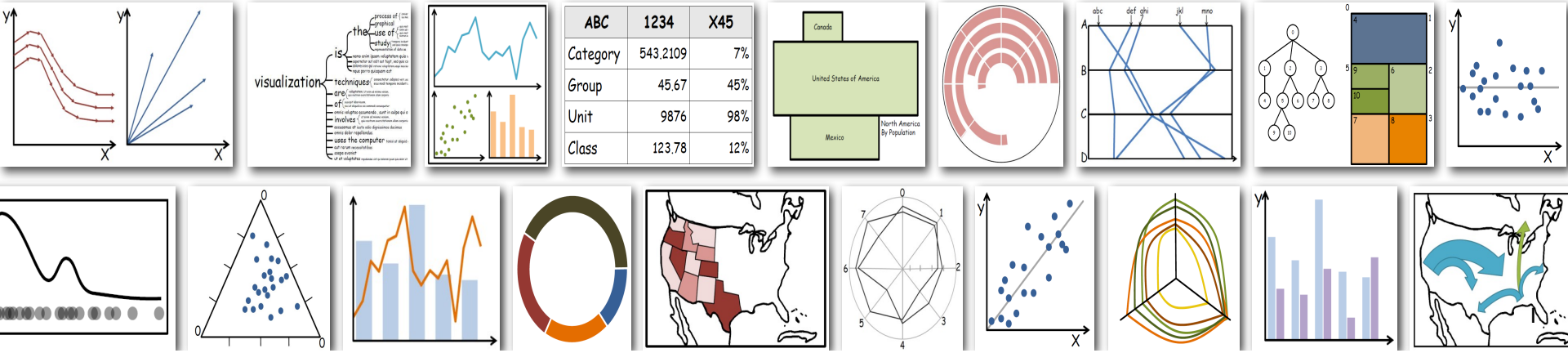


Lecture: Visualizing Text

DATA ANALYSIS & VISUALIZATION

FALL 2021

Dr. Muhammad Faisal Cheema

FAST NU

Text visualization:
What-Why-How

What is text data?

Documents

Articles, books and novels

E-mails, web pages, blogs

Text Snippets

Tweets, SMS messages

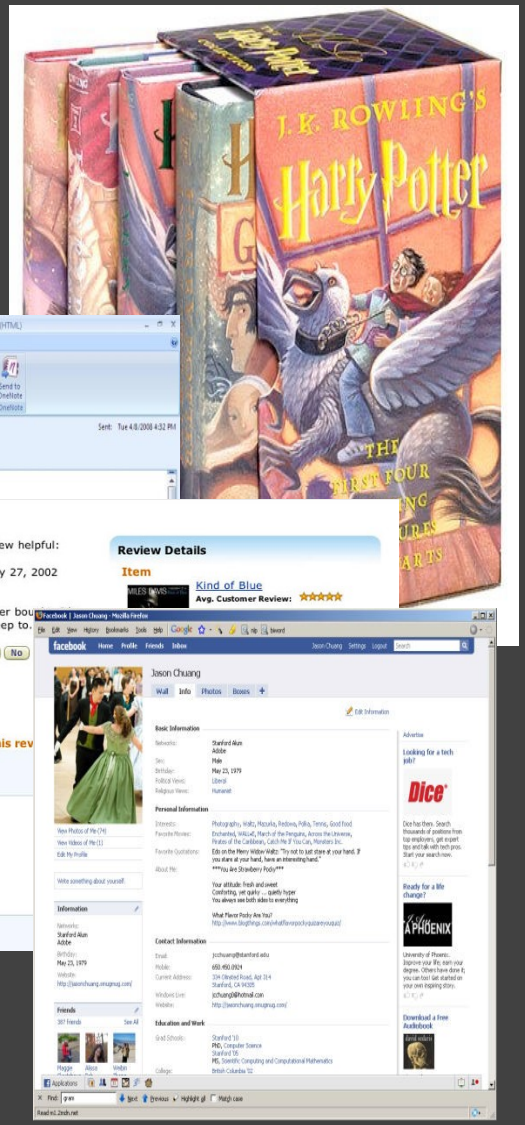
Tags, comments, profiles

And More...

Computer programs, logs

This slide!

Collections of documents



Why Text is Tough

Text is not pre-attentive

Text consists of abstract concepts

- which are difficult to visualize

Text represents similar concepts in many different ways

- space ship, flying saucer, UFO, figment of imagination

Text has very high dimensionality

- Tens or hundreds of thousands of features
- Many subsets can be combined together

Text Meaning is NOT pre-attentive

SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC

Why Text is Tough

Abstract concepts are difficult to visualize
Combinations of abstract concepts are even more difficult to visualize

- time
- shades of meaning
- social and psychological concepts
- causal relationships

Why Text is Tough

The Dog.

Why Text is Tough

The Dog.

The dog cavorts.

The dog cavorted.

Why Text is Tough

The man.

The man walks.

Why Text is Tough

The man walks the cavorting dog.

So far, we can sort of show this in pictures.

Why Text is Tough

As the man walks the cavorting dog, thoughts arrive unbidden of the previous spring, so unlike this one, in which walking was marching and dogs were baleful sentinals outside unjust halls.

How do we visualize this?

Why Text is Tough

Language only hints at meaning
Most meaning of text lies within our minds and common understanding

- "How much is that kitty in the window?"
  - "how much" social system of barter and trade (not the size of the cat)
  - "kitty" implies a kitten here
  - "in the window" implies behind a store window, not really inside a window, requires notion of window shopping

Why Text is Tough

General categories have no standard ordering (nominal data)
Categorization of documents by single topics misses important distinctions
Consider an article about

- NAFTA
- The effects of NAFTA on truck manufacture
- The effects of NAFTA on productivity of truck manufacture in the neighboring cities of El Paso and Juarez

Why Text is Tough

Other issues about language

ambiguous (many different meanings for the same words and phrases)

different combinations imply different meanings

Why Text is Tough

I saw Pathfinder on Mars with a telescope.

Pathfinder photographed Mars.

The Pathfinder photograph mars our perception of a lifeless planet.

The Pathfinder photograph from Ford has arrived.

The Pathfinder forded the river without marring its paint job.

Why Text is Easy

Text is easier when you have a lot of it

- Highly redundant
- Because people are good at finding associations, just about *any* simple algorithm can get "good" results for coarse tasks
  - Pull out "important" phrases
  - Find "meaningfully" related words
  - Create "summary" from document
- **Major problem: Evaluation**

People usually search on relatively coarse meanings

Why Text is Easy

Pretty much any simple technique can pull out phrases that seem to characterize a document
E.g. Most frequent words from an example lecture:

109	slide	69	to	37	view	37	version	37	graphic	37	first								
37	back	36	previous	36	next	32	of	31	the										
30	recall	28	relevant	27	precision	25	retrieved	25	documents										
21	and	18	evaluate	15	a	13	what	13	vs	13	how								
12	trec	12	is	12	high	12	for	10	relevance										
10	queries	10	on	9	information	8	x	8	why										
8	as	8	answer	7	search	7	maron	7	document										
7	blair	6	top	6	results	6	measure												
6	length	6	in	6	evaluation	6	curves												

Why Text is Easy

Same text, removing most frequent words in language and most frequent in this text:

30 recall 28 relevant 27 precision 25 retrieved 25 documents
18 evaluate 13 vs 12 trec 12 high 10 relevance
10 queries 9 information 8 x 8 answer 7 search
7 maron 7 document 7 blair 6 top 6 results
6 measure 6 length 6 evaluation 6 curves

These words can act as a simple summary of the document

people are good at inferring the relations
redundancy in the word meanings

Why visualize text?

Understanding – read a document

Summaries – get the "gist" of a document

Clustering – group together similar contents

Quantify – convert to numerical measures

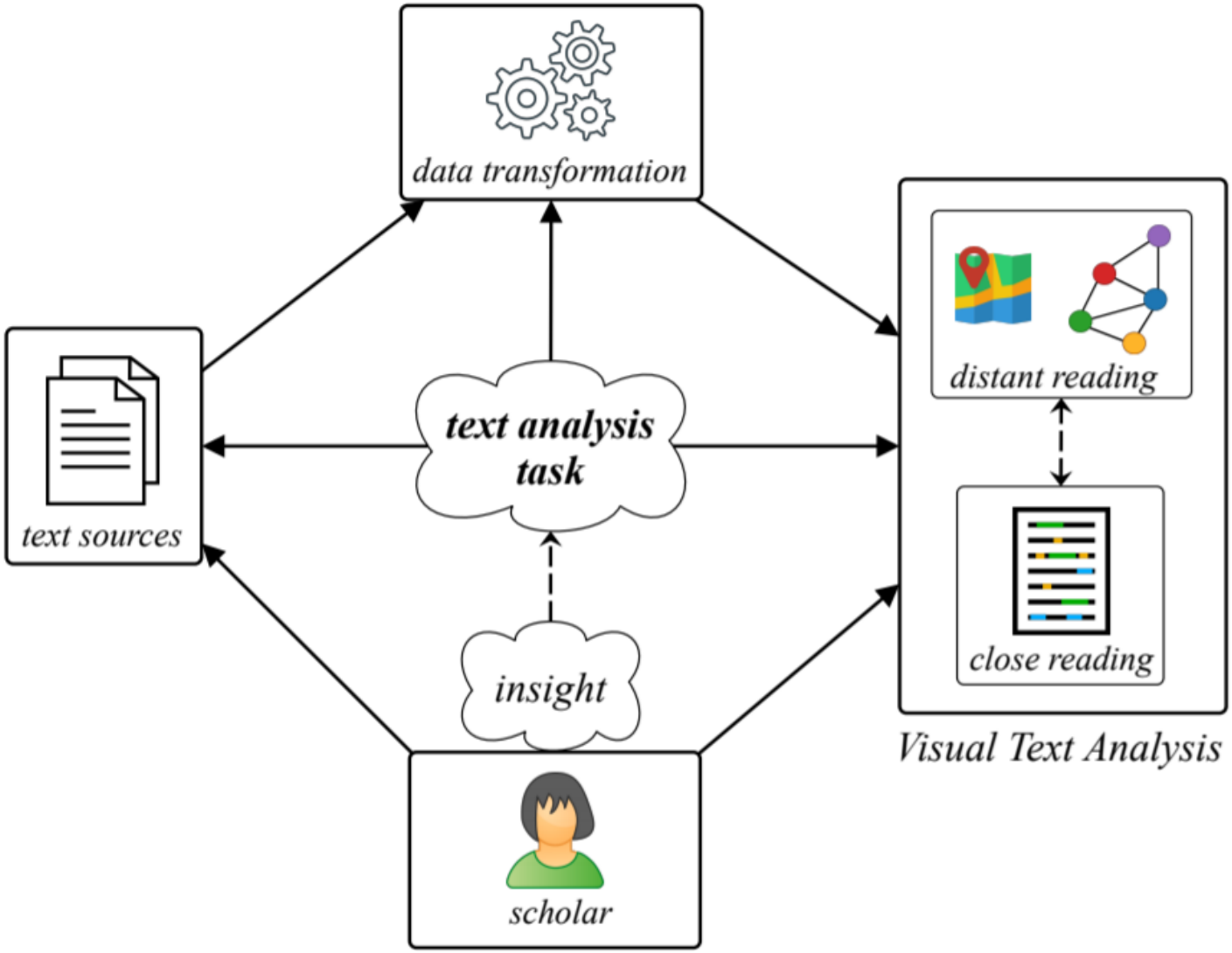
Correlate – compare patterns in text to those in other data, e.g., correlate with social network

Why visualize text?

