CS-5630 / CS-6630 Visualization for Data Science
Text Visualization

Alexander Lex
alex@sci.utah.edu

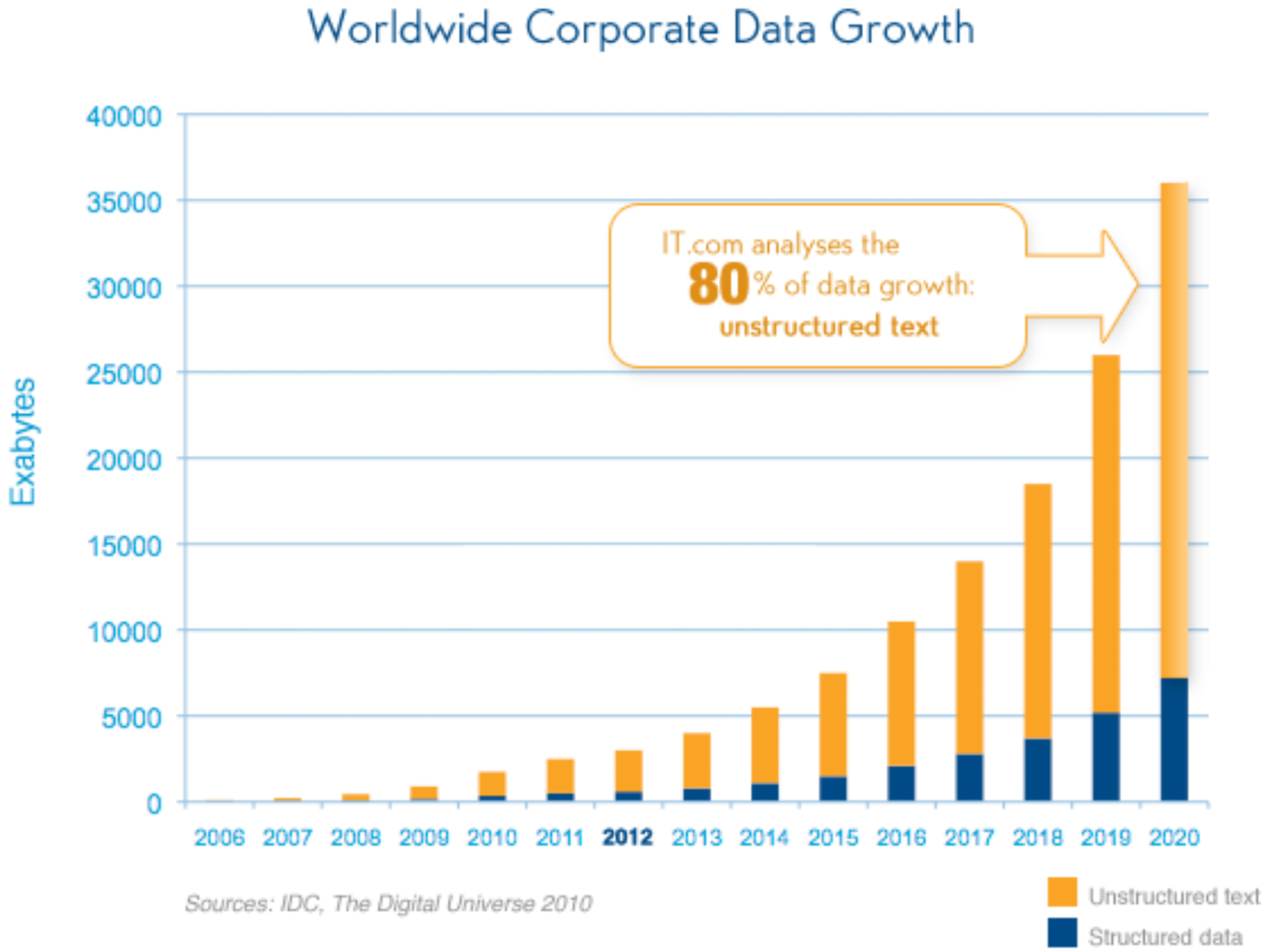
[xkcd]

Text / Language

Features of Text as representation language

- abstract, general
- extremely expressive
- different across population groups (countries, accents, religions,…)
- linear perception
- semi-structured (content: grammar, words, sentences, paragraphs,.. ; appearance: typography, calligraphy,..)

Why Visualize Text?

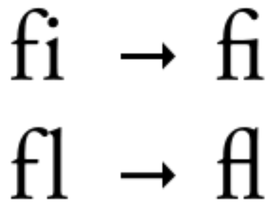


Design and Text

Typography:

- **typefaces** (serif, sans-serif, **bold**, *italic*)
- **point size** (10pt, 12pt, 24pt, 36pt..)
- **line length** (alignment: left, right, justified)
- **vertical**: line spacing (leading)
- **horizontal**: spaces between groups of letters (tracking)
- **Kerning** – space between pairs of letters
- **Ligatures** – combining letters to a glyph

Creating a font type is an art
that requires profound design knowledge



Oscars and Typography

Wrong Movie announced for Best Picture

Failure of Typography

Larger Failures in a Complicated System

A detailed look: https://podcasts.apple.com/us/podcast/la-la-land-galileos-warning/id1484511465?i=1000457519798

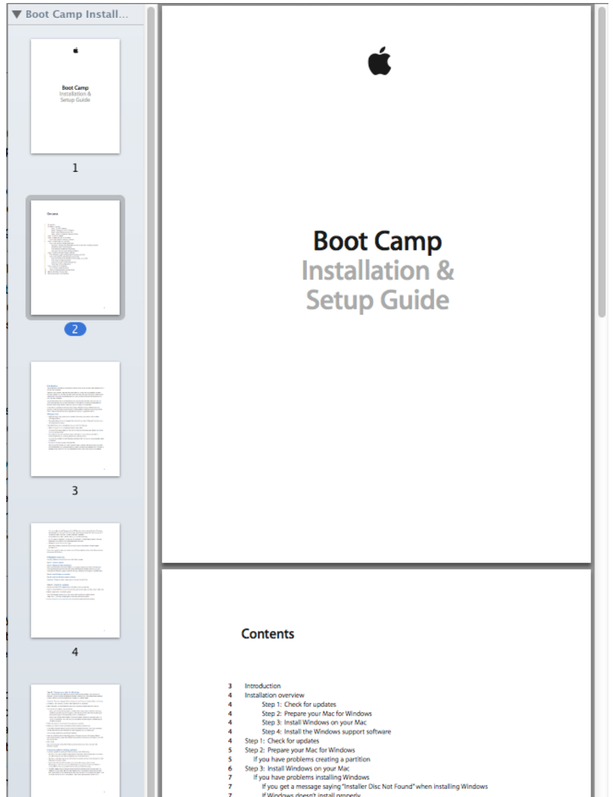
https://www.freecodecamp.org/news/why-typography-matters-especially-at-the-oscars-f7b00e202f22/

Visualization for “Raw” Text

in daily use..

enriched text – hypertext linking (graph navigation)

overview & detail



highlighting semantics

```
void base64_encode(const uint8_t * data, size_t length, char * dst)
{
    size_t src_idx = 0;
    size_t dst_idx = 0;
    for (; (src_idx + 2) < length; src_idx += 3, dst_idx += 4)
    {
        uint8_t s0 = data[src_idx];
        uint8_t s1 = data[src_idx + 1];
        uint8_t s2 = data[src_idx + 2];

        dst[dst_idx + 0] = charset[(s0 & 0xfc) >> 2];
        dst[dst_idx + 1] = charset[((s0 & 0x03) << 4) | ((s1 & 0xf0) >> 4)];
        dst[dst_idx + 2] = charset[((s1 & 0x0f) << 2) | (s2 & 0xc0) >> 6];
        dst[dst_idx + 3] = charset[(s2 & 0x3f)];
    }

    if (src_idx < length)
    {
        uint8_t s0 = data[src_idx];
        uint8_t s1 = (src_idx + 1 < length) ? data[src_idx + 1] : 0;

        dst[dst_idx++] = charset[(s0 & 0xfc) >> 2];
        dst[dst_idx++] = charset[((s0 & 0x03) << 4) | ((s1 & 0xf0) >> 4)];
        if (src_idx + 1 < length)
            dst[dst_idx++] = charset[((s1 & 0x0f) << 2)];
    }
```

