VISUALIZING TEXT

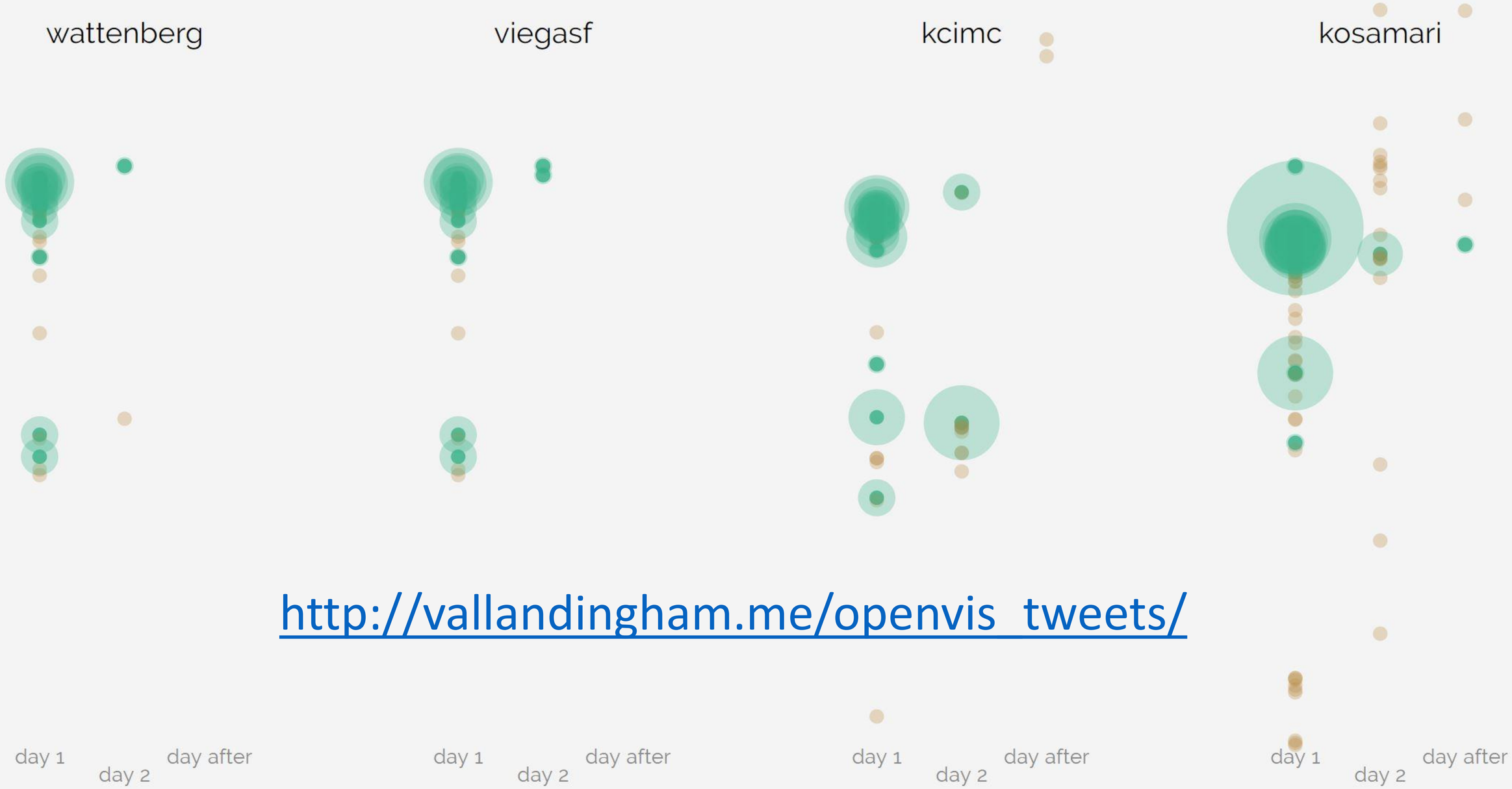
Slides adapted
from Hendrik Strobelt, Irene Ross, Jeff Herr and Alex Lex

You are scrapping twitter for tweets about **Bob Ross** to create a visualization communicating the **time, number of retweets** and whether the tweet is **positive, negative, or neutral**.
So far you've scraped:

Topic	**# of tweets**	**sentiment**
Happy trees	5	3 neutral, 2 positive
Bob's hair	6	3 positive, 3 negative
Bob being a former drill Sargent (Yes this is true).	7	1 positive, 4 neutral, 2 negative

This captures the popularity of the content produced by or for a speaker, but not the popularity of their talk.

2016 ▼



http://vallandingham.me/openvis_tweets/

Challenges of text and language

abstract and unstructured

Structured Data

- Graphs
- Tables

Unstructured Data

Video

Images

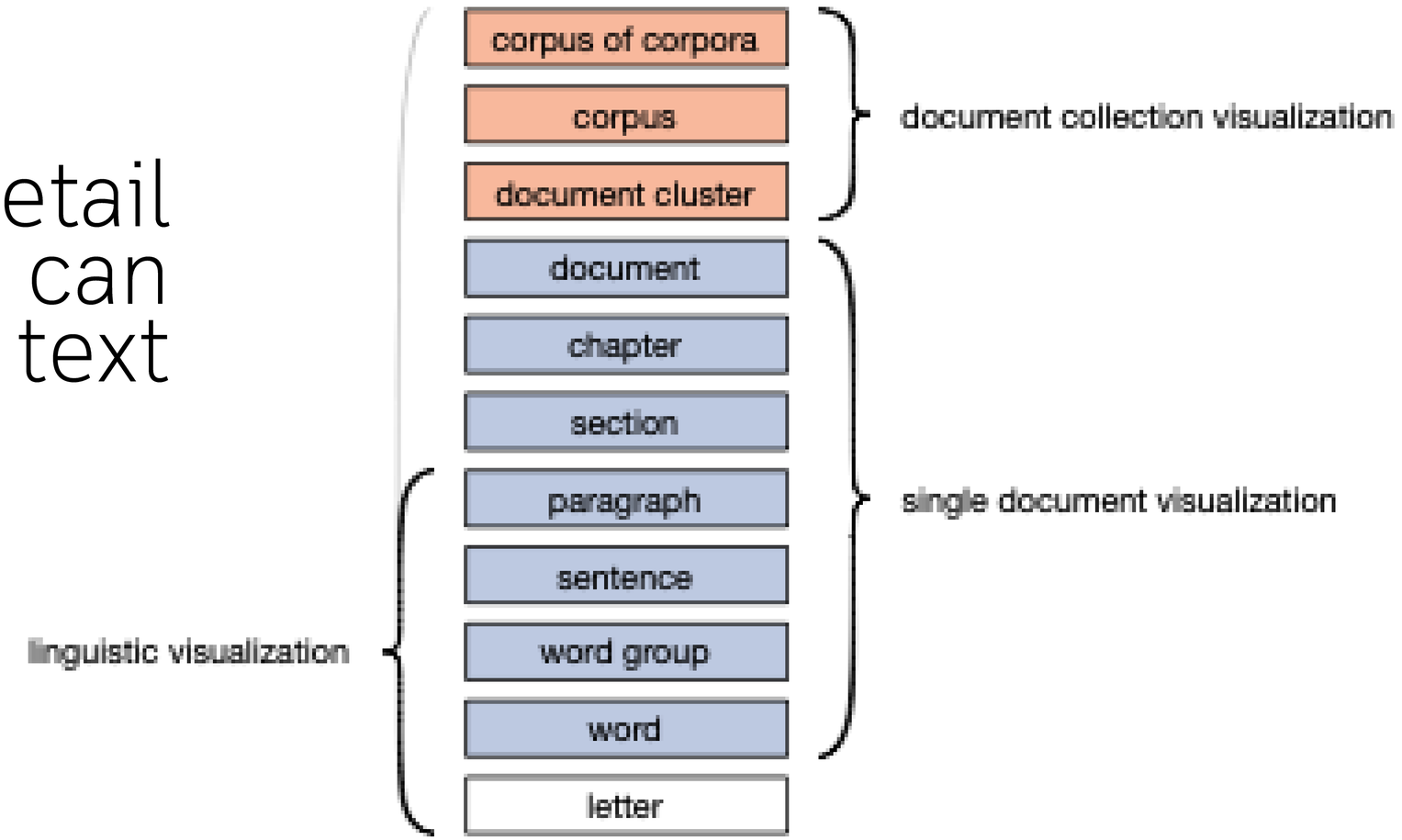
Text

Text is semi-structured

- Grammatical rules
- linear perception
- Words → Sentences → Paragraphs → Documents

Text Unit Hierarchy

Level of detail that we can analyze in text

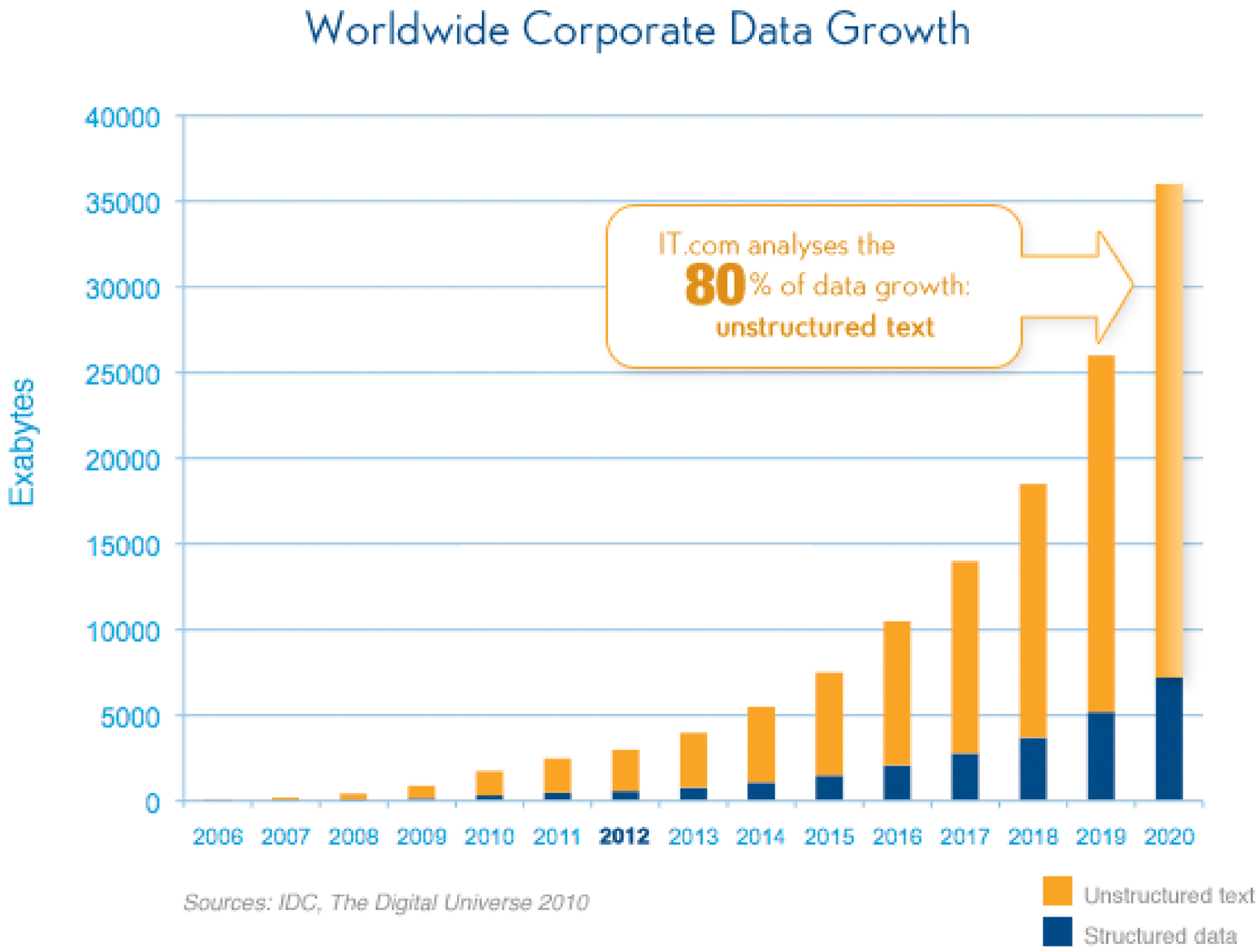


Language
Main medium for communication

- Extremely expressive for **communicating complex ideas**
- **More expressive** than visualization
- different across population groups (countries, accents, religions,…)

Why visualize text?

Why Visualize Text?



