PNNL-SA-70791

Intelligent Visual Interfaces for Text Analysis: Selected Observations and Challenges Ahead

Pak Chung Wong
Pacific Northwest National Laboratory
Richland, WA 99352, USA
http://www.pnl.gov/wong
pak.wong@pnl.gov

Pacific Northwest
NATIONAL LABORATORY

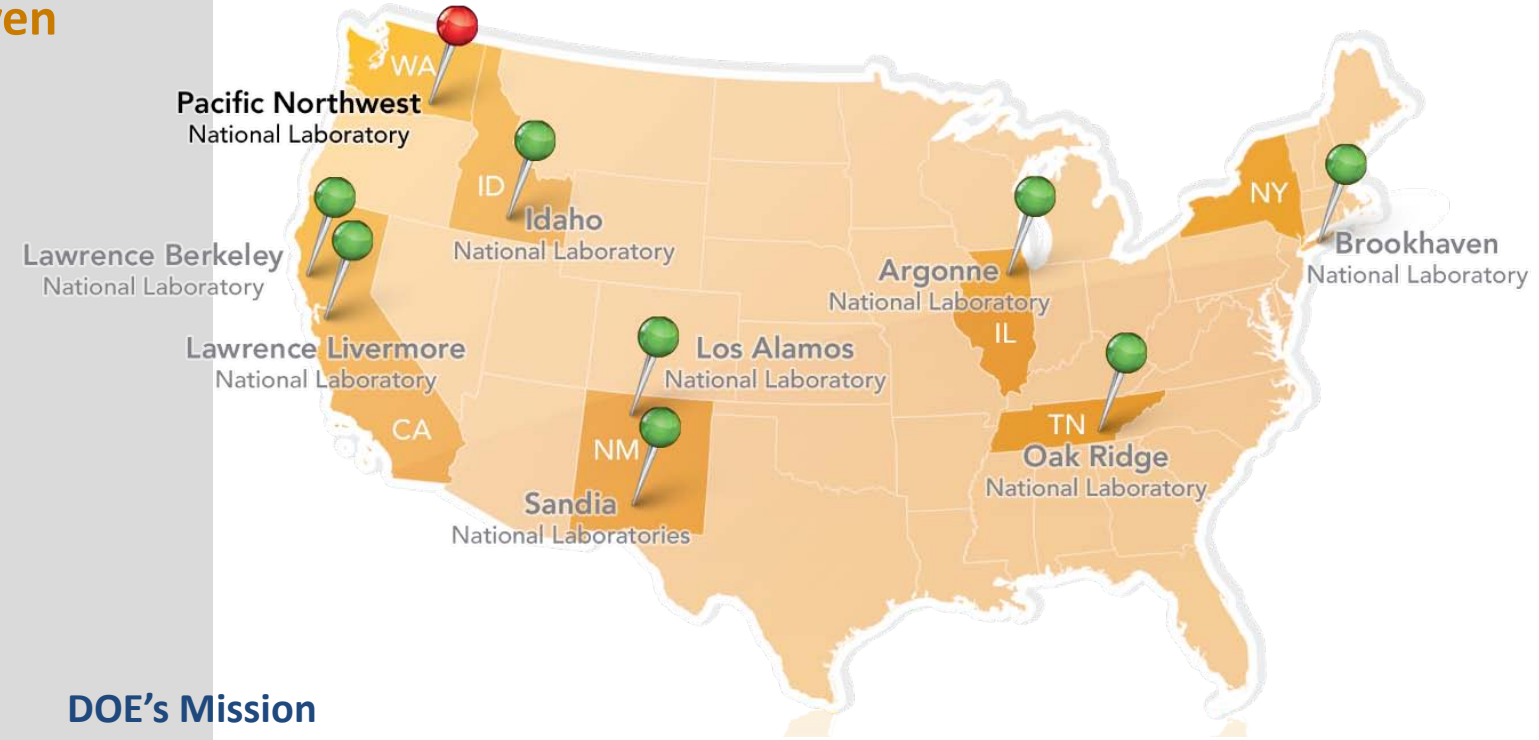
Proudly Operated by Battelle *Since 1965*

Outline

- Text analytics research and development (R&D) at the Pacific Northwest National Laboratory
- Selected observations and lessons learned (5)
- Selected R&D trends and challenges ahead (3)
- Panel discussion

The National Laboratory System

Washington is one of only seven states to host a Department of Energy (DOE) multi-program national laboratory



DOE's Mission

"Discovering the solutions to power and securing America's future."

With overarching missions:

"To advance the national, economic, and energy security of the United States; to promote scientific and technological innovation in support of that mission; and to ensure the environmental cleanup of the national nuclear weapons complex."

Office of Science
U.S. DEPARTMENT OF ENERGY

Pacific Northwest
NATIONAL LABORATORY

Text Analytics at PNNL

- Almost two decades of text analytics R&D at PNNL
- Categorization of large structured and unstructured document corpora in the 1990s
- Predictive analytics in the 2000s
- Different design requirements, different missions, different mindsets, but human participation has been part of the solutions
- IN-SPIRE is the flagship product
  - Still undergoing intensive R&D (with obstacles still to overcome)
  - http://in-spire.pnl.gov/

IN-SPIRE Visualization

- Continuously co-developed with information analysts
- Window, icon, menu, pointing device (WIMP)-based human-computer interface
- Icon, scatterplot, graph, chart, and browser-based visualization designs
- Primary role of visualization is to assist human beings, not to replace them
  - Ingest, navigation, reasoning, collaboration, and report

