G53FIV: Fundamentals of Information Visualization

Lecture 11: Visualizing Text and Documents

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

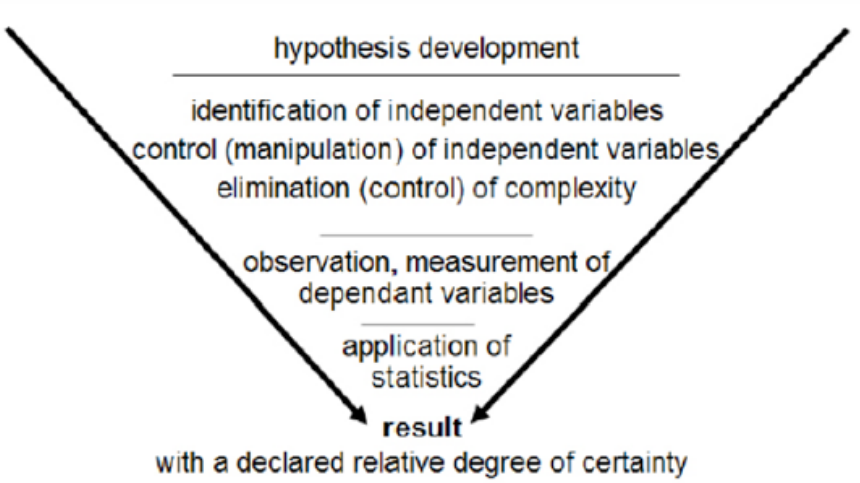
https://moodle.nottingham.ac.uk/course/view.php?id=68644

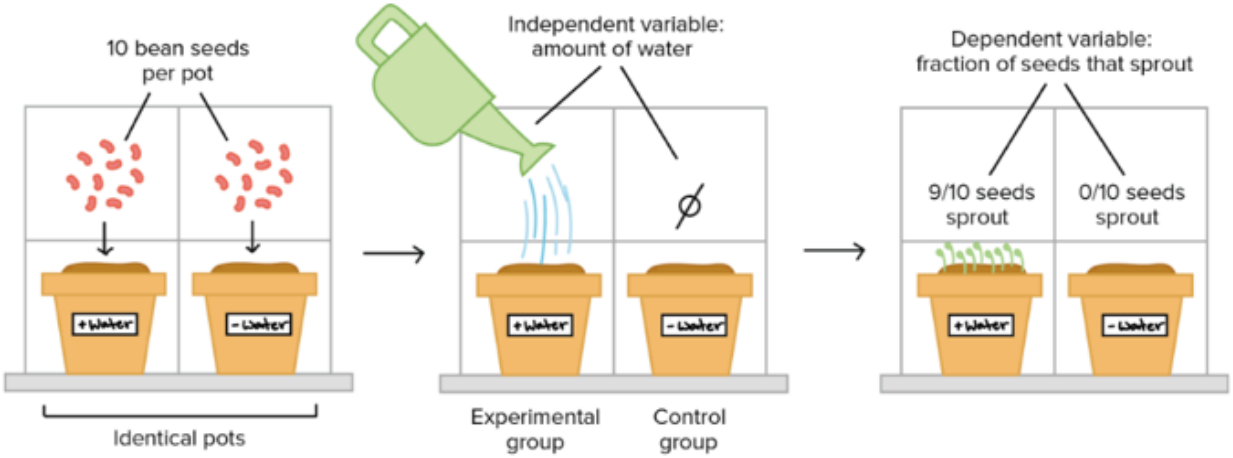
Last Lecture

Evaluation

Quantitative Methods: Controlled Experiments

- Good for measuring performance or comparing multiple techniques
- What do we measure?
  - Performance, time, errors, …





An Example: Controlled Experiment

- Run an experiment to evaluate three visualization strategies
  - Animation
  - Small multiples
  - Traces
- Especially interested in examining whether animated bubble charts are beneficial for analysis and presentation

Do you remember Hans Rosling's TED talk? (Lecture 2)

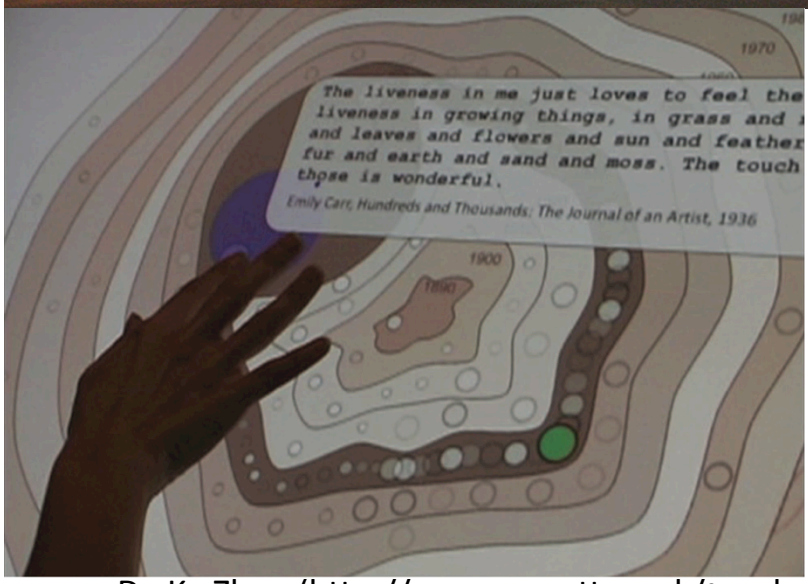
Qualitative Methods

- Types
  - Nested methods
    - Experimenter observation, think-aloud protocol, collecting participant opinions
  - Inspection evaluation methods
    - Heuristics to judge
- Observational context
  - In situ, laboratory, participatory
  - Contextual interviews is important

An Example: Subjective Assessments

- Evaluating a newly developed visualization system (EMDialog) at the museum (Emily Carr exhibit)

- Discourse visualization
  - Time
  - Context



Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Methodology vs. Desirable Features

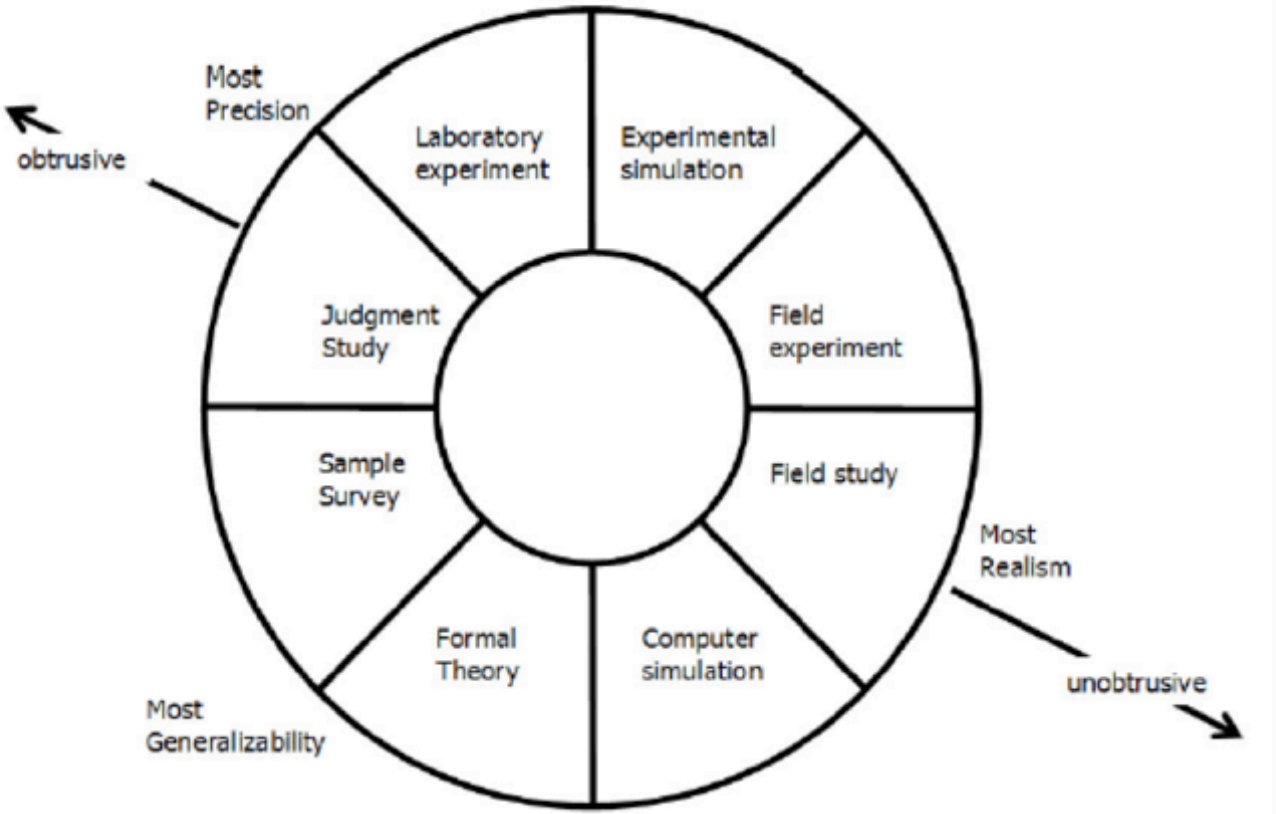
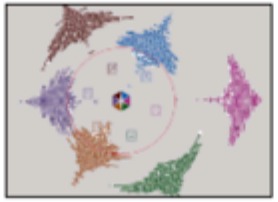


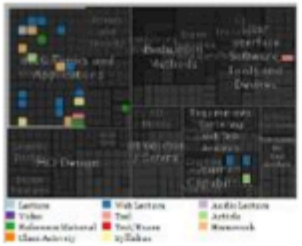
Fig. 1. Types of methodologies organized to show relationships to precision, generalizability and realism. (adapted, simplified from McGrath 1995)

Overview

Visualizing text
Showing words, phrases, and sentences



Visualization for IR
Helping search



- Content
- Context
- Relationship to others

Why Visualize Text?

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

- **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

Challenges

- High Dimensionality
  - Where possible use text to represent text...
  - ... which terms are the most descriptive?
- Context and 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.

Visualizing Text

How do we represent the words, phrases, and sentences in a document or set of documents?

Visualizing text
Showing words, phrases, and sentences

An Example

- Health care speech transcripts
  - Clinton in 1993
  - Obama in 2009

- What questions might you want to answer?
- What visualizations might help?

