

November 14, 2011

# Analyzing Documents and Text through Visualization

John Stasko

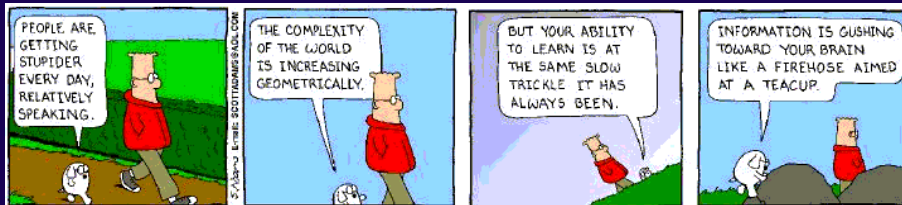
Information Interfaces Research Group  
School of Interactive Computing  
Georgia Institute of Technology

Killam Lecture  
Dalhousie Univ.



## Data Overload

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



3

## Human Vision

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

Impressive. Lets use it!



4

## Visualization

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



5

## Purpose

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



6

## 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



7



## To Help Convince You

- Why visualization helps...



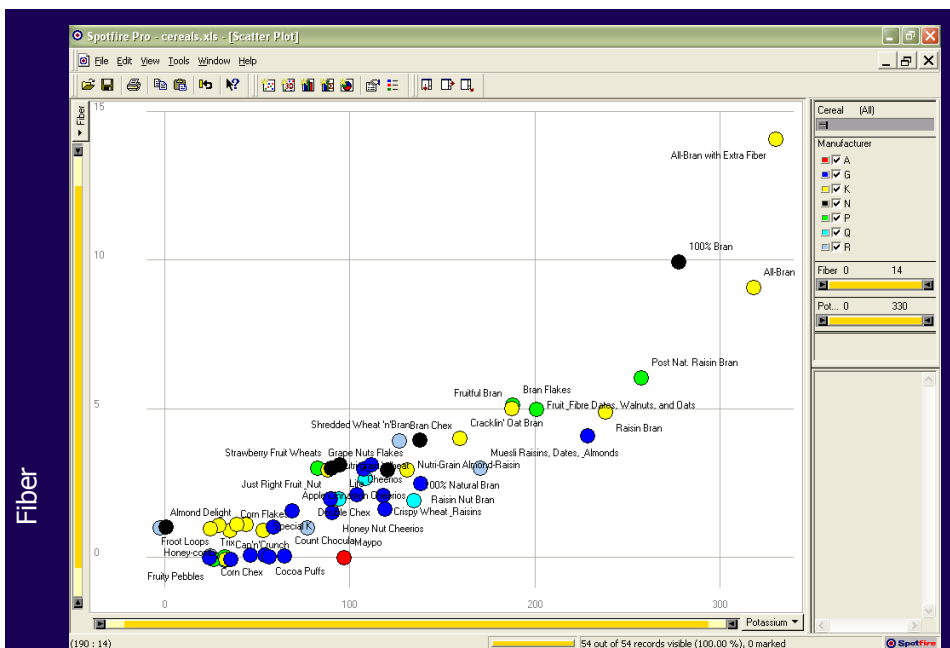
8



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



## 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?



11

## Some Examples

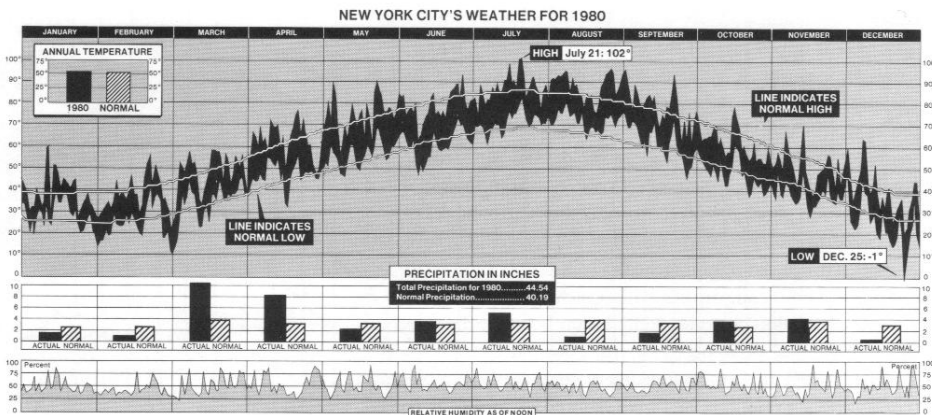
- Static "InfoGraphics"



12

# NYC Weather

2220 numbers



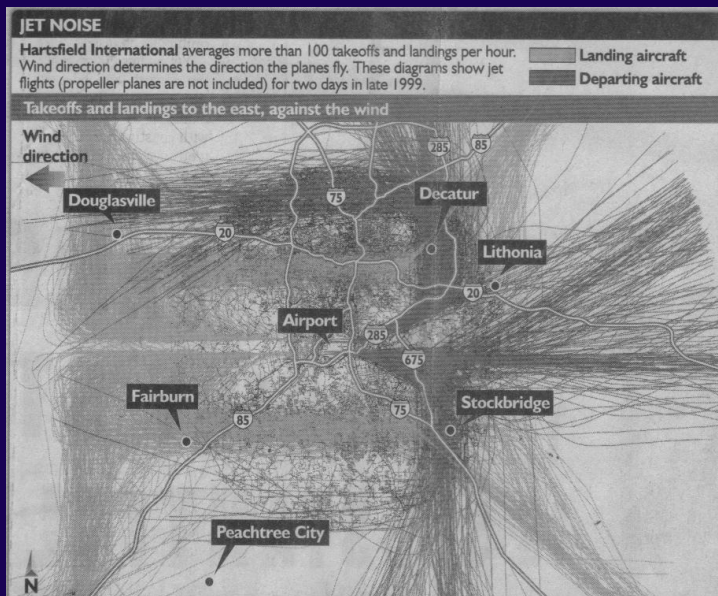
New York Times, January 11, 1981, p. 32.

