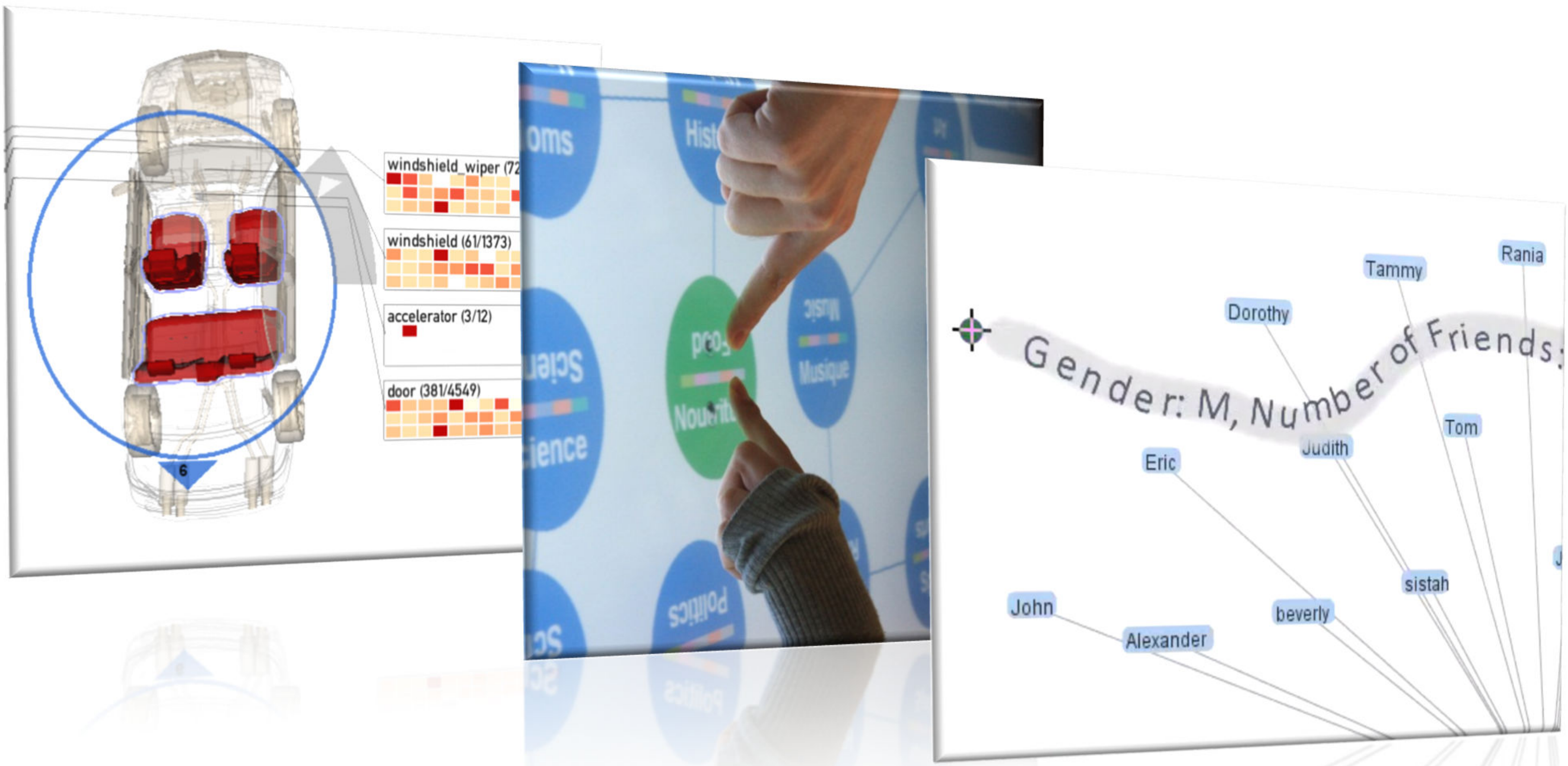


Semantics and Sentiment in Visual Text Analytics

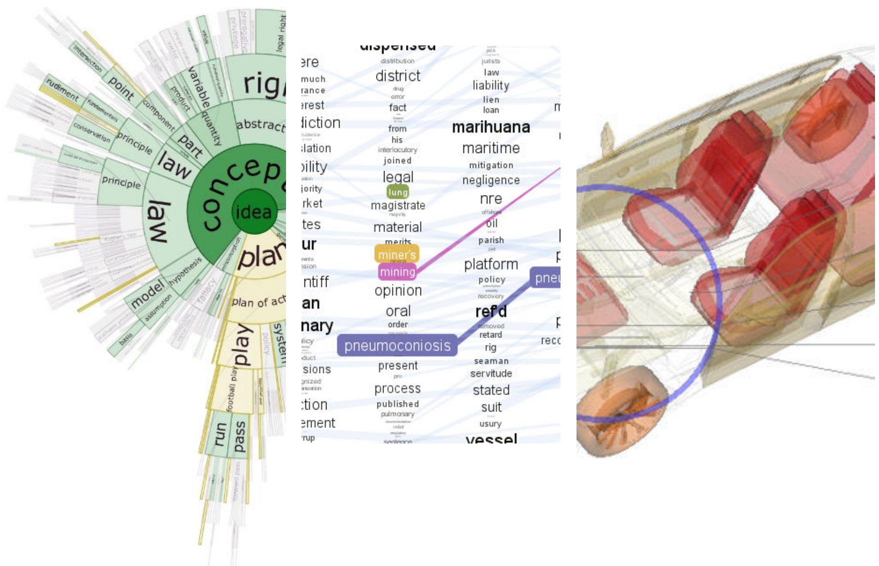
Christopher Collins

Assistant Professor, Canada Research Chair

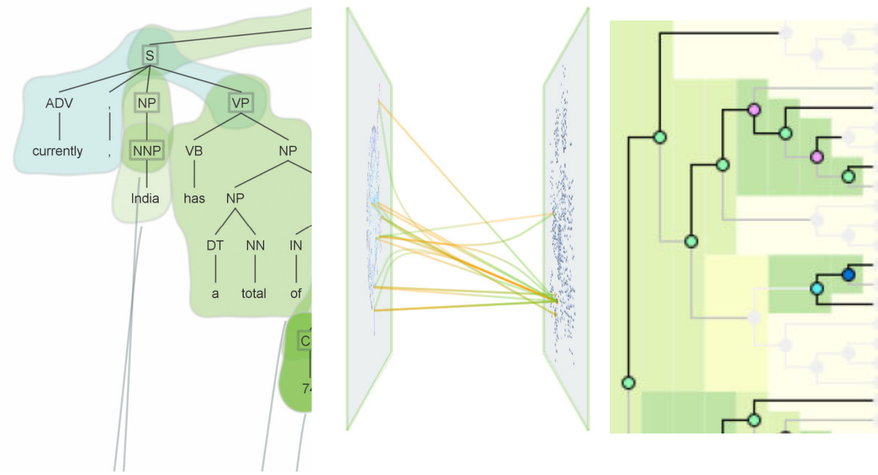
University of Ontario Institute of Technology

vialab visualization for information analysis

UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY

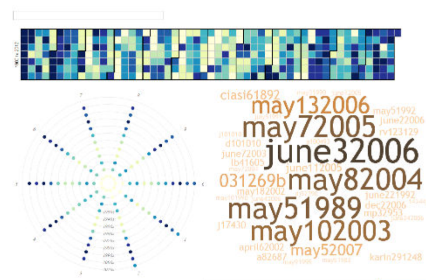


Linguistic Information Visualization

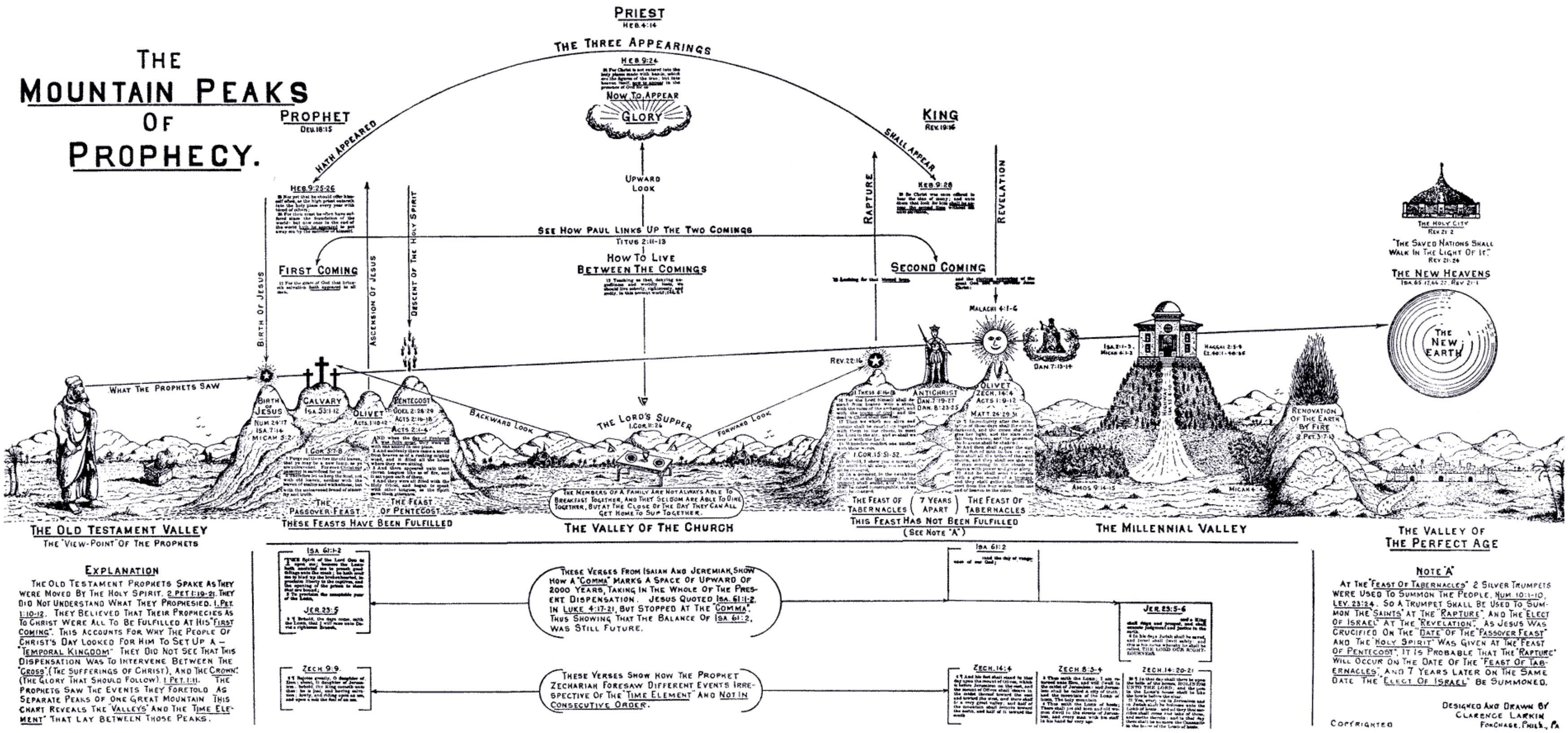


Visualization Technique
and Interaction Design

NUIs for visualization:
tables, walls, gestures



Applied visualization:
software, security,
humanities, healthcare

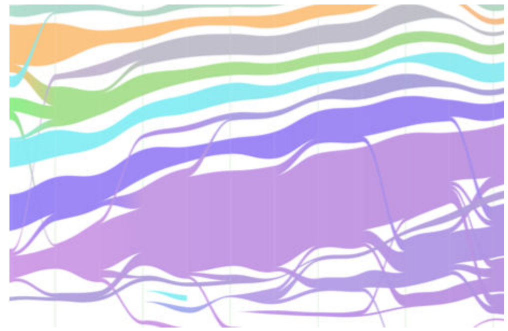


Mountain Peaks of Prophecy (Larkin, 1918)

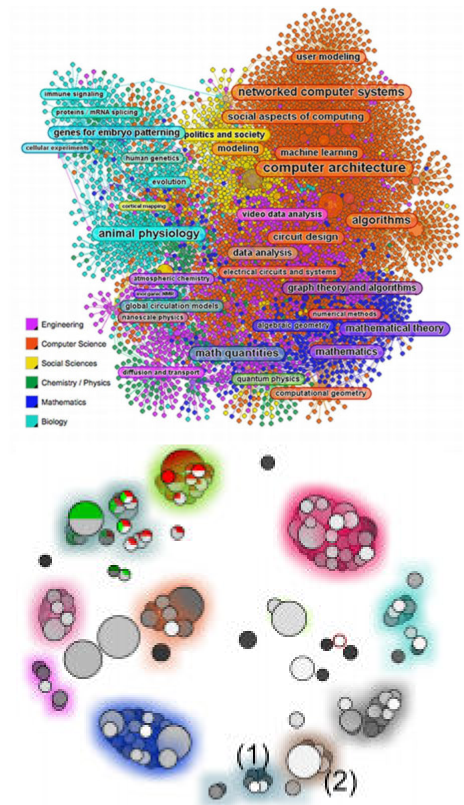
Visual Text Analytics

- Visual techniques for words, documents, sets of documents to support rapid summarization, trend analysis, exploration, search, comparative analysis, …
- Application areas include market analysis, legal studies, e-discovery, readability, literary studies, personal reflection, information retrieval and exploration, intelligence analysis

