Case Study: Jigsaw
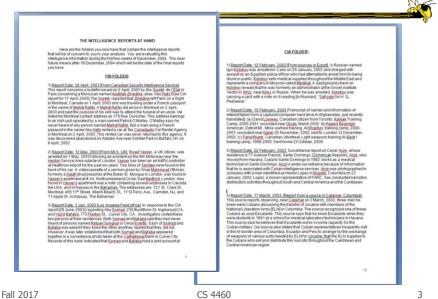
CS 4460 – Intro. to Information Visualization
September 15, 2017
John Stasko

Learning Objectives

- Become familiar with investigative analysis process carried out by various types of analysts
- Learn about visualizations used to represent text and document data
- Gain familiarity with how Jigsaw system works and how to use it to explore document collections

1

An Illuminating Exercise

HW 3

- 50 synthetic intelligence documents
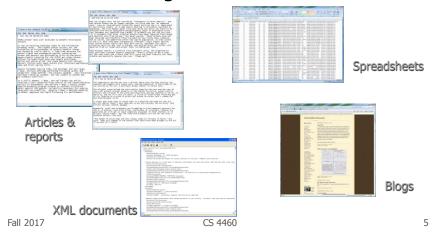- Identify and describe the hidden threat

2

Problem Addressed

Help "investigators" explore, analyze and understand large document collections

Spreadsheets

Articles & reports

XML documents

Blogs

Example Documents

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.

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.

Challenge

- Suppose you are an intelligence analyst and you have hundreds or thousands of such documents
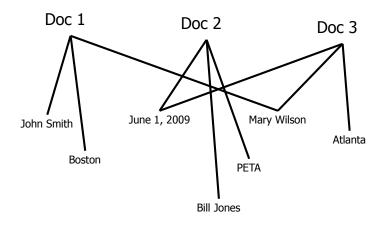- How would you explore for hidden plots and threats?

One Emphasis

- 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

Doc 1   Doc 2   Doc 3

John Smith   June 1, 2009   Mary Wilson

                                         Atlanta

Boston

                              PETA

Bill Jones

Multiple Domains

Visualization for Investigative Analysis
  across Document Collections
        Law enforcement & intelligence community
        Fraud (finance, accounting, banking)
        Academic research
        Product reviews
        Blogs, stories
        Journalism & reporting
        Consumer research

Jigsaw

Computational analysis of documents' text
Entity identification, document similarity, clustering, summarization, sentiment

Multiple visualizations of documents, analysis results, entities, and their connections
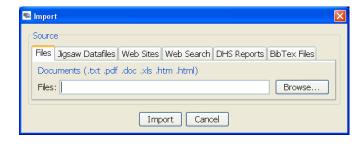Views are highly coordinated

Analysis Process

1. Document import
2. Entity identification
3. Computational text mining
4. Exploration of visualizations

1. Document Import

- Text, csv, pdf, Word, html
- Jigsaw data file format
  - Our own xml

2. 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
  - GATE, Lingpipe, OpenCalais, Illinois (Roth)

Sample Document 1

```
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.

8

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
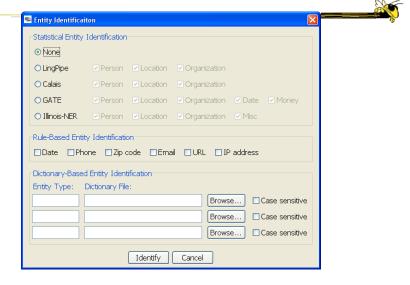Organization: Northwest Central Univ.
Amount: 123,456
Program Mgr: Ephraim Glinert
Division: IIS
ProgramElementCode: 2860

Entities Already Identified

Title: Proving Columbus was Wrong   **Text**
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   **Entities**
Division: IIS
ProgramElementCode: 2860

Entity Identification

3. Computational Text Analyses

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

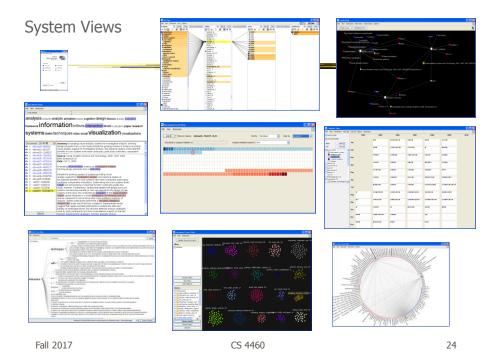# 4. Visualization Exploration

Entity Connections

- Core Principle
  - 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

View Coordination

- Actions (selections, expansions) in one view are propagated to others that update appropriately
- Views can be "frozen"

System Views

Demo

Documents

231 reviews of 2009 Hyundai Genesis

A Review

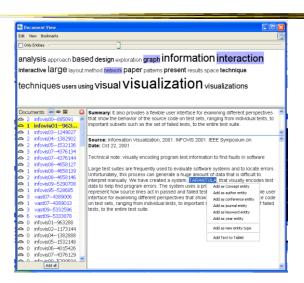Rating from 1 to 10 on multiple categories:
Reliability, Overall

Free narrative text

Entities
- Overall rating, value is integer score
- Car makes and models mentioned in narrative text
  (Ford, Audi, BMW, Corvette, Sonata, ES300)
- Car "features" mentioned in narrative text
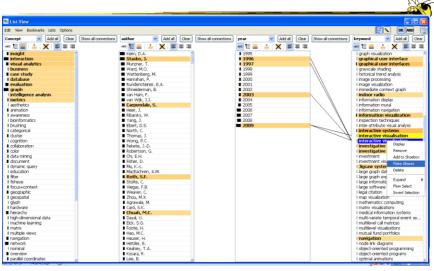 (brakes, transmission, stereo, seats)

EI Correction

Entity Aliasing
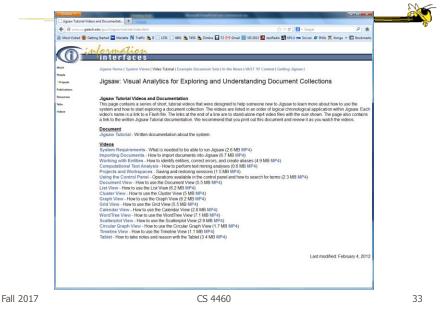
Alias Representation

Tablet

Website Items
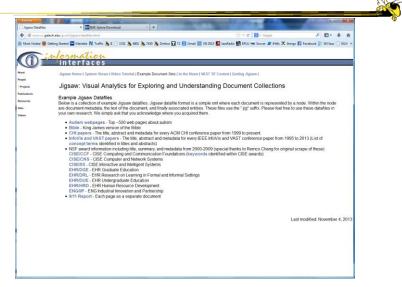
- Tutorial videos
- Sample data sets
- Scenario videos
- ...

16

Video Tutorials

Sample Data Sets

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, wine reviews
  - 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

Learning Objectives

- Become familiar with investigative analysis process carried out by various types of analysts
- Learn about visualizations used to represent text and document data
- Gain familiarity with how Jigsaw system works and how to use it to explore document collections

HW 3

- 50 synthetic intelligence documents
  - In T-square
  - As one big document, as 50 documents
  - As a Jigsaw data file
- Identify and describe the hidden threat
  - Write a 3-4 sentence hypothesis
- Write a few paragraphs about your analysis process
- Due next Friday -Don't wait til last minute

Upcoming

- Multivariate visual representations 1
  - Prep: Watch system demo videos

- Multivariate visual representations 2