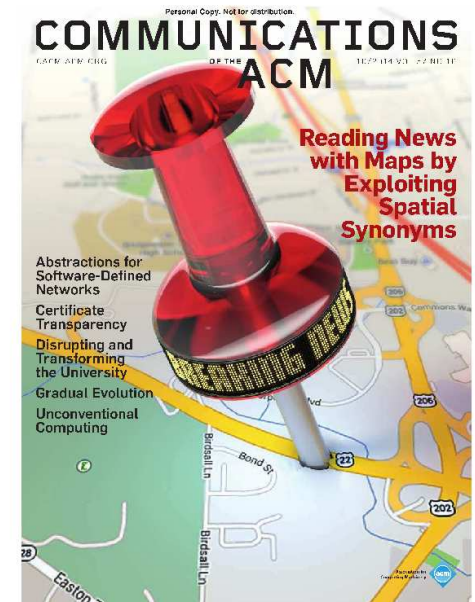


Reading News with Maps by Exploiting Spatial Synonyms

Hanan Samet*

`hjs@cs.umd.edu`

Department of Computer Science
Institute for Advanced Computer Studies
Center for Automation Research
University of Maryland
College Park, MD 20742, USA



Video

*Based on Joint Work with Marco D. Adelfio, Brendan C. Fruin, Jack Lotkowski, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Jon Sperling, and Ben E. Teitler

Application

■ Questions

1. Do you travel?
2. Do you want to know what is going on in the town you are traveling to?
3. Do you want to keep up with the latest news in the town you have left
 - Especially when it is your own hometown?
 - E.g., keep up with the local sports team

■ Answer: NewsStand

- Enables search with a map query interface instead of by keyword
- Advantage: a map, coupled with ability to vary the zoom level at which it is viewed, provides a granularity to the search and facilitates an approximate search
 - Can do an approximate search with a group of keywords (e.g., synonyms) in the query formulation
 - But users often have no clue as to which keyword to use
 - Would welcome the search automatically taking them into account
- Map query interface is a step in this direction
 - Pointing at a location and making the interpretation of the precision of this positioning dependent on the zoom level is equivalent to using spatial synonyms

Power of Spatial Synonyms

- Enables search for data when not exactly sure of what we are seeking, or what should be the answer to the query
- Ex: Seek a “Rock Concert in Manhattan”
 - “Rock Concerts” in “Harlem”, “New York City”, or “Brooklyn” are good answers when no such events can be found in “Manhattan” as they correspond to spatial synonyms:
 - “Harlem” by virtue of being contained in Manhattan
 - “New York City” by virtue of containing Manhattan
 - “Brooklyn” by virtue of proximity and a sibling relationship (neighboring borough)

Conventional Search Engines and Spatial Synonyms

- Use page rank method and good at finding documents containing keywords that we are looking for, but cannot be easily modified to handle spatial proximity query
- Primary utility is based on popularity in sense of ensuring that web pages in response are ordered by a measure incorporating their frequency of being linked to so results are same as provided to other users
 - “Democratization of search”
 - All users are treated equally
 - They all get the same bad (or good!) answers
- Effectively means that if nobody ever looked for some data before or linked to it, then it will never be found and, hence, never presented to users
- In case of synonyms, if no links to similar pages on account of being equivalent but for the use of the same words, then similarity will never be found by the search engine as the web crawler will never be able to find the similar pages when building the index to the web pages

Taking Advantage of Spatial Synonyms: Location Specification


- Explicit via geometry (latitude-longitude pairs of numbers)
 - Used in programs and calculations
 - Not in search engines or mobile devices
 - Users don't know them in this way or used to communicate in this way
- Accustomed to textual specification
 - Easy to communicate on smartphone devices with soft keyboard
 - Can capture verbally by speech recognition (e.g., Siri)
 - Behave like a polymorphic type
 - One size fits all
 - "Los Angeles" can be interpreted as a point or an area and user need not be concerned about it
 - Supports use of spatial synonyms
 - Drawback is ambiguity
 - Is "London" reference a person or a location? (toponym recognition)
 - If "London" is a location, which of many? (toponym resolution)

Decoding or Forming an E-mail Address



Emailing in London is like emailing in London.

Use your phone abroad like you do in Kentucky with AT&T's international data packages.


Rethink Possible® 

Determining Performance of a Team in a Sports League



Checking scores in Dublin is like checking scores in Dublin.

Use your phone abroad like you do in California with AT&T's international data packages.


Rethink Possible® 

Interpreting Weather Temperature Measurement Unit



Getting the weather in Mexico is like getting the weather in Mexico.

Use your phone abroad like you do in New York with AT&T's international data packages.


Rethink Possible® 

Finding Local Food



Finding restaurants in China is like finding restaurants in China.

Use your phone abroad like you do in Texas with AT&T's international data packages.

Rethink Possible® 

33 Different Plymouths

(The Numbers)



Goal: Change News Reading Paradigm

- Use map to read news for all media (e.g., text, photos, tweets, videos)
- Choose place of interest and find topics/articles relevant to it
- Topics/articles determined by location and level of zoom
- No predetermined boundaries on sources of articles
- Application: monitoring hot spots
 1. Investors
 2. National security
 3. Disease monitoring
- One-stop shopping for spatially-oriented news reading
 1. Summarize the news
 - What are the top stories happening?
 2. Explore the news
 - What is happening in Darfur?
 3. Discover patterns in the news
 - How are the Olympics and Darfur related?
- Overall goal: make map medium for presenting all spatially-referenced information

Mapping the News

1. Cluster articles on same topic using TF-IDF and associate clusters with the mentioned locations
 - Same cluster can be associated with many locations
2. As zoom-in, the cluster populations will be smaller as fewer articles refer to the viewing window
 - Location plays a larger role in the clustering algorithm
 - Geotagging errors are less likely to be filtered out
3. Cluster rank vs: cluster spread
 - Don't want to have empty areas on the map with no articles implying that less important articles are displayed with some regions than others and some important articles are not displayed unless zoom-in
 - As zoom-in and pan want to make sure that once an article has been displayed, it persists until its location is no longer in the viewing window
4. Zoom-In and Pan are expensive as much redrawing
 - Use "Home", "Local (L)", and "World (W)" as navigation shortcuts
 - Can use an inset "overview window" to control zoom and pan with little symbolic information that needs to be redrawn

Existing News Readers

1. Popular news aggregators such as Google News, Yahoo! News, and Microsoft Bing News have only a rudimentary understanding of the implicit geographic content of news articles, usually based on the address of the publishing news source (e.g., newspaper)
2. Usually a linear presentation format
 - Articles grouped by keyword or topic, rather than by geography
3. Ex: Google News Reader
 - Classifies articles by topic
 - Local news search
 - Aggregates articles by zip code or city, state specification
 - E.g., articles mentioning “College Park, MD”
 - Provides a limited number of articles (about 20 at the moment)
 - Seems to be based on the host of the articles
 - E.g., “LA Times” provides local articles for “Los Angeles, CA”
 - Seems to use Google Search with location names as search keys
 - E.g., articles for ZIP 20742 are those mentioning “College Park, MD” or “University of Maryland”
 - Has no notion of story importance in the grand scheme
 - International versions use international news sources

NewsStand: Spatio-Textual Aggregation of News and Display

1. Crawls the web looking for news sources and feeds
 - Indexing 10,000 news sources
 - About 50,000 news articles per day
2. Aggregate news articles by both content similarity and location
 - Articles about the same event are grouped into clusters
3. Rank clusters by importance which is based on:
 - Number of articles in cluster
 - Number of unique newspapers in cluster
 - Event's rate of propagation to other newspapers
4. Associate each cluster with its geographic focus or foci
5. Display each cluster at the positions of the geographic foci
6. Other options:
 - Category (e.g., General, Business, SciTech, Entertainment, Health, Sports)
 - Image and video galleries
 - Map stories by disease, brands, people, etc.
 - User-generated news (e.g., Social networks such as Twitter)

NewsStand: Map Mode

- NewsStand is at <http://newsstand.umiacs.umd.edu/>
- Query: What is happening at location Y?

