Beyond Search: Statistical Topic Models for Text Analysis

ChengXiang Zhai

Department of Computer Science
University of Illinois at Urbana-Champaign
http://www.cs.uiuc.edu/homes/czhai

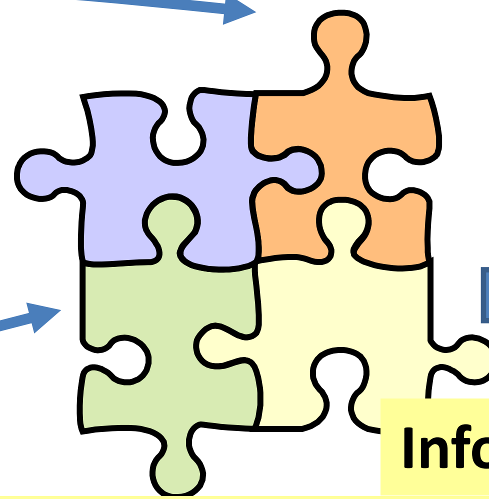
Search is a means to the end of finishing a task

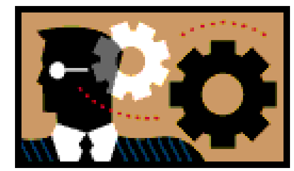
Search → **Information Synthesis & Analysis** → **Task Completion**

Search 1

Search 2

…





Decision Making
Learning
…

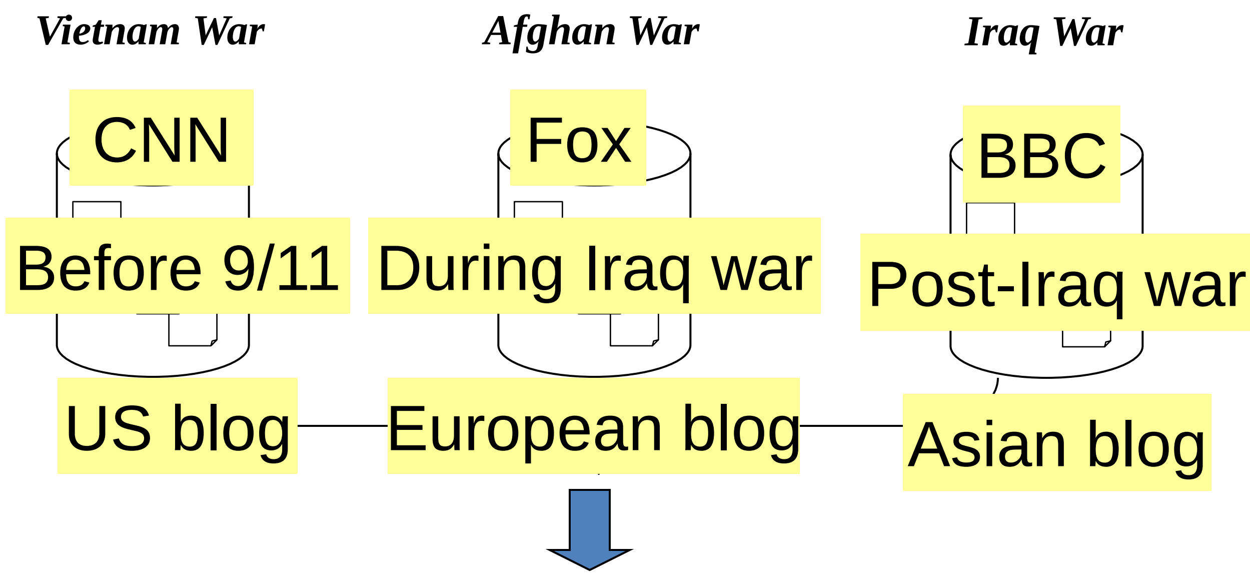
Information Interpretation

Information Synthesis

Potentially iterate…

Multiple Searches

Example Task 1: Comparing News Articles



What's in common? What's unique?

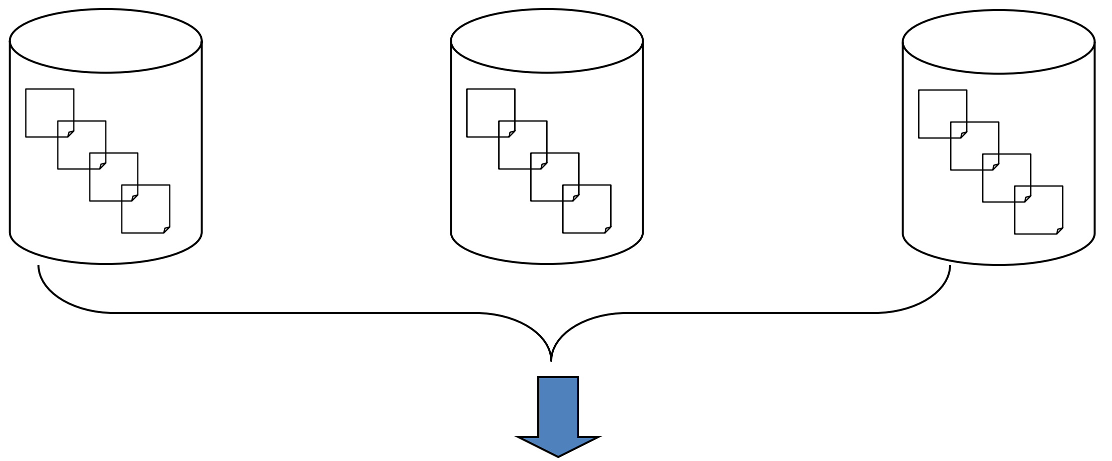
Common Themes	"Vietnam" specific	"Afghan" specific	"Iraq" specific
United nations	…	…	…
Death of people	…	…	…
…	…	…	…

Example Task 2: Compare Customer Reviews

IBM Laptop Reviews

APPLE Laptop Reviews

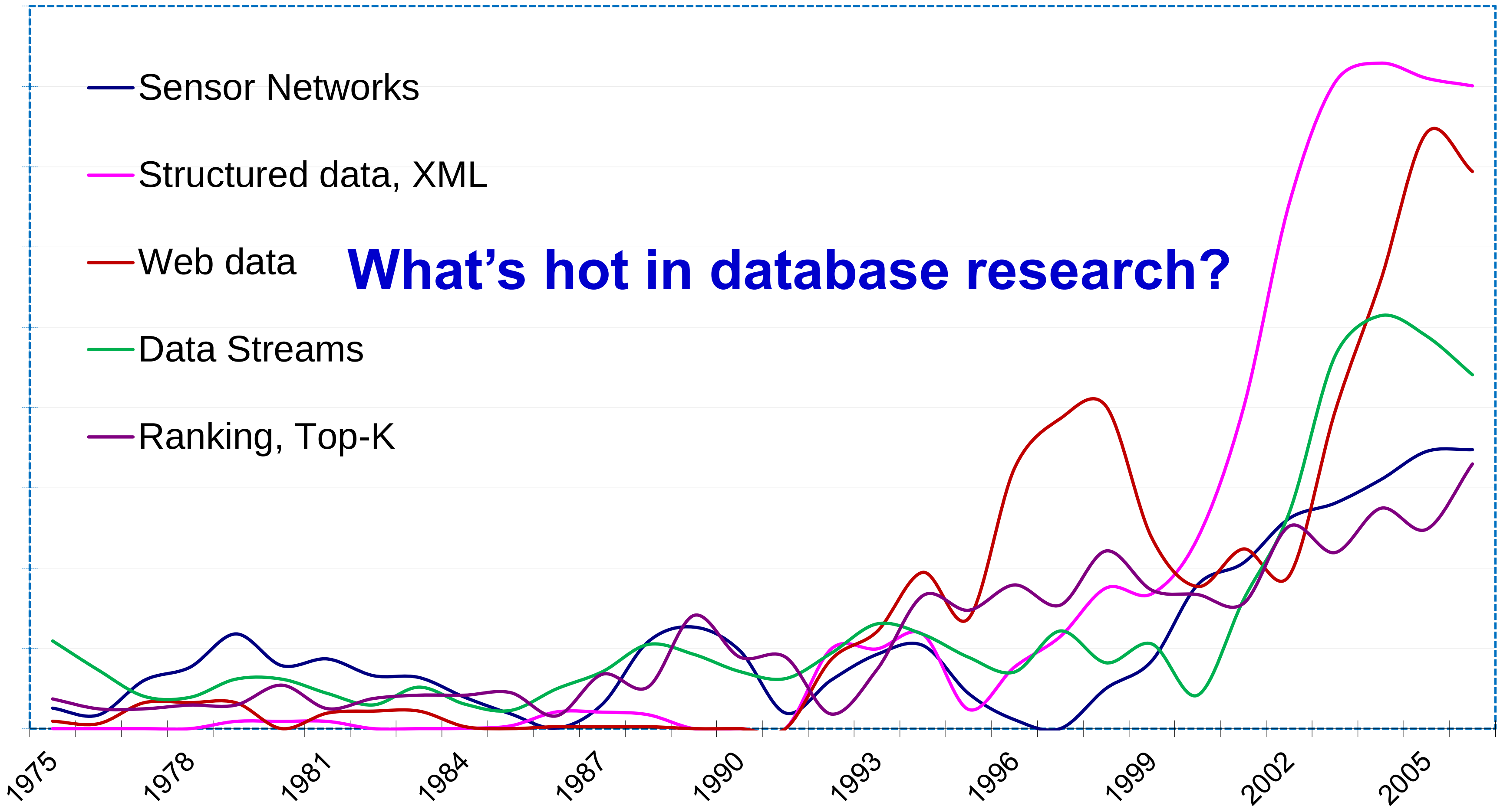
DELL Laptop Reviews



Common Themes	“IBM” specific	“APPLE” specific	“DELL” specific
Battery Life	**….**	**…**	**….**
Hard disk	**…**	**…**	**…**
Speed	**…**	**…**	**…**

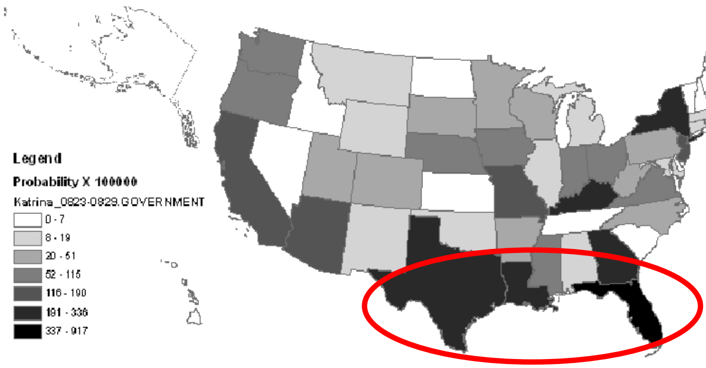
Which laptop to buy?

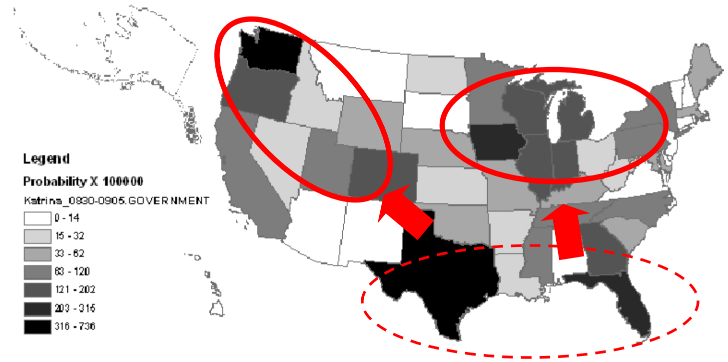
Example Task 3: Identify Emerging Research Topics



Example Task 4: Analysis of Topic Diffusion

One Week Later





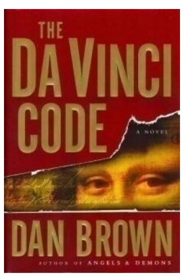
How did a discussion of a topic in blogs spread?

Sample Task 5: Opinion Analysis on Blog Articles

Query="Da Vinci Code"

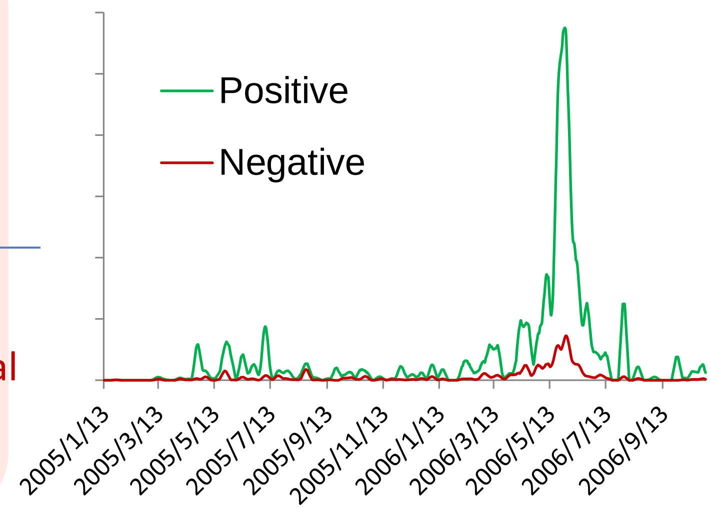
Tom Hanks, who is my favorite movie star act the leading role.

protesting... will lose your faith by watching the movie.



a good book to past time.

