HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

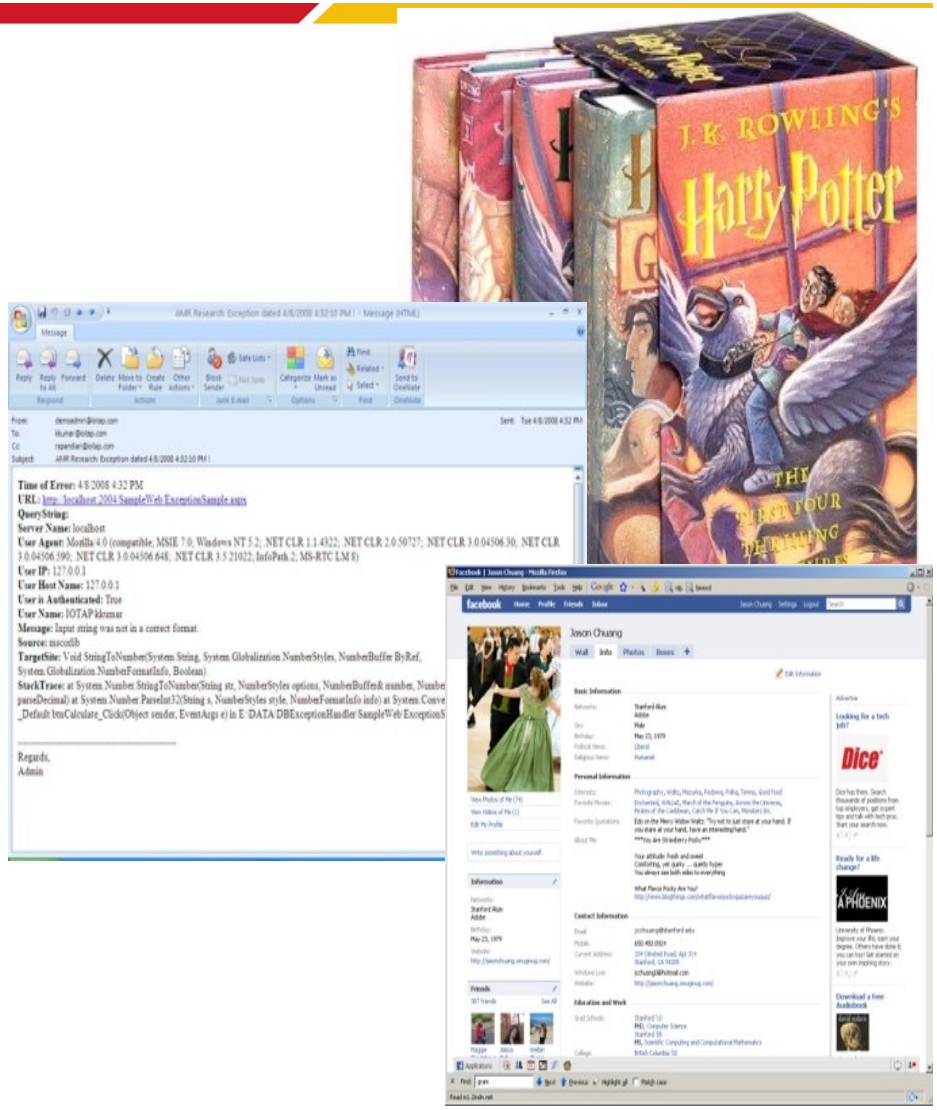
Lecture 16 - Visualization for text data

Visualization goals

- Understanding
  - Get the “gist” of a document
- Grouping
  - Cluster for overview or classification
- Comparison
  - Compare document collections, or inspect evolution of collection over time
- Correlation
  - Compare patterns in text to those in other data, e.g., correlate with social network

Text as data

- Documents
  - Articles, books and novels
    E-mails, web pages, blogs
    Tags, comments Computer programs, logs
- Collections of Documents
  - Messages (e-mail, blogs, tags, comments) Social networks (personal profiles)
    Academic collaborations (publications)



Example: Health care reform

- Recent History
  - Initiatives by President Clinton
  - Overhaul by President Obama
- Text Data
  - News articles
  - Speech transcriptions
  - Legal documents
- What questions might you want to answer?
- What visualizations might help?

A concrete example

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of you

Tag Clouds: Word Count

- President Obama's Health Care Speech to Congress [NY Times]

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

Gulfs of evaluation

- Many text visualizations do not represent the text directly. They represent the output of a language model (word counts, word sequences, etc.).
- Can you interpret the visualization?
  - How well does it convey the properties of the model?
- Do you trust the model?
  - How does the model enable us to reason about the text?

Text visualization challenges

- High Dimensionality
  - Where possible use text to represent text…
  - … which terms are the most descriptive?
- Context & Semantics
  - Provide relevant context to aid understanding.
  - Show (or provide access to) the source text.
- Modeling Abstraction
  - Determine your analysis task.
  - Understand abstraction of your language models. Match analysis task with appropriate tools and models.