Visualization for “Raw” Text

Document Lens

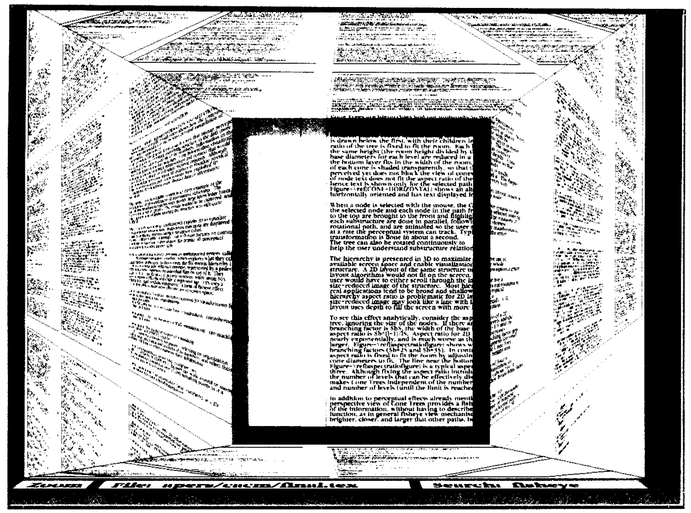
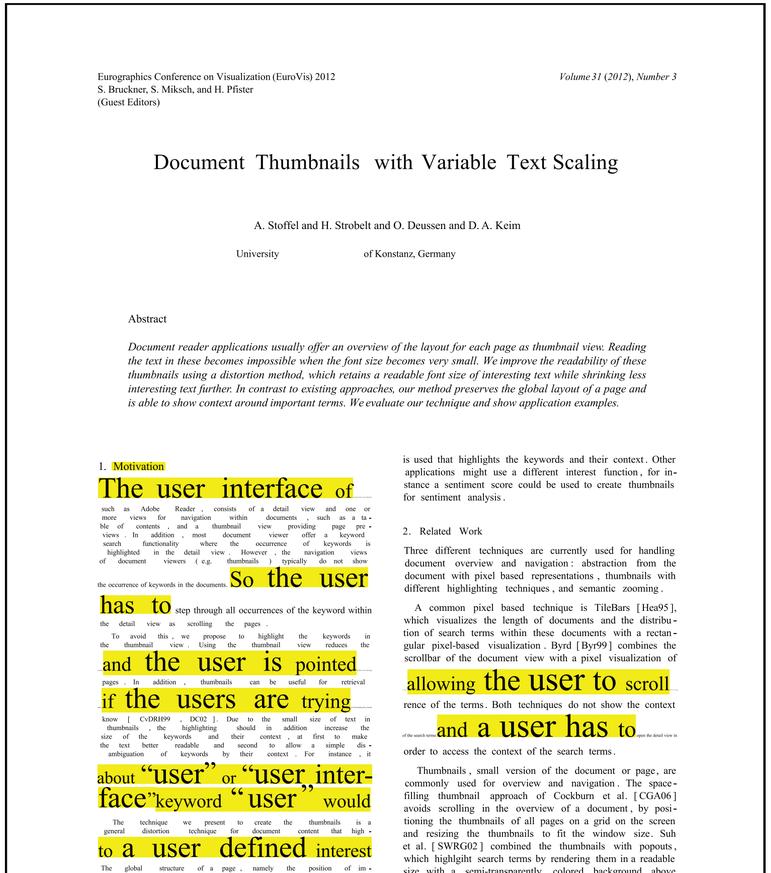


Figure 3: Document Lens with lens pulled toward the user. The resulting truncated pyramid makes text near the lens' edges readable.

Robertson, George G., and Jock D. Mackinlay
The document lens
Proceedings of the 6th annual ACM symposium on User interface software and technology. ACM, 1993.

Visualizing Search Results



Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

Document Thumbnails with Variable Text Scaling

A. Stoffel and H. Strobelt and O. Deussen and D. A. Keim

University of Konstanz, Germany

Abstract

Document reader applications usually offer an overview of the layout for each page as thumbnail view. Reading the text in these becomes impossible when the font size becomes very small. We improve the readability of these thumbnails using a distortion method, which retains a readable font size of interesting text while shrinking less interesting text further. In contrast to existing approaches, our method preserves the global layout of a page and is able to show context around important terms. We evaluate our technique and show application examples.

1. Motivation

The user interface of such as Adobe Reader , consists of a detail view and one or more views for navigation within documents , such as a table of contents , and a thumbnail view providing page previews . In addition , most document viewer offer a keyword search functionality where the occurrence of keywords is highlighted in the detail view . However , the navigation views of document viewers (e.g. thumbnails) typically do not show the occurrence of keywords in the document. So the user has to step through all occurrences of the keyword within the detail view as scrolling the pages .

To avoid this , we propose to highlight the keywords in the thumbnail view . Using the thumbnail view reduces the and the user is pointed pages . In addition , thumbnails can be useful for retrieval if the users are trying know [CvDRI99 , DC02]. Due to the small size of text in thumbnails , the highlighting should in addition increase the size of the keywords and their context , at first to make the text better readable and second to allow a simple disambiguation of keywords by their context . For instance , it about “user” or “user interface”keyword “user” would

The technique we present to create the thumbnails is a general distortion technique for document content that high - to a user defined interest

The global structure of a page , namely the position of im -

is used that highlights the keywords and their context . Other applications might use a different interest function , for instance a sentiment score could be used to create thumbnails for sentiment analysis .

2. Related Work

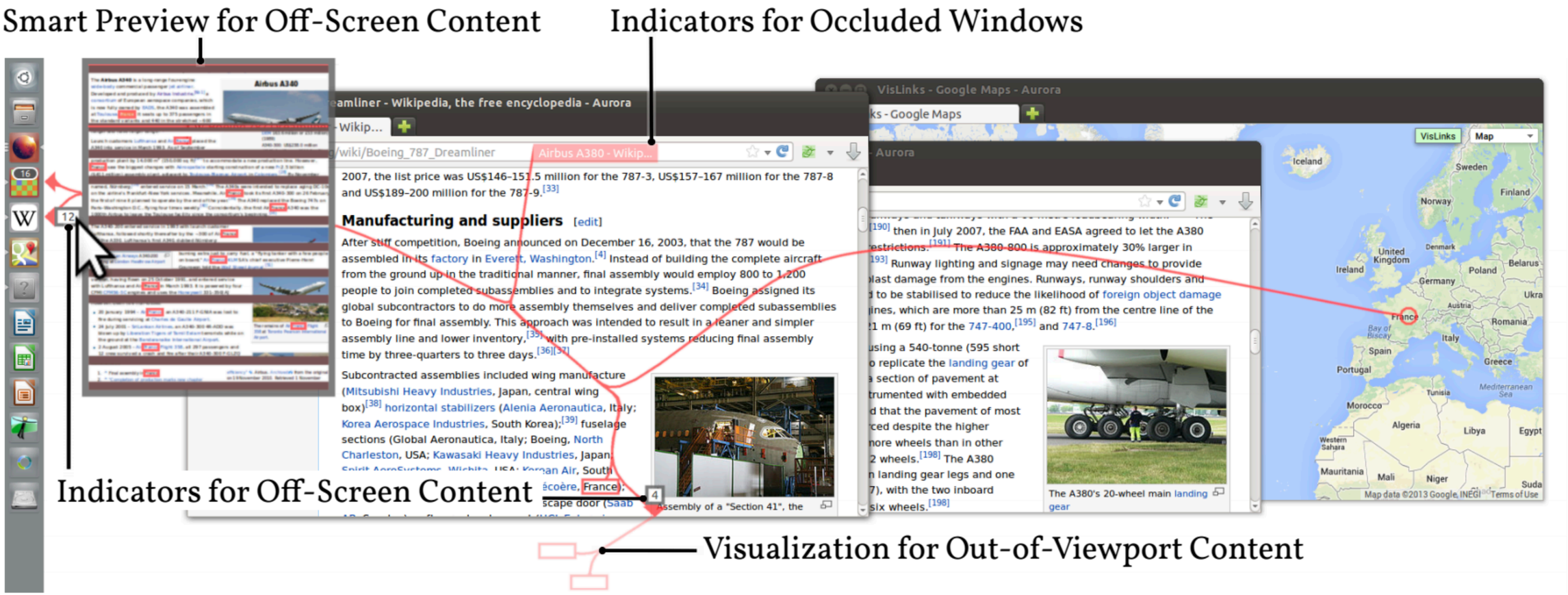
Three different techniques are currently used for handling document overview and navigation : abstraction from the document with pixel based representations , thumbnails with different highlighting techniques , and semantic zooming .

A common pixel based technique is TileBars [Hea95], which visualizes the length of documents and the distribution of search terms within these documents with a rectangular pixel-based visualization . Byrd [Byr99] combines the scrollbar of the document view with a pixel visualization of allowing the user to scroll rence of the terms . Both techniques do not show the context of the search terms and a user has to see the detail view in order to access the context of the search terms .

Thumbnails , small version of the document or page , are commonly used for overview and navigation . The space-filling thumbnail approach of Cockburn et al. [CGA06] avoids scrolling in the overview of a document , by positioning the thumbnails of all pages on a grid on the screen and resizing the thumbnails to fit the window size . Suh et al. [SWRG02] combined the thumbnails with popouts , which highlgiht search terms by rendering them in a readable size with a semi-transparently colored background above

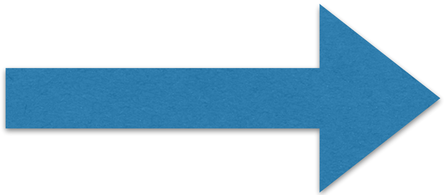
Document Thumbnails with Variable Text Scaling
A. Stoffel, H. Strobelt, O. Deussen, D. A. Keim
Computer Graphics Forum, volume 31 issue 3 pp.

