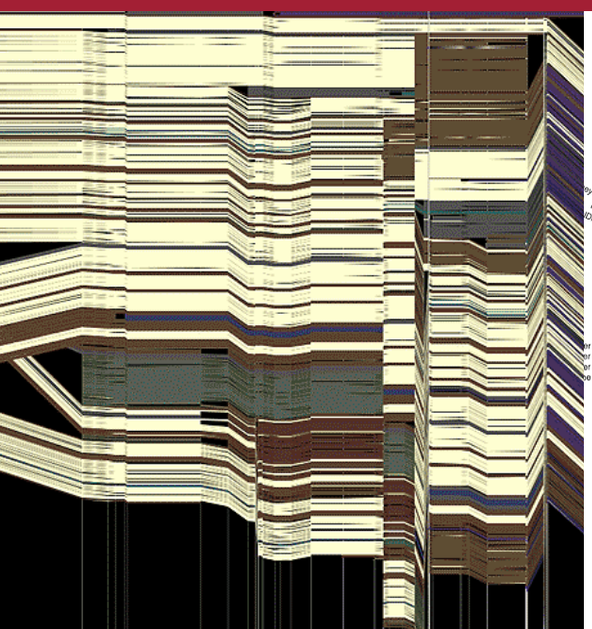
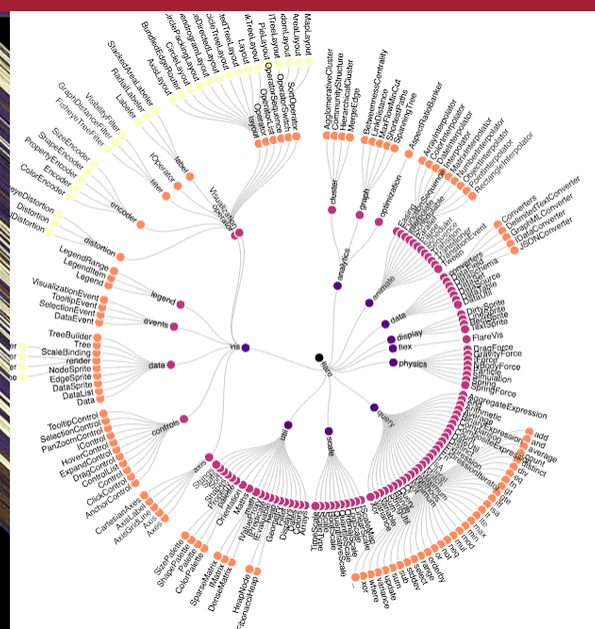
6.894: Interactive Data Visualization

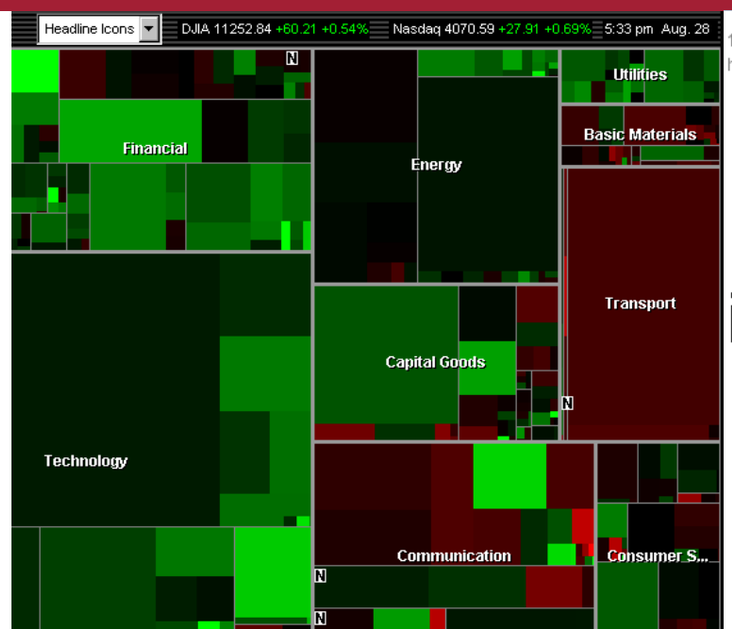
Networks & Text

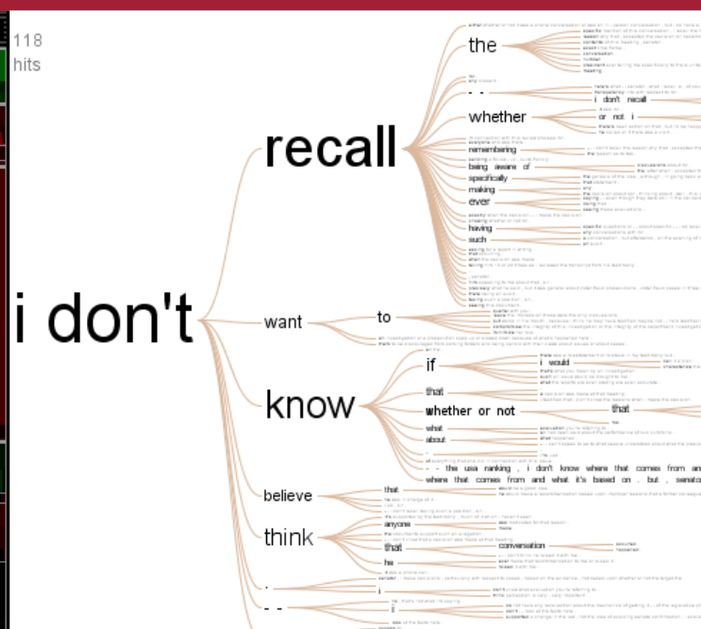
Slides from Jeff Heer

Arvind Satyanarayan









Final Project

Monday, 5/13, 2:30–5pm under origami birds.

Free-standing **32-inch monitor w/HDMI** hookup (bring your own dongle!) + **table for laptops**.

Interactive poster: **60s video** or HTML web page w/embedded visualization.

Final paper now due **Wednesday, 5/15, 11:59pm.**

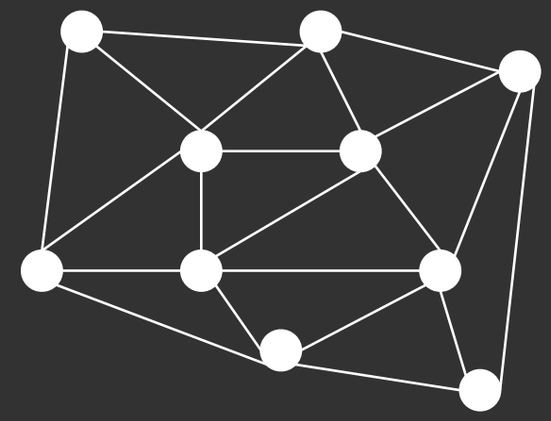
Networks, Trees and Hierarchies

Graphs and Trees

Graphs

Model relations among data

Nodes and *edges*

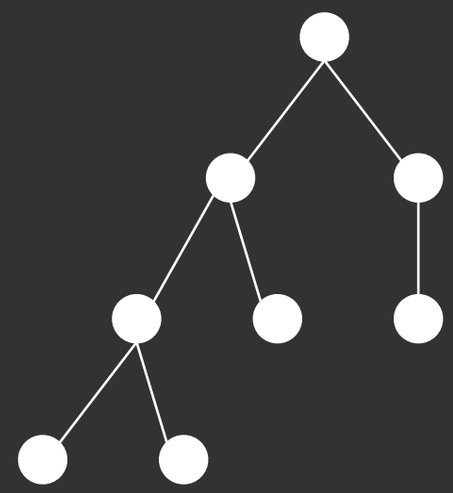


Trees

Graphs with hierarchical structure

Connected graph with N-1 edges

Nodes as *parents* and *children*



Spatial Layout

A primary concern of tree/graph drawing is the spatial arrangement of nodes and edges.

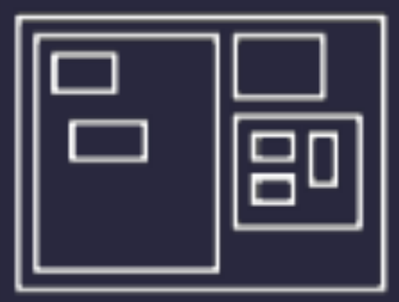
Often (but not always) the goal is to effectively depict the graph structure:

- Connectivity, path-following
- Topological distance
- Clustering / grouping
- Ordering (e.g., hierarchy level)

Tree Visualization

Indentation

Linear list, indentation encodes depth



Enclosure diagrams

Represent hierarchy by enclosure

Layering

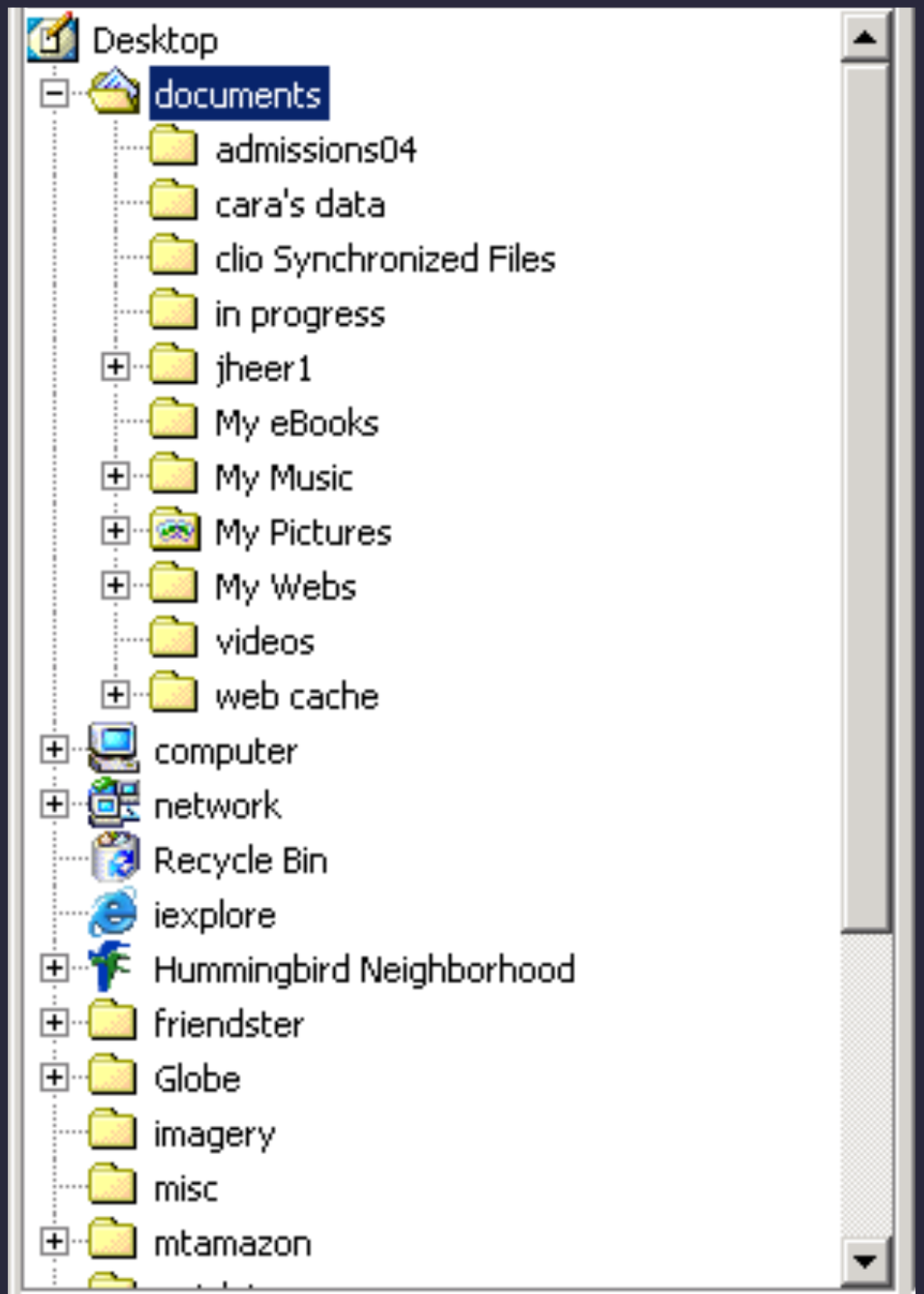
Relative position and alignment

Node-Link diagrams

Nodes connected by lines/curves

Typically fast: O(n) or O(n log n), interactive layout

Indentation



Places all items along vertically spaced rows

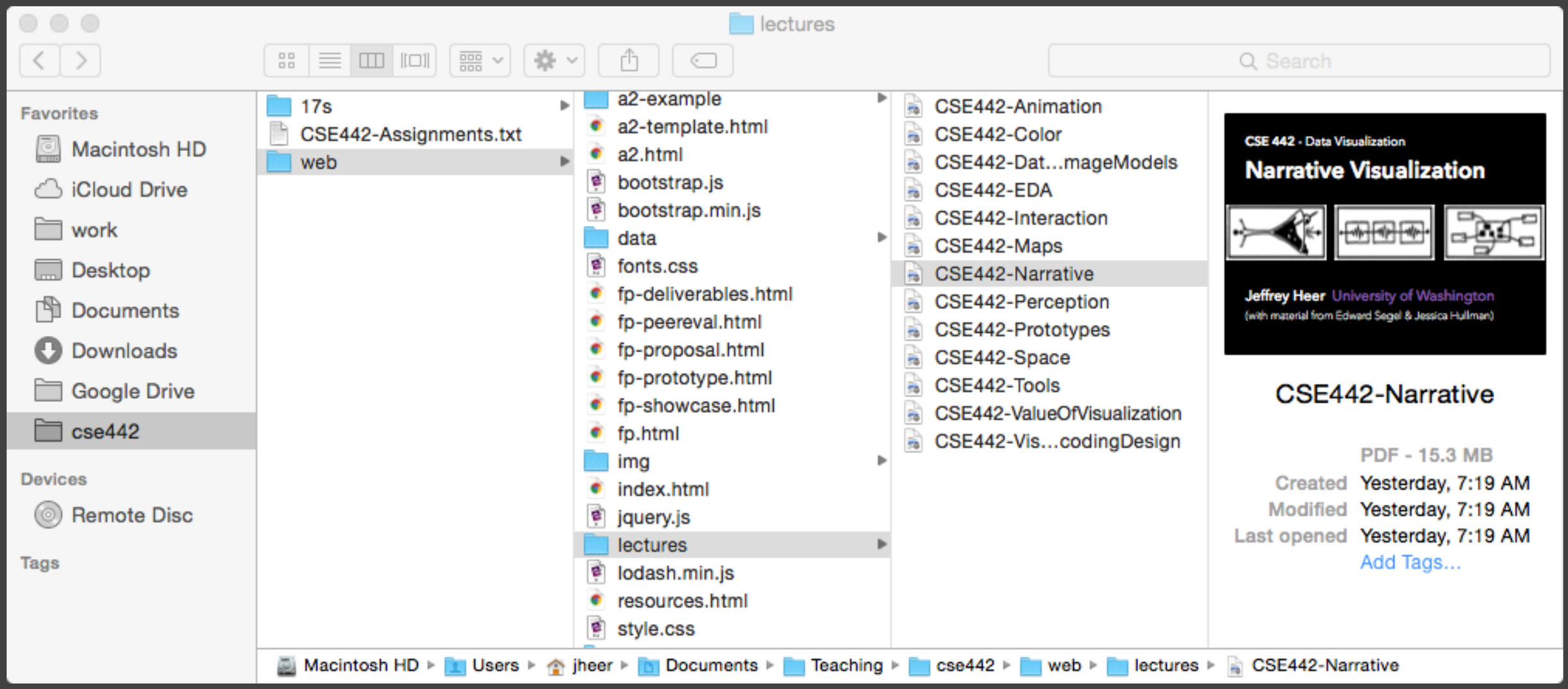
Indentation used to show parent/child relationships

Commonly used as a component in an interface

Breadth and depth contend for space

Often requires a great deal of scrolling

Single-Focus (Accordion) List



Separate breadth & depth along 2D.
Focus on a single path at a time.

Enclosure Diagrams

Encode structure using **spatial enclosure**
Popularly known as **treemaps**

Benefits

Provides a single view of an entire tree
Easier to spot large/small nodes

Problems

Difficult to accurately read structure / depth

Circle Packing Layout

Nodes are represented as sized circles.

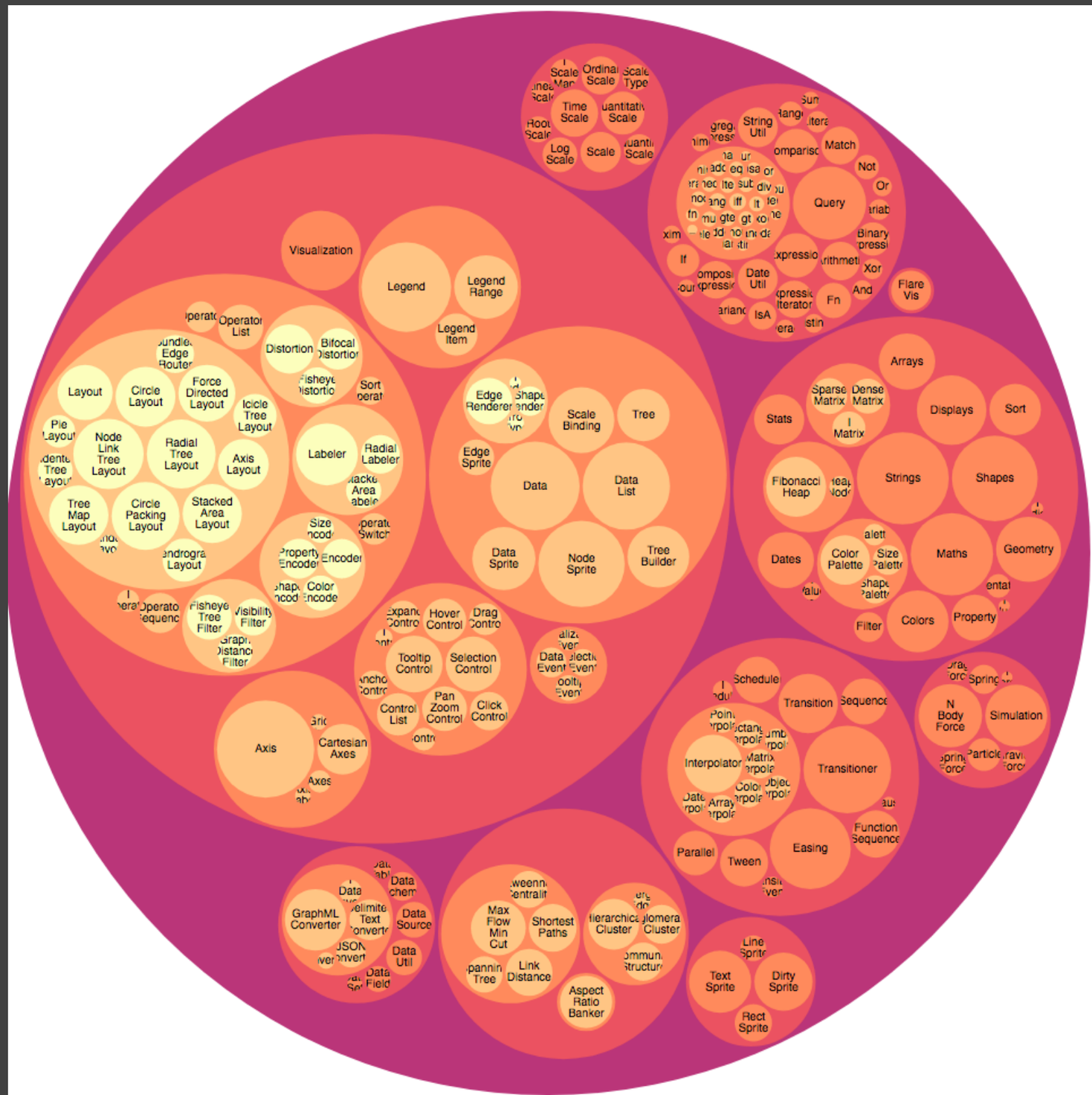
Nesting shows parent-child relationships.