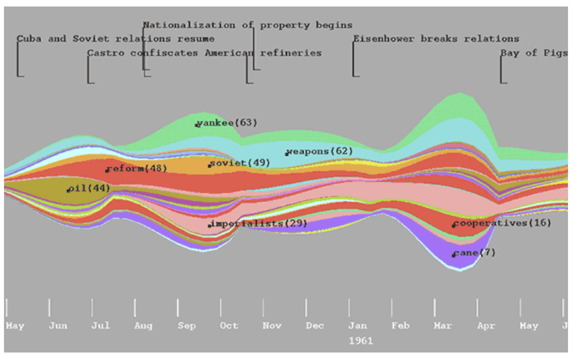
Word Clouds



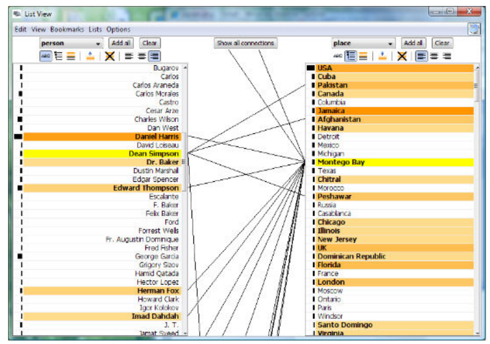
TextFlow



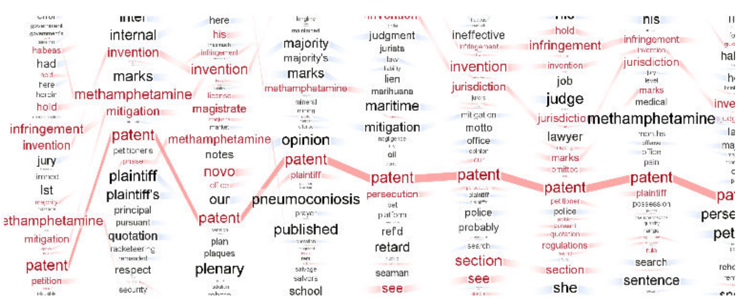
Topic Models



Theme River

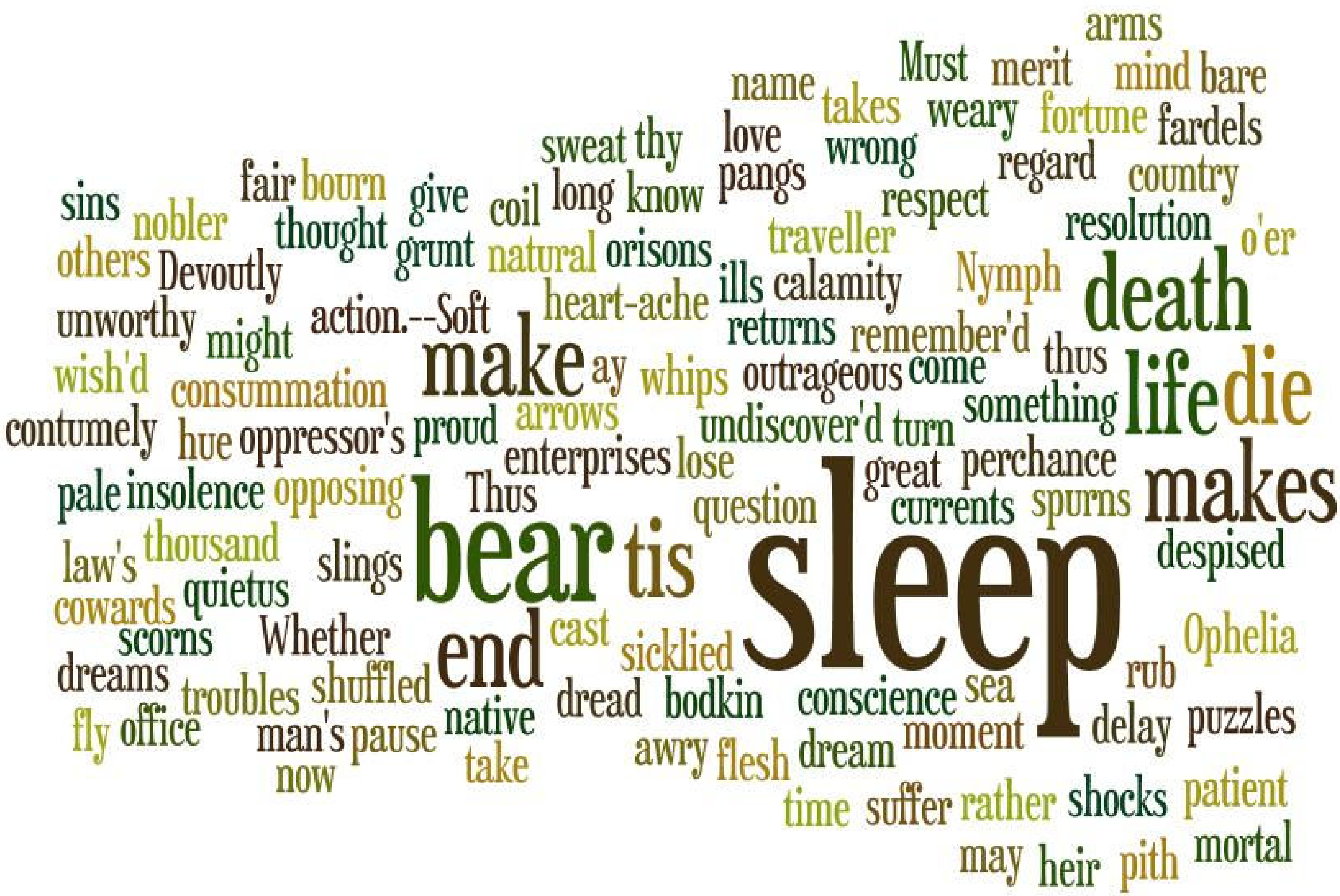


Jigsaw

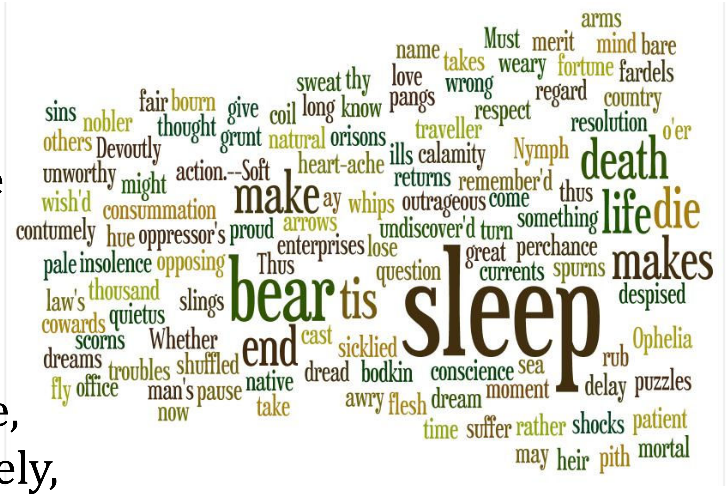


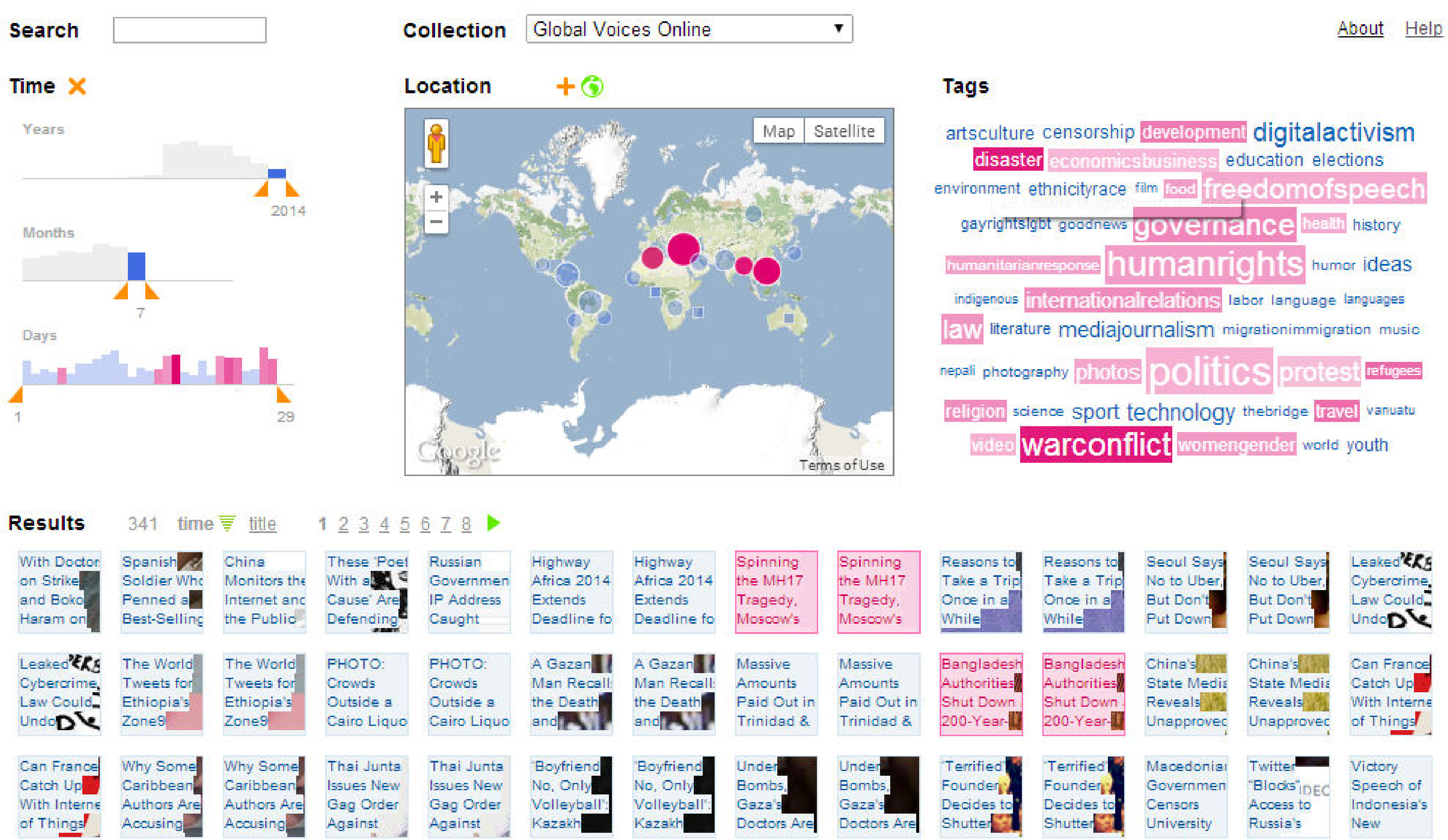
Parallel Tag Clouds

Themescape


arms
Must merit mind bare
name takes weary fortune fardels
sweat thy love wrong regard
fair bourn give coil long know pangs respect country
sins nobler thought grunt natural orisons traveller resolution o'er
others Devoutly ills calamity Nymph
unworthy might action.--Soft heart-ache returns remember'd death
wish'd consummation make ay whips outrageous come thus
contumely hue oppressor's proud arrows undiscover'd turn something life die
enterprises lose great perchance
pale insolence opposing Thus question currents spurns makes
law's thousand slings bear tis despised
cowards quietus sleep
scorns Whether end cast sicklied Ophelia
dreams troubles shuffled rub
fly office man's pause native dread bodkin conscience sea delay puzzles
now take awry flesh dream moment
time suffer rather shocks patient
may heir pith mortal

To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death what dreams may come
When we have shuffled off this mortal coil,
Must give us pause: there's the respect
That makes calamity of so long life;
For who would bear the whips and scorns of time,
The oppressor's wrong, the proud man's contumely,
The pangs of despised love, the law's delay,
The insolence of office and the spurns
That patient merit of the unworthy takes,
When he himself might his quietus make
With a bare bodkin? who would fardels bear,
...





Marian Dörk et al. VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205-1212, November-December, 2008.

Linguistic Methods

- Word Counting
- Word Scoring
- Stemming
- Stop Word Removal
- Part of Speech Tagging
- Parsing
- Word Sense Disambiguation
- Named Entity Recognition
- Semantic Categorization
- Sentiment Analysis
- Topic Modeling (some caveats)

NLTK: Natural Language Toolkit

- NLTK.org
- Python

NLTK 3.0 documentation

NEXT | MODULES | INDEX

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

TABLE OF CONTENTS

SEARCH

Go

Enter search terms or a module, class or function name.

Stemming

- Reduce words to their ‘stems’ by removing endings (morphology)
  - running -> run
  - runs -> run
- A good way to increase signal and reduce fracturing of the corpus if there aren’t many words.
- Note: Keep the original words somewhere! Also keep the case if you choose to lowercase the word; you never know when you’ll need this data

Stop Word Removal

- Common words such as “and”, “the”, “I” are removed from view to highlight content words
- Domain specific stop words, e.g. in legal domain:
  - Court, attorney, honour, plaintiff, etc.
- Caution! These words have been shown to be useful for stylistic analysis! When working with text corpora, KEEP EVERYTHING.

Part of Speech Tagging

- Assign grammatical roles to words
- Conventional tagsets and representation:
  - The/AT grand/JJ jury/NN commented/VBD on/IN a/AT number/NN of/IN …
- Many words are ambiguous: fly, chair, run, store, table, and more!
  - Fly/NN
  - Fly/VB

Fly/NN


Fly/VB

Term / Concept Ambiguity

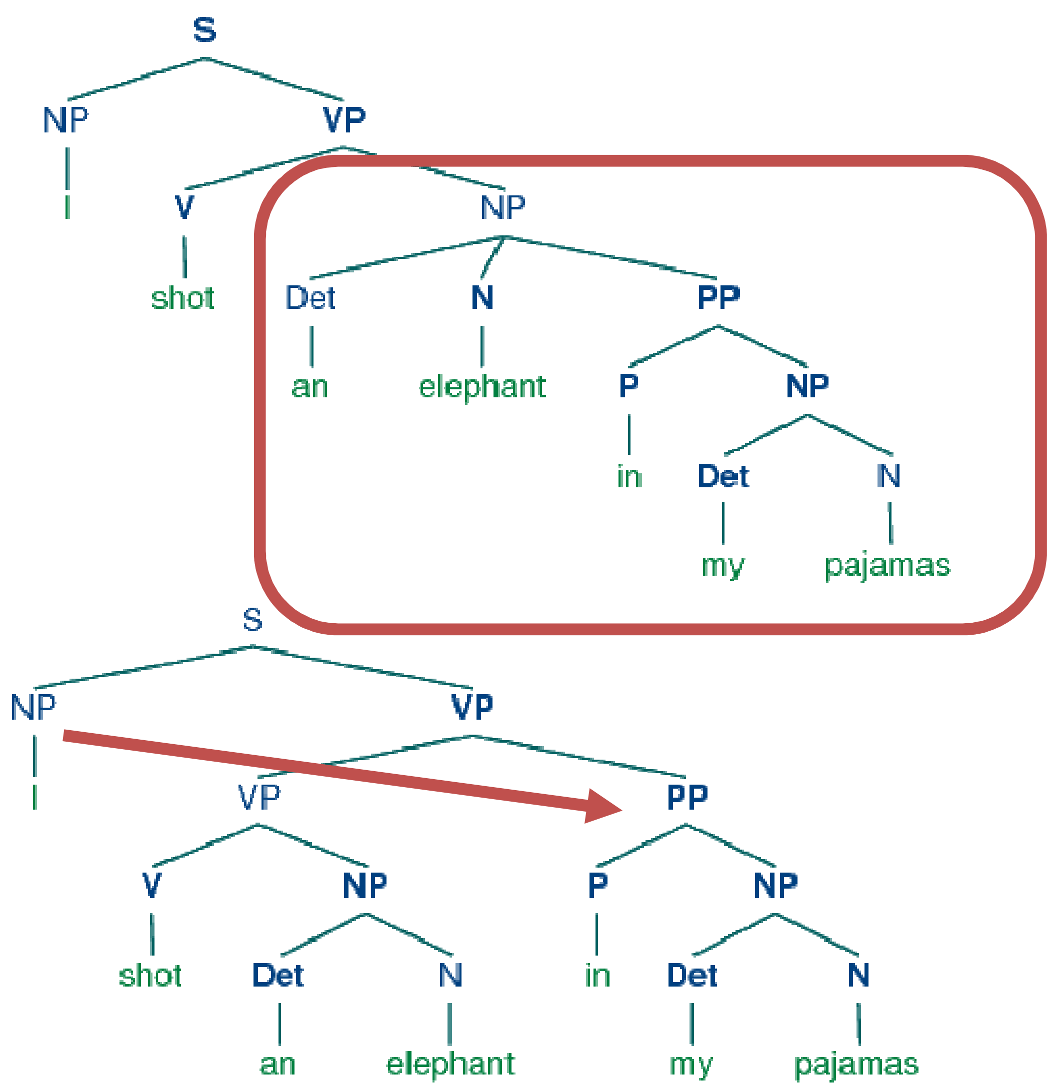
- Most meaning comes from our minds and common understanding.

- “How much is that doggy in the window?”
  - how much: social system of barter and trade (not the size of the dog)
  - “doggy” implies childlike, plaintive, probably cannot do the purchasing on their own
  - “in the window” implies behind a store window, not really inside a window, requires notion of window shopping

(Hearst, 2006)

Parsing

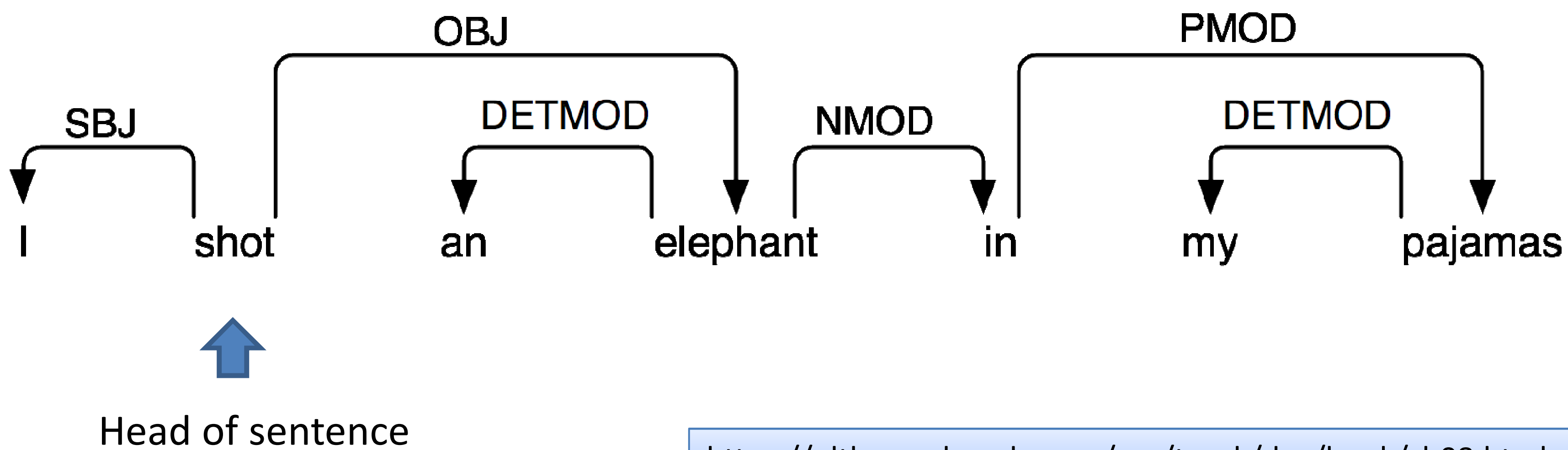
- Determining language structure
- Can reveal word-word relationships
- Useful for processing negation



https://nltk.googlecode.com/svn/trunk/doc/book/ch08.html

Dependency Parsing

- Labelled directed graph
- Arcs represent relationships from heads to dependents



https://nltk.googlecode.com/svn/trunk/doc/book/ch08.html

Word Sense Disambiguation

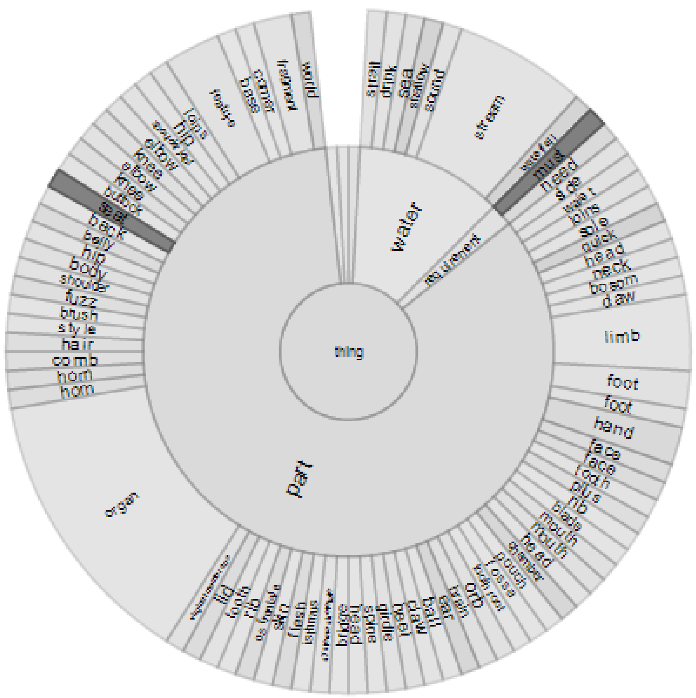
- Susan, the meeting chair, chaired the meeting well from the big chair in the front of the room.
  - Leader of a meeting
  - Action of leading a meeting
  - An object to sit upon

Word Sense Disambiguation

- This is VERY difficult for a computer.
- Contexts are often the same and meanings can be quite fine-grained:
  - bank the financial institution, bank the building in which the financial institution is housed
- Annual contest: SENSEVAL
- My method: assume the most common sense

Named Entity Recognition

- What are the people, places in the text?
- Use NLTK – it's very good at this.



http://vialab.science.uoit.ca/docuburst

Semantic Categorization

- Placing a word into an ontology or sense thesaurus based on *meaning.*

- Common resources include:
  - WordNet
  - Roget’s Thesaurus

WordNet

- A large lexical database, or “digital dictionary”
- Covers most English nouns, verbs, adjectives, adverbs
- Organizes *synsets* by *meaning*
- Words are related to one another through many different relationship types:
- X is a kind of Y, X has part Y, an X Ys, X is Y/has property Y

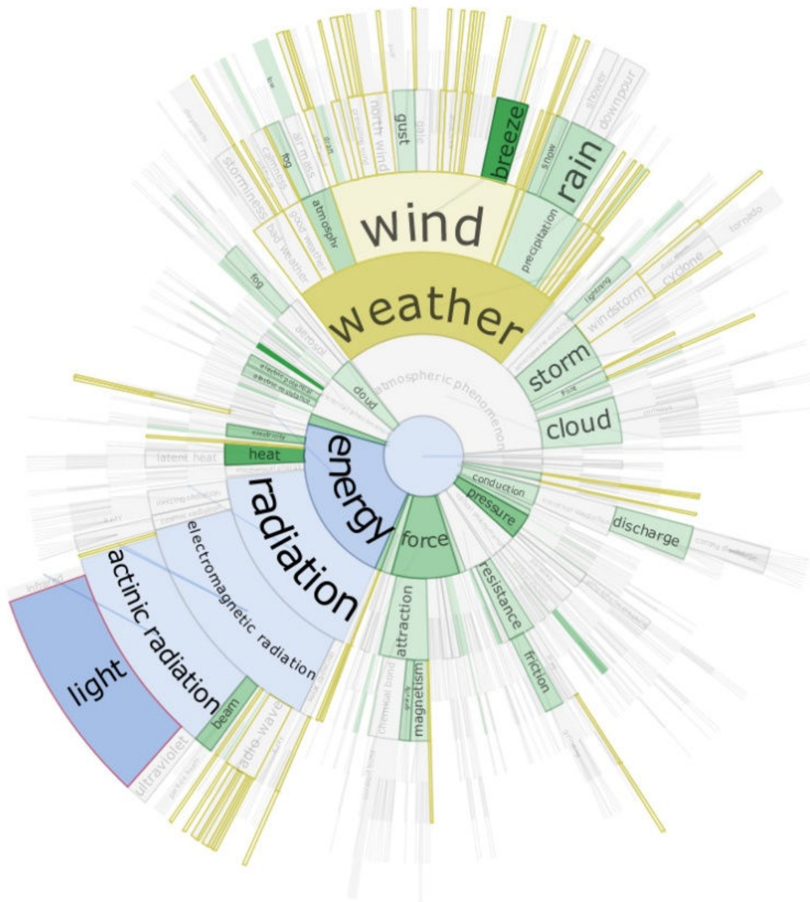
Hyponymy

- The “IS-A” relation for nouns

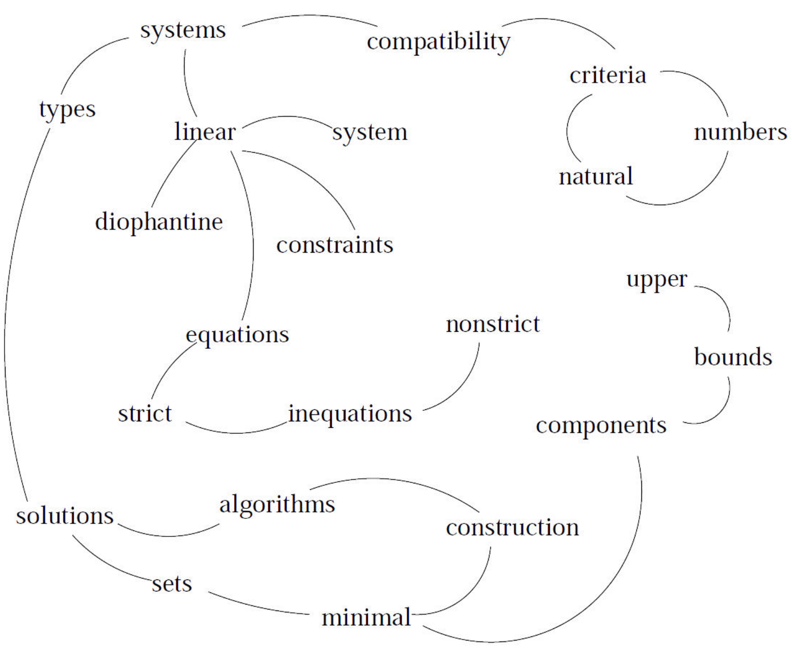
```
                {vehicle}
                  /   \
  {car, automobile}     {bicycle, bike}
      /           \               \
    {convertible}  {SUV}  {mountain bike}
```

SEMANTIC VISUALIZATIONS

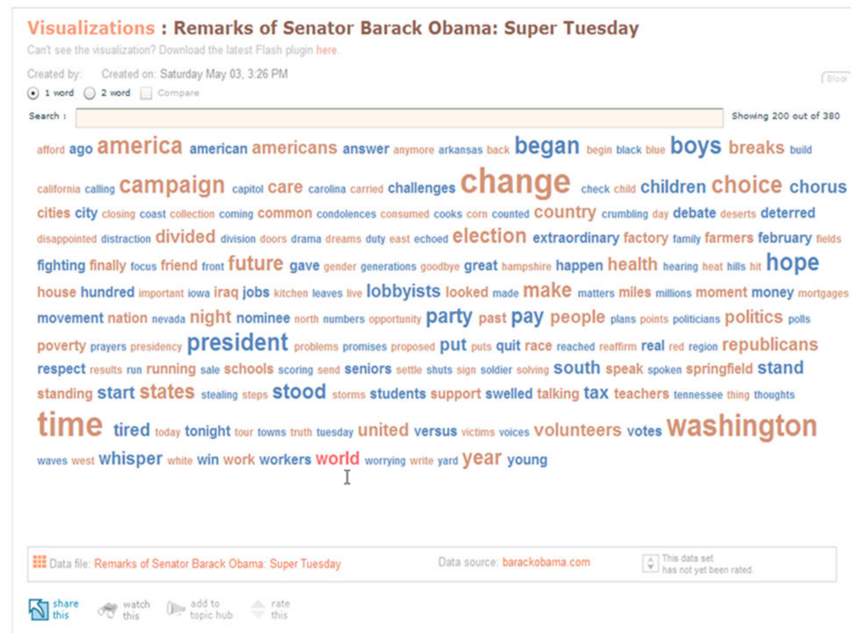
DocuBurst



Collins, C.; Carpendale, S.; Penn, G. DocuBurst: Visualizing Document Content using Language Structure. Proceedings of Eurographics/IEEE VGTC Symposium on Visualization, June, 2009.