IN-SPIRE Text Engine

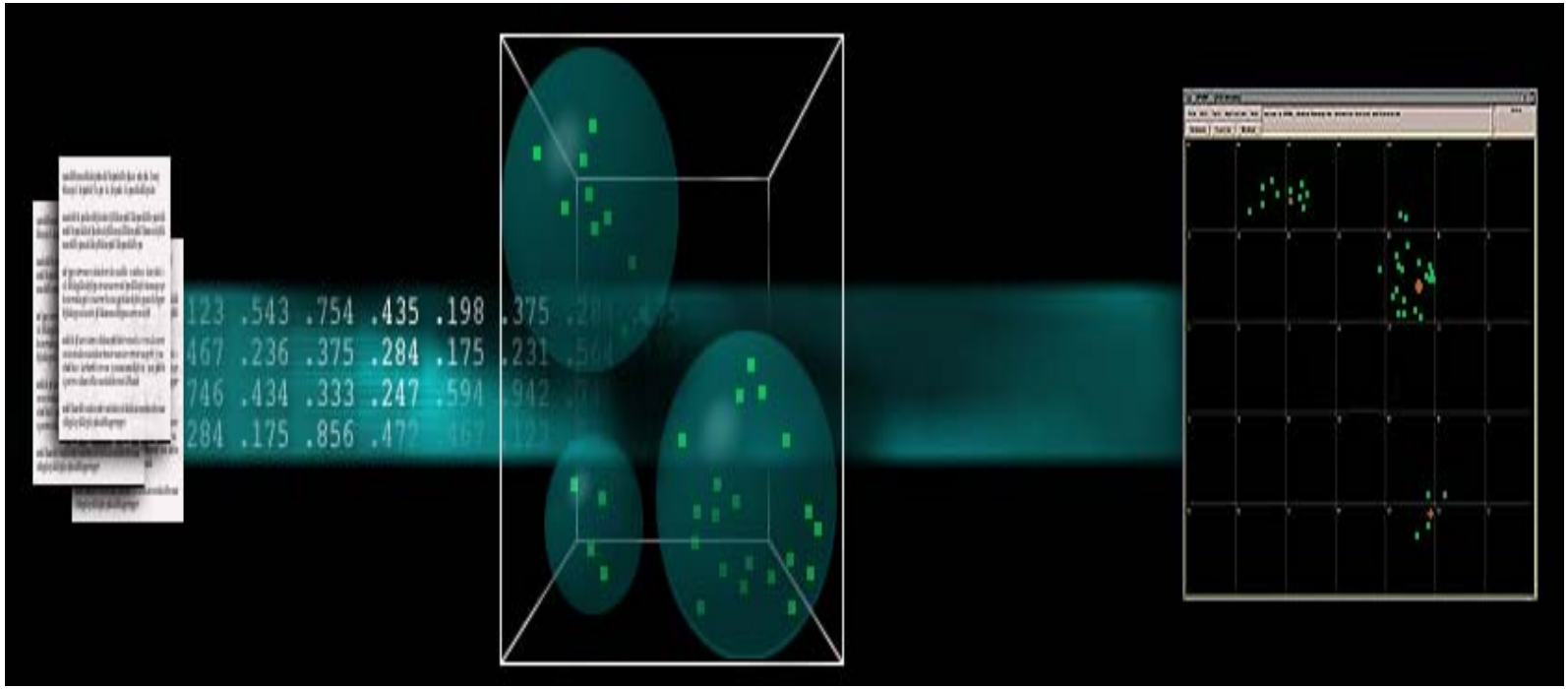
- Continuously co-developed with information analysts
- Salton's bag-of-words model, no built-in natural language processing (NLP) capability
- Additional efforts to develop multi-language capabilities
- Multiple architectures for different privacy and security needs and operation modes

IN-SPIRE Core Concept

Create data signature

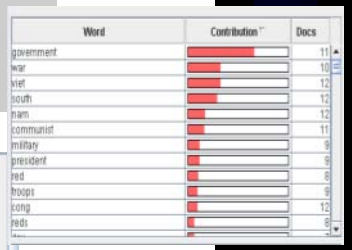
Synthesize into high-dimensional discovery space