... so sick of people making such a big deal about a fiction **book**



What did people like/dislike about "Da Vinci Code"?

Questions

- **Can we model all these analysis problems in a general way?** **Yes!**
- **Can we solve these problems with a unified approach?** **Yes!**
- **How can we bring users into the loop?** **Yes!**

Solutions: Statistical Topic Models

Rest of the talk

- **Overview of Statistical Topic Models**
- **Contextual Probabilistic Latent Semantic Analysis (CPLSA)**
- **Text Analysis Enabled by CPLSA**
- **From Search Engines to Analysis Engines**

What is a Statistical LM?

- A probability distribution over word sequences
  - p("*Today is Wednesday*") ≈ 0.001
  - p("*Today Wednesday is*") ≈ 0.0000000000001
  - p("*The eigenvalue is positive*") ≈ 0.00001
- Context/topic dependent!
- Can also be regarded as a probabilistic mechanism for "generating" text, thus also called a "generative" model

The Simplest Language Model
(Unigram Model)

- **Generate a piece of text by generating each word <u>independently</u>**
- **Thus, $p(w_1\ w_2\ ...\ w_n)=p(w_1)p(w_2)\ldots p(w_n)$**
- **Parameters: $\{p(w_i)\}$** $p(w_1)+\ldots+p(w_N)=1$ **(N is voc. size)**
- **Essentially a multinomial distribution over words**
- **A piece of text can be regarded as a sample drawn according to this word distribution**

Text Generation with Unigram LM

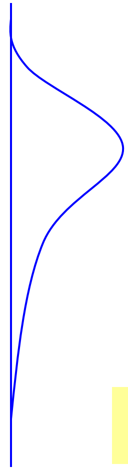
(Unigram) Language Model θ
p(w| θ)

Sampling

Document d

Topic 1:
Text mining

…
text 0.2
mining 0.1
assocation 0.01
clustering 0.02
…
food 0.00001
…

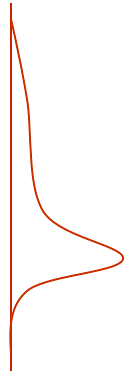


Text mining paper

Given θ, p(d| θ) varies according to d

Topic 2:
Health

…
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
…



Food nutrition paper

Estimation of Unigram LM

(Unigram) Language Model θ **Estimation** **Document**

p(w| θ)=?

…
10/100 → text ?
5/100 → mining ?
3/100 → assocation ?
3/100 → database ?
…
1/100 → query ?
…

text 10
mining 5
association 3
database 3
algorithm 2
…
query 1
efficient 1

Total #words =100

language model as topic representation?

Language Model as Text Representation: Early Work

- **1961: H. P. Luhn's early idea of using relative frequency to represent text** [Luhn 61]
- **1976: Robertson & Sparck Jones' BIR model** [Robertson & Sparck Jones 76]
- **1989: Wong & Yao's work on multinomial distribution representation** [Wong & Yao 89]

Luhn, H. P (1961) The automatic derivation of information retrieval encodements from machine-readable texts. In A. Kent (Ed.), *Information Retrieval and Machine Translation*, Vol. 3, Pt 2., pp. 1021-1028.

S. Robertson and K. Sparck Jones. (1976). *Relevance Weighting of Search Terms*. JASIS, 27, 129-146.

S. K. M. Wong and Y. Y. Yao (1989), A probability distribution model for information retrieval. *Information Processing and Management*, 25(1):39--53.

Language Model as Text Representation: Two Important Milestones in 1998~1999

- **1998: Language model for retrieval (i.e., query likelihood scoring** [Ponte & Croft 98] **(and also independently** [Hiemstra & Kraaij 99]**)**
- **1999: Probabilistic Latent Semantic Analysis (PLSA)** [Hofmann 99]

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of ACM-SIGIR 1998*, pages 275-281.
D. Hiemstra and W. Kraaij, Twenty-One at TREC-7: Ad-hoc and Cross-language track, In *Proceedings of the Seventh Text REtrieval Conference (TREC-7),* 1999.
Thomas Hofmann: Probabilistic Latent Semantic Analysis. UAI 1999: 289-296

Probabilistic Latent Semantic Analysis (PLSA)

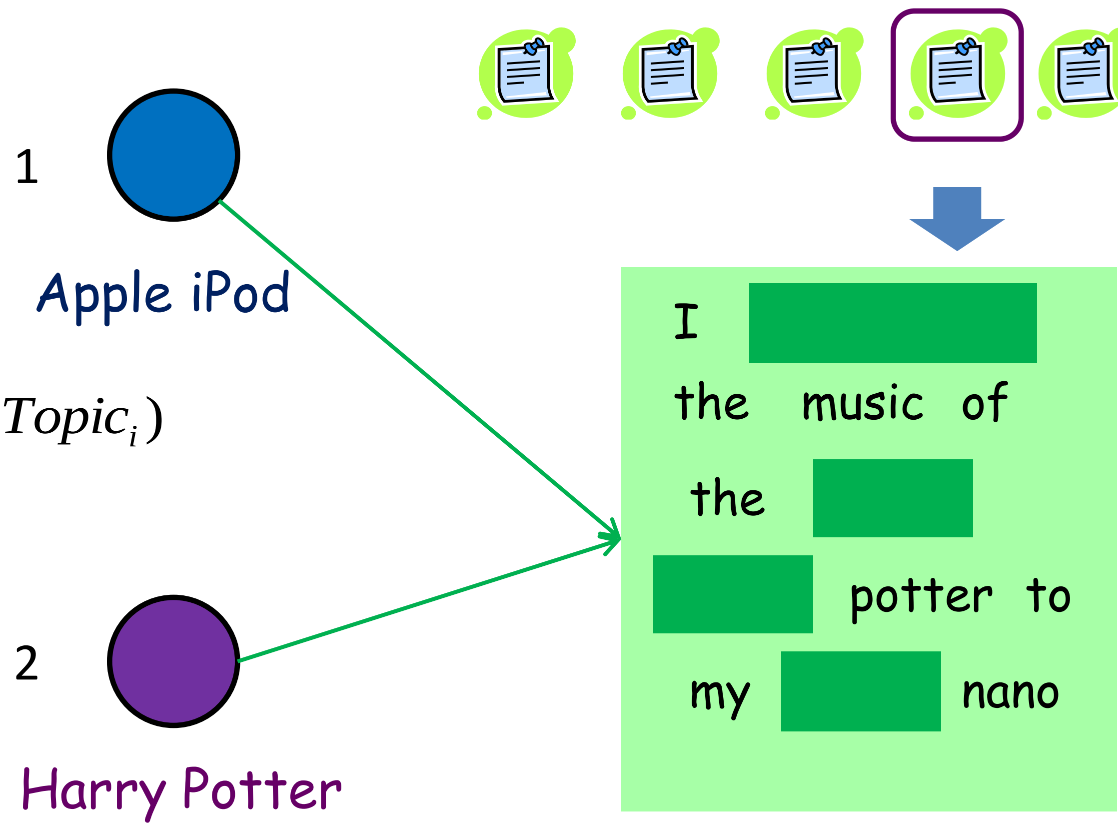
Topic 1

ipod	**0.15**
nano	0.08
music	0.05
download	0.02
apple	0.01

$$P(w) = \sum_{i=1..K} P(z=i)P(w\,|\,Topic_i)$$

Topic 2

movie	0.10
harry	**0.09**
potter	0.05
actress	0.04
music	0.02



Parameter Estimation

- **Maximizing data likelihood:**

$$\Lambda^* = \arg\max_{\Lambda} \log(P(Data \mid Model))$$

- **Parameter Estimation using EM algorithm**

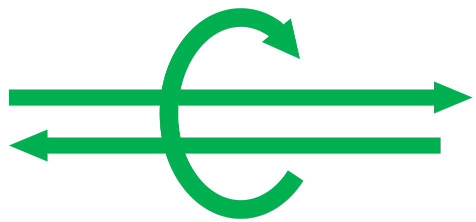
Prior set by users

Pseudo-Counts

ipod	?
nano	?
music	?
download	?
apple	?

movie	?
harry	?
potter	?
actress	?
music	?

