November 14, 2011

Analyzing Documents and Text through Visualization

John Stasko

Information Interfaces Research Group
School of Interactive Computing
Georgia Institute of Technology

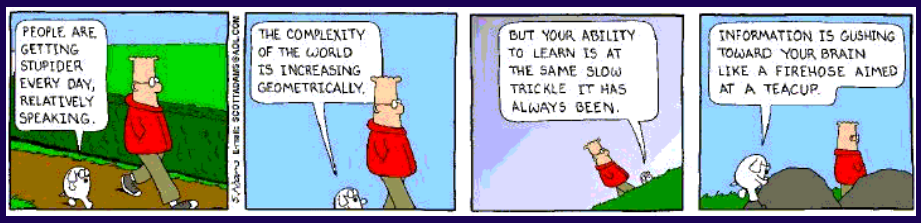
Killam Lecture
Dalhousie Univ.

Georgia Tech College of Computing
School of Interactive Computing

information interfaces

Data Overload

- How do we make use of the data rather than being overwhelmed by it?



information interfaces

Human Vision

- Highest bandwidth sense
- Fast, parallel
- Pattern recognition
- Pre-attentive
- Extends memory and cognitive capacity
- People think visually

Impressive. Lets use it!

information interfaces

Visualization

- "The use of computer-supported, interactive visual representations of data to amplify cognition."
  - Card, Mackinlay, Shneiderman '98

5

information interfaces

Purpose

- Cognition, not graphics
- "The purpose of visualization is insight, not pictures"

6

information interfaces

How?

- Visuals help us think
  - Provide a frame of reference, a temporary storage area
- Cognition → Perception
- Pattern matching
- External cognition aid
  - Role of external world in thinking and reason

Larkin & Simon '87
Card, Mackinlay, Shneiderman '98

To Help Convince You

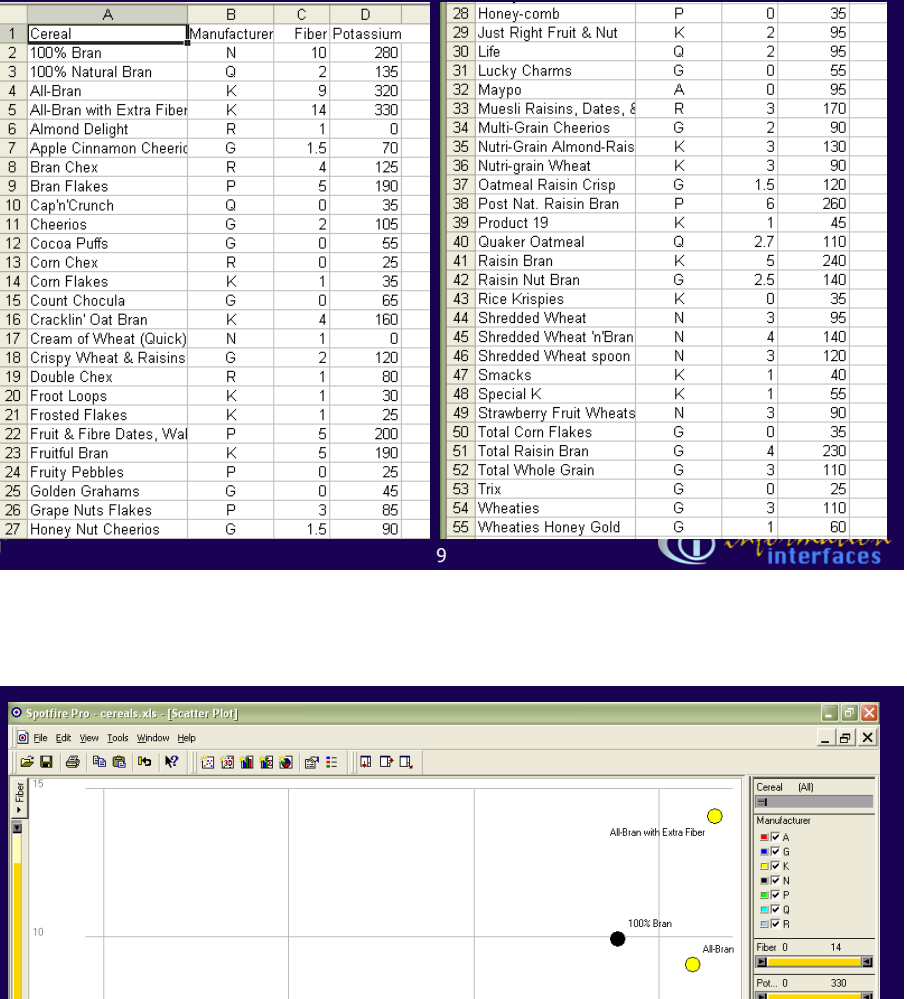
- Why visualization helps…

Questions:

Which cereal has the most/least potassium?
Is there a relationship between potassium and fiber?
If so, are there any outliers?
Which manufacturer makes the healthiest cereals?

	A	B	C	D
1	Cereal	Manufacturer	Fiber	Potassium
2	100% Bran	N	10	280
3	100% Natural Bran	Q	2	135
4	All-Bran	K	9	320
5	All-Bran with Extra Fiber	K	14	330
6	Almond Delight	R	1	0
7	Apple Cinnamon Cheeric	G	1.5	70
8	Bran Chex	R	4	125
9	Bran Flakes	P	5	190
10	Cap'n'Crunch	Q	0	35
11	Cheerios	G	2	105
12	Cocoa Puffs	G	0	55
13	Corn Chex	R	0	25
14	Corn Flakes	K	1	35
15	Count Chocula	G	0	65
16	Cracklin' Oat Bran	K	4	160
17	Cream of Wheat (Quick)	N	1	0
18	Crispy Wheat & Raisins	G	2	120
19	Double Chex	R	1	80
20	Froot Loops	K	1	30
21	Frosted Flakes	K	1	25
22	Fruit & Fibre Dates, Wal	P	5	200
23	Fruitful Bran	K	5	190
24	Fruity Pebbles	P	0	25
25	Golden Grahams	G	0	45
26	Grape Nuts Flakes	P	3	85
27	Honey Nut Cheerios	G	1.5	90
28	Honey-comb	P	0	35
29	Just Right Fruit & Nut	K	2	95
30	Life	Q	2	95
31	Lucky Charms	G	0	55
32	Maypo	A	0	95
33	Muesli Raisins, Dates, &	R	3	170
34	Multi-Grain Cheerios	G	2	90
35	Nutri-Grain Almond-Rais	K	3	130
36	Nutri-grain Wheat	K	3	90
37	Oatmeal Raisin Crisp	G	1.5	120
38	Post Nat. Raisin Bran	P	6	260
39	Product 19	K	1	45
40	Quaker Oatmeal	Q	2.7	110
41	Raisin Bran	K	5	240
42	Raisin Nut Bran	G	2.5	140
43	Rice Krispies	K	0	35
44	Shredded Wheat	N	3	95
45	Shredded Wheat 'n'Bran	N	4	140
46	Shredded Wheat spoon	N	3	120
47	Smacks	K	1	40
48	Special K	K	1	55
49	Strawberry Fruit Wheats	N	3	90
50	Total Corn Flakes	G	0	35
51	Total Raisin Bran	G	4	230
52	Total Whole Grain	G	3	110
53	Trix	G	0	25
54	Wheaties	G	3	110
55	Wheaties Honey Gold	G	1	60

9



Potassium

10

Even Tougher?

