Dec. 3, 2009

Developing Visual Analytics Applications: Lessons Learned from the Trenches

John Stasko & Carsten Görg

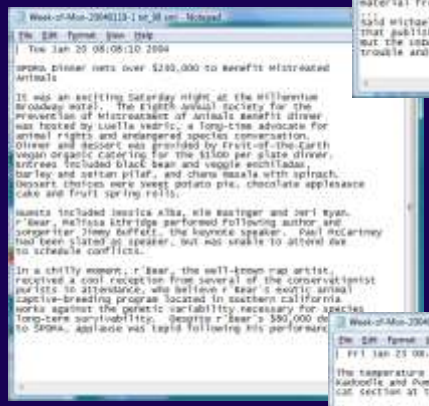
Information Interfaces Research Group
School of Interactive Computing
Georgia Institute of Technology

Georgia Tech College of Computing
School of Interactive Computing

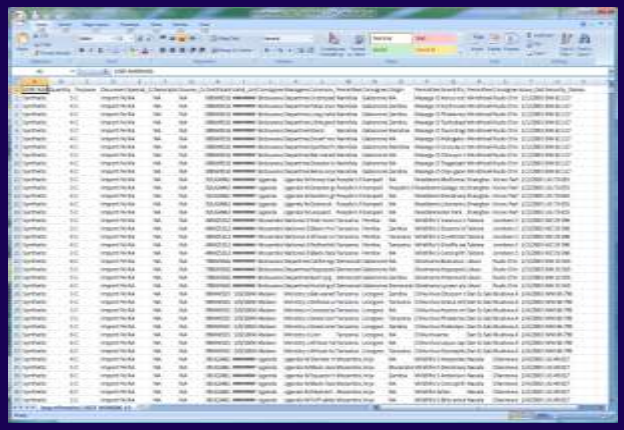
Jigsaw

- Visualization for Investigative Analysis and Sense-making across Document Collections

"Putting the pieces together"



Documents/
case reports



DBs

Blogs

Example Documents

2) Report Date: 22 February, 2003. Surveillance report on Cesar Arze, whose residence is 77 Avenue Francis, Santo Domingo, Dominican Republic. Arze, who moved from Havana, Cuba to Santo Domingo in 1992, works as a medical technician in Santo Domingo. Arze is under surveillance because of information that he is associated with Cuban intelligence services. Arze was photographed in company with a man identified as Hector Lopez in Bogota, Columbia on 23 January, 2003. Lopez, a known representative of FARC, has conducted narcotics distribution activities throughout South and Central America and the Caribbean.

1) Report Date: 12 July, 2004. A routine customs inspection was performed on a package that was sent by a person named Pieter Dopple, 22 Hoveniersstraat, Antwerp, Belgium. This package was addressed to A. Hijazi, 1212 Lyons Ave., Newark, NJ. The customs form stated that this package contained two decorative clocks having a commercial value of $250.00. The package instead contained 111 polished diamonds, whose value is estimated to be $47,000. Discussions with the FBI resulted in a decision not to pursue a customs violation charge against Hljazi. The reason is that Hljazi is currently under FBI surveillance. This package was resealed and delivered to Hljazi by USPS.

1) Report Date: 11 November, 2003. On 1 November, 2003 Mousa Salah, of no fixed address in Herndon, VA., was arrested at Dulles International Airport as he attempted to board a flight to London Heathrow. At the time of his arrest he was using a Jordanian passport in the name Shadi abu Hoshar. Salah has been wanted in connection with Hamas fund-raising and other activities in Northern Virginia. It is believed that Salah entered the USA illegally from Windsor, Canada into Detroit, MI in 2002. Salah was in possession of a Virginia driver's license # T21-23-8820 registered in the name Mousa Salah. Virginia DOT records show that this same license number was issued on 10 August, 2001 to a woman named April Stevens of Roanoke, VA. Ms Stevens is not the subject of further investigations; she teaches kindergarten classes in Roanoke. The address shown on Salah's driver's license does not exist. Several numbers and initials were written on the back of an envelope in possession of Salah at the time of his arrest; they are: (i) J. T., Detroit, (ii) A. H., Newark, (iii) M. M., Laurel.

