文本数据可视化

大数据可视分析导论

陈晴
https://idvxlab.com
同济大学

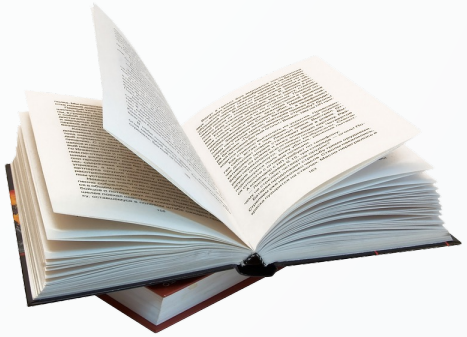
智能大数据可视化实验室

课程大纲

- 什么是文本数据？
- 为什么要对文本数据可视化？
- 文本数据分析技术
- 文本数据可视化技术

文本数据无处不在

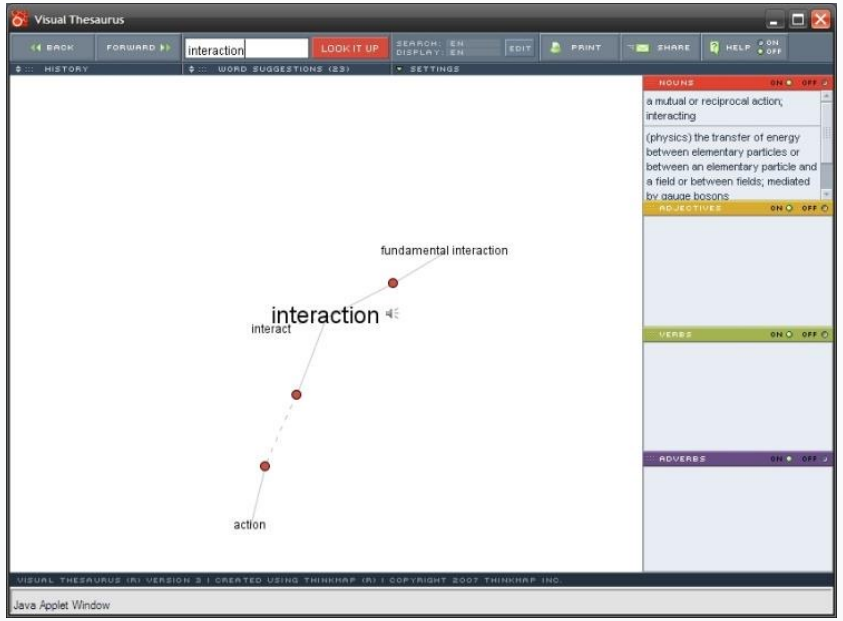
- 文章、文献、书籍
- 新闻媒体
- 邮件、网页、博客
- 社交网络评论、标签
- 计算机编程语言
- .....



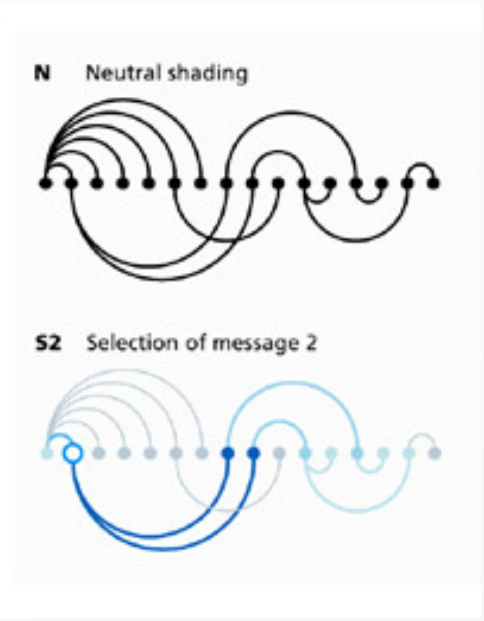
为什么要对文本数据可视化？

- 协助信息检索与理解
- 支持语义情感的分析
- 增强对文本数据之间关联的分析

Themescape



Visual Thesaurus



Thread Arcs

文本数据可视化应用领域

电子商务

社会计算

商务智能

用户体验

预测分析

公共关系

文本可视分析任务

- 文本数据内容的可视化
  - 总结文本内容
  - 提取文本要义
  - 展现文本所包含的情感
  - 辅助大规模文本数据集的浏览
- 文本数据关系的可视化
  - 挖掘文本结构
  - 分析文本之间的关联
  - 展现文本语义等内在信息的联系

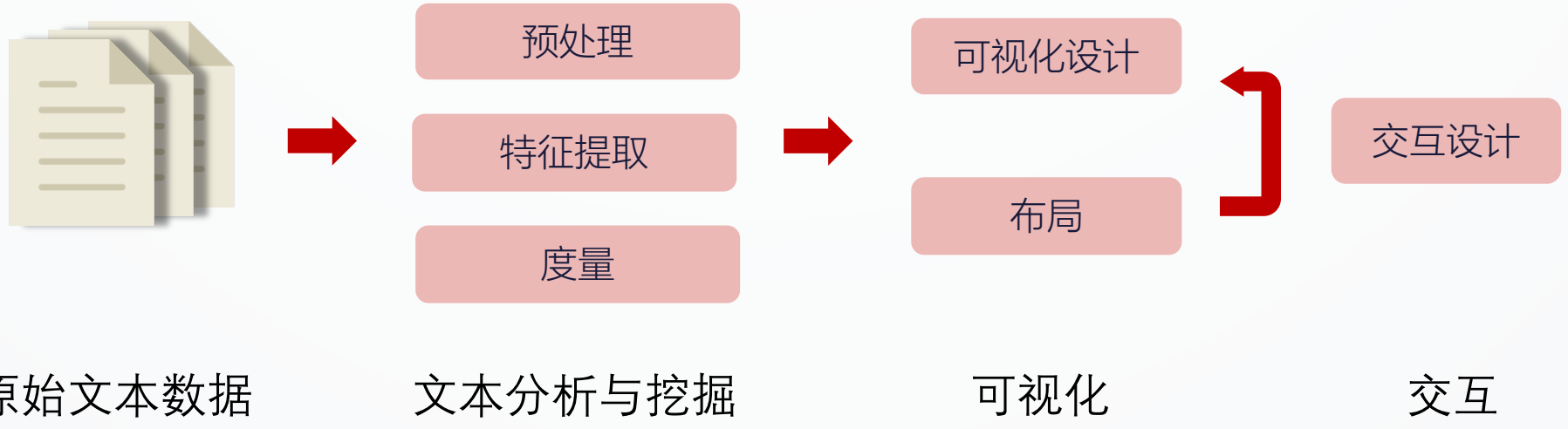
文本可视分析任务

- 文本数据内容的可视化
  - 总结文本内容
  - 提取文本要义
  - 展现文本所包含的情感
  - 辅助大规模文本数据集的浏览
- 文本数据关系的可视化
  - 挖掘文本结构
  - 分析文本之间的关联
  - 展现文本语义等内在信息的联系

文本可视分析任务

- 文本数据内容的可视化
  - 总结文本内容
  - 提取文本要义
  - 展现文本所包含的情感
  - 辅助大规模文本数据集的浏览
- 文本数据关系的可视化
  - 挖掘文本结构
  - 分析文本之间的关联
  - 展现文本中内在信息的联系

文本可视分析流程



文本可视分析流程

- **文本分析与挖掘**
  - 预处理 - 数据整理
  - 特征提取 - 关键词、词频、主题、情感
  - 度量 - 相似性

- **文本可视化**
  - 可视化设计 / 布局考量
  - 分析任务的支持
  - 交互式探索

文本数据分析技术

- **分词 Tokenization**

"I have a dream that one day this nation will rise up and live out the true meaning of its creed: We hold these truths to be self-evident, that all men are created equal."

- 移除停顿词: a, the, that, etc.
- 单复数转换: men->man, truths->truth

I, dream, one, day, nation, rise, up, live, out, true, meaning, creed, hold, truth, be, self-evident, all, man, created, equal

文本数据分析技术

- **向量空间模型Vector-space Model**
  - Bag-of-words

Word	I	dream	color	skin	nation	slave	injustice	owner
Frequency	4	4	I	I	2	2	I	I

    - 计算文档相似性：比较两个词向量之间的相似性
      - Cosine: $\cos\theta = \frac{\mathbf{v}_1\mathbf{v}_2^{\mathrm{T}}}{\|\mathbf{v}_1\|\cdot\|\mathbf{v}_2\|}$
      - TF-IDF

文本数据分析技术

- **TF-IDF (term frequency–inverse document frequency)**

$$词频(TF) = \frac{某个词在文章中的出现次数}{文章的总词数}$$

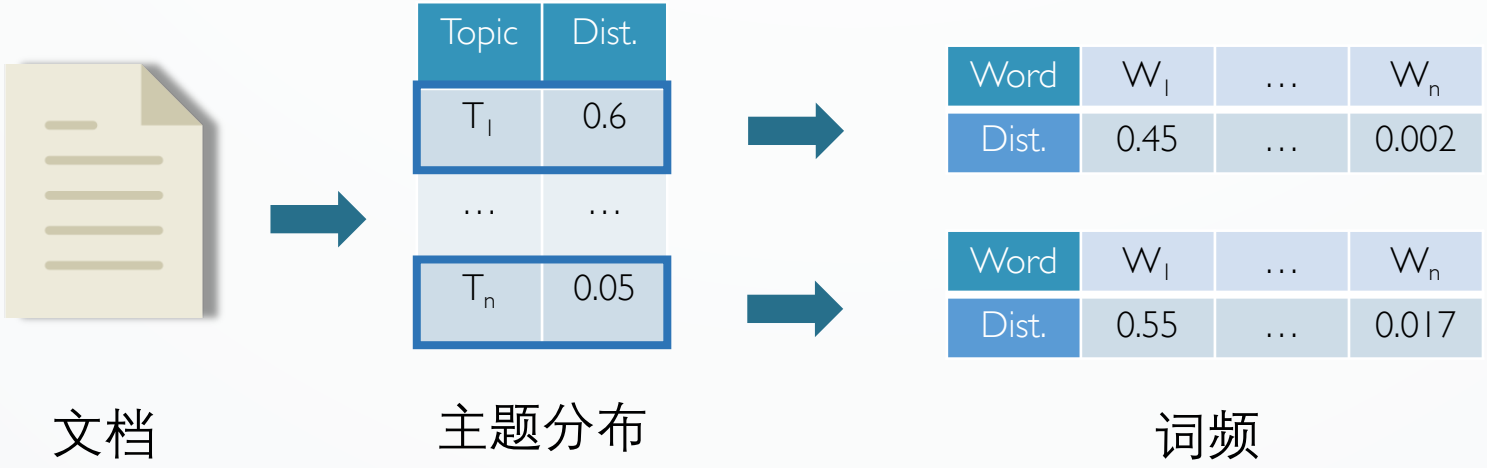
$$逆文档频率(IDF) = \log(\frac{语料库的文档总数}{包含该词的文档数 + 1})$$

	包含该词的文档数（亿）	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

TF-IDF 计算词在文档中的出现次数成正比，
与该词在整个语言中的出现次数成反比。

文本数据分析技术

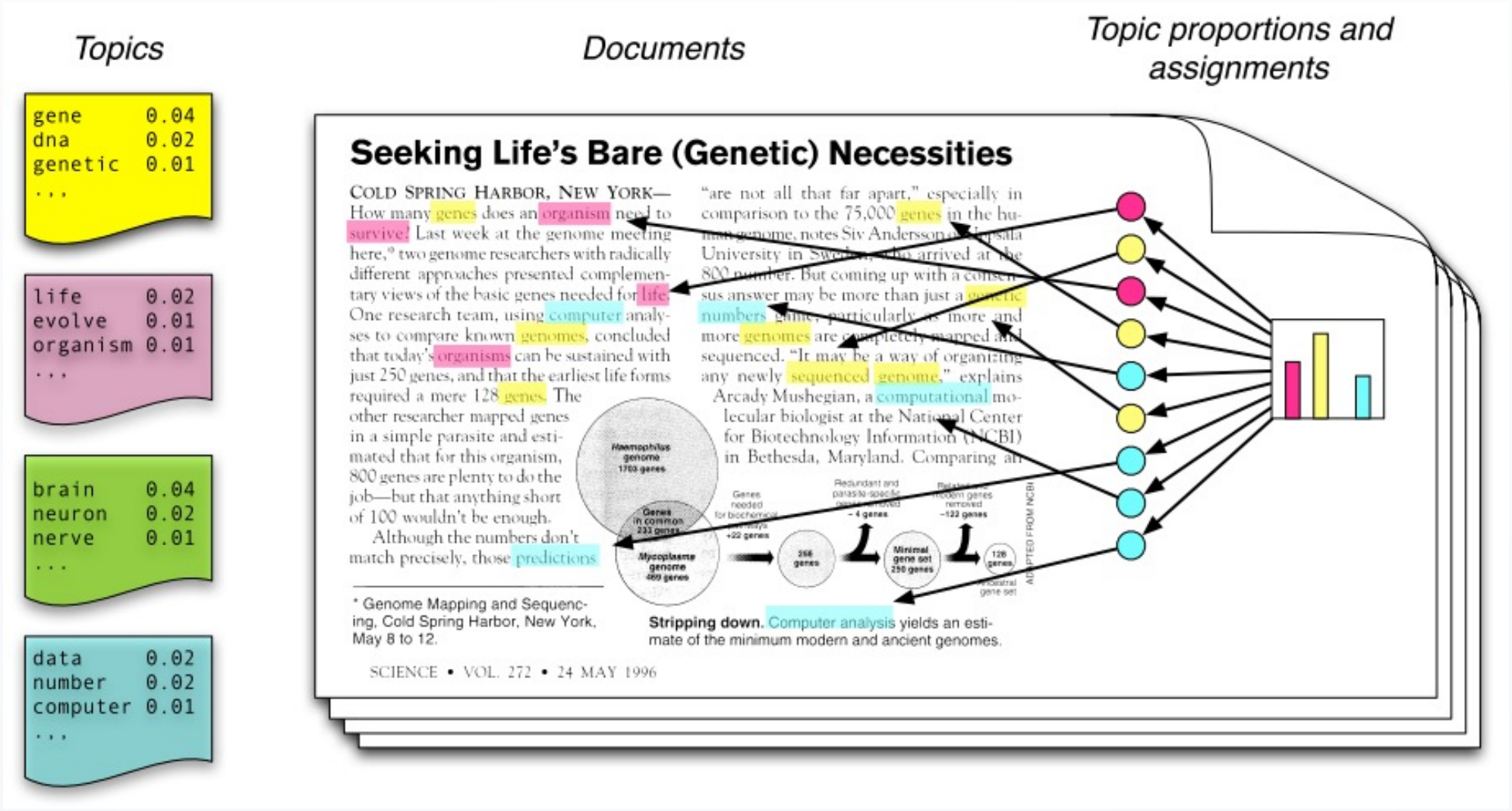
- **主题模型**



方法：

- Latent Semantic Indexing
- pLSI
- LDA Latent Dirichlet Allocation

文本数据分析技术


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

文本可视分析流程

- **文本分析与挖掘**
  - 预处理 - 数据整理
  - 特征提取 - 关键词、词频、主题、情感
  - 度量 - 相似性

- **文本可视化**
  - 可视化设计 / 布局考量
  - 分析任务的支持
  - 交互式探索

文本数据可视化技术

- **根据分析任务特征分类**
  - 可视化文本数据内容
  - 可视化文本数据关联
  - 多层级文本数据可视化

文本数据可视化技术

- **根据分析任务特征分类**
  - 可视化文本数据内容
    - 基于关键词的可视化
    - 时序文档的可视化
    - 基于特征分布的可视化
  - 可视化文本数据关联
  - 多层级文本数据可视化

文本数据可视化技术

- **根据分析任务特征分类**
  - 可视化文本数据内容
    - 基于关键词的可视化
    - 时序文档的可视化
    - 基于特征分布的可视化
  - 可视化文本数据关联
  - 多层级文本数据可视化

基于关键词的文本内容可视化技术

- **词云 word cloud/tag cloud/text cloud**
- 计算关键词 – 频率 – 大小/颜色

Tag cloud of “I have a dream”.

基于关键词的文本内容可视化技术

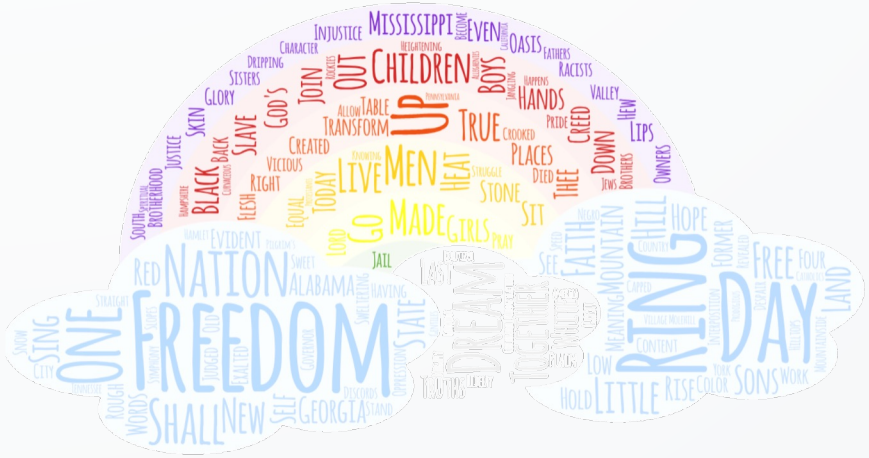
- **词云衍生 wordle**
  - 支持更加复杂的样式
  - 可以控制形状
  - 紧致排布

Wordle of “I have a dream”.

Created by: https://wordart.com/create

基于关键词的文本内容可视化技术

- **词云衍生 wordle**
  - 可以控制形状
  - 紧致排布

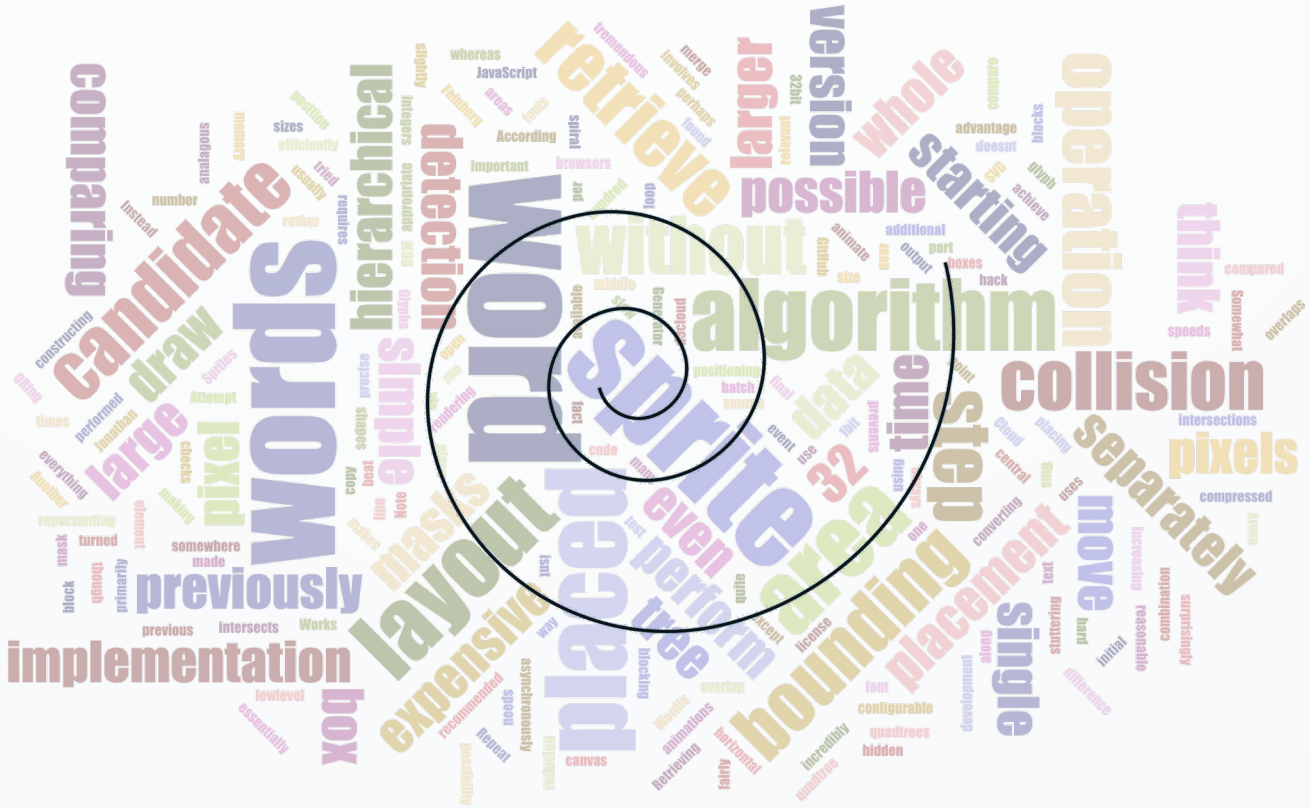


Wordle of “I have a dream”.