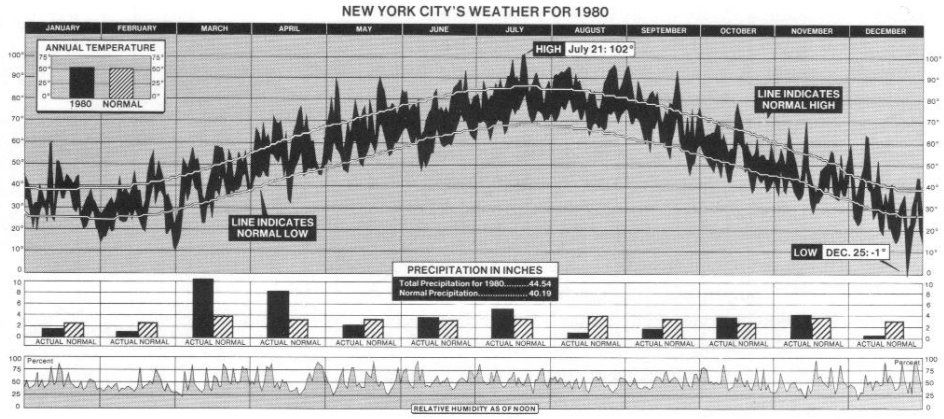
- What if you could only see one cereal's data at a time? (e.g. some websites)
- What if I read the data to you?

Some Examples

- Static "InfoGraphics"

NYC Weather

2220 numbers



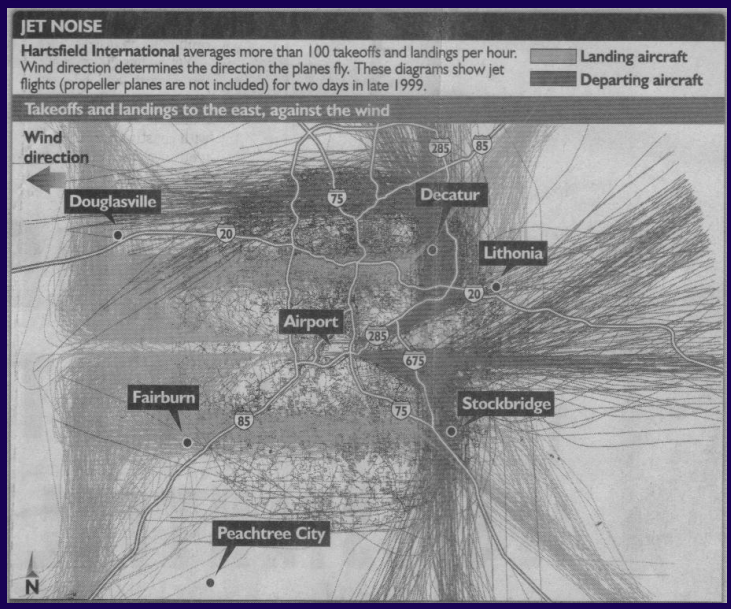
New York Times, January 11, 1981, p. 32.

Tufte, Vol. 1

information interfaces

Atlanta Flight Traffic

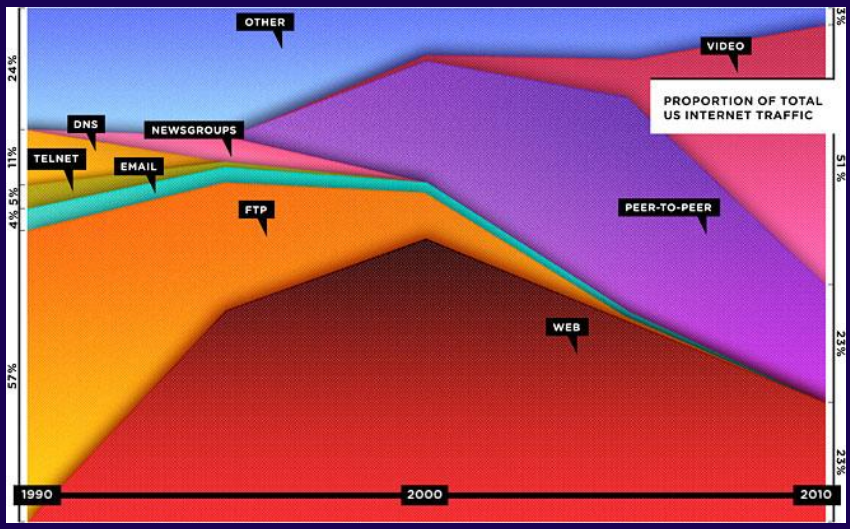
Atlanta Journal
April 30, 2000



information interfaces

Internet Traffic

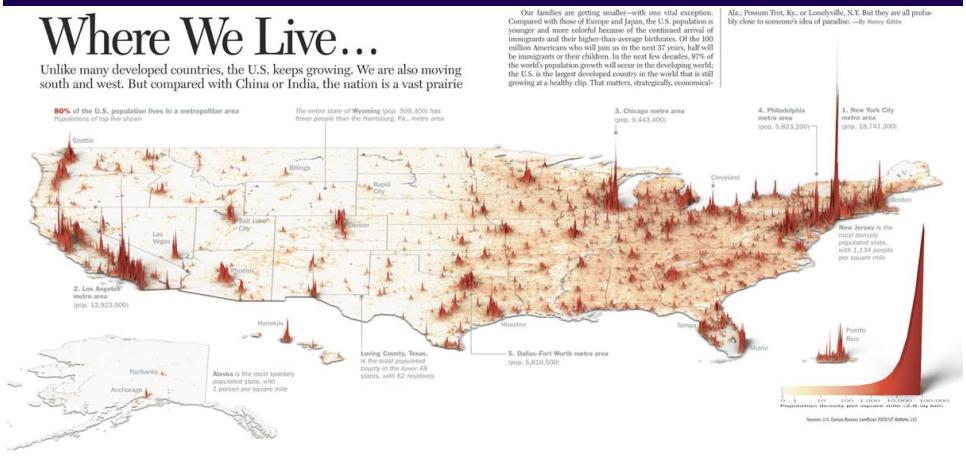
http://www.wired.com/magazine/2010/08/ff_webrip/all/1



15

Population

http://infographicsnews.blogspot.com/2009/04/mantras-joe-lertolas-maps.html



16

http://infosthetics.com/archives/2008/09/funniest_pie_chart_ever.html

17

information interfaces

Purpose

- Two main uses of visualization
  - Analysis – Understand your data better and act upon that understanding
  - Presentation – Communicate and inform others more effectively

18

information interfaces

1. Analysis

- Given all the data, then
  - understand, compare, decide, judge, evaluate, assess, determine, …
- Ultimately, about solving problems

19

information interfaces

When to Apply?

- Many other techniques for data analysis
  - Statistics, DB, data mining, machine learning…
- Visualization most useful in exploratory data analysis
  - Don't know what you're looking for
  - Don't have a priori questions
  - Want to know what questions to ask

20

information interfaces

EDA Example

- Business
  - Why has Hyundai made such great strides in the US market?
  - How influential was their "Lose your job, we'll buy the car back" campaign?
  - Have their cars improved in quality? If so, in what major ways?
  - Is the Genesis as good of a car as the Lexus ES?

21

information interfaces

2. Presentation – Tell a story

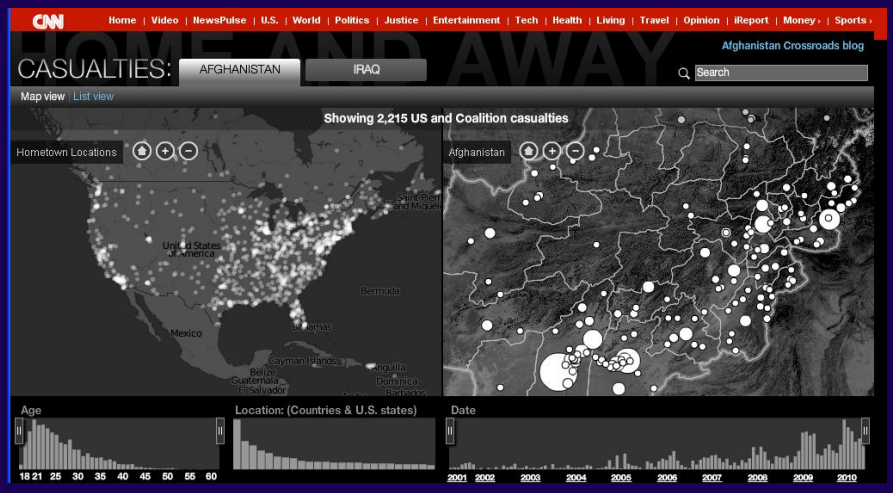
- Use visualization to communicate ideas, influence, explain
- Visuals can serve as evidence or support
- More and more news articles are accompanied by a data visualization

