



Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Conversation 2: Digital News and its Scholarly Uses

Mary Feeney, Moderator

Librarian for the Social Sciences, Research & Learning Department
University of Arizona Libraries

eDesiderata®
FORUM



Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Digital News and its Scholarly Uses

Mary Feeney
Social Sciences Librarian
The University of Arizona Libraries

Scholars Using News Content

- **Many different disciplines**
 - History
 - Journalism
 - Communication
 - Political Science
 - Sociology
 - Linguistics
 - Education
 - Literature
 - Music
 - and so on...
- **Many different uses**
 - Entire news source or specific pieces:
 - Ads
 - Editorials
 - Obituaries
 - Political cartoons
 - Literary works
 - Photographs
 - Data
 - Broadcast transcripts
 - News video

Scholars Accessing News Content

- **Formats/types**

- Print
- Microforms
- Digitized archival collections – open access and commercial databases
- Commercial news aggregator databases
- News websites
- News apps

- **Methods**

- Browse
- Search
- Download
- Extract
- Text analysis/text mining
- Content analysis
- Network analysis
- and much, much more!



Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Conversation 2: Digital News and its Scholarly Uses

James Danowski

Communication Professor Emeritus
University of Illinois at Chicago
jdanowski@catsci.org

eDesiderata®
FORUM

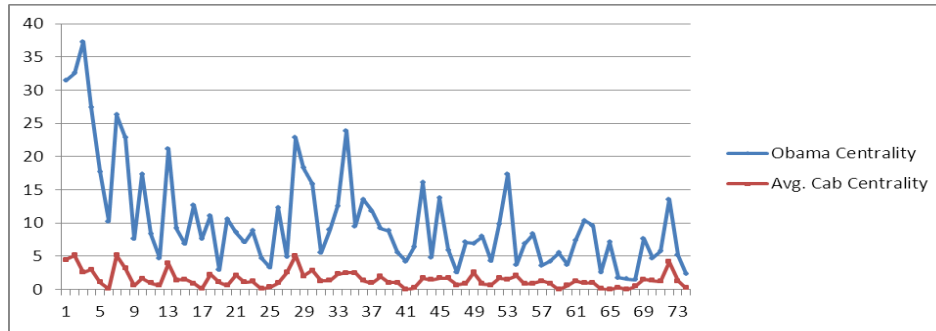
Extracting News Documents for Political Communication Research

1. News Extraction #1: Presidential network centrality and job approval
2. News Extraction #2: Arab Spring jihad semantic networks
3. News Extraction #3: Optimal messages for community resilience
4. **Six challenges in extracting news documents**

News Extraction #1

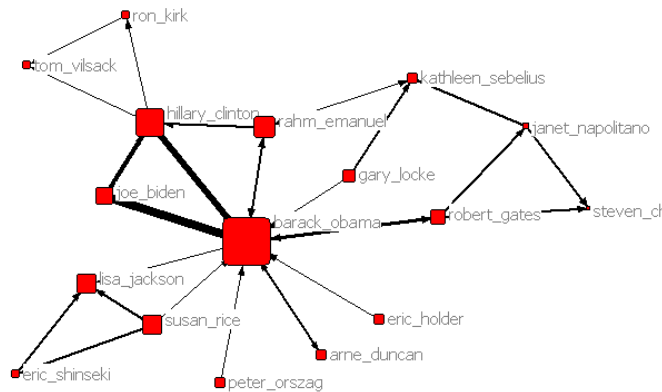
Centrality of President among cabinet members

News Extraction #1



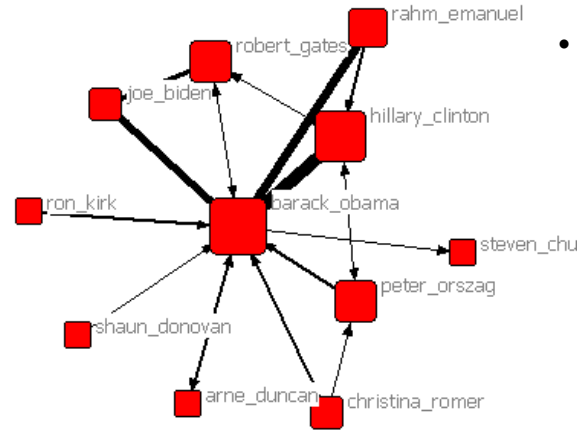
- Obama's centrality relative to average cabinet centrality significantly correlated with job approval.

Debt Crisis



Obama **centrality high** relative to the group

Libyan Civil War



Obama **centrality low** relative to the group News Extraction #1

- Occurs at a lag of $L=3$, a period of 6 weeks.

News Extraction #2

Changing semantic networks for *jihad* among majority-Muslim nations before and after the Early Arab Spring Uprisings