Visualizing Hidden Text



Working with Text

unstructured text



4 x ’t'
3 x ‘u’
2 x ‘r’
2 x ‘e’
…
structured data

Structured Text Features

simple counts (bag of words)
used for similarity measures

	princess	dragon	castle
doc1	1	1	1
doc2	0	0	1

Processing to Derive Features

Typical steps are:

- cleaning (regular expressions)
- sentence splitting
- change to lower case
- stopword removal (most frequent words in a language)
- stemming
- POS tagging (part of speech)
- noun chunking
- NER (name entity recognition)
- deep parsing - try to “understand” text.

Text features are complicated

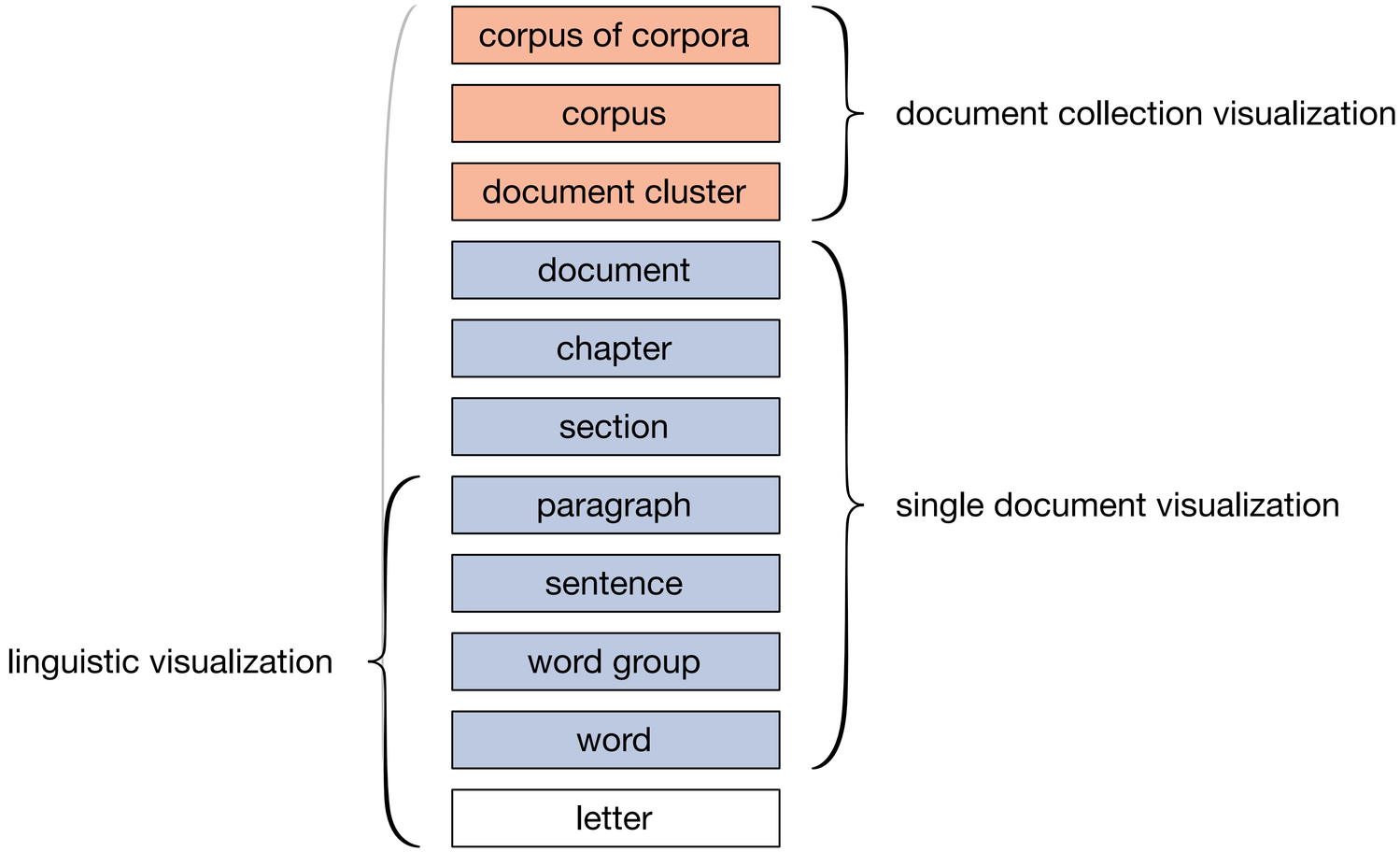
Toilet out of order. Please use floor below.

One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know.

Did you ever hear the story about the blind carpenter who picked up his hammer and saw?

http://en.wikipedia.org/wiki/List_of_linguistic_example_sentences

Text Units Hierarchy



Types of Text Visualizations

Document Visualization

Corpus Visualization

Visualization for NLP

Creativity Support

Document Visualization

Wordle

Frequency-based

words that occur often are large

Can vary font type,
size, color, etc.

Uses stop-word removal

Create a Wrodle at http://www.edwordle.net/index.html

[Viegas 2009]

Wordle vs Tag Cloud

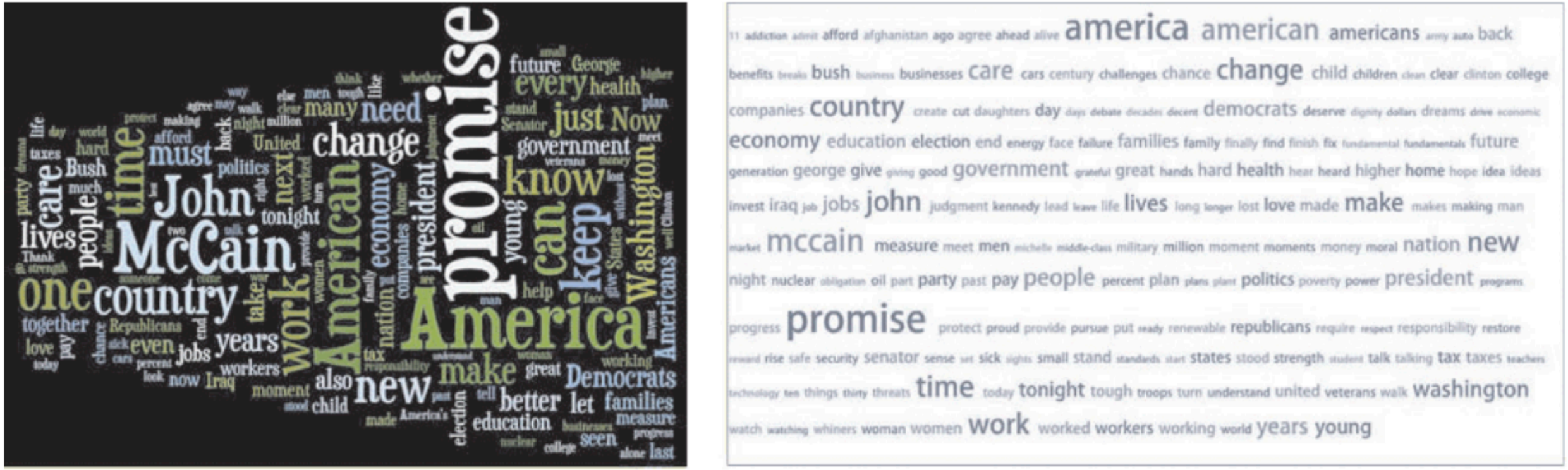


Fig 2: Wordle vs. Tag Cloud of Barack Obama's speech at the Democratic Convention in 2008.

Opinion

Use Wordle if you want something evocative.

Don’t use Tag Cloud! (Looks bad, not very useful)

Use structured approach instead

- Top keywords with counts
- Maybe group by topics


PARTY
CRAFT
BIRTHDAY
BAGS
BOX
KIDS
COLORED
ACTIVITY
TWEEN
CRAFTING
PAPER
GIFT
SALT
BAG
MAKING
CRAFTY
FUN
PLASTIC
CHALK
USED
EASY
happy
NIECE
JUST
SUBSCRIBE
WEEKEND
PERSONAL
POST
LIKED
MAKE
PACKED
WELL
MAILING

ideas
great
idealings
idea
imagination
friends
environment
Napoleon
creative
good
Steven
like
Johnson
time
one
Hill
mind
bouncing
person
new
come
word
flash

malice
towards
none
God
war
shall
nation
Union
right
offenses
years
Woe
wills
come
gives
came
extend
hope
men
public
needs
cause

Word Tree

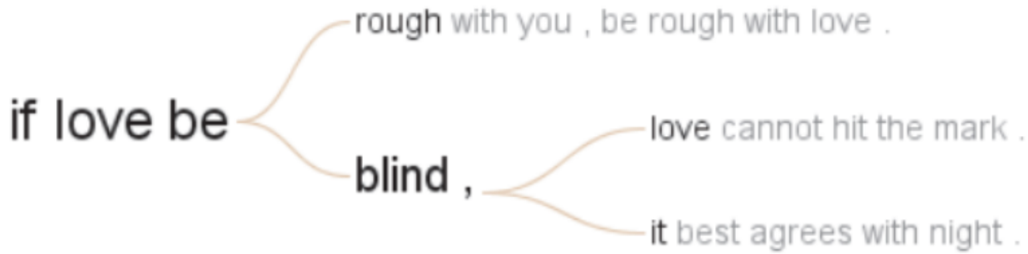
Text

if love be rough with you , be rough with love .

if love be blind , love cannot hit the mark .

