Integrating Social with Search

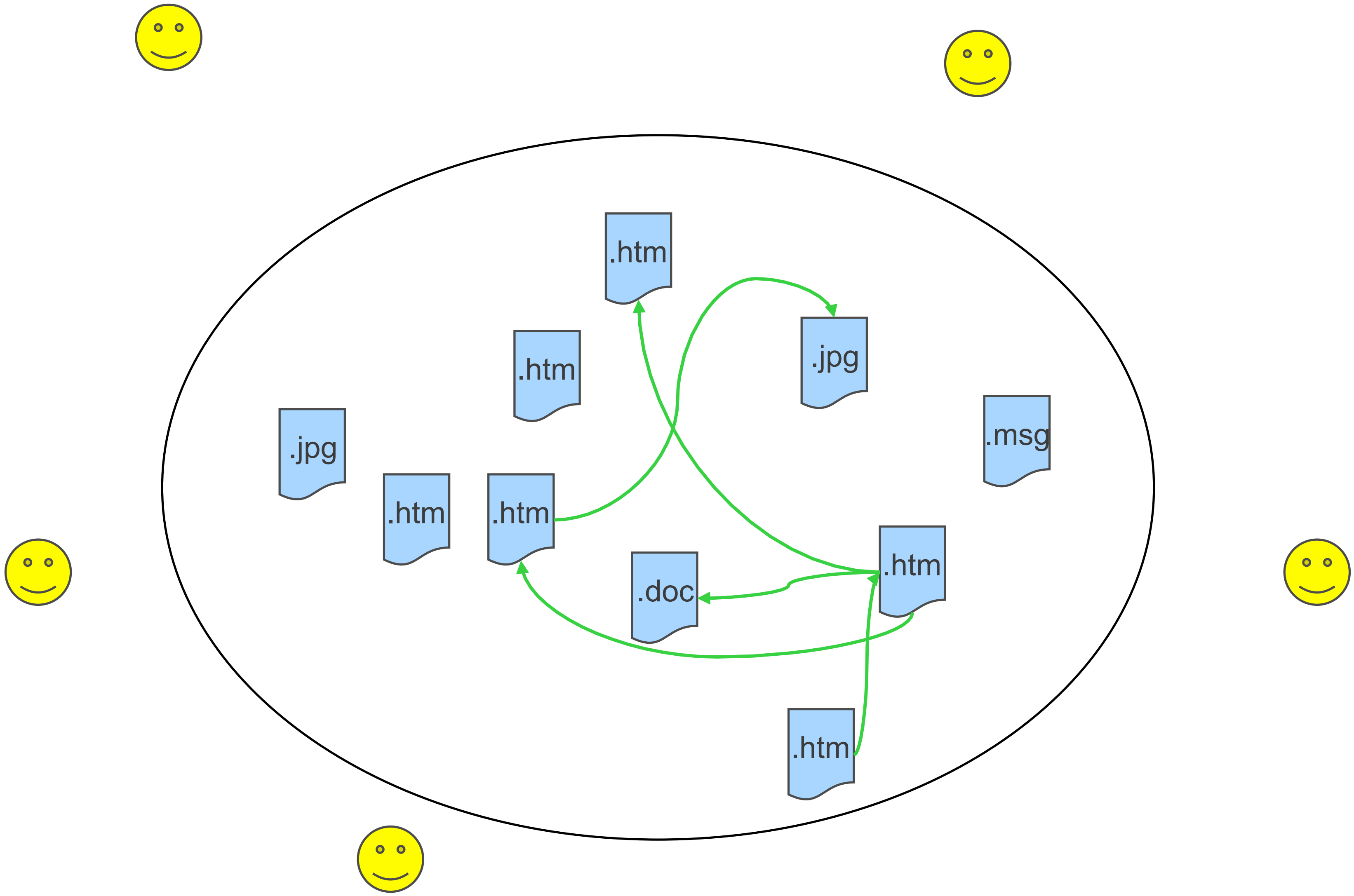
Edward Chang

Director, Google Research, Beijing

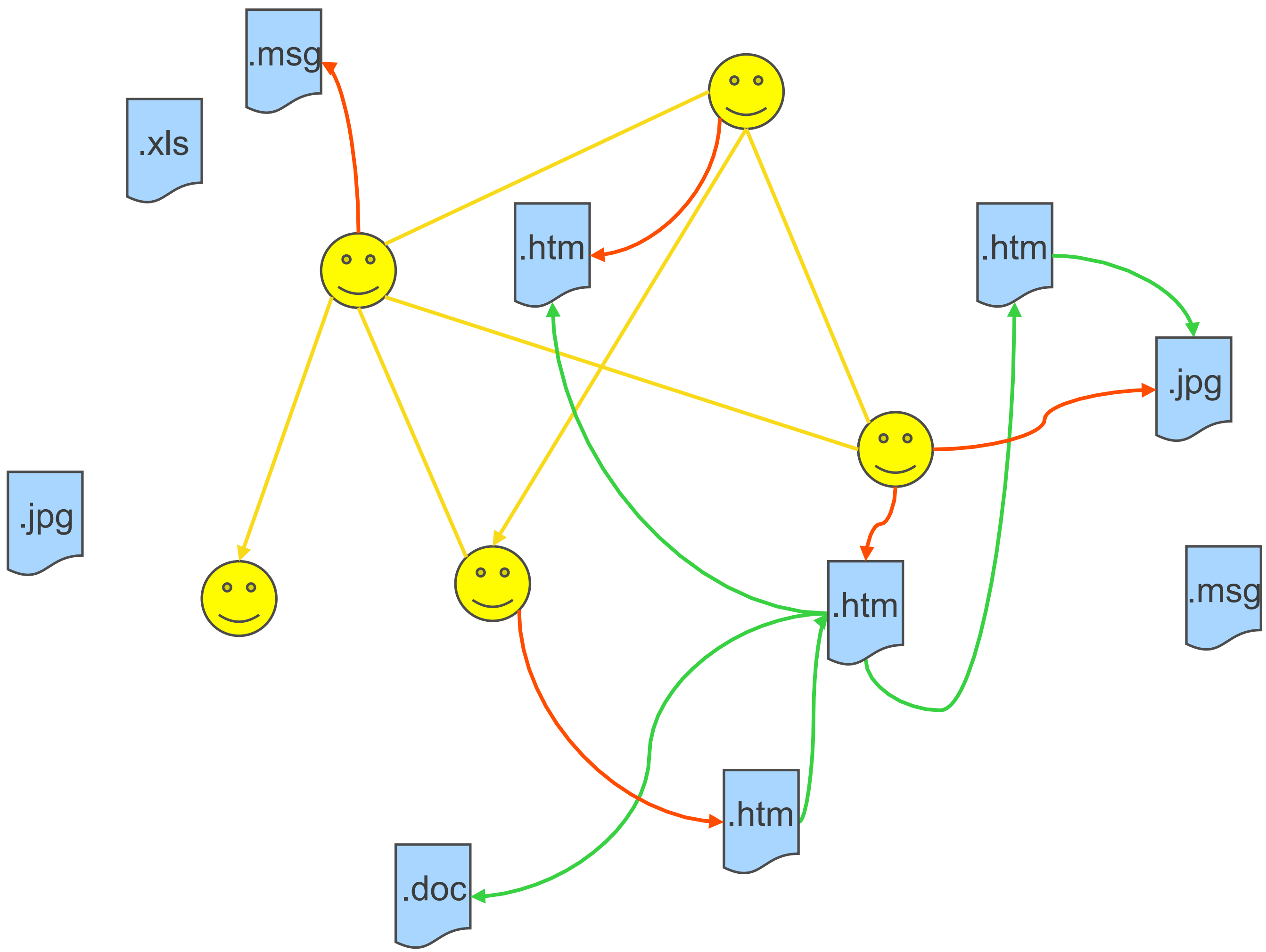
Related Papers

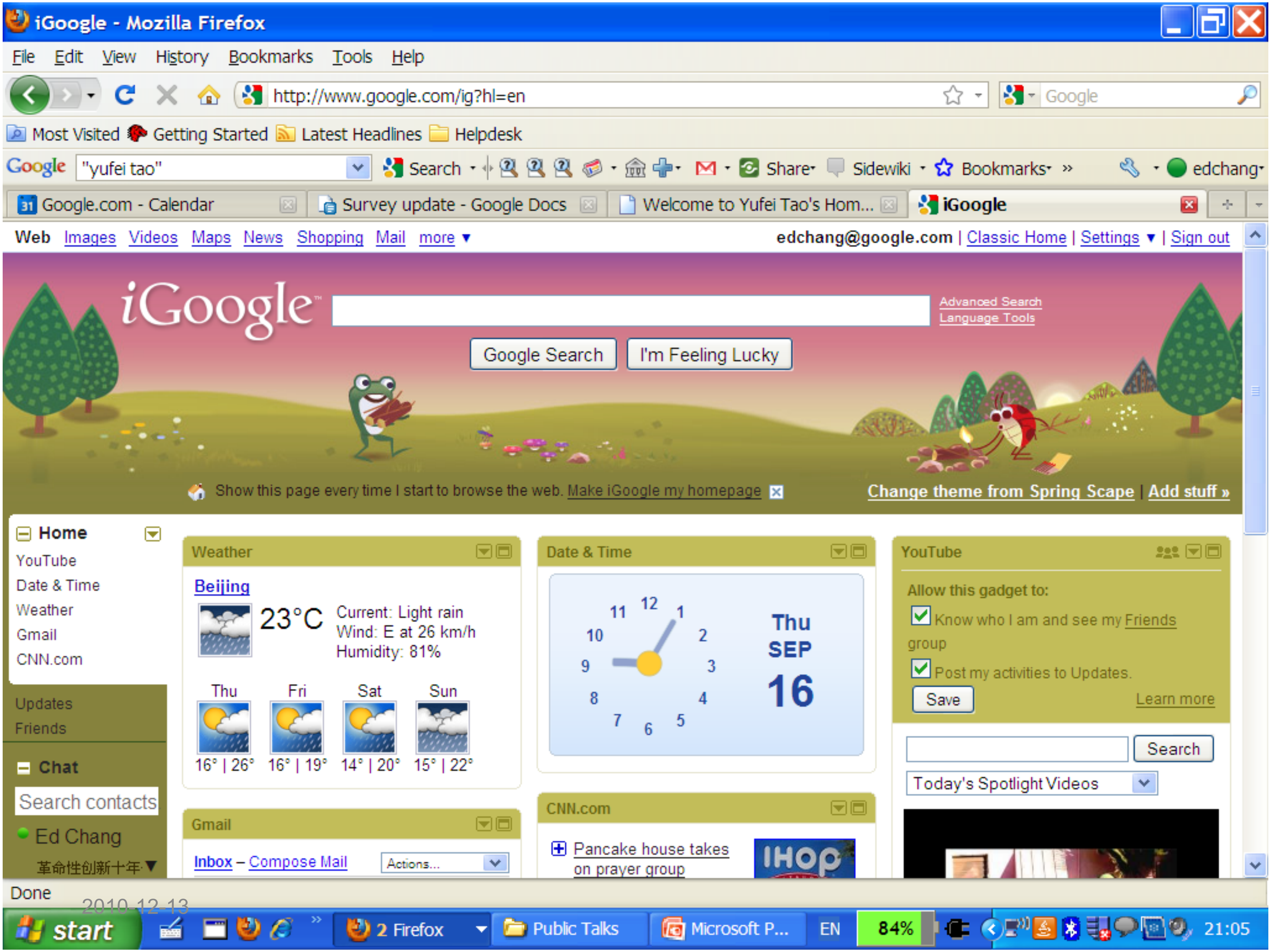
- **AdHeat (Social Ads)**:
  - AdHeat: An Influence-based Diffusion Model for Propagating Hints to Match Ads, H.J. Bao and E. Y. Chang, WWW 2010 (best paper candidate), April 2010.
  - Parallel Spectral Clustering in Distributed Systems, Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and E. Y. Chang, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2010.
- **UserRank:**
  - Confucius and its Intelligent Disciples, X. Si, E. Y. Chang, Z. Gyongyi, VLDB, September 2010 .
  - Topic-dependent User Rank, Xiance Si, Z. Gyongyi, E. Y. Chang, and M.S. Sun, Google Technical Report.
- **Large-scale Collaborative Filtering**:
  - PLDA+: Parallel Latent Dirichlet Allocation for Large-Scale Applications, ACM Transactions on Internet Technology, 2011.
  - Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior, W.-Y. Chen, J. Chu, E. Y. Chang, WWW 2009: 681-690.
  - Combinational Collaborative Filtering for Personalized Community Recommendation, W.-Y. Chen, E. Y. Chang, KDD 2008: 115-123.
  - PSVM: Parallelizing SVMs on distributed machines, E. Y. Chang, et al., NIPS 2007.

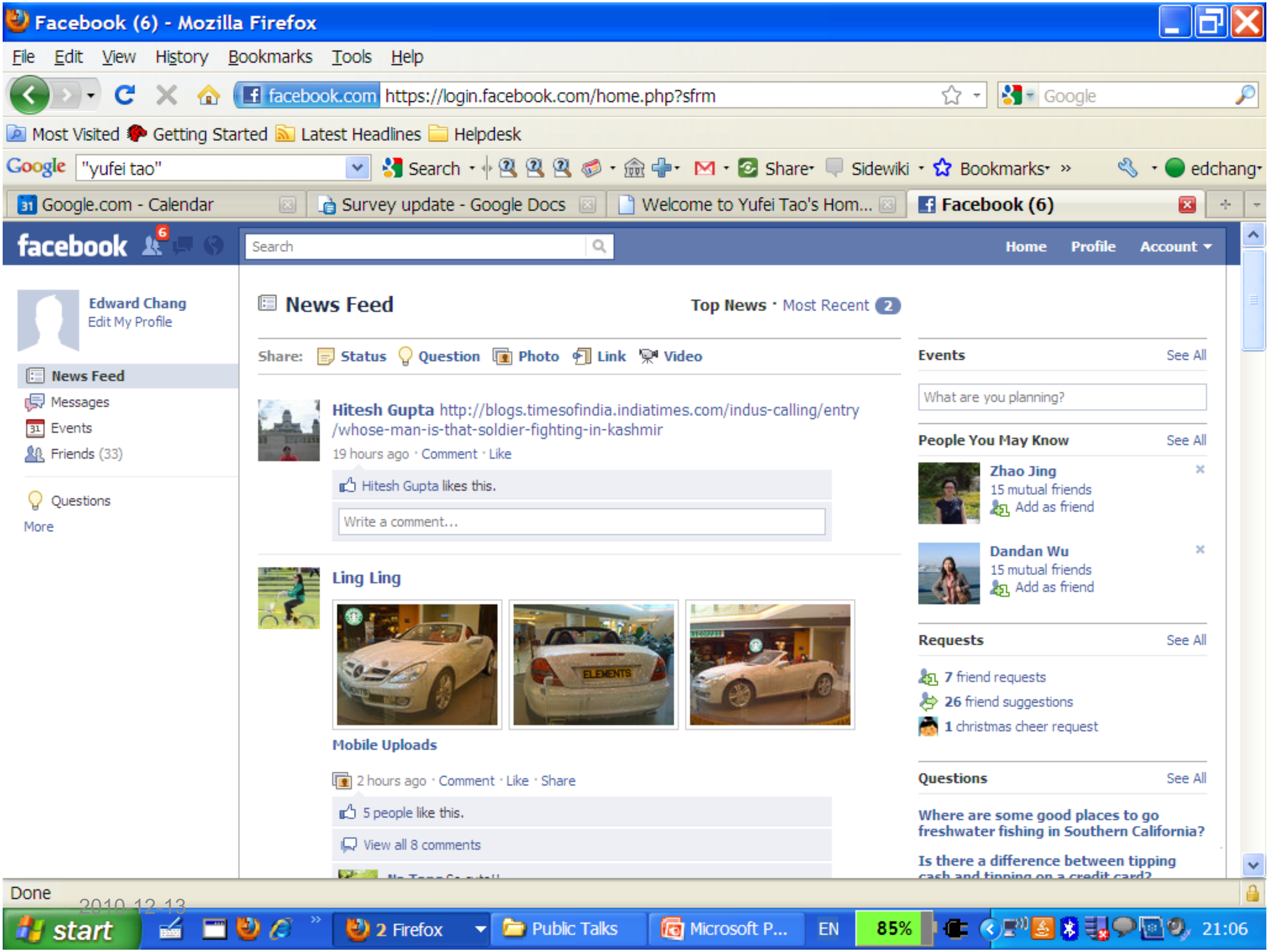
Web 1.0



Web 2.0 --- Web with People




iGoogle - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/ig?hl=en
Google
Most Visited Getting Started Latest Headlines Helpdesk
Google "yufei tao" Search Share Sidewiki Bookmarks edchang
Google.com - Calendar
Survey update - Google Docs
Welcome to Yufei Tao's Hom...
iGoogle
Web Images Videos Maps News Shopping Mail more
edchang@google.com | Classic Home | Settings | Sign out
iGoogle
Advanced Search
Language Tools
Google Search
I'm Feeling Lucky
Show this page every time I start to browse the web. Make iGoogle my homepage
Change theme from Spring Scape | Add stuff »
Home
YouTube
Date & Time
Weather
Gmail
CNN.com
Updates
Friends
Chat
Search contacts
Ed Chang
革命性创新十年
Weather
Beijing
23°C
Current: Light rain
Wind: E at 26 km/h
Humidity: 81%
Thu Fri Sat Sun
16° | 26° 16° | 19° 14° | 20° 15° | 22°
Date & Time
Thu
SEP
16
YouTube
Allow this gadget to:
Know who I am and see my Friends group
Post my activities to Updates.
Save
Learn more
Search
Today's Spotlight Videos
Gmail
Inbox – Compose Mail
Actions...
CNN.com
Pancake house takes on prayer group
IHOP
Done
2010-12-13
start
2 Firefox
Public Talks
Microsoft P...
EN
84%
21:05


Facebook (6) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
facebook.com https://login.facebook.com/home.php?sfrm
Google
Most Visited Getting Started Latest Headlines Helpdesk
Google "yufei tao" Search Share Sidewiki Bookmarks edchang
Google.com - Calendar
Survey update - Google Docs
Welcome to Yufei Tao's Hom...
Facebook (6)
facebook
Search
Home Profile Account
Edward Chang
Edit My Profile
News Feed
Messages
Events
Friends (33)
Questions
More
News Feed
Top News · Most Recent 2
Share: Status Question Photo Link Video
Hitesh Gupta http://blogs.timesofindia.indiatimes.com/indus-calling/entry/whose-man-is-that-soldier-fighting-in-kashmir
19 hours ago · Comment · Like
Hitesh Gupta likes this.
Write a comment...
Ling Ling
Mobile Uploads
2 hours ago · Comment · Like · Share
5 people like this.
View all 8 comments
Events See All
What are you planning?
People You May Know See All
Zhao Jing
15 mutual friends
Add as friend
Dandan Wu
15 mutual friends
Add as friend
Requests See All
7 friend requests
26 friend suggestions
1 christmas cheer request
Questions See All
Where are some good places to go freshwater fishing in Southern California?
Is there a difference between tipping cash and tipping on a credit card?
Done
start
2 Firefox
Public Talks
Microsoft P...
EN
85%
21:06
2010-12-13