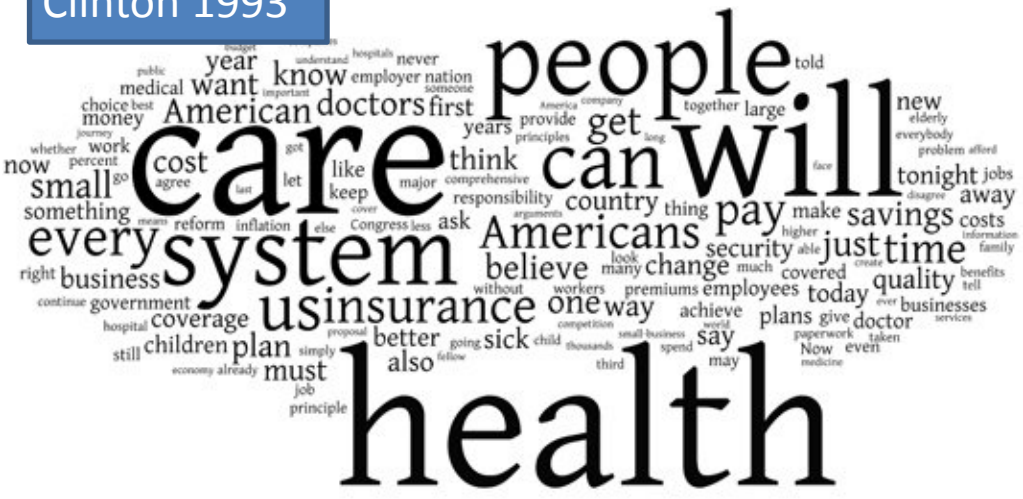
Bill Clinton on Health Care, 1993

Now we are in a time of profound change and opportunity — the end of the cold war, **the information age, and the global economy have brought us both opportunity and hope and strife and uncertainty. Our purpose in this dynamic age must be to change — to make change our friend and not our enemy.**

To achieve that goal, we must face all our challenges with confidence, with faith and with discipline, whether we're reducing the deficit, creating tomorrow's jobs and training our people to fill them, converting from a high-tech defense to a high-tech domestic economy, expanding trade, reinventing government, making our streets safer, or rewarding work over idleness, all of these challenges require us to change.

Tag Clouds: Word Count

Clinton 1993



Obama 2009

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

Word Tree: Word Sequences

Language Model

- 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?

Challenges

- High Dimensionality
  - Where possible use text to represent text...
  - ... which terms are the most descriptive?
- Context and 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.

Words as nominal data?

- High dimensional (10,000+)

- 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, ...

Text Processing Pipeline

- Tokenization
  - Segment text into terms.
  - Remove stop words? a, an, the, of, to, be
  - Numbers and symbols? #gocard, @nottinghamforestfball
  - Entities? Nottingham, Trump.
- 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

Content

Bag of Words Model

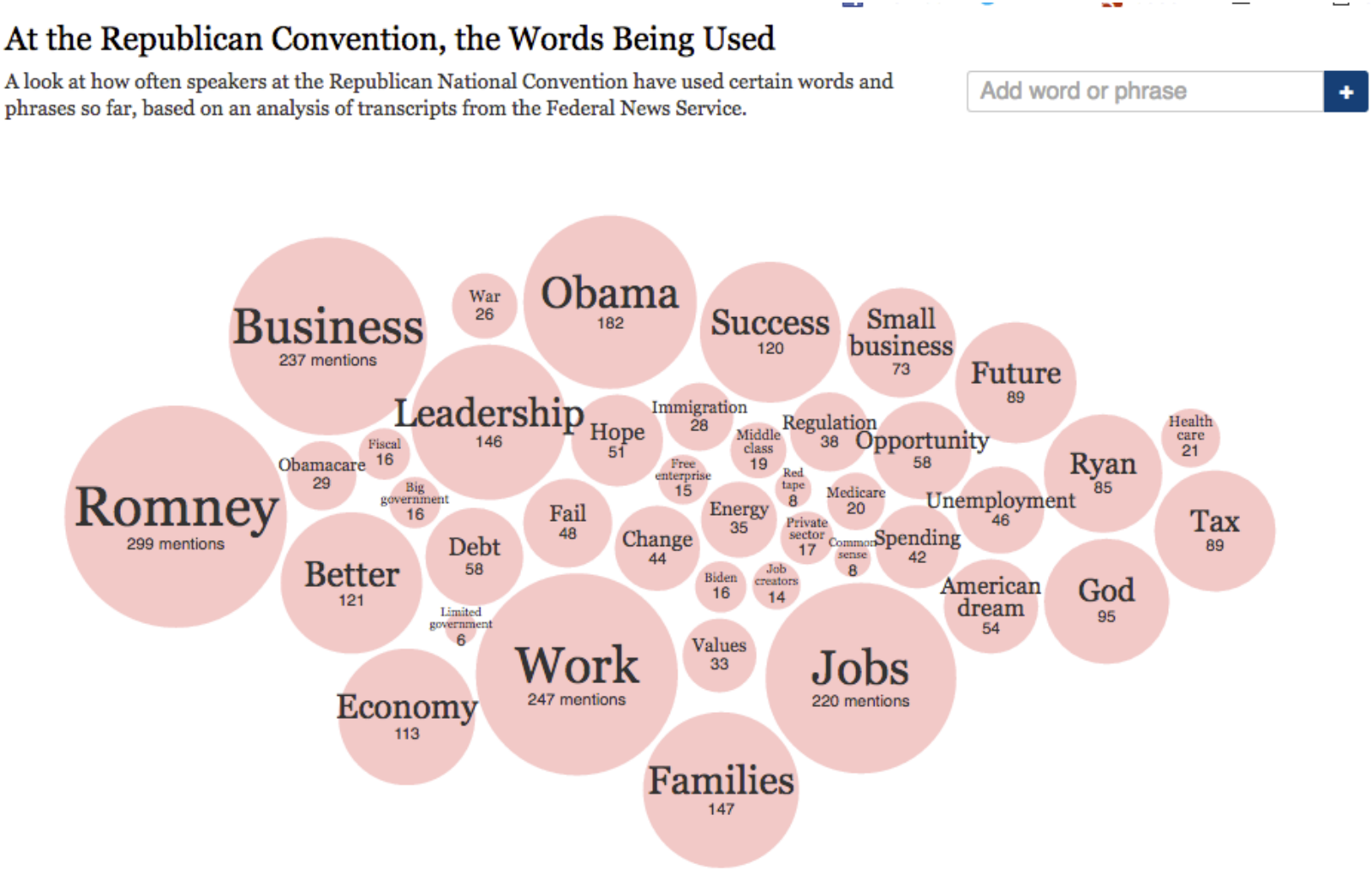
- Ignore ordering relationships within the text
- A document ≈ vector of term weights
  - Each dimension corresponds to a term (10,000+)
  - Each value represents the relevance
    - For example, simple term counts
- Aggregate into a document-term matrix
  - Document vector space model

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 Counts



http://www.nytimes.com/interactive/2012/08/28/us/politics/convention-word-counts.html

Word Counts

http://www.wordcount.org

Tag/Word Clouds

http://www.wordle.net

Wordle Tag Clouds

- Tightly packed words, sometimes vertical or diagonal
- Word size is linearly correlated with frequency (typically square root in cloud)
- Multiple color palettes
- User gets some control

Viegas, Wattenberg, &
Feinberg TVCG (InfoVis) '09

Tag Clouds: Pros and Cons

- Strengths
  - Can help with gisting and initial query formation.
- Weaknesses
  - Sub-optimal visual encoding (size vs. position)
  - Inaccurate size encoding (long words are bigger)
  - May not facilitate comparison (unstable layout)
  - Term frequency may not be meaningful
  - Does not show the structure of the text

Descriptive Words

- Given a text, what are the best descriptive words?

Keyword Weighting

- Term Frequency
  - tf_{td} = count(t) in d
  - Can take log frequency: $\log(1 + tf_{td})$
  - Can normalize to show proportion ($tf_{td} / \Sigma_t\ tf_{td}$)

- TF.IDF: Term Freq by Inverse Document Freq
  - $tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$
    - df_t = # docs containing t;
    - N = # of docs

An Example

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\text{tf}(\text{"this"}, d_1) = \frac{1}{5} = 0.2$$

$$\text{tf}(\text{"this"}, d_2) = \frac{1}{7} \approx 0.14$$

$$\text{idf}(\text{"this"}, D) = \log\left(\frac{2}{2}\right) = 0$$

$$\text{tfidf}(\text{"this"}, d_1) = 0.2 \times 0 = 0$$

$$\text{tfidf}(\text{"this"}, d_2) = 0.14 \times 0 = 0$$

$$\text{tf}(\text{"example"}, d_1) = \frac{0}{5} = 0$$

$$\text{tf}(\text{"example"}, d_2) = \frac{3}{7} \approx 0.429$$

$$\text{idf}(\text{"example"}, D) = \log\left(\frac{2}{1}\right) = 0.301$$

$$\text{tfidf}(\text{"example"}, d_1) = 0 \times 0.301 = 0$$

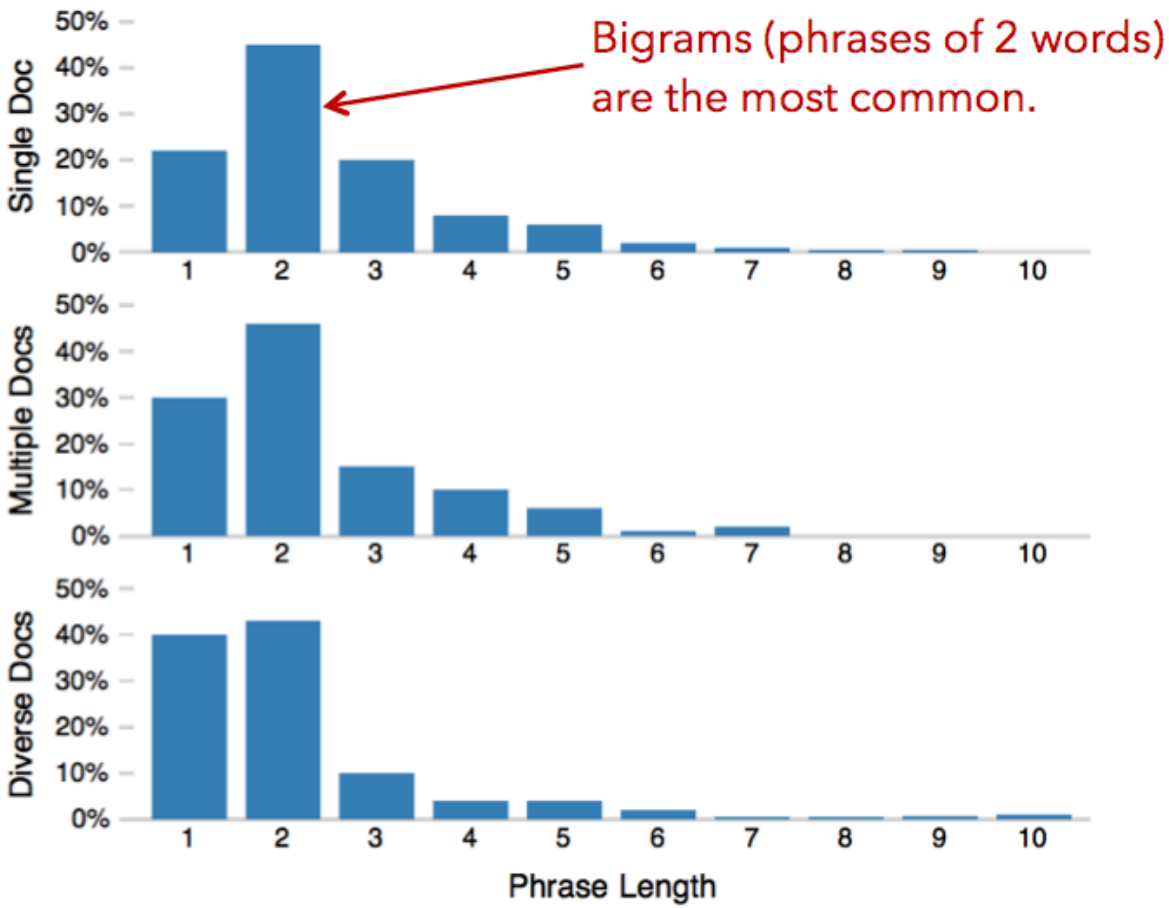
$$\text{tfidf}(\text{"example"}, d_2) = 0.429 \times 0.301 \approx 0.13$$

Limitations of Frequency Statistics

- Typically focus on unigrams (single terms)
- Often favors frequent (TF) or rare (IDF) terms
  - Not clear that these provide best description
- A "bag of words" ignores additional information
  - Grammar / part-of-speech
  - Position within document
  - Recognizable entities

How do people describe text?