Issues?

Inefficient use of space.

Labeling.

Parent size misleading?

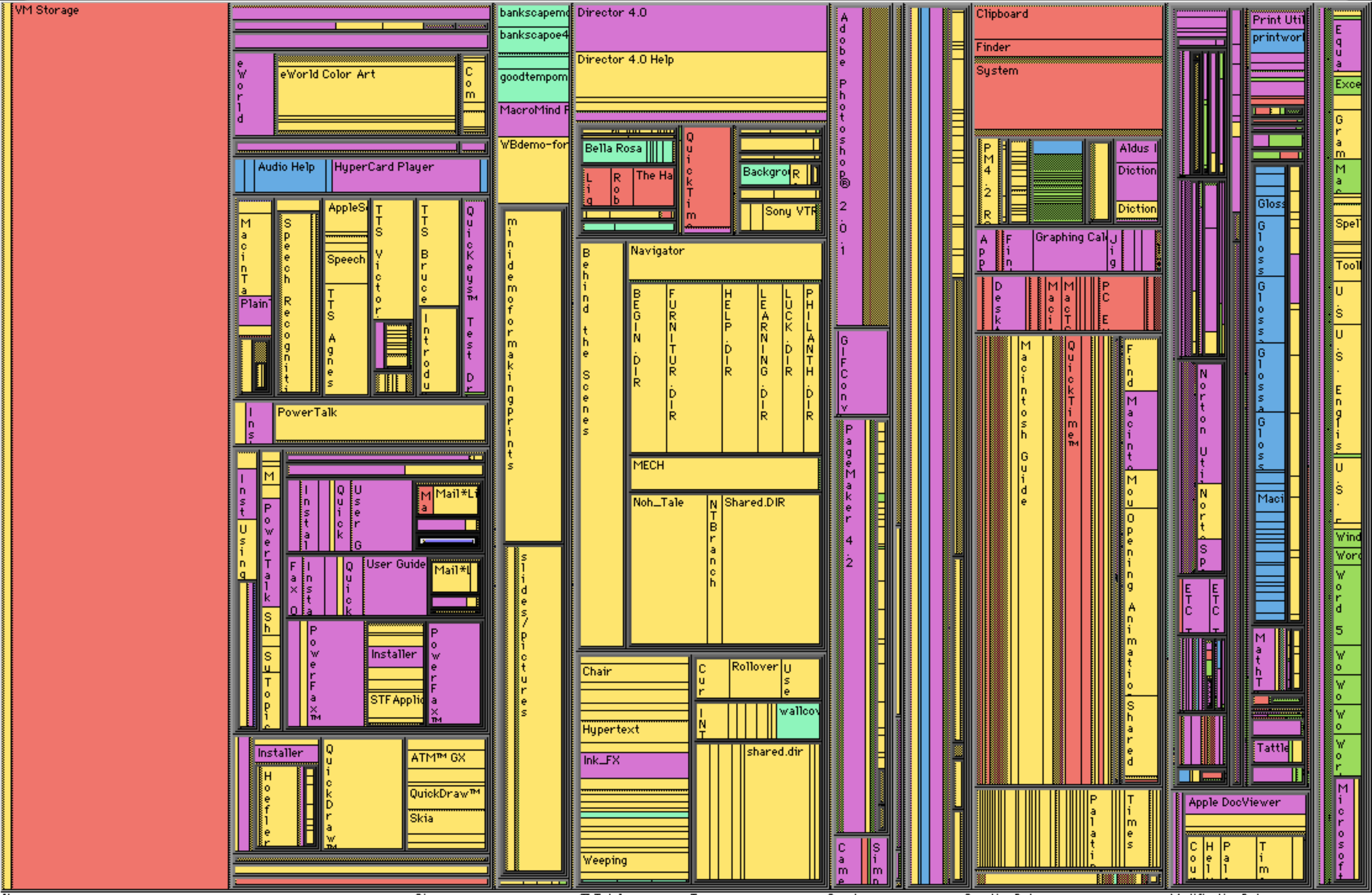


Treemaps

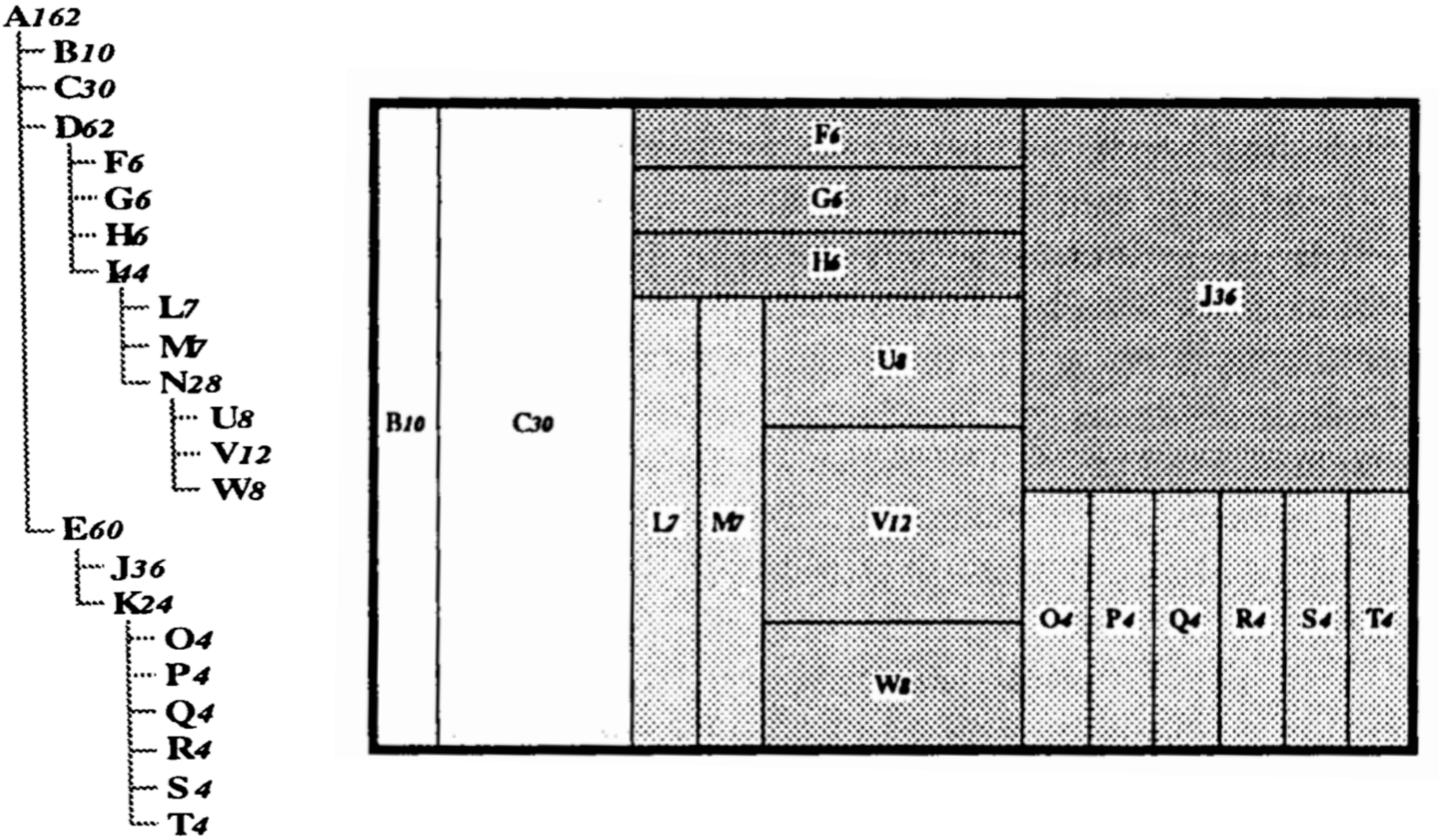
Hierarchy visualization that emphasizes values of nodes via area encoding.

Partition 2D space such that leaf nodes have sizes proportional to data values.

First layout algorithms proposed by Shneiderman et al. in 1990, with focus on showing file sizes on a hard drive.

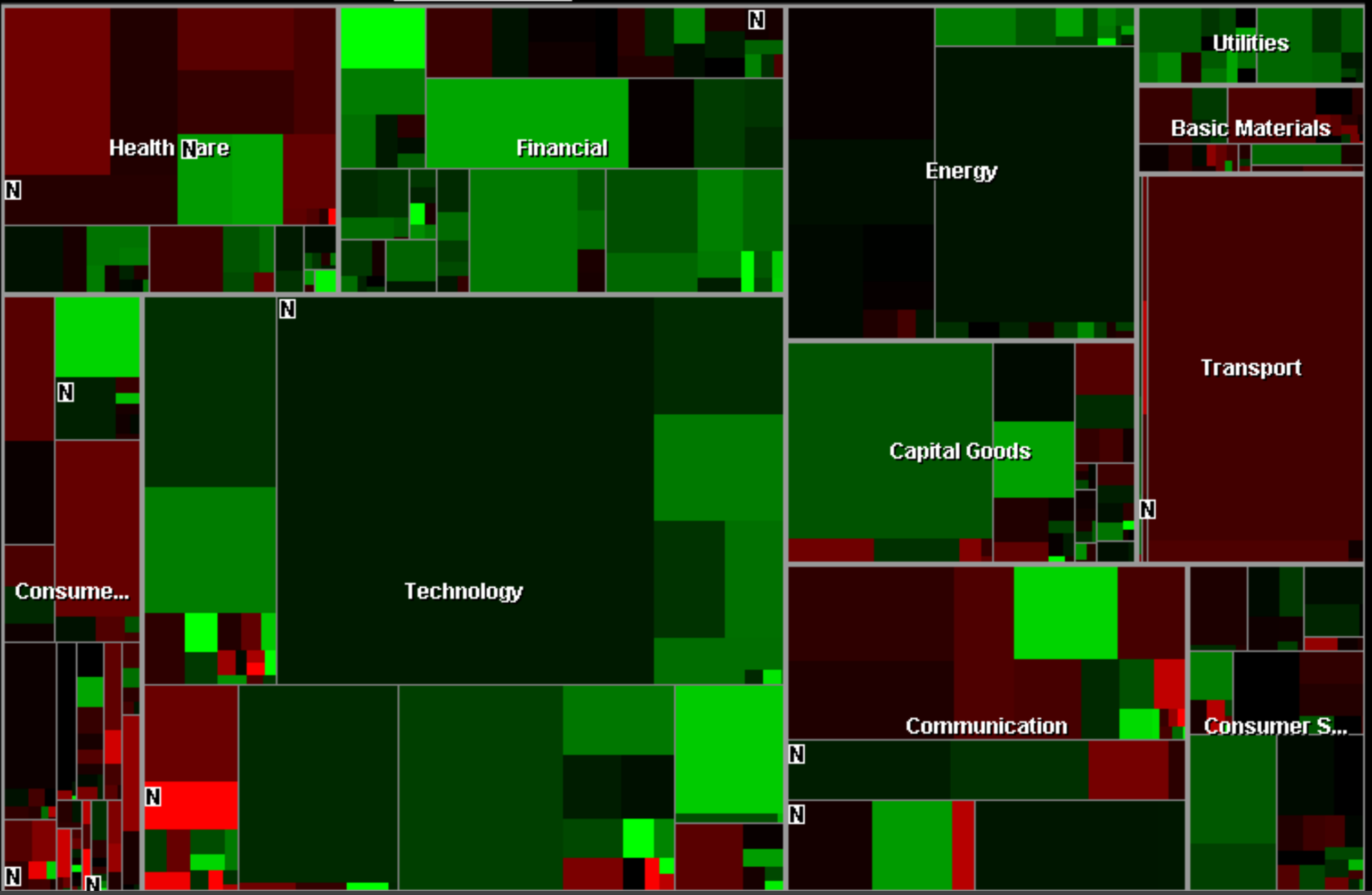


Slice & Dice layout: Alternate horizontal / vertical partitions.



Slice & Dice layout: Alternate horizontal / vertical partitions.

Controls | Instructions | Headline Icons | DJIA 11252.84 +60.21 +0.54% | Nasdaq 4070.59 +27.91 +0.69% | 5:33 pm Aug. 28

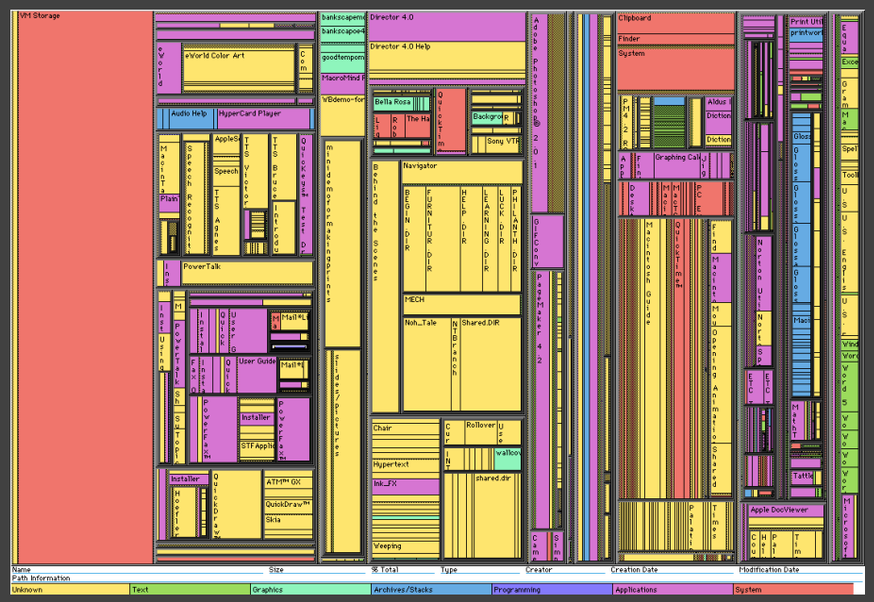


Squarifed layout: Try to produce square (1:1) aspect ratios

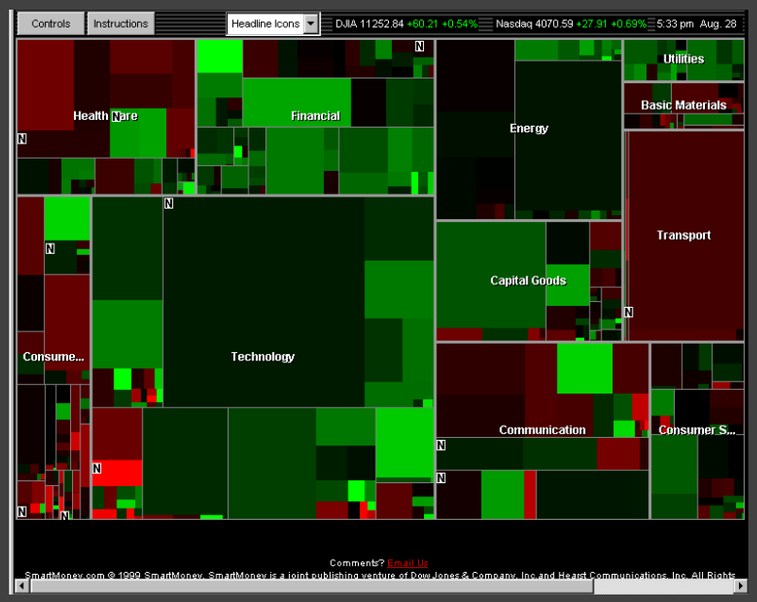
Squarified Treemaps [Bruls et al. ‘00]

Slice & Dice layout suffers from extreme aspect ratios. How might we do better?

Squarified layout: greedy optimization for objective of square rectangles. Slice/dice within siblings; alternate whenever ratio worsens.



vs.



Why Squares? [Bruls et al. '00]

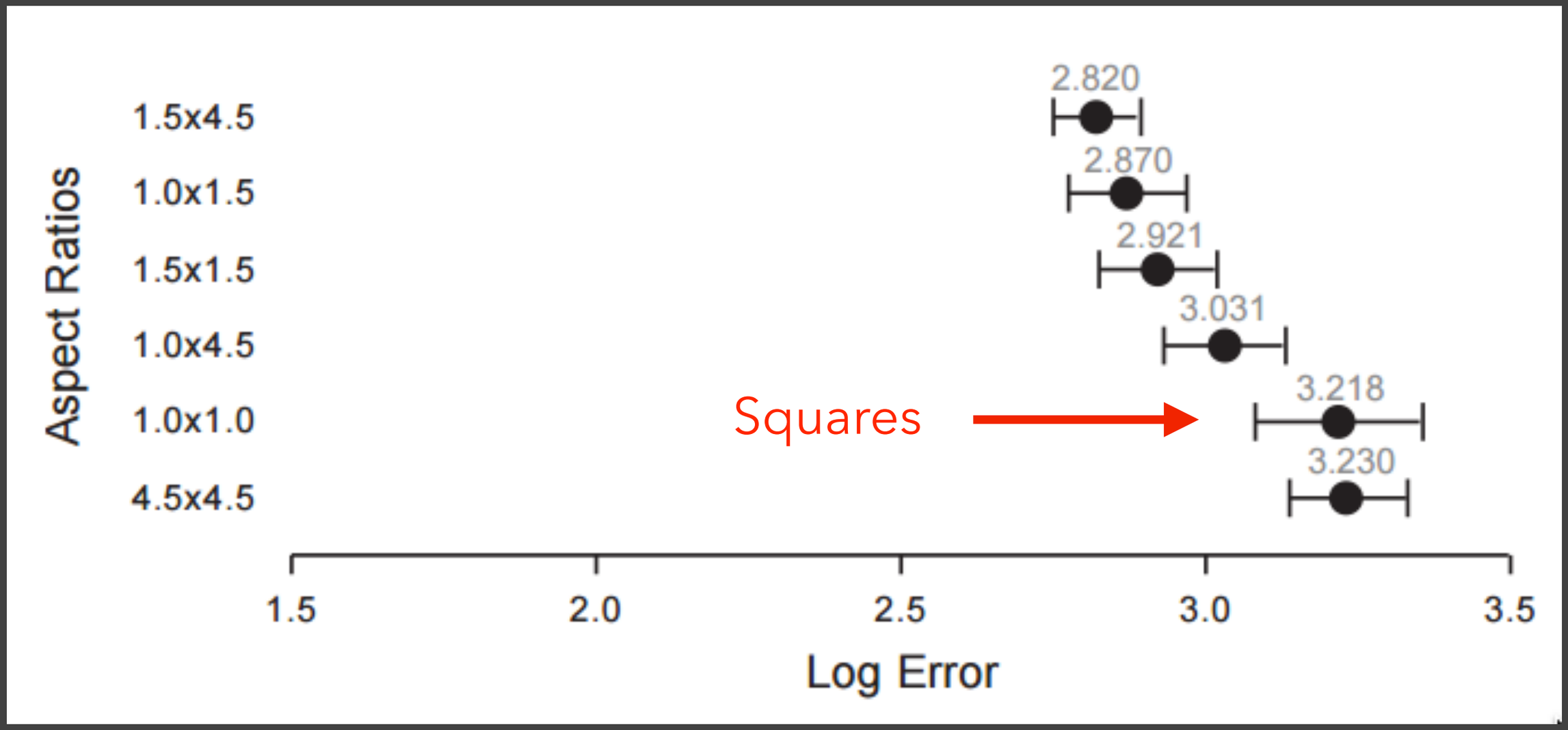
Posited Benefits of 1:1 Aspect Ratios

1. Minimize perimeter, reducing border ink.
   Mathematically true!

2. Easier to select with a mouse cursor.
   Validated by empirical research & Fitt's Law!

3. Similar aspect ratios are easier to compare.
   Seems intuitive, but is this true?

Comparison Error vs. Aspect Ratio



Study by Kong, Heer & Agrawala, InfoVis ‘10.
Comparison of squares has higher error!
“Squarify” works because it fails to meet its objective?