News Extraction #2

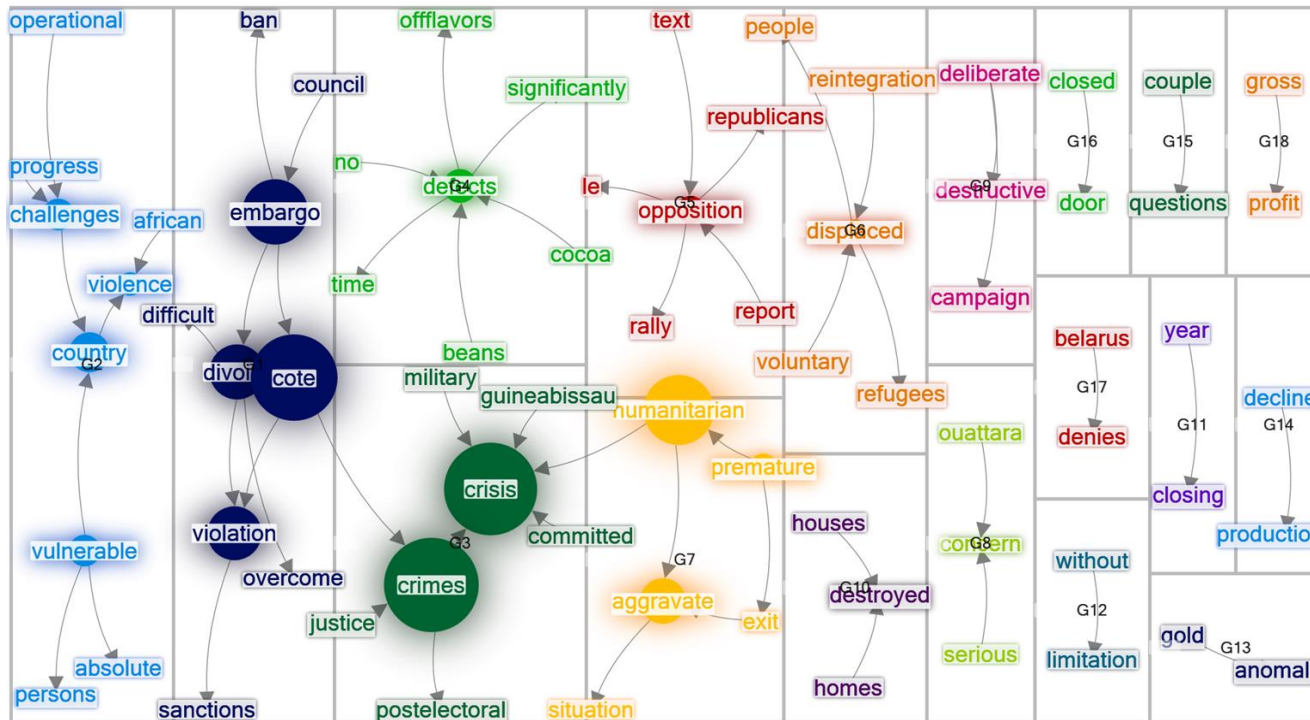
News Extraction #3

Ivory Coast Community Resilience Study after Muslim-Christian Civil War

News Extraction #3

High Mobility Sentiment Predictors Lags 0-3

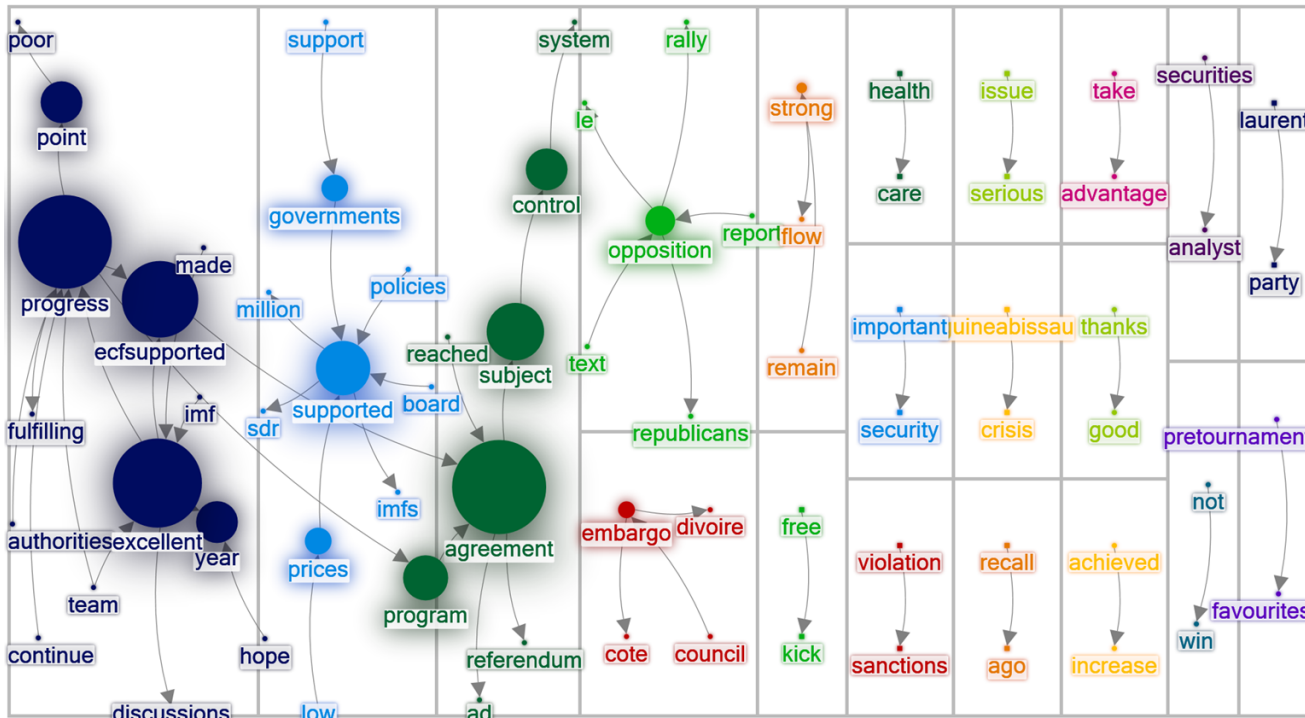
(n=255 subprefectures)



☹️ **Negative word pairs** → **increased mobility (to plan or to engage in violence?)**

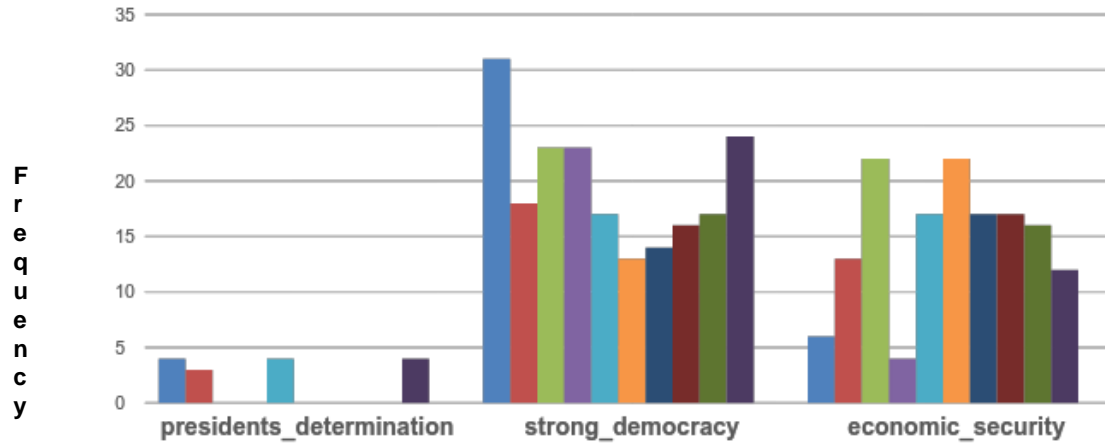
Low Mobility Sentiment Predictors Lags 0-3

(n=255 subprefectures)



☺ **Positive media words** → **low mobility (stay home and be happy).** News Extraction #3

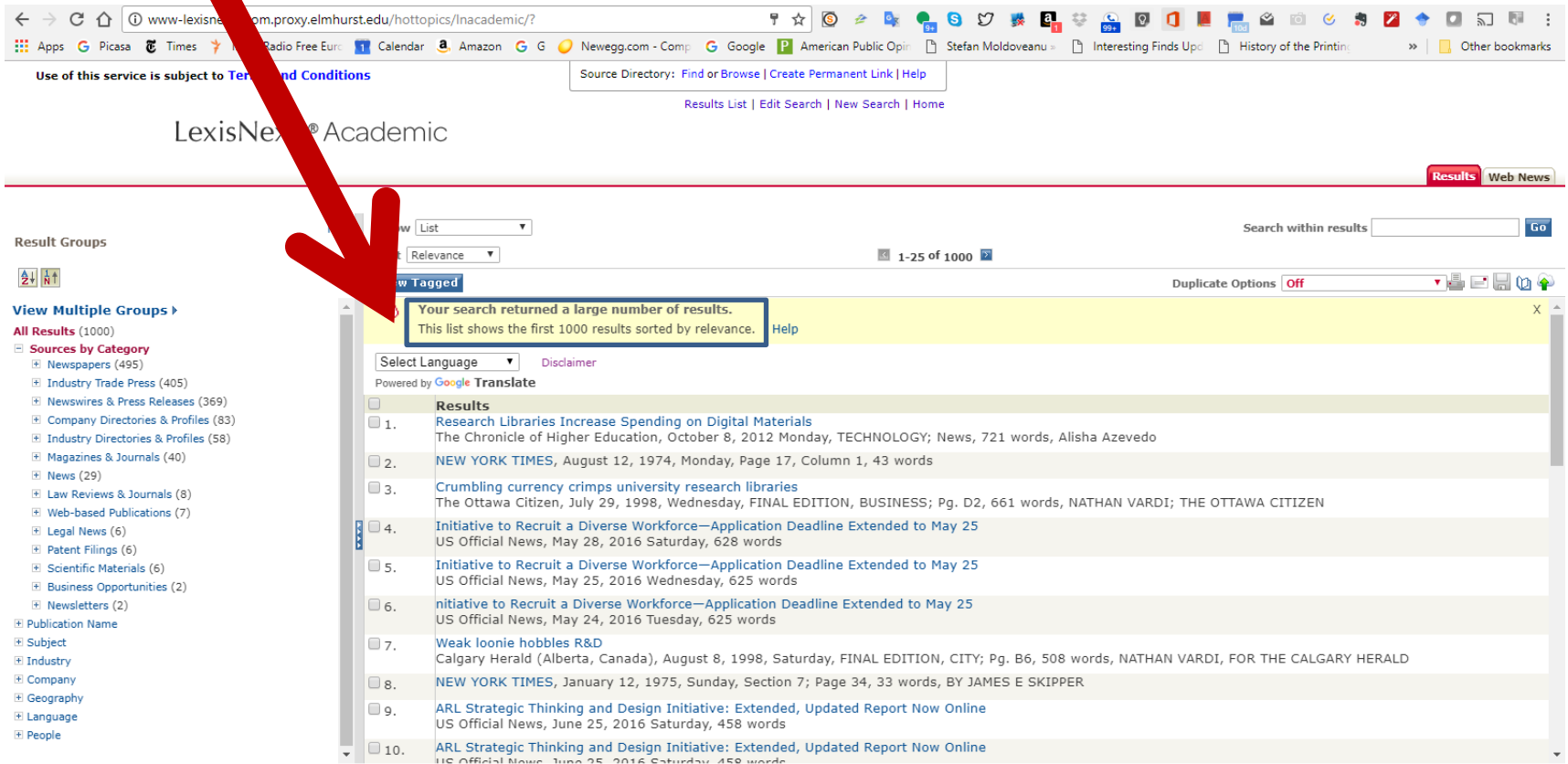
Optimal Word Pairs Activated by Photos



4. Six Challenges Extracting News Documents

Four Problems

Challenge #1: More than 1000 Hits Problem



The screenshot shows a web browser window displaying the LexisNexis Academic search results page. The address bar shows the URL: www.lexisnexis.com.proxy.elmhurst.edu/hottopics/lnacademic/. The page header includes the LexisNexis Academic logo and navigation links like "Results List", "Edit Search", "New Search", and "Home".