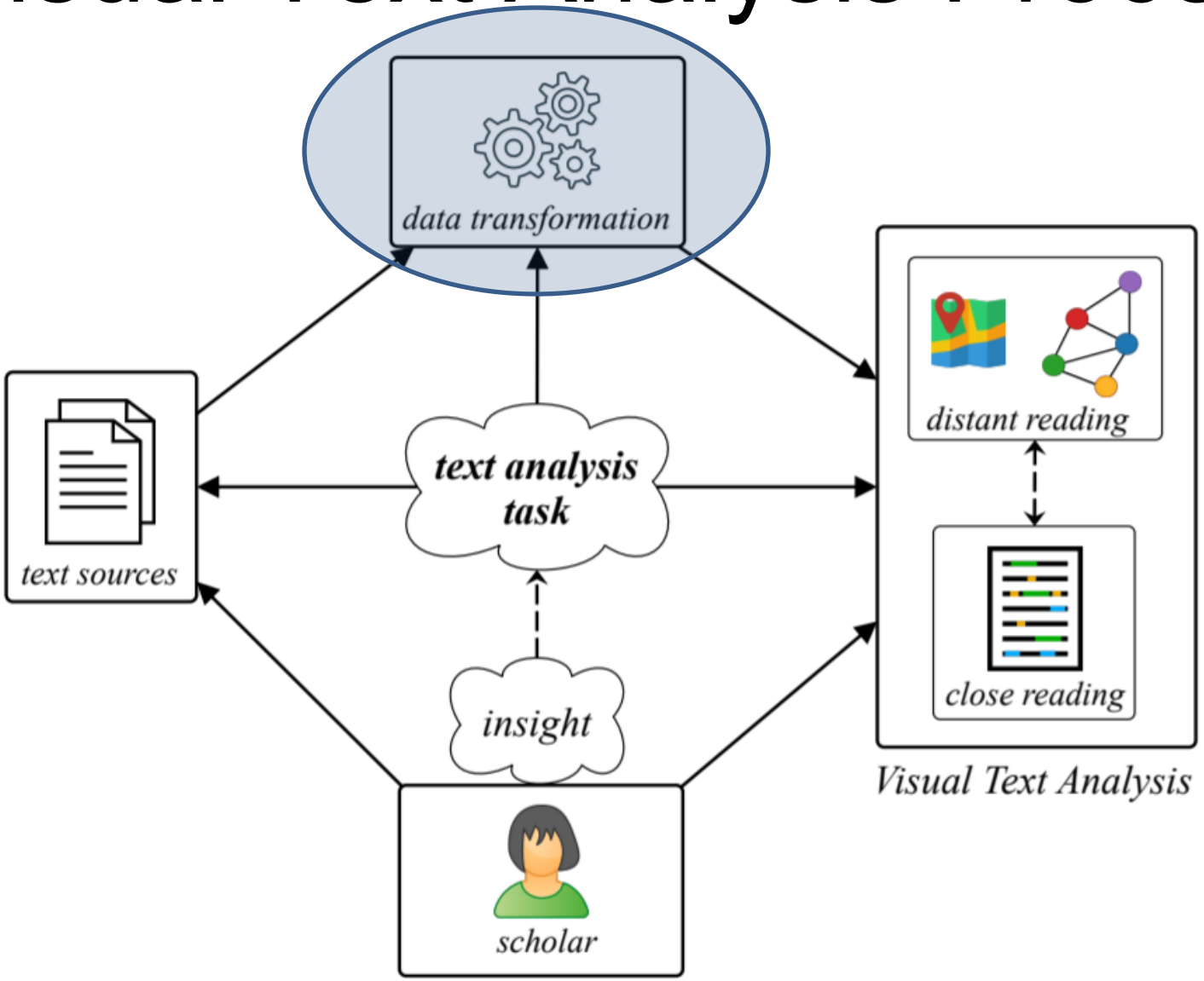
=

Visual **Text Analysis Task**

Visual Text Analysis Process



Visual Text Analysis Process



Data/Text Transformation

In other words ..

Text pre-processing

Example of a text pre-processing process

- Tokenization of text
- Stemming of words
- Ordered lists of words
- Bag of word model
- Document vector space model

Text pre-processing steps

Tokenization

- Segment text into terms.
- Remove stop words? *a, an, the, of, to, be*
- Numbers and symbols? *#gocard, @stanfordfball, Beat Cal!!!!!!!!*
- Entities? *San Francisco, O'Connor, U.S.A.*

Stemming

- Group together different forms of a word.
- Porter stemmer? *visualization(s), visualize(s), visually* → *visual*
- Lemmatization? *goes, went, gone* → *go*

Ordered list of terms

Bag of Words Model

Ignore ordering relationships within the text

A document ≈ vector of term weights

- Each dimension corresponds to a term (10,000+)
- Each value represents the relevance For example, simple term counts

