

YAHOO!

# Graphical Models for the Internet

Alexander Smola

Yahoo! Research, Santa Clara, CA

Australian National University

[alex@smola.org](mailto:alex@smola.org) [blog.smola.org](http://blog.smola.org)

credits to Amr Ahmed, Yahoo Research, CA

# Outline

## 1. Systems

- Hardware (computer architecture / networks / data centers)
- Storage and processing (file systems, MapReduce, Dryad, S4)
- Communication and synchronization (star, ring, hashtable, distributed star, tree)

## 2. Applications on the internet

- User modeling (clustering, abuse, profiling)
- Content analysis (webpages, links, news)
- Search / sponsored search

## 3. Probabilistic modeling

- Basic probability theory
- Naive Bayes
- Density estimation (exponential families)

## 4. Directed graphical models

- Directed graph semantics (independence, factorization)
- Clustering and Markov models (basic model, EM, sampling)
- Dirichlet distribution

# Outline

## 5. Scalable topic modeling

- Latent Dirichlet Allocation
- Sampling and parallel communication
- Applications to user profiling

## 6. Applications of latent variable models

- Time dependent / context dependent models
- News articles / ideology estimation
- Recommendation systems

## 7. Undirected graphical models

- Conditional independence and graphs (vertices, chains, trellis)
- Message passing (junction trees, variable elimination)
- Variational methods, sampling, particle filtering, message passing

## 8. Applications of undirected graphical models (time permitting)

- Conditional random fields
- Information extraction
- Unlabeled data

# Part 1 - Systems





# Hardware

# Computers

- CPU
  - 8-16 cores (Intel/AMD servers)
  - 2-3 GHz (close to 1 IPC per core peak) - 10-100 GFlops/socket
  - 8-16 MB Cache (essentially accessible at clock speed)
  - Vectorized multimedia instructions (128bit wide, e.g. add, multiply, logical)
  - Deep pipeline architectures (branching is expensive)
- RAM
  - 8-128 GB depending on use
  - 2-3 memory banks (each 32bit wide - atomic writes!)
  - DDR3 (10GB/s per chip, **random access >10x slower**)
- Harddisk
  - 2-3 TB/disk
  - 100 MB/s sequential read from SATA2
  - 5ms latency (**no change over 10 years**), i.e. random access is slow
- Solid State Drives
  - 500 MB/s sequential read
  - Random writes are really expensive (read-erase-write cycle for a block)
  - Latency is 0.5ms or lower (controller & flash cell dependent)
- **Anything you can do in bulk & sequence is at least 10x faster**



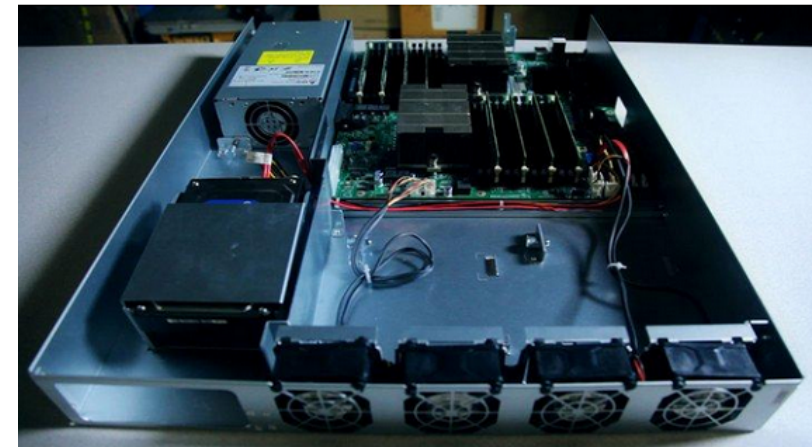
# Computers

- Network interface
  - Gigabit network (10-100MB/s throughput)
  - Copying RAM across network takes 1-10 minutes
  - Copying 1TB across network takes 1 day
  - Dozens of computers in a rack (on same switch)
- Power consumption
  - 200-400W per unit (don't buy the fastest CPUs)
  - Energy density big issue (cooling!)
  - GPUs take much more power (150W each)
- Systems designed to fail
  - Commodity hardware
  - Too many machines to ensure 100% uptime
  - Design software to deal with it (monitoring, scheduler, data storage)
- Systems designed to grow



# Server Centers

- 10,000+ servers
- Aggregated into rack (same switch)
- Reduced bandwidth between racks
- Some failure modes
  - OS crash
  - Disk dies (one of many)
  - Computer dies
  - Network switch dies (lose rack)
  - Packet loss / congestion / DNS
- Several server centers worldwide
  - Applications running distributed (e.g. ranking, mail, images)
  - Load balancing
  - Data transfer between centers



# Some indicative prices (on Amazon)

## Standard On-Demand Instances

Small (Default)	\$0.085 per hour
Large	\$0.34 per hour
Extra Large	\$0.68 per hour

## Micro On-Demand Instances

Micro	\$0.02 per hour
-------	-----------------

## Hi-Memory On-Demand Instances

Extra Large	\$0.50 per hour
Double Extra Large	\$1.00 per hour
Quadruple Extra Large	\$2.00 per hour

## Hi-CPU On-Demand Instances

Medium	\$0.17 per hour
Extra Large	\$0.68 per hour

## Cluster Compute Instances

Quadruple Extra Large	\$1.60 per hour
-----------------------	-----------------

## Cluster GPU Instances

Quadruple Extra Large	\$2.10 per hour
-----------------------	-----------------

server costs

storage

### Amazon EBS Volumes

- \$0.10 per GB-month of provisioned storage
- \$0.10 per 1 million I/O requests

### Amazon EBS Snapshots to Amazon S3

- \$0.15 per GB-month of data stored
- \$0.01 per 1,000 PUT requests (when saving a snapshot)
- \$0.01 per 10,000 GET requests (when loading a snapshot)

data transfer

### Data Transfer IN

All data transfer in	\$0.100 per GB
----------------------	----------------

### Data Transfer OUT

First 1 GB / month	\$0.000 per GB
Up to 10 TB / month	\$0.150 per GB
Next 40 TB / month	\$0.110 per GB
Next 100 TB / month	\$0.090 per GB
Greater than 150 TB / month	\$0.080 per GB

**Processing in the cloud**



# Data storage

- Billions of webpages
- Billions of queries in query / click log
- Millions of data ranked by editors
- **Storing data is cheap**
  - less than 100 Billion interesting webpages
  - assume 10kB per webpage - **1PB total**
  - Amazon S3 price (1 month) \$10k (at \$0.01/GB)
- **Processing data is expensive**
  - 10 Billion webpages
  - 10ms per page (only simple algorithms work)
  - 10 days on 1000 computers (\$24k-\$240k at \$0.1-\$1/h)
- **Crawling the data is very expensive**
  - Assume 10 Gb/s link - takes >100 days to gather  
APCN2 cable has 2.5 Tb/s bandwidth
  - Amazon EC2 price \$100k (at \$0.1/GB), with overhead \$1M)



# File Systems (GoogleFS/HDFS)

name node  
server  
(replicated)



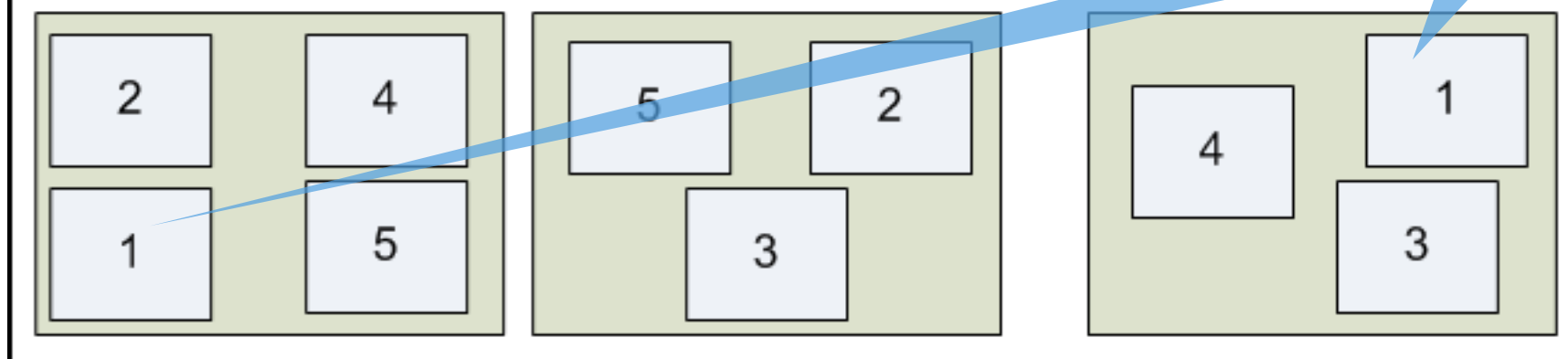
not quite so cheap

NameNode:  
Stores metadata only

METADATA:  
/user/aaron/foo → 1, 2, 4  
/user/aaron/bar → 3, 5

replicate  
blocks 3x

DataNodes: Store blocks from files



cheap  
servers

Ghemawat, Gobioff, Leung, 2003

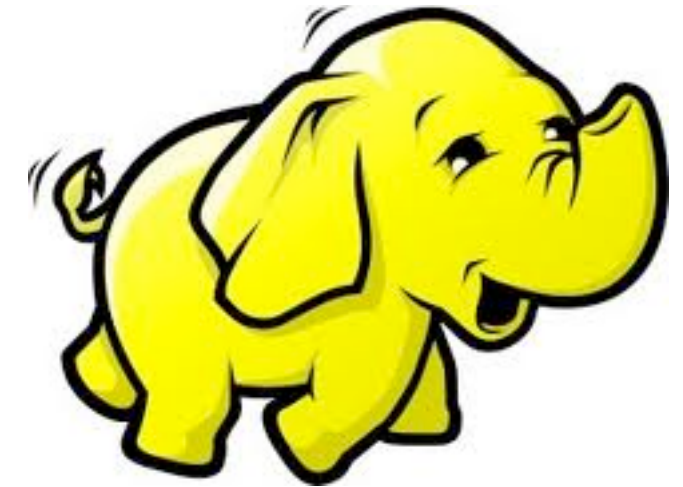


# File Systems Details

- Chunkservers
  - store 64MB (or larger) blocks
  - write to one, replicate transparently (3x)
- Name node
  - keeps directory of chunks
  - replicated
- Write
  - get chunkserver ID from name node
  - write block(s) to chunkserver (expensive to write parts of chunk)
  - largely write once scenario (read many times)
  - distributed write for higher bandwidth (each server 1 chunk)
- Read
  - from any of the replicated chunks
  - higher replication for hotspots (many reads of a chunk)
- Elasticity
  - Add additional chunkservers - name node migrates chunks to new machine
  - Node failure (name node keeps track of chunk server status) - requests replication

# Comparison

- **HDFS/GoogleFS**
  - Fast block writes / reads
  - Fault tolerant
  - Flexible size adjustment
  - Terrible for random writes / bad for random writes
  - Not really a filesystem
- **NFS & co.**
  - No distributed file management
- **Lustre**
  - Proper filesystem
  - High bandwidth
  - Explicitly exploits fast interconnects
  - Cabinet servers replicated with RAID5
  - **Fails if cabinet dies**
  - Difficult to add more storage



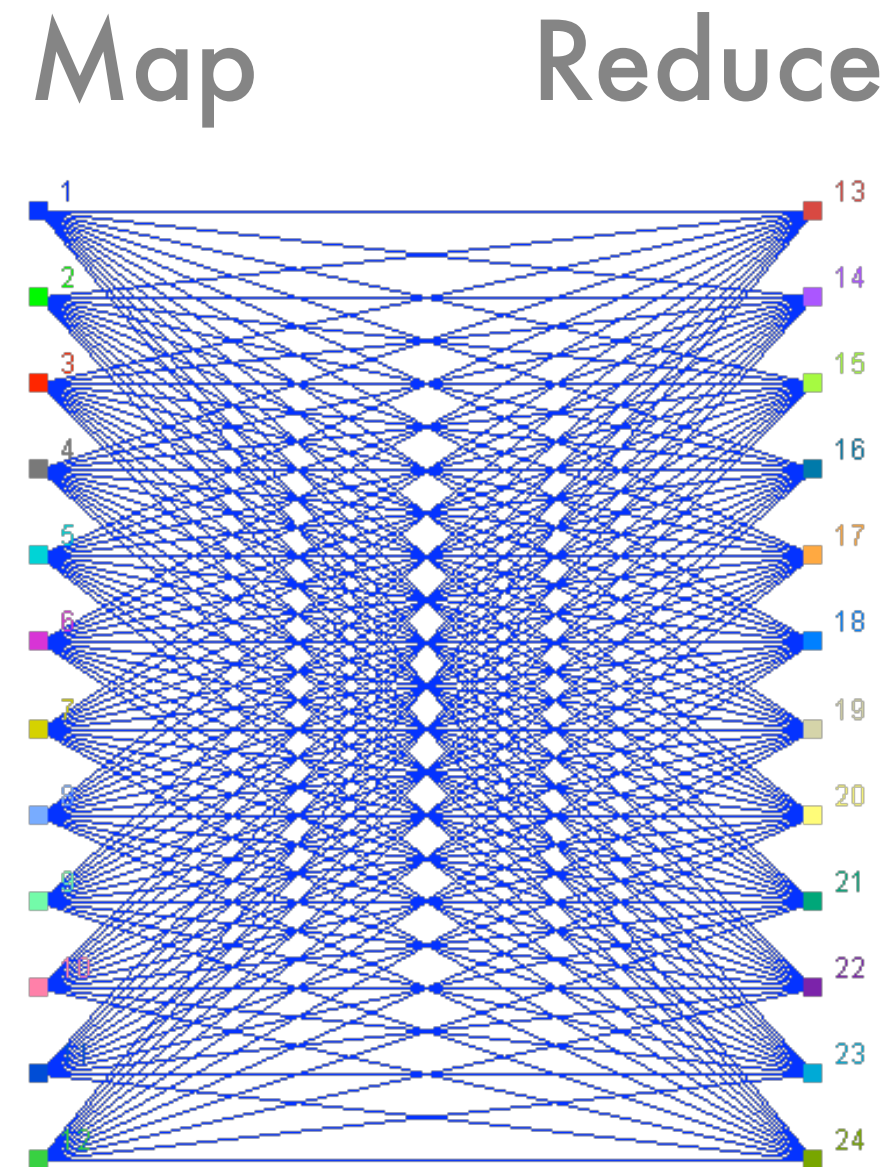
lustre

# MapReduce

- **Scenario**
  - Lots of data (much more than what a single computer can process)
  - Stored in a distributed fashion
  - Move computation to data
- **Map**

Apply function  $f$  to data as distributed over the mappers  
This runs (if possible) where data sits
- **Reduce**

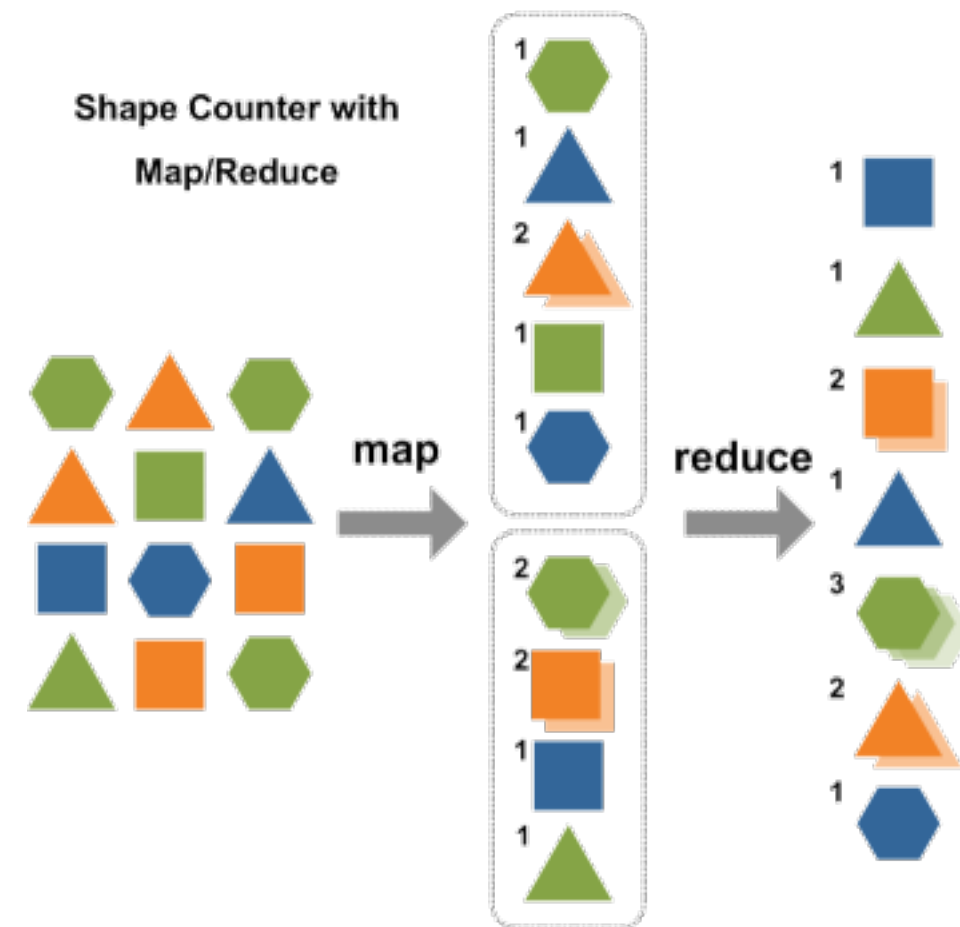
Combine data given keys generated in the Map phase
- **Fault tolerance**
  - If mapper dies, re-process the data
  - If reducer dies, re-send the data (cached) from mappers (requires considerable storage)



Dean, Ghemawat, 2004

# Item count in MapReduce

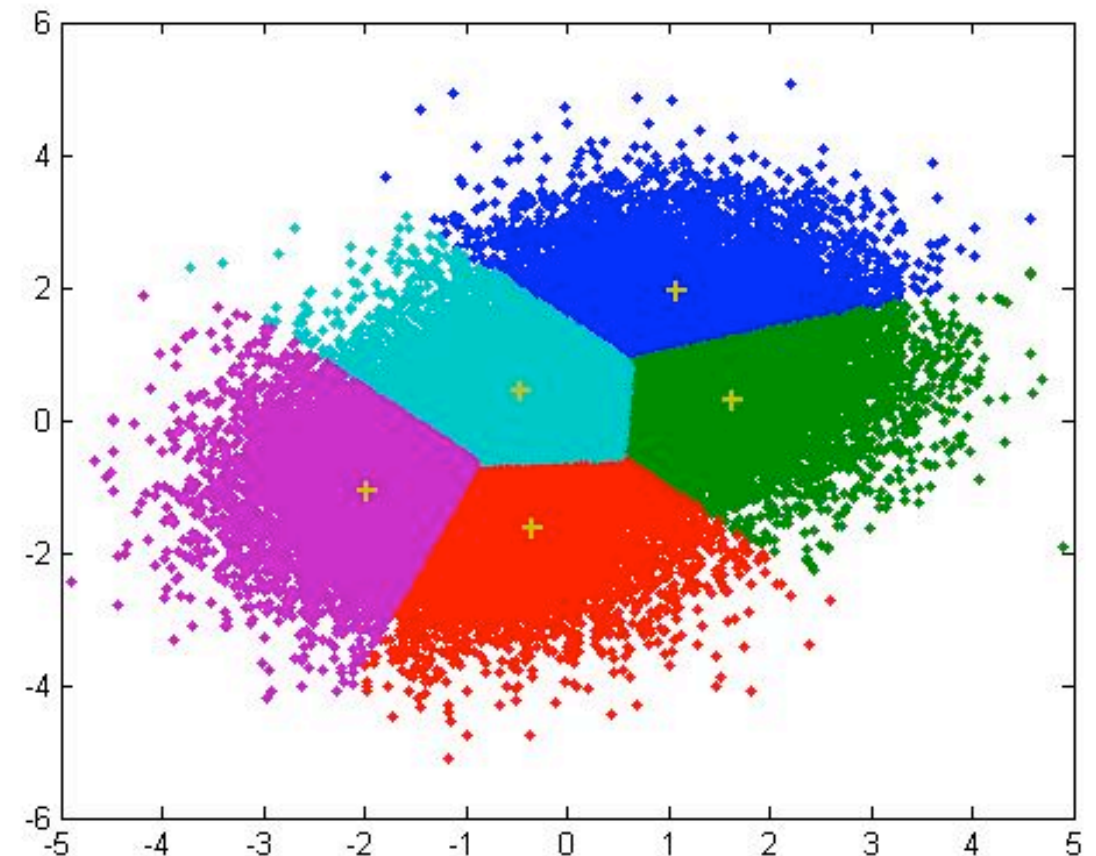
- Task: object counting
- Map
  - Each mapper gets (unsorted) chunk of data
  - Preferably local on chunkservers
  - Perform local item counts
  - Emit (item, count) data
- Reduce
  - Aggregate all counts for a given item (all end up at same reducer)
  - Emit aggregate counts



(image: gridgain.com)

# k-means in MapReduce

- Initialize random cluster centers
- Map
  - Assign each data point to a cluster based on current model
  - Aggregate data per cluster
  - Send cluster aggregates to reducer (e.g. mean, variance, size)
- Reduce
  - Aggregate all data per cluster
  - Update cluster centers
  - Send new cluster centers to new reducers
- Repeat until converged

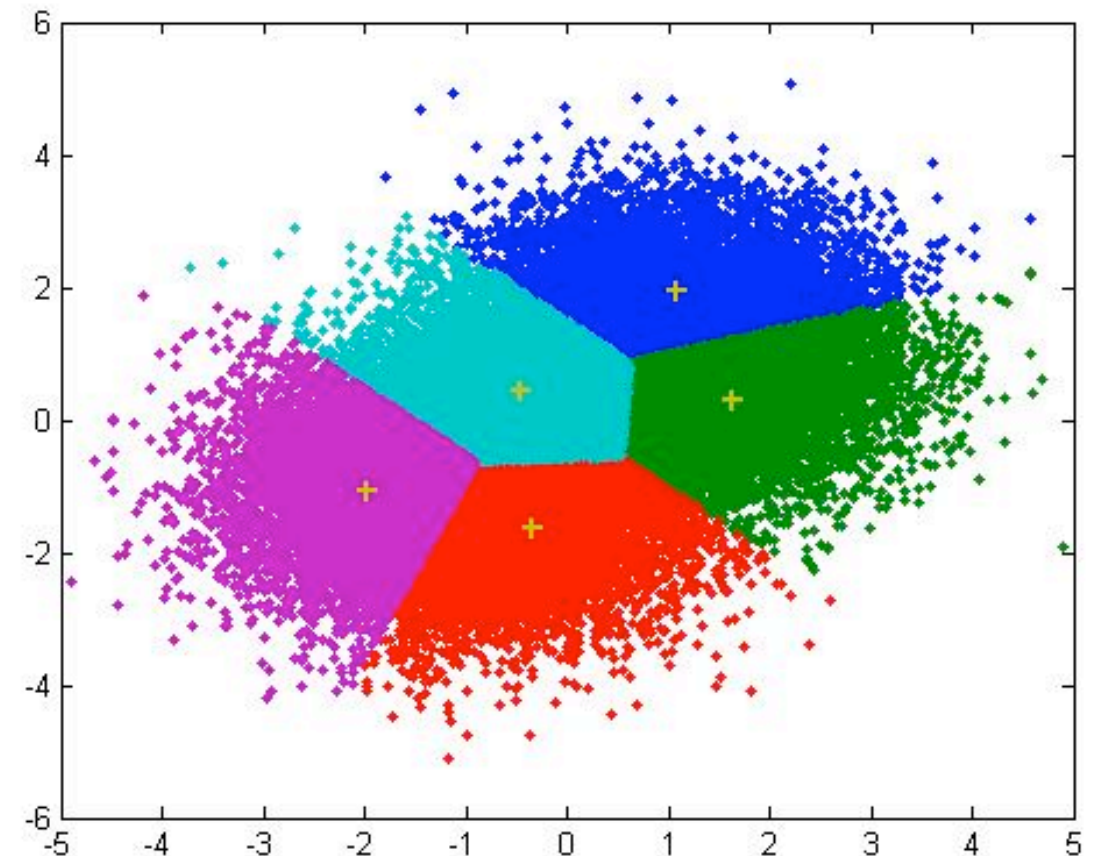


(image: mathworks.com)



# k-means in MapReduce

- Initialize random cluster centers
- Map
  - Assign each data point to a cluster based on current model
  - Aggregate data per cluster
  - Send cluster aggregates to reducer (e.g. mean, variance, size)
- Reduce
  - Aggregate all data per cluster
  - Update cluster centers
  - Send new cluster centers to new reducers
- Repeat until converged

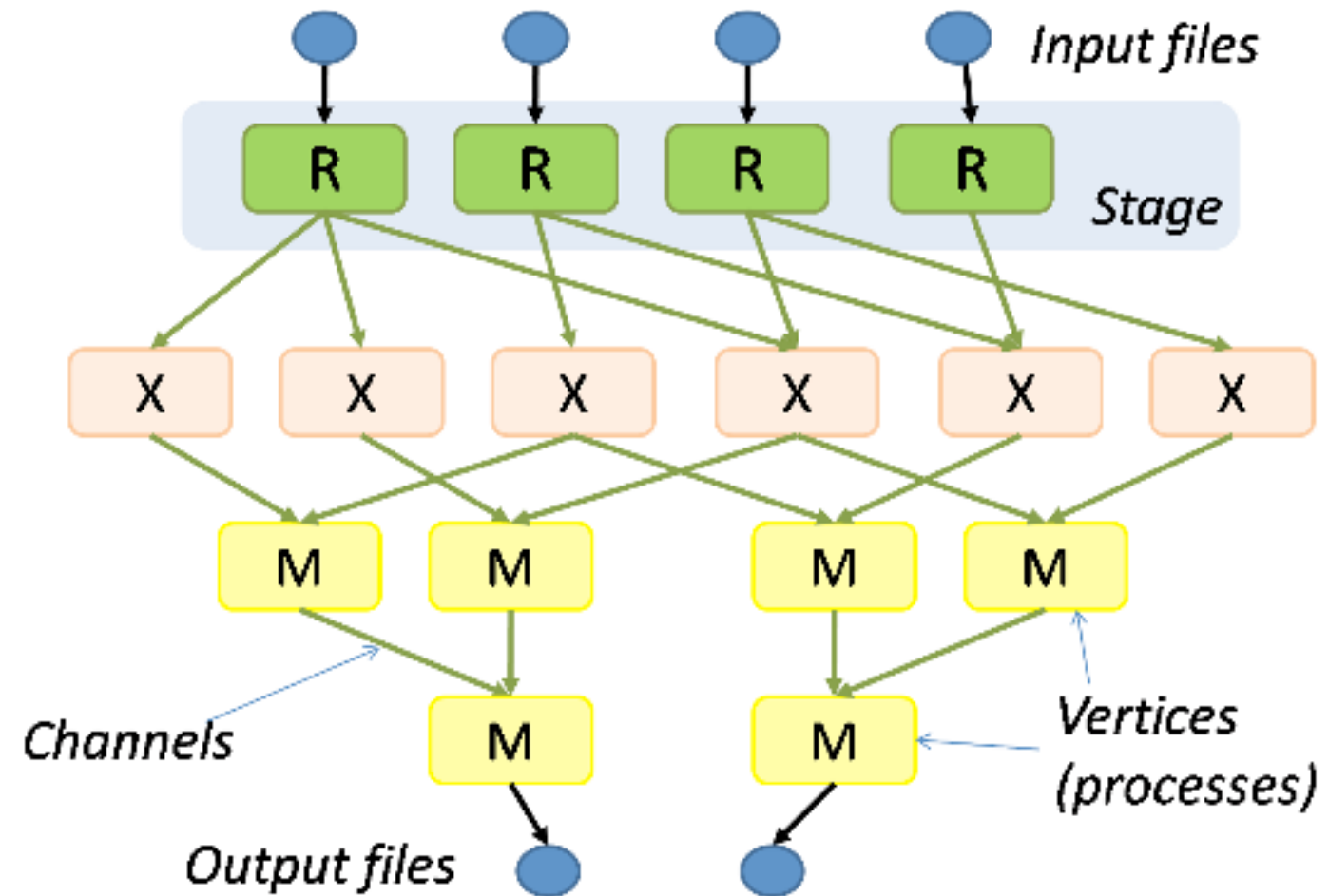


(image: mathworks.com)

needs to re-read data from disk  
for each MapReduce iteration

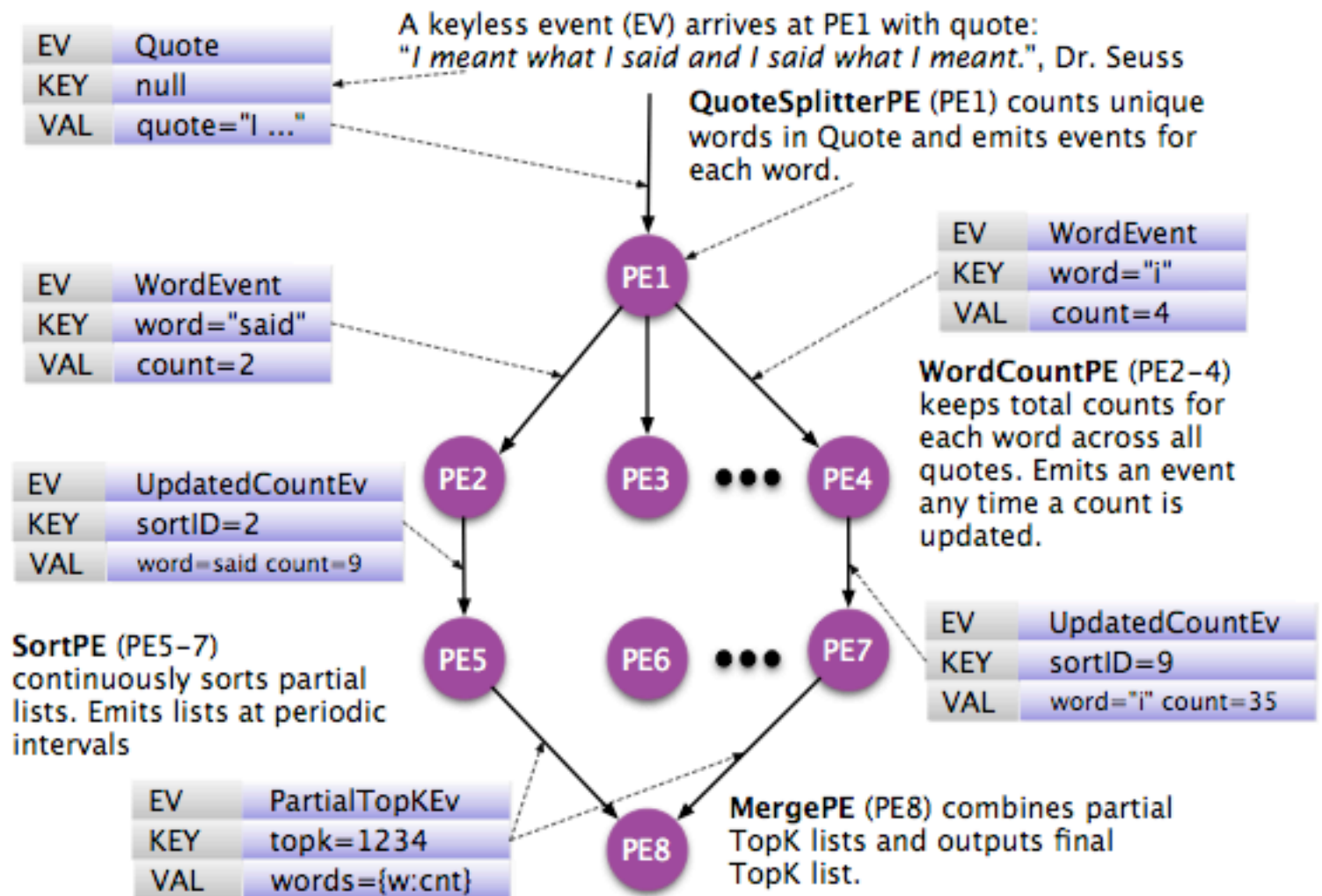
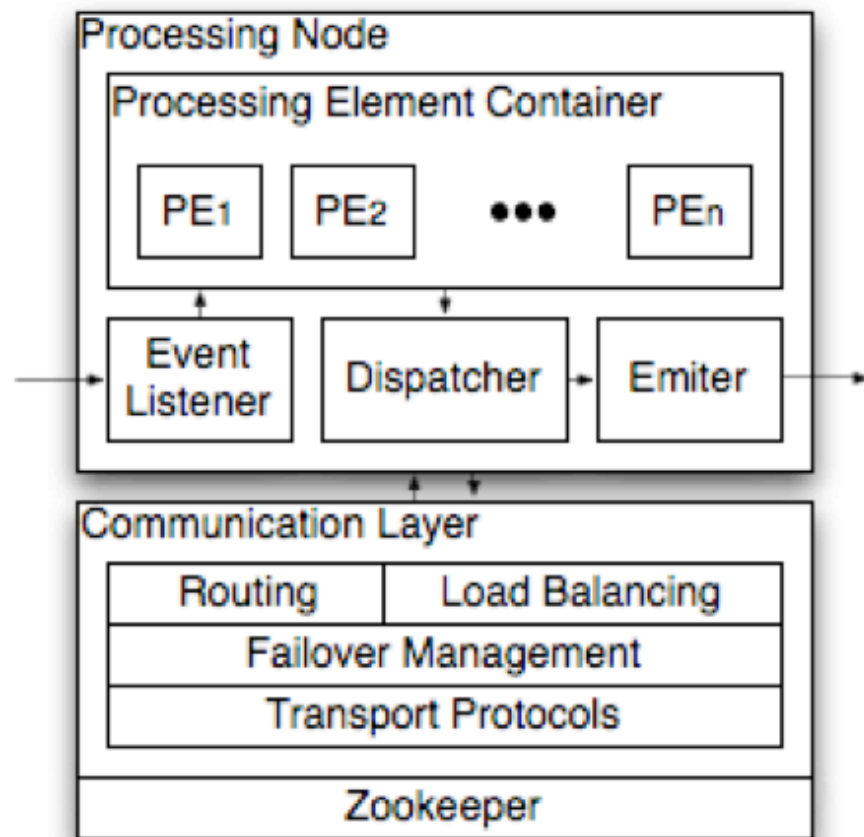
# Dryad

- Data flow graph (DAG rather than bipartite graph)
- Interface variety
  - Memory FIFO
  - Disk
  - Network
- Modular composition of computation graph



(image: Microsoft Research)

# S4 - Online processing



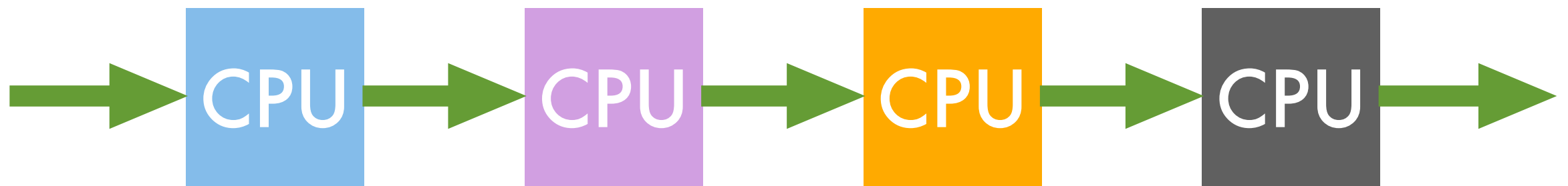
PE ID	PE Name	Key Tuple
PE1	QuoteSplitterPE	null
PE2	WordCountPE	word="said"
PE4	WordCountPE	word="i"
PE5	SortPE	sortID=2
PE7	SortPE	sortID=9
PE8	MergePE	topK=1234

dispatch with distributed hash



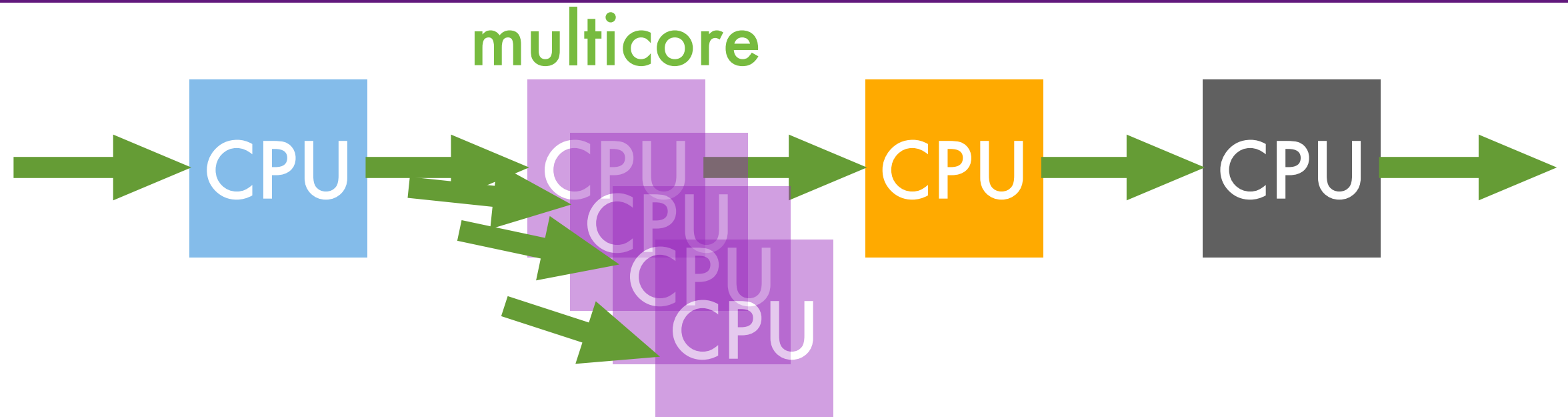
# Dataflow Paradigms

# Pipeline



- Process data sequentially
- Parallelizes up to number of tasks  
(disk read, feature extraction, logging, output)
- Reasonable for a single machine
- Parallelization per filter possible  
(see e.g. Intel Threading Building Blocks)

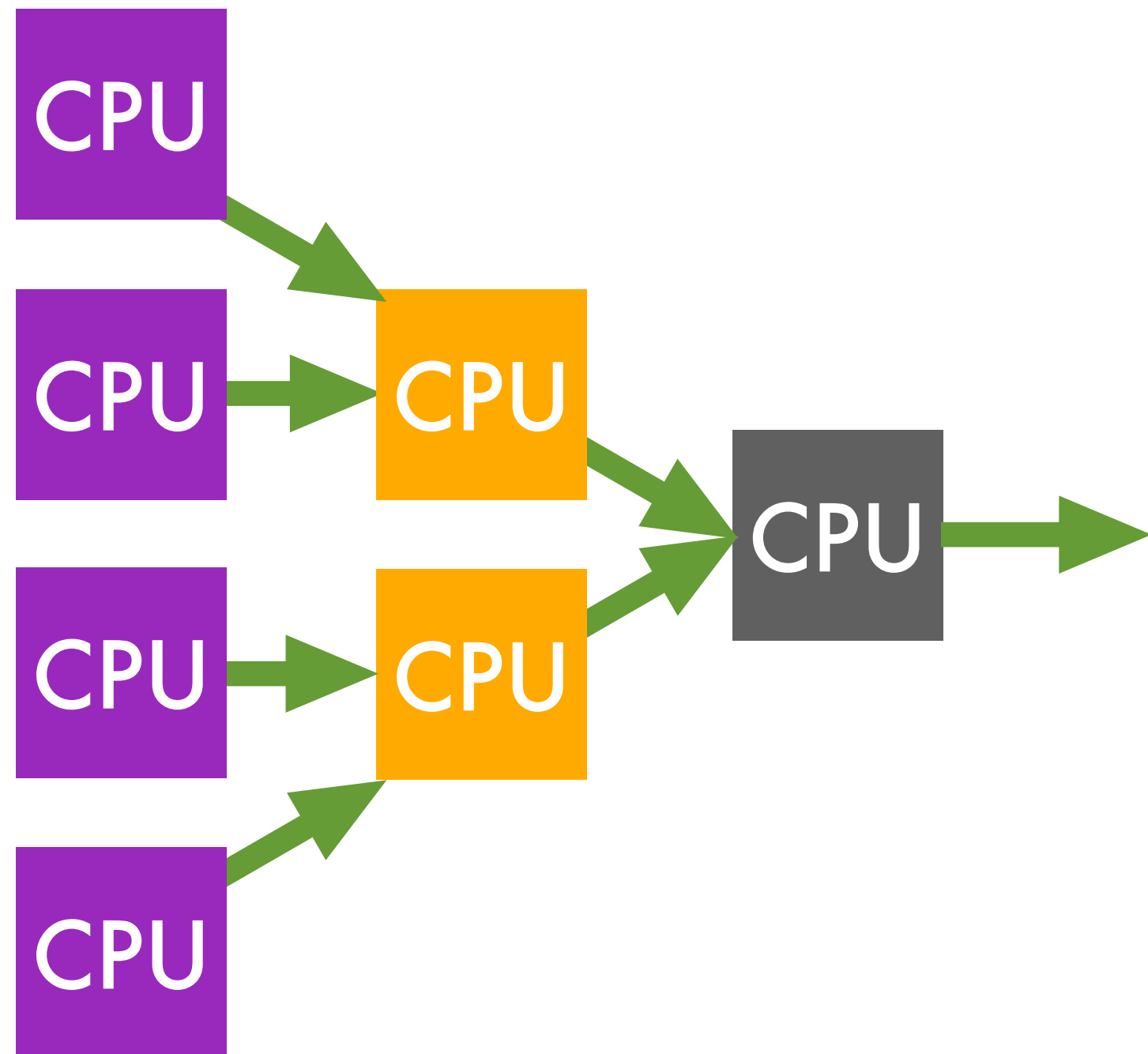
# Pipeline



- Process data sequentially
- Parallelizes up to number of tasks  
(disk read, feature extraction, logging, output)
- Reasonable for a single machine
- Parallelization per filter possible  
(see e.g. Intel Threading Building Blocks)

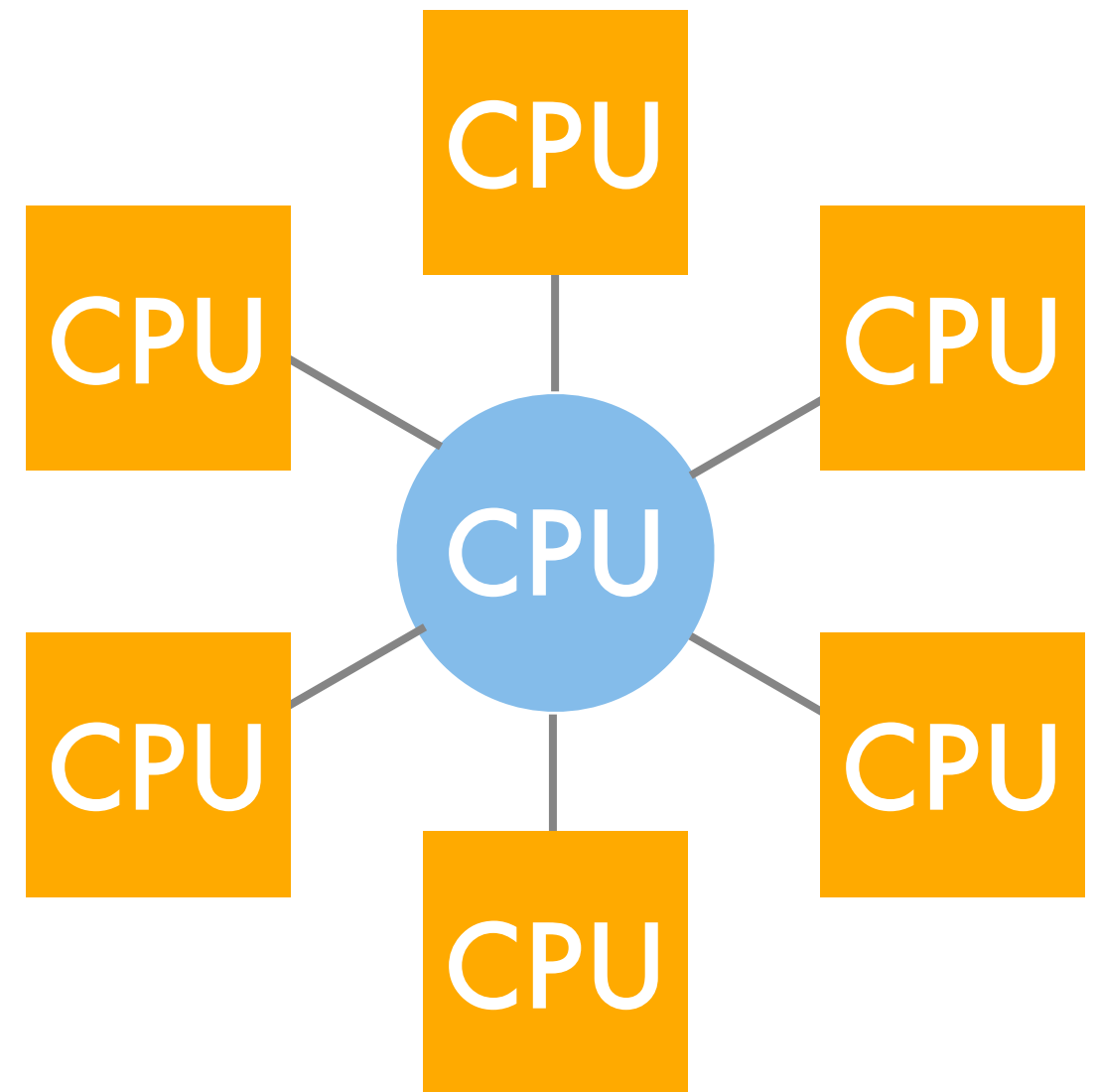
# Tree

- Aggregate data hierarchically
- Parallelizes at  $O(\log n)$  cost (tree traversal)
- Communication at the interface is important (e.g. network latency)
- Good dataflow processing
- Poor efficiency for batched processing  $O(1/\log n)$
- Poor fault tolerance
- Does not map well onto server centers (need to ensure that we have lower leaves on rack)



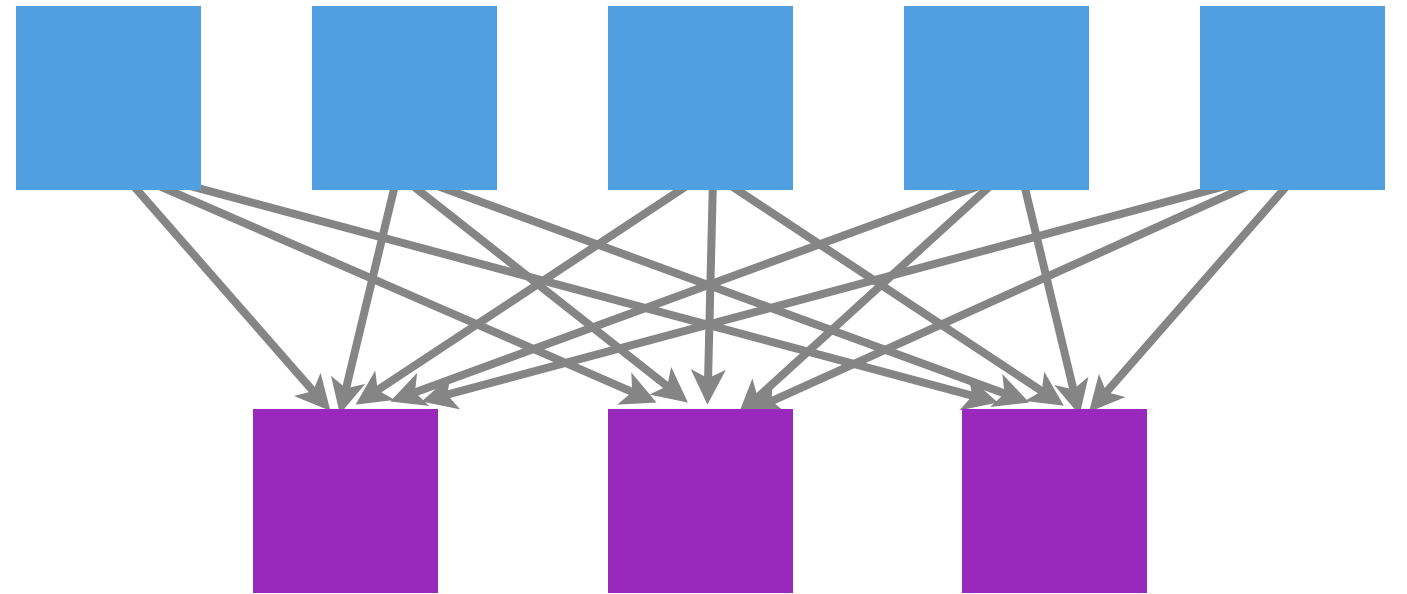
# Star

- Aggregate data centrally
- Does not parallelize at all if communication cost is high
- Perfect parallelization if CPU bound
- Latency is  $O(1)$  unless the network is congested.
- Network requirements are  $O(n)$
- Central node becomes hotspot
- Synchronization is very easy
- Trivial to add more resources (just add leaves)
- Difficult to parallelize center



# Distributed (key,value) storage

- Caching problem
  - Store many (key,value) pairs
  - Linear scalability in clients and servers
  - Automatic key distribution mechanism
- memcached
  - (key,value) servers
  - client access library distributes access patterns
  - randomized  $O(n)$  bandwidth
  - aggregate  $O(n)$  bandwidth
  - load balancing via hashing
  - no versioned writes

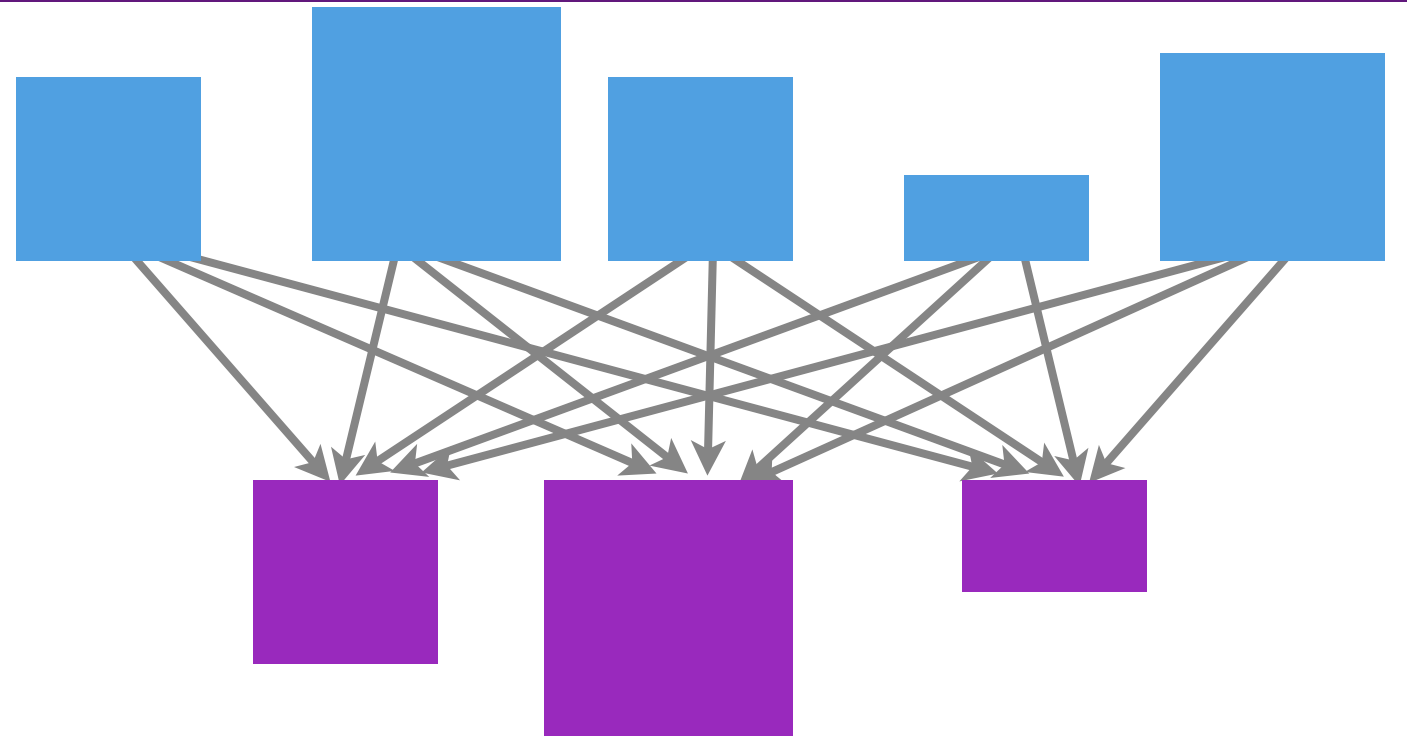


$$m(\text{key}, M) = \underset{m \in M}{\operatorname{argmin}} h(\text{key}, m)$$

P2P uses similar routing tables

# Proportional hashing/caching

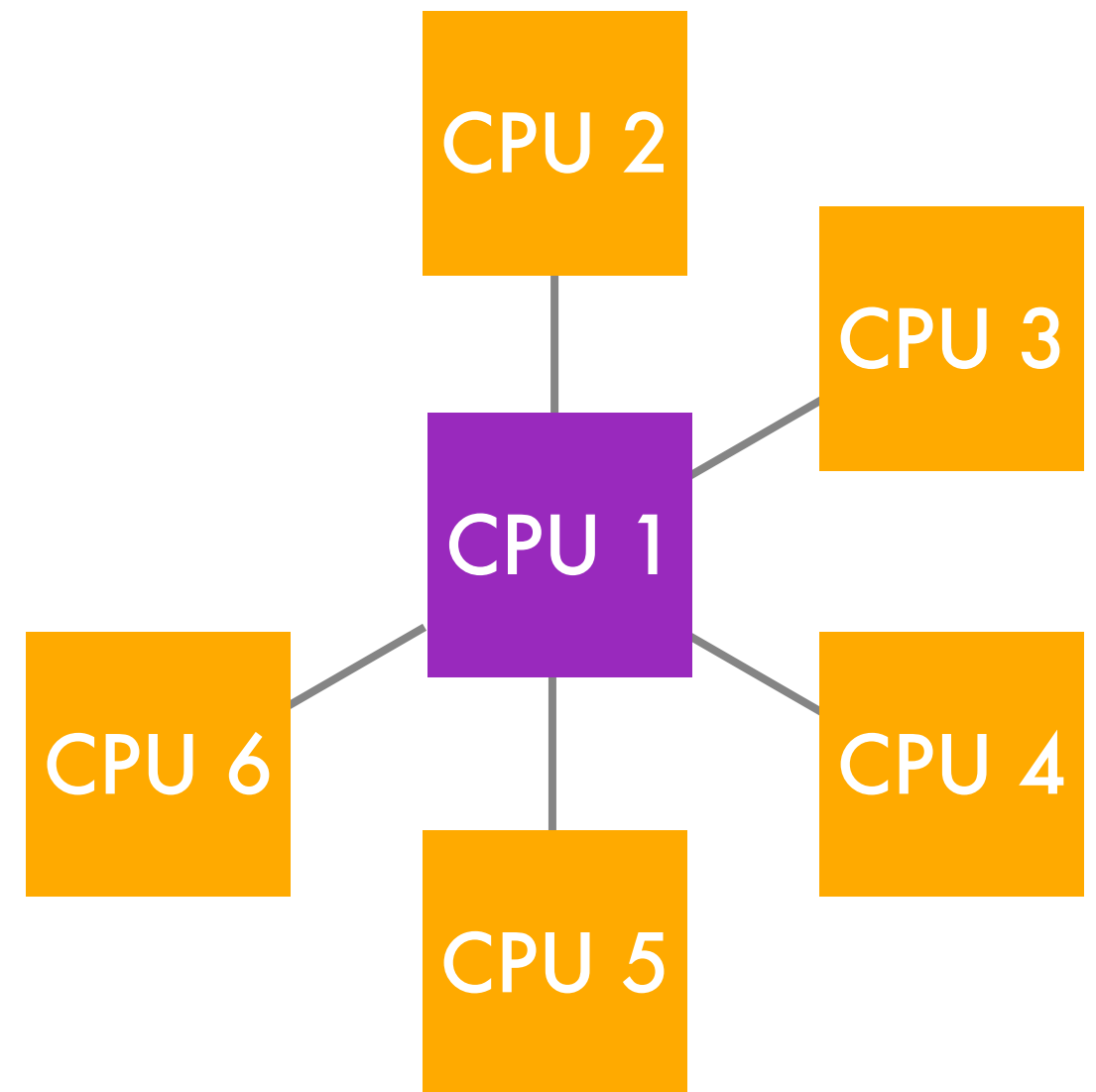
- Machines with different capacity
- Hotspots too hot for a single machine to handle
- Retain nearest neighbor hashing properties



- Sparse cover of keyspace proportional to machine capacities
- Repeatedly hash a key until it hits a key-range, i.e.  $h^n(\text{key})$
- Keep busy table and advance to next hash if machine already used

# Distributed Star

- Aggregate data centrally
- Use different center for each key, as selected by distributed hashing
- Linear bandwidth for synchronization
- Perfect scalability  $O(n)$  bandwidth required
- Each CPU performs local computation and stores small fraction of global data
- Works best if all nodes on the same switch / rack

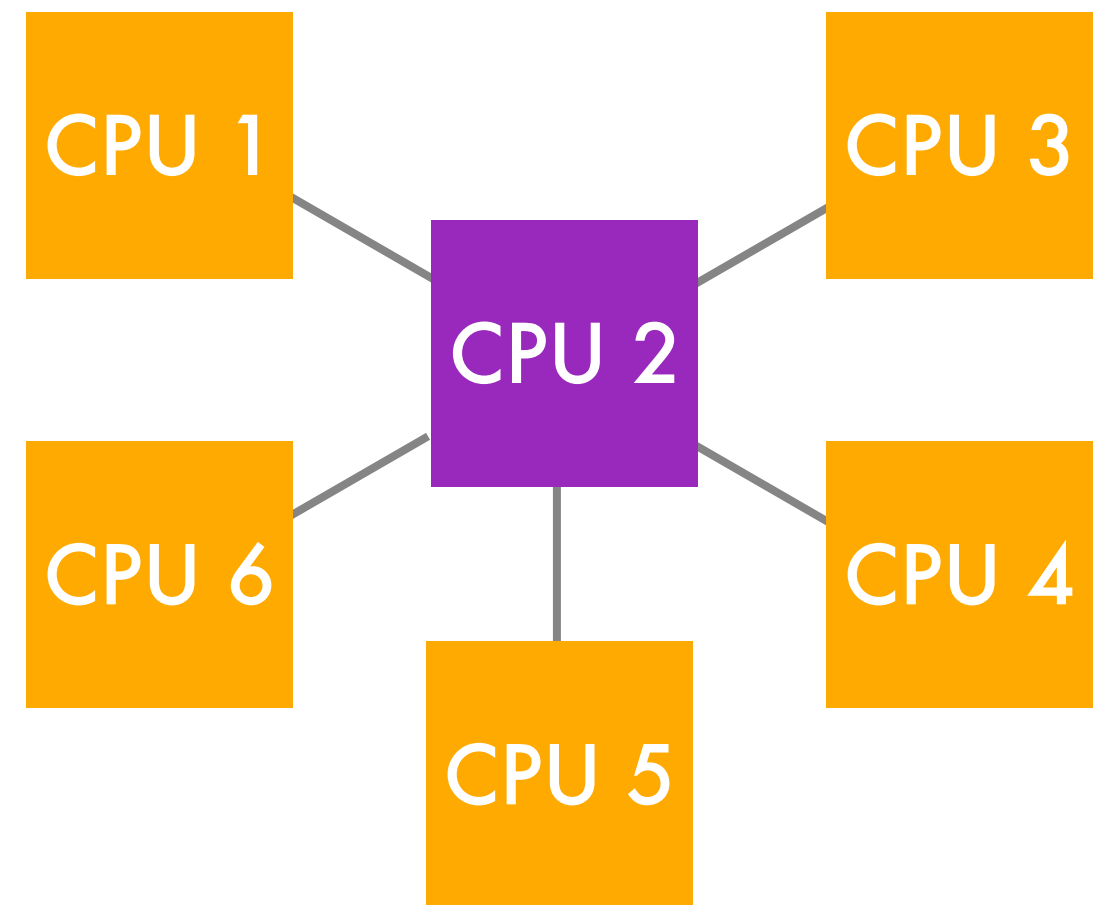


$$m(\text{key}, M) = \underset{m \in M}{\operatorname{argmin}} h(\text{key}, m)$$



# Distributed Star

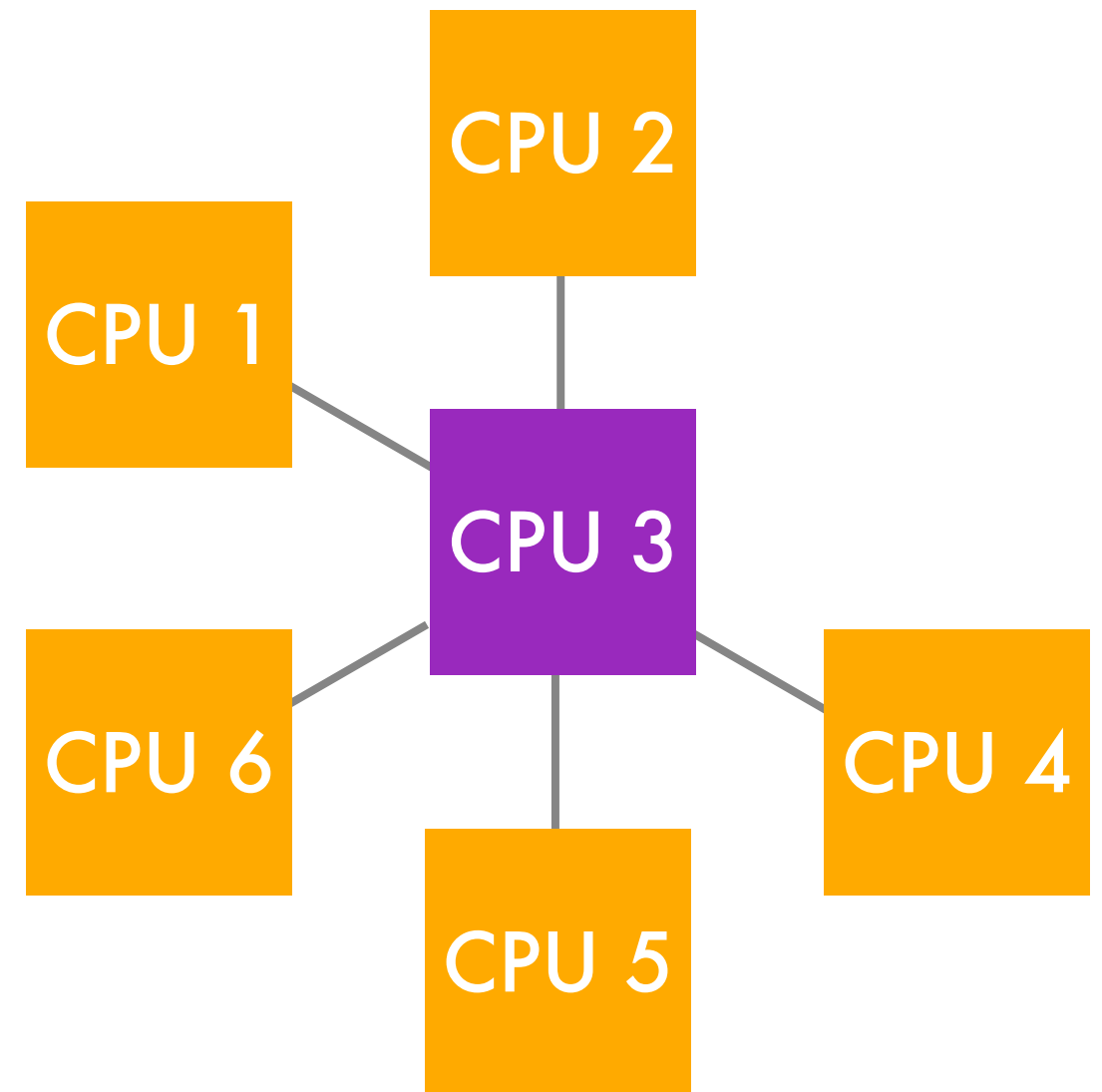
- Aggregate data centrally
- Use different center for each key, as selected by distributed hashing
- Linear bandwidth for synchronization
- Perfect scalability  $O(n)$  bandwidth required
- Each CPU performs local computation and stores small fraction of global data
- Works best if all nodes on the same switch / rack



$$m(\text{key}, M) = \underset{m \in M}{\operatorname{argmin}} h(\text{key}, m)$$

# Distributed Star

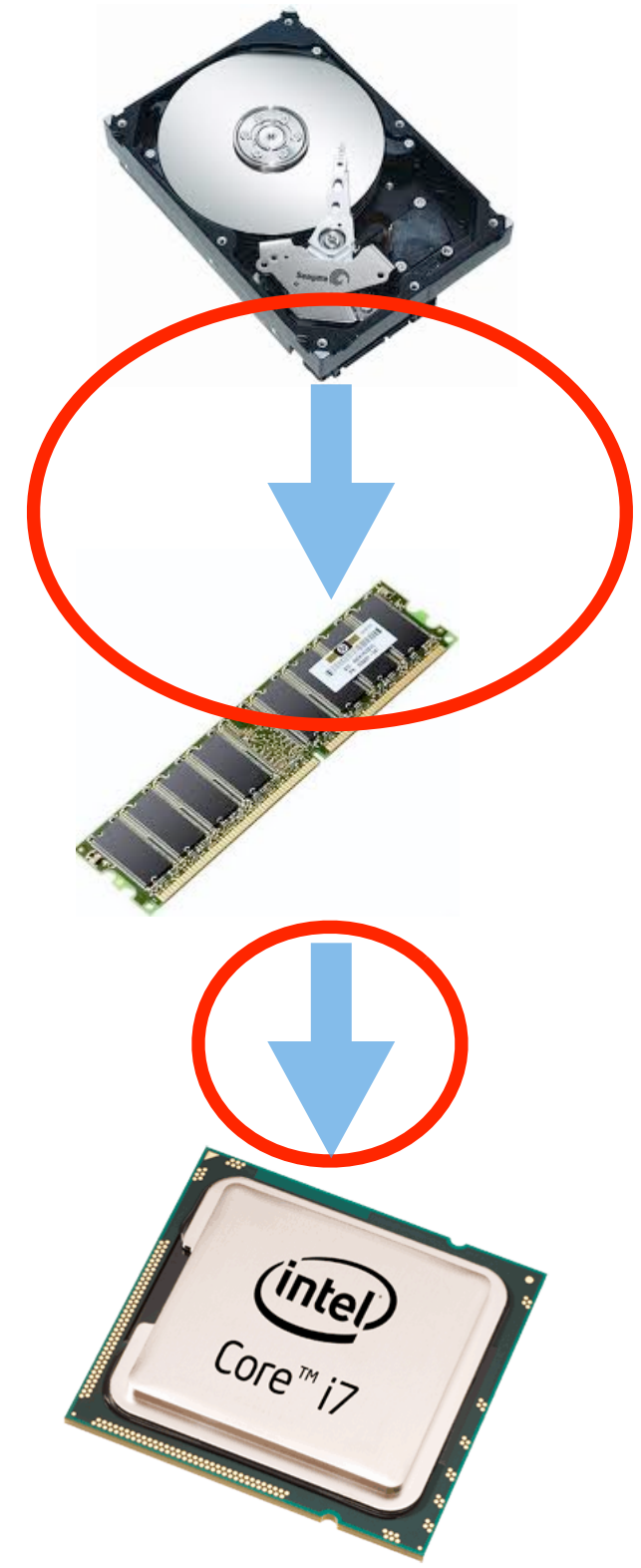
- Aggregate data centrally
- Use different center for each key, as selected by distributed hashing
- Linear bandwidth for synchronization
- Perfect scalability  $O(n)$  bandwidth required
- Each CPU performs local computation and stores small fraction of global data
- Works best if all nodes on the same switch / rack



$$m(\text{key}, M) = \underset{m \in M}{\operatorname{argmin}} h(\text{key}, m)$$

# Ringbuffer

- Problem
  - Disk, RAM and CPU operate at different speeds ( $>10x$  difference)
  - Want to do maximum data processing (e.g. optimization)
- Idea
  - Load data from disk into ringbuffer
  - Process data continuously on buffer
  - Chain ringbuffers
- Yields consistently **maximum throughput for each resource**



# Summary

- **Hardware**  
**Servers, networks, amounts of data**
- **Processing paradigms**  
**MapReduce, Dryad, S4**
- **Communication templates**  
**Stars, pipelines, distributed hash table, caching**

# Part 2 - Motivation



**MOTIVATION**

if a pretty poster and a cute saying are all it takes to motivate you  
you probably have a very easy job. the kind robots will be doing soon.

# Data on the Internet

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

## Finite resources

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

**Invest in modeling and naming rather than data generation**

# Data on the Internet

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

## Finite resources

- Editors are expensive
- Editors / know users

unlimited amounts  
of data

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

Invest in modeling and naming  
rather than data generation

# Unsupervised Modeling



# Hierarchical Clustering



NIPS 2010  
Adams,  
Ghahramani,  
Jordan

# Topics in text

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

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



# Word segmentation

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



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

Mochihashi, Yamada, Ueda, ACL 2009

# Language model

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

---

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

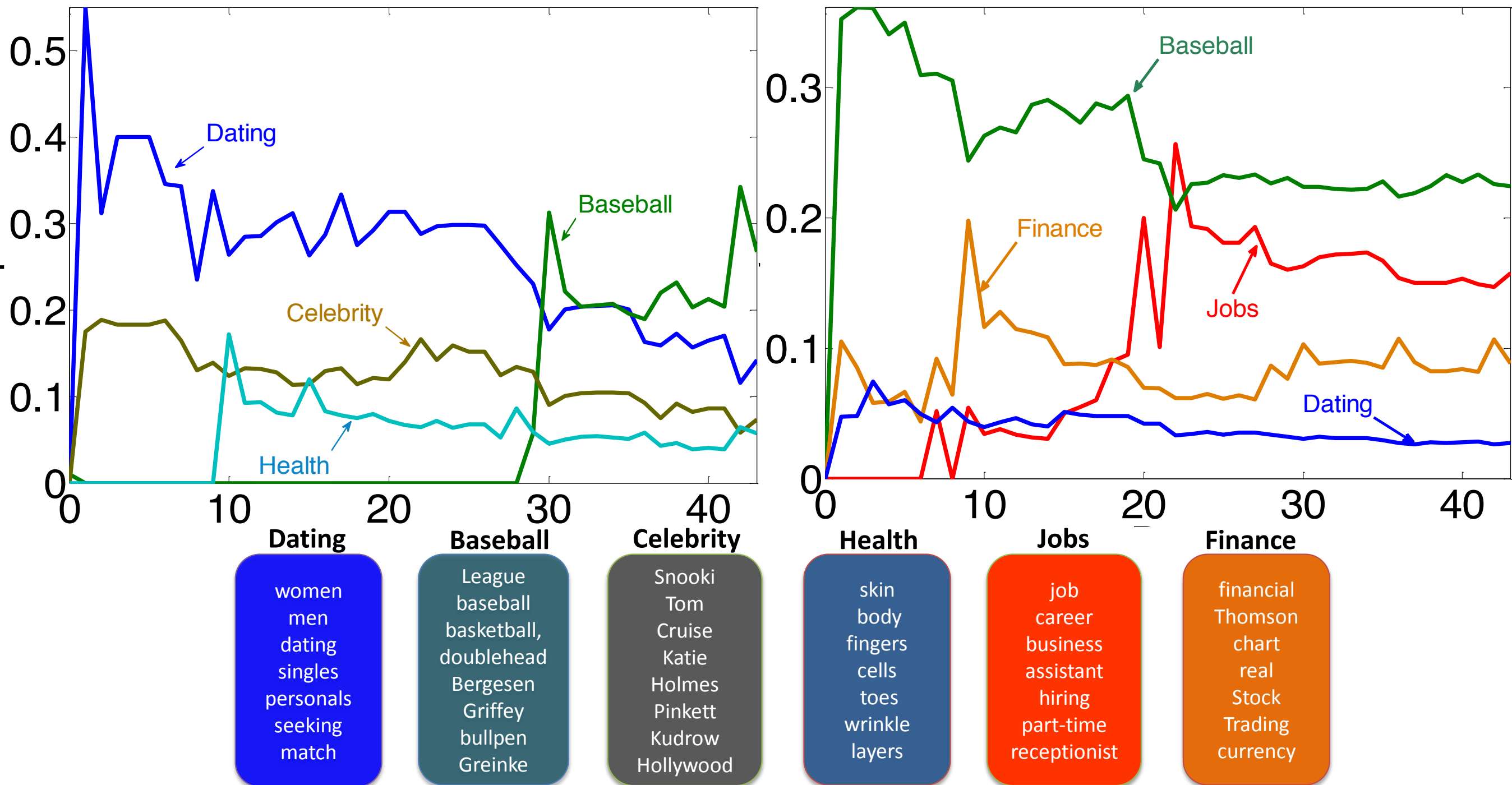
---

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

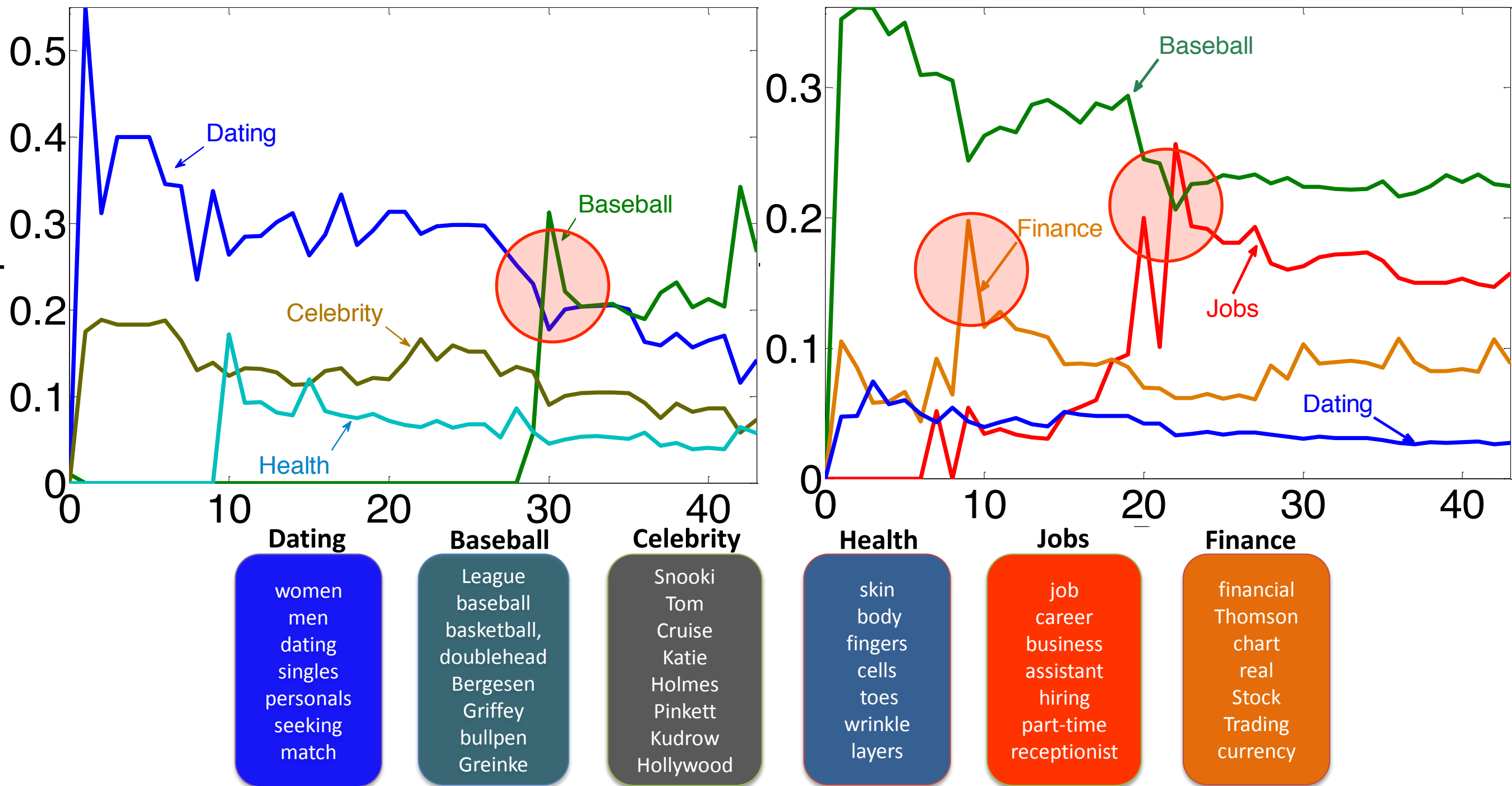
**automatically** synthesized  
from Penn Treebank

Mochihashi, Yamada, Ueda  
ACL 2009

# User model over time



# User model over time





# Face recognition from captions



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



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



# Storylines from news

TOPICS

## Sports

games  
won  
team  
final  
season  
league  
held

## Politics

government  
minister  
authorities  
opposition  
officials  
leaders  
group

## Unrest

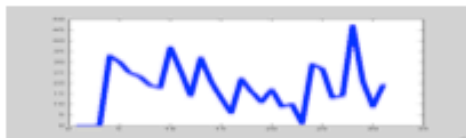
police  
attack  
run  
man  
group  
arrested  
move

Ahmed et al,  
AISTATS 2011

STORYLINES

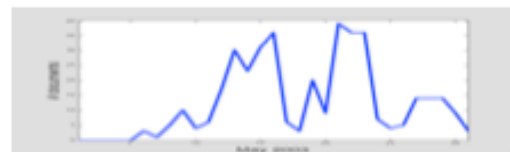
## UEFA-soccer

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



## Tax bills

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



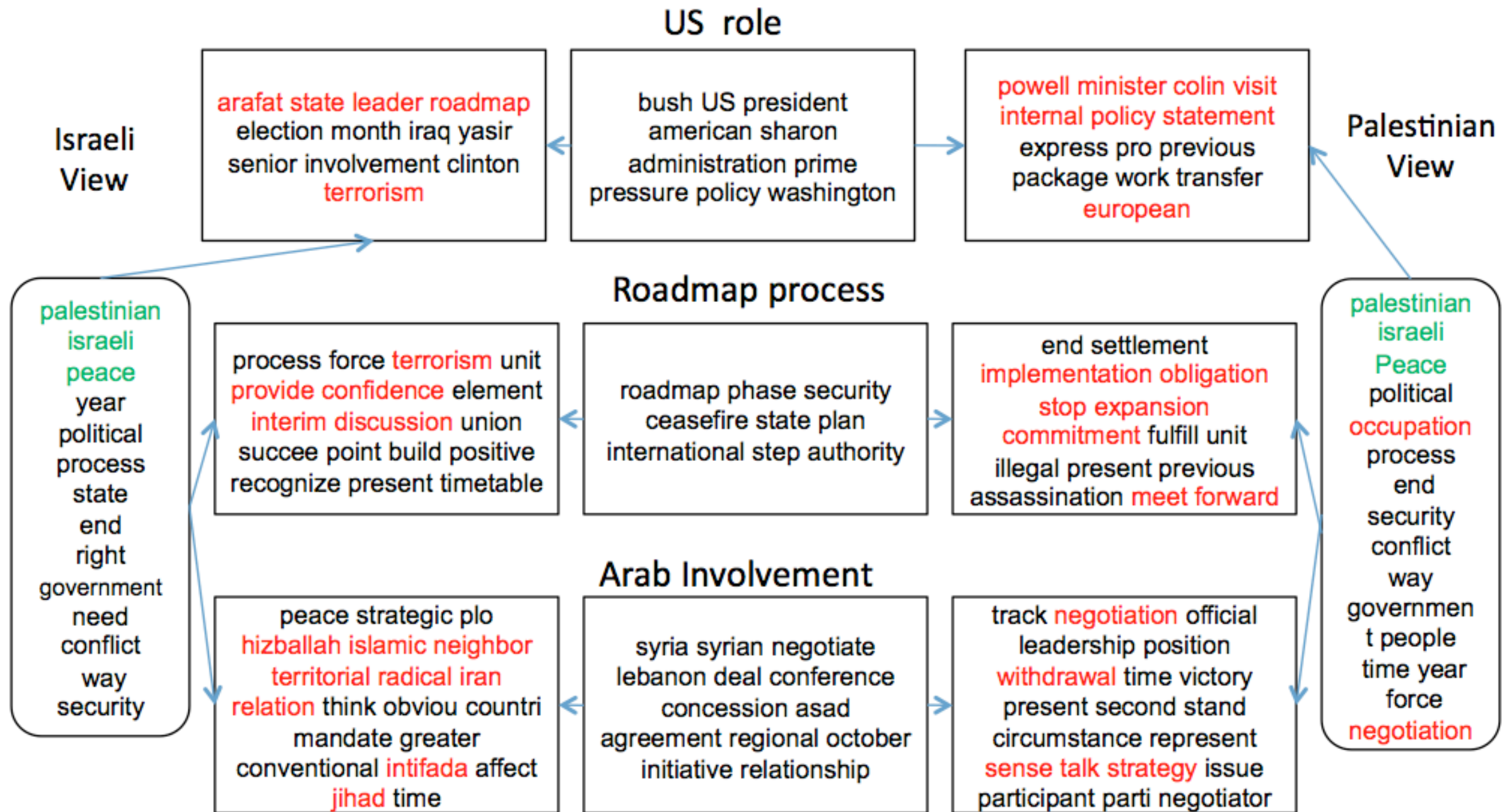
## India-Pakistan tension

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







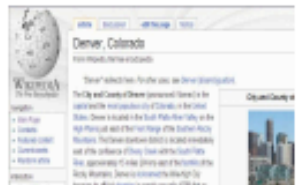





# Ideology detection



Ahmed et al, 2010; Bitterlemons collection

# Hypertext topic extraction

Topic 1			Topic 2		
neural	0.067	Artificial neural network  0.004	recognition	0.058	Speech recognition  0.004
network	0.047		speech	0.033	
networks	0.039	Neural network  0.003	language	0.015	Pattern recognition  0.004
learning	0.027		pattern	0.012	
artificial	0.017		handwriting	0.011	
data	0.015		evaluation	0.010	
models	0.014		robots	0.010	
function	0.014		systems	0.009	
Topic 3			Topic 4		
vancouver	0.051	Denver, Colorado  0.0008	brain	0.047	Cognitive science 0.003 
denver	0.043			cognitive	
city	0.041	Vancouver  0.0002	science	0.016	Neuroscience 0.002 
retrieved	0.024			press	
colorado	0.011		neurons	0.010	
area	0.009		mind	0.010	
population	0.009		systems	0.010	
canada	0.008		human	0.010	

Gruber, Rosen-Zvi, Weiss; UAI 2008

# Supervised Modeling

# Ontologies


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

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

[advanced](#)

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


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



Copyright © 2011 Netscape

- continuous maintenance
- no guarantee of coverage
- difficult categories


# Ontologies

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

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

[advanced](#)

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

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

Copyright © 2011 Netscape

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

- continuous maintenance
- no guarantee of coverage
- difficult categories



# Face Classification/Recognition



**Iranian Face Database**  
Pose, Age and Expression

|| HOME ||

**Sampels**

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

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

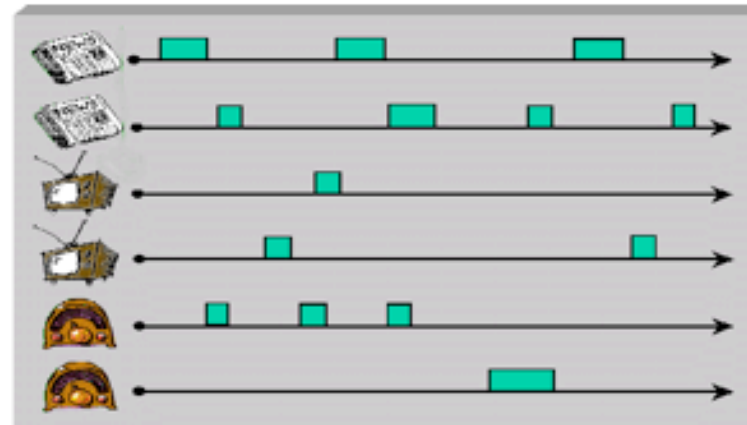
# Topic Detection & Tracking

Information Technology Laboratory

Information Access Division (IAD)

**NIST**  
National Institute of  
Standards and Technology

## Topic Detection and Tracking Evaluation



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

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

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

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

• [Multimodal Information Group Home](#)

• [Benchmark Tests](#)

• [Tools](#)

• [Test Beds](#)

• [Publications](#)

• [Links](#)

• [Contacts](#)

# Advertising Targeting

## Browse Ad Solutions

### Media Spotlight

#### AUDIENCE

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

#### Your categories

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

#### Category

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

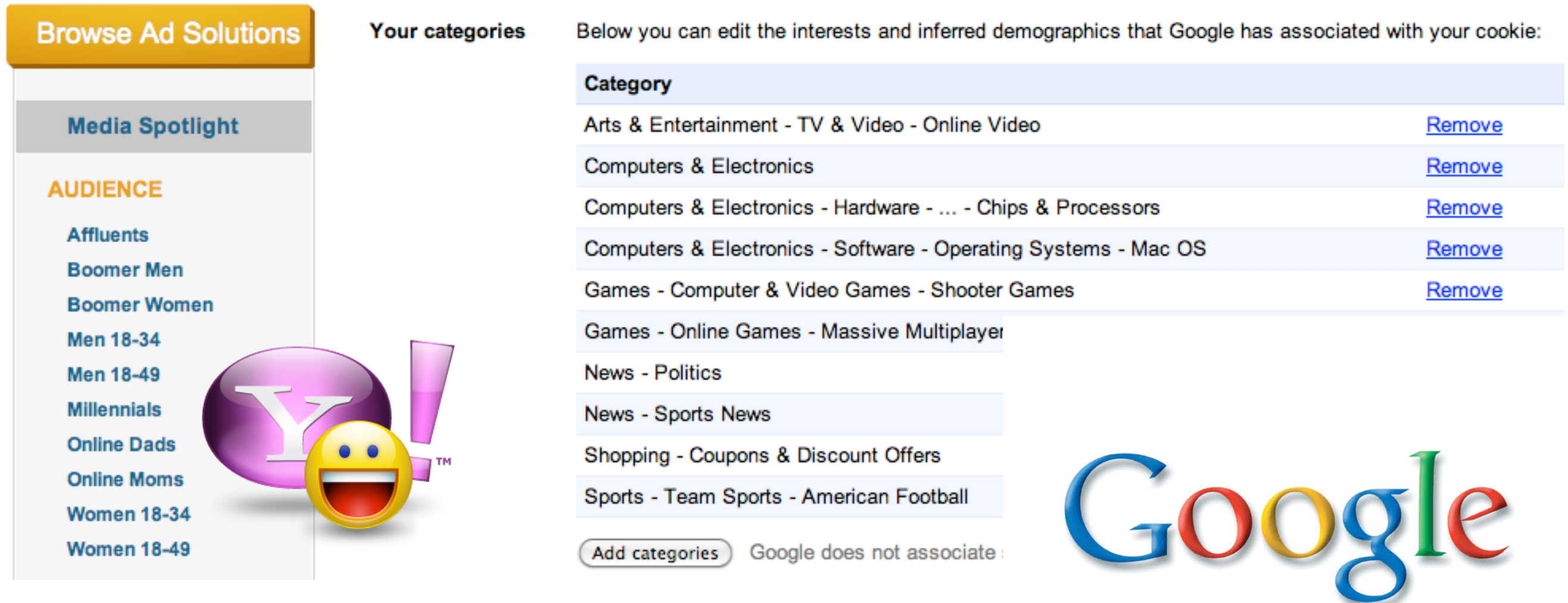
[Add categories](#)

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

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



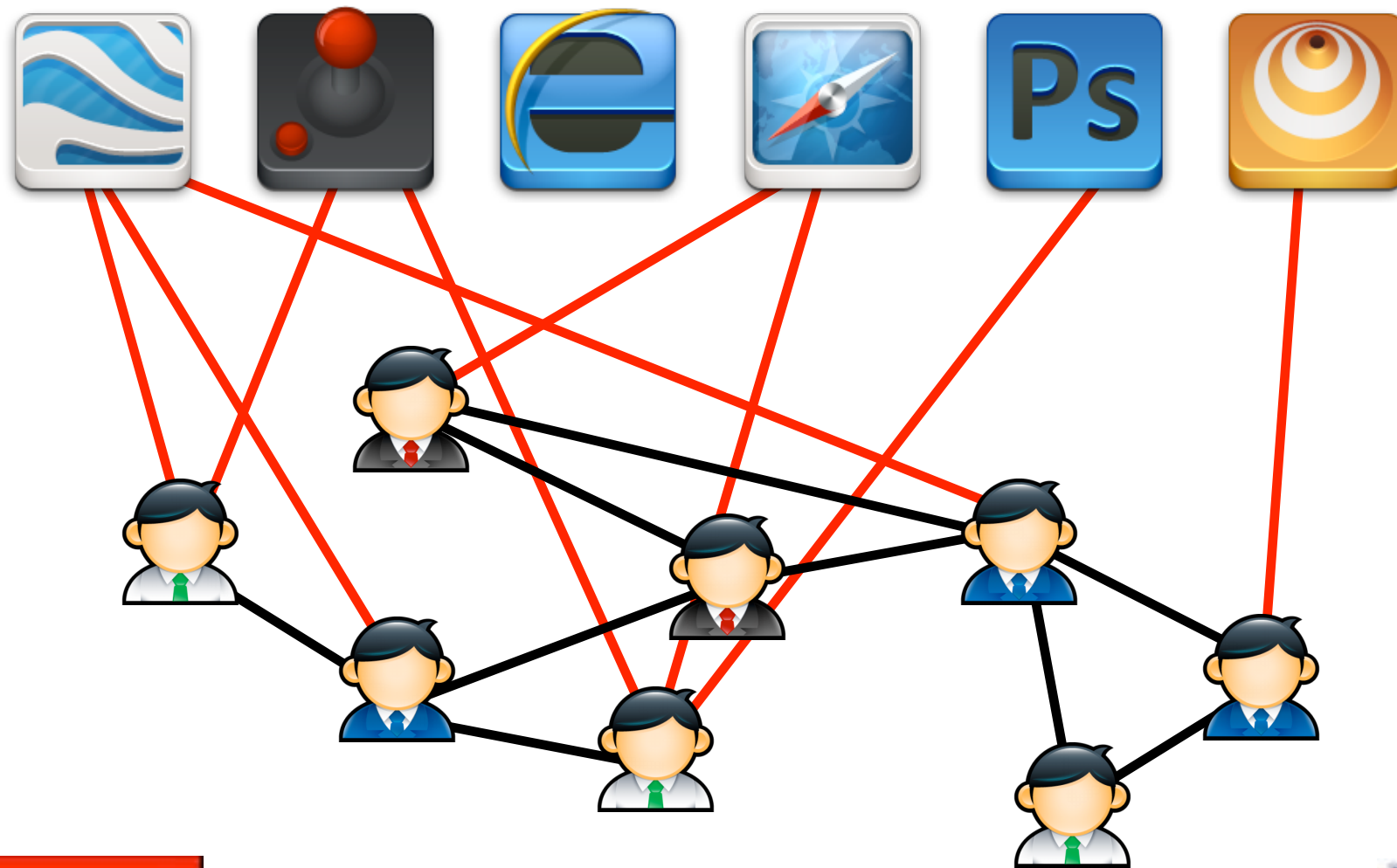
# Advertising Targeting



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

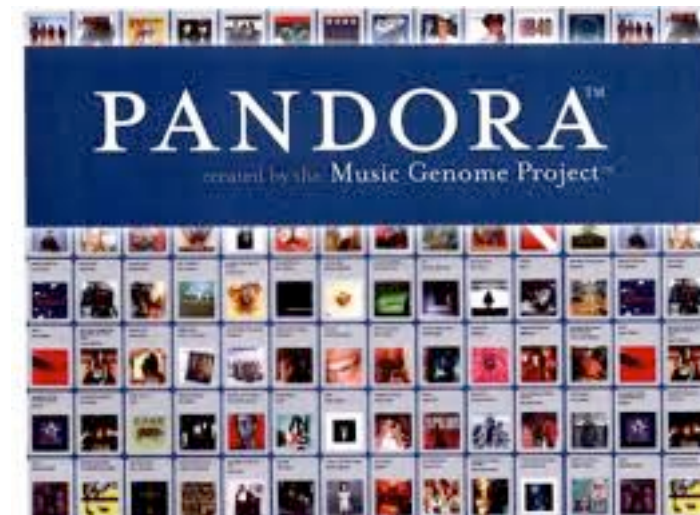
- Needs training data **in every language**
- Is it really relevant for better ads?
- Does it **cover** relevant areas?

# Collaborative Filtering



**NETFLIX**

amazon.com



# 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

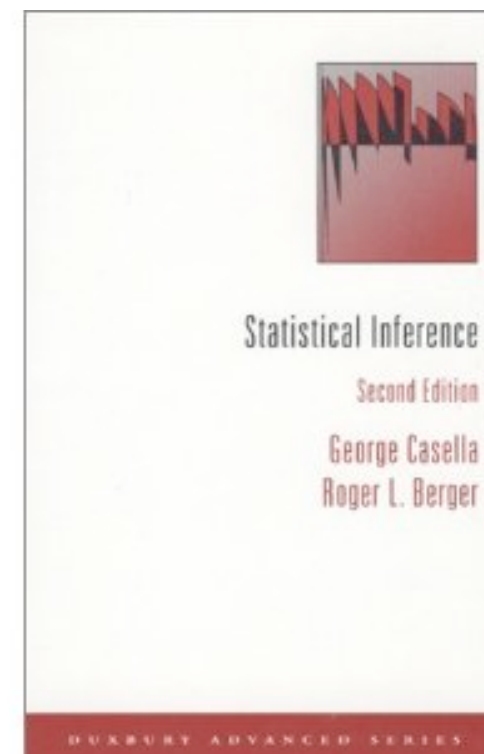
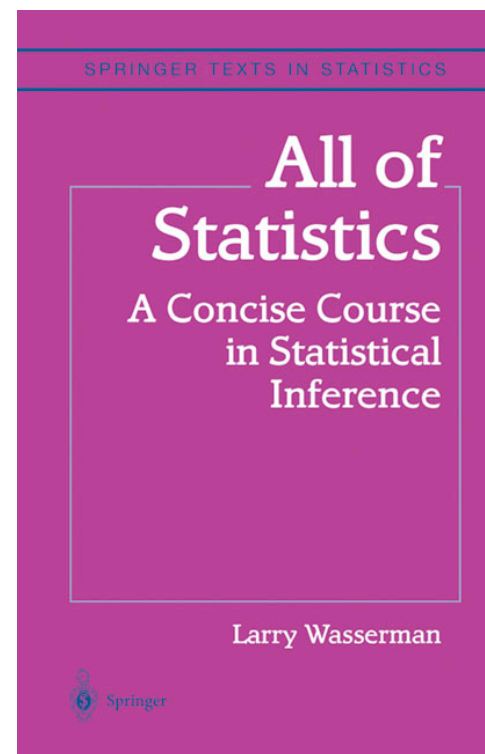
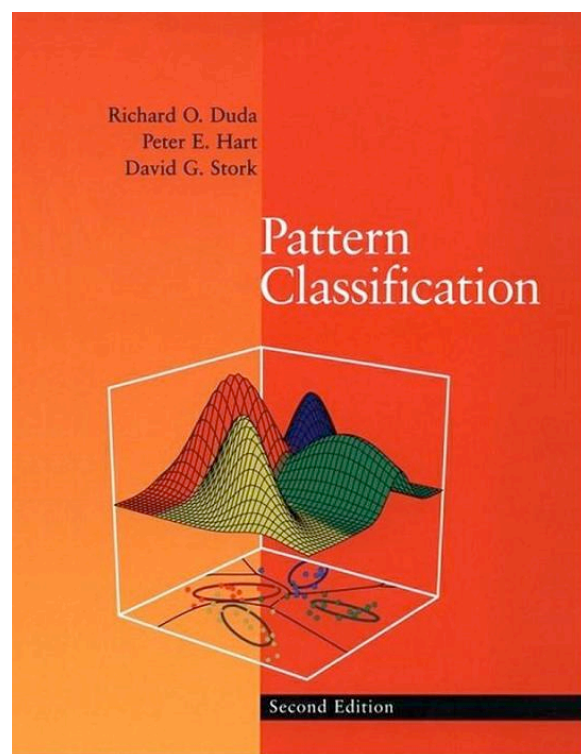
# Summary

- Essentially infinite amount of data
  - Labeling is (in many cases) prohibitively expensive
  - Editorial data not scalable for  $10^8$
  - 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 3 - Basic Tools



# Statistics 101



# Probability

- Space of events  $X$ 
  - server working; slow response; server broken
  - income of the user (e.g. \$95,000)
  - query text for search (e.g. "statistics tutorial")

- Probability axioms (Kolmogorov)

$$\Pr(X) \in [0, 1], \Pr(\mathcal{X}) = 1$$

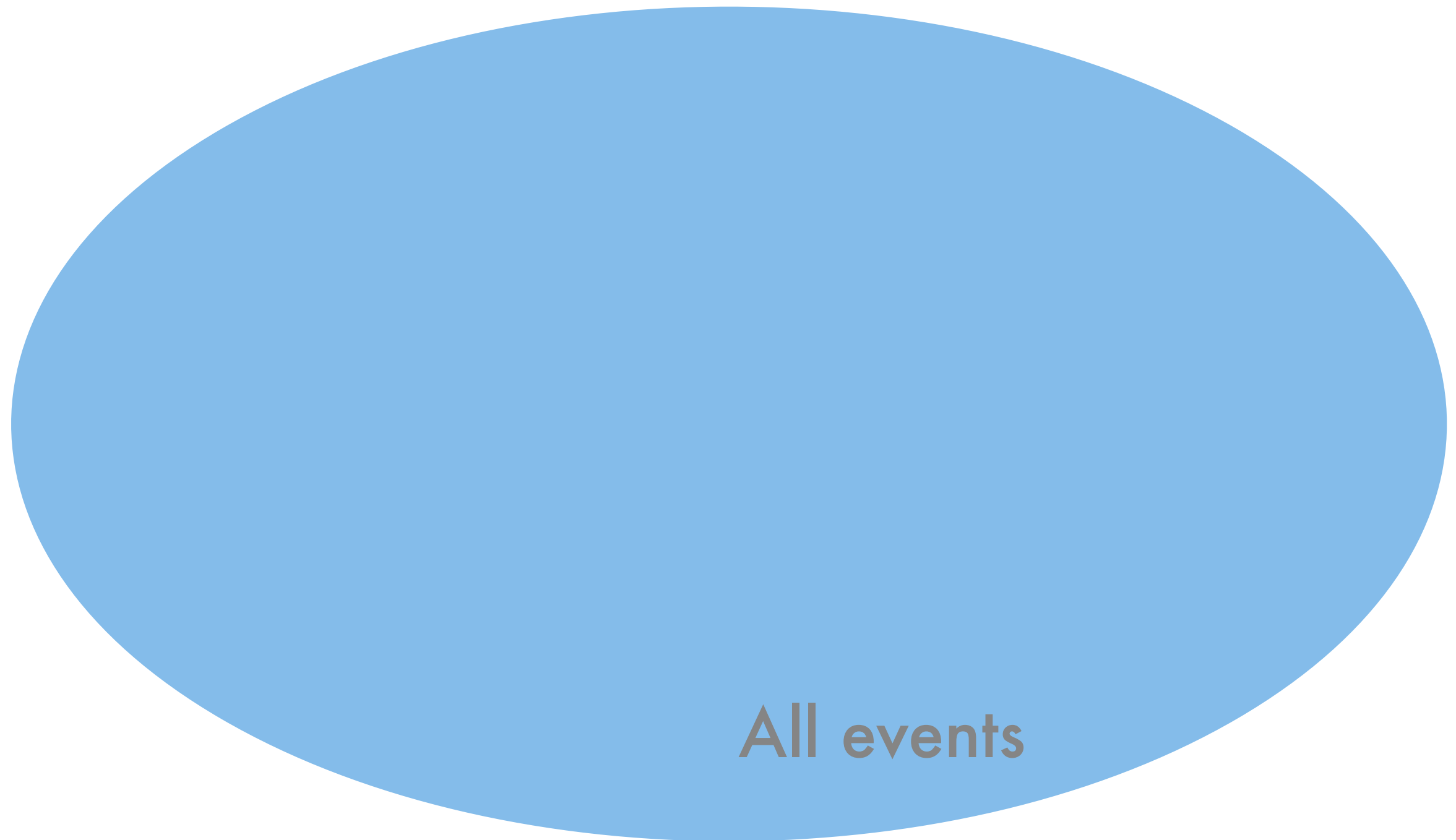
$$\Pr(\cup_i X_i) = \sum_i \Pr(X_i) \text{ if } X_i \cap X_j = \emptyset$$

- Example queries

- $P(\text{server working}) = 0.999$

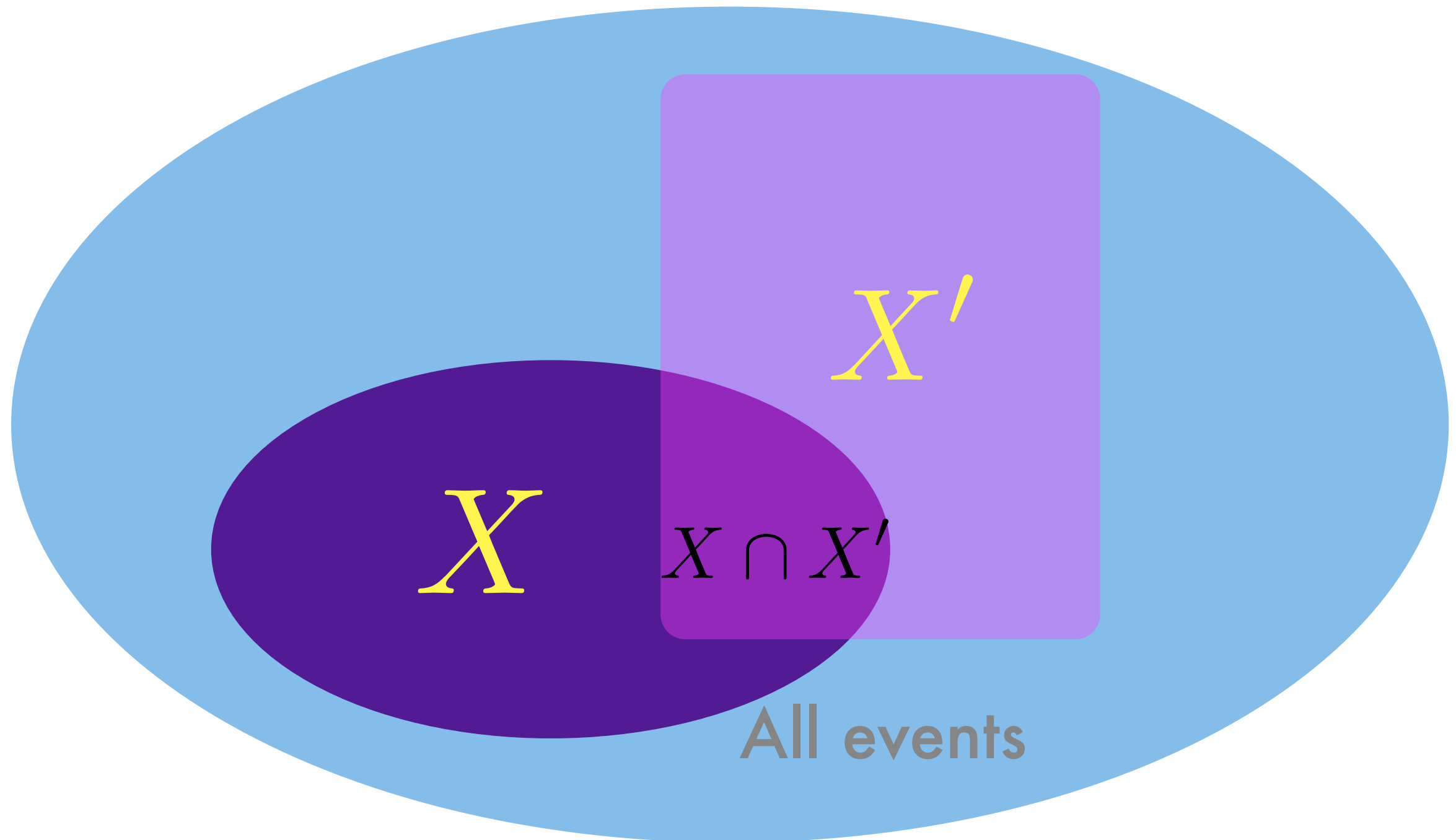
- $P(90,000 < \text{income} < 100,000) = 0.1$

# Venn Diagram

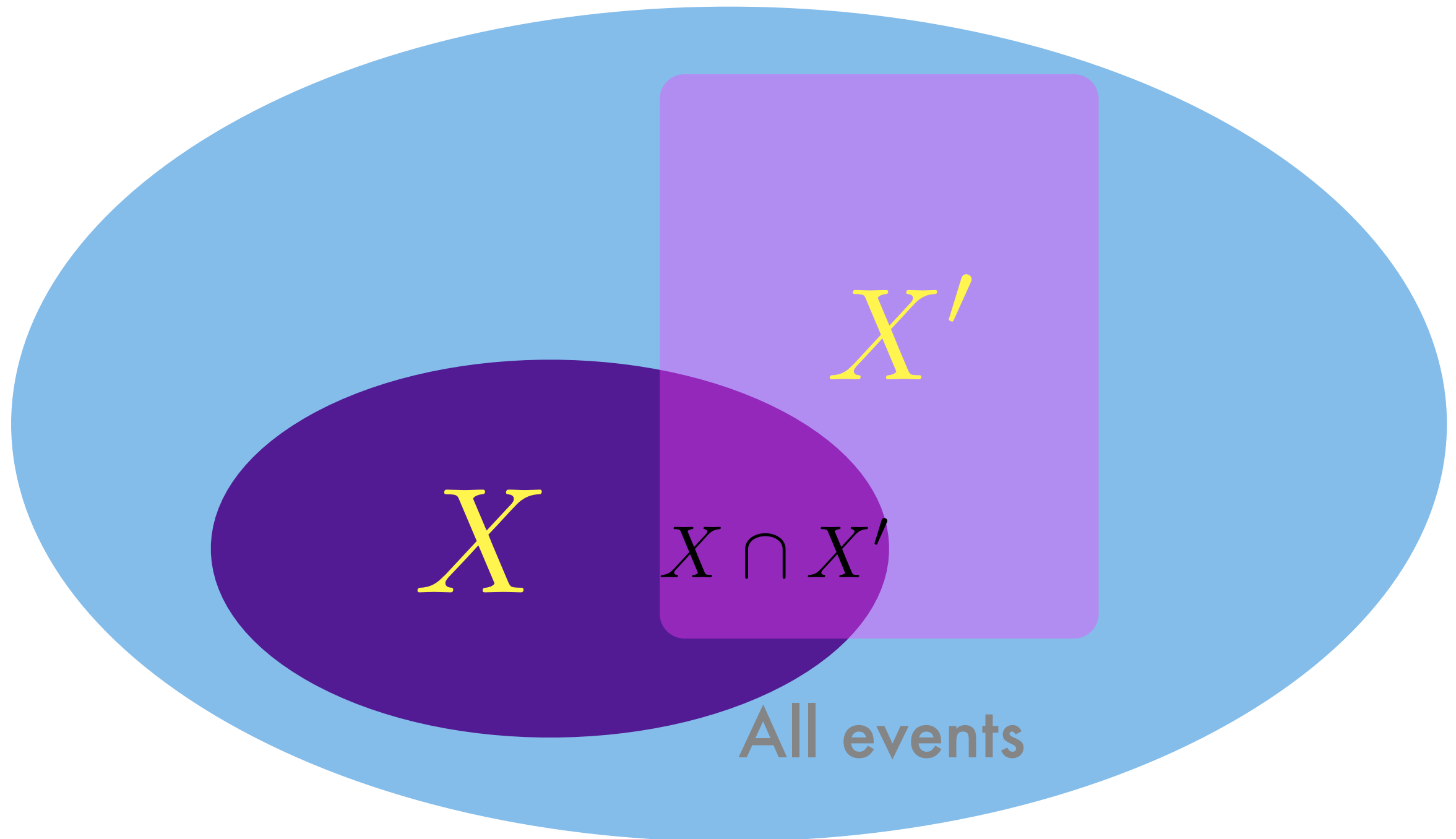




# Venn Diagram



# Venn Diagram



$$\Pr(X \cup X') = \Pr(X) + \Pr(X') - \Pr(X \cap X')$$

# (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



0.25

0.25



0.25

0.25

# Dependence



0.45

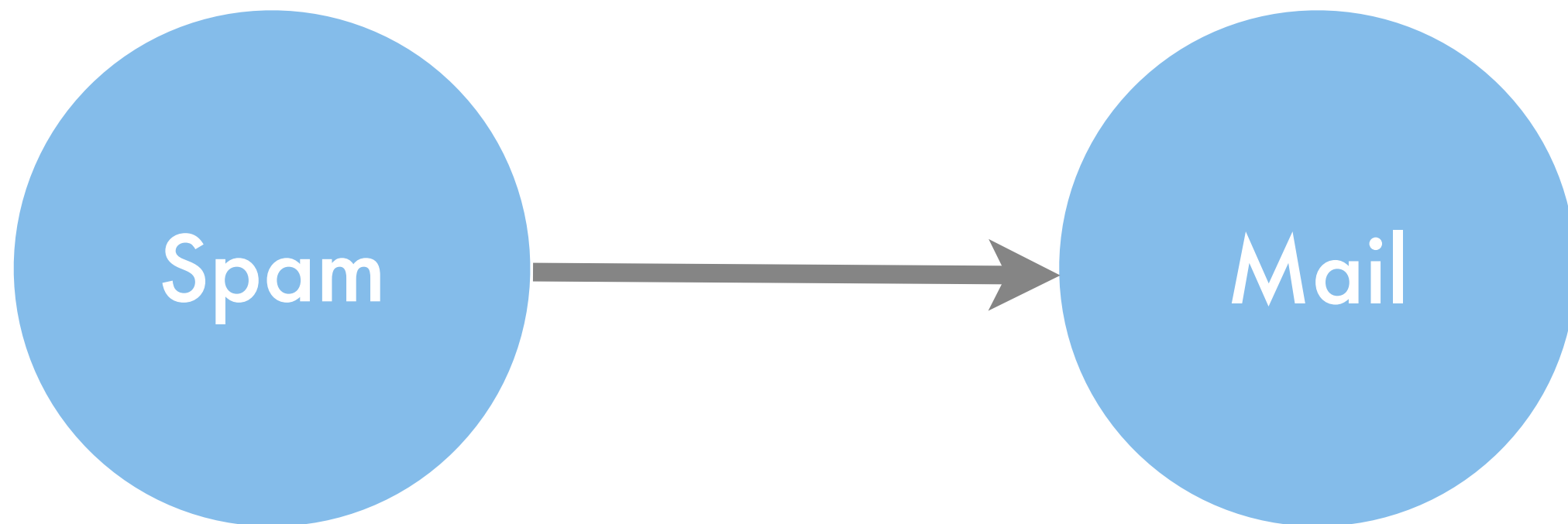
0.05



0.05

0.45

# A Graphical Model



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



# Bayes Rule

- **Joint Probability**

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

- **Bayes Rule**

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

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

# AIDS test (Bayes rule)

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

# AIDS test (Bayes rule)

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

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

# Improving the diagnosis

# Improving the diagnosis

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

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

# Improving the diagnosis

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

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

- **Why can't we use Test 1 twice?**  
Outcomes are **not** independent but tests 1 and 2 are **conditionally independent**

# Improving the diagnosis

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

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

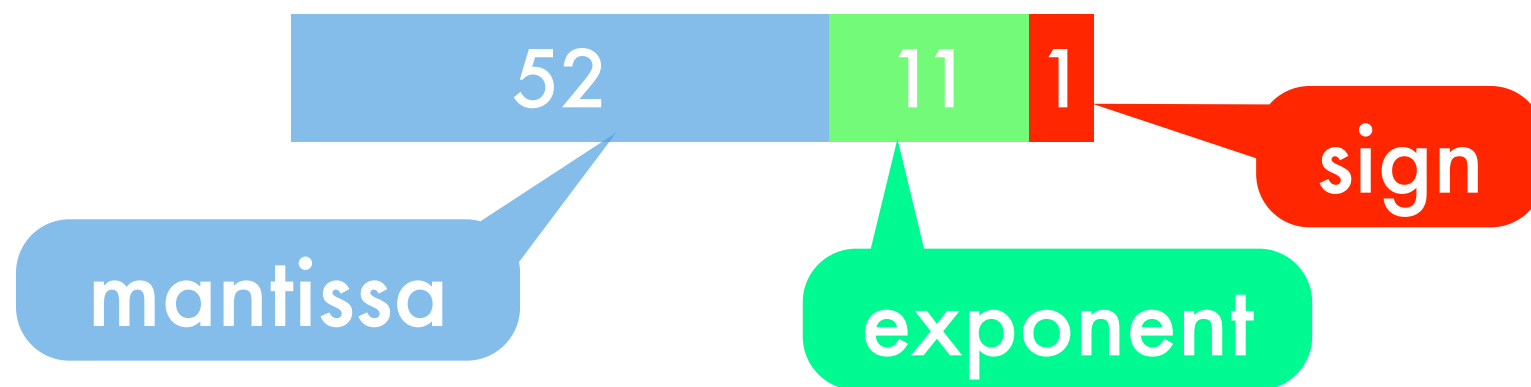
- **Why can't we use Test 1 twice?**

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

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

# Logarithms are good

- Floating point numbers



$$\pi = \log p$$

- Probabilities can be very small. In particular products of many probabilities. **Underflow!**
- Store data in **mantissa**, not **exponent**

$$\prod_i p_i \rightarrow \sum_i \pi_i \qquad \sum_i p_i \rightarrow \max \pi + \log \sum_i \exp [\pi_i - \max \pi]$$

- **Known bug e.g. in Mahout Dirichlet clustering**



# Application: Naive Bayes



# Naive Bayes Spam Filter

# Naive Bayes Spam Filter

- **Key assumption**

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

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

# Naive Bayes Spam Filter

- **Key assumption**

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

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

- **Spam classification via Bayes Rule**

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

# Naive Bayes Spam Filter

- **Key assumption**

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

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

- **Spam classification via Bayes Rule**

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

- **Parameter estimation**

Compute spam probability and word distributions for spam and ham

# Naive Bayes Spam Filter

## Equally likely phrases

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

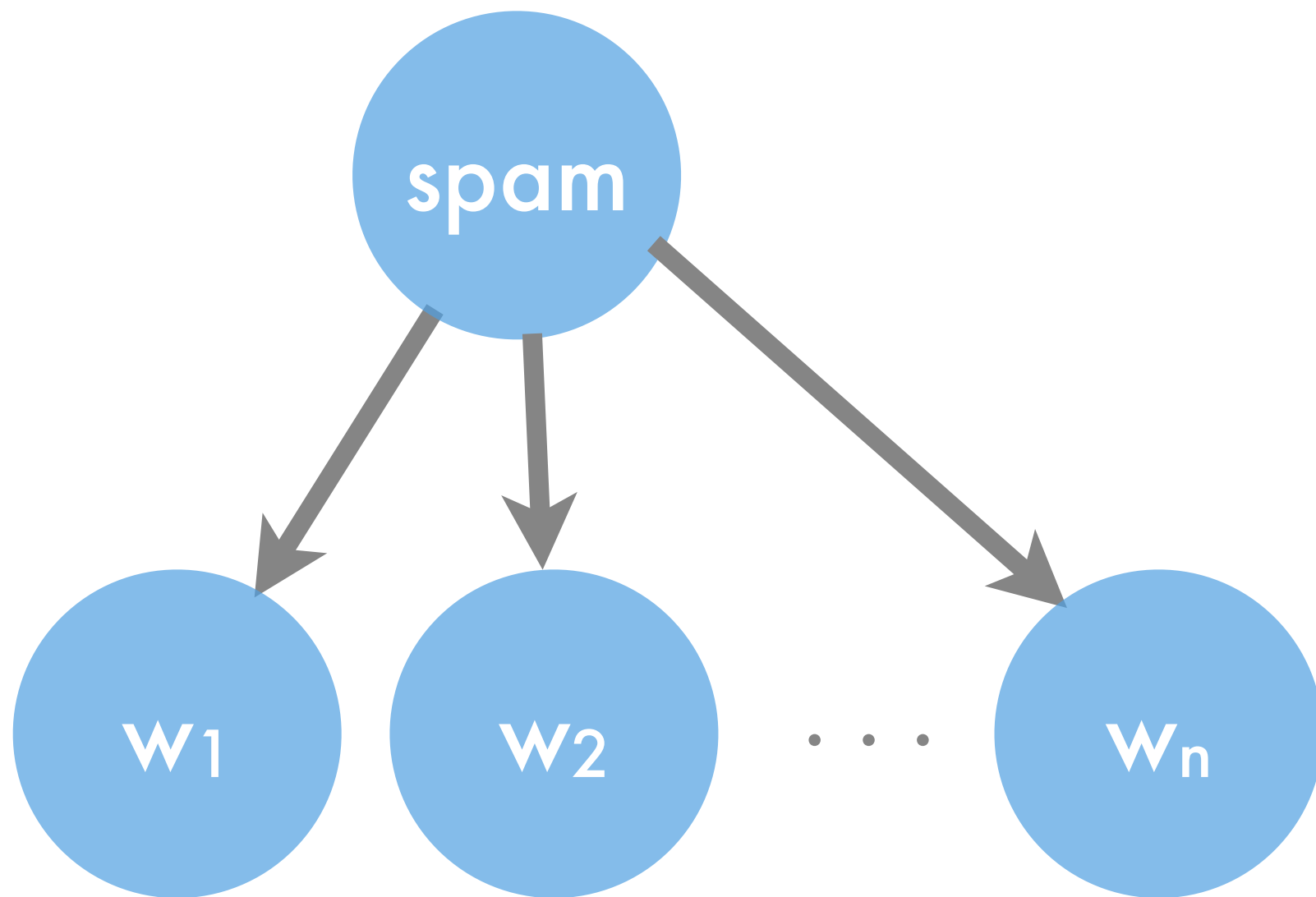
# Naive Bayes Spam Filter

## Equally likely phrases

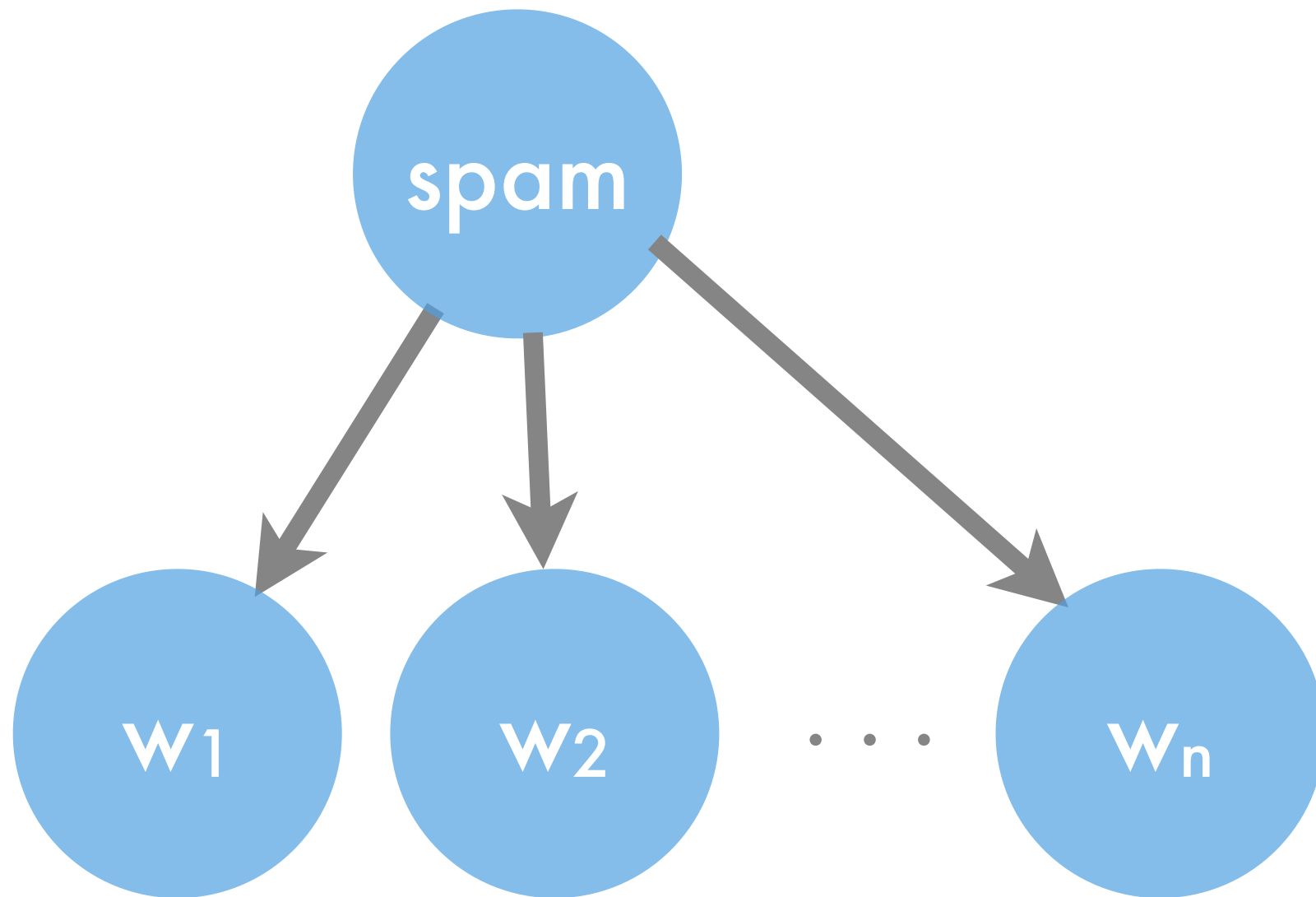
- Get rich quick. Buy WWW stock.
- Buy Viagra. Make your WWW experience last longer.
- You deserve a PhD from WWW University.  
We recognize your expertise.
- Make your rich WWW PhD experience last longer.



# A Graphical Model

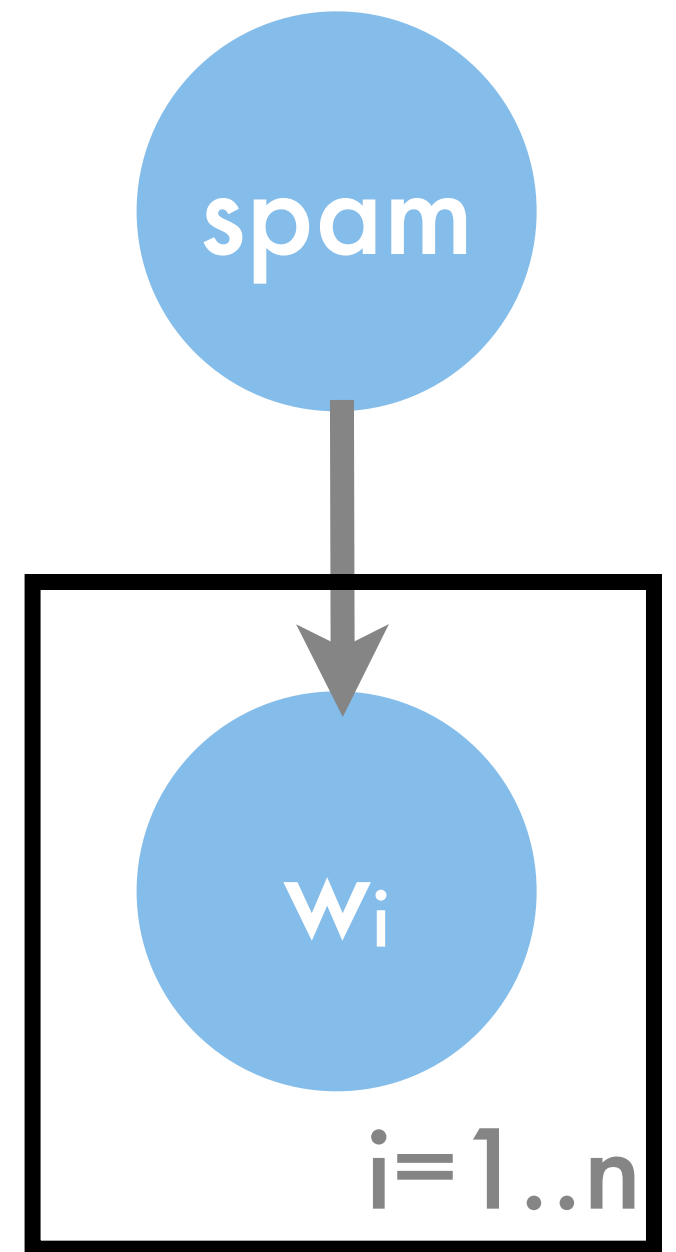
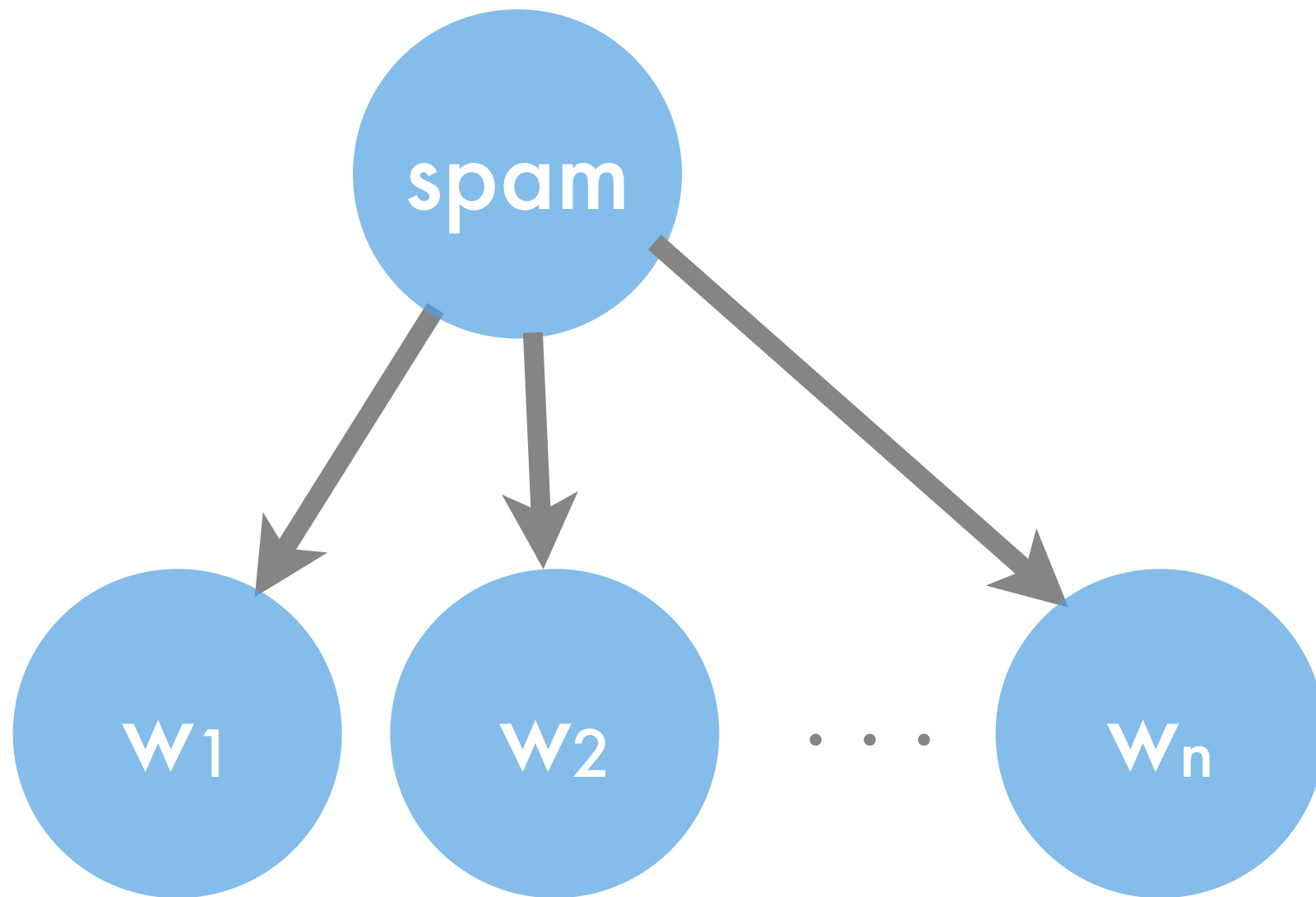


# A Graphical Model



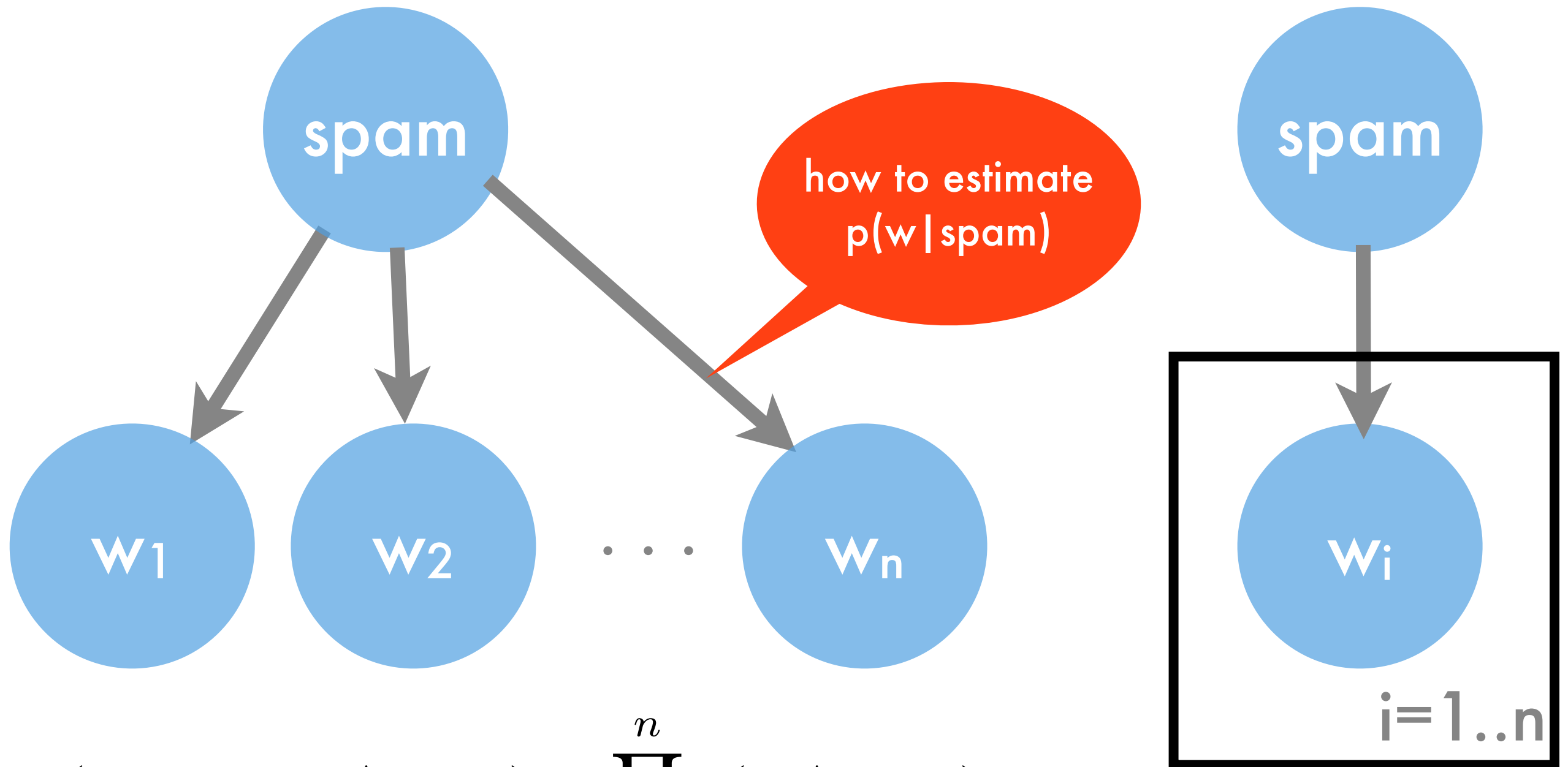
$$p(w_1, \dots, w_n | \text{spam}) = \prod_{i=1}^n p(w_i | \text{spam})$$

# A Graphical Model



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

# A Graphical Model



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

# Simple Algorithm

- For each document  $(x,y)$  do
  - Aggregate label counts given  $y$
  - For each feature  $x_i$  in  $x$  do
    - Aggregate statistic for  $(x_i, y)$  for each  $y$
- For  $y$  estimate distribution  $p(y)$
- For each  $(x_i,y)$  pair do  
Estimate distribution  $p(x_i | y)$ , e.g. Parzen Windows, Exponential family (Gauss, Laplace, Poisson, ...), Mixture
- Given new instance compute

$$p(y|x) \propto p(y) \prod_j p(x_j|y)$$

# Simple Algorithm

- For each document  $(x,y)$  do
  - Aggregate label counts given  $y$  **pass over all data**
  - For each feature  $x_i$  in  $x$  do
    - Aggregate statistic for  $(x_i, y)$  for each  $y$
- For  $y$  estimate distribution  $p(y)$
- For each  $(x_i,y)$  pair do **trivially parallel**  
Estimate distribution  $p(x_i | y)$ , e.g. Parzen Windows, Exponential family (Gauss, Laplace, Poisson, ...), Mixture
- Given new instance compute

$$p(y|x) \propto p(y) \prod_j p(x_j|y)$$



# MapReduce Algorithm

- **Map(document (x,y))**
  - For each mapper for each feature  $x_i$  in  $x$  do
    - Aggregate statistic for  $(x_i, y)$  for each  $y$
  - Send aggregate statistics to reducer
- **Reduce( $x_i, y$ )**
  - Aggregate over all messages from mappers
  - Estimate distribution  $p(x_i|y)$ , e.g. Parzen Windows, Exponential family (Gauss, Laplace, Poisson, ...), Mixture
  - Send coordinate-wise model to global storage
- **Given new instance compute**

$$p(y|x) \propto p(y) \prod_j p(x_j|y)$$

# MapReduce Algorithm

- Map(document (x,y))
  - For each mapper for each feature  $x_i$  in  $x$  do
    - Aggregate statistic for  $(x_i, y)$  for each  $y$
  - Send aggregate statistics to reducer
- Reduce( $x_i, y$ )
  - Aggregate over all messages from mappers
  - Estimate distribution  $p(x_i|y)$ , e.g. Parzen Windows, Exponential family (Gauss, Laplace, Poisson, ...), Mixture
  - Send coordinate-wise model to global storage
- Given new instance compute

$$p(y|x) \propto p(y) \prod_j p(x_j|y)$$

# Naive NaiveBayes Classifier

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

feature probability

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

spam probability

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

# Naive NaiveBayes Classifier

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

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

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

# Naive NaiveBayes Classifier

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

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

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

# Estimating Probabilities





# Binomial Distribution

- Two outcomes (head, tail); (0,1)

- Data likelihood

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

- Maximum Likelihood Estimation



- Constrained optimization problem  $\pi \in [0, 1]$

- Incorporate constraint via  $p(x; \theta) = \frac{e^{x\theta}}{1 + e^\theta}$

- Taking derivatives yields

$$\theta = \log \frac{n_1}{n_0 + n_1} \iff p(x = 1) = \frac{n_1}{n_0 + n_1}$$

... in detail ...

$$p(X; \theta) = \prod_{i=1}^n p(x_i; \theta) = \prod_{i=1}^n \frac{e^{\theta x_i}}{1 + e^{\theta}}$$

$$\implies \log p(X; \theta) = \theta \sum_{i=1}^n x_i - n \log [1 + e^{\theta}]$$

$$\implies \partial_{\theta} \log p(X; \theta) = \sum_{i=1}^n x_i - n \frac{e^{\theta}}{1 + e^{\theta}}$$

$$\iff \frac{1}{n} \sum_{i=1}^n x_i = \frac{e^{\theta}}{1 + e^{\theta}} = p(x = 1)$$

# ... in detail ...

$$p(X; \theta) = \prod_{i=1}^n p(x_i; \theta) = \prod_{i=1}^n \frac{e^{\theta x_i}}{1 + e^{\theta}}$$

$$\implies \log p(X; \theta) = \theta \sum_{i=1}^n x_i - n \log [1 + e^{\theta}]$$

$$\implies \partial_{\theta} \log p(X; \theta) = \sum_{i=1}^n x_i - n \frac{e^{\theta}}{1 + e^{\theta}}$$

$$\iff \frac{1}{n} \sum_{i=1}^n x_i = \frac{e^{\theta}}{1 + e^{\theta}} = p(x = 1)$$

empirical probability of  $x=1$

# Discrete Distribution

- n outcomes (e.g. USA, Canada, India, UK, NZ)
- Data likelihood

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

- Maximum Likelihood Estimation

- Constrained optimization problem ... or ...
- Incorporate constraint via  $p(x; \theta) = \frac{\exp \theta_x}{\sum_{x'} \exp \theta_{x'}}$
- Taking derivatives yields

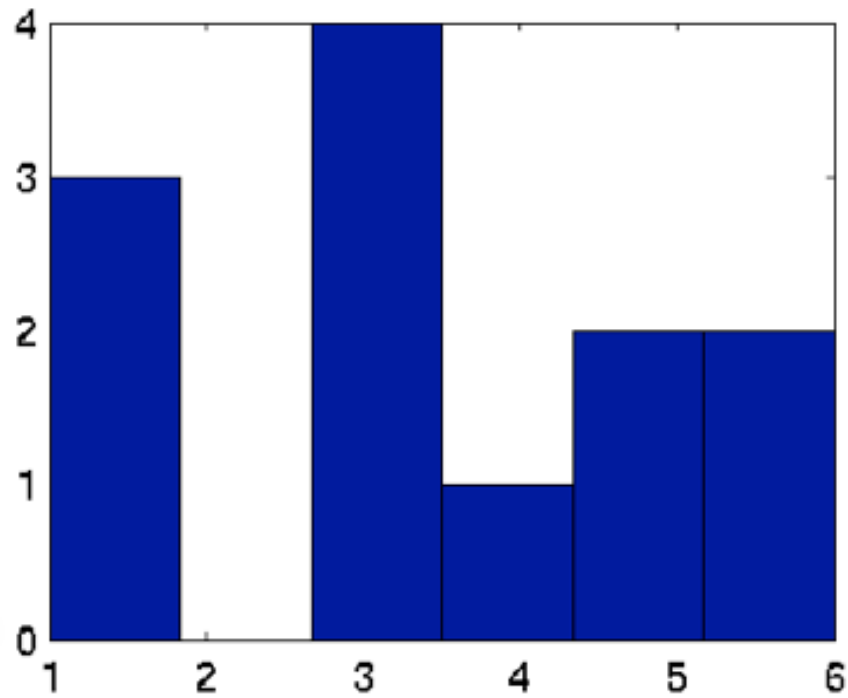
$$\theta_i = \log \frac{n_i}{\sum_j n_j} \iff p(x = i) = \frac{n_i}{\sum_j n_j}$$

# Tossing a Dice

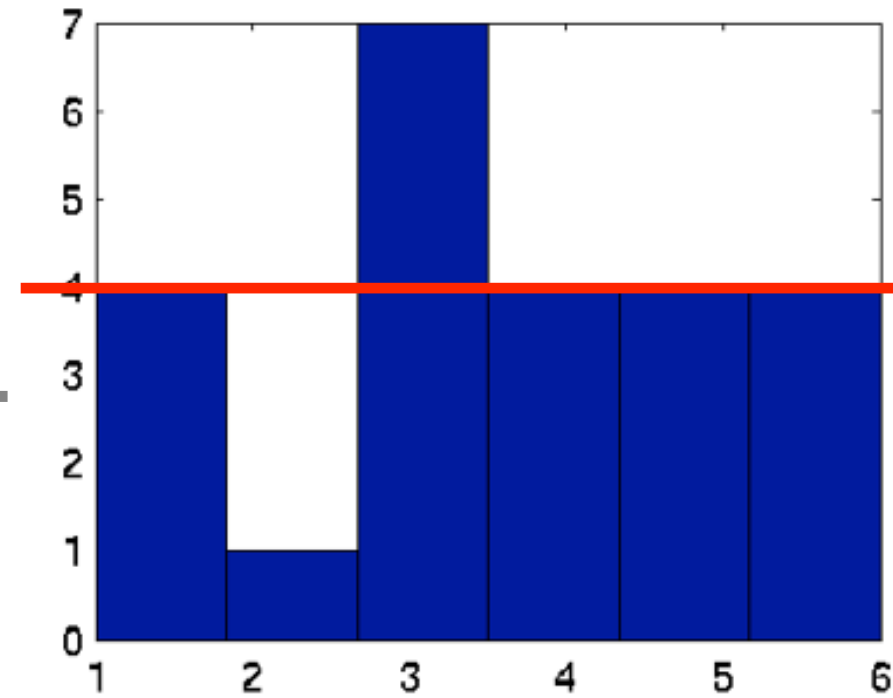
12



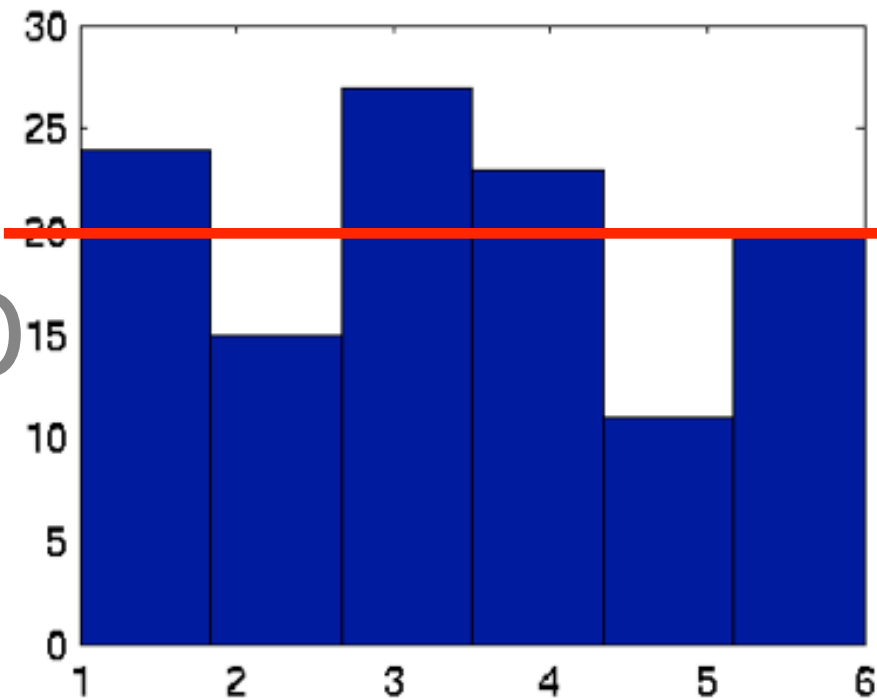
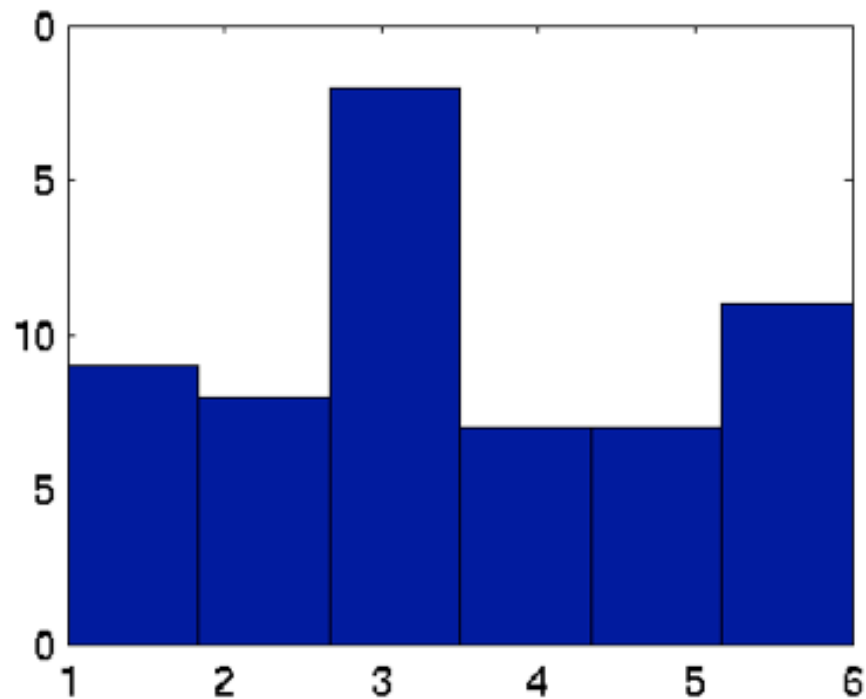
60