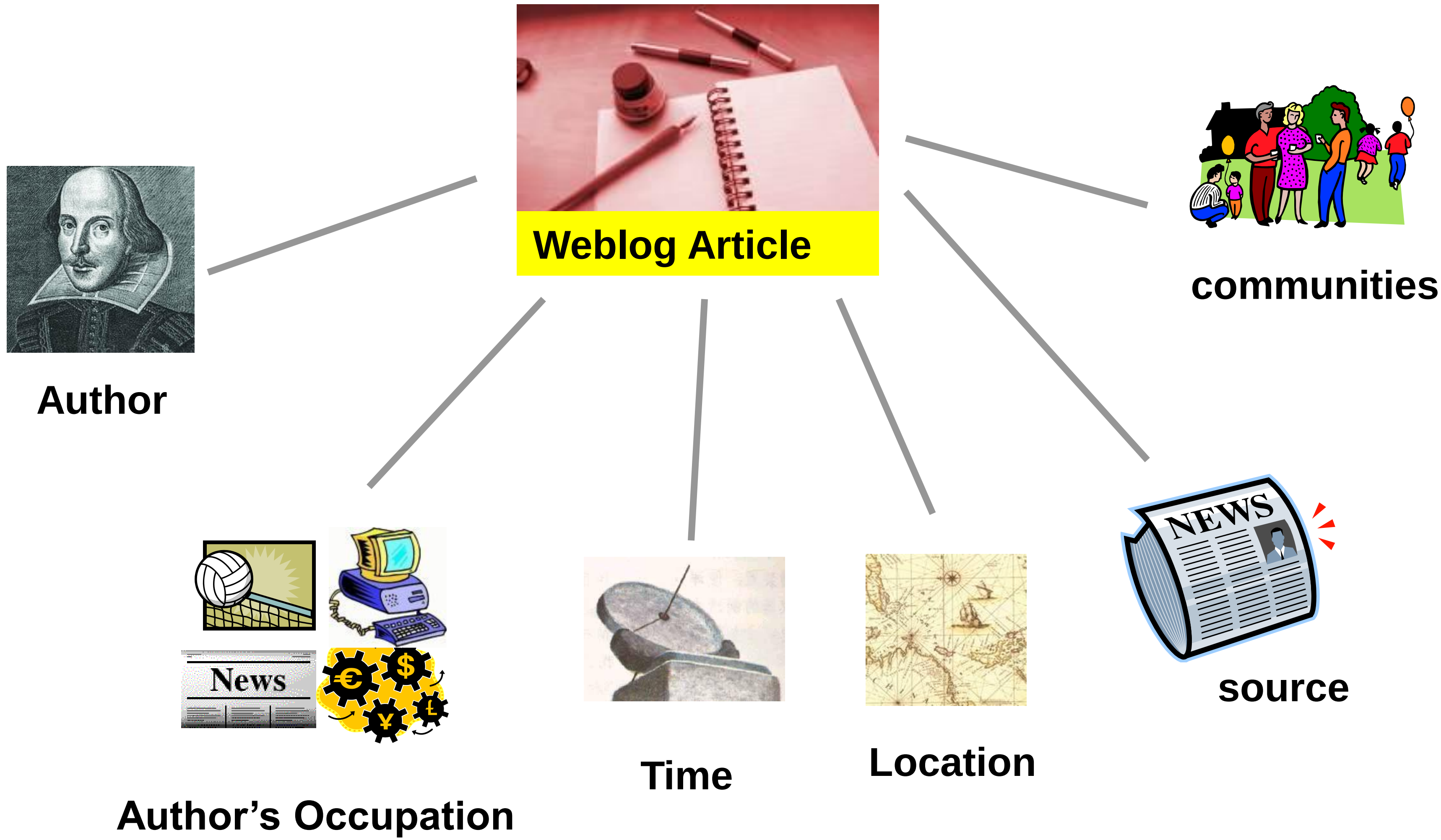
Guess the affiliation



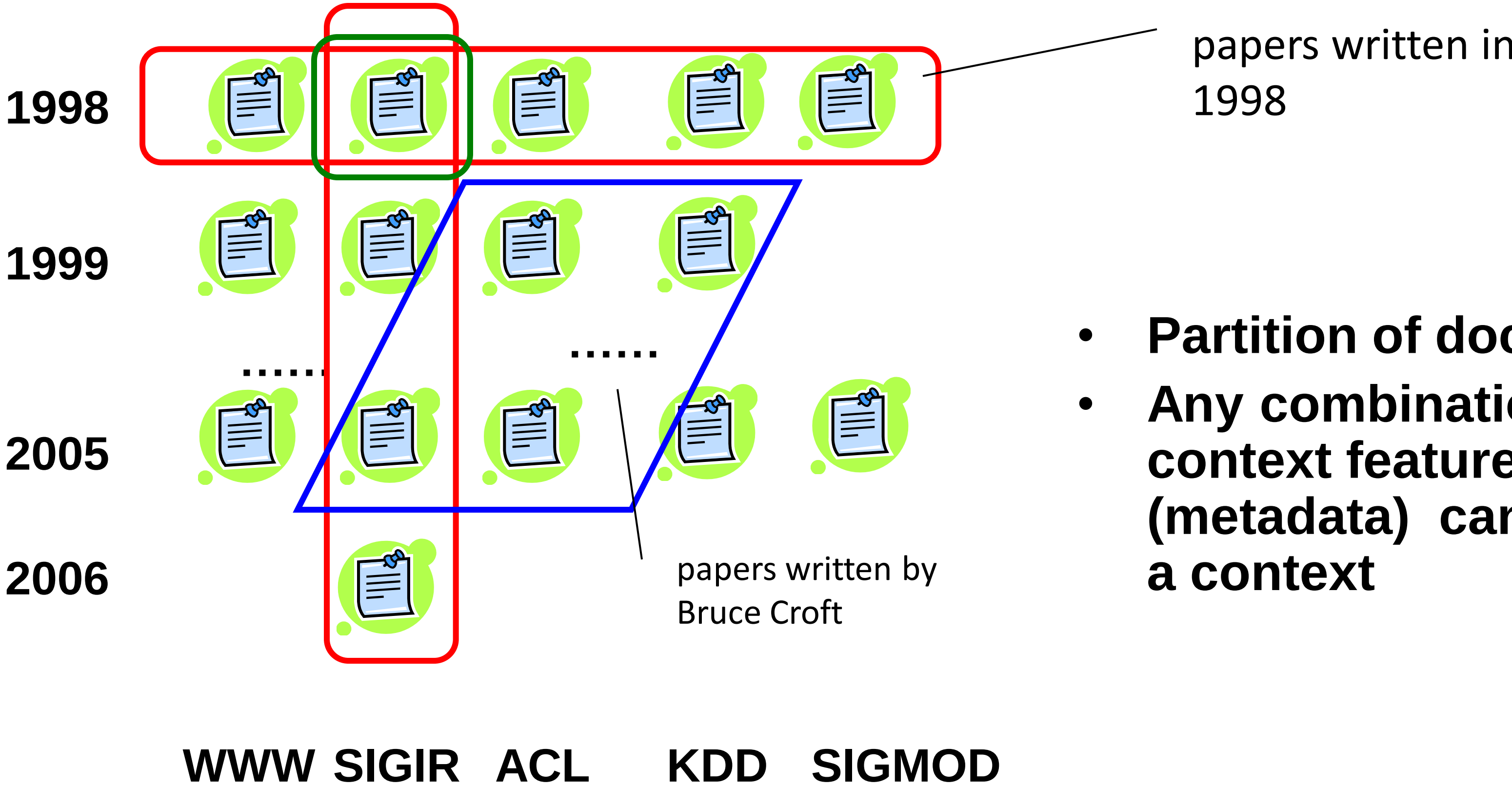
Estimate the params

I downloaded
the music of
the movie
harry potter to
my ipod nano

Context Features of a Document



A General View of Context



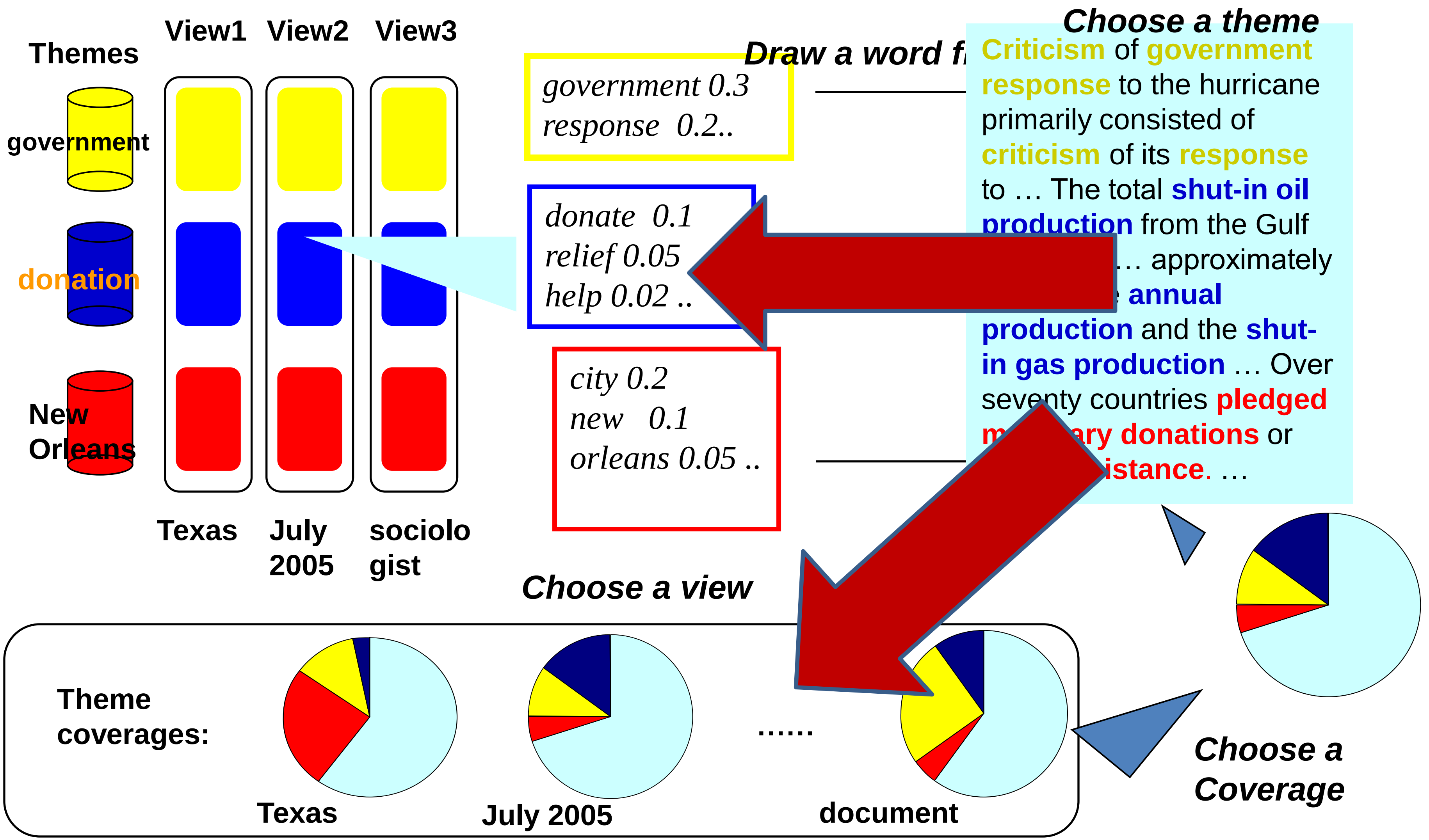
- **Partition of documents**
- **Any combination of context features (metadata) can define a context**

Empower PLSA with Context [Mei & Zhai 06]

- **Make topics depend on context variables**
- **Text is generated from a contextualized PLSA model (CPLSA)**
- **Fitting such a model to text enables a wide range of analysis tasks involving topics and context**

Qiaozhu Mei, ChengXiang Zhai, **A Mixture Model for Contextual Text Mining**, *Proceedings of the 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , (**KDD'06**), pages 649-655

Contextual Probabilistic Latent Semantics Analysis



Comparing News Articles

Iraq War (30 articles) vs. Afghan War (26 articles)

The common theme indicates that "United Nations" is involved in both wars

	Cluster 1	**Cluster 2**	**Cluster 3**
Common Theme	united 0.042 nations 0.04 …	killed 0.035 month 0.032 deaths 0.023 …	…
Iraq Theme	n 0.03 Weapons 0.024 Inspections 0.023 …	troops 0.016 hoon 0.015 sanches 0.012 …	…
Afghan Theme	Northern 0.04 alliance 0.04 kabul 0.03 taleban 0.025 aid 0.02 …	taleban 0.026 rumsfeld 0.02 hotel 0.012 front 0.011 …	…

Collection-specific themes indicate different roles of "United Nations" in the two wars

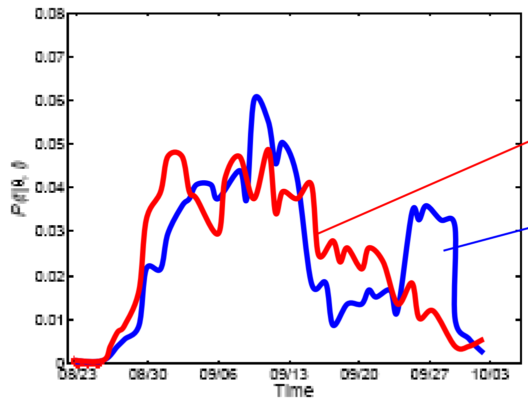
Spatiotemporal Patterns in Blog Articles

- **Query= "Hurricane Katrina"**
- **Topics in the results:**

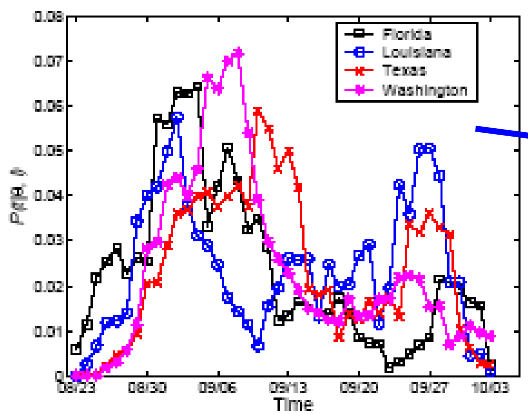
Government Response	New Orleans	Oil Price	Praying and Blessing	Aid and Donation	Personal
bush 0.071	city 0.063	price 0.077	god 0.141	donate 0.120	i 0.405
president 0.061	orleans 0.054	oil 0.064	pray 0.047	relief 0.076	my 0.116
federal 0.051	new 0.034	gas 0.045	prayer 0.041	red 0.070	me 0.060
government 0.047	louisiana 0.023	increase 0.020	love 0.030	cross 0.065	am 0.029
fema 0.047	flood 0.022	product 0.020	life 0.025	help 0.050	think 0.015
administrate 0.023	evacuate 0.021	fuel 0.018	bless 0.025	victim 0.036	feel 0.012
response 0.020	storm 0.017	company 0.018	lord 0.017	organize 0.022	know 0.011
brown 0.019	resident 0.016	energy 0.017	jesus 0.016	effort 0.020	something 0.007
blame 0.017	center 0.016	market 0.016	will 0.013	fund 0.019	guess 0.007
governor 0.014	rescue 0.012	gasoline 0.012	faith 0.012	volunteer 0.019	myself 0.006

- **Spatiotemporal patterns**

Theme Life Cycles ("Hurricane Katrina")



(a) Theme life cycles in Texas (Hurricane Katrina)



(b) Theme "New Orleans" over states (Hurricane Katrina)

Oil Price

price 0.0772
oil 0.0643
gas 0.0454
increase 0.0210
product 0.0203
fuel 0.0188
company 0.0182
…

New Orleans

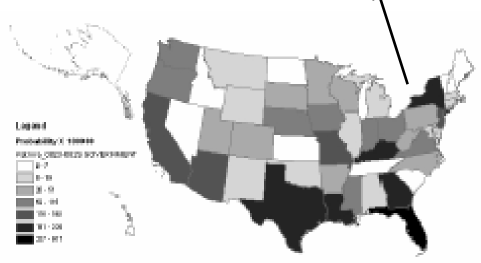
city 0.0634
orleans 0.0541
new 0.0342
louisiana 0.0235
flood 0.0227
evacuate 0.0211
storm 0.0177
…

Theme Snapshots ("Hurricane Katrina")

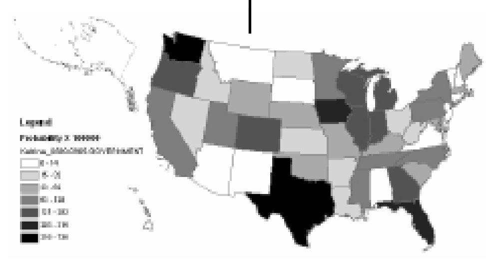
Week2: The discussion moves towards the north and west

Week1: The theme is the strongest along the Gulf of Mexico

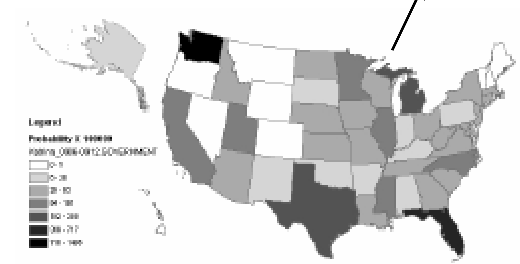
Week3: The theme distributes more uniformly over the states



(a) Week1: 08/23-08/29

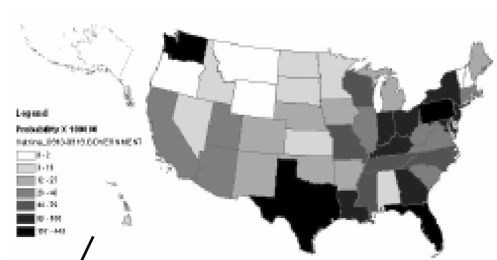


(b) Week Two: 08/30-09/05

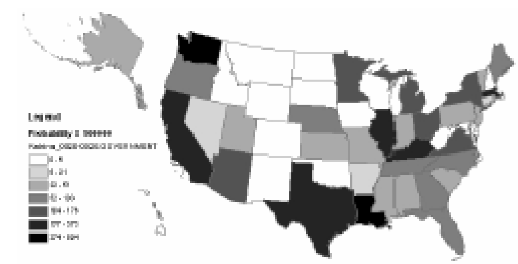


(c) Week Three:09/06-09/12

Theme 1
Government Response
bush 0.0716374
president 0.0610942
federal 0.0514114
govern 0.0476977
fema 0.0474692
administrate 0.0233903
response 0.0208351
brown 0.0199573
blame 0.0170033
governor 0.0142153