Tufte, Vol. 1



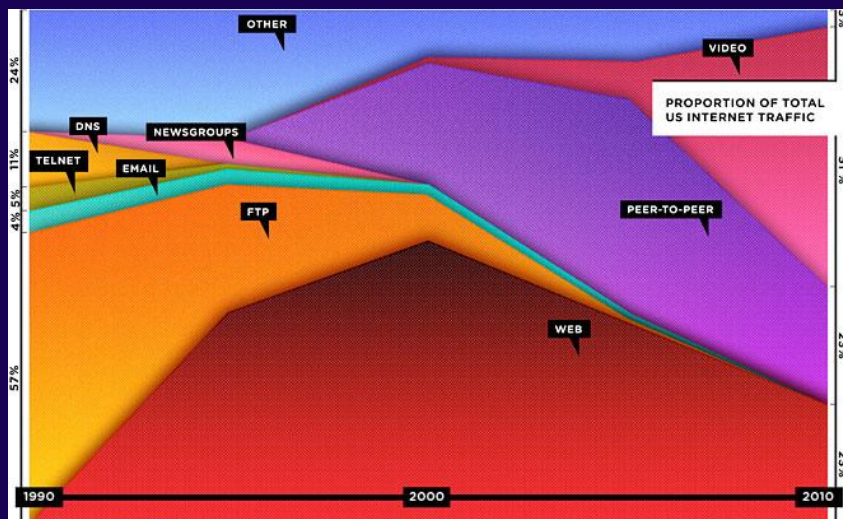
# Atlanta Flight Traffic

Atlanta Journal  
April 30, 2000



# Internet Traffic

[http://www.wired.com/magazine/2010/08/ff\\_webrip/all/1](http://www.wired.com/magazine/2010/08/ff_webrip/all/1)



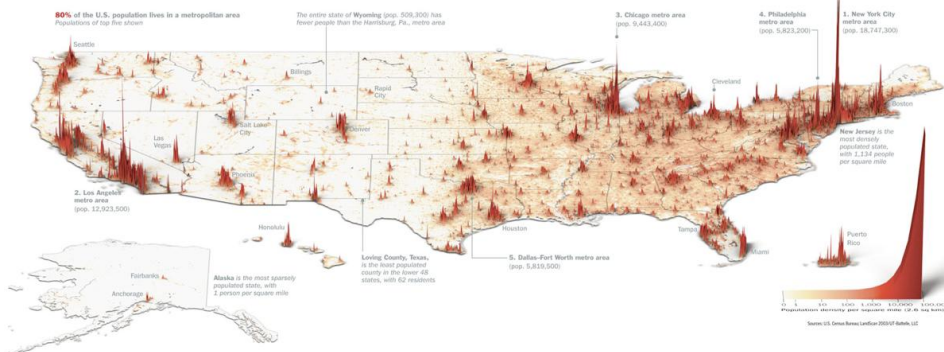
# Population

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

## Where We Live...

Unlike many developed countries, the U.S. keeps growing. We are also moving south and west. But compared with China or India, the nation is a vast prairie.

Our families are getting smaller—with one vital exception. Compared with those of Europe and Japan, the U.S. population is younger and more colorful because of the continued arrival of immigrants and their higher-than-average birthrates. Of the 100 million Americans who will join us in the next 27 years, half will be immigrants or their children. In the next few decades, 87% of the world's population growth will occur in the developing world; the U.S. is the largest developed country in the world that is still growing at a healthy clip. That matters, statistically, economically. ...  
 Ala., Fresno Dist., So., or Louisville, N.Y. But they are all probably close to someone's idea of paradise. ...by Nancy Gates







**■ Pie I have eaten**  
**■ Pie I have not yet eaten**

[http://infosthetics.com/archives/2008/09/funniest\\_pie\\_chart\\_ever.html](http://infosthetics.com/archives/2008/09/funniest_pie_chart_ever.html)

17

## Purpose

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

18

## 1. Analysis

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



19

## 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

## 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

## 2. Presentation – Tell a story

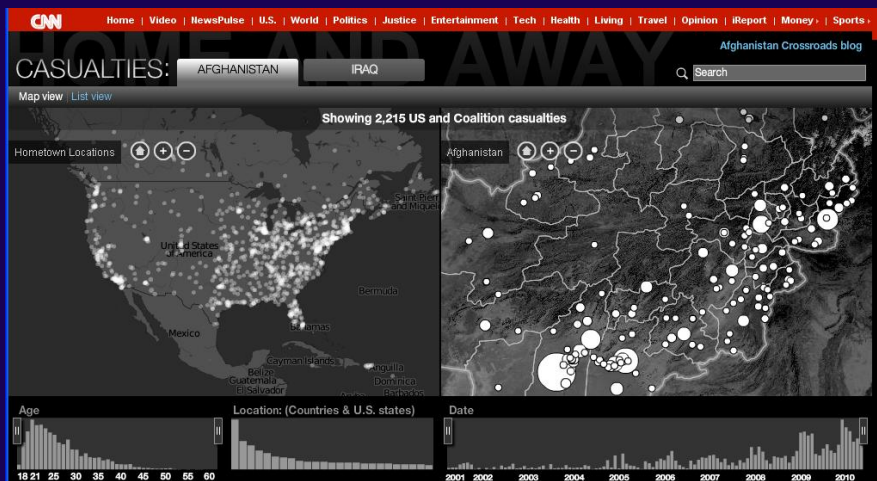
- Use visualization to communicate ideas, influence, explain
- Visuals can serve as evidence or support
- More and more news articles are accompanied by a data visualization