Mihalcea and Tarau, 2004

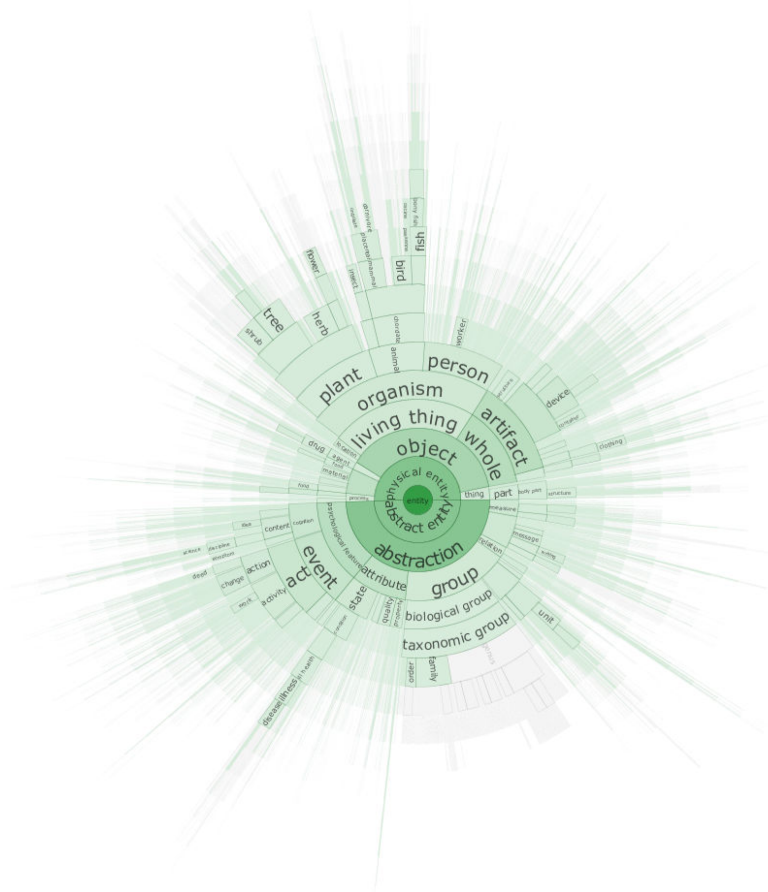


Wattenberg et al., 2008

DocuBurst

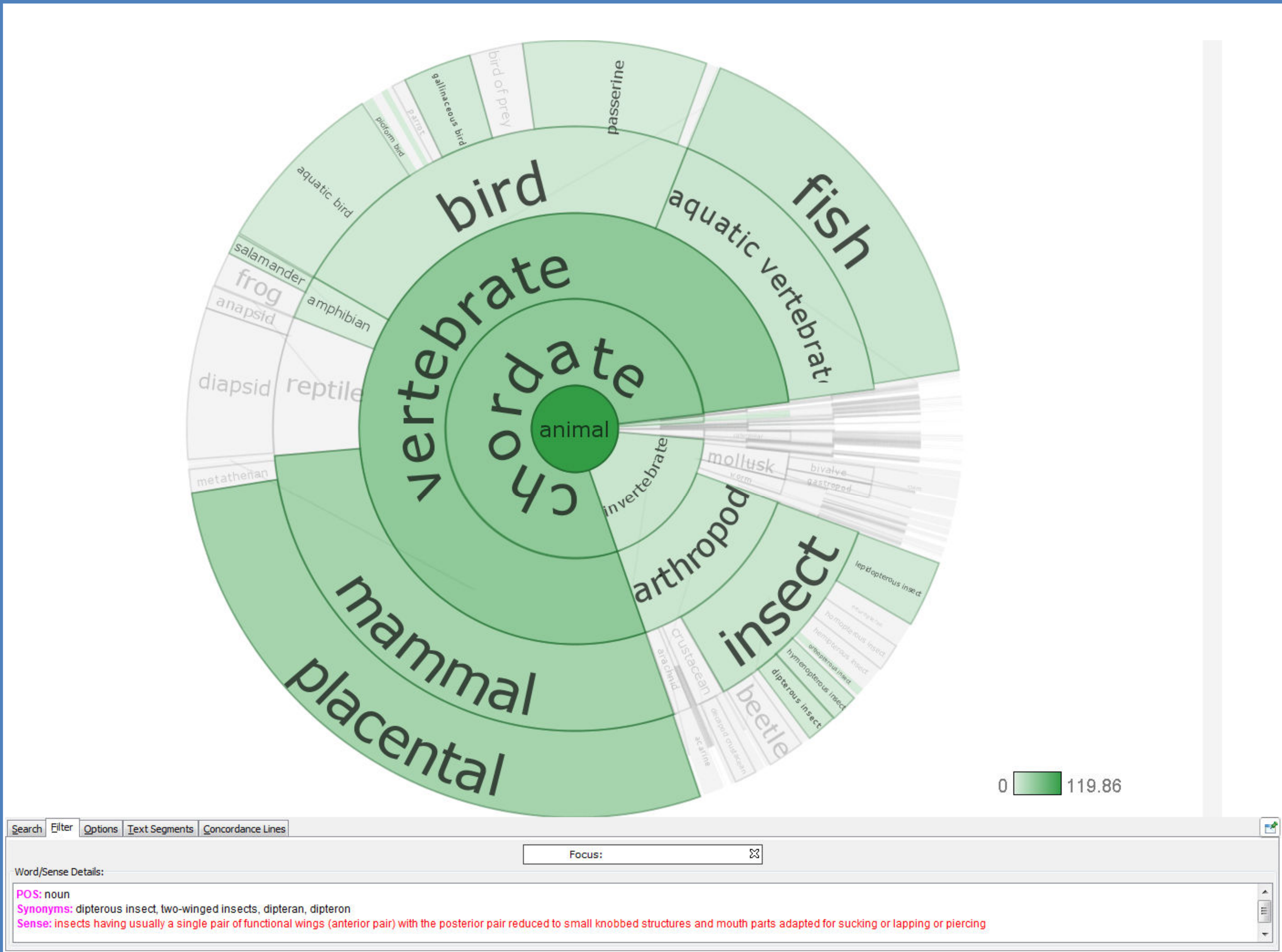
games→game
taken→take

absolute,noun,10
chair,noun,2
moment,noun,11
game,noun,30
reality,noun,3
take,verb,13
represent,verb,17
...

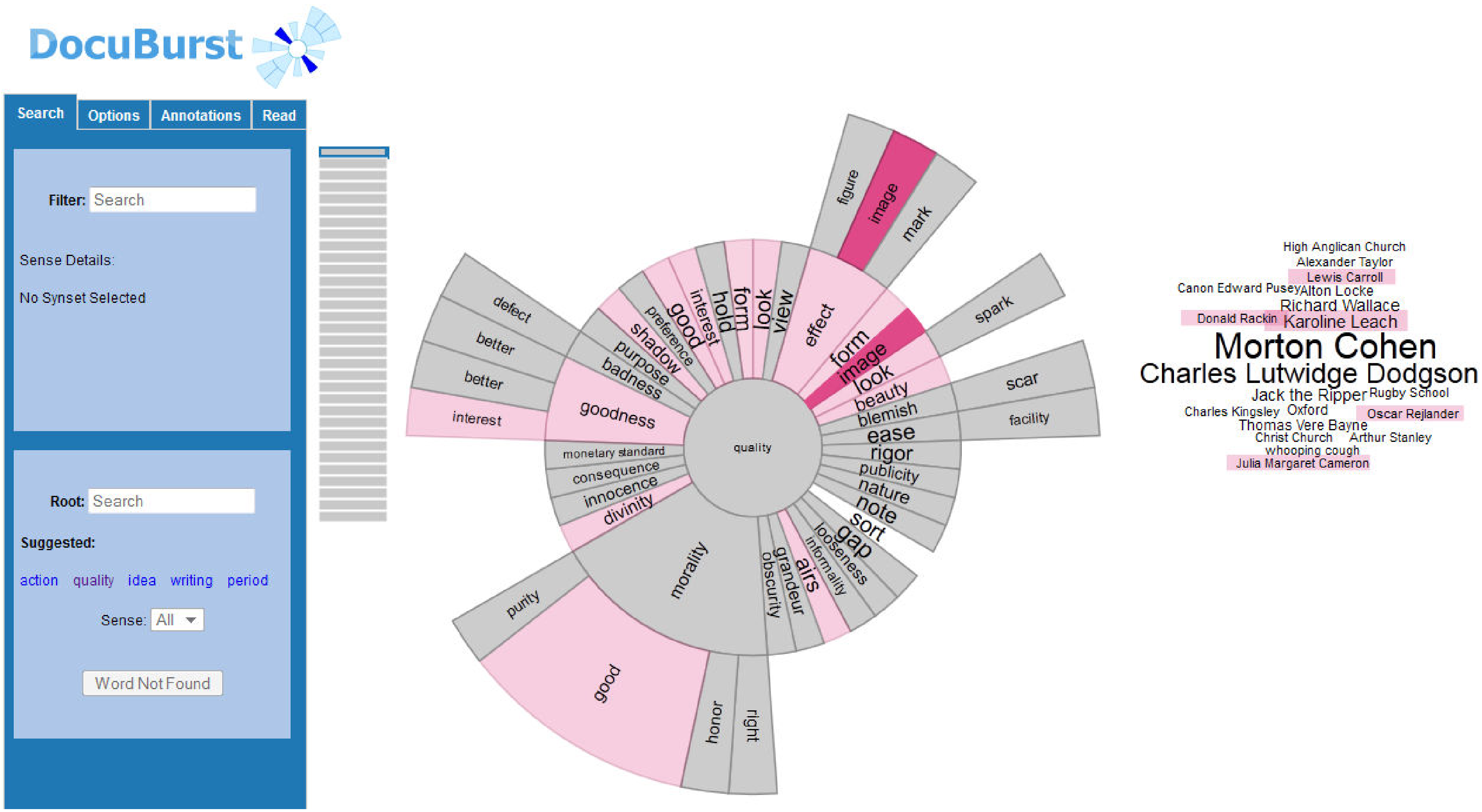


WordNet

game IS activity
chair IS furniture

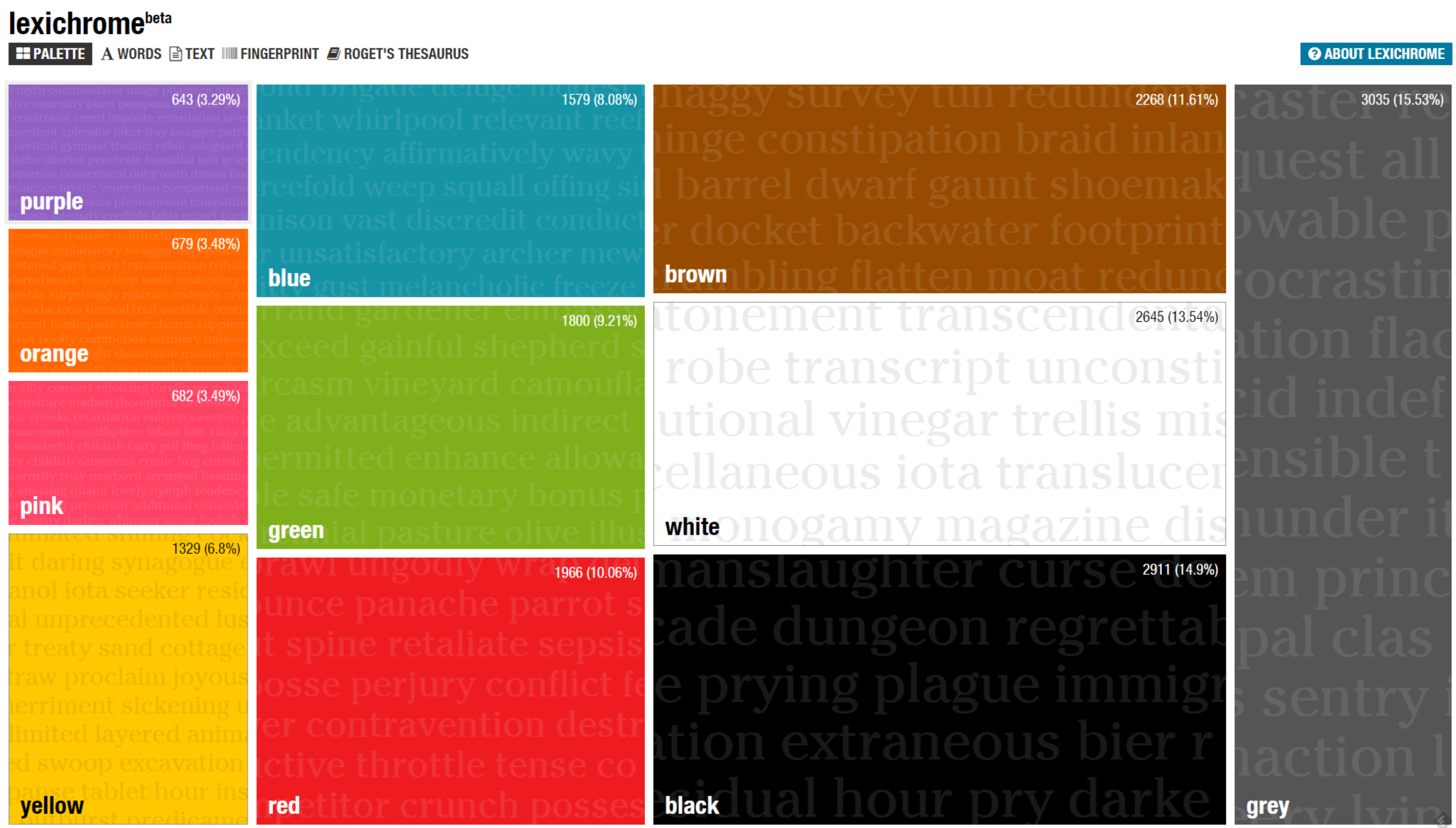


animal
chordate
vertebrate
invertebrate
bird
aquatic vertebrate
fish
mammal
placental
arthropod
insect
mollusk
reptile
diapsid
anapsid
frog
salamander
amphibian
aquatic bird
piciform bird
gallinaceous bird
bird of prey
passerine
metatherian
crustacean
arachnid
beetle
dipterous insect
hymenopterous insect
lepidopterous insect
bivalve
gastropod
0
119.86
Search
Filter
Options
Text Segments
Concordance Lines
Focus:
Word/Sense Details:
POS: noun
Synonyms: dipterous insect, two-winged insects, dipteran, dipteron
Sense: insects having usually a single pair of functional wings (anterior pair) with the posterior pair reduced to small knobbed structures and mouth parts adapted for sucking or lapping or piercing



Try it! http://vialab.science.uoit.ca/docuburst

Lexichrome



http://lexichrome.com

Work in Progress with Chris Kim and Saif Mohammed

‹ all words associated with yellow

PALETTE | WORDS | TEXT

ABOUT LEXICHROME

RELEVANCE (DESC) | ALPHABETICAL

cowardly 10 out of 10	nugget 7 out of 7	sun 7 out of 7	sunny 9 out of 10
saffron 8 out of 9	treasure 7 out of 8	lion 6 out of 7	mustard 6 out of 7
radiant 6 out of 7	bee 11 out of 13	butter 11 out of 13	insecure 6 out of 8
sandy 6 out of 8	scatter 6 out of 8	lightning 8 out of 11	beehive 10 out of 14
practically 5 out of 7	radiate 5 out of 7	enlighten 7 out of 10	sunshine 7 out of 10

lexichrome^alpha

PALETTE | WORDS | TEXT

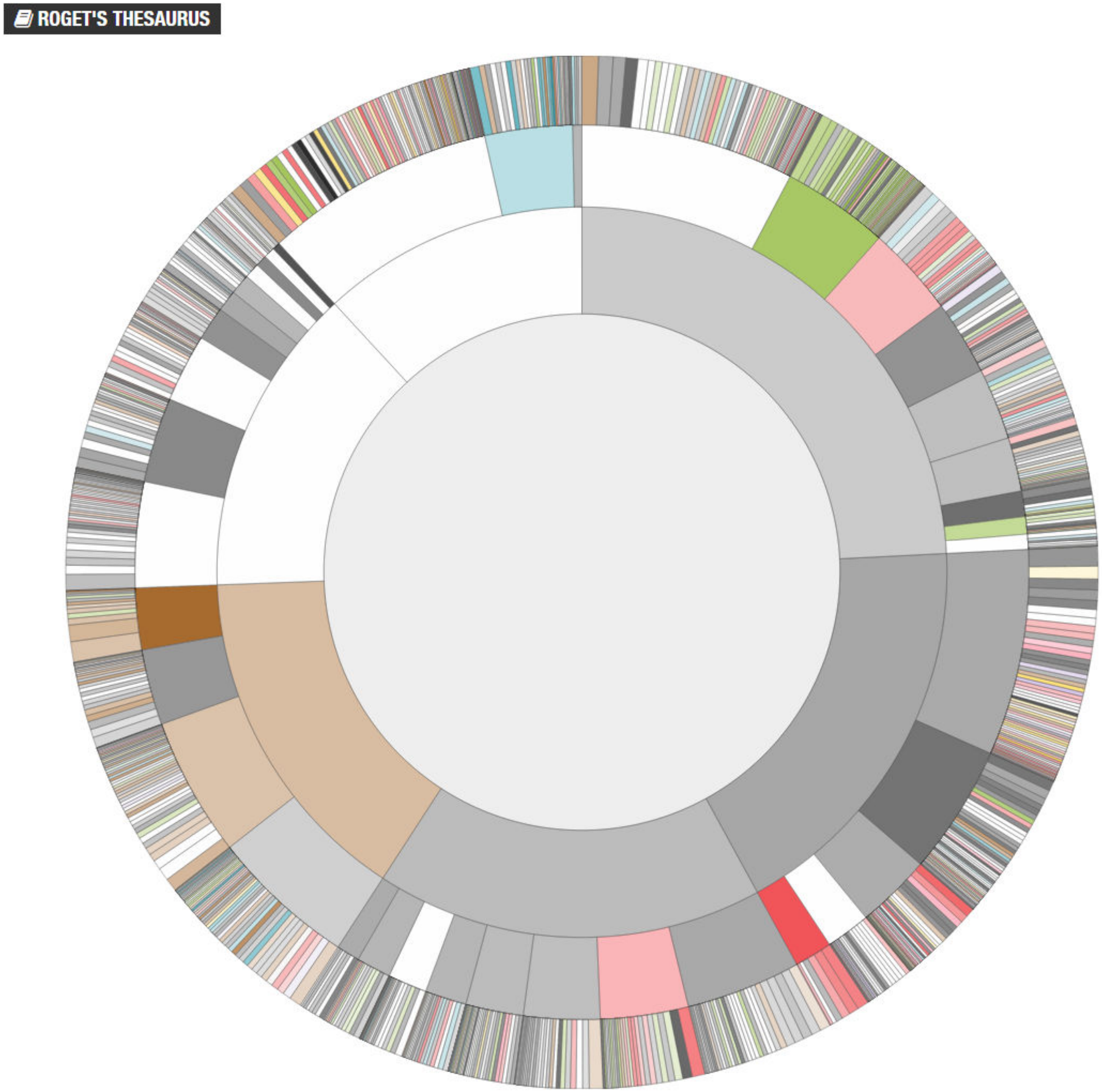
ABOUT LEXICHROME

Nameless here for evermore.