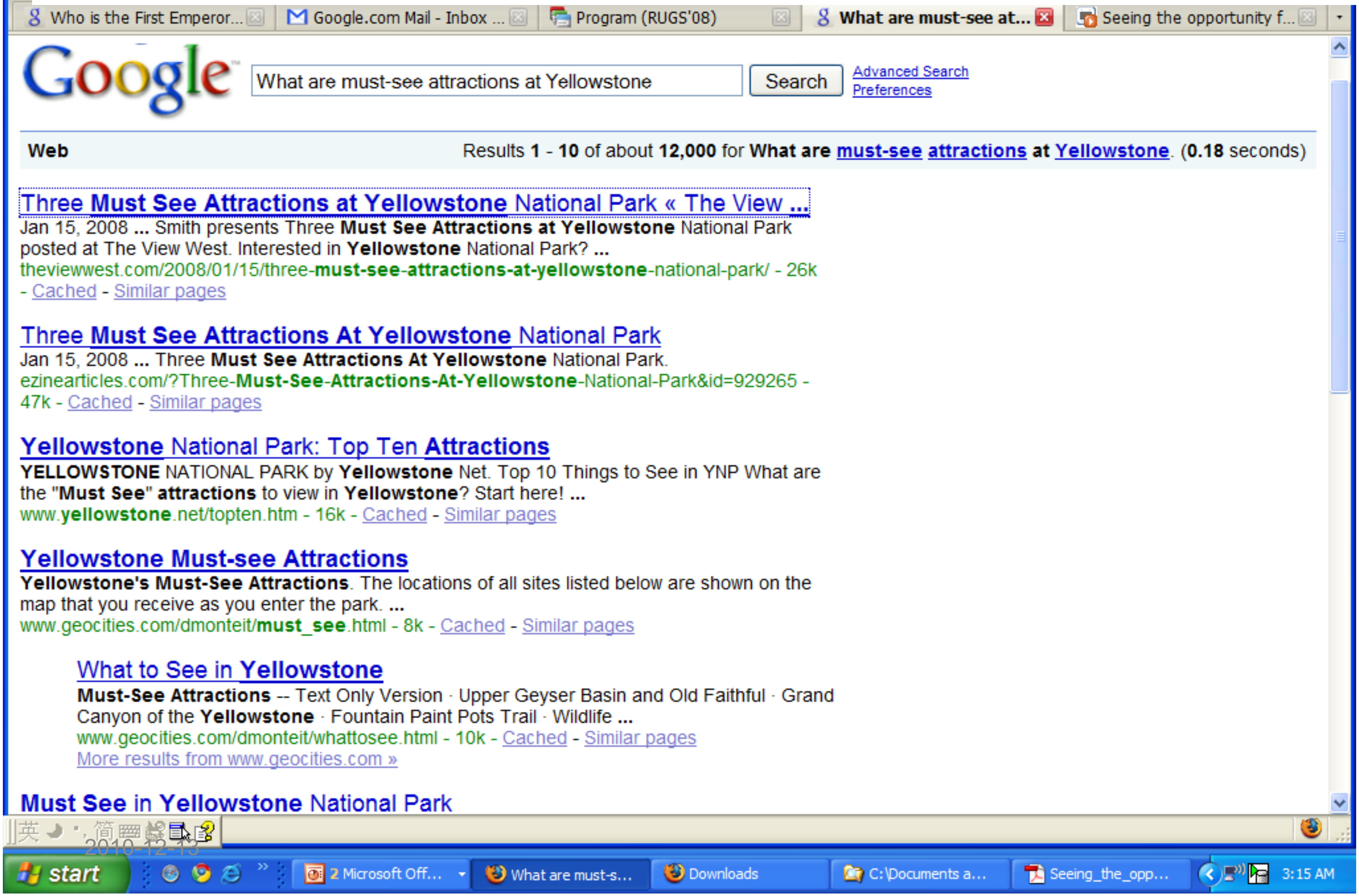
Outline

- Search + Social Synergy
- Search → Social
- Social → Search
- Scalability

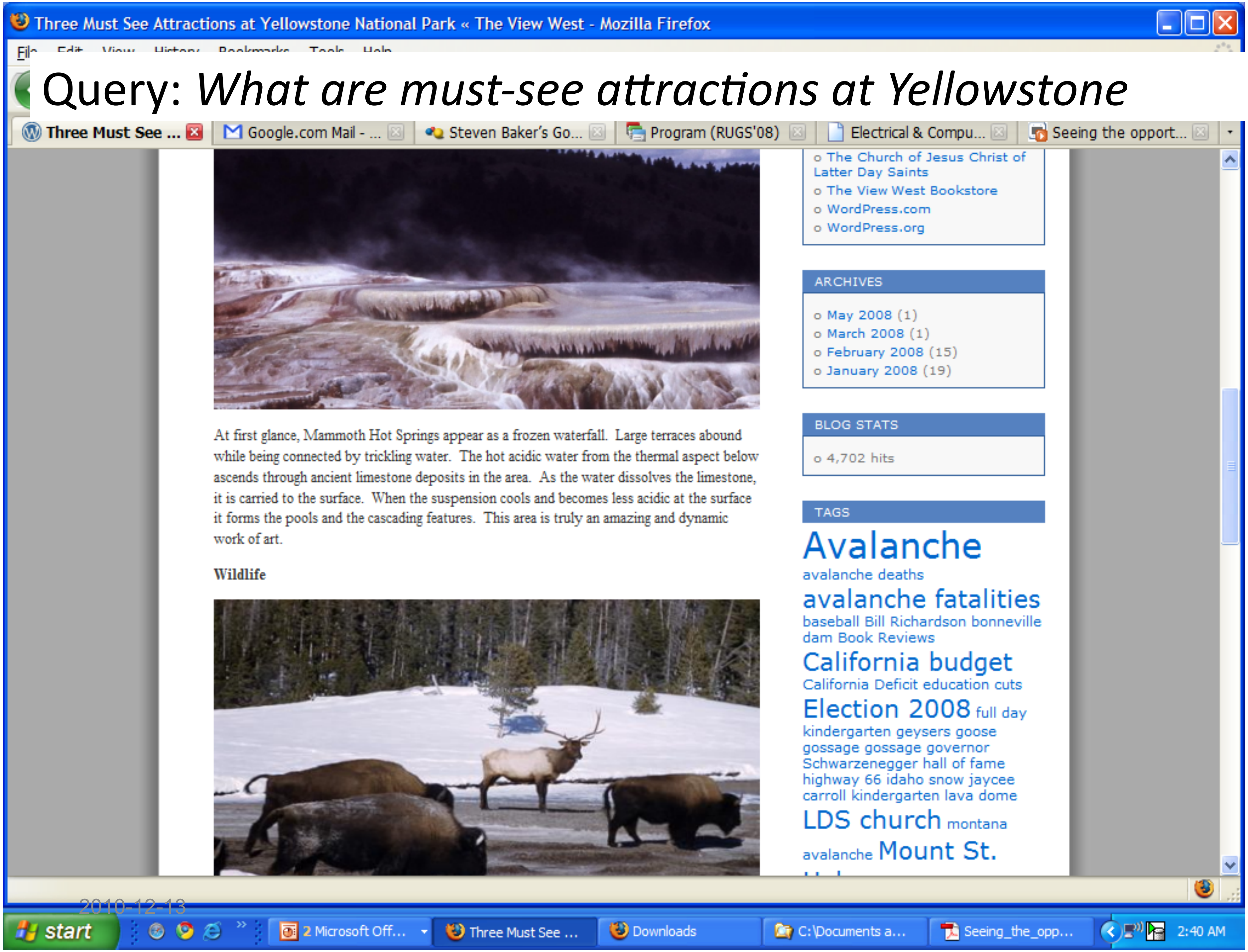
Google Q&A (Confucius)

- Developed from 2007 till now @ Beijing
- Launched in more than 60 courtiers
  - Russia
  - HK
  - Southeast Asia
  - Arab World
  - Sub-Saharan Africa (Baraza)

Query: *What are must-see attractions at Yellowstone*



Query: *What are must-see attractions at Yellowstone*



Query: *What are must-see attractions at Yosemite*

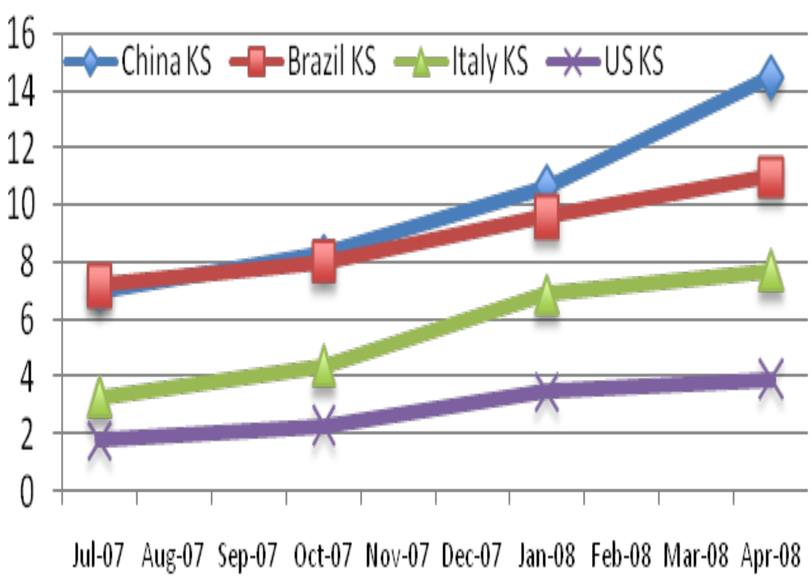
Query: *What are must-see attractions at Beijing*

Strategic Issue

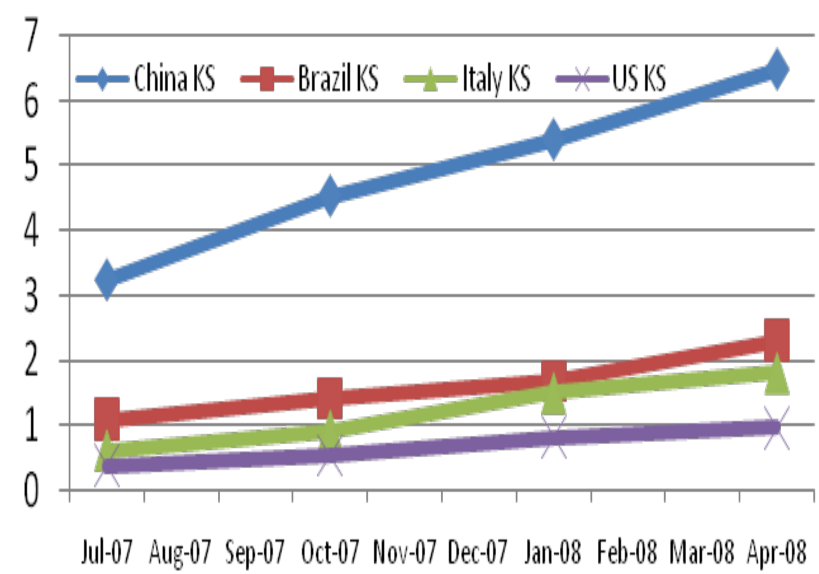
Search Quality at Stake.

61 countries have Q&A or advanced forums as top 10 most clicked destination
(out of 115 countries with more than 1M session)

% of First Result Page with >=1 Q&A Result from Yahoo or Baidu



% of Referral Traffic From 1st Page Sent to Yahoo / Baidu

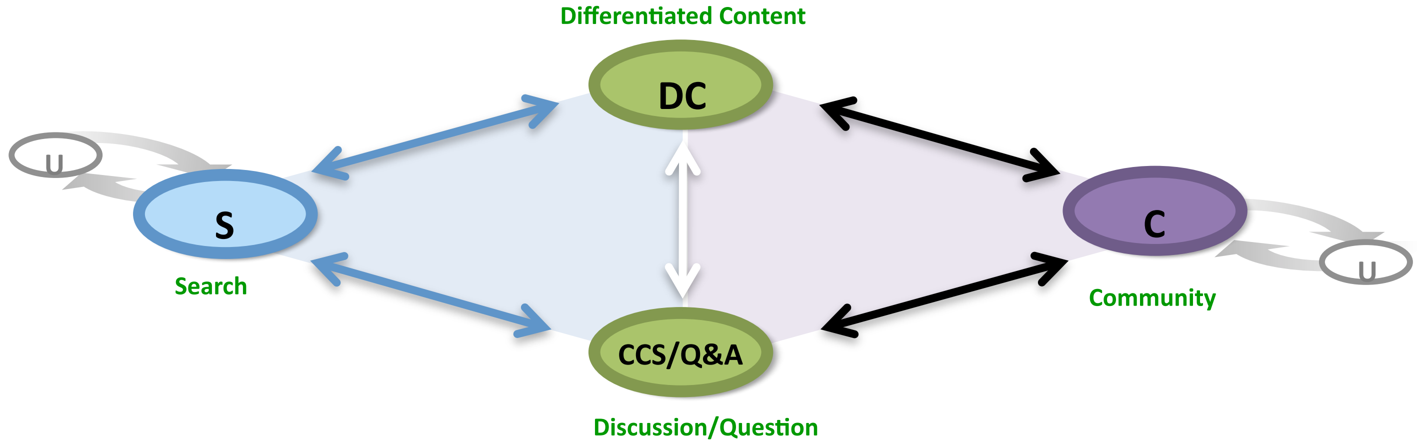


SNS & Mobile Also Need Q&A

- Social Networks
  - Difficult to find user intent to match ads
  - Q&A is a perfect app to learn users' problems
- Mobile Search
  - Voice is the most convenient user interface
  - Succinct search result (or rich snippets) is desirable

Confucius: Google Q&A

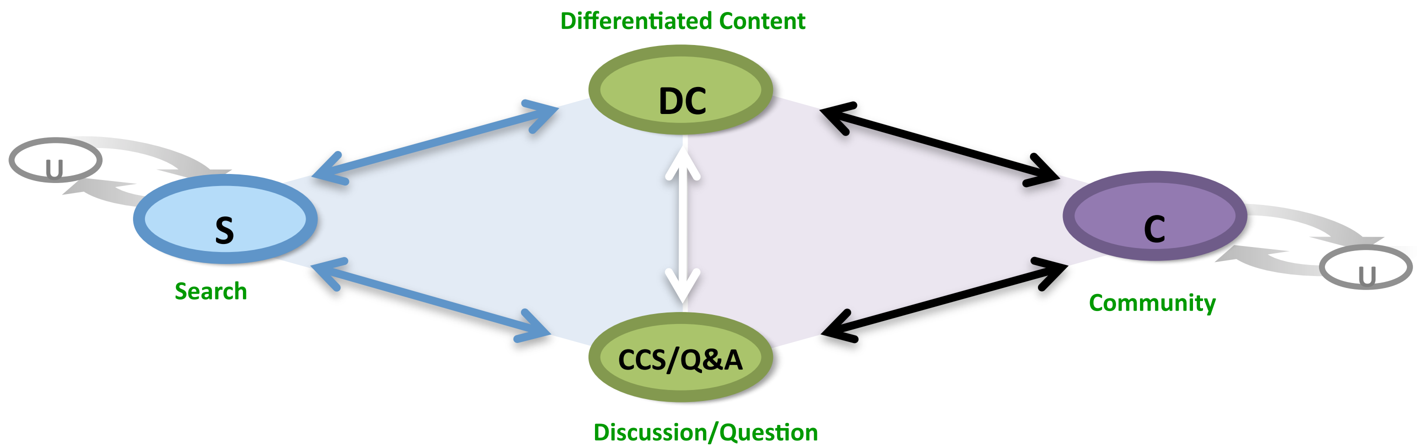
Providing High-Quality Answers in a Timely Fashion



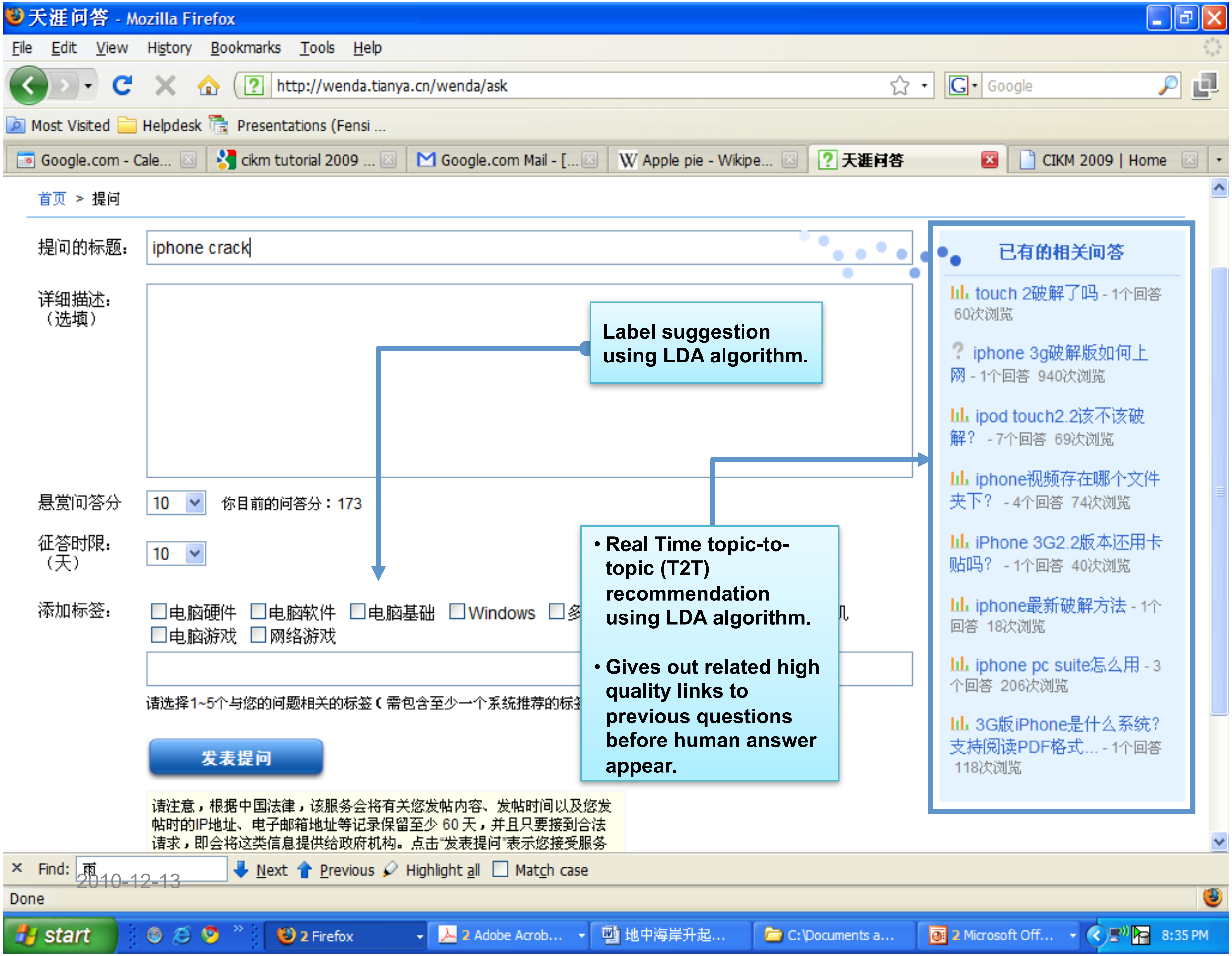
- Trigger a discussion/question session during search
- Provide labels to a post (semi-automatically)
- Given a post, find similar posts (automatically)
- Evaluate quality of a post, relevance and originality
- Evaluate user credentials in a topic sensitive way
- Route questions to experts
- Provide most relevant, high-quality content for Search to index
- Generate answers using NLP

Confucius: Google Q&A

Providing High-Quality Answers in a Timely Fashion



- ❑ Trigger a discussion/question session during search
- ❑ Provide labels to a question (semi-automatically)
- ❑ Given a question, find similar questions (automatically)
- ❑ Evaluate quality of an answer, relevance and originality
- ❑ Evaluate user credentials in a topic sensitive way
- ❑ Route questions to experts
- ❑ Provide most relevant, high-quality content for Search to index
- ❑ Generate answers using NLP