Our Focus

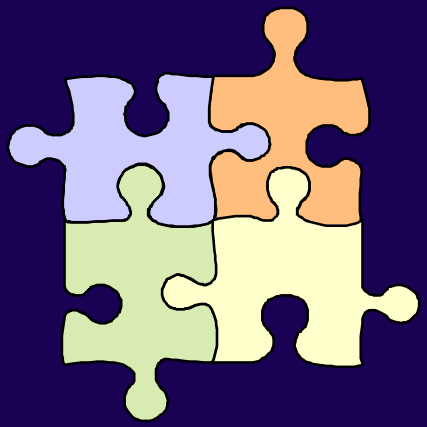
- Entities within the documents
  - Person, place, organization, phone number, date, license plate, etc.
- Thesis: A bigger story or insight from the documents will involve a set of interconnected entities working in coordination
- (We use entity identification software from others)

Connections

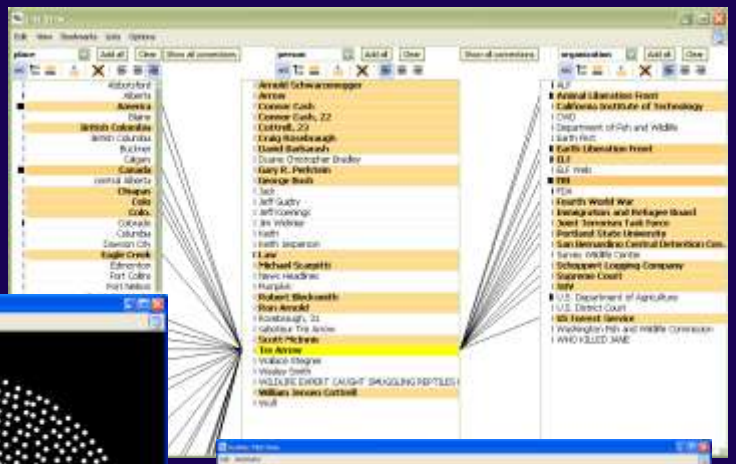
- Entities relate/connect to each other to make a larger "story"
- Connection definition:
  - Two entities are connected if they appear in a document together
  - The more documents they appear in together, the stronger the connection

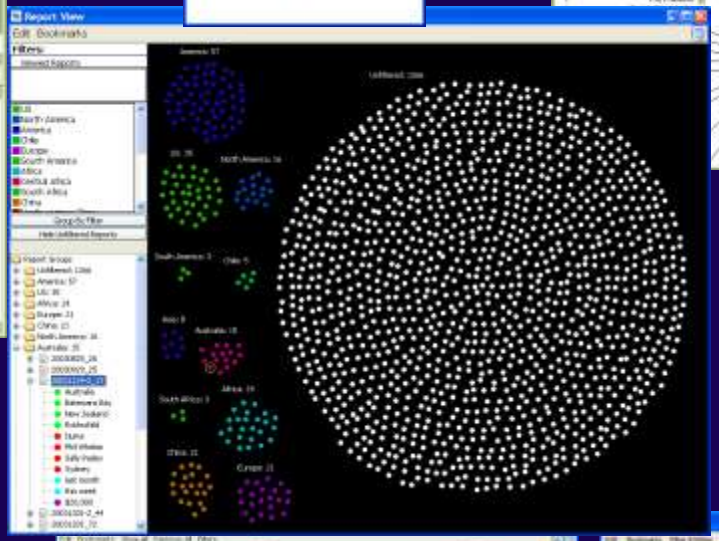
Jigsaw

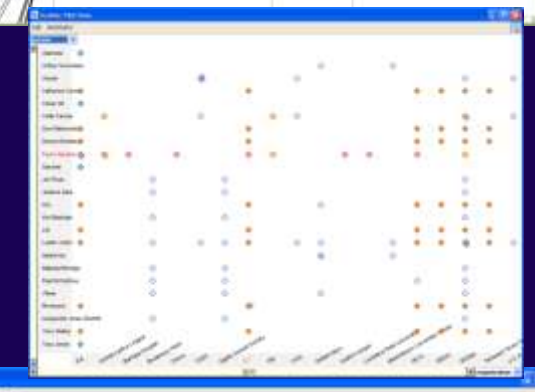
- Multiple visualizations (views) of documents, entities, & their connections
- Views are highly interactive and coordinated
- User actions generate events that are transmitted to and (possibly) reflected in other views