(d) Week Four: 09/13-09/19

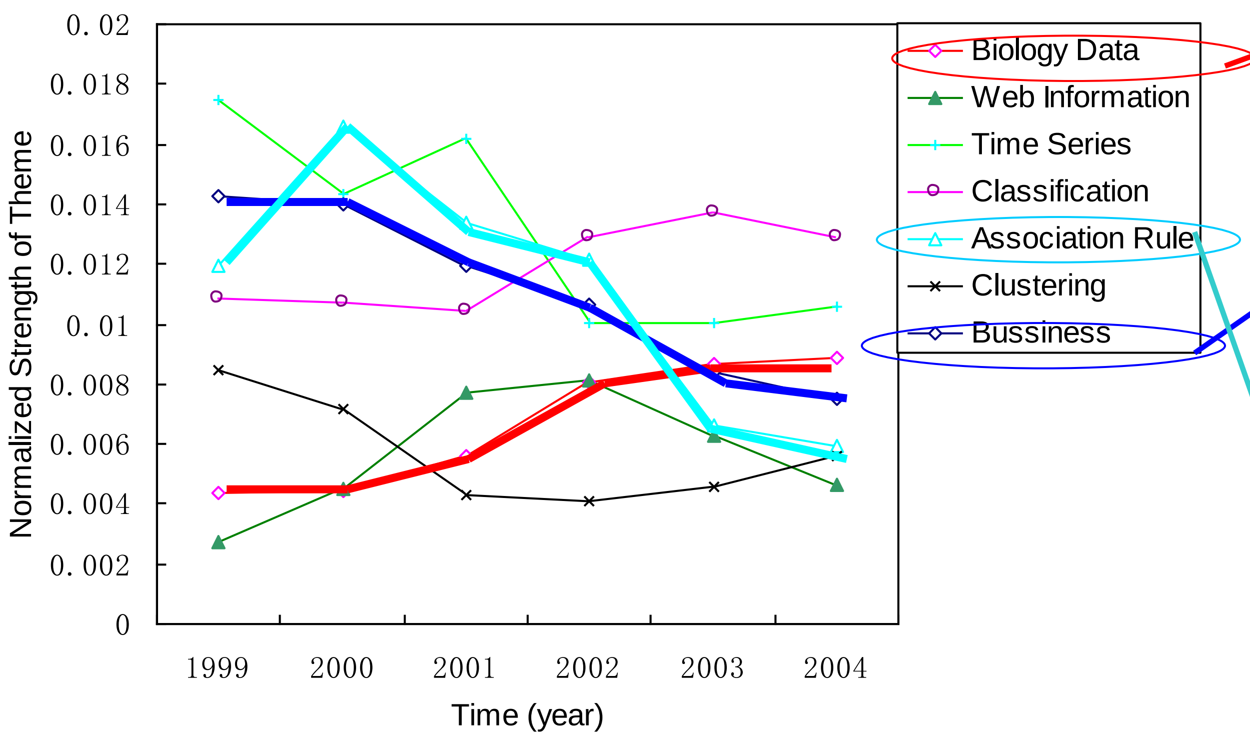


(e) Week Five: 09/20-09/26

Week4: The theme is again strong along the east coast and the Gulf of Mexico

Week5: The theme fades out in most states

Theme Life Cycles (KDD Papers)

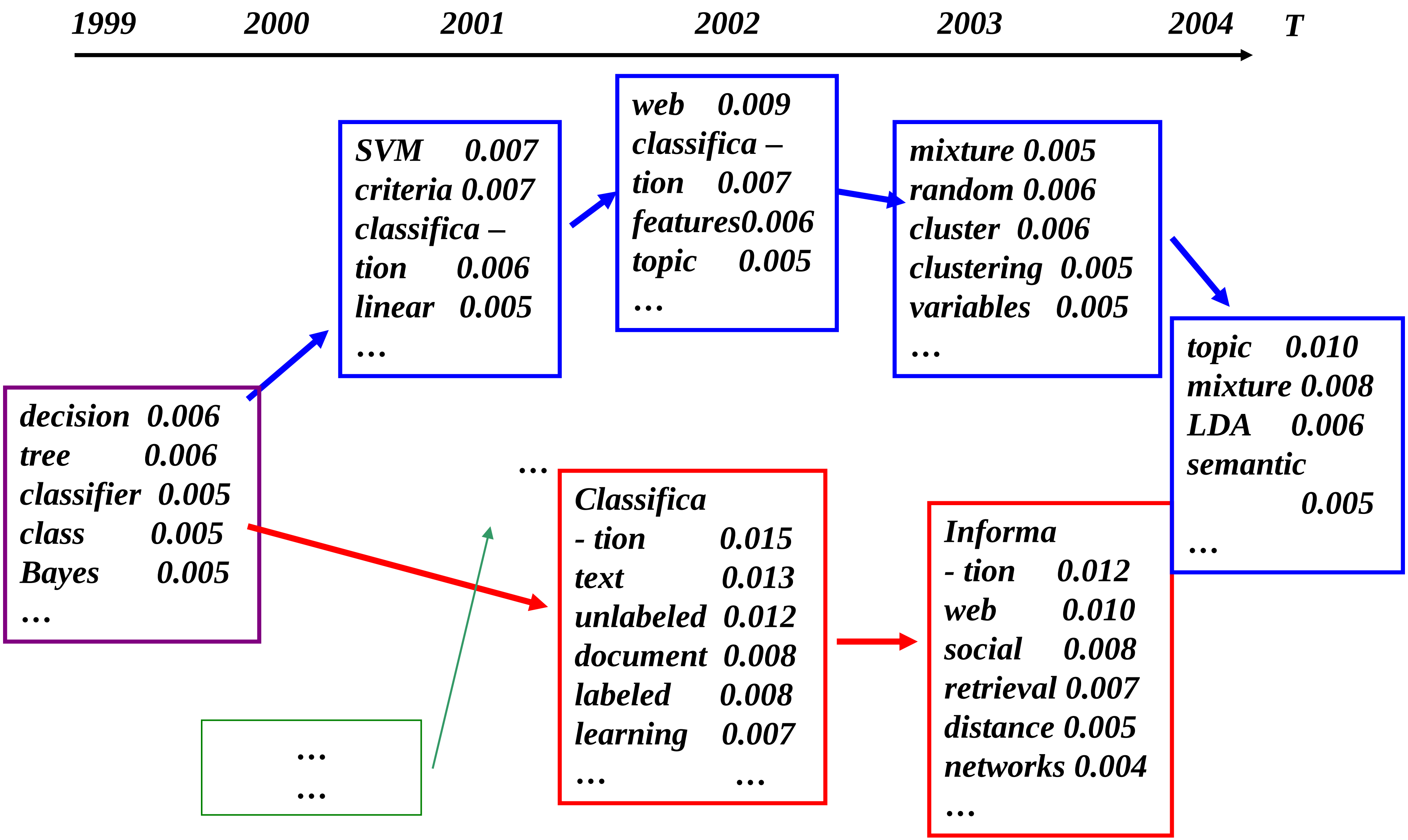


gene 0.0173
expressions 0.0096
probability 0.0081
microarray 0.0038
…

marketing 0.0087
customer 0.0086
model 0.0079
business 0.0048
…

rules 0.0142
association 0.0064
support 0.0053
…

Theme Evolution Graph: KDD

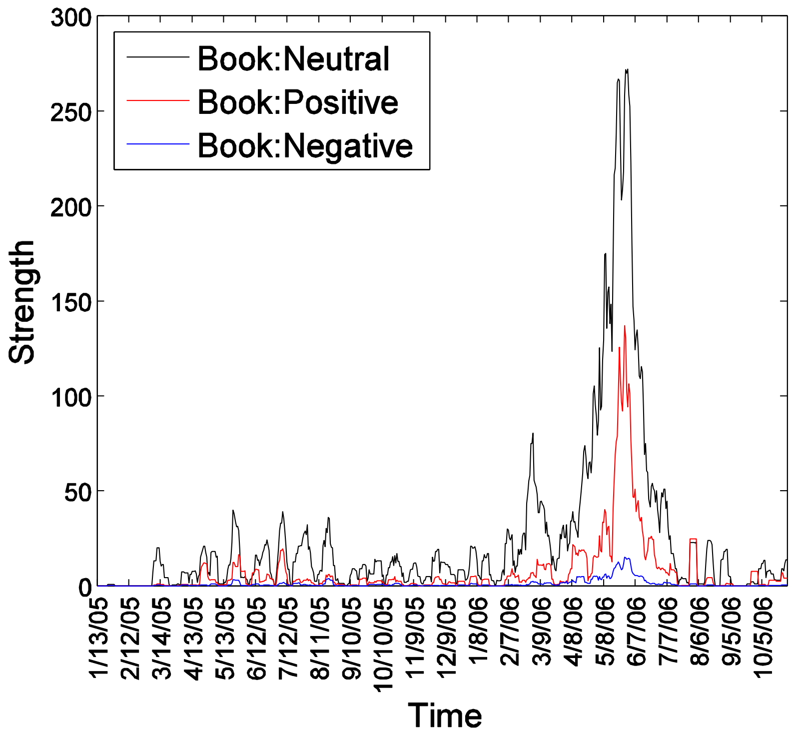


Multi-Faceted Sentiment Summary (query="Da Vinci Code")

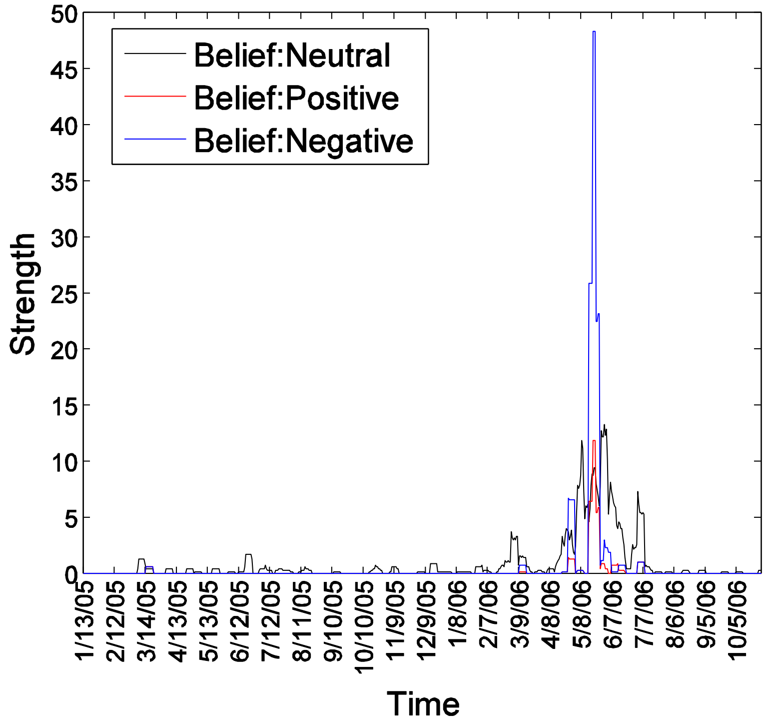
	Neutral	Positive	Negative
Facet 1: Movie	... Ron Howards selection of Tom Hanks to play Robert Langdon.	Tom Hanks stars in the movie,who can be mad at that?	But the movie might get delayed, and even killed off if he loses.
	Directed by: Ron Howard Writing credits: Akiva Goldsman ...	Tom Hanks, who is my favorite movie star act the leading role.	protesting ... will lose your faith by ... watching the movie.
	After watching the movie I went online and some research on ...	Anybody is interested in it?	... so sick of people making such a big deal about a FICTION **book** and **movie**.
Facet 2: Book	I remembered when i first read the book, I finished the book in two days.	Awesome book.	... so sick of people making such a big deal about a FICTION **book** and **movie**.
	I'm reading "Da Vinci Code" now. …	So still a good book to past time.	This controversy book cause lots conflict in west society.

Separate Theme Sentiment Dynamics

"book"



"religious beliefs"



Event Impact Analysis: IR Research

SIGIR papers

Theme:
retrieval models

term	*0.1599*
relevance	*0.0752*
weight	*0.0660*
feedback	*0.0372*
independence	*0.0311*
model	*0.0310*
frequent	*0.0233*
probabilistic	*0.0188*
document	*0.0173*
…	

1992

Starting of the TREC conferences

vector	*0.0514*
concept	*0.0298*
extend	*0.0297*
model	*0.0291*
space	*0.0236*
boolean	*0.0151*
function	*0.0123*
feedback	*0.0077*
…	

probabilist	*0.0778*
model	*0.0432*
logic	*0.0404*
ir	*0.0338*
boolean	*0.0281*
algebra	*0.0200*
estimate	*0.0119*
weight	*0.0111*
…	

1998

Publication of the paper "A language modeling approach to information retrieval"

xml	*0.0678*
email	*0.0197*
model	*0.0191*
collect	*0.0187*
judgment	*0.0102*
rank	*0.0097*
subtopic	*0.0079*
…	

model	*0.1687*
language	*0.0753*
estimate	*0.0520*
parameter	*0.0281*
distribution	*0.0268*
probable	*0.0205*
smooth	*0.0198*
markov	*0.0137*
likelihood	*0.0059*
…	

year

Many Other Variations

- **Latent Dirichlet Allocation (LDA) [Blei et al. 03]**
  - Impose priors on topic choices and word distributions
  - Make PLSA a generative model
- **Many variants of LDA!**
- **In practice, LDA and PLSA variants tend to work equally well for text analysis** [Lu et al. 11]

[Blei et al. 02] D. Blei, A. Ng, and M. Jordan. *Latent dirichlet allocation*. In T G Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14, Cambridge, MA, 2002. MIT Press.

