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
- **Future work**

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	How customers' sentiment have changed toward my hotels	How do customers feel about my new product launch
• 3000+ customer reviews	• 3000+ customer-posted reviews	• 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"

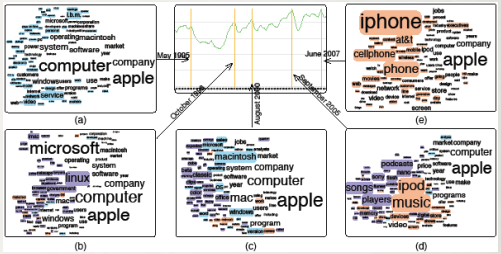
Machines are *not* just smart enough.

Outline

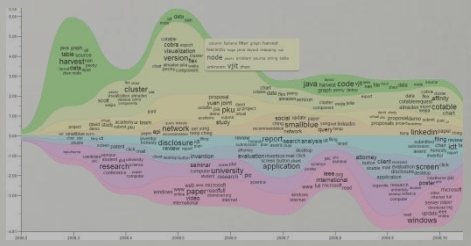
- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic word cloud
  - TIARA: topic-based visual text summarization
  - TextFlow: towards better understanding of evolving topics in text
- **Future work**

Selected Projects

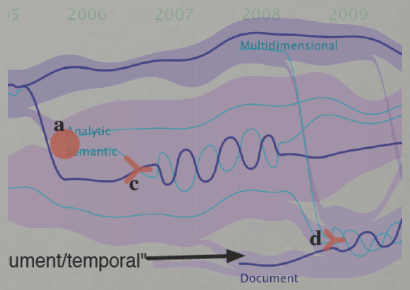
- **Dynamic word cloud**
  - Illustrate content evolution trend



- **TIARA**
  - Topic-based visual text summarization and analysis



- **TextFlow**
  - Towards better understanding of evolving topics in text



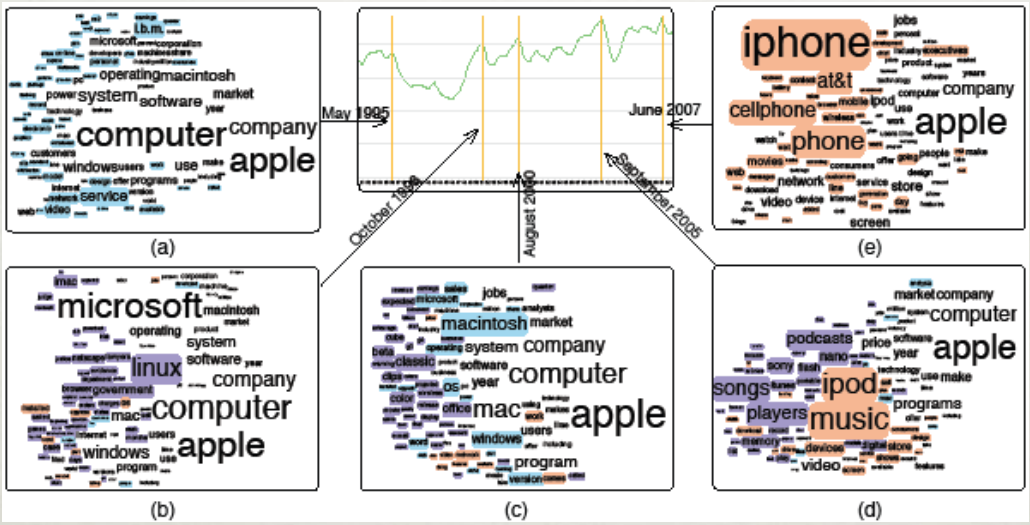
Cui et al. IEEE CG&A 2011

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



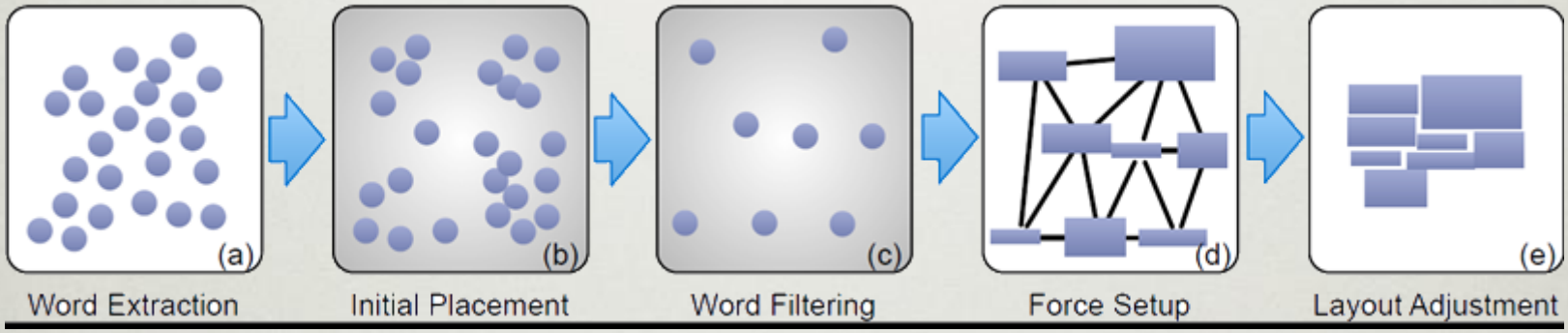
Evolution Measurement

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

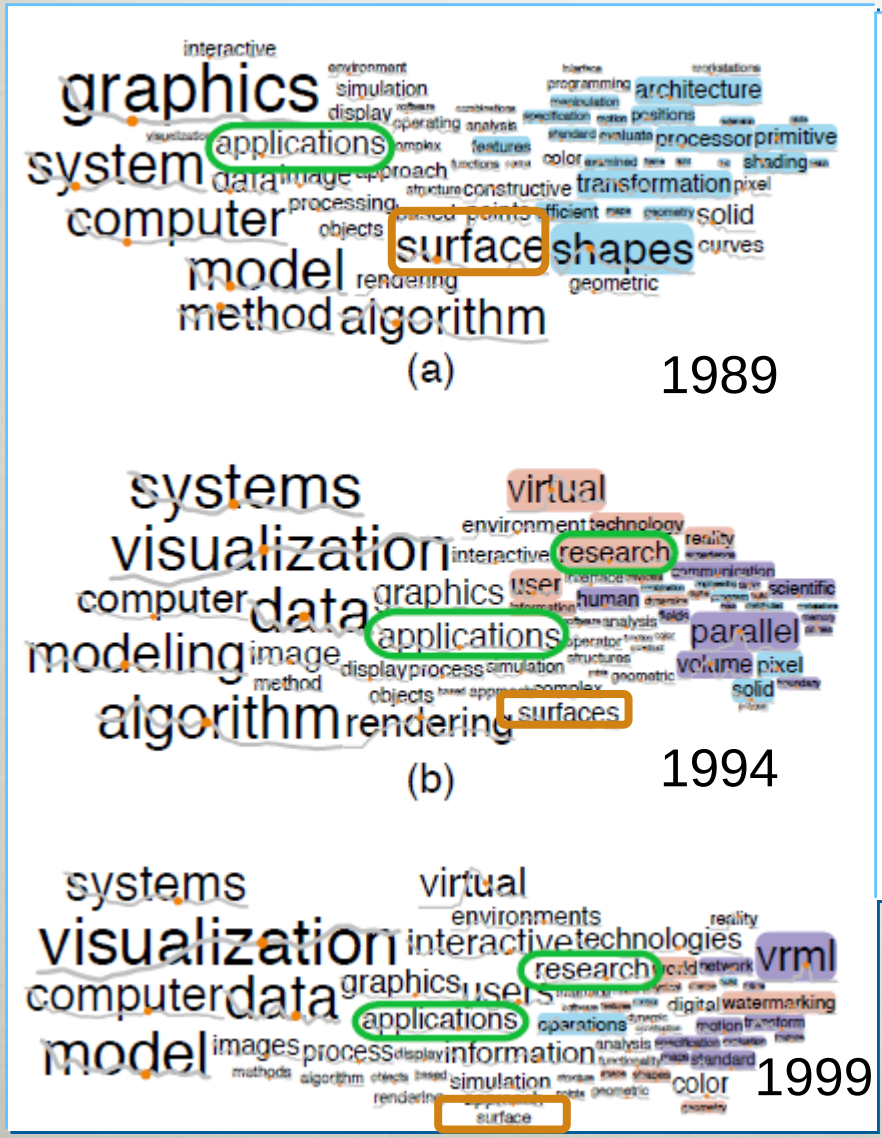
$$S(X_i) = \sum_{j=-w/2}^{w/2} t_j H(X_i|X_{i+j}) = \sum_{j=-w/2}^{w/2} t_j (H(X_i) - H(X_i; X_{i+j}))$$

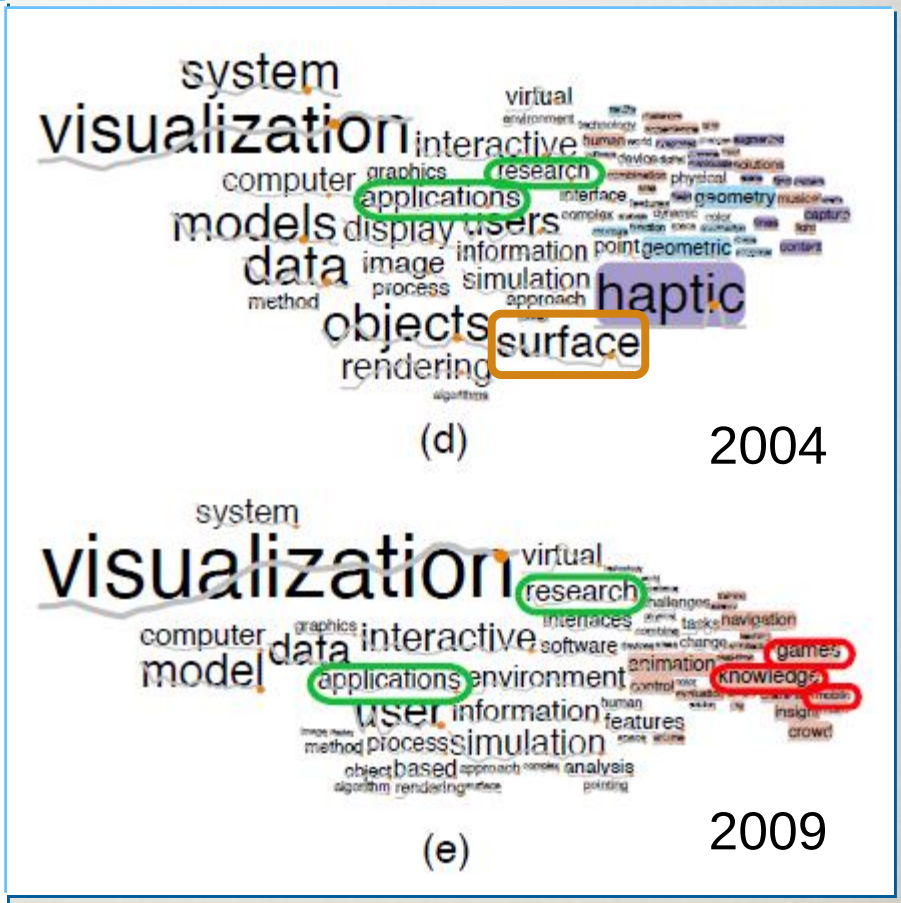
Word Cloud Layout

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



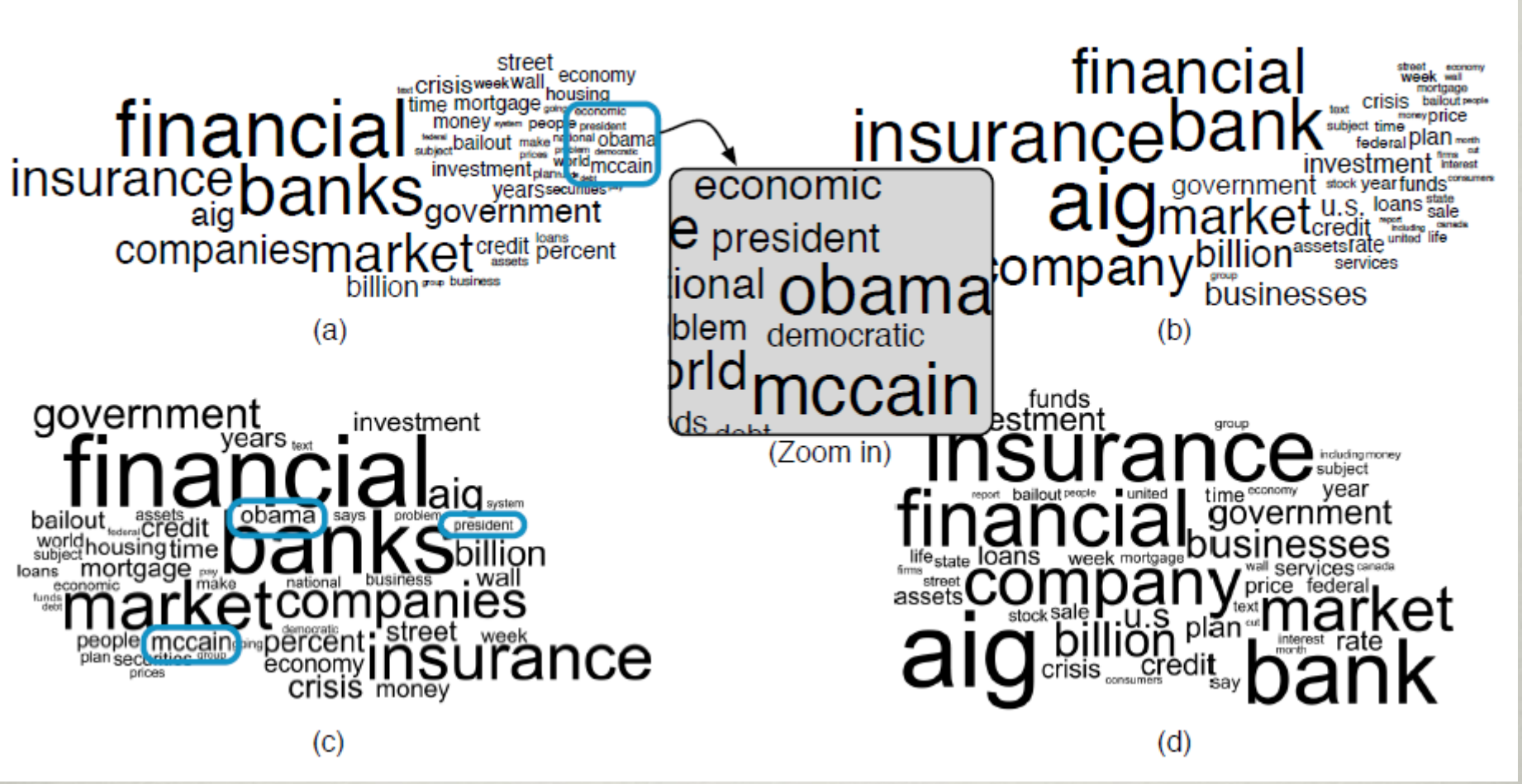
Example: CG&A Abstracts





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

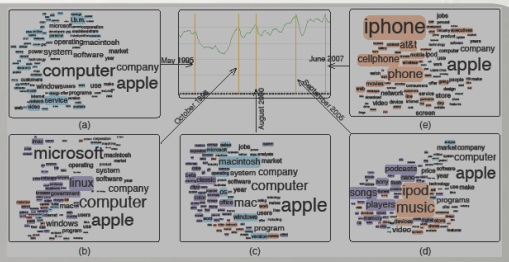
Comparison with Wordle



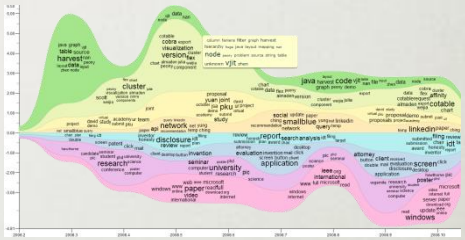
13,828 news articles

Selected Projects

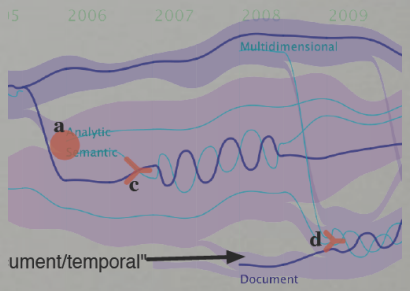
- **Dynamic word cloud**
  - Illustrate content evolution trend