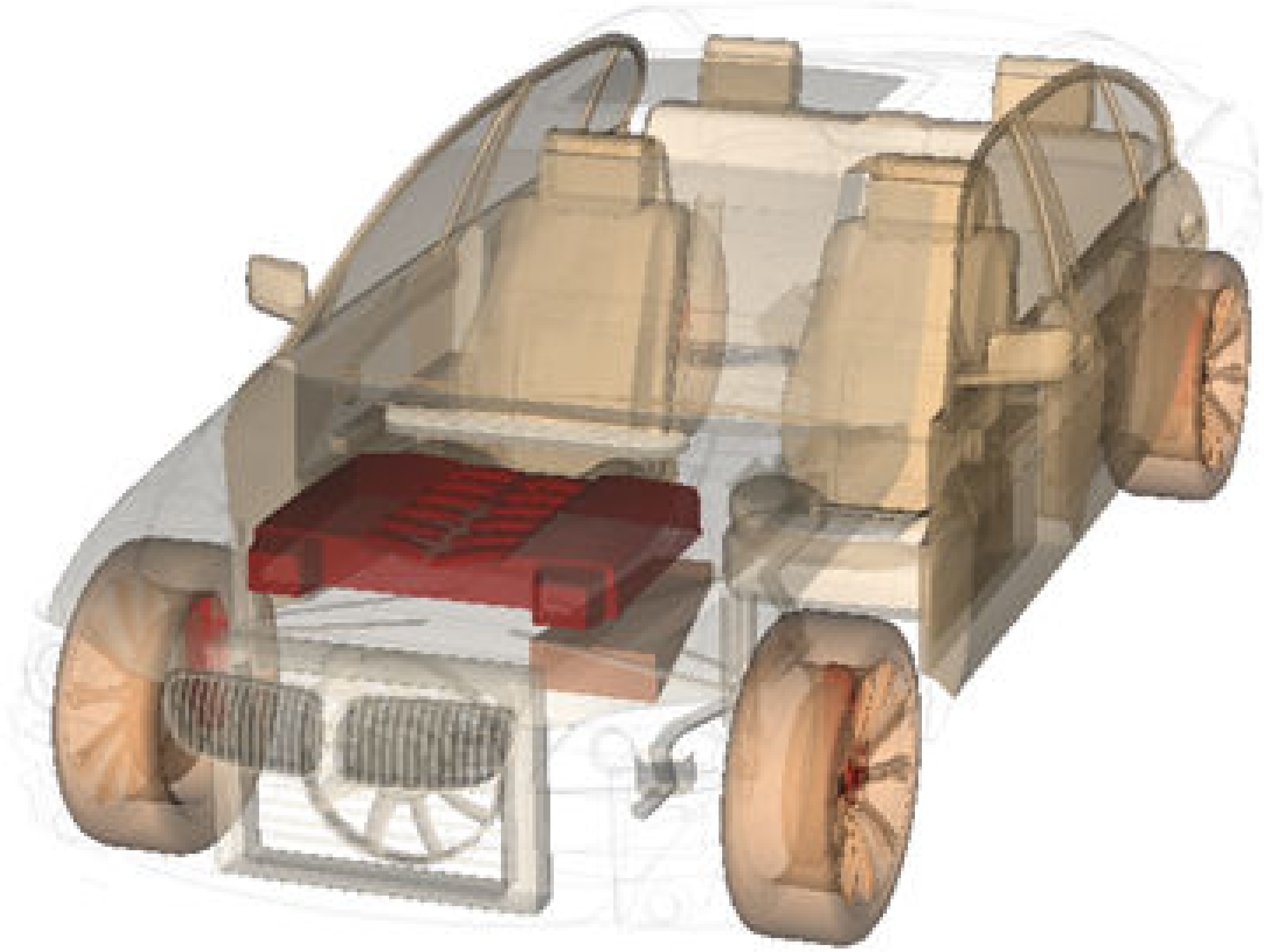
And the silken sad uncertain rustling
of each purple curtain
Thrilled me - filled me with fantastic
terrors never felt before;
So that now, to still the beating of my
heart, I stood repeating
`'Tis some visitor entreating entrance
at my chamber door -
Some late visitor entreating entrance
at my chamber door; -
This it is, and nothing more,'

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.'

ANALYZE

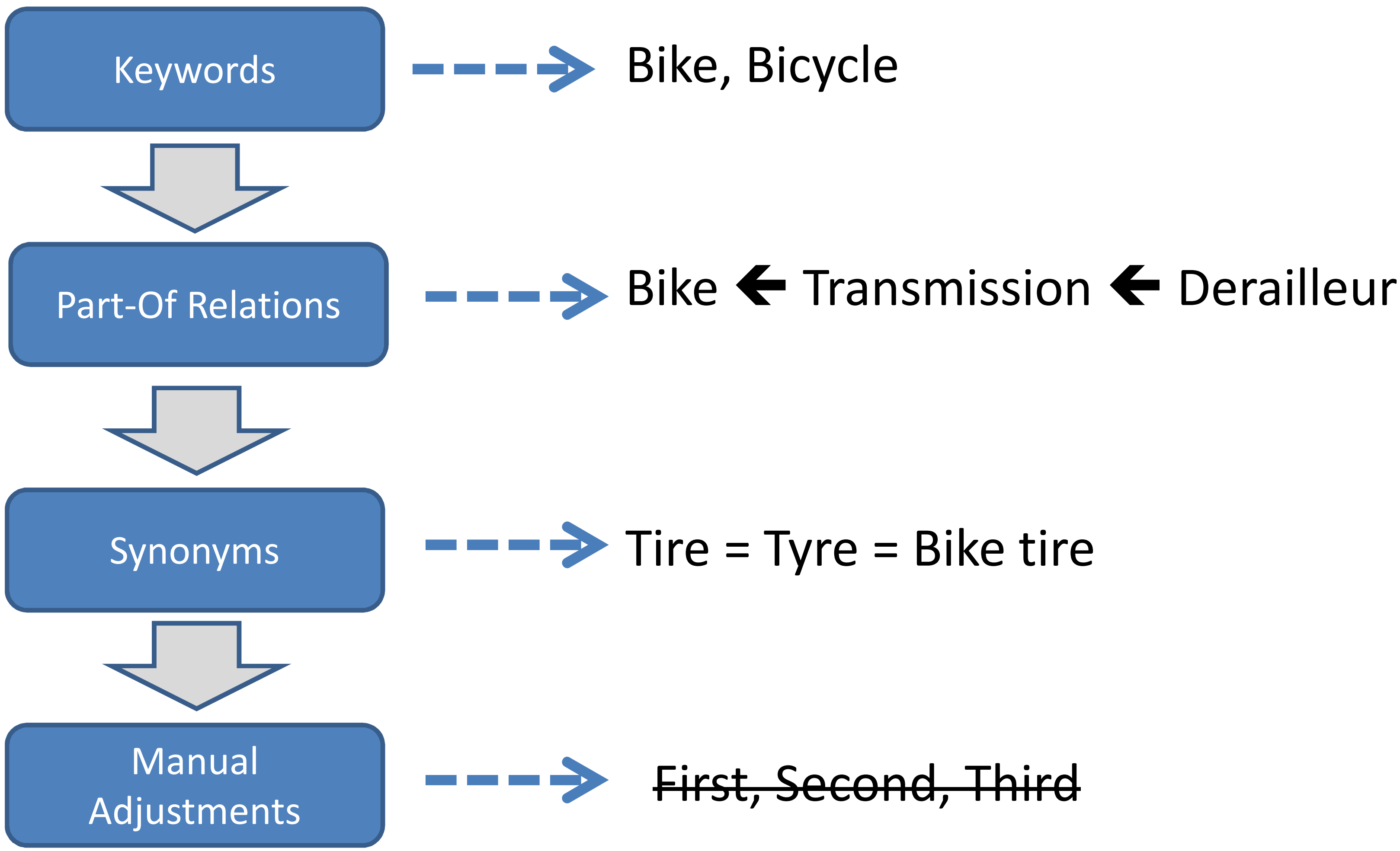

ROGET'S THESAURUS

Descriptive Non-Photorealistic Rendering



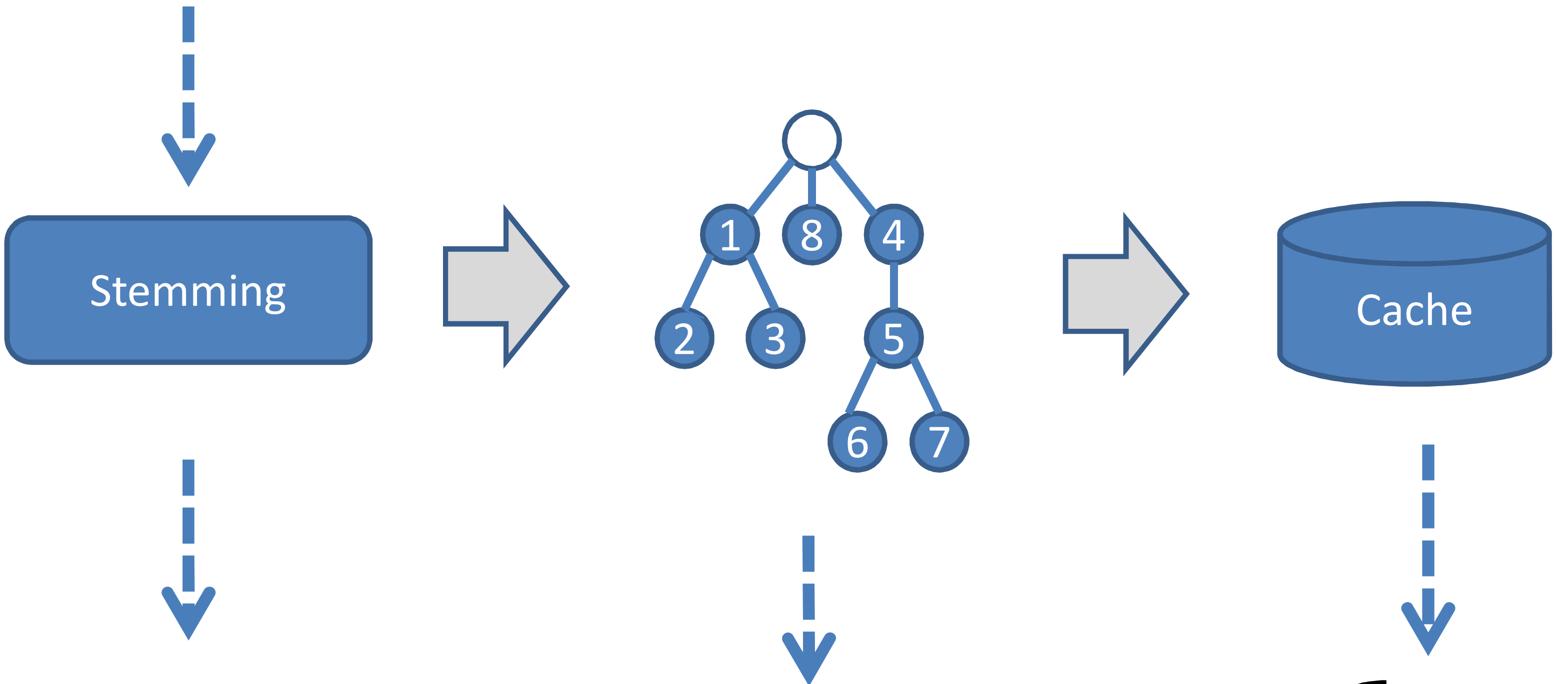
M. Chang and C. Collins, "Exploring Entities in Text with Descriptive Non-photorealistic Rendering," in *Proc. of the 2013 IEEE Pacific Visualization Symposium (PACIFICVIS '13)*, 2013.

Ontology Generation



Entity Extraction

"**Brakes** failed going at 35 mph."

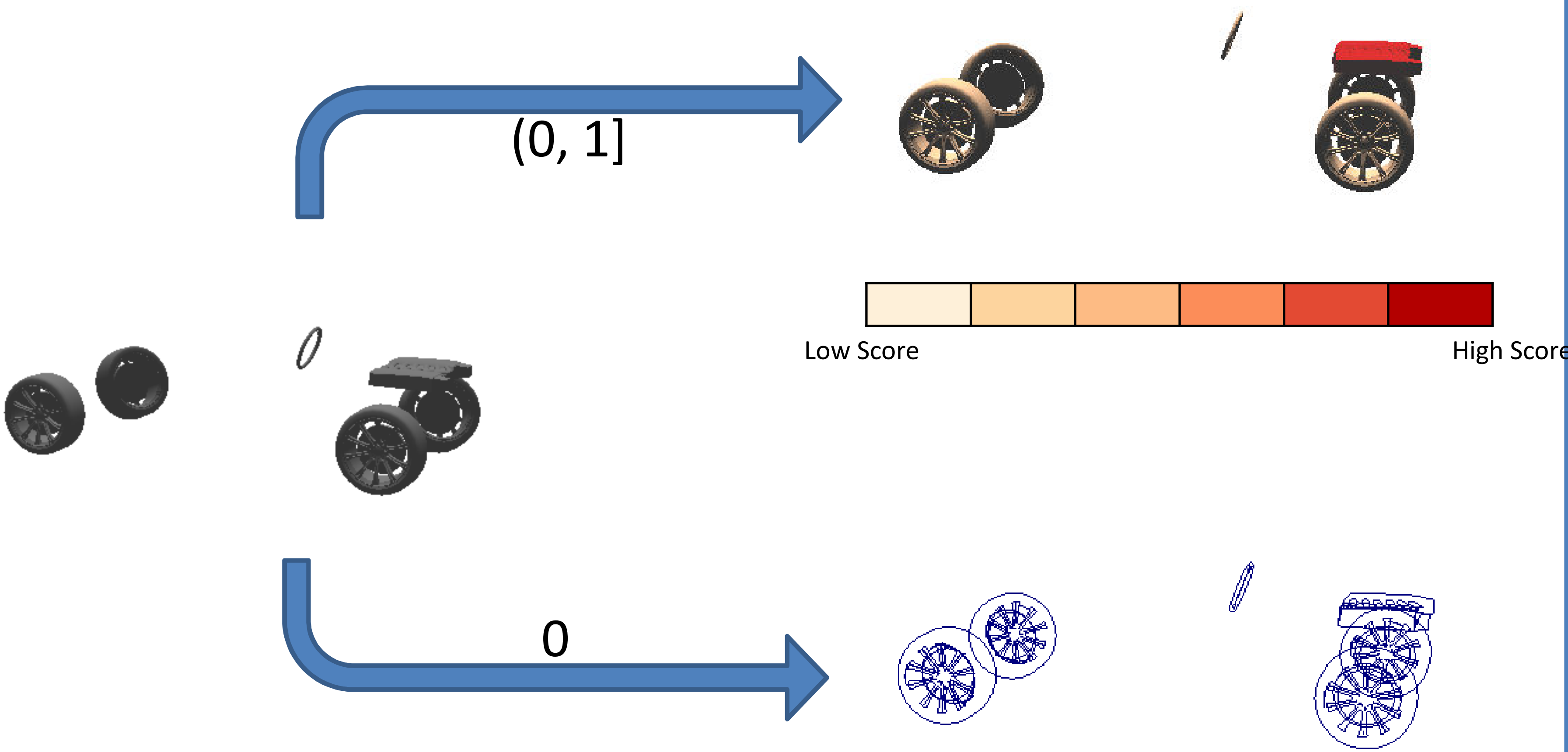


"brak"