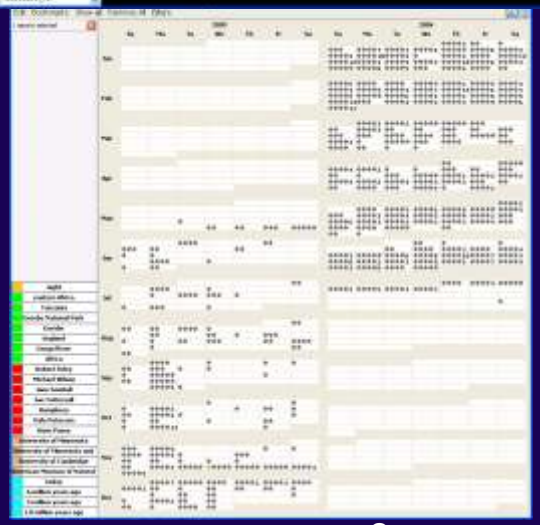


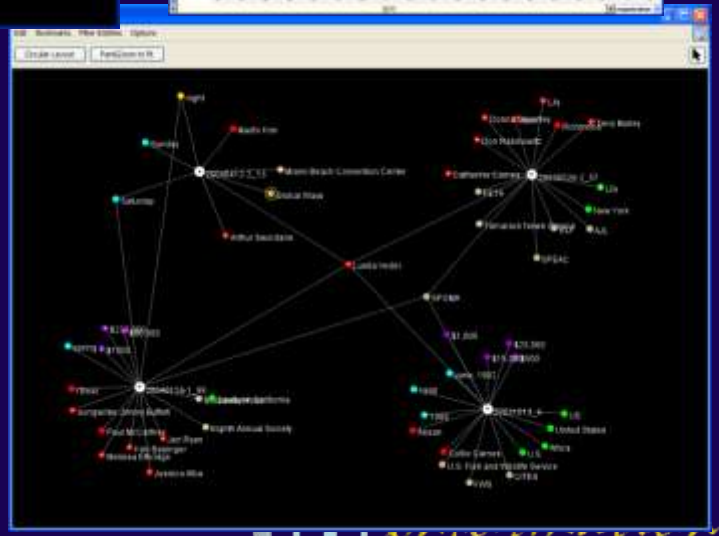
System Views









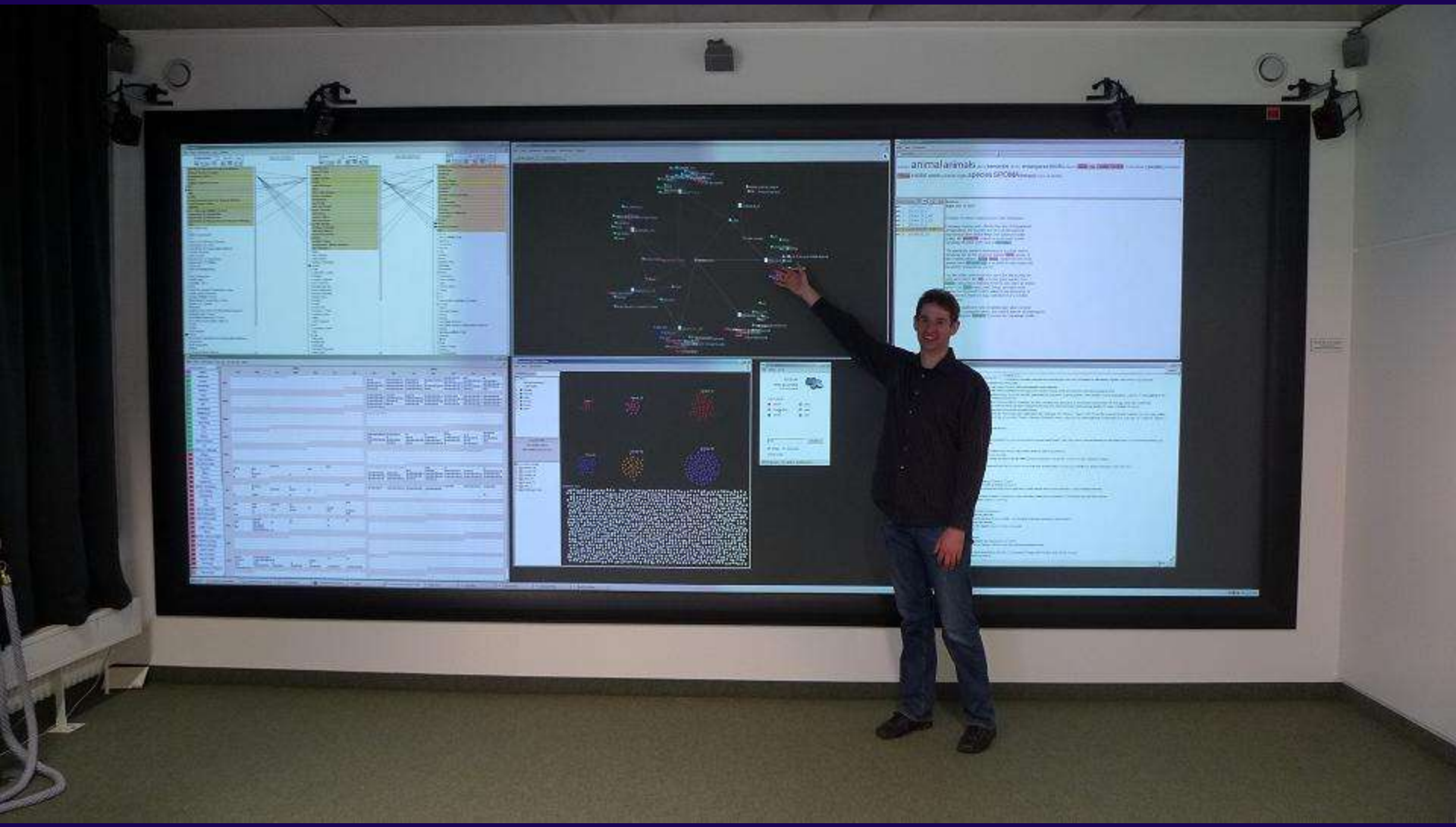


The Need for Pixels

Even More!

information interfaces

Even More!



Demo Data

- NSF awards from 2000-now (abstract & key meta-data)
- Big thanks to Remco Chang & UNCC folks

- Demo focus: CSE awards (12,243)

information interfaces

IIS_2006-2010.xml - WordPad

File Edit View Insert Format Help

```
<award>
<awardnumber>0915788</awardnumber>
<title>III: Small: Supporting Investigative Analysts and Researchers in Sense-making across Large Document
<nsforganization>IIS </nsforganization>
<programs>GRAPHICS &amp;amp; VISUALIZATION</programs>
<startdate>August 1, 2009</startdate>
<lastamendmentdate>July 13,  2009</lastamendmentdate>
<principalinvestigator>Stasko, John</principalinvestigator>
<state>GA</state>
<organization>GA Tech Research Corporation - GA Institute of Technology       </organization>
<awardinstrument>Continuing grant                                  </awardinstrument>
<programmanager>Ephraim P. Glinert                          </programmanager>
<expirationdate>July 31,  2010</expirationdate>
<awardedamounttodate>218691</awardedamounttodate>
<co_pinames></co_pinames>
<piemailaddress>stasko@cc.gatech.edu                                                    </piemaila
<organizationstreetaddress>Office of Sponsored Programs    </organizationstreetaddress>
<organizationcity>Atlanta                    </organizationcity>
<organizationstate>GA</organizationstate>
<organizationzip>30332</organizationzip>
<organizationphone>4048944819</organizationphone>
<nsfdirectorate>CSE </nsfdirectorate>
<programelementcodes>7453</programelementcodes>
<programreferencecodes>HPCC|9215|7923|7453</programreferencecodes>
<fieldofapplications>0116000    Human Subjects                              |</fieldofapplications>
<awardnumber>0915788</awardnumber>
<abstract>People routinely encounter and seek to make sense of large collections of data that include both
</award>
```