Yue Lu, Qiaozhu Mei, ChengXiang Zhai. Investigating Task Performance of Probabilistic Topic Models - An Empirical Study of PLSA and LDA, *Information Retrieval*, vol. 14, no. 2, April, 2011.

Other Uses of Topic Models for Text Analysis

- **Topic analysis on social networks** [Mei et al. 08]
- **Opinion Integration** [Lu & Zhai 08]
- **Latent Aspect Rating Analysis** [Wang et al. 10]

Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. **Topic Modeling with Network Regularization**, *Proceedings of the World Wide Conference 2008* (**WWW'08**), pages 101-110.

Yue Lu, ChengXiang Zhai. **Opinion Integration Through Semi-supervised Topic Modeling**, *Proceedings of the World Wide Conference 2008* (**WWW'08**), pages 121-130.

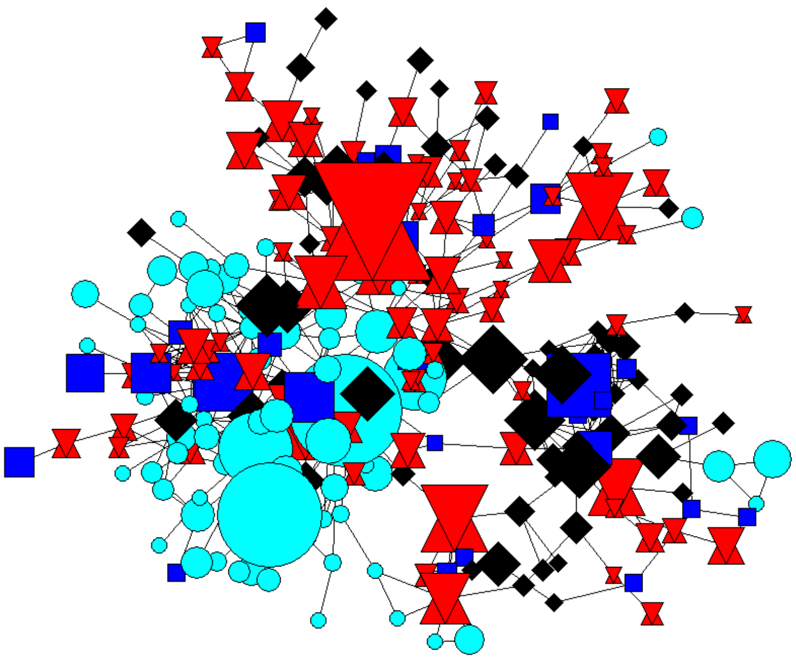
Hongning Wang, Yue Lu, ChengXiang Zhai. **Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach**, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (**KDD'10**), pages 115-124, 2010.

Topic Modeling + Social Networks: who work together on what?

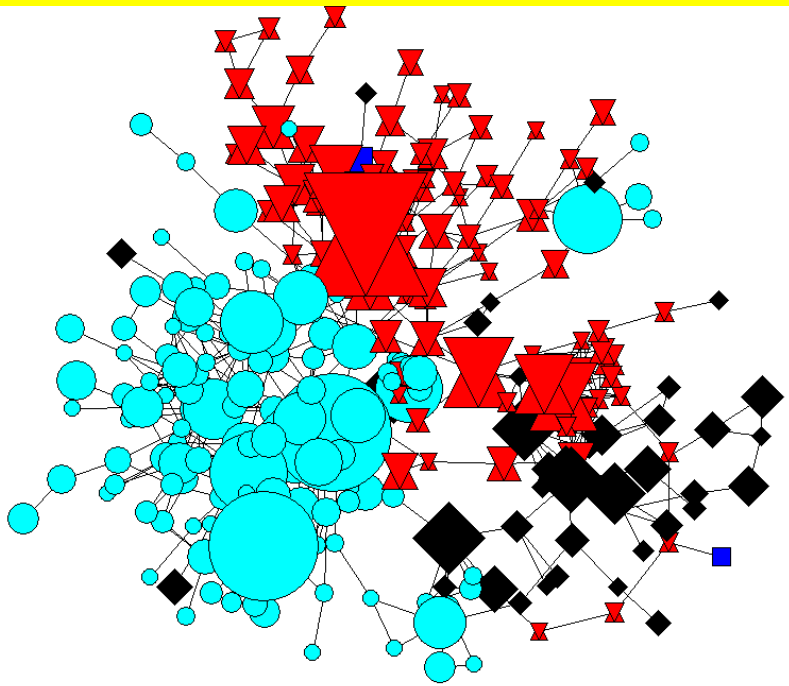
Authors writing about the same topic form a community

Separation of 3 research communities: IR, ML, Web

Topic Model Only



Topic Model + Social Network



Topic Model for Opinion Integration

How to digest all?

Two Kinds of Opinions

How to benefit from both?

Expert opinions
• CNET editor's review • Wikipedia article
• Well-structured • Easy to access
• Maybe biased • Outdated soon

Ordinary opinions
• Forum discussions • Blog articles
• Represent the majority • Up to date
• Hard to access • fragmental

Generate an Integrative Summary

Input

Topic: iPod

Expert review with aspects

DesignBatteryPrice..

Text collection of ordinary opinions, e.g. Weblogs

Output

Review Aspects

	Similar opinions	Supplementary opinions
Design	cute... tiny...	..thicker..
Battery	last many hrs	die out soon
Price	could afford it	still expensive

Extra Aspects

iTunes	... easy to use...
warranty	...better to extend..

Integrated Summary

Methods

- **Semi-Supervised Probabilistic Latent Semantic Analysis (PLSA)**
  - The aspects extracted from expert reviews serve as clues to define a conjugate prior on topics
  - Maximum a Posteriori (MAP) estimation
  - Repeated applications of PLSA to integrate and align opinions in blog articles to expert review

Results: Product (iPhone)

- **Opinion Integration with review aspects**

Review article	Similar opinions	Supplementary opinions
You can make emergency calls, but you can't use any other functions…	N/A	… methods for **unlocking** the iPhone have emerged on the past few weeks, involve tinkering with the iPhone hardware…
rated battery life **hours talk time**, **24 hours of music** playback, **7 hours of video** playback, and 6 hours on Internet use.	…ature Up to **8 Hours of Talk Time**, 6 Hours of Internet Use, **7 Hours of Video** Playback or **24 Hours of Audio** Playback	Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly **9 freaking hours** of continuous playback with cell and WiFi on (but Bluetoot…

Activation

Confirm the opinions from the review

Unlock/hack iPhone

Battery

Additional info under real usage

Results: Product (iPhone)

- **Opinions on extra aspects**

support	Supplementary opinions on extra aspects
15	You may have heard of **iASign** … an iPl … at allows you to **activate** your phone wit … e iTunes rigamarole.
13	**Cisco** has owned the **trademark** on the name "**iPhone**" since 2000, when it acquired InfoGe … hich originally registered the name …
13	With the imminent availabilit … hone, a look at 10 things current smartphones like the **Nokia N95** have been able … the **iPhone can't currently match**...

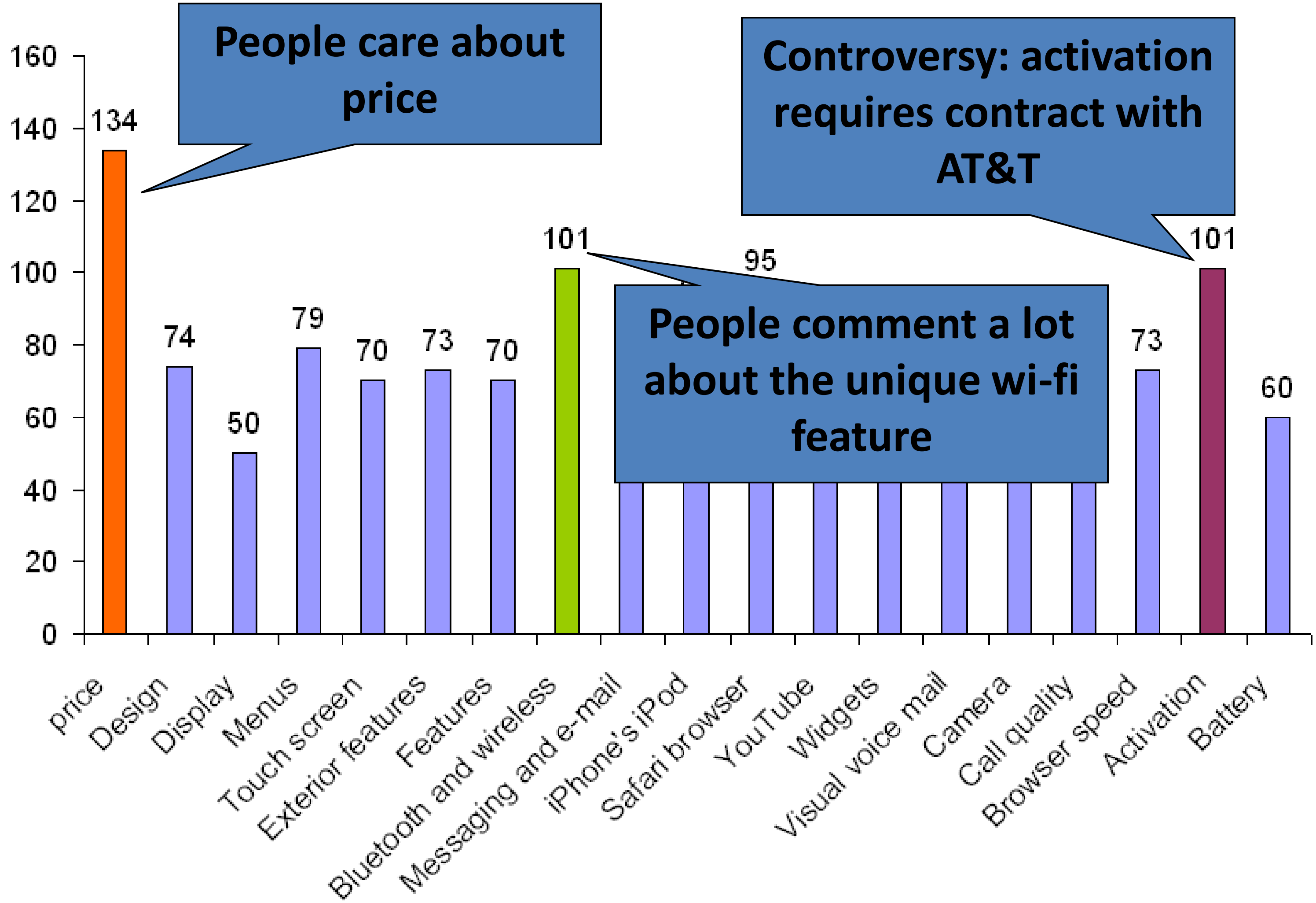
Another way to activate iPhone

iPhone trademark originally owned by Cisco

A better choice for smart phones?

Results: Product (iPhone)

- **Support statistics for review aspects**



Latent Aspect Rating Analysis

Hotel Palomar Chicago: Traveler Reviews

How to infer aspect ratings?

