## **Visual Analytics 2**

CS 7450 - Information Visualization December 1, 2011 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

Review

# **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



CS 7450

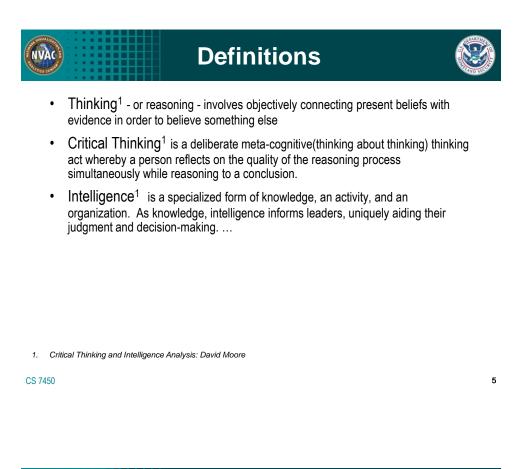
Keim et al, chapter in Information Visualization: Human-Centered Issues and Perspectives, 2008

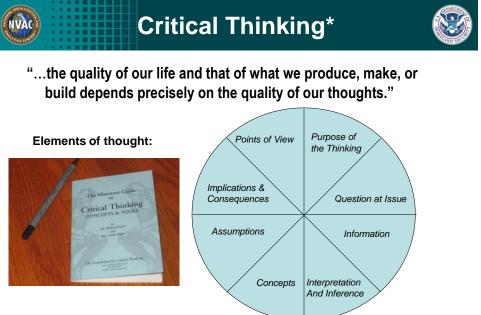
3

Fall 2011

## **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



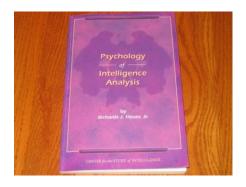


\* Foundations of Critical Thinking www.criticalthinking.org CS 7450

## Example: Heuer's Central Ideas



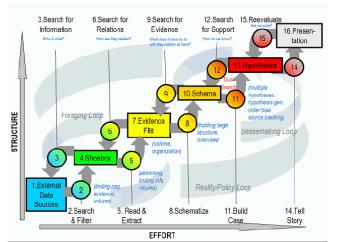
7



 "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

NVAC



Intelligence Process

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

Pirolli & Card Intl Conf Intelligence Analysis '05

Fall 2011

## **Pain Points**

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

CS 7450

#### HW 8 – Investigative Analysis

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

CS 7450

## Jigsaw



11

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 2011

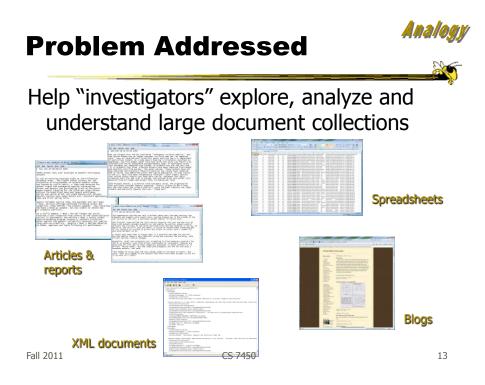
CS 7450

## **The Jigsaw Team**

Current:

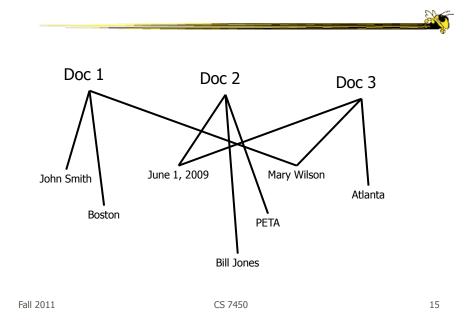
Carsten Görg Zhicheng Liu

Youn-ah Kang Jaeyeon Kihm Chad Stolper and many alumni



## **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 -  ${\tt A}$  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.

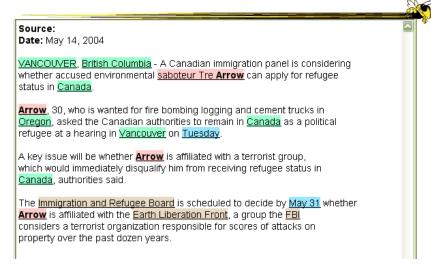
The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.