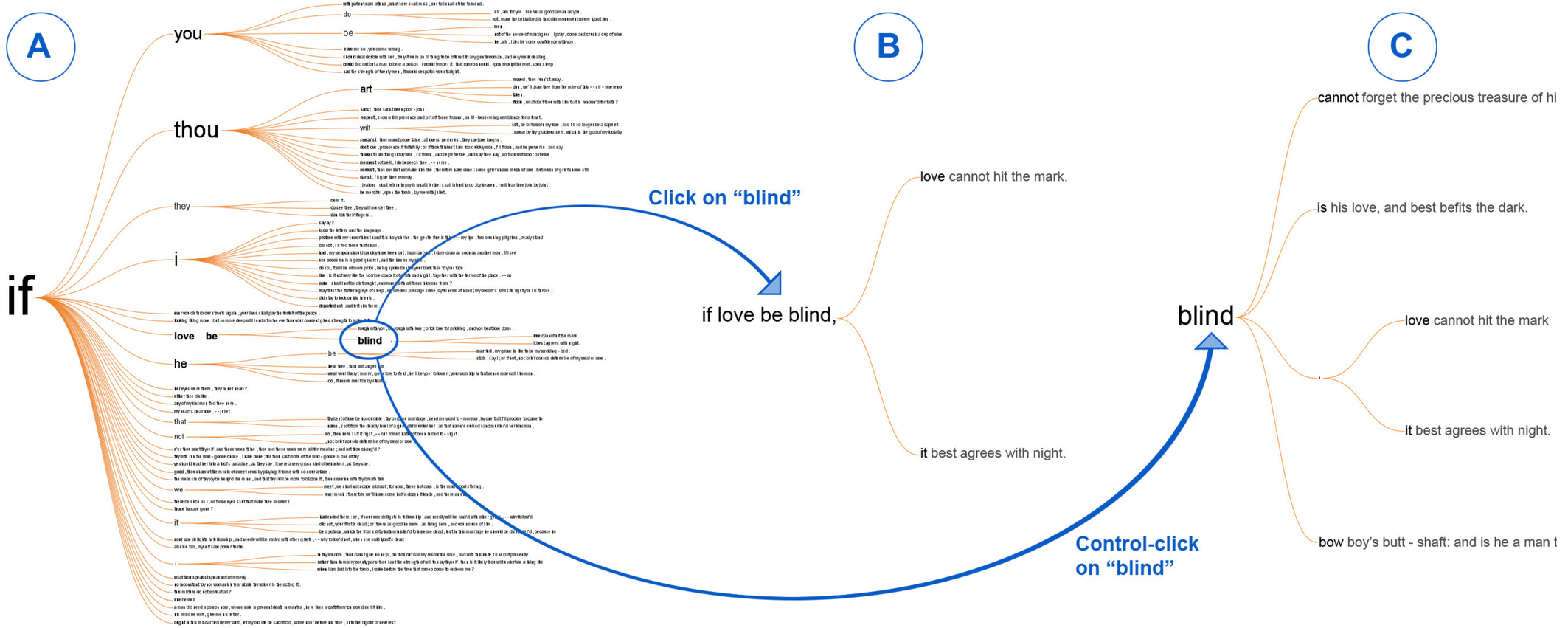
if love be blind , it best agrees with night .

WordTree



[Wattenberg 2008]

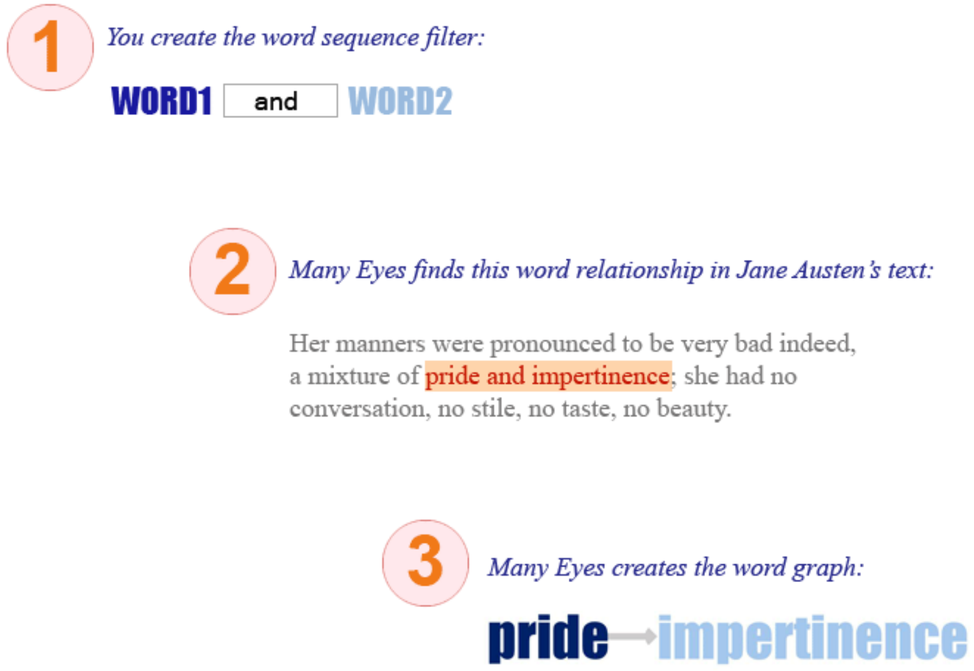
Search for “if” in romeo & Juliet



The word tree, an interactive visual concordance
M Wattenberg, FB Viégas
Visualization and Computer Graphics, IEEE Transactions on 14 (6), 1221-1228

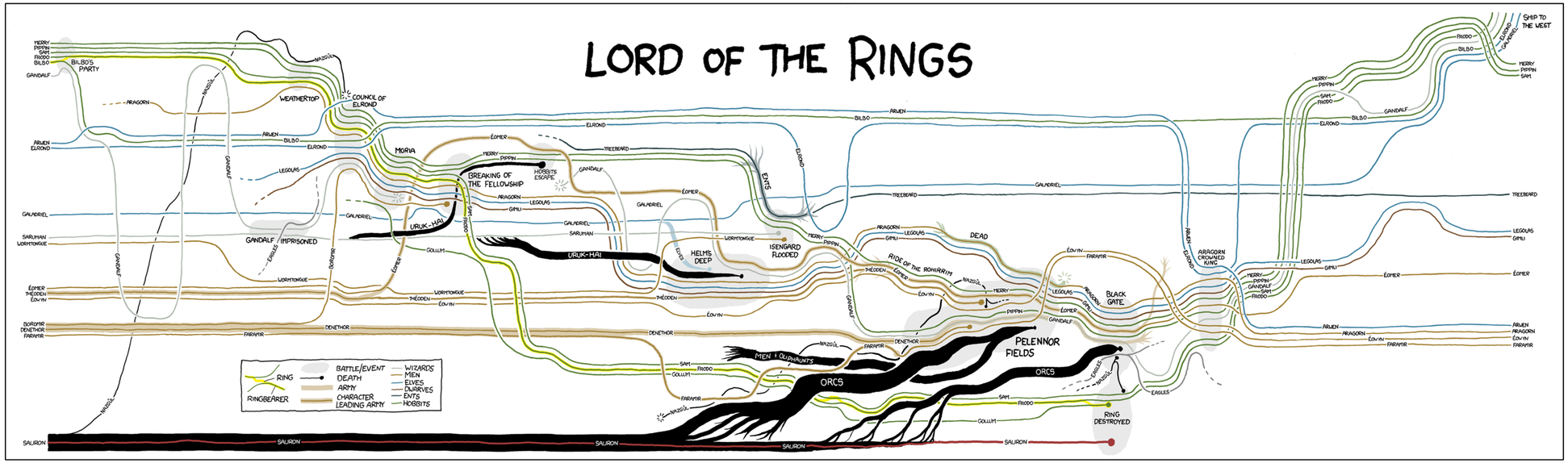
Phrase Net for Bible and “begat”