"Great location+spacious room =happy traveler"

leos_10 3 contributions
Boston

Save Review

Jul 11, 2010 | Trip type: Couples NEW

Stayed for a weekend in July. Walked everywhere, enjoyed the comfy bed and quiet hallways. more

My ratings for this hotel

- Value
- Rooms
- Location
- Cleanliness
- Service
- Sleep Quality

"terrific service and gorgeous facility"

ahickling 1 contribution
Greensboro, North Carolina

Save Review

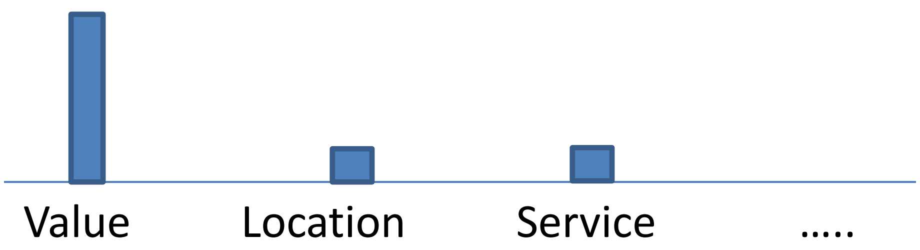
Jul 7, 2010 | Trip type: Family NEW

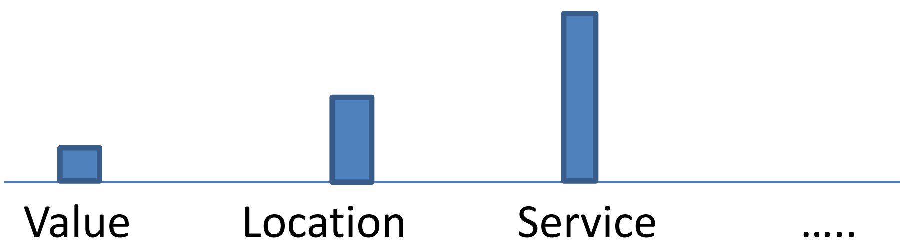
I stayed at the Palomar with my young daughter for three nights June 17-20, 2010 and absolutely loved the hotel. The room was one of the nicest I've ever stayed in (My daughter loved the Fuji jetted tub so much that she wanted to take 2 baths a day!) in terms of decor, design, and size. (It compared favorably to... more

My ratings for this hotel

- Value
- Rooms
- Location
- Cleanliness
- Service
- Sleep Quality

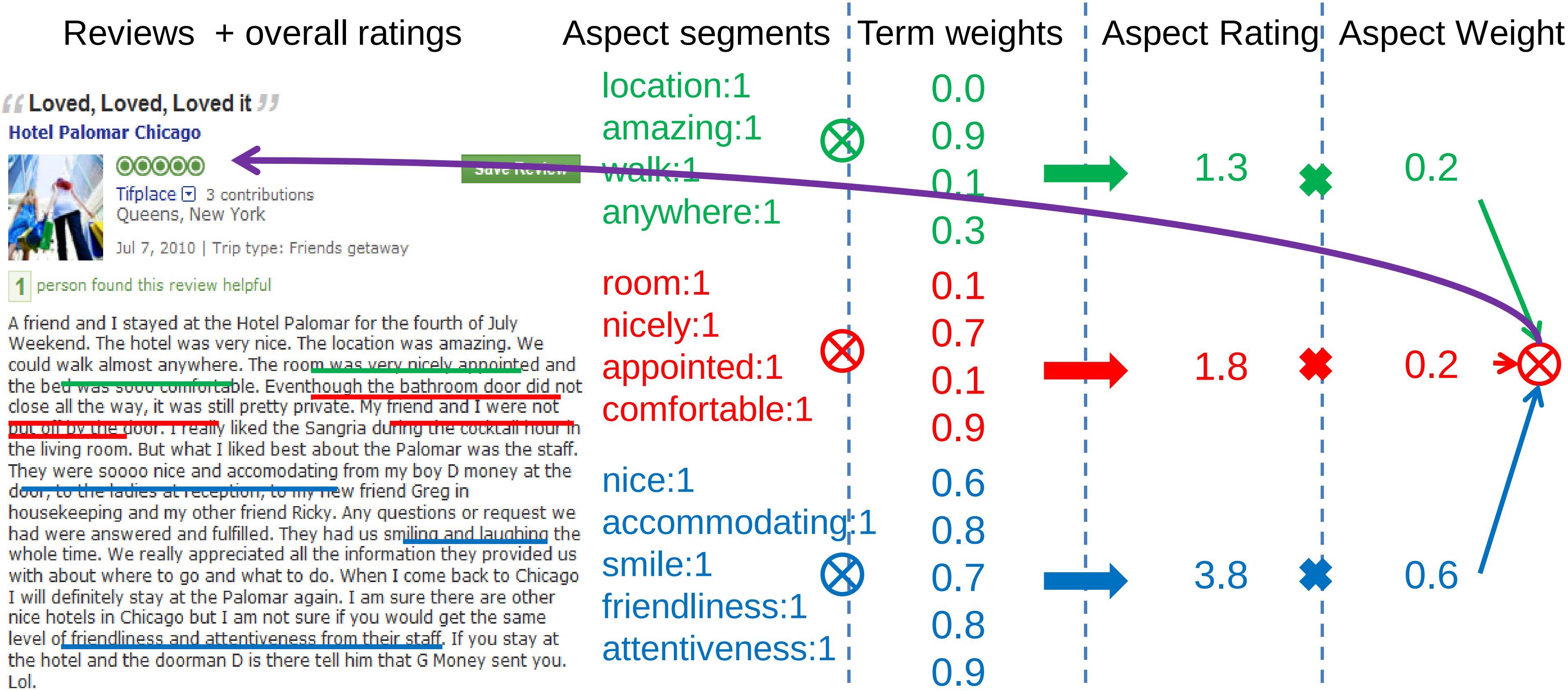
How to infer aspect weights?





Solution: Latent Rating Regression Model

Aspect Segmentation + *Latent Rating Regression*



Topic model for aspect discovery

Aspect-Based Opinion Summarization

Table 6: Aspect-based Comparative Summarization (Hotel Max in Seattle)

Aspect	Summary	Rating
Value	Truly unique character and a great location at a reasonable price Hotel Max was an excellent choice for our recent three night stay in Seattle.	3.1
	Overall not a negative experience, however considering that the hotel industry is very much in the impressing business there was a lot of room for improvement.	1.7
Room	We chose this hotel because there was a Travelzoo deal where the Queen of Art room was $139.00/night.	3.7
	Heating system is a window AC unit that has to be shut off at night or guests will roast.	1.2
Location	The location ,a short walk to downtown and Pike Place market , made the hotel a good choice.	3.5
	when you visit a big metropolitan city, be prepared to hear a little traffic outside!	2.1
Business Service	You can pay for wireless by the day or use the complimentary Internet in the business center behind the lobby though.	2.7
	My only complaint is the daily charge for internet access when you can pretty much connect to wireless on the streets anymore.	0.9

Reviewer Behavior Analysis & Personalized Ranking of Entities

People like expensive hotels because of good service

People like cheap hotels because of good value

Table 4: User behavior analysis

	Expensive Hotel		Cheap Hotel	
Aspect	5 Star	3 Star	5 Star	1 Star
Value	0.134	0.148	0.171	0.093
Room	0.098	0.162	0.126	0.121
Location	0.171	0.074	0.161	0.082
Cleanliness	0.081	0.163	0.116	0.294
Service	0.251	0.101	0.101	0.049

Query: 0.9 value
0.1 others

Non-Personalized

Personalized

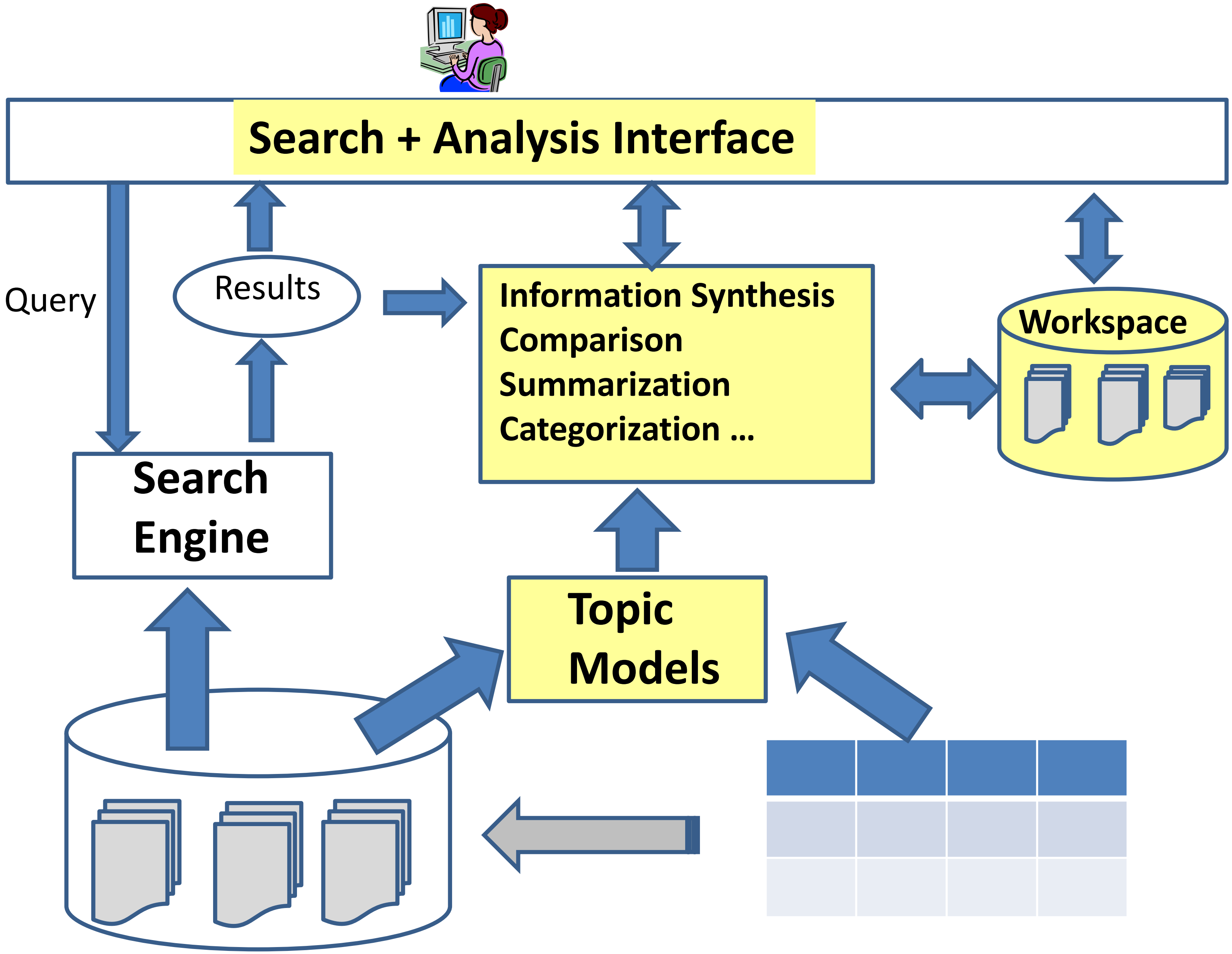
Table 10: Personalized Hotel Ranking

Hotel	Overall Rating	Price	Location
Majestic Colonial	5.0	339	Punta Cana
Agua Resort	5.0	753	Punta Cana
Majestic Elegance	5.0	537	Punta Cana
Grand Palladium	5.0	277	Punta Cana
Iberostar	5.0	157	Punta Cana
Elan Hotel Modern	5.0	216	Los Angeles
Marriott San Juan Resort	4.0	354	San Juan
Punta Cana Club	5.0	409	Punta Cana
Comfort Inn	5.0	155	Boston
Hotel Commonwealth	4.5	313	Boston

How can we extend a search engine to leverage topic models for text analysis?

How should we extend a search engine to support text analysis in general?

Analysis Engine based on Topic Models



Beyond Search: Toward a General Analysis Engine

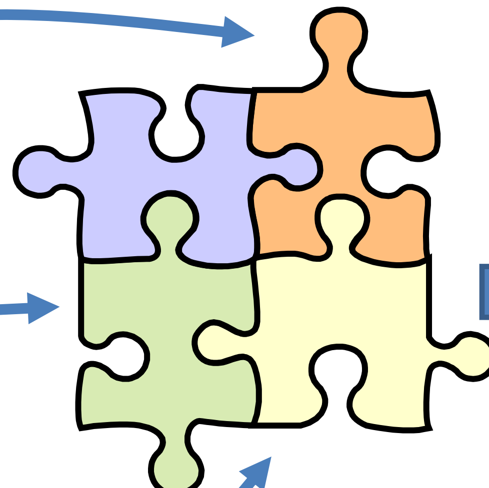
Task Completion

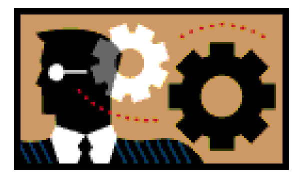
Search → Information Synthesis & Analysis

Search 1

Search 2

...