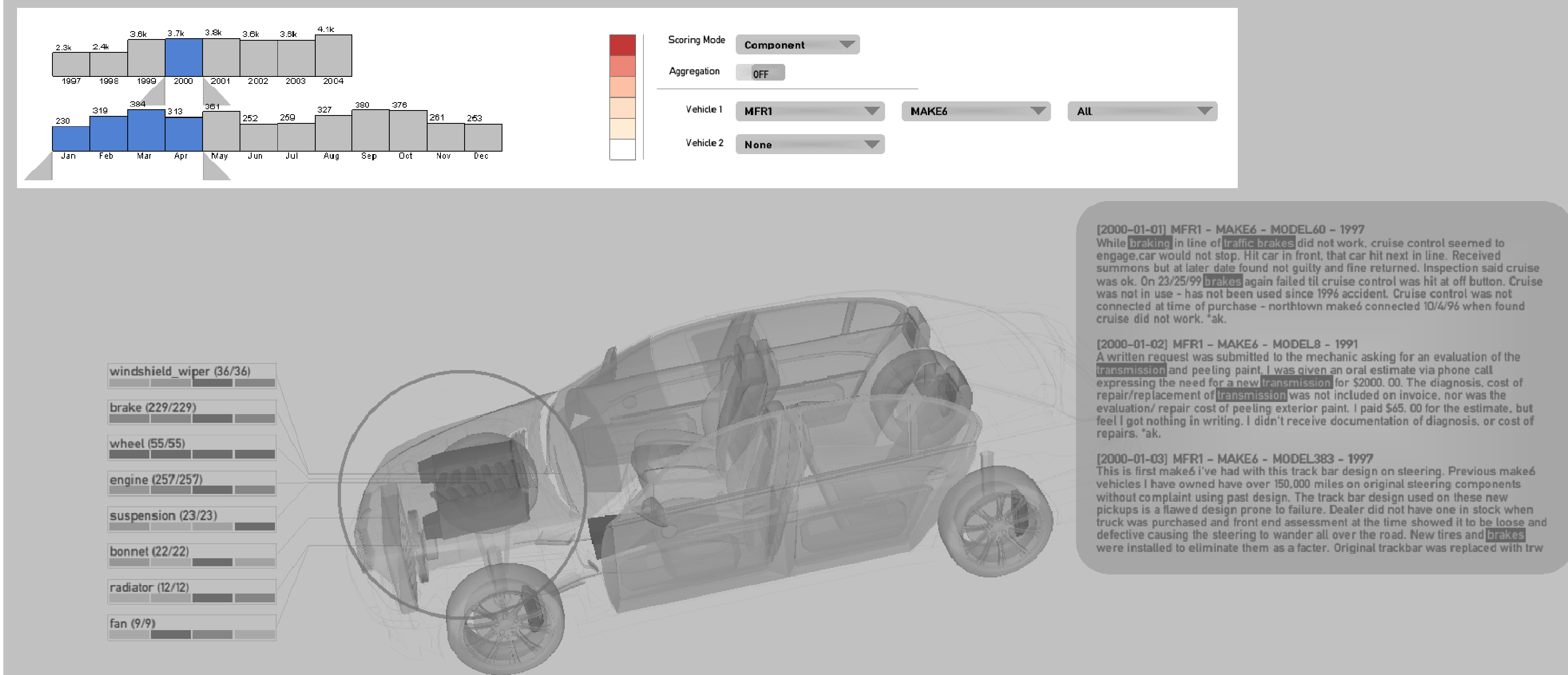
6: {brak, brak system…}

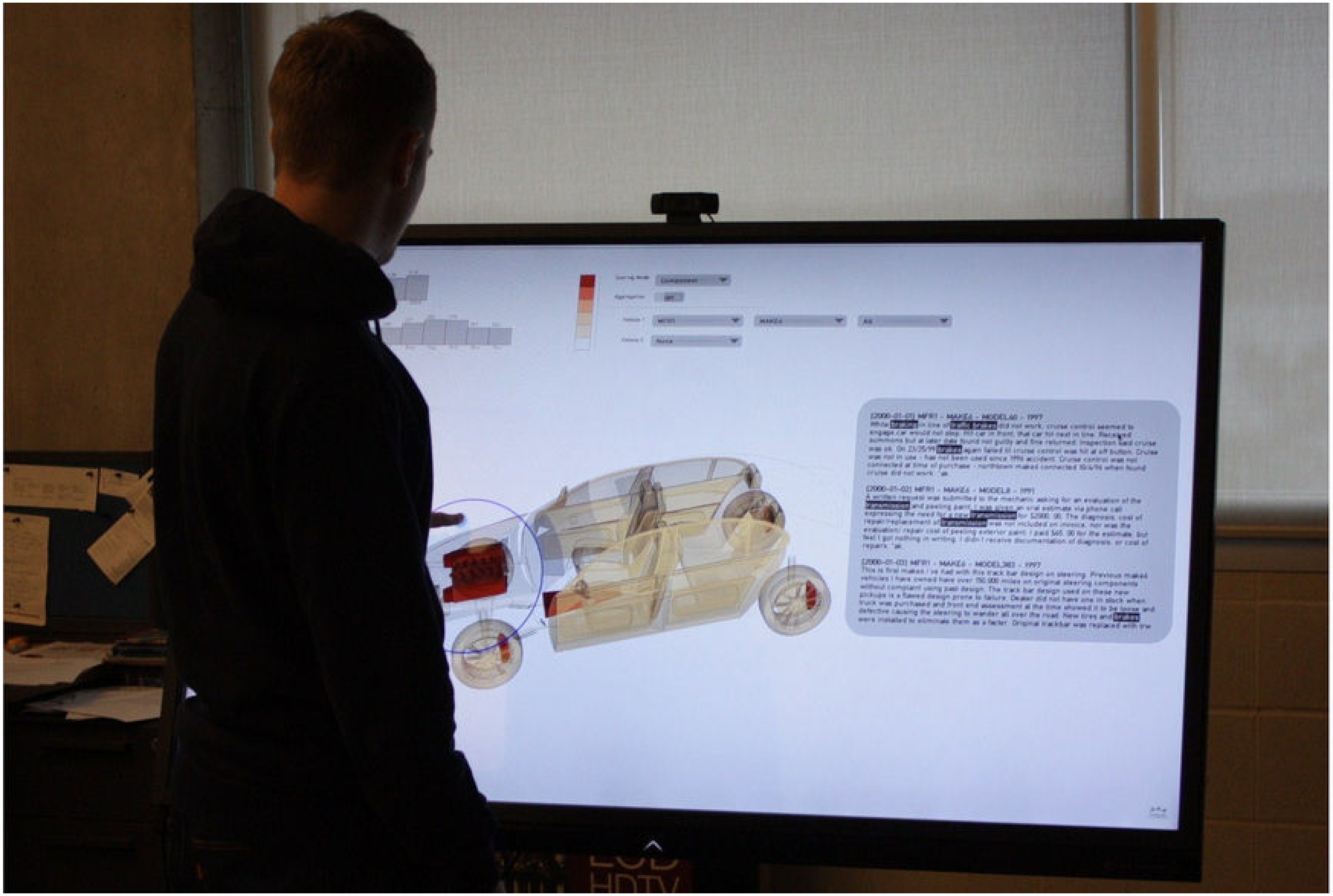
Document: 123
Entity: 6
Location: 0

Visual Representation

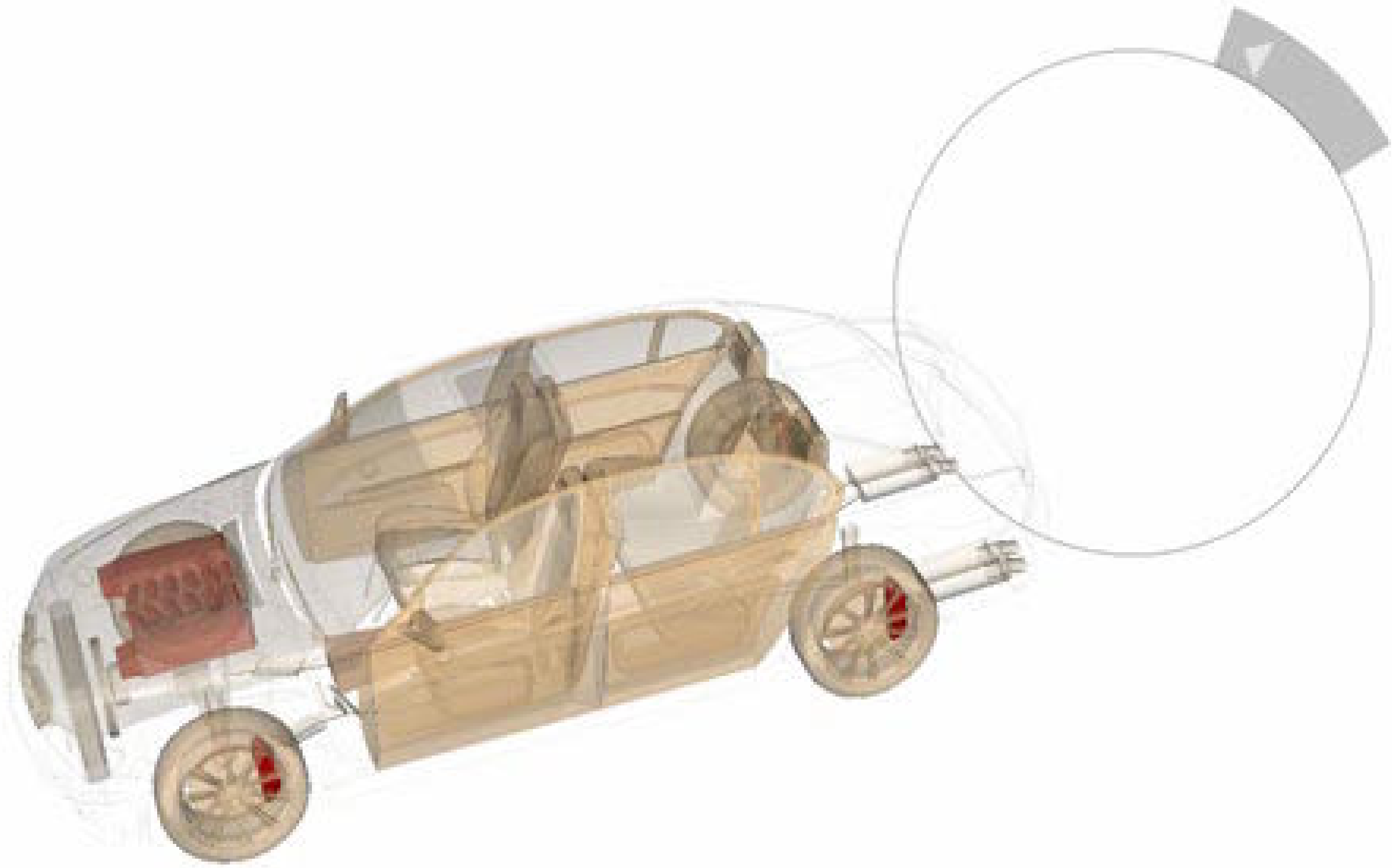


Main Interface





Exploration with Lens



Semantic Password Analysis

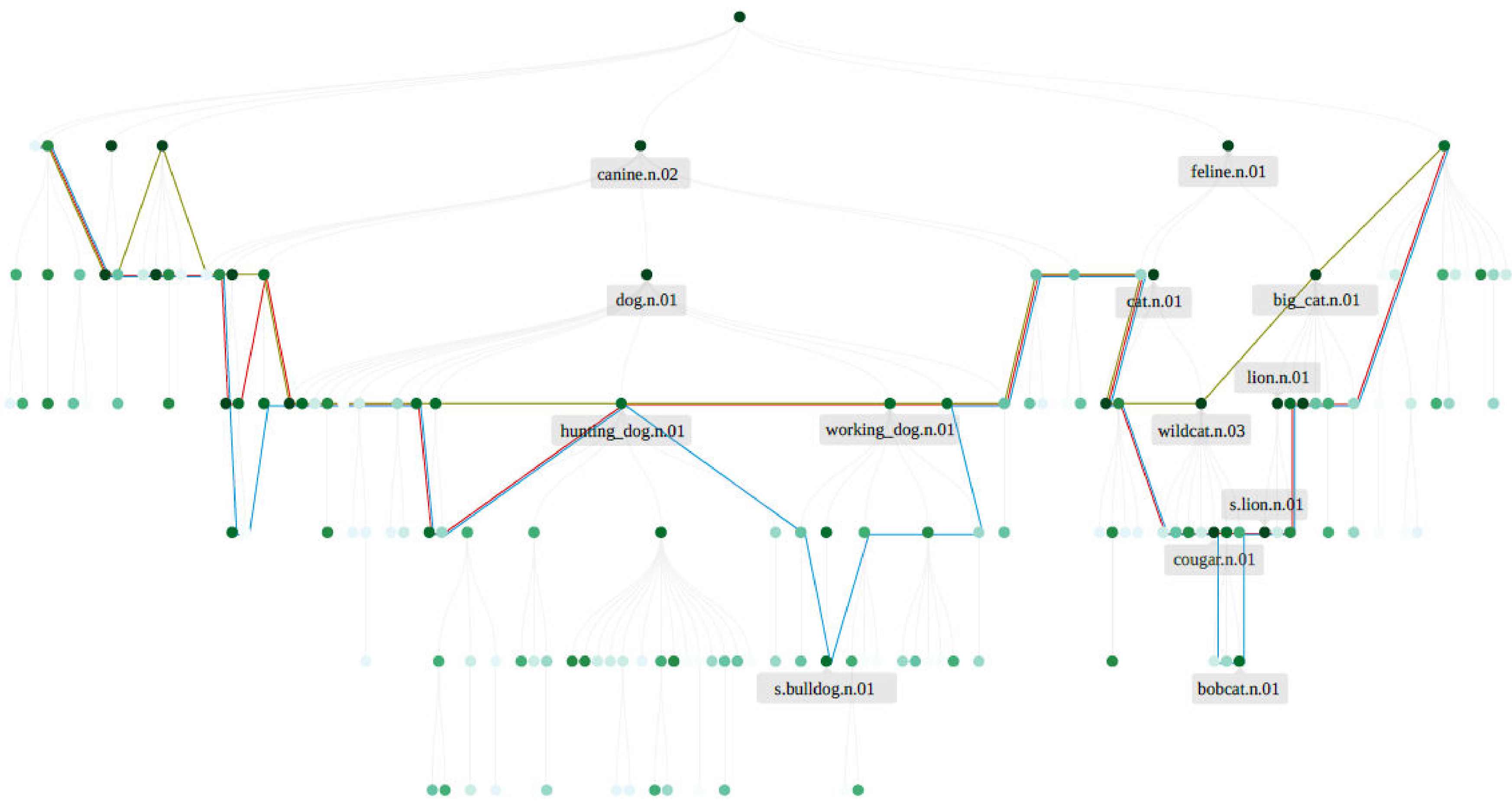
- What types of words do people use in their passwords?
- Do the patterns of word use represent security vulnerabilities?

R. Veras, C. Collins, and J. Thorpe, "On Semantic Patterns of Passwords and their Security Impact," *In Proceeding of the Network and Distributed System Security Symposium (NDSS'14)*, 2014.

- Extract words from 32 million passwords
- Categorize them
- Parse the results to find structure
- Create a password guessing system based on the model

Password	Segment	Semantic tag
hope87	hope	wish.v.01
hope87	87	number
serenity	serenity	trait.n.01
bishop5	bishop	status.n.01
bishop5	5	number
goblue0507	go	s.travel.v.01
goblue0507	blue	
goblue0507	507	number
looted	looted	take.v.21
drift21	drift	force.n.02
drift21	21	number
candysinger	candy	s.candy.n.01
candysinger	singer	musician.n.01
671soldier	671	number
671soldier	soldier	worker.n.01
bravo100	bravo	murderer.n.01
bravo100	100	number
egobrain	ego	pride.n.01
egobrain	brain	structure.n.04
pitcher9	pitcher	athlete.n.01
pitcher9	9	number
puppies	puppies	puppy.n.01
church	church	religion.n.02
‘ale‘8	‘	special
‘ale‘8	ale	alcohol.n.01
‘ale‘8	‘8	num+special

Appropriate Levels of Detail



Results

- Created best cracker on several measures, including percent correct guesses
- Designing strategies to help people make passwords more *semantically* secure – keep the meaning but lower the probability

Results

- Created best cracker on measure of % correct guesses
- Place names, male names very popular
- "Cute" animals more common:
  - Monkey, dogs, cats, dolphins
- Emotional verbs like "love" are common
  - People "love" male names 4x more often than female!
- Profanity is very common

thestar.com
GTA

News / GTA

Is there 'love' in your online passwords?

After analyzing 32 million leaked passwords, a team of researchers from the University of Ontario Institute of Technology has discovered that "love" is the most common password verb.

Tweet 55 g+1 1 reddit this!

By: Daniel Otis News Reporter, Published on Fri Feb 13 2015

People are putting a little too much "love" into their online passwords.

The New York Times | http://nyti.ms/1xqfNJL

MAGAZINE | NYT NOW

The Secret Life of Passwords

We despise them – yet we imbue them with our hopes and dreams, our dearest memories, our deepest meanings. They unlock much more than our accounts.

By IAN URBINA Video by LESLYE DAVIS

WordsEye.com

SENTIMENT VISUALIZATION

Sentiment Analysis

- Business intelligence:
  - Do people like my product/restaurant/movie/hotel?
  - Why or why not?
- Forensics and medicine:
  - State of mind analysis based on social media
- Emotional profiling / psycholinguistics
  - Understanding users -> individualization
  - Targeted advertising

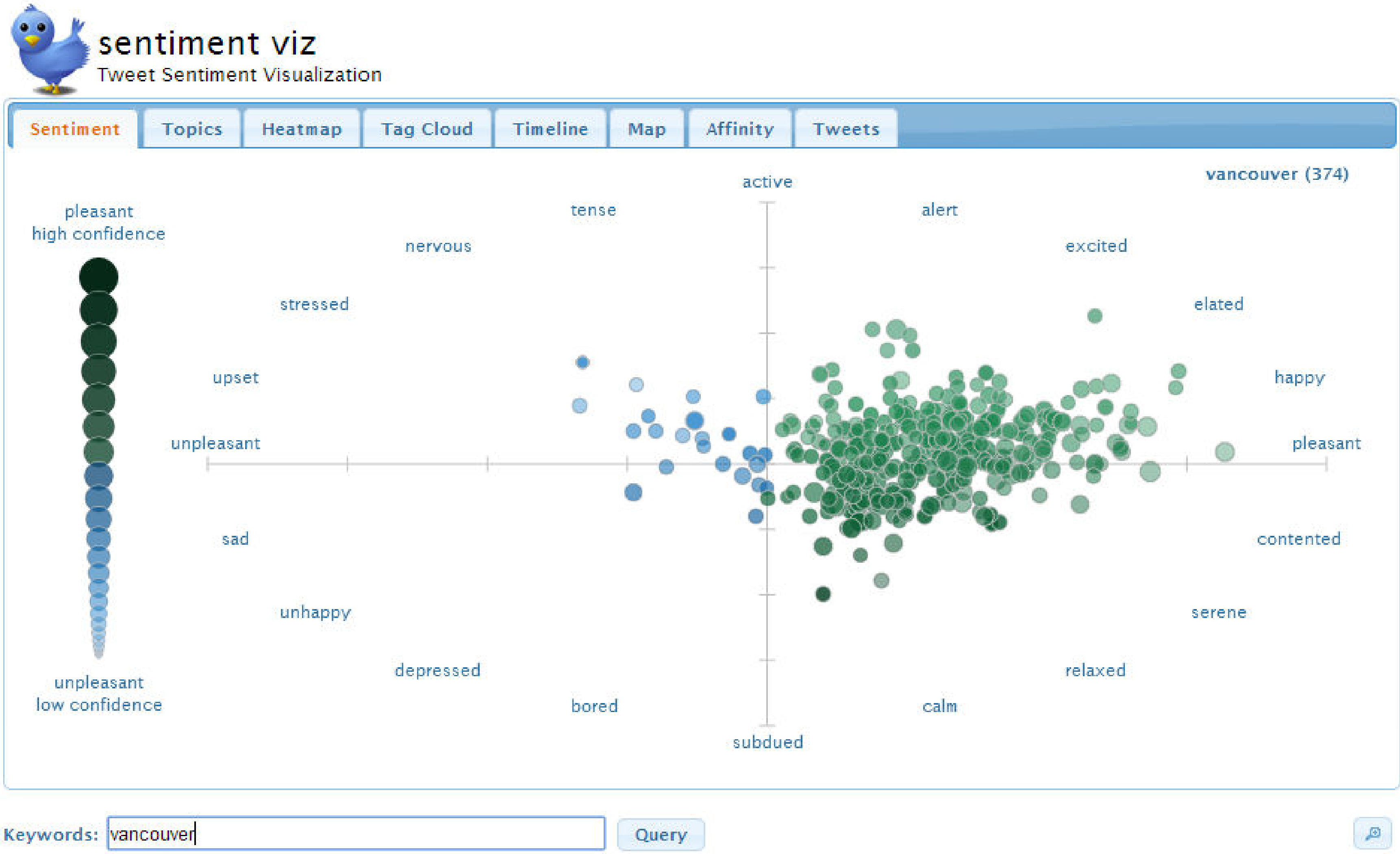
Sentiment Analysis

- Language Processing:
  - Stemming
  - POS Tagging
  - Dependency Parsing
  - Named Entity Detection
- Granularity:
  - Positive/negative/uncertain
  - 8+ emotions
  - Word, sentence, paragraph, document, corpus level

Resources and Datasets

- NRC Word-Emotion Lexicon:
  - Saif Mohammad, 2013
    http://www.saifmohammad.com/WebPages/ResearchInterests.html
- LIWC:
  - James Pennybaker et al., 2007:
    http://www.liwc.net/
- Opinion Mining Dataset:
  - Bing Liu, 2004—current
    http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Twitter Sentiment Viz



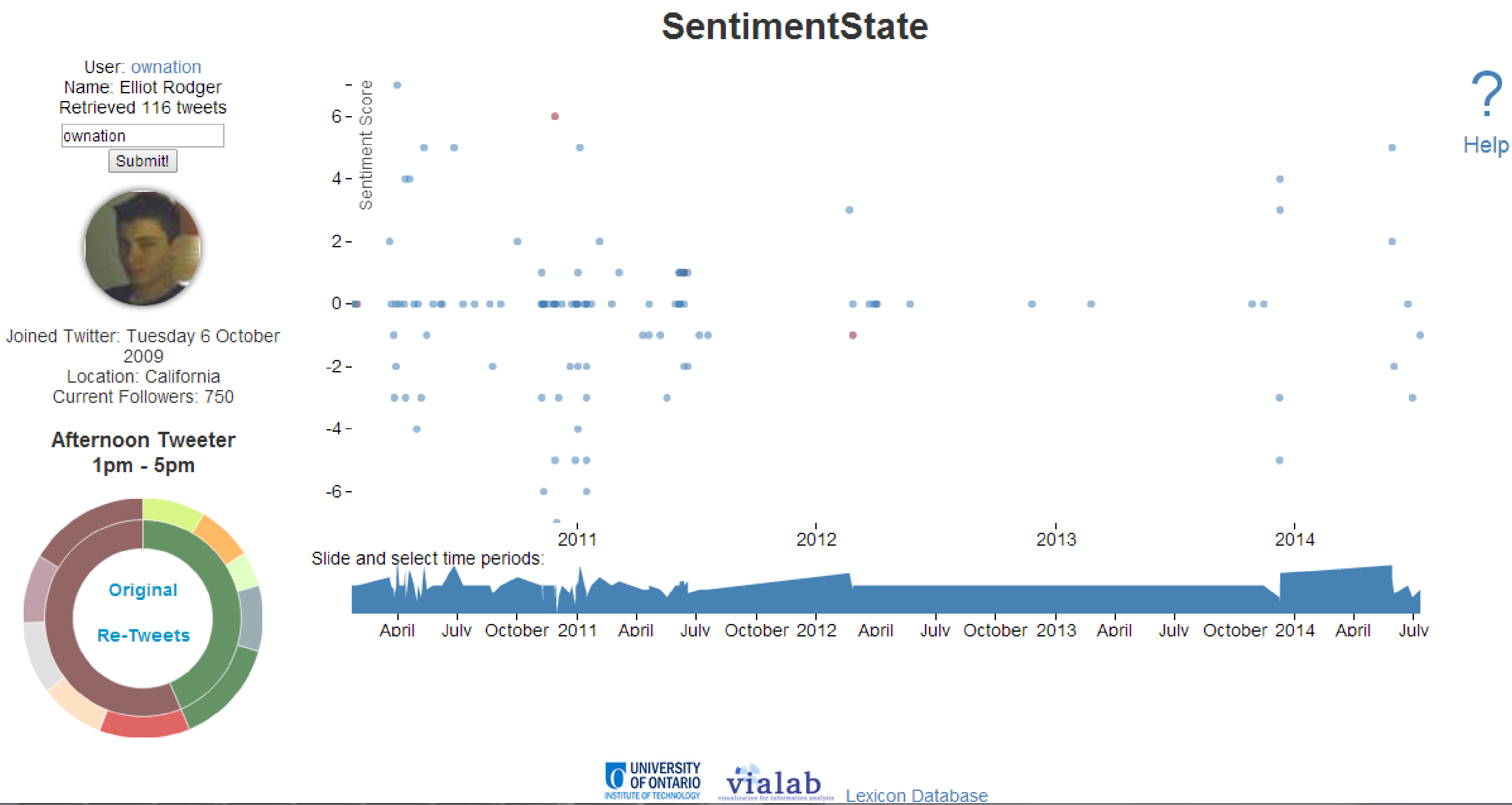
Healey and Ramaswamy, 2013. http://www.csc.ncsu.edu/faculty/healey/tweet_viz/

SentimentState

- Tweets over time, categorized using an emotion lexicon
- Examine Tweets in context, filter based on time and emotions

Scantlebury and Collins, 2014. http://vialab.science.uoit.ca/sentimentstate

SentimentState



Scantlebury and Collins, 2014. http://vialab.science.uoit.ca/sentimentstate

This movie was actually neither that funny, nor super witty.