22

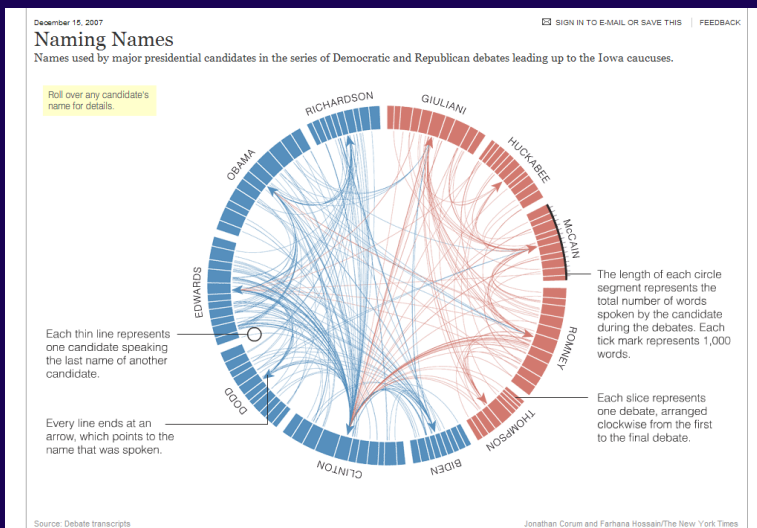
CNN

http://www.cnn.com/homeandaway



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

NY Times



## 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

## Example Domains for InfoVis

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



26

## Example Domains for InfoVis

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



27

## 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

## 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.



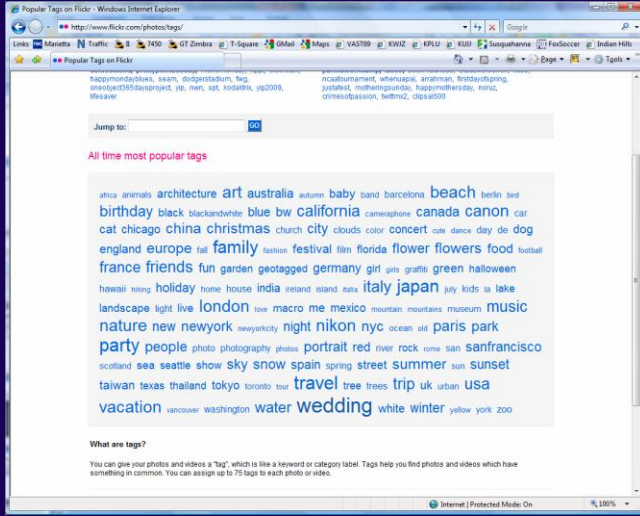
29

## A Little Tour



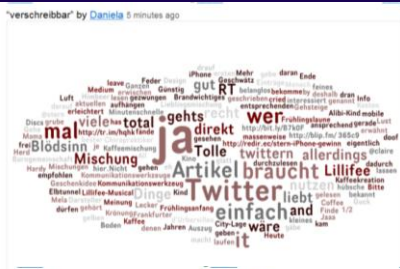
30

## Tag Cloud (Flickr)



## Wordle

<http://www.wordle.net>





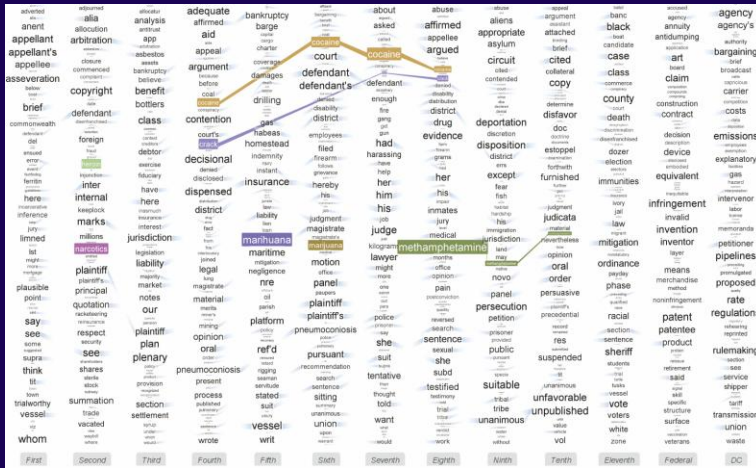
# SotU Wordles



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



# Parallel Tag Clouds



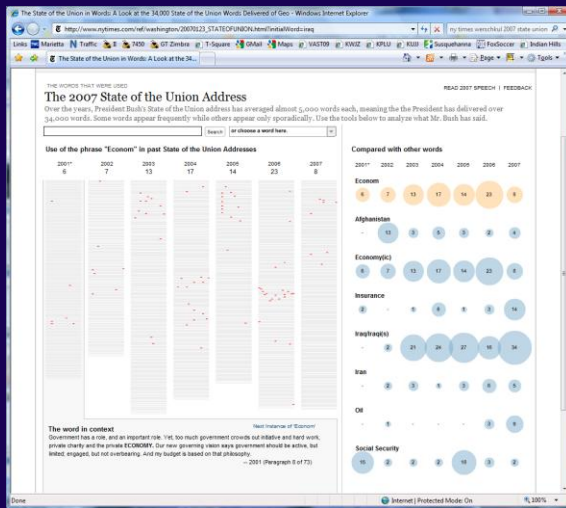
Showing word patterns in multiple documents

Collins et al VAST '09

Different circuit courts



# NY Times State of the Union Reviews



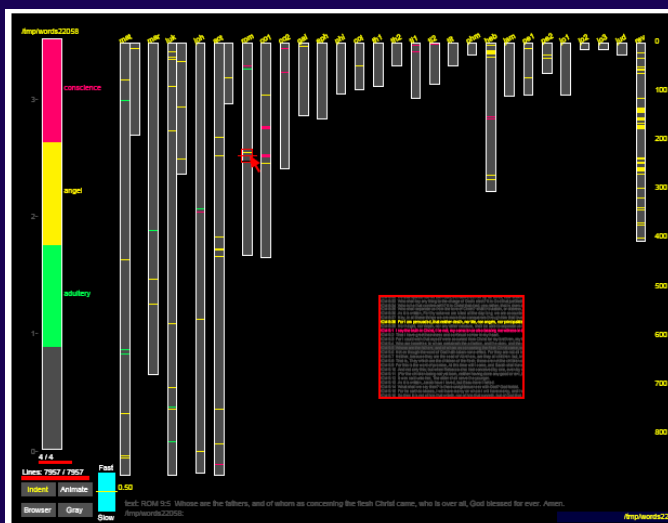
Adding search and queries

[http://www.nytimes.com/ref/washington/20070123\\_STATEOFUNION.html?initialWord=iraq](http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq)