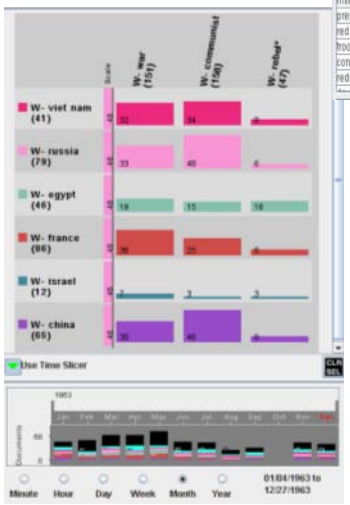
Visual discourse for discovery

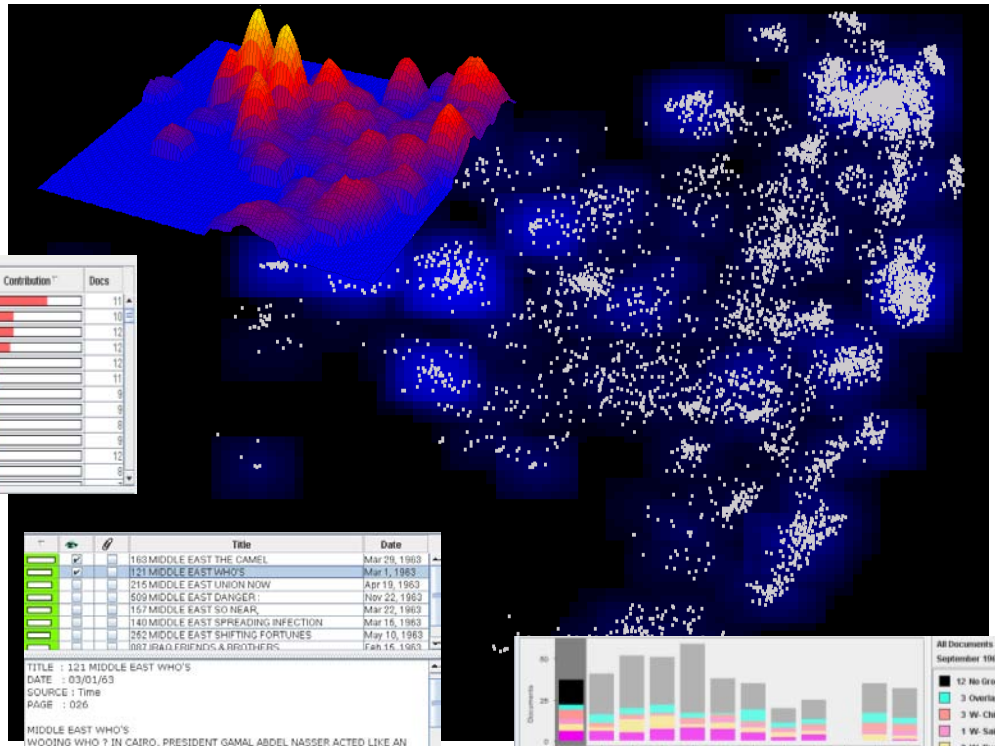


http://in-spire.pnl.gov

Eight Primary IN-SPIRE Visualization Tools

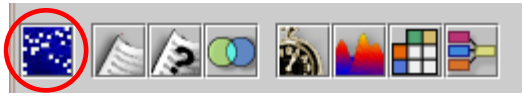


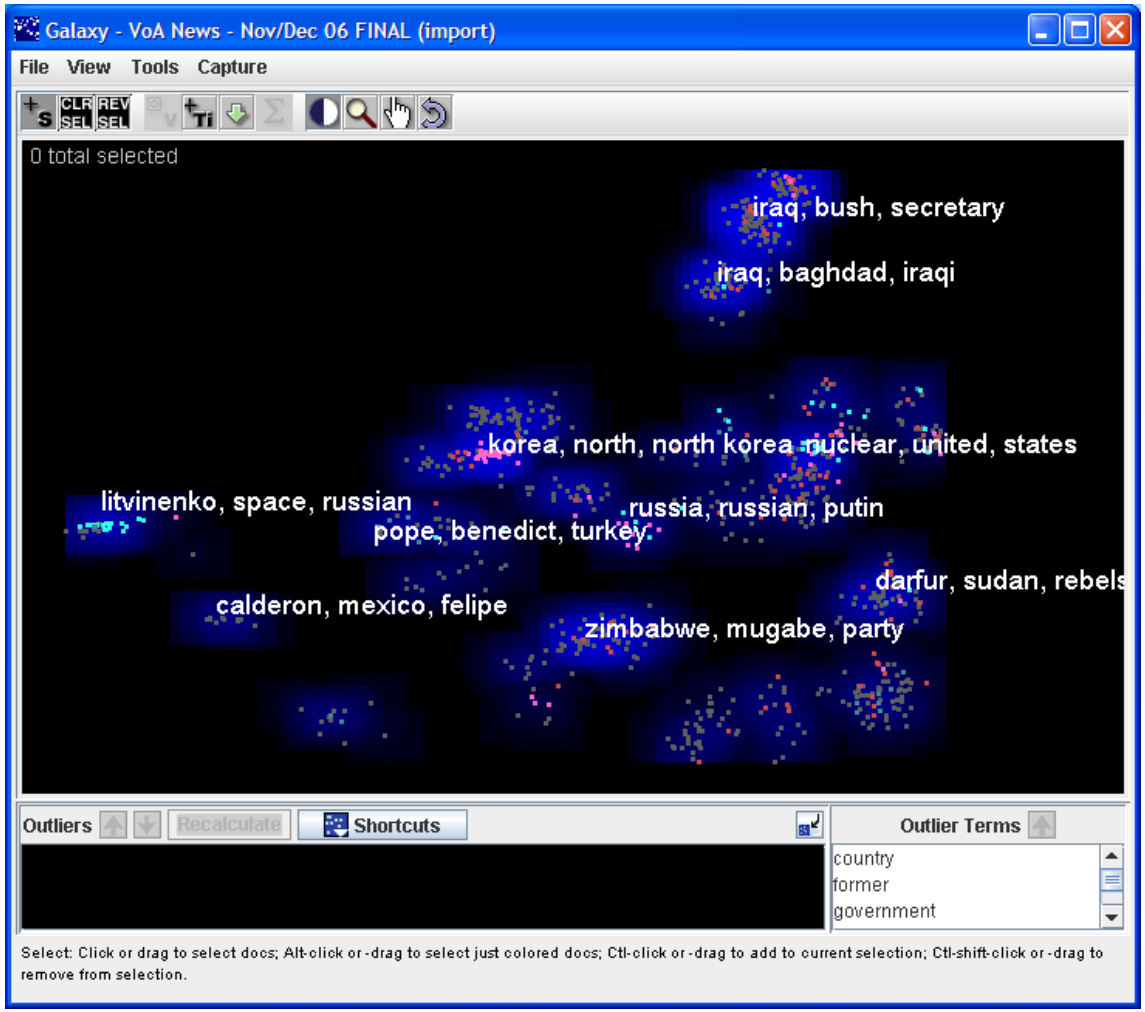




Pacific Northwest
NATIONAL LABORATORY

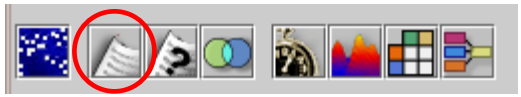
1. Galaxy





Pacific Northwest
NATIONAL LABORATORY

2. Document/Evidence Viewer



Document Viewer: 5 selected, 8 more from groups

File Edit View

		Title	Date	D	Chechen ...	Russian...	Yukos Oil Do...	!
✔	✔	Britain: Radioactive Material Found...	Nov 24, 2006					
	✔	Deaths of Russians Critical of Gove...	Nov 27, 2006					
	✔	Murdered Former Russian Spy Buri...	Dec 7, 2006					
	✔	New Report Finds Journalists Facin...	Dec 11, 2006					
	✔	Of Politics and Poison	Dec 8, 2006	★				
		Poisoned Russian Spy Litvinenko ...	Nov 23, 2006					
		Former Russian Spy Dies of Appar...	Nov 24, 2006					
	✔	British Minister Blasts Russia's Pres...	Nov 26, 2006					

TITLE: Britain: Radioactive Material Found in Former Russian Spy's Body
TEXT:

Britain: Radioactive Material Found in Former Russian Spy's Body
By VOA News
24 November 2006

[imagealt Alexander Litvinenko, former KGB spy and author of the book 'Blowing Up Russia: Terror From Within']Alexander Litvinenko, former KGB spy and author of the book 'Blowing Up Russia: Terror From Within'

The British government says former Russian spy Alexander Litvinenko was poisoned by radiation. In London Friday, the country's Health Protection Agency said a large quantity of the radioactive element polonium 210 were found in Litvinenko's urine. They said it is not yet clear how the material entered his body. Litvinenko died in a London hospital Thursday after a three-week fight for his life. He fell ill at the beginning of November after meeting two Russians at a hotel in London. In a statement read following his death, the former KGB agent accused Russian President Vladimir Putin of being responsible for his death. He called Mr. Putin "barbaric and ruthless," with "no respect for life, liberty or any civilized value." In Helsinki Friday for an EU summit, Mr. Putin described Litvinenko's death as a tragedy, but he said there is no reason for speculation that he had any role in the incident. The former Russian agent had been investigating the recent killing in Moscow of Russian journalist Anna Politkovskaya, a frequent critic of the Kremlin. Litvinenko defected to Britain six years ago after accusing Russia's Federal Security Service, the successor to the Soviet KGB, of ordering the killing of