The search interface shows a search bar with the text "w List" and a dropdown menu set to "Relevance". Below the search bar, it indicates "1-25 of 1000" results. A yellow message box with a black border and a red arrow pointing to it contains the text: "Your search returned a large number of results. This list shows the first 1000 results sorted by relevance." A "Help" link is also visible next to the message.

The results list is titled "Results" and is powered by Google Translate. It contains 10 search results, each with a checkbox, a title, and a brief description. The results include:

1. **Research Libraries Increase Spending on Digital Materials**
The Chronicle of Higher Education, October 8, 2012 Monday, TECHNOLOGY; News, 721 words, Alisha Azevedo
2. **NEW YORK TIMES**, August 12, 1974, Monday, Page 17, Column 1, 43 words
3. **Crumbling currency crimps university research libraries**
The Ottawa Citizen, July 29, 1998, Wednesday, FINAL EDITION, BUSINESS; Pg. D2, 661 words, NATHAN VARDI; THE OTTAWA CITIZEN
4. **Initiative to Recruit a Diverse Workforce—Application Deadline Extended to May 25**
US Official News, May 28, 2016 Saturday, 628 words
5. **Initiative to Recruit a Diverse Workforce—Application Deadline Extended to May 25**
US Official News, May 25, 2016 Wednesday, 625 words
6. **initiative to Recruit a Diverse Workforce—Application Deadline Extended to May 25**
US Official News, May 24, 2016 Tuesday, 625 words
7. **Weak loonie hobbles R&D**
Calgary Herald (Alberta, Canada), August 8, 1998, Saturday, FINAL EDITION, CITY; Pg. B6, 508 words, NATHAN VARDI, FOR THE CALGARY HERALD
8. **NEW YORK TIMES**, January 12, 1975, Sunday, Section 7; Page 34, 33 words, BY JAMES E SKIPPER
9. **ARL Strategic Thinking and Design Initiative: Extended, Updated Report Now Online**
US Official News, June 25, 2016 Saturday, 458 words
10. **ARL Strategic Thinking and Design Initiative: Extended, Updated Report Now Online**
US Official News, June 25, 2016 Saturday, 458 words

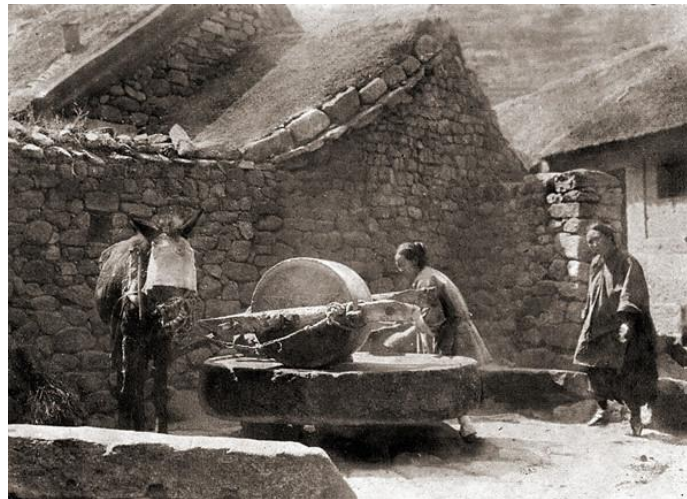
Challenge #2: Download Only 500 Documents per File

The screenshot displays the LexisNexis Academic web interface. A red arrow points from the title 'Challenge #2: Download Only 500 Documents per File' to a dialog box that has appeared over the download options. The dialog box, titled 'www.lexisnexis-com.proxy.elmhurst.edu says:', contains the message: 'You have exceeded the maximum number of documents allowed (500) for delivery. Please select fewer documents and try again.' Below the message is an 'OK' button. The background interface shows a search results page with a list of documents, a 'Download' button, and various options for document delivery and formatting. The browser's address bar shows the URL 'www.lexisnexis-com.proxy.elmhurst.edu/hottopics/Inacademic?'. The browser's taskbar at the top shows several open applications, including Amazon, Google, and various news and public opinion sites.

Four Problems

Challenge #2 Example: Downloads Needed for the Presidential Cabinet Centrality Study

- Nixon 61
- Ford
- Carter
- Reagan
- GHW Bush
- Clinton 229
- GW Bush 180
- Obama 144



Manually grinding out **1,051 files**, one rotation at a time.

Four Problems

Challenge 3: Stripping Meta Data



Five Problems

Challenge #4 Duplicate Documents



Five Problems

Challenge #5: No Photos



Five Problems

Challenge #6: No Chinese in Lexis-Nexis Academic

没有中文文件



About LexisNexis | Contact Us | Worldwide: China | Language: 中文 | Site Feedback | Product Sign-In

[Products & Services](#) | [Support](#) | [Get Involved](#) | [Media Center](#) | [E-Library](#)

[Home](#) > [Products](#) > [Wisers](#)

[Print This Page](#)

Greater China News Database

Get the information and insight you need to get ahead of the competition in Greater China, with 2,500 sources of news and business information from top-tier to local media

Contact Customer Service

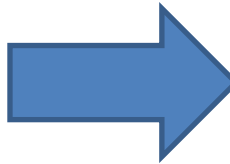


Tel: +86 [400 600 8680](tel:+864006008680)
service.china@lexisnexis.com

钱



Lexis-Nexis User Needs Analyst



Five Problems





Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Conversation 2: Digital News and its Scholarly Uses