More abundant than structured data

Typography

Style, arrangement, or appearance of printed letters on a page

Visual medium for language

Typography

Sans Serif, **Serif**, **Bold**, *Italic*

point size (10pt, 12pt, 24pt, 36pt..)
line length (alignment: left, right, justified)

Leading: vertical line spacing

Tracking: spacing between groups of letters

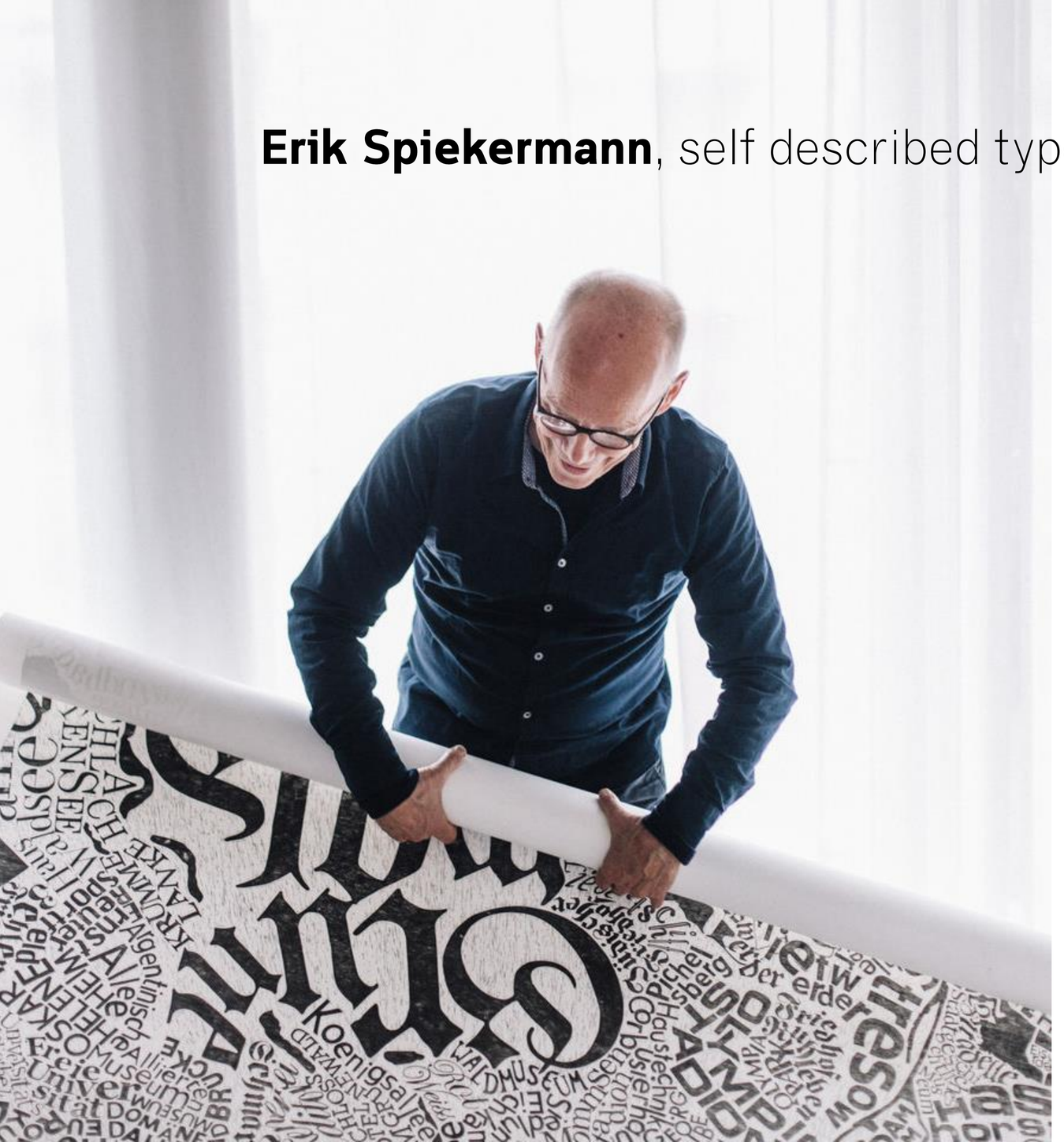
Kerning: space between actual letters

combining letters to a glyph ligatures

ß

fi → fi
fl → fl

Erik Spiekermann, self described typomaniac

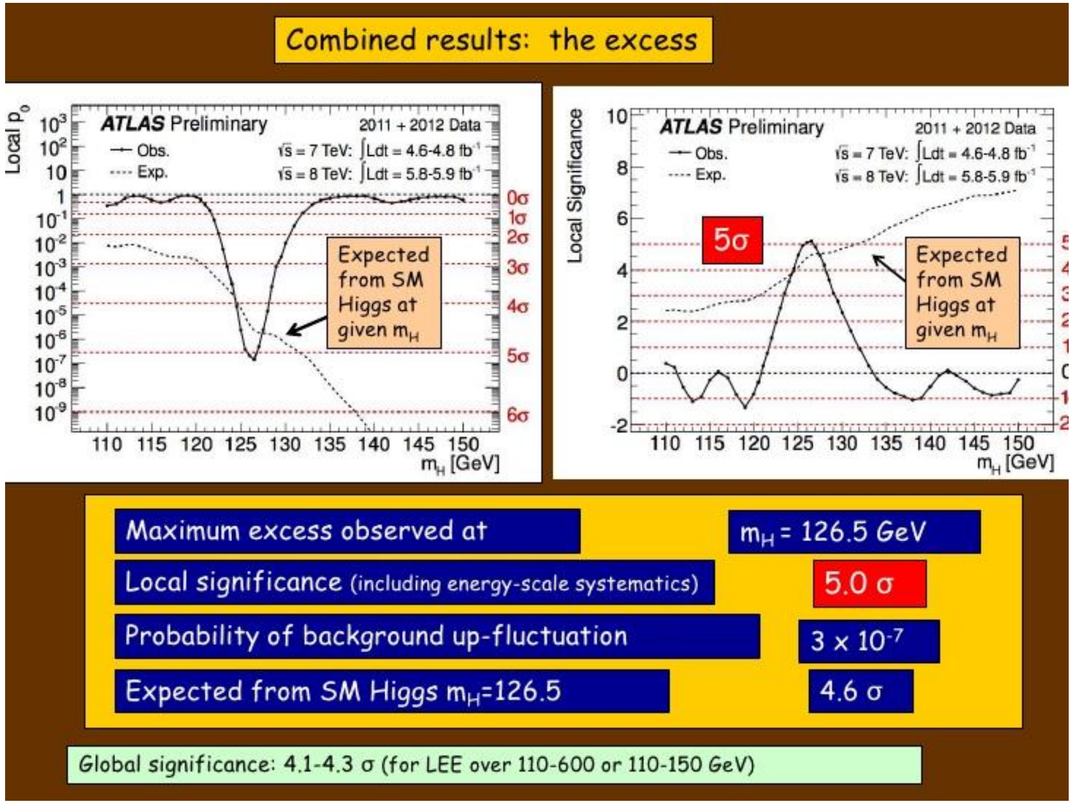


Creating a font requires profound design knowledge

We [designers] are interpreters, not merely translators, between sender and receiver. **What we say and how we say it makes a difference.** If we want to speak to people, we need to know their language. In order to design for understanding, we need to understand design.

Design choices influence your message

Comic Sans/Higgs Boson catastrophe of 2012



Taking the god particle seriously

One of the most important scientific discoveries in the last 100 years

Presented their work in Comic sans.

Does the medium fit the message?

http://www.comicsanscriminal.com/

Other recorded offences of bad font choices

Follow doespap

Does Papyrus Belong Here?

Hi, I'm Emily. Let's talk about Papyrus.

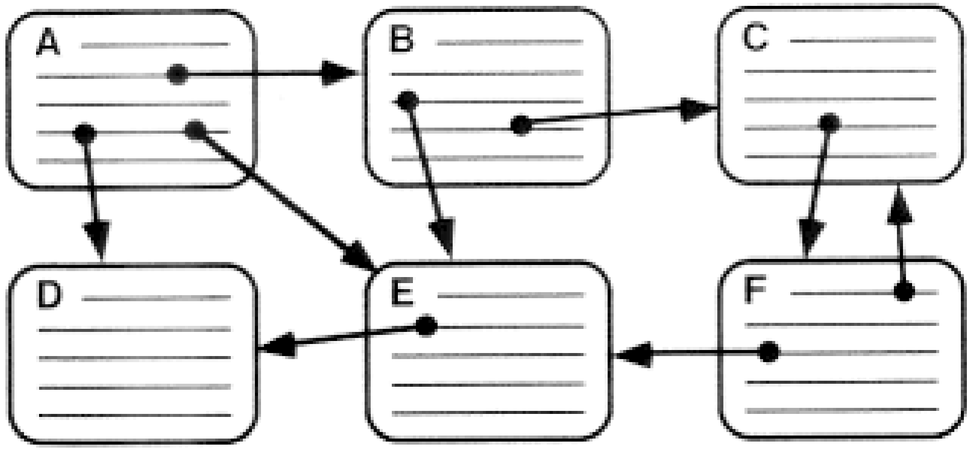
No.

Nov 4th, 2011
1 note

doespapyrusbelonghere posted this

We use text visualization on a daily basis…

Visualization for “Raw” Text

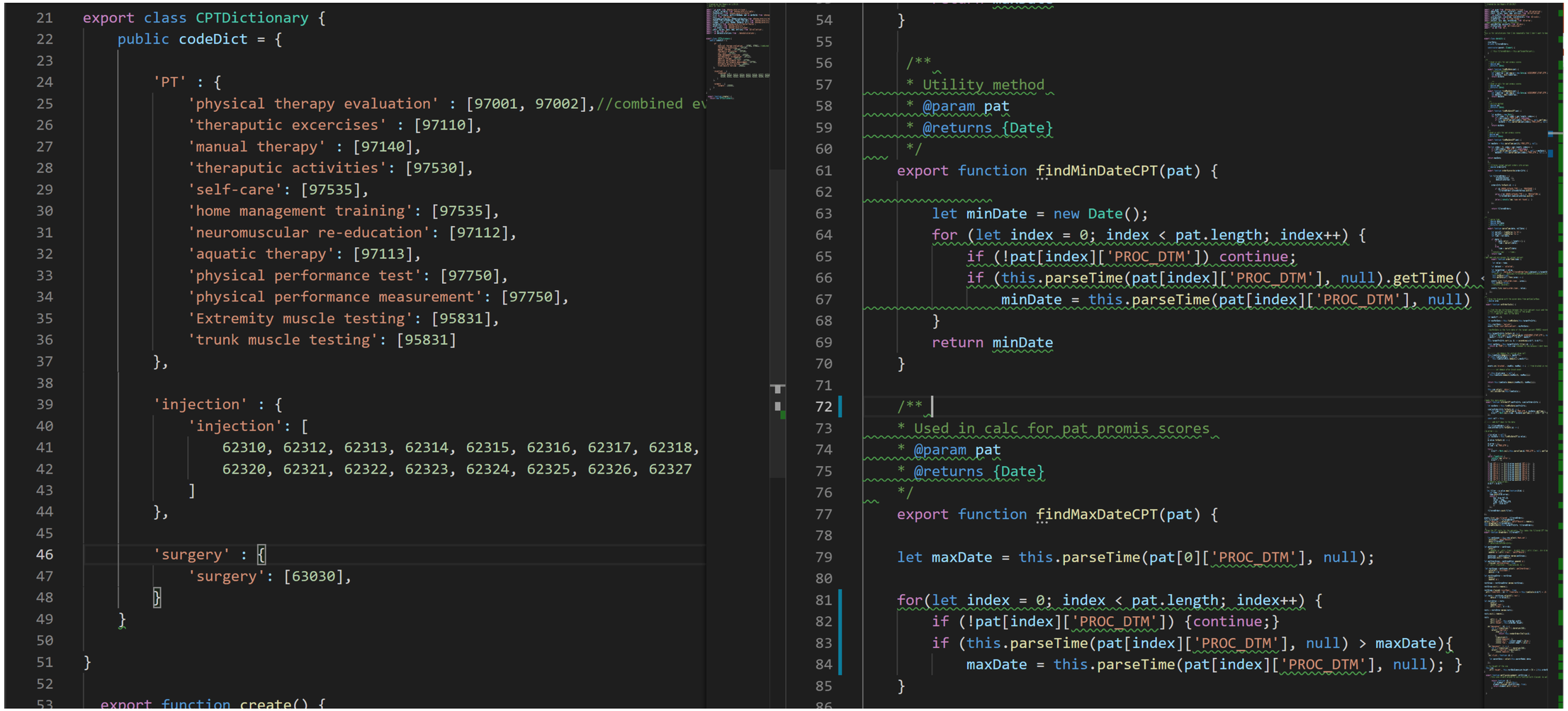


Screenshot

Enriched text – hypertext linking (graph navigation)