1 Russia could seek retribution

2 Litvinenko has written about attacks against Chechens

Highlights (0)

Previous

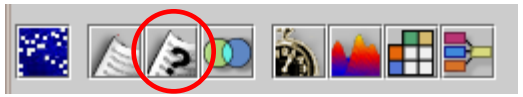
Next

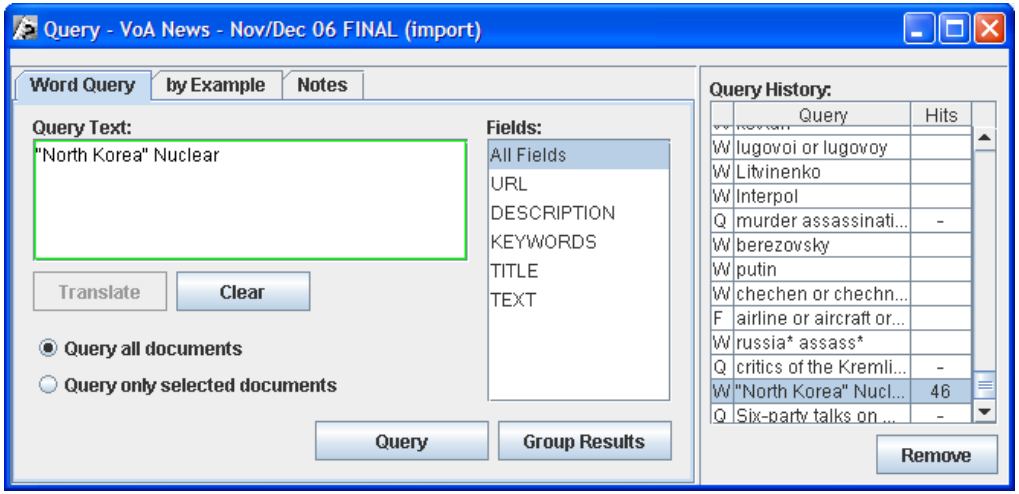
Type Comment

Marking

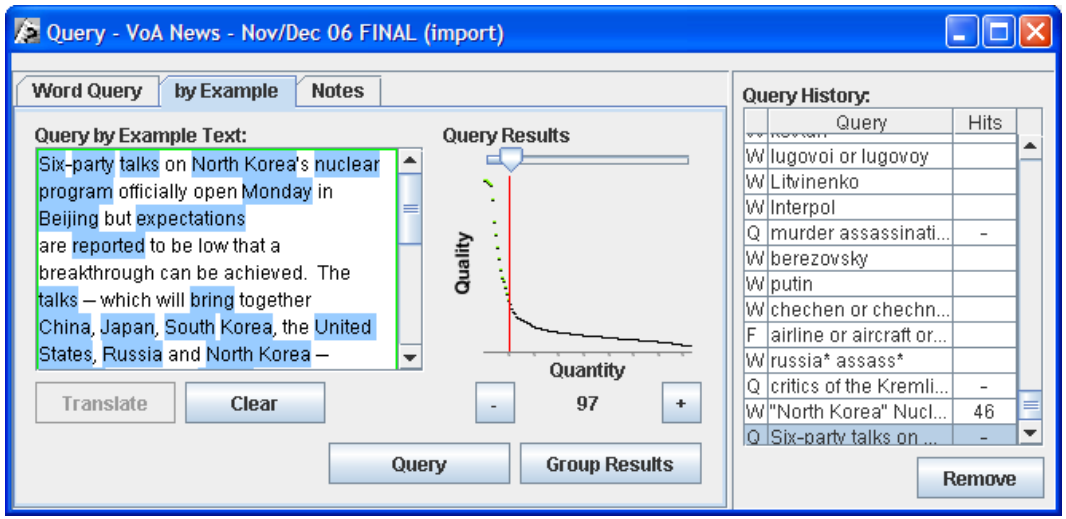
Pacific Northwest NATIONAL LABORATORY

3. Query





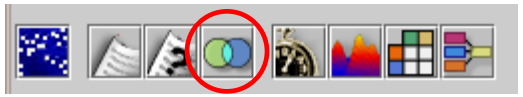
Boolean

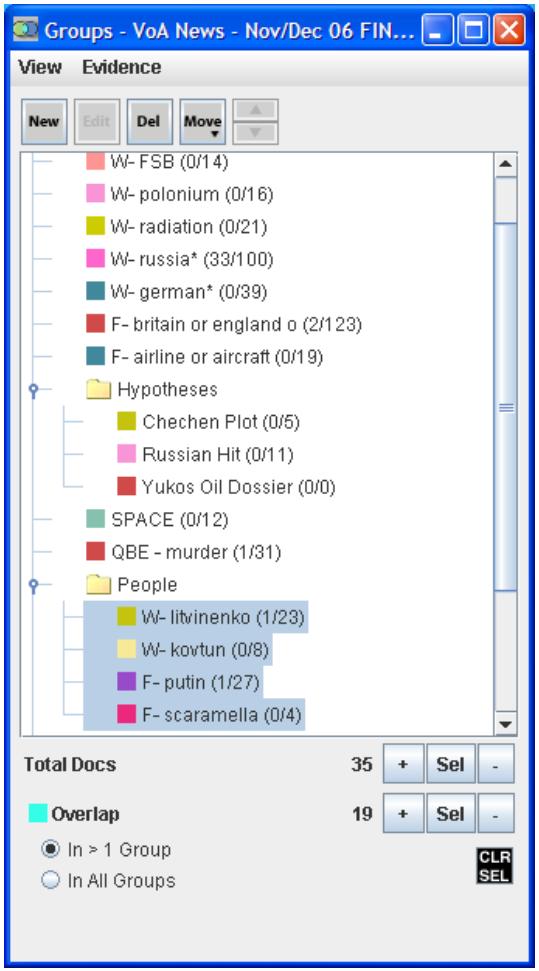


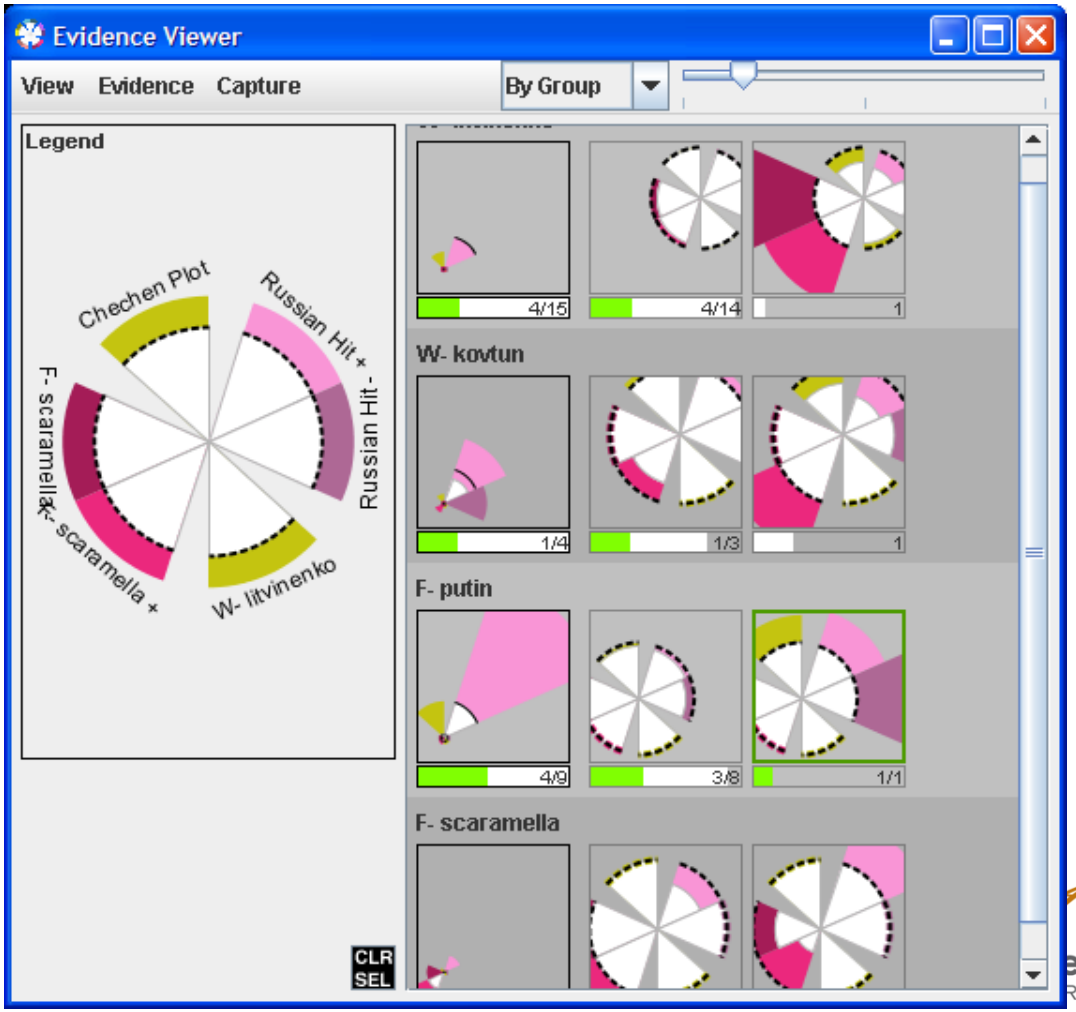
By Example

Pacific Northwest
NATIONAL LABORATORY

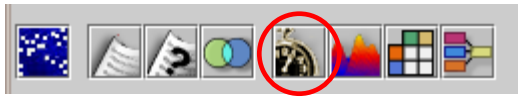
4. Groups and Evidence

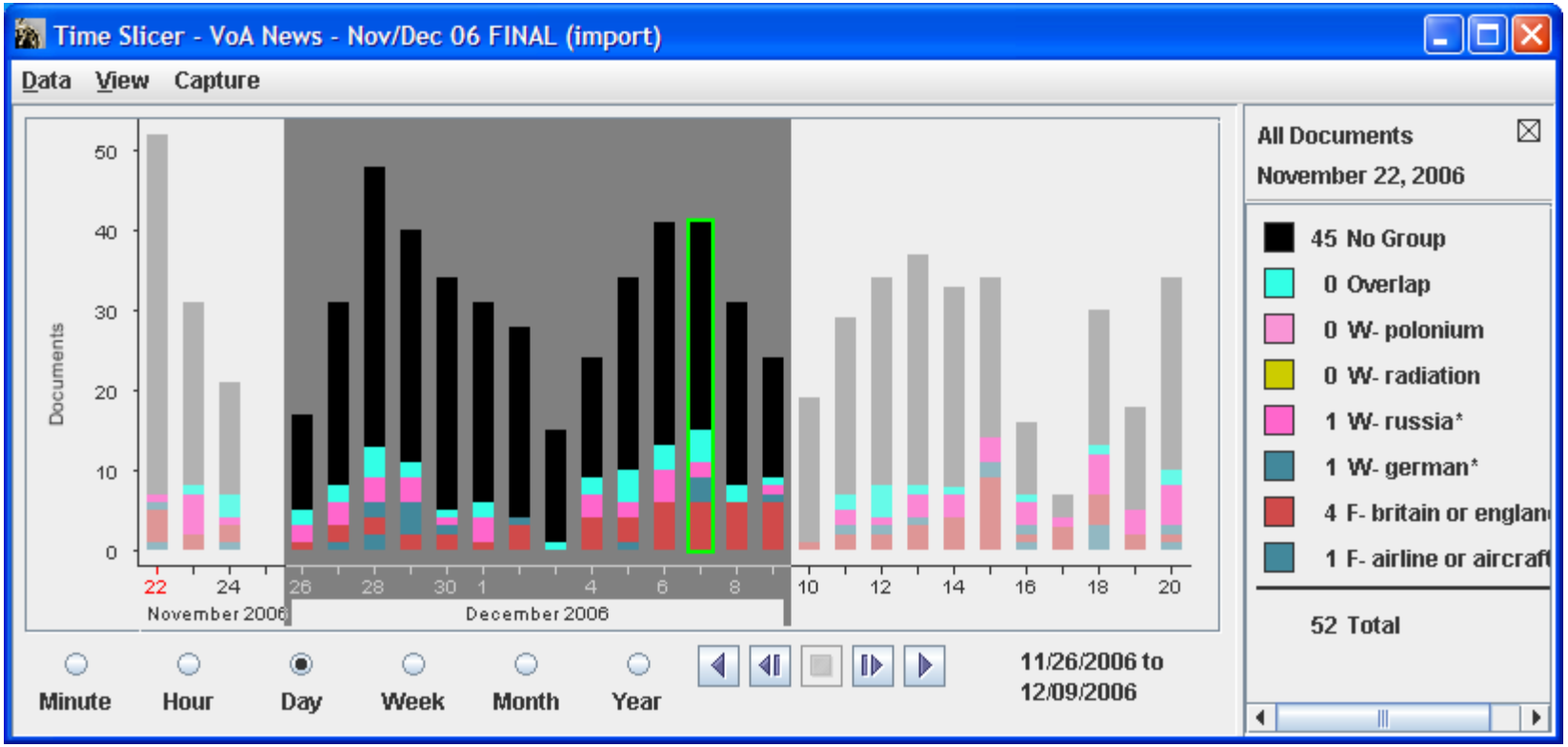






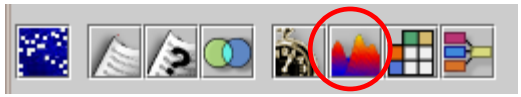
5. Time Slicer

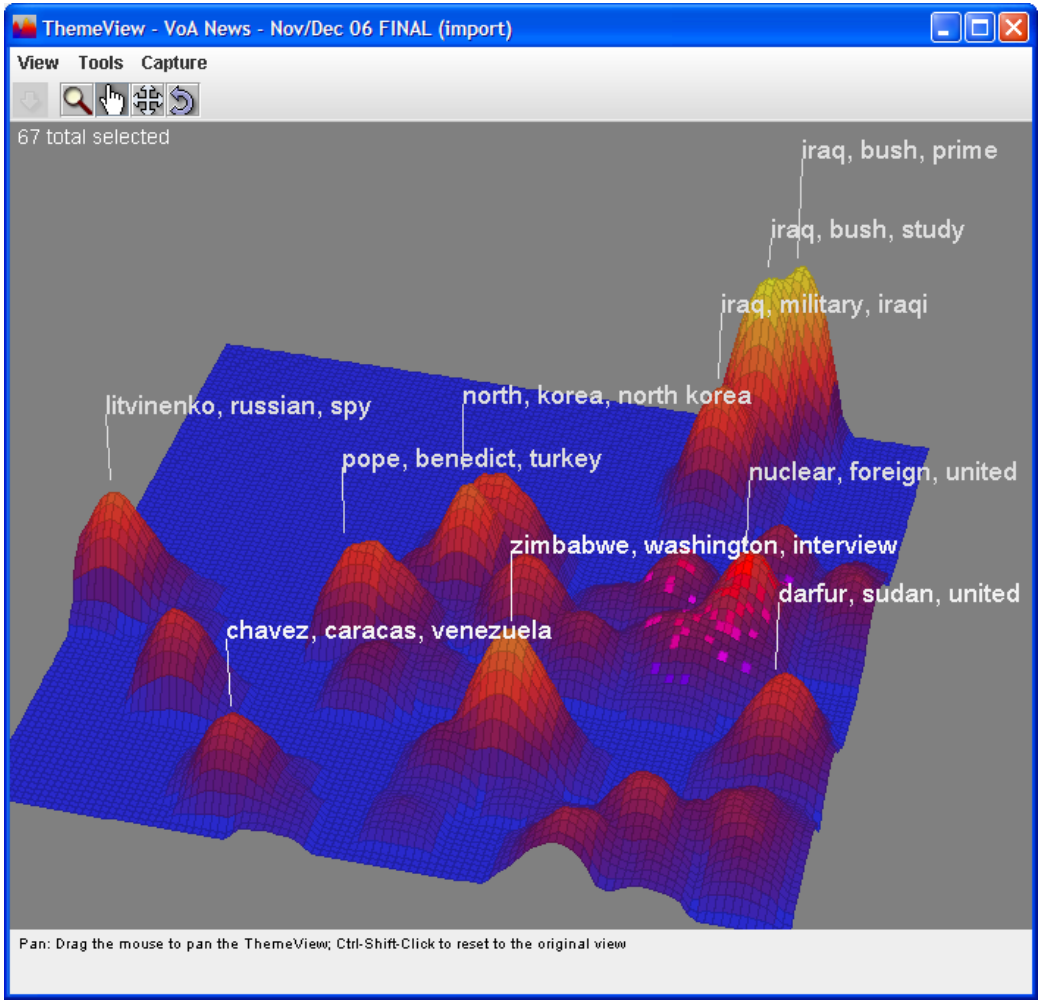




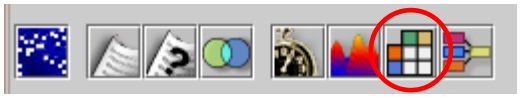
Pacific Northwest
NATIONAL LABORATORY

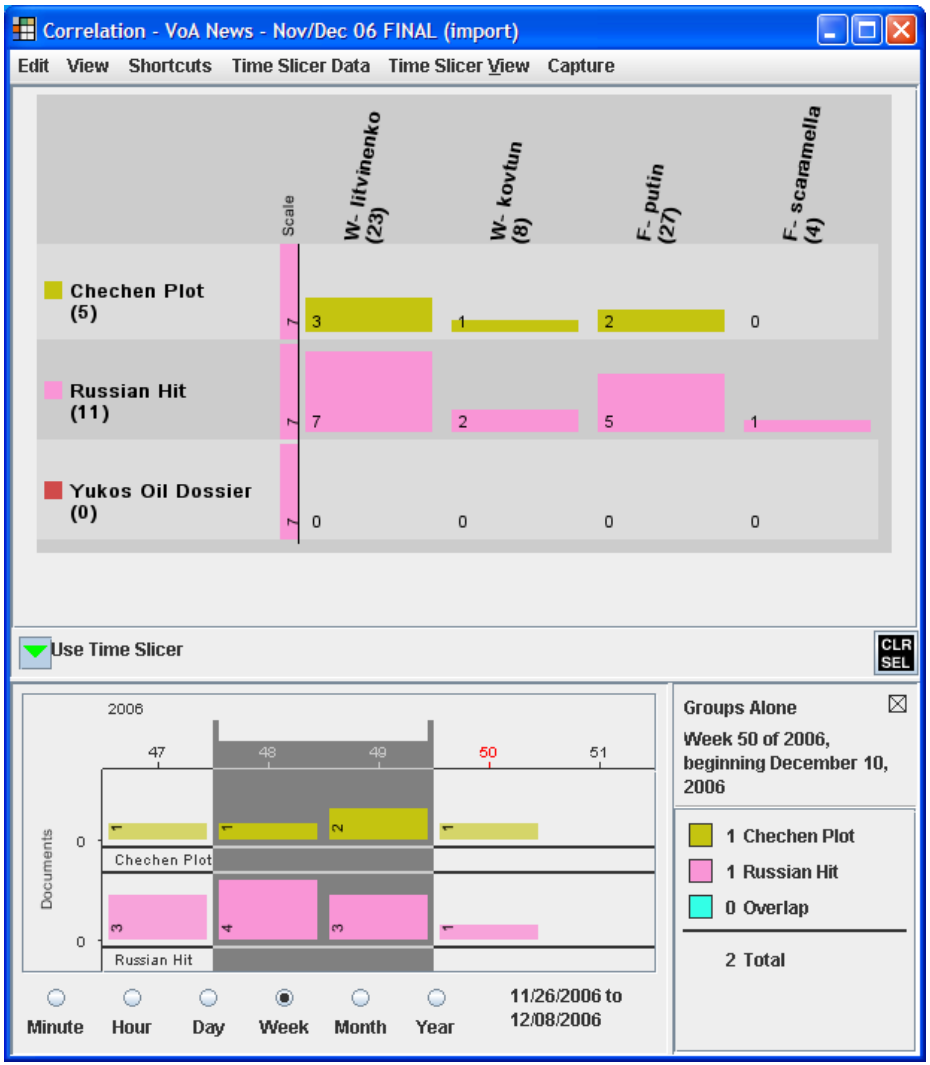
6. ThemeView





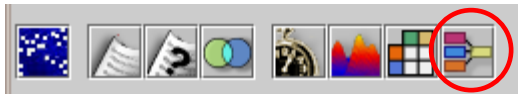
7. Correlation

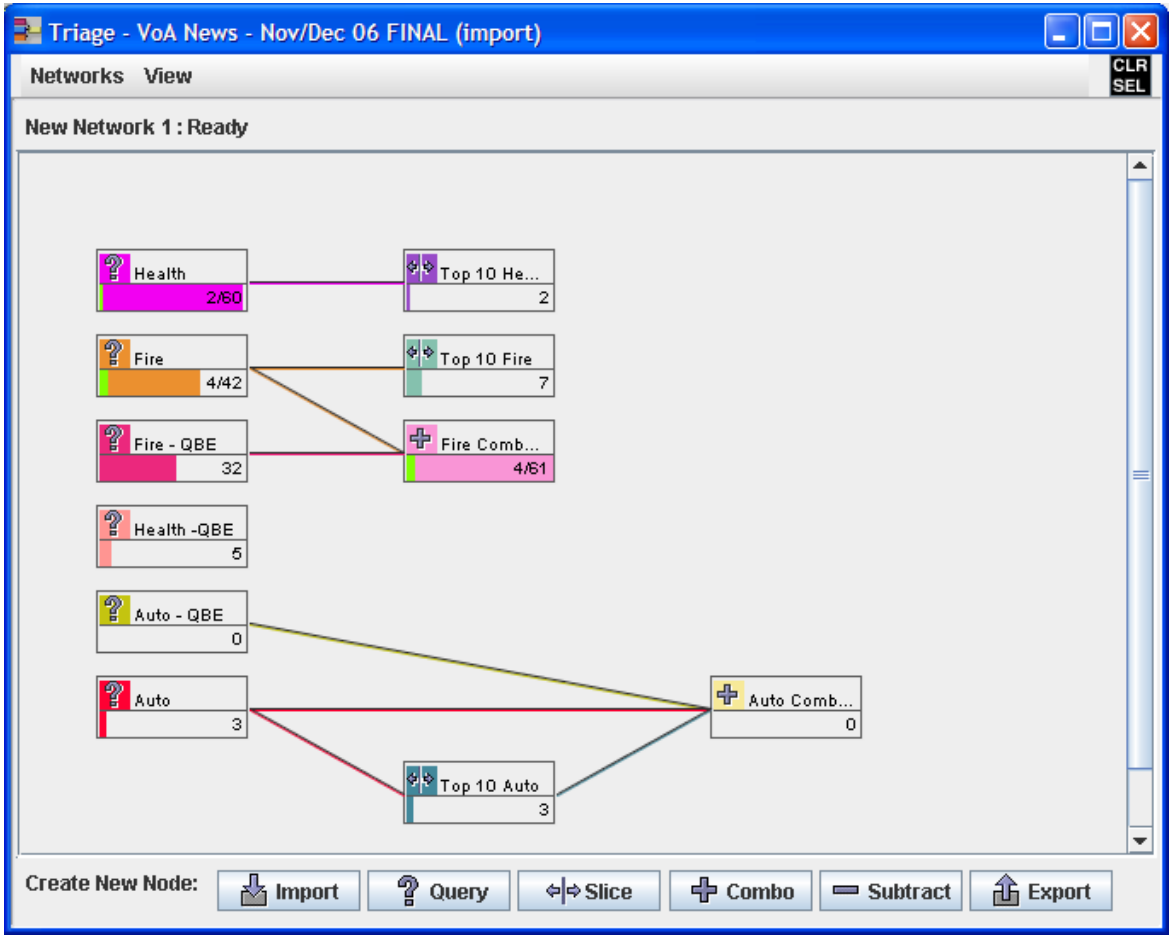




Pacific Northwest NATIONAL LABORATORY

8. Triage





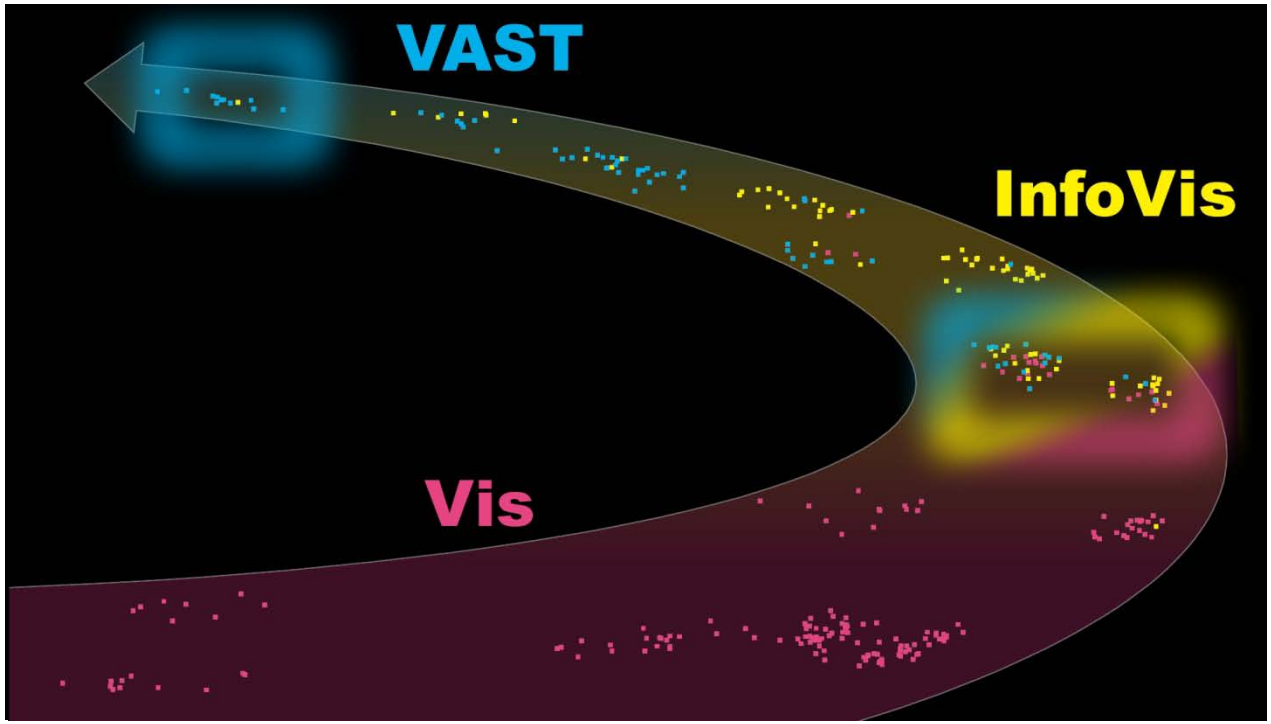