Decision Making

Learning

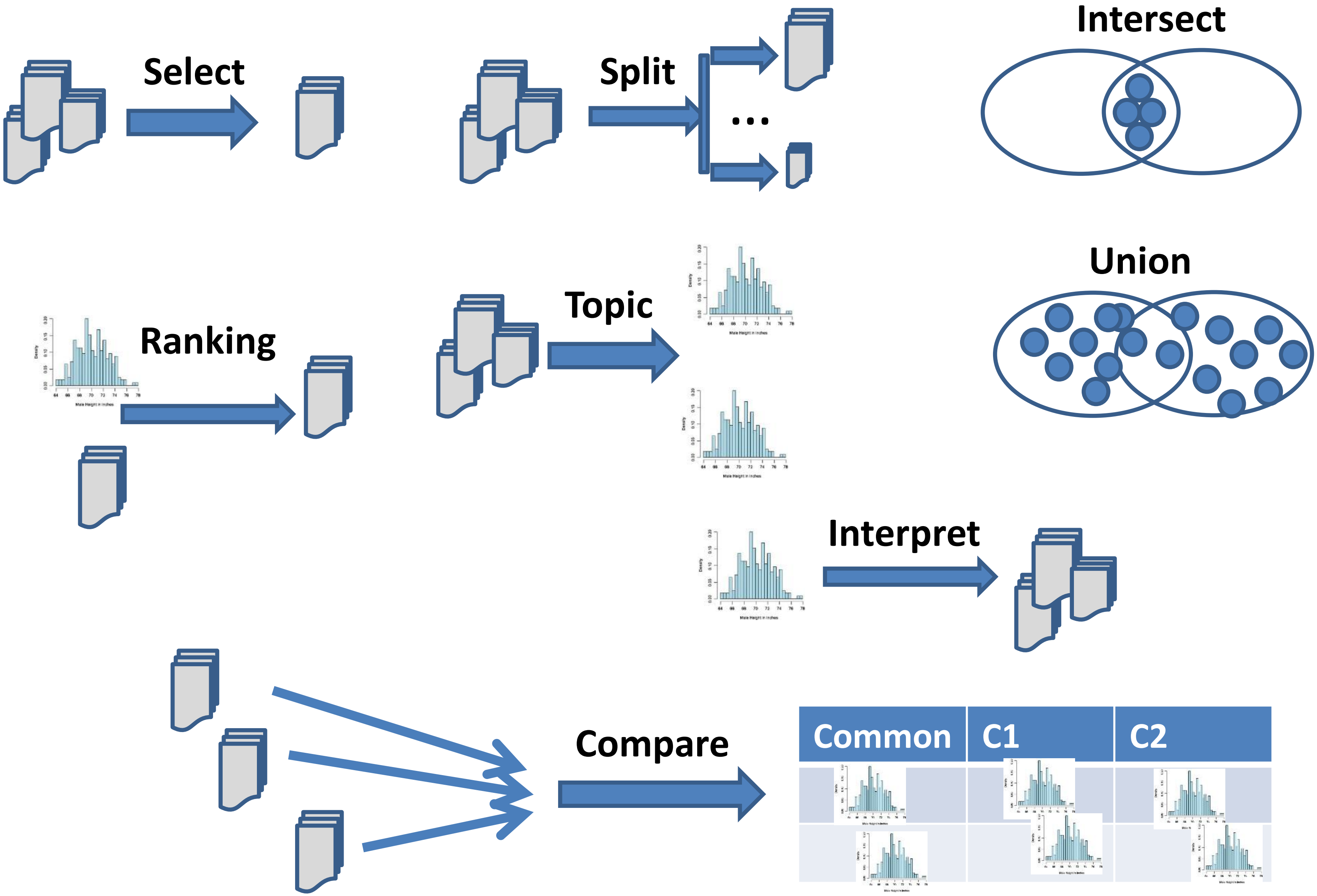
…

Analysis Engine

Challenges in Building a General Analysis Engine

- **What is a "task" and how can we formally model a task? (task vs. intent vs. information needs)**
- **How to design a task specification language?**
- **How do we design a set of general analysis operators to accommodate many different tasks?**
- **What does ranking mean in an analysis engine (ranking terms, documents, topics, operators)?**
- **What should the user interface look like?**
- **How can we seamlessly integrate search and analysis?**
- **How should we evaluate an analysis engine?**
- **…**

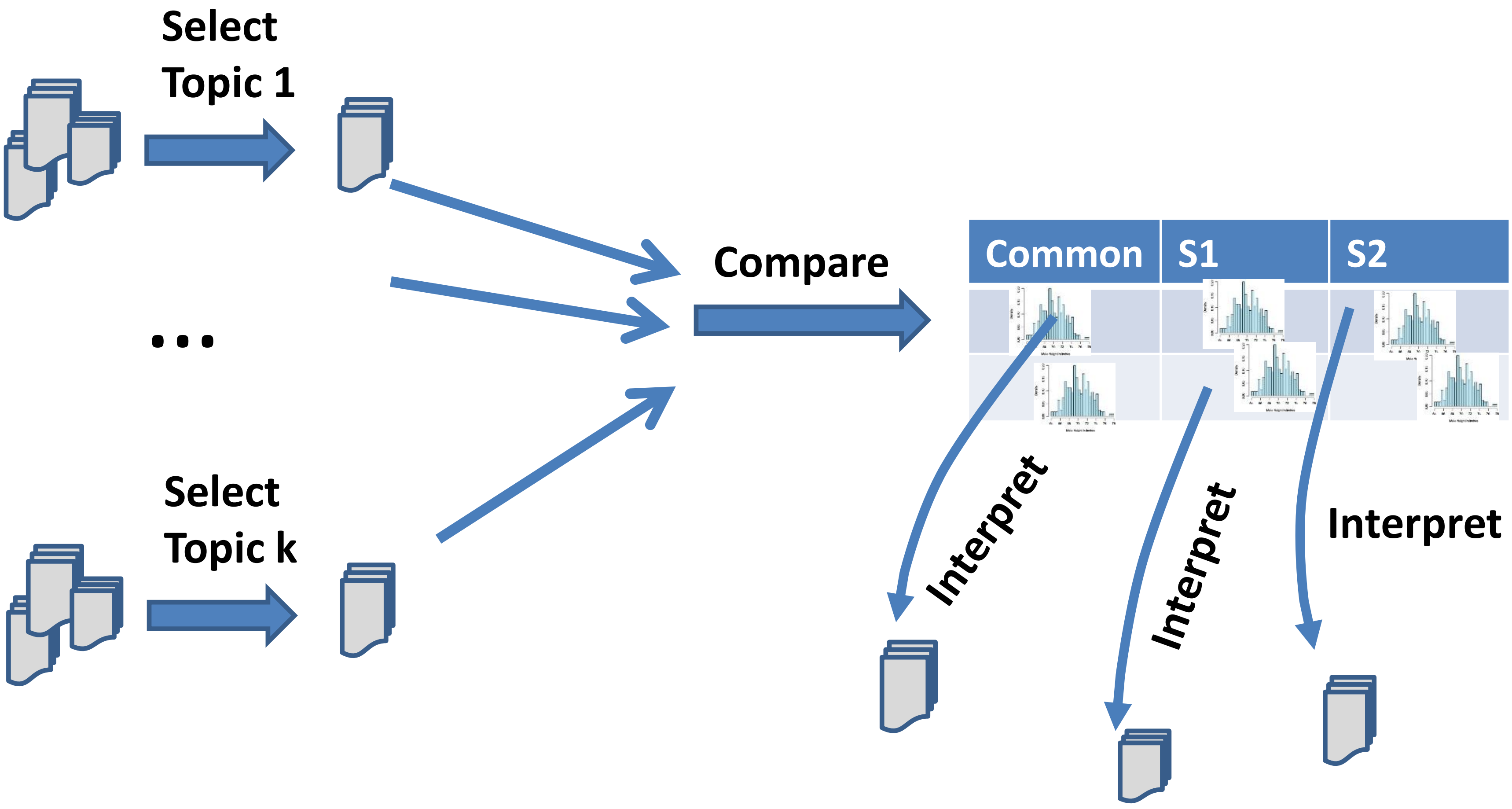
Analysis Operators



Examples of Specific Operators

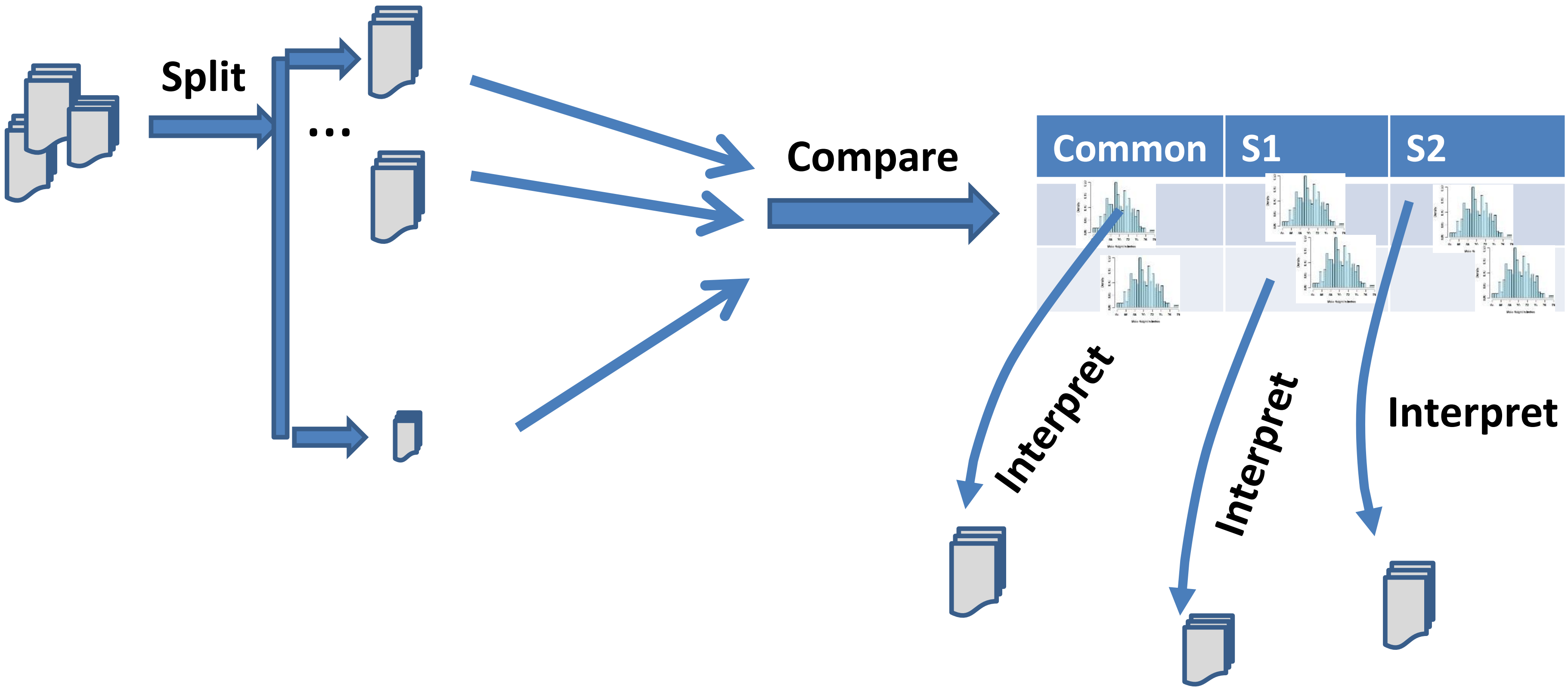
- **C={D1, …, Dn}; S, S1, S2, …, Sk subset of C**
- **Select Operator**
  - Querying(Q): C→ S
  - Browsing: C→S
- **Split**
  - Categorization (supervised): C→ S1, S2, …, Sk
  - Clustering (unsupervised): C→ S1, S2, …, Sk
- **Interpret**
  - C x θ →S
- **Ranking**
  - θ x Si → ordered Si

Compound Analysis Operator: Comparison of K Topics



Interpret(Compare(Select(T1,C), Select(T2,C),…Select(Tk,C)),C)

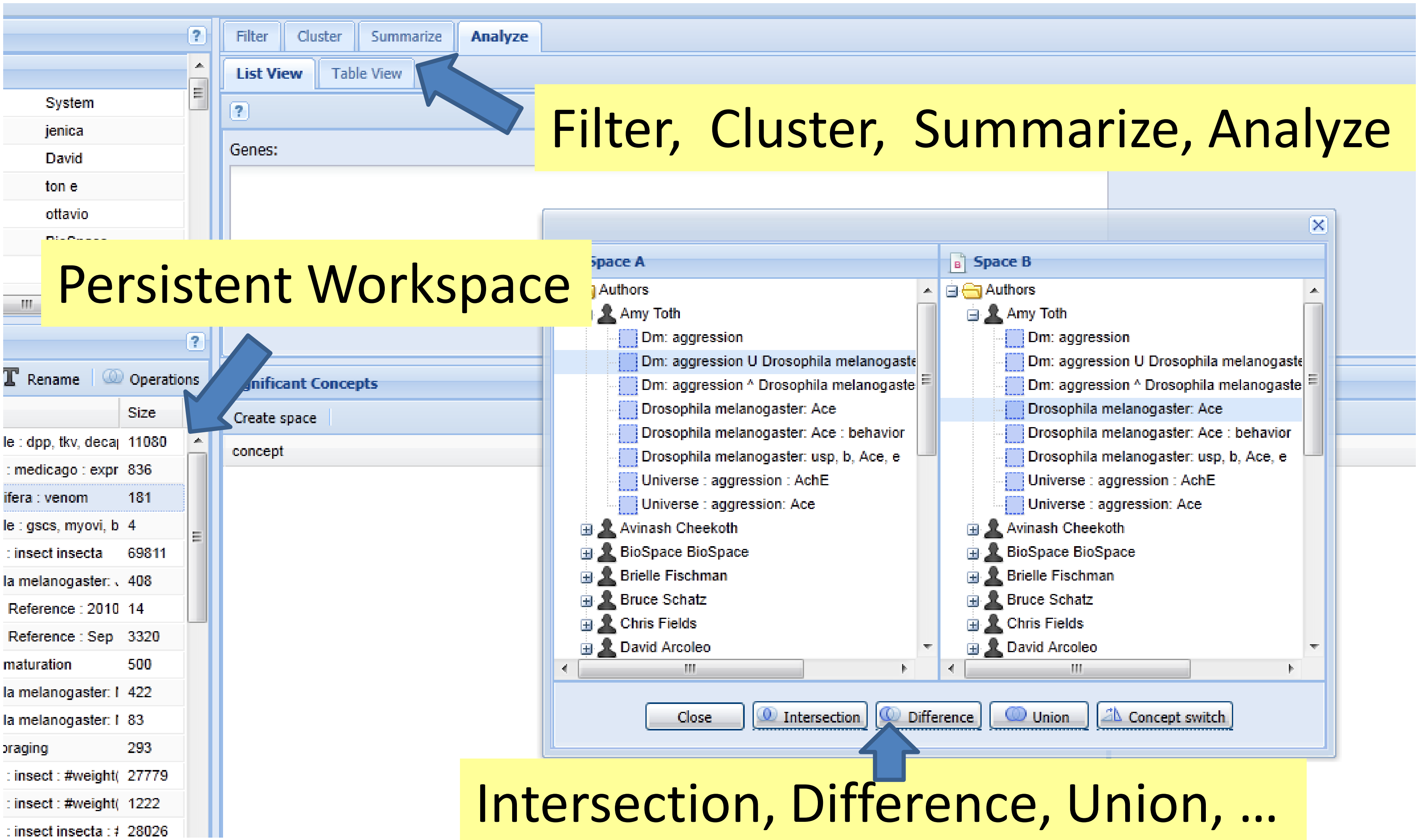
Compound Analysis Operator: Split and Compare



