



Scalable Machine Learning

Models, Architectures and Algorithms

Alexander Smola
Carnegie Mellon University & Google
[@smolix](mailto:alex.smola.org)

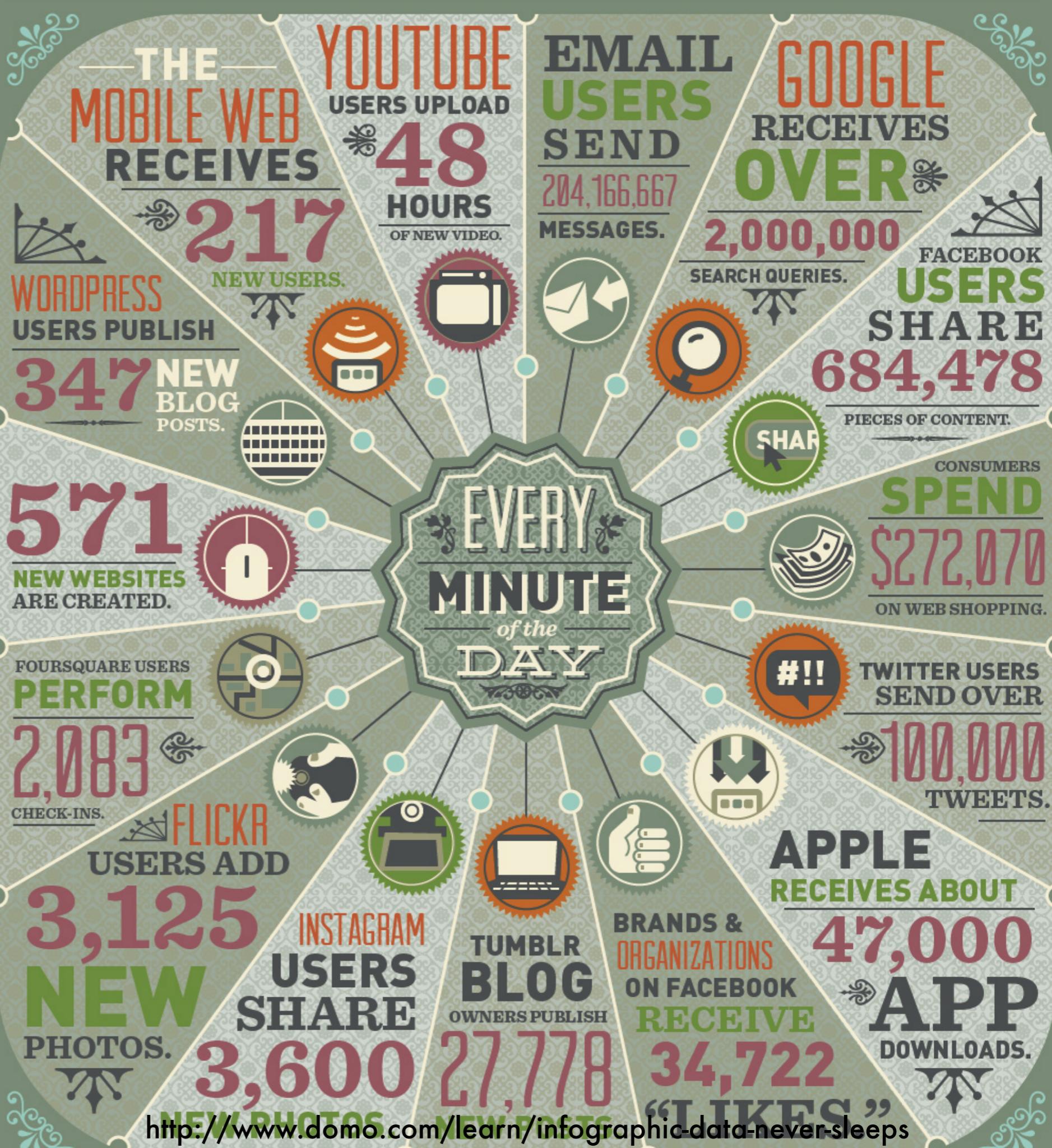
Outline

- **Data**
Actions, Interactions, User generated content
- **Architectures**
MapReduce, Graphs, Streams, Parameterserver
- **Models and Algorithms**
 - Logistic regression (advertising, search)
Distributed proximal gradient
 - Topic models (personalization, profiling)
Stochastic variational
 - Randomized algorithms for large models
Hash kernel, Fastfood, Alias sampling, Sketches

A portrait of the character Data from Star Trek: The Next Generation. He is wearing his iconic dark blue uniform with a yellow collar and a blue-striped shirt underneath. His characteristic white hair is visible. The background is dark, with some glowing blue lights visible in the upper right corner.

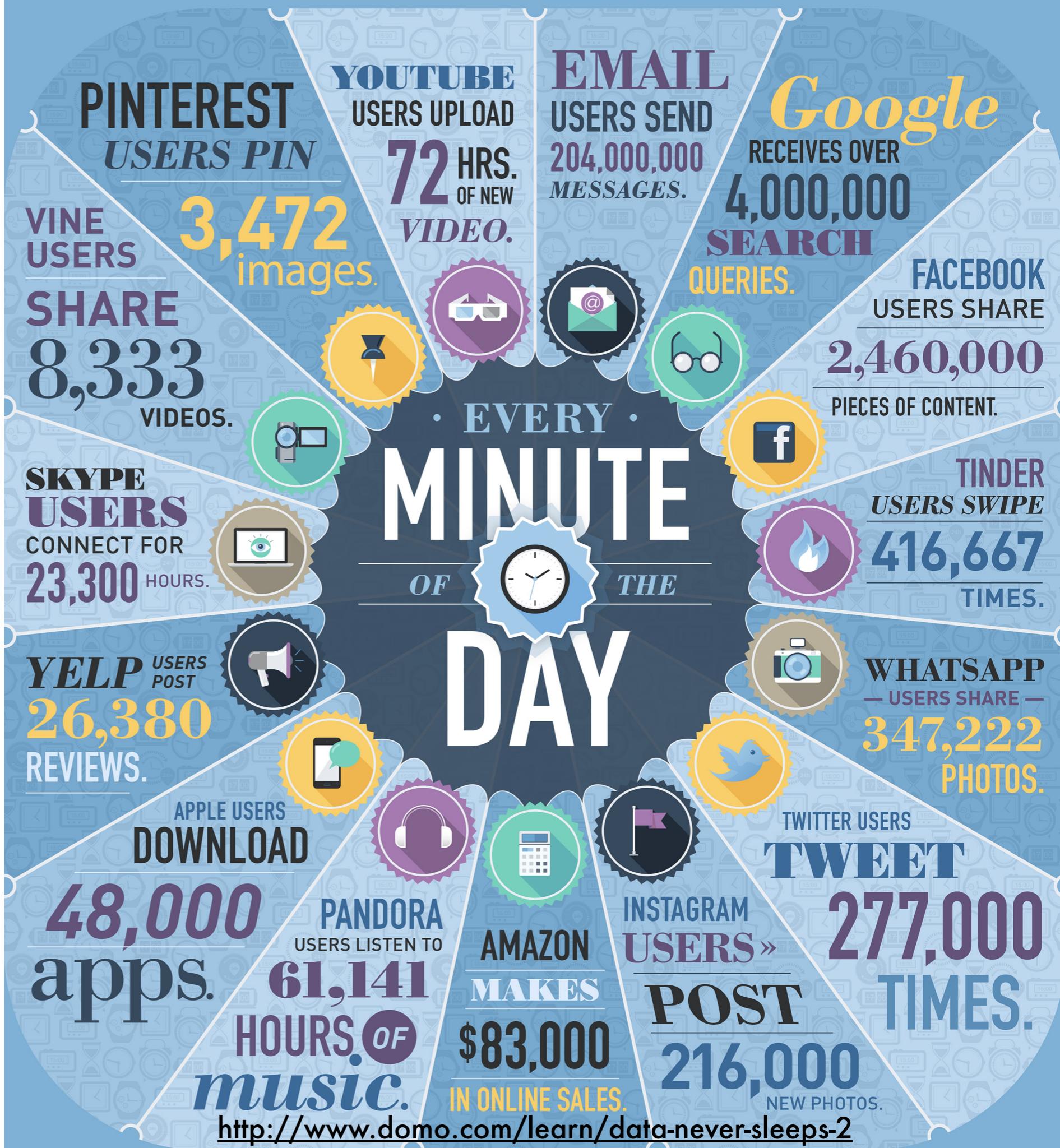
Data

Data per minute 2012



Data per minute 2014

Google



Computational Advertising

A screenshot of a Google search results page for the query "mesothelioma". The search bar at the top contains "mesothelioma". Below it, a navigation bar offers options: Web (selected), Videos, News, Images, Books, More, and Search tools. A message indicates "About 2,970,000 results (0.21 seconds)". The results are organized into two columns. The left column contains five organic search results and one sponsored ad. The right column contains four organic search results and one sponsored ad. The sponsored ads are highlighted with a black border.

Mesothelioma Compensation
Ad www.nationalmesotheliomaclaims.com/
The Money's Already There. \$30 Billion Asbestos Trust Fund
What Is Mesothelioma? - National Claims Center - Mesothelioma Claims

Mesothelioma Symptoms - Mesothelioma-Answers.org
Ad www.mesothelioma-answers.org/
By Anna Kaplan, M.D. 101 Facts about Mesothelioma.
Asbestos - Treatments - Top Doctors - Free Mesothelioma Book

CA Mesothelioma Resource - californiamesothelioma.com
Ad www.californiamesothelioma.com/ (800) 259-9249
Learn about mesothelioma & receive a free book of helpful answers.
What is Mesothelioma? - Asbestos Exposure in CA - California Legal Rights

Mesothelioma Cancer - Mesothelioma.com
www.mesothelioma.com/mesothelioma/
by Dr. Howard Jack West - Apr 2, 2014 - Mesothelioma is an aggressive cancer affecting the membrane lining ... Between 50 and 70% of all mesotheliomas are of the epithelial variety.
Mesothelioma Symptoms - Mesothelioma Prognosis - Mesothelioma Survival Rate

Mesothelioma - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mesothelioma Wikipedia
Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Asbestos - Mesothelium - Paul Kraus - Category:Mesothelioma

Mesothelioma
Ads ⓘ www.mesothelioma-attorney-locators.com/
Easily Find Mesothelioma Attorneys.
Locations Across The United States

CA Mesothelioma
www.mesotheliomatreatmentcenters.org/
Mesothelioma? Get the Money you Deserve Fast-Help Filing your Claim

Mesothelioma Compensation
www.mesotheliomaclaimscenter.info/ (877) 456-3935
Mesothelioma? Get Money You Deserve Fast! Get Help with Filing a Claim.

California Mesothelioma
[\(888\) 707-4525](http://www.mesotheliomaattorney-usa.com/Legal)
100% Free Mesothelioma Legal Help!
\$30 Billion Trust Fund Available.

Mesothelioma
meso.lawyers.local.alotresults.com/
Seasoned Lawyers in your Area.
In your Local Lawyer Listings!

sponsored
search picks
position of
ad using

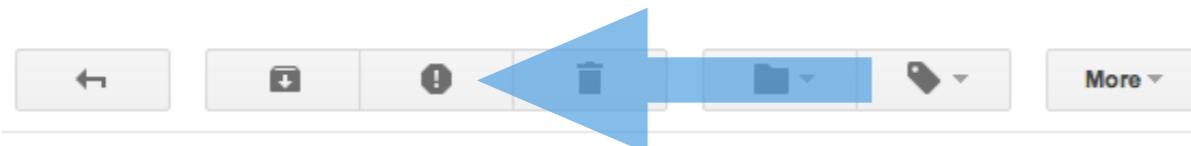
$$p(\text{click}|\text{ad}) \cdot \text{bid}(\text{ad})$$

estimate it

4 million/minute

Carnegie Mellon University

Spam filtering



200 million/minute

Upcoming MLSS Volunteer tasks Inbox

Mallory Deptola Hi Volunteers! Again, thanks for volunteering to help out during MLSS 2014. W... 8:04 am (2 days ago)

Mallory Deptola via smola.org to Alex, Zico 8:51 am (2 days ago)

Hi Guys,

I was wondering if you could add the audio/visual tasks to the volunteer spreadsheet. I am not sure how you would like to go about handling them – would it be based on the speaker schedule that they would need to man the camera? How many people per recording?

If you just wanted to those tasks to the list, that'd be great! <https://docs.google.com/spreadsheets/d/1fawSYWppJARcvmk-PpOfvoli9XvVq6fcY6fAF4-2Qno/edit#gid=0>

imbalanced
dataset

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)		
<input type="checkbox"/> EM Office	Donation To You: - Hello Dear, This is a personal email directed to you by Chris and Colin Weir. Chris and Colin Weir	11:48 am
<input type="checkbox"/> 钟	{Spam?} hwitek: 请审批 - Hi hwitek; 研发人员的考核与激励是企业高层领导、研发经理、人力资源经理最为头疼的问题之一，	8:49 am
<input type="checkbox"/> GMEE2014	[Conference Notification] (July 3, 2014 G--M--E--E--2014 EI & ISTP) - GMEE2014 September 21-22 2014 Internation	8:10 am
<input type="checkbox"/> Esther	TTP_EI Compendex and ISTP index GMEE (Green Materials and Environmental Engineering) - GMEE2014 September	8:03 am
<input type="checkbox"/> EachBuyer	EachBuyer Deals For Jun.2014.Vol.19 - If you are unable to see the message below, click here to view. If you do not wish to	Jun 27
<input type="checkbox"/> CUUE2014	-Civil, Urban and Environment→ EI&ISTP ---C-U-E-E-2014- * →submission due: July 12- - 2014 International Conference	Jun 27
<input type="checkbox"/> EachBuyer	Up to 90% off! End This Week. - To unsubscribe please click here. EachBuyer Email not displaying correctly? Click here to	Jun 27
<input type="checkbox"/> Call For Papers	IEEE Big Data 2014 paper submission deadline is extended to July 13, 2014 - We have received many requests to extenc	Jun 27
<input type="checkbox"/> Tara Alngindabu	Up up and away - Make sure you always get all of our most sensational deals. Add JS Design to your address book today.	Jun 27
<input type="checkbox"/> Dan Roy	page for Alex - http://xn-12cu8ak3e3dxdde4cn.com/akl/	Jun 27
<input type="checkbox"/> kisman@emirates.ae	Please I want to propose something important to you. - Dear Friend, I am Barrister Krishnan Al-Qassimi, an attorney at law	Jun 26
<input type="checkbox"/> Delta Air Lines	Flights Reminder: Your June Medallion STATEMENT - An Insider look at travel, deals and your account. JUNE 2014 Hell	Jun 26
<input type="checkbox"/> ACSIJ Journal	ACSIJ Journal Call for Papers July 2014 - Call for Papers Advances in Computer Science : an International Journal (ACSIJ)	Jun 26

Recommendation & Ranking

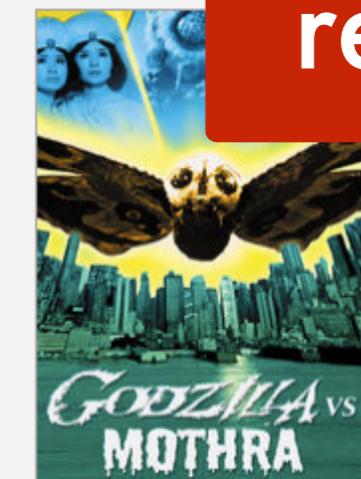
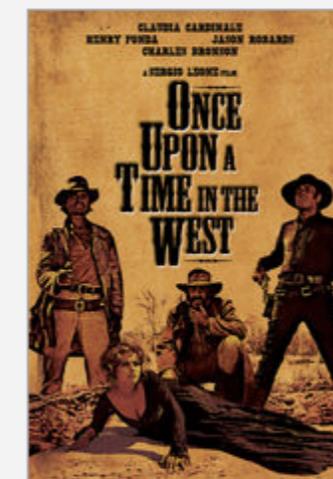
maximize interaction probability for whole page

Foreign Movies >

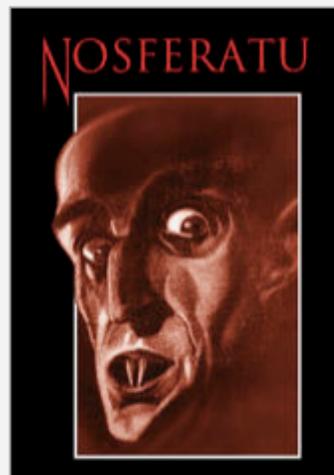
Classic Foreign Movies

Subgenres ▾

Sort by Suggestions for You ▾



really?



Time series & trends

Topics

[Subscribe](#)



machine learning

Search term

data mining

Search term

big data

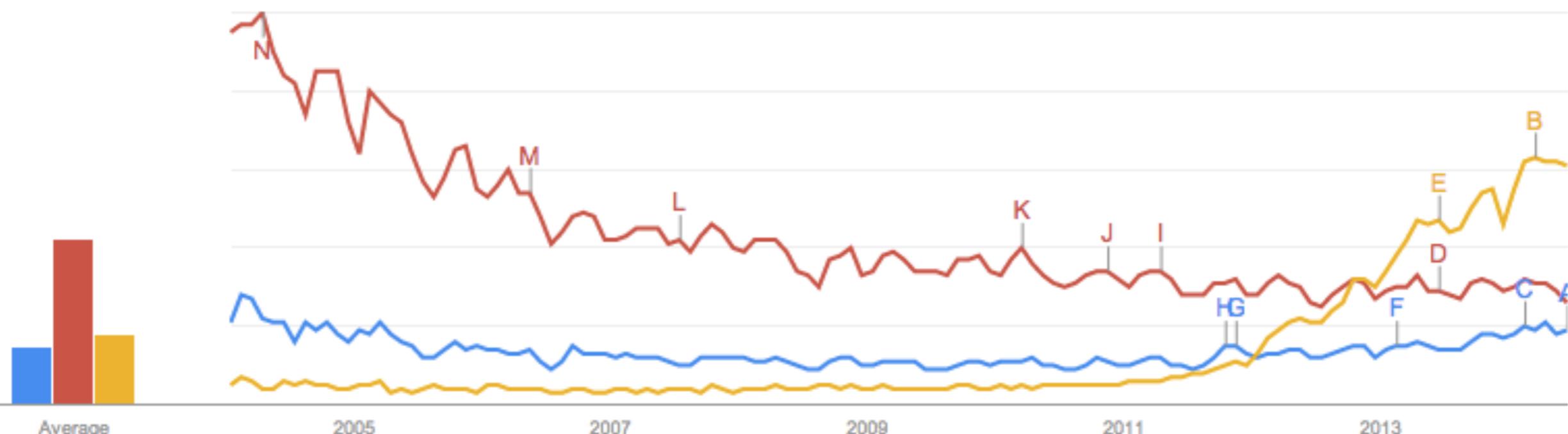
Search term

+Add term

Interest over time

News headlines

Forecast



More data

- News articles & events (NY Times, GNews)
- Blogs / microblogs (Tumblr, Twitter, Weibo)
- Reviews (IMDB, Yelp, Amazon)
- Comments (YouTube, Reddit)
- Messages (Facebook, Hangouts, SMS)
- Graphs (Friends, Followers, Webpages)
- Information diffusion (Meme tracking)
- Spatiotemporal (GMaps, Foursquare, Twitter)

Lots more data

- Bioinformatics
DNA Microarrays, High throughput sequencing
- Astronomy
Square Kilometer Array, Radio telescopes
- Medicine
MRI / MEG scans, Connectome, Health records
- Finance (e.g. high frequency trading)
- Geophysics (e.g. oil discovery)
- Industrial process monitoring

User generated content

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



flickr™



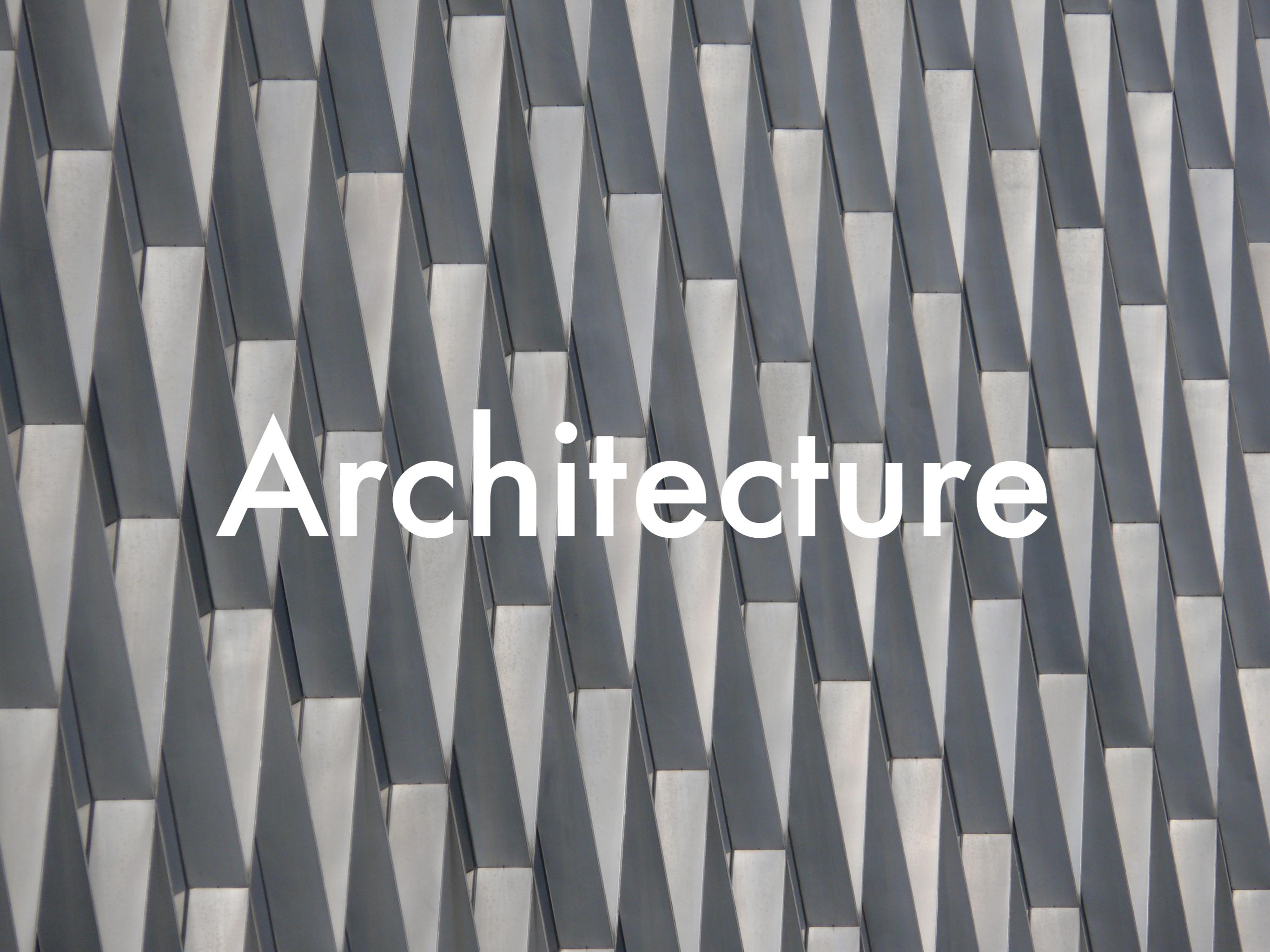
You Tube

yelp

DISQUS

Summary

- **Expensive data ≠ big data**
(1000 brain scans are expensive)
- **Big data requires big models**
(1000 parameter model on TB of data)
- Big data needs systems built for it
(don't ship data to computation)
- **Vast range of problem domains**
- **Vast range of statistical models**



Architecture

Real Hardware



Machines

- CPU

- 8-64 cores (Intel/AMD servers)
 - 2-3 GHz (close to 1 IPC per core peak) - over 100 GFlops/socket
 - 8-32 MB Cache (essentially accessible at clock speed)
 - Vectorized multimedia instructions (AVX 256bit wide, e.g. add, multiply, logical)

Bulk transfer is at least 10x faster



- RAM

- 16-256 GB depending on use
 - 3-8 memory banks (each 32bit wide - atomic writes!)
 - DDR3 (up to 100GB/s per board, random access 10x slower)



- Harddisk

- 4 TB/disk
 - 100 MB/s sequential read from SATA2
 - 5ms latency for 10,000 RPM drive, 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)



The real joy of hardware

Typical first year for a new cluster:

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

Jeff Dean's Stanford slides

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

Numbers everyone should know

L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	100 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	10,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from network	10,000,000 ns
Read 1 MB sequentially from disk	30,000,000 ns
Send packet CA->Netherlands->CA	150,000,000 ns

Why a single machine is not enough

- Data (lower bounds)
 - 10-100 Billion documents (webpages, e-mails, ads, tweets)
 - 100-1000 Million users on Google, Facebook, Twitter, Hotmail
 - 1 Million days of video on YouTube
 - 100 Billion images on Facebook
- Processing capability for single machine 1TB/hour
But we have much more data
- Parameter space for models is too big for a single machine
Personalize content for many millions of users
- Process on **many cores** and **many machines simultaneously**

Cloud pricing

- Google Compute Engine and Amazon EC2

Instance type	Virtual Cores	Memory	Price (US\$)/Hour (US hosted)
n1-standard-1	1	3.75GB	\$0.070
n1-standard-2	2	7.5GB	\$0.140
n1-standard-4	4	15GB	\$0.280
n1-standard-8	8	30GB	\$0.560
n1-standard-16	16	60GB	\$1.120

\$10,000/year

- Storage

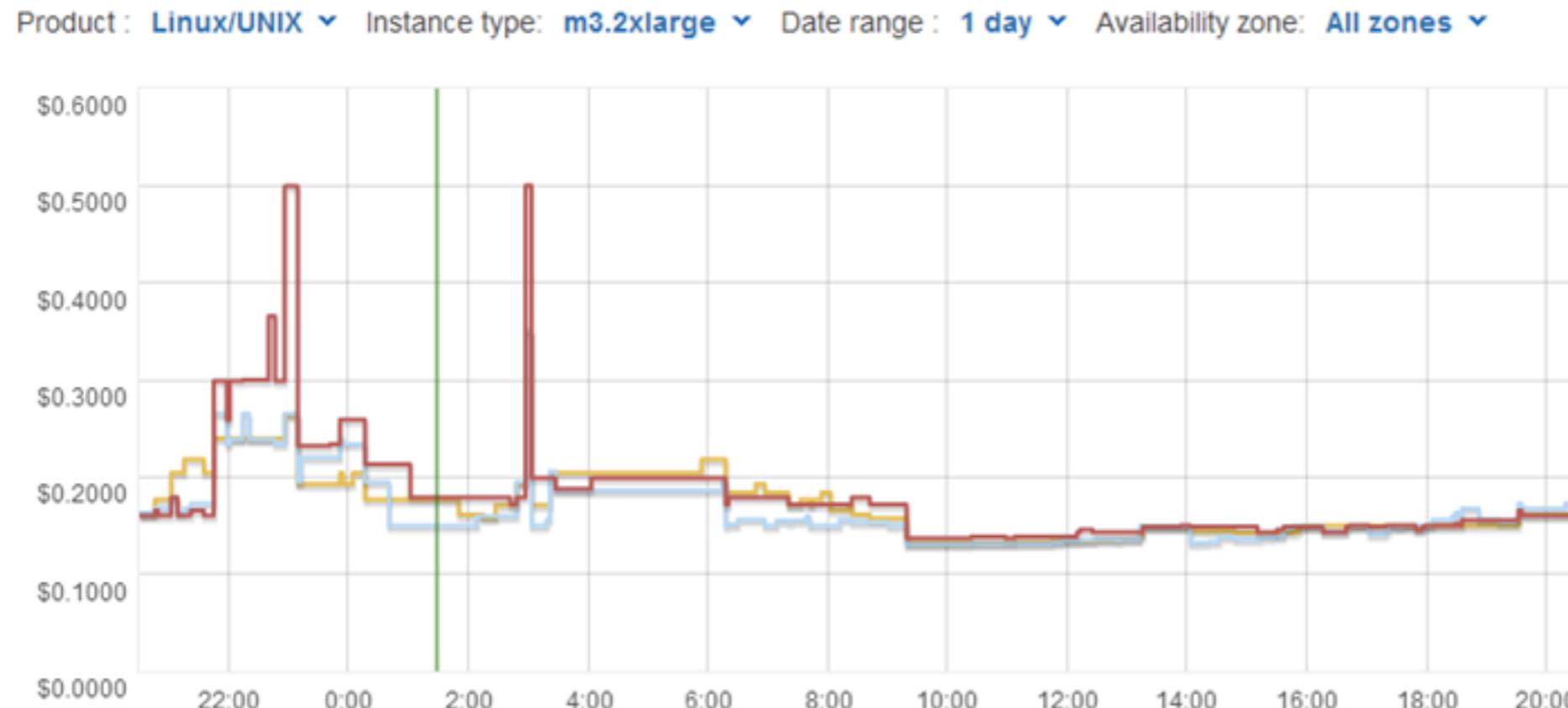
Standard Provisioned Space	\$0.04 GB / month
SSD Provisioned Space	\$0.325 GB / month
Snapshot storage	\$0.125 GB / month
IO operations	No additional charge

Spot instances
much cheaper

- Amazon EBS General Purpose (SSD) volumes
 - \$0.10 per GB-month of provisioned storage
- Amazon EBS Provisioned IOPS (SSD) volumes
 - \$0.125 per GB-month of provisioned storage
 - \$0.10 per provisioned IOPS-month
- Amazon EBS Magnetic volumes
 - \$0.05 per GB-month of provisioned storage
 - \$0.05 per 1 million I/O requests
- Amazon EBS Snapshots to Amazon S3
 - \$0.095 per GB-month of data stored

Real Hardware

- can and will fail
- Spot instances much cheaper (but can lead to preemption). Design algorithms for it!





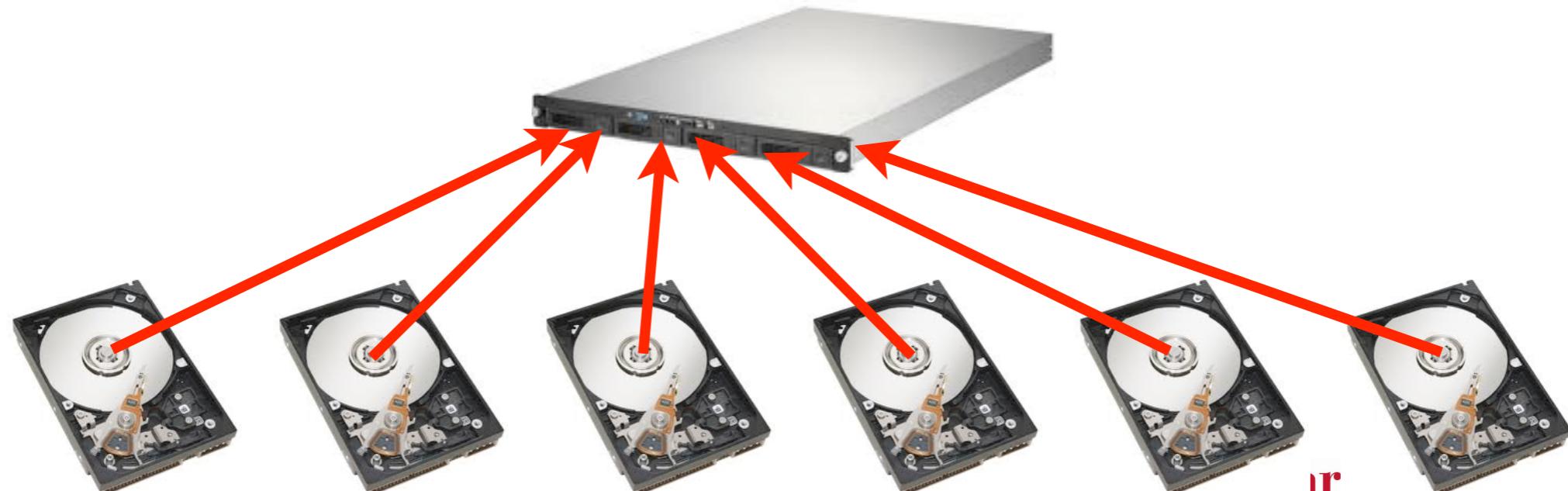
work & storage



File systems

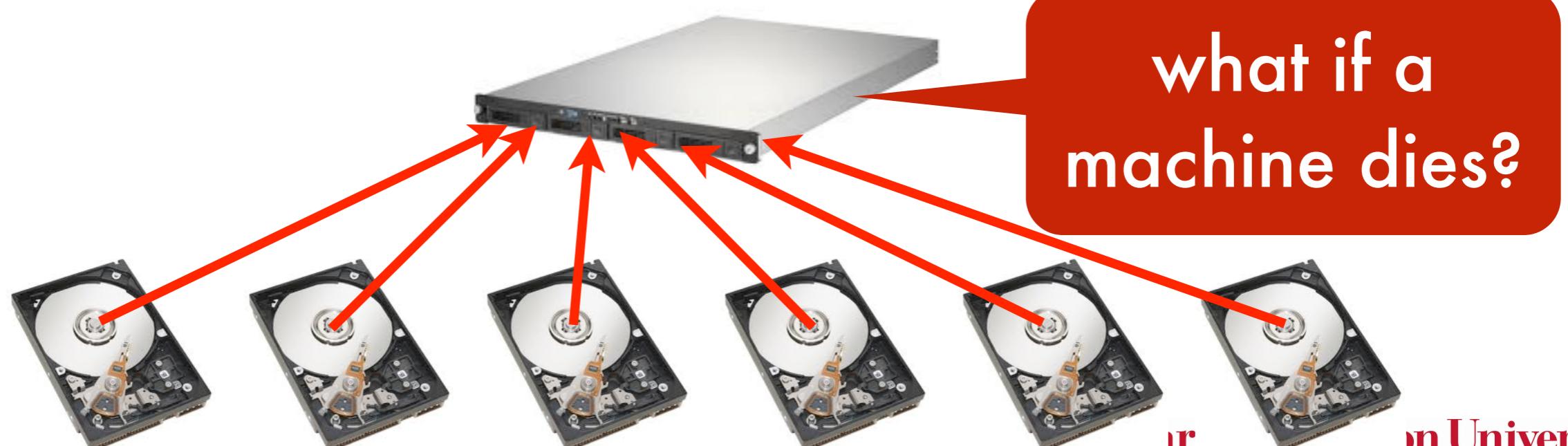
RAID

- Redundant array of inexpensive disks (optional fault tolerance)
 - Aggregate storage of many disks
 - Aggregate bandwidth of many disks
- RAID 0 - stripe data over disks (**good bandwidth, faulty**)
- RAID 1 - mirror disks (mediocre bandwidth, **fault tolerance**)
- RAID 5 - stripe data with 1 disk for parity (**good bandwidth, fault tolerance**)
- Even better - use error correcting code for fault tolerance,
e.g. (4,2) code, i.e. two disks out of 6 may fail



RAID

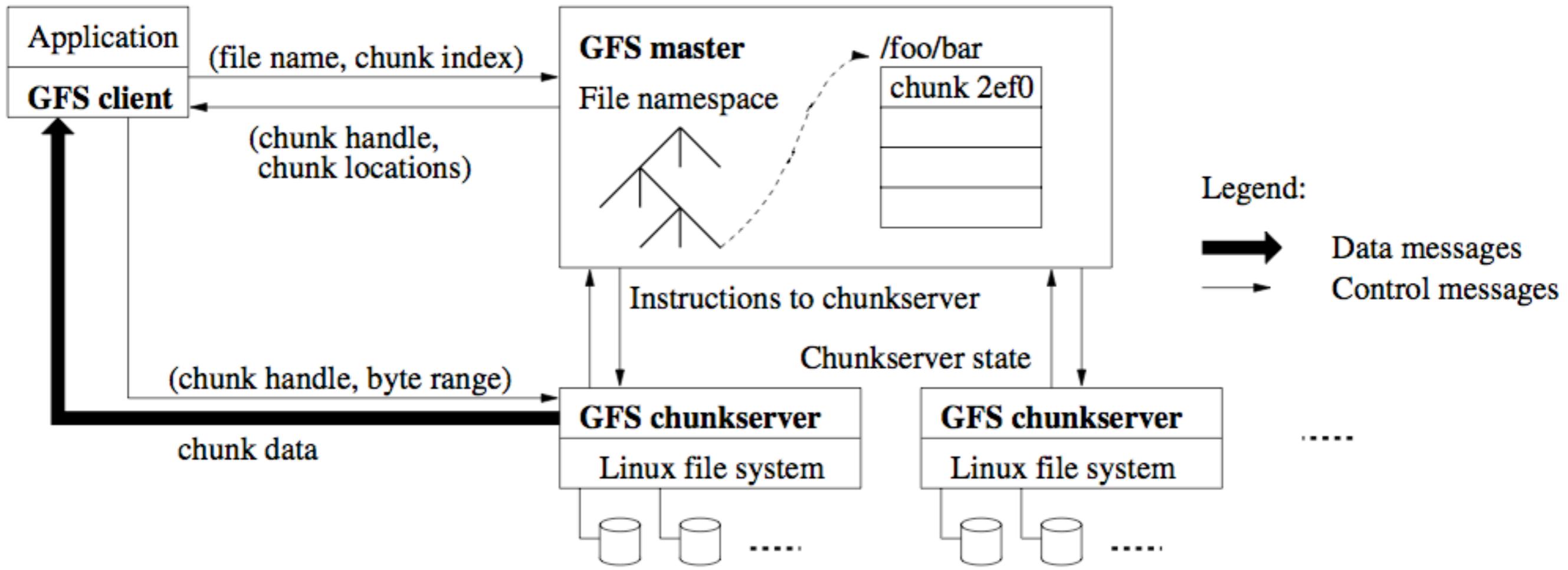
- Redundant array of inexpensive disks (optional fault tolerance)
 - Aggregate storage of many disks
 - Aggregate bandwidth of many disks
- RAID 0 - stripe data over disks (**good bandwidth, faulty**)
- RAID 1 - mirror disks (mediocre bandwidth, **fault tolerance**)
- RAID 5 - stripe data with 1 disk for parity (**good bandwidth, fault tolerance**)
- Even better - use error correcting code for fault tolerance,
e.g. (4,2) code, i.e. two disks out of 6 may fail



Distributed replicated file systems

- Internet workload
 - Bulk sequential writes
 - Bulk sequential reads
 - **No random writes (possibly random reads)**
 - High bandwidth requirements per file
 - High availability / replication
- Non starters
 - Lustre (high bandwidth, but no replication outside racks)
 - Gluster (POSIX, more classical mirroring, see Lustre)
 - NFS/AFS/whatever - doesn't actually parallelize

Google File System / HadoopFS

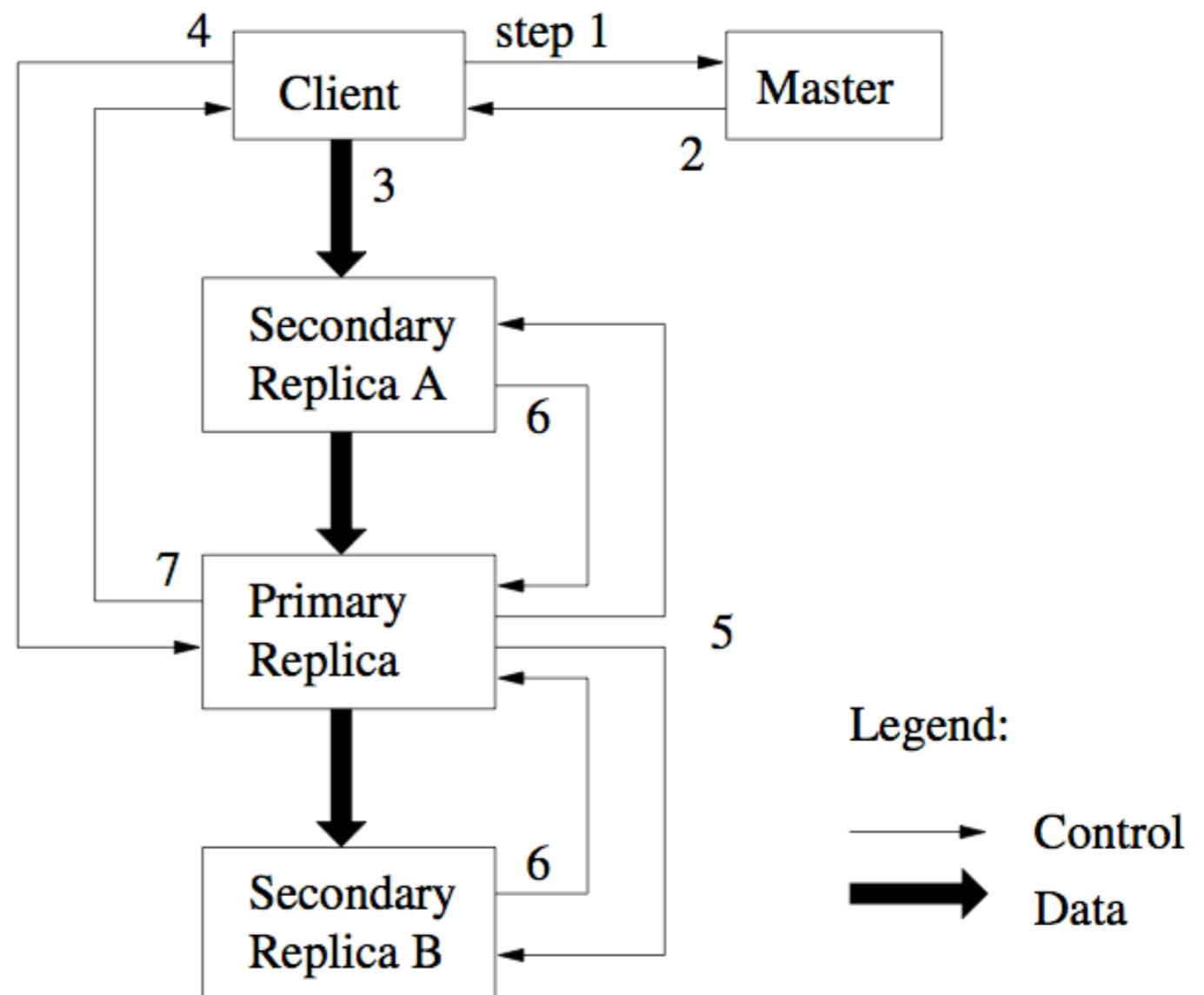


Ghemawat, Gobioff, Leung, 2003

- Chunk servers hold blocks of the file (64MB per chunk)
- Replicate chunks (chunk servers do this autonomously). **More bandwidth and fault tolerance**
- **Master distributes, checks faults, rebalances (Achilles heel)**
- Client can do bulk read / write / random reads

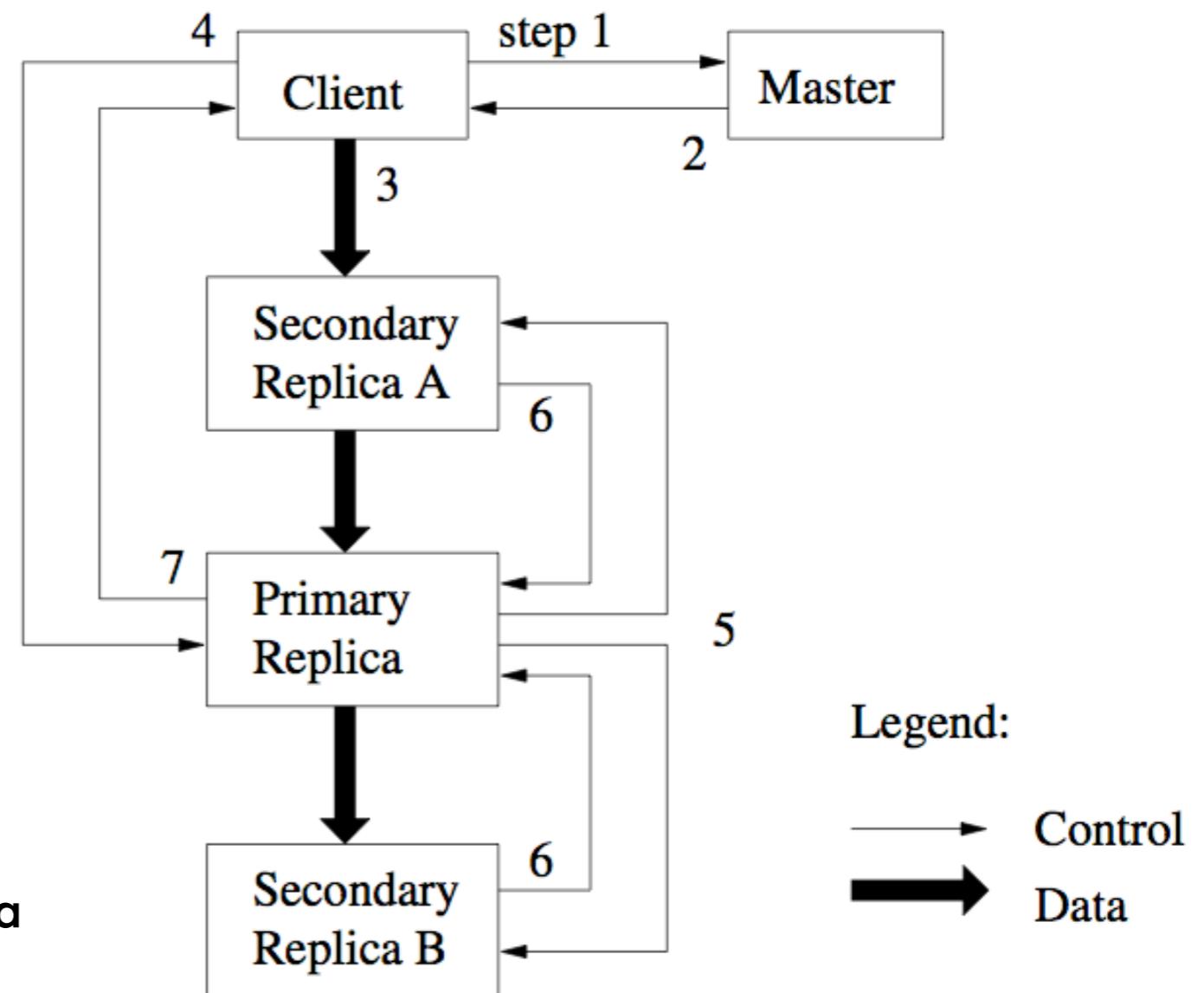
Google File System / HDFS

- Client requests chunk from master
- Master responds with replica location
- Client writes to replica A
- Client notifies primary replica
- Primary replica requests data from replica A
- Replica A sends data to Primary replica
- Primary replica confirms write to client