24



120

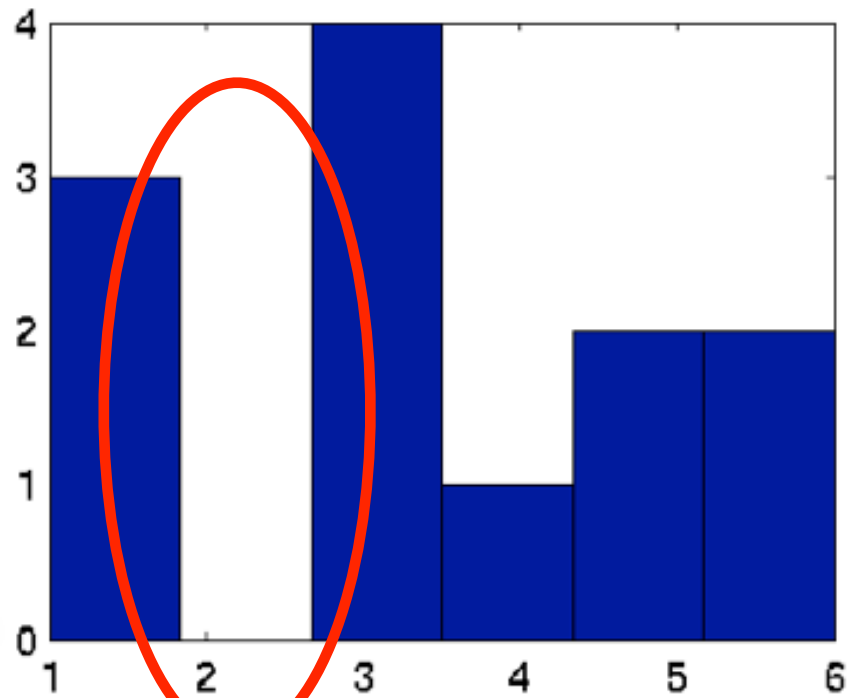


# Tossing a Dice

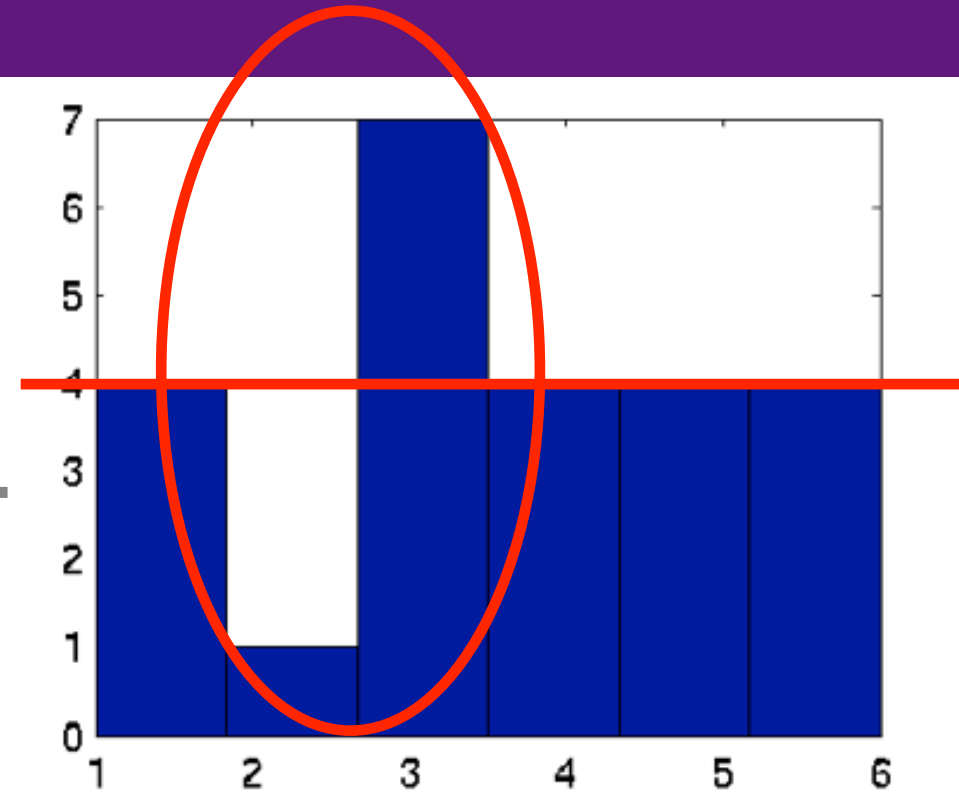
12



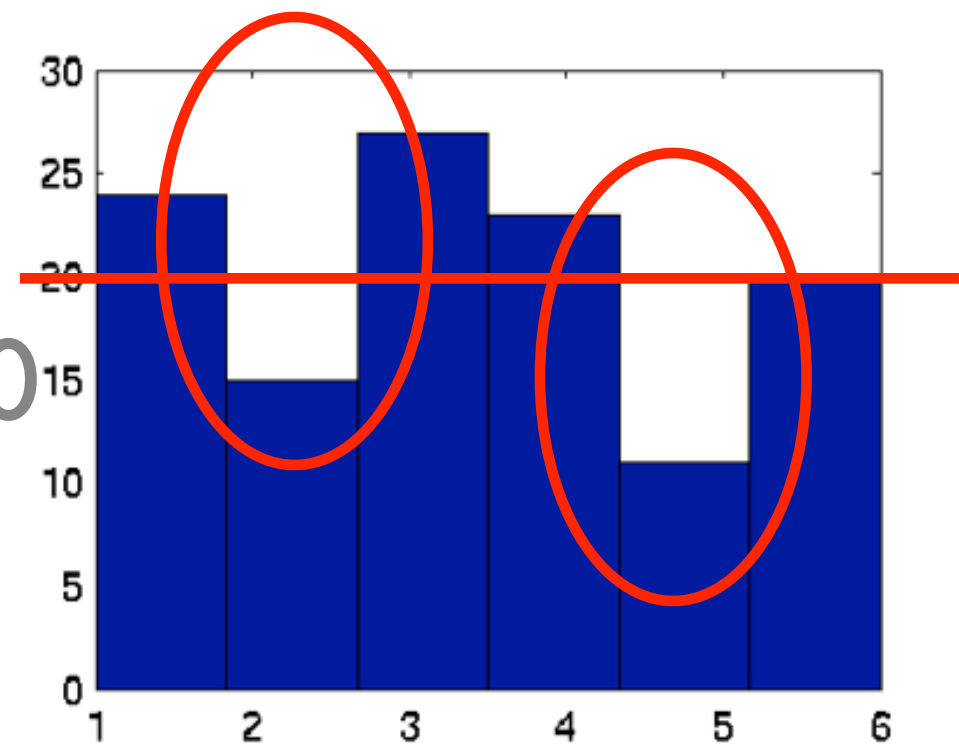
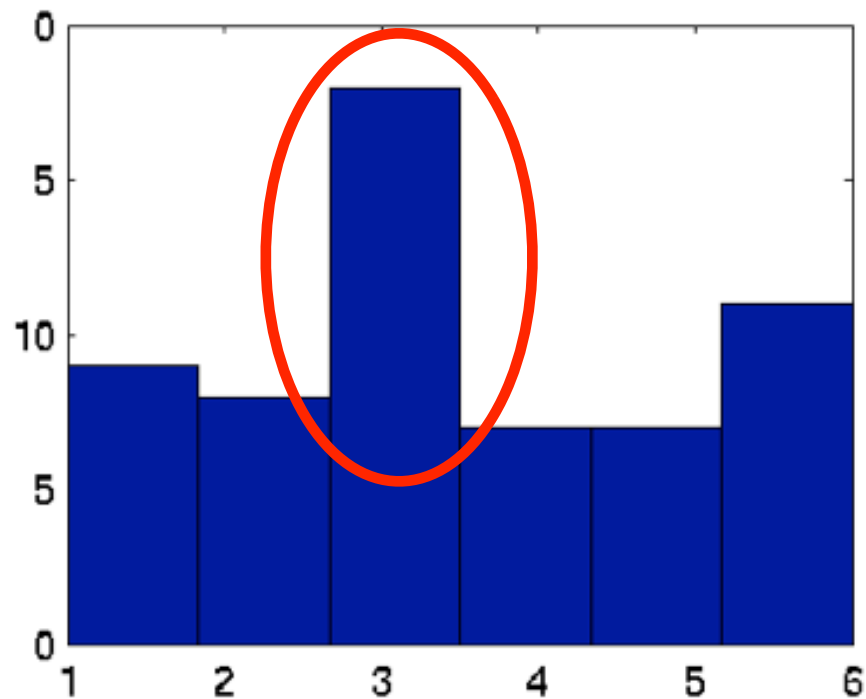
60



24



120



# Exponential Families





# Exponential Families

# Exponential Families

- **Density function**

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

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

# Exponential Families

- **Density function**

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

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

- **Log partition function generates cumulants**

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

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

# Exponential Families

- **Density function**

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

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

- **Log partition function generates cumulants**

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

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

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

# Examples

- **Binomial Distribution**

$$\phi(x) = x$$

- **Discrete Distribution**

$$\phi(x) = e_x$$

( $e_x$  is unit vector for  $x$ )

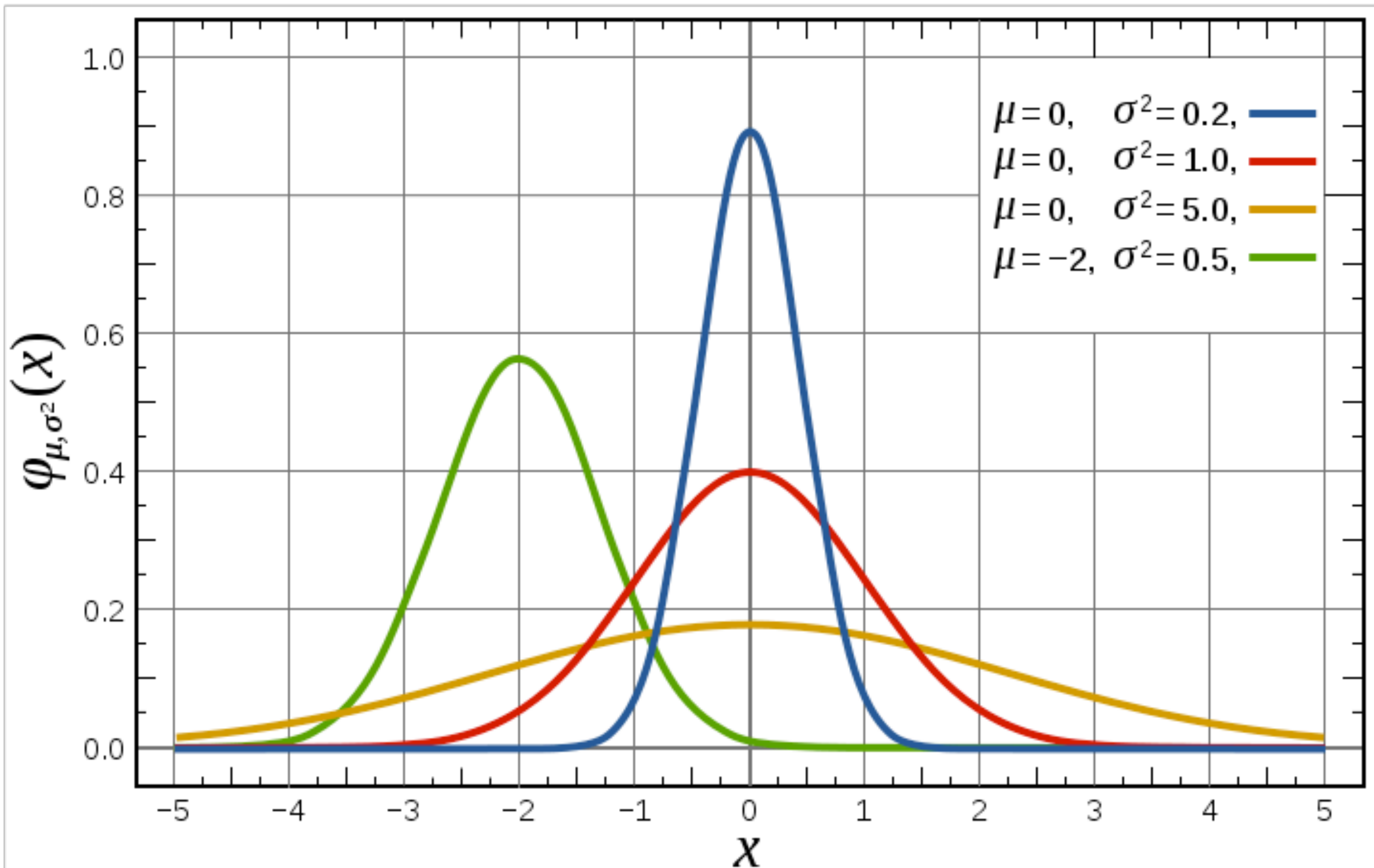
- **Gaussian**

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

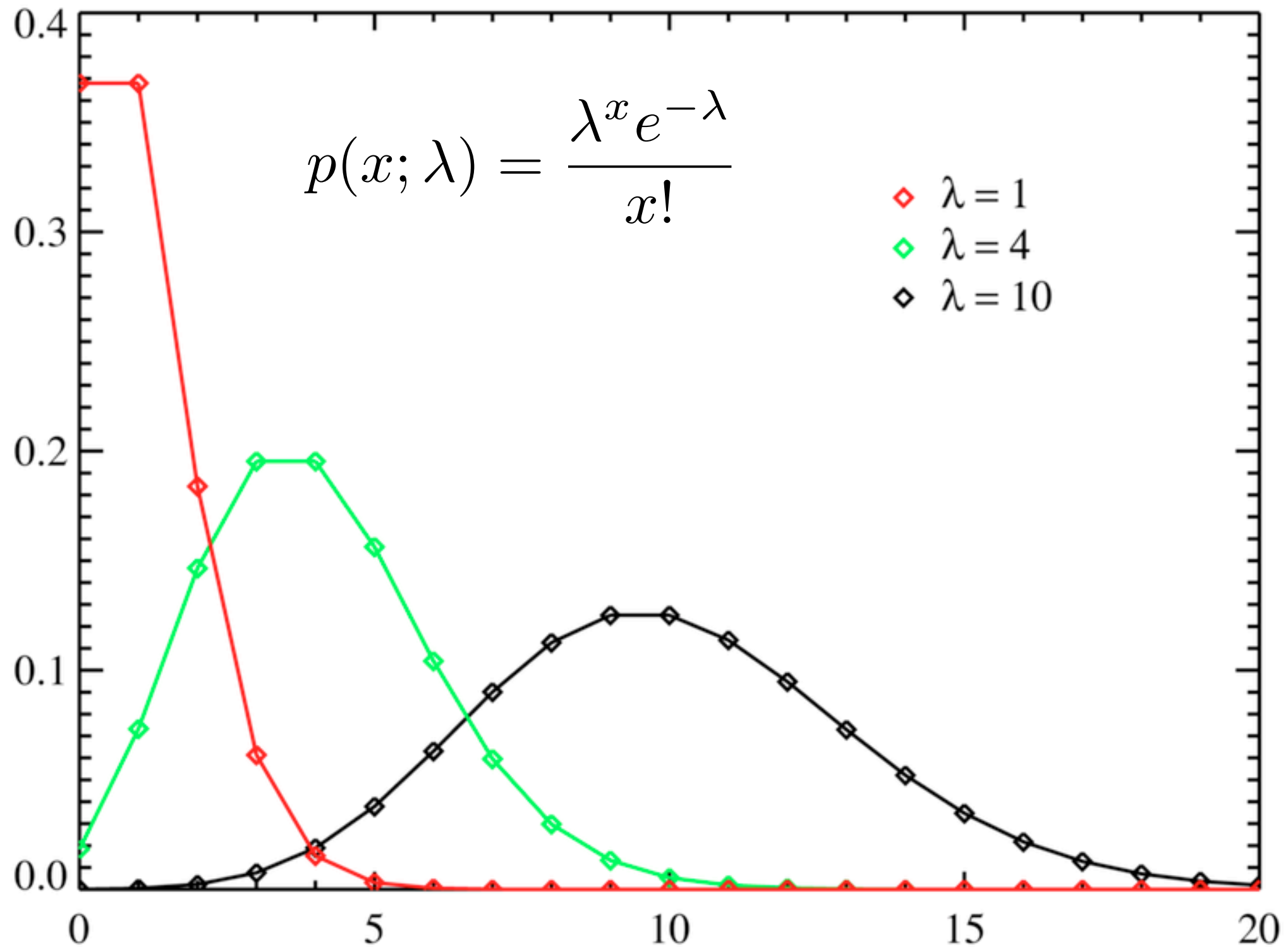
- **Poisson (counting measure  $1/x!$ )**  $\phi(x) = x$

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

# Normal Distribution

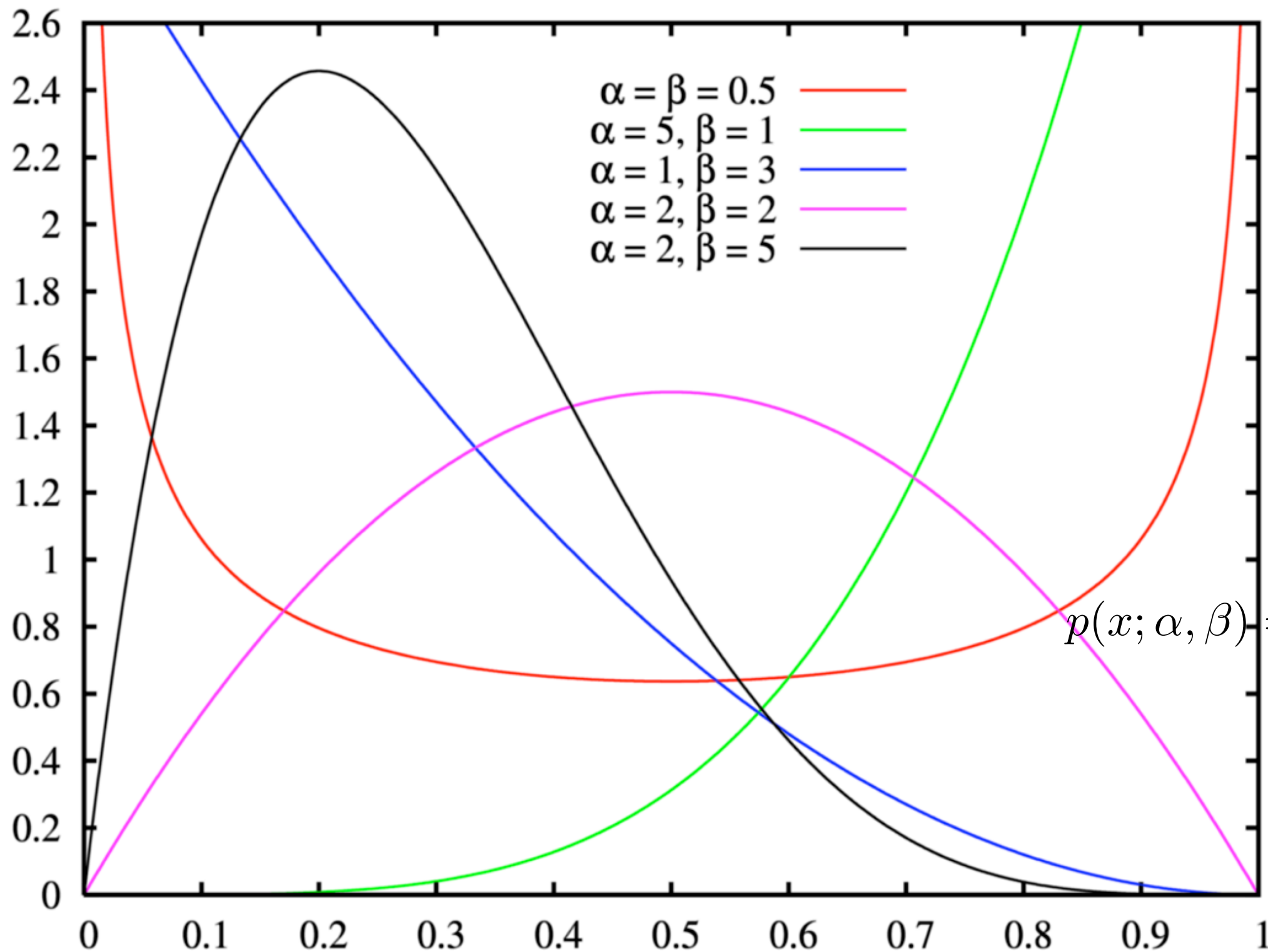


# Poisson Distribution



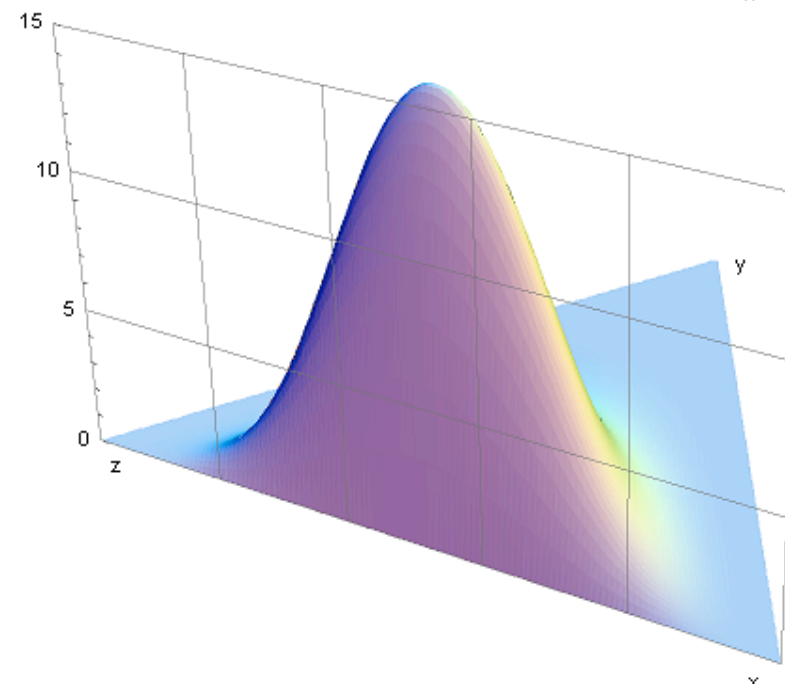
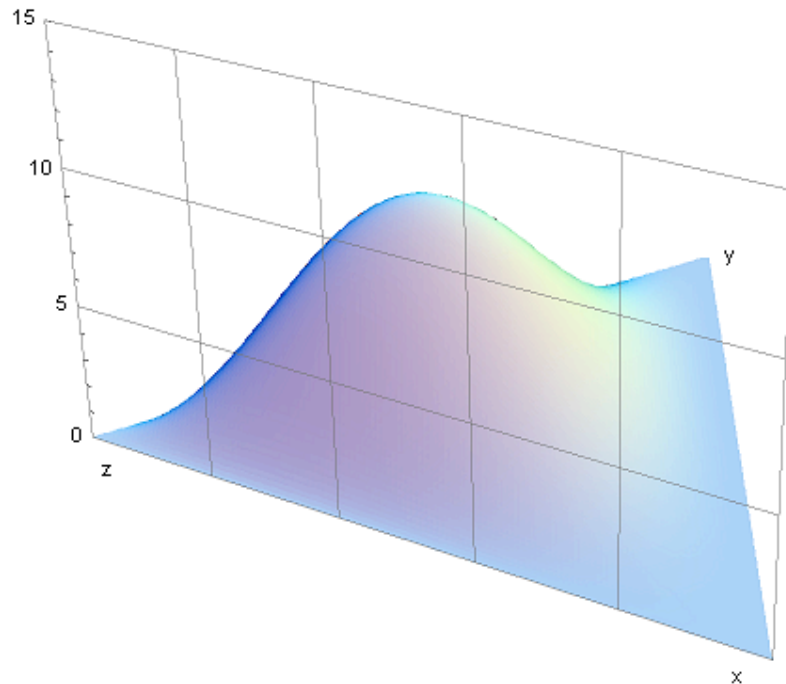
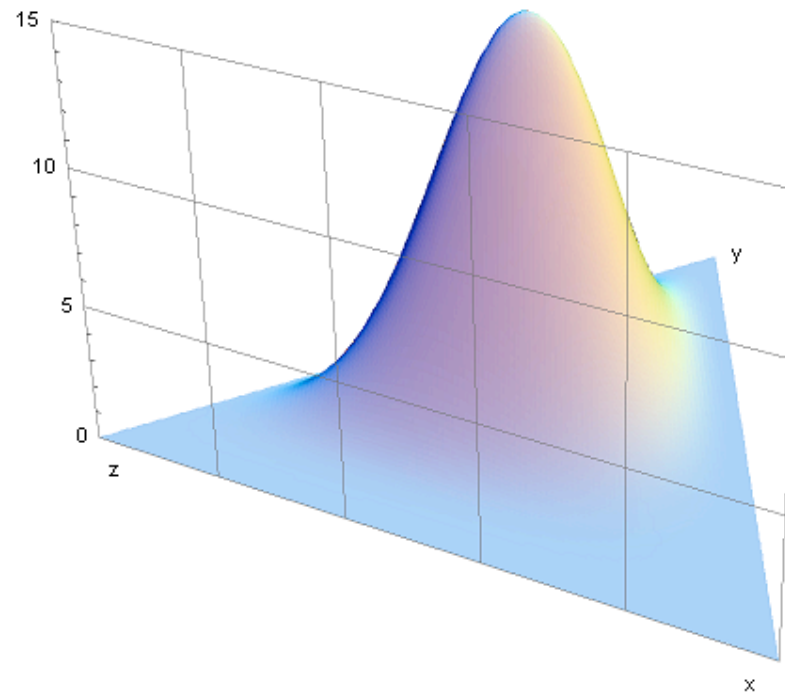
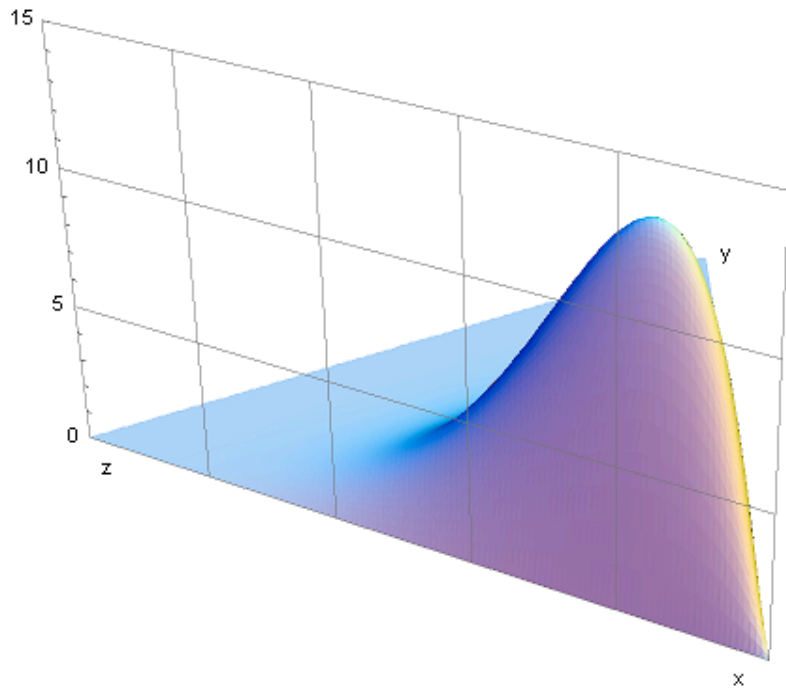


# Beta Distribution



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

# Dirichlet Distribution



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

# Maximum Likelihood

# Maximum Likelihood

- **Negative log-likelihood**

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

# Maximum Likelihood

- Negative log-likelihood

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

- Taking derivatives

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

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

# 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

# Example: Gaussian Estimation

- Sufficient statistics:  $x, x^2$

- Mean and variance given by

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

- Maximum Likelihood Estimate

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

- Maximum a Posteriori Estimate

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

smoother

smoother



# Collapsing

- Conjugate priors

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

Hence we know how to compute normalization

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

(Beta, binomial)

(Dirichlet, multinomial)

(Gamma, Poisson)

(Wishart, Gauss)

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

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

look up closed  
form expansions

# Conjugate Prior in action

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

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

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

# Conjugate Prior in action

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

- **Discrete Distribution**

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

- **Tossing a dice**

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

# Conjugate Prior in action

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

- **Discrete Distribution**

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

- **Tossing a dice**

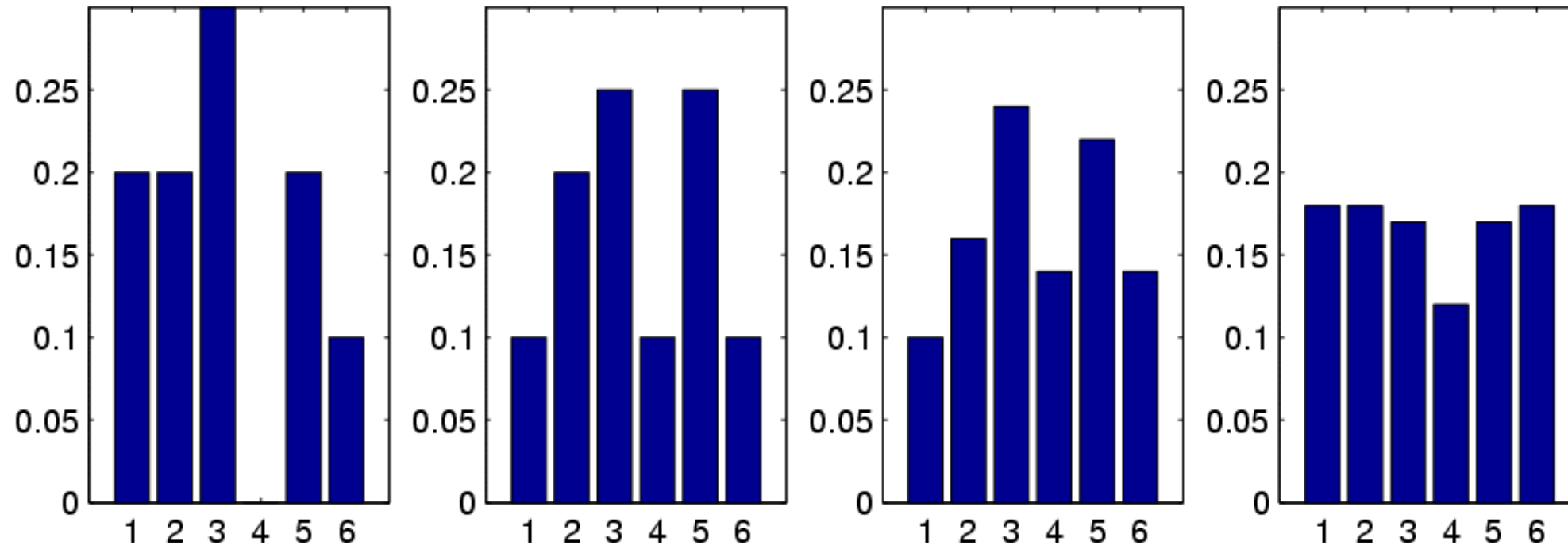
Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ( $m_0 = 6$ )	0.15	0.27	0.12	0.08	0.19	0.19
MAP ( $m_0 = 100$ )	0.16	0.19	0.16	0.15	0.17	0.17

- **Rule of thumb**

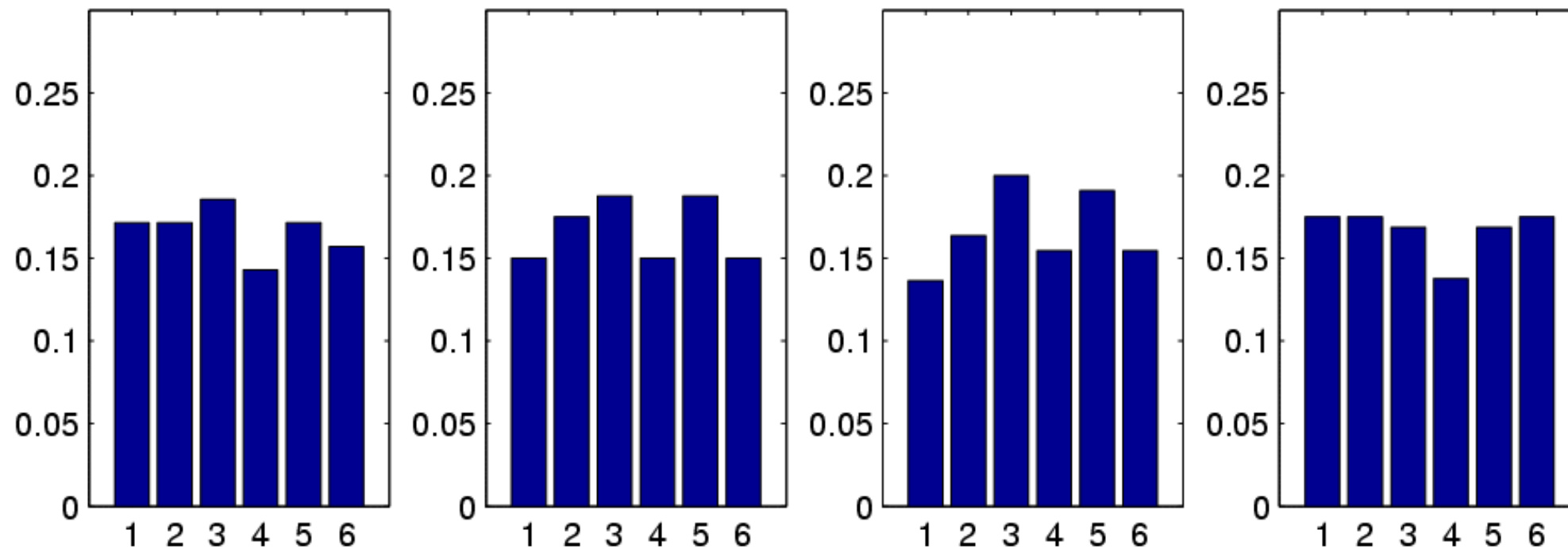
**need 10 data points (or prior) per parameter**

# Honest dice

MLE

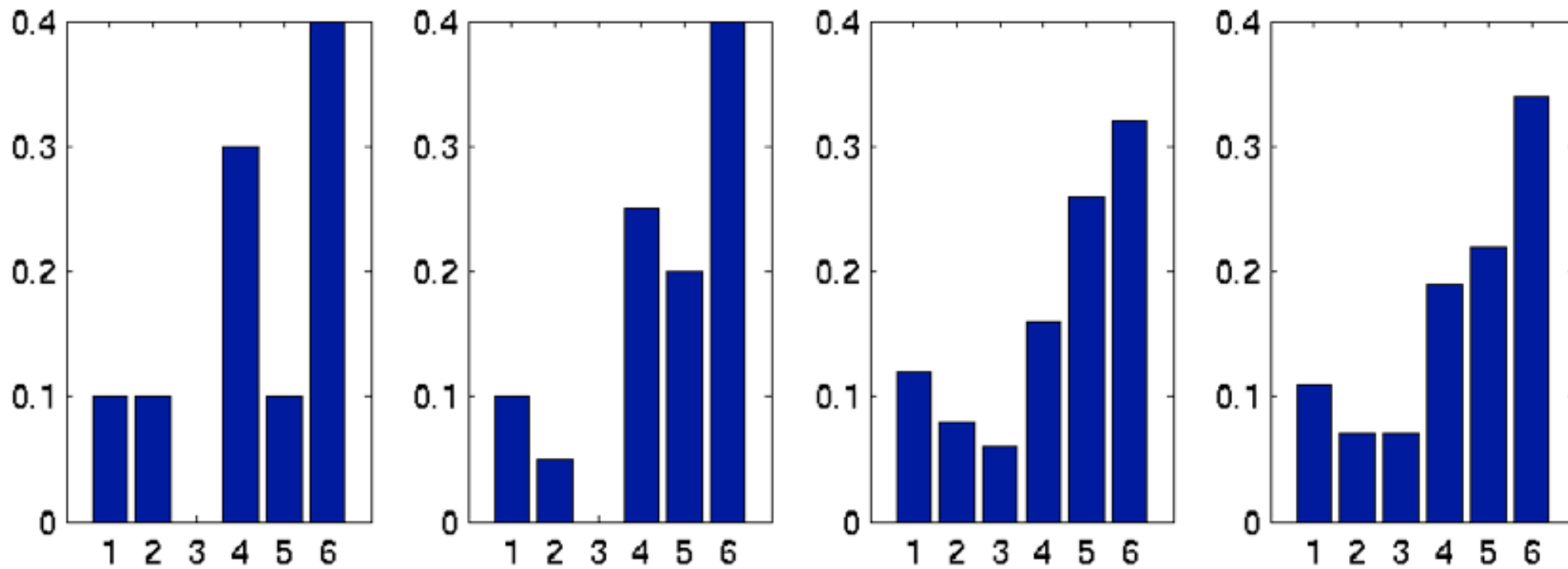


MAP

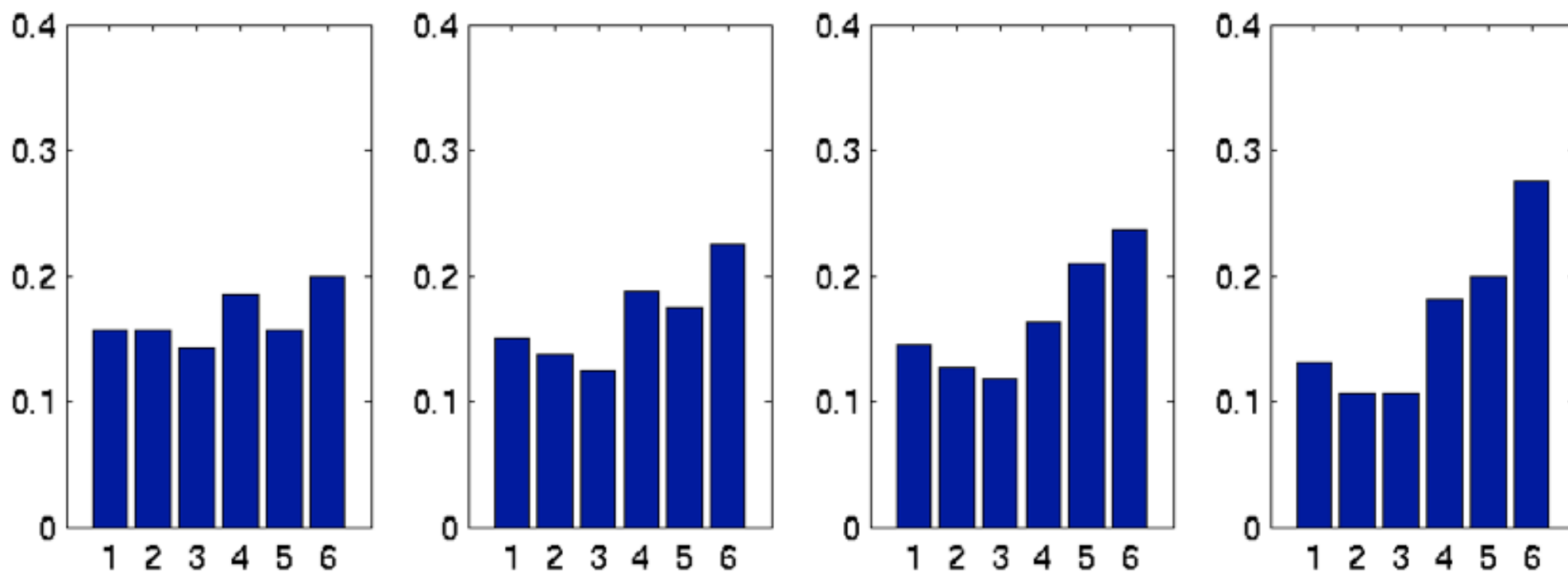


# Tainted dice

MLE



MAP



# Priors (part deux)

- **Parameter smoothing**

$$p(\theta) \propto \exp(-\lambda \|\theta\|_1) \text{ or } p(\theta) \propto \exp(-\lambda \|\theta\|_2^2)$$

- **Posterior**

$$\begin{aligned} p(\theta|x) &\propto \prod_{i=1}^m p(x_i|\theta)p(\theta) \\ &\propto \exp\left(\sum_{i=1}^m \langle \phi(x_i), \theta \rangle - mg(\theta) - \frac{1}{2\sigma^2} \|\theta\|_2^2\right) \end{aligned}$$

- **Convex optimization problem (MAP estimation)**

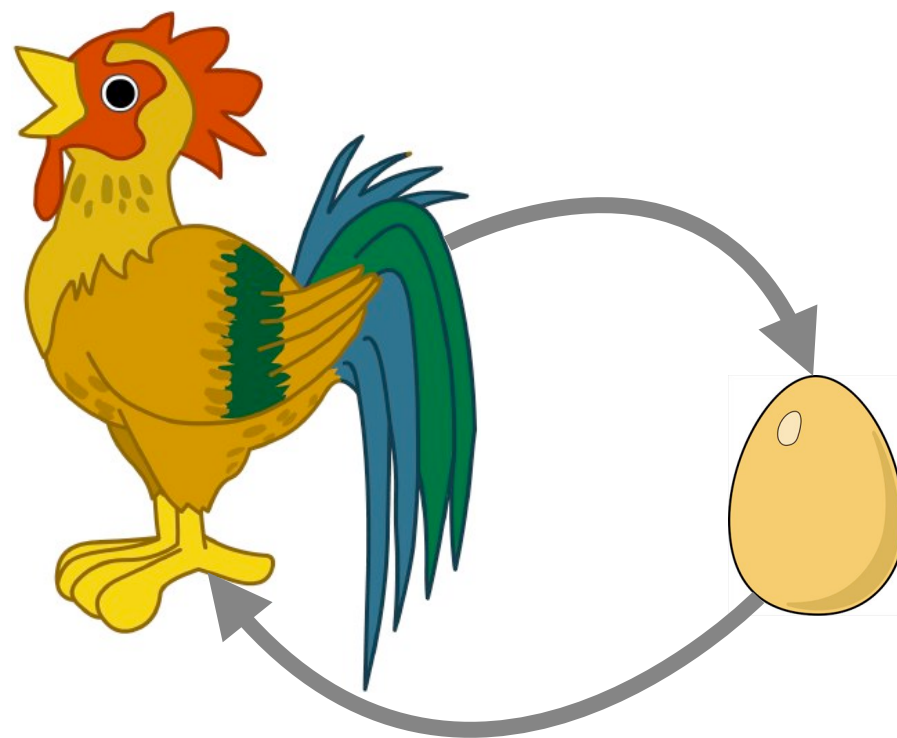
$$\underset{\theta}{\text{minimize}} \quad g(\theta) - \left\langle \frac{1}{m} \sum_{i=1}^m \phi(x_i), \theta \right\rangle + \frac{1}{2m\sigma^2} \|\theta\|_2^2$$



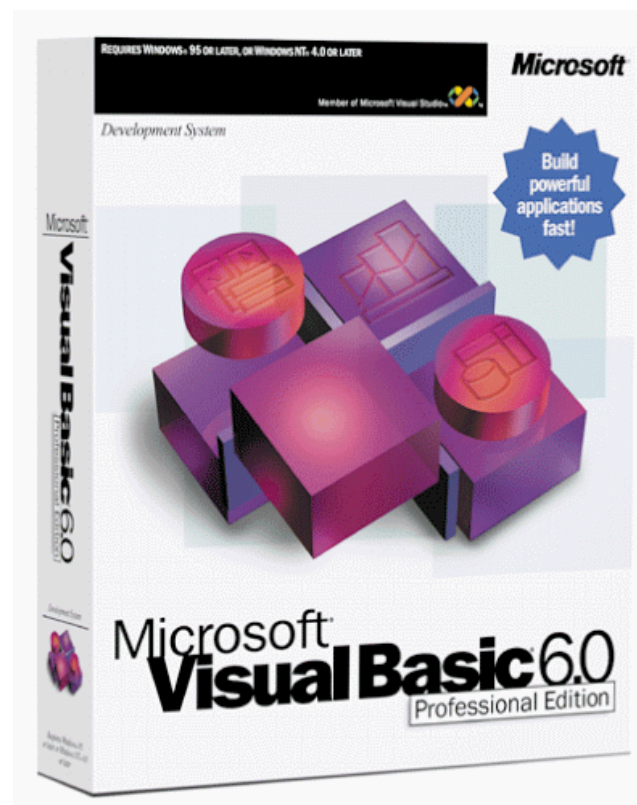
# Summary

- **Basic statistics tools**
- **Estimating probabilities (mainly scalar)**
- **Exponential family introduction**

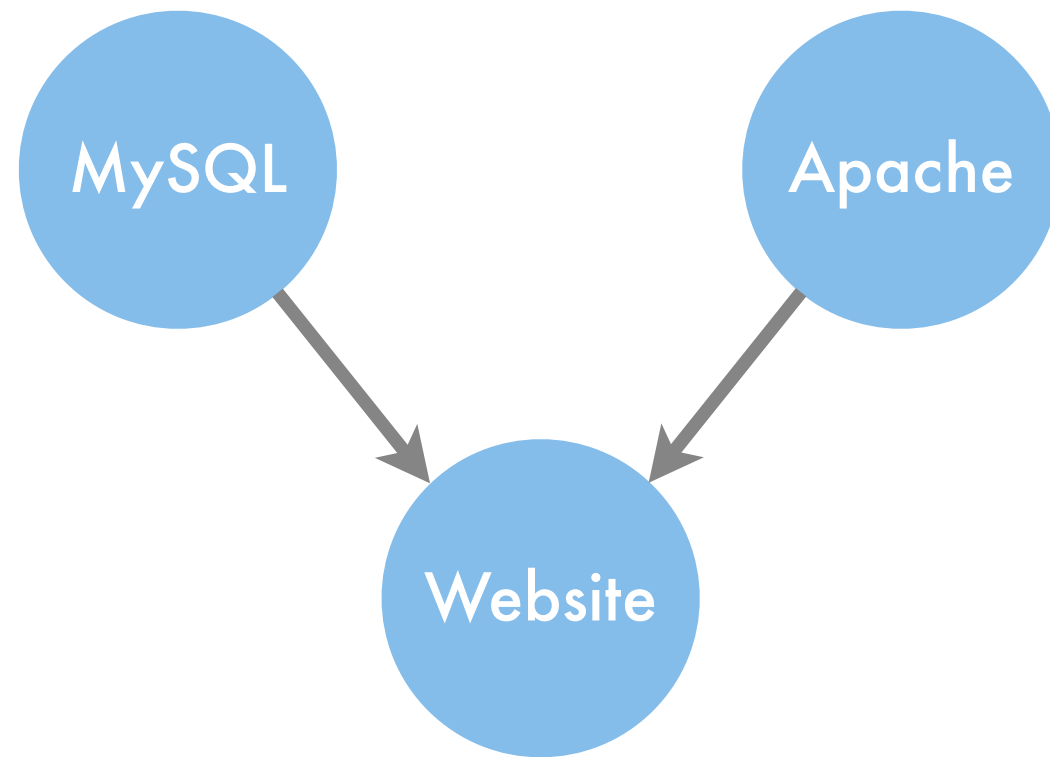
# Part 4: Directed Graphical



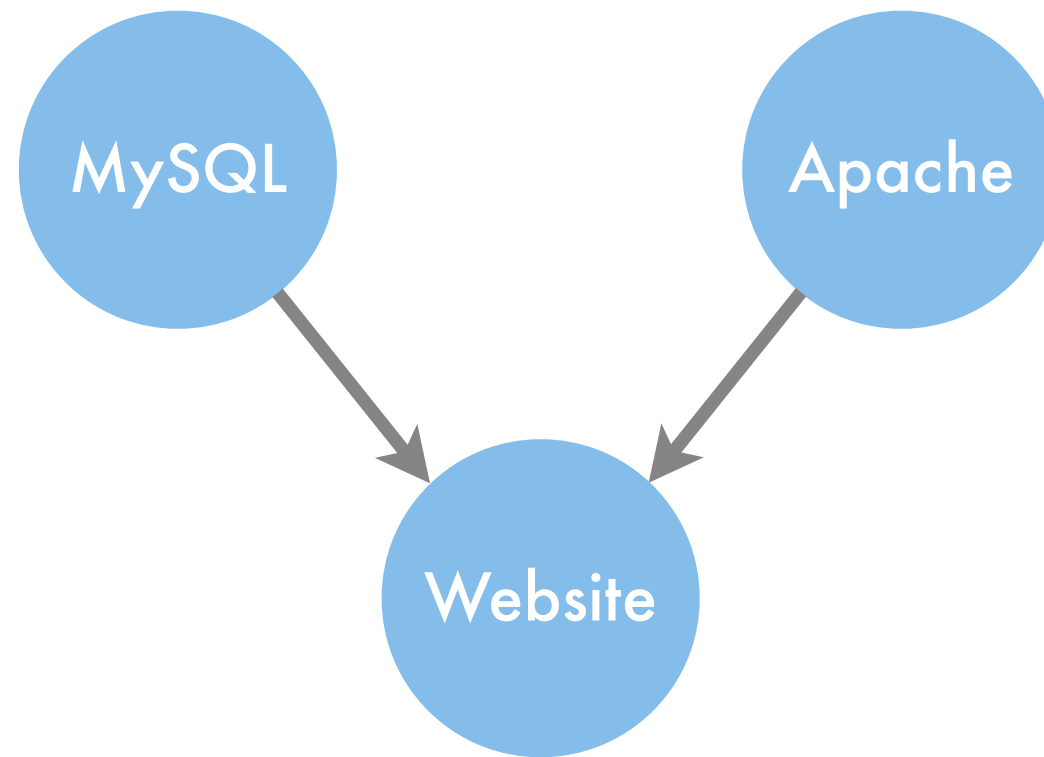
# Basics



... some Web 2.0 service



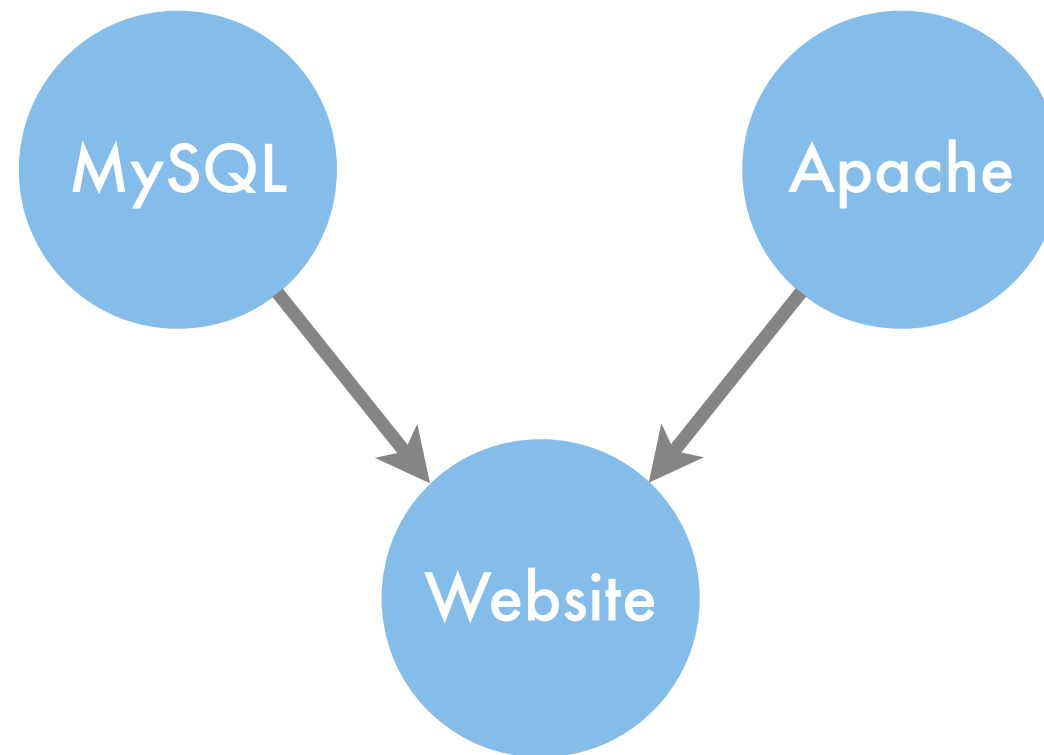
# ... some Web 2.0 service



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

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

# ... some Web 2.0 service



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

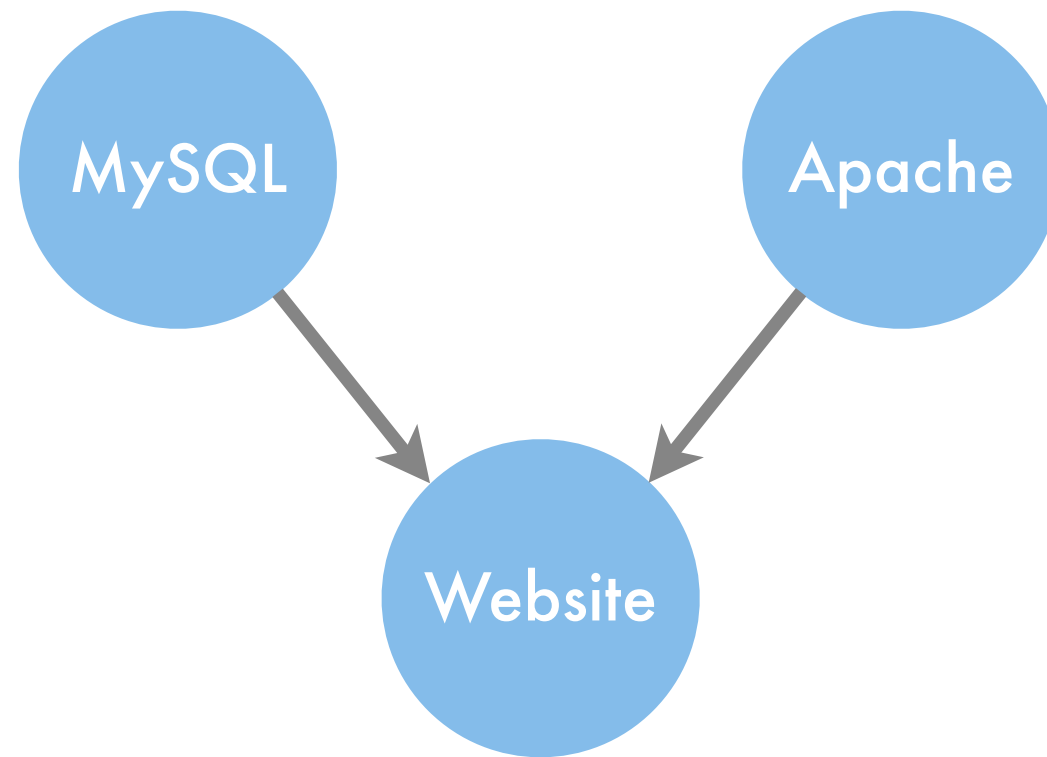
$$p(m, a, w) = p(w|m, a)p(m)p(a)$$

- **Explaining away**

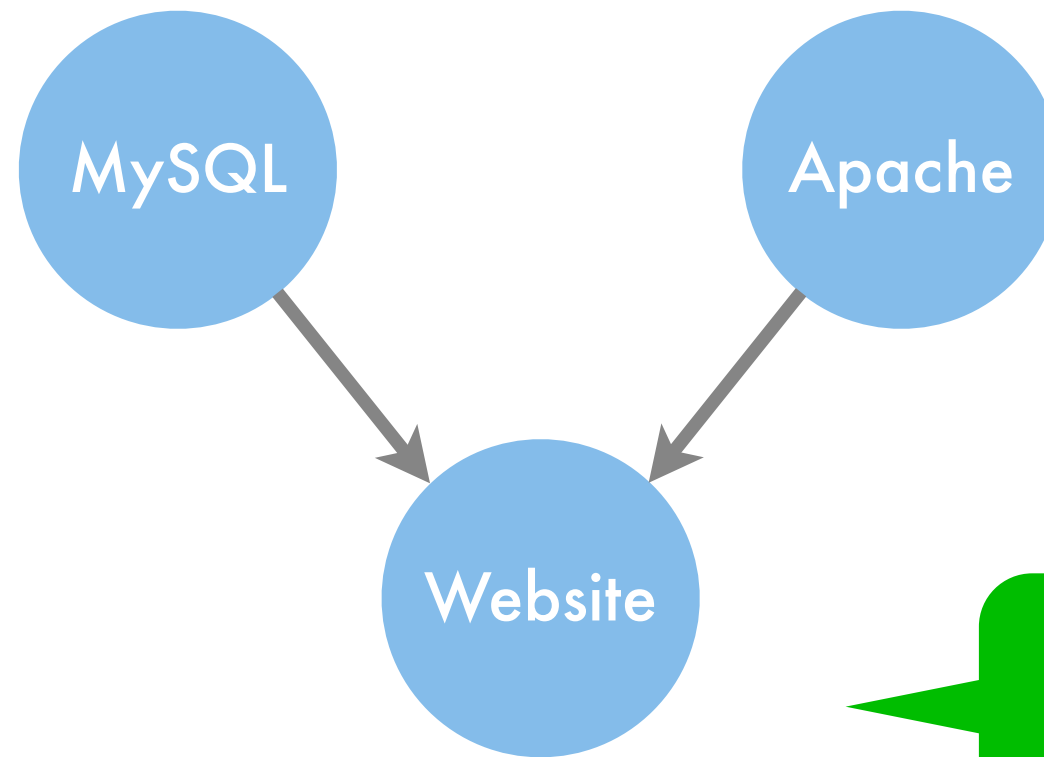
$$p(m, a|w) = \frac{p(w|m, a)p(m)p(a)}{\sum_{m', a'} p(w|m', a')p(m')p(a')}$$

**$a$  and  $m$  are dependent conditioned on  $w$**

... some Web 2.0 service



# ... some Web 2.0 service

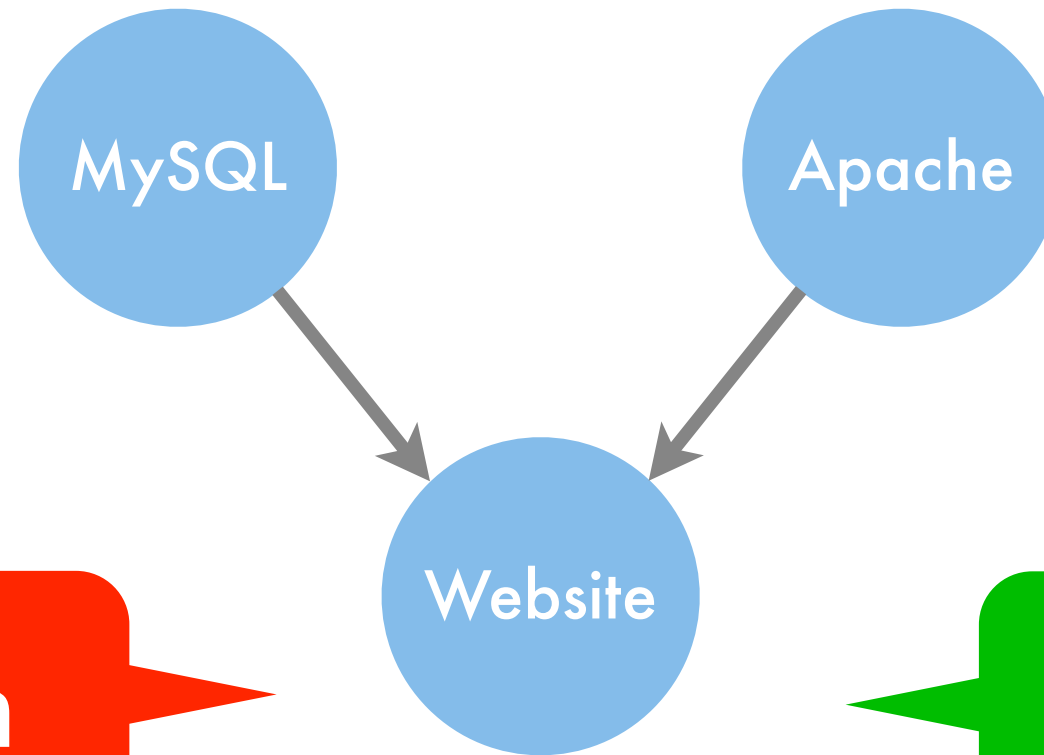


is working

MySQL is working  
Apache is working



# ... some Web 2.0 service



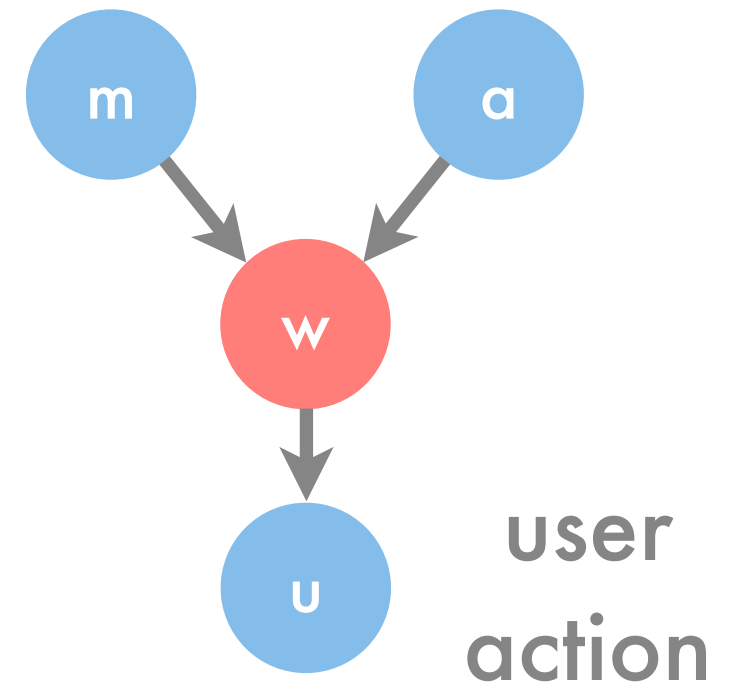
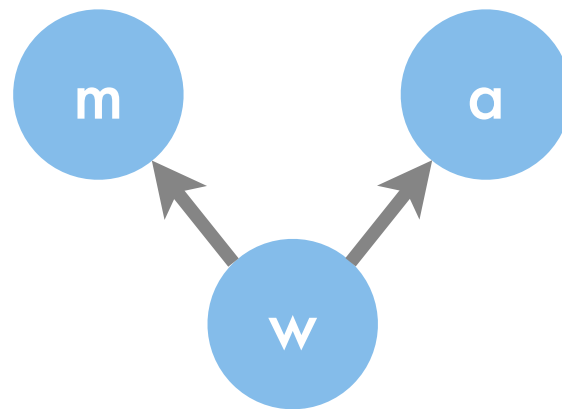
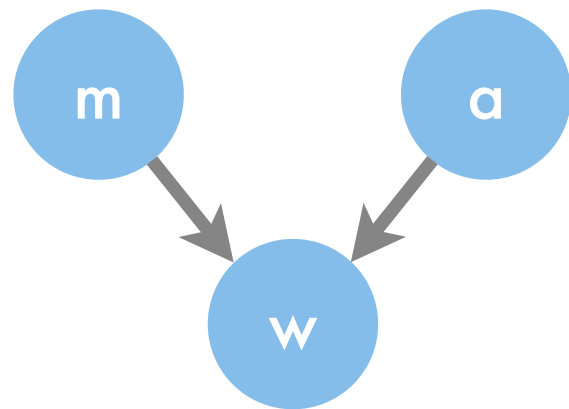
is broken

is working

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

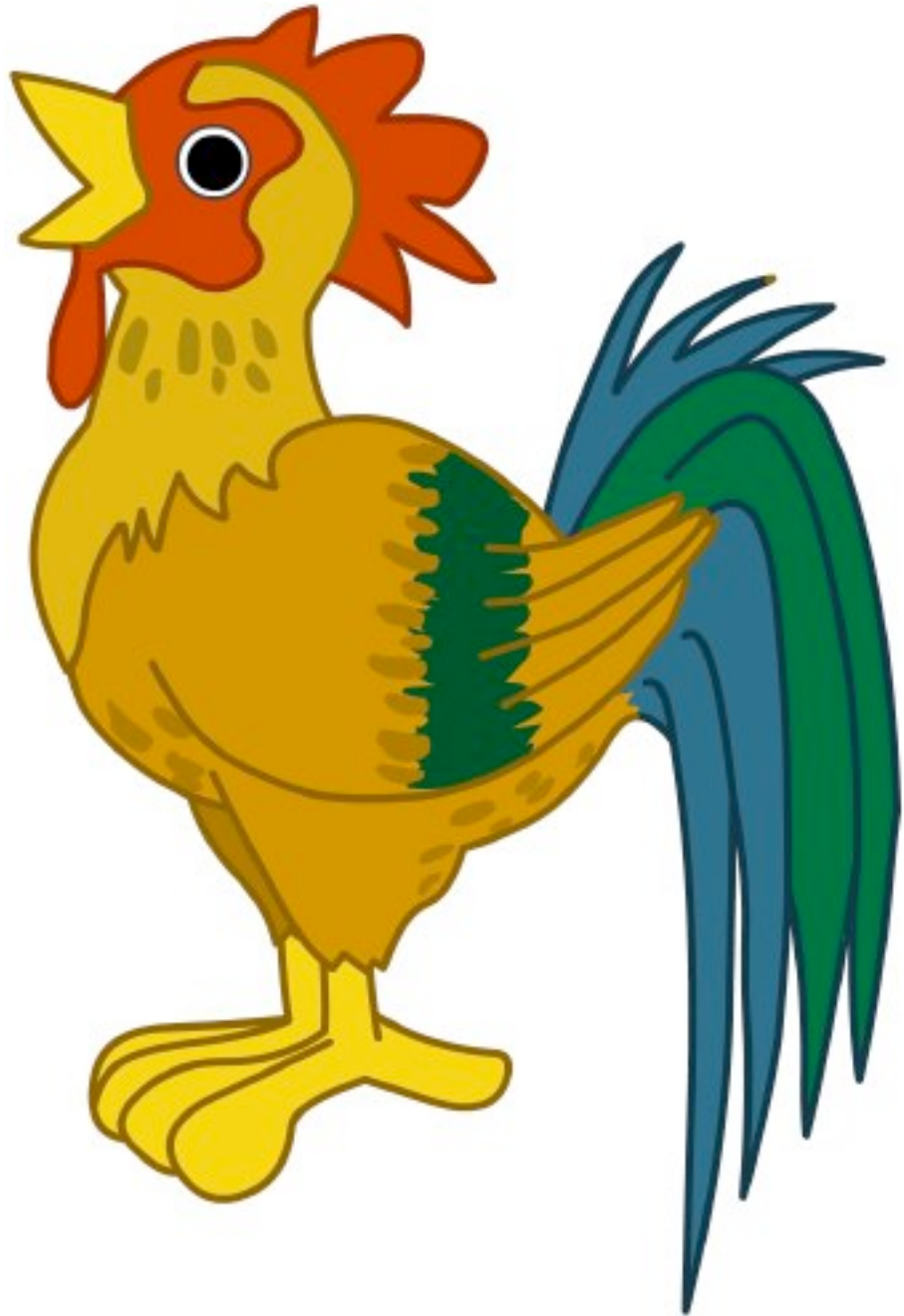
MySQL is working  
Apache is working

# Directed graphical model

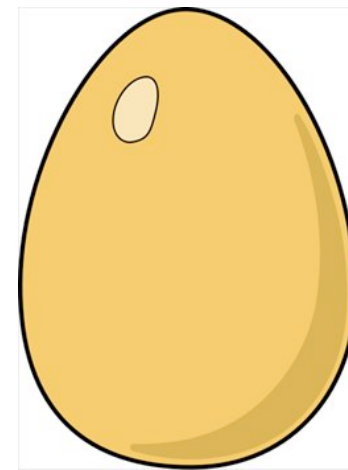
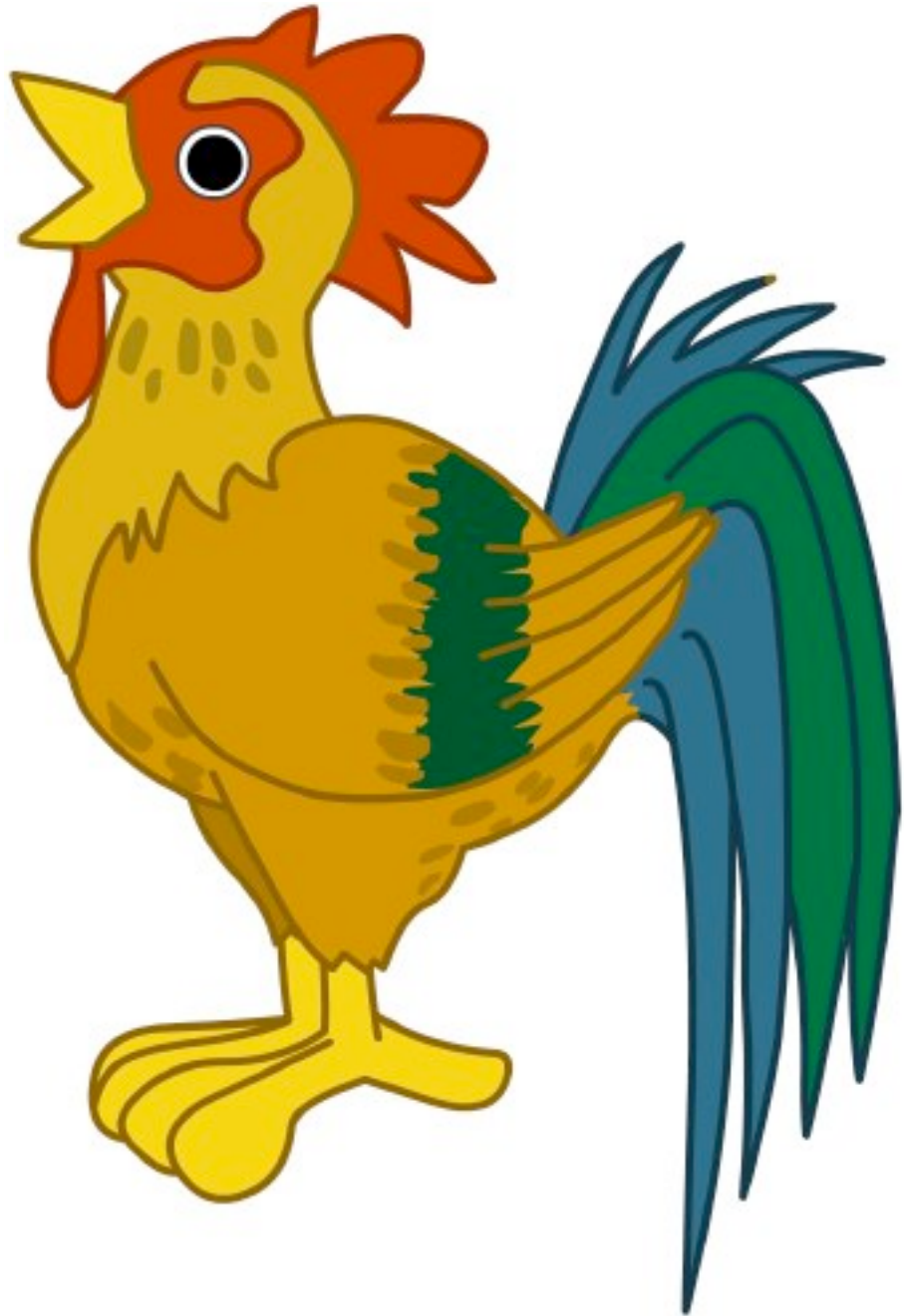


- **Easier estimation**
  - 15 parameters for full joint distribution
  - $1+1+3+1$  for factorizing distribution
- **Causal** relations
- Inference for unobserved variables

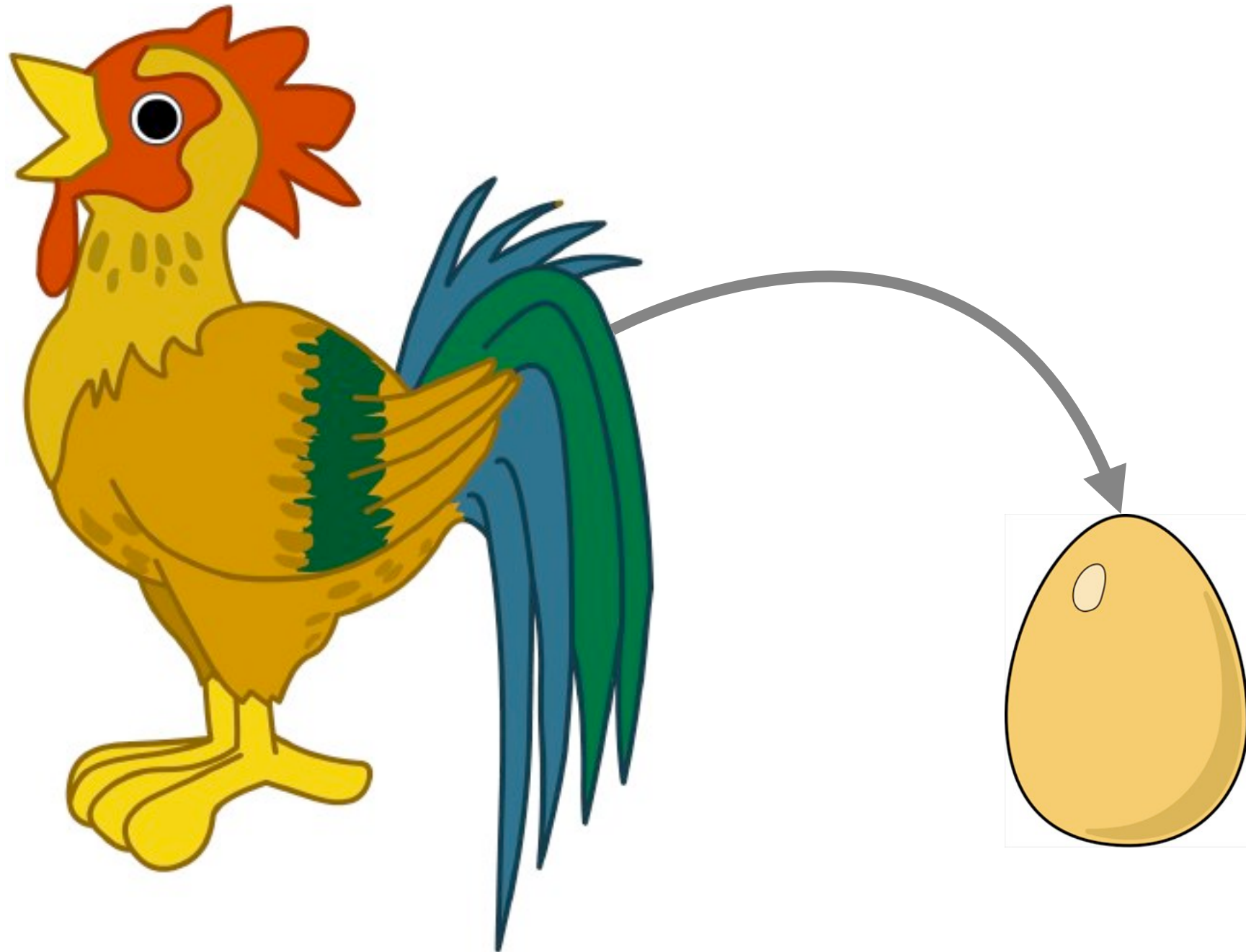
# No loops allowed



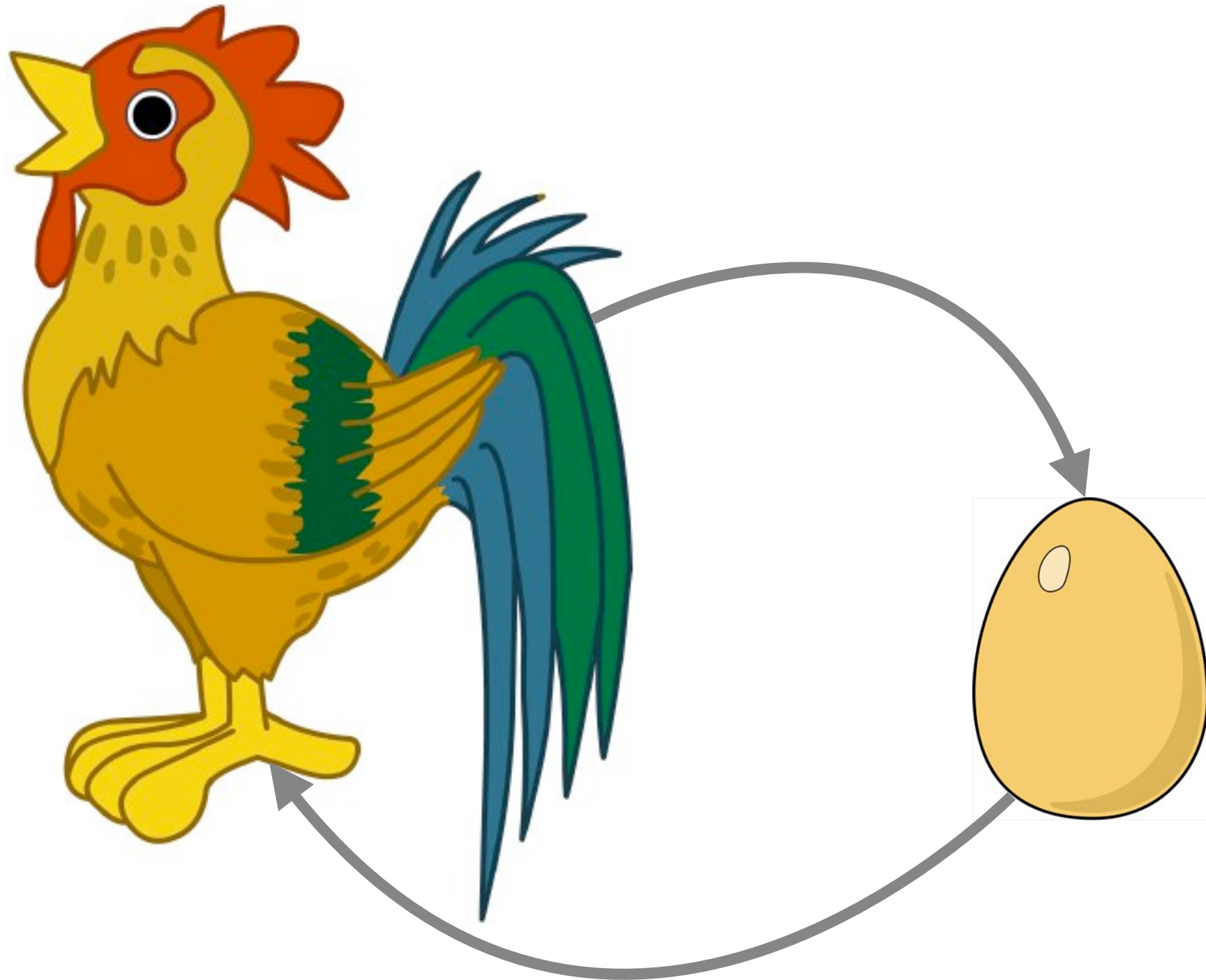
# No loops allowed



# No loops allowed



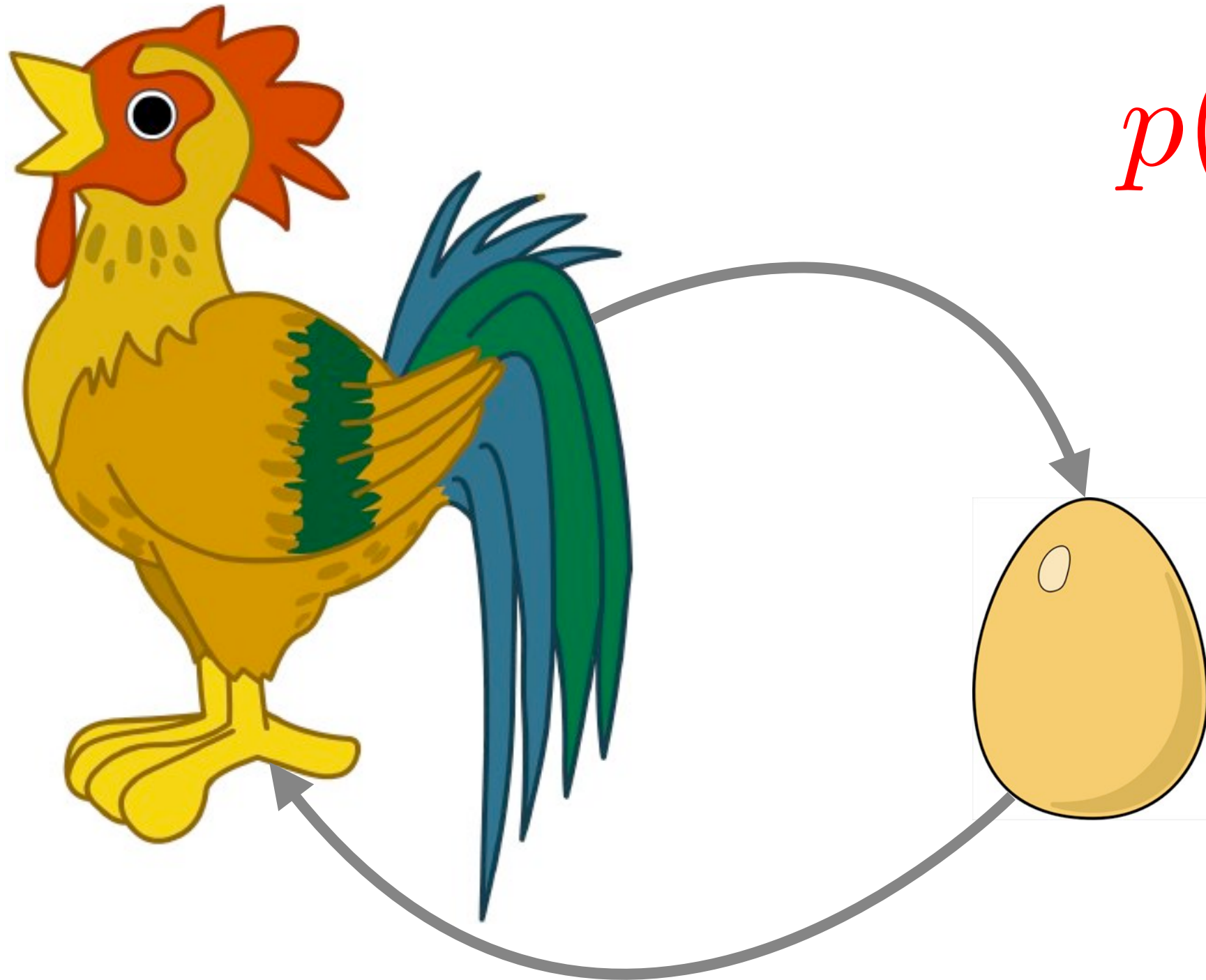
# No loops allowed





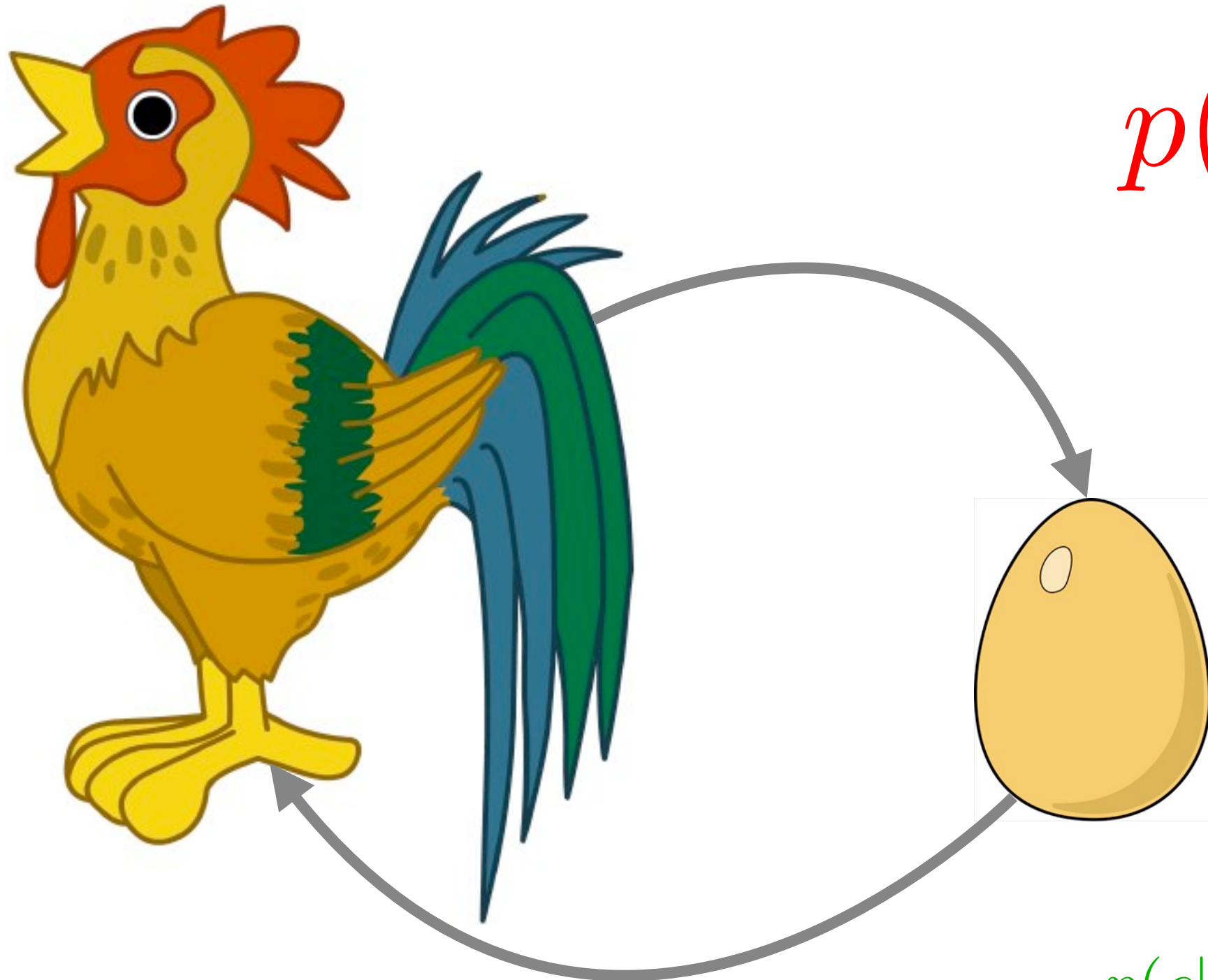
# No loops allowed

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



# No loops allowed

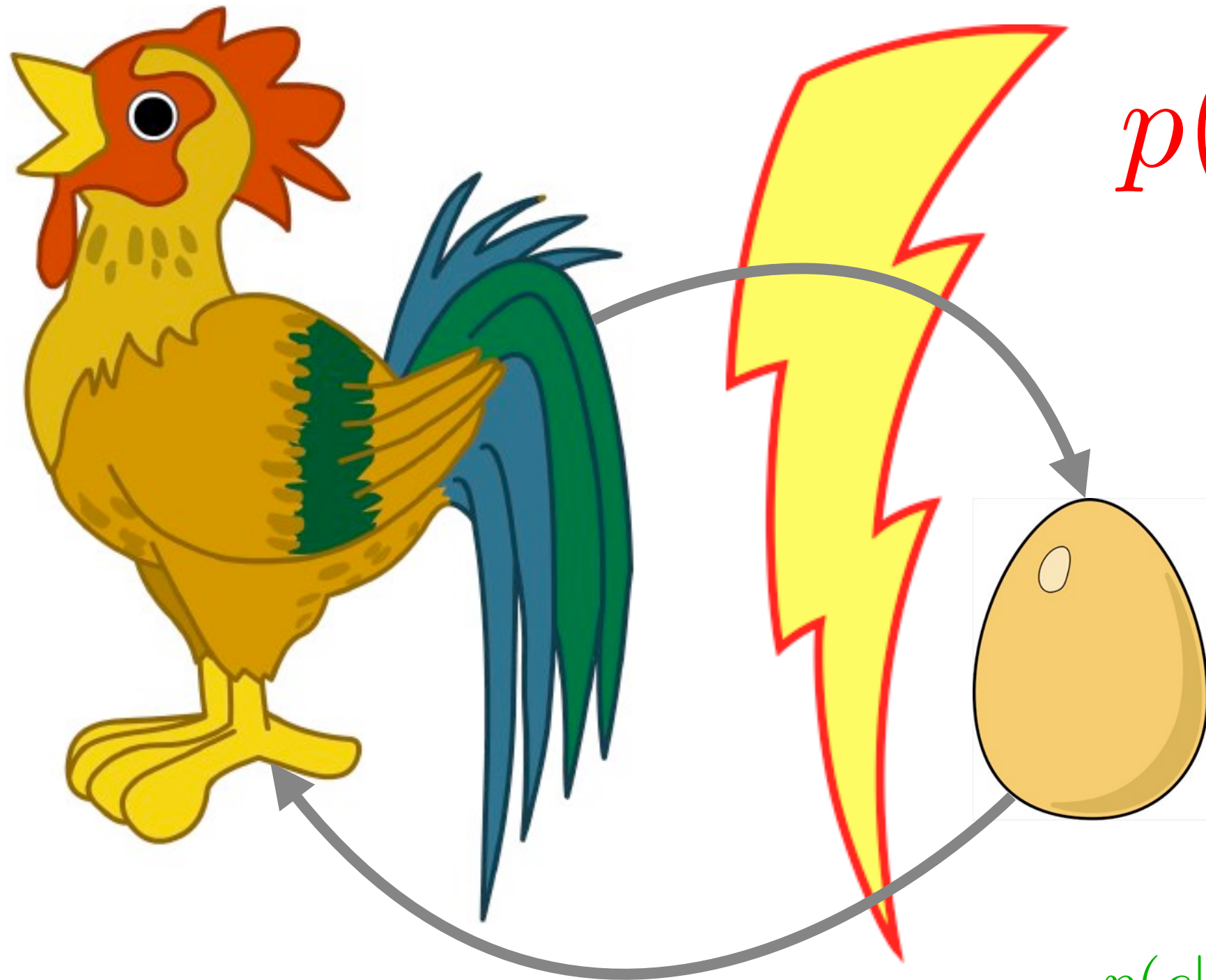
$$p(c|e)p(e|c)$$



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



# No loops allowed



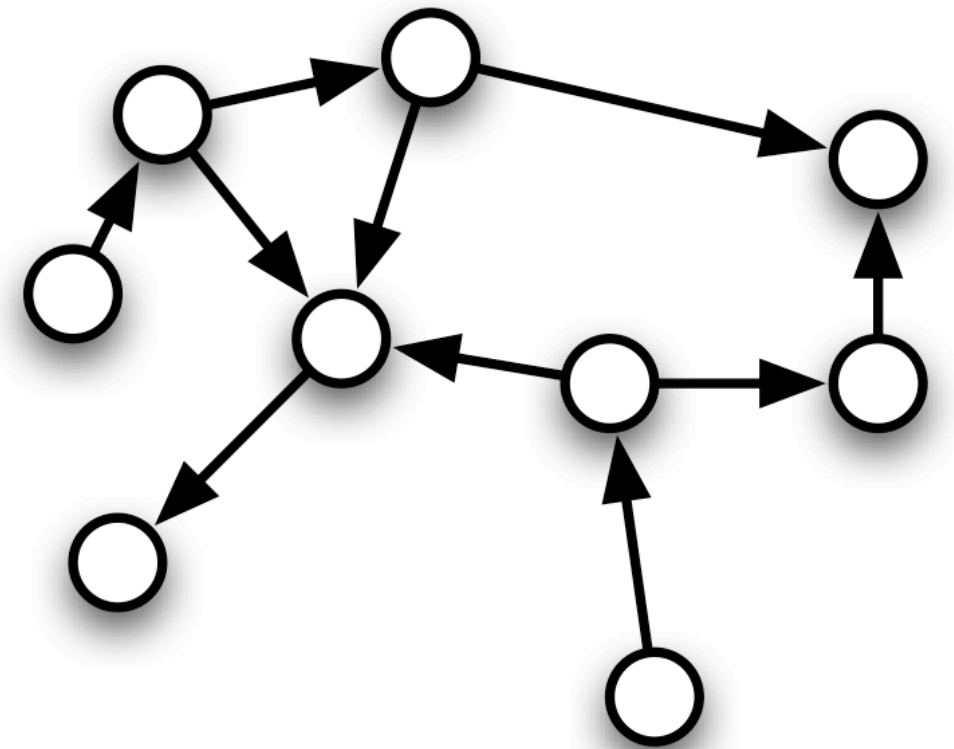
$$p(c|e)p(e|c)$$

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

# Directed Graphical Model

- Joint probability distribution

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



- Parameter estimation

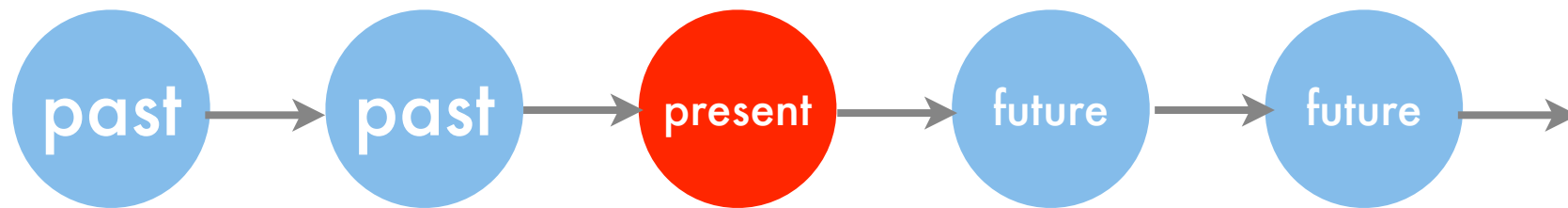
- If  $x$  is fully observed the likelihood breaks up

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

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

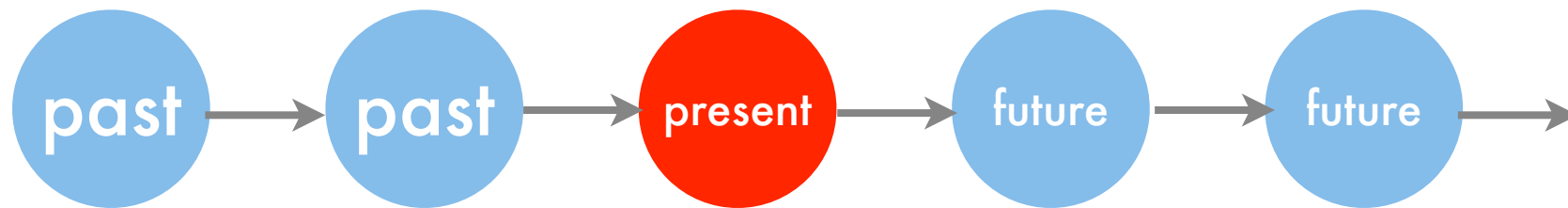
# Chains

## Markov Chain

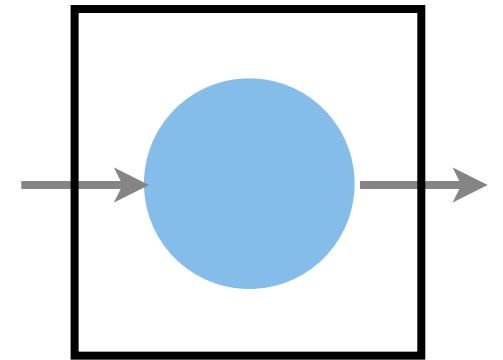


# Chains

Markov Chain

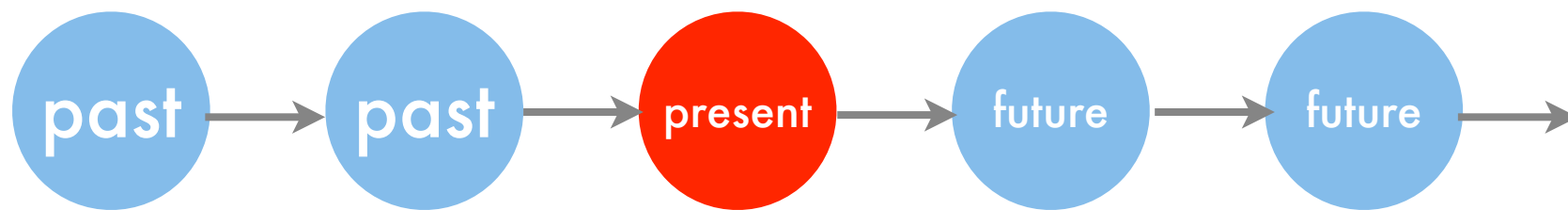


Plate

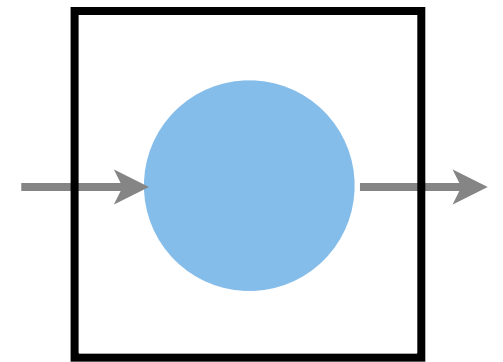


# Chains

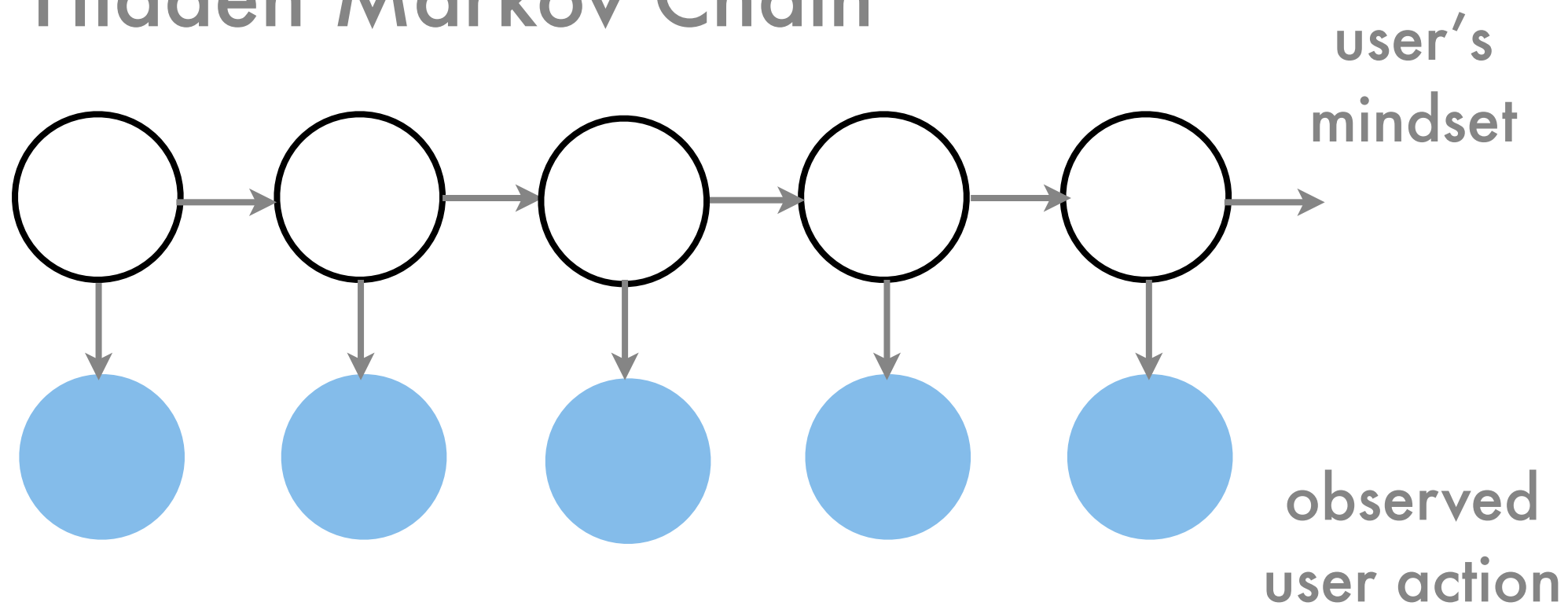
## Markov Chain



## Plate

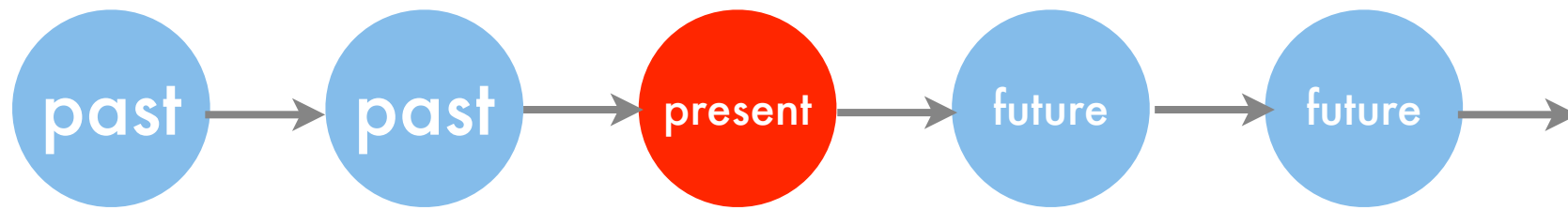


## Hidden Markov Chain

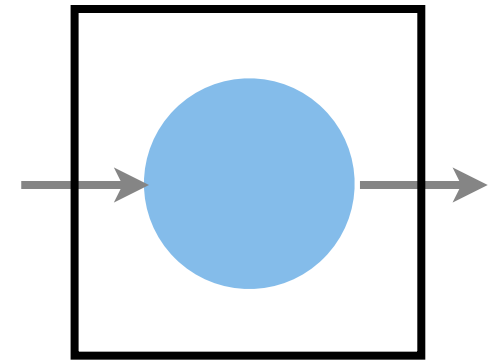


# Chains

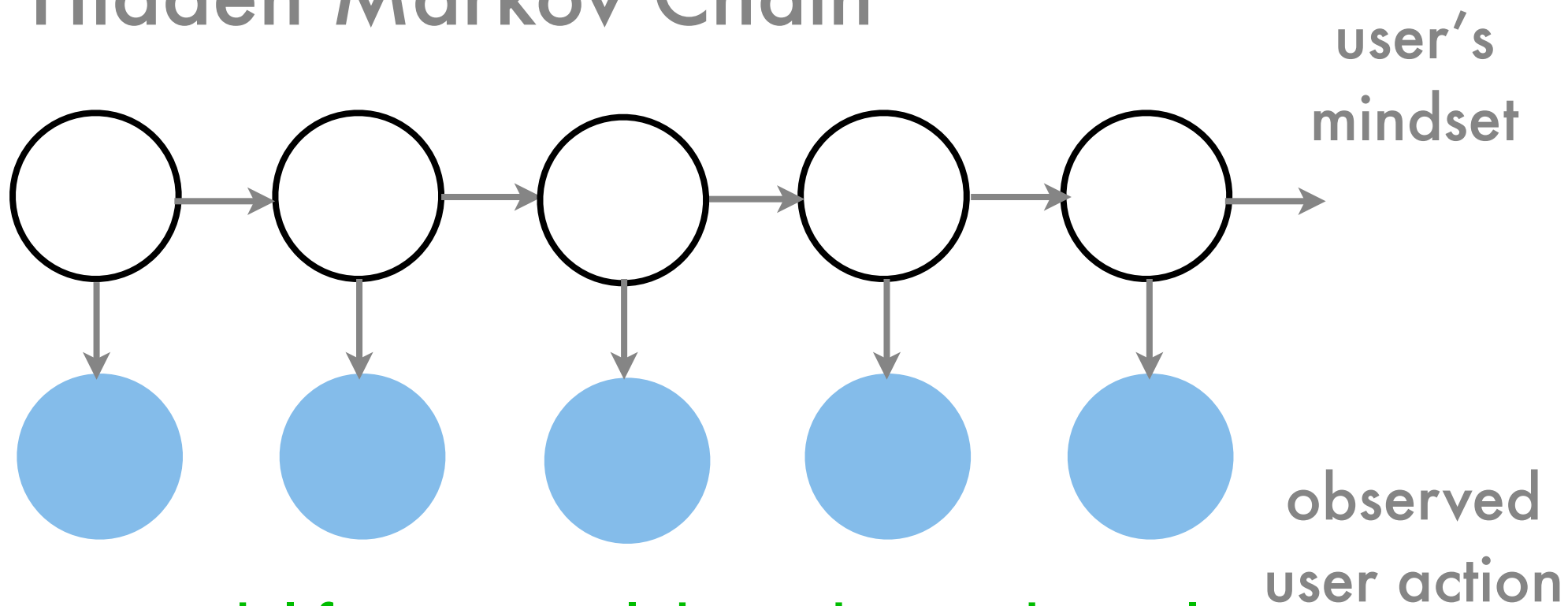
## Markov Chain



## Plate



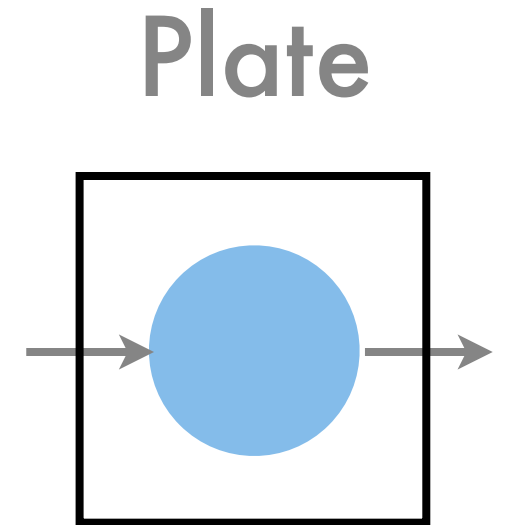
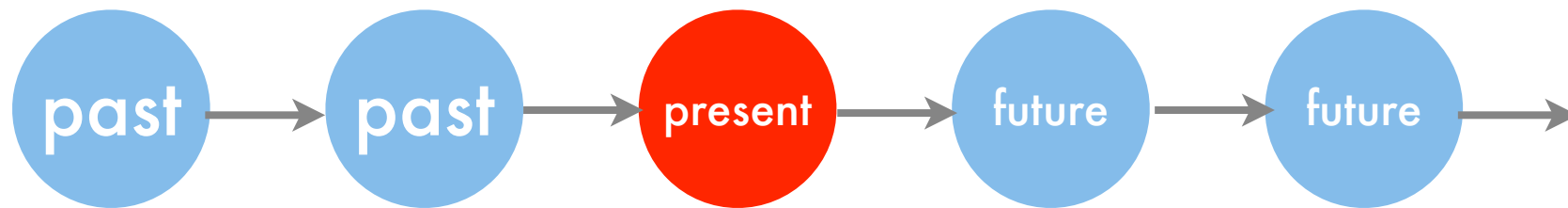
## Hidden Markov Chain



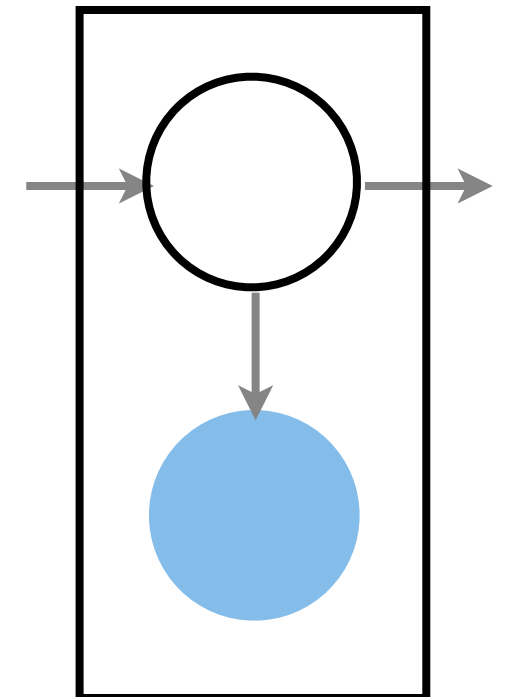
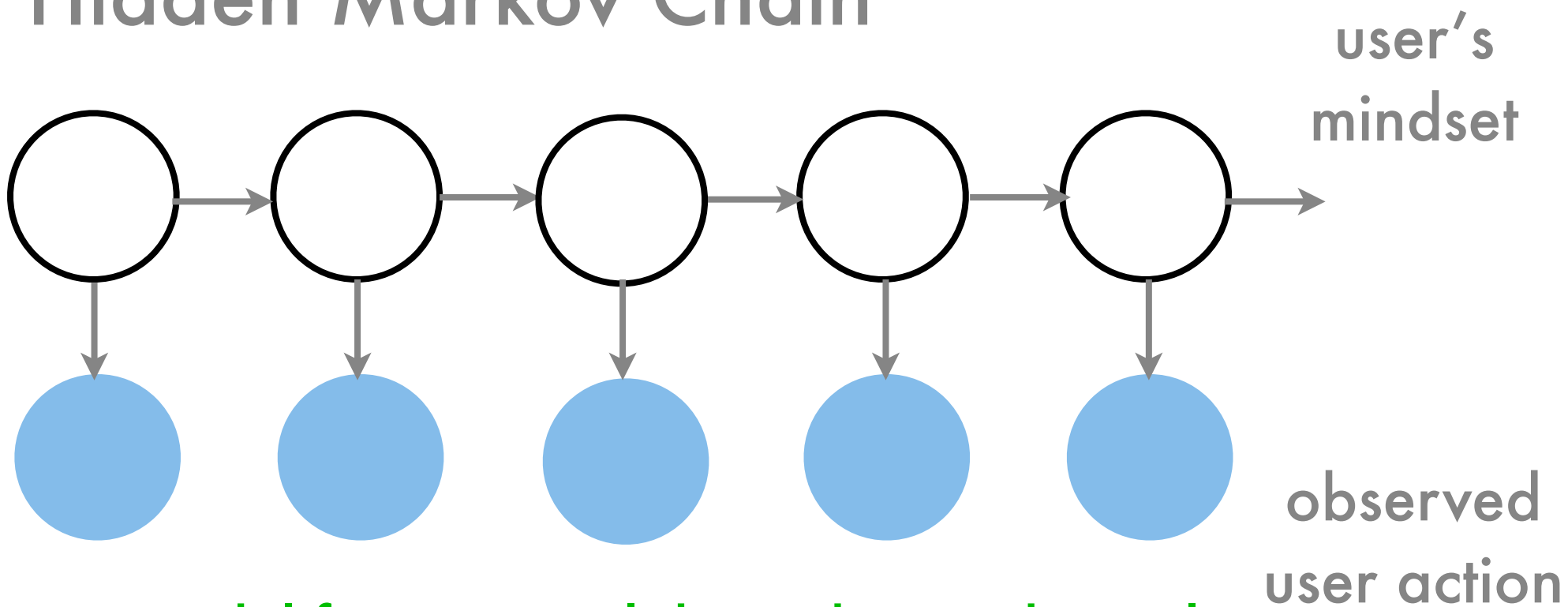
user model for traversal through search results

# Chains

## Markov Chain



## Hidden Markov Chain



user model for traversal through search results

# Chains

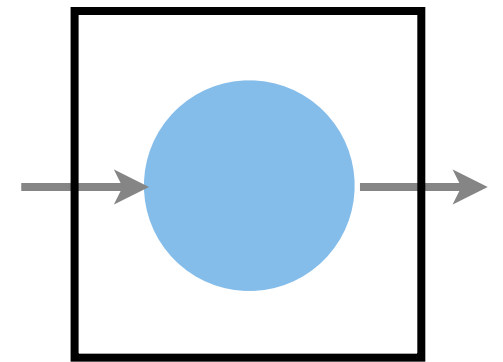
## Markov Chain

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

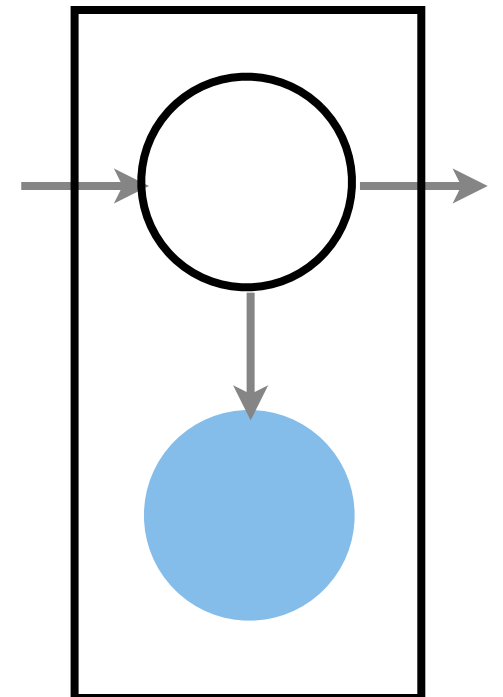
## Hidden Markov Chain

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

Plate



user's  
mindset

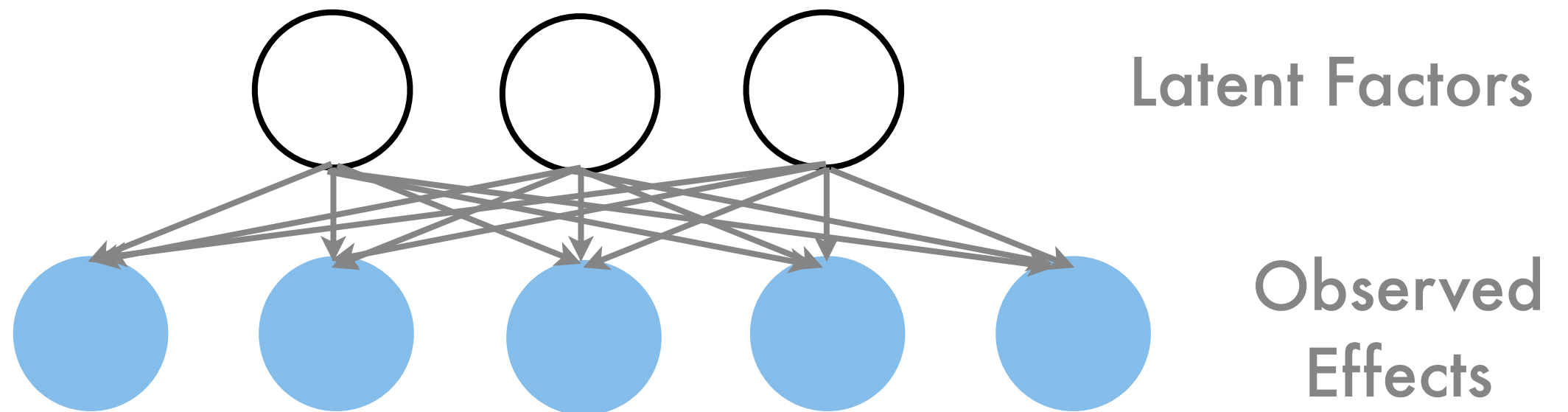


observed  
user action

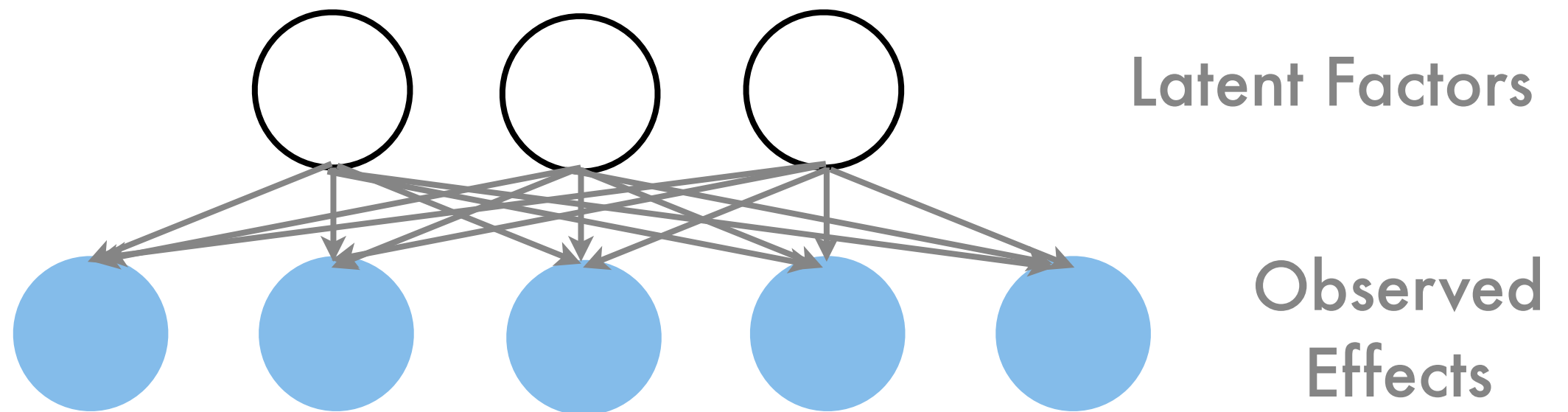
user model for traversal through search results



# Factor Graphs

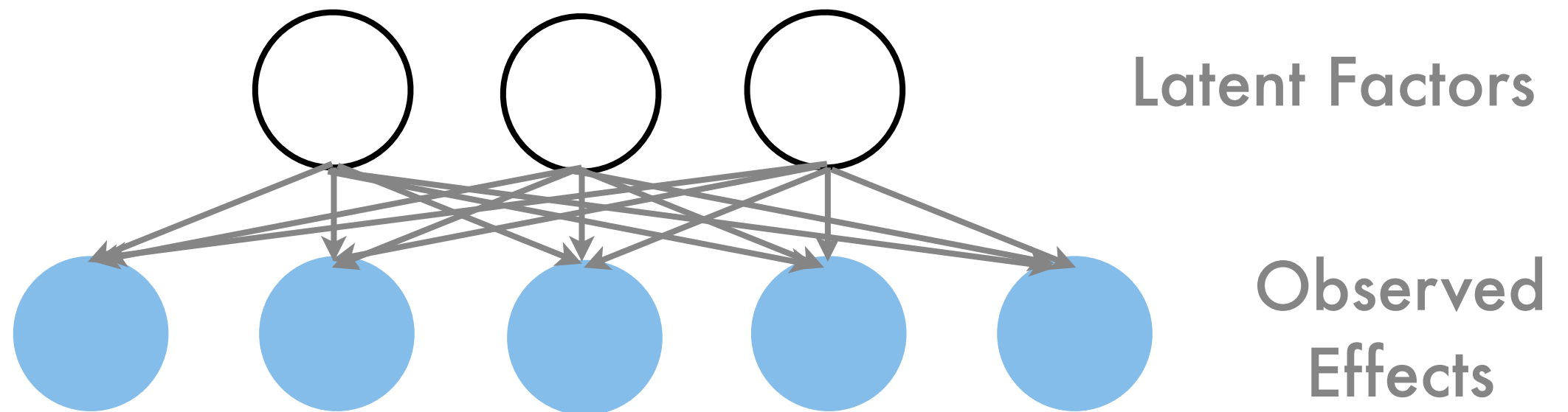


# Factor Graphs



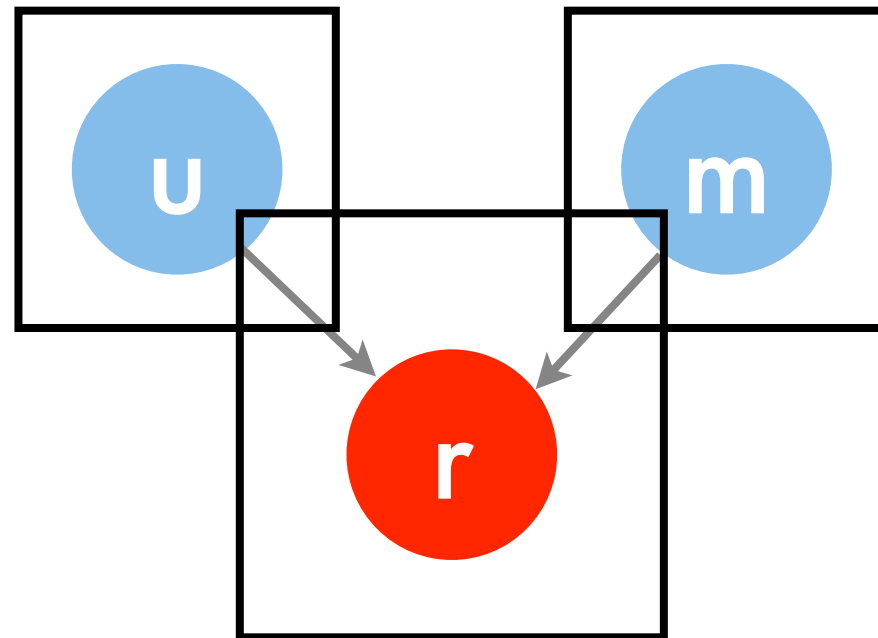
- **Observed effects**  
Click behavior, queries, watched news, emails

# Factor Graphs

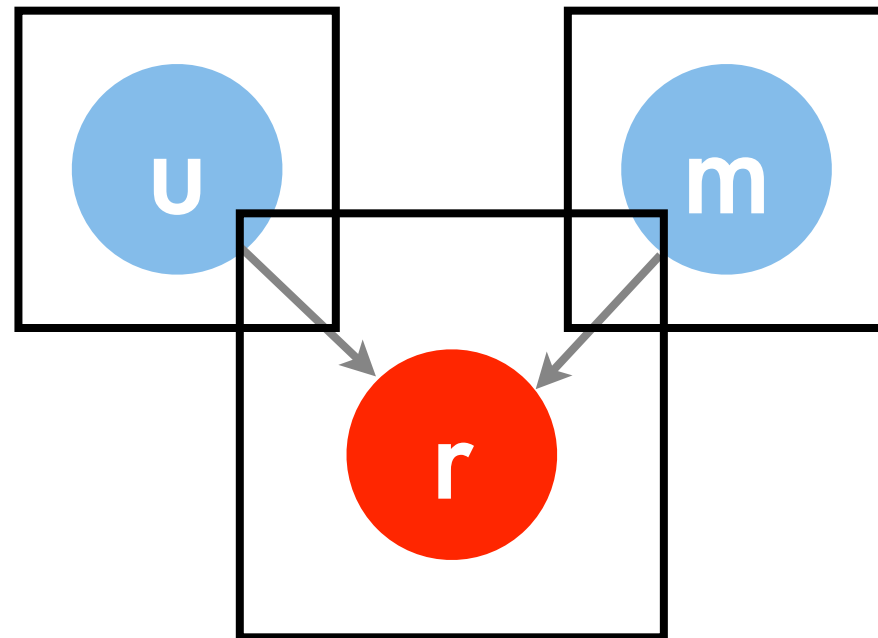


- **Observed effects**  
Click behavior, queries, watched news, emails
- **Latent factors**  
User profile, news content, hot keywords, social connectivity graph, events

# Recommender Systems

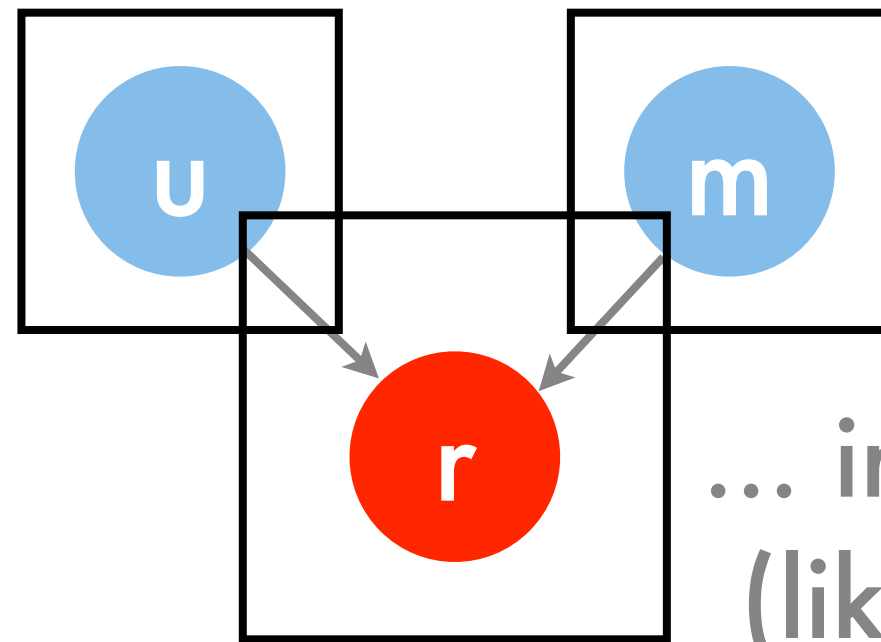


# Recommender Systems



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

# Recommender Systems

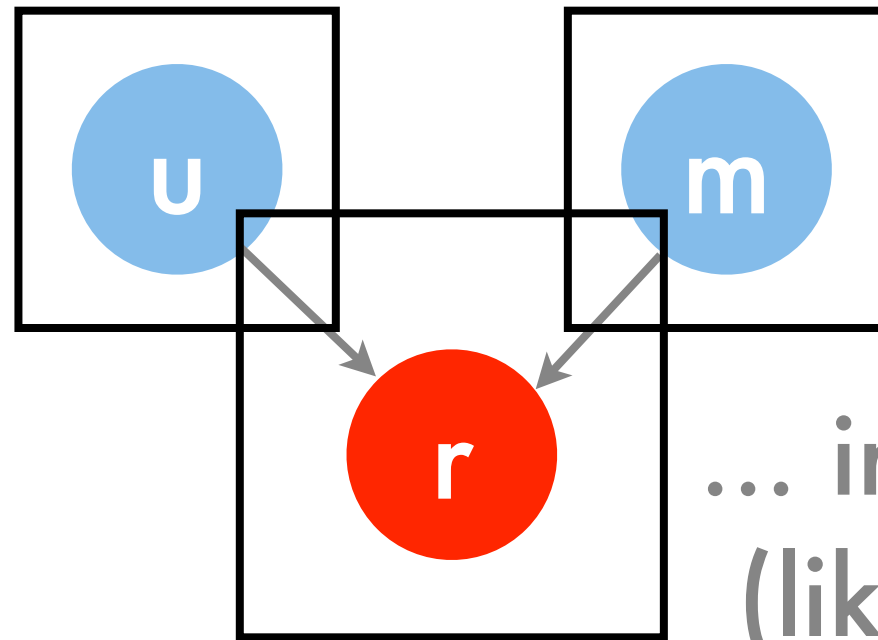


... intersecting plates ...  
(like nested for loops)

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

# Recommender Systems

news,  
SearchMonkey  
answers  
social  
ranking  
OMG  
personals



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

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

# Challenges



your job



my job



# Challenges

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



your job



my job

# Challenges

- How to design models
  - Common (engineering) sense
  - Computational tractability
- Dependency analysis
  - Bayes ball (not in this lecture)





your job



my job

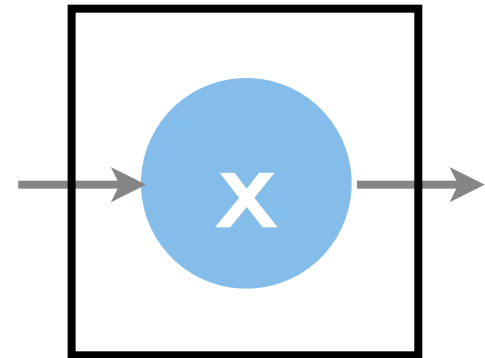
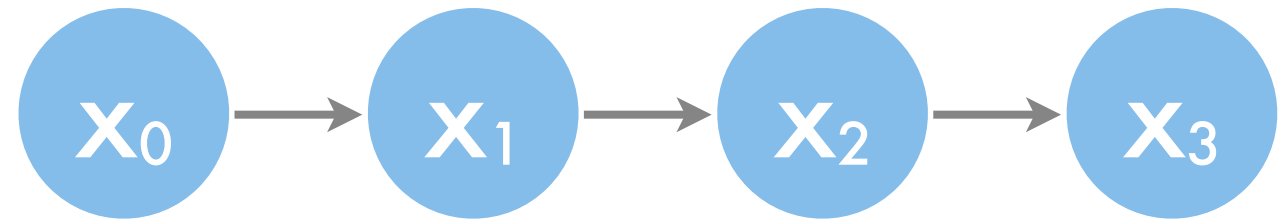
# Challenges

- How to design models
  - Common (engineering) sense  your job
  - Computational tractability  my job
- Dependency analysis
  - Bayes ball (not in this lecture)
- Inference
  - Easy for fully observed situations
  - Many algorithms if not fully observed
  - Dynamic programming / message passing

# Dynamic Programming 101

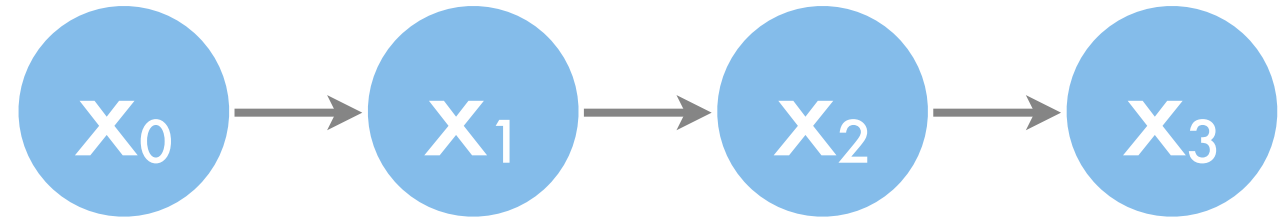
# Chains

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

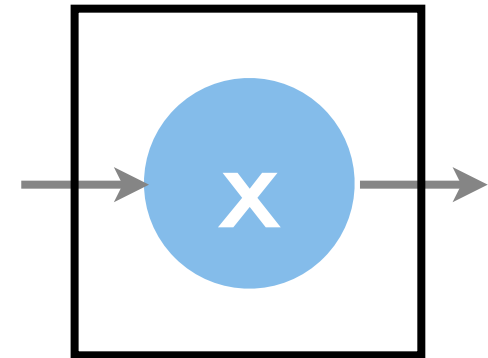


# Chains

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

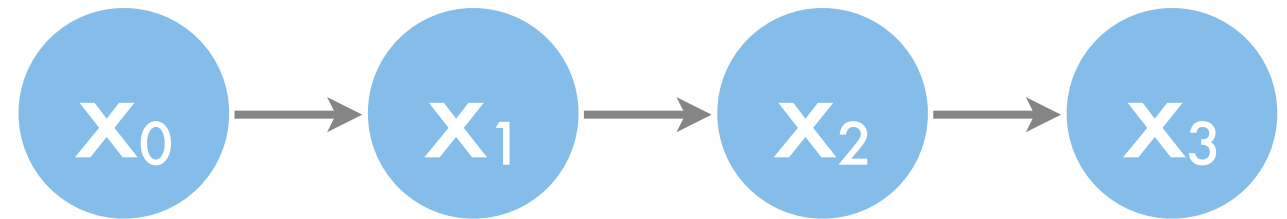


$$p(x_i) = \sum_{x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \underbrace{p(x_0)}_{:=l_0(x_0)} \prod_{j=1}^n p(x_j | x_{j-1})$$



# Chains

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

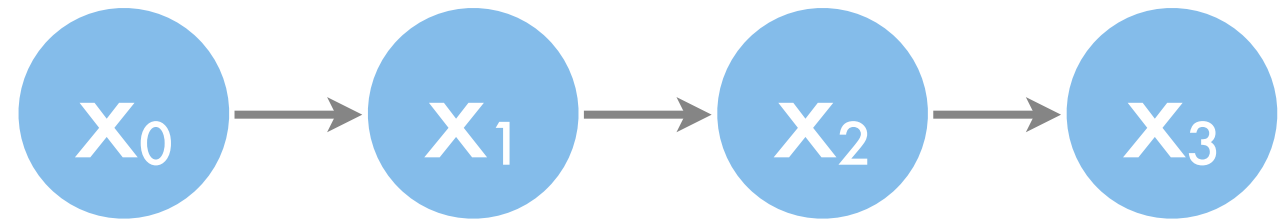


$$p(x_i) = \sum_{x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \underbrace{p(x_0)}_{:=l_0(x_0)} \prod_{j=1}^n p(x_j | x_{j-1})$$

$$= \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \underbrace{\sum_{x_0} [l_0(x_0) p(x_1 | x_0)]}_{:=l_1(x_1)} \prod_{j=2}^n p(x_j | x_{j-1}) \rightarrow \boxed{x}$$

# Chains

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



$$p(x_i) = \sum_{x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \underbrace{p(x_0)}_{:=l_0(x_0)} \prod_{j=1}^n p(x_j | x_{j-1})$$

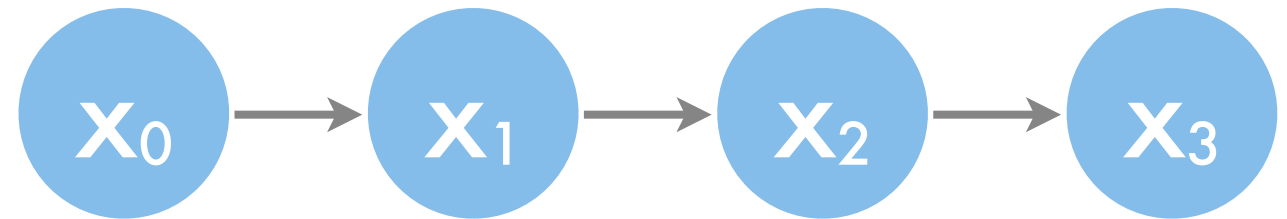
$$= \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \underbrace{\sum_{x_0} [l_0(x_0) p(x_1 | x_0)]}_{:=l_1(x_1)} \prod_{j=2}^n p(x_j | x_{j-1}) \rightarrow \boxed{x}$$

$$= \sum_{x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n} \underbrace{\sum_{x_1} [l_1(x_1) p(x_2 | x_1)]}_{:=l_2(x_2)} \prod_{j=3}^n p(x_j | x_{j-1})$$



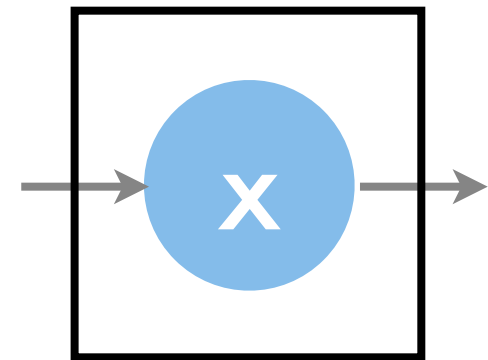
# Chains

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



$$p(x_i) = l_i(x_i) \sum_{x_{i+1} \dots x_n} \prod_{j=i}^{n-1} p(x_{j+1} | x_j)$$

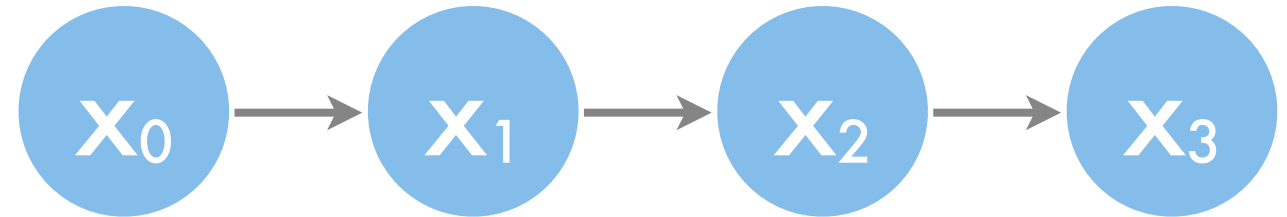
$$= l_i(x_i) \sum_{x_{i+1} \dots x_{n-1}} \prod_{j=i}^{n-2} p(x_{j+1} | x_j) \underbrace{\sum_{x_n} p(x_n | x_{n-1})}_{:=r_{n-1}(x_{n-1})}$$



$$= l_i(x_i) \sum_{x_{i+1} \dots x_{n-2}} \prod_{j=i}^{n-3} p(x_{j+1} | x_j) \underbrace{\sum_{x_{n-1}} p(x_{n-1} | x_{n-2}) r_{n-1}(x_{n-1})}_{:=r_{n-2}(x_{n-2})}$$

# Chains

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



- **Forward recursion**

$$l_0(x_0) := p(x_0) \text{ and } l_i(x_i) := \sum_{x_{i-1}} l_{i-1}(x_{i-1}) p(x_i | x_{i-1})$$

- **Backward recursion**

$$r_n(x_n) := 1 \text{ and } r_i(x_i) := \sum_{x_{i+1}} r_{i+1}(x_{i+1}) p(x_{i+1} | x_i)$$

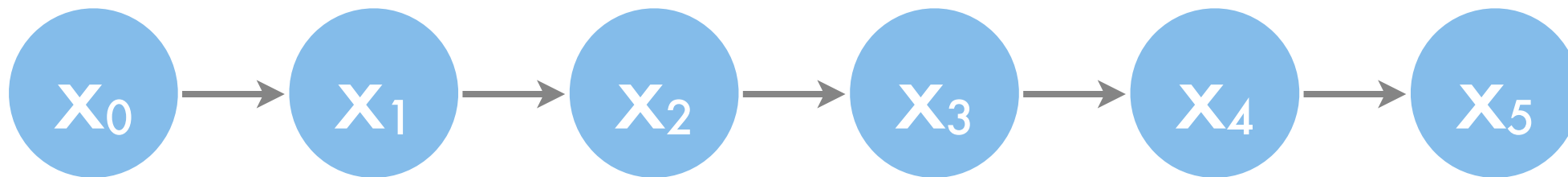
- **Marginalization & conditioning**

$$p(x_i) = l_i(x_i) r_i(x_i)$$

$$p(x_{-i} | x_i) = \frac{p(x)}{p(x_i)}$$

$$p(x_i, x_{i+1}) = l_i(x_i) p(x_{i+1} | x_i) r_i(x_{i+1})$$

# Chains



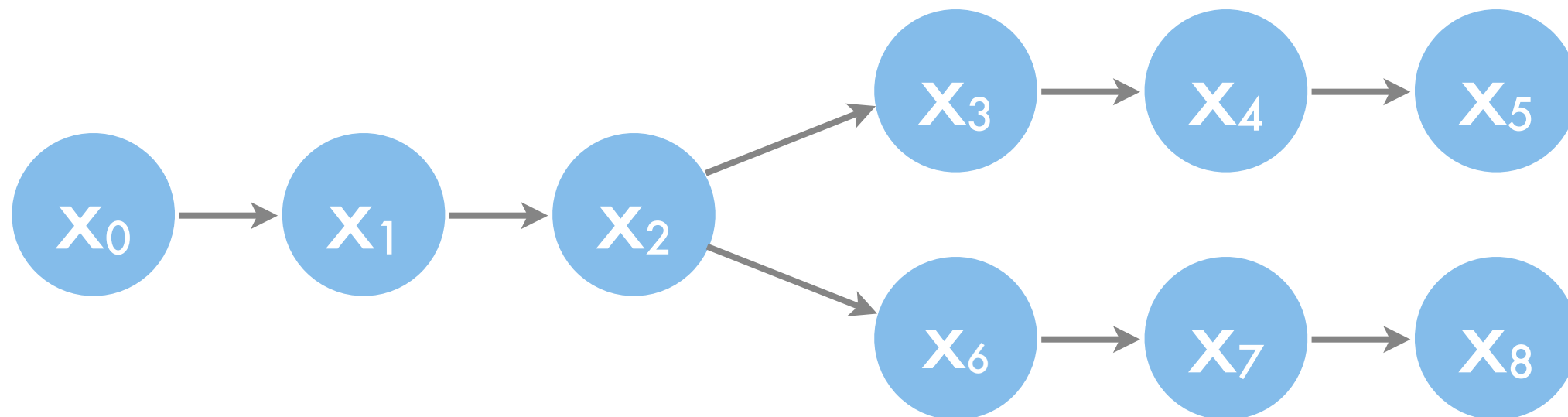
- Send forward messages starting from left node

→ 
$$m_{i-1 \rightarrow i}(x_i) = \sum_{x_{i-1}} m_{i-2 \rightarrow i-1}(x_{i-1}) f(x_{i-1}, x_i)$$

- Send backward messages starting from right node

$$m_{i+1 \rightarrow i}(x_i) = \sum_{x_{i+1}} m_{i+2 \rightarrow i+1}(x_{i+1}) f(x_i, x_{i+1})$$
 ←

# Trees



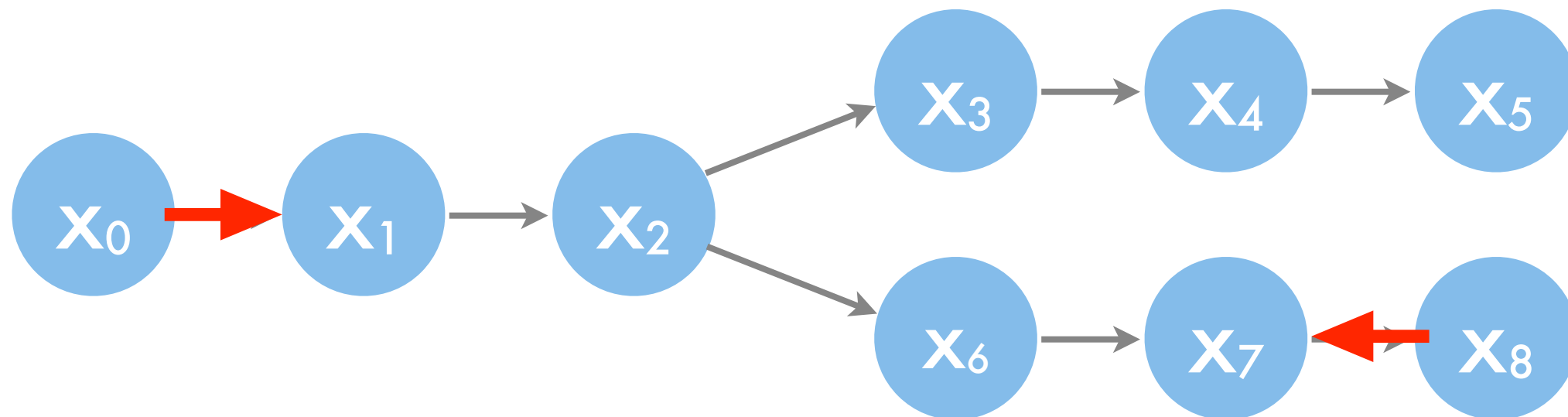
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$

# Trees



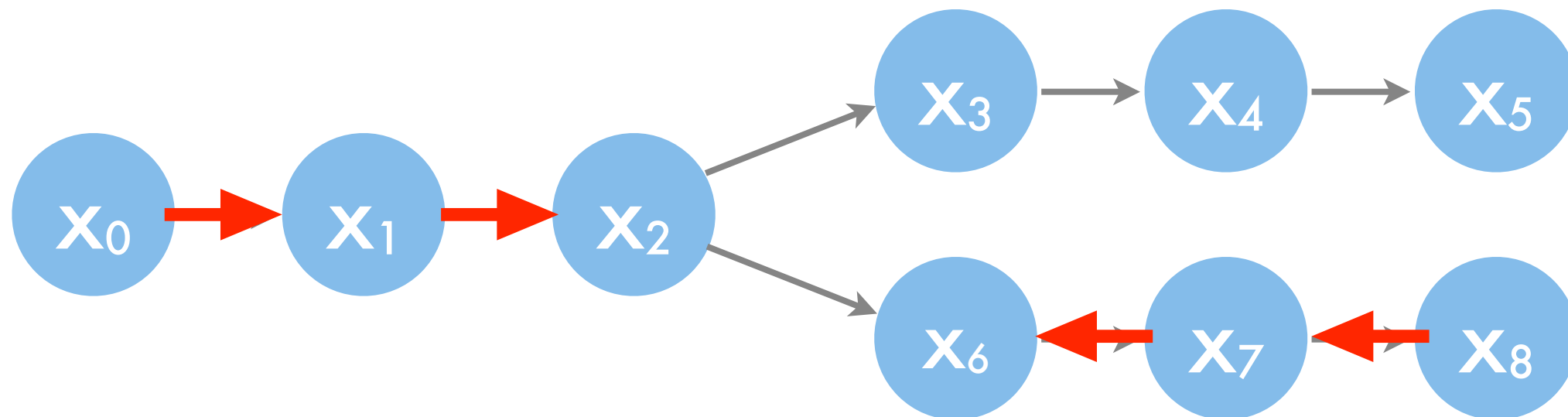
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$

# Trees



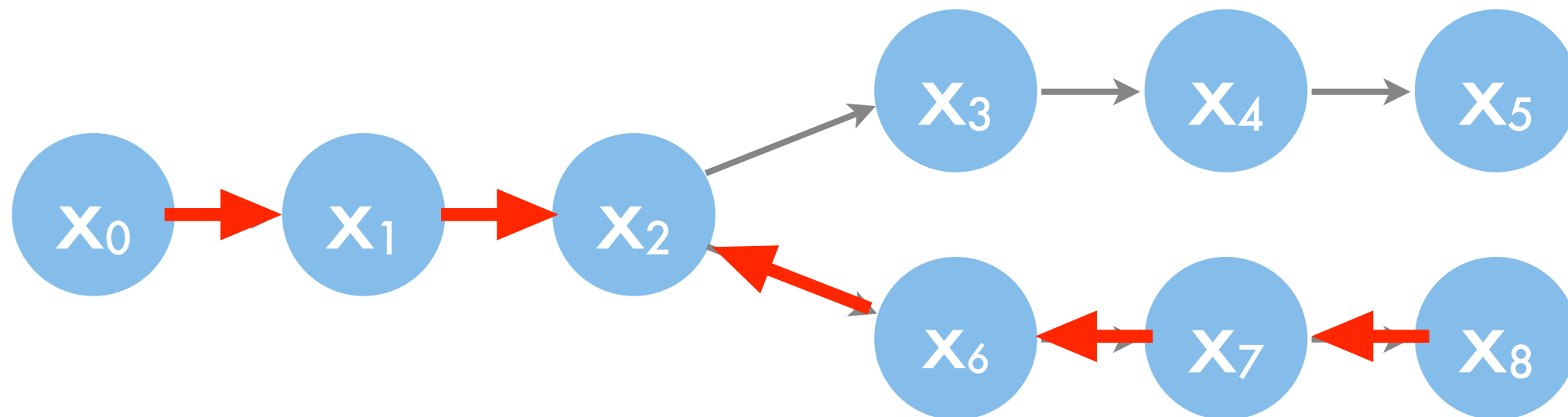
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$

# Trees



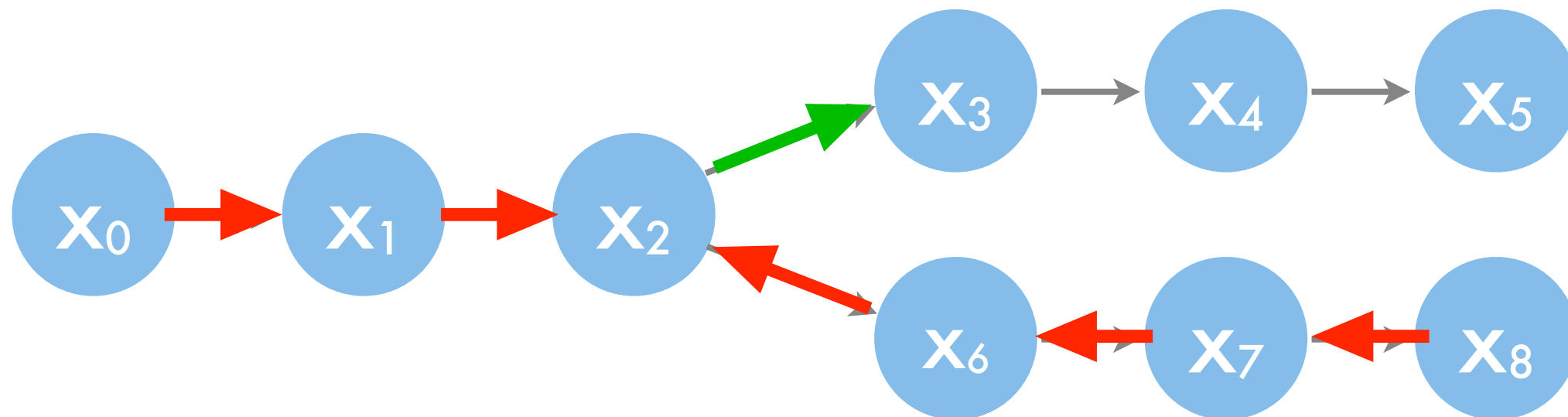
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$