- 69 subjects (graduate students) were asked to read and describe dissertation abstracts.



Chuang, Manning & Heer, 2012

Word vs. Phrases

A fighter jet rain check

Story and video by Chamila Jayaweera

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

CHAMILA JAYAWEERA/BOEING

The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out moreabout how the process works and watch the action unfold, click above to see the video story.

Word vs. Phrases

Word (e.g. TFIDF)

fighter
F/A
Hornet
Super
Boeing
-18
rain
St.
jet
Louis
15-inch-per-hour
douse
hangar
water-tight
Check
Baxter
sea-based
aircraft
Rich
seep
click
Navy
sure
Water
moisture
watch
enormous
stay

Key Phrase Extraction

Super Hornet
F/A -18
fighter jet
Boeing engineers
special process
rain check
electronics suite
Program experts
simulated rain
ultimate customers
enormous hangar
water-tight integrity
Rich Baxter
15-inch-per-hour rate
video story
aircraft
U.S. Navy fighter pilots
Super Hornet final assembly manager

Descriptive Phrases

- Understand the limitations of your language model.
  - Bag of words:
    - Easy to compute
    - Single words
    - Loss of word ordering
- Select appropriate model and visualization
  - Generate longer, more meaningful phrases
  - Adjective-noun word pairs for reviews
  - Show keyphrases within source text

(Optional Reading) Automatic Keyphrase Extraction: A Survey of the State of the Art: http://www.aclweb.org/anthology/P/P14/P14-1119.xhtml
http://bdewilde.github.io/blog/2014/09/23/intro-to-automatic-keyphrase-extraction/

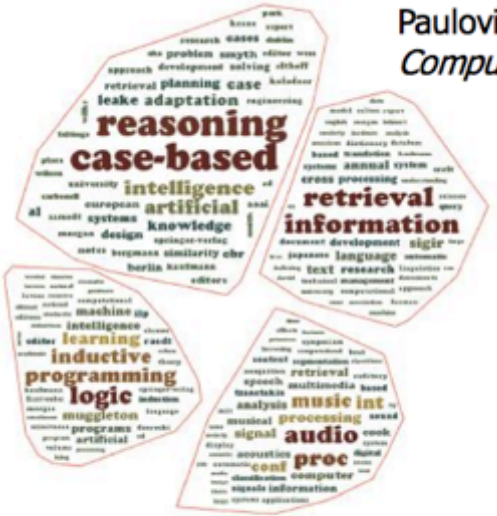
Context

Challenges

- High Dimensionality
  - Where possible use text to represent text...
  - ... which terms are the most descriptive?
- Context and 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.

Semantic/Context Word Clouds

Paulovich et al
Computer Graphics Forum '12

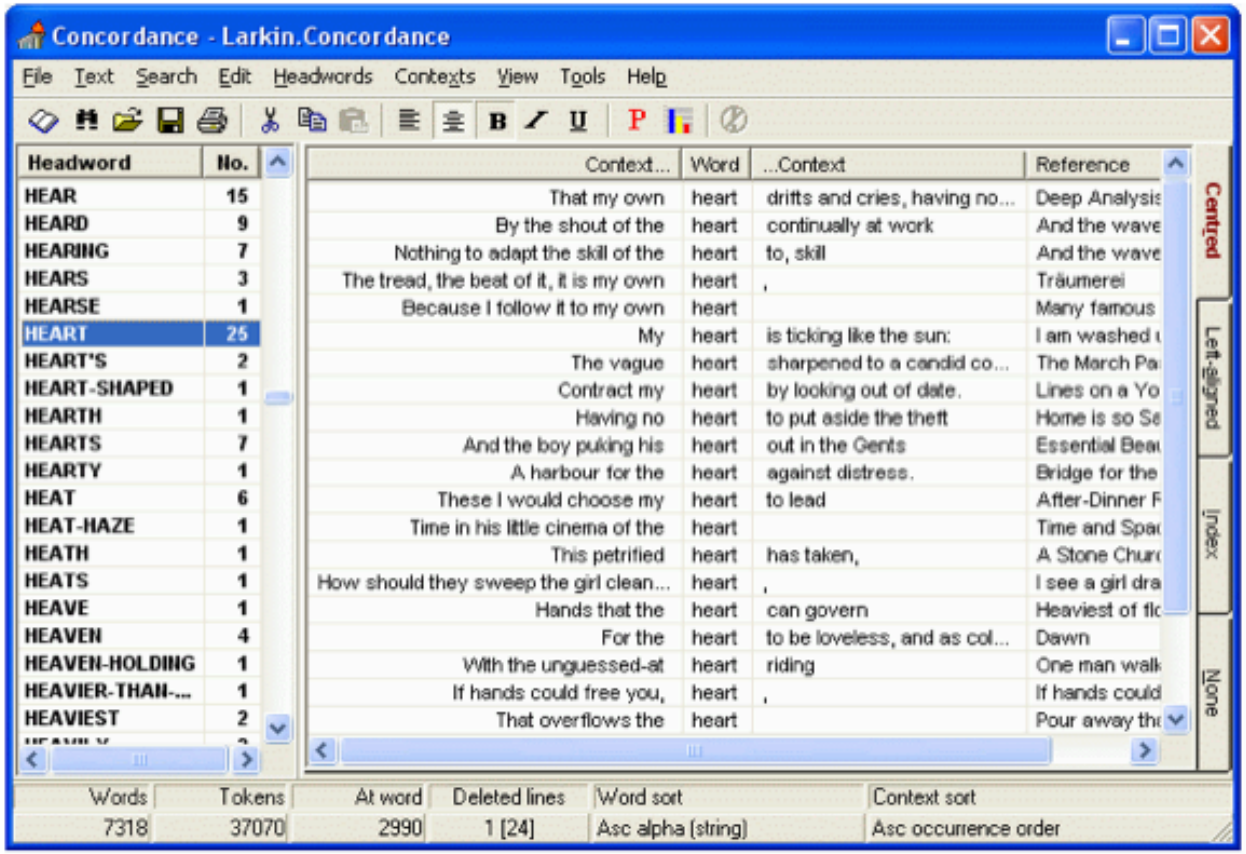


Wang et al
Graphics Interface '14

Wu et al
Computer Graphics Forum '11

Word / Phrase Context

- Concordance
  - What is the common local context of a term?



Word Tree

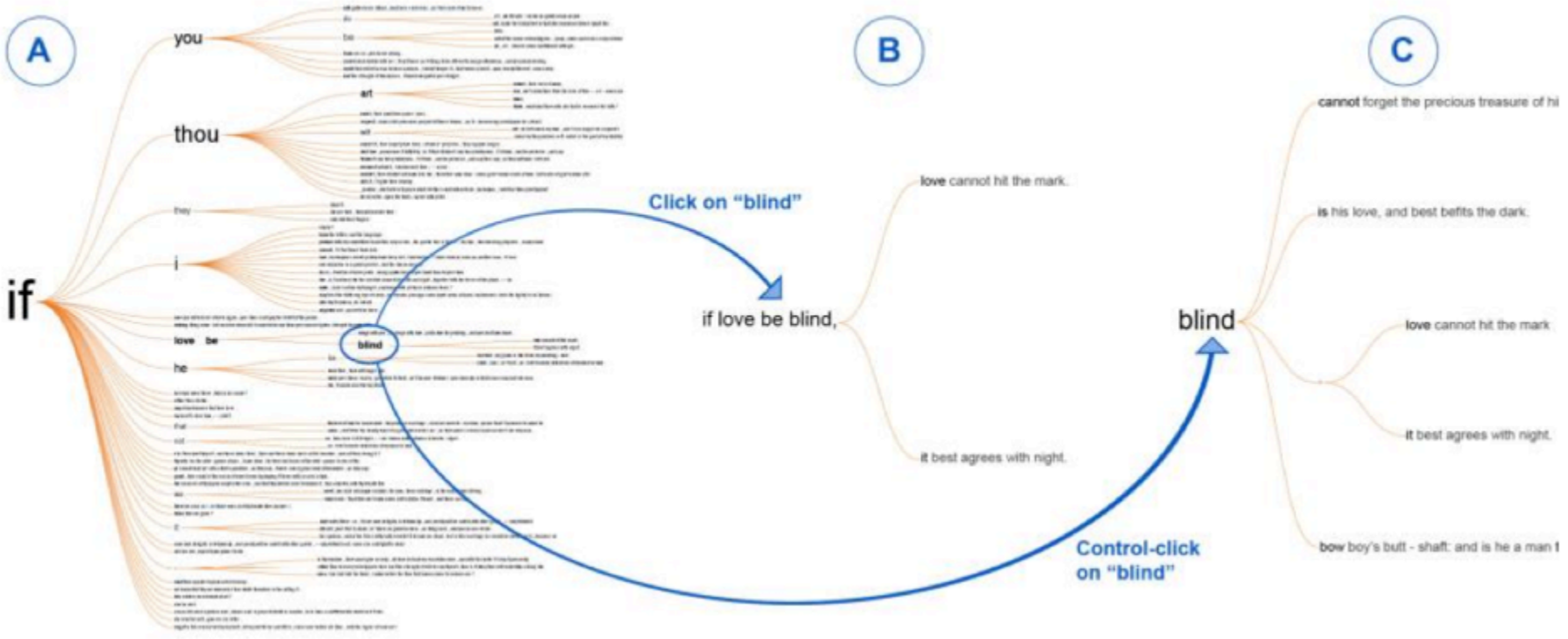

love the
lord
thy god
with all
thine heart , and with all thy soul ,
and with all thy might .
that thou mayest live .
thy heart , and with all thy soul , and with all thy
mind
strength , an
,
and
keep his charge , and his statutes , and his judgments , and his commandments , alway .
to walk ever in his ways ; then shalt thou add three cities more for thee , beside these three : 19 :
that thou mayest obey his voice , and that thou mayest cleave unto him : for he is thy life , and th
to walk in his ways , and to keep his commandments and his statutes and his judgments , that thou mayest live
your god
,
and to
serve him with all your heart and with all your soul , 11 : 14 that i will give you the rain of your lan
walk in all his ways , and to keep his commandments , and to cleave unto him , and to serve him
to walk in all his ways , and to cleave unto him ; 11 : 23 then will the lord drive out all these nations from
with all your heart and with all your soul .
,
all ye his saints : for the lord preserveth the faithful , and plentifully rewardeth the proud doer .
hate evil : he preserveth the souls of his saints ; he delivereth them out of the hand of the wicked .
because he hath heard my voice and my supplications .
name of the lord , to be his servants , every one that keepeth the sabbath from polluting it , and taketh hold of my covenant
good , and establish judgment in the gate : it may be that the lord god of hosts will be gracious unto the remnant of joseph
evil ; who pluck off their skin from off them , and their flesh from off their bones ; 3 : 3 who also eat the
truth and peace .
other ; or else he will hold to the one , and despise the other . ye cannot serve god and mammon .
6 : 25 therefore i say unto
16 : 14 and the pharisees
uppermost
rooms at feasts , and the chief seats in the synagogues , 23 : 7 and greetings in the markets , and to be called of
seats in the synagogues , and greetings in the markets .
father
; and as the father gave me commandment , even so i do .
hath bestowed upon us , that we should be called the sons of god : therefore the world knoweth us not , because it knew him
brotherhood .
world , the love of the father is not in him .
brethren .
children of god , when we love god , and keep his commandments .

