

Visual Analytics 2



CS 7450 - Information Visualization
November 28, 2012
John Stasko

Agenda

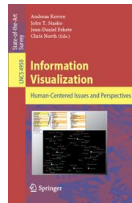


- Last time
 - Overview of what the term means and how it relates to information visualization
 - Some example VA research projects
- Today
 - Specific example, Jigsaw, helping investigative analysis
 - Related systems

VA Definition



- Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets



Keim et al, chapter in
*Information Visualization:
Human-Centered
Issues and Perspectives*, 2008

Application Area



- Investigative & Intelligence Analysis
 - Gather information from various sources then analyze and reason about what you find and know
 - Analyze situations, understand the particulars, anticipate what may happen



Definitions



- Thinking¹ - or reasoning - involves objectively connecting present beliefs with evidence in order to believe something else
- Critical Thinking¹ is a deliberate meta-cognitive(thinking about thinking) thinking act whereby a person reflects on the quality of the reasoning process simultaneously while reasoning to a conclusion.
- Intelligence¹ is a specialized form of knowledge, an activity, and an organization. As knowledge, intelligence informs leaders, uniquely aiding their judgment and decision-making. ...

1. *Critical Thinking and Intelligence Analysis: David Moore*

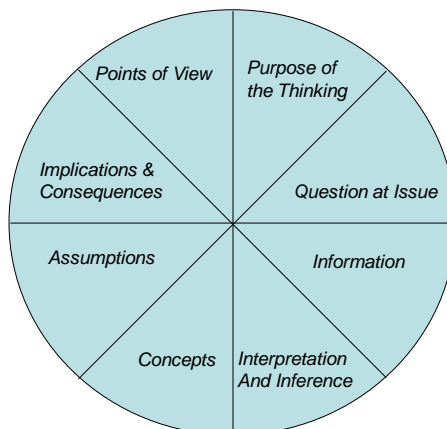
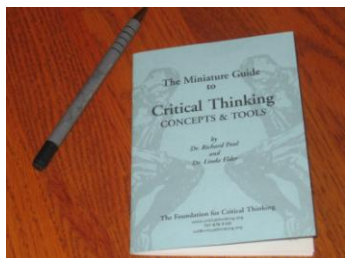


Critical Thinking*



“...the quality of our life and that of what we produce, make, or build depends precisely on the quality of our thoughts.”

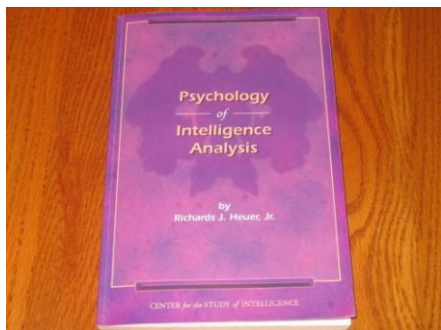
Elements of thought:



* Foundations of Critical Thinking www.criticalthinking.org



Example: Heuer's Central Ideas



- “Tools and techniques that gear the analyst’s mind to apply higher levels of critical thinking can substantially improve analysis... structuring information, challenging assumptions, and exploring alternative interpretations.”

CS 7450

7

Intelligence Process

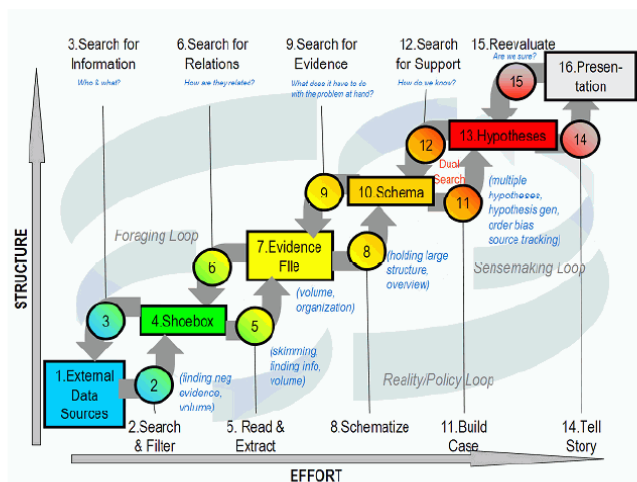


Figure 2.1. Notional model of sensemaking loop for intelligence analysis derived from CTA.

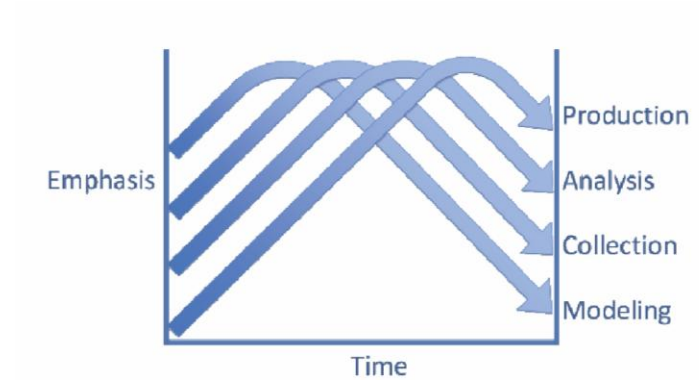
Pirolli & Card
Int'l Conf Intelligence Analysis '05

Fall 2012

CS 7450

8

Intelligence Process



Wheaton
In preparation

Fall 2012

CS 7450

9

Pain Points



- Cost structure of scanning and selecting items for further attention
- Analysts' span of attention for evidence and hypotheses

Fall 2012

CS 7450

10

HW 8 – Investigative Analysis



- Did you encounter those pain points?
- Discuss
 - How did you work on the problem?
 - What were the main challenges?