Created by: https://wordart.com/create

基于关键词的文本内容可视化技术

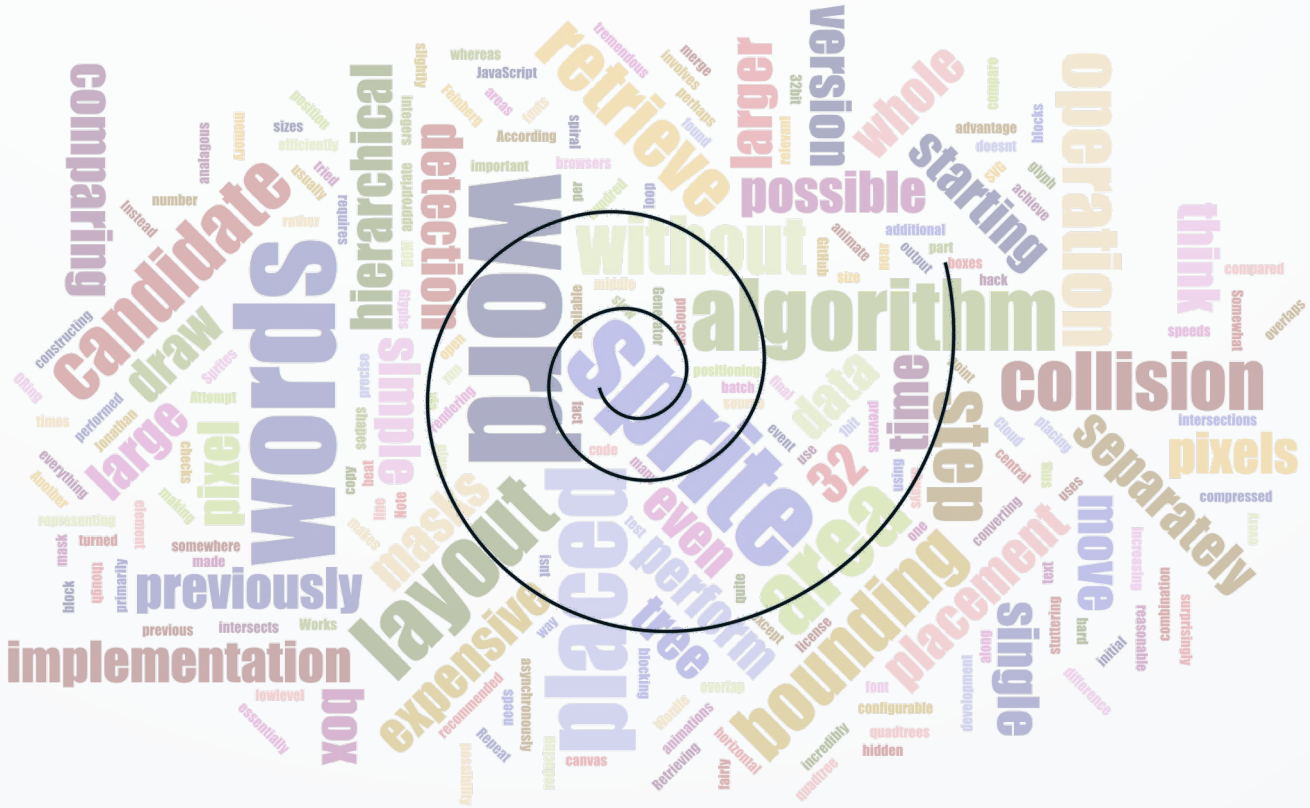
- **词云衍生 wordle**
  - 布局：螺旋



Created by: https://www.jasondavies.com/wordcloud/

基于关键词的文本内容可视化技术

- **词云衍生 wordle**
  - 布局：螺旋



Created by: https://www.jasondavies.com/wordcloud/

基于关键词的文本内容可视化技术

- **词云衍生 wordle**
  - 布局：水平/垂直

Created by: https://www.jasondavies.com/wordcloud/

基于关键词的文本内容可视化技术

- **词云衍生 wordle**
  - 布局：随机

Created by: https://www.jasondavies.com/wordcloud/

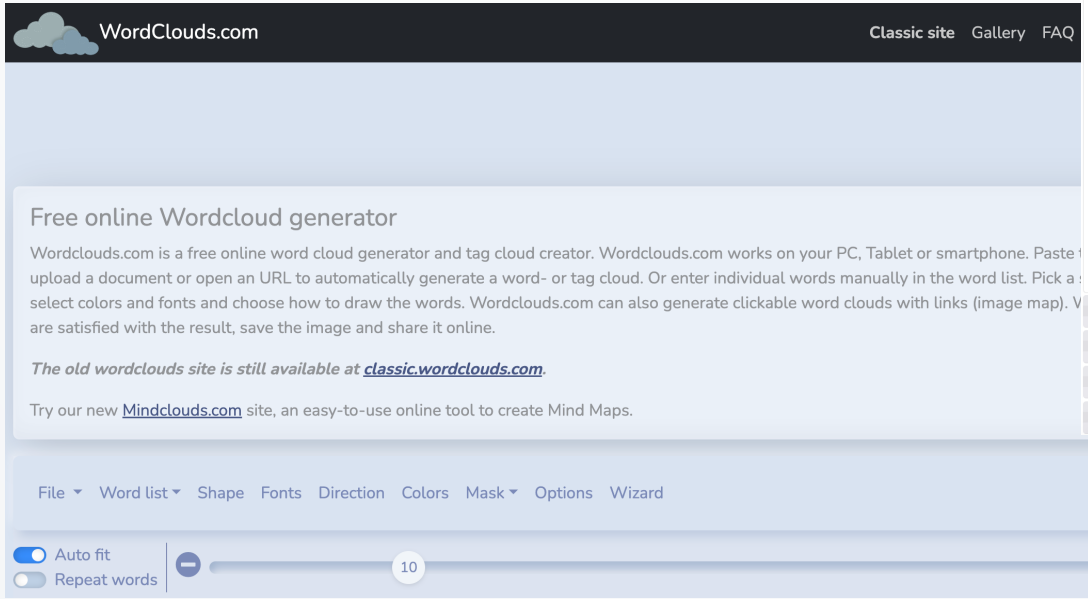
基于关键词的文本内容可视化技术

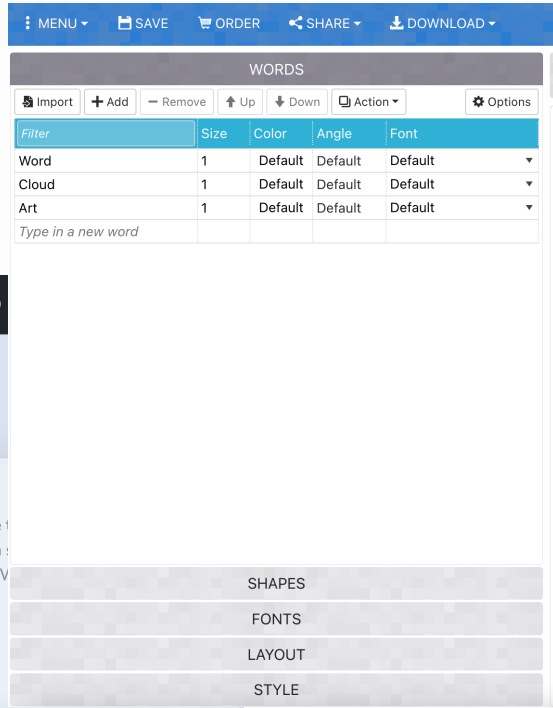
- **词云衍生 wordle**
  - 布局：形状

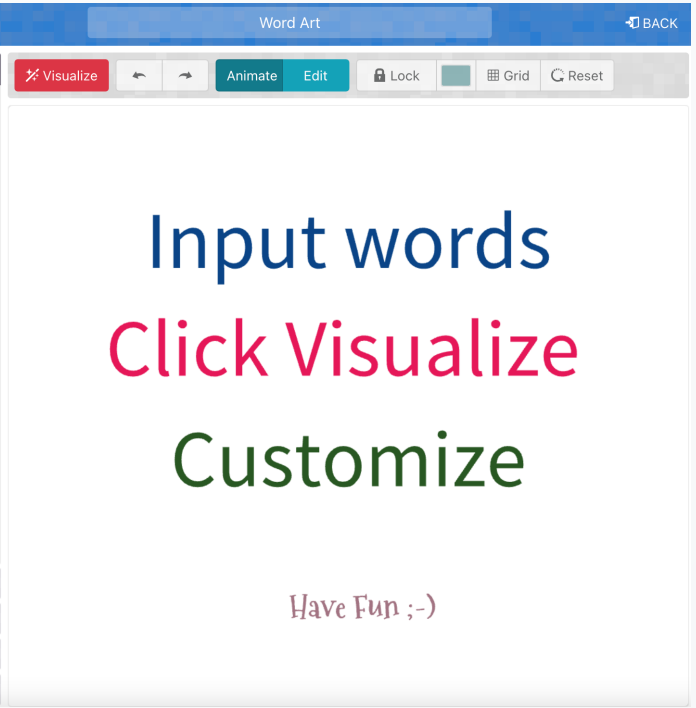
Created by: https://www.jasondavies.com/wordcloud/

基于关键词的文本内容可视化技术

- 练习：动手试一试
- https://www.wordclouds.com/
- https://wordart.com/create





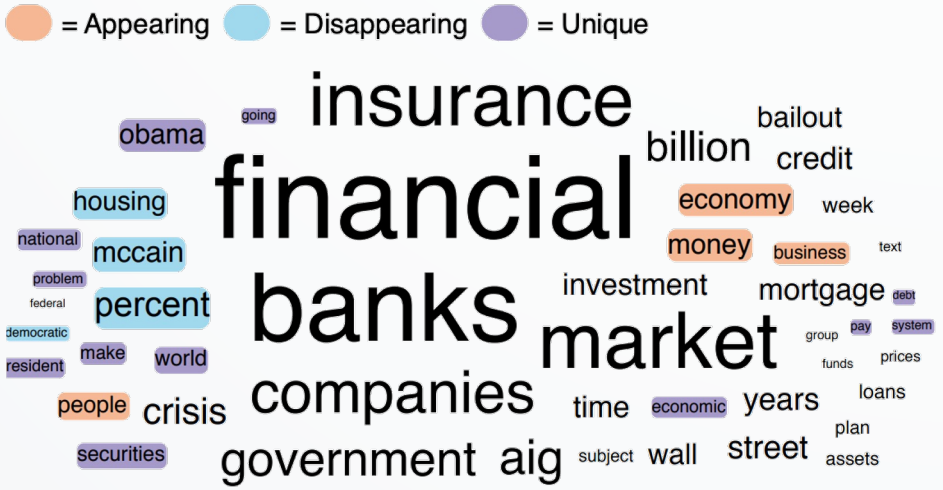


基于关键词的文本内容可视化技术

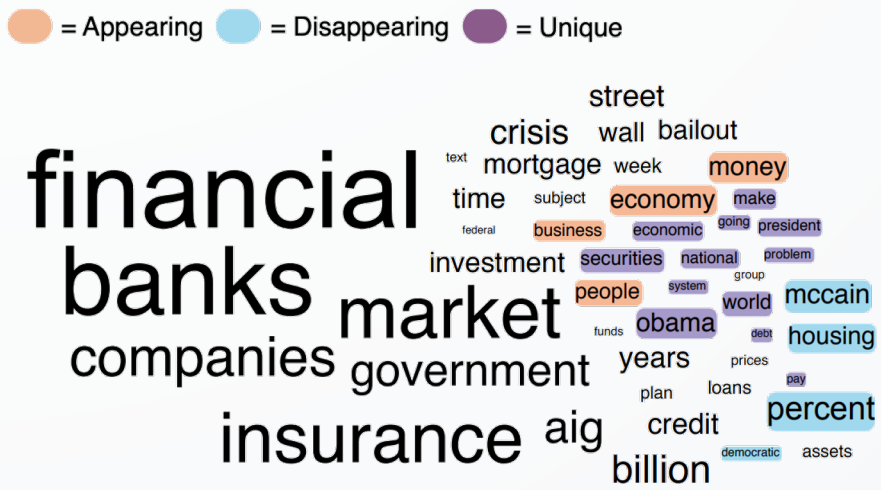
- **词云 wordle**
  - 有什么问题？

基于关键词的文本内容可视化技术

- **词云 wordle**
  - 上下文保持 Context-preserving Word Cloud



The importance criterion



The co-occurrence criterion

Weiwei Cui et al. Context preserving dynamic word cloud visualization. (IEEE PacificVis 2010)

基于关键词的文本内容可视化技术

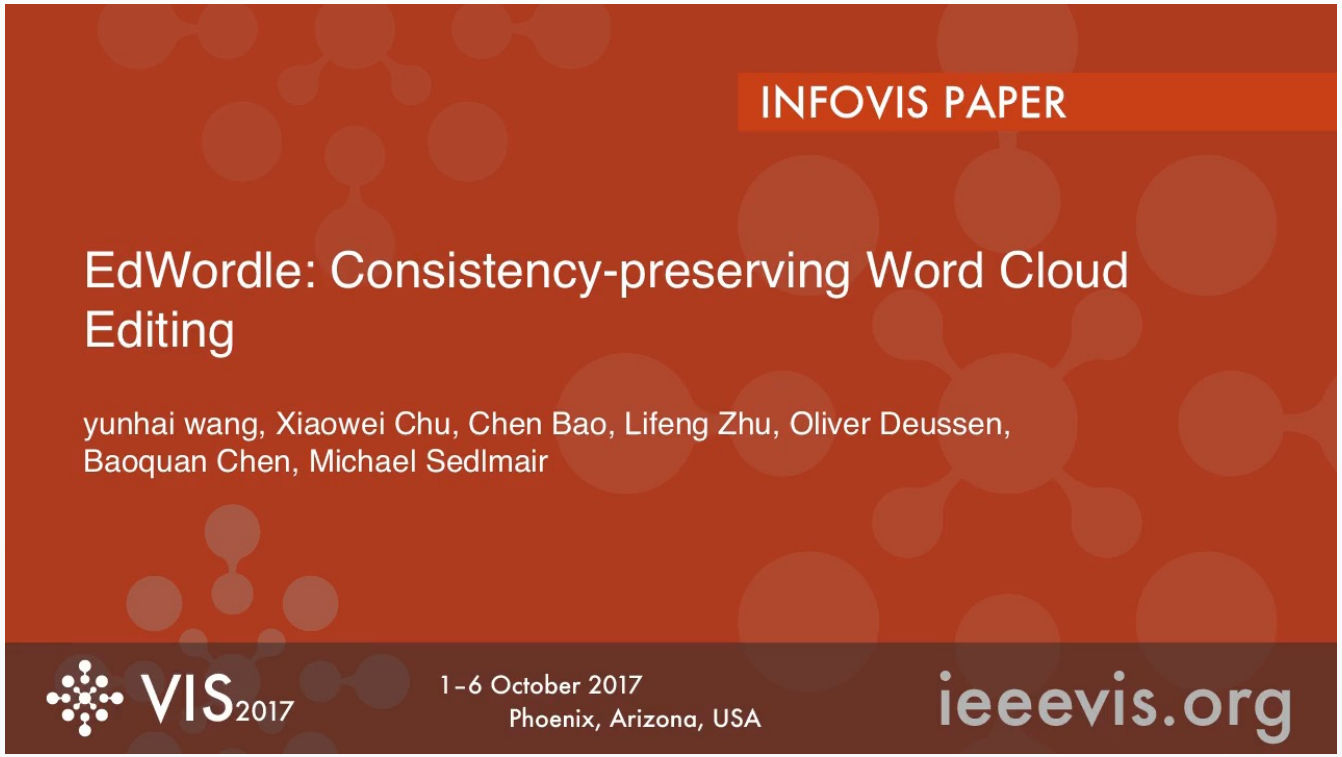
- **词云 wordle**
  - 连续性保持 Consistency-preserving Word Cloud

Word cloud of five articles

Yunhai Wang et al. EdWordle: Consistency-preserving Word Cloud Editing. (IEEE TVCG 2018)

基于关键词的文本内容可视化技术

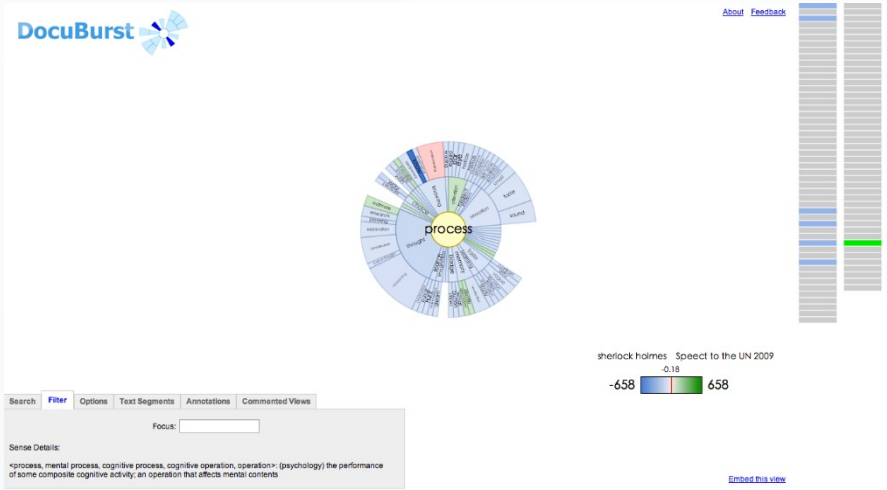
- **词云 wordle**
  - 连续性保持 Consistency-preserving Word Cloud

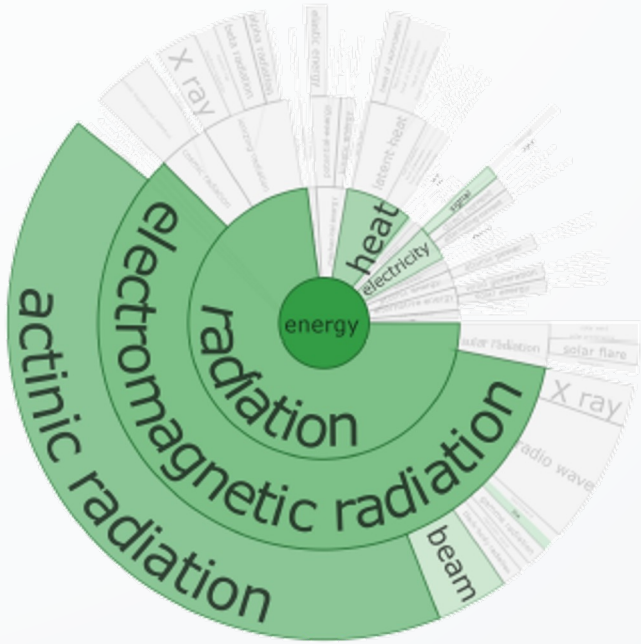


Yunhai Wang et al. EdWordle: Consistency-preserving Word Cloud Editing. (IEEE TVCG 2018)

基于关键词的文本内容可视化技术

- **DocuBurst**
  - 基于关键字的方法。
  - 放射状布局：外圈中的词是内圈中的词的下位词。
  - 树状结构

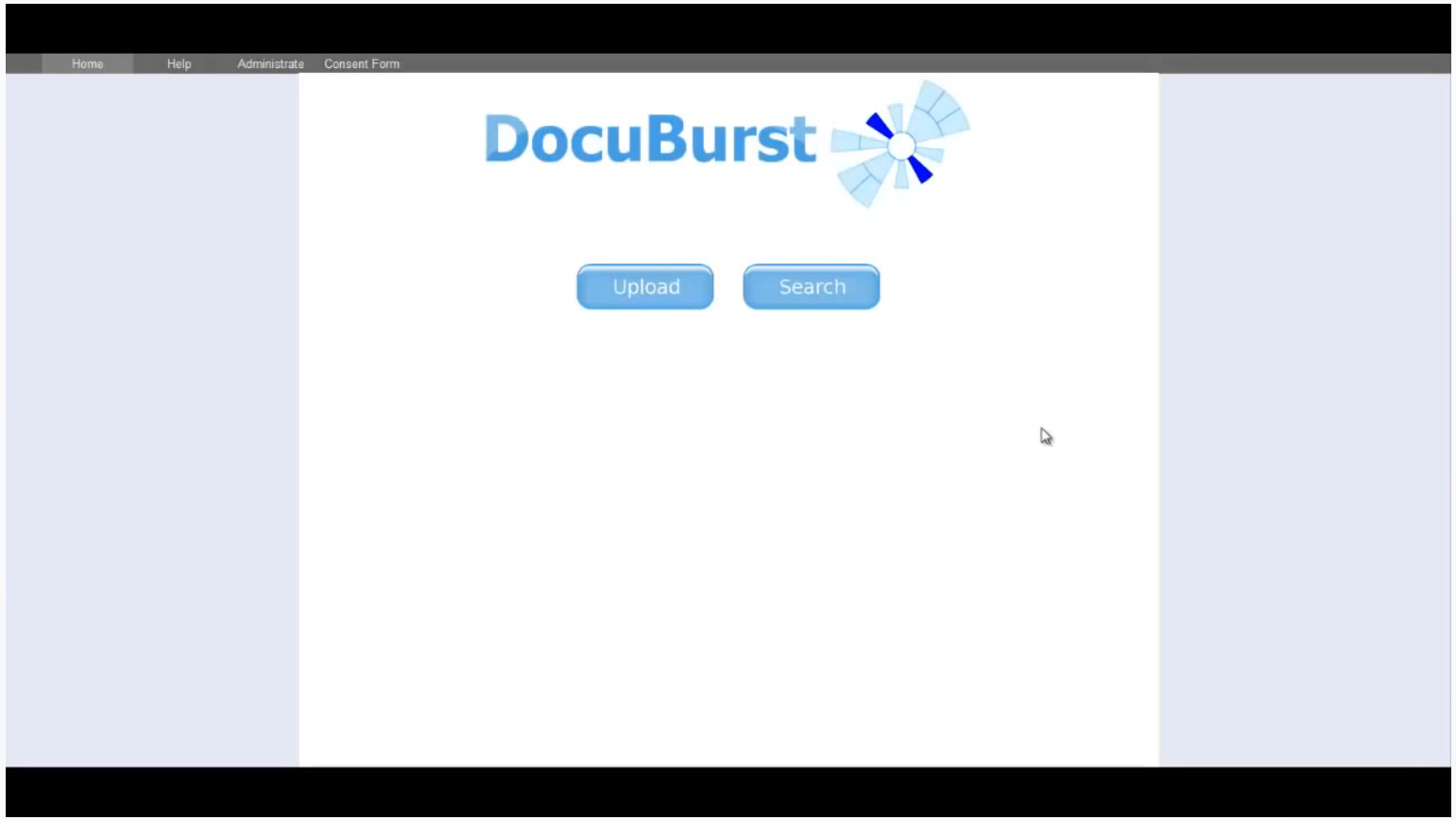




Christopher Collins et al. DocuBurst:Visualizing Document Content using Language Structure. (IEEE CGF 2009)