22

information interfaces

CNN

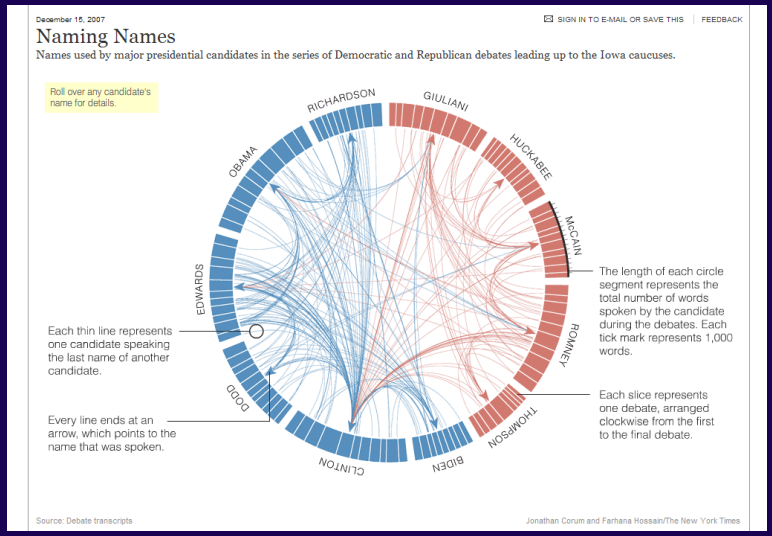
http://www.cnn.com/homeandaway



information interfaces

http://www.nytimes.com/interactive/2007/12/15/us/politics/DEBATE.html#

NY Times



information interfaces

Strongest Benefits of Visualization

- Facilitating awareness and understanding
- Helping to raise new questions and supply answers
- Generating insights
- Telling a story and making a point

25

information interfaces

Example Domains for InfoVis

- Text & documents
- Statistics
- Financial/business data
- Internet information
- Software
- …

26

information interfaces

Example Domains for InfoVis

- Text & documents
- Statistics
- Financial/business data
- Internet information
- Software
- …

27

information interfaces

Text is Everywhere

- We use documents as primary information artifact in our lives
- Our access to documents has grown tremendously in recent years due to networking infrastructure
  - WWW
  - Digital libraries
  - …

28

information interfaces

Example Tasks & Goals

- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- Which documents have an angry tone?
- How are certain words or themes distributed through a document?
- Identify "hidden" messages or stories in this document collection.
- Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
- Find connections between documents.

29

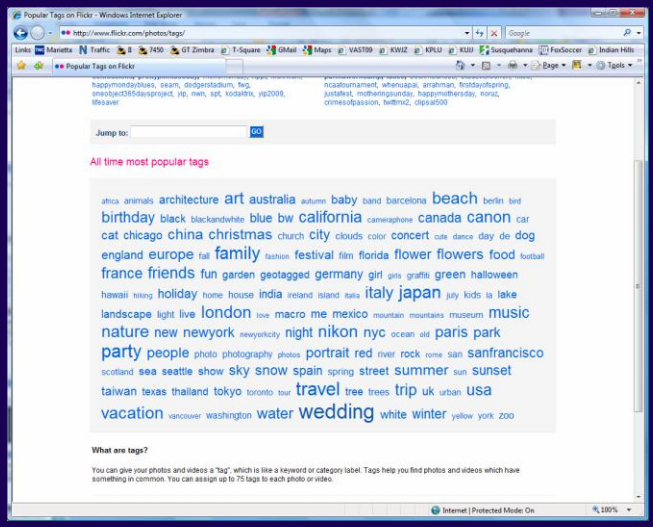
information interfaces

A Little Tour

30

information interfaces

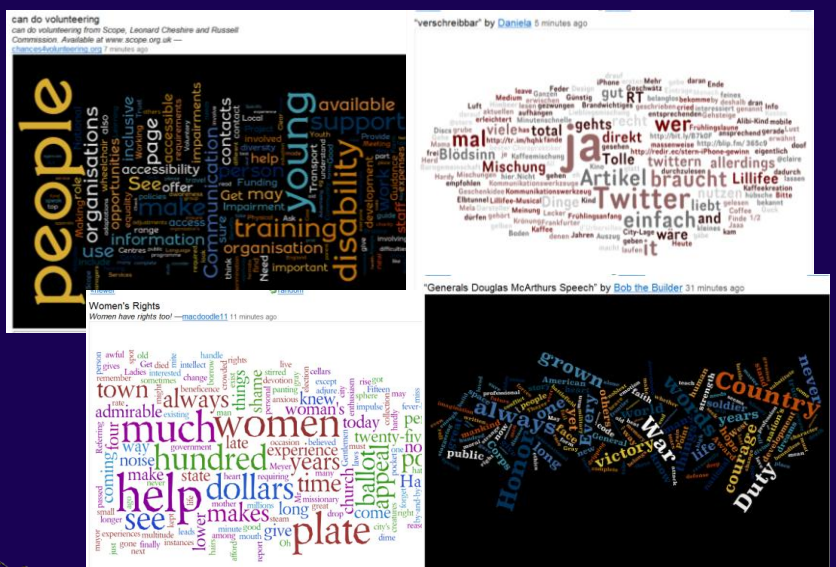
Tag Cloud (Flickr)



information interfaces

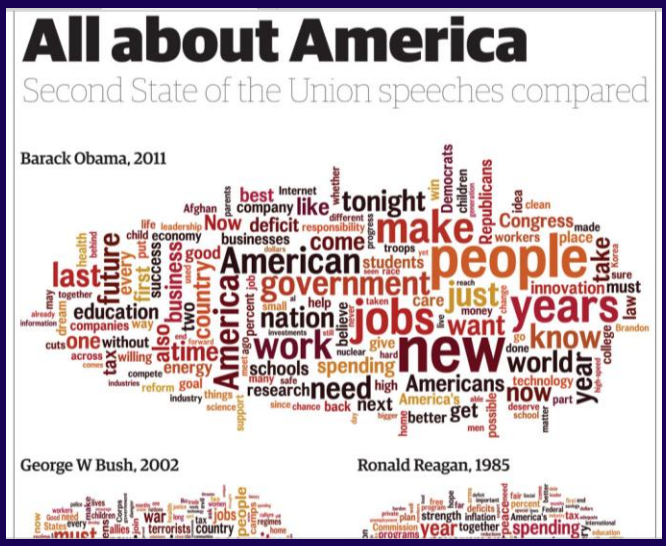
Wordle

http://www.wordle.net



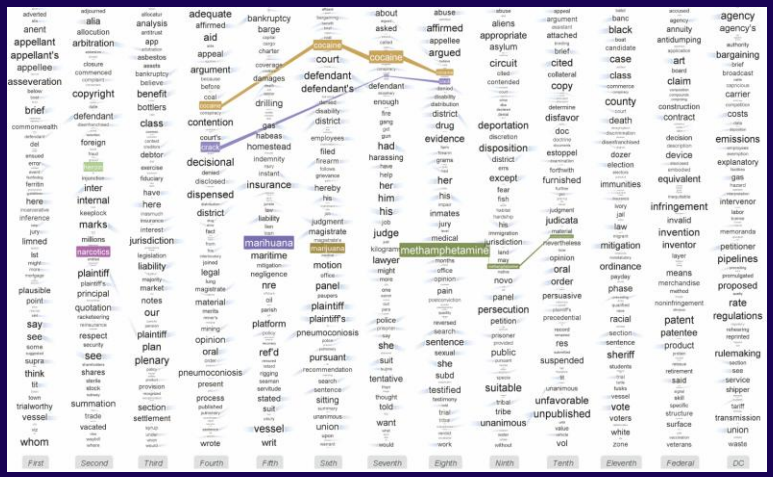
information interfaces

SotU Wordles



http://www.guardian.co.uk/news/datablog/2011/jan/25/state-of-the-union-text-obama#