NewsStand: Top Stories Mode

The screenshot displays the NewsStand interface. On the left, a list of news stories is shown, including:

- How Obama, Bush, Clinton viewed Russia's Putin over time** (5 hours ago from usatoday.com)
- Questions, answers about college union ruling** (2 hours ago from miamiherald.com)
- How Obama, Bush, Clinton viewed Russia's Putin over time** (5 hours ago from usatoday.com)
- Renewed search for plane debris** (2 hours ago from bbc.co.uk)
- VIDEO: Why did Facebook spend \$2bn on VR?** (6 hours ago from bbc.co.uk)
- Two firefighters killed in Boston** (1 hour ago from bbc.co.uk)
- Grace period on 'Obamacare' deadline** (12 hours ago from bbc.co.uk)

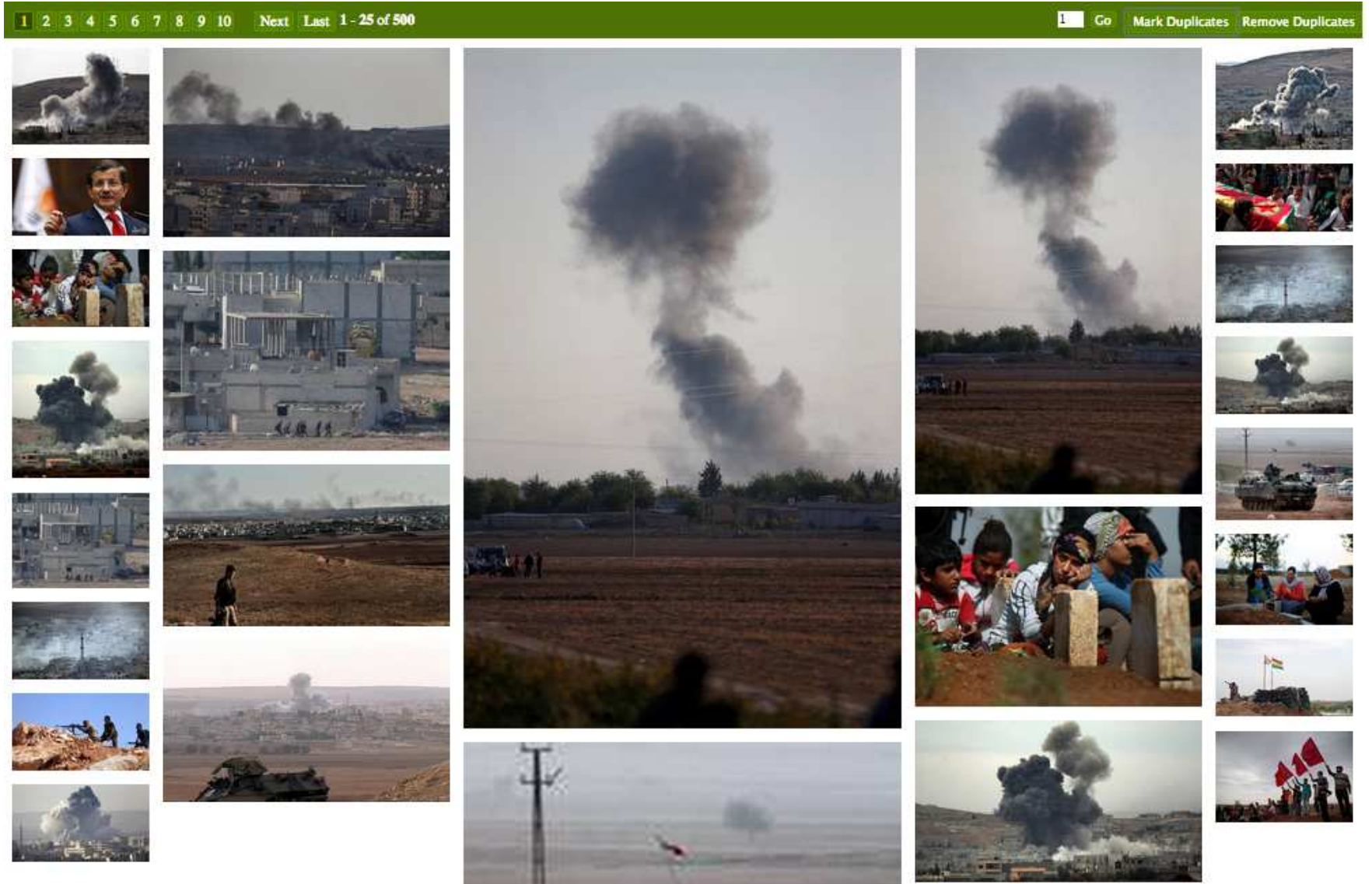
On the right, a map shows the location of **Moscow, Pennsylvania, United States**. A popup window for this location provides details:

- Category:** General
- Reference:** ... from early warning systems and notifications of missile launches at in **Moscow**, Russia on June 4, 2000. "[The] world demands that we take every ...
- Description:** Presidents Obama, George W. Bush and Bill Clinton held an optimistic view of Russian leader Vladimir Putin during their early interactions.

The interface also includes navigation options like 'Top Stories', 'Map Mode', and 'Display: 5', along with search and filter controls.

- NewsStand is at <http://newsstand.umiacs.umd.edu/>
- Query: Where is topic X occurring (spatial data mining)?

Image Gallery



Video Gallery

1 2 3 4 5 6 1 - 25 of 126 Go Show Information

Horror of Kobani: Headless corpses left in the street and victims with their eyes cut out, the savagery of Isis laid bare 2:02

CNN: Australia joins ISIS fight 1:55

NewsBlaze: Countdown From ISIS to IS-Less - Obama Says We Underestimated Strength 6:58

CNN: Paul Ryan open to boots on the ground against ISIS 5:22

Truthdig: Top 7 Surprising Reasons Turkey Is Entering War on Islamic State 1:14

CNN: The great American freakout 5:32

Centre Daily Times: Activists: Kurds halt jihadi advance in Syria town 0:31

Centre Daily Times: Despite fierce struggle, Kurds losing ground in besieged Syrian town 1:30

CNN: Coalition airstrikes fail to halt ISIS advance in Iraq, Syria 1:54

Turkey says it wont launch ground action alone against ISIS - CNN 1:58

CNN: Civilian casualty rules dont apply 1:49

NewsBlaze: Countdown From ISIS to IS-Less - Obama Says We Underestimated Strength 0:42

CNN: As coalition strikes by air, Kurds fight ISIS in Syria 2:22

Hot Air: Panetta memoir blames Obama for collapse in Iraq 2:26

CNN: Why was ISIS threat misjudged? 2:14

Centre Daily Times: US-led coalition ramps up strikes on Syrian town 1:24

Truthdig: The Last Days of Kobane Loom as IS Closes in on Syrian Kurds with Murder on its Mind 2:41