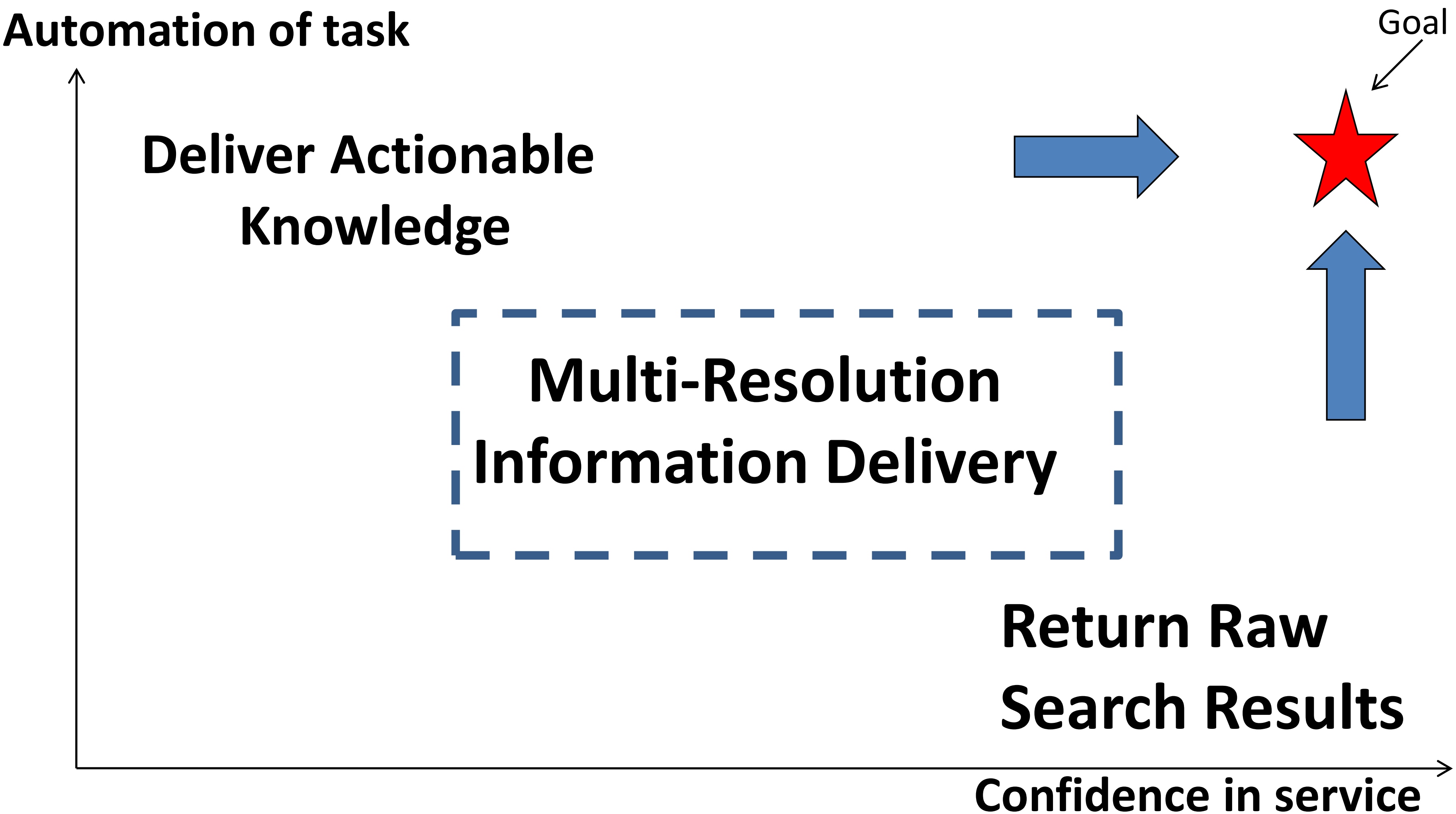
Interpret(Compare(Split(S,k)),C)

BeeSpace System

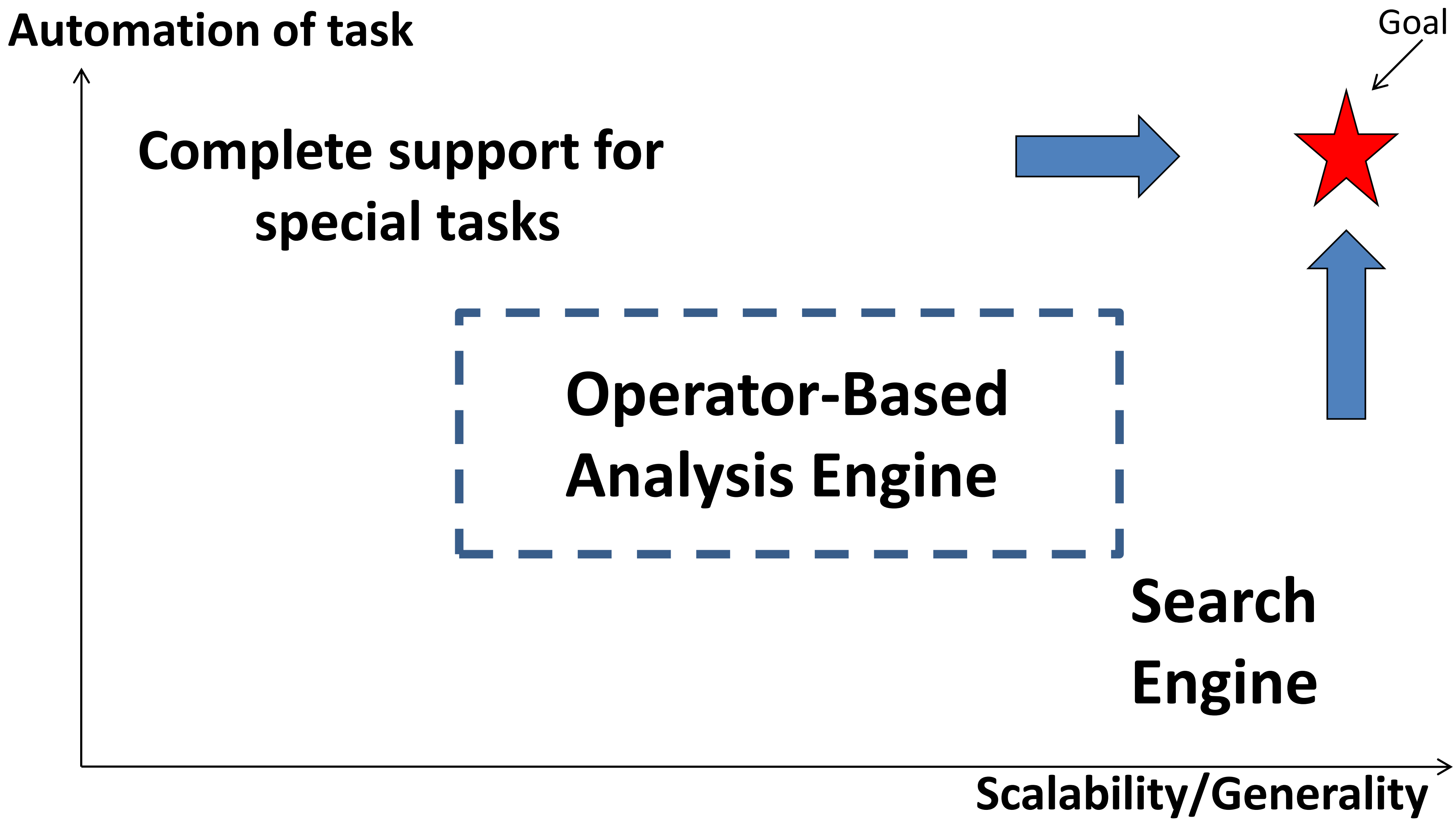
Sarma, M.S., et al. (2011) BeeSpace Navigator: exploratory analysis of gene function using semantic indexing of biological literature. *Nucleic Acids Research*, 2011, 1-8, doi:10.1093/nar/gkr285.



Automation-Confidence (AC) Tradeoff



Automation-Generality (AG) Tradeoff



Automation-Confidence Tradeoff: Dining Analogy

Serve Raw-Food

Need further processing,
but flexible for making different dishes

Serve Cooked Dishes

Directly useful for a task,
But would be worse if it's not the right dish

?

Automation-Generality Tradeoff: Dining Analogy

What's the right paradigm? Need both paradigms?

Buffet Paradigm

Basic Components + Infinite Combination

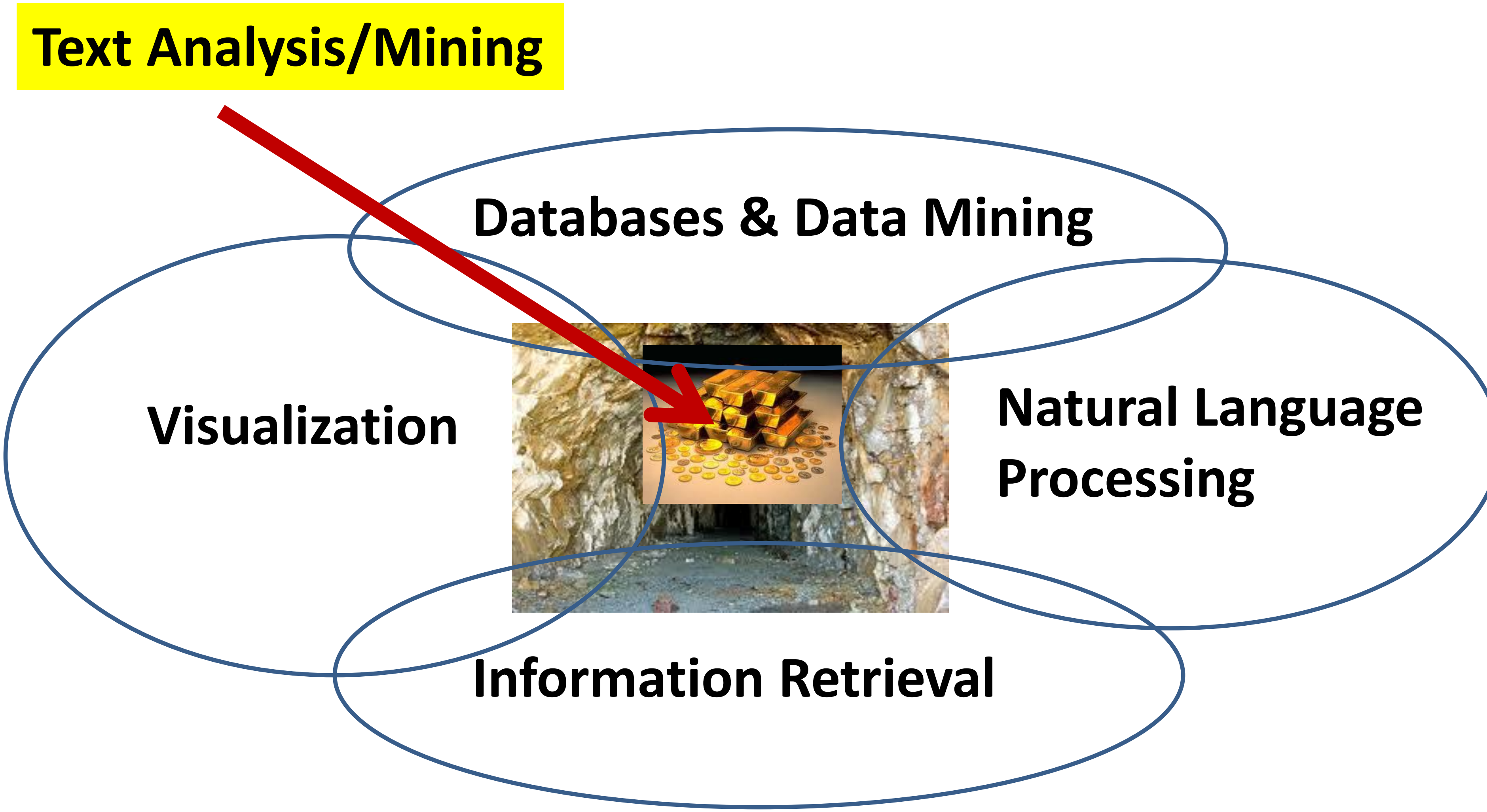
Food Court Paradigm

Finite Choices of Complete Packages

Summary

- **Statistical topic models are promising general tools for supporting text analysis**
- **Next-generation search engines should go beyond search to seamlessly support text analysis and better help users complete their tasks**
- **Many challenges to be solved:**
  - Task modeling
  - Task specification language
  - New analysis operators
  - New ranking models
  - New interface issues
  - New evaluation challenges
  - Automation-Generality (AG) tradeoff & Automation-Confidence (AC) tradeoff
  - …

Looking Ahead…



Acknowledgments

- **Collaborators: Qiaozhu Mei, Yue Lu, Hongning Wang, Jiawei Han, Bruce Schatz, and many others**
- **Funding**

Thank You!

Questions/Comments?