# Trees



- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

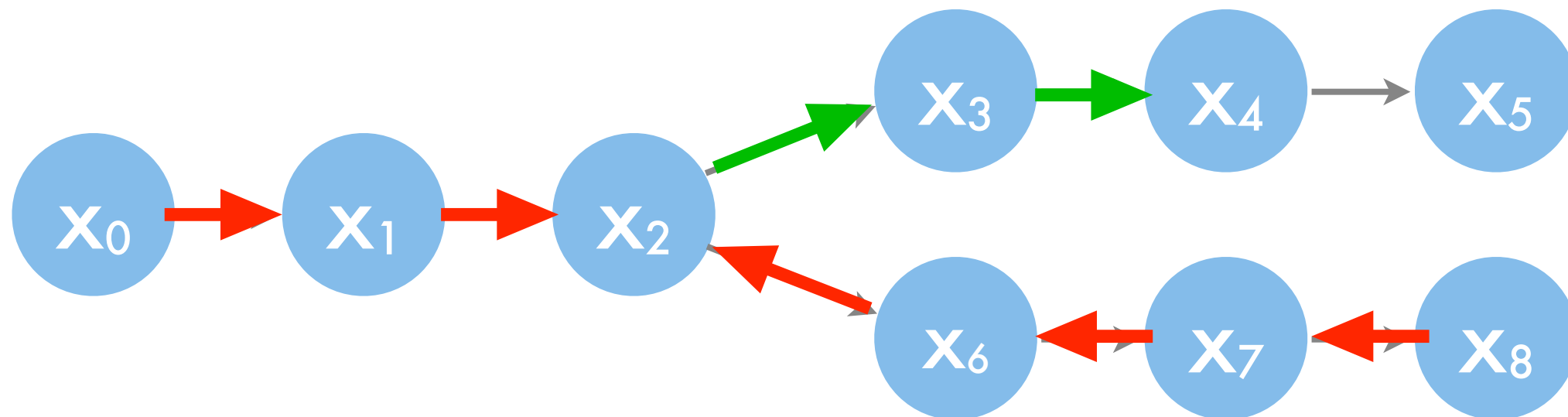
$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$



# Trees



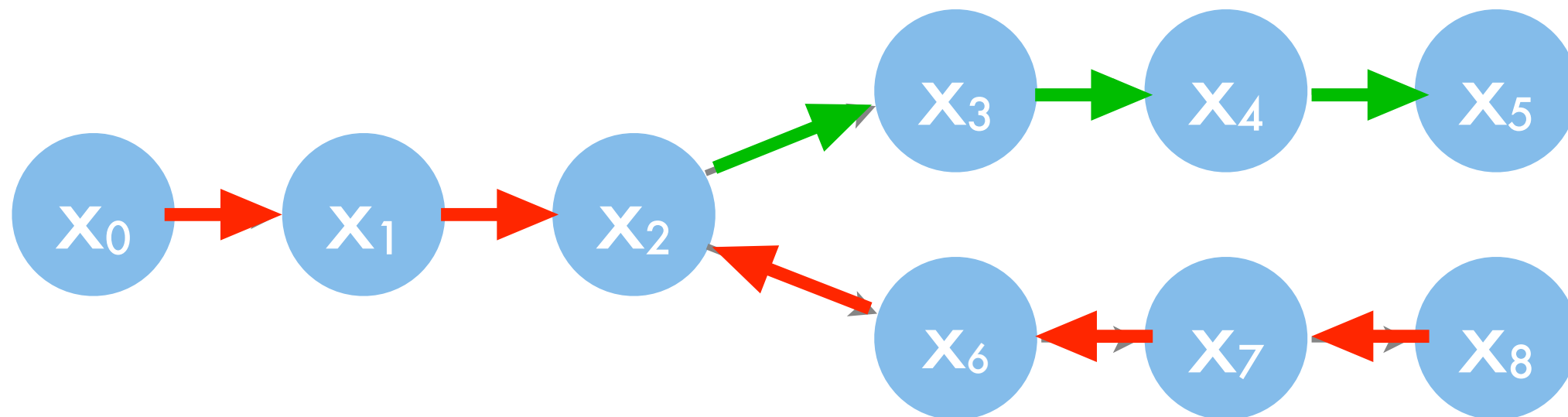
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$

# Trees



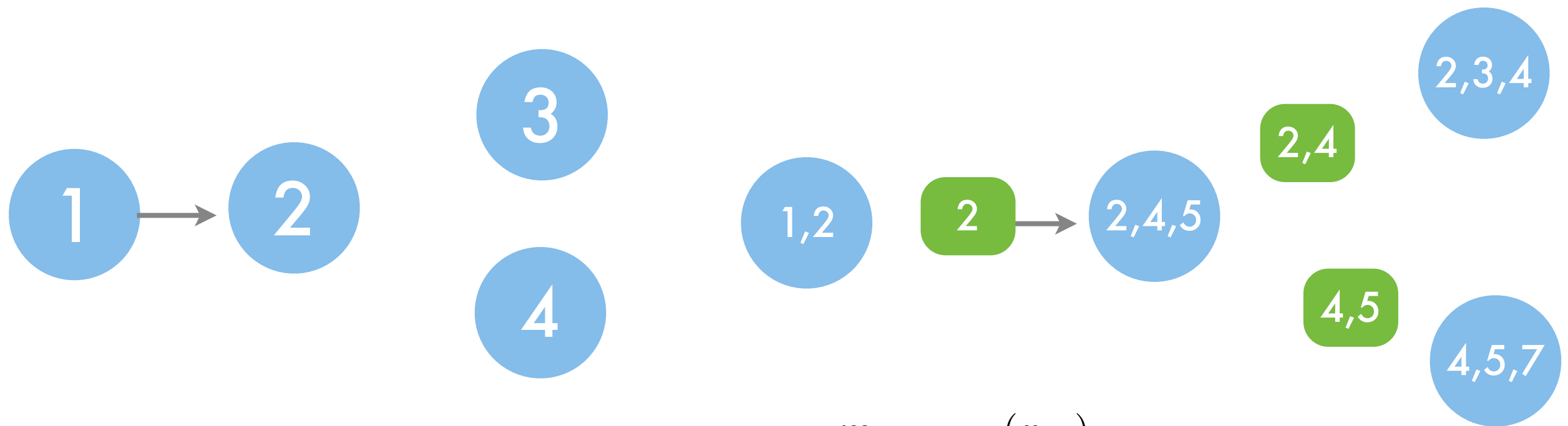
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use ...

$$m_{2 \rightarrow 3}(x_3) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_2, x_3)$$

$$m_{2 \rightarrow 6}(x_6) = \sum_{x_2} m_{1 \rightarrow 2}(x_2) m_{3 \rightarrow 2}(x_2) f(x_2, x_6)$$

$$m_{2 \rightarrow 1}(x_1) = \sum_{x_2} m_{3 \rightarrow 2}(x_2) m_{6 \rightarrow 2}(x_2) f(x_1, x_2)$$

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

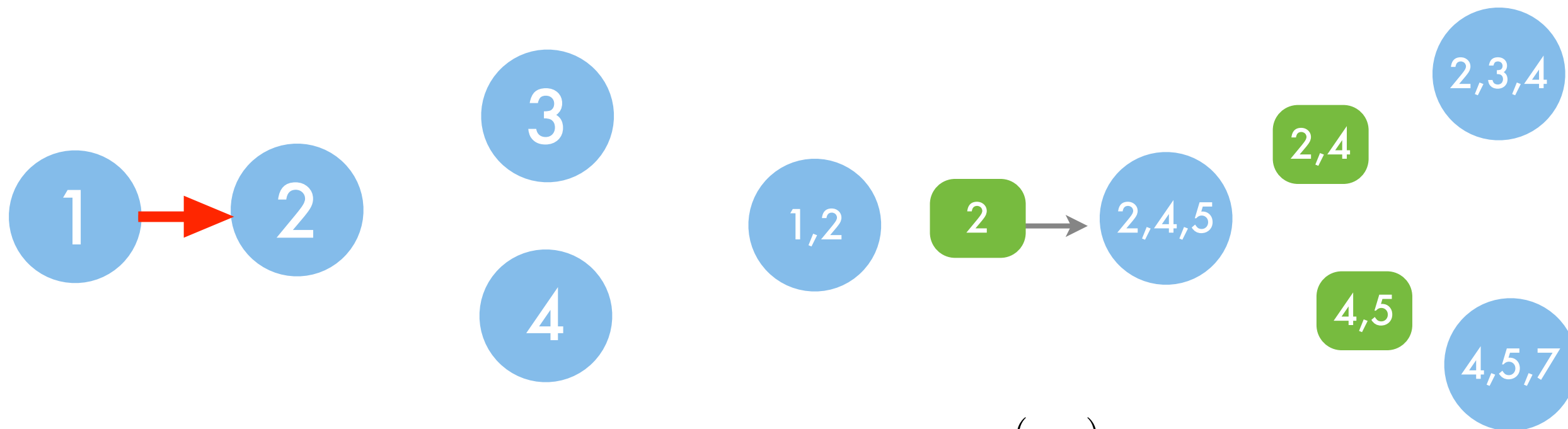
clique potential

$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique potential

separator set

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

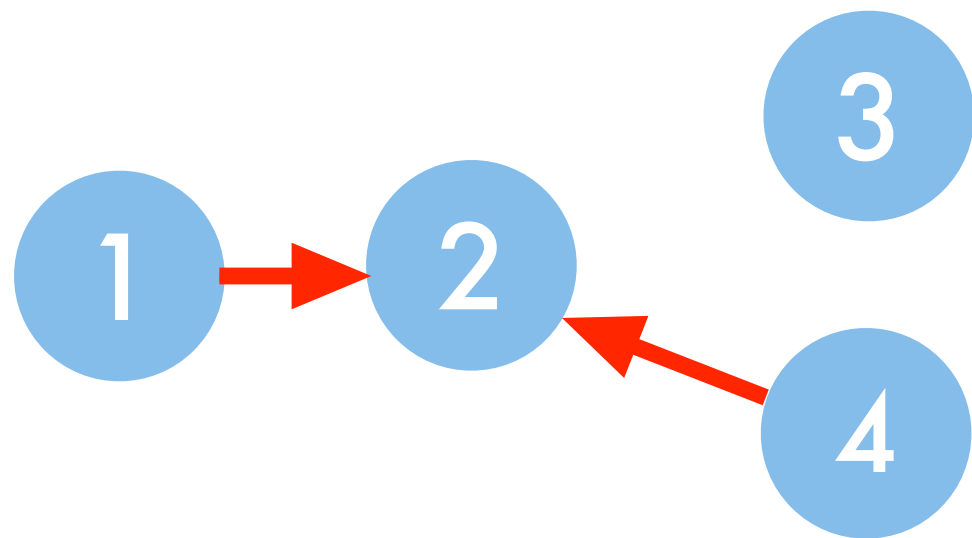
clique  
potential

$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique  
potential

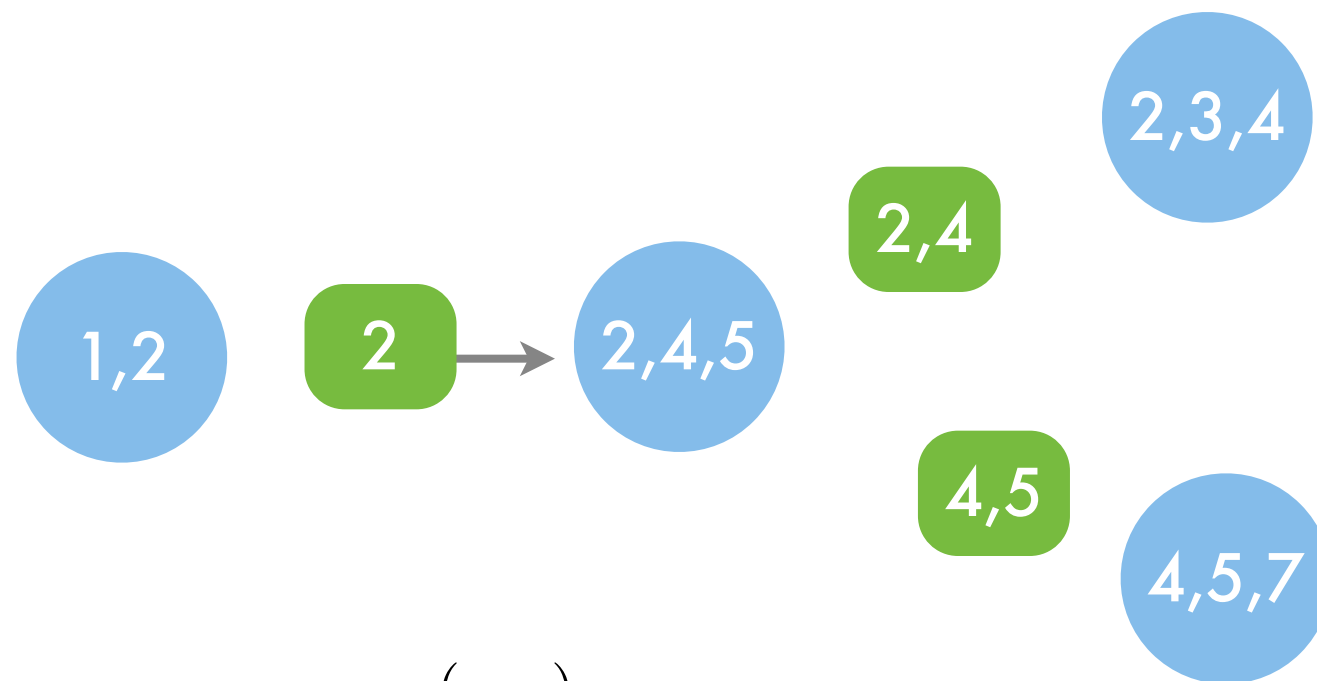
separator  
set

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique potential

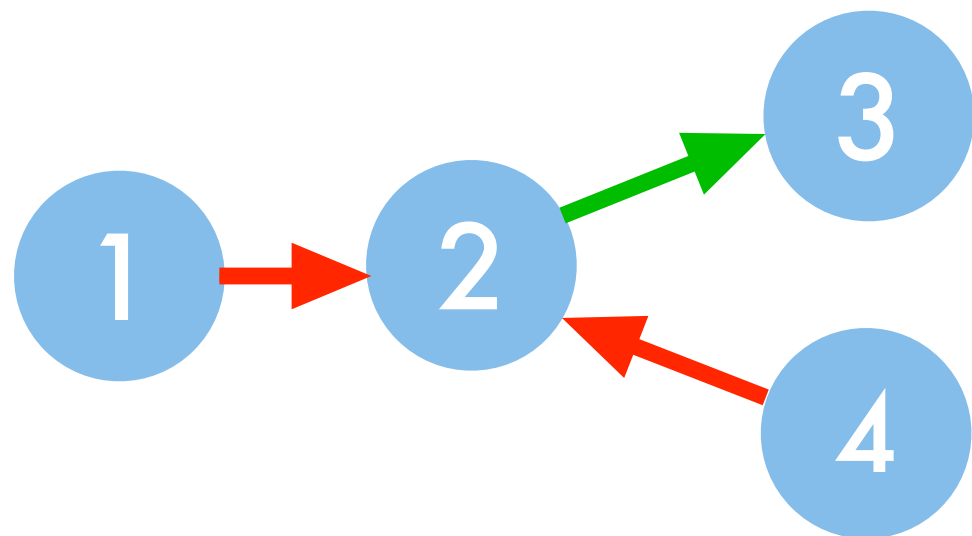


$$m_{245 \rightarrow 234}(x_{24}) = \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45})$$

clique potential

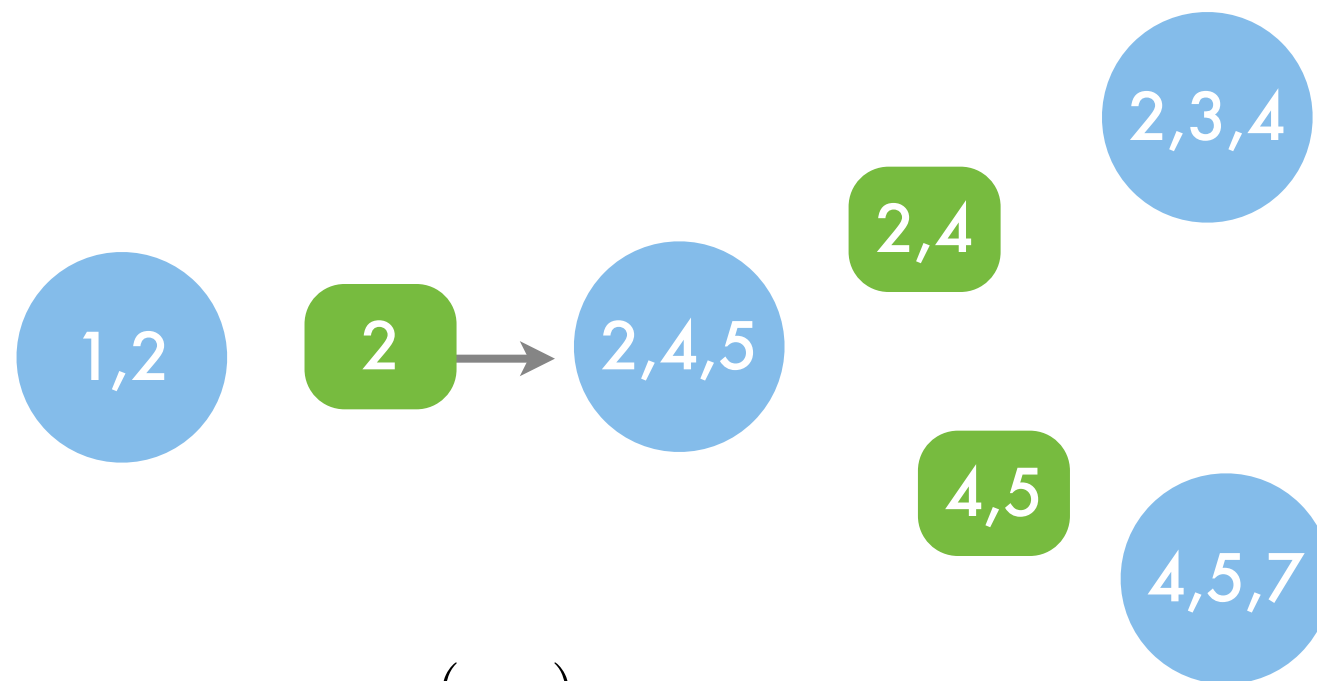
separator set

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

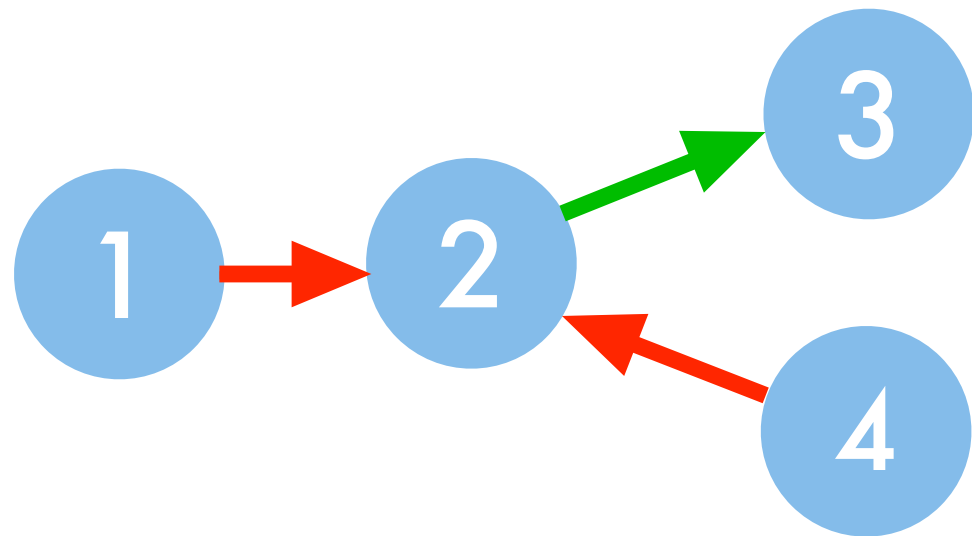


$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique  
potential

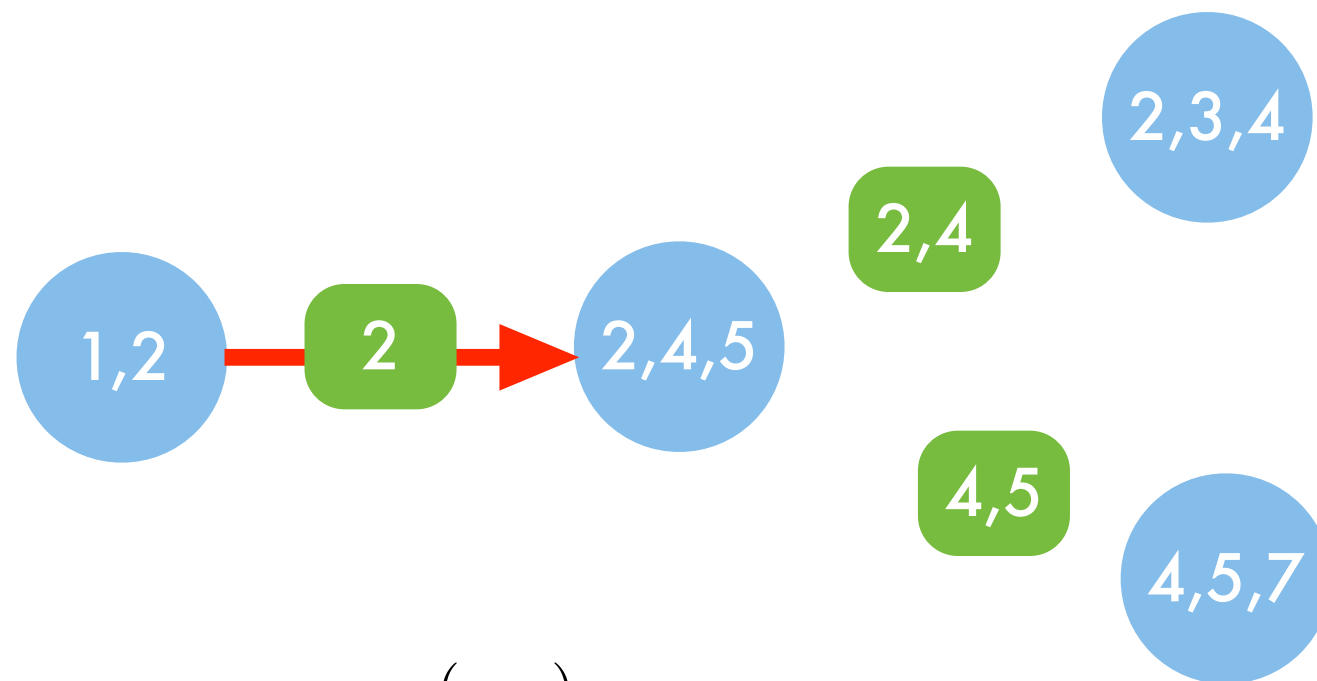
separator  
set

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

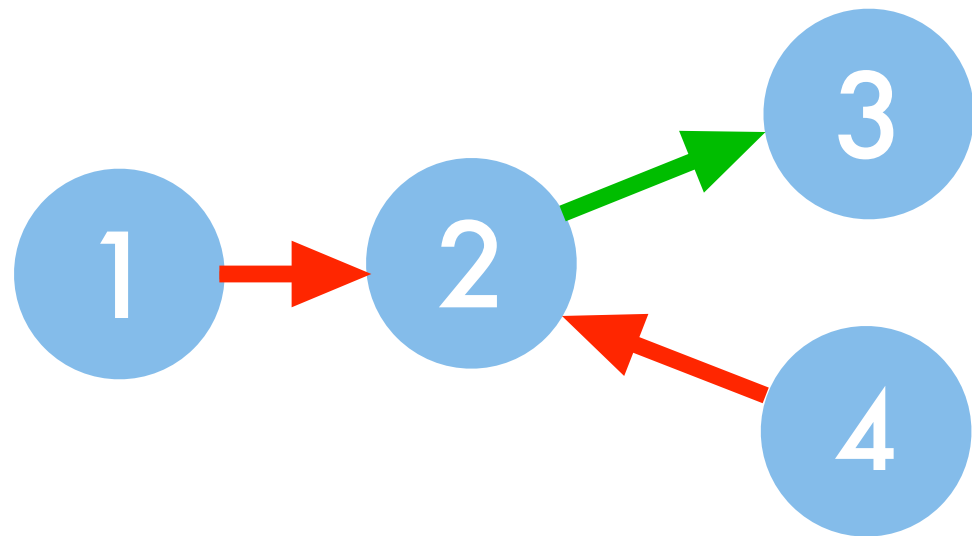


$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique  
potential

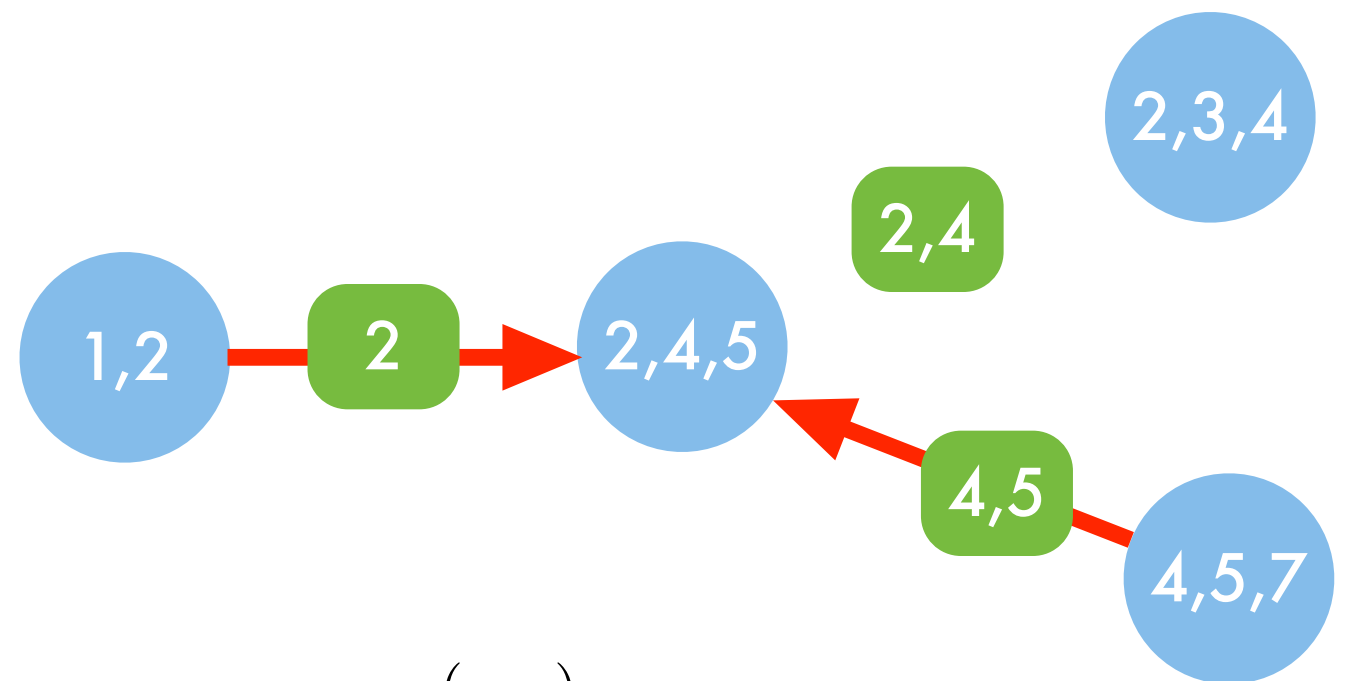
separator  
set

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential



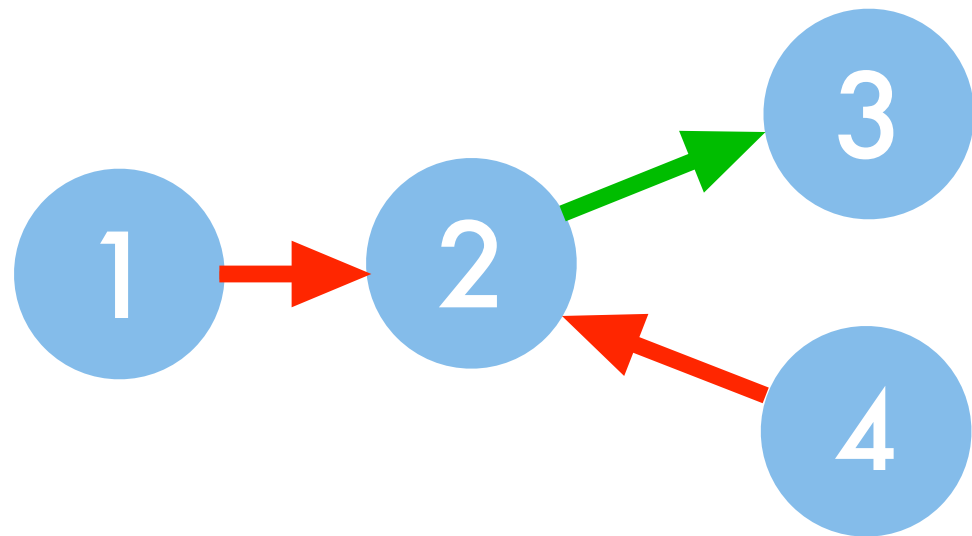
$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique  
potential

separator  
set

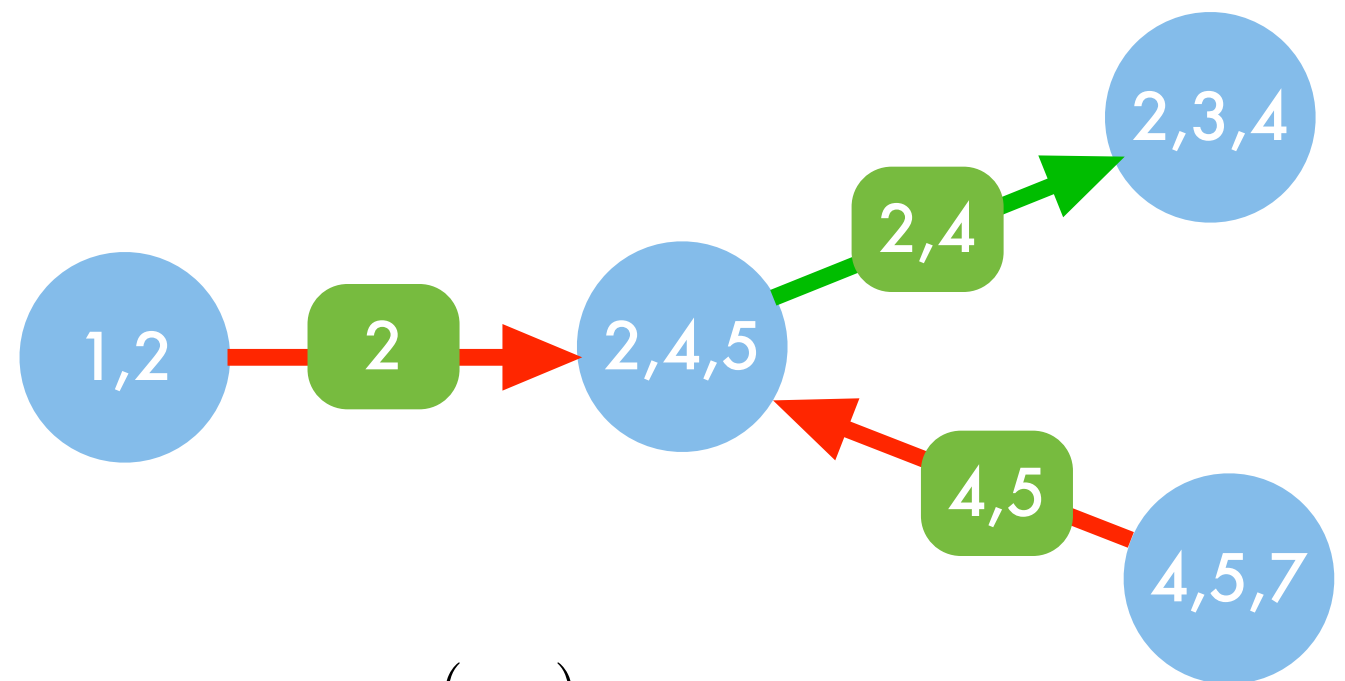


# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

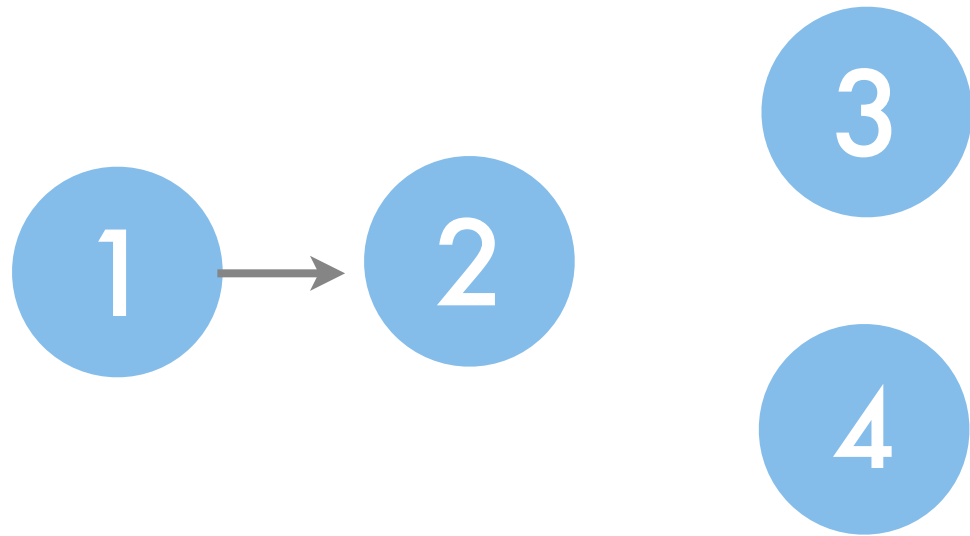


$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique  
potential

separator  
set

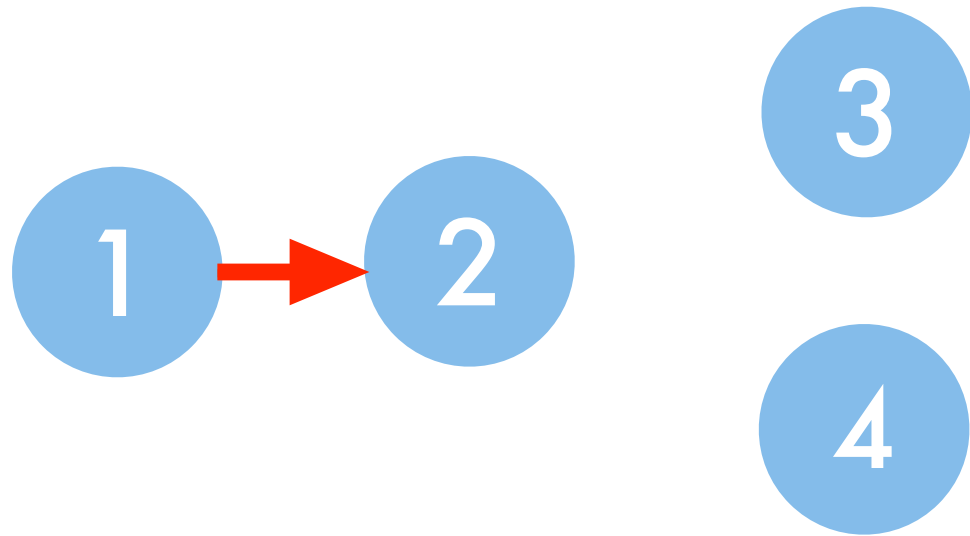
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

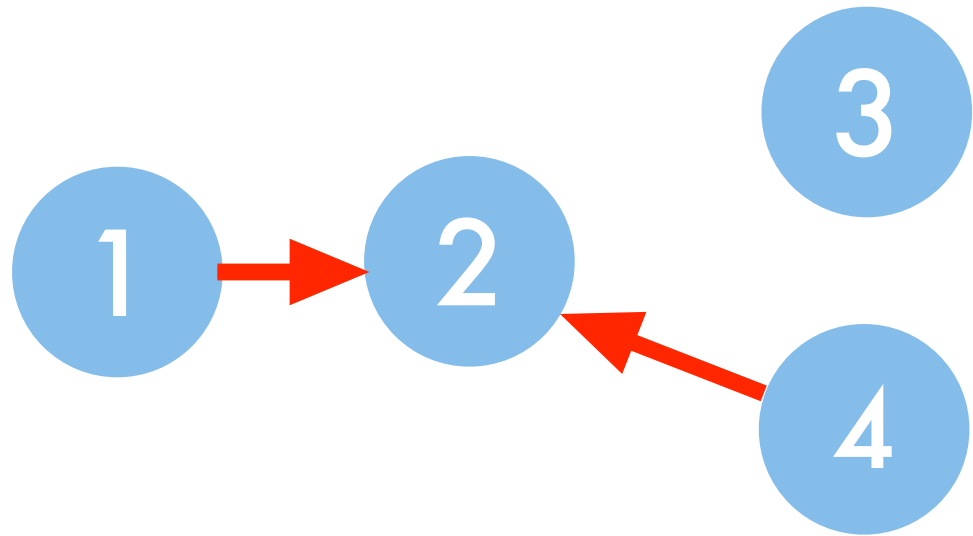
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

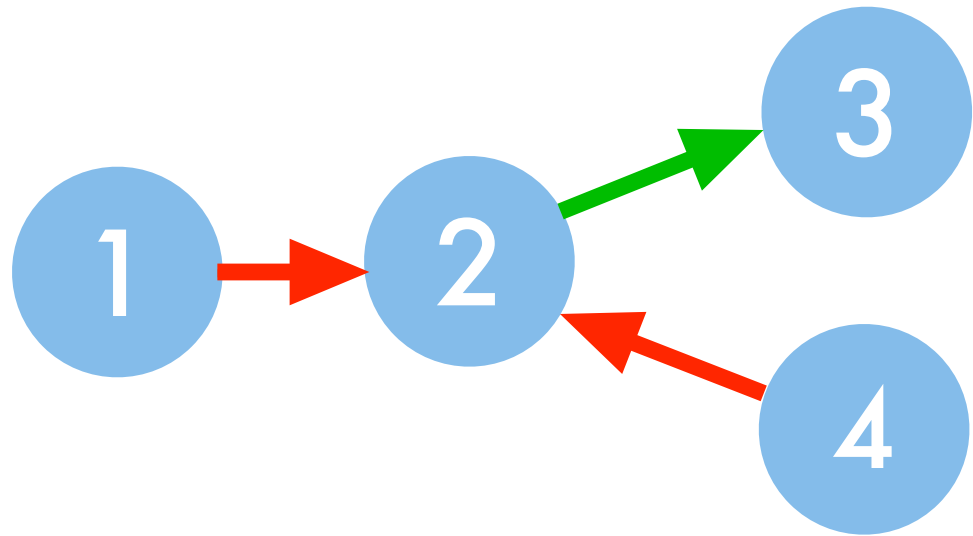
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

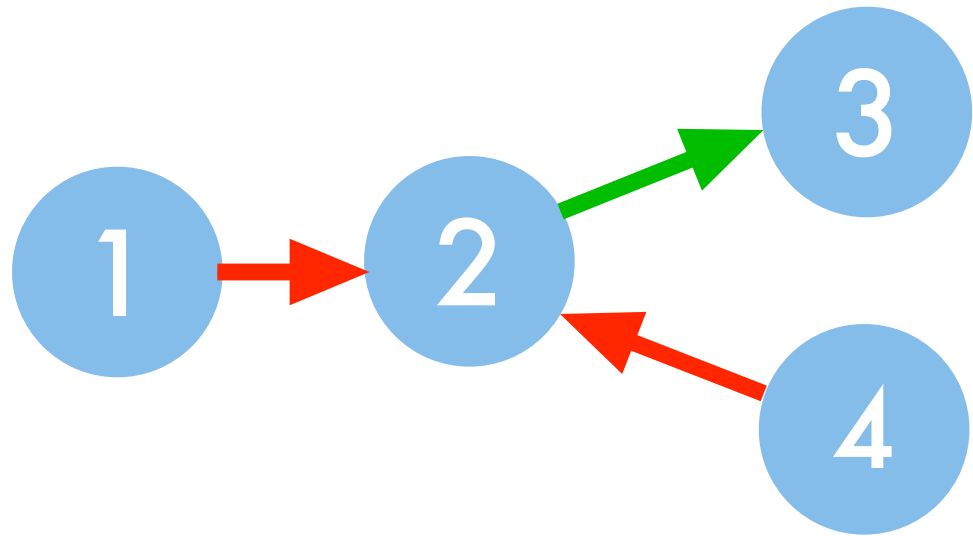
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

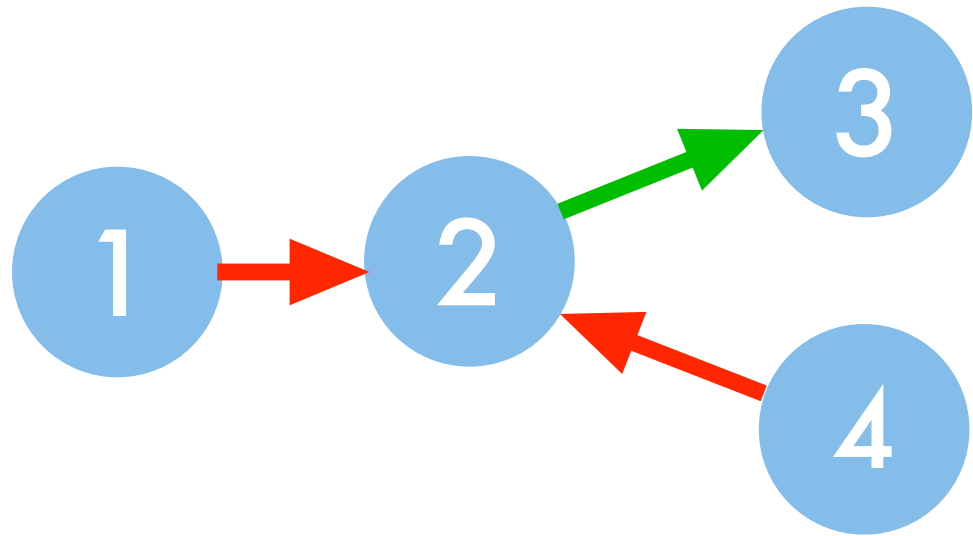
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

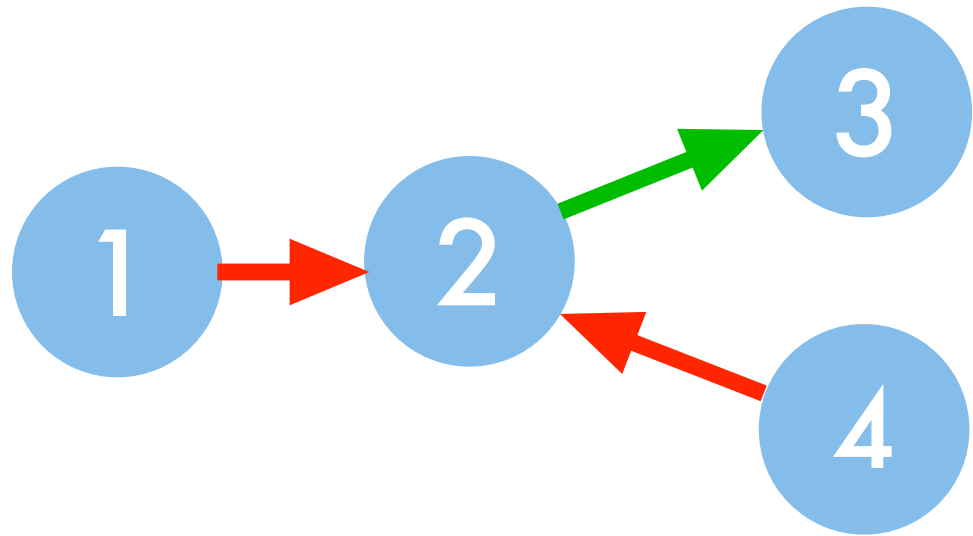
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

# Junction Trees

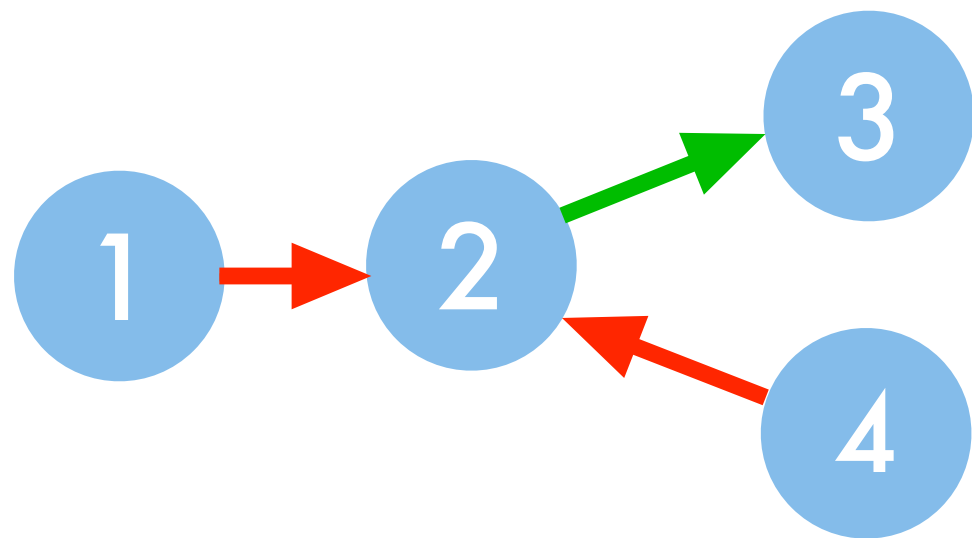


$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

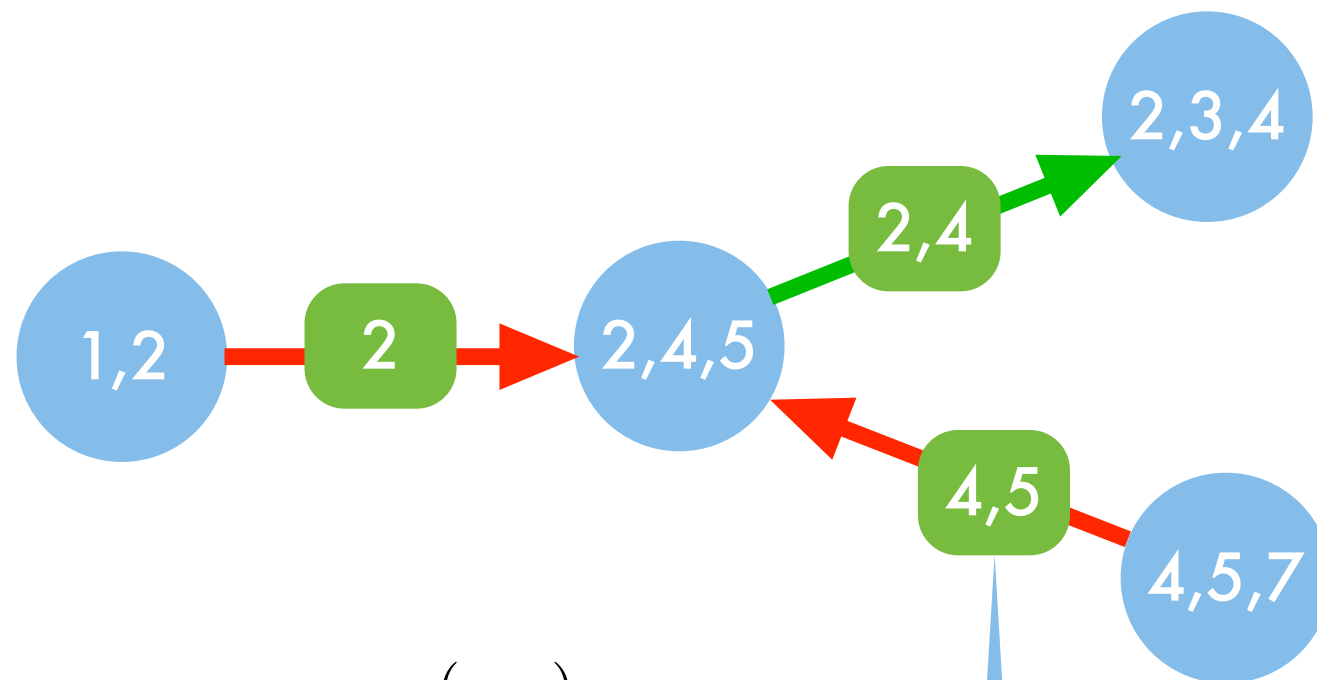


# Junction Trees



$$m_{i \rightarrow j}(x_j) = \sum_{x_i} f(x_i, x_j) \prod_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique potential



$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \sum_{x_5} f(x_{245}) m_{12 \rightarrow 245}(x_2) m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique potential

separator set

# Generalized Distributive Law

- **Key Idea**

Dynamic programming uses only sums and multiplications, hence replace them with equivalent operations from other semirings

- **Semiring**

- 'addition' and 'summation' equivalent

- Associative law:  $(a+b)+c = a+(b+c)$

- Distributive law:  $a(b+c) = ab + ac$

# Generalized Distributive Law

- Integrating out probabilities (sum, product)

$$a \cdot (b + c) = a \cdot b + a \cdot c$$

- Finding the maximum (max, +)

$$a + \max(b, c) = \max(a + b, a + c)$$

- Set algebra (union, intersection)

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

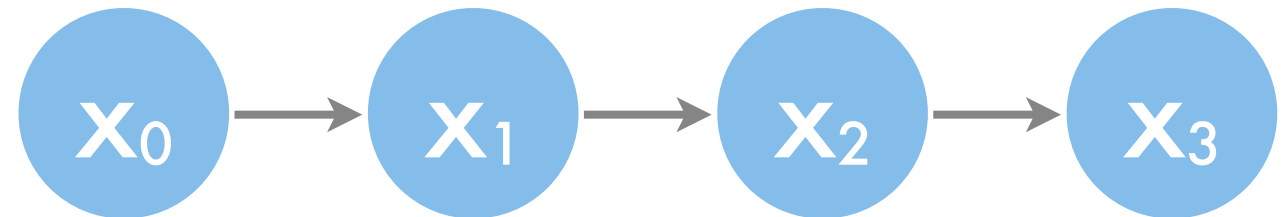
- Boolean semiring (AND, OR)

- Probability semiring (log +, +)

- Tropical semiring (min, +)

# Chains ... again

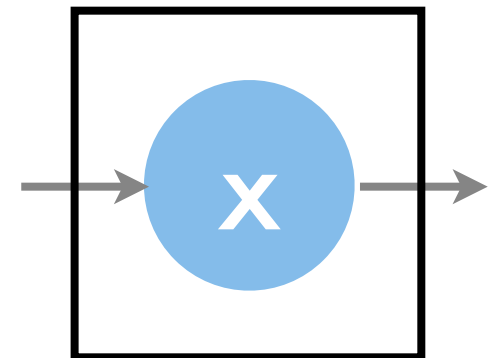
$$\bar{s} = \max_x s(x_0) + \sum_{i=1}^{n-1} s(x_{i+1}|x_i)$$



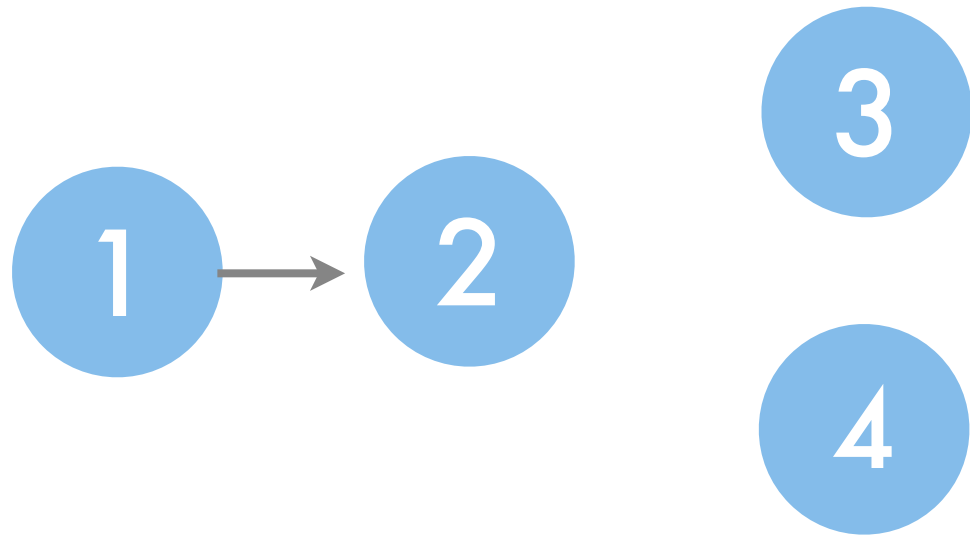
$$\bar{s} = \max_{x_0 \dots n} \underbrace{s(x_0)}_{:=l_0(x_0)} + \sum_{j=1}^n s(x_j|x_{j-1})$$

$$= \max_{x_1 \dots n} \underbrace{\max_{x_0} [l_0(x_0) s(x_1|x_0)]}_{:=l_1(x_1)} + \sum_{j=2}^n s(x_j|x_{j-1})$$

$$= \max_{x_2 \dots n} \underbrace{\max_{x_1} [l_1(x_1) s(x_2|x_1)]}_{:=l_2(x_2)} + \sum_{j=3}^n s(x_j|x_{j-1})$$



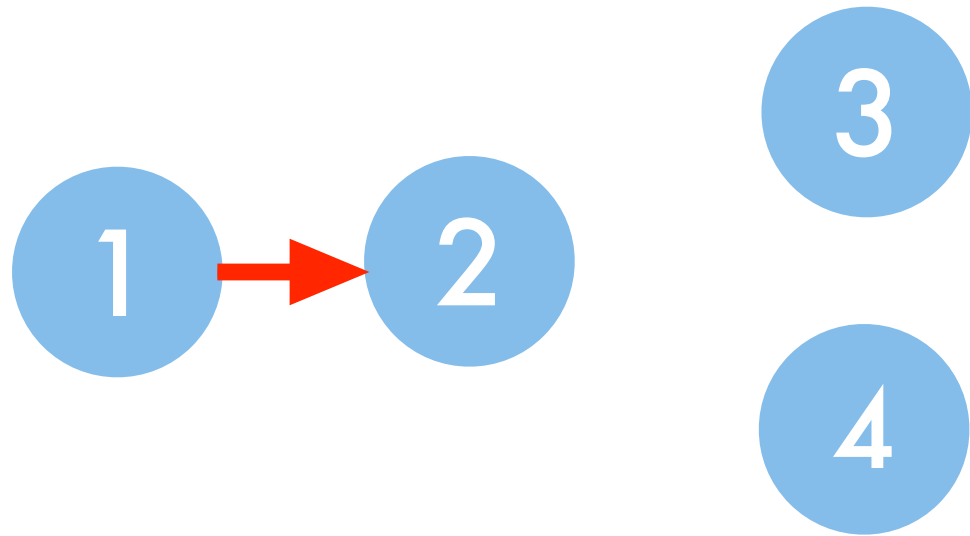
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

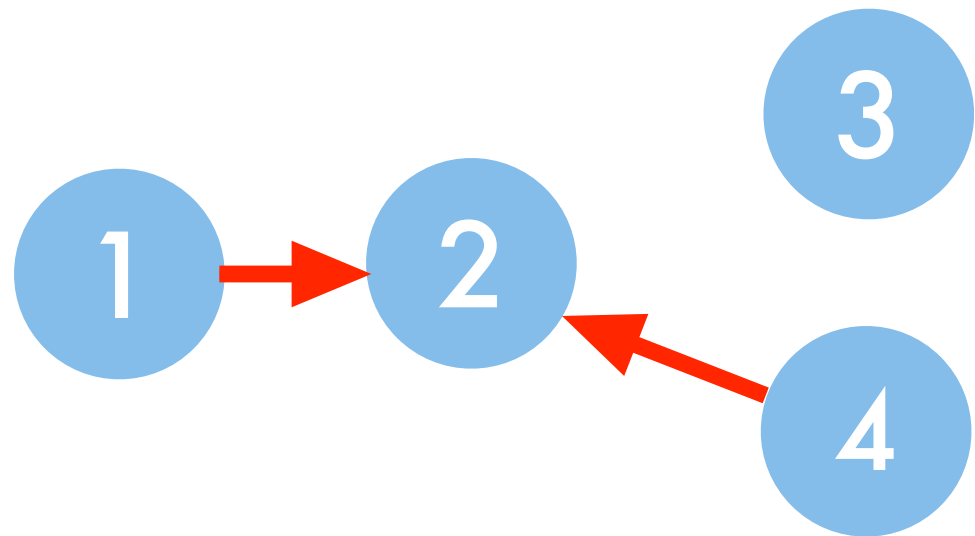
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

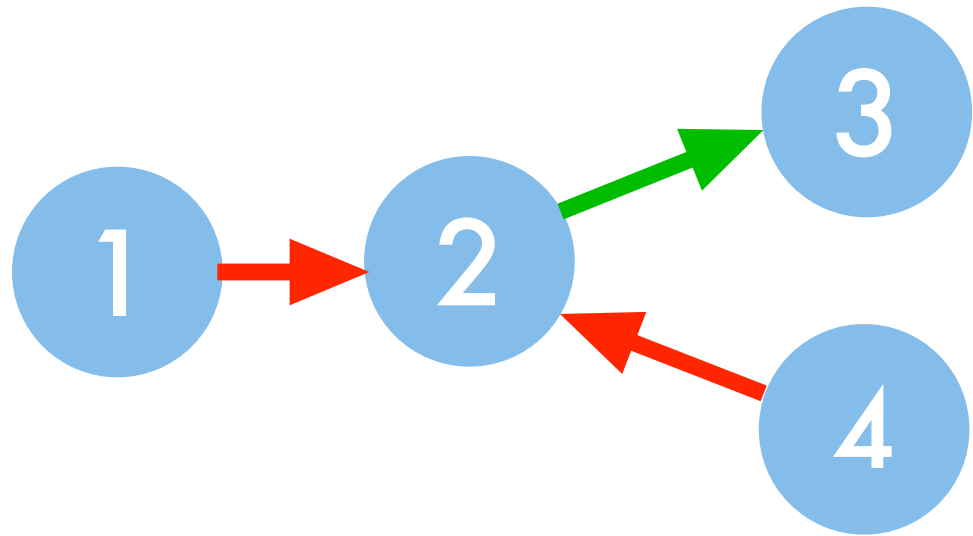
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

# Junction Trees

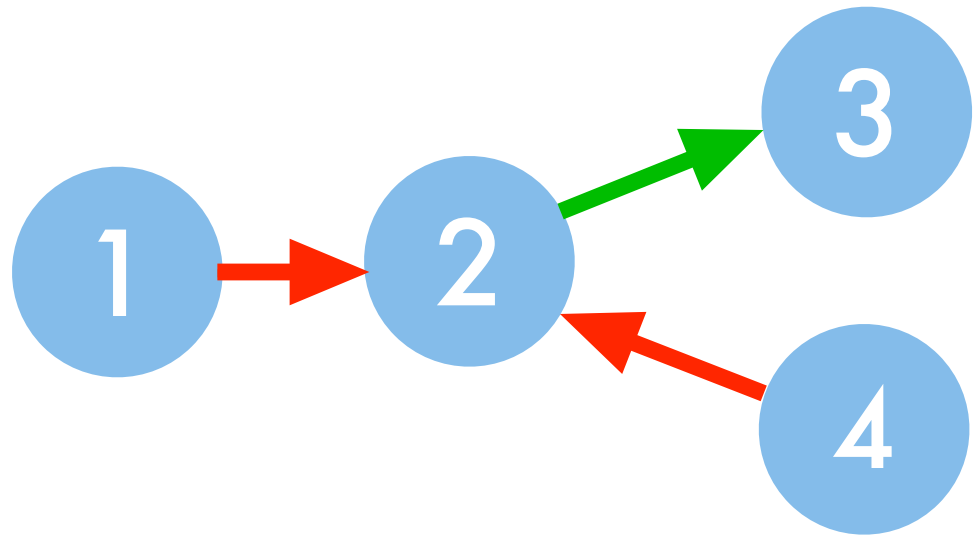


$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential



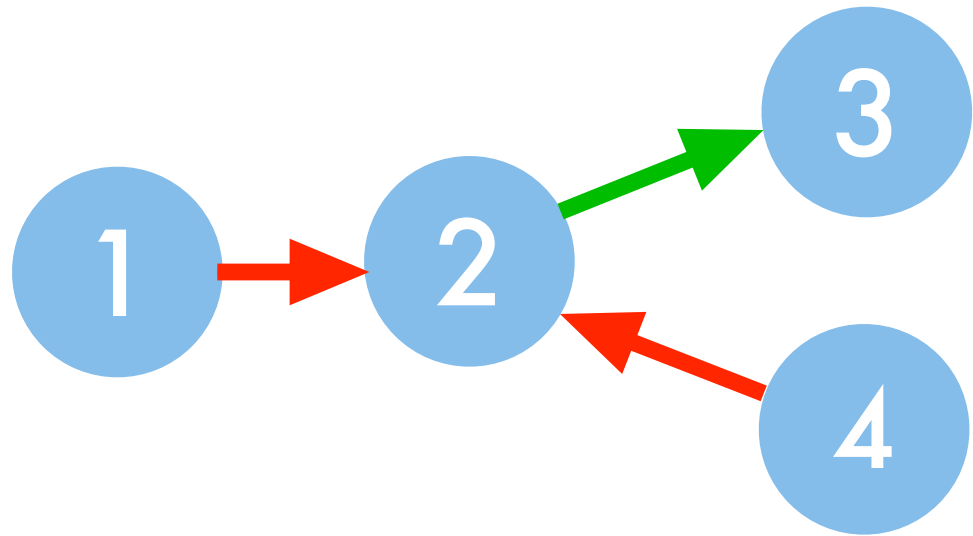
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

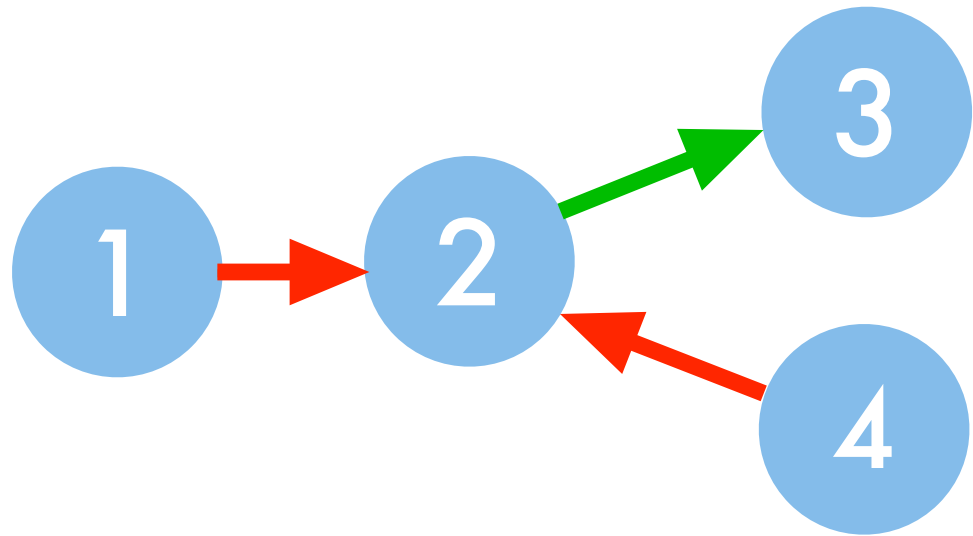
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential

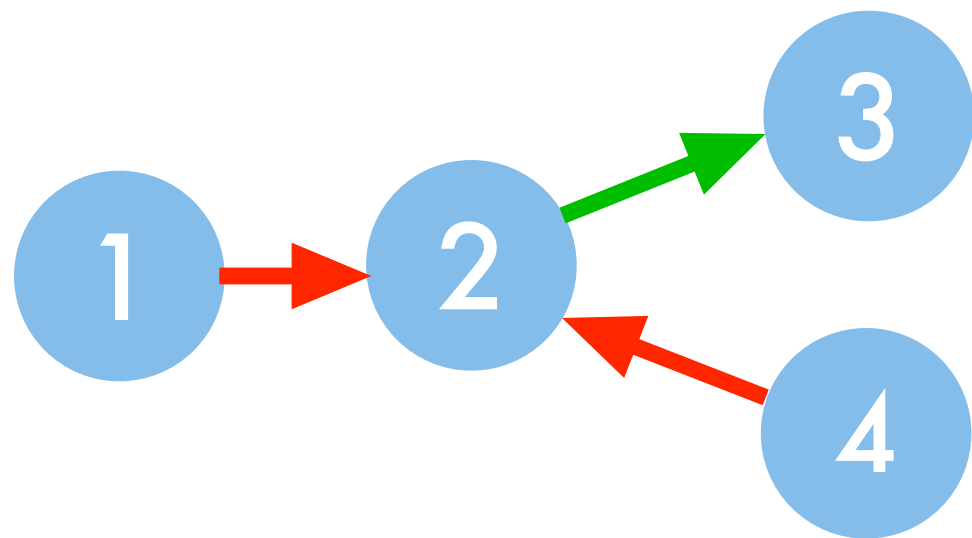
# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

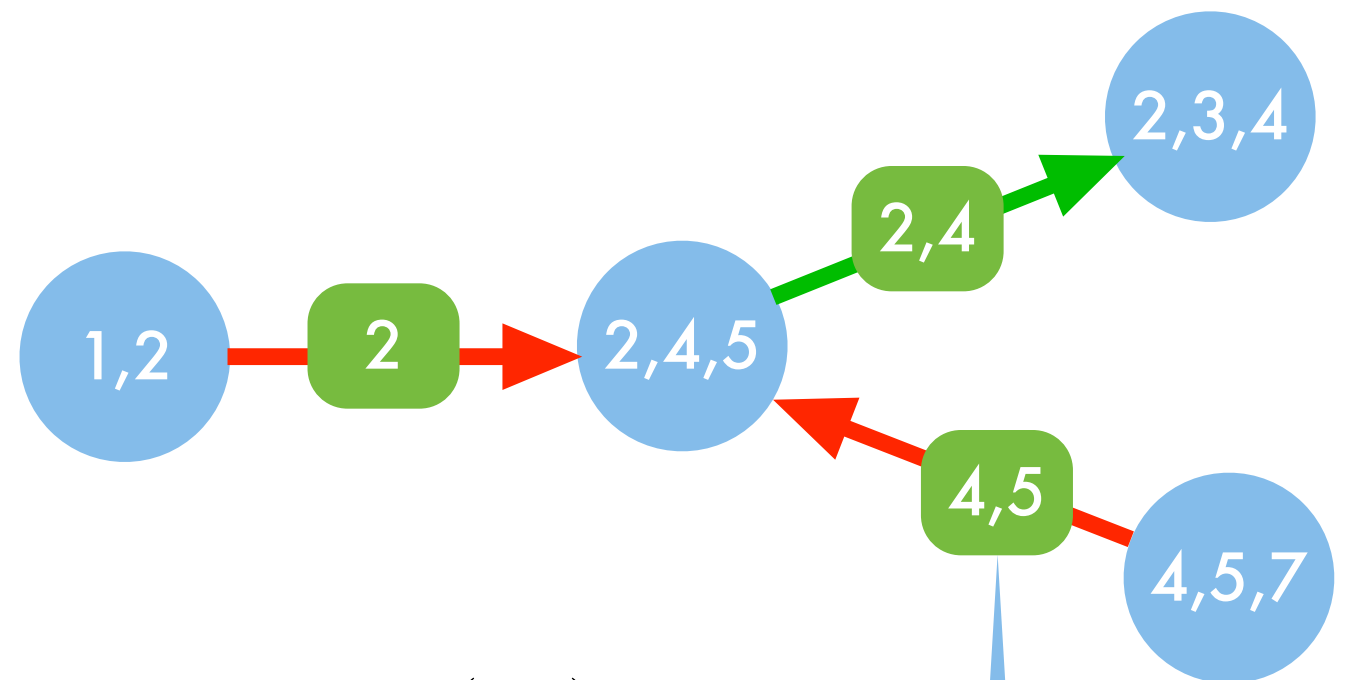
clique  
potential

# Junction Trees



$$m_{i \rightarrow j}(x_j) = \max_{x_i} f(x_i, x_j) + \sum_{l \neq j} m_{l \rightarrow i}(x_j)$$

clique  
potential



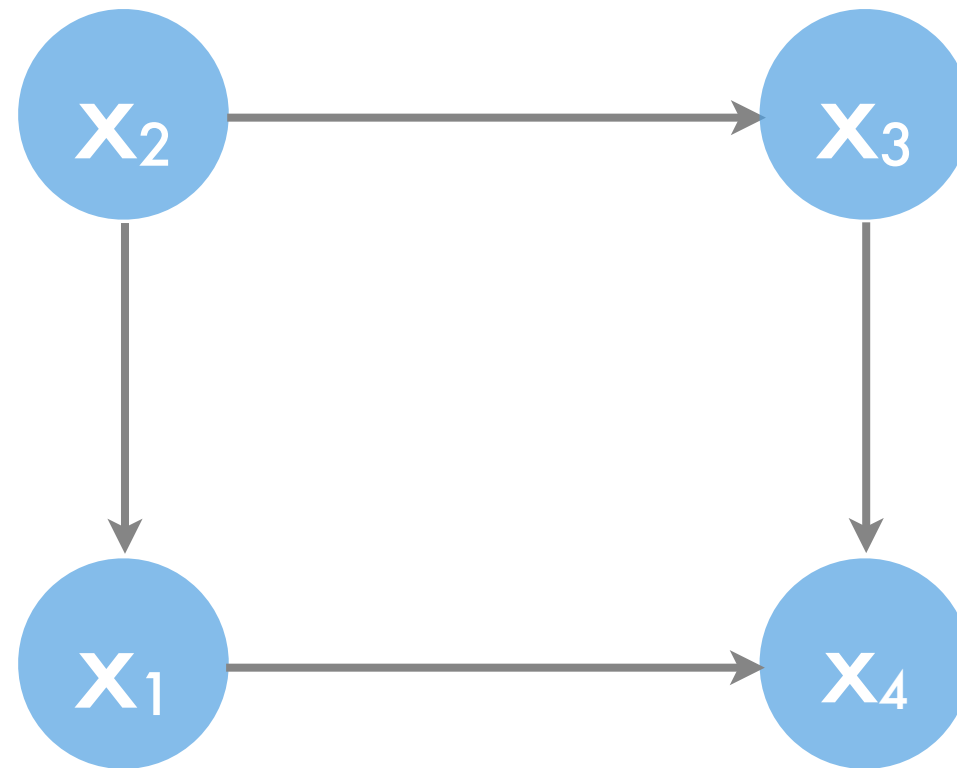
$$\begin{aligned} & m_{245 \rightarrow 234}(x_{24}) \\ &= \max_{x_5} f(x_{245}) + m_{12 \rightarrow 245}(x_2) + m_{457 \rightarrow 245}(x_{45}) \end{aligned}$$

clique  
potential

separator  
set

# No loops allowed

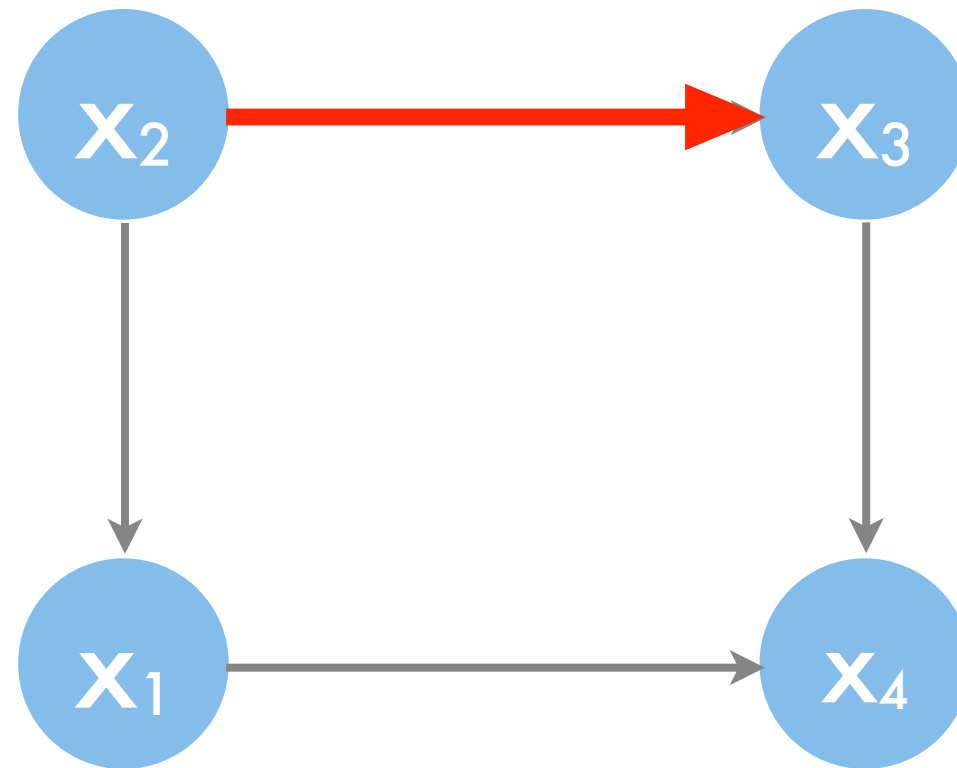
$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$



If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)

# No loops allowed

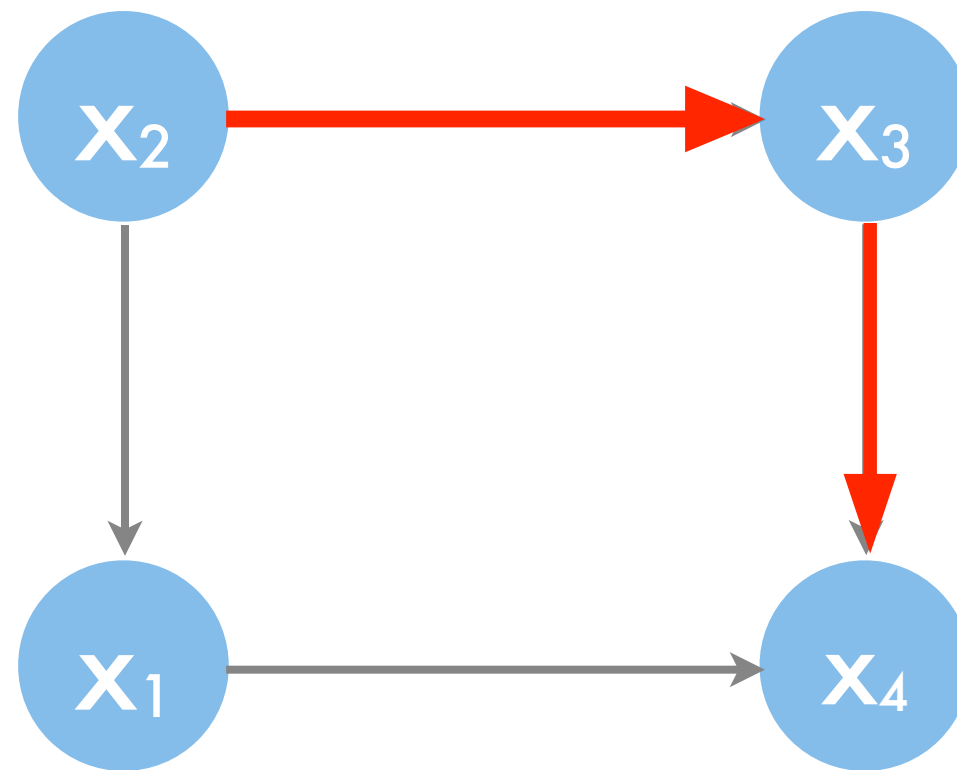
$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$



If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)

# No loops allowed

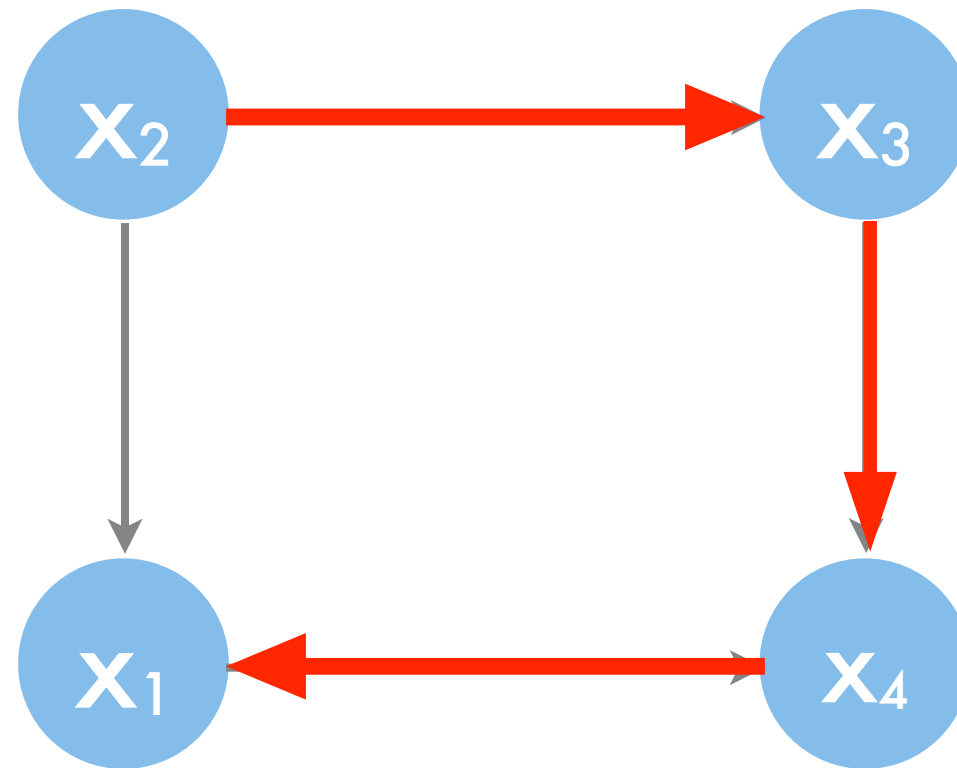
$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$



If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)

# No loops allowed

$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$

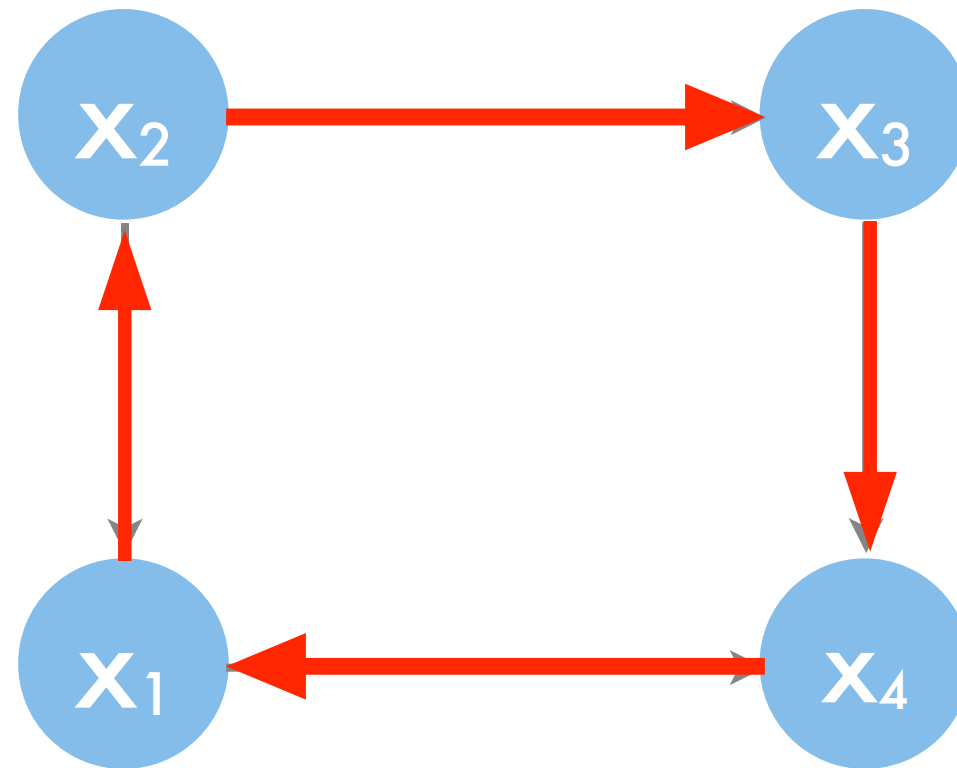


If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)



# No loops allowed

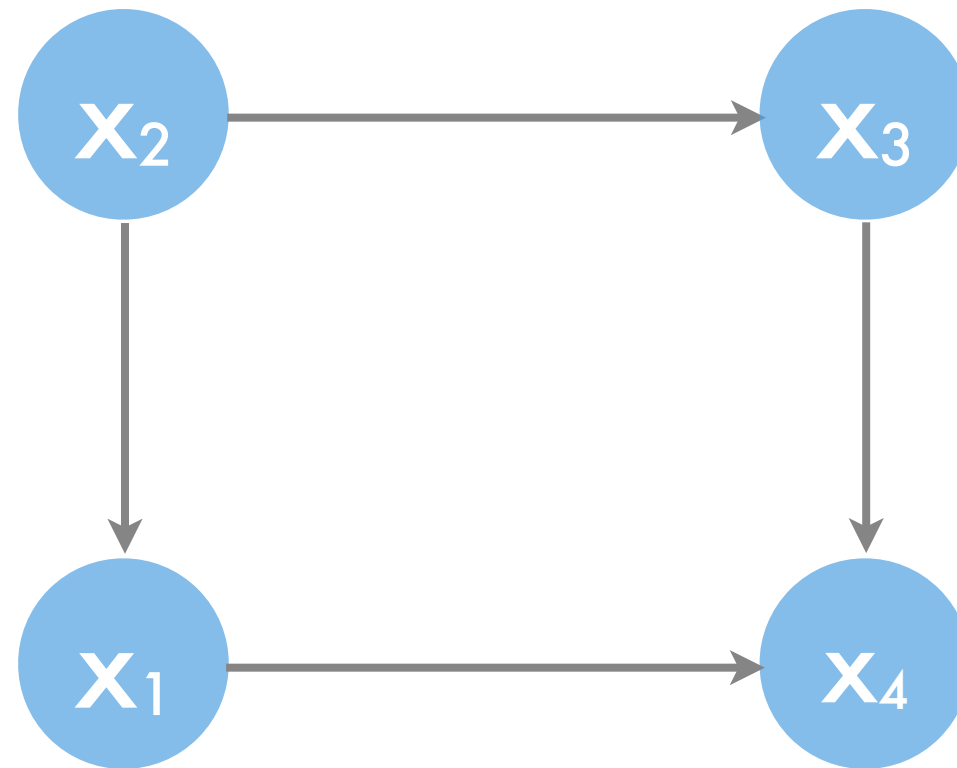
$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$



If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)

# No loops allowed

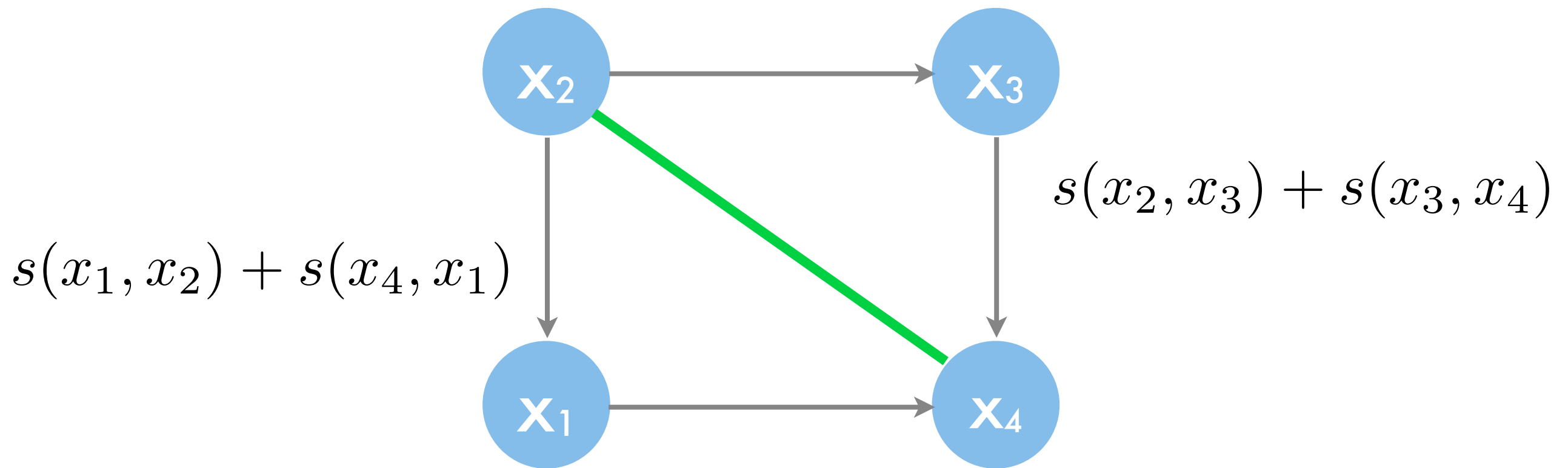
$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$



If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)

# No loops allowed

$$s(x_1, x_2) + s(x_2, x_3) + s(x_3, x_4) + s(x_4, x_1)$$



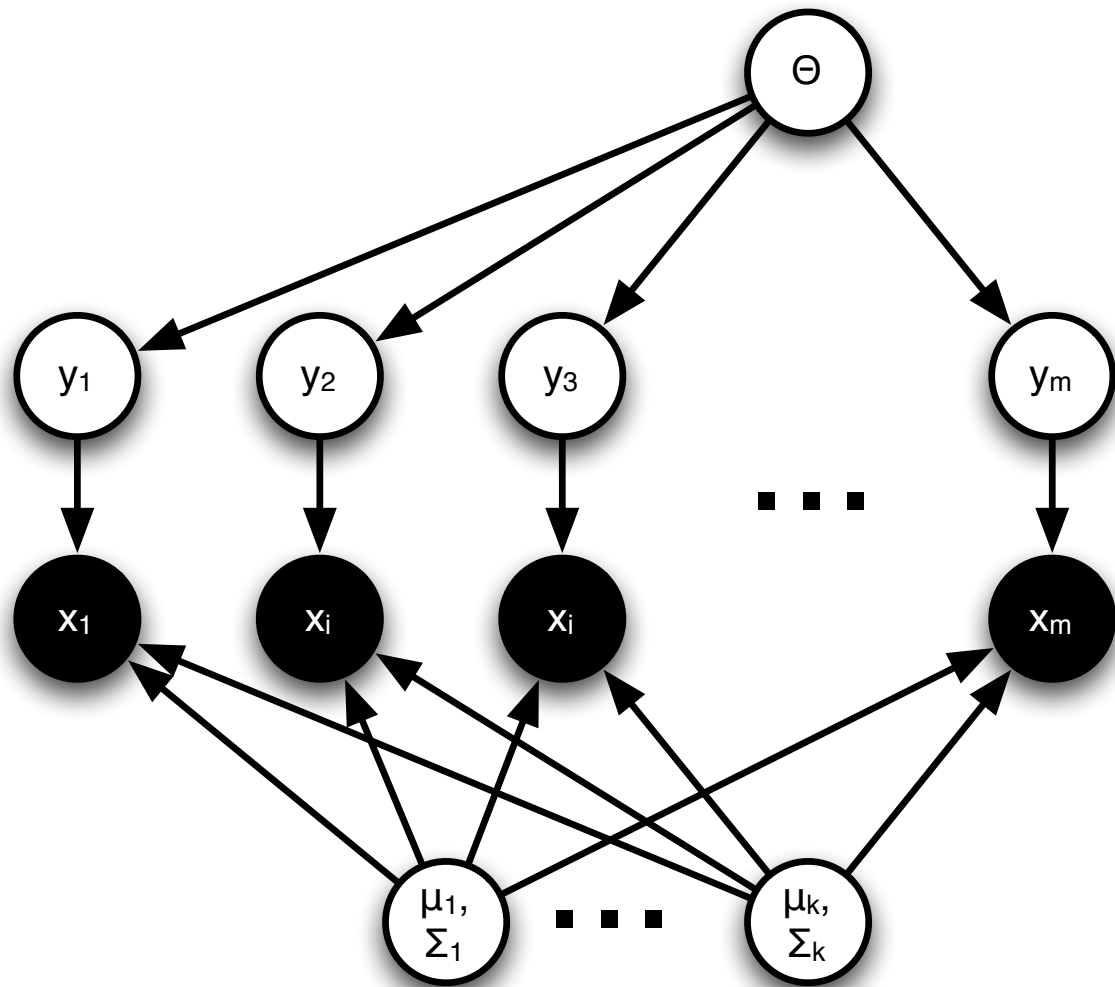
If we use it anyway – Loopy Belief Propagation  
(Turbo Codes, Markov Random Fields, etc.)



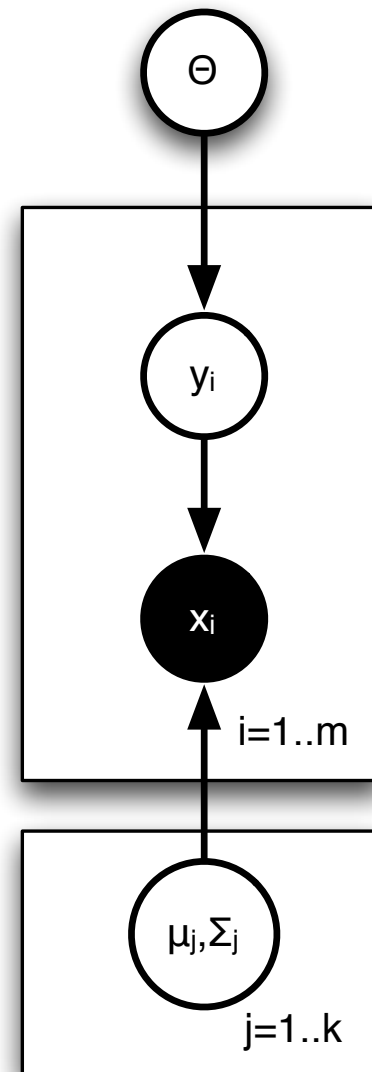
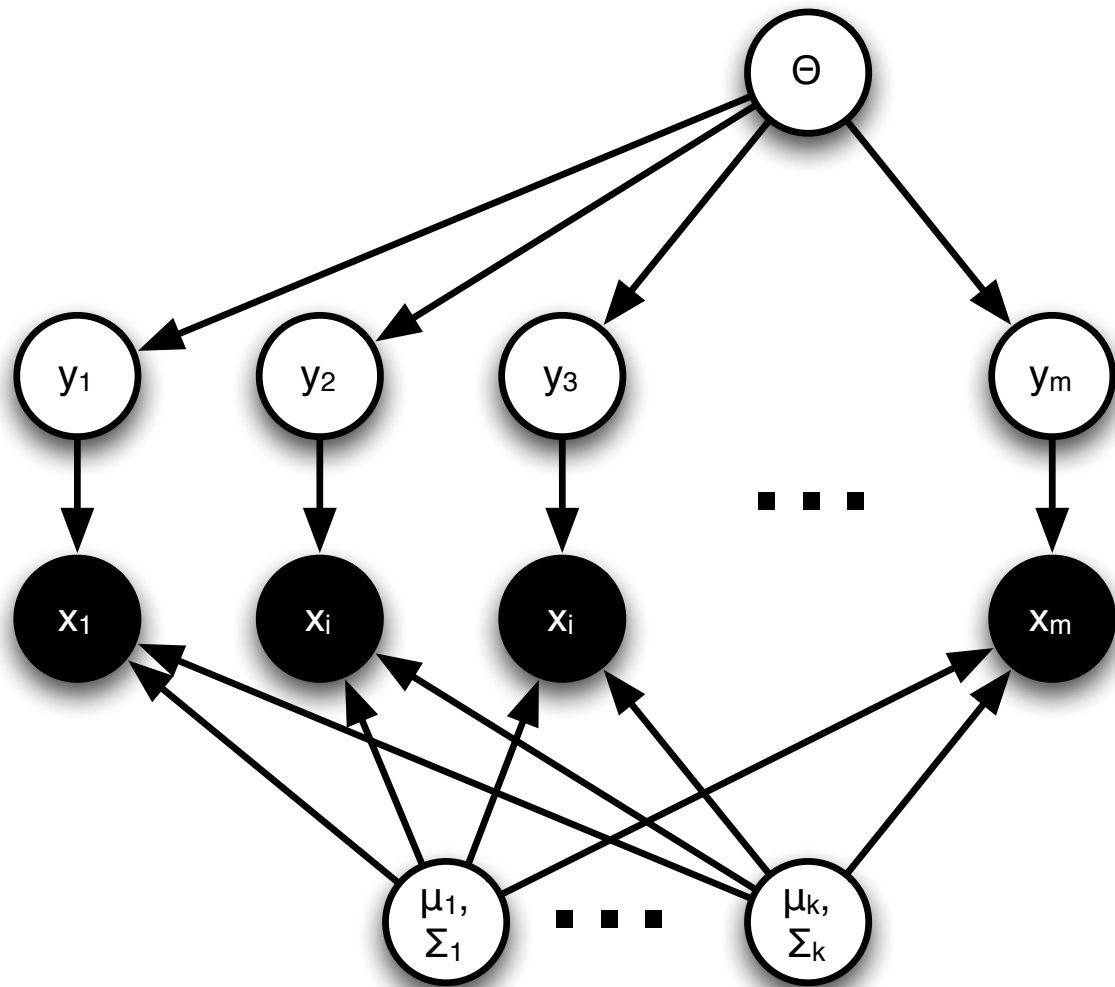
# Clustering



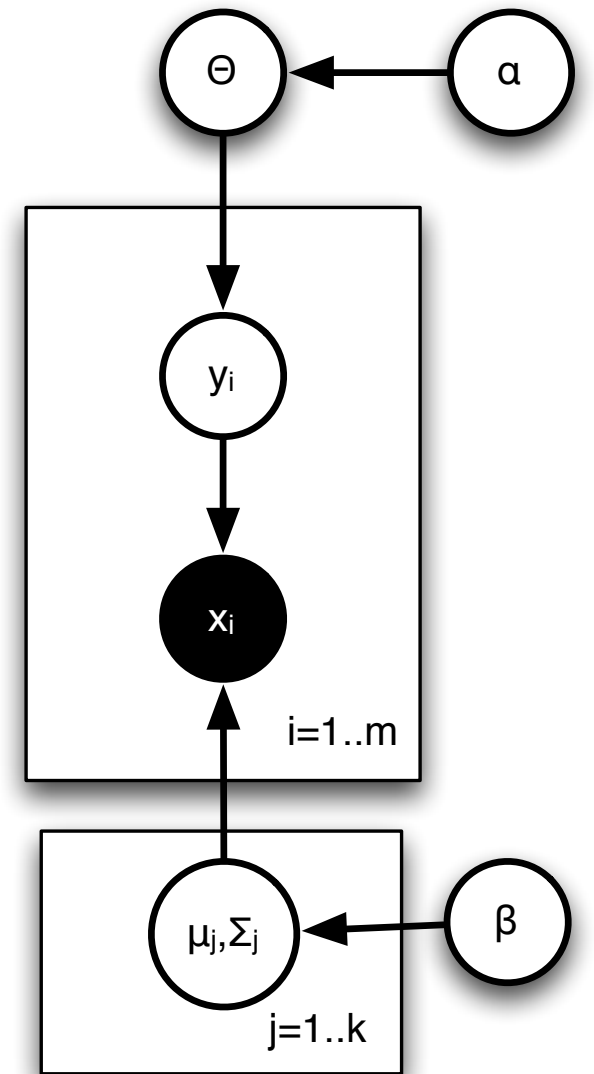
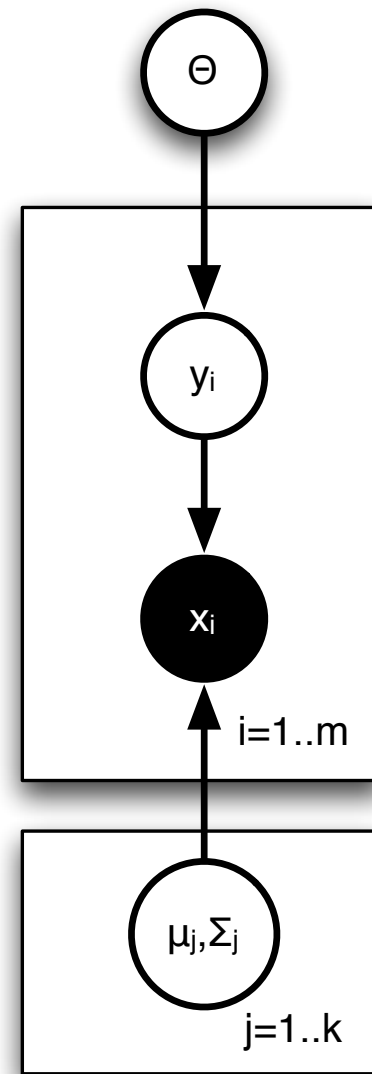
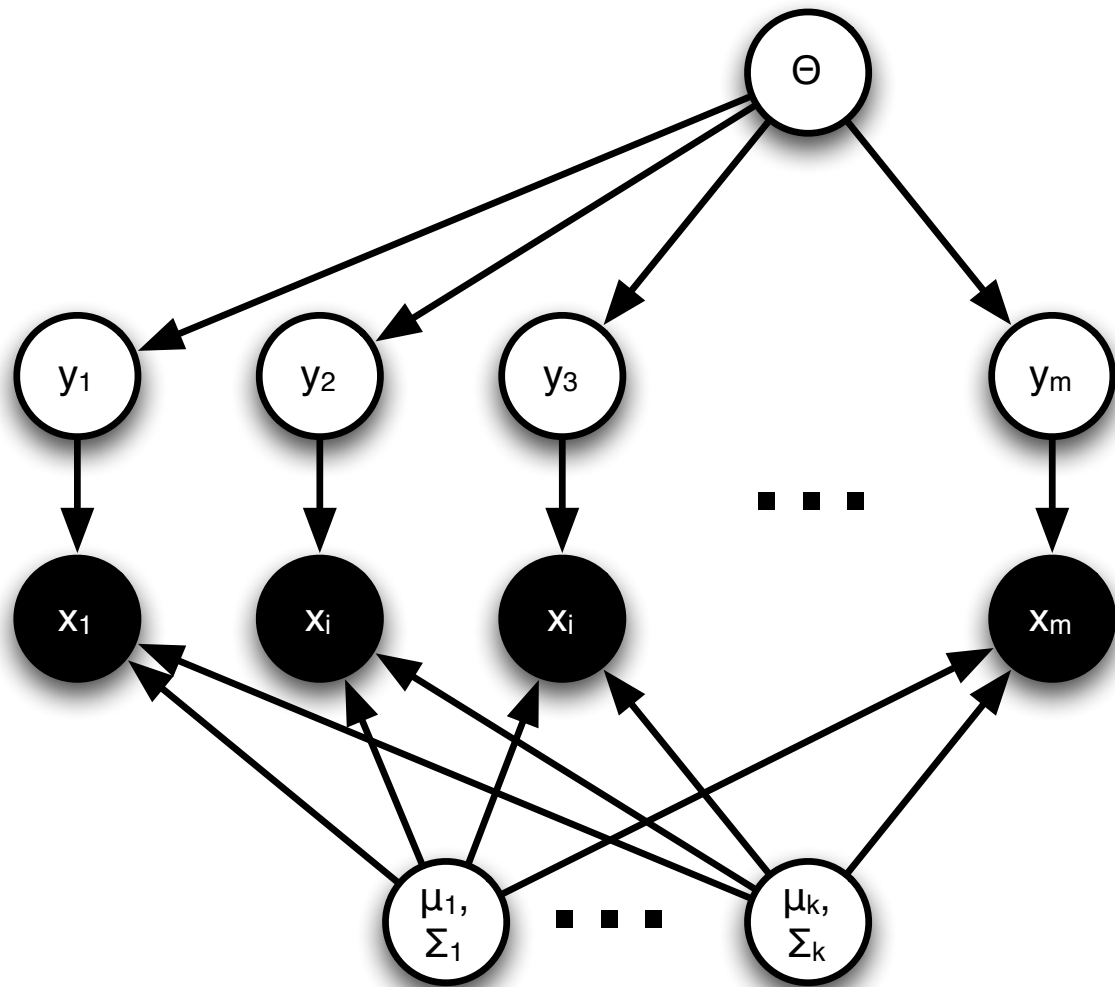
# Basic Idea



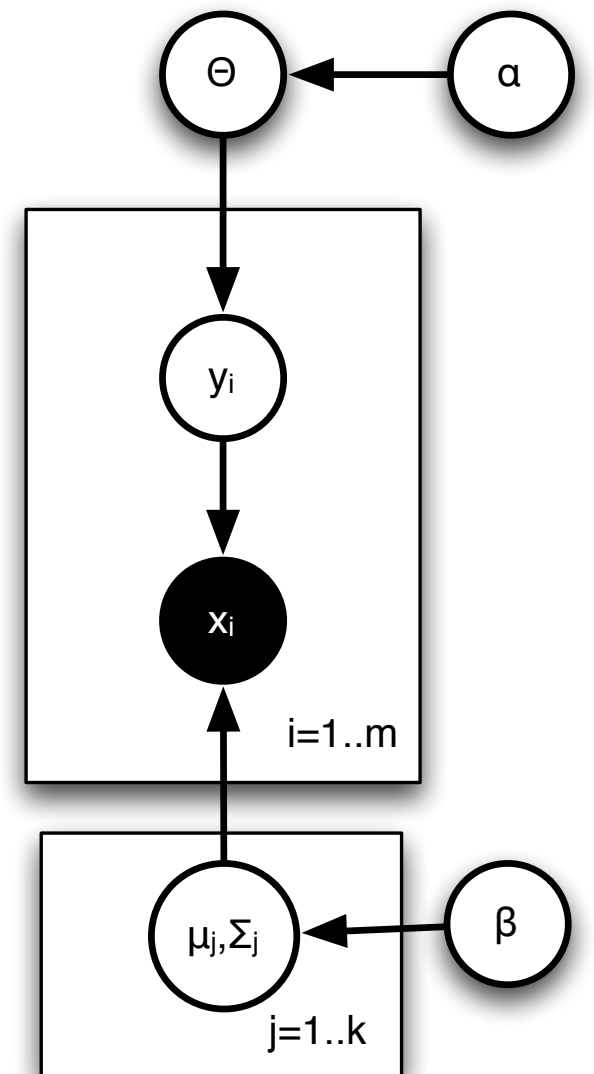
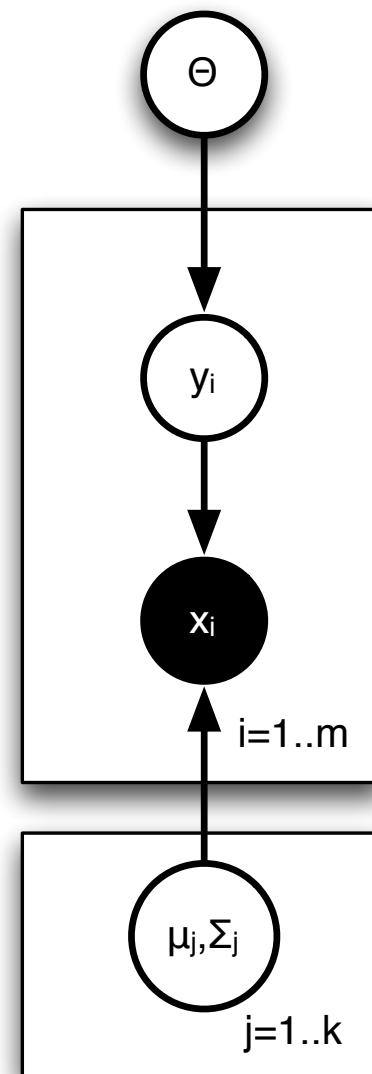
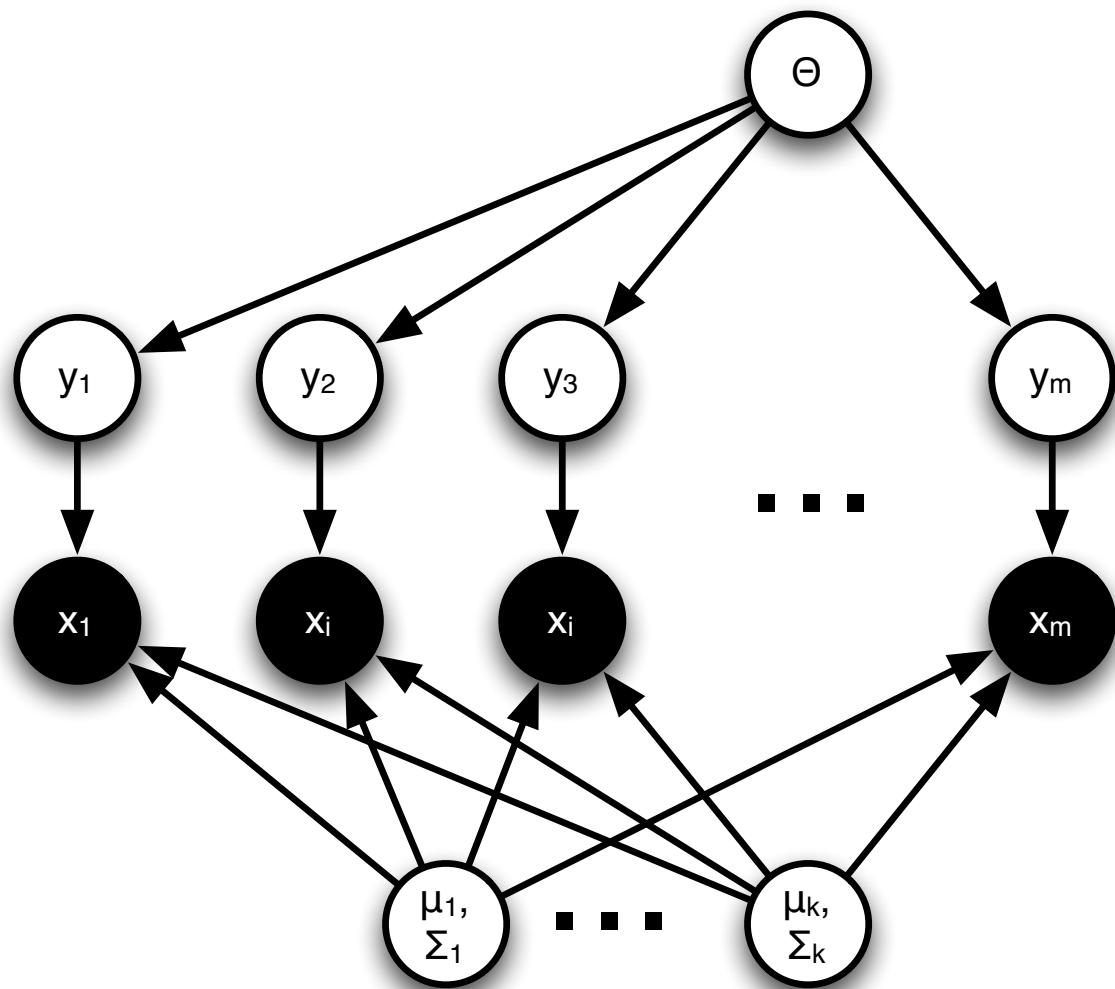
# Basic Idea



# Basic Idea



# Basic Idea



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



# What can we cluster?

# What can we cluster?

mails      text      urls      products

news      queries      users

spammers      ads      locations

abuse      events

# Mixture of Gaussians

- Draw cluster label  $y$  from discrete distribution
- Draw data  $x$  from Gaussian for given cluster  $y$
- Prior for discrete distribution - Dirichlet
- Prior for Gaussians - Gauss-Wishart distribution
- Problem: we don't know  $y$ 
  - If we knew the parameters we could get  $y$
  - If we knew  $y$  we could get the parameters

# k-means

- Fixed uniform variance for all Gaussians
- Fixed uniform distribution over clusters
- Initialize centers with random subset of points
- Find most likely cluster  $y$  for  $x$

$$y_i = \operatorname{argmax}_y p(x_i | y, \sigma, \mu)$$

- Find most likely center for given cluster

$$\mu_y = \frac{1}{n_y} \sum_i \{y_i = y\} x_i$$

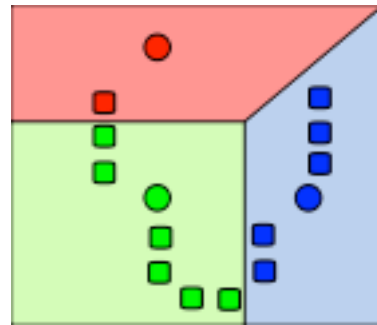
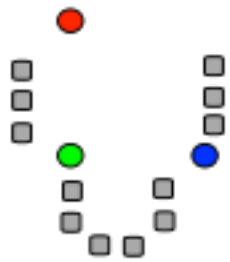
- Repeat until converged

# k-means

- **Pro**
  - simple algorithm
  - can be implemented by *MapReduce* passes
- **Con**
  - no proper probabilistic representation
  - can get stuck easily in local minima

# k-means

partitioning

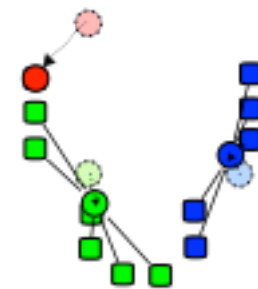
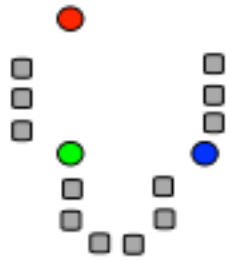


initialization

# k-means

partitioning

partitioning



initialization

update

# Expectation Maximization



# Expectation Maximization

- **Optimization Problem**

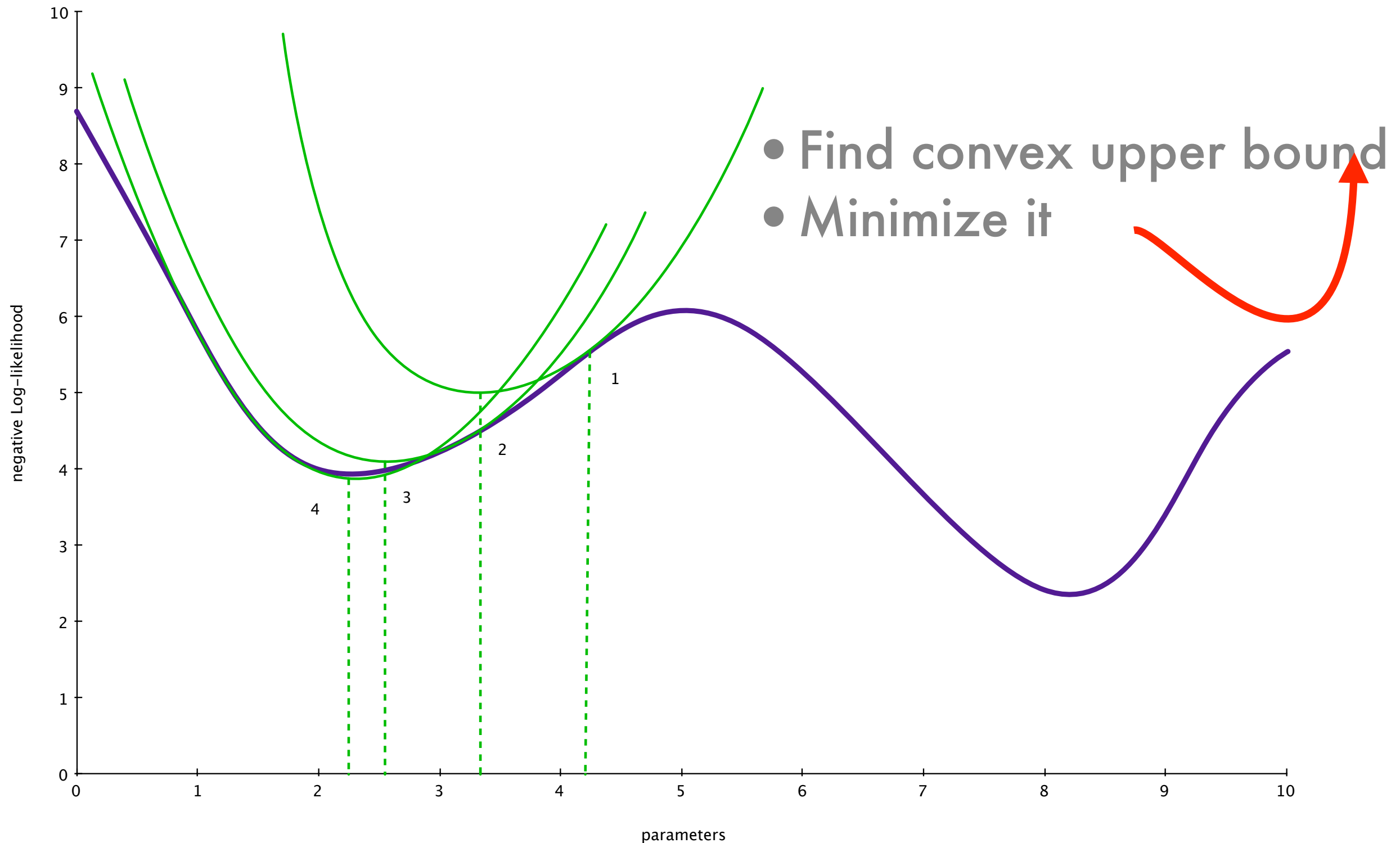
$$\text{maximize}_{\theta, \mu, \sigma} p(X|\theta, \sigma, \mu) = \text{maximize}_{\theta, \mu, \sigma} \sum_Y \prod_{i=1}^n p(x_i|y_i, \sigma, \mu)p(y_i|\theta)$$

**This problem is nonconvex and difficult to solve**

- **Key idea**

**If we knew  $p(y|x)$  we could estimate the remaining parameters easily and vice versa**

# Nonconvex Minimization



# Expectation Maximization

# Expectation Maximization

- **Variational Bound**

$$\begin{aligned}\log p(x; \theta) &\geq \log p(x; \theta) - D(q(y) \| p(y|x; \theta)) \\ &= \int dq(y) [\log p(x; \theta) + \log p(y|x; \theta) - \log q(y)] \\ &= \int dq(y) \log p(x, y; \theta) - \int dq(y) \log q(y)\end{aligned}$$

**This inequality is tight for  $p(y|x) = q(y)$**

# Expectation Maximization

- **Variational Bound**

$$\begin{aligned}\log p(x; \theta) &\geq \log p(x; \theta) - D(q(y) \| p(y|x; \theta)) \\ &= \int dq(y) [\log p(x; \theta) + \log p(y|x; \theta) - \log q(y)] \\ &= \int dq(y) \log p(x, y; \theta) - \int dq(y) \log q(y)\end{aligned}$$

**This inequality is tight for  $p(y|x) = q(y)$**

- **Expectation step**

$$q(y) = p(y|x; \theta)$$

# Expectation Maximization

- **Variational Bound**

$$\begin{aligned}\log p(x; \theta) &\geq \log p(x; \theta) - D(q(y) \| p(y|x; \theta)) \\ &= \int dq(y) [\log p(x; \theta) + \log p(y|x; \theta) - \log q(y)] \\ &= \int dq(y) \log p(x, y; \theta) - \int dq(y) \log q(y)\end{aligned}$$

**This inequality is tight for  $p(y|x) = q(y)$**

- **Expectation step**

$$q(y) = p(y|x; \theta)$$

**find bound**

# Expectation Maximization

- **Variational Bound**

$$\begin{aligned}\log p(x; \theta) &\geq \log p(x; \theta) - D(q(y) \| p(y|x; \theta)) \\ &= \int dq(y) [\log p(x; \theta) + \log p(y|x; \theta) - \log q(y)] \\ &= \int dq(y) \log p(x, y; \theta) - \int dq(y) \log q(y)\end{aligned}$$

This inequality is tight for  $p(y|x) = q(y)$

- **Expectation step**

$$q(y) = p(y|x; \theta)$$

find bound

- **Maximization step**

$$\theta^* = \operatorname{argmax}_{\theta} \int dq(y) \log p(x, y; \theta)$$

# Expectation Maximization

- **Variational Bound**

$$\begin{aligned}\log p(x; \theta) &\geq \log p(x; \theta) - D(q(y) \| p(y|x; \theta)) \\ &= \int dq(y) [\log p(x; \theta) + \log p(y|x; \theta) - \log q(y)] \\ &= \int dq(y) \log p(x, y; \theta) - \int dq(y) \log q(y)\end{aligned}$$

This inequality is tight for  $p(y|x) = q(y)$

- **Expectation step**

$$q(y) = p(y|x; \theta)$$

find bound

- **Maximization step**

maximize it

$$\theta^* = \operatorname{argmax}_{\theta} \int dq(y) \log p(x, y; \theta)$$



# Expectation Step

- Factorizing distribution

$$q(Y) = \prod_i q_i(y)$$

- E-Step

$q_i(y) \propto p(x_i|y_i, \mu, \sigma)p(y_i|\theta)$  hence

$$m_{iy} := \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_y|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (x_i - \mu_y) \Sigma_y^{-1} (x_i - \mu_y) \right] p(y)$$

$$q_i(y) = \frac{m_{iy}}{\sum_{y'} m_{iy'}}$$

# Maximization Step

- **Log-likelihood**

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

- **Cluster distribution  
(weighted Gaussian MLE)**

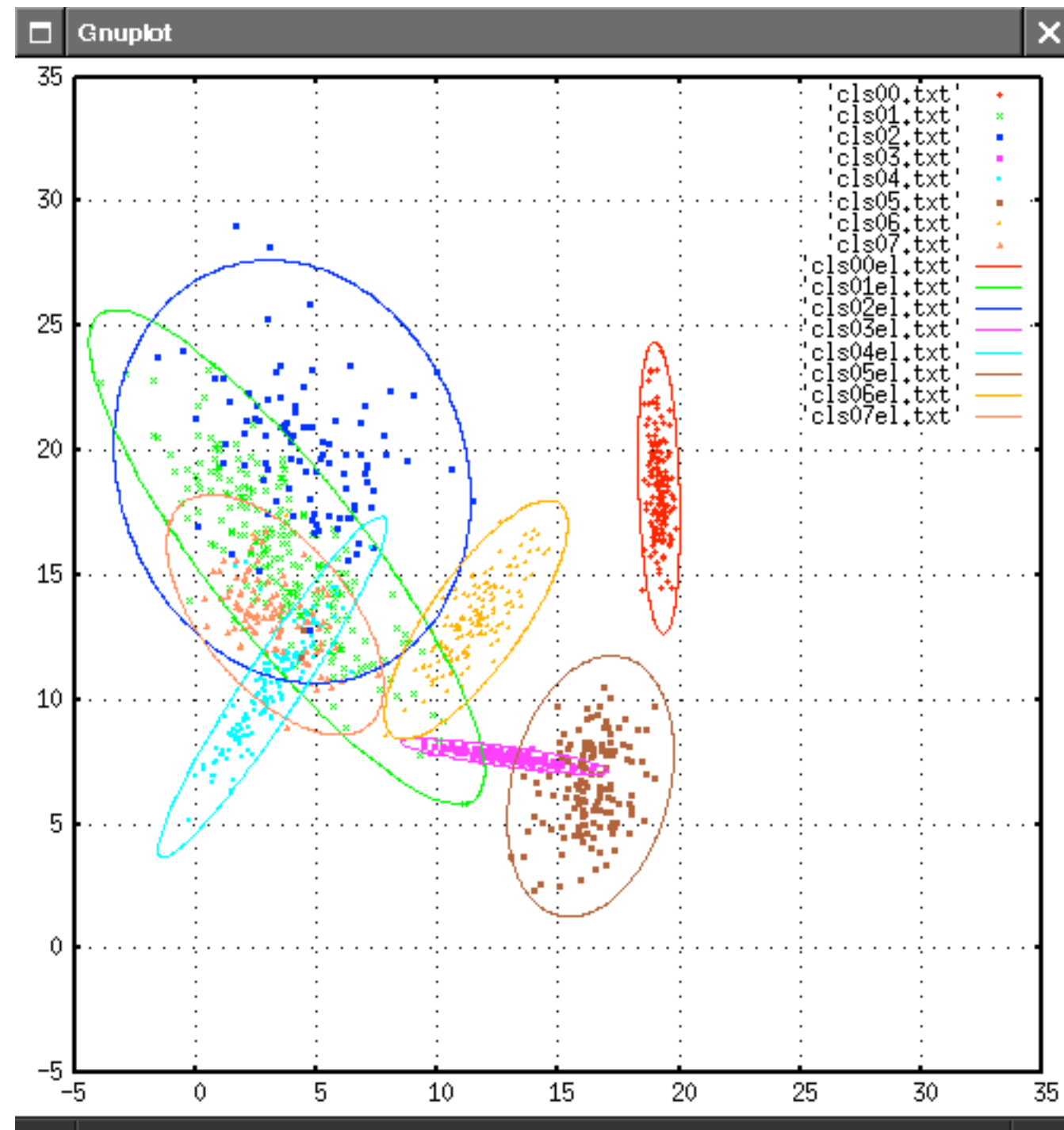
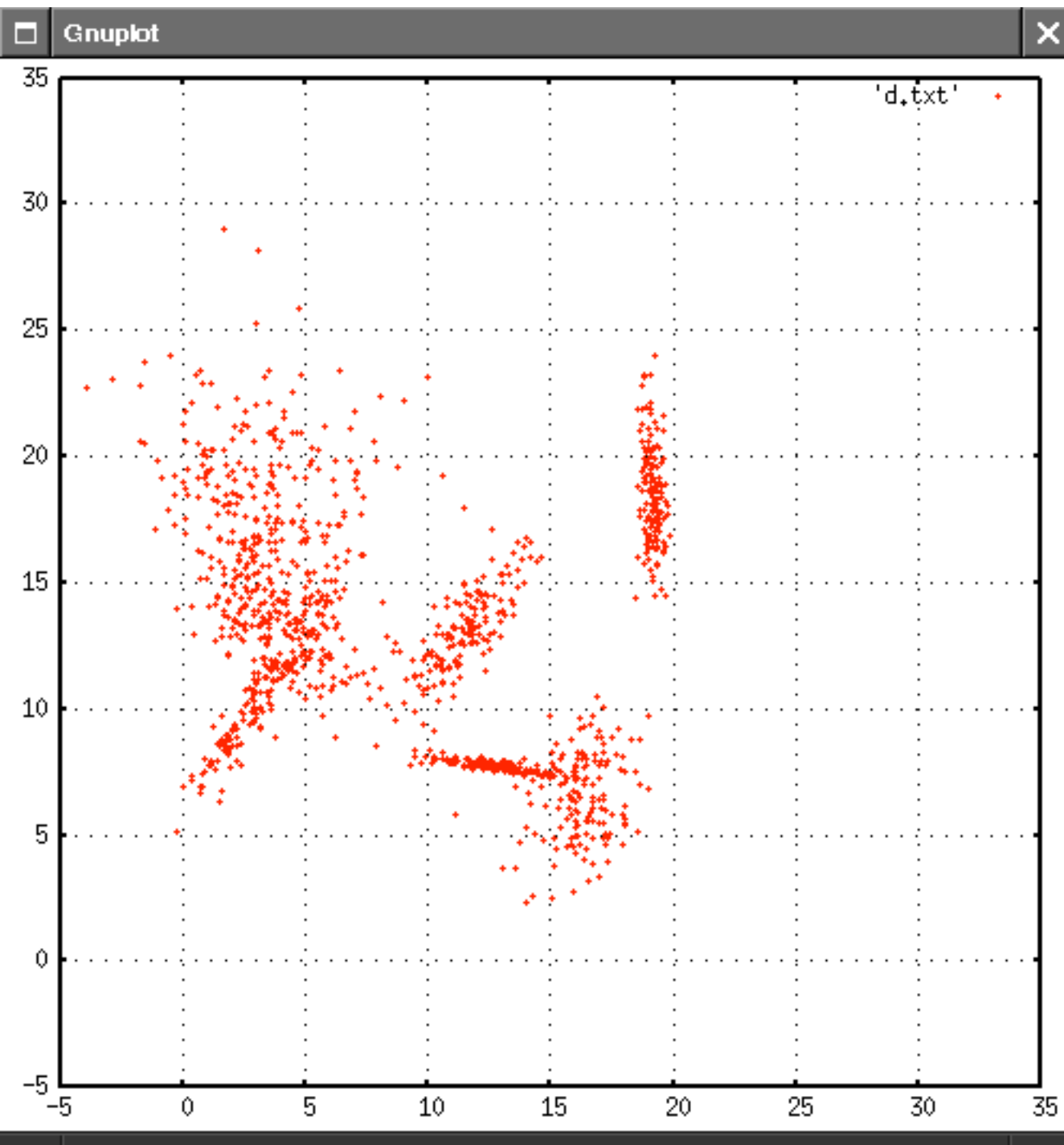
$$n_y = \sum_i q_i(y)$$

$$\mu_y = \frac{1}{n_y} \sum_{i=1}^n q_i(y) x_i$$
$$\Sigma_y = \frac{1}{n_y} \sum_{i=1}^n q_i(y) x_i x_i^\top - \mu_y \mu_y^\top$$

- **Cluster probabilities**

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{i=1}^n \sum_y q_i(y) \log p(y_i | \theta) \text{ hence } p(y | \theta) = \frac{n_y}{n}$$

# EM Clustering in action



# Problem

**Estimates will diverge  
(infinite variance, zero probability, tiny clusters)**

# Solution

- Use priors for  $\mu, \sigma, \theta$
- Dirichlet distribution for cluster probabilities
- Gauss-Wishart for Gaussian

- Cluster distribution

$$n_y = n_0 + \sum_i q_i(y)$$

$$\mu_y = \frac{1}{n_y} \sum_{i=1}^n q_i(y) x_i$$

$$\Sigma_y = \frac{1}{n_y} \sum_{i=1}^n q_i(y) x_i x_i^\top + \frac{n_0}{n_y} \mathbf{1} - \mu_y \mu_y^\top$$

- Cluster probabilities

$$p(y|\theta) = \frac{n_y}{n + k \cdot n_0}$$

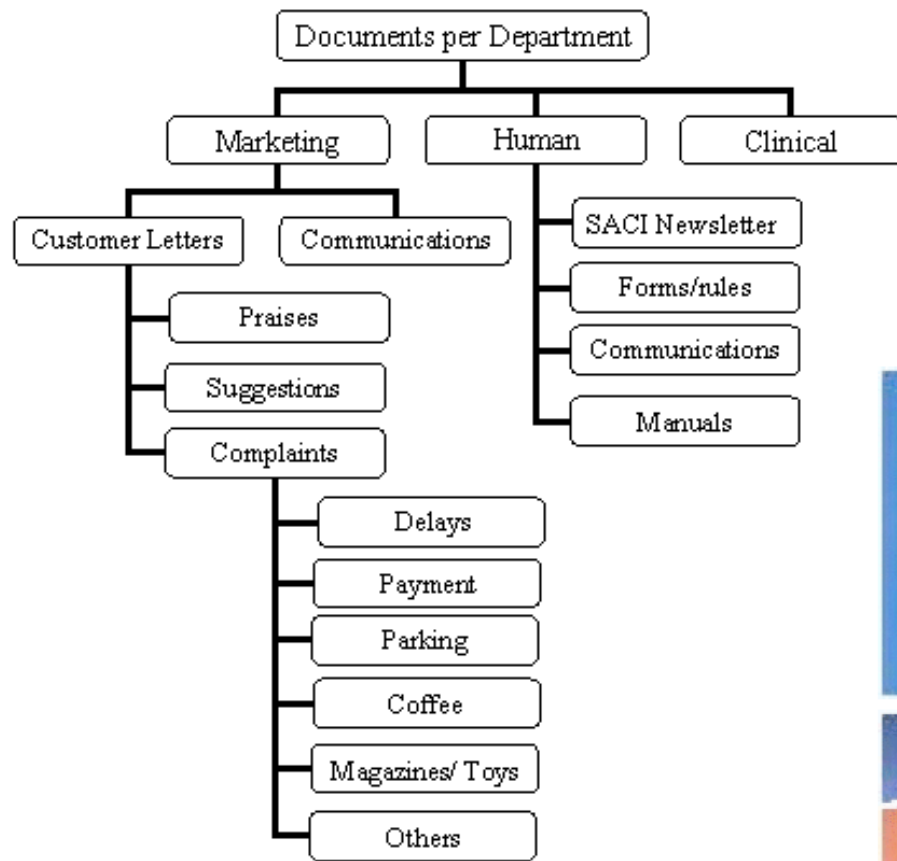
# Variational Approximation

- Lower bound on log-likelihood

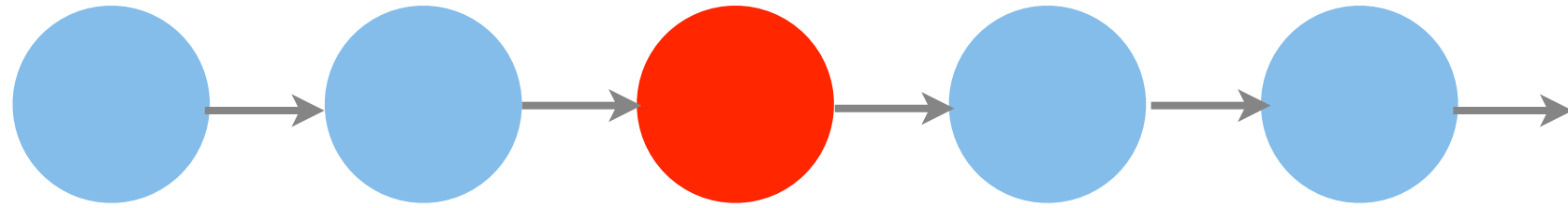
$$\log p(x; \theta) \geq \int dq(y) \log p(x, y; \theta) - \int dq(y) \log q(y)$$

- Key insight - inequality holds for any  $q$ 
  - Find  $q$  within subset  $Q$  to tighten inequality
  - Find parameters to maximize for fixed  $q$
- Inference for graphical models where joint probability computation is infeasible

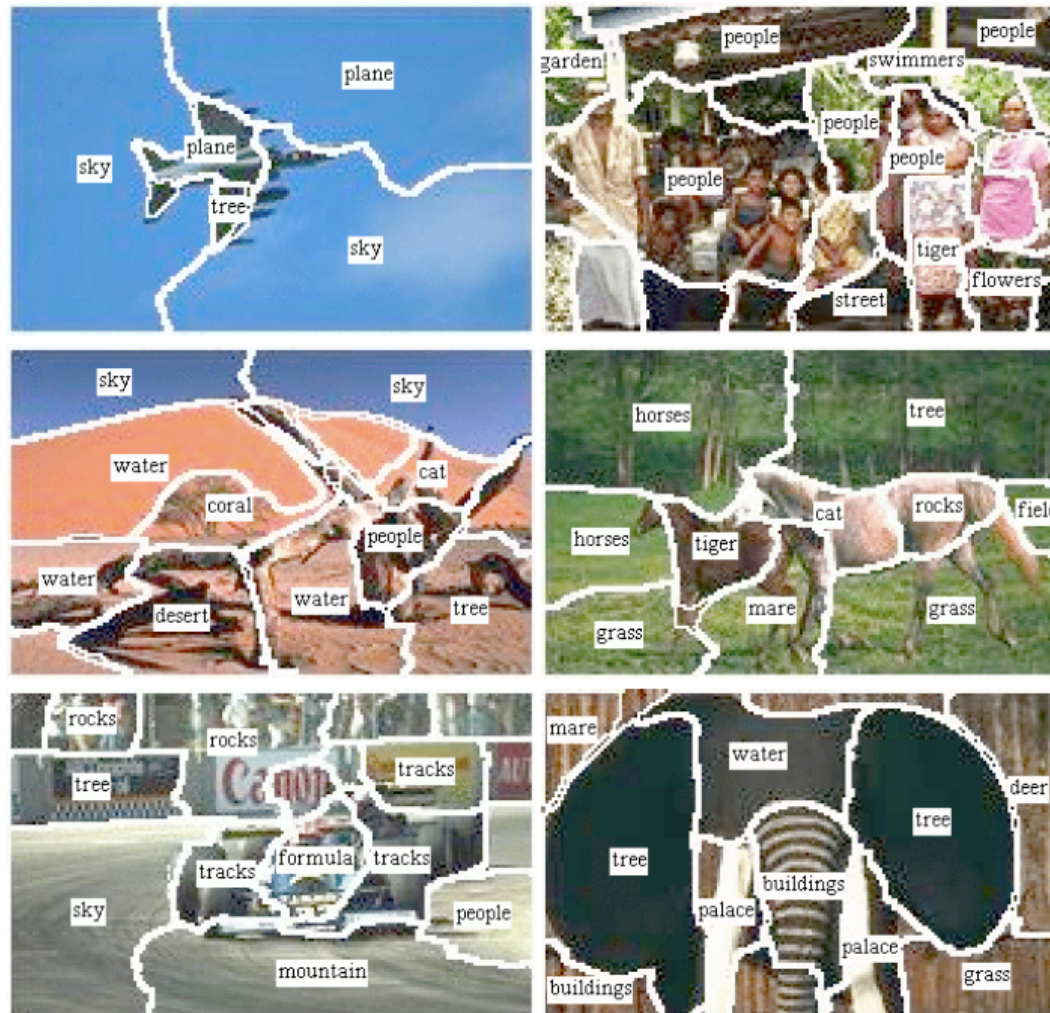
# Beyond mixtures



taxonomies



chains



topics

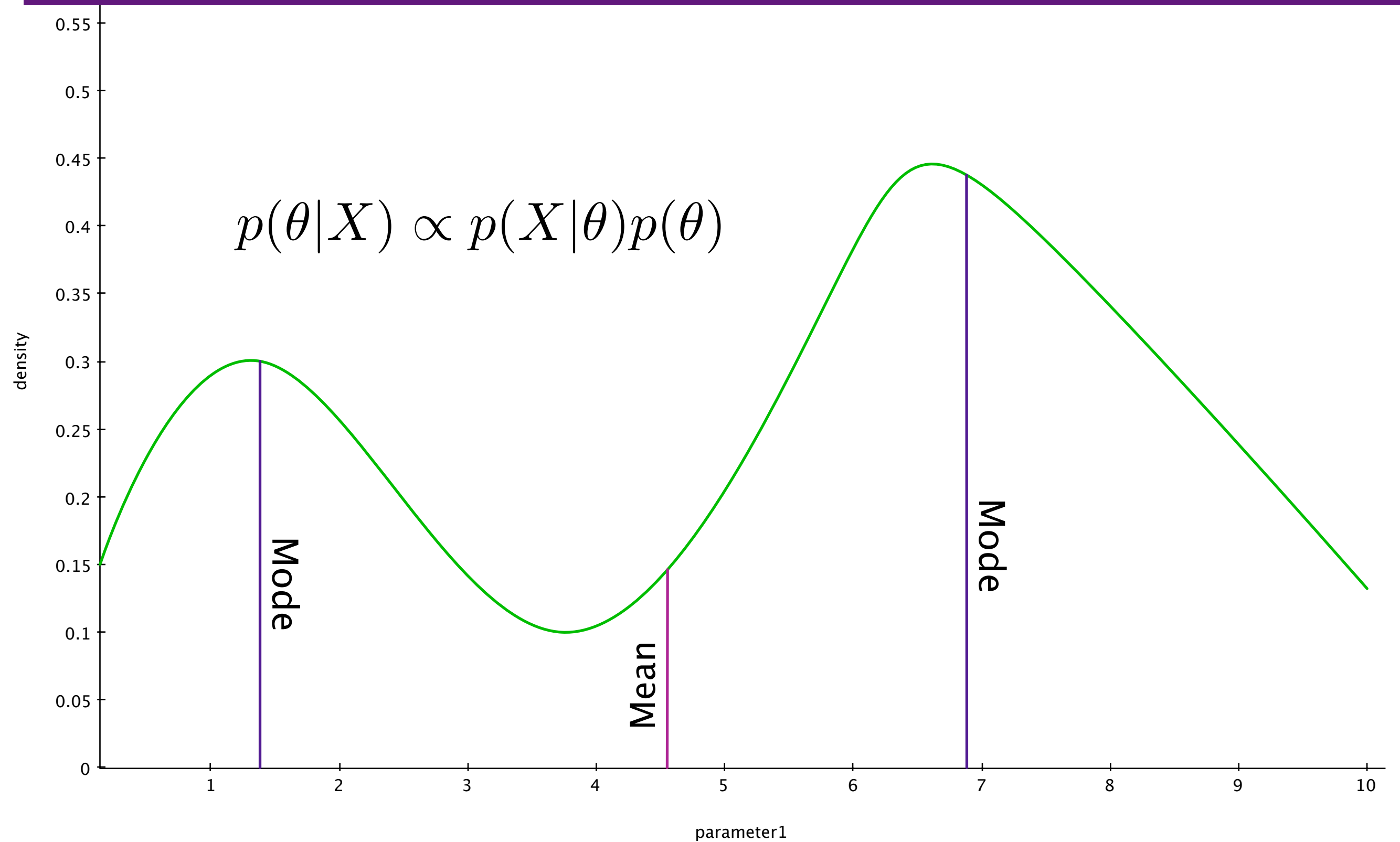


# Sampling



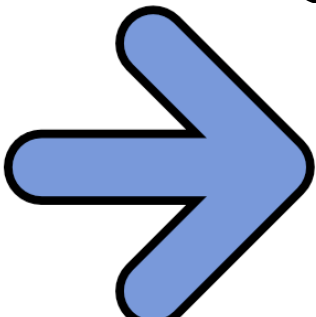



# Is maximization (always) good?








# Sampling

- Key idea
  - Want **accurate** distribution of the posterior
  - **Sample** from posterior distribution rather than **maximizing** it
- Problem - direct sampling is usually intractable
- Solutions
  - Markov Chain Monte Carlo (complicated)
  - Gibbs Sampling (somewhat simpler)


$$x \sim p(x|x')$$
 and then  $x' \sim p(x'|x)$ 






# Gibbs sampling

- 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






- 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






- 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






- 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  $p(\cdot, g)$   
 $(g, g)$  - draw  $p(g, \cdot)$   
 $(g, g)$  - draw  $p(\cdot, g)$

# Gibbs sampling






- 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

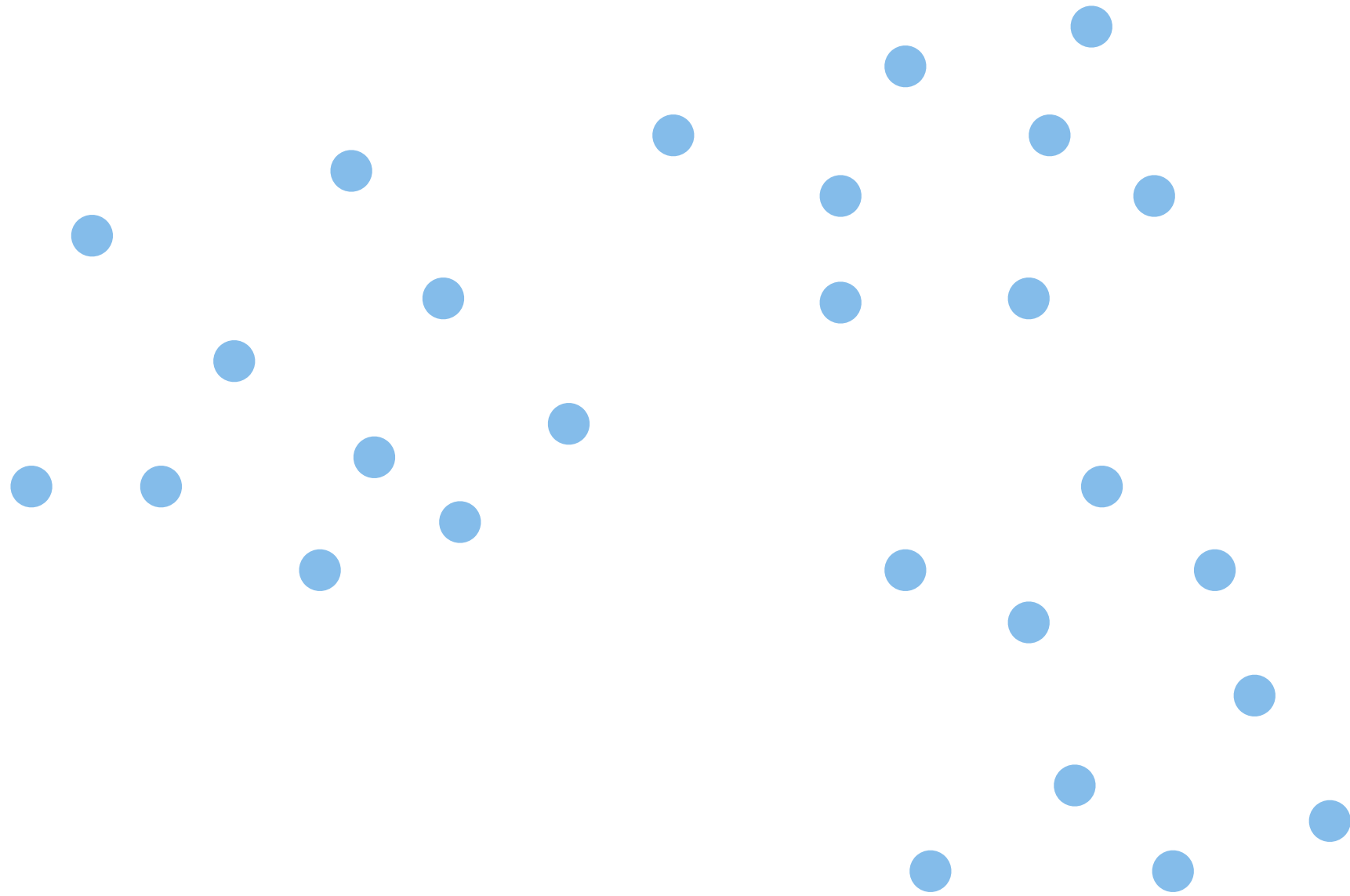
- Gibbs sampling:
  - In most cases direct sampling not possible
  - Draw one set of variables at a time

		
	0.45	0.05
	0.05	0.45

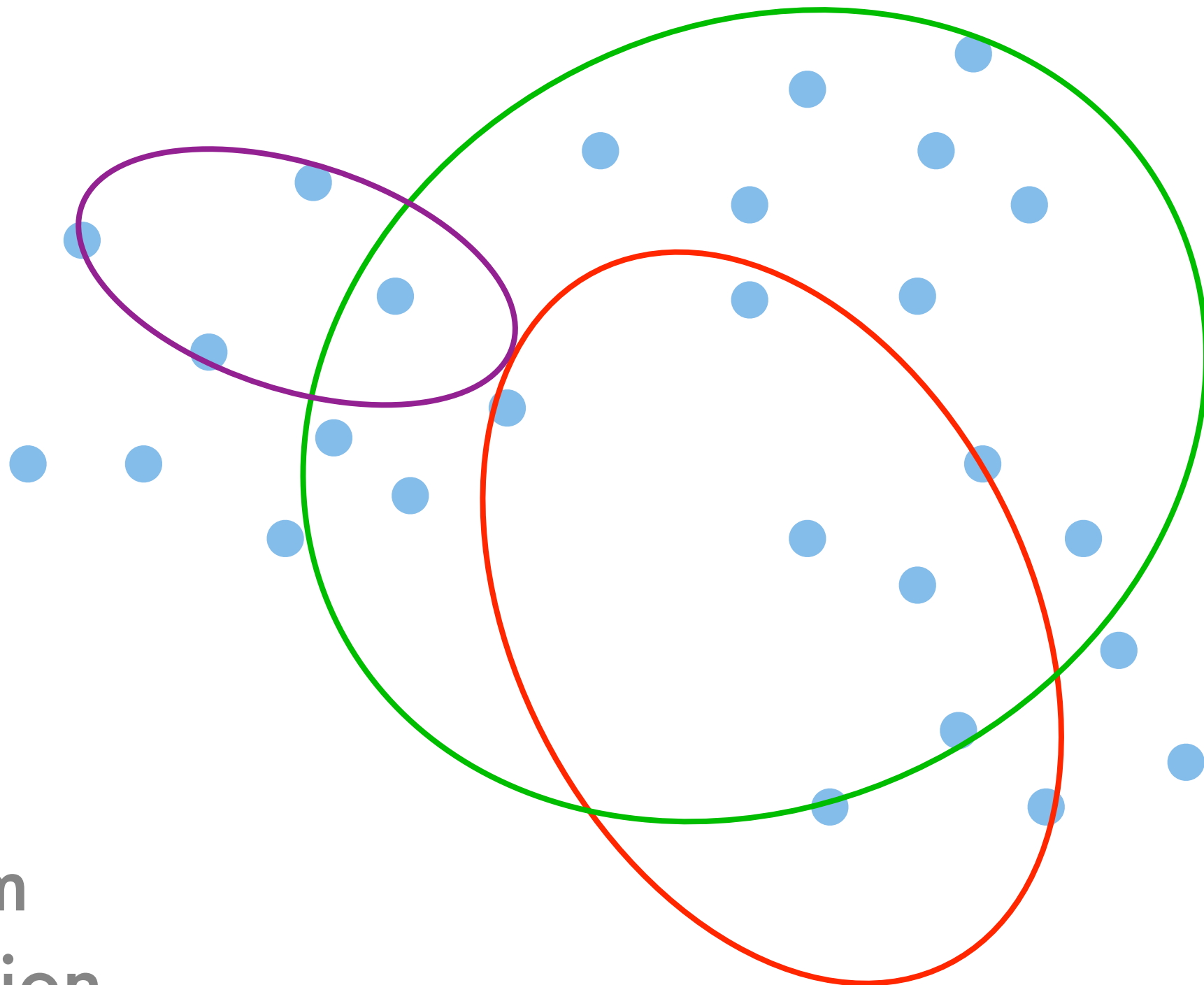
(b,g) - draw  $p(.,g)$   
(g,g) - draw  $p(g,.)$   
(g,g) - draw  $p(.,g)$   
(b,g) - draw  $p(b,.)$   
(b,b) ...



