Text and Document Visualization 1

CS 4460 – Intro. to Information Visualization
October 28, 2014
John Stasko

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?

Tasks/Goals

- What kinds of analysis questions might a person ask about text & documents?

Example Tasks & Goals

- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- Which documents have an angry tone?
- How are certain words or themes distributed through a document?
- Identify "hidden" messages or stories in this document collection.
- How does one set of documents differ from another set?
- Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
- Understand the history of changes in a document.
- Find connections between documents.

Related Topic - IR

- Information Retrieval
  - Active search process that brings back particular/specific items (will discuss that some today, but not always focus)
  - I think InfoVis and HCI can help some…
- InfoVis, conversely, seems to be most useful when
  - Perhaps not sure precisely what you're looking for
  - More of a browsing task than a search one

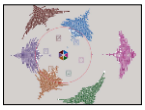
Related Topic - Sensemaking

- Sensemaking
  - Gaining a better understanding of the facts at hand in order to take some next steps
  - (Better definitions in VA lecture)
- InfoVis can help make a large document collection more understandable more rapidly

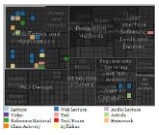
Challenge

- Text is nominal data
  - Does not seem to map to geometric/graphical presentation as easily as ordinal and quantitative data
- The "Raw data --> Data Table" mapping now becomes more important

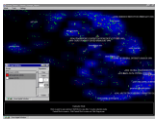
This Week's Agenda



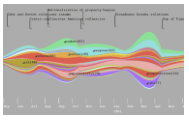
Visualization for IR
Helping search



Visualizing text
Showing words, phrases, and sentences



Visualizing document sets
Words, entities & sentences
Analysis metrics
Concepts & themes



Information Retrieval

- Can InfoVis help IR?
- Assume there is some active search or query
  - Show results visually
  - Show how query terms relate to results
  - …

Generalize More

- How about the "holy grail" of a visual search engine?
  - Hot idea for a while
- My personal view: It's a mistake in the general case. Text is just better for this.

Search Visualization

`http://www.kartoo.com`

Defunct

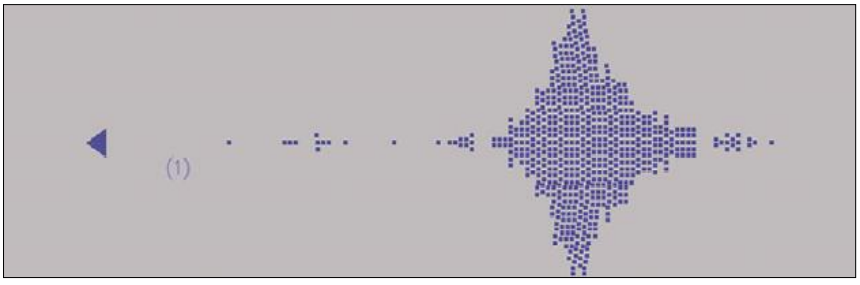
Sparkler

- Abstract result documents more
- Show "distance" from query in order to give user better feel for quality of match(es)
- Also shows documents in responses to multiple queries

Havre et al
InfoVis '01

Visualizing One Query

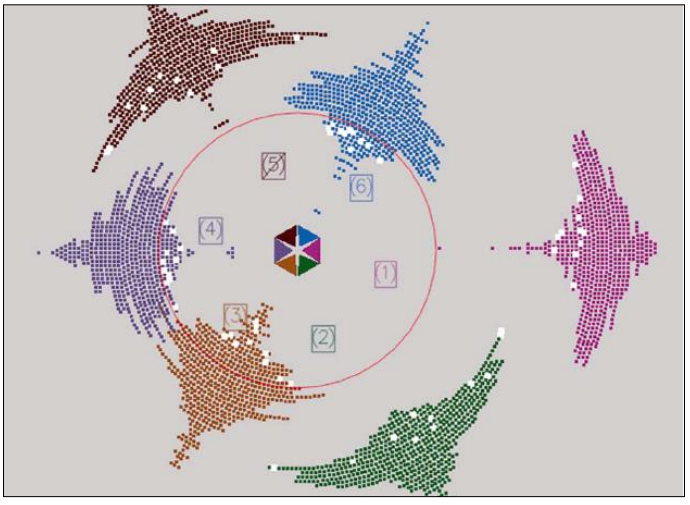
- Triangle – query
- Square – document
- Distance between query and documents represents their relevance



Visualizing Multiple Queries

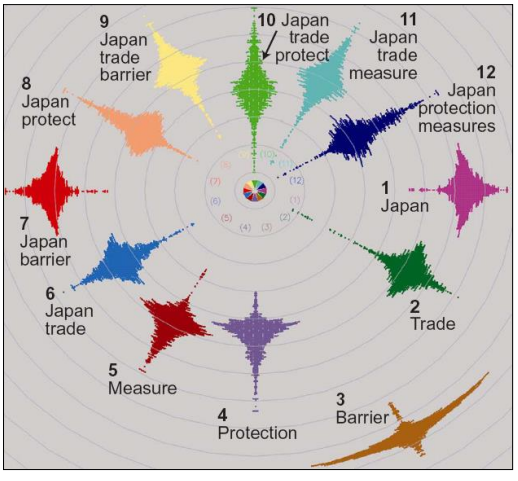
Six queries here

Bullseye allows viewer to select quality results

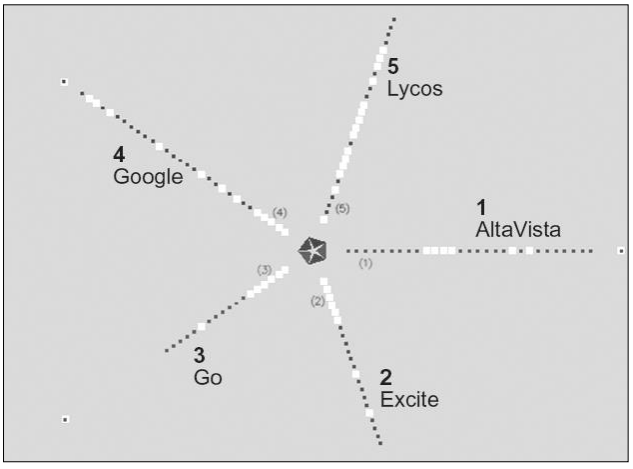


Test Example

- Text Retrieval Conference (TREC-3) test document collection
- AP news stories from June 24–30, 1990
- TREC topic: Japan Protectionist Measures
- Sparkler found 16 of 17 relevant documents



