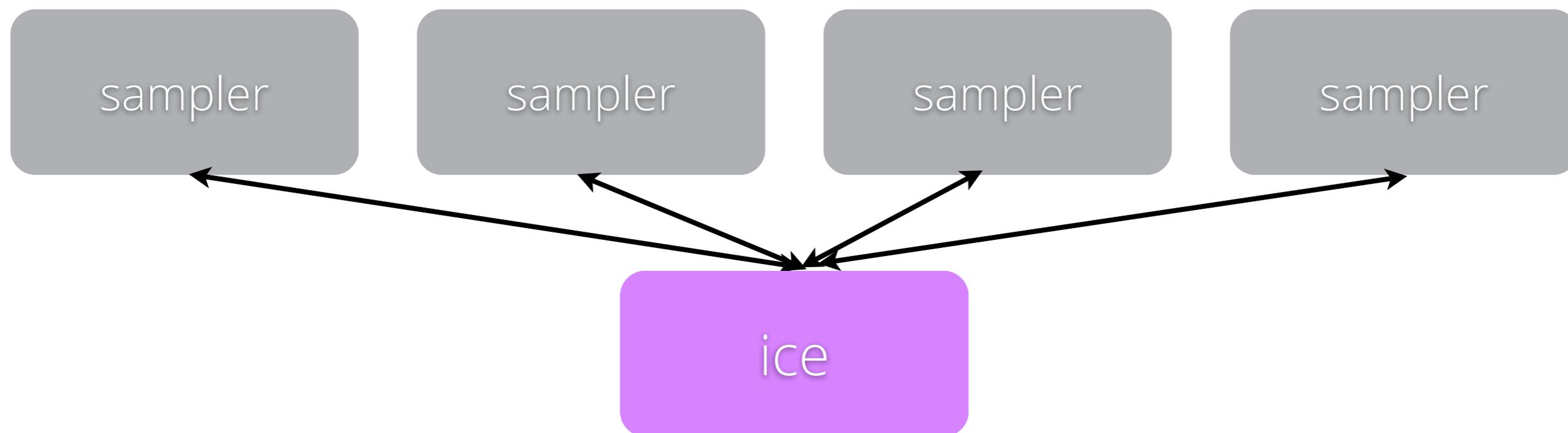
barrier free

# Multicore Architecture



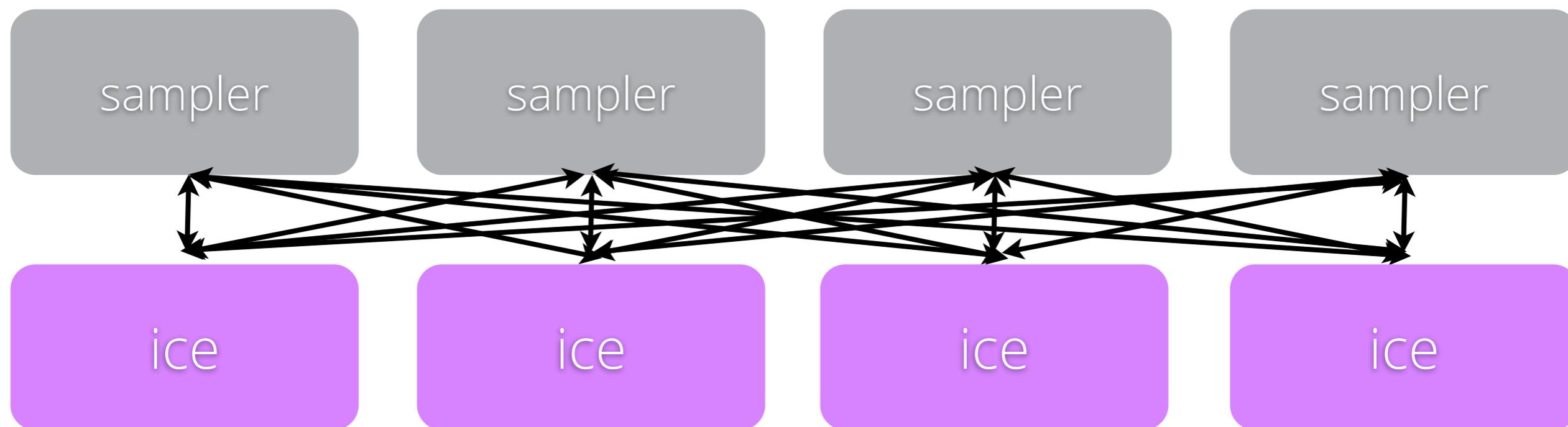
- Decouple multithreaded sampling and updating (almost) avoids stalling for locks in the sampler
- Joint state table
  - much less memory required
  - samplers synchronized (10 docs vs. millions delay)
- Hyperparameter update via stochastic gradient descent
- No need to keep documents in memory (streaming)