This movie was actually neither that funny, nor super witty.

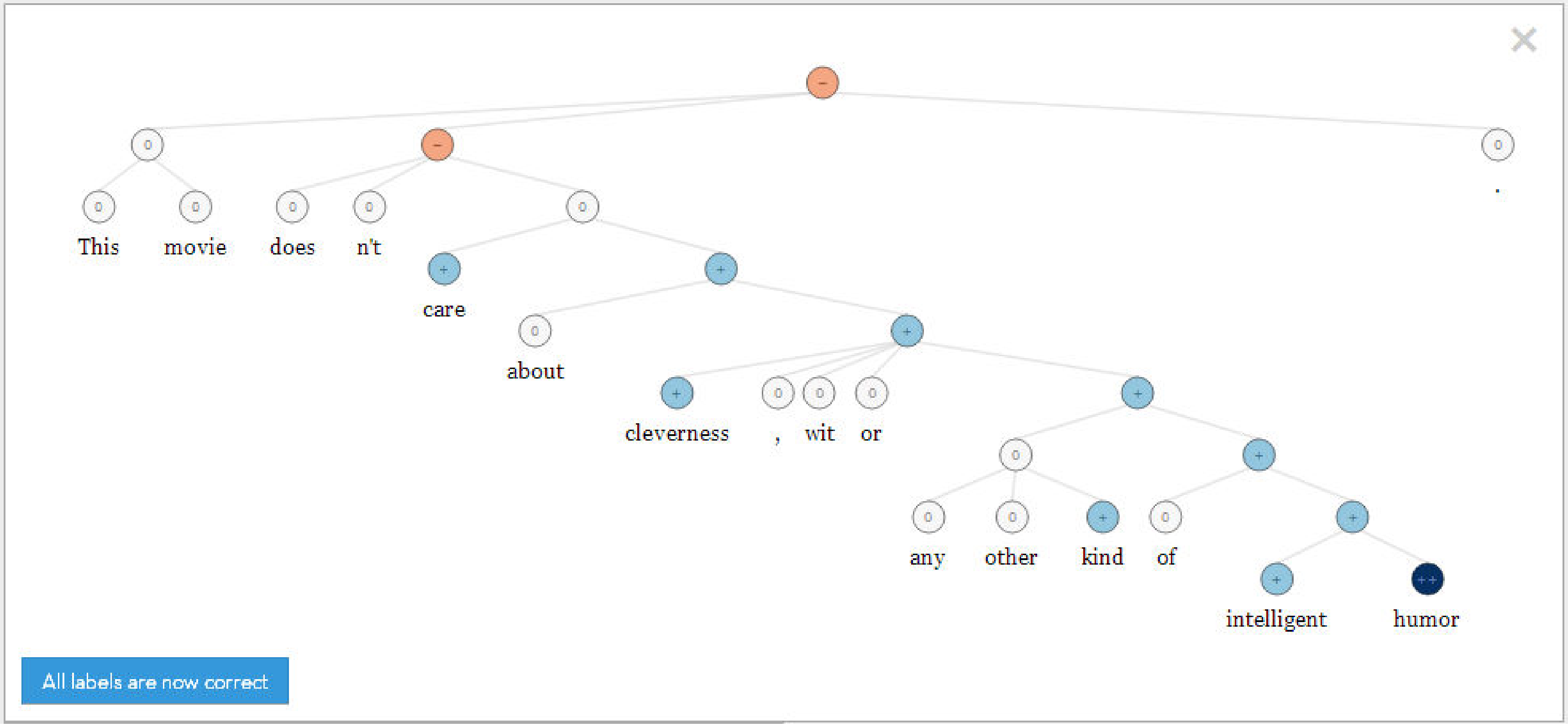
Stanford Sentiment Parser

- Recursive neural network built on top of grammatical structures
- Trained on Stanford Sentiment Treebank
  – Parse trees labelled with sentiment scores
  – Crowed-sourced and editable

Socher et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).

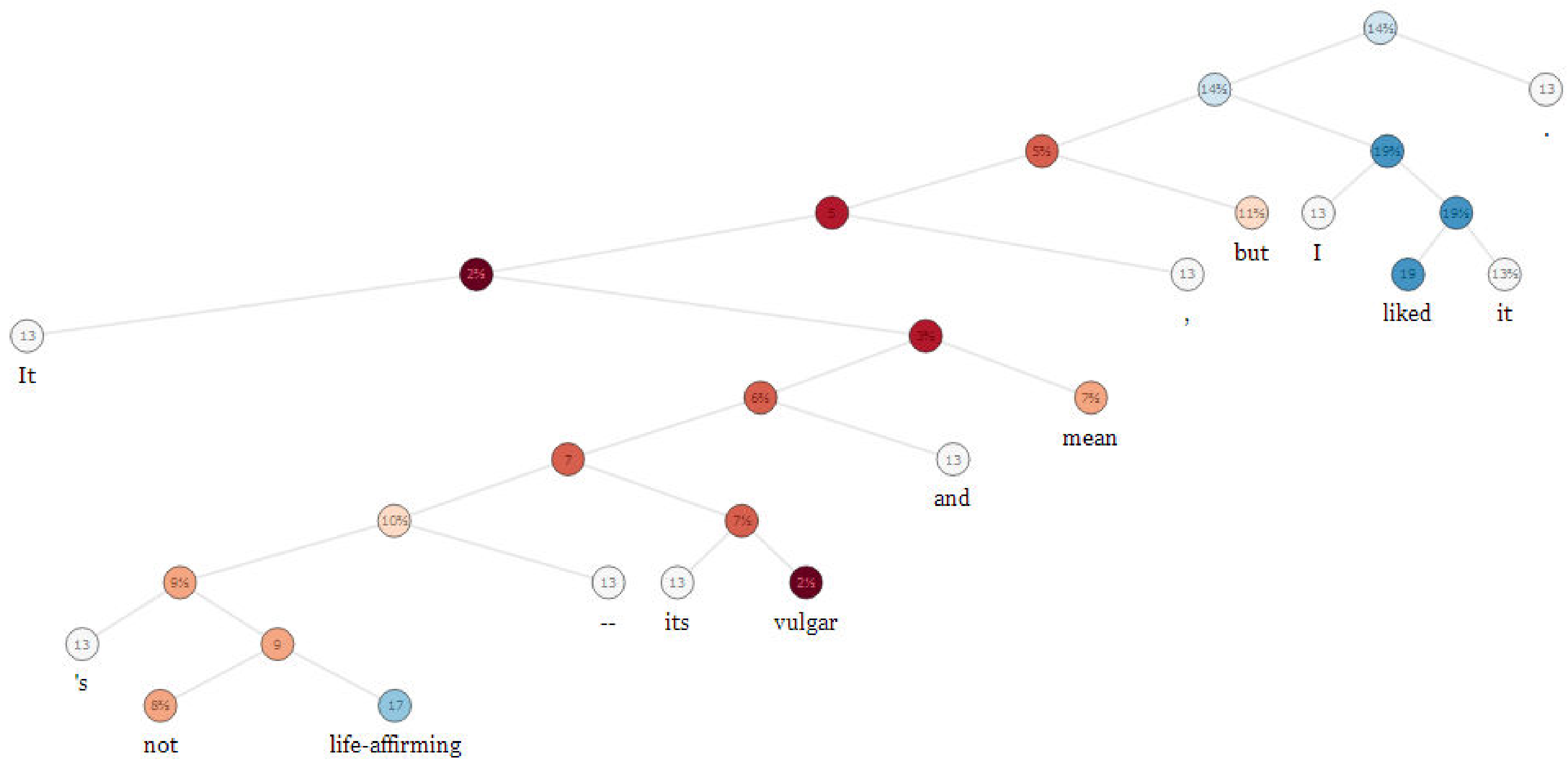
Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: very negative, negative, neutral, positive, and very positive.



Parsing is Needed!

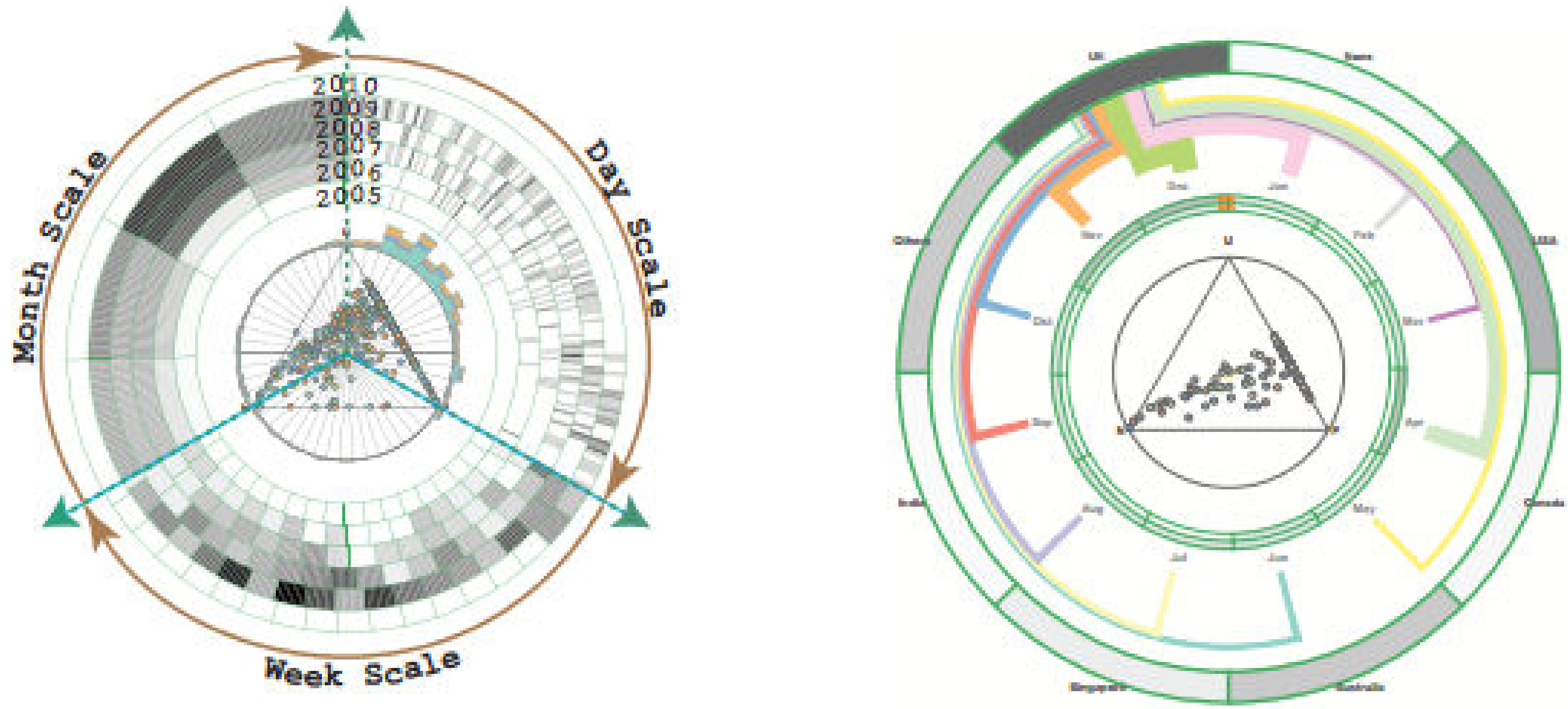
- Stanford Sentiment Treebank:
  - http://nlp.stanford.edu/sentiment/treebank.html



Challenges

- Word-counting techniques are fast, but inaccurate
  - Sarcasm, quotes, metaphorical language
- Accurate methods are slow/difficult to run over big data

Opinion Seer



(a) (b)

Yingcai Wu et al. 2010. OpinionSeer: Interactive Visualization of Hotel Customer Feedback. *IEEE Transactions on Visualization and Computer Graphics* 16 (6), November 2010.

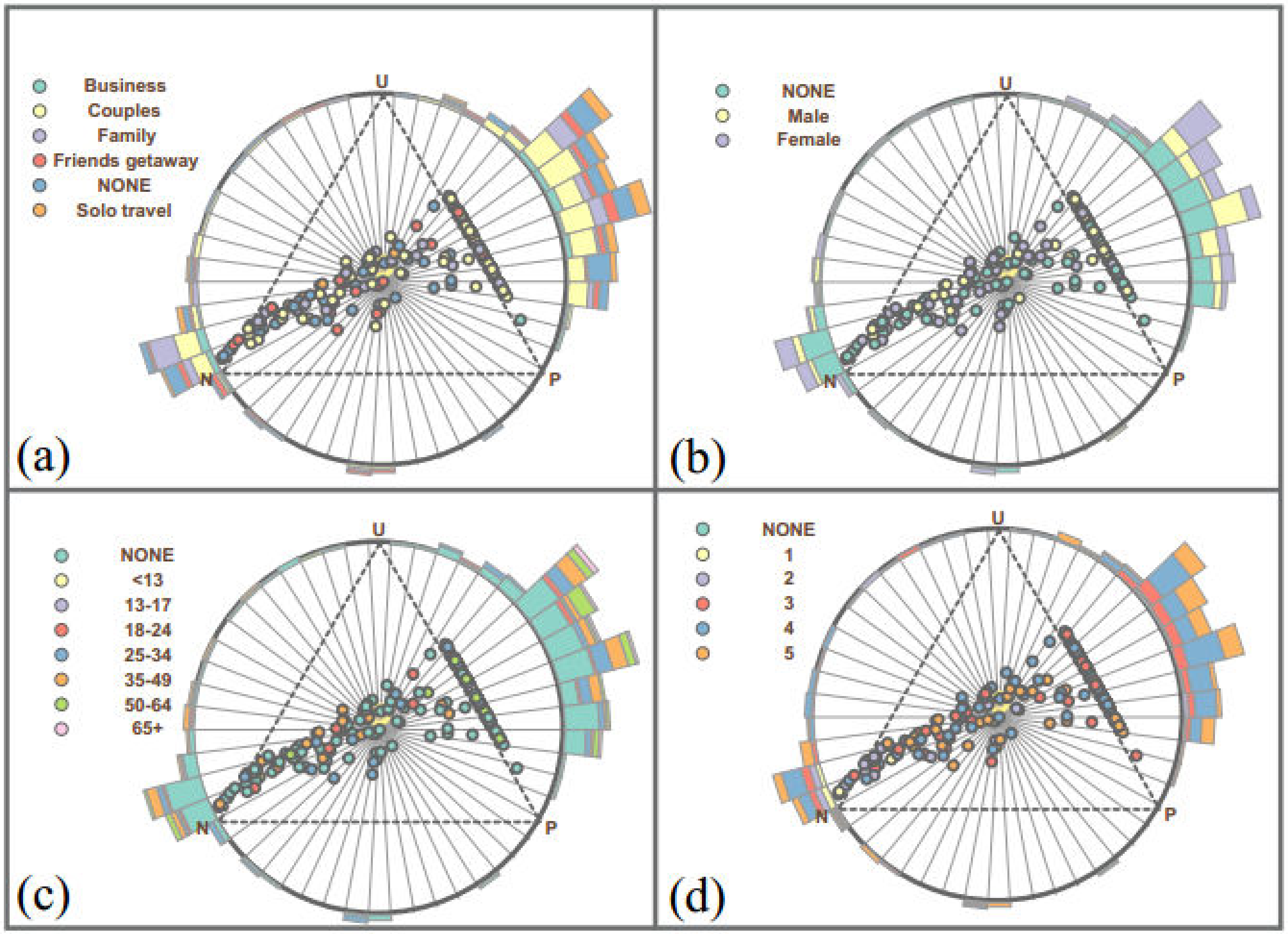


Fig. 8. OpinionSeer results showing how customer opinions are correlated with trip type, gender, age range, and ratings.

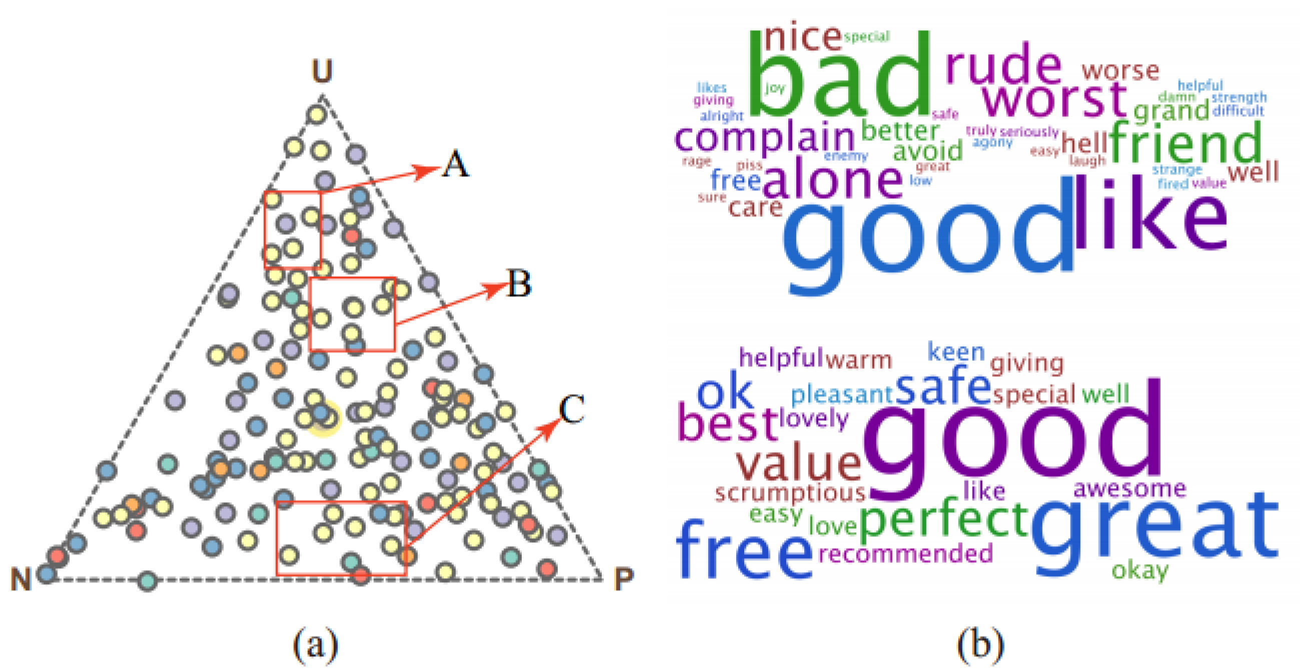
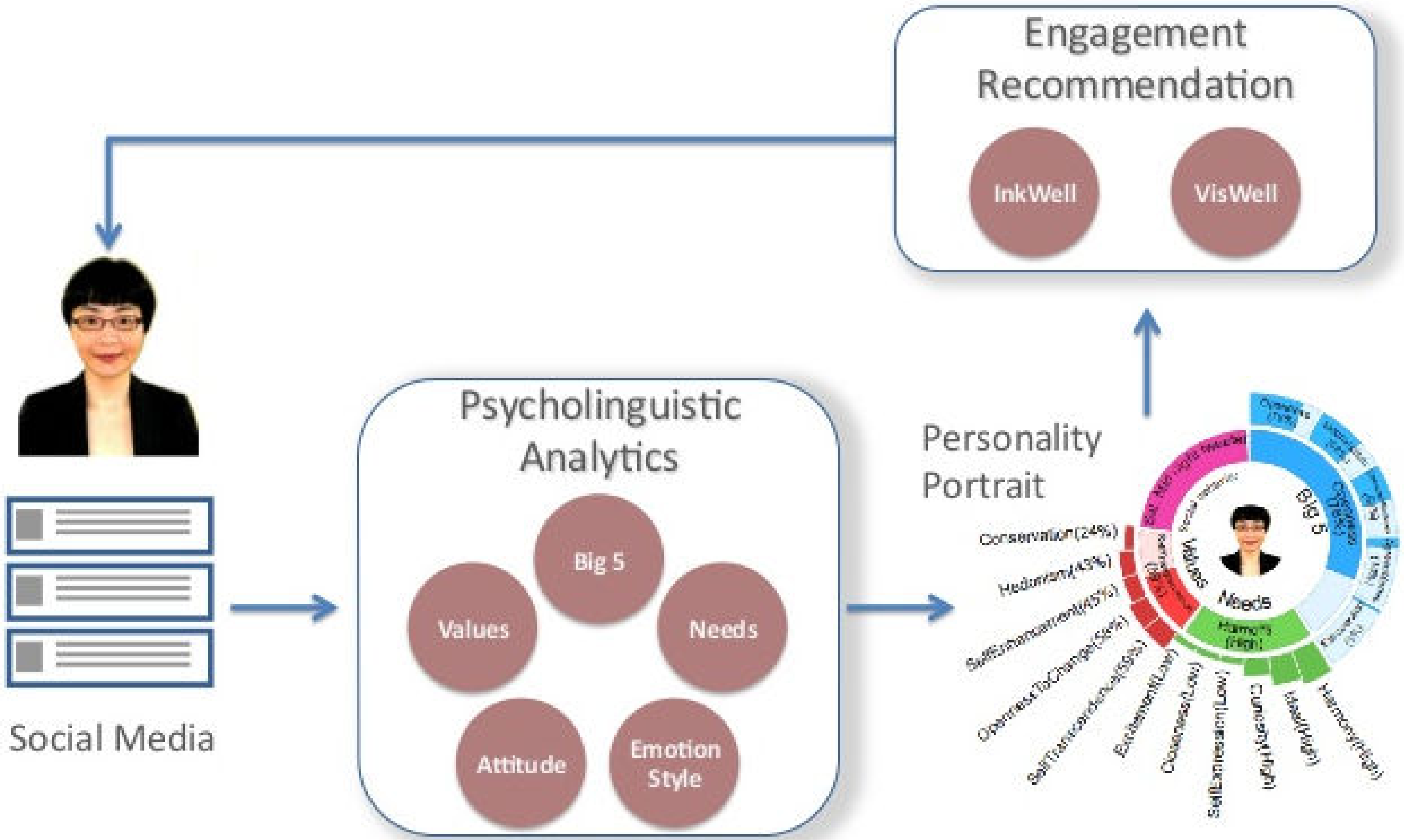


Fig. 7. (a) An opinion triangle where three regions A, B, and C are selected; (b) Top and bottom: two tag clouds of the opinion words associated with Region A and B in (a), respectively.