CNN: ISIS and Kurds battle for Syrian city of Kobani 2:12

Centre Daily Times: Turkey: Syrian town about to fall to jihadists 1:18

Hot Air: State's Marie Harf: ISIS couldn't even have predicted how powerful ISIS would become 0:51

CNN: Australia joins ISIS fight 1:22

CNN: Paul Ryan open to boots on the ground against ISIS 1:51

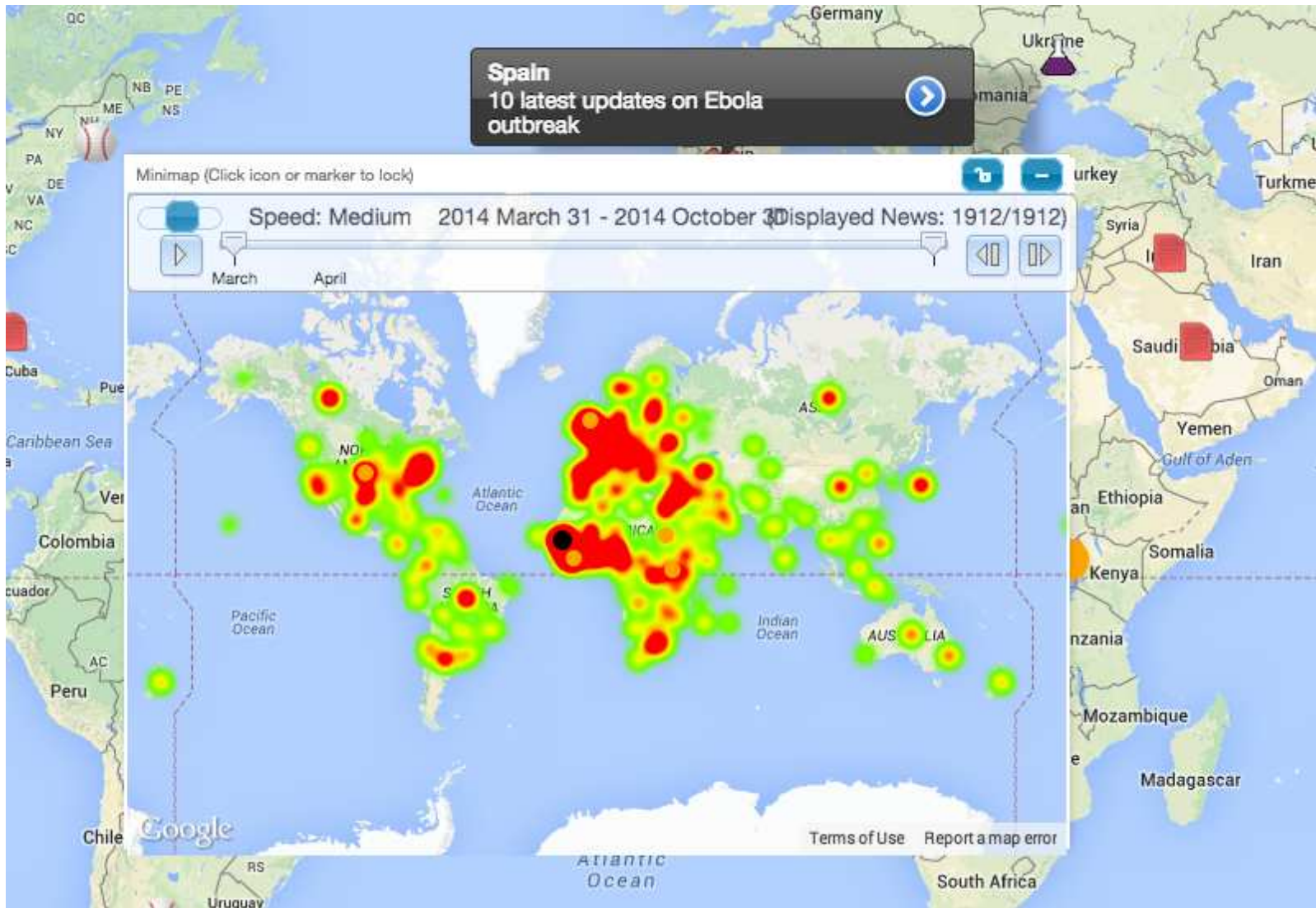
CNN: Opinion: Exploit ISIS biggest fear 1:06