# Gibbs sampling for clustering

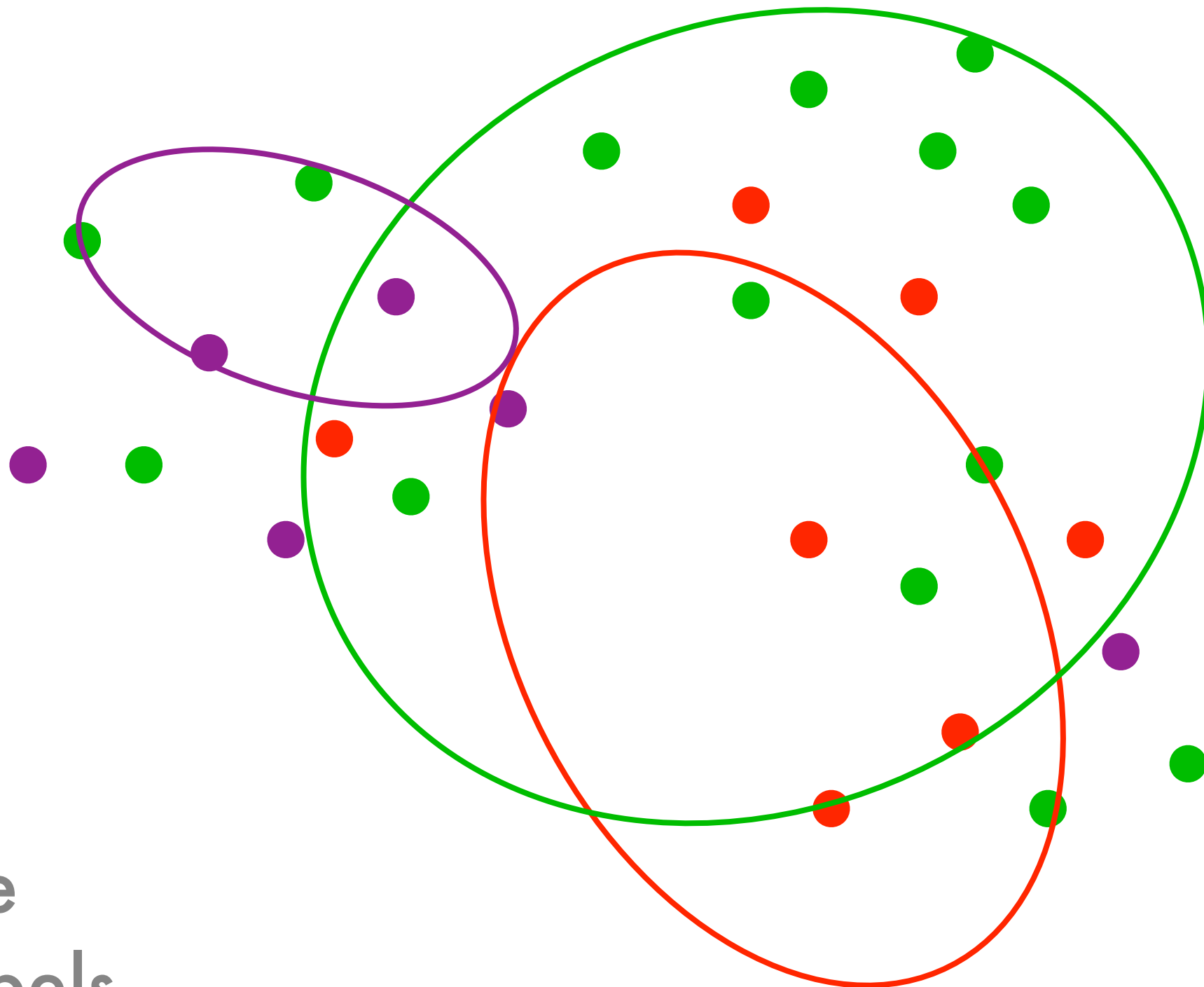


# Gibbs sampling for clustering



random  
initialization

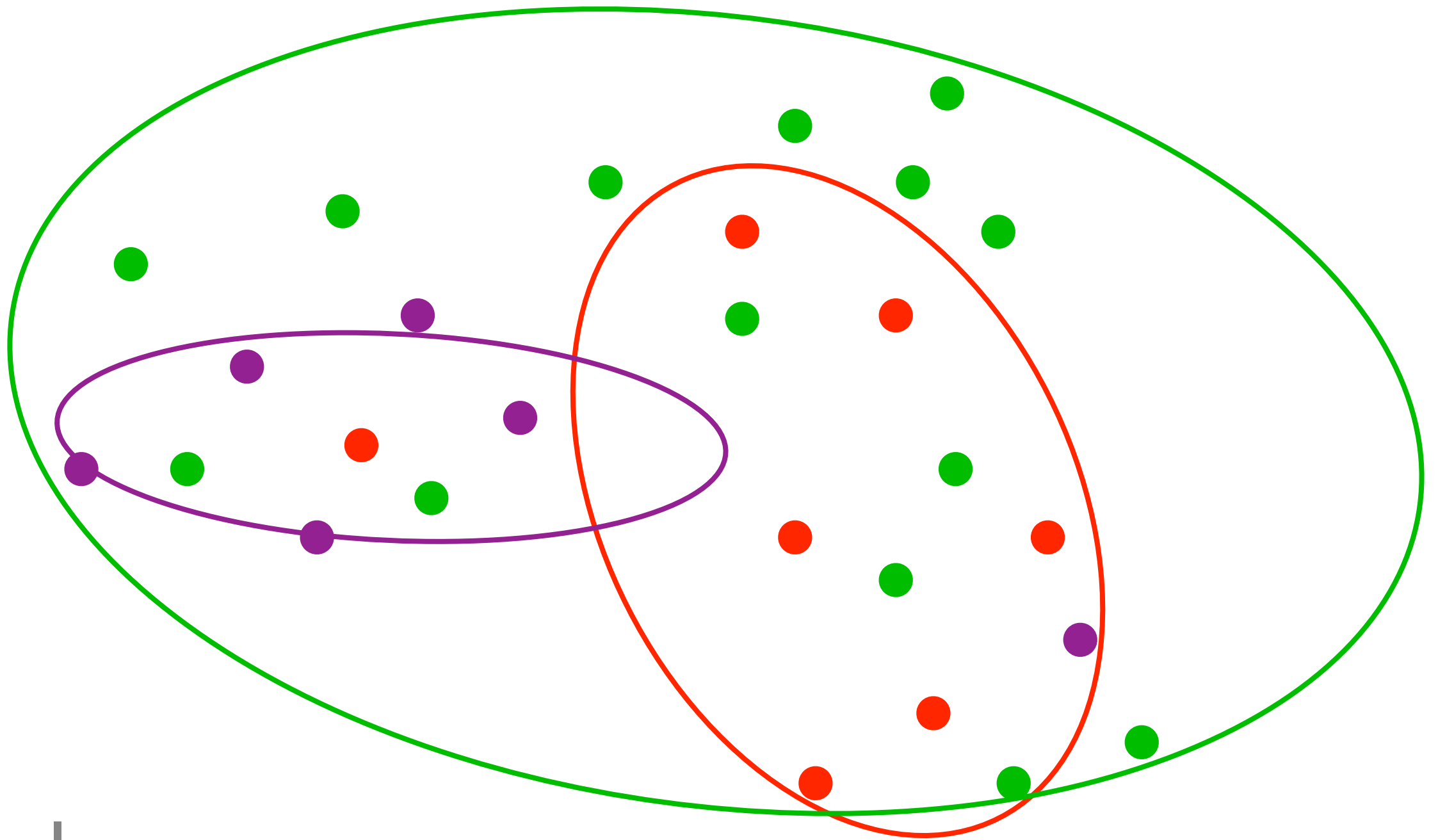
# Gibbs sampling for clustering



sample

cluster labels

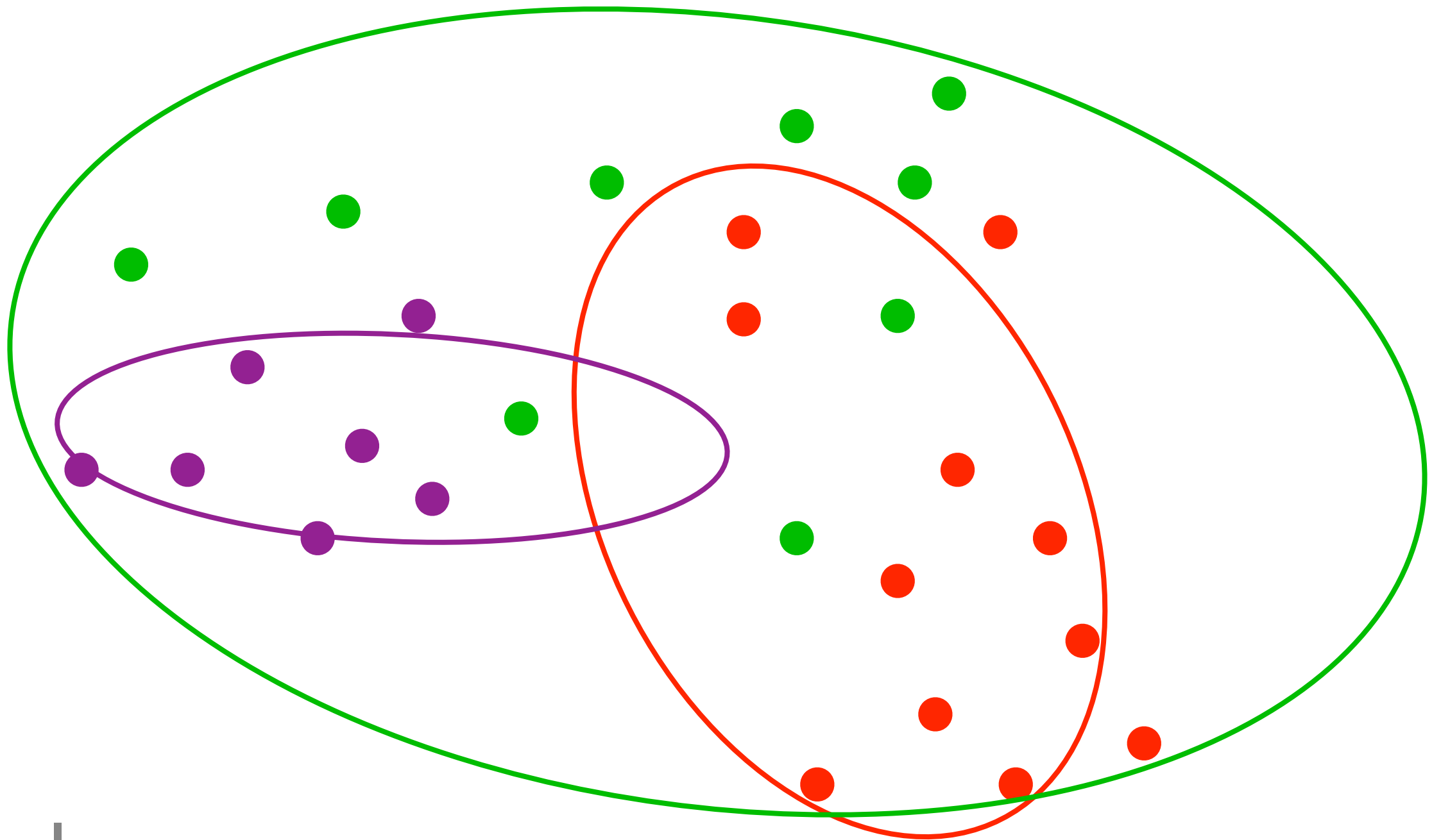
# Gibbs sampling for clustering



resample

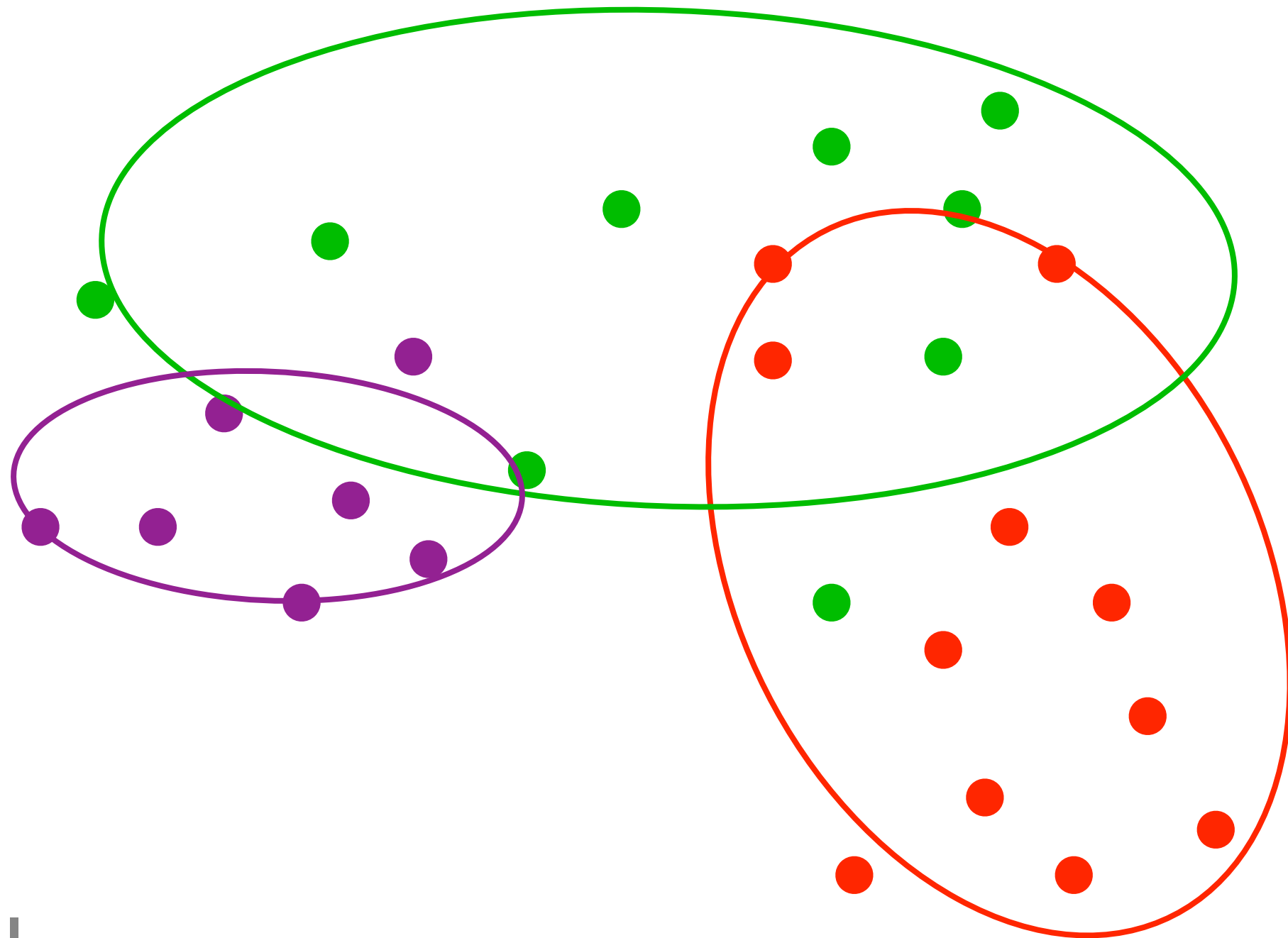
cluster model

# Gibbs sampling for clustering



resample  
cluster labels

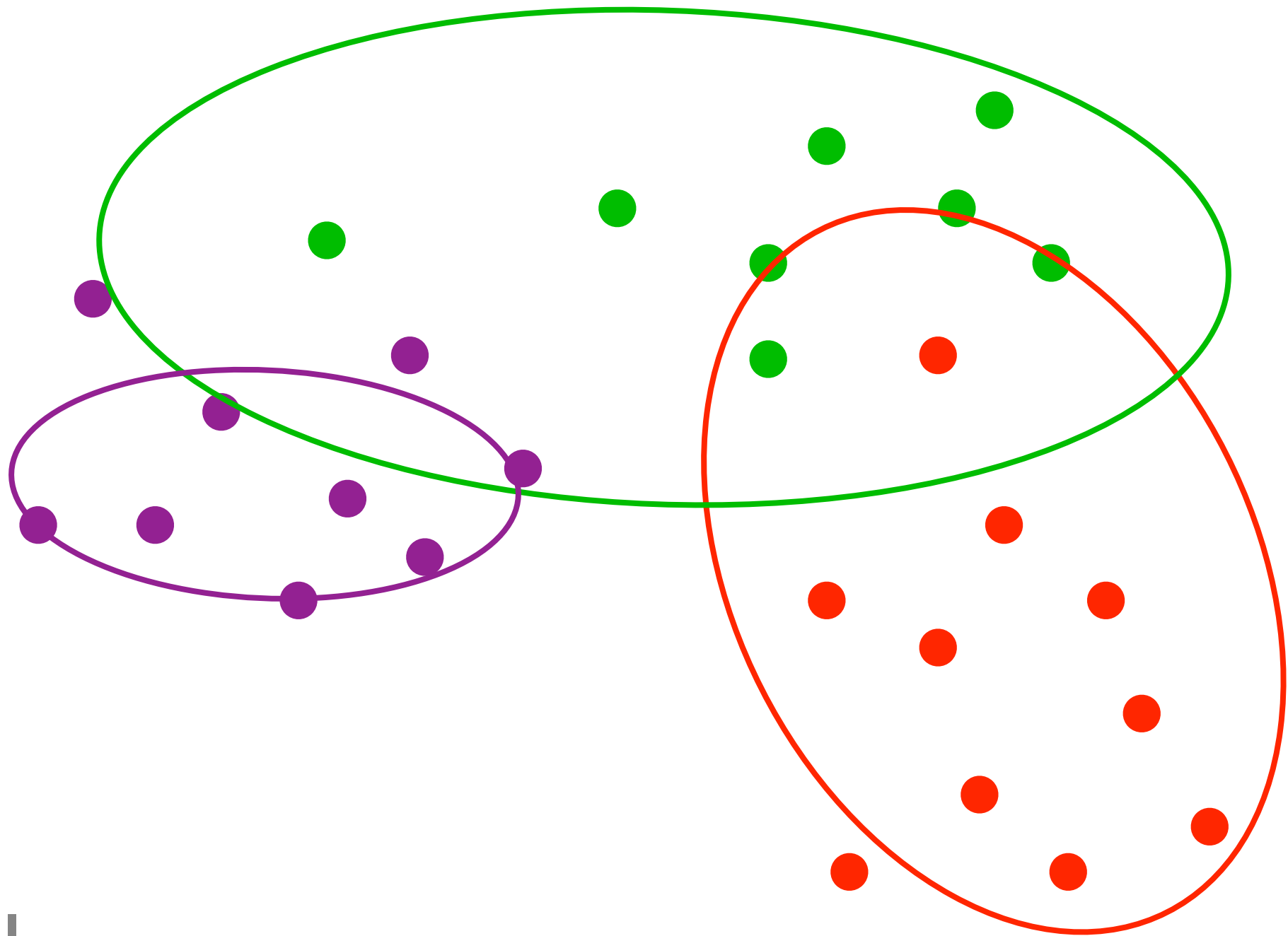
# Gibbs sampling for clustering



resample

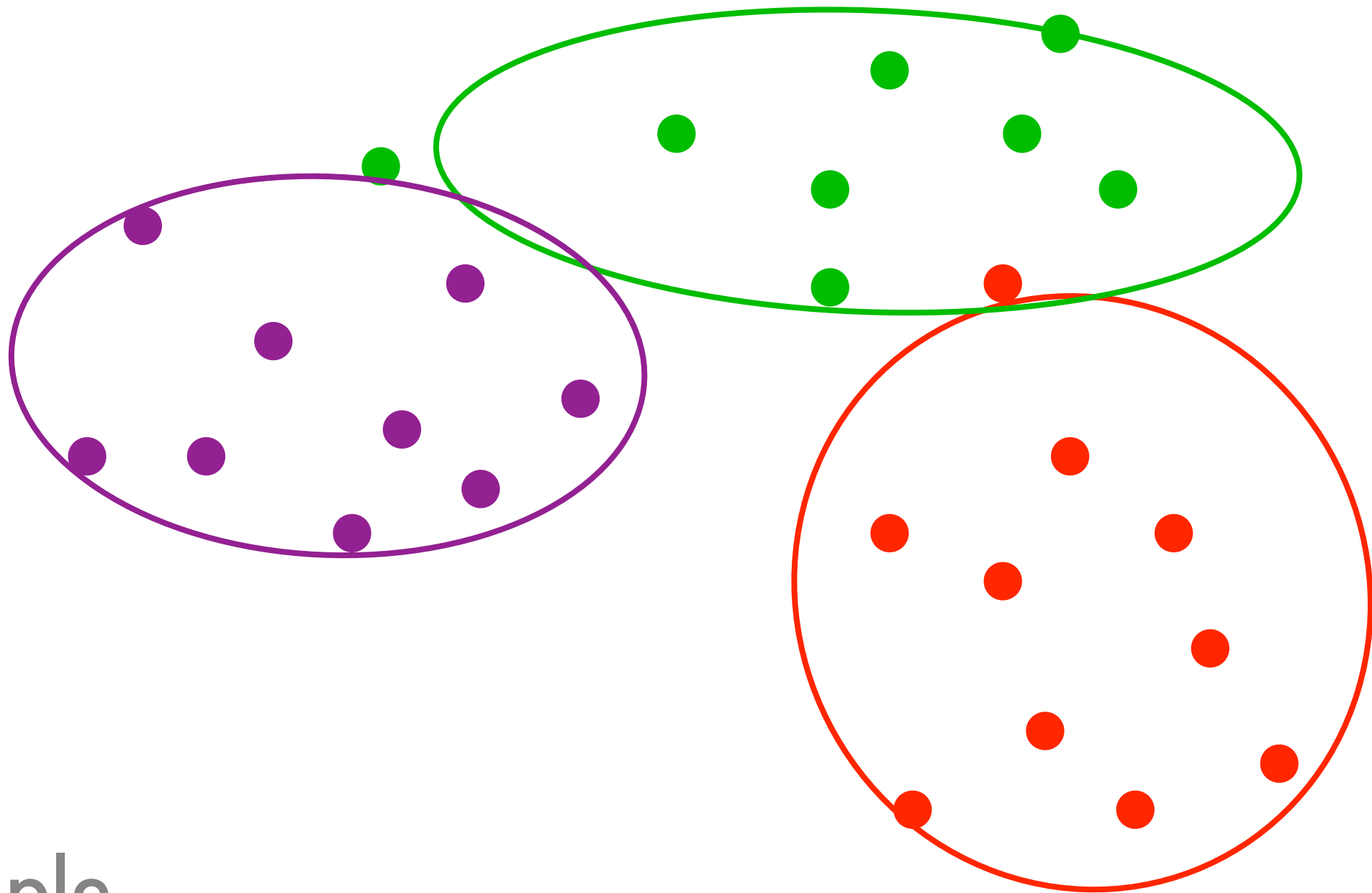
cluster model

# Gibbs sampling for clustering



resample  
cluster labels

# Gibbs sampling for clustering



resample

cluster model

e.g. Mahout Dirichlet Process Clustering



**Inference Algorithm  $\neq$  Model**

**Inference Algorithm  $\neq$  Model**

**Corollary: EM  $\neq$  Clustering**

# Graphical Models Zoology



**YAHOO!**<sup>®</sup>

Models

Statistics

**YAHOO!**<sup>®</sup>

Inference  
Methods

Efficient  
Computation

Models

Statistics

$l_1, l_2$  Priors

Conjugate Prior

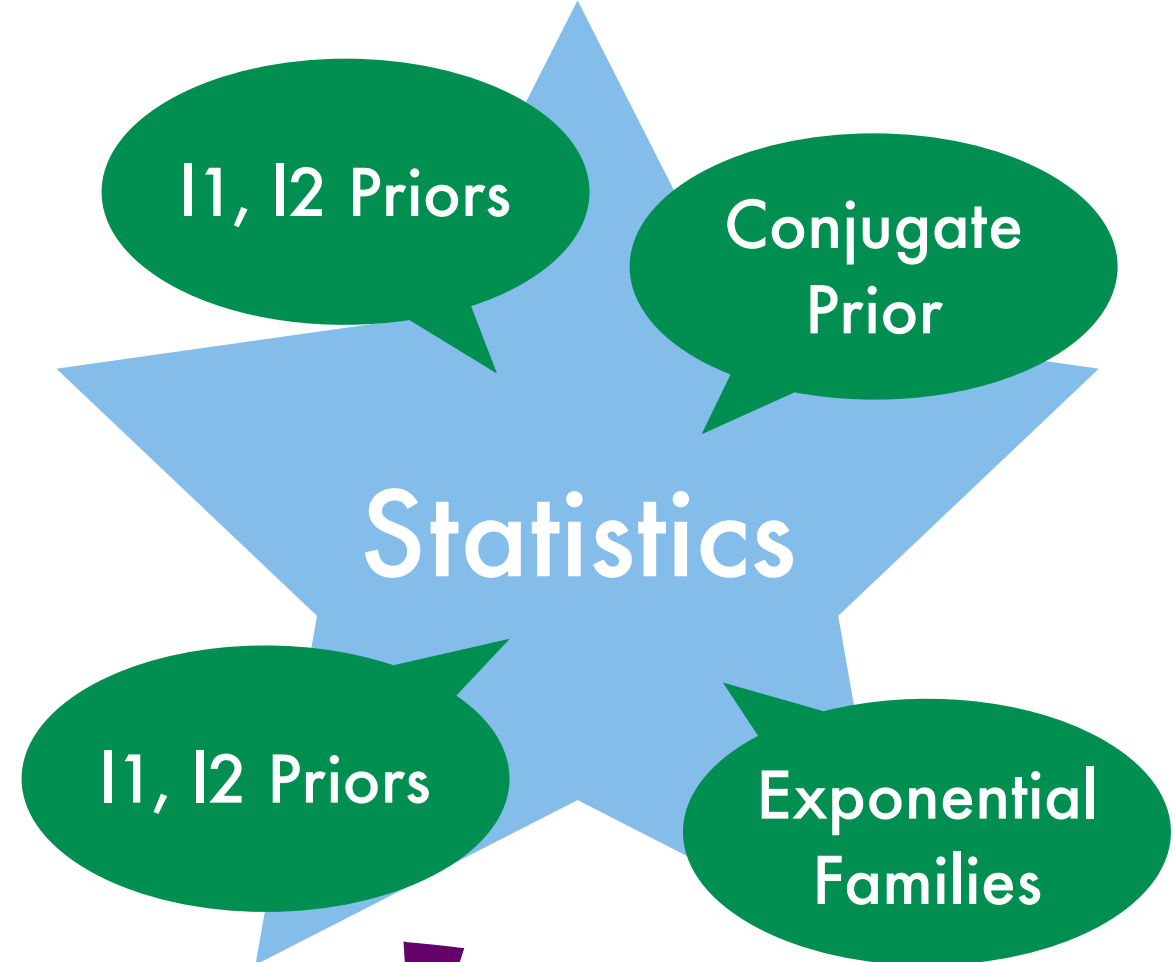
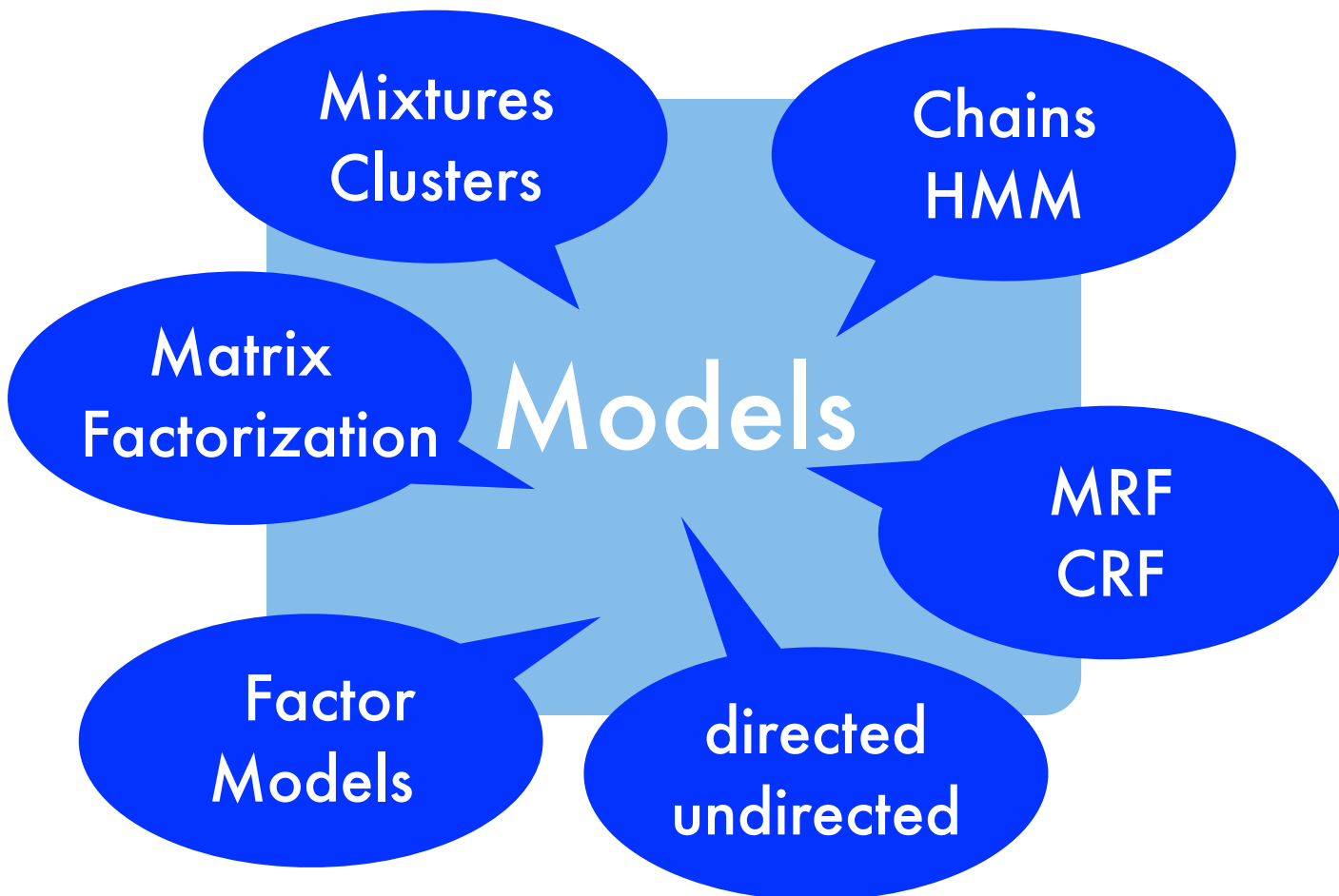
$l_1, l_2$  Priors

Exponential Families

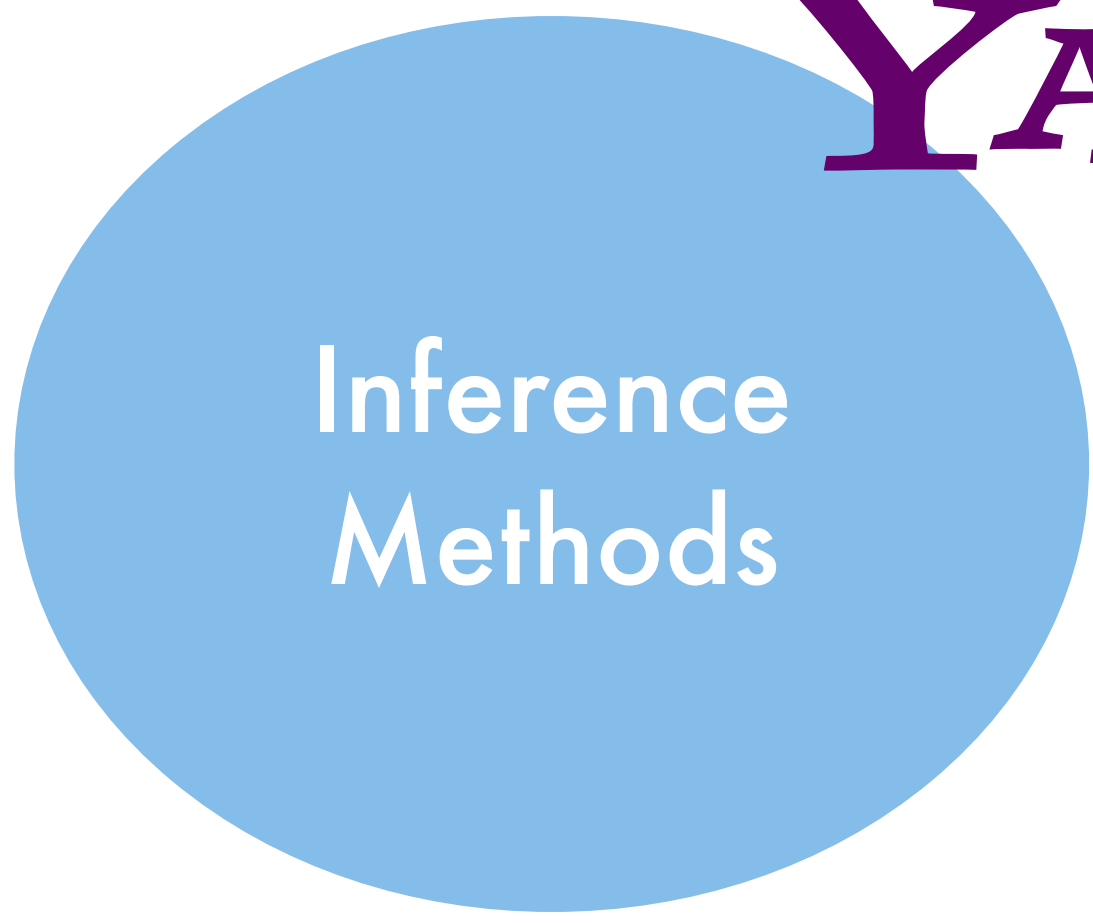
YAHOO!

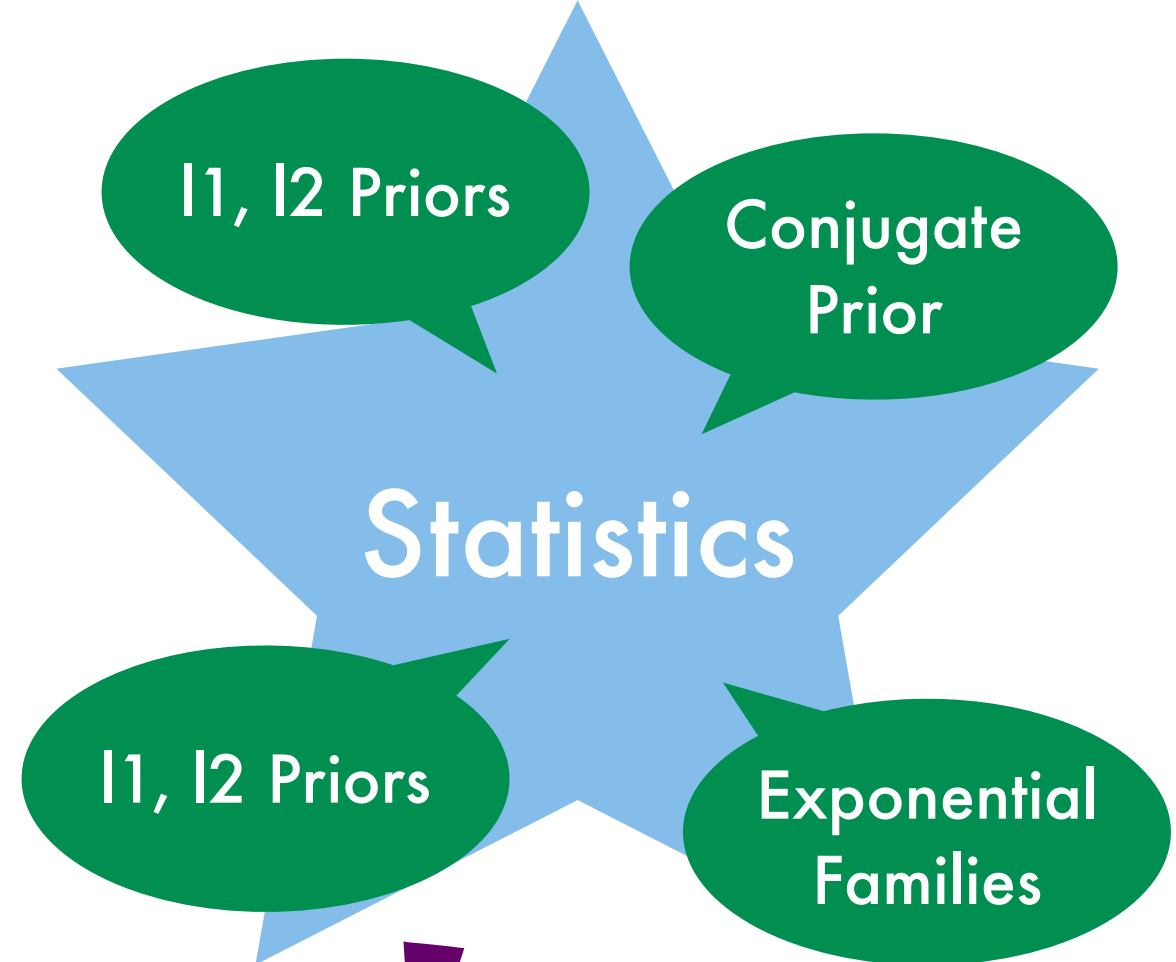
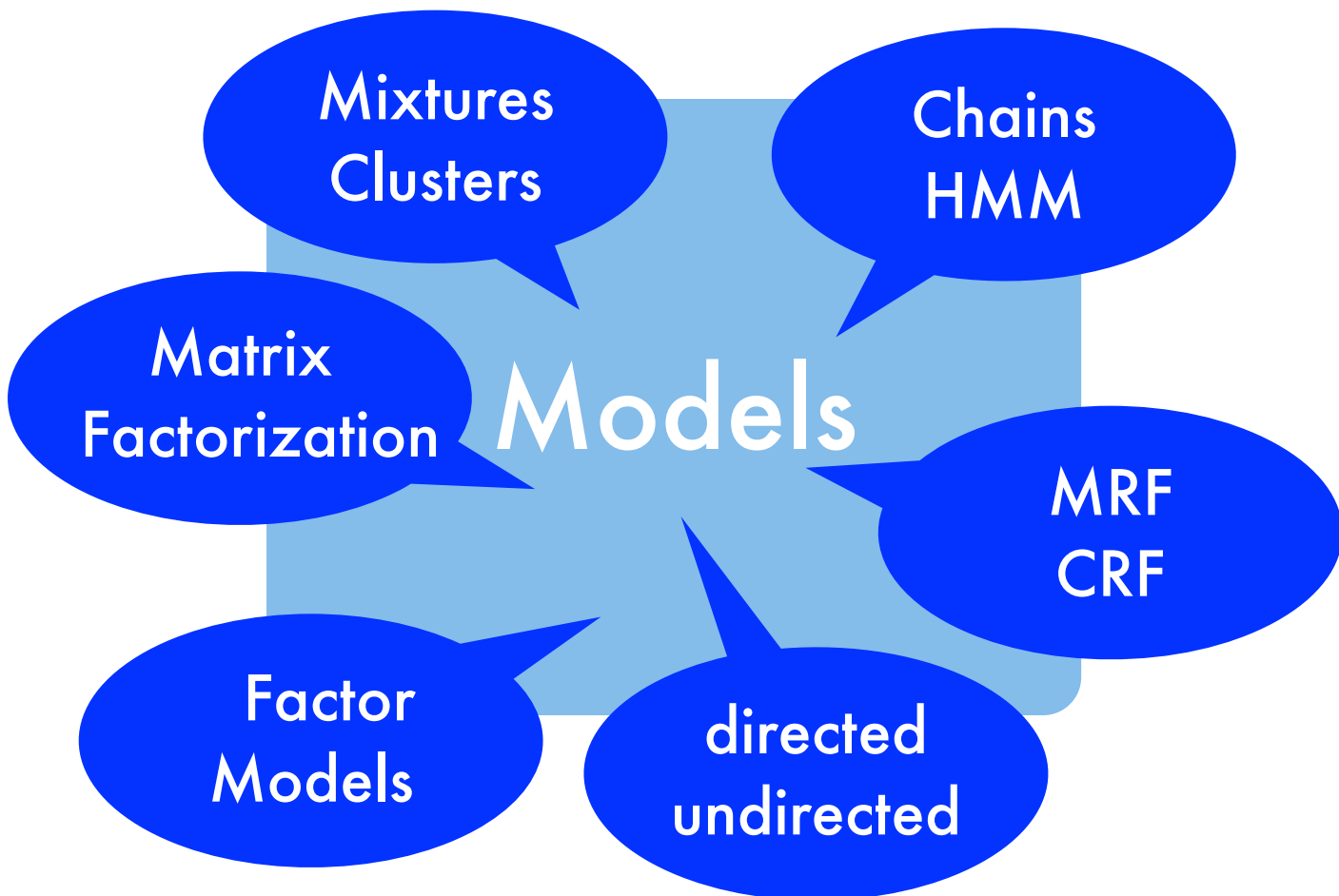
Inference Methods

Efficient Computation

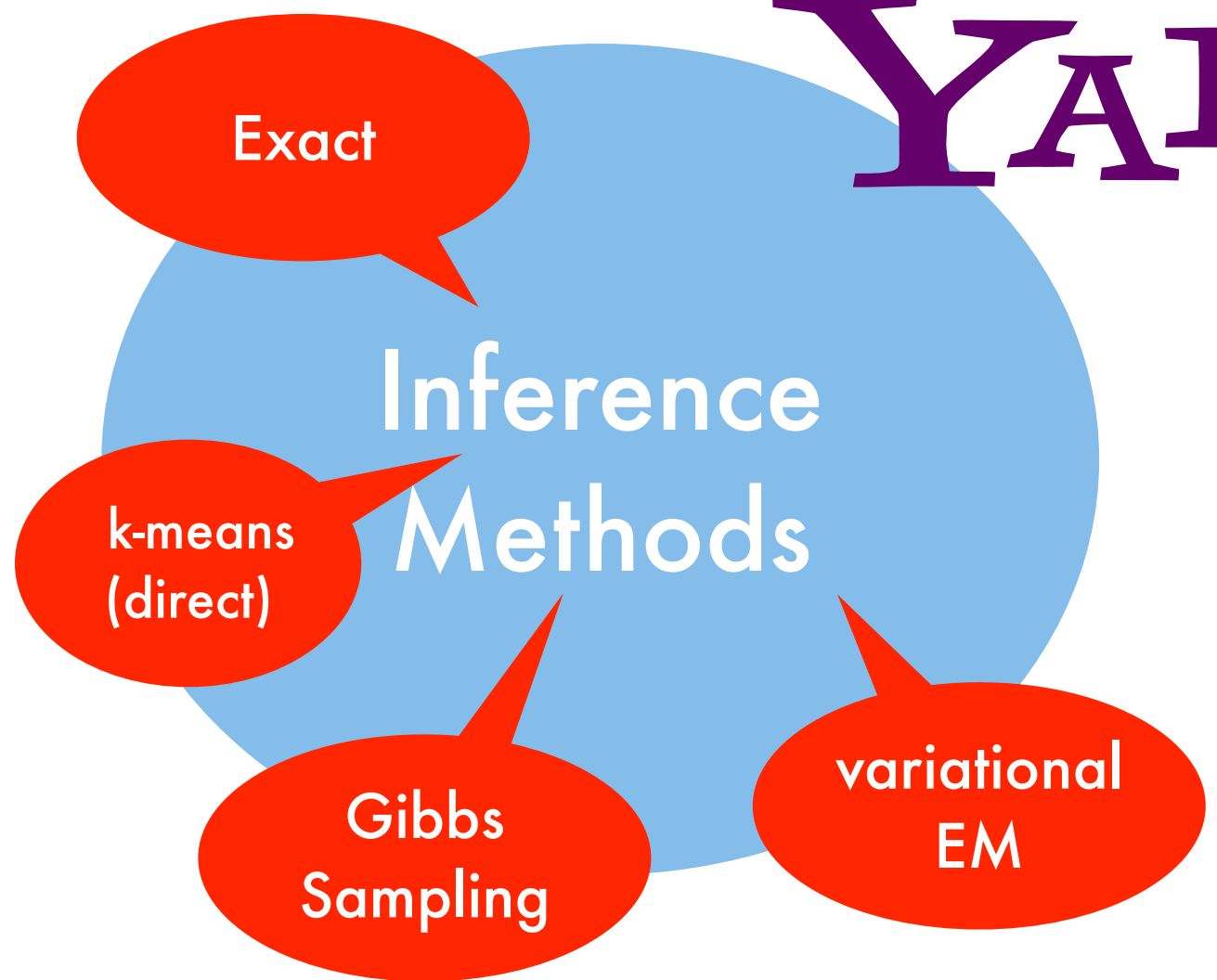


**YAHOO!**<sup>®</sup>

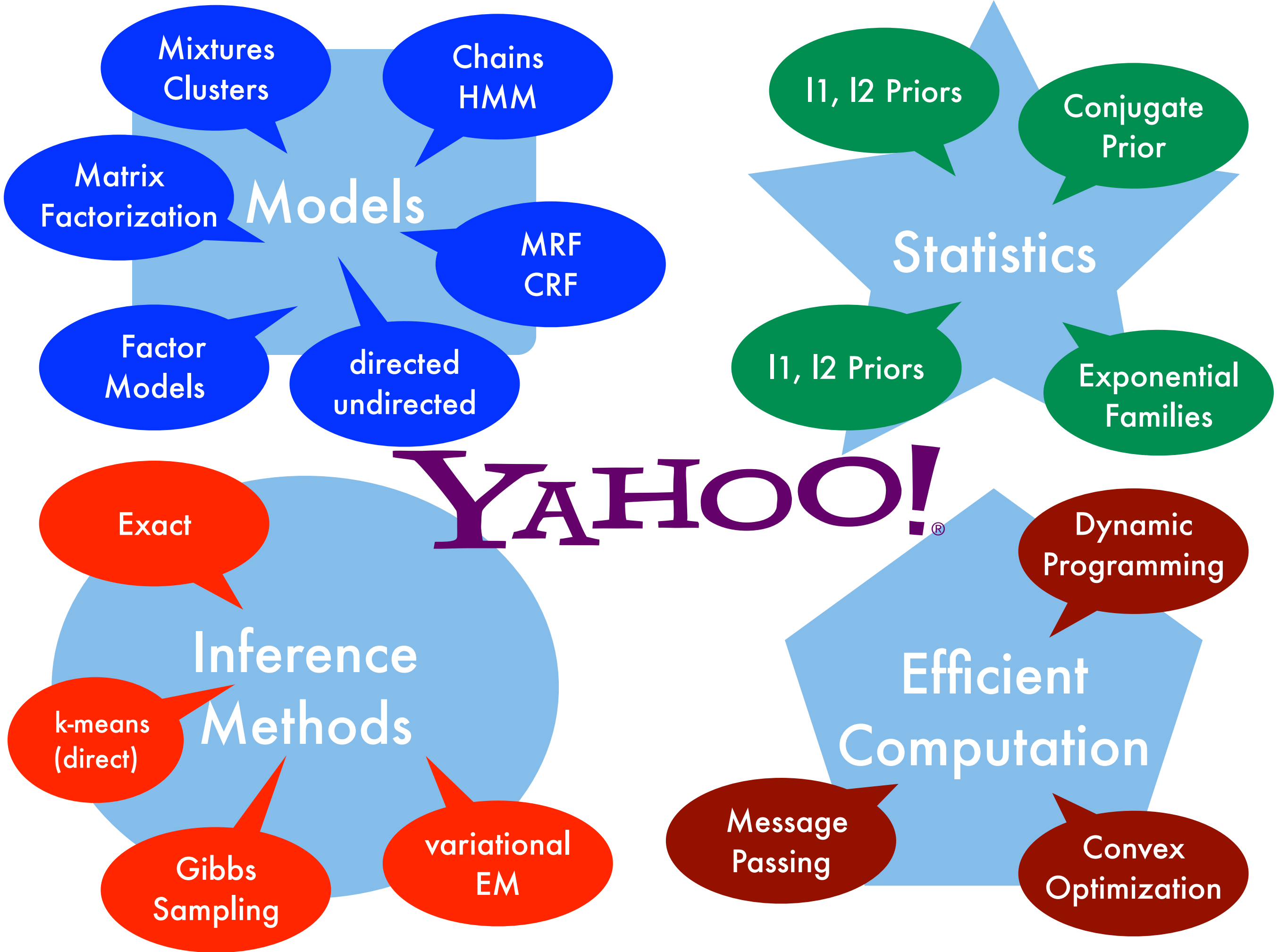




**YAHOO!**<sup>®</sup>







YAHOO!

# YAHOO!

Spam  
Filtering

Classification

Exploration

Segmentation

Annotation

System  
Design

Prediction  
(time series)

Clustering

Document  
Understanding

User  
Modeling

Advertising

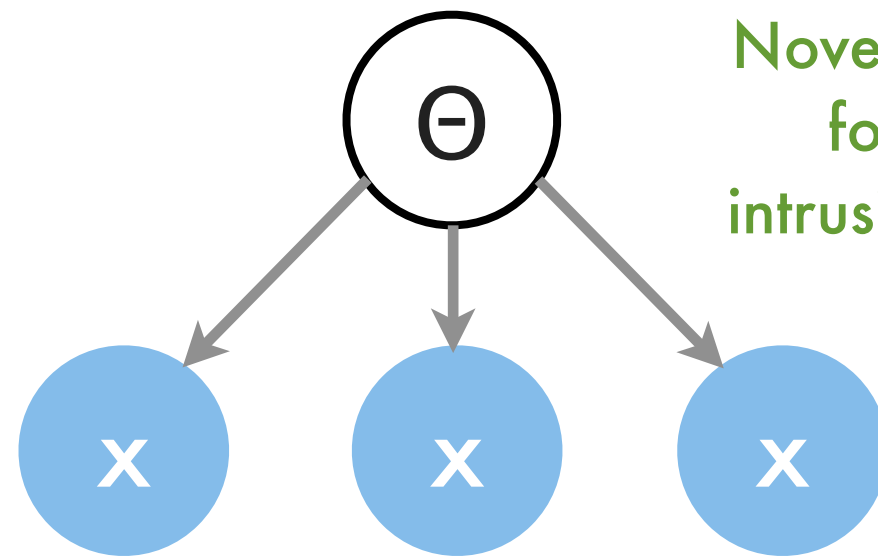
Novelty  
Detection

Performance  
Tuning

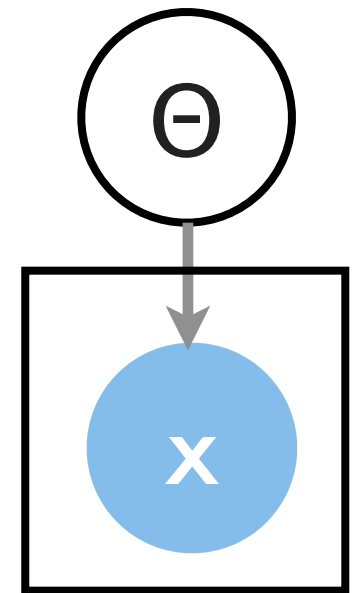
Debugging

# 'Unsupervised' Models

Density Estimation



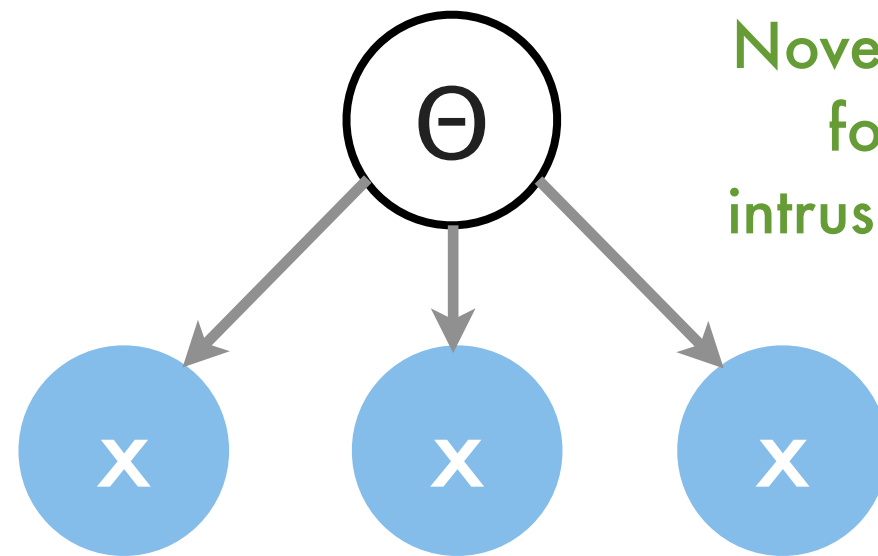
Novelty Detection  
forecasting  
intrusion detection



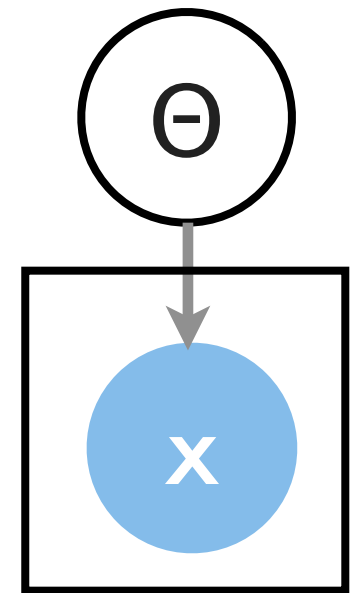
webpages  
news  
users  
ads  
queries  
images

# 'Unsupervised' Models

Density Estimation

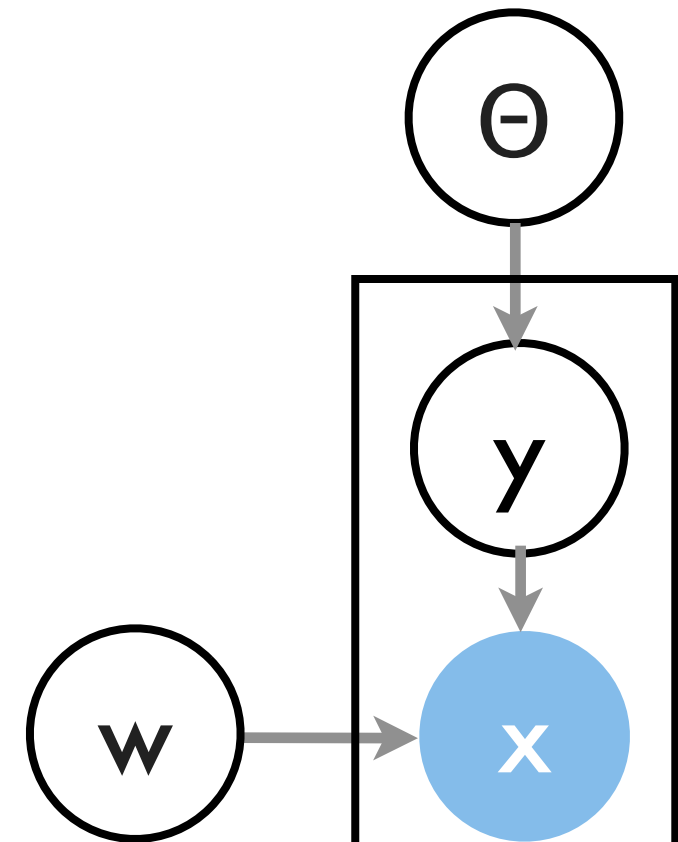
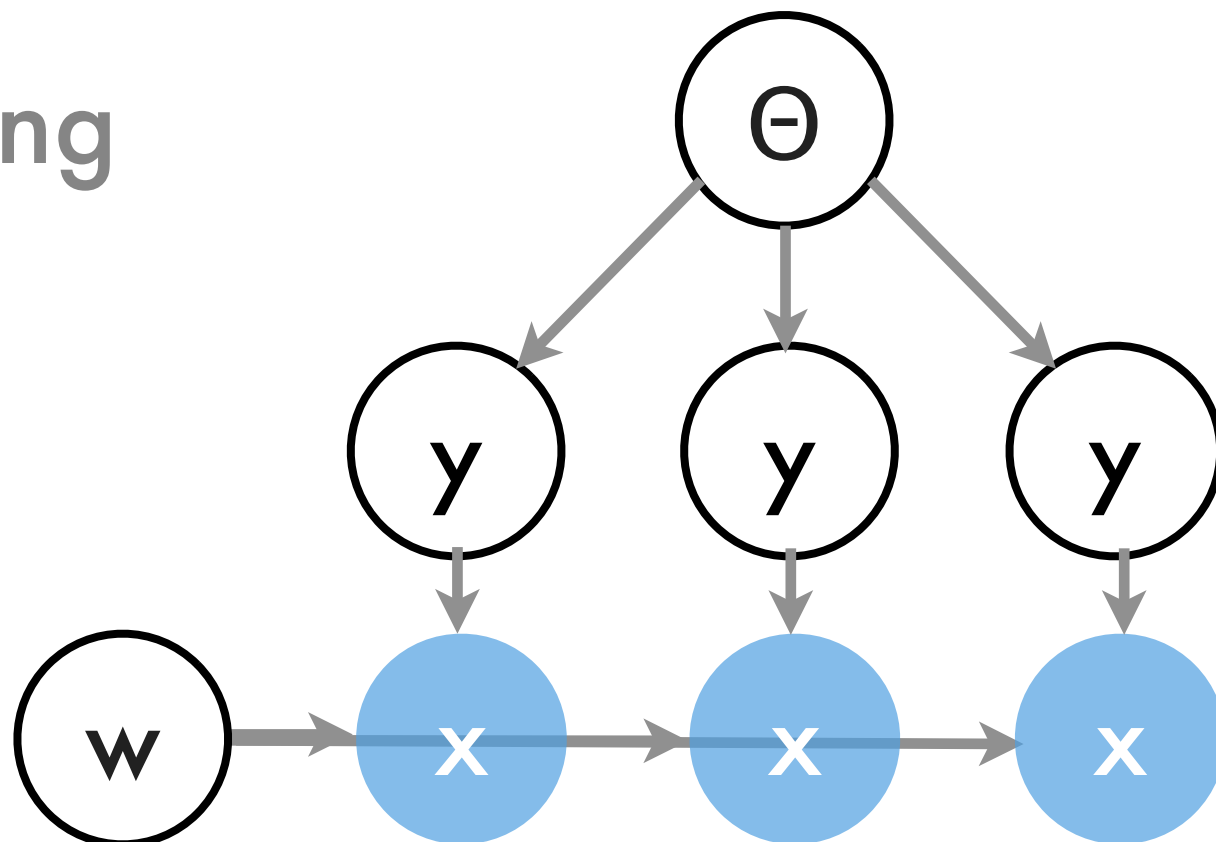


Novelty Detection  
forecasting  
intrusion detection



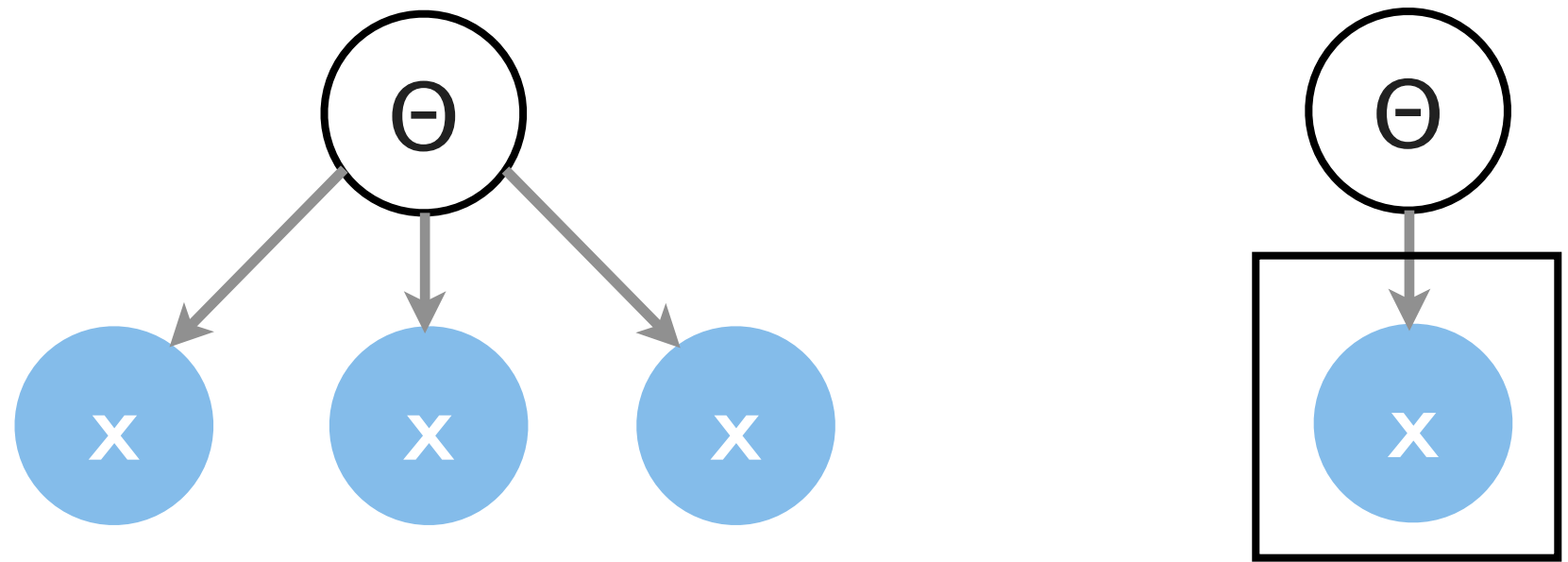
Clustering

webpages  
news  
users  
ads  
queries  
images

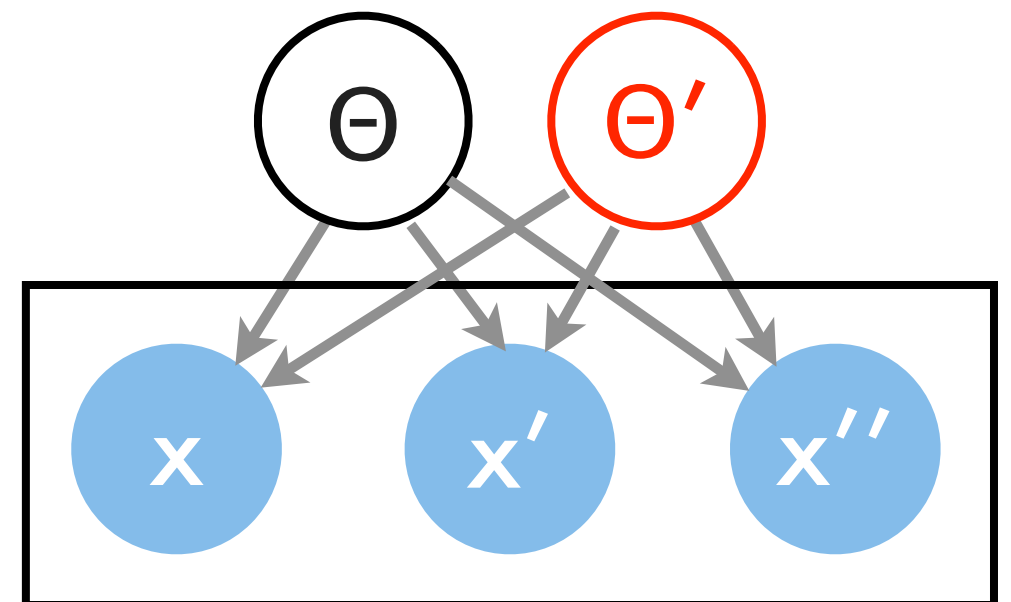


# 'Unsupervised' Models

Density Estimation



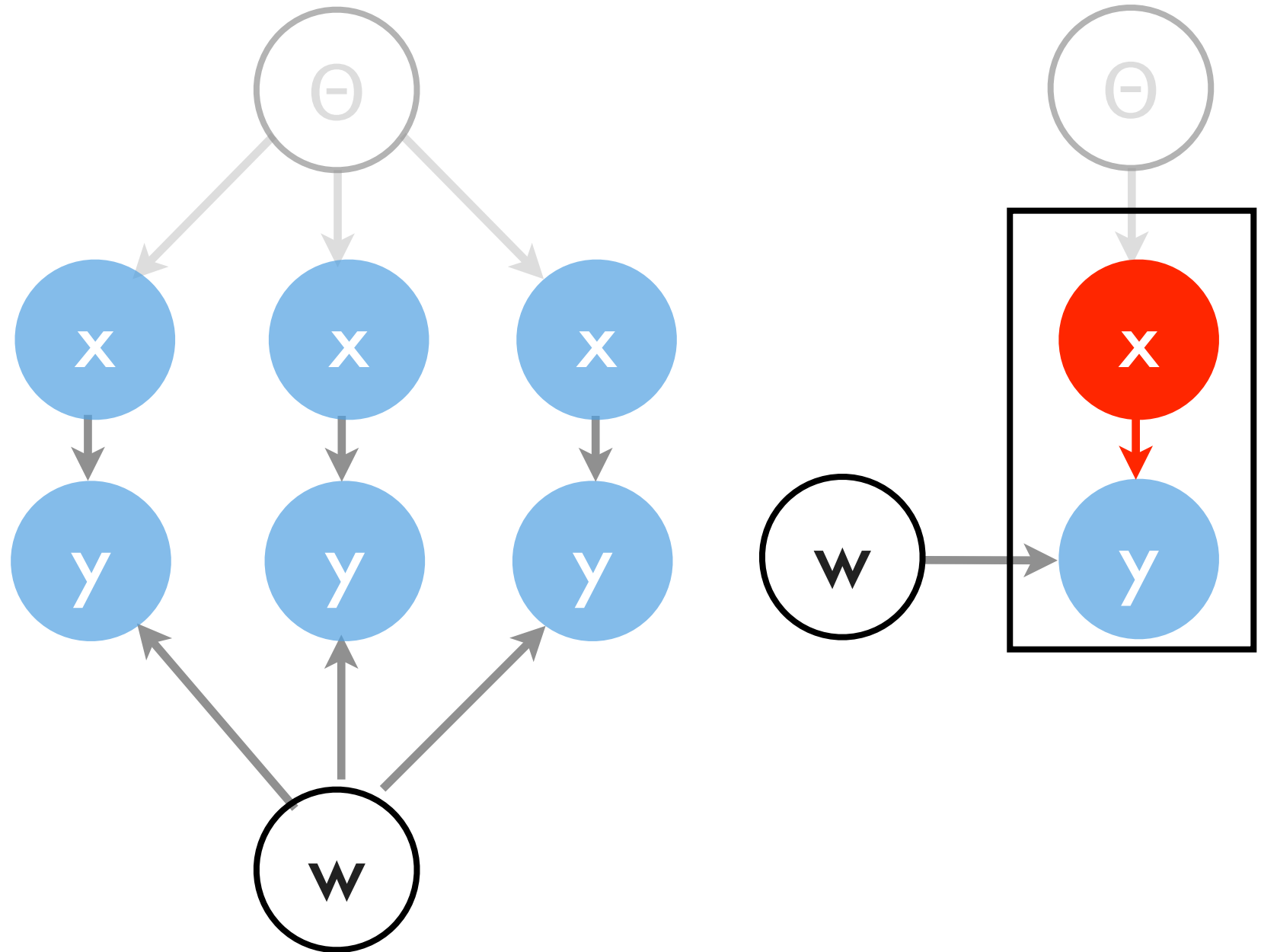
Factor Analysis



# 'Supervised' Models

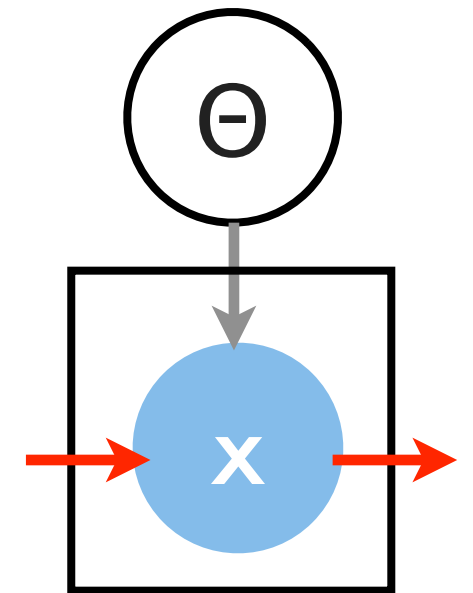
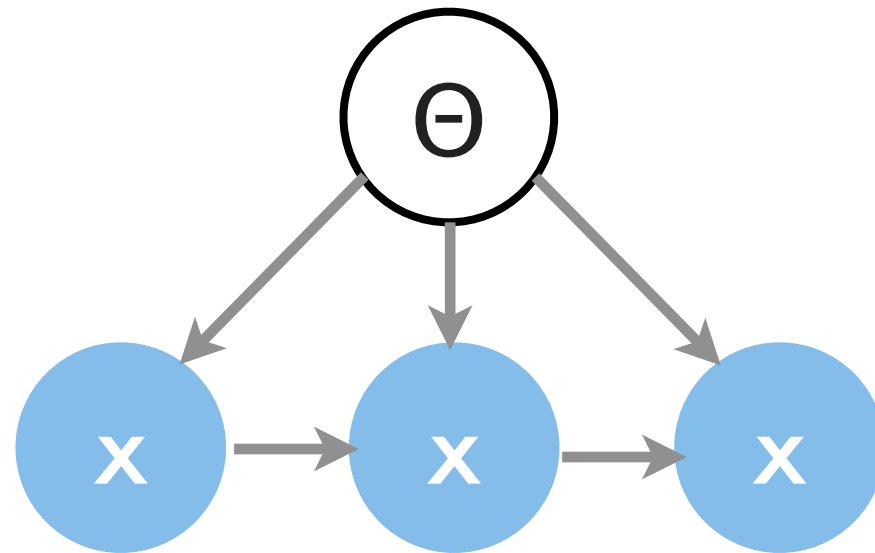
Classification  
Regression

spam filtering  
tiering  
crawling  
categorization  
bid estimation  
tagging



# Chains

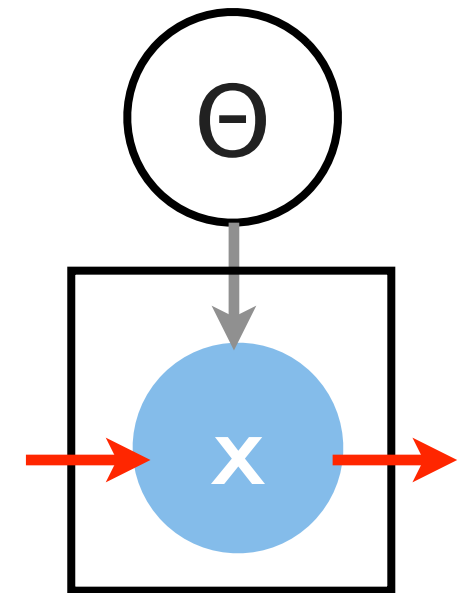
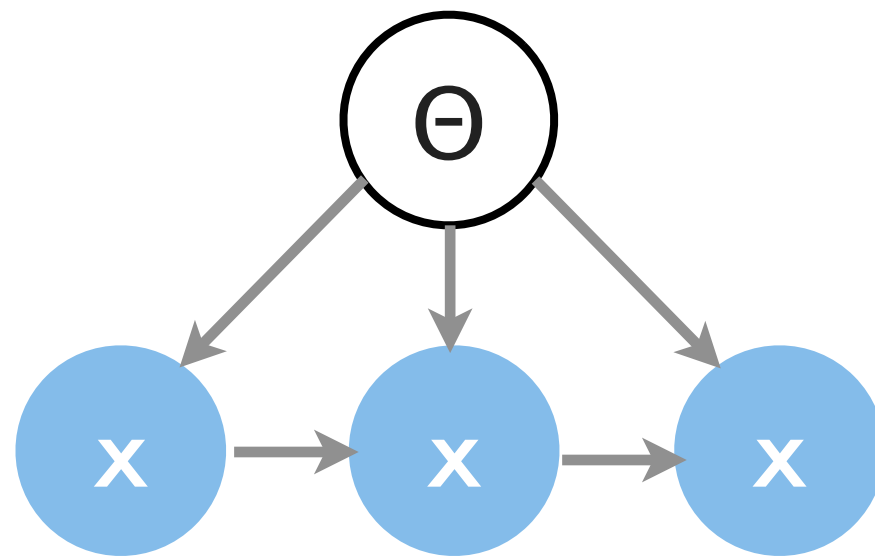
Markov  
Chain



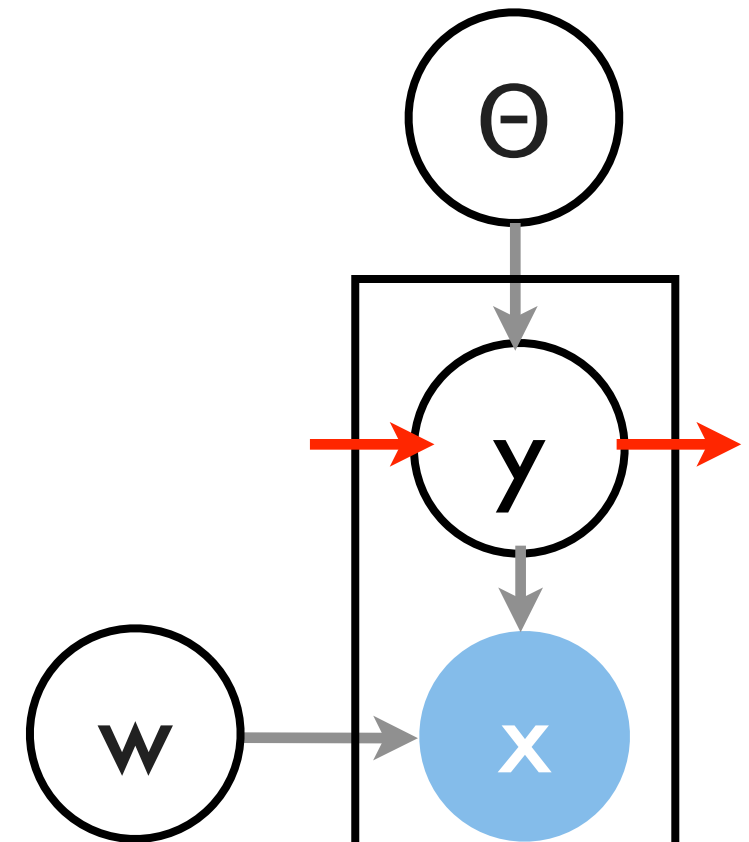
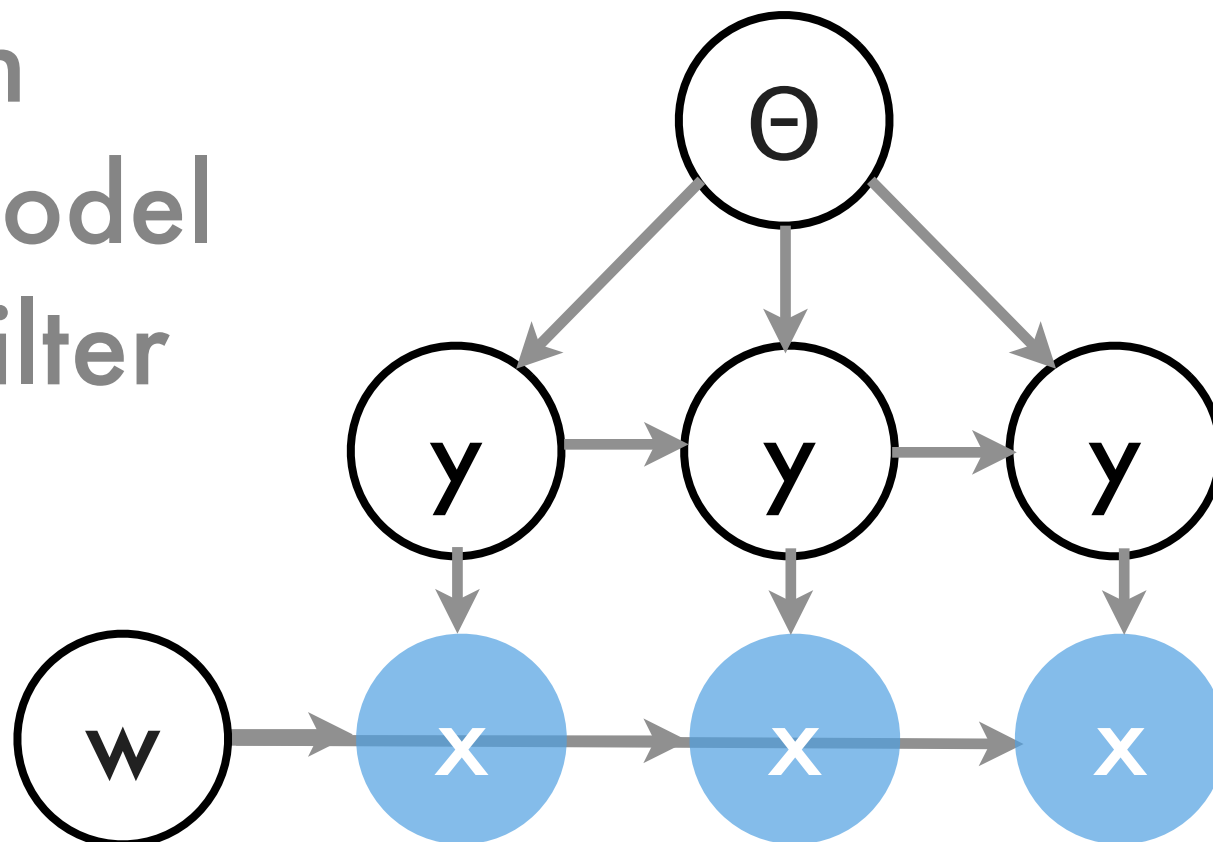


# Chains

Markov  
Chain



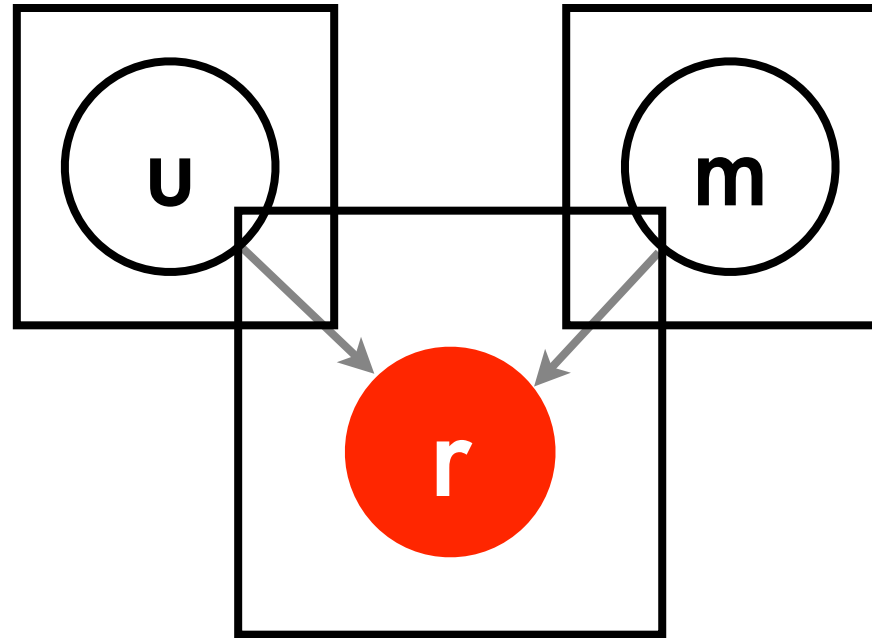
Hidden  
Markov Model  
Kalman Filter



# Collaborative Models

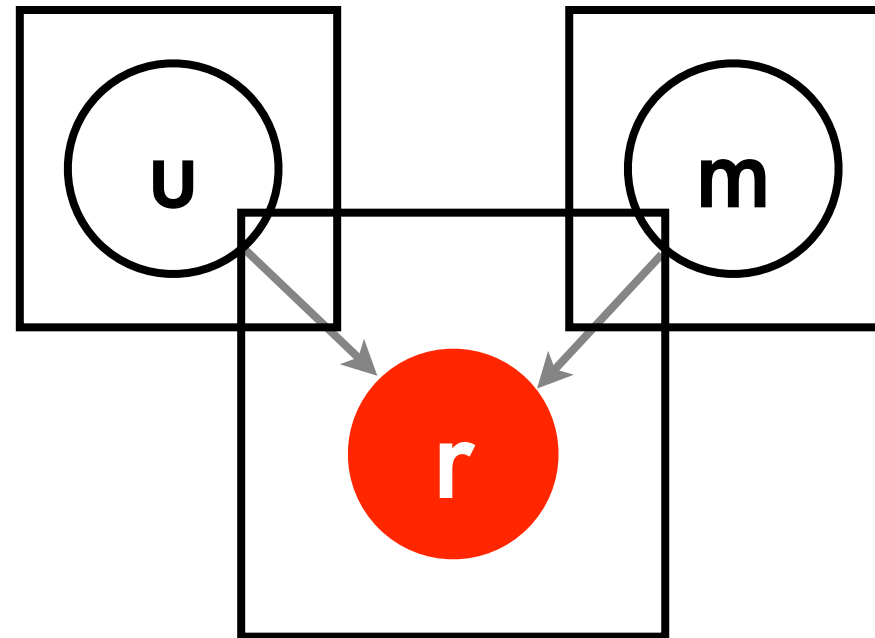
# Collaborative Models

Collaborative  
Filtering

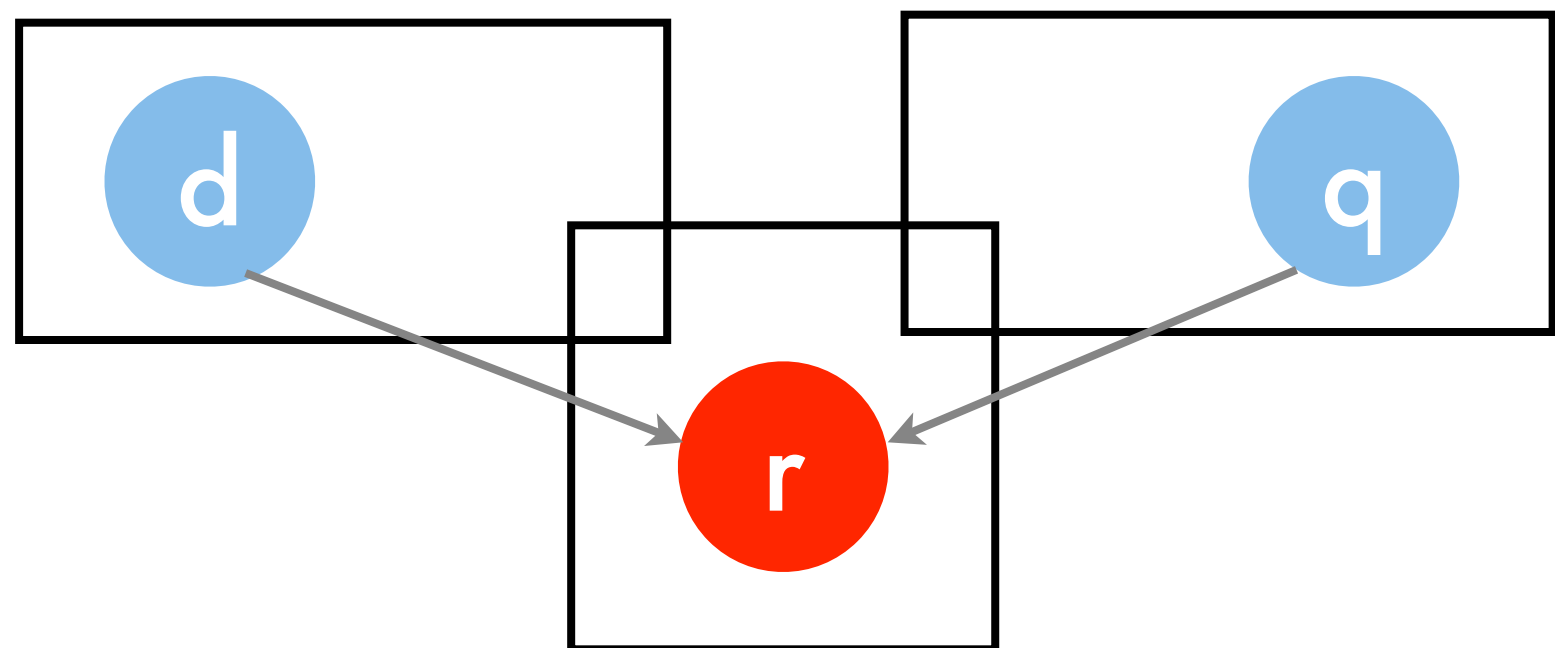


# Collaborative Models

Collaborative  
Filtering

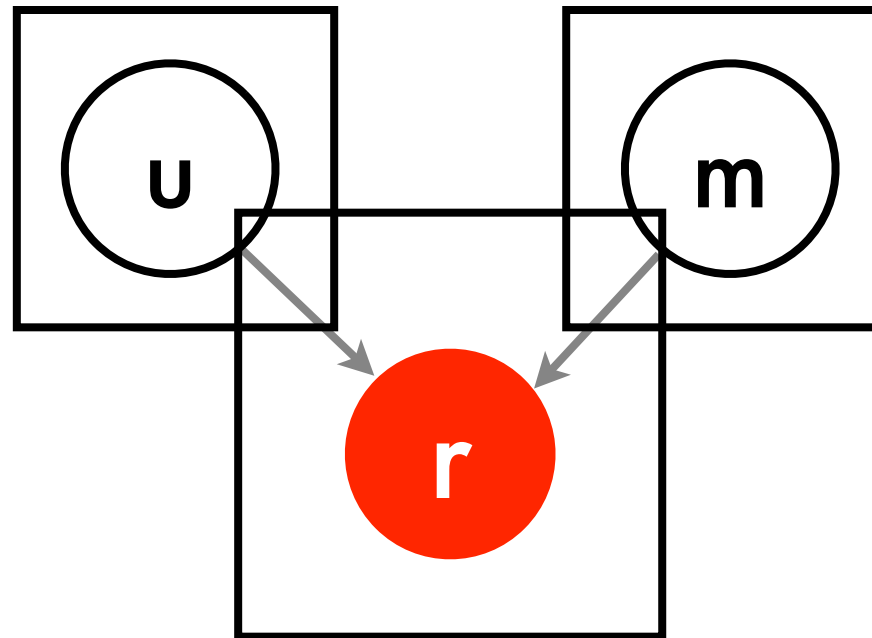


Current  
Webpage  
Ranking

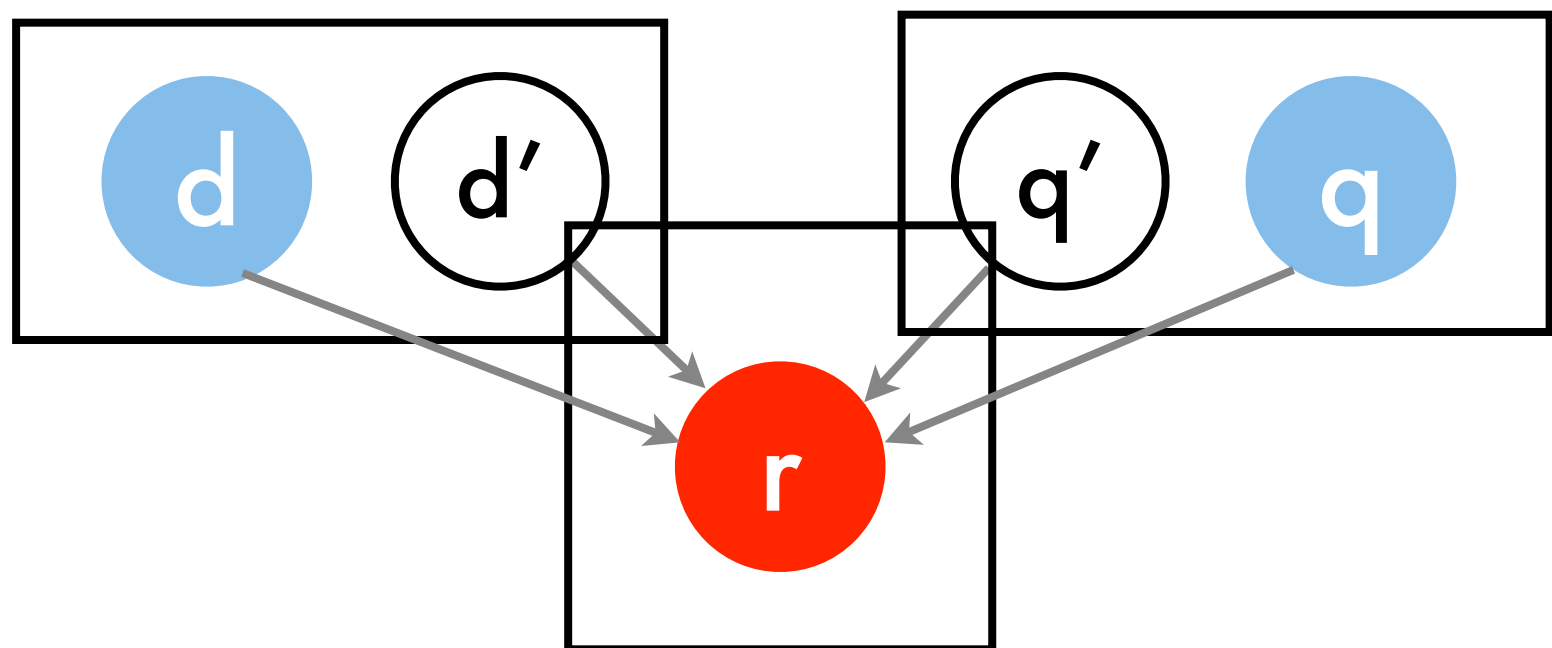


# Collaborative Models

Collaborative  
Filtering

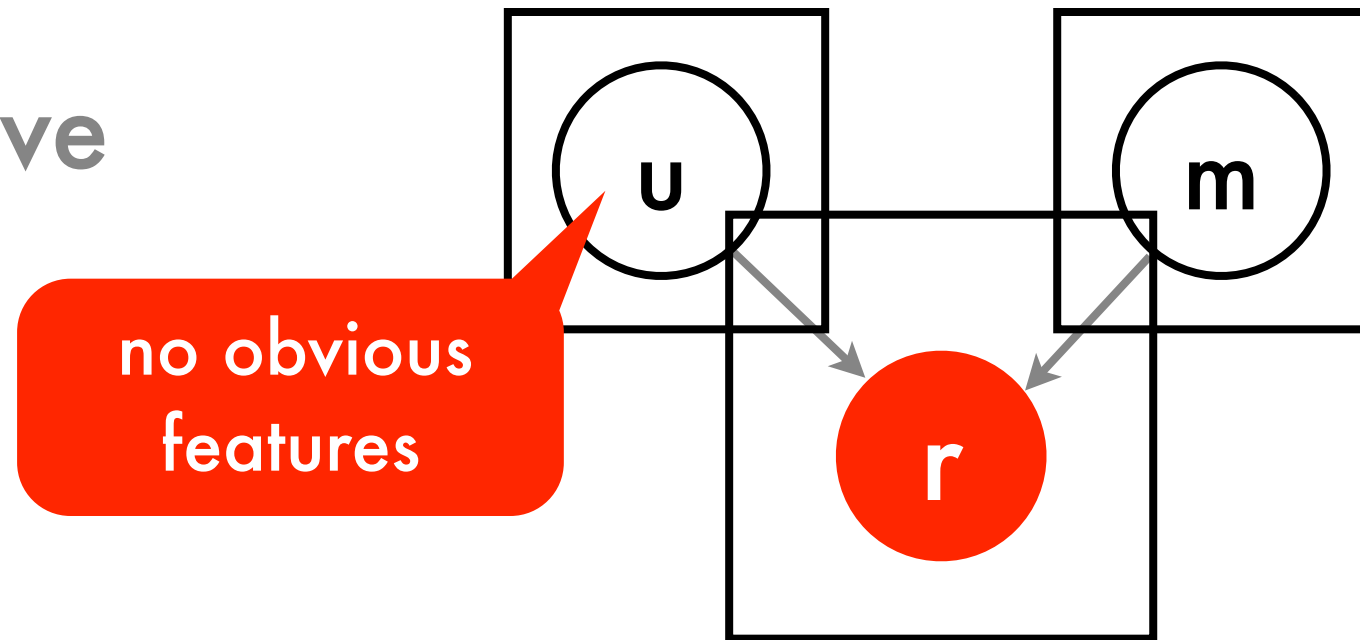


Webpage  
Ranking

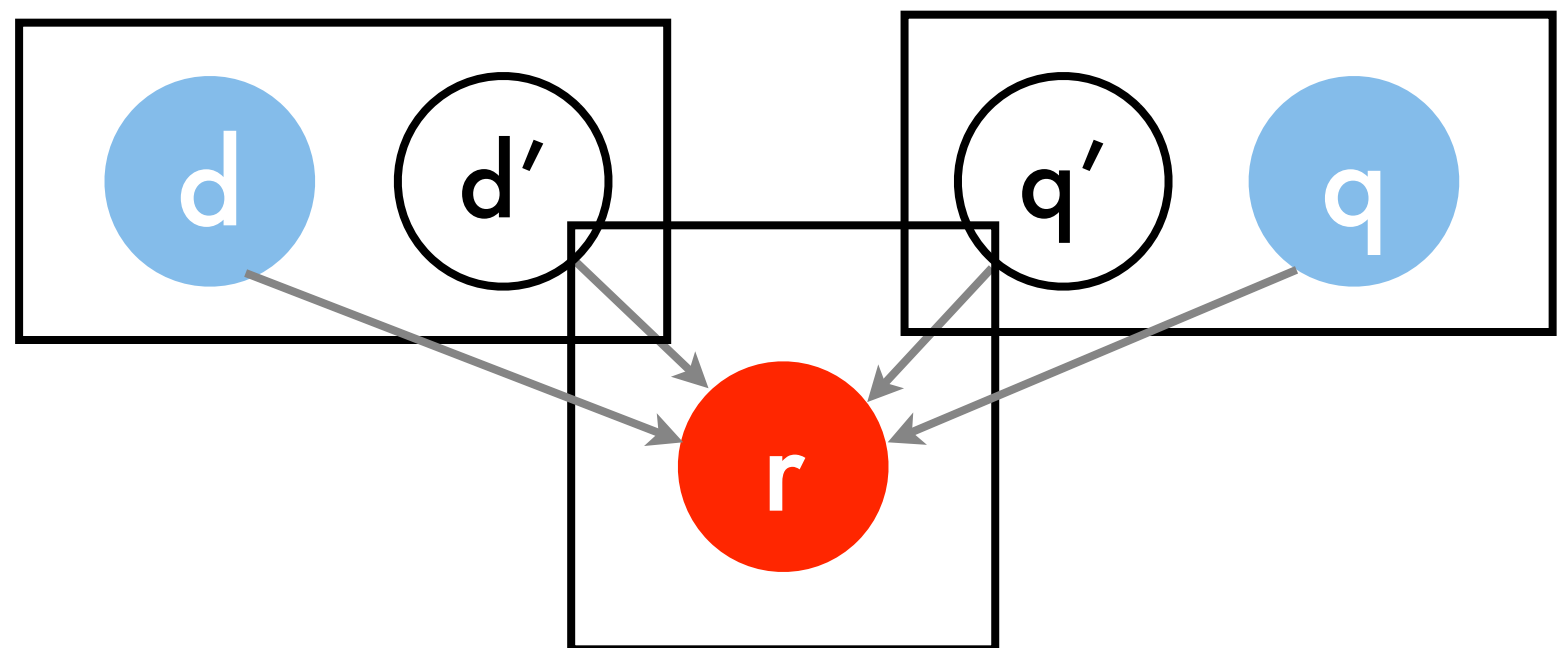


# Collaborative Models

Collaborative  
Filtering

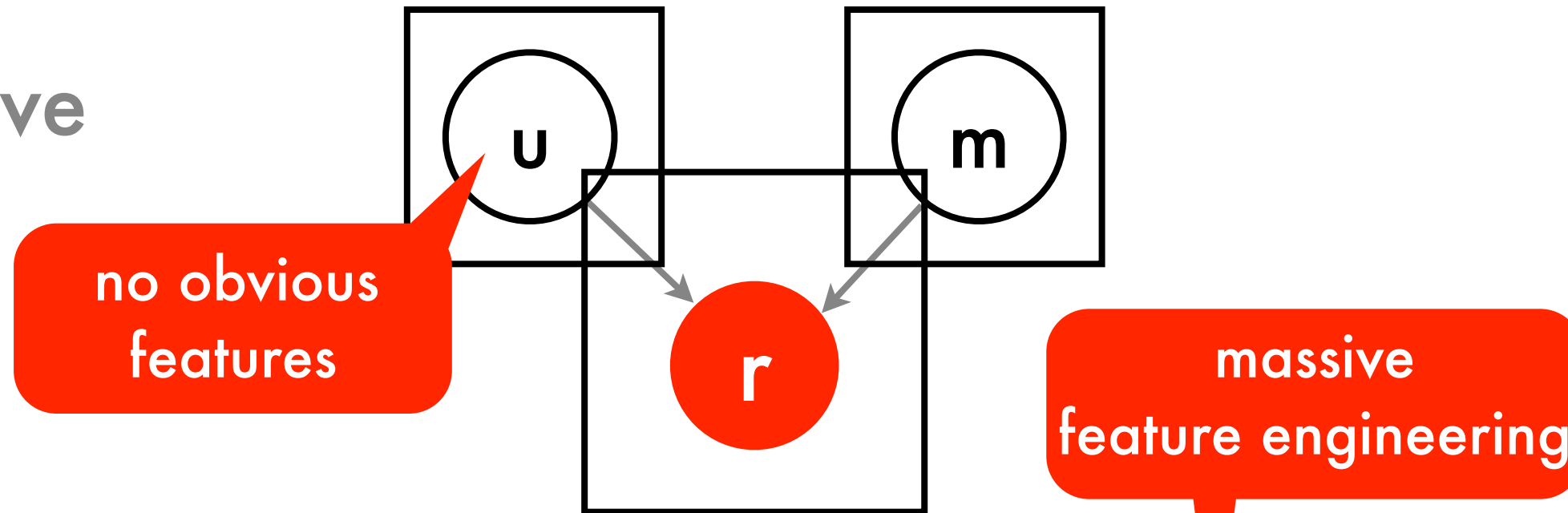


Webpage  
Ranking

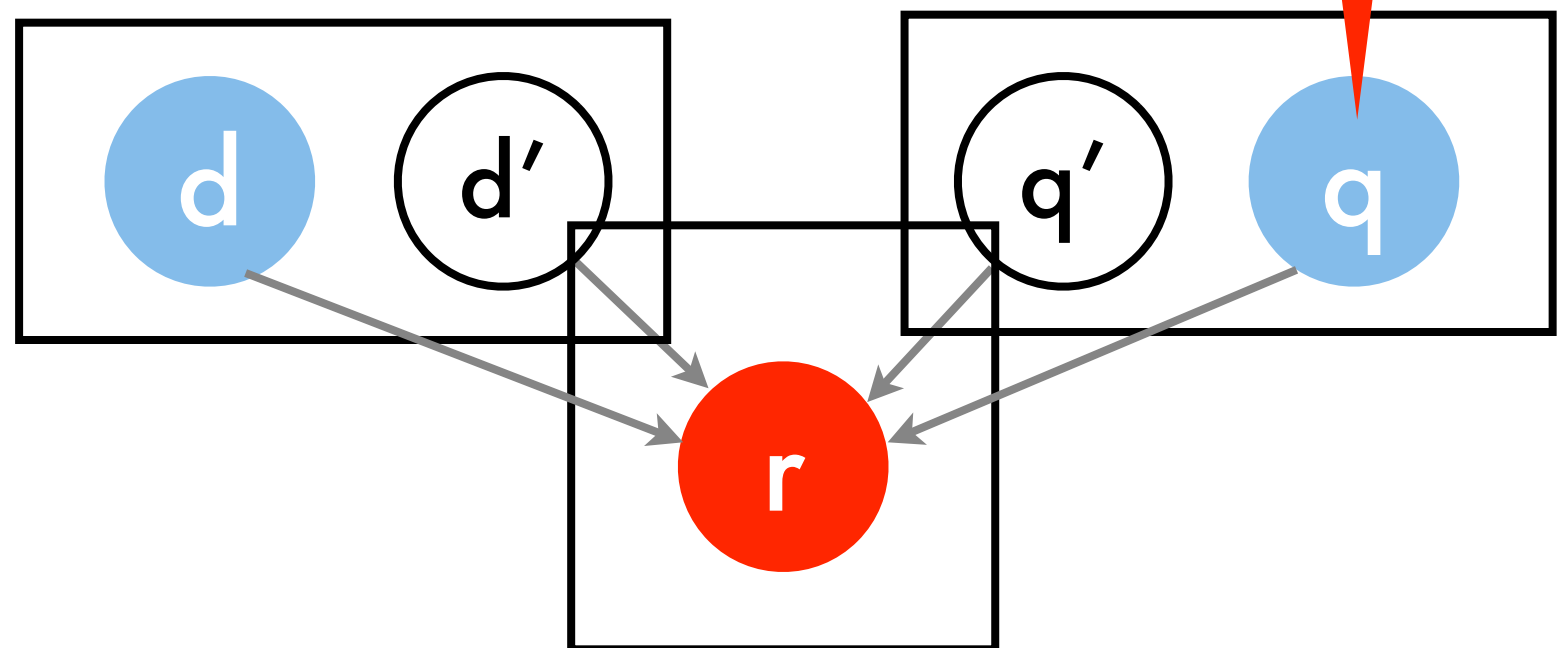


# Collaborative Models

Collaborative  
Filtering

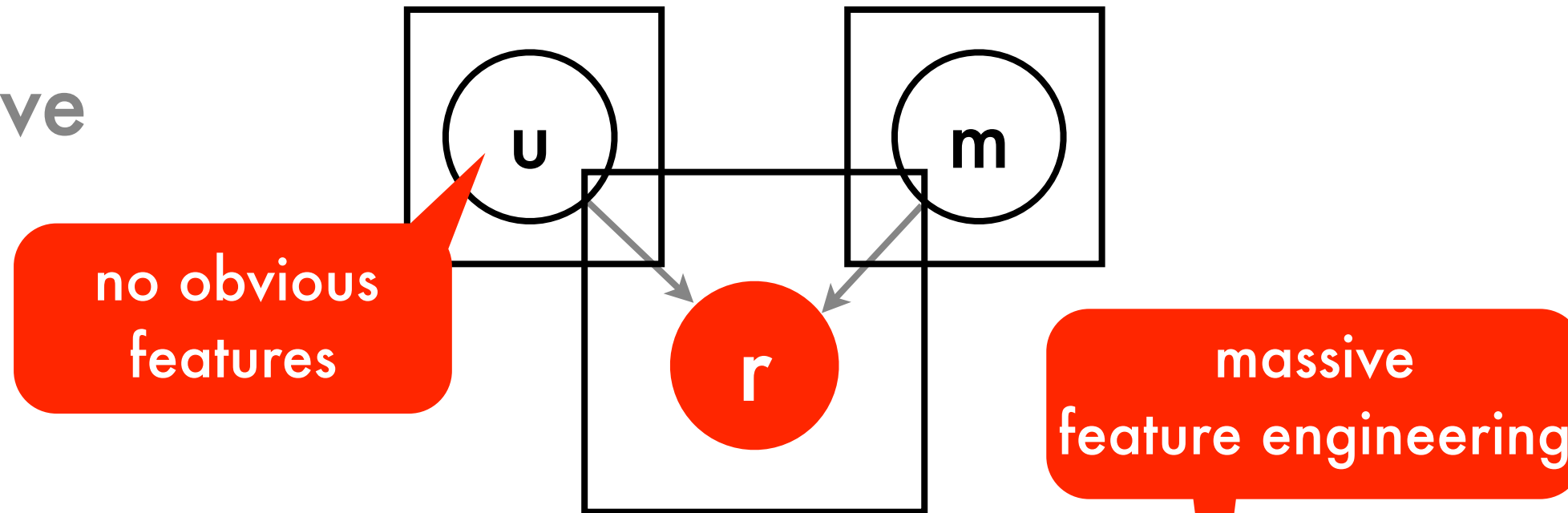


Webpage  
Ranking

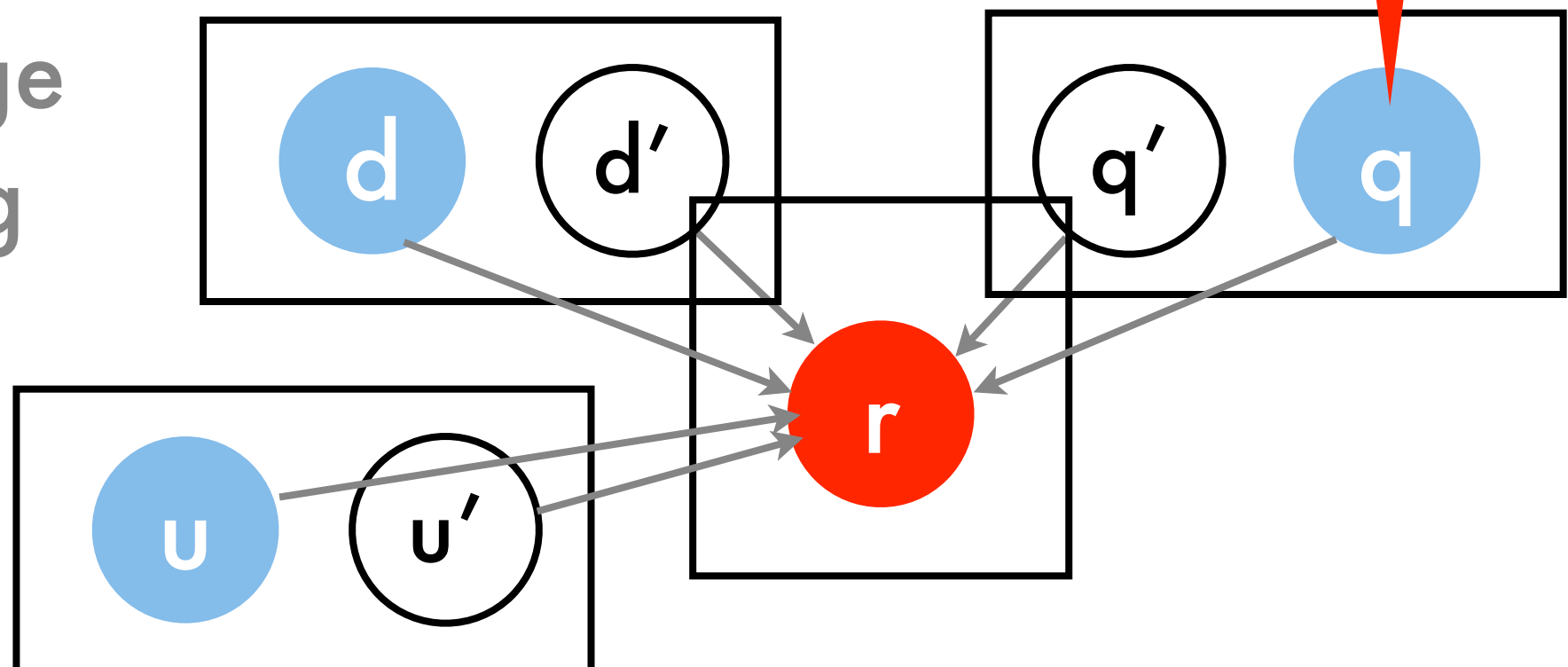


# Collaborative Models

Collaborative  
Filtering



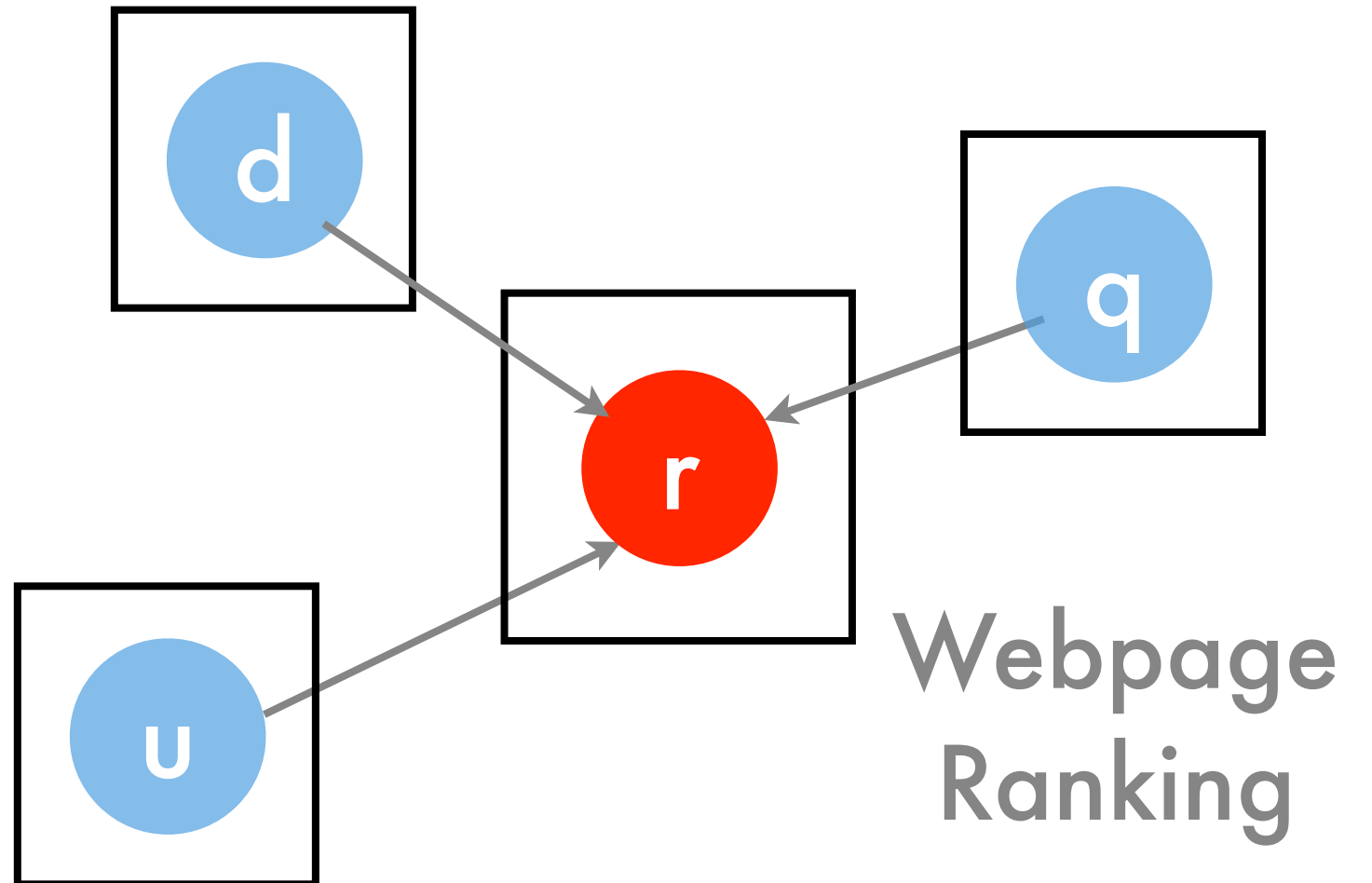
Webpage  
Ranking



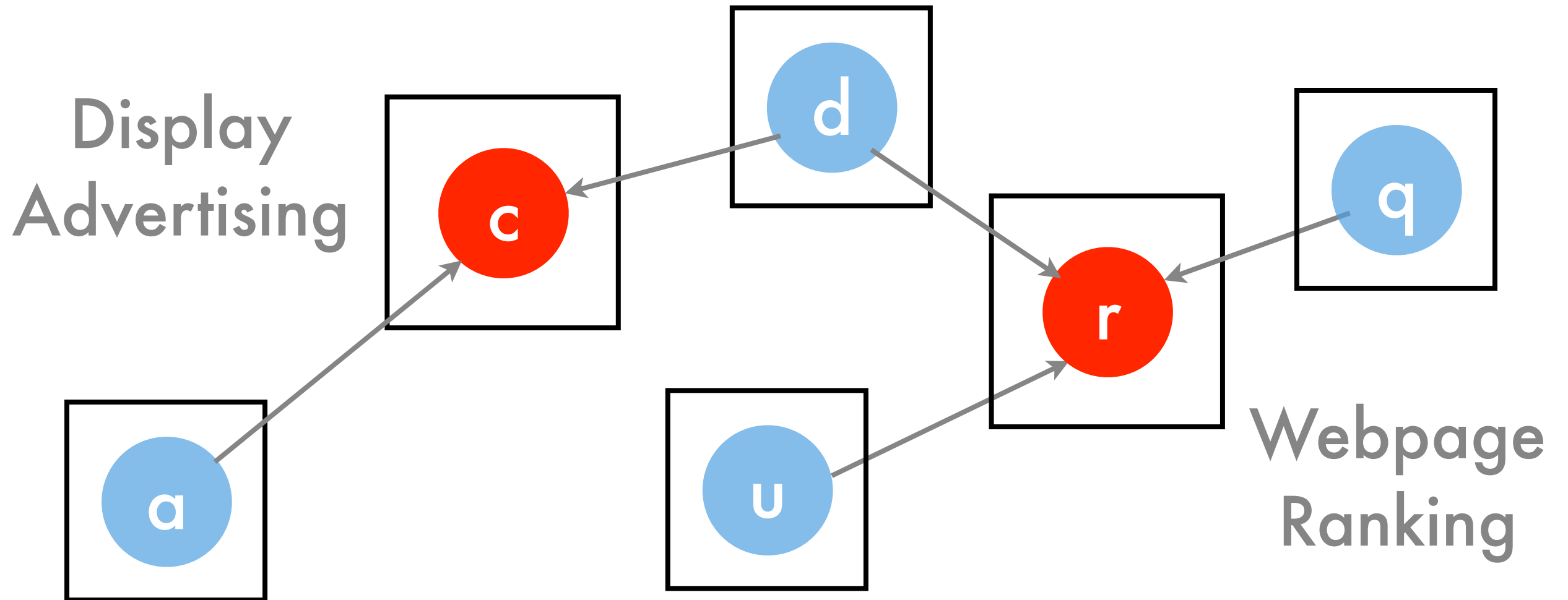
personalized



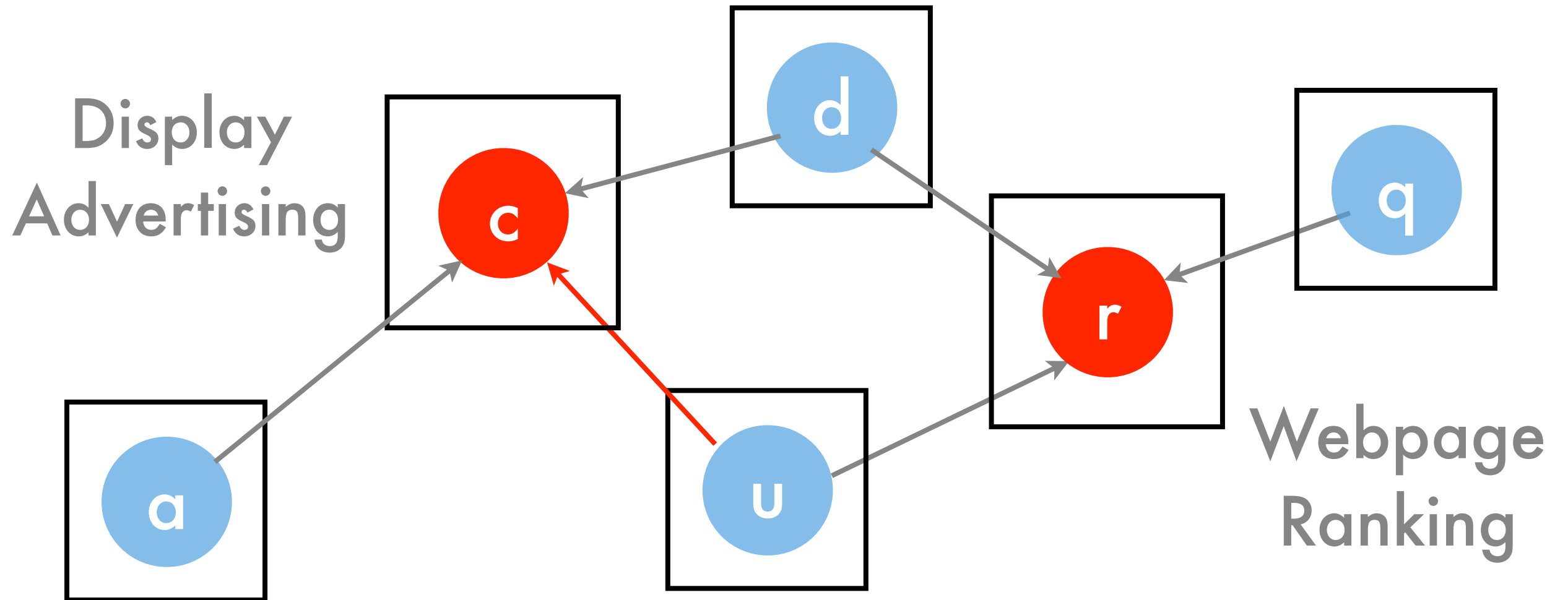
# Data Integration



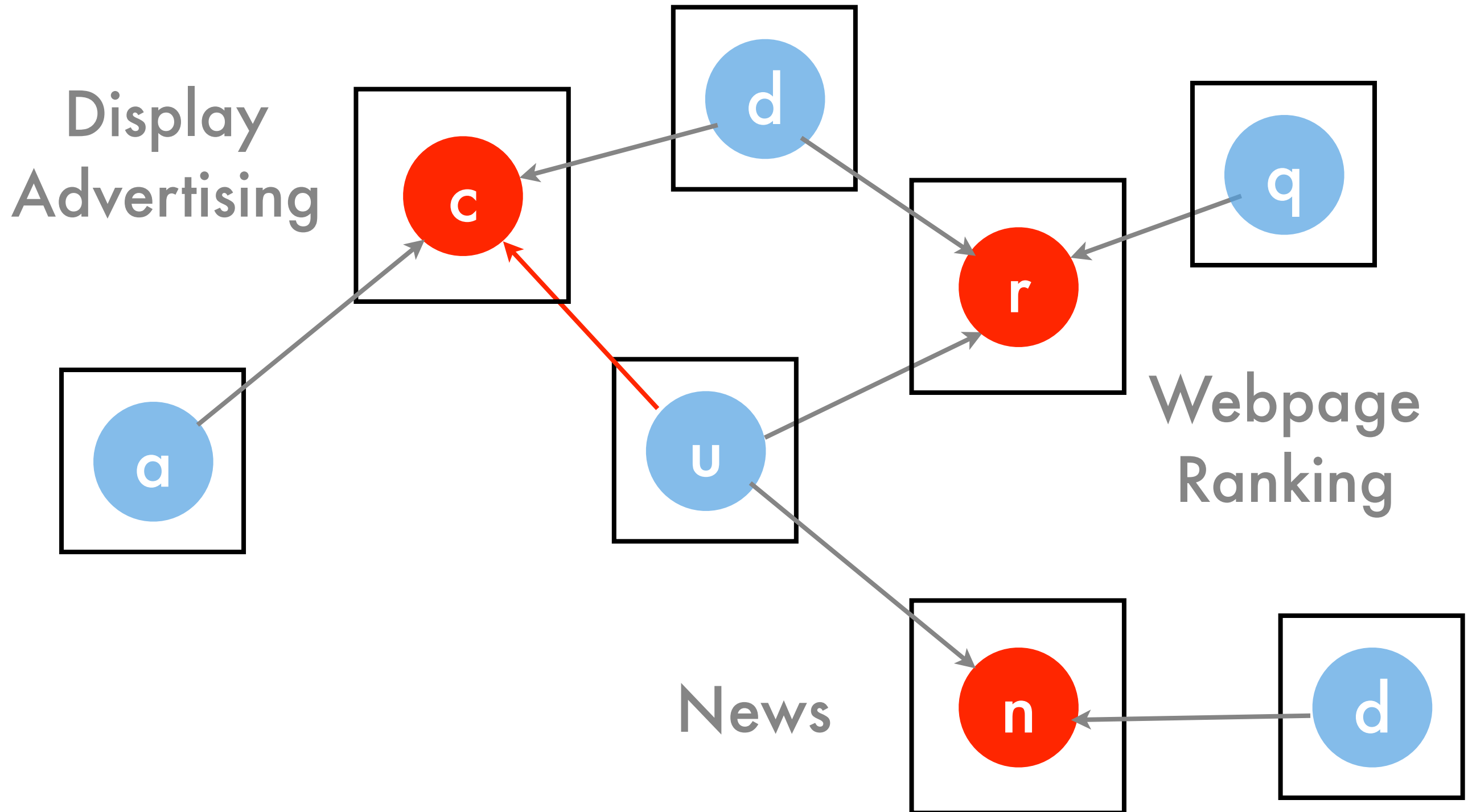
# Data Integration



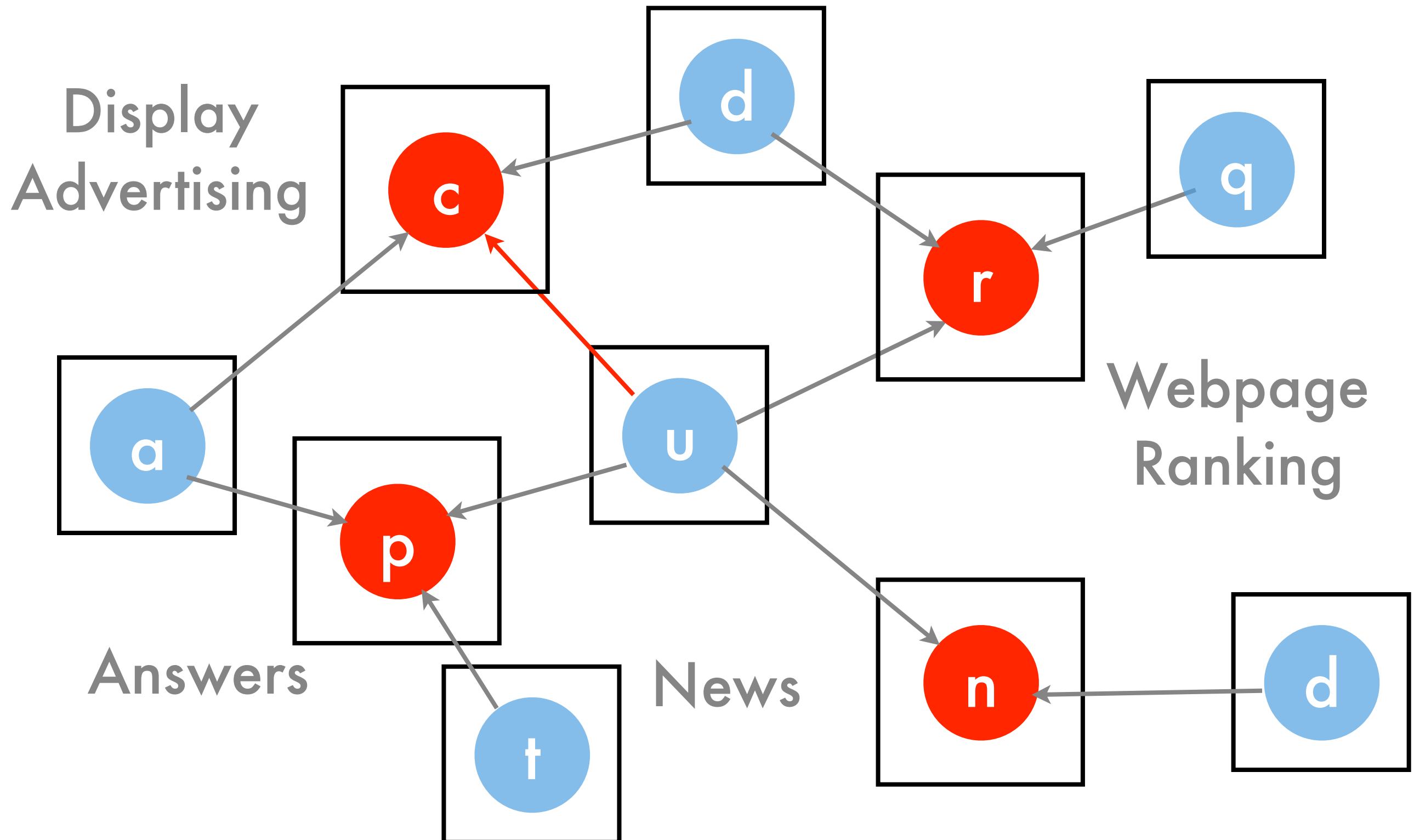
# Data Integration



# Data Integration



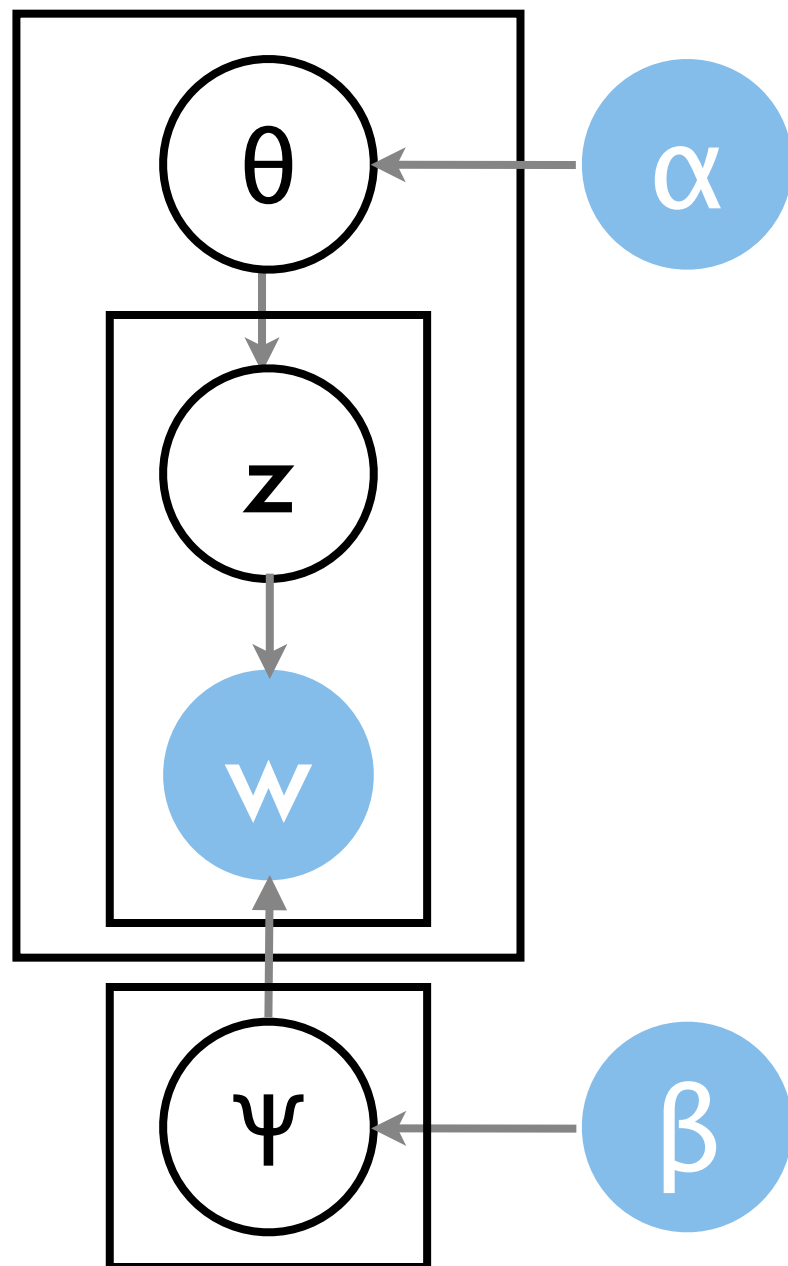
# Data Integration



# Topic Models

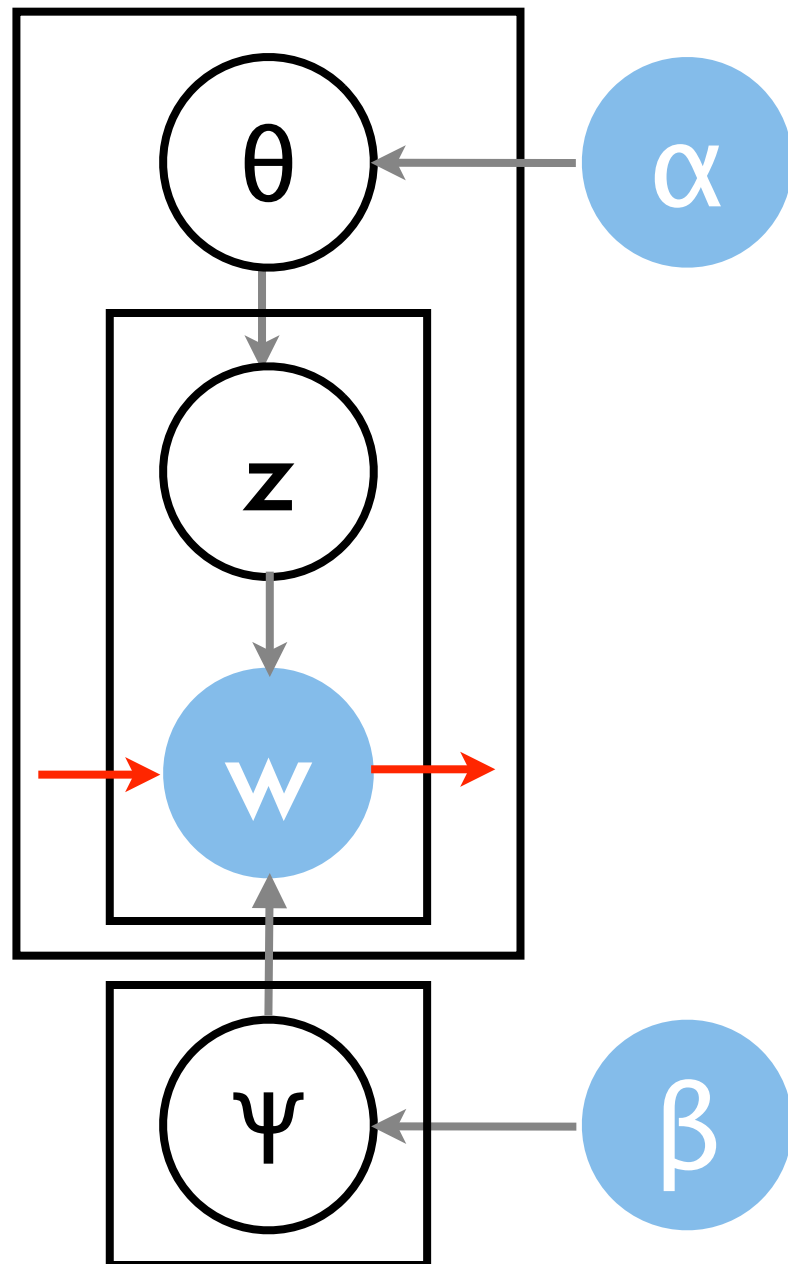
# Topic Models

Topic  
Models



# Topic Models

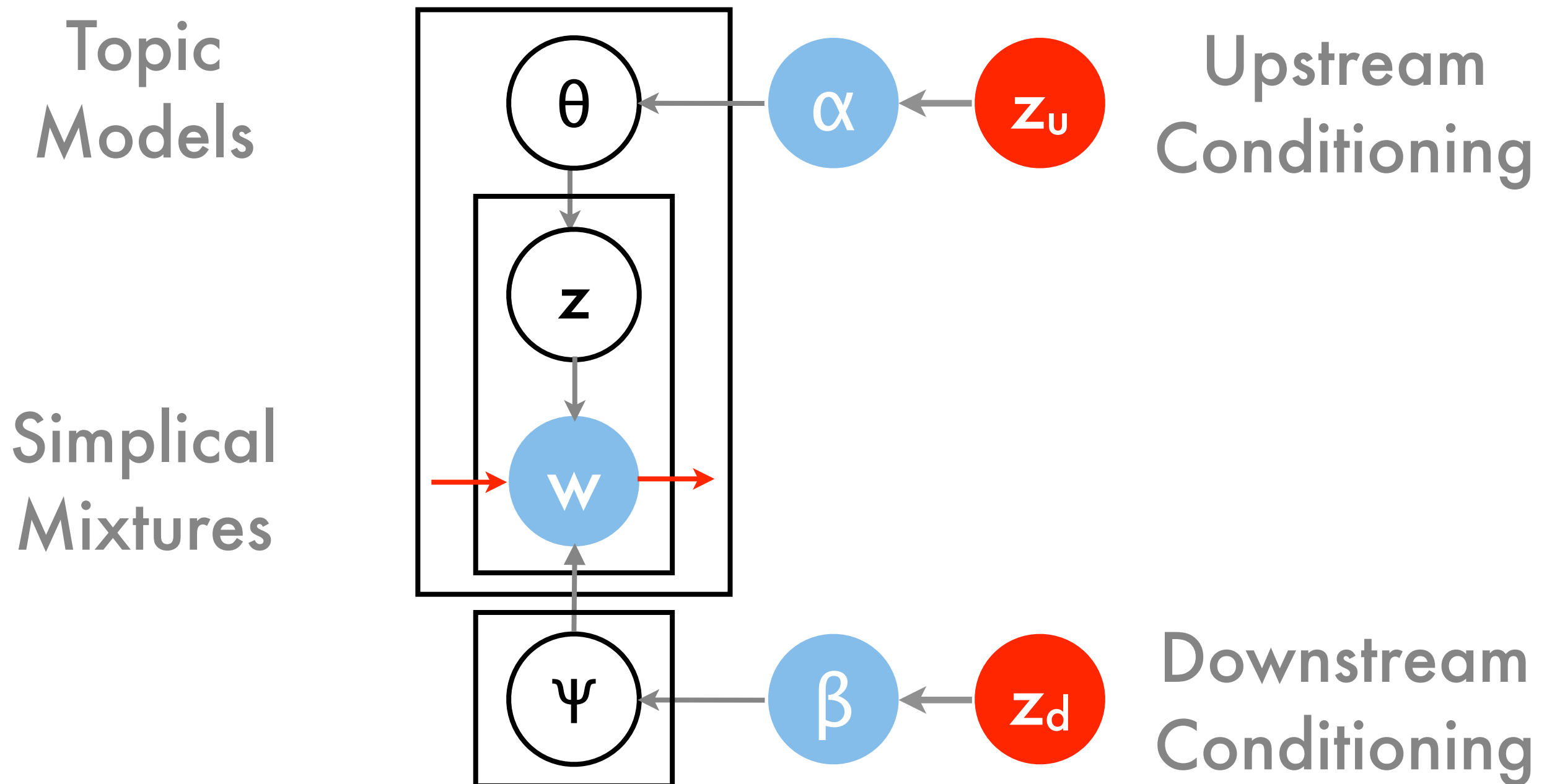
Topic Models



Simplicial Mixtures



# Topic Models



# Part 5 - Scalable Topic Models

# Topic models

# Grouping objects

# Grouping objects

SINGAPORE AIRLINES

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

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

Book a Flight | Check In

Round Trip  One Way

From:



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

Search  in  GO

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

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

Singapore

CHIJMES  
restaurants • bars • shops

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

Chijmes, a premier lifestyle destination in Singapore

Owned by: Managed by: Property Manager:



Copyright © 2006 Chijmes. All rights reserved.

Feedback | Terms & Conditions



Flame Arrival Ceremony at NUS

WATCH THE VIDEO

Joint Evacuation Exercises

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

MORE DETAILS

STAFF

ALUMNI

VISITORS

YAHOO!



# Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's the United logo and navigation links for 'My profile', 'Worldwide sites', and 'Customer service'. Below that are dropdown menus for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information', along with a search bar. A blue banner highlights '#1 ON TIME' and 'United. #1 in on-time arrivals. Details'. The main content area is divided into several sections: a flight booking form with fields for 'From', 'To', 'Departing', and 'Returning'; a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'; a promotional banner for 'Use 30% fewer miles on your next United flight.' featuring a large orange percentage sign; and a 'United news and deals' section with links to various travel-related news items. At the bottom, there are links for 'Cars', 'Hotels', and 'Vacations'.

The screenshot shows the website for The Australian National University (ANU). At the top, there's a search bar and navigation links for 'Change Location', 'Search', 'Before You Fly', 'Loyalty Programmes', and 'Promotions'. Below that are links for 'CALENDAR', 'SITEMAP', 'CONTACT', and 'e-CARDS'. A search bar for 'Search ANU...' is visible. The main banner features the text 'The Australian National University' in a large, light blue font. Below the banner are navigation links for 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The background of the banner shows a close-up of a tree trunk with a small plant growing from it.

This footer section lists the ownership and management of the website. It includes the following information:  
Owned by: SUNTEC  
Managed by: ARA  
Property Manager: APC  
Below the logos, there is a small image of a tree trunk with a small plant growing from it.



# Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's a navigation bar with 'UNITED' logo and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below that, there are tabs for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', and 'Services & information'. A search bar is present. The main content area features a large promotional banner: 'Use 30% fewer miles on your next United flight.' with a large orange percentage sign. To the right, there's a 'Log in' section with fields for 'Mileage Plus # or email address' and 'Password'. Below the login section, there are links for 'Start with' (My Mileage Plus, My reservations) and 'Start earning miles today'. A 'Change Location' search box is also visible. At the bottom, there's a 'Need Help?' section with links for 'Book A Flight Guide', 'SIA Holidays', and 'Hotel Bookings'.

The screenshot shows the Australian National University (ANU) website. The top navigation bar includes 'EXPLORE ANU', 'A-Z INDEX', and a search bar. The ANU logo and name are prominently displayed. Below the header, there are navigation tabs: '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 four featured sections: 'Forests renew after Black Saturday fires', 'School of Music at Floriade', 'Undergraduate studies', and 'Higher Degree Research'. At the bottom, there are navigation buttons for 'PROSPECTIVE STUDENTS', 'CURRENT STUDENTS', 'STAFF', 'ALUMNI', and 'VISITORS'.

The screenshot shows the Chez Panisse website. The top navigation bar includes 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. The main content area features a large image of a restaurant interior. Below the image, there's 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', 'CONTACT', 'INFORMATION', and 'DIRECTIONS • MAILING LIST'.



Home | Wining & Dining | Contact | Sitemap | About Suntec REIT



Feedback | Terms & Conditions

YAHOO!



# Grouping objects

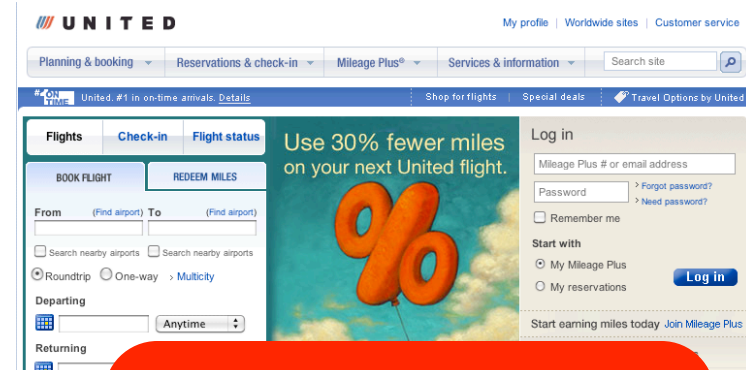
The image shows a screenshot of the United Airlines website. The page features a navigation bar with 'UNITED' and links for 'My profile', 'Worldwide sites', and 'Customer service'. Below the navigation bar are tabs for 'Planning & booking', 'Reservations & check-in', 'Mileage Plus', 'Services & information', and a search bar. The main content area includes a 'Flights' section with a 'BOOK FLIGHT' button, a 'REDEEM MILES' section, and a 'Log in' section. A large red speech bubble with the word 'airline' is overlaid on the page.

The image shows a screenshot of the Australian National University (ANU) website. The page features a navigation bar with 'EXPLORE ANU', 'A-Z INDEX', and a search bar. Below the navigation bar are tabs for 'HOME', 'FUTURE STUDENTS', 'CURRENT STUDENTS', 'RESEARCH & EDUCATION', 'ABOUT ANU', and 'STAFF'. The main content area includes a featured article titled 'Ash forests rise and rise again' and a section for 'Higher Degree Research'. A large red speech bubble with the word 'university' is overlaid on the page.

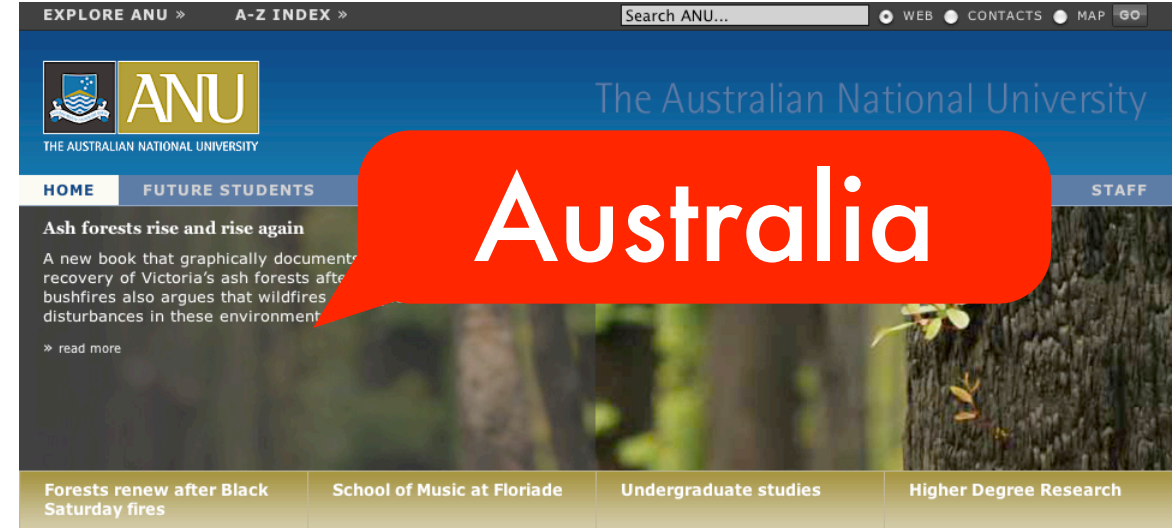
The image shows a screenshot of the Chez Panisse restaurant website. The page features a navigation bar with 'Home', 'Wining & Dining', 'Contact', 'Sitemap', and 'About Suntec REIT'. Below the navigation bar is a large image of the restaurant's exterior at night. A large red speech bubble with the word 'restaurant' is overlaid on the page.



# Grouping objects



USA



Australia



Singapore



YAHOO!

# Topic Models

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

Use 30% fewer miles on your next United flight.