Another Idea



Use it to compare search results from different search engines

RankSpiral

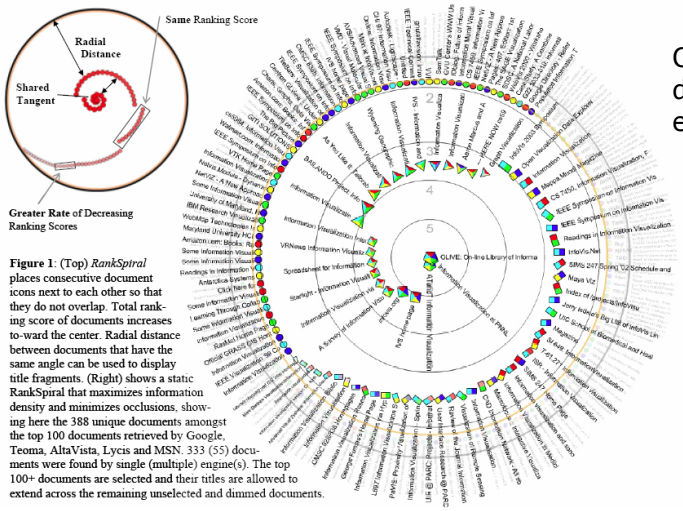
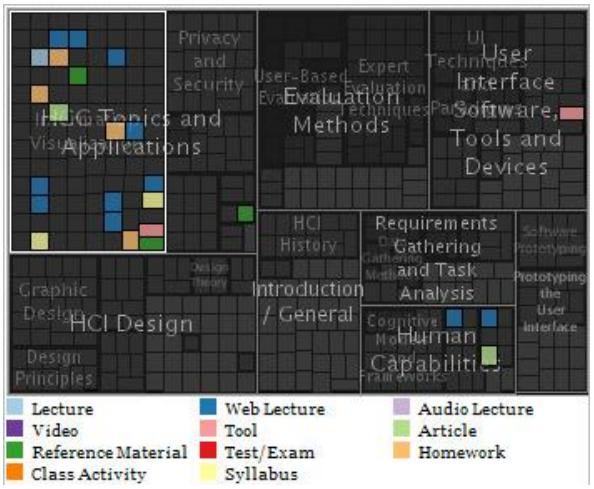


Figure 1: (Top) *RankSpiral* places consecutive document icons next to each other so that they do not overlap. Total ranking score of documents increases to-ward the center. Radial distance between documents that have the same angle can be used to display title fragments. (Right) shows a static RankSpiral 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

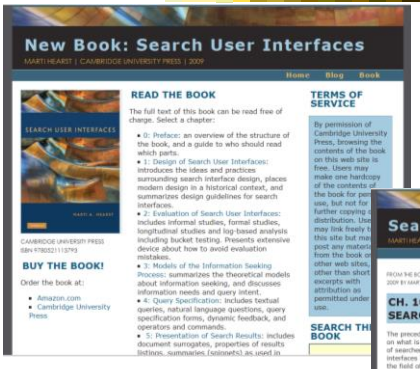
ResultMaps



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

Clarkson, Desai & Foley *TVCG* (InfoVis) '09

To Learn More



Marti Hearst's Book

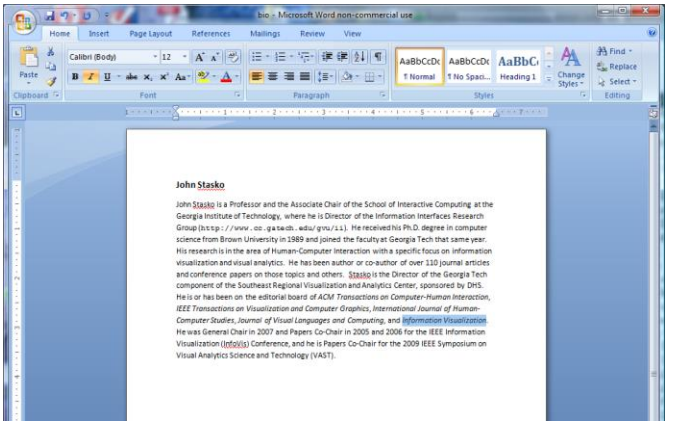
Chapter 10

http://searchuserinterfaces.com/book/

Transition 1

- OK, let's move up beyond just search/IR
- How do we represent the words, phrases, and sentences in a document or set of documents?
  - Main goal of *understanding* versus search