Why Squares? [Bruls et al. '00]

Posited Benefits of 1:1 Aspect Ratios

1. Minimize perimeter, reducing border ink.
   Mathematically true!

2. Easier to select with a mouse cursor.
   Validated by empirical research & Fitt's Law!

3. Similar aspect ratios are easier to compare.
   Seems intuitive, but is this true?

Why Squares? [Bruls et al. '00]

Posited Benefits of 1:1 Aspect Ratios

1. Minimize perimeter, reducing border ink.
   Mathematically true!

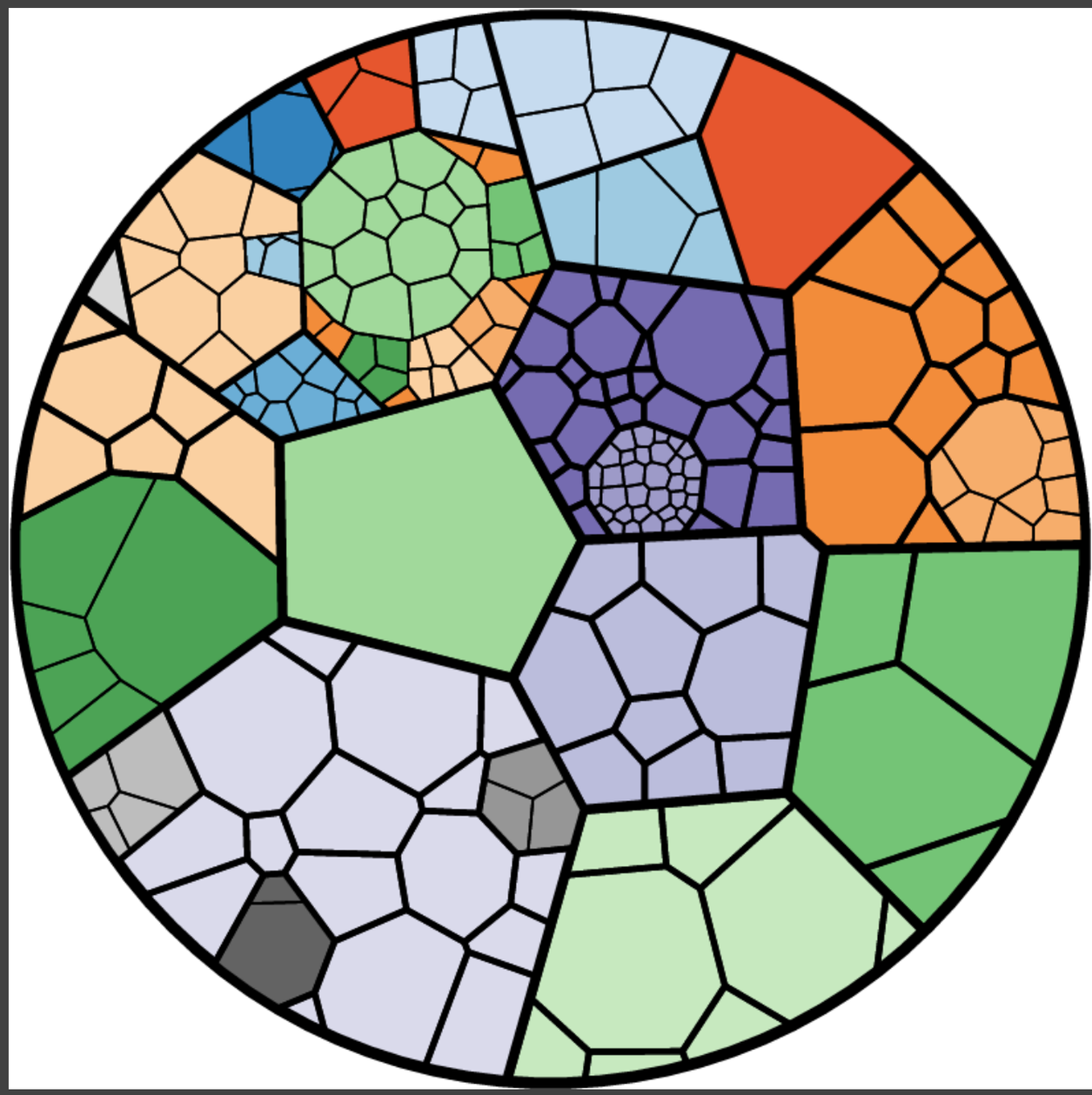
2. Easier to select with a mouse cursor.
   Validated by empirical research & Fitt's Law!

3. ~~Similar aspect ratios are easier to compare.~~
   Extreme ratios & squares-only more inaccurate.
   Balanced ratios better? Target golden ratio?

Voronoi Treemaps [Balzer et al. ’05]

Instead of rectangles, create treemaps with arbitrary polygonal shapes and boundary.

Use iterative, weighted Voronoi tessellations to achieve cells with value-proportional areas.



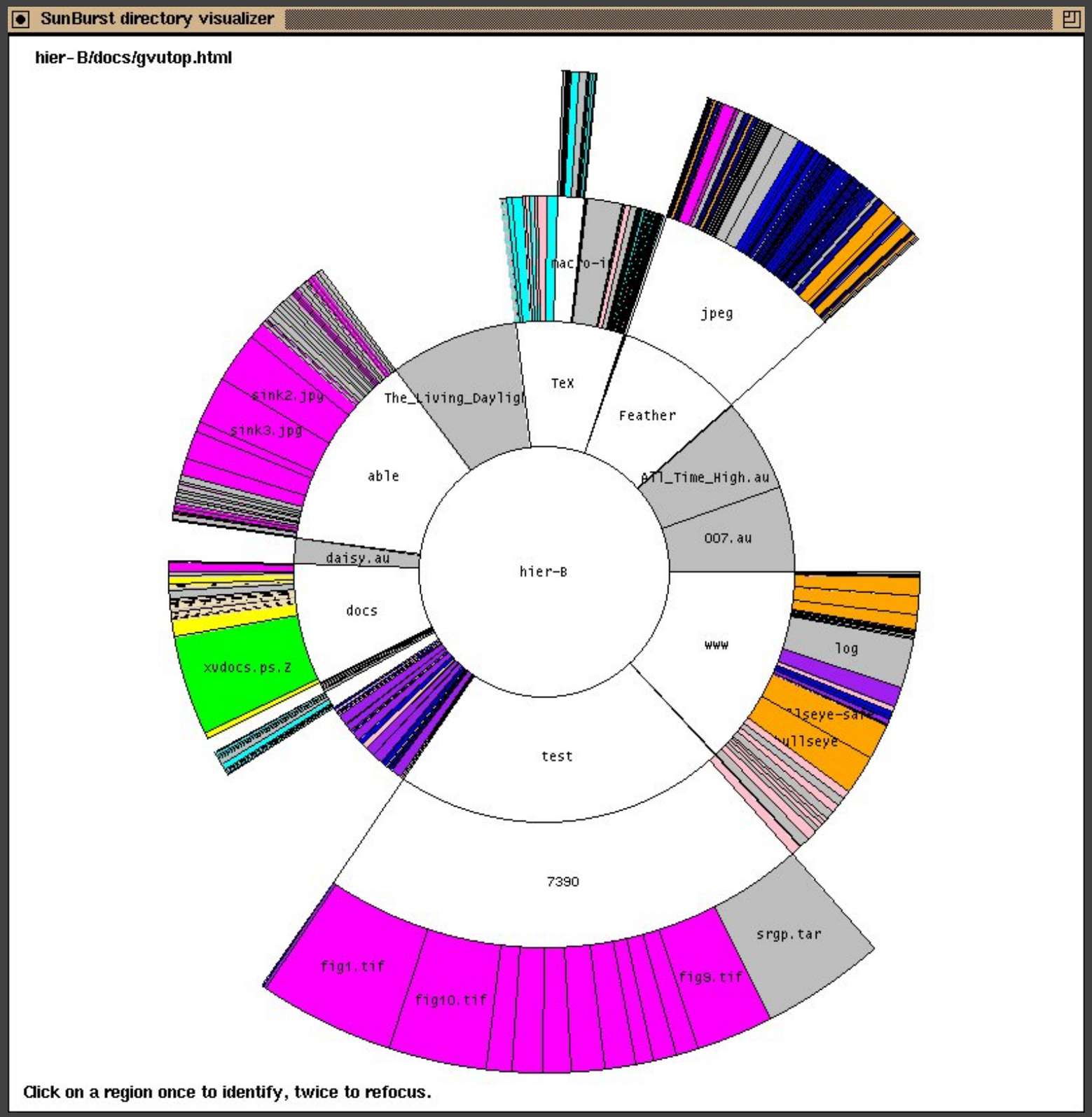
Layered Diagrams

Signify tree structure using:
- Layering
- Adjacency
- Alignment

Involves recursive sub-division of space.

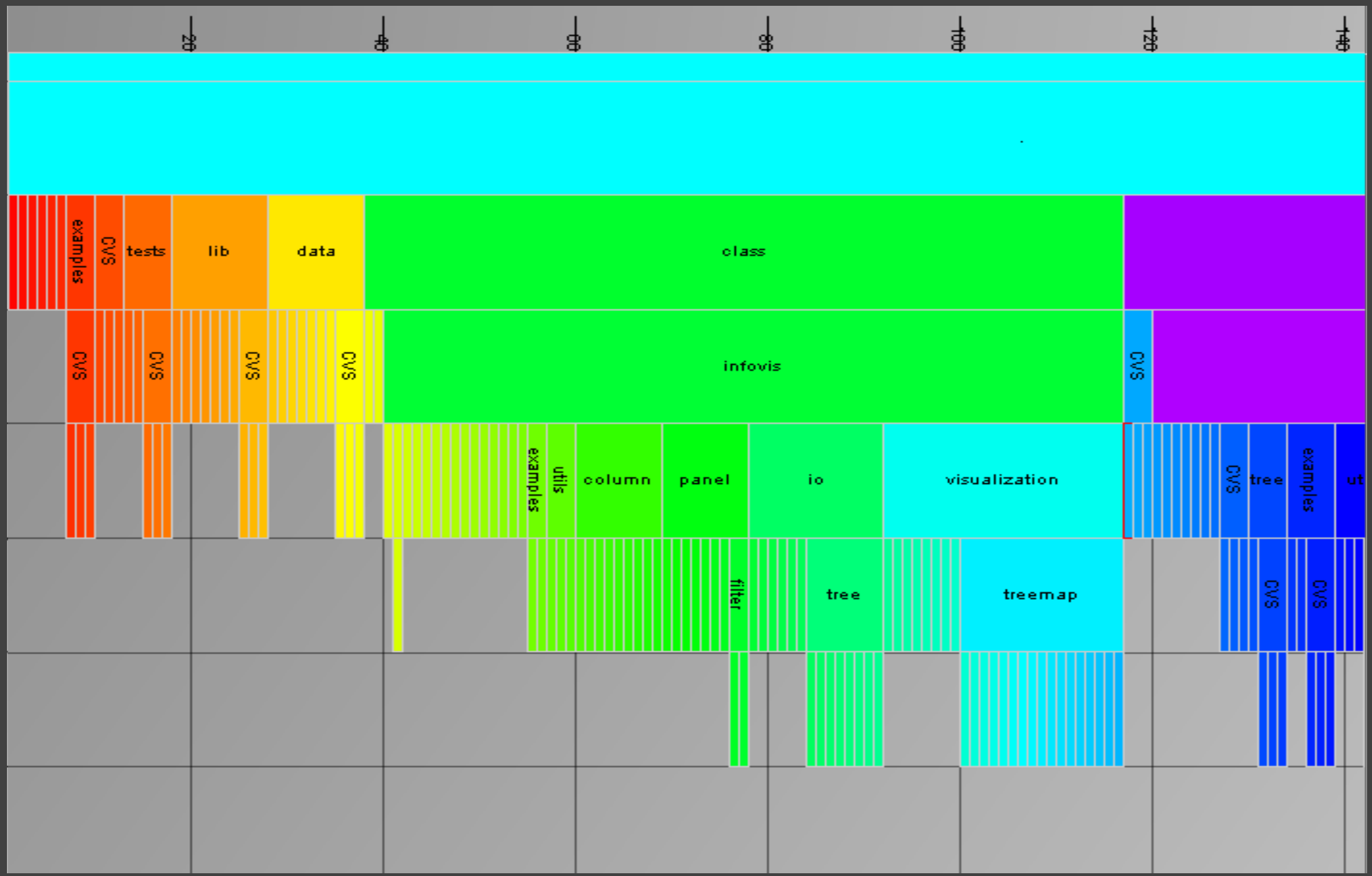
Leaf nodes may be sized by value, parent size visualizes sum of descendant leaf values.

“Sunburst” Trees: Polar Partition



[Stasko et al., IJHCS 2000]

Icicle Trees: Cartesian Partition



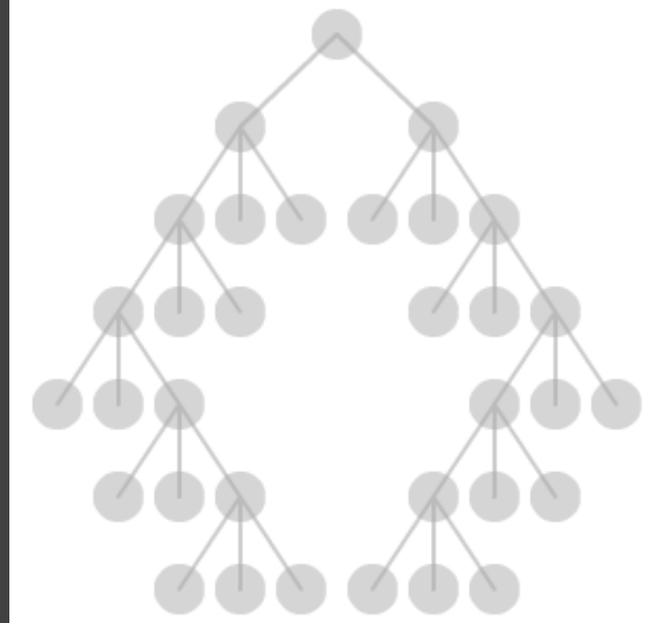
Node-Link Diagrams

Nodes are distributed in space, connected by straight or curved lines

Typical approach is to use 2D space to break apart breadth and depth

Often space is used to communicate hierarchical orientation (e.g., towards authority or generality)

Trees



Reingold-Tilford Layout

Breadth along one dimension, depth along the other.

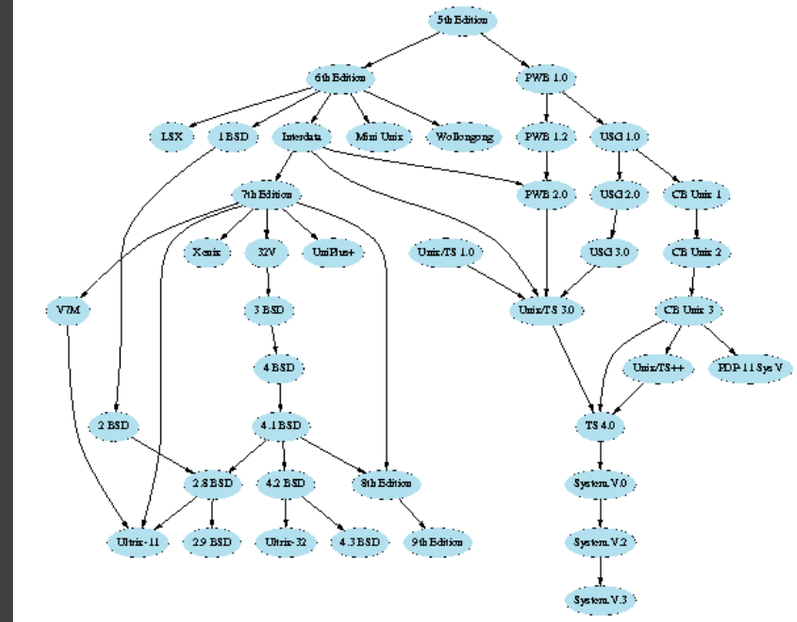
No edge crossings.

Identical isomorphic subtrees.

Compact, symmetric layout.

Available as part of `d3.tree`

Graphs



Layered/Sugiyama Layout

Hierarchical layering based on descent.

Assumes a directionality/flow.

Minimize edge crossings.

Available as part of GraphVis!

Cycles may mislead, long edges can be difficult to perceive.

Force-Directed Layout

Nodes = charged particles $F = q_i * q_j / d_{ij}^2$

with air resistance $F = -b * v_i$

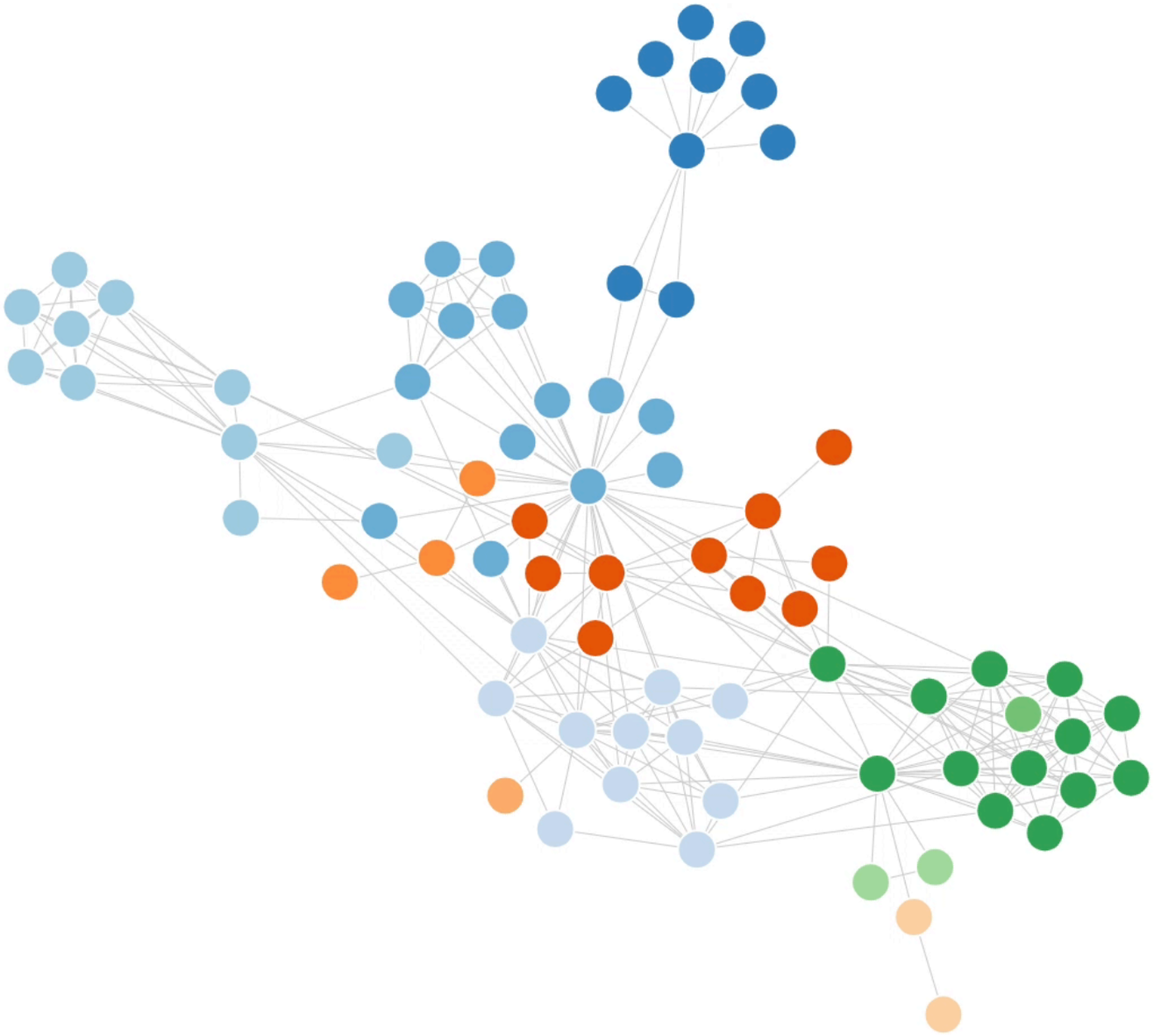
Edges = springs $F = k * (L - d_{ij})$

At each timestep, calculate forces acting on nodes.
Integrate for updated velocities and positions.

