Interactive Visual Text Analytics for Decision Making

Shixia Liu

Microsoft Research Asia

Text is Everywhere

- **We use documents as primary information artifact in our lives**
- **Our access to documents has grown tremendously in recent years due to networking infrastructure**
  - WWW
  - Digital libraries
  - …

Big Question

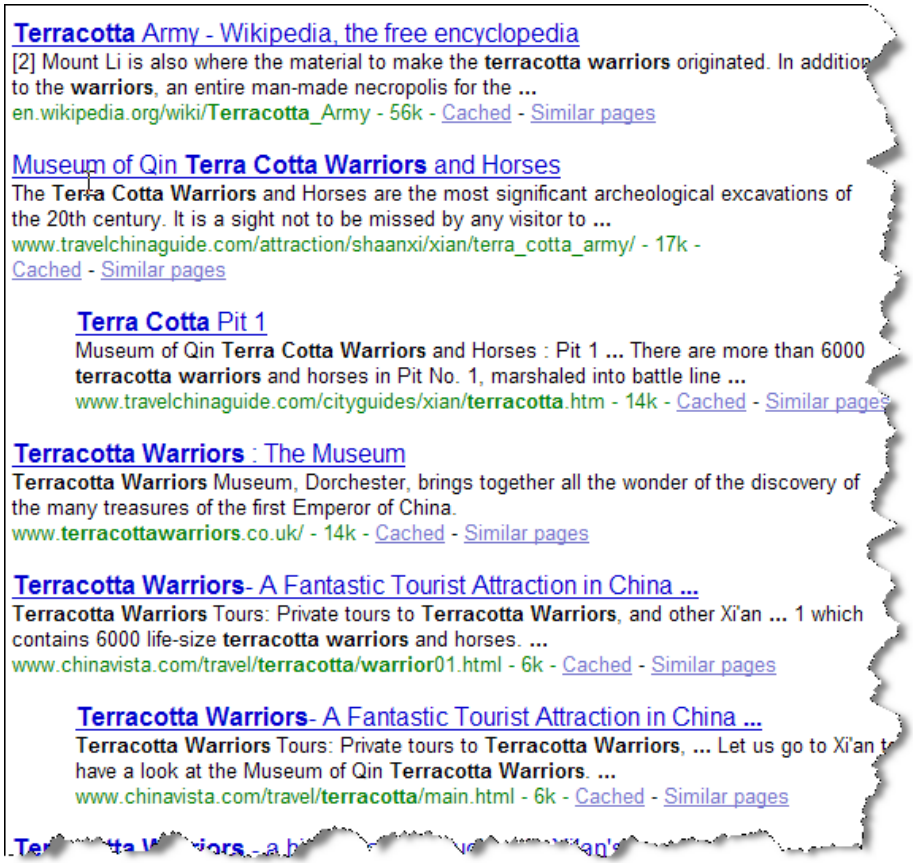
- What can information visualization provide to help users in understanding and gathering information from text and document collections?

Outline

- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text
- **Text Visualization Perspectives**

Text Analytics: Our Understanding

How can I find information buried inside the piles of text?



Terracotta Army - Wikipedia, the free encyclopedia
[2] Mount Li is also where the material to make the **terracotta warriors** originated. In addition to the **warriors**, an entire man-made necropolis for the ...
en.wikipedia.org/wiki/**Terracotta**_Army - 56k - Cached - Similar pages

Museum of Qin **Terra Cotta Warriors** and Horses
The **Terra Cotta Warriors** and Horses are the most significant archeological excavations of the 20th century. It is a sight not to be missed by any visitor to ...
www.travelchinaguide.com/attraction/shaanxi/xian/terra_cotta_army/ - 17k - Cached - Similar pages

Terra Cotta Pit 1
Museum of Qin **Terra Cotta Warriors** and Horses : Pit 1 ... There are more than 6000 **terracotta warriors** and horses in Pit No. 1, marshaled into battle line ...
www.travelchinaguide.com/cityguides/xian/**terracotta**.htm - 14k - Cached - Similar pages

Terracotta Warriors : The Museum
Terracotta Warriors Museum, Dorchester, brings together all the wonder of the discovery of the many treasures of the first Emperor of China.
www.**terracottawarriors**.co.uk/ - 14k - Cached - Similar pages

Terracotta Warriors- A Fantastic Tourist Attraction in China ...
Terracotta Warriors Tours: Private tours to **Terracotta Warriors**, and other Xi'an ... 1 which contains 6000 life-size **terracotta warriors** and horses. ...
www.chinavista.com/travel/**terracotta**/**warrior**01.html - 6k - Cached - Similar pages

Terracotta Warriors- A Fantastic Tourist Attraction in China ...
Terracotta Warriors Tours: Private tours to **Terracotta Warriors**, ... Let us go to Xi'an to have a look at the Museum of Qin **Terracotta Warriors**. ...
www.chinavista.com/travel/**terracotta**/main.html - 6k - Cached - Similar pages

Information finding

Text Analytics: Our Understanding

What is in my text?

What's inside the NHTSA Data:	**What are the major causes of injuries**	**What did my customers say about my hotels**
450,000+ documents	70,000+ patient emergency room records	3000+ customer-posted reviews

Information Understanding: Text Summarization

Text Analytics: Our Understanding

What is in my text?

Which hotel features do my customers like/dislike 3000+ customer reviews	**How customers' sentiment have changed toward my hotels** 3000+ customer-posted reviews	**How do customers feel about my new product launch** thousands of e-opinion postings

Insight Discovery: Sentiment Analysis

Text Analytics: Our Understanding

What is in my text?

What are the correlations of tire problems and highway death in the NHTSA Data: 450,000+ documents	**What are the correlations of patient gender and the cause of injury** 70,000+ patient emergency room records	**Compare the customers' attitude toward our product with theirs for our competitors** thousands of e-opinion postings

Decision Making and Problem Solving: Text Analysis++

Major Challenges

- **Huge amounts of complex information**
  - Understanding the meanings of free text is just hard
  - Performing analysis on top of that is harder
- **Different people want different things**
  - No one-size-fits-all solutions
- **People may not know what they want**
  - “Tell me something I don’t know”
  - “I will tell you when I see it”

Machines are *not* just smart enough.

Outline

- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text
- **Text Visualization Perspectives**

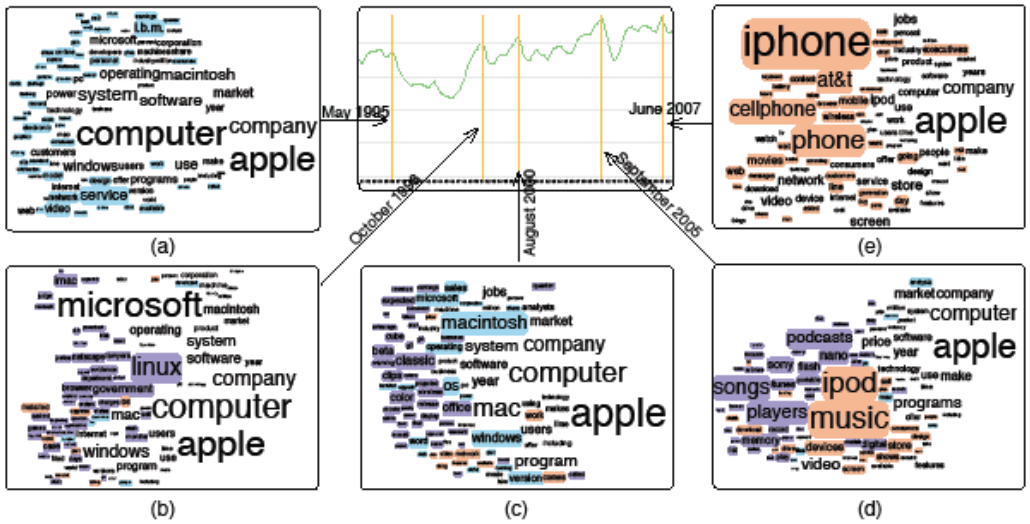
Cui et al. IEEE CG&A 2011

Dynamic Word Cloud

- **Word clouds for content overview**
  - Aesthetic issues
  - Inadequate for temporal patterns
- **Standard time chart: trend**
  - Inadequate for correlations

Our Solution

- A evolution trend chart + word clouds
  - Measure the evolution
  - Ensure the semantic coherence between clouds



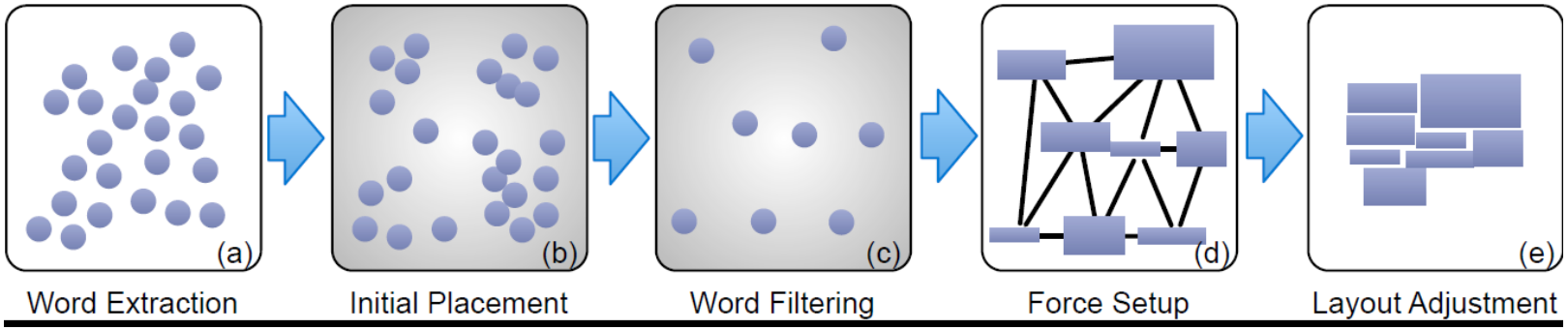
Evolution Measurement

- Conditional entropy: measure the amount of information contained by X_i but not by X_j

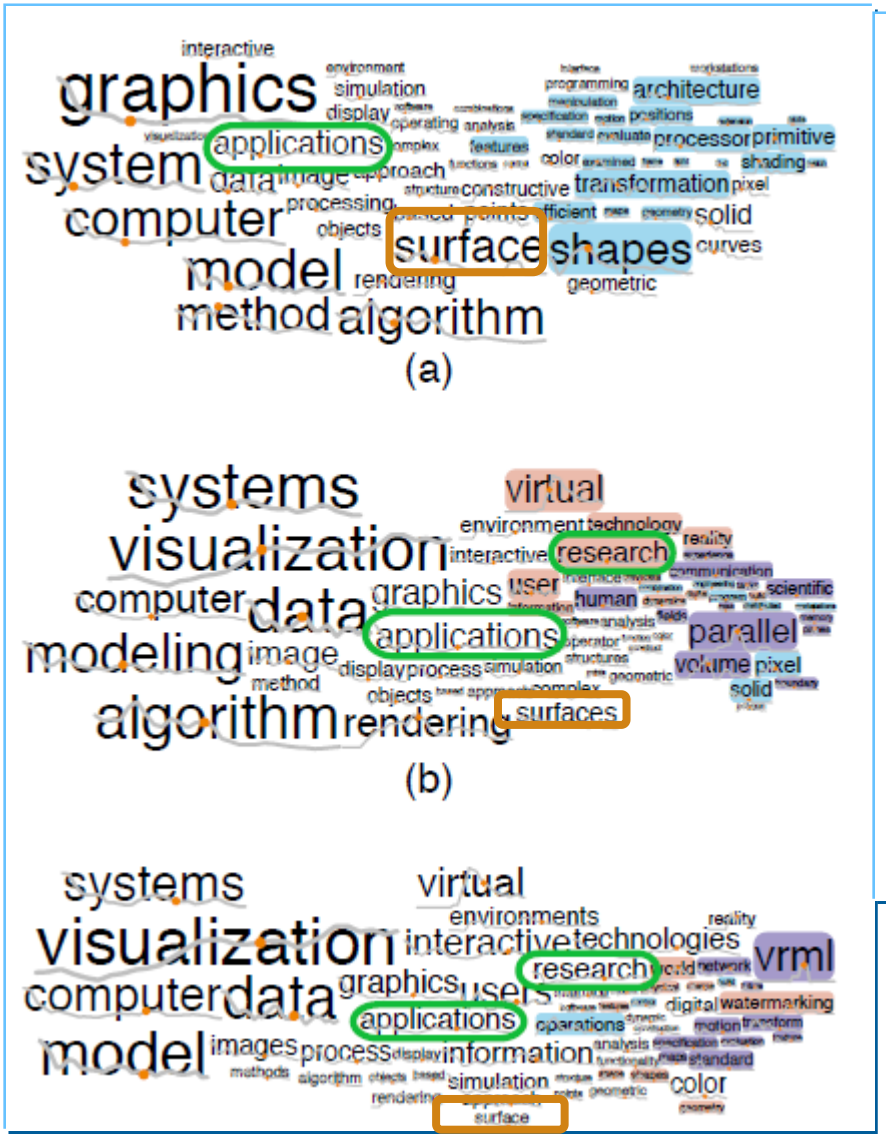
$$S(X_i) = \sum_{j=-w/2}^{w/2} t_j H(X_i|X_{i+j}) = \sum_{j=-w/2}^{w/2} t_j (H(X_i) - H(X_i; X_{i+j}))$$

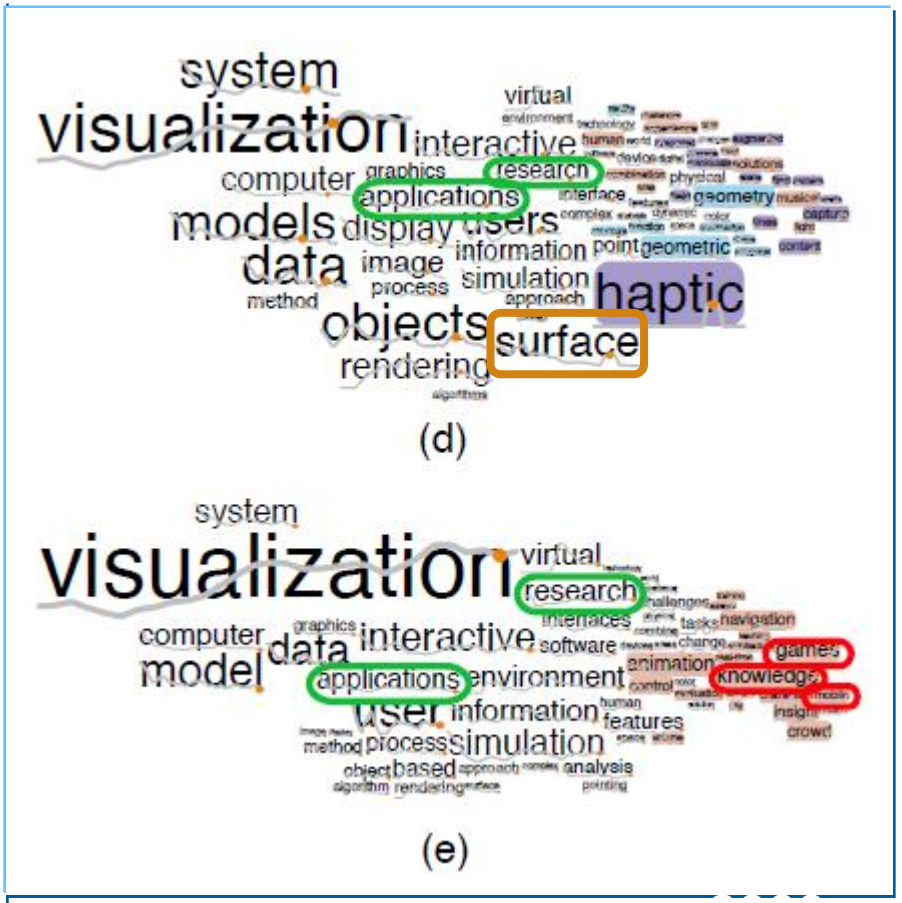
Word Cloud Layout

- Geometry meshes to ensure the semantic coherence
  - Semantically related words stay together
  - The same word in different clouds stay at the similar place



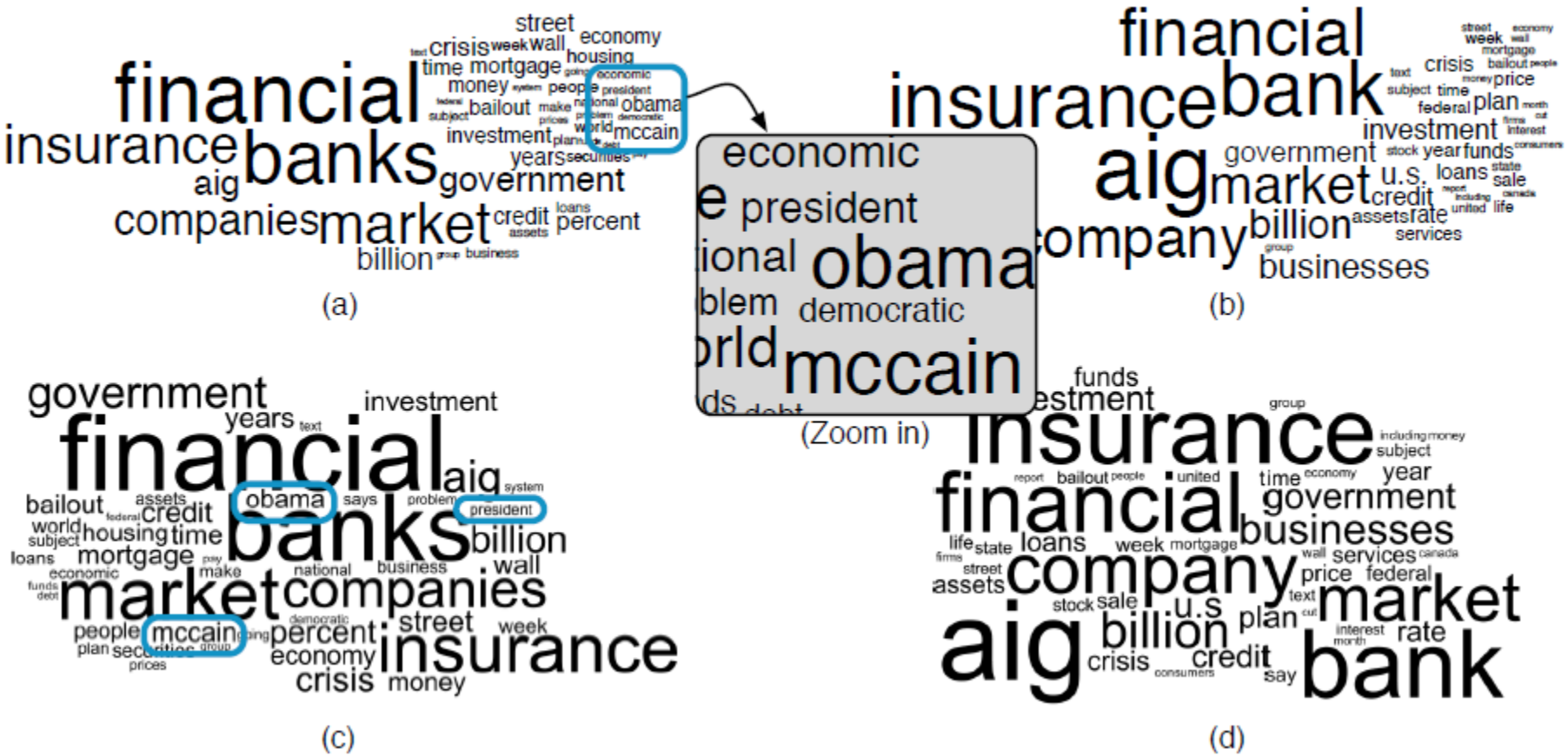
Example: CG&A Abstracts





1,984 abstracts IEEE Computer Graphics and Applications (CG&A) from 1981 to 2009

Comparison with Wordle



13,828 news articles

Outline

- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text
- **Text Visualization Perspectives**