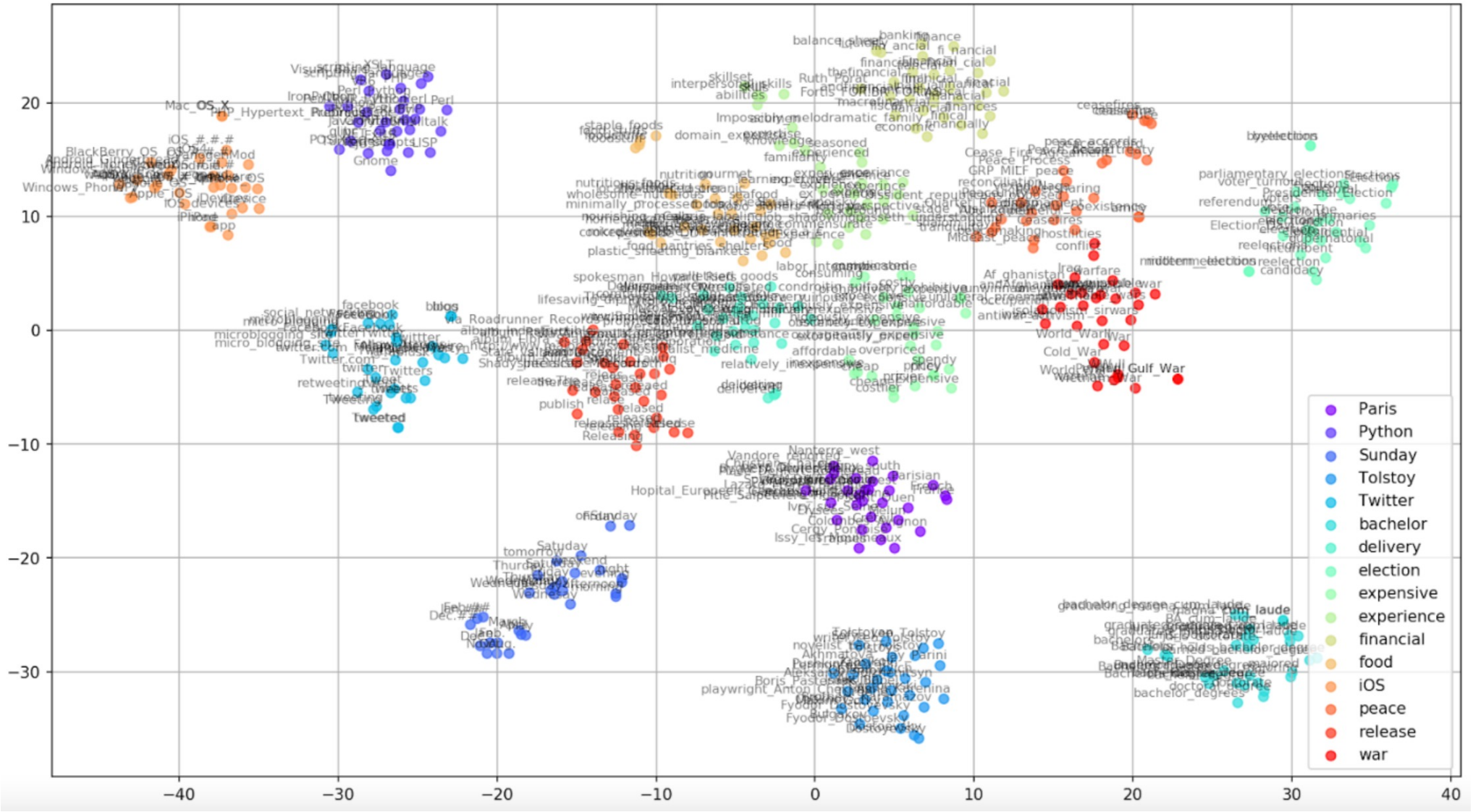
Topics

- Word visualization
- Topic model visualization
- Other visualization techniques

Word visualization

- Number of words is massive
- Words have meanings and relations
  - Correlations: Hong Kong, San Francisco, Bay Area
  - Order: April, February, January, June, March, May
  - Membership: Tennis, Running, Swimming, Hiking, Piano
  - Hierarchy, antonyms & synonyms, entities, …

Word embeddings



Text processing pipeline

- Tokenization
  - Segment text into terms.
  - Remove stop words? a, an, the, of, to, be
  - Numbers and symbols? #gocard, @stanfordfball, Beat Cal!!!!!!!!
  - Entities? San Francisco, O'Connor, U.S.A.
- Stemming
  - Group together different forms of a word.
  - Porter stemmer? visualization(s), visualize(s), visually -> visual
  - Lemmatization? goes, went, gone -> go
- Ordered list of terms

Bag of words model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
  - Each dimension corresponds to a term (10,000+)
  - Each value represents the relevance
  - For example, simple term counts
- Aggregate into a document-term matrix
  - Document vector space model

Document-Term Matrix

- Each document is a vector of term weights
- Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Word count

http://wordcount.org

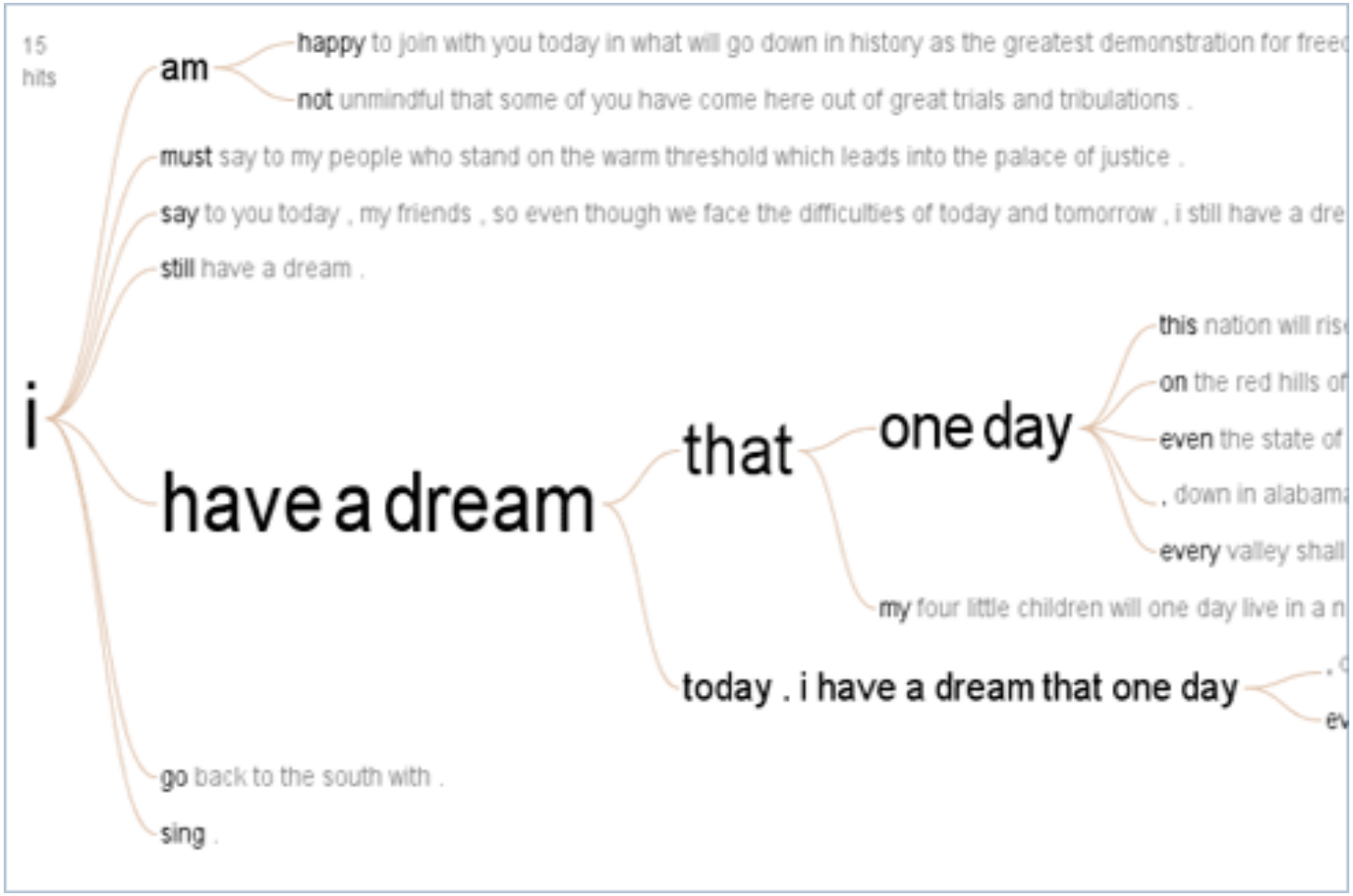
Word cloud


Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech
Creator: Anonymous
Tags:
Edit Language Font Layout Color
McCain
America
country
man
people
just
one
John
president
Washington
nominee
election
going
American
state
good
reform
energy
things
Americans
small
Senator
interests
oil
will
like
world
better
fellow
special
taxes
citizens
might
change
office
opponent
raise
way
back
Alaska
Congress
States
ones
families
kind
great
serve
listening
family
proud
tax
wants
make
put
politics
next
lobbyists
seek
right
need
control
gas
Iraq
United
children
grow
much
fought
always
pipeline
time
fight
years
came
can
go
us
vice
Senate
governor's
brought
spending
door
business
power
even
towns
day
Data file: Sarah Palin speaks at the Republican National Convention, 9/3/2008 Data source: SFGate / AP This data set has not yet been rated