基于关键词的文本内容可视化技术

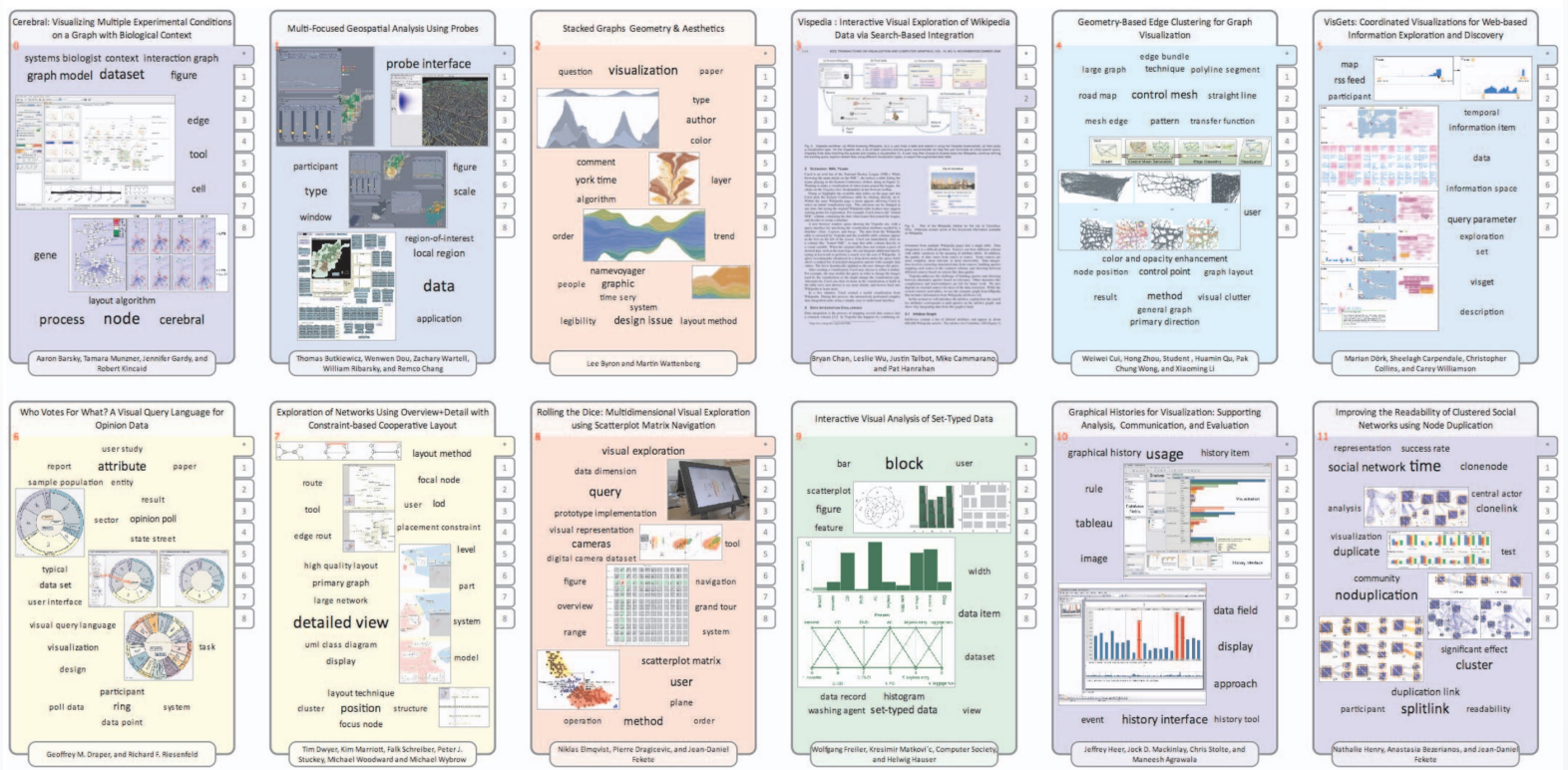
- **DocuBurst**



Christopher Collins et al. DocuBurst:Visualizing Document Content using Language Structure. (IEEE CGF 2009)

基于关键词的文本内容可视化技术

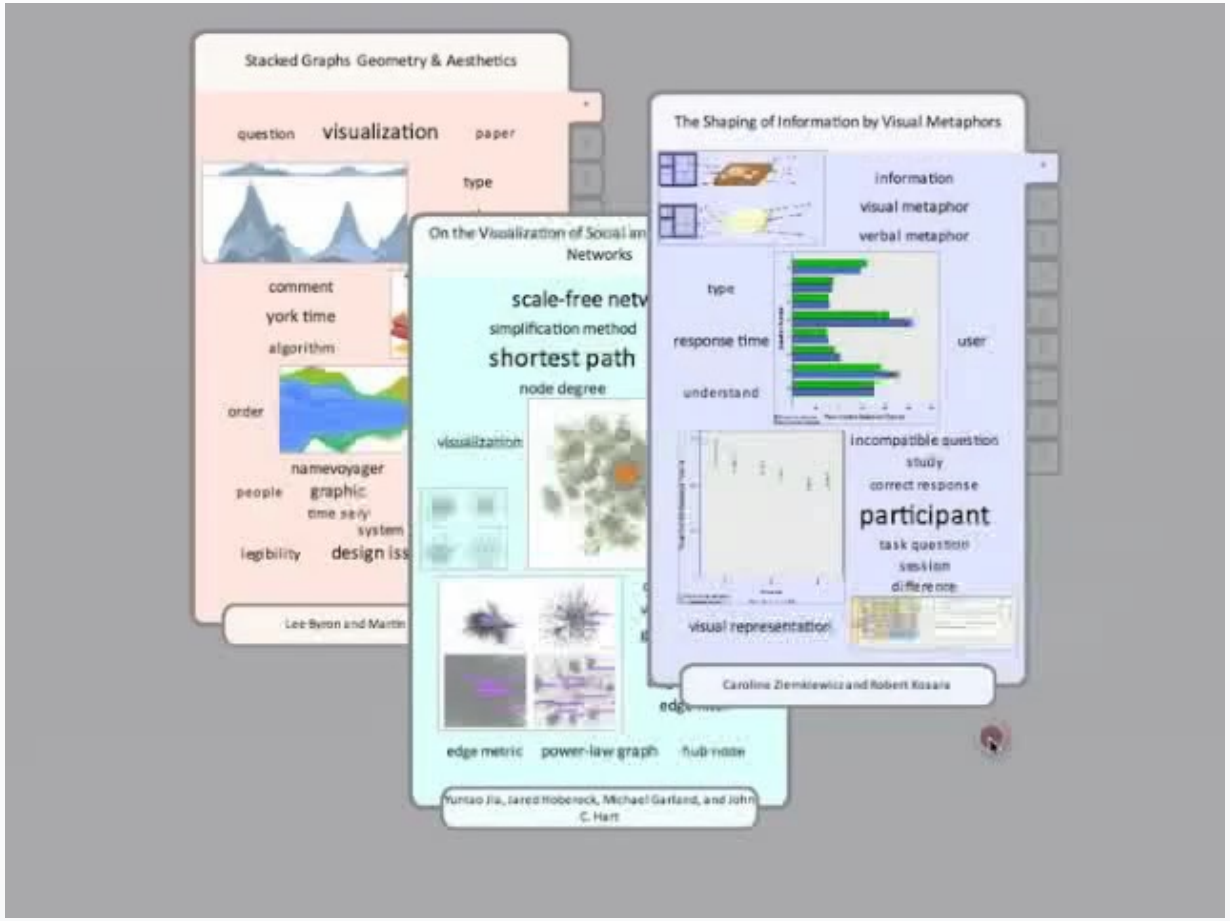
- **文档卡片 Document Card**
  - 使用关键字和图像总结文档
  - 像卡片一样布局



Hendrik Strobelt et al. Document Cards: A Top Trumps Visualization for Documents. (IEEE TVCG 2009)

基于关键词的文本内容可视化技术

- **文档卡片 Document Card**
  - 使用关键字和图像总结文档
  - 像卡片一样布局



Hendrik Strobelt et al. Document Cards: A Top Trumps Visualization for Documents. (IEEE TVCG 2009)

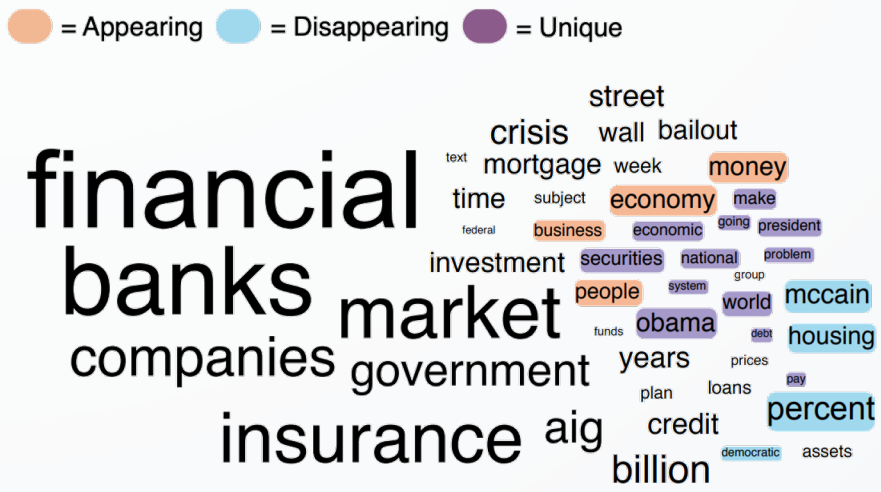
文本数据可视化技术

- **根据特征分类**
  - 可视化文本数据内容
    - 基于关键词的可视化
    - 时序文档的可视化
    - 基于特征分布的可视化
  - 可视化文本数据关联
  - 多层级文本数据可视化

时序文档的文本内容可视化技术

- **词云 wordle**
  - 上下文保持 Context-preserving Word Cloud

The importance criterion

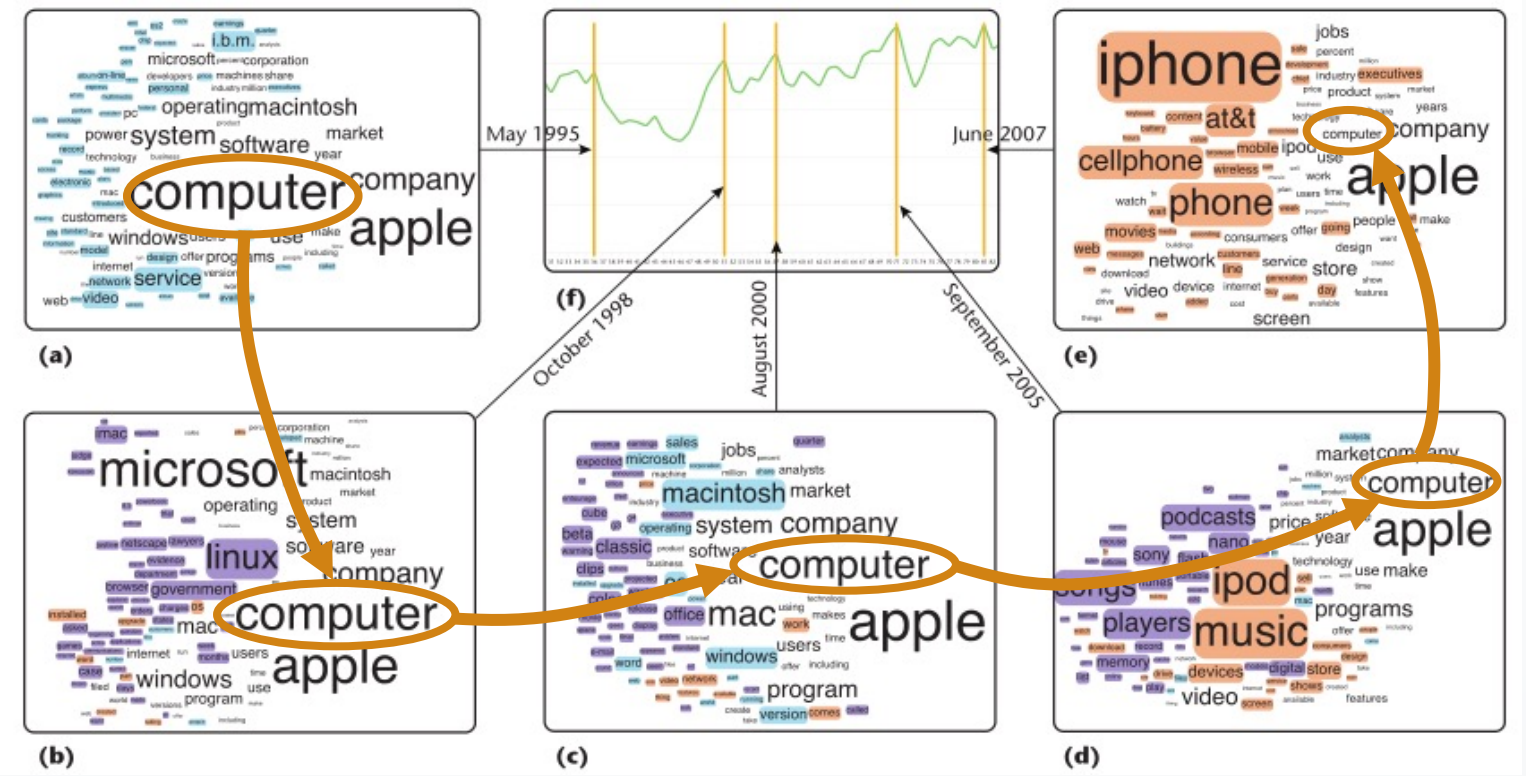


The co-occurrence criterion

Weiwei Cui et al. Context preserving dynamic word cloud visualization. (IEEE PacificVis 2010)

时序文档的文本内容可视化技术

- **词云 wordle**
  - 上下文保持 Context-preserving Word Cloud



Weiwei Cui et al. Context preserving dynamic word cloud visualization. (IEEE PacificVis 2010)

时序文档的文本内容可视化技术

- **时变词云 Morphable Word Cloud**

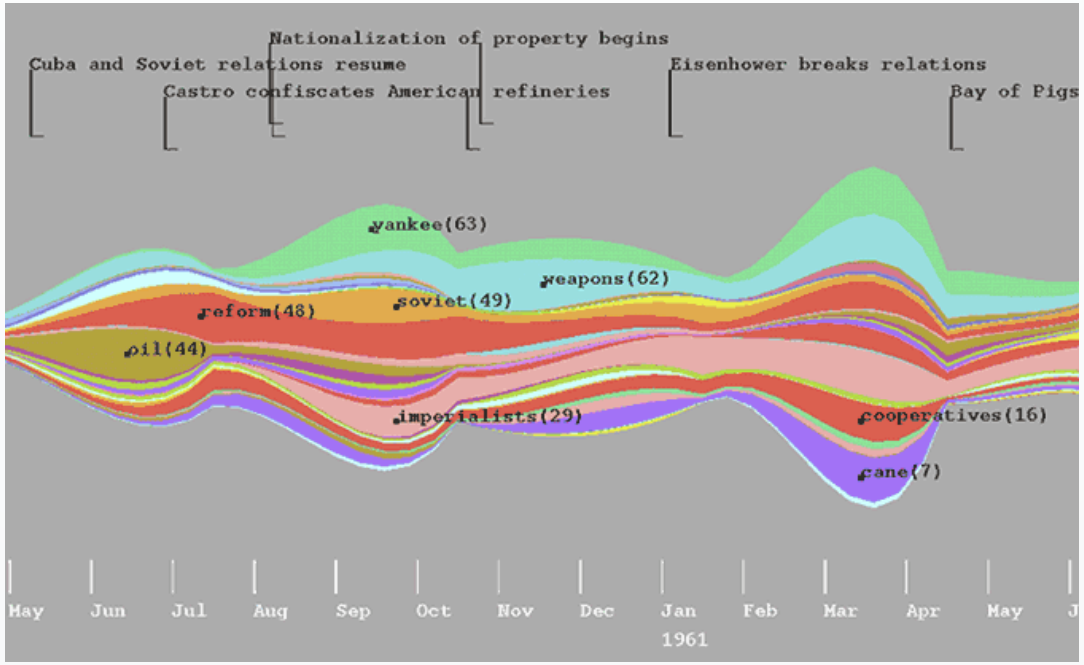
人类进化过程

青蛙的生命周期

Ming-Te Chi et al. Morphable Word Clouds for Time-varying Text Data Visualization. (IEEE TVCG 2015)

时序文档的文本内容可视化技术

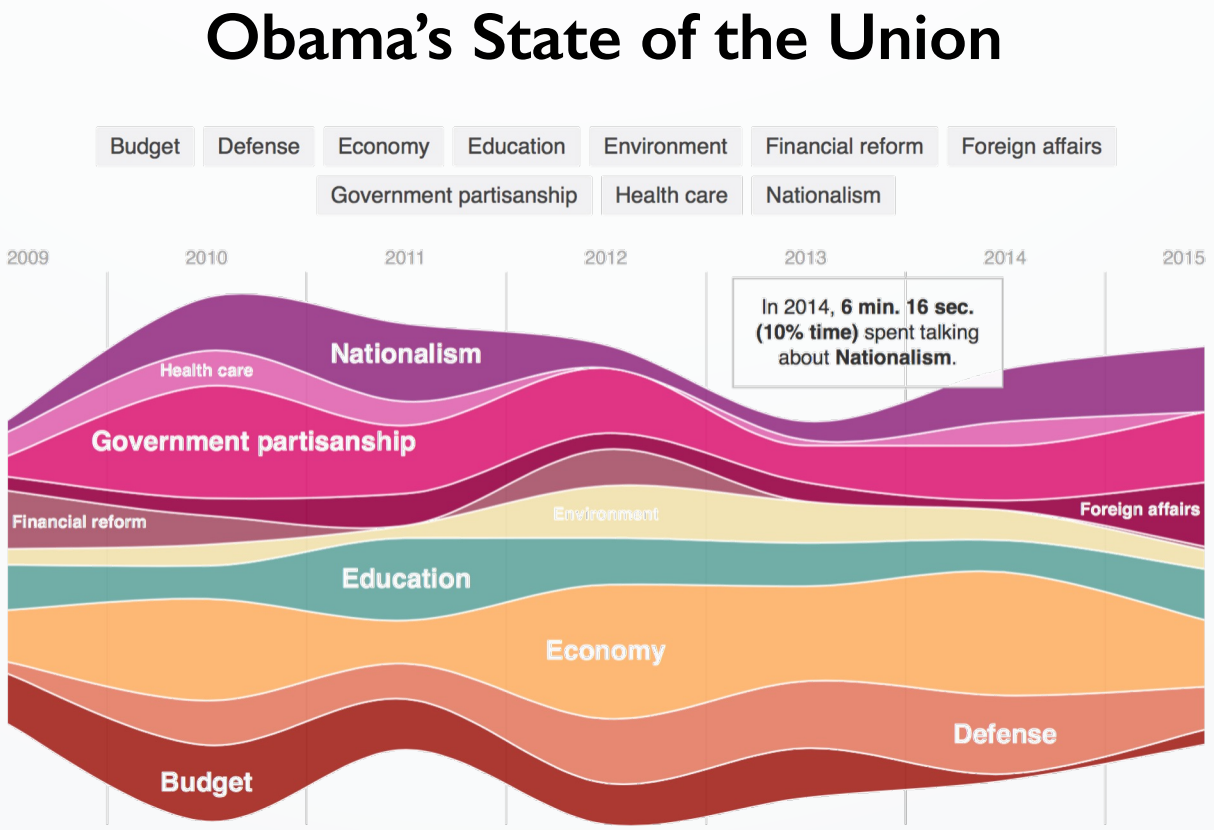
- **主题河 ThemeRiver**
  - 河流，水如时间
  - 不同的河流带 – 表示不同的类别，可以是主题、关键词、人物等等



Susan Havre et al. ThemeRiver: Visualizing Theme Changes over Time. (IEEE SIV 2000)

时序文档的文本内容可视化技术

- 主题河 ThemeRiver



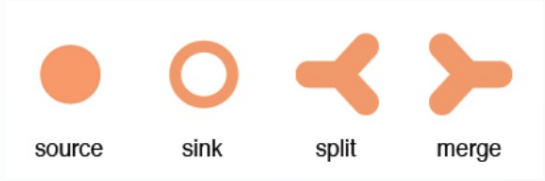
http://interactives.ap.org/2015/sotu-2015/

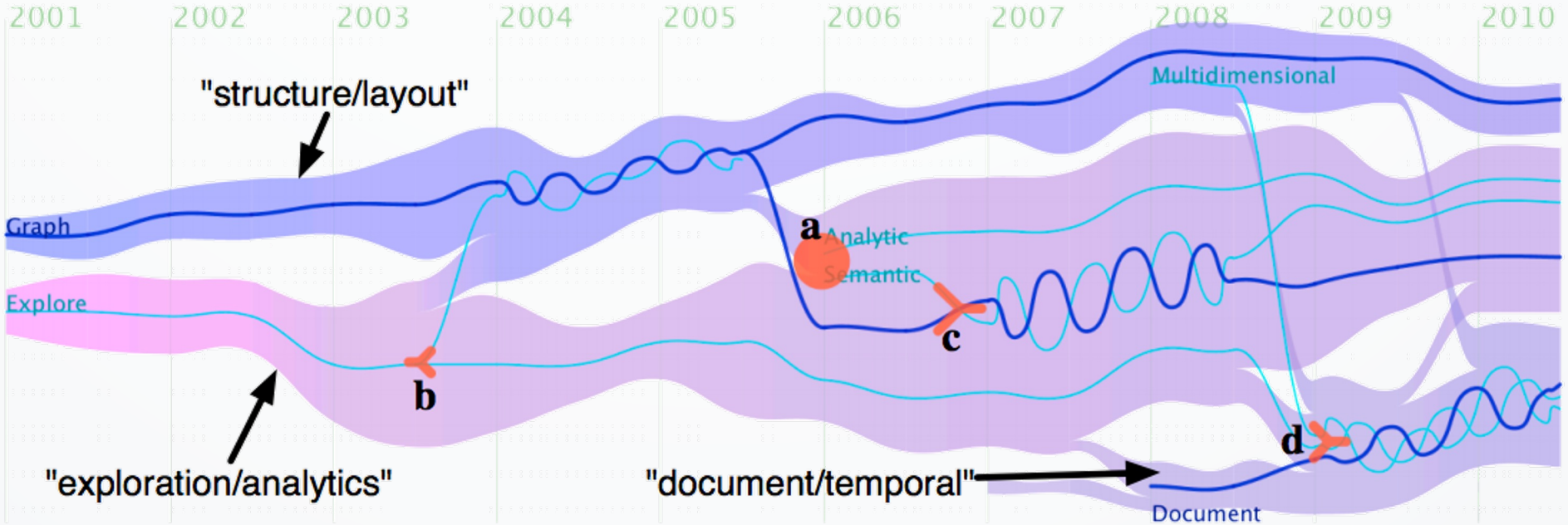
时序文档的文本内容可视化技术

- 主题河 ThemeRiver

时序文档的文本内容可视化技术

- **TextFlow**

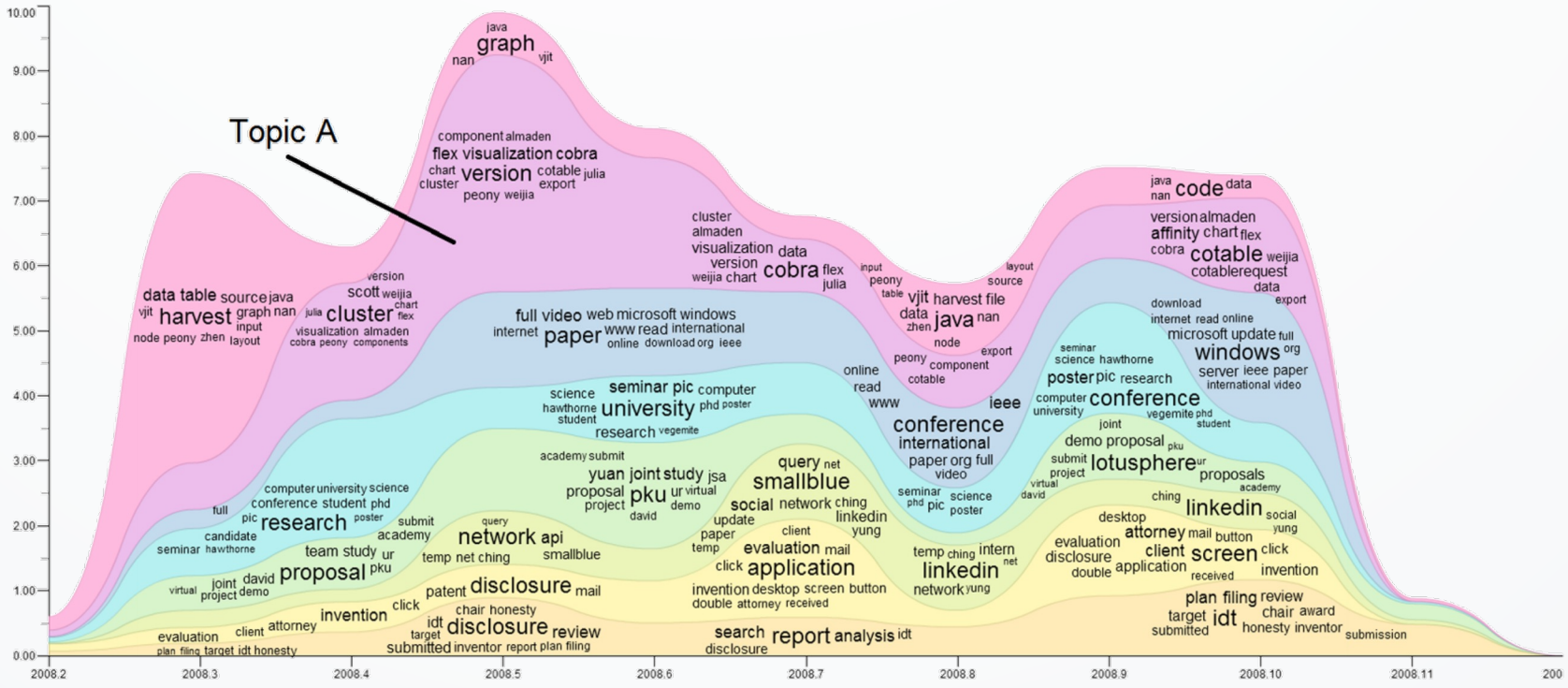




Weiwei Cui et al. TextFlow: Towards Better Understanding of Evolving Topics in Text. (IEEE TVCG 2011)