# SeeSoft

Eick  
*Journal Comput. & Graph. Stats '94*



Large text overview

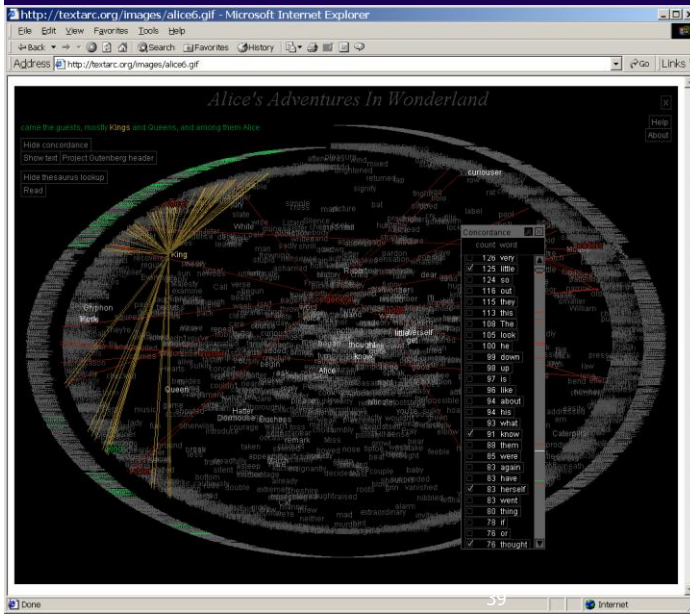
New Testament





# TextArc

<http://textarc.org>



Sentences laid out in order of appearance

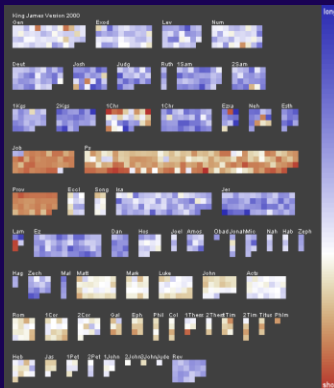
Words near to where they appear

Significant interaction

Brad Paley



# Keim's Group's Work

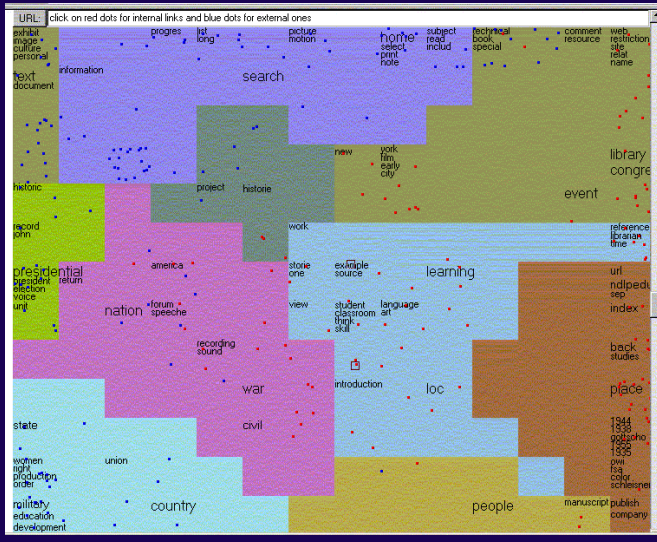


Keim & Oelke  
VAST '07

Oelke & Keim  
VAST '10



# Self-Organizing (Kohonen) Maps



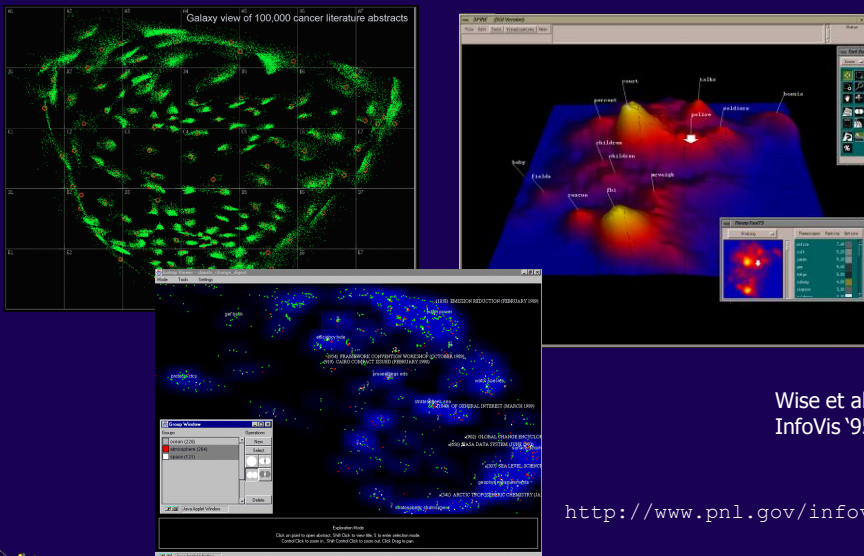
Regions represent concepts

Dots are the documents

by Xia Lin



# Galaxy View and ThemeScape



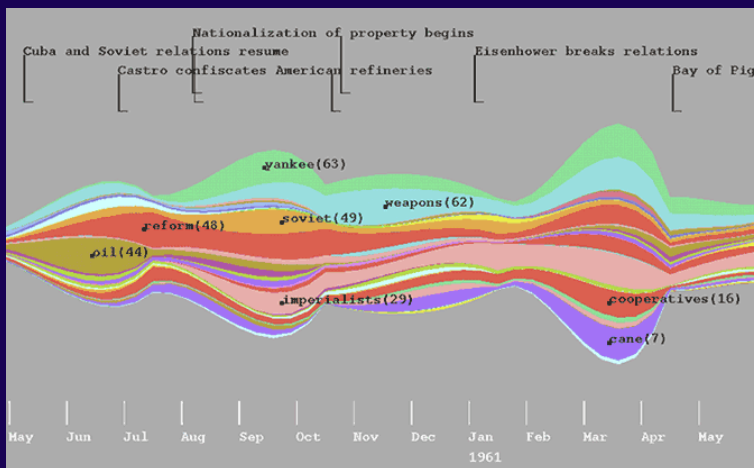
Wise et al  
InfoVis '95

<http://www.pnl.gov/infviz>

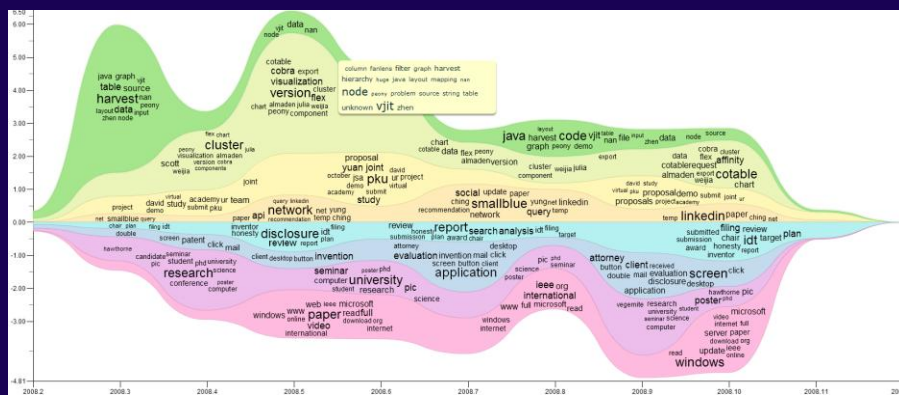


# ThemeRiver

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

## The Jigsaw Team

Current:

Carsten Görg  
Zhicheng Liu  
Youn-ah Kang  
Jaeyeon Kihm

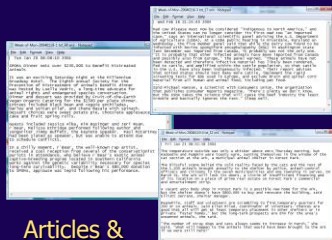
and many alumni



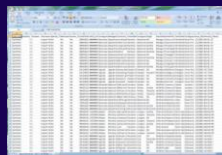
46

## Problem Addressed

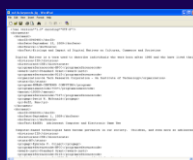
Help “investigators” explore, analyze and understand large document collections



Articles & reports



Spreadsheets



XML documents



Blogs