- **TIARA**
  - Topic-based visual text summarization and analysis

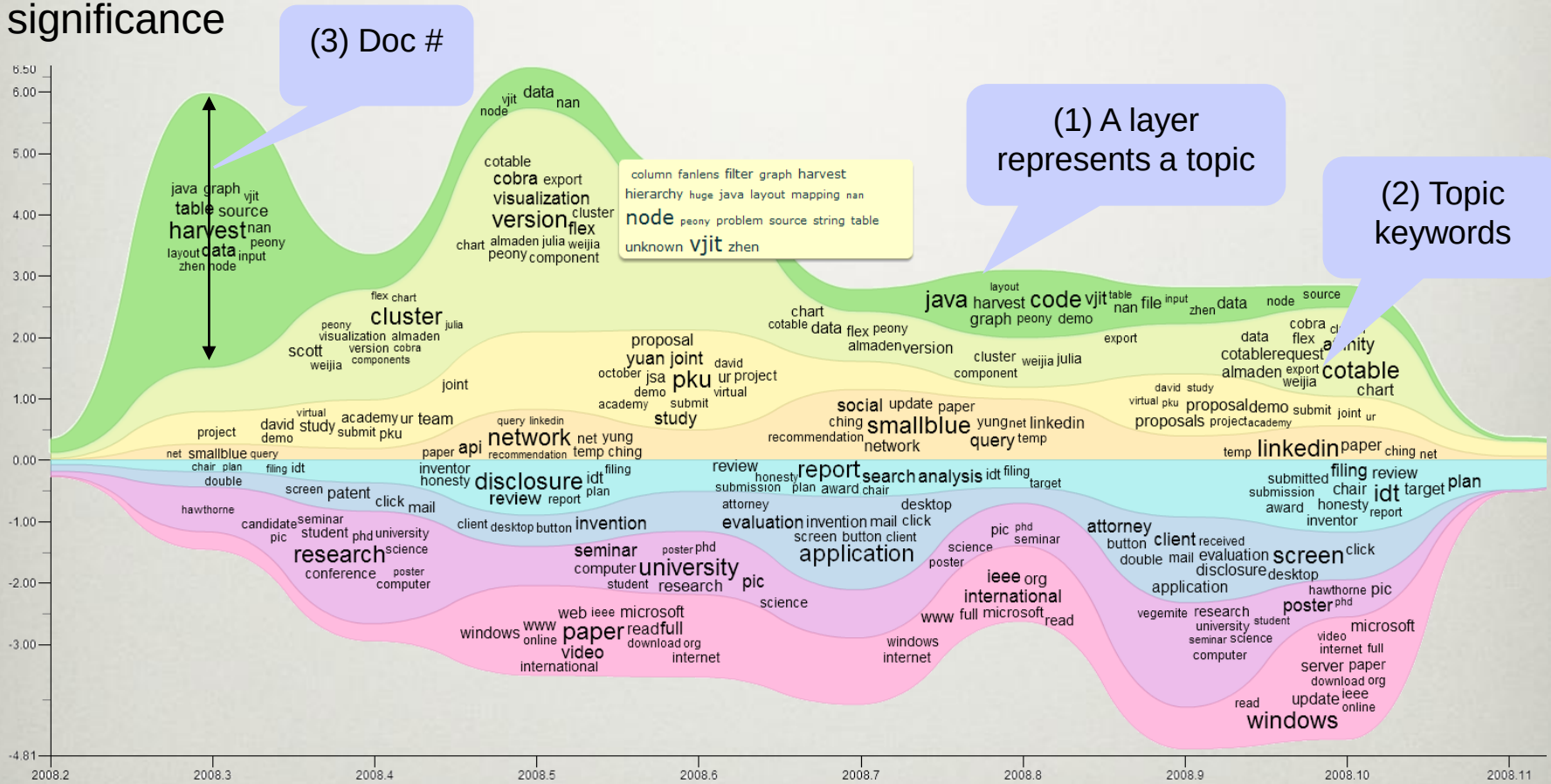


- **TextFlow**
  - Towards better understanding of evolving topics in text



Demo

Y axis encodes topic significance



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

Key Challenges

- **Summarize text corpora**
  - Huge amounts of complex information
  - Time-varying
- **Visually explain summarization results**
  - Consistent visualization
- **Provide feedback or articulate their needs**
  - Imperfect summarization results or varied user needs

TIARA Overview

Text collection

Text Preprocessing

Text content + meta data

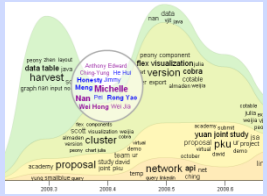
Text Summarization

Summarization results

Data Transformation

Transformed results

Visualization



User Interaction

TIARA Technical Focus

Text collection

Text Preprocessing

Text content + meta data

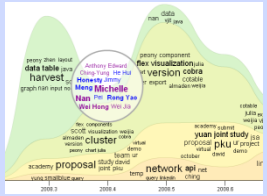
Text Summarization

Summarization results

Data Transformation

Transformed results

Visualization



User Interaction

Text Summarization

- **Latent Dirichlet Allocation (LDA) model [Blei et al. 03]**
  - High portability
  - High compaction rate for scalability
  - A finer grained model

*k*th document in the collection

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

A set of topic *probabilities*

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

A set of topics

A set of keywords

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

A set of word *probabilities*

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

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"**

Domain-dependent activeness metric

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

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

Experiments

- **Goal**
  - Measure which metric produces more “important” topics
- **Data sets**
  - Email
    - 8326 email messages
  - News
    - 34,690 documents
- **Method**
  - Users indicate the importance of top-K ranked topics
    - Very important, somewhat important, Unimportant

Results

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

Retrieved	Top 5	Top 10
Strength	0.800 ± 0.000	0.620 ± 0.028
Distinctiveness	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.780 \pm 0.028}$
M.I.	0.760 ± 0.106	0.740 ± 0.035
T.S.	0.440 ± 0.057	0.480 ± 0.028

- **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

Text collection

Text Preprocessing

Text content + meta data

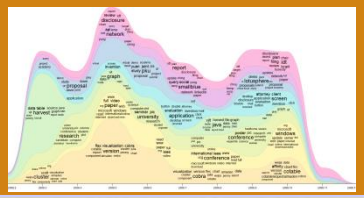
Text Summarization

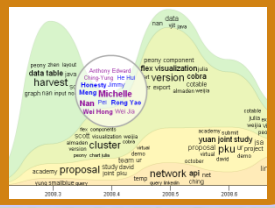
Summarization results

Data Transformation

Transformed results

Visualization

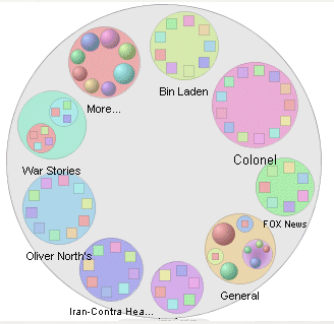


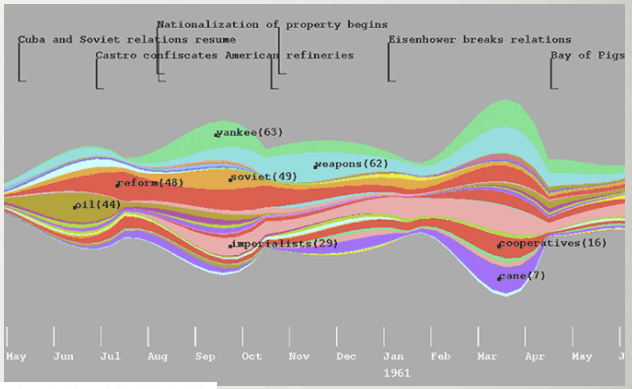


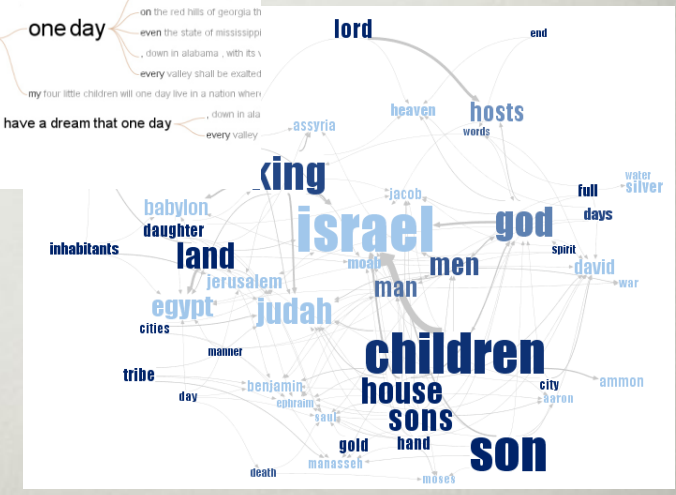
User Interaction

Visualizing Text: Existing Work

- Visualize text at a high level
- Visualize text at a low level
- Few on explaining advanced text analysis results





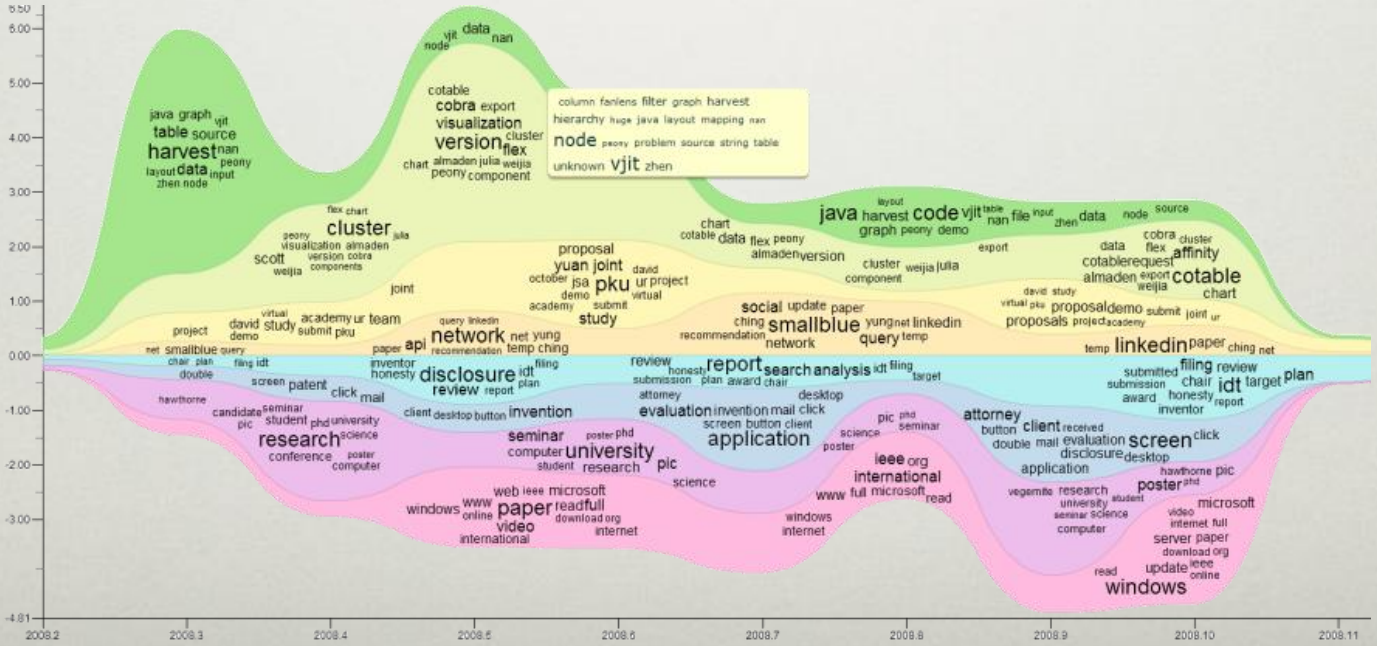


Visual Text Summary Metaphor

Data to be visualized:

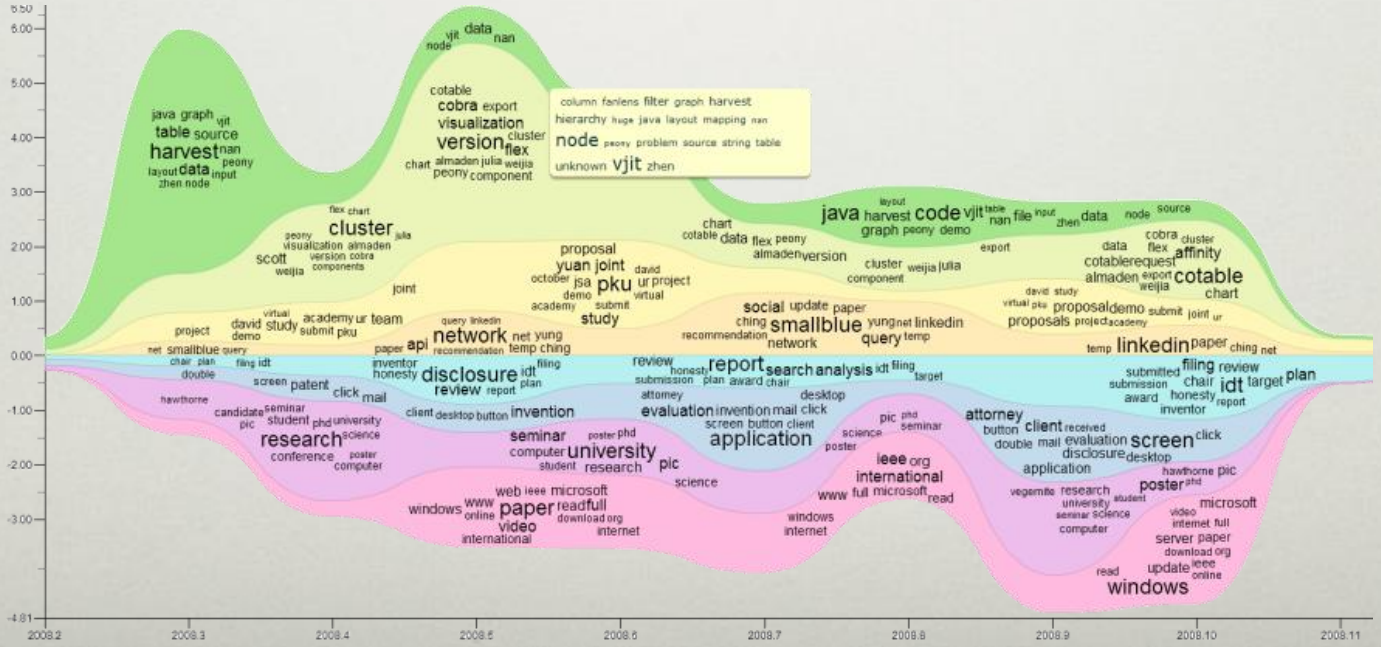
1. Topics: $\{T_1, \ldots T_i, \ldots T_N\}$ and their probabilities
2. For each T_i ,Topic keywords by time: … $\{\ldots, w^i_k, \ldots, \}_t$, … and their probabilities over time
3. For each T_i , Topic strength: $\{\ldots, S^i(t), \ldots\}$ over time