N-gram cloud

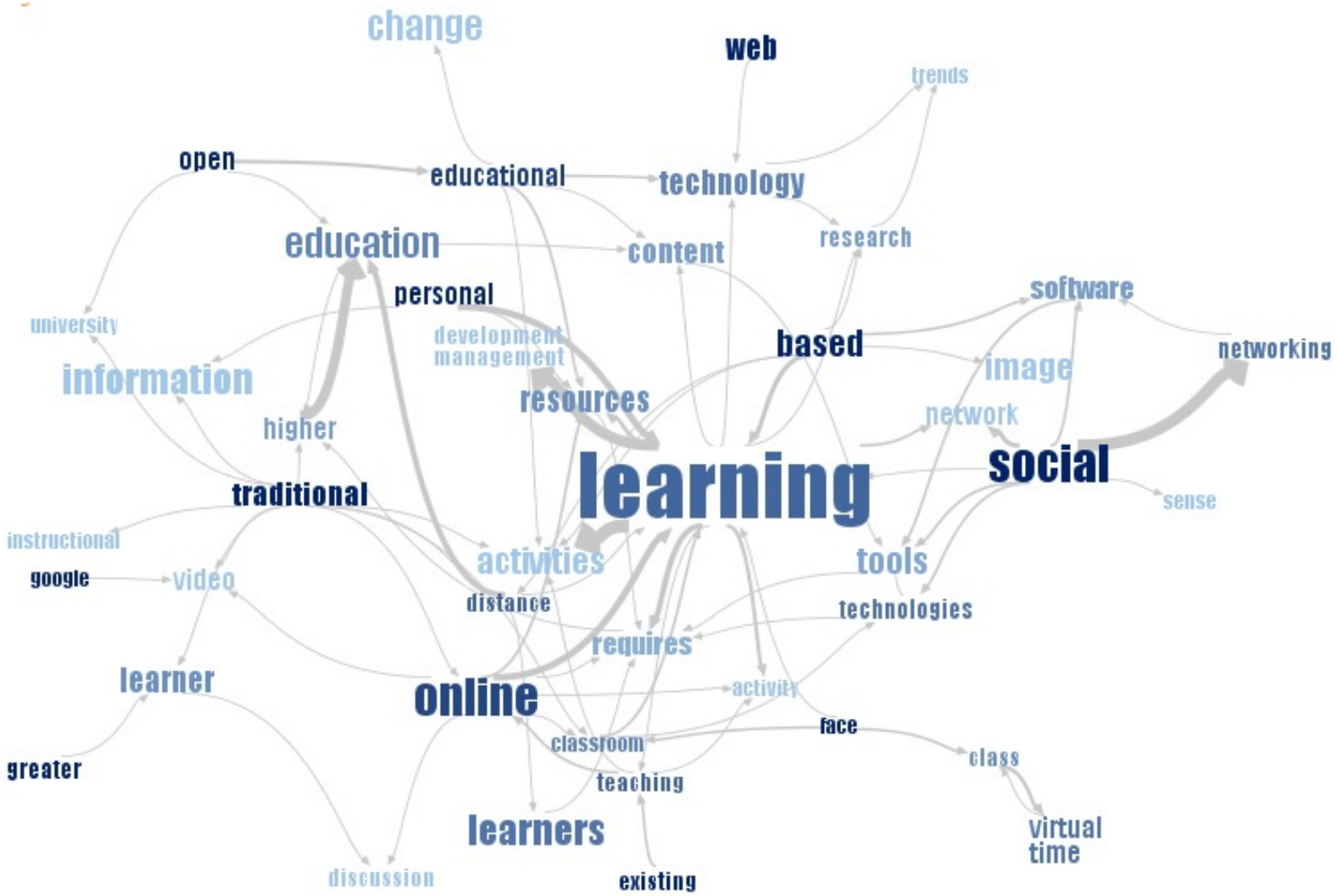

Most popular phrases in 1980
made known the engagement of
took place yesterday at the
previous marriage ended in divorce
in the New York law
Her father is president of
was graduated magna cum laude
will keep her name professionally
a JD degree from the
& Company in New York
and chief executive officer of
the late Mr and Mrs
a daughter of Mr and
performed the ceremony at the
an associate in the
is a granddaughter of
son of Mr and Mrs
New York law firm of
matron of
by the Rev
who will retain her name
parents of the future bride
The future bride is a
a graduate
The wedding is planned for
was his brothers best man
was graduated from
best man for his brother
Episcopal Church in
was maid of honor for
and cum laude from
bride who will keep her
of the Junior League of
bridegroom is a grandson of
the Wharton School of the
and magna cum laude from
Announcement has been made by
elected to Phi Beta Kapha
of the University of Pennsylvania
received a masters degree from
daughter of Dr and Mrs
received an MBA degree from
the engagement of their daughter
is a vice president of
father is chairman of the
of honor for her sister
father is executive vice president
is a partner in the
the University of California at
Roman Catholic Church
Graduate School of Business Administration
daughter also of the late
an alumnus of
father who is retired was