Aggregate into a document-term matrix

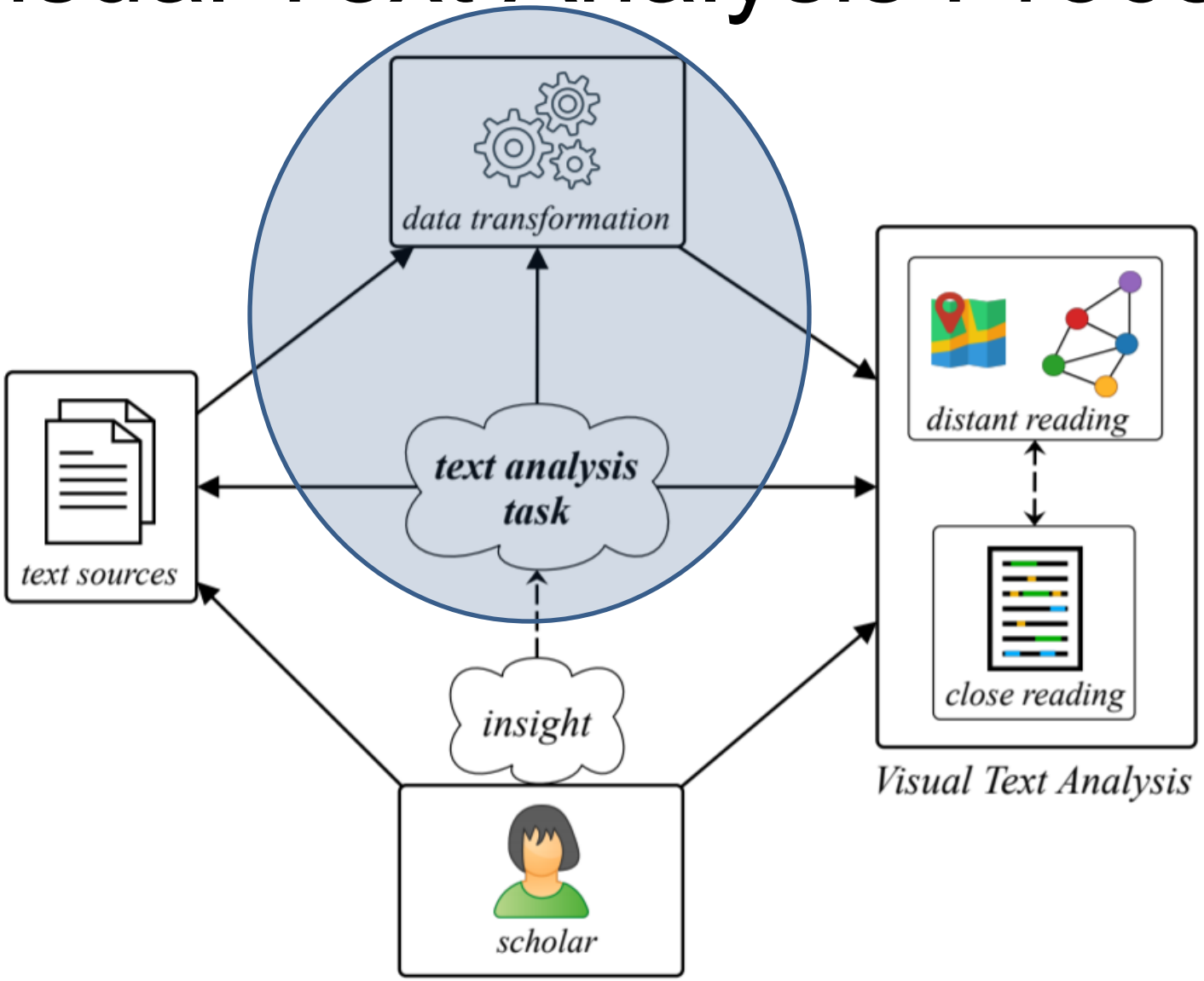
- Document vector space model

Document-Term Matrix

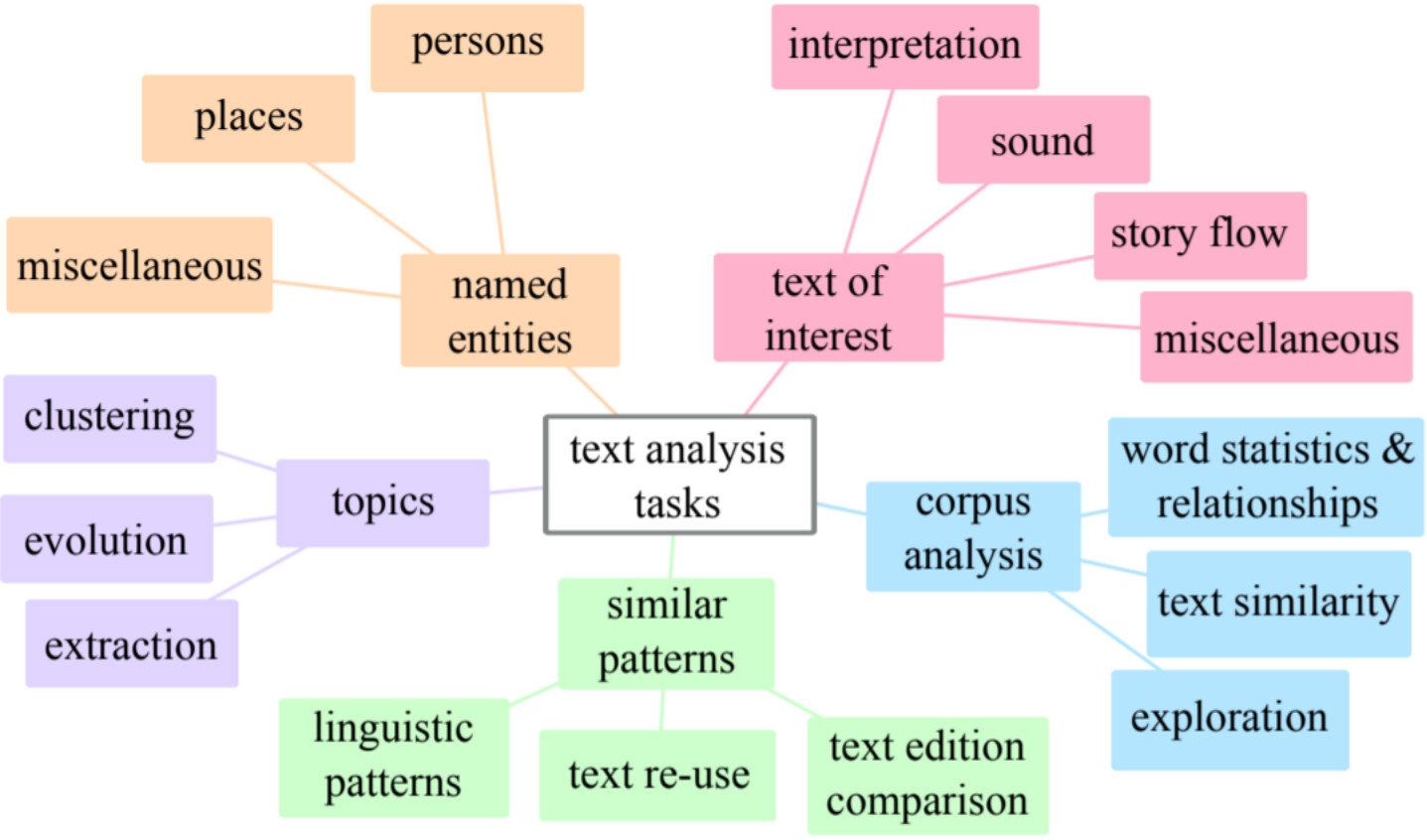
Each document is a vector of term weights
Simplest weighting is to just count occurrences

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

Visual Text Analysis Process



Taxonomy of Text analysis tasks



Text analysis tasks

Named Entities

The analysis of named entities is a common text analysis task to extract places (fictional or reported geographies), persons / characters in a story or miscellaneous entities from a single text or a whole collection of text.

- Persons
- Places
- Miscellaneous

Topics

The analysis of topics consists of tasks like topic extractions, so that major topics in the source texts can be tracked and topic-related text passages can be discovered. The presence of temporal data allows for the analysis of topical evolution, and on the basis of the found topics, a topical clustering of a corpus is possible.

- Clustering
- Evolution
- Extraction

Text analysis tasks

Corpus Analysis

An important text analysis task is explore text corpora containing a high number of texts. Tasks such as the analysis of word statistics & relationships among text, text similarity between the texts etc. play important role in corpus analysis tasks

- Word statistics & similarity
- Text similarity
- Exploration

Text of interest

Some tasks focus an individual literary work, which we call text of interest. The underlying research tasks vary from visualizing text interpretations to the analysis of sound of literary works (mostly poems), Interpretation and to the story flow analysis of a given source text.

- Sound
- Story flow
- Miscellaneous

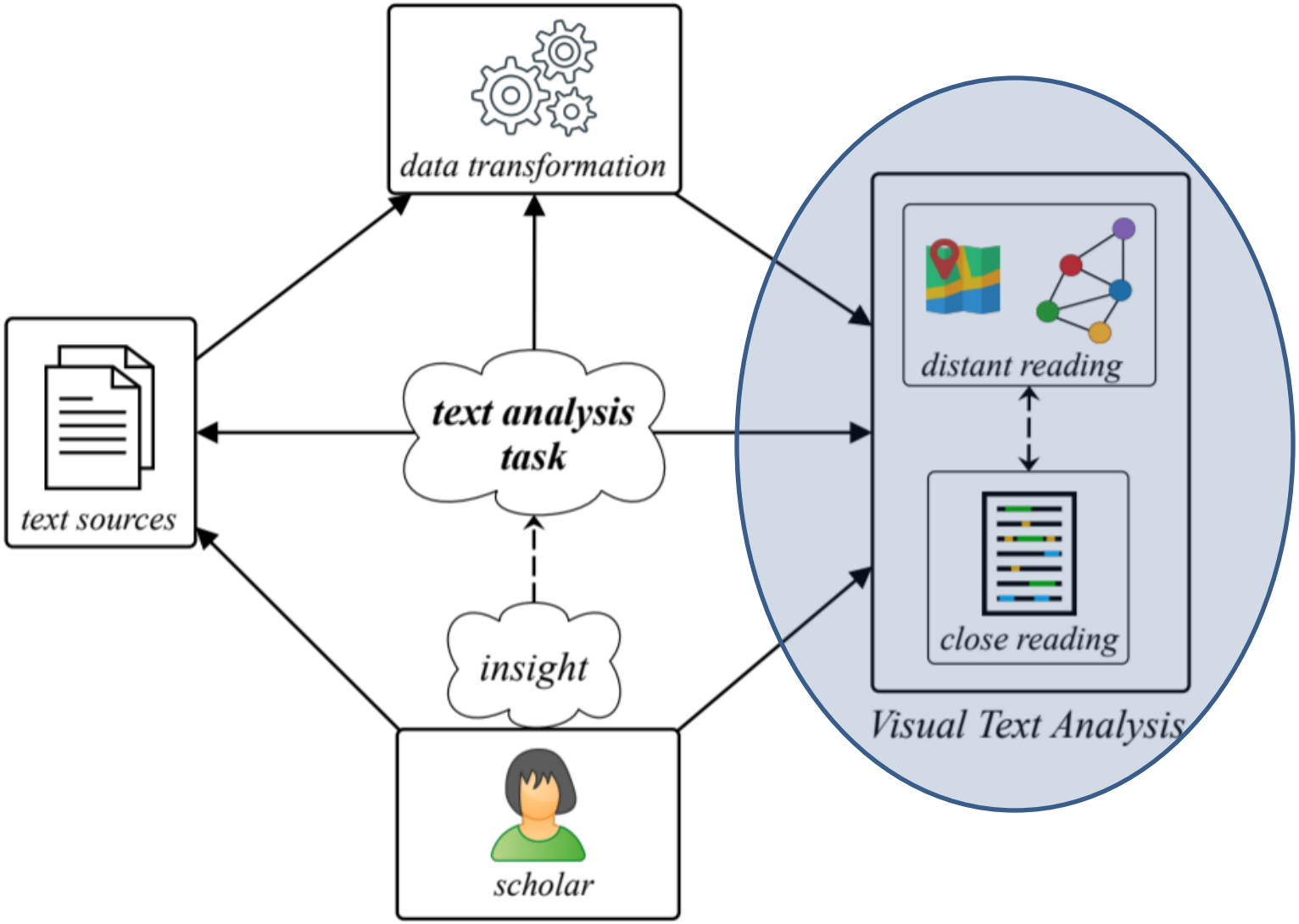
Text analysis tasks

Similar Patterns

An analysis of similar patterns that includes the discovery, the alignment and the visualization of similar text segments among the texts of a given collection is a typical text analysis task. Dependent on the length of patterns, the tasks are divided into three sets. While the analysis of linguistic patterns concerns short phrases, text re-use analysis focuses on determining deliberately re-used text segments (e.g., quotes or plagiarized passages)., whereas text editions focus on comparison of whole text documents.

- Linguistic Patterns
- Text re-use
- Text edition comparison

Visual Text Analysis Process



Visual Text analysis

Close reading

Close reading is the thorough interpretation of a text passage (usually single document) by determination of central themes and the analysis of their development. E.g.

- individuals, events, and ideas, their development and interaction
- used words and phrases
- vocabulary (word frequency, distribution, structure)
- semantic structure
- content

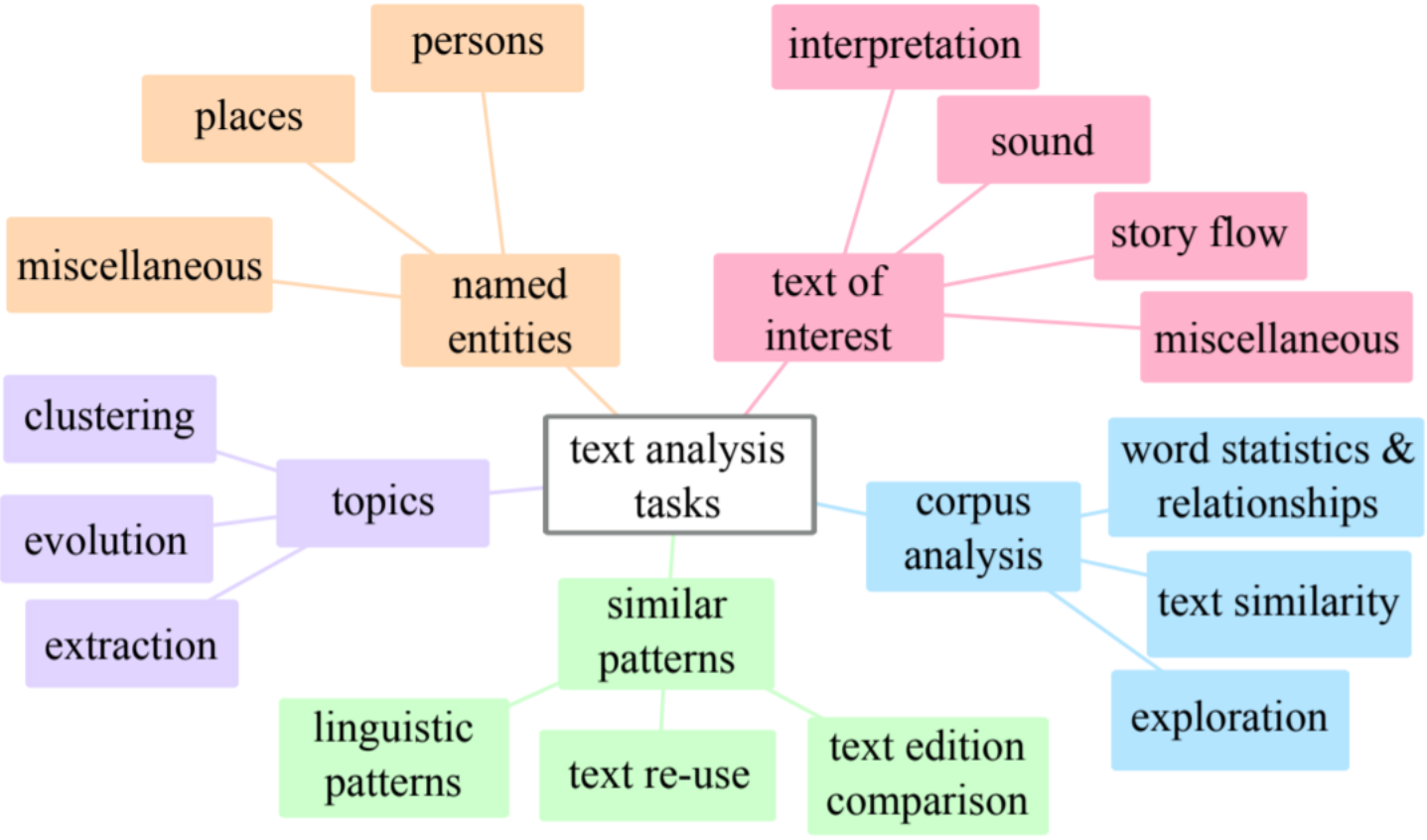
Visual Text analysis

Distant reading

While close reading retains the ability to read the source text without dissolving its structure, distant reading does the exact opposite. It aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s). E.g.