Fall 2012

CS 7450

11

Jigsaw

Stasko, Görg, Liu
Information Visualization '08



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"



Fall 2012

CS 7450

12

The Jigsaw Team



Current:

and many alumni

Carsten Görg
Zhicheng Liu
Youn-ah Kang
Jaeyeon Kihm
Chad Stolper

Fall 2012

CS 7450

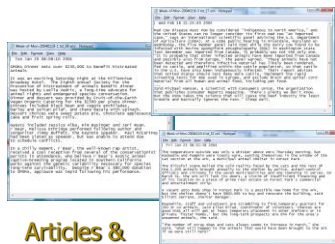
13

Problem Addressed

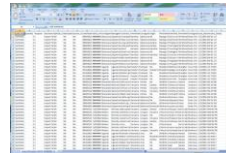
Analogy



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



Articles & reports



Spreadsheets



Blogs



XML documents

Fall 2012

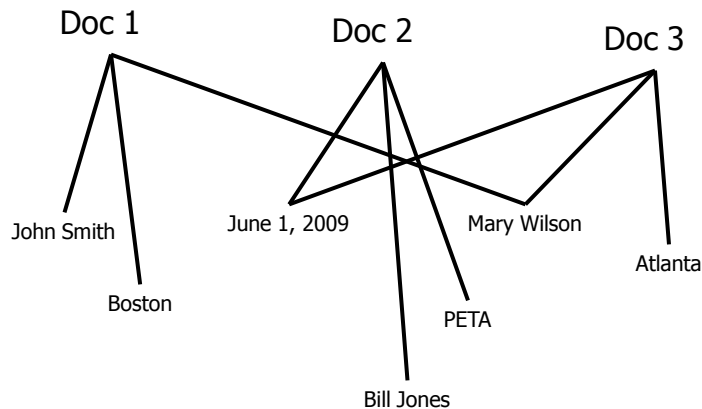
CS 7450

14

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



Entity Identification



- Must identify and extract entities from plain text documents
 - Crucial for our work
- Not our main research focus – We use tools from others

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.

Sample Document 2



Title: Proving Columbus was Wrong
Abstract: In this work, we show the world is really flat. To do this, we build a bunch of ships. Then we...
PI: Amerigo Vespucci
Co-PI: Vasco de Gama, Ponce de Leon
Organization: Northwest Central Univ.
Amount: 123,456
Program Mgr: Ephraim Glinert
Division: IIS
ProgramElementCode: 2860

Entities Already Identified



Title: Proving Columbus was Wrong

Abstract: In this work, we show the world is really flat. To do this, we build a bunch of ships. Then we...

PI: Amerigo Vespucci

Co-PI: Vasco de Gama, Ponce de Leon

Organization: Northwest Central Univ.

Amount: 123,456

Program Mgr: Ephraim Glinert

Division: IIS

ProgramElementCode: 2860

Entities

Connections



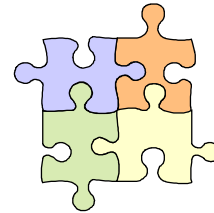
- Entities relate/connect to each other to make a larger “story”
- Connection definition:
 - Two entities are connected if they appear in a document together
 - The more documents they appear in together, the stronger the connection

Jigsaw

“Putting the pieces together”



- 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



Fall 2012

CS 7450

23

System Views

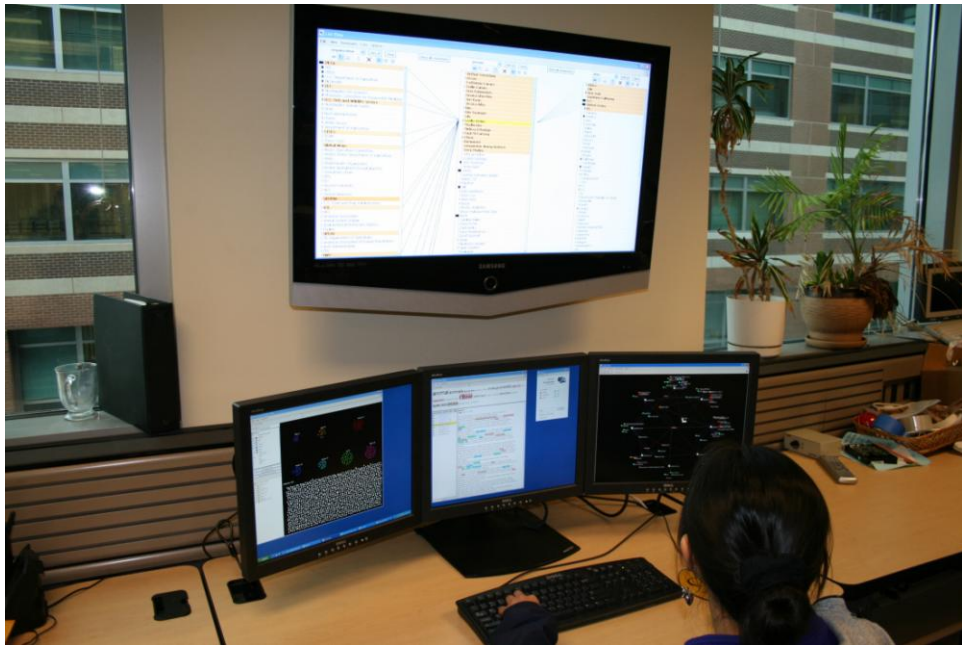
The collage displays several views from the Jigsaw system:

- Top Left:** A small overview window showing a document structure.
- Top Middle:** A hierarchical tree view of document sections.
- Top Right:** A network graph visualization with nodes and edges.
- Middle Left:** A text analysis window with a list of terms and their frequencies.
- Middle Center:** A window showing a grid of colored bars representing data points.
- Middle Right:** A window displaying a table of data with multiple columns.
- Bottom Left:** A window showing a detailed view of a document's content.
- Bottom Center:** A window showing a network graph with nodes colored by cluster.
- Bottom Right:** A window showing a circular network graph visualization.

Fall 2012

CS 7450

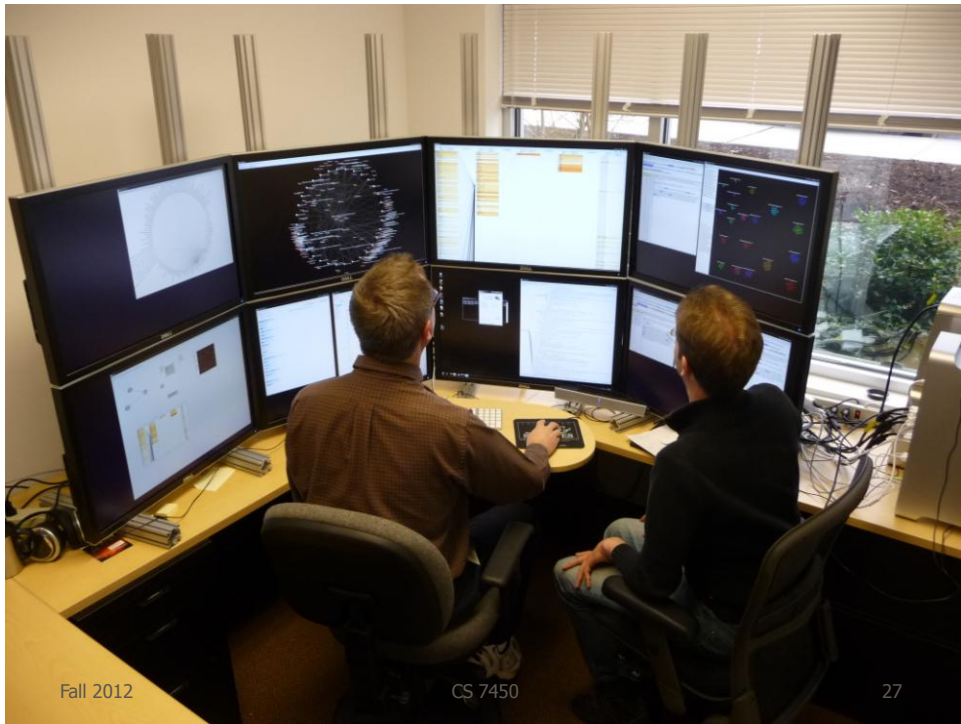