Google File System / HDFS

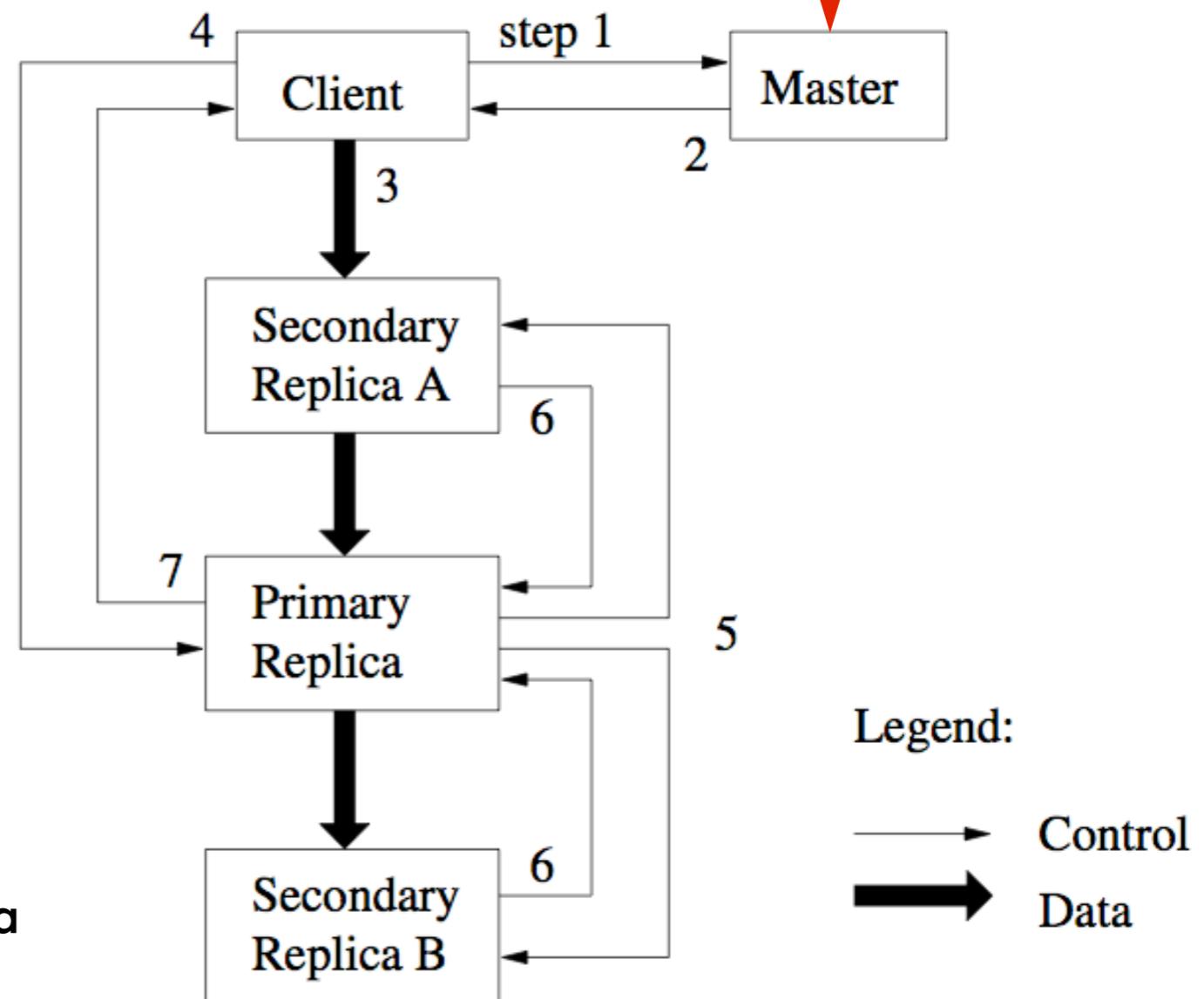
- Client requests chunk from master
- Master responds with replica location
- Client writes to replica A
- Client notifies primary replica
- Primary replica requests data from replica A
- Replica A sends data to Primary replica
- Primary replica confirms write to client
- Master ensures nodes are live
 - Chunks are checksummed
 - Can control replication factor for hotspots / load balancing
- Deserialize master state by loading data structure as flat file from disk (fast)



Google File System / HDFS

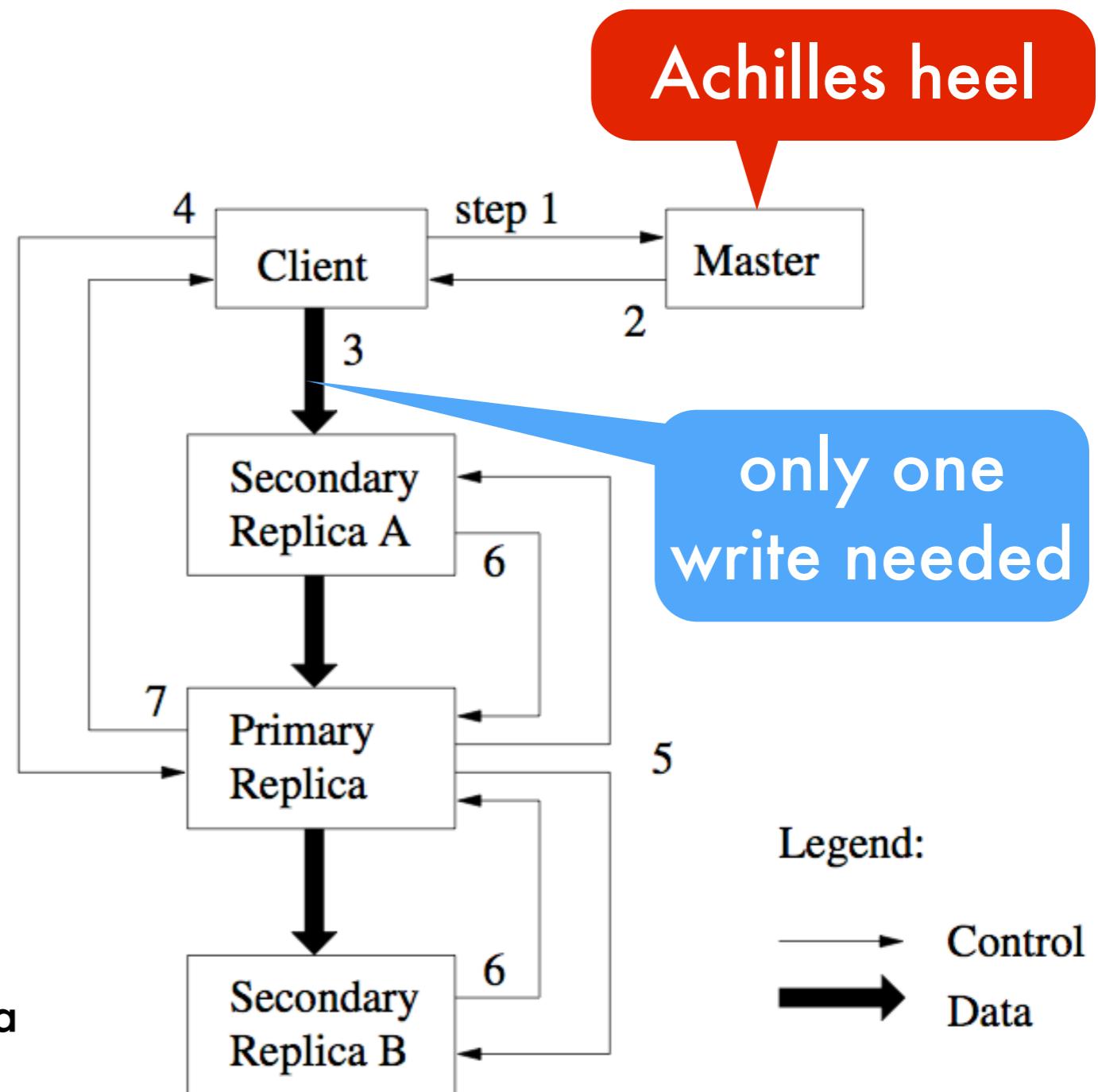
- Client requests chunk from master
- Master responds with replica location
- Client writes to replica A
- Client notifies primary replica
- Primary replica requests data from replica A
- Replica A sends data to Primary replica
- Primary replica confirms write to client
- Master ensures nodes are live
 - Chunks are checksummed
 - Can control replication factor for hotspots / load balancing
- Deserialize master state by loading data structure as flat file from disk (fast)

Achilles heel



Google File System / HDFS

- Client requests chunk from master
- Master responds with replica location
- Client writes to replica A
- Client notifies primary replica
- Primary replica requests data from replica A
- Replica A sends data to Primary replica
- Primary replica confirms write to client
- Master ensures nodes are live
 - Chunks are checksummed
 - Can control replication factor for hotspots / load balancing
- Deserialize master state by loading data structure as flat file from disk (fast)



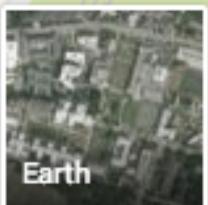
Carnegie Mellon University, Pittsburgh, PA

Carnegie Mellon University

(412) 268-2000

4.7 ★★★★☆ 117 reviews

Map Reduce & Processing



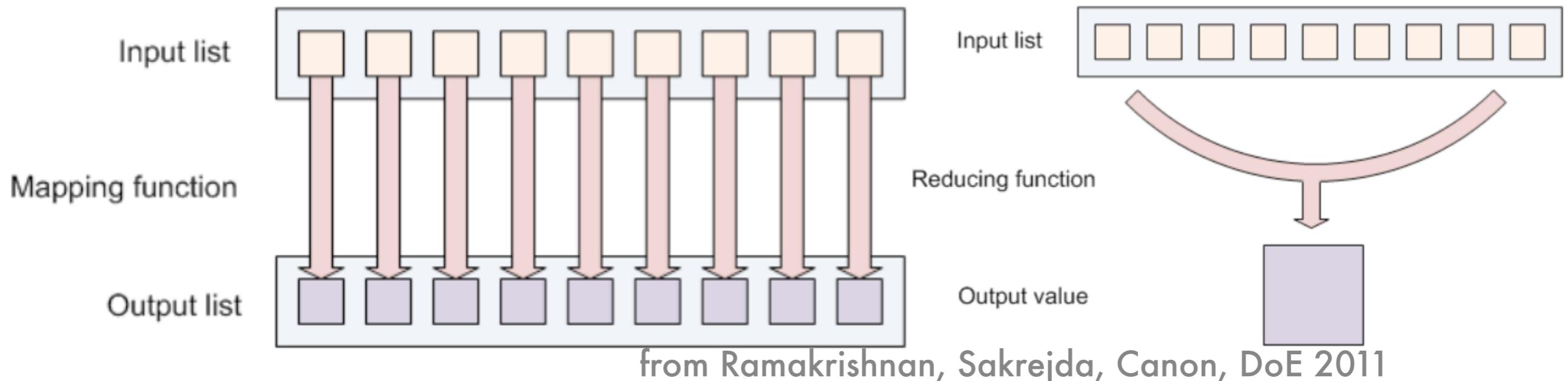
Flagstaff Hill

Google

Map data ©2014 Google

Map Reduce

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



Map Reduce

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

Map Reduce

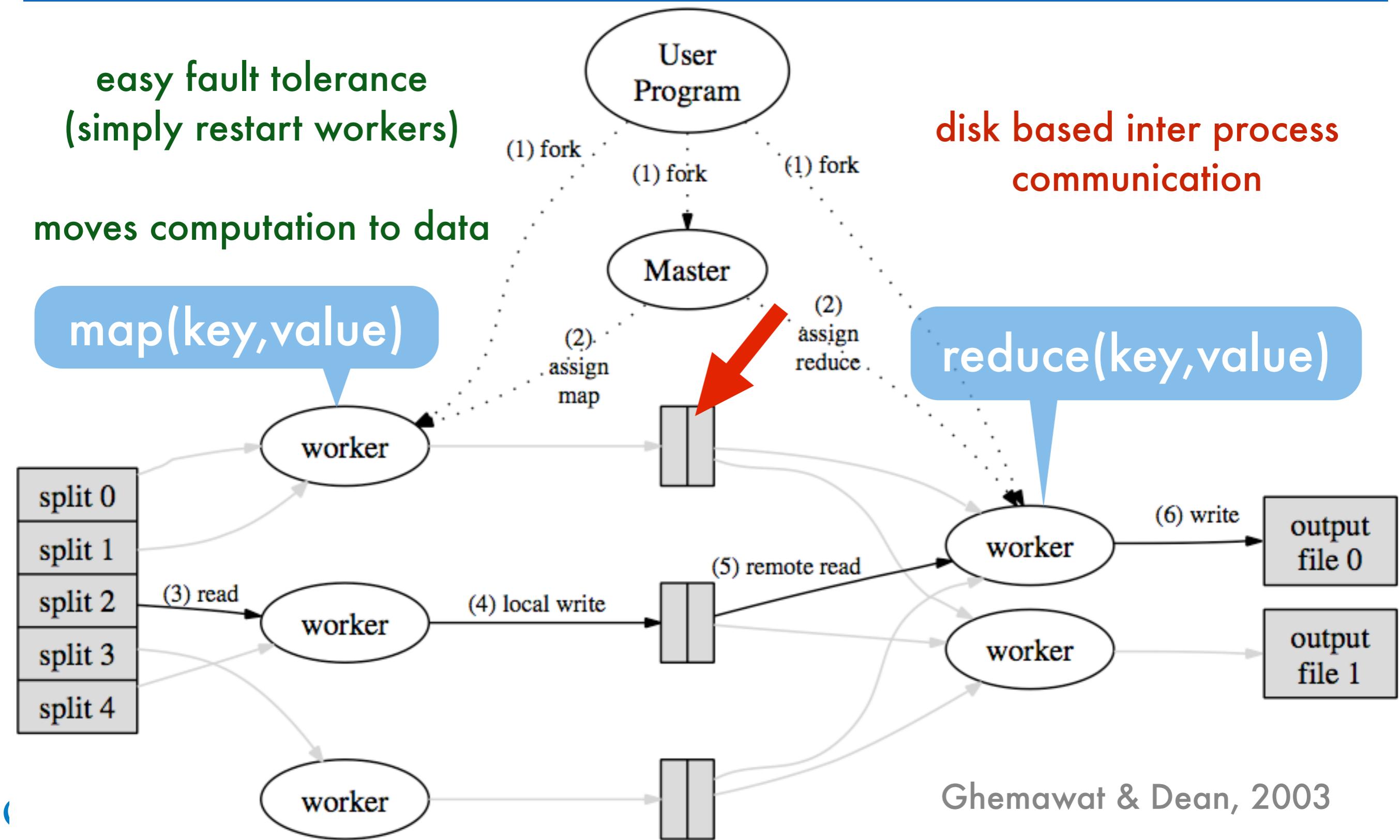
easy fault tolerance
(simply restart workers)

moves computation to data

map(key,value)

disk based inter process
communication

reduce(key,value)



Map Combine Reduce

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

Example - Gradient Descent

- Objective

$$\underset{w}{\text{minimize}} \sum_{i=1}^m l(x_i, y_i, w) + \frac{\lambda}{2} \|w\|^2$$

- Algorithm

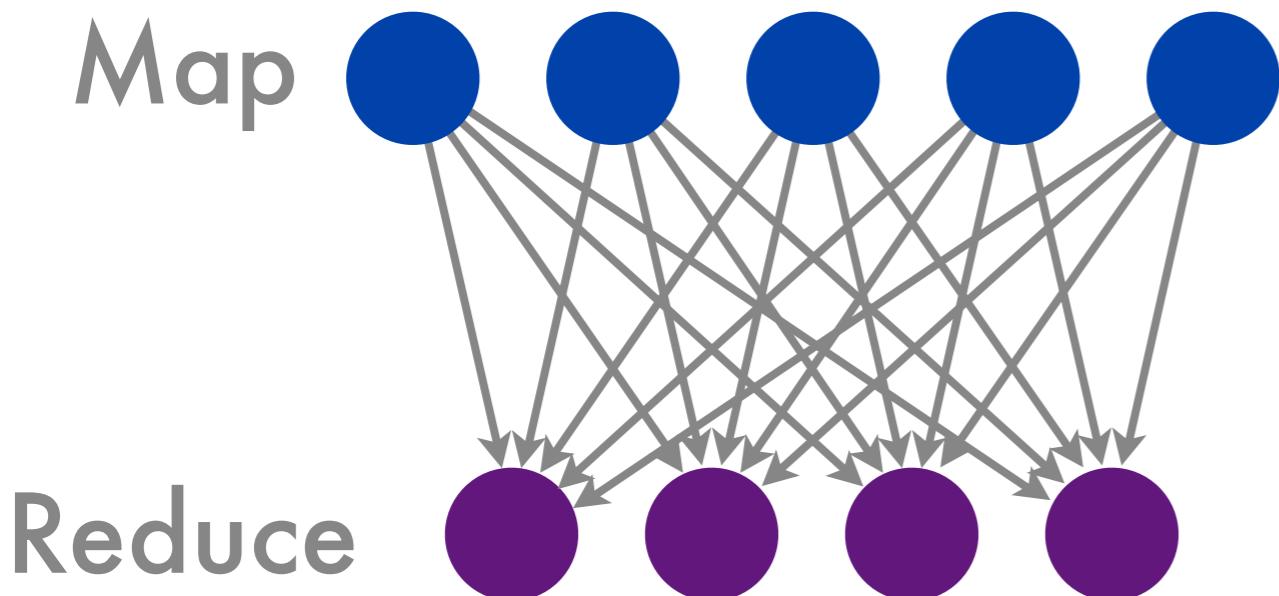
- compute gradient

$$g := \sum_{i=1}^m \partial_w l(x_i, y_i, w)$$

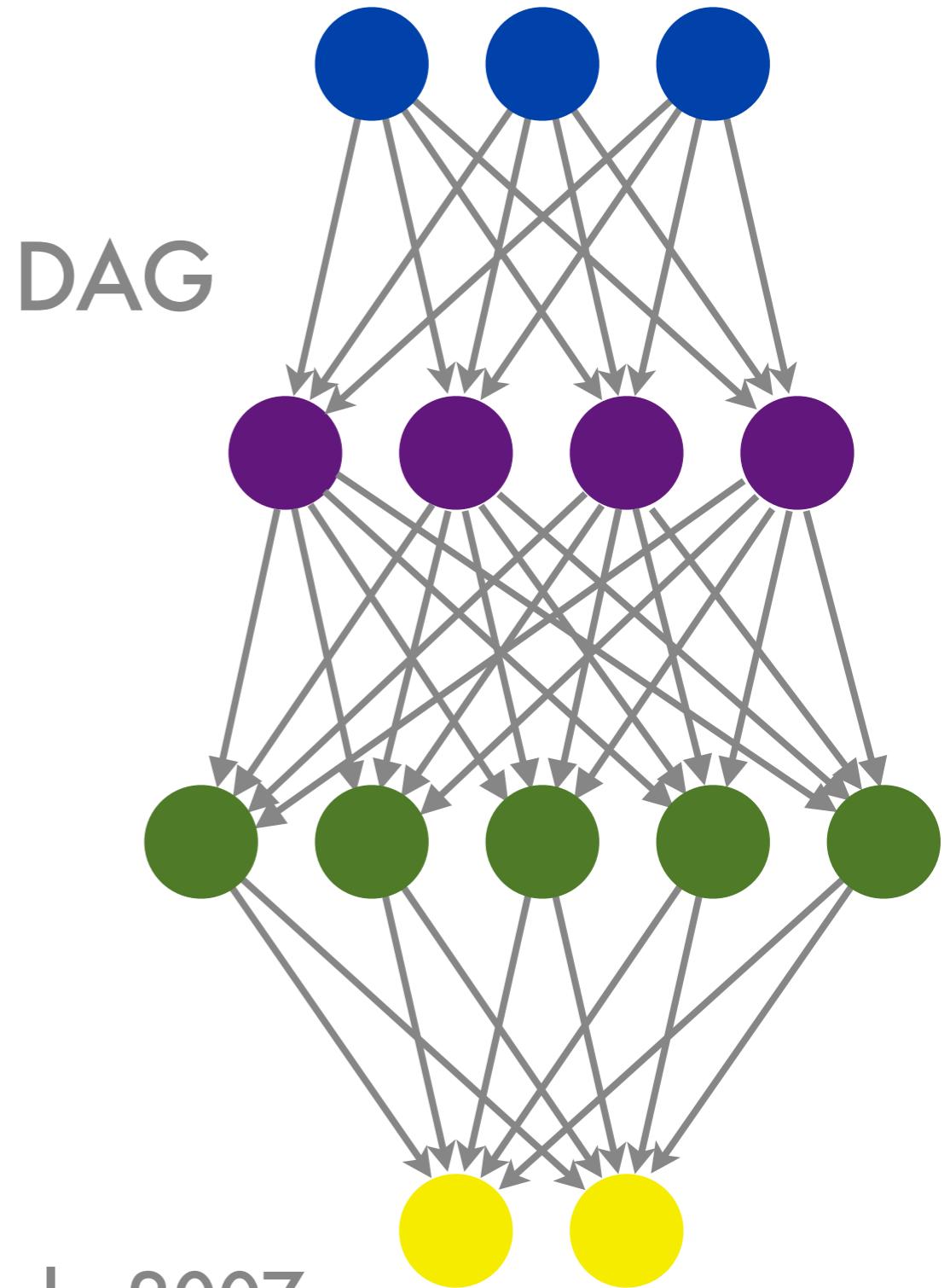
- On each data point via `Map(i,data)`
 - Sum gradient via `Reduce(coordinate)`
 - perform update step
(much better with line search)
 - repeat

$$w \leftarrow w - \eta(g + \lambda w)$$

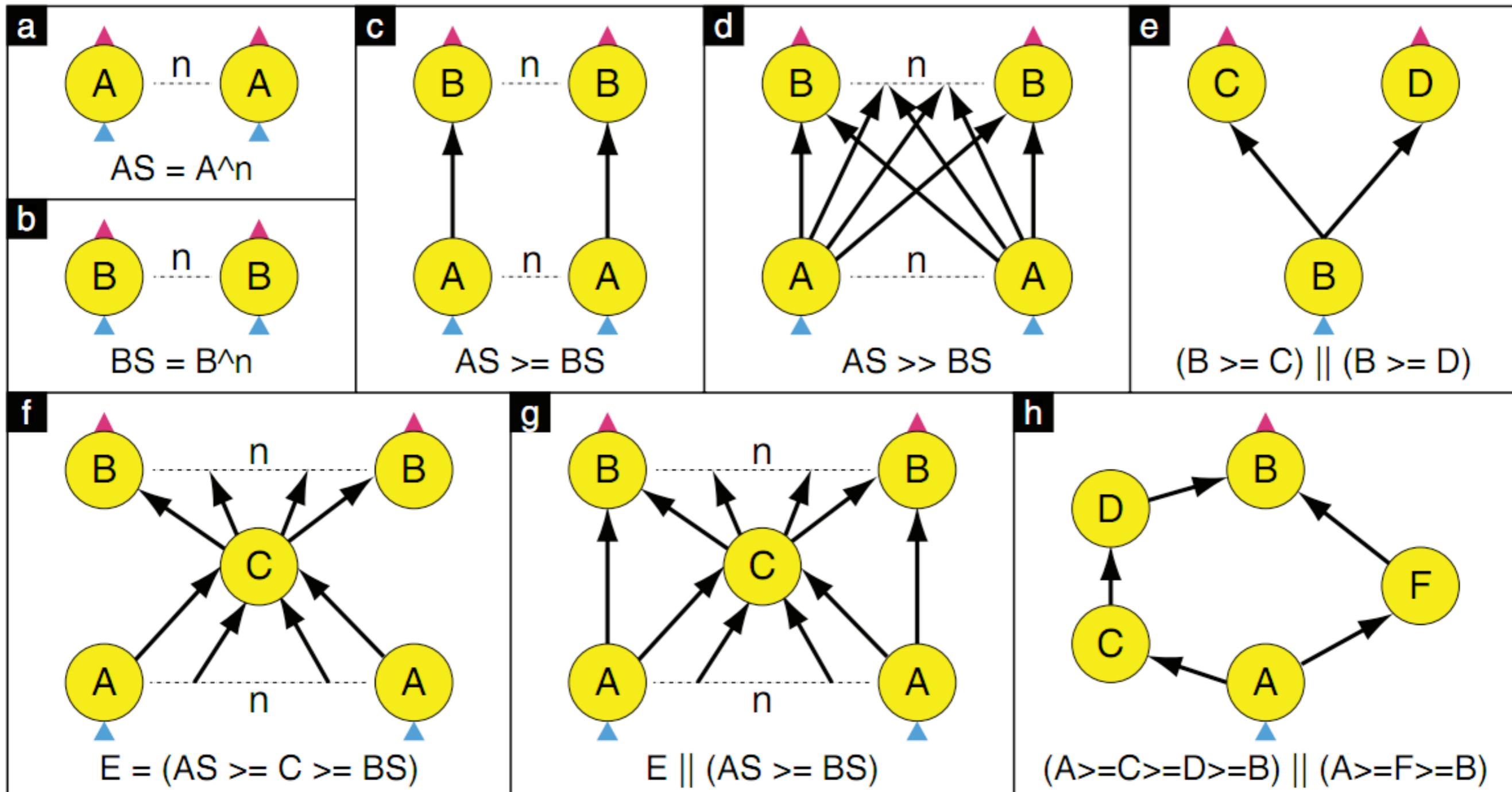
Dryad



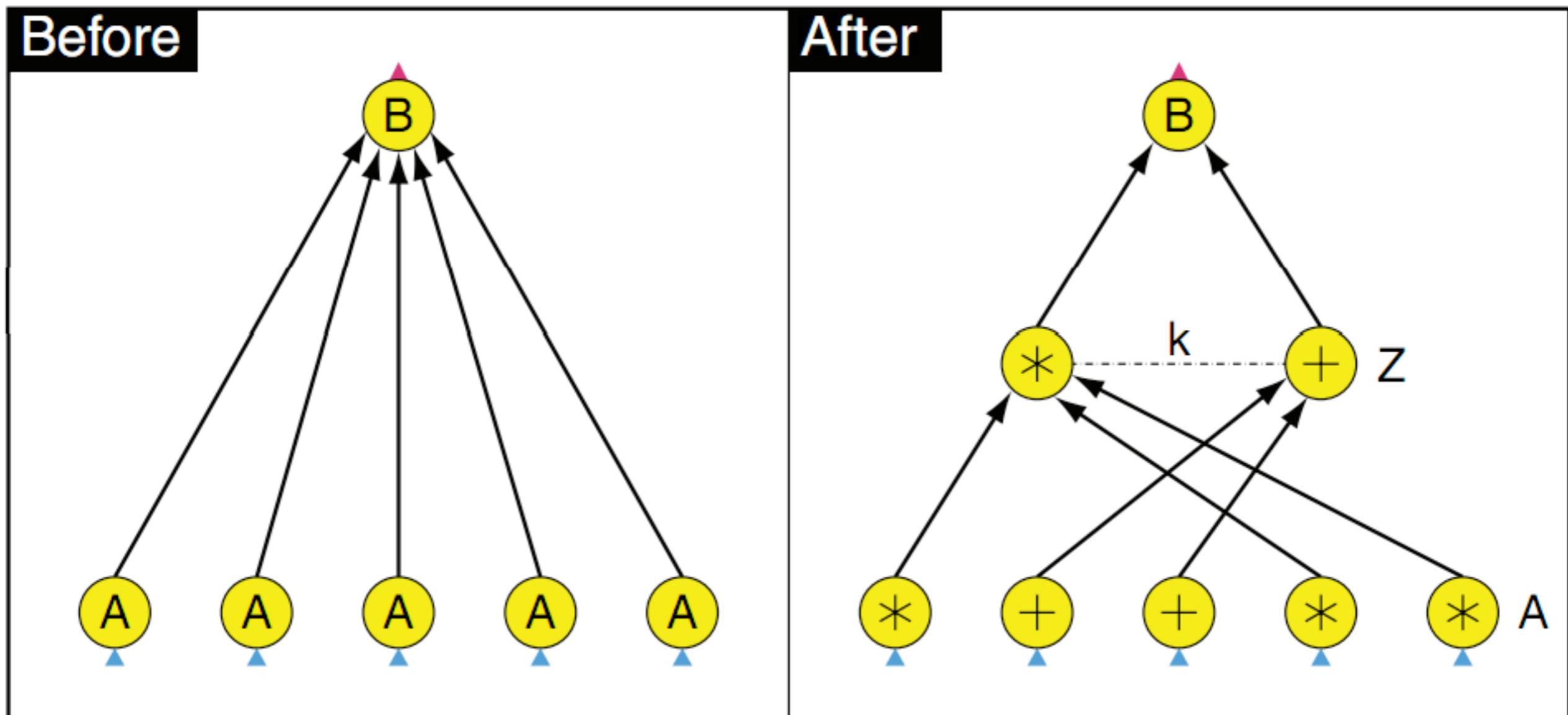
- Directed acyclic graph
- System optimizes parallelism
- Different types of IPC
(memory FIFO/network/file)
- Tight integration with .NET
(allows easy prototyping)



DRYAD



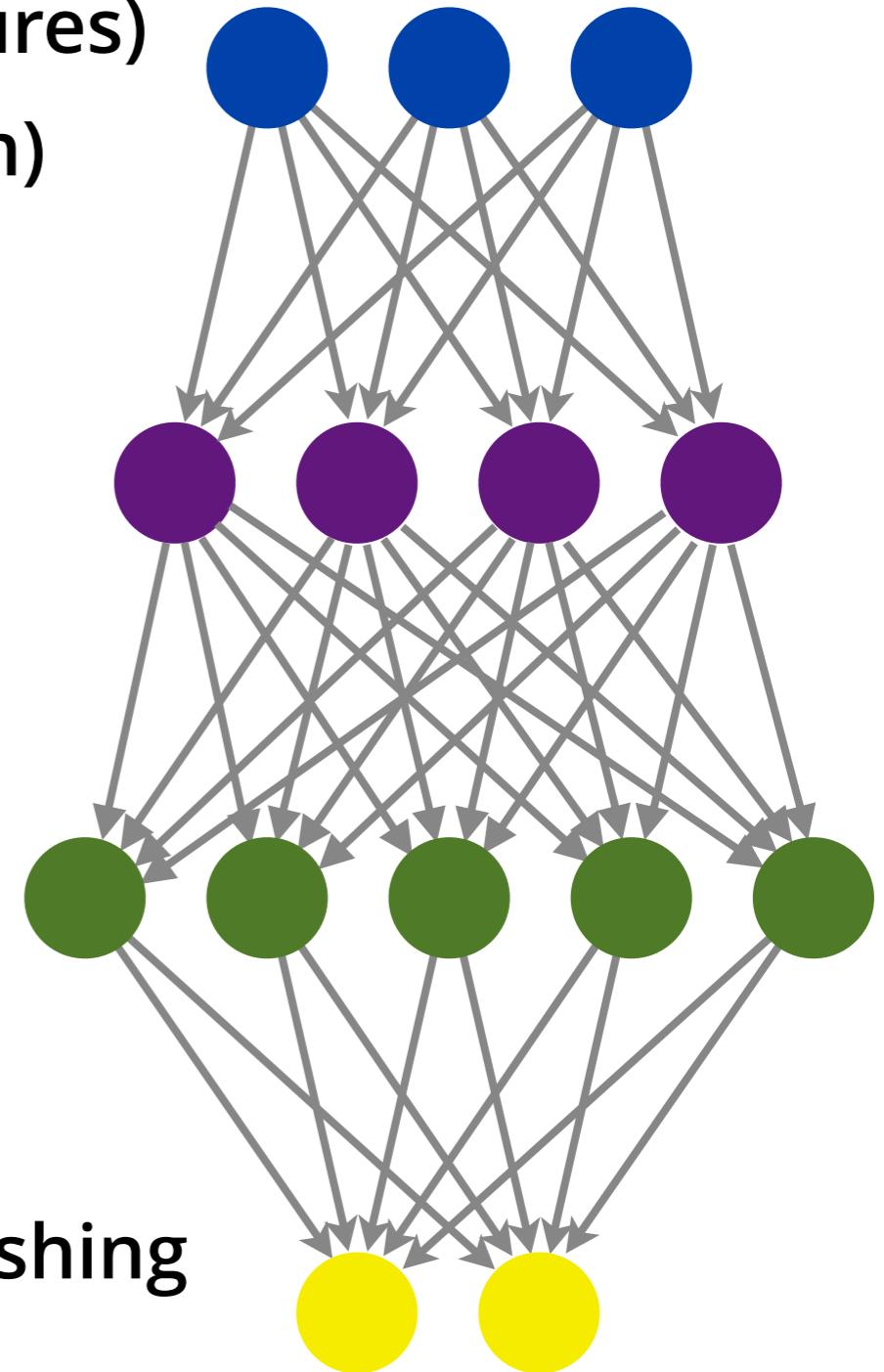
DRYAD



automatic graph refinement

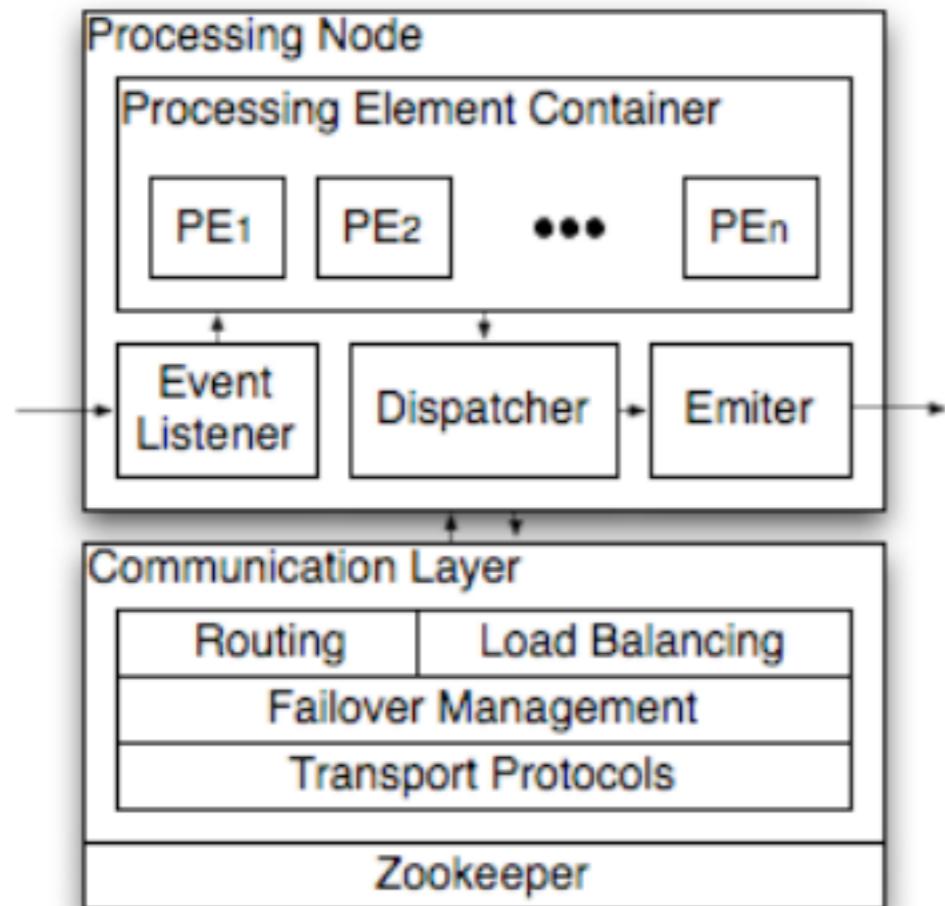
S4

- Directed acyclic graph (Dryad-like features)
- Real-time processing of data (as stream)
- Scalability (decentralized & symmetric)
- Fault tolerance
- Consistency for keys
- Processing elements
 - Ingest (key, value) pair
 - Capabilities tied to ID
 - Clonable (for scaling)
- Simple implementation - consistent hashing

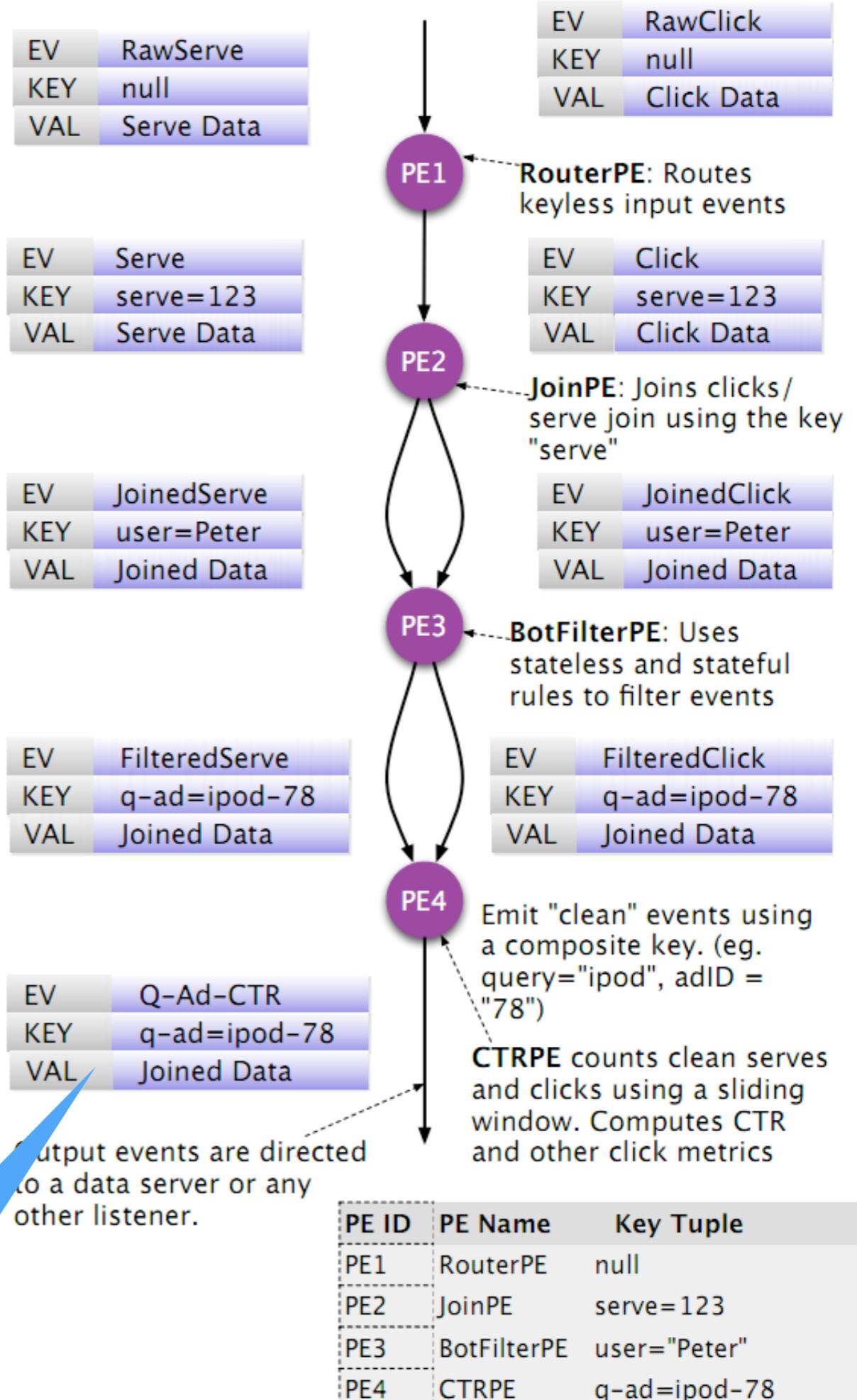


S4

processing element



click through rate estimation

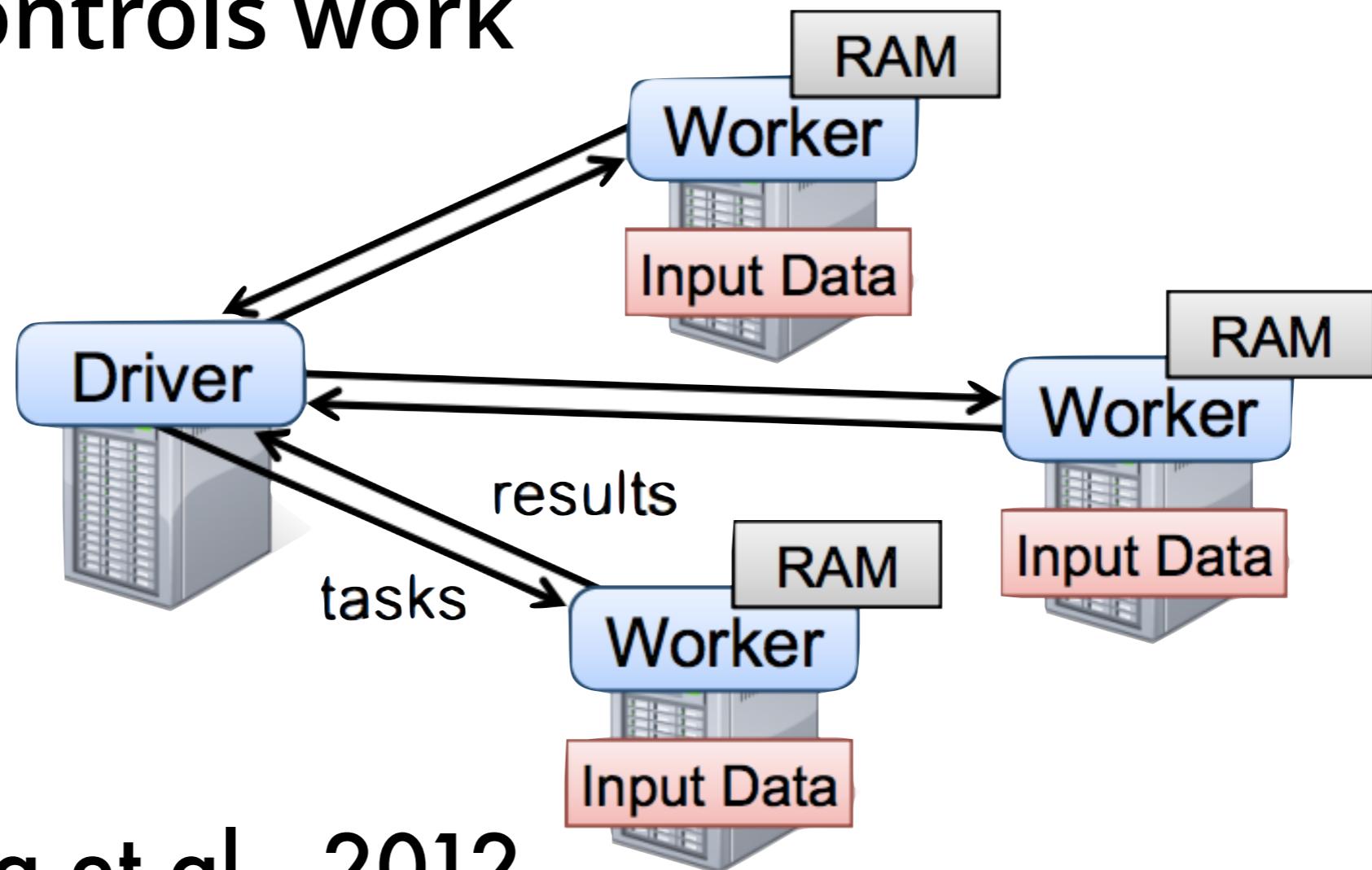


Spark



Resilient Distributed Datasets

- Data is transformed by processing
- Store intermediate data using lineage
- Driver controls work



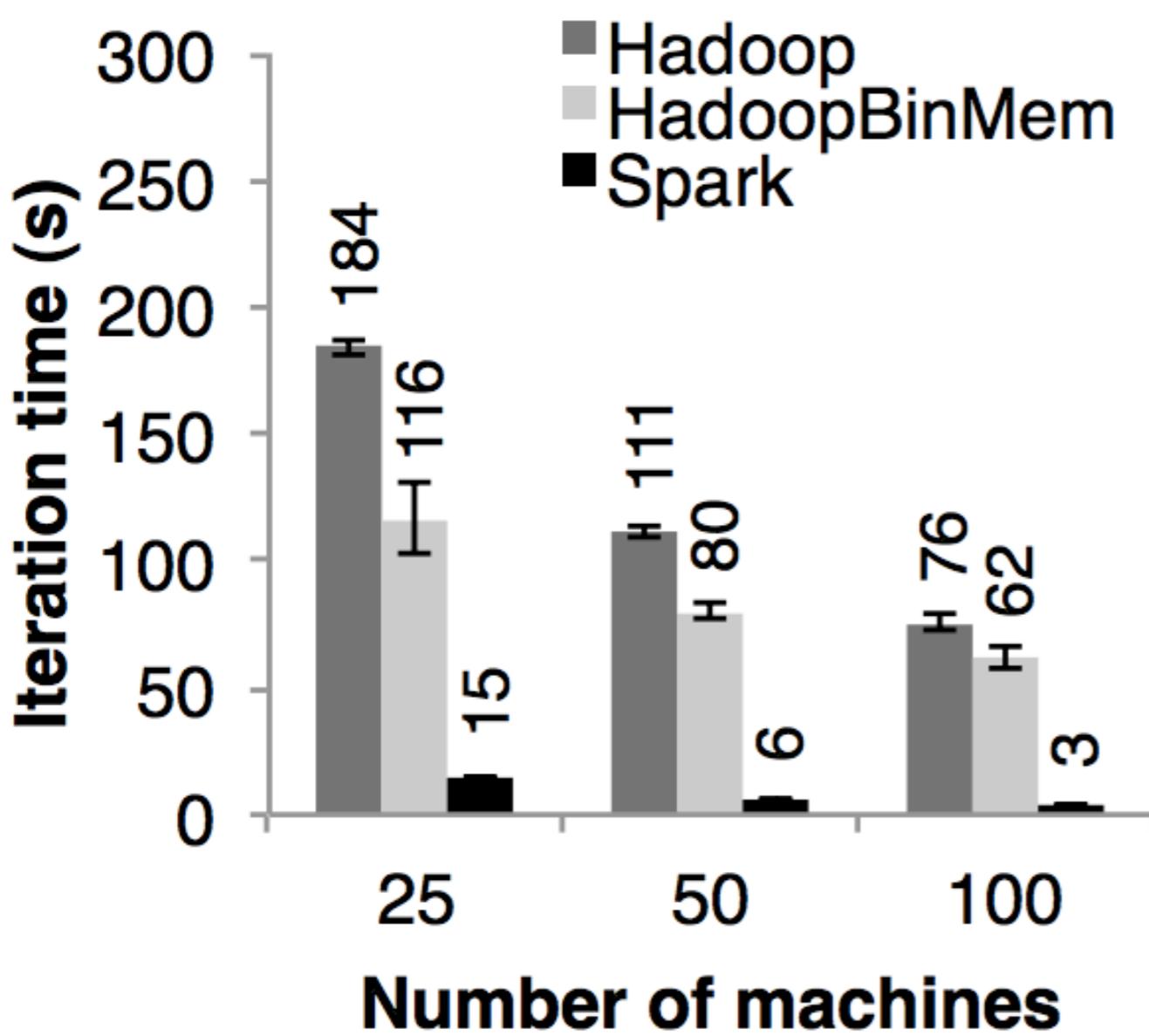
Zaharia et al., 2012

Beyond MapReduce

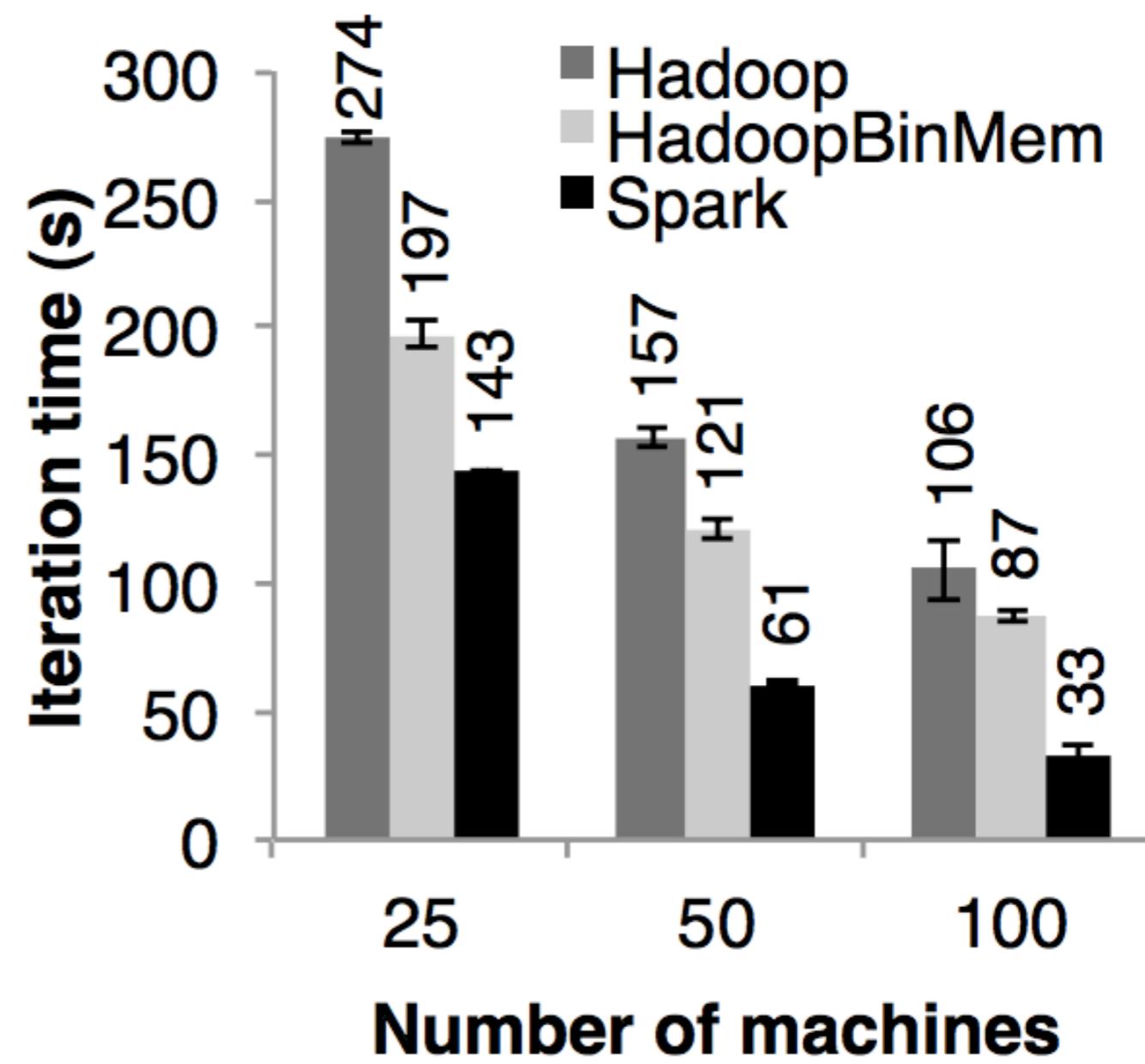
Transformations	$map(f : T \Rightarrow U)$: $RDD[T] \Rightarrow RDD[U]$ $filter(f : T \Rightarrow Bool)$: $RDD[T] \Rightarrow RDD[T]$ $flatMap(f : T \Rightarrow Seq[U])$: $RDD[T] \Rightarrow RDD[U]$ $sample(fraction : Float)$: $RDD[T] \Rightarrow RDD[T]$ (Deterministic sampling) $groupByKey()$: $RDD[(K, V)] \Rightarrow RDD[(K, Seq[V])]$ $reduceByKey(f : (V, V) \Rightarrow V)$: $RDD[(K, V)] \Rightarrow RDD[(K, V)]$ $union()$: $(RDD[T], RDD[T]) \Rightarrow RDD[T]$ $join()$: $(RDD[(K, V)], RDD[(K, W)]) \Rightarrow RDD[(K, (V, W))]$ $cogroup()$: $(RDD[(K, V)], RDD[(K, W)]) \Rightarrow RDD[(K, (Seq[V], Seq[W])))]$ $crossProduct()$: $(RDD[T], RDD[U]) \Rightarrow RDD[(T, U)]$ $mapValues(f : V \Rightarrow W)$: $RDD[(K, V)] \Rightarrow RDD[(K, W)]$ (Preserves partitioning) $sort(c : Comparator[K])$: $RDD[(K, V)] \Rightarrow RDD[(K, V)]$ $partitionBy(p : Partitioner[K])$: $RDD[(K, V)] \Rightarrow RDD[(K, V)]$
Actions	$count()$: $RDD[T] \Rightarrow Long$ $collect()$: $RDD[T] \Rightarrow Seq[T]$ $reduce(f : (T, T) \Rightarrow T)$: $RDD[T] \Rightarrow T$ $lookup(k : K)$: $RDD[(K, V)] \Rightarrow Seq[V]$ (On hash/range partitioned RDDs) $save(path : String)$: Outputs RDD to a storage system, e.g., HDFS

rich language & preprocessor

Improvement over MapReduce



(a) Logistic Regression



(b) K-Means



parameterserver.org



The background of the image is a photograph of a rolling green hillside under a bright blue sky with scattered white clouds. The grass is a vibrant green with some yellow flowers visible in the foreground.