天涯问答 - Mozilla Firefox
http://wenda.tianya.cn/wenda/ask
首页 > 提问
提问的标题： iphone crack
详细描述：（选填）
Label suggestion using LDA algorithm.
悬赏问答分 10 你目前的问答分：173
征答时限：（天） 10
添加标签： 电脑硬件 电脑软件 电脑基础 Windows 电脑游戏 网络游戏
请选择1~5个与您的问题相关的标签（需包含至少一个系统推荐的标签
发表提问
• Real Time topic-to-topic (T2T) recommendation using LDA algorithm.
• Gives out related high quality links to previous questions before human answer appear.
已有的相关问答
touch 2破解了吗 - 1个回答 60次浏览
iphone 3g破解版如何上网 - 1个回答 940次浏览
ipod touch2.2该不该破解？ - 7个回答 69次浏览
iphone视频存在哪个文件夹下？ - 4个回答 74次浏览
iPhone 3G2.2版本还用卡贴吗？ - 1个回答 40次浏览
iphone最新破解方法 - 1个回答 18次浏览
iphone pc suite怎么用 - 3个回答 206次浏览
3G版iPhone是什么系统？支持阅读PDF格式... - 1个回答 118次浏览
请注意，根据中国法律，该服务会将有关您发帖内容、发帖时间以及您发帖时的IP地址、电子邮箱地址等记录保留至少 60 天，并且只要接到合法请求，即会将这类信息提供给政府机构。点击"发表提问"表示您接受服务
2010-12-13

Collaborative Filtering

Based on *membership* so far,
and *memberships* of others

↓

Predict further *membership*

Labels/Qs

Questions

		1	1	1						
	1		1	1		1		1		1
					1		1			1
	1		1		1	1				
		1								
						1	1			
			1					1		
1	1									
	1								1	
1										1
	1	1	1	1	1					

FIM-based Recommendation

To grow the base, we need association rules

- An association rule: $a, b, c \longrightarrow d$
- A Bayesian interpretation: $P(d \mid a, b, c) = \frac{N(a,b,c,d)}{N(a,b,c)}$
- The key is to count the occurrences (*support*) of itemsets $N(\ldots)$

Distributed Latent Dirichlet Allocation (LDA)

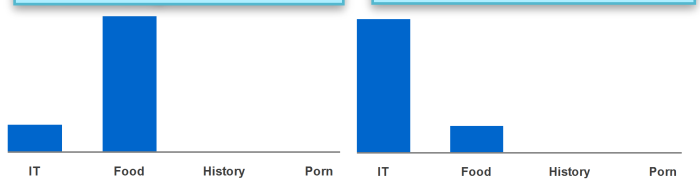
- Search
  - Construct a latent layer for better for semantic matching
- Example:
  - iPhone crack
  - Apple pie

Documents

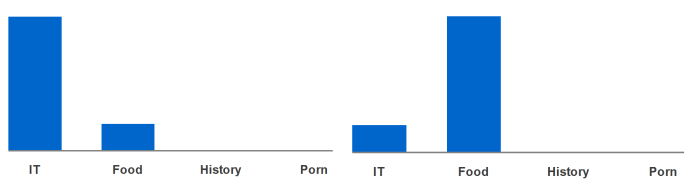
1 recipe pastry for a 9 inch double crust 9 apples, 2/1 cup, brown sugar

How to install apps on Apple mobile phones?

Topic Distribution



Topic Distribution



User quries

iPhone crack

Apple pie

Users/Music/Ads/Question

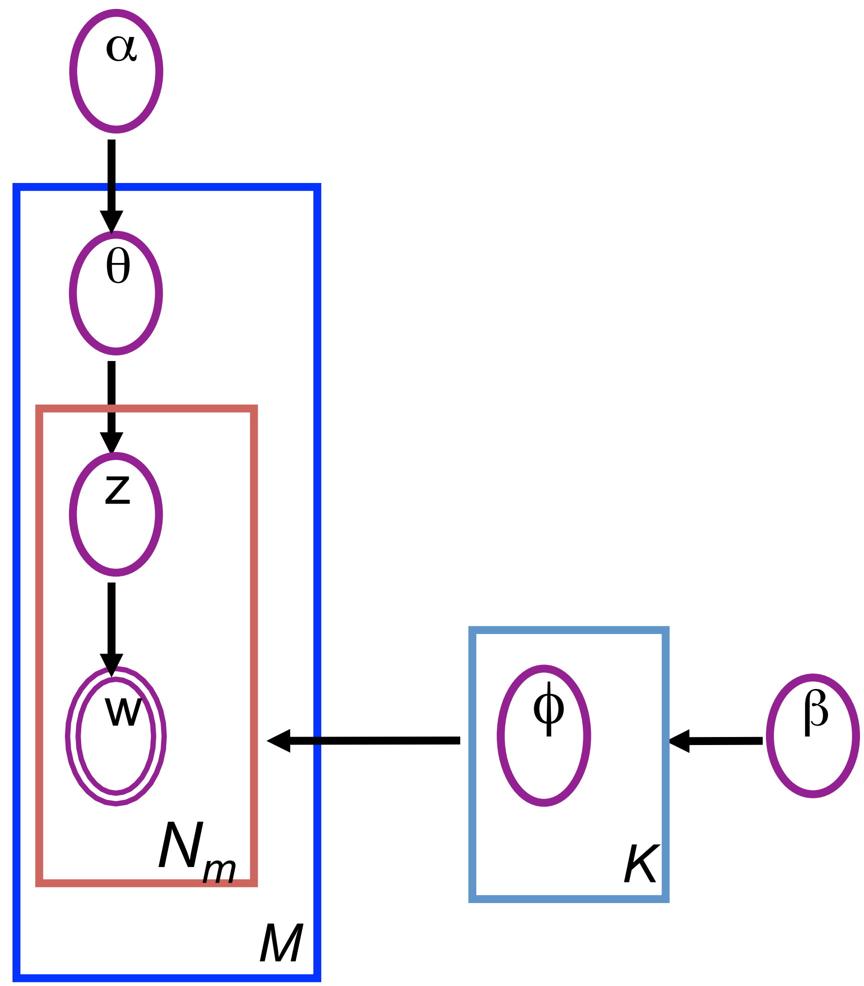
Users/Music/Ads/Answers

?	?	1	3	1	?	?	?	?	?	?
?	2	?	1	2	?	1	?	3	?	1
?	?	?	?	?	1		5			1
	5		3		1	1				
		1								
						1	4			
			2					1		
1	2									
	1								5	
1										1
	1	4	1	3	6					

- Other Collaborative Filtering Apps
  - Recommend Users → Users
  - Recommend Music → Users
  - Recommend Ads → Users
  - Recommend Answers → Q
- Predict the ? In the light-blue cells

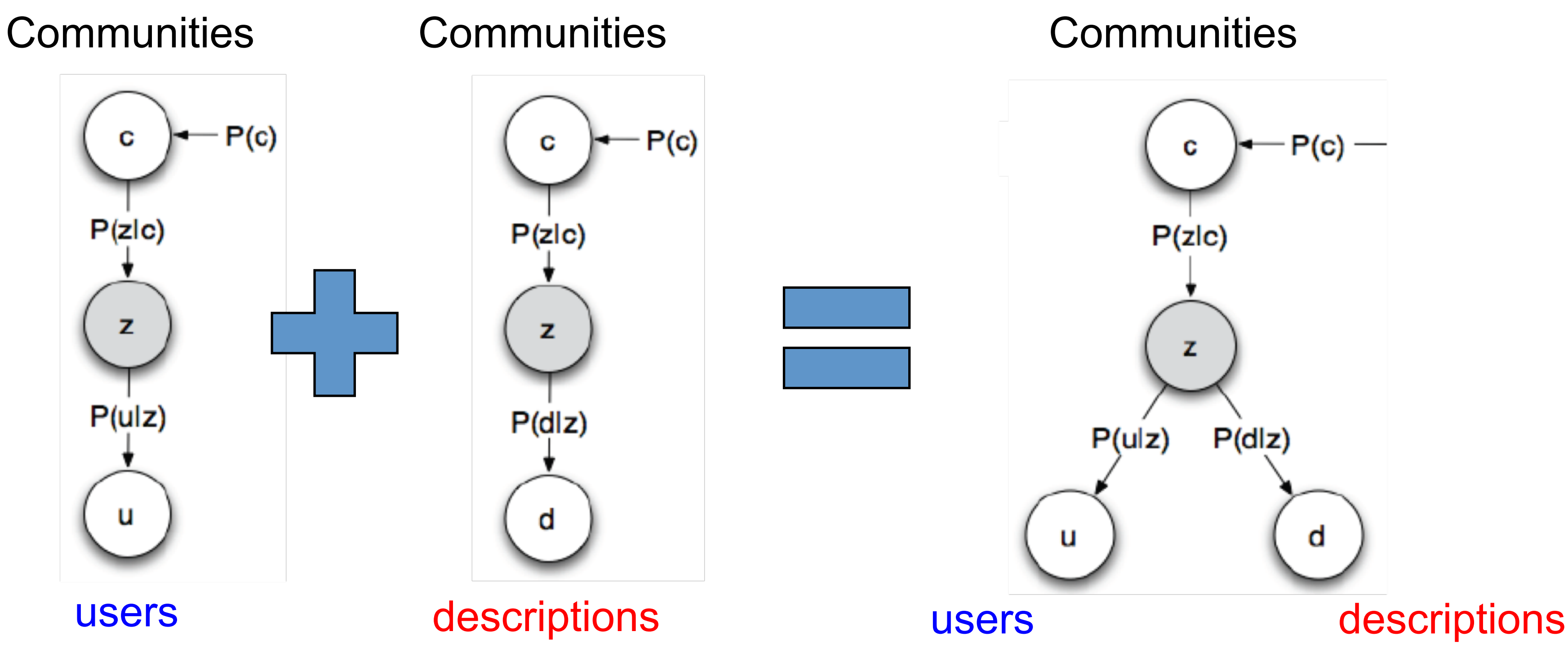
Latent Dirichlet Allocation [D. Blei, M. Jordan 04]

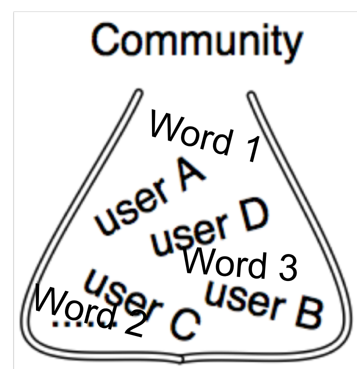
- α: uniform Dirichlet ϕ prior for per document d topic distribution (corpus level parameter)
- β: uniform Dirichlet ϕ prior for per topic z word distribution (corpus level parameter)
- θ_d is the topic distribution of document d (document level)
- z_{dj} the topic if the j^{th} word in d, w_{dj} the specific word (word level)

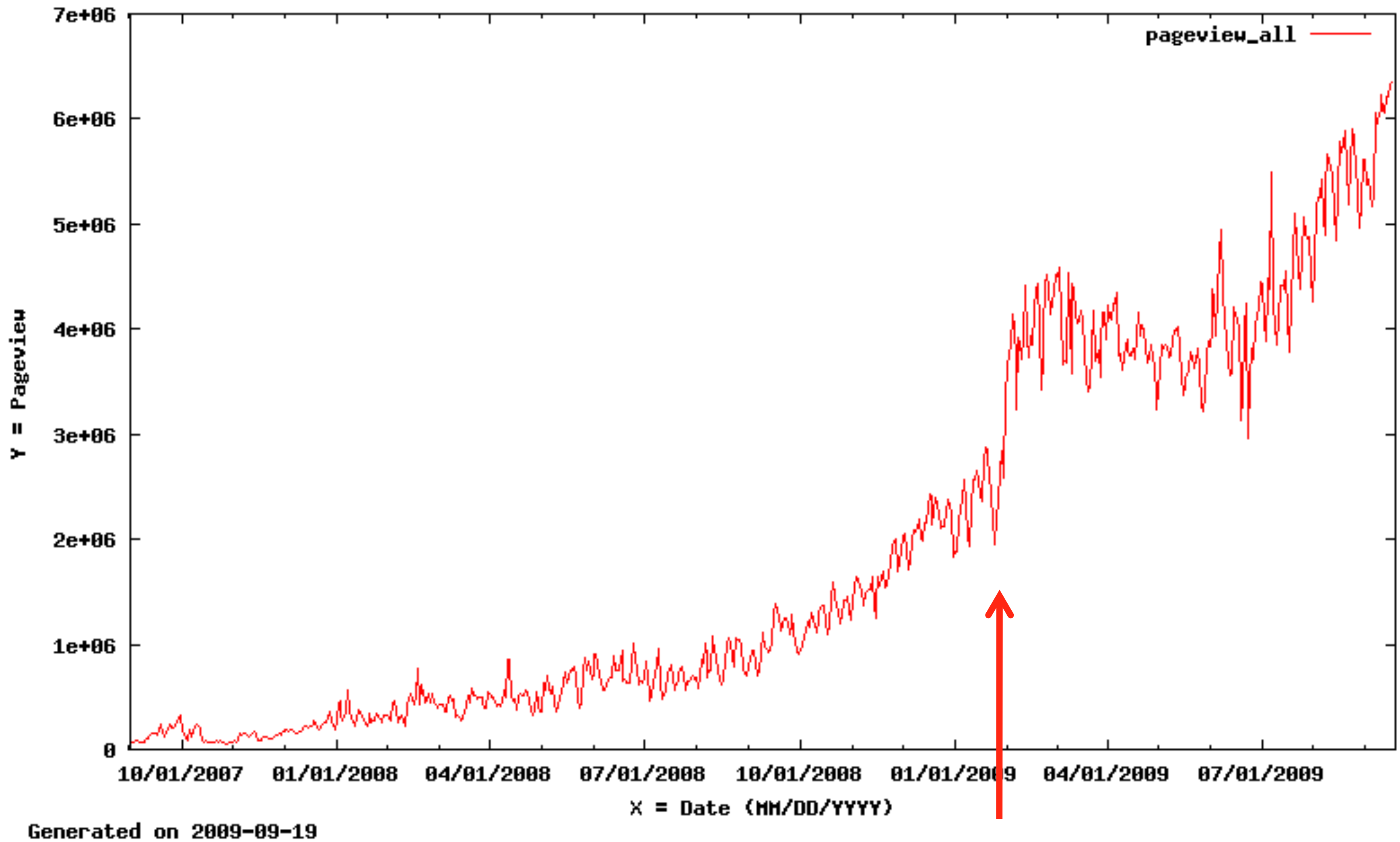


Combinational Collaborative Filtering Model (CCF)

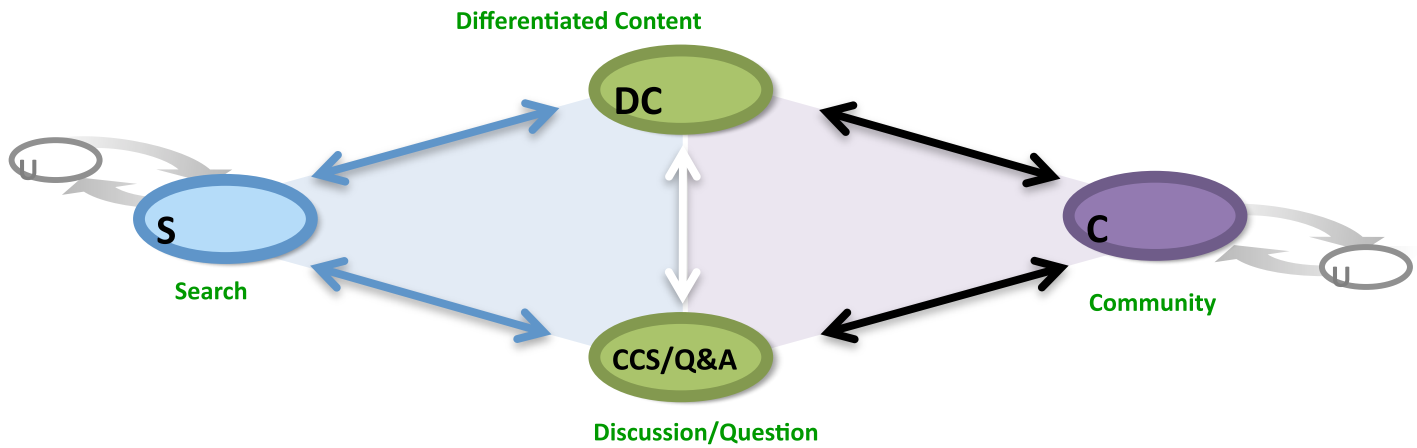
[W.-Y. Chen, et al, KDD2008]





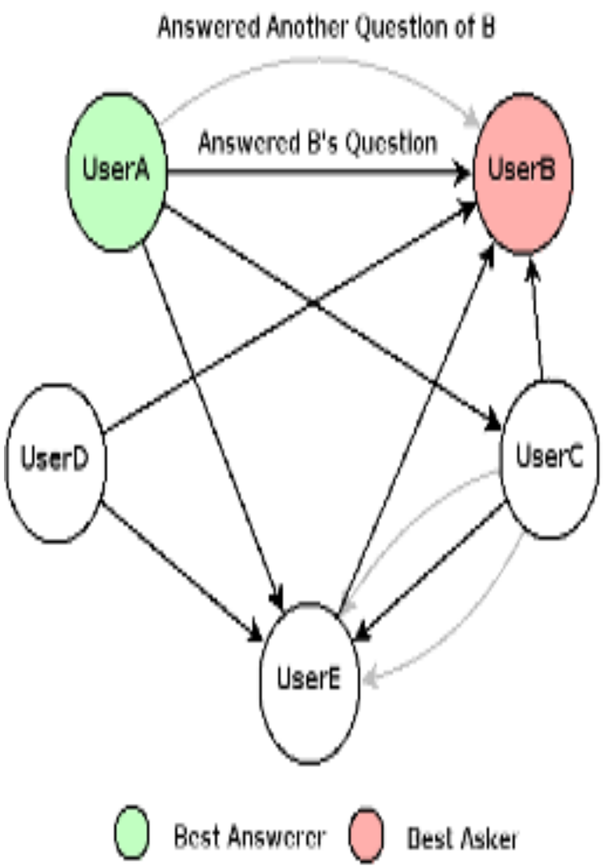

pageview_all
7e+06
6e+06
5e+06
4e+06
3e+06
2e+06
1e+06
0
Y = Pageview
10/01/2007
01/01/2008
04/01/2008
07/01/2008
10/01/2008
01/01/2009
04/01/2009
07/01/2009
X = Date (MM/DD/YYYY)
Generated on 2009-09-19

Confucius: Google Q&A



- Trigger a discussion/question session during search
- Provide labels to a post (semi-automatically)
- Given a post, find similar posts (automatically)
- Evaluate quality of a post, relevance and originality
- Evaluate user credentials in a topic sensitive way
- Route questions to experts
- Provide most relevant, high-quality content for Search to index
- NLQA

UserRank



- Rank users by quantity **(number of links)** and quality **(weights on links)** of contributions

  Quality include:
  - **Relevance.** Is an answer relevant to the Q? Measured by KL divergence between *latent-topic vectors* of A and Q
  - **Coverage.** Compared among different answers
  - **Originality.** Detect potential plagiarism and spam
  - **Promptness.** Time between Q and A posting time

Outline

- Search + Social Synergy
- Social → Search
- Search → Social
- Scalability

Outline

- Search + Social Synergy
- Social → Search
- Search → Social
- Scalability

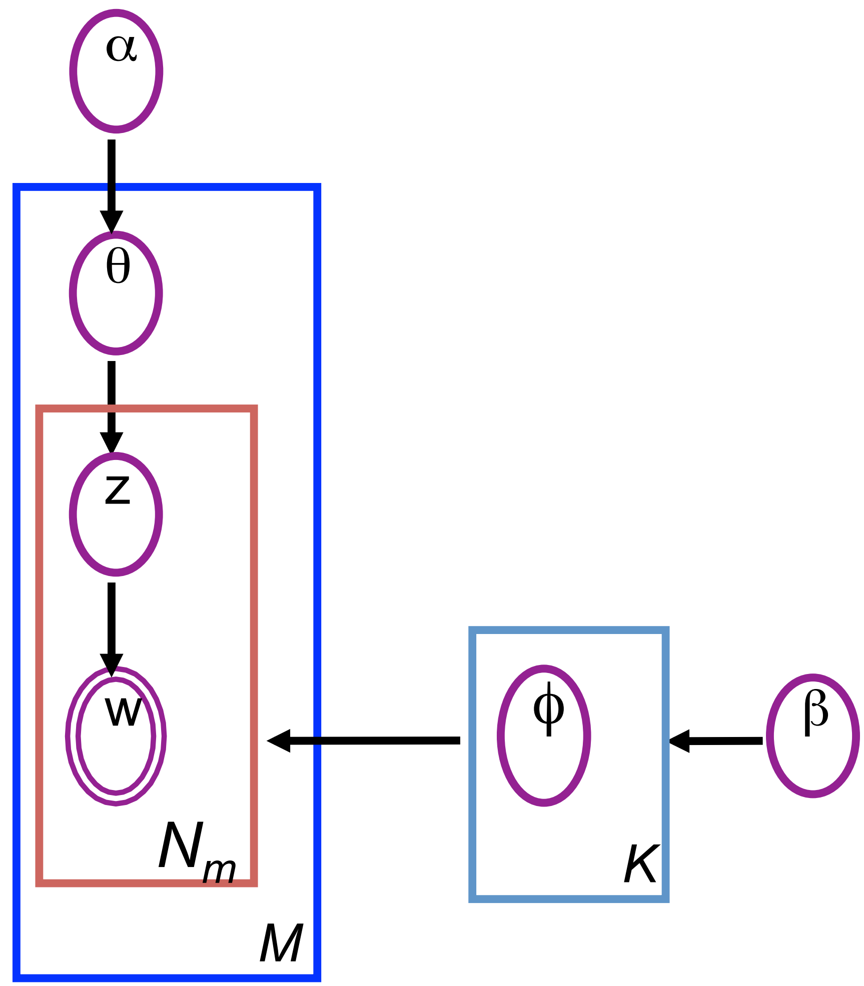
Social?