Parallel Tag Clouds

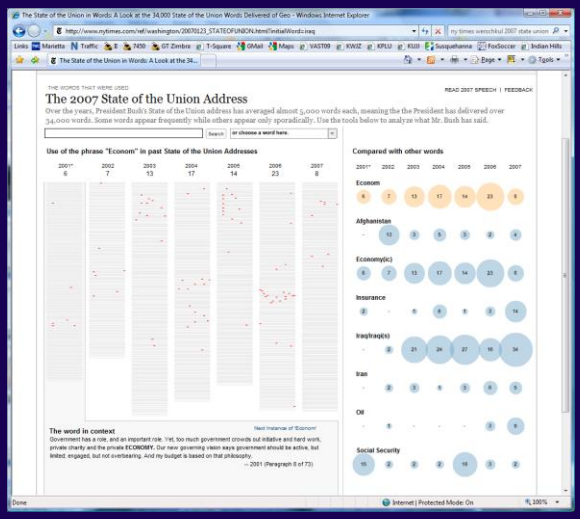


Showing word patterns in multiple documents

Collins et al VAST '09

Different circuit courts

NY Times State of the Union Reviews



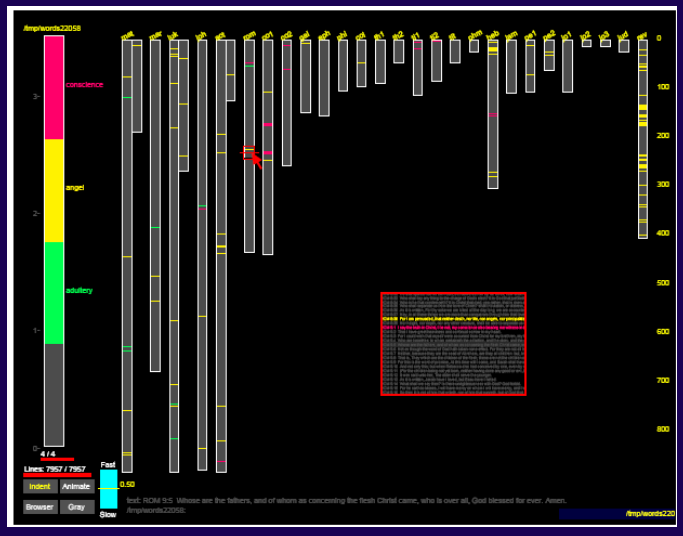
Adding search and queries

http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq

35

SeeSoft

Eick
Journal Comput. & Graph. Stats '94



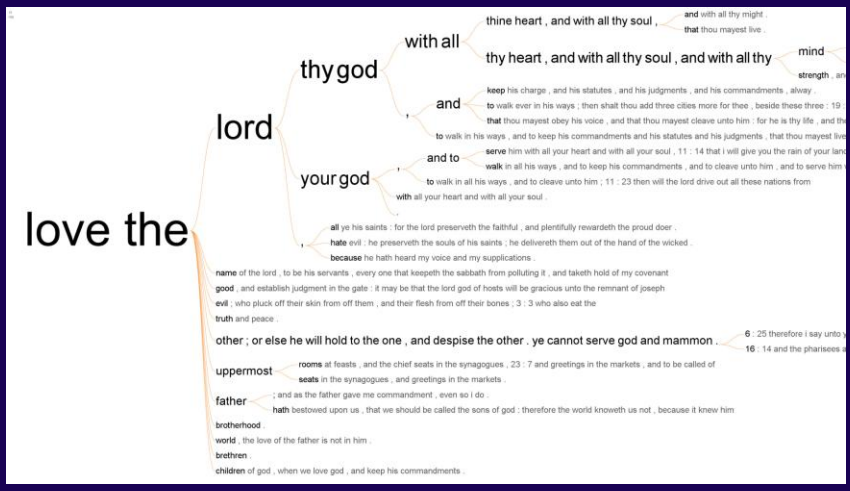
Large text overview

New Testament

36

Many Eyes' Word Tree

From King James Bible



http://www-958.ibm.com/software/data/cognos/manyeyes/

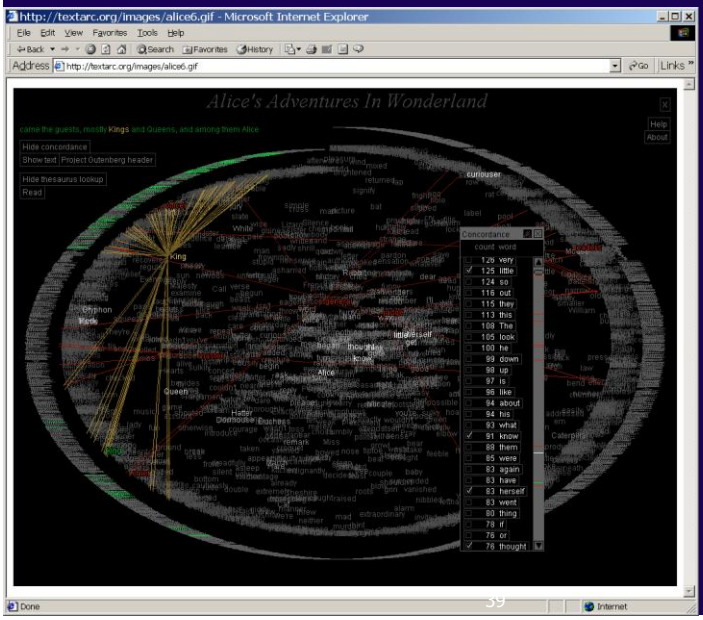
Phrase Nets

Fig 5. Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

van Ham et al
TVCG (InfoVis) '09

TextArc

http://textarc.org



Sentences laid out in order of appearance

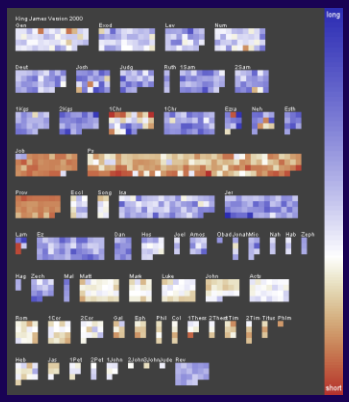
Words near to where they appear

Significant interaction

Brad Paley

information interfaces

Keim's Group's Work

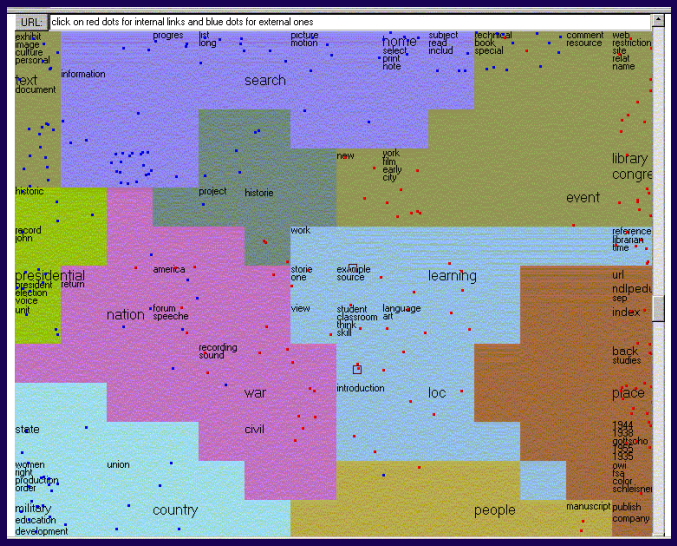


Keim & Oelke
VAST '07

Oelke & Keim
VAST '10

information interfaces

Self-Organizing (Kohonen) Maps

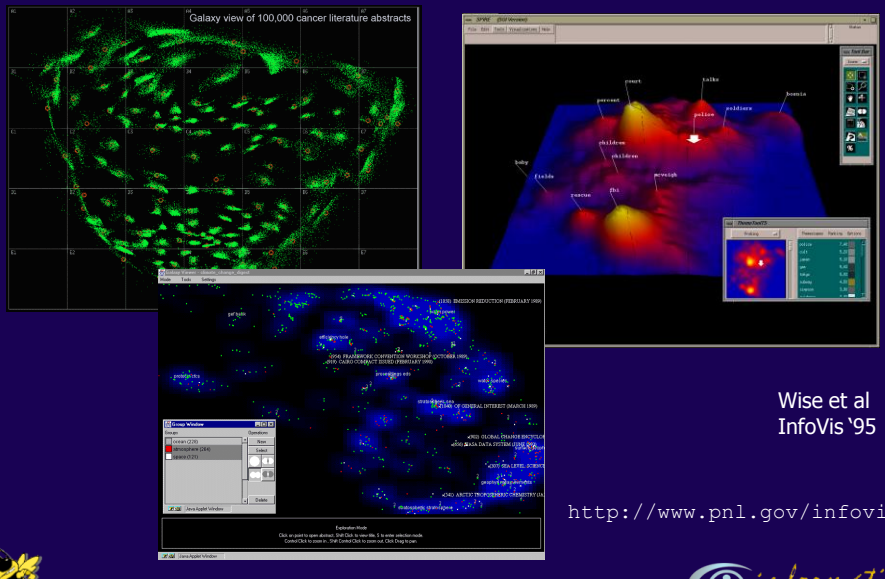


Regions represent concepts

Dots are the documents

by Xia Lin

Galaxy View and ThemeScape

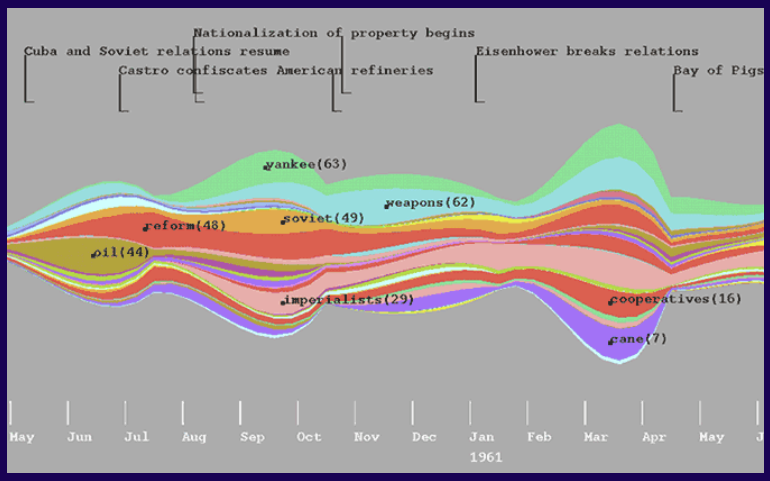


Wise et al
InfoVis '95

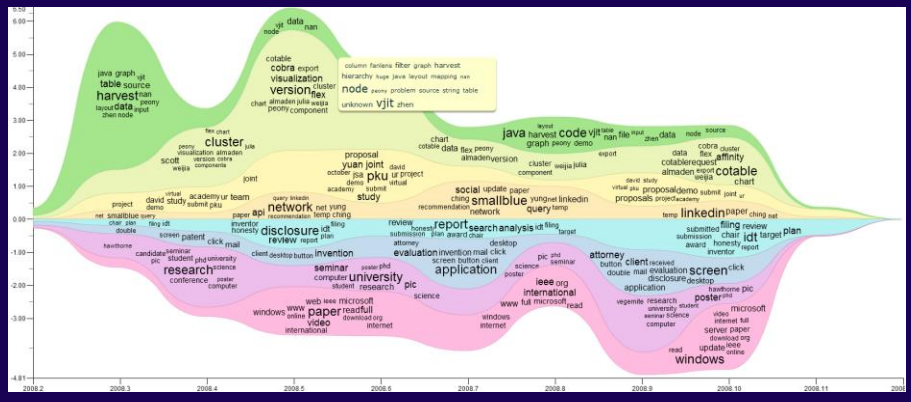
http://www.pnl.gov/infoviz

ThemeRiver

Havre, Hetzler, & Nowell
InfoVis '00



TIARA



Liu et al
CIKM '09, KDD '10, VAST '10

Jigsaw

Visualization for Investigative Analysis across Document Collections

- Law enforcement & intelligence community
- Fraud (finance, accounting, banking)
- Academic research
- Journalism & reporting
- Consumer research

"Putting the pieces together"

45

information interfaces

The Jigsaw Team

Current:

Carsten Görg
Zhicheng Liu
Youn-ah Kang
Jaeyeon Kihm

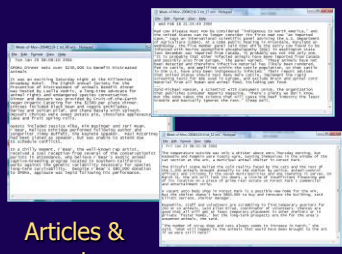
and many alumni

46

information interfaces

Problem Addressed

Help "investigators" explore, analyze and understand large document collections



Articles & reports

Spreadsheets

Blogs

XML documents

47