Background

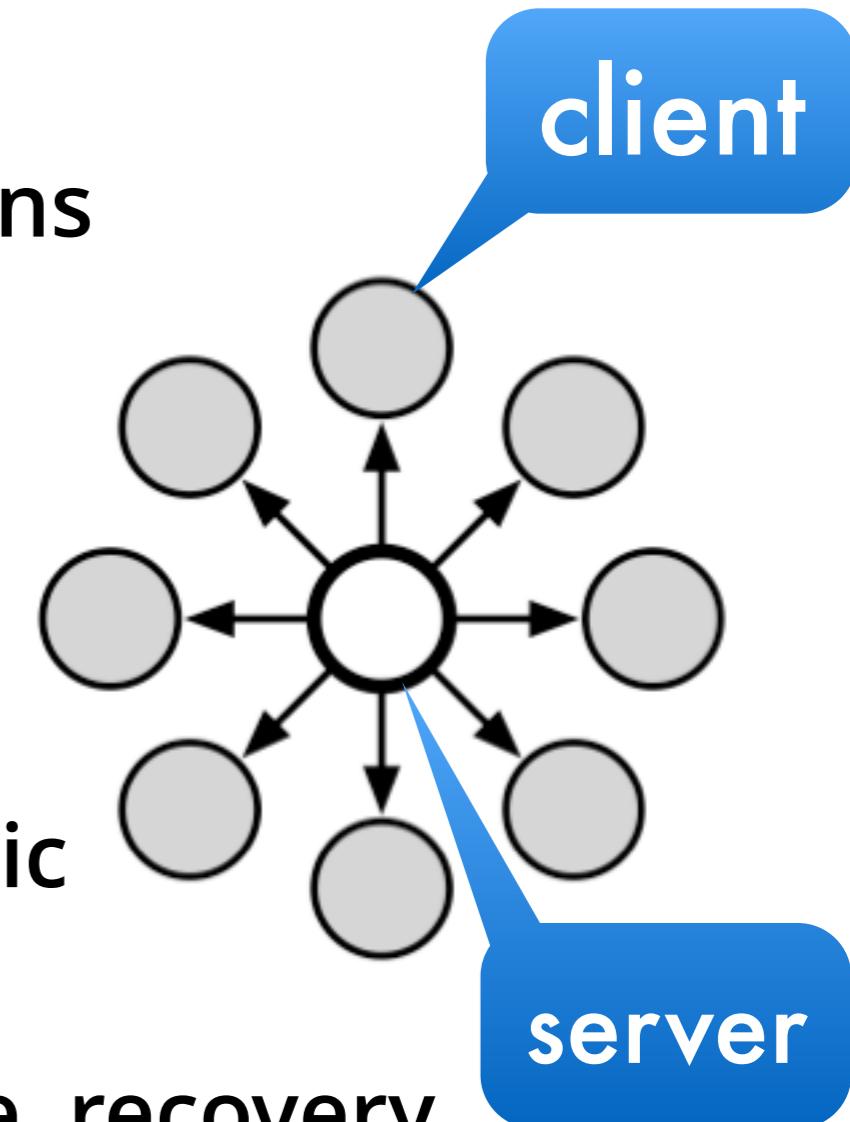
The Challenge

- Scale
 - Terabytes of data
 - 1000s of computers
 - Billions of parameters
- Reality
 - Faulty machines
 - Shared cluster
- Performance
 - Front end serving machines
 - Real time response



General parallel algorithm template

- Clients have local view of parameters
- P2P is infeasible since $O(n^2)$ connections
- Synchronize* with parameter server
 - Reconciliation protocol
average parameters, lock variables
 - Synchronization schedule
asynchronous, synchronous, episodic
 - Load distribution algorithm
uniform distribution, fault tolerance, recovery

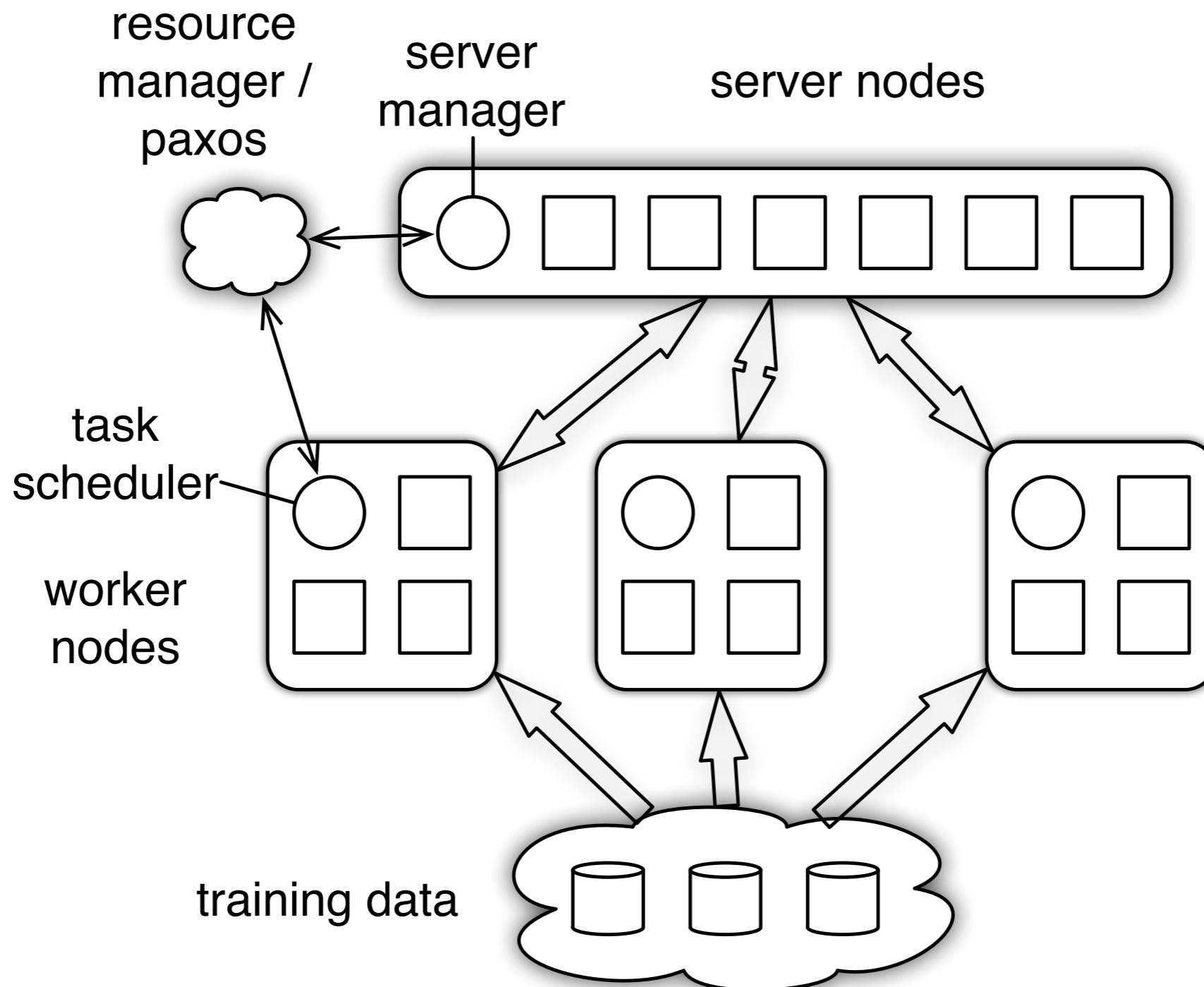


Smola & Narayananurthy, 2010, VLDB

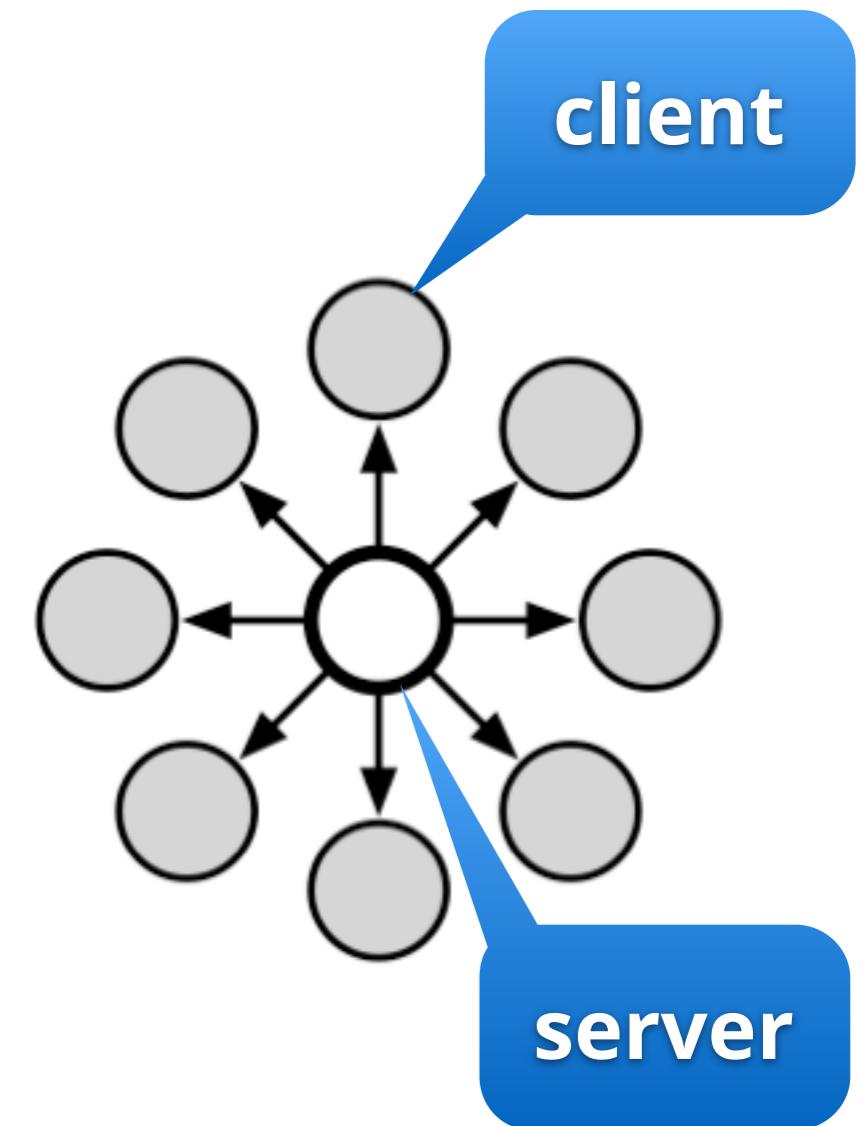
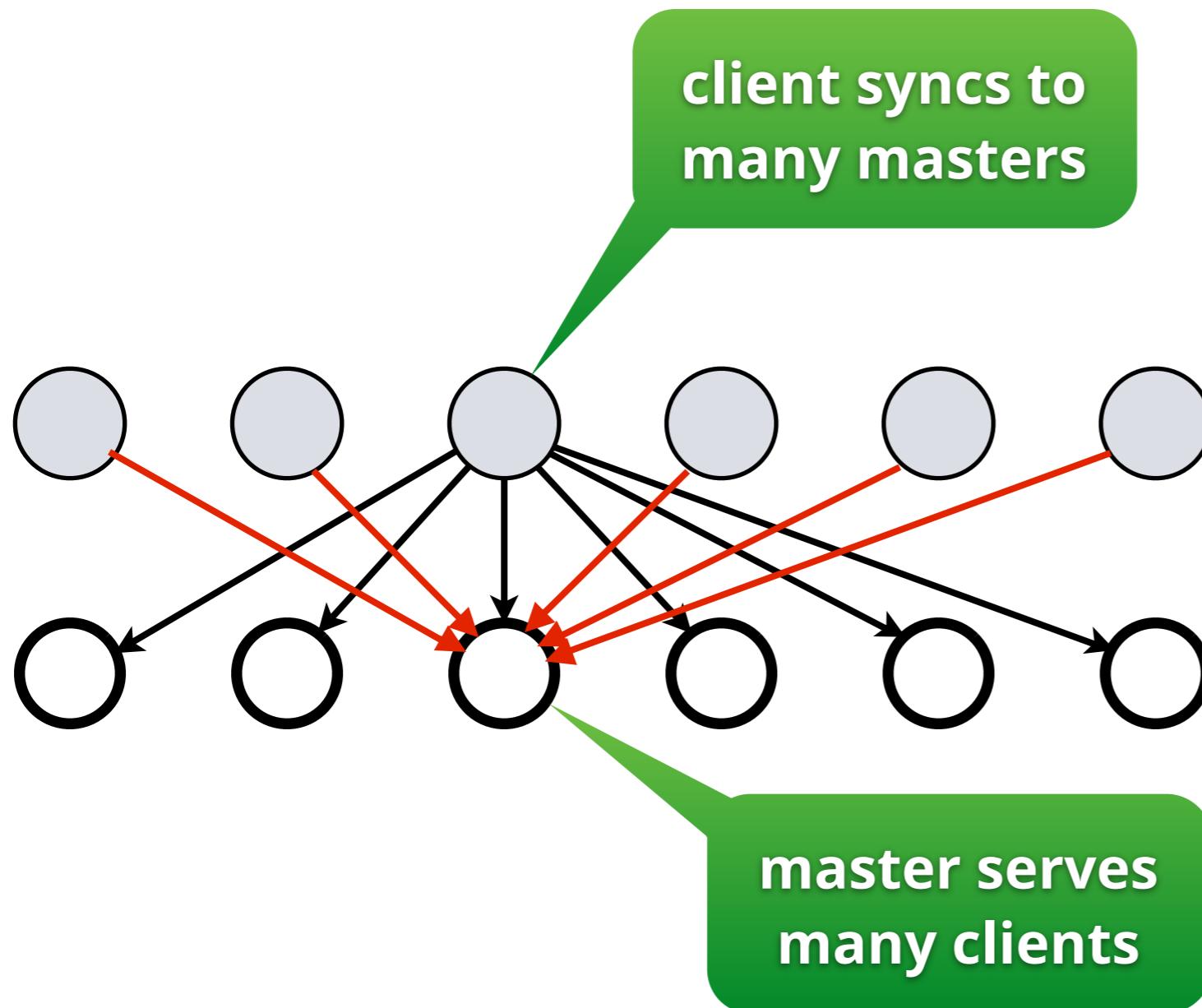
Gonzalez et al., 2012, WSDM

Shervashidze et al., 2013, WWW

Architecture



Communication pattern

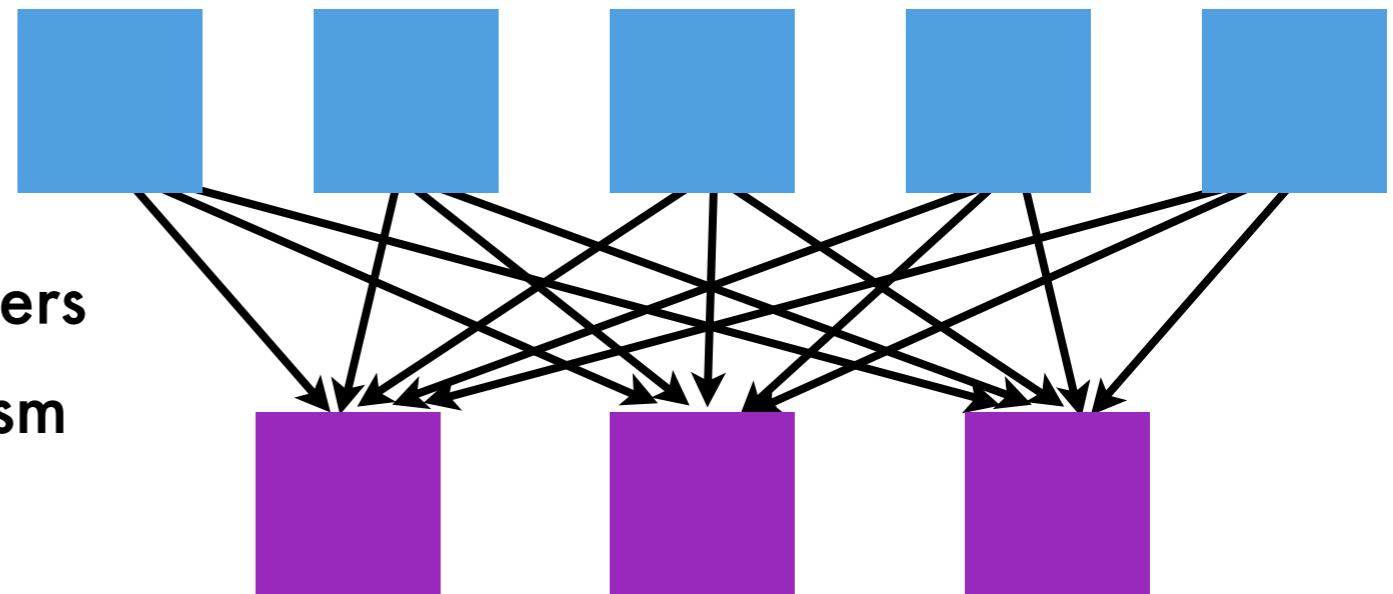




Key layout & recovery

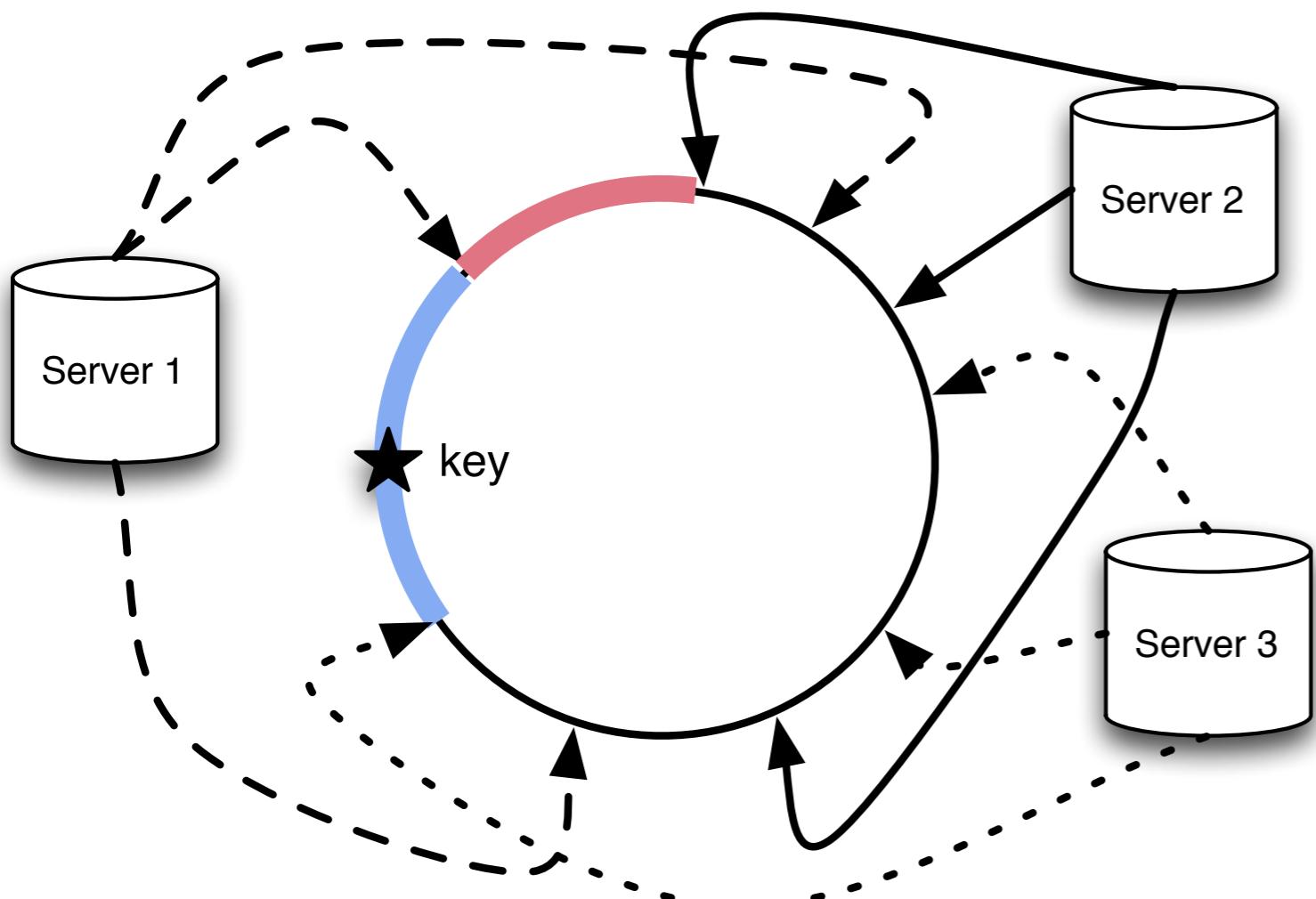
Consistent Hashing

- 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
 - very expensive to iterate over all keys for a given server



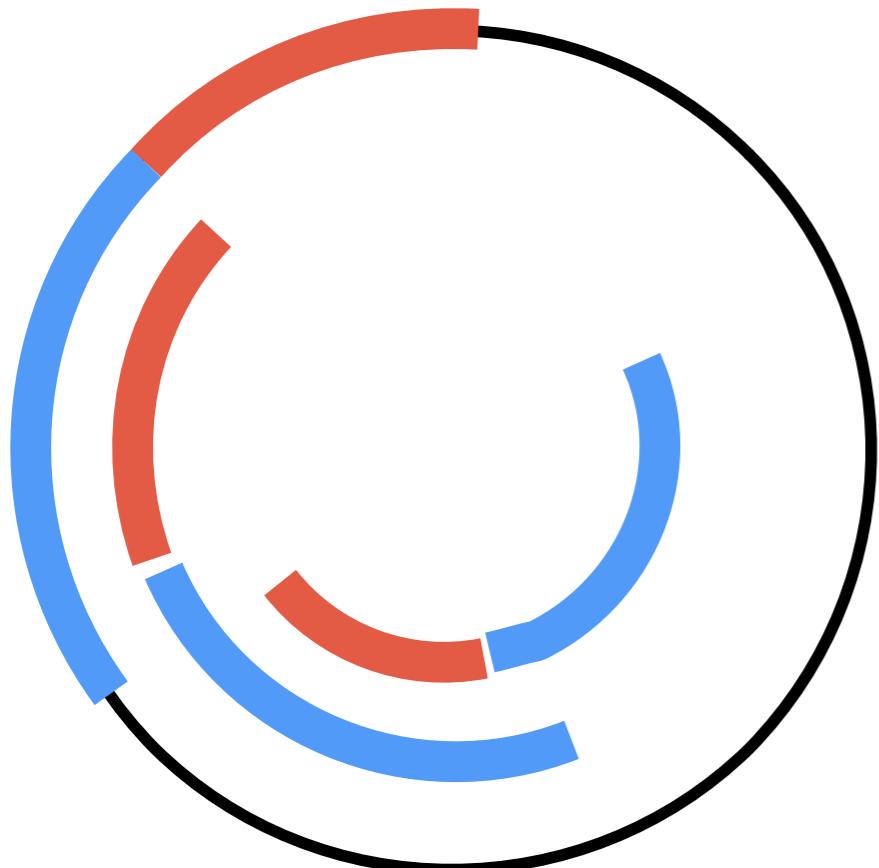
$$m(\text{key}, \mathcal{M}) = \operatorname{argmin}_{m' \in \mathcal{M}} h(\text{key}, m')$$

Keys arranged in a DHT



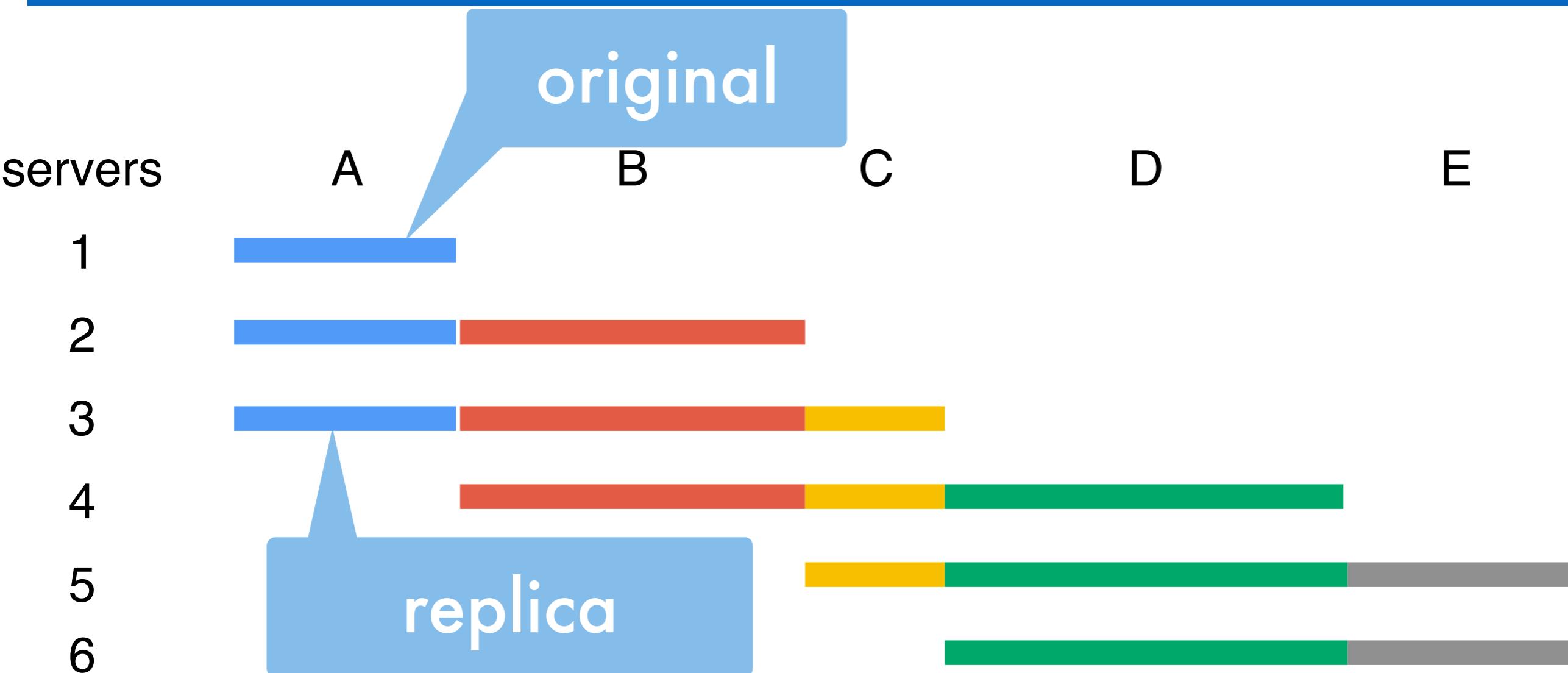
- Virtual servers
 - loadbalancing
 - multithreading
- DHT
 - contiguous key range for clients
 - easy bulk sync
 - easy insertion of servers

Key Replication

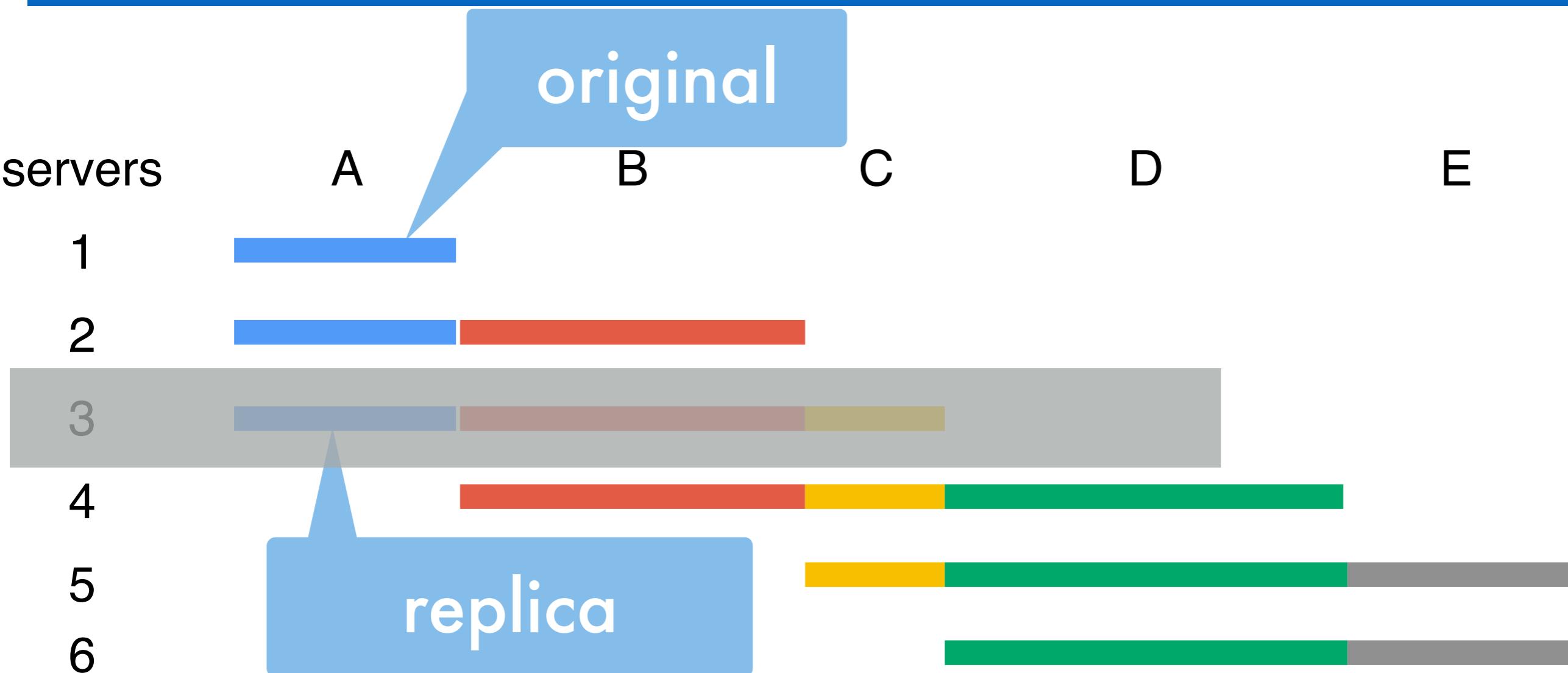


- Each segment is owned by one virtual server
- Subsequent machines hold replicas
- Easy fallback
- Easy insertion / repair
- Dynamic load balancing (important e.g. on EC2)

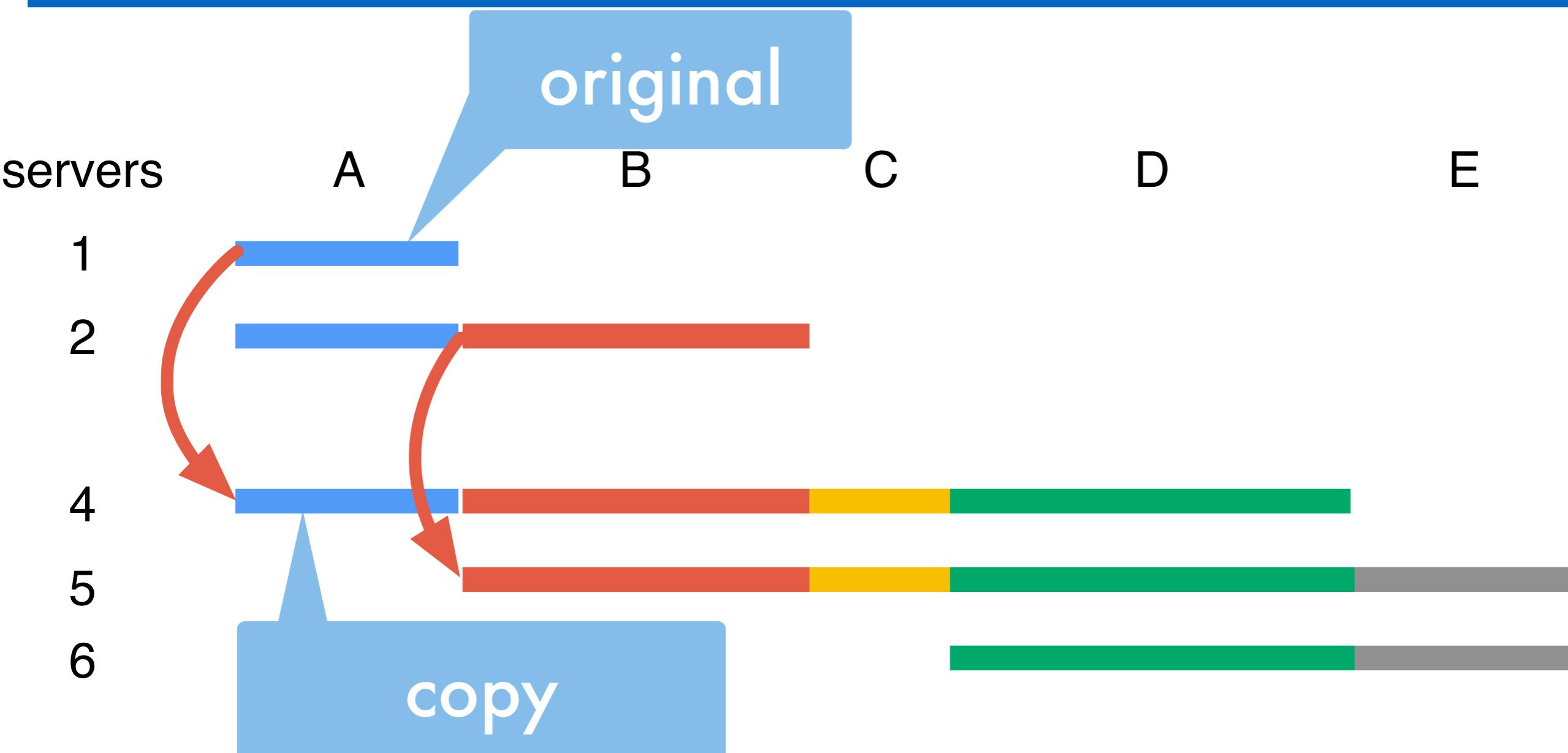
Key layout



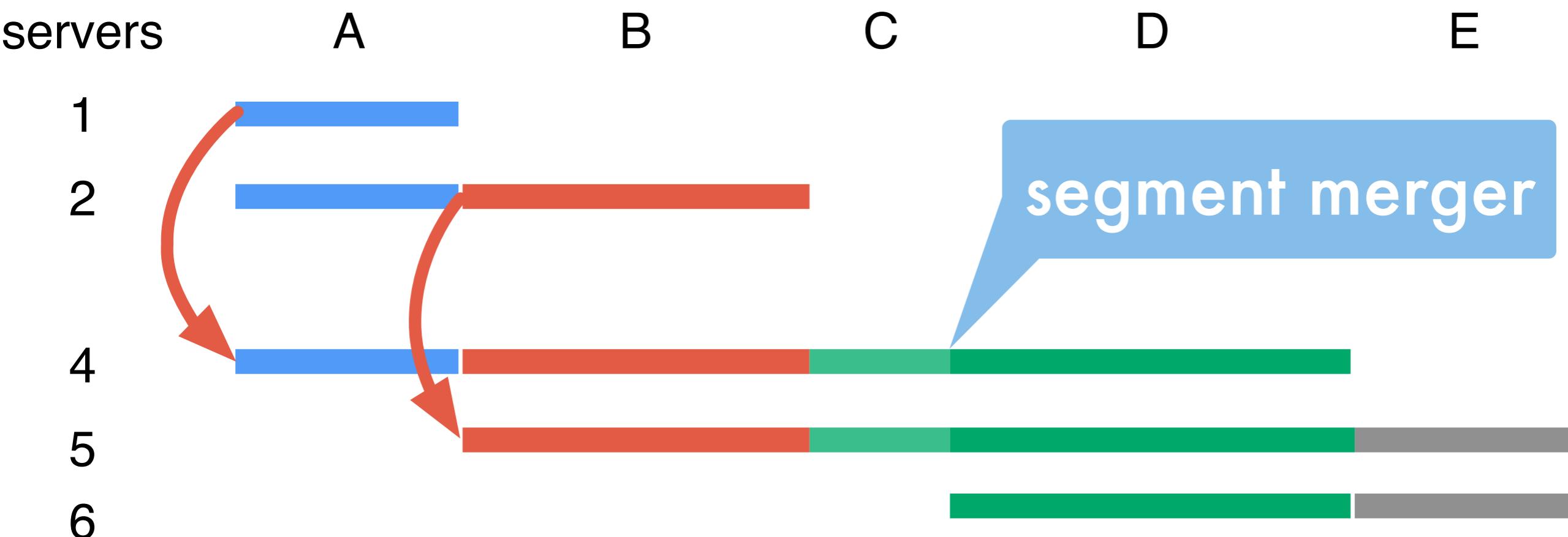
Key layout



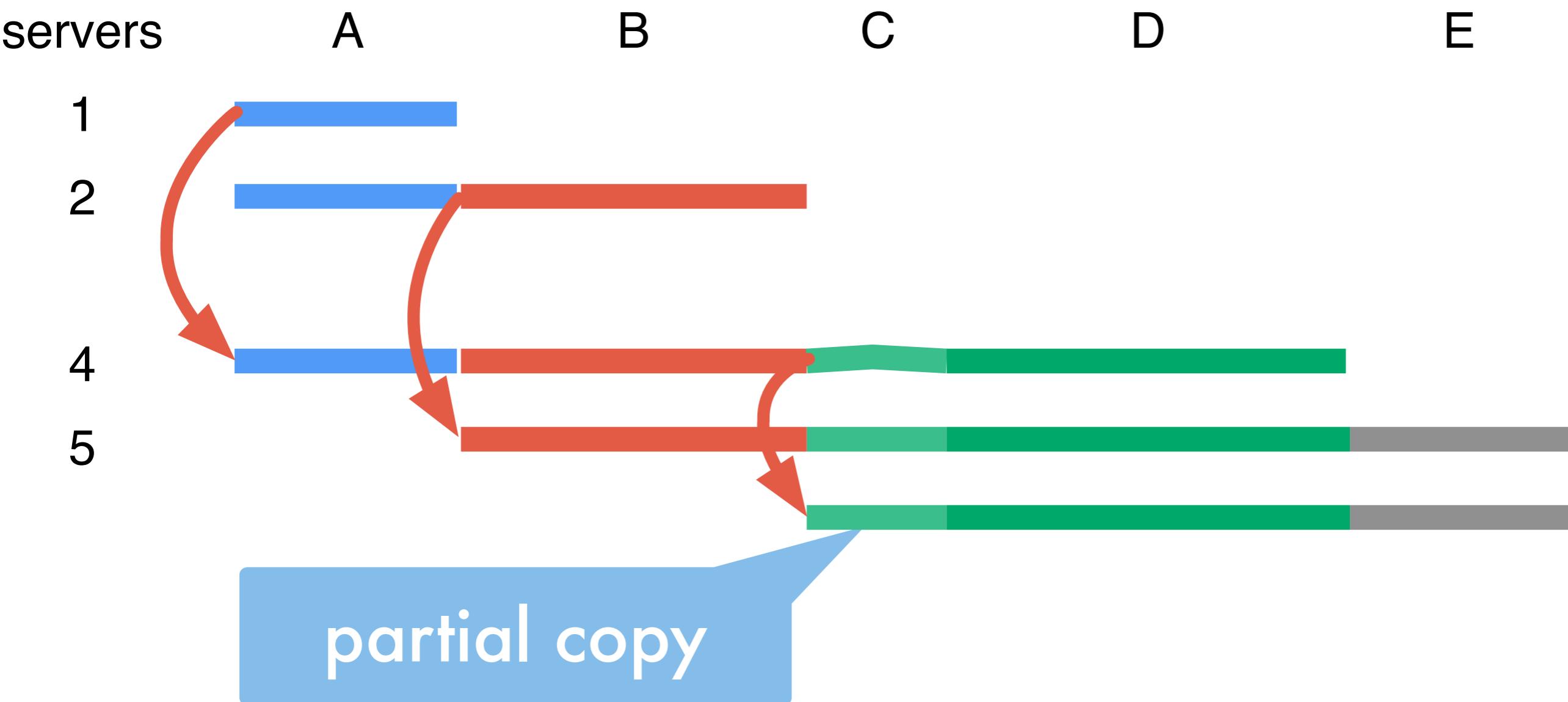
Key layout



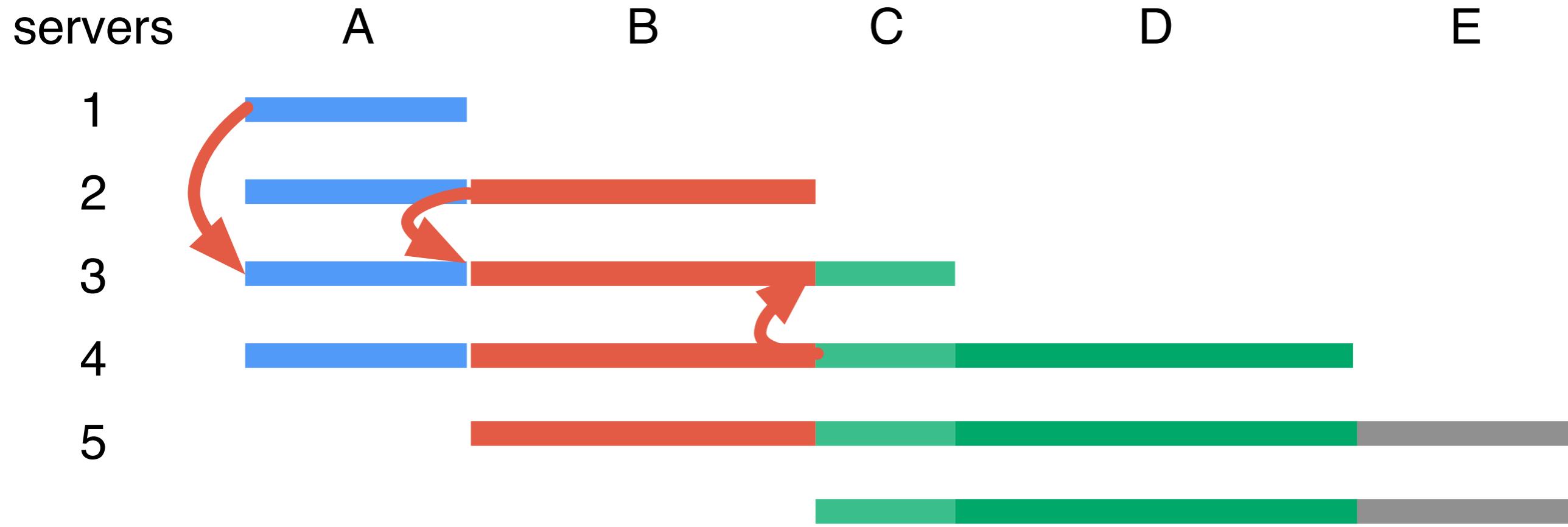
Key layout



Key layout



Recovery / server insertion



- Precopy server content to new candidate (3)
- After precopy ended, send log
- For k virtual servers this causes $O(k^2)$ delay