24



Fall 2012

CS 7450

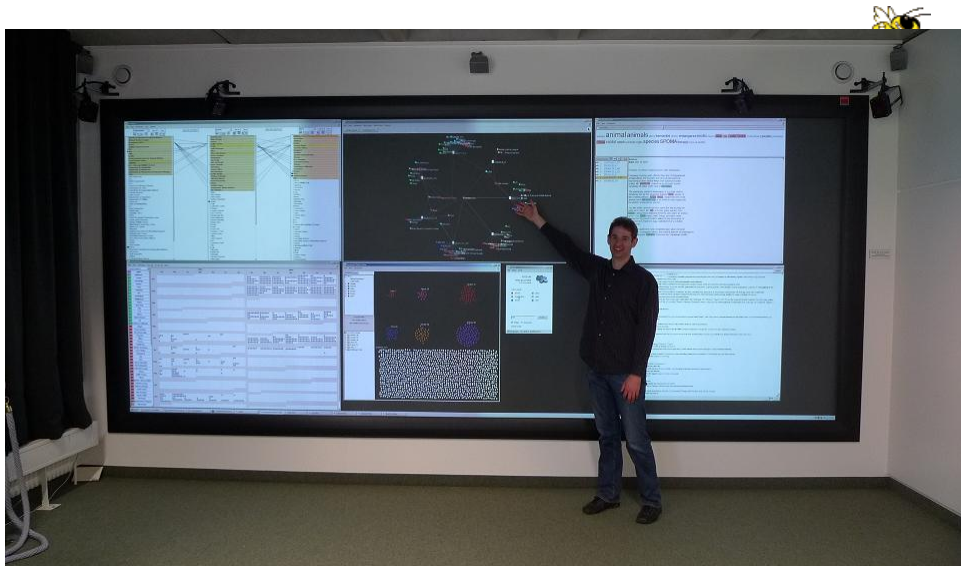
26



Fall 2012

CS 7450

27



Fall 2012

CS 7450

28

Console

Entity types

Fall 2012

CS 7450

29

Document View

Important words in loaded docs

Automatic summary

Entities identified

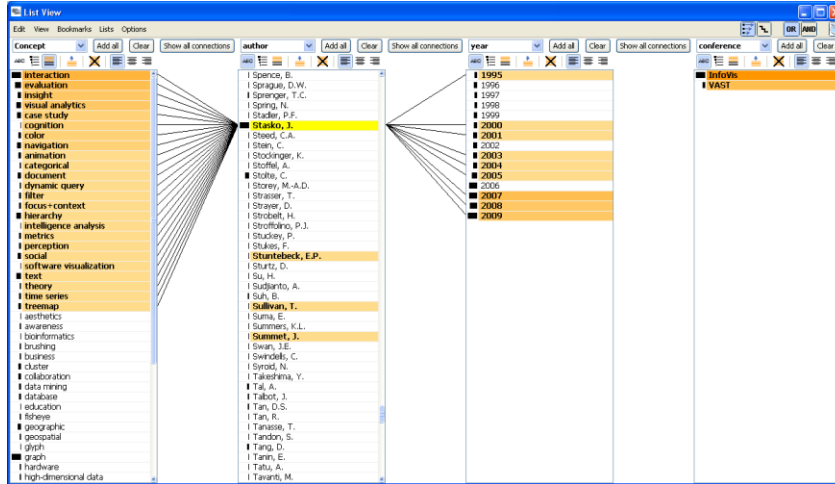
Fall 2012

CS 7450

30

List View

Lists of entities by type
Connections highlighted



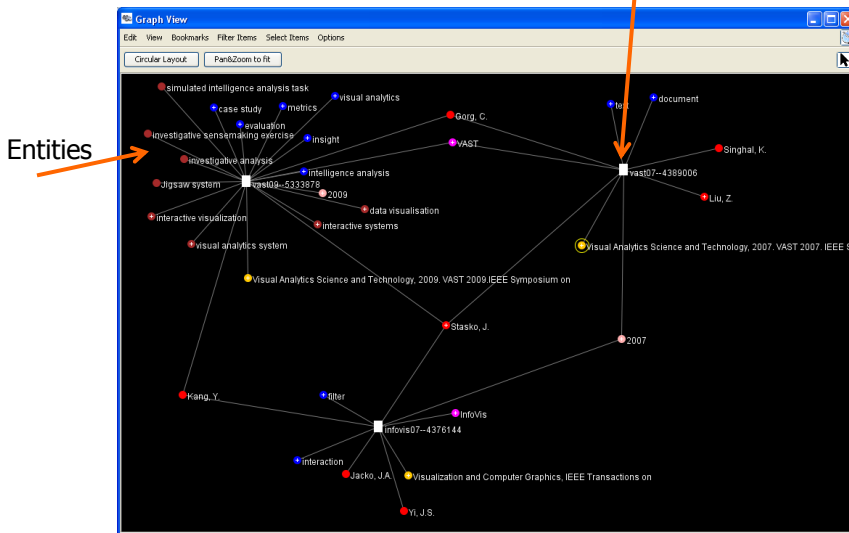
Fall 2012

CS 7450

31

Graph View

Document



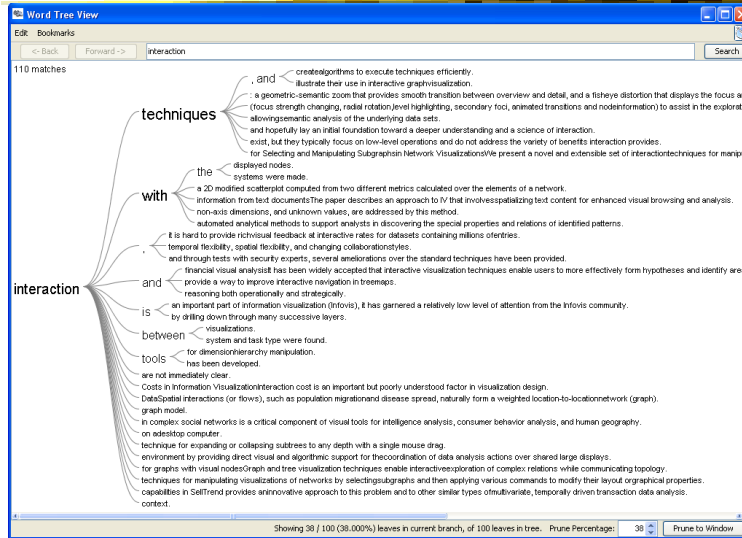
Fall 2012

CS 7450

32

WordTree View

Context of a word in the collection



Fall 2012

CS 7450

33

Document Cluster View

Clustered by document text or by entities

Summarized by three words