Visualization for “Raw” Text

Highlighting semantics

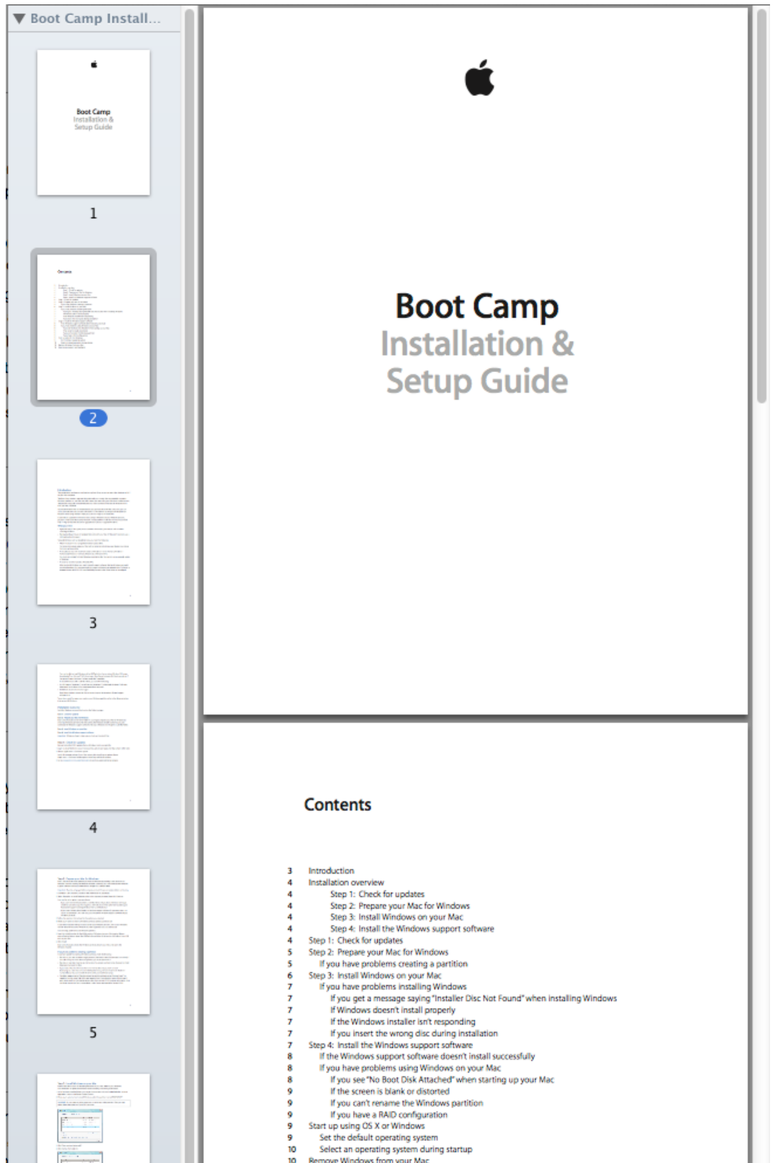


```
export class CPTDictionary {
    public codeDict = {

        'PT' : {
            'physical therapy evaluation' : [97001, 97002],//combined ev
            'theraputic excercises' : [97110],
            'manual therapy' : [97140],
            'theraputic activities': [97530],
            'self-care': [97535],
            'home management training': [97535],
            'neuromuscular re-education': [97112],
            'aquatic therapy': [97113],
            'physical performance test': [97750],
            'physical performance measurement': [97750],
            'Extremity muscle testing': [95831],
            'trunk muscle testing': [95831]
        },

        'injection' : {
            'injection': [
                62310, 62312, 62313, 62314, 62315, 62316, 62317, 62318,
                62320, 62321, 62322, 62323, 62324, 62325, 62326, 62327
            ]
        },

        'surgery' : {
            'surgery': [63030],
        }
    }

}

  export function create() {
```

```
    }

    /**
     * Utility method
     * @param pat
     * @returns {Date}
     */
    export function findMinDateCPT(pat) {

        let minDate = new Date();
        for (let index = 0; index < pat.length; index++) {
            if (!pat[index]['PROC_DTM']) continue;
            if (this.parseTime(pat[index]['PROC_DTM'], null).getTime() <
                minDate = this.parseTime(pat[index]['PROC_DTM'], null)
        }
        return minDate
    }

    /**
     * Used in calc for pat promis scores
     * @param pat
     * @returns {Date}
     */
    export function findMaxDateCPT(pat) {

    let maxDate = this.parseTime(pat[0]['PROC_DTM'], null);

    for(let index = 0; index < pat.length; index++) {
        if (!pat[index]['PROC_DTM']) {continue;}
        if (this.parseTime(pat[index]['PROC_DTM'], null) > maxDate){
            maxDate = this.parseTime(pat[index]['PROC_DTM'], null); }
    }
```

Visualization for “Raw” Text

Overview and detail views



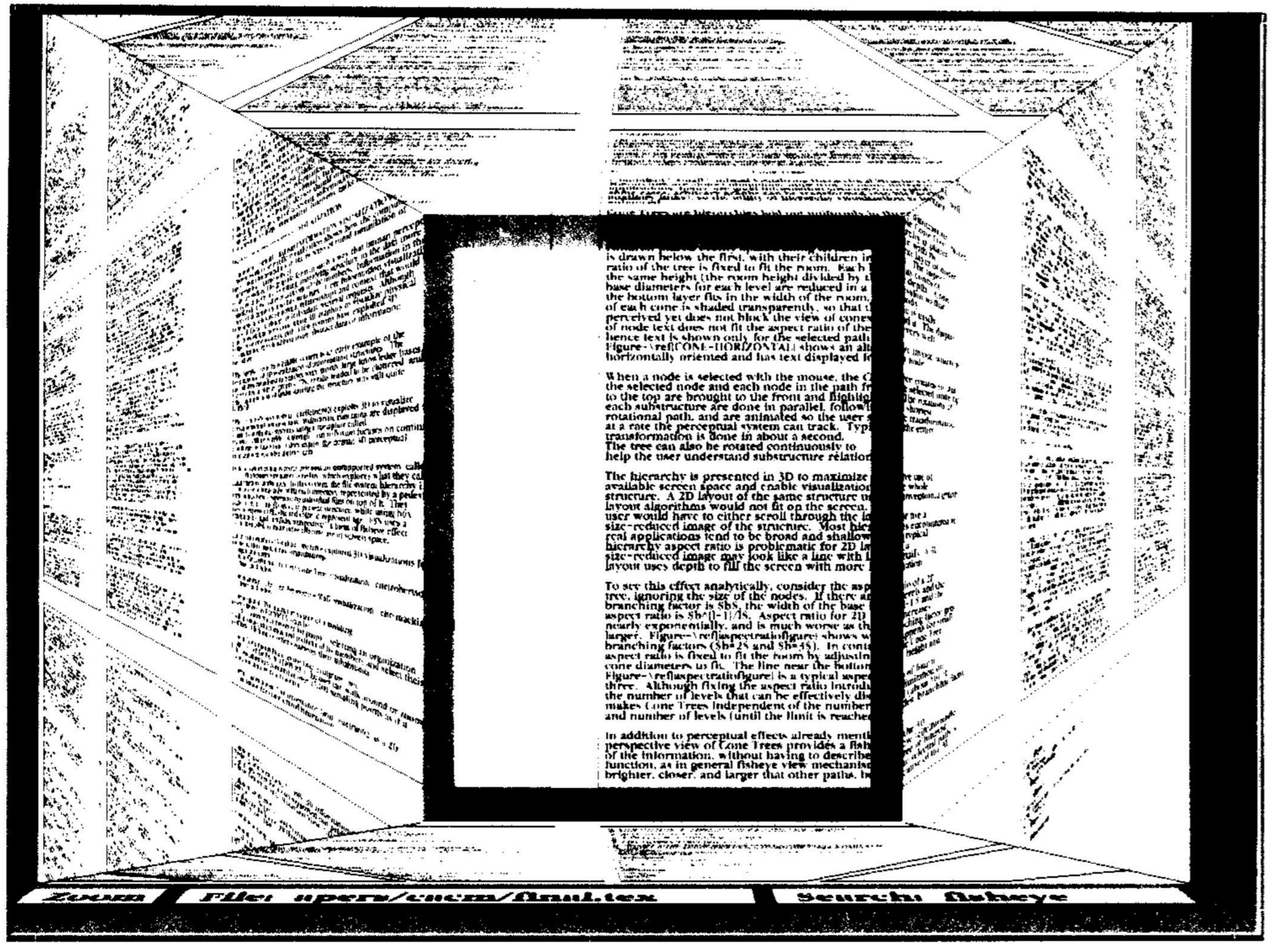
Visualization for “Raw” Text

Document Lens

‘93

Focus and Context

Zoomed area of interest
Without loosing context
of the whole document



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.

Visualization for “Raw” Text

Visualizing Search Results

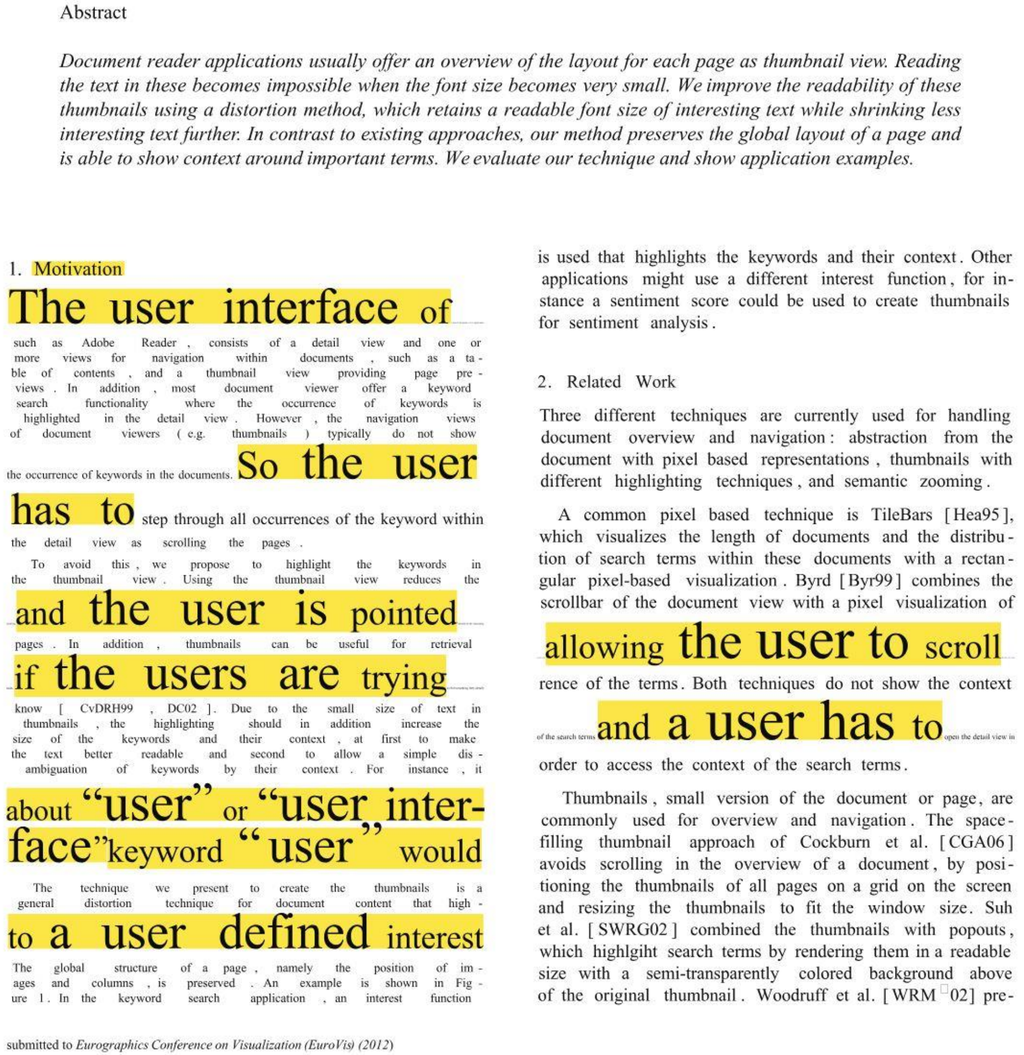
To find keywords in an overview

Thumbnail view shows location of keywords or phrases

Document Thumbnails with Variable Text Scaling

2012

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



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 documents. 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 [CvDRH99 , 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 images and columns , is preserved . An example is shown in Figure 1 . In the keyword search application , an interest function 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 .