CNN: ISIS closes in on Kurdish town in Syria; Turkey debates sending troops 2:06

Static Disease Tracking Application

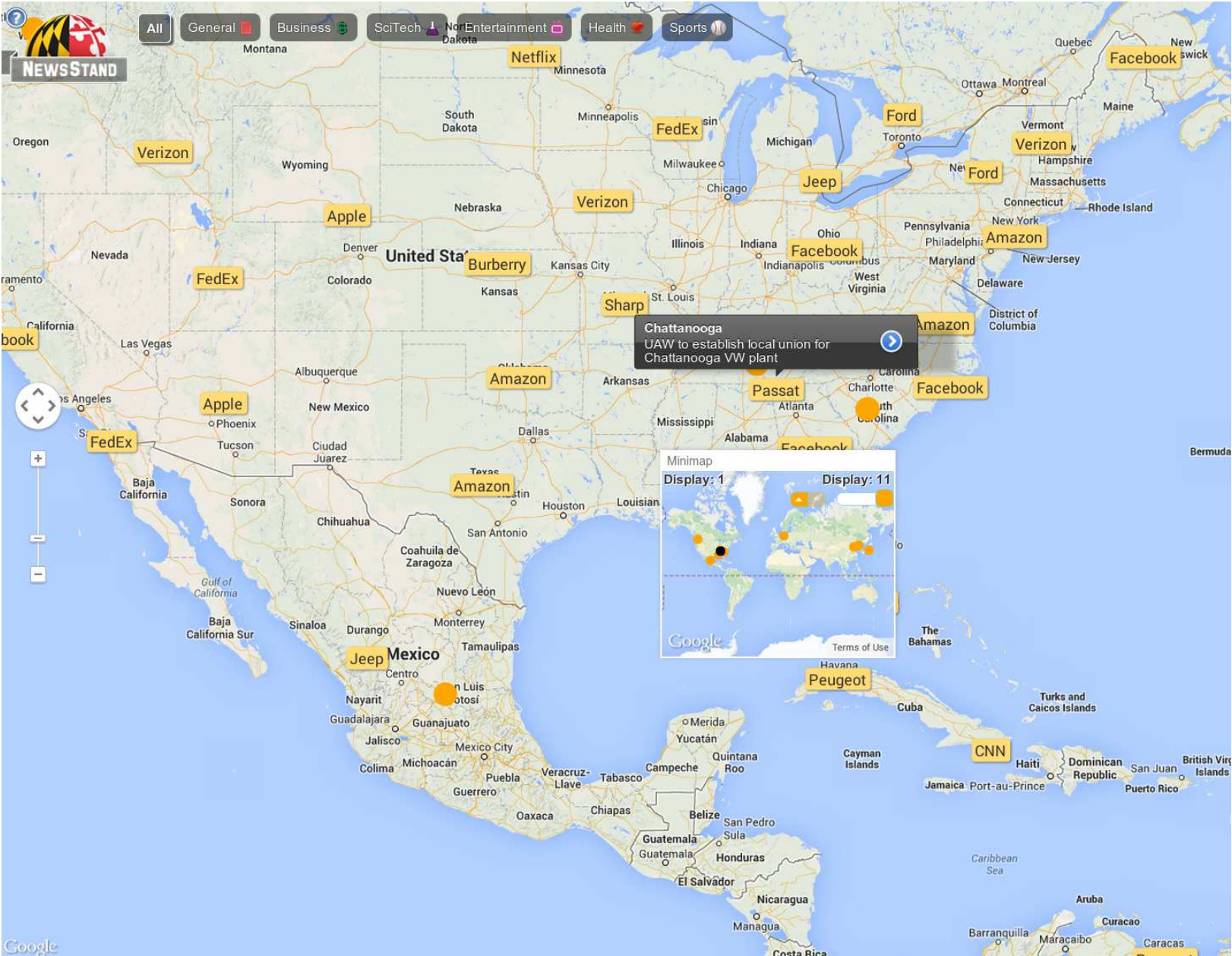


Dynamic Disease Tracking Application: Time Mode



- E.g., track ebola

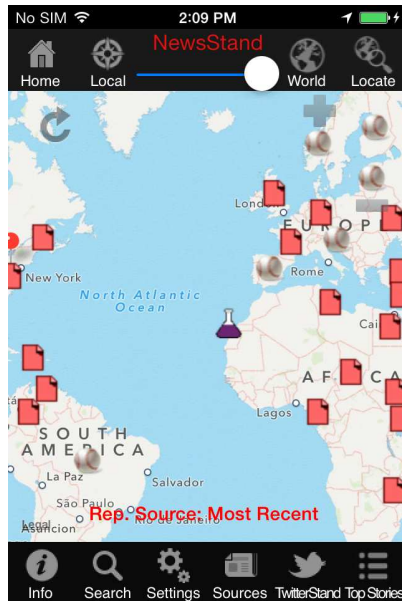
Brand Remediation



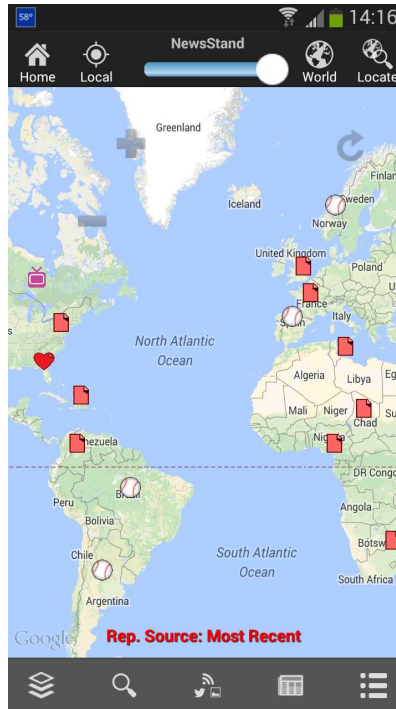
People on Map



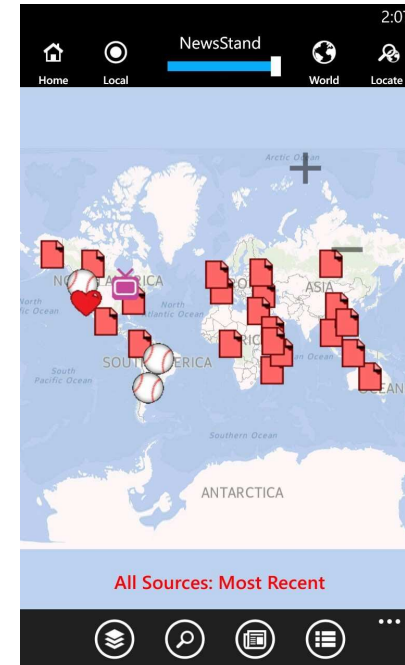
Port to Mobile Platforms (Apps)



(IOS)

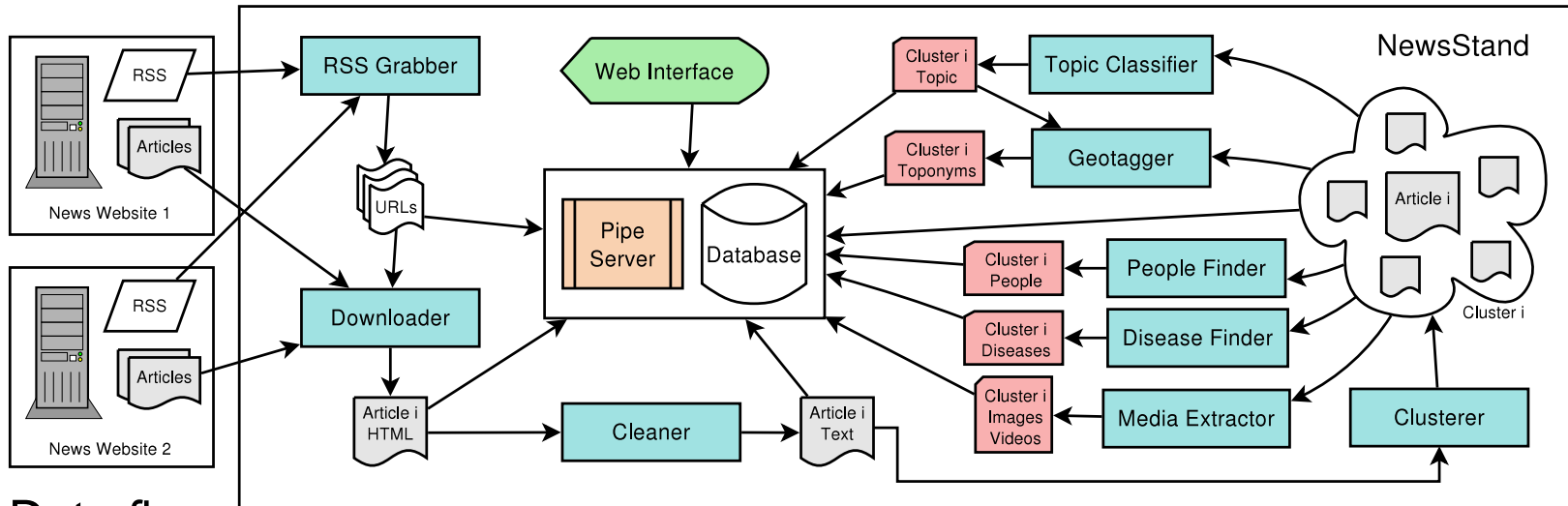


(Android)