BOOK FLIGHT | REDEEM MILES

From (Find airport) To (Find airport)

Roundtrip | One-way | Multicity

Departing: Anytime

Returning: Anytime

Log in: Mileage Plus # or email address, Password, Remember me, My Mileage Plus, My reservations

USA  
airline

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

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

HOME | FUTURE STUDENTS | CURRICULUM | ABOUT ANU

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

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

Australia  
university

SINGAPORE AIRLINES

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

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

Round Trip | One Way | Stopover/Multi-city

Departure City: Singapore

Destination City: Bangkok

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

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

Singapore  
airline

NUS National University of Singapore

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

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

A Leading Global University

Game Arrival Ceremony

Joint Evacuation Exercises

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

USA  
food

PROSPECTIVE STUDENTS | CURRENT STUDENTS | STAFF | ALUMNI | VISITORS

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

Chijmes  
restaurants • bars • shops

Discover a century of resplendent living history behind the cloisters

Chijmes, a premier lifestyle destination in Singapore

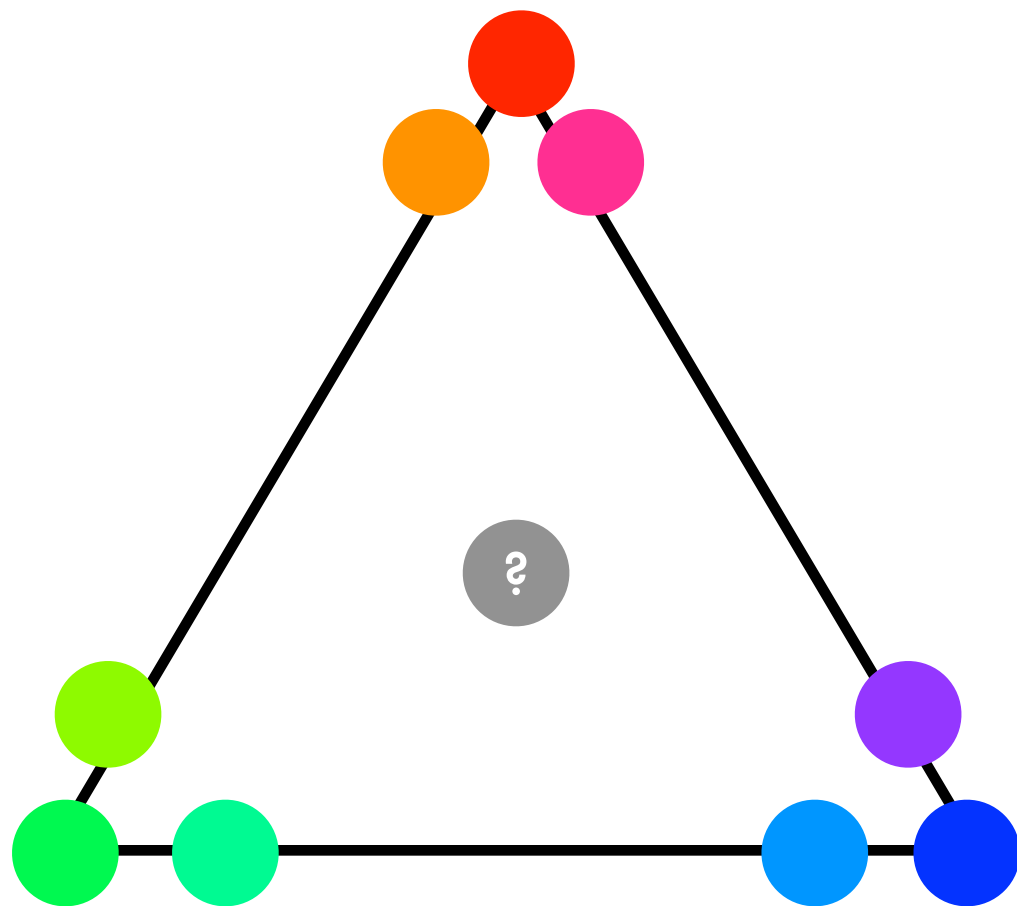
Owned by: SUNTEC | Managed by: ARA | Property Manager: APC

Singapore  
food



# Clustering & Topic Models

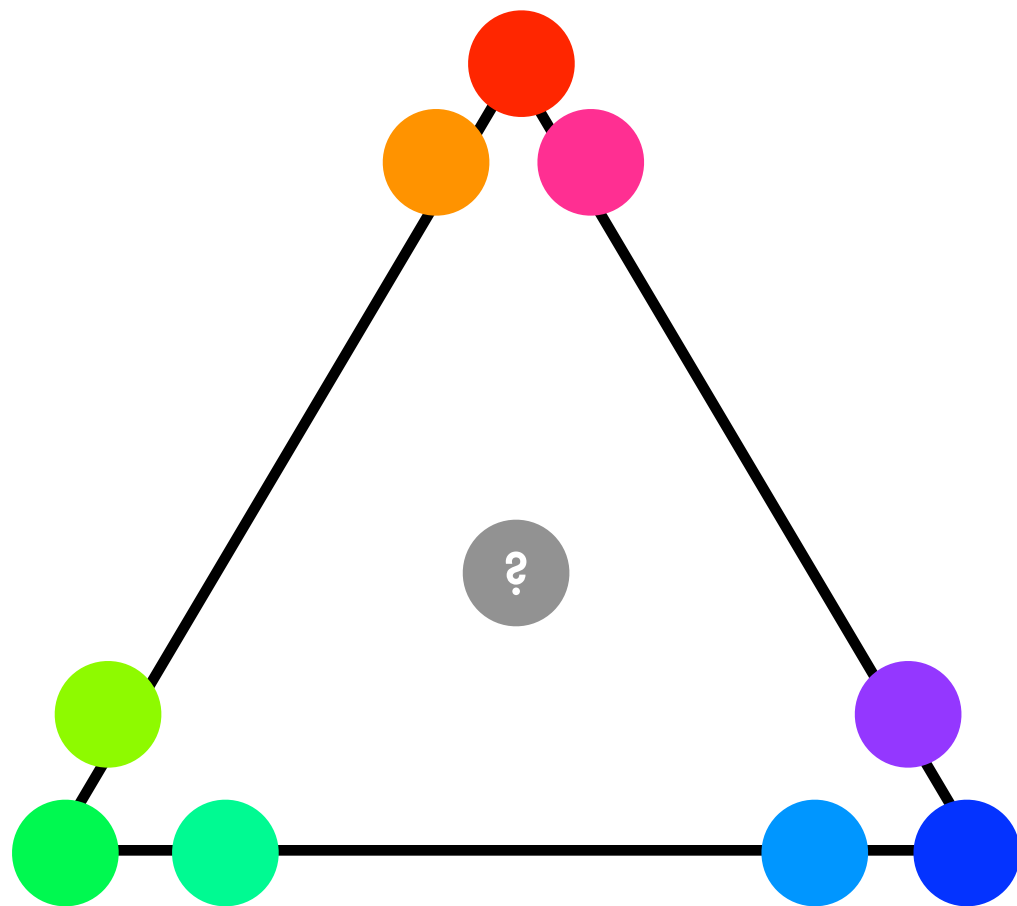
Clustering



group objects  
by prototypes

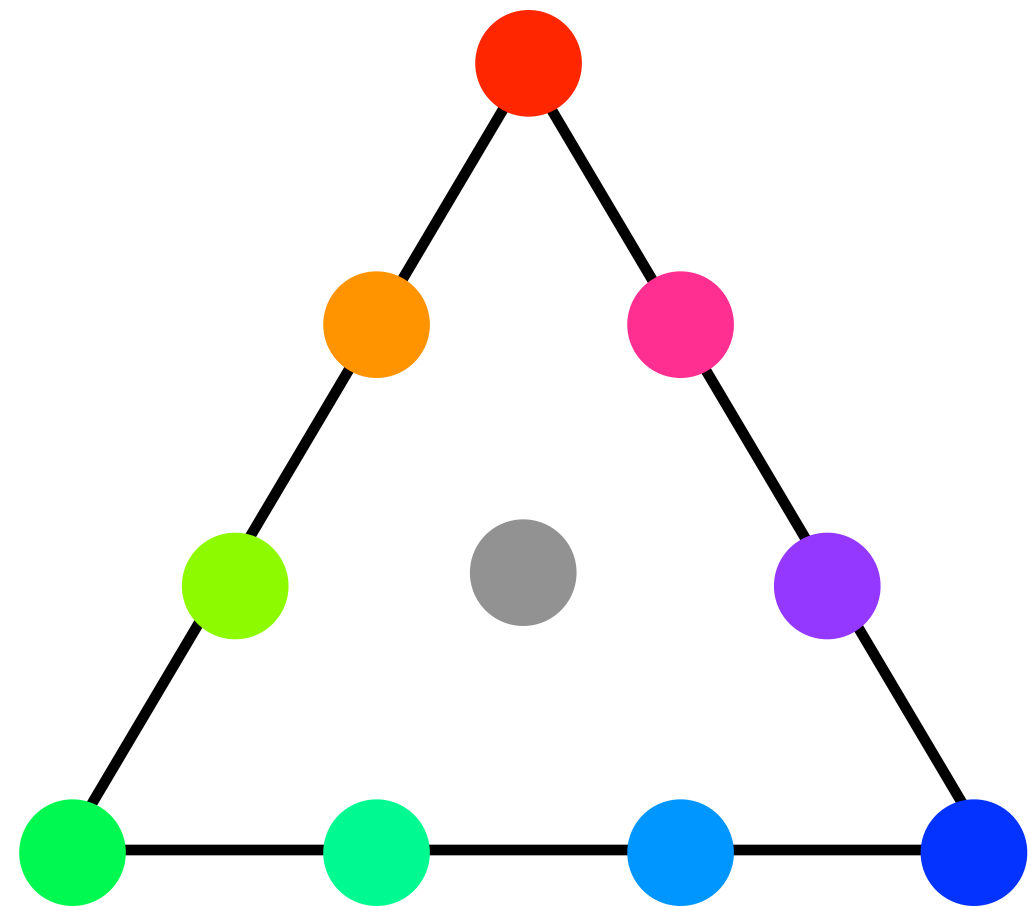
# Clustering & Topic Models

Clustering



group objects  
by prototypes

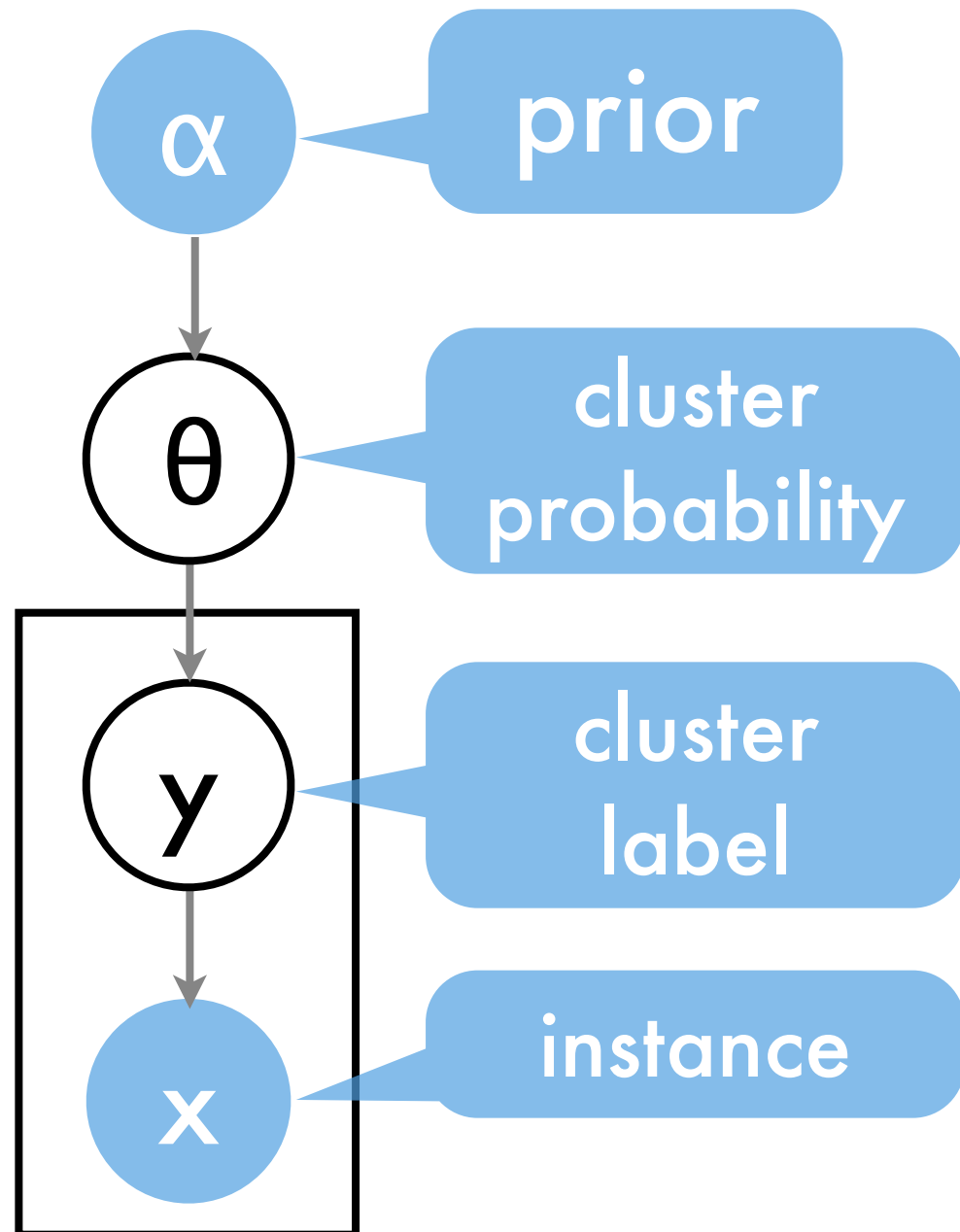
Topics



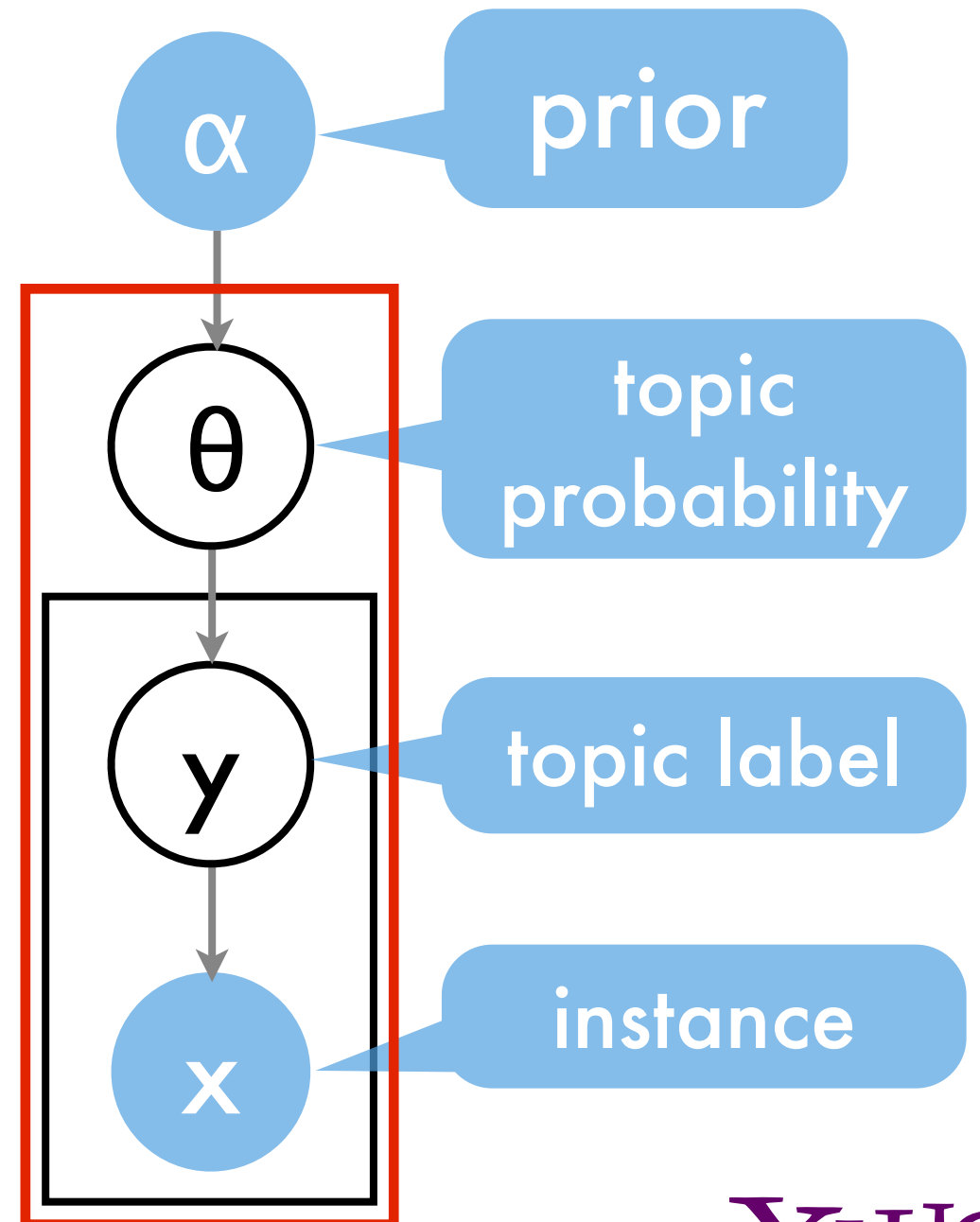
decompose objects  
into prototypes

# Clustering & Topic Models

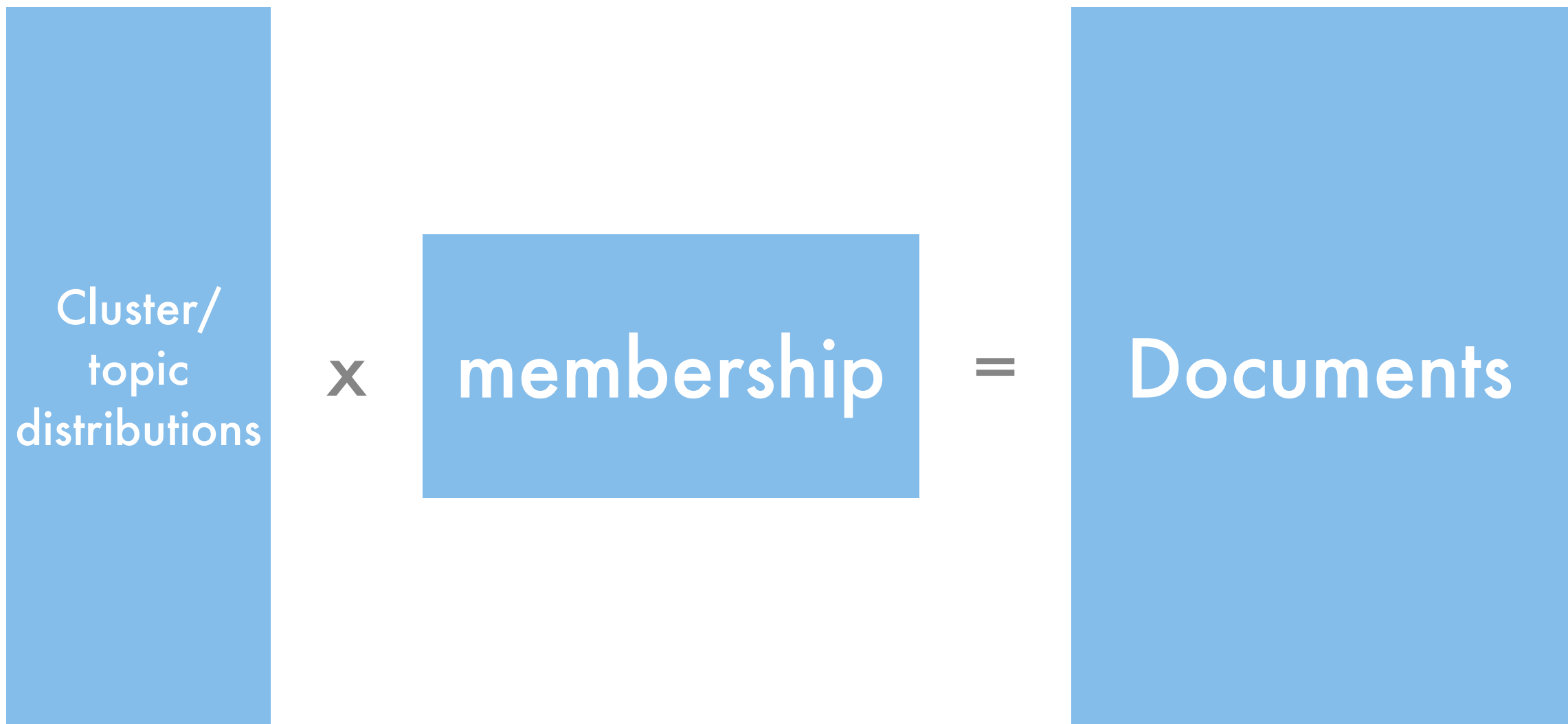
clustering



Latent Dirichlet Allocation



# Clustering & Topic Models



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

# Topics in text

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

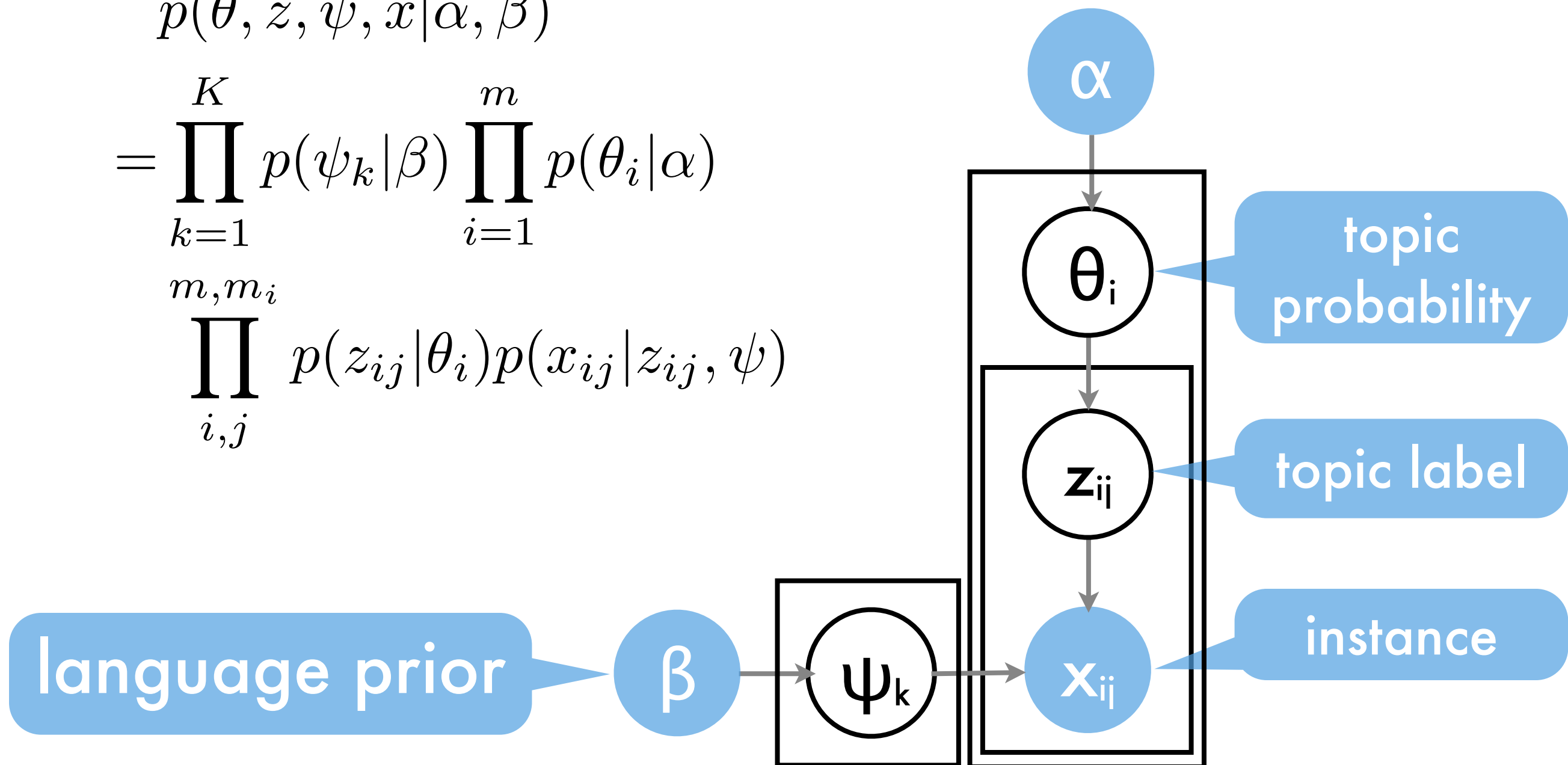
Latent Dirichlet Allocation; Blei, Ng, Jordan, JMLR 2003

# Collapsed Gibbs Sampler



# Joint Probability Distribution

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



# 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

$\beta$

$\Psi_k$

$\alpha$

$\theta_i$

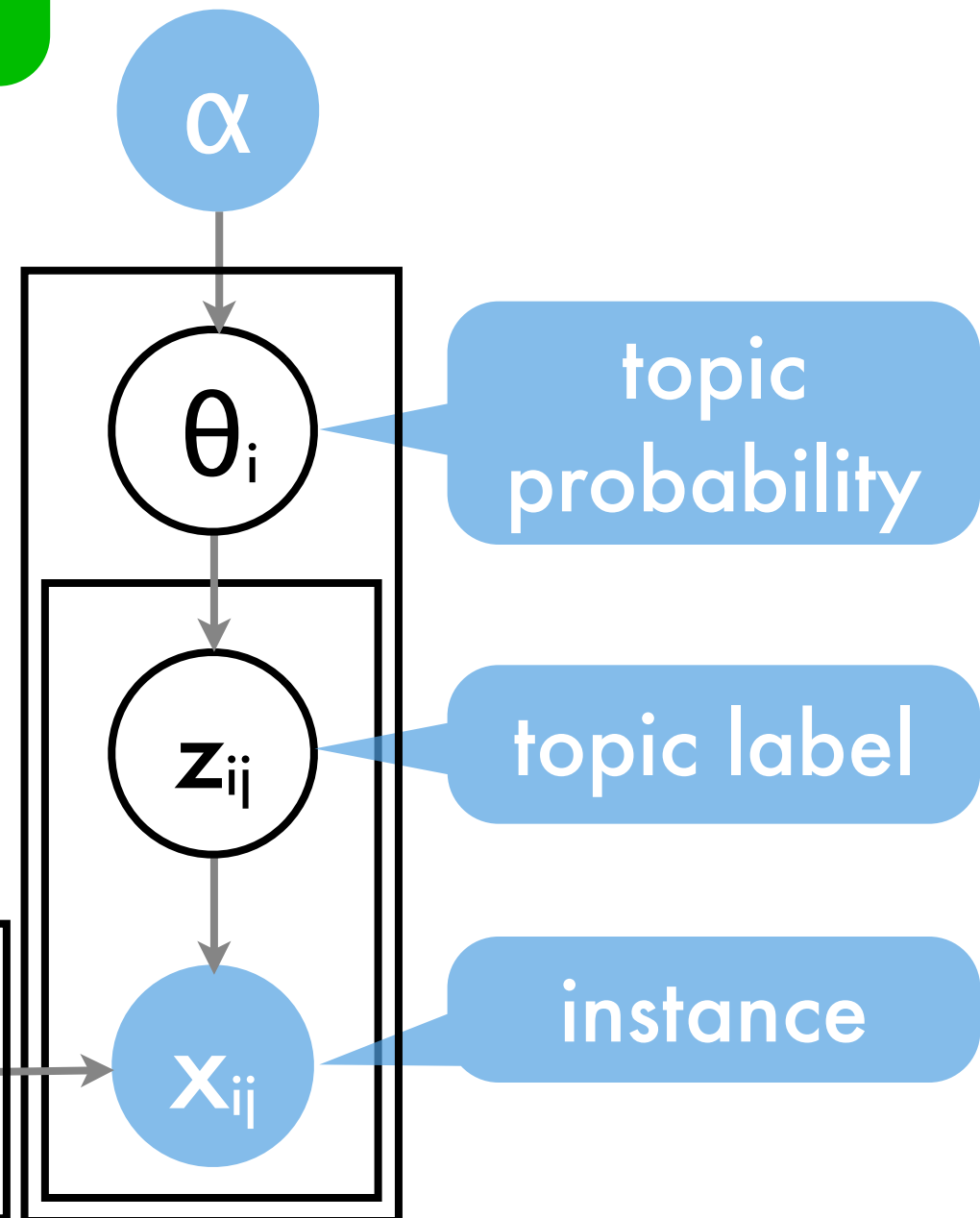
$z_{ij}$

$x_{ij}$

topic probability

topic label

instance



# Joint Probability Distribution

sample  $\Psi$   
independently

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

sample  $\theta$   
independently

sample  $z$   
independently

language prior

$\beta$

$\Psi_k$

$\alpha$

$\theta_i$

$z_{ij}$

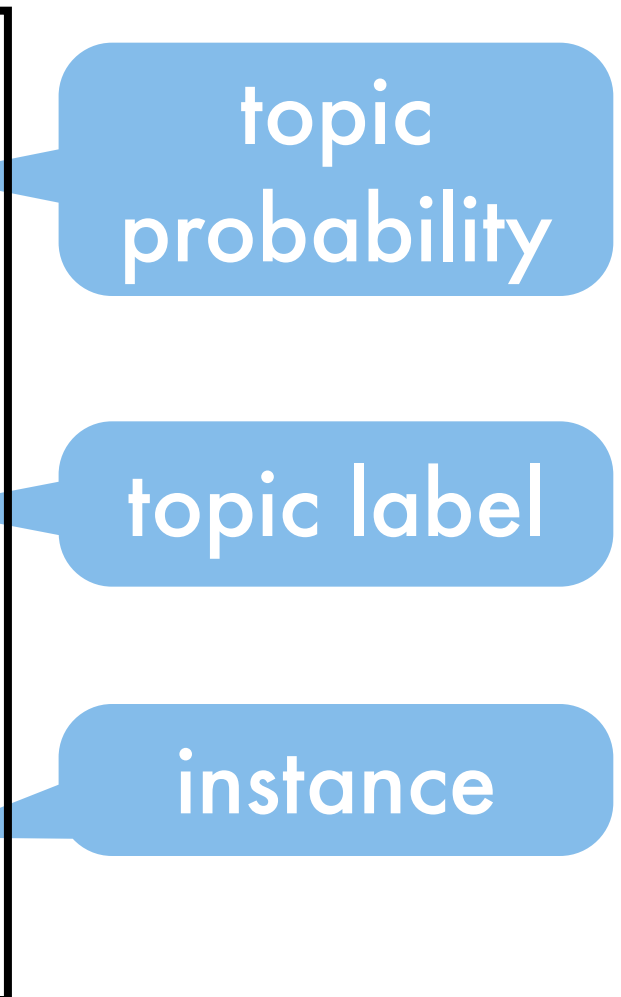
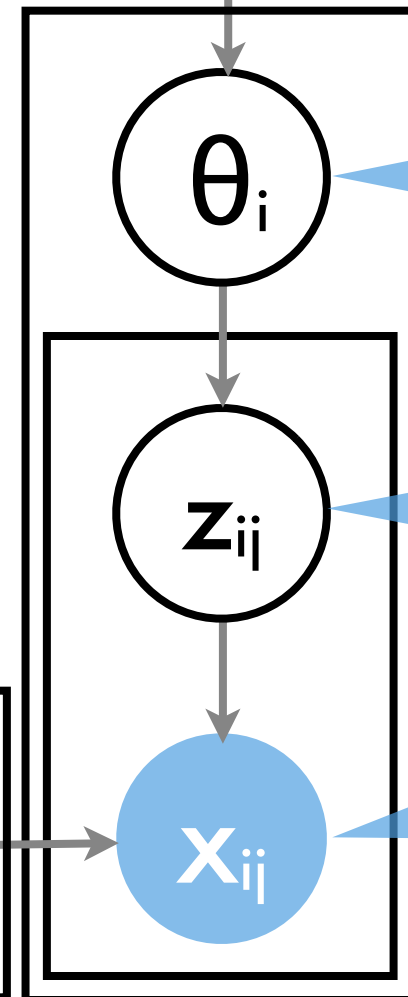
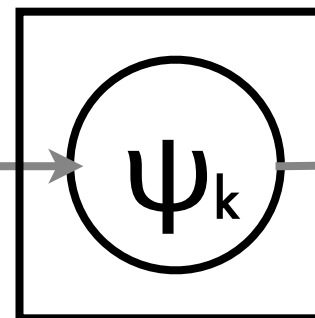
$x_{ij}$

slow

topic  
probability

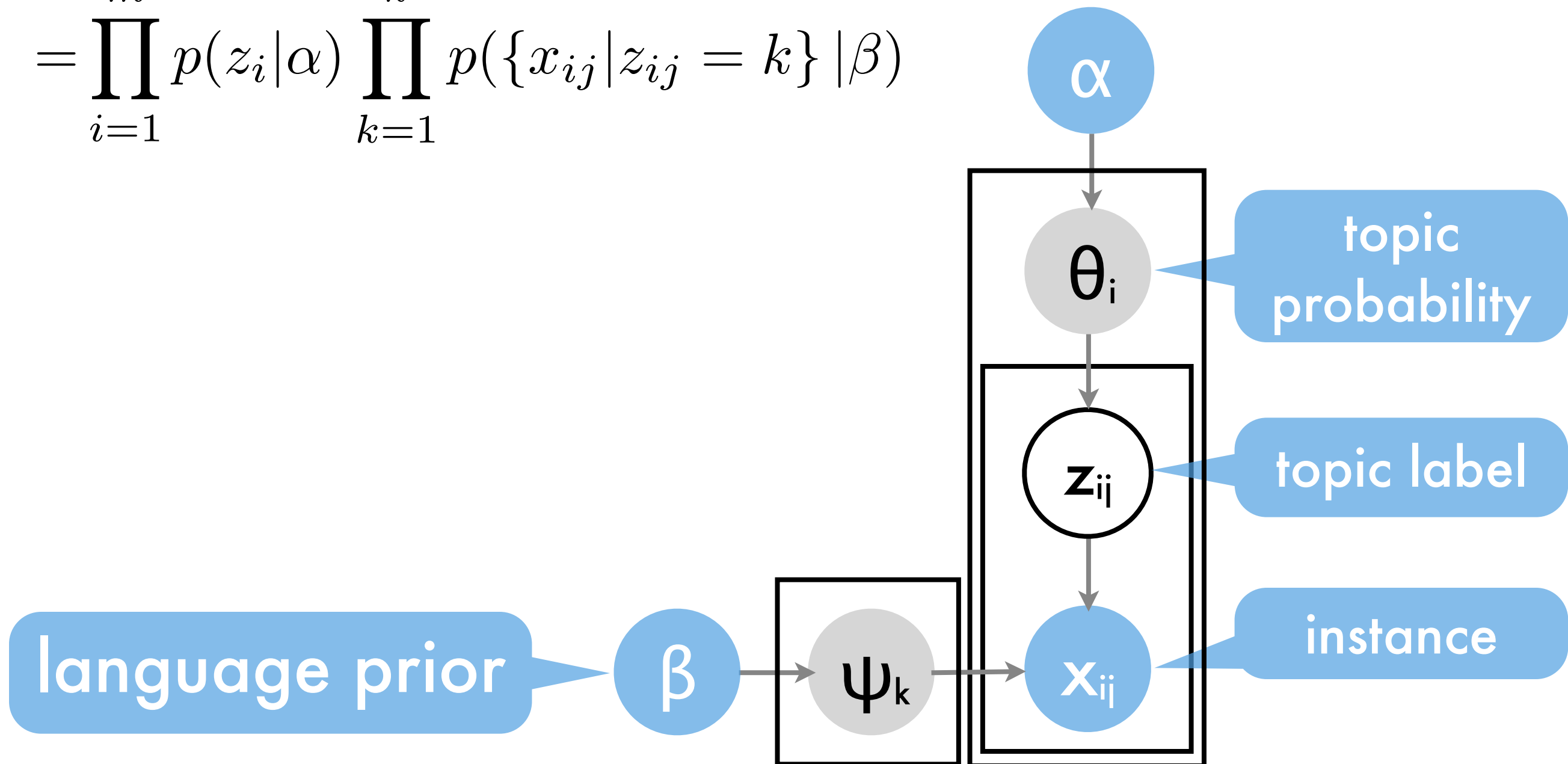
topic label

instance



# Collapsed Sampler

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



# Collapsed Sampler

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

sample  $z$   
sequentially

language prior

$\beta$

$\psi_k$

$x_{ij}$

$\alpha$

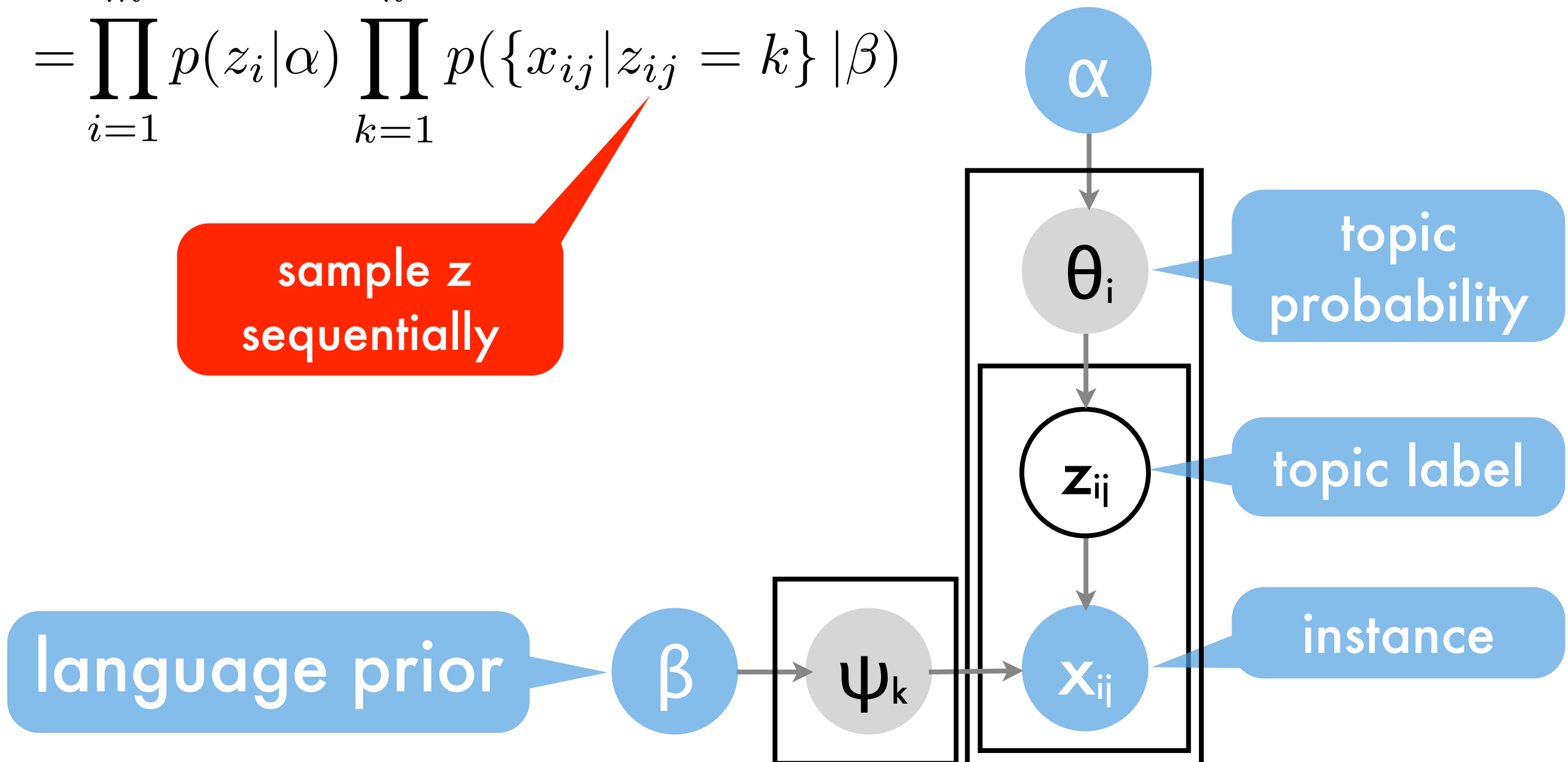
$\theta_i$

$z_{ij}$

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

$\beta$

$\psi_k$

$x_{ij}$

$z_{ij}$

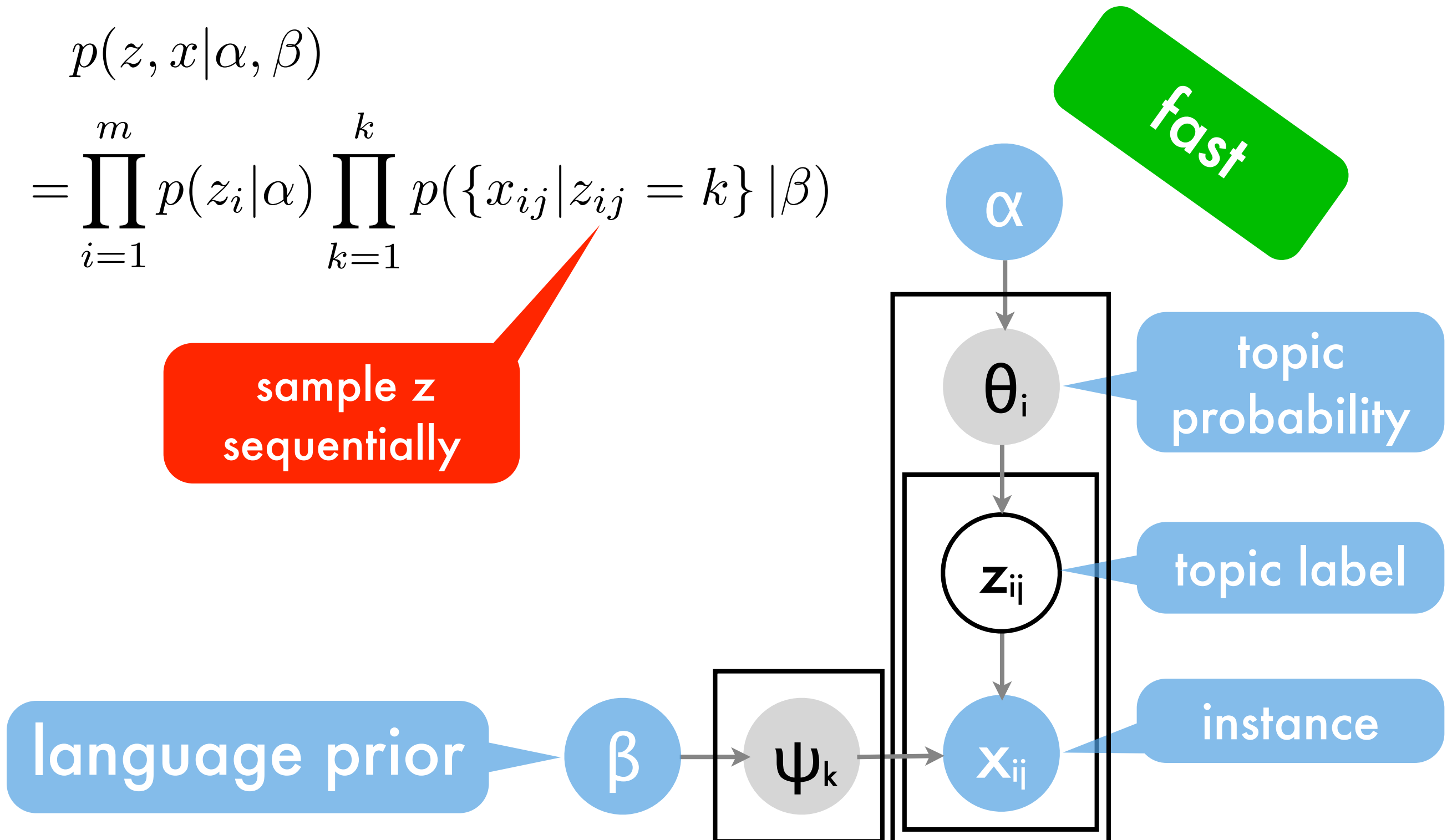
$\theta_i$

$\alpha$

topic probability

topic label

instance



# Collapsed Sampler

Griffiths & Steyvers, 2005

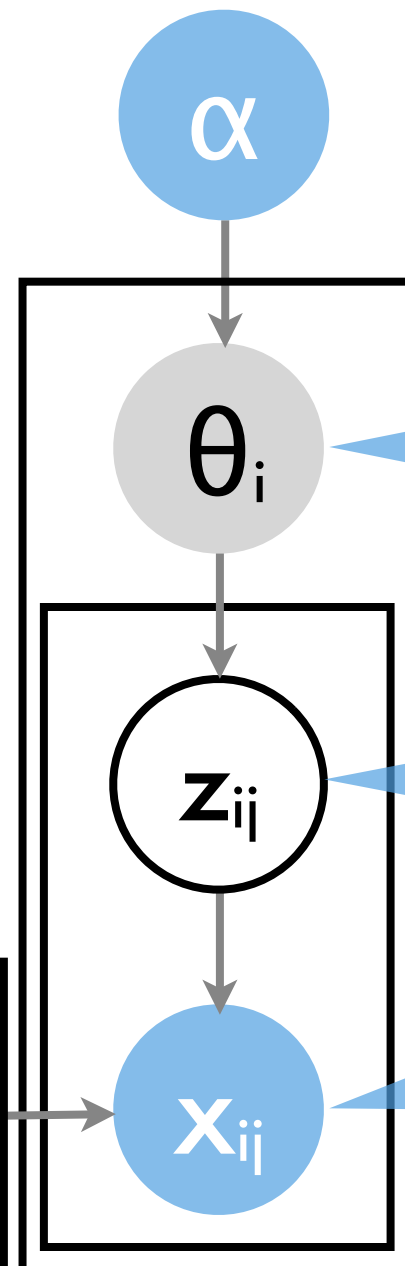
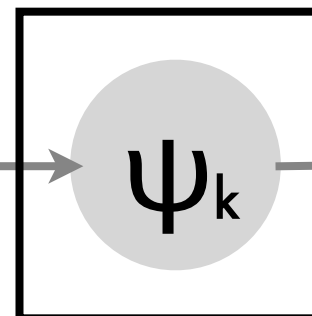
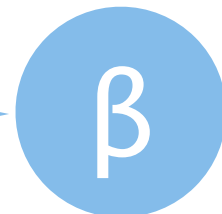
$$p(z, x | \alpha, \beta)$$

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

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

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

language prior



topic probability

topic label

instance

# Collapsed Sampler

Griffiths & Steyvers, 2005

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

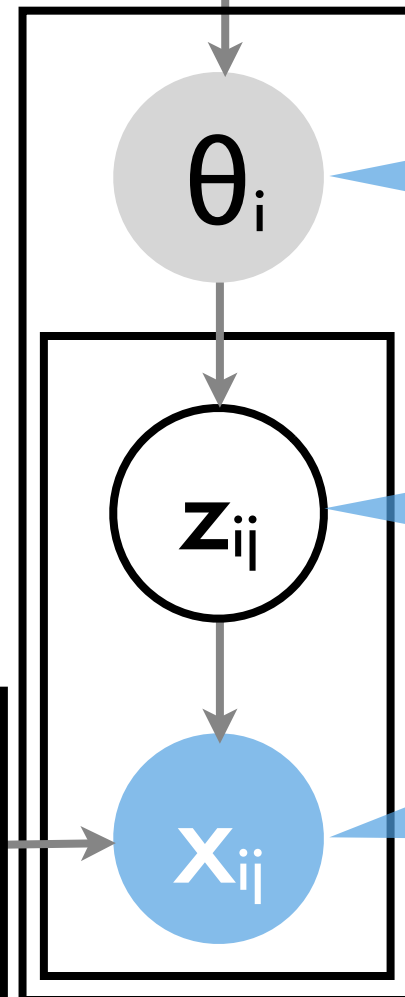
$$\frac{n^{-ij}(t, d) + \alpha_t}{n^{-i}(d) + \sum_t \alpha_t}$$

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

language prior

$\beta$

$\psi_k$



topic probability

topic label

instance

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

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

# State of the art

## UMass Mallet, UC Irvine, Google

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

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

slow

# State of the art

## UMass Mallet, UC Irvine, Google

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

changes rapidly

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

slow

# State of the art

## UMass Mallet, UC Irvine, Google

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

changes rapidly

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

slow

moderately fast

# State of the art

## UMass Mallet, UC Irvine, Google

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

table out of sync

memory inefficient

blocking

network bound

changes rapidly

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

slow

moderately fast

# Our Approach

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



# Our Approach

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

network  
bound

concurrent  
cpu hdd net

# Our Approach

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

network  
bound

memory  
inefficient

concurrent  
cpu hdd net

minimal  
view

# Our Approach

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

network  
bound

memory  
inefficient

table out  
of sync

concurrent  
cpu hdd net

minimal  
view

continuous  
sync

# Our Approach

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

network  
bound

memory  
inefficient

table out  
of sync

blocking

concurrent  
cpu hdd net

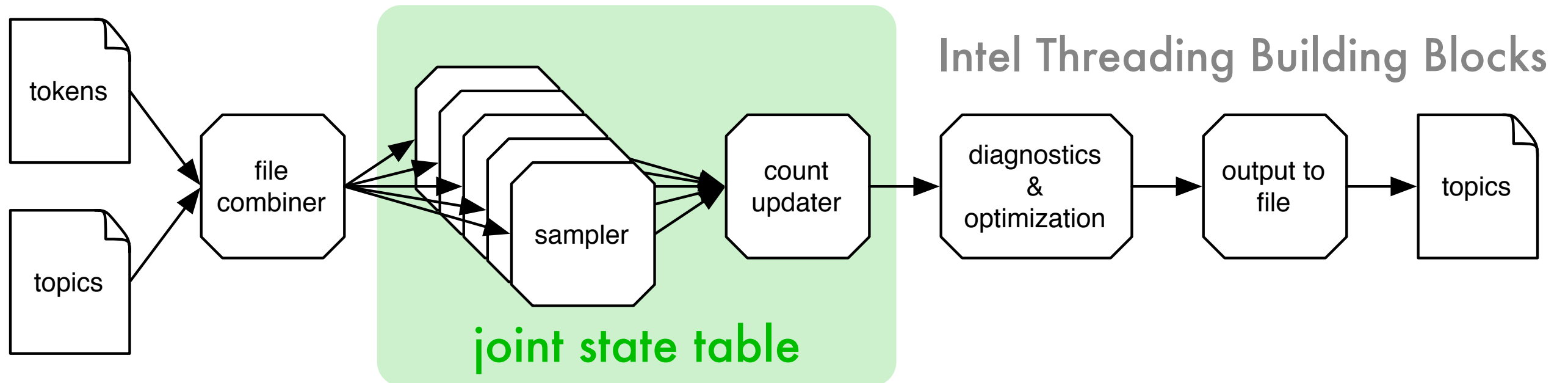
minimal  
view

continuous  
sync

barrier  
free

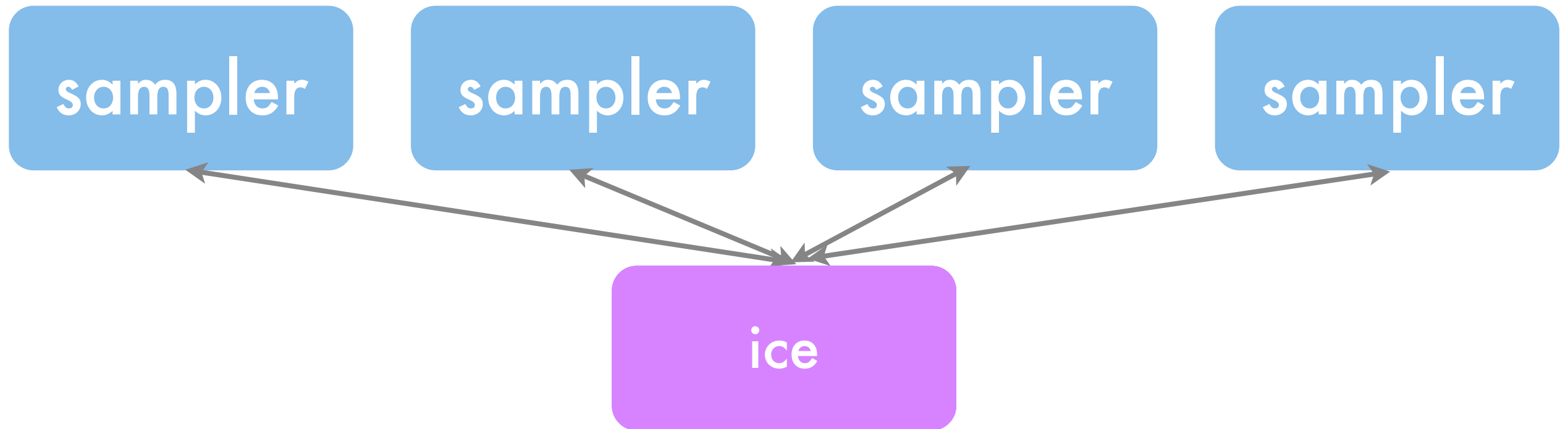
# Architecture details

# Multicore Architecture



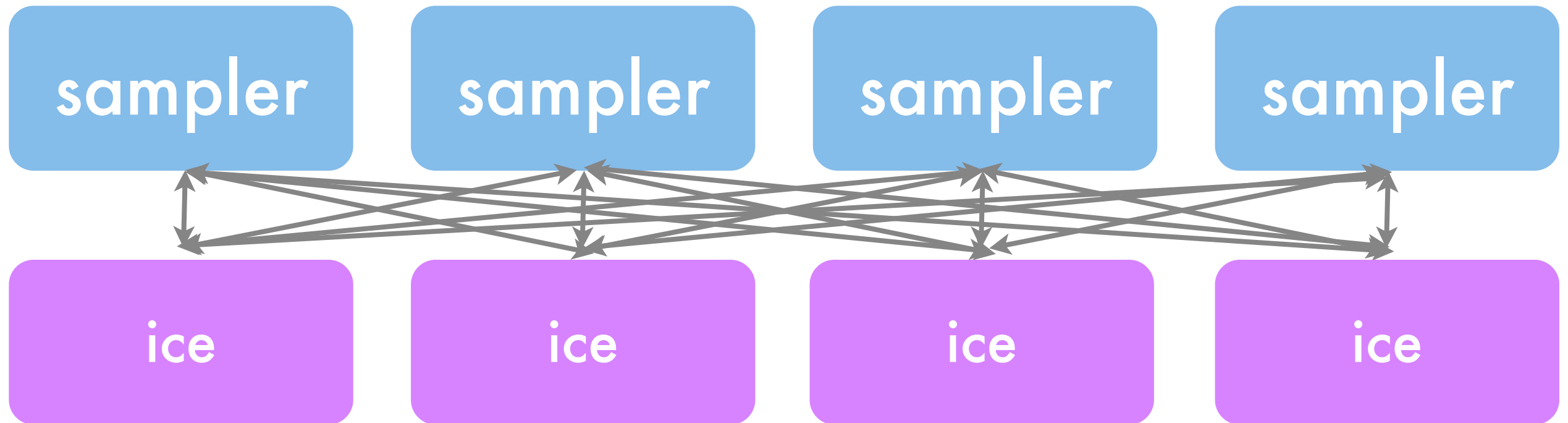
- Decouple multithreaded sampling and updating (almost) avoids stalling for locks in the sampler
- Joint state table
  - much less memory required
  - samplers synchronized (10 docs vs. millions delay)
- Hyperparameter update via stochastic gradient descent
- No need to keep documents in memory (streaming)

# Cluster Architecture



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

# Cluster Architecture



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



# Making it work

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

# Easily extensible

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

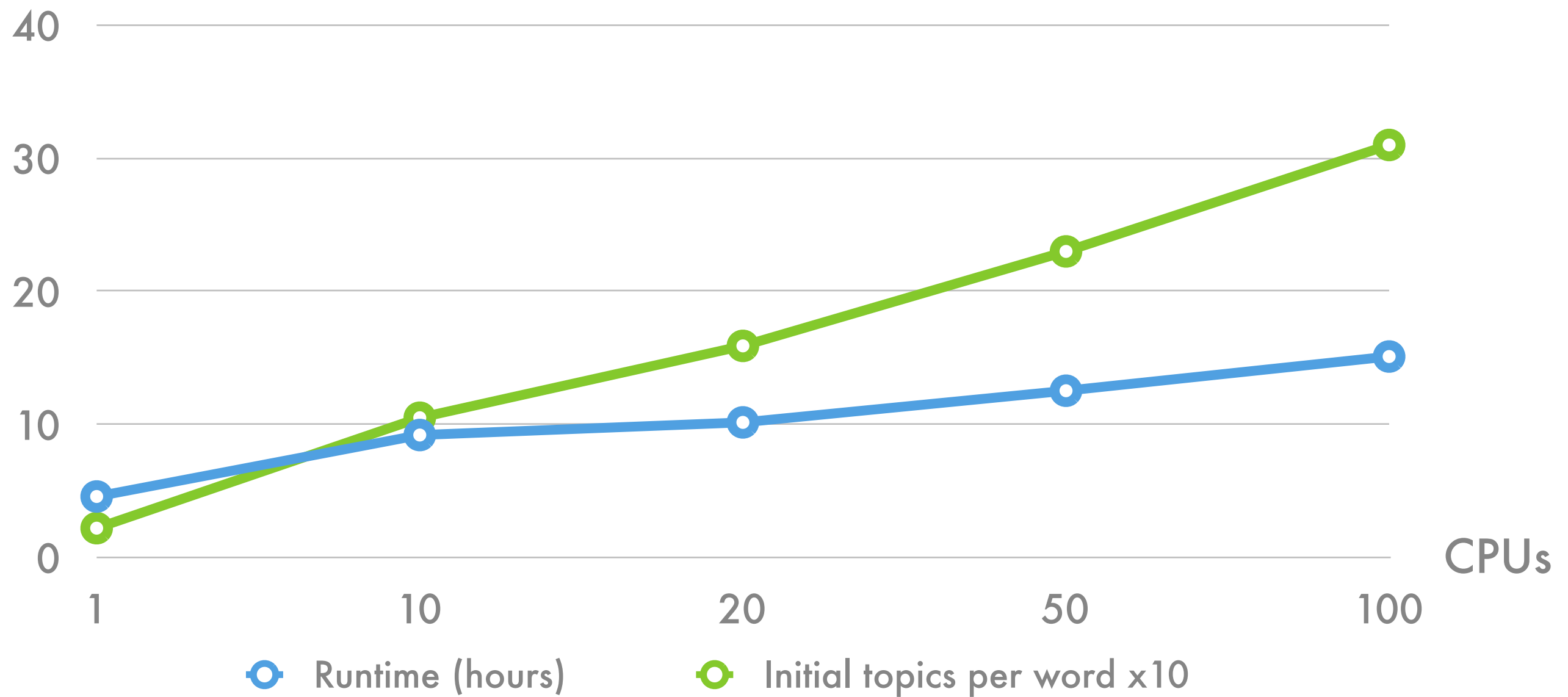
	Google LDA	Mallet	Irvine'08	Irvine'09	Yahoo LDA
Multicore	no	yes	yes	yes	yes
Cluster	MPI	no	MPI	point 2 point	memcached
State table	dictionary split	separate sparse	separate	separate	joint sparse
Schedule	synchronous exact	synchronous exact	synchronous exact	asynchronous approximate messages	asynchronous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

# Speed

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

# Scalability

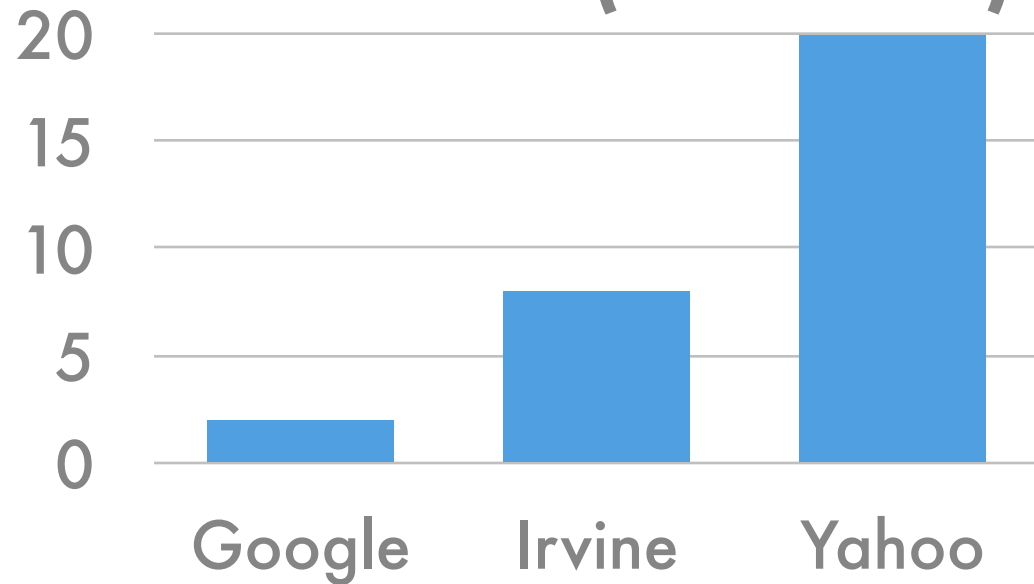
200k documents/computer



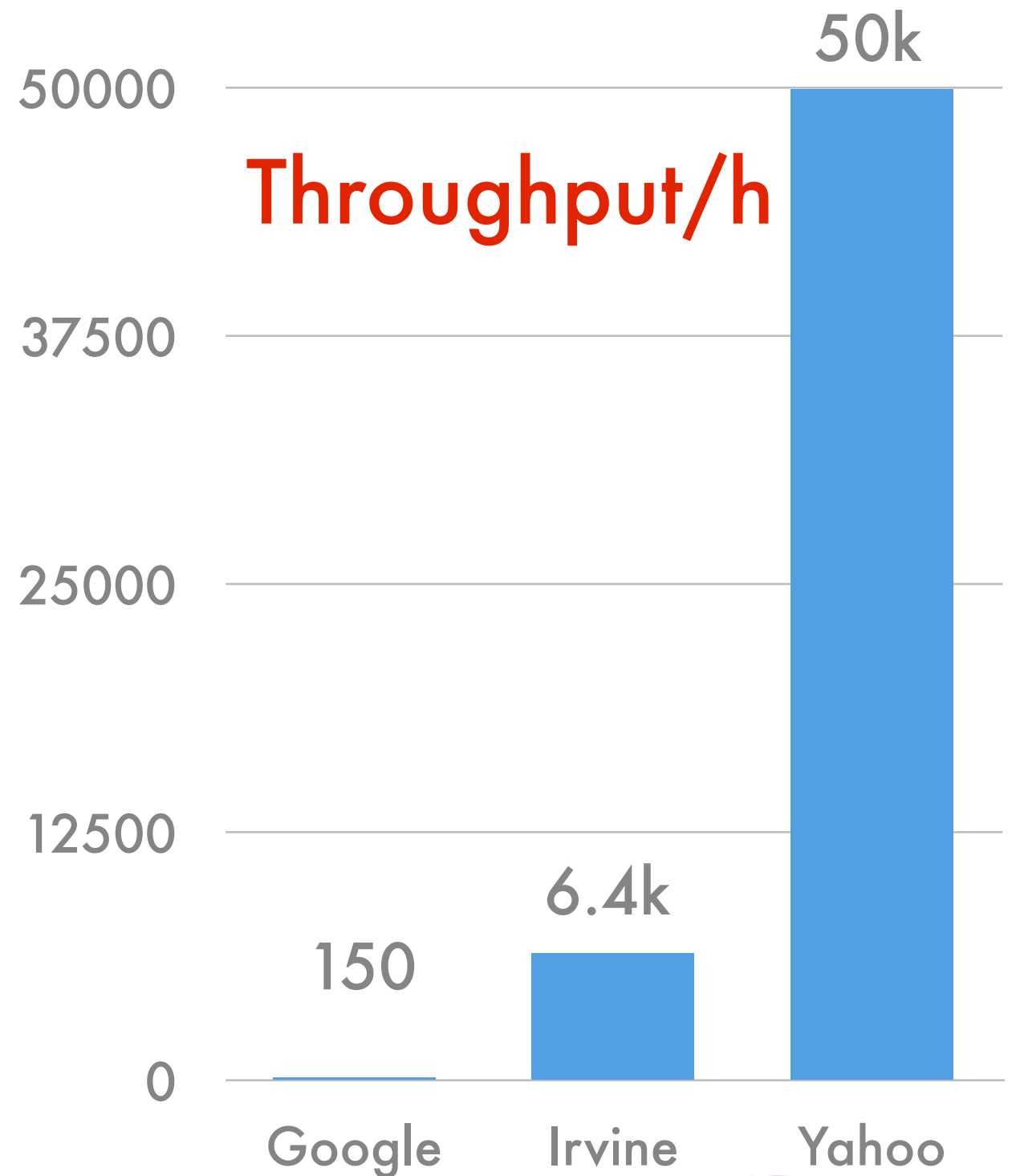
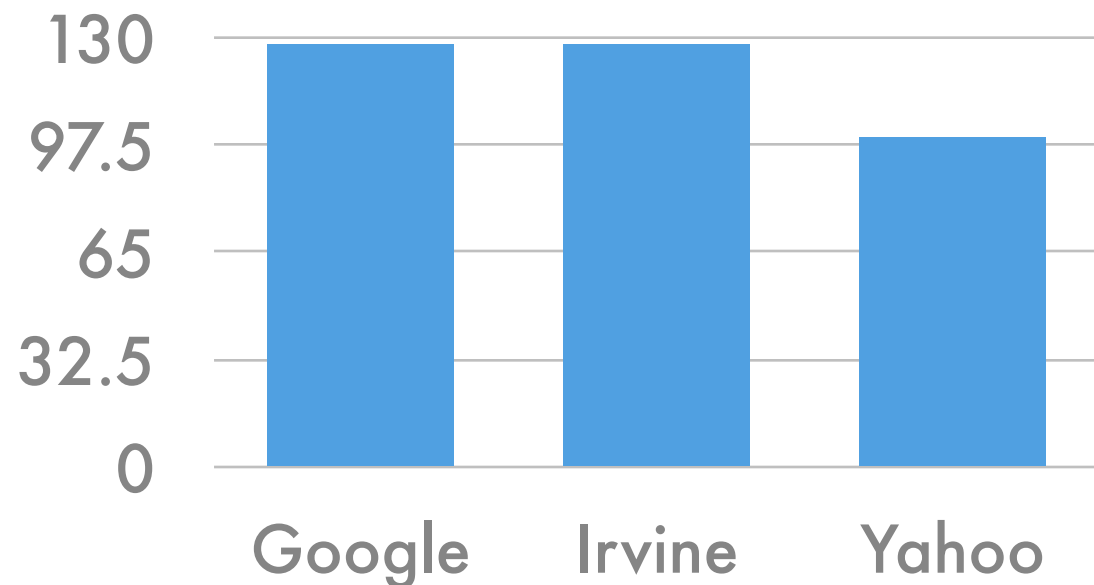
**Likelihood even improves with parallelism!**  
**-3.295 (1 node) -3.288 (10 nodes) -3.287 (20 nodes)**

# The Competition

Dataset size (millions)



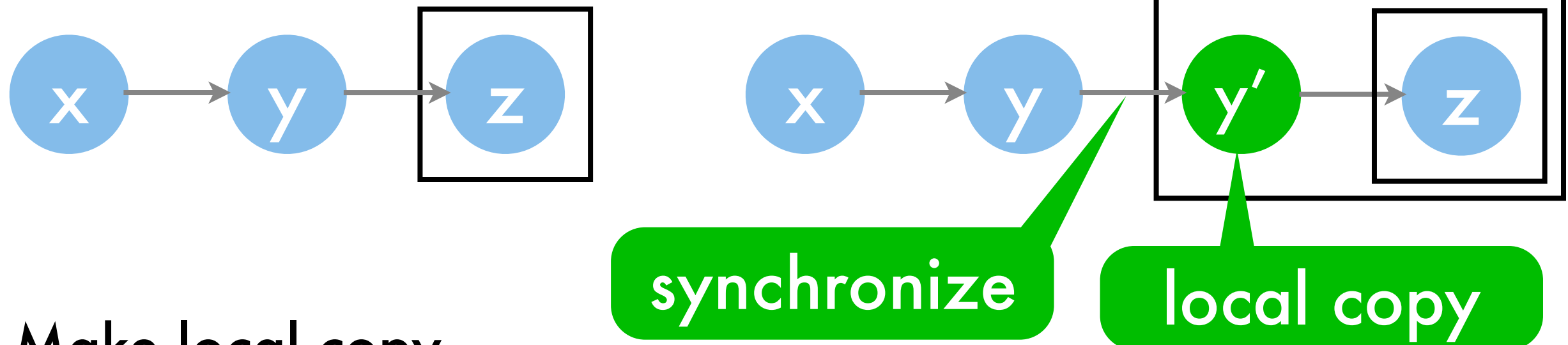
Cluster size



# Design Principles

# Variable Replication

- Global shared variable

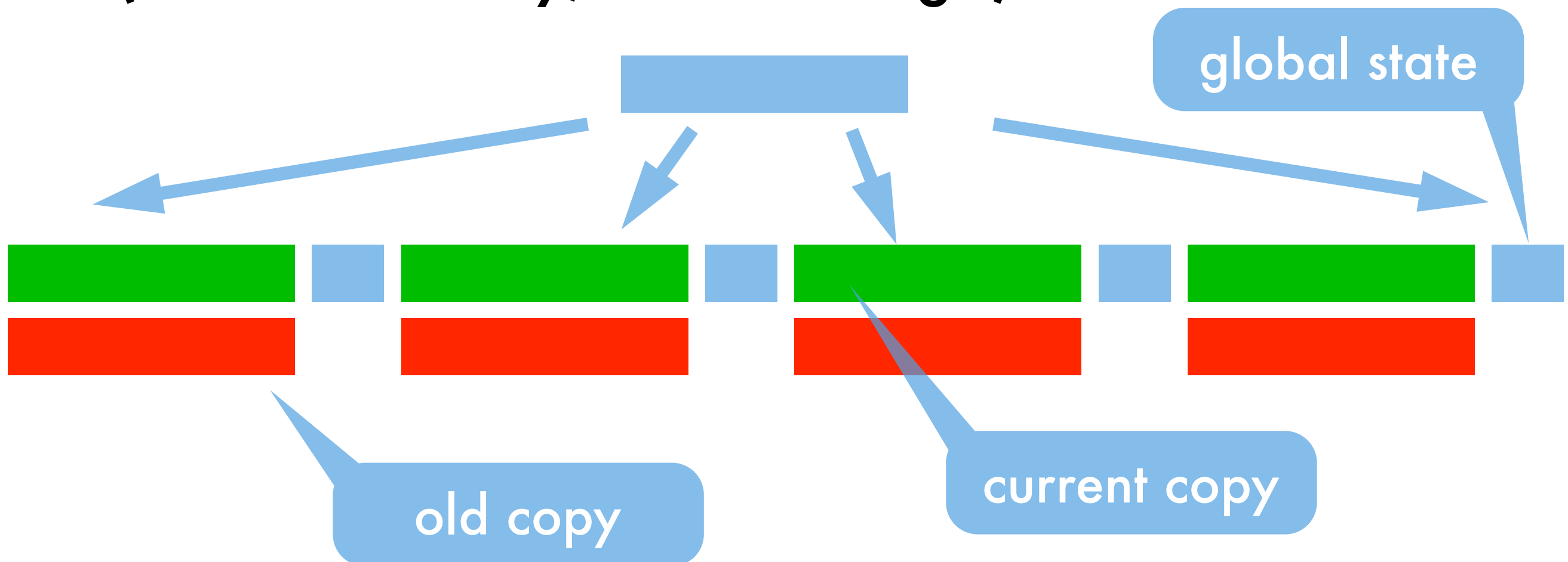


- Make local copy
  - Distributed (key,value) storage table for global copy
  - Do all bookkeeping locally (store old versions)
  - Sync local copies asynchronously using message passing (no global locks are needed)
- **This is an approximation!**



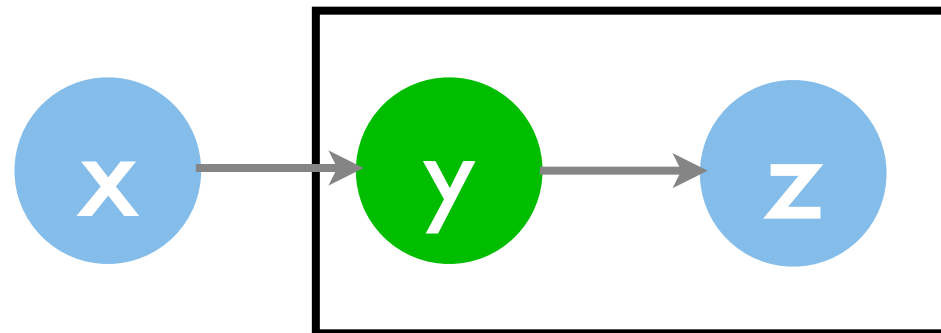
# Asymmetric Message Passing

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

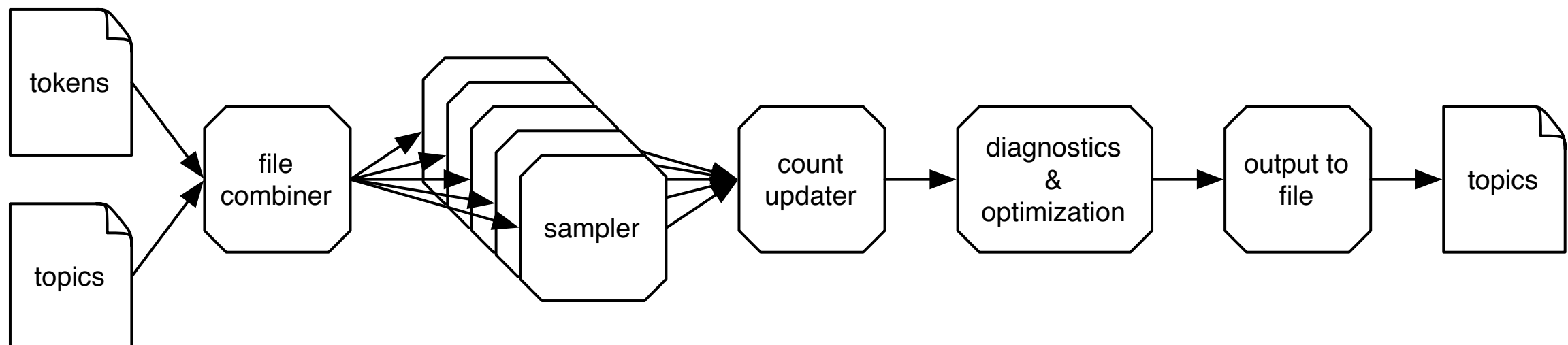


# Out of core storage

- Very large state space



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

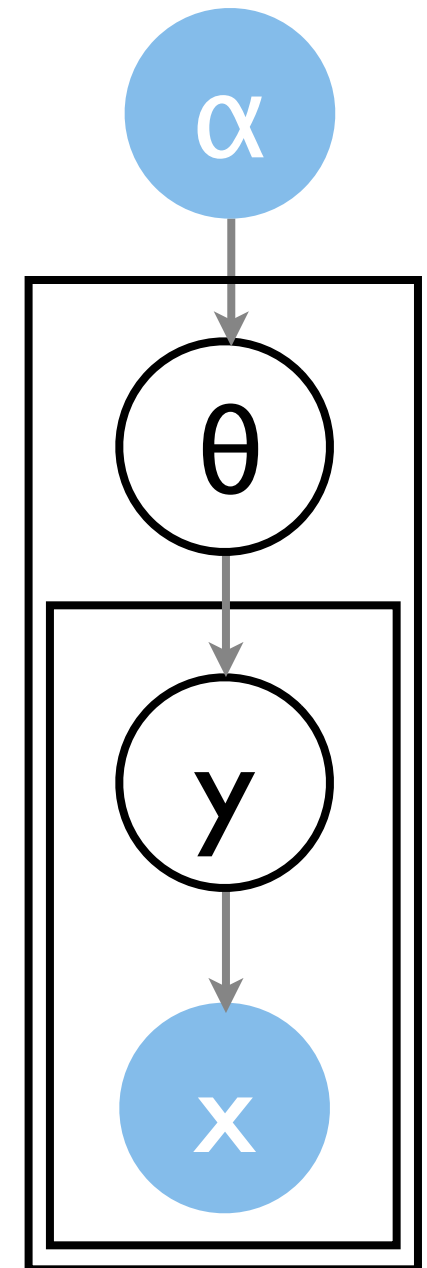


# Part 6 - Advanced Modeling

# Advances in Representation

# Extensions to topic models

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



# Correlated topic models

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

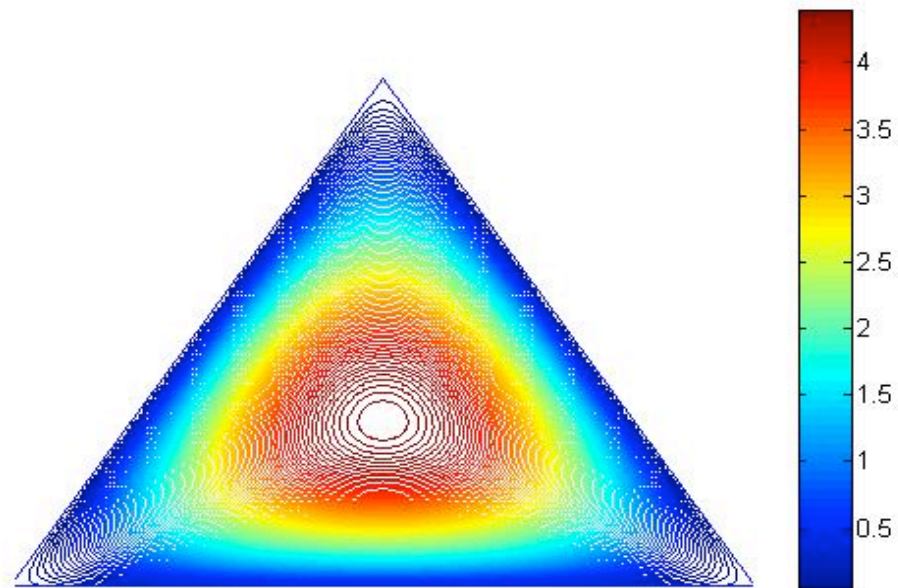
# Correlated topic models

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

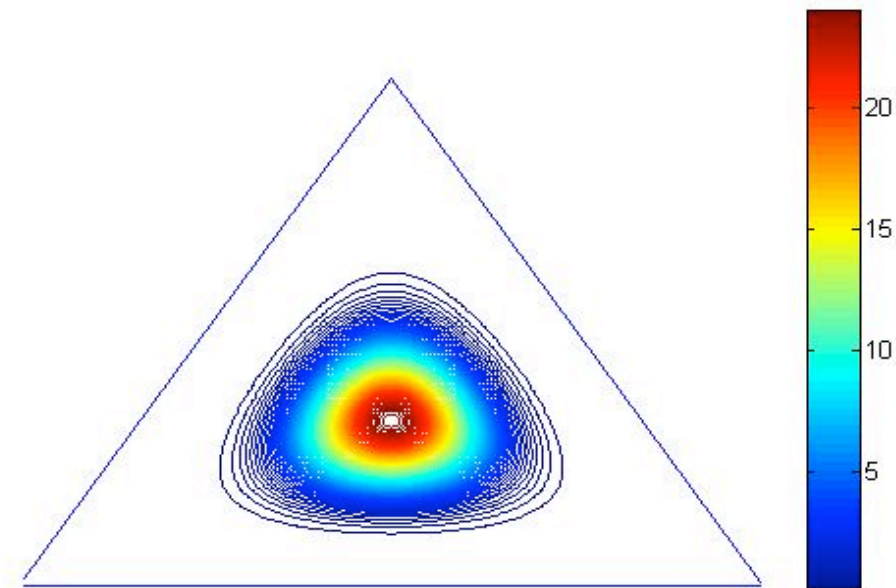
Blei & Lafferty 2005; Ahmed & Xing 2007

# Dirichlet prior on topics

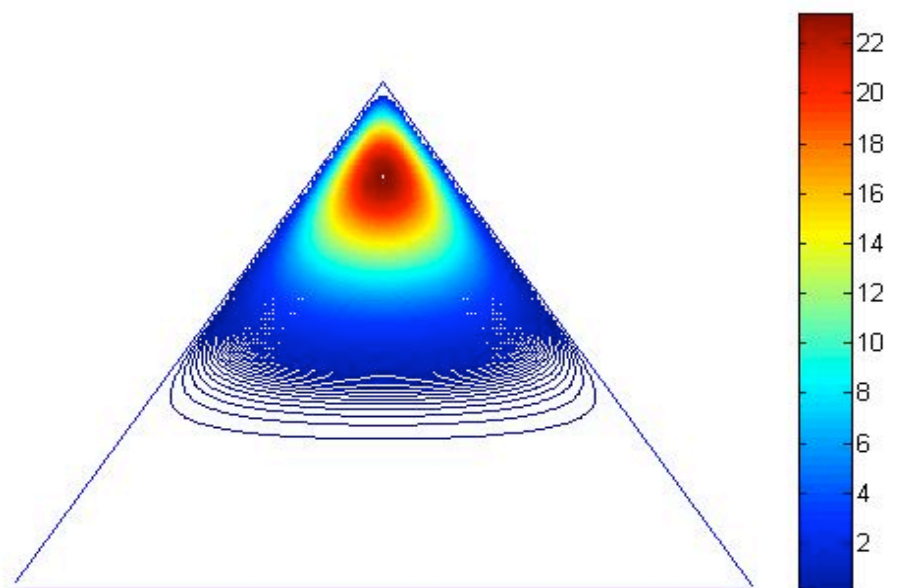
Alpha = [2.00 2.00 2.00]



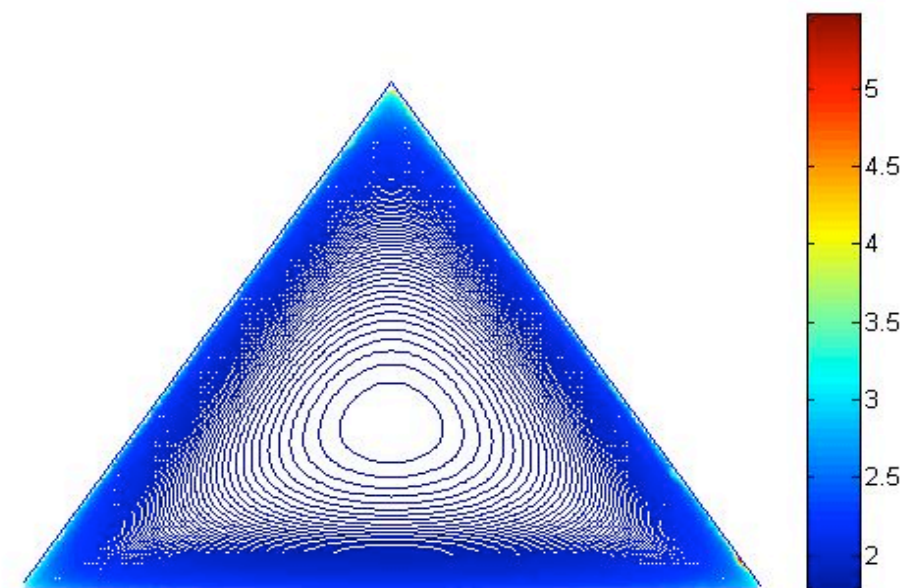
Alpha = [10.00 10.00 10.00]



Alpha = [2.00 10.00 2.00]

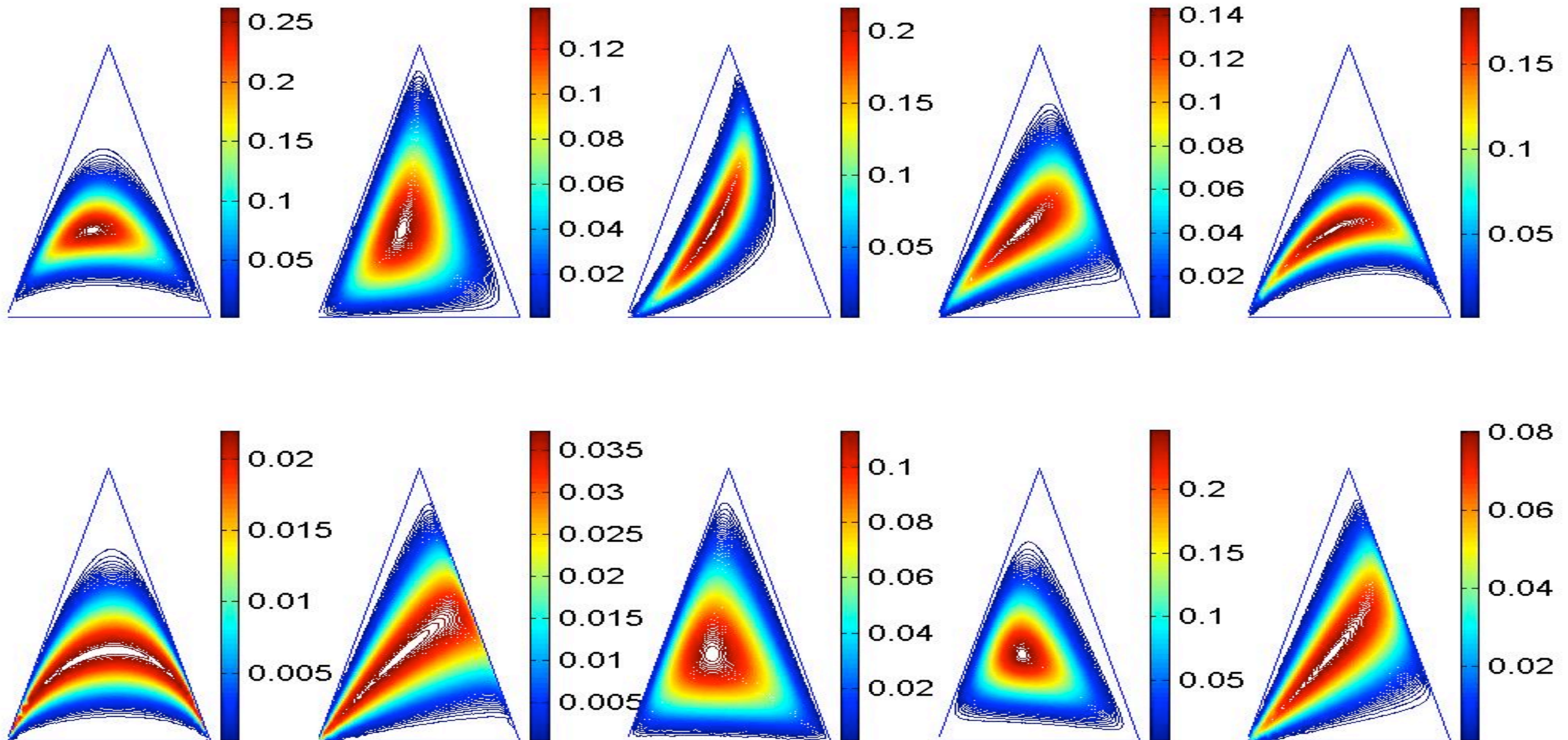


Alpha = [0.90 0.90 0.90]



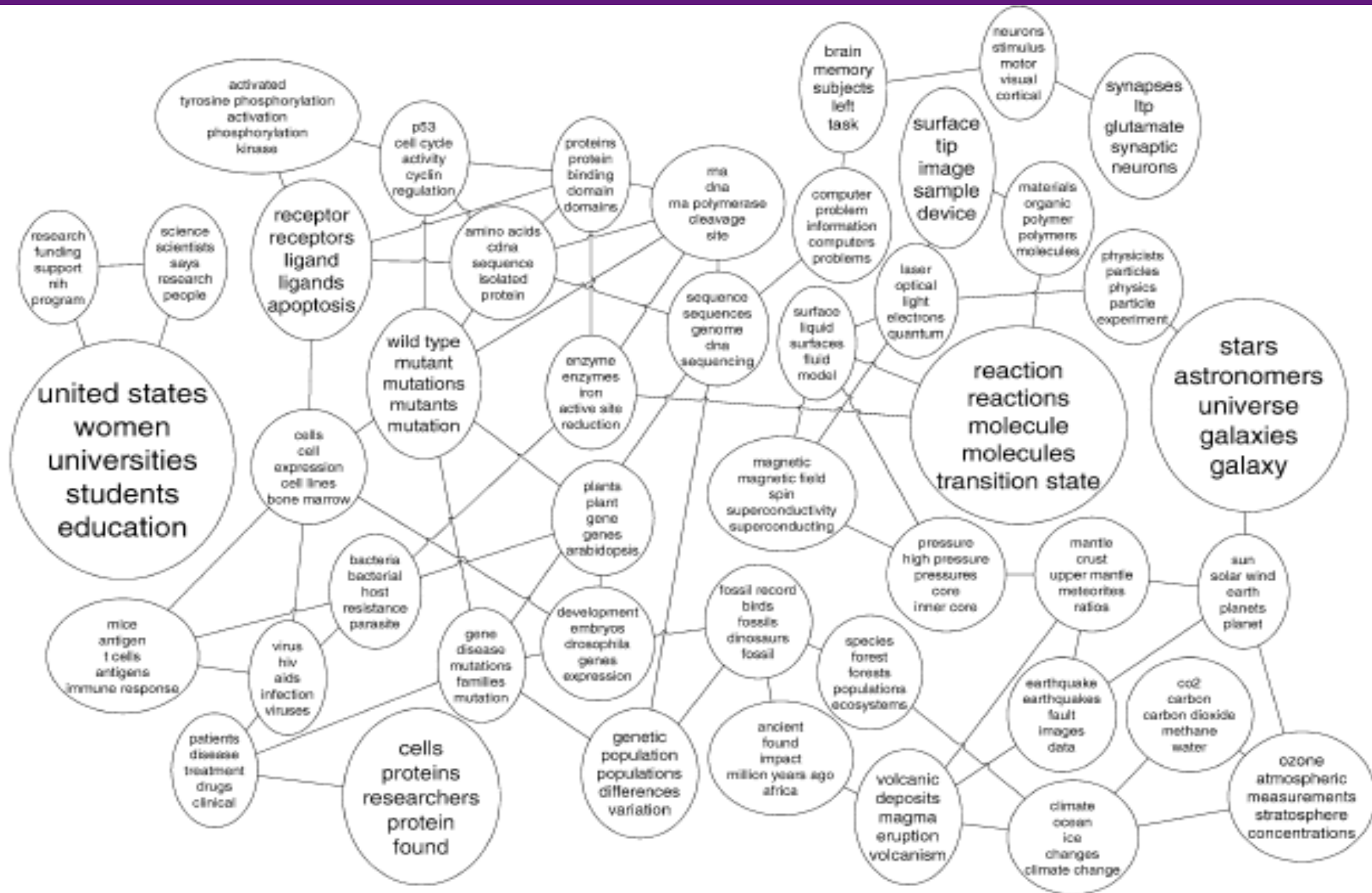


# Log-normal prior on topics



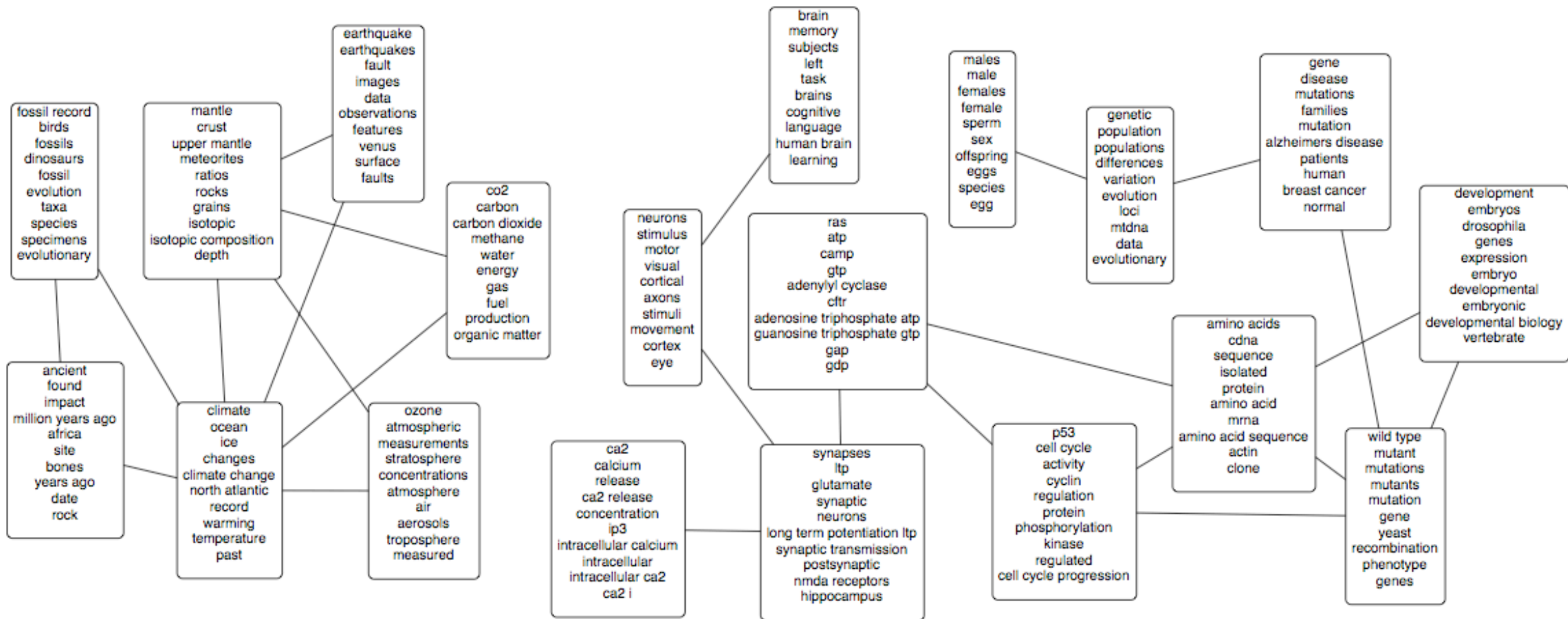
$$\theta = e^{\eta - g(\eta)} \quad \text{with} \quad \eta \sim \mathcal{N}(\mu, \Sigma)$$

# Correlated topics



Blei and Lafferty 2005

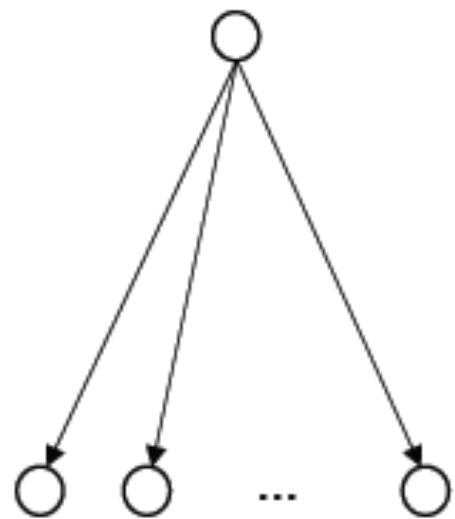
# Correlated topics



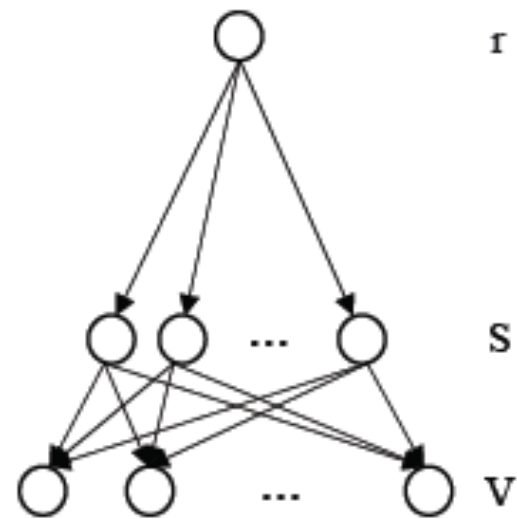


# Pachinko Allocation

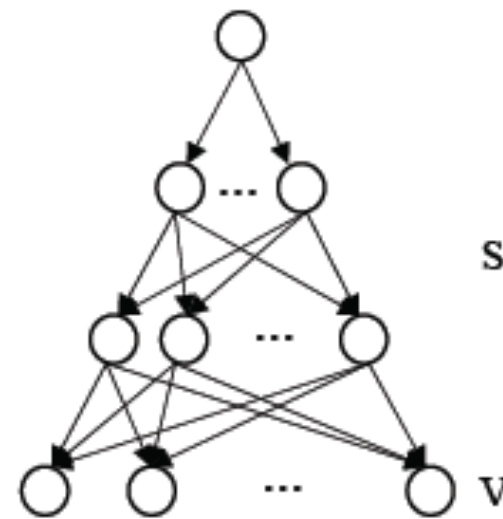
- Model the prior as a Directed Acyclic Graph
- Each document is modeled as multiple paths
- To sample a word, first select a path and then sample a word from the final topic
- The topics reside on the leaves of the tree



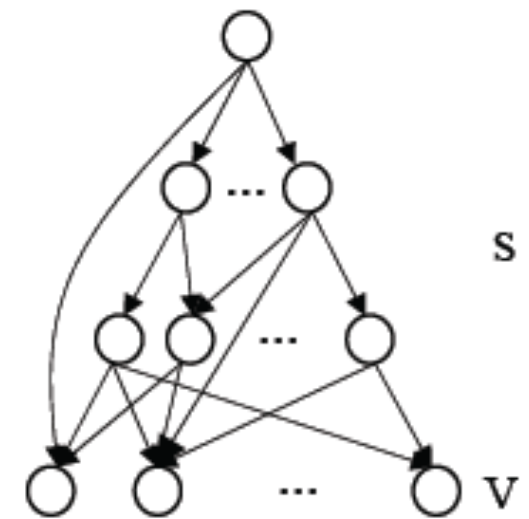
(a) Dirichlet Multinomial



(b) LDA

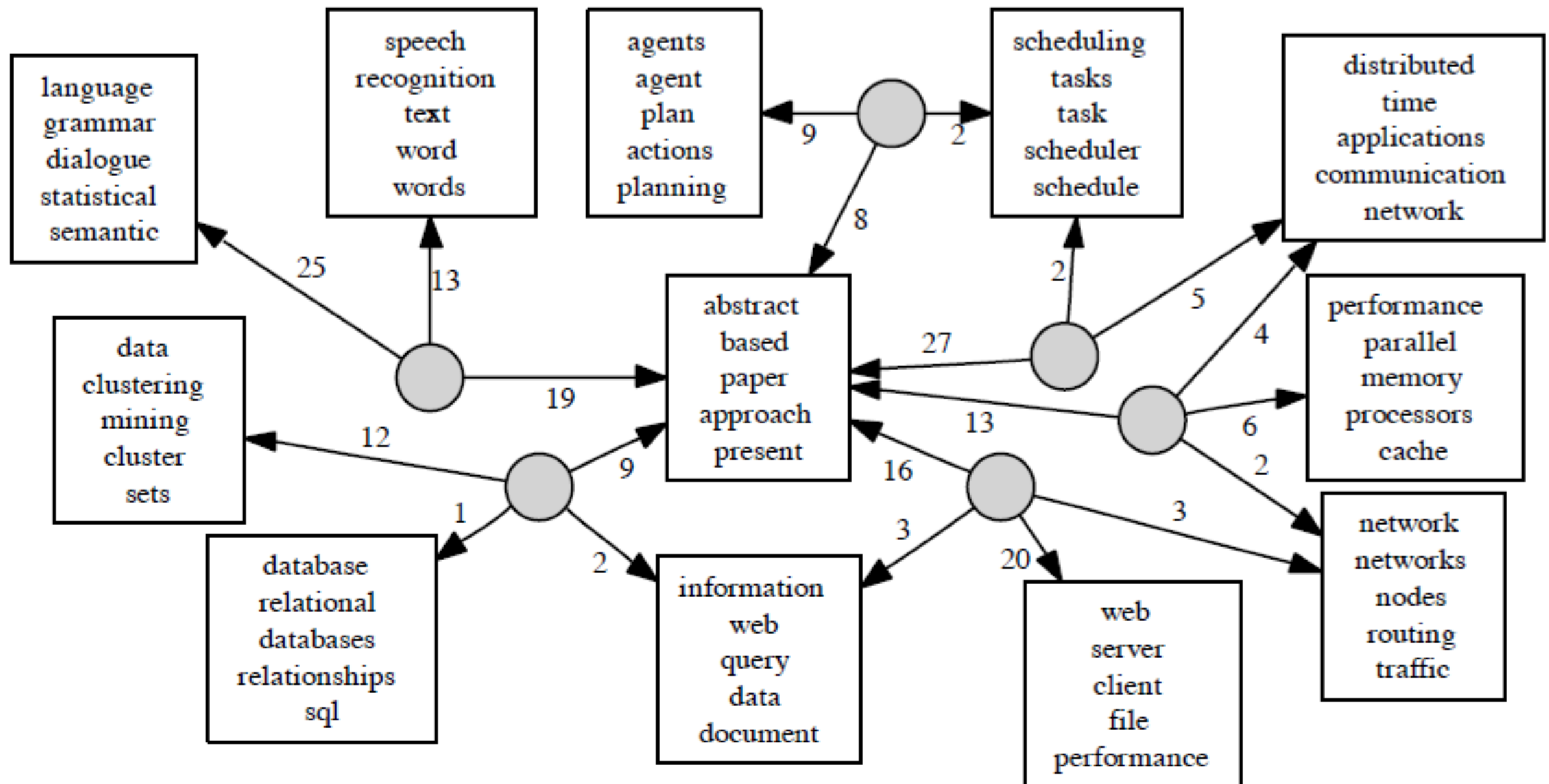


(c) Four-Level PAM



(d) Arbitrary PAM

# Pachinko Allocation

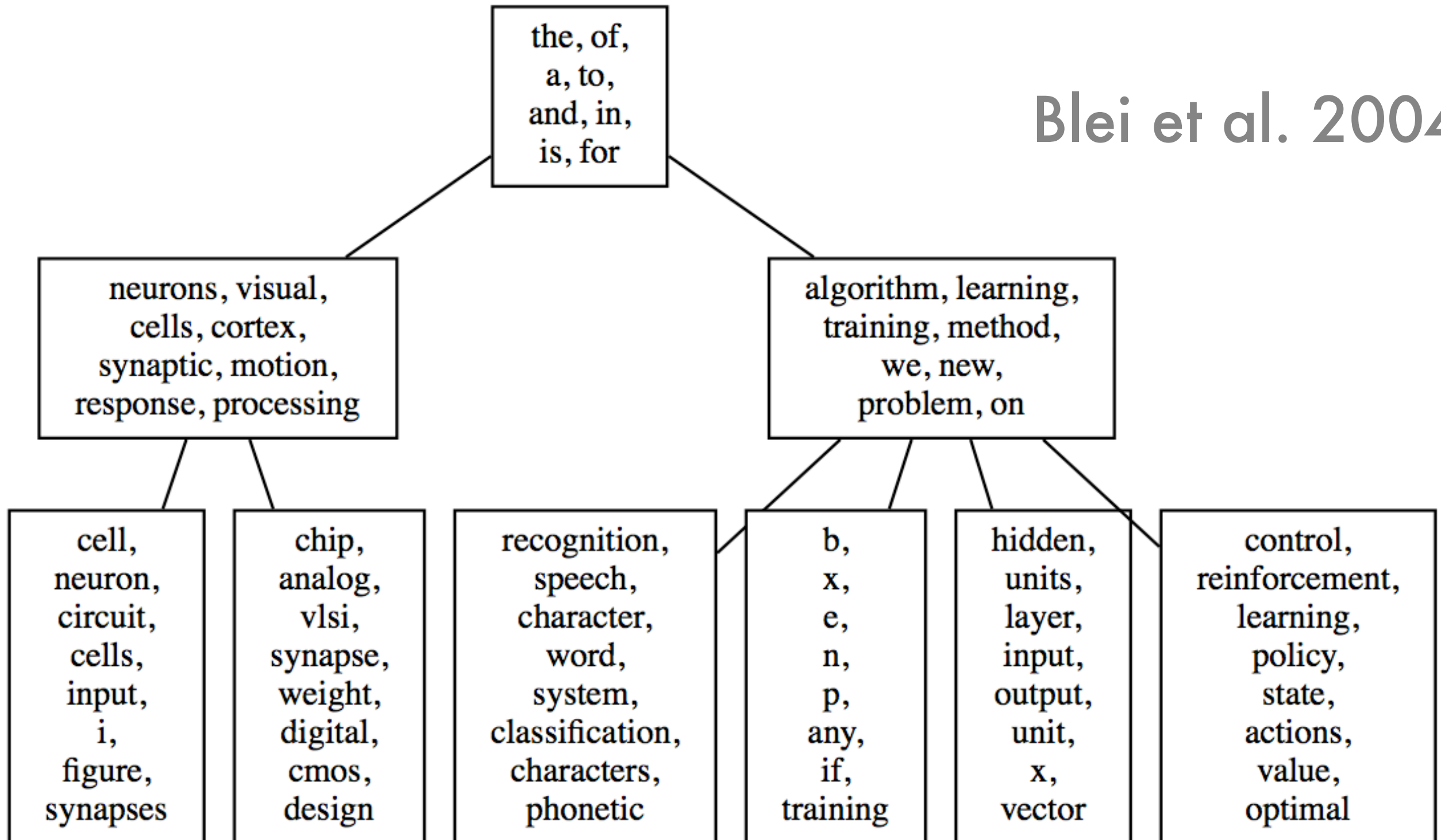


# Topic Hierarchies

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

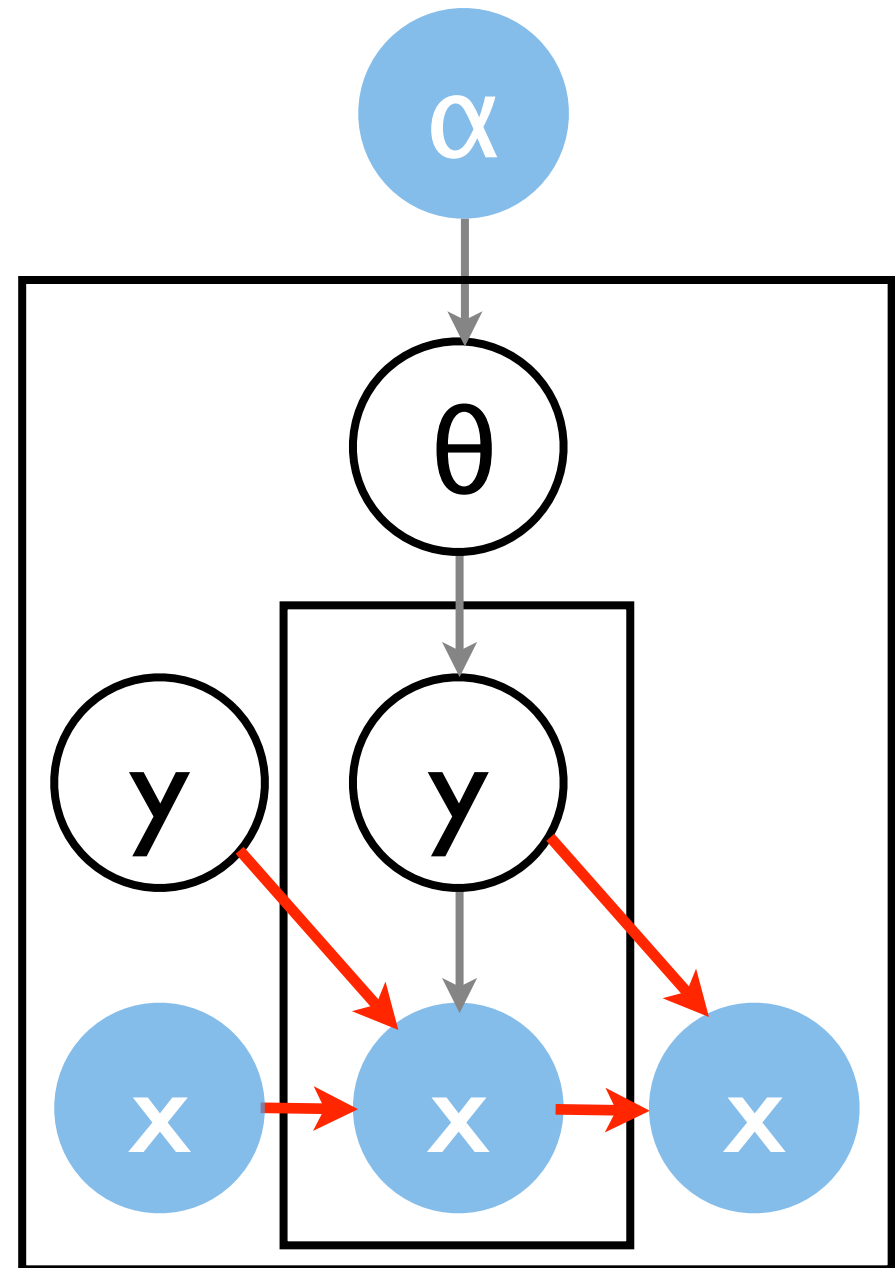
# Topic Hierarchies

Blei et al. 2004



# Topical n-grams

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





# Topic n-grams

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

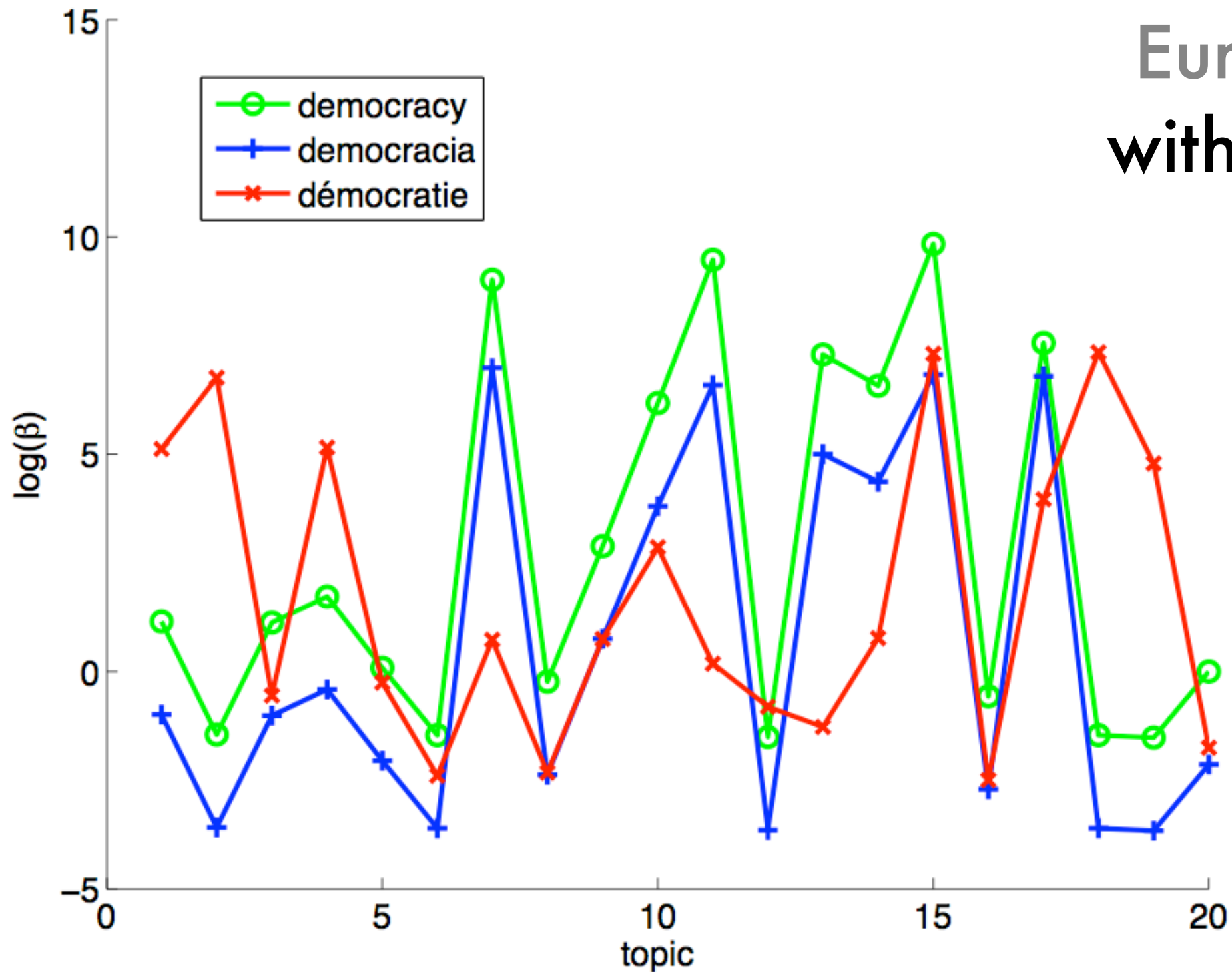
# Side information

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

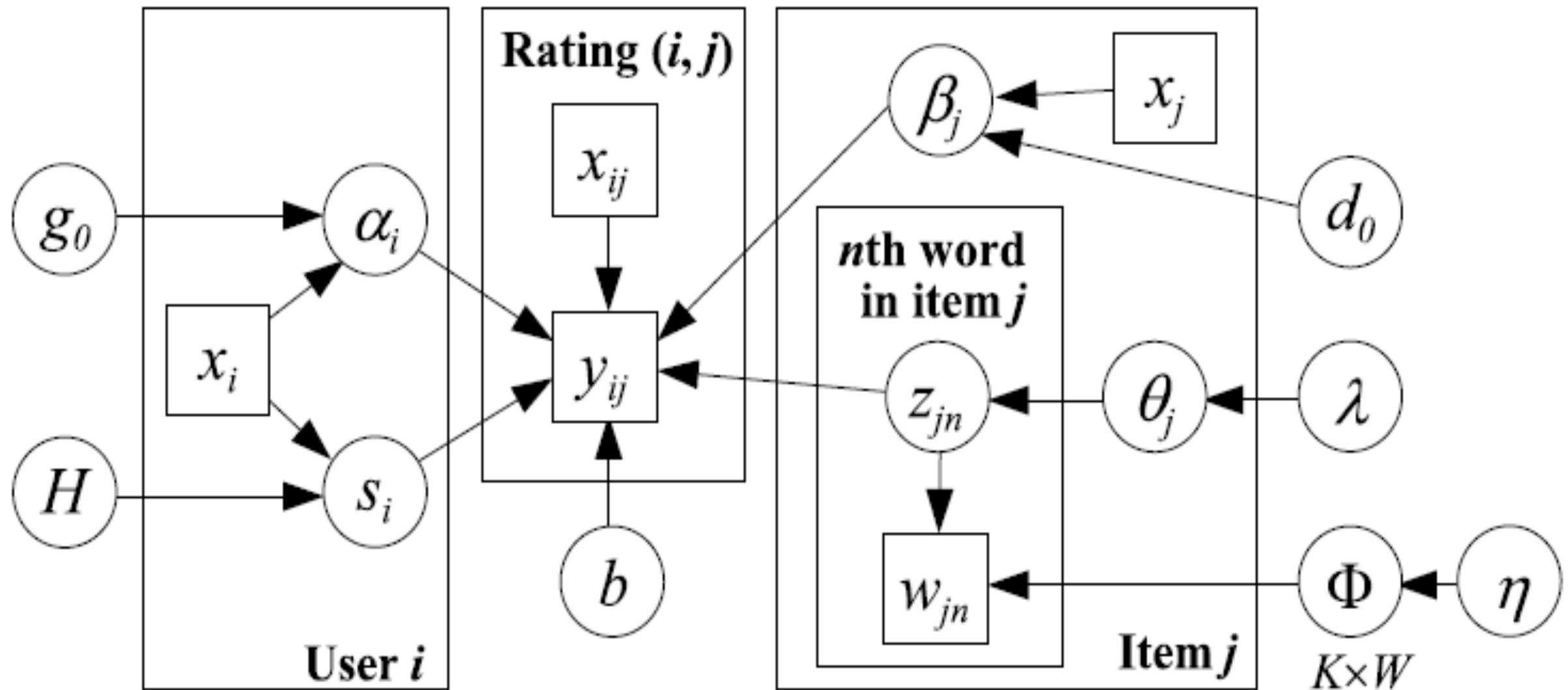
# Downstream conditioning

DC

Europarl corpus  
**without alignment**

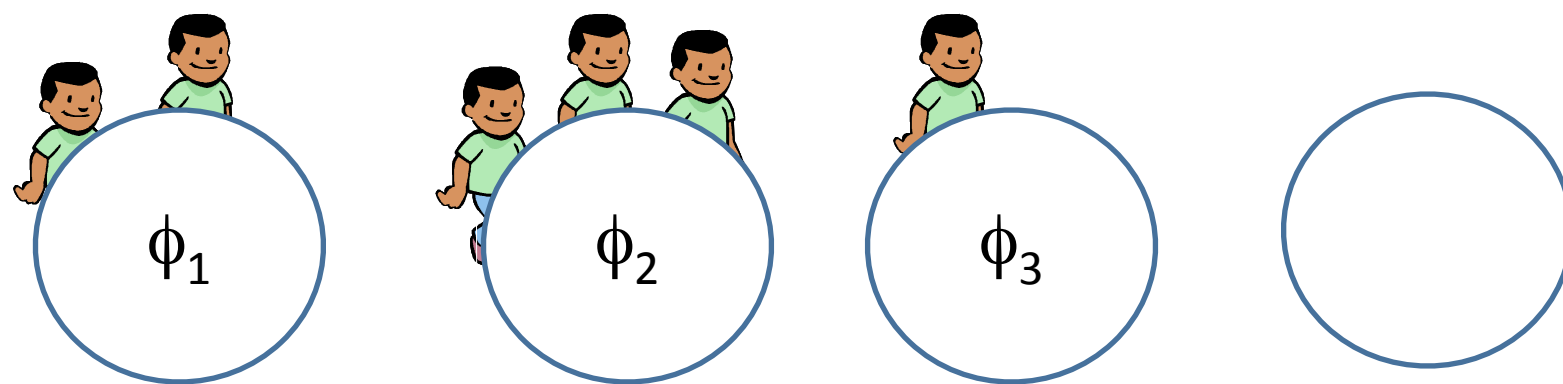


# Recommender Systems



Agarwal & Chen, 2010

# Chinese Restaurant Process



# Problem

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

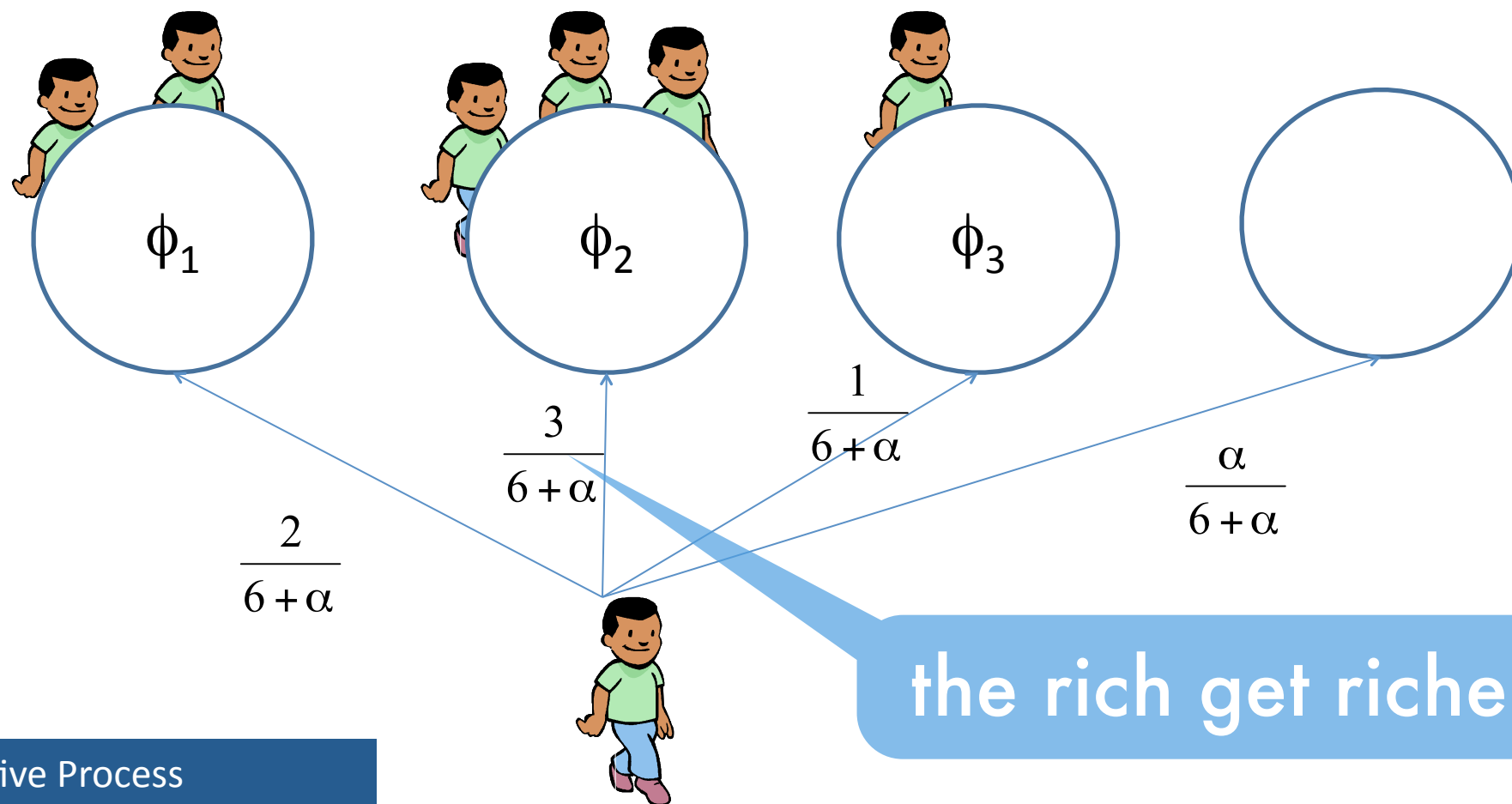
$$p(y|Y, \alpha) = \frac{n(y) + \alpha_y}{n + \sum_{y'} \alpha_{y'}}$$

- Infinite model

**Assume that the total smoother weight is constant**

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

# Chinese Restaurant Metaphor



## Generative Process

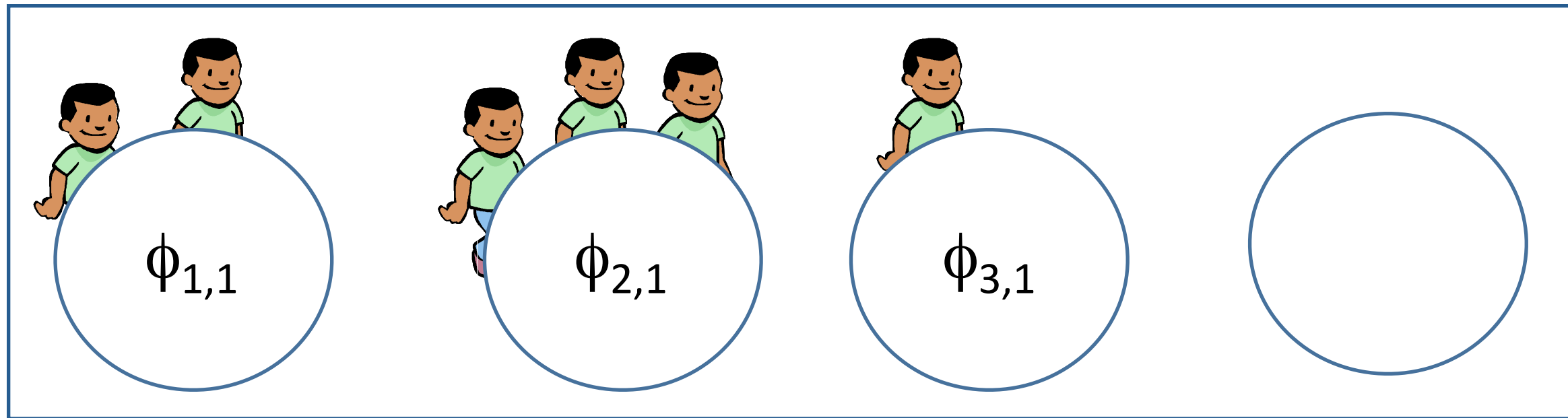
- For data point  $x_i$ 
  - Choose table  $j \propto m_j$  and Sample  $x_i \sim f(\phi_j)$
  - Choose a new table  $K+1 \propto \alpha$ 
    - Sample  $\phi_{K+1} \sim G_0$  and Sample  $x_i \sim f(\phi_{K+1})$

# Evolutionary Clustering

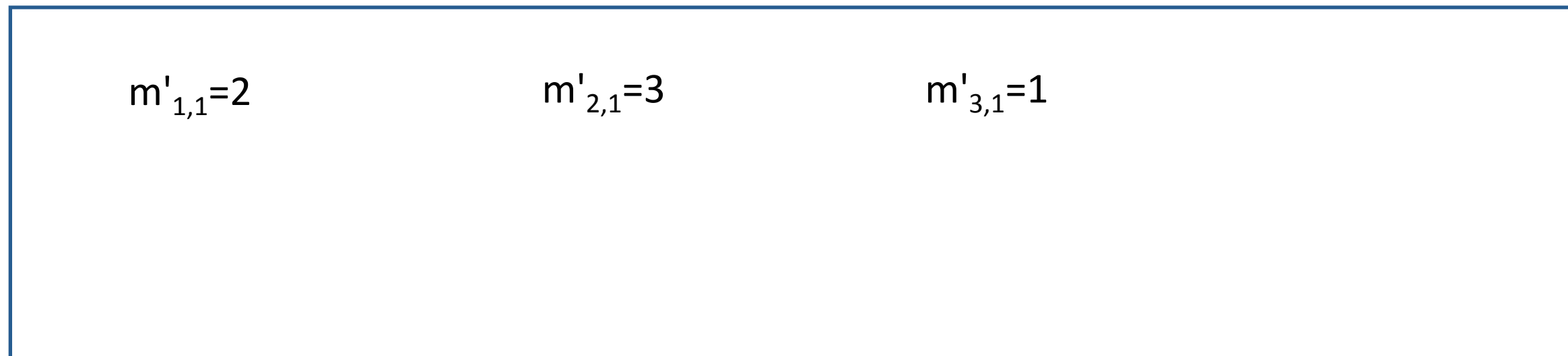
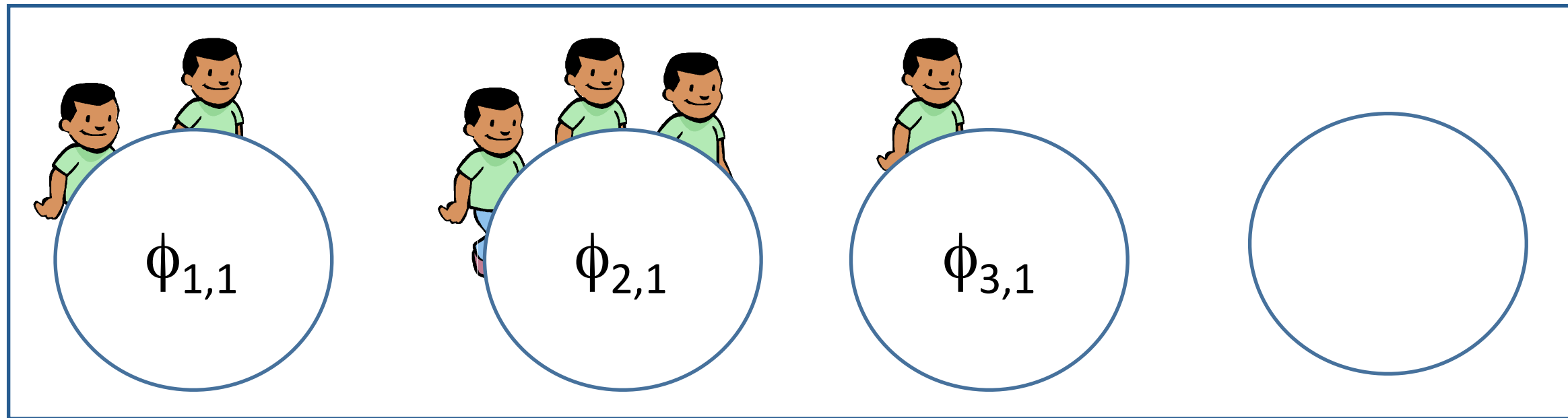
- **Time series of objects, e.g. news stories**
- **Stories appear / disappear**
- **Want to keep track of clusters automatically**



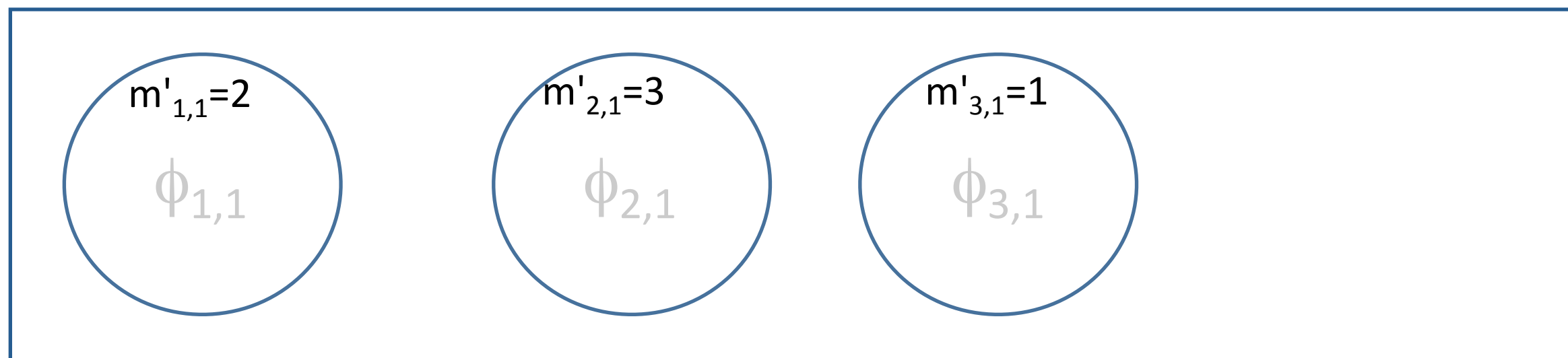
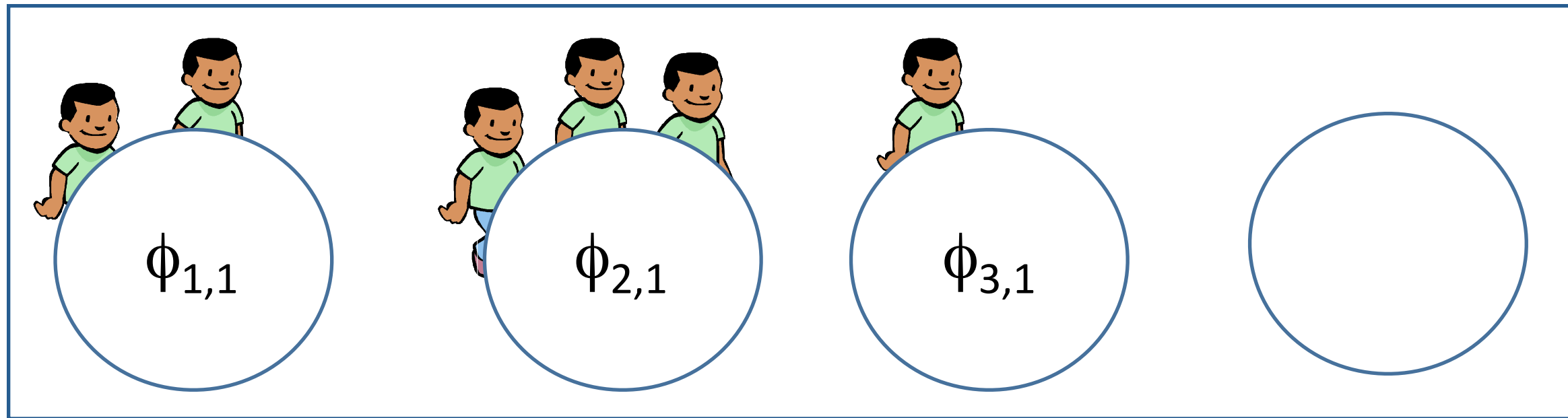
# Recurrent Chinese Restaurant Process



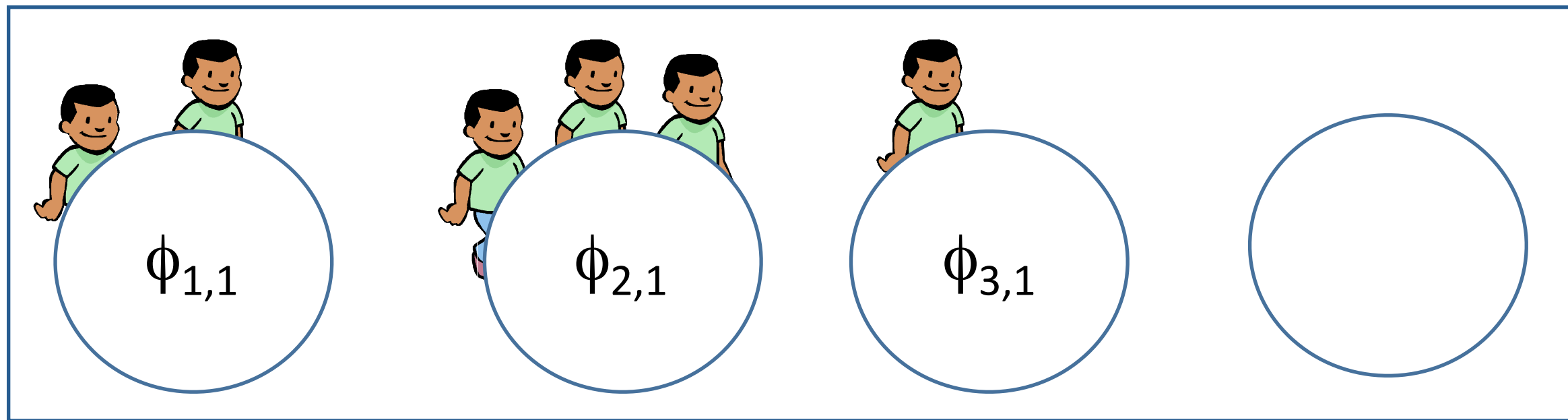
# Recurrent Chinese Restaurant Process



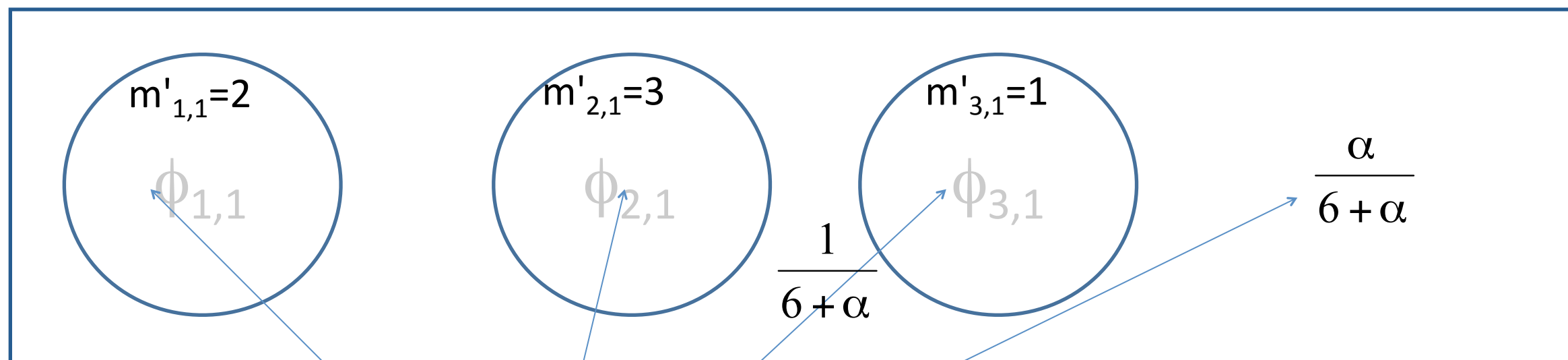
# Recurrent Chinese Restaurant Process



# Recurrent Chinese Restaurant Process



$T=1$



$T=2$

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

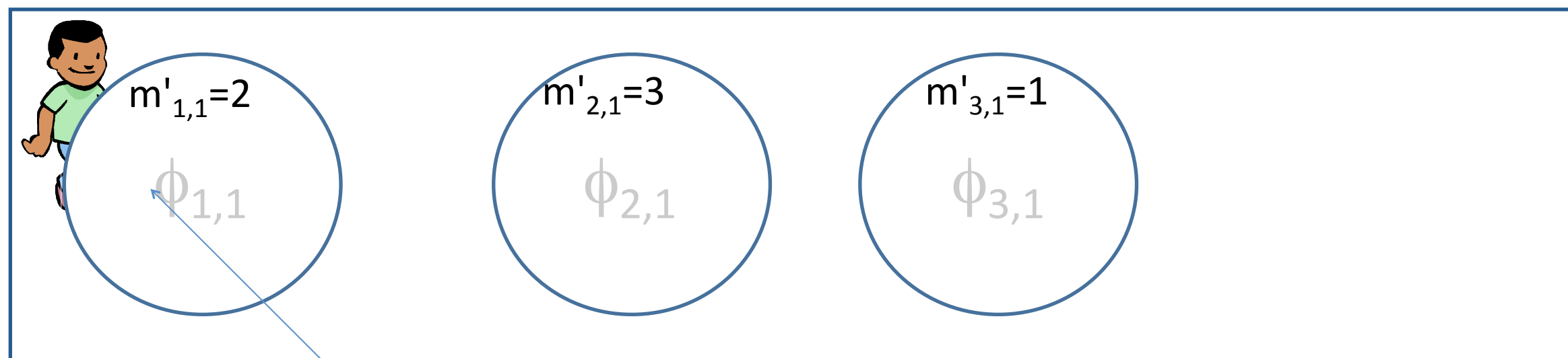
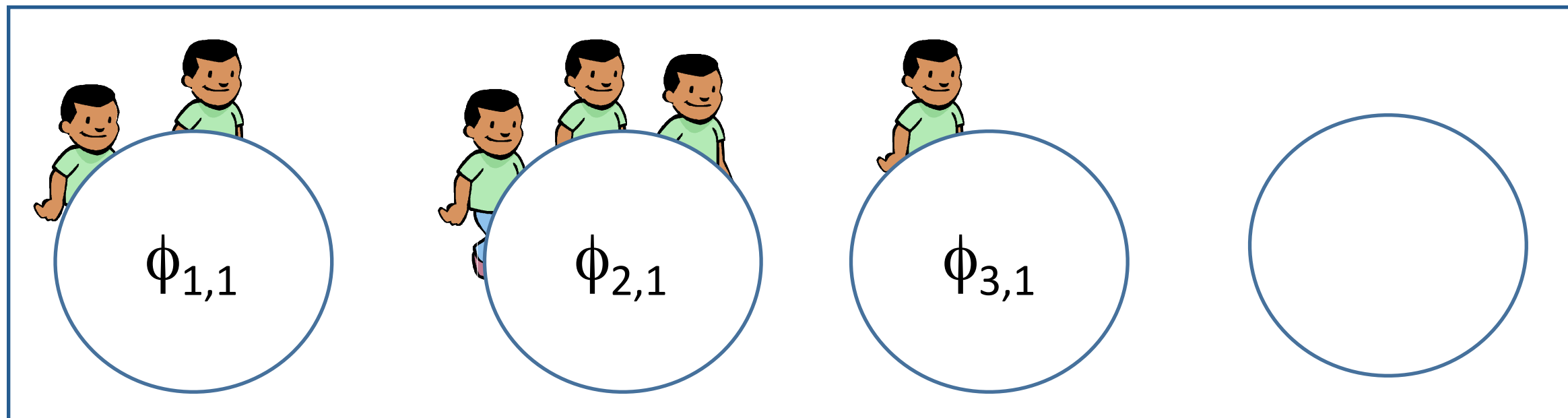
$$\frac{3}{6+\alpha}$$

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

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



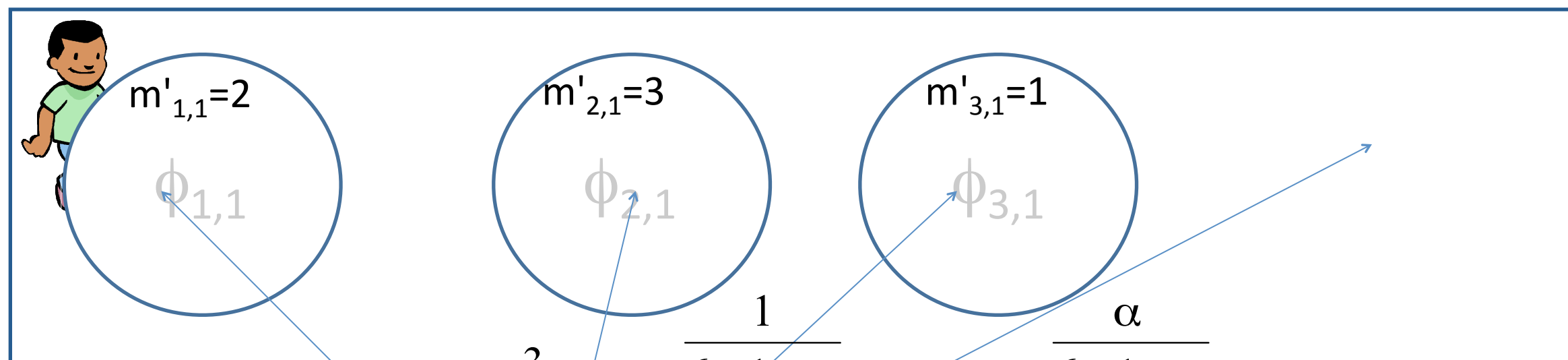
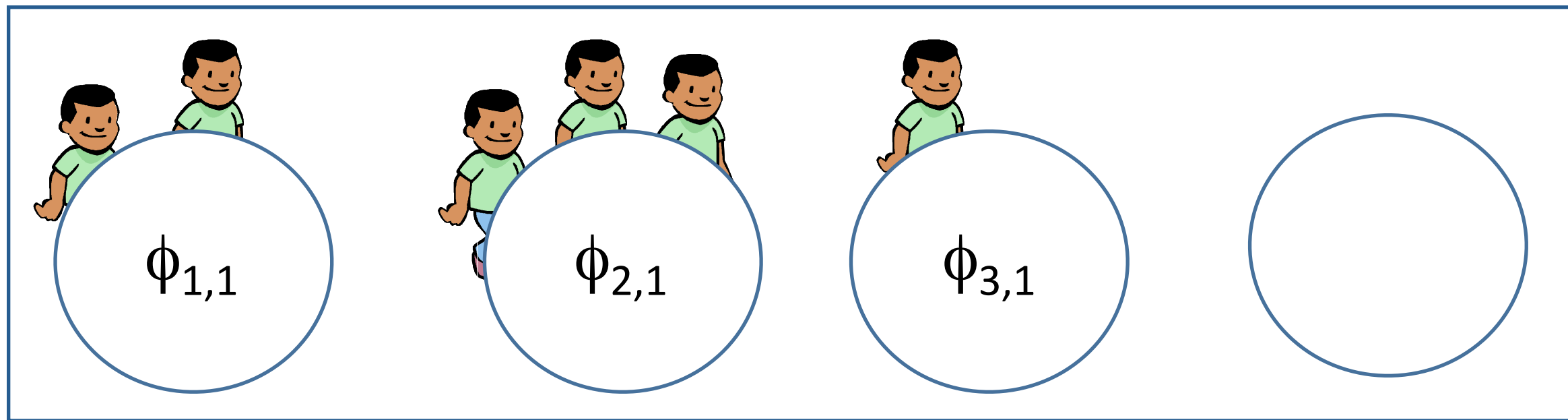
# Recurrent Chinese Restaurant Process



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

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

# Recurrent Chinese Restaurant Process



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

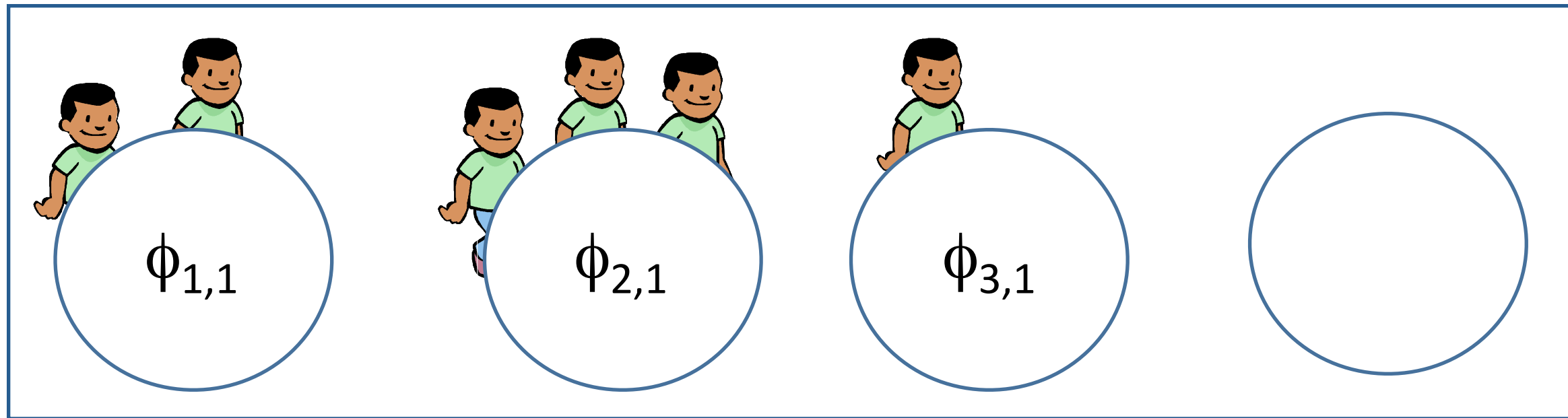
$$\frac{3}{6+1+\alpha}$$

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

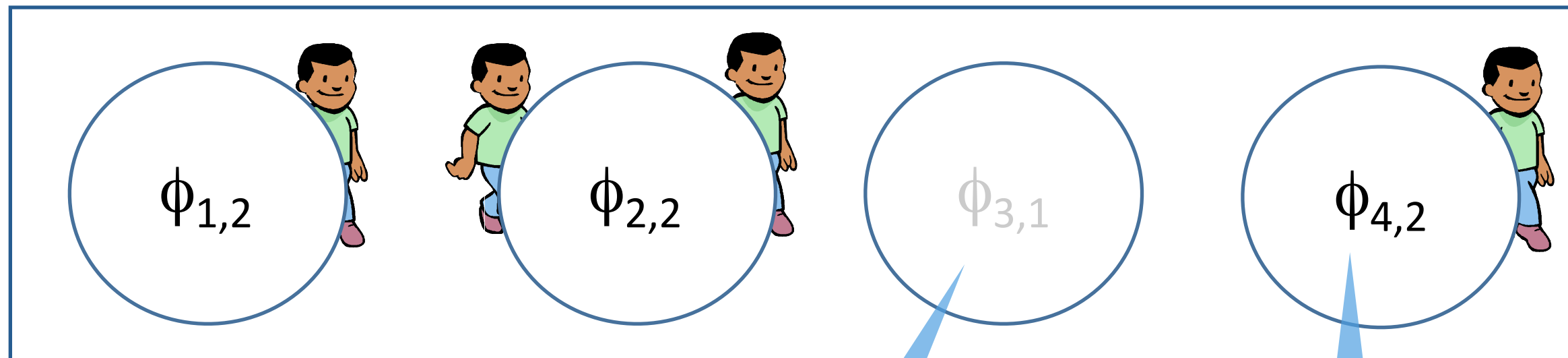
$$\frac{\alpha}{6+1+\alpha}$$



# Recurrent Chinese Restaurant Process



$T=1$

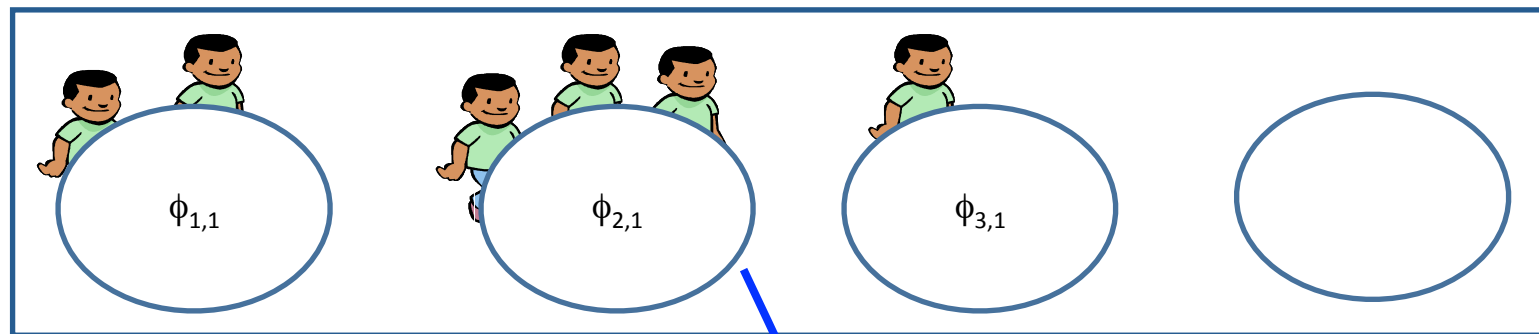


$T=2$

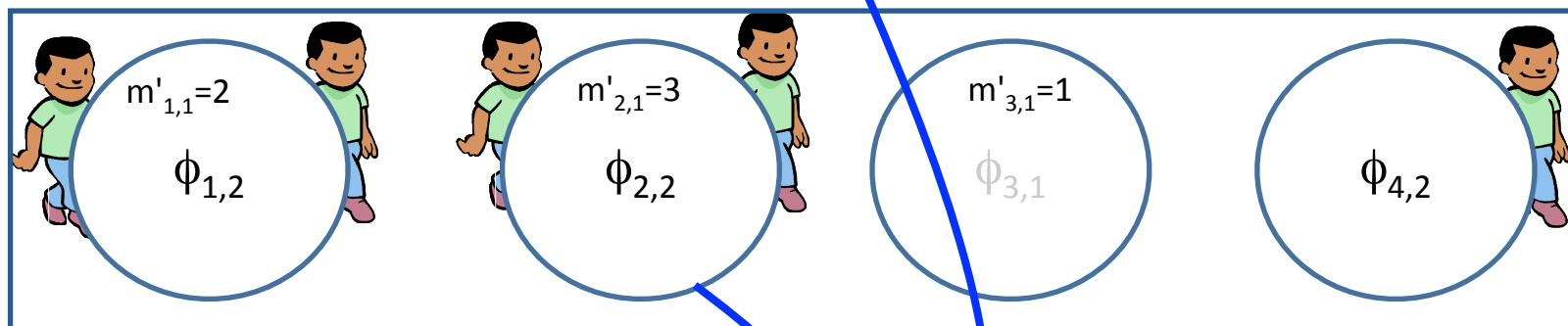
dead cluster

new cluster

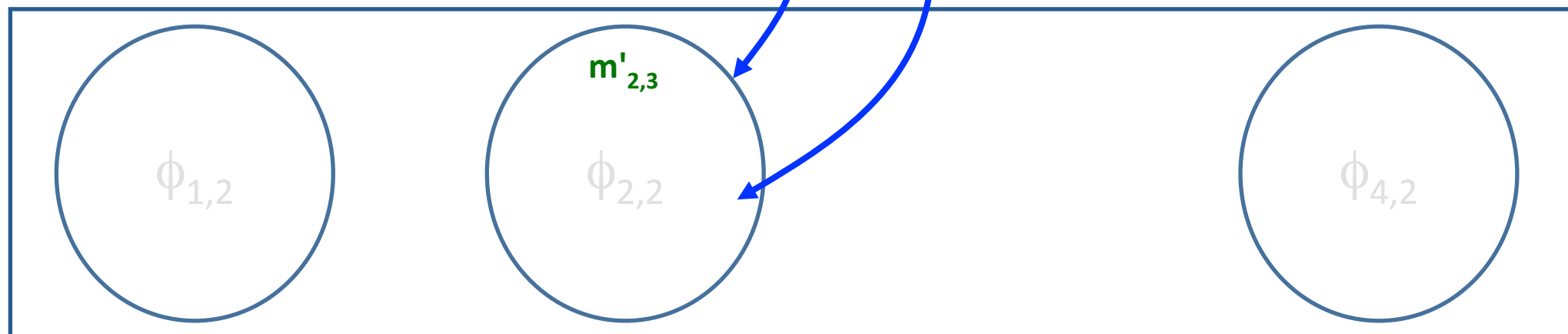
# Longer History



T=1



T=2



T=3

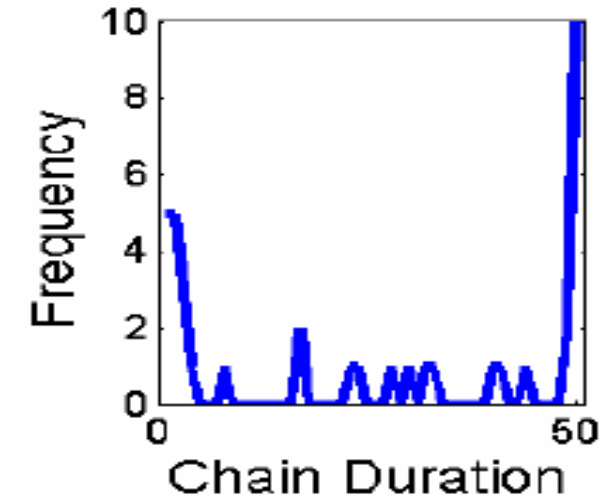
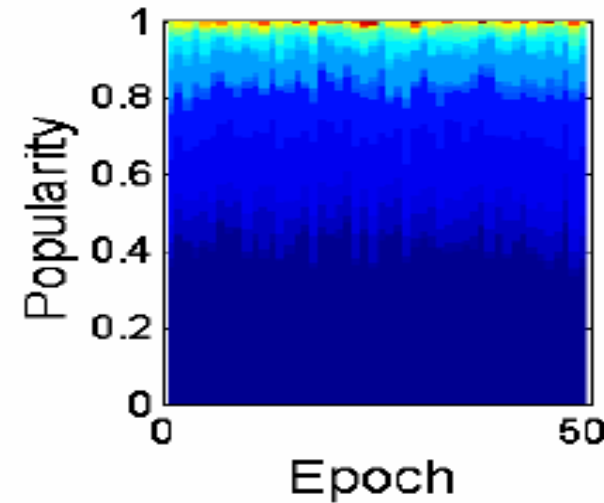
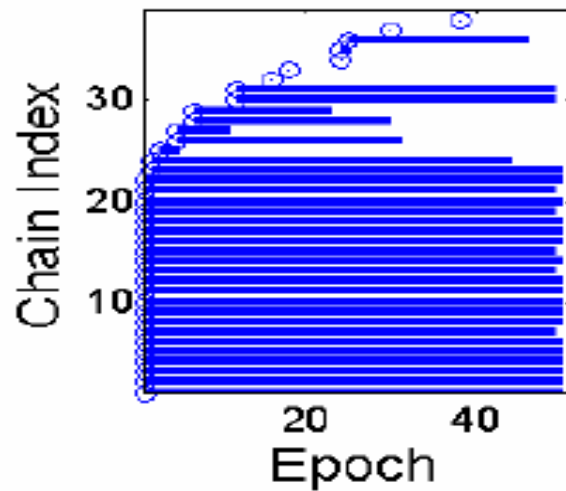