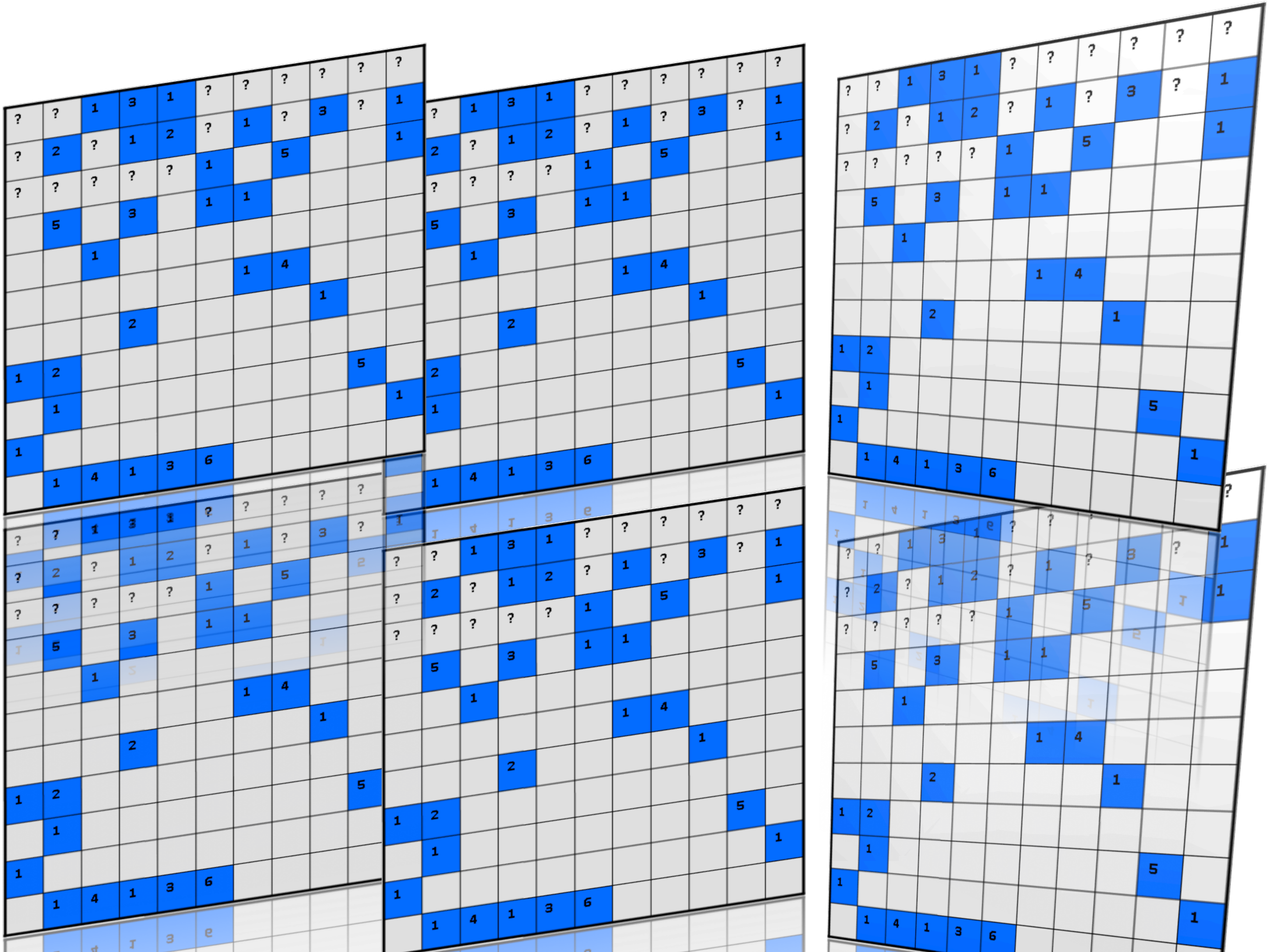
- Connecting to friends
- Knowing what friends are up to
- Connecting to strangers
  - Dating, Gaming
  - Shopping

- Making recommendations based on activities

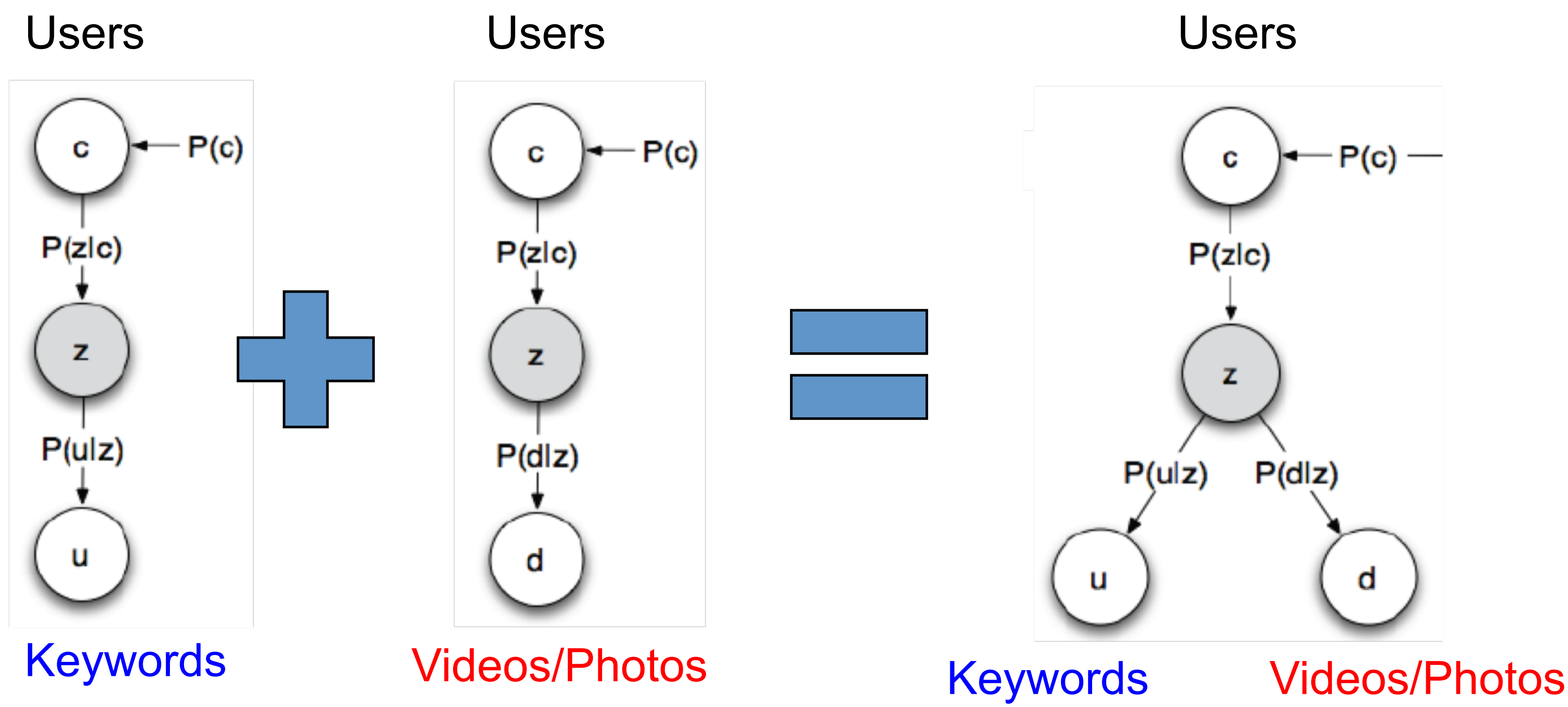
User Latent Model

- α: uniform Dirichlet ϕ prior for per user u interest distribution (population level parameter)
- β: uniform Dirichlet ϕ prior for per interest z activity distribution (population level parameter)
- θ_d is the interest distribution of user u (user level)
- z_{uj} the interest of the j^{th} activity in u, w_{uj} the specific activity (activity level)

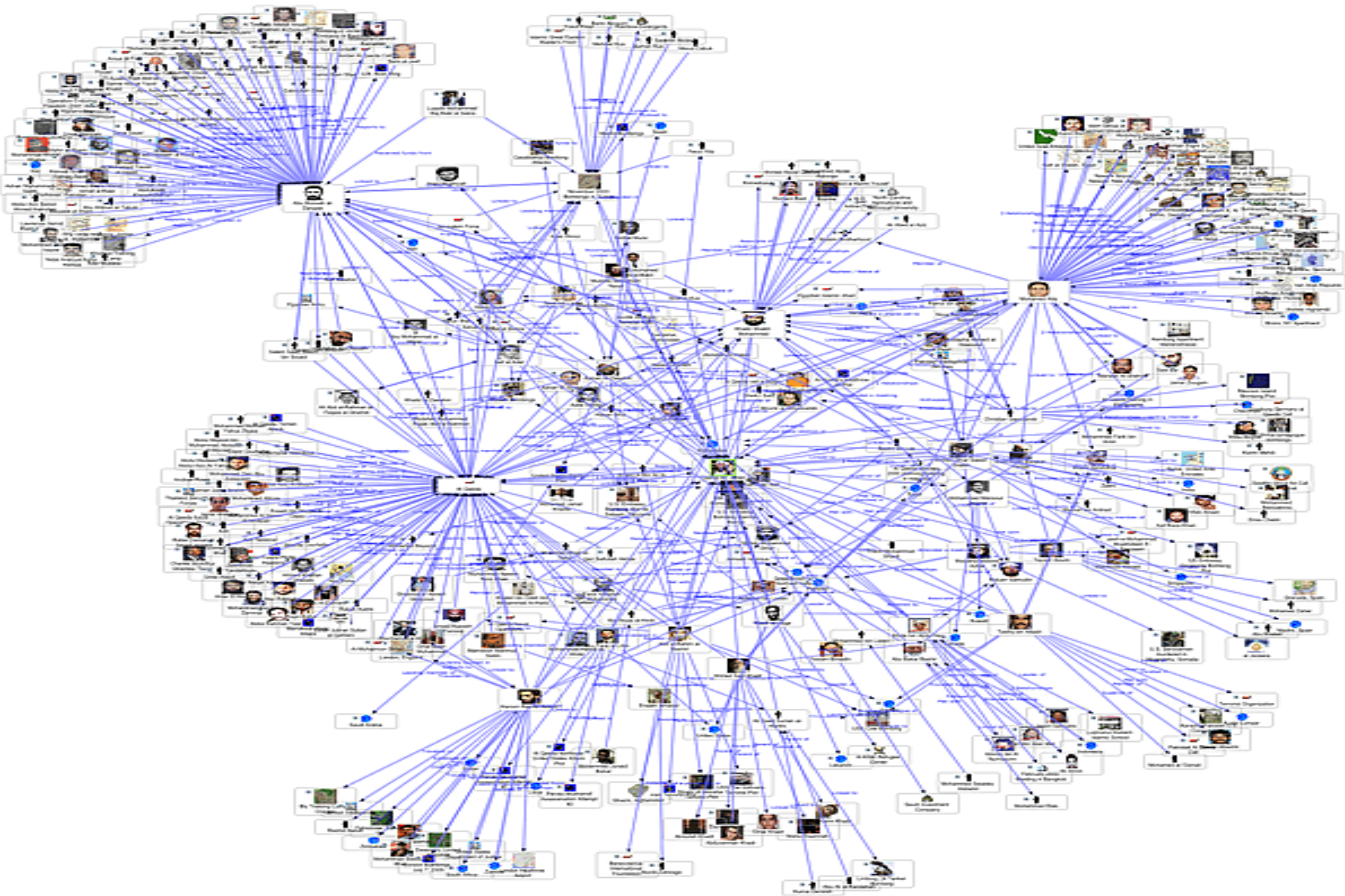




Combinational Collaborative Filtering Model (CCF)



Interest Networks



Outline

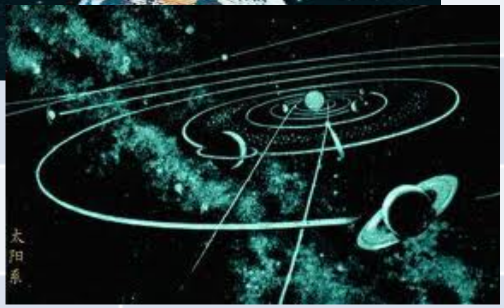
- Search + Social Synergy
- Social → Search
  - Mobilize users to improve search quality
  - Google *Q&A*, Facebook *Like*
- Search → Social
  - Use query log to help social
    - Activities → Interests → Social
    - Groupcom
- Scalability

Prefixes

SI prefix	Name	Power of 10 or 2	
k kilo	thousand	10^{3} 2^{10}	
M mega	million	10^{6} 2^{20}	
G giga	billion	10^{9} 2^{30}	
T tera	trillion	10^{12} 2^{40}	
P peta	quadrillion	10^{15} 2^{50}	
E exa	quintillion	10^{18} 2^{60}	
Z zetta	sextillion	10^{21} 2^{70}	
Y yotta	septillion	10^{24} 2^{80}	

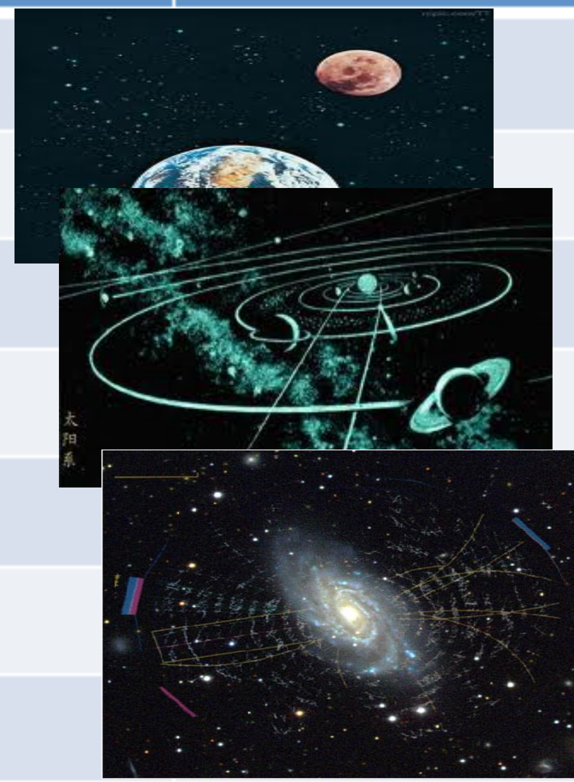
Prefixes

SI prefix	Name	Power of 10 or 2	
k kilo	thousand	10^{3} 2^{10}	
M mega	million	10^{6} 2^{20}	
G giga	billion	10^{9} 2^{30}	
T tera	trillion	10^{12} 2^{40}	
P peta	quadrillion	10^{15} 2^{50}	
E exa	quintillion	10^{18} 2^{60}	
Z zetta	sextillion	10^{21} 2^{70}	
Y yotta	septillion	10^{24} 2^{80}	



Prefixes

SI prefix	Name	Power of 10 or 2	
k kilo	thousand	10^{3} 2^{10}	
M mega	million	10^{6} 2^{20}	
G giga	billion	10^{9} 2^{30}	
T tera	trillion	10^{12} 2^{40}	
P peta	quadrillion	10^{15} 2^{50}	
E exa	quintillion	10^{18} 2^{60}	
Z zetta	sextillion	10^{21} 2^{70}	
Y yotta	septillion	10^{24} 2^{80}	