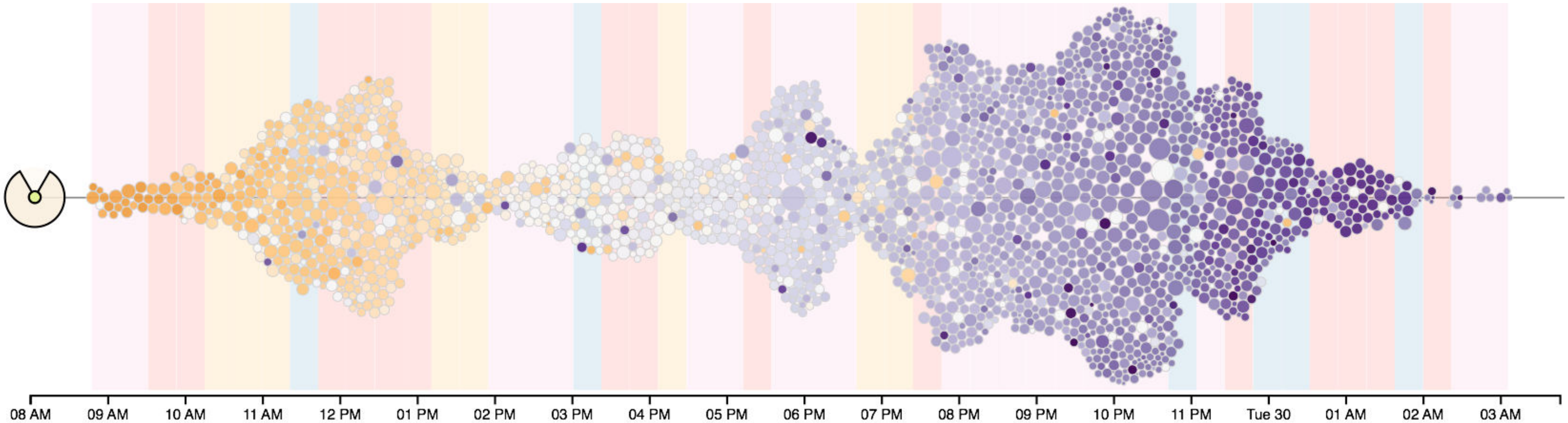
IBM System U



Michelle Zhou, System U: Computational Discovery of Personality Traits from Social Media for Individualized Experience, 2014.

Big 5
Social behavi
Values
Needs
Openness (55%)
Conscientiousne (55%)
Extraversion (62%)
Agreeableness (54%)
Neuroticism (41%)
Afternoon tweeter (EST) (10%)
Wed. Afternoon tweeter
Wed. twe (21%)
Self-expression (100%)
Adventurousness (58
Artistic interests (71
Emotionality (55%)
Imagination (55%)
Intellect (51%)
Liberalism (42%)
Achievement striving (83%)
Cautiousness (50%)
Dutifulness (53%)
Orderliness (39%)
Self-discipline (44%)
Self-efficacy (55%)
Activity level (70%)
Assertiveness (81%)
Cheerfulness (65%)
Excitement-seeking (65%)
Friendliness (45%)
Gregariousness (75%)
Altruism (55%)
Cooperation (62%)
Modesty (54%)
Morality (45%)
Sympathy (61%)
Trust (47%)
Anger (40%)
Anxiety (53%)
Depression (45%)
Immoderation (36%)
Self-consciousness (40%)
Vulnerability (50%)
Challenge (97%)
Closeness (1%)
Curiosity (1%)
Excitement (1%)
Harmony (1%)
Ideal (96%)
Liberty (1%)
Love (3%)
Practicality (3%)
Self-expression (100%)
Stability (1%)
Structure (37%)
Conservation (27%)
Openness to change (41%)
Hedonism (12%)
Self-enhancement (95%)
Self-transcendence (40%)
Sunday (9%)
Monday (17%)
Tuesday (16%)
Wednesday (21%)
Thursday (18%)
Friday (13%)
Saturday (7%)

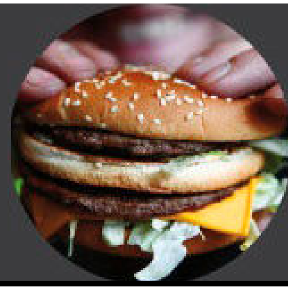
#FluxFlow



Jian Zhao et al., 2014. #FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media. *IEEE Transactions on Visualization and Computer Graphics* 20 (12), December, 2014.

London Riots 2011

Rioters attack London and release animals



Rioters cook their own food in McDonald's

Police 'beat a 16-year-old girl'

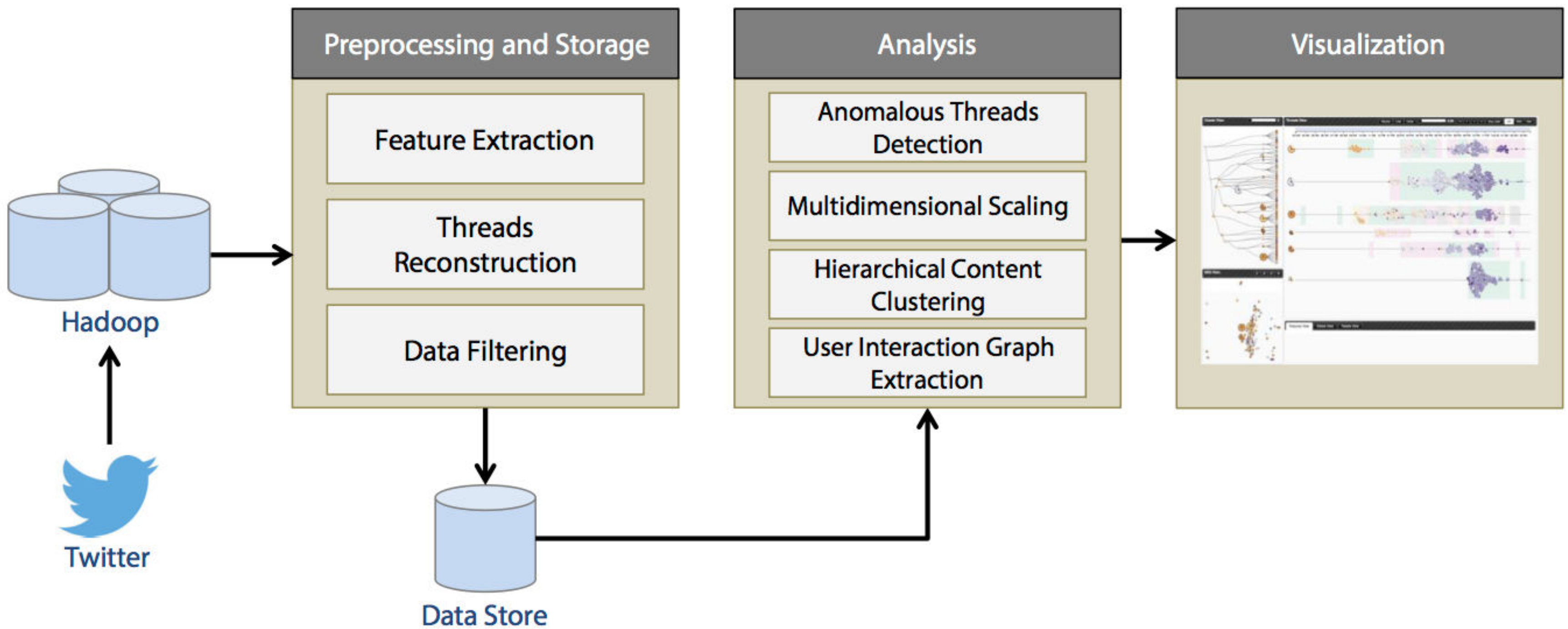
London Eye set on fire

Rioters attack a children's hospital

Army deployed in Bank

Miss Selfridge set on fire

http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter



Preprocessing and Storage
Feature Extraction
Threads Reconstruction
Data Filtering
Analysis
Anomalous Threads Detection
Multidimensional Scaling
Hierarchical Content Clustering
User Interaction Graph Extraction
Visualization
Hadoop
Twitter
Data Store

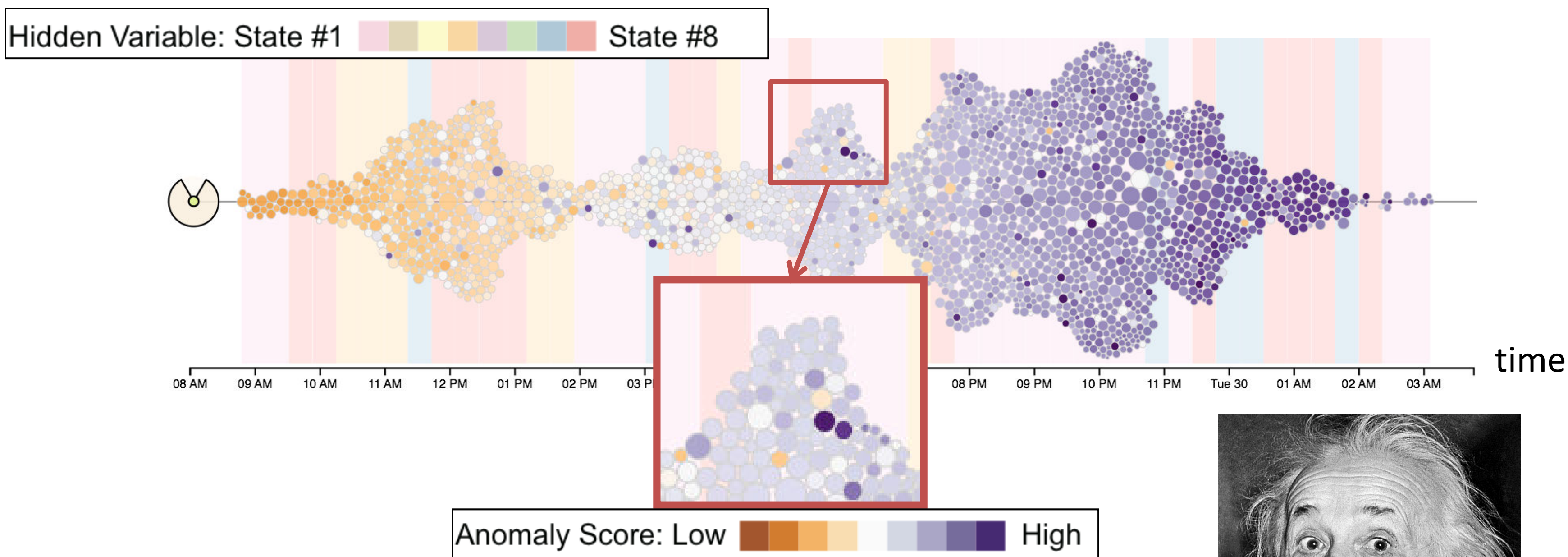
Abnormal RT Threads Detection

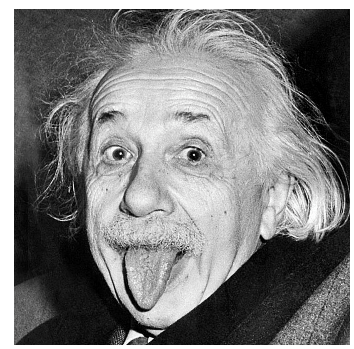
- One-class conditional random fields (OCCRF) [Song et al. 2013]
  - *One-class nature,* i.e., little is known about true anomalies
  - *High time-dependency,* due to the RT mechanism in Twitter
- Input features (239 dimensions in total)
  - User profile features → followers count, …
  - User network features → in/out-degrees in a user's ego-

Examples:
- Threads expressing unusual temporal patterns of user volumes.
- Threads containing many groups of users who do not normally tweet each other.

  - Tweet content features → emotional keywords, …

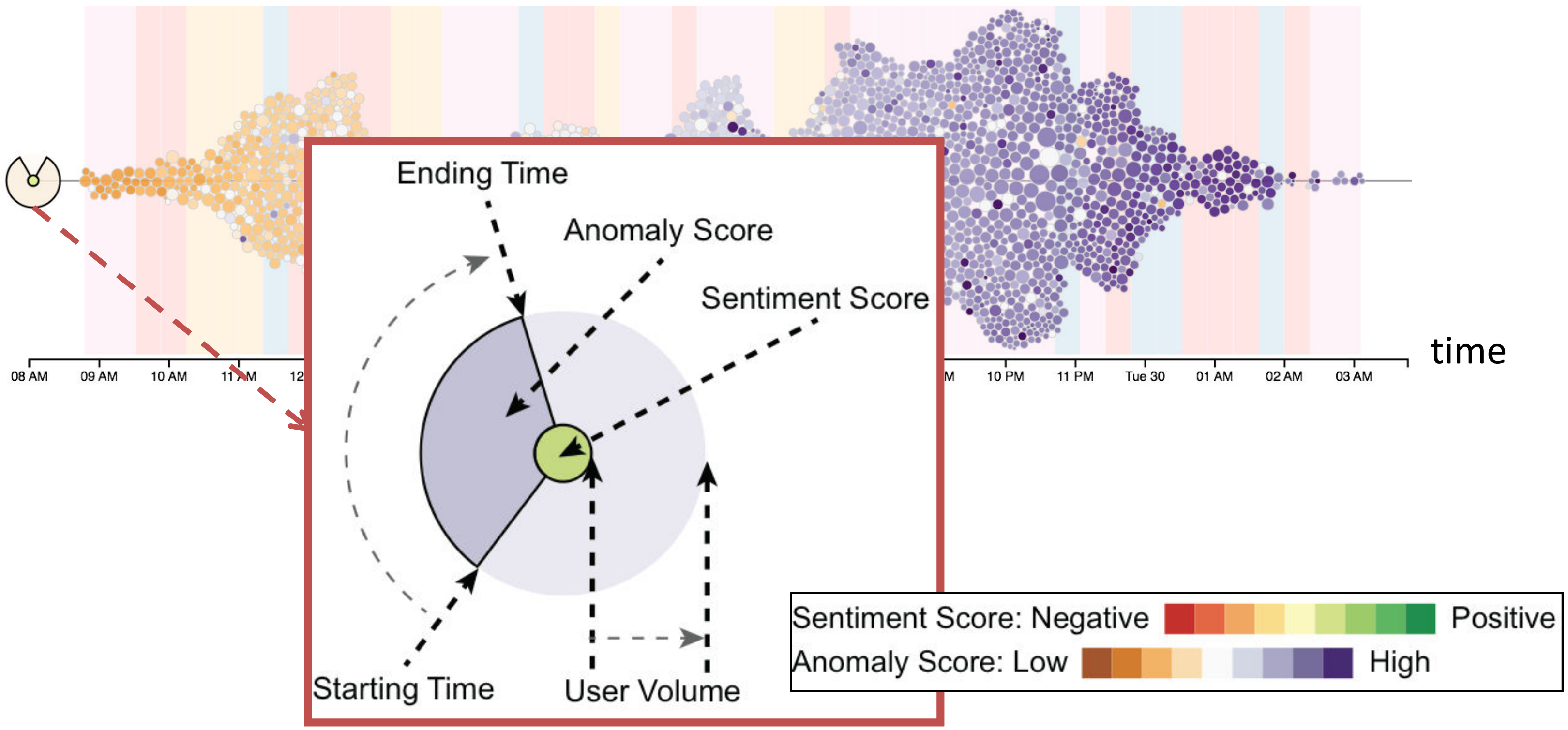
RT Thread Timeline





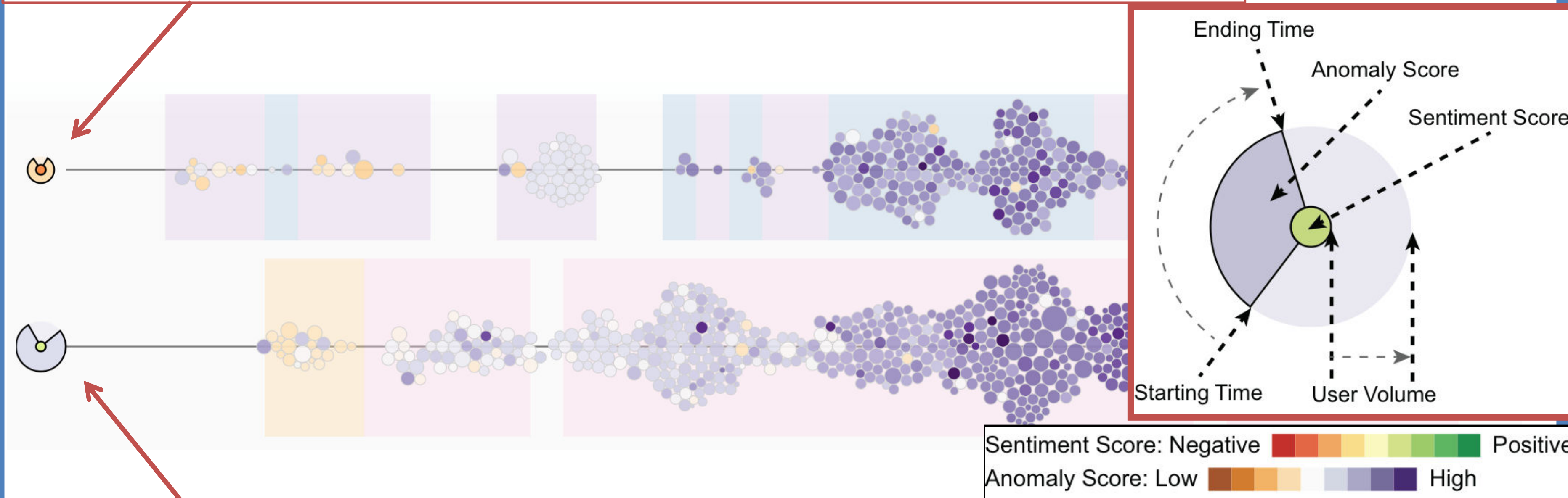
"It [FluxFlow] has the wave-particle duality, as the light."
---- Albert Einstein

