

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

Jigsaw

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



“Putting the pieces together”



Week-of-Mon-20040216-3 bt_17.xml - Notepad

File Edit Format View Help

Wed Feb 18 21:24:40 2004

Mad cow disease must now be considered "indigenous to North America," and the United States can no longer consider its first mad cow "an imported case," says an international scientific panel advising the U.S. Department of Agriculture (USDA), at a USDA public hearing in Silverdale, Maryland on Wednesday, the five member panel said that while the dairy cow found to be infected with bovine spongiform encephalopathy (BSE) in Washington state last December was imported from Canada, it probably was not the only one. "It is probable that other infected animals have been imported from Canada and possibly also from Europe," the panel warned. "These animals have not been detected and therefore infective material has likely been rendered, fed to cattle, and amplified within the cattle population, so that cattle in the U.S. have also been indigenously infected." Their report advised that United States should test many more cattle, implement the rapid screening tests for BSE used in Europe, and exclude brain and spinal cord material from all human and animal food, including pet food.

*** Michael Hanson, a scientist with Consumers Union, the organization that publishes Consumer Reports magazine, "There's plenty we don't know, but the USDA takes the science that will cause the beef industry the least trouble and basically ignores the rest." Sleep well.

Week-of-Mon-20040119-1 bt_98.xml - Notepad

File Edit Format View Help

Tue Jan 20 08:08:10 2004

SPOKA Dinner nets over \$230,000 to Benefit Mistreated Animals

It was an exciting Saturday night at the Millennium Broadway Hotel. The Eighth Annual Society for the Prevention of Mistreatment of Animals benefit dinner was hosted by Luelia Vedic, a long-time advocate for animal rights and endangered species conversation. Dinner and dessert was provided by Fruit-of-the-Earth Vegan Organic Catering for the \$1500 per plate dinner. Entrees included black bean and veggie enchiladas, barley and seitan pilaf, and chana masala with spinach. Dessert choices were sweet potato pie, chocolate apple/cake and fruit spring rolls.

Guests included Jessica Alba, Kim Basinger and Jeri Ryan. r Bear, Melissa Ethridge performed following author and songwriter Jimmy Buffett, the keynote speaker. Paul McCartney had been slated as speaker, but was unable to attend due to schedule conflicts.

In a chilly moment, r Bear, the well-known rap artist, received a cool reception from several of the conservationists purists in attendance, who believe r Bear's exotic animal captive-breeding program located in Southern California works against the genetic variability necessary for species long-term survivability. Despite r Bear's \$80,000, due to SPOKA, applause was tepid following his performance.

Week-of-Mon-20040119-4 bt_32.xml - Notepad

File Edit Format View Help

Fri Jan 23 08:52:58 2004

The temperature outside was only a whisker above zero Thursday morning, but Kadoodle and Pumpkin were toasty warm, sunning themselves in the window of the cat section at the Ark, a municipal animal shelter in Forest Park.

The blissful scene belied the cold reality faced by the cats and the rest of the 1,100 animals brought annually to the shelter by police, animal-control officers and citizens in the seven municipalities and one township it serves. On March 21, the Ark will lock its doors, a victim of insufficient financing and of its location on a piece of prime real estate on Forest Park's commercial and entertainment strip.

A vacant auto body shop in Forest Park is a possible new home for the Ark, but the shelter doesn't have \$600,000 to buy and renovate the building, said Elliott Serrano, shelter manager.

Meanwhile, staff and volunteers are scrambling to find temporary quarters for 200 or so animals, said Ellen Milad, coordinator of volunteers. Chances are good that all will get at least temporary placement in other shelters or in private "foster homes," but the long-term prospects are dim for the area's unwanted animals, she said.

"The number of stray dogs and cats always seems to increase in March," she said. "What will happen to the animals that would have been brought to the Ark if we were still here?"