More Data vs. Better Algorithms

Banko & Brill, 2001

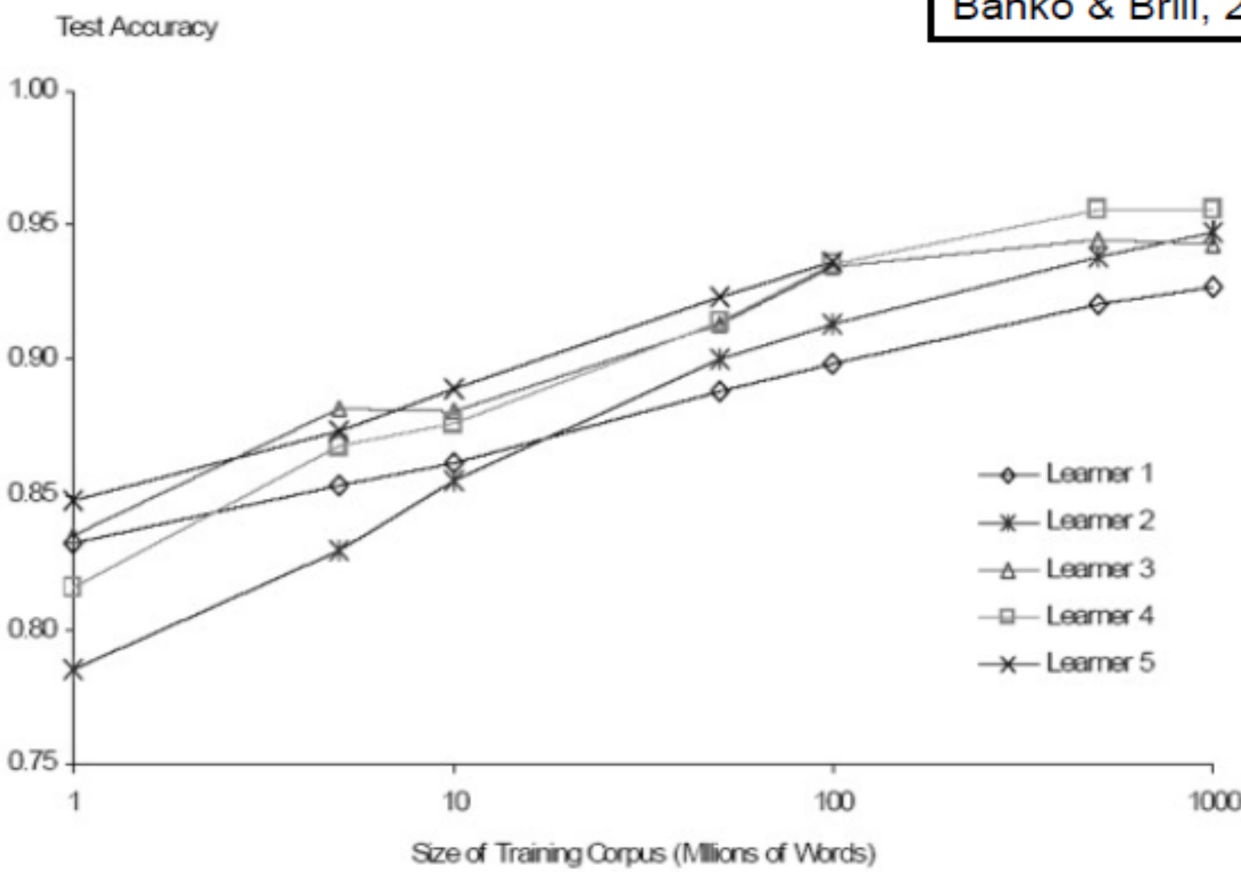


Figure 2. Learning Curves for Confusable Disambiguation

More Data vs. Better Algorithms

Banko & Brill, 2001

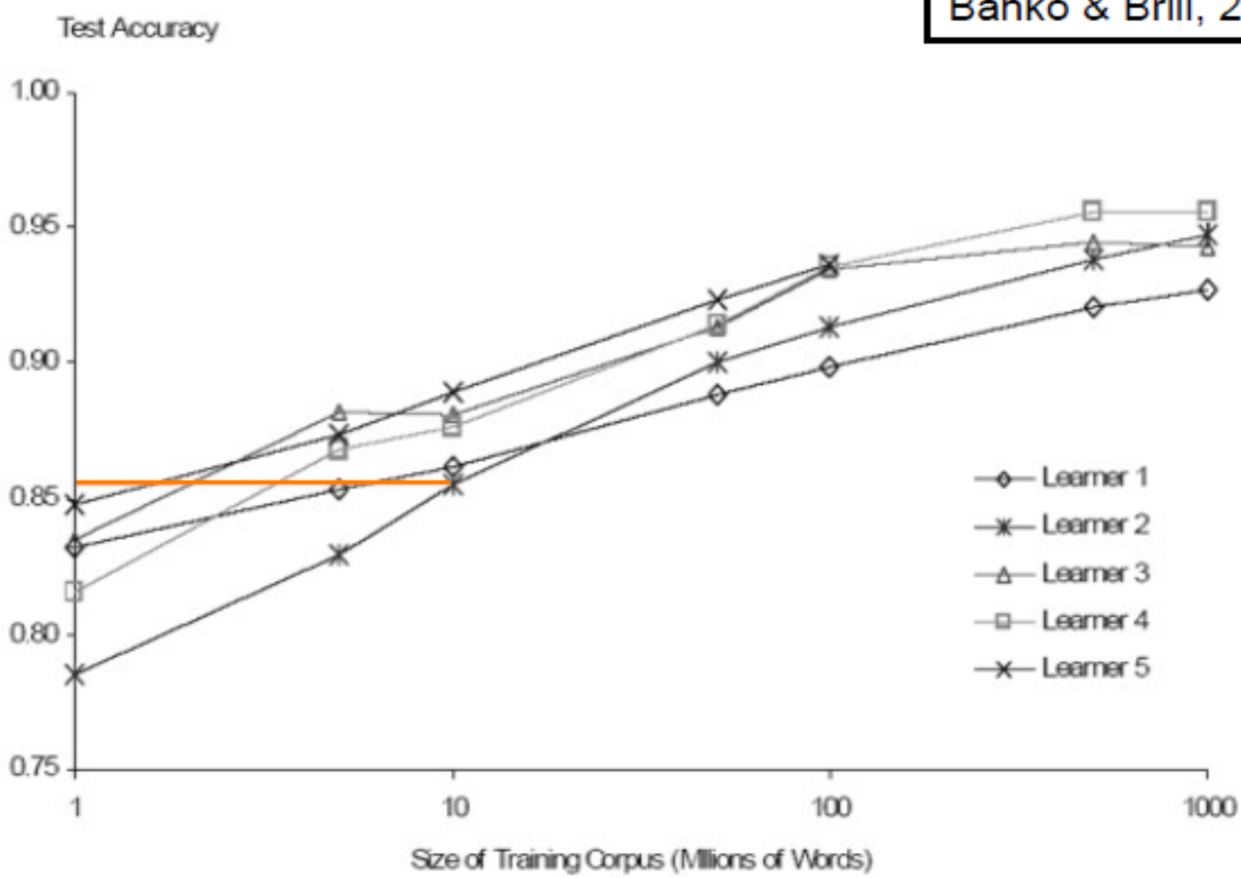


Figure 2. Learning Curves for Confusable Disambiguation

More Data vs. Better Algorithms

Banko & Brill, 2001

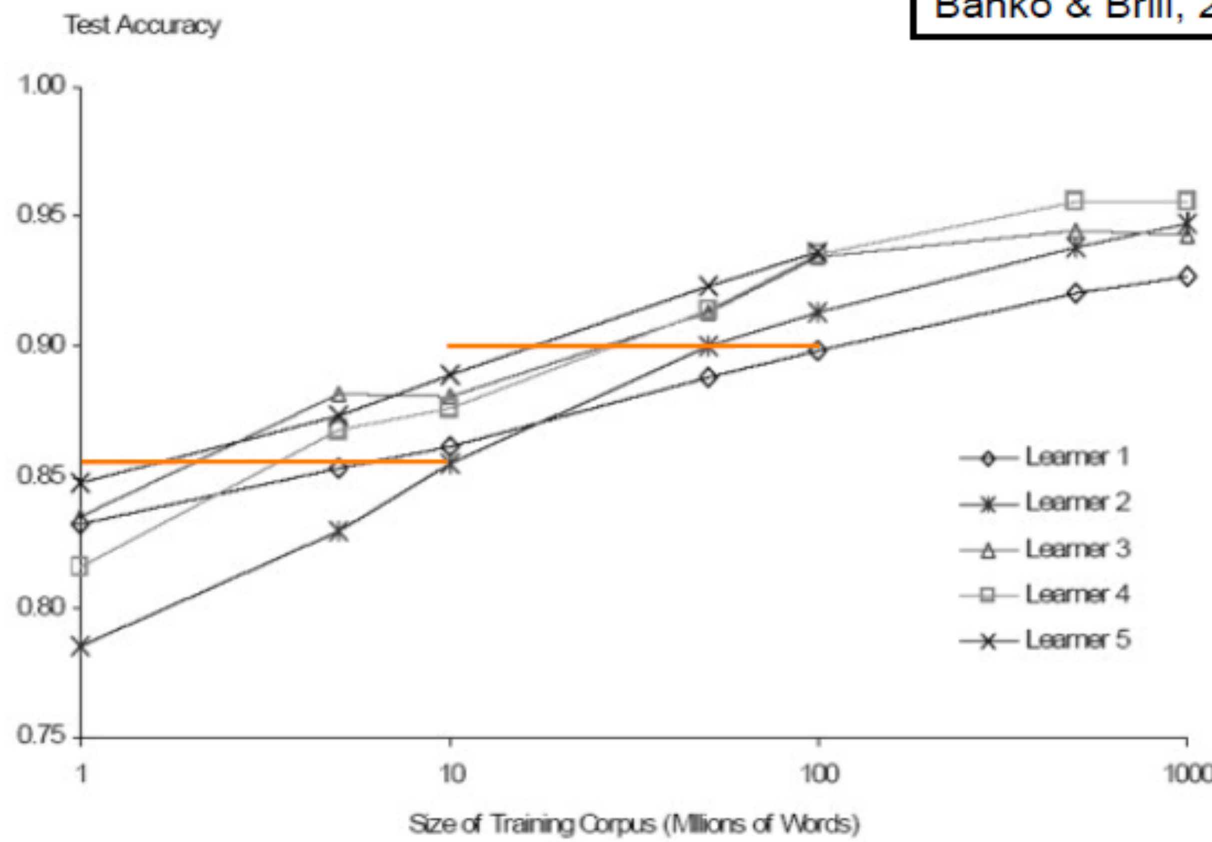


Figure 2. Learning Curves for Confusable Disambiguation

More Data vs. Better Algorithms

Banko & Brill, 2001

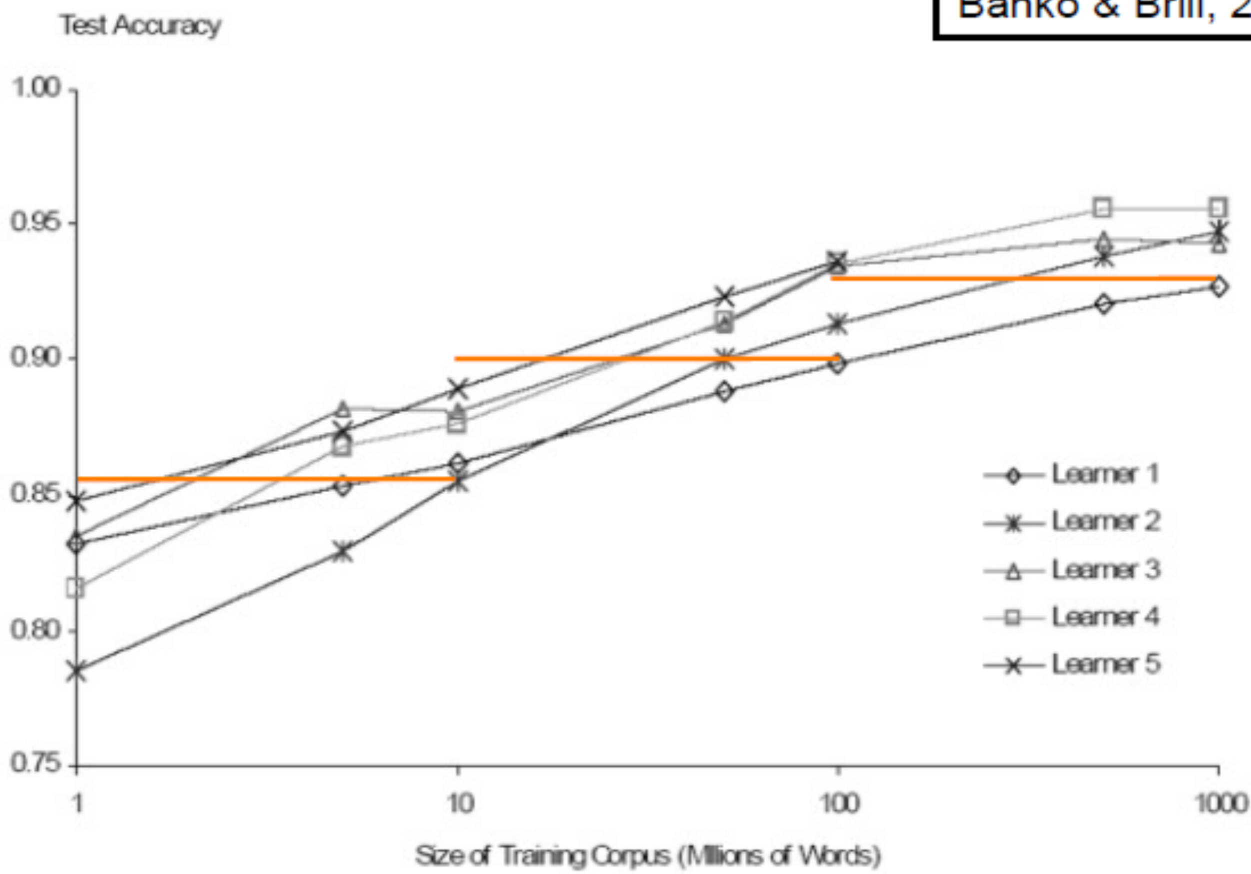
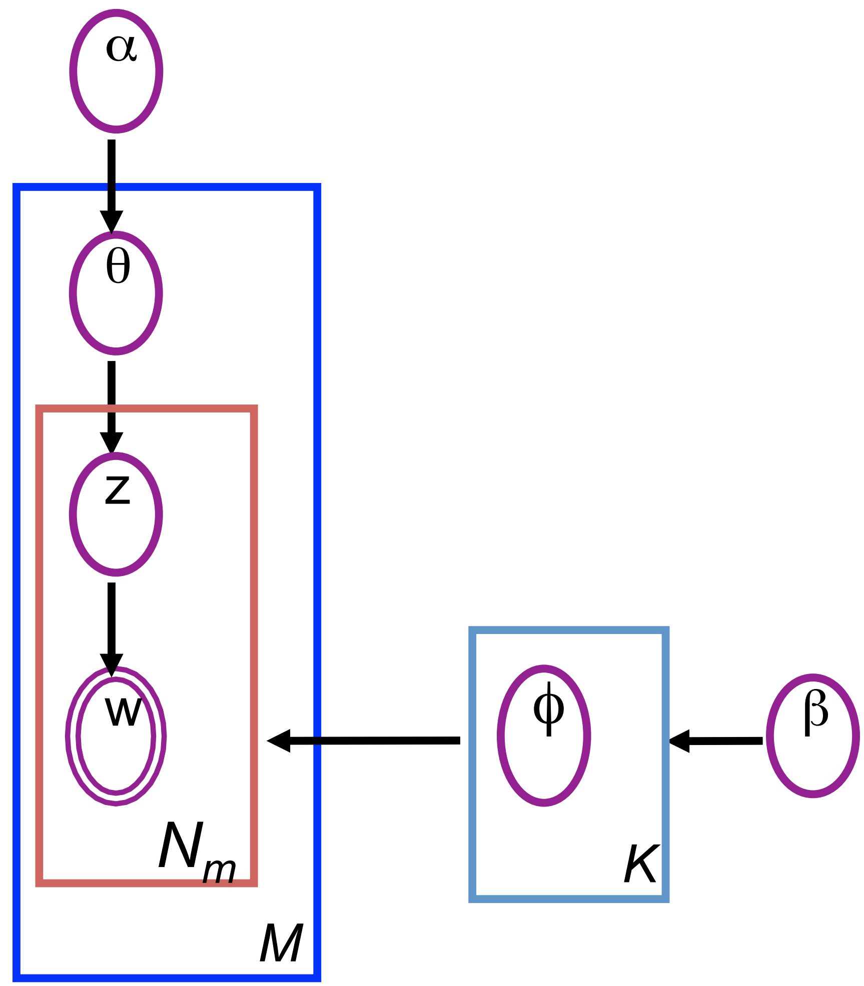


Figure 2. Learning Curves for Confusable Disambiguation

User Latent Model

- α: uniform Dirichlet ϕ prior for per user u interest distribution (population level parameter)
- β: uniform Dirichlet ϕ prior for per interest z activity distribution (population level parameter)
- θ_d is the interest distribution of user u (user level)
- z_{uj} the interest of the j^{th} activity in u, w_{uj} the specific activity (activity level)



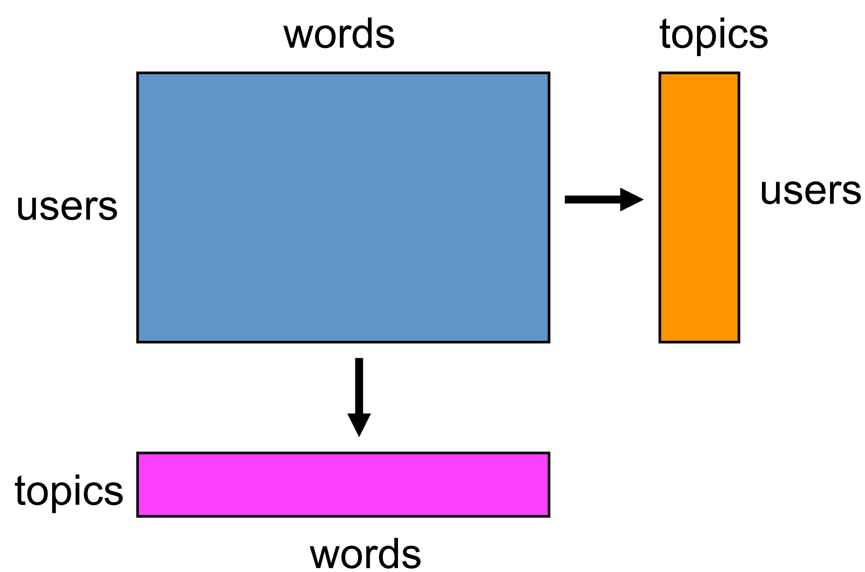
LDA Gibbs Sampling: Inputs & Outputs

Inputs:

1. training data: users as bags of words
2. parameter: the number of topics

Outputs:

1. model parameters: a co-occurrence matrix of topics and words.
2. by-product: a co-occurrence matrix of topics and users.



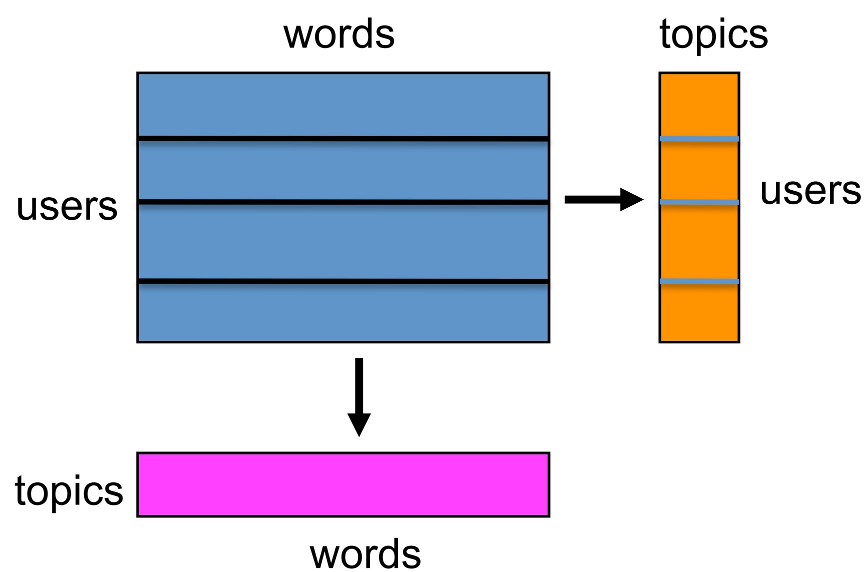
Parallel Gibbs Sampling

Inputs:

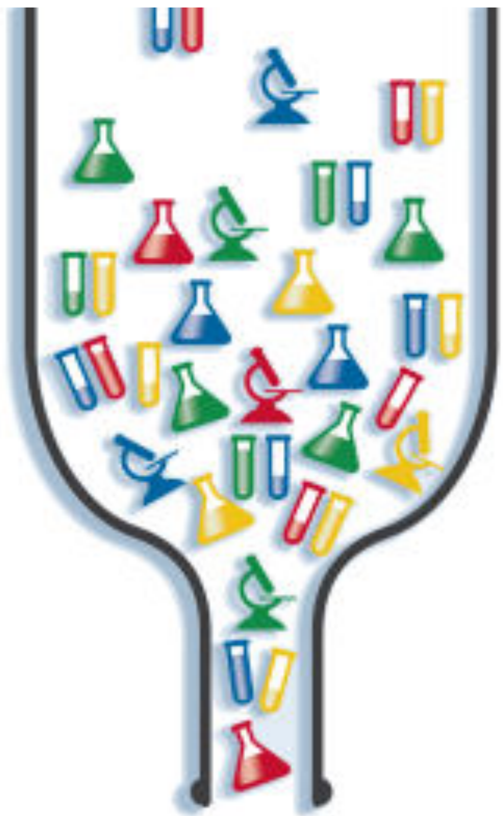
1. training data: users as bags of words
2. parameter: the number of topics

Outputs:

1. model parameters: a co-occurrence matrix of topics and words.
2. by-product: a co-occurrence matrix of topics and users.