Word Tree

- Shows context of a word or words
  - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus
- Ordered alphabetically, by frequency, or by first appearance

(Optional) Wattenberg & Viégas TVCG (InfoVis) '08

Word Tree Interaction

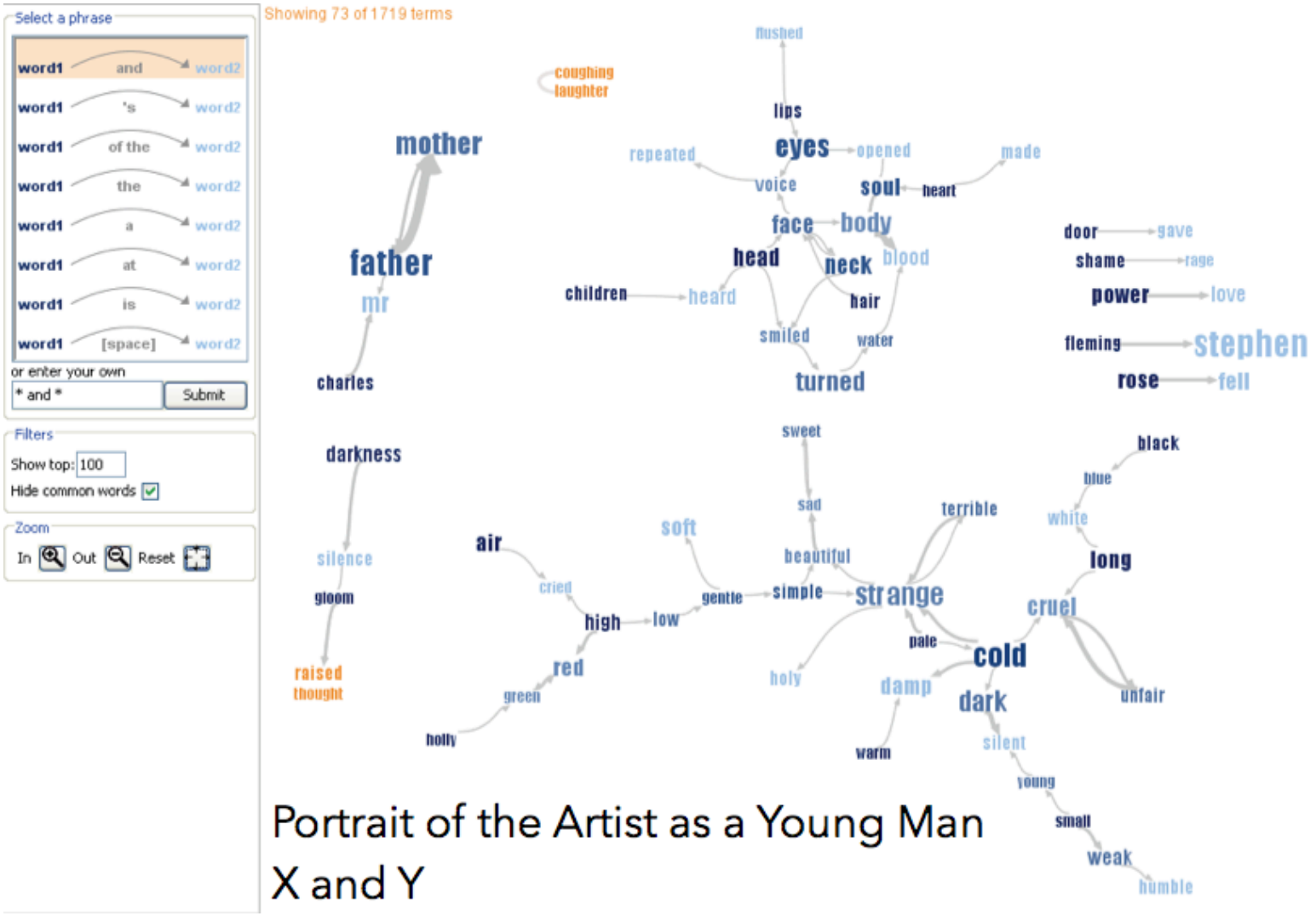


Phrase Nets

- Concordances show local, repeated structure, but what about other types of patterns?
  - Lexical: <A> at <B>
  - Syntactic: <Noun> <Verb> <Object>
- Look for specific linking patterns in the text:
  - ‘A and B’, ‘A at B’, ‘A of B’, etc
  - Could be output of regexp or parser.
- Visualize patterns in a node-link view
  - Occurrences -> Node size
  - Pattern position -> Edge direction

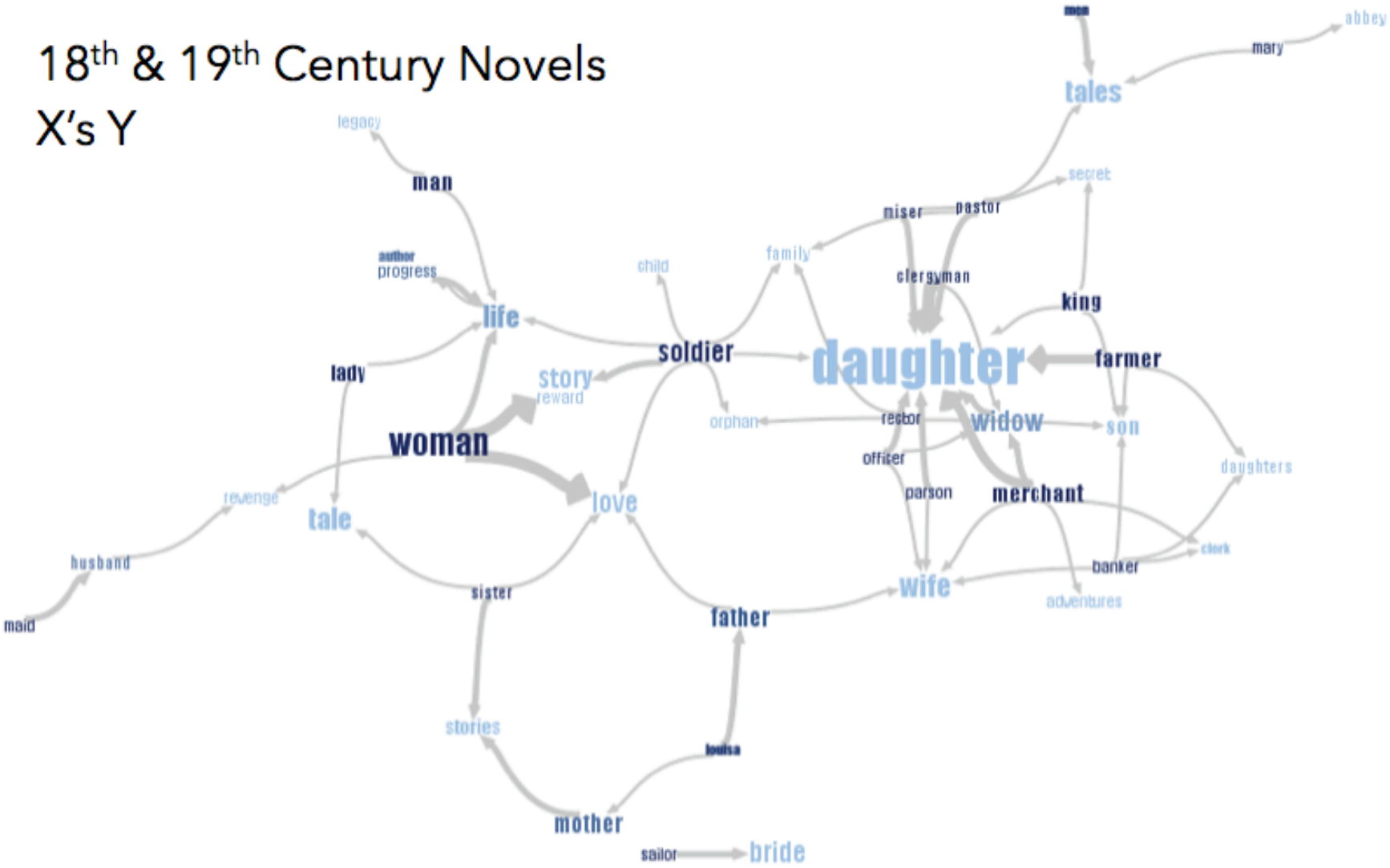
van Ham et al TVCG (InfoVis) ‘09

Phrase Nets



Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Phrase Nets



Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

SentenTree

- Elements of word clouds and word trees
  - Highlight keywords using size
  - Show sentence fragments
  - Provide a summary of the dataset
  - Enable drill-down into details