DBs

Documents/
case reports

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 Hijazi. The reason is that Hijazi is currently under FBI surveillance. This package was resealed and delivered to Hijazi 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)



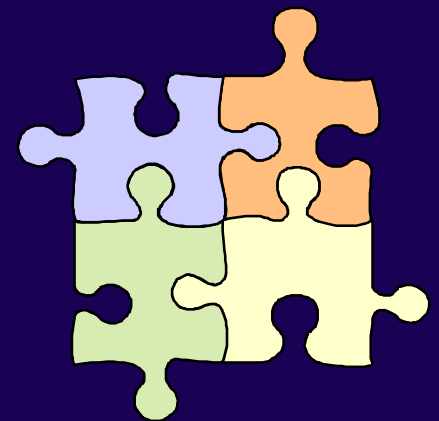
Connections

- 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

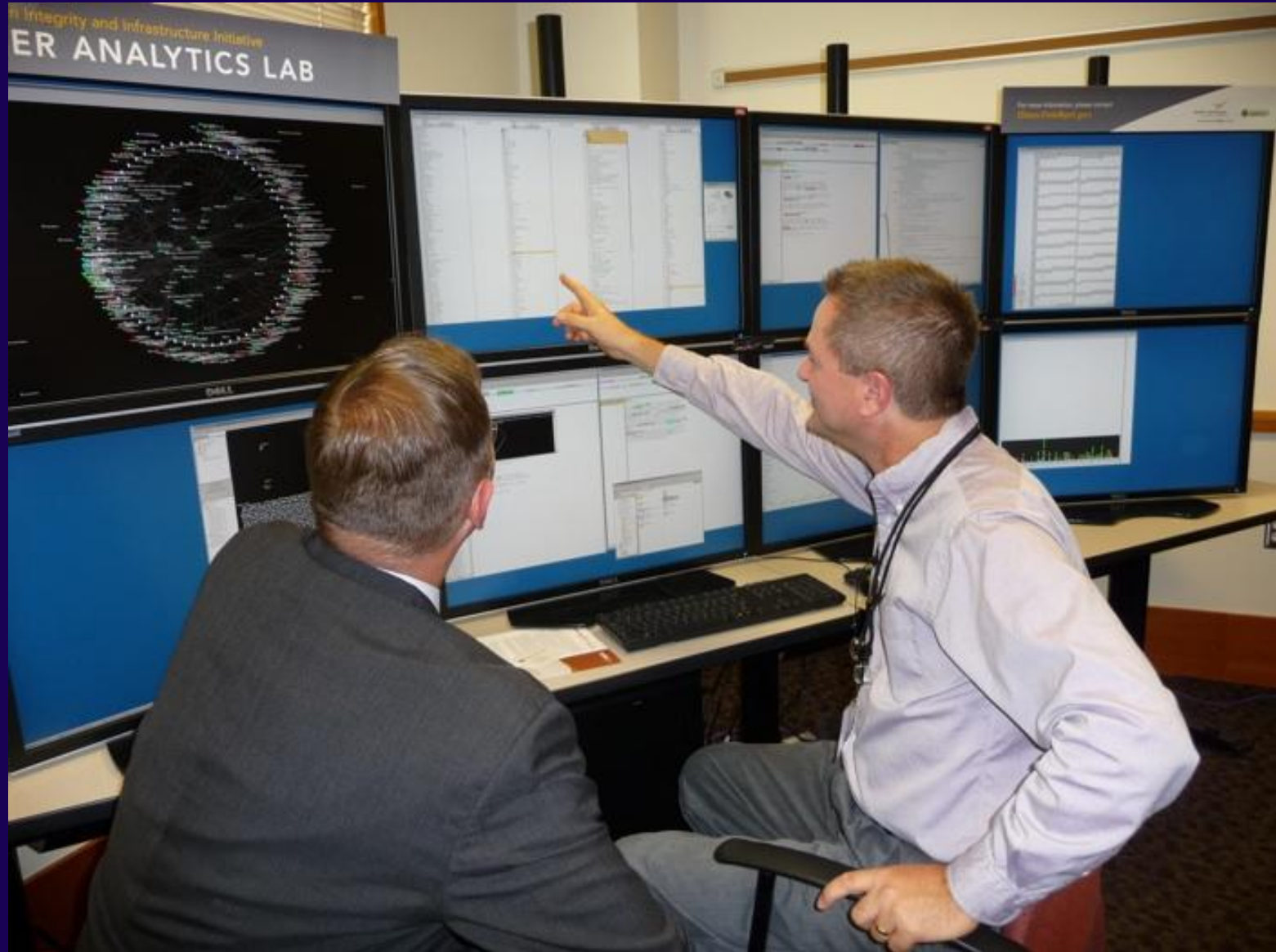
The image displays a collection of software windows from a system named "Jigsaw" by Blue Iguanodon. The windows include:

- Document View:** Shows search results for terms like "arrow", "arson", "authorities", "being", "canada", "canadian", "charges", "construction", "elf", "enforcement", "environmental", "fbi", "federal", "government", "group", "logging". It includes a list of documents with dates and a summary of an article about "Arrow 30" and a key issue regarding the "Immigration and Refugee Board".
- Report View:** Displays a "Report Groups" list with categories like "US", "North America", "Africa", etc., and a "Viewed Reports" section.
- Circular Graph View:** A circular network graph showing relationships between various entities and dates.
- Word Tree View:** A hierarchical tree structure showing the frequency of words in a document, with "dog" being the root.
- Scatter Plot View:** A large scatter plot showing data points distributed across a coordinate system, with labels for regions like "America: 57", "North America: 16", etc.
- Report Calendar:** A calendar view showing reports scheduled for various days of the month.
- List View:** A list of names and organizations, including "Arnold Schwarzenegger", "Dennis Quaid", "Cottrell, Z3", "Craig Rosebraugh", "David Barbash", "Duane Christopher Bradley", "Gary R. Perlstein", "George Bush", "Jack", "Jeff Gaudy", "Jeff Koenig", "Jim Widmer", "Kath", "Keith Jaspersen", "Law", "Michael Scarpitti", "News Headlines", "Robert Blacksmith", "Ron Arnold", "Scott McInnis", "Tobit Arrow", "Wilbur Stogier", "Wesley Smith", "WILL-LIFE EXPERT 'CAUGHT' SMUGGLING REPTILES", and "William Jensen Cottrell".
- Scatter Plot View (Circular Layout):** A circular network graph showing relationships between various entities and dates.