## 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





## 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.



49

## 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.



50

## Connections

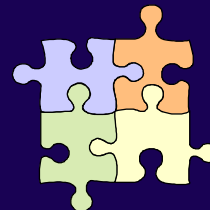
- Entities relate/connect to each other to make a larger “story”
- Connection definition:
  - Two entities are connected if they appear in a document together
  - The more documents they appear in together, the stronger the connection



51

## 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

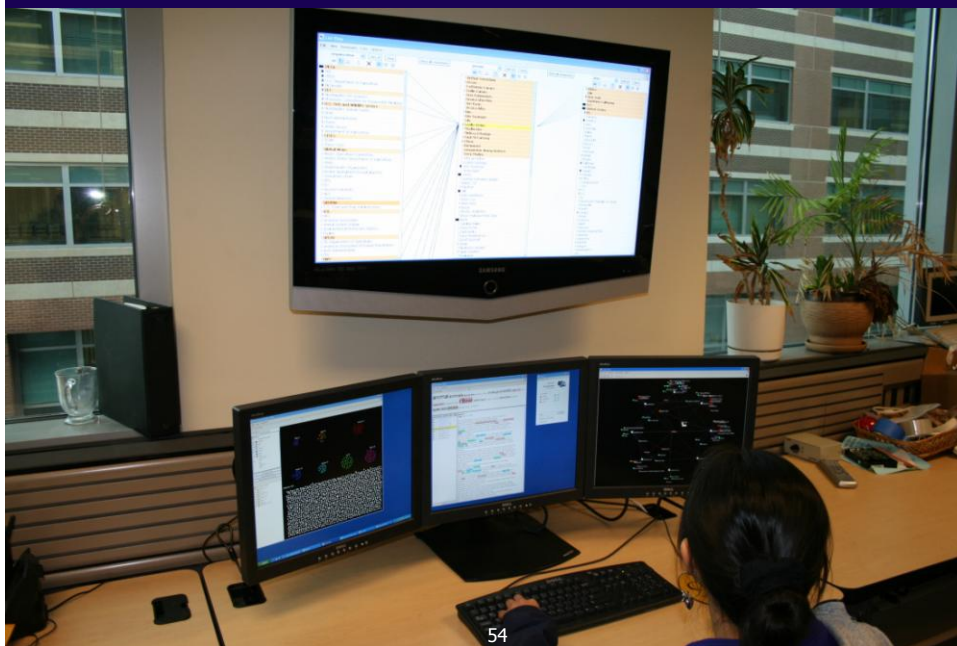
# System Views

The collage includes the following windows:

- Top Left:** A small window with a tree view.
- Top Middle:** A window showing a complex network graph with nodes and edges.
- Top Right:** A window displaying a large, dense network graph.
- Middle Left:** A window with a list of items and a detailed text view on the right.
- Middle Middle:** A window showing a grid or timeline visualization with colored bars.
- Middle Right:** A window displaying a large table with many columns and rows of data.
- Bottom Left:** A window showing a hierarchical tree structure.
- Bottom Middle:** A window displaying a network graph with nodes colored in various colors.
- Bottom Right:** A window showing a circular network graph with many nodes.