- document themes
- changes over time
- document relationships
- document similarity

Lets map text analysis tasks to Close and Distant reading visualizations

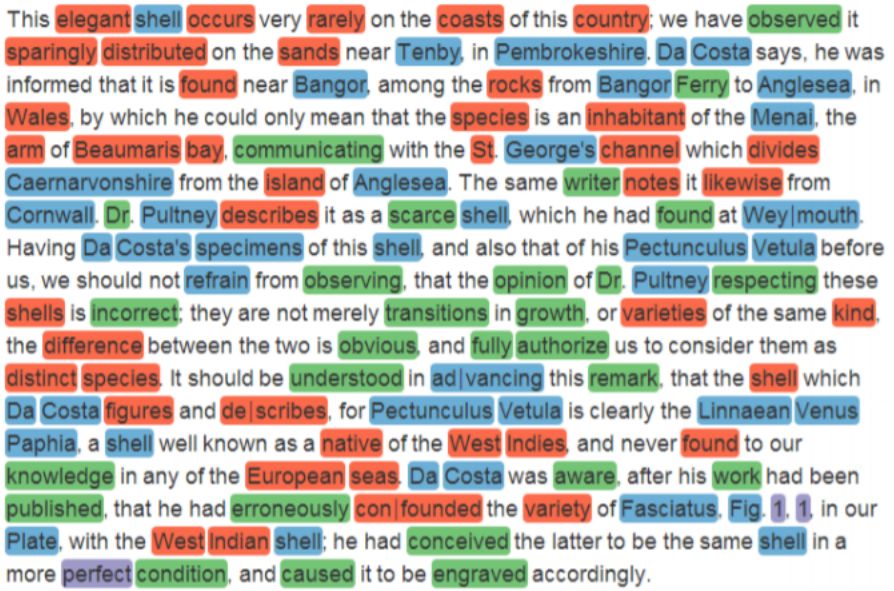


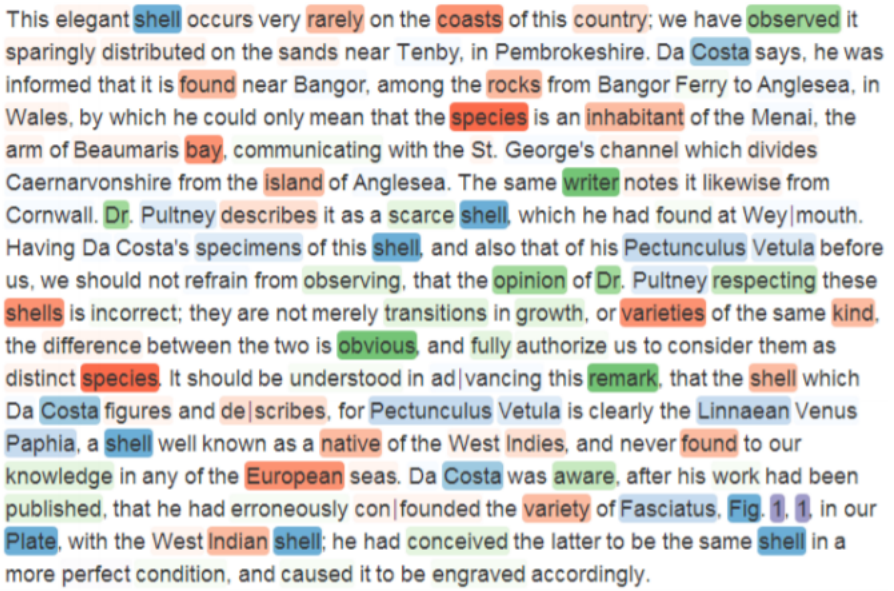
Close reading visualization techniques

IMPORTANT: Focus is on highlighting actual TEXT

- Colors
- Font sizes
- Glyphs
- Connections

Close reading techniques
Color

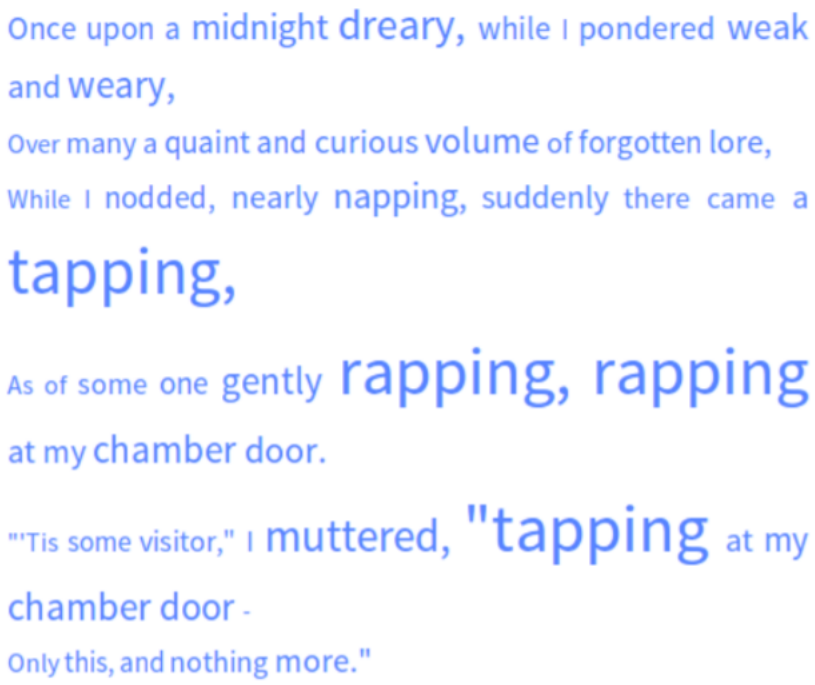




Close reading techniques
Font size

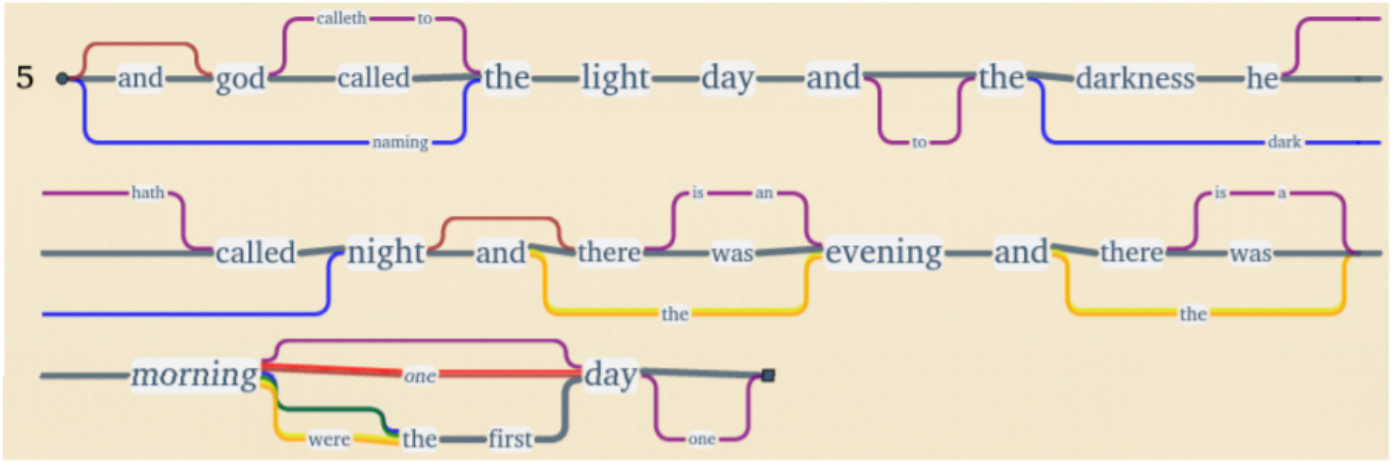
Once upon a midnight dreary, while I pondered weak and weary,
Over many a quaint and curious volume of forgotten lore,
While I nodded, nearly napping, suddenly there came a tapping,
As of some one gently rapping, rapping at my chamber door.
"'Tis some visitor," I muttered, "tapping at my chamber door -
Only this, and nothing more."

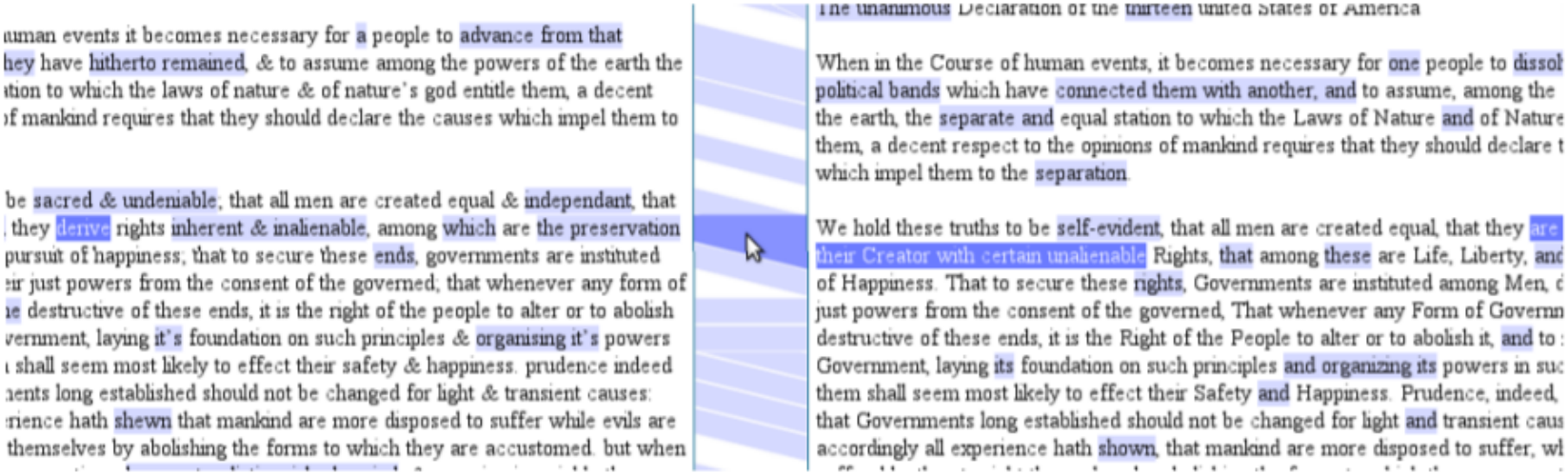
Ah, distinctly I remember it was in the bleak December,
And each separate dying ember wrought its ghost upon the floor.
Eagerly I wished the morrow; - vainly I had sought to borrow
From my books surcease of sorrow - sorrow for the lost Lenore -
For the rare and radiant maiden whom the angels named Lenore -
Nameless here for evermore.