# Cluster Architecture



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

# Cluster Architecture



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

# Making it work

- Startup

- Naive: randomly initialize topics on each node (read from disk if already assigned - hotstart)
- Forward sampling for startup **much faster**
- Aggregate changes on the fly

- Failover

- State constantly being written to disk (worst case we lose 1 iteration out of 1000)
- Restart via standard startup routine

- Achilles heel: need to restart from checkpoint if even a single machine dies.

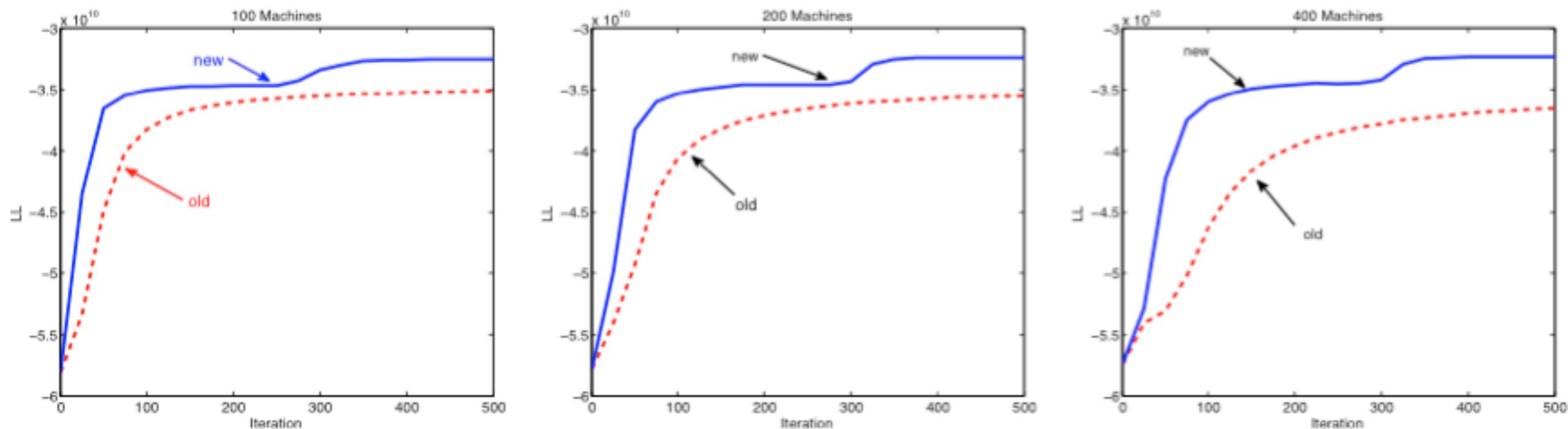
# Easily extensible

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

# Speed (2010 numbers)

- **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**

# Fast sampler



- 8 Million documents, 1000 topics, {100,200,400} machines, LDA
- Red (symmetric latency bound message passing)
- Blue (asynchronous bandwidth bound message passing & message scheduling)
  - 10x faster synchronization time
  - 10x faster snapshots
  - Scheduling improves 10% already on 150 machines

# Summary

- Data
- Hardware
- Distributed latent variable inference
- Many models
  - User profiling
  - Multi-domain analysis
  - Social network analysis