Fall 2011

CS 7450

17

#### **Entities Identified**



## 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

Fall 2011

CS 7450

19

## **Entities Already Identified**

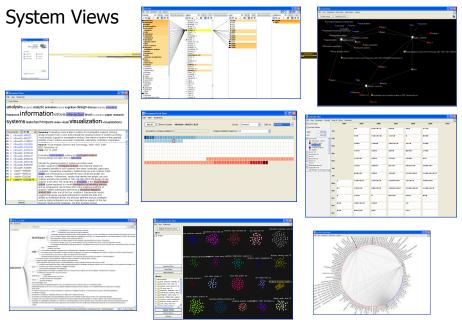
Title: Proving Columbus was Wrong Abstract: In this work, we show the world is really flat. To do this, we build a bunch of ships. Then we... PI: Amerigo Vespucci Co-PI: Vasco de Gama, Ponce de Leon Organization: Northwest Central Univ. Amount: 123,456 Program Mgr: Ephraim Glinert Division: IIS ProgramElementCode: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

## Connections



- Connection definition:
  - Two entities are connected if they appear in a document together
  - The more documents they appear in together, the stronger the connection

Fall 2011	CS 7450	21
Jigsaw	"Putting the pi	eces together"
<ul> <li>Entity id clustering</li> <li>Multiple document and their</li> </ul>	ational analysis of docum dentification, document sim ng, summarization, sentime visualizations (views) of nts, analysis results, enti- r connections e highly interactive and ted	ilarity, ent
Fall 2011	CS 7450	22



Fall 2011

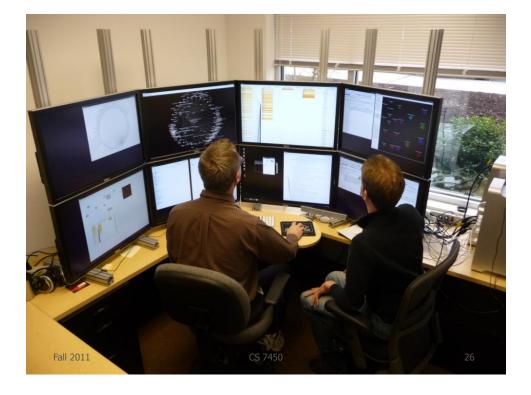
23





Fall 2011

CS 7450





# 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

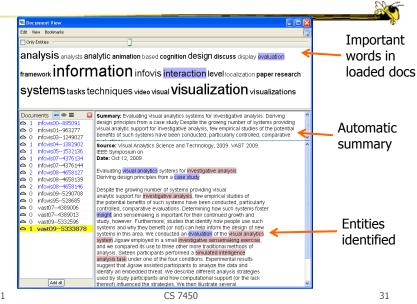
CS 7450

29

#### Console

	🛸 Jigsaw	
	<u>File Views Entities Tools</u>	
	JIGSAW infovis-vast 512 documents	S.
	Color Legend:	
Entity	author (1017)	ept (77)
Entity types	conference (2)	xterm (1790)
, p	📕 journal (17) 🔲 keyu	vord (1202)
	🔲 year (16)	
		Search
	Entities     Documents	
	Workspace: no active workspace	

#### **Document View**

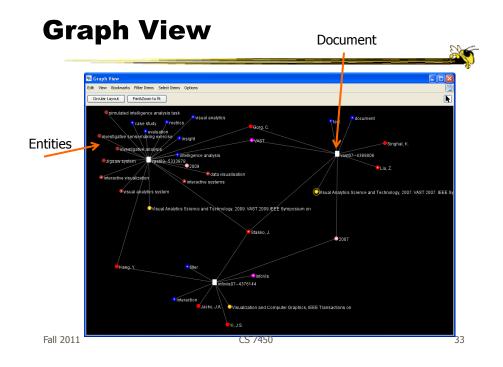


Fall 2011

**List View** 

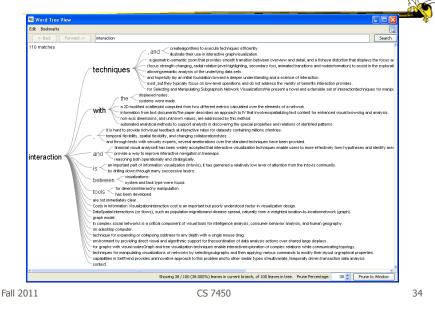
#### Lists of entities by type Connections highlighted