Communication



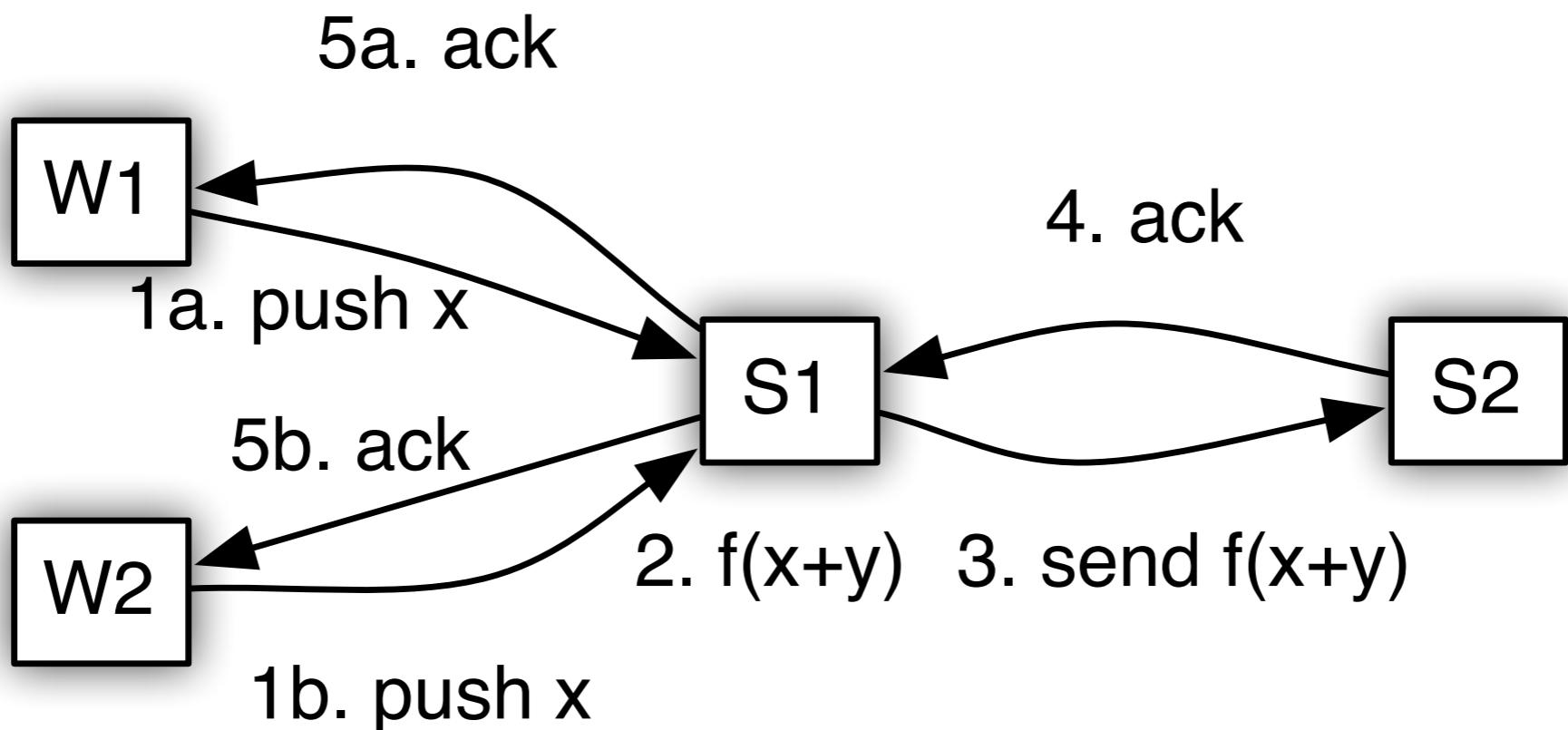
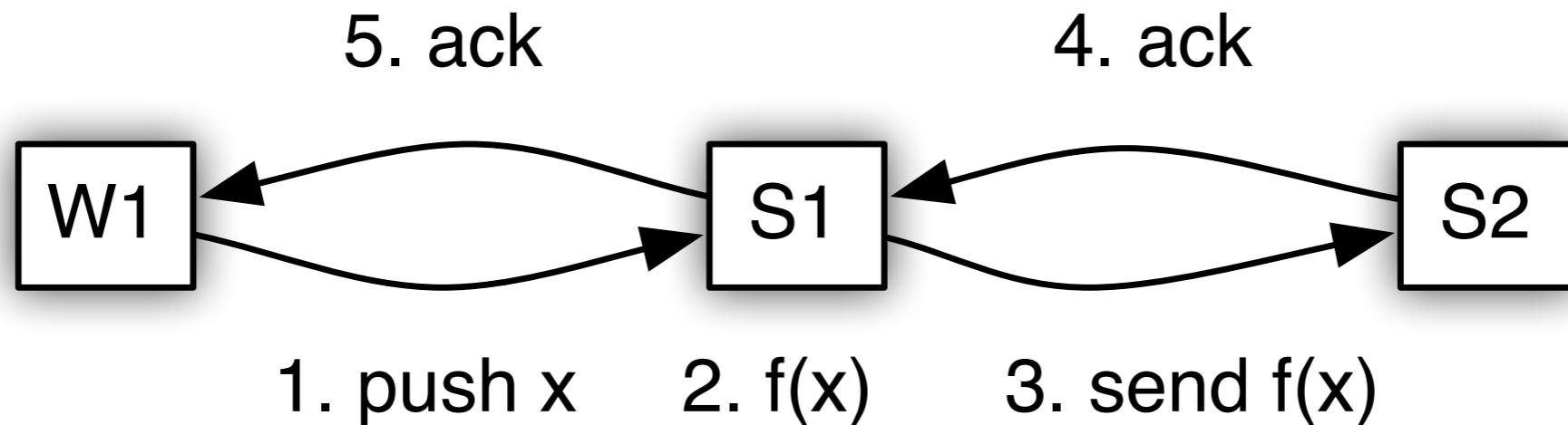
Communication tools

- ZeroMQ for messages
 - memcached way too slow (and no locks)
 - ICE is slow (YahooLDA uses it ... mea culpa)
 - MPI is stone age awkward
- Google Protobuf for serialization
 - good message compression
- Common pitfalls
 - Too small messages
 - Large metadata overhead

aggregation

compression
+ sparsity

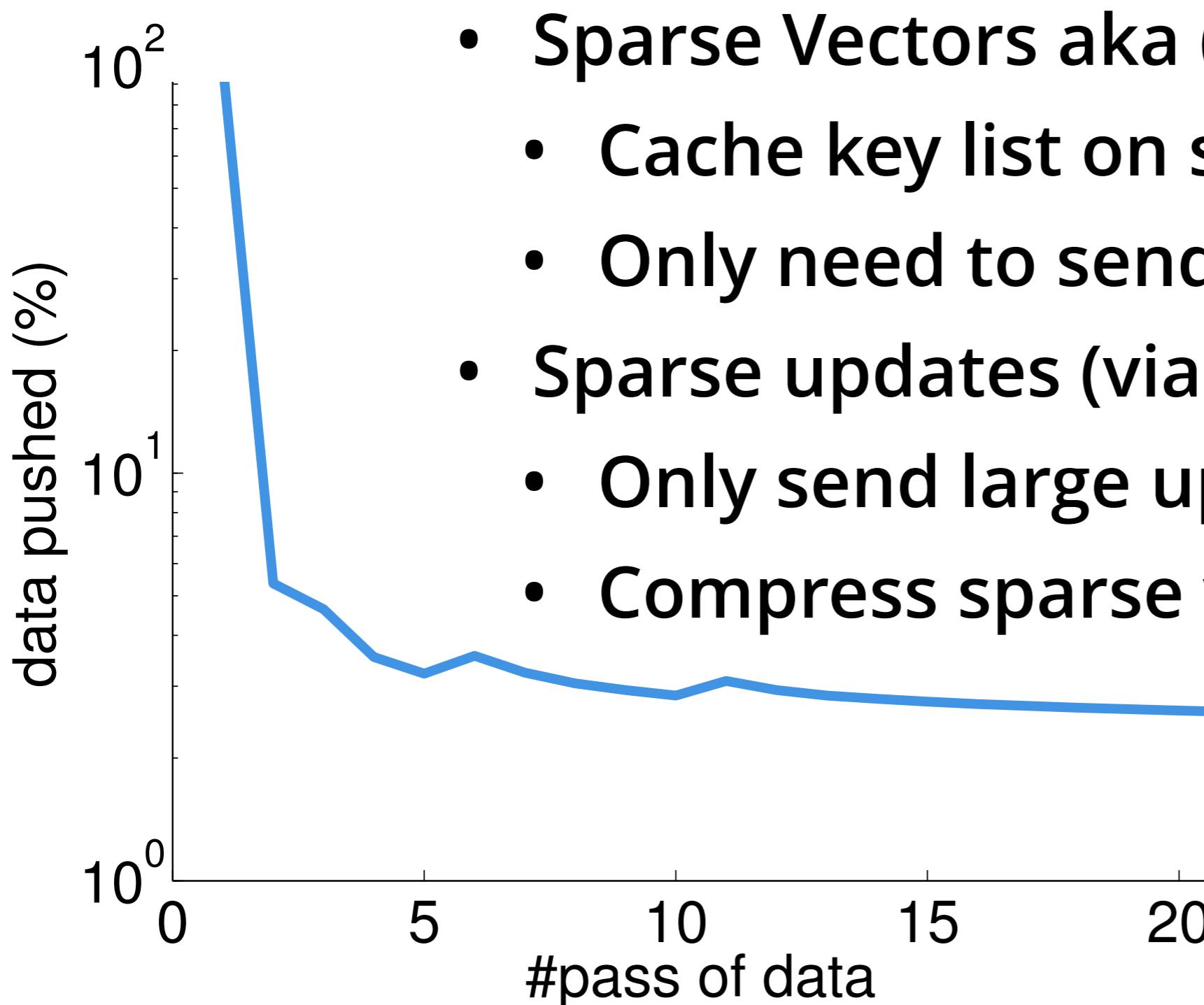
Message Aggregation on Server



Message Compression

- Convergence speed depends on communication efficiency
 - Sending (key,value) pairs is inefficient
Send only values (cache key list) instead
 - Sending small gradients is inefficient
Send only sufficiently large ones instead
 - Updating near-optimal values is inefficient
Send only large violators of KKT conditions
 - Filter data before sending

Message Compression



- Sparse Vectors aka (key,value) pairs
 - Cache key list on server
 - Only need to send values
- Sparse updates (via user defined filter)
 - Only send large updates
 - Compress sparse value list

Messaging

- Datatypes are eigen3 native
 - Dense vectors
 - Sparse vectors
- Push(Header flag)
- Pull(Header flag)
Flag may specify
 - Value or delta update
 - key range
 - recipient (all server, all clients, particular node)

Mailbox design

- Postmaster handles server liveness (i.e. master controller)
- Postoffice messages arrive here
- Van holds the message
- Box holds the actual datatype

Shared pointer. No copy on queue (by default)!

Consistency models

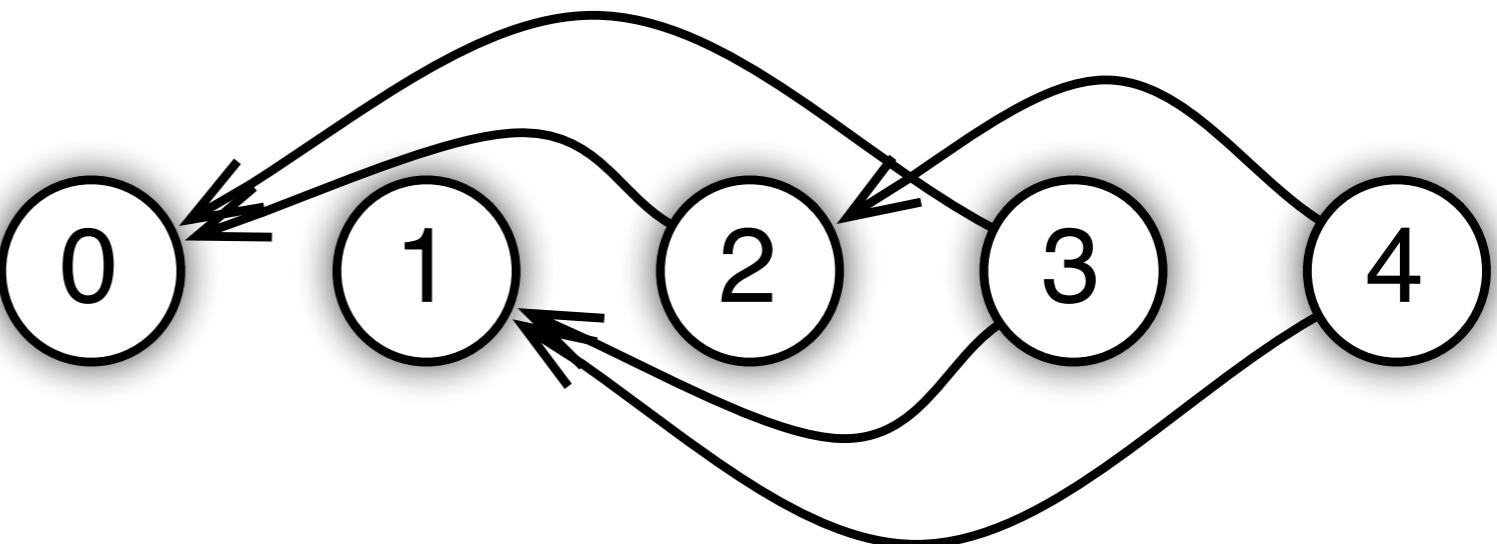
(a) Sequential



(b) Eventual



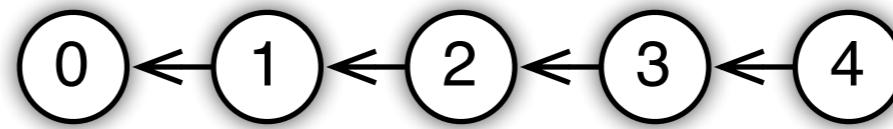
(c) Bounded delay



via task processing engine on client/controller

Consistency models

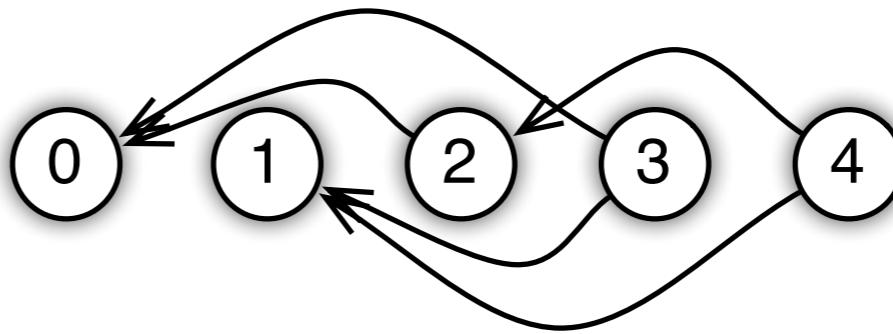
(a) Sequential



(b) Eventual



(c) Bounded delay



- Change dependency on the fly
- Task granularity programmatically defined (small or large tasks)
- Subtree controlled by worker

A dense collection of model trains on multiple levels of tracks. The trains are of various types, including steam locomotives, electric locomotives, and passenger cars, all in different colors and designs. The tracks are arranged in a complex, overlapping pattern across several levels.

Models



Logistic Regression

Recall - Computational Advertising

A screenshot of a Google search results page for the query "mesothelioma". The search bar at the top contains "mesothelioma". Below it, the navigation bar includes "Web" (which is highlighted in red), "Videos", "News", "Images", "Books", "More", and "Search tools". A message indicates "About 2,970,000 results (0.21 seconds)". The search results are divided into two columns. The left column contains five organic search results and one sponsored ad. The right column contains four organic search results and three sponsored ads. The sponsored ads are outlined with a thick black border.

Left Column (Organic Results):

- Mesothelioma Compensation**
Ad www.nationalmesotheliomaclaims.com/
The Money's Already There. \$30 Billion Asbestos Trust Fund
What Is Mesothelioma? - National Claims Center - Mesothelioma Claims
- Mesothelioma Symptoms - Mesothelioma-Answers.org**
Ad www.mesothelioma-answers.org/
By Anna Kaplan, M.D. 101 Facts about Mesothelioma.
Asbestos - Treatments - Top Doctors - Free Mesothelioma Book
- CA Mesothelioma Resource - californiamesothelioma.com**
Ad www.californiamesothelioma.com/ (800) 259-9249
Learn about mesothelioma & receive a free book of helpful answers.
What is Mesothelioma? - Asbestos Exposure in CA - California Legal Rights
- Mesothelioma Cancer - Mesothelioma.com**
www.mesothelioma.com/mesothelioma/
by Dr. Howard Jack West - Apr 2, 2014 - Mesothelioma is an aggressive cancer affecting the membrane lining ... Between 50 and 70% of all mesotheliomas are of the epithelial variety.
Mesothelioma Symptoms - Mesothelioma Prognosis - Mesothelioma Survival Rate
- Mesothelioma - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Mesothelioma Wikipedia
Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Asbestos - Mesothelium - Paul Kraus - Category:Mesothelioma

Right Column (Organic Results and Ads):

- Mesothelioma**
Ads www.mesothelioma-attorney-locators.com/
Easily Find Mesothelioma Attorneys.
Locations Across The United States
- CA Mesothelioma**
www.mesotheliomatreatmentcenters.org/
Mesothelioma? Get the Money you Deserve Fast-Help Filing your Claim
- Mesothelioma Compensation**
www.mesotheliomaclaimscenter.info/ (877) 456-3935
Mesothelioma? Get Money You Deserve Fast! Get Help with Filing a Claim.
- California Mesothelioma**
[\(888\) 707-4525](http://www.mesotheliomaattorney-usa.com/Legal)
100% Free Mesothelioma Legal Help!
\$30 Billion Trust Fund Available.
- Mesothelioma**
meso.lawyers.local.alotresults.com/
Seasoned Lawyers in your Area.
In your Local Lawyer Listings!

sponsored
search picks
position of
ad using

$$p(\text{click}|\text{ad}) \cdot \text{bid}(\text{ad})$$

estimate it

4 million/minute

Carnegie Mellon University

Recall - Computational Advertising

A Google search results page for the query "mesothelioma". The results are filtered by "Web". The page shows approximately 2,970,000 results found in 0.21 seconds. The results are organized into two columns. The left column contains organic search results and the right column contains sponsored ads.

Organic Search Results:

- Mesothelioma Compensation**
Ad www.nationalmesotheliomaclaims.com/
The Money's Already There. \$30 Billion Asbestos Trust Fund
What Is Mesothelioma? - National Claims Center - Mesothelioma Claims
- Mesothelioma Symptoms - Mesothelioma-Answers.org**
Ad www.mesothelioma-answers.org/
By Anna Kaplan, M.D. 101 Facts about Mesothelioma.
Asbestos - Treatments - Top Doctors - Free Mesothelioma Book
- CA Mesothelioma Resource - californiamesothelioma.com**
Ad www.californiamesothelioma.com/ (800) 259-9249
Learn about mesothelioma & receive a free book of helpful answers.
What is Mesothelioma? - Asbestos Exposure in CA - California Legal Rights
- Mesothelioma Cancer - Mesothelioma.com**
www.mesothelioma.com/mesothelioma/
by Dr. Howard Jack West - Apr 2, 2014 - Mesothelioma is an aggressive cancer affecting the membrane lining ... Between 50 and 70% of all mesotheliomas are of the epithelial variety.
Mesothelioma Symptoms - Mesothelioma Prognosis - Mesothelioma Survival Rate
- Mesothelioma - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Mesothelioma Wikipedia
Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Asbestos - Mesothelium - Paul Kraus - Category:Mesothelioma

Sponsored Ads:

- Mesothelioma**
Ads www.mesothelioma-attorney-locators.com/
Easily Find Mesothelioma Attorneys.
Locations Across The United States
- CA Mesothelioma**
www.mesotheliomatreatmentcenters.org/
Mesothelioma? Get the Money you Deserve Fast-Help Filing your Claim
- Mesothelioma Compensation**
www.mesotheliomaclaimscenter.info/ (877) 456-3935
Mesothelioma? Get Money You Deserve Fast! Get Help with Filing a Claim.
- California Mesothelioma**
[\(888\) 707-4525](http://www.mesotheliomaattorney-usa.com/Legal)
100% Free Mesothelioma Legal Help!
\$30 Billion Trust Fund Available.
- Mesothelioma**
meso.lawyers.local.alotresults.com/
Seasoned Lawyers in your Area.
In your Local Lawyer Listings!

sponsored
search picks
position of
ad using

$$p(\text{click}|\text{ad}) \cdot \text{bid}(\text{ad})$$

estimate it

4 million/minute

Carnegie Mellon University

Estimating Probabilities

- Logistic model (exponential family)

$$p(y|t) \propto \exp\left(\frac{1}{2}yt\right) \text{ where } y \in \{\pm 1\}$$

y will tend to agree with the sign of t (**find t**)

- Normalizing terms

$$\begin{aligned} p(y|t) &= \frac{\exp\left(\frac{1}{2}yt\right)}{\exp\left(\frac{1}{2}t\right) + \exp\left(-\frac{1}{2}t\right)} = \frac{\exp\left(\frac{1}{2}yt\right)}{\exp\left(\frac{1}{2}yt\right) + \exp\left(-\frac{1}{2}yt\right)} \\ &= \boxed{\frac{1}{1 + \exp(-yt)}} \end{aligned}$$

(Penalized) Maximum Likelihood

- Goal

Find t that correlates with y

- Strategy

Use covariates x and function $f(x)$

$$p(y|x) = \frac{1}{1 + \exp(-yf(x))}$$

- Penalty against overfitting / Bayes rule

$$p(f|X, Y) \propto p(f) \prod_{i=1}^m \frac{1}{1 + \exp(-y_i f(x_i))}$$

Penalized Maximum Likelihood

- Picking a function class

$$f(x) = \langle w, x \rangle$$

we want sparse
models for advertising

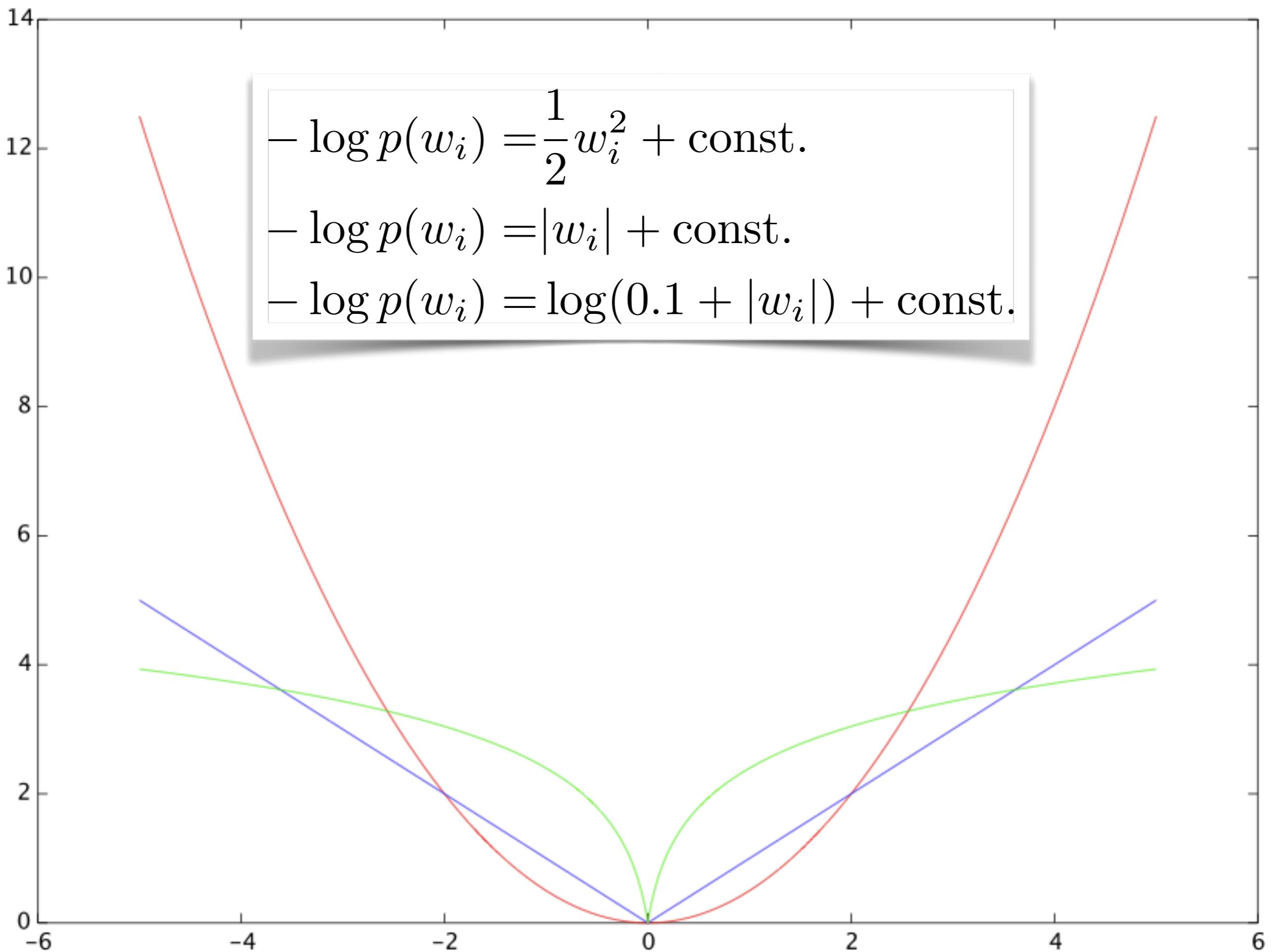
- Picking a prior

$$\log p(f) = \lambda \|w\|_1 + \text{const.}$$

- Picking an inference strategy

$$\underset{w}{\text{minimize}} -\log p(f|X, Y)$$

$$\underset{w}{\text{minimize}} \sum_{i=1}^m \log(1 + \exp(-y_i \langle w, x_i \rangle)) + \lambda \|w\|_1$$



Proximal Algorithm

- Problem - ℓ_1 norm is non-smooth
- Proximal operator

$$\operatorname{argmin}_w \|w\|_1 + \frac{\gamma}{2} \|w - (w_t - \eta g_t)\|$$

(more generally use penalty on w)

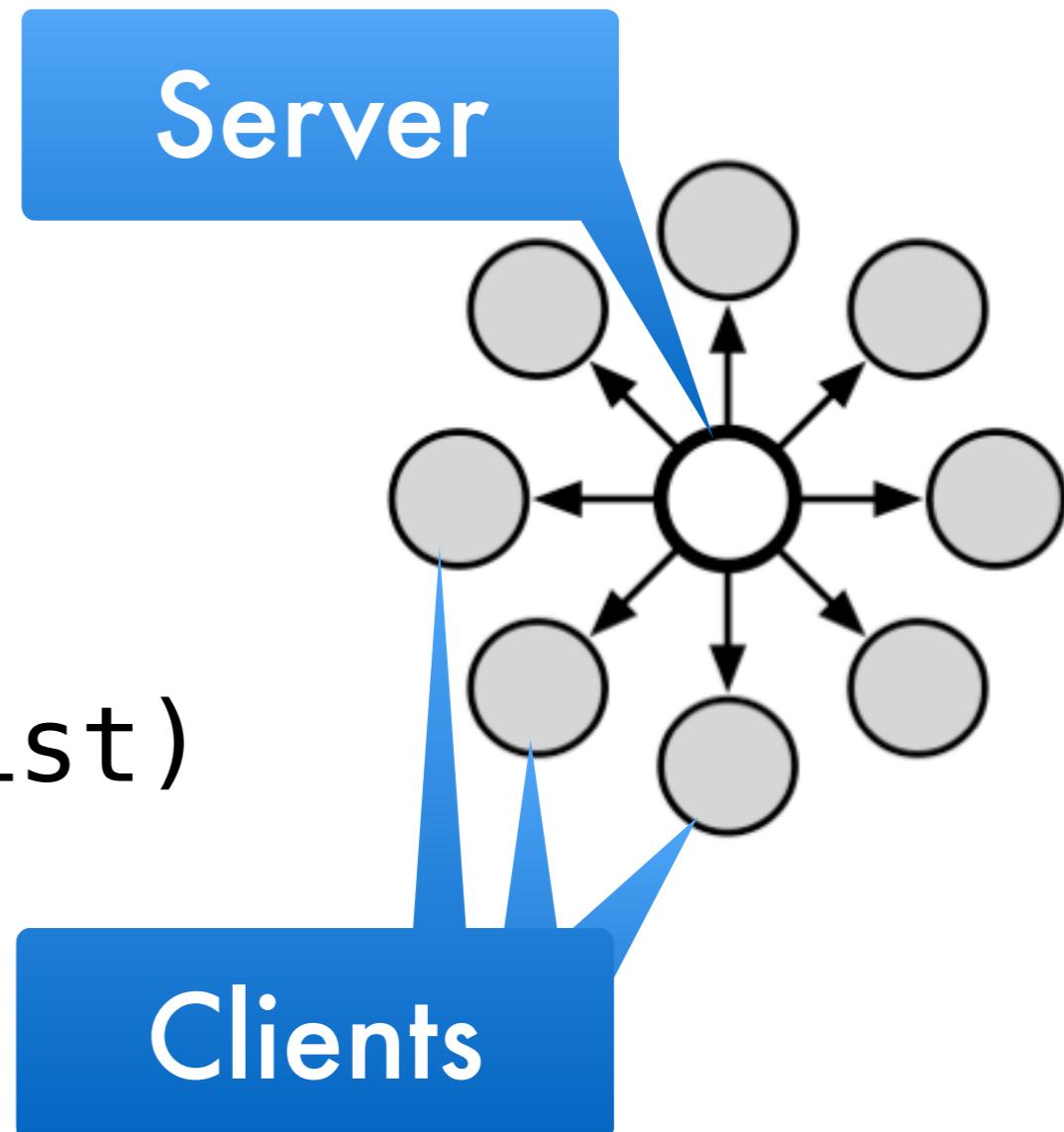
- Updates for ℓ_1 are

$$w_i \leftarrow \operatorname{sgn}(w_i) \max(0, |w_i| - \epsilon)$$

more detail in Mu's tutorial

Generic Parallel Template

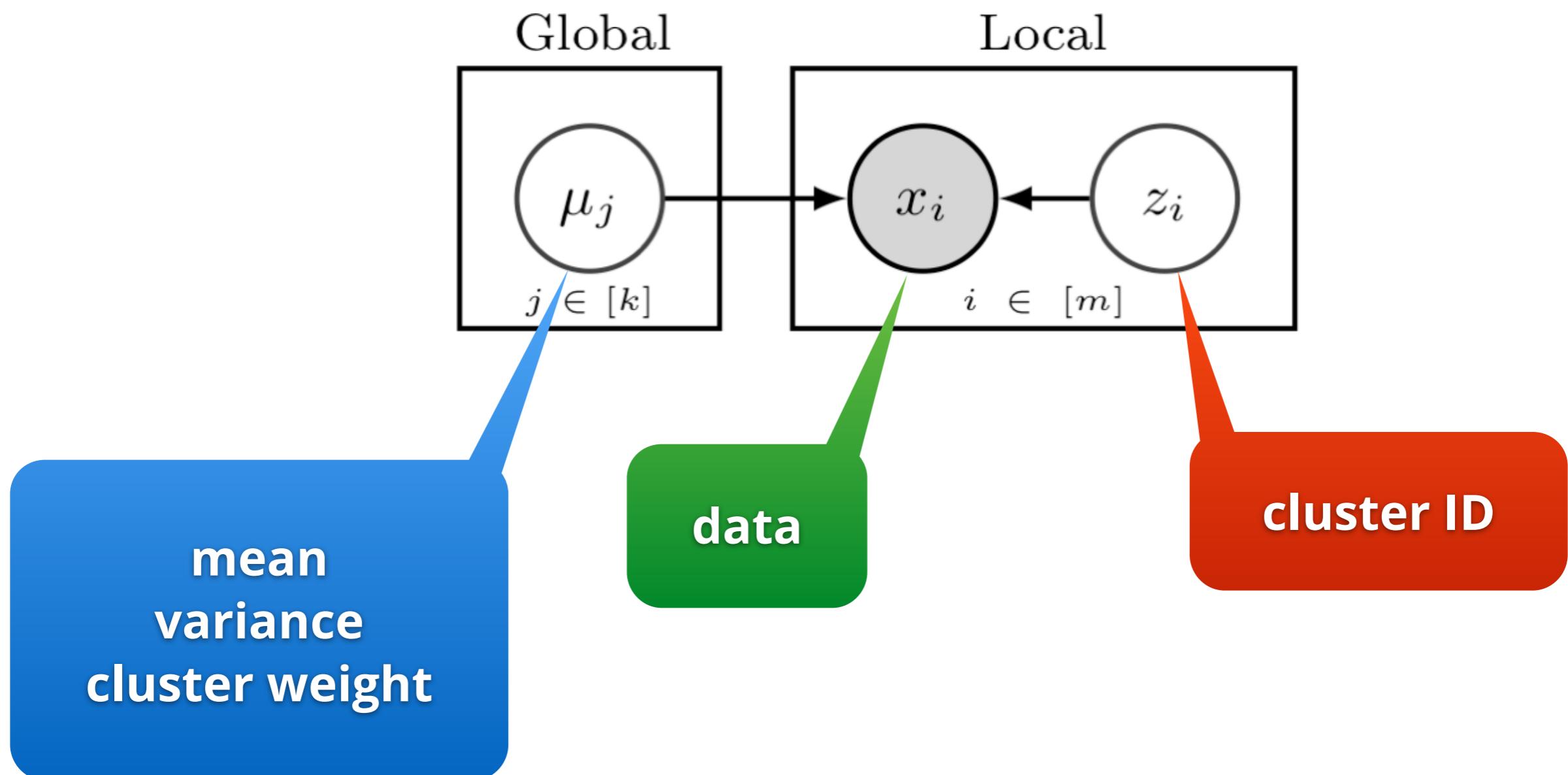
- Compute gradient on (subset of data) on each client
- Send gradient from client to server asynchronously
`push(key_list,value_list)`
- Proximal gradient update on server
- Server returns parameters
`pull(key_list,value_list)`