D3's force layout uses **velocity Verlet** integration.
Assume uniform mass ***m*** and timestep ***Δt***:
$F = ma \rightarrow F = a \rightarrow F = \Delta v / \Delta t \rightarrow F = \Delta v$
Forces simplify to velocity offsets!

Force Directed Layout Example

Network layout by force-directed placement. Uses Vega's force transform to simulate physical forces such as charge repulsion and edge constraint. Drag nodes to reposition them.



nodeRadius 12
nodeCharge -36
linkDistance 30
static

View in Online Vega Editor

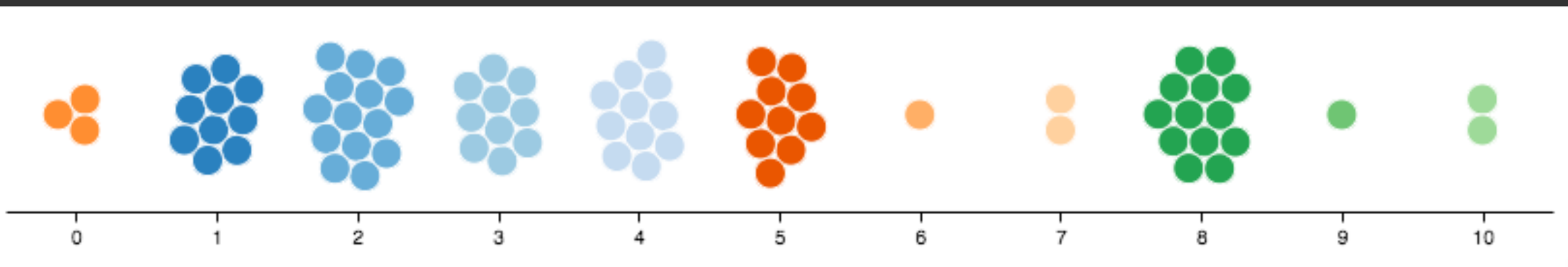
Customized Force Layouts

Different forces can be composed to create an expressive space of custom layouts.

A **beeswarm plot** can be made by combining:

Attractive **X** and **Y** forces to draw nodes of a certain category to a desired point

Collide force to detect collision & remove overlap



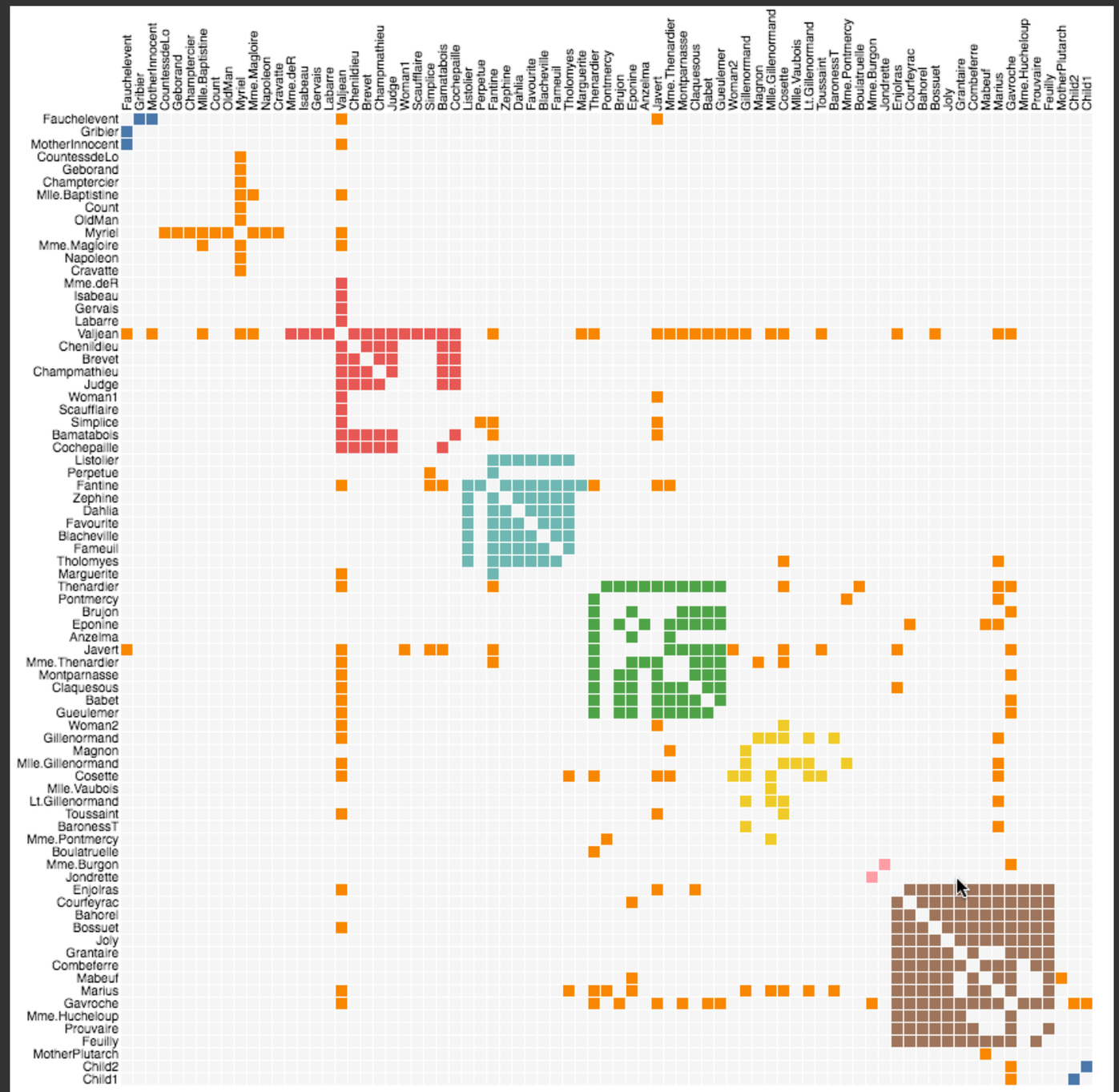
Adjacency Matrices

More scaleable visual form.

Predictable and stable (physical size, interactive reordering, adding/removing elements).

Can require training to read and identify patterns (e.g., clusters and cliques).

Topological structure is visualized indirectly.



Text Visualization

Text as Data

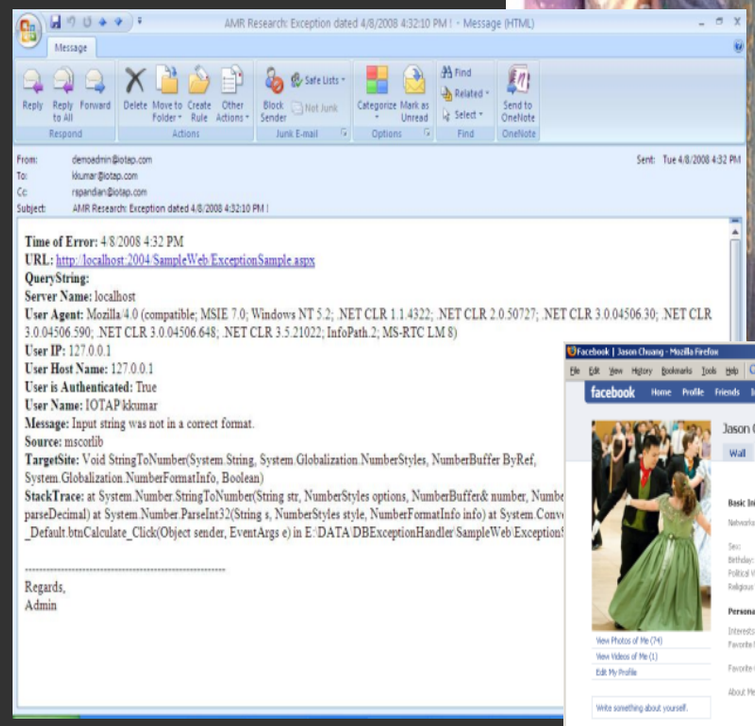
Documents

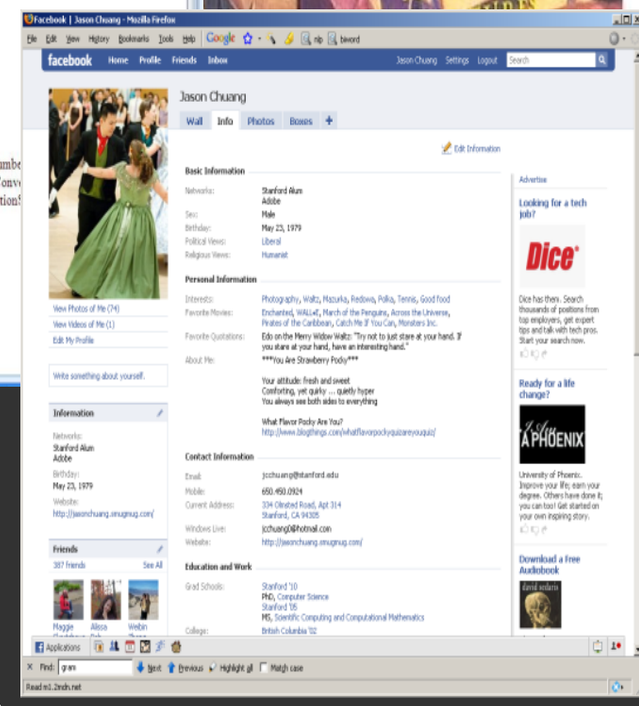
Articles, books and novels

E-mails, web pages, blogs

Tags, comments

Computer programs, logs





Collections of Documents

Messages (e-mail, blogs, tags, comments)

Social networks (personal profiles)

Academic collaborations (publications)

Why visualize text?

Why Visualize Text?

Understanding – get the "gist" of a document

Grouping – cluster for overview or classification

Comparison – compare document collections, or inspect evolution of collection over time

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

Example: Health Care Reform

Background

Initiatives by President Clinton
Overhaul by President Obama

Text Data

News articles
Speech transcriptions
Legal documents

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

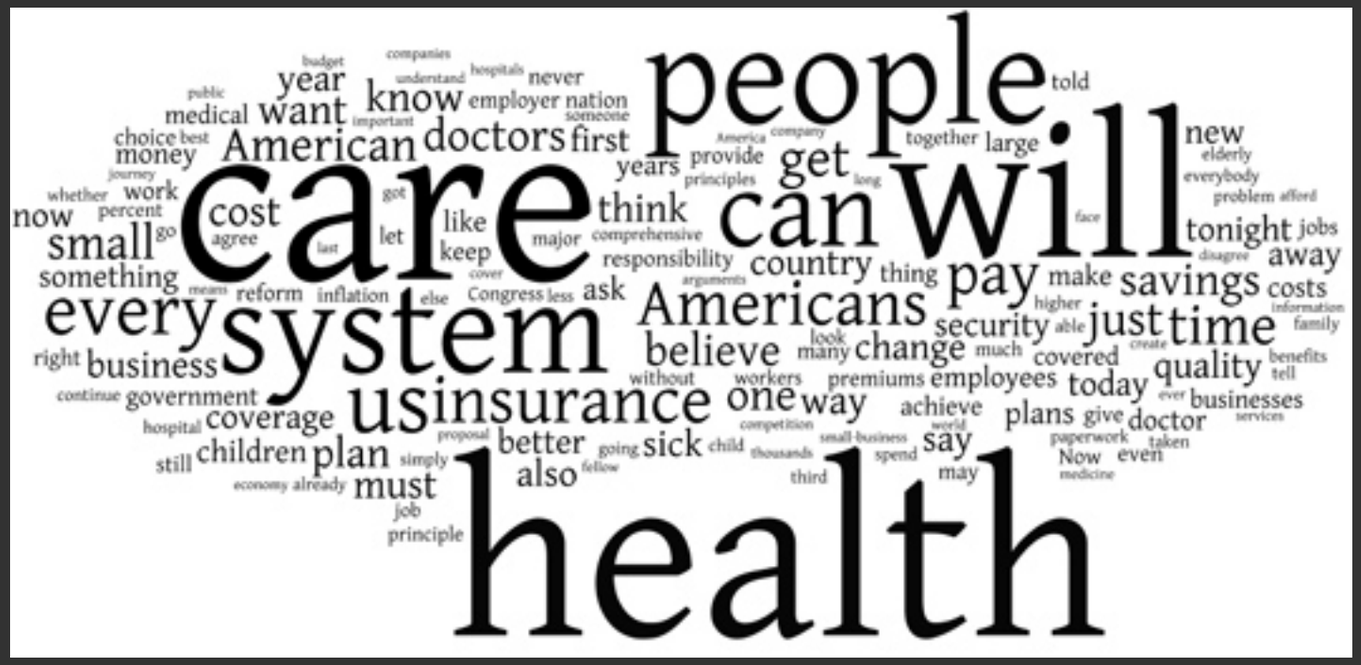
Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

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



Bill Clinton, 1993

Barack Obama, 2009

Charles Boustany (Rep. LA, official Republican response), 2009

Word Tree: Word Sequences

Gulfs of Evaluation

Many text visualizations do not represent the text directly. They represent the output of a **language model** (word counts, word sequences, etc.).

- Can you interpret the visualization? How well does it convey the properties of the model?
- Do you trust the model? How does the model enable us to reason about the text?

Text as Data

Words as nominal data?

High dimensional (10,000+)

More than equality tests

Words have meanings and relations

- Correlations: *Hong Kong, Puget Sound, Bay Area*
- Order: *April, February, January, June, March, May*
- Membership: *Tennis, Running, Swimming, Hiking, Piano*
- Hierarchy, antonyms & synonyms, entities, …

Text Processing Pipeline

1. Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#huskies, @UW, OMG!!!!!!!!*

Entities? *Washington State, O'Connor, U.S.A.*

Text Processing Pipeline

1. Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#huskies, @UW, OMG!!!!!!!!*

Entities? *Washington State, O'Connor, U.S.A.*

2. Stemming

Group together different forms of a word.

Porter stemmer? *visualization(s), visualize(s), visually* -> *visual*

Lemmatization? *goes, went, gone* -> *go*

Text Processing Pipeline

1. Tokenization

   Segment text into terms.

   Remove stop words? *a, an, the, of, to, be*

   Numbers and symbols? *#huskies, @UW, OMG!!!!!!!!*

   Entities? *Washington State, O'Connor, U.S.A.*

2. Stemming

   Group together different forms of a word.

   Porter stemmer? *visualization(s), visualize(s), visually* -> *visual*

   Lemmatization? *goes, went, gone* -> *go*

3. 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

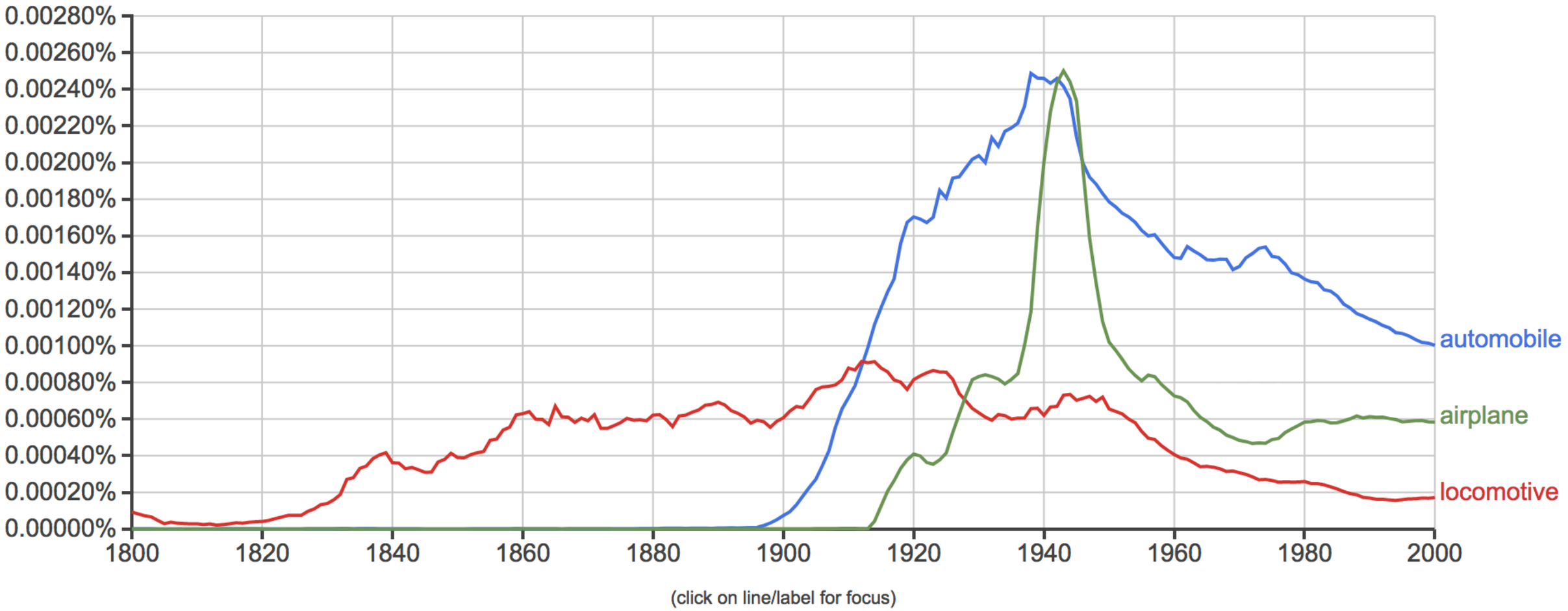
WordCounts (Harris '04)

http://wordcount.org

Google Books Ngram Viewer

Graph these comma-separated phrases: automobile,locomotive,airplane case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books



https://books.google.com/ngrams/

Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech

Creator: Anonymous
Tags:

Edit Language Font Layout Color

Data file: Sarah Palin speaks at the Republican National Convention, 9/3/2008 Data source: SFGate / AP This data set has not yet been rated

Tag Clouds

Strengths

Can help with gisting and initial query formation.