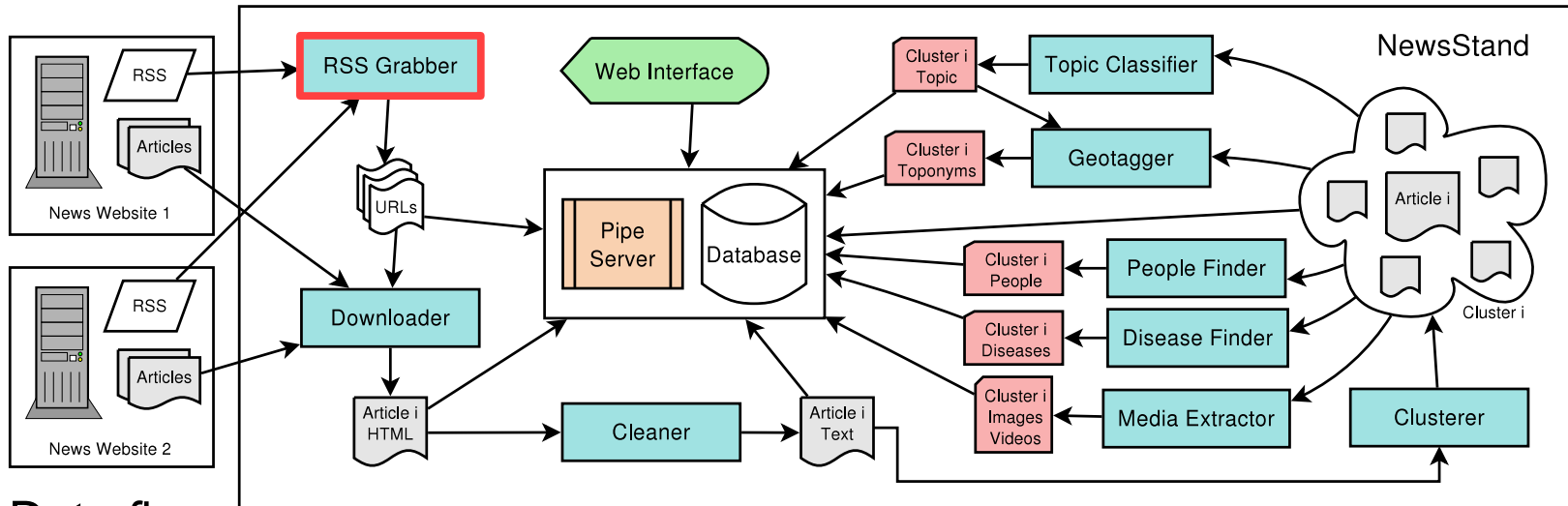
(Windows)

NewsStand's Architecture



Data flow:

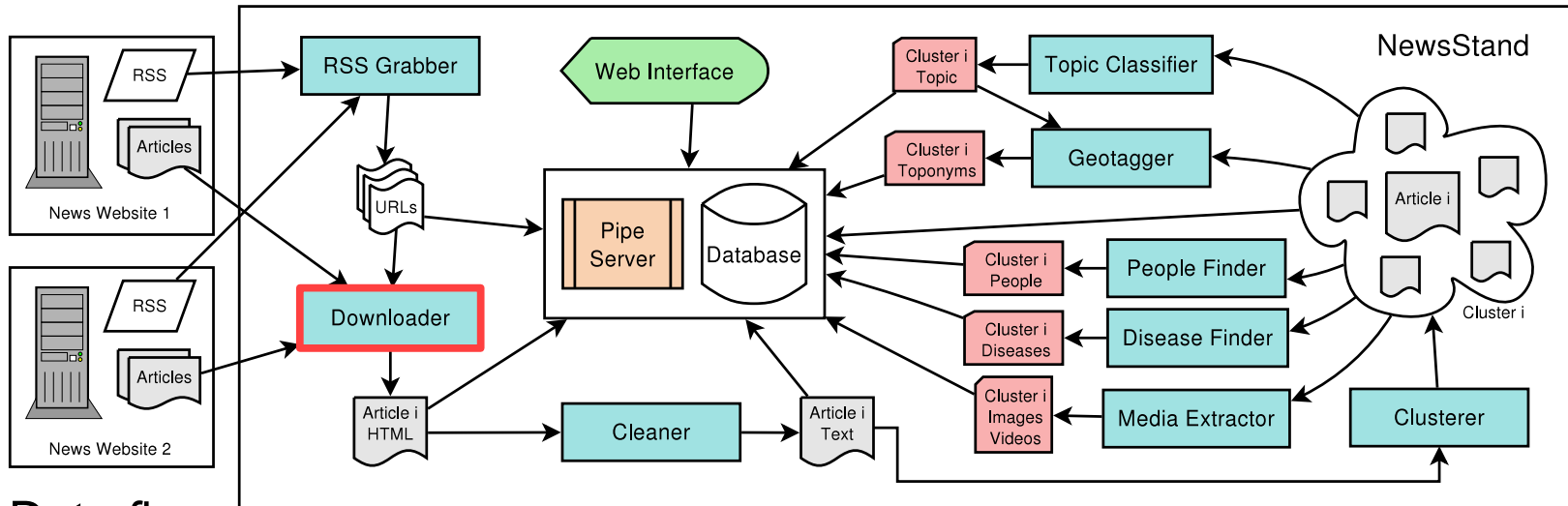
NewsStand's Architecture



Data flow:

1. **RSS Grabber:** Polls RSS feeds and retrieves URLs to news articles.

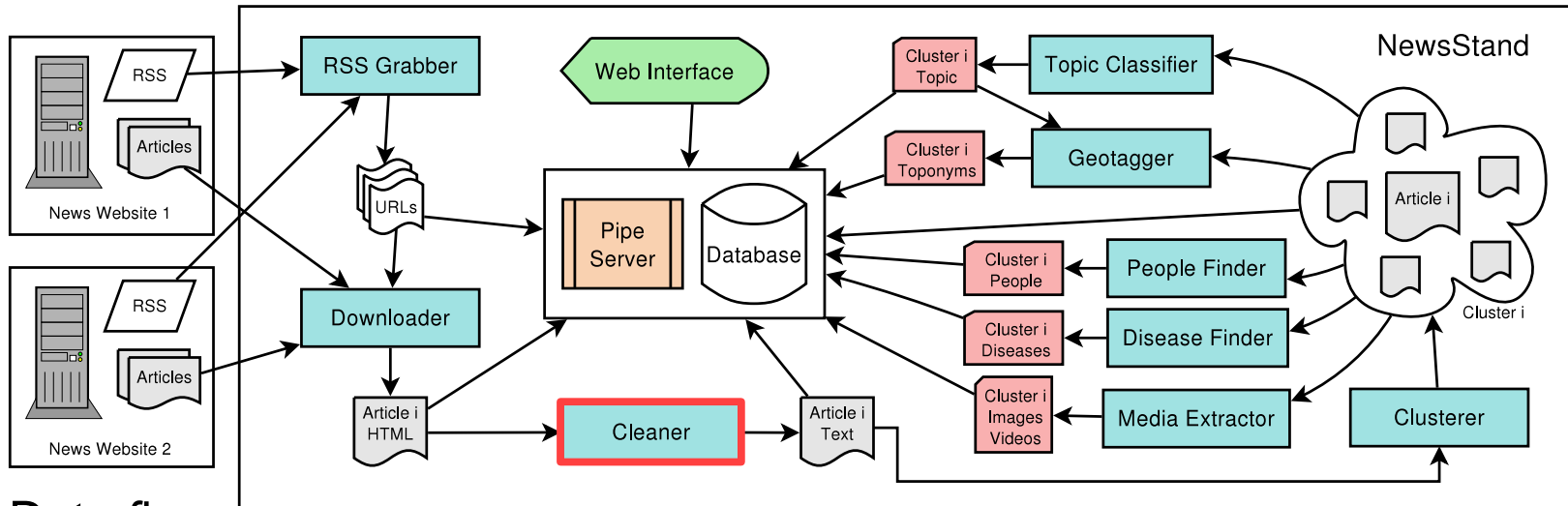
NewsStand's Architecture



Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.