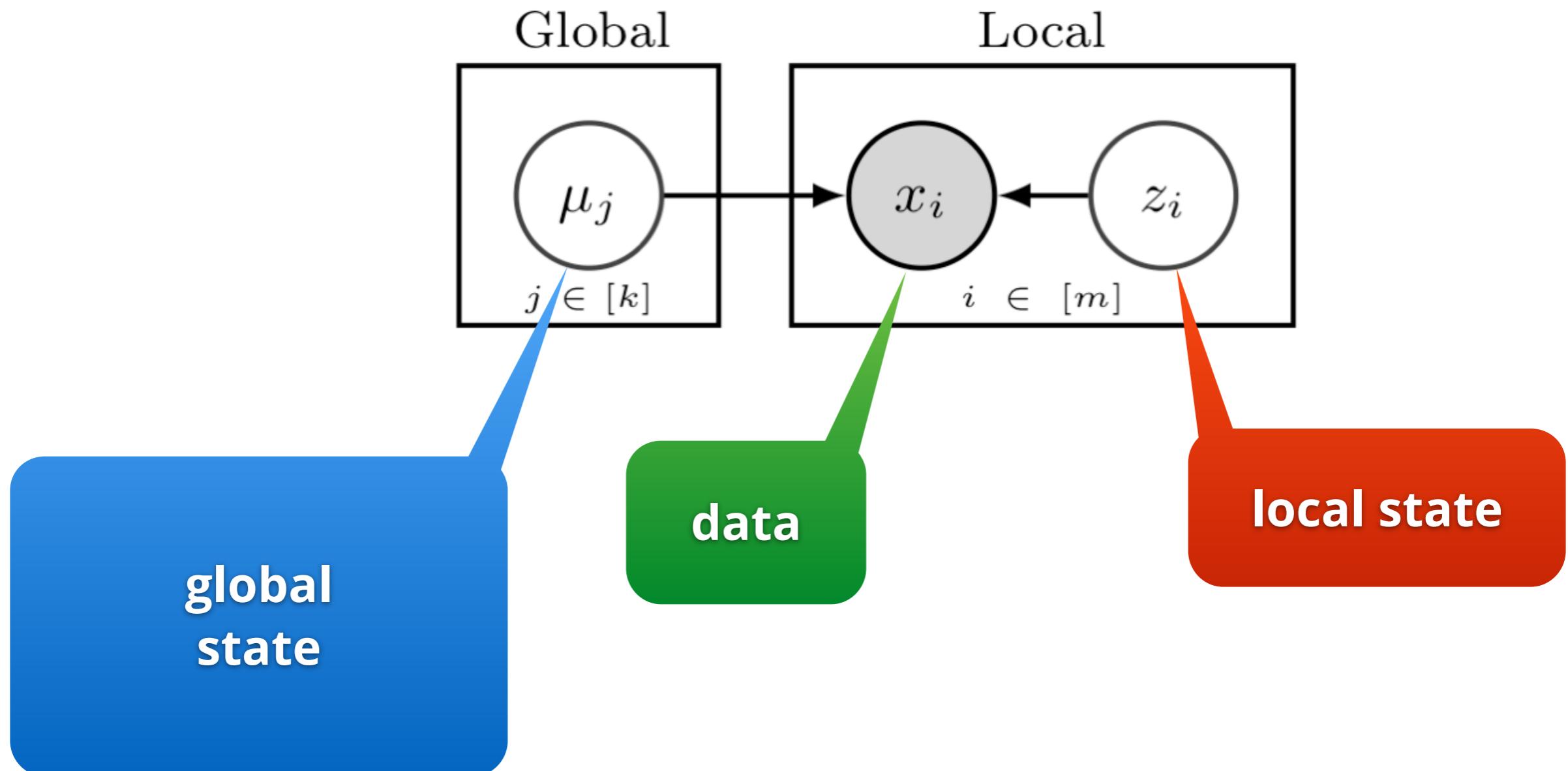


Scaling Latent Variable Models

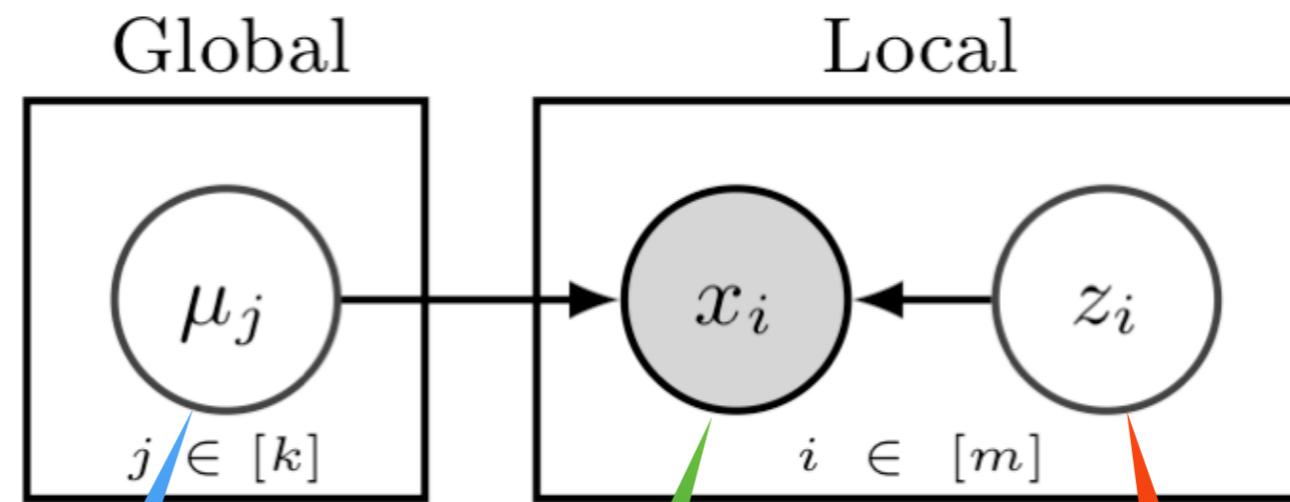
Clustering



Clustering



Clustering

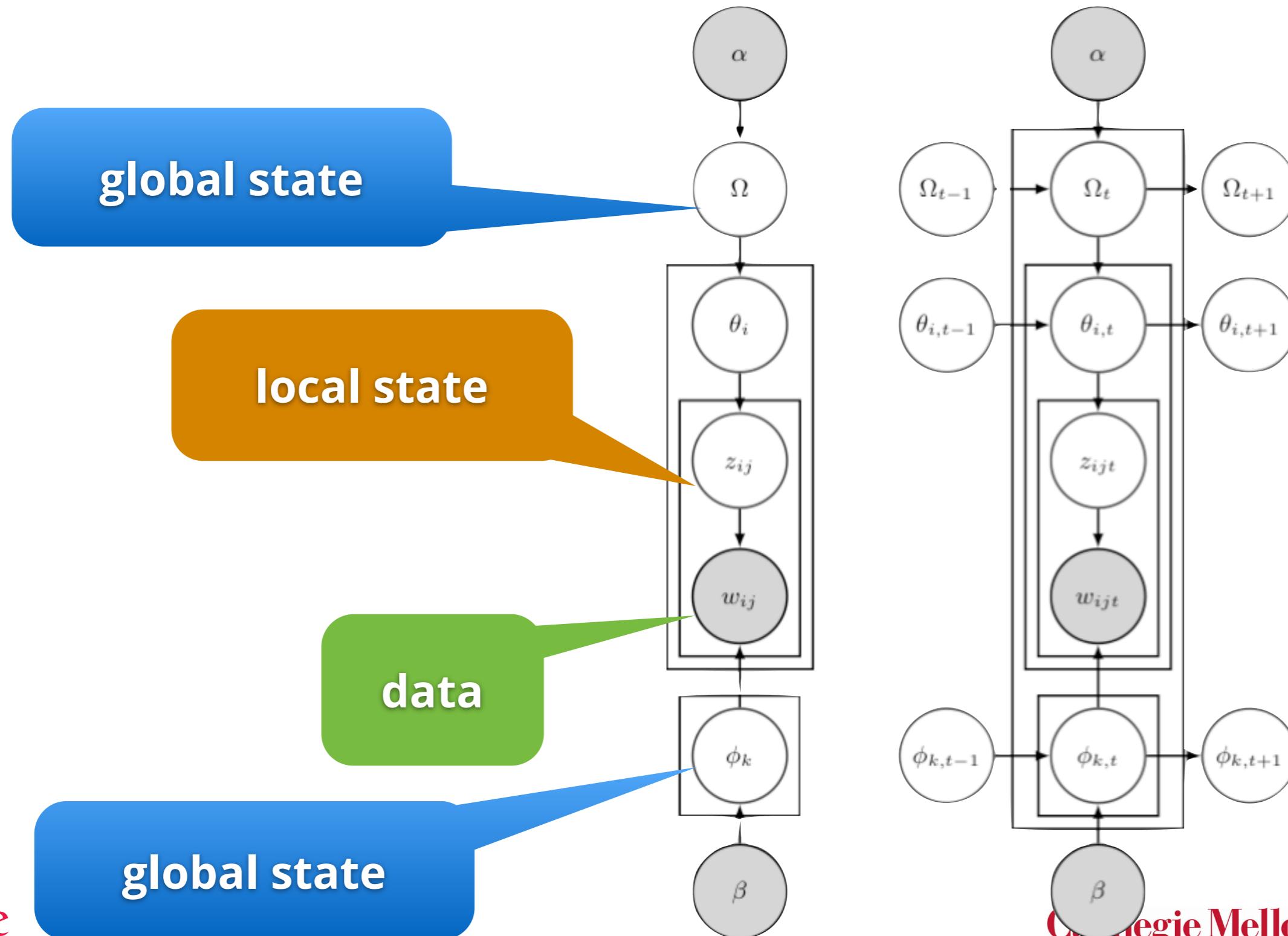


too big for
a single machine

huge

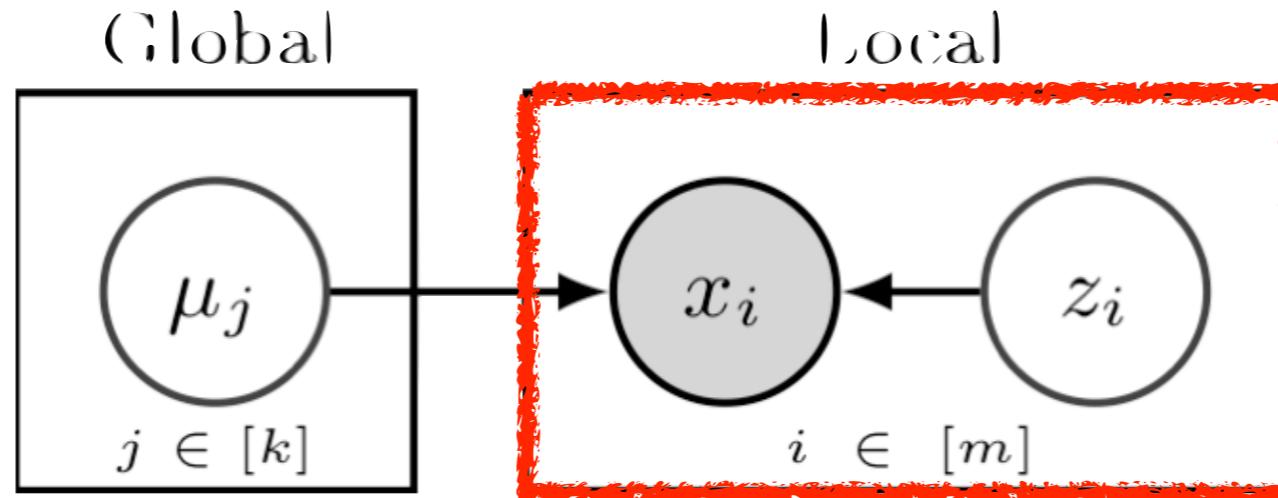
only local

Fancy models



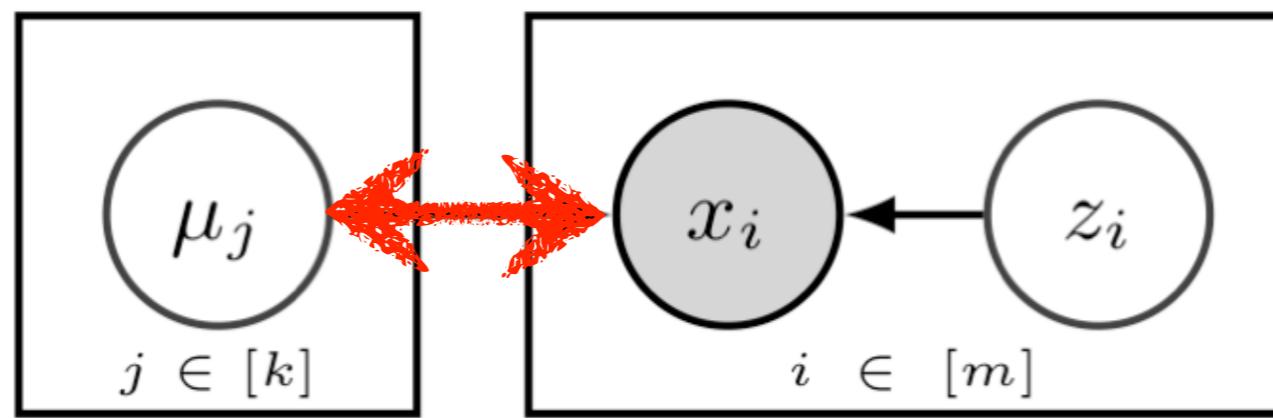
ParameterServer to the rescue

local state
is too large

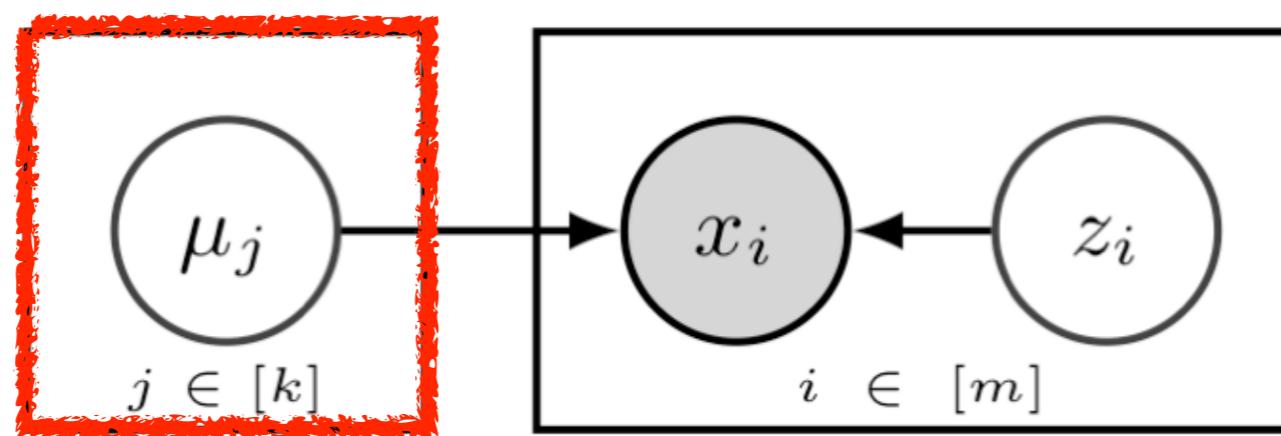


does not fit
into memory

global state
is too large



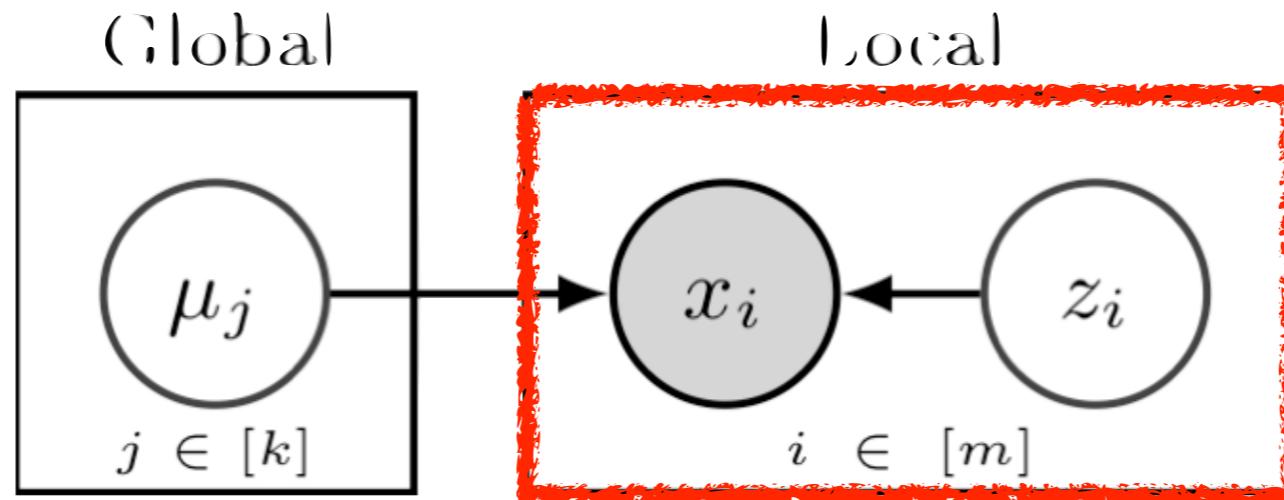
network load
& barriers



does not fit
into memory

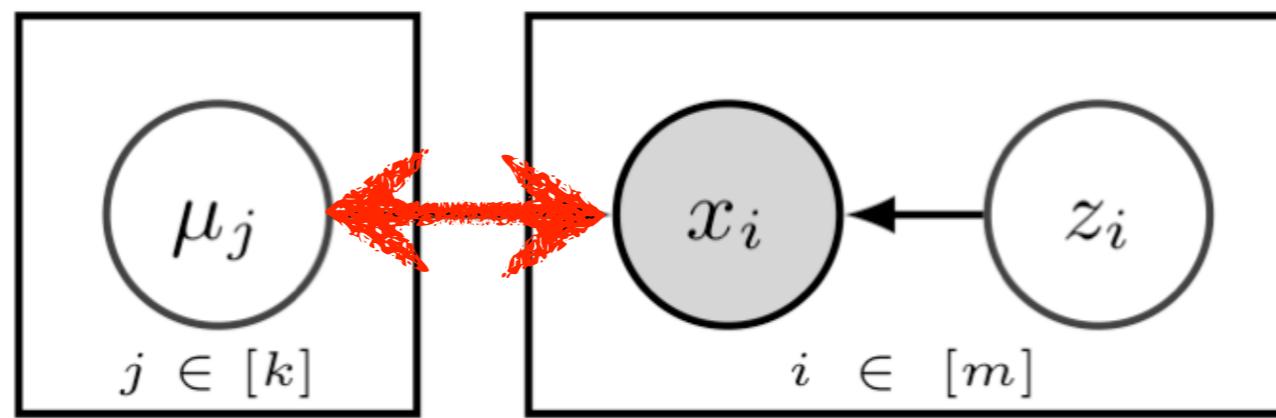
ParameterServer to the rescue

local state
is too large

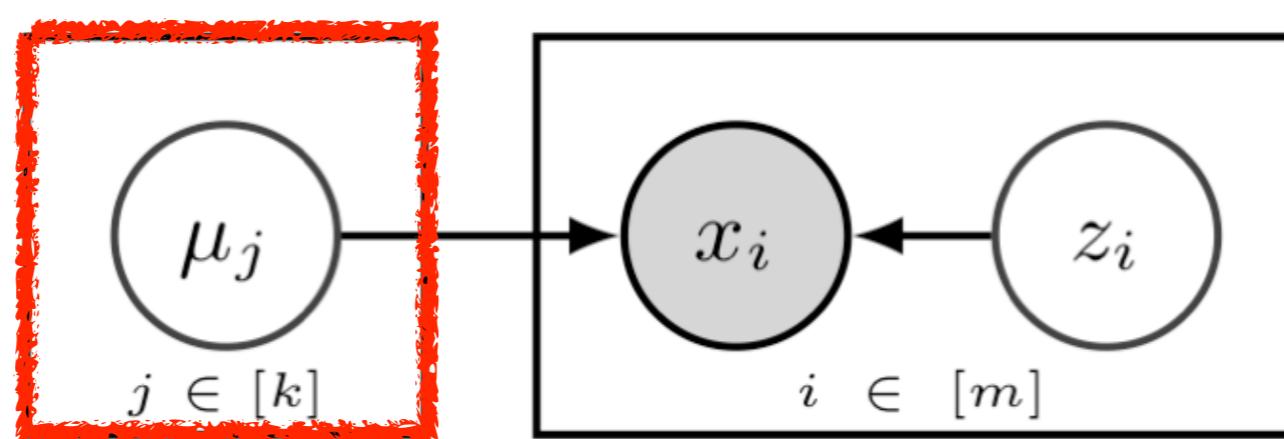


stream local
data from disk

global state
is too large



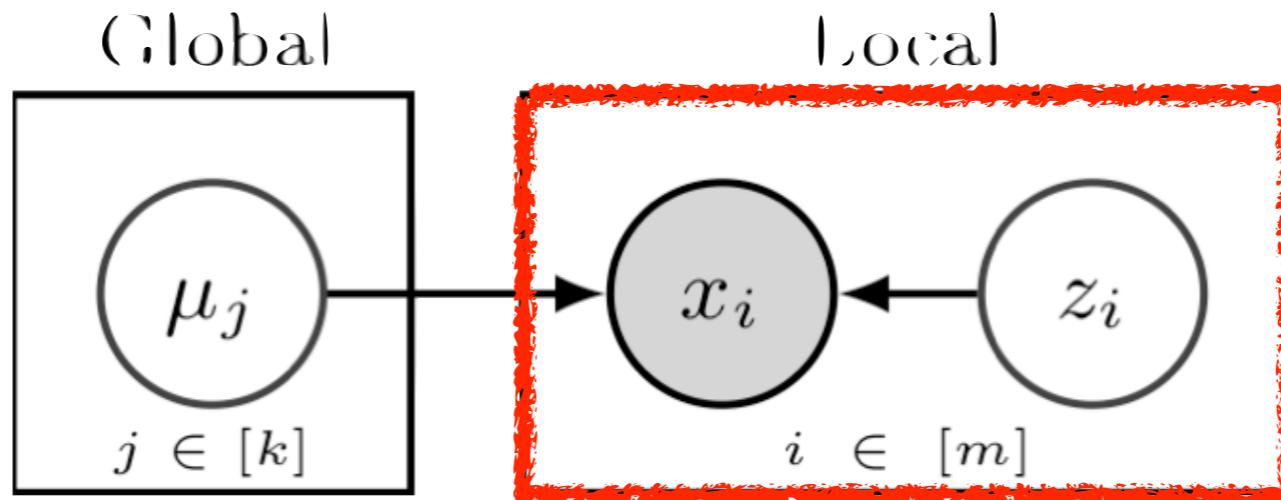
network load
& barriers



does not fit
into memory

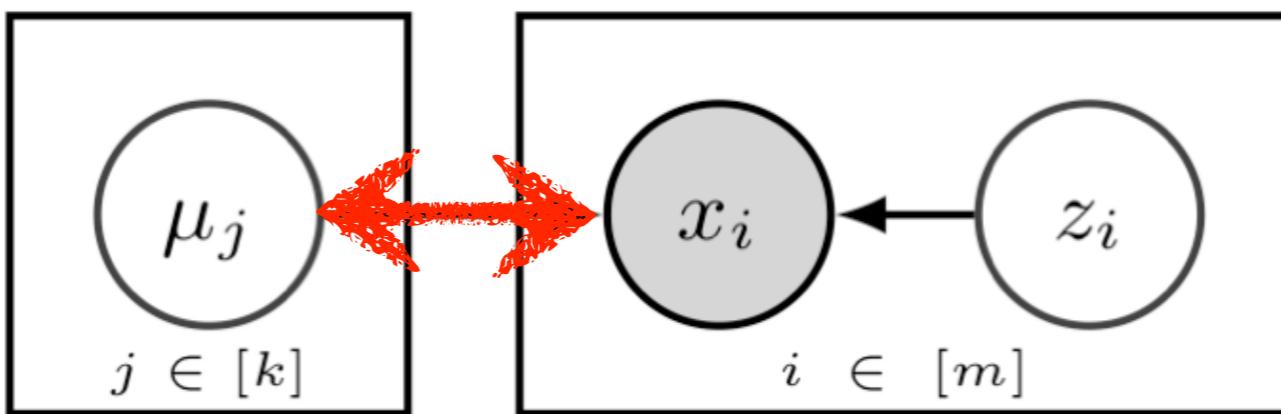
ParameterServer to the rescue

local state
is too large

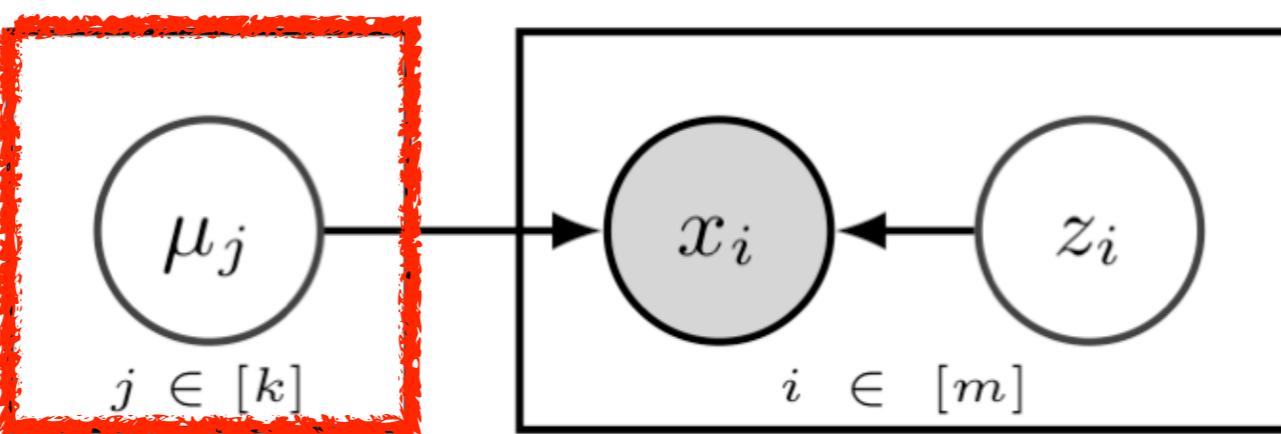


stream local
data from disk

global state
is too large



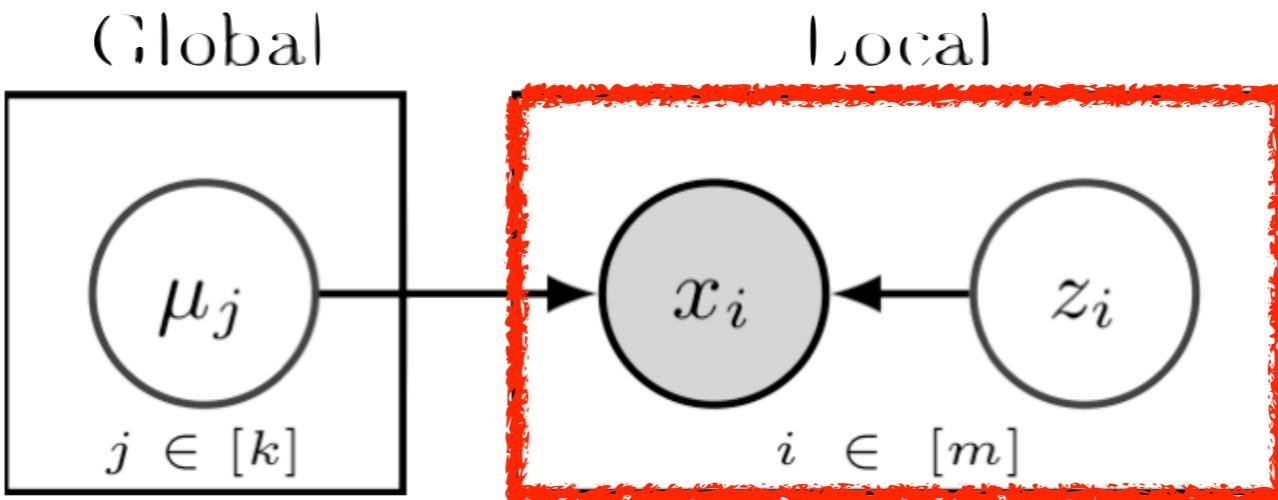
asynchronous
synchronization



does not fit
into memory

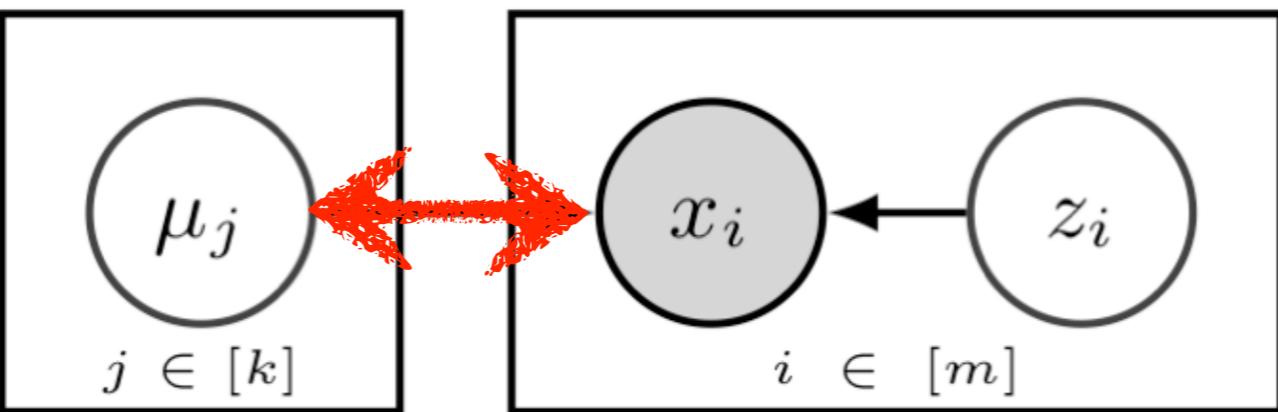
ParameterServer to the rescue

local state
is too large

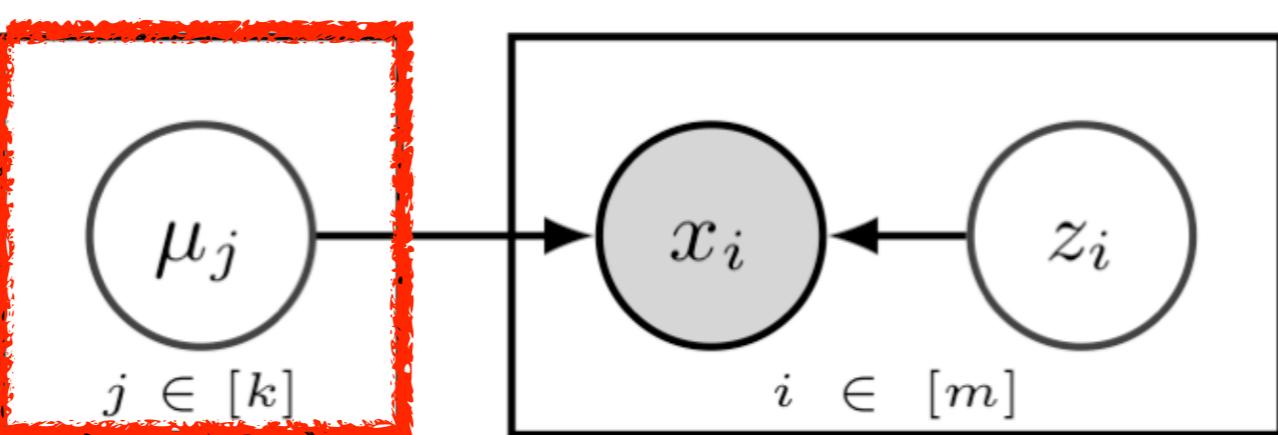


stream local
data from disk

global state
is too large

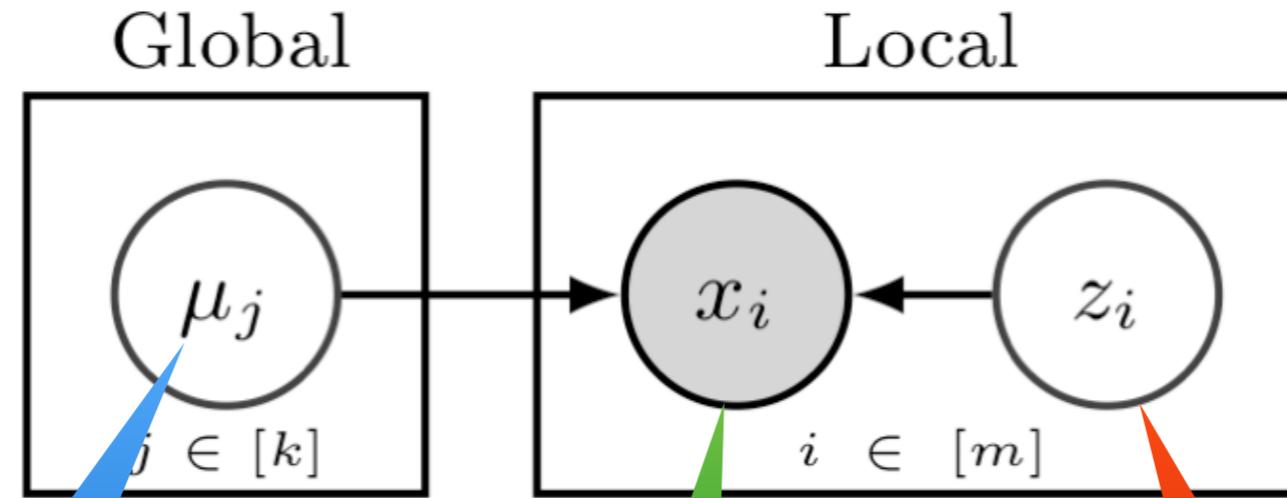


asynchronous
synchronization



partial view
shared between
threads

Synchronization Strategy



mean
variance
cluster weight

data

cluster ID

- Locally Gibbs Sample cluster ID

$$p(z_i|x_i, \text{rest}) \propto p(z_i|Z^{-i})p(x_i|X^{-i}, Z^{-i}, z_i)$$

- Communicate changes in statistics of data to server
(mean, variance, cluster size)

Mixture of Gaussians

- Multinomial with Dirichlet for cluster ID

$$p(Z|\theta) = \prod_{i=1}^m \theta_{z_i} \text{ and } p(\theta|\alpha) = \text{Dir}(\alpha)$$

- Integrating out multinomial yields collapsed

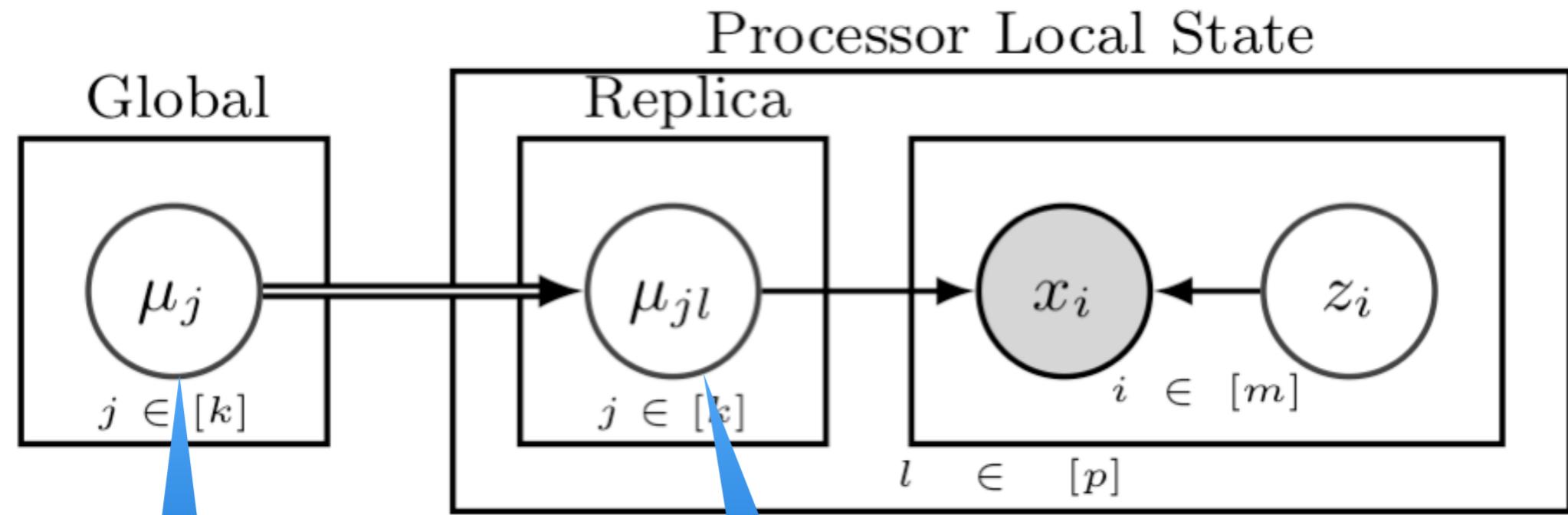
$$p(z_i = z | Z^{-i}) = \frac{n_z^{-i} + \alpha_z}{n - 1 + \sum_{z'} \alpha_{z'}}$$

- Gaussian with Gauss-Wishart for data

$$x_i | z_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i}) \text{ and } (\mu_{z_i}, \Sigma_{z_i}) \sim \text{GaussWishart}(m_0, \mu_0, Q_0)$$

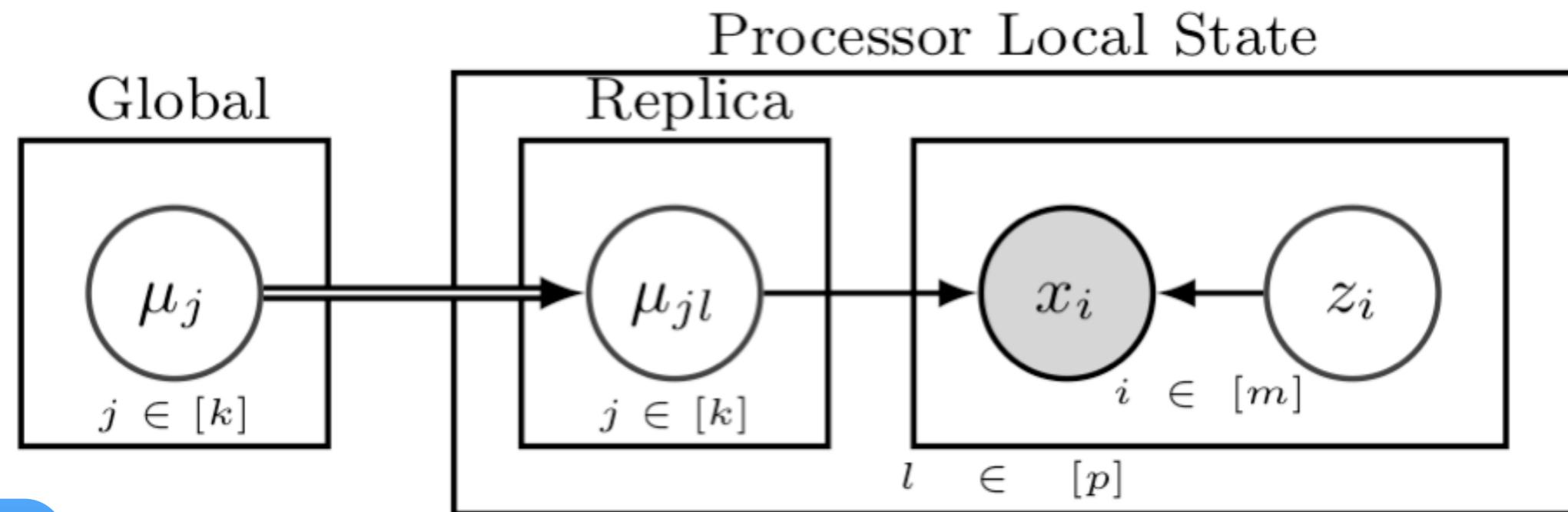
- Only need to sync $(n_z, l_z, Q_z) := \sum_{z_i=z} (1, x_i, x_i x_i^\top)$

Local and Global Variables

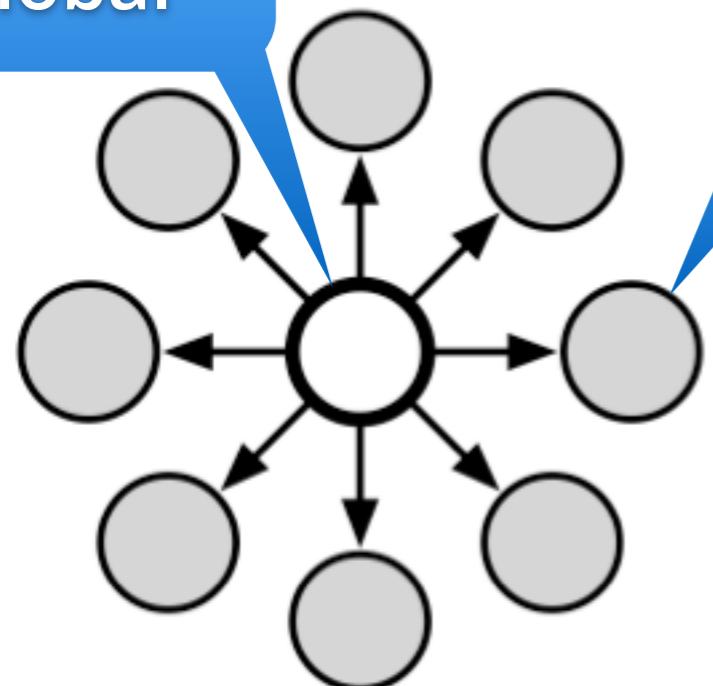


- No locks between machines to access z
- Synchronization mechanism for global μ needed
- In LDA this is the local copy of the (topic,word) counts

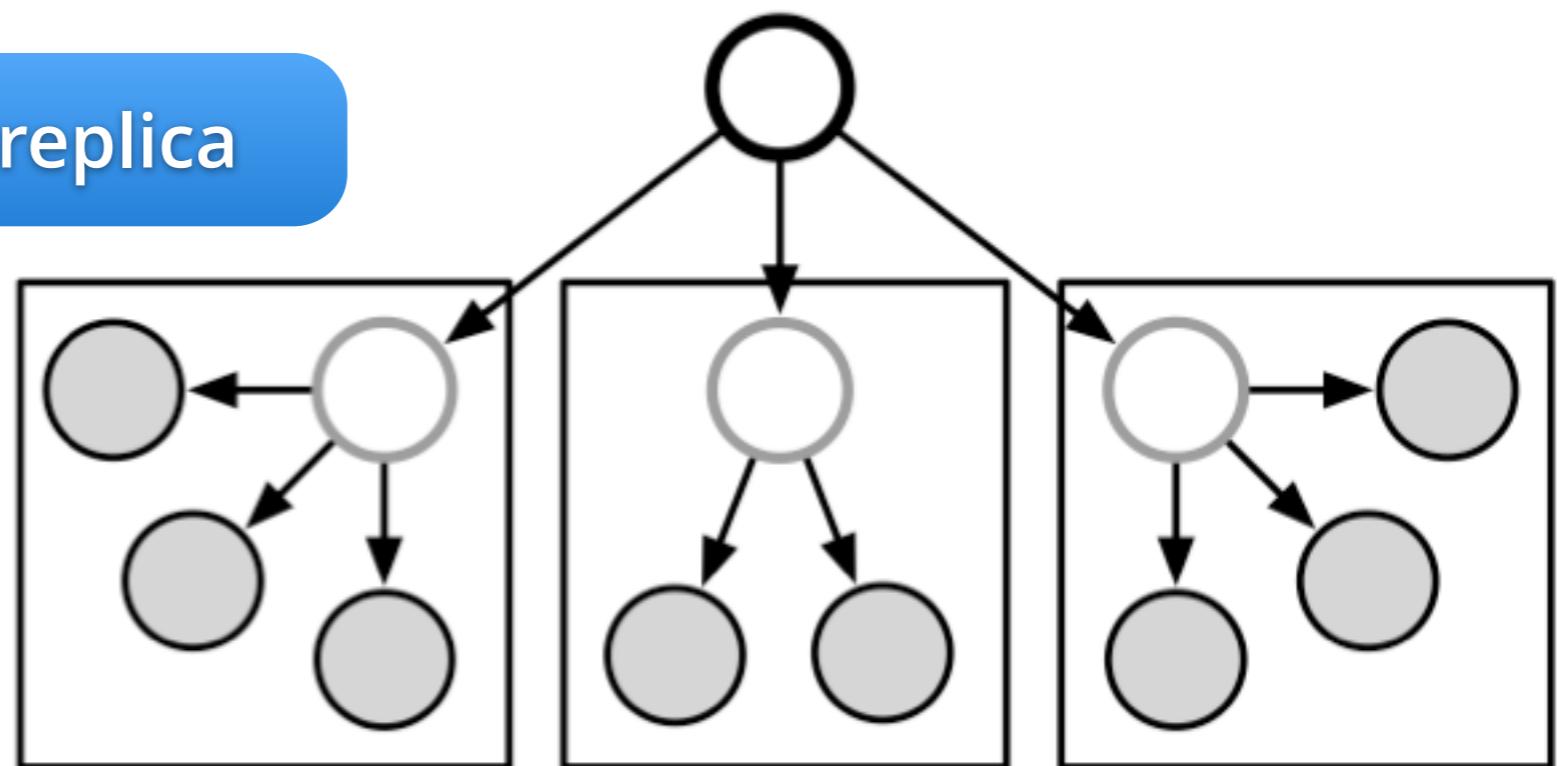
Local and Global Variables



global



replica

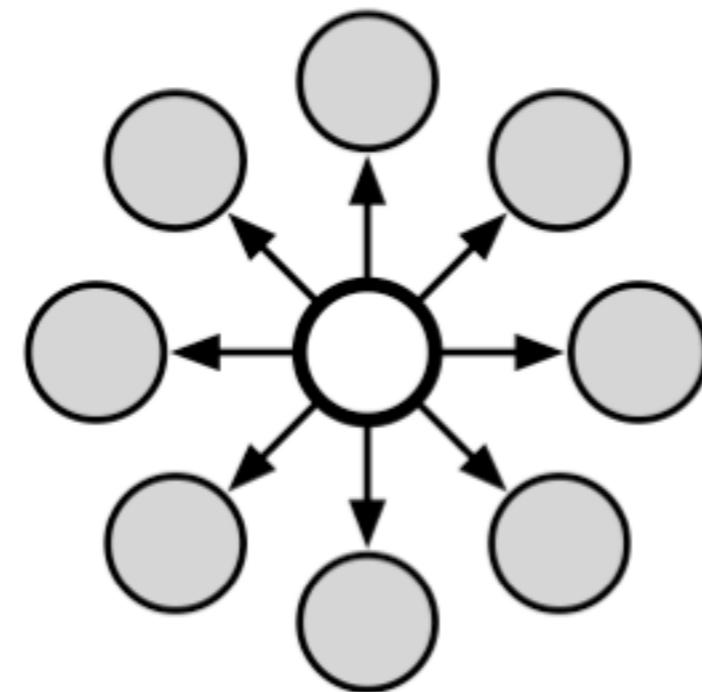


Message Passing

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

local to global

$$\begin{aligned}\delta &\leftarrow x - x^{\text{old}} \\ x^{\text{old}} &\leftarrow x \\ x^{\text{global}} &\leftarrow x^{\text{global}} + \delta\end{aligned}$$



global to local

$$\begin{aligned}x &\leftarrow x + (x^{\text{global}} - x^{\text{old}}) \\ x^{\text{old}} &\leftarrow x^{\text{global}}\end{aligned}$$



Models

Grouping objects

Grouping objects

The image displays three distinct web pages side-by-side:

- Singapore Airlines:** The top navigation bar includes links for Help, Site Map, Contact Us, Singapore (highlighted in yellow), Change Location, Search, and several menu items: The Experience, Flights & Fares, Before You Fly, Loyalty Programmes, and Promotions.
- National University of Singapore (NUS):** The header features the NUS logo and navigation links for myEMAIL, IVLE, LIBRARY, MAPS, CALENDAR, SITEMAP, CONTACT, and e-CARDS. Below the header, there's a search bar and a booking interface for flights. The main menu includes ABOUT NUS, GLOBAL, ADMISSIONS, EDUCATION, RESEARCH, ENTERPRISE, CAMPUS LIFE, GIVING, and CAREERS@NUS.
- Chijmes:** This page promotes Chijmes as a lifestyle destination. It features a large image of a historic building at night, text about its history, and a photo of a smiling couple. It also lists partners like Suntec, ARA, and PAC, along with staff, alumni, and visitors buttons.

A blue speech bubble highlights the word "Singapore" on the NUS website's main menu.

Chijmes Text:

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

Chijmes, a premier lifestyle destination in Singapore

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

Copyright © 2006 Chijmes. All rights reserved.

Feedback | Terms & Conditions

Carnegie Mellon University:

Carnegie Mellon University

Grouping objects

UNITED

My profile | Worldwide sites | Customer service

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

#**ON TIME** United. #1 in on-time arrivals. [Details](#)

Flights | Check-in | Flight status

BOOK FLIGHT | REDEEM MILES

From (Find airport) To (Find airport)

Search nearby airports Search nearby airports

Roundtrip One-way > Multicity

Departing Anytime

Returning Anytime

Search by Schedule & price Price > Flexible

Adult (child or senior?)

Cabin Refundable

Promotion code or Electronic certificate More info

[Log in to view all seating options](#)

[Advanced Search](#)

Cars | Hotels | Vacations

Use 30% fewer miles on your next United flight.



Save now on Saver Awards for flights 700 miles or less. [Learn more](#)

3 of 6

United news and deals

- > Travel waiver issued due to Hurricane Earl
- > E-Fares: Save on weekend getaways
- > Opt to send your bags ahead
- > Wireless check-in, paperless boarding
- > Receive deal alerts: Follow us on Twitter
- > Take our survey & you could win miles

United-Continental merger [Learn more about the merger](#)

© 1998-2010 Chez Panisse Restaurant

A STAR ALLIANCE MEMBER 

Singapore Change Location Search

Before You Fly Loyalty Programmes Promotions

Log in

Mileage Plus # or email address

Password [Forgot password?](#) [Need password?](#)

Remember me

Start with My Mileage Plus My reservations

Start earning miles today [Join Mileage Plus](#)

united.com benefits and features

- > Low Fare Guarantee
- > Why united.com? *New!*

Travel information

- > Updates to baggage & standby policies
- > View travel requirements and regulations

Play now. Win now. Optathlon mobile and online games. The gold medal is a million miles.

Mileage Plus UNITED 4411 123 8888 0000 VISA 

Earn up to 30,000 Bonus Miles

Search ANU...

CALENDAR SITEMAP CONTACT e-CARDS

GIVING CAREERS@NUS

WEB CONTACTS MAP GO

The Australian National University

About United | Investor relations | Business resources | Careers | Site map

Owned by:  Managed by:  Property Manager: 

© 1998-2010 Chez Panisse Restaurant

Forests renew after Black Saturday fires

School of Music at Floriade

Undergraduate studies

Higher Degree Research

Copyright © 2006 Chijmes. All rights reserved.

Grouping objects

UNITED

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

ON TIME United, #1 in on-time arrivals. Details

Flights | Check-in | Flight status

BOOK FLIGHT | REDEEM MILES

From (Find airport) To (Find airport)

Search nearby airports | Roundtrip | One-way | Multicity

Departing Anytime

Returning Anytime

Search by Schedule & price | Price | Flexible

Adult (child or senior?)

Cabin Economy | Refundable

Promotion code or Electronic certificate More info

Log in to view all seating options

Advanced Search | Search

Cars | Hotels | Vacations

Use 30% fewer miles on your next United flight.

%

Save now on Saver Awards for flights 700 miles or less. Learn more

United news and deals

- > Travel waiver issued due to Hurricane Earl
- > E-Fares: Save on weekend getaways
- > Opt to send your bags ahead
- > Wireless check-in, paperless boarding
- > Receive deal alerts: Follow us on Twitter
- > Take our survey & you could win miles

United-Continental merger Learn more about the merger

KRISFLYER

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

Need Help? View Book A Flight Guide

→ SIA Holidays | Book Now | Show Schedule

→ Hotel Bookings

Log in

Mileage Plus # or email address

Forgot password?

Password

Need password?

Remember me

Start with

My Mileage Plus

My reservations

Log in

Start earning miles today Join Mileage Plus

united.com benefits and features

- > Low Fare Guarantee
- > Why united.com? New!

Travel information

- > Updates to baggage & standby policies
- > View travel requirements and regulations

Play now. Win now.

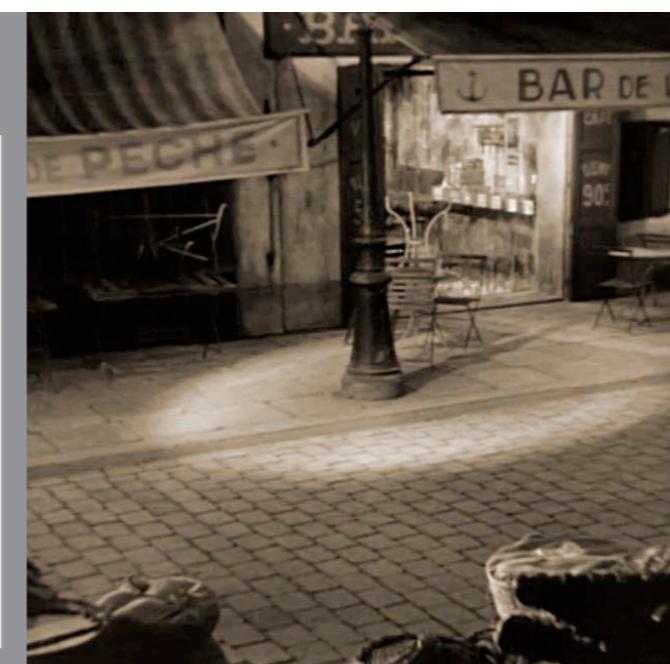
Opt-in to mobile and online games. The gold medal is a million miles.

Earn up to 30,000 Bonus Miles

Learn more

6-Digit PIN

Log-In Help | Log In →



EXPLORE ANU » A-Z INDEX »

Search ANU... WEB CONTACTS MAP GO

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

HOME FUTURE STUDENTS CURRENT STUDENTS RESEARCH & EDUCATION ABOUT ANU STAFF

Ash forests rise and rise again

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.

» read more

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

A Leader at NUS WATCH THE VIDEO

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

PROSPECTIVE STUDENTS CURRENT STUDENTS STAFF ALUMNI VISITORS



Google

e Mellon University

Grouping objects

The screenshot shows the United Airlines website interface. At the top, there's a navigation bar with links like "Planning & booking", "Reservations & check-in", "Mileage Plus®", "Services & information", and a search bar. Below this, a banner for "Saver Awards" offers 30% fewer miles on flights 700 miles or less. Another banner for "KrisFlyer" promotes earning up to 30,000 bonus miles. The main content area includes sections for "United news and deals" and "Travel information". At the bottom, there are links for "SIA Holidays" and "Hotel Bookings".

The screenshot shows the Chez Panisse website. It features a "RESERVATIONS" section with a form, a "MENUS" section listing "RESTAURANT • CAFÉ", "MONDAY NIGHTS • WINE LIST", and a "ABOUT" section with links to "CHEZ PANISSE • ALICE WATERS", "OUR CHEFS", "FRIENDS", "PRESS", "FOUNDATION & MISSION", "SPECIAL EVENTS", "CALENDAR", "STORE", "BOOKS • POSTERS • GIFTS", and "CONTACT" information.



The screenshot shows the Australian National University (ANU) website. The header includes links for "EXPLORE ANU", "A-Z INDEX", and a search bar. The main content area features a large image of a tree trunk with new green leaves sprouting from it. Below the image, there are several news and feature articles, such as "Ash forests rise and rise again" and "Forests renew after Black Saturday fires". The footer contains links for "PROSPECTIVE STUDENTS", "CURRENT STUDENTS", "STAFF", "ALUMNI", and "VISITORS".



Google

Grouping objects

UNITED

My profile | Worldwide sites | Customer service

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

#1 in on-time arrivals. Details

Flights | Check-in | Flight status

BOOK FLIGHT | REDEEM MILES

From (Find airport) To (Find airport)

Search nearby airports | One-way | Multicity

Departing Anytime

Returning Anytime

Use 30% fewer miles on your next United flight.

%

Log in

Mileage Plus # or email address | Password | Forgot password? | Need password?

Remember me

Start with My Mileage Plus | My reservations | Log in

Start earning miles today. Join Mileage Plus

united.com benefits and features

Chez Panisse USA

RESERVATIONS RESTAURANT & CAFÉ

MENUS RESTAURANT • CAFÉ MONDAY NIGHTS • WINE LIST

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

SPECIAL EVENTS CALENDAR

STORE BOOKS • POSTERS • GIFTS

CONTACT INFORMATION DIRECTIONS • MAILING LIST

Directions Reservations Contact

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

EXPLORE ANU » A-Z INDEX »

Search ANU... WEB CONTACTS MAP GO

THE AUSTRALIAN NATIONAL UNIVERSITY

ANU

HOME FUTURE STUDENTS CURRENT STUDENTS RESEARCH & EDUCATION ABOUT ANU STAFF

Ash forests rise and rise again

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.

» read more

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

NUS National University of Singapore

myEMAIL IVLE LIBRARY MAPS CALENDAR SITEMAP CONTACT e-CARDS

Search search for... in NUS Websites GO

The Experience Flights

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

A Leading Global University Centred in Asia

KrisFly

CLICK HERE TO FIND OUT MORE

Flame Arrival Ceremony at NUS WATCH THE VIDEO

Joint Evacuation Exercises

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

More details

ALUMNI VISITORS

CHIJMES restaurants • bars • shops

Singapore

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

Chijmes, a premier lifestyle destination in Singapore

Owned by: SUNTEC Real Estate Investment Trust Managed by: ARA Property Manager: APAC Investment Management Pte Ltd

Copyright © 2006 Chijmes. All rights reserved.

Feedback | Terms & Conditions

ellen University

Google

Topic Models

UNITED

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

From (Find airport) To (Find airport)

Start with: My Mileage Plus | My reservations

Log in | Book Flight | REDEEM MILES

Use 30% fewer miles on your next United flight.