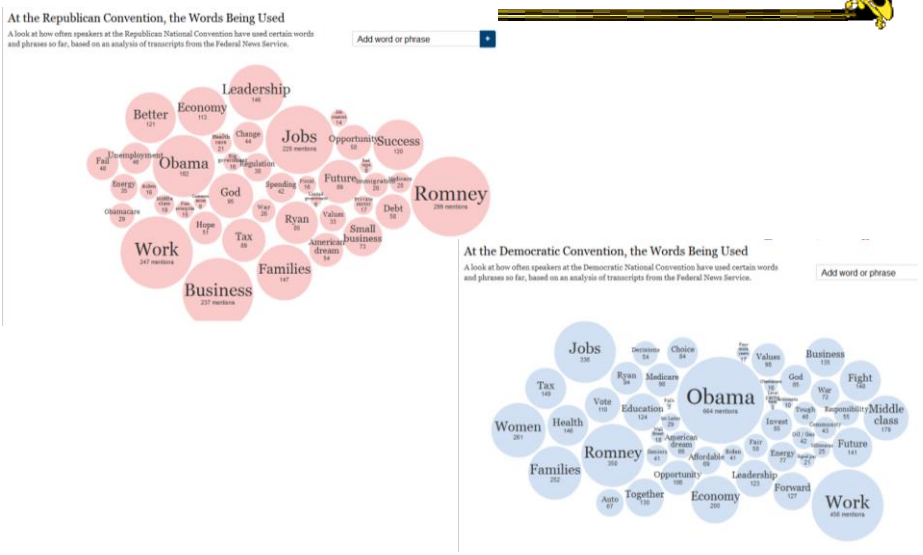
One Text Visualization



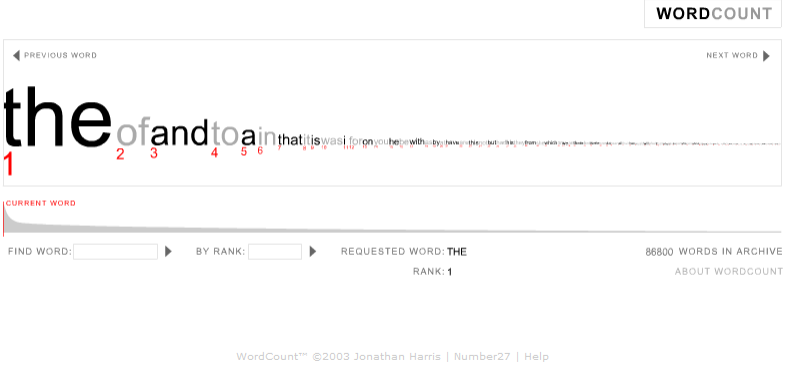
Uses:
Layout
Font
Style
Color
…

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

Word Counts



More Word Counting



http://www.wordcount.org

Tag/Word Clouds

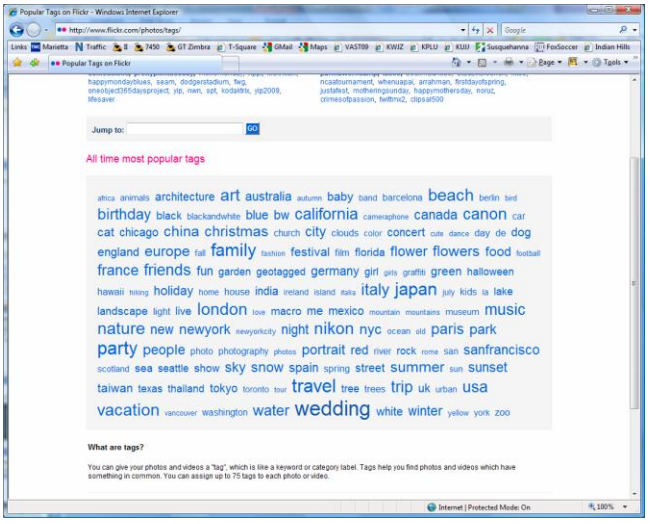
- Currently very "hot" in research community
- Have proven to be very popular on web
- Idea is to show word/concept importance through visual means
  - Tags: User-specified metadata (descriptors) about something
  - Sometimes generalized to just reflect word frequencies

History

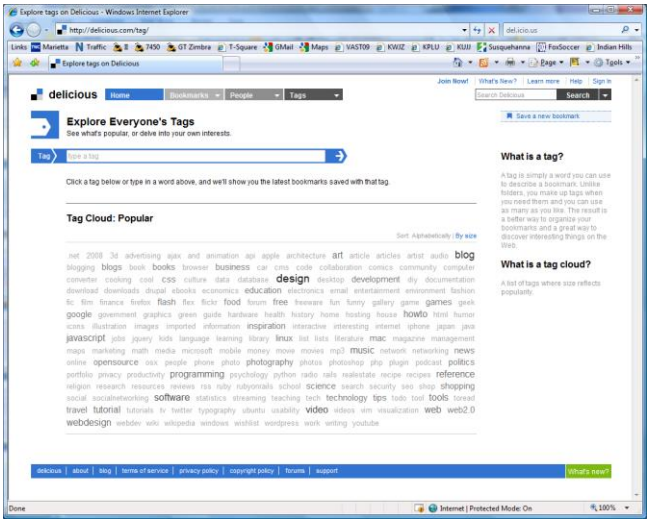
- 90-year old Soviet Constructivism
- Milgram's '76 experiment to have people label landmarks in Paris
- Flanagan's '97 "Search referral Zeitgeist"
- Fortune's '01 Money Makes the World Go Round