information interfaces

Our Focus

- Entities within the documents
  - Person, place, organization, phone number, date, license plate, etc.
- Thesis: A story/narrative/plot/threat within the documents will involve a set of entities in coordination

48

information interfaces

Sample Document

```
Report: 20040510-4_16
May 14 2004

VANCOUVER, British Columbia - A Canadian immigration panel is considering
whether accused environmental saboteur Tre Arrow can apply for refugee
status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in
Oregon, asked the Canadian authorities to remain in Canada as a political
refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether Arrow is affiliated with a terrorist group,
which would immediately disqualify him from receiving refugee status in
Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether
Arrow is affiliated with the Earth Liberation Front, a group the FBI
considers a terrorist organization responsible for scores of attacks on
property over the past dozen years.
```

Entities Identified

Source:
Date: May 14, 2004

VANCOUVER, British Columbia - A Canadian immigration panel is considering whether accused environmental saboteur Tre **Arrow** can apply for refugee status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether **Arrow** is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether **Arrow** is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.

Connections

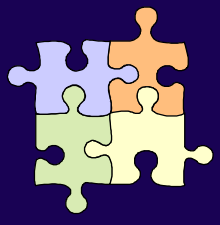
- Entities relate/connect to each other to make a larger "story"
- Connection definition:
  - Two entities are connected if they appear in a document together
  - The more documents they appear in together, the stronger the connection

51

information interfaces

Jigsaw

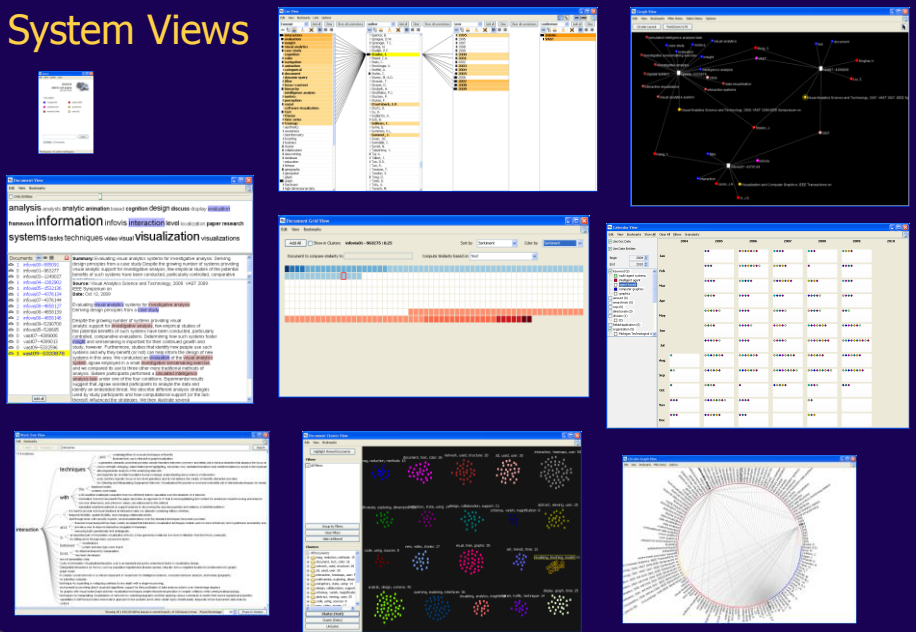
- Computational analysis of document text
  - Entity identification, document similarity, clustering, summarization, sentiment
- Multiple visualizations (views) of documents, analysis results, entities and their connections
- Views are highly interactive and coordinated