Nicholas Adams

Data Science Fellow

Berkeley Institute for Data Science

eDesiderata®
FORUM



Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

News as Data: Analyzing Events at Scale

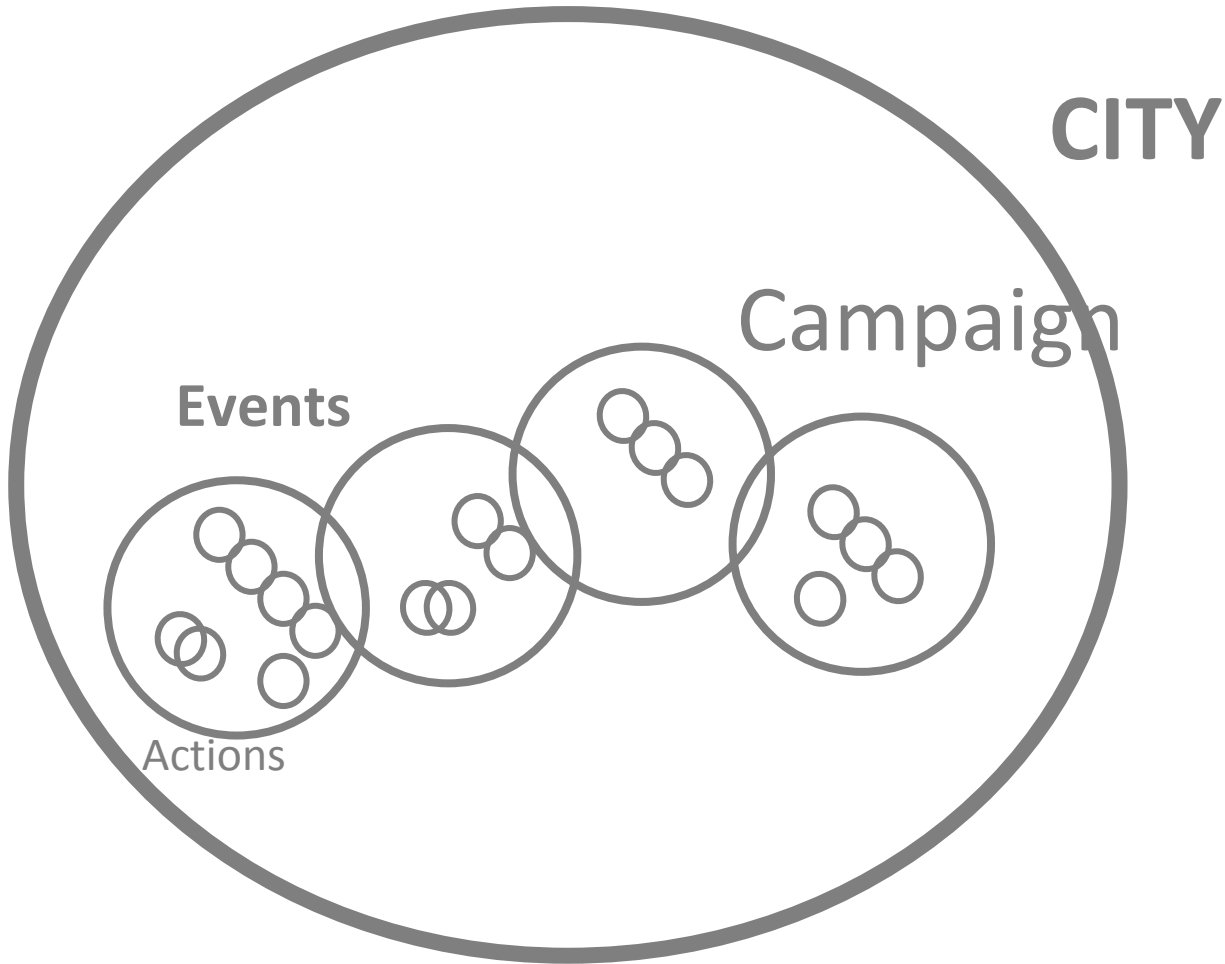
Nicholas Adams, Ph.D.
Berkeley Institute for Data Science
Goodly Labs
nickbadams@berkeley.edu

A Longstanding Challenge

How can we identify and extract reliable and richly linked data about events reported in newspapers?

Even though:

- Events are described across non-contiguous text.
- Journalists do not write according to a temporal sequence.
- Multiple events are reported in a single document.
- A single sentence can report multiple events.
- Multiple events can occur on a single day.



The Reality

Just some of the people working on this...

Human

Soule and Davenport 2009; McAdam & Su 2002; McCarthy & McPhail 2006; Olzak & Soule 2009; Franzosi, De Fazio, & Vicari 2012; Tilly 2008; King, Bentele, & Soule 2007; Davenport 1997; Della Porta & Tarrow 2012

Machine

Hammand and Weidmann 2014; Gao et al. 2013; Keertipati 2014; Kwak & An 2014; Leetaru and Schrodtt 2013; Schrodtt and Yonamine

Note: Researchers tend to use one or the other.

Strengths (and weaknesses)

Human

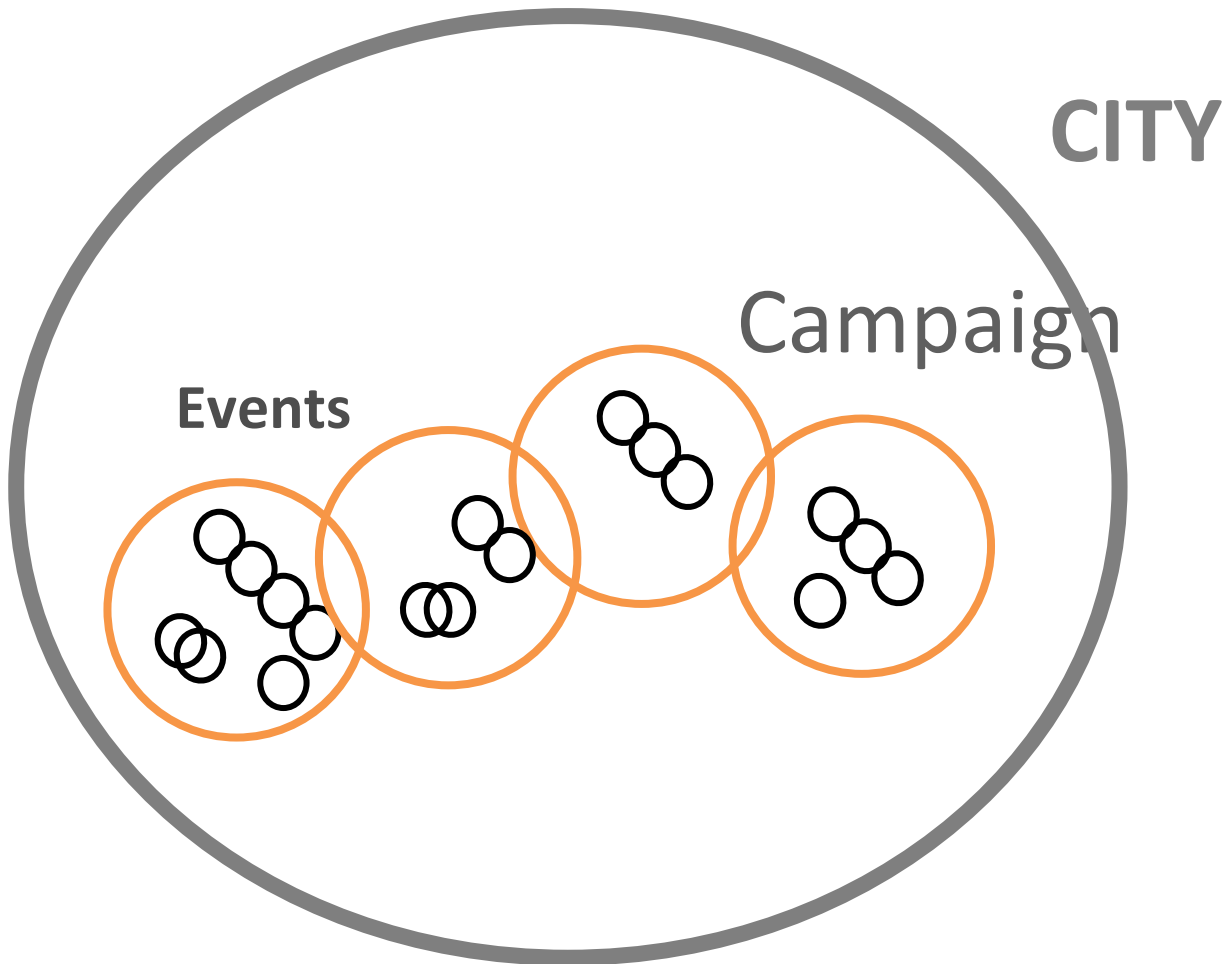
- Valid
- Interpretable
- Easily communicated
- Trusted by Soc. Science & DH

Example: Dynamics of Collective Action (DOCA)