Close reading techniques
Glyphs

Close reading techniques
Connections





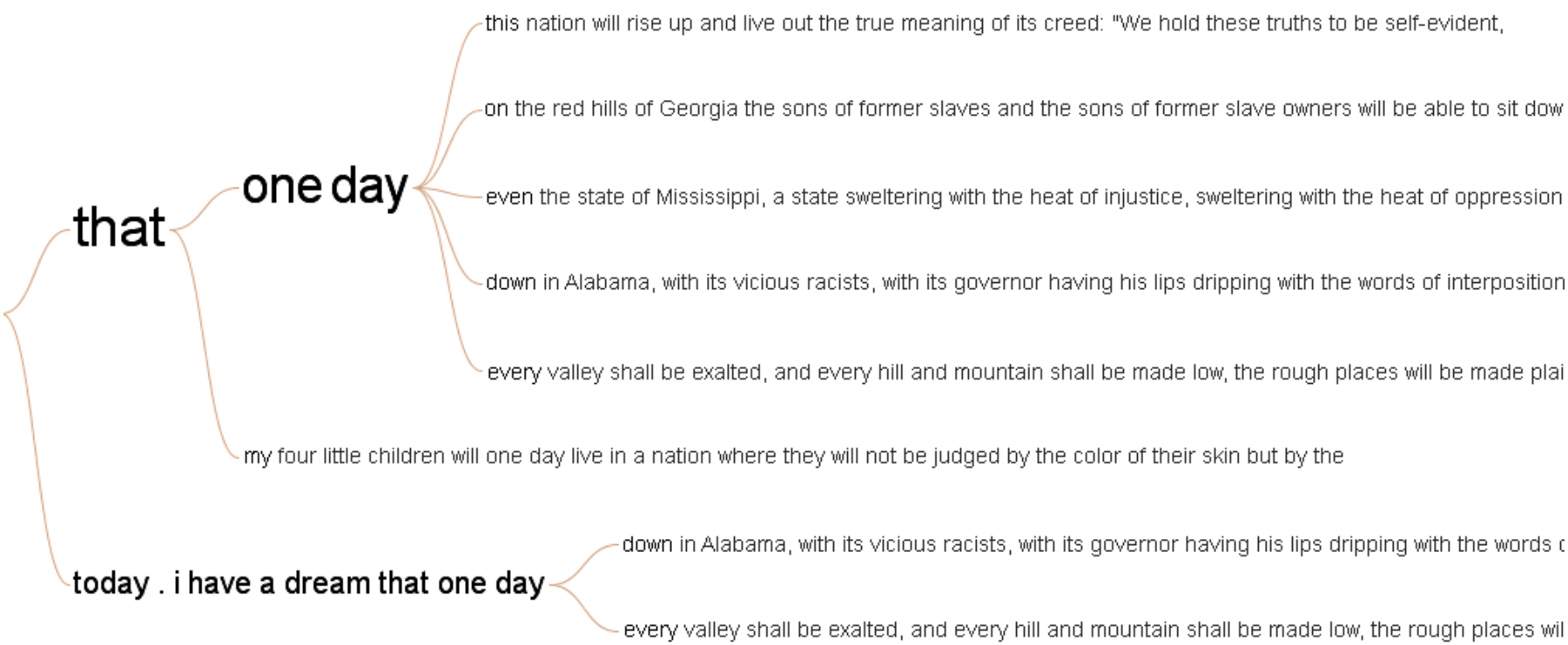
TextArc (One Document)

[textarc.org]

TextArc

- Three rules:
  - Show the entire text in an ellipse around the page: line-by-line and word-by-word
  - Like tag clouds, use larger font-size and brighter text for frequent words
  - Central words move to the middle (links to its mentions)

Word Tree (One Document)

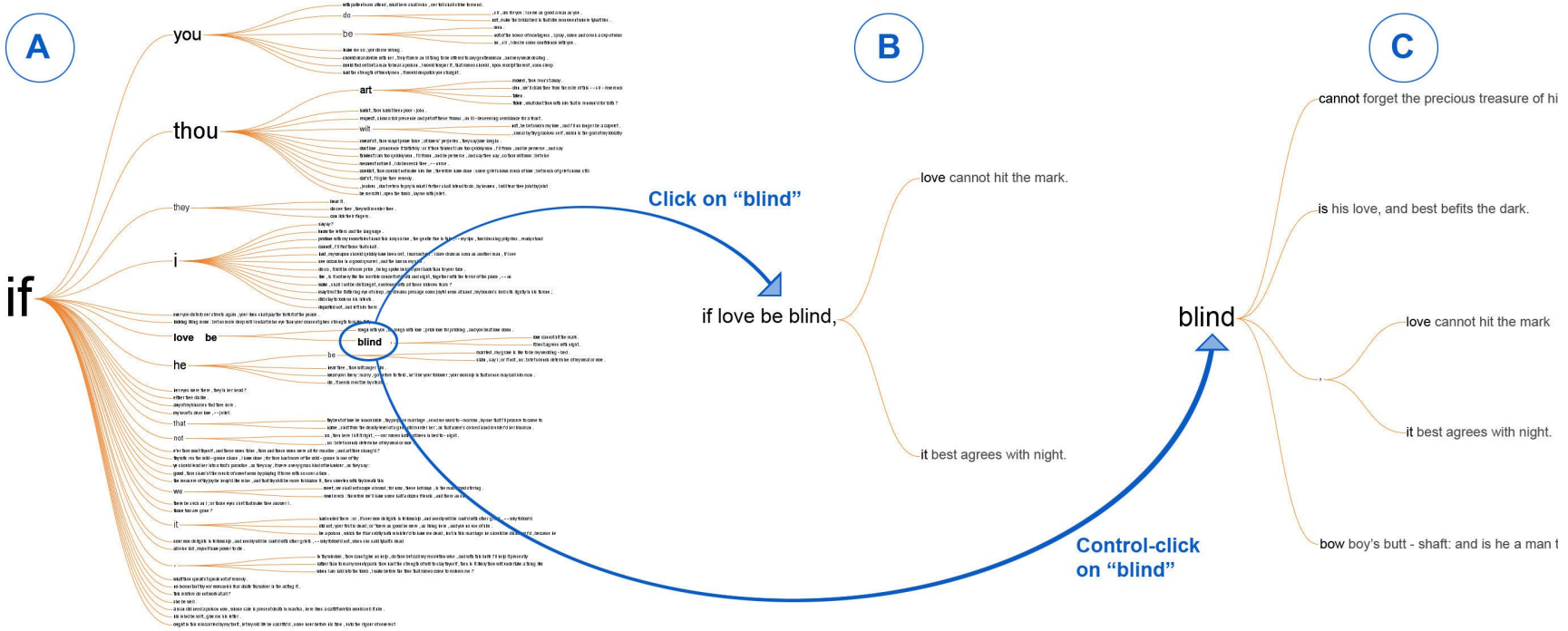


[Wattenberg & Viegas, 2007]

Word Tree

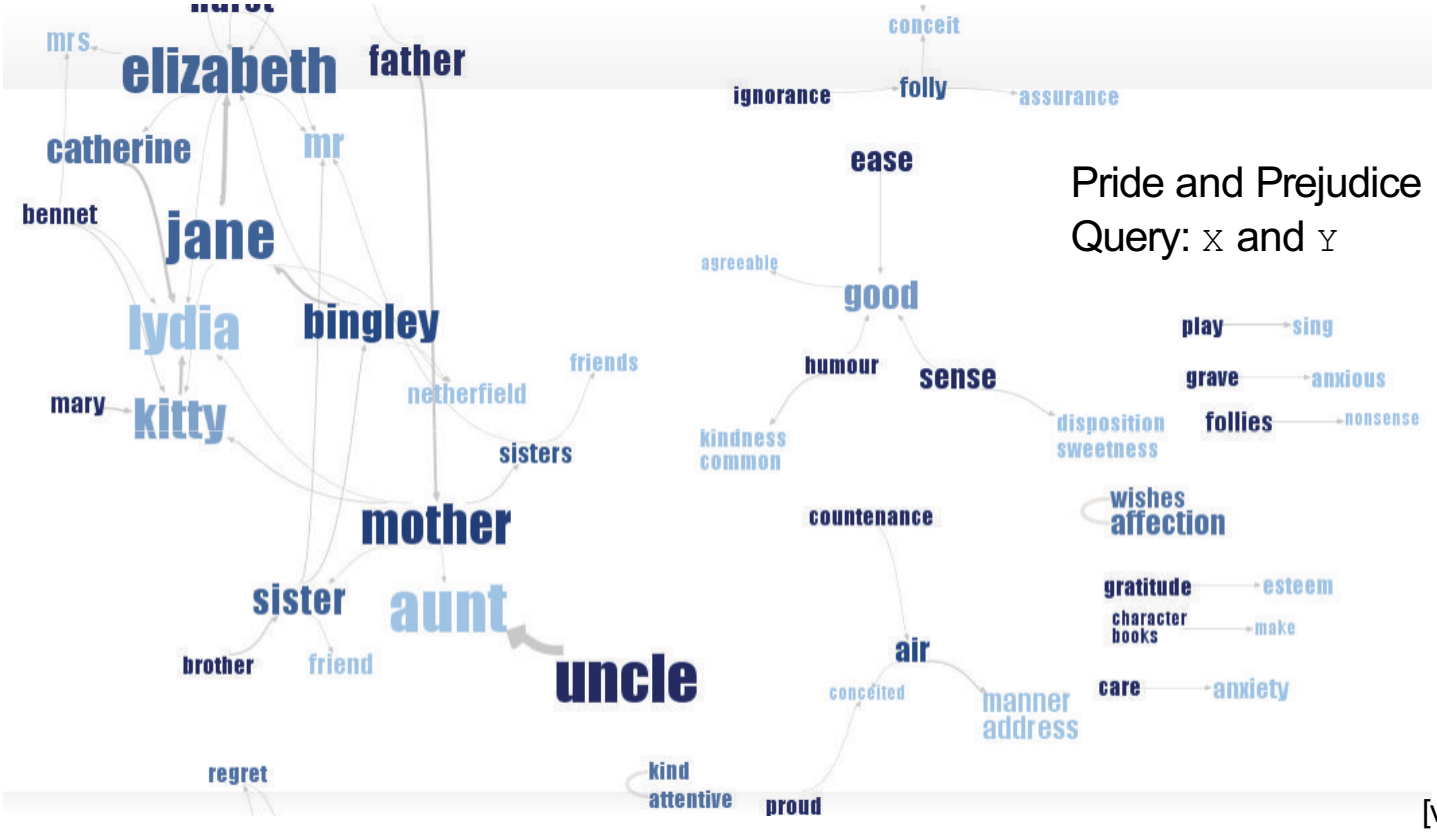
- A "Visual Concordance"
- Shows phrasing, relationships between words
- Starting point is a single word or snippet
- Branches to show common words/phrases that follow
- Goal is to show context: "keyword-in-context"