时序文档的文本内容可视化技术

- **TIARA**



Shixia Liu et al. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. (ACM TIST 2012)

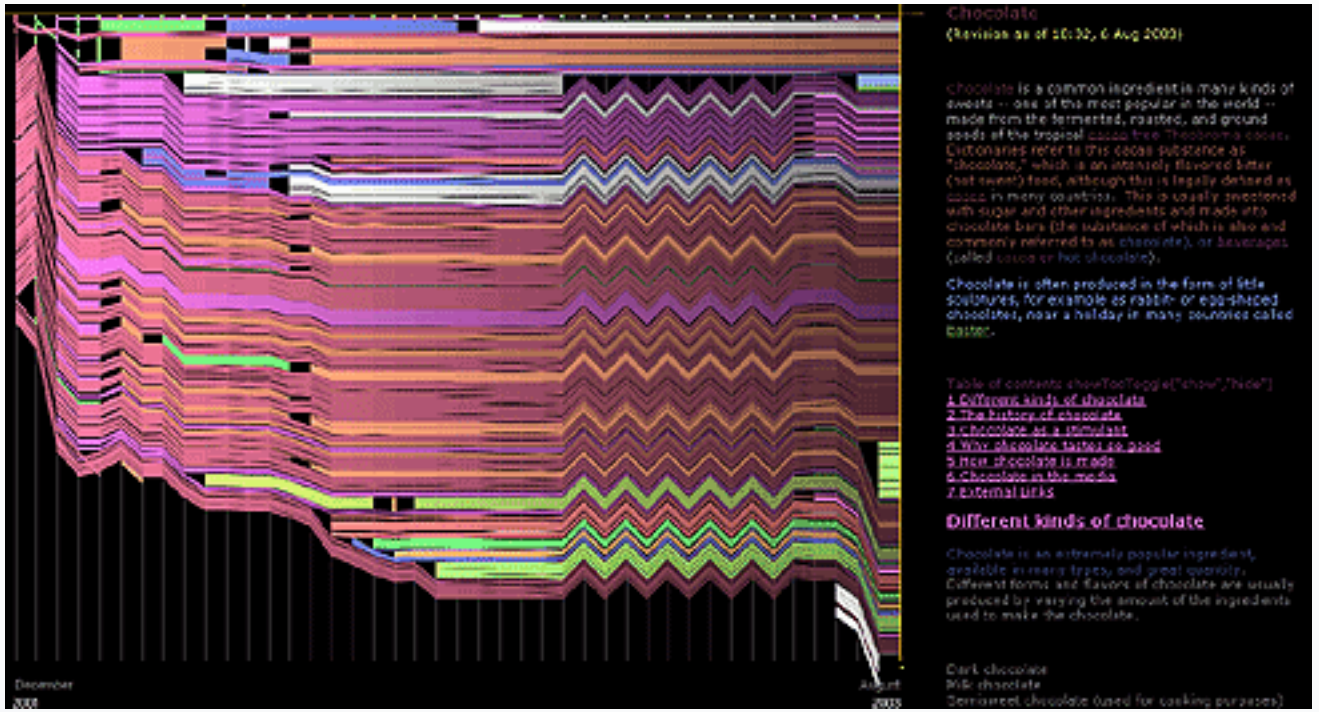
时序文档的文本内容可视化技术

- EvoRiver

Sun et al. "**EvoRiver**: Visual analysis of topic coopetition on social media." TVCG 2014)

时序文档的文本内容可视化技术

- **History Flow**



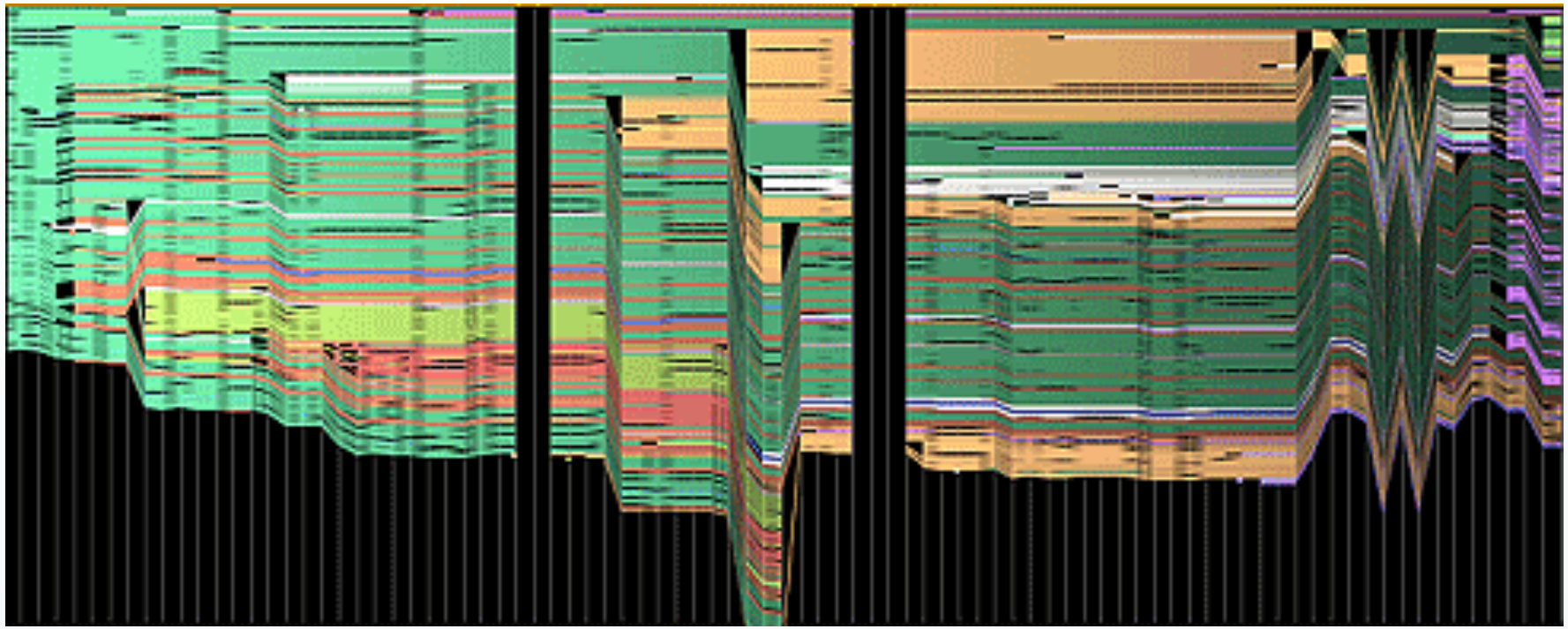
Wikipedia: Chocolate
Maintainers and modifications in each version

History Flow at MoMA

http://hint.fm/projects/historyflow/

时序文档的文本内容可视化技术

- History Flow

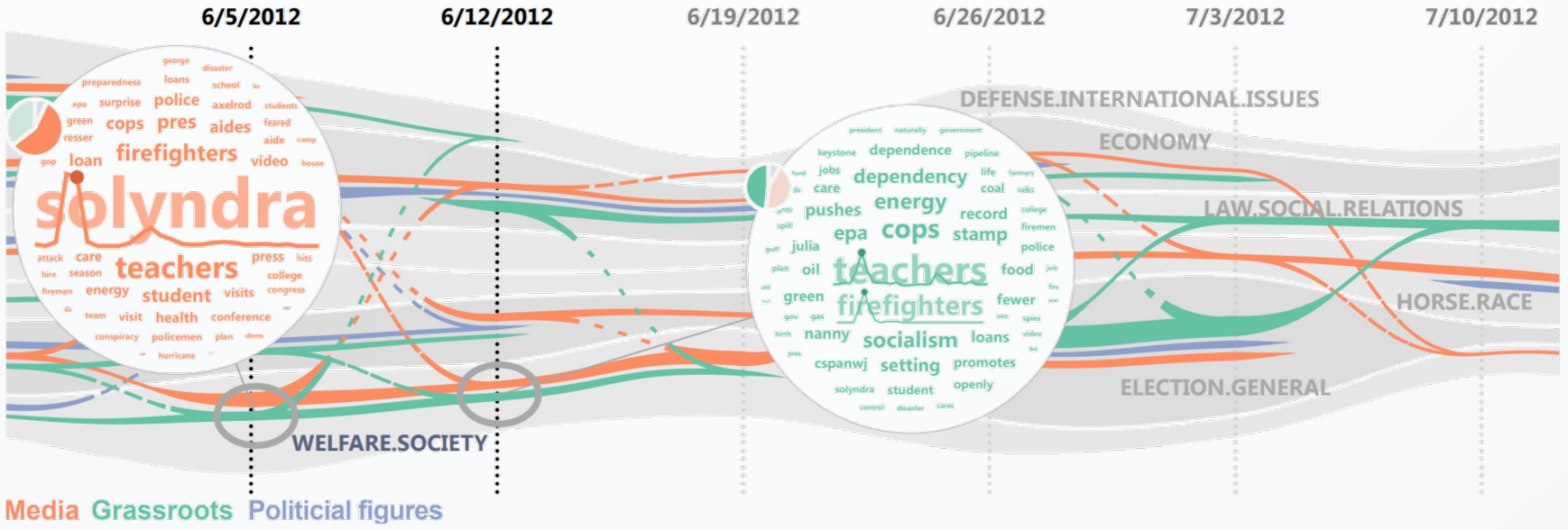


Wikipedia: Abortion

http://hint.fm/projects/historyflow/

时序文档的文本内容可视化技术

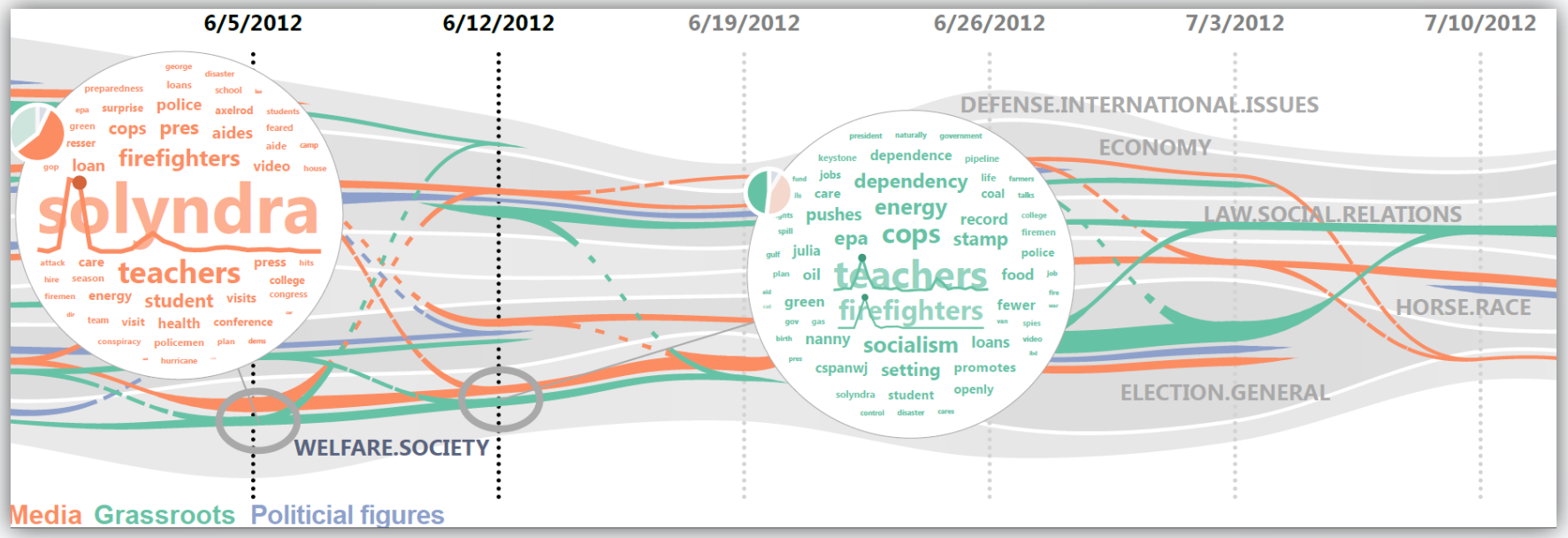
- 社交媒体数据可视分析案例SocialFlow



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

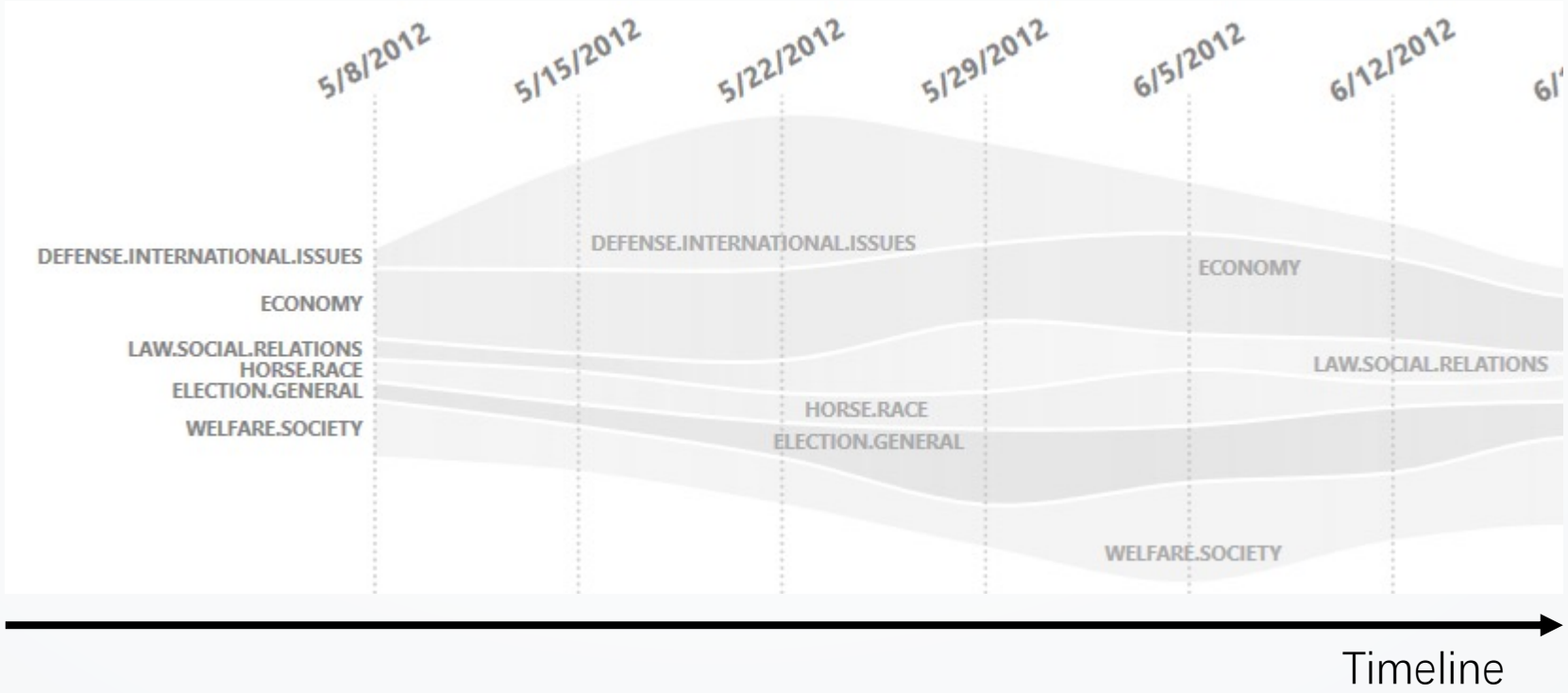
- **社交媒体数据可视分析案例SocialFlow**
- 2012 US presidential election Twitter Data
  - 89,174,308 tweets classified into six major topics
  - May 01, 2012 to November 20, 2012



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

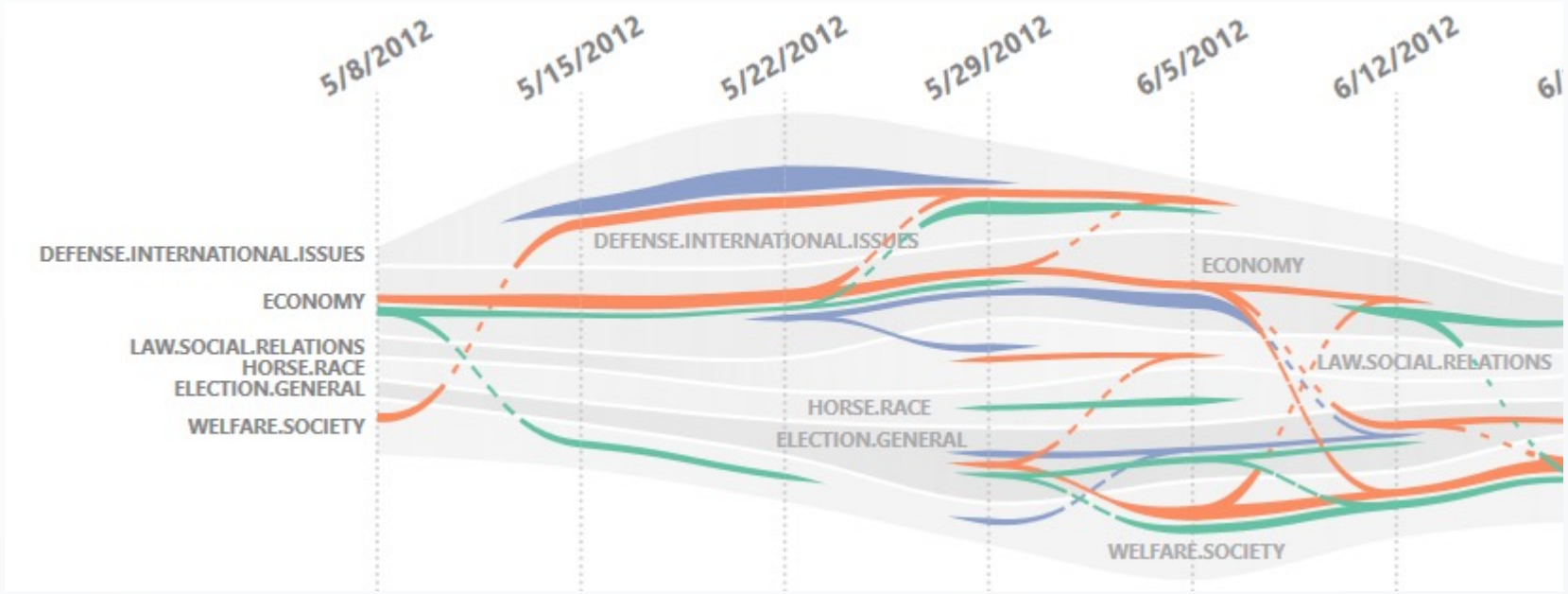
- **社交媒体数据可视分析案例SocialFlow**
- Overall trend of the competitiveness of topics



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

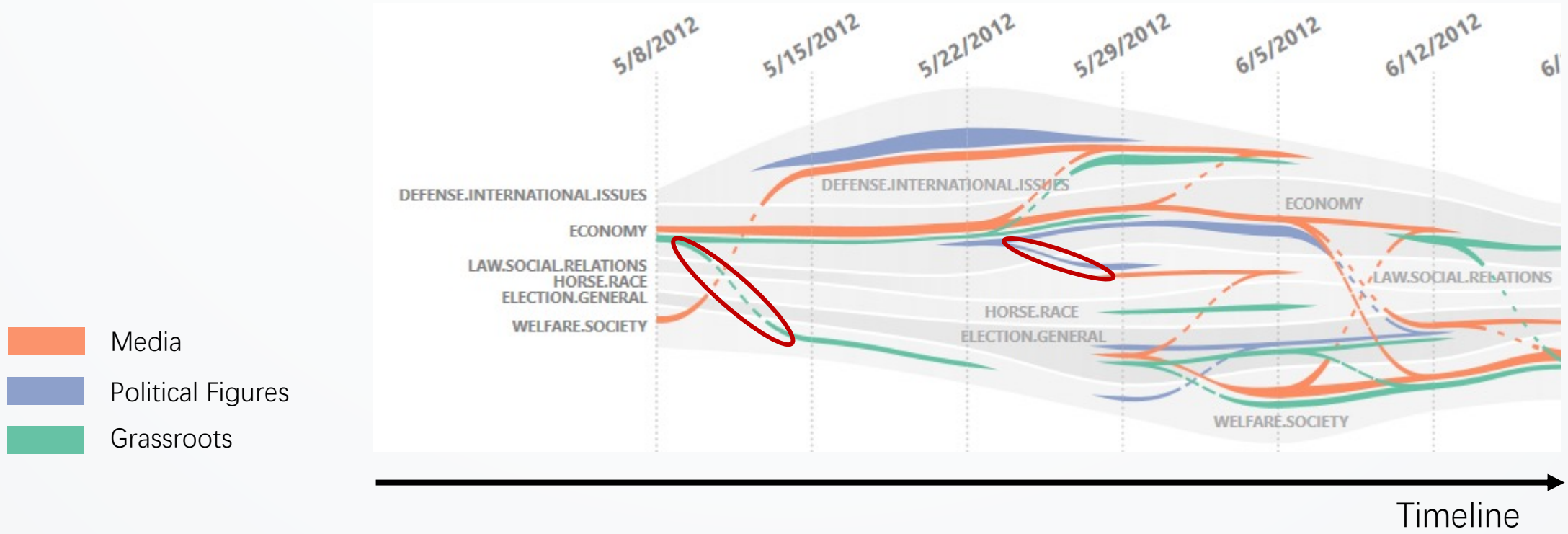
- **社交媒体数据可视分析案例SocialFlow**
- Influences of opinion leaders