nsf-iis.jig - WordPad

File Edit View Insert Format Help

```
<report>
  <reportId>0915788</reportId>
  <reportDate>August 1, 2009</reportDate>
  <referencedReport></referencedReport>
  <reportSource></reportSource>
  <reportDescription>
 III: Small: Supporting Investigative Analysts and Researchers in Sense-making across Large Document Col

 People routinely encounter and seek to make sense of large collections of data that include both unstru
  </reportDescription>
  <program>GRAPHICS &amp;amp; VISUALIZATION</program>
  <division>IIS</division>
  <directorate>CSE</directorate>
  <pi>Stasko, John</pi>
  <organization>GA Tech Research Corporation - GA Institute of Technology</organization>
  <state>GA</state>
  <progmgr>Ephraim P. Glinert</progmgr>
  <award-instr>Continuing grant</award-instr>
  <amount>218691</amount>
  <programelementcode>7453</programelementcode>
  <programreferencecode>HPCC</programreferencecode>
  <programreferencecode>9215</programreferencecode>
  <programreferencecode>7923</programreferencecode>
  <programreferencecode>7453</programreferencecode>
  <fieldofapplication>0116000 Human Subjects</fieldofapplication>
</report>
```

For Help, press F1

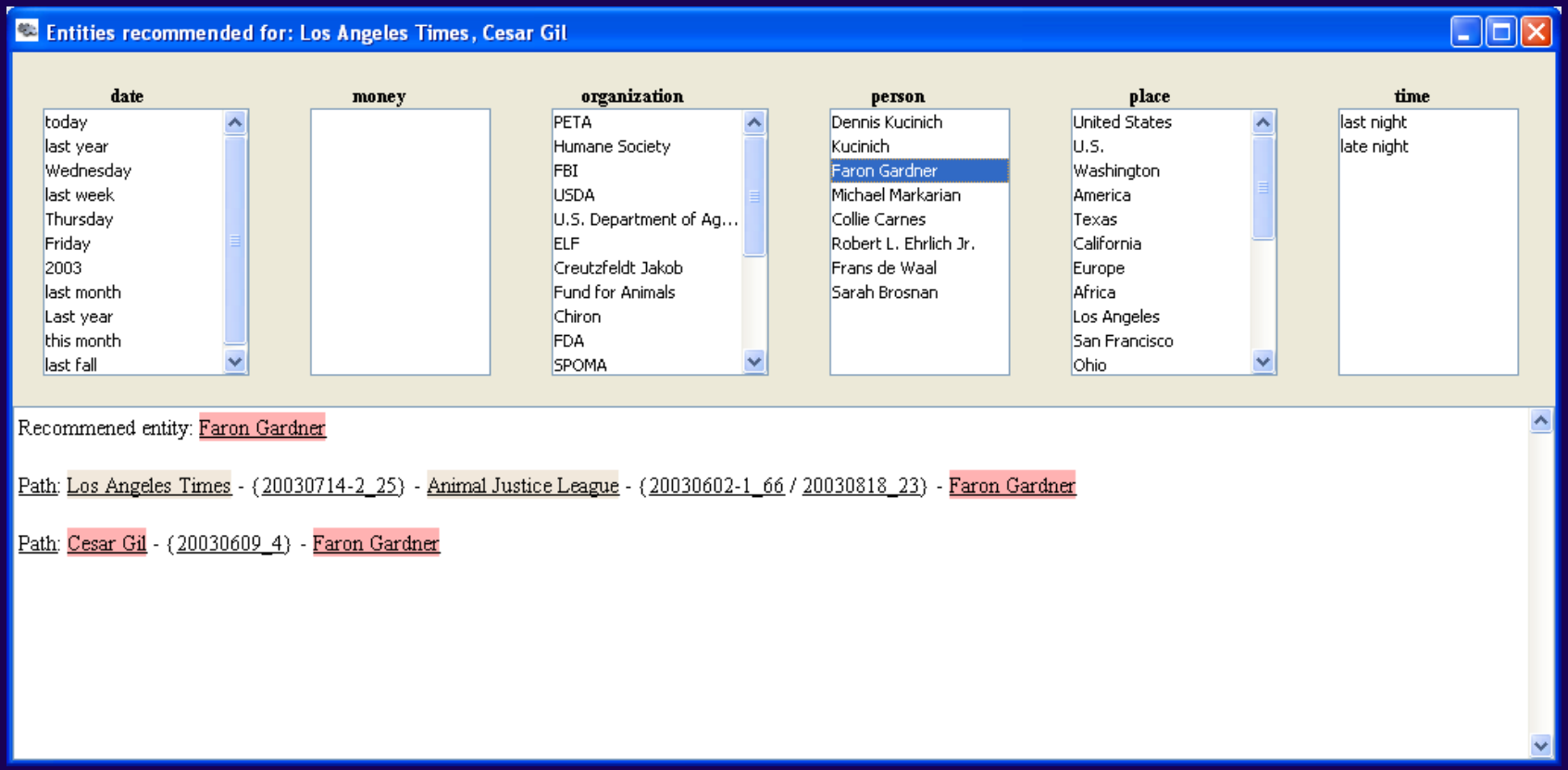
Demo

information interfaces

What We're Working On

- Jigsaw is heavy on the user-directed visual exploration, but light on the automated computational analysis