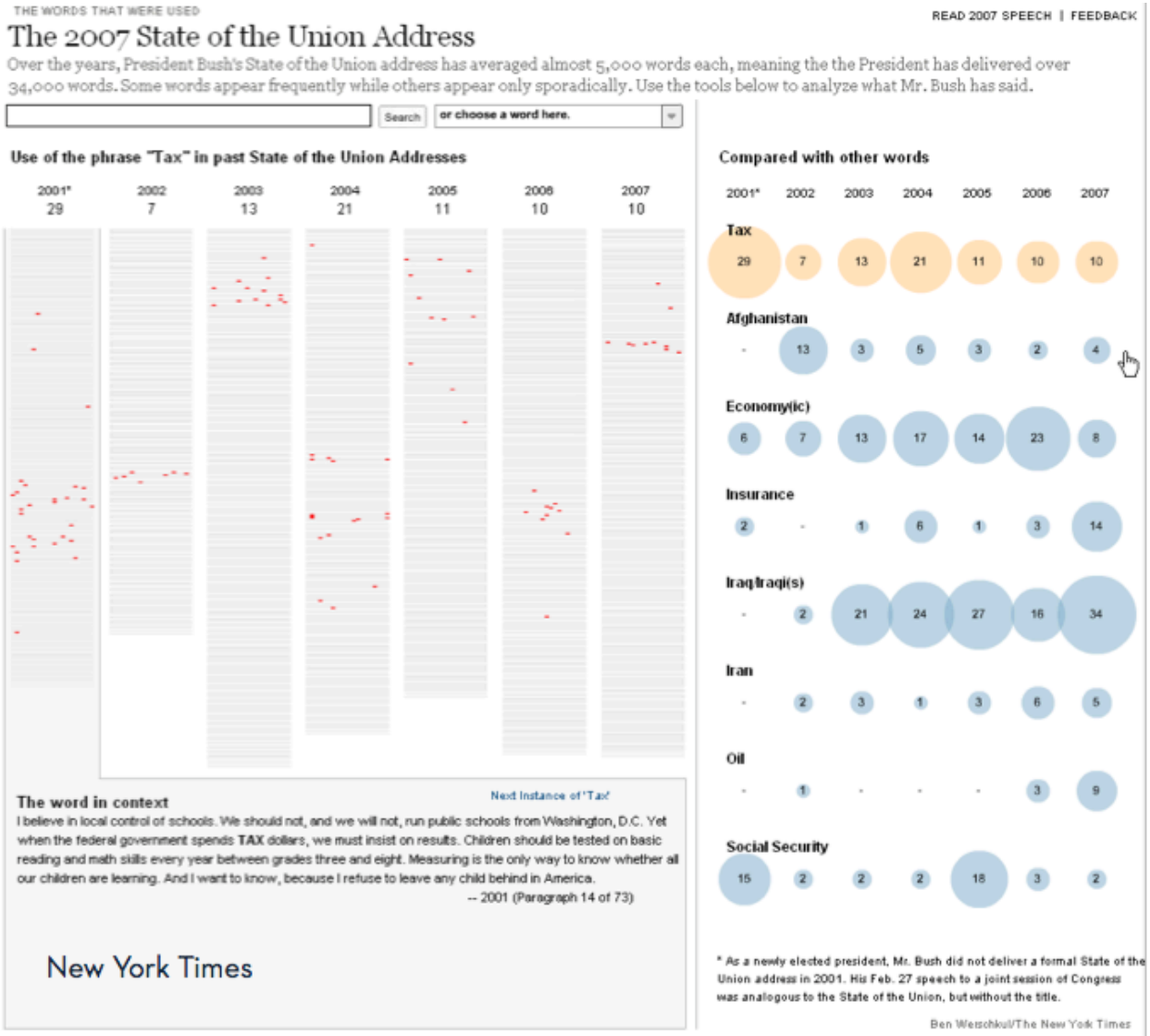
Hu, et al. TVCG '17 (InfoVis '16)

Summary of 189,450 tweets (108,702 unique) posted in a 15 minute time window around the first goal of the opening game of the 2014 Soccer World Cup

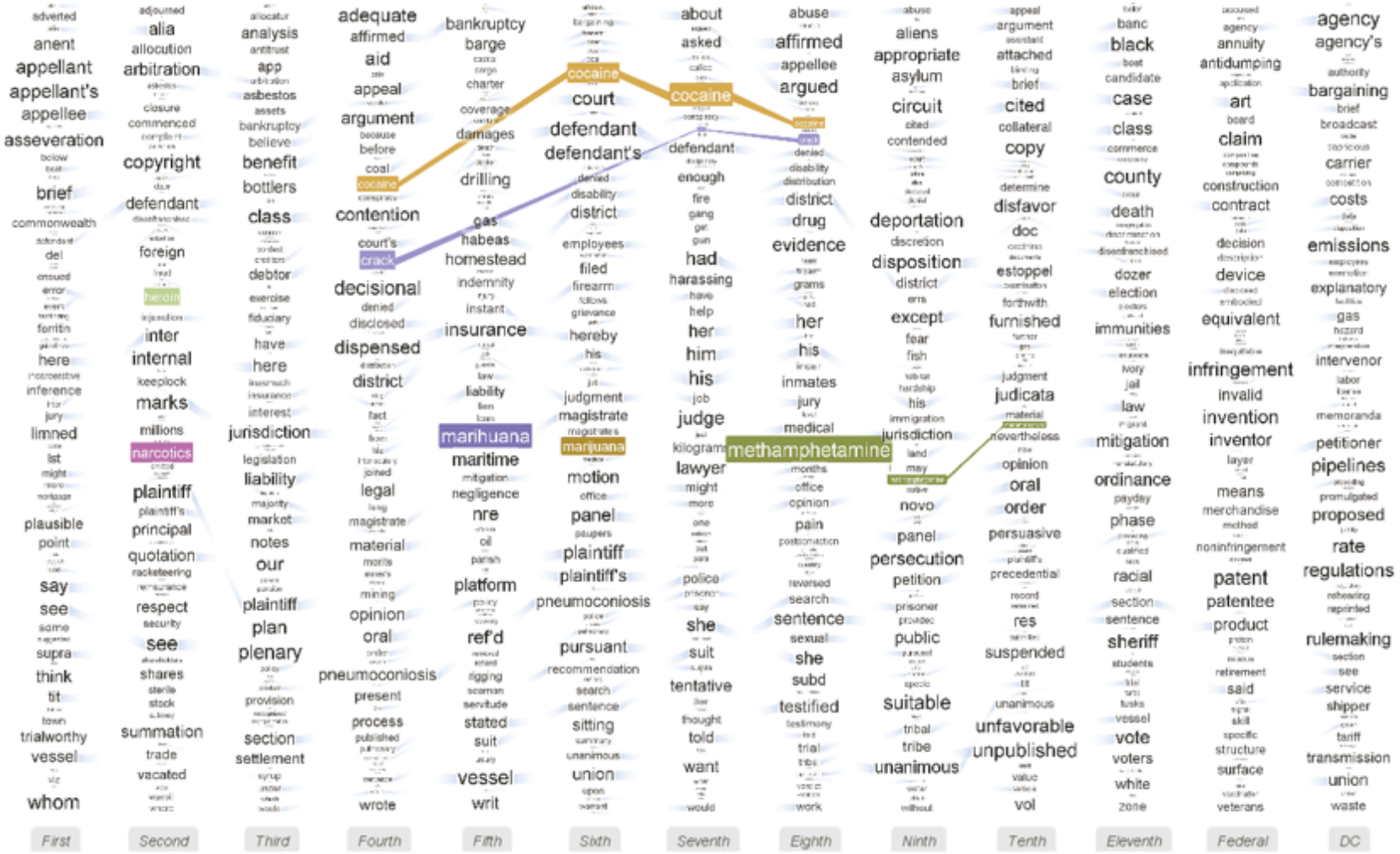
Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Relationship to Other Texts

Compare to Others



Parallel Tag Clouds



Collins et al VAST '09

Video: https://www.youtube.com/watch?v=rL3Ga6xBgLw

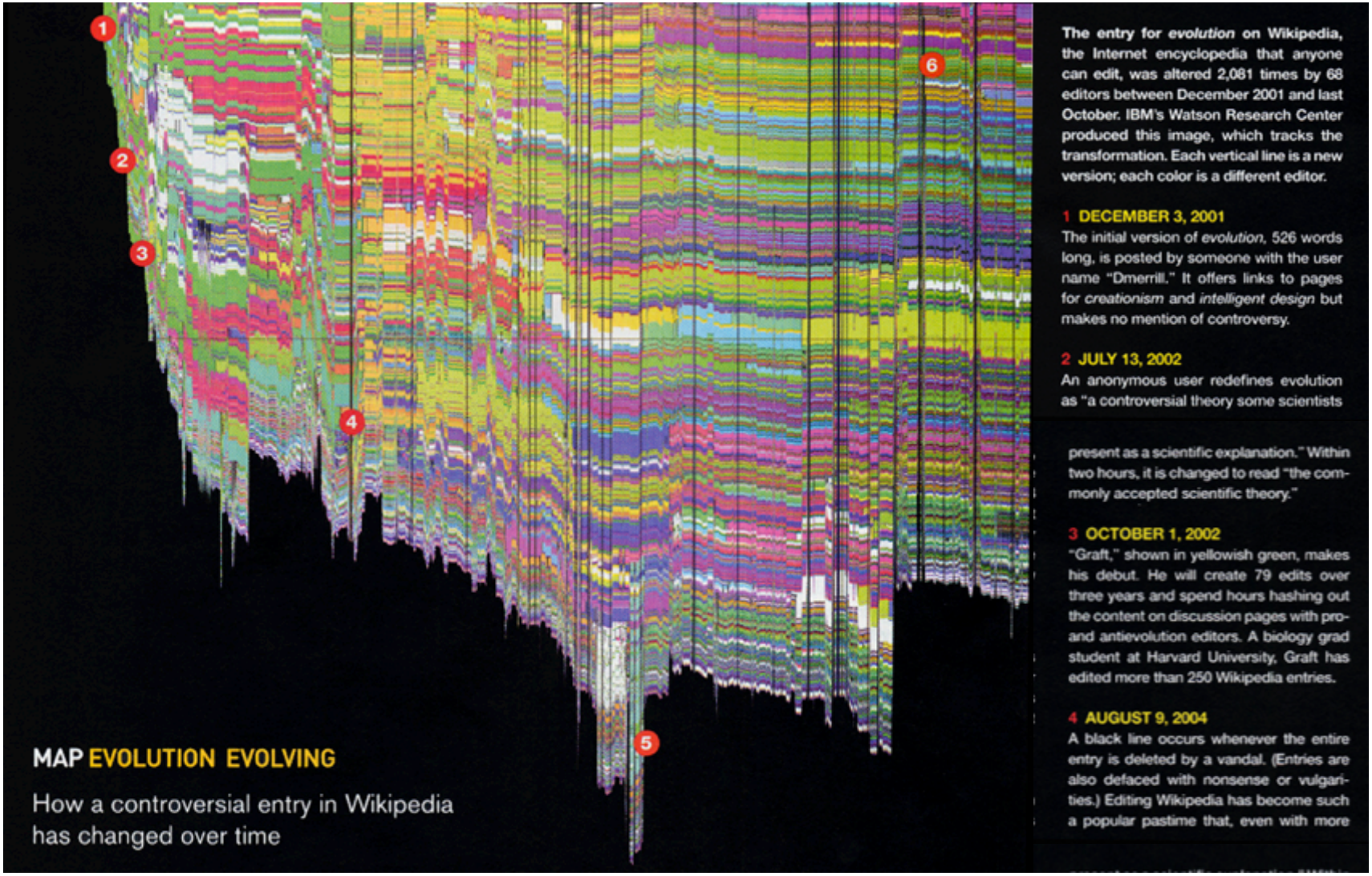
Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Animated Traces: evolving documents