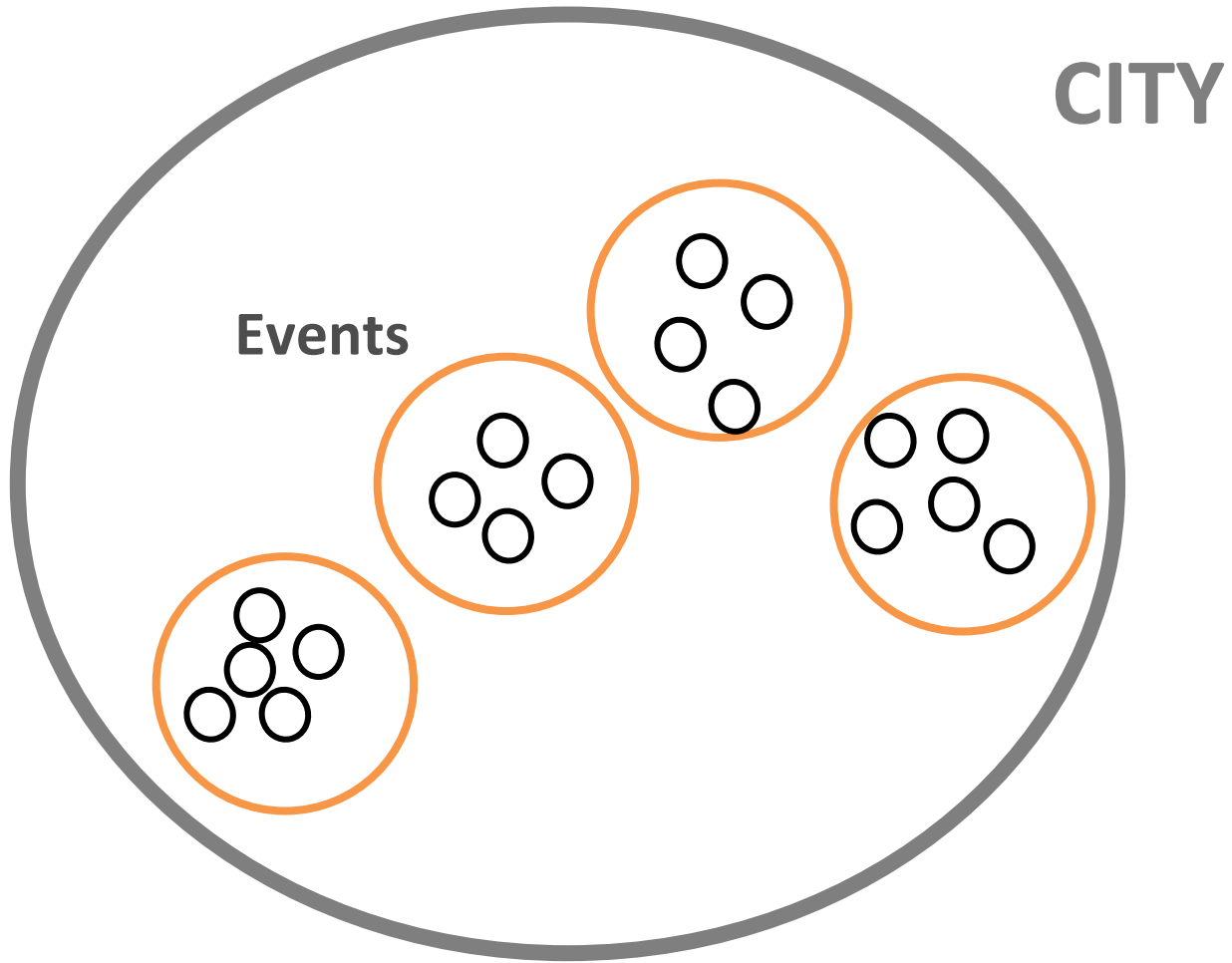
Machine

- Reliable, Replicable, Reproducible
- FAST!!
- Scalable to “Big Data”
- Trusted by CS & CL

Ex: Global Data on Events, Language, and Tone (GDELT)