WordTree

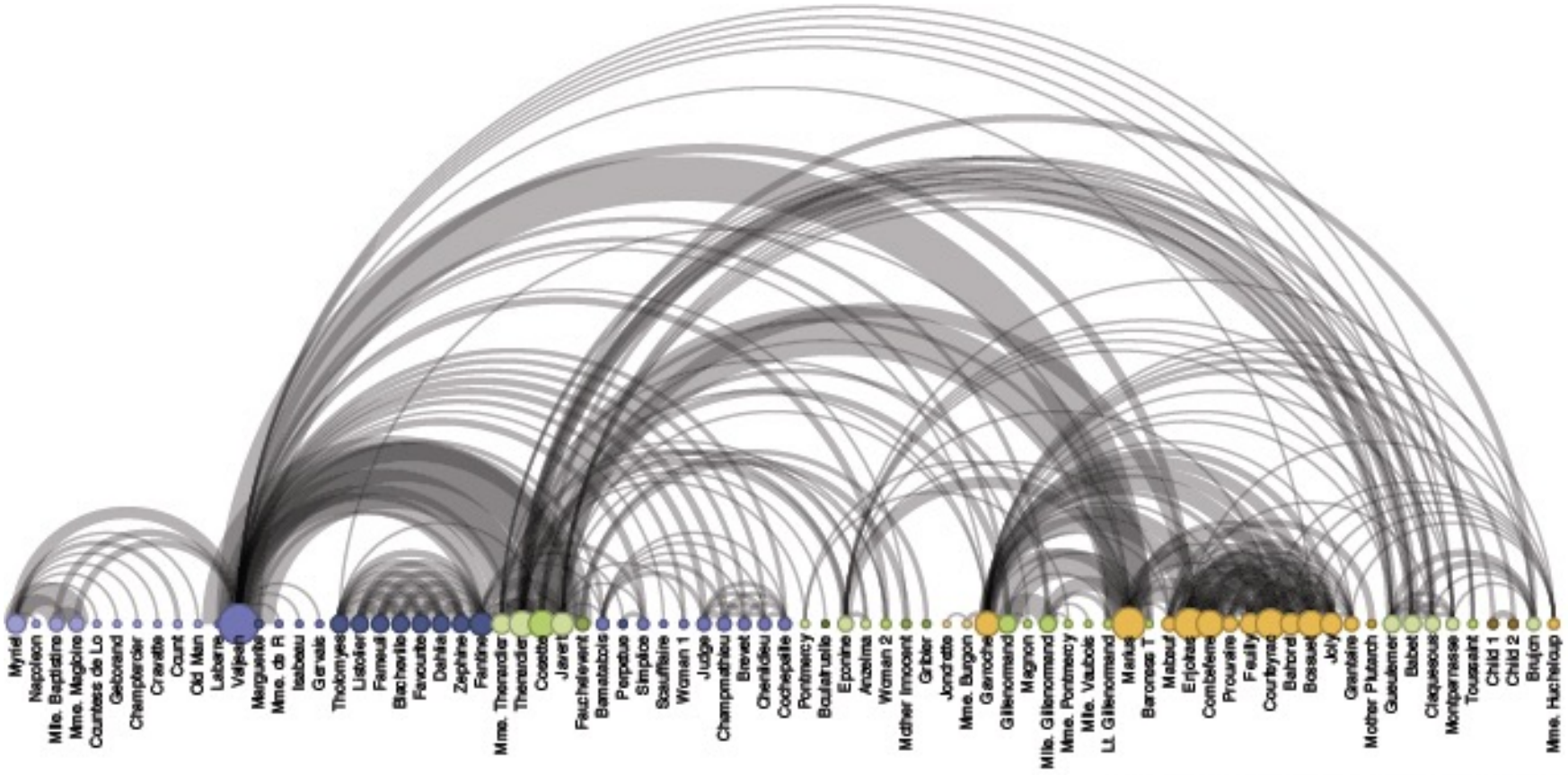


Wattenberg, Martin, and Fernanda B. Viégas. "The word tree, an interactive visual concordance." *IEEE transactions on visualization and computer graphics* 14.6 (2008): 1221-1228. http://hint.fm/papers/wordtree_final2.pdf

PhraseNet


change
web
trends
open
educational
technology
education
content
research
personal
software
university
development
management
based
networking
information
image
resources
network
higher
learning
social
traditional
sense
instructional
activities
tools
google
video
distance
technologies
requires
learner
online
activity
face
classroom
class
greater
teaching
learners
virtual
time
discussion
existing

Arc diagram

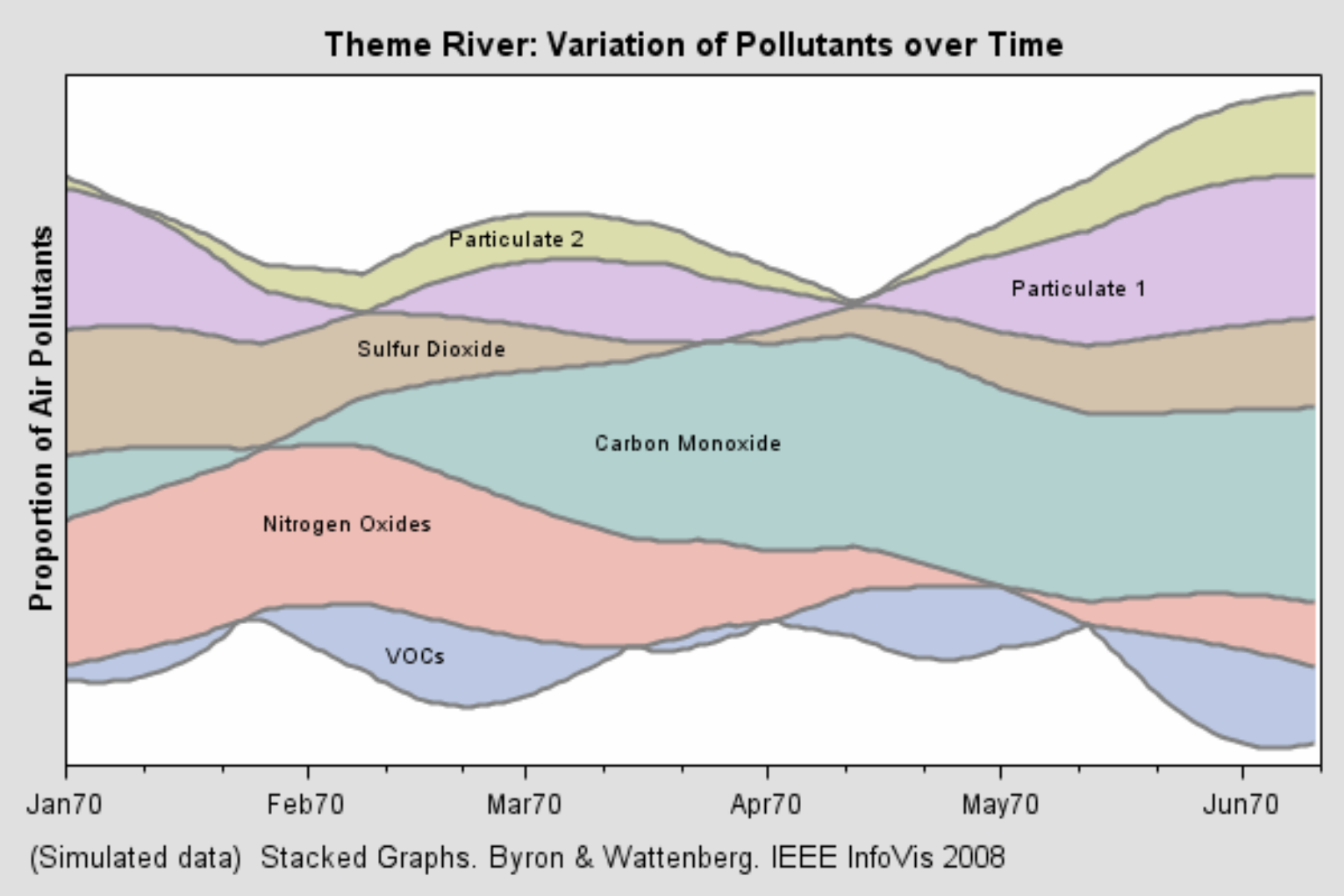


http://hci.stanford.edu/jheer/files/zoo/ex/networks/arc.html

Visualization for document topics

- Topic visualization problem
  - Large amount of documents
  - Need to visualize the general topics of the documents
- Method
  - Static: ContentTour, IN-SPIRE
  - Dynamic: ThemeRiver, RoseRiver
  - Hidden: Termite

ThemeRiver


Theme River: Variation of Pollutants over Time
Proportion of Air Pollutants
Particulate 2
Particulate 1
Sulfur Dioxide
Carbon Monoxide
Nitrogen Oxides
VOCs
Jan70
Feb70
Mar70
Apr70
May70
Jun70
(Simulated data) Stacked Graphs. Byron & Wattenberg. IEEE InfoVis 2008