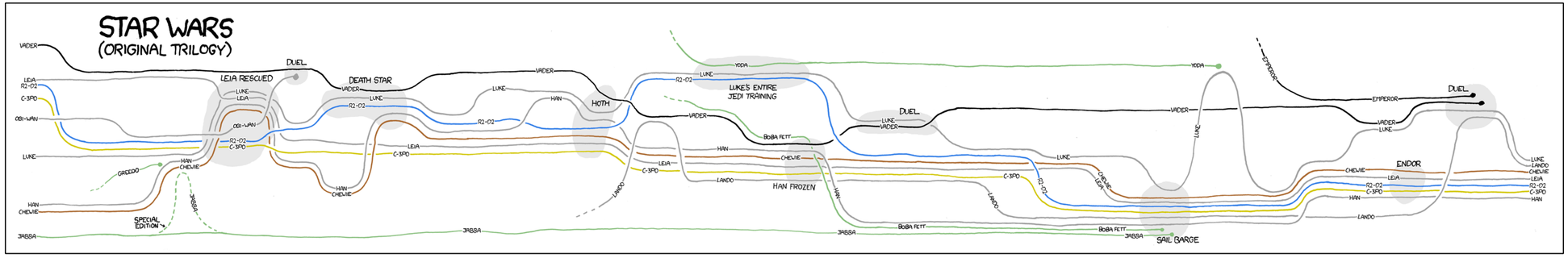
PhraseNets

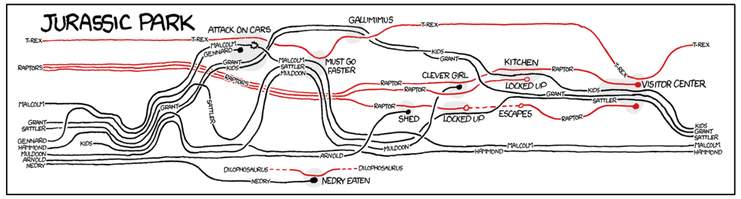


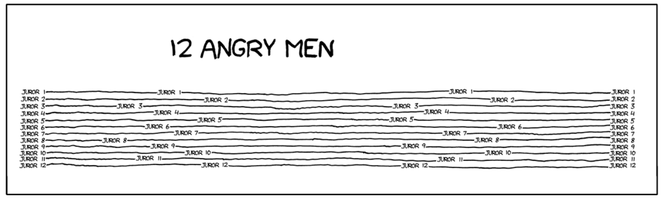
Frank van Ham, Martin Wattenberg, and Fernanda B. Viegas.
Mapping Text with Phrase Nets.
IEEE Transactions on Visualization and Computer Graphics 15, 6 (November 2009)

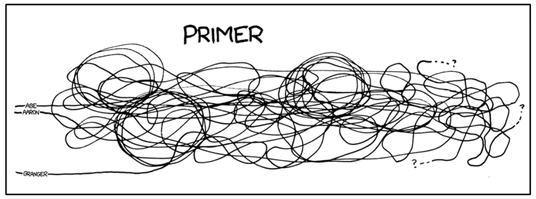
THESE CHARTS SHOW MOVIE CHARACTER INTERACTIONS.
THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE
LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.





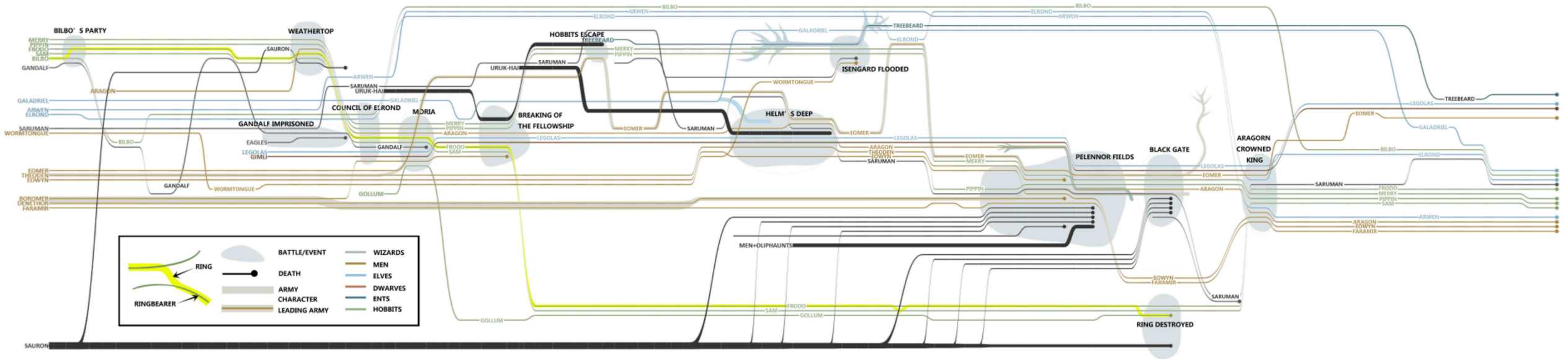






https://xkcd.com/657/

StoryFlow: Tracking the Evolution of Stories



[Liu 2013]

Visualizing Corpora

Text Corpora

Varied Goals:

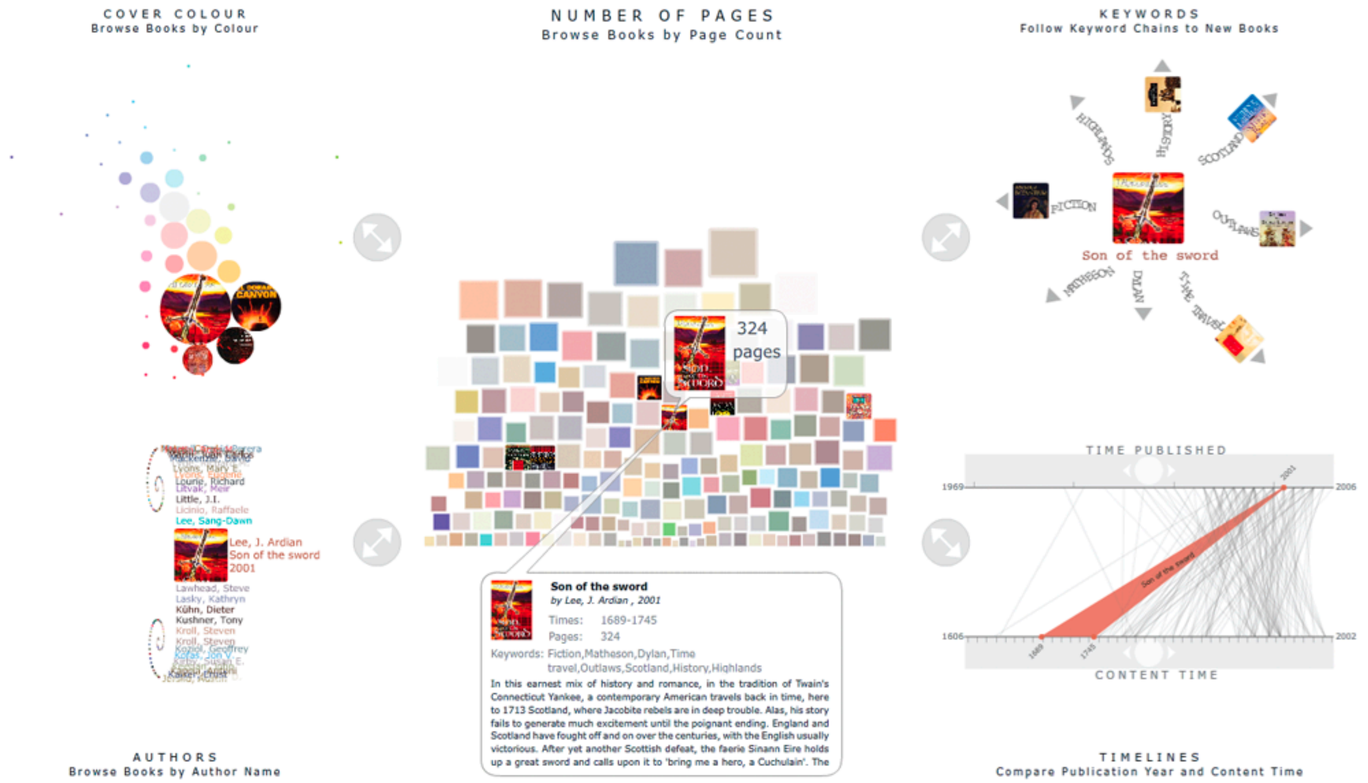
- Discover interesting documents
- Summarize Documents
- Classify Documents
- Extract Facts (Intelligence Analysis)

Rich Information:

- Document Metadata
  - Authors, date, type,
- Paragraphs, figures…
- Revisions, annotations, comments,

Visualize Library Catalogs

Bohemian Bookshelf

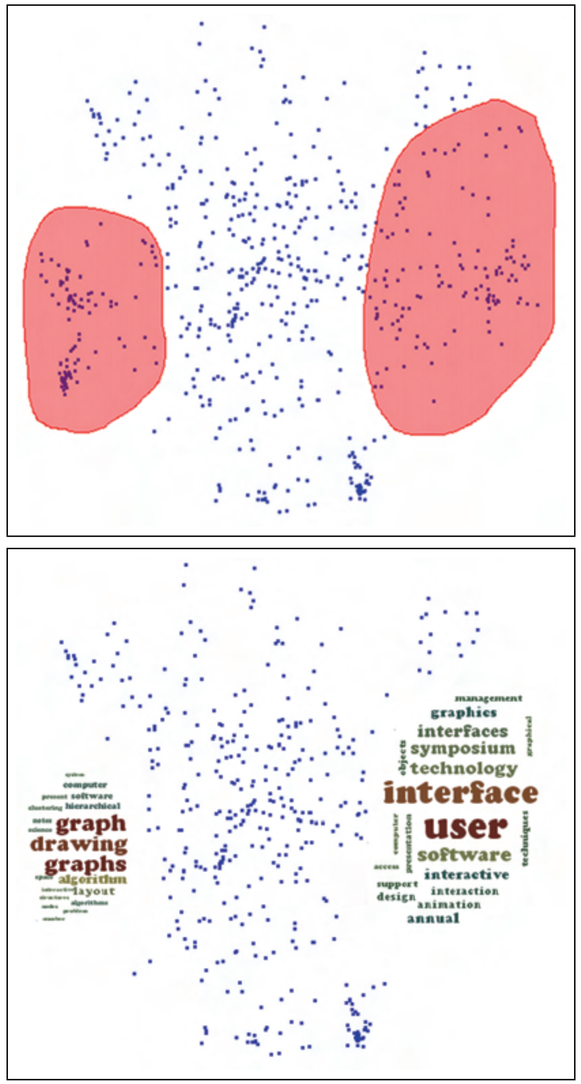


[Thudt, Hinrichs, Carpendale, 2012]