Visual encoding: Augmented stacked graph



Enhanced Stacked Graph: Key Steps

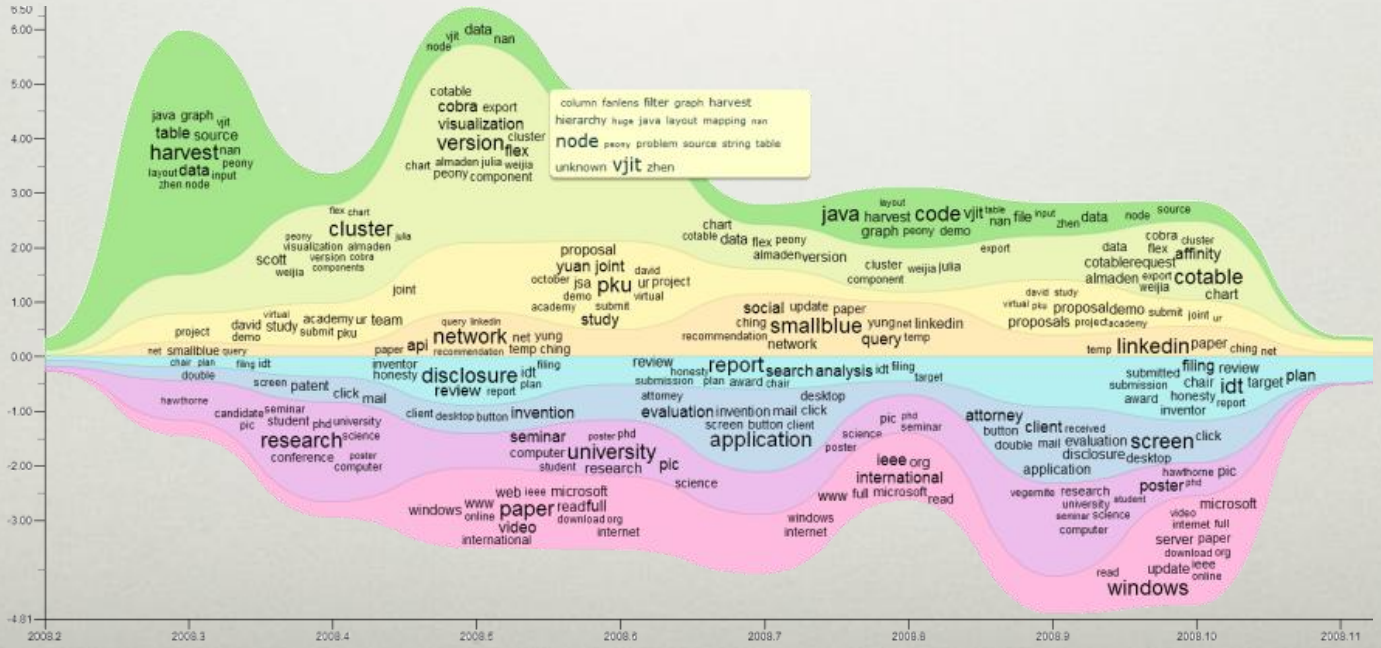
- Computing geometry of layers
- Layer coloring
- Layer ordering
- Layer labeling



Enhanced Stacked Graph: Key Steps

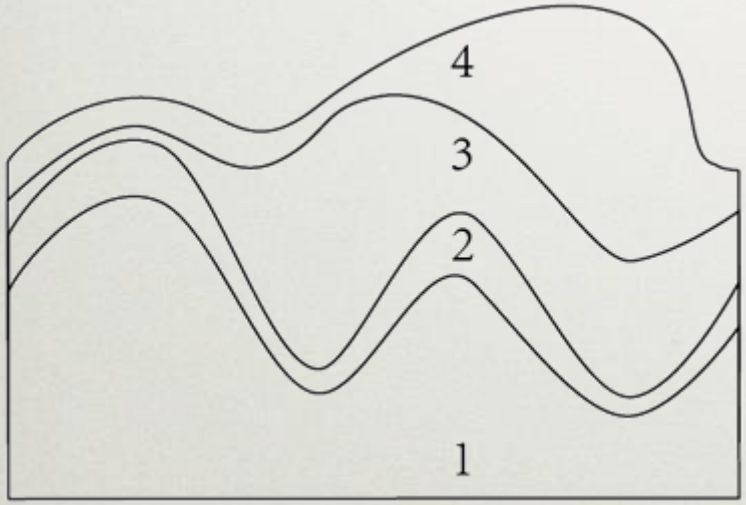
- Computing geometry of layers
- Layer coloring
- **Layer ordering**
- **Layer labeling**