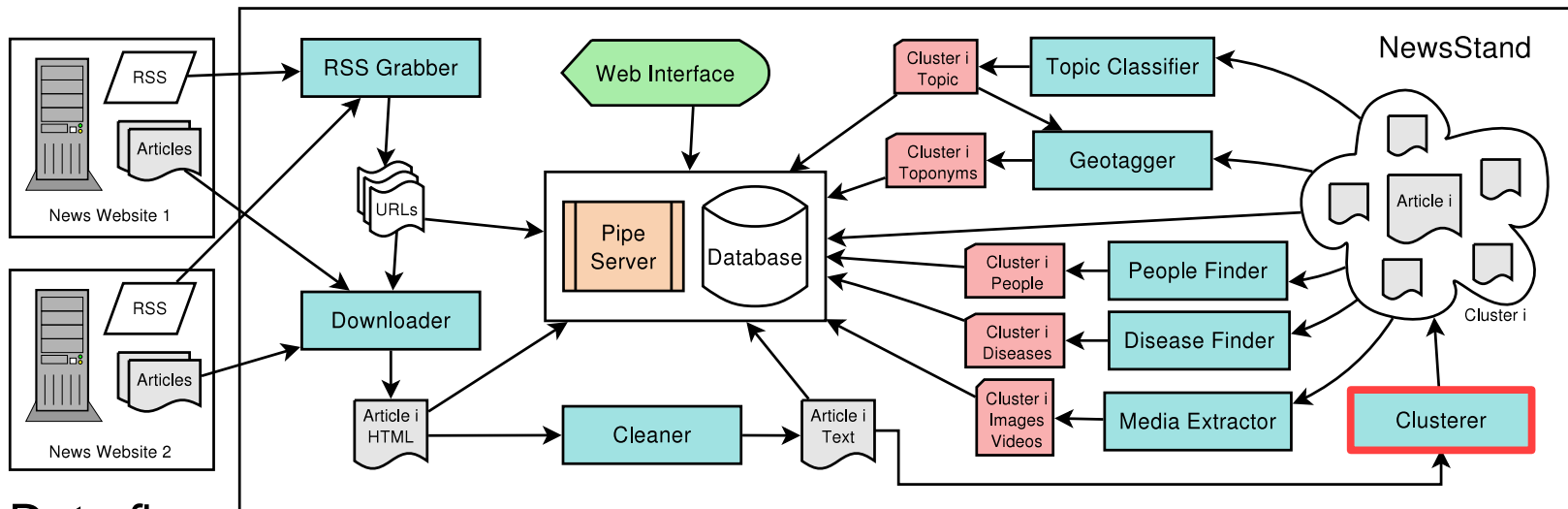
NewsStand's Architecture



Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.

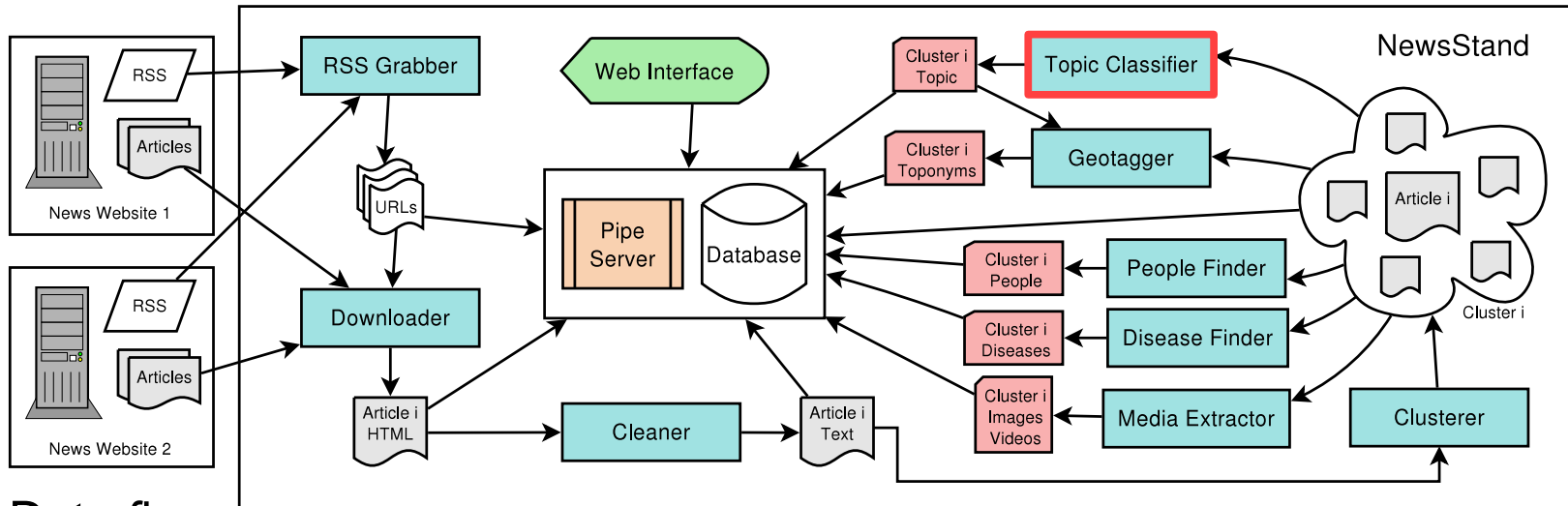
NewsStand's Architecture



Data flow:

1. **RSS Grabber:** Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader:** Downloads news articles from URLs.
3. **Cleaner:** Extracts article content from source HTML.
4. **Clusterer:** Groups together articles about the same story.

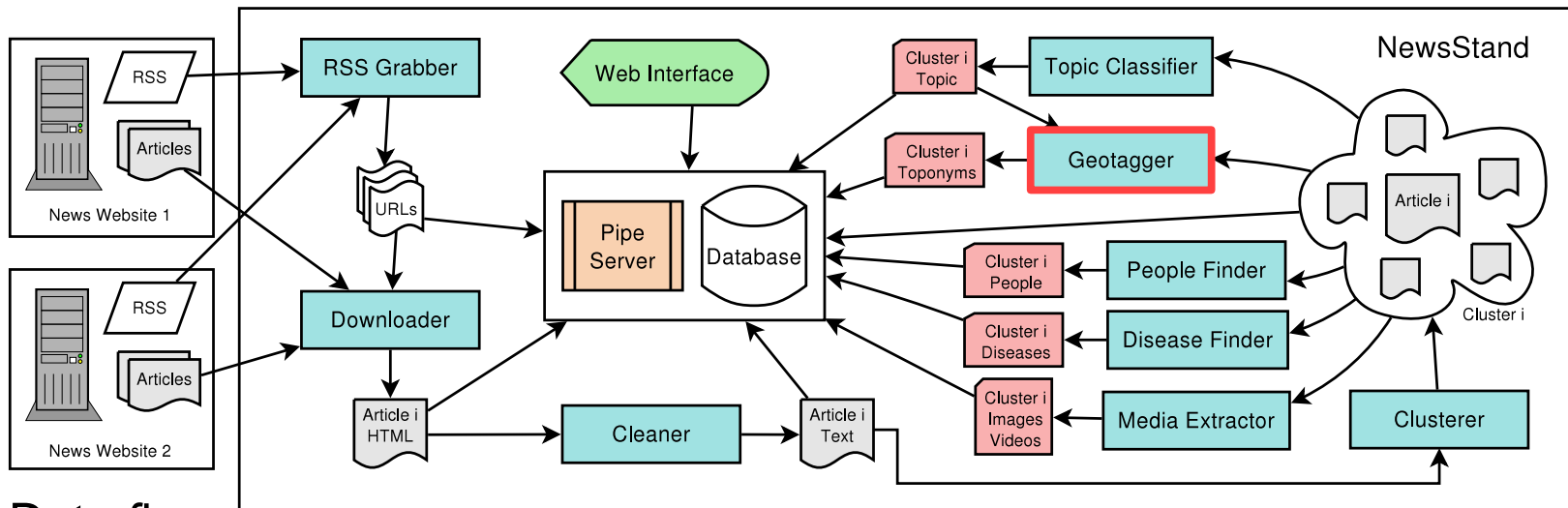
NewsStand's Architecture



Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., "Sports").