Interaction in Word Tree



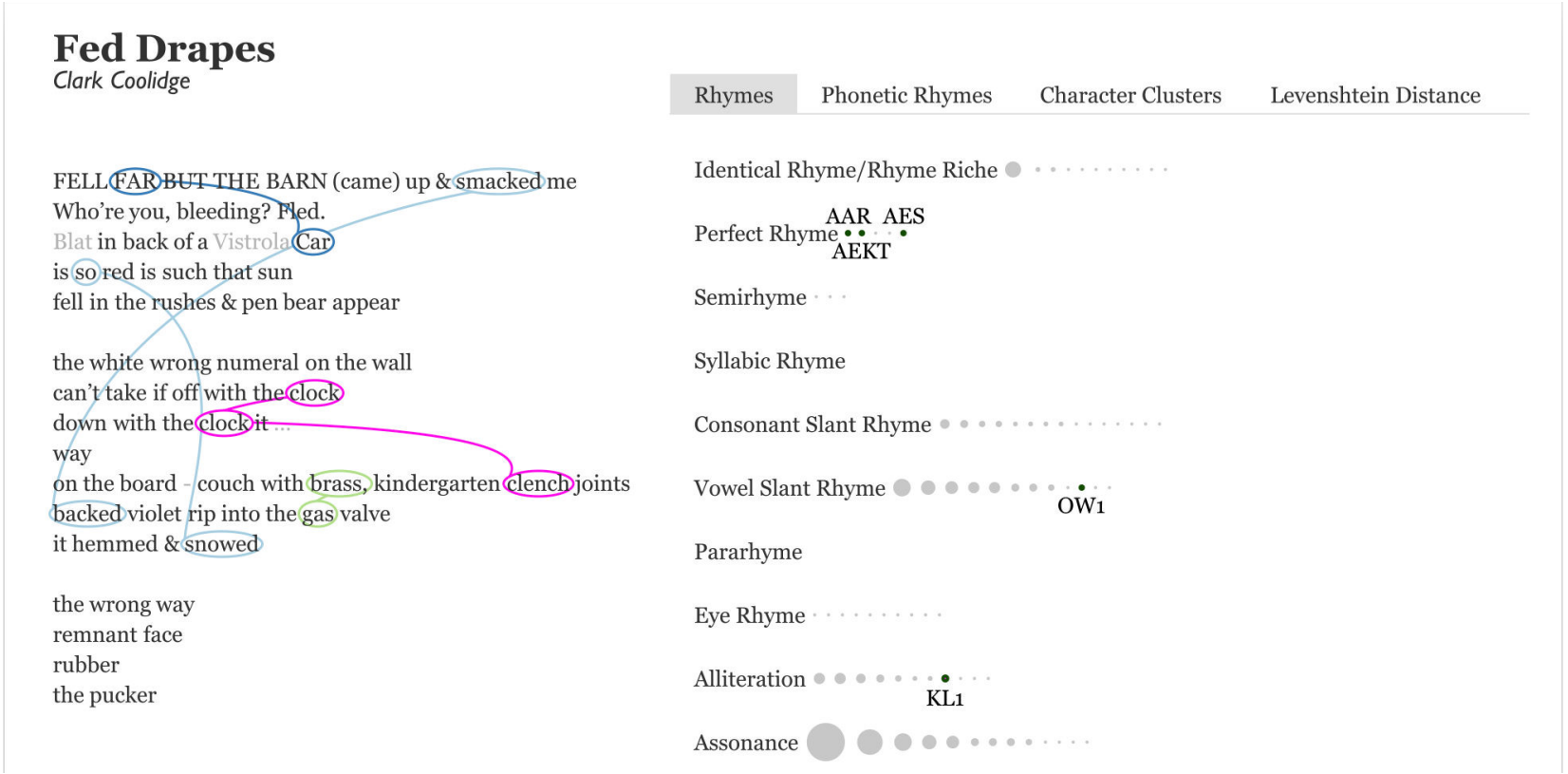
[Wattenberg & Viegas, 2007]

Phrase Nets



[van Ham, 2009]

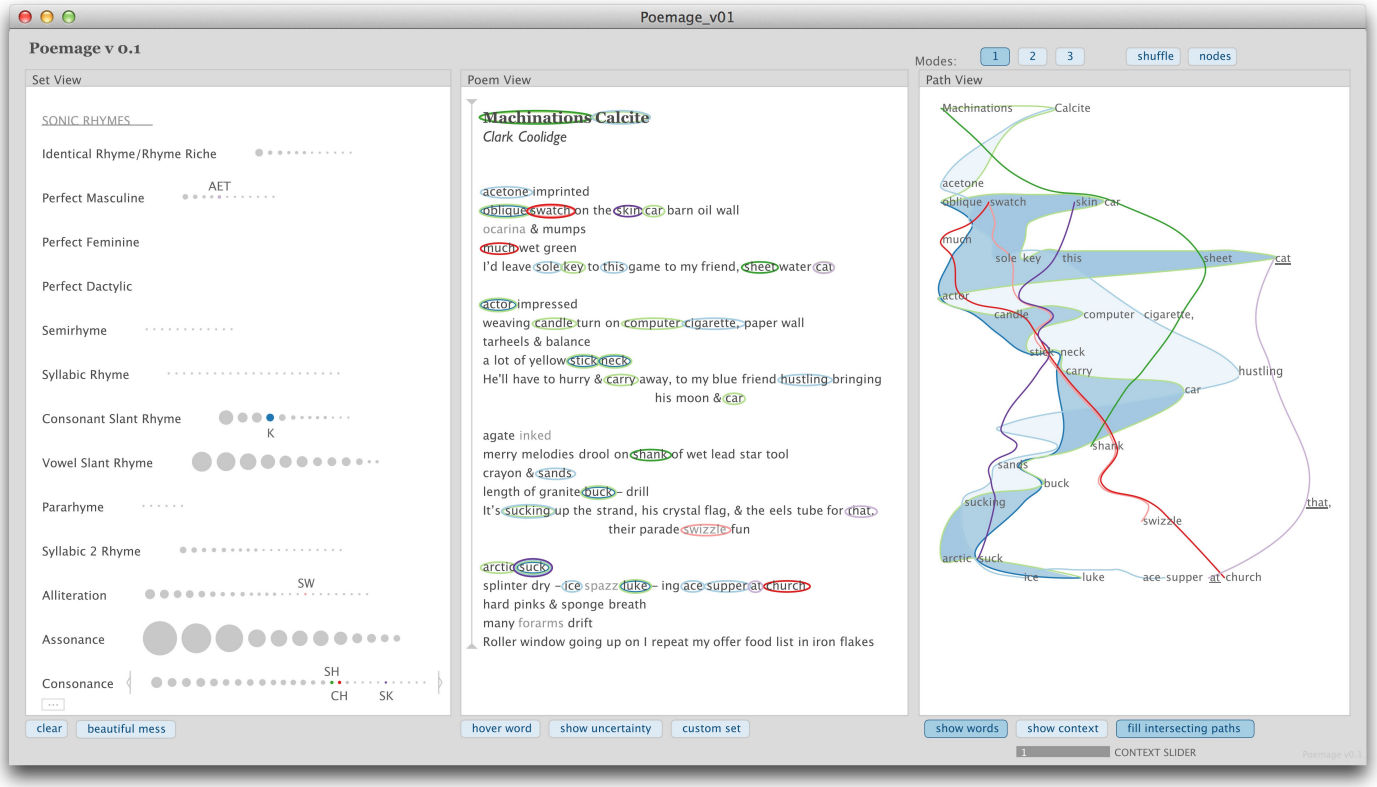
Words are more than just character sequences



Poemage

- Support close reading—in-depth reading to generate as much productive meaning as possible
- Search for poetic devices: affect, imagery, pun, metaphor
- Sound and linguistic devices → Rhyming
  - Identical: pare/pair
  - Perfect: picky/tricky
  - Assonance & consonance: blue/estuaries, shell/chiffon
  - Eye rhyme: cough/bough
- Support exploration: scholars do not want computers to "solve" poems

Interface


Poemage_v01
Poemage v 0.1
Modes:
1
2
3
shuffle
nodes
Set View
SONIC RHYMES
Identical Rhyme/Rhyme Riche
Perfect Masculine
AET
Perfect Feminine
Perfect Dactylic
Semirhyme
Syllabic Rhyme
Consonant Slant Rhyme
K
Vowel Slant Rhyme
Pararhyme
Syllabic 2 Rhyme
Alliteration
SW
Assonance
Consonance
SH
CH
SK
clear
beautiful mess
Poem View
Machinations Calcite
Clark Coolidge
acetone imprinted
oblique swatch on the skin car barn oil wall
ocarina & mumps
much wet green
I'd leave sole key to this game to my friend, sheet water cat
actor impressed
weaving candle turn on computer cigarette, paper wall
tarheels & balance
a lot of yellow stick neck
He'll have to hurry & carry away, to my blue friend hustling bringing his moon & car
agate inked
merry melodies drool on shank of wet lead star tool
crayon & sands
length of granite buck- drill
It's sucking up the strand, his crystal flag, & the eels tube for that, their parade swizzle fun
arctic suck
splinter dry -ice spazz luke- ing ace supper at church
hard pinks & sponge breath
many forarms drift
Roller window going up on I repeat my offer food list in iron flakes
hover word
show uncertainty
custom set
Path View
Machinations
Calcite
acetone
oblique
swatch
skin
car
much
sole
key
this
sheet
cat
actor
candle
computer
cigarette,
stick
neck
carry
hustling
car
shank
sands
buck
sucking
that,
swizzle
arctic
suck
ice
luke
ace
supper
at
church
show words
show context
fill intersecting paths
CONTEXT SLIDER