Byron_Infovis08

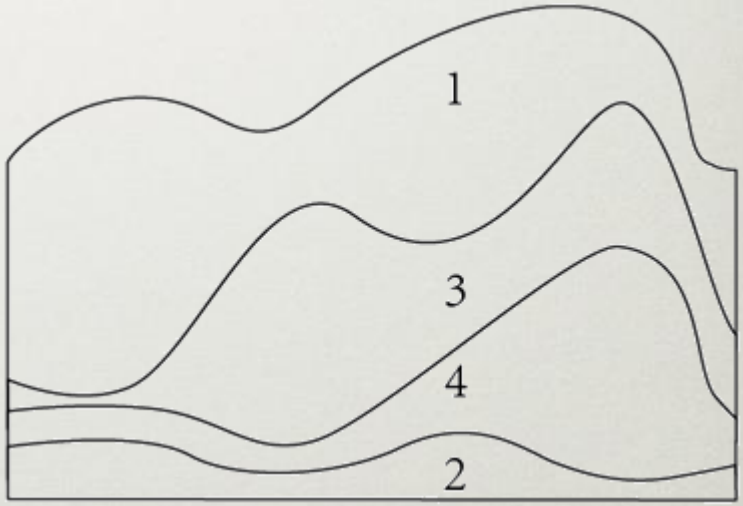


Layer Ordering

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

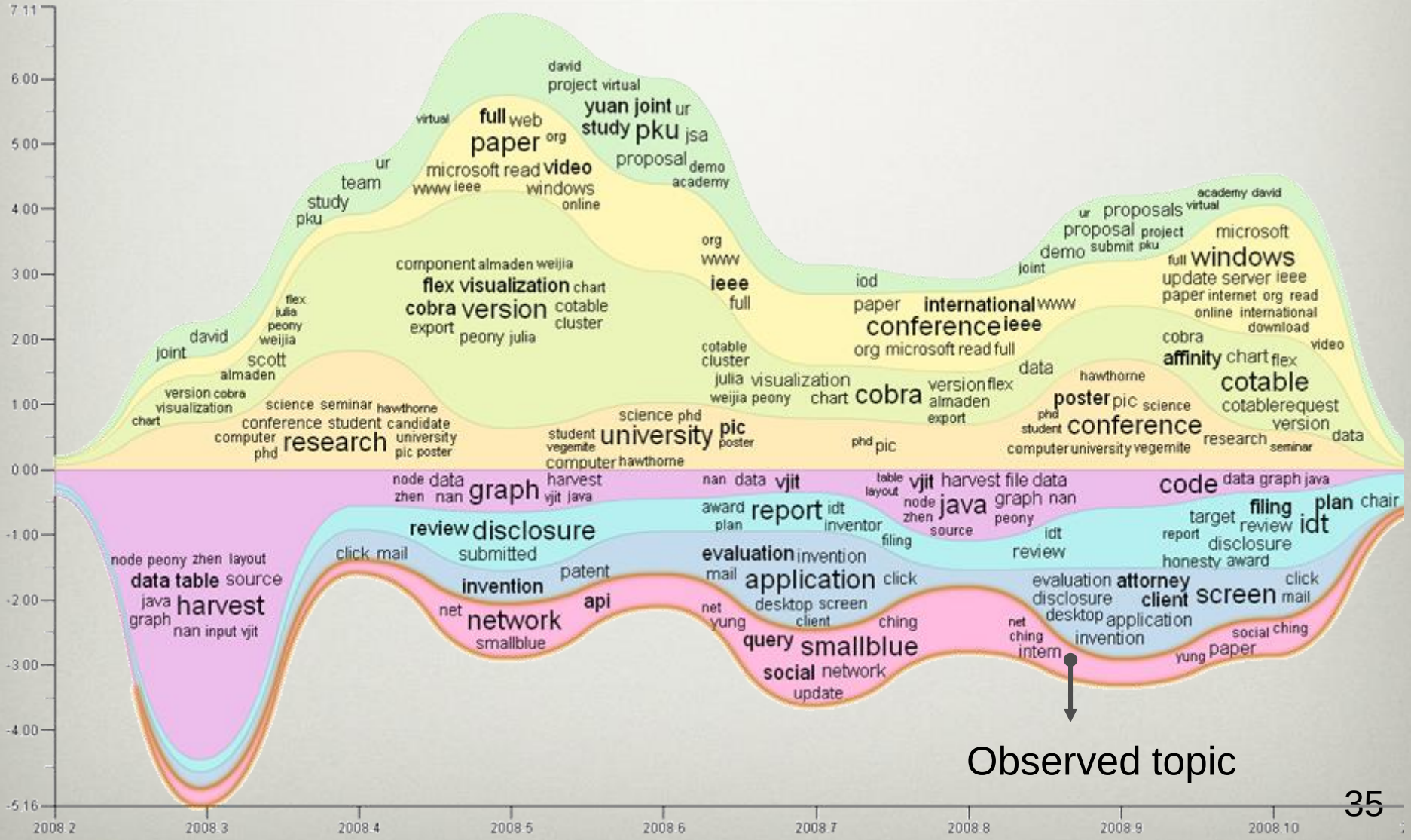


unordered

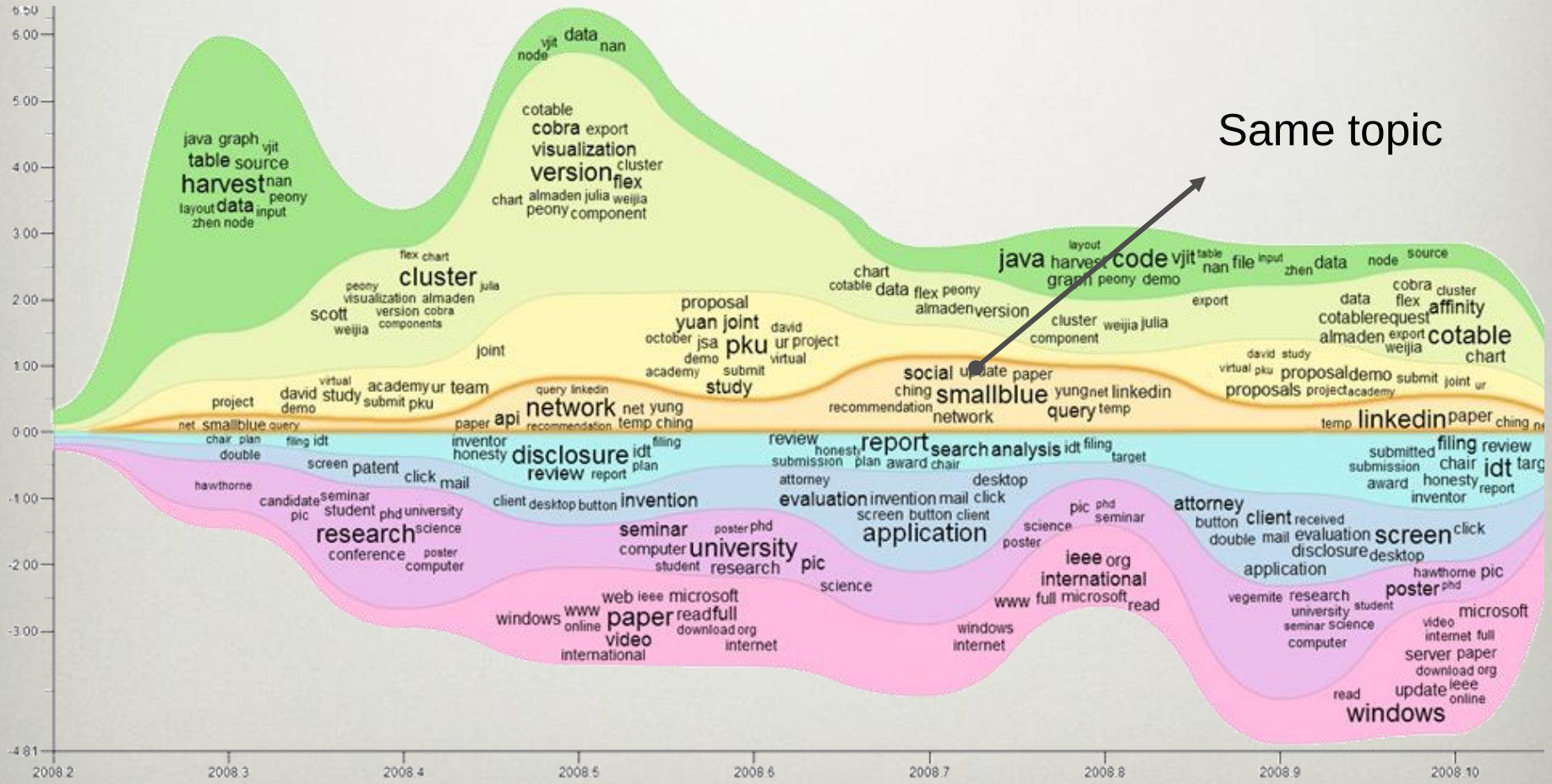


ordered

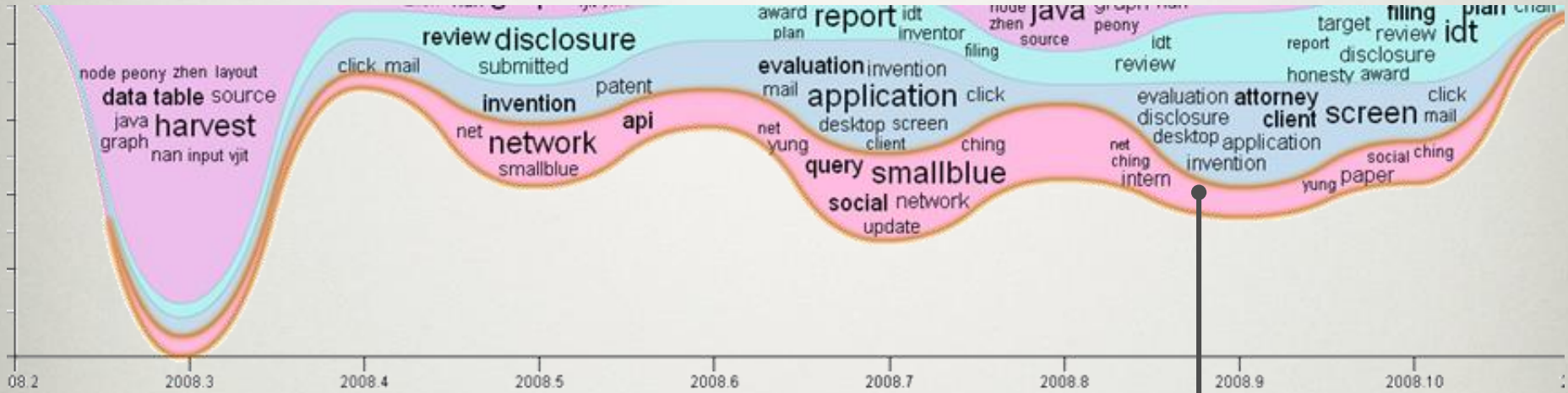
Layer Ordering - Comparison



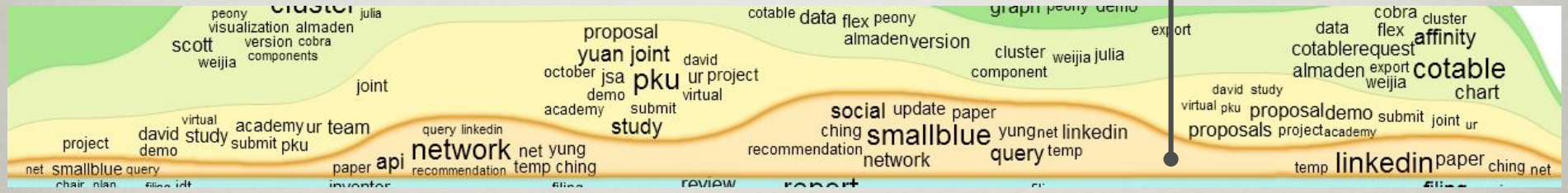
Layer Ordering - Comparison



Layer Ordering - Comparison



Same topic

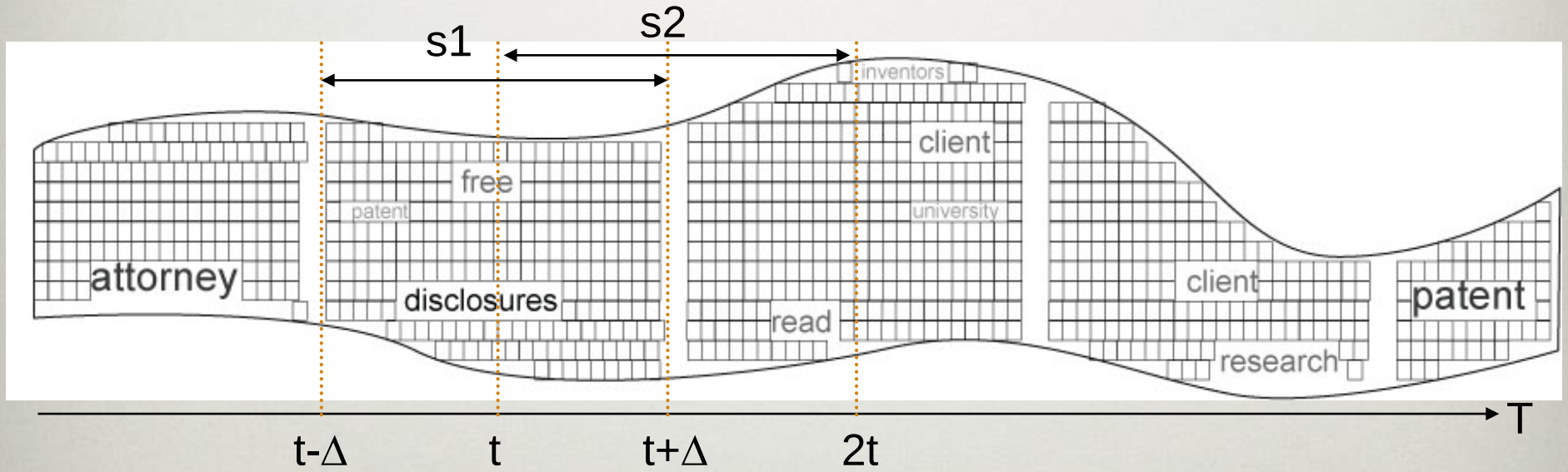


Enhanced Stacked Graph: Layer Labeling

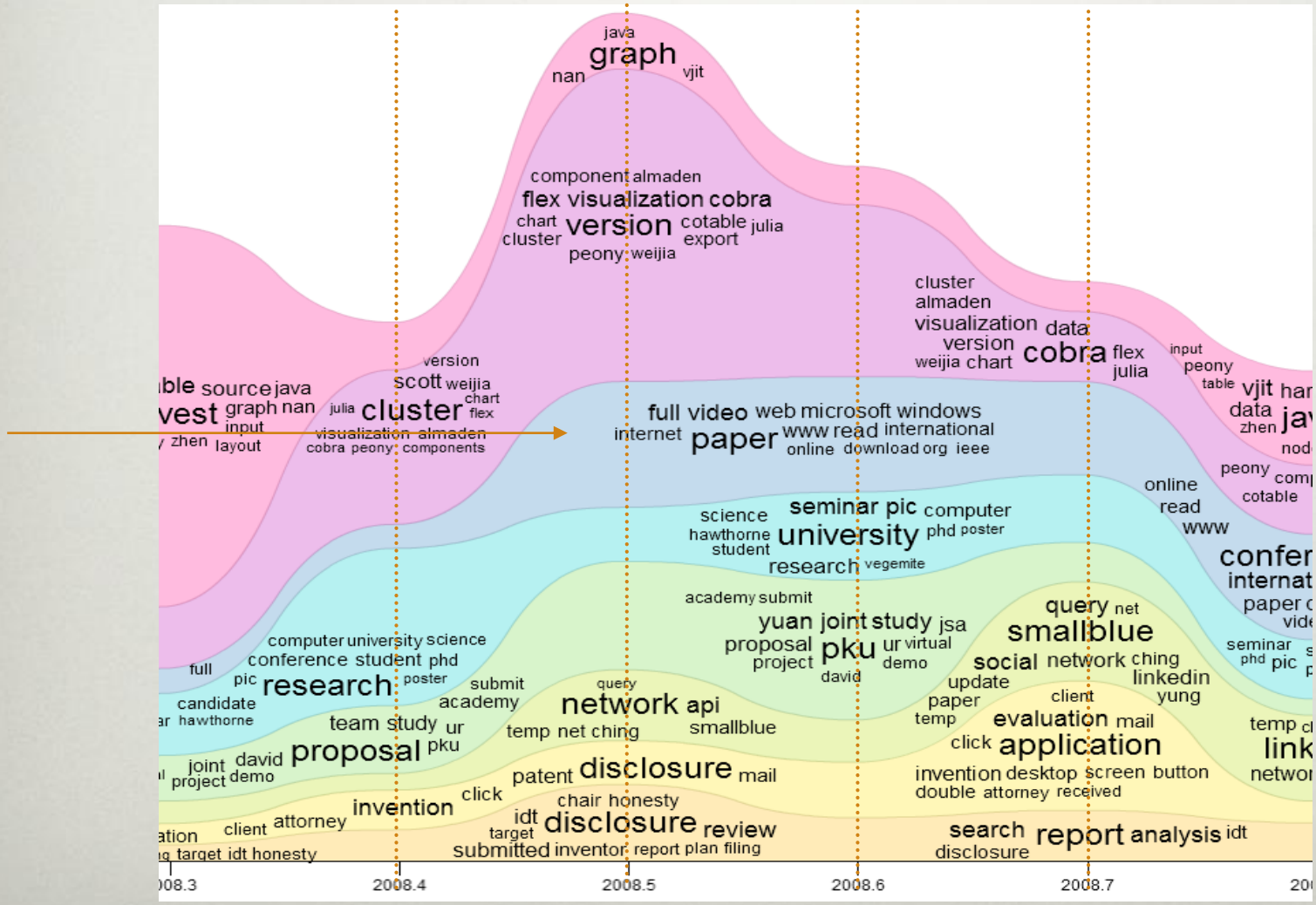
- **Goals**
  - Temporal proximity
  - Informativeness

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

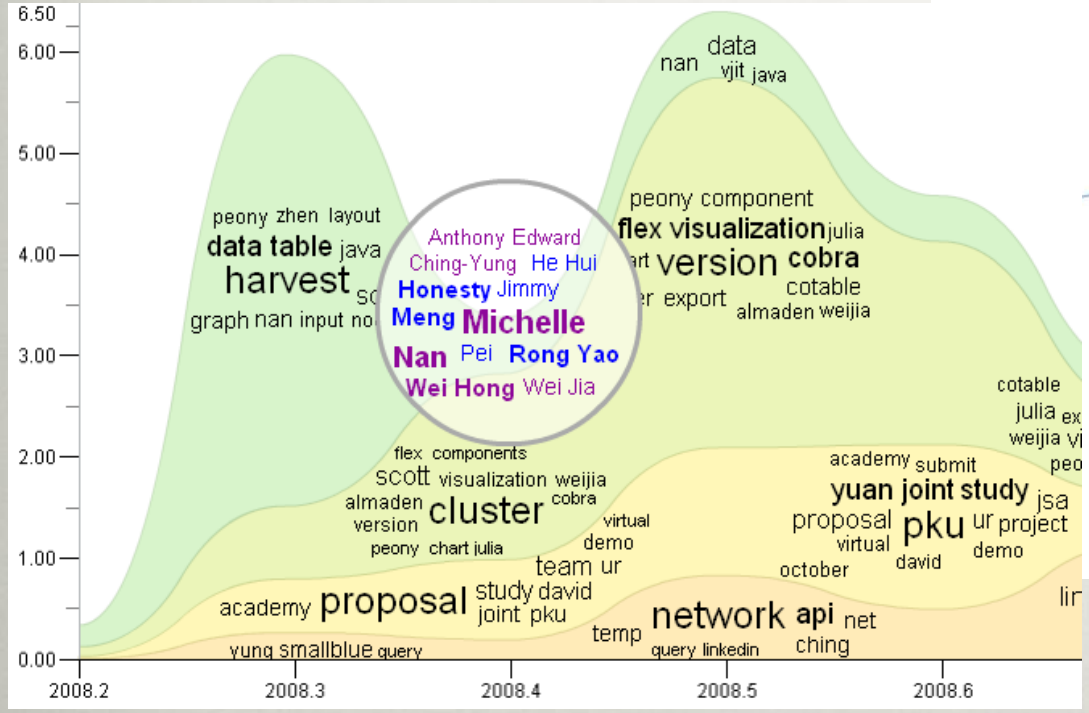


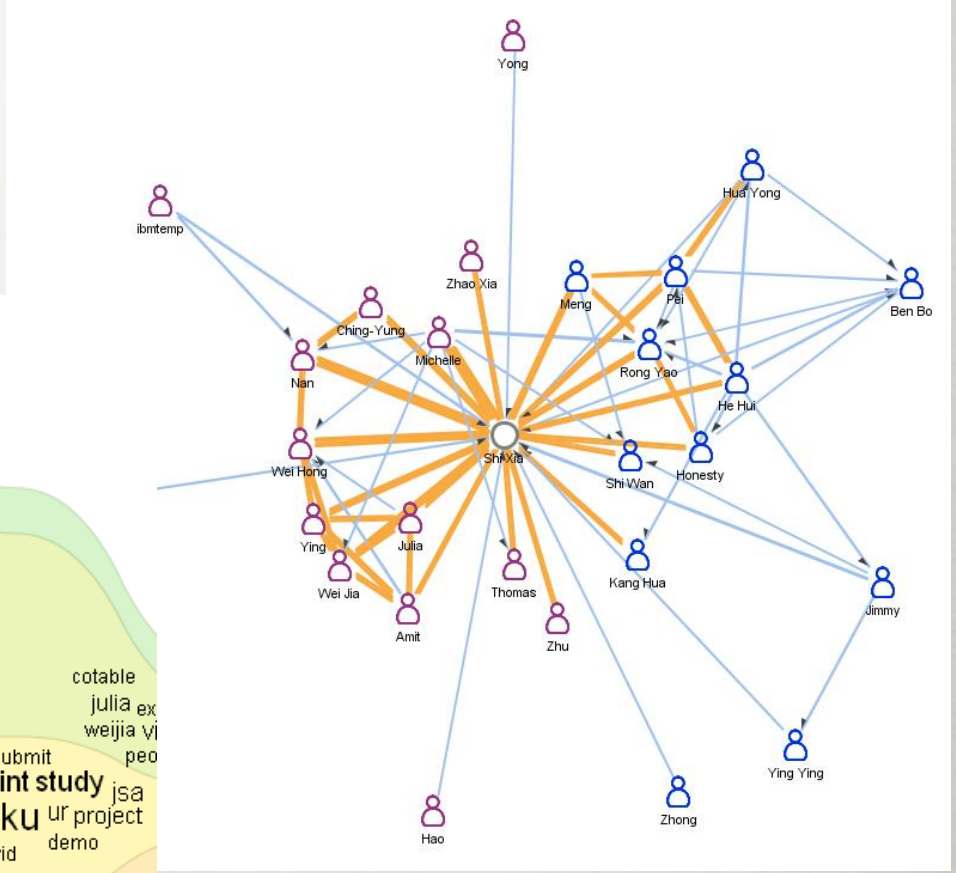
Enhanced Stacked Graph: Layer Labeling (cont'd)