Liu et al. CIKM 09

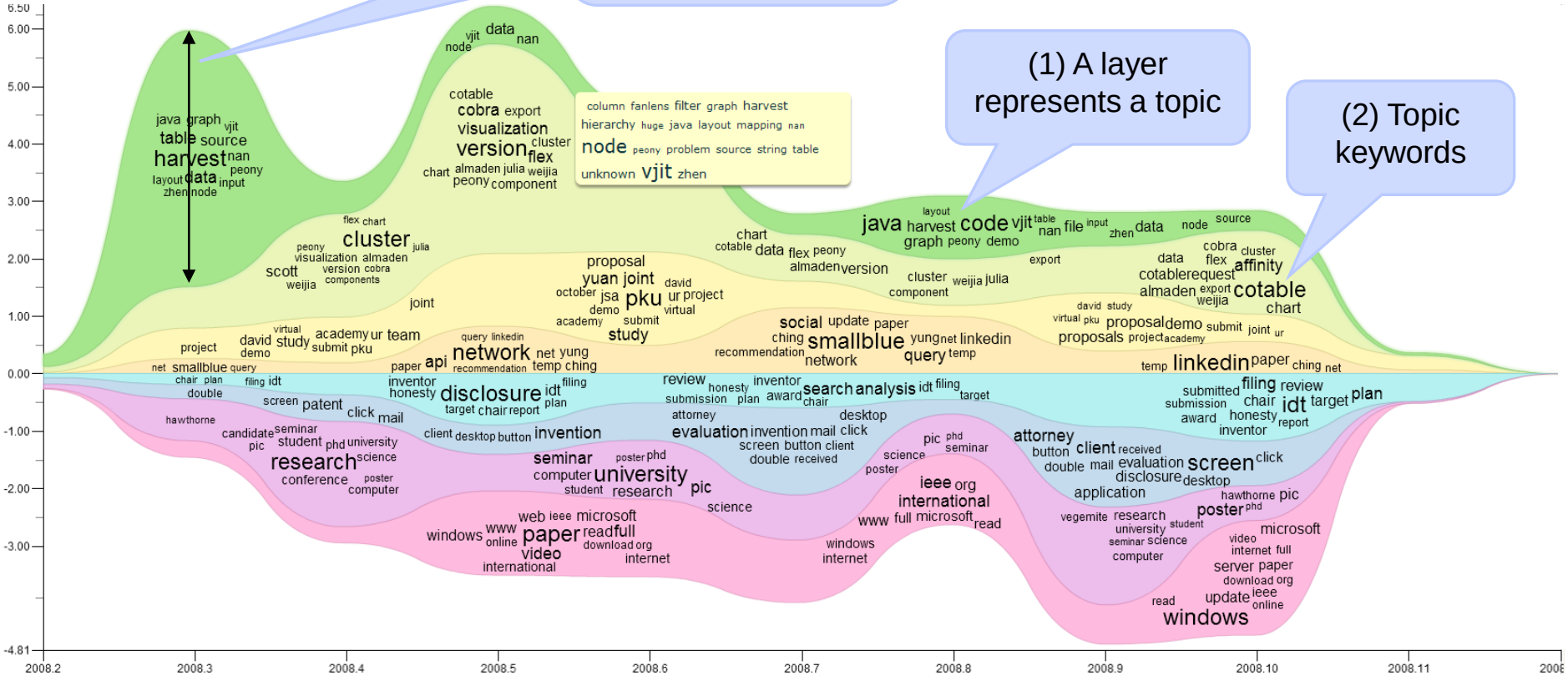
Demo: Topic-based Visual Text Summarization

Y axis encodes topic significance

(3) Height encodes the number of emails in the topic at this time

(1) A layer represents a topic

(2) Topic keywords



X axis encodes time

~10,000 emails in 2008

Demo

Interactive, Time-based
Visual Email Summarization

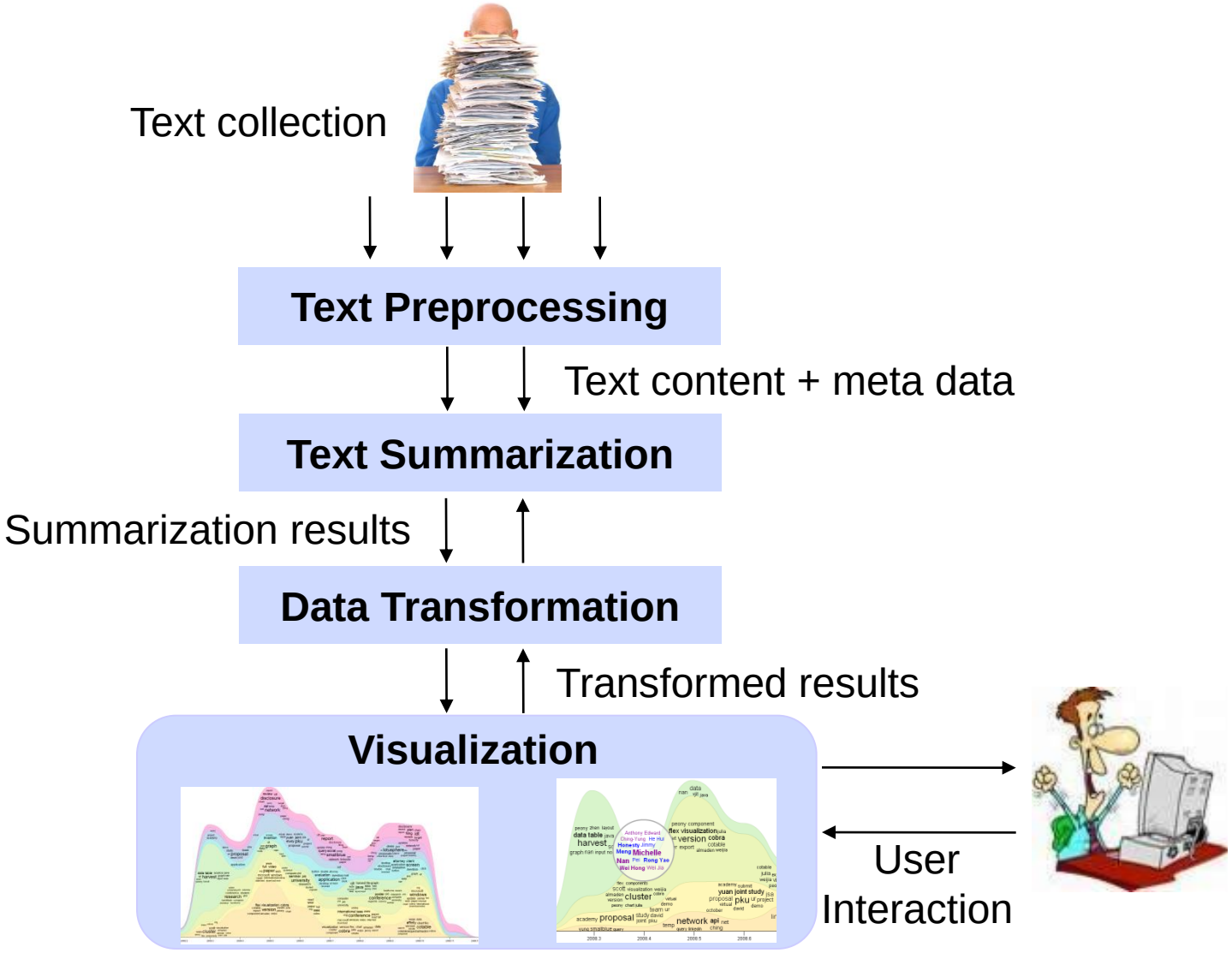
Shixia Liu, Michelle X Zhou, Shimei Pan,
Weihong Qian, Weijia Cai, Xiaoxiao Lian

IBM Research

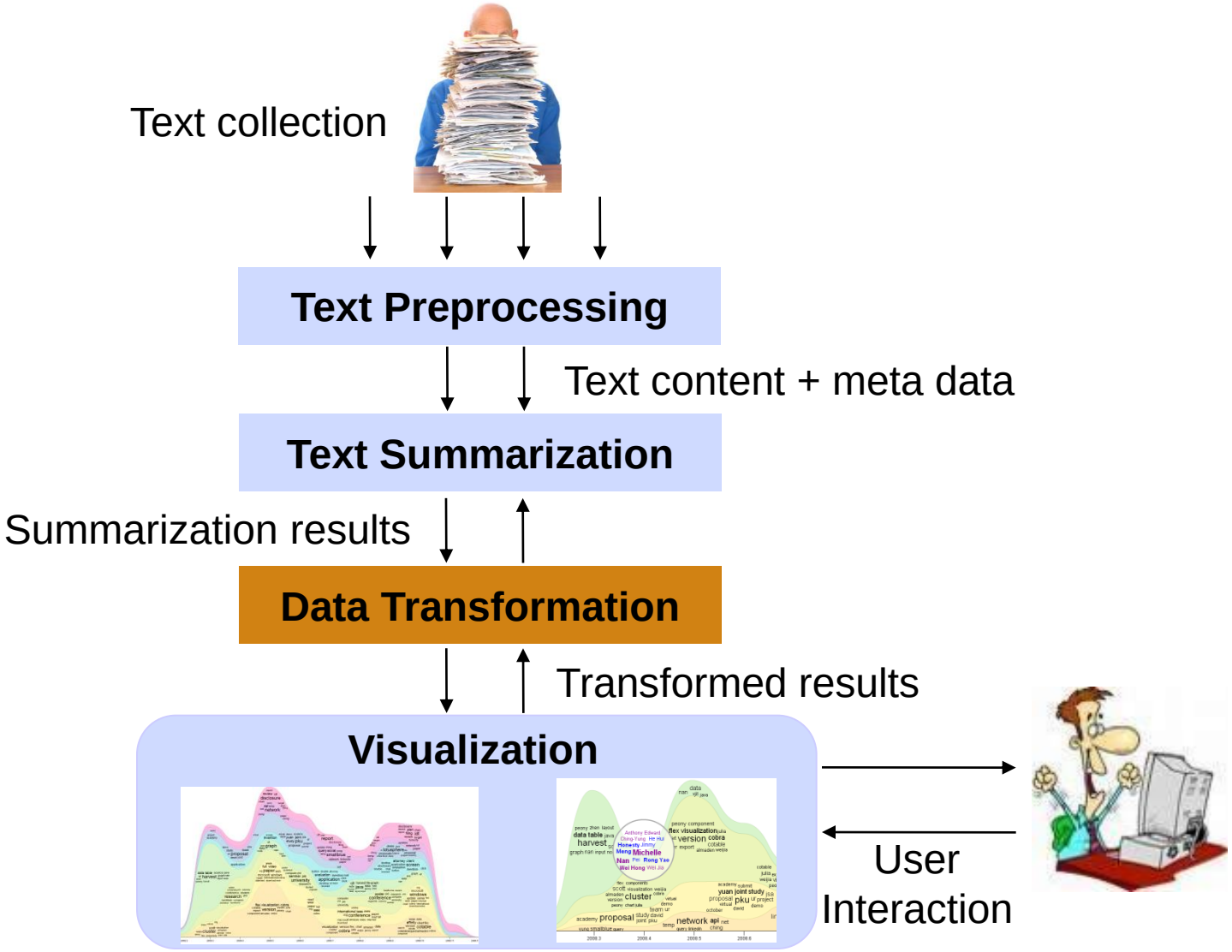
Visual Text Summarization: Key Challenges

- **How to summarize massive, time-varying text corpora**
  - Huge amounts of complex information
  - Accuracy + extensibility
  - Time-varying
  - Effectiveness
- **How to visually explain text summarization results**
  - Which visual metaphors to use
  - How to switch between visualizations
- **How to allow users to provide feedback or articulate their needs**
  - Incorrect summarization results or varied user needs

TIARA Overview

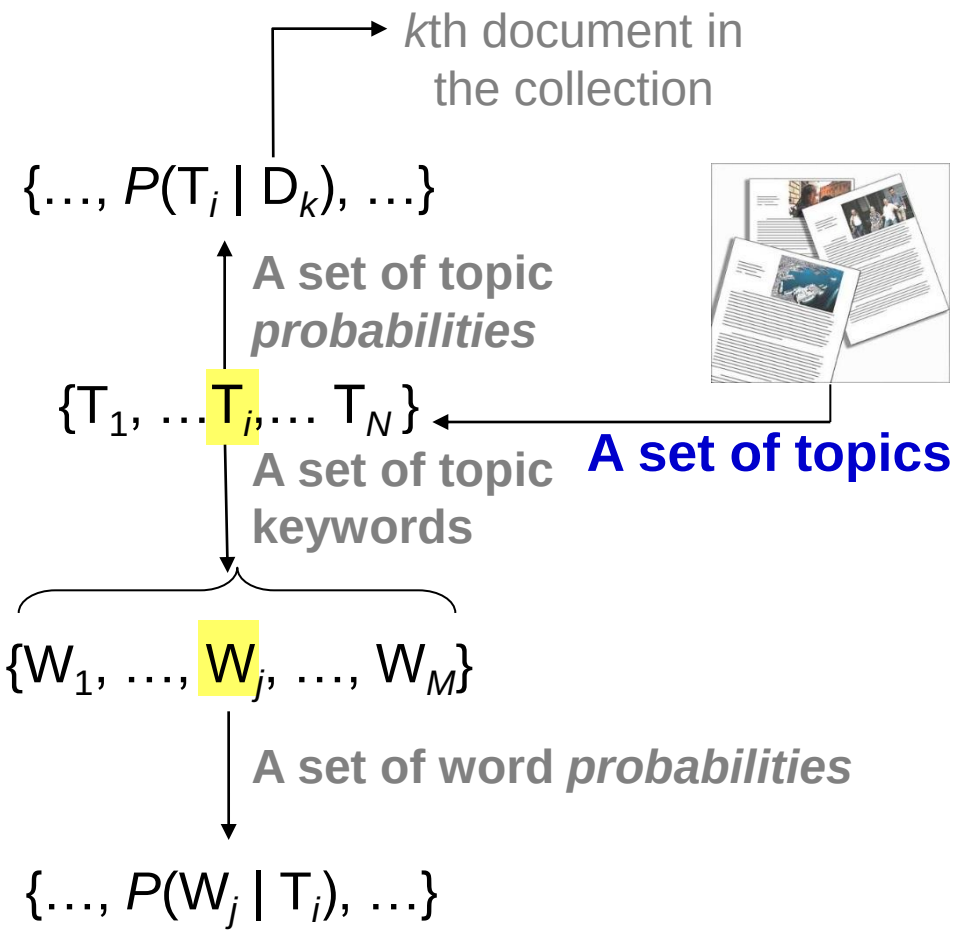


TIARA Technical Focus



Adopted Text Summarization: LDA-based Topic Analysis

- **Latent Dirichlet Allocation (LDA) model [Blei et al. 03]**
  - Statistical analysis that requires no extra knowledge about the language or the world (**high portability**)
  - Keyword-based description of thematic structure of a text collection (**high compaction rate for scalability**)
  - A **finer grained model** than clustering
    - One document → multiple topics → finer-grained summarization



LDA Data Transformation

A set of topics

$\{T_1, \ldots T_i, \ldots T_N\}$

A set of topic *probabilities*

$\{\ldots, P(T_i \mid D_k), \ldots\}$

*k*th document in the collection

A set of keywords

$\{W_1, \ldots, W_j, \ldots, W_M\}$

A set of word *probabilities*

$\{\ldots, P(W_j \mid T_i), \ldots\}$

Rank the topics to present most interesting ones first

Select keyword sub-sets for time-based summary

$\{\ldots\}_{t-1}, \{\ldots, W_j, \ldots\}_t, \{\ldots\}_{t+1},$

[Song et al. CIKM 09]

Topic Ranking by User Interests

- **Rank topics by "strength"**

$$rank(T_k) = f\left(\mu(T_k), \sigma(T_k), \alpha(T_k)\right)$$

Domain-dependent activeness metric

topic coverage

$$\mu(T_k) = \frac{\sum_{m=1}^{M} N_m \hat{\theta}_{m,k}}{\sum_{m=1}^{M} N_m}$$

topic variance

$$\sigma(T_k) = \sqrt{\frac{\sum_{m=1}^{M} N_m \left(\hat{\theta}_{m,k} - \mu(T_k)\right)}{\sum_{m=1}^{M} N_m}}$$

- **Rank topics by "distinctiveness"**

$$rank(T_k) = l(T_k) = \frac{\tilde{v}_k^T L \tilde{v}_k}{\tilde{v}_k^T D \tilde{v}_k}$$

graph Laplacian

graph degree matrix

doc-topic distribution

for each T_k, $v_k = \left(\hat{\theta}_{1,k}, \hat{\theta}_{2,k}, \ldots, \hat{\theta}_{M,k}\right)^T$