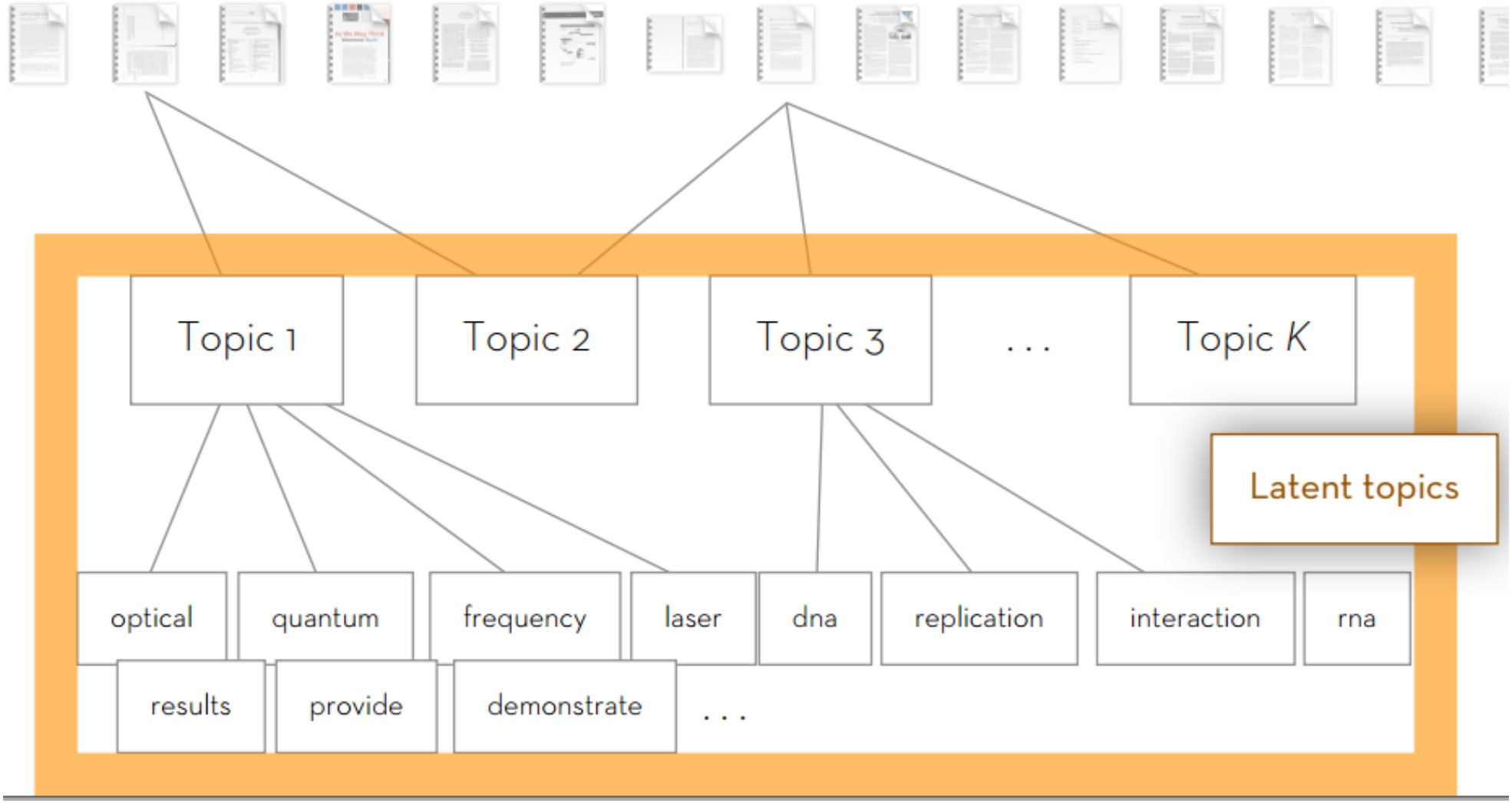
Termite: Topic model visualization

- http://vis.stanford.edu/topic-diagnostics/model/silverStandards/
- Termite
  - Overview of topics
  - Words in a topic
  - Identifying topics that don't help
  - Compare topics

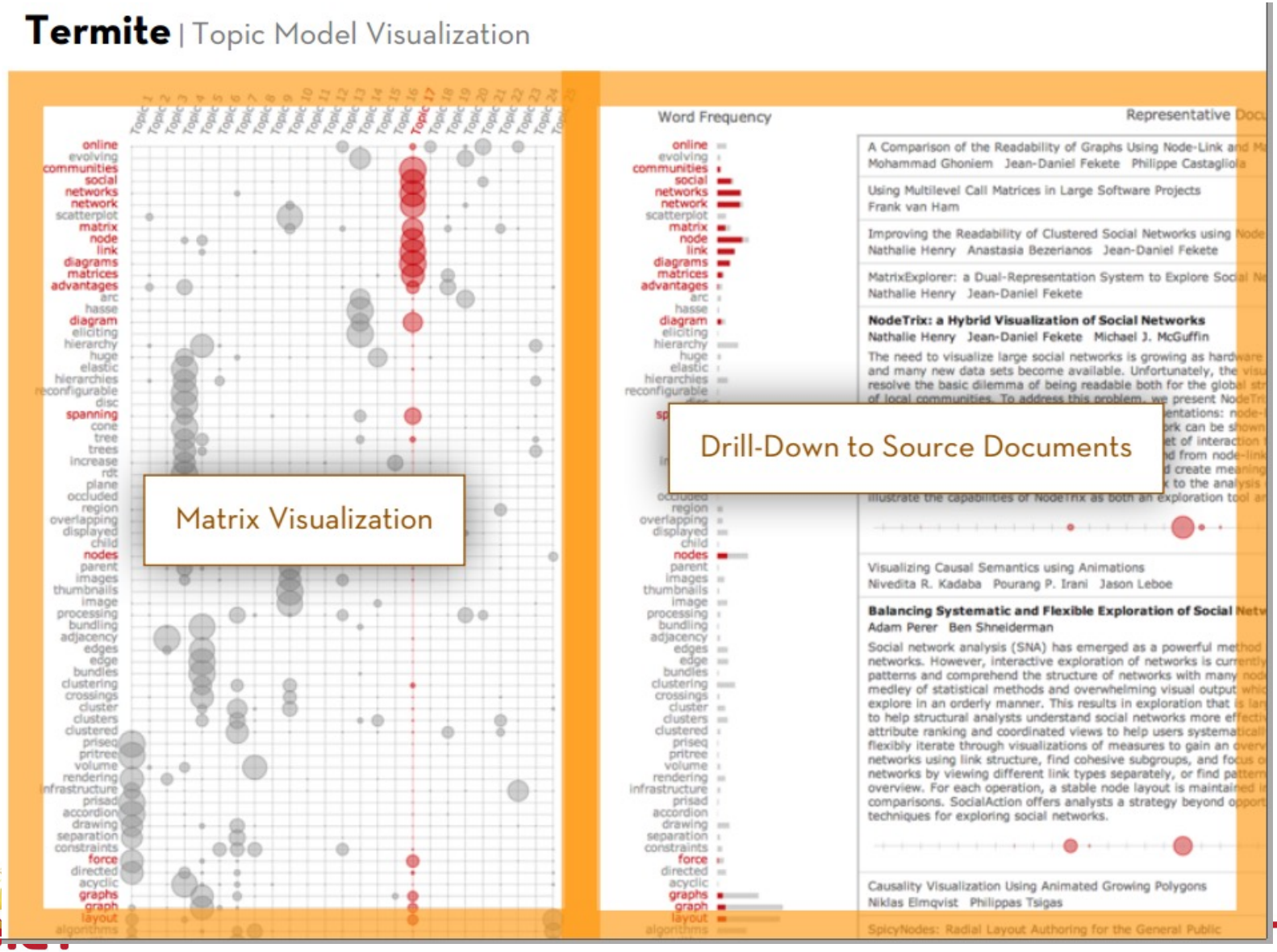
Topic model visualization

- Filter words in a topic
  - Words that appear frequently are not necessarily useful because they are not distinctive
  - `saliency(w) = frequency(w) x distinctiveness(w)`
- Order words in topics
  - Cluster related words
  - Preserve the order of appearance in the document
  - Similarities between the words