Observations



- Bottleneck: Communication
- Amdahl's law caps speedup

- Words in a bag have no order
- Words on a computer node can be reordered

Example Bags / Node A

- Bag #1 w1, w2, w3, w1, w2, w3, w1, w2, w3
- Bag #2 w1, w2, w1, w2, w1, w2, w1, w2
- Bag #3 w3, w1, w3, w1, w3, w1, w3, w1

- Bundle #1 w1, w1, w1, w1, w1, w1, …
- Bundle #2 w2, w2, w2,… ,
- Bundle #3 w3, w3, w3,…

Two Nodes

Node A	Node B
W1	W2
W2	W3
W3	W1

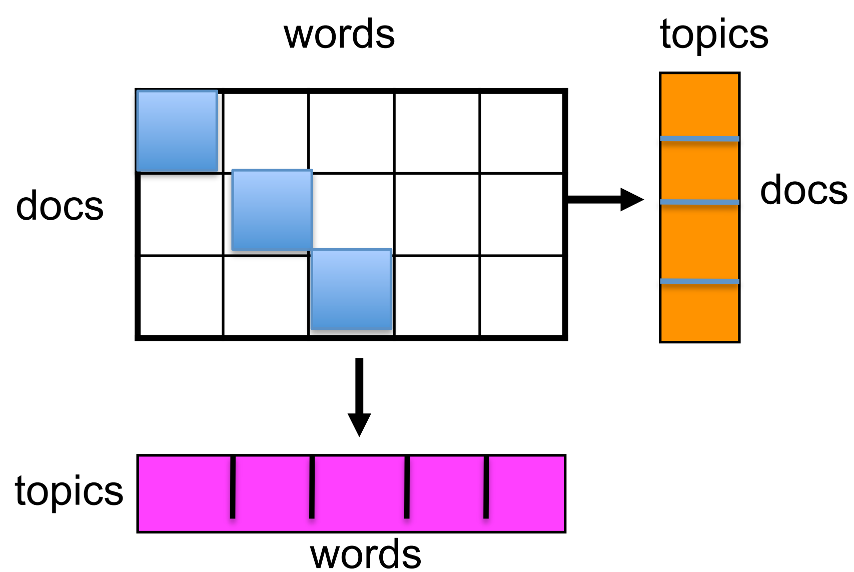
Parallel Gibbs Sampling

Inputs:

1. training data: documents as bags of words
2. parameter: the number of topics

Outputs:

1. model parameters: a co-occurrence matrix of topics and words.
2. by-product: a co-occurrence matrix of topics and documents.



PLDA -- enhanced parallel LDA

- Take advantage of bag of words modeling: each Pw machine processes vocabulary in a word order
- Pipelining: fetching the updated topic distribution matrix while doing Gibbs sampling

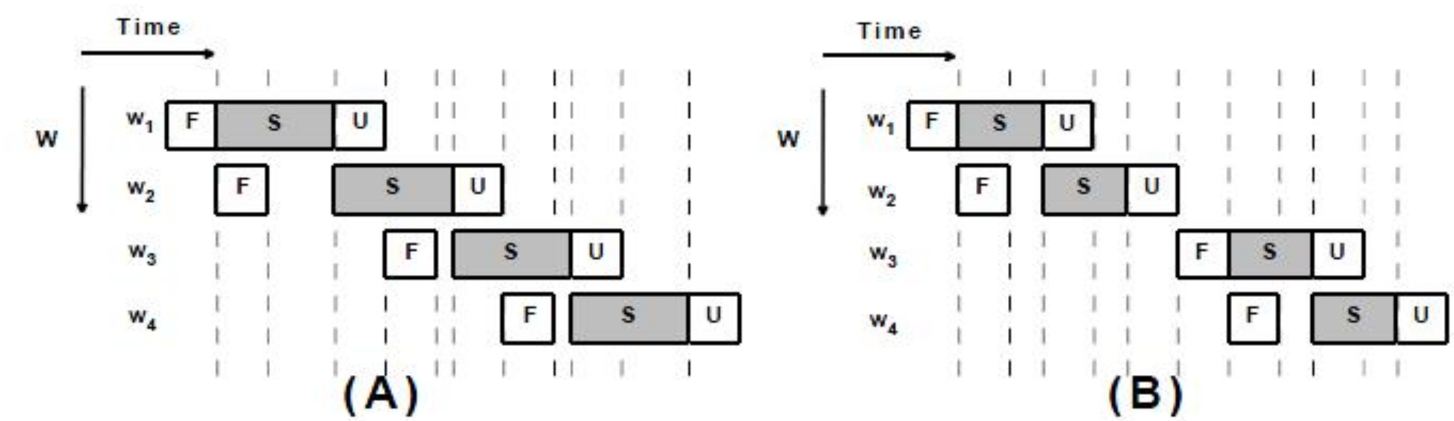
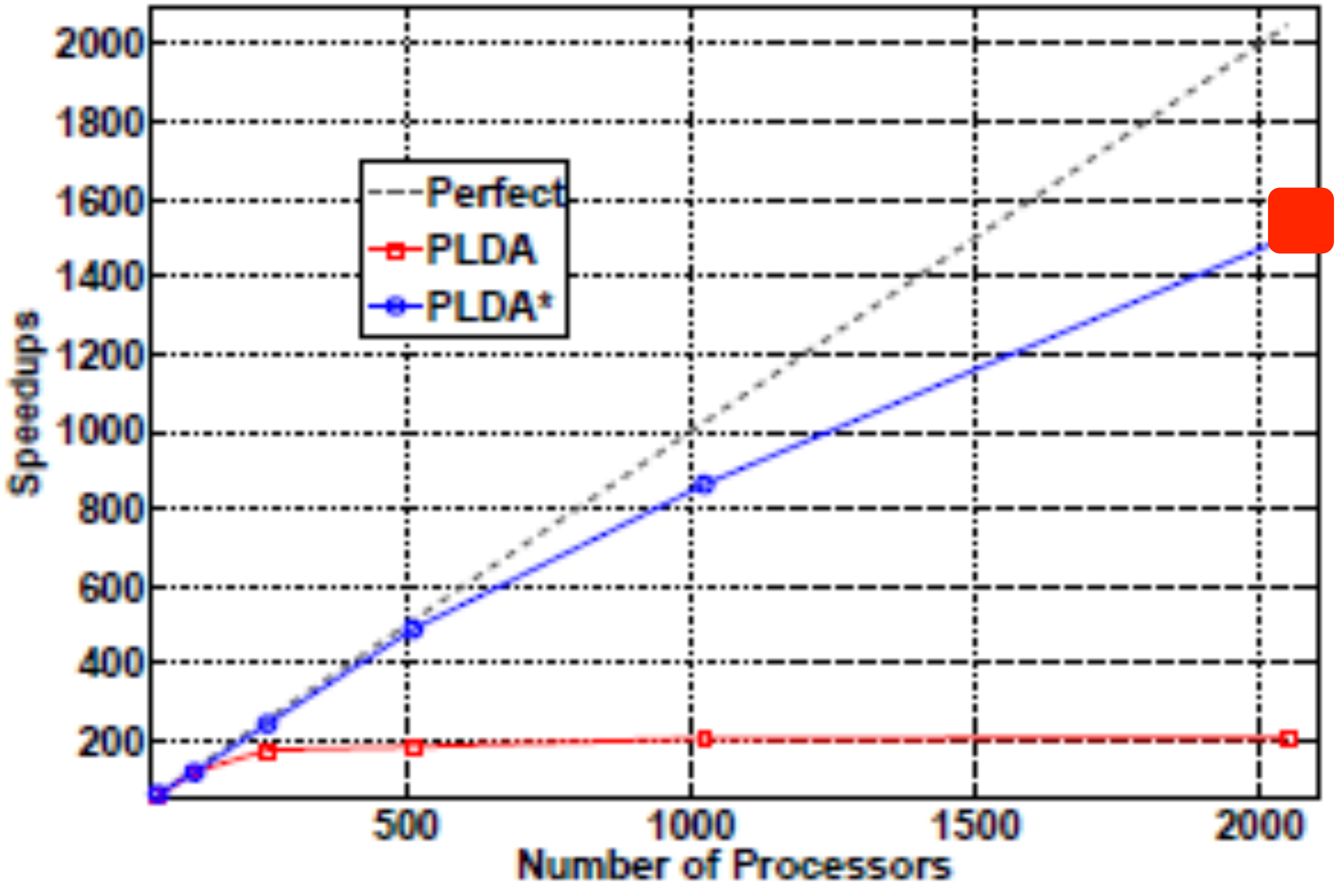


Fig. 4: Pipeline-based Gibbs Sampling in PLDA*. (A): $t_s \geq t_f + t_u$. (B): $t_s < t_f + t_u$.

Speedup

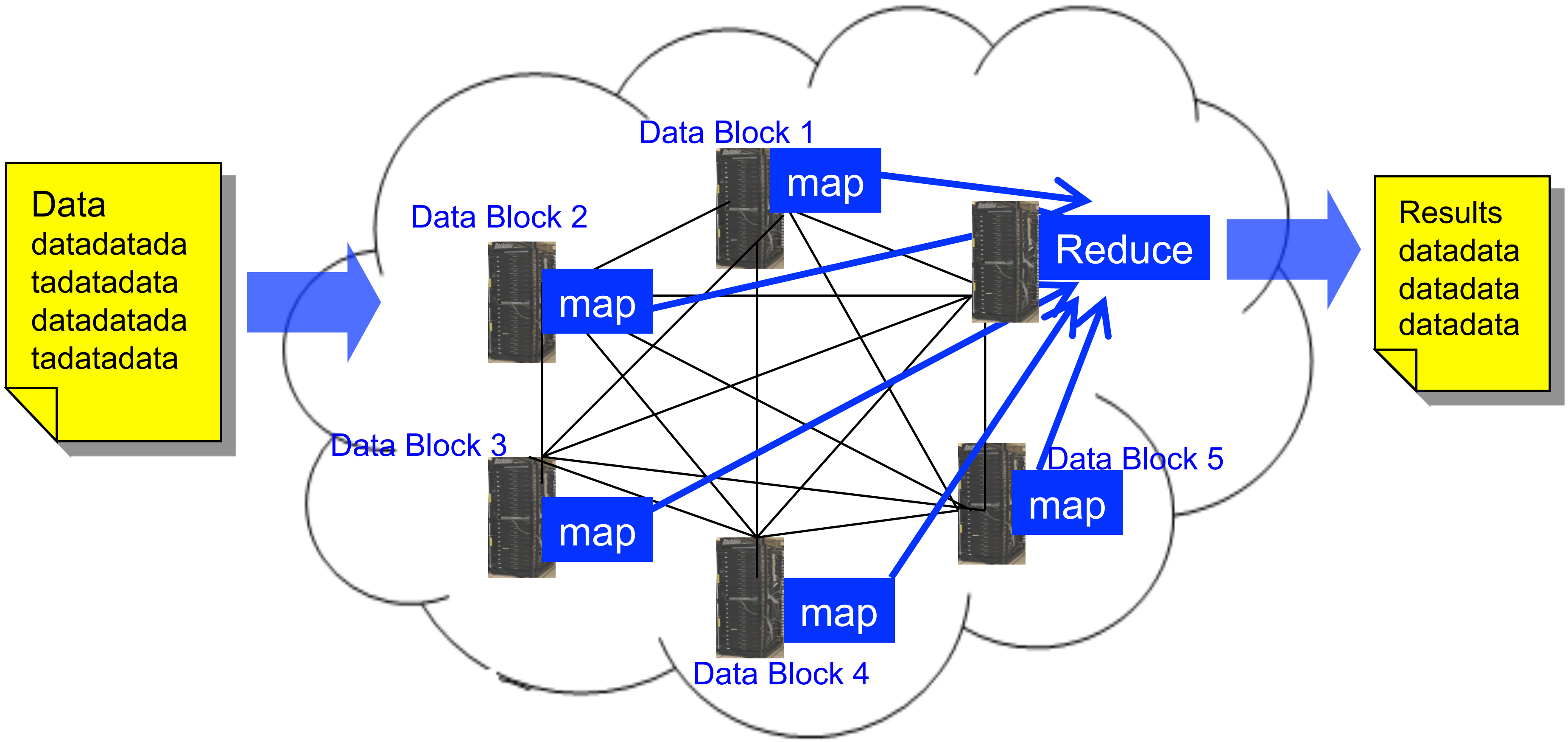
1,500x using 2,000 machines



Lessons Learned

- Bottleneck Matters
- Inter-iteration Matters

MapReduce



Data
datadatada
tadatadata
datadatada
tadatadata
Data Block 1
map
Data Block 2
map
Reduce
Data Block 3
map
map
Data Block 4
Data Block 5
map
Results
datadata
datadata
datadata

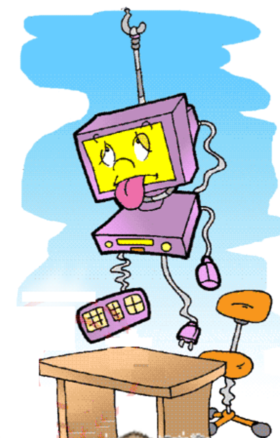
Parallel Programming Models

	MapReduce	Project +	MPI
GFS/IO and task rescheduling overhead between iterations	Yes	No +1	No +1
Flexibility of computation model	AllReduce only +0.5	? +1	Flexible +1
Efficient AllReduce	Yes +1	Yes +1	Yes +1
Recover from faults between iterations	Yes +1	Yes +1	Apps
Recover from faults within each iteration	Yes +1	Yes +1	Apps
Final Score for scalable machine learning	3.5	5	4

SVM Bottlenecks

Time consuming – 1M dataset, 8 days

... .
.



Memory consuming – 1M dataset, 10G

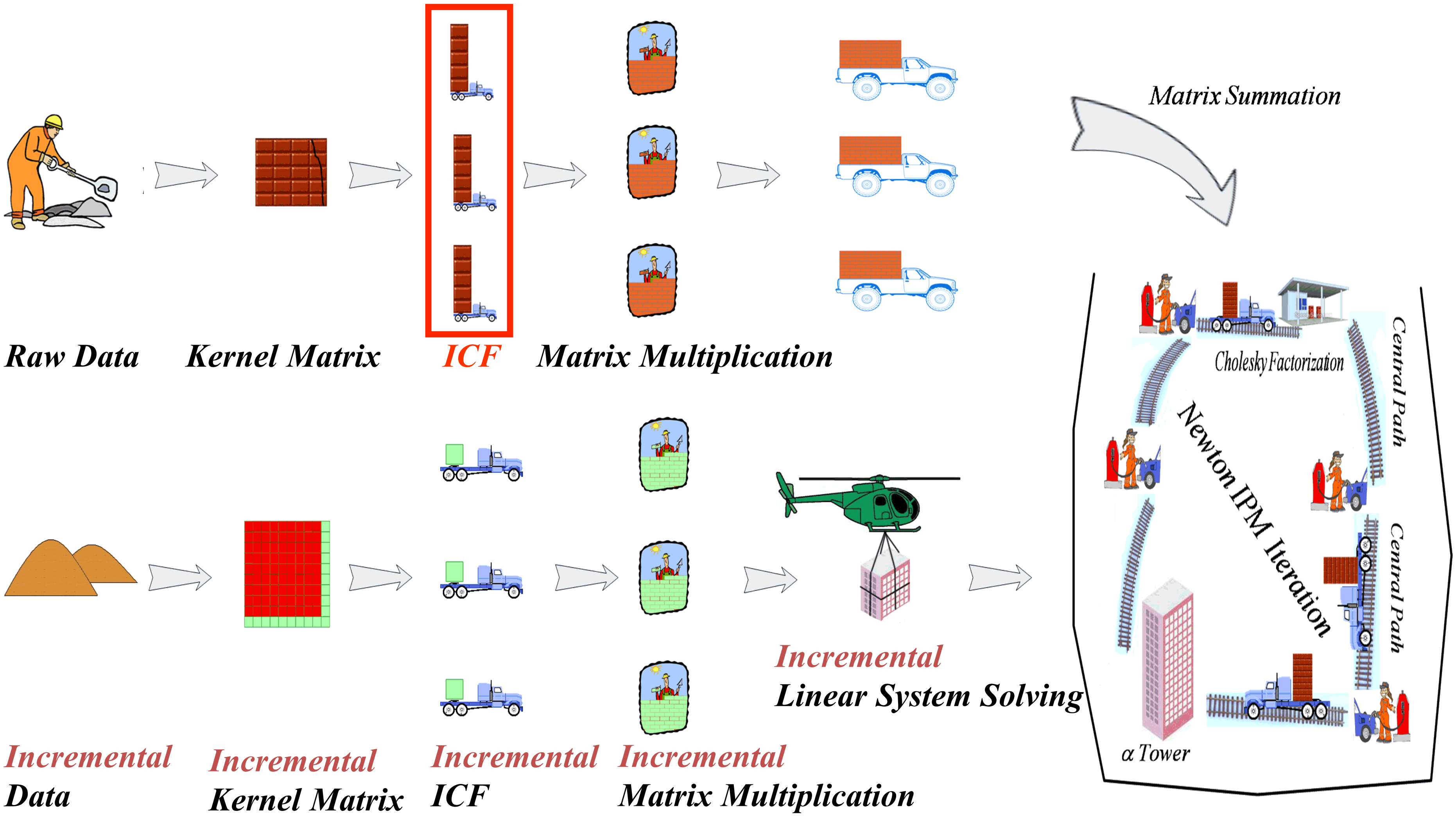
Matrix Factorization Alternatives

Factorization	Cost
QR	$O(\frac{4}{3}n^3)$
LU	$O(\frac{2}{3}n^3)$
Cholesky	$O(\frac{1}{3}n^3 + 2n^2)$
LDL^T	$O(\frac{1}{3}n^3)$
Incomplete Cholesky	$O(p^2n)$
Kronecker	$O(2n^2)$