$\tilde{v}_k$ is normalized v_k

Topic Ranking by User Interests: Experiments

- **Goal**
  - Measure which metric produces more “important” topics
- **Data sets**
  - Email
    - 8326 email messages
    - 958,069 word tokens
  - News
    - 34,690 documents
    - 11,491,246 word tokens
- **Method**
  - Users indicate the importance of top-K ranked topics
    - Very important
    - Somewhat important
    - Unimportant

Topic Ranking by User Interests: Results

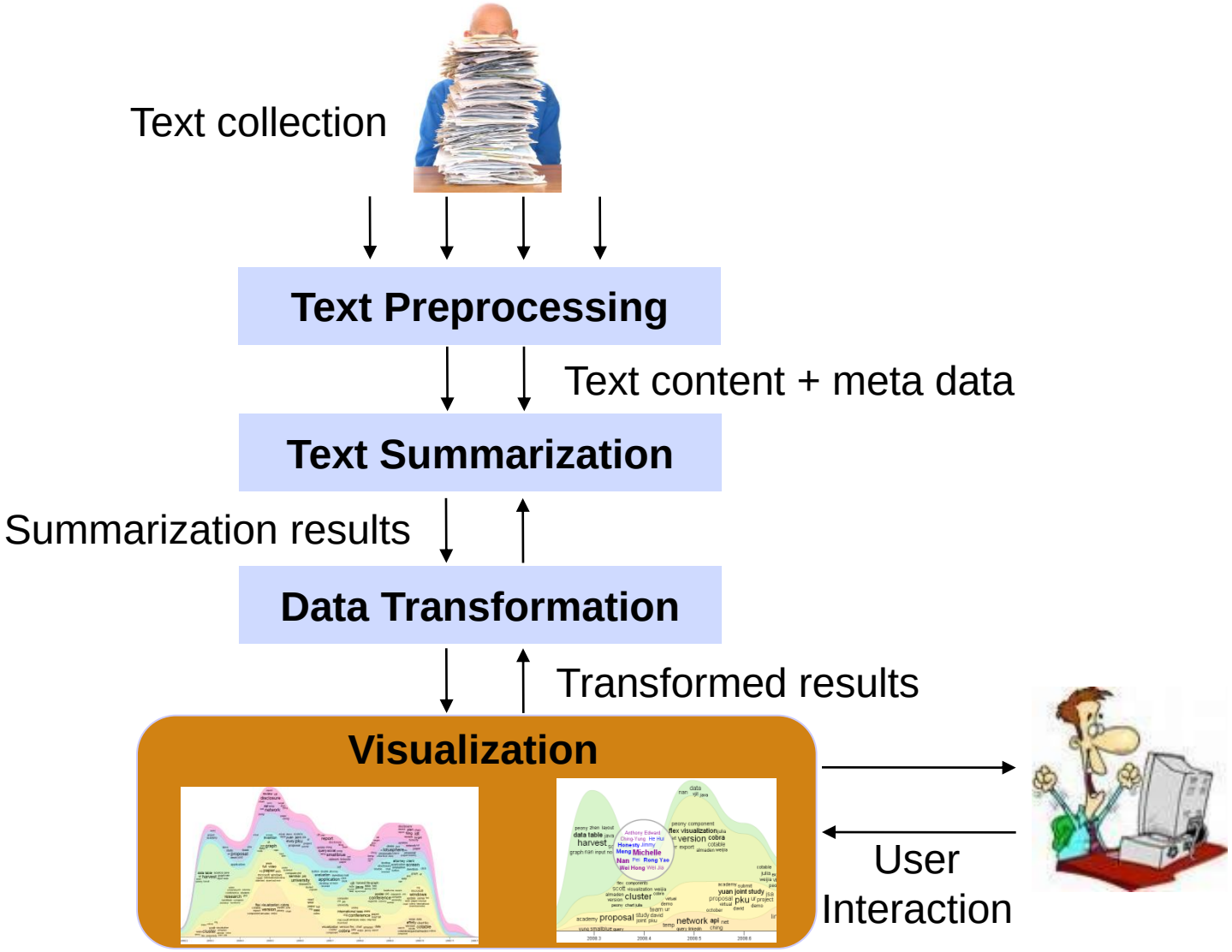
- **Email data (by F1 measure)**

Retrieved	Top 5	Top 10
strength	0.760 ± 0.057	0.640 ± 0.035
distinctiveness	0.920 ± 0.069	**0.900 ± 0.000**
M.I.	**0.960 ± 0.057**	0.860 ± 0.035
T.S.	0.520 ± 0.056	0.560 ± 0.035

- **News data (by F1 measure)**

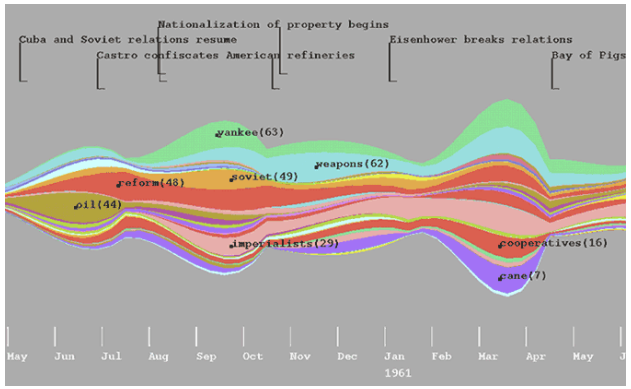
Retrieved	Top 5	Top 10
strength	0.640 ± 0.057	0.68 ± 0.028
distinctiveness	**0.760 ± 0.057**	**0.76 ± 0.035**
M.I.	**0.760 ± 0.057**	0.74 ± 0.035
T.S.	0.720 ± 0.069	0.70 ± 0.045

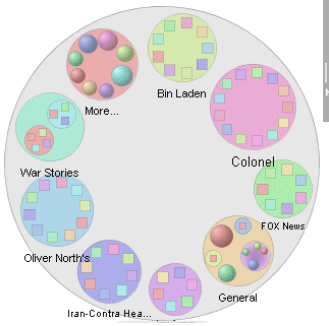
TIARA Technical Focus

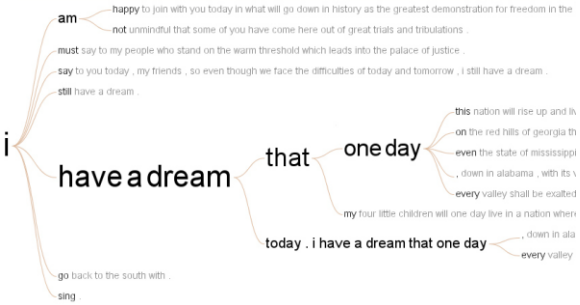


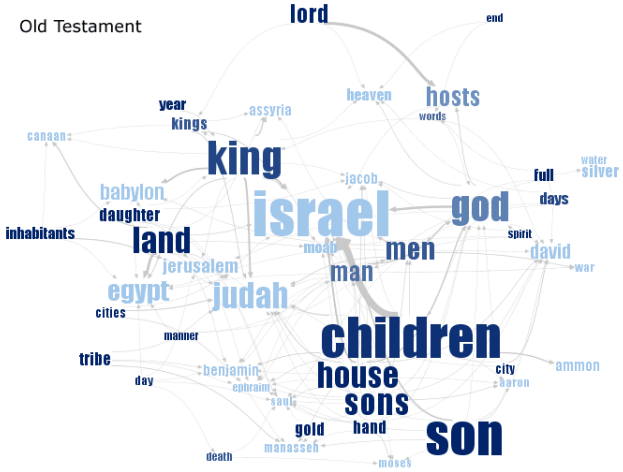
Visualizing Text: Existing Work

- **Visualize text at a high level (e.g., documents)**
  - Issues
    - Inadequate information
- **Visualize text at a low level (e.g., words and phrases)**
  - Issues
    - Missing a big picture
    - Scalability
- **Few on explaining advanced text analysis results**





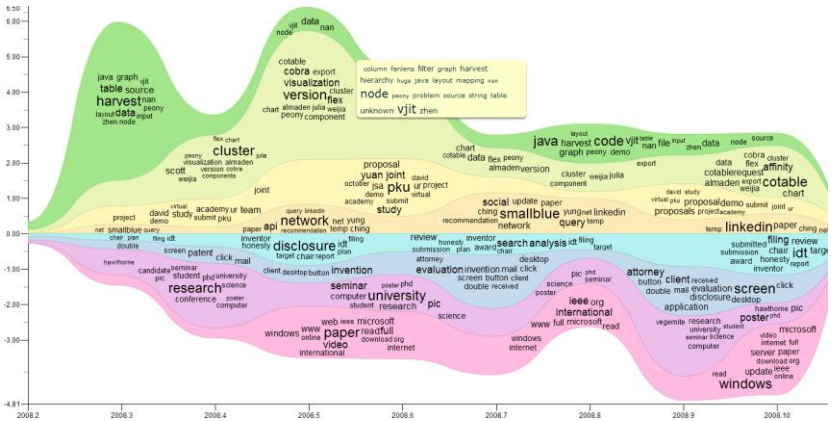


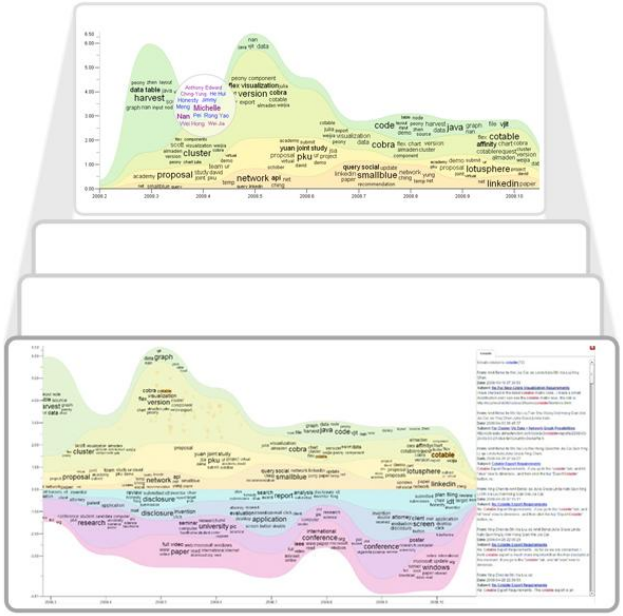


[Liu et al. CIKM 09]

Our Approach

- **Create visual metaphors to explain text summarization results**
- **Use interaction to let people**
  - Examine data from multiple aspects
- **Support both top-down and bottom-up analysis**
  - Forest → tree → leaf
  - Leaf → tree → forest
- **Design principle**
  - Augment common visual metaphors to convey complex summarization results



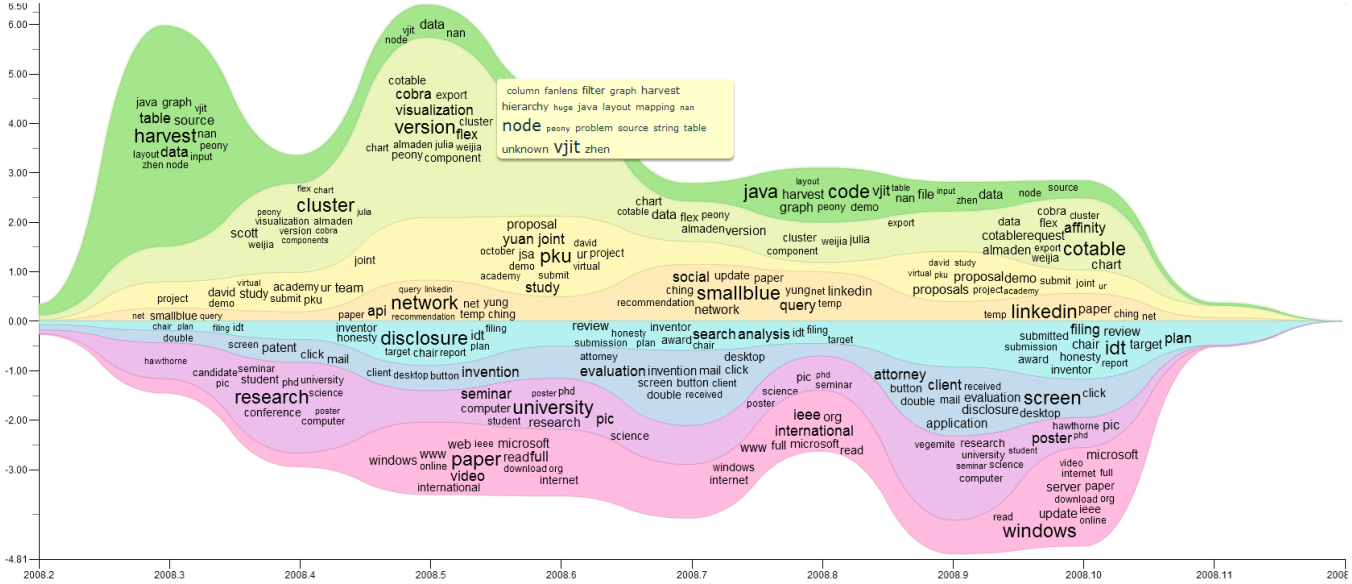


Visual Text Summary Metaphor

Data to be visualized:

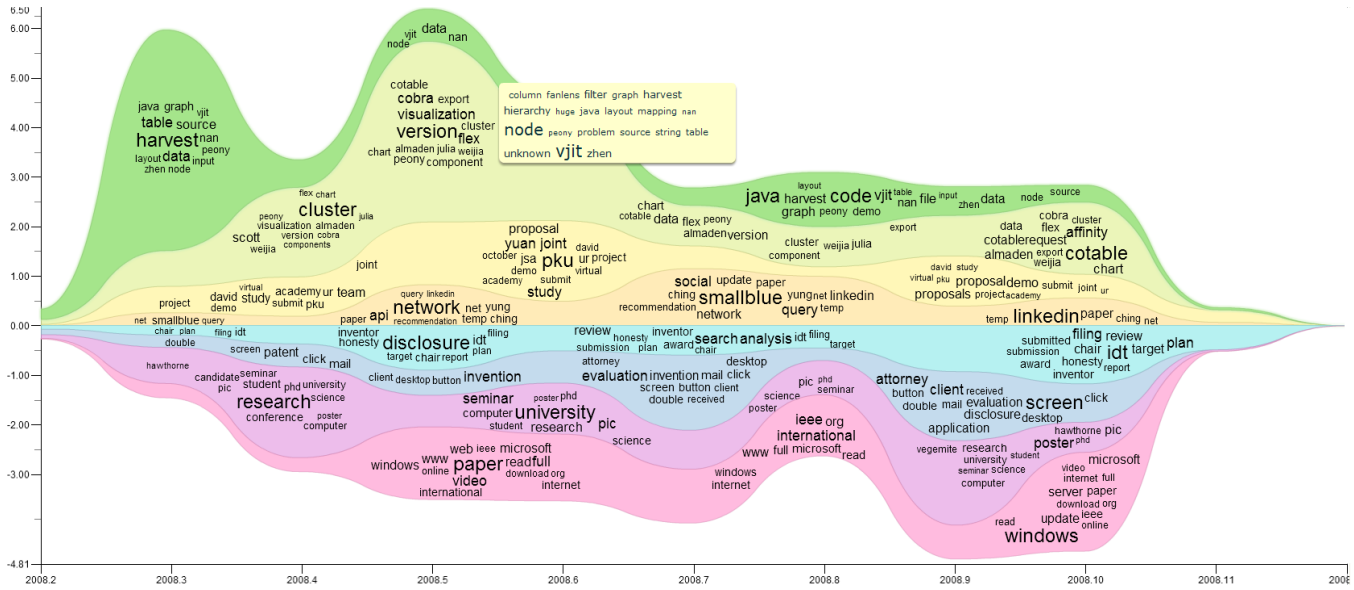
1. Topics: $\{T_1, \ldots T_i, \ldots T_N\}$ and their probabilities
2. For each T_i ,Topic keywords by time: … $\{\ldots, w^i_k, \ldots, \}_t$, … and their probabilities over time
3. For each T_i , Topic strength: $\{\ldots, S^i(t), \ldots\}$ over time

Visual encoding: Augmented stacked graph



Enhanced Stacked Graph: Key Steps

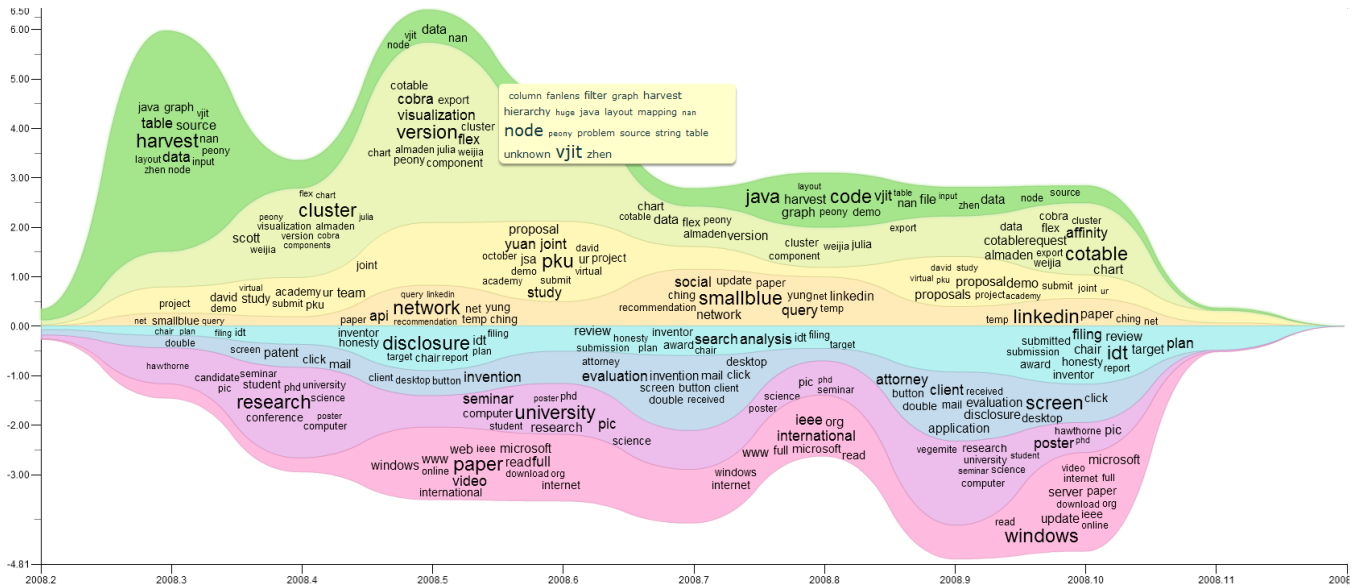
1. **Computing geometry of layers**
2. **Layer coloring**
3. **Layer ordering**
4. **Layer labeling**