Latent topics

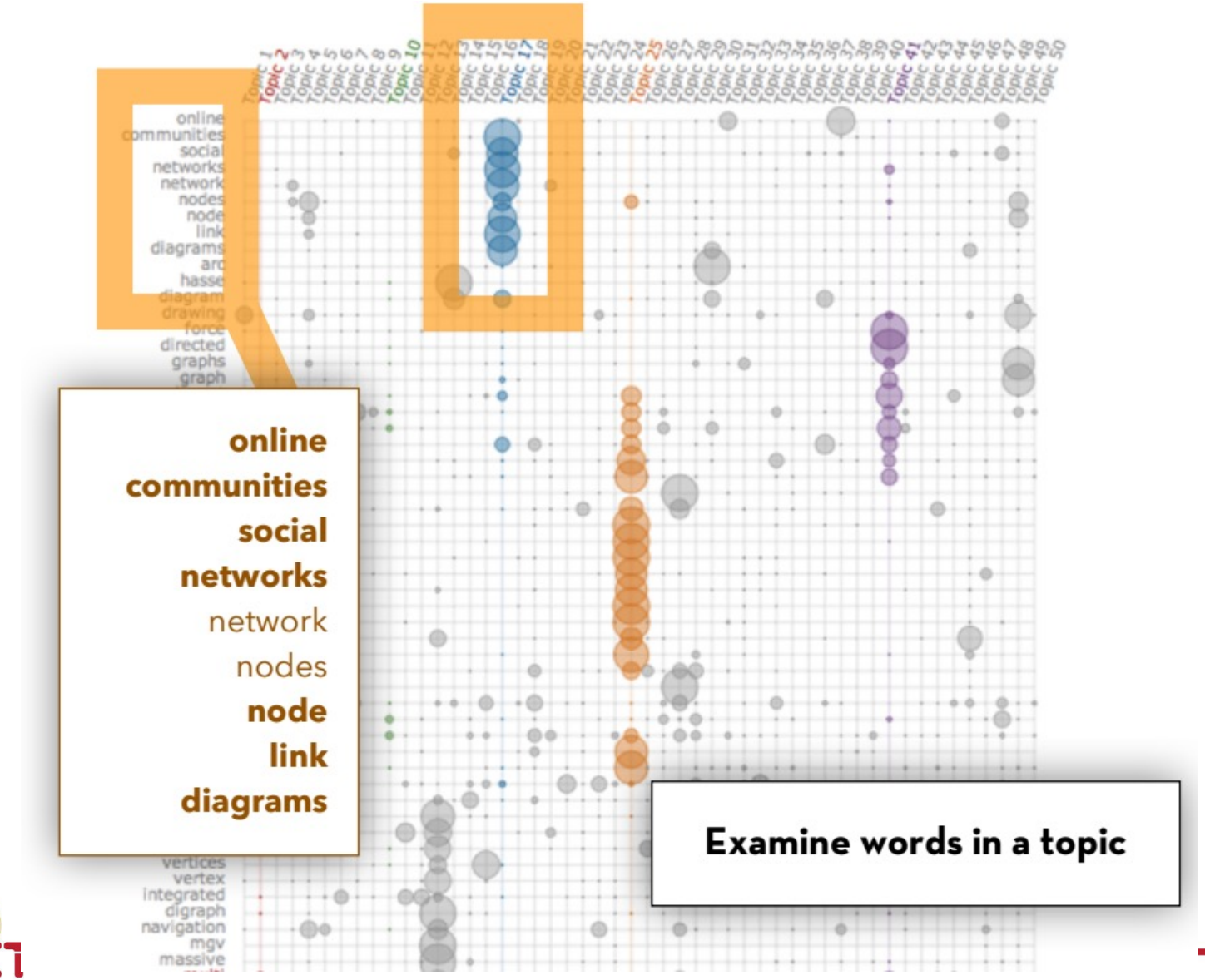


Topic model visualization

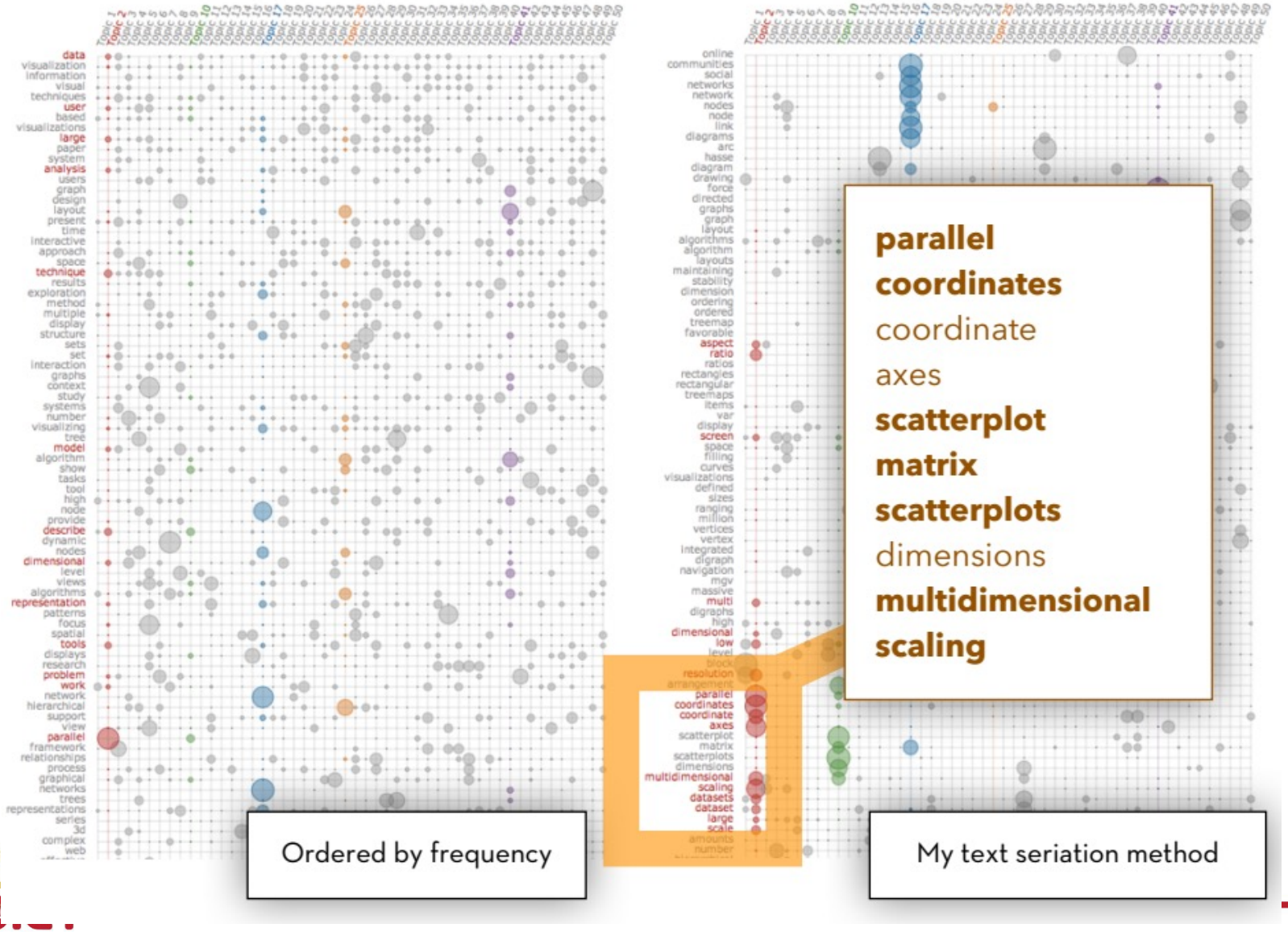


Topic model visualization

Topic model visualization



Sắp xếp từ trong chủ đề

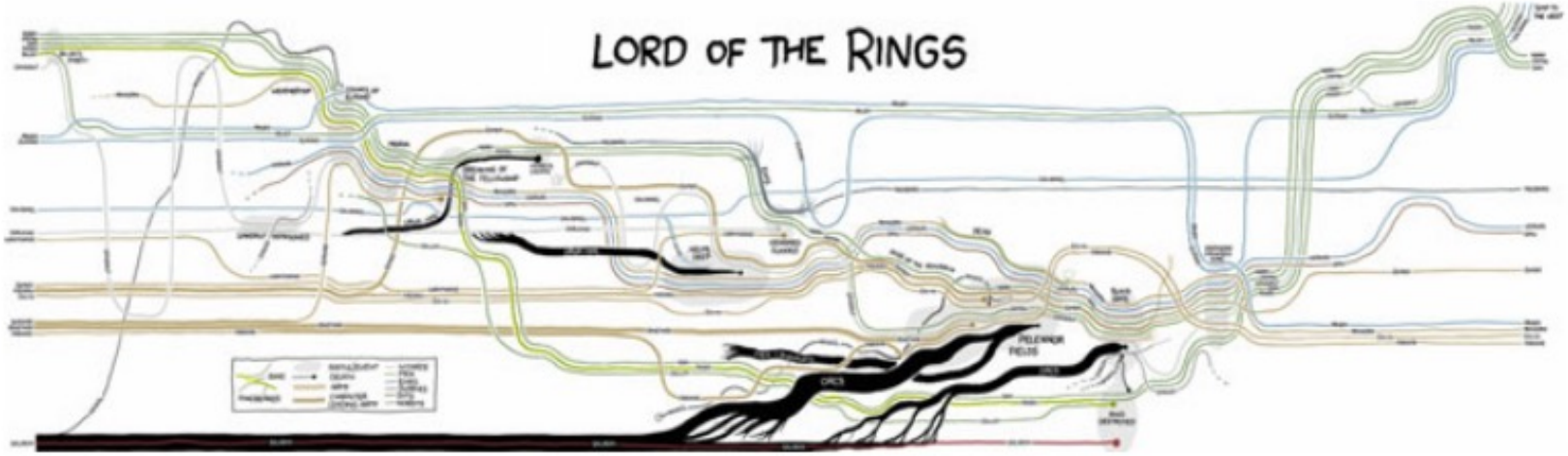


Other visualization techniques