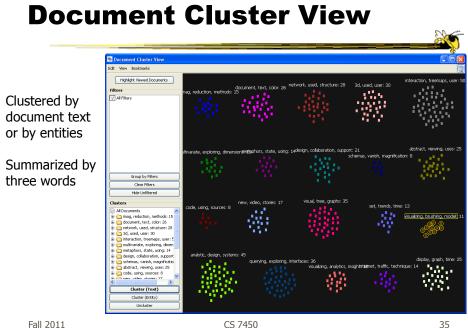
List View									
dit View Bookmarks Lists Options									😴 🐛 🛛 🛛 👌
oncept 🛛 🖌 Add all 🛛 Clear	Show all connections	author 🔽	Add all Clear	Show all connections	year	× A	dd all Clear	Show all connections	conference 🗸 Add al Clear
• H 🔳 📥 🗙 🔳 🗐 🗉		🗝 🗄 🚞 🔔 🚺			A40 🗄 🚞	1 ×			
Interaction		I Spence, B.		a .	1995				InfoVis
evaluation		I Sprague, D.W.			1996				VAST
insight		I Sprenger, T.C.			1997				1 1951
visual analytics		I Spring, N.			1998				
case study		I Stadler, P.F.			1999				
cognition		Stasko, J.		<u> </u>	2000				
color		I Steed, C.A.			2001				
navigation		I Stein, C.			2002				
animation		I Stockinger, K.			2003				
I categorical		I Stoffel, A.			2004				
document	///////////////////////////////////////	Stolte, C.			2005				
dynamic query	////////	I Storey, MA.D.		///	2006				
filter	////////	I Strasser, T.			2007				
focus+context	///////	I Strayer, D.			2008				
hierarchy	///////	I Strobelt, H.			2009				
intelligence analysis		I Stroffolino, P.J. I Stuckey, P.							
metrics perception	(/////	I Stuckey, P. I Stukes, F.							
sorial	/////	Stuntebeck, E.P.							
software visualization	(////	I Sturtz, D.							
text	///	I Su. H.							
theory	//	I Sudjanto, A.							
time series	/	Sub, B.							
treemap		Sullivan, T.							
Laosthotics		I Suma, E.							
awareness		I Summers, K.L.							
I bioinformatics		Summet, J.							
I brushing		I Swan, J.E.							
I business		I Swindels, C.							
cluster		I Syroid, N.							
collaboration		I Takeshima, Y.							
I data mining		Tal, A.							
database		I Tabot, J.							
education		I Tan, D.S.							
fisheye		I Tan, R.							
geographic		I Tanasse, T.							
geospatial		I Tandon, S. Tang, D.							
l glyph I graph		Tang, D. Tanin, E.							
l graph I hardware		Tanin, E. Tatu, A.							
I high-dimensional data		I Tatu, A. I Tavanti, M.							
i righ-cimensional casa		T Tavano, M.							