Enhanced Stacked Graph: Key Steps

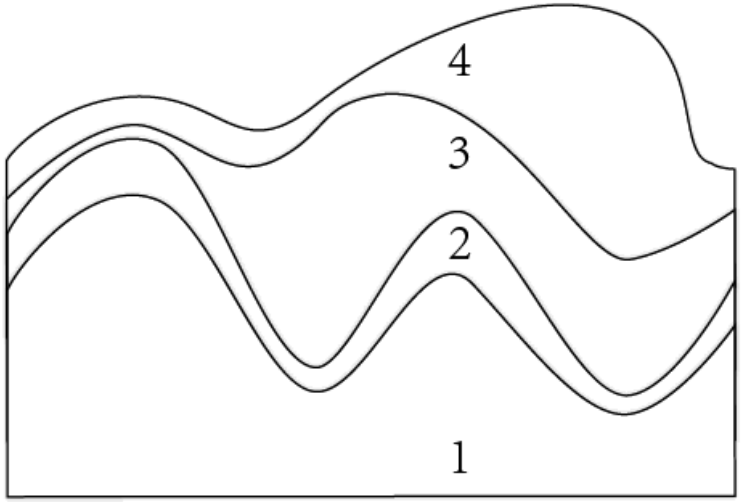
1. Computing geometry of layers
2. Layer coloring
3. **Layer ordering**
4. **Layer labeling**

} Byron_Infovis08

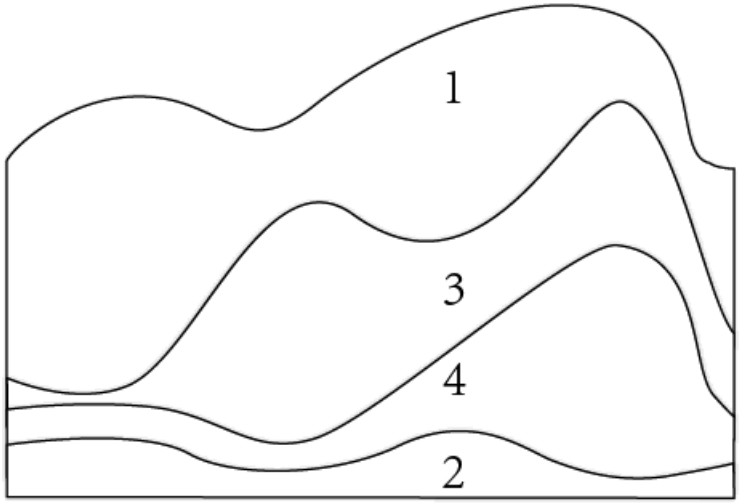


Enhanced Stacked Graph: Layer Ordering

- **Goals**
  - Minimize distortion
  - Maximize usable space
  - Ensure semantic coherence



unordered



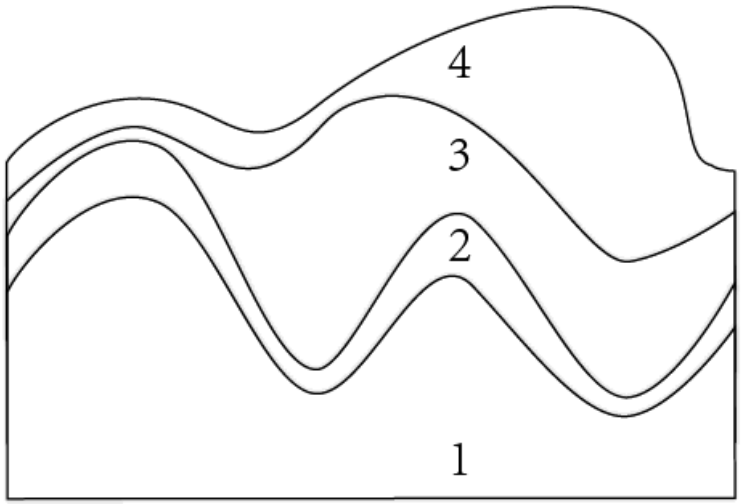
ordered

Enhanced Stacked Graph: Layer Ordering (cont'd)

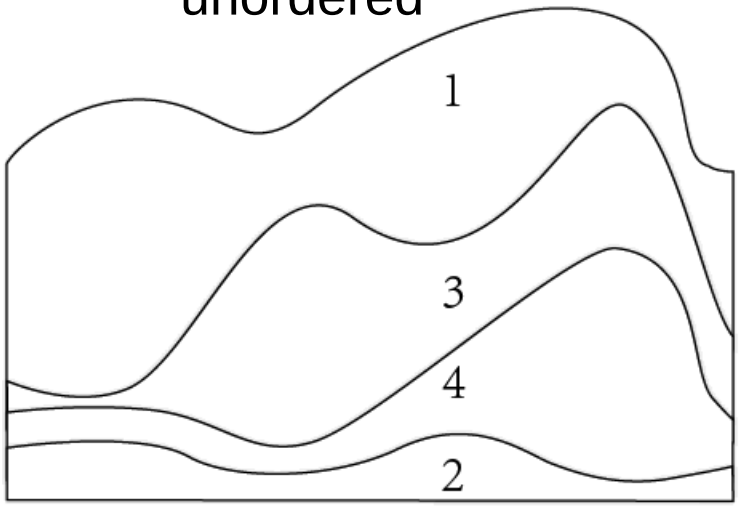
- **Our approach**
  - Optimization-based approach to
    - Minimize volatility (curvature) of topics
    - Maximize geometric complementariness of adjacent topics
    - Ensure semantic proximity of topics

for topics T_i and T_j

$$F_\sigma(p'_{ij}(t));\ p'_{ij}(t) = p_i(t) + p_j(t)$$

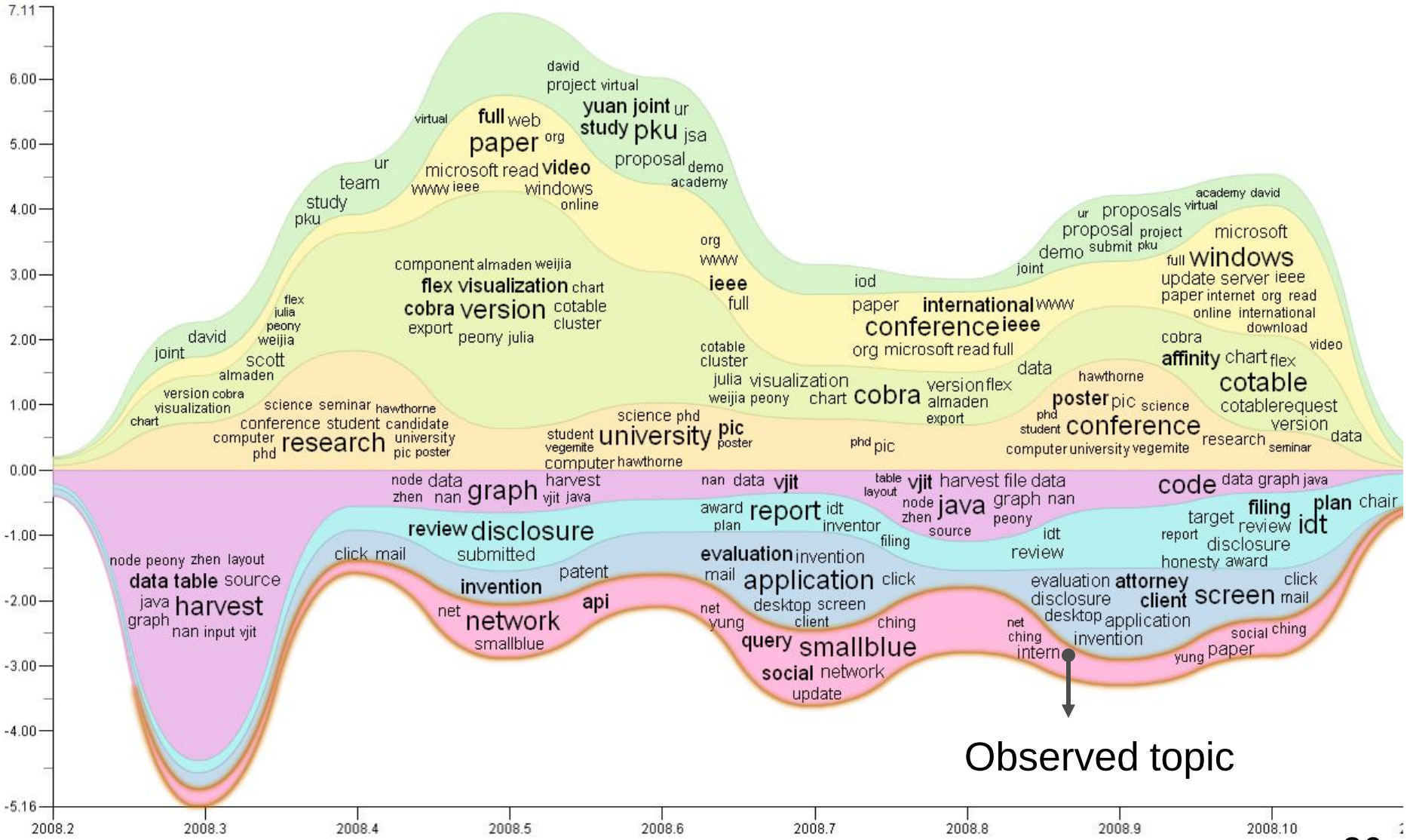


unordered

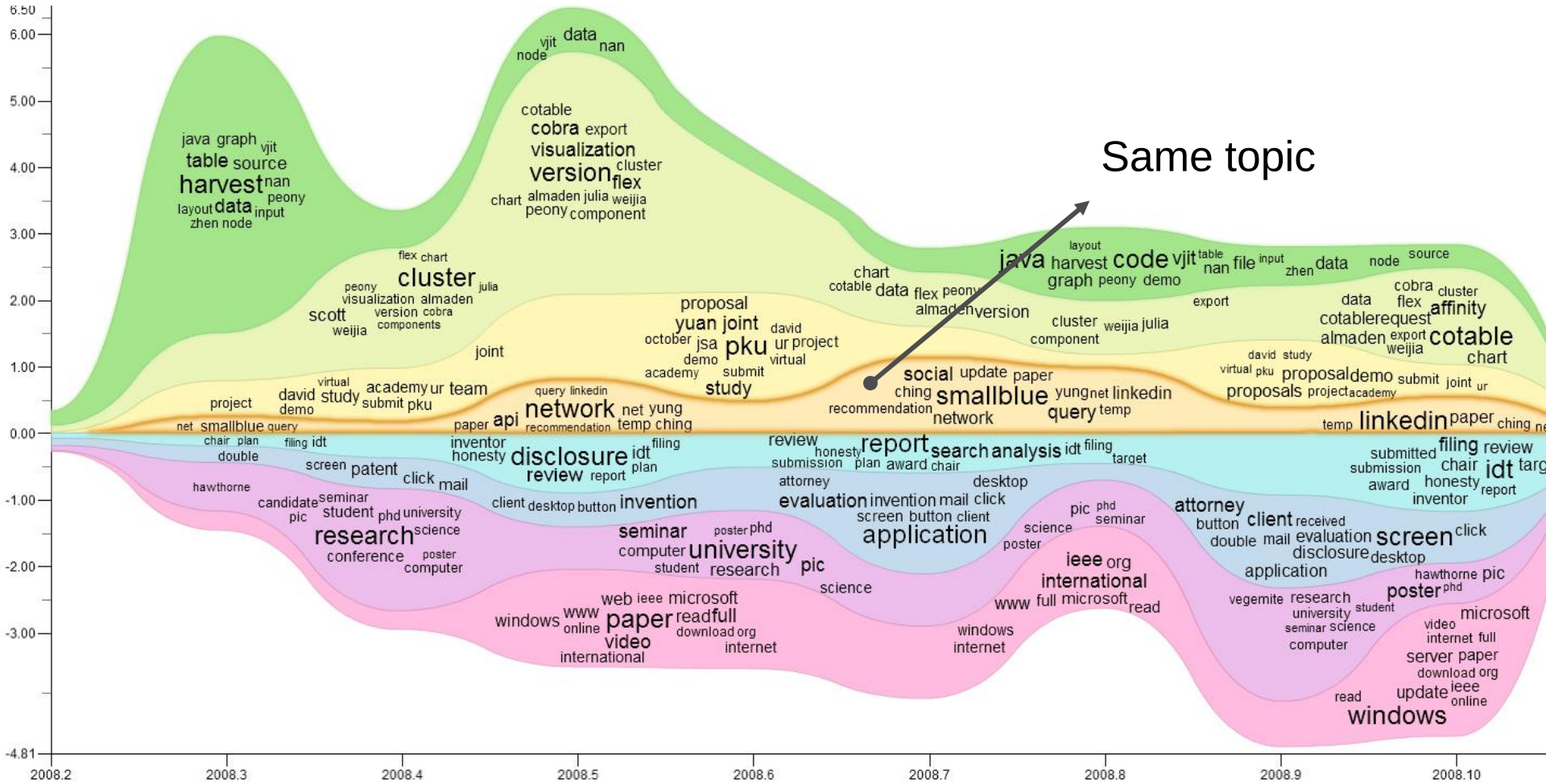


ordered

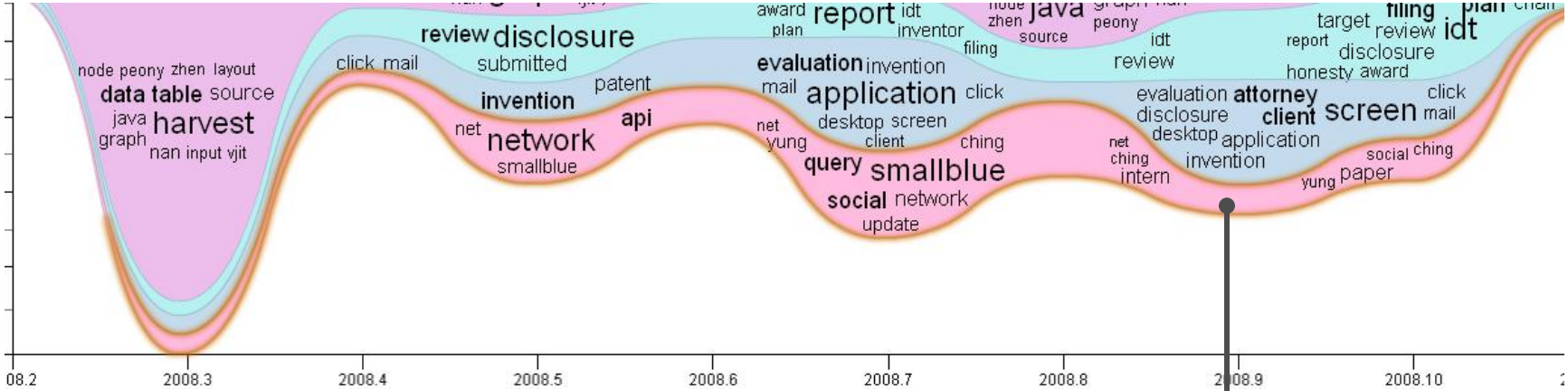
Enhanced Stacked Graph: Layer Ordering (cont'd)



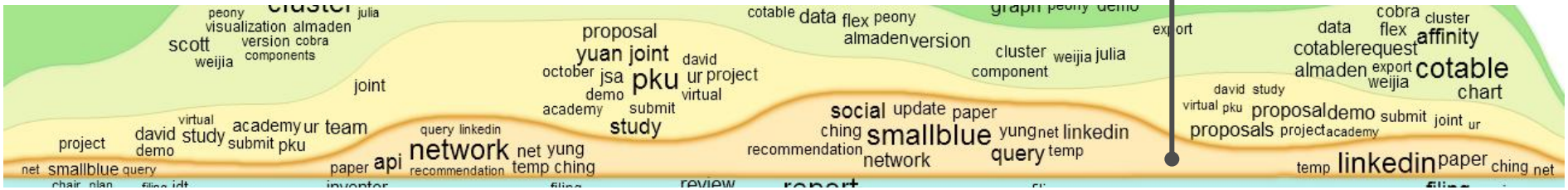
Enhanced Stacked Graph: Layer Ordering (cont'd)



Enhanced Stacked Graph: Layer Ordering (cont'd)



Same topic



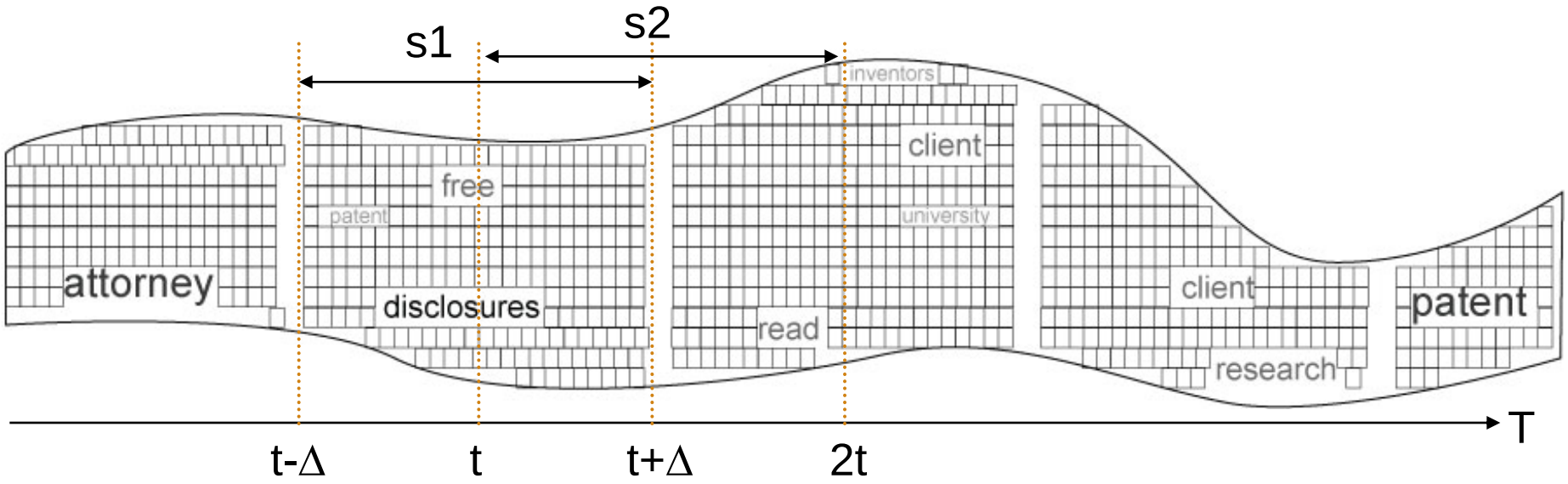
Alternative solution:
Interactive reordering