https://fathom.info/traces/

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Wikipedia Edit Evolution

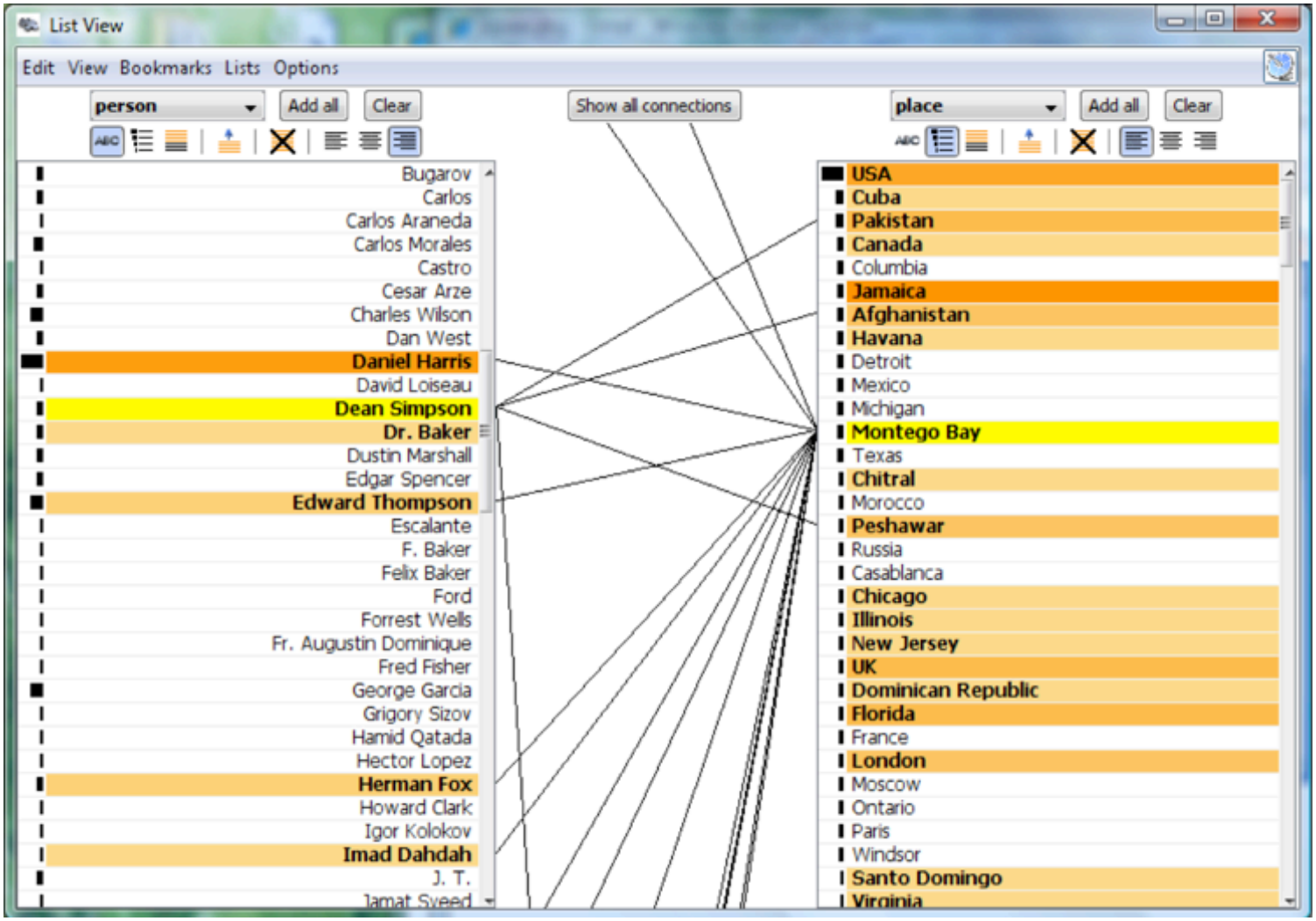


*Produced by IBM Watson Research Center

Visualizing Document Collections
Named Entity Recognition

- Label named entities in text:
  - John Smith -> PERSON
  - Soviet Union -> COUNTRY
  - 353 Serra St -> ADDRESS
  - (555) 721-4312 -> PHONE NUMBER
- Entity relations: how do the entities relate?
- Simple approach: do they co-occur in a small window of text?

Entity Linkage

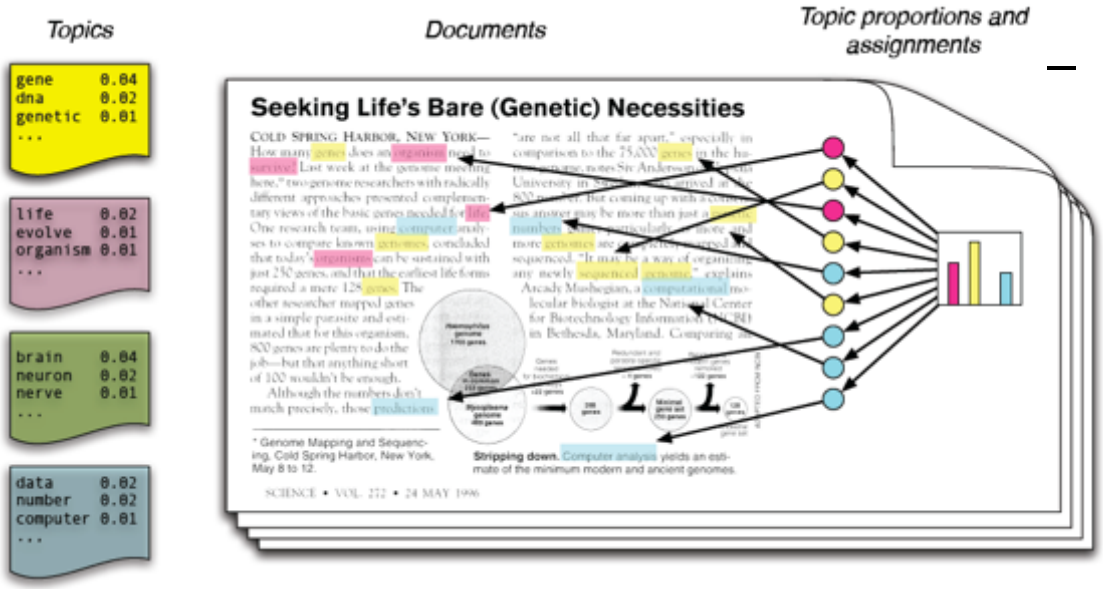


Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

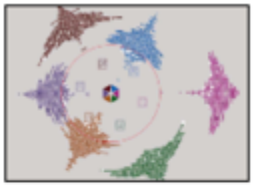
Similarity & Clustering

- Compute vector distance among docs
  - For TF.IDF, typically cosine distance Similarity measure can be used to cluster

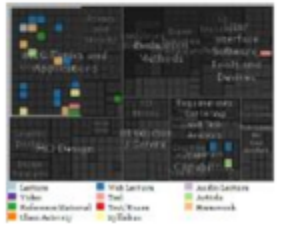
- Topic modeling
  - Assume documents are a mixture of topics
  - Topics are (roughly) a set of co-occurring terms
  - Latent Semantic Analysis (LSA): reduce term matrix
  - Latent Dirichlet Allocation (LDA): statistical model



Visualization for Information Retrieval



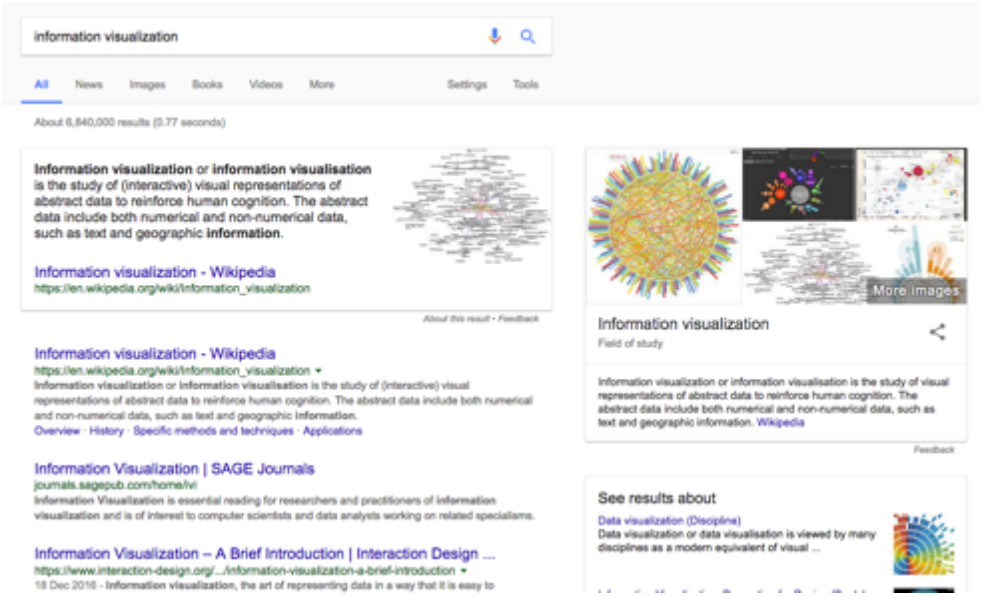
Visualization for IR
Helping search



Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

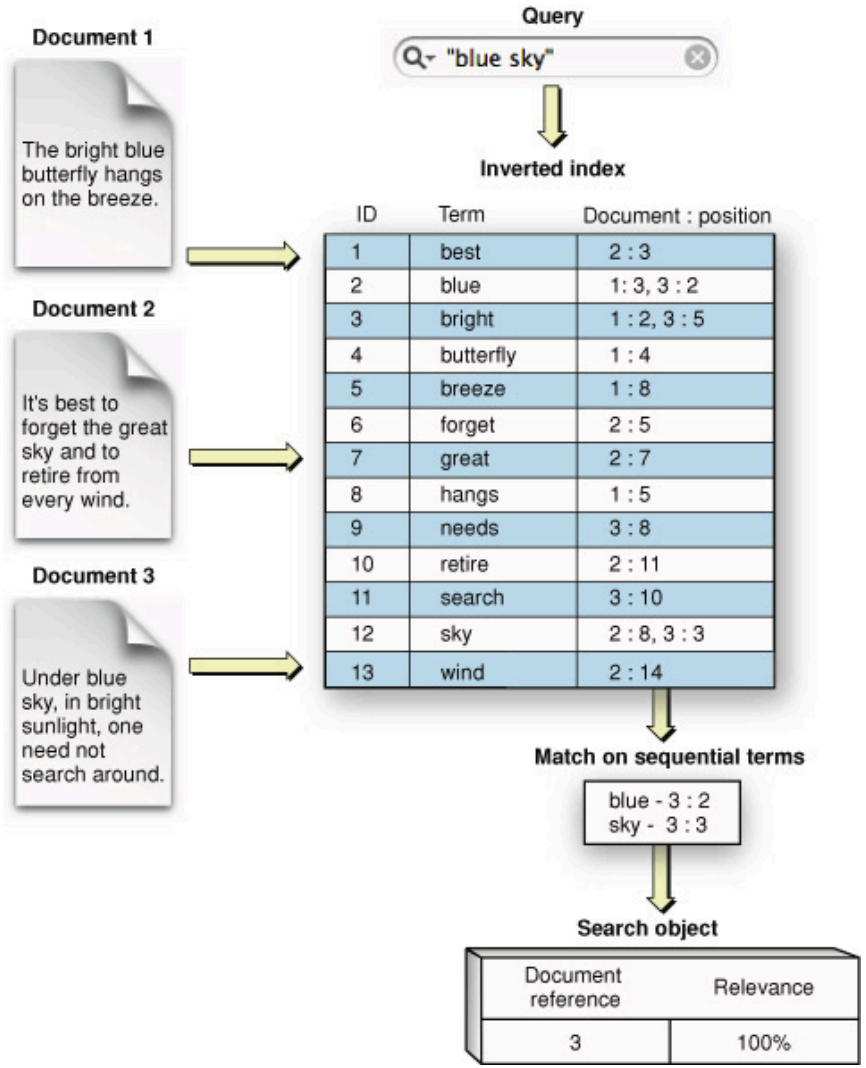
Information Retrieval (IR)