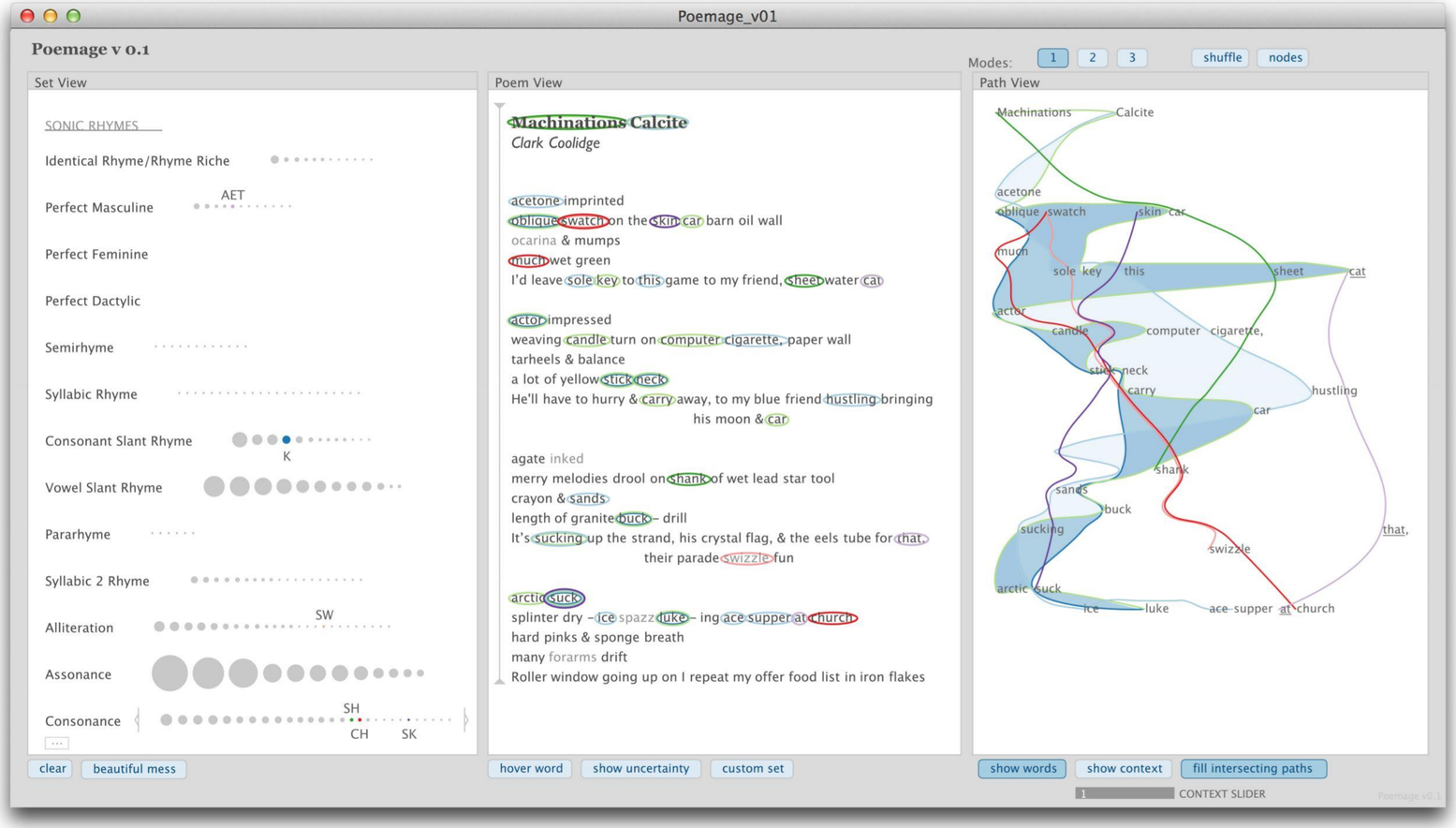
A common pixel based technique is TileBars [Hea95], which visualizes the length of documents and the distribution of search terms within these documents with a rectangular pixel-based visualization . Byrd [Byr99] combines the scrollbar of the document view with a pixel visualization of allowing the user to scroll rence of the terms . Both techniques do not show the context of the search terms and a user has to open 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 of the original thumbnail . Woodruff et al. [WRM+02] pre-

submitted to *Eurographics Conference on Visualization (EuroVis) (2012)*

Visualization for “Raw” Text

Poemage: Visualizing the Sonic Topology of a Poem



complex structures from the interaction of **sonic patterns** in the poem

‘15

Working with Text

Call me Ishmael. Some years ago -- never mind how long precisely -- having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen, and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off -- then, I account it high time to get to sea as soon as I can. This is my substitute for pistol and ball. With a philosophical flourish Cato throws himself

4 x 't'
3 x 'u'
2 x 'r'
2 x 'e'
...

structured data

Generally want to convert raw text to features

Steps of Processing to derive Text Features

The cleanup:

- Sentence splitting
- change to lower case
- Removing punctuation
- Stop word removal (most frequent words in a language)
- Stemming - demo porter stemmer

Steps of Processing to derive Text Features

Understanding the structure:

- **N-grams, bag of words**
- Concordance: Keyword in context
- Co-occurrence : Phrase Net
- POS tagging (part of speech)
- Sentiment analysis for twitter
- NER (name entity recognition)
- deep parsing - try to “understand” text.

Measuring Structured Text Features

- Simple counts (bag of words) used for similarity measures
- One of the most basic measures for text analysis
- Divide text into n-grams
- If texts share similar words, may be similar in content

	princess	dragon	castle
doc1	1	1	1
doc2	0	0	1

Wordle

- Frequency-based
  - words that occur often are large
- Can vary font type, size, color, etc.
- One of the most popular text level vis tools

http://www.wordle.net

[Viegas 2009]


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


ideas
great
idealings
idea
imagination
creative
friends
good
like
time
Napoleon
Johnson
Steven
Hill
mind
environment
flash
bouncing
person
new
come
word
one
little
just
appears
born
play
individuals
Champions
example
found


Union
War
right
shall
God
nation
years
Woe
malice
towards
none
offenses
men
hope
judge
needs
public
gives
come
extend
came
wills
conflicts
laves
drawn
ago
parties
care
cause
pray

Wordle vs Tag Cloud

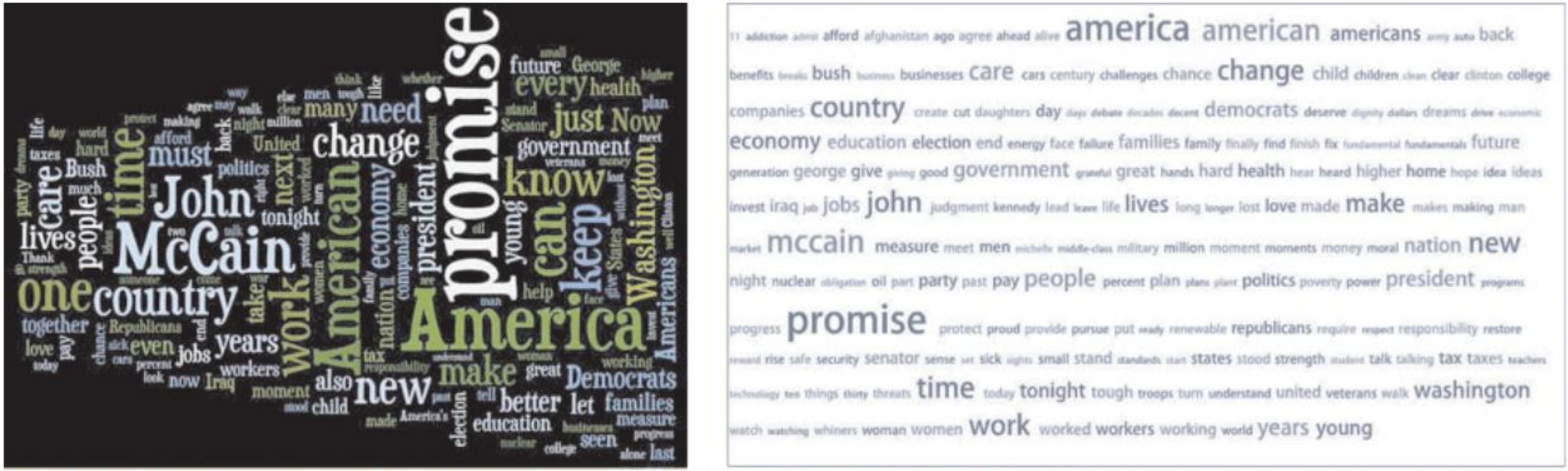


Fig 2: Wordle vs. Tag Cloud of Barack Obama's speech at the Democratic Convention in 2008.

Can quickly see dominant words

Not for precise comparisons and similarity measures

Weighting

Frequency count

log frequency : **log(1 + tftd)**

Normalized for proportion

Text Frequency by inverse document frequency - **TF IDF**

Limitations

Frequency may not be meaningful

Does not show the structure

Does not explain the context/grammar/POS

Steps of Processing to derive Text Features

Understanding the structure:

- N-grams, bag of words
- **Concordance: Keyword in context**
- Co-occurrence : Phrase Net
- POS tagging (part of speech)
- Sentiment analysis for twitter
- NER (name entity recognition)
- deep parsing - try to “understand” text.

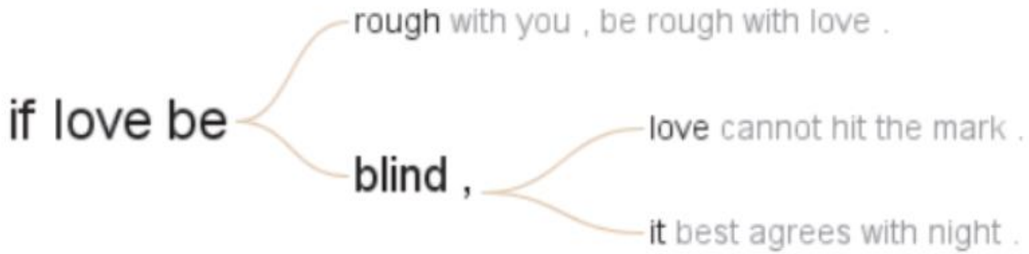
Concordance: Keyword in context

Word Tree

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

if love be blind , love cannot hit the mark .

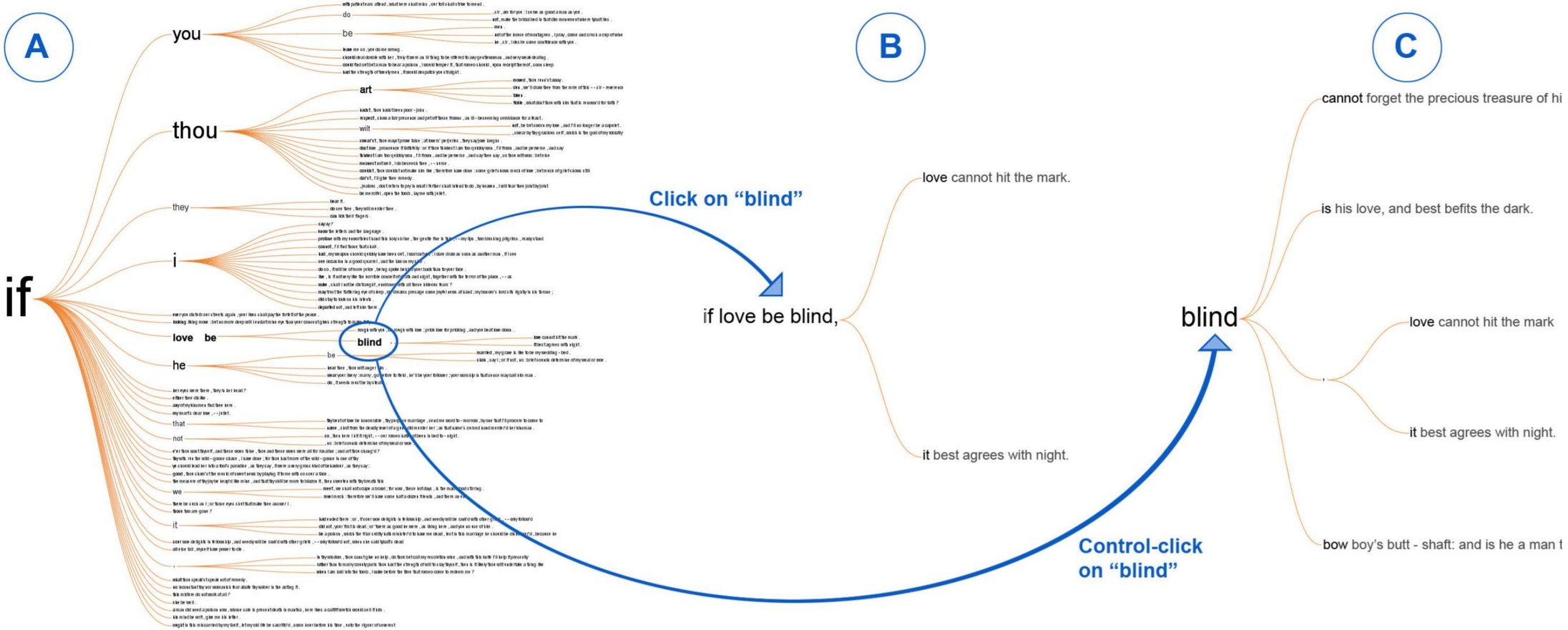
if love be blind , it best agrees with night .



What is the local context of the word?

[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

Shows local repeated structure

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

Steps of Processing to derive Text Features

Understanding the structure:

- N-grams, bag of words
- Concordance: Keyword in context
- **Co-occurrence : Phrase Net**
- POS tagging (part of speech)
- Sentiment analysis for twitter
- NER (name entity recognition)
- deep parsing - try to “understand” text.