RT Thread Glyph

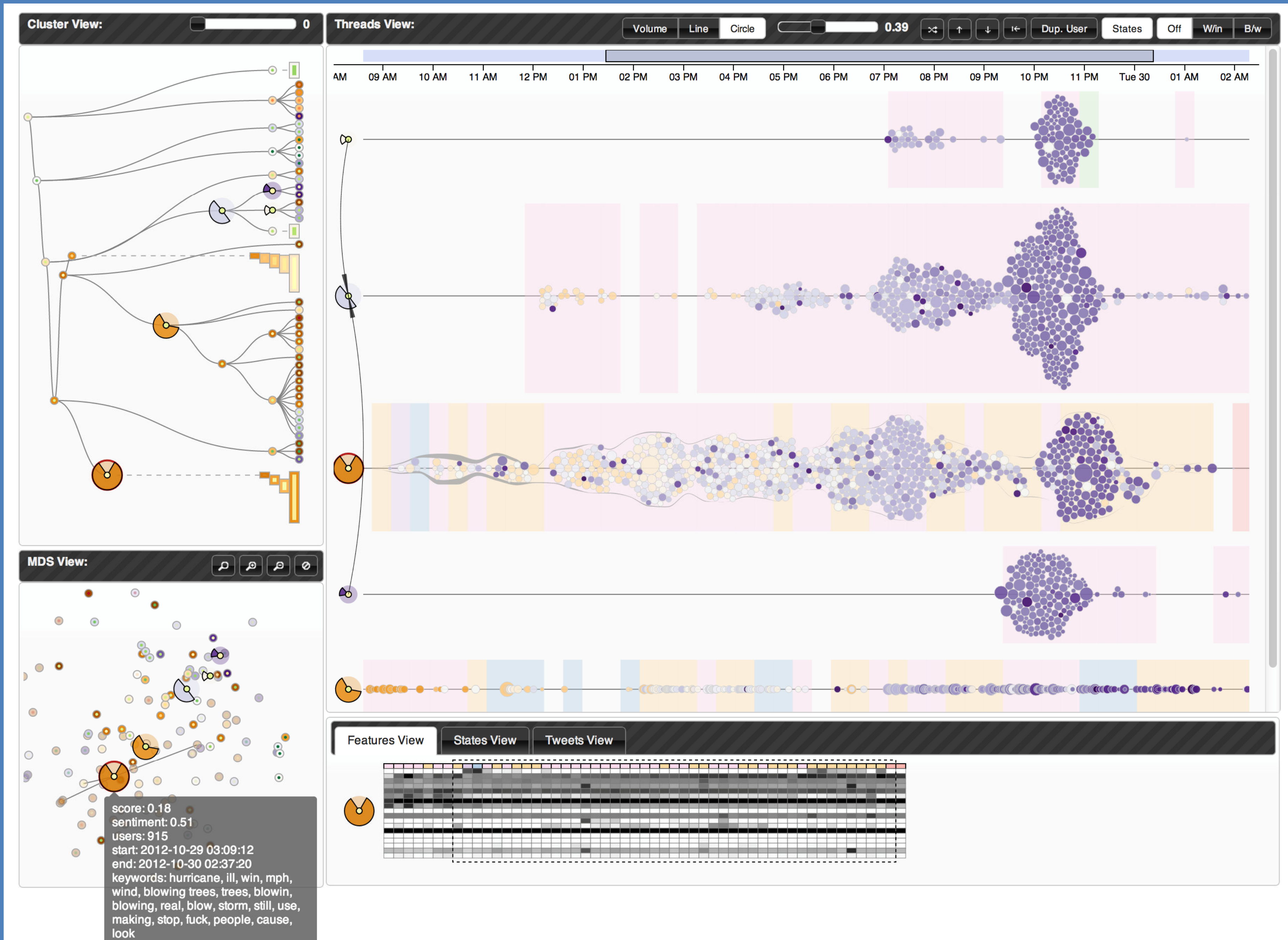


Examples

Lower anomaly score; Lower user volume; Negative sentiment; Started earlier and ended later



Higher anomaly score; Higher user volume; Positive sentiment; Started later and ended earlier

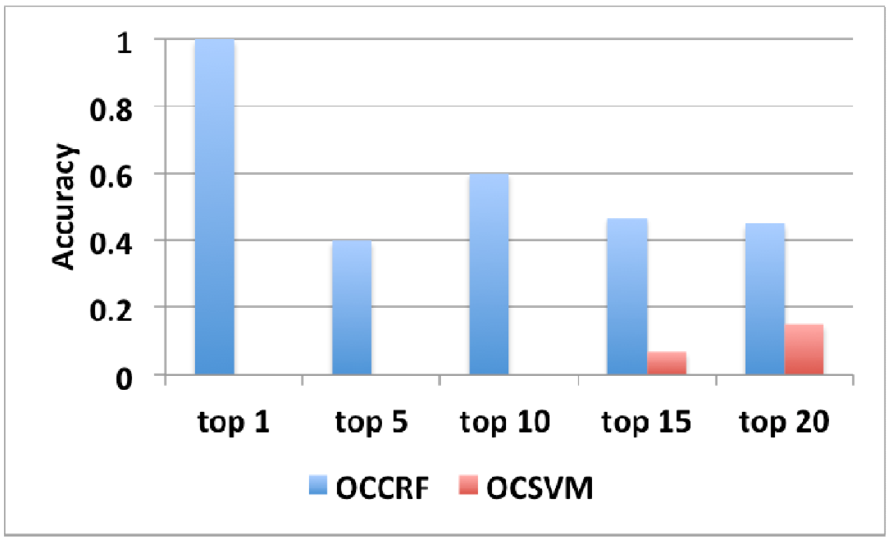

Cluster View:
0
Threads View:
Volume
Line
Circle
0.39
Dup. User
States
Off
W/in
B/w
AM
09 AM
10 AM
11 AM
12 PM
01 PM
02 PM
03 PM
04 PM
05 PM
06 PM
07 PM
08 PM
09 PM
10 PM
11 PM
Tue 30
01 AM
02 AM
MDS View:
score: 0.18
sentiment: 0.51
users: 915
start: 2012-10-29 03:09:12
end: 2012-10-30 02:37:20
keywords: hurricane, ill, win, mph, wind, blowing trees, trees, blowin, blowing, real, blow, storm, still, use, making, stop, fuck, people, cause, look
Features View
States View
Tweets View

Evaluation of Model Performance

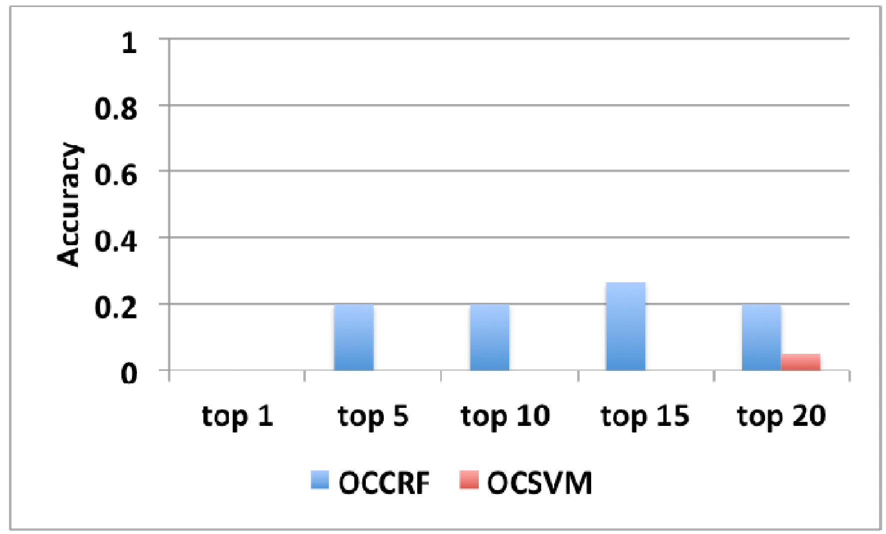
- Datasets: two Twitter feeds captured during:
  - Hurricane Sandy (Oct 29-30, 2012; 52 million tweets)
  - Boston Marathon Bombing (Apr 15-19, 2013; 242 million tweets)
- Techniques
  - Our OCCRF-based model
  - One-Class SVM (OCSVM) model [Scholkopf et al., 2001]
- Ground truth (in the top 500 abnormal threads ranked by models)
  - Manually label if a retweeting thread is misinformation by 3 annotators based on reports after the events

Model Comparison Results

Hurricane Sandy



Boston Bombing



Interactive Visualization is critical!

Accuracies in correctly detecting rumors in the top-N retweeting threads ranked by the OCCRF and OCSVM.

TOPIC MODELING VISUALIZATIONS

10 Topics from Psych Review

'memory recognition retrieval recall items item list'

'representations representation bias single scale known relation'

'information processing action levels automatic components controlled'

'theory theories predictions account explain formation esteem'

'visual attention brain mechanism selection color attentional'

'speech system target motor masking neural relative'

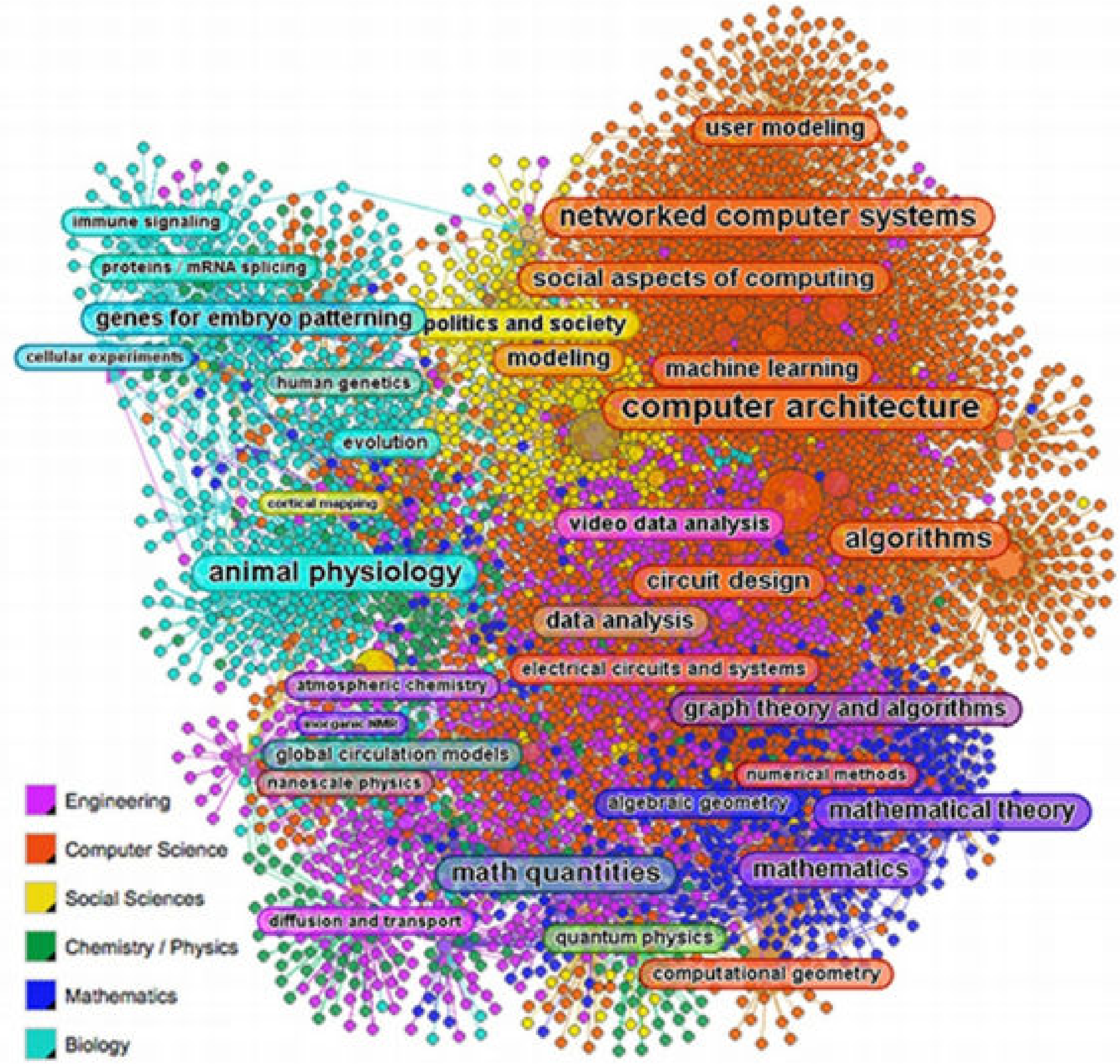
'network semantic ability test lexical predictions normal'

'states related emotional positive primary arousal motivation'

'control conditioning responses conditions avoidance procedures suggested'

'stimulus effects shown concepts trial free generalization'

http://psiexp.ss.uci.edu/research/programs_data/exampleLDA1.html



Gretarsson et al. 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Trans. Intell. Syst. Technol.* 3, 2, Article 23 (February 2012).

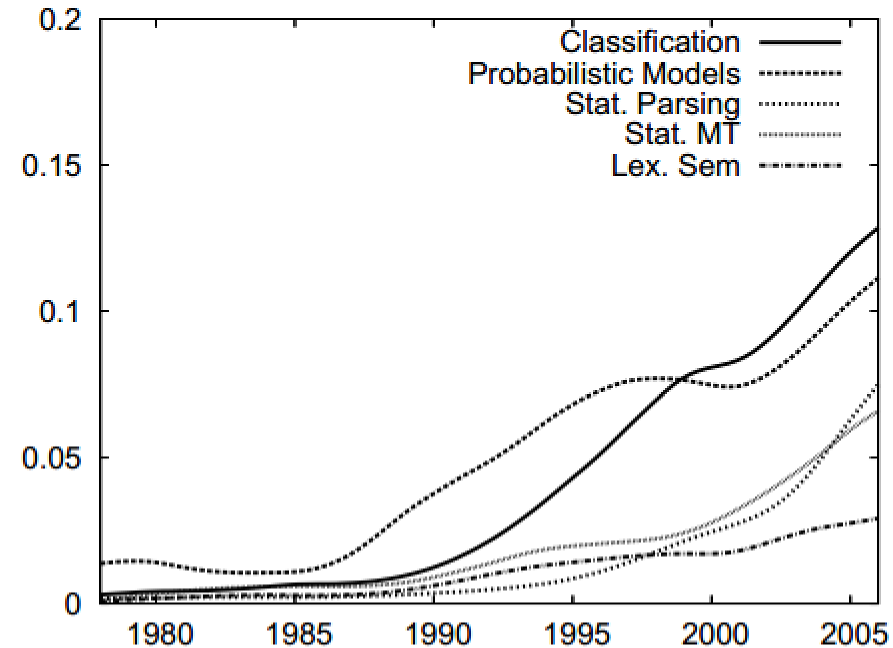
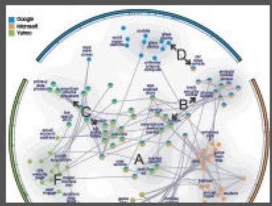


Figure 1: Topics in the ACL Anthology that show a strong recent increase in strength.

Anaphora Resolution	resolution anaphora pronoun discourse antecedent pı
Automata	string state set finite context rule algorithm strings la
Biomedical	medical protein gene biomedical wkh abstracts medi
Call Routing	call caller routing calls destination vietnamese route
Categorial Grammar	proof formula graph logic calculus axioms axiom th
Centering*	centering cb discourse cf utterance center utterances
Classical MT	japanese method case sentence analysis english dicti
Classification/Tagging	features data corpus set feature table word tag al test
Comp. Phonology	vowel phonological syllable phoneme stress phoneti
Comp. Semantics*	semantic logical semantics john sentence interpretat
Dialogue Systems	user dialogue system speech information task spoke
Discourse Relations	discourse text structure relations rhetorical relation u
Discourse Segment.	segment segmentation segments chain chains bound
Events/Temporal	event temporal time events tense state aspect referen
French Function	de le des les en une est du par pour
Generation	generation text system language information knowle
Genre Detection	genre stylistic style genres fiction humor register bib
Info. Extraction	system text information muc extraction template names patterns pattern domain
Information Retrieval	document documents query retrieval question information answer term text web
Lexical Semantics	semantic relations domain noun corpus relation nouns lexical ontology patterns
MUC Terrorism	slot incident tgt target id hum phys type fills perp
Metaphor	metaphor literal metonymy metaphors metaphorical essay metonymic essays qualia analogy
Morphology	word morphological lexicon form dictionary analysis morphology lexical stem arabic
Named Entities*	entity named entities ne names ner recognition ace nes mentions mention
Paraphrase/RTE	paraphrases paraphrase entailment paraphrasing textual para rte pascal entailed dagan

David Hall et al. 2008. Studying the history of ideas using topic models. In Proc. *EMNLP*, pp. 363-371.

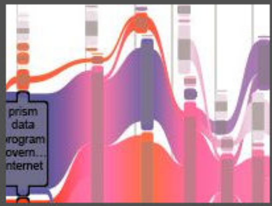


TopicPanorama: a Full Picture of Relevant Topics

Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu, Baining Guo.
Proceedings of IEEE VAST 2014

"Develop a full picture of relevant topics discussed in multiple text sources."

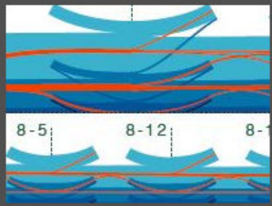
project | paper | video(English) | video(Chinese)



How Hierarchical Topics Evolve in Large Text Corpora

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei.
IEEE Transactions on Visualization and Computer Graphics (IEEE InfoVis 2014). 20(12): 2281-2290.

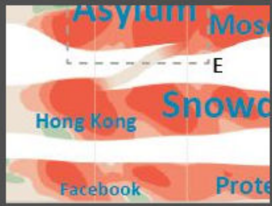
paper | video



LoyalTracker: Visualizing Loyalty Dynamics in Search Engines

Conglei Shi, Yingcai Wu, Shixia Liu, Hong Zhou, Huamin Qu.
IEEE Transactions on Visualization and Computer Graphics (IEEE VAST 2014). 20 (12): 1733-1742 (**Honorable Mention**).

paper | video



OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media

Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, Fangzhao Wu.
IEEE Transactions on Visualization and Computer Graphics (IEEE VAST 2014). 20 (12): 1763-1772

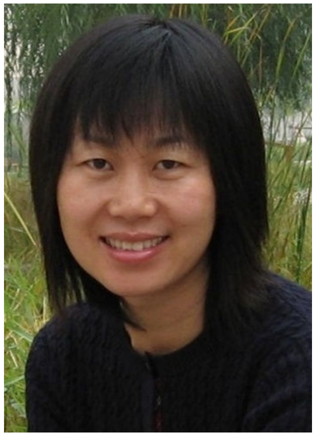
paper | video

EvoRiver: Visual Analysis of Topic Coopetition on Social Media

Guodao Sun, Yingcai Wu, Shixia Liu, Tai-Quan Peng, Jonathan J.H. Zhu, Ronghua Liang.
IEEE Transactions on Visualization and Computer Graphics (IEEE VAST 2014). 20 (12): 1753-1762

paper | video

Topic Modeling Visualization (done right)



Shixia Liu
Tsinghua University

Ongoing work

- Visualizing the health of online developer communities (GitHub) through sentiment in commits and issues, and other signals
- Suggesting analytic starting points – "Overview first" is just too general to be useful

Research Directions

- Putting it all together: NNPs + sentiment + semantic structure
- Statistical rigour: visualizing changes in sentiment, uncertainty in sentiment, significance of differences

Visual Text Analytics Best Practice

- Problem-driven, real questions about real data
- Generalizable techniques
- Human-understandable outputs of linguistic processing (issues of trust, transparency, usability)
- Interactive linking to original text

- "Clarify, don't simplify" – Marti Hearst

Undergrad

Zachary Cook
Taurean Scantlebury
Kalev Sikes
Ammar Sidhu

Grad

Rafael Veras
Erik Paluka
Brittany Kondo
Hrim Mehta
Stephen McIntyre
Daniel Chang
Chris Kim

UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY

Sheelagh Carpendale
Martin Wattenberg
Saif Mohammad
Mark Hancock
Jian Zhao
Jeremy Bradbury
Fraser Shein
Gerald Penn
Fernanda Viégas
Petra Isenberg
Fanny Chevalier
Julie Thorpe
Marian Dörk
Vivian Tsang

NSERC CRSNG
Federal Economic Development Agency for Southern Ontario
Quillsoft
GRAND
Canada Research Chairs www.chairs-chaires.gc.ca
SurfNet
SHARCNET™
Ontario
INNOVATION.CA

Christopher.Collins@uoit.net
http://vialab.science.uoit.ca
@chrisnf
https://github.com/vialab