Corpora: MDS Approaches

use bag-of-word to project documents w.r.t. text similarity into a landscape

(only) one example



Fernando V. Paulovich, Franklina M. B. Toledo, Guilherme P. Telles, Rosane Minghim, and Luis Gustavo Nonato.
Semantic Wordification of Document Collections.
Comp. Graph. Forum 31, 3pt3 (June 2012)

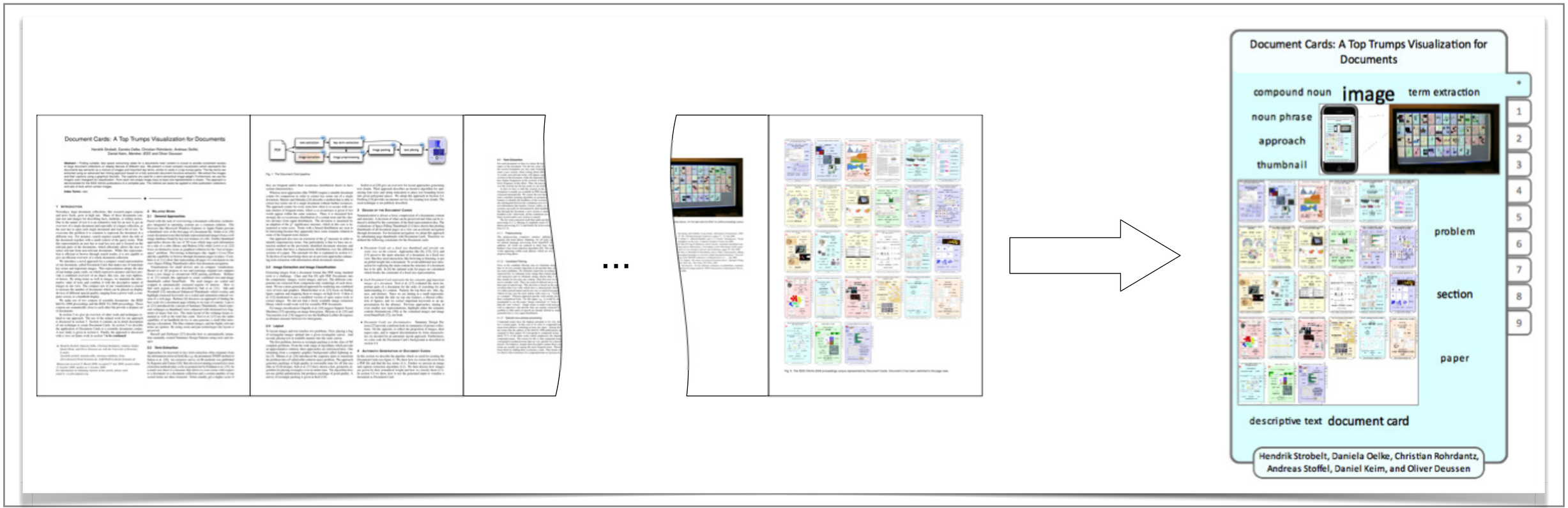
Figure 5: *A user can interactively draw a region (polygon) containing a subset of documents of interest (top figure). Keywords are extracted from the selected document and their corresponding word could is built inside the user-defined region (bottom figure).*

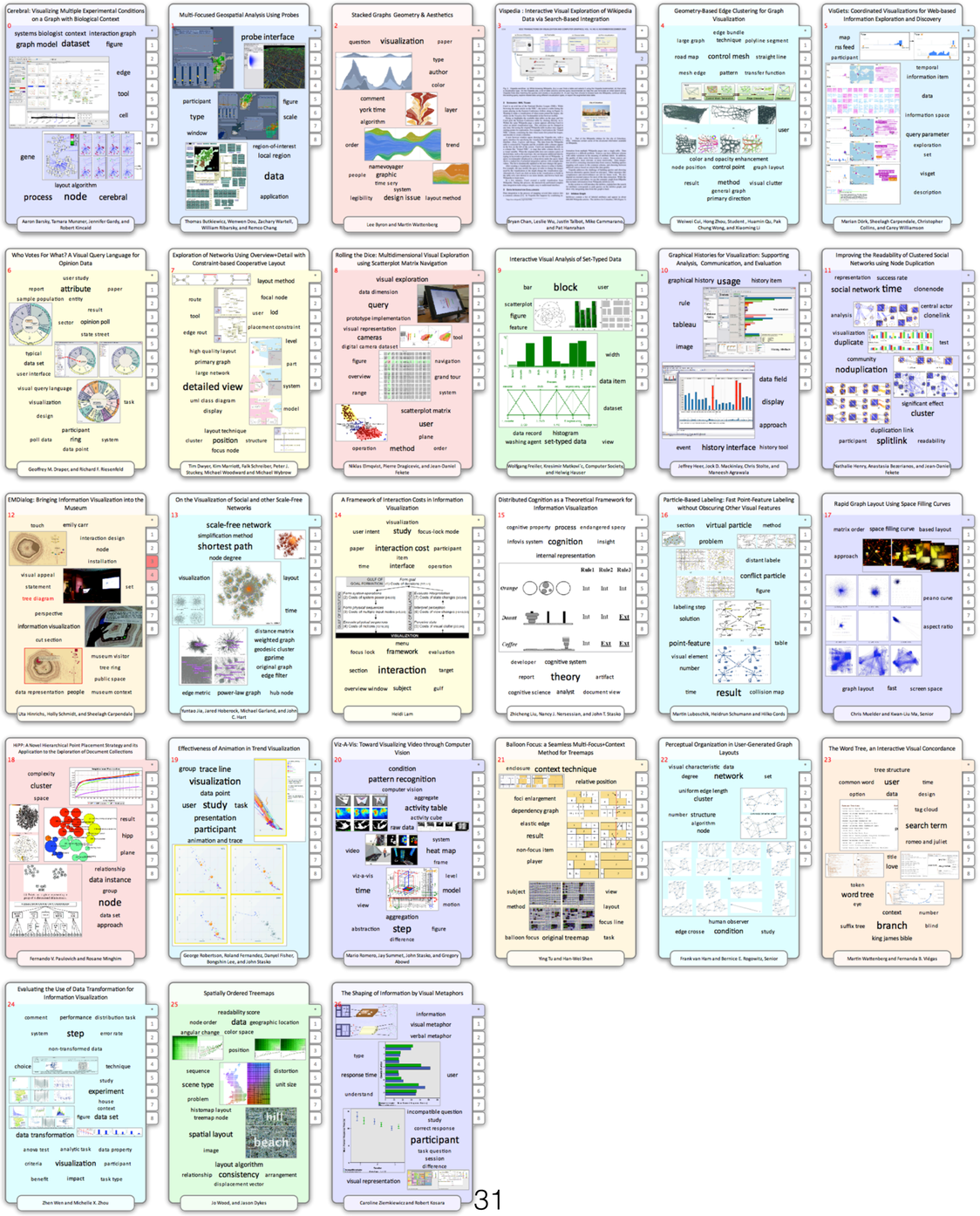
DocumentCards

[Strobelt et al]