Phrase Net

We are looking for specific linking patterns

1 *You create the word sequence filter:*

WORD1 [and] WORD2

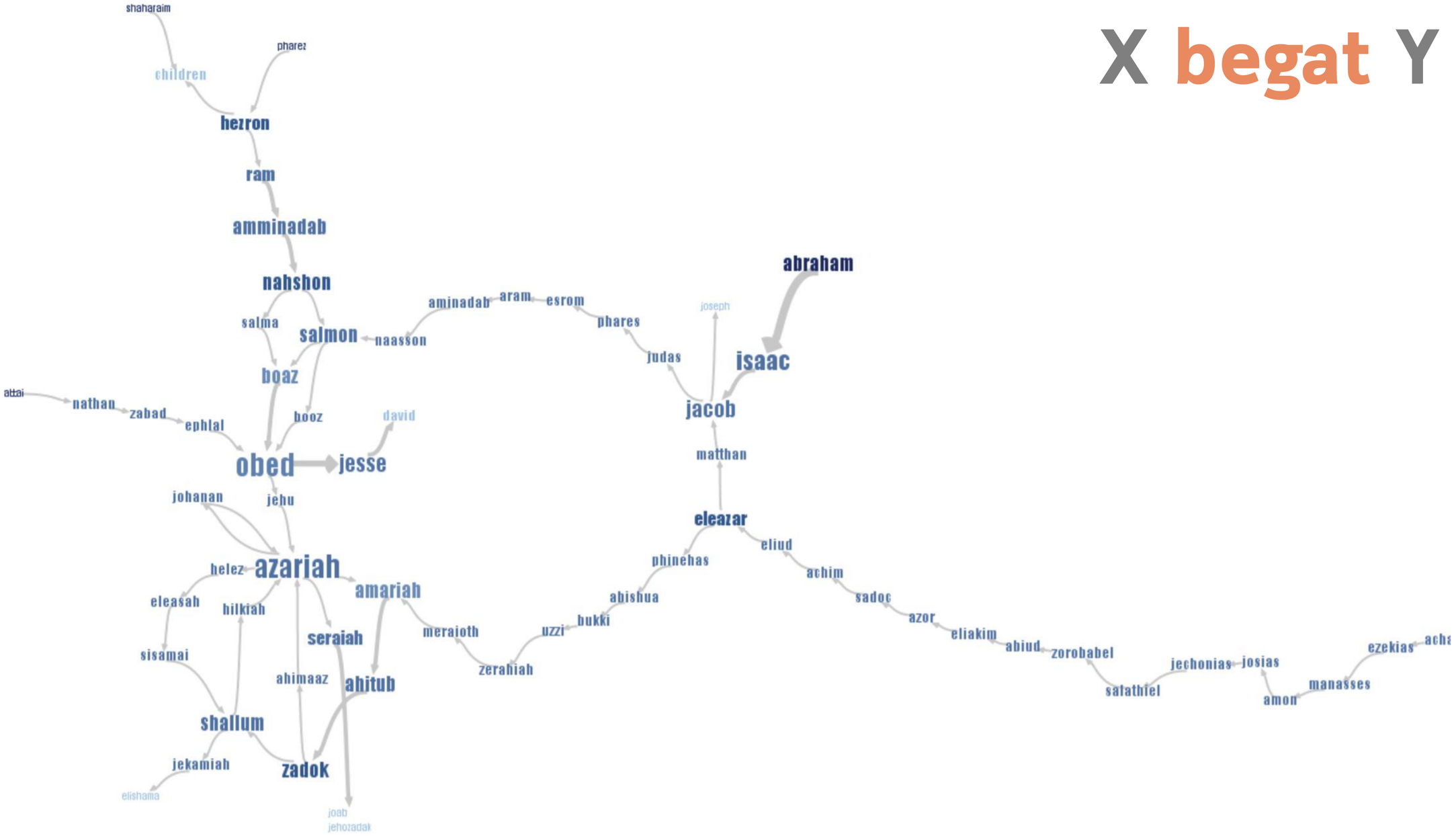
2 *Many Eyes finds this word relationship in Jane Austen's text:*

Her manners were pronounced to be very bad indeed,
a mixture of pride and impertinence; she had no
conversation, no stile, no taste, no beauty.

3 *Many Eyes creates the word graph:*

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)

X begat Y


shaharaim
pharez
children
hezron
ram
amminadab
nahshon
salma
salmon
naasson
aminadab
aram
esrom
phares
judas
joseph
abraham
isaac
jacob
boaz
attai
nathan
zabad
ephlal
obed
booz
david
jesse
matthan
eleazar
johanan
jehu
azariah
helez
eliud
phinehas
achim
sadoc
azor
eliakim
abiud
zorobabel
salathiel
jechonias
josias
amon
manasses
ezekias
acha
amariah
abishua
bukki
uzzi
meraioth
zerahiah
eleasah
hilkiah
seraiah
sisamai
ahimaaz
ahitub
shallum
jekamiah
elishama
zadok
joab
jehozadak

Steps of Processing to derive Text Features

Understanding the structure:

- N-grams, bag of words
- Concordance: Keyword in context
- Co-occurrence : Phrase Net
- **POS tagging (part of speech)**
- Sentiment analysis
- NER (name entity recognition)
- deep parsing - try to “understand” text.

Part of Speech Tagging

Labeling words in text as a specific part of speech

How is a word **used in the scope** of a phrase?

Distinguish meaning of the word.

“I am going to the fair.” VS **“That is fair”**

Explain the **syntactic role** of a word

Infer semantic information from this due to our knowledge of syntactic role

Part of Speech Tagging Demo

Steps of Processing to derive Text Features

Understanding the structure:

- N-grams, bag of words
- Concordance: Keyword in context
- Co-occurrence : Phrase Net
- POS tagging (part of speech)
- **Sentiment analysis**
- NER (name entity recognition)
- deep parsing - try to “understand” text.

Sentiment Analysis

Subjective information (opinions and attitudes) from text.

Social media is a huge data resource for this.

Basic task is identifying **polarity**:

positive, neutral, negative

Tweet sentiment visualization

Steps of Processing to derive Text Features

Understanding the structure:

- N-grams, bag of words
- Concordance: Keyword in context
- Co-occurrence : Phrase Net
- POS tagging (part of speech) – demo
- Sentiment analysis
- **NER (name entity recognition)**
- deep parsing - try to “understand” text.

Name Entity Recognition

Reveals major people, organizations, and places.

Used for **tagging and categorizing** of documents and articles

Demo for NER

Steps of Processing to derive Text Features

Understanding the structure:

- N-grams, bag of words
- Concordance: Keyword in context
- Co-occurrence : Phrase Net
- POS tagging (part of speech)
- Sentiment analysis
- NER (name entity recognition)
- **deep parsing - try to “understand” text.**

Text features are complicated

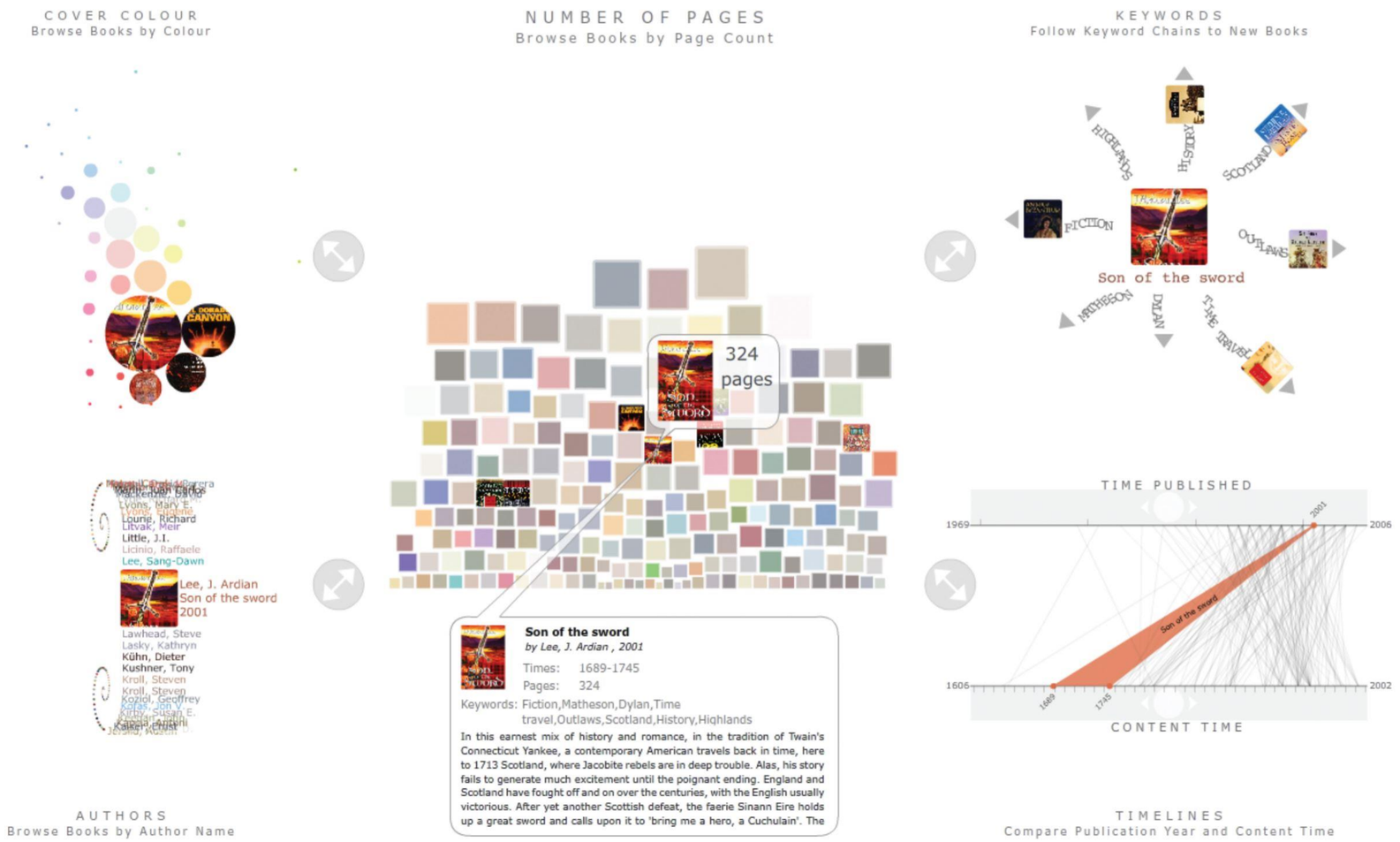
- *Toilet out of order. Please use floor below.*
- *One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know.*
- *Did you ever hear the story about the blind carpenter who picked up his hammer and saw?*

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

Visualizing Collections of Documents

- Identify **themes** across documents
- Identify **groupings** of documents
- Identify **differences** between collections
- Understand adjacent information about a **particular topic** in the collection

The Bohemian Bookshelf



Open ended book exploration in the library

Highlight adjacencies between books

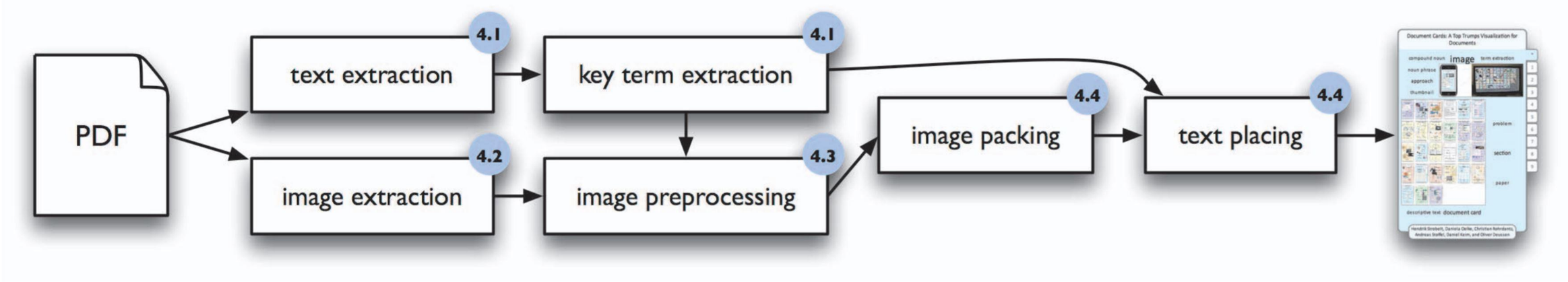
Flexible visual pathways

Alice Thudt, Uta Hinrichs and Sheelagh Carpendale.
The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization.
CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012