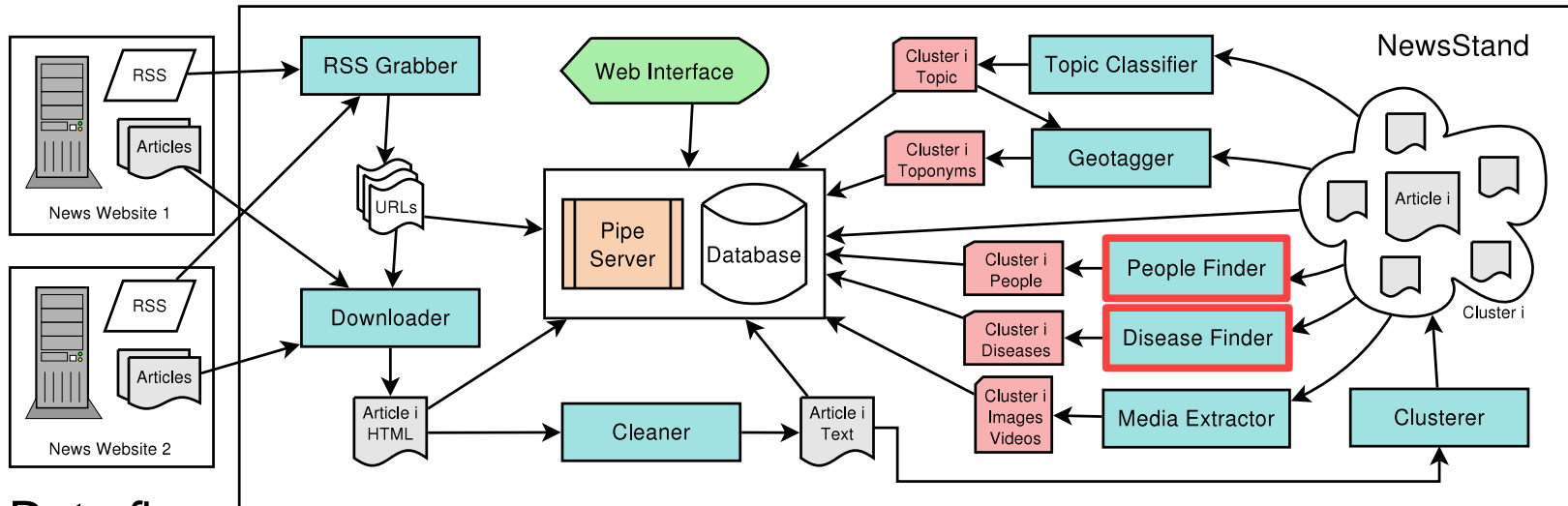
NewsStand's Architecture



Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., "Sports").
6. **Geotagger**: Finds toponyms and assigns lat/long values to each.

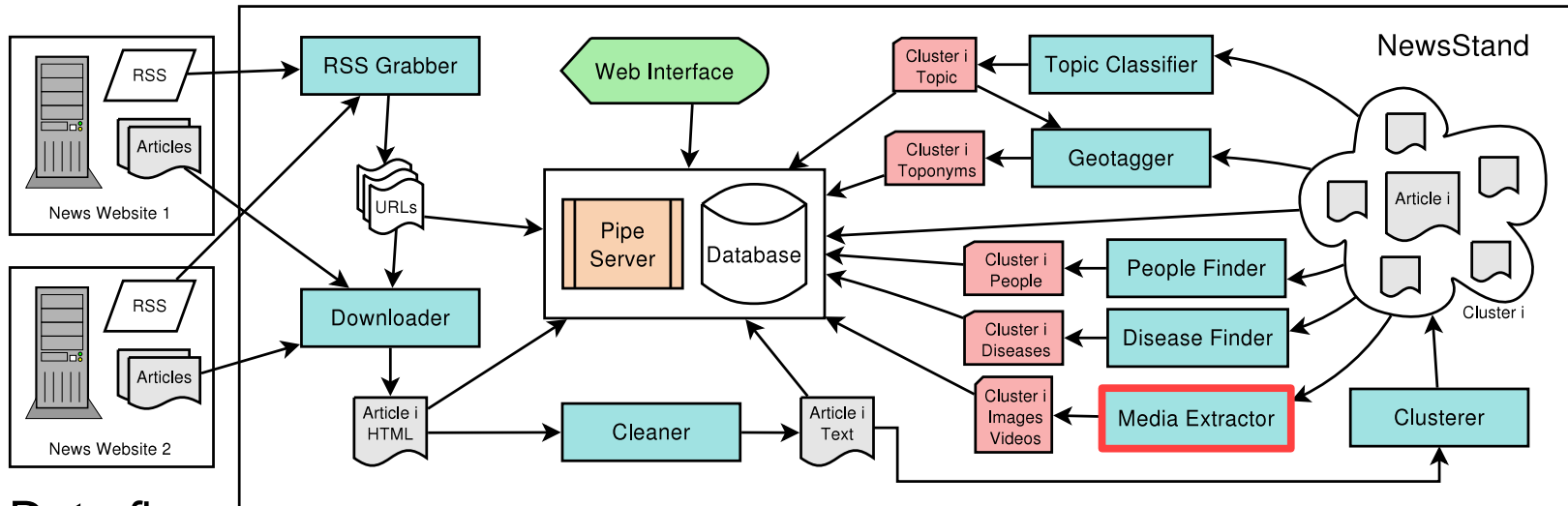
NewsStand's Architecture



Data flow:

1. **RSS Grabber:** Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader:** Downloads news articles from URLs.
3. **Cleaner:** Extracts article content from source HTML.
4. **Clusterer:** Groups together articles about the same story.
5. **Topic Classifier:** Assigns general topics to articles (e.g., "Sports").
6. **Geotagger:** Finds toponyms and assigns lat/long values to each.
7. **People/Disease Finder:** Finds mentions of people/diseases.