# TDPM Generative Power

DPM

$$W = \infty$$

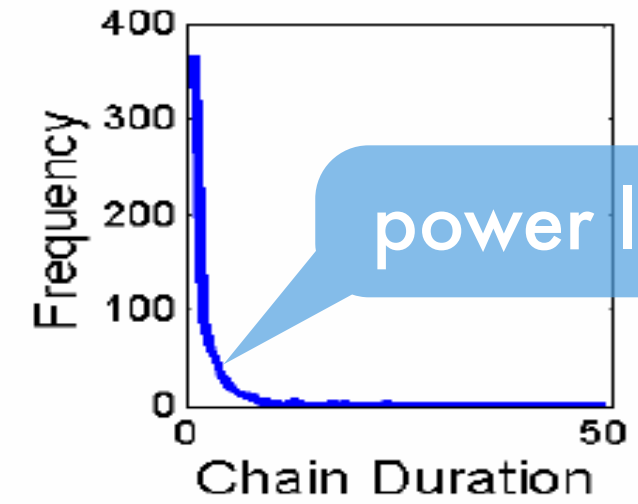
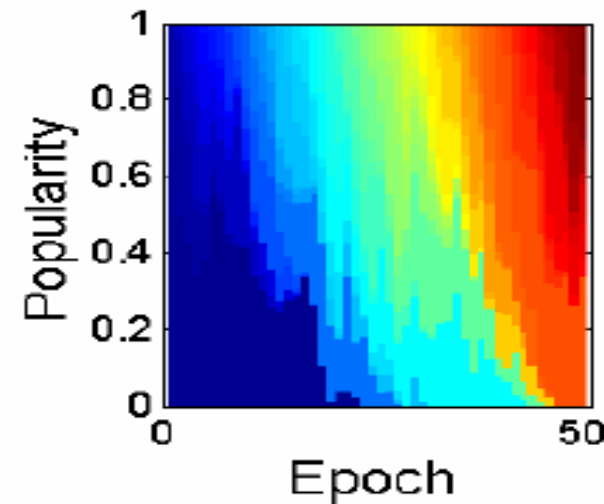
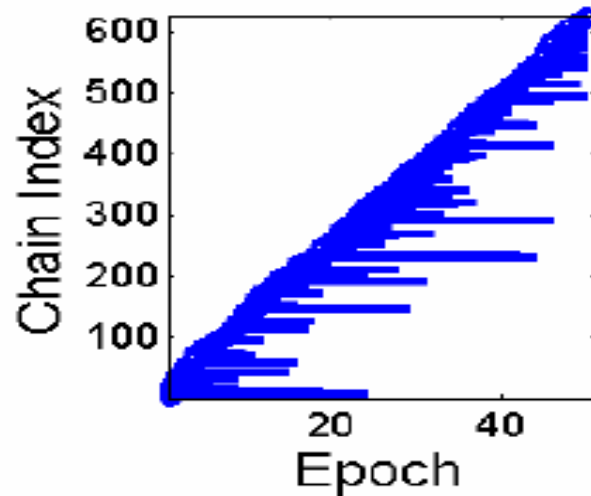
$$\lambda = \infty$$



TDPM

$$W = 4$$

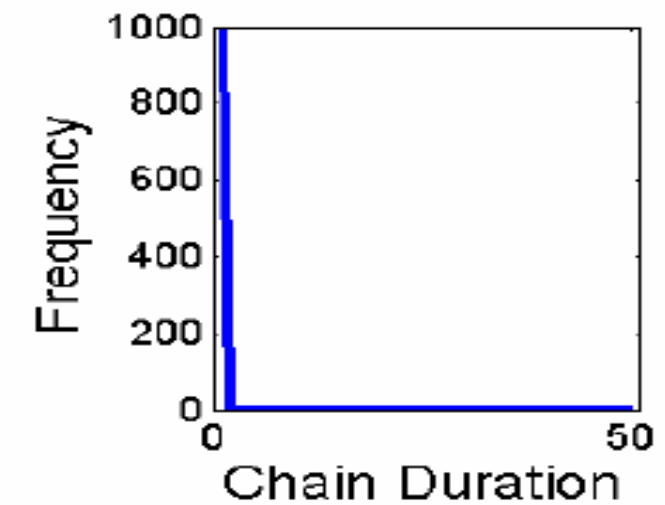
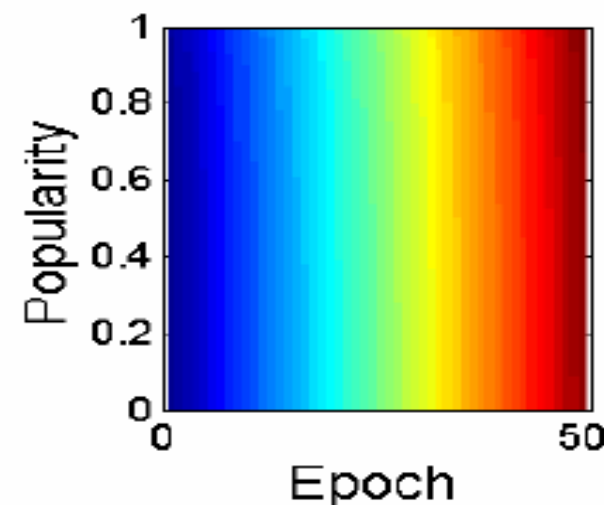
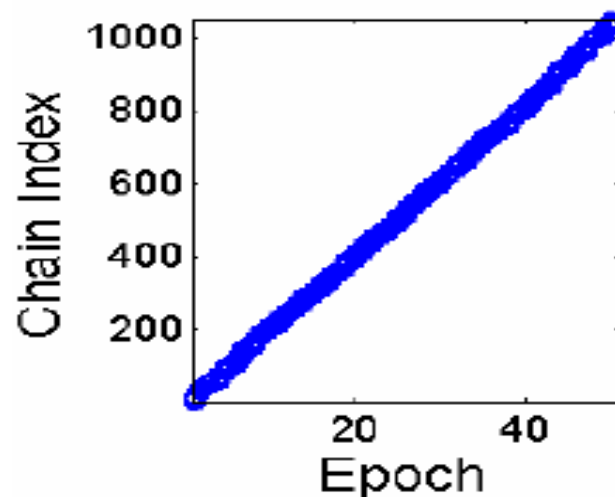
$$\lambda = .4$$



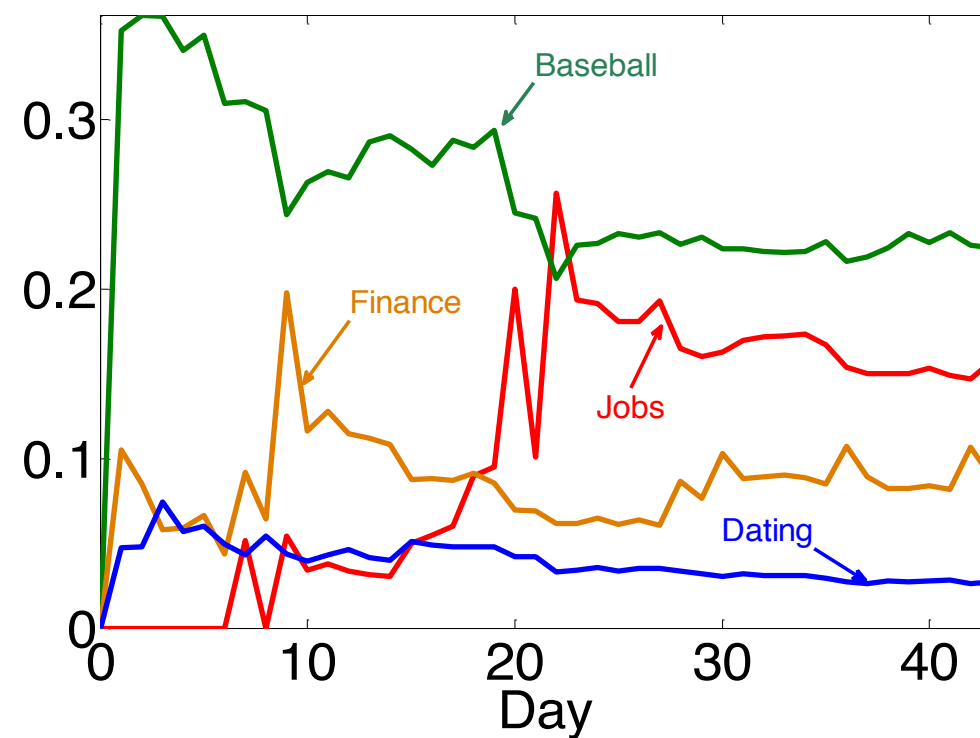
Independent DPMs

$$W = 0$$

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



# User modeling



# Buying a camera



time

# Buying a camera

**YAHOO!** Web Images Video Local Shopping News More ▾

panasonic lx5

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



 Sponsor Results

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

### Panasonic LX5 Cheap

Best Value for **Panasonic LX5**. Find NexTag Sellers  
[www.NexTag.com](http://www.NexTag.com)

### Panasonic Lumix DMC-LX5 Review (white

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

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

### Panasonic LX5 | Get The Lowest Price On

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

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

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

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

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

Sponsored Results



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

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

Shop All Departments ▾

Search

Camera & Photo

All Electronics

Brands

Bestsellers

Digital SLRs & Lenses

Point-And-Shoots

Camcorders

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

Color: Black

**Prime**

Member: Alexander Smola

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

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

by [Panasonic](#)

★★★★☆  (40 customer reviews)

List Price: ~~\$499.00~~

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

**Amazon Prime**

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



[new](#)

time

# Buying a camera

**YAHOO!** Web Images Video Local Shopping News More ▾

panasonic lx5

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



Also try: [panasonic lz5](#), [more...](#)

### Panasonic LX5 Cheap

Best Value for **Panasonic LX5**. Find NexTag Sellers  
[www.NexTag.com](http://www.NexTag.com)

### Panasonic Lumix DMC-LX5 Review (white

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

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

### Panasonic LX5 | Get The Lowest Price On

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

### Panasonic Lumix DMC-LX5 White Digital (

[shopping.yahoo.com](http://shopping.yahoo.com)

Sponsor Results

Sponsored Results



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

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

Shop All Departments ▾

Search

Camera & Photo

All Electronics

Brands

Bestsellers

Digital SLRs & Lenses

Point-And-Shoots

Camcorders

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

Color: **Black**

**Prime**

Member: Alexander Smola

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

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

by [Panasonic](#)

★★★★☆  (40 customer reviews)

List Price: ~~\$499.00~~

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

**Amazon Prime**

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



[new](#)

show ads now

time



# Buying a camera

**YAHOO!** Web Images Video Local Shopping News More ▾  
panasonic lx5  
Search In:  the Web  pages in English, French, German, Italian and Spanish



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

### Panasonic LX5 Cheap

Best Value for **Panasonic LX5**. Find NexTag Sellers  
[www.NexTag.com](http://www.NexTag.com)

### Panasonic Lumix DMC-LX5 Review (white

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

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

### Panasonic LX5 | Get The Lowest Price On

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

### Panasonic Lumix DMC-LX5 White Digital (

[shopping.yahoo.com](http://shopping.yahoo.com)

Sponsor Results

Sponsored Results



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

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

Shop All Departments ▾

Search

Camera & Photo

All Electronics

Brands

Bestsellers

Digital SLRs & Lenses

Point-And-Shoots

Camcorders

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

Color: Black

**Prime**

Member: Alexander Smola

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

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

by [Panasonic](#)

★★★★☆ (40 customer reviews)

List Price: ~~\$499.00~~

Price: **\$444.95** & eligible for

**Amazon Prime**

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



show ads now

too late

time

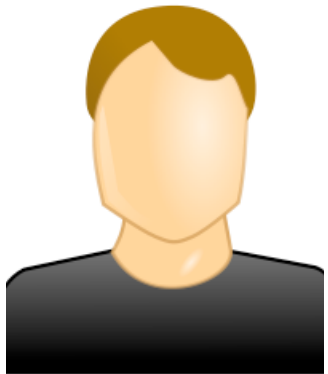






Car  
Deals  
van

---



job  
Hiring  
diet

---



---

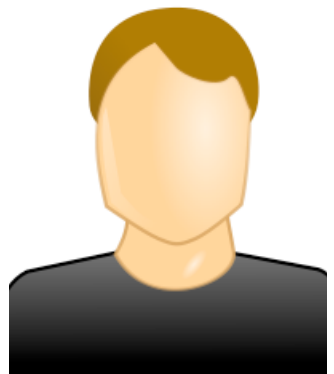


Car  
Deals  
van

Auto  
Price  
Used  
inspection

Movies  
Theatre  
Art  
gallery

---



job  
Hiring  
diet

Hiring  
Salary  
Diet  
calories

Diet  
Calories  
Recipe  
chocolate

---

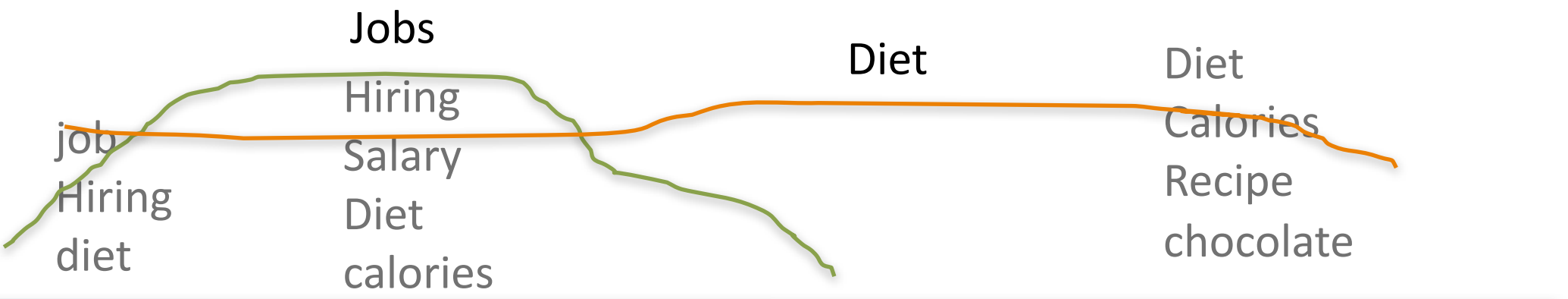
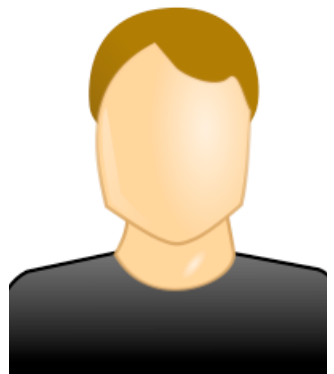
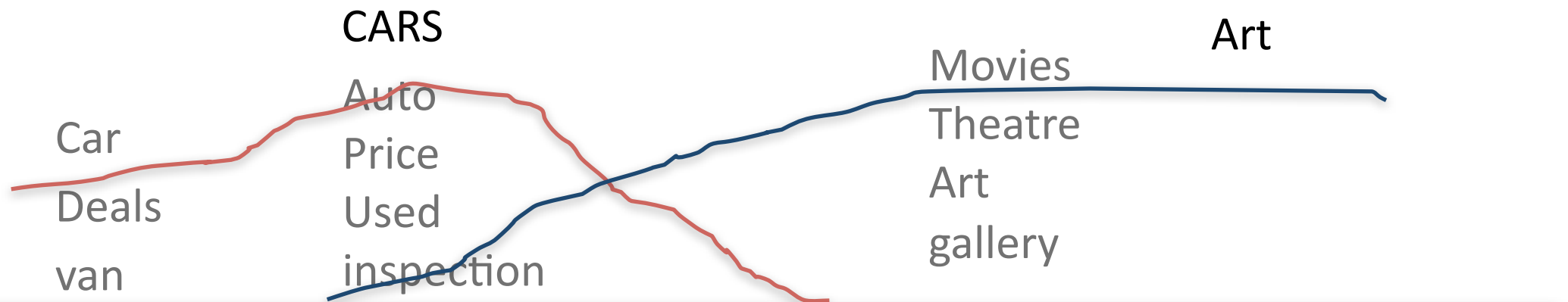


Flight  
London  
Hotel  
weather

School  
Supplies  
Loan  
college

---





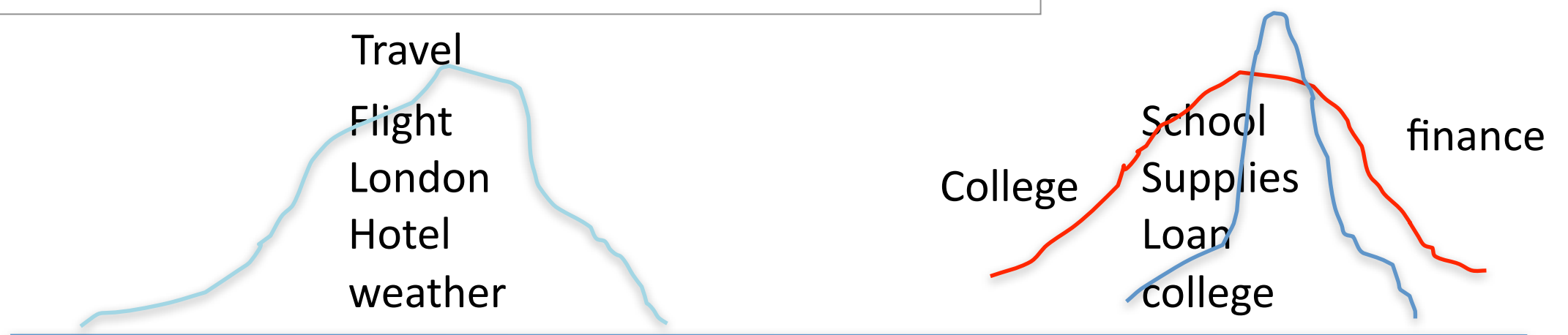
# User modeling

## Input

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

## Output

- Users' daily distribution over intents
- Dynamic intent representation

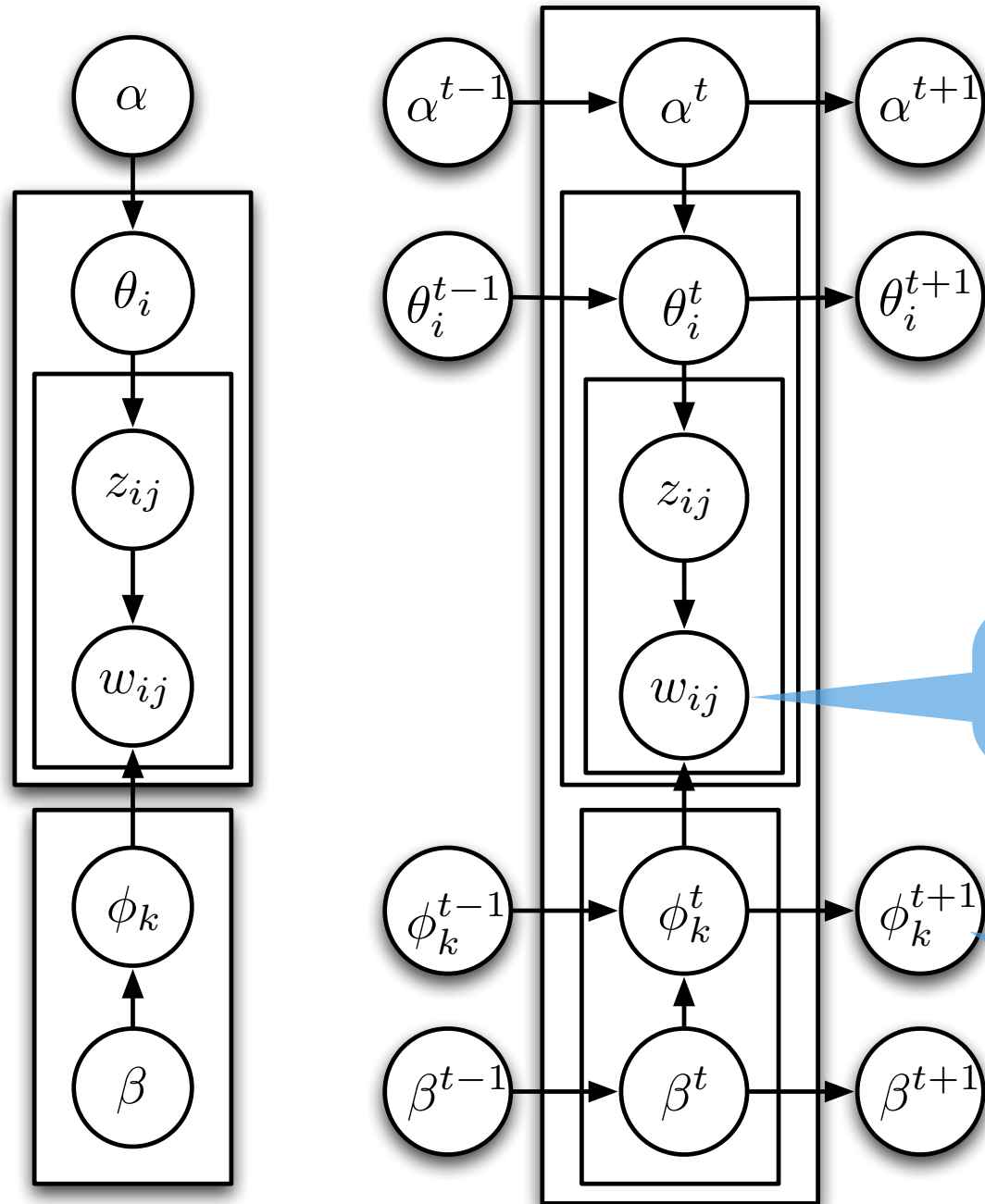


# Time dependent models

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

# Graphical Model

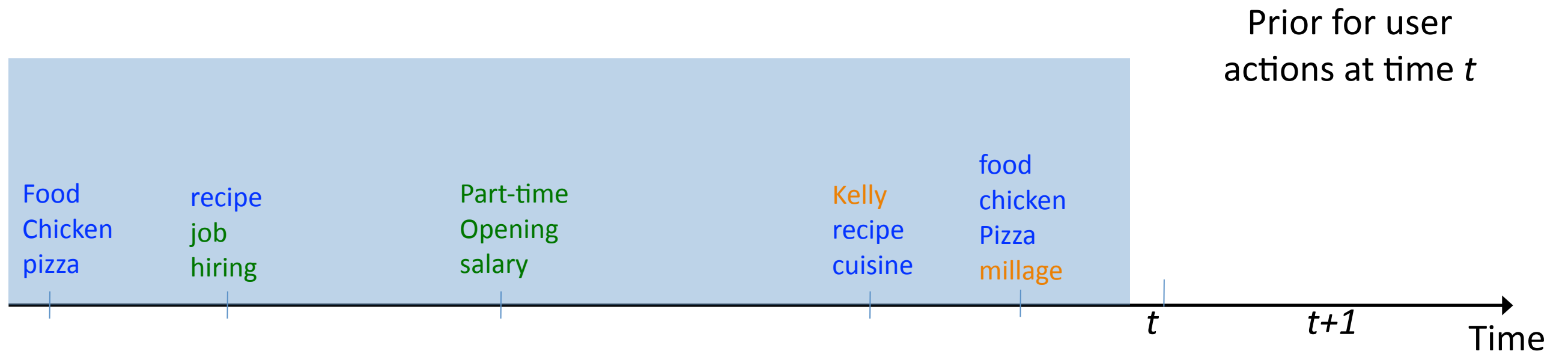
plain  
LDA



time dependent  
user interest

user actions

actions per topic



### Diet

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

### Cars

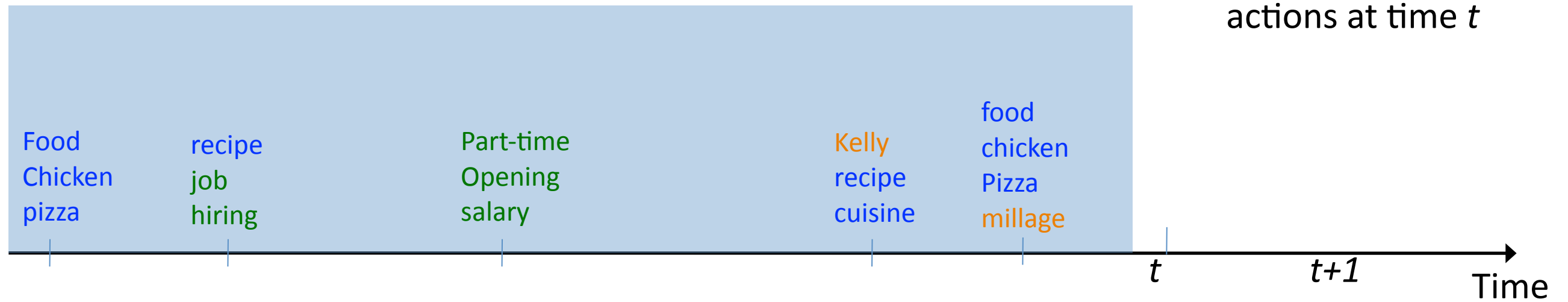
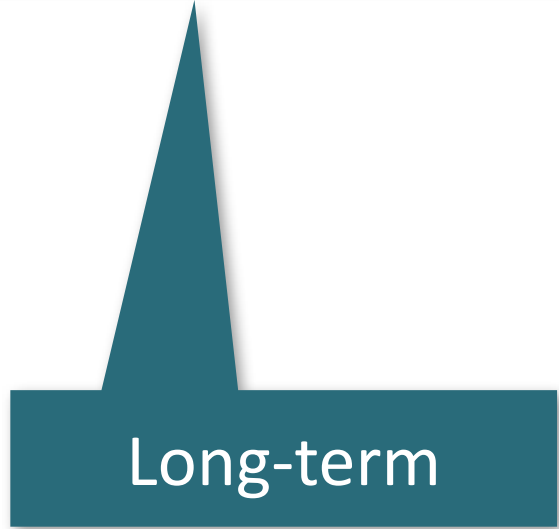
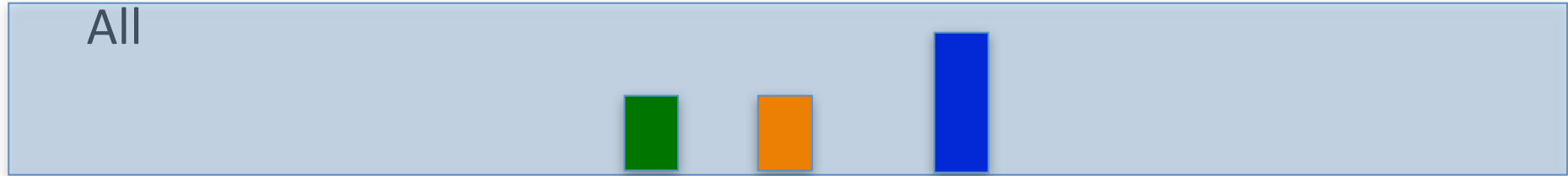
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

### Job

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

### Finance

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



Diet

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

Cars

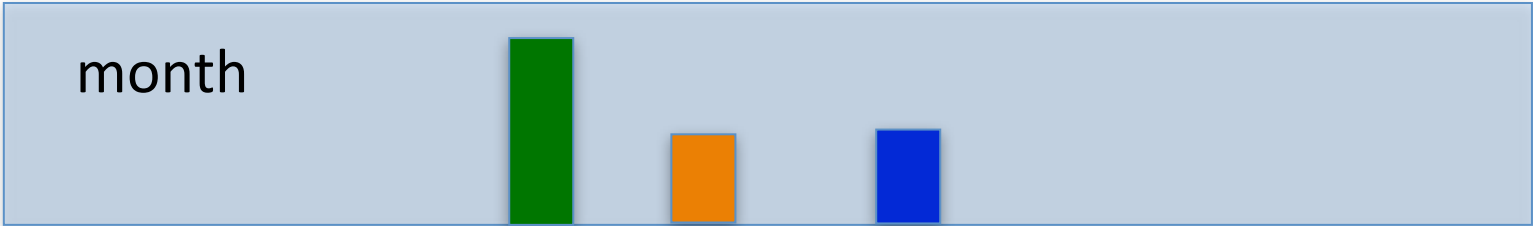
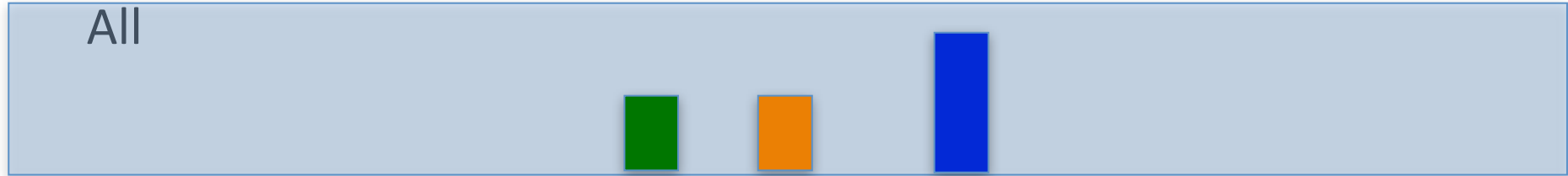
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

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

Finance

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



Long-term



Prior for user actions at time  $t$

$t$   $t+1$  Time

Diet

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

Cars

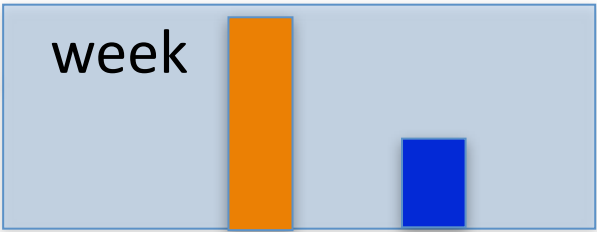
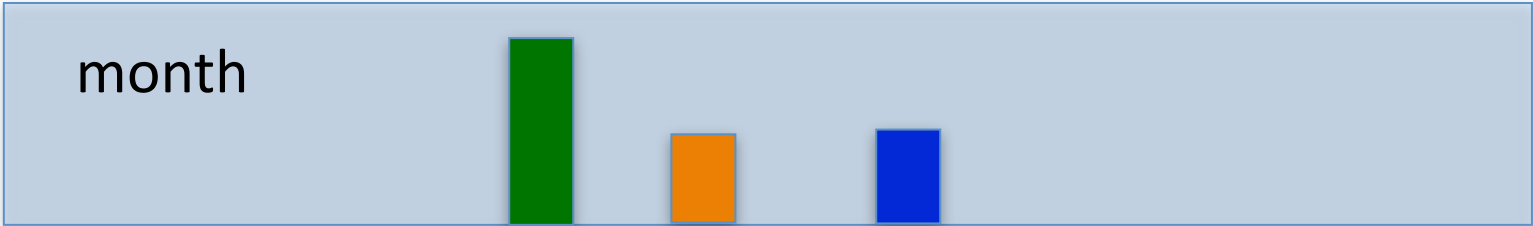
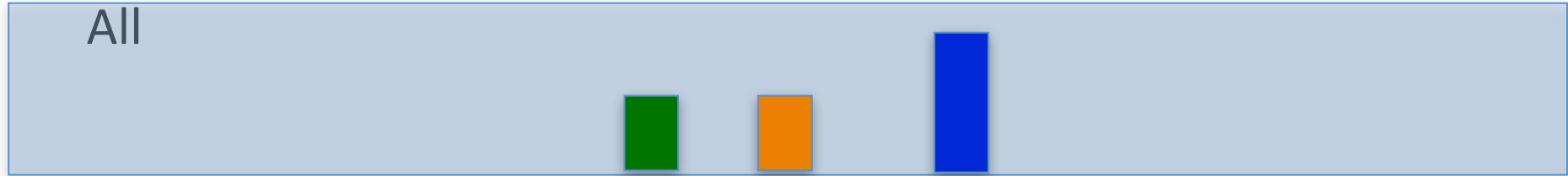
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

Job

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

Finance

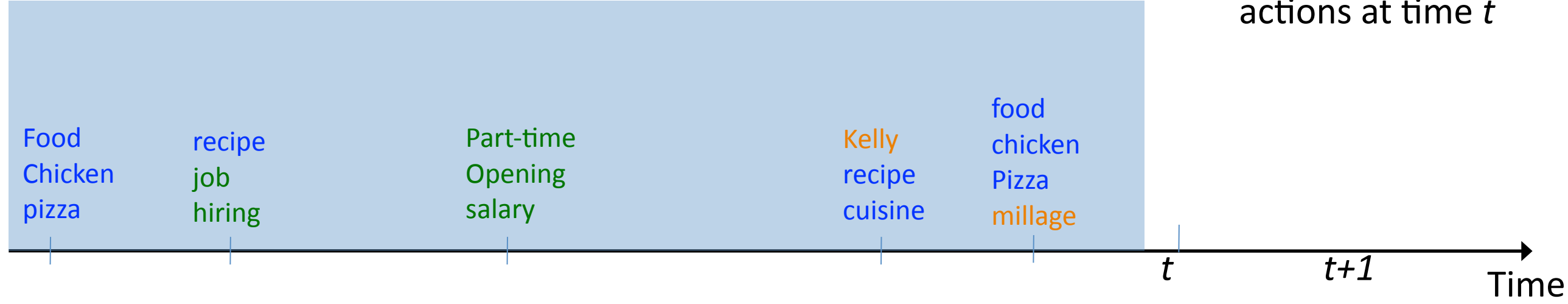
- Bank
- Online
- Credit
- Card
- debt
- portfolio
- Finance
- Chase



Long-term

short-term

Prior for user actions at time  $t$



Diet

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

Cars

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

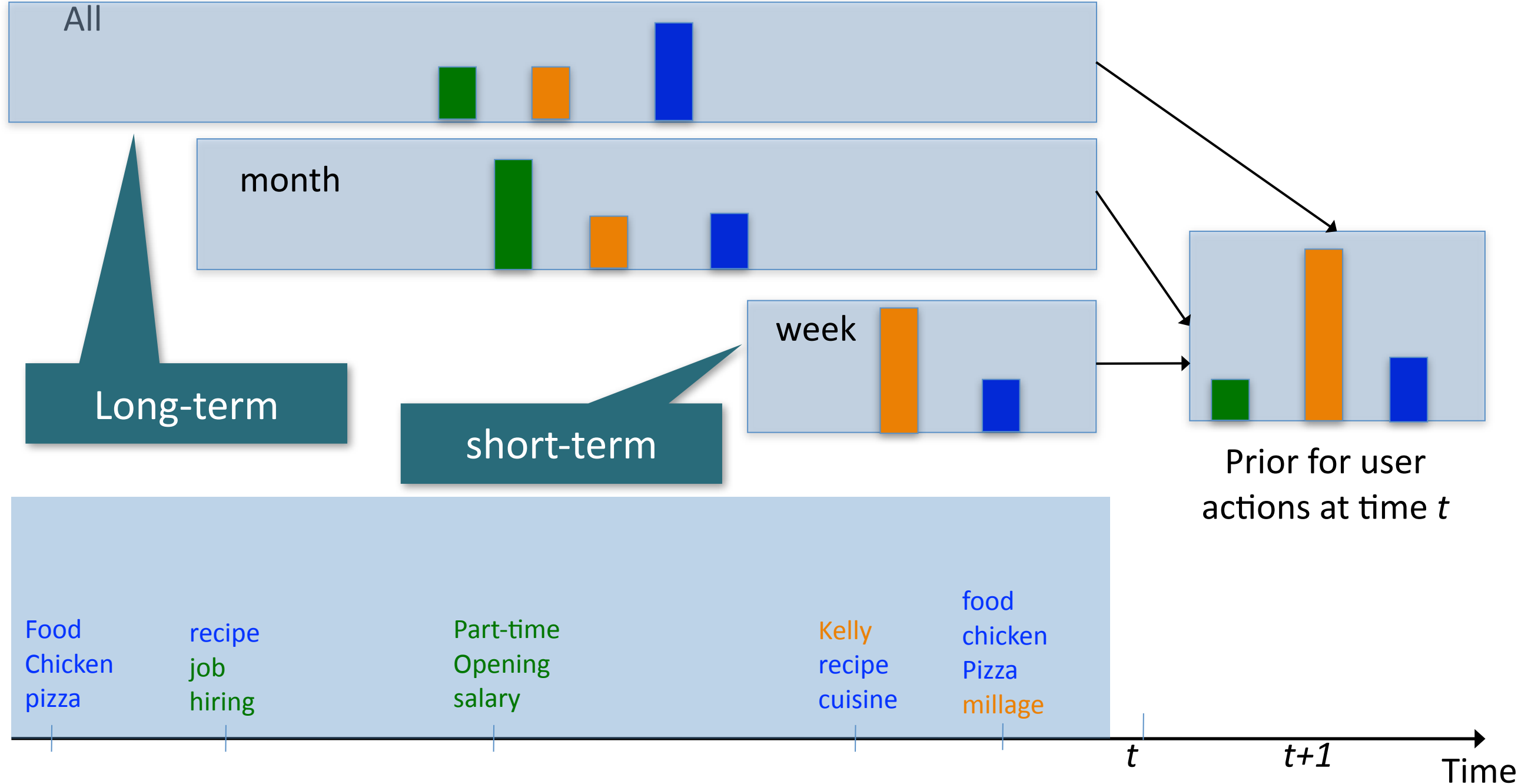
Job

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

Finance

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





### Diet

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

### Cars

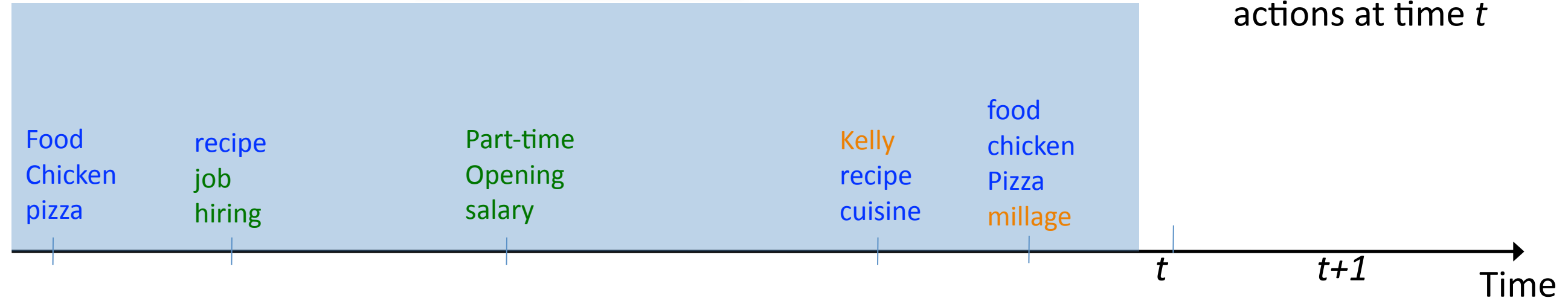
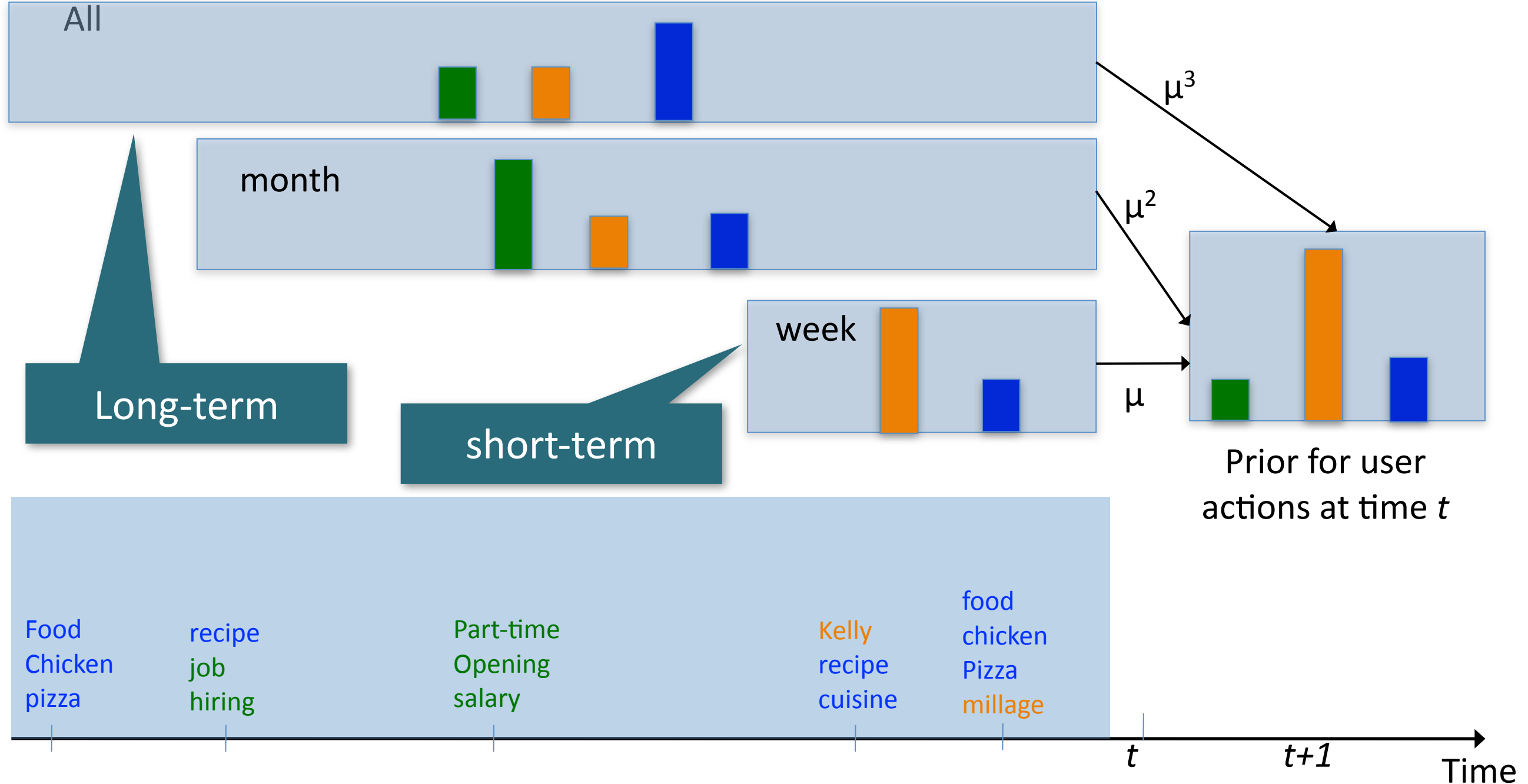
- Car
- Blue
- Book
- Kelley
- Prices
- Small
- Speed
- large

### Job

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

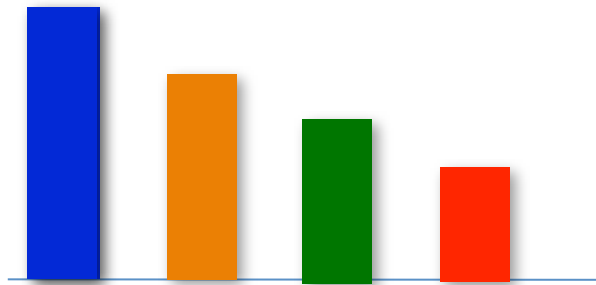
### Finance

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

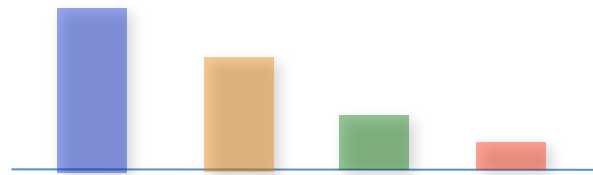


- | Diet      | Cars   | Job          | Finance   |
|-----------|--------|--------------|-----------|
| Recipe    | Car    | job          | Bank      |
| Chocolate | Blue   | Career       | Online    |
| Pizza     | Book   | Business     | Credit    |
| Food      | Kelley | Assistant    | Card      |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

### At time t



### At time t+1



Recipe  
Chocolate  
Pizza  
Food  
Chicken  
Milk  
Butter  
Powder

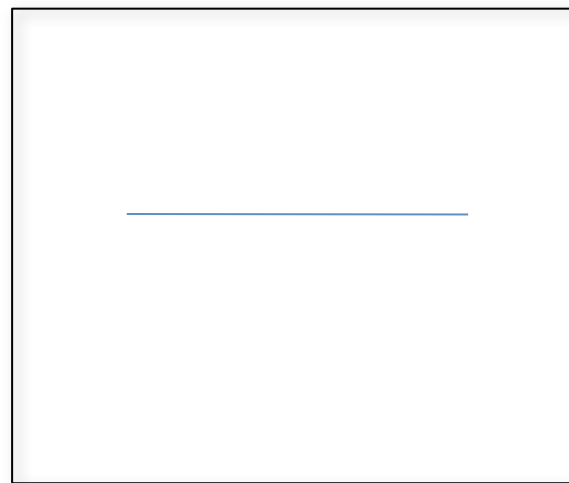
Car  
Altima  
Accord  
Blue  
Book  
Kelley  
Prices  
Small  
Speed

job  
Career  
Business  
Assistant  
Hiring  
Part-time  
Receptioni  
st

Bank  
Online  
Credit  
Card  
debt  
portfolio  
Finance  
Chase

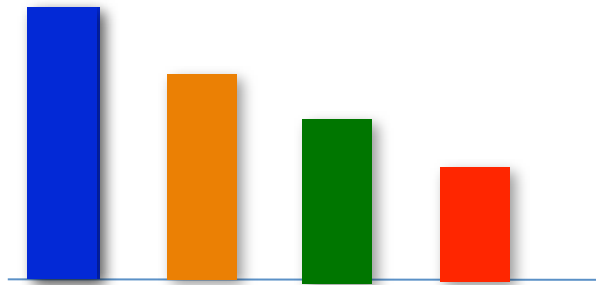


Food Chicken  
Pizza mileage

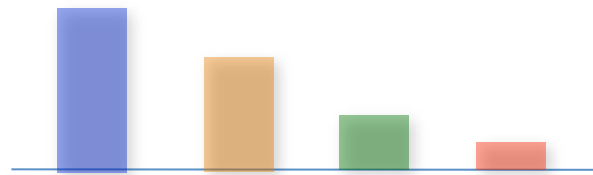


Car speed offer  
Camry accord career

### At time t



### At time t+1



Recipe  
Chocolate  
Pizza  
Food  
Chicken  
Milk  
Butter  
Powder

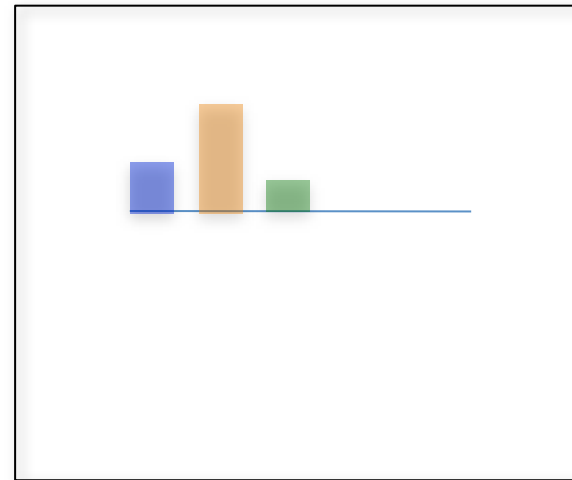
Car  
Altima  
Accord  
Blue  
Book  
Kelley  
Prices  
Small  
Speed

job  
Career  
Business  
Assistant  
Hiring  
Part-time  
Receptioni  
st

Bank  
Online  
Credit  
Card  
debt  
portfolio  
Finance  
Chase

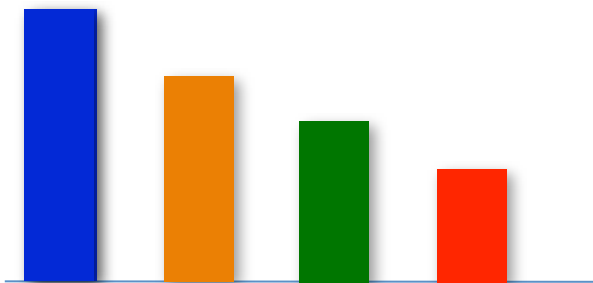


Food Chicken  
Pizza mileage

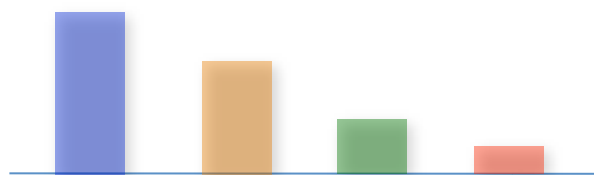


Car speed offer  
Camry accord career

### At time t



### At time t+1



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

- Car
- Altima
- Accord
- Blue
- Book
- Kelley
- Prices
- Small
- Speed

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

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



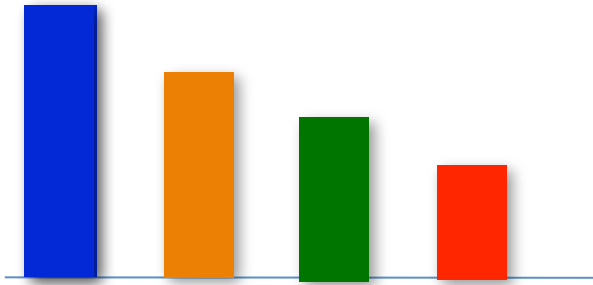
Food Chicken  
Pizza mileage

priors

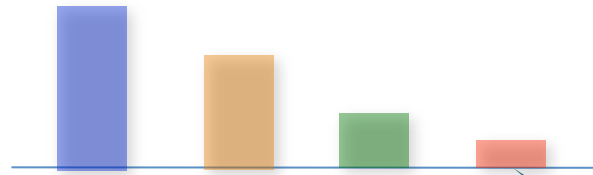


Car speed offer  
Camry accord career

At time t



At time t+1



Recipe  
Chocolate  
Pizza  
Food  
Chicken  
Milk  
Butter  
Powder

Car  
Altima  
Accord  
Blue  
Book  
Kelley  
Prices  
Small  
Speed

job  
Career  
Business  
Assistant  
Hiring  
Part-time  
Receptioni  
st

Bank  
Online  
Credit  
Card  
debt  
portfolio  
Finance  
Chase



Food Chicken  
Pizza mileage

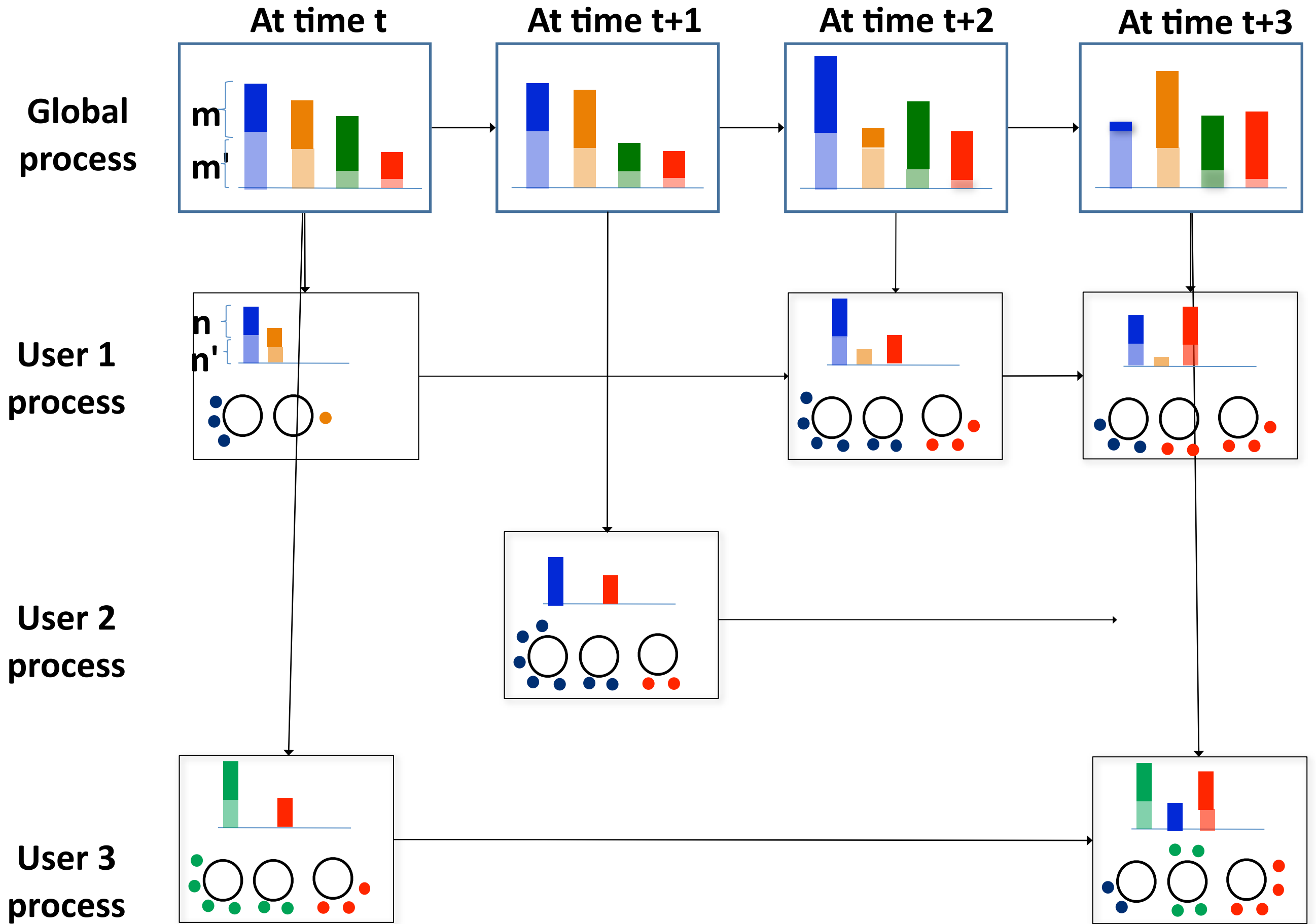
priors



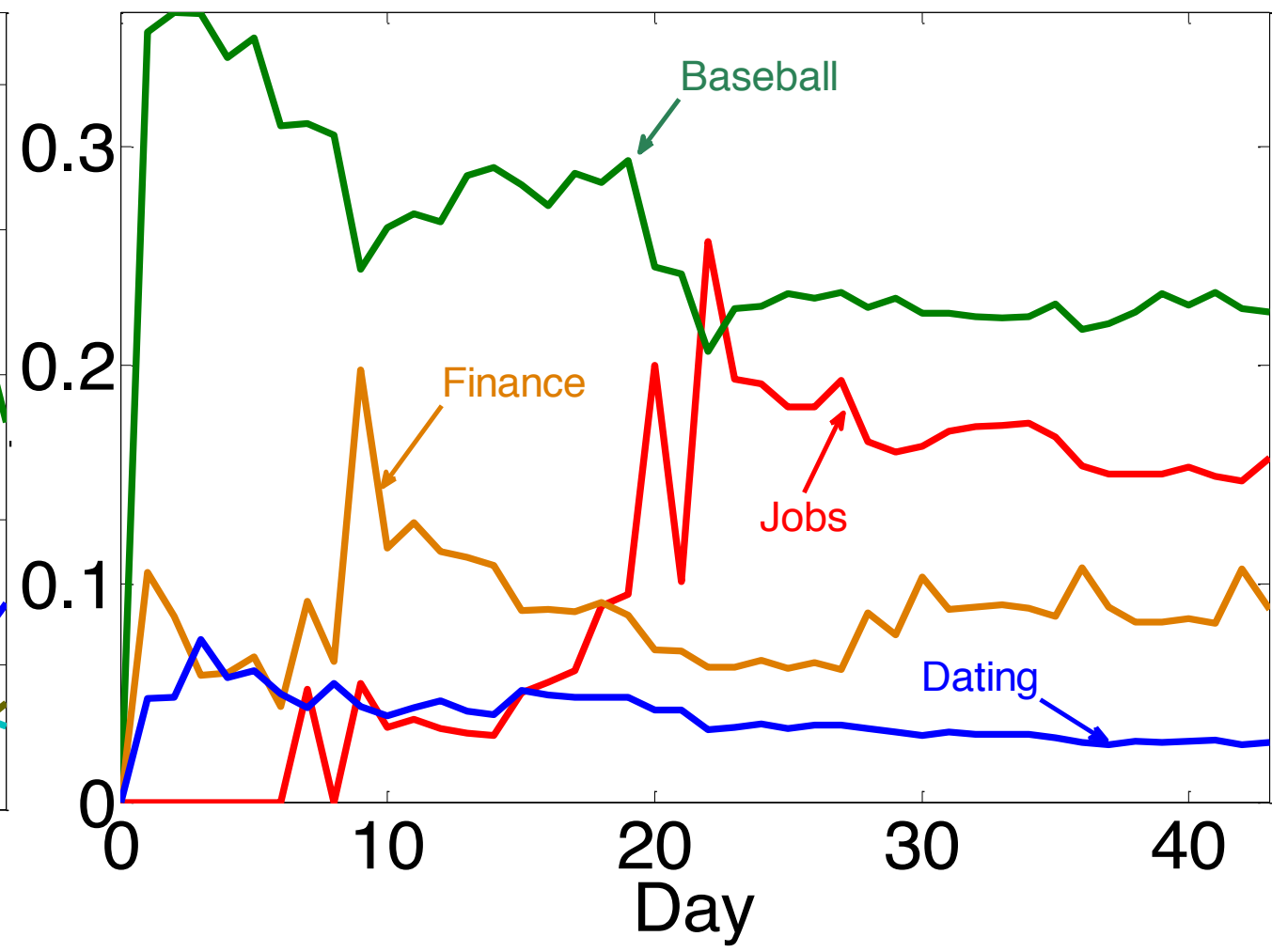
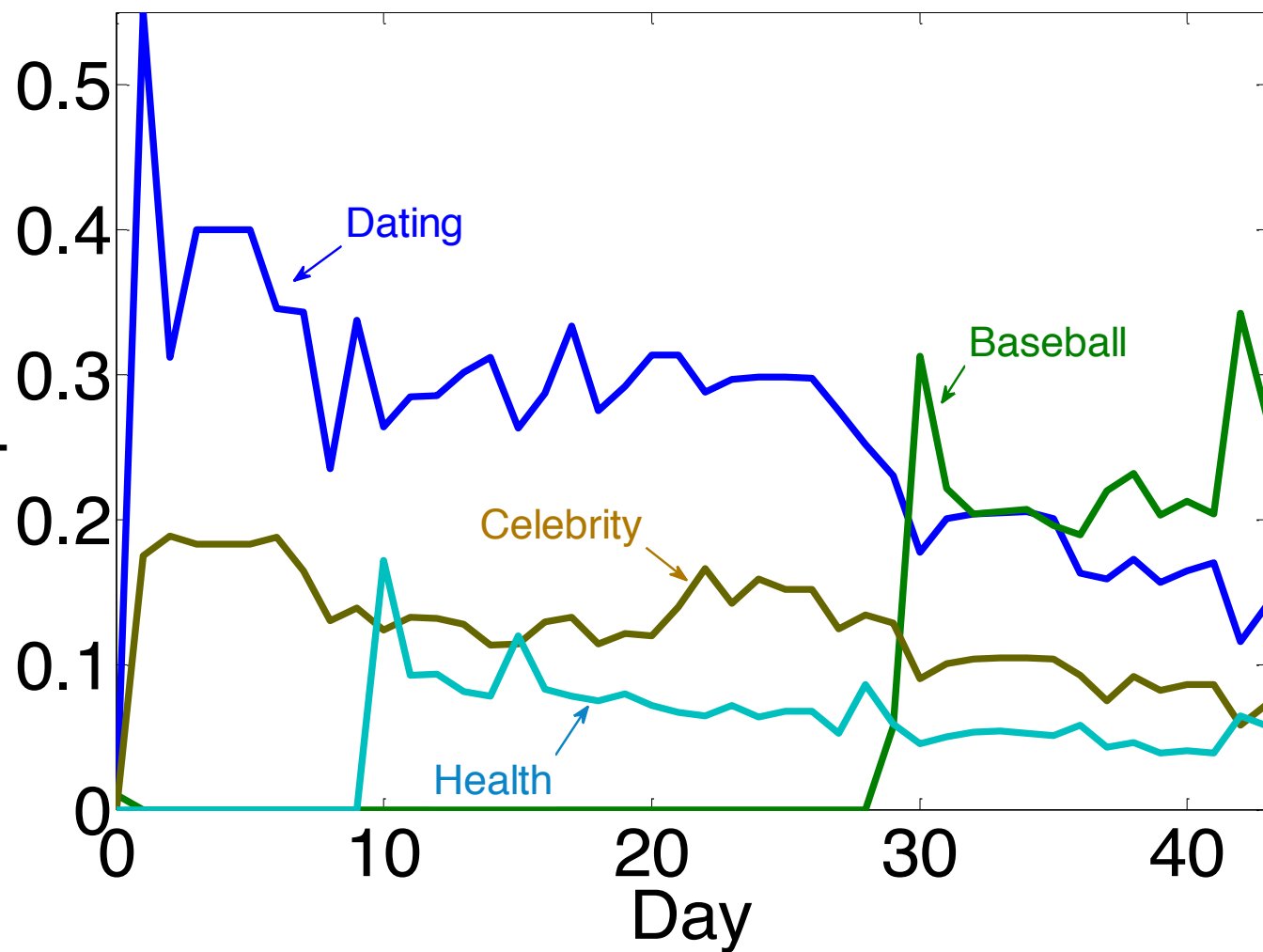
Car speed offer  
Camry accord career

### Generative Process

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



# Sample users



## Dating

women  
men  
dating  
singles  
personals  
seeking  
match

## Baseball

League  
baseball  
basketball,  
doublehead  
Bergesen  
Griffey  
bullpen  
Greinke

## Celebrity

Snooki  
Tom  
Cruise  
Katie  
Holmes  
Pinkett  
Kudrow  
Hollywood

## Health

skin  
body  
fingers  
cells  
toes  
wrinkle  
layers

## Jobs

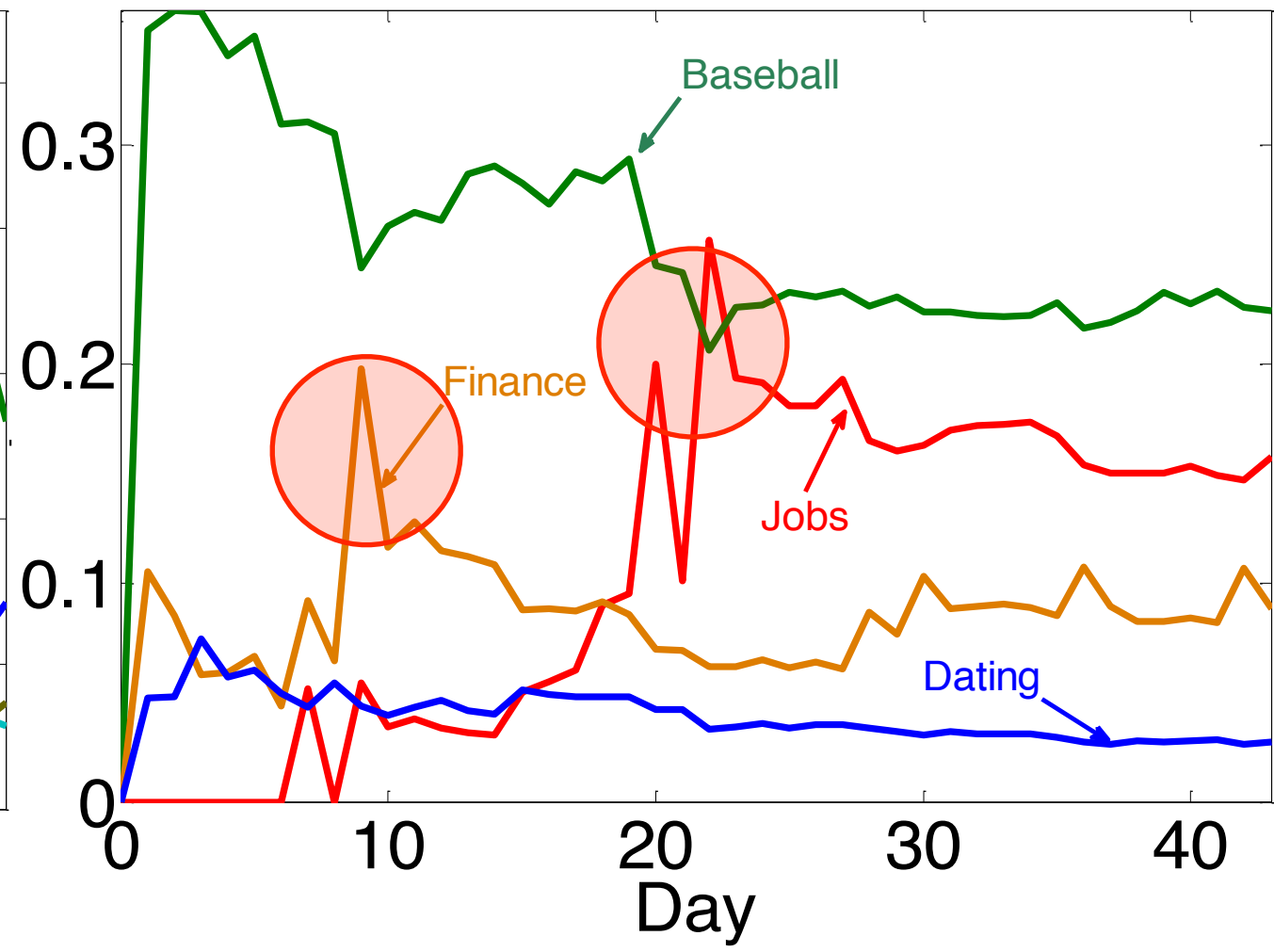
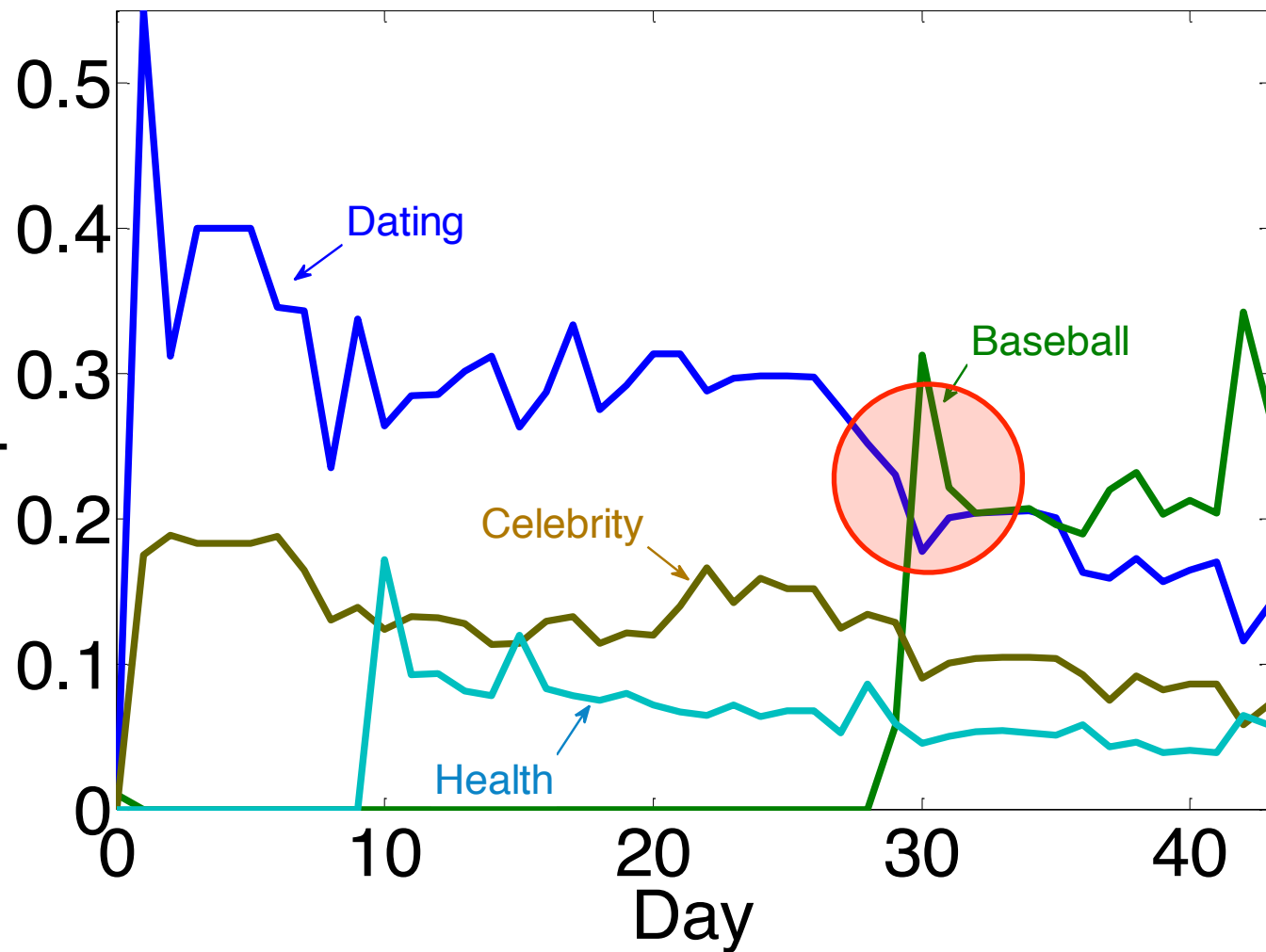
job  
career  
business  
assistant  
hiring  
part-time  
receptionist

## Finance

financial  
Thomson  
chart  
real  
Stock  
Trading  
currency



# Sample users



## Dating

women  
men  
dating  
singles  
personals  
seeking  
match

## Baseball

League  
baseball  
basketball,  
doublehead  
Bergesen  
Griffey  
bullpen  
Greinke

## Celebrity

Snooki  
Tom  
Cruise  
Katie  
Holmes  
Pinkett  
Kudrow  
Hollywood

## Health

skin  
body  
fingers  
cells  
toes  
wrinkle  
layers

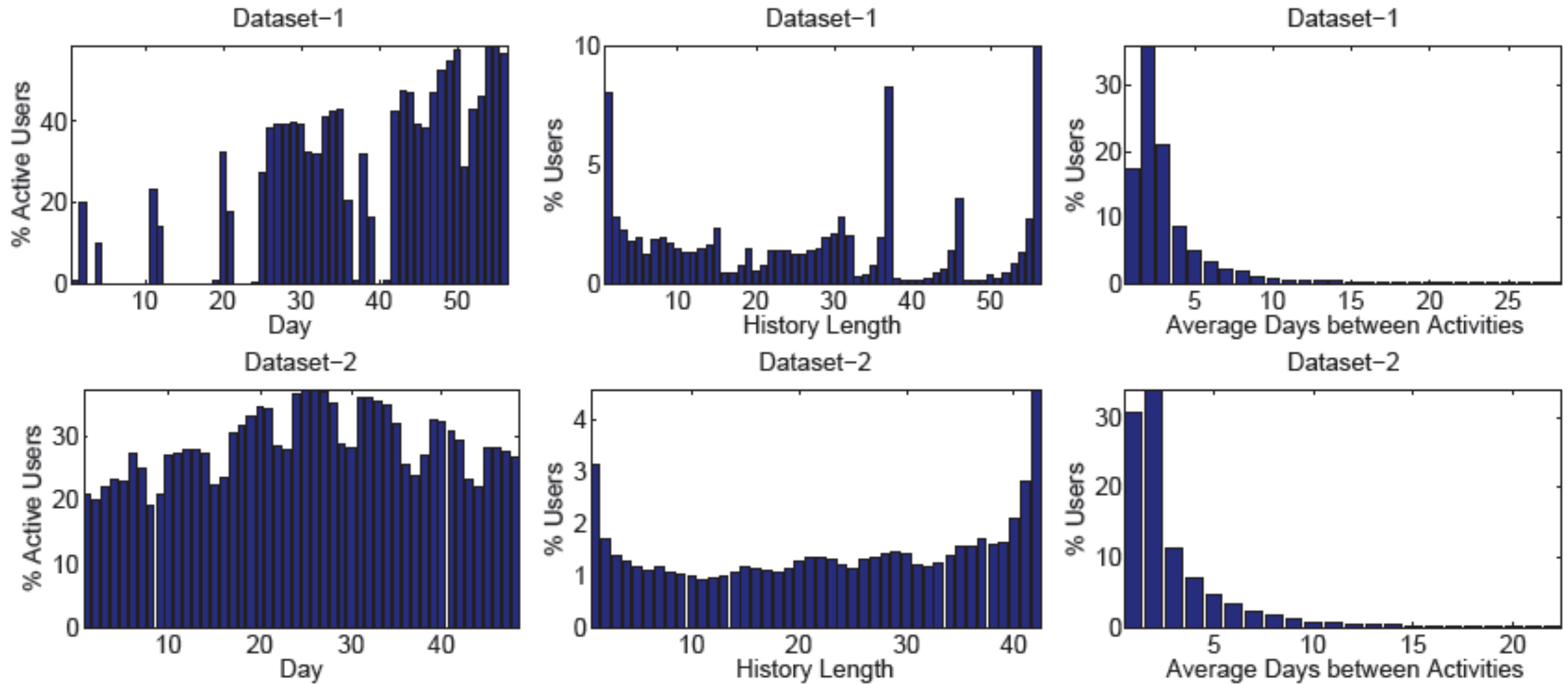
## Jobs

job  
career  
business  
assistant  
hiring  
part-time  
receptionist

## Finance

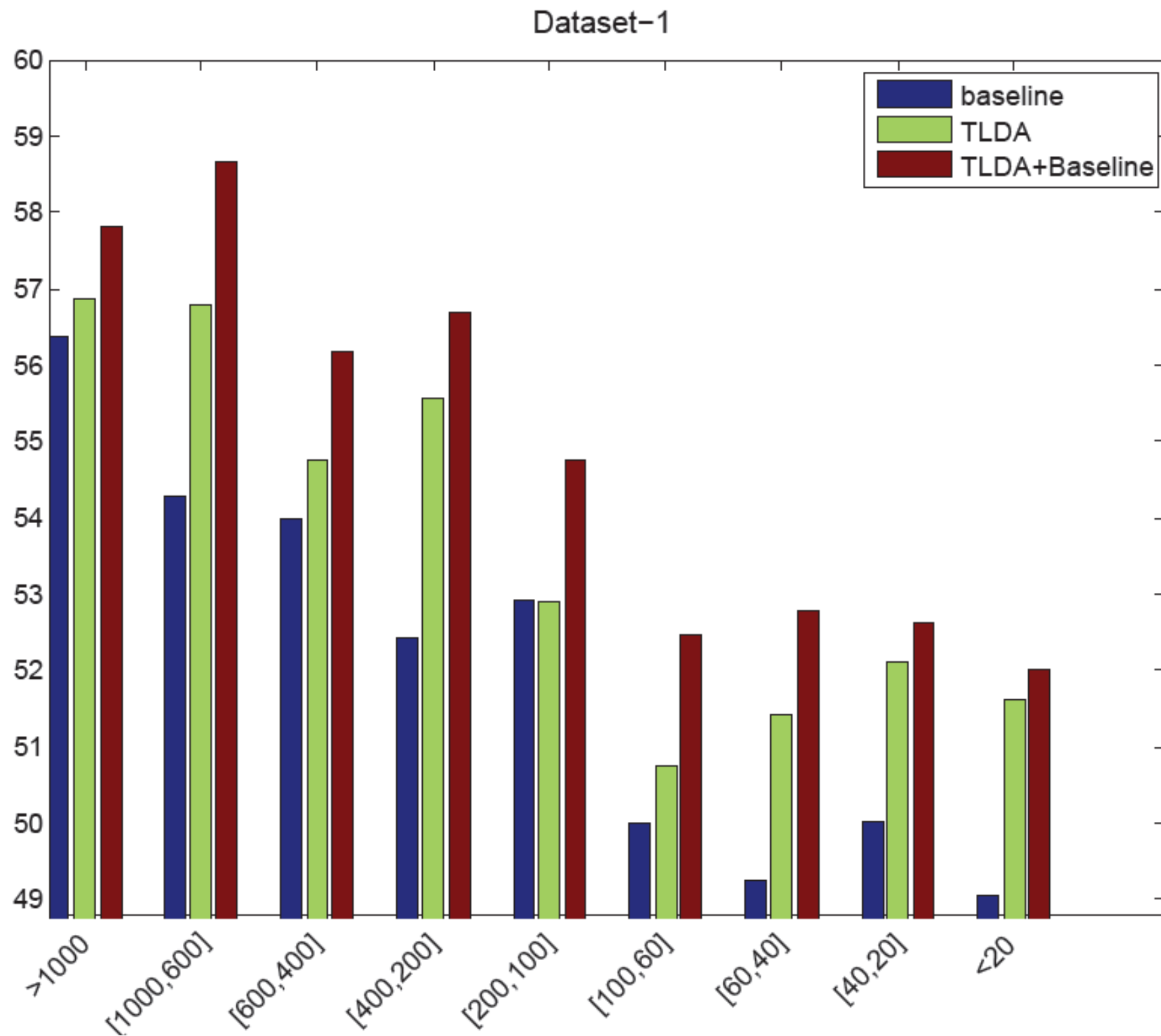
financial  
Thomson  
chart  
real  
Stock  
Trading  
currency

# Data



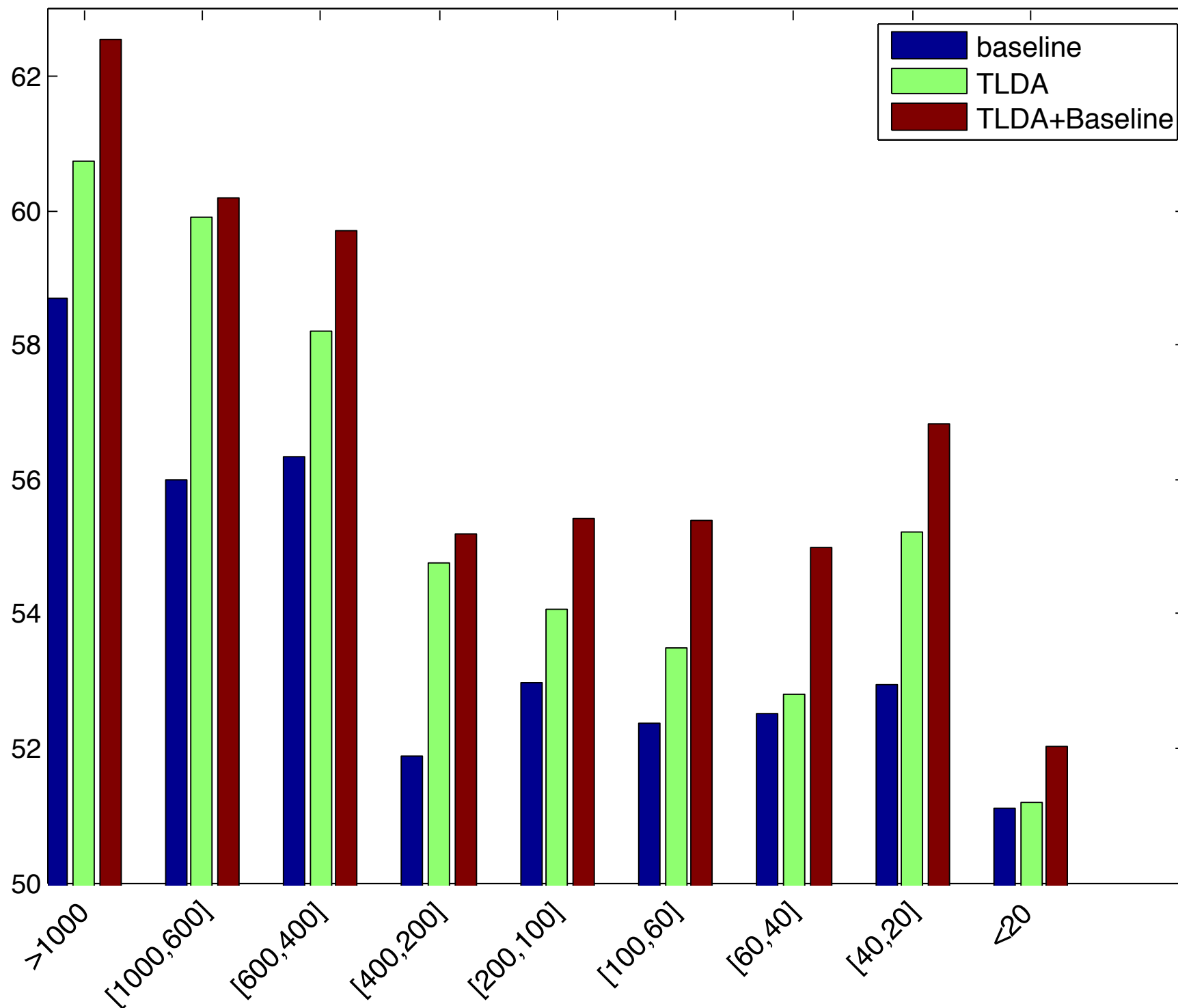
dataset	# days	# users	# campaigns	size
1	56	13.34M	241	242GB
2	44	33.5M	216	435GB

# ROC score improvement

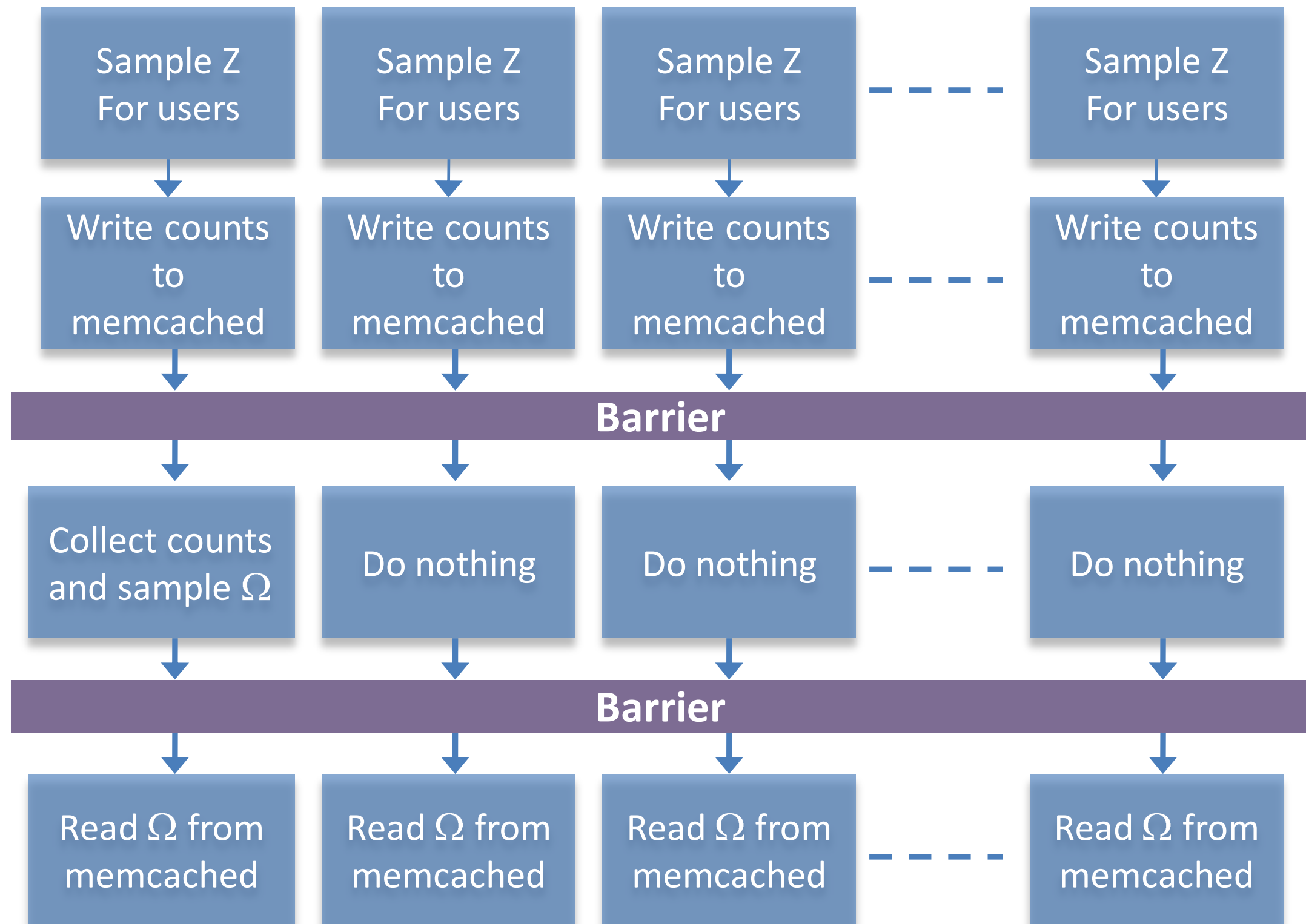


# ROC score improvement

Dataset-2



# LDA for user profiling



News

# News Stream



# News Stream



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

AP - 1 hr 32 mins ago  
WASHINGTON - In the

BEYOND FOSSIL FUELS

## Using Waste, Swedish



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

## China says inflation up 5.1 per cent

Associated Press

Buzz up! 19 votes | Share



Wall Street Video: [Charting Consumer Sentiment](#) CNBC



Wall Street Video: [Bright Future](#) TheStreet.com

### RELATED QUOTES

<b>^DJI</b>	11,410.32	<b>+40.26</b>
<b>^GSPC</b>	1,240.40	<b>+7.40</b>
<b>^IXIC</b>	2,637.54	<b>+20.87</b>

By CARA ANNA, Associated Press

BEIJING - China's inflation surged Saturday, despite supplies and end diesel shortages

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

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

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

## Suit to Recover Madoff's Money Calls Austrian an Accomplice

By DIANA B. HENRIQUES and PETER LATTMAN

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

Post a Comment

er

Print

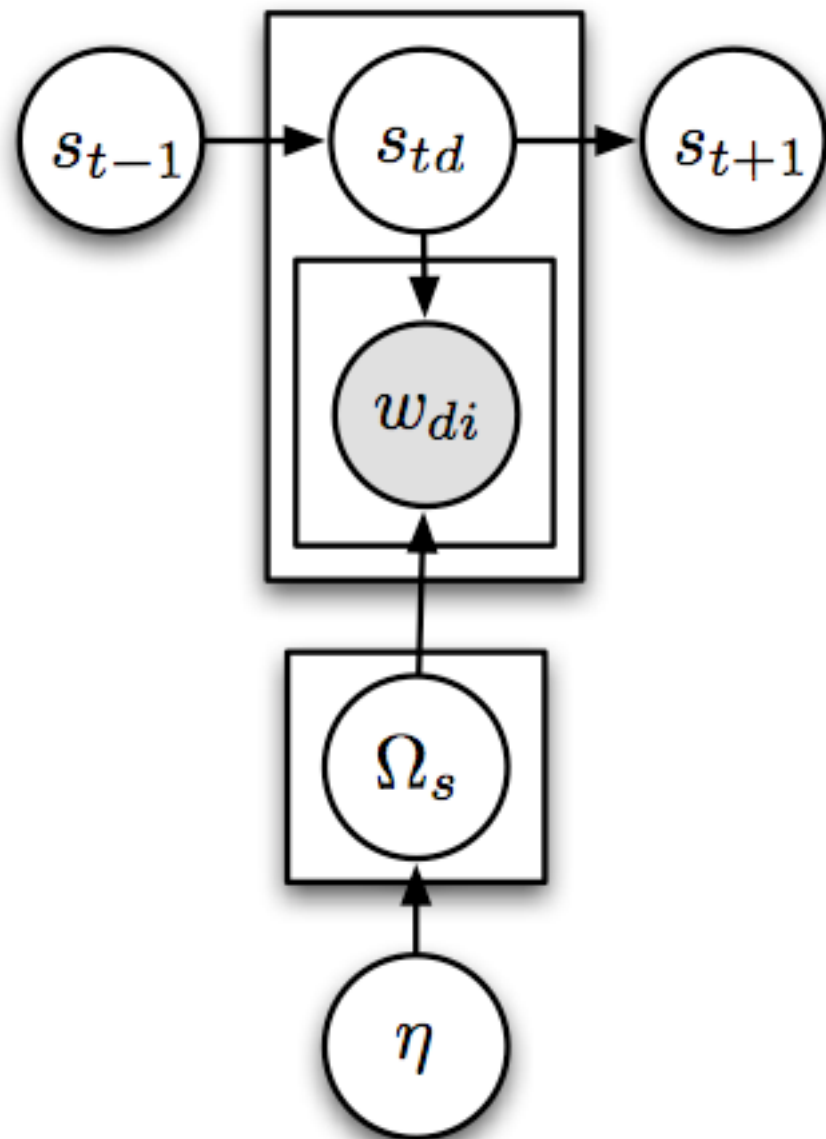
November, use food



# 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



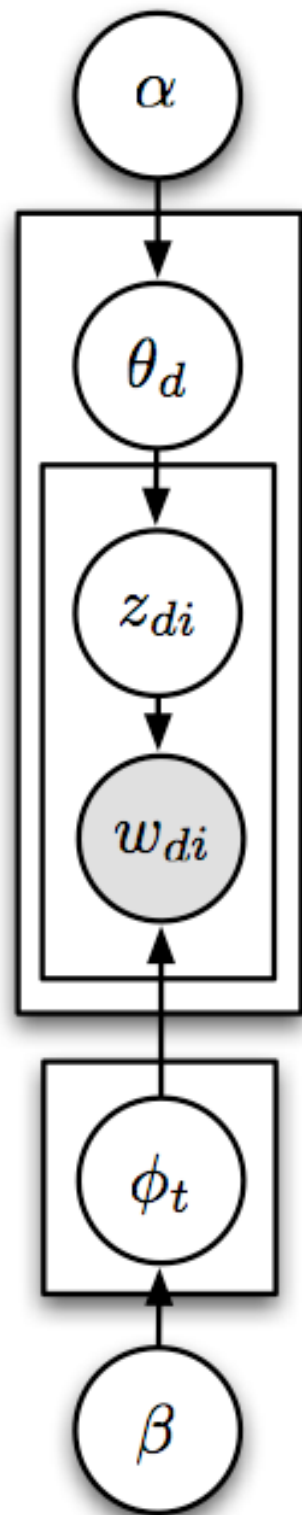
- Assume active story distribution at time  $t$
- Draw story indicator
- Draw words from story distribution
- Down-weight story counts for next day

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



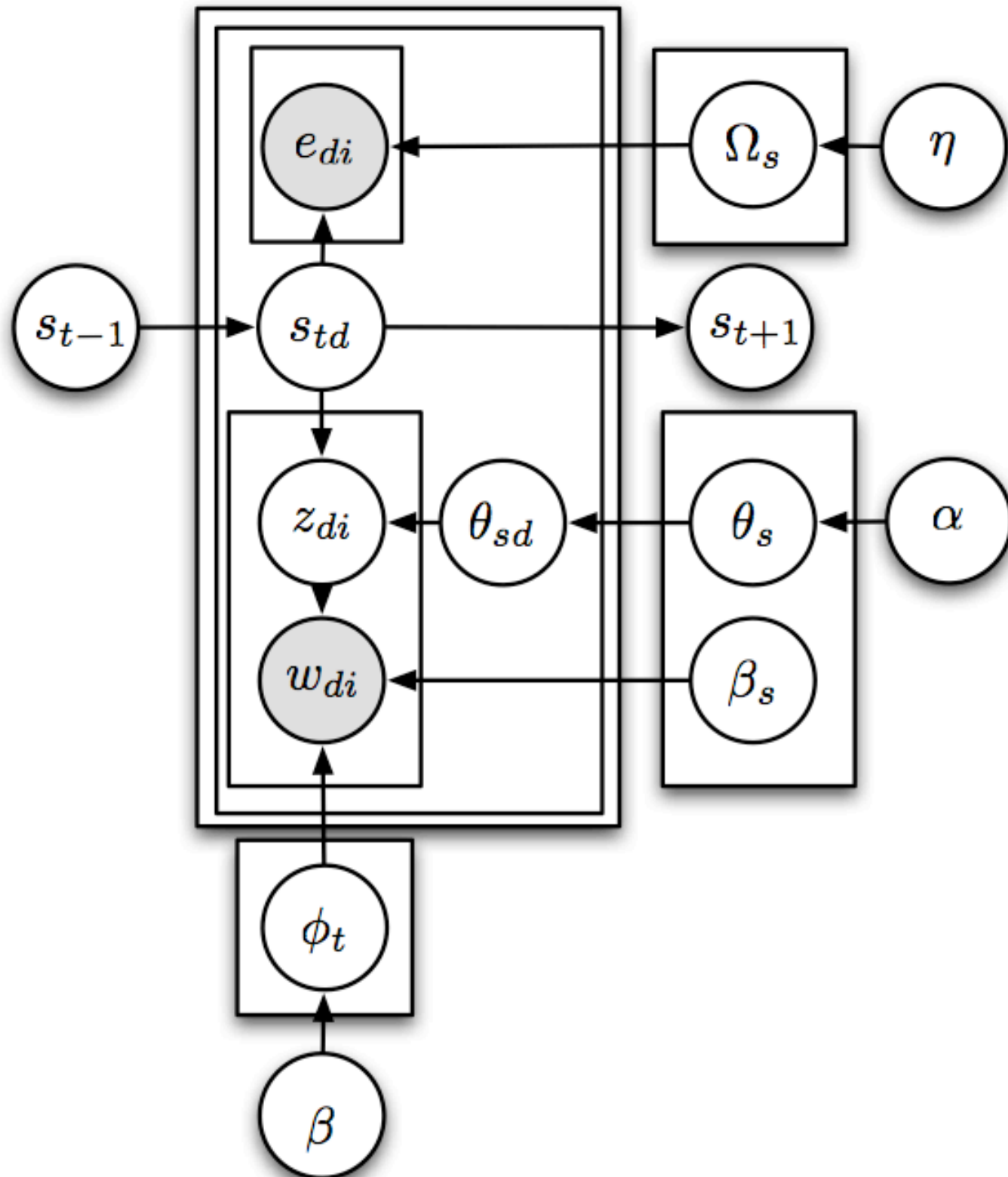
# More Issues

- **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

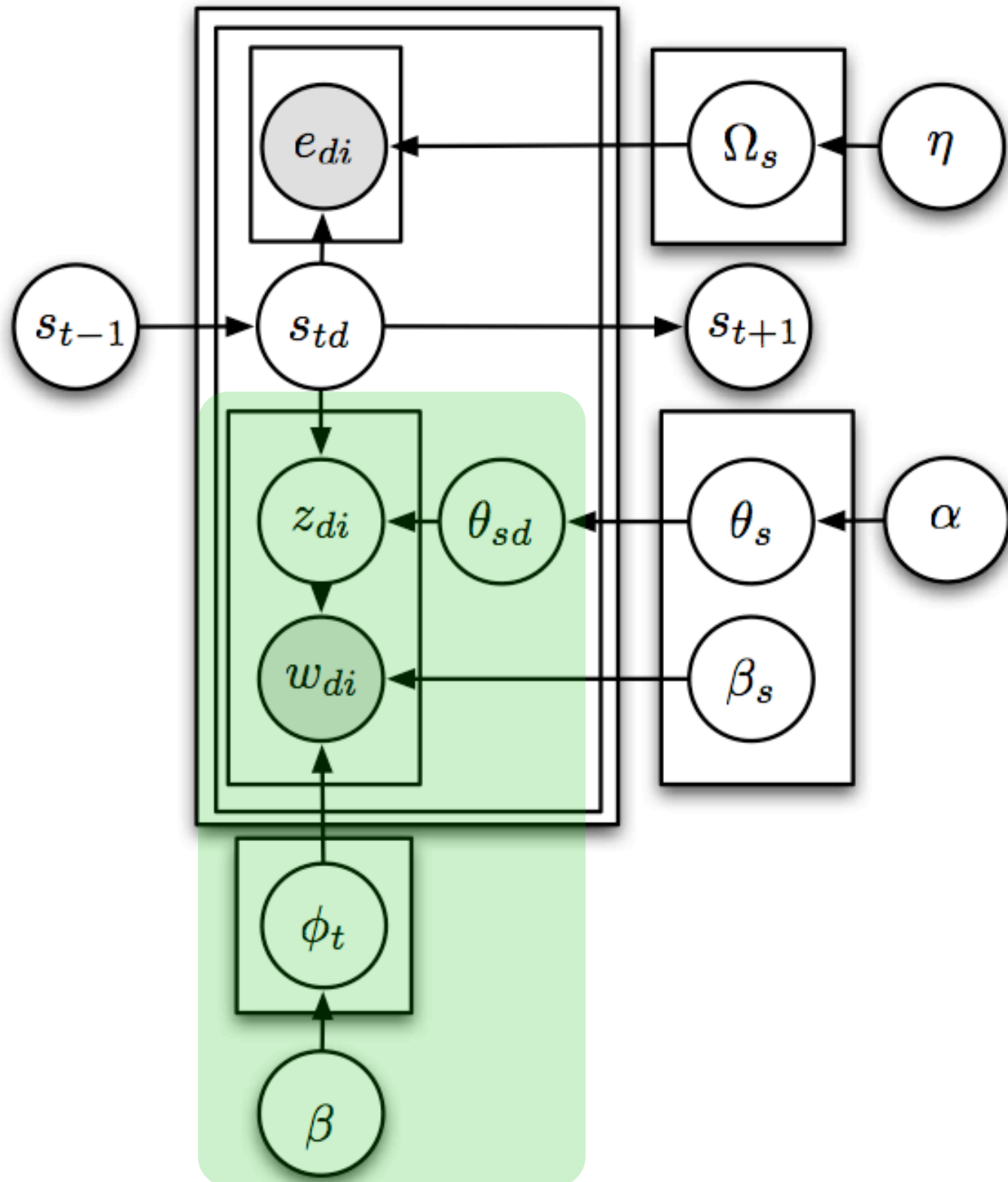


# Storylines Model



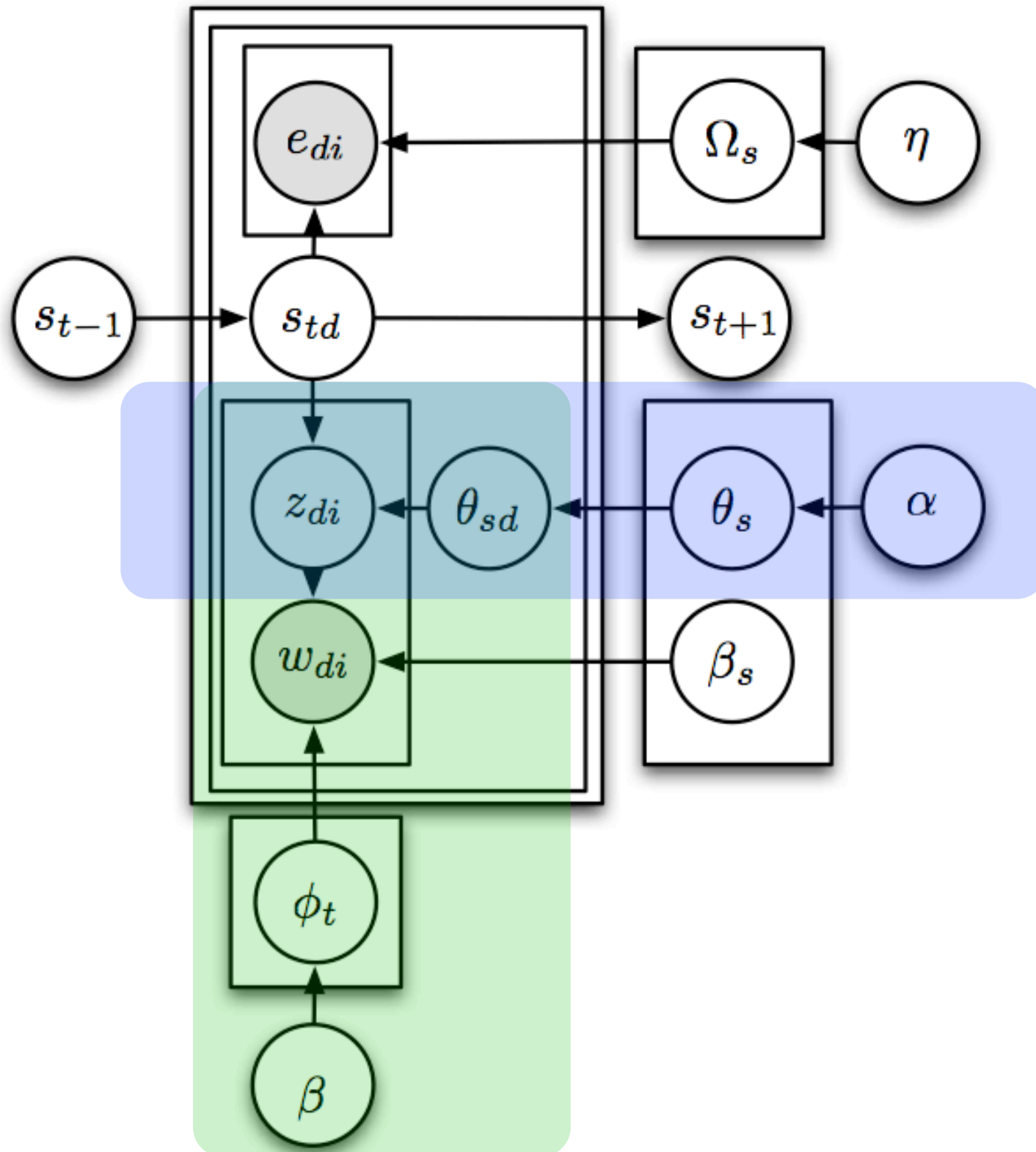
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

# Storylines Model



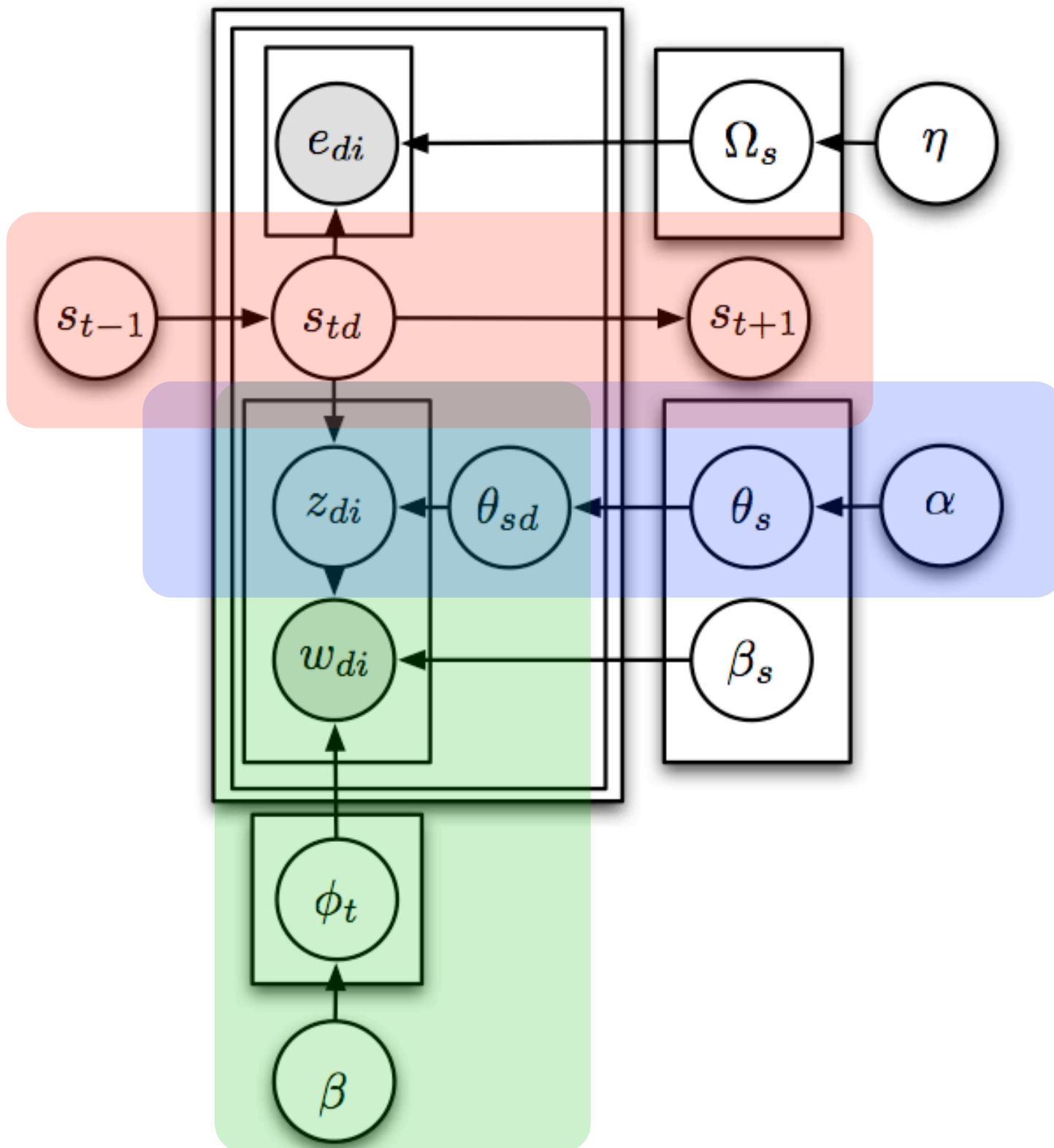
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

# Storylines Model



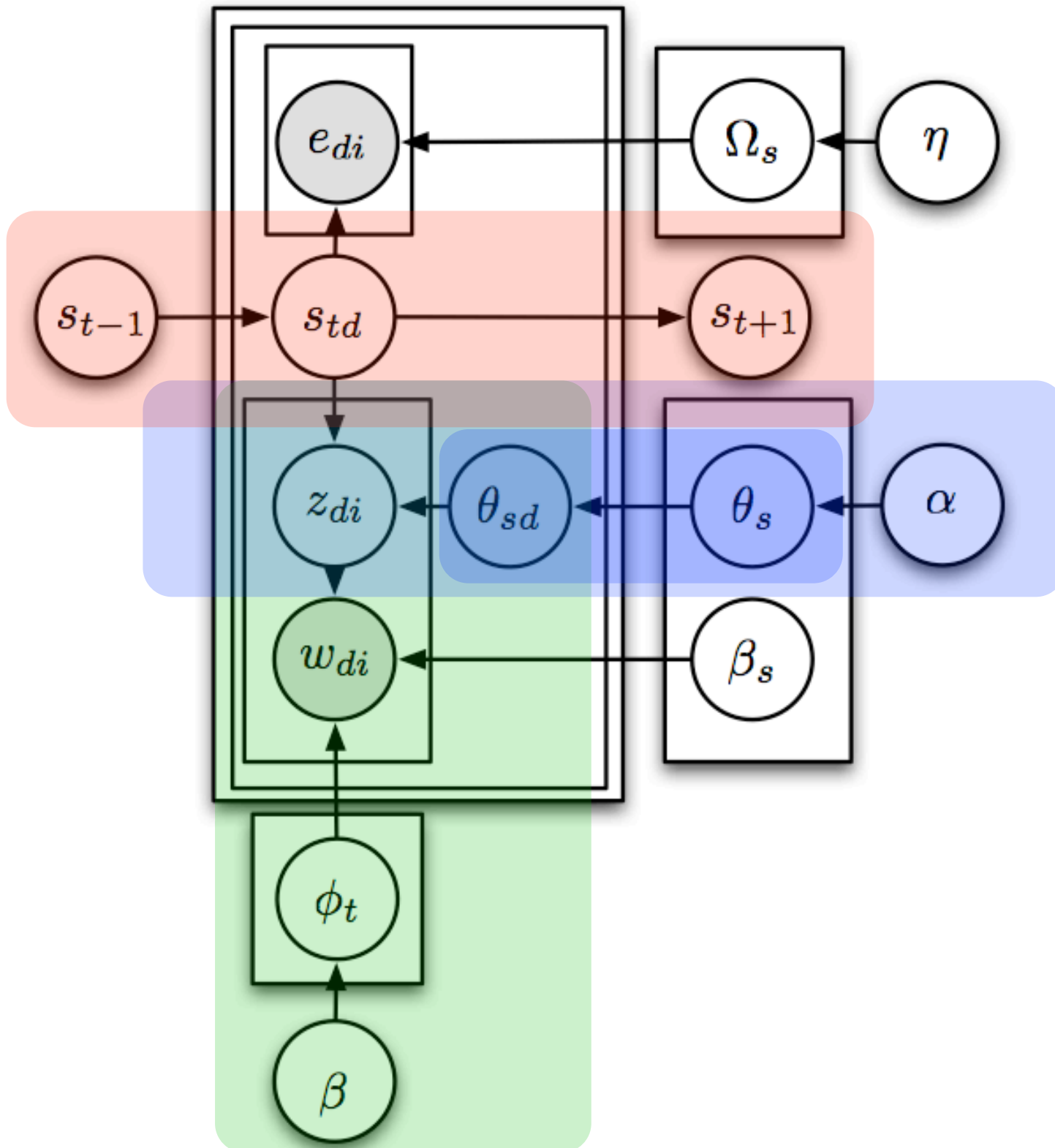
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

# Storylines Model



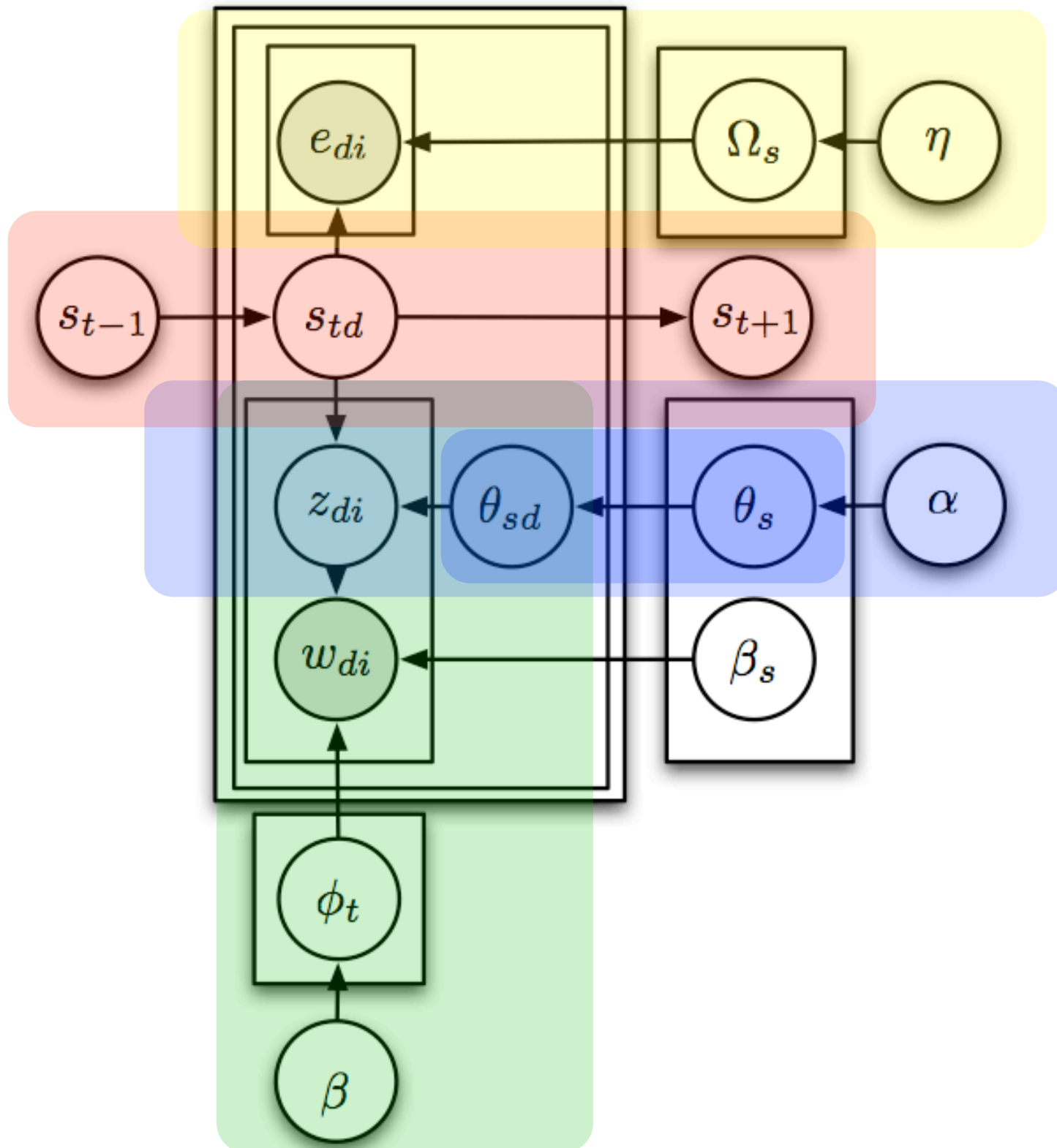
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

# Storylines Model



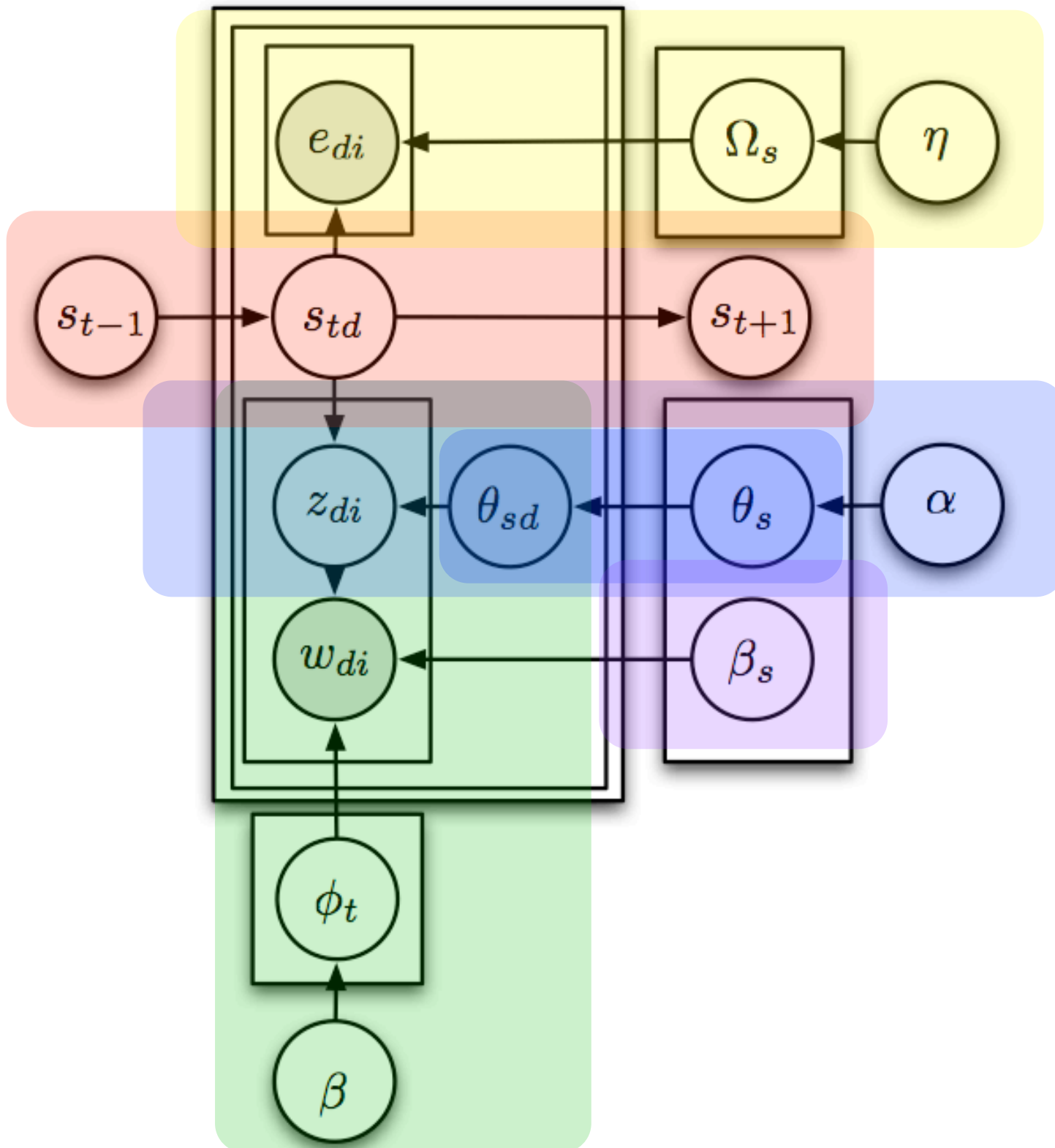
- Topic model
- Topics per cluster
- RCRP for cluster
- Hierarchical DP for article
- Separate model for named entities
- Story specific correction

# Storylines Model



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

# Storylines Model



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



# Dynamic Cluster-Topic Hybrid

**Sports**  
games  
Won  
Team  
Final  
Season  
League  
held

**Politics**  
Government  
Minister  
Authorities  
Opposition  
Officials  
Leaders  
group

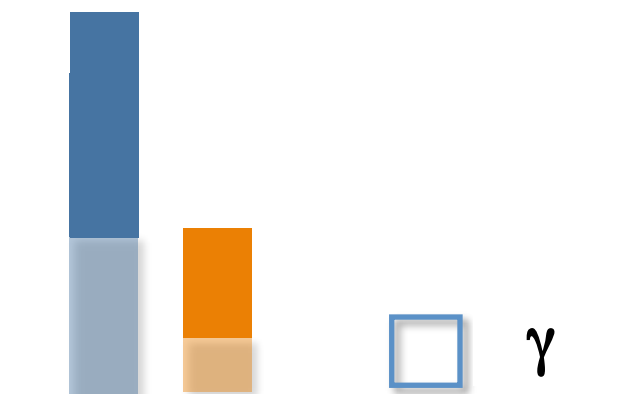
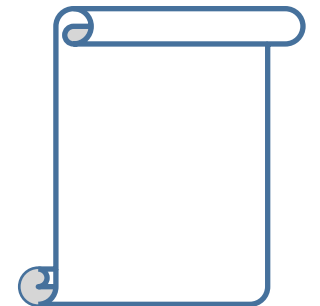
**Accidents**  
Police  
Attack  
run  
man  
group  
arrested  
move

## UEFA-soccer

Champions	Juventus
Goal	AC Milan
Coach	Lazio
Striker	Ronaldo
Midfield	Lyon
penalty	

## Tax-Bill

Tax	Bush
Billion	Senate
Cut	Fleischer
Plan	White House
Budget	Republican
Economy	





# Dynamic Cluster-Topic Hybrid

**Sports**  
games  
Won  
Team  
Final  
Season  
League  
held

**Politics**  
Government  
Minister  
Authorities  
Opposition  
Officials  
Leaders  
group

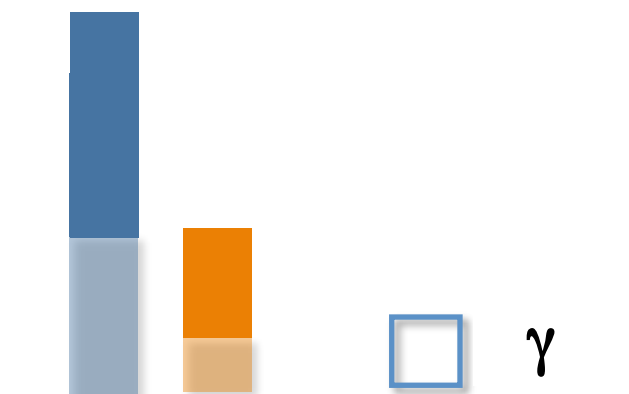
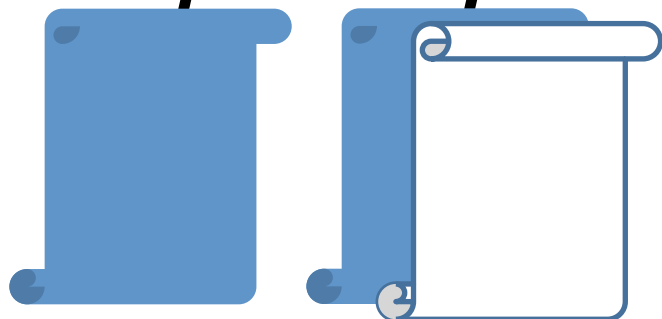
**Accidents**  
Police  
Attack  
run  
man  
group  
arrested  
move

## UEFA-soccer

Champions	Juventus
Goal	AC Milan
Coach	Lazio
Striker	Ronaldo
Midfield	Lyon
penalty	

## Tax-Bill

Tax	Bush
Billion	Senate
Cut	Fleischer
Plan	White House
Budget	Republican
Economy	



# Dynamic Cluster-Topic Hybrid

**Sports**  
games  
Won  
Team  
Final  
Season  
League  
held

**Politics**  
Government  
Minister  
Authorities  
Opposition  
Officials  
Leaders  
group

**Accidents**  
Police  
Attack  
run  
man  
group  
arrested  
move

## UEFA-soccer

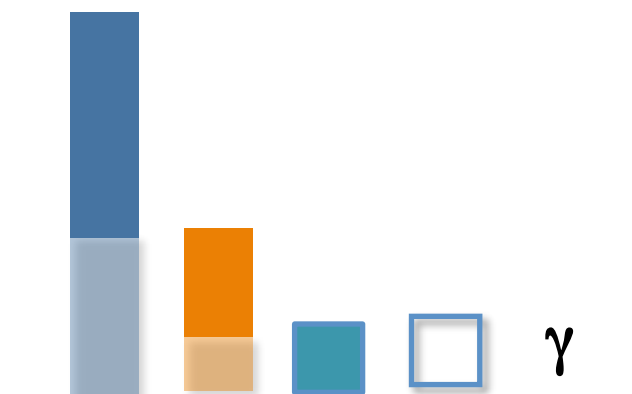
Champions	Juventus
Goal	AC Milan
Coach	Lazio
Striker	Ronaldo
Midfield	Lyon
penalty	

## Tax-Bill

Tax	Bush
Billion	Senate
Cut	Fleischer
Plan	White House
Budget	Republican
Economy	

## Border-Tension

Nuclear	Pakistan
Border	India
Dialogue	Kashmir
Diplomatic	New Delhi
militant	Islamabad
Insurgency	Musharraf
missile	Vajpayee



# Inference

- We receive articles as a stream
  - Want topics & stories now
- Variational inference infeasible
  - (RCRP, sparse to dense, vocabulary size)
- We have a 'tracking problem'
  - Sequential Monte Carlo
  - Use sampled variables of surviving particle
  - Use ideas from Cannini et al. 2009

# Particle Filter

- Proposal distribution - draw stories  $s$ , topics  $z$

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

using Gibbs Sampling for each particle

- Reweight particle via

past state

new data

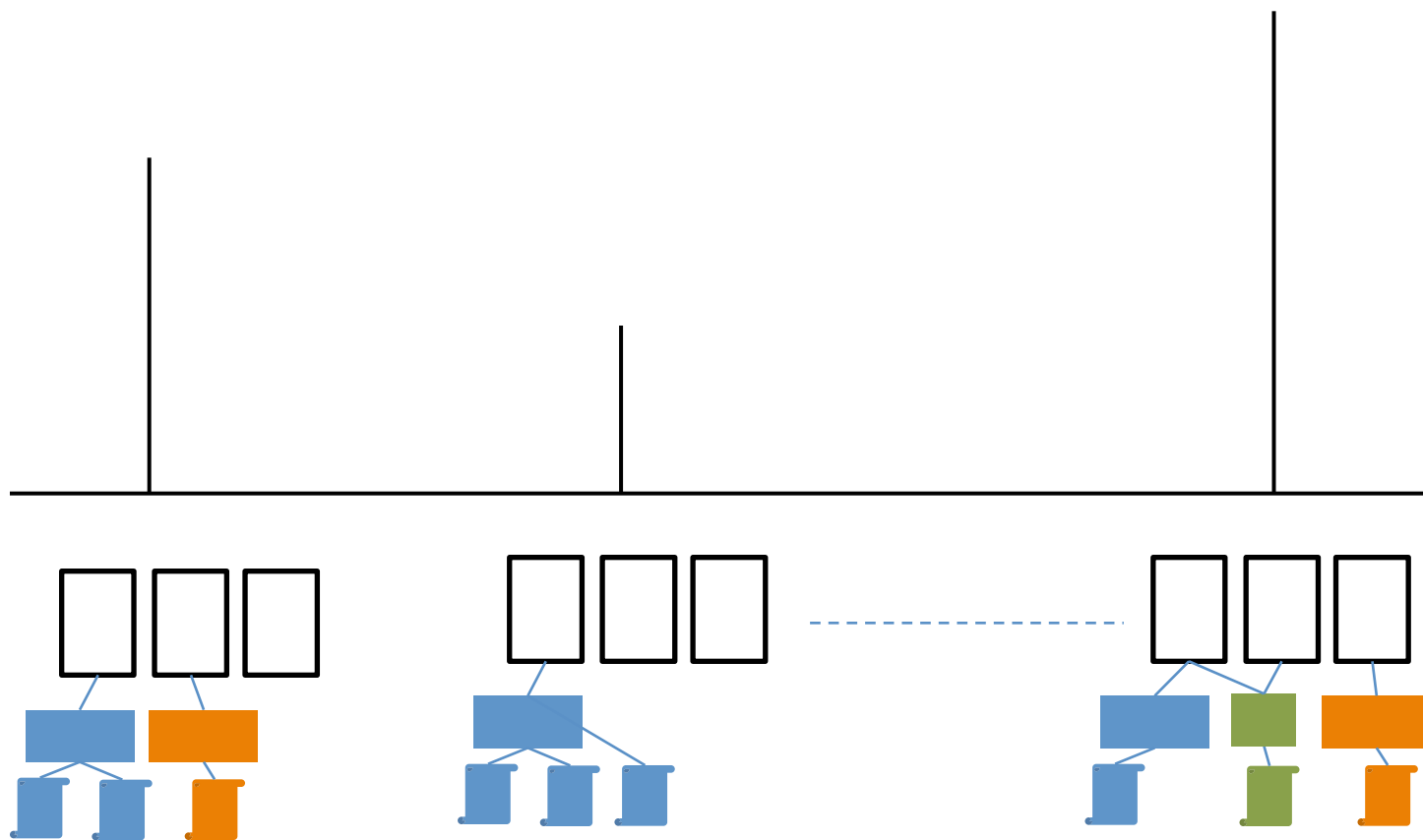
$$p(x_{t+1} | x_{1..t}, s_{1..t}, z_{1..t})$$

- Resample particles if  $l_2$  norm too large  
(resample some assignments for diversity, too)
- Compare to multiplicative updates algorithm  
In our case predictive likelihood yields weights

# Particle Filter

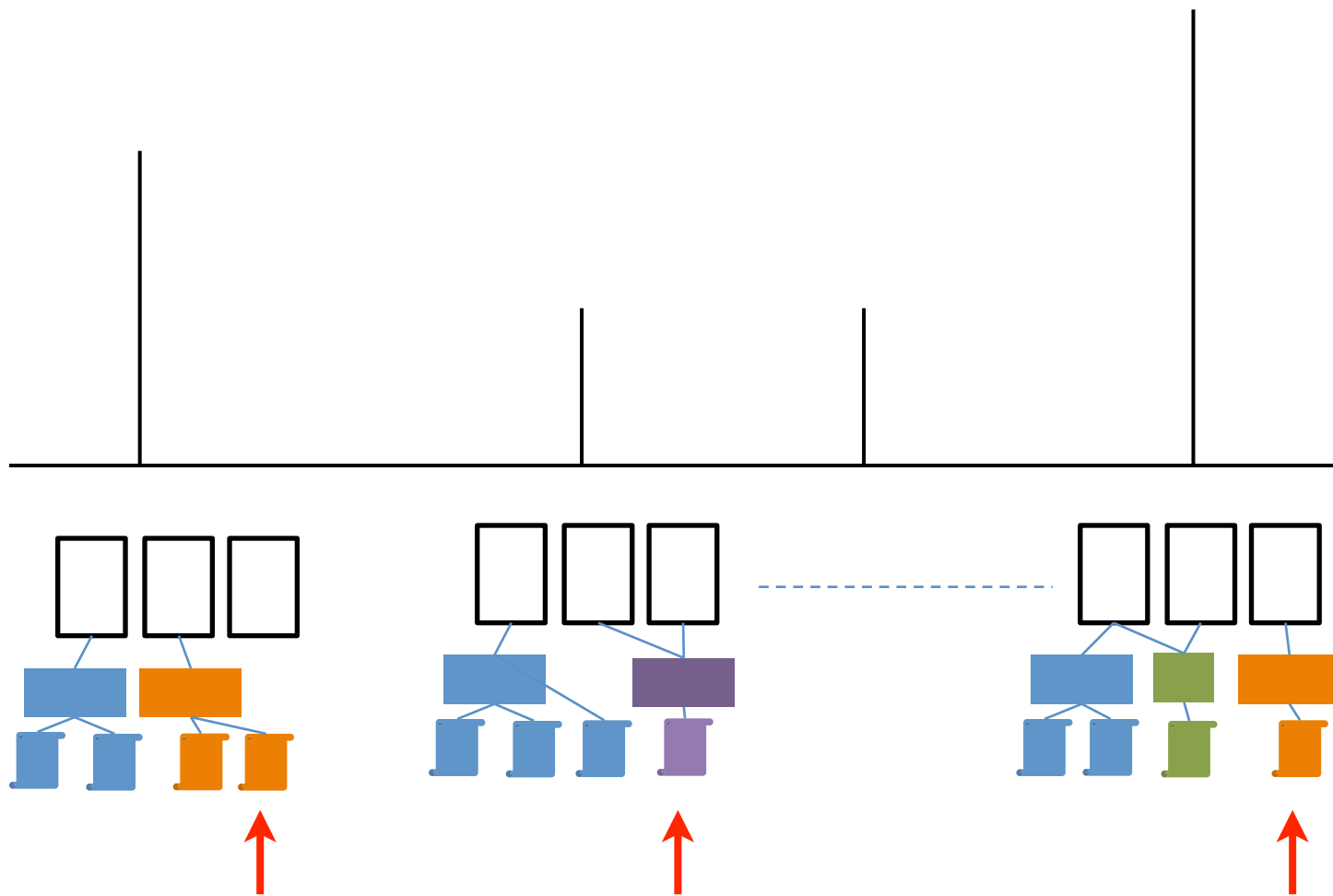
## Algorithm 1 A Particle Filter Algorithm

```
Initialize  $\omega_1^f$  to  $\frac{1}{F}$  for all  $f \in \{1, \dots, F\}$ 
for each document  $d$  with time stamp  $t$  do
  for  $f \in \{1, \dots, F\}$  do
    Sample  $s_{td}^f, z_{td}^f$  using MCMC
     $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t,d-1})$ 
  end for
  Normalize particle weights
  if  $\|\omega_t\|_2^{-2} < \text{threshold}$  then
    resample particles
    for  $f \in \{1, \dots, F\}$  do
      MCMC pass over 10 random past documents
    end for
  end if
end for
```



- $\mathbf{s}$  and  $\mathbf{z}$  are tightly coupled
- Alternative to MCMC
  - Sample  $\mathbf{s}$  then sample  $\mathbf{z}$  (high variance)
  - Sample  $\mathbf{z}$  then sample  $\mathbf{s}$  (doesn't make sense)
- Idea (following a similar trick by Jain and Neal)
  - Run a few iterations of MCMC over  $\mathbf{s}$  and  $\mathbf{z}$
  - Take last sample as the proposed value

# Particle Filter



---

## Algorithm 1 A Particle Filter Algorithm

---

```
Initialize  $\omega_1^f$  to  $\frac{1}{F}$  for all  $f \in \{1, \dots, F\}$   
for each document  $d$  with time stamp  $t$  do  
  for  $f \in \{1, \dots, F\}$  do  
    Sample  $s_{td}^f, z_{td}^f$  using MCMC  
     $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t, d-1})$   
  end for  
  Normalize particle weights  
  if  $\|\omega_t\|_2^{-2} < \text{threshold}$  then  
    resample particles  
    for  $f \in \{1, \dots, F\}$  do  
      MCMC pass over 10 random past documents  
    end for  
  end if  
end for
```

---

# Particle Filter

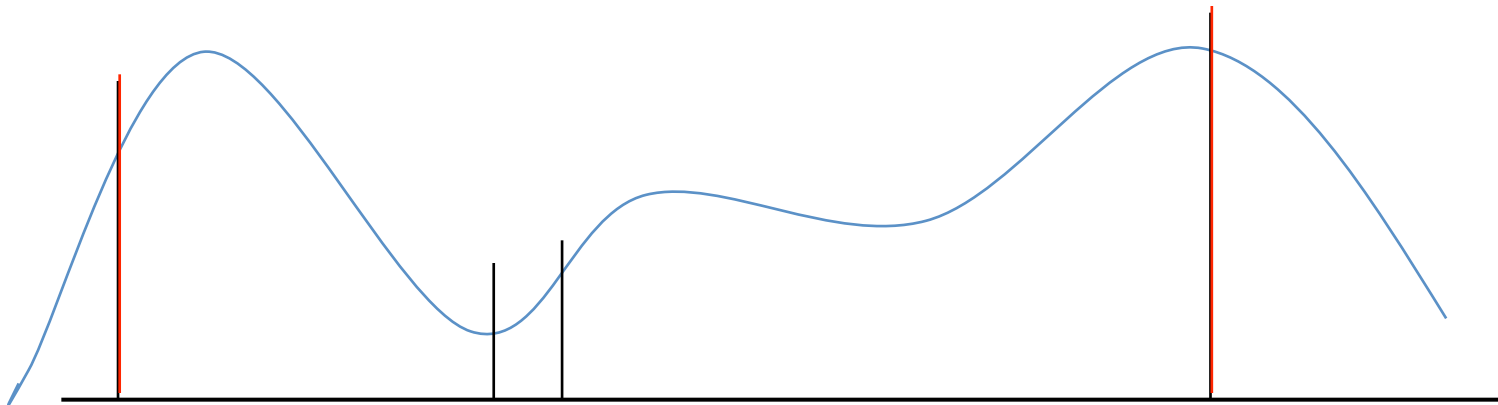
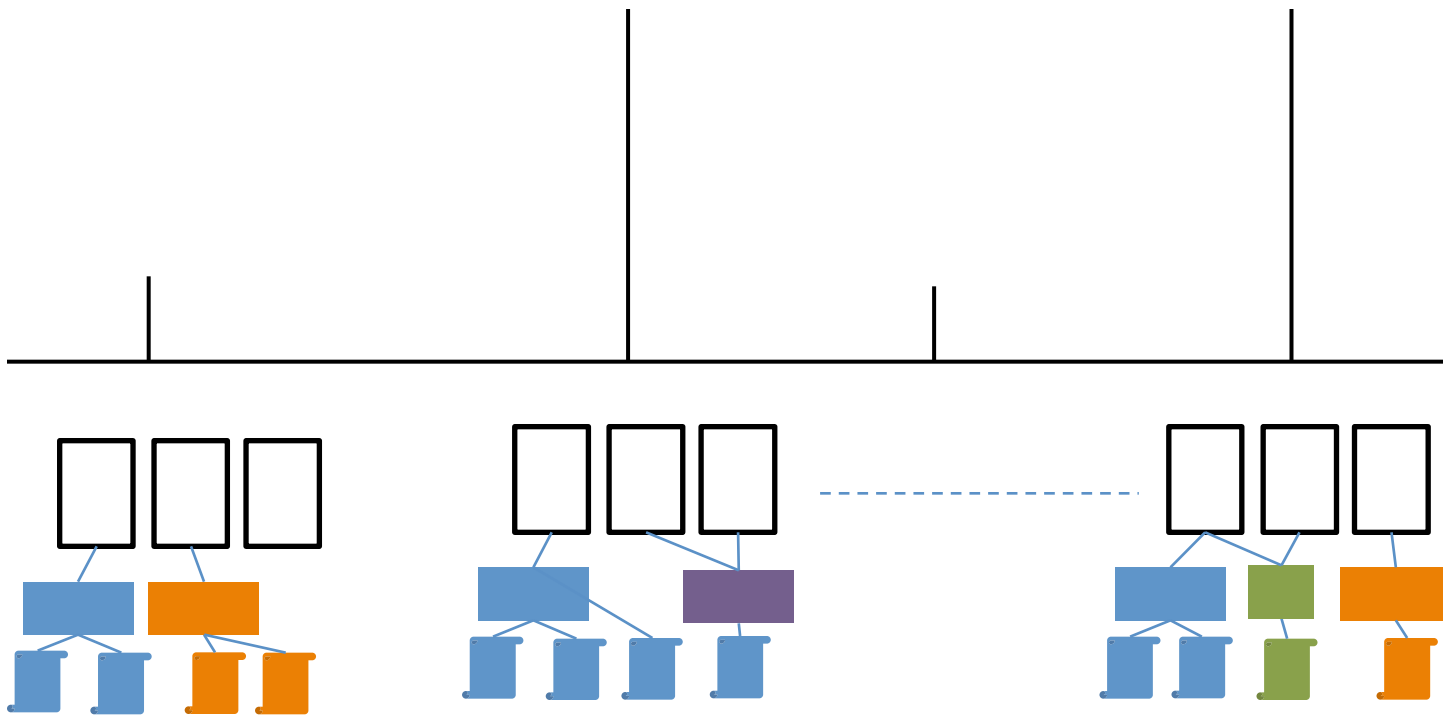
---

**Algorithm 1** A Particle Filter Algorithm

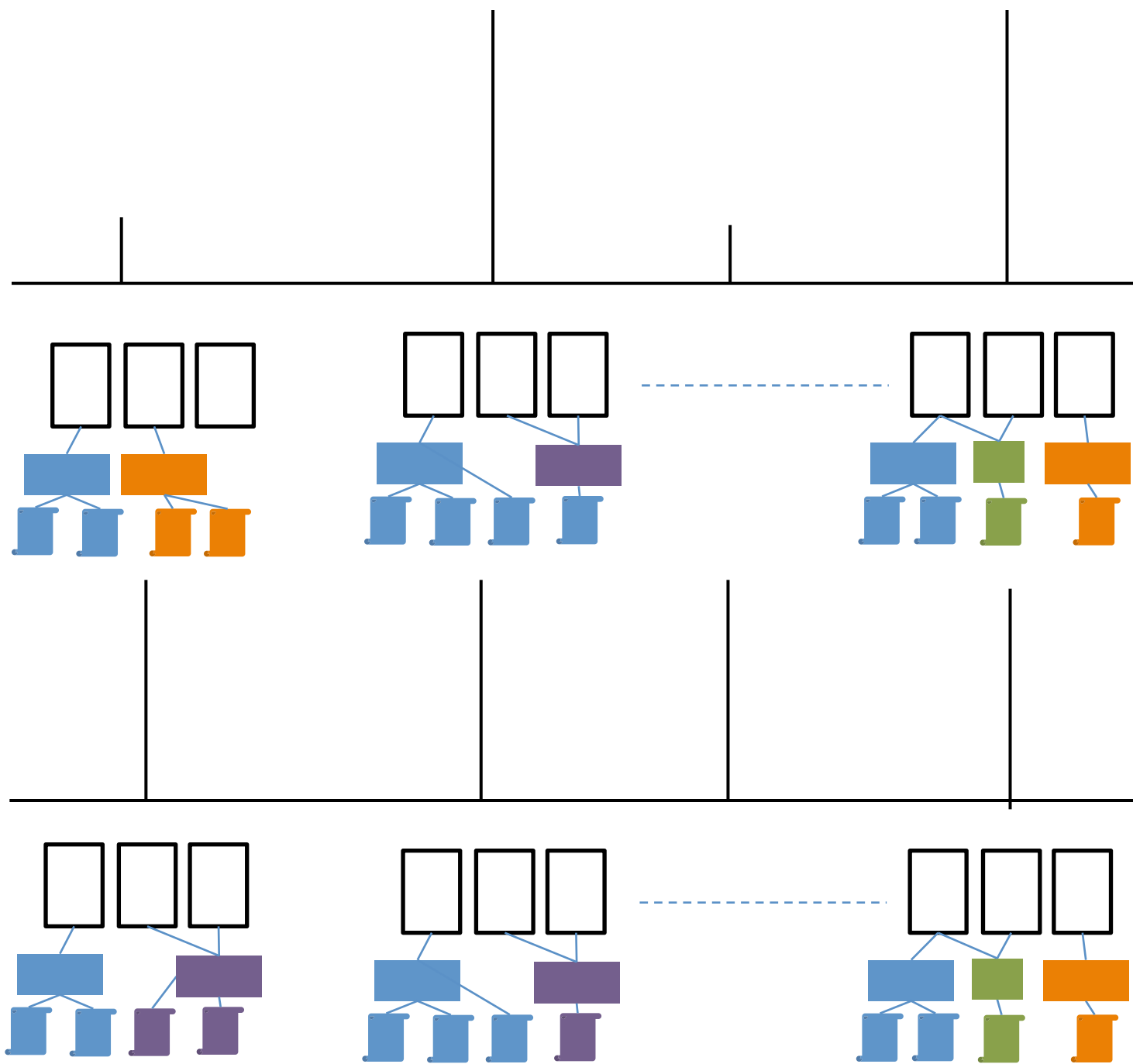
---

```
Initialize  $\omega_1^f$  to  $\frac{1}{F}$  for all  $f \in \{1, \dots, F\}$ 
for each document  $d$  with time stamp  $t$  do
  for  $f \in \{1, \dots, F\}$  do
    Sample  $s_{td}^f, z_{td}^f$  using MCMC
     $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t, d-1})$ 
  end for
  Normalize particle weights
  if  $\|\omega_t\|_2^{-2} < \text{threshold}$  then
    resample particles
    for  $f \in \{1, \dots, F\}$  do
      MCMC pass over 10 random past documents
    end for
  end if
end for
```

---



# Particle Filter



---

## Algorithm 1 A Particle Filter Algorithm

---

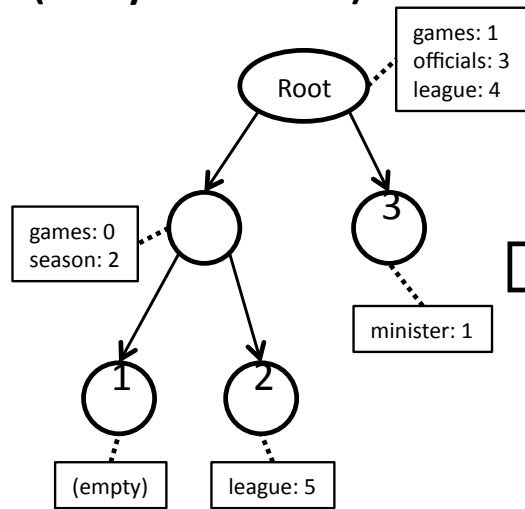
```
Initialize  $\omega_1^f$  to  $\frac{1}{F}$  for all  $f \in \{1, \dots, F\}$   
for each document  $d$  with time stamp  $t$  do  
  for  $f \in \{1, \dots, F\}$  do  
    Sample  $s_{td}^f, z_{td}^f$  using MCMC  
     $\omega^f \leftarrow \omega^f P(\mathbf{x}_{td} | \mathbf{z}_{td}^f, \mathbf{s}_{td}^f, \mathbf{x}_{1:t, d-1})$   
  end for  
  Normalize particle weights  
  if  $\|\omega_t\|_2^{-2} < \text{threshold}$  then  
    resample particles  
    for  $f \in \{1, \dots, F\}$  do  
      MCMC pass over 10 random past documents  
    end for  
  end if  
end for
```