Viégas & Wattenberg
interactions '08

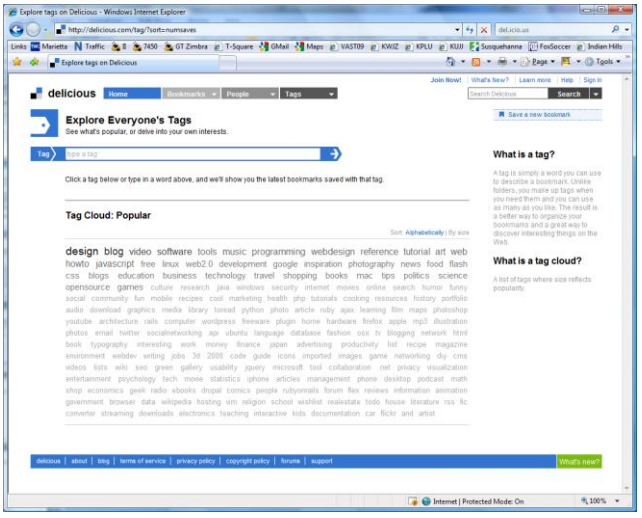
Flickr Tag Cloud



delicious Tag Cloud

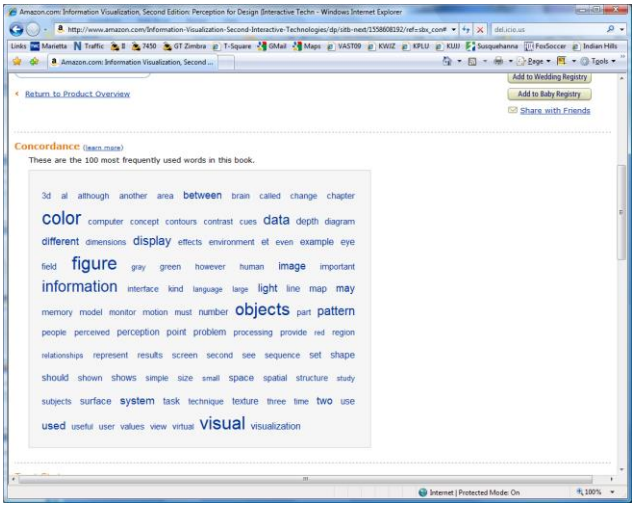


Alternate Order



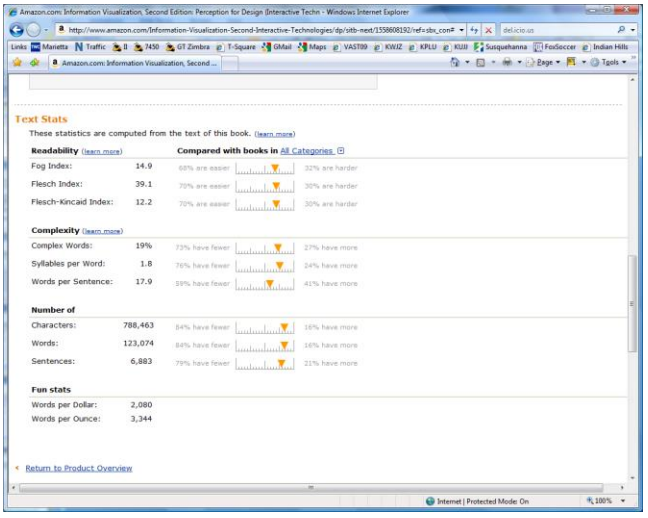
Amazon's Product Concordance

Maybe now a "word cloud"



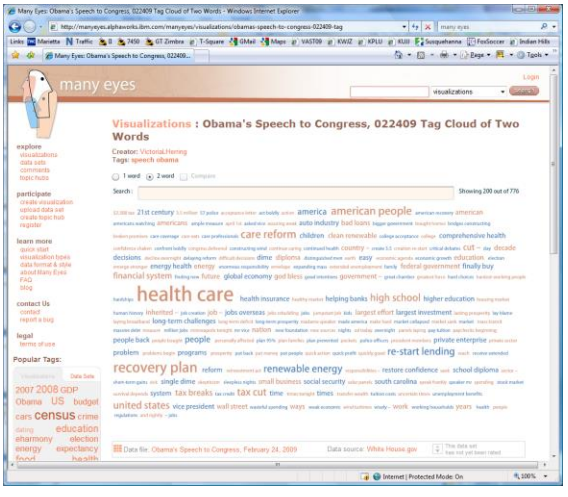
Sidenote

There are other types of info about a document on Amazon



Many Eyes Tag Cloud

Here, pairs of words are shown



Problems

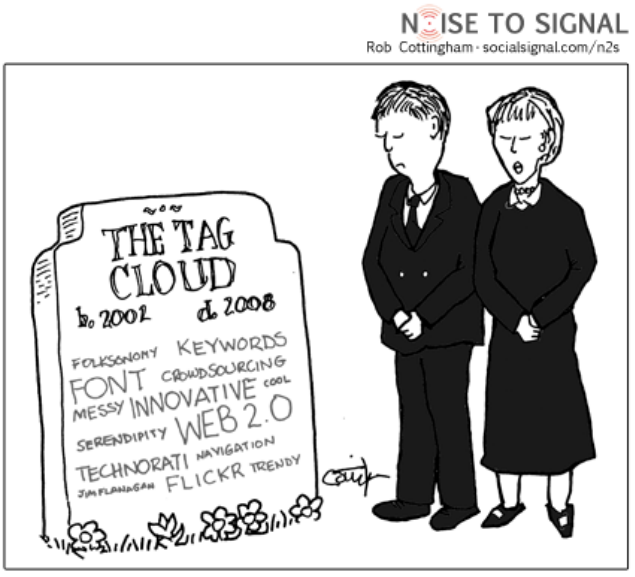
- Actually not a great visualization. Why?
  - Hard to find a particular word
  - Long words get increased visual emphasis
  - Font sizes are hard to compare
  - Alphabetical ordering not ideal for many tasks
- Studies have even shown they underperform

Gruen et al
CHI '06

Why So Popular?

- Serve as social signifiers that provide a friendly atmosphere that provide a point of entry into a complex site
- Act as individual and group mirrors
- Fun, not business-like

Hearst & Rosner
HICSS '08



http://www.socialsignal.com/system/files/images/2008-08-01-tagcloud.gif

Wordle

http://www.wordle.net

Wordle

- 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

Layout Algorithm

- Details not published
- Idea:
  - sort words by weight, decreasing order
    for each word w
      w.position := makeInitialPosition(w);
      while w intersects other words:
        updatePosition(w);
  - Init position randomly chosen according to distribution for target shape
  - Update position moves out radially

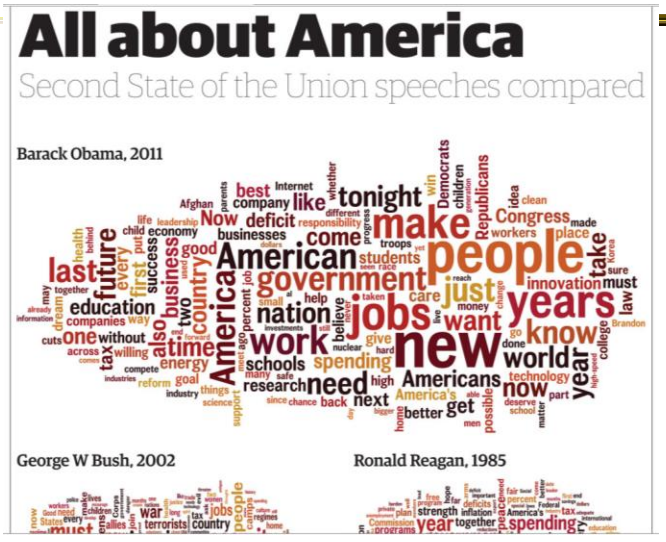
Fun Uses

- Political speeches
- Songs and poems
- Love letters (for "boyfriend points")
- Wedding vows
- Course syllabi
- Teaching writing
- Gifts