Weaknesses

Sub-optimal visual encoding (size vs. position)

Inaccurate size encoding (long words are bigger)

May not facilitate comparison (unstable layout)

Term frequency may not be meaningful

Does not show the structure of the text

Keyword Weighting

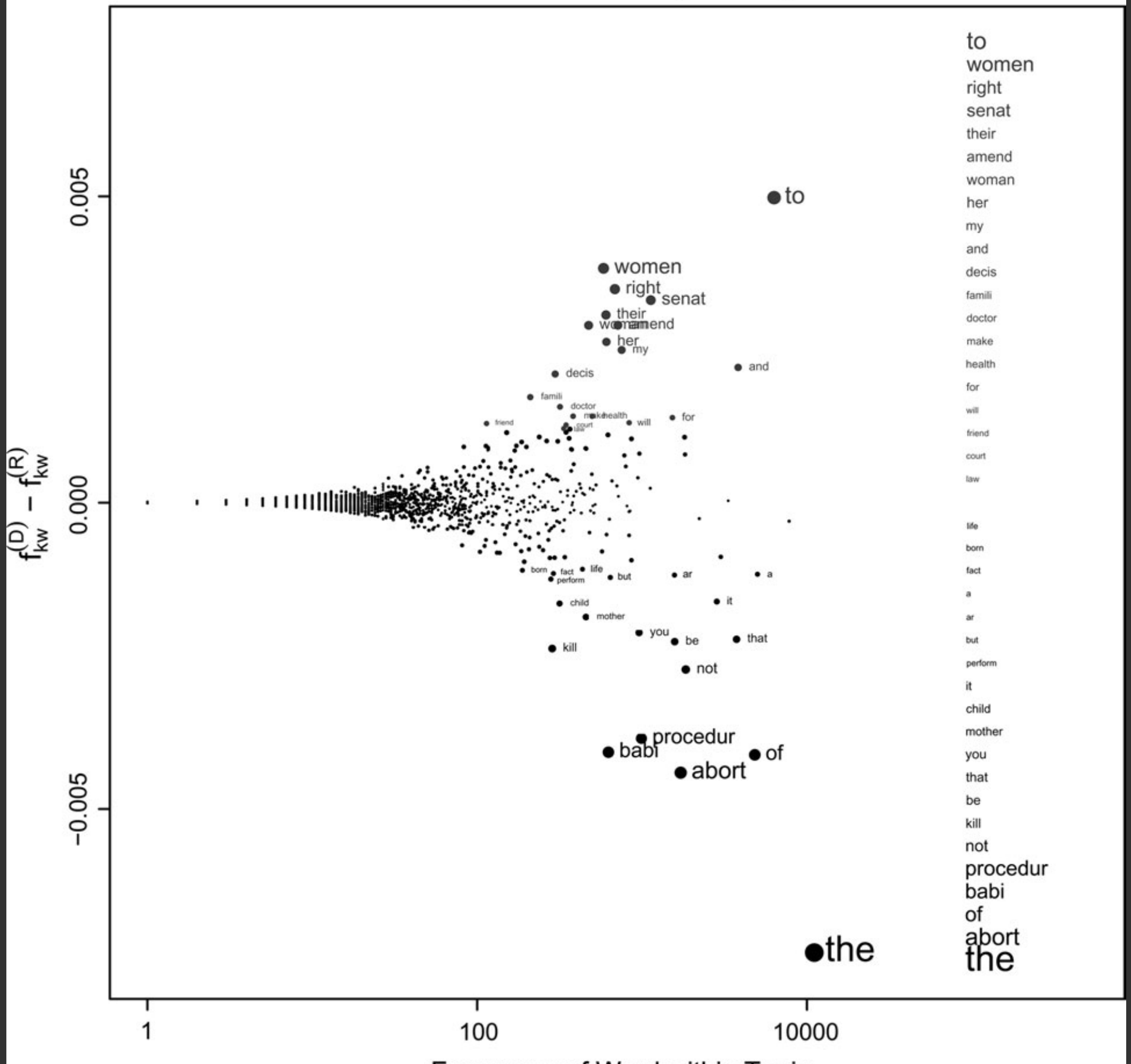
Term Frequency

tf_{td} = count(t) in d

Can normalize to show proportion: $tf_{td} / \Sigma_t tf_{td}$

Can take log frequency: $\log(1 + tf_{td})$

Partisan Words, 106th Congress, Abortion
(Difference of Proportions)



Keyword Weighting

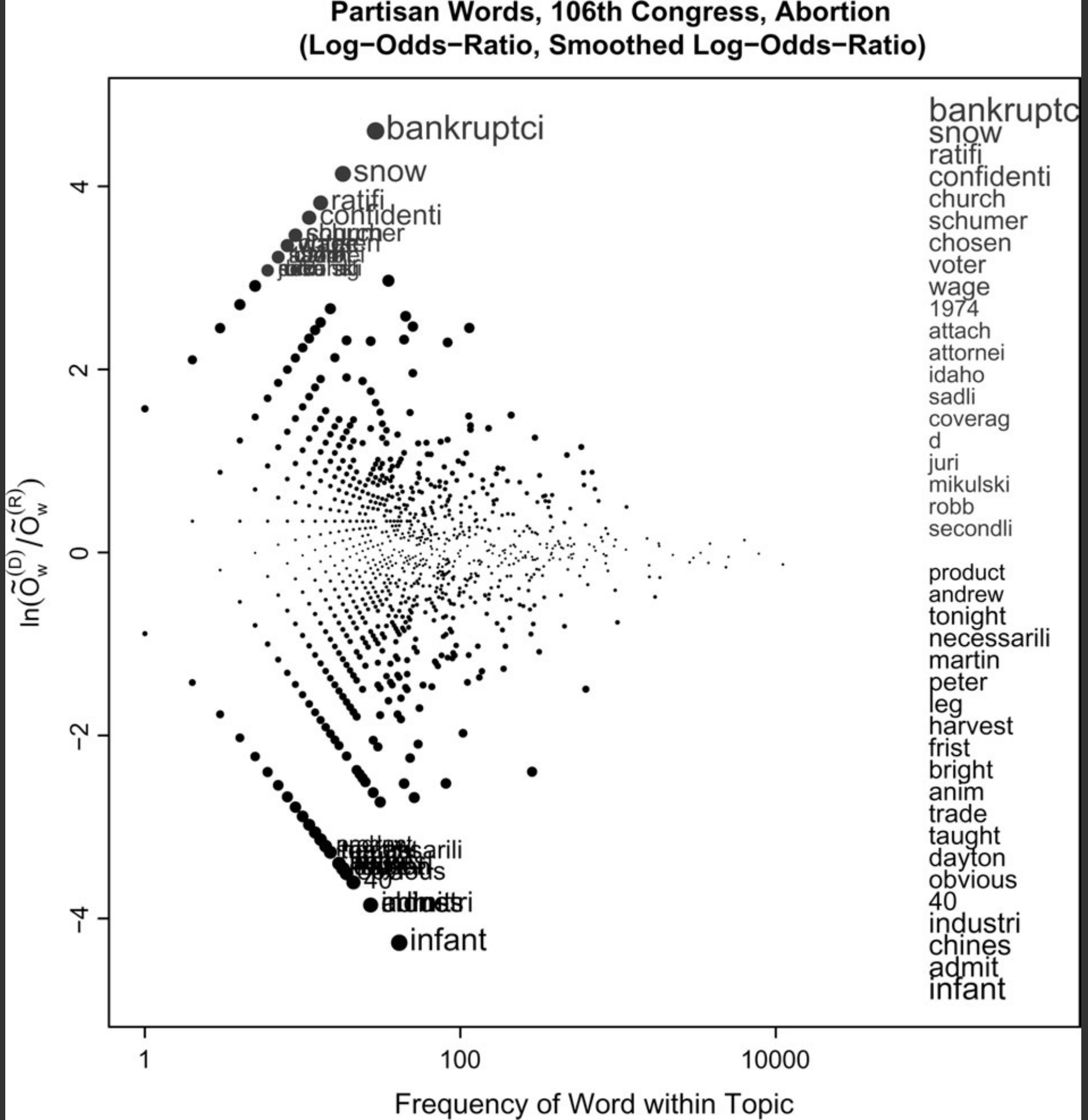
Term Frequency

tf_{td} = count(t) in d

TF.IDF: Term Freq by Inverse Document Freq

$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

df_t = # docs containing t; N = # of docs


Partisan Words, 106th Congress, Abortion
(Log−Odds−Ratio, Smoothed Log−Odds−Ratio)
bankruptci
snow
ratifi
confidenti
schumer
infant
bankruptc
snow
ratifi
confidenti
church
schumer
chosen
voter
wage
1974
attach
attornei
idaho
sadli
coverag
d
juri
mikulski
robb
secondli
product
andrew
tonight
necessarili
martin
peter
leg
harvest
frist
bright
anim
trade
taught
dayton
obvious
40
industri
chines
admit
infant
4
2
0
−2
−4
1
100
10000
Frequency of Word within Topic
$\ln(\tilde{O}_w^{(D)}/\tilde{O}_w^{(R)})$

Limitations of Freq. Statistics

Typically focus on unigrams (single terms)

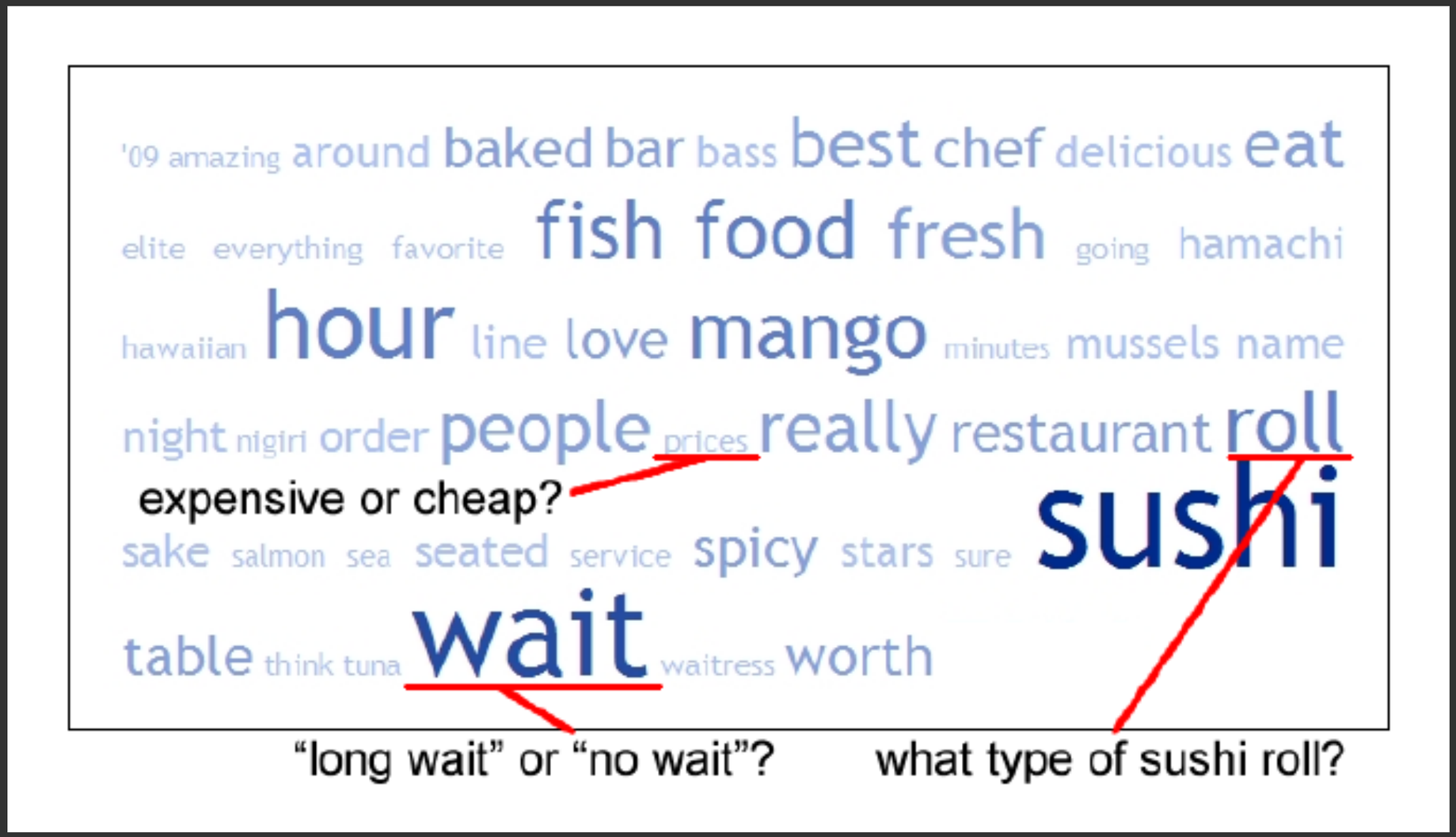
Often favors frequent (TF) or rare (IDF) terms

- Not clear that these provide best description

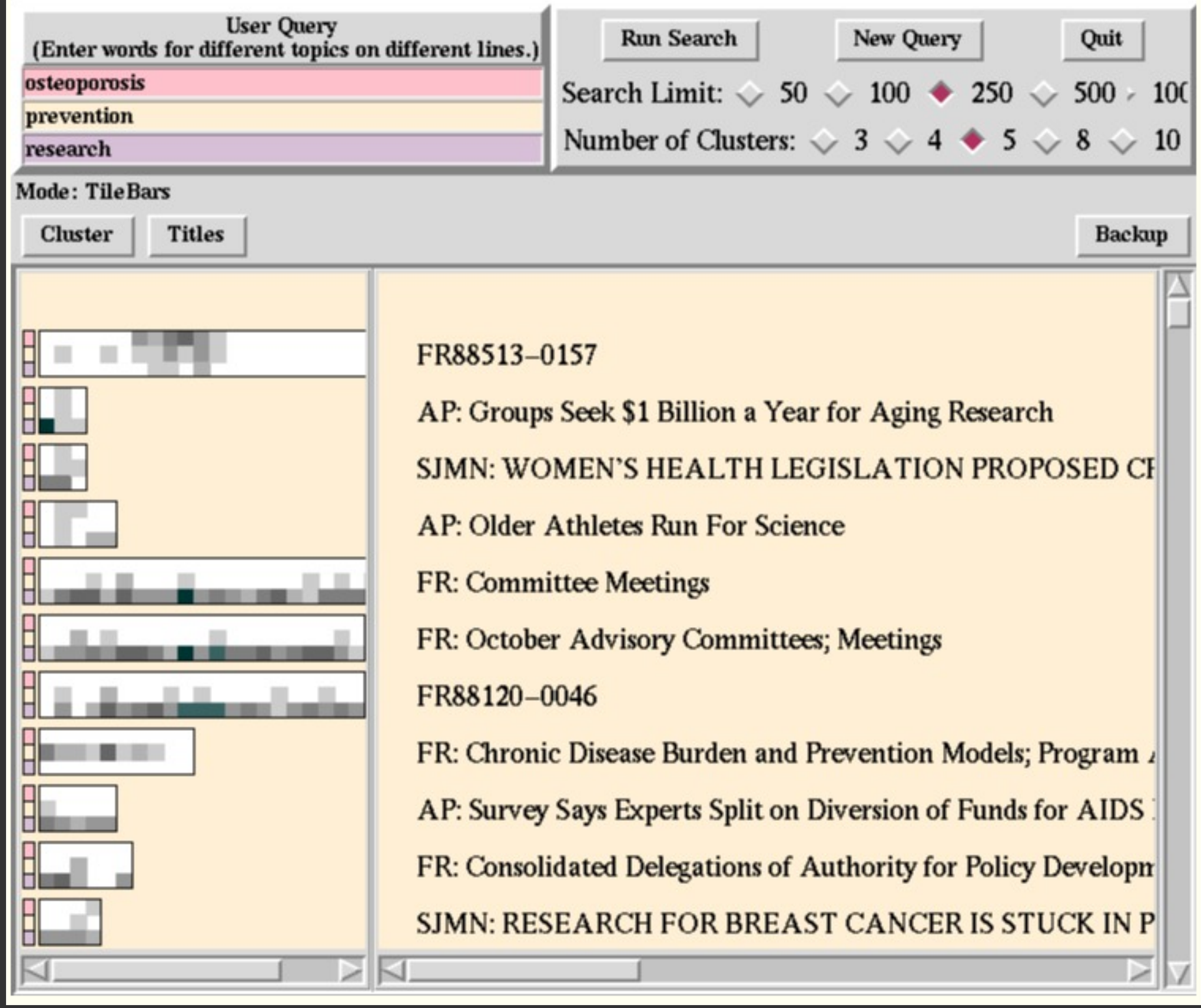
A “bag of words” ignores information

- Grammar / part-of-speech
- Position within document
- Recognizable entities

Yelp Review Spotlight (Yatani 2011)



Yelp Review Spotlight (Yatani 2011)



TileBars [Hearst]

THE WORDS THAT WERE USED

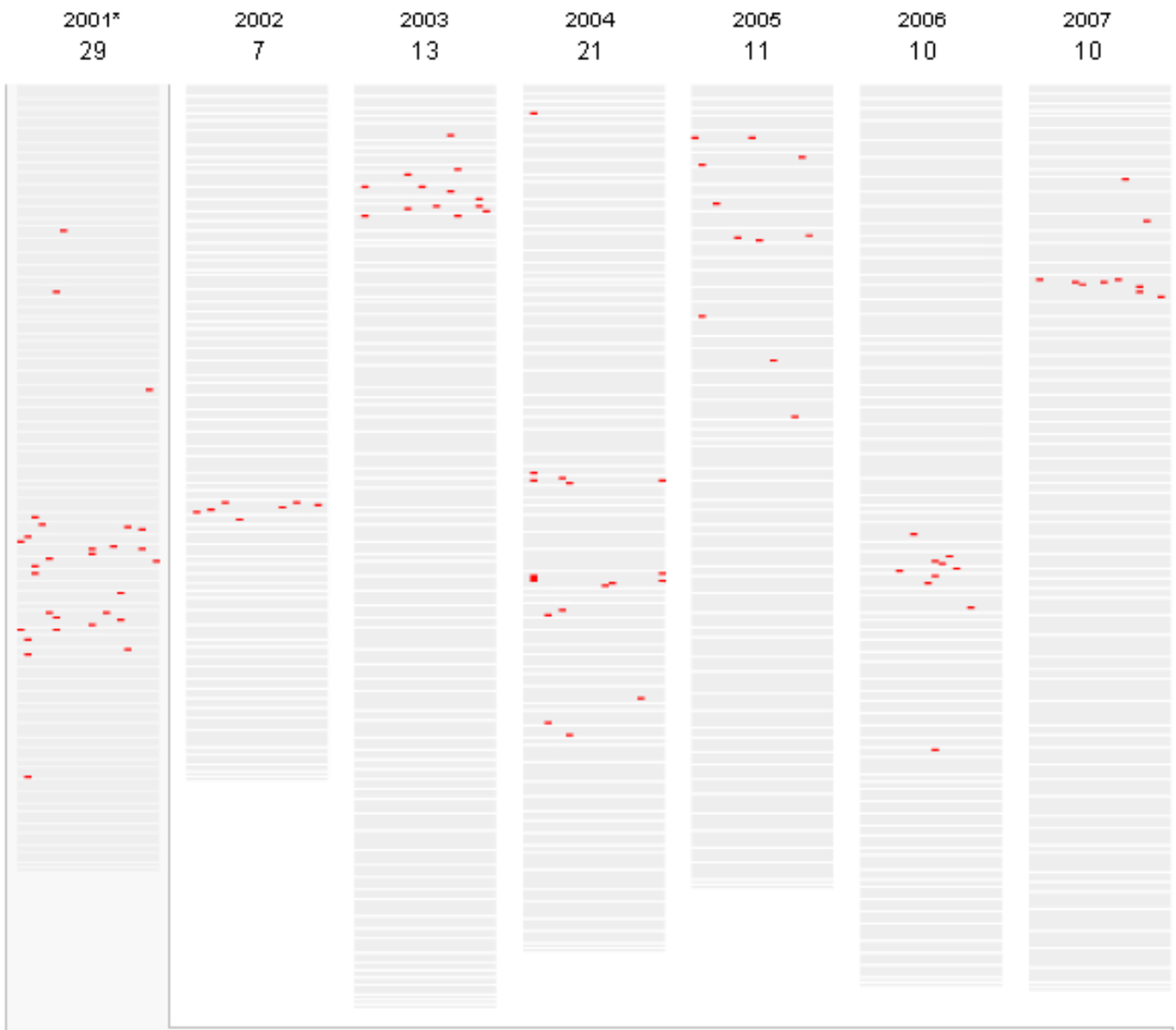
READ 2007 SPEECH | FEEDBACK

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

Search | or choose a word here.