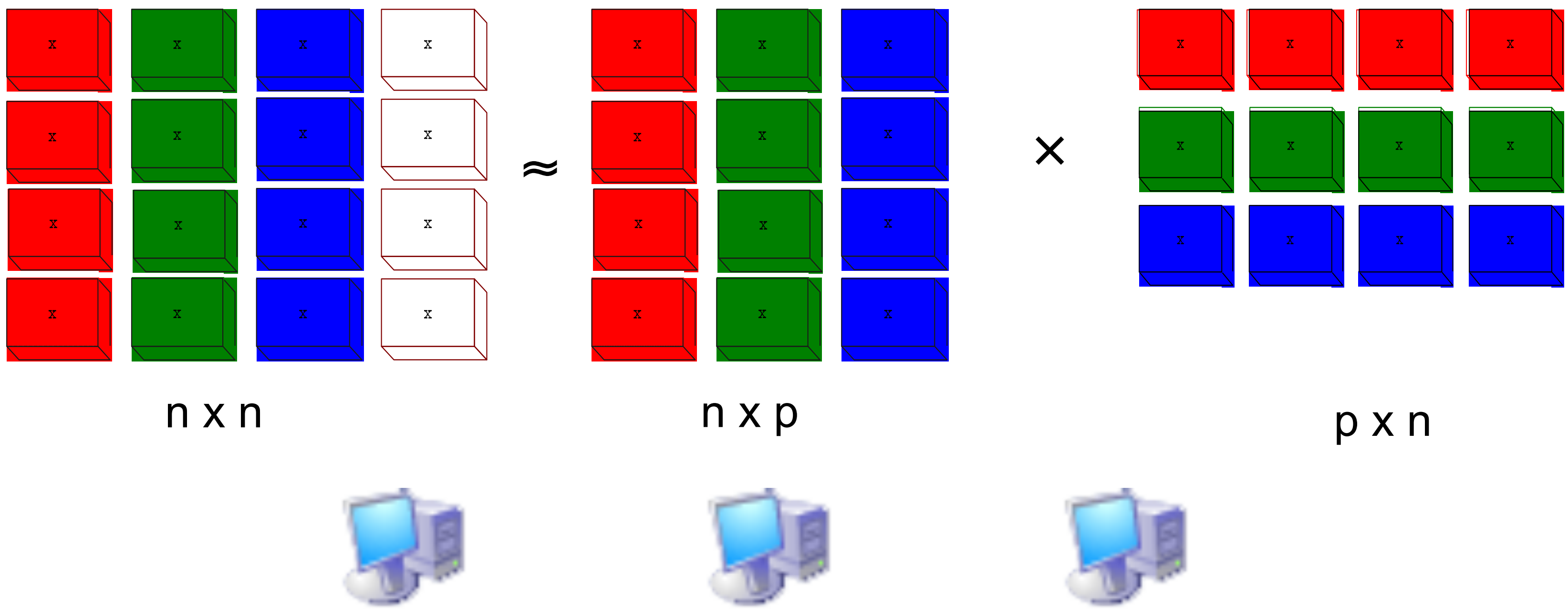
exact

approximate

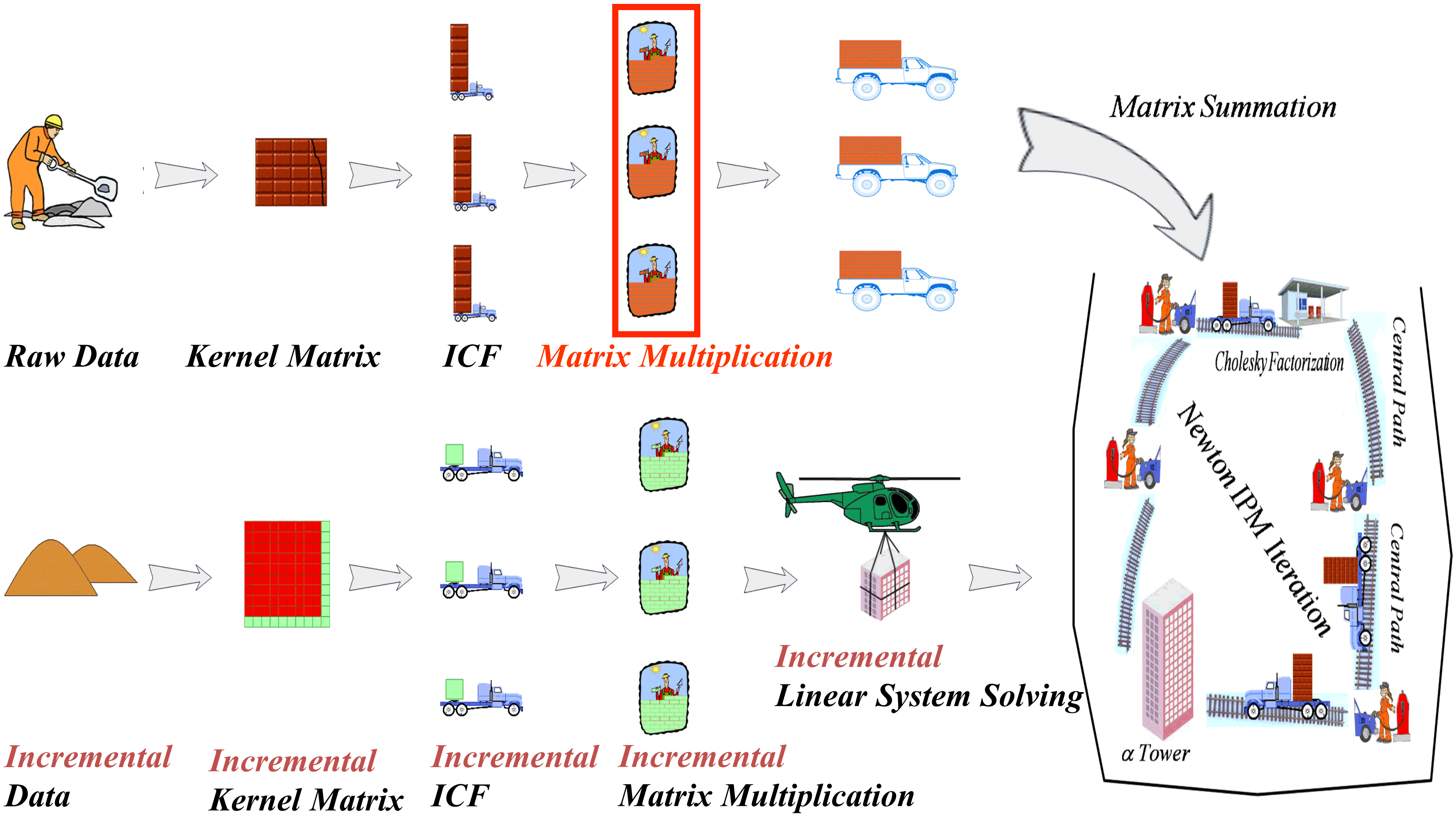
Parallelizing SVM [E. Chang, et al, NIPS 07]



Incomplete Cholesky Factorization (ICF)

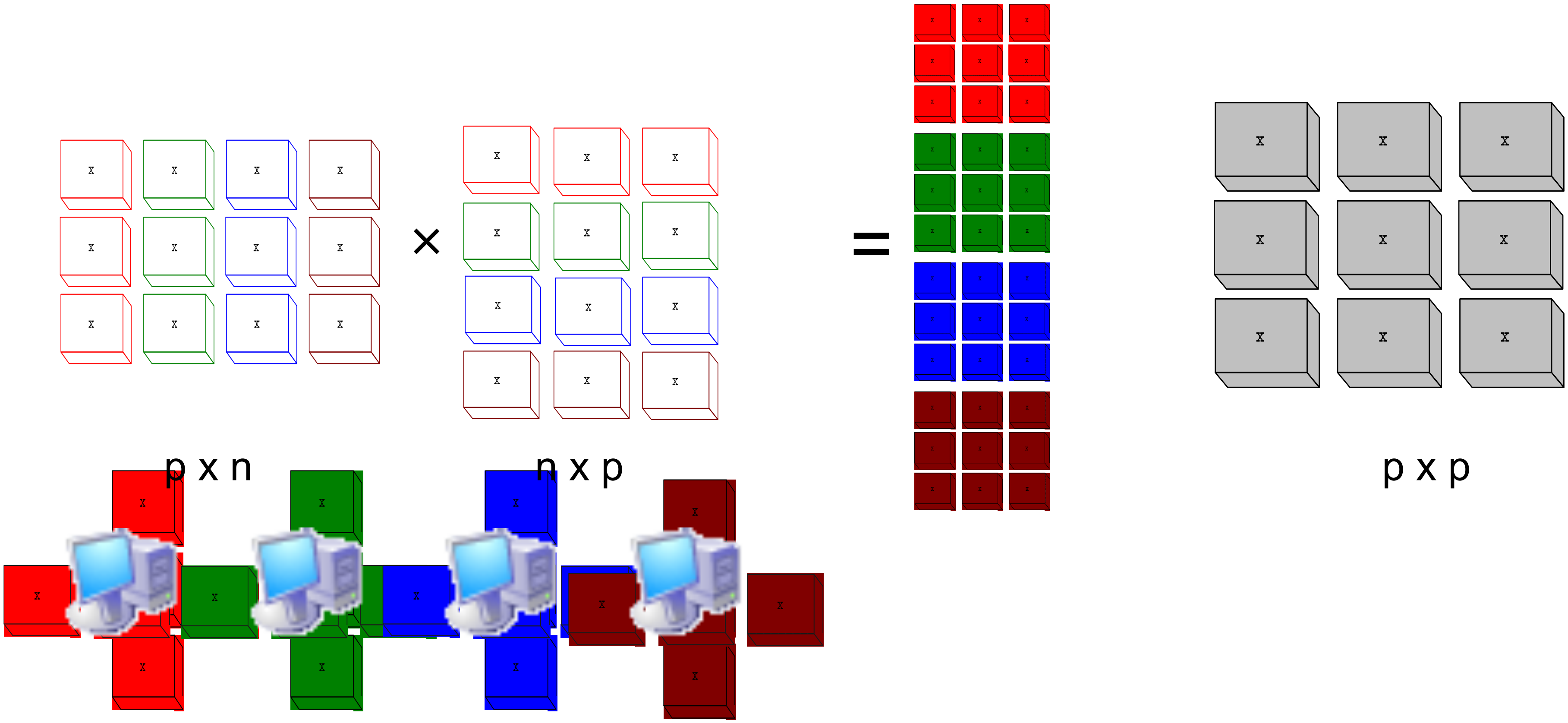


PSVM



Raw Data
Kernel Matrix
ICF
Matrix Multiplication
Matrix Summation
Cholesky Factorization
Central Path
Newton IPM Iteration
Central Path
Incremental
Linear System Solving
α Tower
Incremental
Data
Incremental
Kernel Matrix
Incremental
ICF
Incremental
Matrix Multiplication

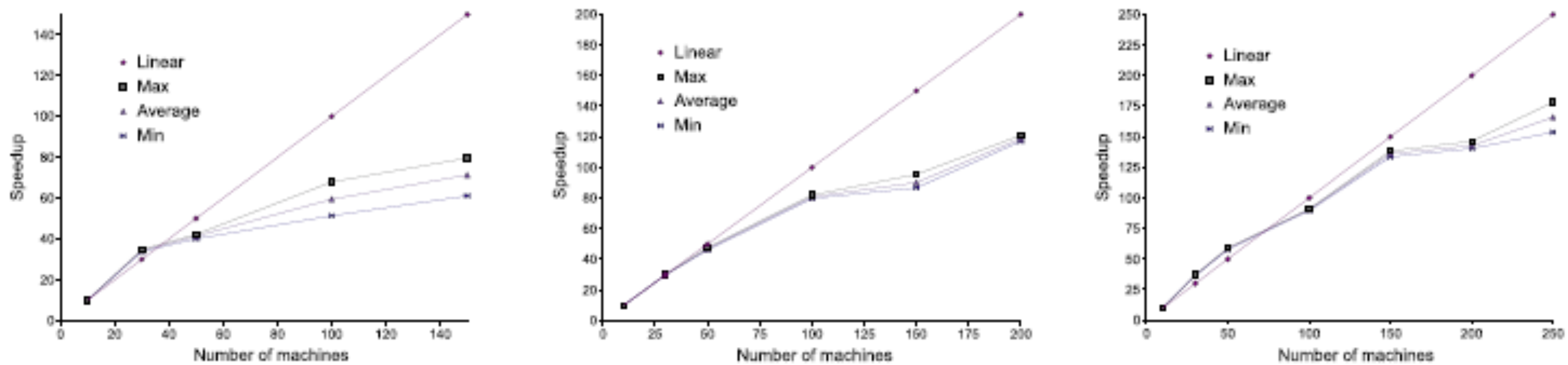
Matrix Product

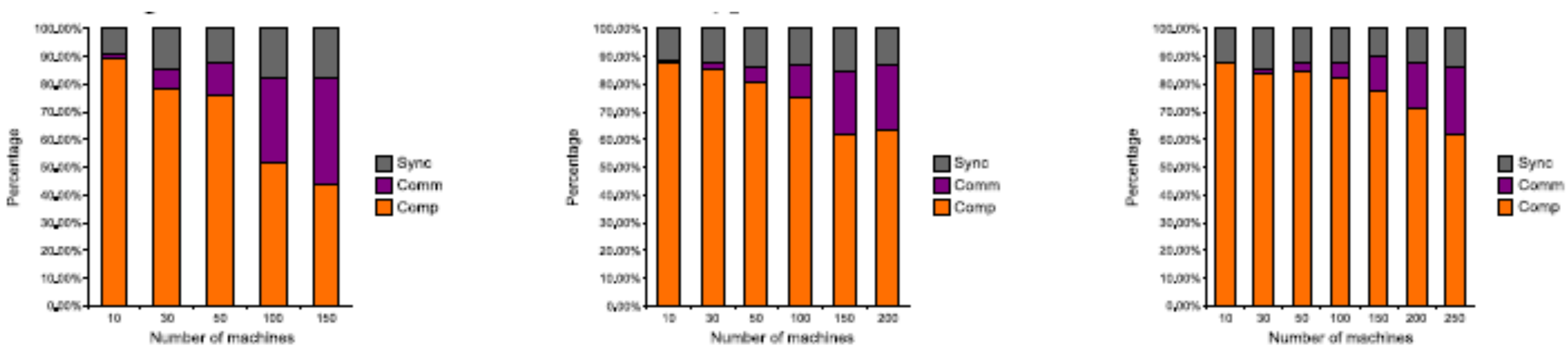


PSVM [E. Chang, et al, NIPS 07]

- Column-based ICF
  - Slower than row-based on single machine
  - Parallelizable on multiple machines
- Changing IPM computation order to achieve parallelization

Overheads


Linear
Max
Average
Min
Speedup
Number of machines


Percentage
Sync
Comm
Comp
Number of machines

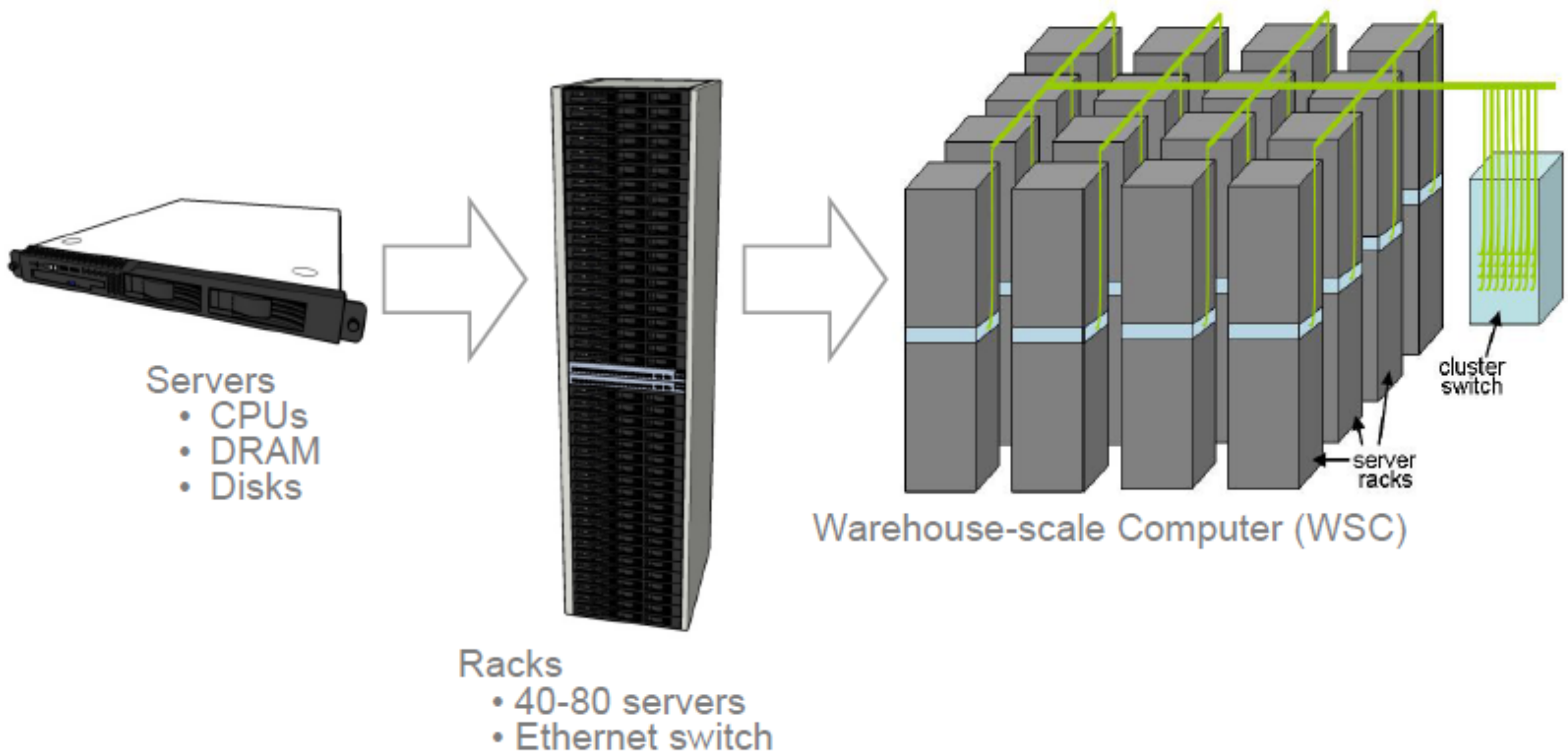
Speedup

	Image (200k)			CoverType (500k)			RCV (800k)		
Machines	Time (s)		Speedup	Time (s)		Speedup	Time (s)		Speedup
10	1,958	(9)	10*	16,818	(442)	10*	45,135	(1373)	10*
30	572	(8)	34.2	5,591	(10)	30.1	12,289	(98)	36.7
50	473	(14)	41.4	3,598	(60)	46.8	7,695	(92)	58.7
100	330	(47)	59.4	2,082	(29)	80.8	4,992	(34)	90.4
150	274	(40)	71.4	1,865	(93)	90.2	3,313	(59)	136.3
200	294	(41)	66.7	1,416	(24)	118.7	3,163	(69)	142.7
250	397	(78)	49.4	1,405	(115)	119.7	2,719	(203)	166.0
500	814	(123)	24.1	1,655	(34)	101.6	2,671	(193)	169.0
LIBSVM	4,334	NA	NA	28,149	NA	NA	184,199	NA	NA

Scalability

- Computation
  - Parallelization
  - Approximation
- File Systems
  - Latency
  - Recovery
- Power Management