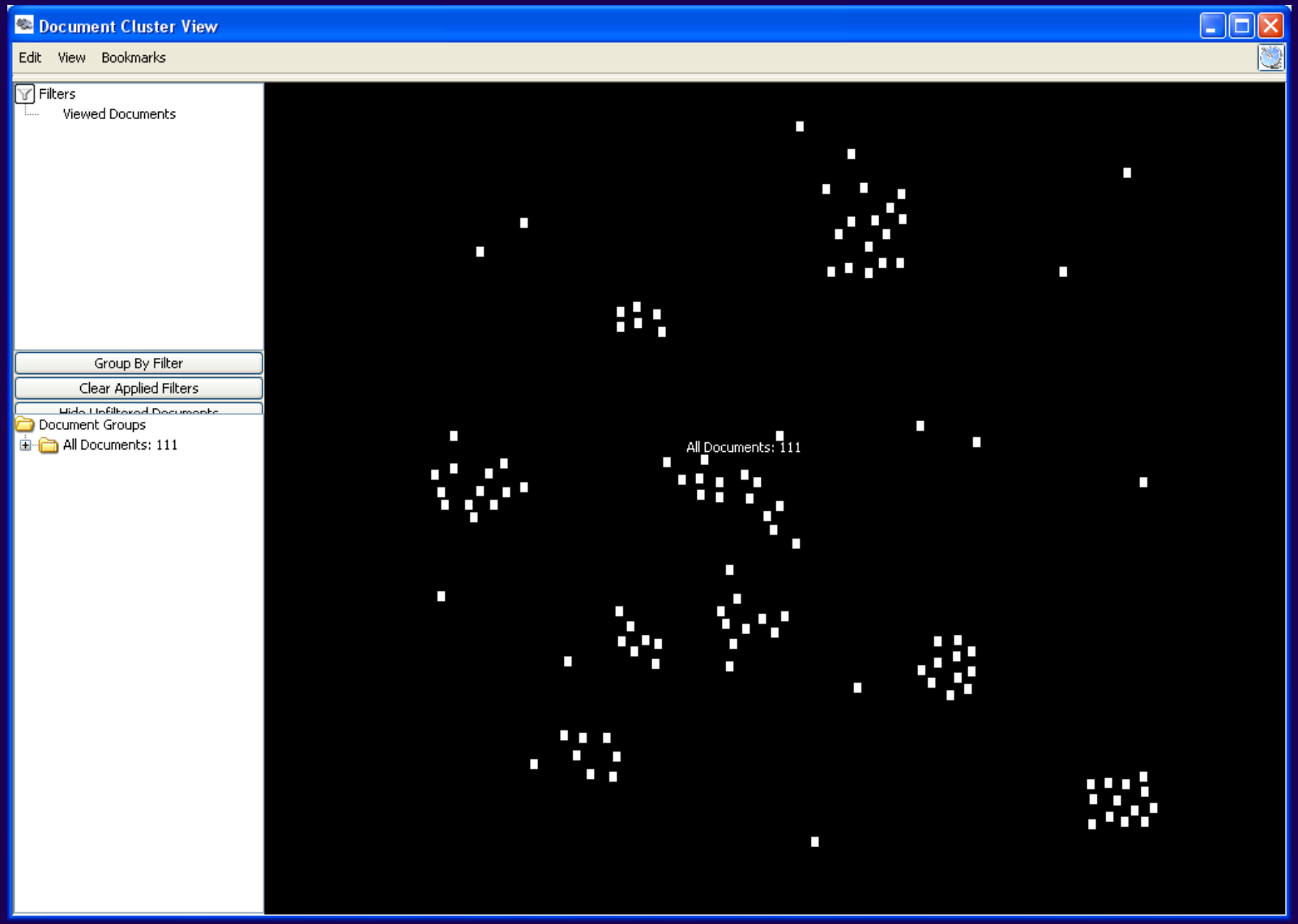
Recommend Related Entities



Adding Computational Analysis

- Document themes & clustering

information interfaces


Document Cluster View
Edit
View
Bookmarks
Filters
Viewed Documents
Group By Filter
Clear Applied Filters
Document Groups
All Documents: 111
All Documents: 111