---

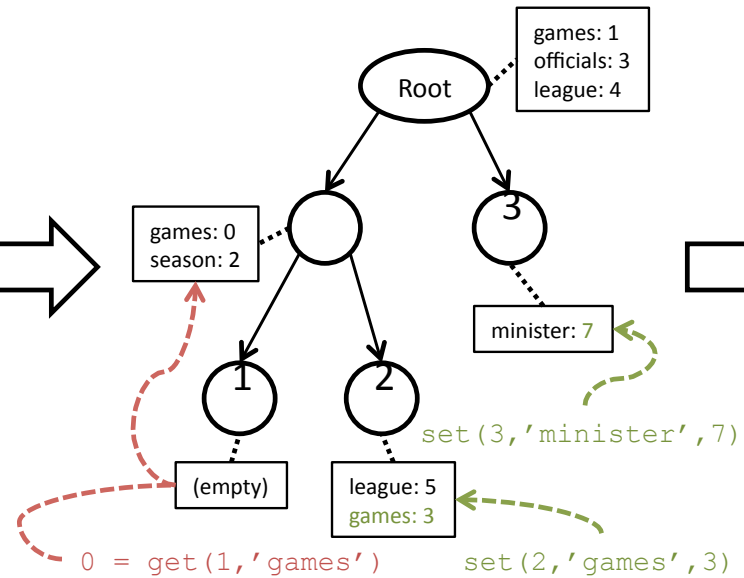


# Inheritance Tree

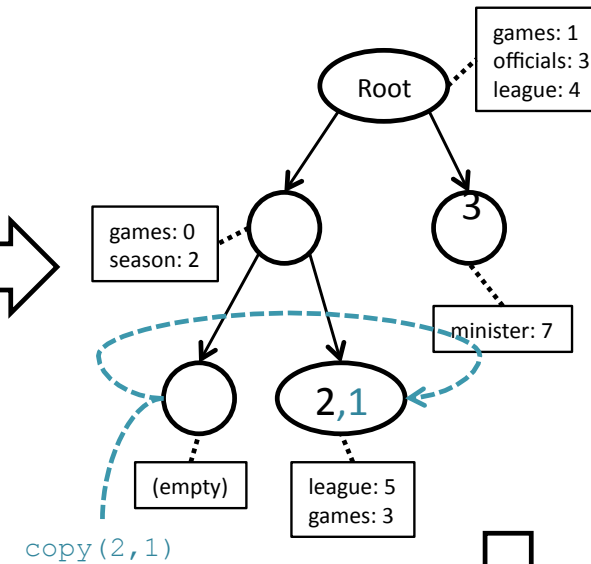
Initial tree  
(ready for threads)



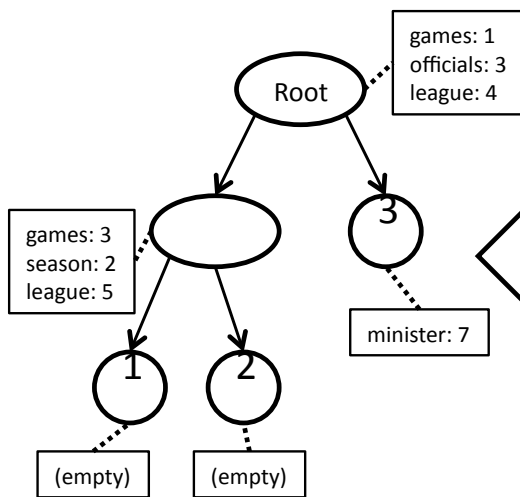
Filter threads *update* particles



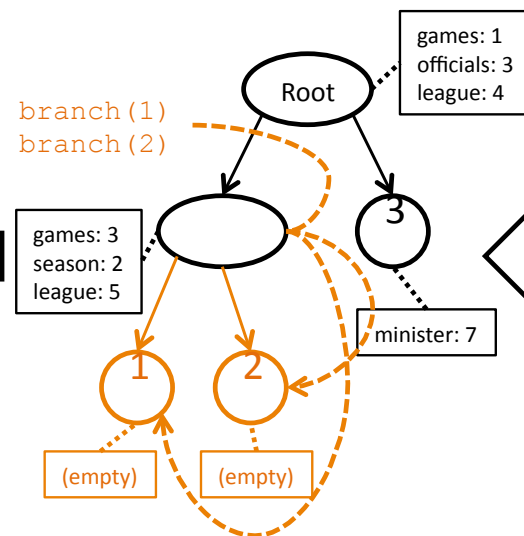
Resampling *copies* particles



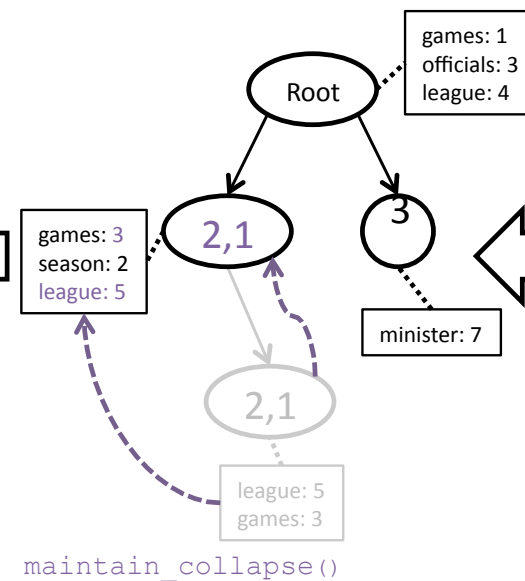
New initial tree  
(ready for threads)



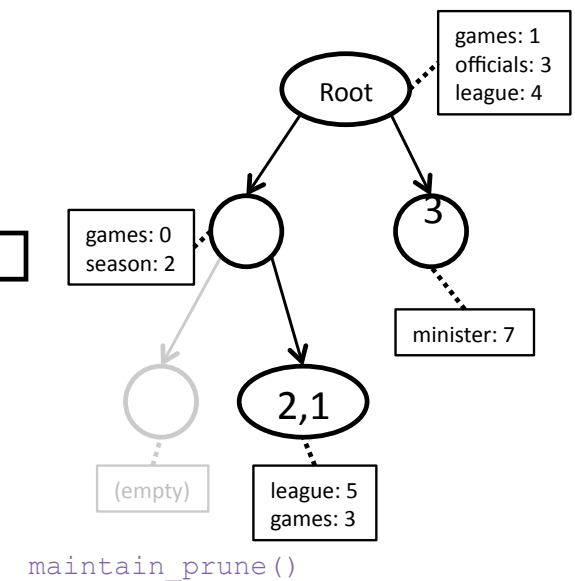
Create *new* leaves



Collapse long branches

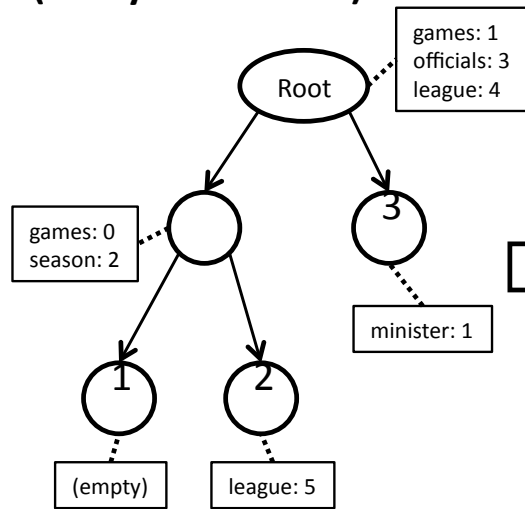


Prune unused branches

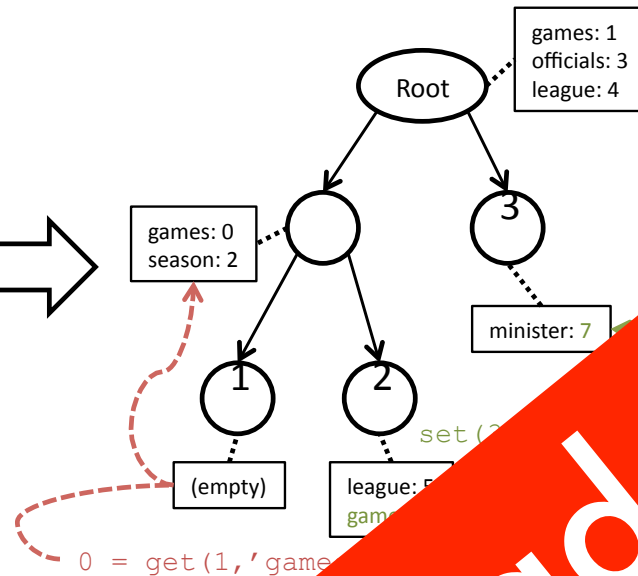


# Inheritance Tree

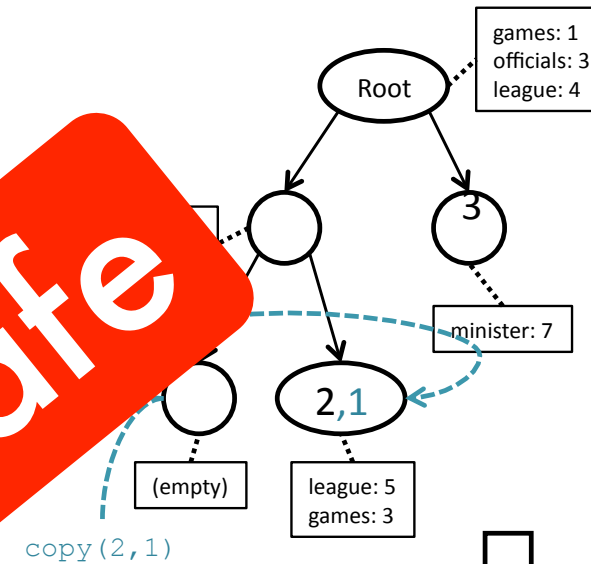
Initial tree  
(ready for threads)



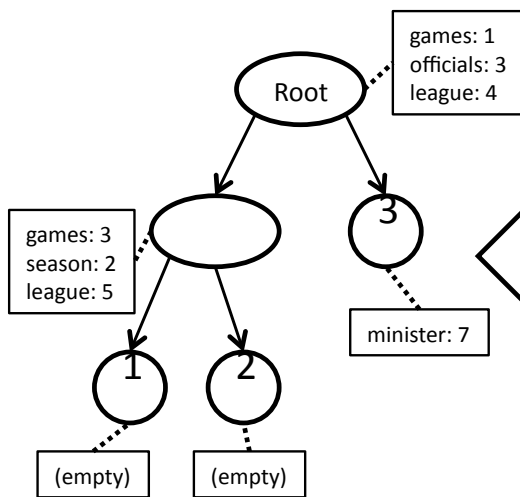
Filter threads *update* particles



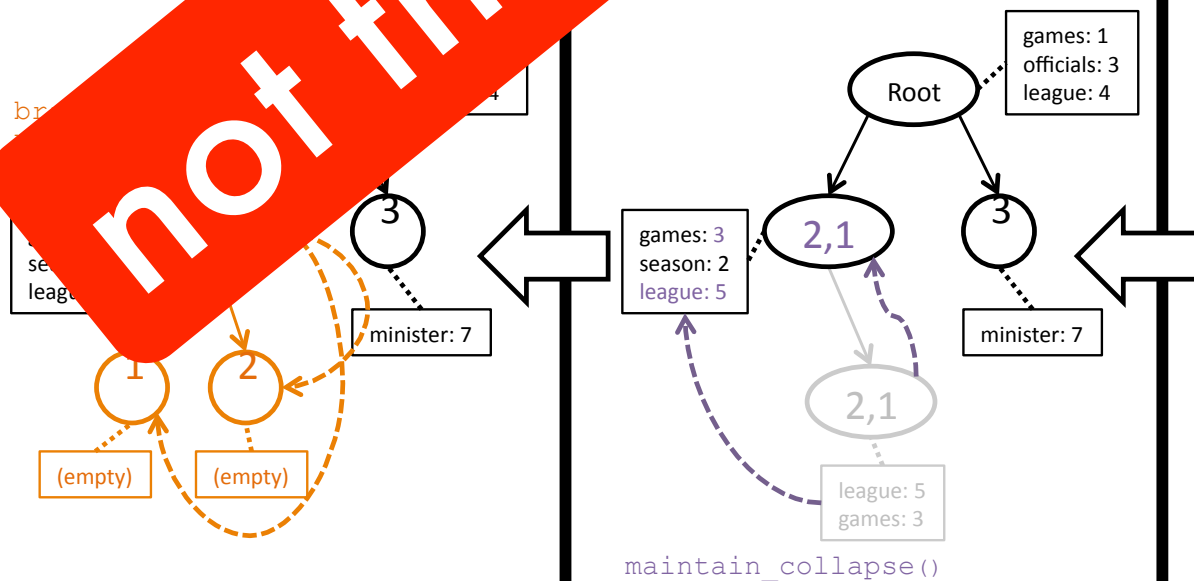
Resampling *copies* particles



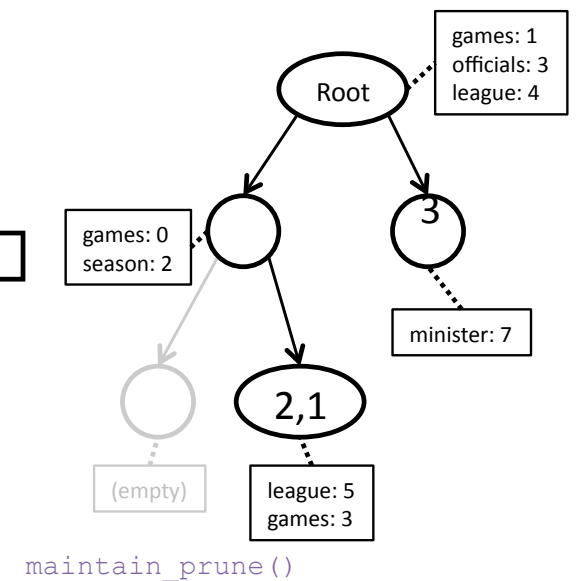
New initial tree  
(ready for threads)



Create *collapse* long branches



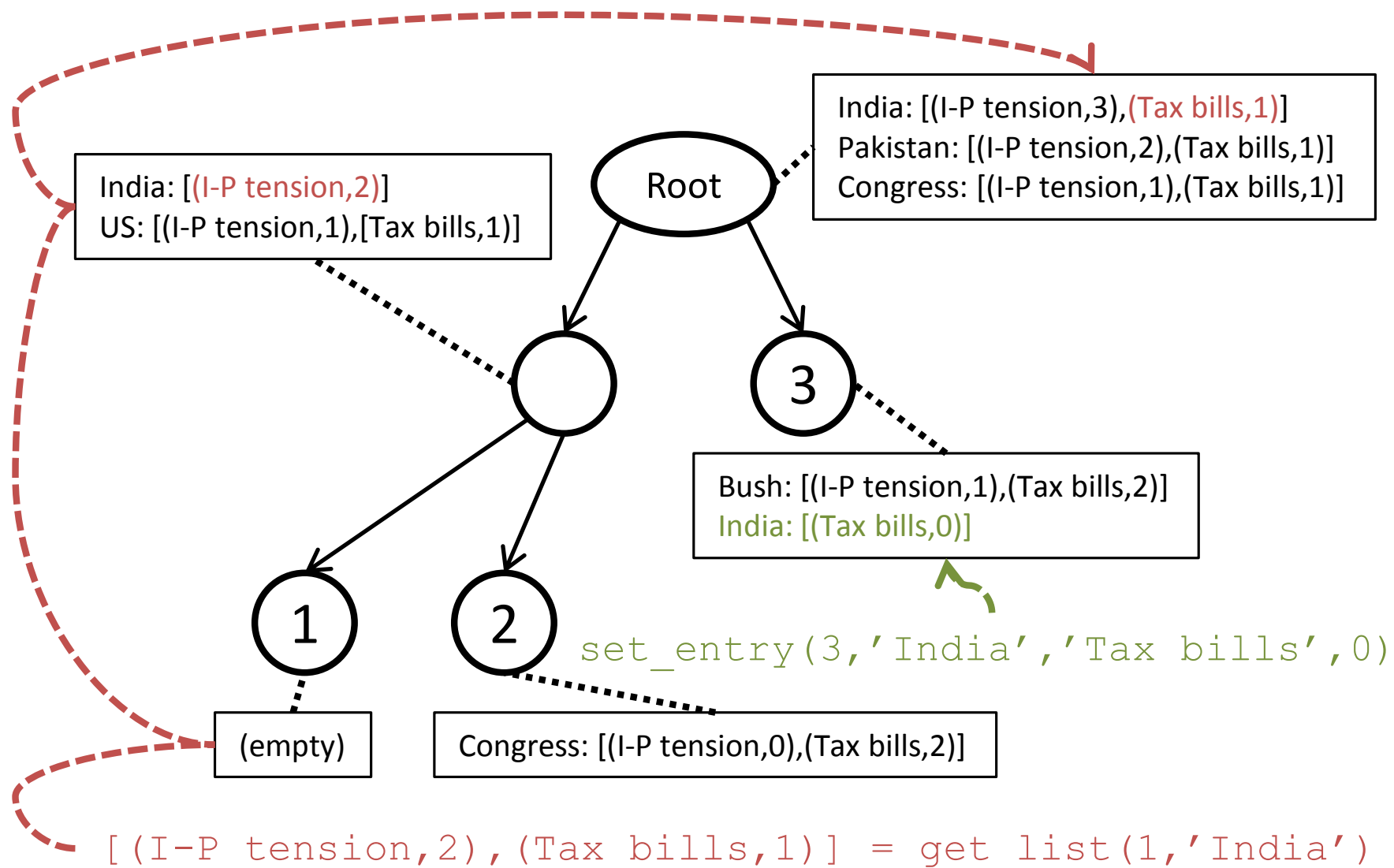
Prune unused branches



**not thread safe**

# Extended Inheritance Tree

## Extended Inheritance Tree



write only in  
the leaves  
(per thread)

Note: "I-P tension" is short for "India-Pakistan tension"

# Results

# Ablation studies

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

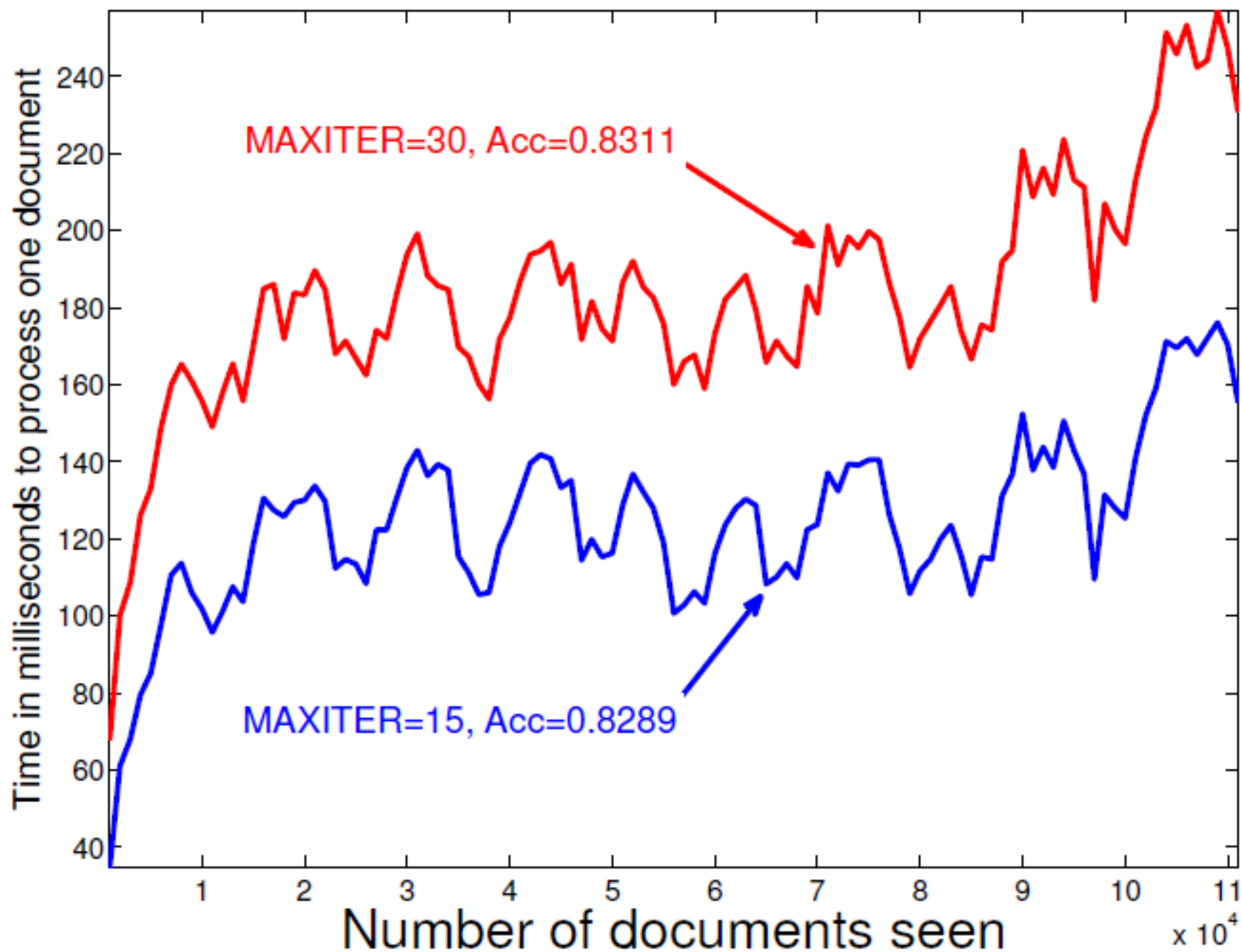
time	entities	topics	story words
0.84	0.90	0.86	0.75

# Comparison

Sample No.	Sample size	Num Words	Num Entities	Story Acc.	LSHC Acc.
1	111,732	19,218	12,475	<b>0.8289</b>	0.738
2	274,969	29,604	21,797	<b>0.8388</b>	0.791
3	547,057	40,576	32,637	<b>0.8395</b>	0.800

Hashing &  
correlation clustering

# Time-Accuracy trade off



# Stories

TOPICS

## Sports

games  
won  
team  
final  
season  
league  
held

## Politics

government  
minister  
authorities  
opposition  
officials  
leaders  
group

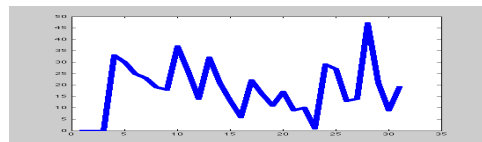
## Unrest

police  
attack  
run  
man  
group  
arrested  
move

STORYLINES

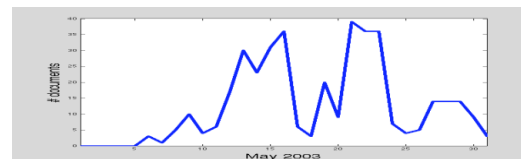
## UEFA-soccer

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



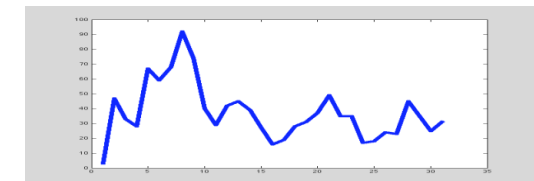
## Tax bills

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



## India-Pakistan tension

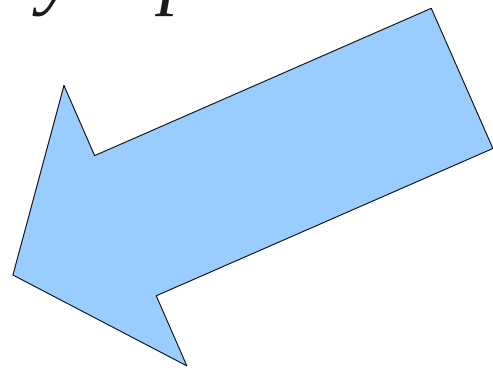
nuclear	<i>Pakistan</i>
border	<i>India</i>
dialogue	<i>Kashmir</i>
diplomatic	<i>New Delhi</i>
militant	<i>Islamabad</i>
insurgency	<i>Musharraf</i>
missile	<i>Vajpayee</i>





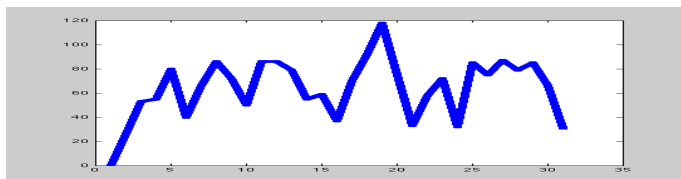
# Related Stories

“Show similar stories by topic”



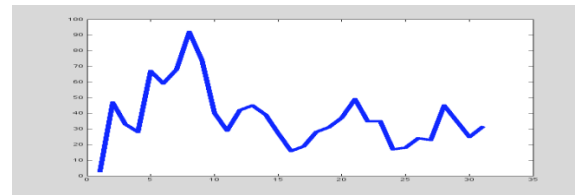
## Middle-east conflict

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

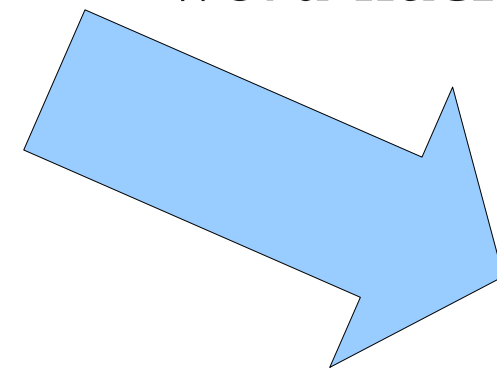


## India-Pakistan tension

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

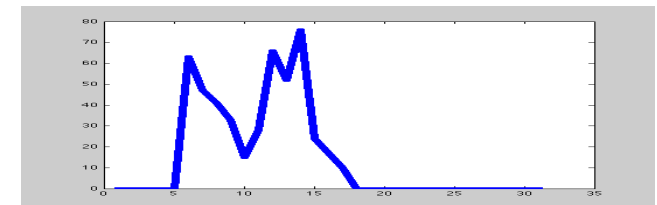


“Show similar stories, require the word nuclear”



## North Korea nuclear

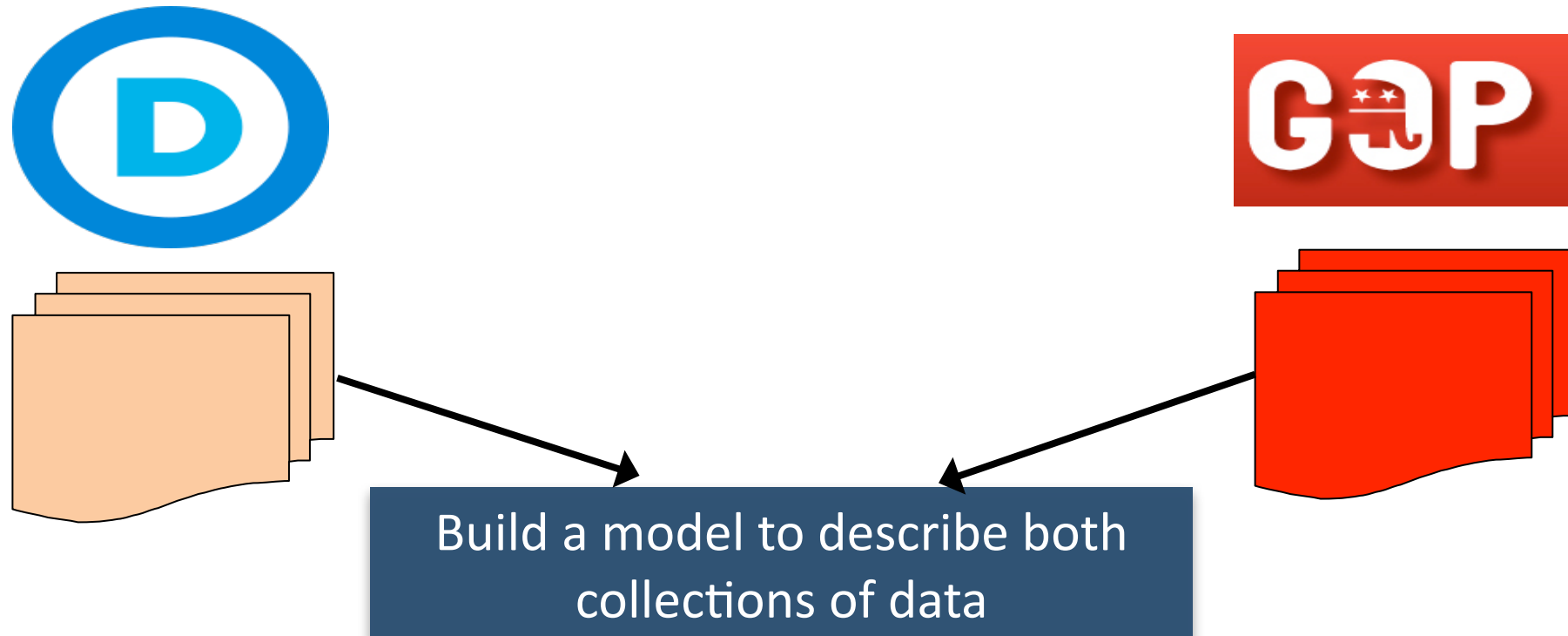
<b>nuclear</b>	<i>North Korea</i>
summit	<i>South Korea</i>
warning	<i>U.S</i>
policy	<i>Bush</i>
missile	<i>Pyongyang</i>
program	



# Detecting Ideologies

Ahmed and Xing, 2010

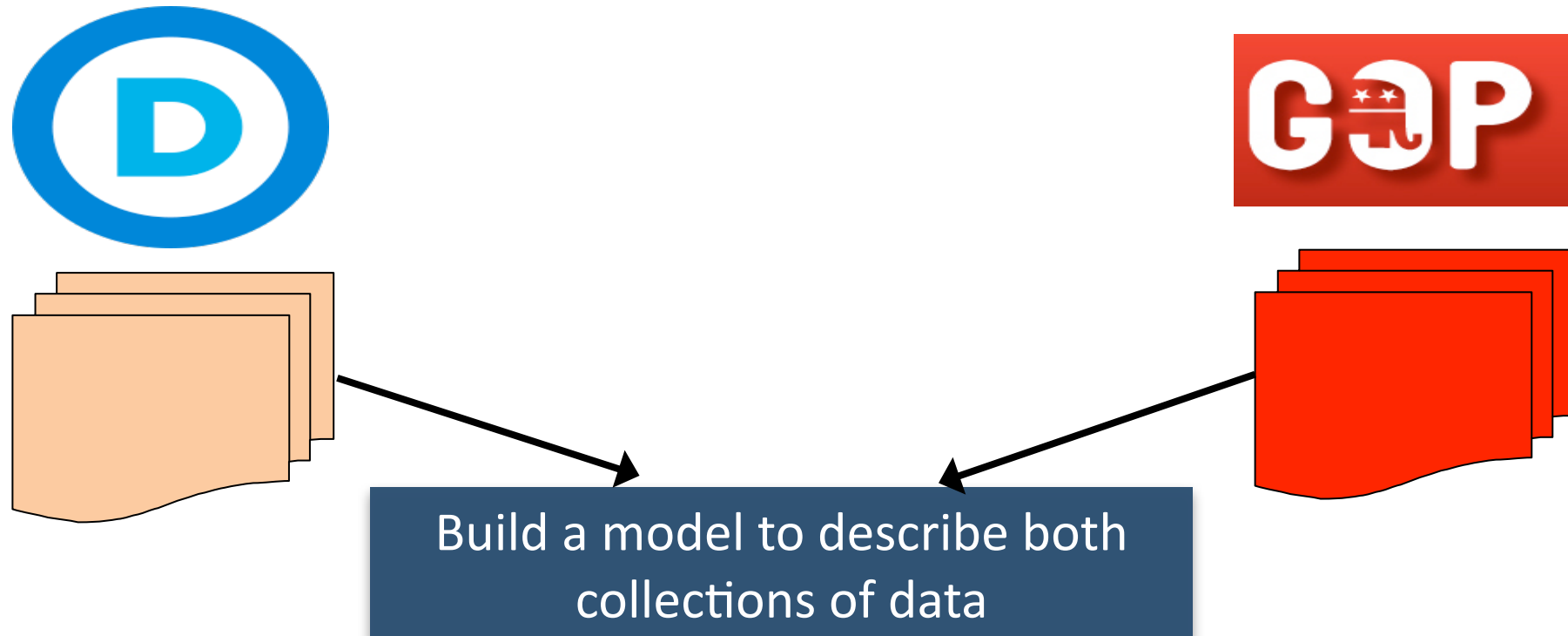
# Ideologies



## Visualization

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

# Ideologies

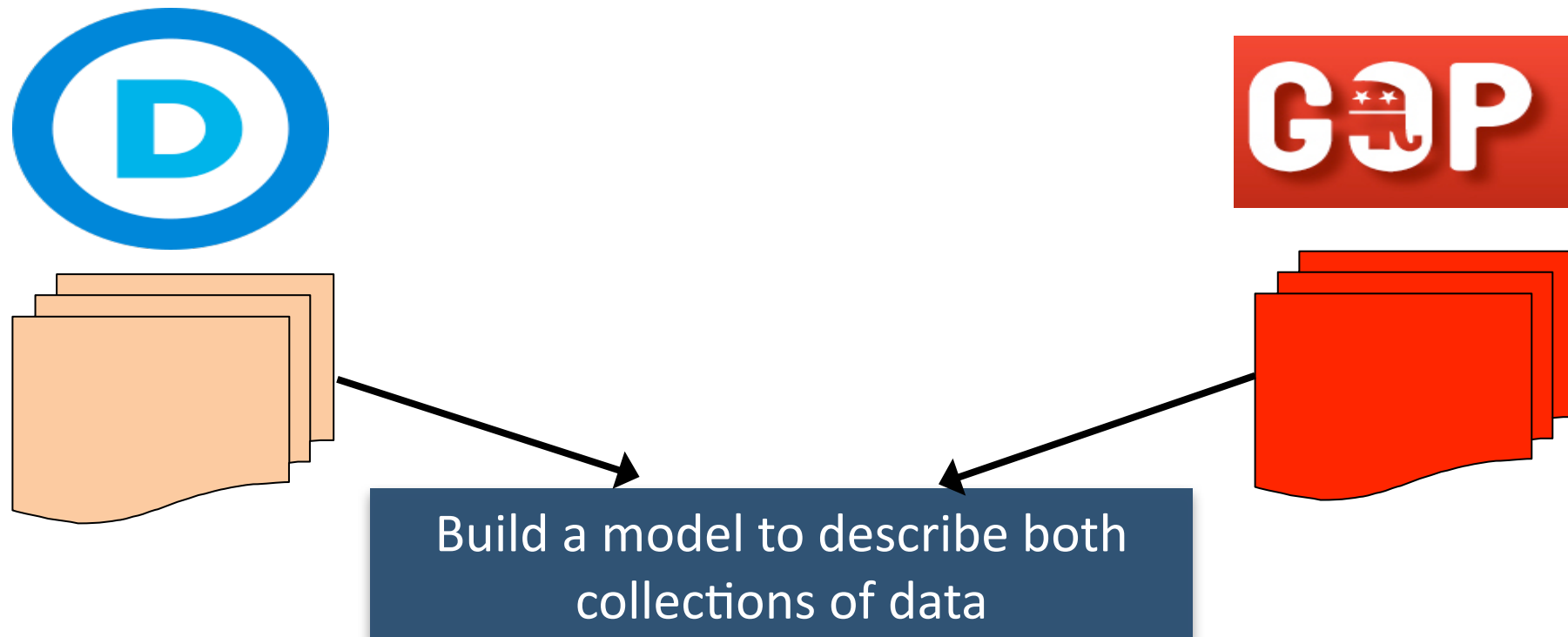


Visualization

Classification

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

# Ideologies



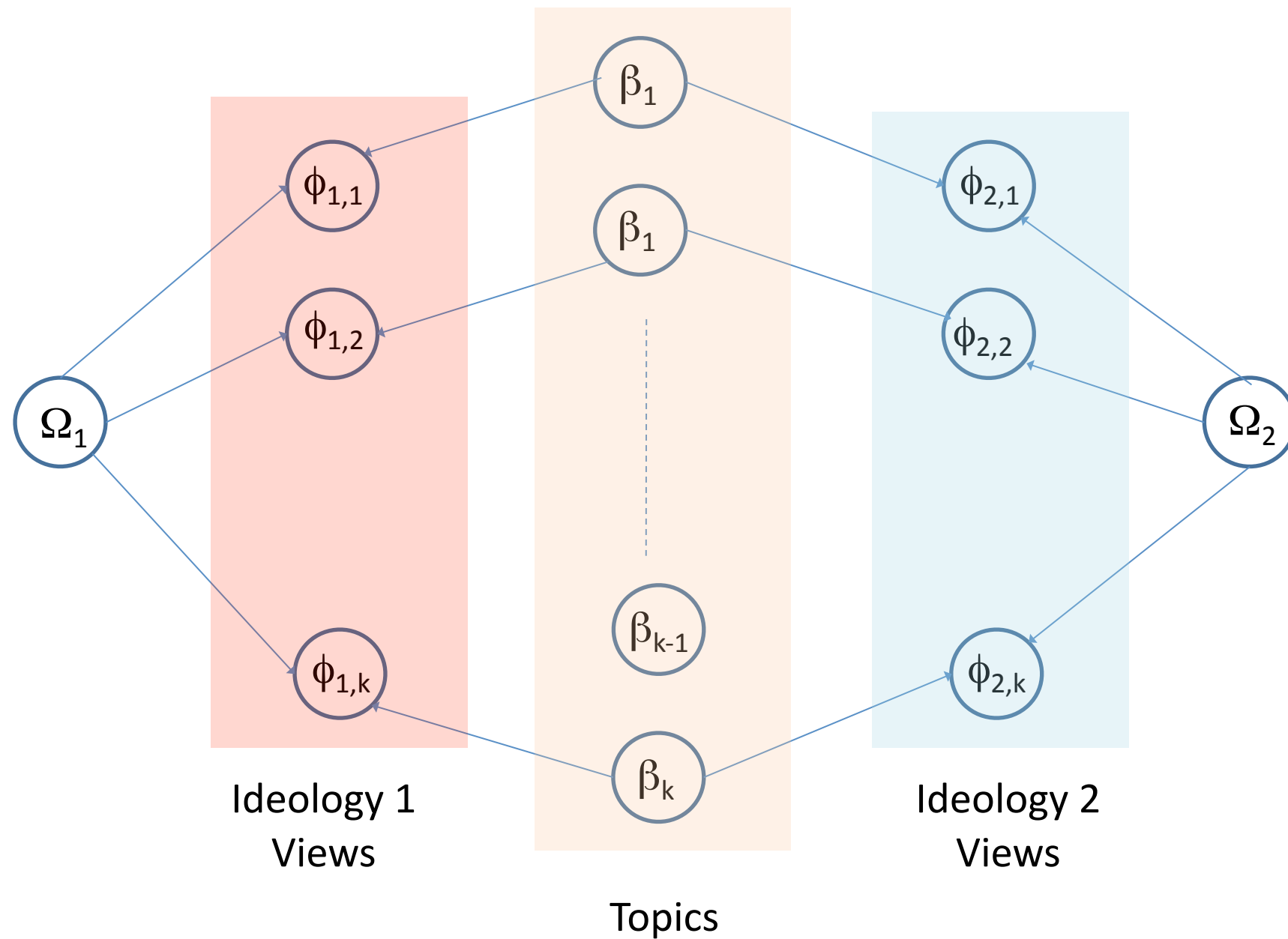
Visualization

Classification

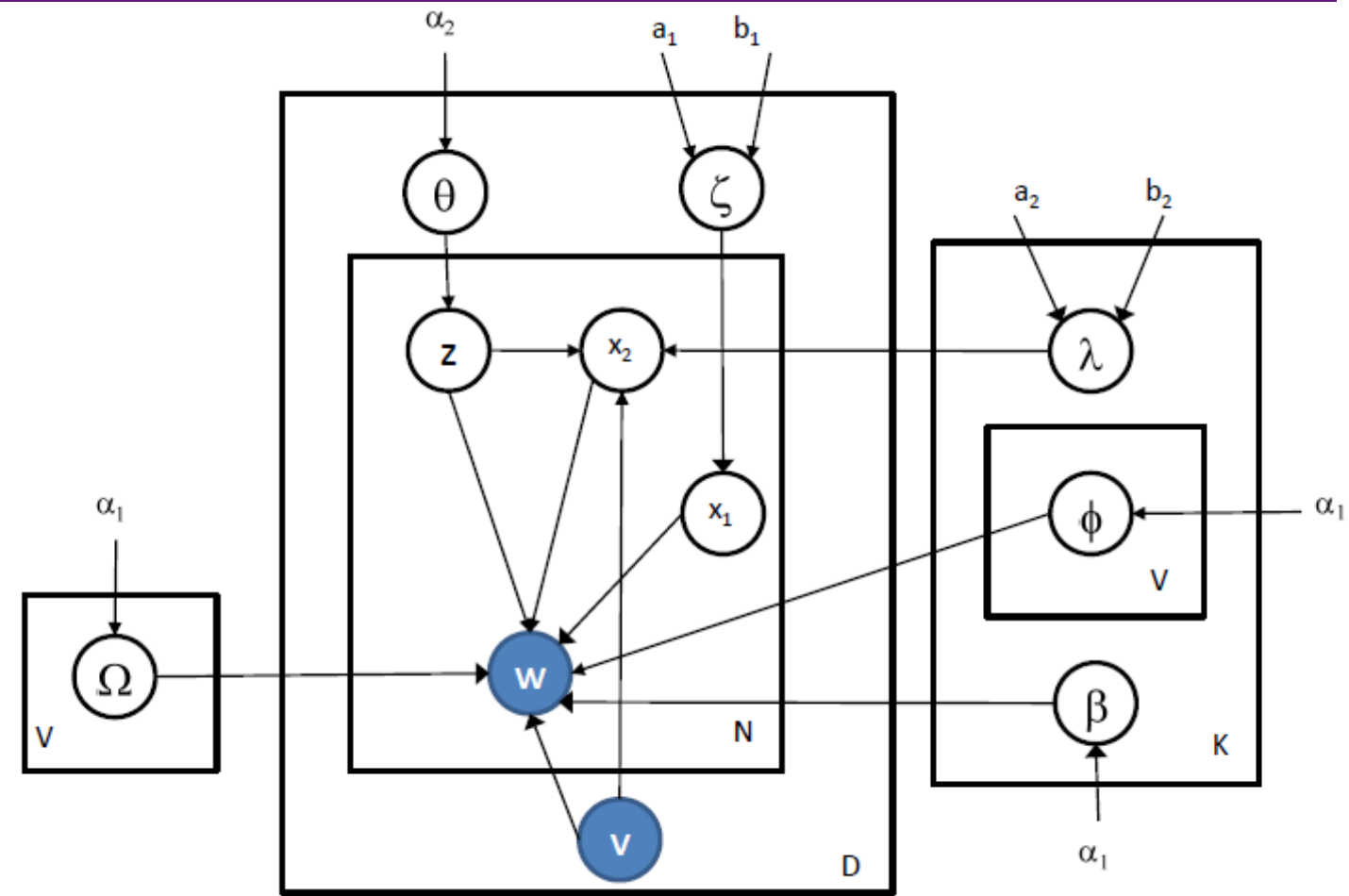
Structured browsing

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

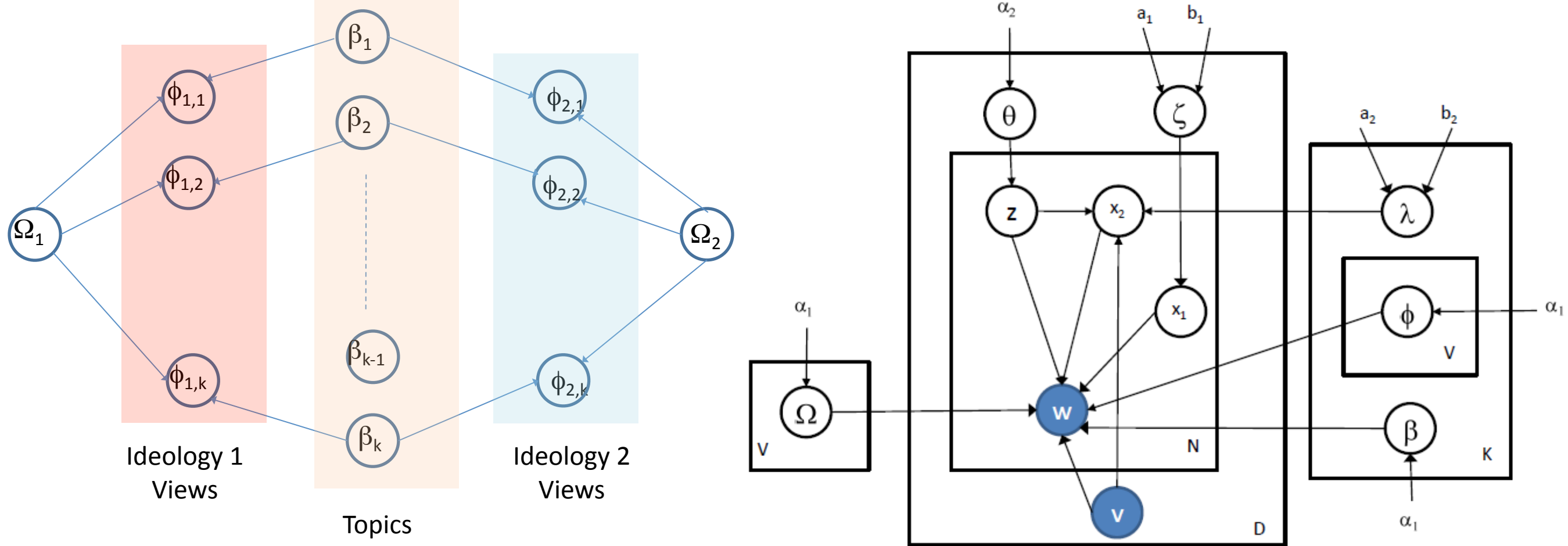
# Building a factored model



# Building a factored model

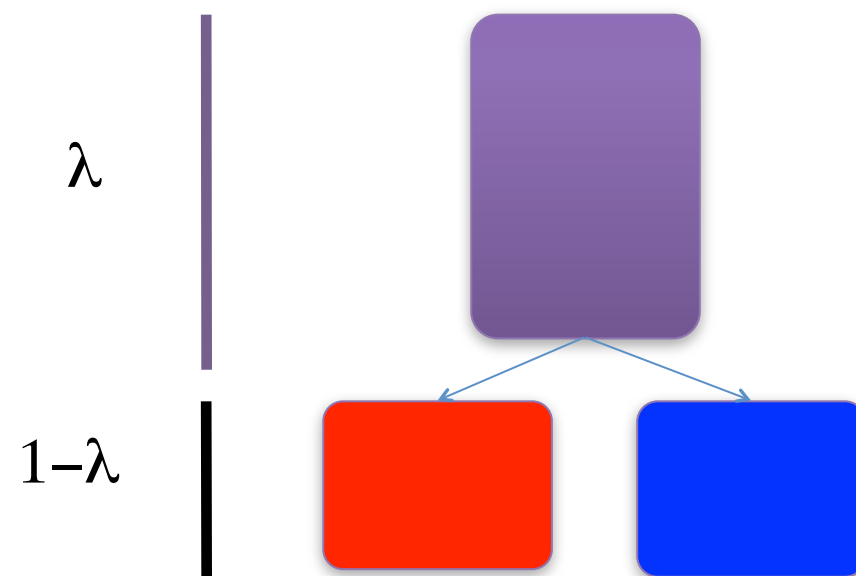
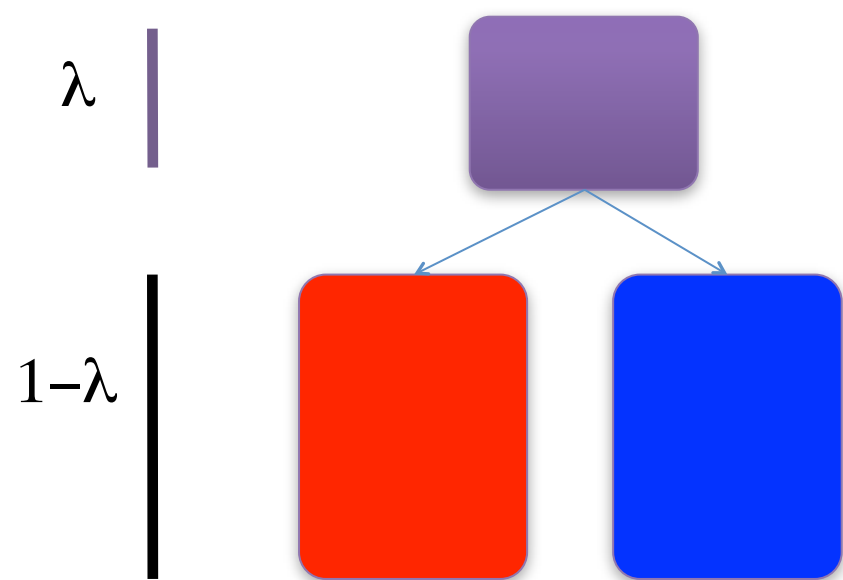
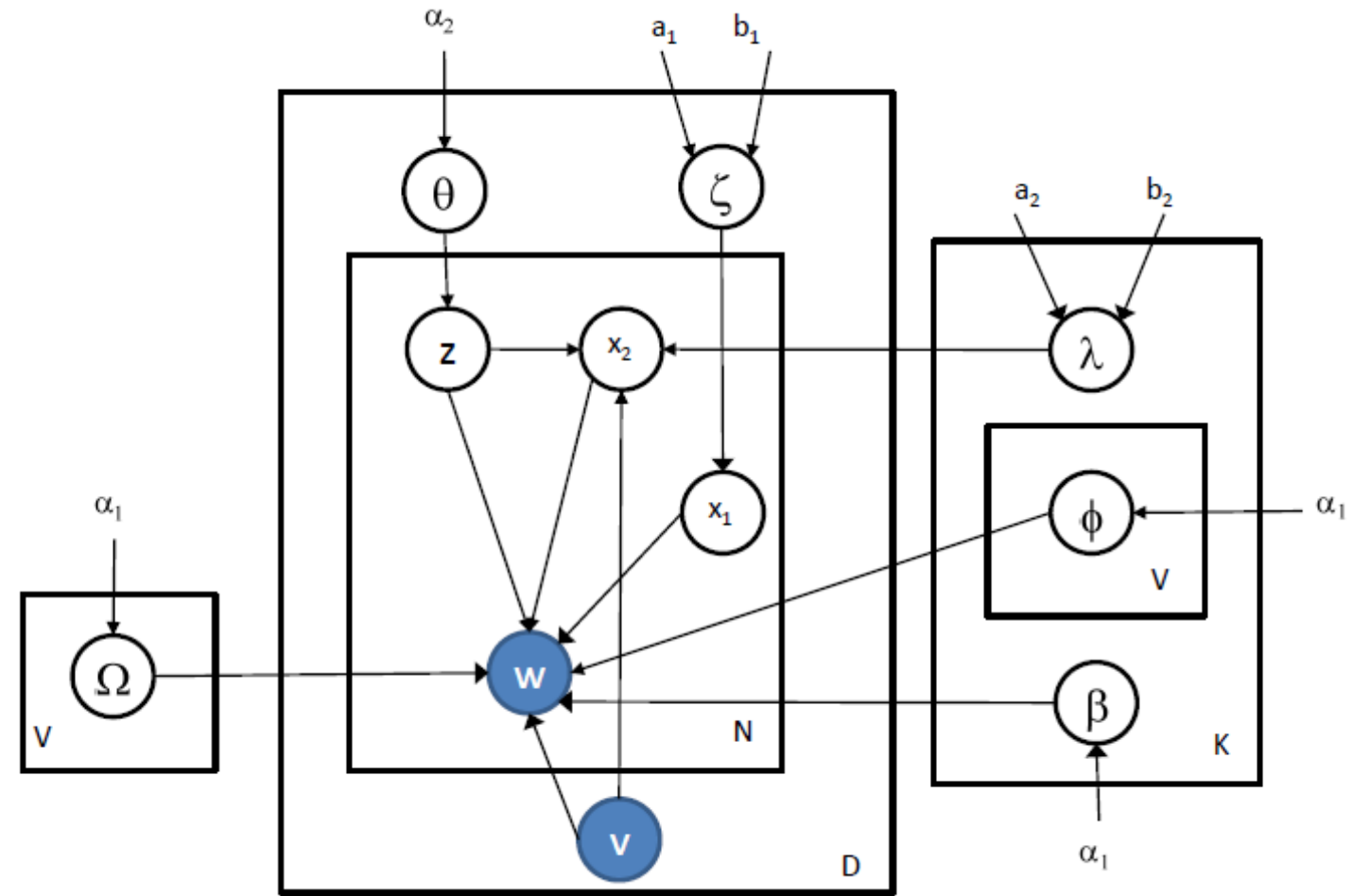
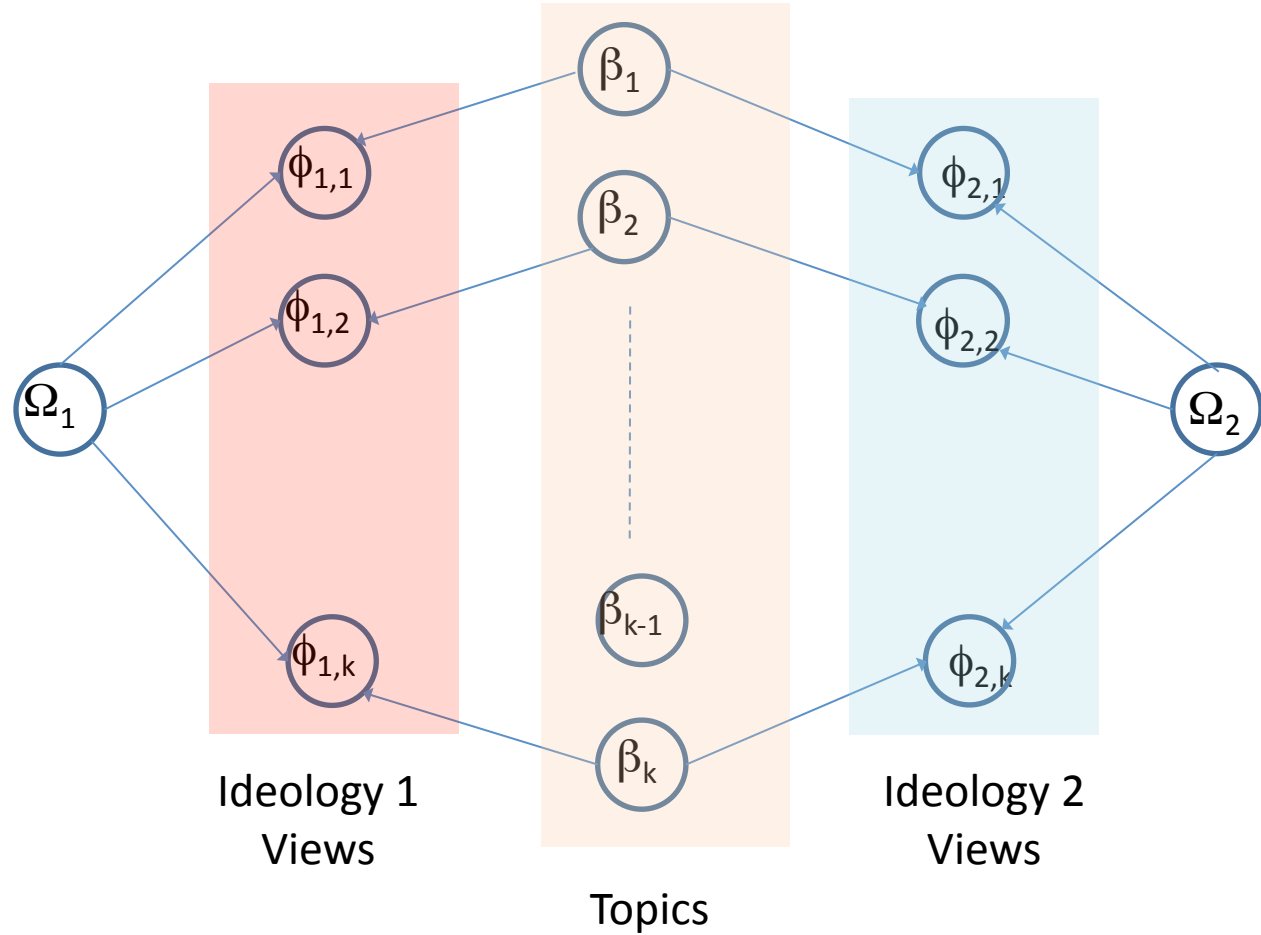


# Building a factored model





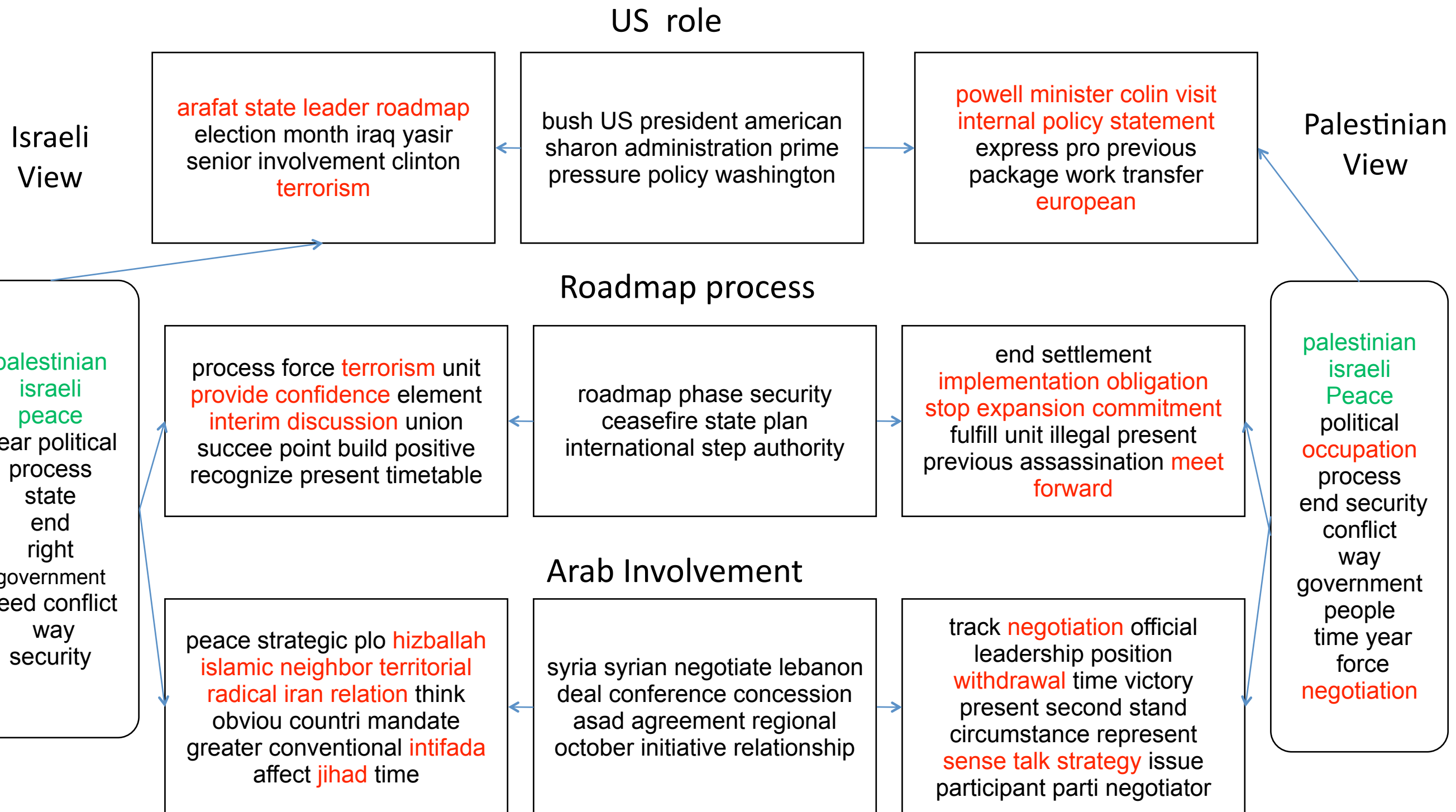
# Building a factored model



# Data

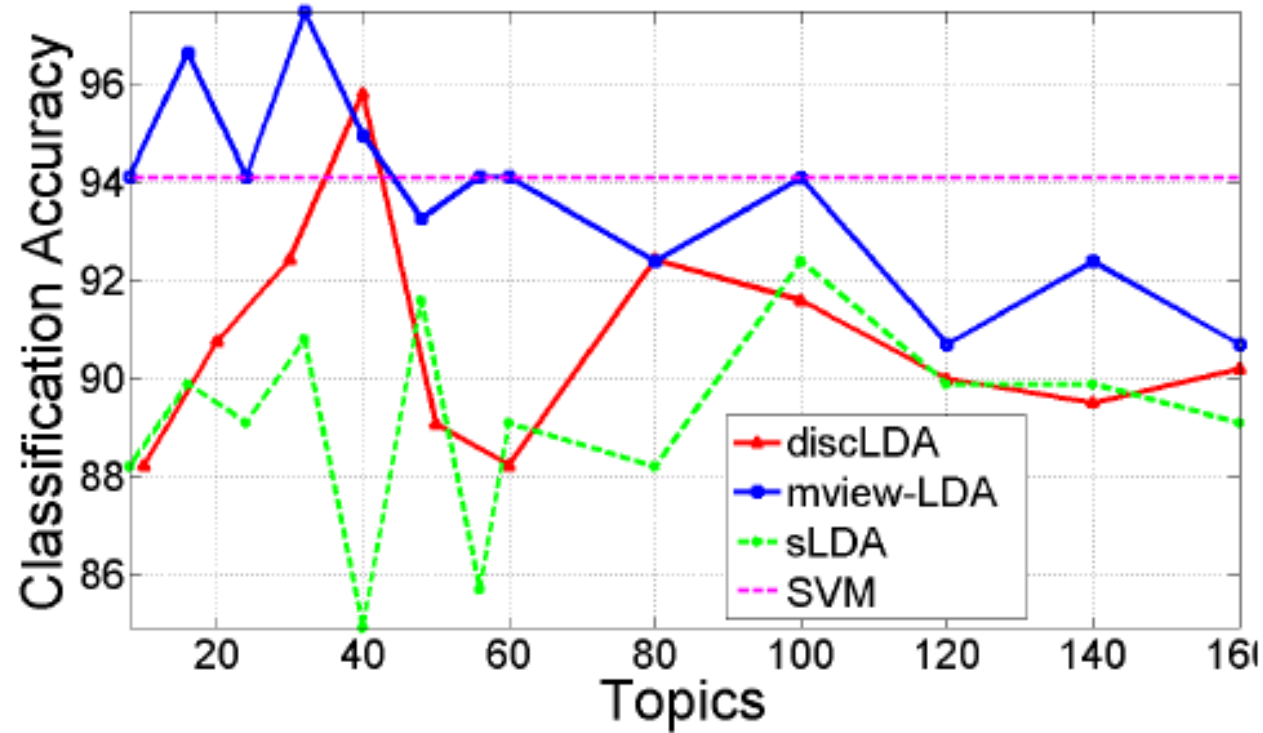
- **Bitterlemons:**
  - Middle-east conflict, document written by Israeli and Palestinian authors.
  - ~300 documents form each view with average length 740
  - Multi author collection
  - 80-20 split for test and train
- **Political Blog-1:**
  - American political blogs (Democrat and Republican)
  - 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

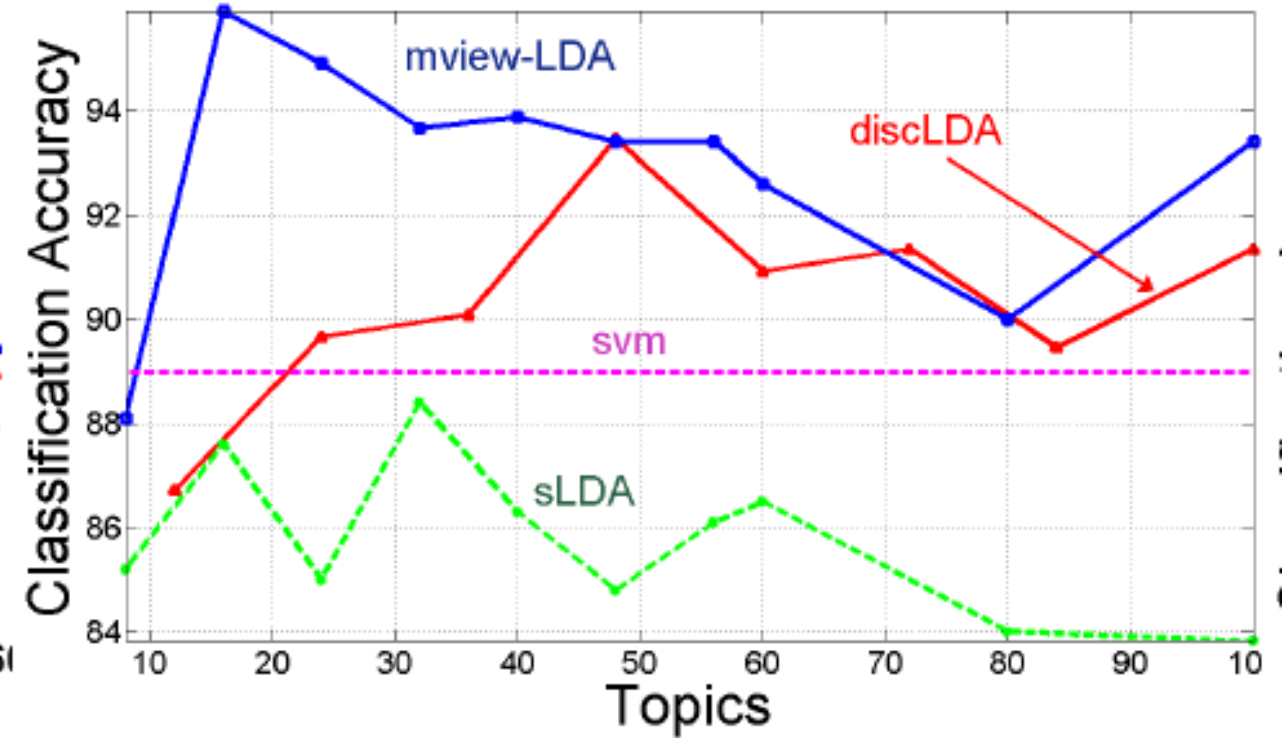


# Classification accuracy

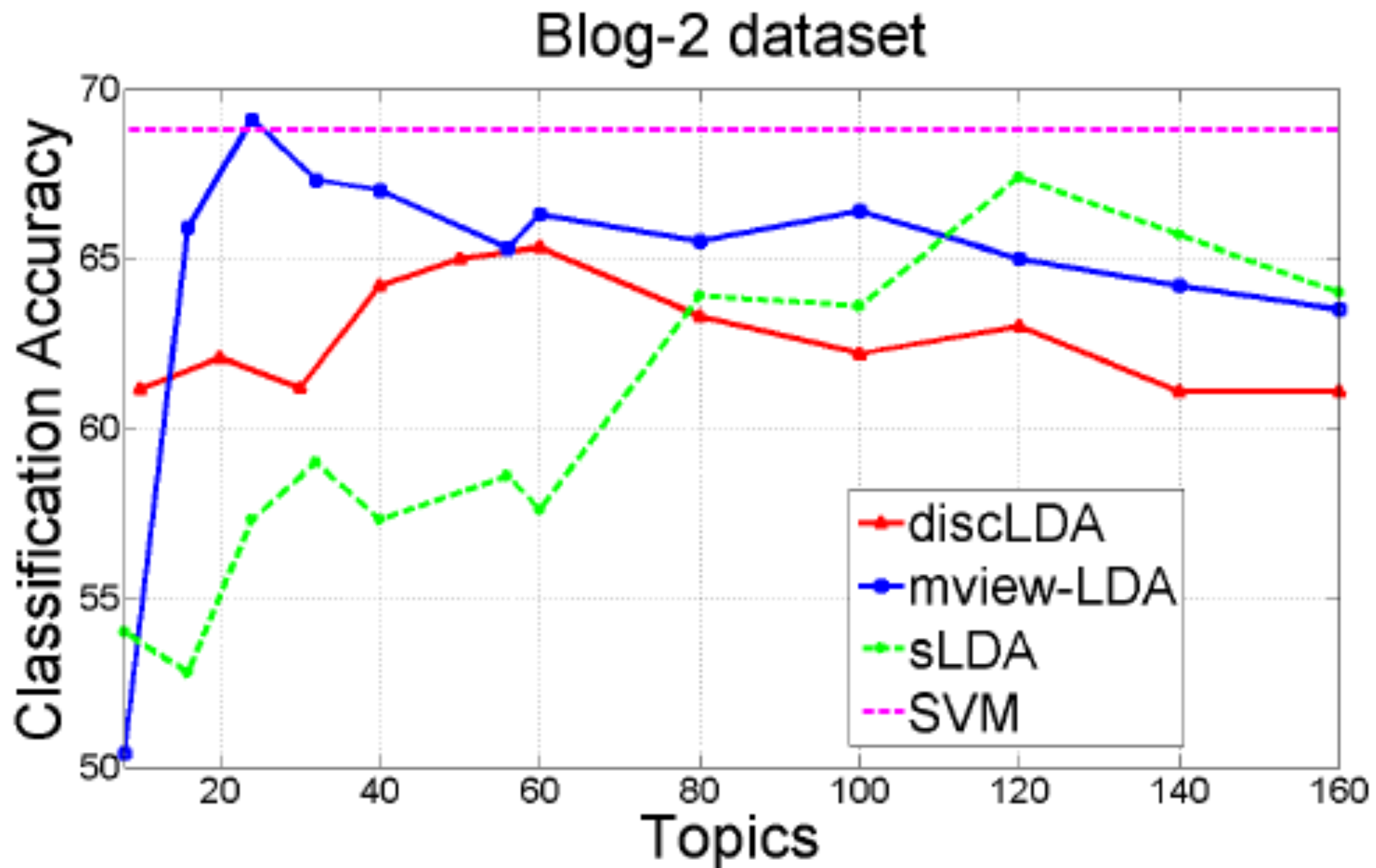
Bitterlemons dataset



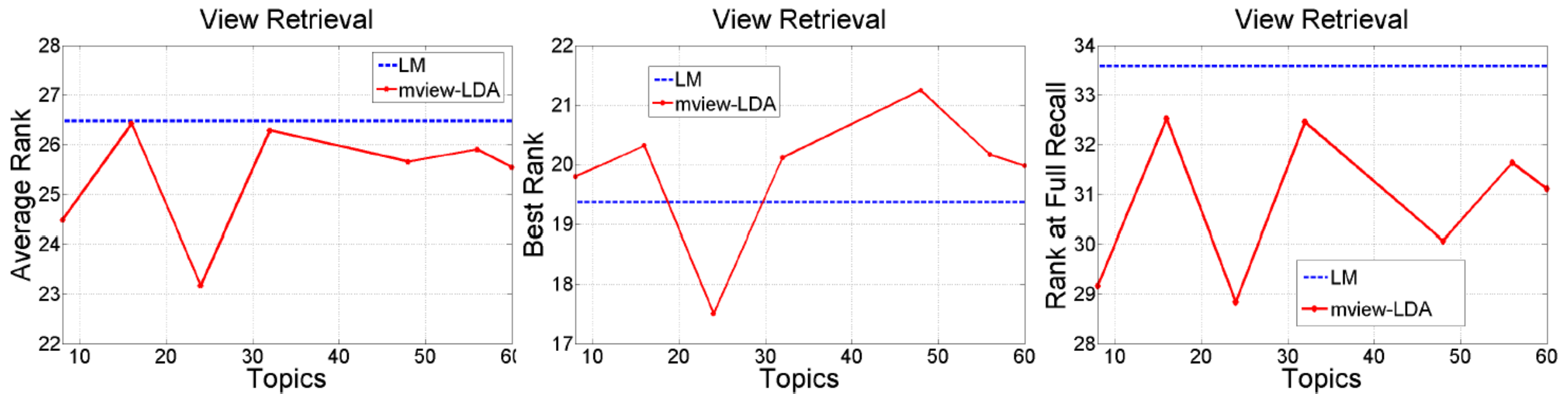
Blog-1 dataset



# Generalization to new blogs

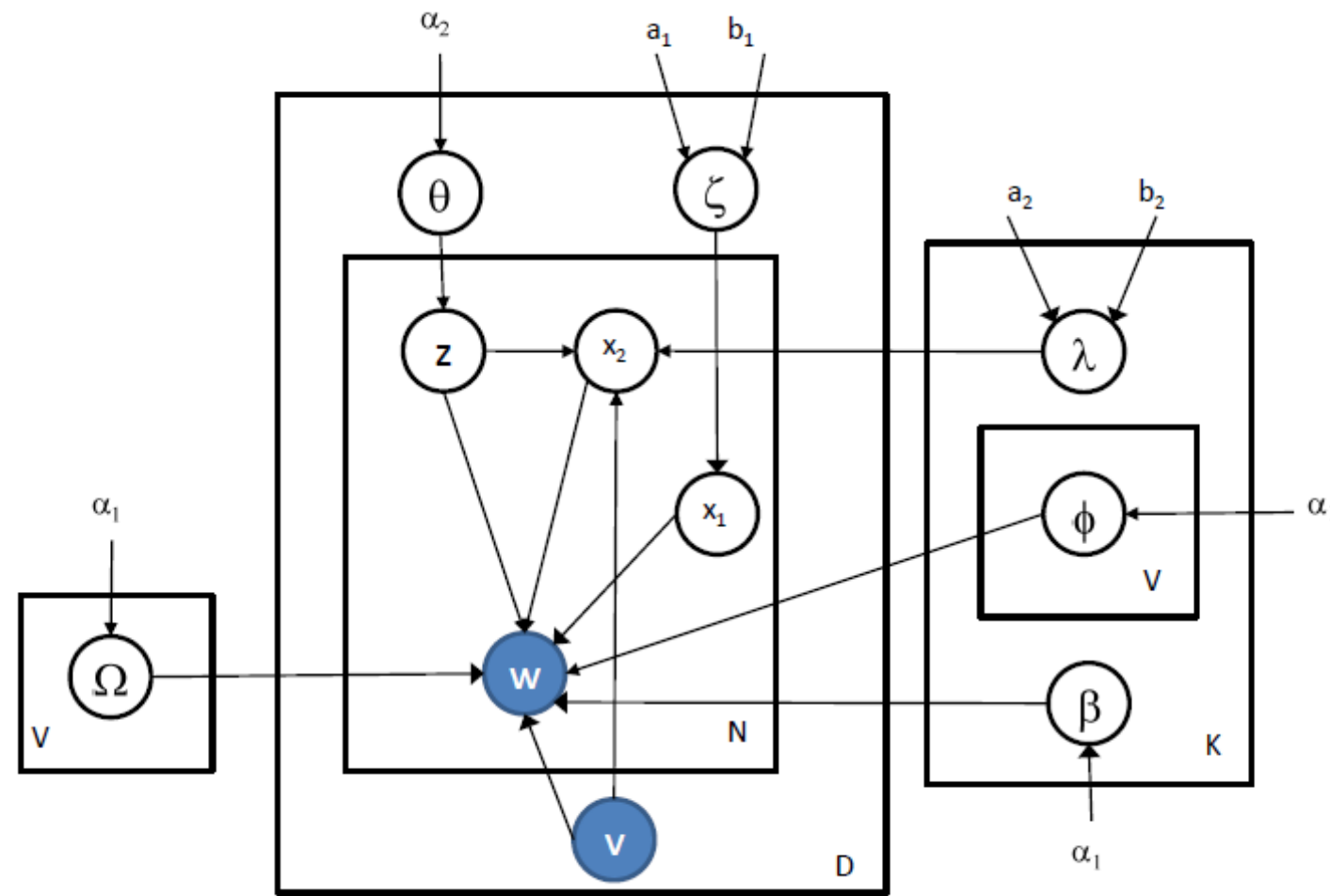


# Finding alternate views

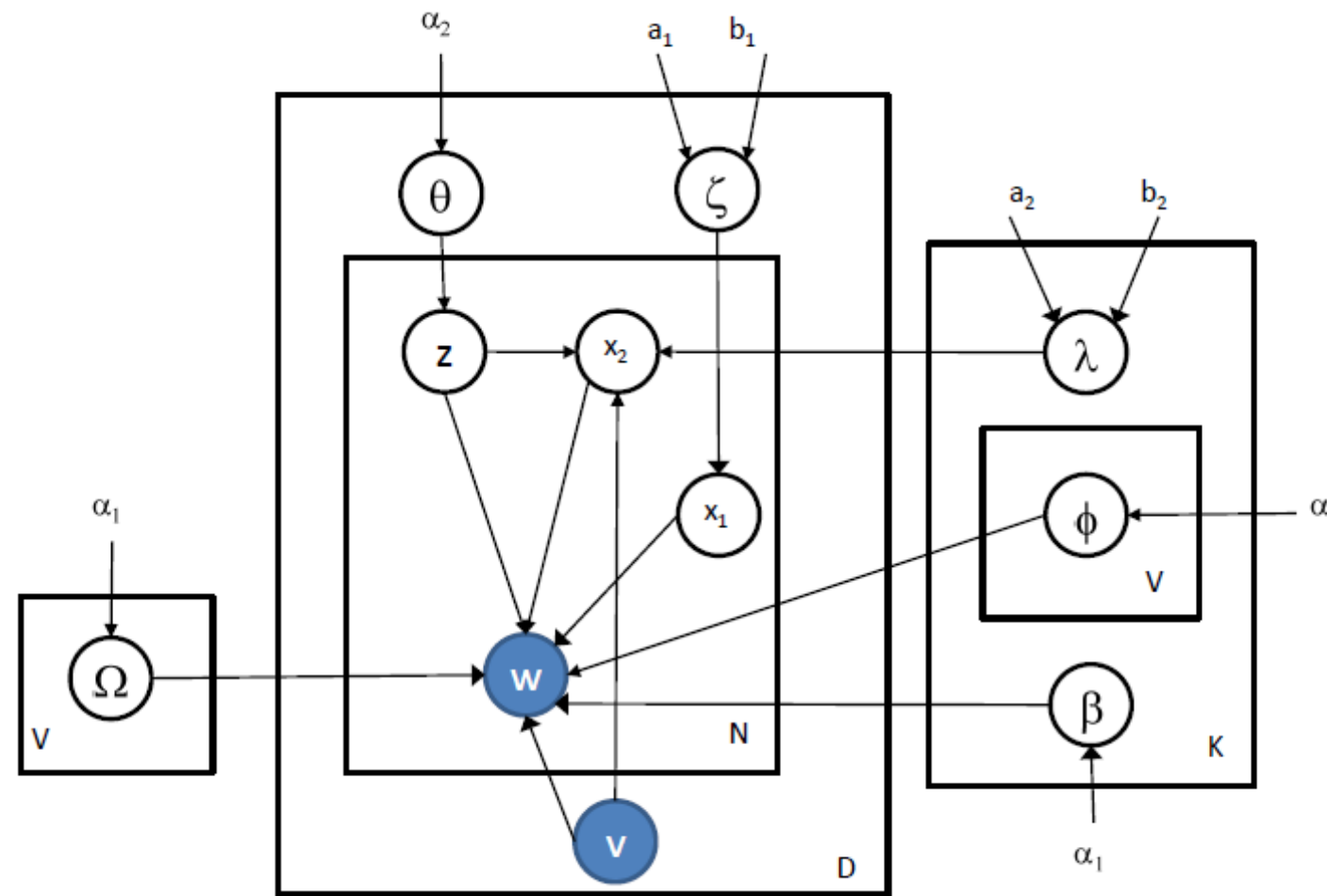


- Given a document written in one ideology, retrieve the equivalent
- Baseline: SVM + cosine similarity

# Unlabeled data



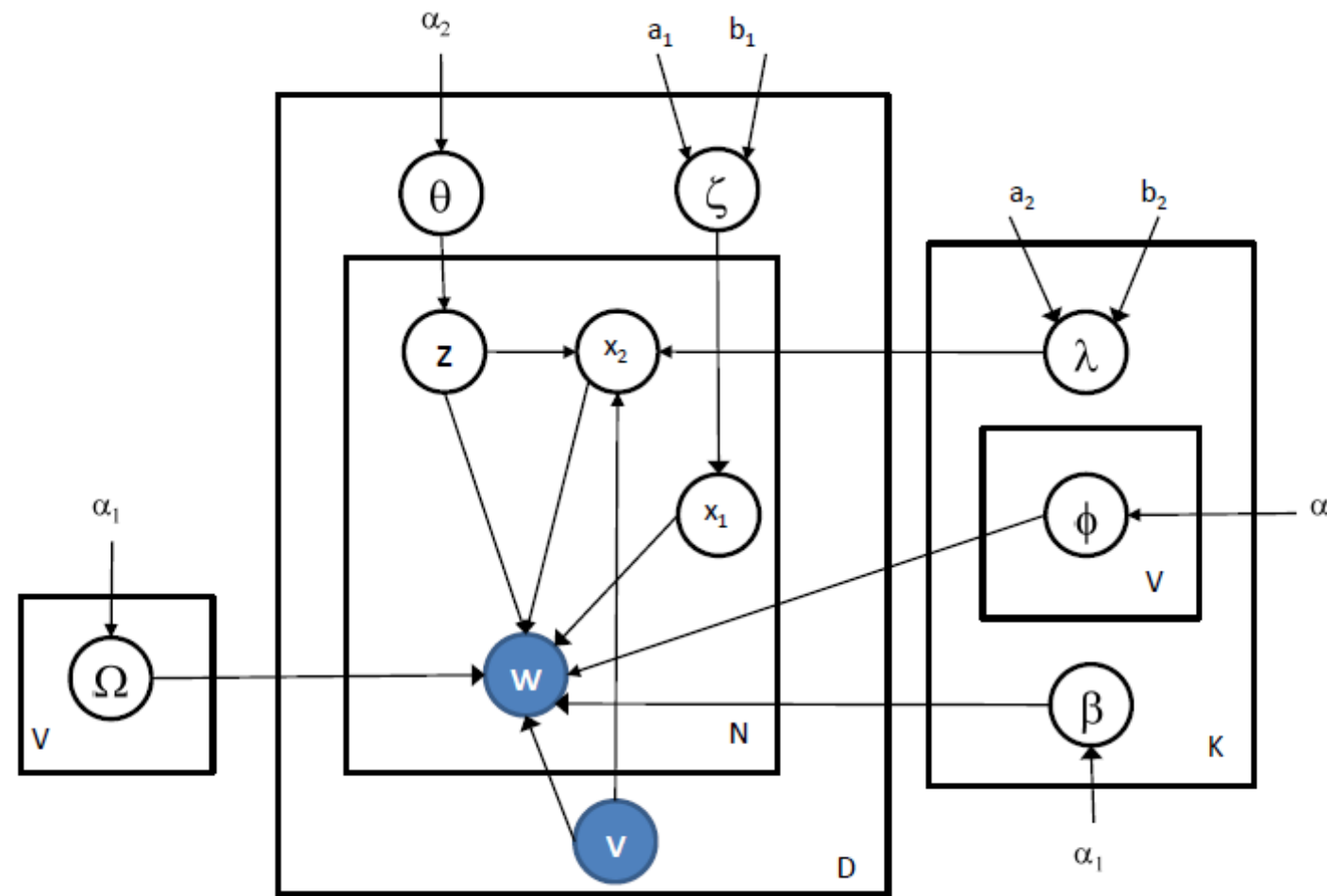
# Unlabeled data



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

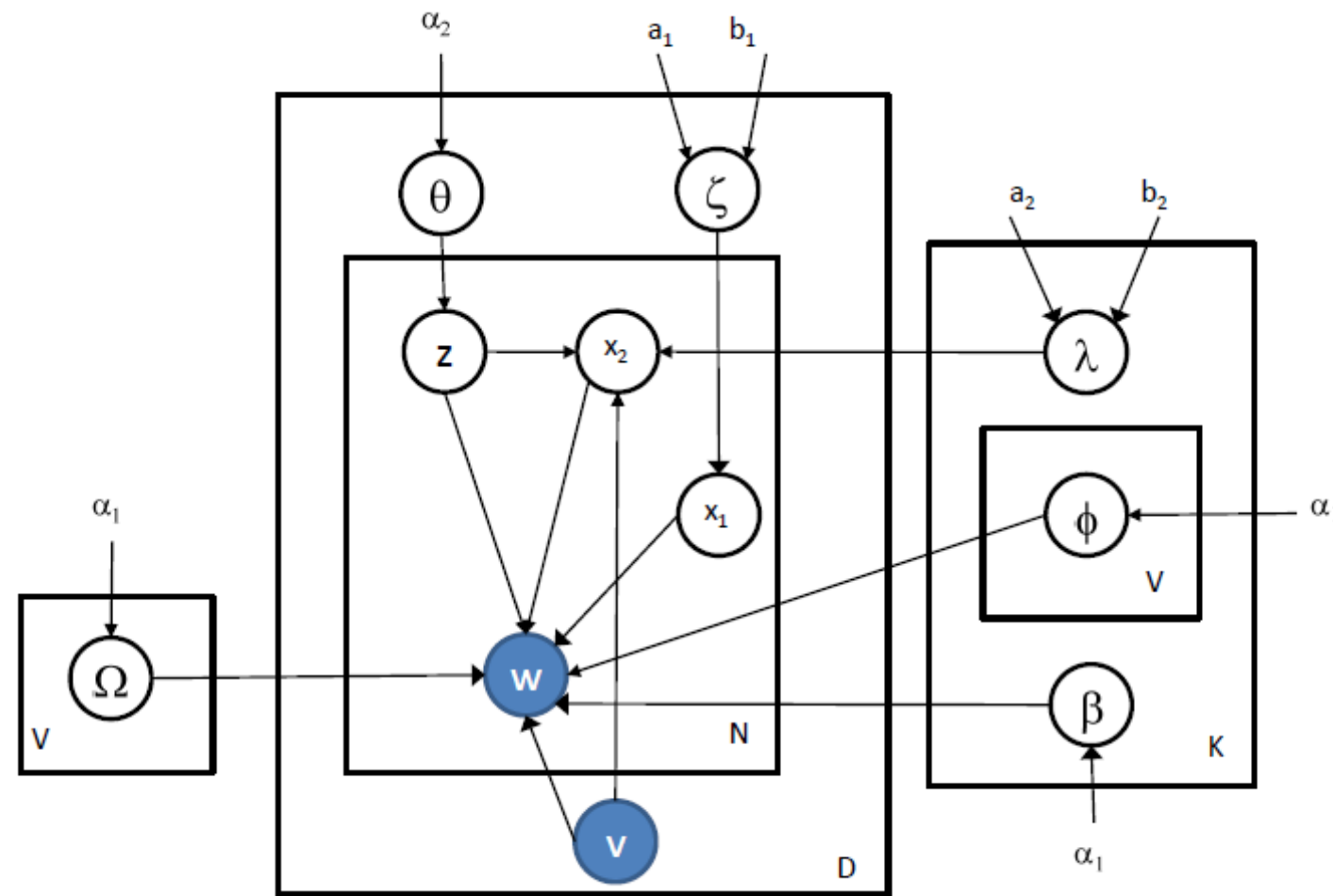


# Unlabeled data



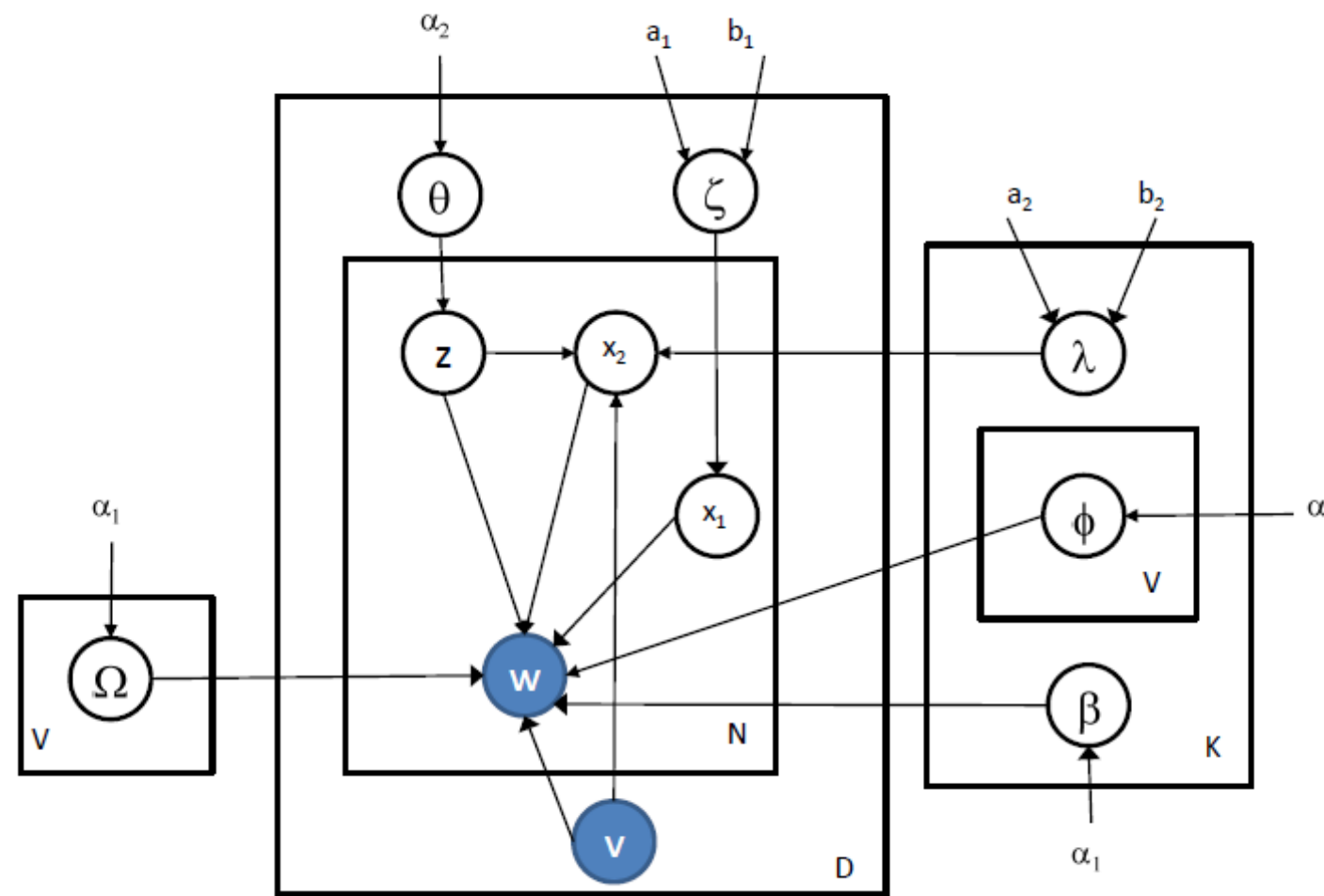
- In theory this is **simple**
  - Add a step that samples the document view ( $v$ )
  - **Doesn't mix** in practice because tight coupling between  $v$  and  $(x_1, x_2, z)$
- Solution

# Unlabeled data



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

# Unlabeled data



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

# Part 7 - Undirected Graphical Models

**Review**

# YAHOO!

Spam  
Filtering

Classification

Exploration

Segmentation

Annotation

System  
Design

Prediction  
(time series)

Clustering

Document  
Understanding

User  
Modeling

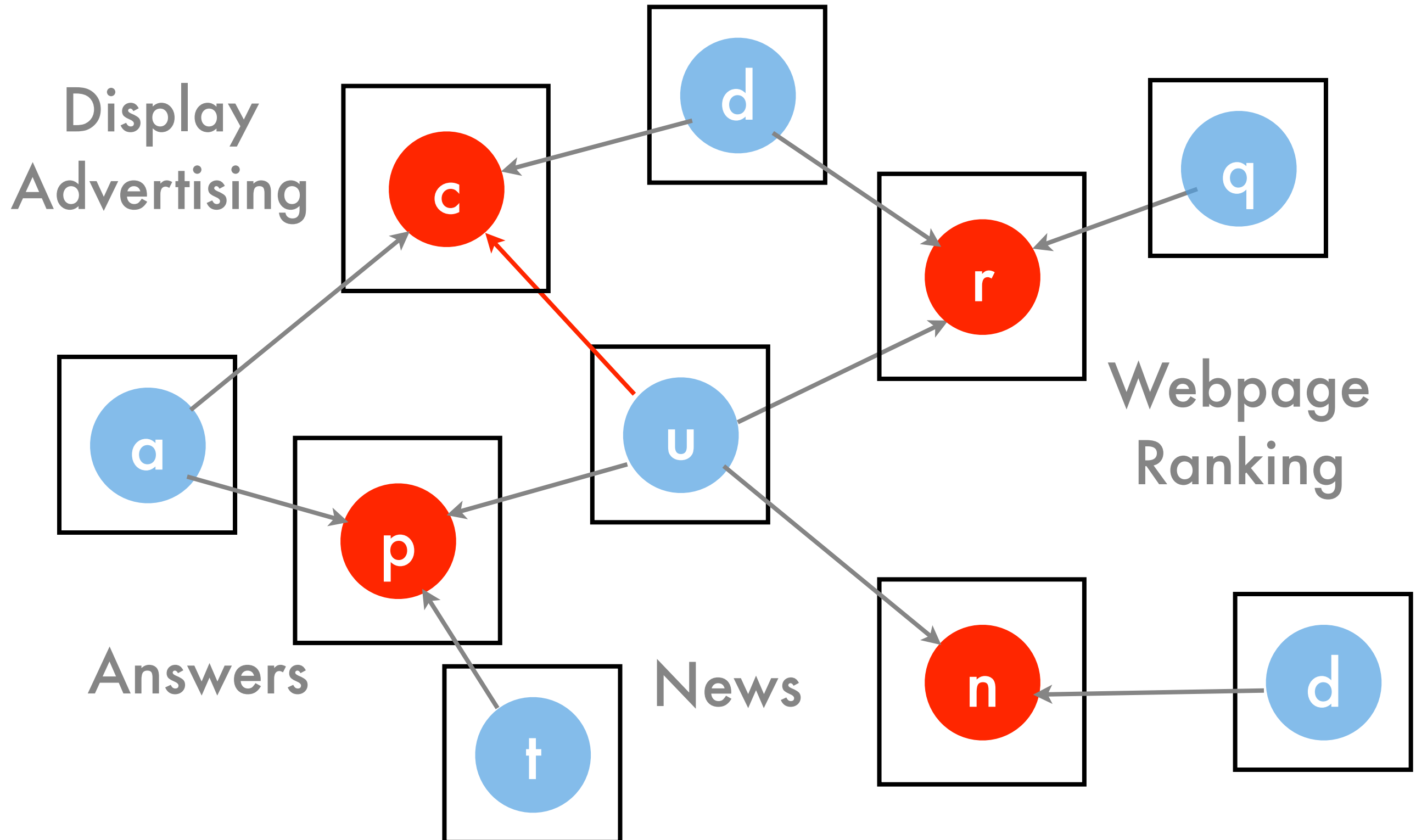
Advertising

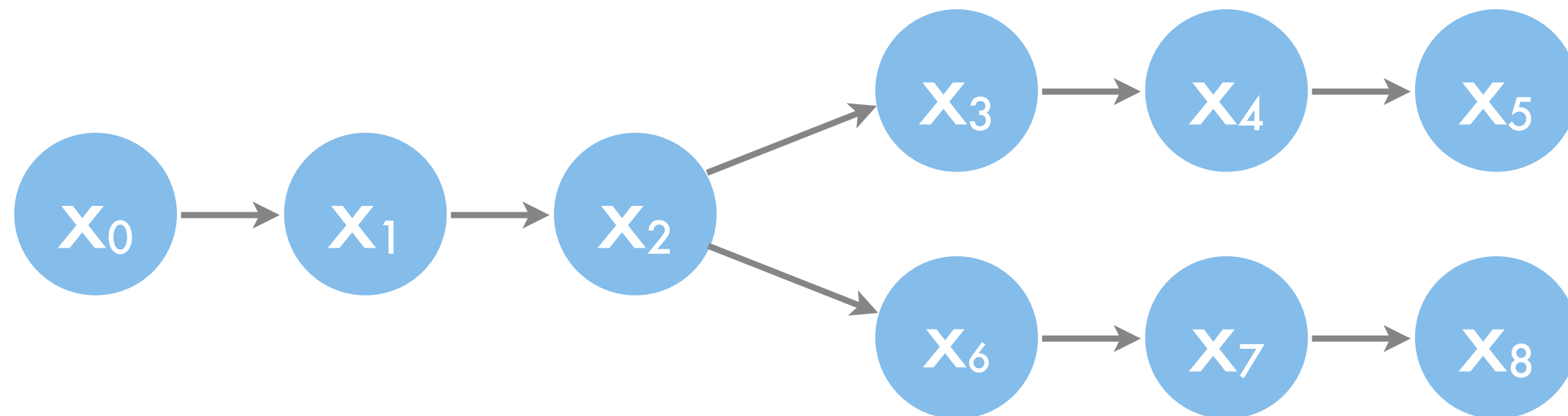
Novelty  
Detection

Performance  
Tuning

Debugging

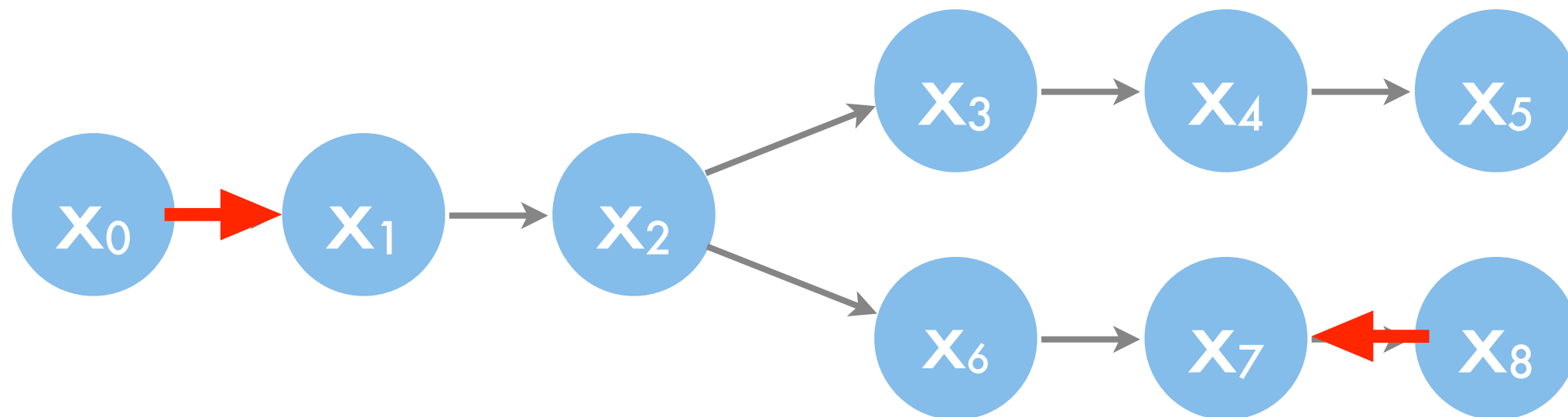
# Data Integration



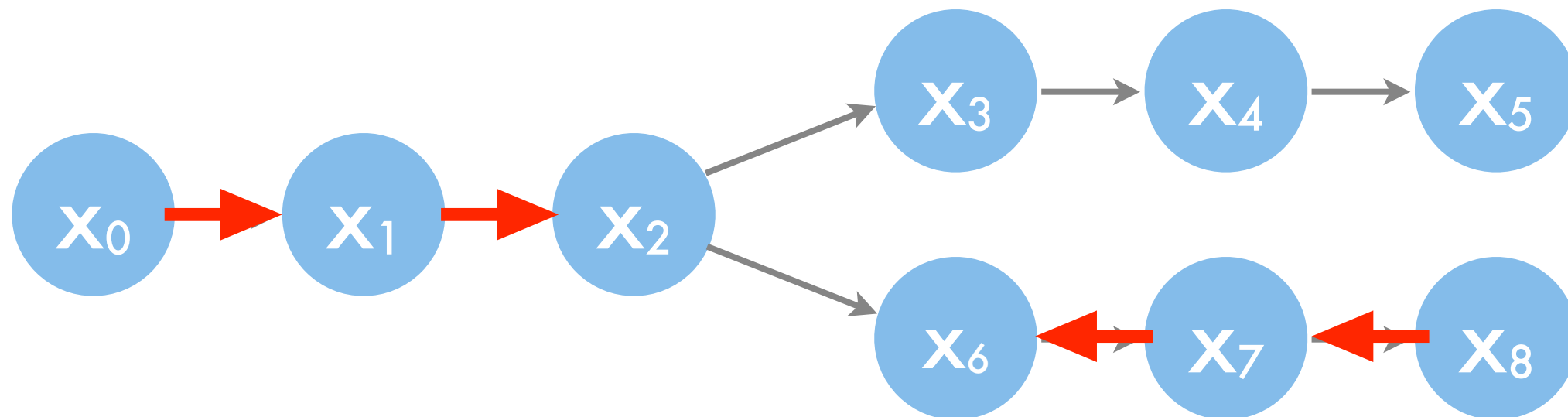


- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether

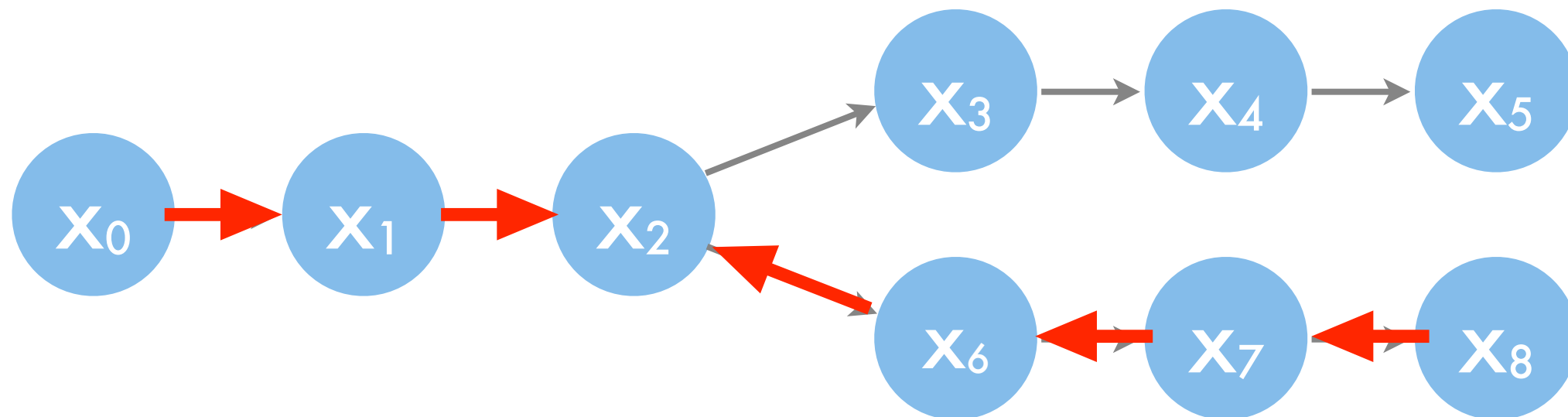




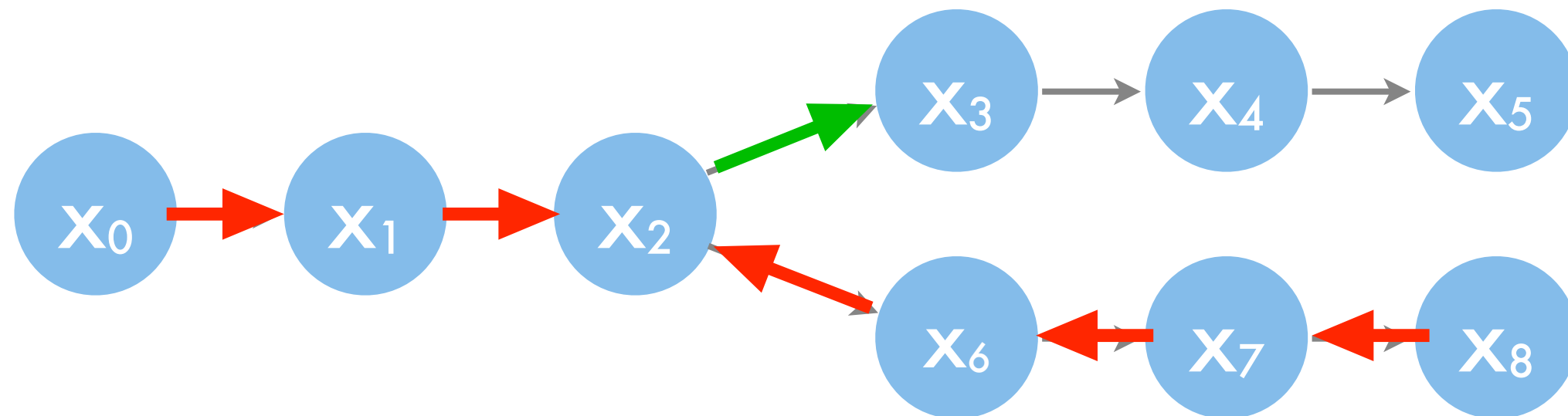
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether



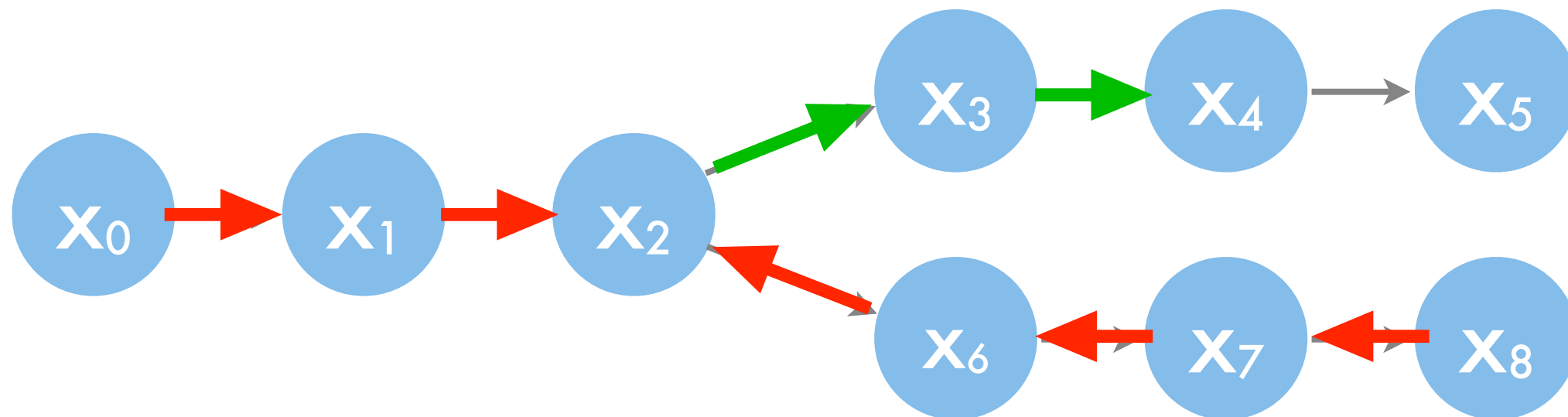
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether



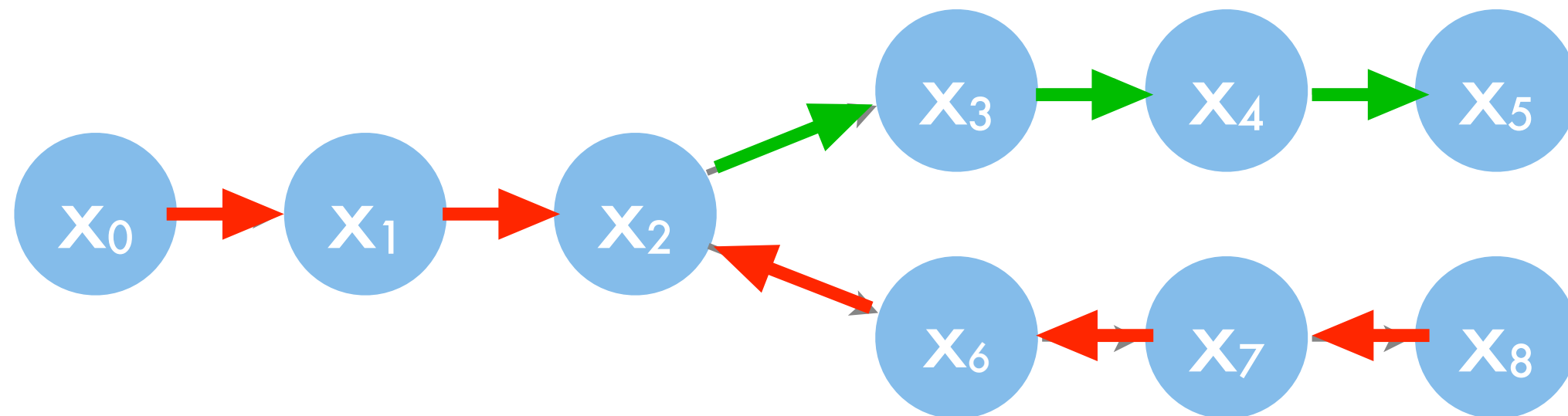
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether



- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether



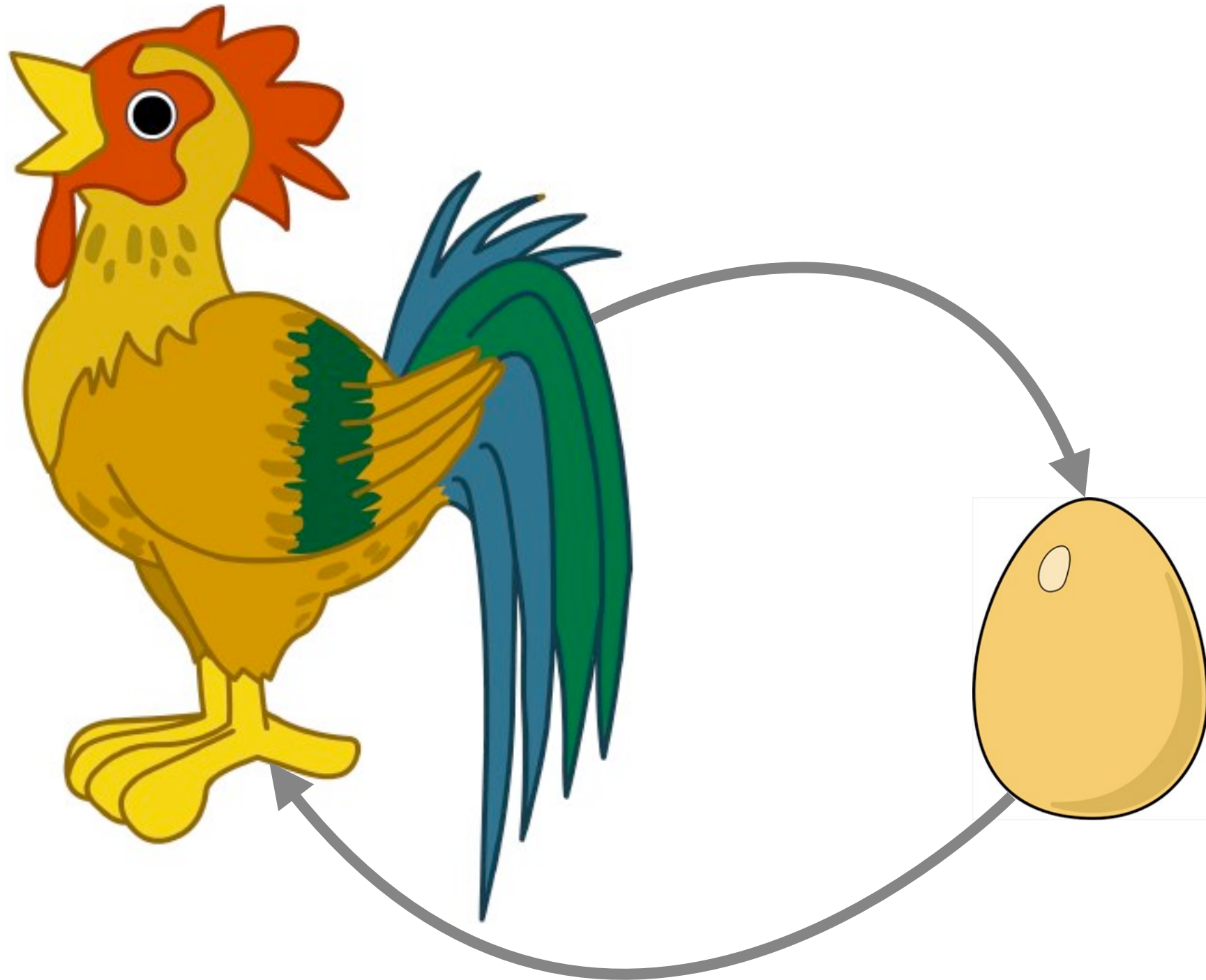
- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether



- Forward/Backward messages as normal for chain
- When we have more edges for a vertex use
  - For each outgoing message, send it once you have all other incoming messages
  - **PRINCIPLED HACK**  
If no message received yet, set it to 1 altogether

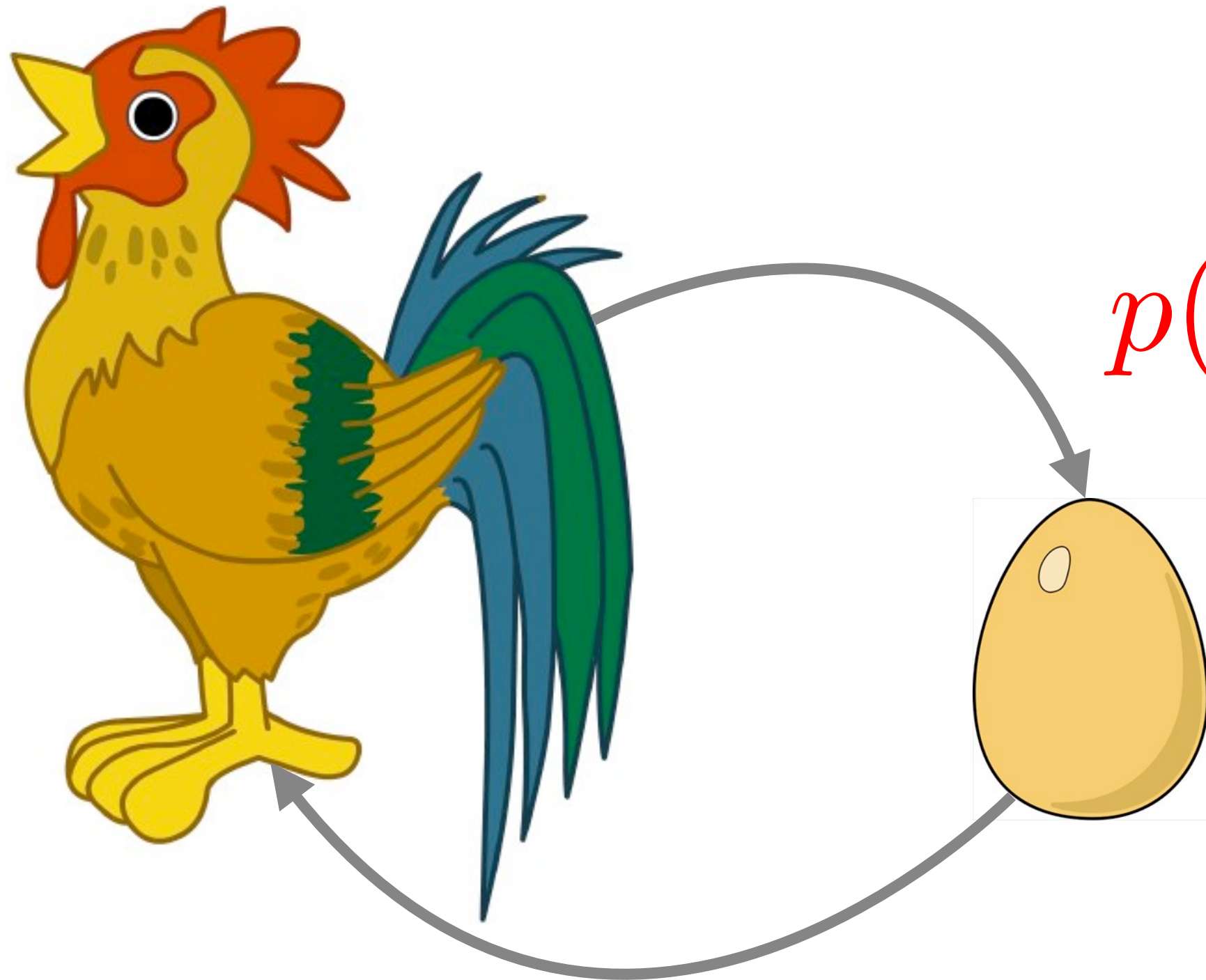
**Blunting the arrows ...**

# Chicken and Egg



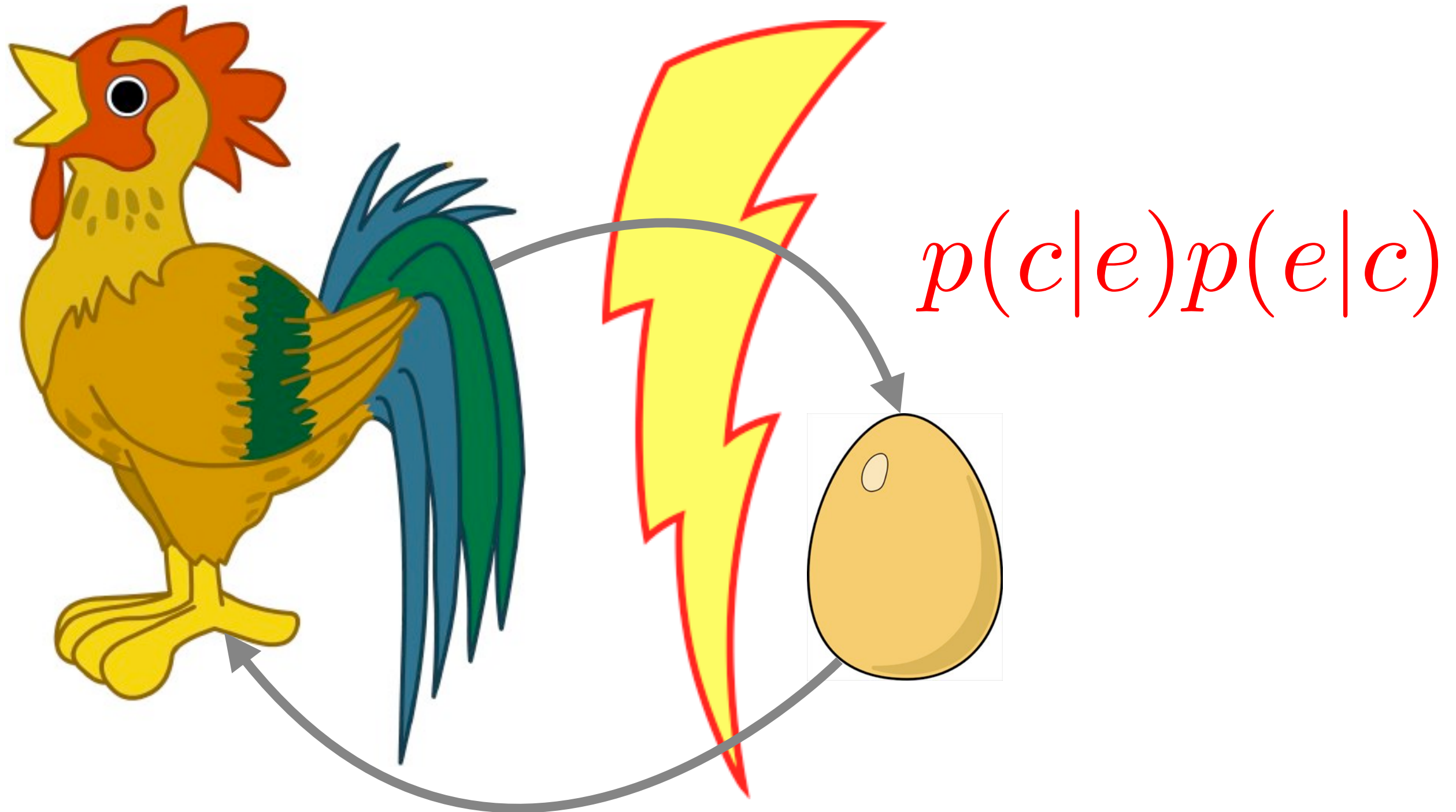


# Chicken and Egg

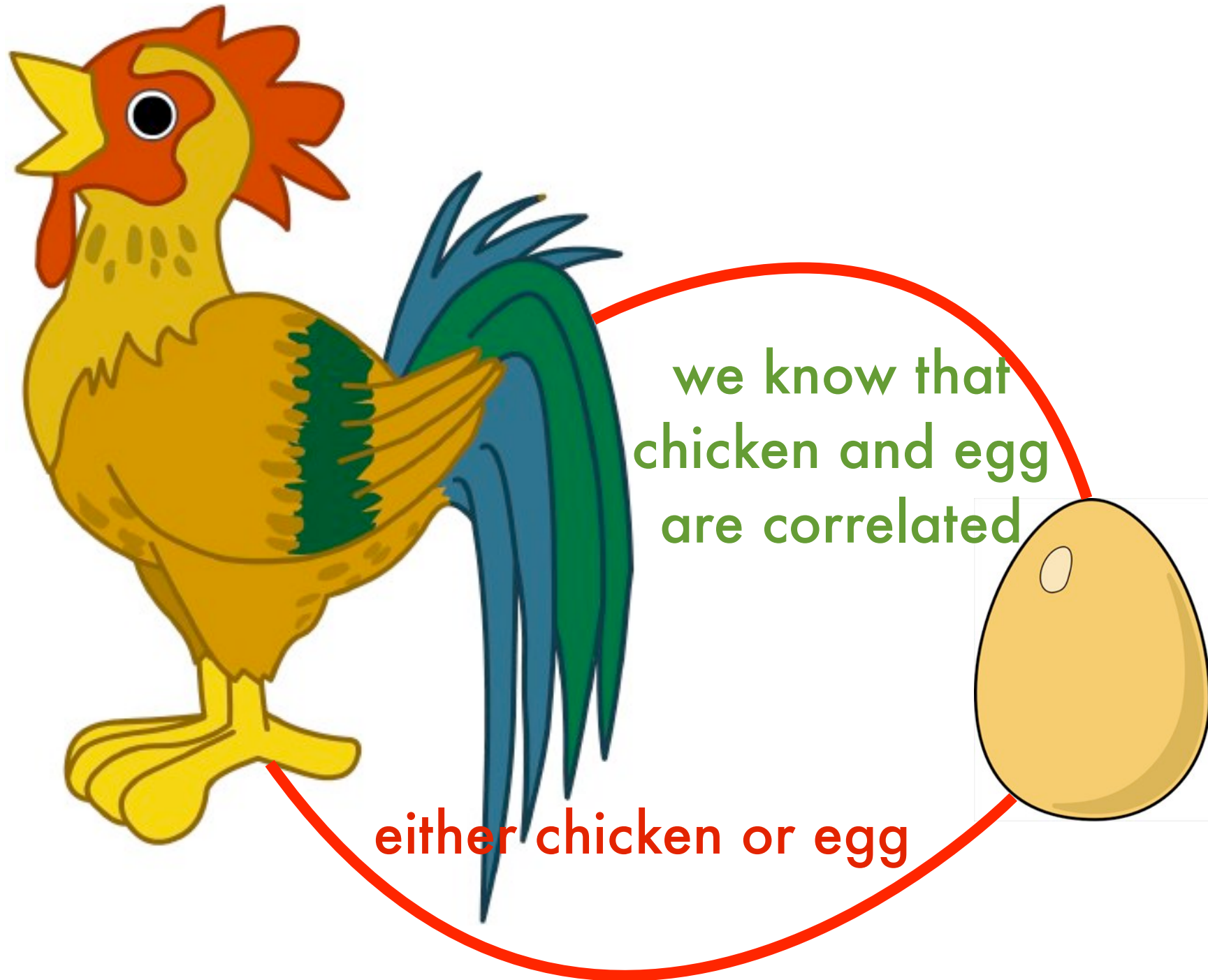


$$p(c|e)p(e|c)$$

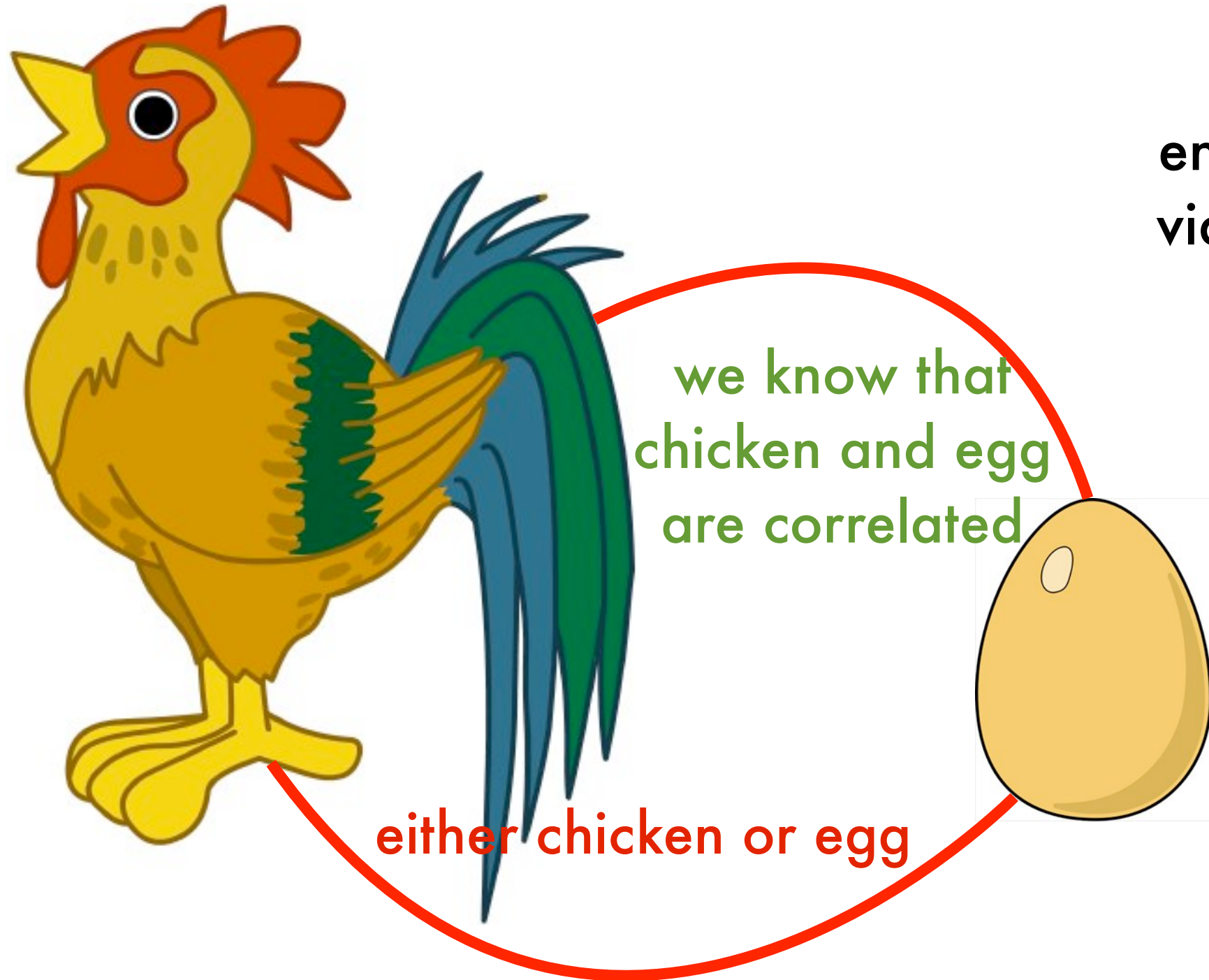
# Chicken and Egg



# Chicken and Egg



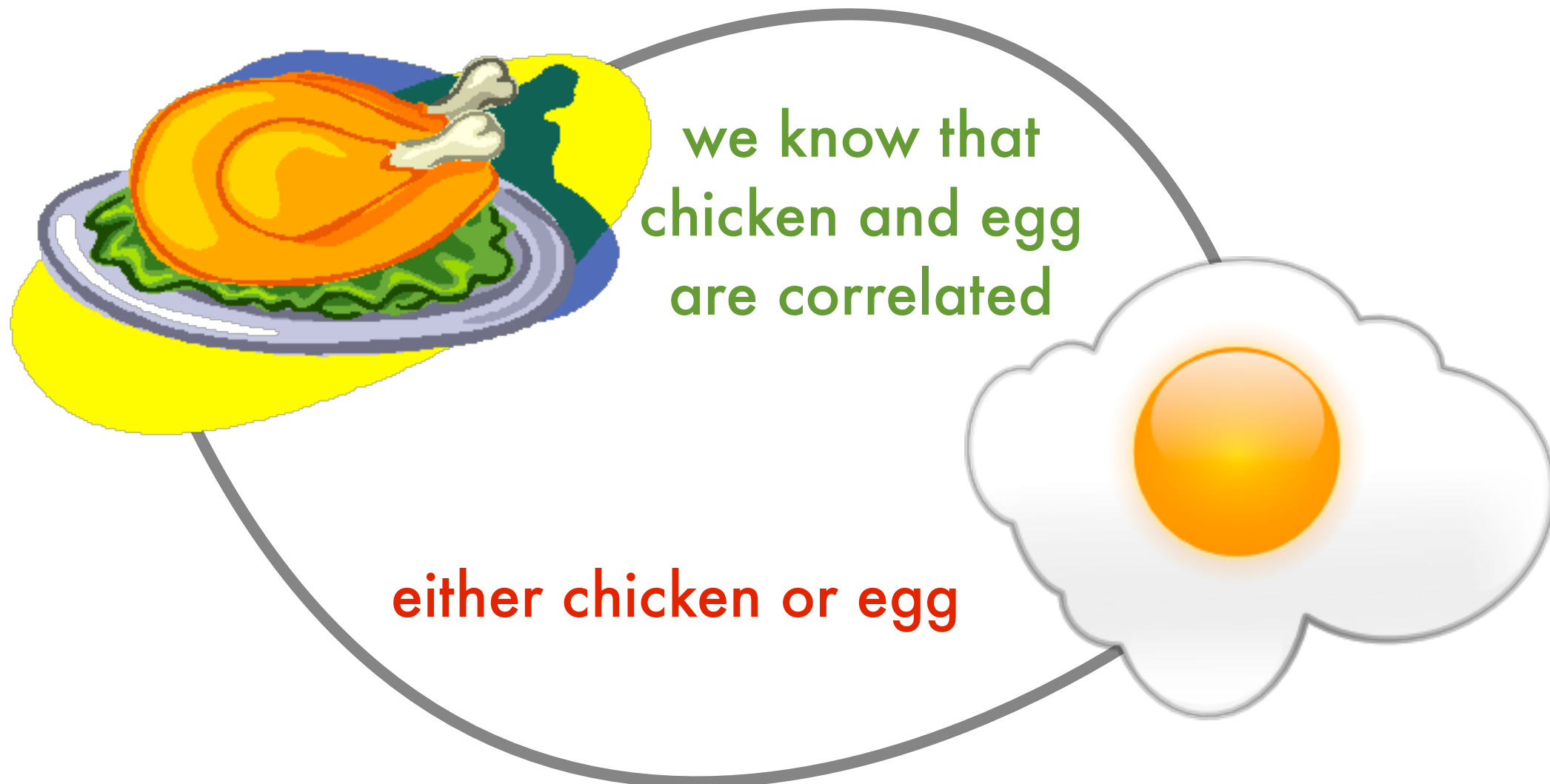
# Chicken and Egg



encode the correlation  
via the clique potential  
between  $c$  and  $e$

$$p(c, e) \propto \exp \psi(c, e)$$

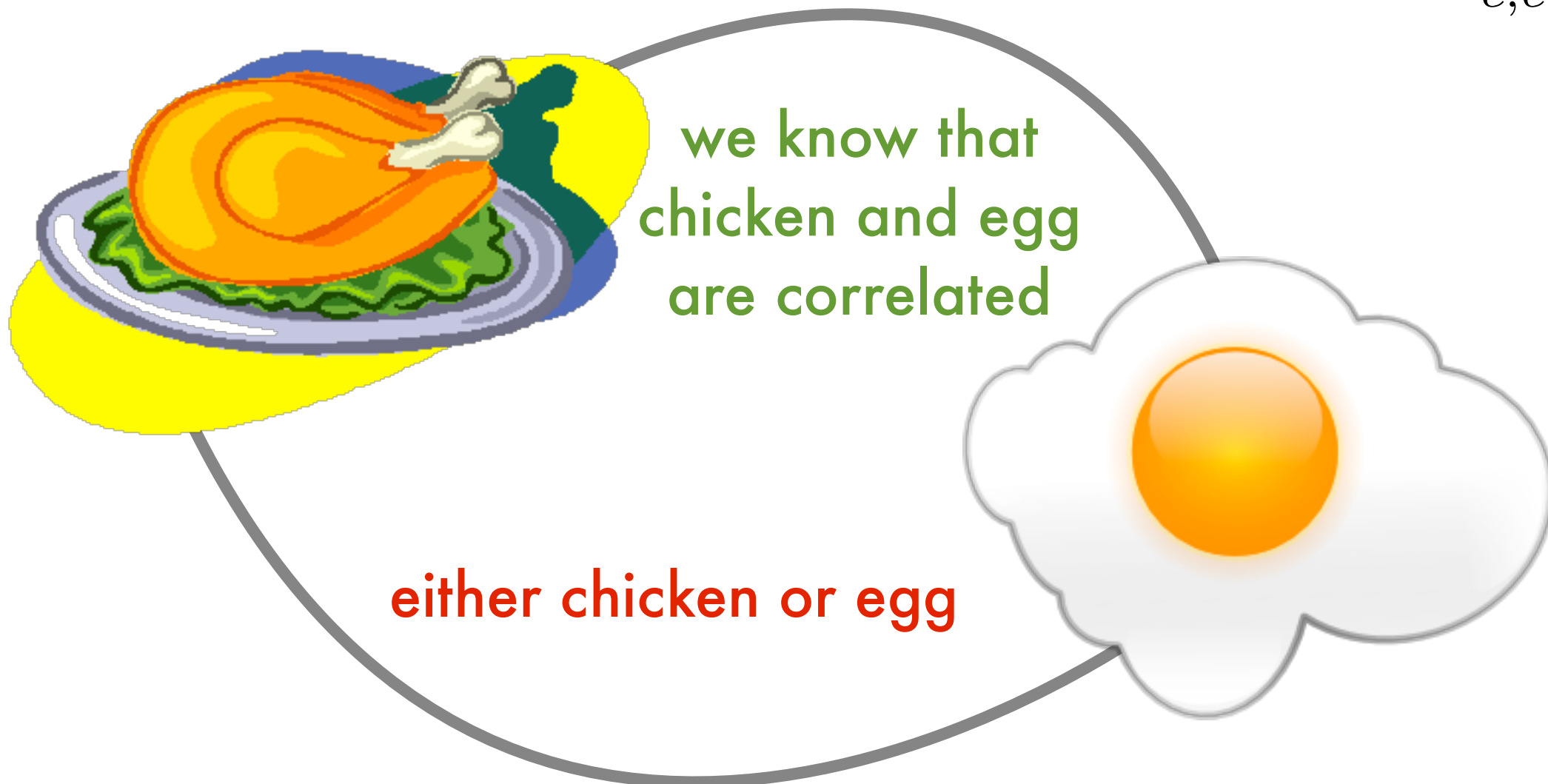
# Chicken and Egg



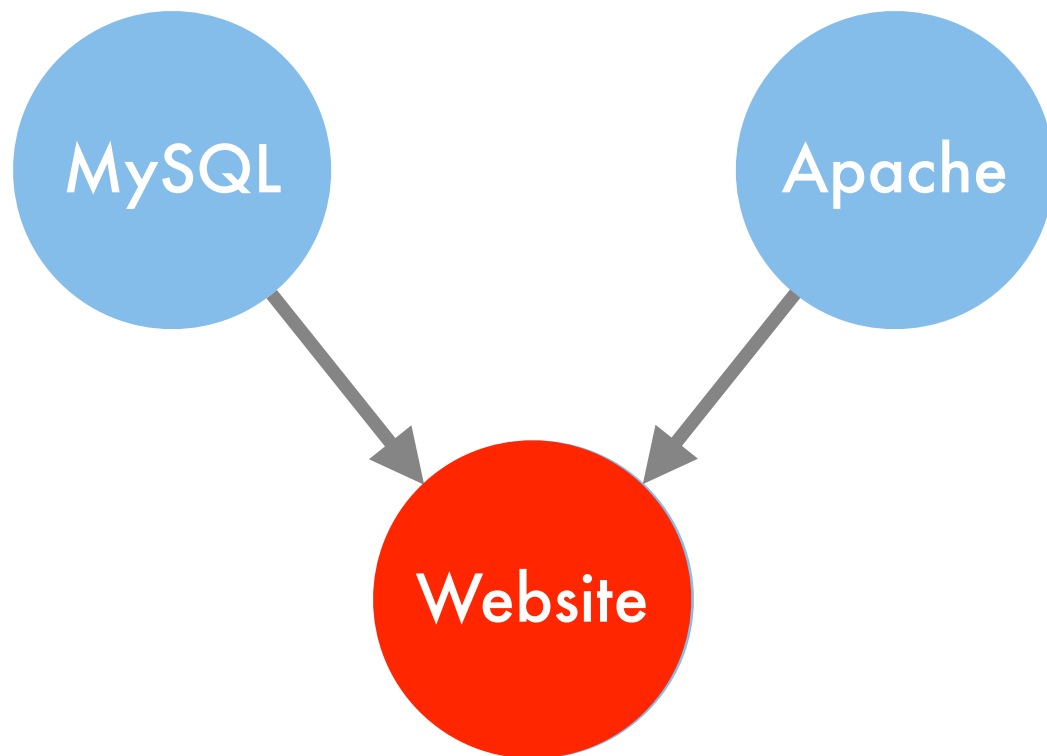


# Chicken and Egg

$$p(c, e) = \frac{\exp \psi(c, e)}{\sum_{c', e'} \exp \psi(c', e')}$$
$$= \exp [\psi(c, e) - g(\psi)] \quad \text{where } g(\psi) = \log \sum_{c, e} \exp \psi(c, e)$$



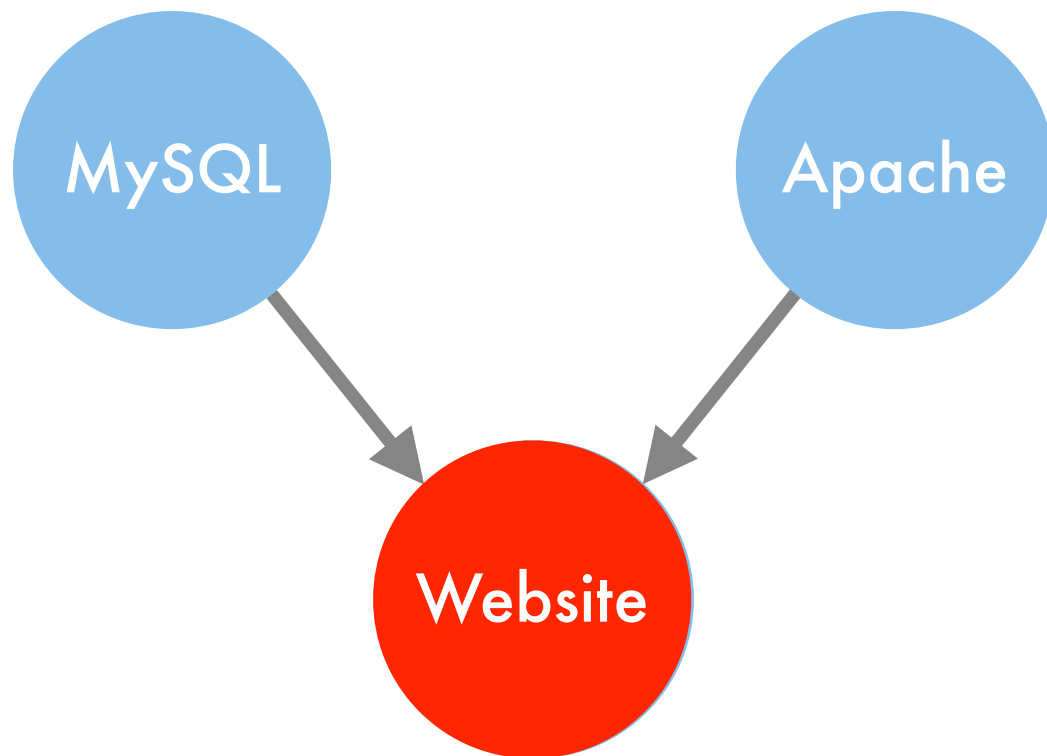
# ... some Yahoo service



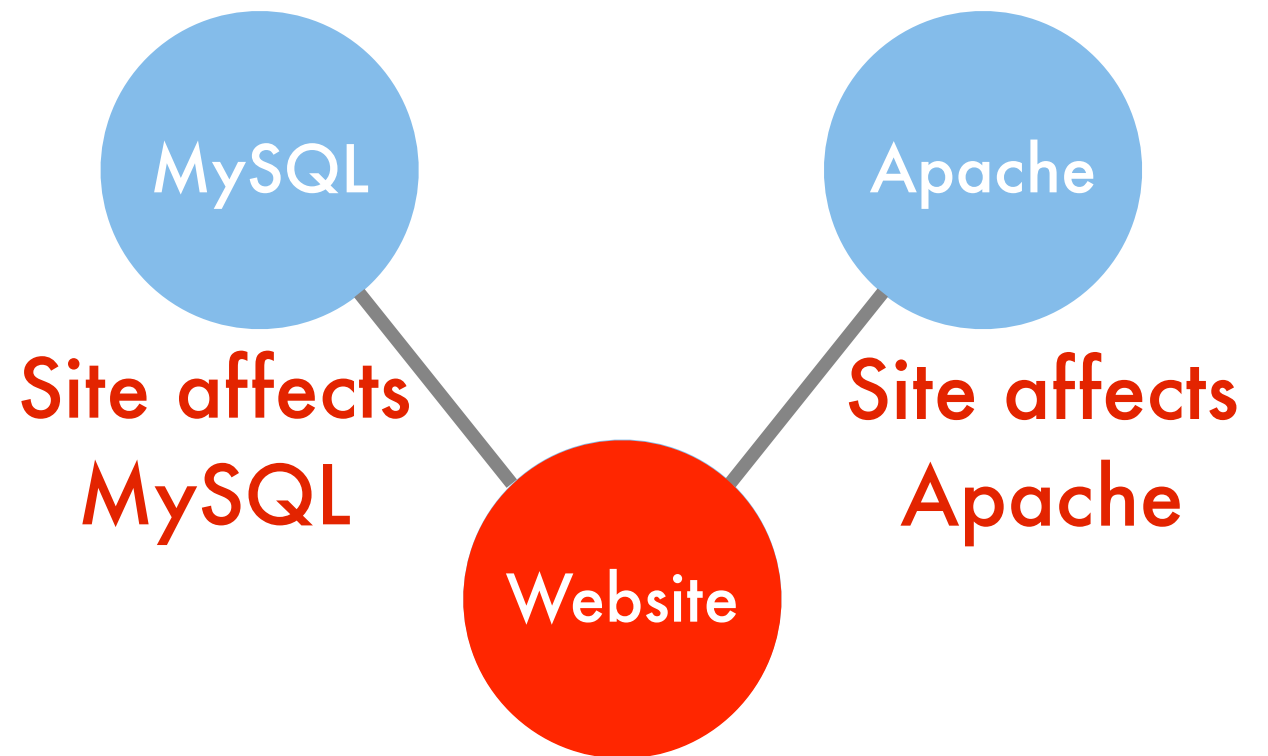
$$p(w|m, a)p(m)p(a)$$

$$m \perp\!\!\!\perp a|w$$

# ... some Yahoo service



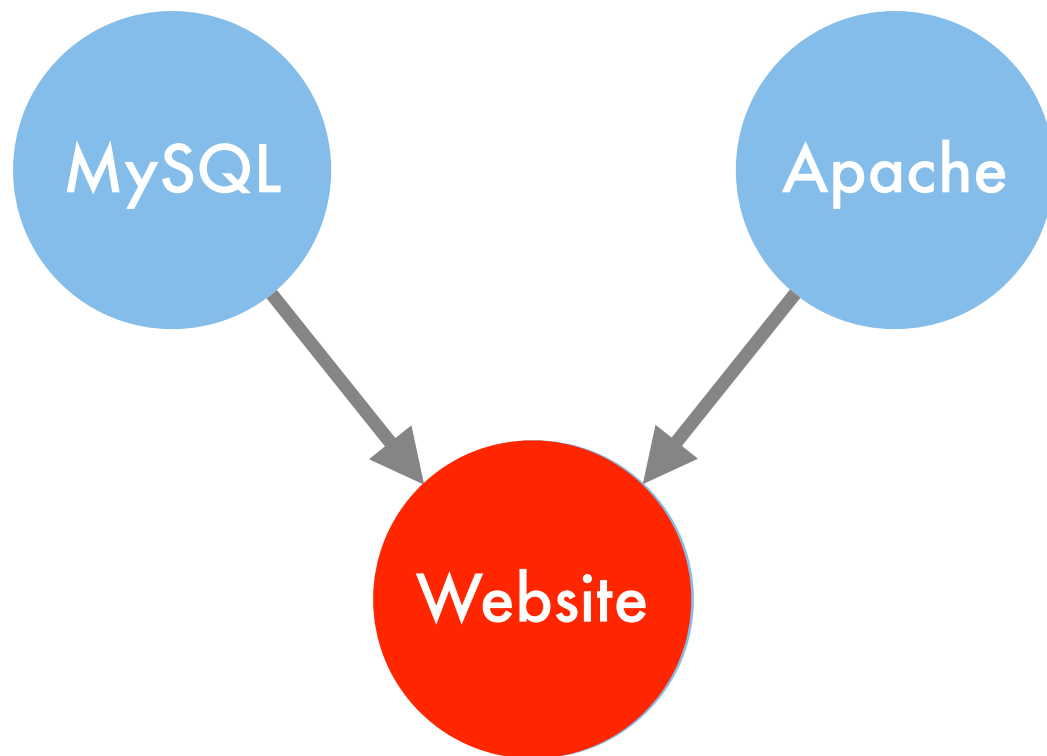
$$p(w|m, a)p(m)p(a)$$
$$m \not\perp a|w$$



$$p(m, w, a) \propto \phi(m, w)\phi(w, a)$$
$$m \perp a|w$$



# ... some Yahoo service

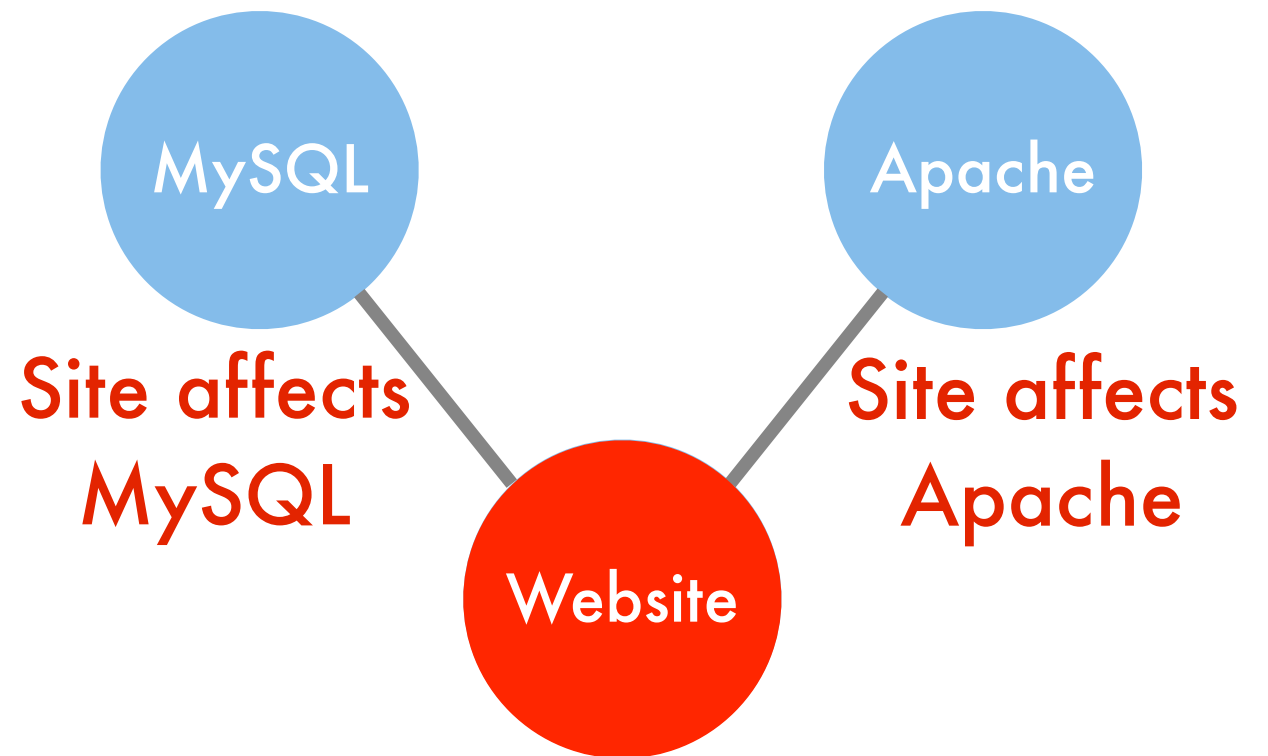


$$p(w|m, a)p(m)p(a)$$

$$m \not\perp a|w$$

easier

"debugging"