Interacting with Visual Summary

"people" involved and their relationships

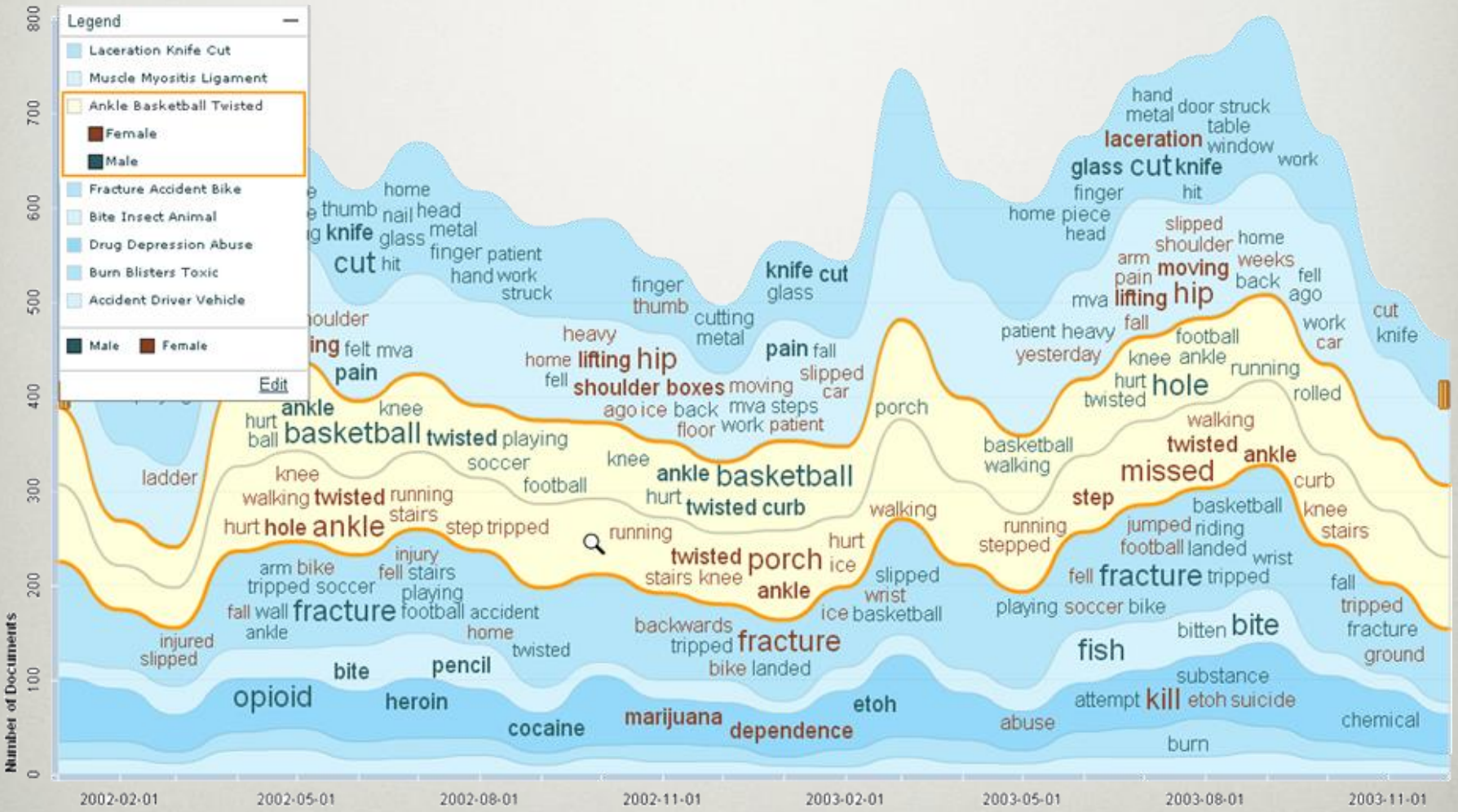




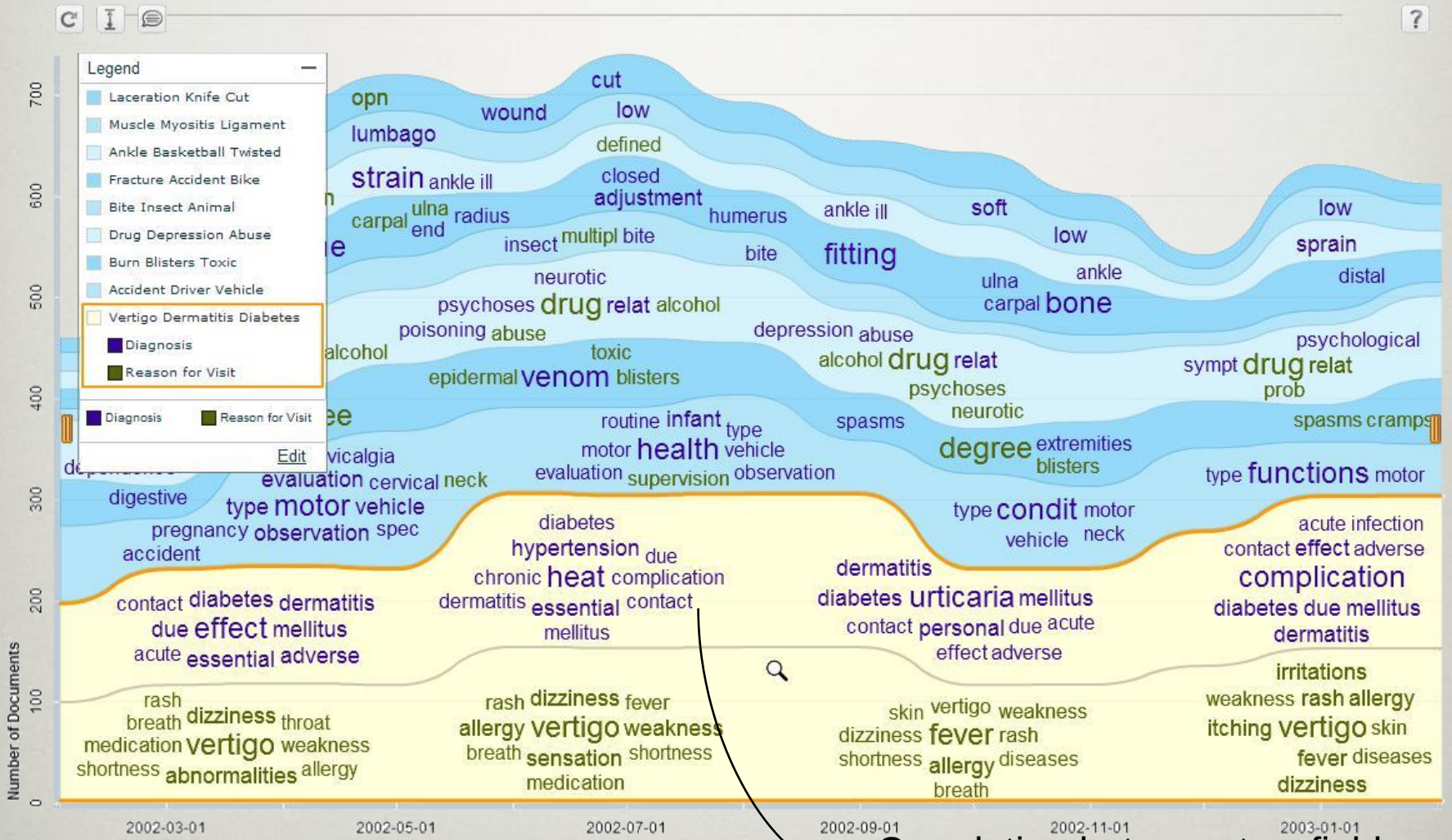
Application Example: Healthcare

- **Visualize text to facilitate analysis**
  - Cause of injury
  - Reason for visit
  - Diagnosis
- **Multiple fields of text data and their correlation**
- **Leverage structured data to help better illustrate text information**
  - Gender + Cause of injury

Correlation between Structured and Text Fields



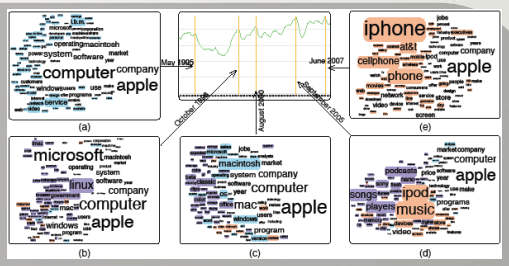
Correlation between Text Fields



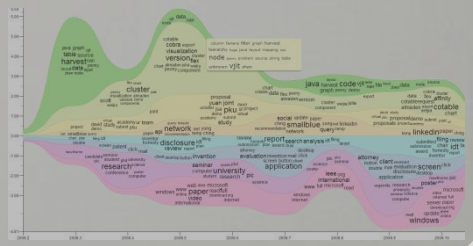
Correlation between two fields, *diagnosis* and *reason for visit*

Selected Projects

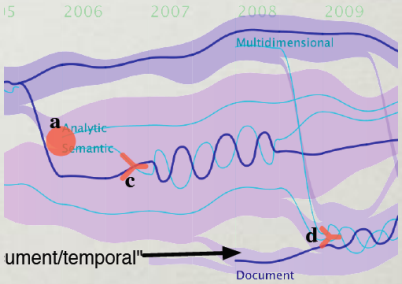
- **Dynamic word cloud**
  - Illustrate content evolution trend



- **TIARA**
  - Topic-based visual text summarization and analysis



- **TextFlow**
  - Towards better understanding of evolving topics in text



Problems

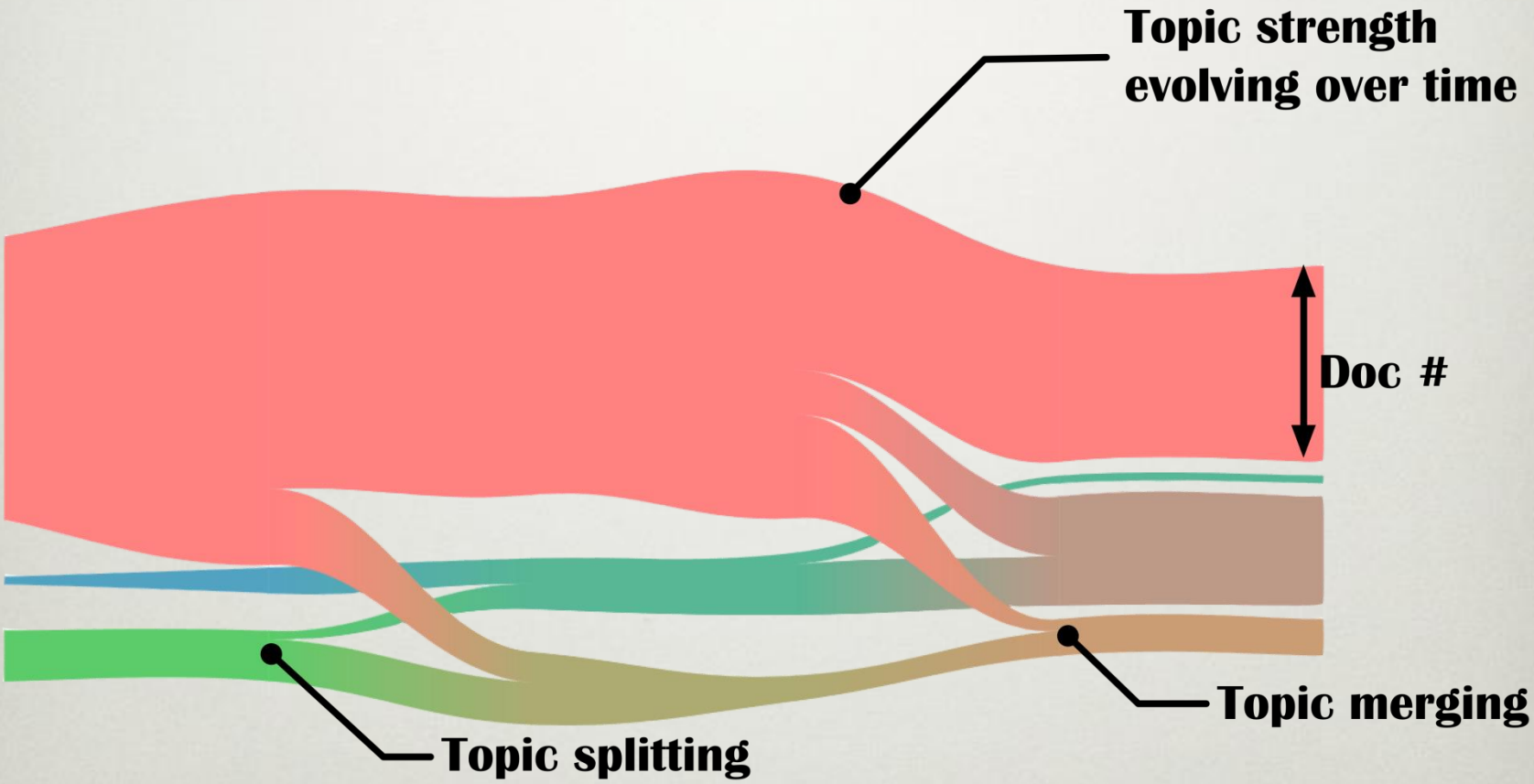
Understanding topic evolution in large text collections is important

- **Keep abreast** of hot, new, and intertwining topics
- **Gain insight** into the latent topics

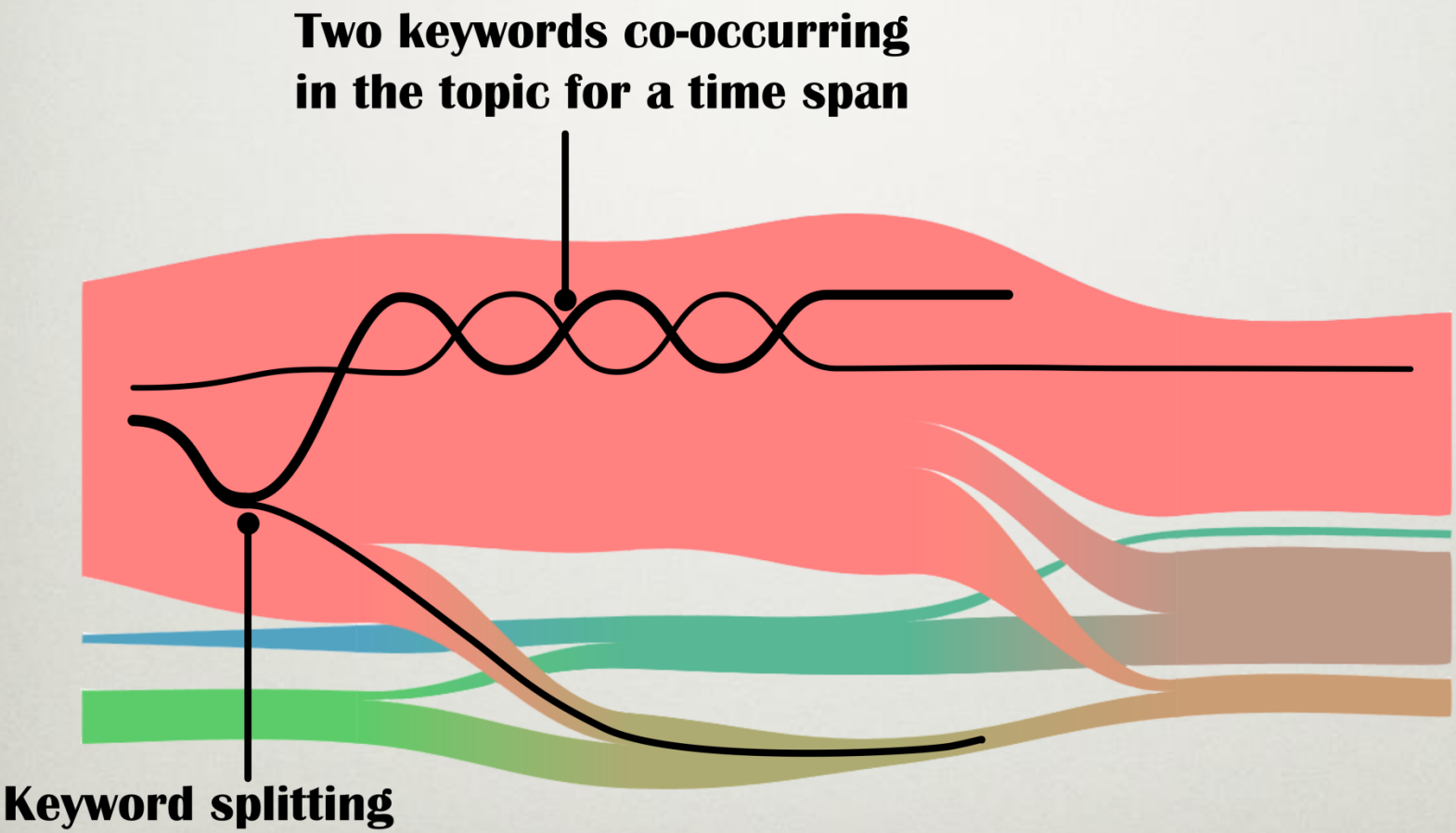
Applications

- **Scholars**
  - Find related works in a publication set
- **Business professional**
  - Examine a large collection of emails and instant messages
- **Politicians**
  - Examine online posts to identify the key public opinion and concern