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

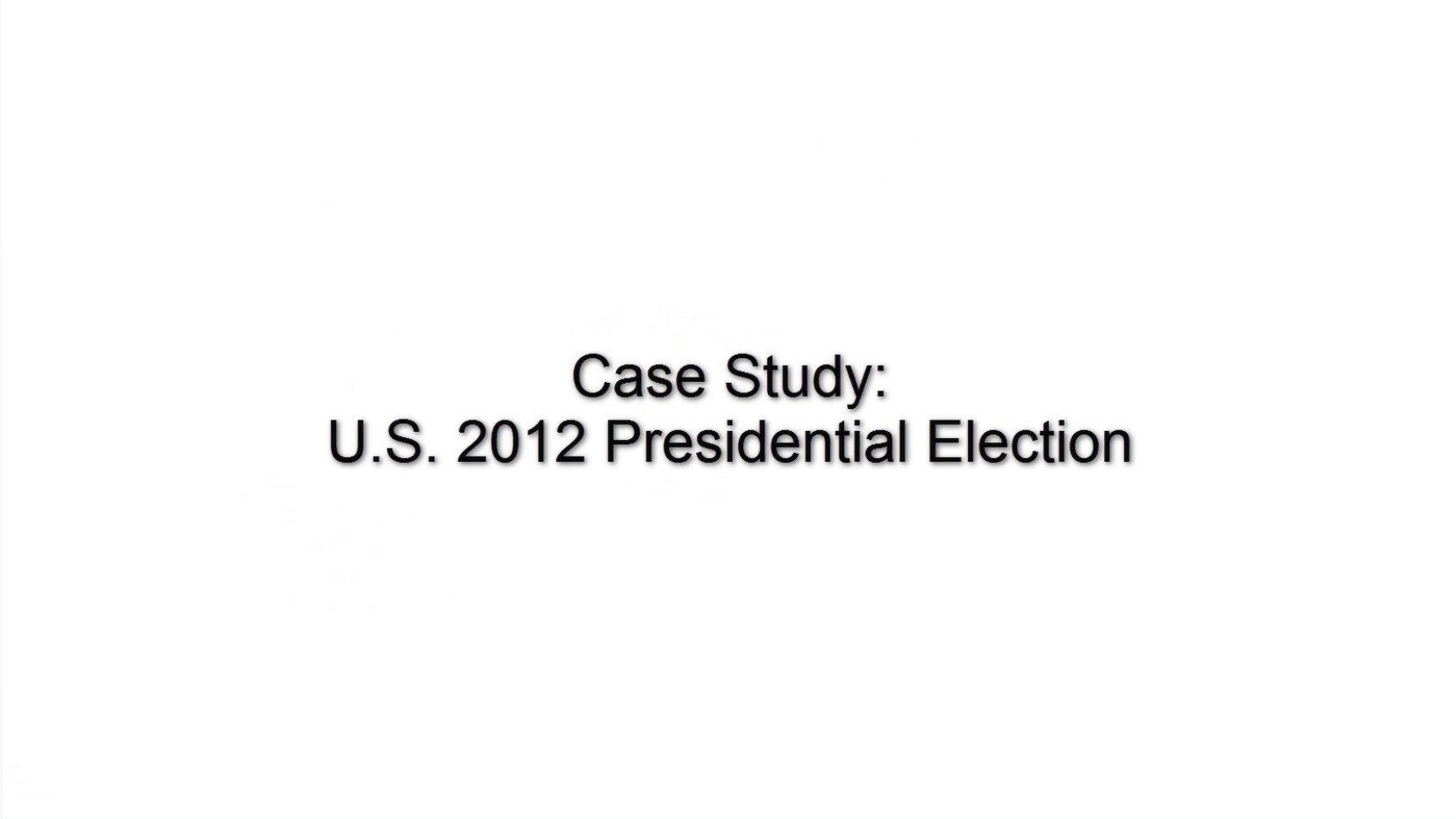
- 社交媒体数据可视分析案例SocialFlow



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

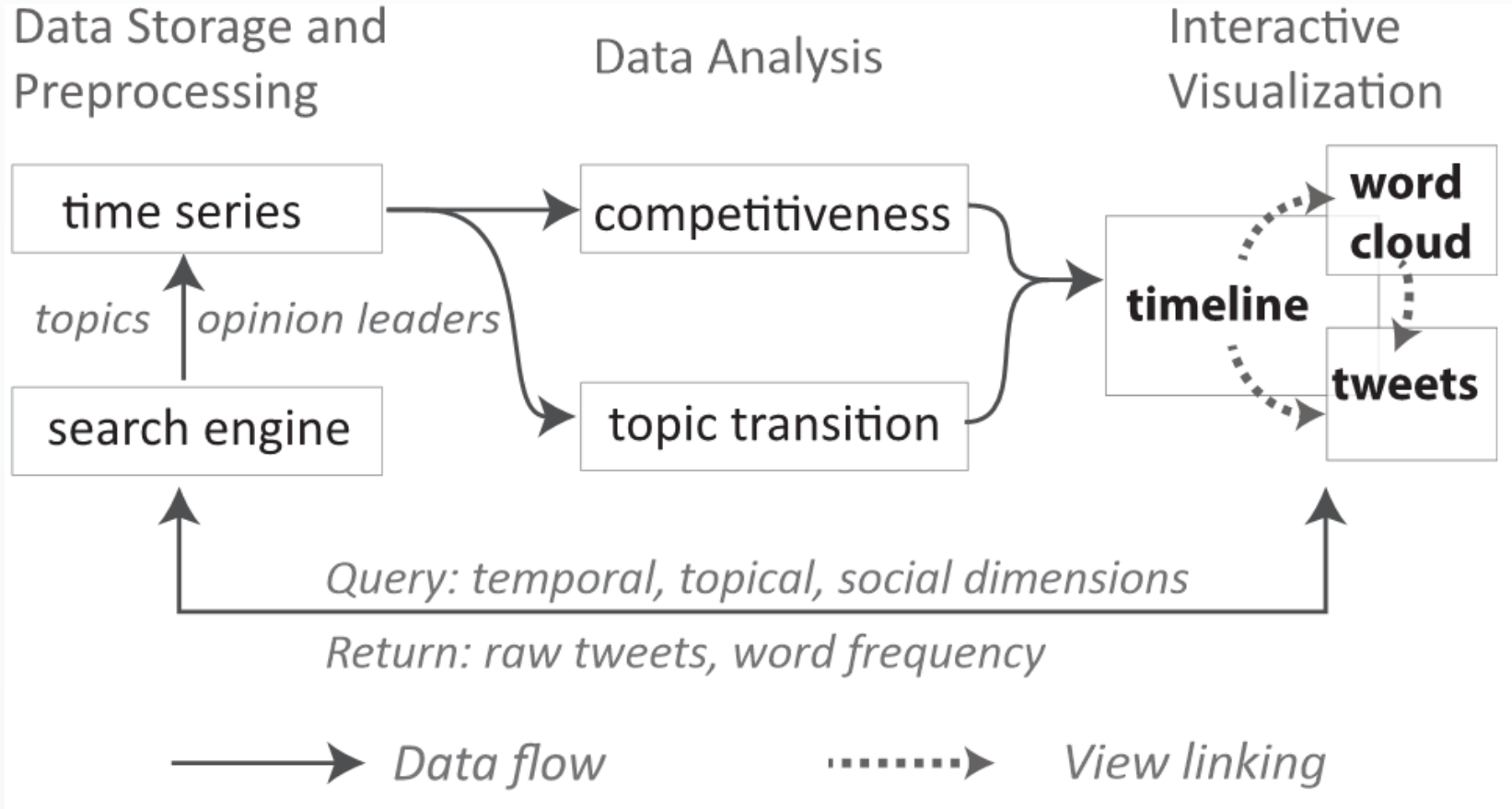
- 社交媒体数据可视分析案例SocialFlow



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

- 社交媒体数据可视分析案例SocialFlow

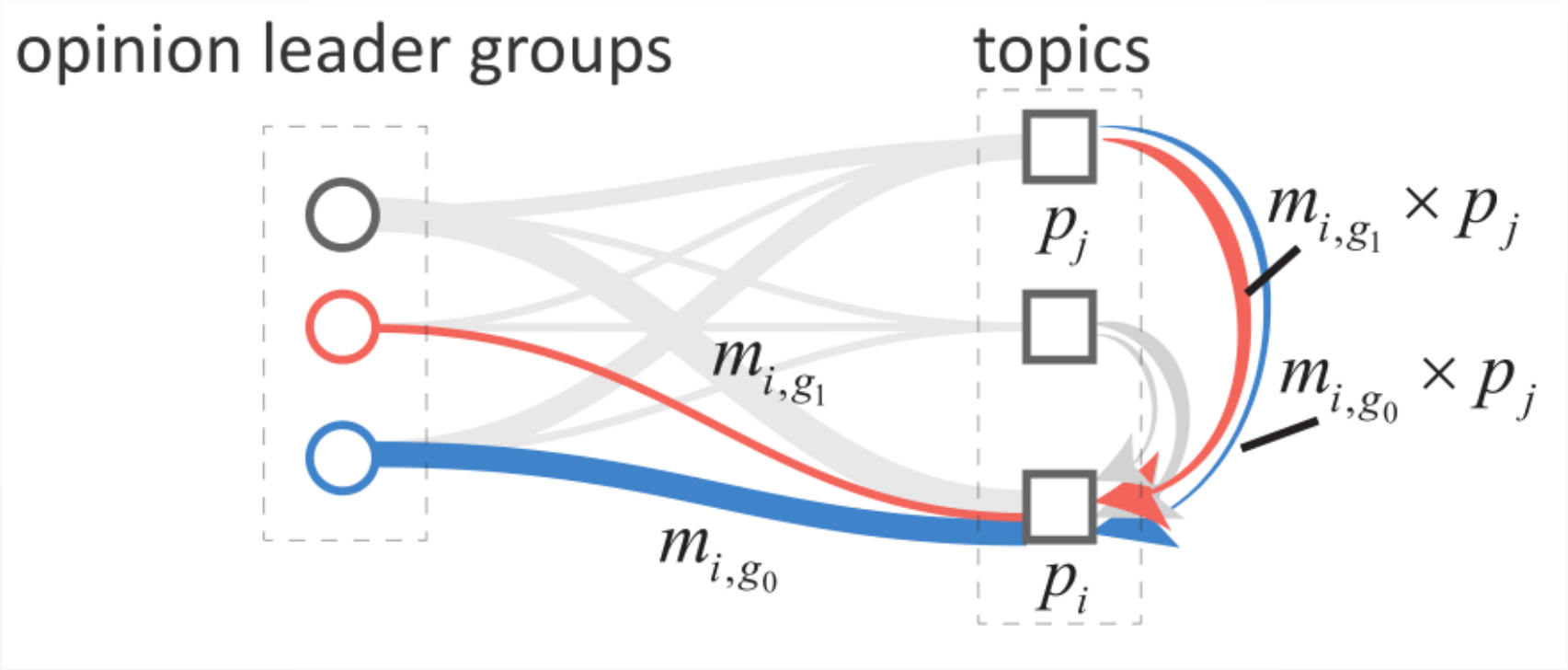


Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

- 社交媒体数据可视分析案例SocialFlow

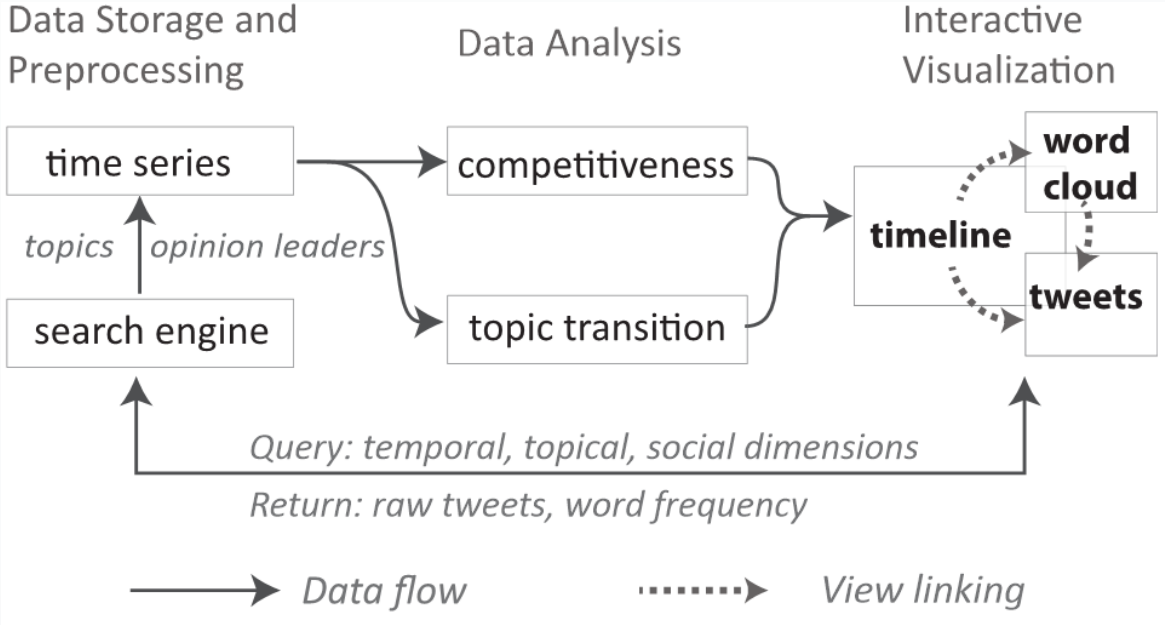
$$p_i^t = \alpha_i p_i^{t-1} + \sum_{g=1}^{n} m_{i,g}^{t-1} \sum_{j=1, j\neq i}^{k} \beta_{i,j,g} p_j^{t-1} - p_i^{t-1} \sum_{j=1, j\neq i}^{k} \left(\sum_{g=1}^{n} \beta_{j,i,g} m_{j,g}^{t-1} \right)$$



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

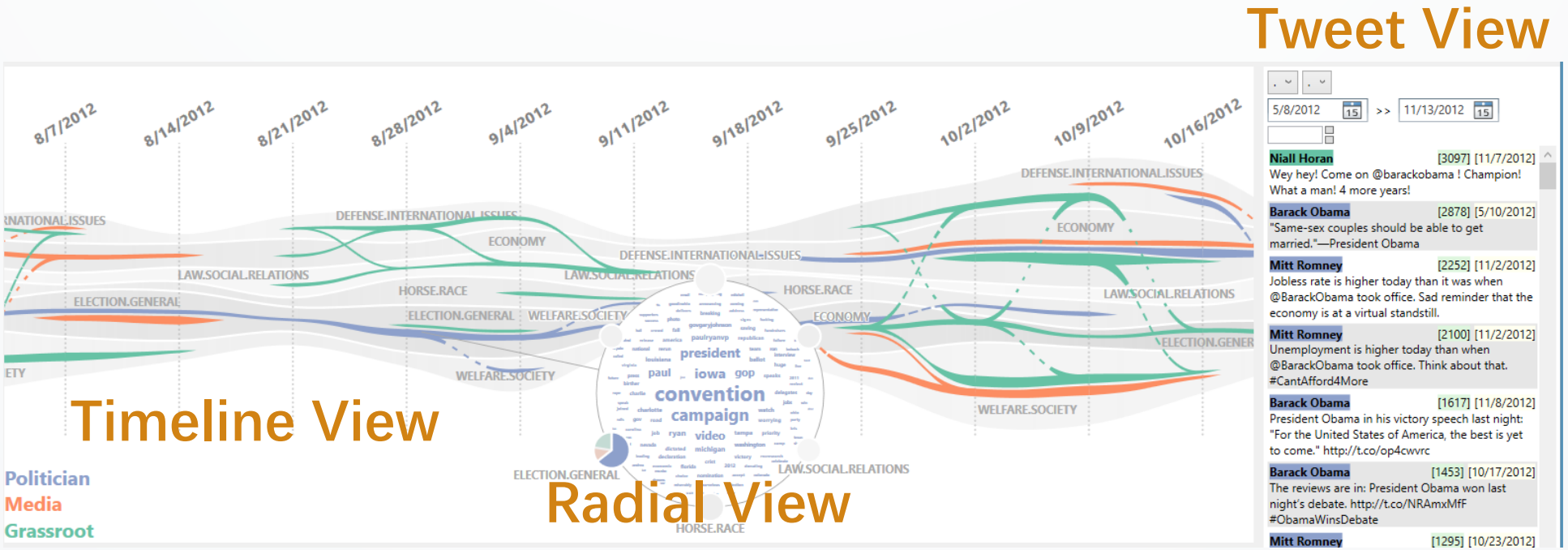
- **社交媒体数据可视分析案例SocialFlow**
- Visualization system
  - Timeline View
  - Radial View
  - Tweet View



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

时序文档的文本内容可视化技术

- 社交媒体数据可视分析案例SocialFlow



Panpan Xu et al. Visual Analysis of Topic Competition on Social Media. (IEEE TVCG 2013)

文本数据可视化技术

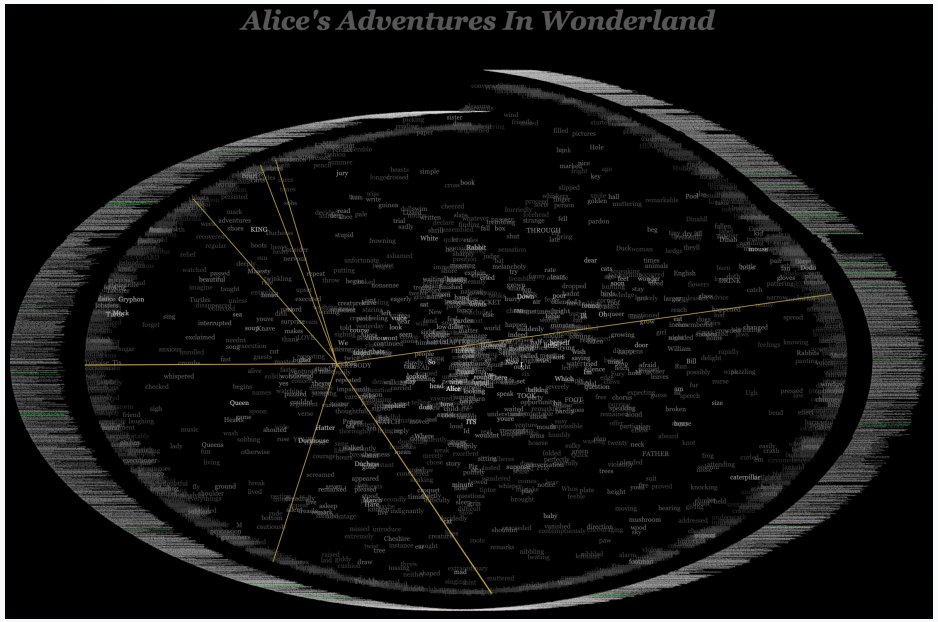
- **根据特征分类**
  - 可视化文本数据内容
    - 基于关键词的可视化
    - 时序文档的可视化
    - 基于特征分布的可视化
  - 可视化文本数据关联
  - 多层级文本数据可视化