# The Need for Pixels





55

## Example Document Collection

A screenshot of the eRobertParker.com website. The page features a navigation bar with links for HOME, SUBSCRIBE, GIFT SUBSCRIPTIONS, FAQs, VIRTUAL TOUR, SUPPORT, SITE MAP, and CONTACT US. The main content area includes a 'What's Inside' section with links to 'Explore Wines', 'The Hedonist's Gazette', 'Article Archive', and 'My Wines'. There is a large featured article titled 'Welcome to Robert Parker Online' with a photo of Robert Parker. Below this, there are several smaller articles and product recommendations, such as '2009 Ridge Chardonnay' and '30% of Wine Offense Apps'. The footer contains a 'Quick Links' section with various utility links like 'Username/Password', 'GSA Subscriptions', and 'About Our Reviews'.



56

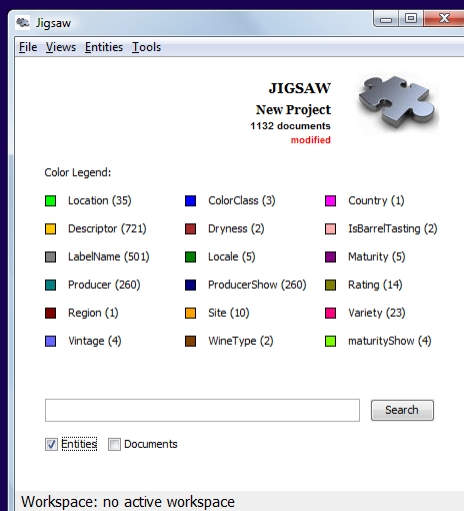
## Demo

- 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

## Console



58

# Document View

2008 2009 american anticipated benseriville boulder clean estate finish flowers gimignano grazia importer importers imports including maturity michael rafael selection skurnik syosset

**Vernaccia white wines**

**Documents**

- 1 188395
- 0 188410
- 0 188433
- 0 212282
- 0 212553
- 0 212554
- 0 234054
- 0 234060
- 0 238556
- 0 238557
- 0 238643
- 0 238685
- 0 264497
- 0 264513
- 0 264514
- 0 264559
- 0 264560
- 0 264630
- 0 264631
- 0 264706
- 0 264765
- 0 264807

**Summary:** Le Calcinai's 2007 Vernaccia di San Gimignano offers notable complexity in its white peaches, minerals, smoke and earthiness.

**Source:**

**Date:**

Le Calcinai's 2007 Vernaccia di San Gimignano offers notable complexity in its white peaches, minerals, smoke and earthiness. Medium in body, it possesses excellent persistence, and a long, refreshing finish. This is just about as good as Vernaccia gets. Anticipated maturity: 2008-2009.

A Marc de Grazia Selection, various American importers, including Michael Skurnik, Syosset, NY, tel: (516) 877-9300; Vin Dwino, Chicago, IL, tel: (773) 354-6700, and Estate Wines, Ltd., San Rafael, CA, tel: (415) 492-9411

**Affiliated entities:**

**ColorClass:** White

**Country:** Italy

**Descriptor:** body complexity earthiness finish medium minerals peaches refreshing smoke white

**Dryness:** 89

**IsBarrelTasting:** 0

**LabelName:** Vernaccia di San Gimignano

**Location:** Vernaccia di San Gimignano

**Maturity:** 4

**maturityShow:** Old

**Producer:** Calcinai, Tenuta le

**ProducerShow:** Tenuta le Calcinai

**Rating:** 87

**Region:** Tuscany

**Variety:** Vernaccia

**Vintage:** 2007



# List View

**Producer**

- Colombina, La
- Coltibuono
- Conti Costanti
- Corsini, Principe
- Corti Corsini, Le
- Corzano E Paterno
- Cupano
- D'Alessandro
- Degli dei, Tenuta
- Dei, Az Agr
- Del Cabreo, Tenute
- Duemani
- Falchini
- Fanti
- Farneto, Tenuta la
- Farnetella, Castello di
- Fattoria di Monsanto
- Fattoria Il Casalone Pepi Lignana
- Felsina, Fattoria di**
- Foloni, Ambrogio & Giovanni
- Fontaleoni
- Fonterutoli, Castello di
- Fonti, Le
- Fontodi
- Forte, Podere
- Fortediga
- Fossacolle
- Frescobaldi
- Fulgni
- Gabbiano, Castello di
- Gagliole
- Gerla, La
- Ghizzano, Tenuta di
- Grattamacco
- Gruppi, I
- Guado Al Melo
- Guado al Tasso
- Icano
- Il Borro
- Il Palazzino, Podere
- Isola e Olena
- Lisini
- Livernano
- Lohsa

**Variety**

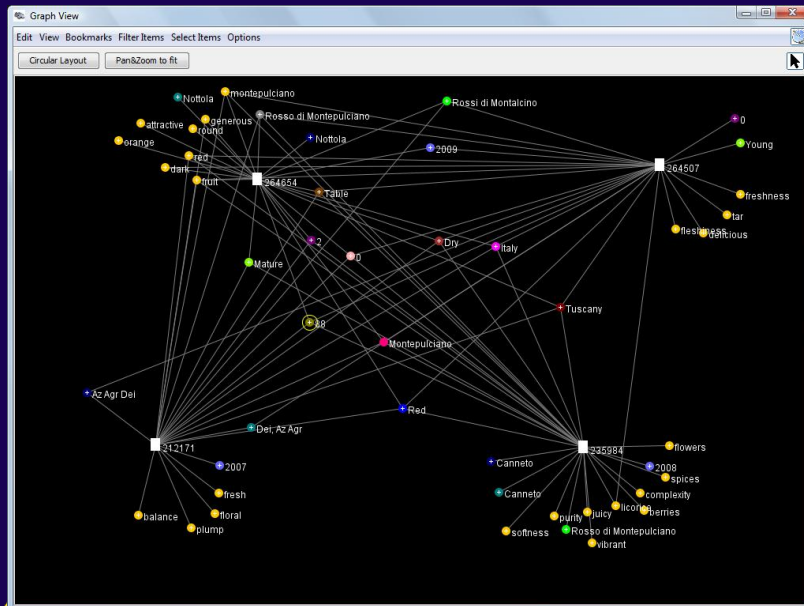
- Ansonica
- Caberlot
- Cabernet Franc
- Cabernet Sauvignon
- Cannaiolo
- Chardonnay
- Cleghio
- Colombo
- Merlot
- Montepulciano
- Monteregio Rosso
- Morelino
- Pinot Noir
- Proprietary Blend
- Pugnitello
- Sangiovese**
- Sangiovese Grosso
- Sauvignon Blanc
- Syrph
- Trebbiano
- Vermentino
- Vernacce
- Vignier