Use of the phrase "Tax" in past State of the Union Addresses



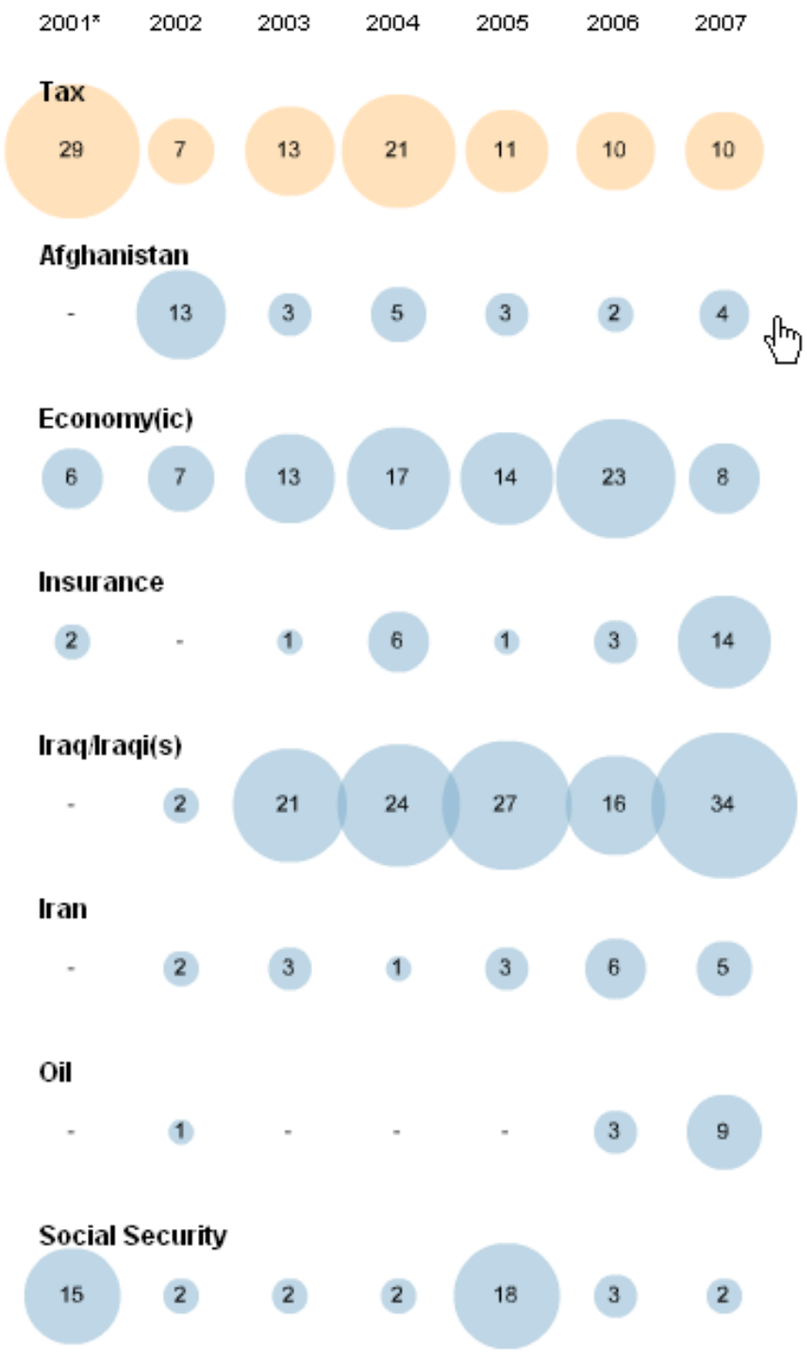
The word in context

Next Instance of 'Tax'

I believe in local control of schools. We should not, and we will not, run public schools from Washington, D.C. Yet when the federal government spends **TAX** dollars, we must insist on results. Children should be tested on basic reading and math skills every year between grades three and eight. Measuring is the only way to know whether all our children are learning. And I want to know, because I refuse to leave any child behind in America.

-- 2001 (Paragraph 14 of 73)

Compared with other words



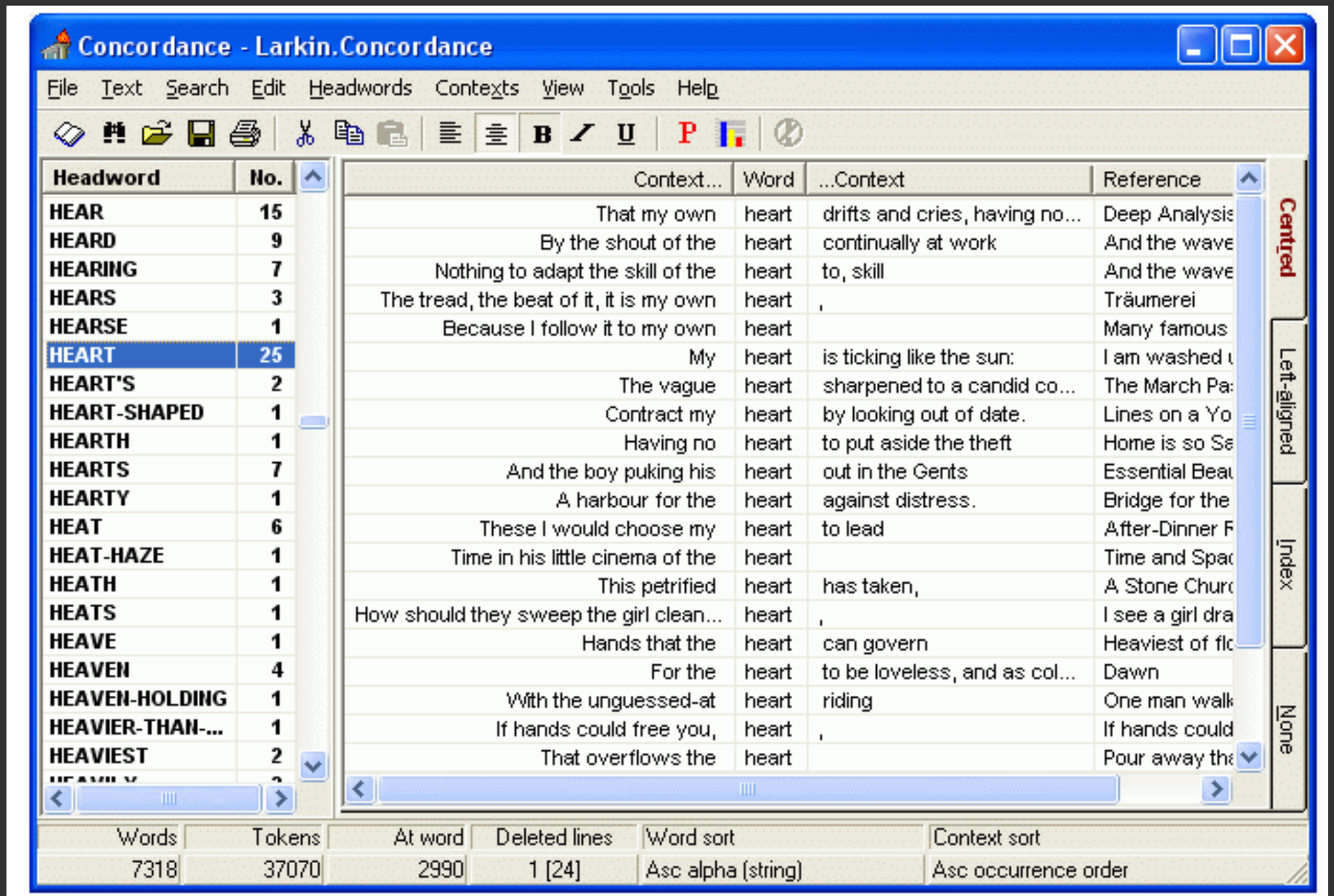
* As a newly elected president, Mr. Bush did not deliver a formal State of the Union address in 2001. His Feb. 27 speech to a joint session of Congress was analogous to the State of the Union, but without the title.

Ben Werschkul/The New York Times

New York Times

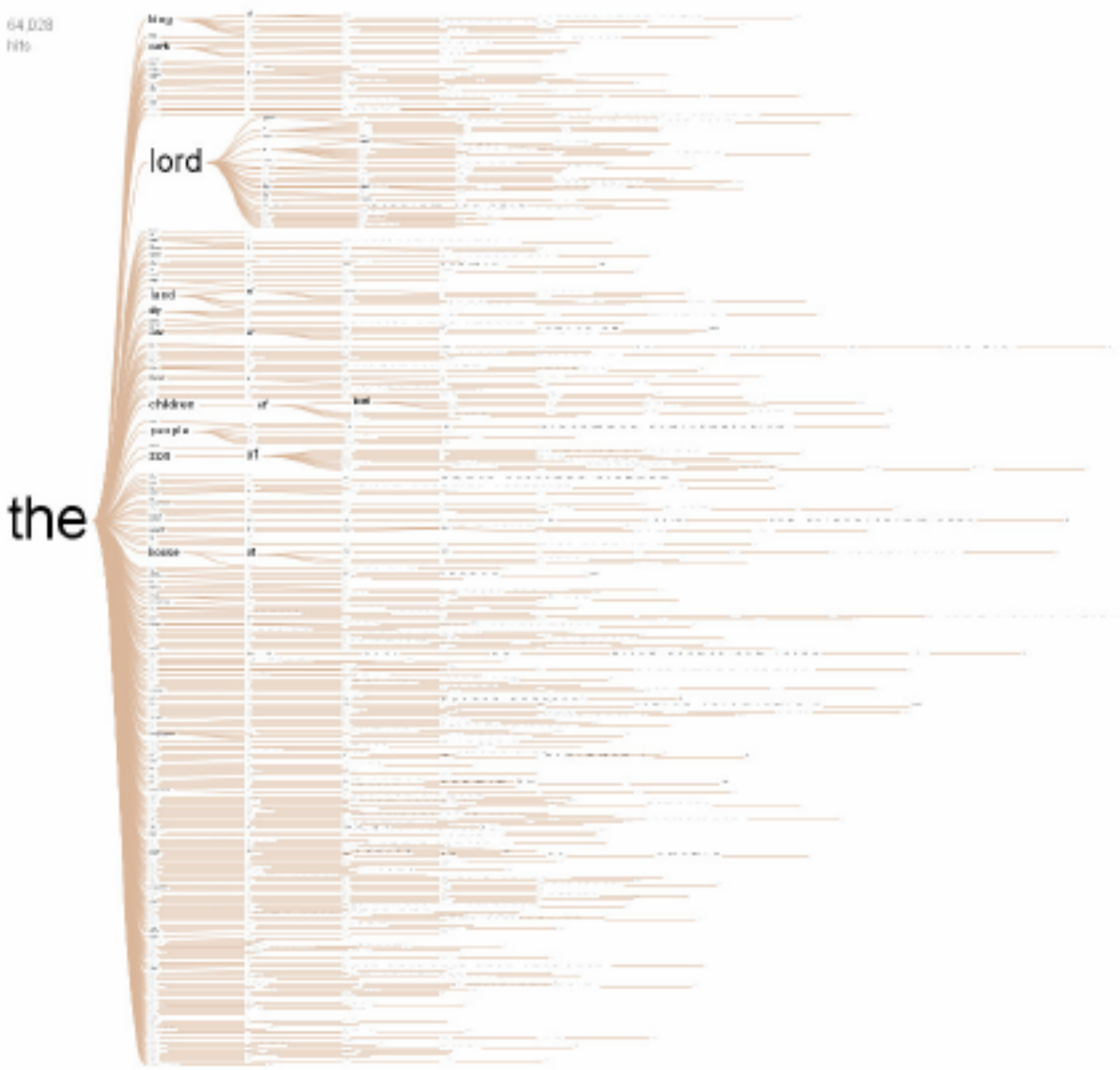
Concordance

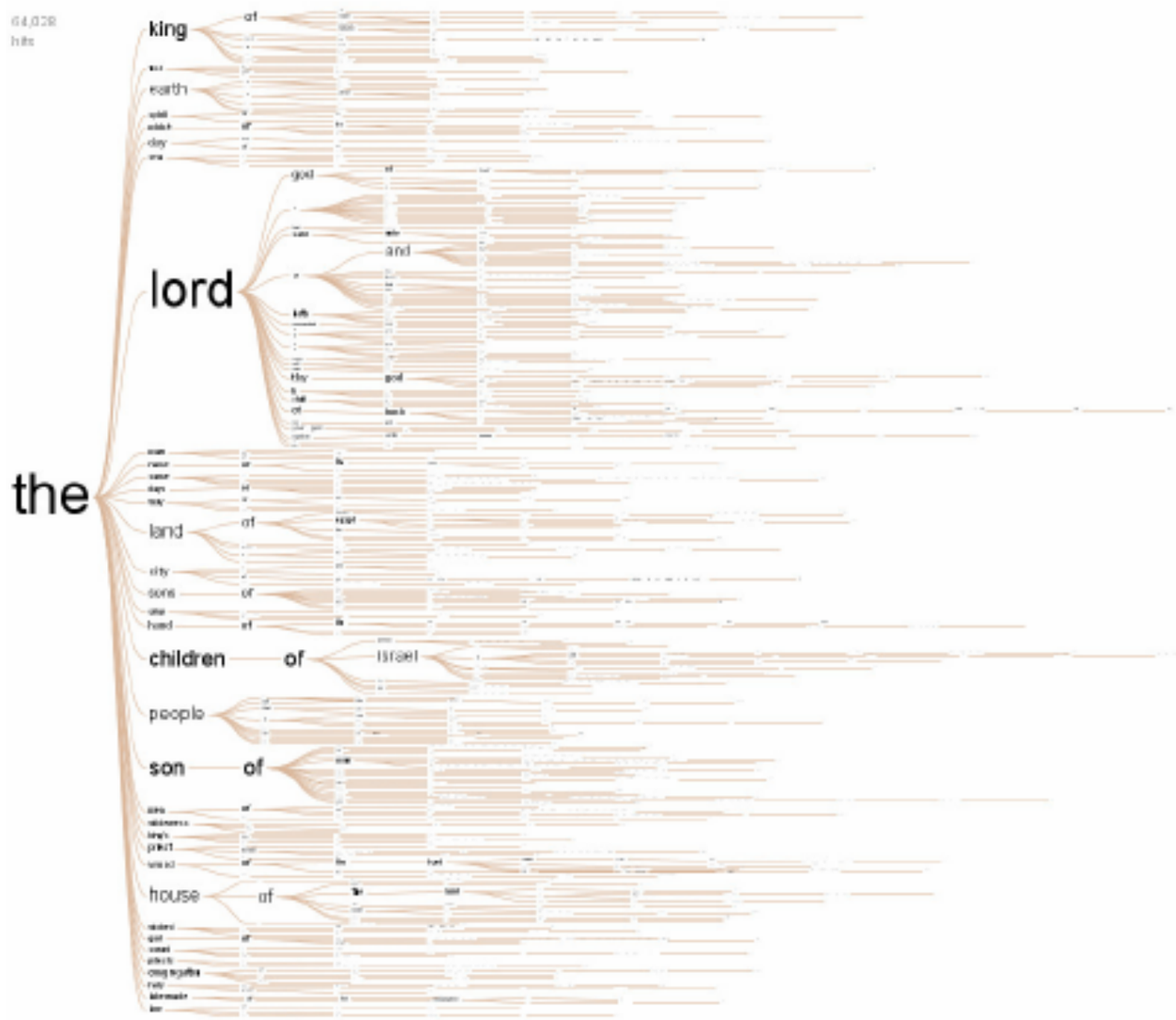
What is the common local context of a term?



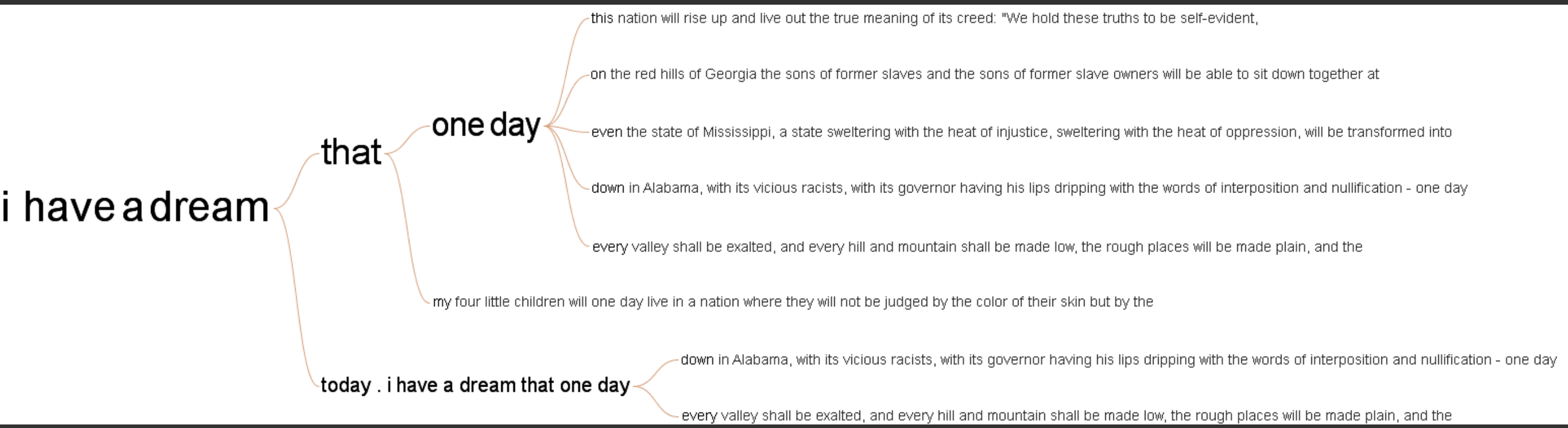
Word Tree [Wattenberg et al.]

Filter Infrequent Runs





Recurrent Themes in Speeches



Visualizations : Word tree / Alberto Gonzales

Creator: Martin Wattenberg
Tags:

explore
visualizations
data sets
comments
topic hubs

participate
create visualization
upload data set
create topic hub
register

learn more
quick start
visualization types
data format & style
about Many Eyes
FAQ
blog

contact Us
contact
report a bug

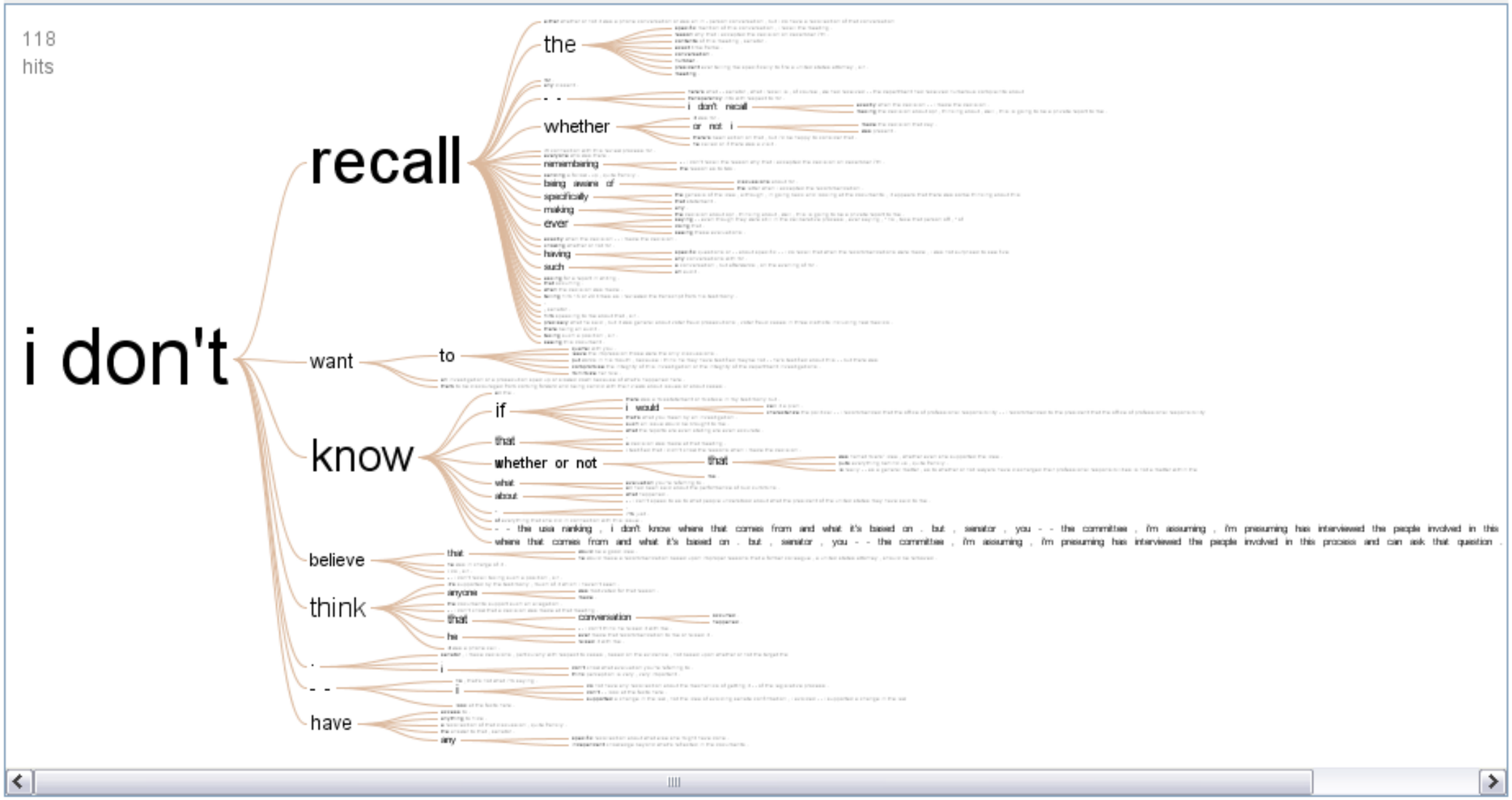
legal
terms of use

Popular Dataset Tags

2007 2008 bible blog books census crime education eharmony election energy food health inauguration internet ireland literature lyrics media music network obama people politics population president prices religion