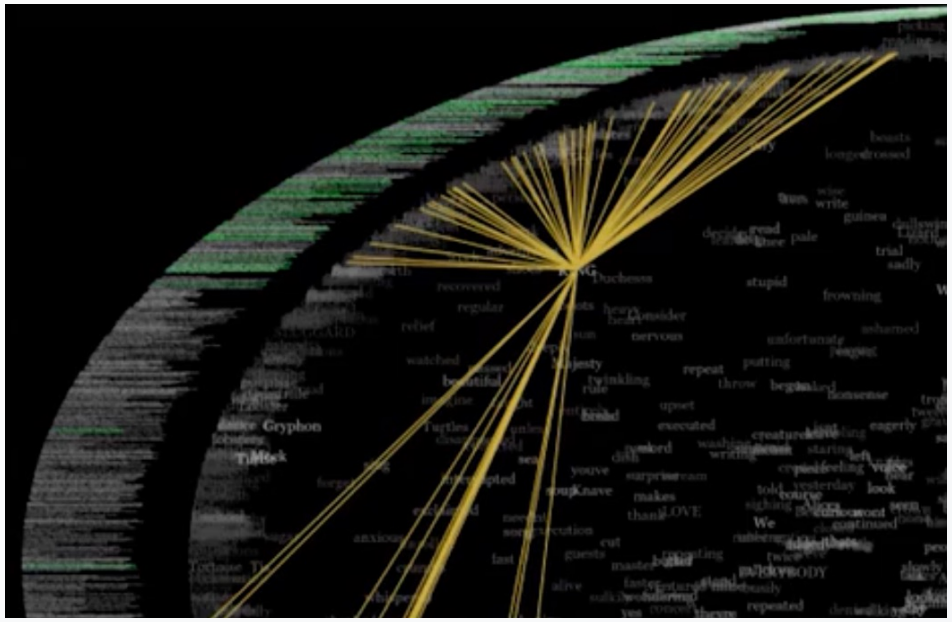
基于特征分布的文本内容可视化技术

- **单个文档或文档集合中的特征分布模式：**
  - 平均句长
  - 词汇量
  - ……

基于特征分布的文本内容可视化技术

- **文本弧 TextArc**
  - 外弧线：文档中的句子
  - 内部单词：文档的单词

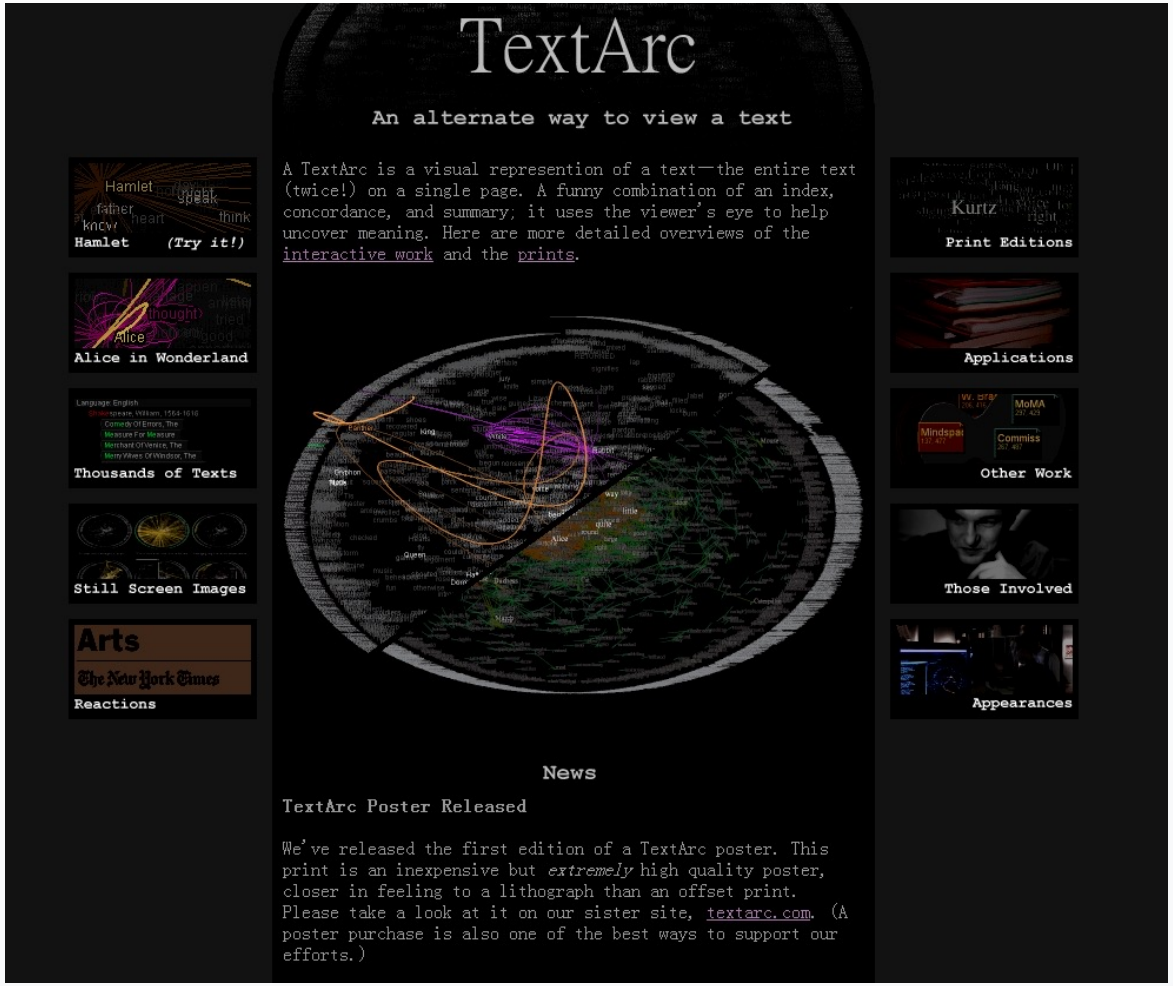




http://vallandingham.me/textarc/

基于特征分布的文本内容可视化技术

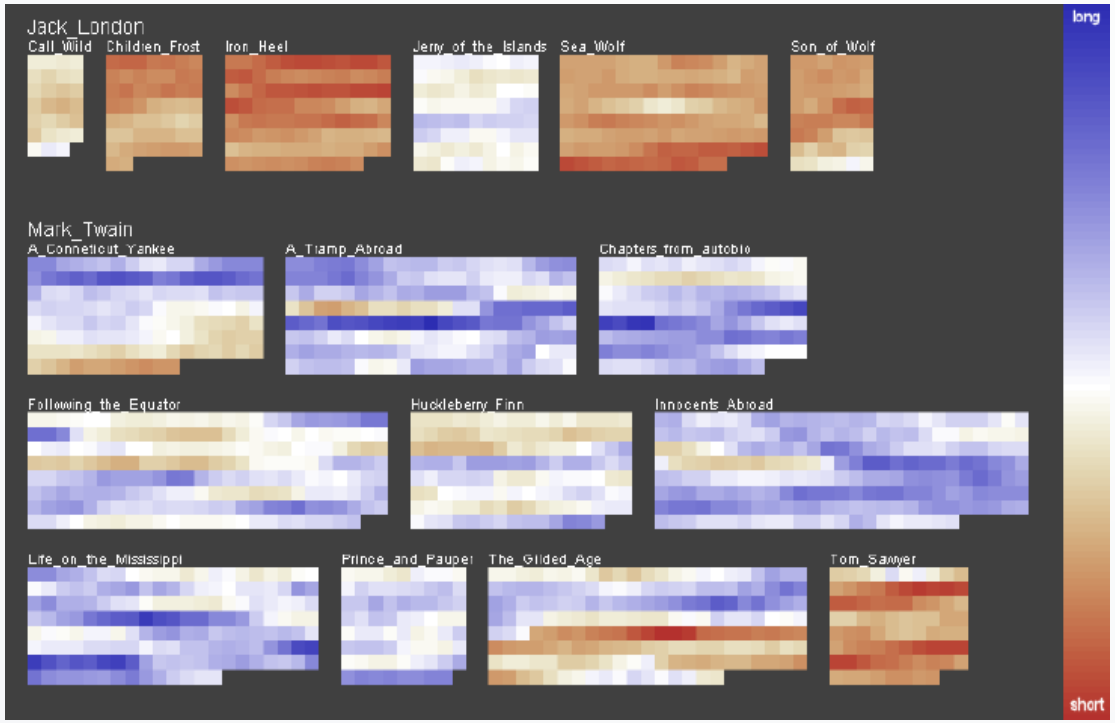
- 文本弧 TextArc



http://vallandingham.me/textarc/

基于特征分布的文本内容可视化技术

- **文本指纹 Literature Fingerprinting**
  - 描述文档的统计特征



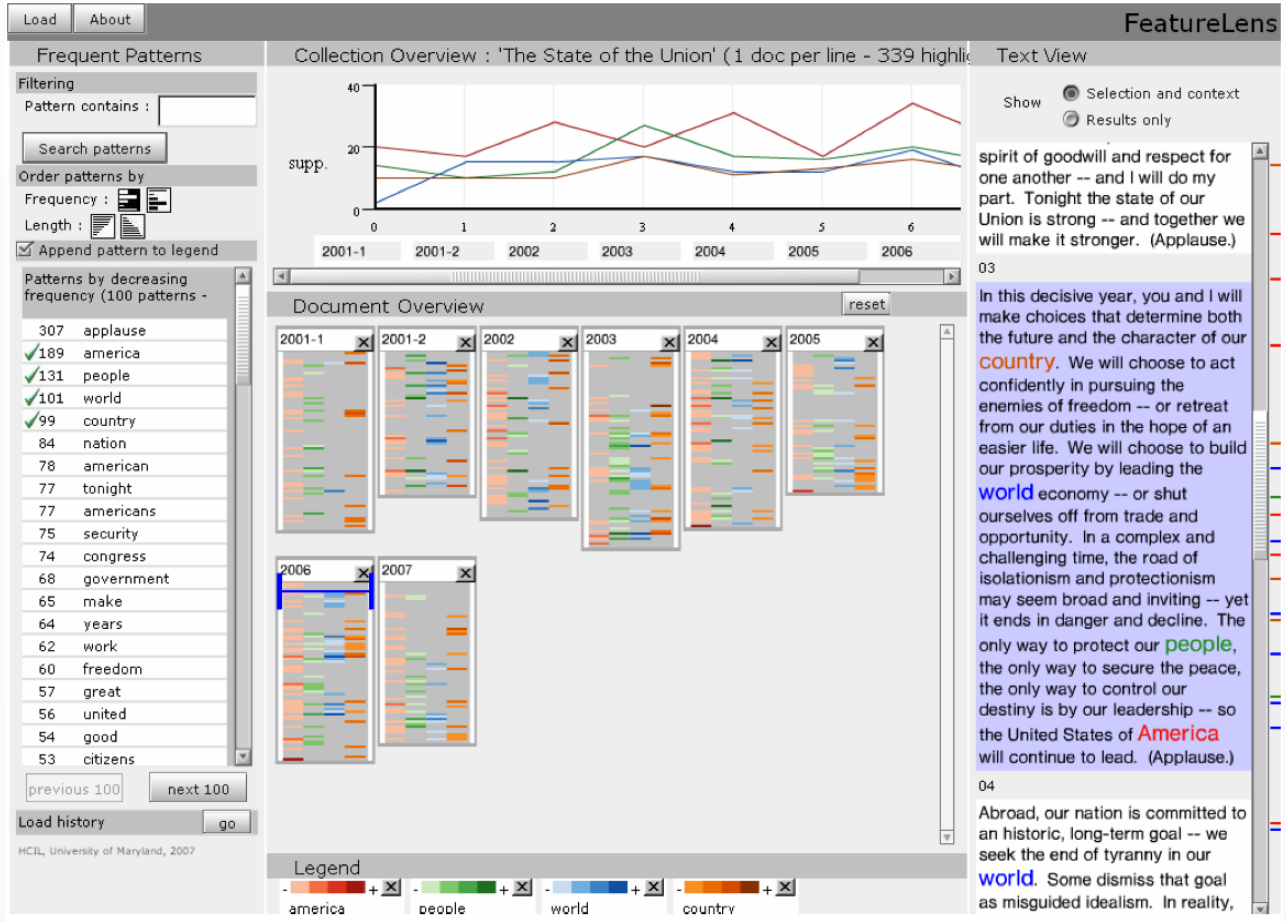
- 马克·吐温与杰克·伦敦写作风格的差异--句子长度。

- 每个像素代表一个段落。
- 一组像素表示一本书。
- 颜色编码句子的长度

Daniel A. Keim & Daniela Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. (IEEE VAST 2007)

基于特征分布的文本内容可视化技术

- **特征透镜 FeatureLens**



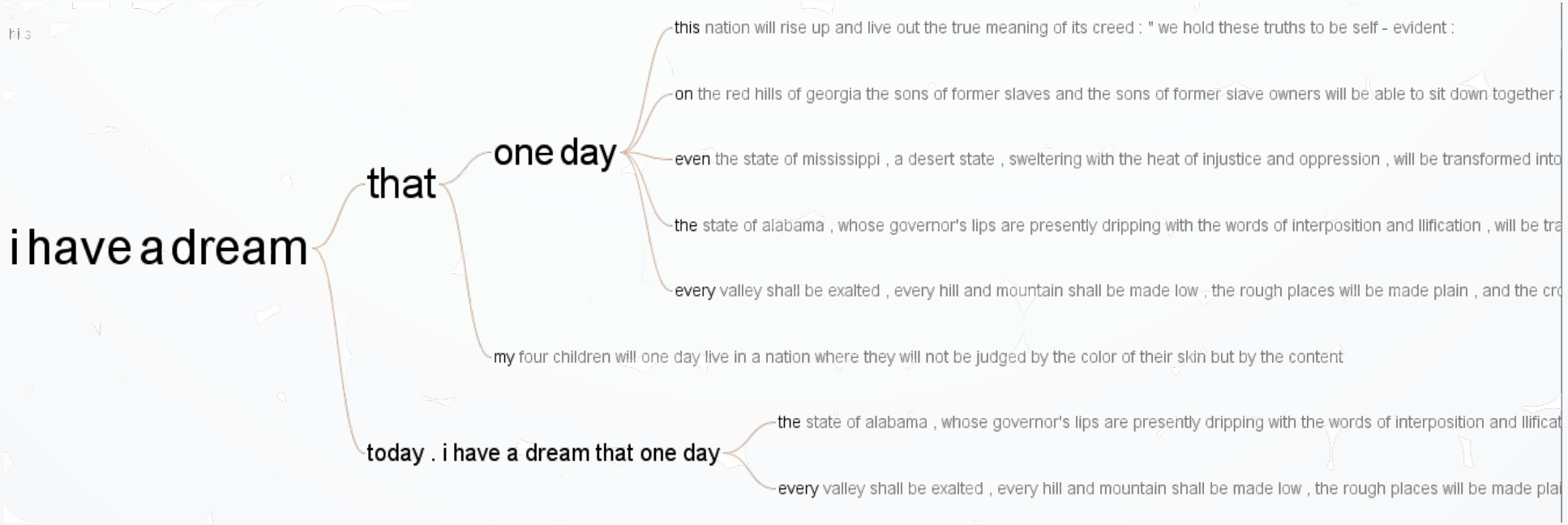
Anthony Don et al. FeatureLens: Interactive Visualization of Text Patterns. (Information Visualization 2007)

文本数据可视化技术

- **根据分析任务特征分类**
  - 可视化文本数据内容
  - 可视化文本数据关联
    - 文本数据集合之间的关系：论文引用、网页超链接、相似性、层级性
    - 可视化方法：图布局、树布局、投影布局
  - 多层级文本数据可视化

文本数据关联的可视化技术

- **单词树 WordTree**
  - 单词大小–>词频

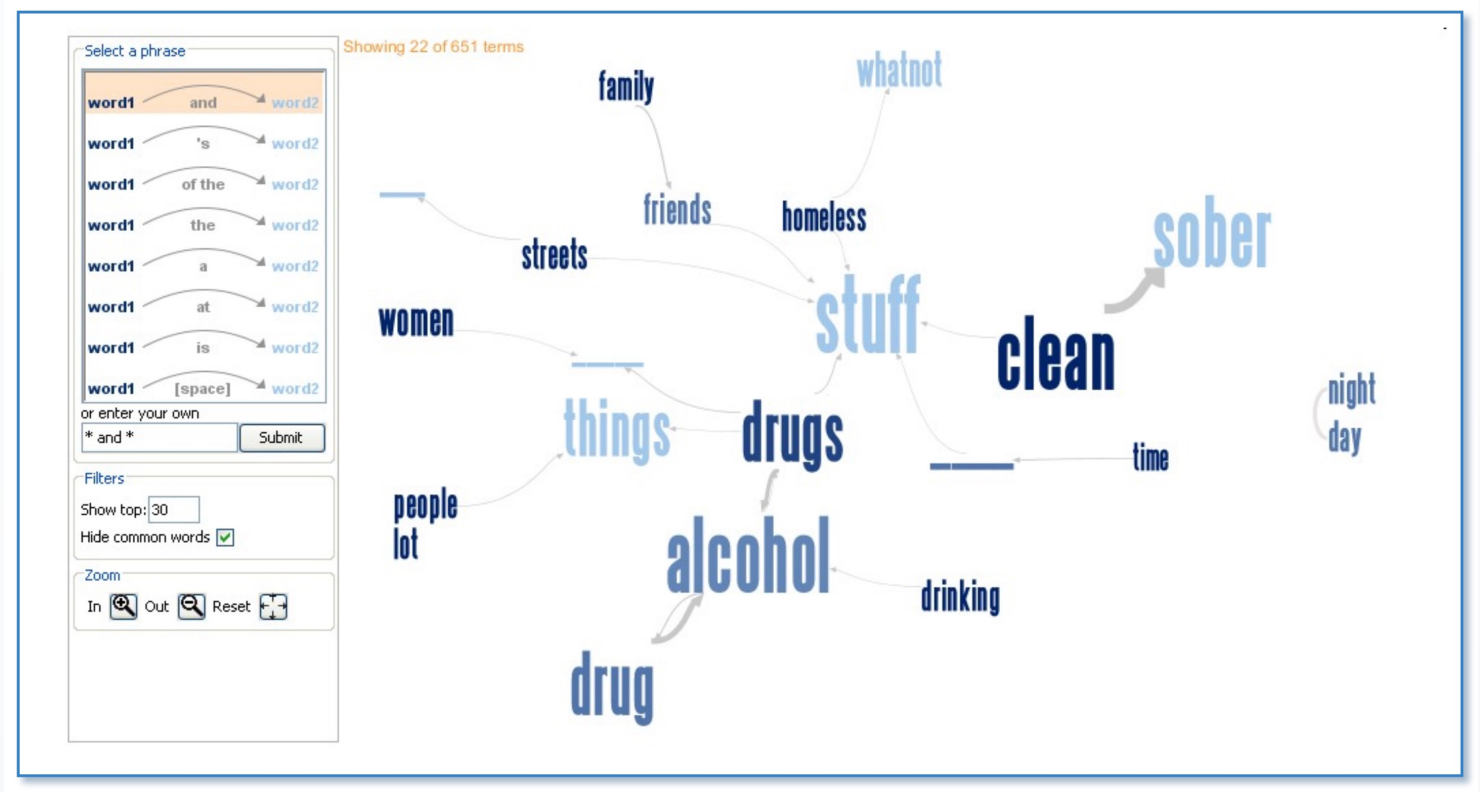


All of the sentences starting with "I" in the speech"I have a dream"

Wattenberg M and Viégas F B. The word tree, an interactive visual concordance. (IEEE TVCG 2008)

文本数据关联的可视化技术

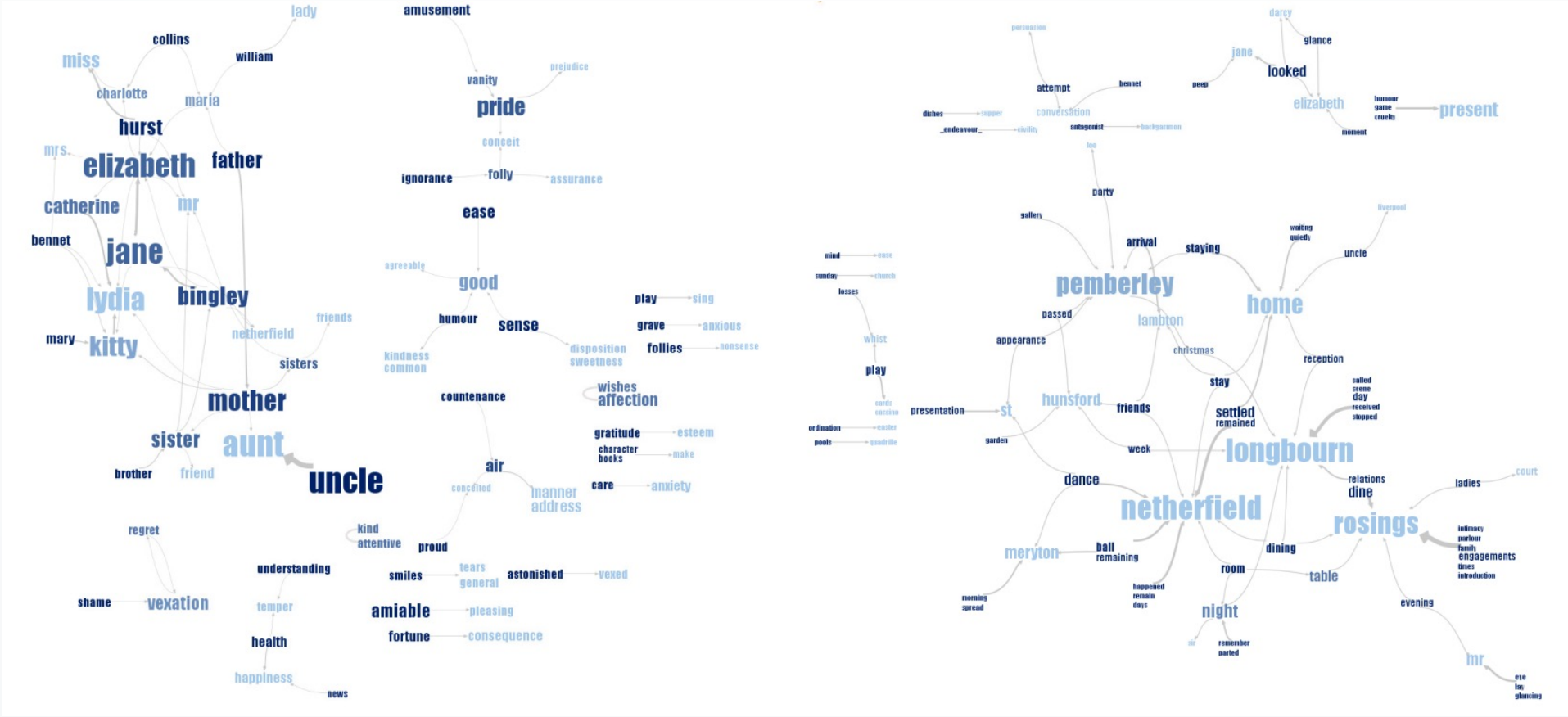
- **短语网络 PhraseNet**
  - 从结点链接图看“A___B”的关系



Van Ham F et al. Mapping text with phrase nets. (IEEE TVCG 2009)

文本数据关联的可视化技术

- 短语网络 PhraseNet



Van Ham F et al. Mapping text with phrase nets. (IEEE TVCG 2009)

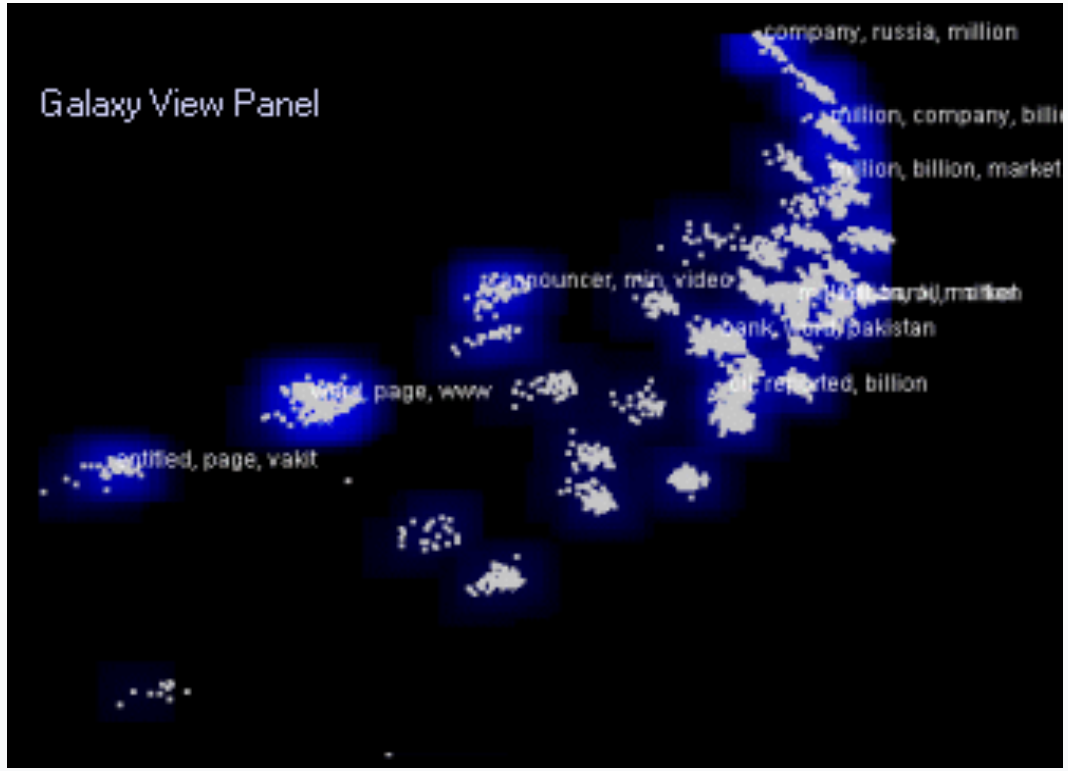
基于特征分布的文本数据可视化技术

- 句子树 SentenTree

Mengdie Hu et al. Visualizing Social Media Content with SentenTree. (IEEE TVCG 2017)