2-day Survey in Jan. 09

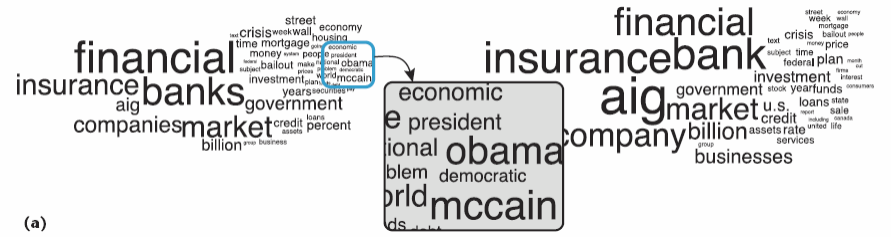
- 2/3 respondents were women
- Interest came from design, visual appeal, beauty
- Why preferred over word clouds:
  - Emotional impact
  - Attention-keeping visuals
  - Organic, non-linear
- Fair percentage didn't know what size signified

SoTU Wordles



http://www.guardian.co.uk/news/datablog/2011/jan/25/state-of-the-union-text-obama#

A Little More Order



Order the words more by frequency

Cui et al
IEEE CG&A '10

Wordle Characteristics

- Layout, words are automatic
- If you had some control, what would you like to change or alter?

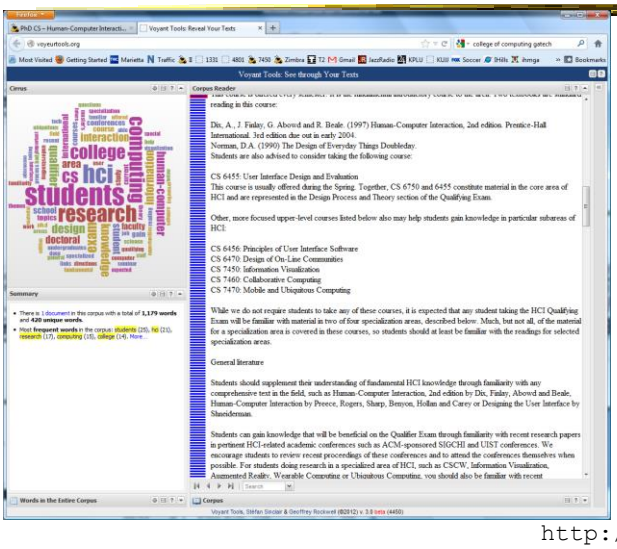
Mani-Wordle

- Start with nice default algorithm
- Give user more control over design
  - Alter color (within a palette)
  - Pin words, redo the rest
  - Move and rotate words
  - Smooth animation and collision detection for tracking changes

Koh et al
TVCG (InfoVis) '10

Video

Text Analysis on Web

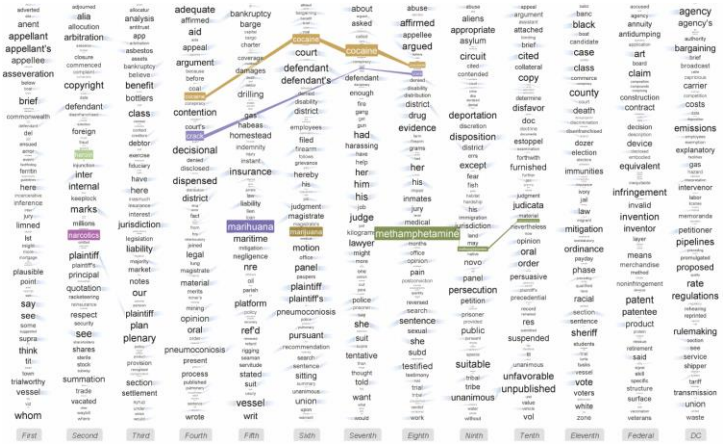


http://voyeurtools.org/

Multiple Documents?

- How to show word frequencies across multiple related documents?

Parallel Tag Clouds



Different circuit courts

Video

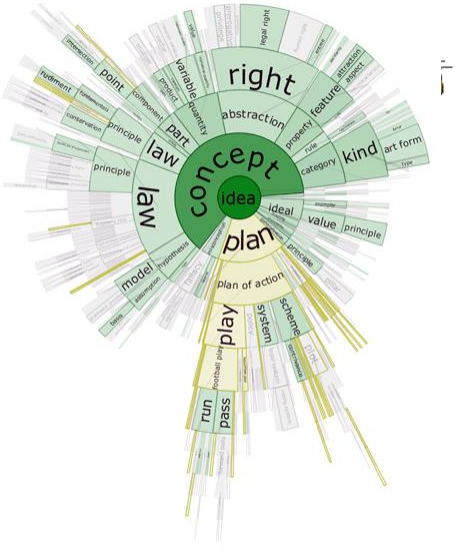
Collins et al
VAST '09

Analytic Support

- Note: Word Clouds and Wordles are really more overview-style visualizations
  - Don't really support queries, searches, drill-down
- How might we also support queries and search?