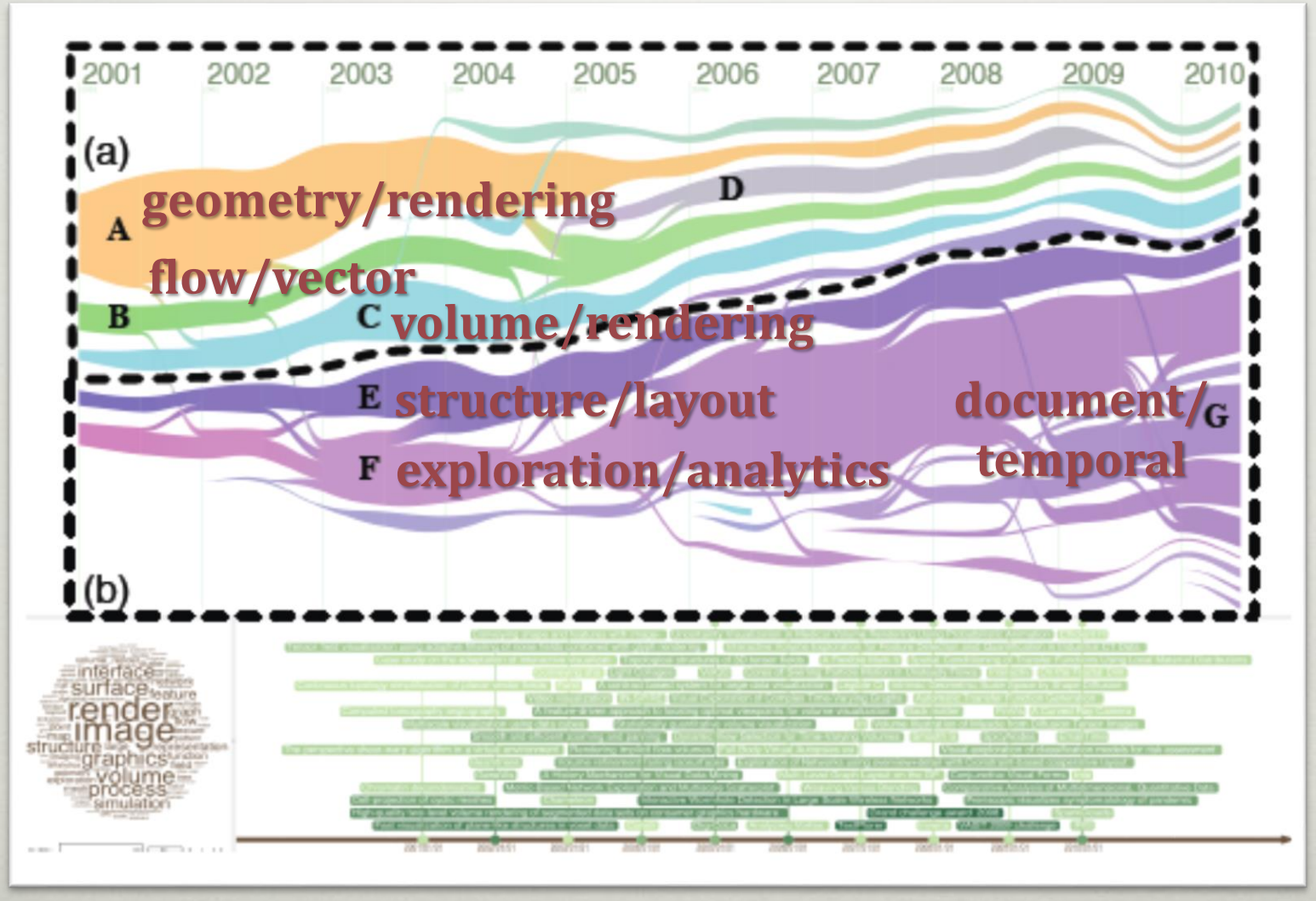
Example


Topic strength
evolving over time
Doc #
Topic merging
Topic splitting

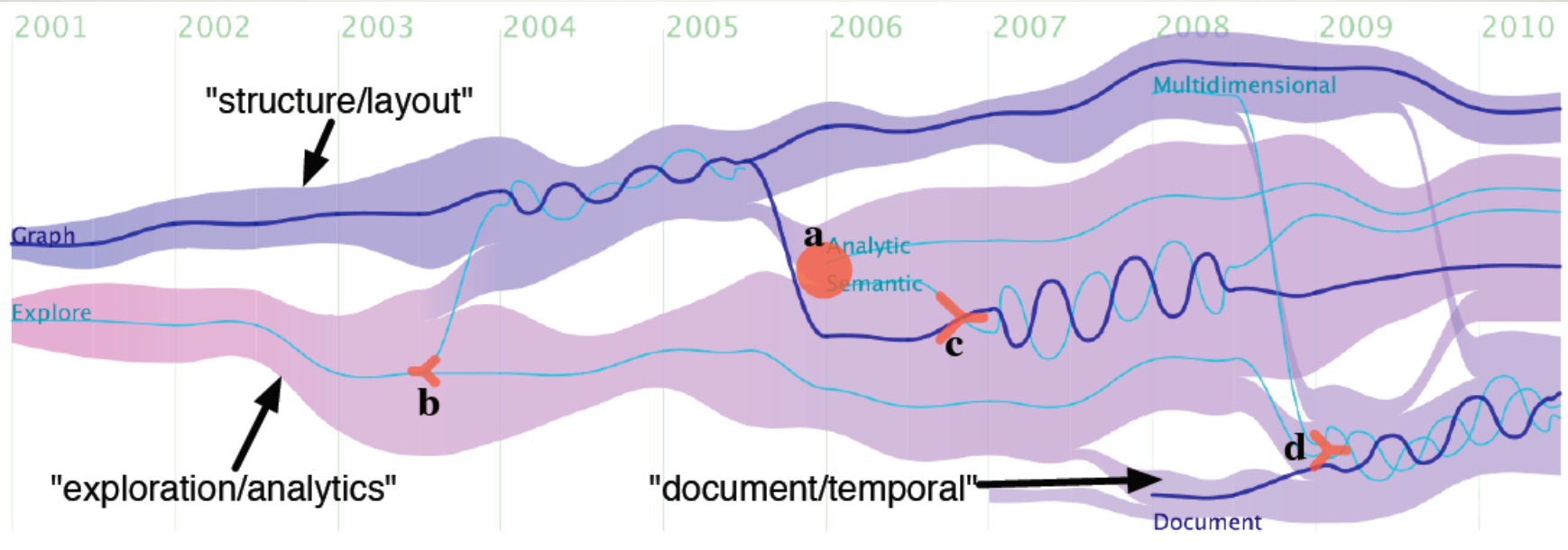
Example


Two keywords co-occurring
in the topic for a time span
Keyword splitting

Application Example: 933 VisWeek Pulications



Application Example: 933 VisWeek Pulications



Challenges

- **Model**
  - Topic merging/splitting patterns
- **Visually convey**
  - Topic merging/splitting patterns in an intuitive way
- **Facilitate**
  - Analytical reasoning

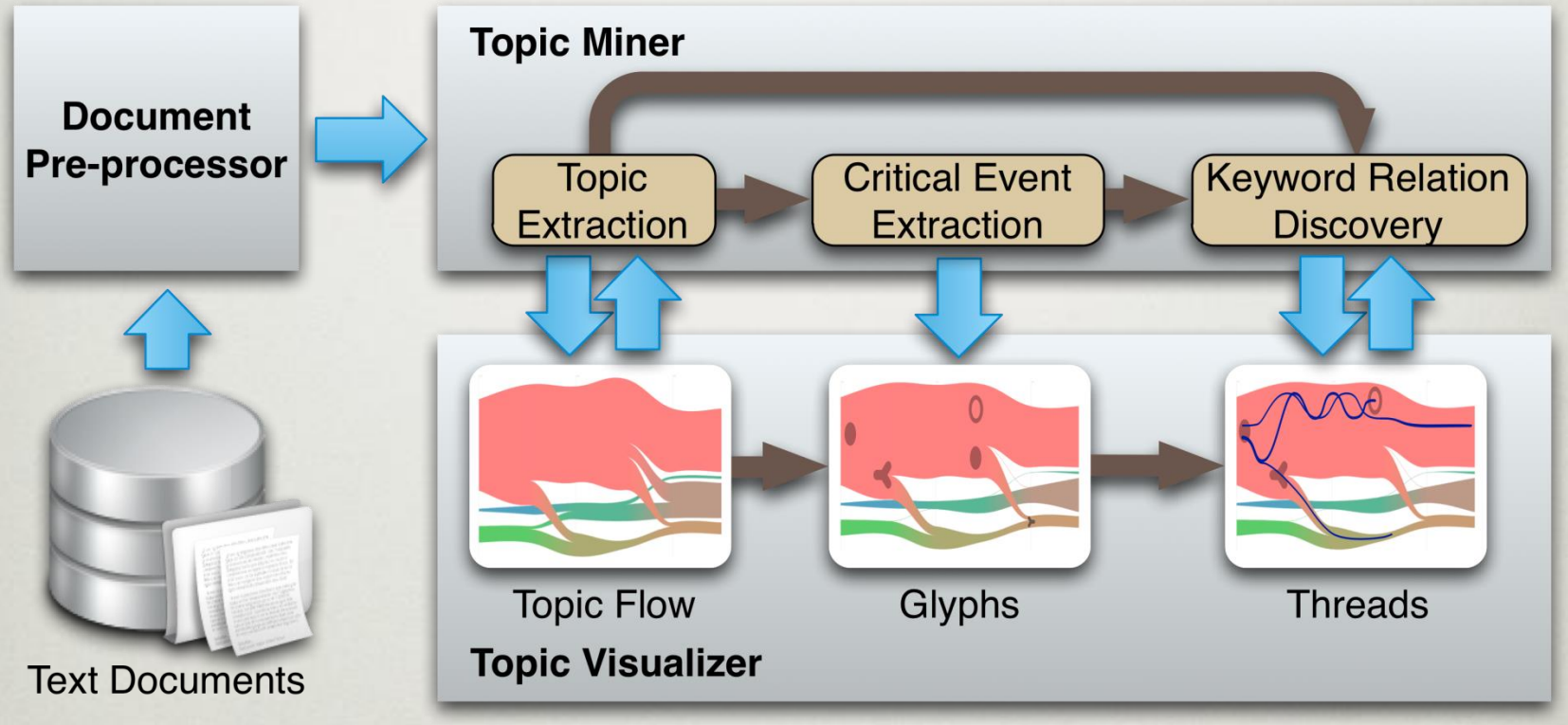
Related Work

- **Most of the existing work**
  - Studying the evolution of individual topics
- **Little work**
  - Studying topic merging and splitting patterns
- **Barely been touched**
  - Using visual analytics techniques to interactively analyze complex topic evolution

Our Solution

- **Leverage *hierarchical Dirichlet processes***
  - To model topic merging/splitting
- **Augment familiar visual metaphors (rivers)**
  - To convey the complex analytic results
- **Support interactions at different levels**
  - Smooth communication between visualization and the topic mining model

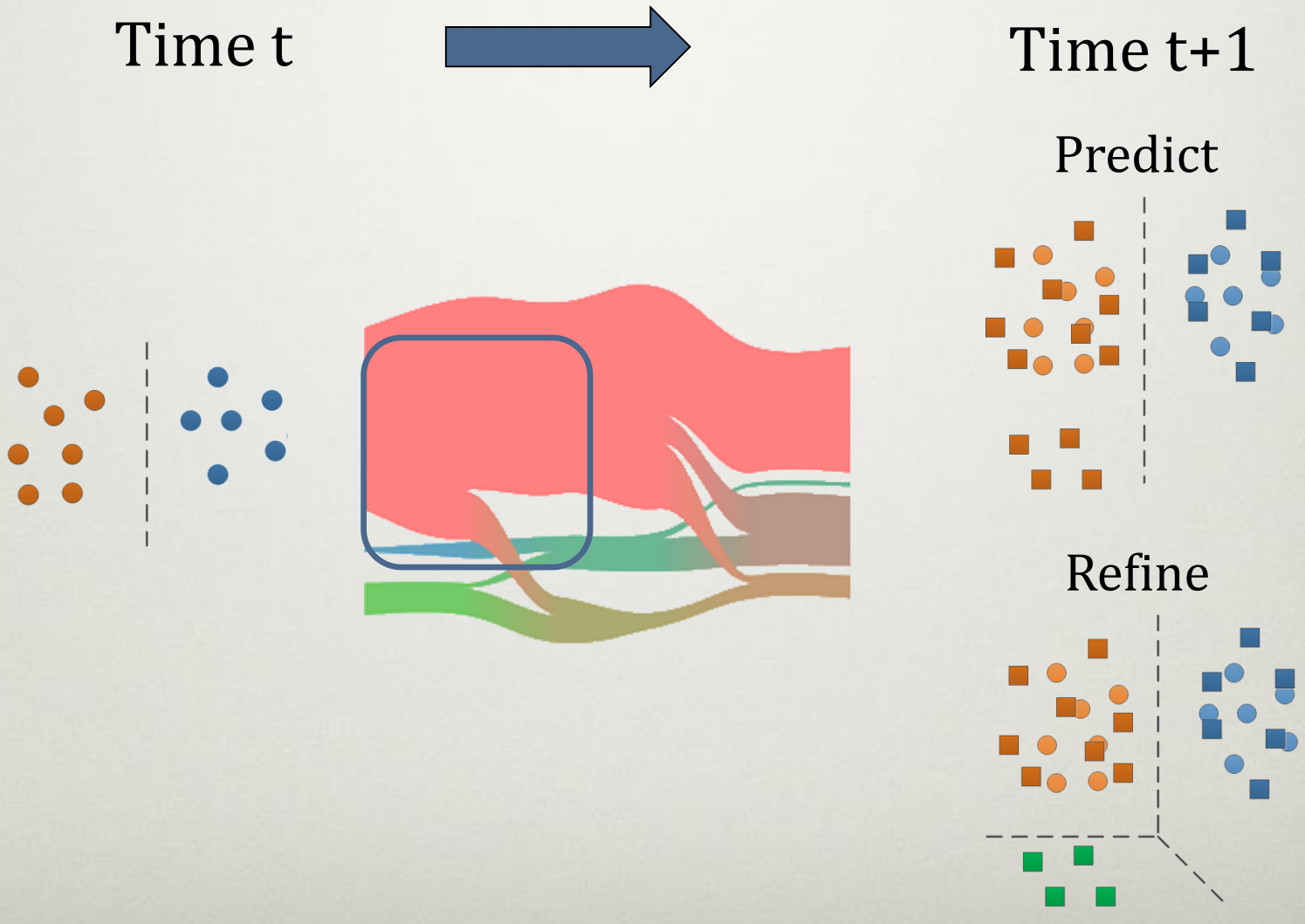
TextFlow Overview



Topic Data and Relationship Extraction

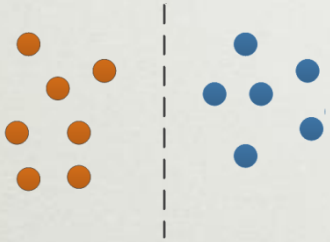
- **Incremental Hierarchical Dirichlet Processes**
  - Online learning of the topics in text
  - Automatically detect the topic numbers
  - Extract the merging/splitting relationships
    - Based on document topic change
    - Online compute the merging/splitting probabilities

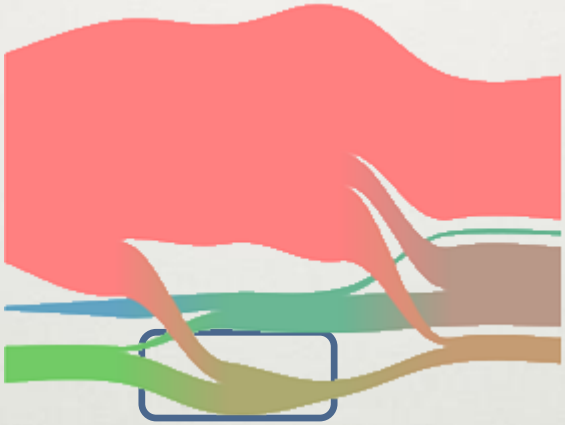
Splitting Relationship



Merging Relationship

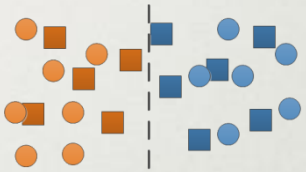
Time t



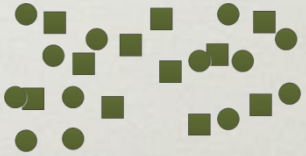


Time t+1

Predict



Refine



Critical Event Extraction

- **Types of critical events**
  - Birth, death, merge, and split
- **Scoring the merging/splitting event**
  - Number of the branches
  - Entropy of the branching probabilities

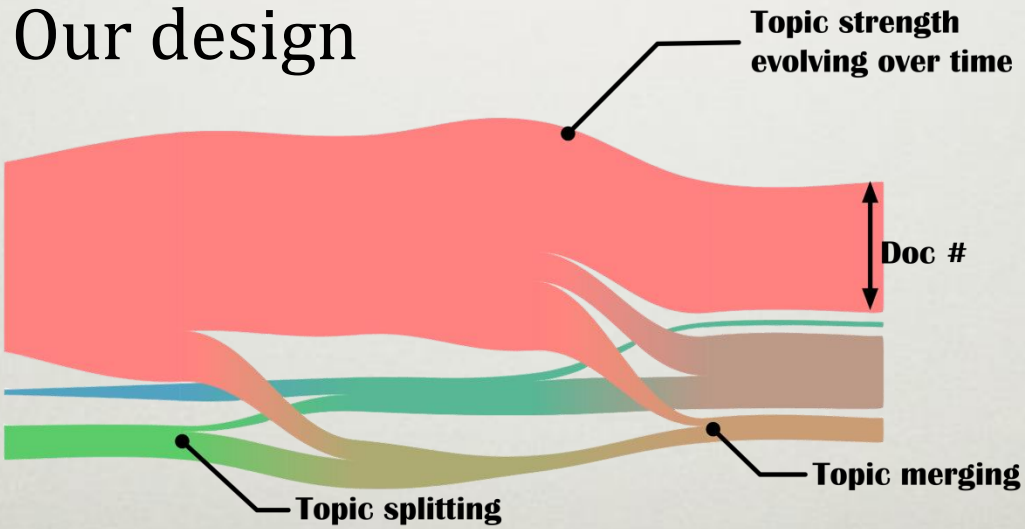
Keyword Correlation Discovery

- **Extract**
  - noun phrases, verb phrases, and named entities in each document
- **Count**
  - Co-occurrences among them
- **Be used to illustrate "why"**