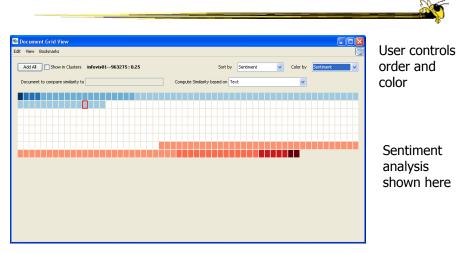
#### WordTree View

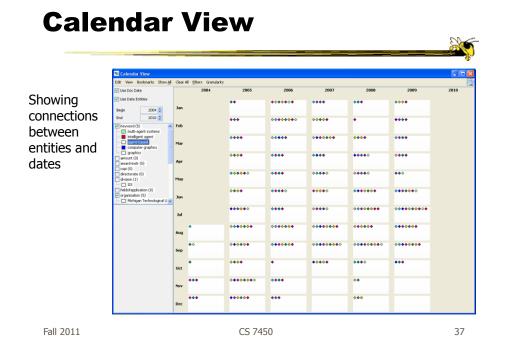
Context of a word in the collection



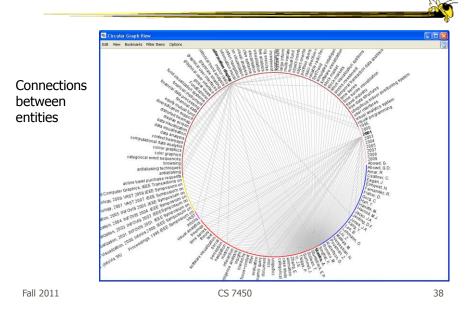


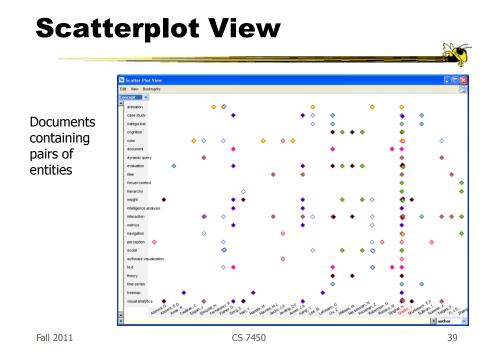






## **Circular Graph View**





## **Computational Analyses**

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

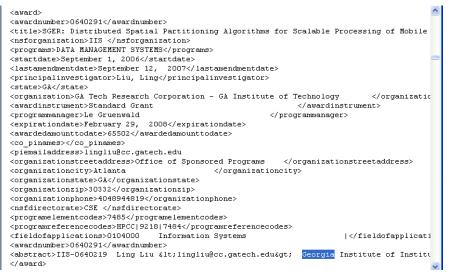
Source Files Jagsaw Datafiles Web Stes Web Search DH5 Reports BbTex I		)rt x
Documents (.txt .pdf .doc .xls .htm .html)	🛸 Entity Identificai	iton 🛛 🔀
Files: Brow	Statistical Entity	Identification
Import Cancel	⊙ None	
		Deres I eacher Commission
	○ LingPipe	Person Location Organization
	<ul> <li>Calais</li> </ul>	🗹 Person 🕑 Location 🕑 Organization
	○ GATE	Person Location V Organization V Date V Money
Various document	◯ Illinois-NER	🗹 Person 🕑 Location 🕑 Organization 🔽 Misc
formats with entity identification		ty Identification one ⊡Zip code □Email □URL □IP address d Entity Identification
		Dictionary File:
	Lindey Type.	Browse Case sensitive
		Browse
		Browse Case sensitive
		Identify Cancel
Fall 2011	CS 745	50 41