52

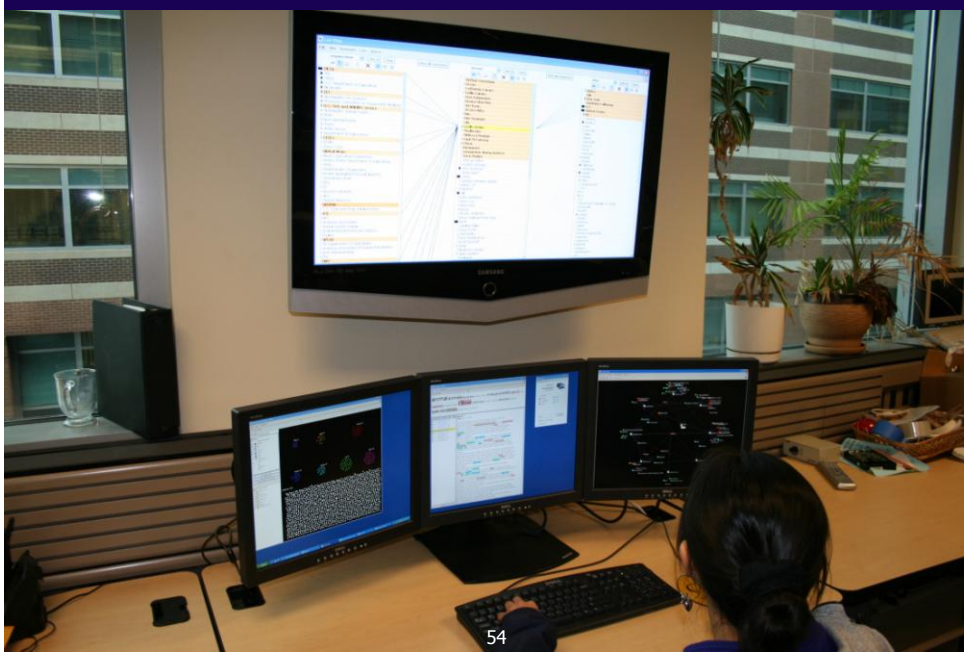
information interfaces

System Views

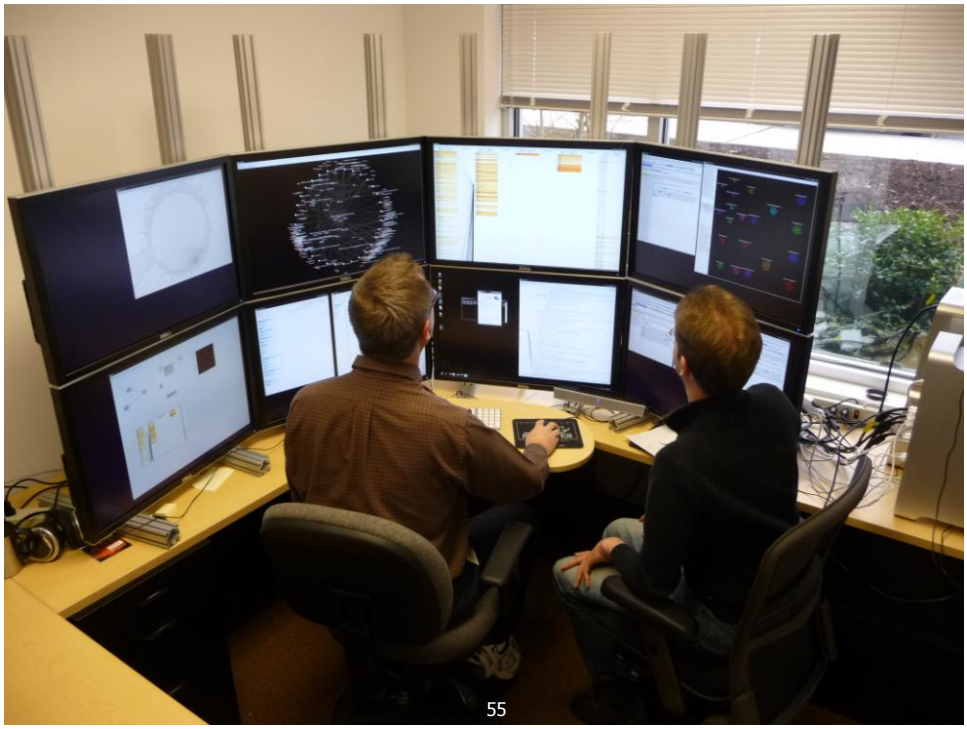


53

The Need for Pixels



54



55

Example Document Collection

56

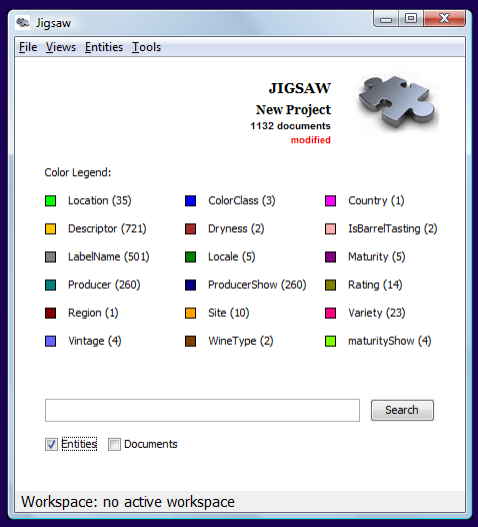
Demo

- Reviews of wines from Tuscany, '07-on
  - Text: review narrative
  - Entities: variety, producer, rating, vintage, color, location, producer, "descriptor", …
- Descriptor (~ 9000)
  - eg: abrasive, oaky, cherry, mocha, textured
- 1132 reviews
  - From database of 150,000 reviews