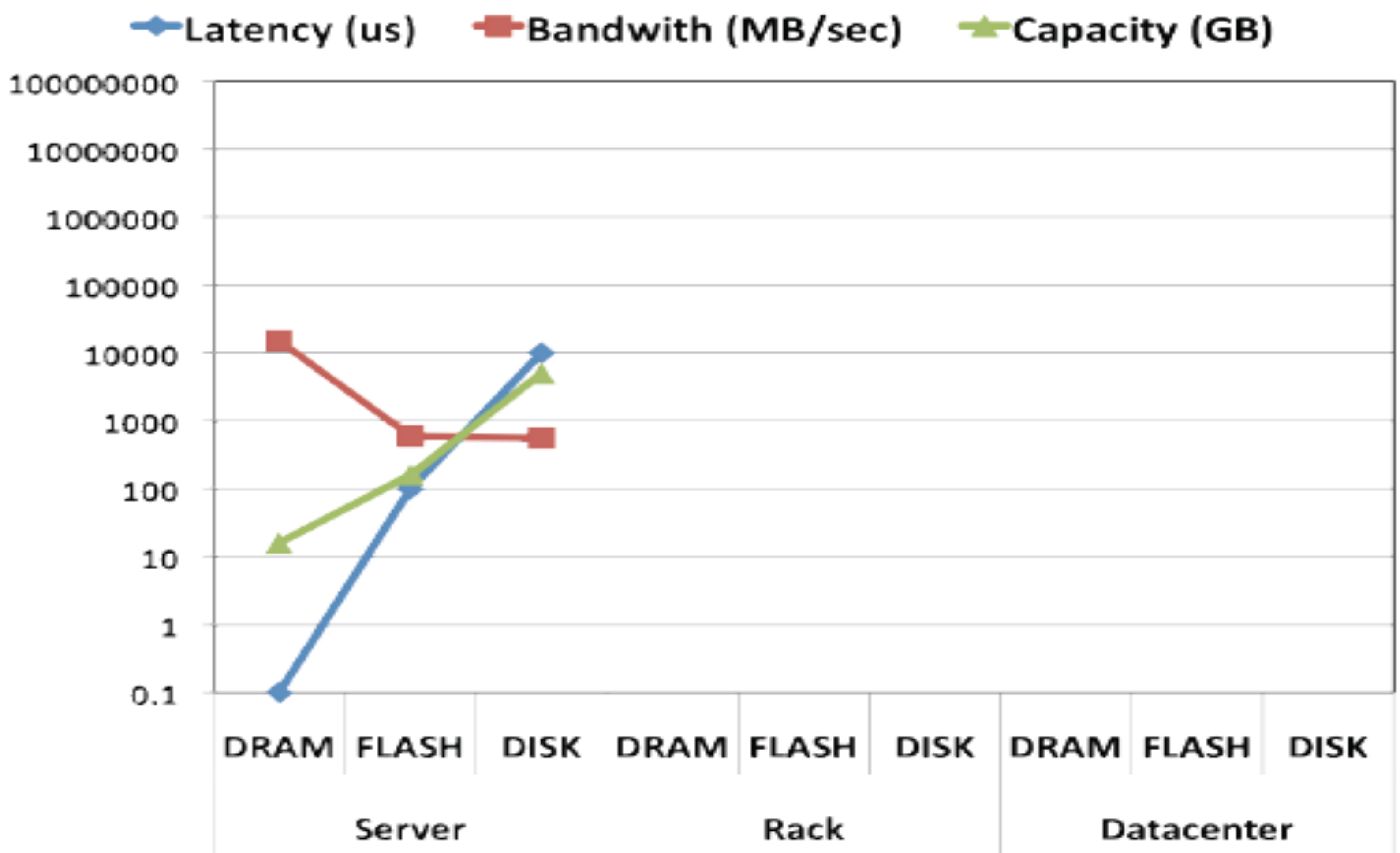
Sample Platforms



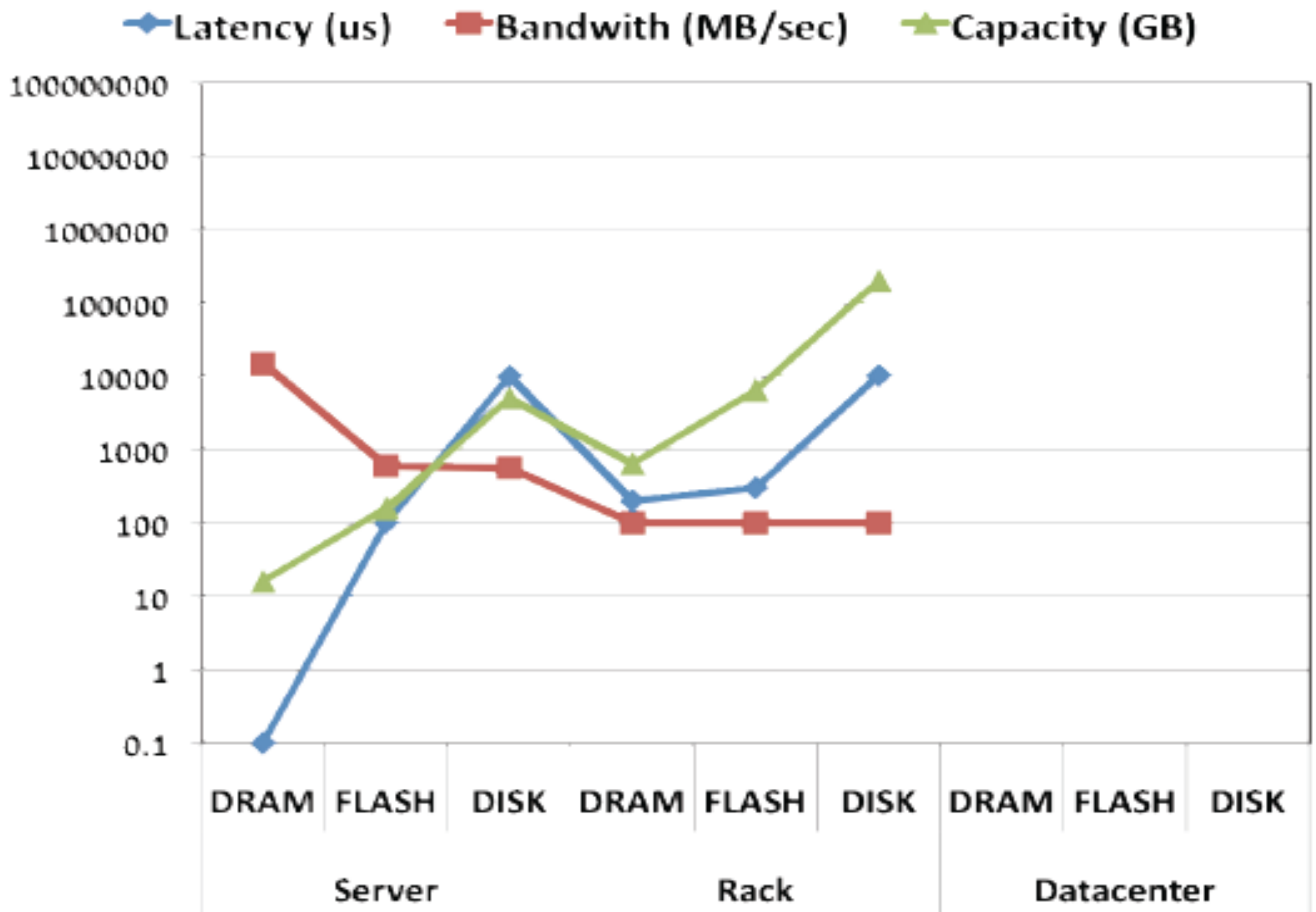
Sample Hierarchy

- Server
  - 16GB DRAM; 160MB Flash; 5 x 1TB disk
- Rack
  - 40 servers
  - 48 port Gigabit Ethernet switch
- Warehouse
  - 10,000 servers (250 racks)
  - 2K port Gigabit Ethernet switch

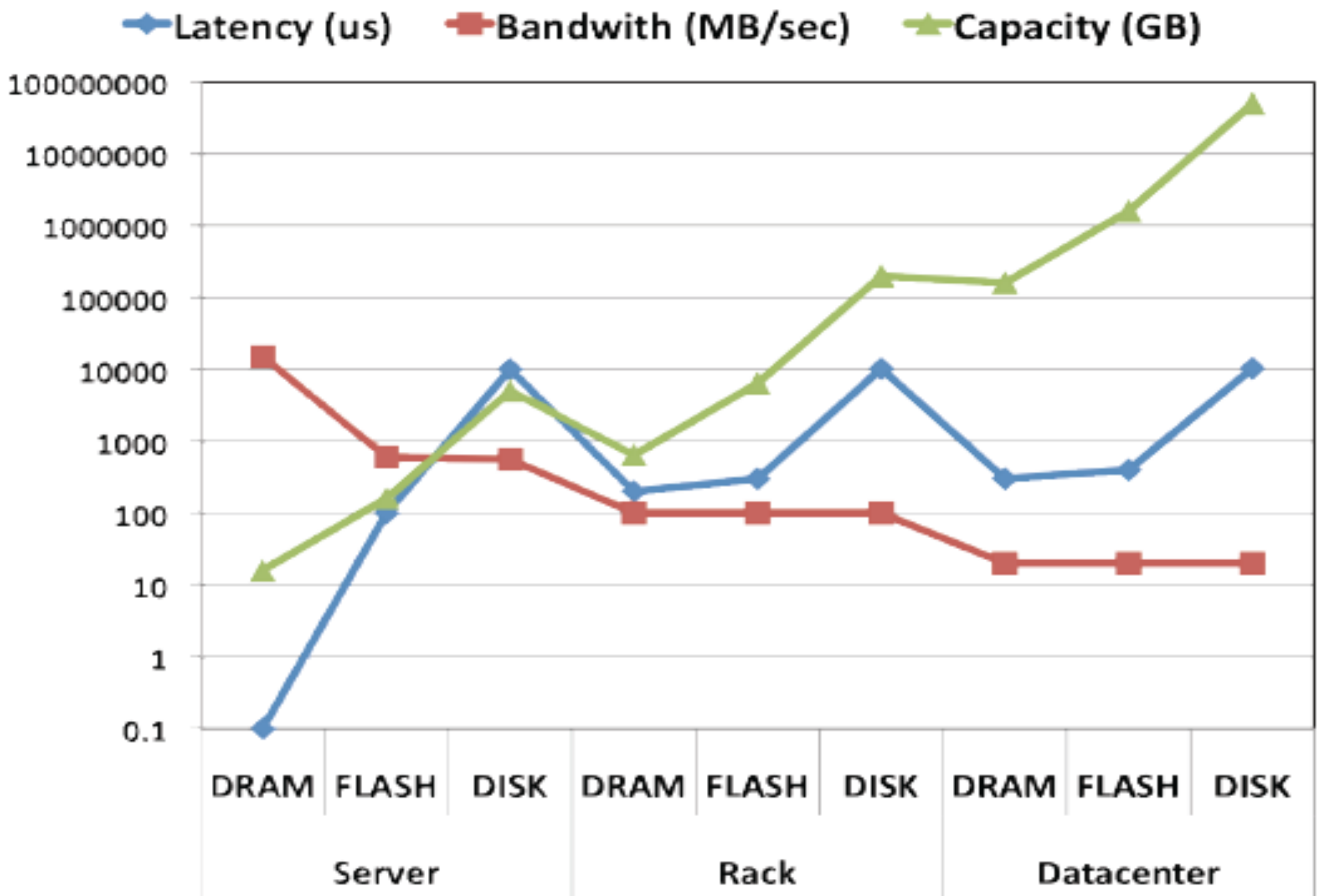
Storage --- One Server



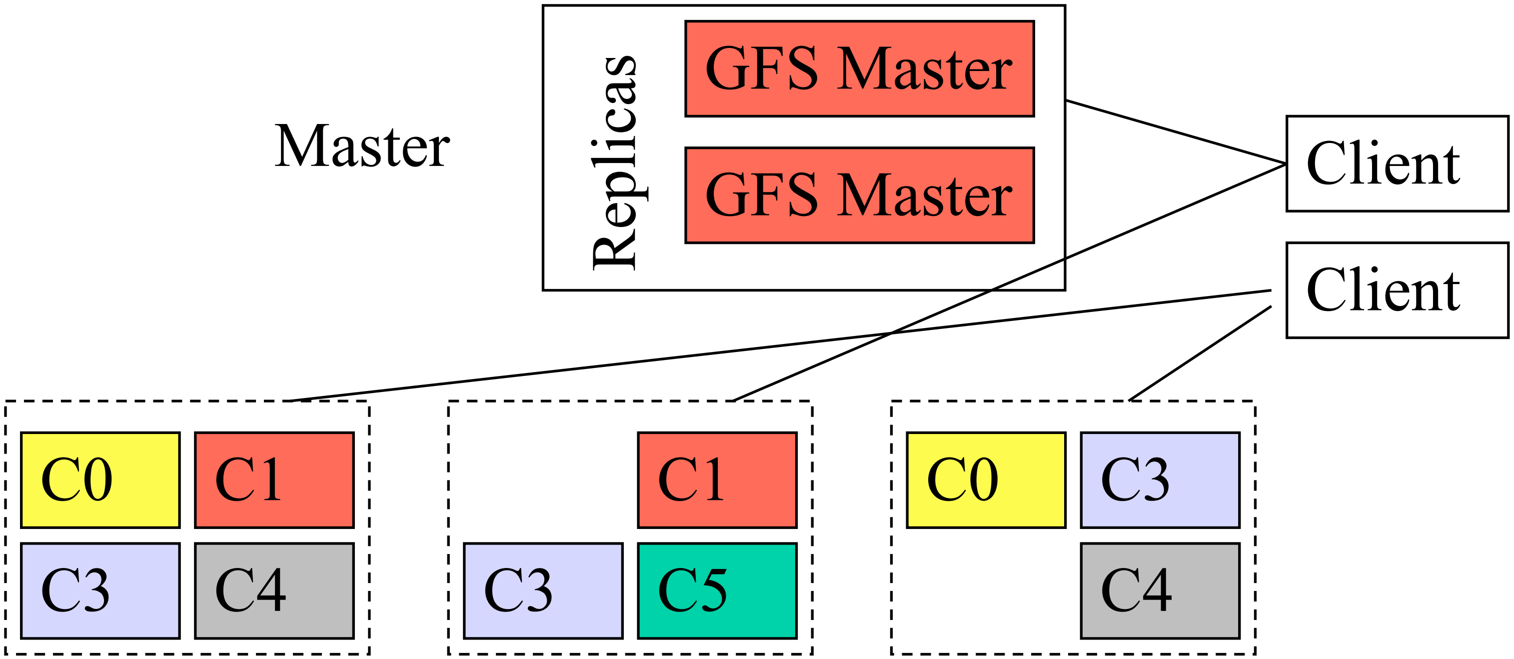
Storage --- One Rack



Storage --- One Center

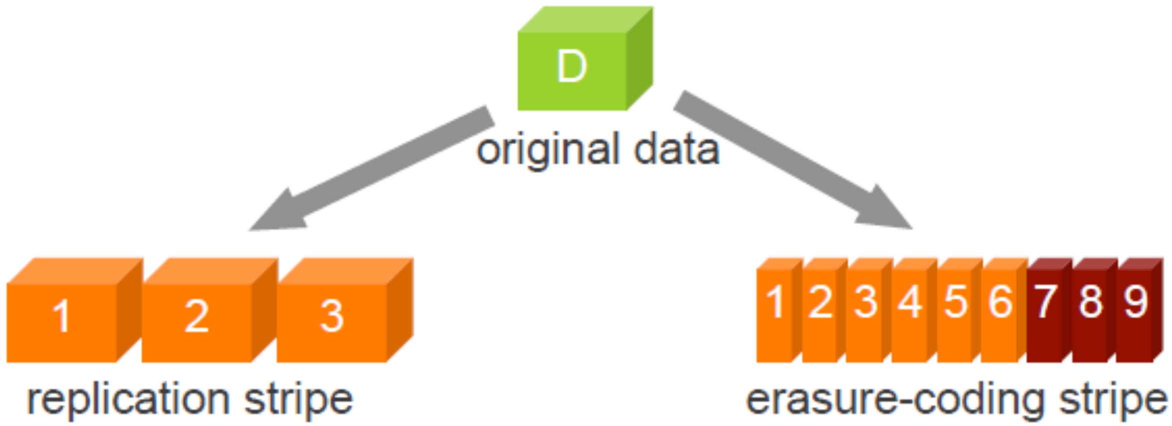


Google File System (GFS)



- Master manages metadata
- Data transfers happen directly between clients/chunkservers
- Files broken into chunks (typically 64 MB)
- Chunks triplicated across three machines for safety
- See SOSP^03 paper at http://labs.google.com/papers/gfs.html

WSC data availability: cluster file systems



- Data blocks of each stripe are placed on different fault domains
  - different disks, servers, racks
  - Data blocks are distributed across the whole WSC
    - read operations are easily load-balanced
    - recovery is highly efficient
- What affects data availability as seen by a client of a cluster file system?

Win in Scale

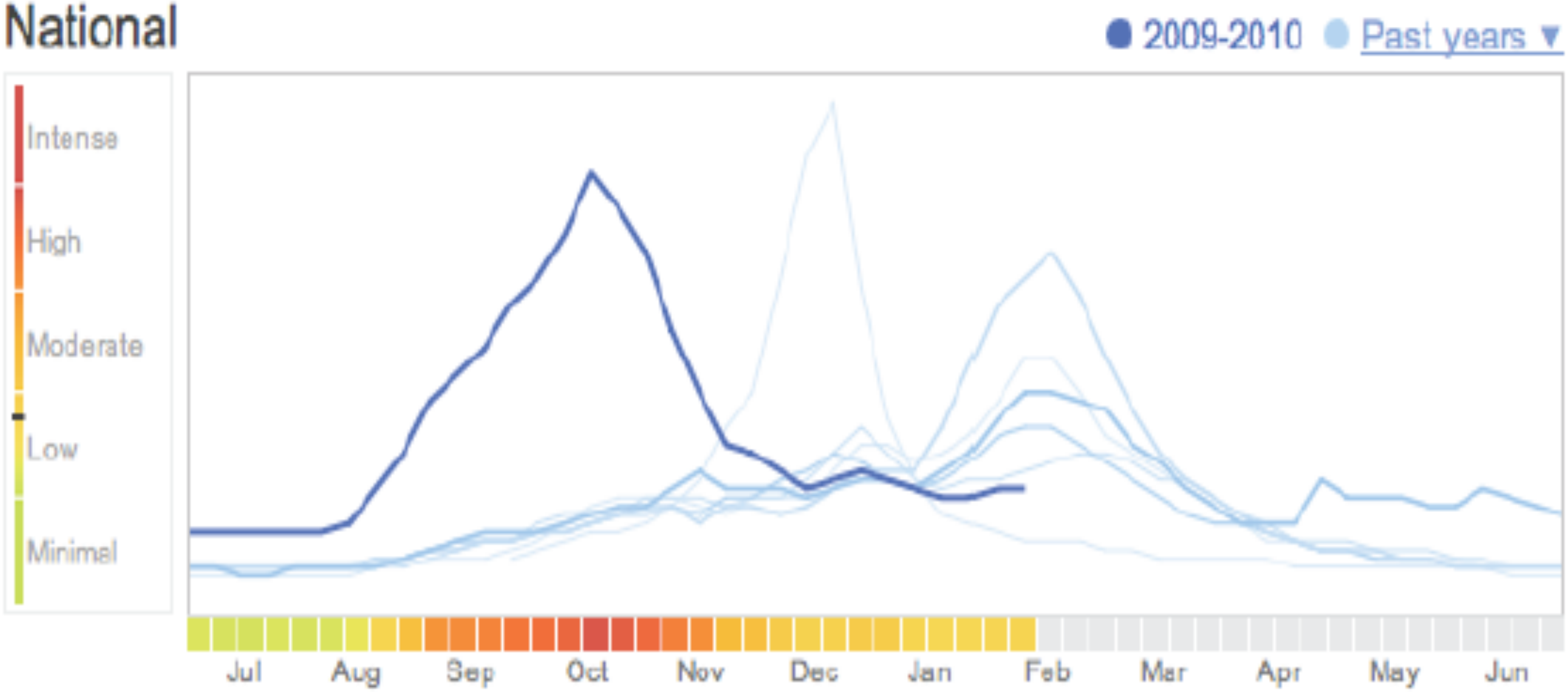
- Google Translate
- Voice
- Trend Prediction
  - An example benefits society

H1N1 United Nation Report

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National



Concluding Remarks

- Search + Social
- Increasing quantity and complexity of data demands scalable solutions
- Have parallelized key subroutines for mining massive data sets
  - Spectral Clustering [ECML 08]
  - Frequent Itemset Mining [ACM RS 08]
  - PLSA [KDD 08]
  - LDA [WWW 09, AAIM 09]
  - UserRank [Google TR 09]
  - Support Vector Machines [NIPS 07]
- Launched Google Q&A (Confucius) in 60+ countries
- Relevant papers
  - http://infolab.stanford.edu/~echang/
- Open Source PSVM, PLDA
  - http://code.google.com/p/psvm/
  - http://code.google.com/p/plda/

Models of Innovation

- Ivory tower
    - Only consider theory but not application
- Build it and they will come
    - Scientists drives product development
- "Research for sale"
    - Research funded by:
        - product groups or customers
- Research & development as equals
    - Research "sells" innovation;
    - Product "requests" innovation
- Google-style innovation