Pacific Northwest NATIONAL LABORATORY

Use Today

- Scientific Research
- Regulatory and Legal Communities
- Intelligence Analysis
- DOE and Department of Defense
- Market Assessments
- Capability Analysis - Resumes
- Medical and Pharmaceutical Communities
- National Security and Law Enforcement
- Information Assurance, Web Analytics
- Technology Scanning, Asset and Intellectual Property Management

IN-SPIRE Demonstration
Technology Convergence & Separation

- Use IN-SPIRE to analyze three years of publications at IEEE Vis, IEEE InfoVis, and IEEE VAST conference from 2007 to 2009
- Vis was first held in 1990, InfoVis in 1995, and VAST in 2006



Pacific Northwest
NATIONAL LABORATORY

Selected Observations & Lessons Learned I

- Current visual technology is far ahead of many (professional) analysts' needs
  - Text analytics is still a human-centric trade in many sectors of our society
  - Some users still flip library cards
  - Computer-savvy users may use spreadsheets and other electronic processing means
  - More advanced users may use text analytics tools such as IN-SPIRE

Selected Observations & Lessons Learned II

- Current visual technology is far, far ahead of the general public's needs
  - After over 15 years of proven success, all major web search engines from Google, to Bing, to Yahoo still use the same text-based user interface
  - Visualization technology is being developed, but the society at large is not ready

Selected Observations & Lessons Learned III

- For those who have successfully adapted modern text analytics technology, basic visualization means such as icon, scatterplot, graph, chart, and browser-based designs are still preferred choices
  - More advanced visualization manifolds have been developed, published, widely recognized, and even imitated, but none of them ever made it to prime time
    - Dozens from PNNL, including ThemeRiver, Topic Islands, etc.
  - Some bring value but also distraction into the analytical process; others simply do not work

Selected Observations & Lessons Learned IV

- Perfect labeling is an unachievable task yet