DocuBurst

Uses WordNet, sets of synonyms grouped together

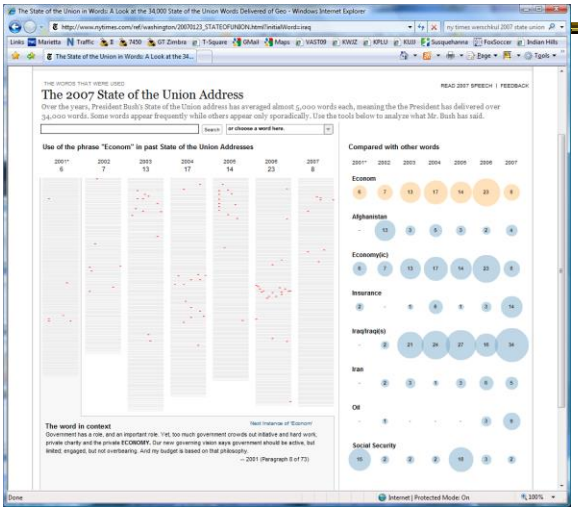


Size – # of leaves in subtree
Hue – diff synsets of word
Shade – frequency of use

Collins et al
EuroVis '09

http://faculty.uoit.ca/collins/research/docuburst

Overview & Timeline

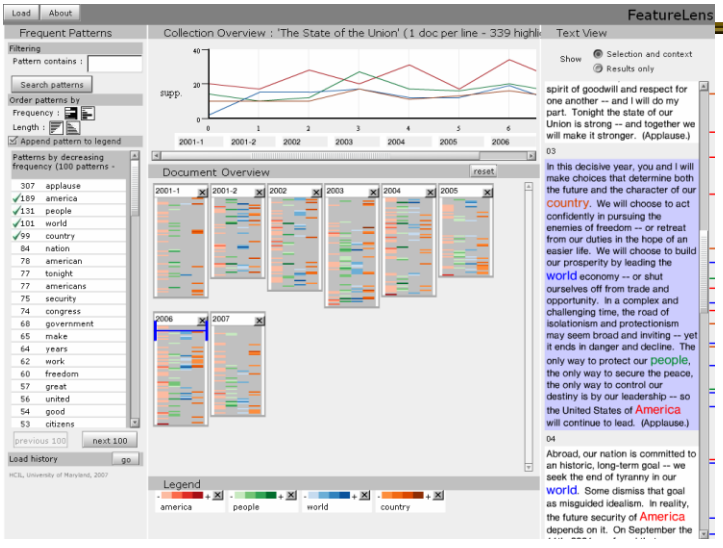


State of the Union Addresses

http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq

FeatureLens

Video

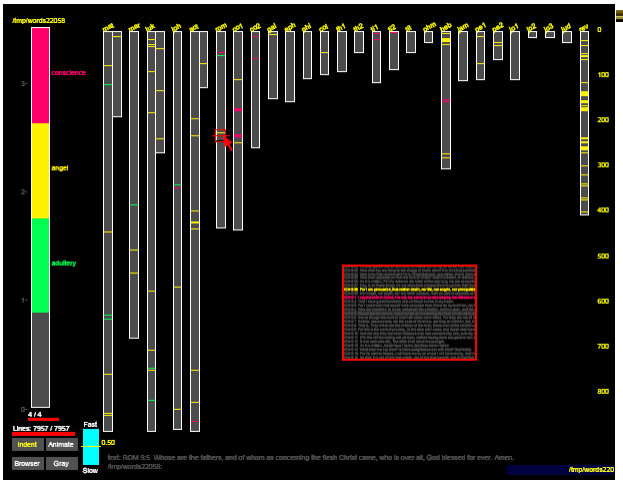


Show patterns of words or n-grams

Don et al CIKM '07

http://www.cs.umd.edu/hcil/textvis/featurelens/

SeeSoft Display



Like taping text to the wall and walking far away

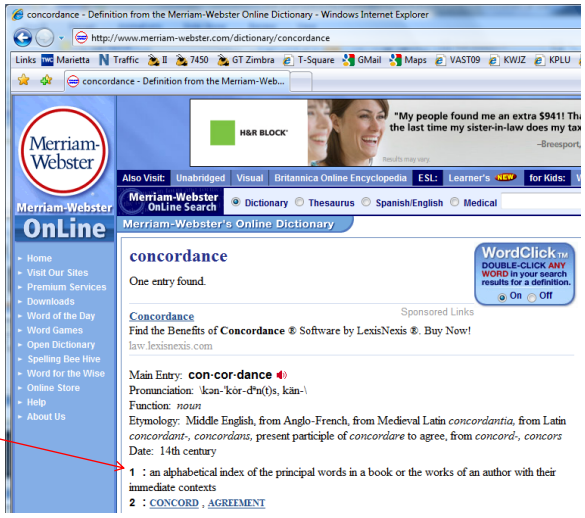
New Testament

Eick
Journal Comput. & Graph. Stats '94

Beyond Individual Words

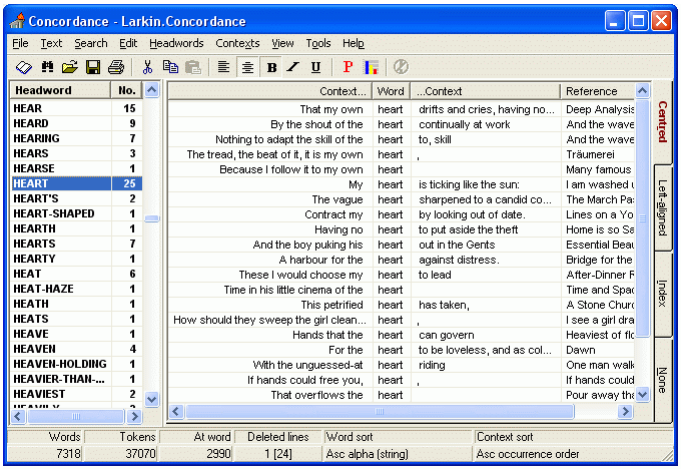
- Can we show combinations of words, phrases, and sentences?