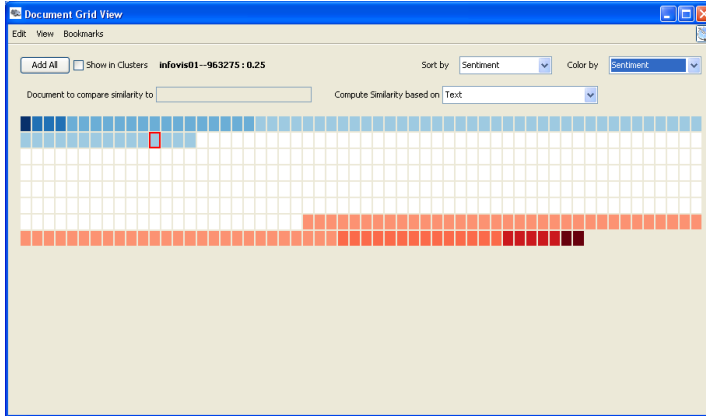


Fall 2012

CS 7450

34

Document Grid View



User controls order and color

Sentiment analysis shown here

Fall 2012

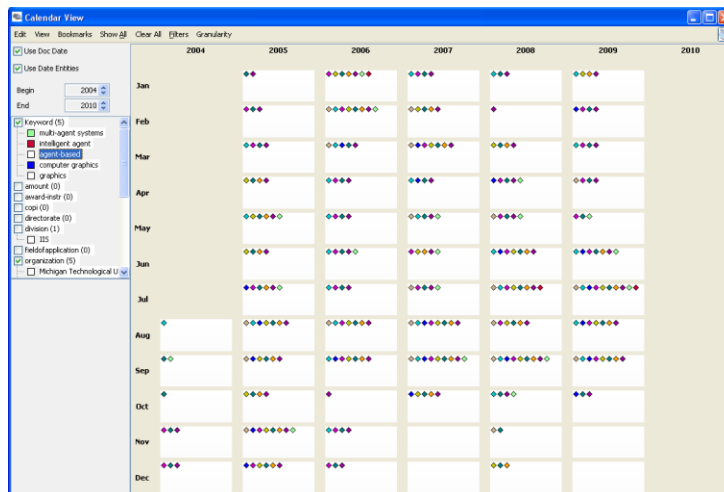
CS 7450

35

Calendar View



Showing connections between entities and dates



Fall 2012

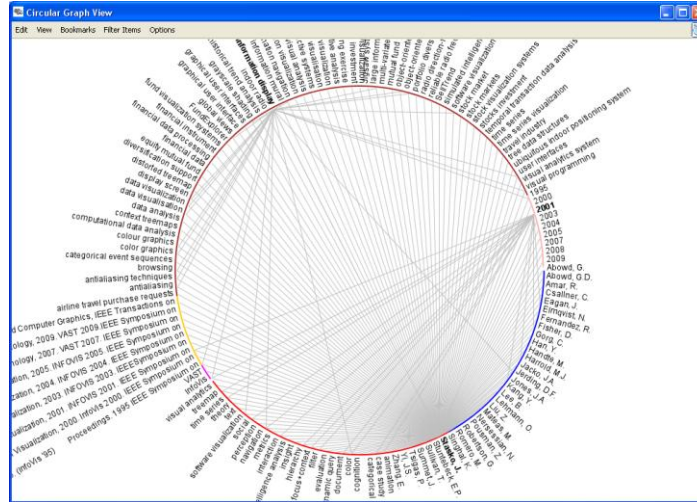
CS 7450

36

Circular Graph View



Connections between entities



Fall 2012

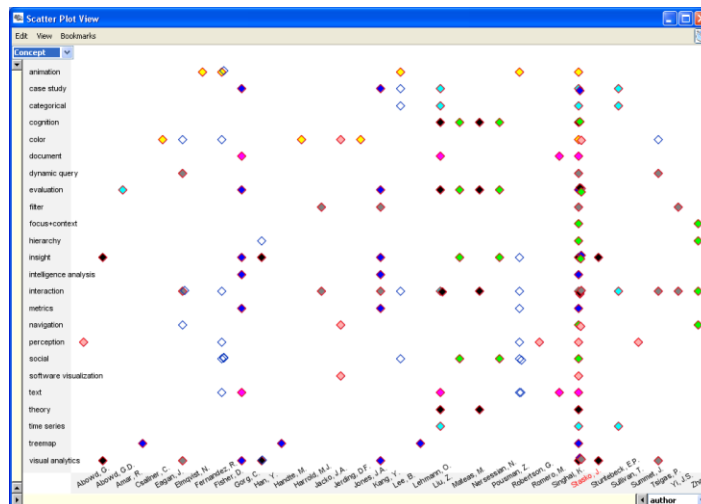
CS 7450

37

Scatterplot View



Documents containing pairs of entities



Fall 2012

CS 7450

38

Demo 1



- Data from HW 8
- Let's find the bad guys!

Demo 2



- Car reviews
 - Text: Consumer's comments
 - Entities: Various ratings (1-10), car features, other makes & models