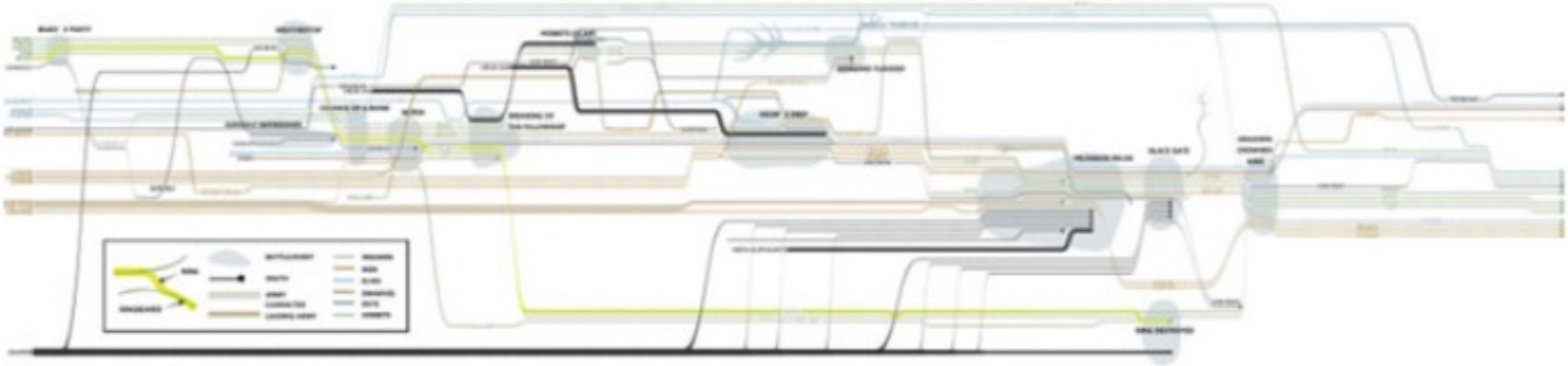
- Visualize event flows
- Emotion visualization
- Document discovery

StoryFlow: Tracking the Evolution of Stories

Liu, Shixia, et al. "Storyflow: Tracking the evolution of stories." *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013)