Concordance



Definition

Concordance in Text



http://www.concordancesoftware.co.uk

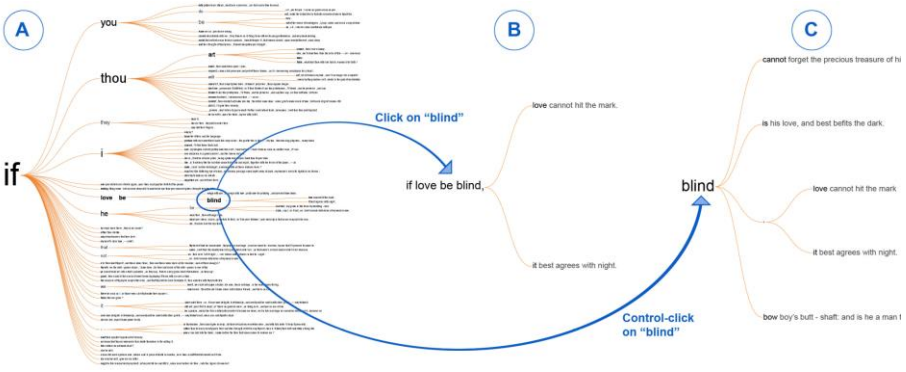
Word Tree

Word Tree

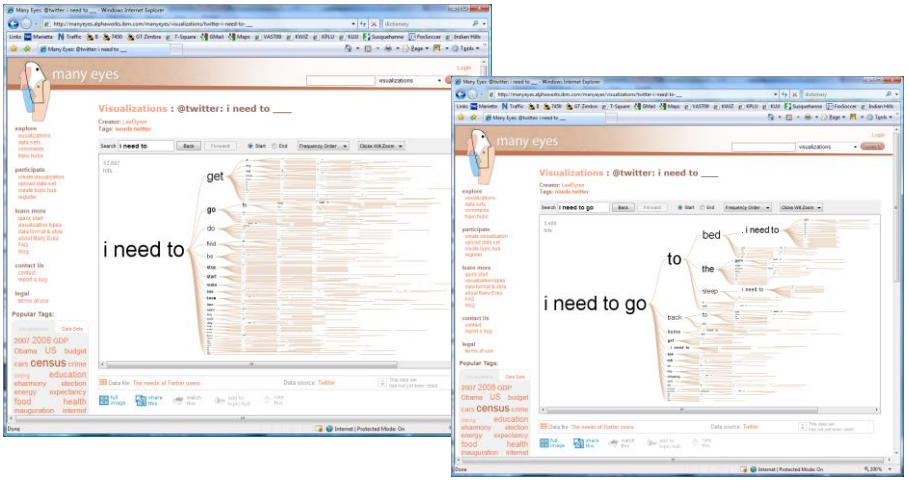
- Shows context of a word or words
  - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus
- Ordered alphabetically, by frequency, or by first appearance

Wattenberg & Viégas
TVCG (InfoVis) '08

Interaction



Many Eyes' WordTree



Phrase Nets

- Examine unstructured text documents
- Presents pairs of terms from phrases such as
  - X and Y
  - X's Y
  - X at Y
  - X (is|are|was|were) Y
- Uses special graph layout algorithm with compression and simplification