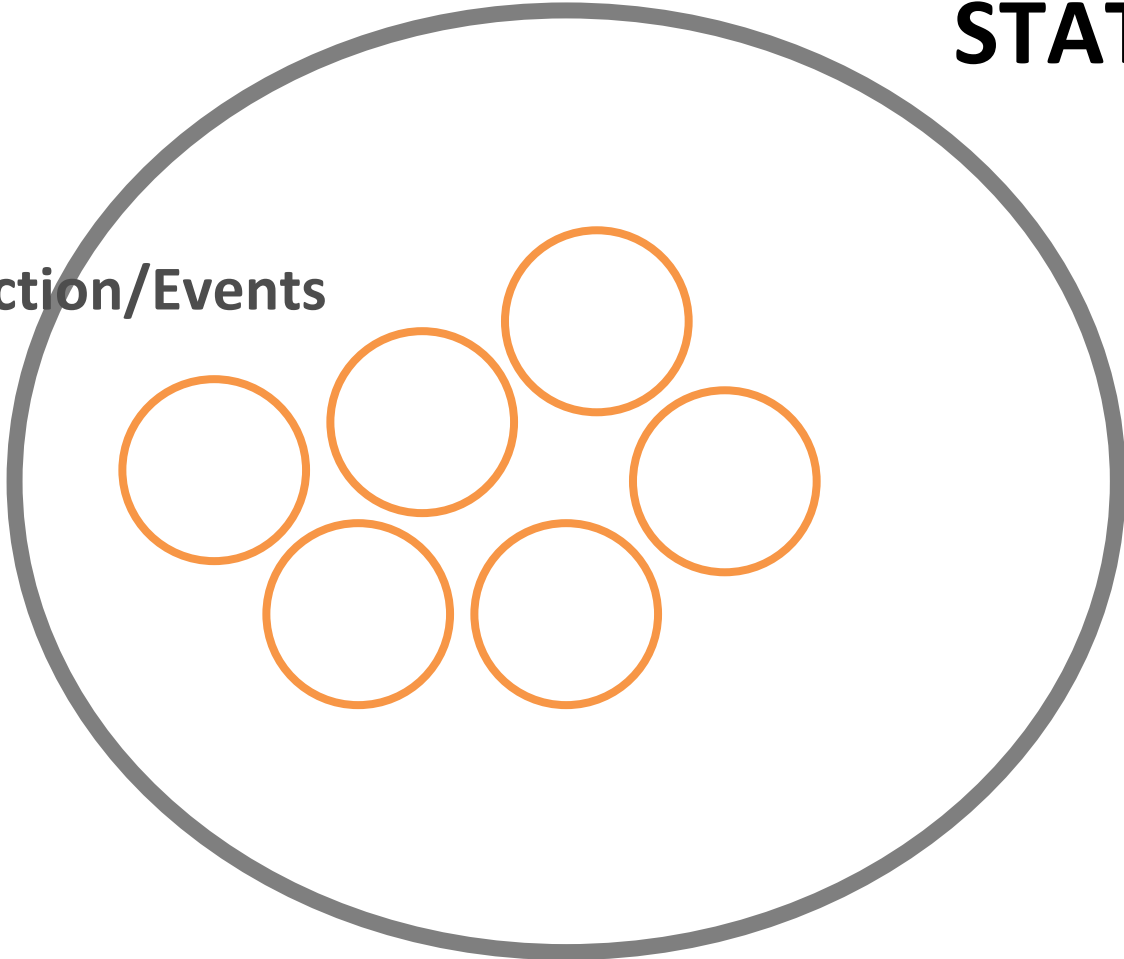
The Reality



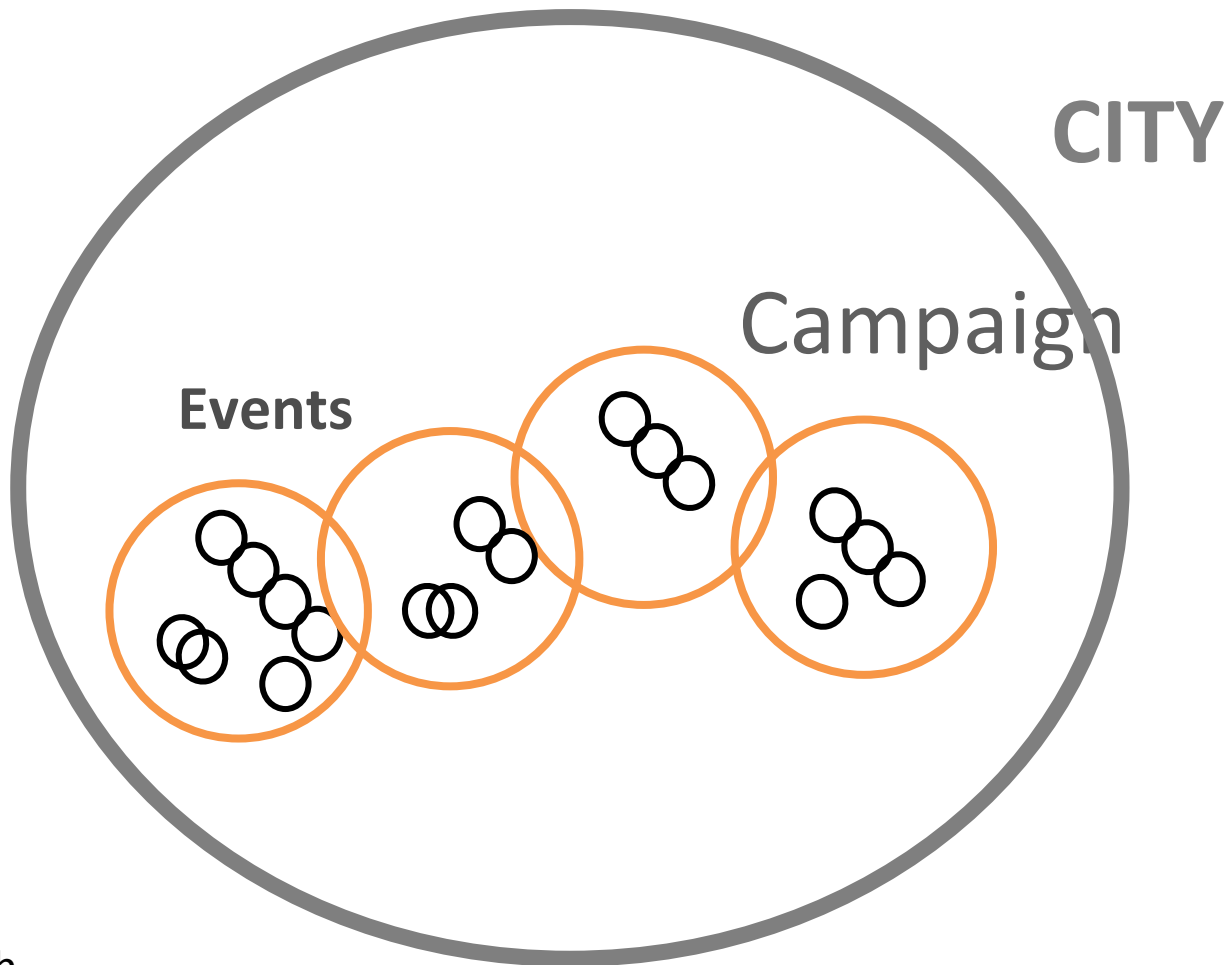
DCA

STATE

Action/Events



GDELT



Our Approach

Refrain: This will be challenging

- Events are described across non-contiguous text.
- Journalists do not write according to a temporal sequence.
- Multiple events are reported in a single document.
- A single sentence can report multiple

One New Hybrid Approach: Performance Modeling

Identify events by hand, get
machines to do the rest

+*+*

10-13-11

CRV, v31, NBA

Originally published Thursday, October 13, 2011 at 9:36 PM

The Associated Press

Bigger Occupy Seattle protest set for Saturday

Occupy Seattle demonstrators **took a break from downtown Seattle's Westlake Park on Thursday to make way for a jobs rally organized by Washington labor unions.**

SEATTLE —

Occupy Seattle demonstrators took a break from downtown Seattle's Westlake Park on Thursday to make way for a jobs rally organized by Washington labor unions.

KING-TV (<http://bit.ly/pCQUJ80>) reports about 10 police officers told some 75 demonstrators they had to leave [some 75 demonstrators they had to leave]. After a brief confrontation, the Occupy Seattle protesters left.

By Thursday evening, about 30 of them were back in the park.

Roughly 25 protesters have moved to City Hall Plaza, where Mayor Mike McGinn has said they are welcome to camp

[Roughly 25 protesters have moved to City Hall Plaza, where Mayor Mike McGinn has said they are welcome to camp]. At midday Thursday, dozens protested outside the downtown hotel where Republican presidential hopeful Mitt Romney had a private fundraiser. They carried signs including one that said, "Romney is the 1 percent" - a reference to what the demonstrators describe as the concentration of wealth among a small slice of the population.

[Occupy Seattle plans a large demonstration Saturday.](#)

Hand-Labeled Occupy Data

Hand-Labeled Occupy Data as JSON on GitHub

df-data

deciding force data

There are currently this many records:

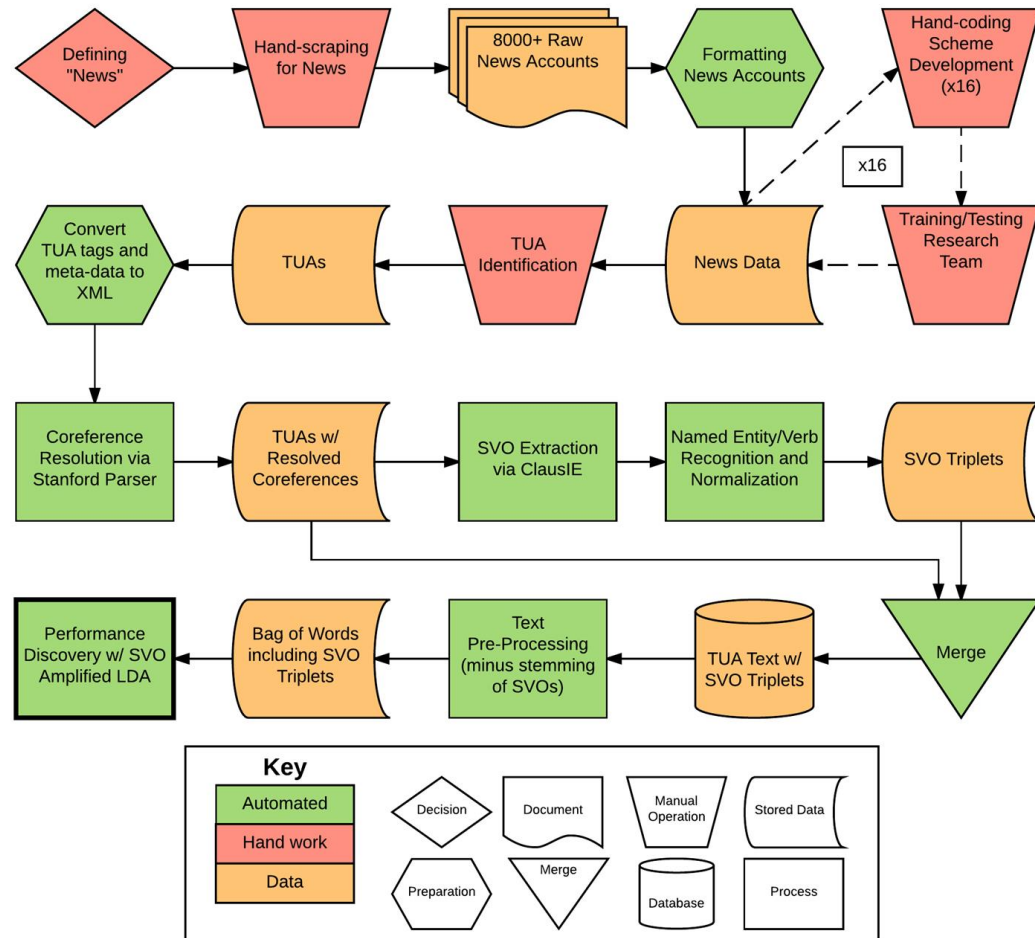
```
gzcat articles.json.gz | jq -c '[] | length'  
5724
```

Example output:

```
gzcat articles.json.gz | jq '.[0] | [.metadata, .tuas.Protester]'
```

```
[  
  {  
    "annotators": [  
      "SLG"  
    ],  
    "article_number": "0",  
    "city": "Albany",  
    "date_published": "2011-11-05",  
    "periodical": "AlbanyDemocratHerald",  
    "periodical_code": "01",  
    "state": "NY",  
    "version": "23"  
  },  
  {  
    "1": [  
      [  
        149,  
        344,  
        "brandishing homemade signs and waving at passing cars, more than 60 people supporting the occupy"  
      ],  
      [  
        345,  
        417,  
        "about five counter-demonstrators were also on hand with signs asking for"  
      ],  
      [  
        445,  
        510,  
        "and occasionally debating with individuals from the larger group."  
      ],  
      [  
        511,
```