From (Find airport) To (Find airport)

Search nearby airports | Departing: Anytime | Returning: Anytime

Search by: Schedule & price | Price | Flex | Adult | Child or senior? | Cabin: Economy | Refundable | Promotion code or Electronic certificate

Log in to view all seating options | Advanced Search | Search

Cars | Hotels | Vacations

About United | Investor relations | Business resources | Careers | Site map | A STAR ALLIANCE MEMBER

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

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

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

HOME FUTURE STUDENTS CURRENT STUDENTS 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 Floriade | Undergraduate studies | Higher Degree Research

Australia university

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

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 | Depart: Departure City | Sat

Must travel on these dates | Adults: Children (2-11): Infants: 1 0 0

KrisFlyer No. | 6-Digit PIN | Log In | KrisFlyer | Book Now | Singapore - Bangkok SGD 395* | Book Now | Singapore - Hong Kong SGD 546* | Book Now | Singapore - Taipei SGD 768* | Book Now | Singapore - Tokyo (Haneda) SGD 983* | Book Now | Singapore - Sydney | Book Now | Singapore - London | Book Now

Singapore airline

National University of Singapore

myEMAIL IVLE LIBRARY MAPS CALENDAR SITEMAP CONTACT e-CARDS

Search search for... in NUS Websites Go

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

A Leading Global University

Game Arrival Ceremony

Joint Evacuation Exercises

PROSPECTIVE STUDENTS CURRENT STUDENTS STAFF ALUMNI VISITORS

Singapore university

I·J·M·E·S·
restaurants • bars • shops

Discover a century of resplendent living history behind the cloisters

Chijmes, a premier lifestyle destination in Singapore

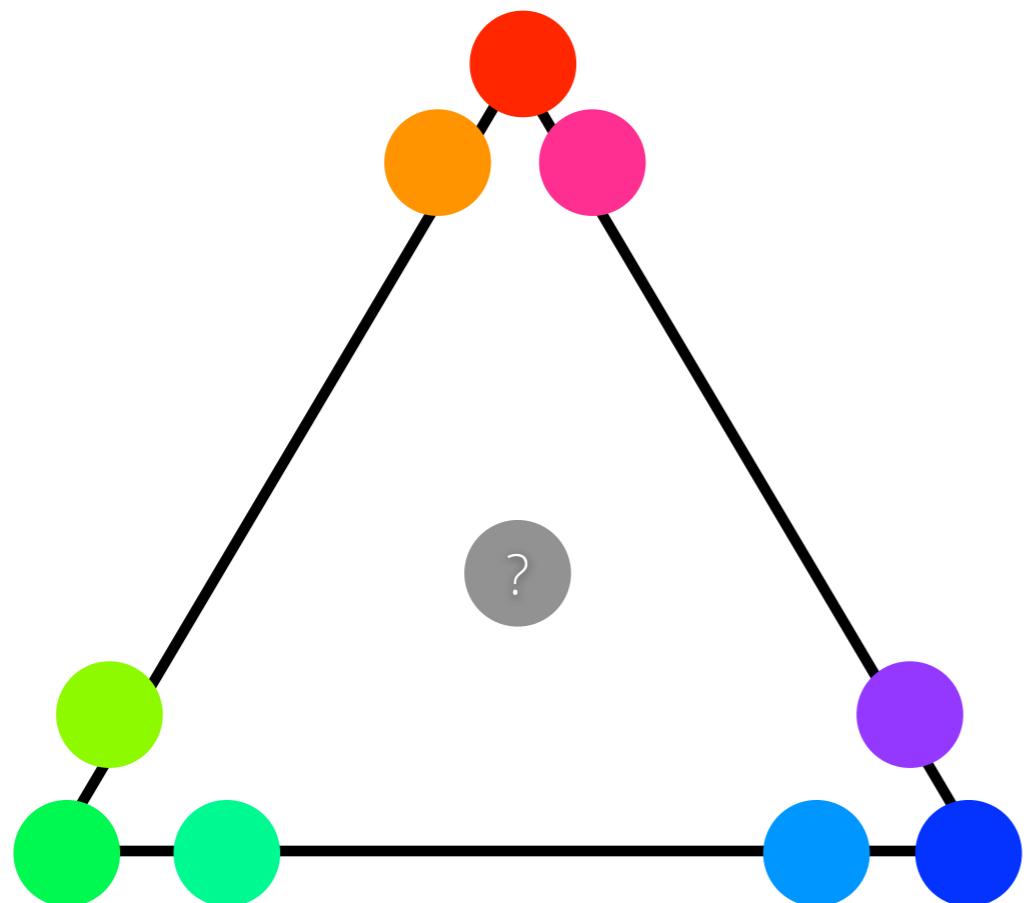
Owned by: SUNTEC Real Estate Investment Trust
Managed by: ARA APAC Investment Management Pte Ltd
Property Manager:

Singapore food



Clustering & Topic Models

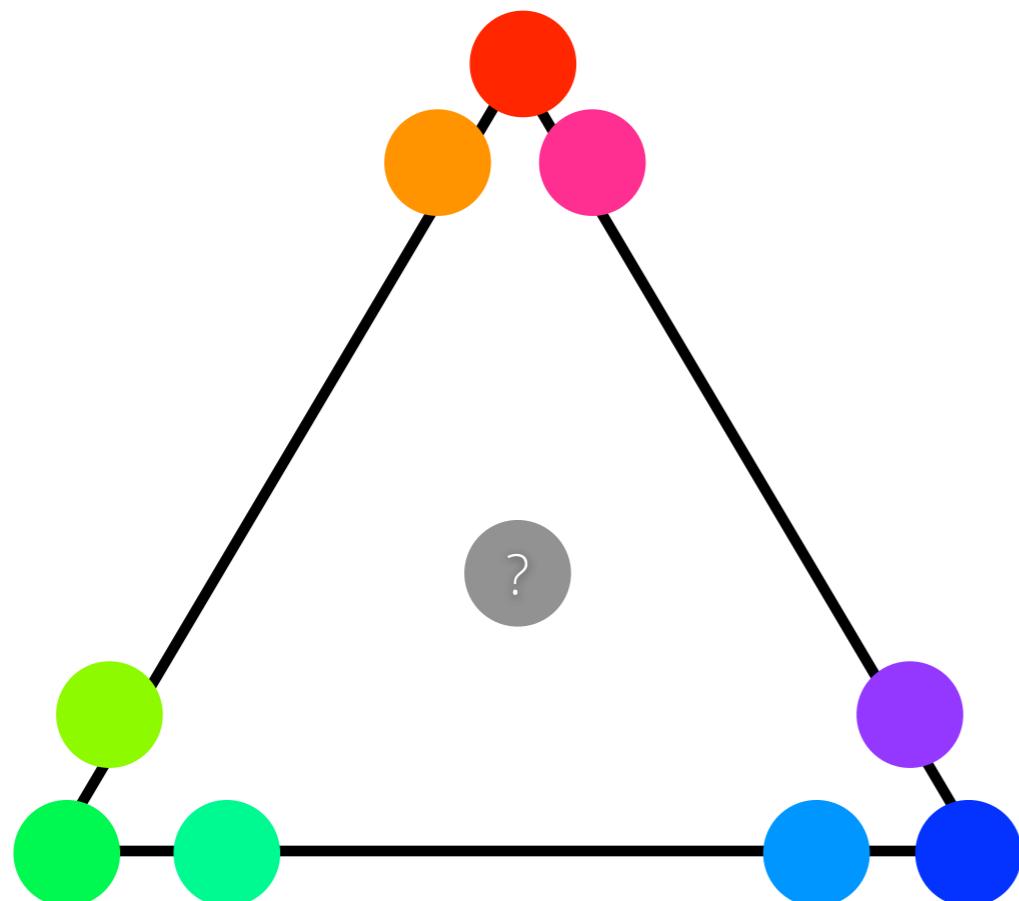
Clustering



group objects
by prototypes

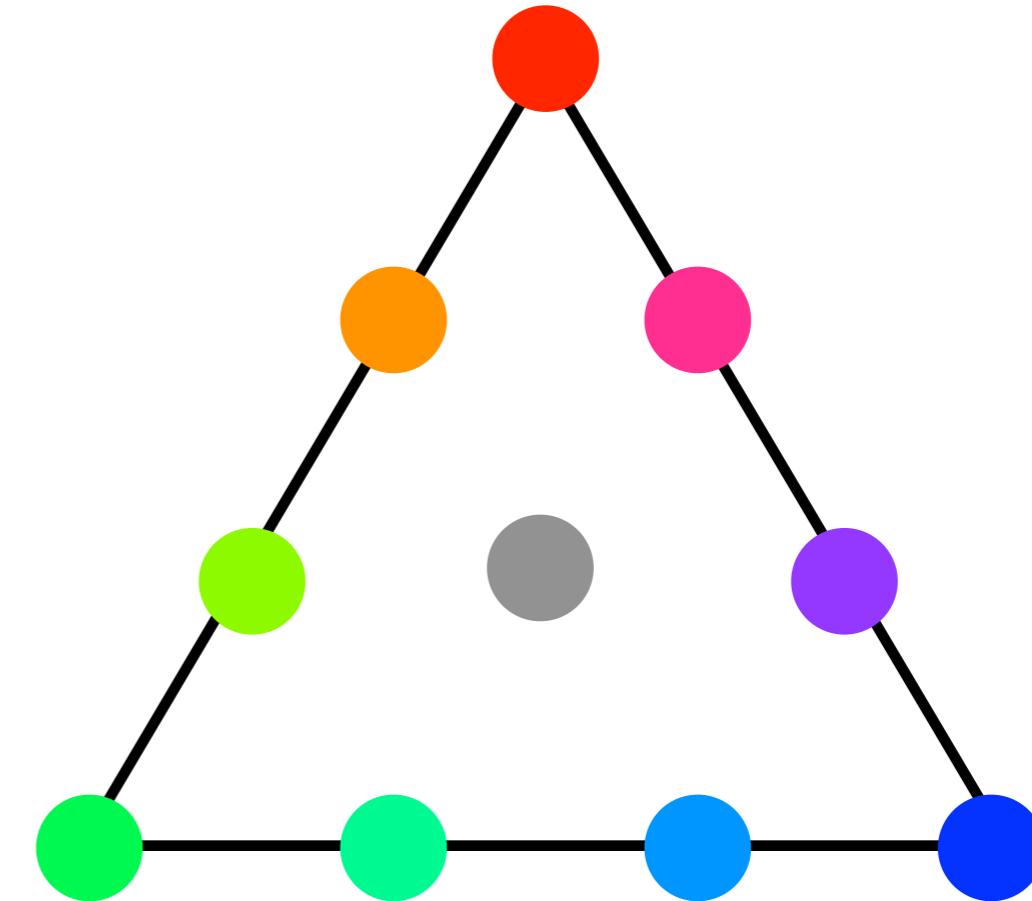
Clustering & Topic Models

Clustering



group objects
by prototypes

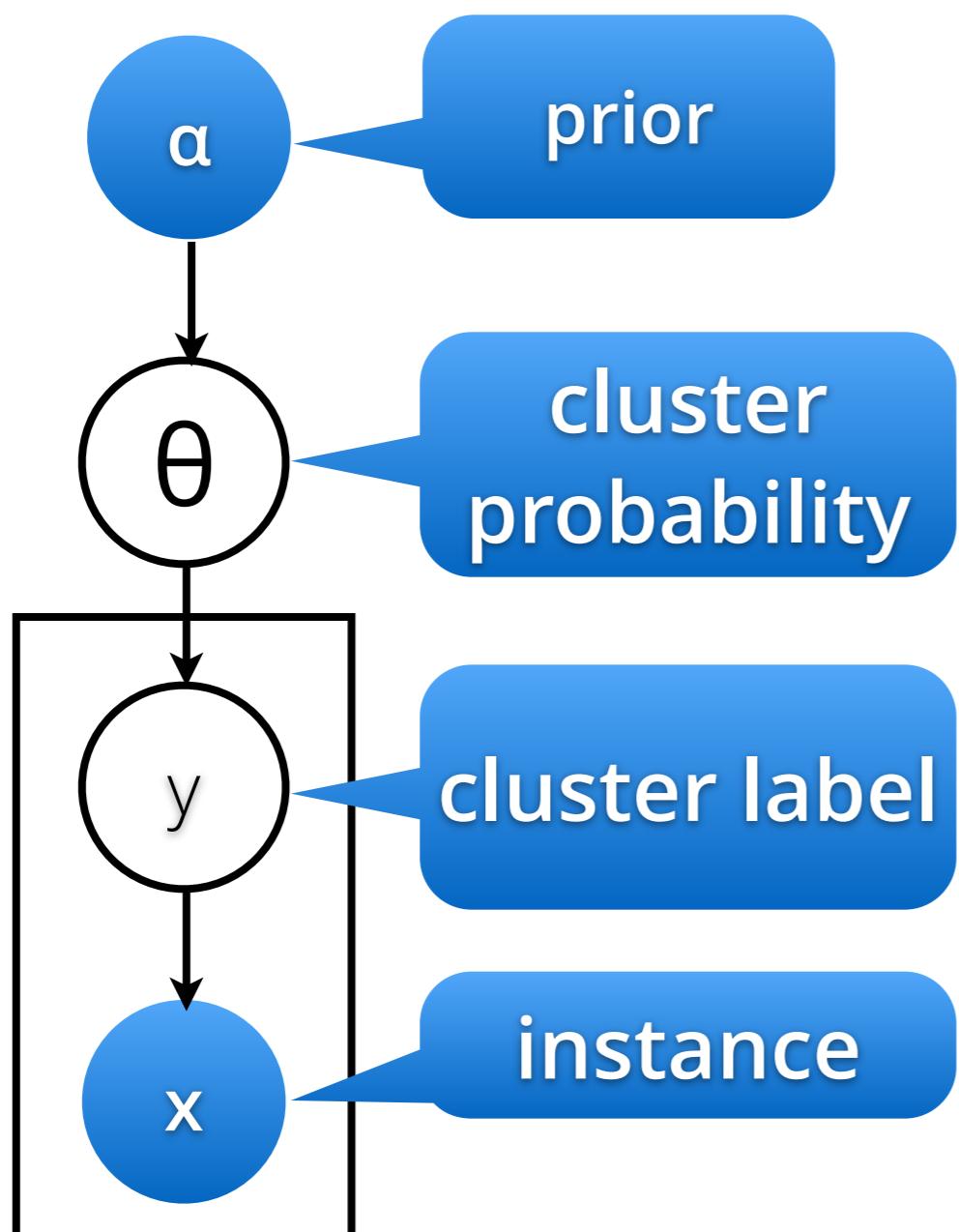
Topics



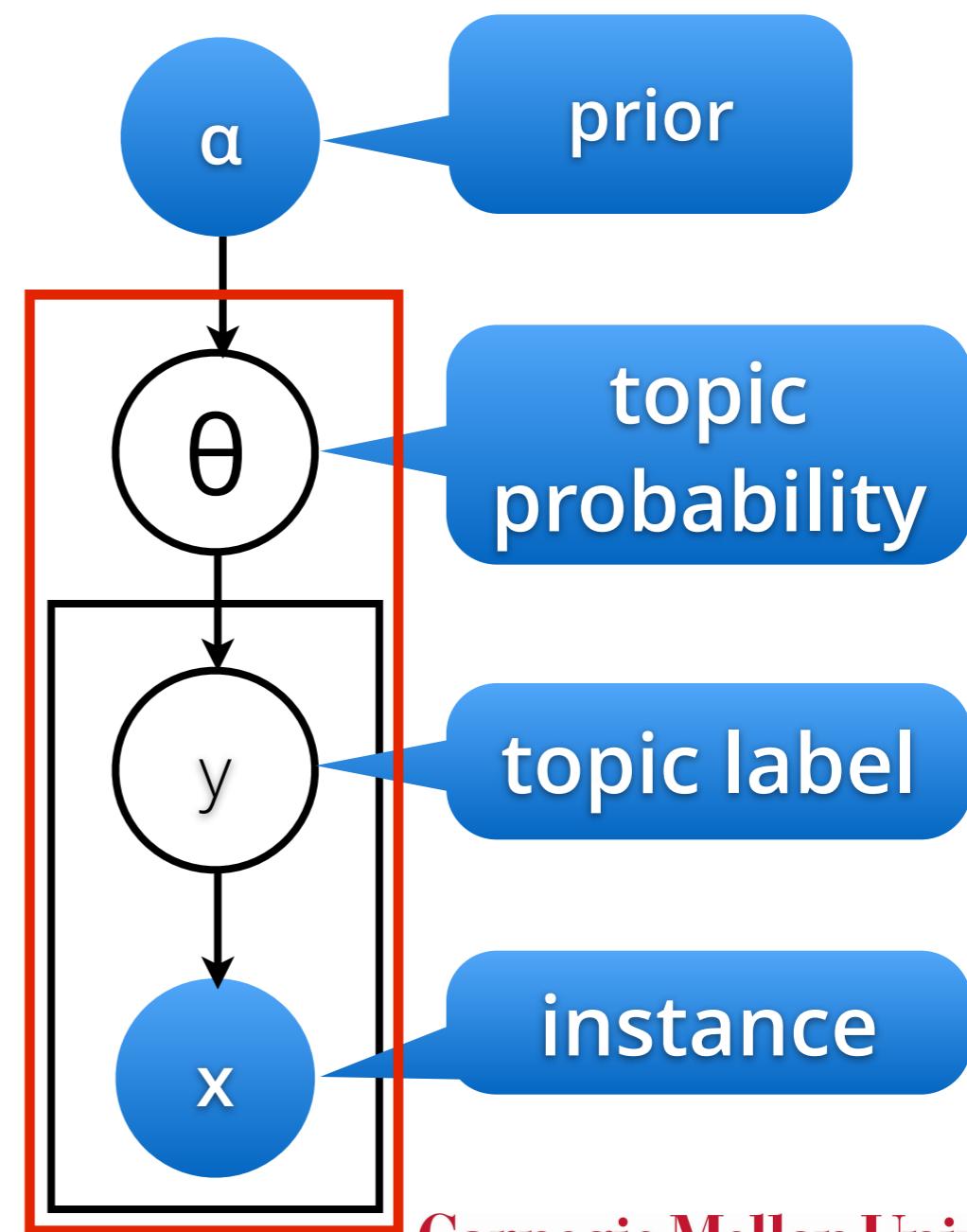
decompose objects
into prototypes

Clustering & Topic Models

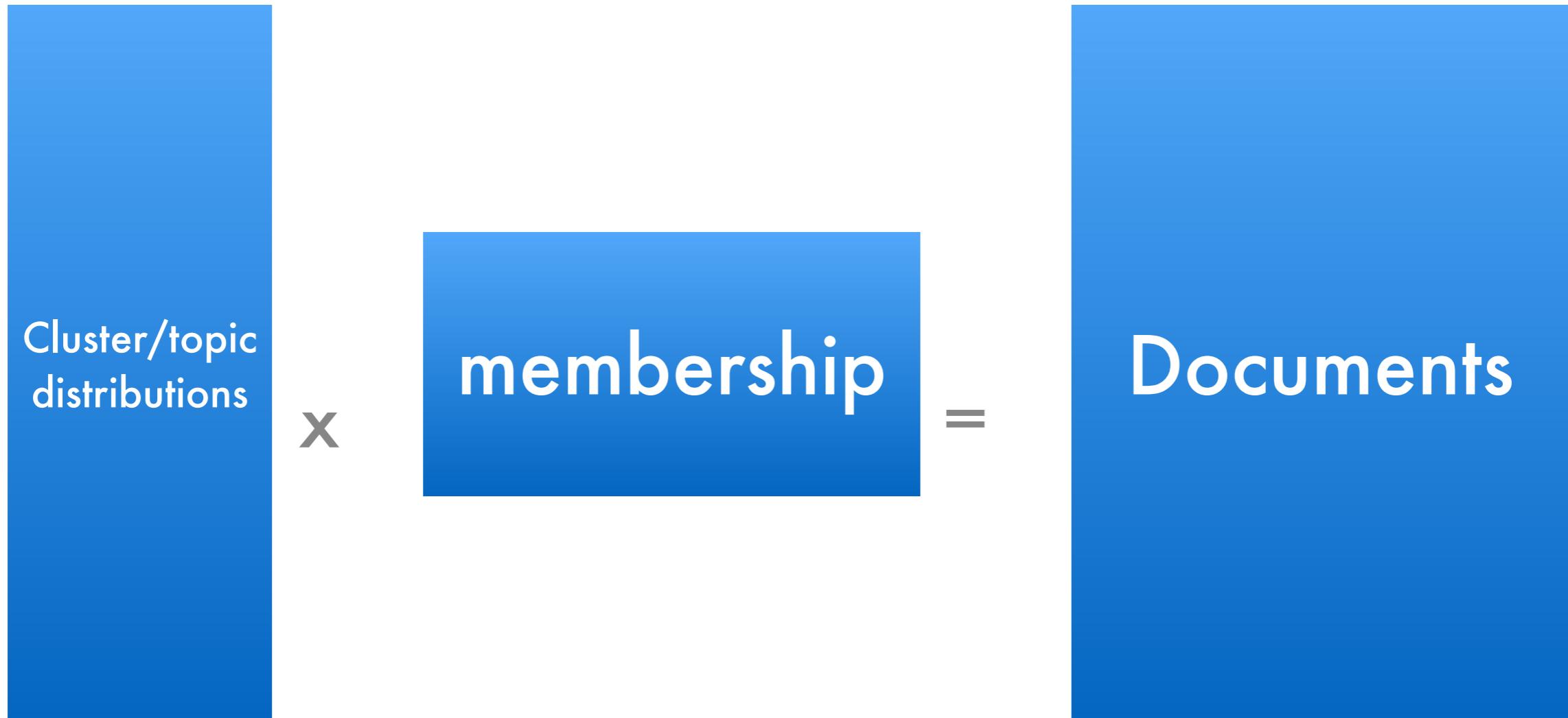
clustering



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

Gibbs Sampling

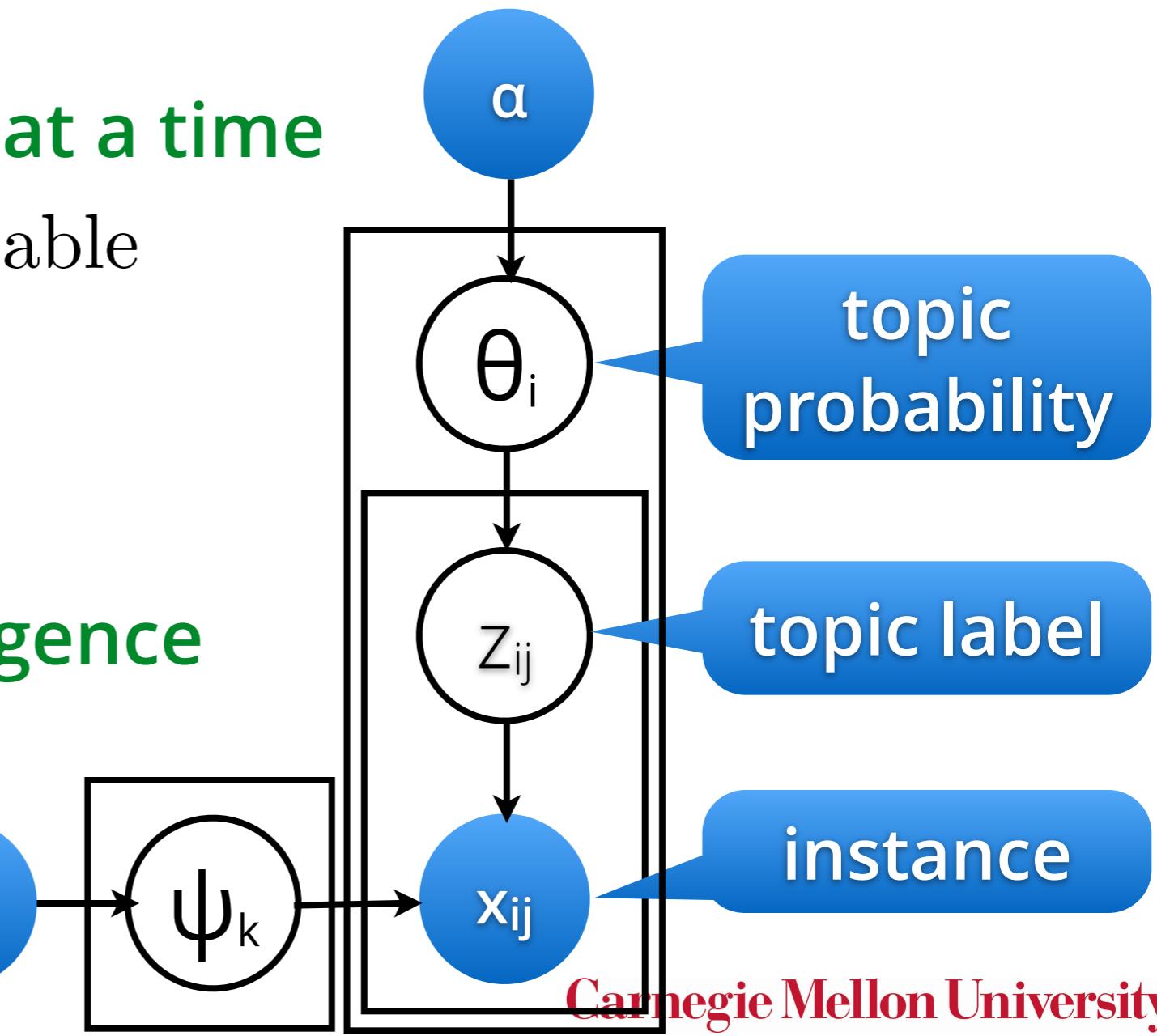
- Goal - sample topics and language model
- Problem - joint distribution intractable
- Solution
 - Sample one variable at a time

$(x, y) \sim p(x, y)$ intractable

$x \sim p(x|y)$

$y \sim p(y|x)$

- Guarantee of convergence



language prior

β

ψ_k

x_{ij}

Joint Probability Distribution

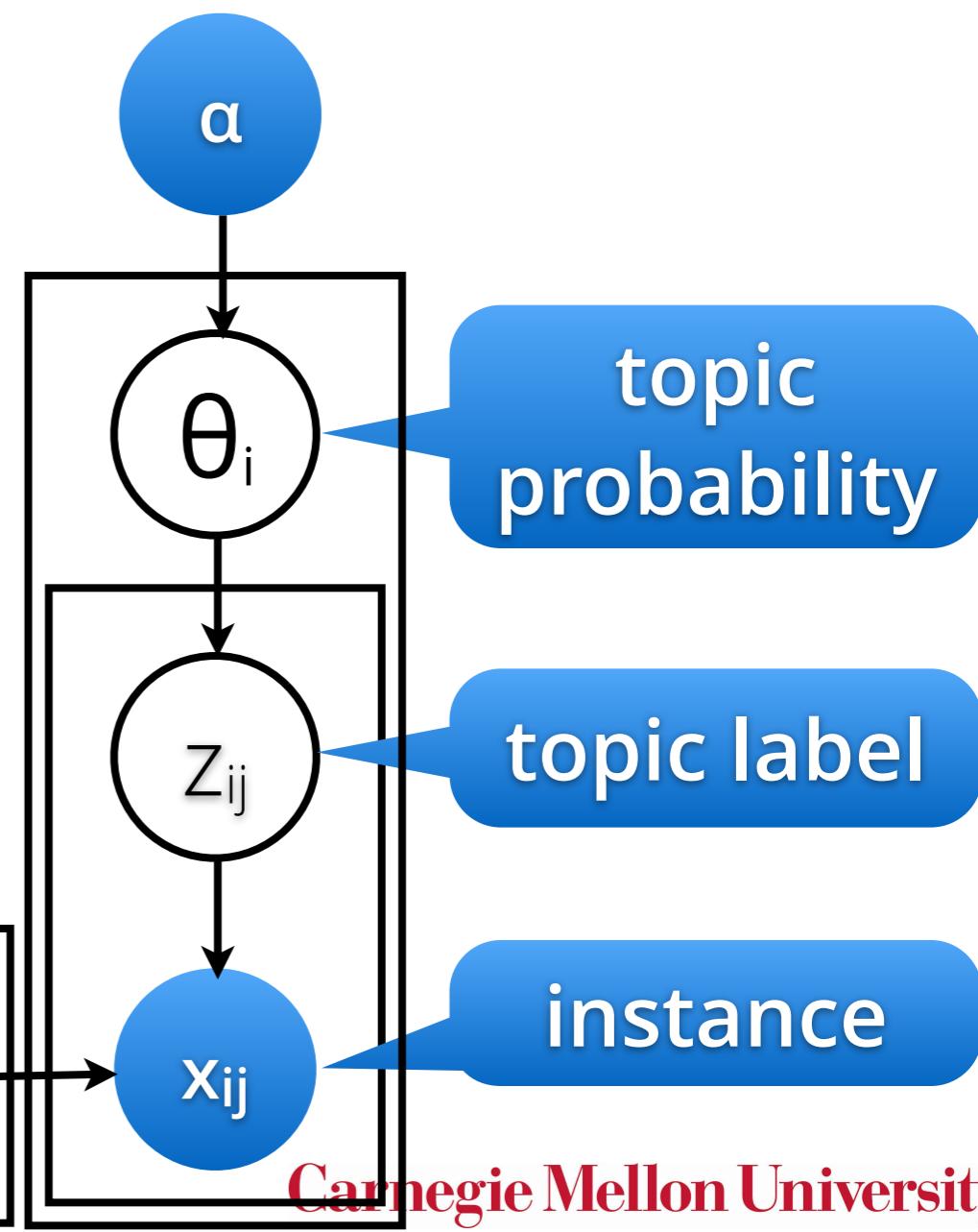
$$p(\theta, z, \psi, x | \alpha, \beta)$$

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

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

language prior

$$\beta \rightarrow \psi_k$$



Joint Probability Distribution

sample Ψ
independently

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

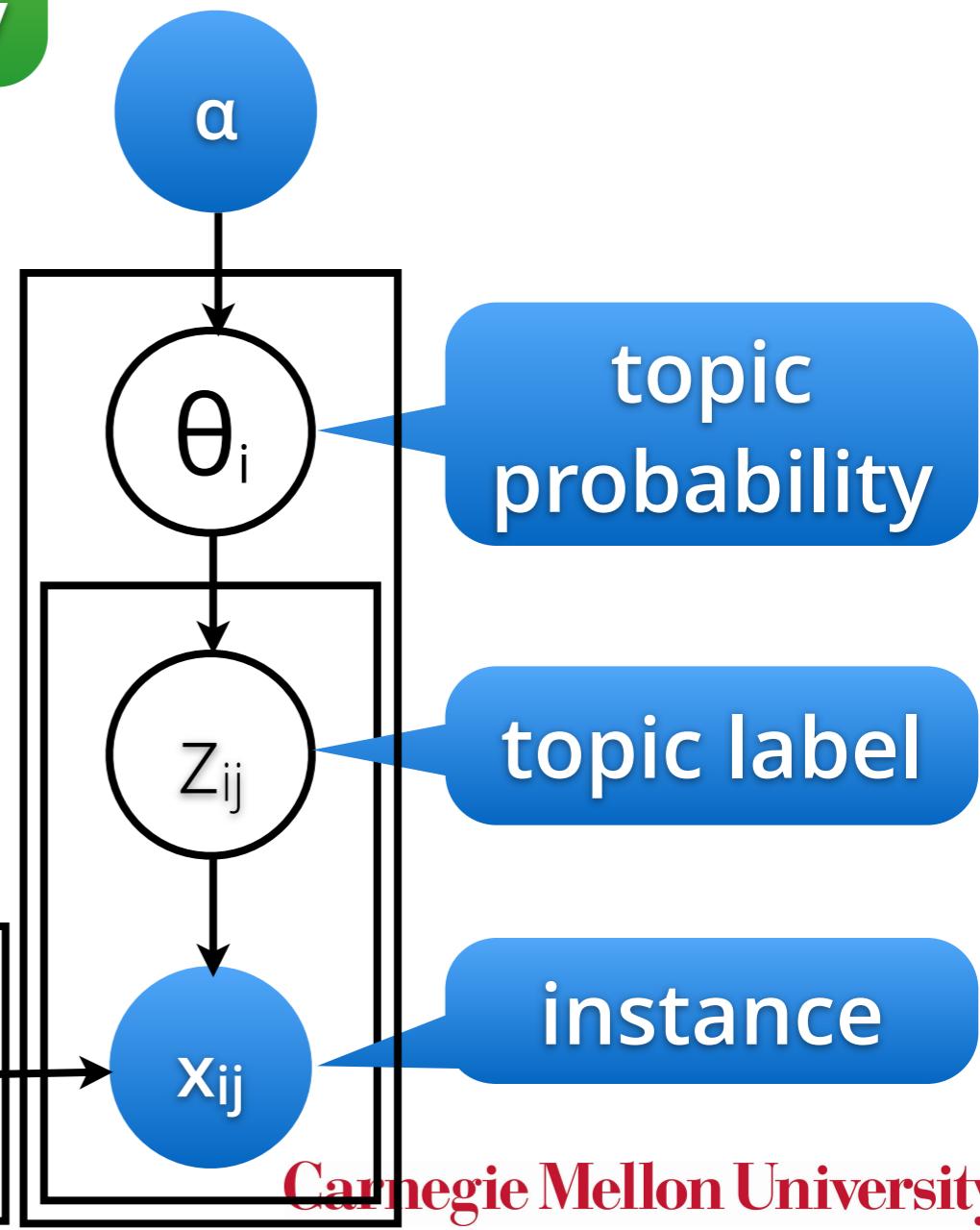
$$= \prod_{k=1}^K p(\psi_k | \beta) \prod_{i=1}^m p(\theta_i | \alpha)$$

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

sample z
independently

language prior

sample θ
independently



Joint Probability Distribution

sample Ψ
independently

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

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

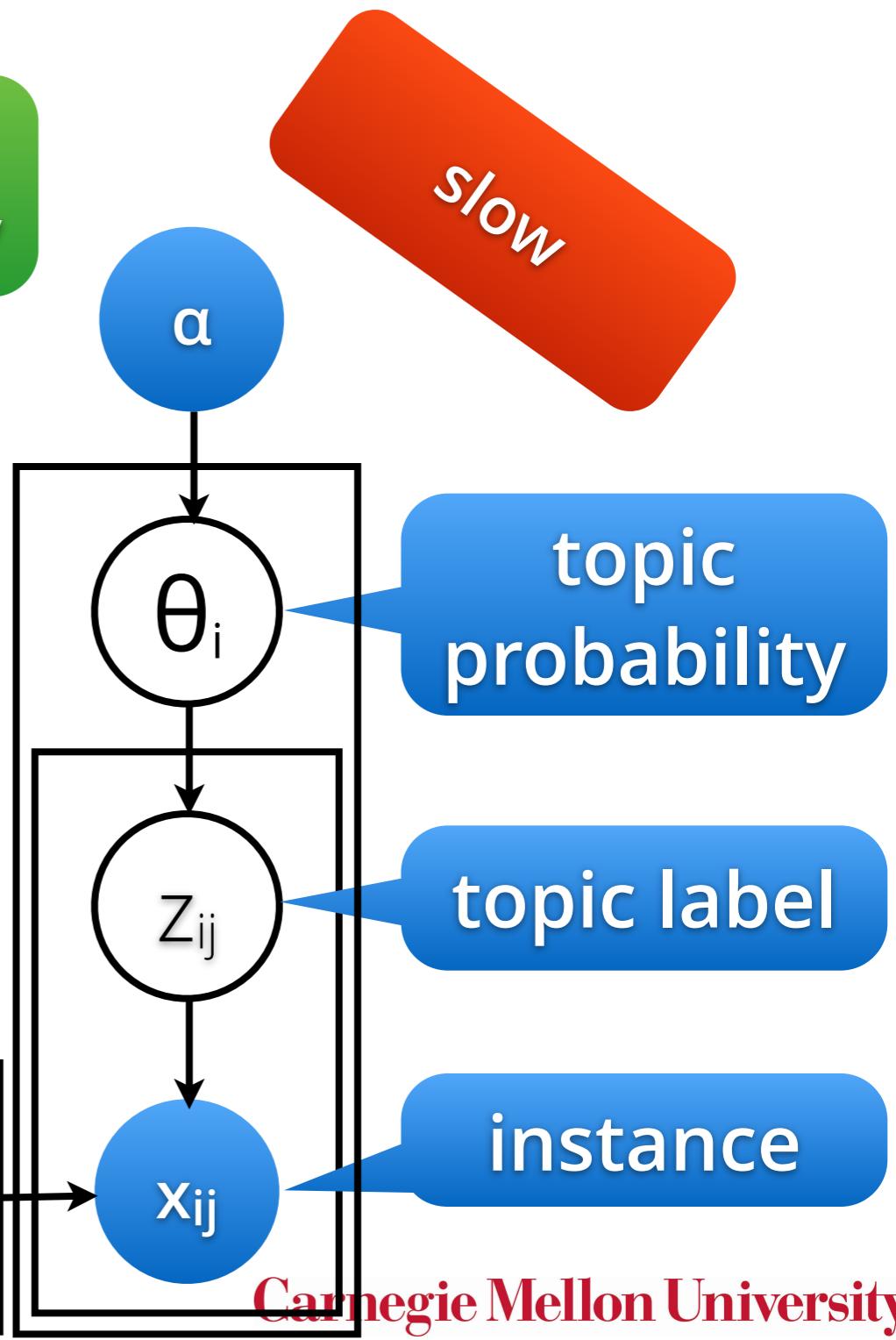
$$\prod_{i,j}^{m, m_i} p(z_{ij} | \theta_i) p(x_{ij} | z_{ij}, \psi)$$

sample z
independently

language prior

$$\beta \rightarrow \psi_k$$

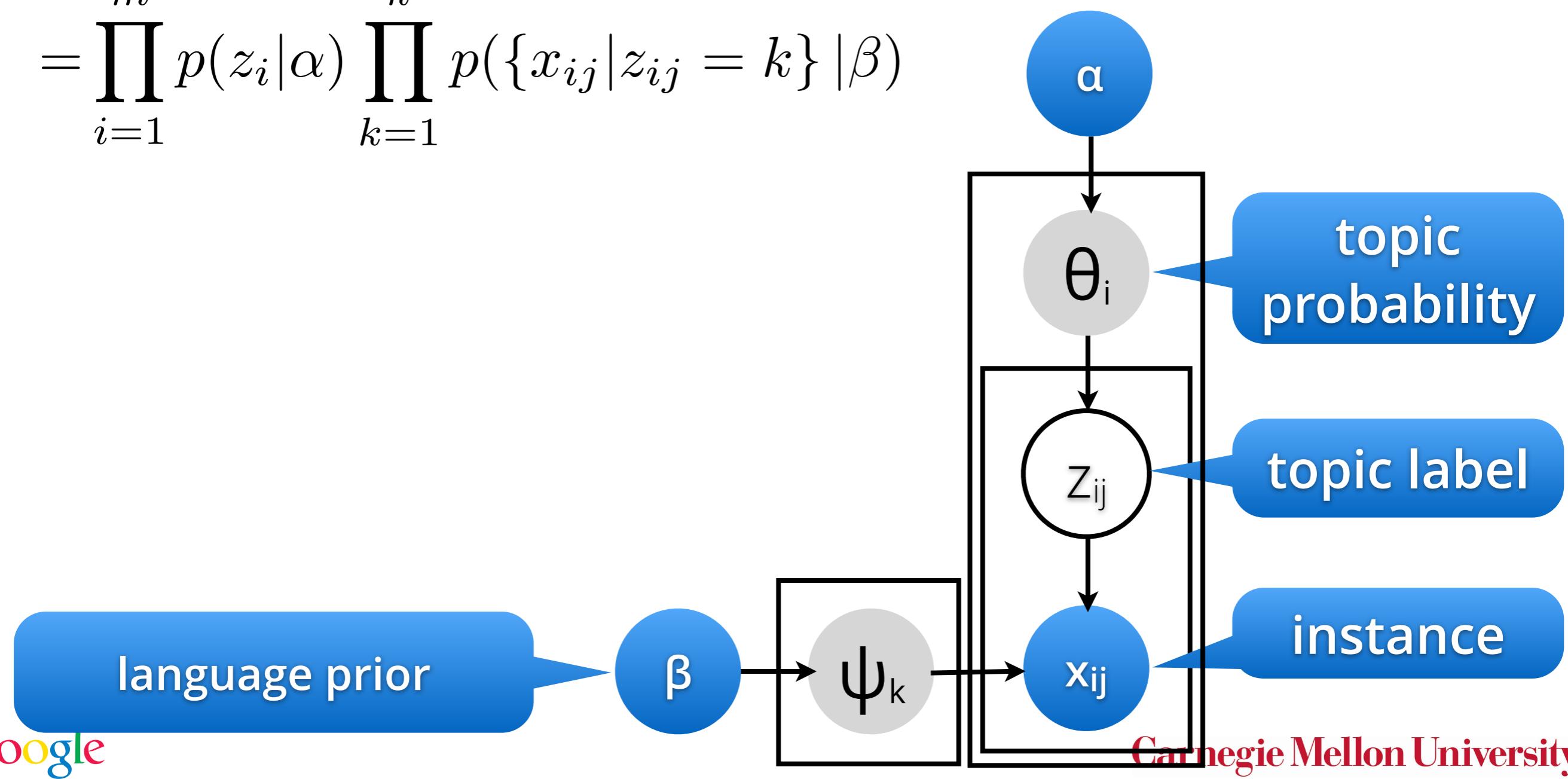
sample θ
independently



Collapsed Sampler

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

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



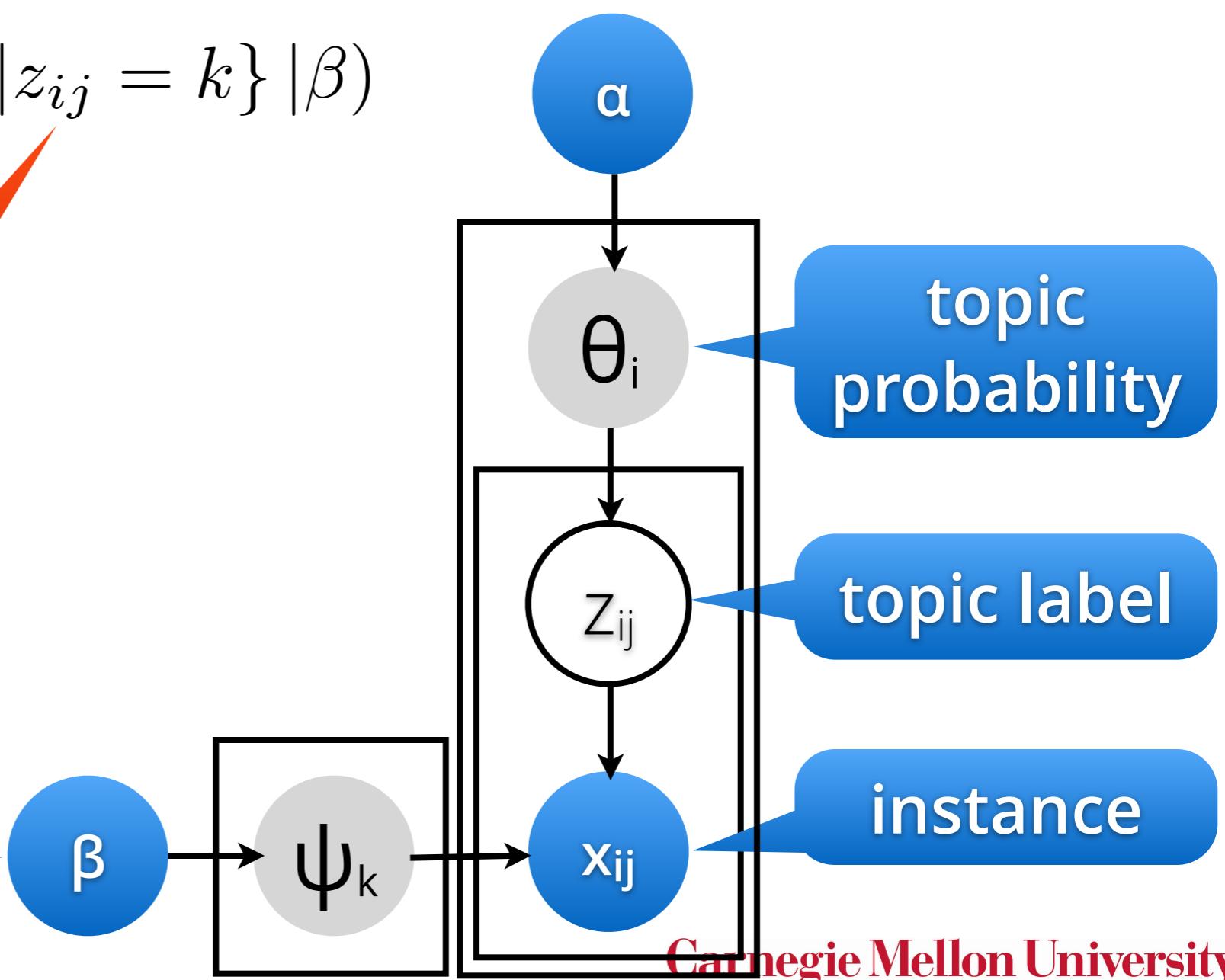
Collapsed Sampler

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

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

sample z sequentially

language prior

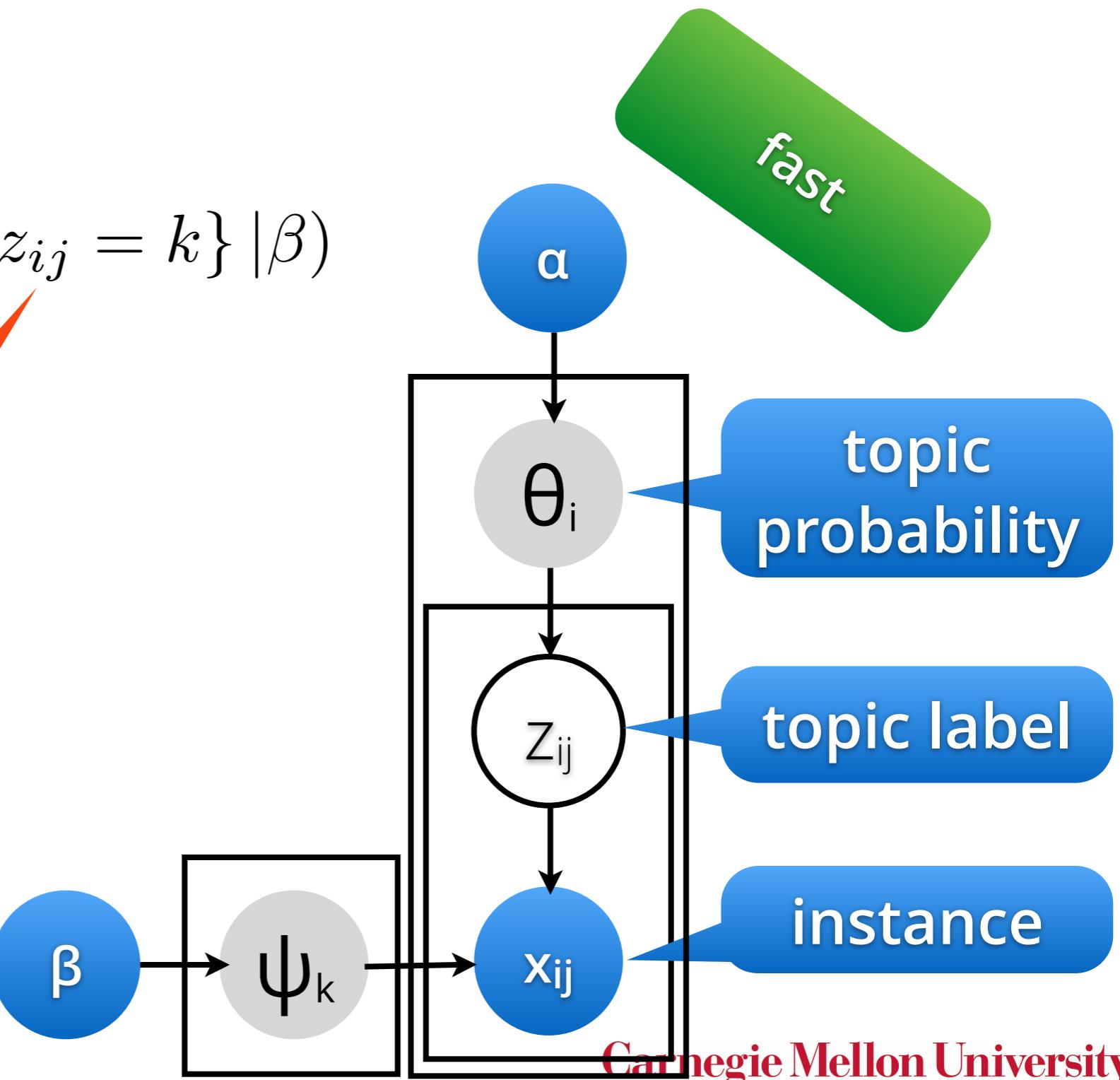


Collapsed Sampler

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

sample z sequentially

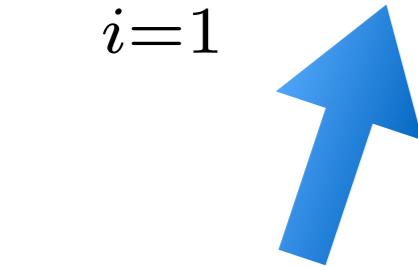
language prior



Collapsed Sampler

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

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



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

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

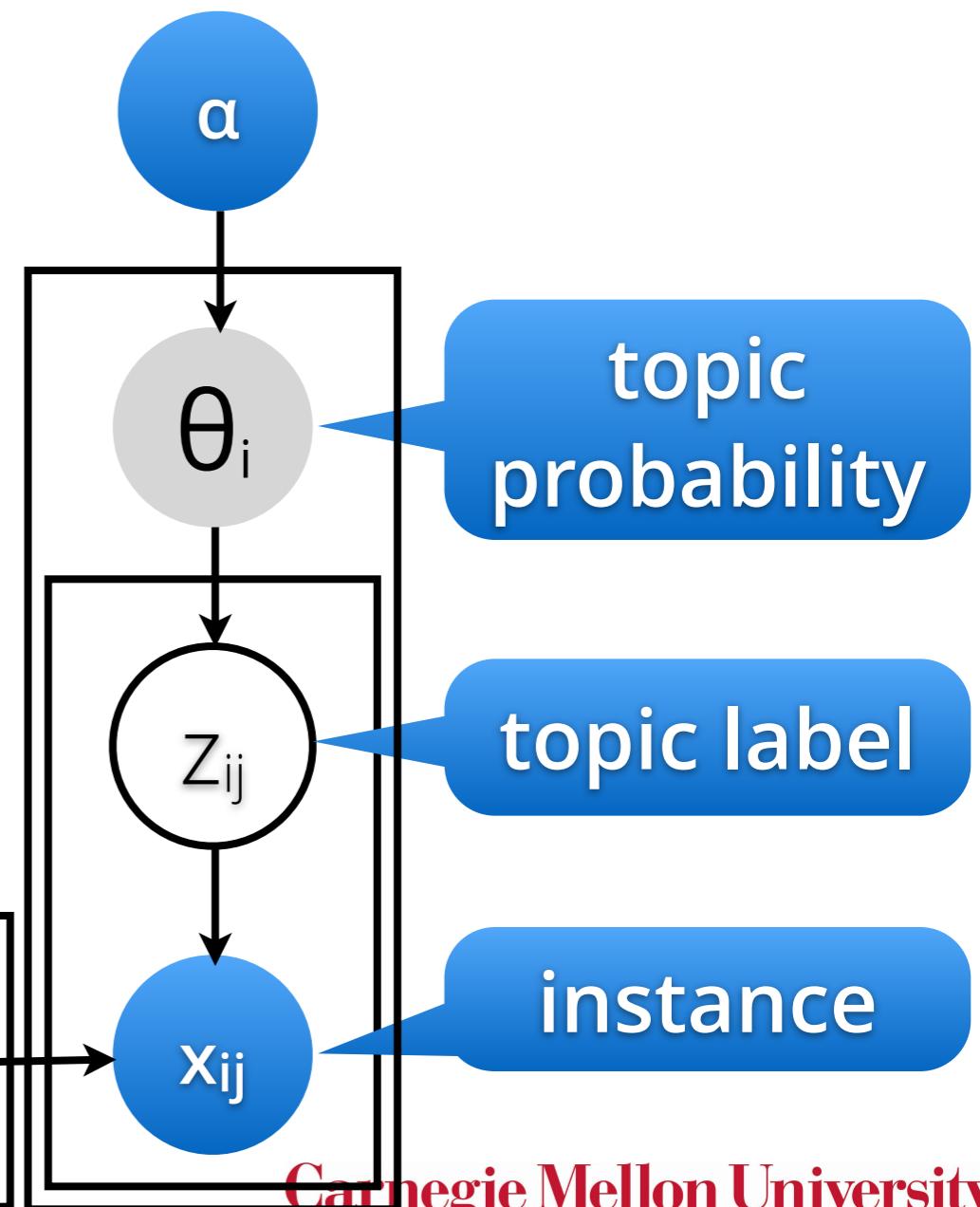
Griffiths & Steyvers, 2005

language prior

β

ψ_k

x_{ij}



Collapsed Sampler

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

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

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

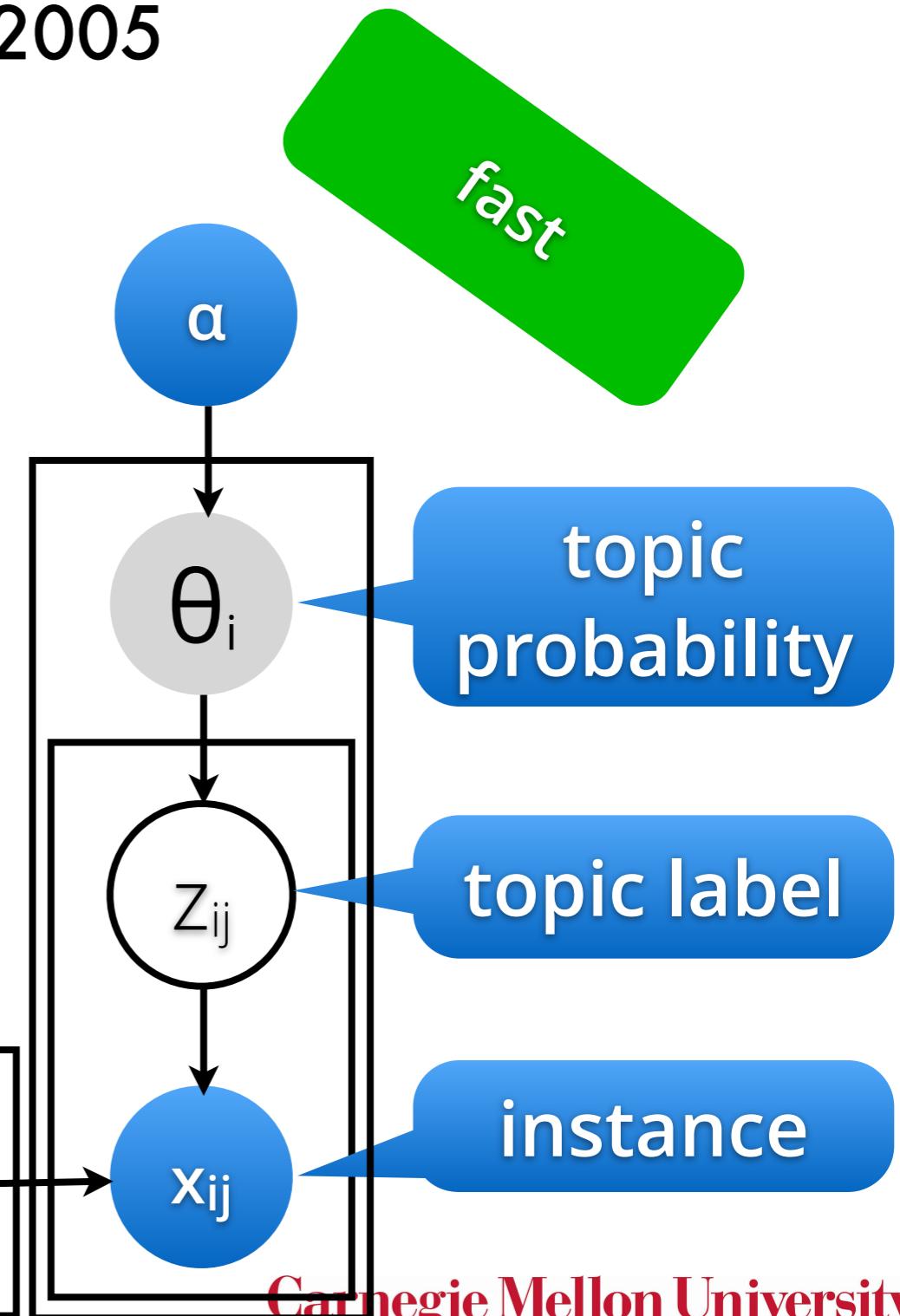
Griffiths & Steyvers, 2005

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

language prior

β

ψ_k



Gibbs Sampler

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



this kills parallelism

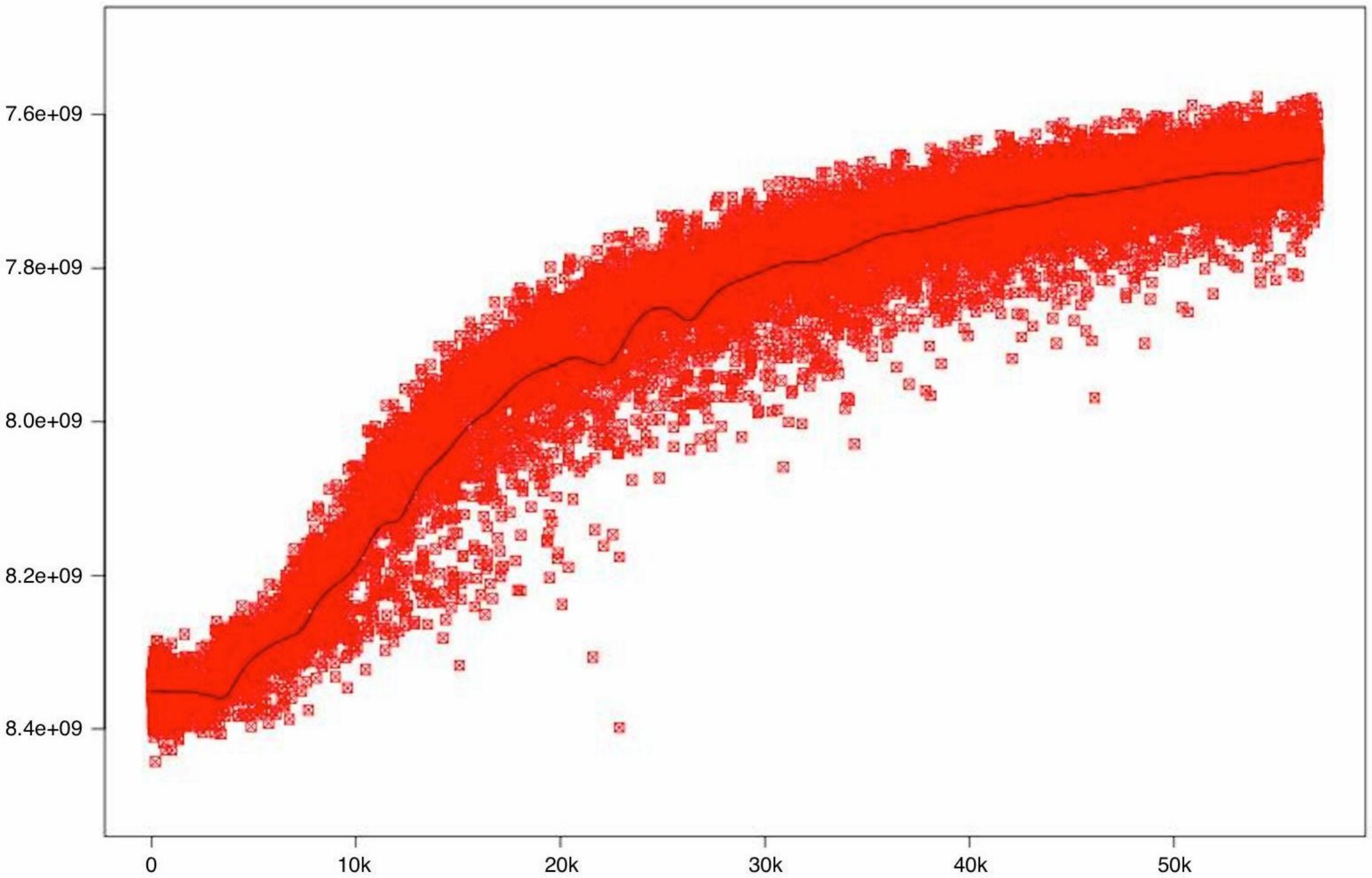
Gibbs Sampler

- For 1000 iterations do
 - For each document do
 - For each word in the document do
 - Resample topic for the word
 - Lock local (word,topic) table
 - Update local (document, topic) table
 - Update local (word,topic) table
 - Unlock local (word,topic) table
 - Synchronize local and global tables