van Ham et al
TVCG (InfoVis) '09

Examples

In Many Eyes now

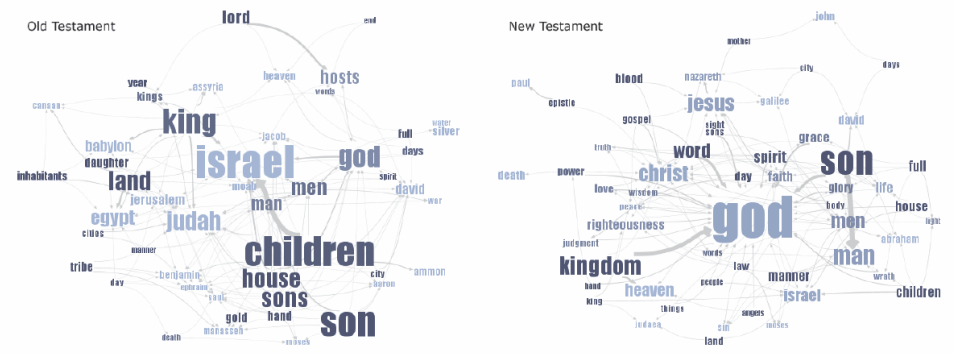


Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

Examples

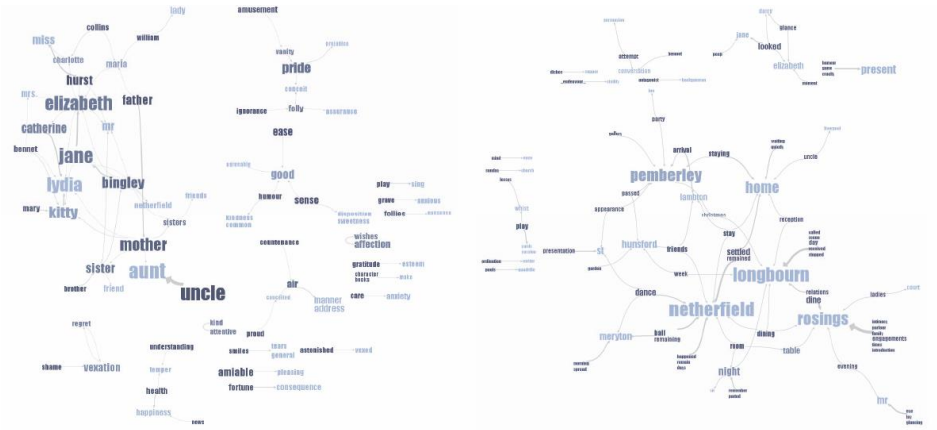


Fig 5. Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

User Interface

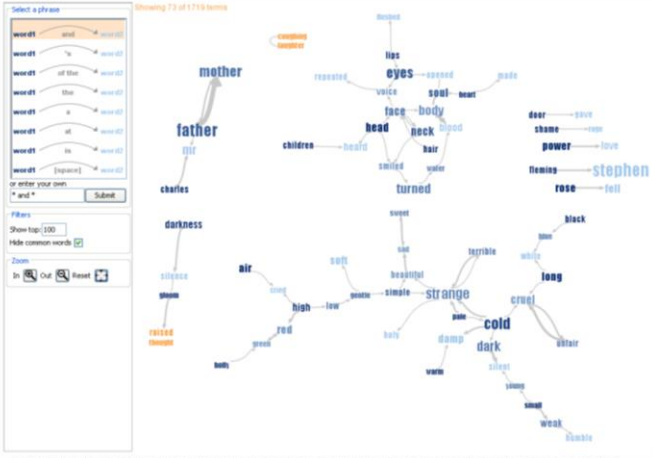


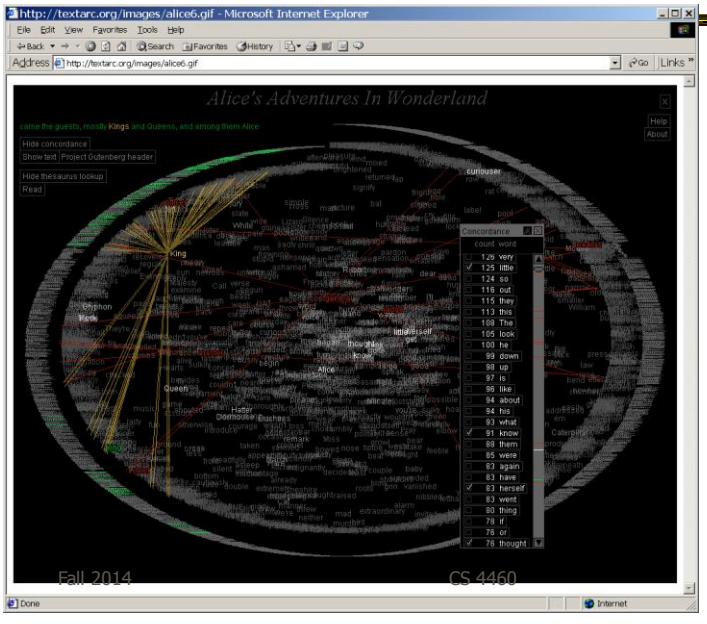
Fig 3. The Phrase Net user interface applied to James Joyce's Portrait of the Artist as a Young Man. The user can select a predefined pattern from the list of patterns on the left or define a custom pattern in the box below. This list of patterns simultaneously serves as a legend, a list of presets and an interactive training mechanism for regular expressions. Here the user has selected "...X and Y...", revealing two main clusters, one almost exclusively consisting of adjectives, the other of verbs and nouns. The highlighted clusters of terms have been aggregated by our edge compression algorithm.

Another Challenge

- Visualize an entire book
- What does that mean?
  - Word appearances
  - Sentences
  - …

TextArc

http://textarc.org



Sentences laid out in order of appearance

Words near to where they appear

Significant interaction

Brad Paley

Next Time

- More about collections of documents and showing other characteristics of documents
  - Analysis metrics
  - Entities
  - Concepts & themes

HW 6

- TV reviews from Amazon
- Design a visualization showing this data
  - Think about what a user would want to know
- Bring 2 copies
- Due Thursday

Project Design Documents

- General thoughts
  - Move beyond just showing data that could be looked up
  - Illuminate trends, patterns, outliners
  - Promote finding insights difficult to discern otherwise
- Grading
  - More about components than judging design

Upcoming

- Text and Documents 2
  - Reading
- Interaction
  - Reading

    Now You See It, chapter 4

    Munzner chapters 11 and 13

References

- Marti Hearst's i247 slides
- All referred to papers

Additional Material

Improving Text Searches

- What's wrong with the common search?
  - Is there really anything wrong?
- Visualizing the results of search queries is one potential important area of text infovis

What Hearst Thinks is Wrong

- Query responses do not include include:
  - How strong the match is
  - How frequent each term is
  - How each term is distributed in the document
  - Overlap between terms
  - Length of document
- Document ranking is opaque
- Inability to compare between results
- Input limits term relationships

Hearst
CHI '95

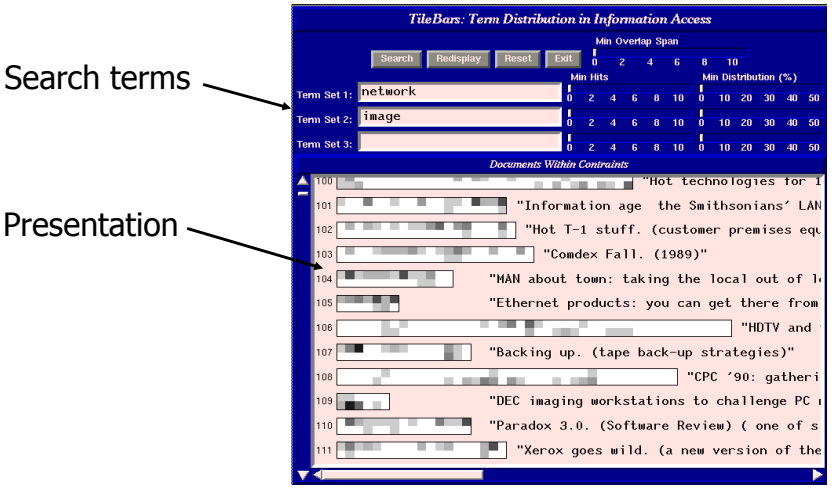
TileBars

- Goal
  - Minimize time and effort for deciding which documents to view in detail
- Idea
  - Show the role of the query terms in the retrieved documents, making use of document structure

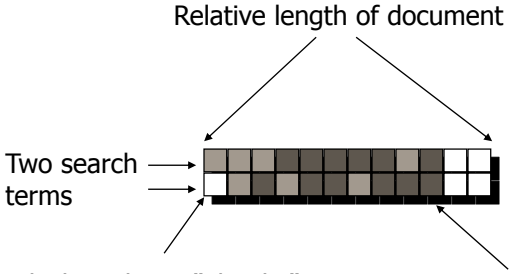
TileBars

- Graphical representation of term distribution and overlap
- Simultaneously indicate:
  - Relative document length
  - Frequency of term sets in document
  - Distribution of term sets with respect to the document and each other

Interface



Technique



Video

Issues

- Horizontal alignment doesn't match mental model
- May not be the best solution for web searches
  - Non-linear material
  - Images? Apps?
- Anything else?