57

information interfaces

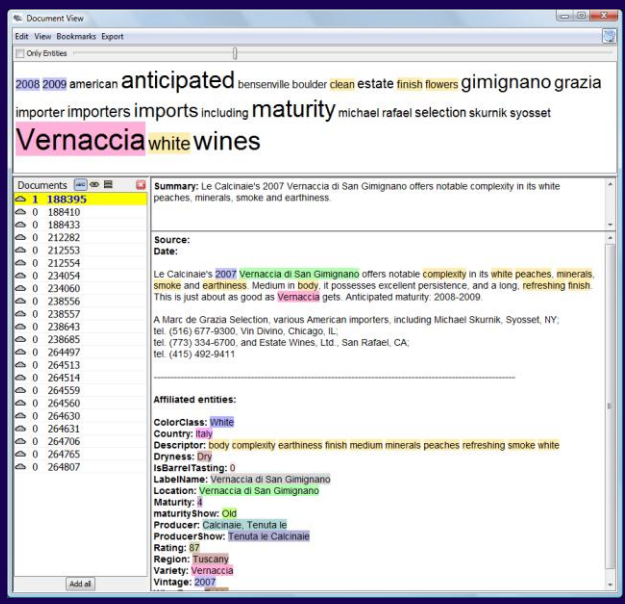
Console



58

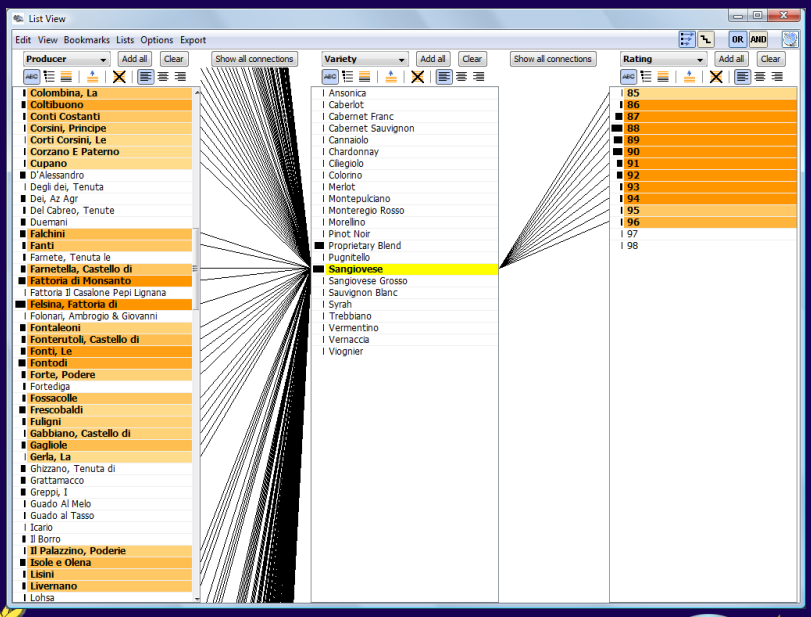
information interfaces

Document View



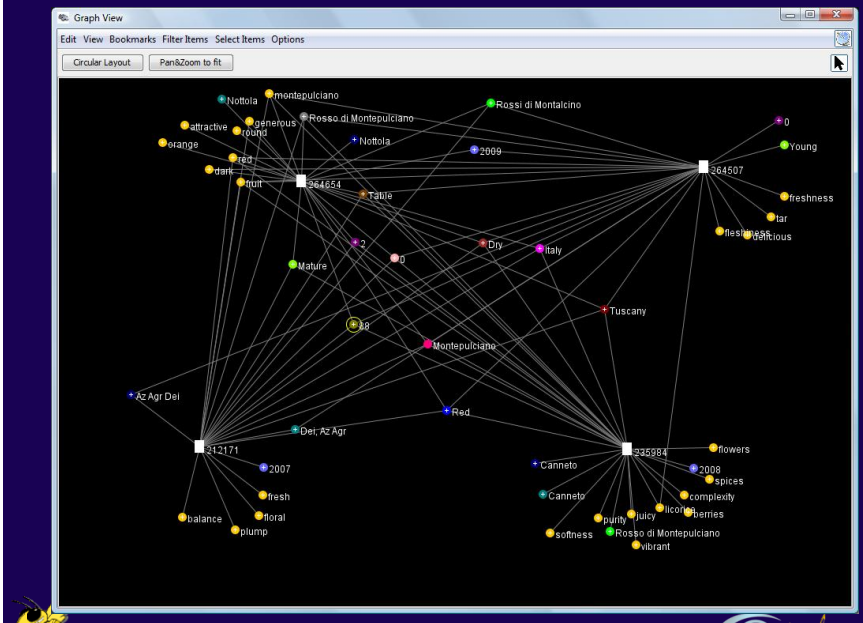
information interfaces

List View

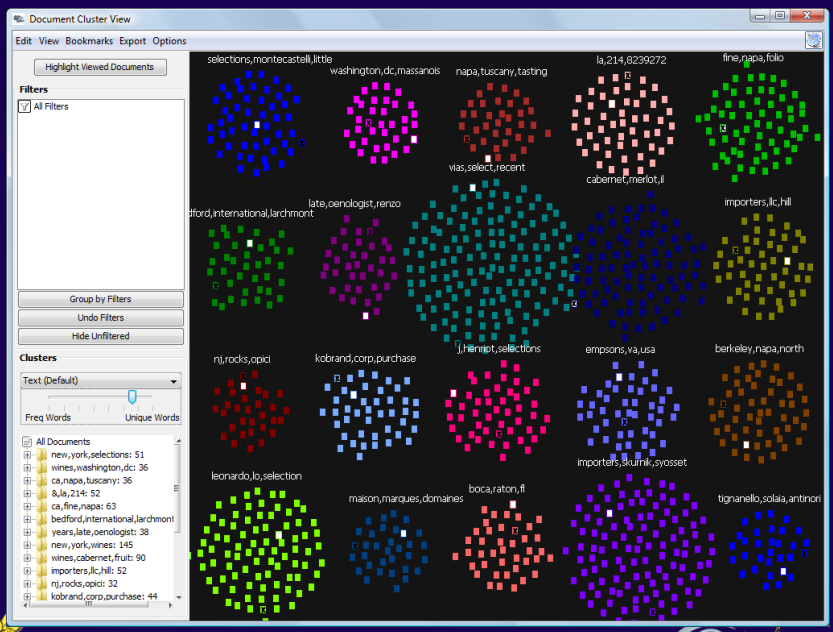


information interfaces

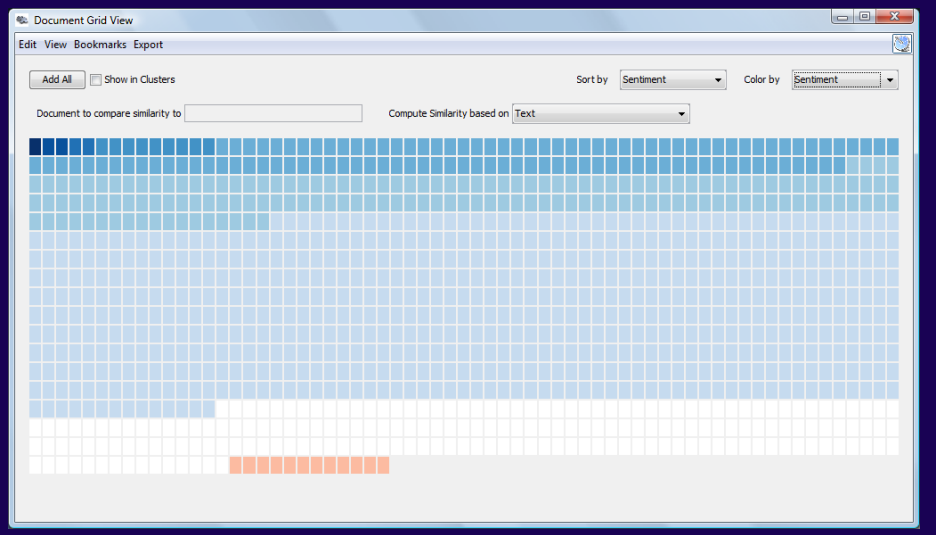
Graph View



Document Cluster View

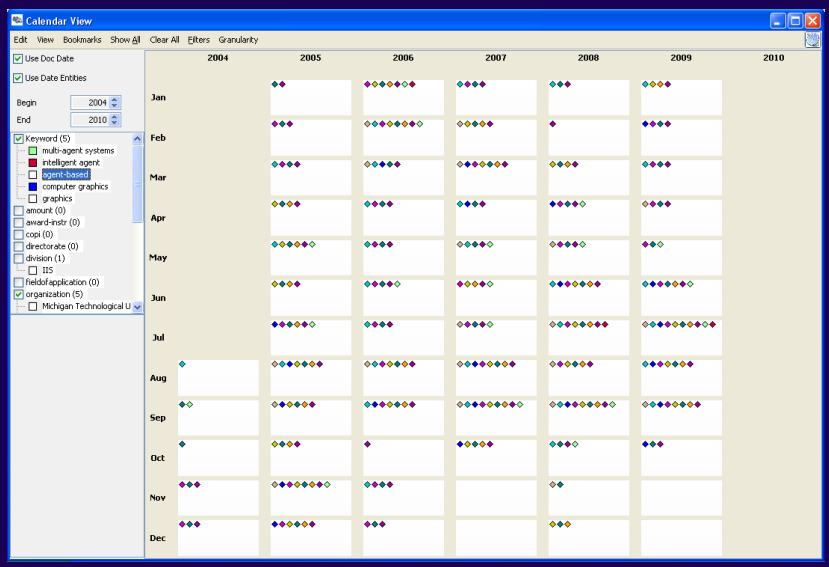


Document Grid View



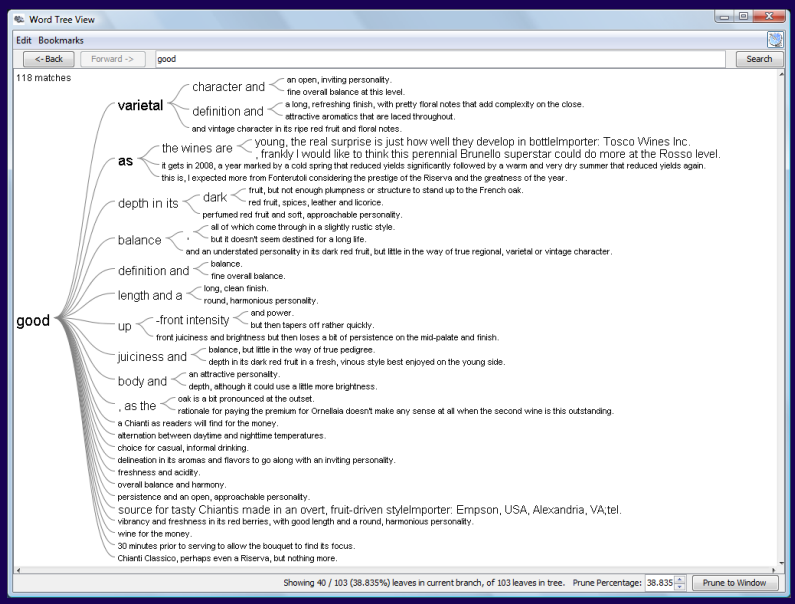
63

Calendar View

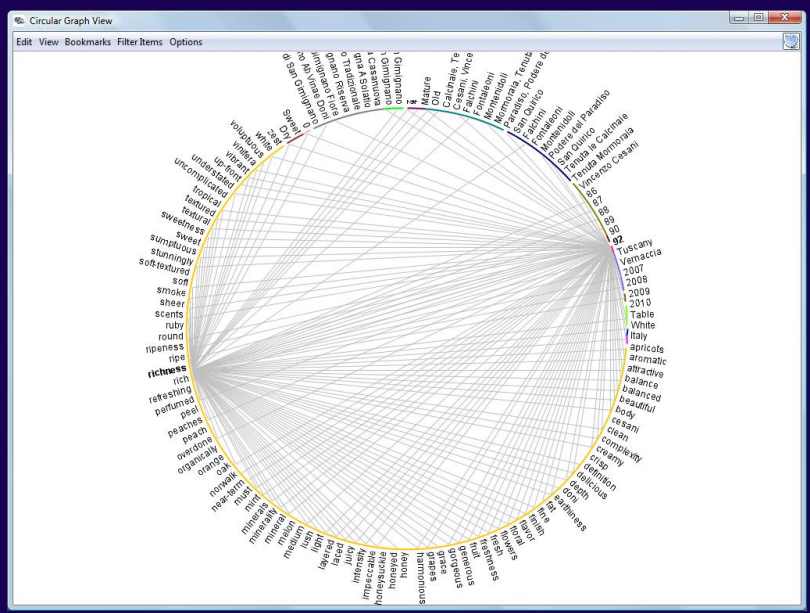


64

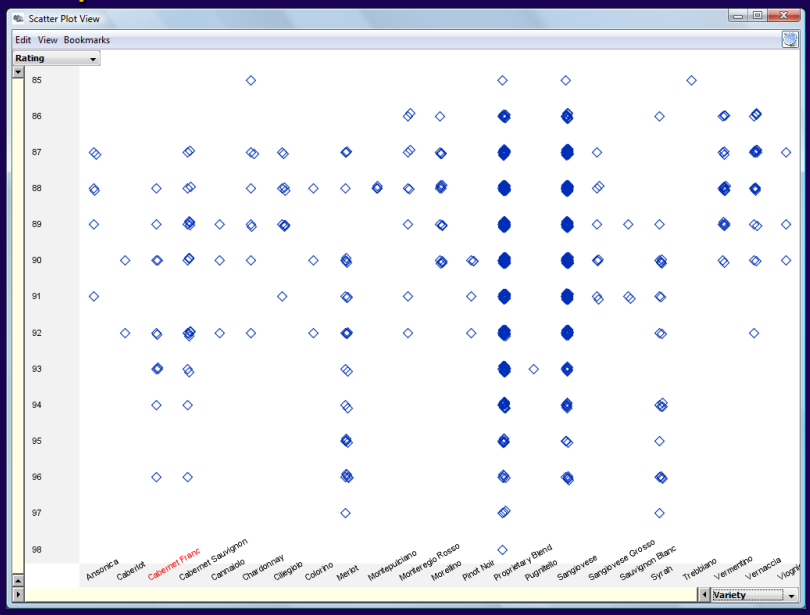
WordTree View



Circular Graph View



Scatterplot View



information interfaces

Application Domains

- Intelligence & law enforcement
  - Police cases
  - Won 2007 VAST Contest
  - Stasko et al, *Information Visualization* '08
- Academic papers, PubMed
  - All InfoVis & VAST papers
  - CHI papers
  - Görg et al, KES '10
- Investigative reporting
- Fraud
  - Finance, accounting, banking
- Grants
  - NSF CISE awards from 2000
- Topics on the web (medical condition)
  - Autism
- Consumer reviews
  - Amazon product reviews, edmunds.com, tripadvisor.com