Enhanced Stacked Graph: Layer Labeling

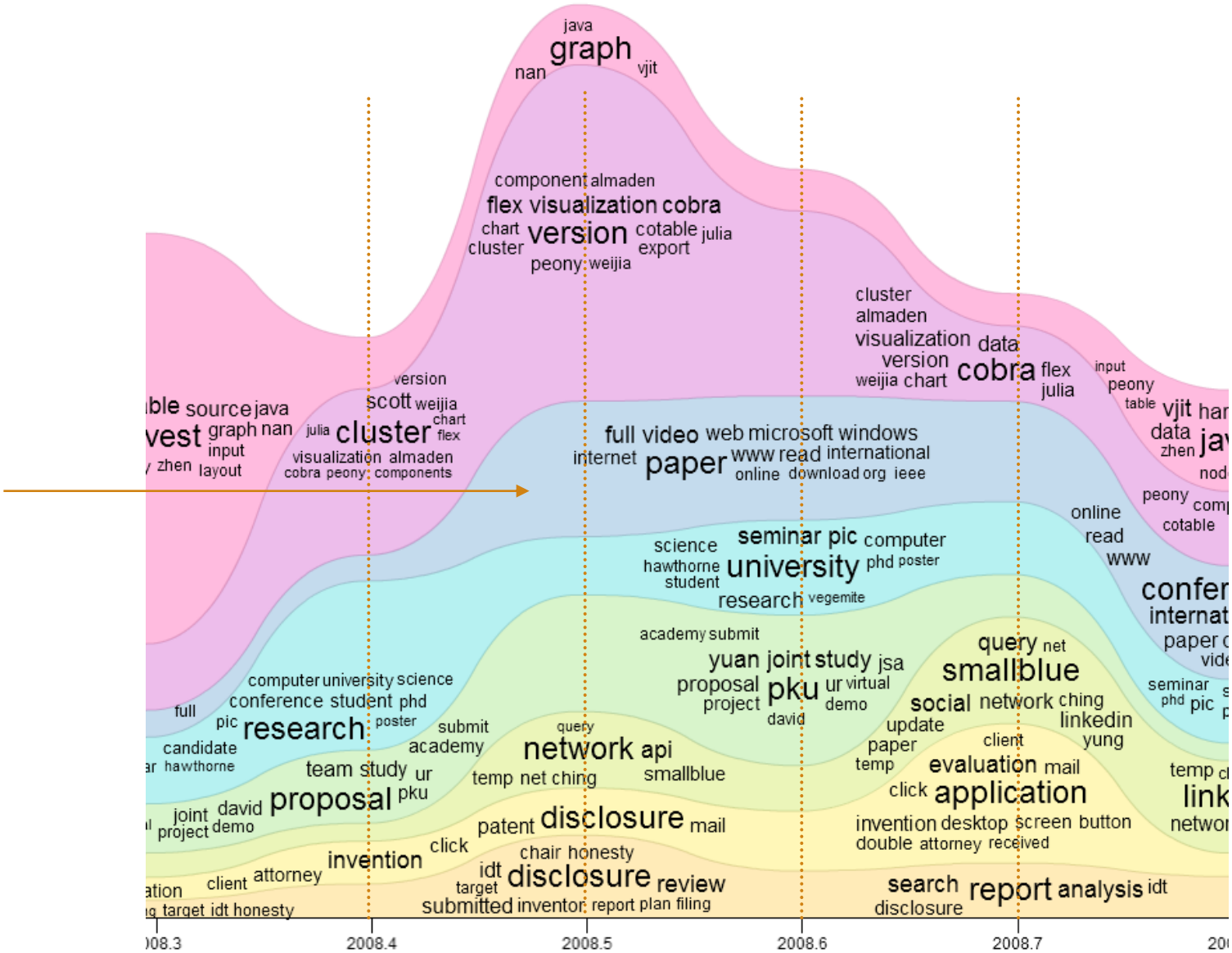
- **Goals**
  - Temporal proximity
  - Informativeness

Enhanced Stacked Graph: Layer Labeling (cont'd)

- **Our approach** [Liu et al. CIKM09]
  - Constraint-based space allocation
  - Particle-based layout [Luboschik et al. 08] + wordle

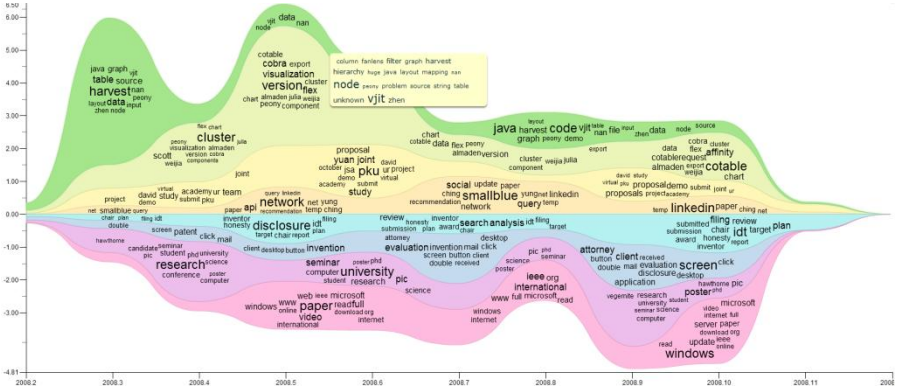


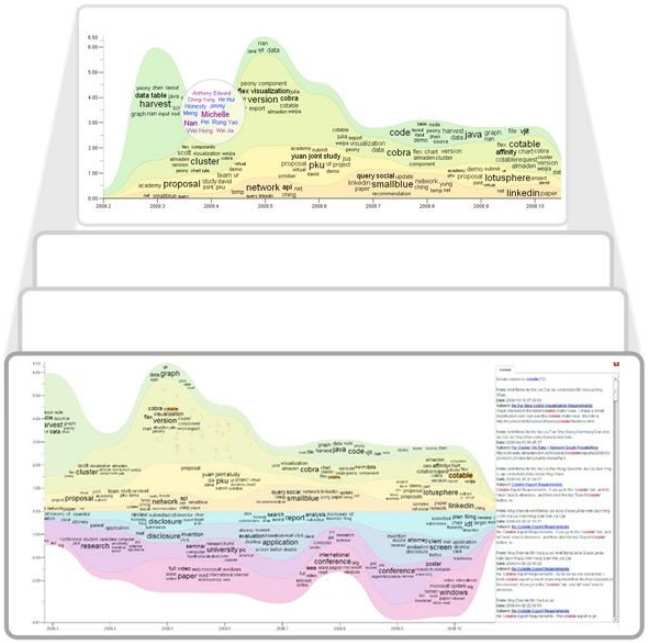
Enhanced Stacked Graph: Layer Labeling (cont'd)



Top-down and Bottom-up Analysis

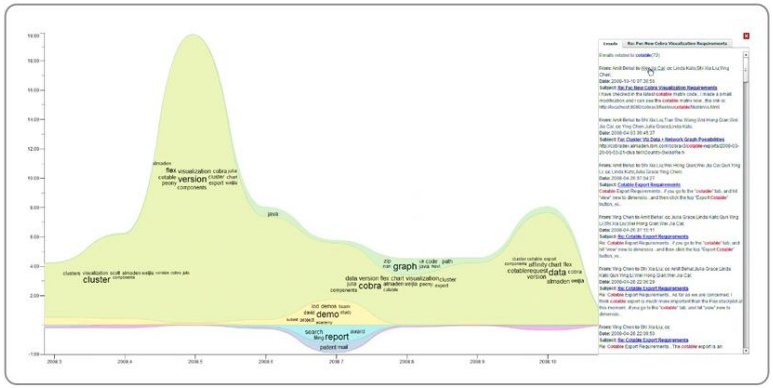
Visual Summary





Top-down navigation

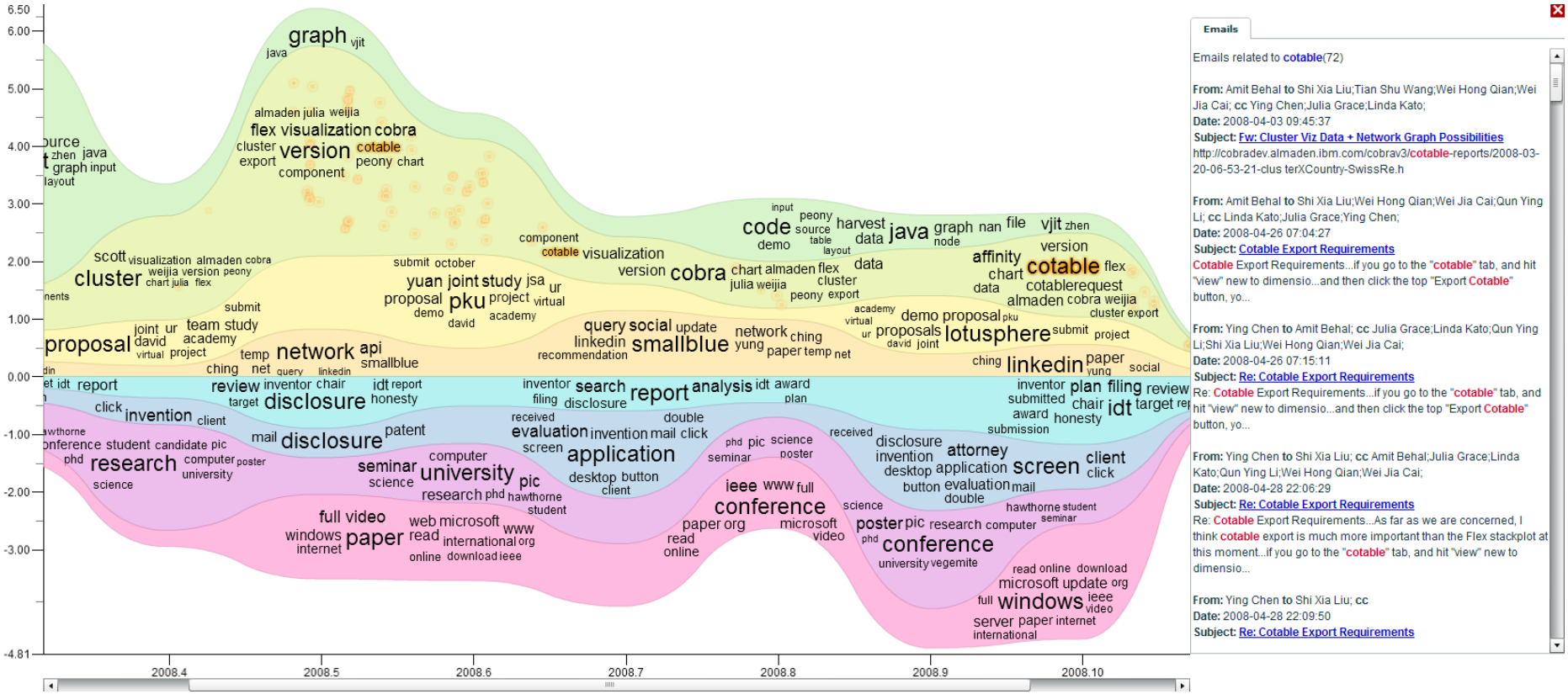
Bottom-up navigation



Summary of selected emails

Iterative analysis

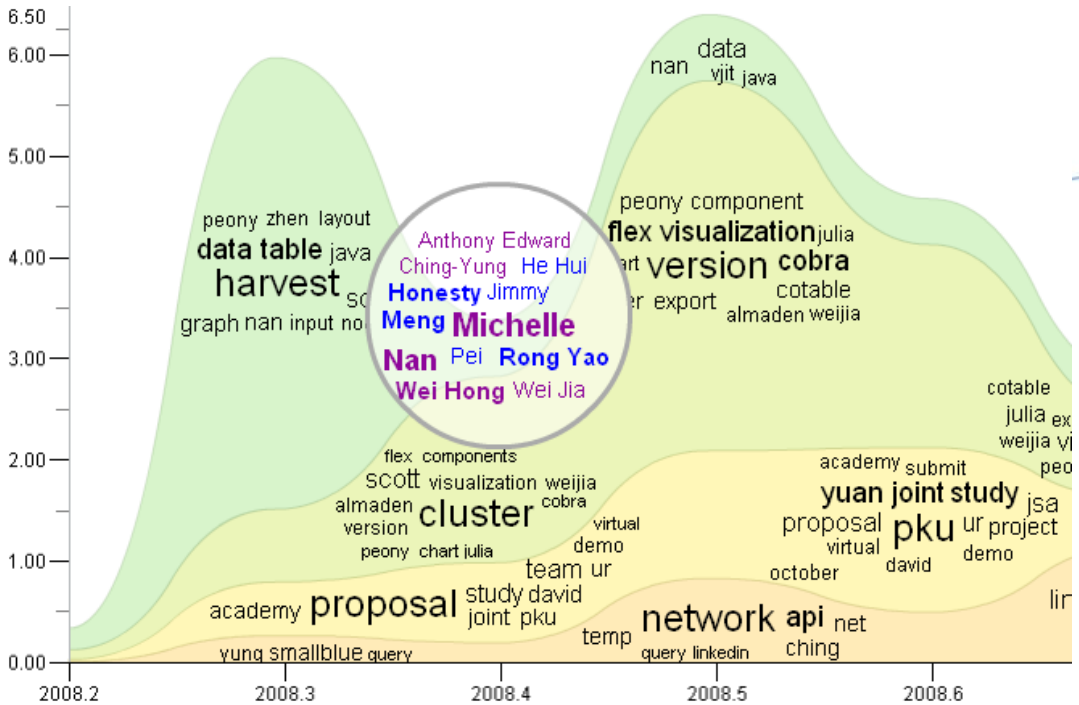
Interacting with Visual Summary



Selected keyword "cotable" and relevant email snippets

Interacting with Visual Summary

"people" involved and their relationships



TIARA Preliminary Evaluation

- **Goal**
  - What kind of tasks can TIARA help users accomplish?
  - What factors impact the use of TIARA?
- **Data set**
  - Emails (~10,000)
- **12 Participants:** 6 familiar with the email owner's work
- **Tasks**
  - Task 1: Analyze emails between two people
  - Task 2: Answer questions about specific projects
  - Task 3: Answer questions characterizing the email owner's work

TIARA Preliminary Evaluation

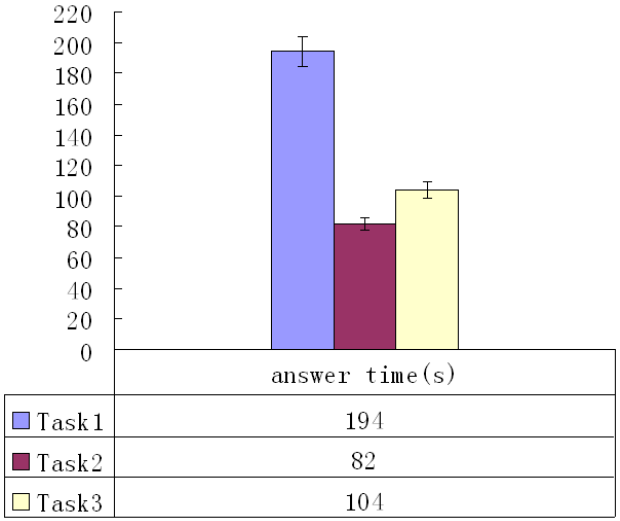
- **Method**
  - Objective measures
    - answer completion rate, answer error rate, and answer time
  - Subjective measures
    - usefulness, usability, and system satisfaction
  - Compared TIARA and Themail for Task 1

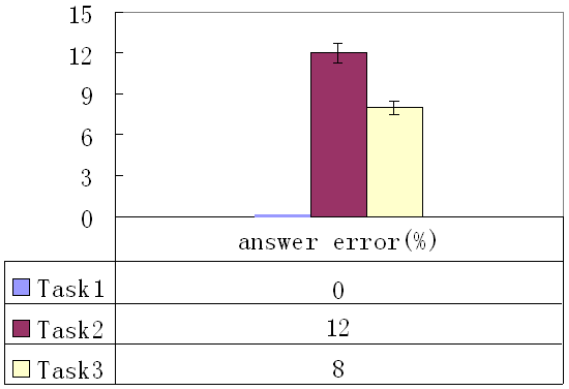
Sample Evaluation Questions

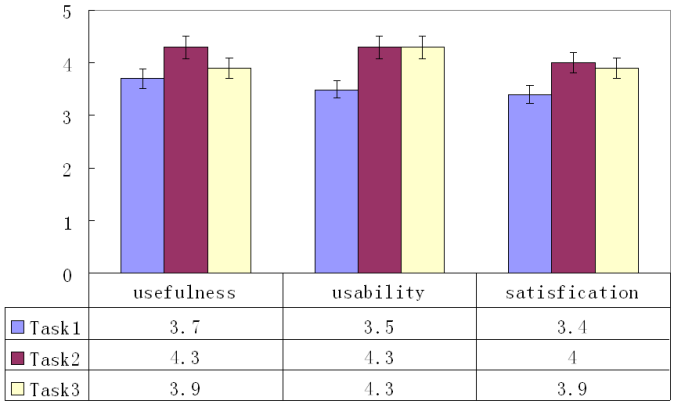
Task1	examining emails between two people
What are the three main concepts mentioned during their June emails?	
Which month of 2008 is most active in their email exchanges?	
Task2	examining emails about a project named "Cobra"
Who were involved in Cobra?	
When was this project most active?	
What was discussed during the active period?	
Task3	examining emails in general
What was the most active topic in May?	
Who were the people involved in this topic?	