  - Text labeling is itself an abstract visualization of the abstract text information
  - No one correct standard for selecting or placing labels
  - Unfortunately, one person's intuition is often at odds with another's
  - Five analysts could have five different interpretations on the same labeled visualization (until they read the actual articles)

Selected Observations & Lesson Learned V

- Always avoid difficult, tedious, and complicated data ingest processes
  - More data ingest overhead → less convenient for analytical tasks → less user would use the tool
  - See Universal Parsing Agent (UPA) http://upa.pnl.gov

Selected Challenges Ahead I

- Multimedia (text) information analytics
  - A combination of text, image captions, closed captioning of audios and videos, semantic contents of images and videos
  - Need a new generation of intelligence technology and thinking to integrate multimedia text information

Selected Challenges Ahead II

- Bag-of-words approach will continue to prevail, even though stronger NLP technology will be available in the future
  - Internet slang words, fragmented sentences, and unorganized and unstructured text (e.g., cut and pasted web pages, OCR translations) will define the future of text analytics

Selected Challenges Ahead III

- Mining multiple corpora
  - Ontology alignment
    - Each corpus dictates the salient term selections through statistical means
    - One topic can be critically important in one corpus; the same topic can be totally ignored in a second corpus
  - Visualization anchoring
    - Additions of new data entries (or a different corpus) can turn a visualization and its ongoing analysis into chaos
      - "Now you see it, now you don't"
    - ∑ parts (i.e., individual corpus) != whole

Panel Discussion

- Similarity or discrepancy among different communities, languages, cultures, applications…
- Share your hurdles and successes of text visualization and analytics
- Other important topics that should have been covered in today's program
- …

Acknowledgments

- Thanks to the workshop organizers at ACM IUI 2010 for inviting me to participate in this panel discussion