summarize scientific documents using important terms and important figures

represent the document's content as a mix of figure and text




Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context
Aaron Barsky, Tamara Munzner, Jennifer Gardy, and Robert Kincaid
Multi-Focused Geospatial Analysis Using Probes
Thomas Butkiewicz, Wenwen Dou, Zachary Wartell, William Ribarsky, and Remco Chang
Stacked Graphs Geometry & Aesthetics
Lee Byron and Martin Wattenberg
Vispedia : Interactive Visual Exploration of Wikipedia Data via Search-Based Integration
Bryan Chan, Leslie Wu, Justin Talbot, Mike Cammarano, and Pat Hanrahan
Geometry-Based Edge Clustering for Graph Visualization
Weiwei Cui, Hong Zhou, Student , Huamin Qu, Pak Chung Wong, and Xiaoming Li
VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery
Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson
Who Votes For What? A Visual Query Language for Opinion Data
Geoffrey M. Draper, and Richard F. Riesenfeld
Exploration of Networks Using Overview+Detail with Constraint-based Cooperative Layout
Tim Dwyer, Kim Marriott, Falk Schreiber, Peter J. Stuckey, Michael Woodward and Michael Wybrow
Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation
Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete
Interactive Visual Analysis of Set-Typed Data
Wolfgang Freiler, Kresimir Matkovi´c, Computer Society, and Helwig Hauser
Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation
Jeffrey Heer, Jock D. Mackinlay, Chris Stolte, and Maneesh Agrawala
Improving the Readability of Clustered Social Networks using Node Duplication
Nathalie Henry, Anastasia Bezerianos, and Jean-Daniel Fekete
EMDialog: Bringing Information Visualization into the Museum
Uta Hinrichs,, Holly Schmidt, and Sheelagh Carpendale
On the Visualization of Social and other Scale-Free Networks
Yuntao Jia, Jared Hoberock, Michael Garland, and John C. Hart
A Framework of Interaction Costs in Information Visualization
Heidi Lam
Distributed Cognition as a Theoretical Framework for Information Visualization
Zhicheng Liu, Nancy J. Nersessian, and John T. Stasko
Particle-Based Labeling: Fast Point-Feature Labeling without Obscuring Other Visual Features
Martin Luboschik, Heidrun Schumann and Hilko Cords
Rapid Graph Layout Using Space Filling Curves
Chris Muelder and Kwan-Liu Ma, Senior
HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections
Fernando V. Paulovich and Rosane Minghim
Effectiveness of Animation in Trend Visualization
George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko
Viz-A-Vis: Toward Visualizing Video through Computer Vision
Mario Romero, Jay Summet, John Stasko, and Gregory Abowd
Balloon Focus: a Seamless Multi-Focus+Context Method for Treemaps
Ying Tu and Han-Wei Shen
Perceptual Organization in User-Generated Graph Layouts
Frank van Ham and Bernice E. Rogowitz, Senior
The Word Tree, an Interactive Visual Concordance
Martin Wattenberg and Fernanda B. Viégas
Evaluating the Use of Data Transformation for Information Visualization
Zhen Wen and Michelle X. Zhou
Spatially Ordered Treemaps
Jo Wood, and Jason Dykes
The Shaping of Information by Visual Metaphors
Caroline Ziemkiewicz and Robert Kosara

Compare Corpora

Compare topics between text collections

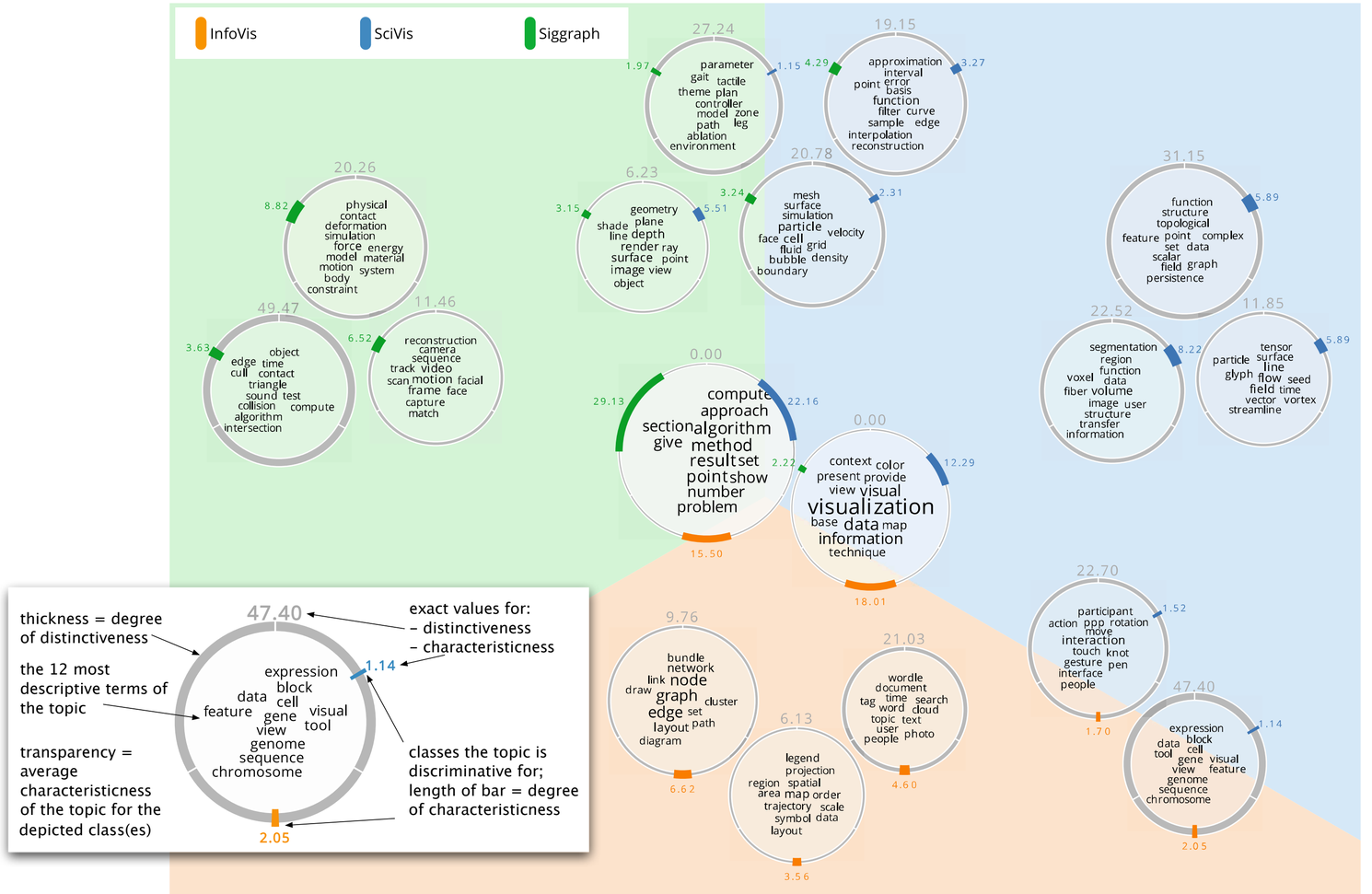
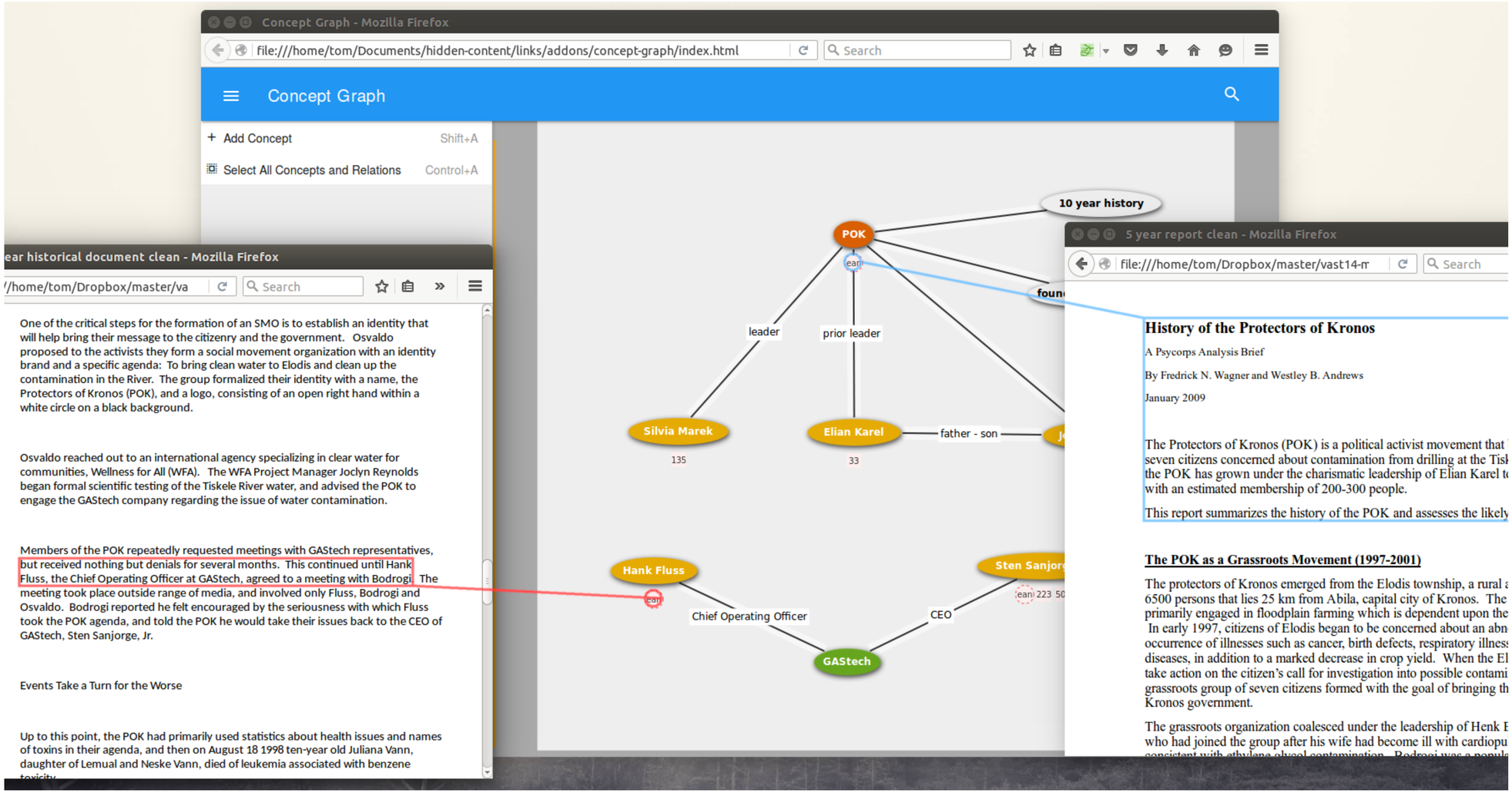