Search **i don't** Back Forward Start End Occurrence Order Clicks Will Zoom

118 hits



Data file: Word in testimony from Gonzales, 4/19/2007 Data source: CQ Transcript Wire via the Washington Post This data set has not yet been rated

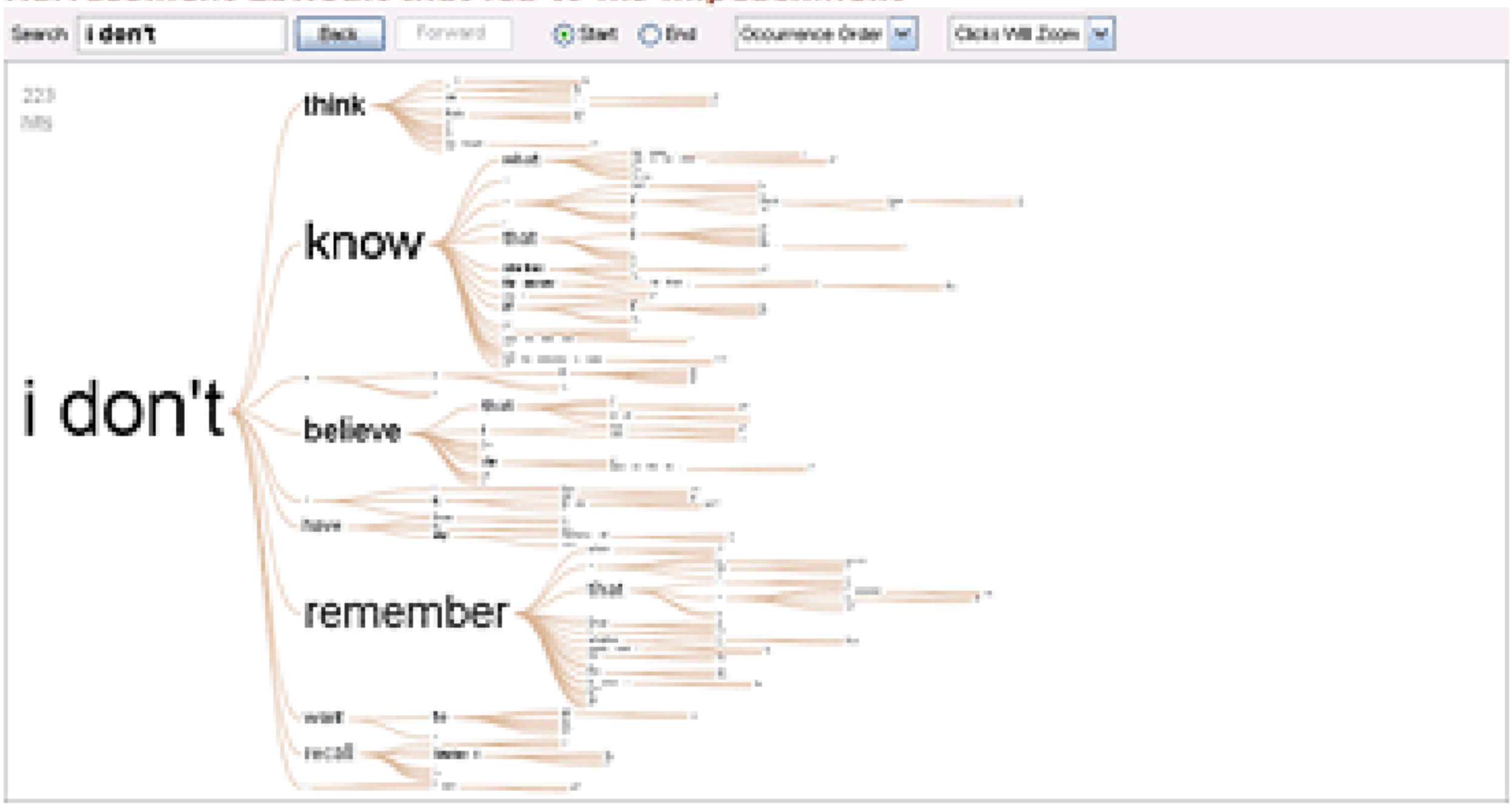
full image share this watch this add to topic hub rate this

Comments (4)

currently showing

This visualization has 4 positive and 0 negative

Visualizations : William "I don't recall" Jefferson Clinton Testimony in Sexual Harrassment Lawsuit that led to his impeachment



Glimpses of Structure…

Concordances show local, repeated structure

But what about other types of patterns?

Lexical: <A> at <B>

Syntactic: <Noun> <Verb> <Object>

Phrase Nets [van Ham et al.]

Look for specific **linking patterns** in the text:

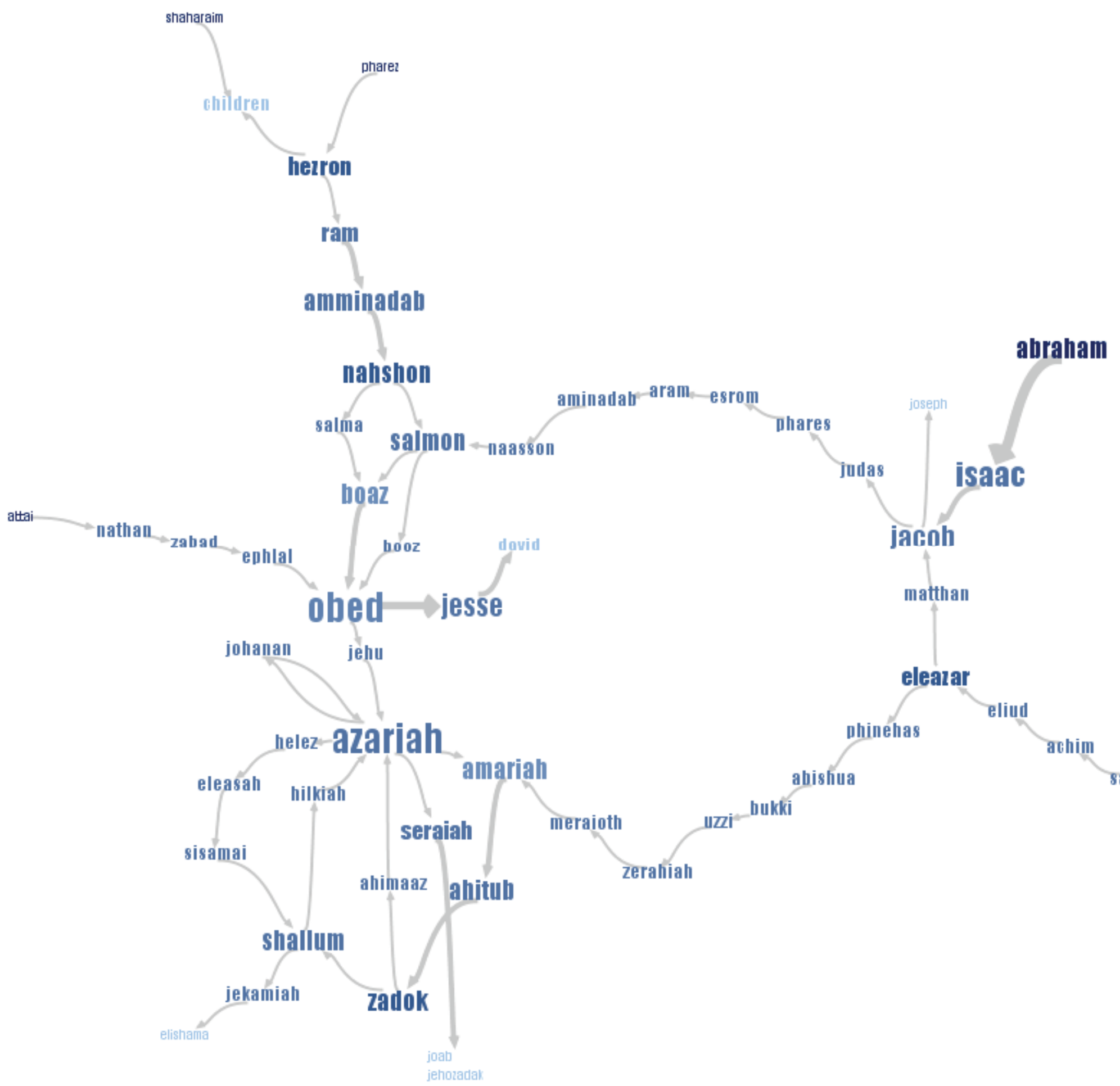
‘A and B’, ‘A at B’, ‘A of B’, etc

Could be output of regexp or parser.

Visualize patterns in a node-link view

Occurrences -> Node size

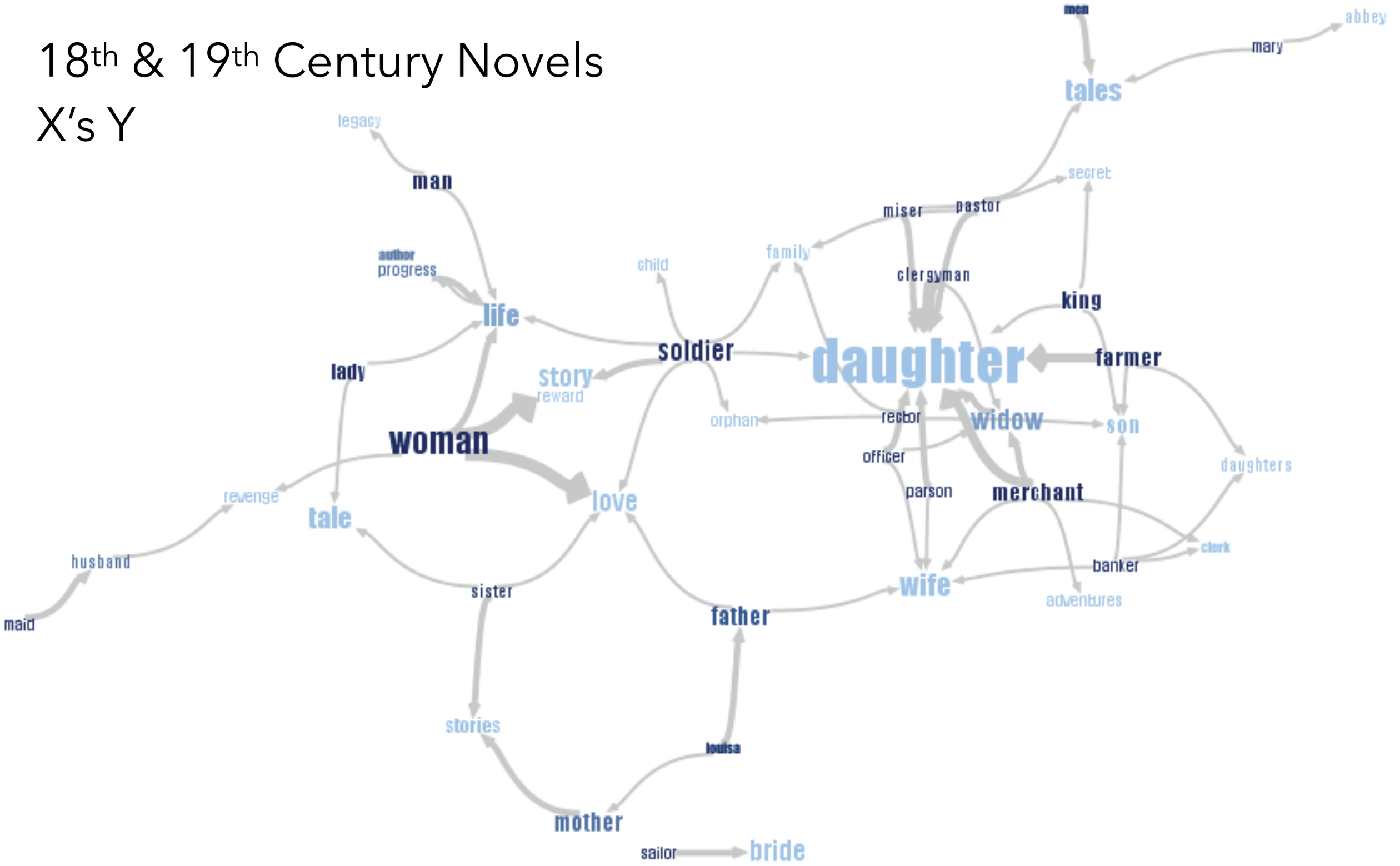
Pattern position -> Edge direction



The Bible
X begat Y

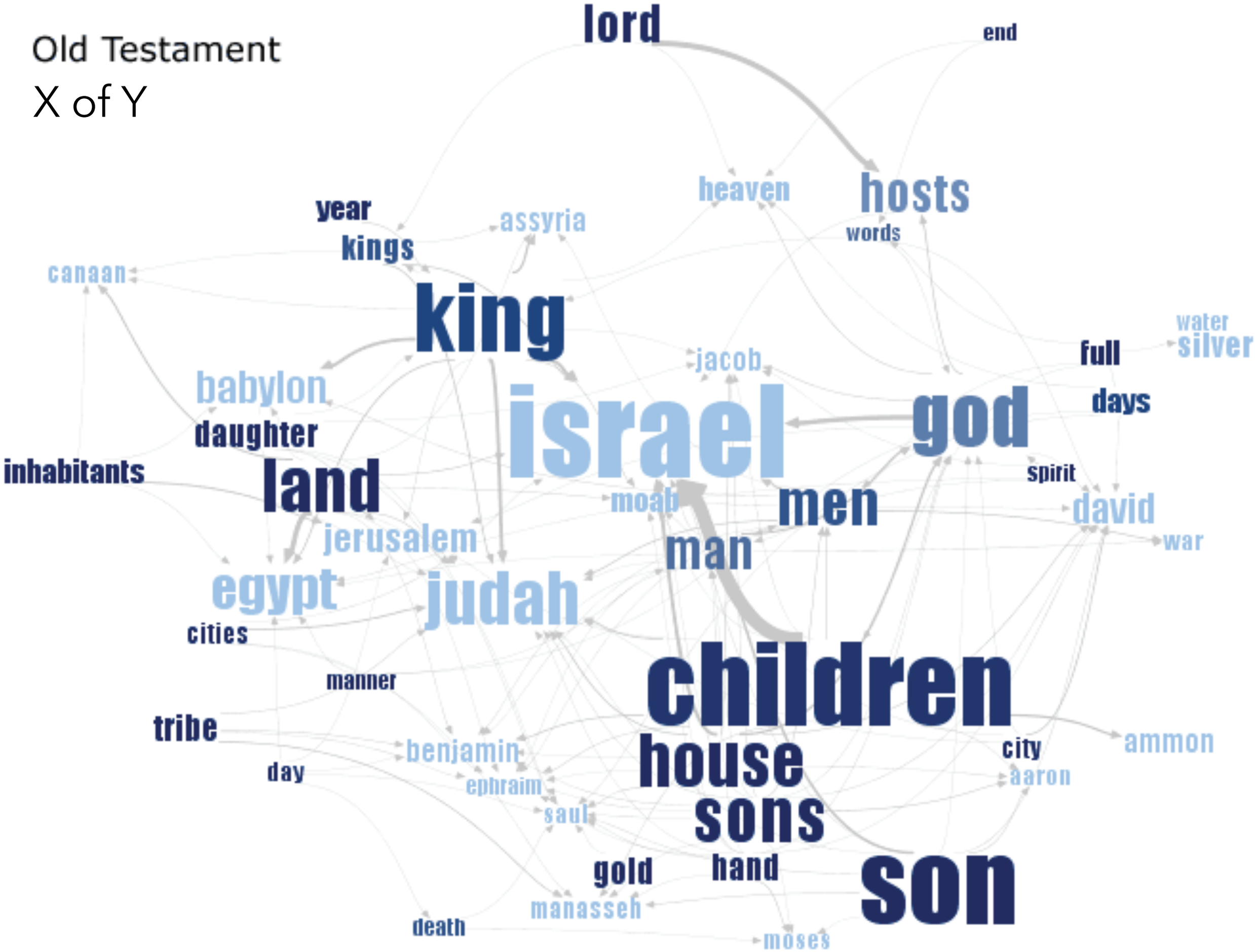
18th & 19th Century Novels

X's Y



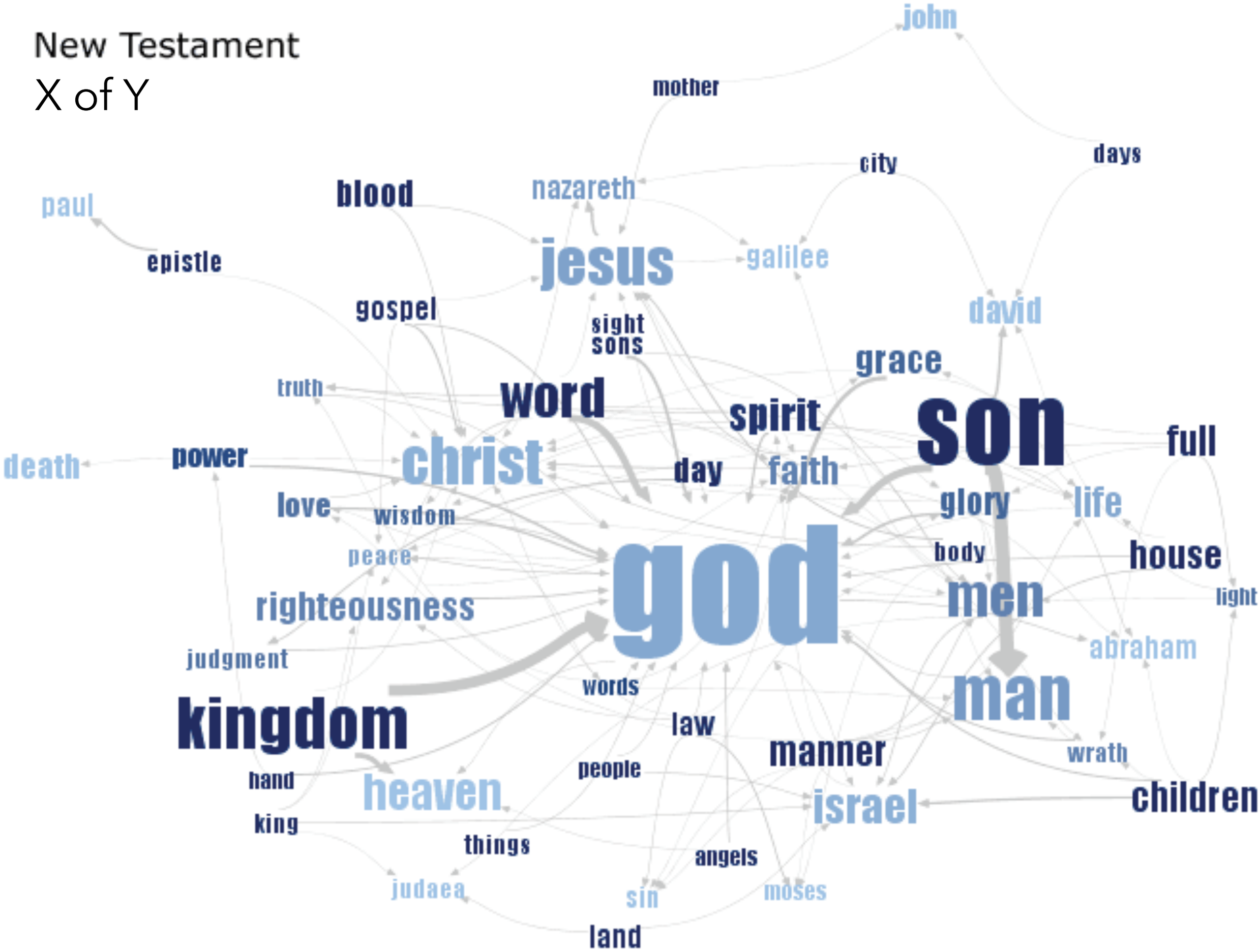
Old Testament

X of Y



New Testament

X of Y



Visualizing Conversation

Many dimensions to consider:
Who (senders, receivers)
What (the content of communication)
When (temporal patterns)