Adding Computational Analysis

- Document themes & clustering
- Document similarity

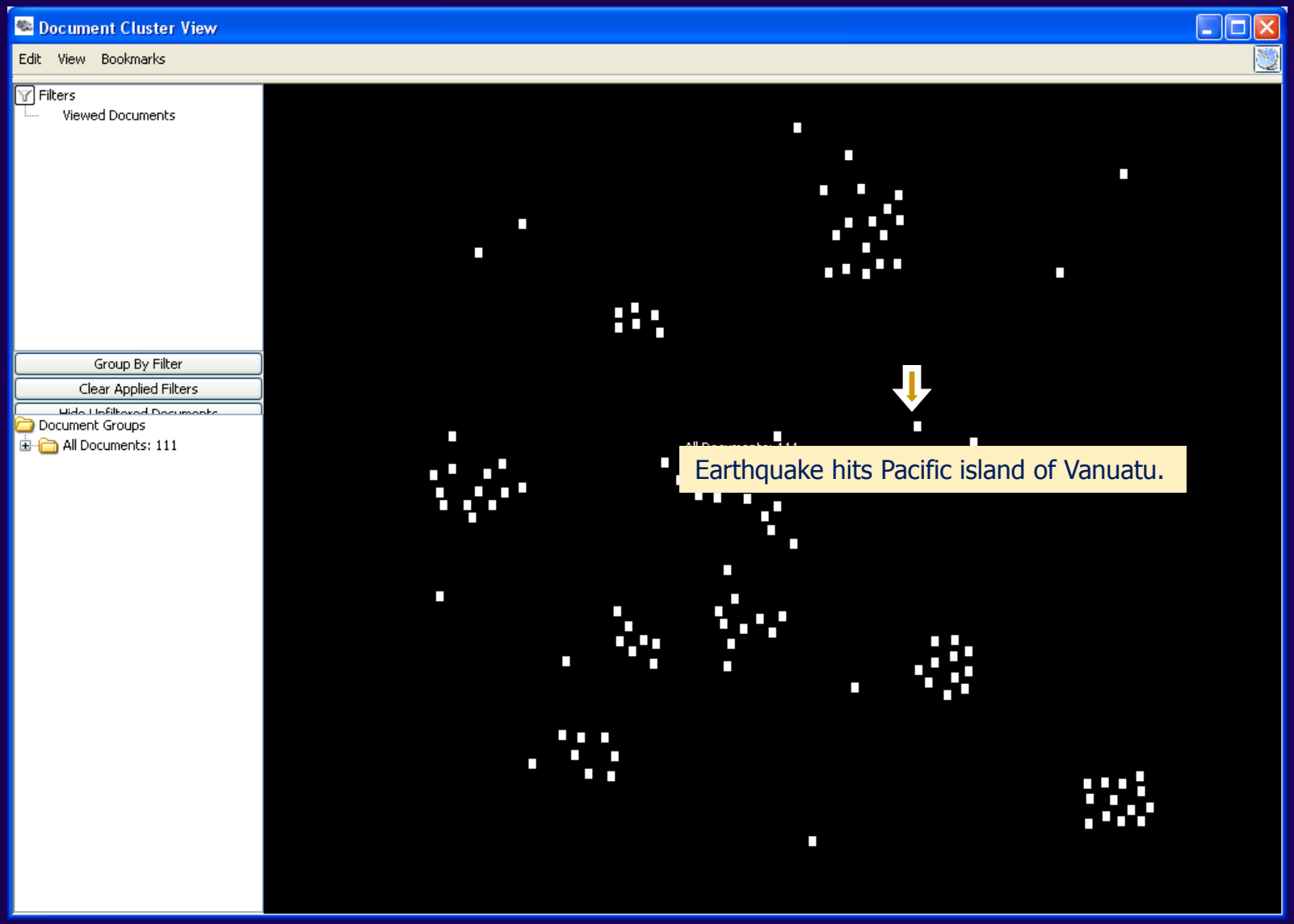
information interfaces

Most similar

Adding Computational Analysis

- Document themes & clustering
- Document similarity
- Document summarization

information interfaces


Document Cluster View
Edit
View
Bookmarks
Filters
Viewed Documents
Group By Filter
Clear Applied Filters
Document Groups
All Documents: 111
Earthquake hits Pacific island of Vanuatu.

Adding Computational Analysis

- Document themes & clustering
- Document similarity
- Document summarization
- Sentiment analysis


Document Grid View
Edit View Bookmarks
Sort by Document Size
Color by Document Size
Add All
Most negative
Most positive

Lessons Learned

- A quick recap of some things we've learned from working with partners from
  - Homeland security
  - Law enforcement
  - Intelligence analysis
  - Investigative reporting
  - Business intelligence

and from our experience developing systems like Jigsaw