Figure 1: Comparison of 495 papers of InfoVis, SciVis, and Siggraph (discrimination threshold = 6, number of topics = 30)

Comparative Exploration of Document Collections: a Visual Analytics Approach (http://ditop.hs8.de)
D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen

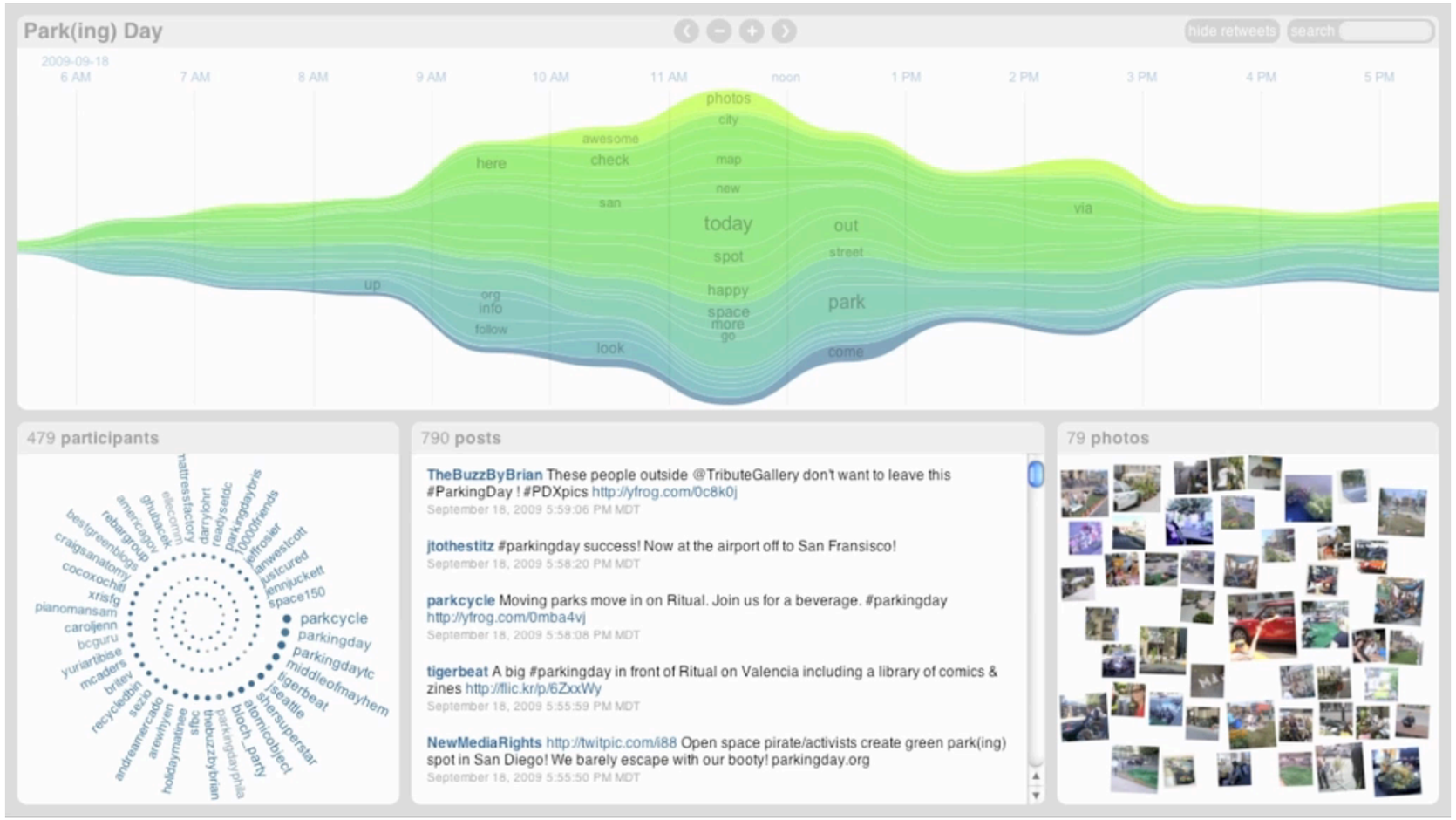
JigSaw – Intelligence Analysis

Extracting and Linking Info From Documents



[Geymayer, 2017]

Collection of Tweets



Marian Dörk, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. **A Visual Backchannel for Large-Scale Events**. TVCG: Transactions on Visualization and Computer Graphics (Proceedings Information Visualization 2010

Visualization for NLP

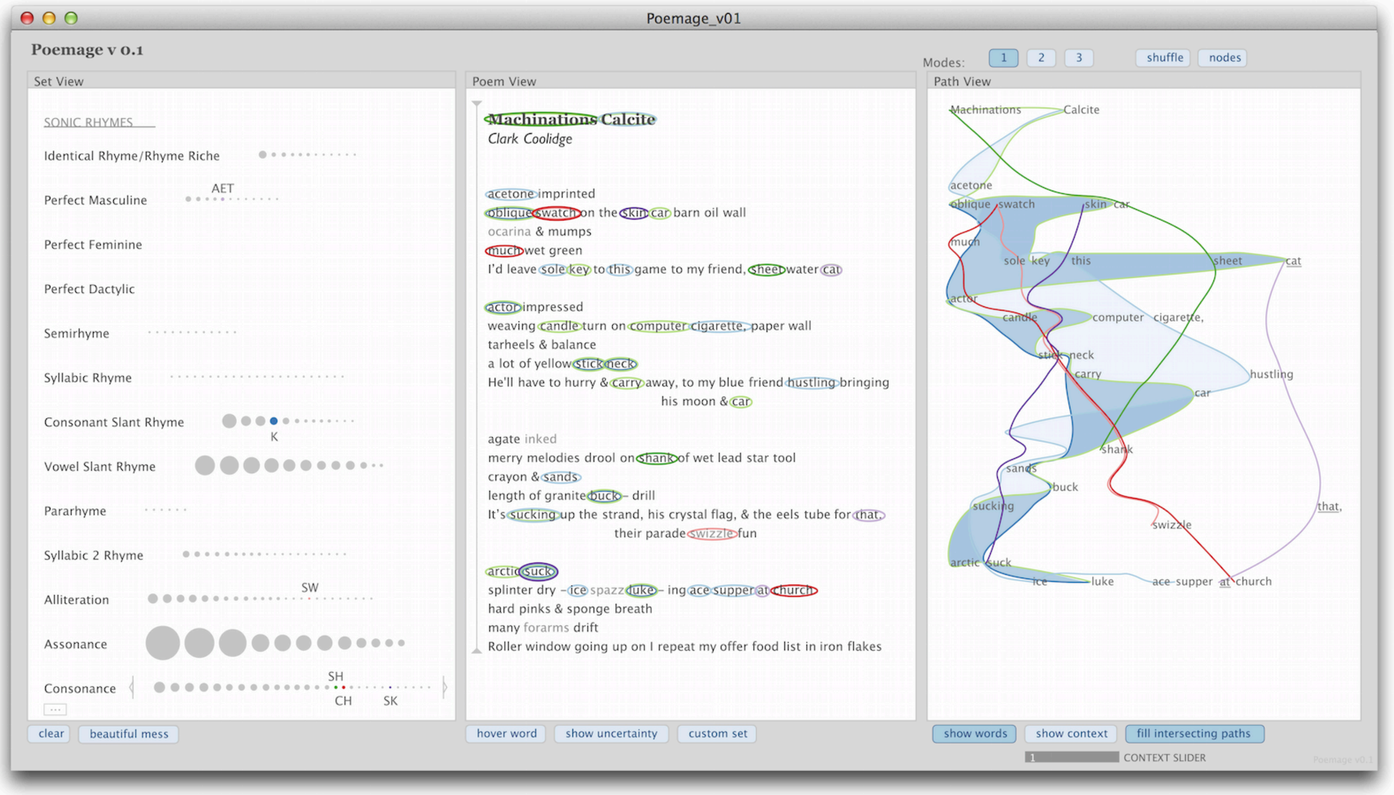
GLTR: Statistical Detection and Visualization of Generated Text, Gehrmann, Strobelt Rush: http://gltr.io/dist/index.html

LSTMVis: Visual Analysis for Recurent Neural Networks, Strobelt et al.: http://lstm.seas.harvard.edu/

Visual Exploration of Semantic Relationships in Neural Word Embeddings. Liu et al.

Visualization for Creativity Support

Poemage: Visualizing the Sonic Topology of a Poem. McCurdy et al. http://www.sci.utah.edu/~nmccurdy/Poemage/



http://textvis.lnu.se/