Interesting cross-products:
What x When -> Topic “Zeitgeist”
Who x Who -> Social network
Who x Who x What x When -> Information flow

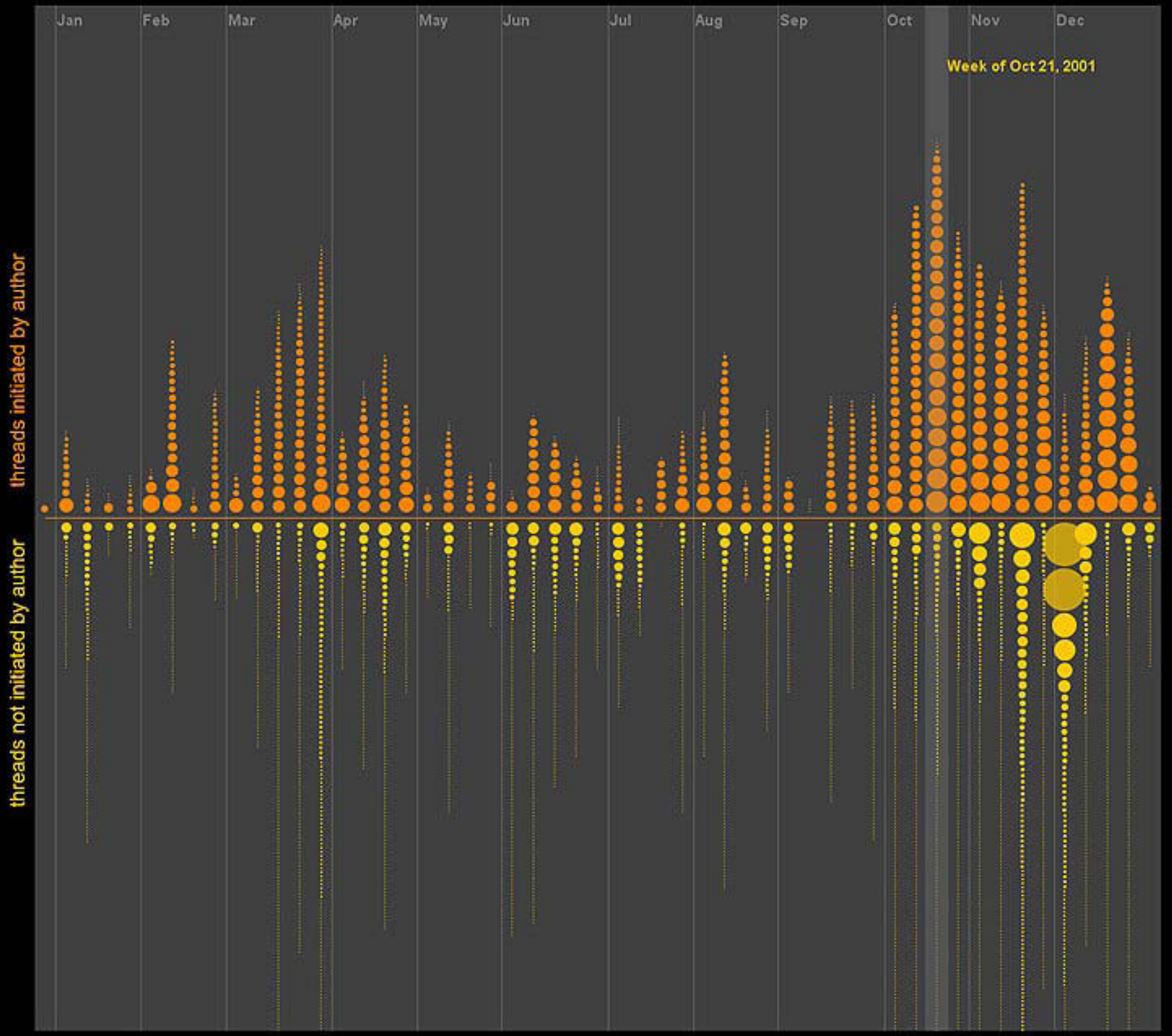
Usenet Visualization [Viegas & Smith]

Show correspondence patterns in text forums

Initiate vs. reply; size and duration of discussion

author: jillyb@mail.com

back to newsgroups



subject	# of posts
Wednesday Spooker ASF	21
WET #3 Anyone for breakfast...	20
Sunny Side Up ASF :)	18
Saturday Ensemble and WET ...	18
Oh no! Watch out!! ASF	18
Thursday Combo-Post WET #...	16
The Yellow Rose Inn.....A gift to...	16
WET #1 JBP The First Time ...	16
We Love the Earth ASF	15
Monday Spooker "The Sight" ...	15
C'mon!!!!	14
Theberge "Le Vent Se Leve" ...	14
Holiday Tog #3 :)	13
Spooker du Jour :)	13
Beginning ASF Short and ...	13
Second Try A Kalie for Suzy...	12
Come On a Safari With Me.....	11
Tuesday Spooker ASF	11
Curses, Foiled Again..... ASF ...	10
Halloween Togs Take Two :)	9
Beauty of the Fury Jim Warren...	9
I thought I saw.....? ASF	7
Wednesday Evening at the Con...	4
Second Try A Kalie for Suzy...	2
Frank Was A Monster ASF...	1

subject	# of posts
Sunday Twofer ASF :)	9
Chopsticks/A Jilly fake	8
Oh no! Trouble in Discworld! ...	7
WET.....your thirst ! ASF ...	6
A pretty for you...Reposted fro...	5
Saturday Spooker ASF	5
Sample Previous install Upgr...	4
Tennessee weather tonite	4
WET - Well I am not smiling!	4
Somethin' mushy <asf>	3
Getting seasonal with workin...	3
A Haunted House :)	3
do you wonder what debi's be...	3
Question: Ethics of posters in ...	3
For Jerry	3
Olu's Tribe - slightly rated	3
WET - Glass Bottles	3
Peace Train<ASF>	2
Arrival at Stewart Island !!	2
WET 195 Wrap-up	2
Cat O'Lantern	2
I Put a Spell on You (Happy H...	2
Goodbye to Summer - A Time!...	2
Two Pumpkins In A Strange B...	2
Still Heading South !!	2
WET- Frank Sinatra - The Man...	2
WET Autumn	2
Purple Martin ASF	2
Opposites Attract . . .	2
Time	2

author: rukbat@pern.org

back to newsgroups

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Week of May 6, 2001

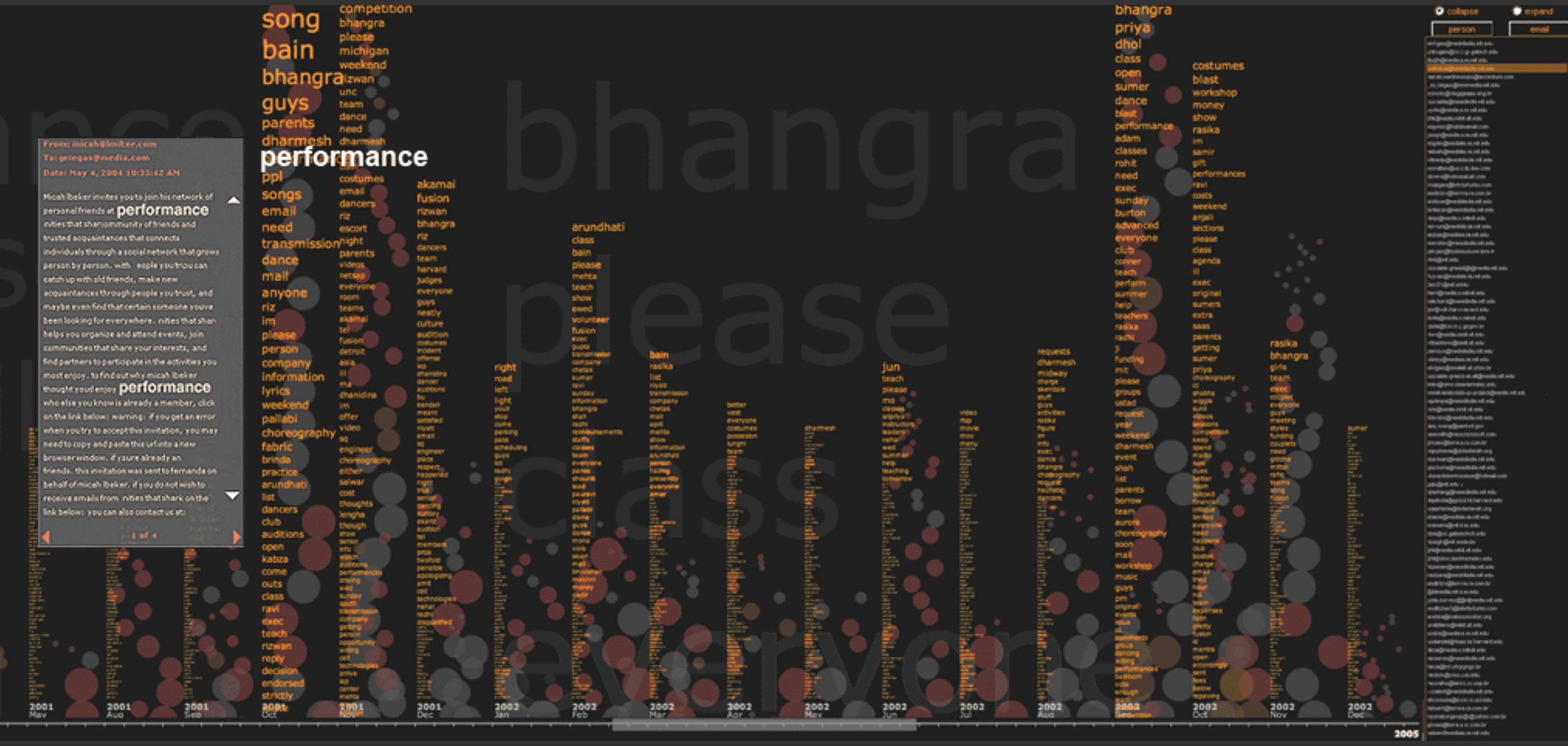
Is Genesis Scientifically Correct?

threads initiated by author

threads not initiated by author

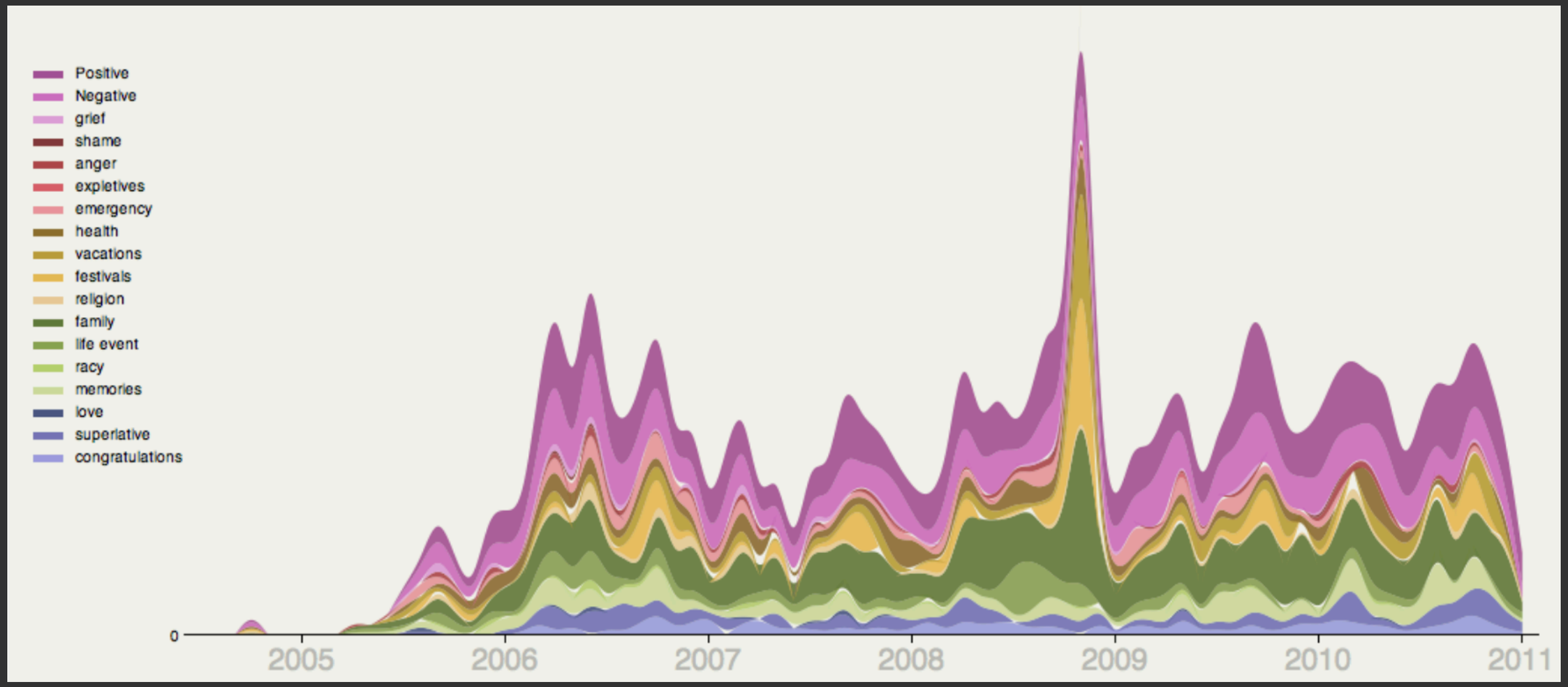
subject	# of posts
Archaeopteryx is	241
Is Genesis Scientif	68
OLD TESTAMENT	37
Please define Ho	33
Evolution flaws	24
Why vouchers wer	23
Scientist against e	15
Bible teaches tha	14
TDMA vs. CDMA	10
God and G-d	7
Darwin is wrong	7
The Atheist Is Guil	7
President Bush Et	7
I got a question	6
A Feathered Dino	6
Freedom from reli	5
Original Sin=Bad F	5
I have Proof that Cr	5
SCIENTIFIC PRO	5
Christians block A	4
Archaeopteryx is	4
An Idiot's respons	4
Car Thief	4
1800MHz vs. 1900	4
Vouchers K-12	4

Themail [Viegas, Golder, & Donath]



One person over time, TF.IDF weighted terms

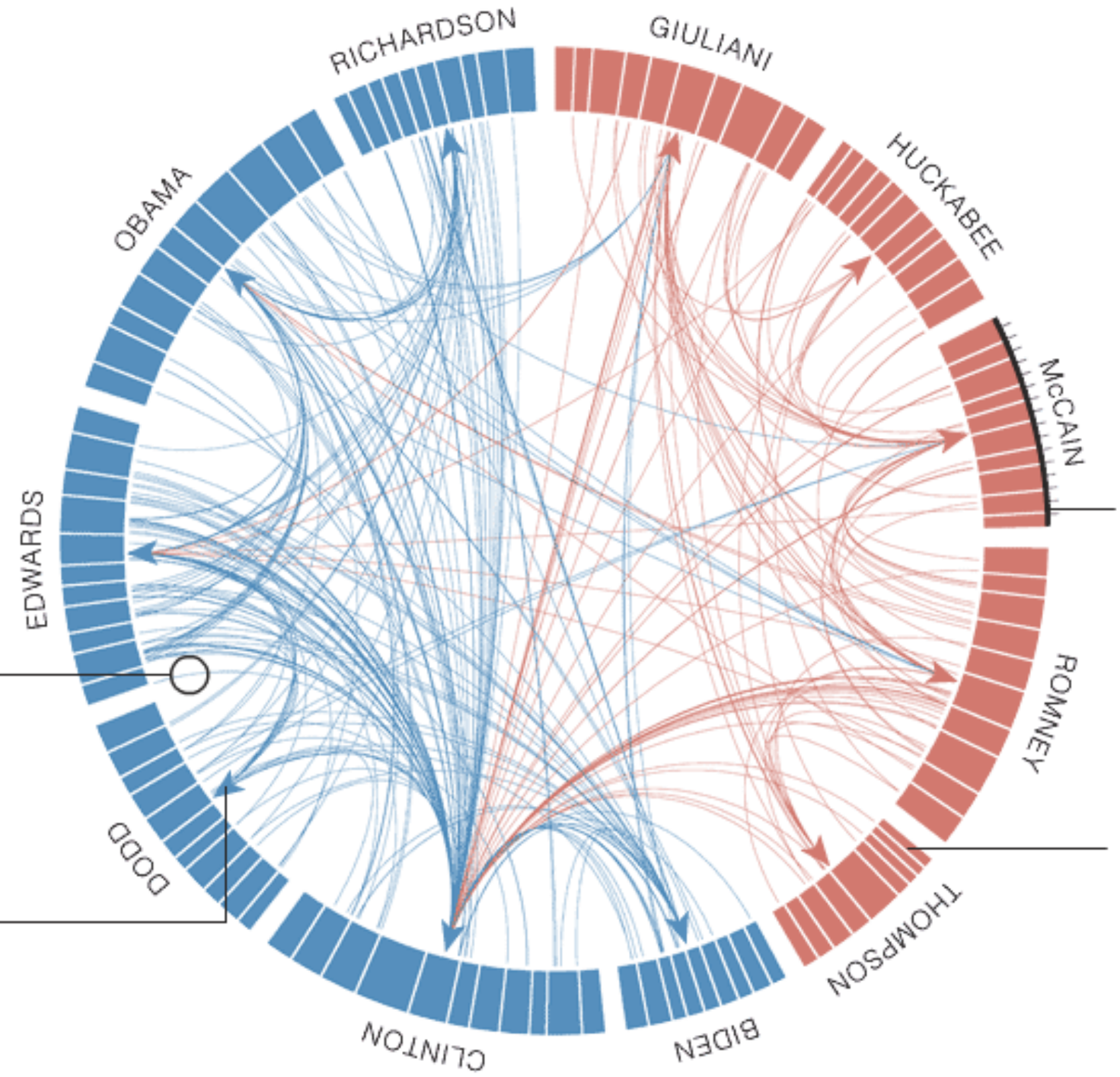
MUSE [Hangal, Lam, & Heer]



Naming Names

Names used by major presidential candidates in the series of Democratic and Republican debates leading up to the Iowa caucuses.

Roll over any candidate's name for details.



Each thin line represents one candidate speaking the last name of another candidate.

Every line ends at an arrow, which points to the name that was spoken.

The length of each circle segment represents the total number of words spoken by the candidate during the debates. Each tick mark represents 1,000 words.

Each slice represents one debate, arranged clockwise from the first to the final debate.

Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times

Summary

High Dimensionality

Where possible use text to represent text…
… which terms are the most descriptive?

Context & Semantics

Provide relevant context to aid understanding.
Show (or provide access to) the source text.

Modeling Abstraction

Understand abstraction of your language models.
Match analysis task with appropriate tools and models.

Currently: from bag-of-words to *vector space embeddings*