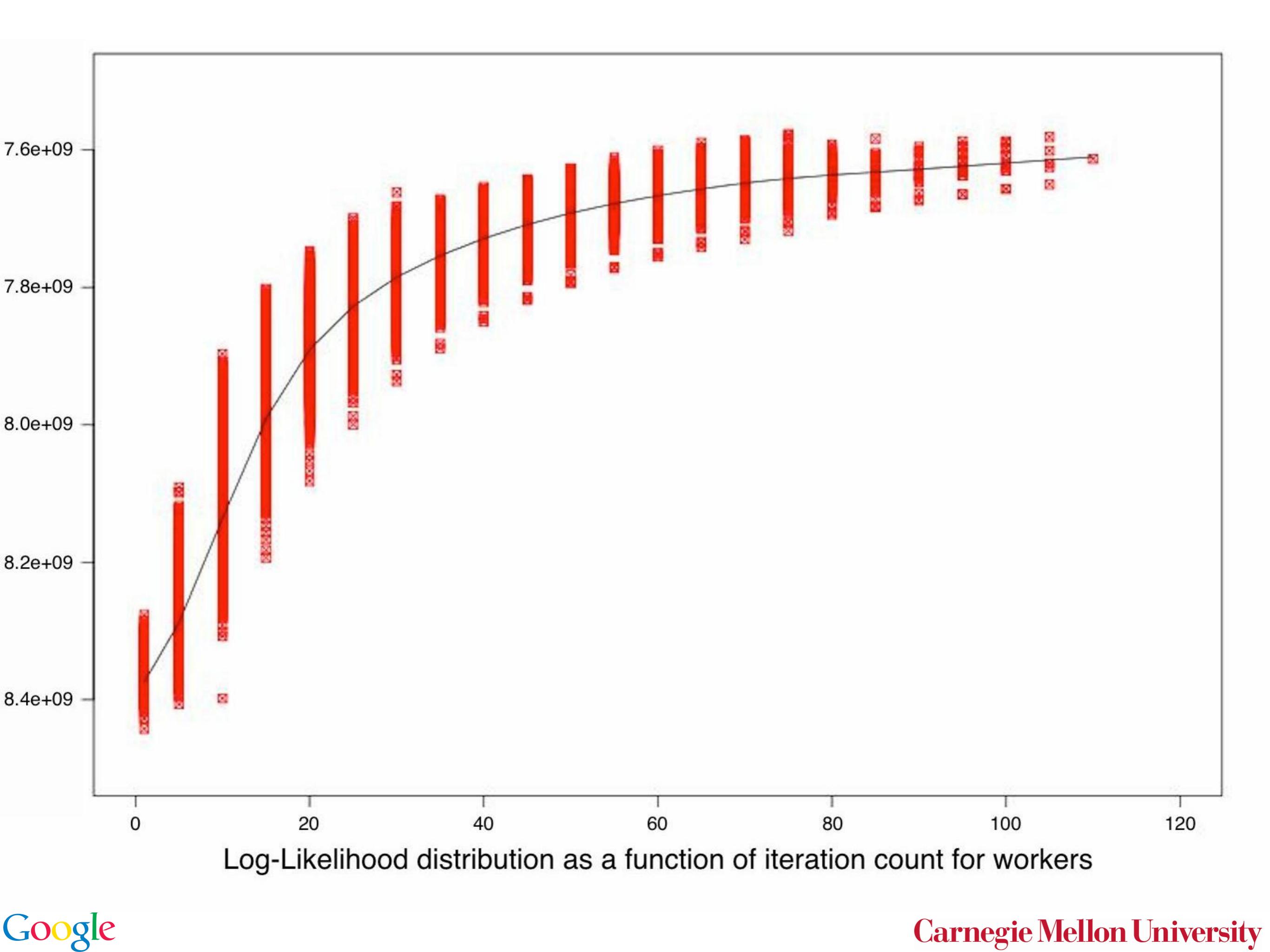
this kills multithreading

Gibbs Sampler

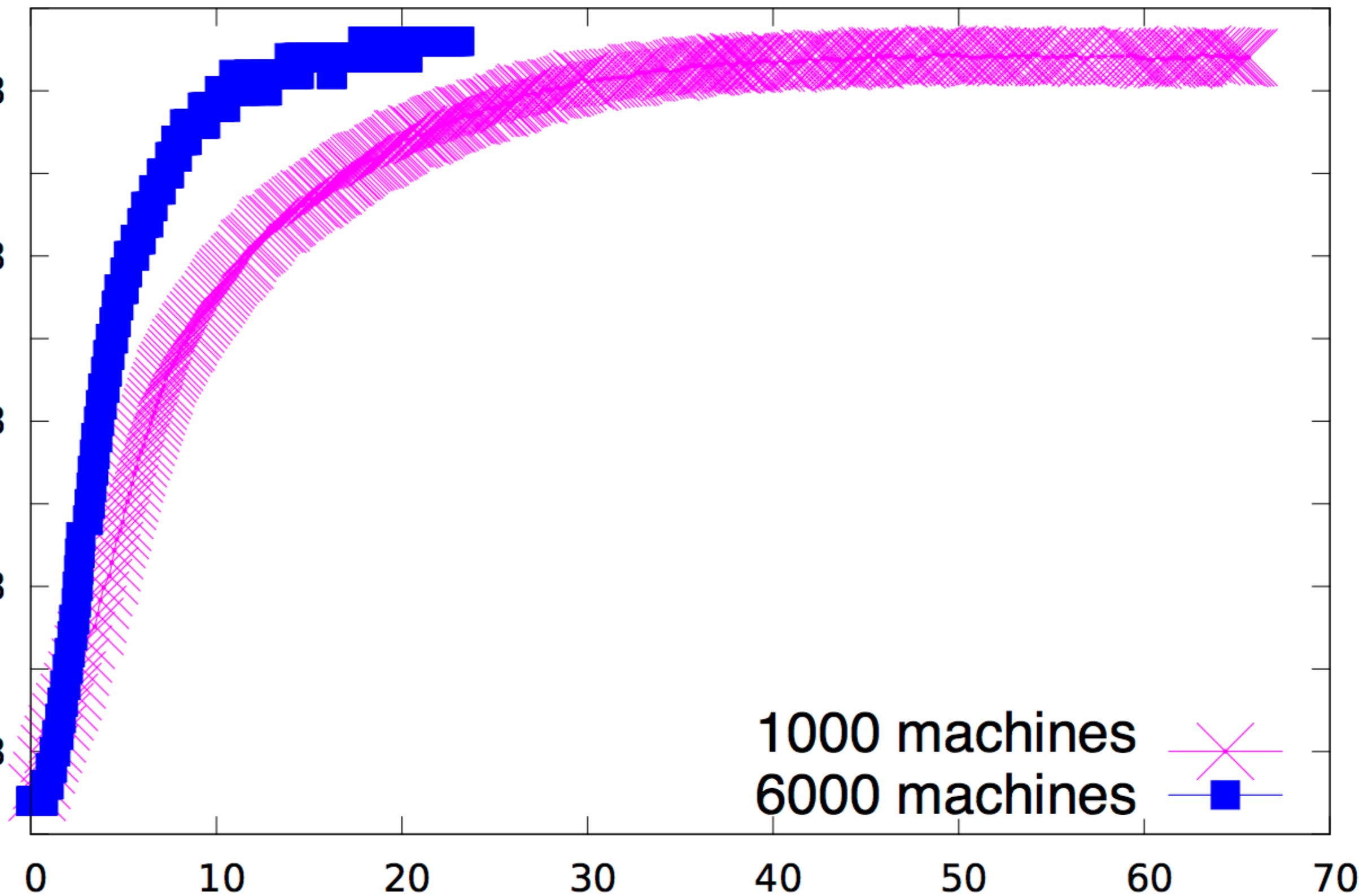
- For 1000 iterations do
 - For each document do
 - For each word in the document do
 - Resample topic for the word
 - Update local (document, topic) table
 - Generate local update message
 - Update local table
 - Lock local (word,topic) table
 - Update local (word,topic) table
 - Unlock local (word,topic) table
 - Synchronize local and global tables



Log-Likelihood distribution as a function of runtime (s) for workers



Log-Likelihood distribution as a function of iteration count for workers



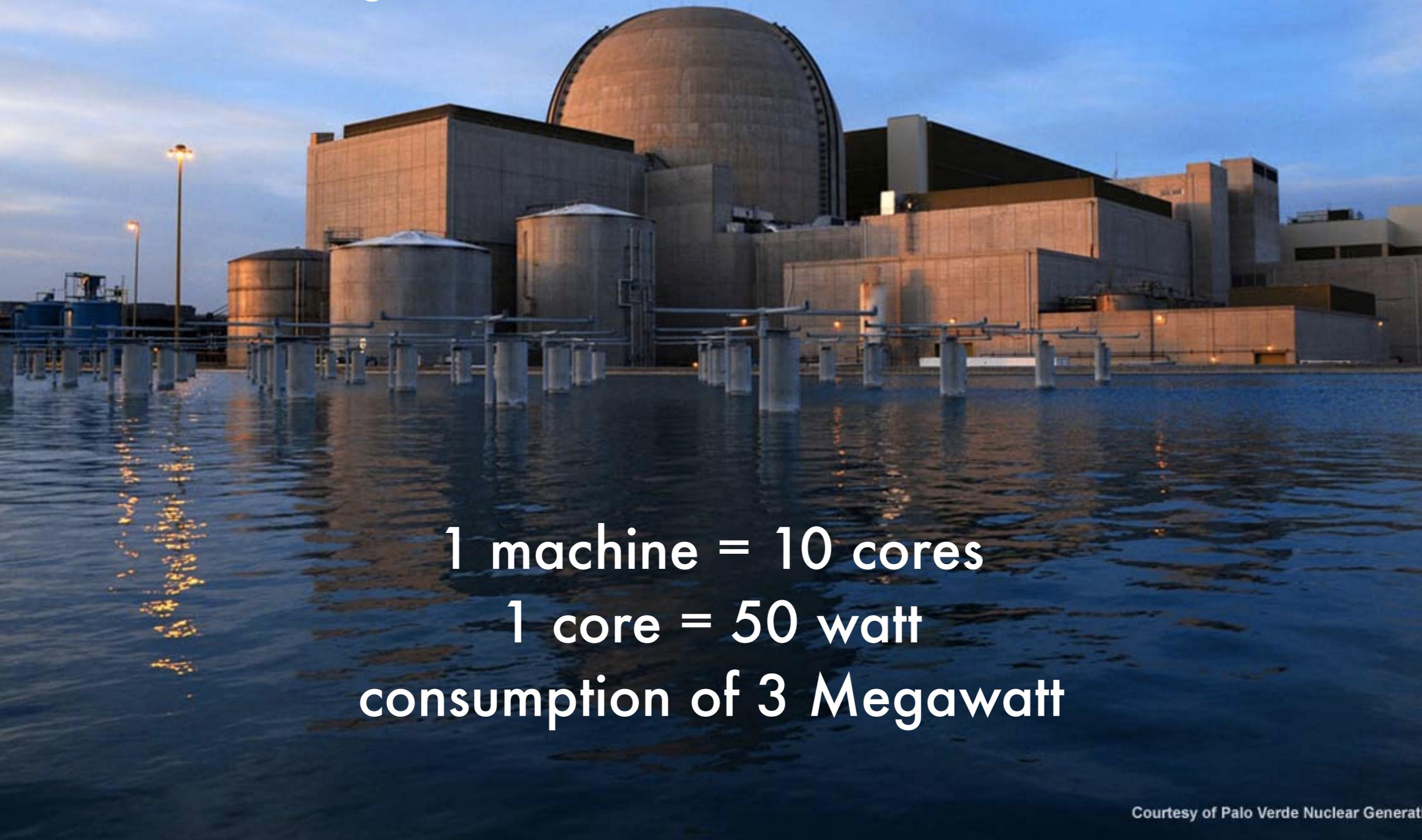
Palo Verde, AZ
3 Gigawatt
Largest nuclear reactor in the USA



Palo Verde, AZ

3 Gigawatt

Largest nuclear reactor in the USA



1 machine = 10 cores

1 core = 50 watt

consumption of 3 Megawatt

User Profiling



Mike

Techie

25-34 single male
living with friends

Mike works as a graphic designer in a small agency and one day wants to run his own agency.

He's got an iPhone and a Vodafone 360 H1 by Samsung phone, one personal and one for work. He follows friends and key people in the design industry via Twitter, blogs, and RSS feeds. He uses his iPhone for work emails and his H1 for Facebook.

He uses Twitter to post updates about what he's up to with his project work as well as using it as a tool to find out what people are up to and to invite them to events. He uses Facebook to share personal photos and video and keeps a Tumblr blog to post interesting things he discovers and share them with his friends and followers.



Zoë

Socialite

18-33 single female
living with friends

Zoë is studying a Masters in International Development unsure of what the future lies ahead of her.

She is constantly using the Facebook app on her Vodafone 360 M1 by Samsung phone as well as on her PC to upload and tag photos and videos from places she's been to with her friends, as well as to find out and comment on who's been where at which club nights and parties.

She regularly texts and messages her friends to find out if they've heard about a new pop-up shop she heard about via a flyer, or one-off warehouse party started by friends of friends.



Geoff

Cost-conscious

35-49 married male
with young kids

Geoff works as a senior architect in a large practice, and has a wife and a young girl and 6-month baby boy. He thinks the time is right to start looking for a bigger home for his family.

Geoff uses his Vodafone 360 H1 to take photos and videos of prospective sites he visits. He purchased the H1 because of its ability to check email, surf the web, use apps, and take photos and video.

He loves the built-in camera and also uses this phone on holiday to take snaps of the family as it fits in his pocket and doesn't want to carry a large SLR around with him. He likes to upload his photos and video to Flickr and share them with his family and friends. He also creates photo books from his holidays snaps to give as gifts to his parents.

User Profiling



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

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

Panasonic LX5 Cheap
Best Value for **Panasonic LX5**. Find NexTag Sellers
[www.NexTag.com](#)

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

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

Panasonic Lumix DMC-LX5 White Digital Camera
shopping.yahoo.com
The Panasonic Lumix DMC-LX5 is a compact digital photo enthusiast's ideal way for capturing professional photos and High Definition video.

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

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

Sponsored Results

amazon.com Prime

Shop All Departments Search Electronics

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

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

Color: Black

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
Color: Black

time

Buying a Camera

YAHOO! Web Images Video Local Shopping News More ▾

panasonic lx5

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

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

Panasonic LX5 Cheap
Best Value for **Panasonic LX5**. Find NexTag Sellers
[www.NexTag.com](#)

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

Panasonic LX5 | Get The Lowest Price On
Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
[www.panasoniclx5.com](#) - [Cached](#) - [More from this s](#)

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**

Sponsor Results

show ads now

Specs

amazon.com Prime

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 Electronics

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

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

time



Buying a Camera

YAHOO! Web Images Video Local Shopping News More ▾

panasonic lx5

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

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

Panasonic LX5 Cheap
Best Value for **Panasonic LX5**. Find NexTag Sellers
[www.NexTag.com](#)

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

Panasonic LX5 | Get The Lowest Price On
Panasonic LX5 with 14.1MP captures enough detail.
Panasonic LX5 Camera
[www.panasoniclx5.com](#) - [Cached](#) - [More from this s](#)

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**

Sponsor Results

show ads now

Specs

amazon.com Prime

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

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

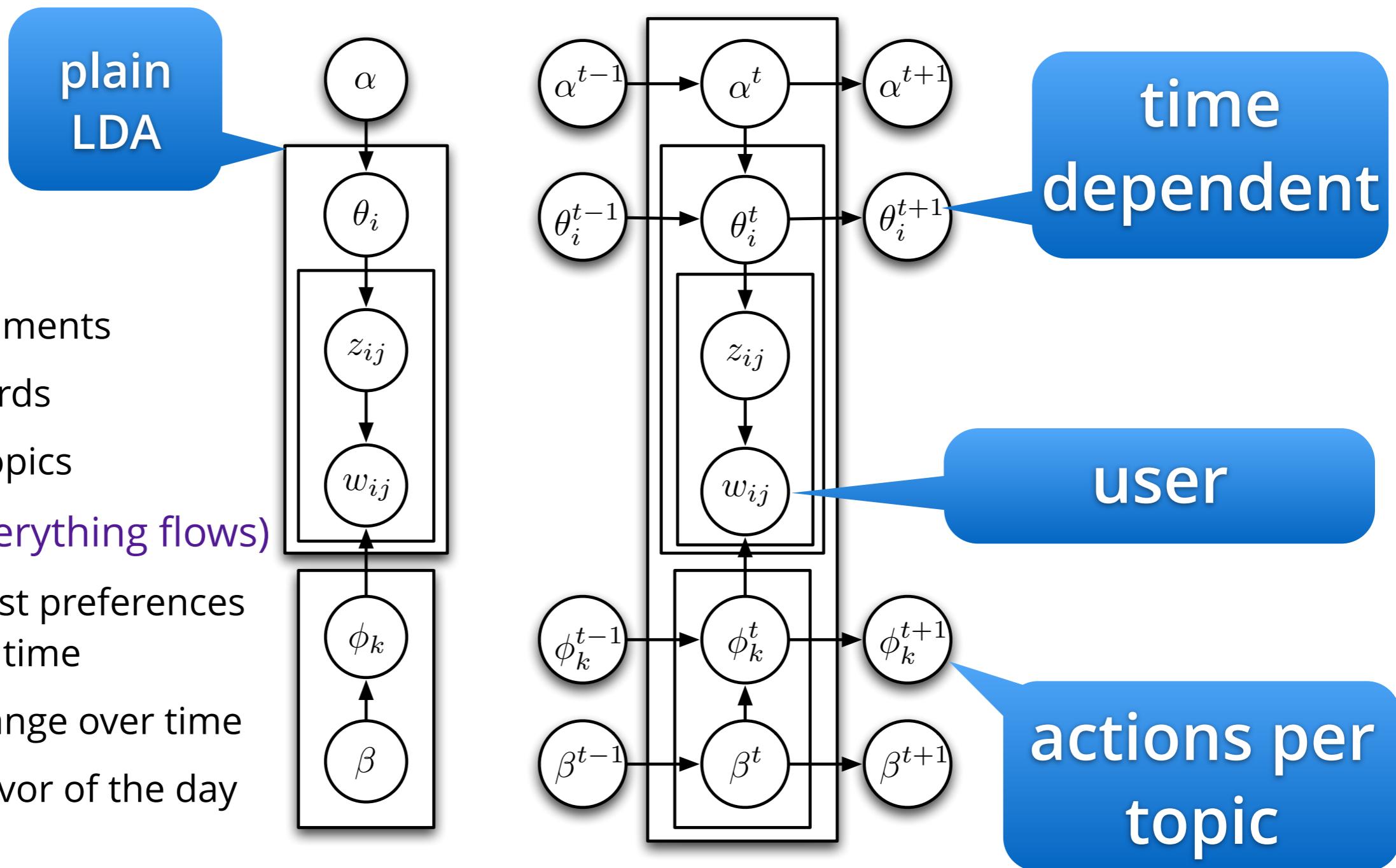
List Price: **\$499.00**
Price: **\$444.95** & eligible for free shipping with **Amazon Prime**
You Save: **\$54.05 (11%)**

too late

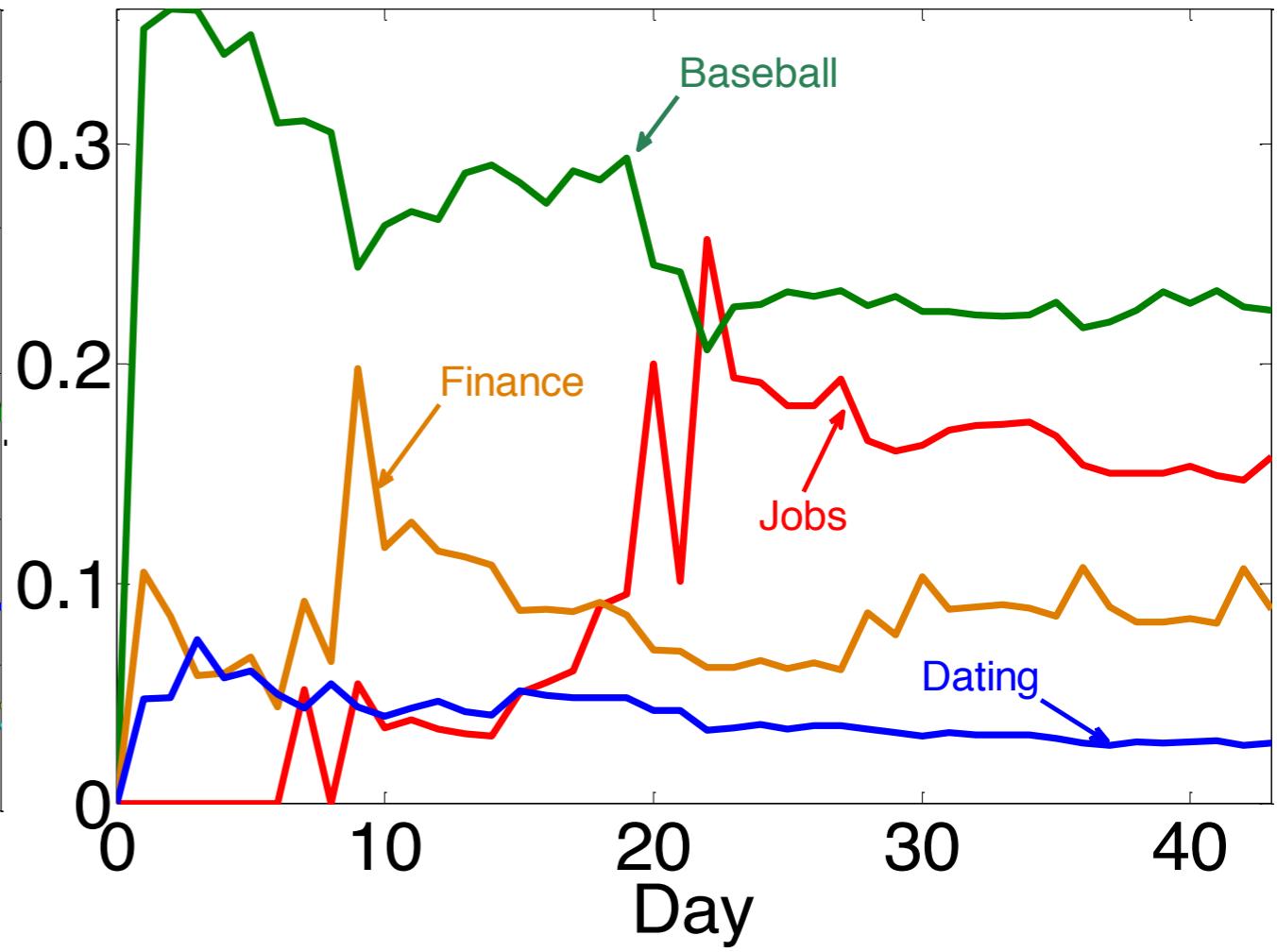
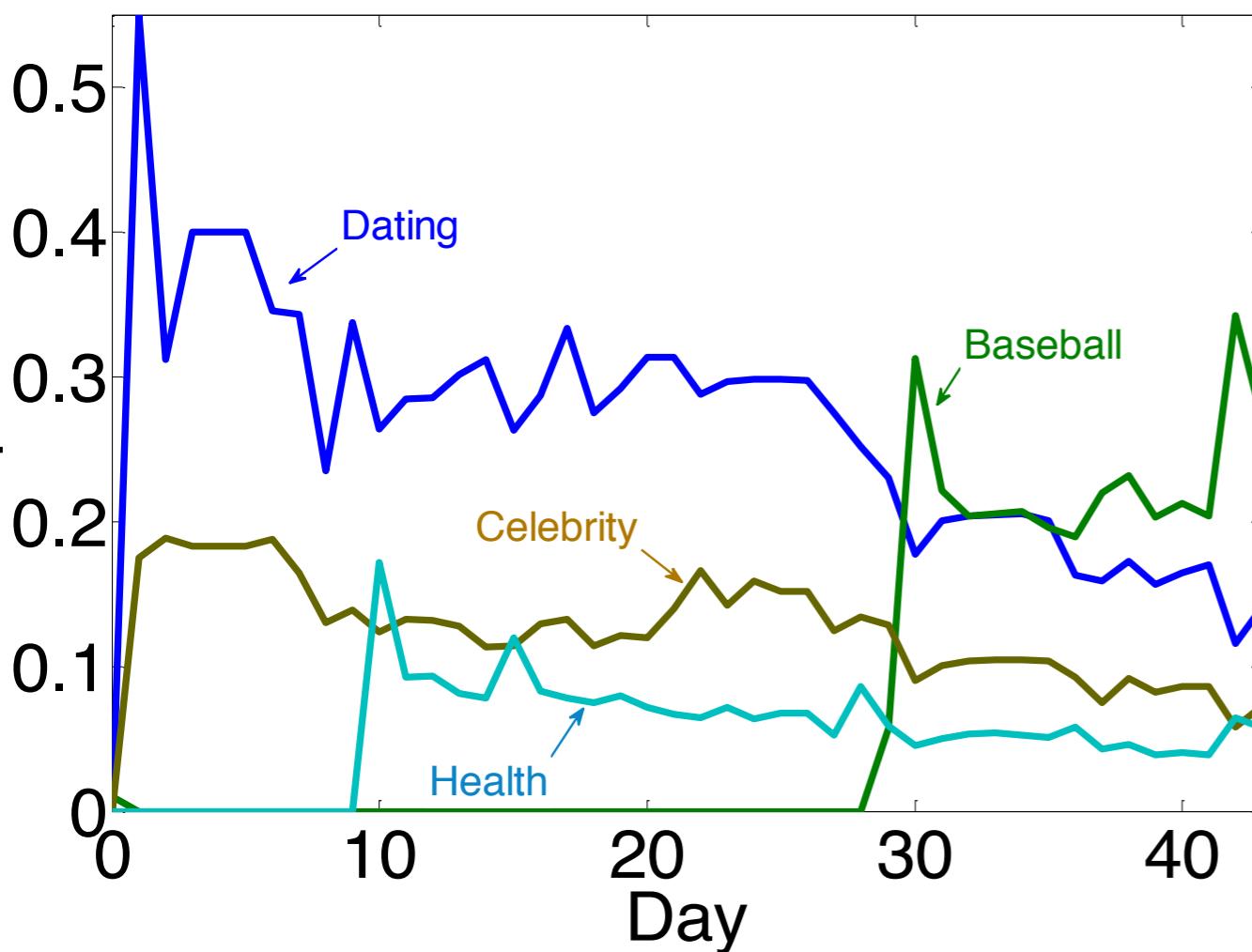
time

Statistical Model

- Topic model
 - Users - Documents
 - Actions - Words
 - Interests - Topics
- Παντα ρει (everything flows)
 - Users' interest preferences change over time
 - Interests change over time
 - Changing flavor of the day



Some 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

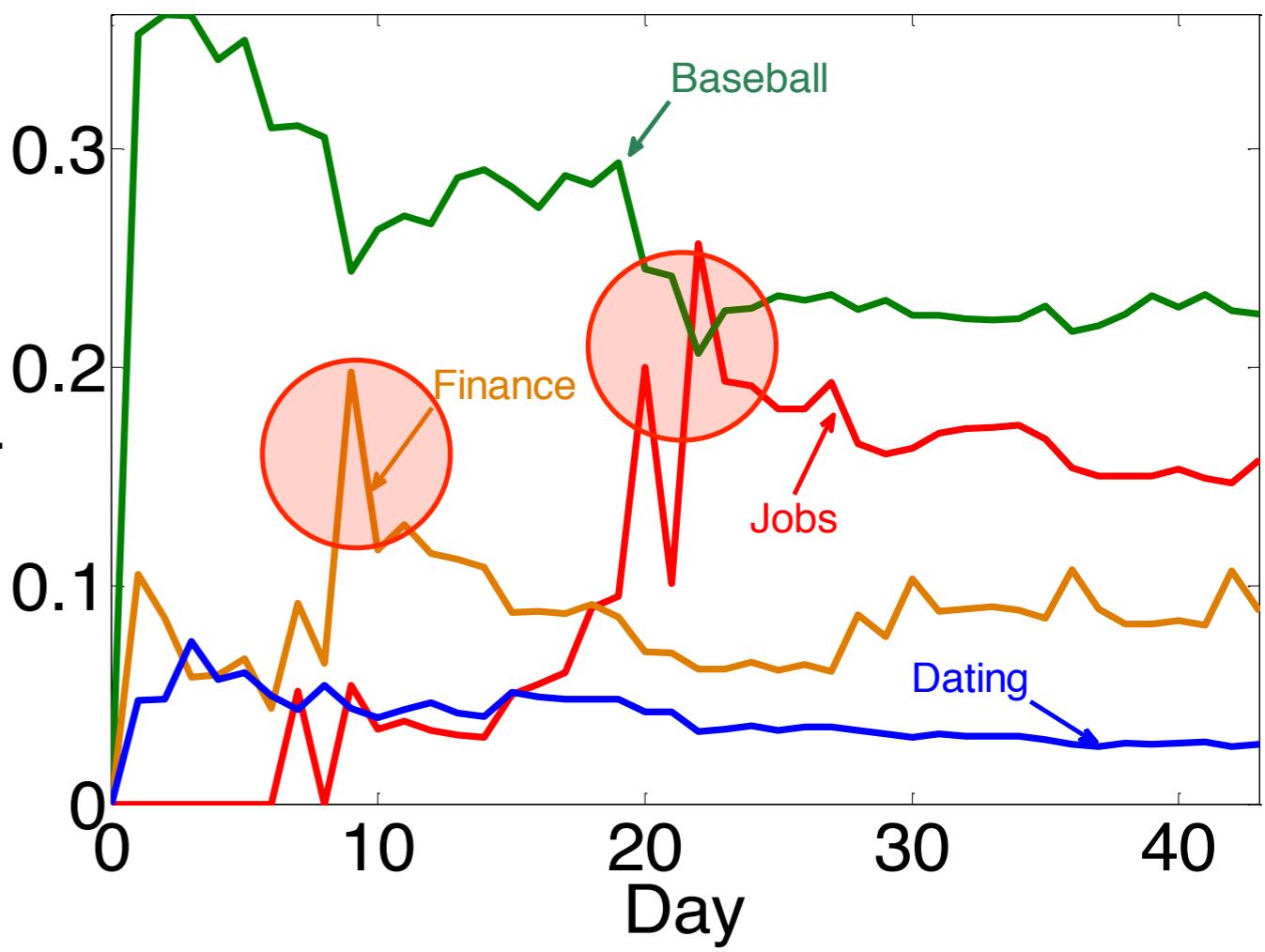
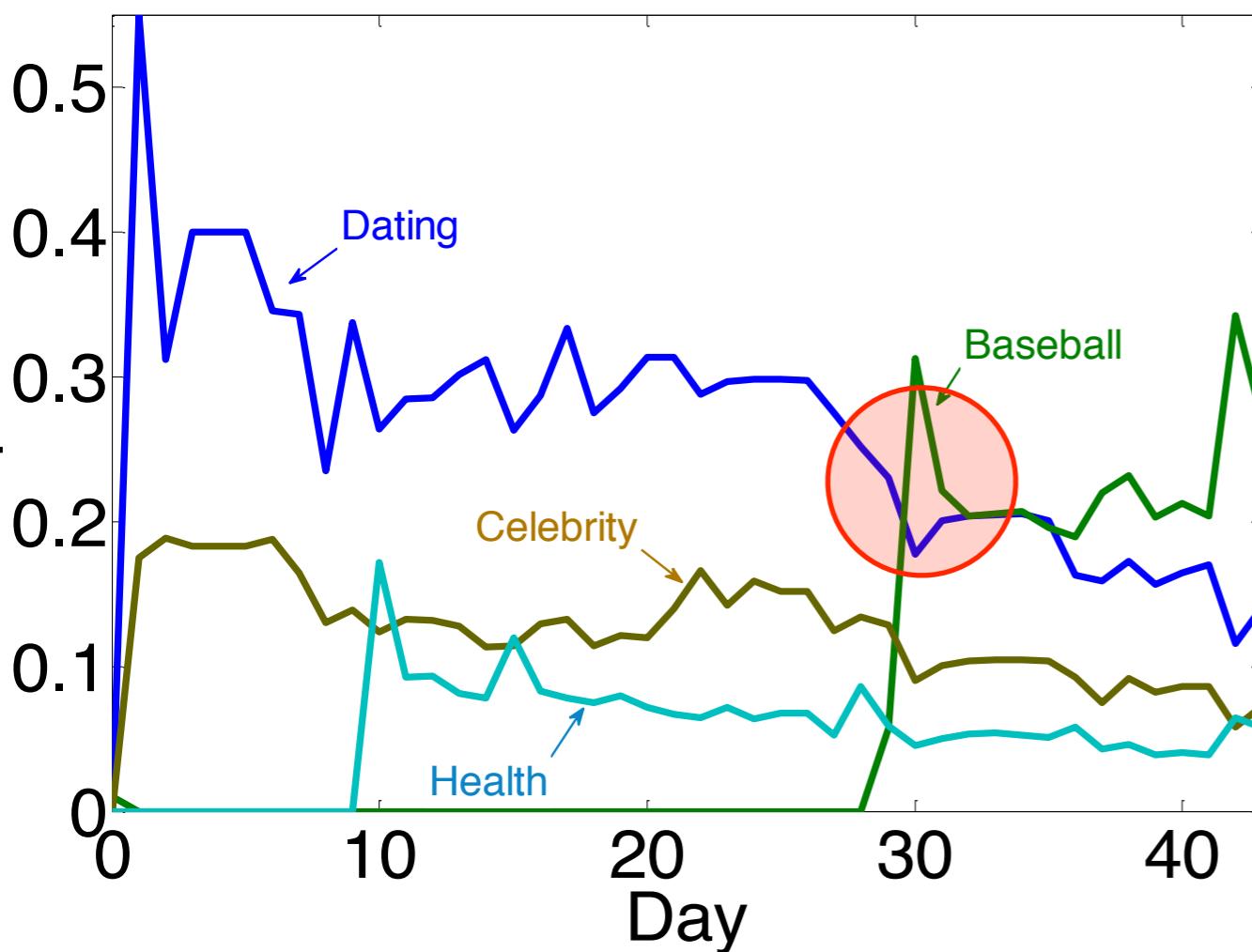
Jobs

job
career
business
assistant
hiring
part-time
receptionist

Finance

financial
Thomson
chart
real
Stock
Trading
currency

Some 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

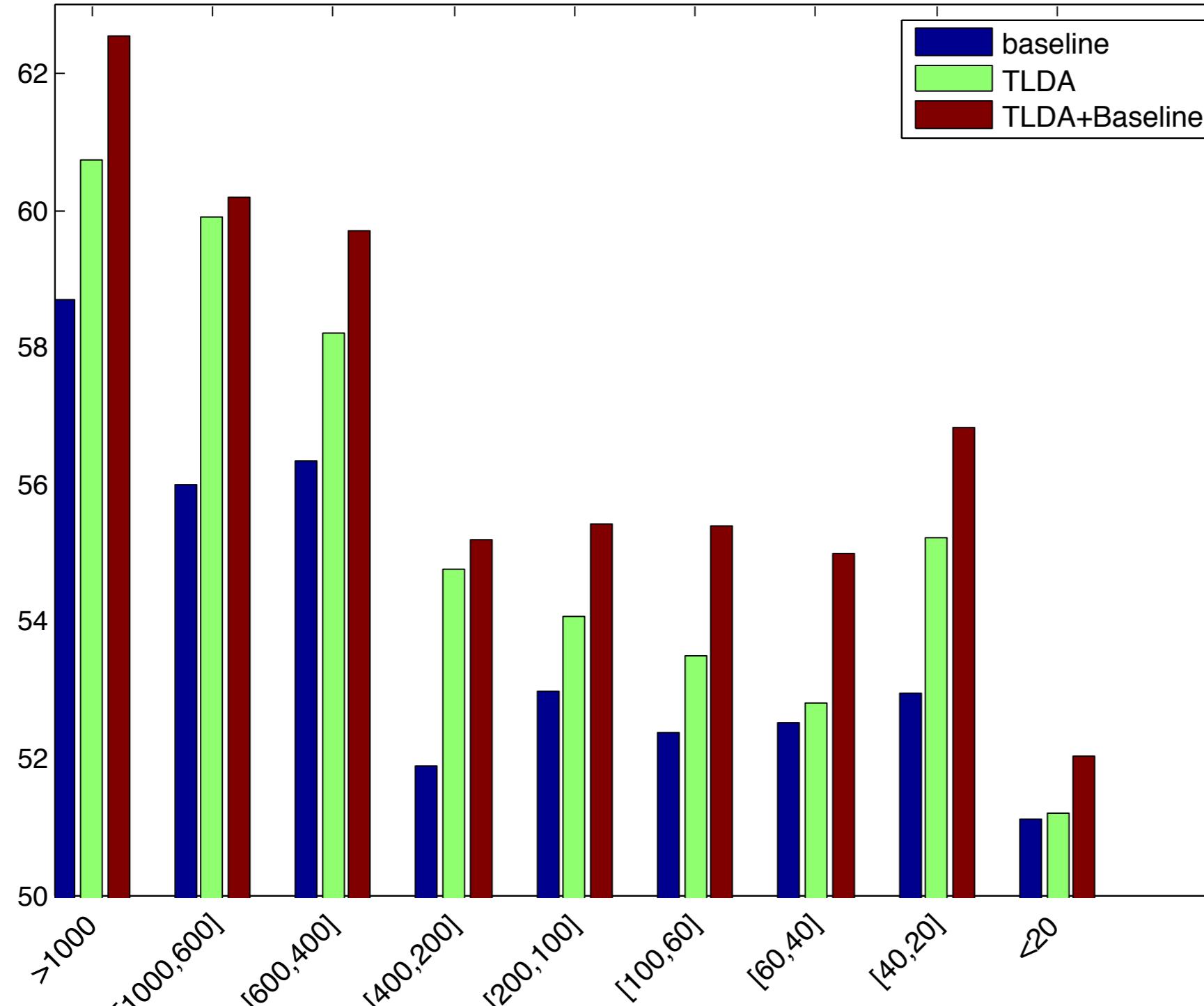
Jobs

job
career
business
assistant
hiring
part-time
receptionist

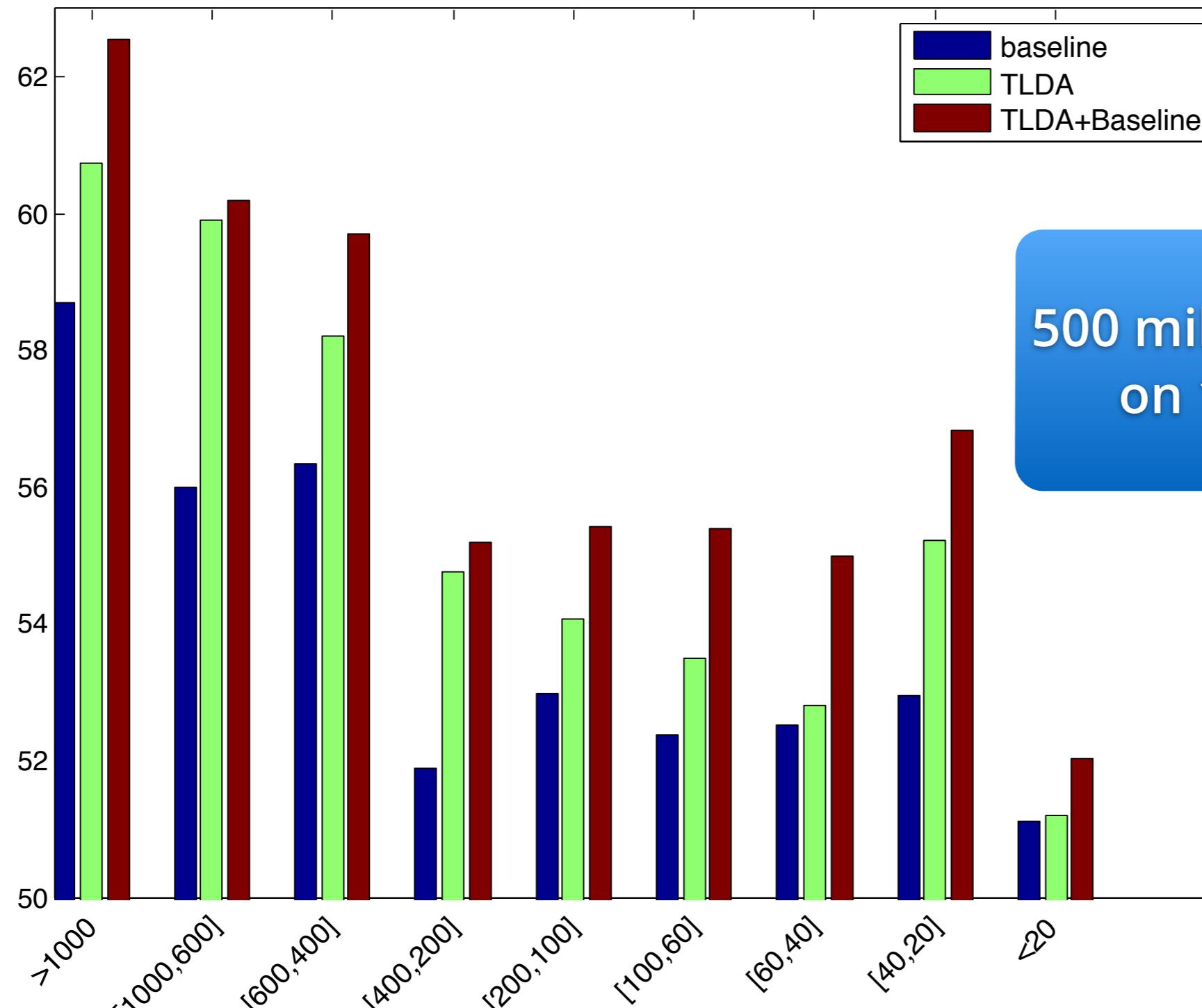
Finance

financial
Thomson
chart
real
Stock
Trading
currency

Improvement (\$\$\$)



Improvement (\$\$\$)