TIARA Evaluation Results

- **Type of tasks impacted the effectiveness of TIARA**
  - TIARA helped users complete complex analytic tasks faster
  - Visual summary too coarse for extracting details (e.g., names)
- **User's knowledge impacted the effectiveness of TIARA**
  - TIARA helped knowledgeable users more



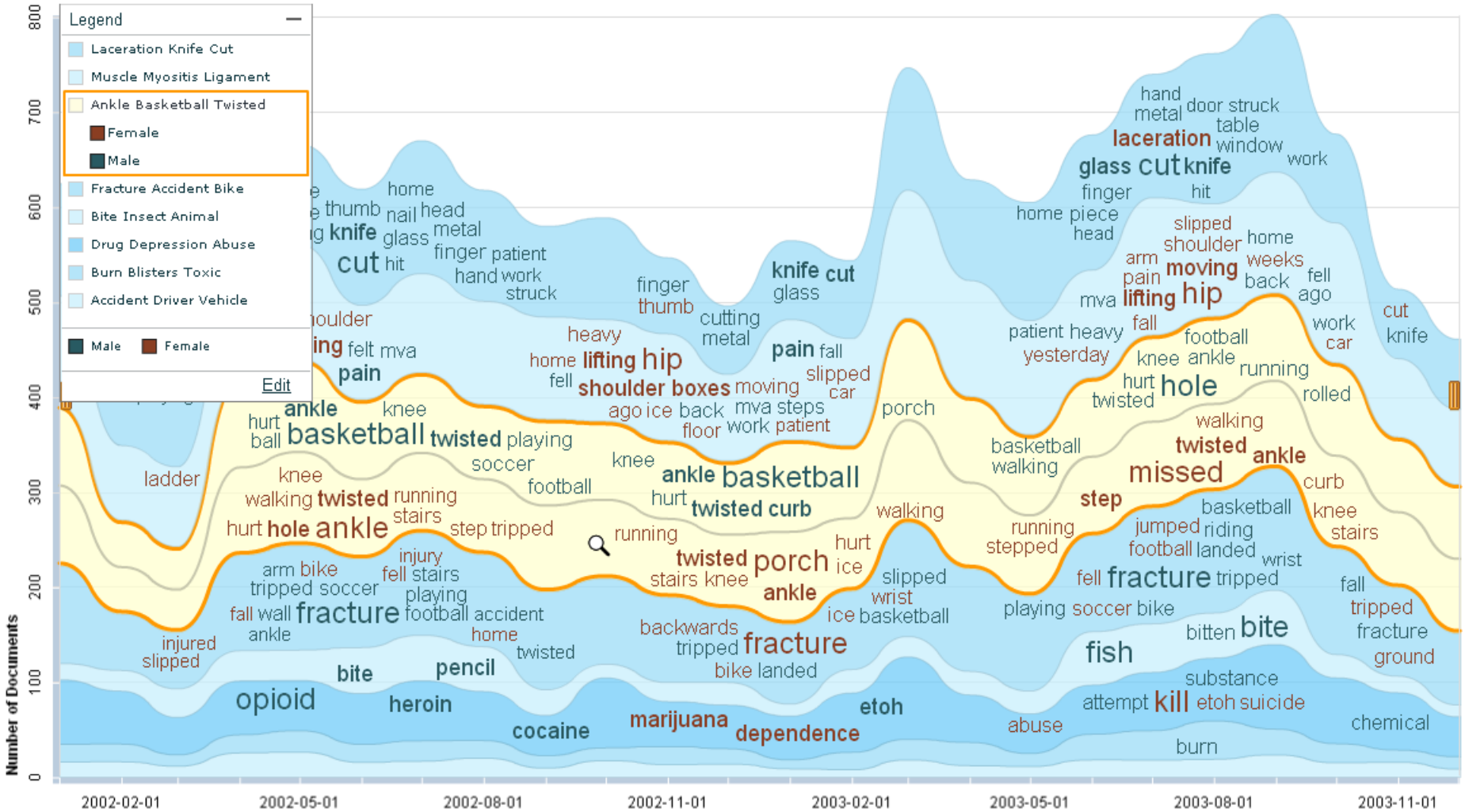




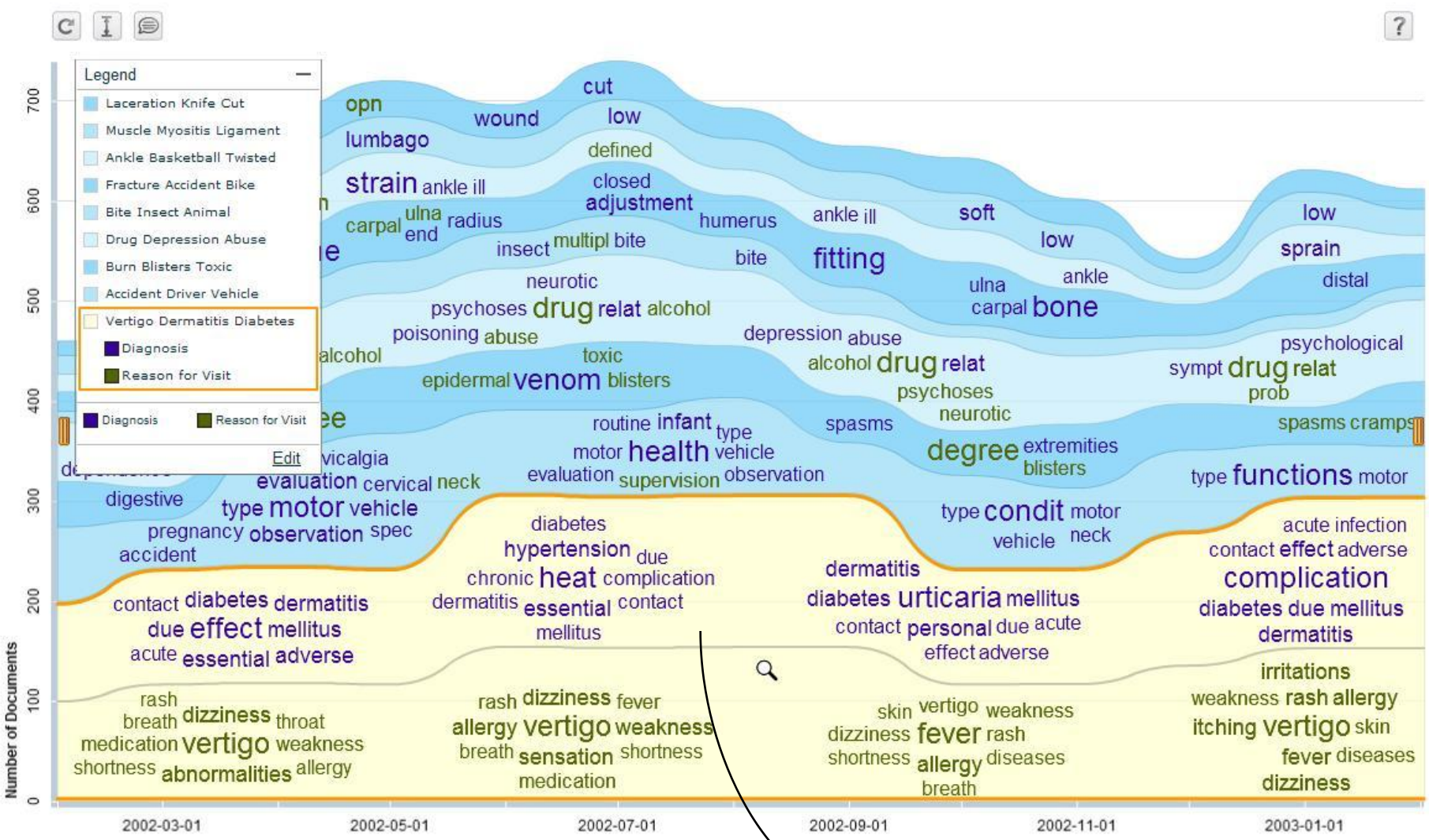
Application Example: Healthcare

- **Visualize unstructured data (text) to facilitate analysis**
  - Cause of injury
  - Reason for visit
  - Diagnosis
- **Handle multiple fields of text data and show their correlation**
- **Leverage structured data to help better illustrate text information**
  - Gender + Cause of injury

Correlation between Structured and Text Fields



Correlation between Two Text Fields



Correlation between two fields, *diagnosis* and *reason for visit*

Outline

- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text
- **Text Visualization Perspectives**

TextFlow: Towards Better Understanding of Evolving Topics in Text

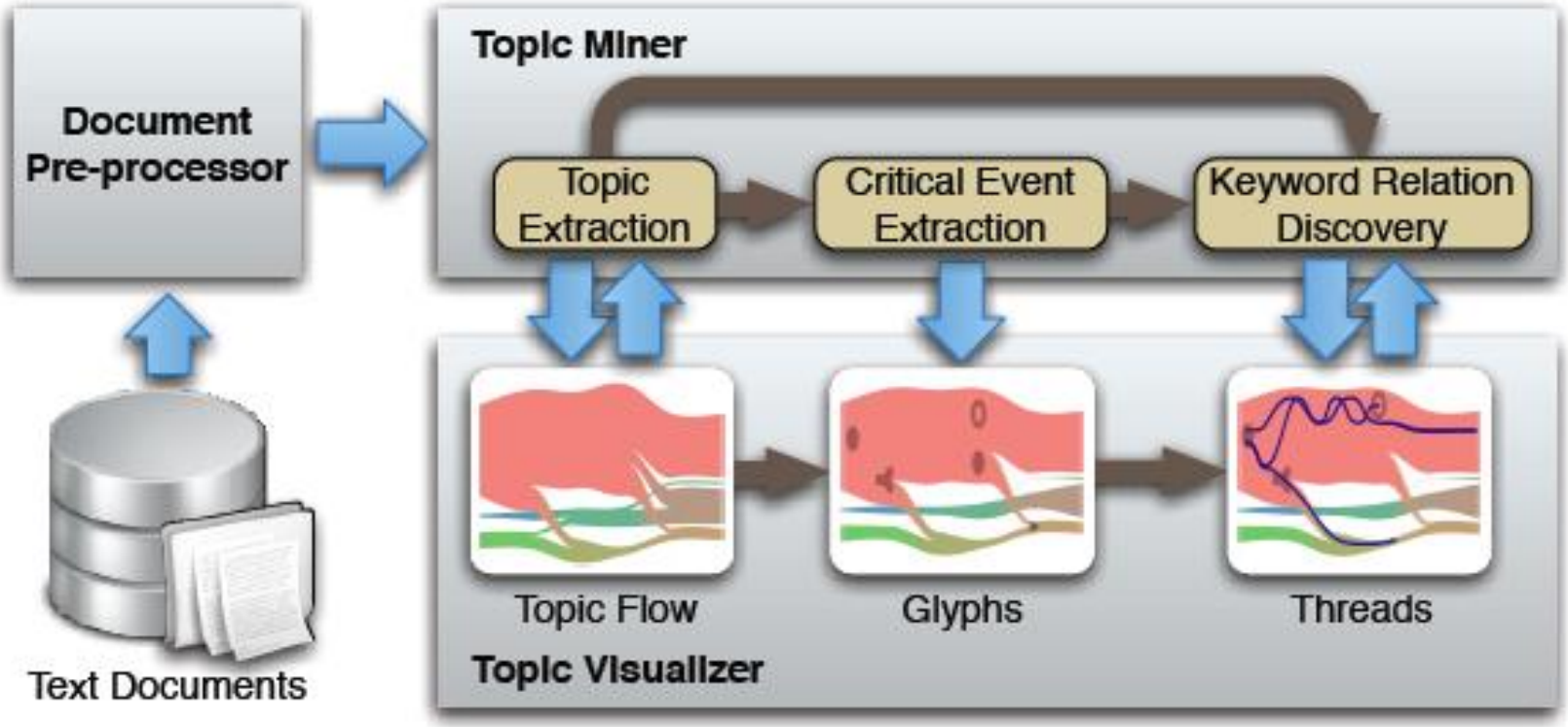
Cui et al. Infovis 11

- **Problems**
  - Understanding topic evolution in large text collections is important
    - Keep abreast of hot, new, and intertwining topics
    - Gain insight into the latent topics
- **Challenges**
  - Model topic merging/splitting patterns
  - Visually convey the topic merging/splitting patterns in an intuitive way
  - Facilitate analytical reasoning
- **Solutions**
  - Leverage *Hierarchical Dirichlet Processes* to model topic merging/splitting
  - Augment the familiar visual metaphor, the river flow, to convey the complex analytic results
  - Interact with the topic from global structure to local salient features

Related Work

- **Little work has focused on studying topic merging and splitting patterns**
- **It has barely been touched by using visual analysis techniques to interactively analyze complex topic evolution from multiple perspectives**

TextFlow Overview



Topic Data and Relationship Extraction

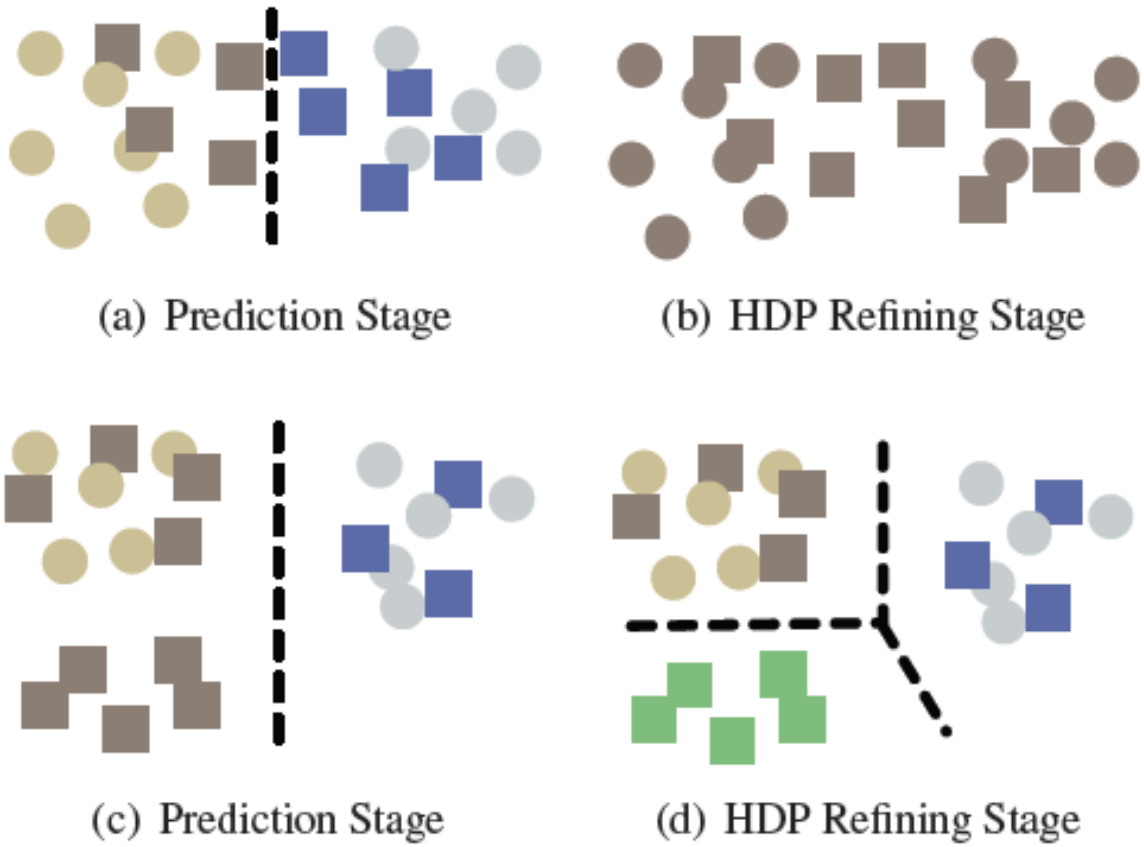


Fig. 3. An example of splitting/merging of clusters: circles representing samples at time $t-1$; rectangles encoding samples at time t.

Merging/Splitting Relationships

- **Merging input from time *t-1* to *t*:**

$$P_t^{in}(s \to r) \triangleq \frac{\sum_{\tau=t-T_{win}+1}^{t} \sum_{i,j} I(z_{ji}^{t,old} = s \,\&\, z_{ji}^{t,new} = r)}{\sum_{\tau=t-T_{win}+1}^{t} \sum_{i=1}^{n^t} I(z_{ji}^{t,new} = r)}$$

- **Splitting output from time *t-1* to *t*:**

$$P_{t-1}^{out}(s \to r) \triangleq \frac{\sum_{\tau=t-T_{win}+1}^{t} \sum_{j,i} I(z_{ji}^{\tau,old} = s \,\&\, z_{ji}^{\tau,new} = r)}{\sum_{\tau=t-T_{win}+1}^{t} \sum_{j,i} I(z_{ji}^{\tau,old} = s)}.$$

Critical Event Extraction

Domain-dependent activeness metric

- **Critical events**
  - Birth, death, merge, and split
- **Score of a merging event**

$$R(r,t) = |\mathcal{N}_r| \cdot H_t(r) = |\mathcal{N}_r| \cdot \kappa_B \sum_{s \in \mathcal{N}_r} -P_t^{in}(s \to r) \ln P_t^{in}(s \to r)$$