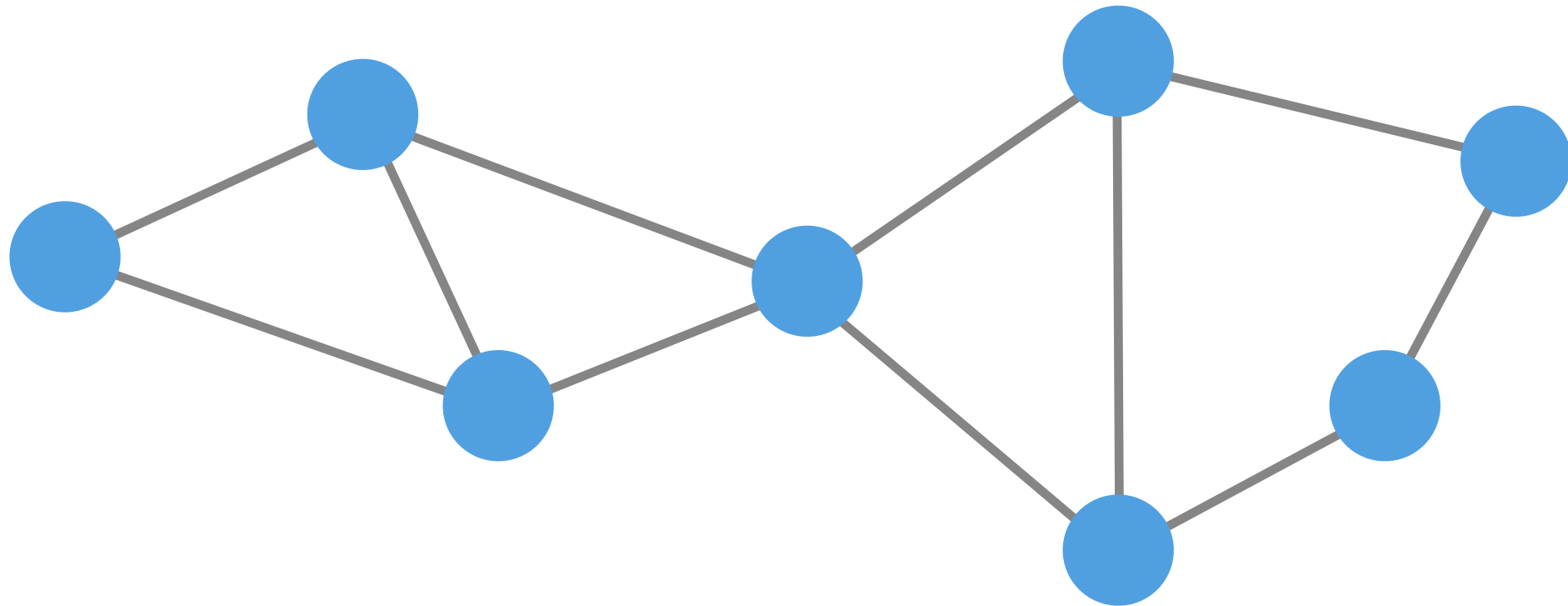
$$p(m, w, a) \propto \phi(m, w)\phi(w, a)$$

$$m \perp a|w$$

easier

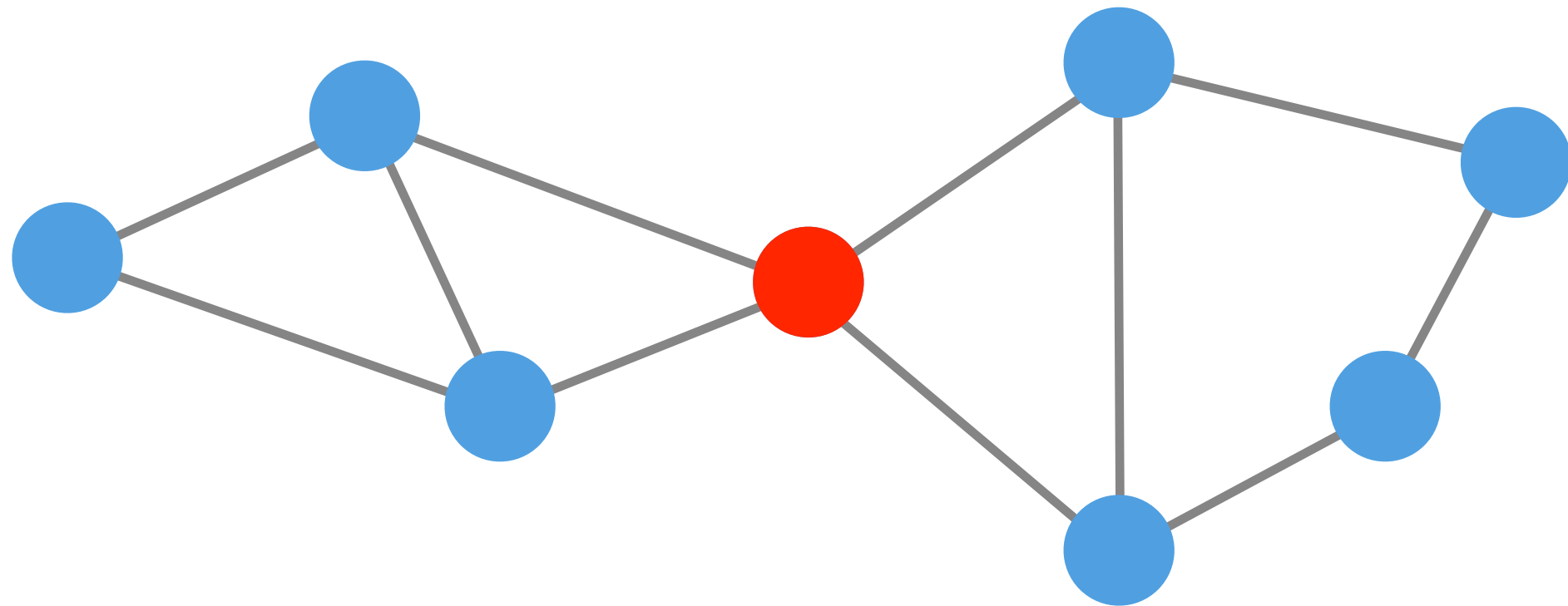
"modeling"

# Undirected Graphical Models



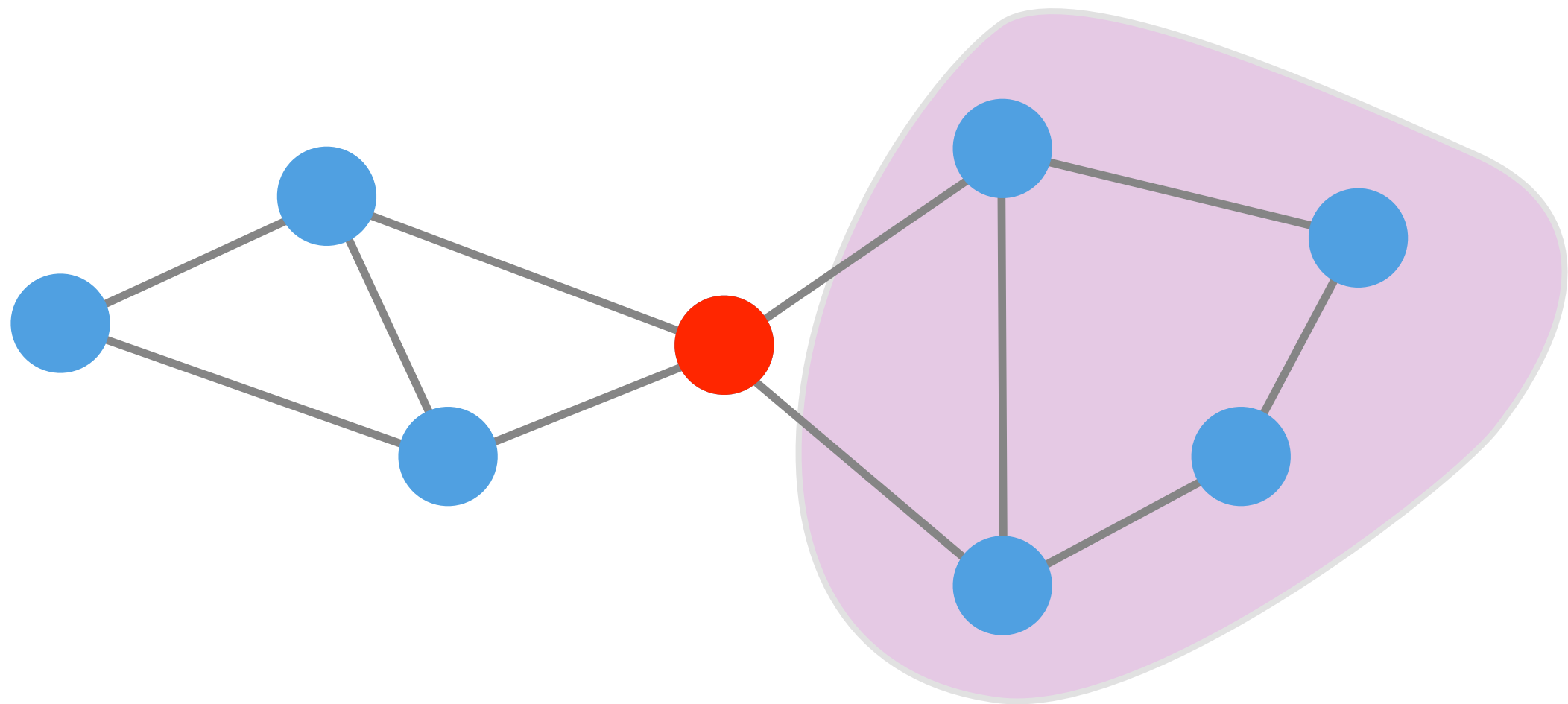
Key Concept  
Observing nodes makes remainder  
conditionally independent

# Undirected Graphical Models



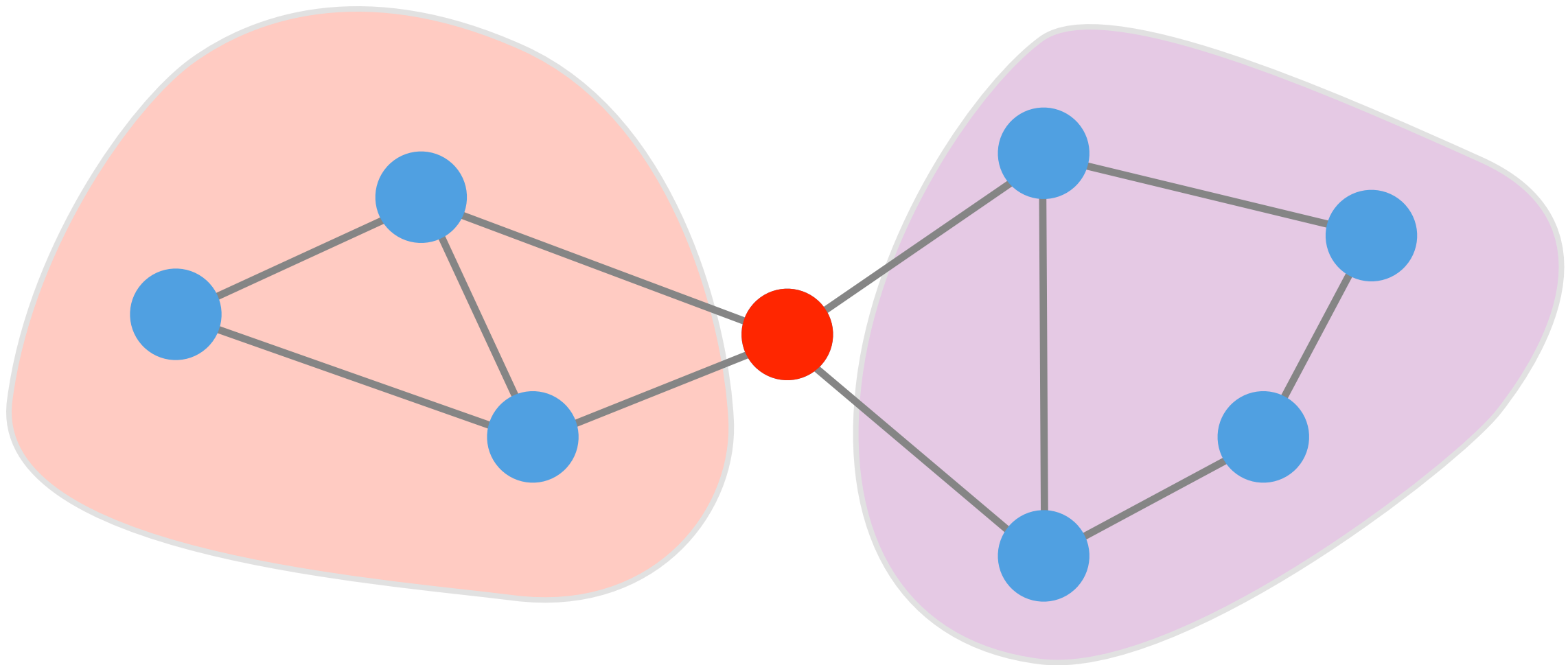
Key Concept  
Observing nodes makes remainder  
conditionally independent

# Undirected Graphical Models



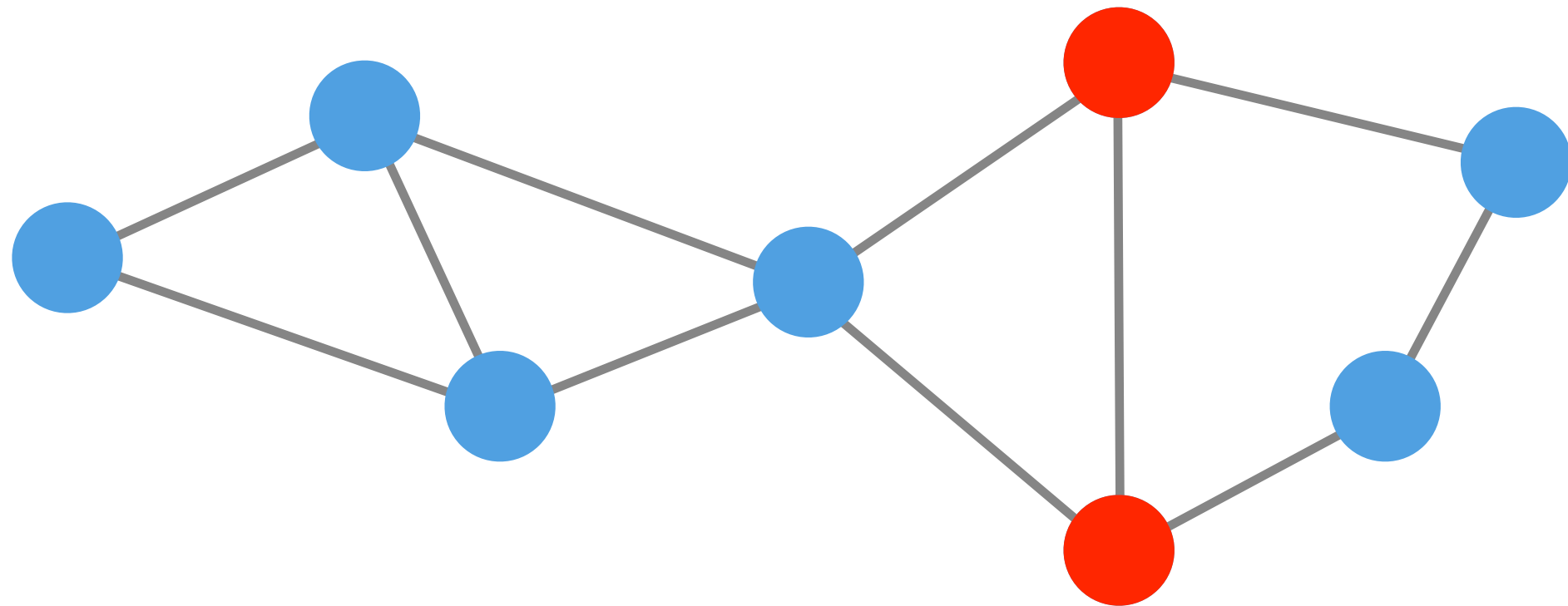
Key Concept  
Observing nodes makes remainder  
conditionally independent

# Undirected Graphical Models



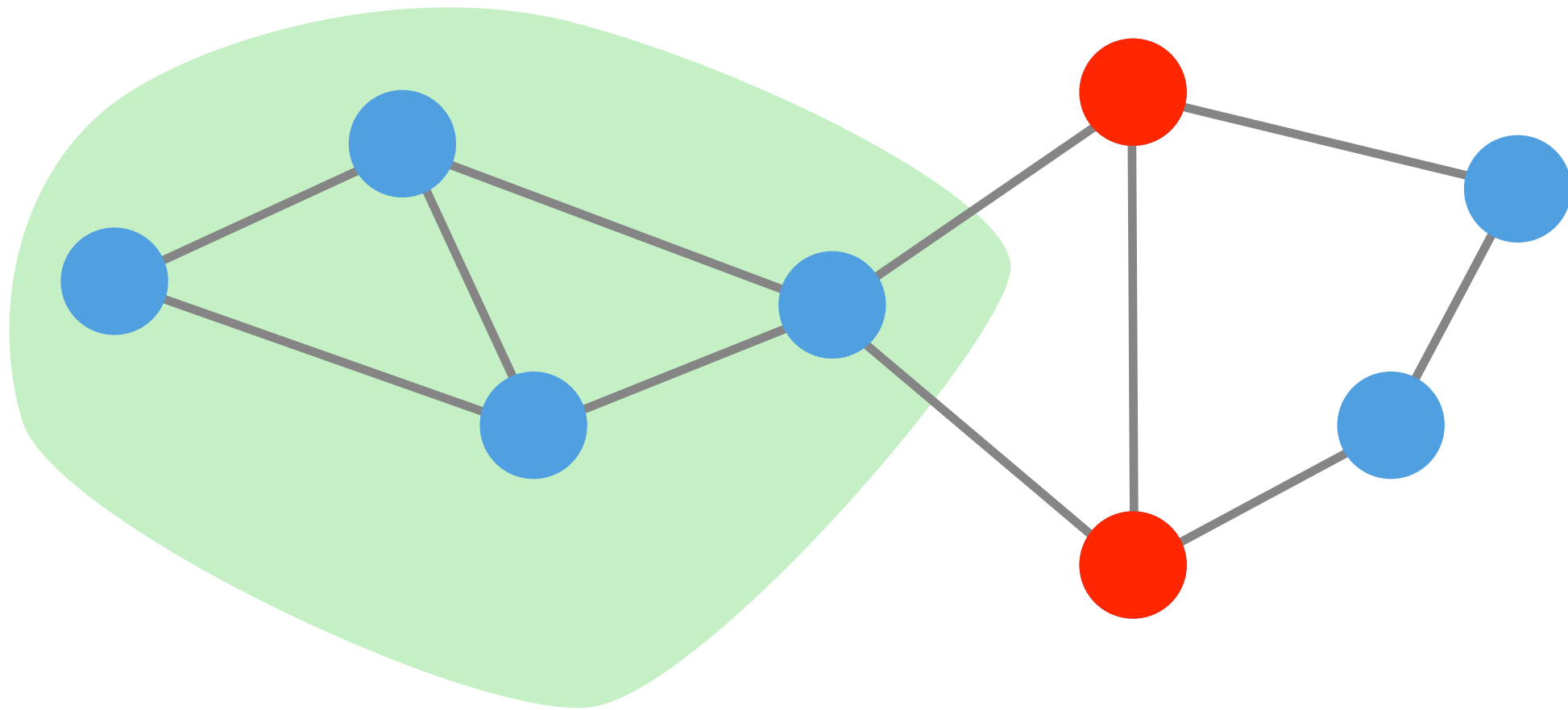
Key Concept  
Observing nodes makes remainder  
conditionally independent

# Undirected Graphical Models



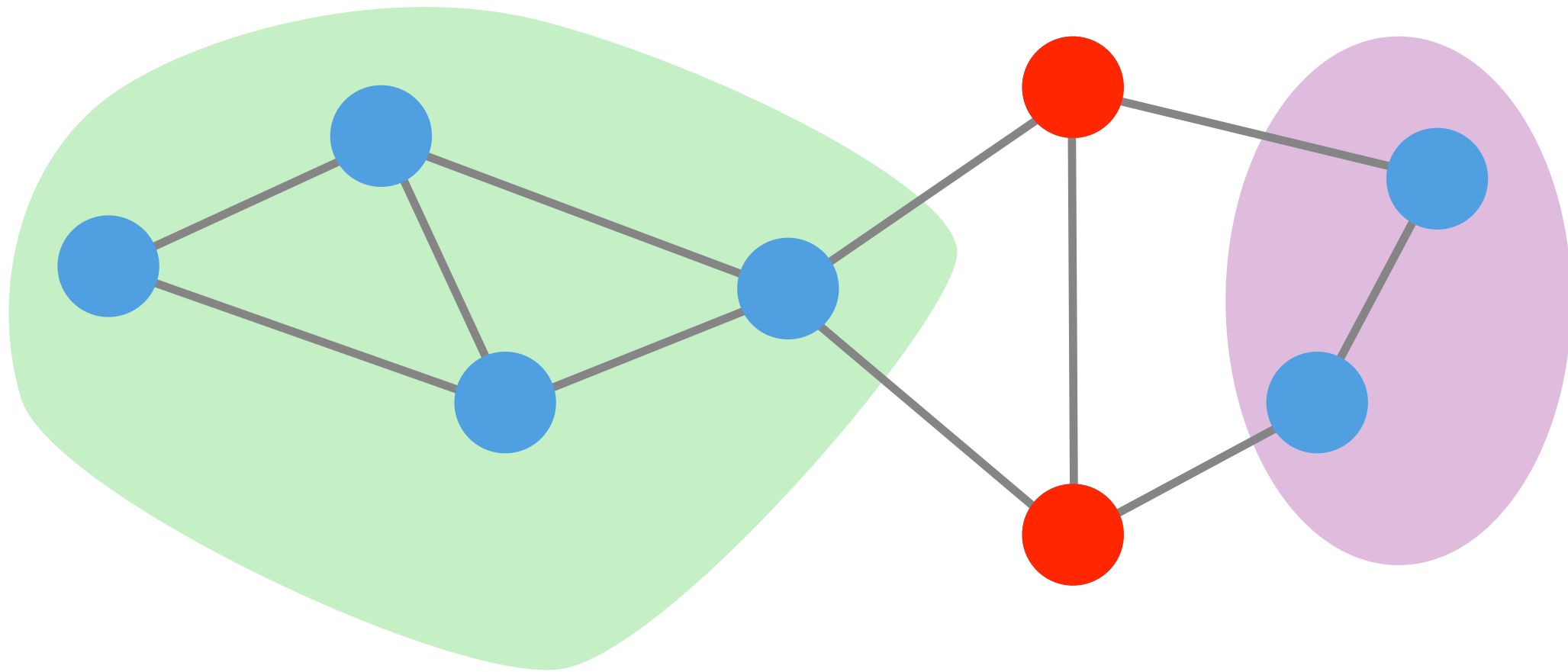
Key Concept  
Observing nodes makes remainder  
conditionally independent

# Undirected Graphical Models



Key Concept  
Observing nodes makes remainder  
conditionally independent

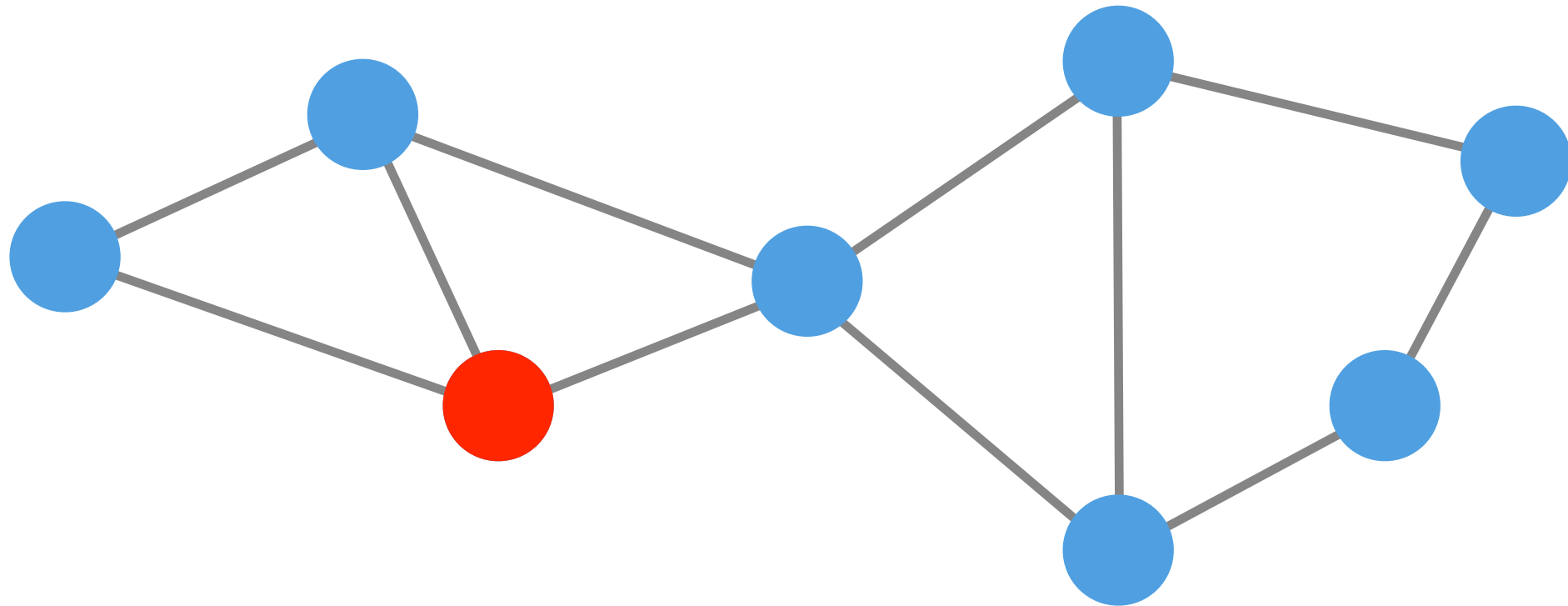
# Undirected Graphical Models



Key Concept  
Observing nodes makes remainder  
conditionally independent



# Undirected Graphical Models



Key Concept  
Observing nodes makes remainder  
conditionally independent

# Cliques



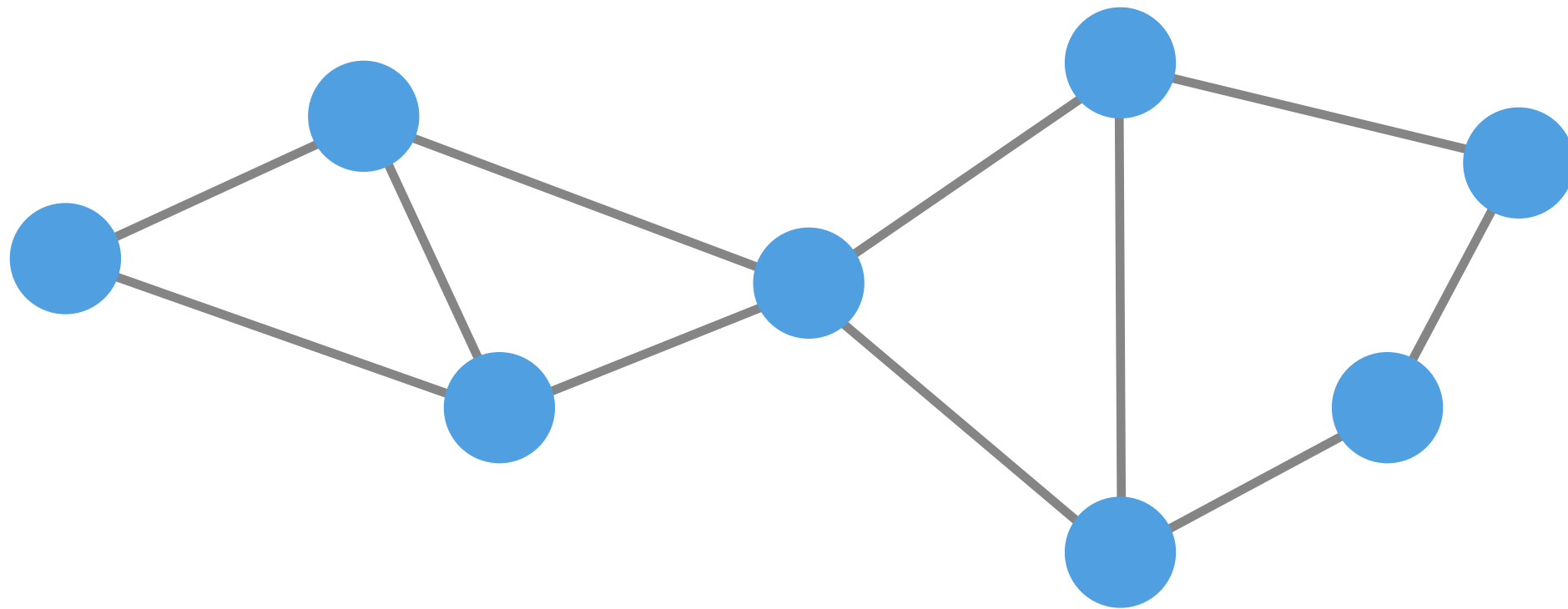


# Cliques



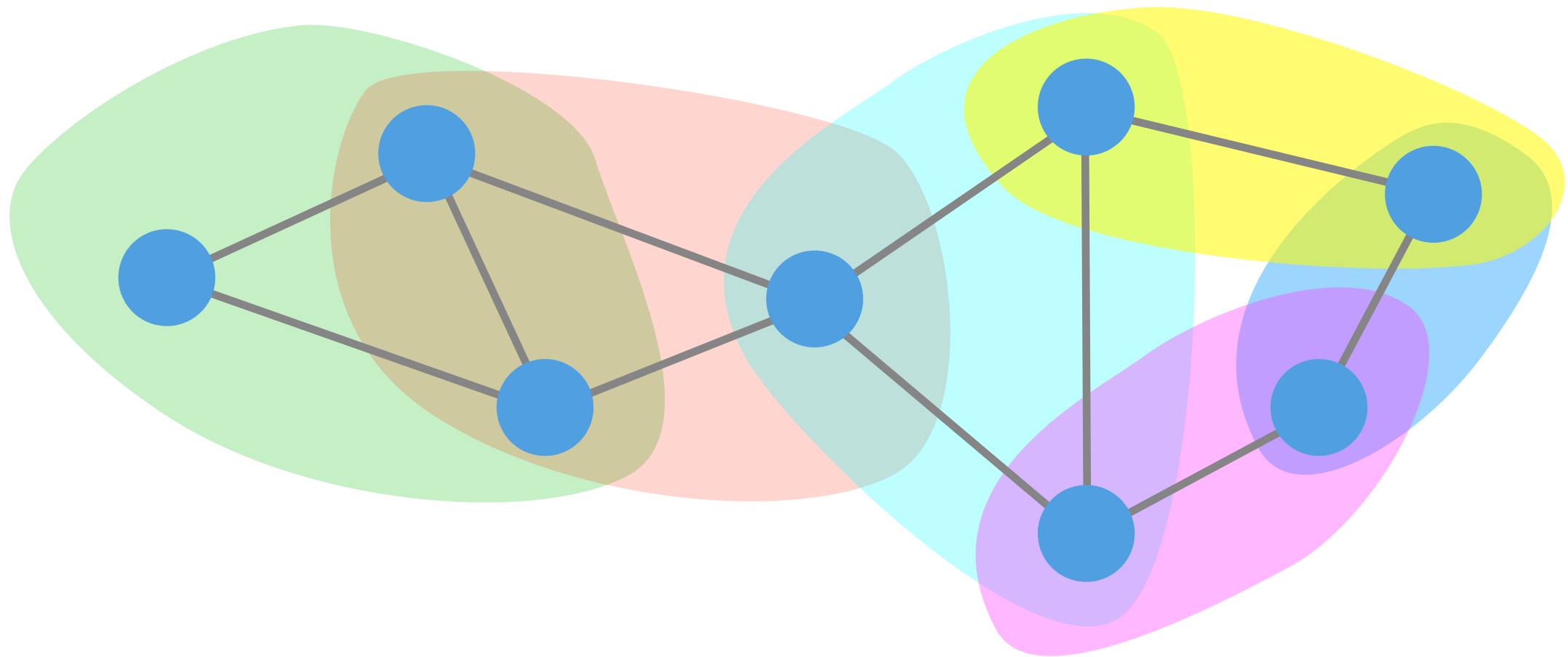
maximal fully connected subgraph

# Cliques



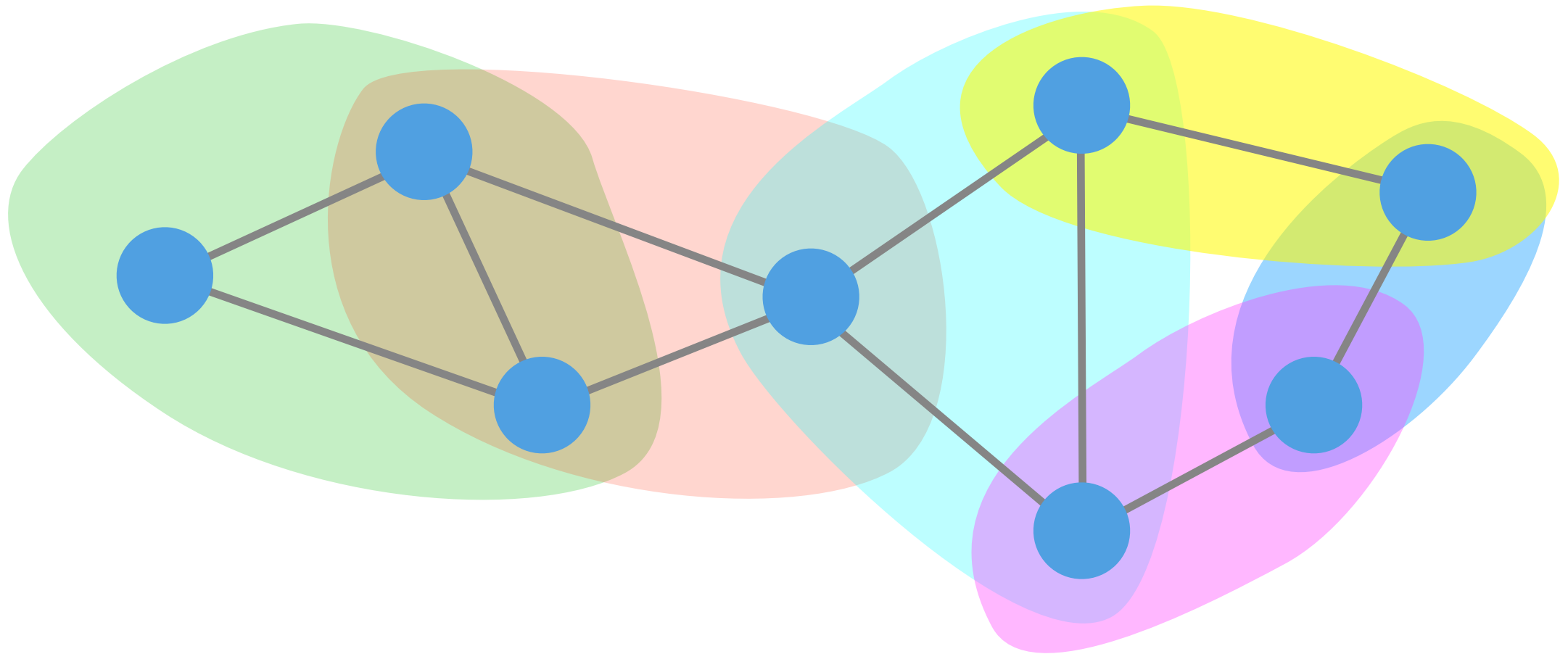
**maximal fully connected subgraph**

# Cliques



**maximal fully connected subgraph**

# Hammersley Clifford Theorem



If density has full support then it decomposes into products of clique potentials

$$p(x) = \prod_c \psi_c(x_c)$$

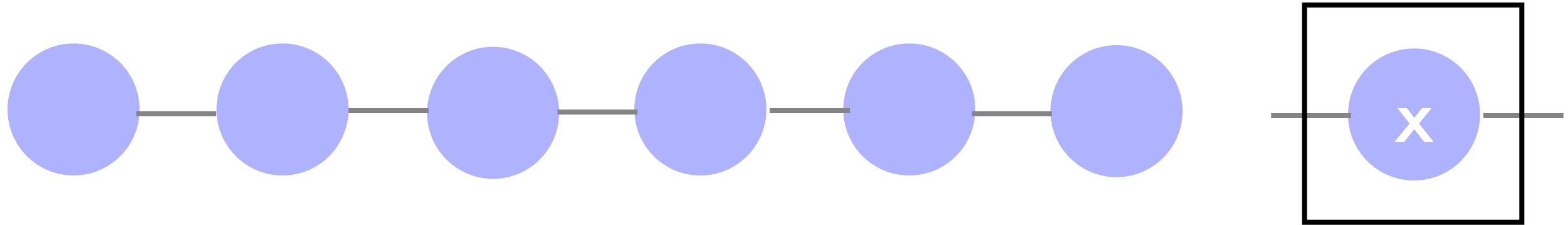
# Directed vs. Undirected

- Causal description
  - Normalization automatic
  - Intuitive
  - Requires knowledge of dependencies
  - Conditional independence tricky (Bayes Ball algorithm)
- Noncausal description (correlation only)
  - Intuitive
  - Easy modeling
  - Normalization difficult
  - Conditional independence easy to read off (graph connectivity)

Examples

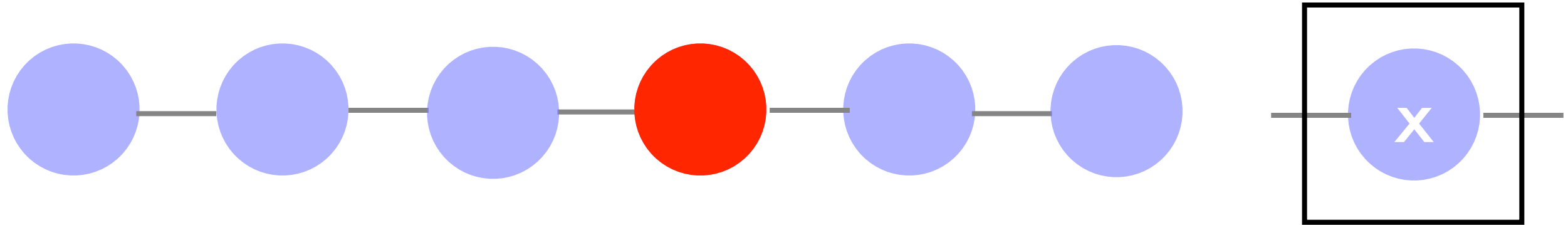


# Chains



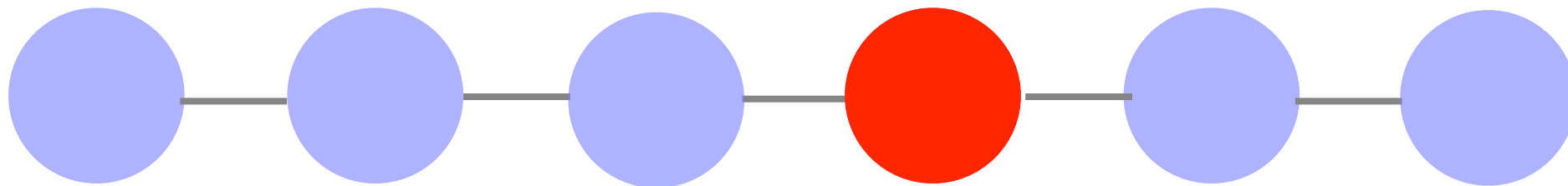
$$p(x) = \prod_i \psi_i(x_i, x_{i+1})$$

# Chains

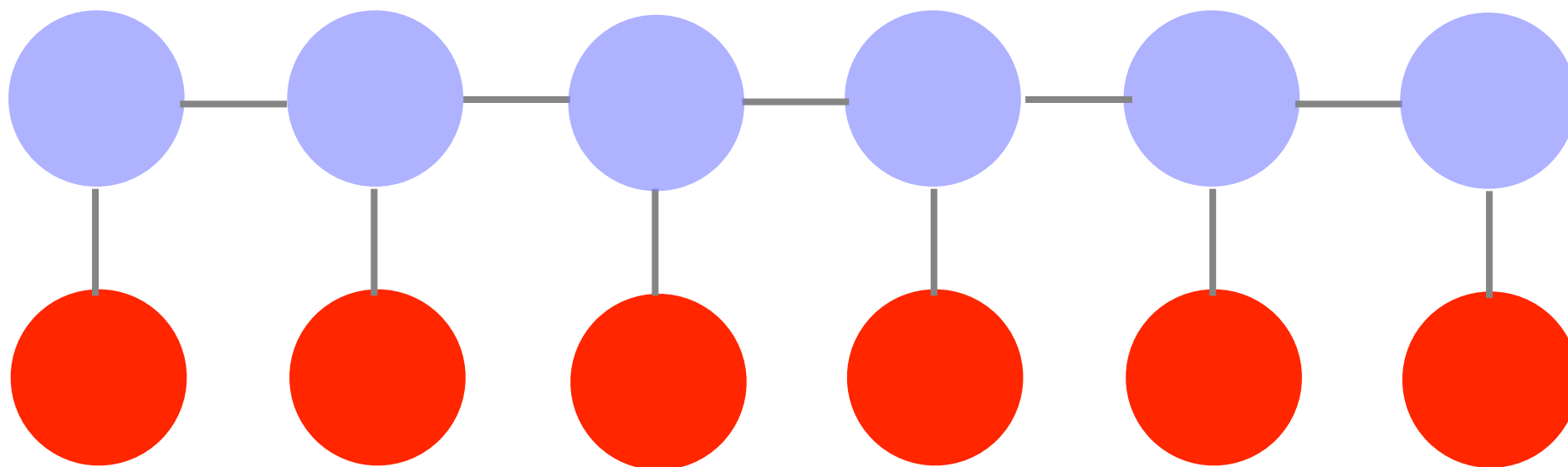
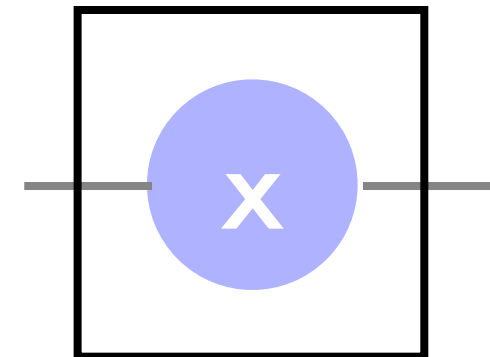


$$p(x) = \prod_i \psi_i(x_i, x_{i+1})$$

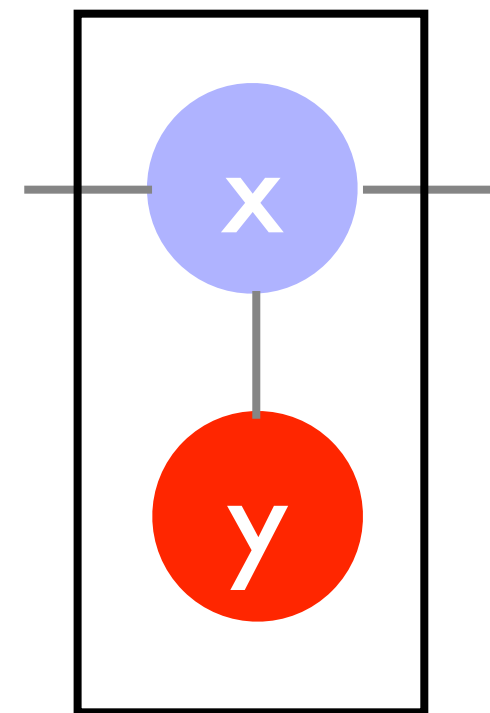
# Chains



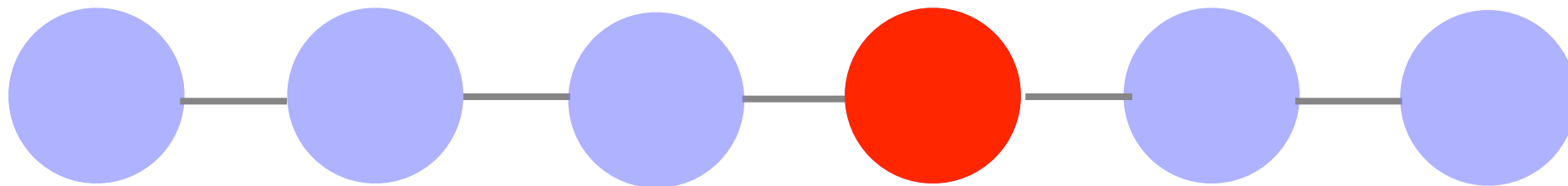
$$p(x) = \prod_i \psi_i(x_i, x_{i+1})$$



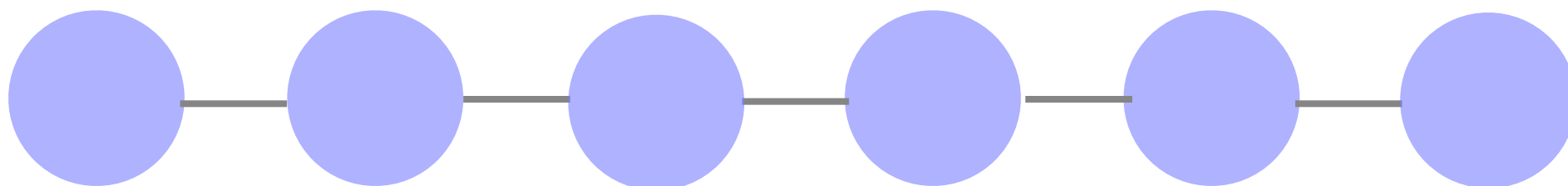
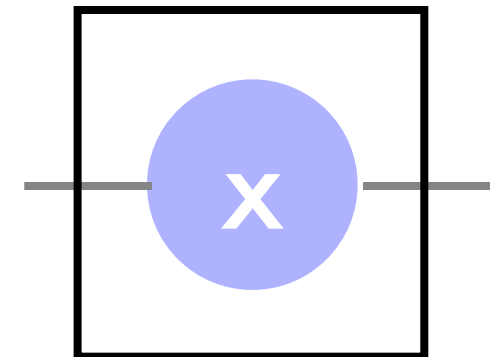
$$p(x, y) = \prod_i \psi_i^x(x_i, x_{i+1}) \psi_i^{xy}(x_i, y_i)$$



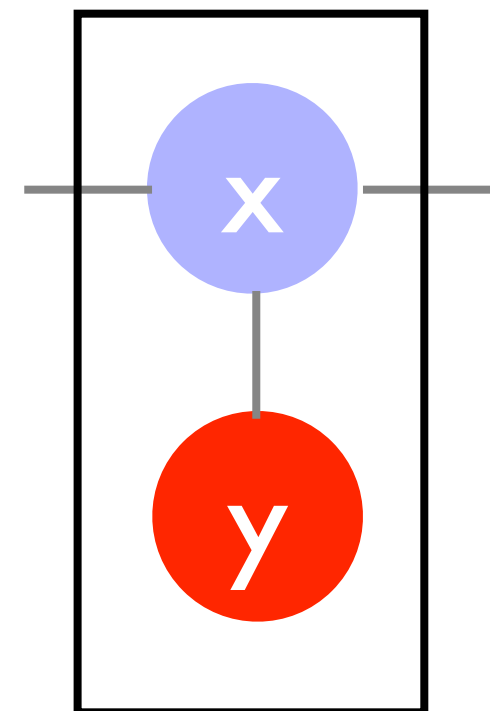
# Chains



$$p(x) = \prod_i \psi_i(x_i, x_{i+1})$$

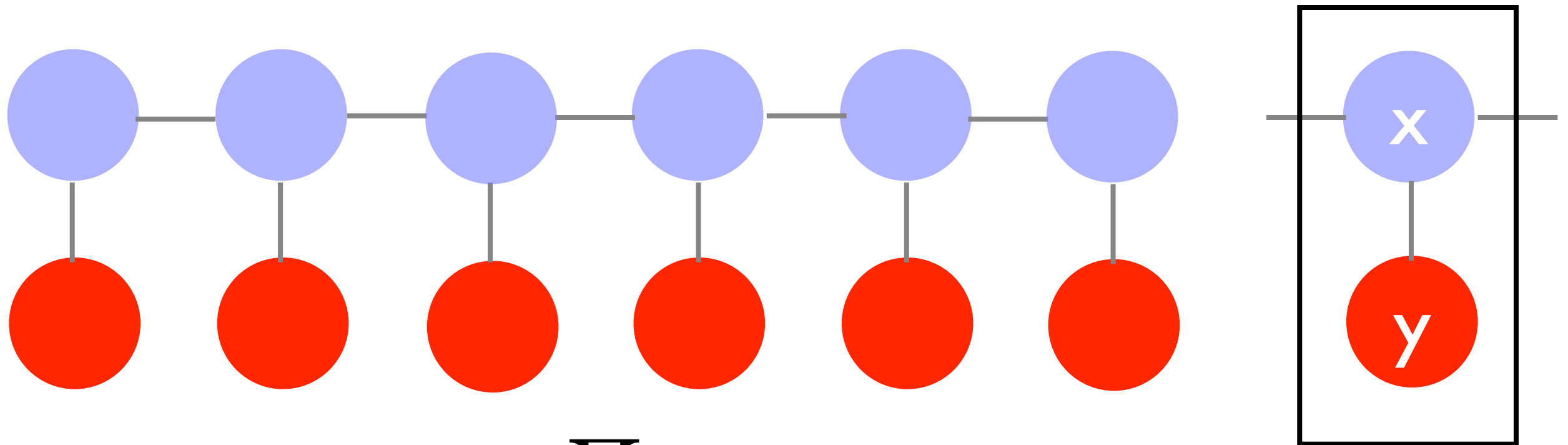


$$p(x|y) \propto \prod_i \underbrace{\psi_i^x(x_i, x_{i+1}) \psi_i^{xy}(x_i, y_i)}_{=: f_i(x_i, x_{i+1})}$$



$$p(x, y) = \prod_i \psi_i^x(x_i, x_{i+1}) \psi_i^{xy}(x_i, y_i)$$

# Chains



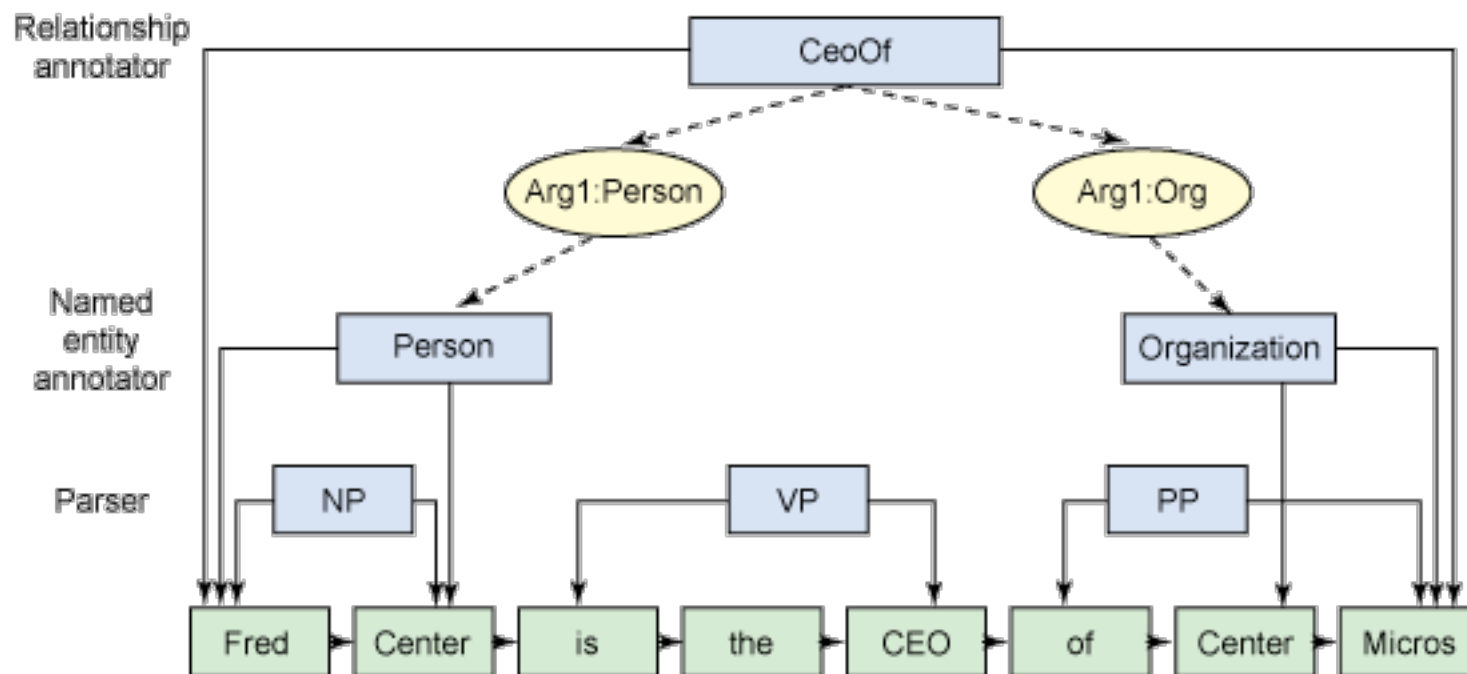
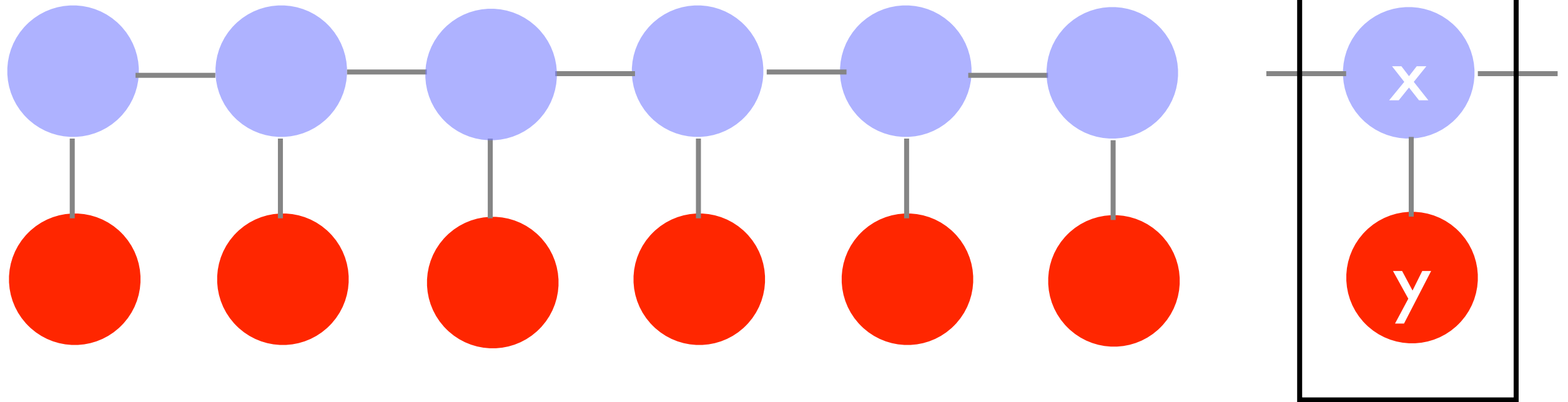
$$p(x|y) \propto \prod_i \underbrace{\psi_i^x(x_i, x_{i+1}) \psi_i^{xy}(x_i, y_i)}_{=: f_i(x_i, x_{i+1})}$$

Dynamic Programming

$$l_1(x_1) = 1 \text{ and } l_{i+1}(x_{i+1}) = \sum_{x_i} l_i(x_i) f_i(x_i, x_{i+1})$$

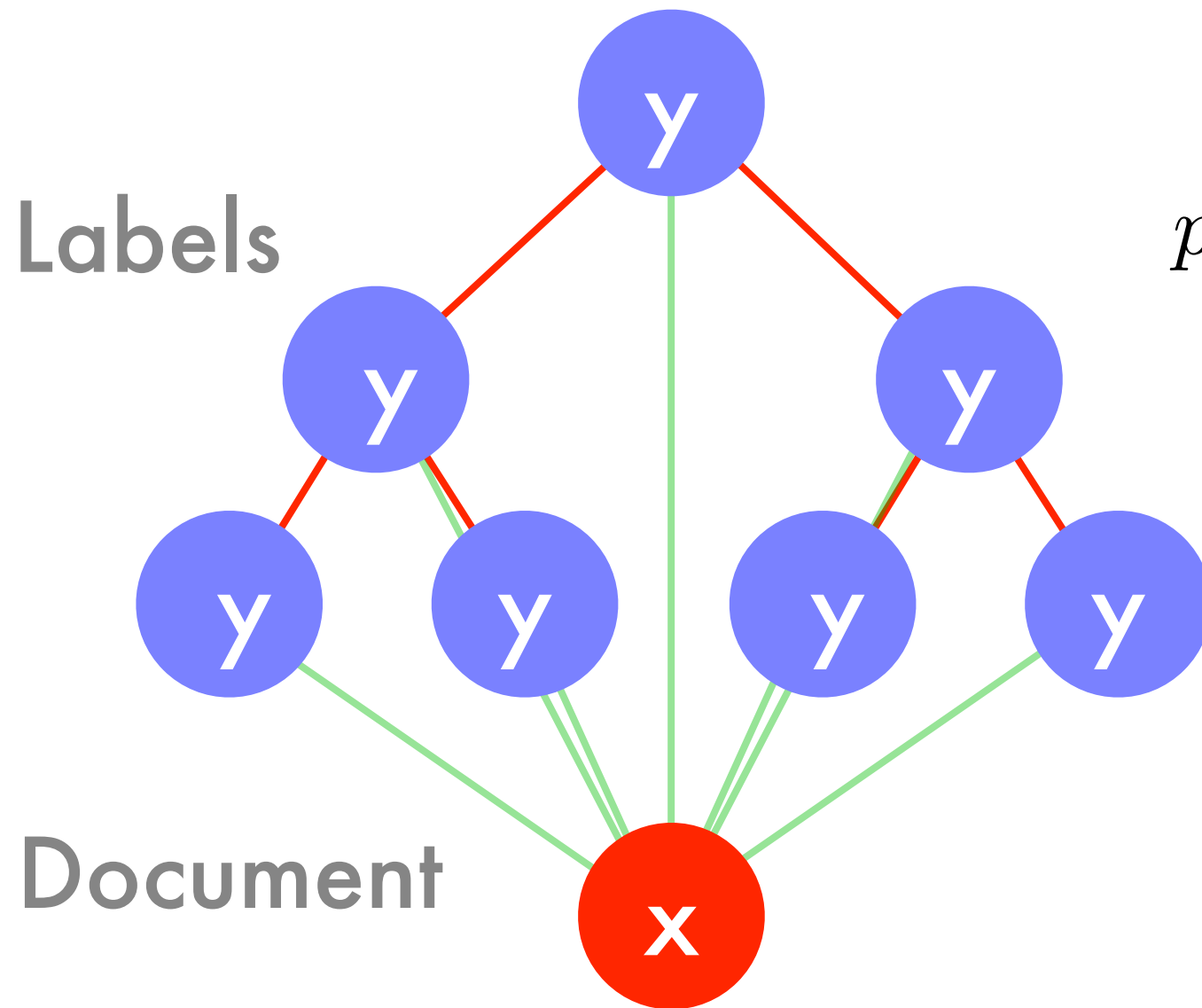
$$r_n(x_n) = 1 \text{ and } r_i(x_i) = \sum_{x_{i+1}} r_{i+1}(x_{i+1}) f_i(x_i, x_{i+1})$$

# Named Entity Tagging



$$p(x|y) \propto \prod_i \underbrace{\psi_i^x(x_i, x_{i+1}) \psi_i^{xy}(x_i, y_i)}_{=: f_i(x_i, x_{i+1})}$$

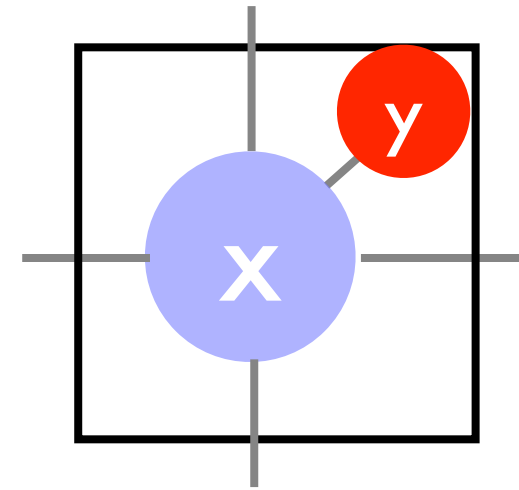
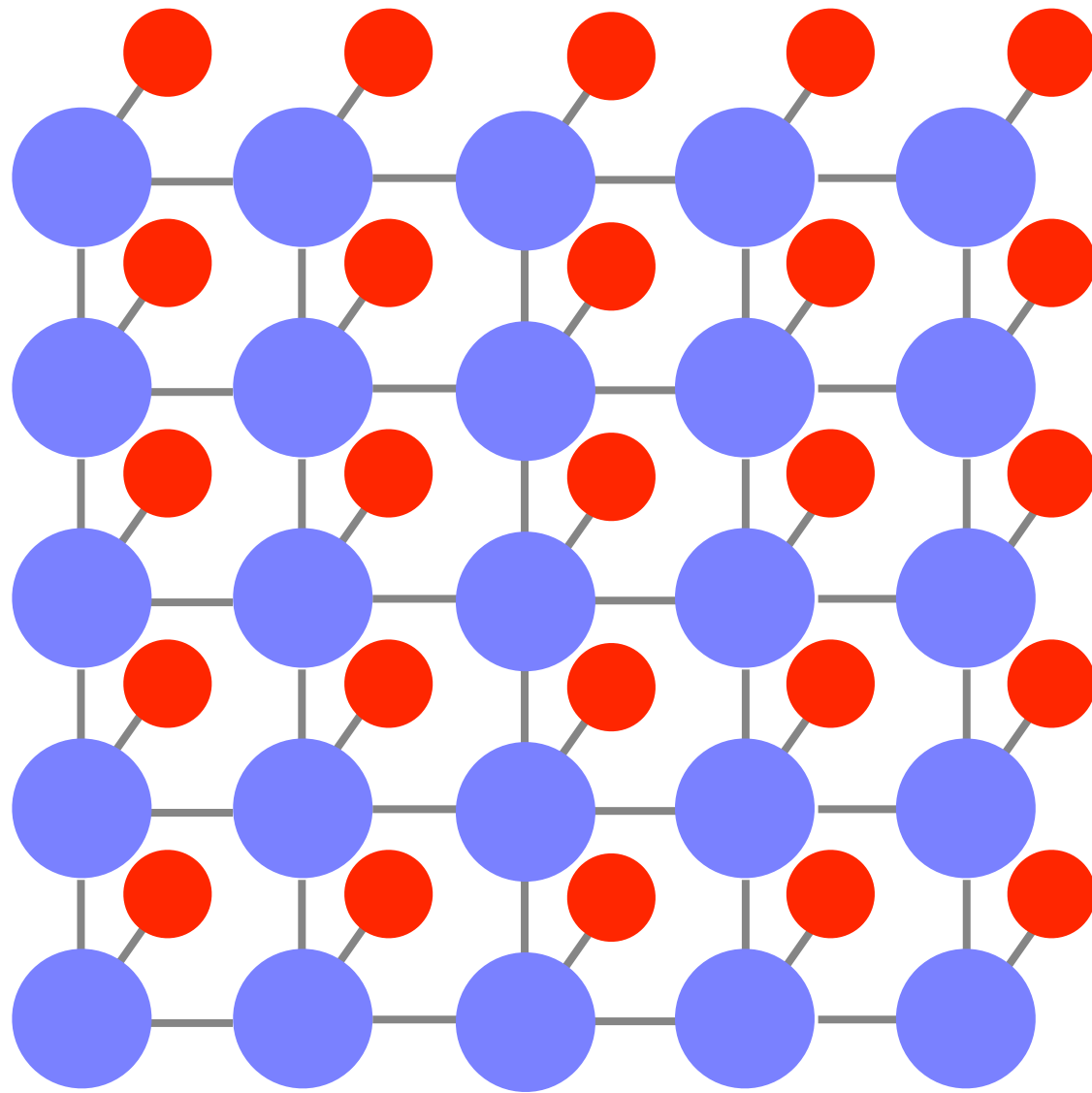
# Trees + Ontologies



$$p(y|x) = \prod_i \psi(y_i, y_{\text{parent}(i)}, x)$$

- **Ontology classification (e.g. YDir, DMOZ)**

# Spin Glasses + Images

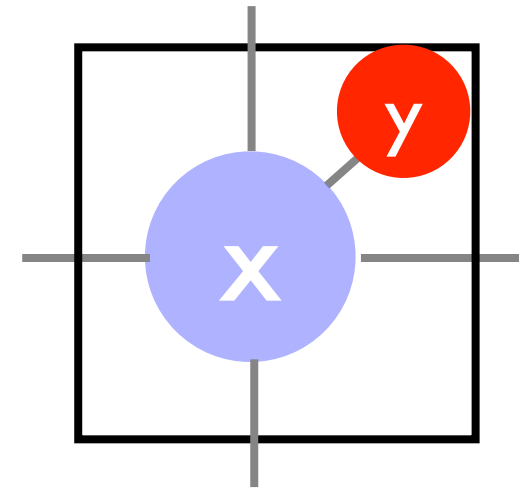
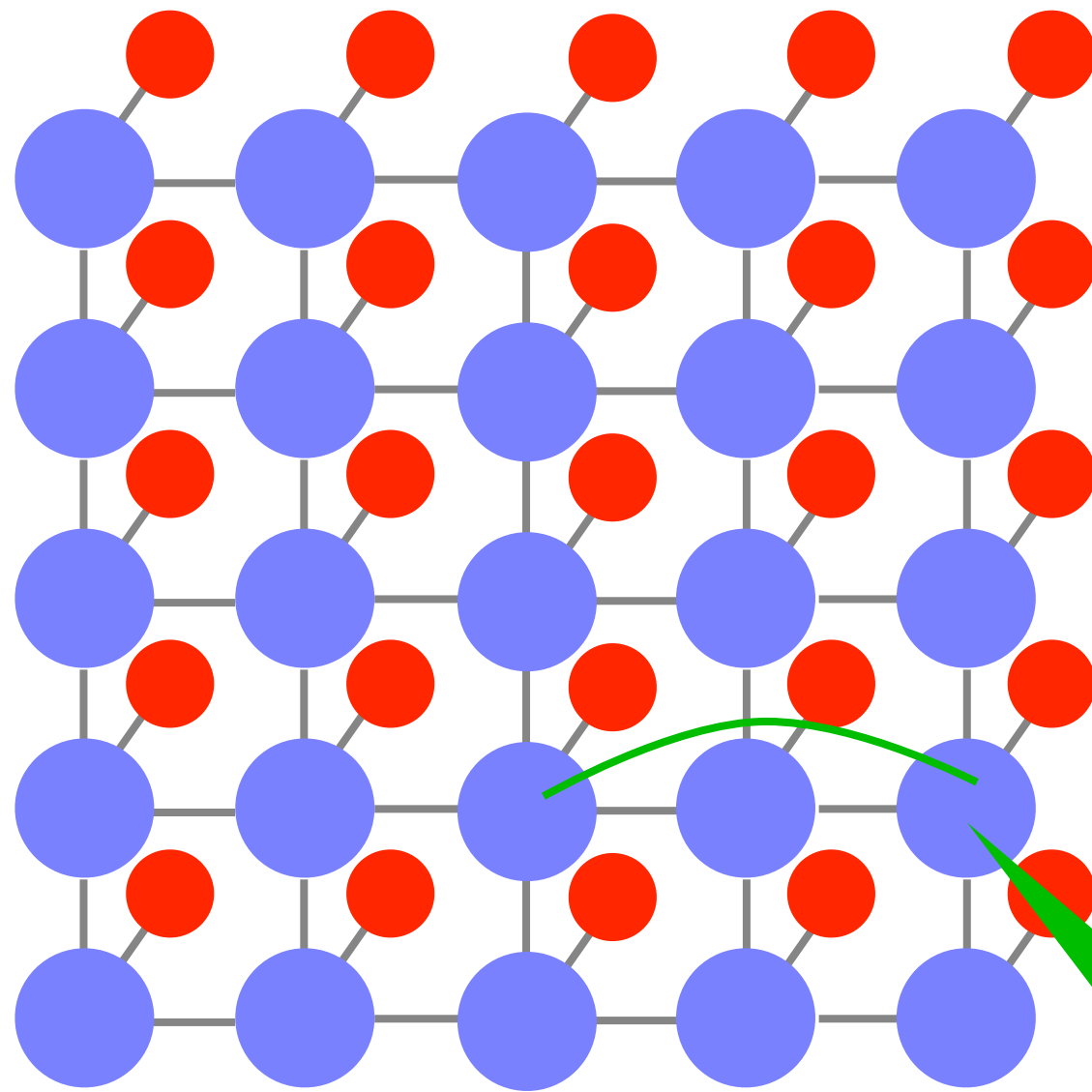


observed pixels  
real image

$$p(x|y) = \prod_{ij} \psi^{\text{right}}(x_{ij}, x_{i+1,j}) \psi^{\text{up}}(x_{ij}, x_{i,j+1}) \psi^{xy}(x_{ij}, y_{ij})$$



# Spin Glasses + Images

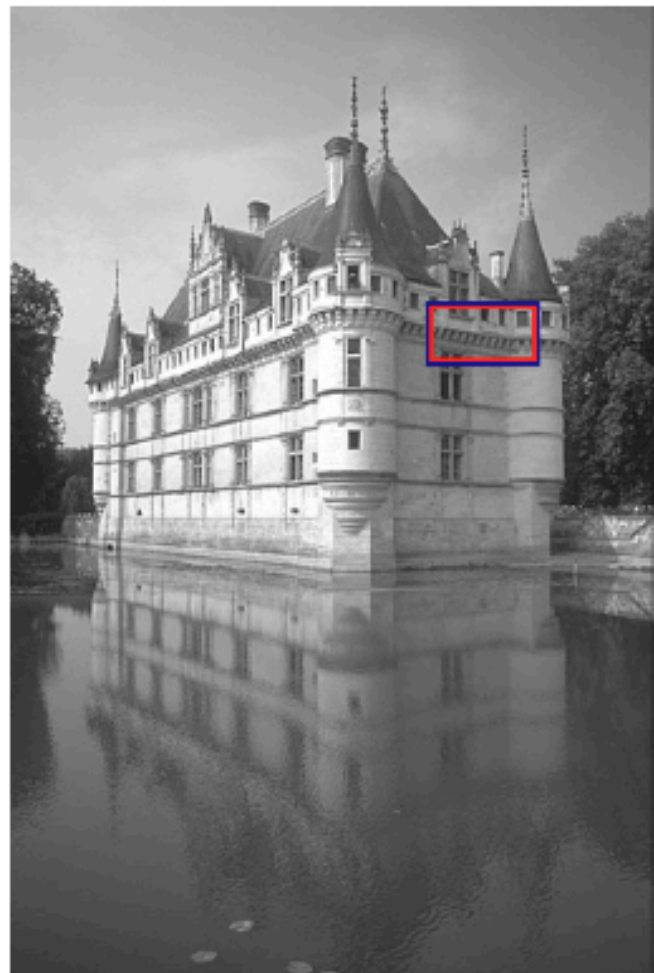


observed pixels  
real image

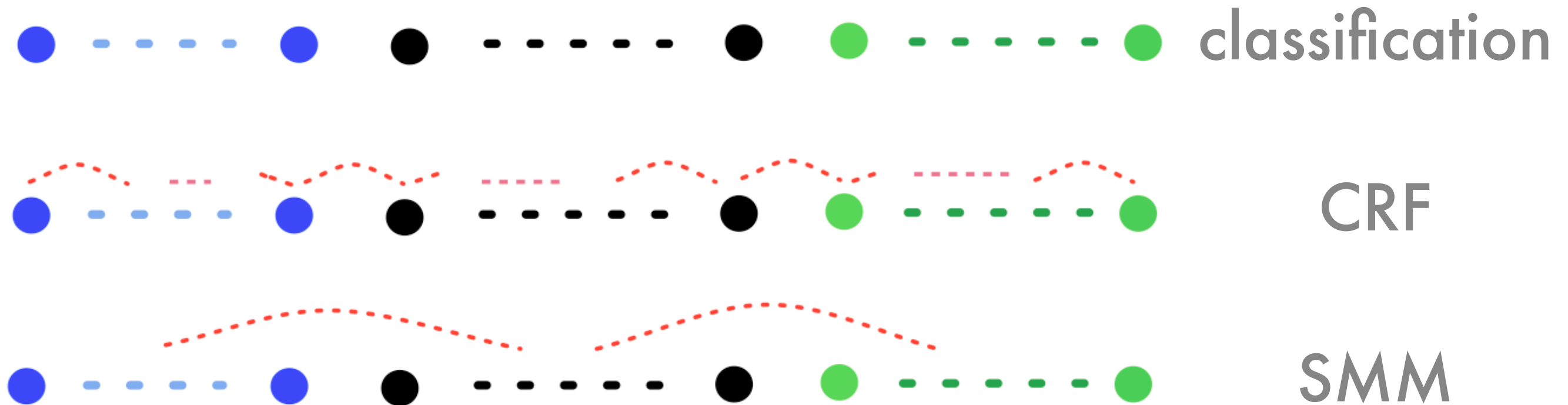
long range interactions

$$p(x|y) = \prod_{ij} \psi^{\text{right}}(x_{ij}, x_{i+1,j}) \psi^{\text{up}}(x_{ij}, x_{i,j+1}) \psi^{xy}(x_{ij}, y_{ij})$$

# Image Denoising



# Semi-Markov Models

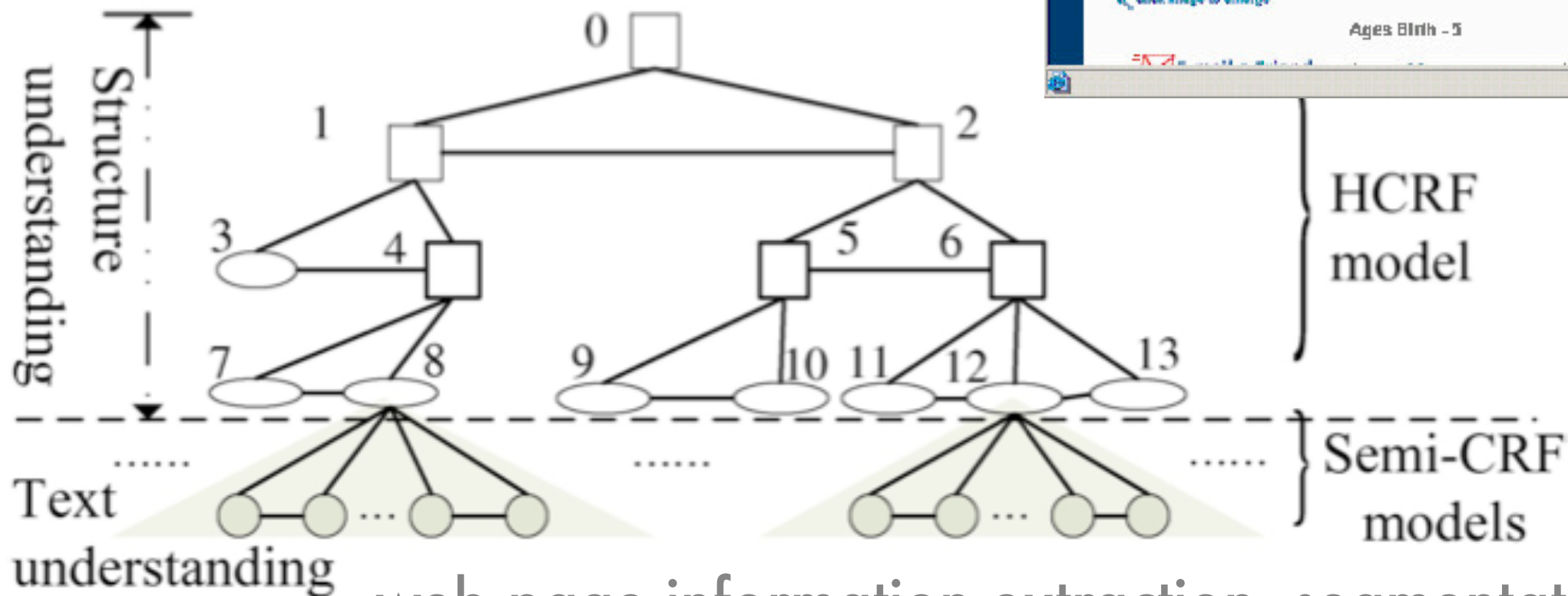
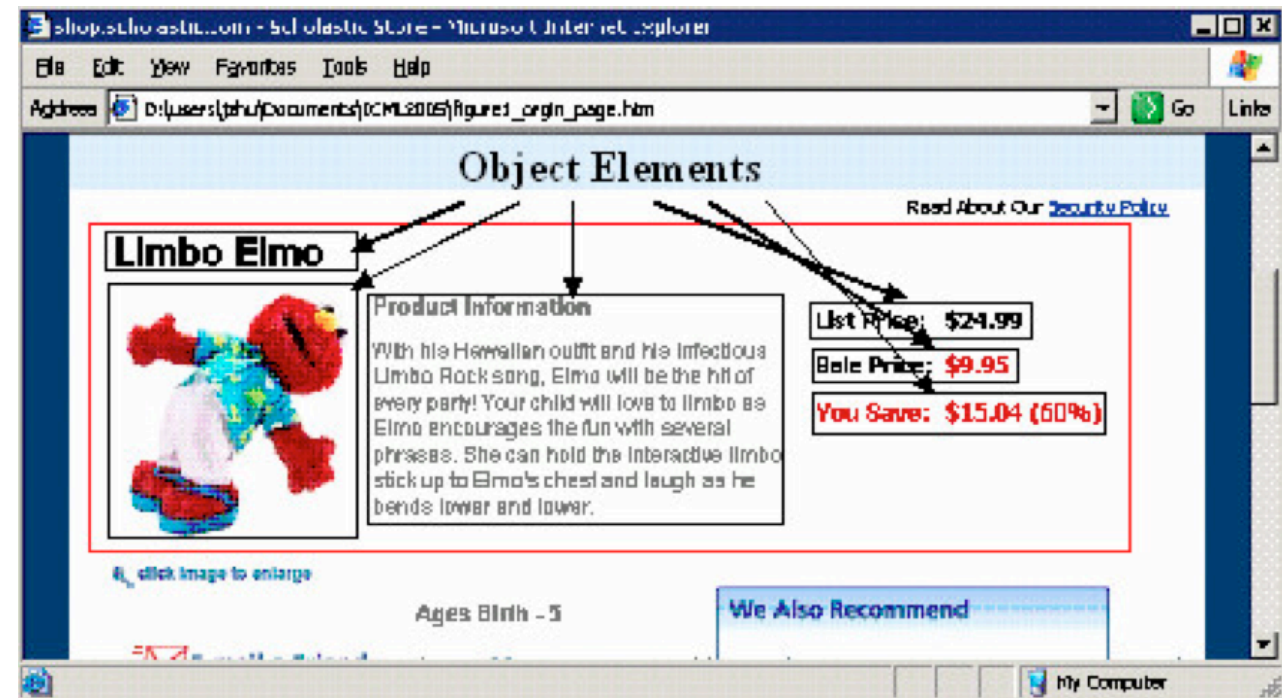


- Flexible length of an episode
- Segmentation between episodes

phrase segmentation, activity recognition, motion data analysis

Shi, Smola, Altun, Vishwanathan, Li, 2007-2009

# 2D CRF for Webpages



web page information extraction, segmentation, annotation

Bo, Zhu, Nie, Wen, Hon, 2005-2007

# Exponential Families and Graphical Models

# Exponential Family Reunion

- **Density function**

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

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

- **Log partition function generates cumulants**

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

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

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



# Log Partition Function

$$p(x|\theta) = e^{\langle \phi(x), \theta \rangle - g(\theta)}$$

Unconditional model

$$g(\theta) = \log \sum_x e^{\langle \phi(x), \theta \rangle}$$

$$\partial_\theta g(\theta) = \frac{\sum_x \phi(x) e^{\langle \phi(x), \theta \rangle}}{\sum_x e^{\langle \phi(x), \theta \rangle}} = \sum_x \phi(x) e^{\langle \phi(x), \theta \rangle - g(\theta)}$$

$$p(y|\theta, x) = e^{\langle \phi(x, y), \theta \rangle - g(\theta|x)}$$

Conditional model

$$g(\theta|x) = \log \sum_y e^{\langle \phi(x, y), \theta \rangle}$$

$$\partial_\theta g(\theta|x) = \frac{\sum_y \phi(x, y) e^{\langle \phi(x, y), \theta \rangle}}{\sum_y e^{\langle \phi(x, y), \theta \rangle}} = \sum_y \phi(x, y) e^{\langle \phi(x, y), \theta \rangle - g(\theta|x)}$$

# Estimation

- **Conditional log-likelihood**

$$\log p(y|x; \theta) = \langle \phi(x, y), \theta \rangle - g(\theta|x)$$

- **Log-posterior (Gaussian Prior)**

$$\log p(\theta|X, Y) = \sum_i \log(y_i|x_i; \theta) + \log p(\theta) + \text{const.}$$

$$= \left\langle \sum_i \phi(x_i, y_i), \theta \right\rangle - \sum_i g(\theta|x_i) - \frac{1}{2\sigma^2} \|\theta\|^2 + \text{const.}$$

- **First order optimality conditions**

expensive

maxent  
model

$$\sum_i \phi(x_i, y_i) = \sum_i \mathbf{E}_{y|x_i} [\phi(x_i, y)] + \frac{1}{\sigma^2} \theta$$

prior



# Logistic Regression

- **Label space**

$$\phi(x, y) = y\phi(x) \text{ where } y \in \{\pm 1\}$$

- **Log-partition function**

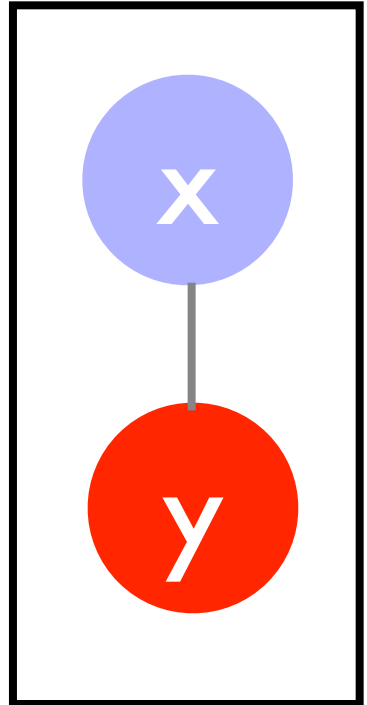
$$g(\theta|x) = \log \left[ e^{1 \cdot \langle \phi(x), \theta \rangle} + e^{-1 \cdot \langle \phi(x), \theta \rangle} \right] = \log 2 \cosh \langle \phi(x), \theta \rangle$$

- **Convex minimization problem**

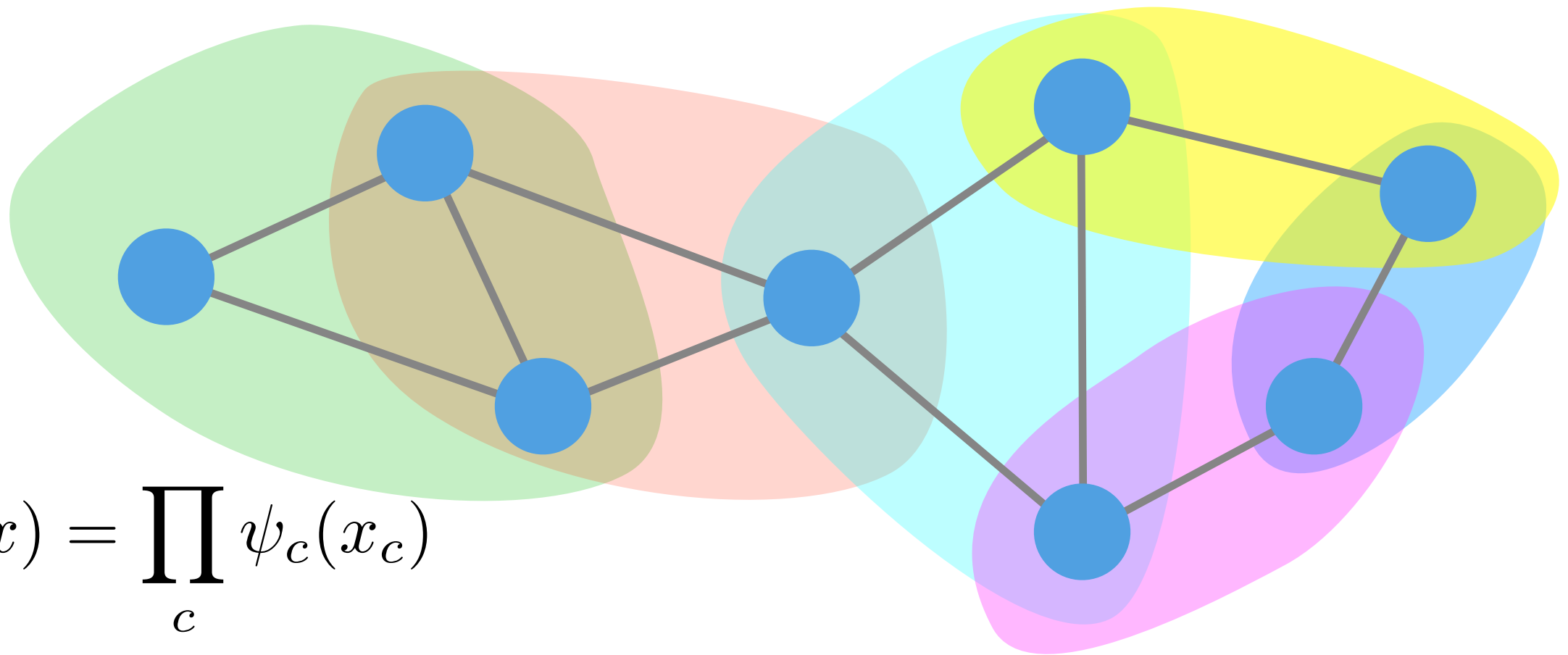
$$\underset{\theta}{\text{minimize}} \frac{1}{2\sigma^2} \|\theta\|^2 + \sum_i \log 2 \cosh \langle \phi(x_i), \theta \rangle - y_i \langle \phi(x_i), \theta \rangle$$

- **Prediction**

$$p(y|x, \theta) = \frac{e^{y\langle \phi(x), \theta \rangle}}{e^{\langle \phi(x), \theta \rangle} + e^{-\langle \phi(x), \theta \rangle}} = \frac{1}{1 + e^{-2y\langle \phi(x), \theta \rangle}}$$



# Exponential Clique Decomposition



$$p(x) = \prod_c \psi_c(x_c)$$

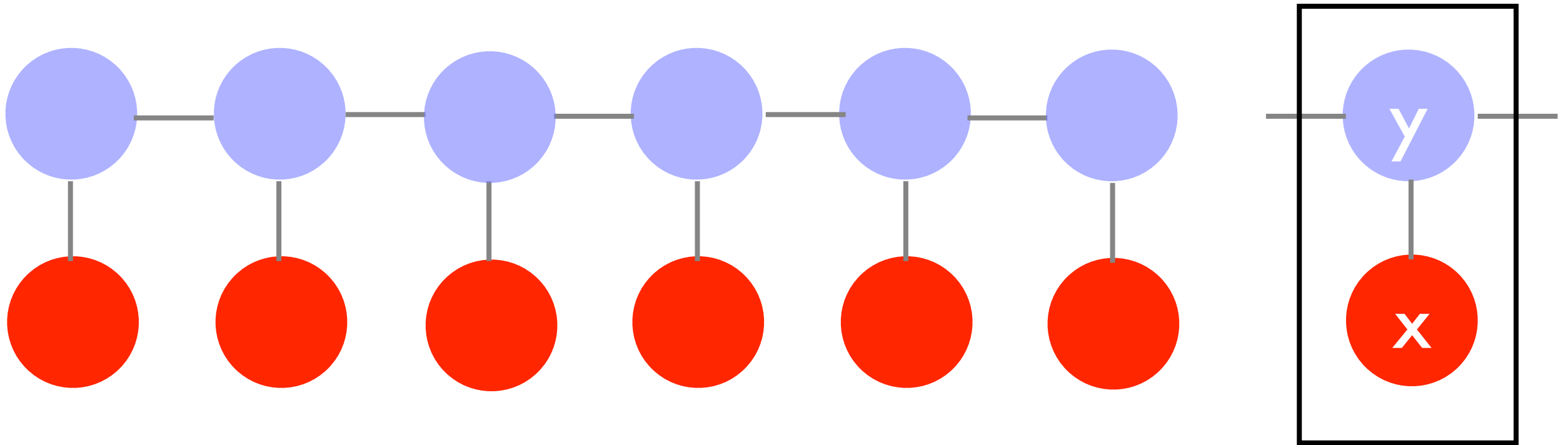
Theorem: Clique decomposition holds in sufficient statistics

$$\phi(x) = (\dots, \phi_c(x_c), \dots) \text{ and } \langle \phi(x), \theta \rangle = \sum_c \langle \phi_c(x_c), \theta_c \rangle$$

Corollary: we only need expectations on cliques

$$\mathbf{E}_x[\phi(x)] = (\dots, \mathbf{E}_{x_c}[\phi_c(x_c)], \dots)$$

# Conditional Random Fields



$$\phi(x) = (y_1 \phi_x(x_1), \dots, y_n \phi_x(x_n), \phi_y(y_1, y_2), \dots, \phi_y(y_{n-1}, y_n))$$

$$\langle \phi(x), \theta \rangle = \sum_i \langle \phi_x(x_i, y_i), \theta_x \rangle + \sum_i \langle \phi_y(y_i, y_{i+1}), \theta_y \rangle$$

$$g(\theta|x) = \sum_y \prod_i f_i(y_i, y_{i+1}) \text{ where}$$

$$f_i(y_i, y_{i+1}) = e^{\langle \phi_x(x_i, y_i), \theta_x \rangle + \langle \phi_y(y_i, y_{i+1}), \theta_y \rangle}$$

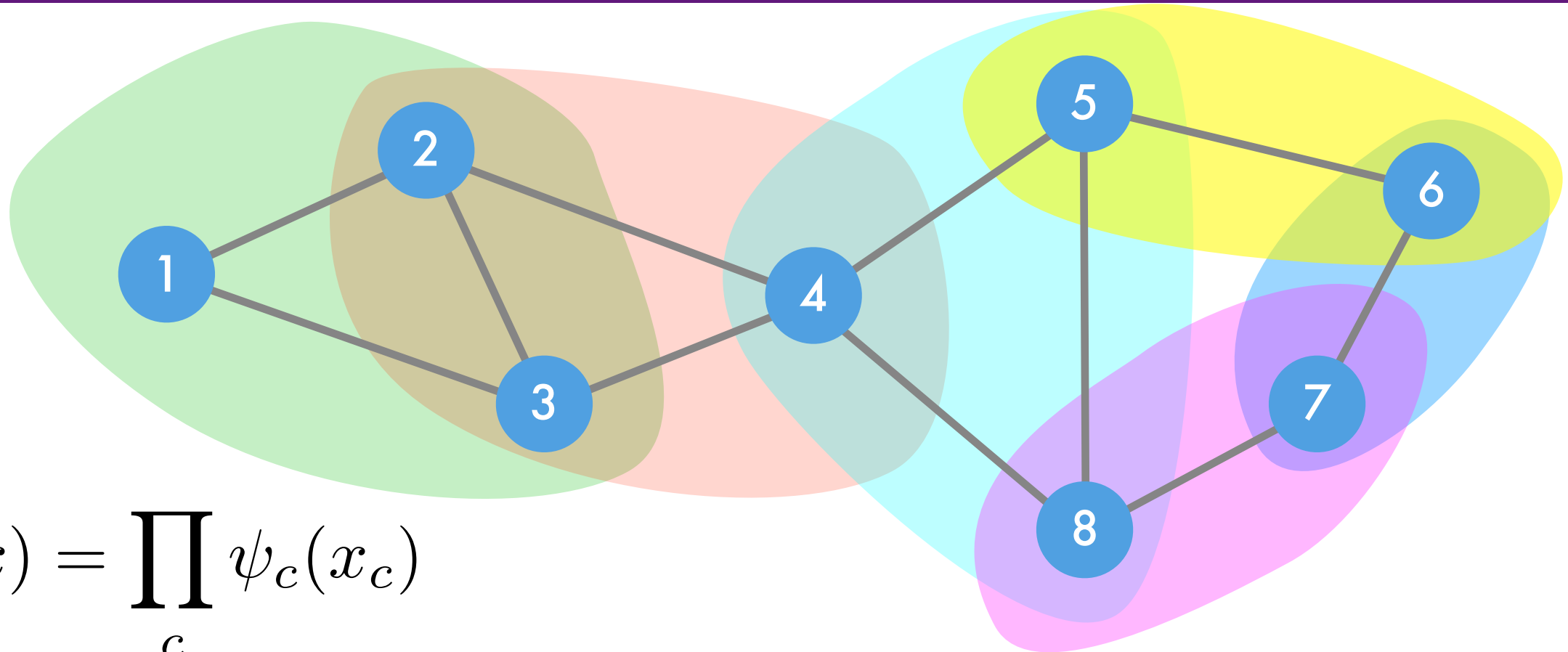
**dynamic  
programming**

# Conditional Random Fields

- Compute distribution over marginal and adjacent labels
- Take conditional expectations
- Take update step (batch or online)
  
- More general techniques for computing normalization via message passing ...

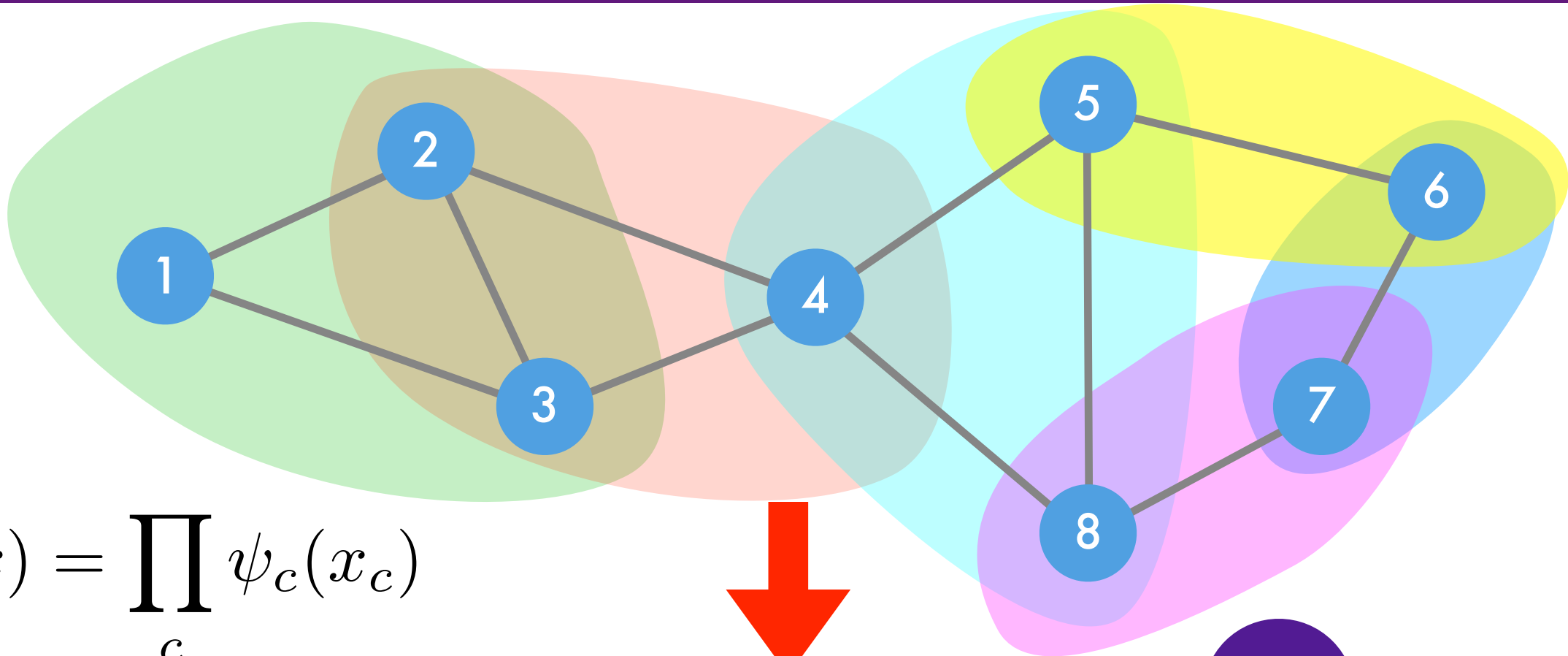
# Dynamic Programming + Message Passing

# Clique Graph

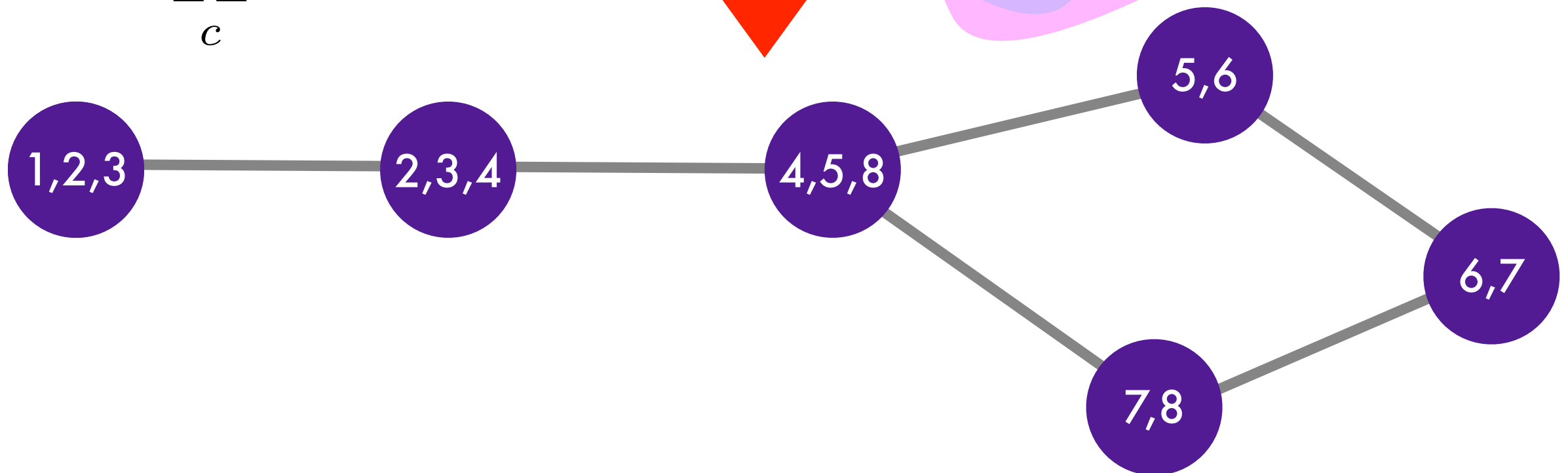
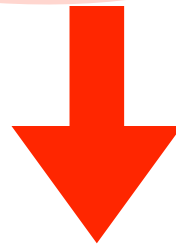


$$p(x) = \prod_c \psi_c(x_c)$$

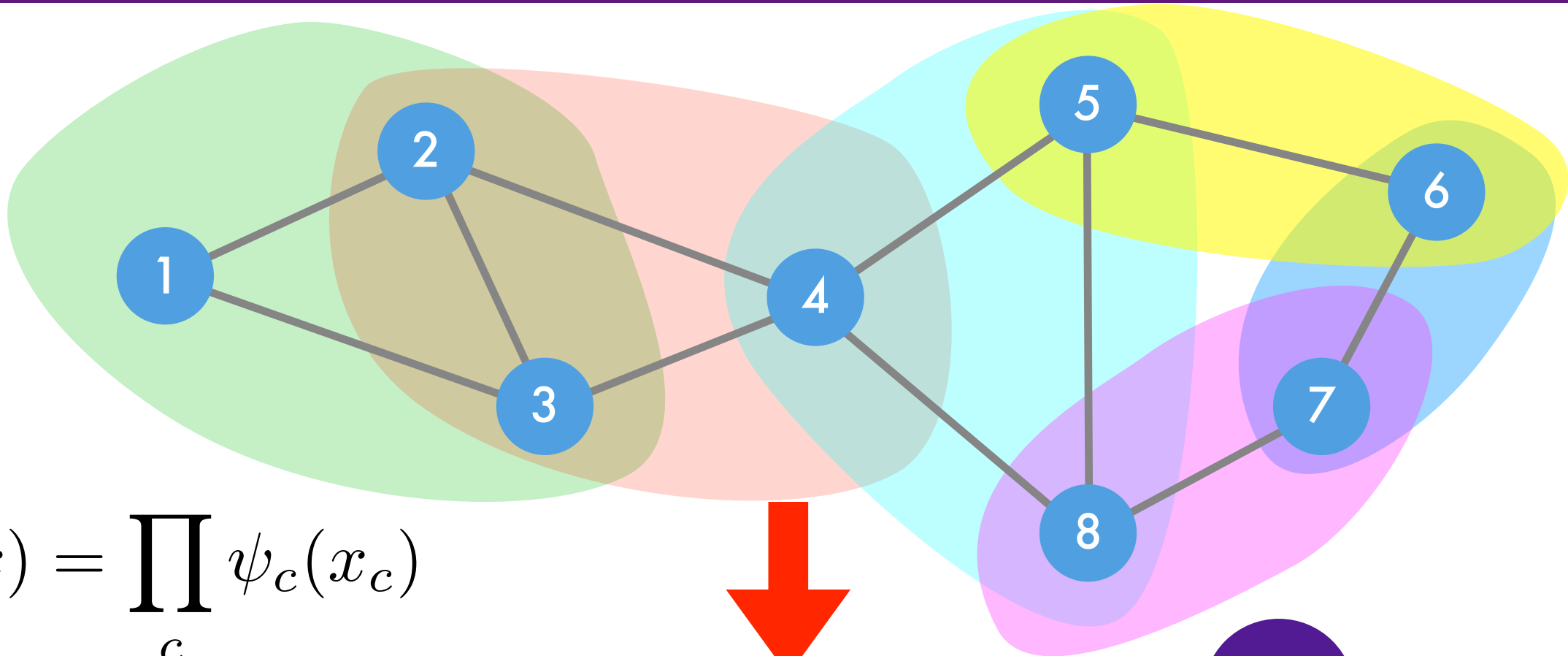
# Clique Graph



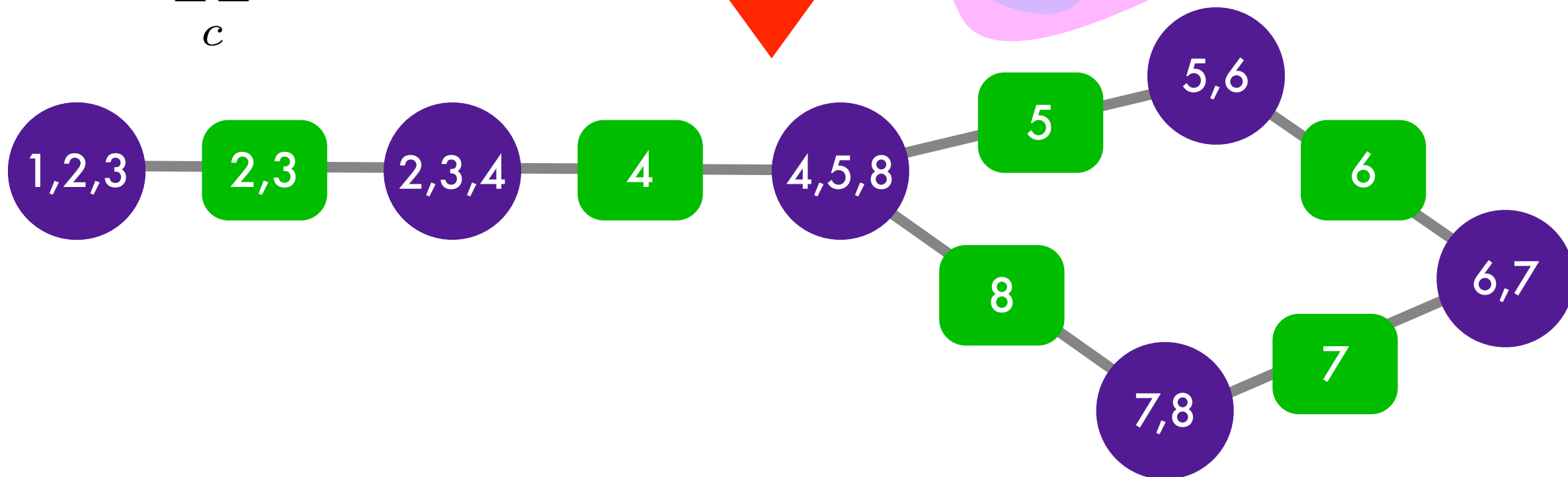
$$p(x) = \prod_c \psi_c(x_c)$$



# Clique Graph

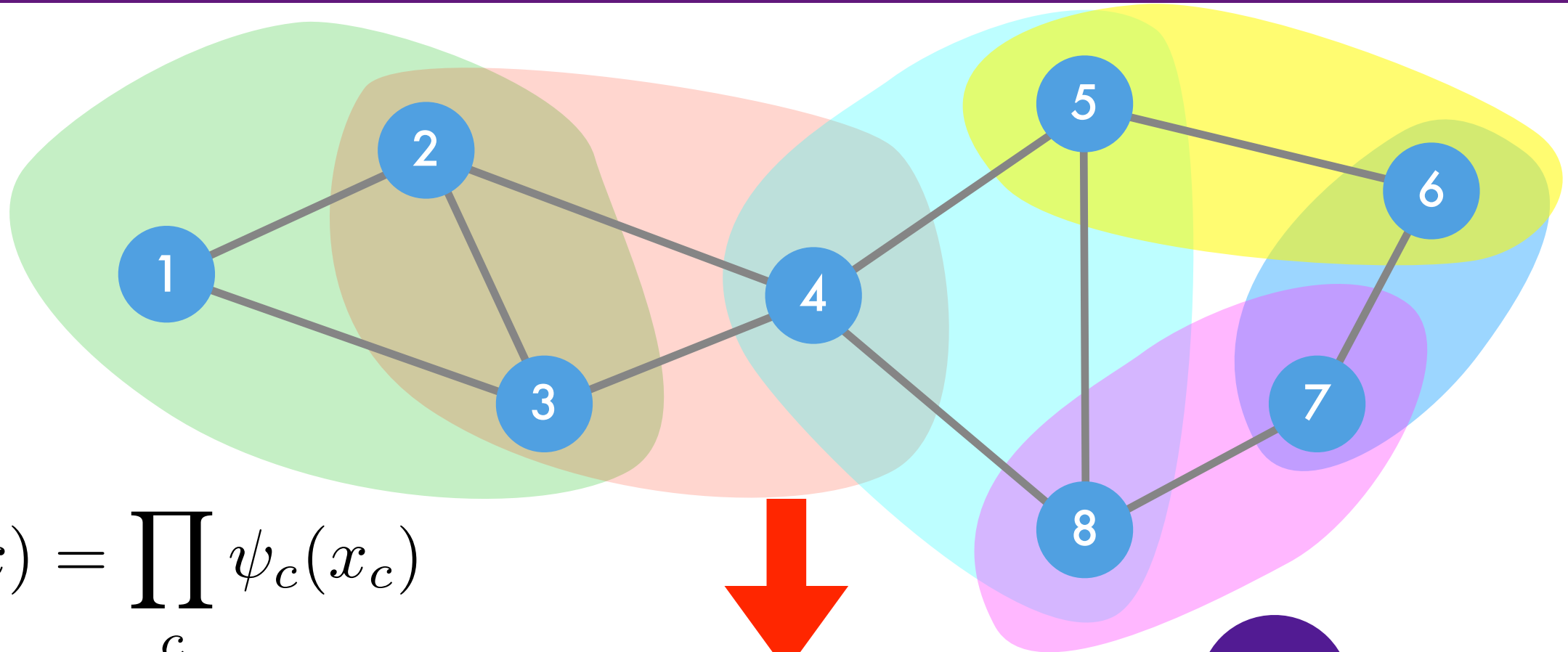


$$p(x) = \prod_c \psi_c(x_c)$$

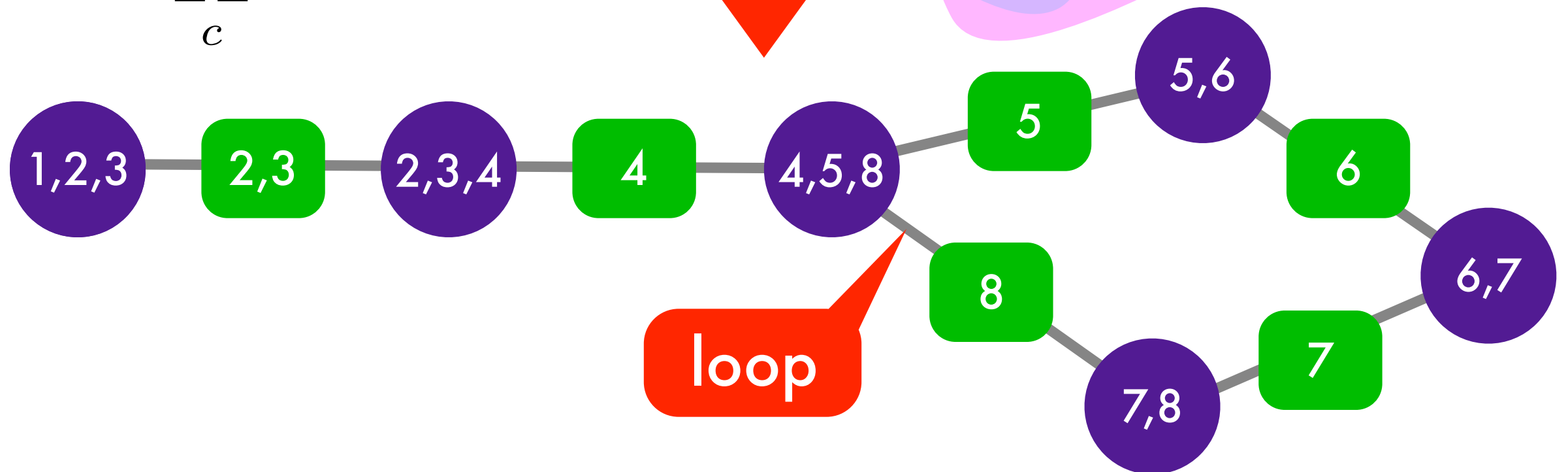




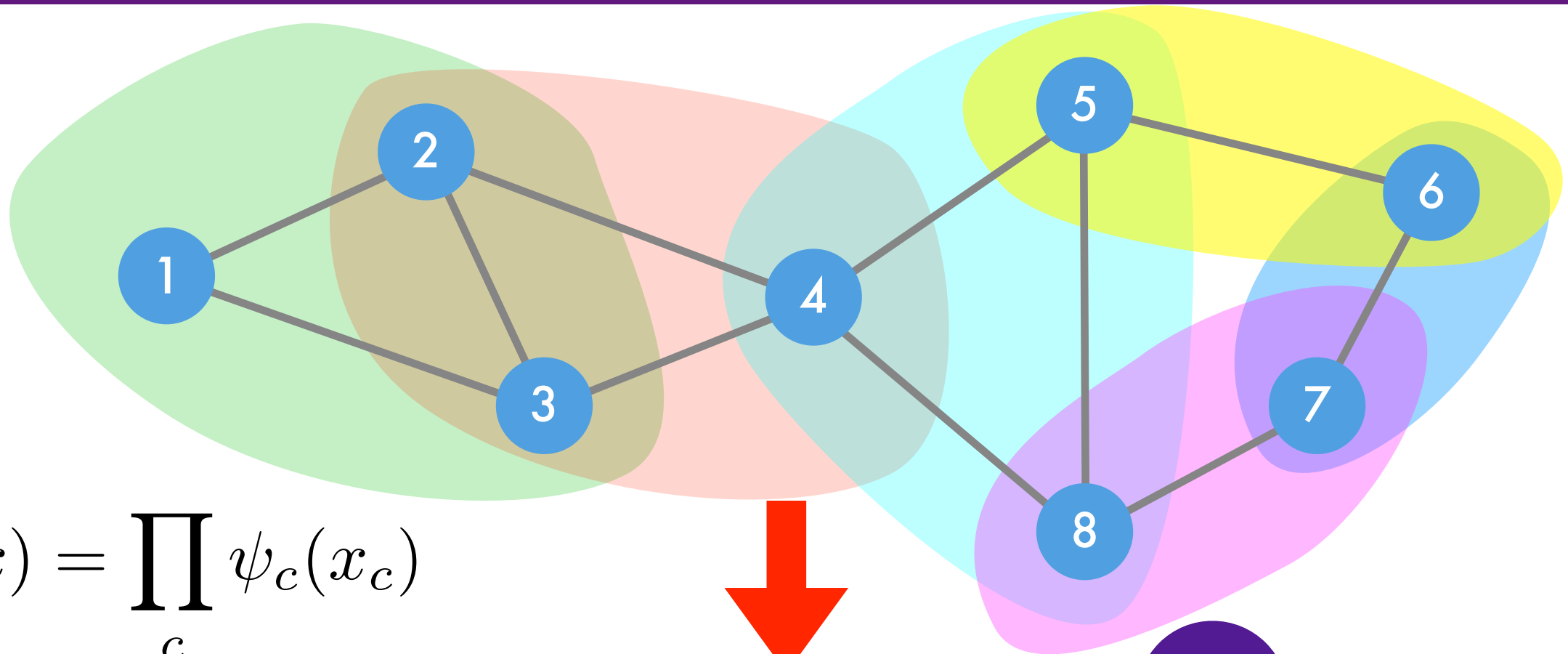
# Clique Graph



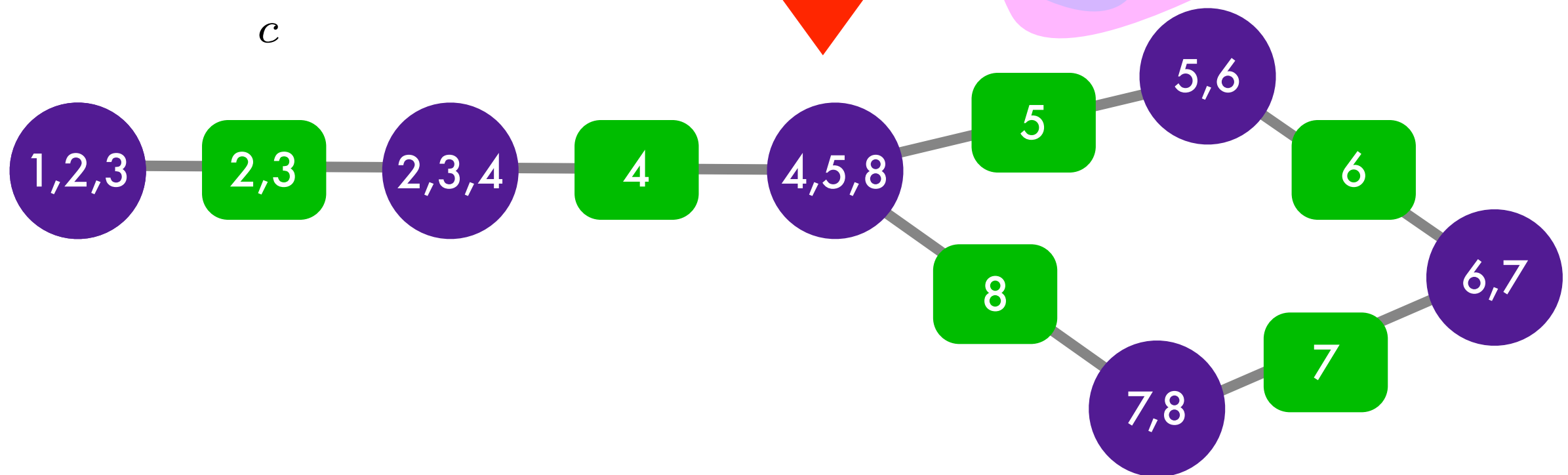
$$p(x) = \prod_c \psi_c(x_c)$$



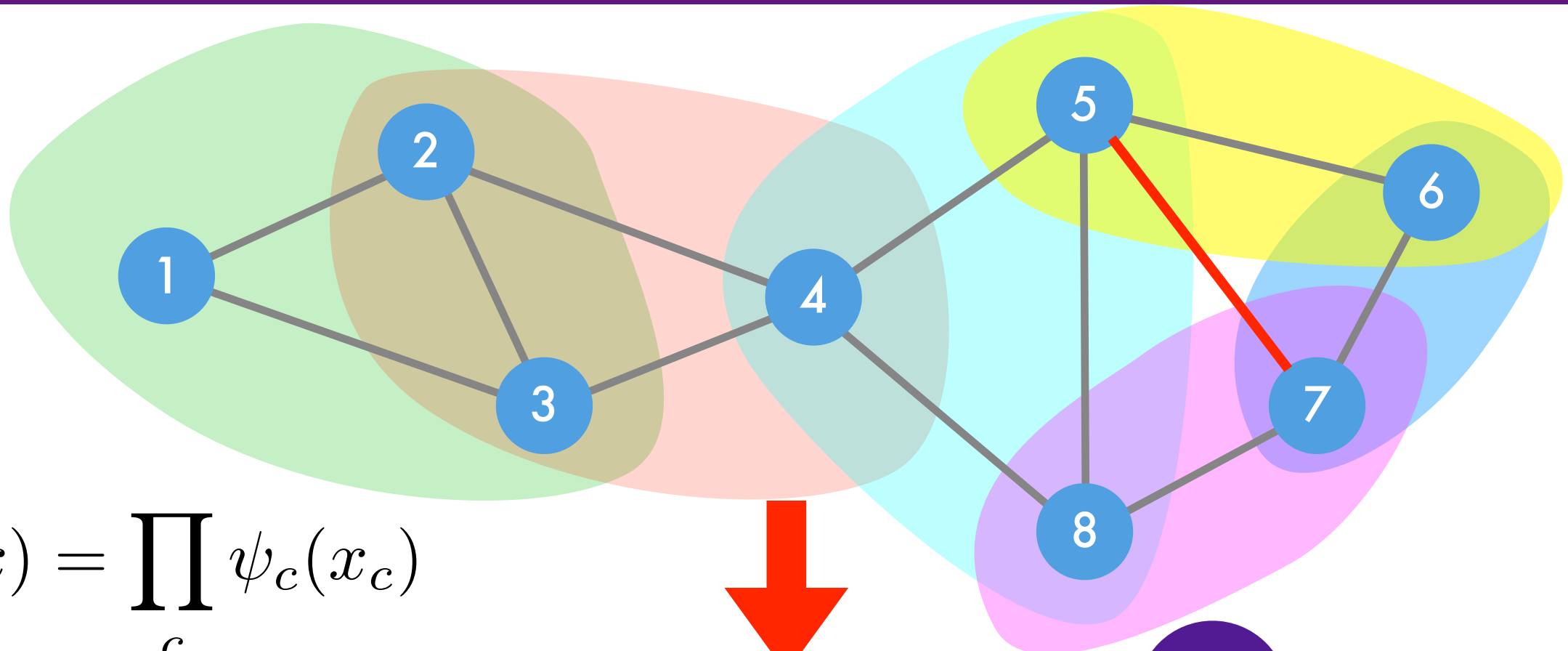
# Junction Tree / Triangulation



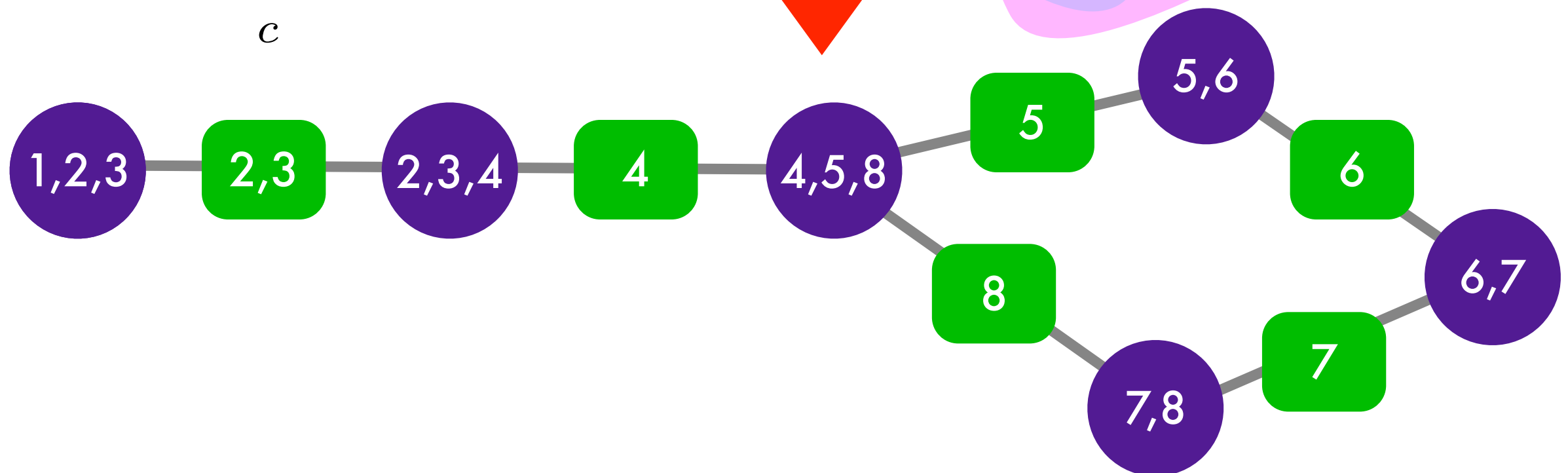
$$p(x) = \prod_c \psi_c(x_c)$$



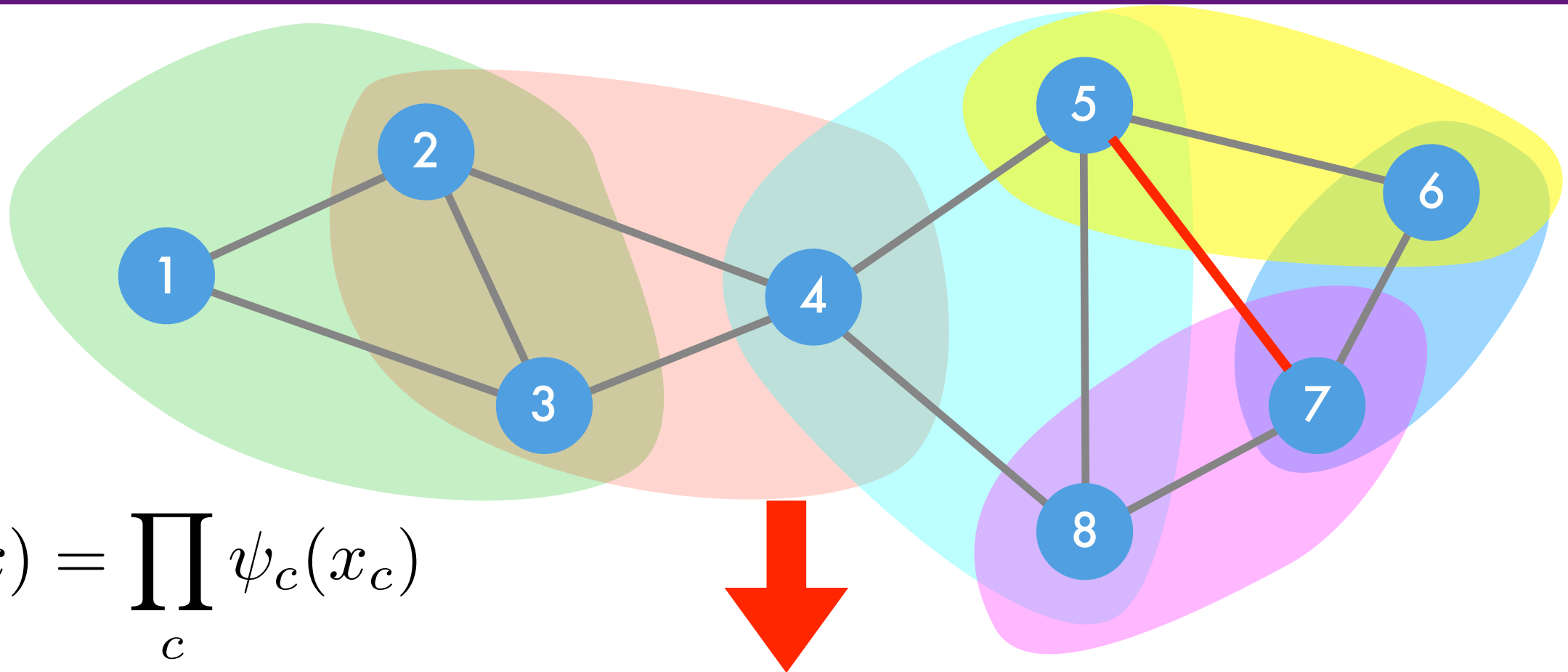
# Junction Tree / Triangulation



$$p(x) = \prod_c \psi_c(x_c)$$



# Junction Tree / Triangulation

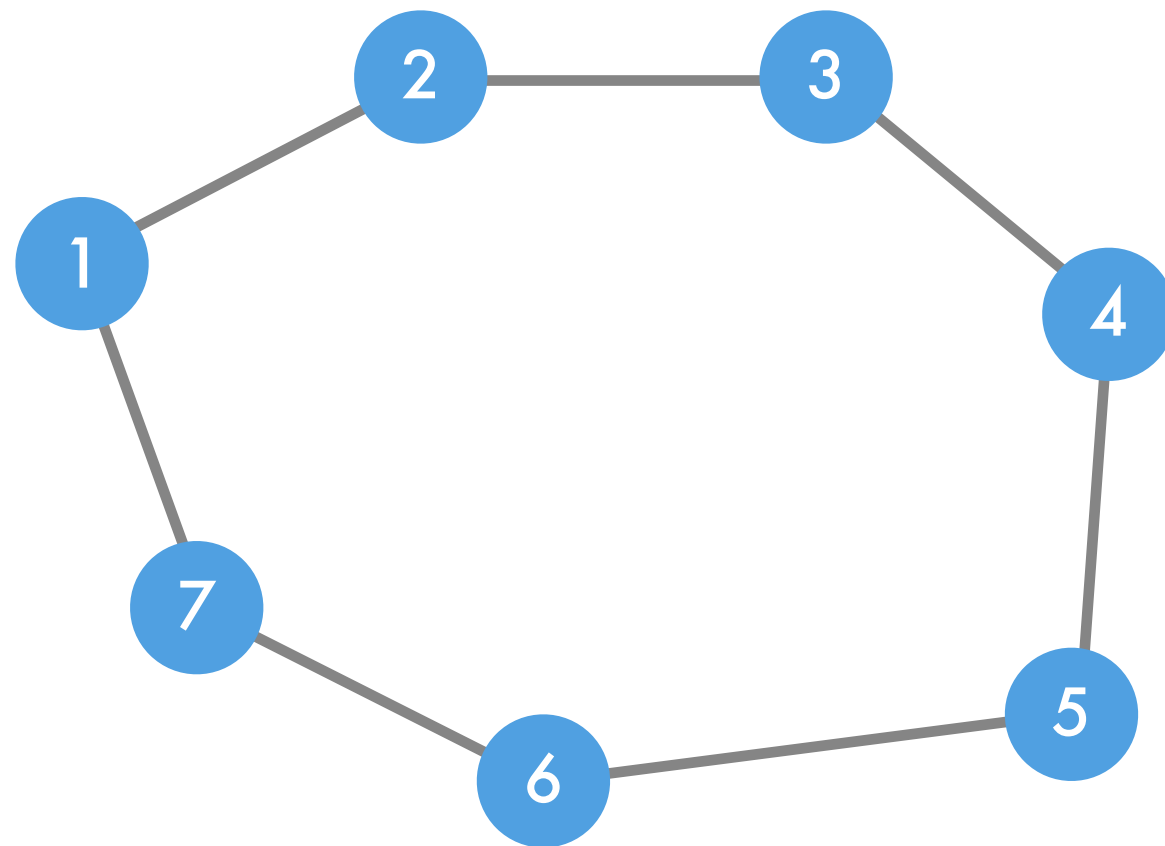


$$p(x) = \prod_c \psi_c(x_c)$$



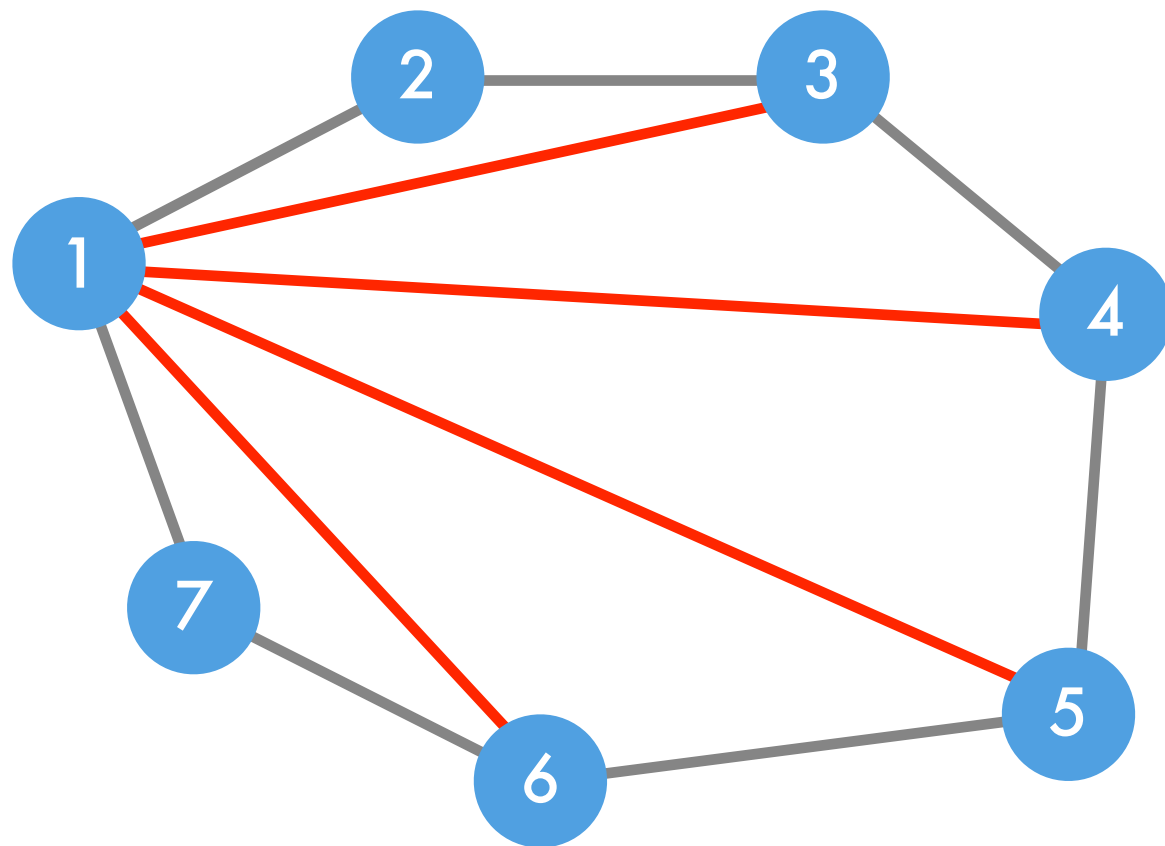
message passing possible

# Triangulation Examples



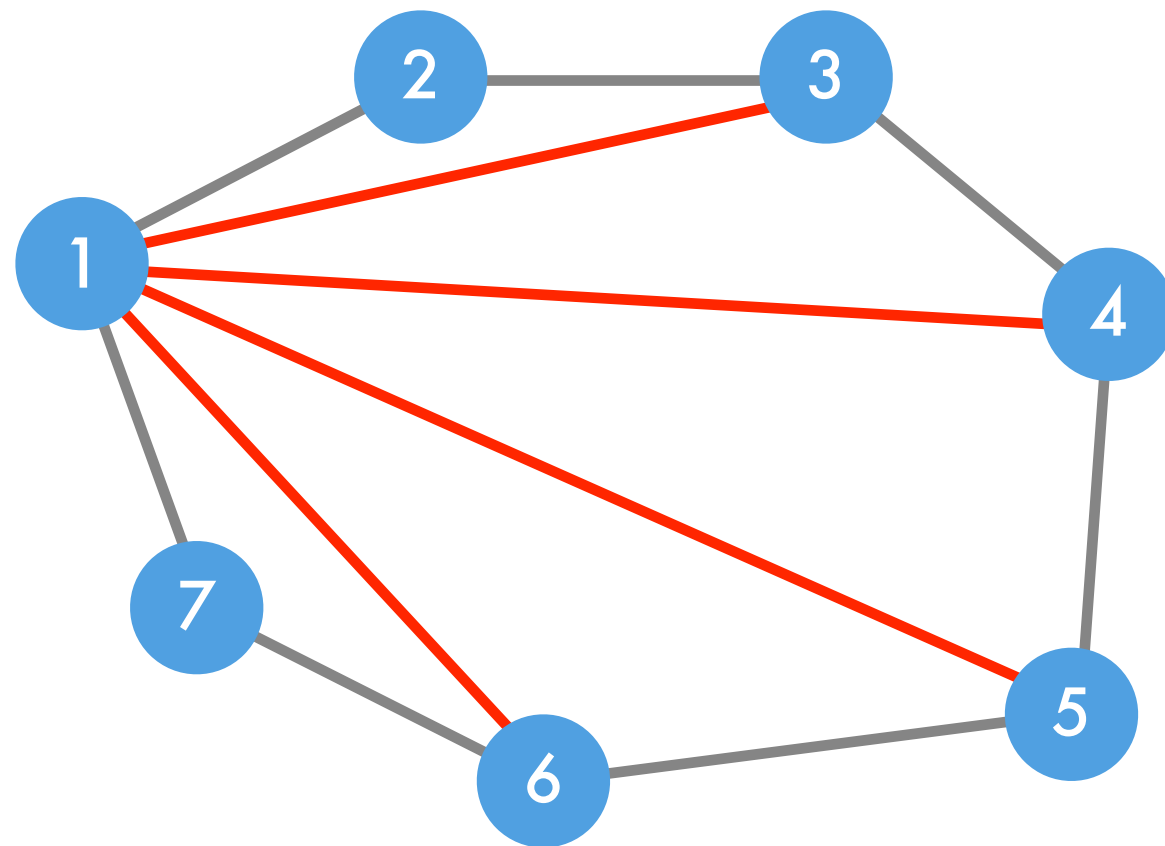
- **Clique size increases**
- **Separator set size increases**

# Triangulation Examples



- **Clique size increases**
- **Separator set size increases**

# Triangulation Examples



- **Clique size increases**
- **Separator set size increases**



# Message Passing



- **Joint Probability**

$$p(x) \propto \psi(x_1, x_2, x_3)\psi(x_1, x_3, x_4)\psi(x_1, x_4, x_5)\psi(x_1, x_5, x_6)\psi(x_1, x_6, x_7)$$

- **Computing the normalization**

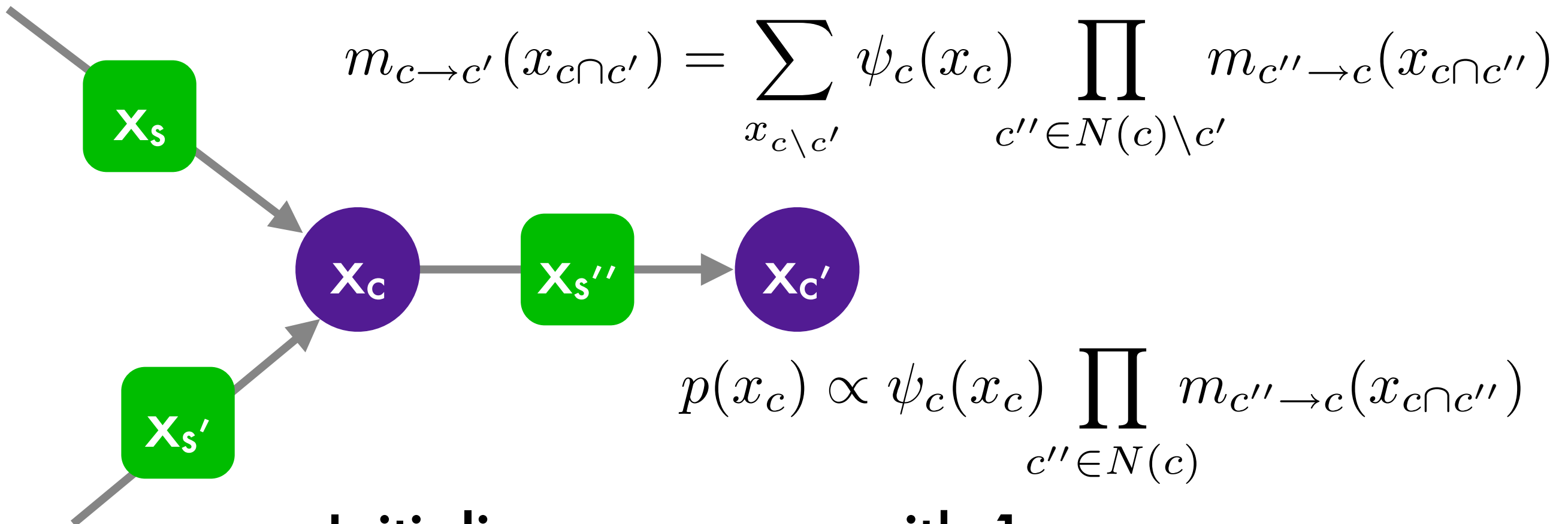
$$m_{\rightarrow}(x_1, x_3) = \sum_{x_2} \psi(x_1, x_2, x_3)$$

$$m_{\rightarrow}(x_1, x_4) = \sum_{x_3} m_{\rightarrow}(x_1, x_3)\psi(x_1, x_3, x_4)$$

$$m_{\rightarrow}(x_1, x_5) = \sum_{x_4} m_{\rightarrow}(x_1, x_4)\psi(x_1, x_4, x_5)$$



# Message Passing



- Initialize messages with 1
- Guaranteed to converge for (junction) trees
- Works well in practice even for loopy graphs
- Only local computations are required

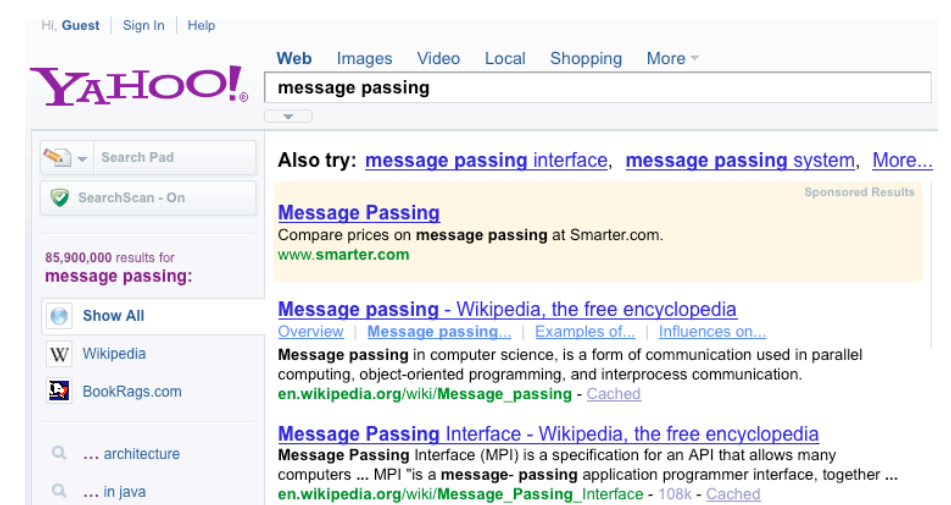
# Message Passing in Practice

- Incoming messages contain aggregate uncertainty from neighboring random variables
- Message passing combines and transmits this information **in both directions**

crawler

phase 1  
ranker

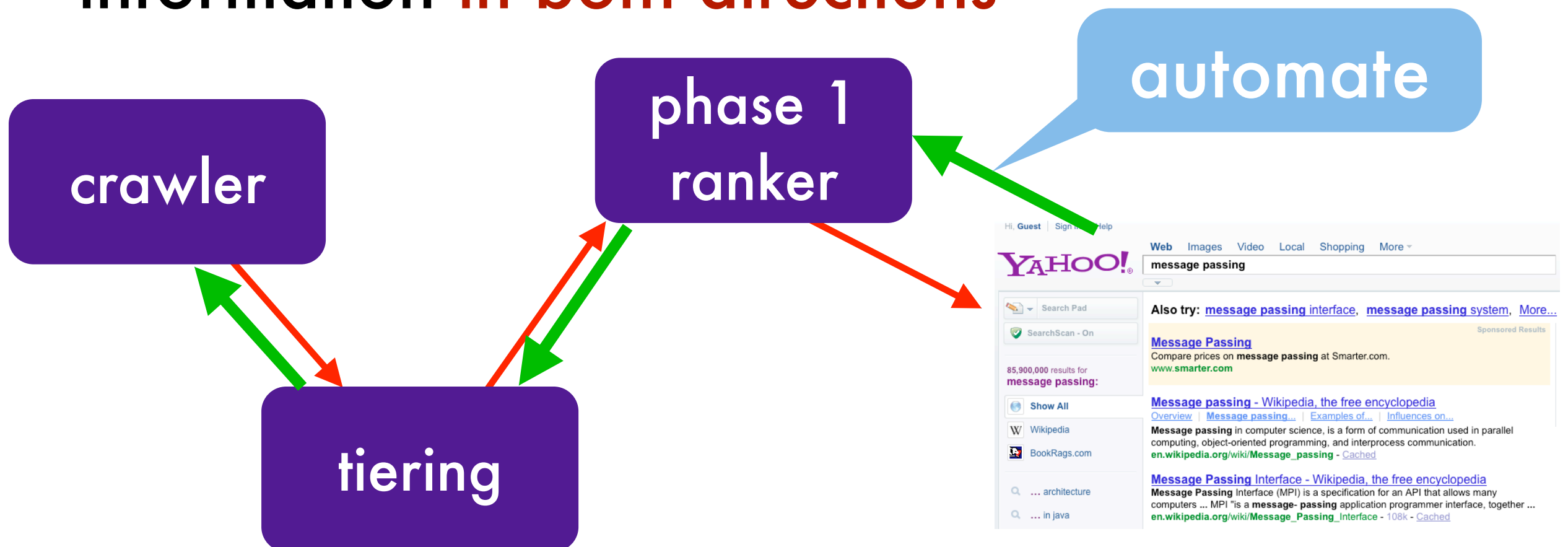
tiering



The screenshot shows a Yahoo! search results page for the query "message passing". The page includes the Yahoo! logo, navigation links (Web, Images, Video, Local, Shopping, More...), and search filters (Search Pad, SearchScan - On). The search results show 85,900,000 results for "message passing". The top result is a sponsored advertisement from Smarter.com titled "Message Passing" with the text "Compare prices on message passing at Smarter.com. www.smarter.com". Below the ad are two organic search results from Wikipedia: "Message passing - Wikipedia, the free encyclopedia" and "Message Passing Interface - Wikipedia, the free encyclopedia".

# Message Passing in Practice

- Incoming messages contain aggregate uncertainty from neighboring random variables
- Message passing combines and transmits this information **in both directions**



# 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](http://alex.smola.org) / [blog.smola.org](http://blog.smola.org)**

# Today's mission

**Find structure in the data**

Human understandable  
Improved knowledge for estimation