  - Görg et al, HCIR '10
- Business Intelligence
  - Patents, press releases, corporate agreements, …
- Emails
  - White House logs
- Software
  - Source code repositories
  - Ruan et al, SoftVis '10

information interfaces

To Learn More & Availability

`http://www.gvu.gatech.edu/ii/jigsaw`

Available for (free) trial use

Send email to:
stasko@cc.gatech.edu

69

information interfaces

Conclusion

- Visualization is about fostering new insights
  - Analysis
  - Presentation
- Interactive exploration instead of sequential reading
- Text/documents is a fascinating new area for visual analytics research

70

information interfaces

Acknowledgments

- Work conducted as part of the Southeastern Regional Visualization and Analytics Center, supported by DHS and NVAC and the DHS Center of Excellence in Command, Control & Interoperability (VACCINE Center)

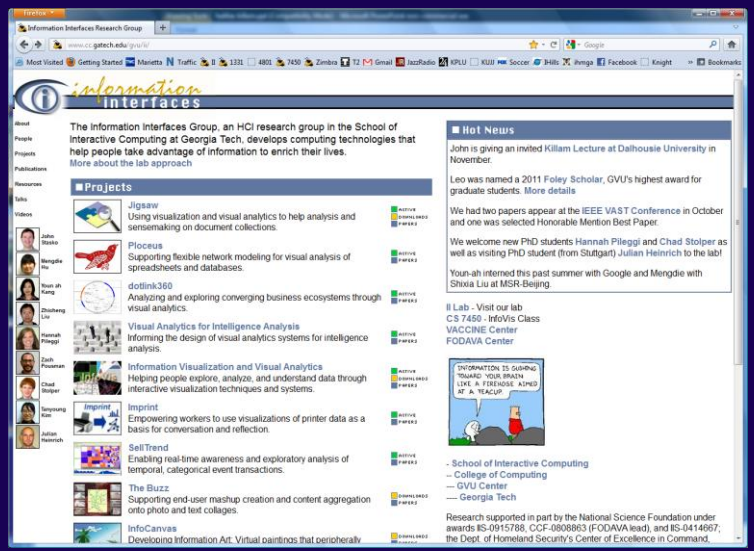
- Supported by NSF IIS-0414667, CCF-0808863 (FODAVA lead), NSF IIS-0915788

This material is supported in part by:
National Science Foundation
Grant #0414667

information interfaces

Thanks!

http://www.cc.gatech.edu/gvu/ii



information interfaces