**Rating**

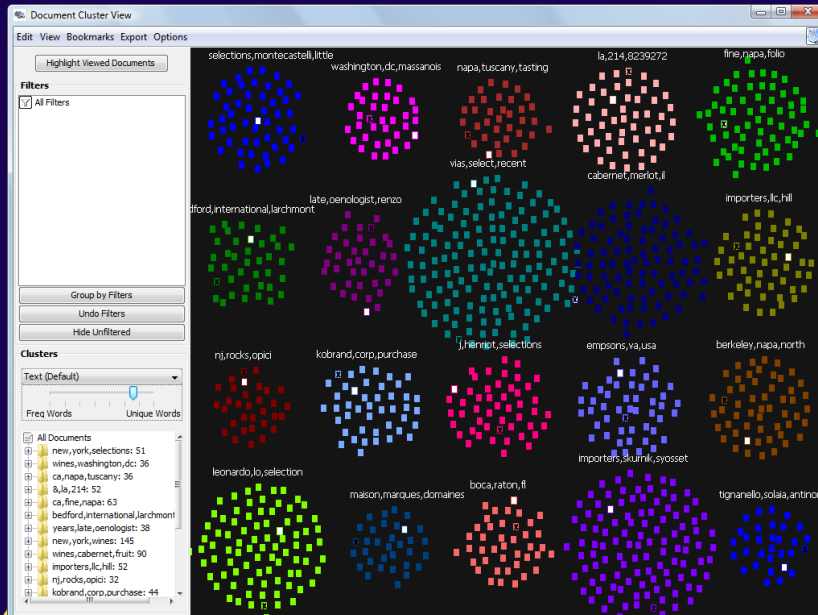
- 85
- 86
- 87
- 88
- 89
- 90
- 91
- 92
- 93
- 94
- 95
- 96
- 97
- 98