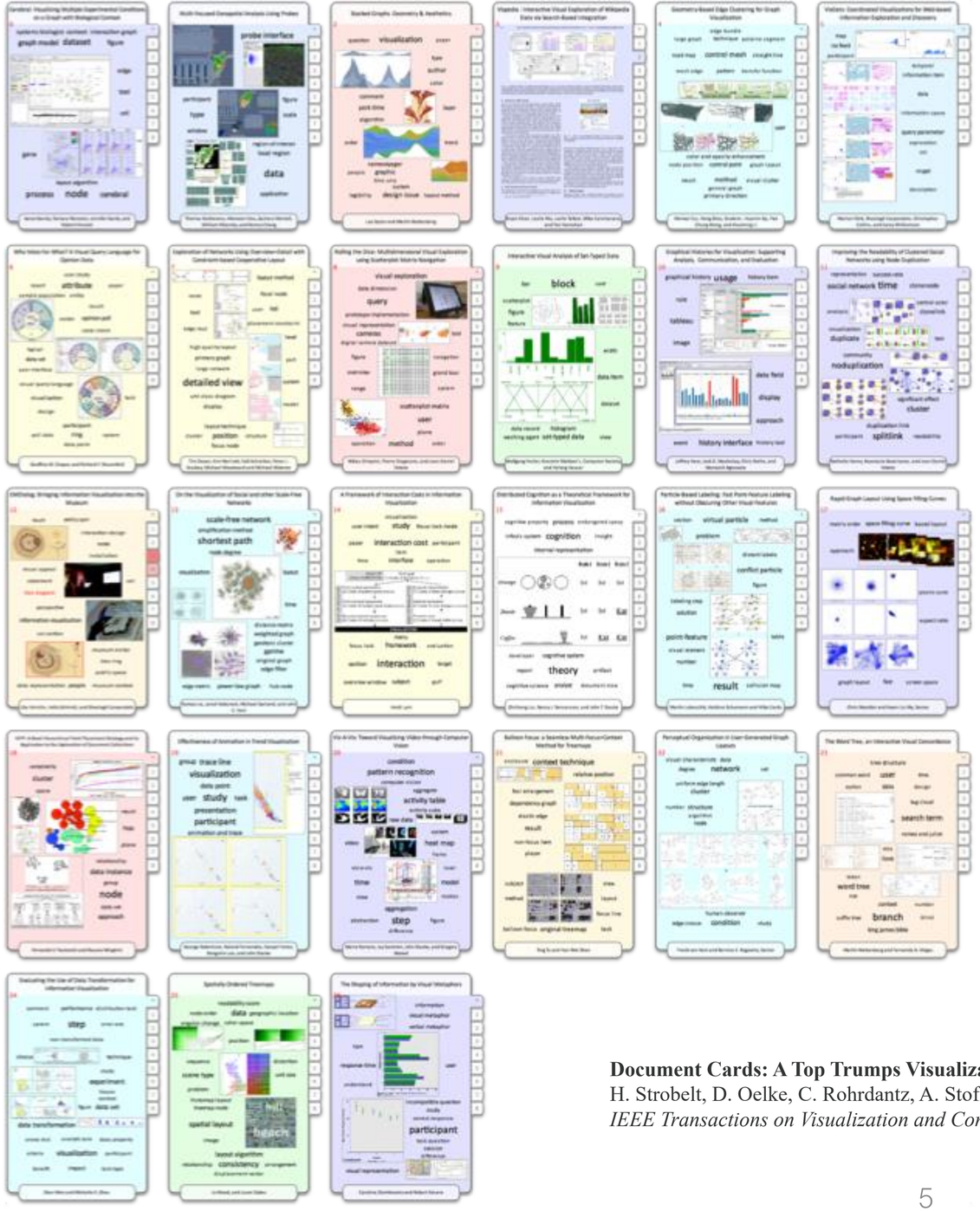
webpage with video

JigSaw- investigate document collections

DocumentCards



summarize using important terms and important figures

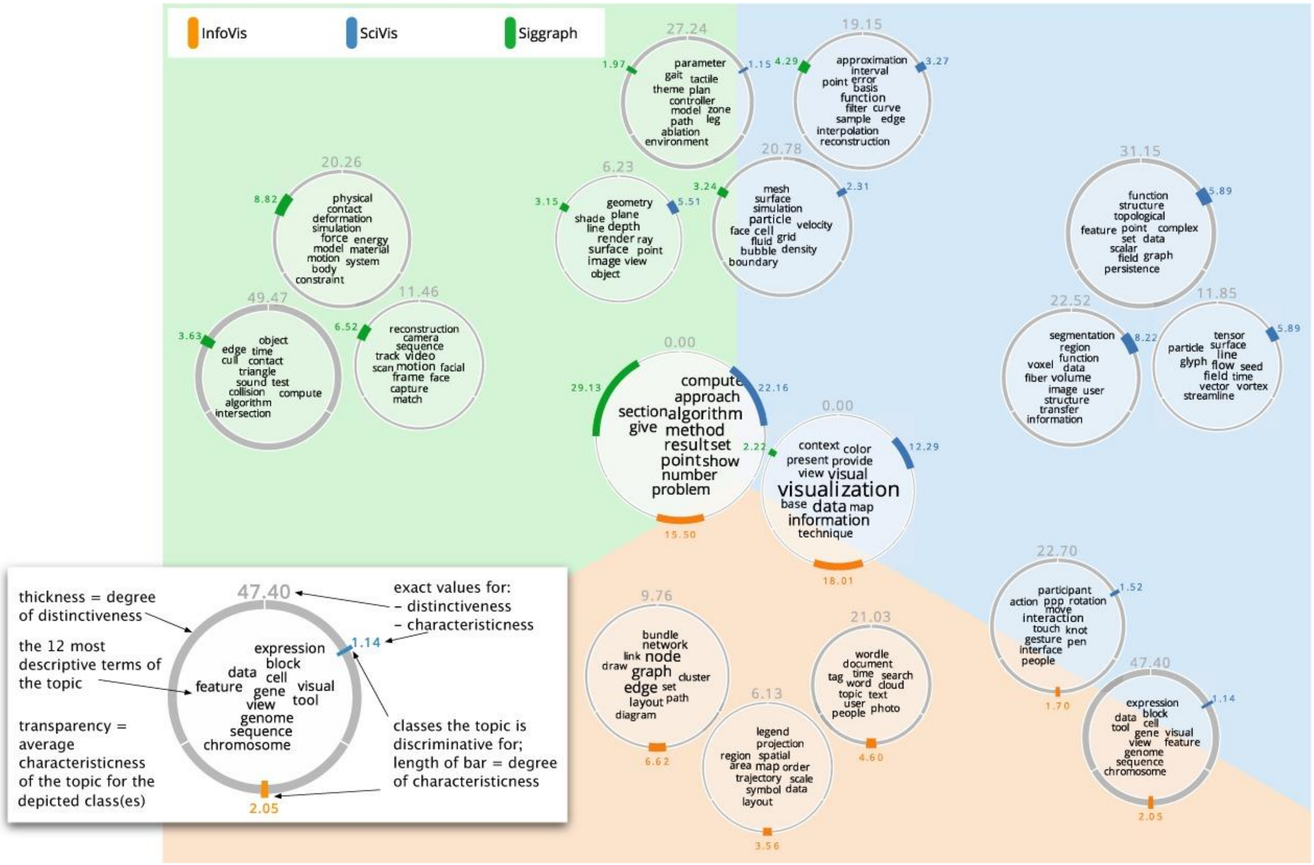


content as a mixture of figure and textual representatives

Document Cards: A Top Trumps Visualization for Documents
H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, O. Deussen, D. Keim
IEEE Transactions on Visualization and Computer Graphics (TVCG - InfoVis), 2009

Compare Corpora

Compare topics between text collections



Use probabilistic topic modeling to identify topics that discriminate one collection from others.

Comparison of papers between conferences.

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

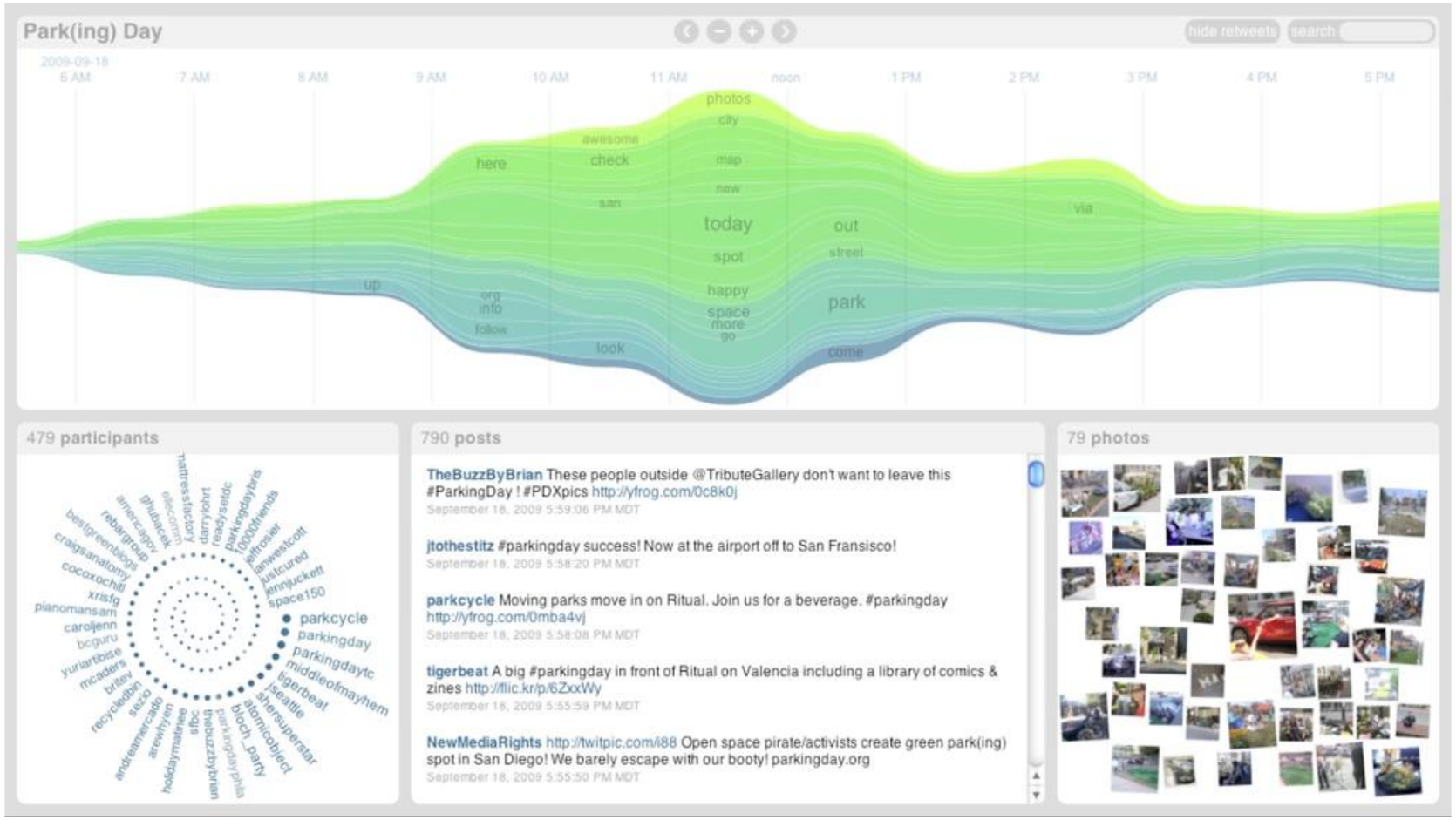
Visualizing and tracking changes

How did this document evolve?