## **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

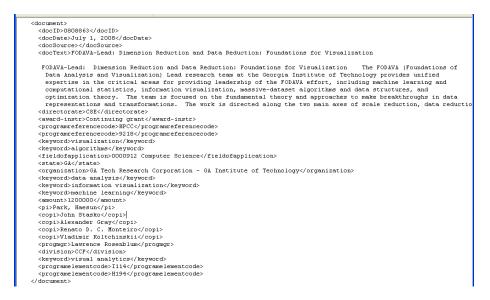


Scraped XML

Fall 2011

CS 7450

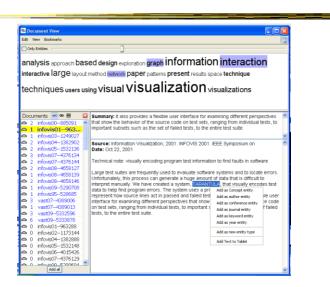
43



#### Jigsaw Datafile Format

CS 7450

## **EI** Correction



Fall 2011

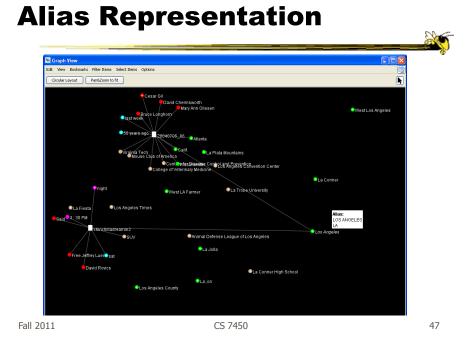
CS 7450

45

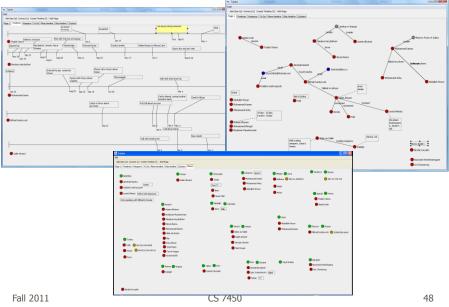
#### **Entity Aliasing**

List View				
it View Bookmarks Lists Options				🖬 ⊾ 🛛 🕅 AND
ncept 🛛 🖌 🖌 🖌	Show all connections author Add all Clear		Add all Clear Show all connection	
• 🗄 🔳 📥 🛛 🗙 🛛 🗮 🗮 🗃	ARC 🗮 🚍   📥   🗙   🌉 🚍 🚍	🛲 🏣 📄 📥	× ≡ = =	🔤 🗮 🗮   📥   🗙   🌉 🗮 🖷
l insight ***	Keim, D.A.	1995		I graph visualization
interaction	Stasko, J.	1996		graphical user interface
visual analytics	Munzner, T.	1997		graphical user interfaces
l business	Ward, M.O.	1998		I grayscale shading
case study	<ul> <li>Wattenberg, M.</li> </ul>	1999		I historical trend analysis
database	Hanrahan, P.	■ 2000		I image processing
evaluation	Rundensteiner, E.A.	2001		I image visualization
graph	Shneiderman, B.	2002		I immediate context graph
intelligence analysis	van Ham, F.	2003		indoor radio
metrics	van Wijk, J.J.	■ 2004		I information display
l aesthetics	Carpendale, S.	2005		Linformation mural
animation	Heer, J.	2006		I information navigation
I awareness	Ribarsky, W.	2007		I information visualization
I bioinformatics	Yang, J.	2008		I inspection techniques
I brushing	Ebert, D.S.	2009		l inter-attribute visual analysis
I categorical	North, C.			I interactive systems
duster	Thomas, J.			Interactive visualisation
I cognition	Wong, P.C.			I interactive visualization Display
	Fekete, JD.			Investigative
color	Robertson, G.     Chi, E.H.			
I data mining document	Fisher, D.			Prog to proceed.
document I dynamic query	Ma, KL.			Linvestment visu Jicsaw system
I education	Ma, KL. MacEachren, A.M.			Large graph dat Delete
l fiter	Roth, S.F.			l large graph dat
I fisheve	Stote, C.			I large informatic Expand
focus+context	Viegas, F.B.			I large software Flow Select
geographic	Weaver, C.			I legal citation Invert Selection
l geographic I geospatial	Zhou, M.X.			I regal citation Invert Selection I map visualization
I dvph	Agrawala, M.			I mathematics computing
I hardware	Card, S.K.			I matrix visualizations
hierarchy	Chuah, M.C.			I medical information systems
I high-dimensional data	Daval, U.			I multi-variate temporal event se
I machine learning	Firk, S.G.			I multilevel call matrices
I matrix	Foote, H.			I multilevel visualizations
multiple views	Hap, M.C.			I mutual fund portfolios
navigation	Hauser, H.			navigation
network	Hetzler, E.			I node link diagrams
Inominal	Keahey, T.A.			I object-oriented programming
oveniew	Kosara, R.			I object-oriented programs
parallel coordinates	Lee, B.			I optimal animations

Fall 2011



## Tablet



## **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

Fall 2011

- 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

CS 7450

**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

CS 7450





## **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 2011

CS 7450

51

## i2's Analyst Notebook



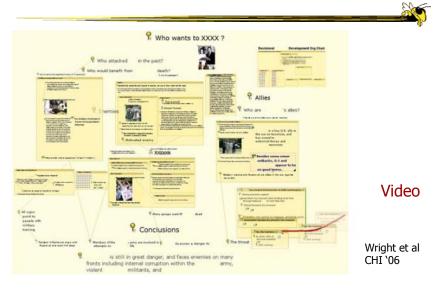
Fall 2011

## **Analyst's Notebook**

**Oculus' Sandbox** 

- 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 2011	CS 7450	53



Fall 2011

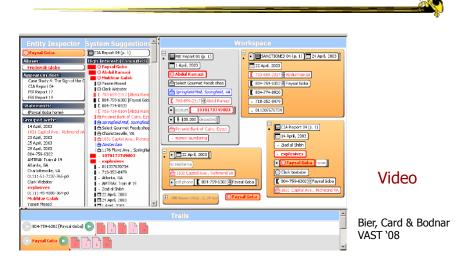
## Sandbox

- 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 2011

CS 7450

## **PARC's Entity Workspace**



CS 7450

## **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	2011	

CS 7450

**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

CS 7450

#### 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

Fall 2011

CS 7450

## **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

CS 7450

## **This Topic**

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

Fall 2011	CS 7450

## Upcoming

- Animation

   Reading
   Heer & Robertson '07
- Review & Wrap-up
  - Reading
     Few chapter 13
     Heer et al '10

CS 7450