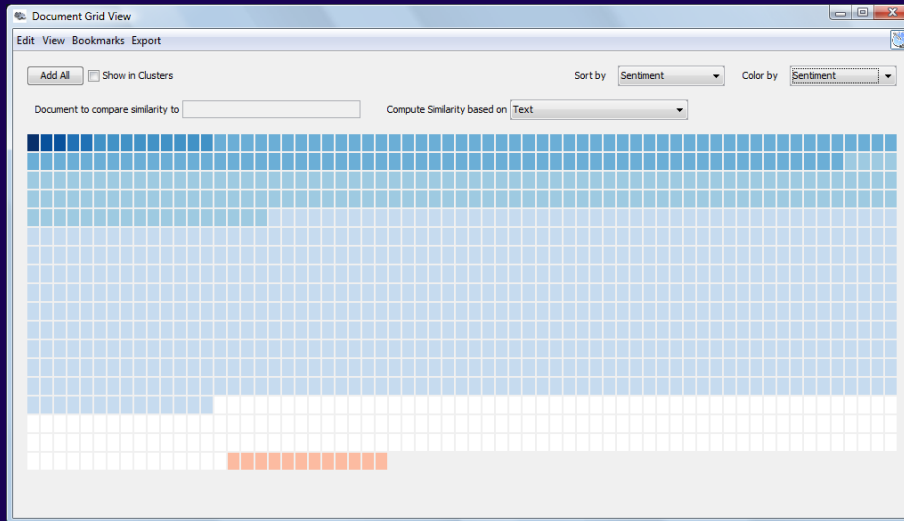
# Graph View



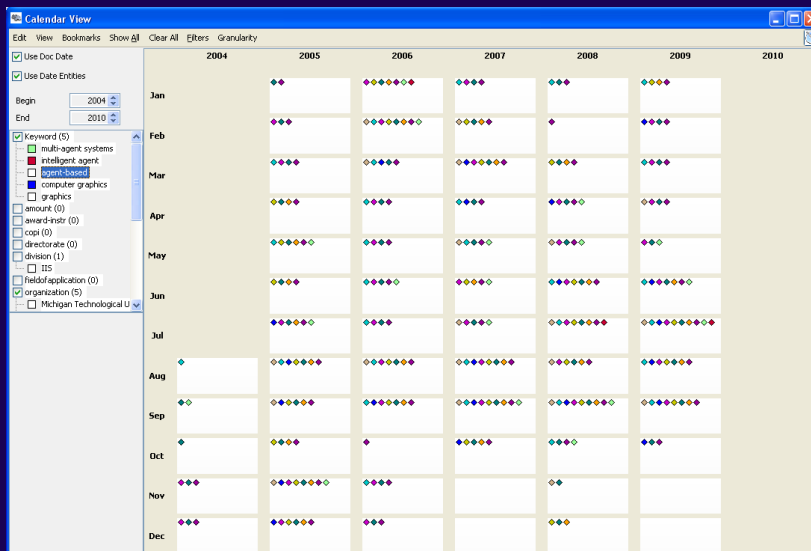
# Document Cluster View



# Document Grid View

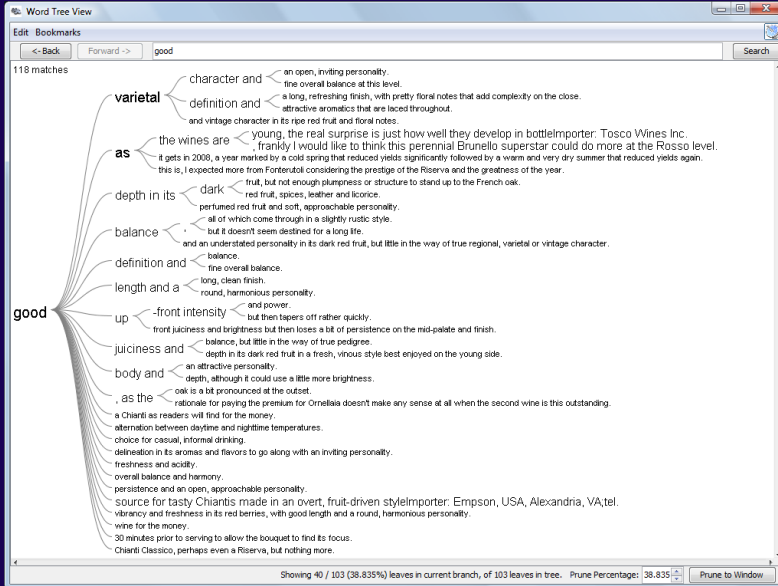


# Calendar View

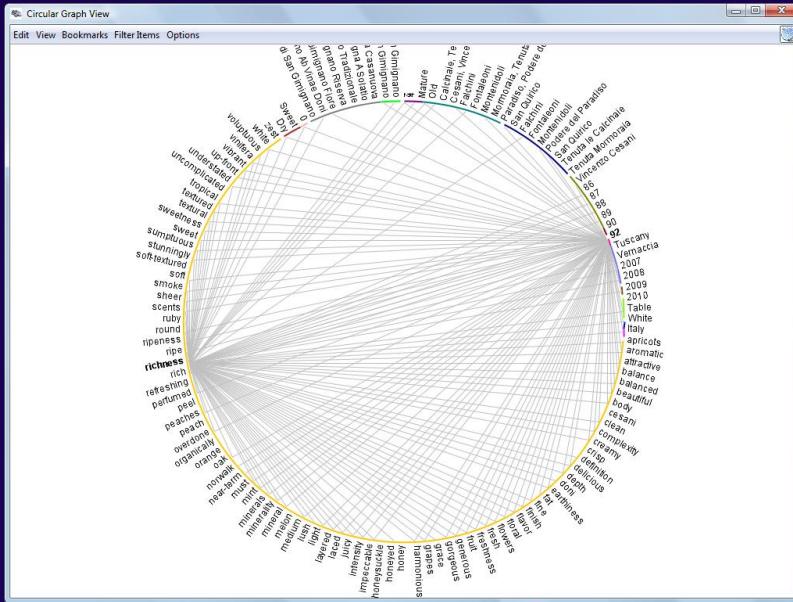




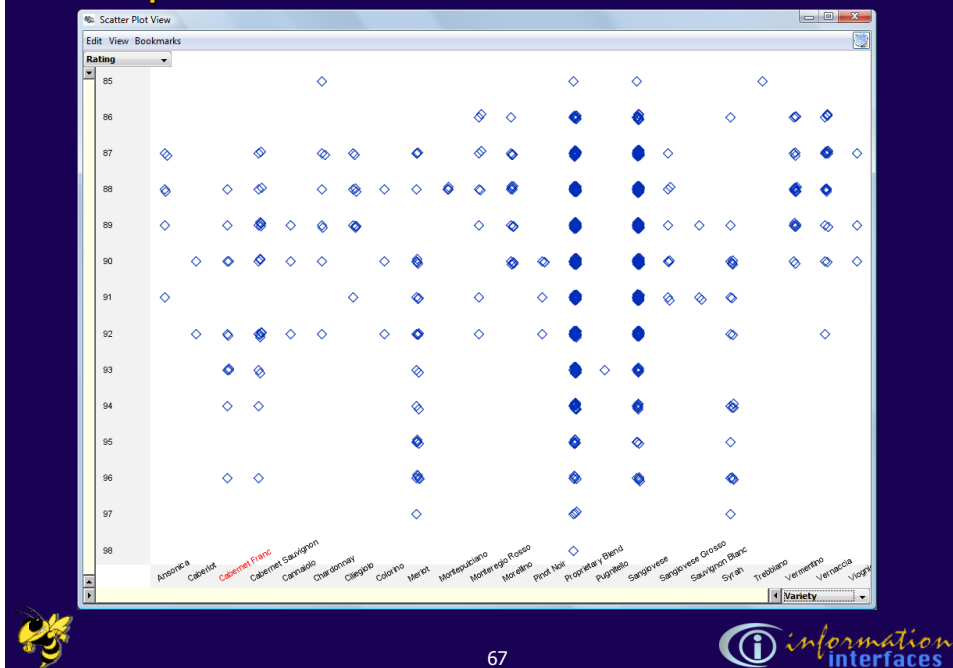
# WordTree View



# Circular Graph View



## Scatterplot View



## 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

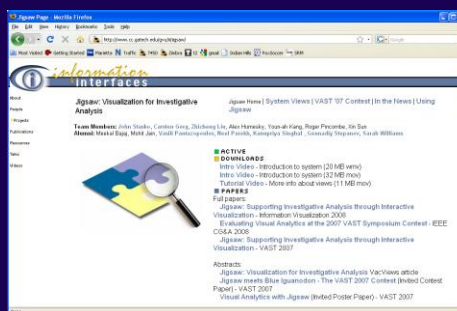


## To Learn More & Availability

<http://www.gvu.gatech.edu/ii/jigsaw>

Available for (free)  
trial use

Send email to:  
[stasko@cc.gatech.edu](mailto:stasko@cc.gatech.edu)



69



## 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



# 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



# Thanks!

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