- Search for documents
- Visualization to contextualize query-doc matching results
- Can Information visualization help IR?
  - A (large) set of documents

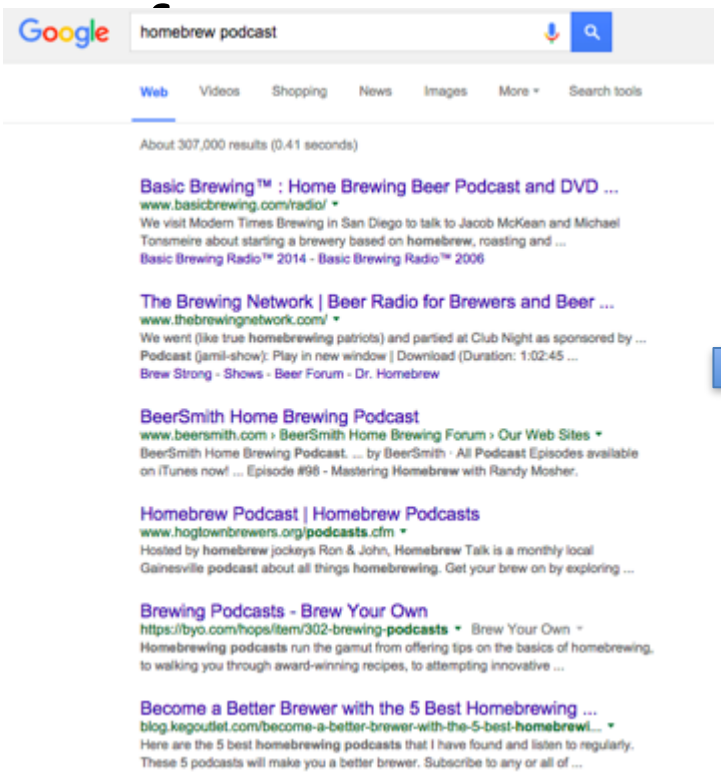


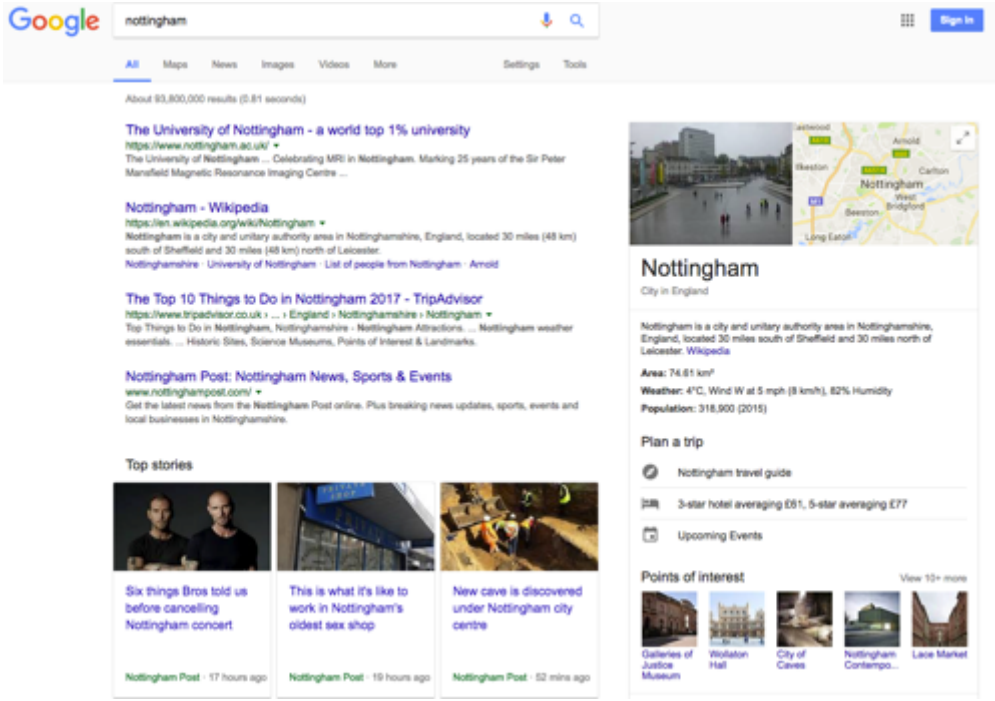
Paper Handout: http://searchuserinterfaces.com/book/, Chapter 10

Information Retrieval



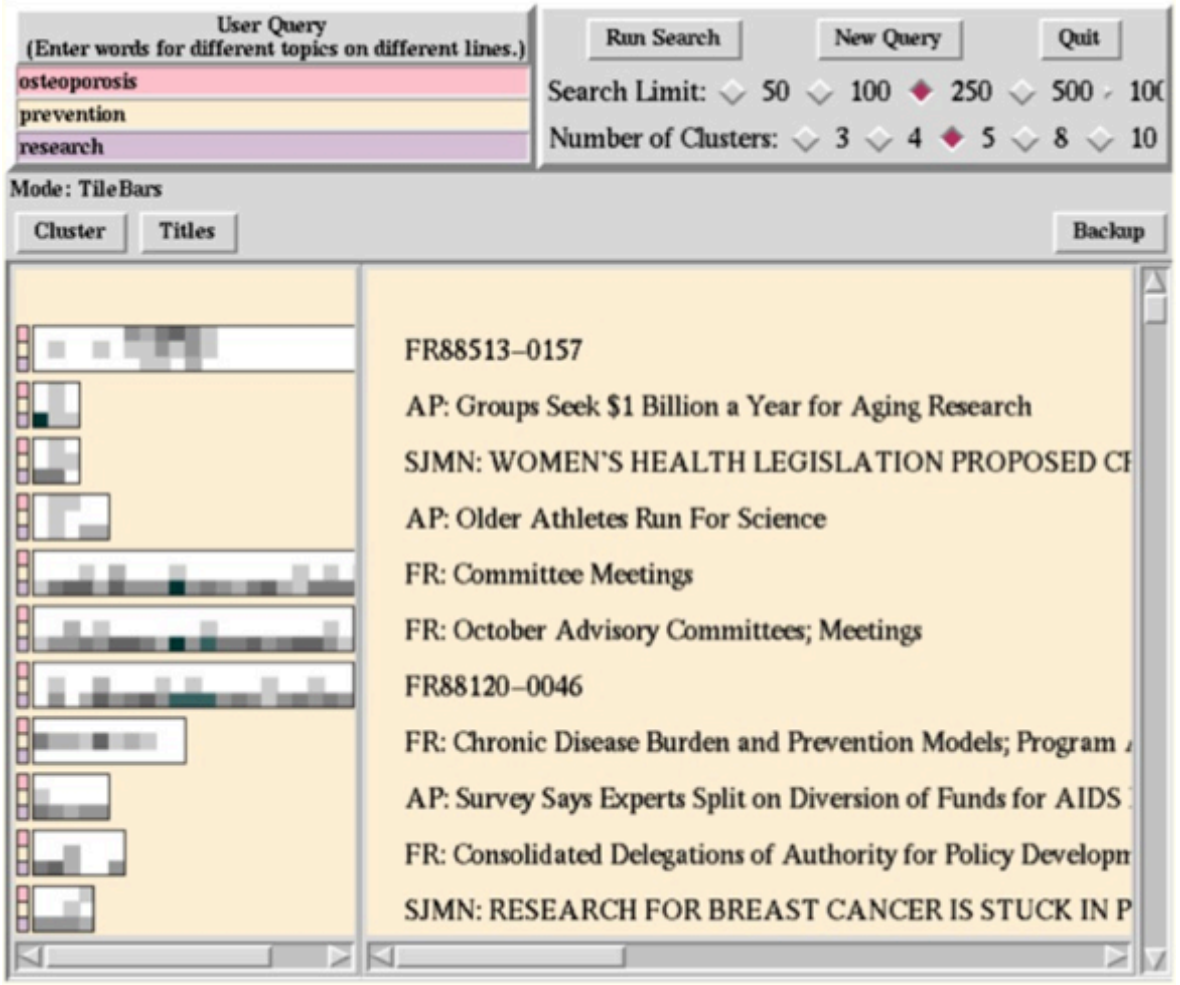
Ten Blue Links to More Complex Data and Visualization





Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

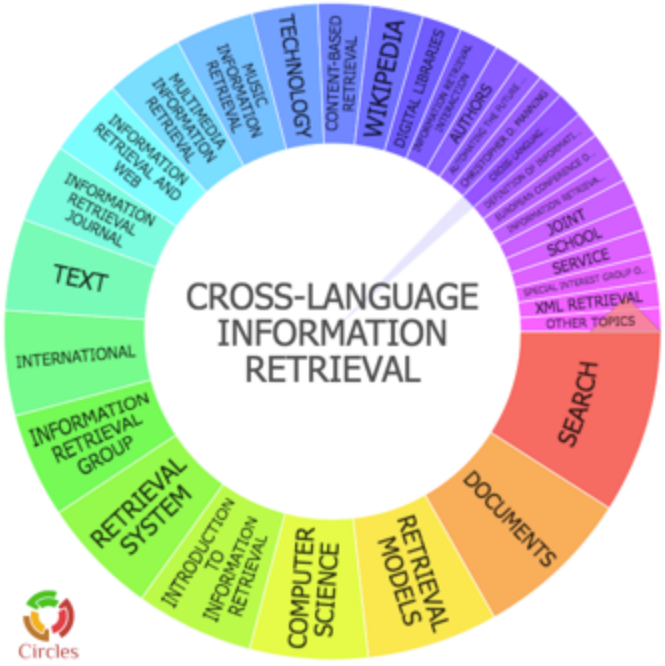
TileBars

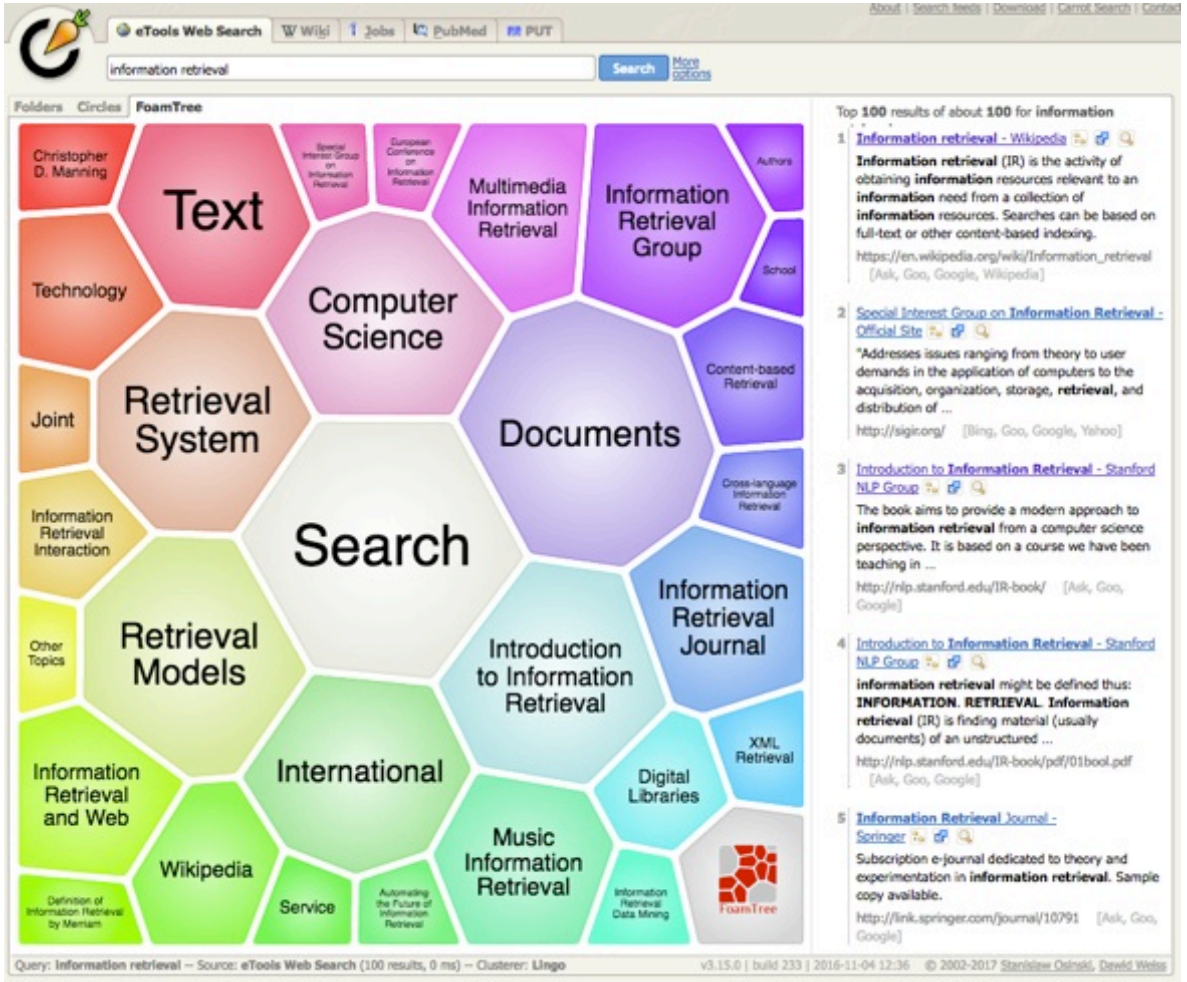


Hearst, 1995

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Cluster Based Search Visualization

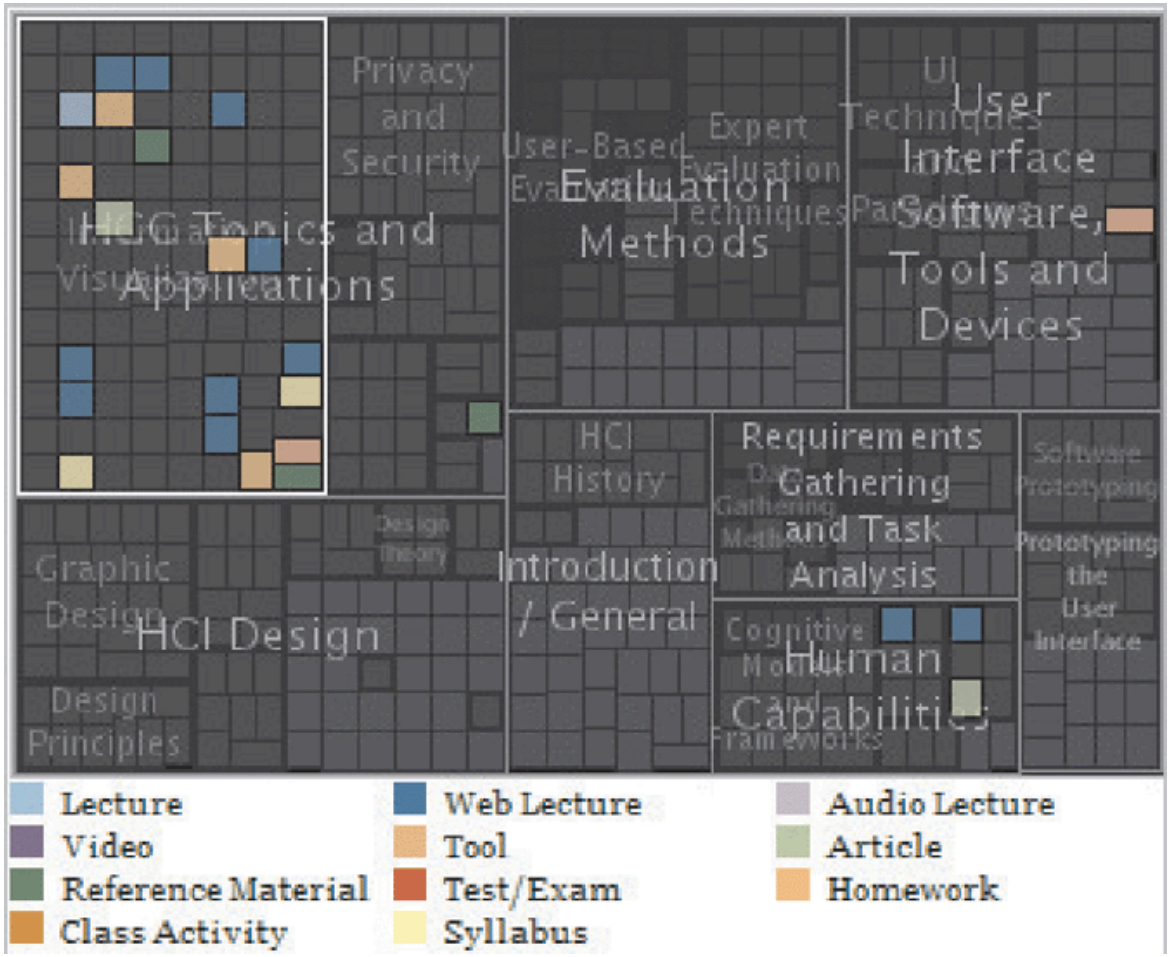




http://search.carrot2.org/

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

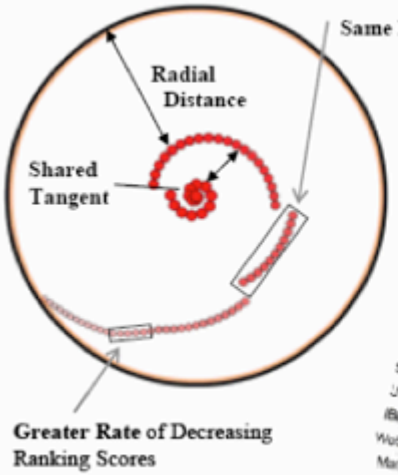
Result Maps



Treemap-style visualization for showing query results in a digital library

Clarkson, Desai & Foley InfoVis'09

RankSpiral



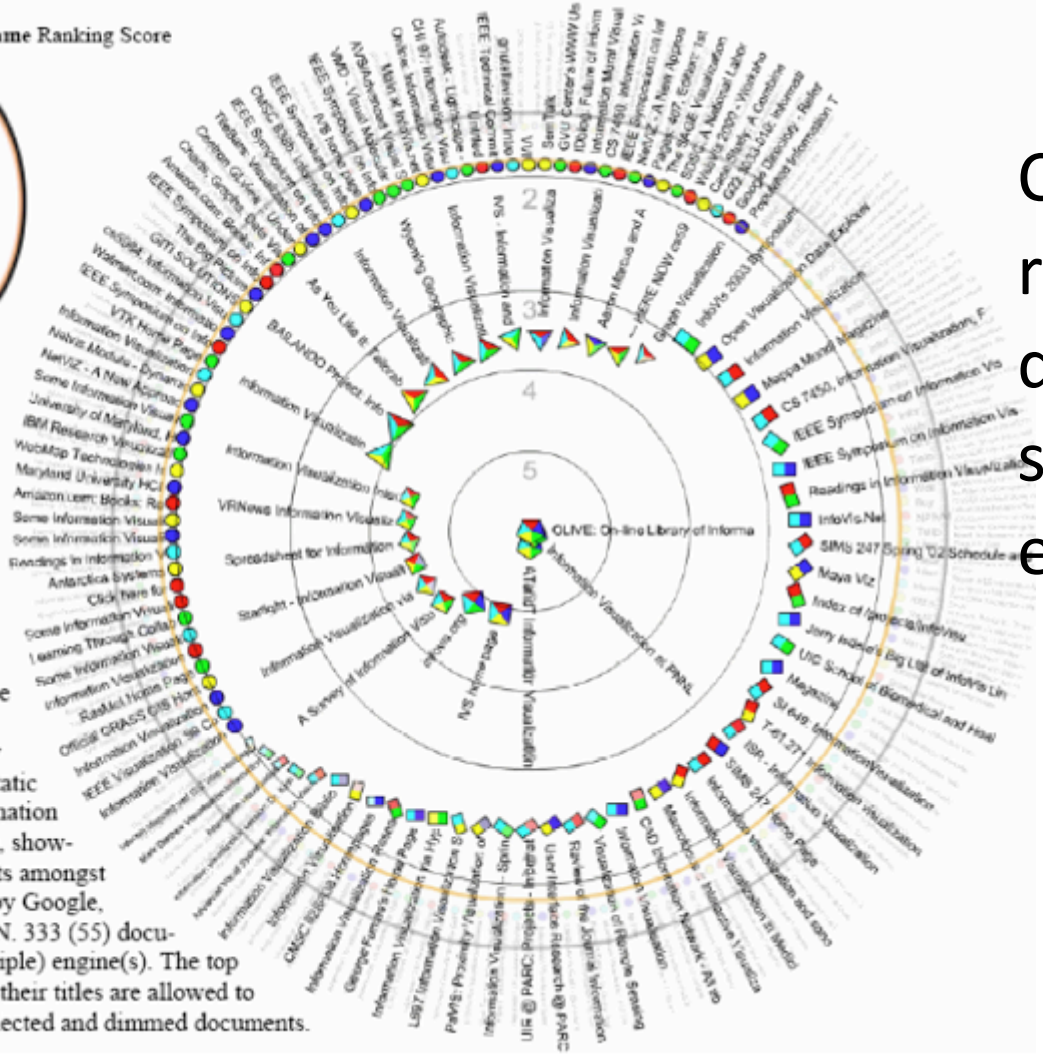


Figure 1: (Top) *RankSpiral* places consecutive document icons next to each other so that they do not overlap. Total ranking score of documents increases to-ward the center. Radial distance between documents that have the same angle can be used to display title fragments. (Right) shows a static RankSpiral that maximizes information density and minimizes occlusions, showing here the 388 unique documents amongst the top 100 documents retrieved by Google, Teoma, AltaVista, Lycis and MSN. 333 (55) documents were found by single (multiple) engine(s). The top 100+ documents are selected and their titles are allowed to extend across the remaining unselected and dimmed documents.

Color represents different search engines

Spoerri InfoVis ’04 poster

Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Summary

- 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
  - Understand abstraction of your language models.
  - Match analysis task with appropriate tools and models.
  - From bag-of-words to vector space embeddings

Next Lecture

- Topic:
  - Visualizing Time Series, Trees and Graphs