Traces Project

Vis for Time-Evolving Document Collections

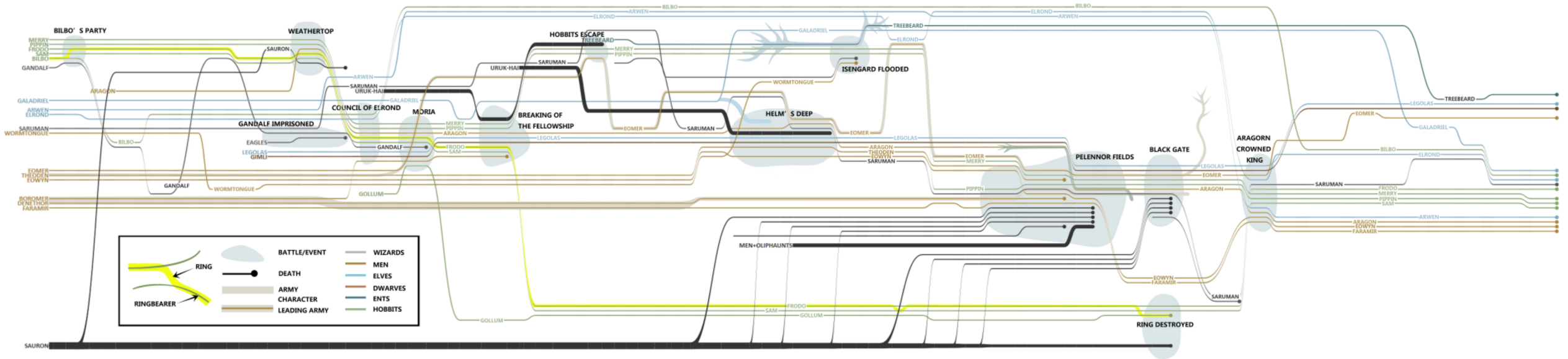
Visualizing tweets in an overview to show saliency of past and present happenings in the context of the evolving conversation



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

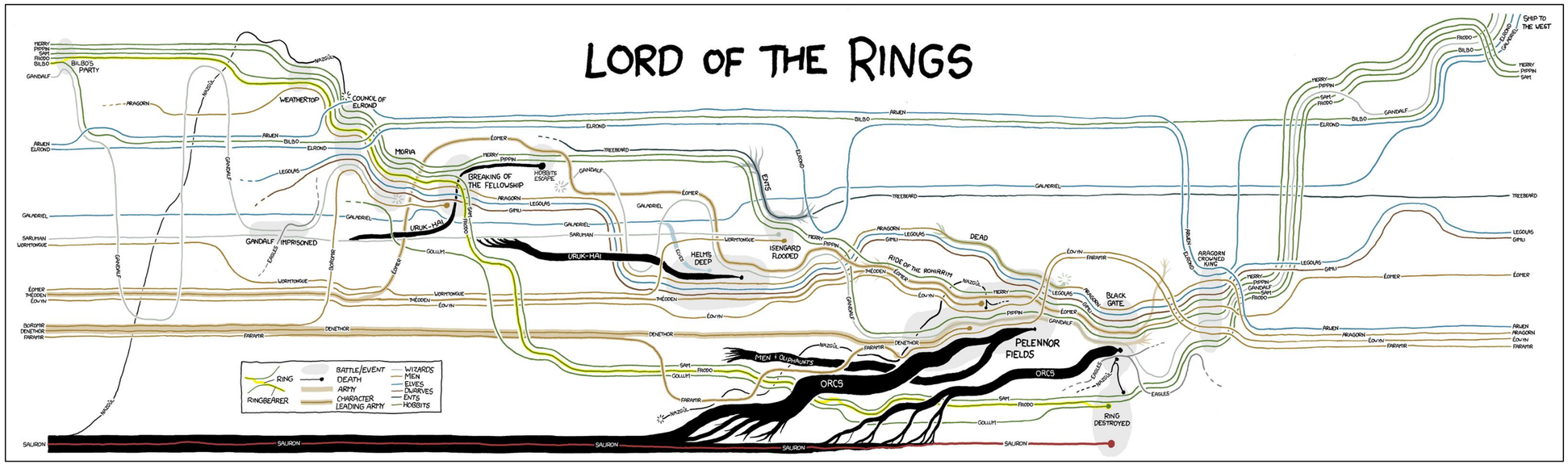
StoryFlow: Tracking the Evolution of Stories

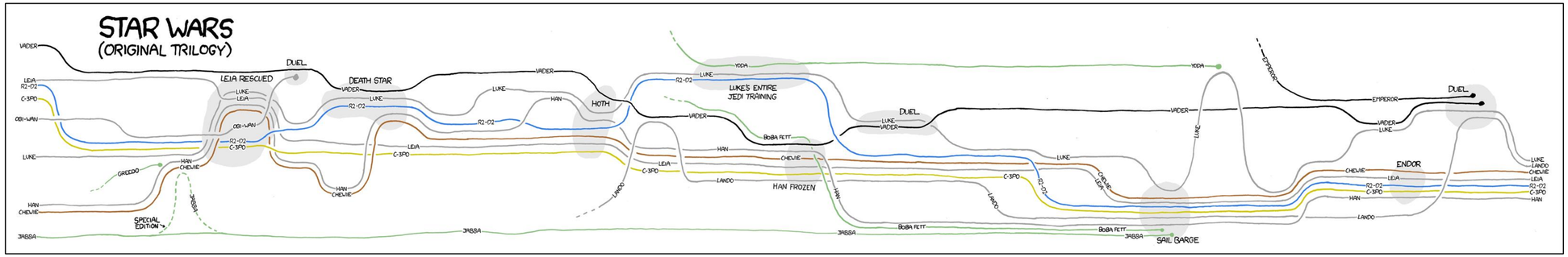
Visual interpretation of a story line.

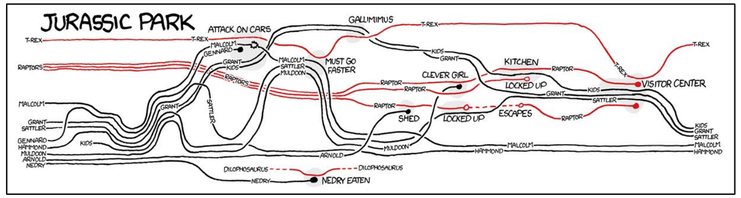


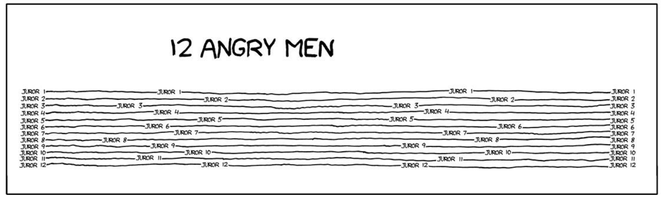
[Liu 2013]

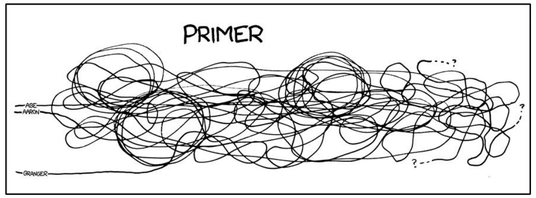
THESE CHARTS SHOW MOVIE CHARACTER INTERACTIONS.
THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE
LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.











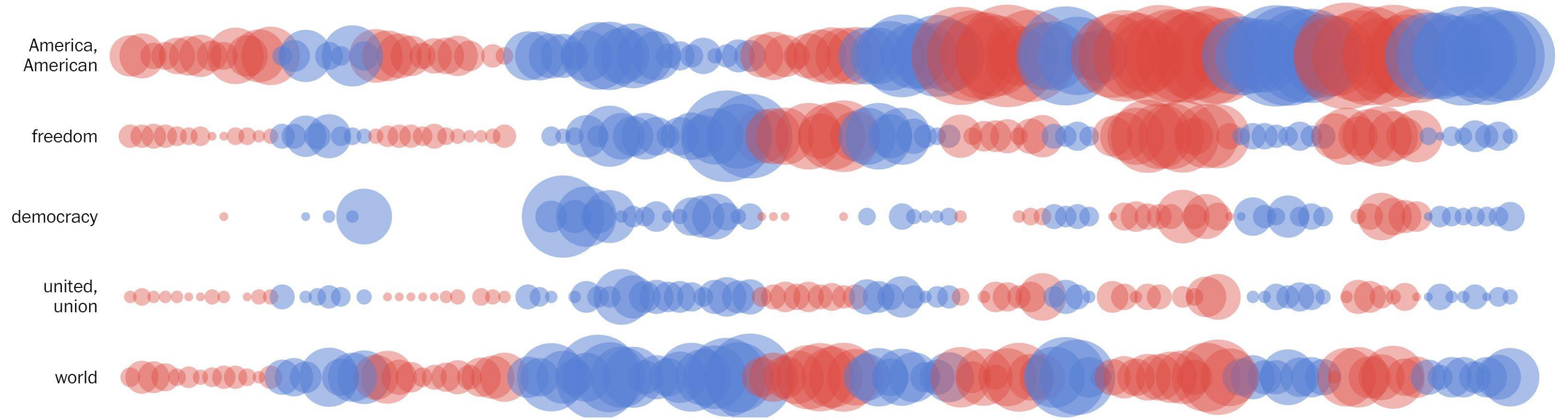
https://xkcd.com/657/

Design Critique: History through the President's words

Nationalism

Liberman shared a colleague's analysis that shows words like "citizens" and "duties" have faded from presidential inaugural addresses, while words like "freedom" and "America" have surged. While it doesn't explain the shift, he suspects that it's likely the trend applies to both types of addresses.

Fields suspected the oscillations in "America" (and variations) shadow the popularity of "God Bless America," which was written by Irving Berlin in 1918 and revised in 1938.



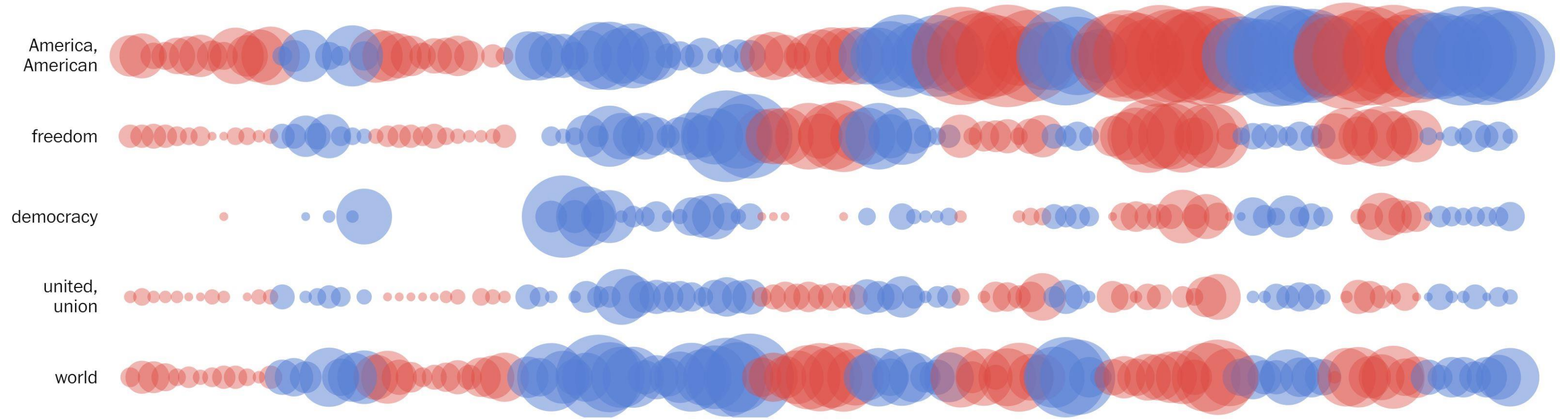
What information is encoded?
What marks and channels are they using?
Is it effective?
What information do you get in overview vs on hover?

Design Critique: History through the President's words

Nationalism

Liberman shared a colleague's analysis that shows words like "citizens" and "duties" have faded from presidential inaugural addresses, while words like "freedom" and "America" have surged. While it doesn't explain the shift, he suspects that it's likely the trend applies to both types of addresses.

Fields suspected the oscillations in "America" (and variations) shadow the popularity of "God Bless America," which was written by Irving Berlin in 1918 and revised in 1938.



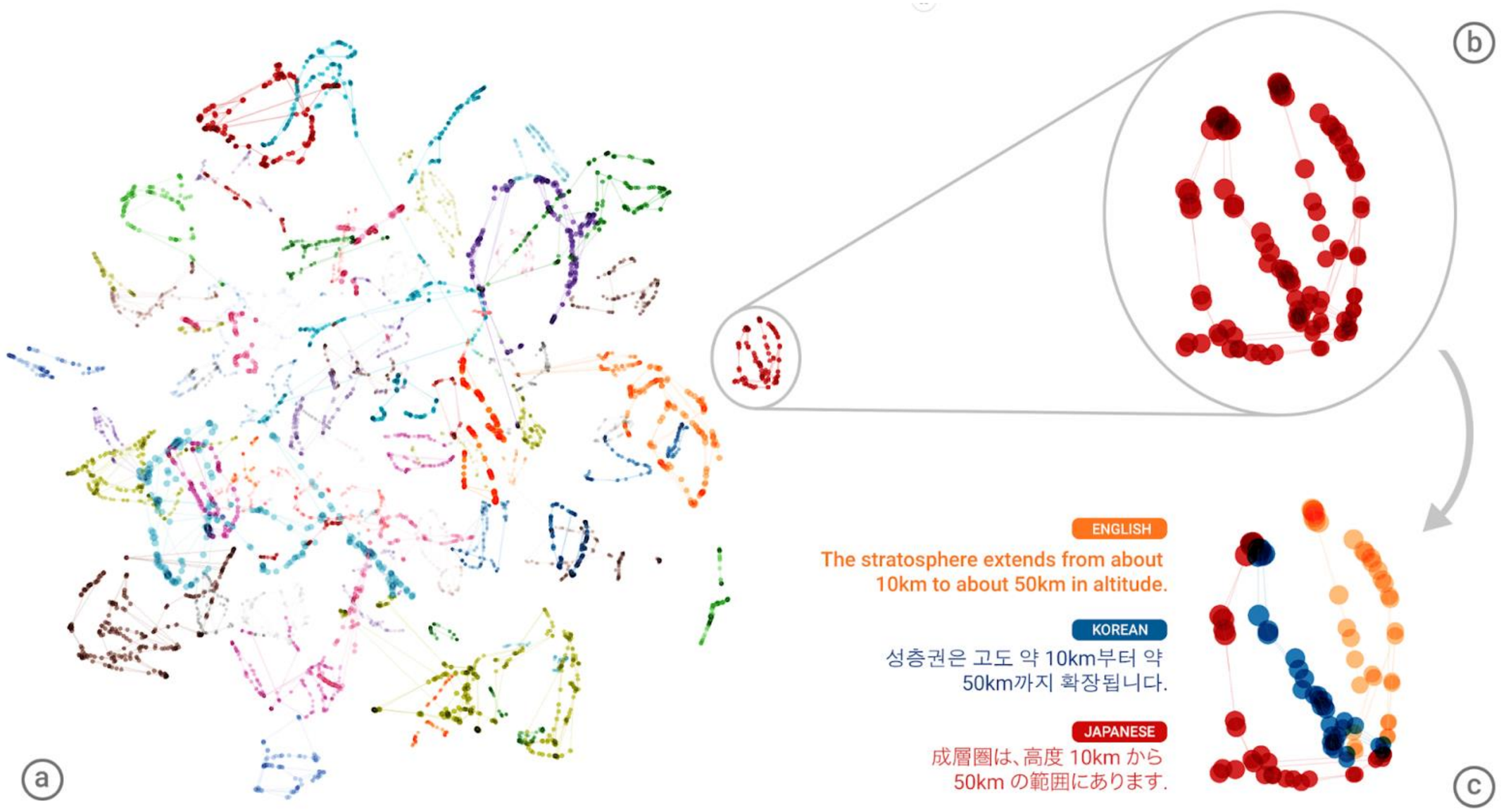
Create a redesign with your group:

Imagine you ran some sentiment analysis on the speech data

Create a redesign encoding polarity of the speech data for listed topics

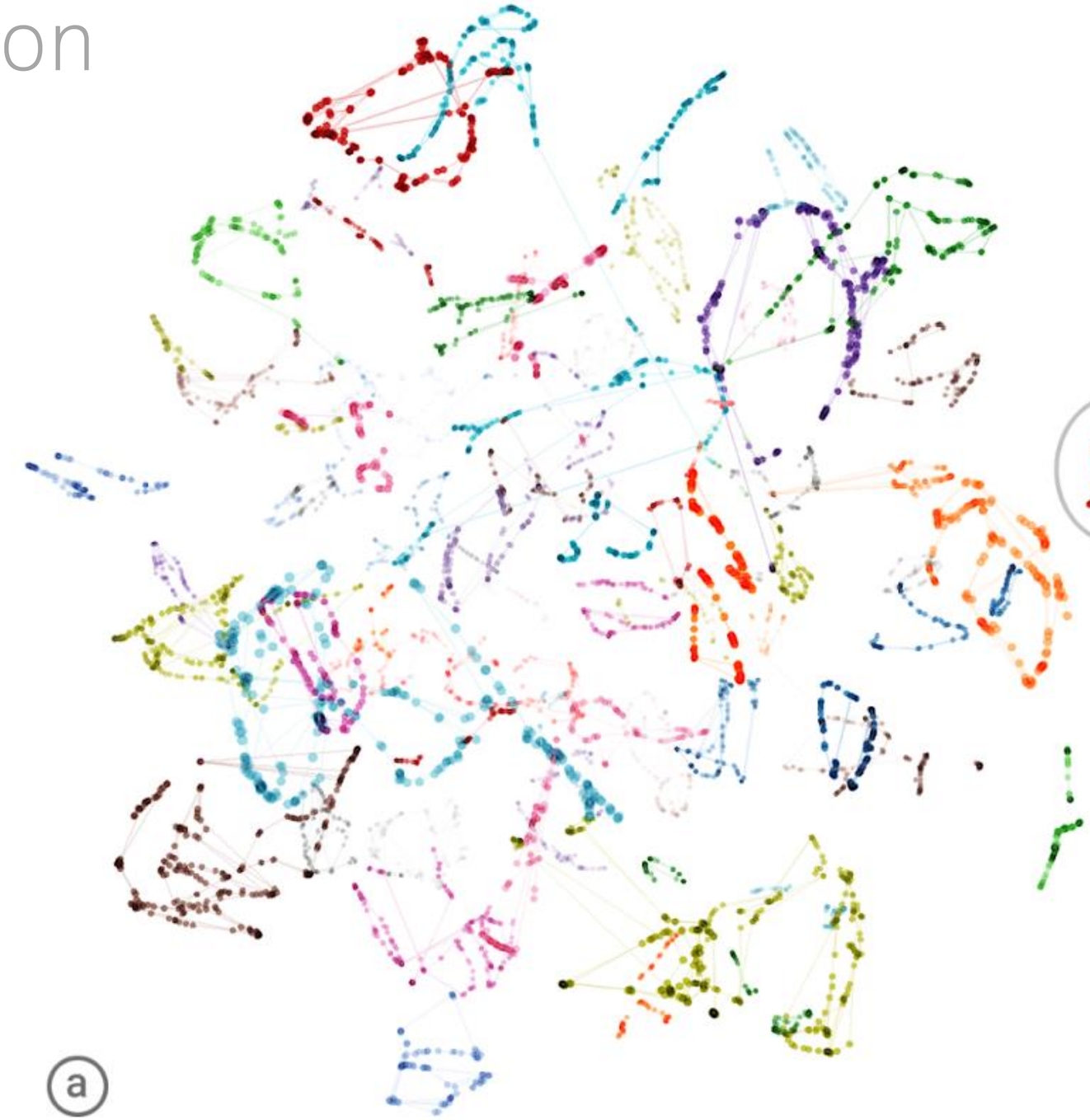
How could you encode sentiment in the visualization

Visualization for Translation

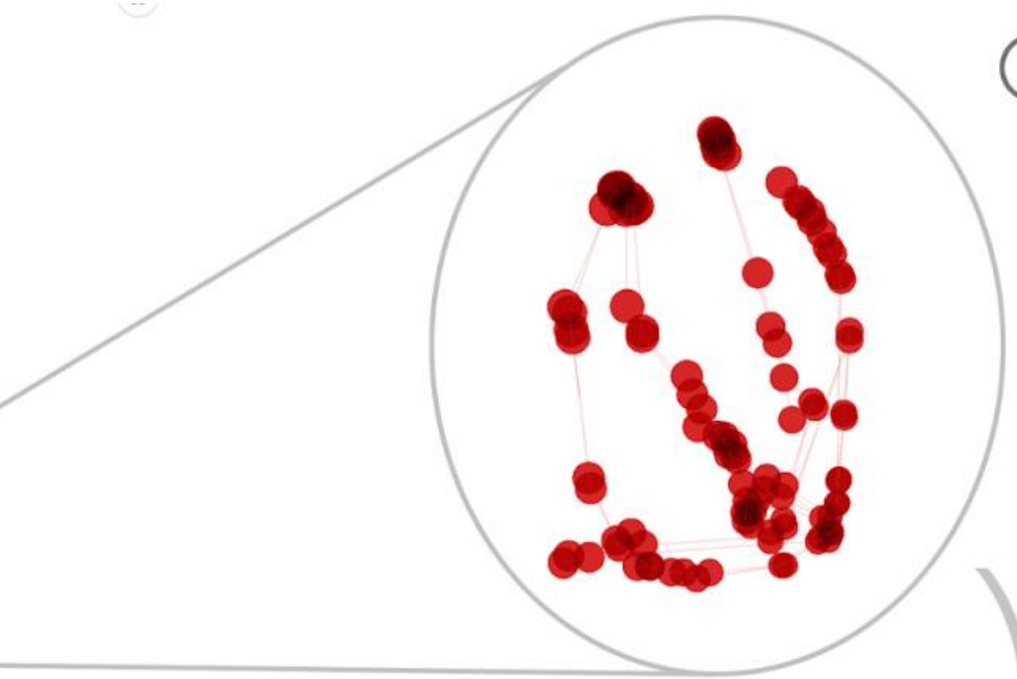


Visualization for Translation

Geometry of translations
Colored by the meaning
A sentence translated from
English → Korean
Japanese → English
share the same color.

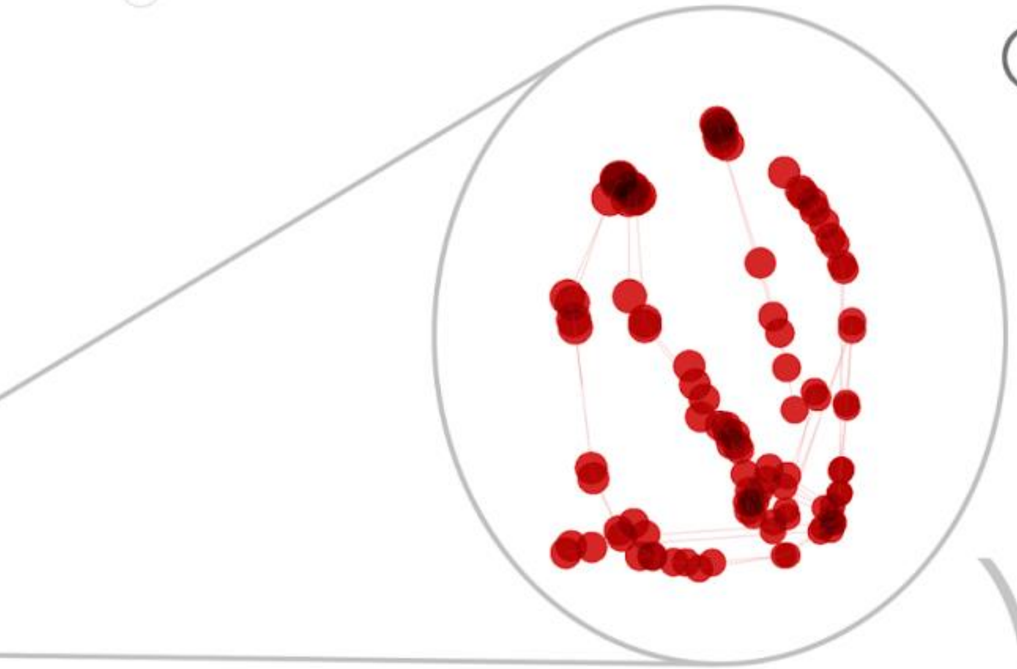


Visualization for Translation



Zoom into one of the groups translated sentence

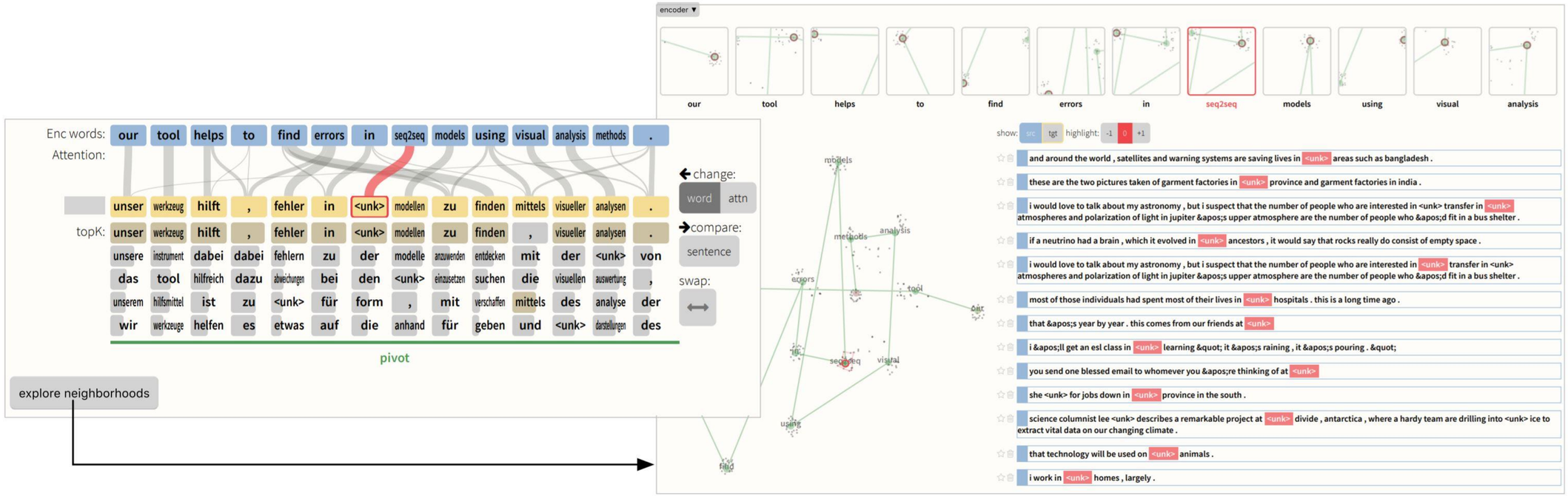
Visualization for Translation



Color is changed to source language,
Network must be encoding semantics rather than phrase to phrase translations
Existence of an interlingua?

Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models

http://seq2seq-vis.io/



http://textvis.lnu.se/