Topic Evolution as Flow

Alternative

Our design

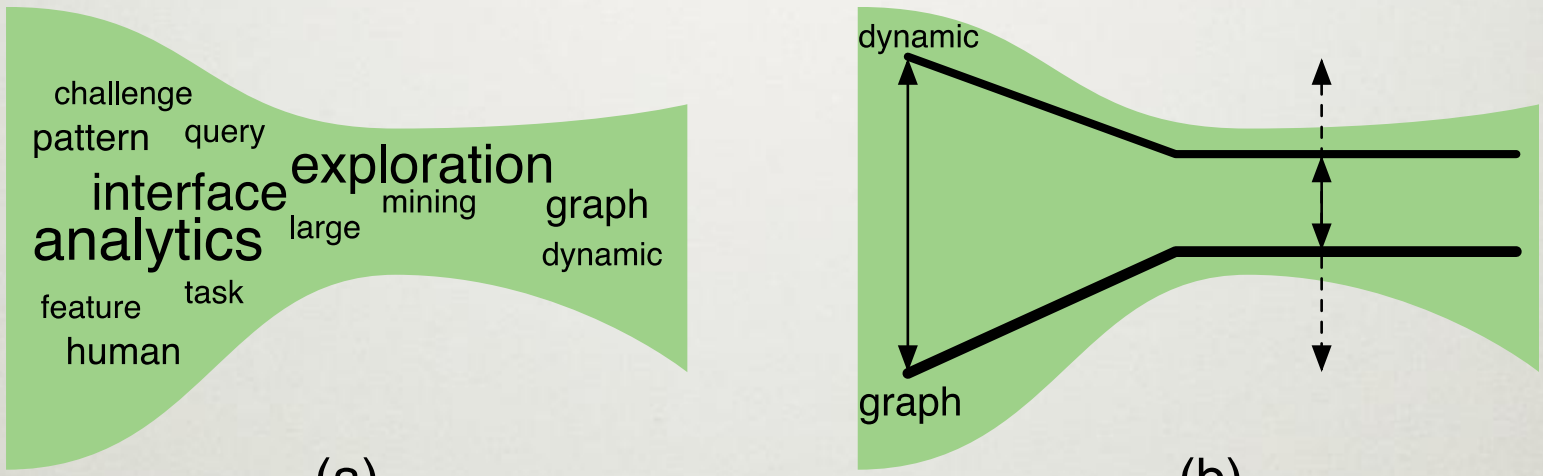


Critical Event as Glyph

- **Emerge, dissolve, split, and merge**

Keyword Correlation as Thread

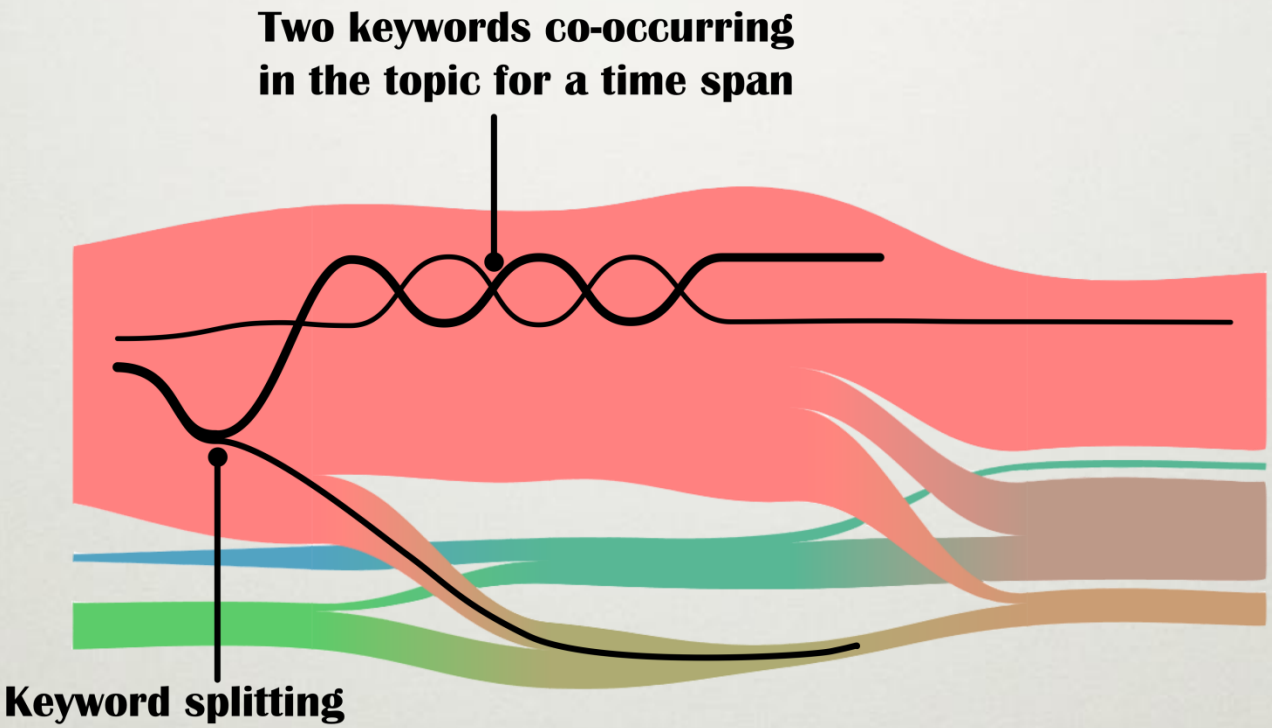
- **Alternatives**



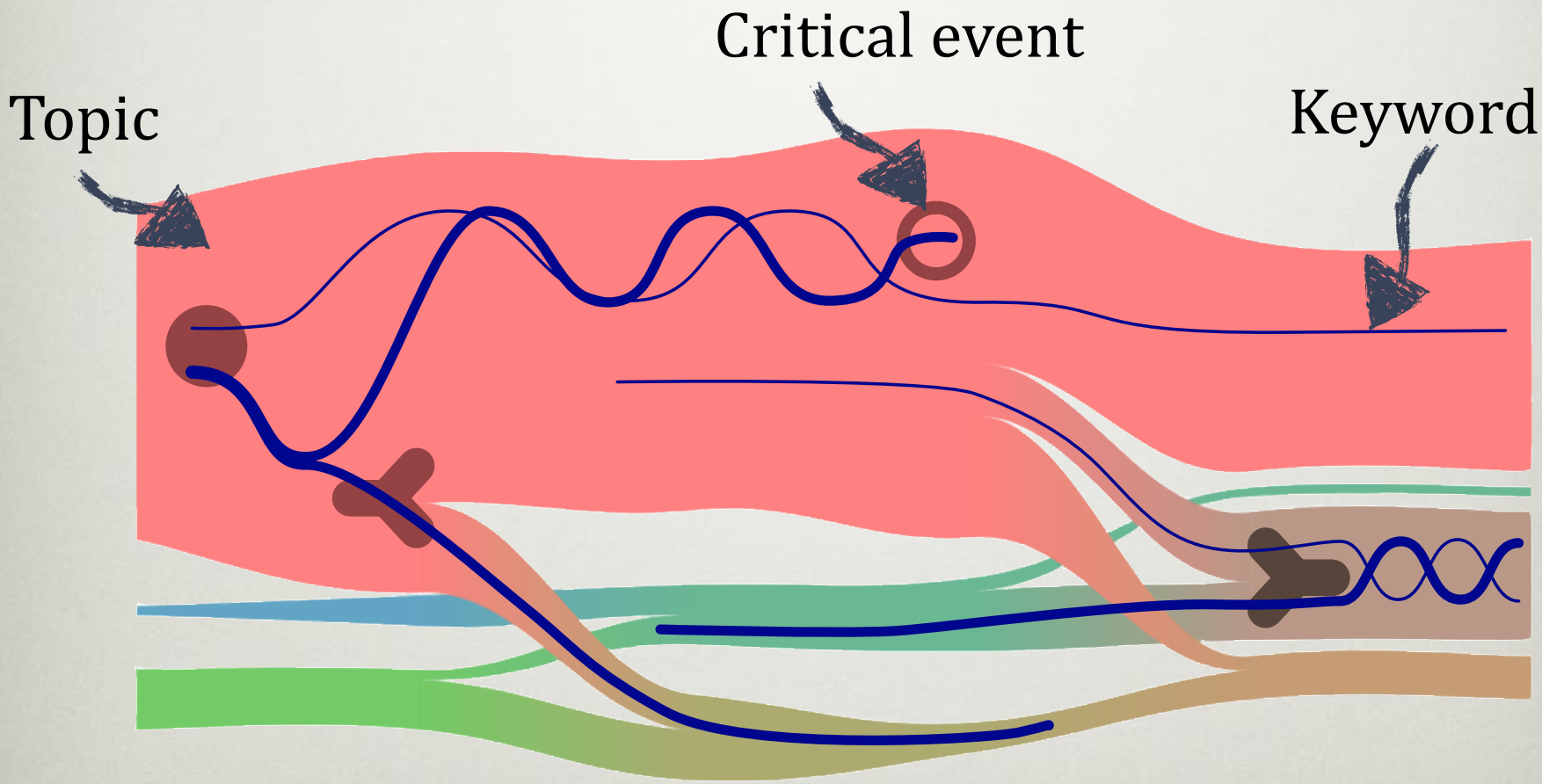
(a) (b)

Keyword Correlation as Thread

- Intertwine to indicate co-occurrences

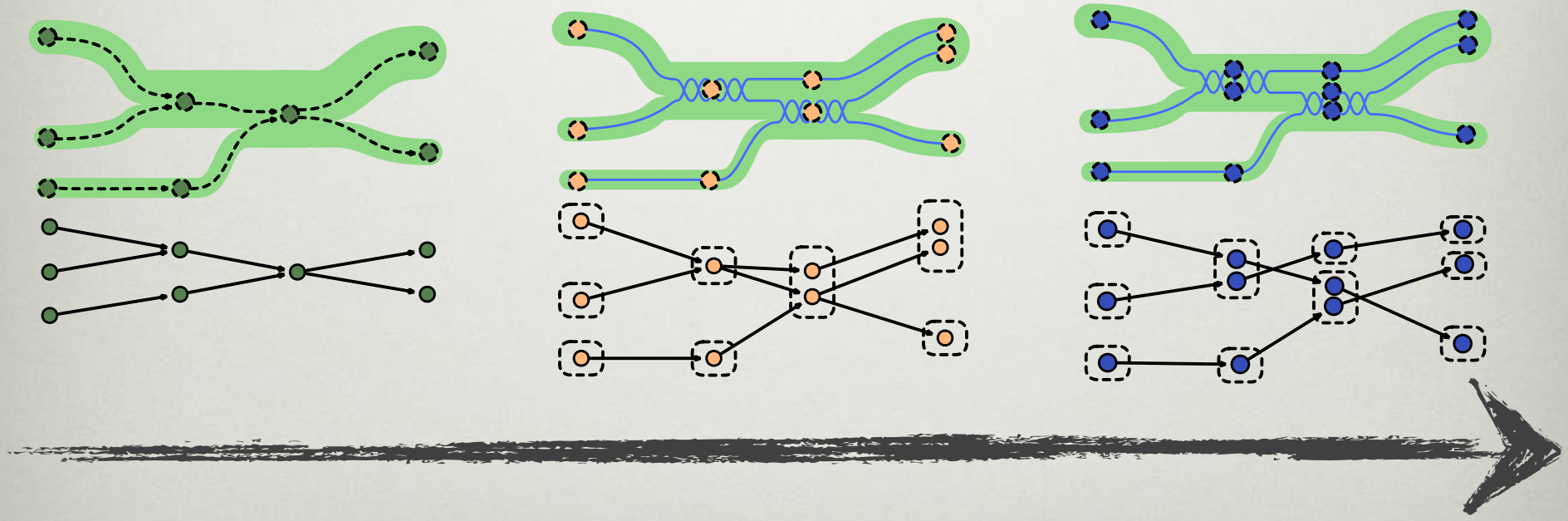


Visualization Design - Consistency



Layout Algorithm

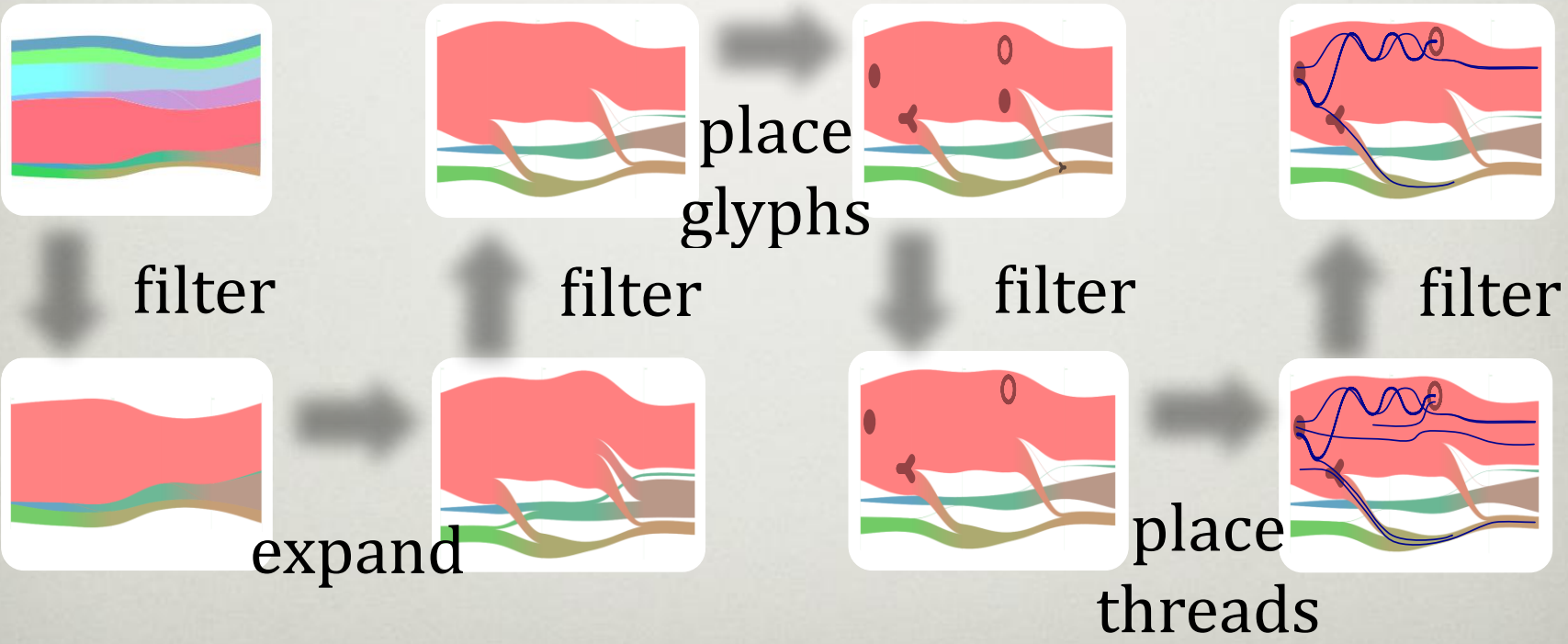
- **Three-level directed acyclic graph (DAG)**