The Need for Pixels



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)



```
IIS_2006-2010.xml - WordPad
File Edit View Insert Format Help
[Icons]

<award>
<awardnumber>0915788</awardnumber>
<title>III: Small: Supporting Investigative Analysts and Researchers in Sense-making across Large Document
<nsfororganization>IIS </nsfororganization>
<programs>GRAPHICS & 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 </piemailaddress>
<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 & 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



What We're Working On

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



Recommend Related Entities

Entities recommended for: Los Angeles Times, Cesar Gil

date	money	organization	person	place	time
today last year Wednesday last week Thursday Friday 2003 last month Last year this month last fall		PETA Humane Society FBI USDA U.S. Department of Ag... ELF Creutzfeldt Jakob Fund for Animals Chiron FDA SPOMA	Dennis Kucinich Kucinich Faron Gardner Michael Markarian Collie Carnes Robert L. Ehrlich Jr. Frans de Waal Sarah Brosnan	United States U.S. Washington America Texas California Europe Africa Los Angeles San Francisco Ohio	last night late night

Recommended entity: **Faron Gardner**

Path: **Los Angeles Times** - (20030714-2_25) - **Animal Justice League** - (20030602-1_66 / 20030818_23) - **Faron Gardner**

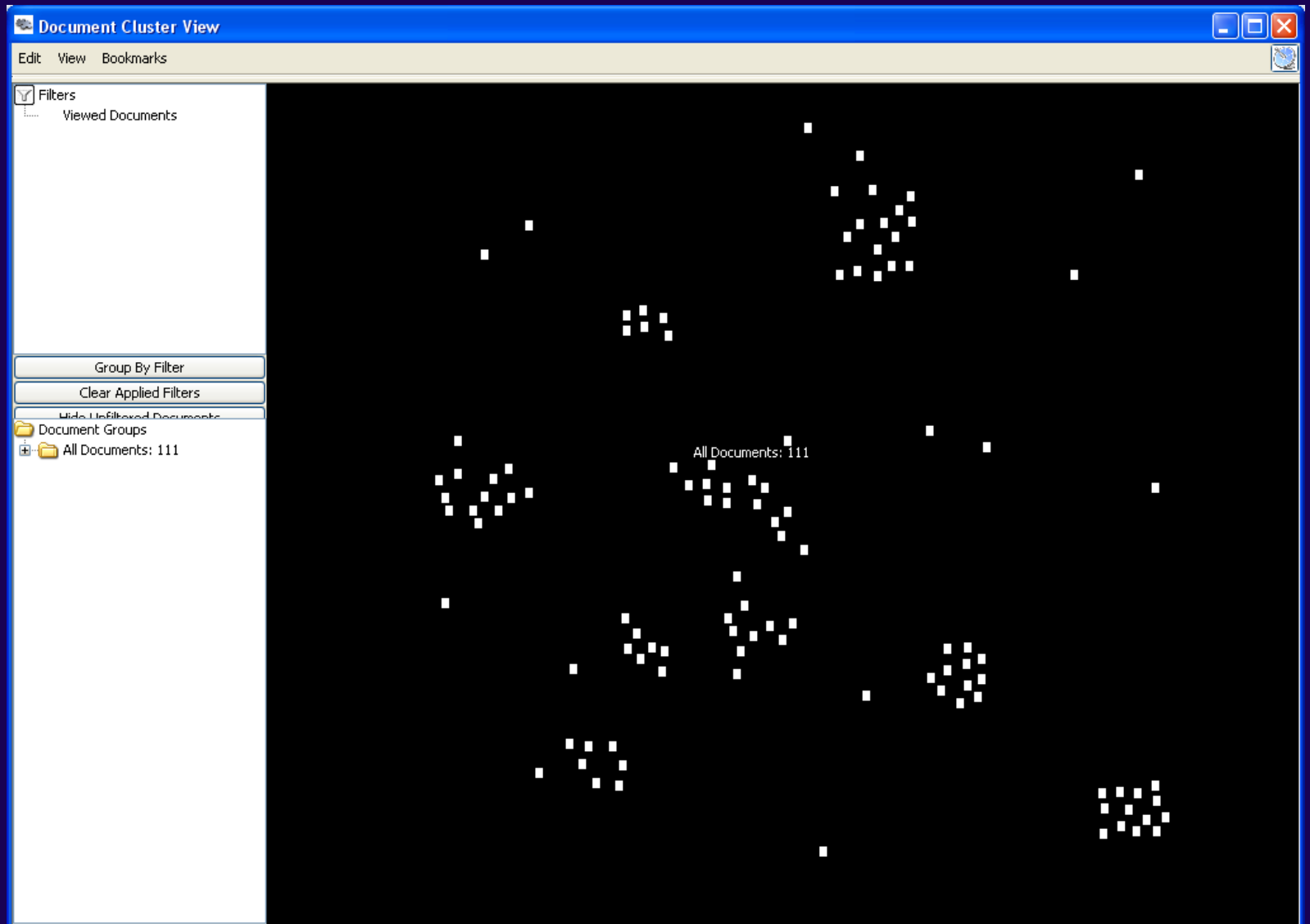
Path: **Cesar Gil** - (20030609_4) - **Faron Gardner**



Adding Computational Analysis

- Document themes & clustering





Adding Computational Analysis