Demo 3



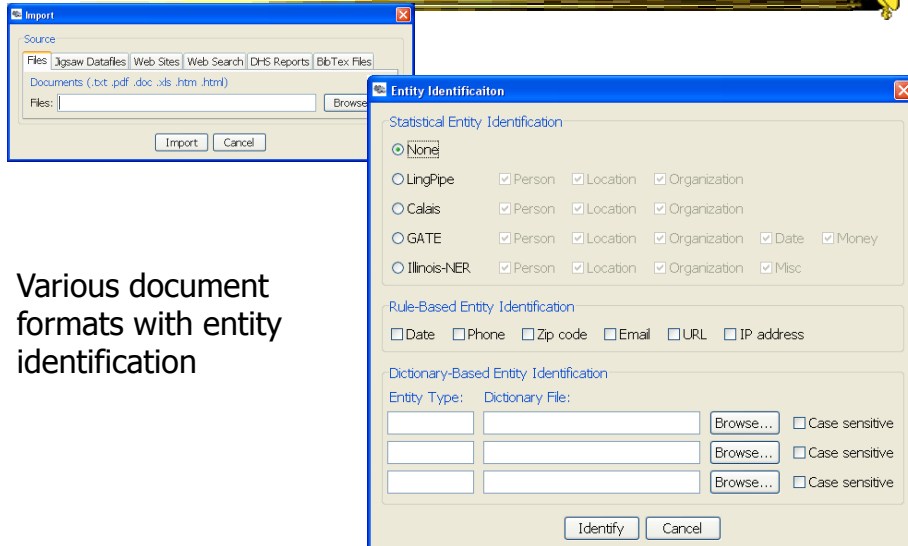
- InfoVis & VAST papers
 - Text: paper title and abstract
 - Entities: author, keyword, year, conference, “concept”

Computational Analyses



- Document summarization
- Document similarity
- Document clustering by content
 - Text or entities
- Sentiment analysis

Document Import



Various document formats with entity identification

Fall 2012

CS 7450

43

Input Data Formats



- Text, pdf, Word, html, Excel
- Jigsaw data file format
 - Our own xml
- DB?
 - Go to Excel
 - Go to text, transform to Jigsaw data file

Fall 2012

CS 7450

44

```

<award>
<awardnumber>0640291</awardnumber>
<title>SGER: Distributed Spatial Partitioning Algorithms for Scalable Processing of Mobile
<nsfororganization>IIS </nsfororganization>
<programs>DATA MANAGEMENT SYSTEMS</programs>
<startdate>September 1, 2006</startdate>
<lastamendmentdate>September 12, 2007</lastamendmentdate>
<principalinvestigator>Liu, Ling</principalinvestigator>
<state>GA</state>
<organization>GA Tech Research Corporation - GA Institute of Technology </organizatio
<awardinstrument>Standard Grant </awardinstrument>
<programmanager>Le Gruenwald </programmanager>
<expirationdate>February 29, 2008</expirationdate>
<awardedamounttodate>65502</awardedamounttodate>
<co_pinames></co_pinames>
<piemailaddress>lingliu@cc.gatech.edu
<organizationstreetaddress>Office of Sponsored Programs </organizationstreetaddress>
<organizationcity>Atlanta </organizationcity>
<organizationstate>GA</organizationstate>
<organizationzip>30332</organizationzip>
<organizationphone>4048944819</organizationphone>
<nsfdirectorate>CSE </nsfdirectorate>
<programelementcodes>7485</programelementcodes>
<programreferencecodes>HPCC|9218|7484</programreferencecodes>
<fieldofapplications>0104000 Information Systems </fieldofapplicati
<awardnumber>0640291</awardnumber>
<abstract>IIS-0640219 Ling Liu <lt;:lingliu@cc.gatech.edu>gt; Georgia Institute of Instit
</award>

```

Scraped XML

Fall 2012

CS 7450

45

```

<document>
<docID>0808863</docID>
<docDate>July 1, 2008</docDate>
<docSource></docSource>
<docText>FODAVA-Lead: Dimension Reduction and Data Reduction: Foundations for Visualization

FODAVA-Lead: Dimension Reduction and Data Reduction: Foundations for Visualization The FODAVA (Foundations of
Data Analysis and Visualization) Lead research team at the Georgia Institute of Technology provides unified
expertise in the critical areas for providing leadership of the FODAVA effort, including machine learning and
computational statistics, information visualization, massive-dataset algorithms and data structures, and
optimization theory. The team is focused on the fundamental theory and approaches to make breakthroughs in data
representations and transformations. The work is directed along the two main axes of scale reduction, data reductio
<directorate>CSE</directorate>
<award-instr>Continuing grant</award-instr>
<programreferencecode>HPCC</programreferencecode>
<programreferencecode>9218</programreferencecode>
<keyword>visualization</keyword>
<keyword>algorithms</keyword>
<fieldofapplication>0000912 Computer Science</fieldofapplication>
<state>GA</state>
<organization>GA Tech Research Corporation - GA Institute of Technology</organization>
<keyword>data analysis</keyword>
<keyword>information visualization</keyword>
<keyword>machine learning</keyword>
<amount>1200000</amount>
<pi>Park, Haesun</pi>
<copi>John Staasko</copi>
<copi>Alexander Gray</copi>
<copi>Renato D. C. Monteiro</copi>
<copi>Vladimir Koltchinskii</copi>
<progmgr>Lawrence Rosenblum</progmgr>
<division>CCF</division>
<keyword>visual analytics</keyword>
<programelementcode>I114</programelementcode>
<programelementcode>H194</programelementcode>
</document>