Neighborhood set

Entropy score

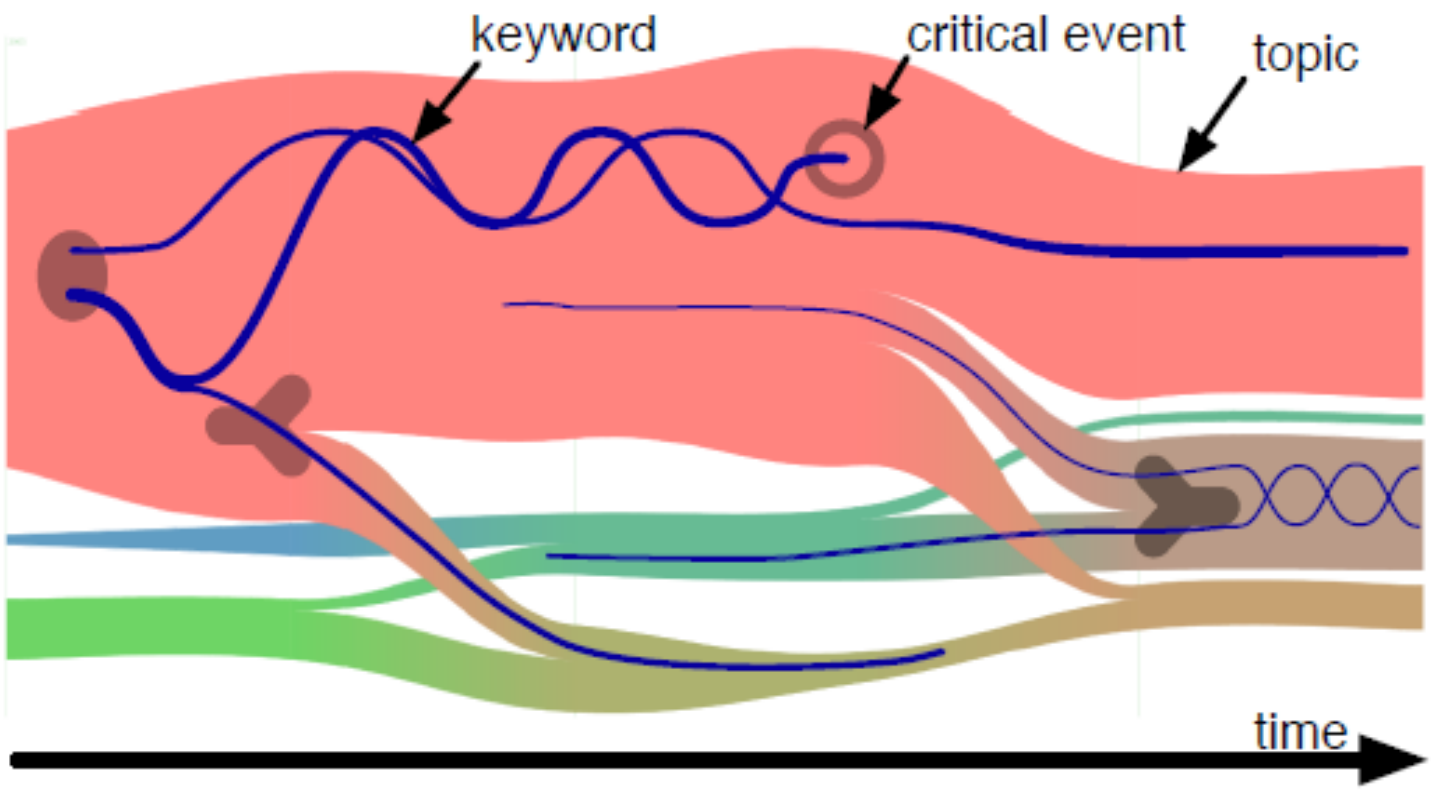
- **Score of a splitting event**

$$R(s,t) = |\mathcal{N}_s| \cdot H_t(r) = |\mathcal{N}_s| \cdot \kappa_B \sum_{r \in \mathcal{N}_s} -P_{t-1}^{out}(s \to r) \ln P_{t-1}^{out}(s \to r)$$

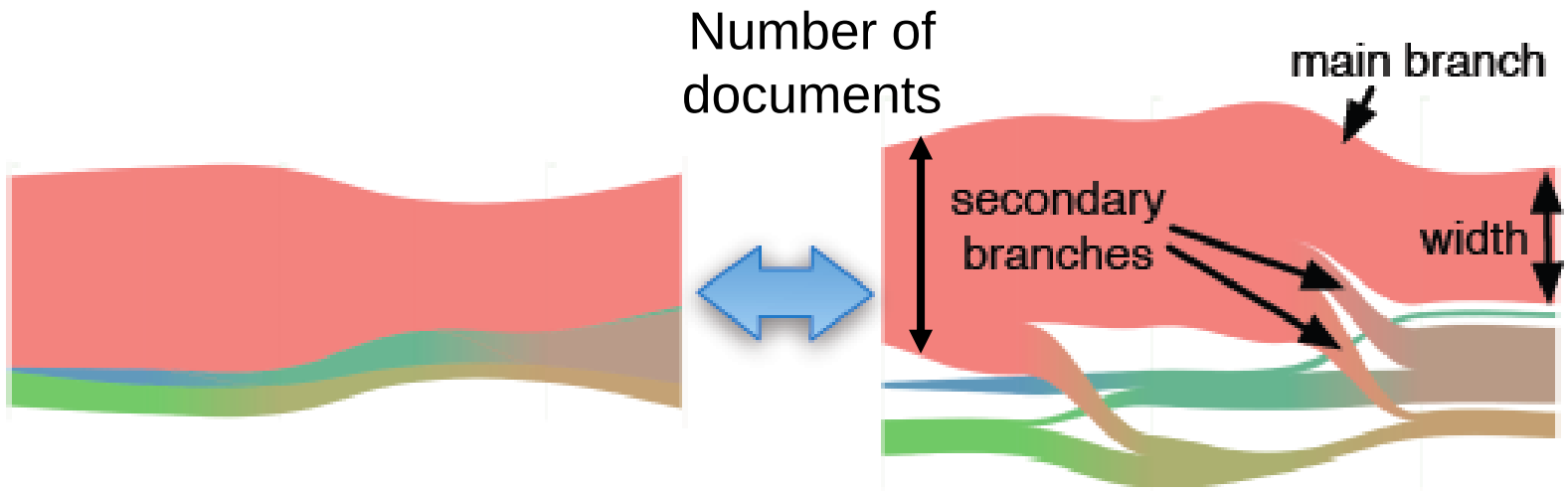
Keyword Correlation Discovery

- **Extract “noun phrases,” “verb phrases”, and “named entities” in each document, and count co-occurrences among them**

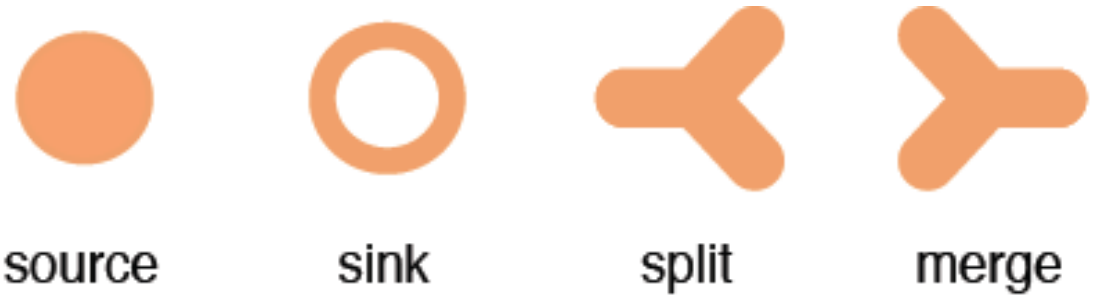
Visualization Design



Topic Evolution as Flow

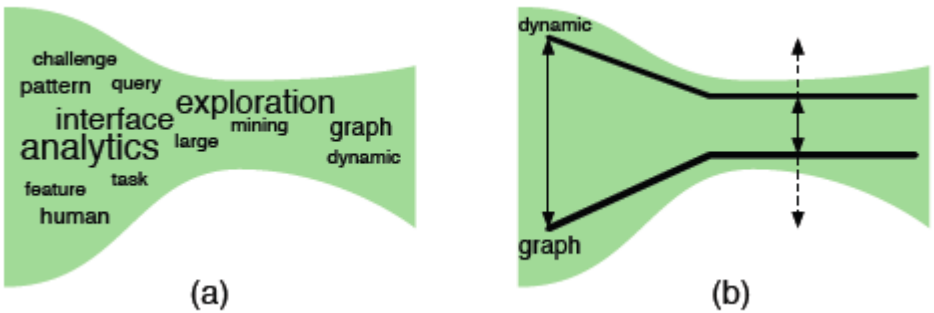


Critical Event as Glyph

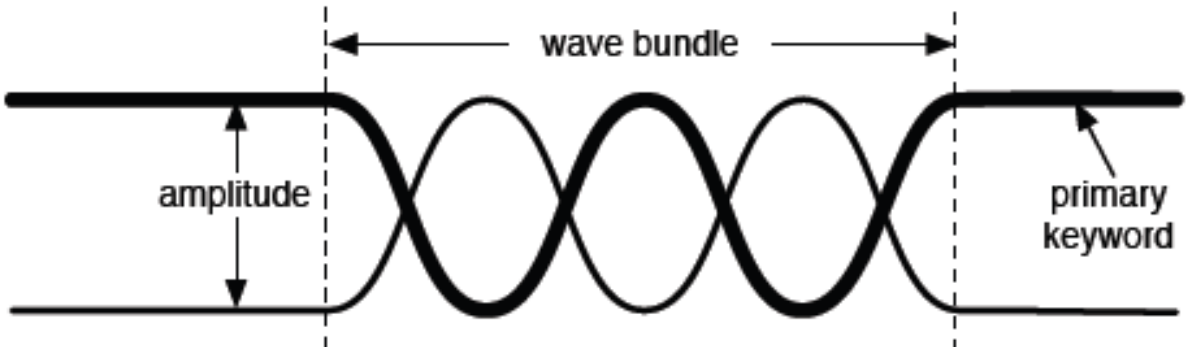


Keyword Correlation as Thread

Alternatives



Keyword thread



Layout Algorithm

- **A three-level Directed Acyclic Graph (DAG)**

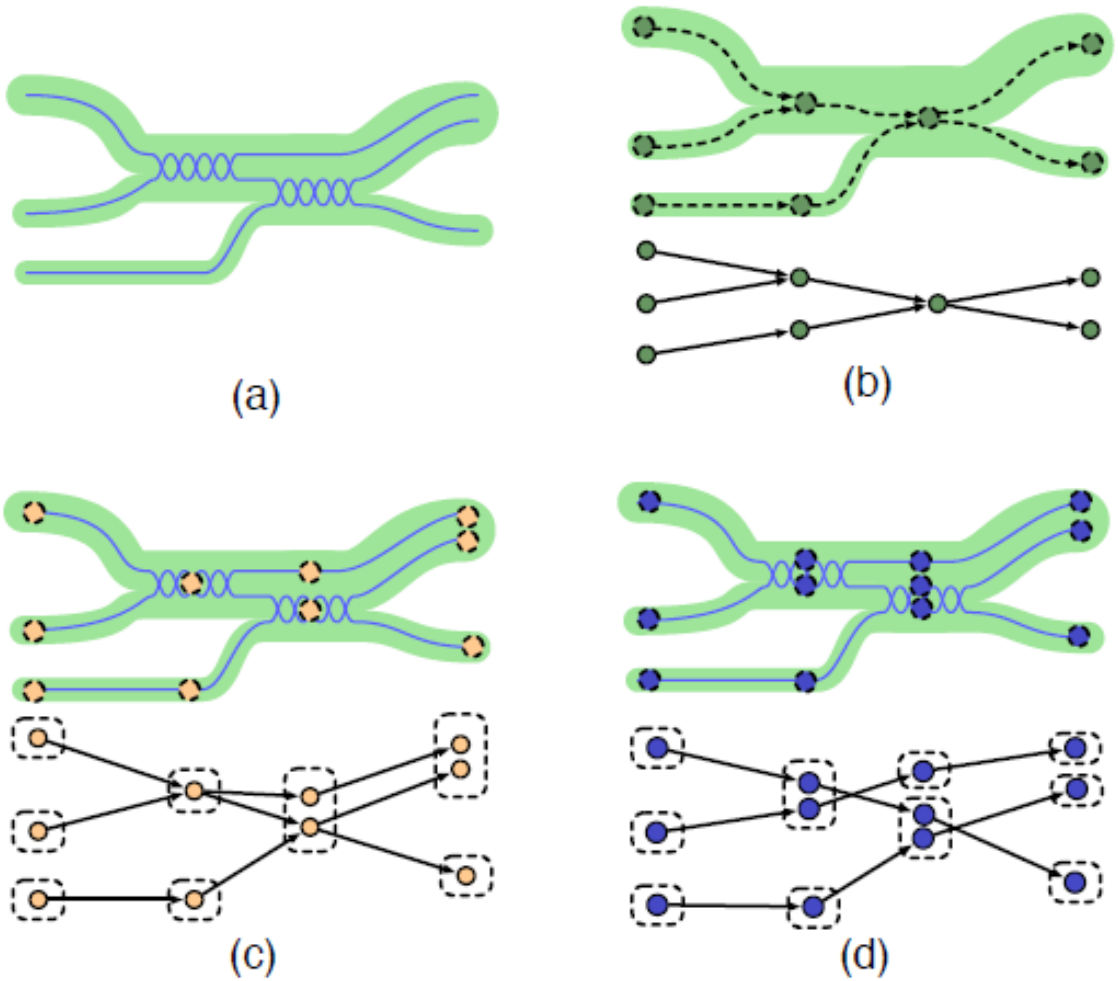


Fig. 9. The three-level model for topic flow graph layout (dotted rectangles indicating boundary constraints): (a) the original structure; (b) first level: topic flows; (c) second level: bundles; (d) third level: threads.

Interactive Exploration

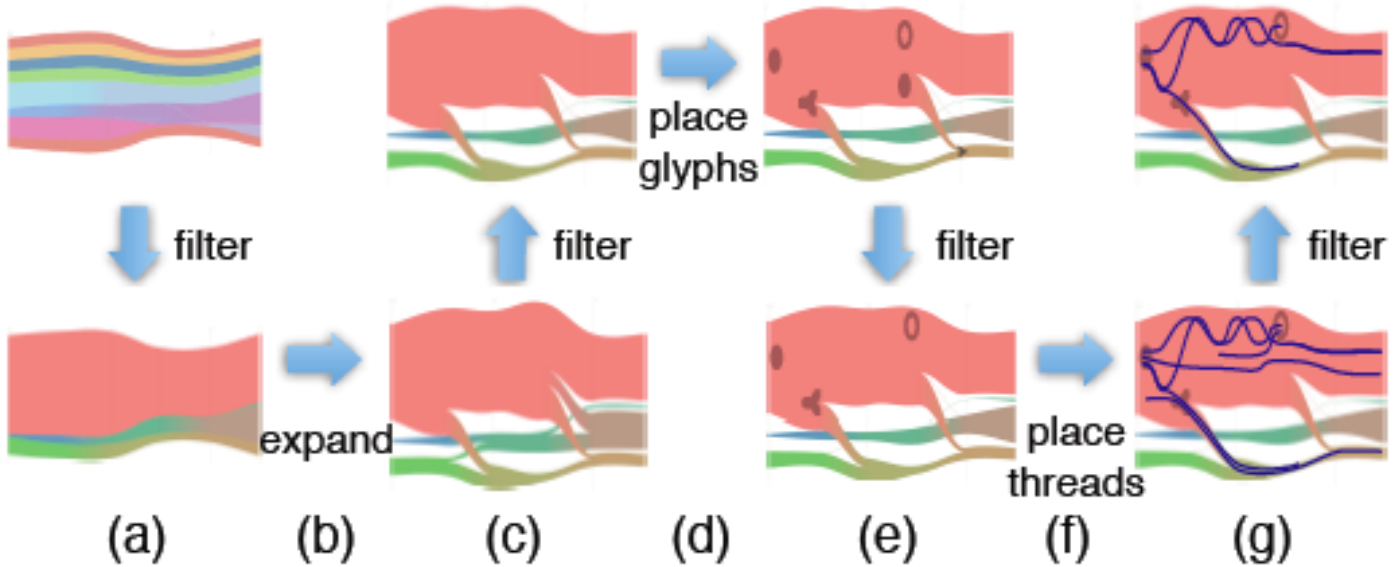
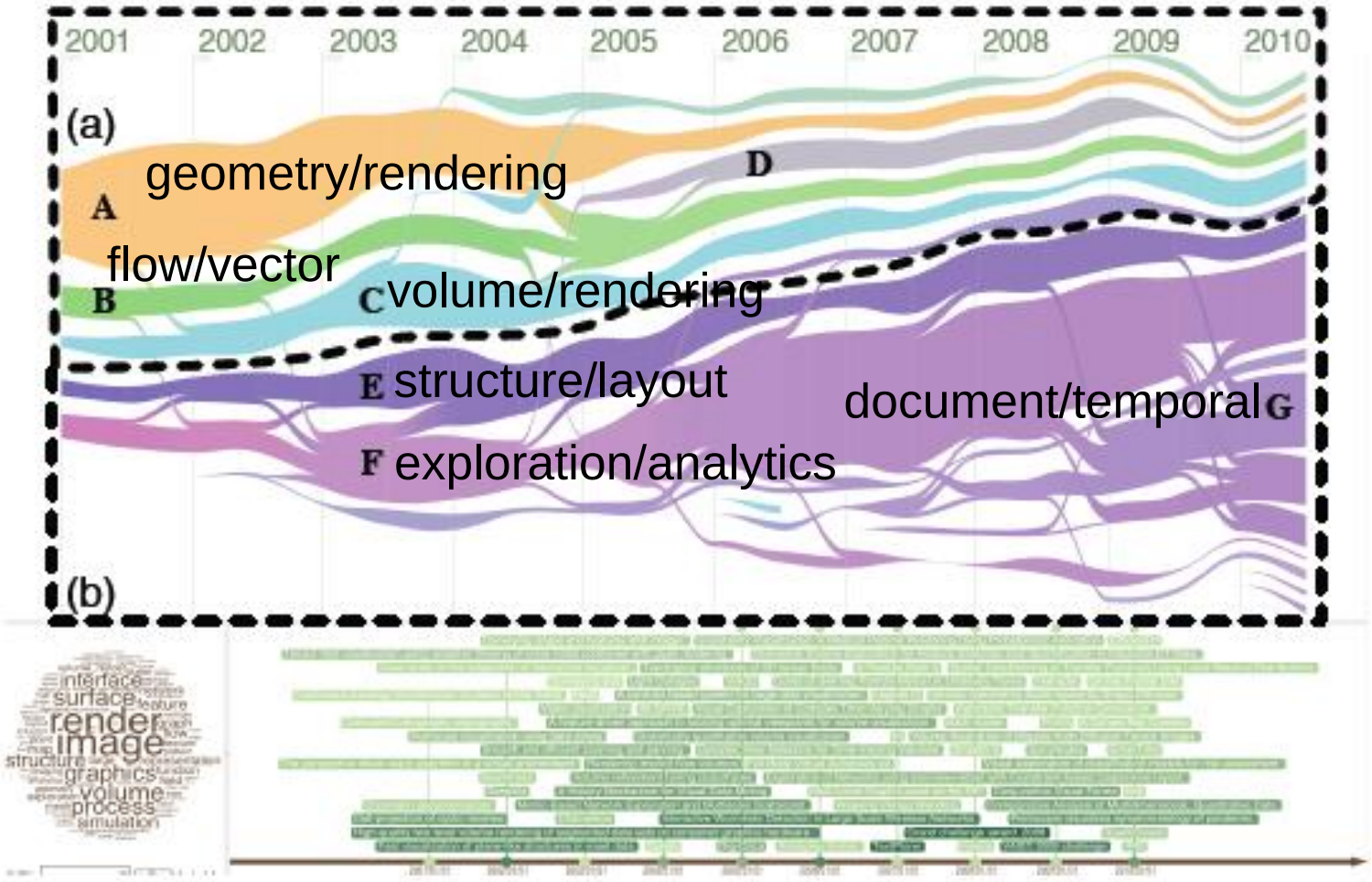