- Document themes & clustering
- Document similarity





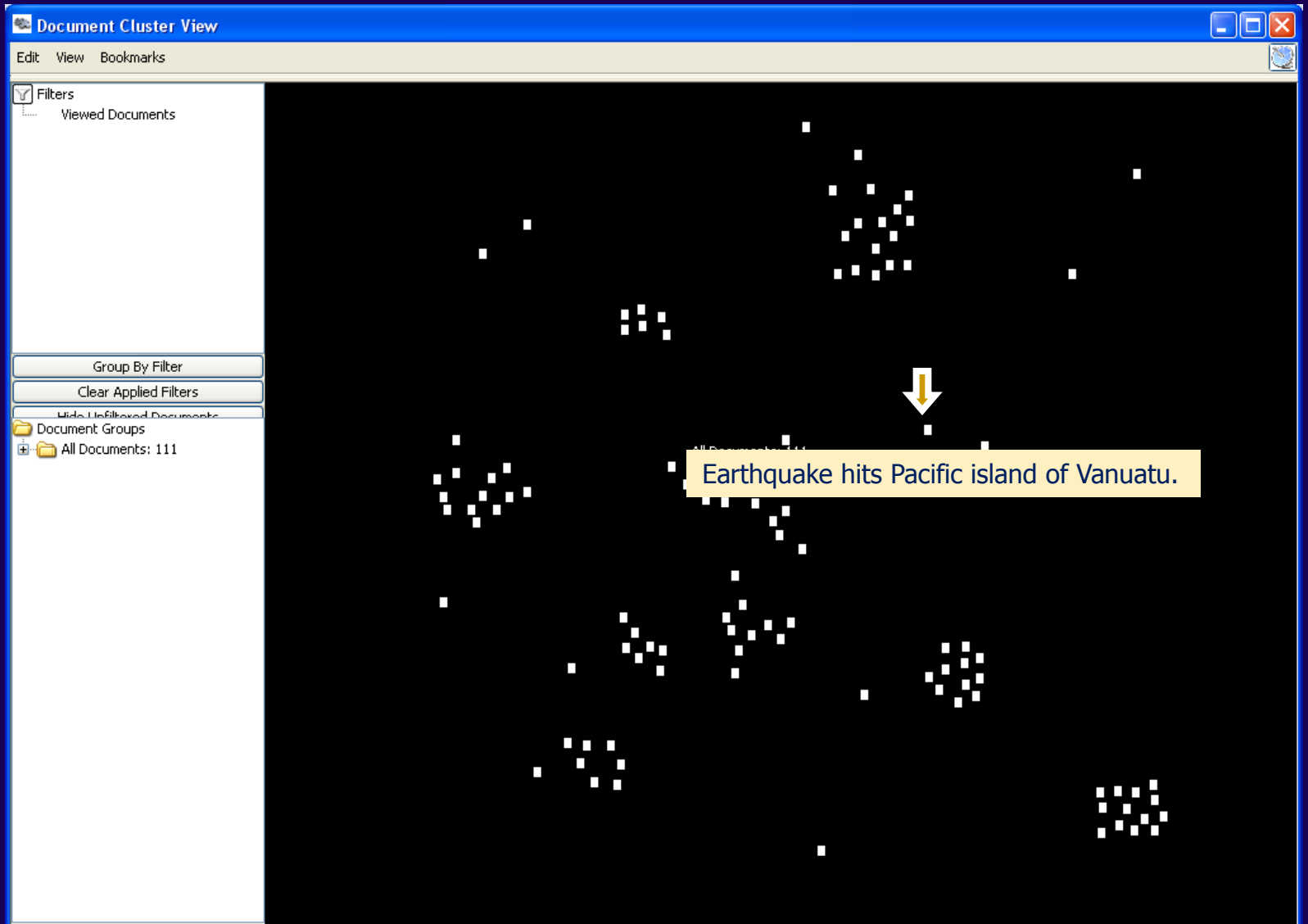
Most similar



Adding Computational Analysis

- Document themes & clustering
- Document similarity
- Document summarization



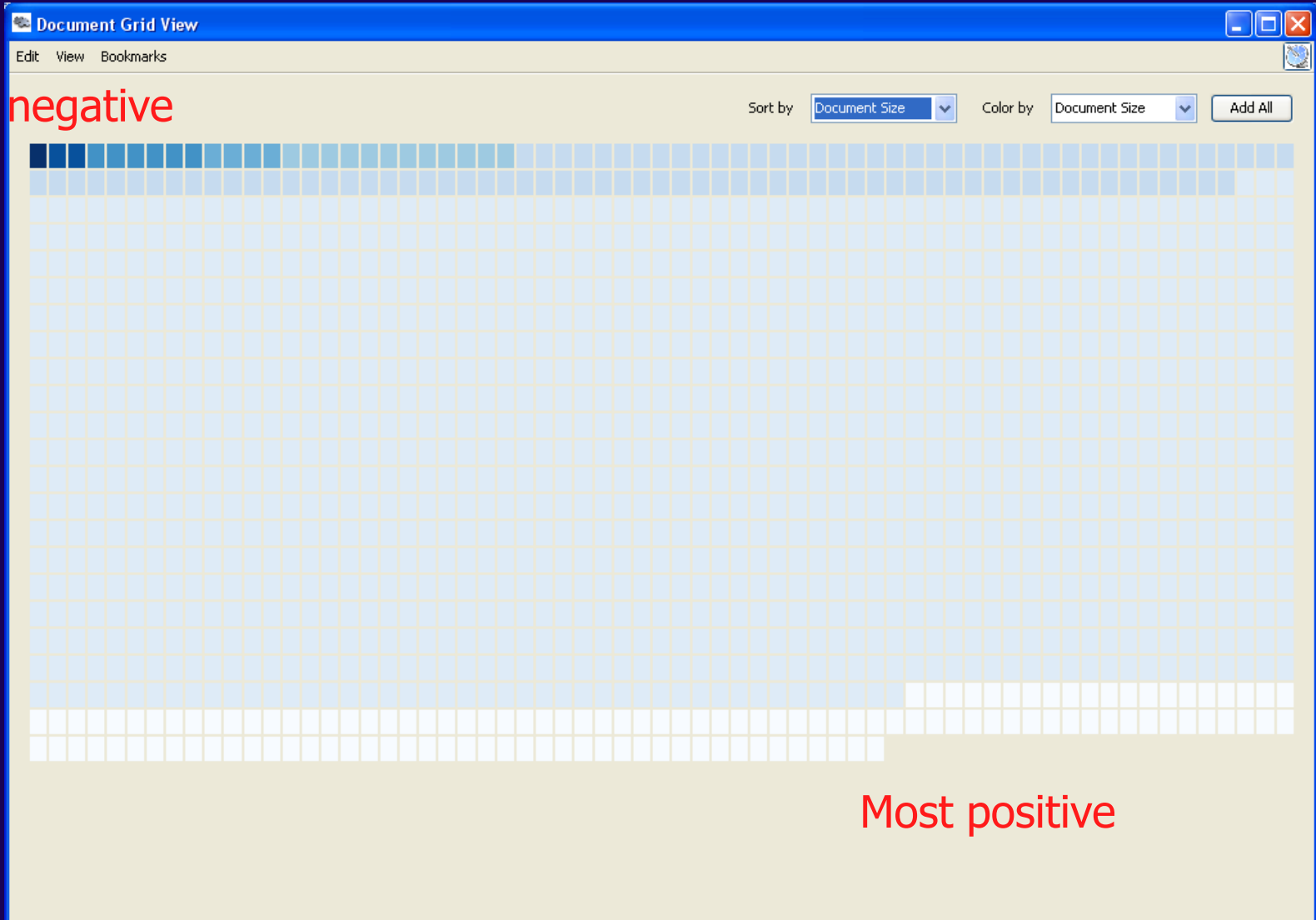


Adding Computational Analysis

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



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

**Intentionally
provocative**

and from our experience developing systems like Jigsaw



Focus on the user's task & goals



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



Real data doesn't necessarily have nice clusters

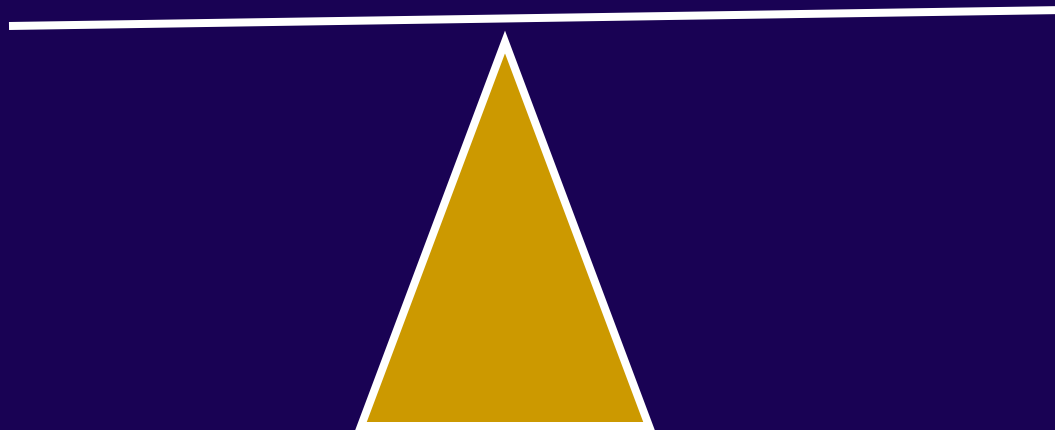


Evaluation doesn't necessarily have a good quantitative measure



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