文本数据关联的可视化技术

- **星系视图 GalaxyView**
  - 基于文档相似度的MDS投影



Wise J A.The ecological approach to text visualization.
(Journal of the Association for Information Science and Technology 1999)

文本数据关联的可视化技术

- 主题地貌 ThemeScape

Wise J A.The ecological approach to text visualization.
(Journal of the Association for Information Science and Technology 1999)

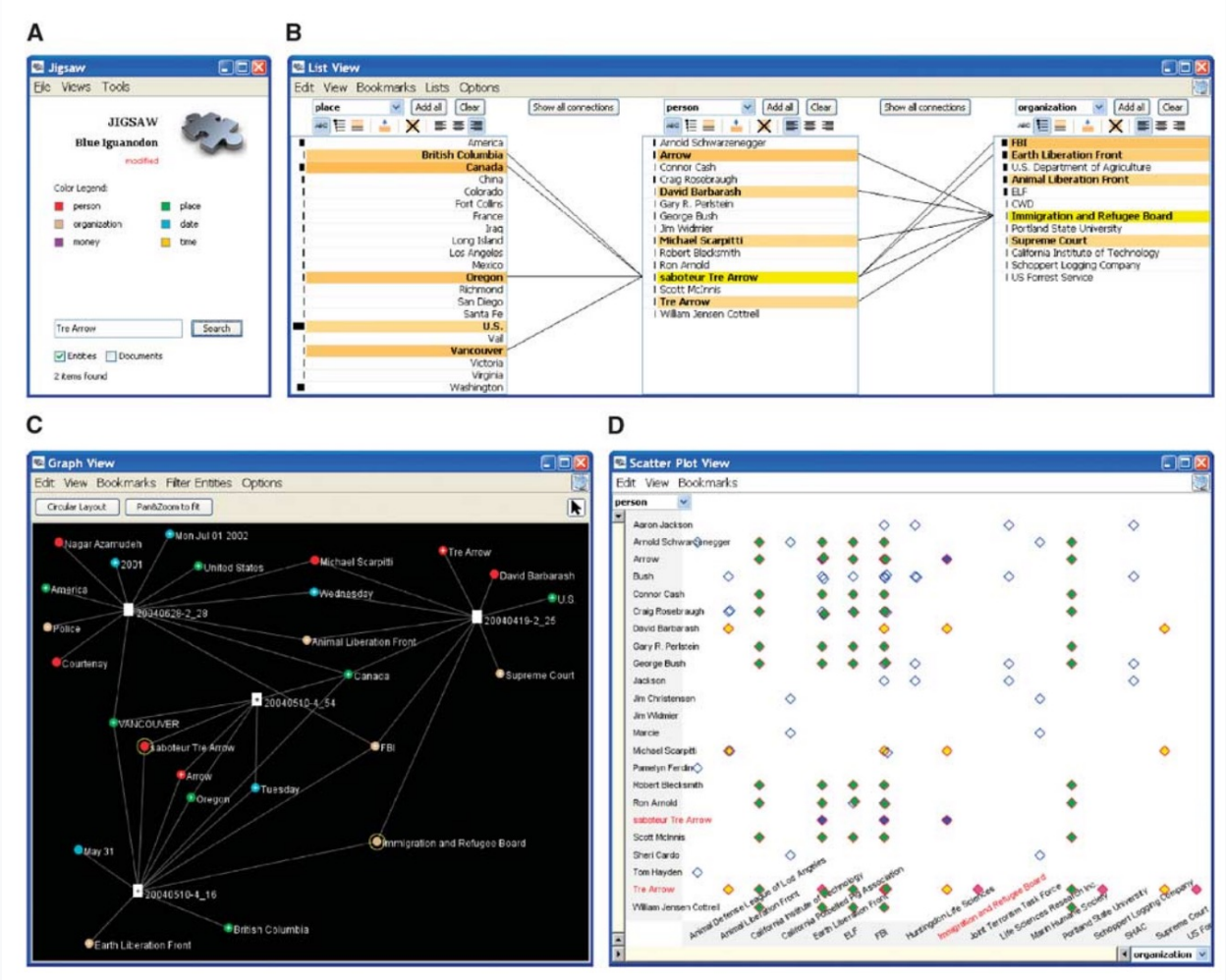
文本数据关系的可视化技术

- **新闻地图 NewsMap**
  - 通过Treemap显示相似性

http://newsmap.jp/

文本数据关系的可视化技术

- Jigsaw 多协同视图



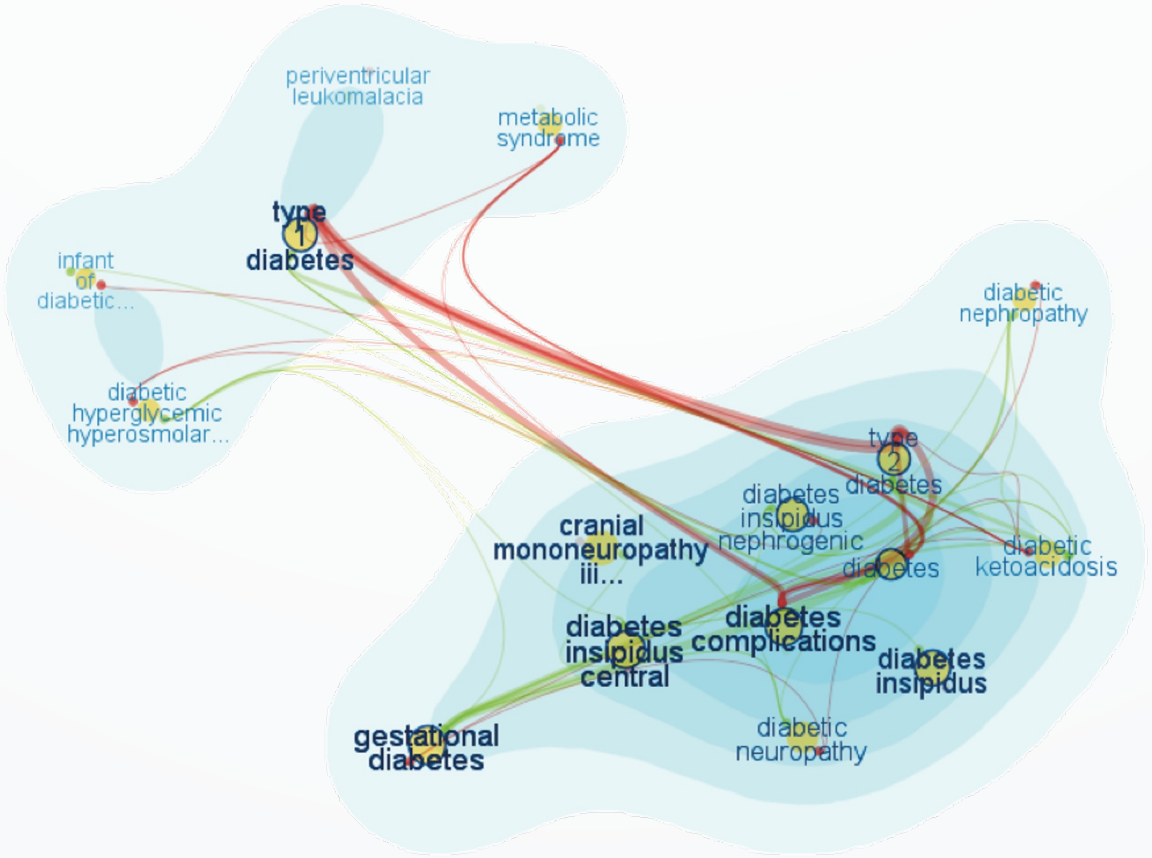
Stasko J et al. Jigsaw: supporting investigative analysis through interactive visualization. (Information Visualization 2008)

文本数据可视化技术

- **根据分析任务特征分类**
  - 可视化文本数据内容
  - 可视化文本数据关联
  - 多层级文本数据可视化
    - 上下文语境
    - 时空信息
    - 作者信息
    - ……

多层级文本数据可视化技术

- **FacetAtlas**



Nan Cao et al. FacetAtlas: Multifaceted Visualization for Rich Text Corpora. (IEEE TVCG 2010)

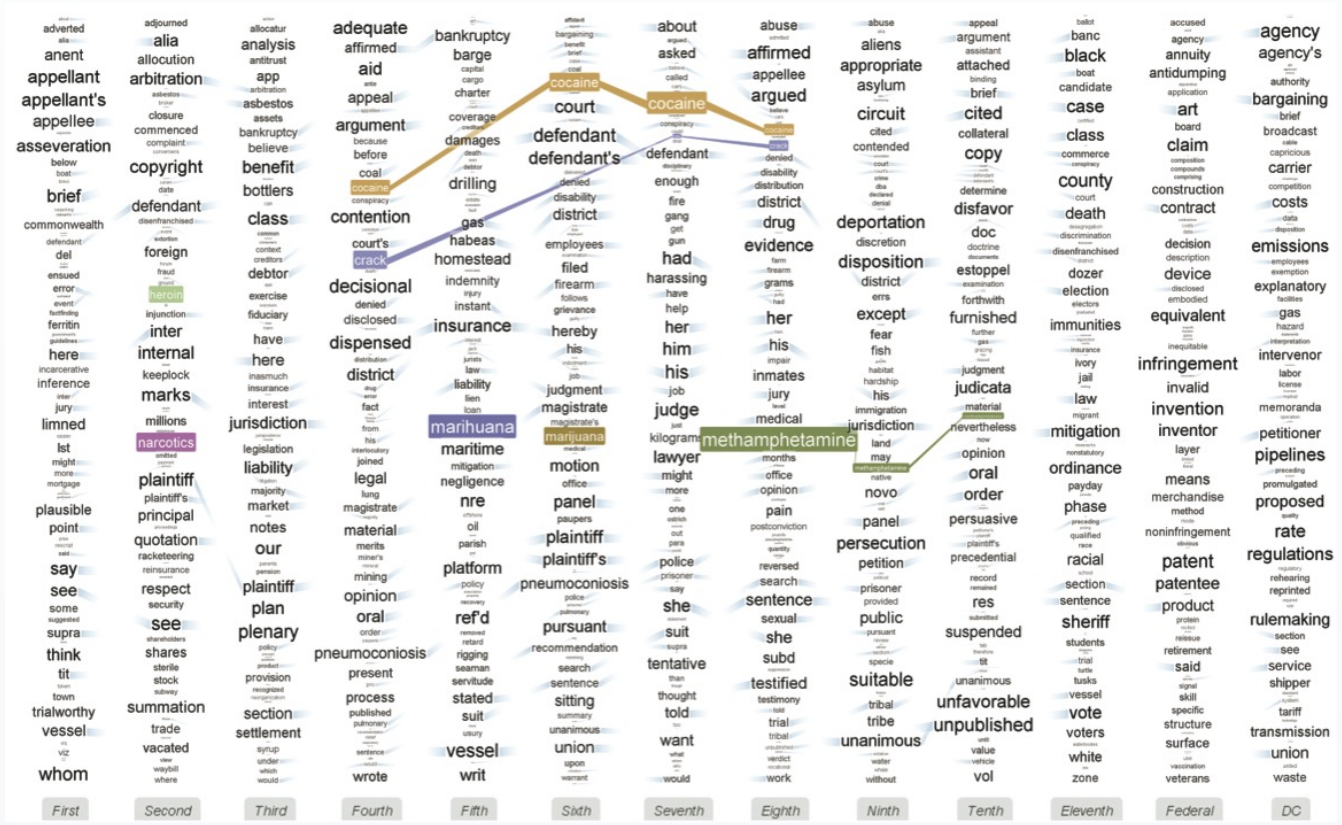
多层级文本数据可视化技术

- FacetAtlas

Nan Cao et al. FacetAtlas: Multifaceted Visualization for Rich Text Corpora. (IEEE TVCG 2010)

多层级文本数据可视化技术

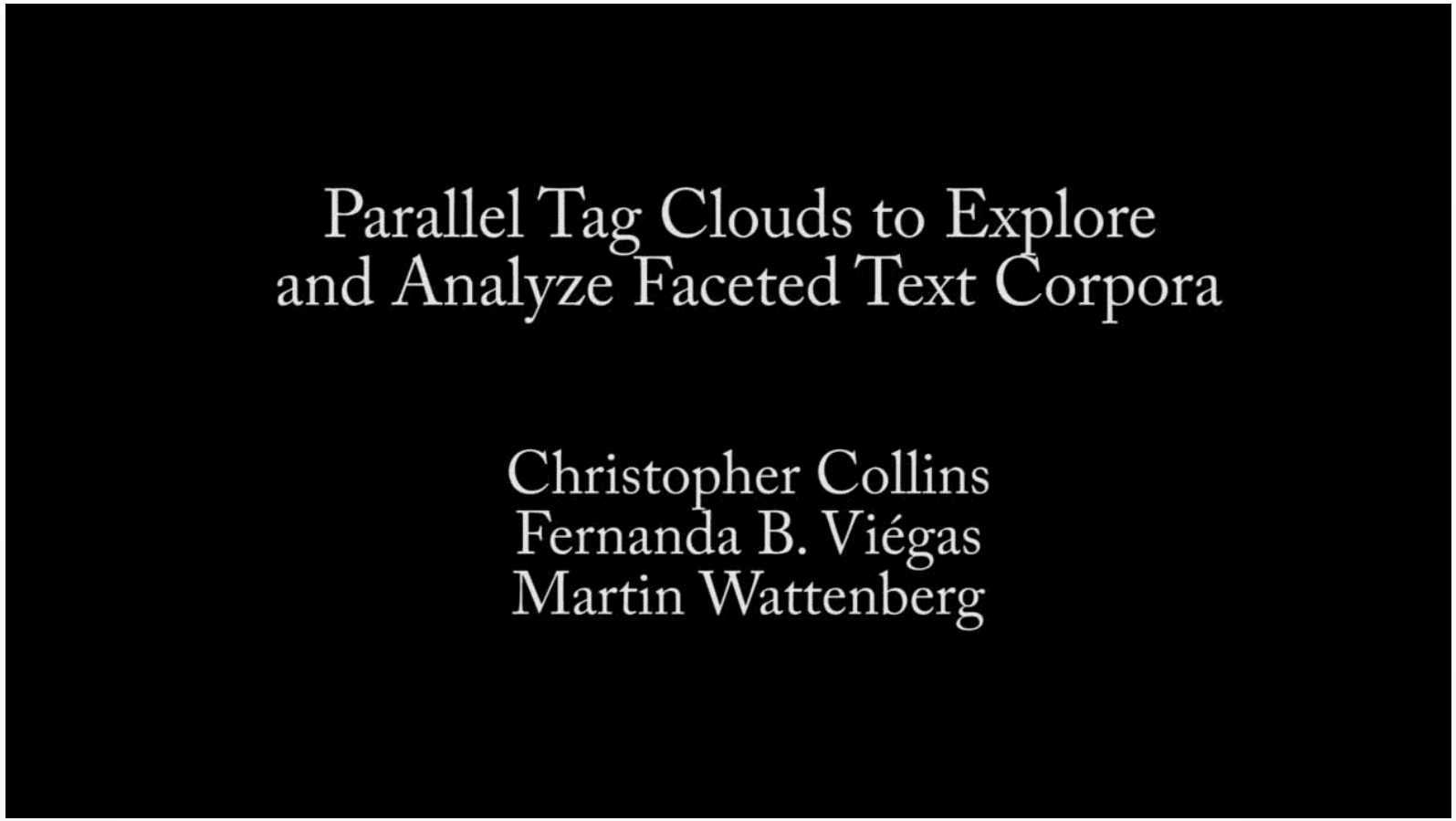
- **Parallel Tag Cloud**



Christopher Collins et al. Parallel Tag Clouds to explore and analyze faceted text corpora. (IEEE VAST 2009)

多层级文本数据可视化技术

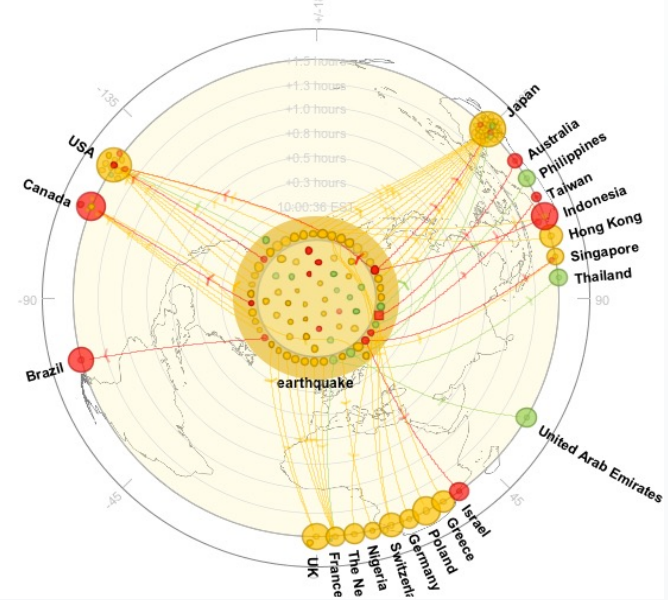
- **Parallel Tag Cloud**

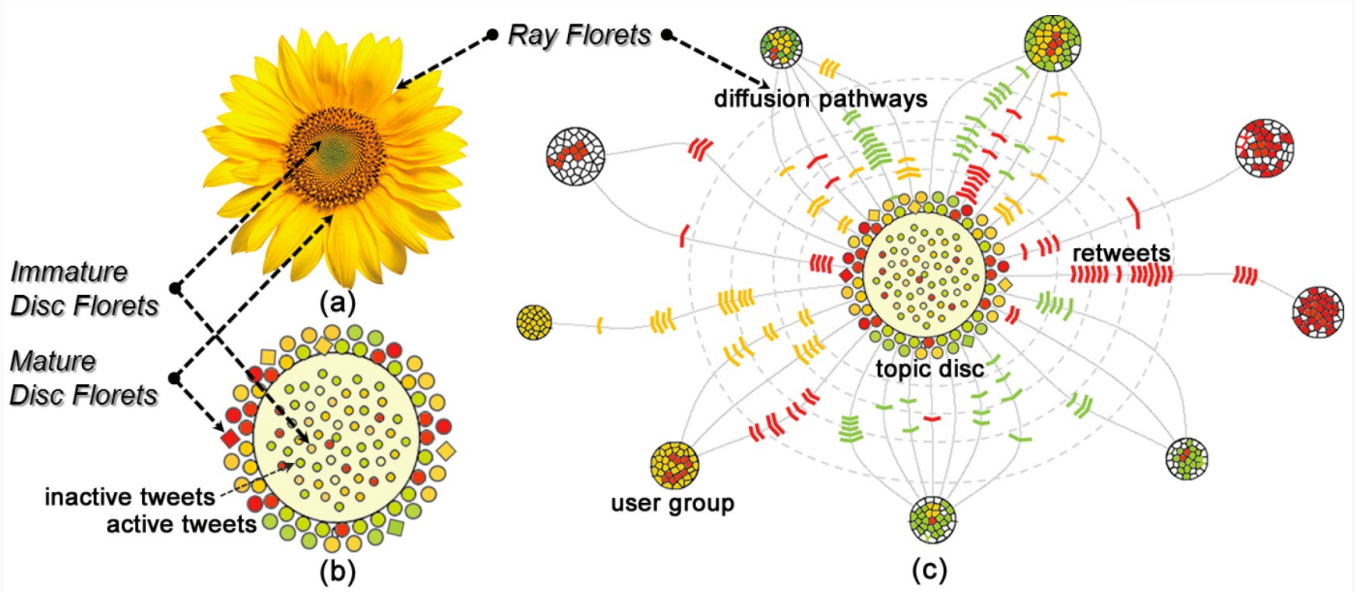


Christopher Collins et al. Parallel Tag Clouds to explore and analyze faceted text corpora. (IEEE VAST 2009)

文本数据可视分析案例（一）

- **文本信息实时扩散 动态监控**
  - 实时感知信息的传播
  - 了解信息传播与主题演化的的动态性

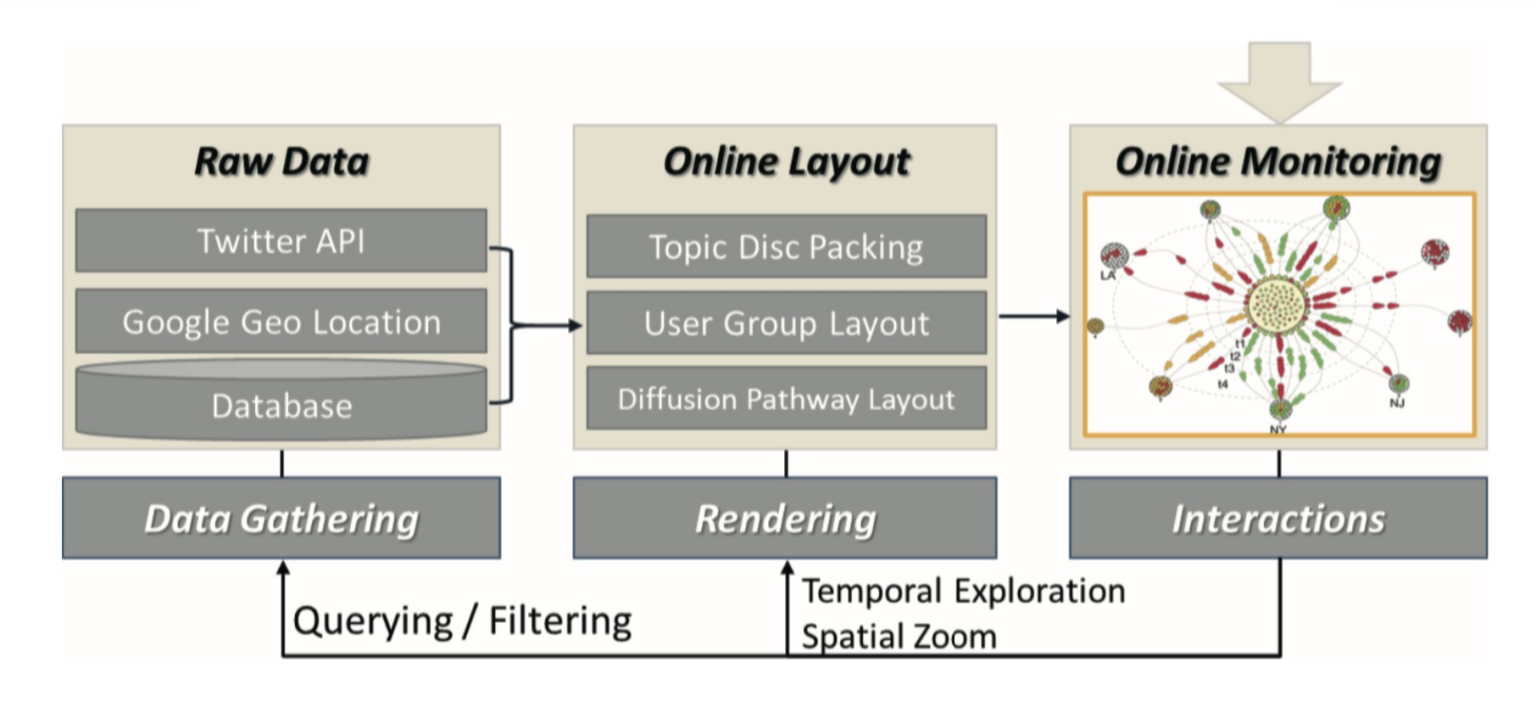




"Whisper: Tracing the Spatiotemporal Process of ***Information Diffusion*** in Real Time",IEEE InfoVis 2012