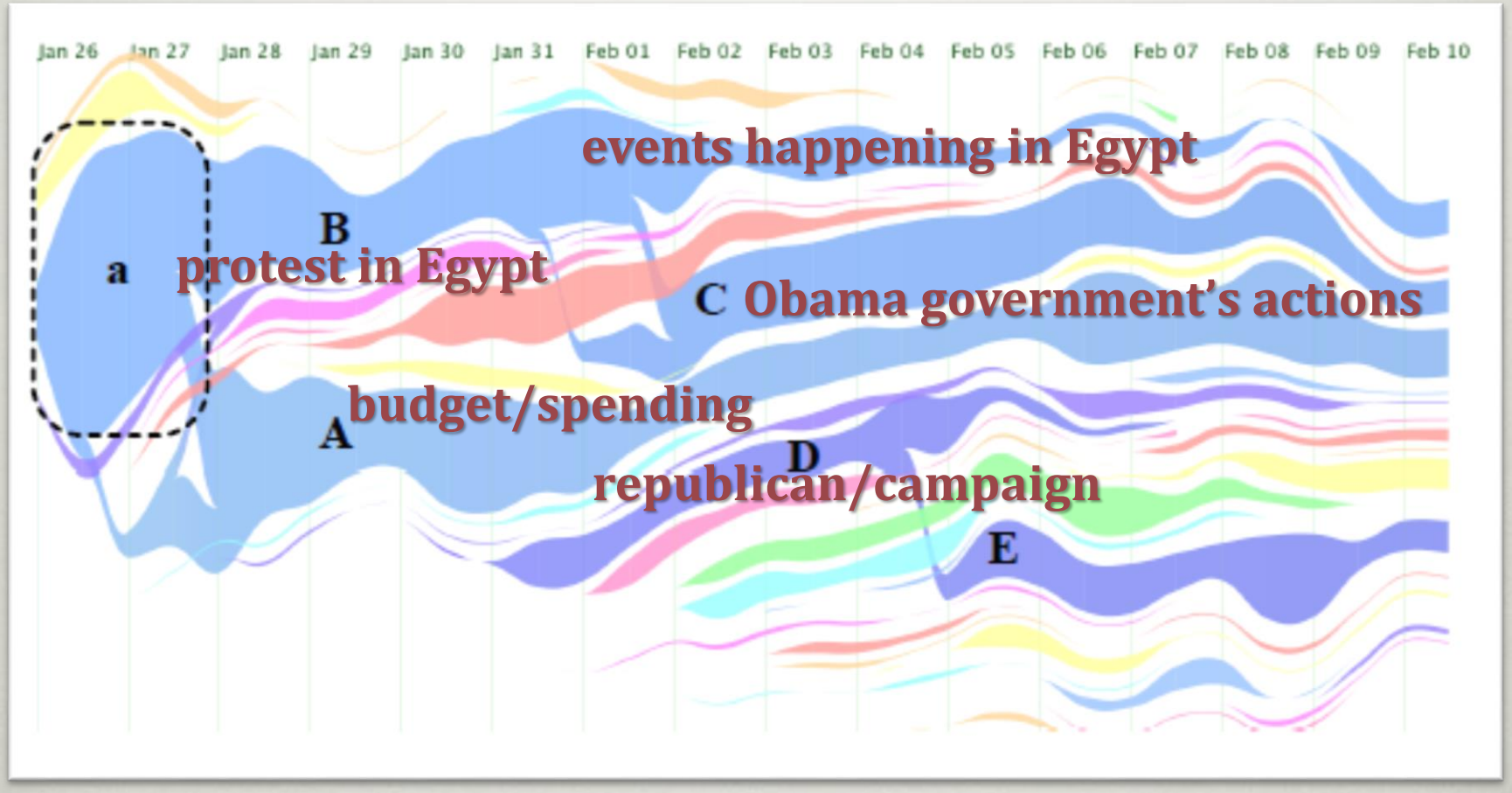
Interactive Exploration

Overview first, zoom and filter, ...

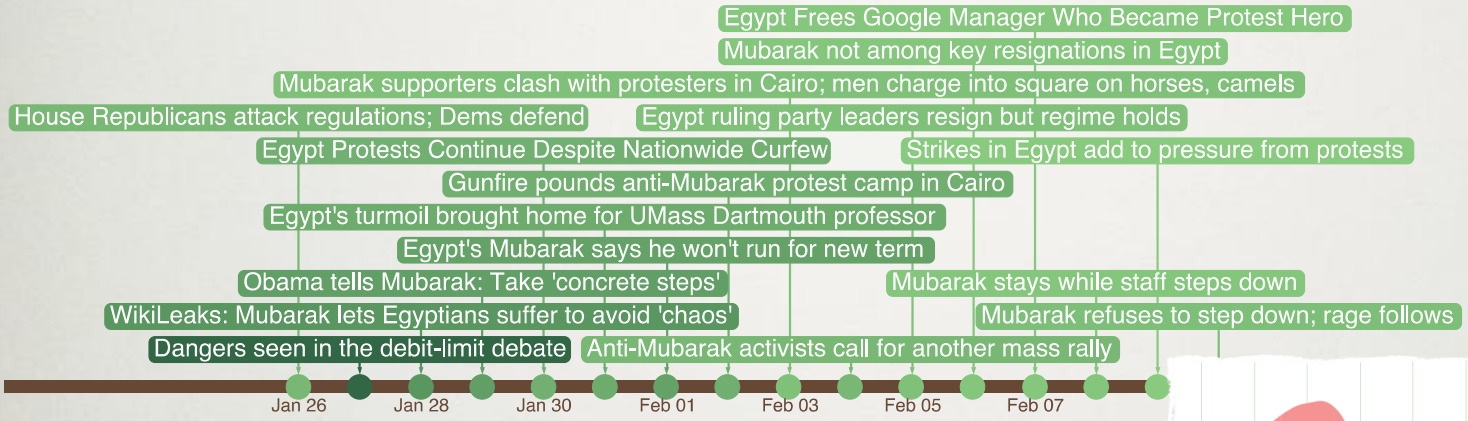
(topic ➡ critical event ➡ keyword)



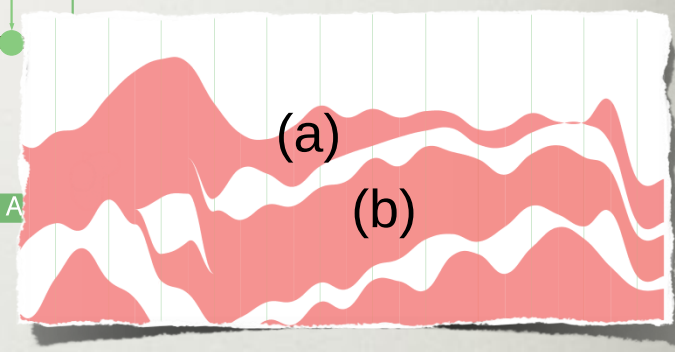
Application Example: Bing News

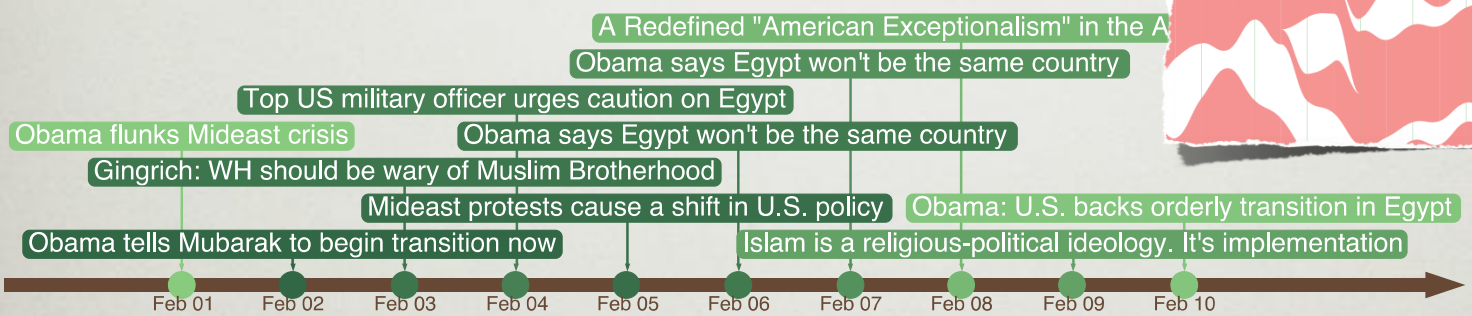


Application Example: Bing News



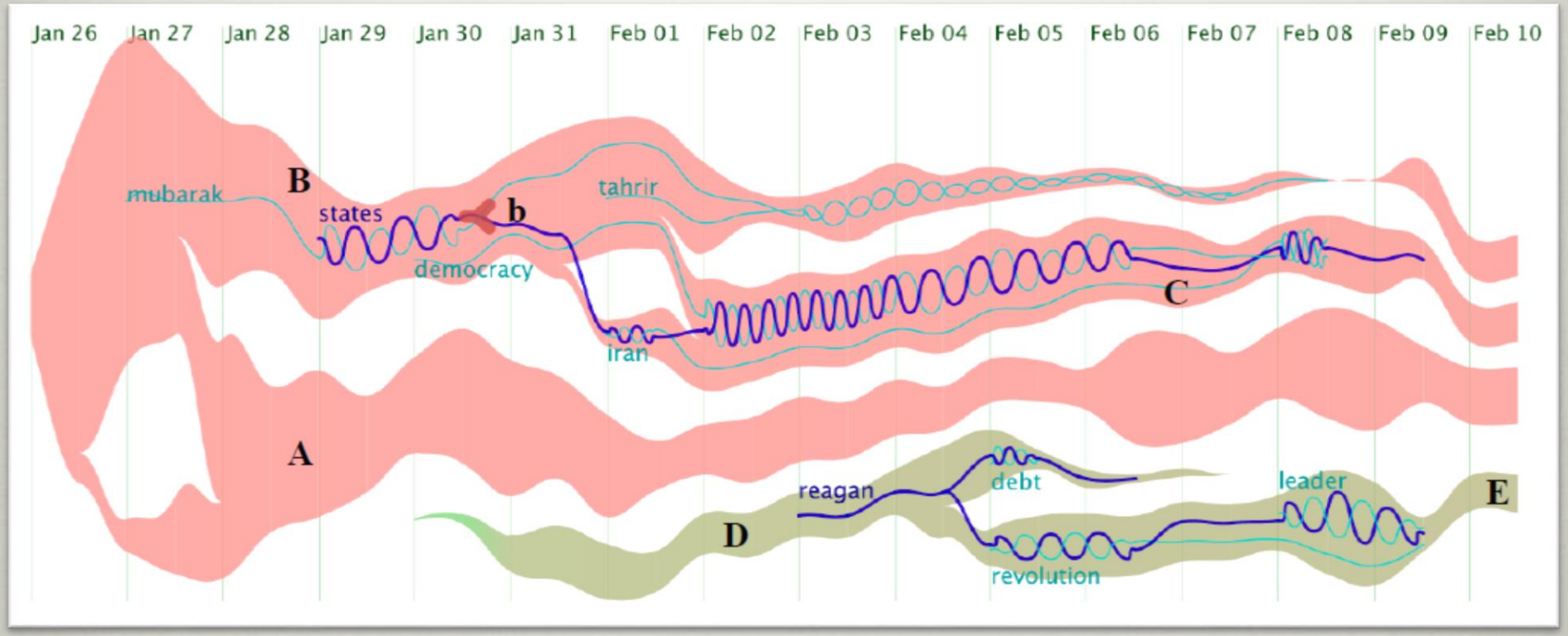
(a)



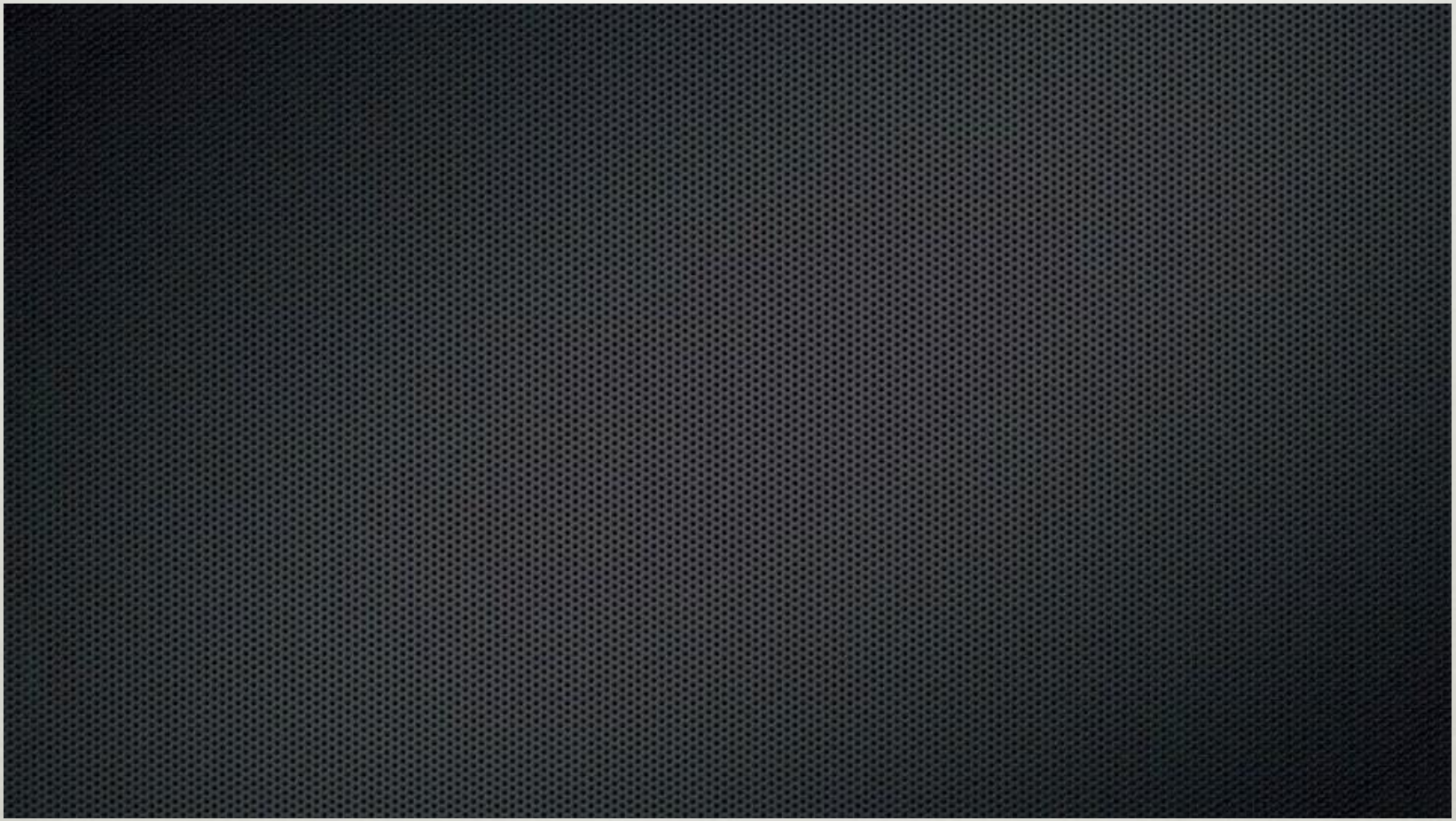


(b)

Application Example: Bing News



Video



Outline

- **Example tasks in text analytics**
- **Visually analyzing textual information**
  - Dynamic word clouds
  - Topic-based visual text summarization
  - TextFlow: towards better understanding of evolving topics in text
- **Future work**

Future Text Visualization Topics

- **Interactive, incremental** text analytics
- **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

Weiwei Cui, Yangqiu Song, Furu Wei, Xin Tong (MSRA)
Nan Cao, Yingcai Wu, Prof. Huamin Qu (HKUST)
Dr. Michelle X Zhou (IBM Almaden Research Center)