```

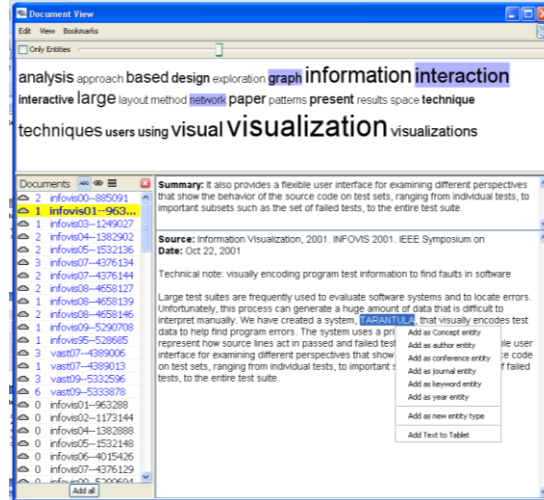
Jigsaw Datafile Format

Fall 2012

CS 7450

46

El Correction

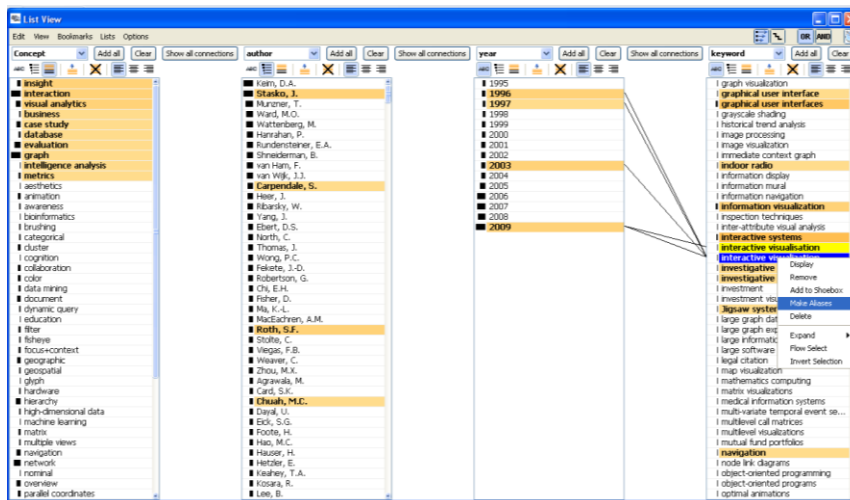


Fall 2012

CS 7450

47

Entity Aliasing

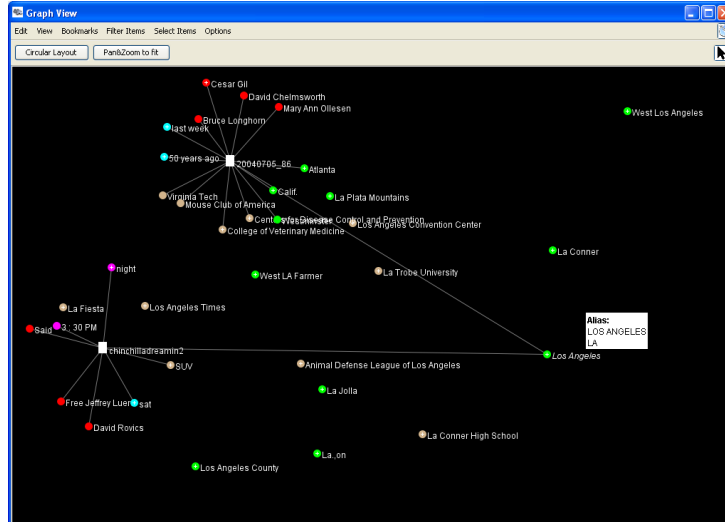


Fall 2012

CS 7450

48

Alias Representation

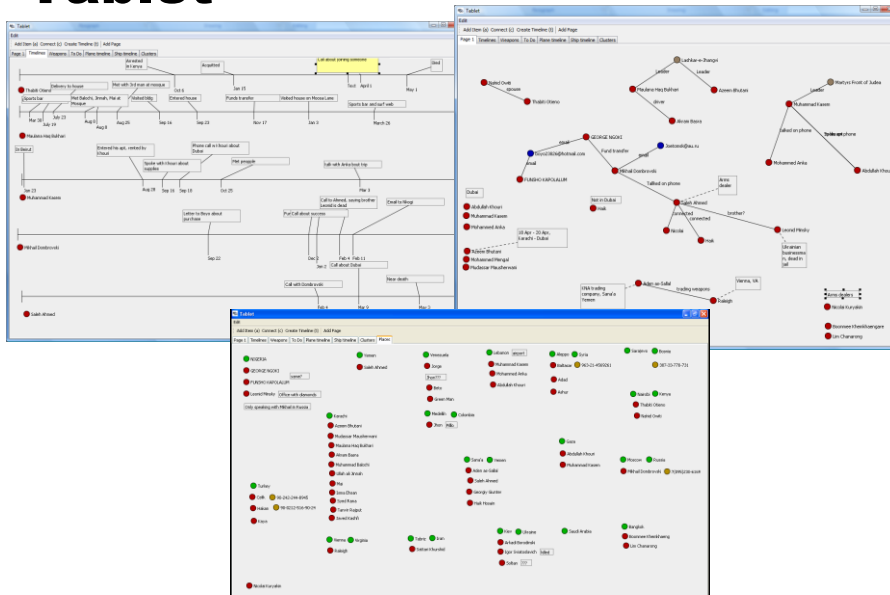


Fall 2012

CS 7450

49

Tablet



Fall 2012

CS 7450

50

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

Fall 2012

CS 7450

51

Potential Jigsaw Future Work



- Collaborative capabilities
- Improved evidence marshalling
- Present/browse investigation history
- Scalability upward
- Web document ingest
- Implement network algorithms
- DB import
- Wikipedia & Intellipedia
- Geospatial view
- Better timeline capabilities
- Reliability/uncertainty
- Other types of data
- Active crawling/RSS ingest
- Try it on display wall
- Deployment to real clients

Fall 2012

CS 7450

52

Room to Improve



- What Jigsaw doesn't do so well now
 - The end-part of the Pirolli-Card model
 - Helping the analyst take notes, organize evidence, generate hypotheses, etc.
(The Tablet is a first step)
 - Sometimes called "evidence marshalling"
 - Others have focused more on that aspect...