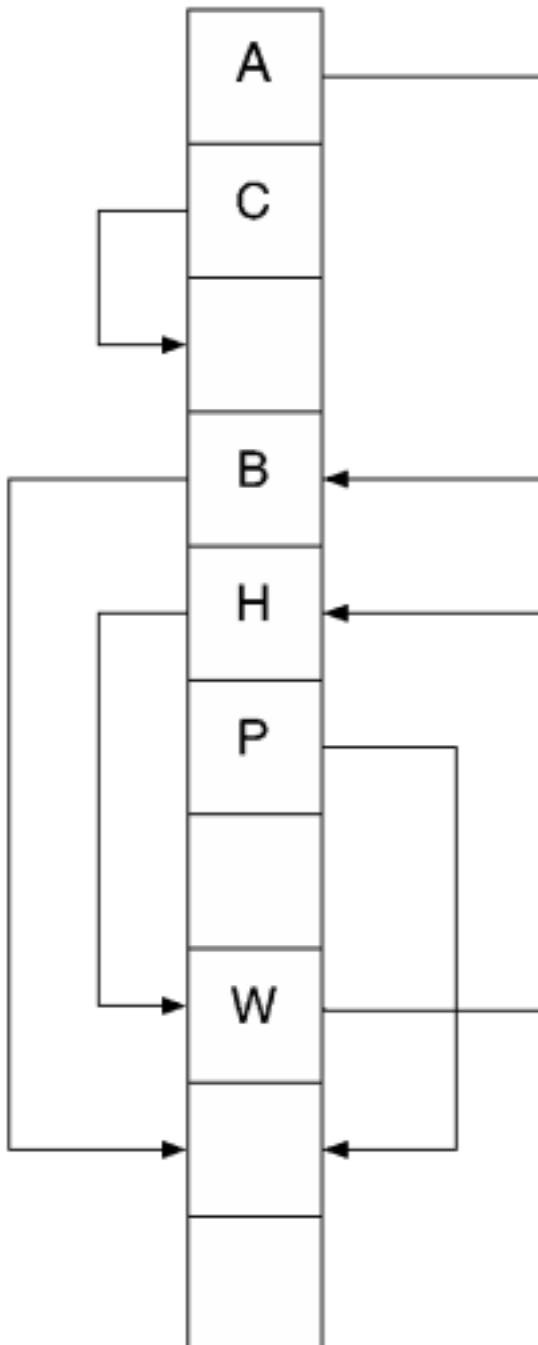
500 million users per day
on 1000 machines



Cuckoo Hashing

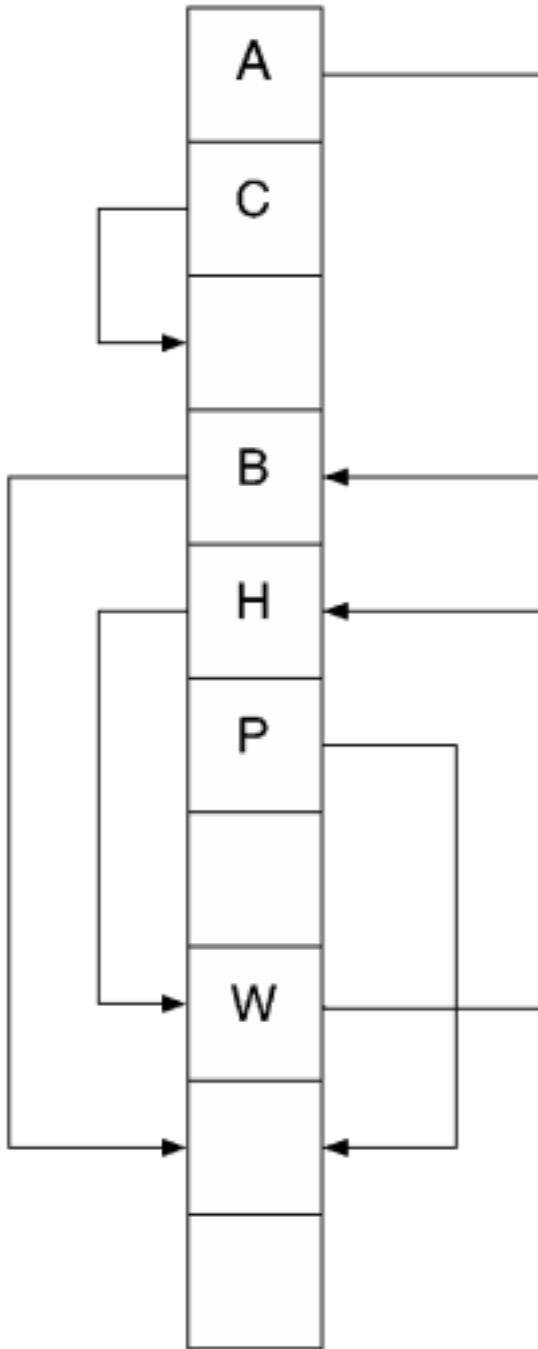
Fast and exact L1 linear algebra

Cuckoo Hash



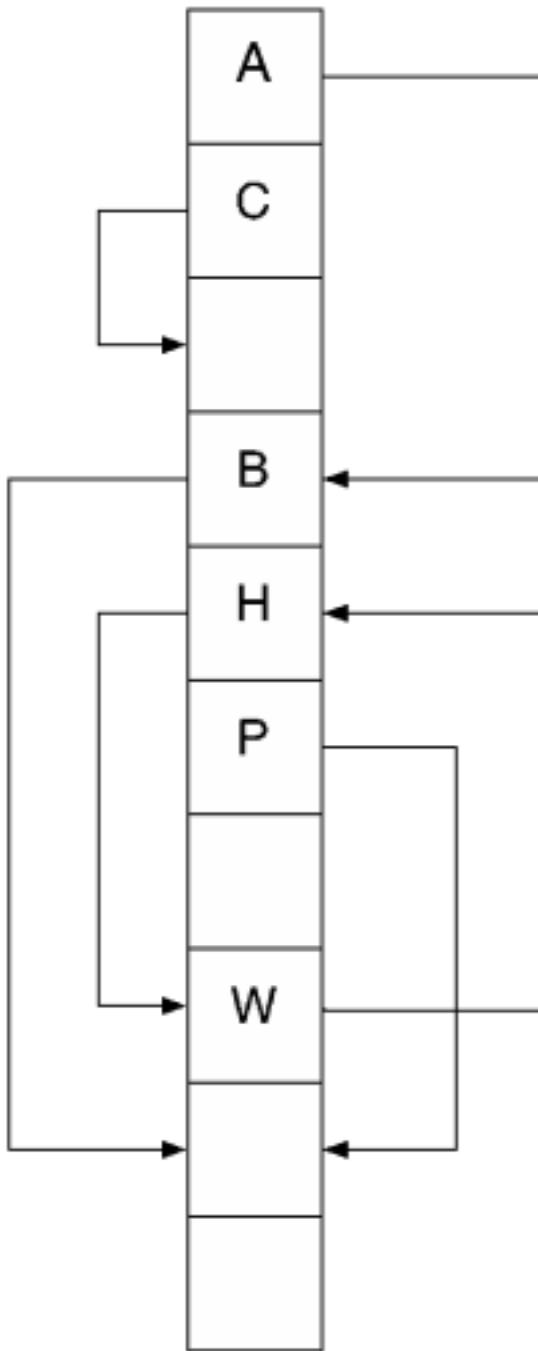
- Use power of two choices trick
 - Given key x insert into array at $h_1(x)$ or $h_2(x)$
 - If both slots are full, evict existing key
 - This key now starts finding a new location
- $O(1)$ lookup for arbitrary keys
- Short eviction chain

Cuckoo Hash



- $O(1)$ lookup for arbitrary keys
- Short eviction chain
- Need to keep key to compute when switching.
- Alternative
$$h_2(x) = h_1(x) \text{ XOR } h(s(x))$$
hence
$$h_1(x) = h_2(x) \text{ XOR } h(s(x))$$
- Poor fill rate
- Use multiple slots per location

Cuckoo hashes



- In practice 80% fill rate is quite reasonable
- Dot products cost $O(n)$ time (this is all we need)
- Merging lists costs $O(n)$ time (for gradient updates)
- Prefetching from RAM to deal with cache misses
- No need to pre-synchronize lists (good for distributed algorithms)

Performance

- Accuracy and timing for l2

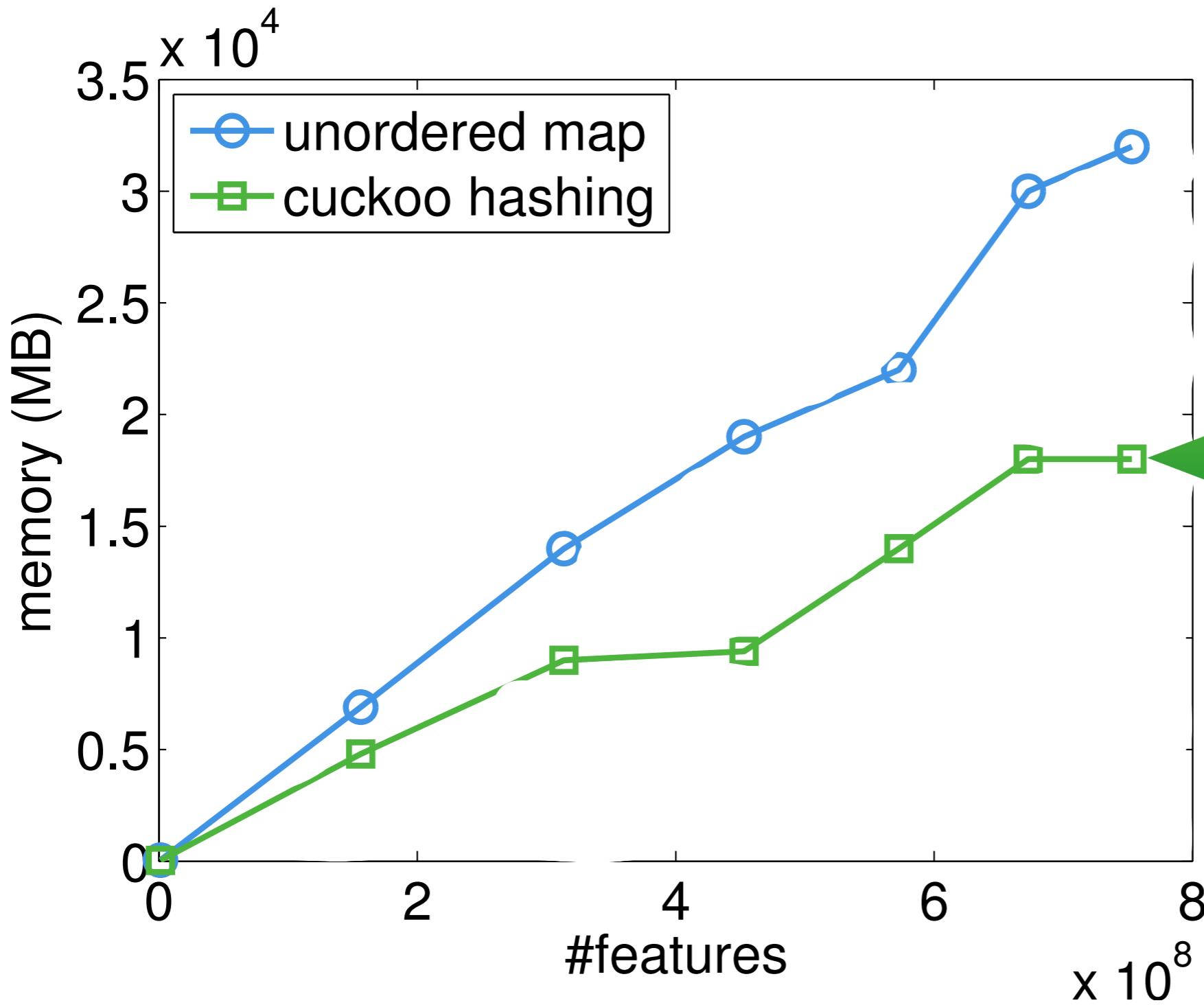
	LibLinear	Cuckoo	Hash Kernel $N = 27$	Hash Kernel $N = 20$
Pre-processing Time (sec)	514	0	0	0
Transpose Time (sec)	32	122	70	50
Training Time (sec)	2077	1520	2373	370
Total Time (sec)	2623	1643	2444	421
Memory Used (GB)	29	21	30	23
Accuray	93.22%	93.23%	93.25%	90.96%

- Reconstruction accuracy for l1
(Jaccard score for different SGD runs)

LibLinear	Cuckoo	hash kernel	
		$N = 20$	$N = 27$
0.8777	0.8773	0.0052	0.2638

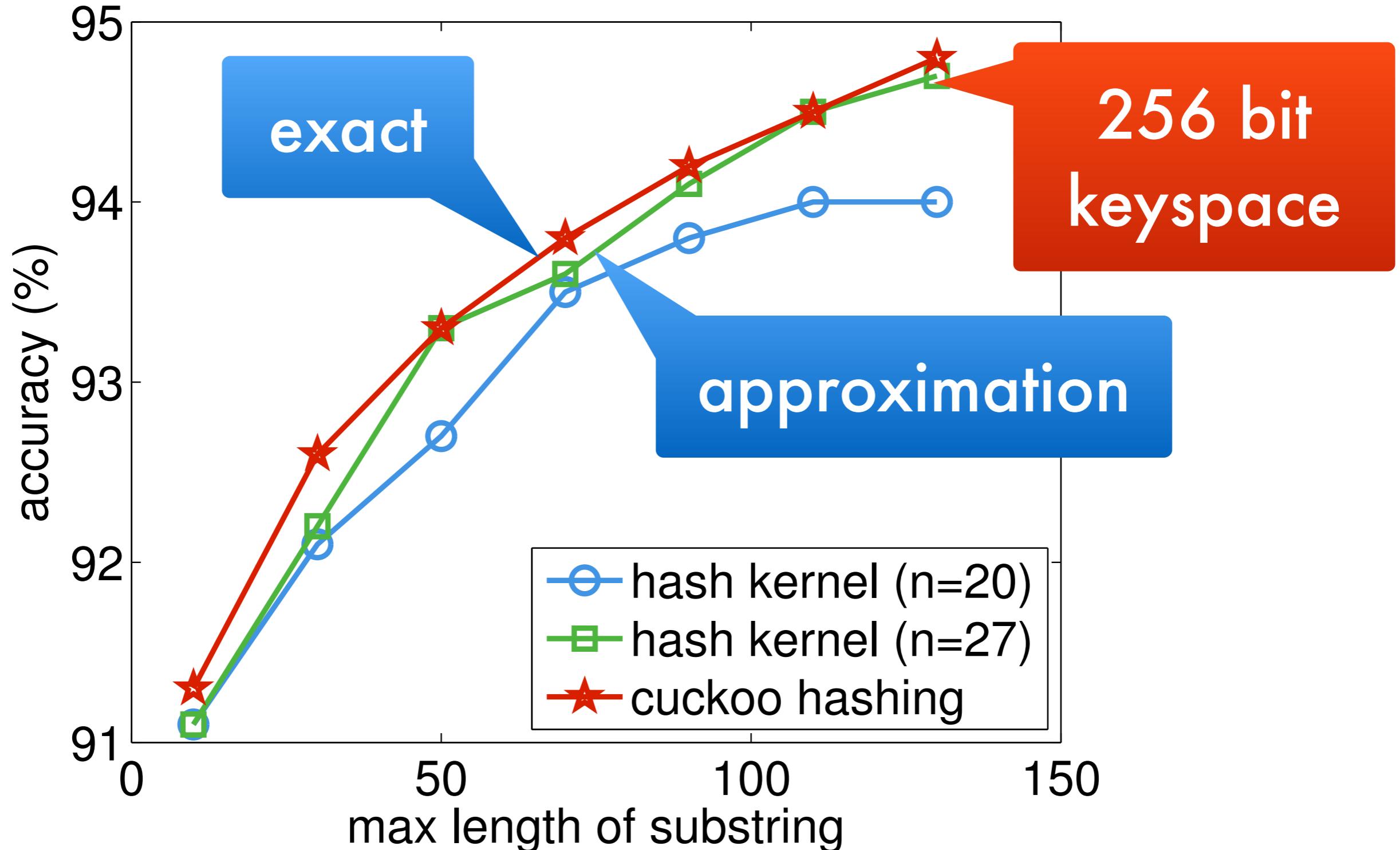
Memory efficiency

(cuckoo hash vs. unordered STL map)



Generating all substrings

(Pascal gene classification data)



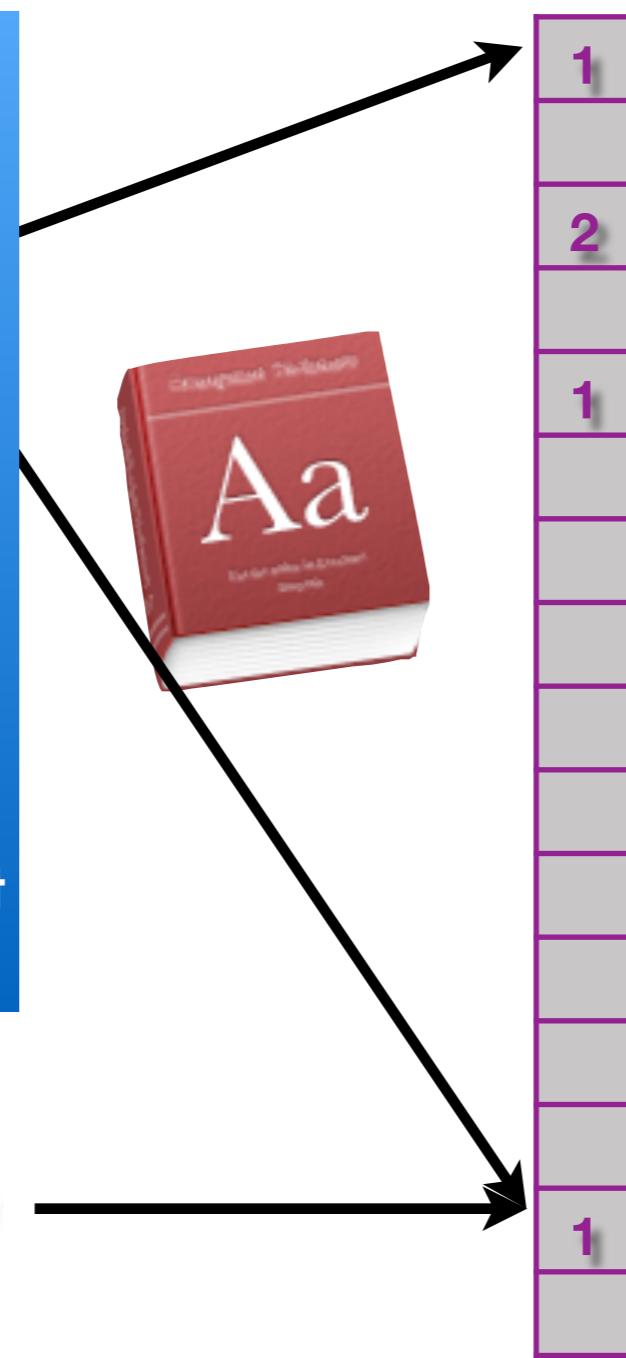
A close-up photograph of a large pile of hash kernels, a type of potato chip. The chips are thin, curly-cut pieces of golden-yellow potato. They are piled high, filling most of the frame. Some larger, irregular pieces of potato are interspersed among the smaller, more uniform hash kernels.

Hash Kernels

Hash Kernels

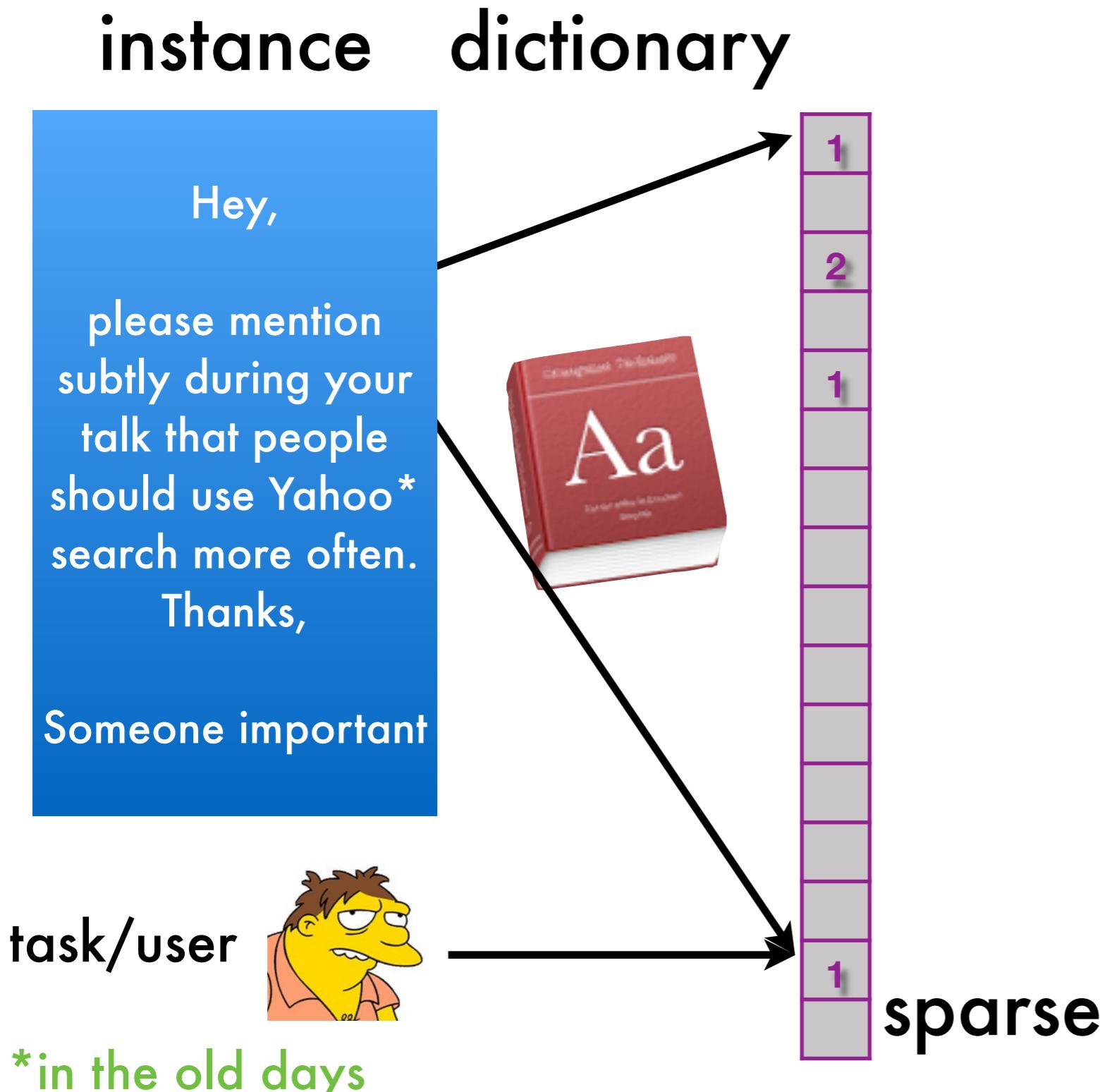
instance dictionary

Hey,
please mention
subtly during your
talk that people
should use Yahoo*
search more often.
Thanks,
Someone important

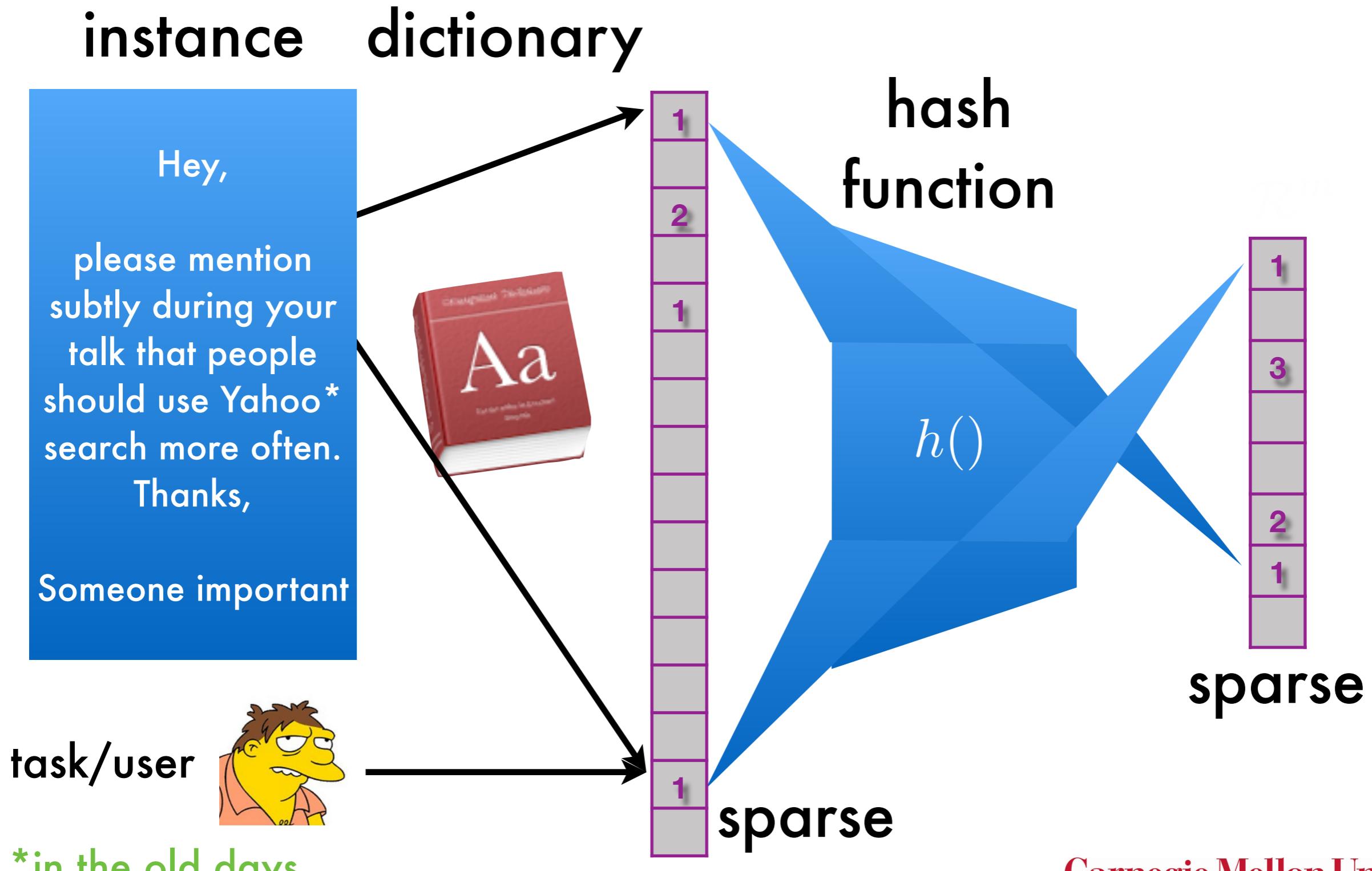


task/user

Hash Kernels



Hash Kernels



Hash Kernels

instance

Hey,
please mention
subtly during your
talk that people
should use Yahoo*
search more often.
Thanks,
Someone important

mention

hash
function

$h()$

1
3
2
1

sparse

task/user



mention_barney

*in the old days

Carnegie Mellon University

Hash Kernels

instance

Hey,
please mention
subtly during your
talk that people
should use Yahoo*
search more often.
Thanks,
Someone important



task/user

*in the old days

hash
function

$h(\text{mention})\sigma(\text{mention})$



sparse

$h(\text{mention}, \text{barney})\sigma(\text{mention}, \text{barney})$

Similar to count sketch (Charikar, Chen, Farrach-Colton, 2002)

Hash Kernels

- Linear model

$$f(x) = \langle w, x \rangle = \sum_i w_i x_i$$

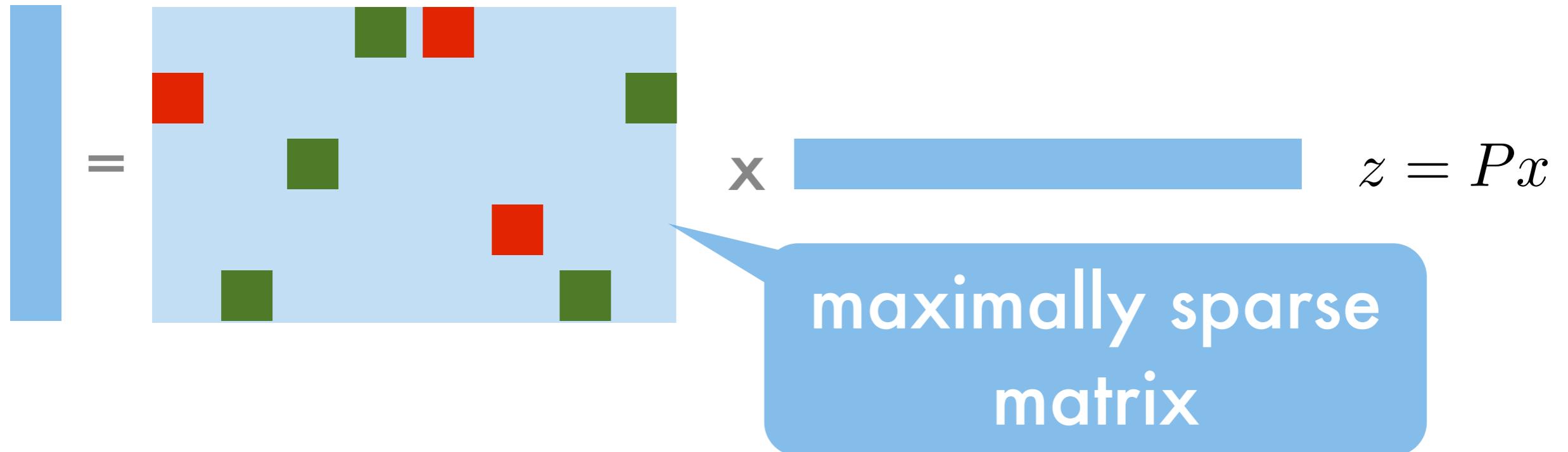
- Hashed model

$$f(x) = \sum_i w_{h(i)} x_i \sigma(i)$$

- Applications

- Multitask learning
- Fast string kernels
- Collaborative filtering (compress matrix)

Hash Kernels - the matrix view

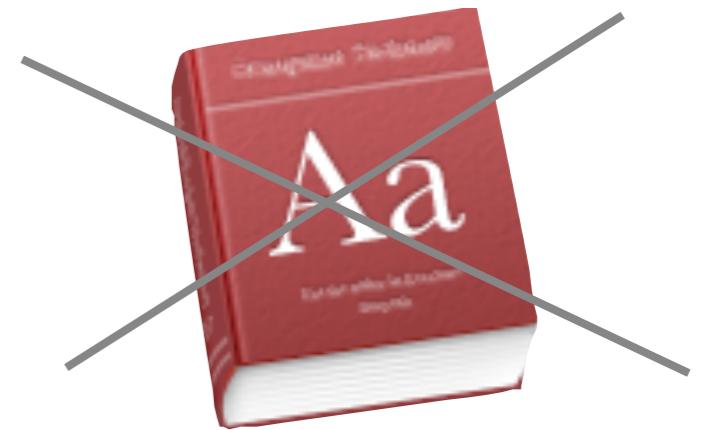


$$\langle w, x \rangle = \sum_i w_i x_i \quad \langle \bar{w}, \bar{x} \rangle = \sum_j \left[\sum_{i:h(i)=j} w_i \sigma(i) \right] \left[\sum_{i:h(i)=j} x_i \sigma(i) \right]$$

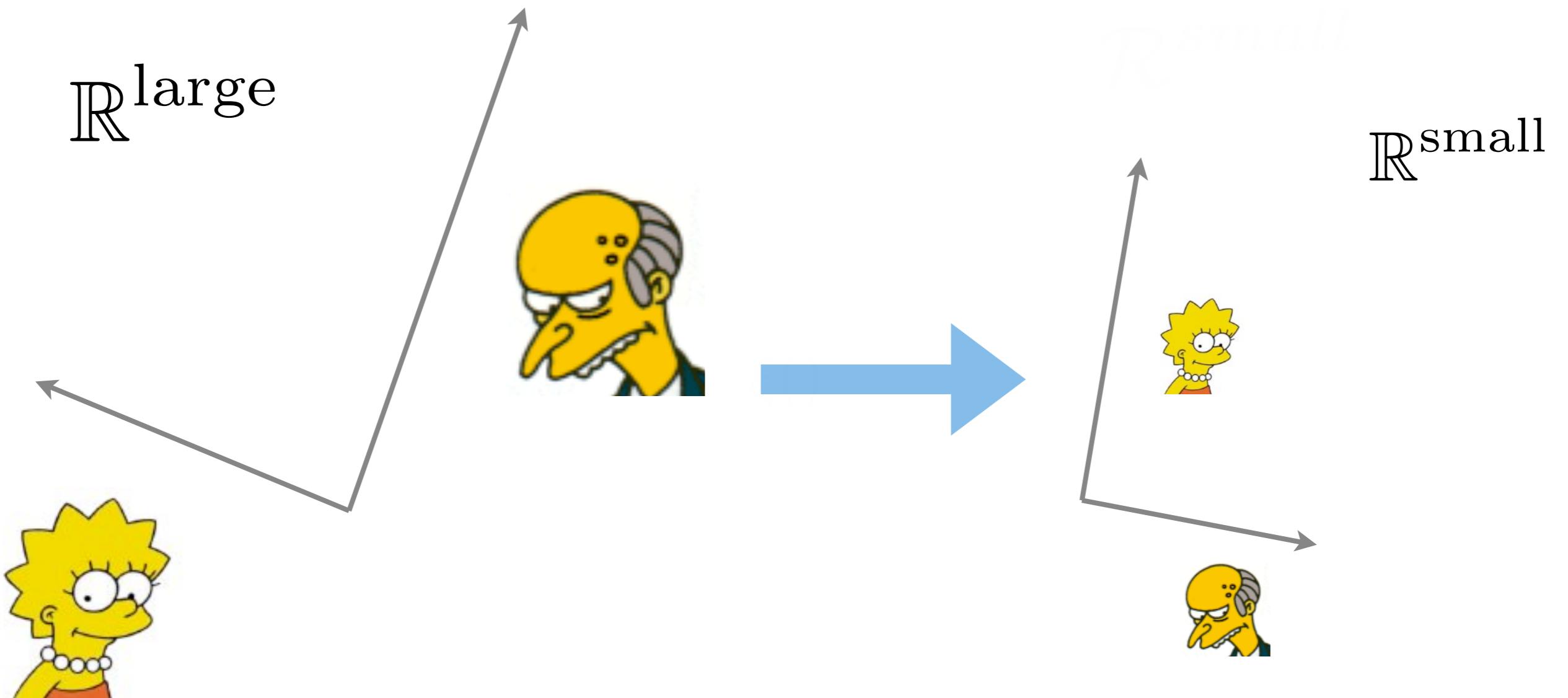
$$\mathbb{E}_\sigma[\sigma(i)\sigma(i')] = \delta_{ii'}$$

Advantages of hashing

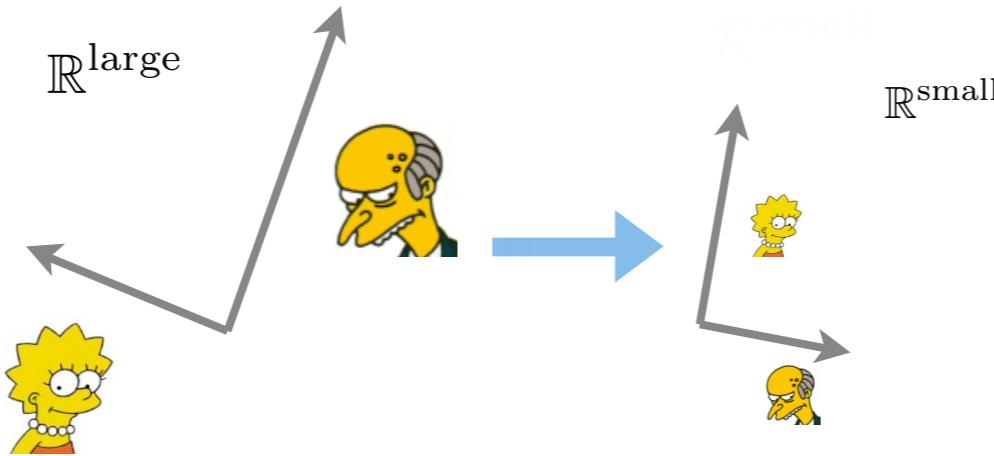
- No dictionary needed
No tokenization table
- Content drift is no problem
- All memory used for classification
No projection matrix (vs. LSH)
- Finite memory footprint
- Sparsity preserving! (vs LSH)



Intuition



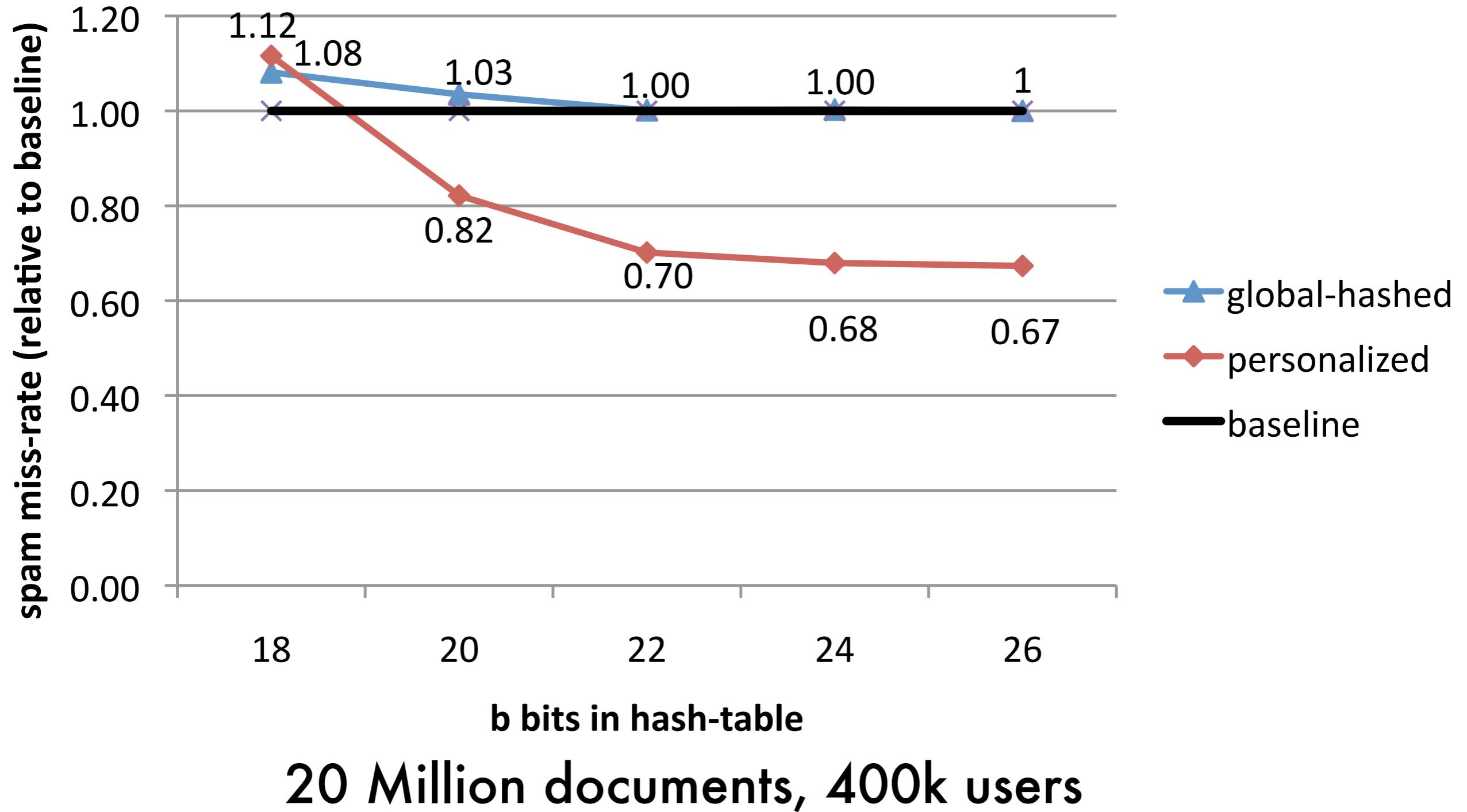
Guarantees



- For a random hash function the inner product vanishes
$$\Pr\{|\langle w_v, h_u(x) \rangle| > \epsilon\} \leq 2e^{-C\epsilon^2 m}$$
- Direct sums in Hilbert space
 - Direct sum in Hilbert Space \rightarrow Sum in Hash Space
- Hashed inner product is unbiased, variance is $O(1/n)$
- Restricted isometry property (Kumar, Sarlos, Dasgupta 2010)

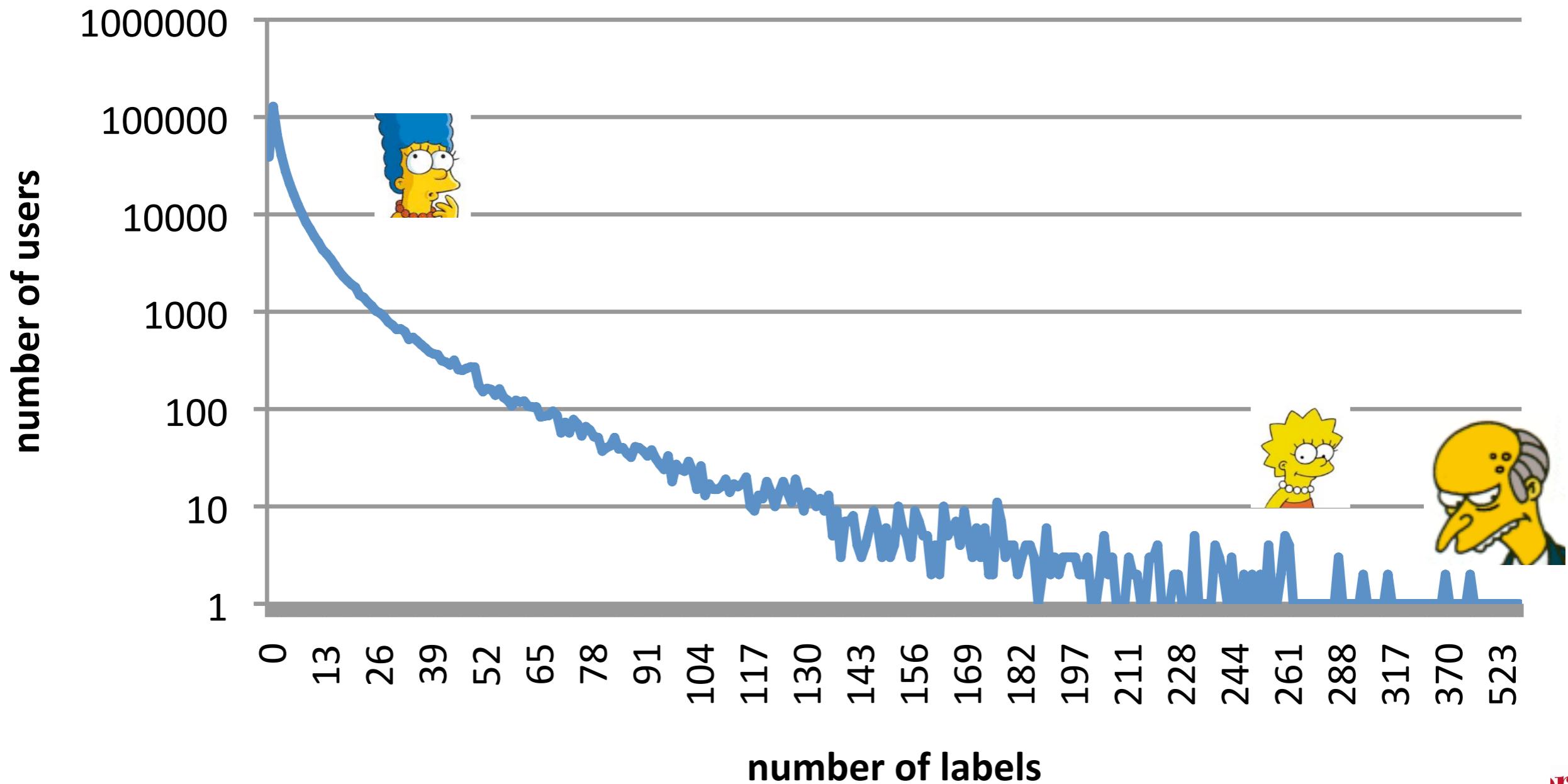
Weinberger, Dasgupta, Attenberg, Langford and Smola
Feature Hashing for Large Scale Multitask Learning, ICML'09

Spam classification results

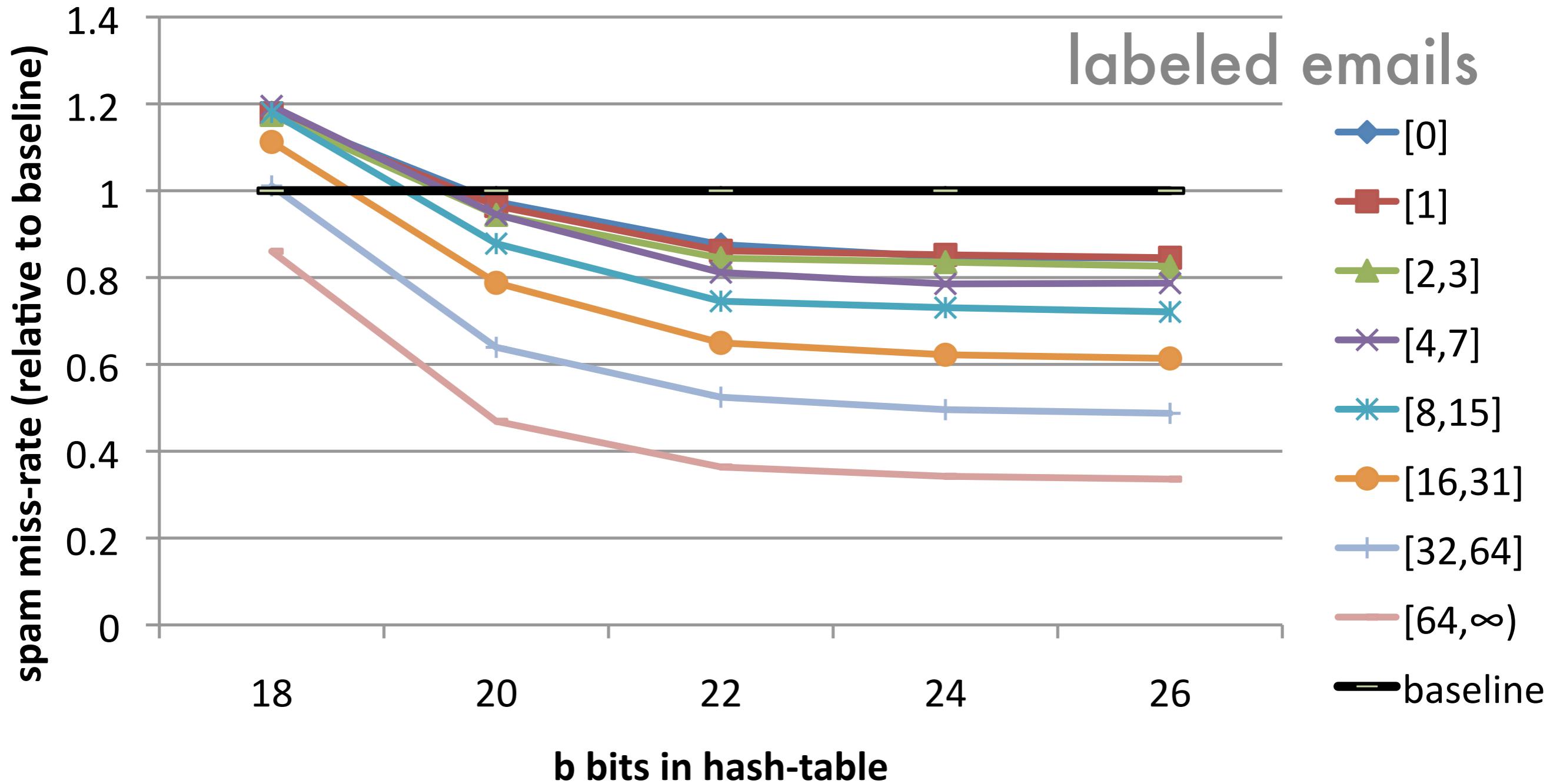


Lazy users ...

Labeled emails per user



Results by user group



Results by user group