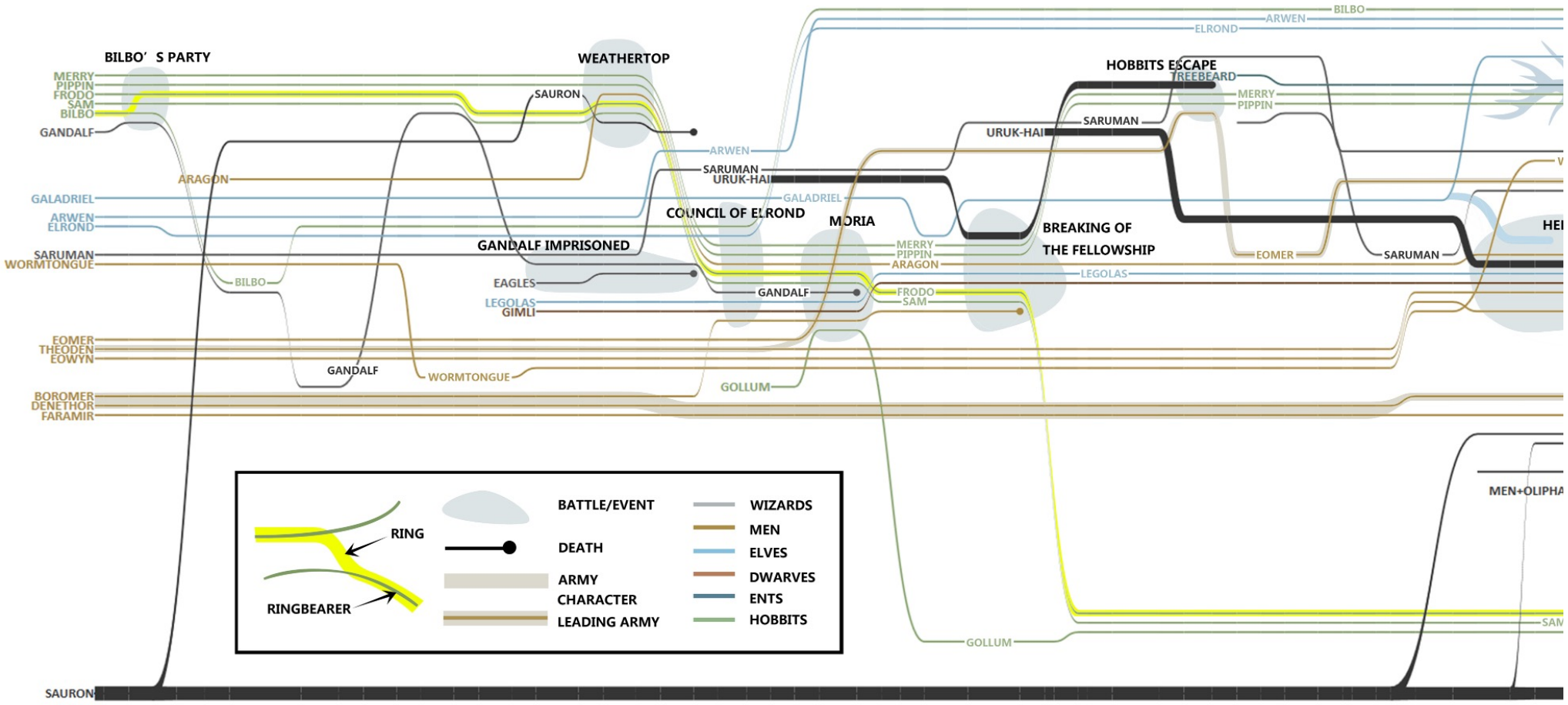


Hand made by XKCD

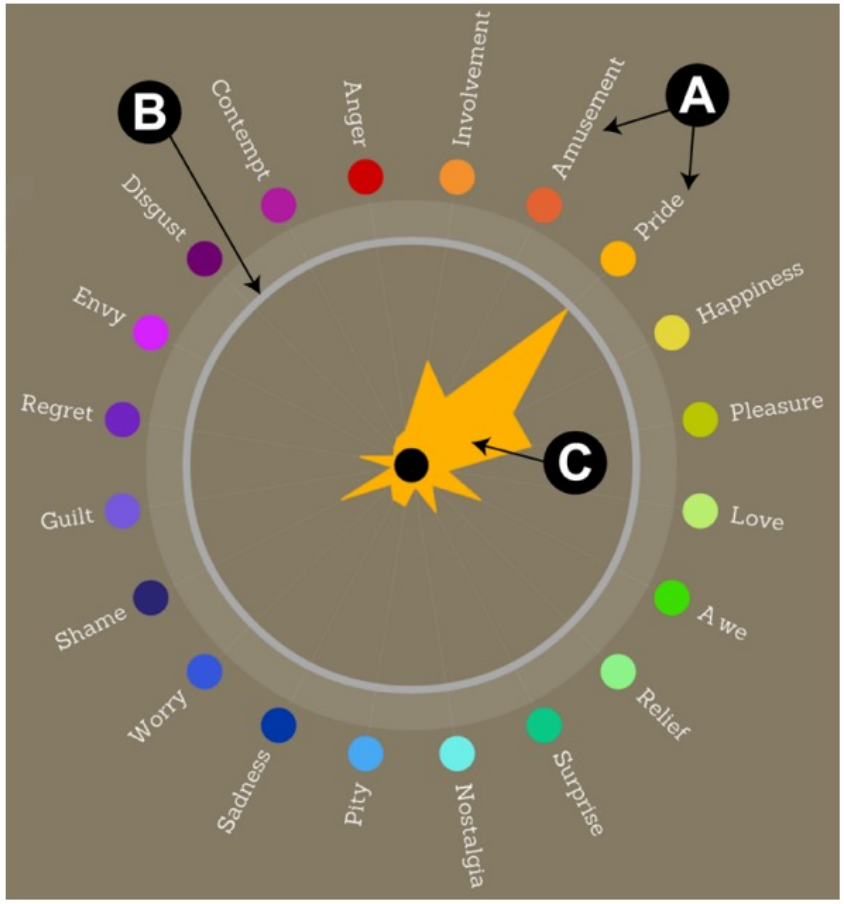


Generation by StoryFlow

Storyflow: Lord of the rings



EmotionWatch



Kempter, Renato, et al. "EmotionWatch: Visualizing fine-grained emotions in event-related tweets." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. No. 1. 2014.

Figure 1: The emotion wheel. A - GEW emotion categories; B - Number of tweets visualized as the ring width; C - Emotion shape visualizing the emotional profile as a star plot.

Document Card

Strobelt, Hendrik, et al. "Document cards: A top trumps visualization for documents." *IEEE transactions on visualization and computer graphics* 15.6 (2009).

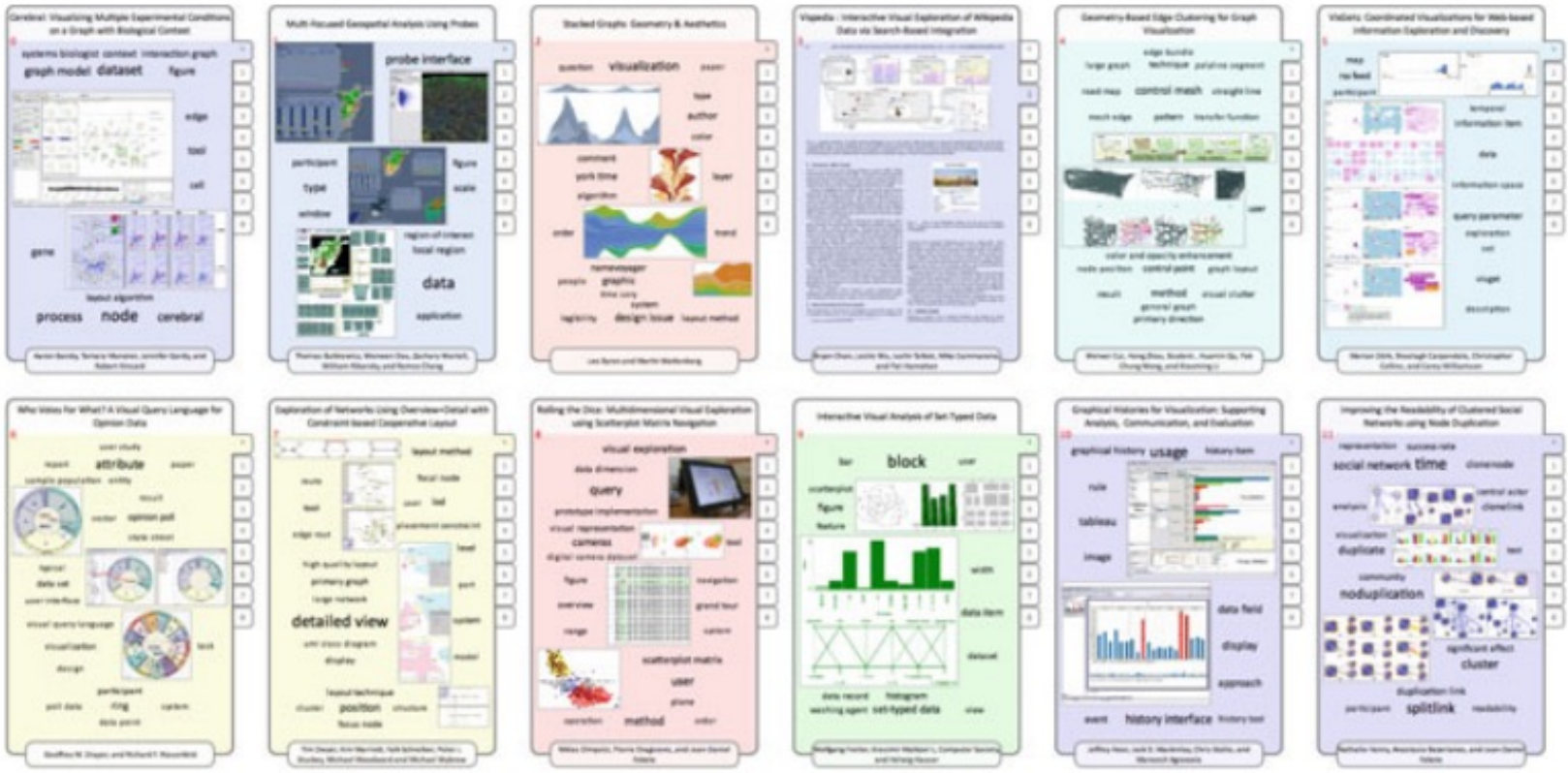


Fig. 2.5 Summarization of the IEEE InfoVis 2008 proceedings corpus in Document Cards (a portion). Referring to [98] for the complete visual summarization of the whole proceeding

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!

soict.hust.edu.vn/ fb.com/groups/soict