Distant reading visualization techniques

- Heat maps
- Tag clouds
- Maps
- Time lines
- Graphs
- Miscellaneous

Distant reading techniques
Heat maps

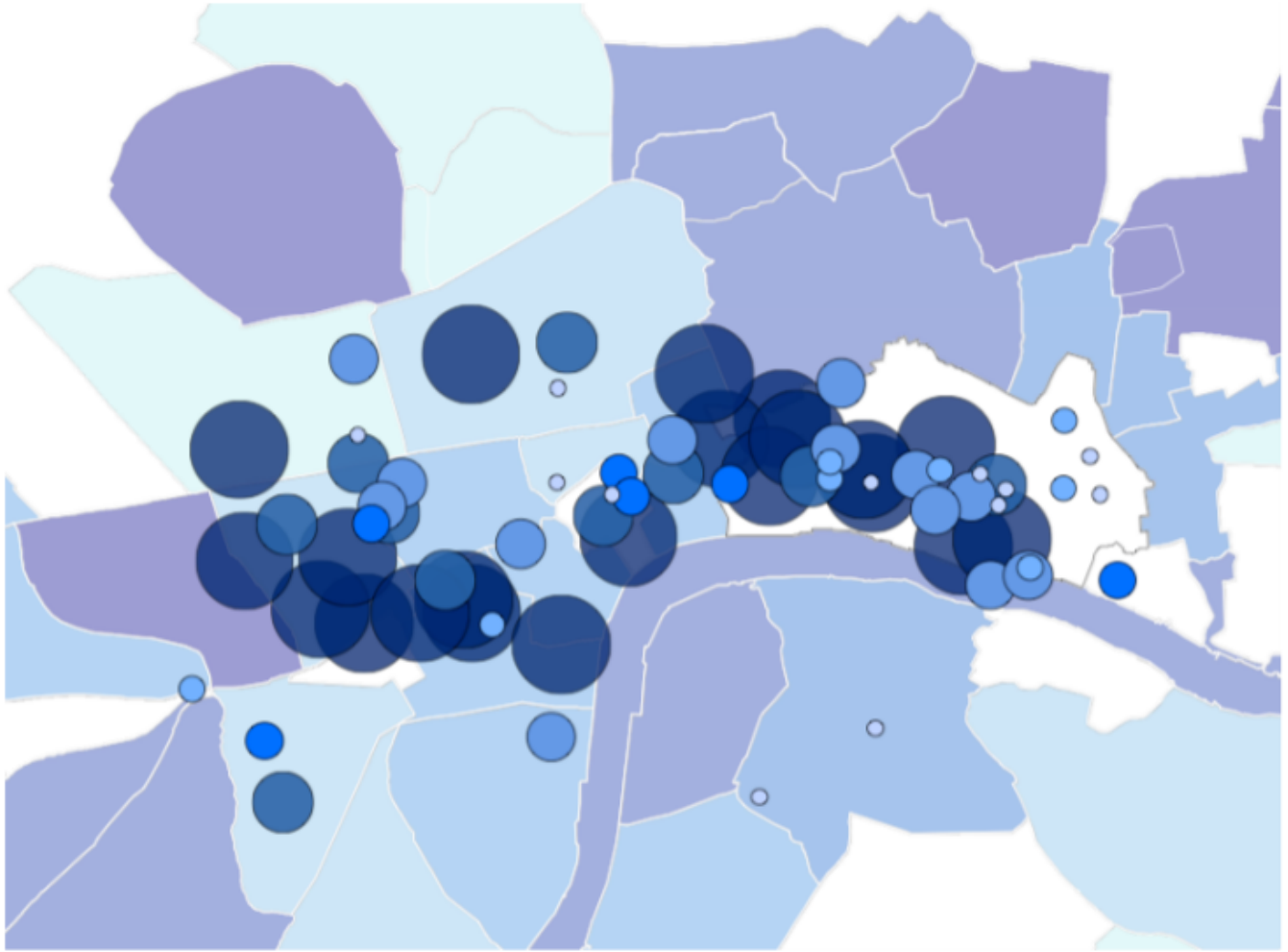
Distant reading techniques
Tag clouds

Tag Cloud (One Document)

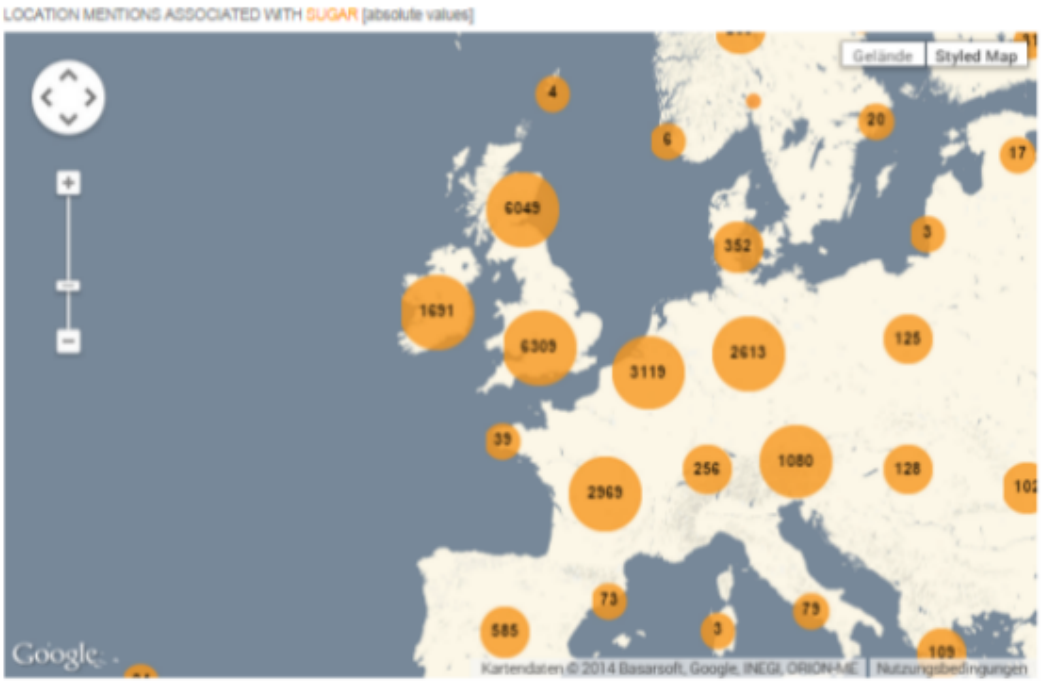
- Derived data: number of occurrences of words
- Channel: Font size
- Potential problem: Think about ink…

[Scray, CC-BY-SA-3.0]