Fall 2012

CS 7450

53

i2's Analyst Notebook



The screenshot shows the i2 Analyst Notebook website. At the top, there is a navigation bar with links for 'Contact Us', 'Download Center', 'Info Request', and a search box. Below this is a secondary navigation bar with 'Home', 'Company', 'Products', 'Solutions', 'Services', 'Partners', and 'Support'. The main content area is titled 'i2 Analyst's Notebook Powering Analysis' and includes a description of the software's capabilities. A sidebar on the right lists various products like 'Analytical Capabilities', 'Online Data Analysis', and 'iBase'. The bottom of the page has a 'Done' button.

Fall 2012

CS 7450

54

Analyst's Notebook



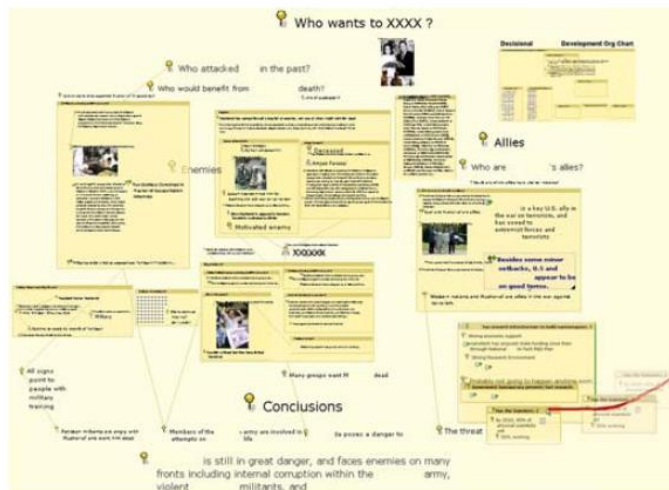
- Leading commercial tool in this space (law enforcement and intelligence agencies)
- Large zooming workspace where analyst creates networks of entities and notes
- Often used to produce presentation or story of analysis done

Fall 2012

CS 7450

55

Oculus' Sandbox



Video

Wright et al
CHI '06

Fall 2012

CS 7450

56

Sandbox



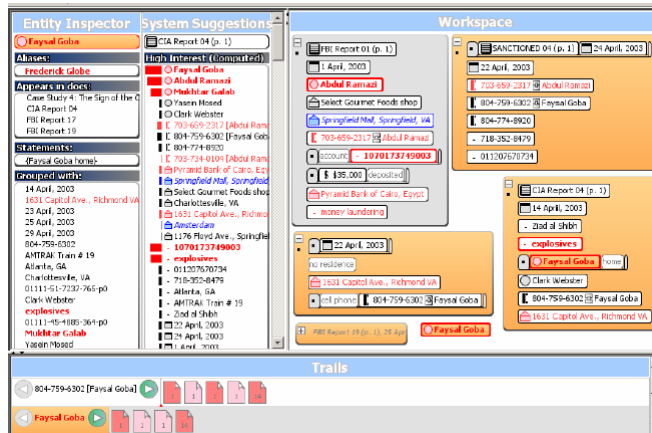
- Flexible space for inserting text and graphics
- Objects can be dragged-and-dropped from their other analysis tools
- Flexible level of detail
- Flexible gestures for making space, inserting, etc.
- Assertions with evidence gates
- Reasoning templates

Fall 2012

CS 7450

57

PARC's Entity Workspace



Video

Bier, Card & Bodnar
VAST '08

Fall 2012

CS 7450

58

Entity Workspace



- Tools for rapid ingest of entities from documents
- Can snap together entities into groups
- Can indicate level of interest in objects
- Four main view panels, with zooming UI

Fall 2012

CS 7450

59

Related Area of Interest



- Sensemaking
- A general term that has been used in a number of different contexts
 - E.g., How large corporations make decisions
- To me, ultimately about people working with data and information to understand it better

Fall 2012

CS 7450

60

Sensemaking



Nice definition:

“A motivated , continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively.”

– Klein, Moon and Hoffman
IEEE Intelligent Systems '06

Alternate Definition



“The process of creating situation awareness in situations of uncertainty”

– D. Leedom, '01 SM Symp. Report

Situation awareness:

“It’s knowing what’s going on so you know what to do”

– B. McGuinness, quoting an Air Force pilot

This Topic



- I work on it a lot now
- Interested in getting more work in this area started

Fall 2012

CS 7450

63

Upcoming



- Evaluation
 - Reading
Carpendale '08
- Review & Wrap-up
 - Reading
Few chapter 13
Heer et al '10

Fall 2012

CS 7450

64