We've learned
 from
 developing
 our
 Performance
 Modeling
 technique



Findings

- **American police are not independent of politics**
 - more accommodating, fewer shows of force when elections are near
- **American police are not merely reactive, but behave strategically, based upon their capacities**
 - When understaffed, they move quickly to shut down campaigns – also avoid force-on-force mass arrests, preferring to arrest individuals or smaller groups, and threaten fewer deadlines
- **Police culture and context affects their response to protest as well**
 - Departments committed to Community Policing focused enforcement on individuals, far fewer Raids. Departments in cities with high violent crime rates rather shrugged at Occupy

Another New Hybrid Approach: Classification by Crowds

Use TextThresher software to
break human-led work out into
an assembly line

Occupy Protesters Announce They Will Shut Down Port Of Long Beach Monday

December 11, 2011 11:01 AM

Filed Under: Longshoreman Workers' Union, Occupation, Occupy, Occupy L.A., Occupy LA, Occupy Long Beach, Occupy Los Angeles, Port of Long Beach



LISTEN LIVE

FOLLOW

Sign Up for

Start saving for your future today.

Sign up for myRA >

myRA U.S. Department of the Treasury

>8,000 Articles

LONG BEACH (CBS) — Occupy activists have announced plans to shut down the Port of Long Beach on Monday in hopes that it will disrupt business for the "1%" whose business relies on the port.

Organizers say the shutdown is a coordinated effort to close cargo ports on the West Coast from Anchorage to San Diego. Dock workers say the move will cost shipping companies millions of dollars in losses.

The organizers' goal is "to disrupt and blockade the economic apparatus of the 1%," according to the [Occupy the Ports website](#).

Activists said they plan on gathering at a park near the Queen Mary at 5 a.m. Monday, and then march to a dock facility owned by SSA Marine, a shipping company that is 51 percent owned by Goldman Sachs. They said they would likely disperse by 11 a.m.



(credit: Michal Czerwonka/Getty Images)

Members of the International Longshoreman Workers' Union were expected by Occupy organizers to honor the community march as a

Advertisements

The #1 "The Alternative Experts are 'liquid gold'"

How Old My healthy living! There's a re Robert Dow

Save with Amex Travel Get the Low Guaranteed

Look &

- Police-Initiated EVENT
 - **EVENT TYPE
 - Raid of camp
 - Showdown at the camp (i.e. po situation that might lead to ar
 - Warnings or Threats at the car inform campers about impend with intention of coercing a ch
 - Surveillance of camp
 - Surveillance of individual prote
 - Arrests of protest organizers a or CAMP
 - Undercover infiltration
 - Propaganda camping against o
 - "Snatch 'n' grab" arrests at car (especially key organizers, at t
 - Space-taking (i.e. establishing lines of police for containment
 - Statement (to public and/or oc eviction order or permit expira city
 - New! Notable Non-event e.g. i Everyone is expecting police to
 - Official statement from police spokesperson
 - Code the act of speakin press release in *blue*, all official spokesperson
 - Code the conten rest of the codin
 - CROWD COMPOSITION (How many
 - Protesters
 - Protesters from different cities
 - From where/
 - How many
 - Counterprotesters
 - Anarchists
 - Union Members
 - Media
 - General Indicator of Diversity
 - Ages
 - Classes
 - Ethnicities
 - ...
 - Police presence characteristics
 - Riot gear
 - Horses
 - Urban Assault Vehicles
 - Brandishing weapons
 - Skirmish lines
 - Actions of police or protesters that
 - Who
 - Police
 - Protester
 - Counter-protester
 - Anarchist
 - How many
 - Tag
 - Estimate
 - What
 - Punch
 - Kick
 - Push/shove
 - To Whom
 - Police
 - Protester
 - Counter-protester
 - Anarchist
 - Media
 - How many
 - Tag
 - Estimate
 - Time
 - Sequence order
 - Date/Time tag
 - START
 - END

- Protester-Initiated EVENT
 - Date/Time tag
 - START
 - END
 - **EVENT TYPE
 - Establishing a camp (wil
 - Voluntary Dissolution of police-Initiated raid
 - Moving a camp to a new
 - NOTE: Implies th count it as a mov process. If the ol move to a new l simply be coded
 - March/Parade (unless p at the start point or end counted as one event.)
 - Rally/Demonstration
 - NEW!! Disrupting an conservative politician's council meeting ONLY if proposal, etc.)
 - Strike
 - Divestment action (e.g. Blocking Action
 - Sidewalk
 - Street
 - Public transport
 - Airport
 - Shipping port
 - Strategic violence
 - Kidnapping
 - Assassination
 - Bombing
 - Assault
 - Strategic sabotage
 - Pre-planned van
 - Pre-planned arse
 - PERMITTED?
 - From when
 - Till when
 - With what conditions
 - CROWD COMPOSITION (How
 - Diversity of...
 - Ages
 - Classes
 - Ethnicities
 - ...
 - Police
 - Protesters
 - New! Protesters from o
 - Which city?
 - Counterprotesters
 - Anarchists
 - Media
 - Union Members
 - Religious leaders or con
 - Occupy the Hood Folks
 - Other ALLIED Groups
 - Police presence characteristi
 - Riot gear
 - Horses
 - Urban Assault Vehicles
 - Brandishing weapons
 - Skirmish lines

- Civilian Government Action
 - Order or threat of enforcement or legal consequences, or s or intention not to enforce
 - Order to disperse
 - New Curfew order (might be bracketed in brick, too)
 - Announcement of intention to enforce or not enforce rea (not permits) relating to encampment, marches, demons etc.
 - New Regulation
 - Warning
 - New! Inspections (Fire, Safety, etc.)
 - Indication of intention NOT to enforce
 - Dispersal order
 - Refusal of basic needs like sanitation services, water, ele (port-a-ion, porta-potty, etc.)
 - Other official statement from city council member or high officer or spokesperson for such an officer
 - Code the act of speaking, writing an open letter, o press release in purple along with the name and/ the official spokesperson
 - Code the content of the speech act accordi rest of the coding scheme
 - Evidence of city ordinances
 - Against semi-permanent structures
 - Against overnight camping
 - ETC...
 - COSTS to City
 - Estimates of damage, cost of extra personnel, etc.
 - Meeting
 - Attendees
 - Mayor
 - City Administrator
 - City council member
 - Governor
 - Police Official
 - City official from other city
 - ...
 - Subjects discussed
 - Strategies for ending protest
 - Strategies for managing protest
 - Accountability
 - Blame
 - ... city officials
 - ... POLICE
 - ... protesters
 - Praise
 - ... city officials
 - ... POLICE
 - ... protesters
 - Date/Time tag
 - Relevant Actors
 - Governor
 - Mayor
 - City Administrator
 - Vice Mayor
 - City Council (as a body)
 - ...
 - Mayor
 - City Administrator
 - City Council/Assembly
 - Governor
 - Attorney General of State
 - City attorney (if speaking as advisor or legal expert on cit ordinances and not as representative of city in an offic

+*+*

10-13-11

CRV, v31, NBA

Originally published Thursday, October 13, 2011 at 9:36 PM

The Associated Press

Bigger Occupy Seattle protest set for Saturday

Occupy Seattle demonstrators **took a break from downtown Seattle's Westlake Park on Thursday to make way for a jobs rally organized by Washington labor unions.**

SEATTLE —

Occupy Seattle demonstrators took a break from downtown Seattle's Westlake Park on Thursday to make way for a jobs rally organized by Washington labor unions.

KING-TV (<http://bit.ly/pCQUJ80>) reports about 10 police officers told some 75 demonstrators they had to leave [some 75 demonstrators they had to leave]. After a brief confrontation, the Occupy Seattle protesters left.