Intentionally provocative

Focus on the user's task & goals

information interfaces

But in many instances, you don’t know the questions

As the data scales up, we still want interactive visualization

Online algorithms are better than offline algorithms

- Real-time is great

For real systems, pragmatics matter

- "Common" programming languages
- Others must be able to access your code
- Documentation, tutorials, …
- "Good" is usually good enough, especially at first
  (great would be wonderful)

You don’t get training data

information interfaces

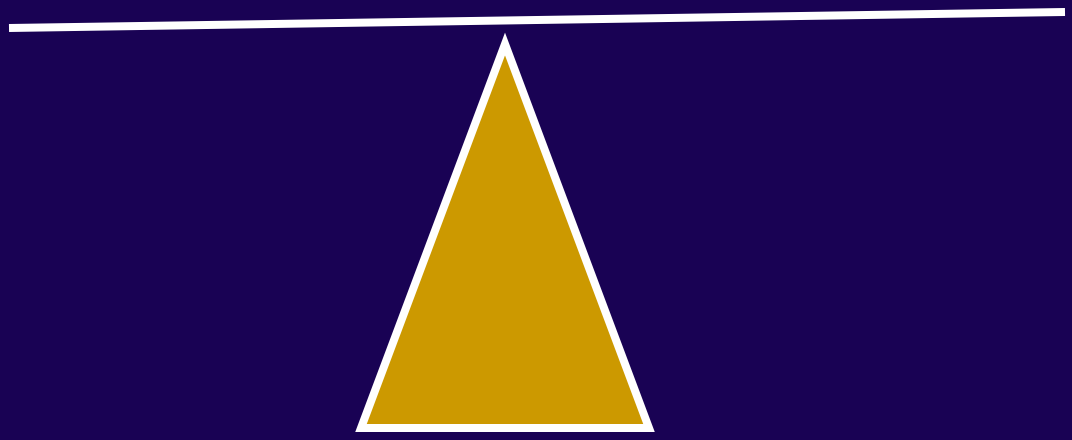
Real data doesn't necessarily have nice clusters

Evaluation doesn't necessarily have a good quantitative measure

information interfaces

Fundamental algorithmic research

Application-driven problems



Acknowledgments

- Work conducted as part of the Southeastern Regional Visualization and Analytics Center, supported by DHS and NVAC and the new 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