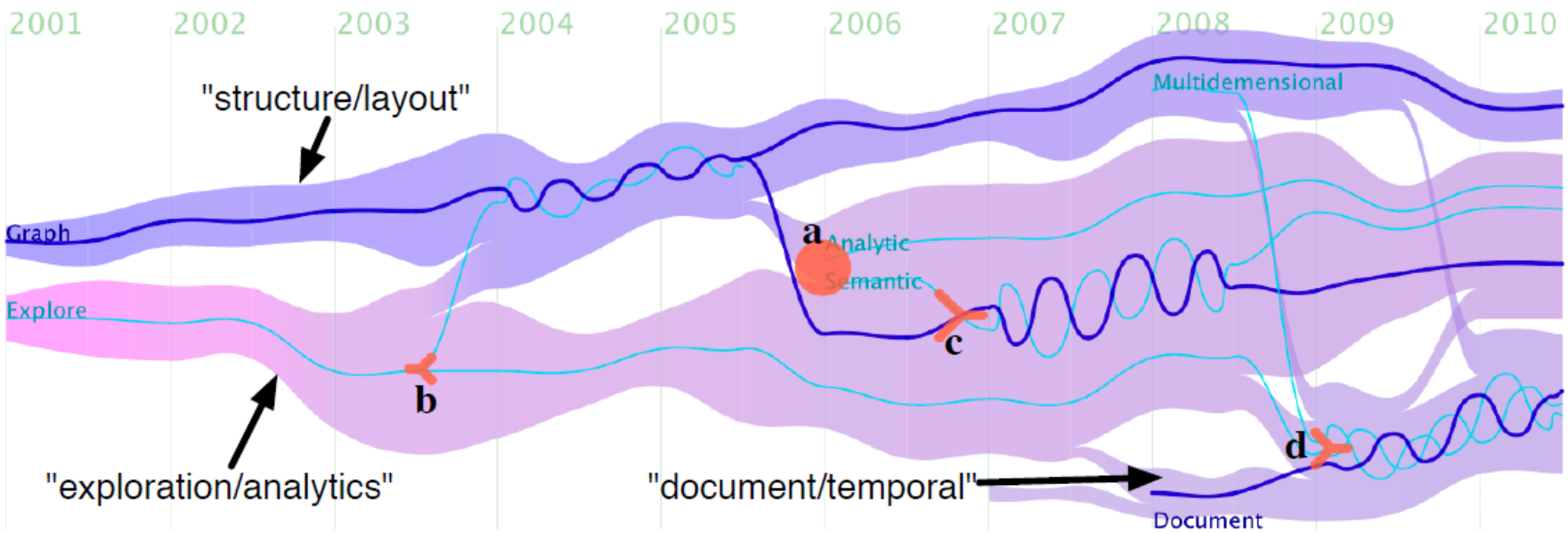
Fig. 11. The exploration pipeline: (a) select a subset of interest based on users' interest or system recommendation; (b) expand the topics to see their splitting and merging patterns; (c) remove trivial or irrelevant branches; (d) and (e) extract major critical events based on current flow patterns; (f) and (g) explore and adjust threads around the selected critical events based on users' interest or system recommendation.

Application Example – VisWeek Pulications

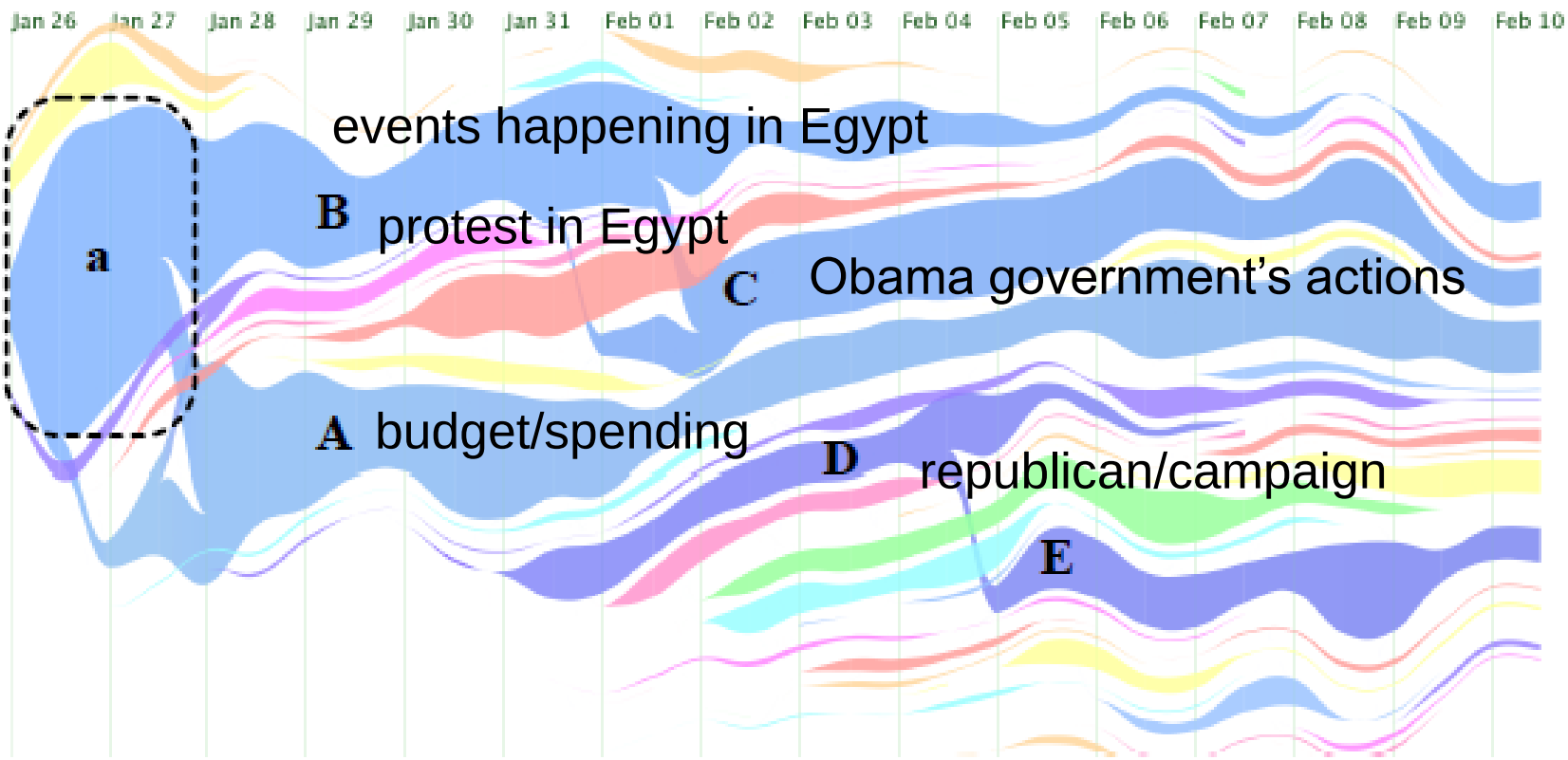


933 research papers

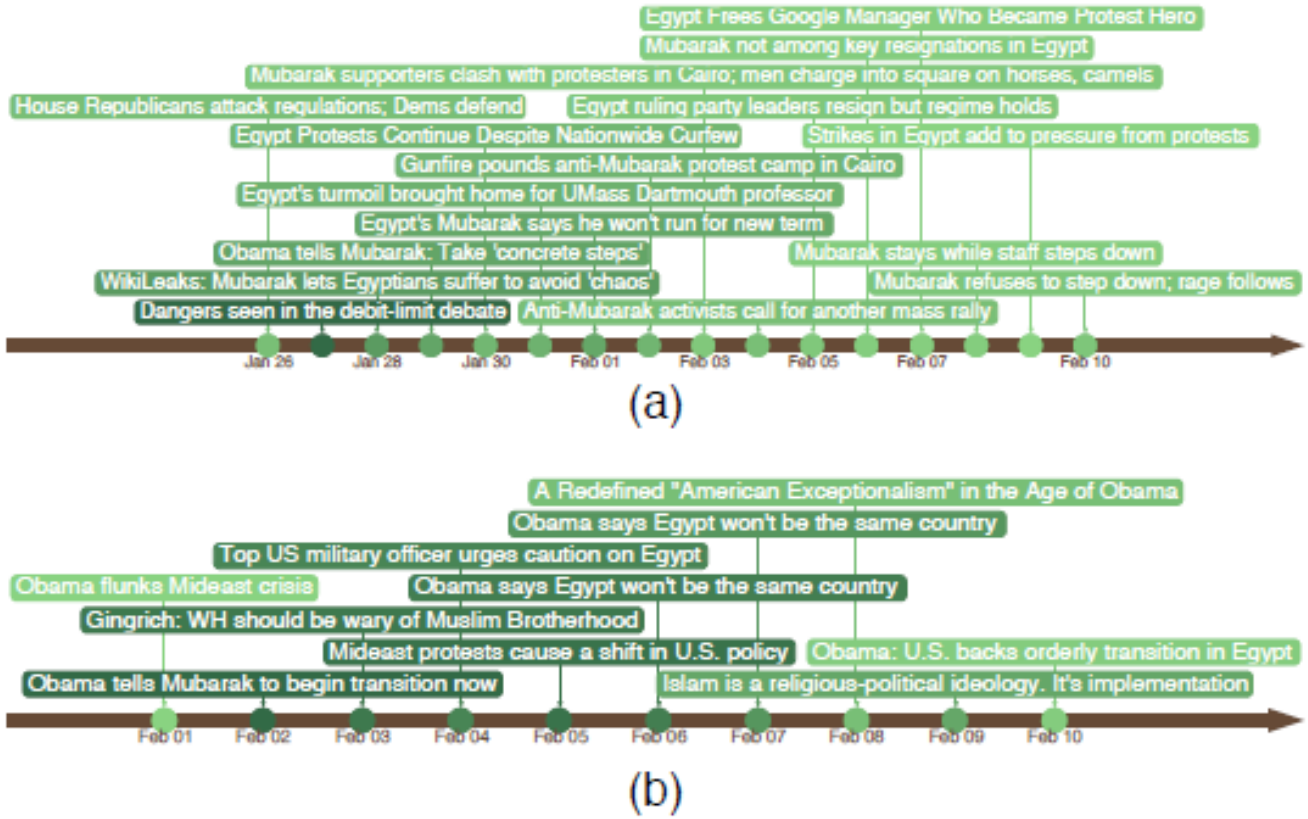
Application Example - VisWeek Pulications



Application Example – Bing News

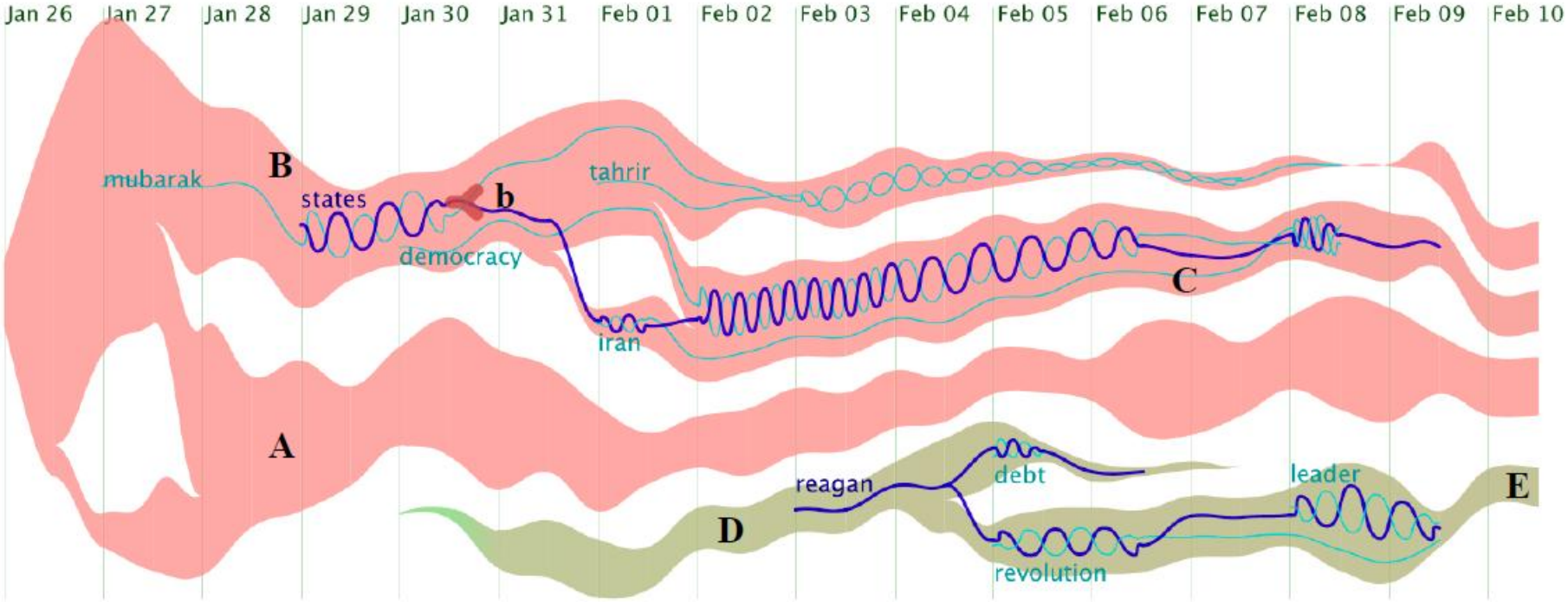


Application Example – Bing News

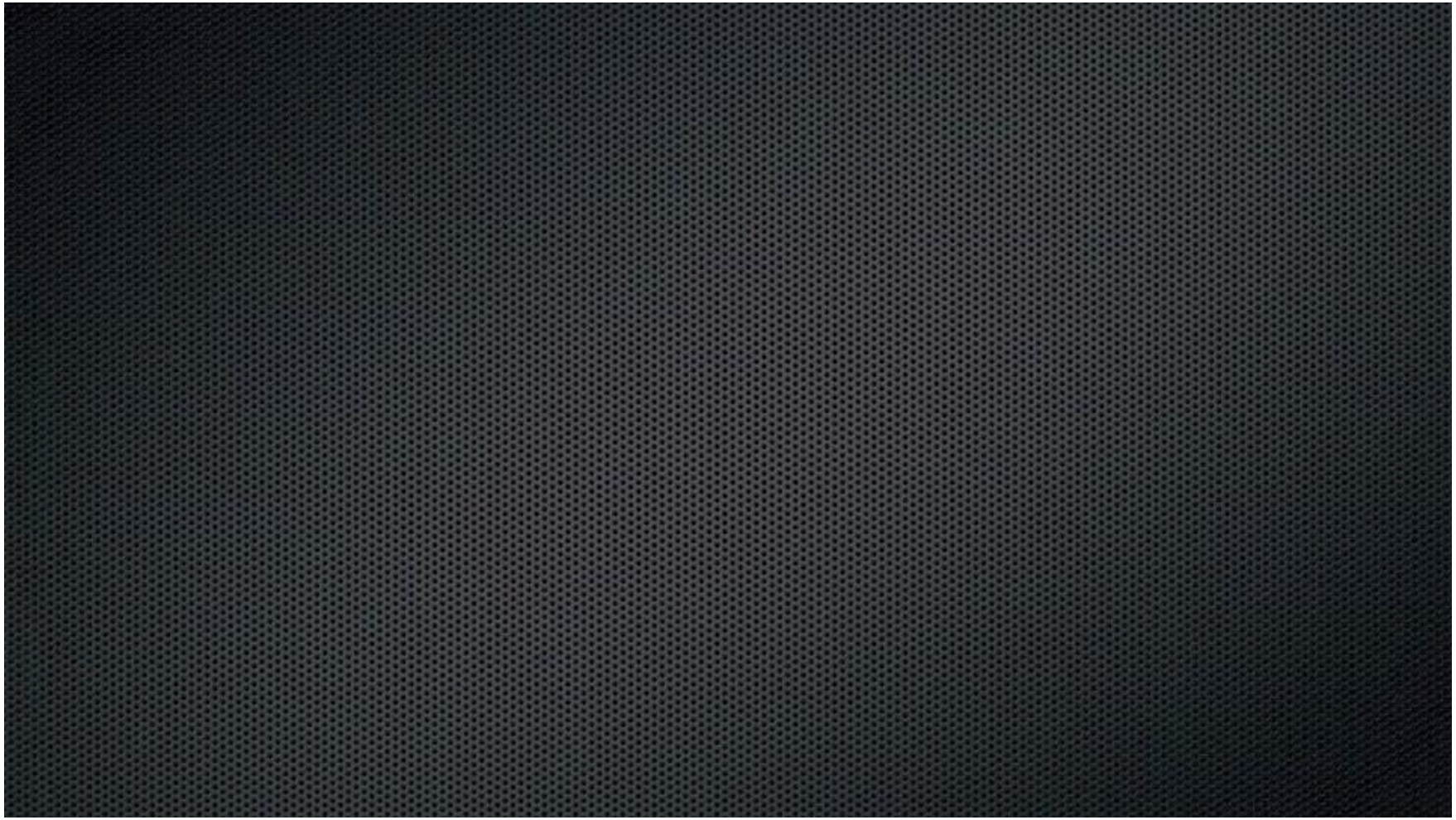


Comparing timelines extracted from different topics: (a) timeline extracted from topic B, which mainly describes protest events happening in Egypt; (b) timeline extracted from topic C, which focuses on the Obama's government's actions on events in Egypt.

Application Example – Bing News



Video



Outline

- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic Word Cloud
  - Topic-based Visual Text Summarization
  - TextFlow: Towards Better Understanding of Evolving Topics in Text
- **Text Visualization Perspectives**

Future Text Visualization Topics

- **Interactive, incremental** text analytics

Topic editing

TIARA LDA latent text semantic ~~David~~ models ~~edt~~ keywords summarization topic

Topic split

bibframe spotfire framemaker bibtex tableau visualization

spotfire tableau visualization ← → **bibframe framemaker bibtex**

- **Multi-level visual** text summarization **(keywords + sentences)**
- **Multi-faceted** text analytics **(e.g., summarization + sentimental analysis)**
- **Multimedia** document summarization **(text + image + video)**
- **Interactive, visual social media analysis**

Acknowledgements

Nan Cao, Yingcai Wu, Weiwei Cui, Prof. Huamin Qu (HKUST)
Dr. Michelle X Zhou (IBM Almaden Research Center)

Thanks a lot for your attention!