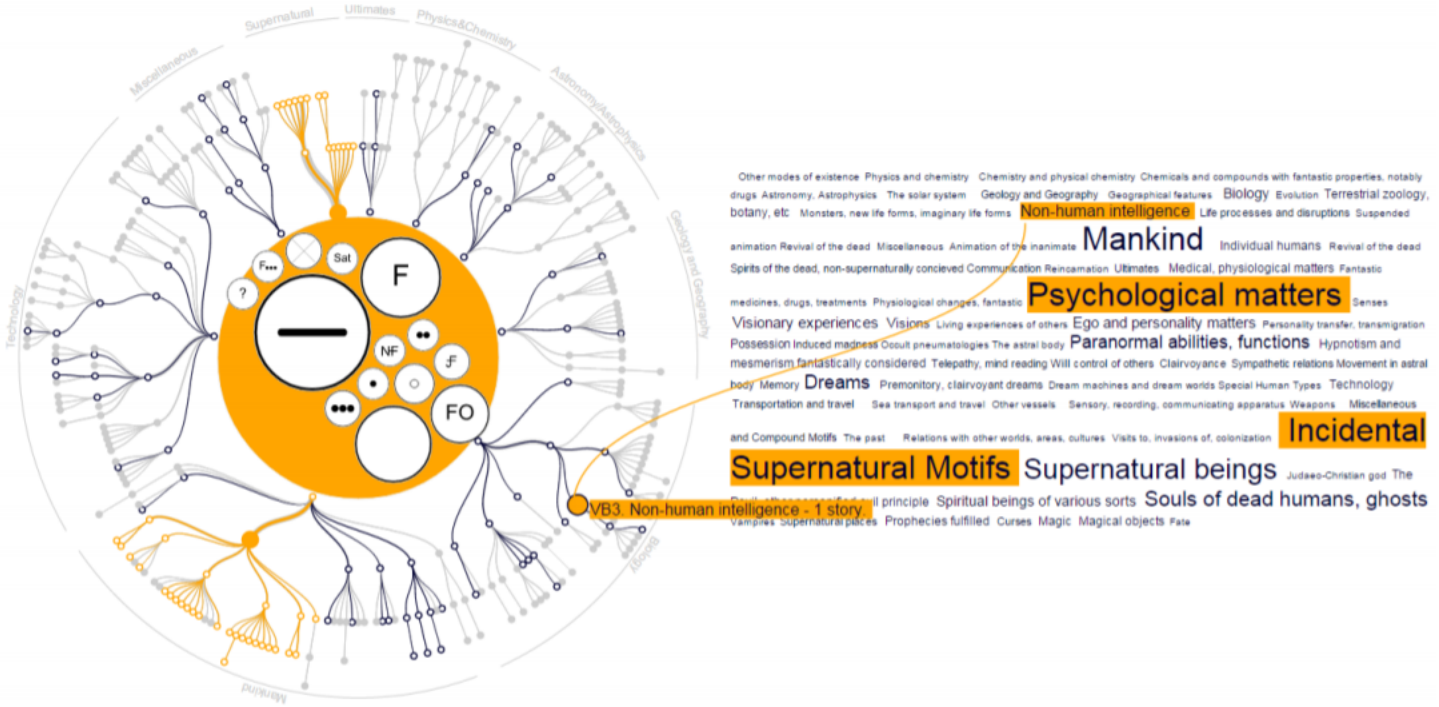
Distant reading techniques
Maps



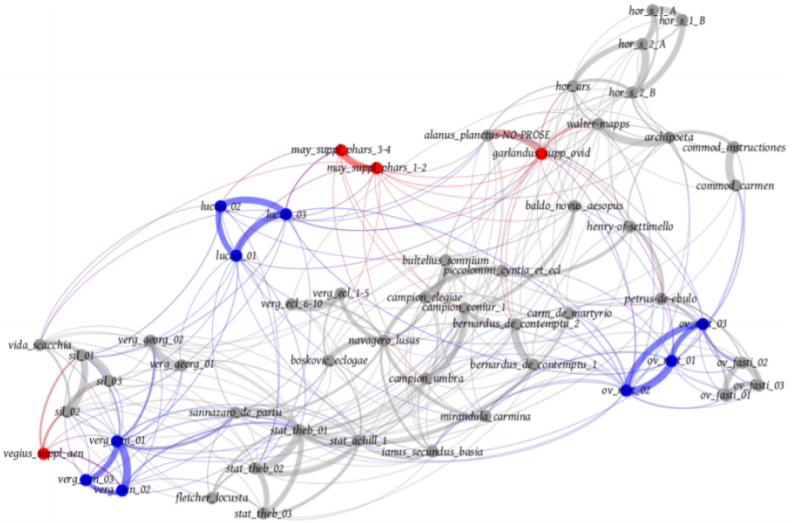
Distant reading techniques
Time lines

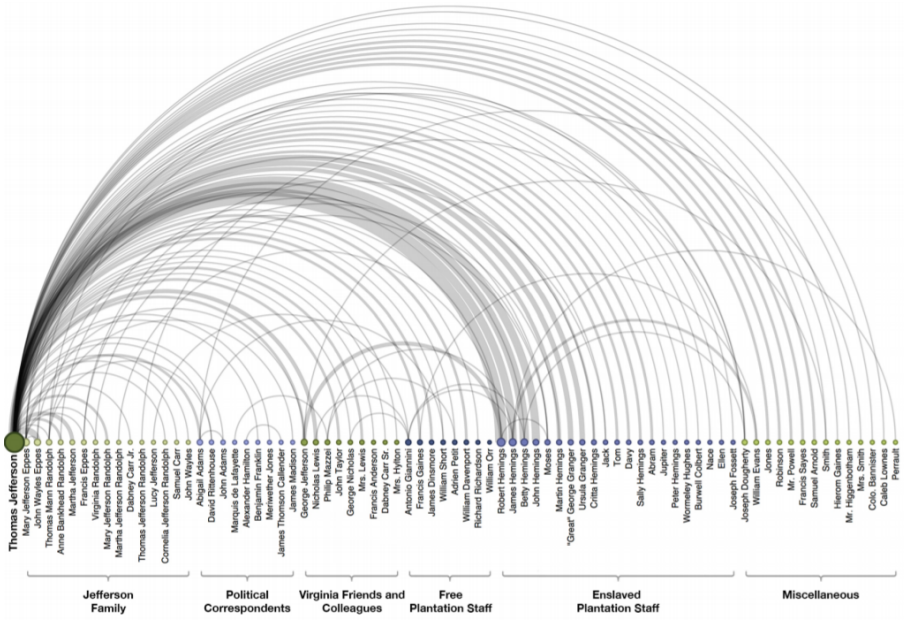


Distant reading techniques
Graphs

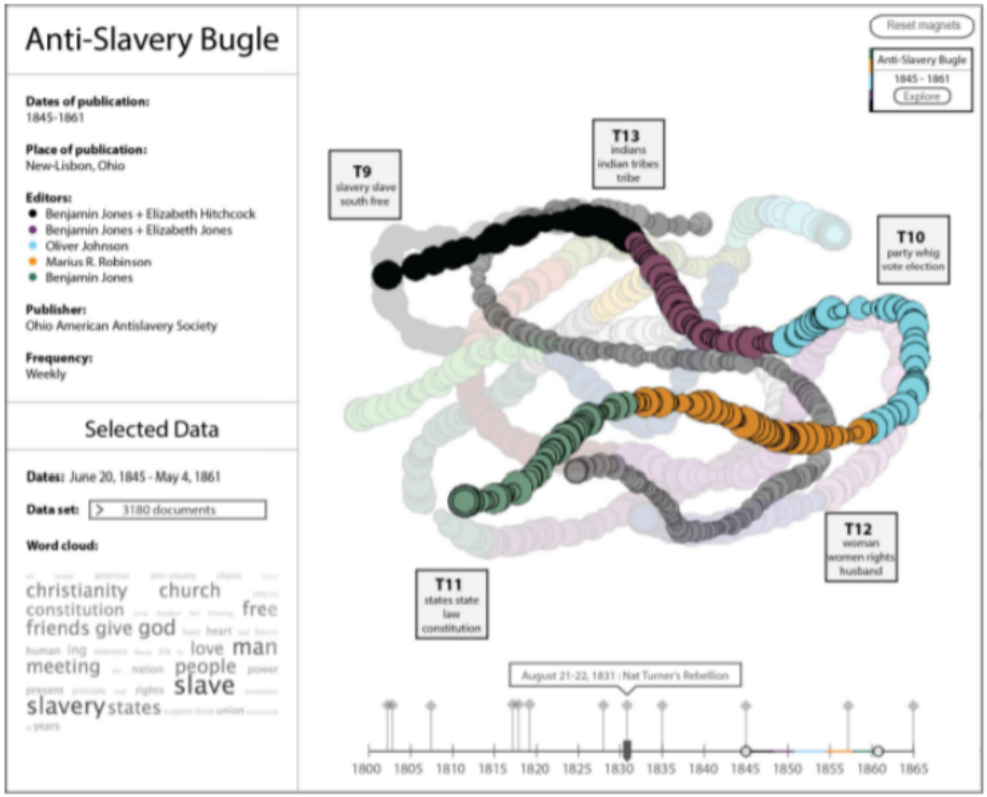


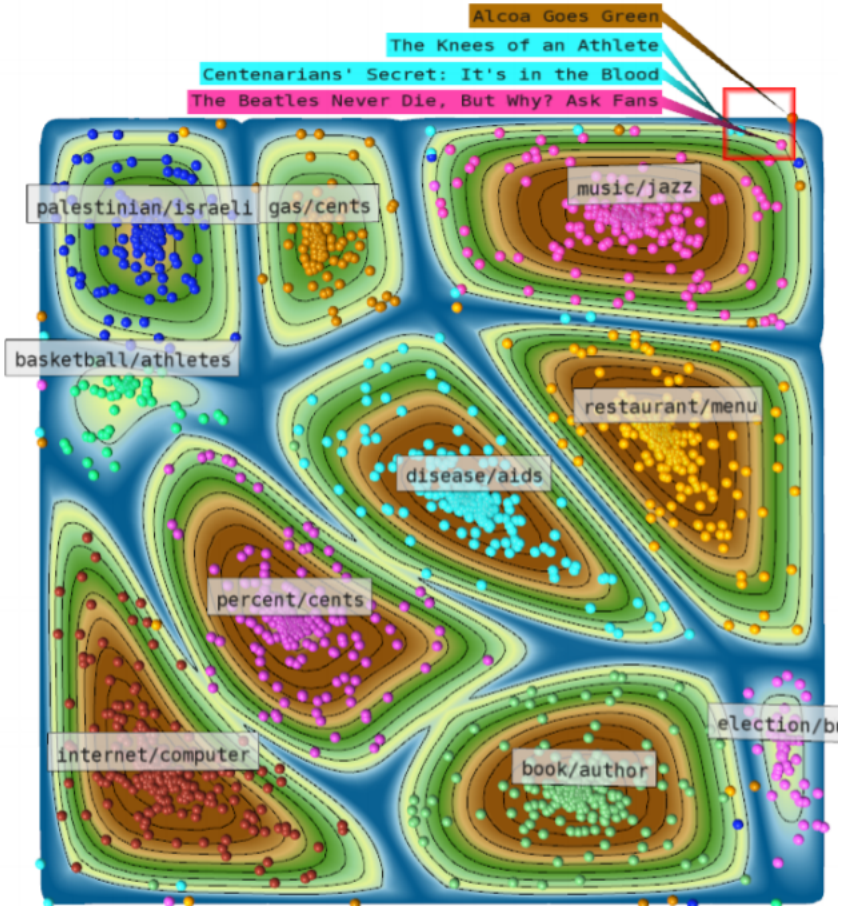
Distant reading techniques
Graphs …





Distant reading techniques
Miscellaneous





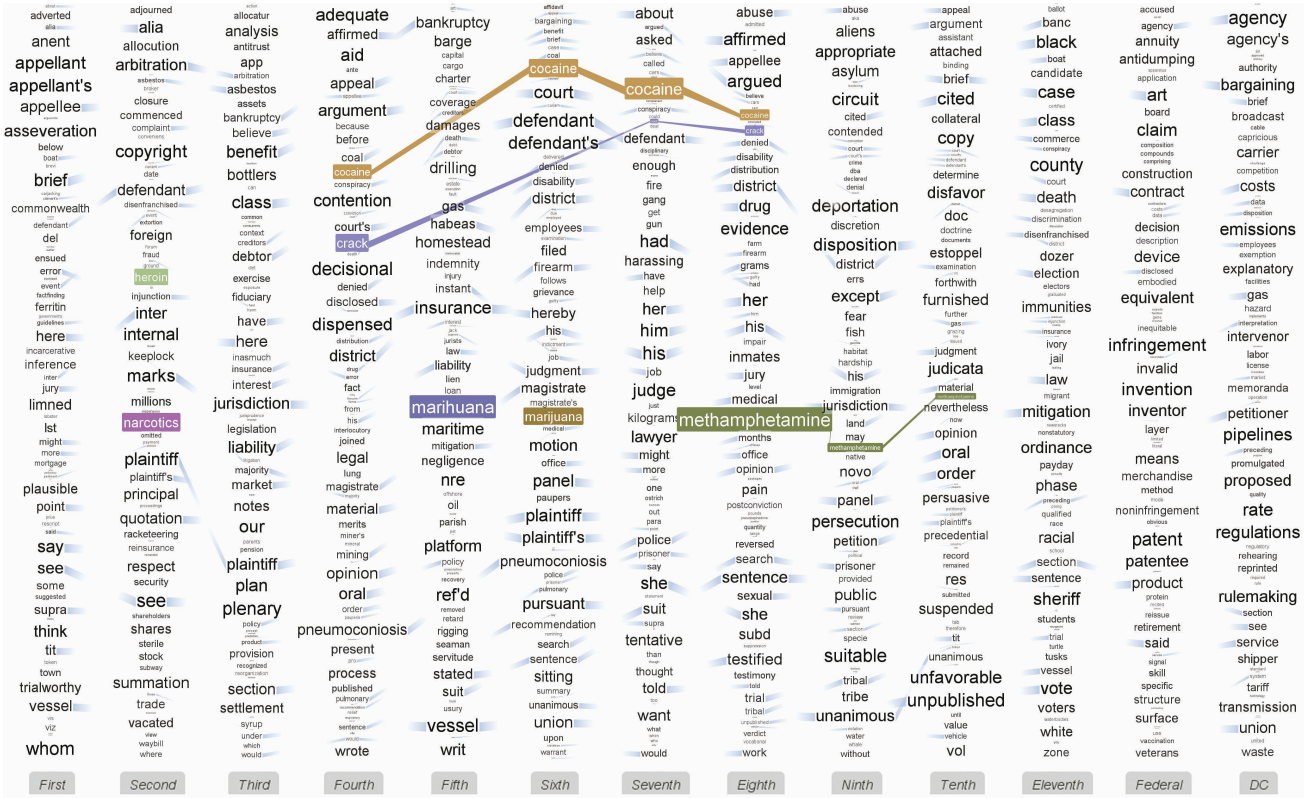
Comparing Documents

- Word choice/usage
- Relationships
- Phrasing

Tag Cloud (Two Documents)

[Pyrsmis, CC-BY-SA-3.0]

Parallel Tag Clouds (Multiple Documents)



[Collins et al., 2009]