By Thursday evening, about 30 of them were back in the park.

Roughly 25 protesters have moved to City Hall Plaza, where Mayor Mike McGinn has said they are welcome to camp

[Roughly 25 protesters have moved to City Hall Plaza, where Mayor Mike McGinn has said they are welcome to camp]. At midday Thursday, dozens protested outside the downtown hotel where Republican presidential hopeful Mitt Romney had a private fundraiser. They carried signs including one that said, "Romney is the 1 percent" - a reference to what the demonstrators describe as the concentration of wealth among a small slice of the population.

[Occupy Seattle plans a large demonstration Saturday.](#)

Hand-Labeled Occupy Data

df.goodlylabs.org/project/dfdemo/task/86

pybossa Community Projects Create About Nick

CAMP

GOVERNMENT

PROTESTER

POLICE

22 of 392 DOCUMENTS

The Associated Press

October 22, 2011 Saturday 08:02 AM GMT

Oakland protesters defy city order to leave

BYLINE: By TERRY COLLINS, Associated Press

SECTION: DOMESTIC NEWS

LENGTH: 132 words

DATELINE: OAKLAND, Calif.

Hundreds of anti-Wall Street protesters defiantly remained at their campsite outside Oakland's City Hall early Saturday, despite a city order to vacate. As the 10 p.m. time of the city's ultimatum passed Friday night, Occupy Oakland demonstrators showed no signs of departing as music blasted from the plaza. More protesters arrived with tents as midnight approached. Earlier, city spokeswoman Karen Boyd said that Oakland gave official notice that the protesters do not have permission to remain overnight and that their encampment is breaking the law. She would not comment on what steps the city would take toward enforcing of the law. There was no indication of significant police presence early Saturday. Boyd says that protesters can legally demonstrate at the plaza from 6 a.m. to 10 p.m.

LOAD-DATE: October 23, 2011

LANGUAGE: ENGLISH

PUBLICATION-TYPE: Newswire

Copyright 2011 Associated Press

All Rights Reserved

Instructions: Highlight text that describes the time, location, and all the actions of a protester-initiated event occurring during the Occupy movement. Remember, a protester-initiated event is one that protesters plan or lead, AND that is not in direct immediate response to an event initiated by the police/government. Many protester activities occur during events that police initiate, and those protester activities should be highlighted as a part of a POLICE-initiated event. If more than one protester-initiated event is described in the article, be sure indicate each by adjusting the number on the highlight flag for each event.

Highlight text corresponding with high-level topic or unit of analysis.

df.goodlylabs.org/project/dfquiz4/task/80

pybossa Community Projects Create About Nick

Look for answers in the bolded text.

words **Raheem DeVaughn says he didn't set out to commit civil disobedience Sunday - he planned only to hear Cornel West speak to the Occupy D.C. protest. But one thing led to another, and the Grammy-nominated crooner got himself arrested, along with West and 17 other protesters, for refusing to leave the steps of the Supreme Court. They were released a day later - no fines, no charges.** So, jail - how was it? Not bad, the Prince George's.....it? Not bad, the Prince George's native told our colleague Aaron Leitko. "It was pretty by-the-book, standard procedure," he said over the phone. The police were "pretty pleasant," for the most part. "They had one bad apple out of the crew." Oooh, what did he do? Well, nothing, really. "It was just his

Rally/Demonstration

Strike

Blocking Action

Establishing a Camp (setting up tents) (Note: text indicates when a camp started)

Moving a camp to a new location

Disrupting an on-going event

Divestment Action (moving \$\$ from banks to credit unions)

Voluntary Dissolution of a camp

Strategic (pre-planned) Violence

Strategic (pre-planned) Sabotage

Building occupation

Re-establishing a Camp (that had previously been evicted)

7. Which of the following information does the text give about the event's setting, time, or attendees?

The number and/or kinds of people who attended

Location

Date or day of week

Time of day

No information given

Police presence and activities

Injuries/Arrests at the event

1. Does the bold text describe any of the following?

Arrests of protesters or others

Injuries to protesters

Injuries to police

None of this information is given

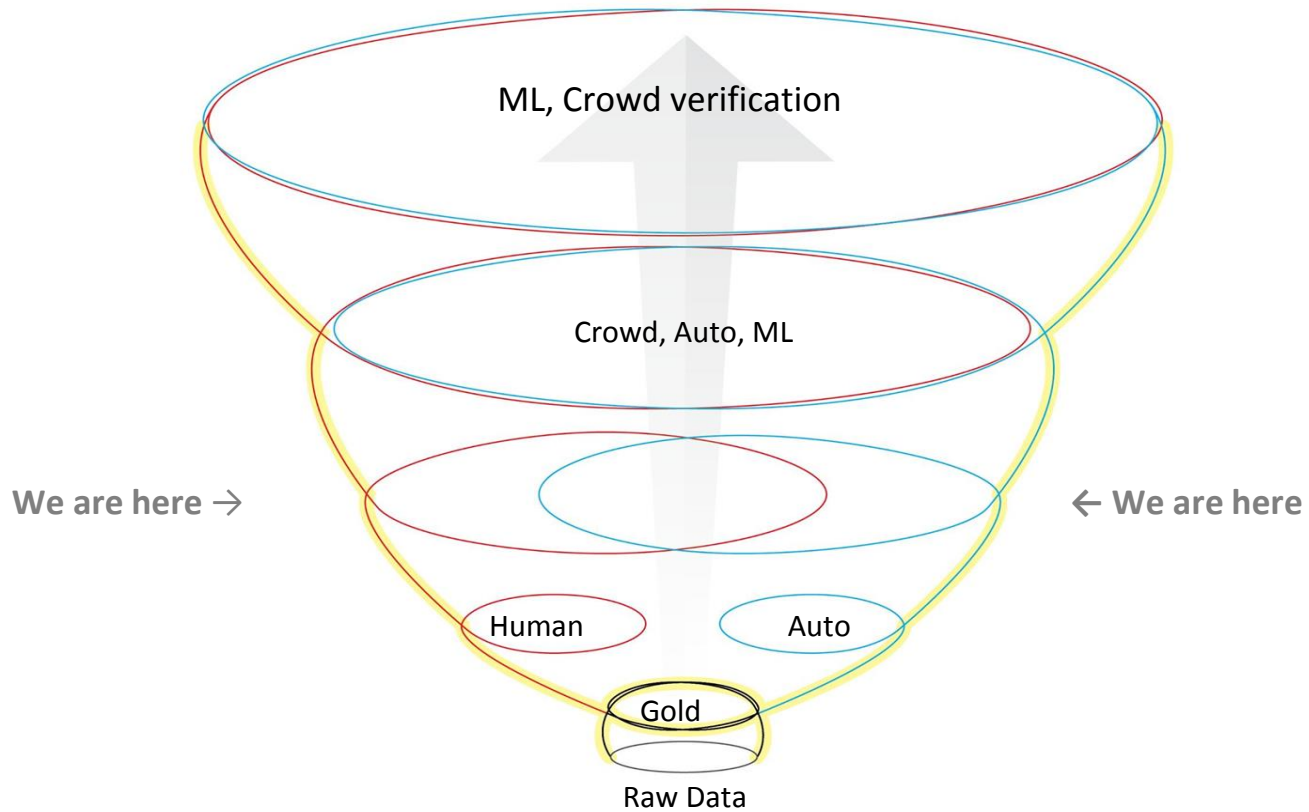
3. Which of the following information does the text give about arrests that occurred during the event?

The exact number

An approximate number

Information not given

Answer "reading comprehension" style questions... AND highlight the text that justifies your answer.



A Hybrid Approach to Scaling Up Social Data Analysis

Organizing Digital Textual Archives

for the coming wave of
computational text analysis
scholarship

CapitolQuery project

CapitolQuery

- Combine separate files appearing across multiple pages of a digital archive website
- Find structure in the documents using computational text analysis techniques
- Organize everything into a query-able database that links to other relevant databases
- *We're writing tutorials so you can learn to do all of this!*



Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Conversation 2: Digital News and its Scholarly Uses

Discussion and Questions

eDesiderata®
FORUM