文本数据可视分析案例（一）

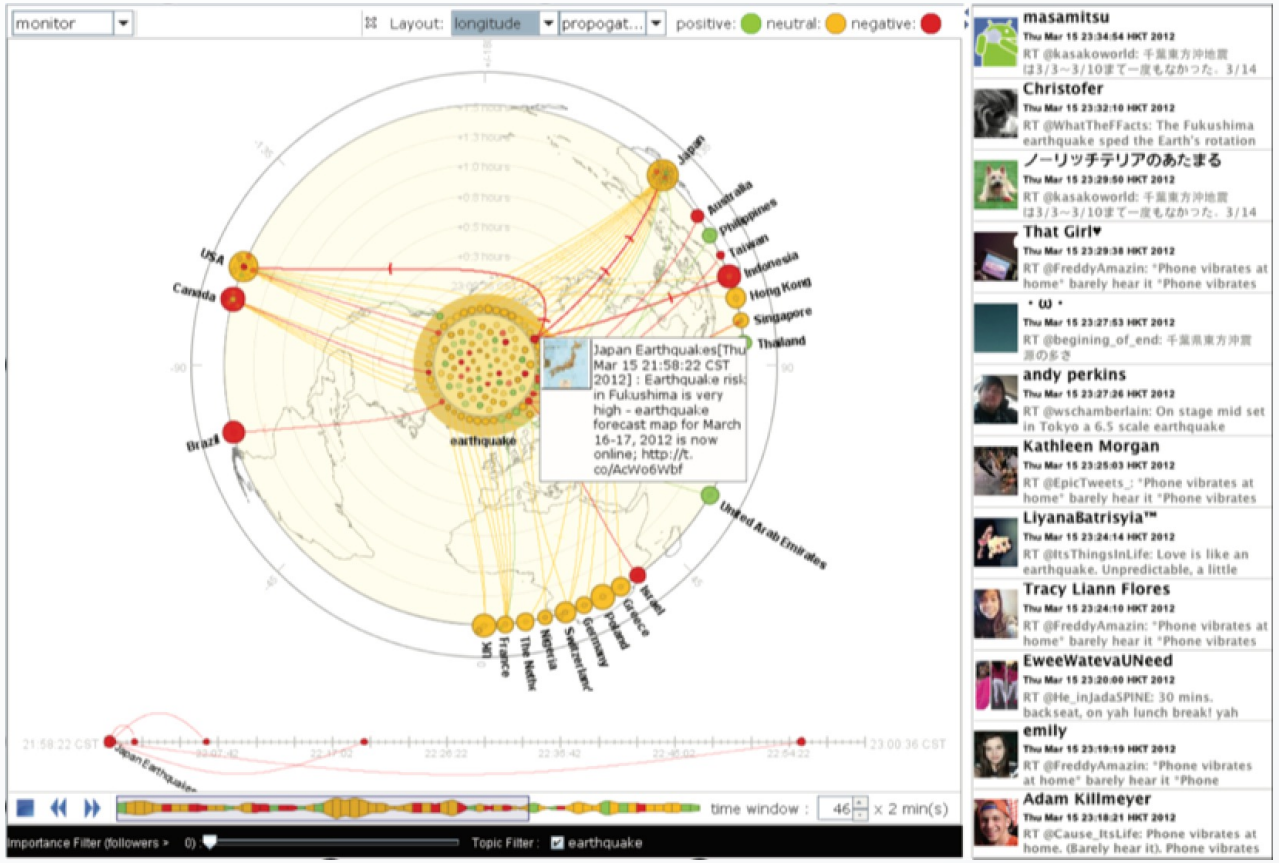
- **文本信息实时扩散 动态监控**
  - 实时感知信息的传播
  - 了解信息传播与主题演化的的动态性



"Whisper: Tracing the Spatiotemporal Process of ***Information Diffusion*** in Real Time",IEEE InfoVis 2012

文本数据可视分析案例（一）

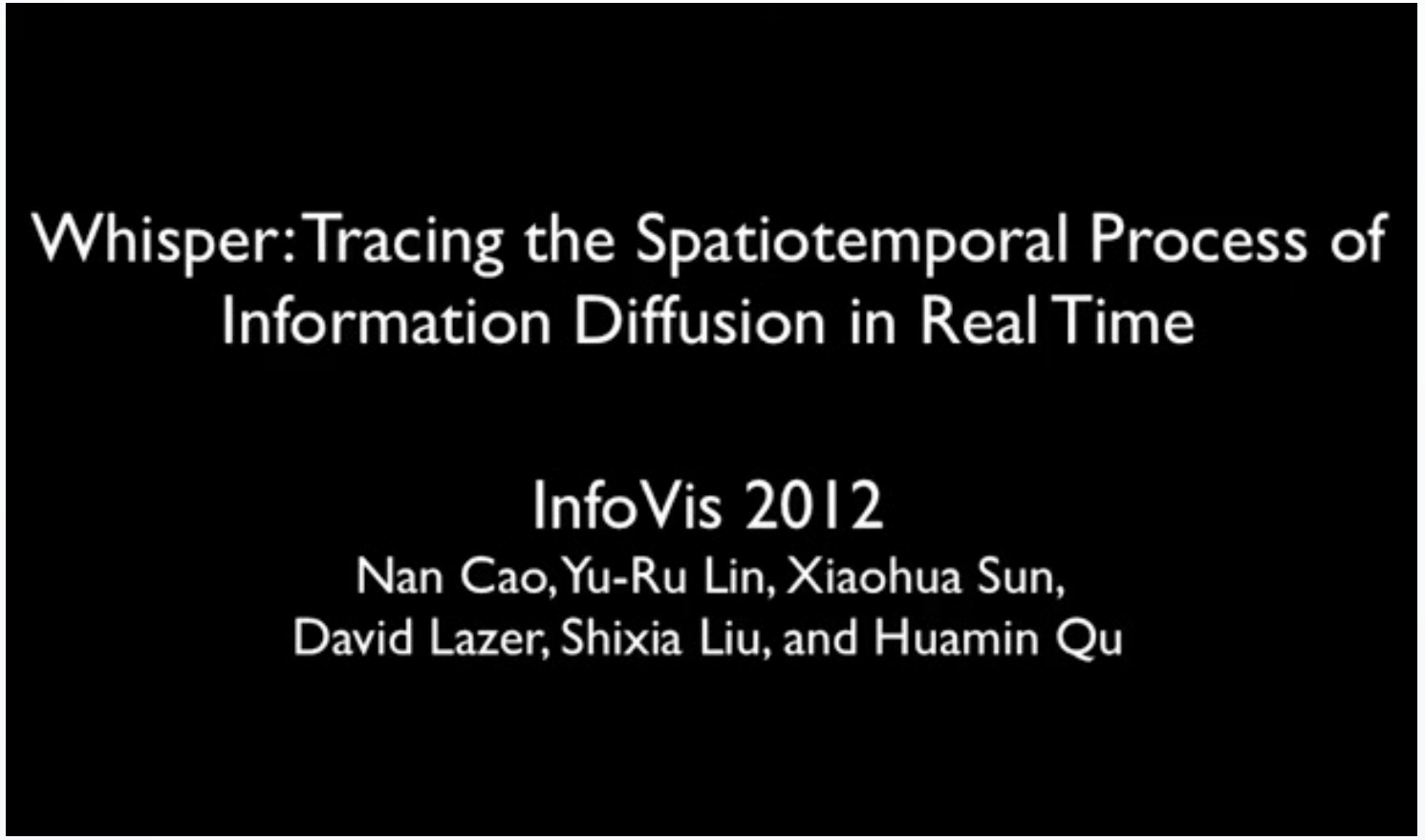
- **文本信息实时扩散 动态监控**
  - 针对热点社会事件的
    实时搜索与追踪
  - 面向信息扩散过程的
    动态捕捉与展示
  - 交互式的数据浏览方式，
    以及直观的数据可视化表达



"Whisper: Tracing the Spatiotemporal Process of ***Information Diffusion*** in Real Time",IEEE InfoVis 2012

文本数据可视分析案例（一）

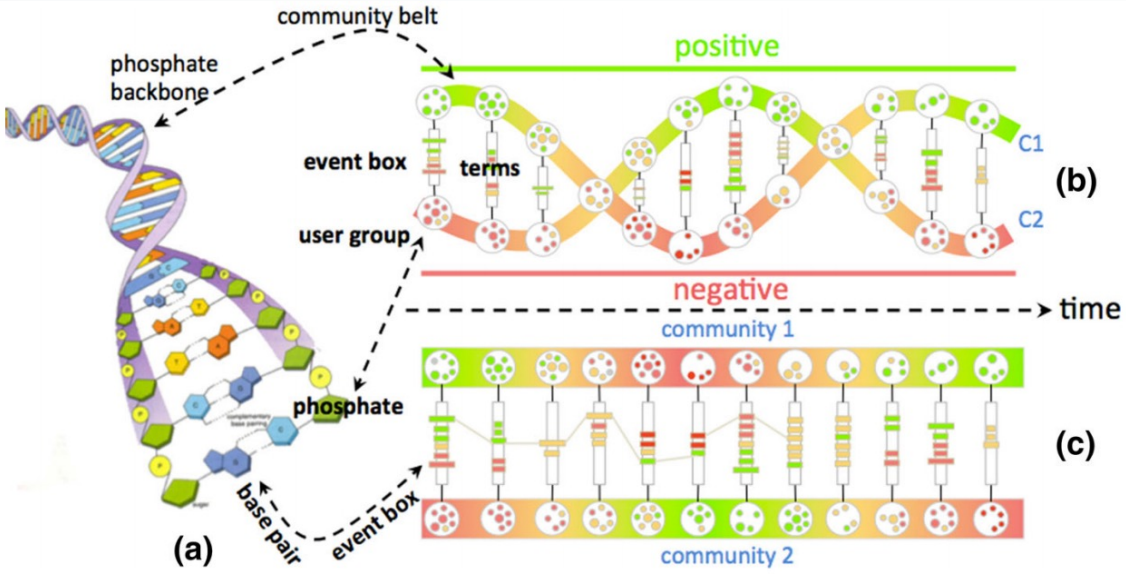
- **文本信息实时扩散 动态监控**

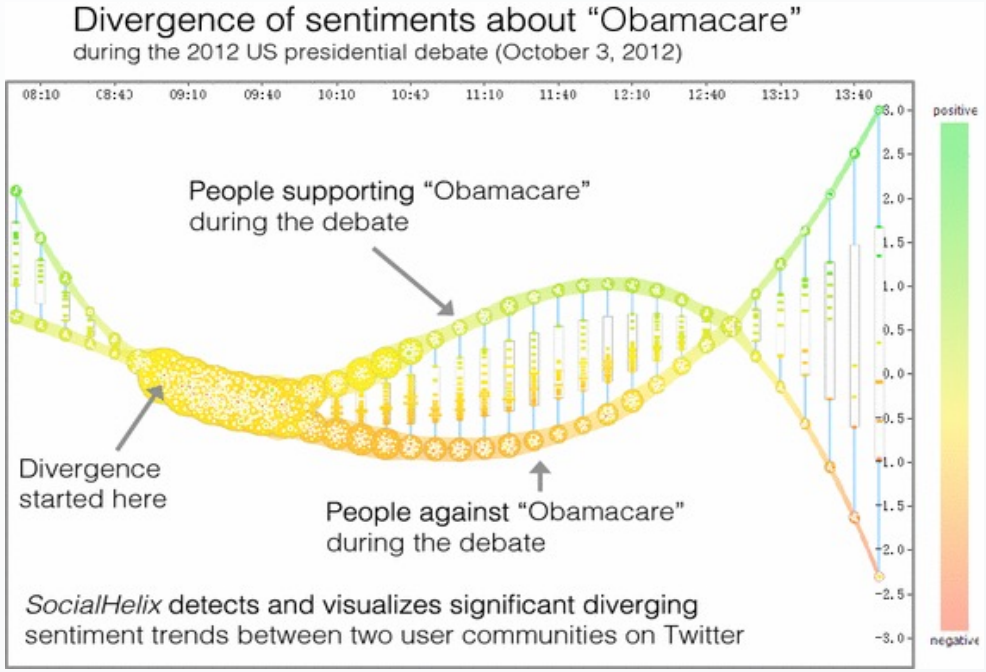


"Whisper: Tracing the Spatiotemporal Process of ***Information Diffusion*** in Real Time",IEEE InfoVis 2012

文本数据可视分析案例（二）

- 分析对立情感的波动变化模式

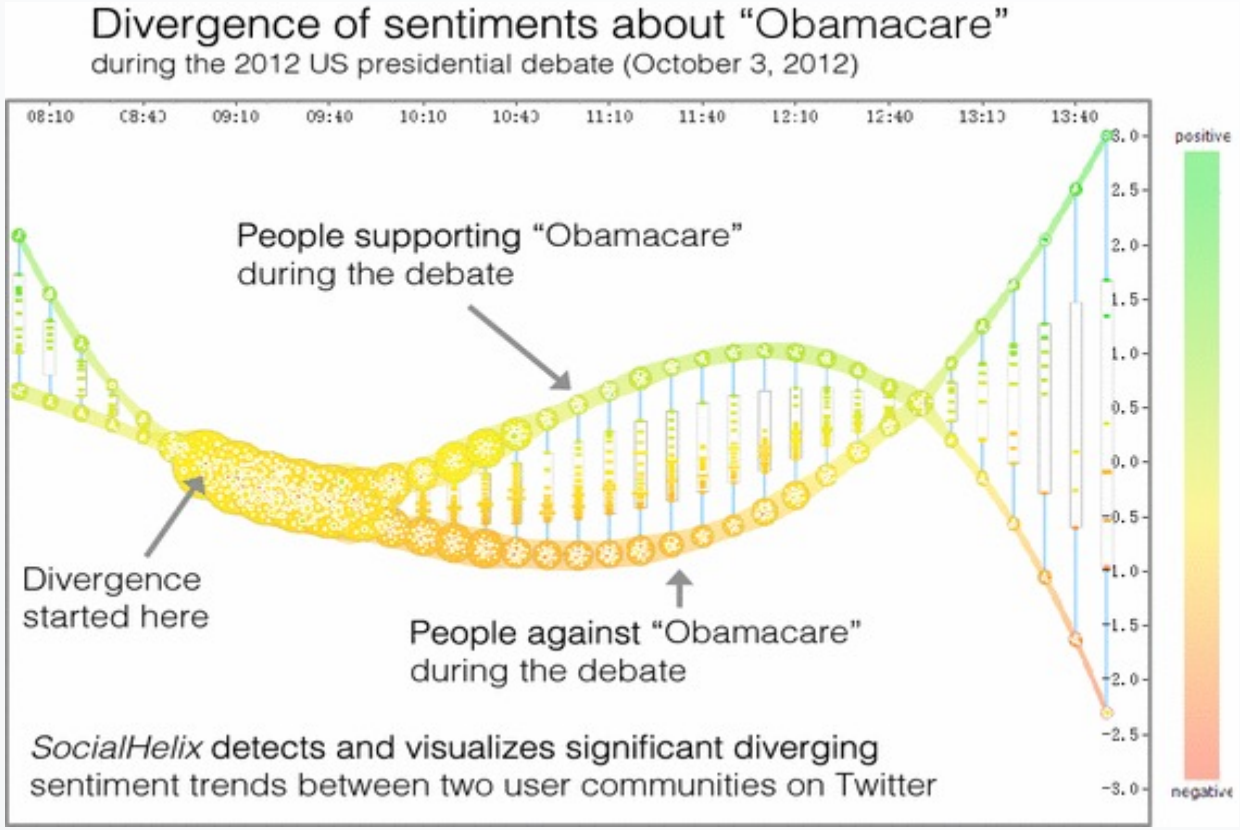




"SocialHelix: Visual Analysis of ***Sentiment Divergence*** in Social Media",
Journal of Visualization, May 2015, Volume 18, Issue 2, pp 221-235

文本数据可视分析案例（二）

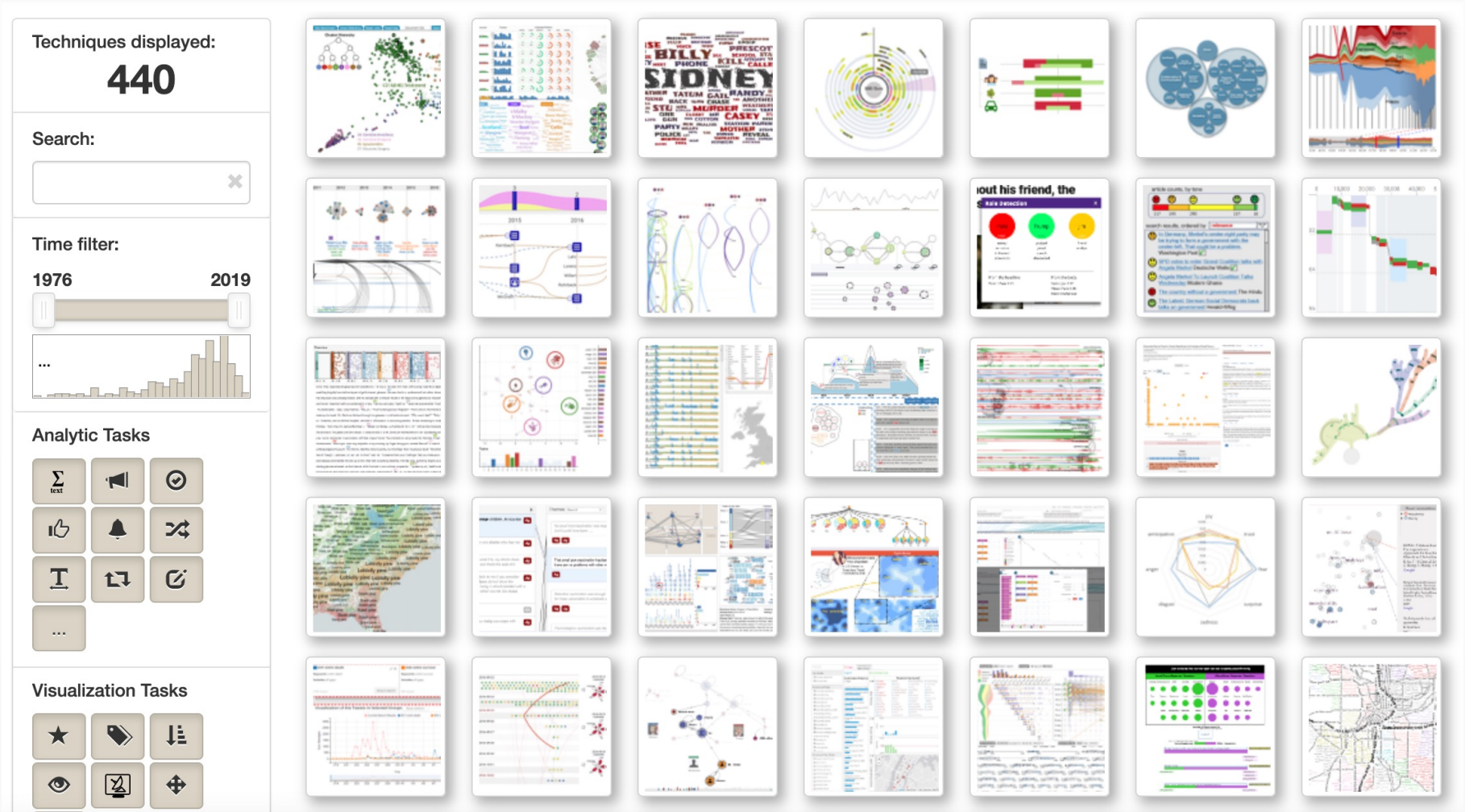
- 分析对立情感的波动变化模式



- 针对特定事件的异常舆情状况检测与分析
- 基于可视化的动态情感数据分析及实时直观展现
- 基于智能交互的异常情感波动模式探索与比对

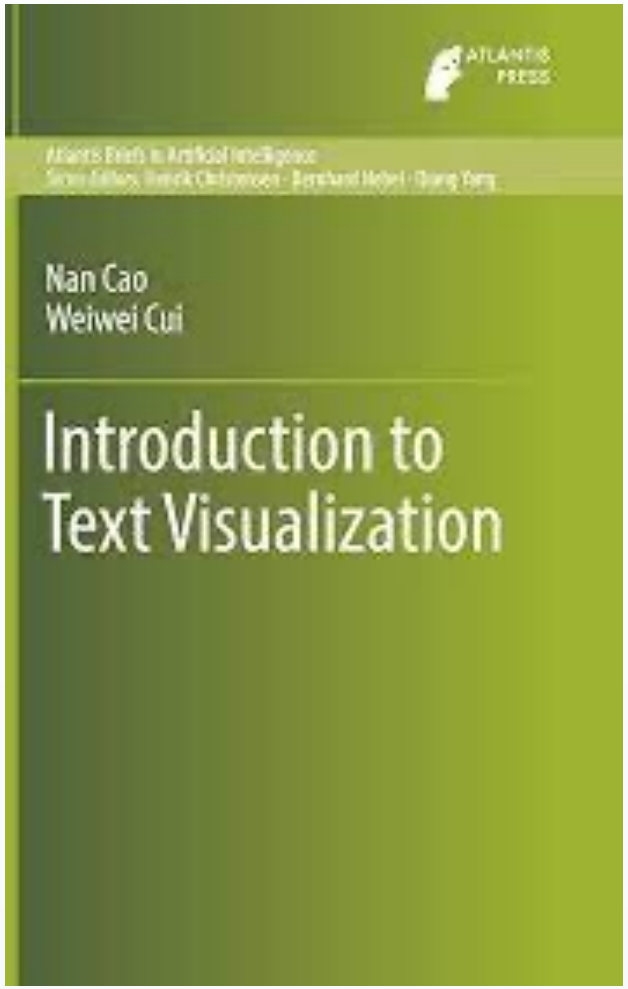
“SocialHelix: Visual Analysis of ***Sentiment Divergence*** in Social Media”，Journal of Visualization 2015

文本数据可视化技术汇总



https://textvis.lnu.se/

文本数据可视化技术汇总



https://link.springer.com/book/10.2991/978-94-6239-186-4

课程大纲

- 什么是文本数据？
- 为什么要对文本数据可视化？
- 文本数据分析技术
- 文本数据可视化技术
  - 可视化文本数据内容
    - 基于关键词的可视化
    - 时序文档的可视化
    - 基于特征分布的可视化
  - 可视化文本数据关联
    - 文本数据集合之间的关系：论文引用、网页超链接、相似性、层级性
    - 可视化方法：图布局、树布局、投影布局
  - 多层级文本数据可视化