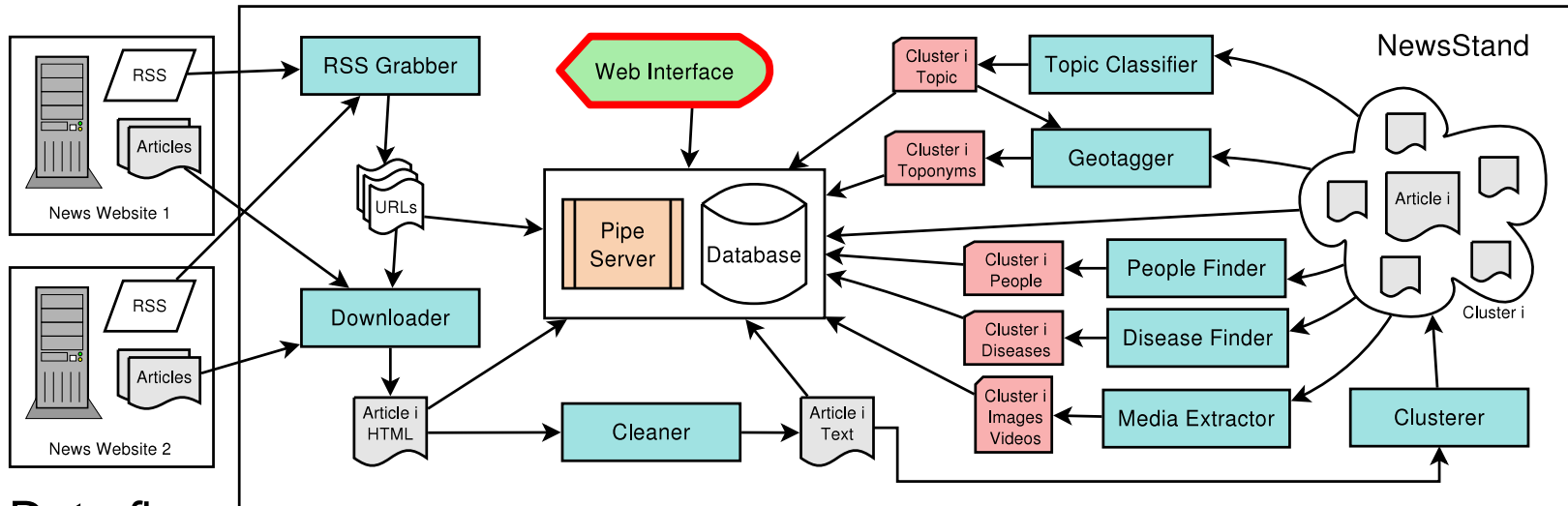
NewsStand's Architecture



Data flow:

1. **RSS Grabber:** Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader:** Downloads news articles from URLs.
3. **Cleaner:** Extracts article content from source HTML.
4. **Clusterer:** Groups together articles about the same story.
5. **Topic Classifier:** Assigns general topics to articles (e.g., “Sports”).
6. **Geotagger:** Finds toponyms and assigns lat/long values to each.
7. **People/Disease Finder:** Finds mentions of people/diseases.
8. **Media Extractor:** Extracts captioned images and videos.

NewsStand's Architecture



Data flow:

1. **RSS Grabber:** Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader:** Downloads news articles from URLs.
3. **Cleaner:** Extracts article content from source HTML.
4. **Clusterer:** Groups together articles about the same story.
5. **Topic Classifier:** Assigns general topics to articles (e.g., "Sports").
6. **Geotagger:** Finds toponyms and assigns lat/long values to each.
7. **People/Disease Finder:** Finds mentions of people/diseases.
8. **Media Extractor:** Extracts captioned images and videos.
9. **Web Interface:** Accesses database to retrieve data for display.

Map Query Interface Requires a Geotagger

- Geotagger: processor that converts a textual specification of a location to a geometric one (i.e., latitude-longitude pair)
- Geotagging issues:
 1. Toponym recognition: identify geographical references in text
 - Does “Jefferson” refer to a person or a geographical location?
 - Known as Geo/Non-Geo Ambiguity
 2. Toponym resolution: disambiguate a geographical reference
 - Does “London” mean “London, UK”, “London, Ontario”, or one of 2570 other instances of “London” in our gazetteer?
 - Known as Geo/Geo Ambiguity
 3. Determine spatial focus of a document
 - Is “Singapore” relevant to a news article about “Hurricane Katrina”?
 - Not so, if article appeared in “Singapore Strait Times”

Geo/Non-Geo Ambiguity Example: Obama



Japan's Obama town overjoyed

Wed, Nov 5 2008

By Toshi Maeda

OBAMA, Japan (Reuters) - The sleepy Japanese fishing town of Obama went wild Wednesday as locals gathered to celebrate namesake Barack Obama's victory in the U.S. presidential election.

More than a hundred residents gathered to watch the vote count on television in a public hall in the middle of the day, and chanted "Obama, Obama!" as the result was announced on a news program.

Some were clad in hula costumes in honor of Obama's birthplace in Hawaii. Others showed up wearing "I love Obama" T-shirts.

The town has taken advantage of the name -- one of many named Obama, or "small beach" in Japanese -- to launch products from fish burgers and steamed cakes to chopsticks.

Buoyed by the victory, locals say they hope Obama, who once mentioned the town in a television interview, will visit.

"The next thing we want to do is to go to the White House and dance the hula at Obama's inauguration ceremony," said Tatsuya Sano, 45, who runs a souvenir shop selling locally made Barack Obama souvenirs.

Chikako Shimizu, 35, the leader of an "Obama Girls" hula dance group launched this year, said she was calm while watching the vote count on television because she had no doubt Obama would win.

"I was convinced that he would win. I couldn't be happier," she said.

Obama City residents plan to dance and party more in the evening.



Geo/Non-Geo Ambiguity Example: Batman



Mayor of Batman sues WB, Nolan

Southeastern city in Turkey fights for name

By ALI JAAFAR

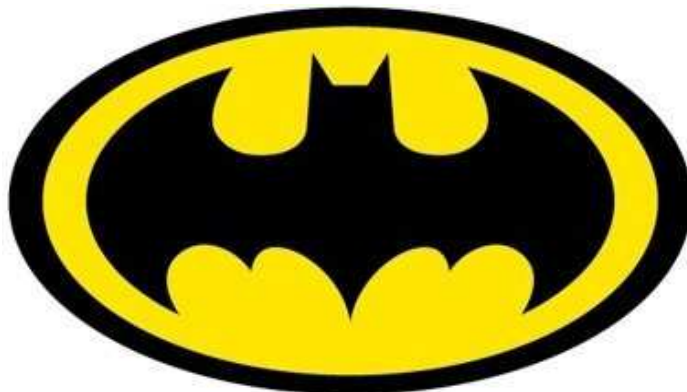
Batman has a new adversary: Batman.

The mayor of an oil-producing city in southeastern Turkey, which has the same name as the Caped Crusader, is suing helmer Christopher Nolan and Warner Bros. for royalties from mega-grosser "The Dark Knight."

Huseyin Kalkan, the pro-Kurdish Democratic Society Party mayor of Batman, has accused "The Dark Knight" producers of using the city's name without permission.

"There is only one Batman in the world," Kalkan said. "The American producers used the name of our city without informing us."

No one from the town of Batman has explained why it took so many years to take legal action. Batman first appeared as a comicbook character in 1939 and the "Batman" TV series started in 1966. Tim Burton's first bigscreen rendition for Warner Bros. came out in 1989. Undoubtedly the fact that "Dark Knight" is about to pass the \$1 billion mark at the B.O. played a part in stirring the ire of the Turkish hamlet.



Geo/Geo Ambiguity Example: Java, Georgia



Georgia accuses Russia of seeking to take over South Ossetia



Russia thrusts into South Ossetia; clashes with Georgia reported

5 hours ago

JAVA, Georgia (AFP) — Russian tanks and troops surged into Georgia's breakaway South Ossetia province on Friday to repel a Georgian offensive to reclaim the region amid fighting said to have left hundreds dead.

"Fierce clashes" between Russian and Georgian troops in the southern suburbs of South Ossetia's capital Tskhinvali were reported by Russian news agencies as night fell on the city.

Moscow had vowed retaliation to defend Russians in Tskhinvali who had come under fire by the Georgian artillery and air assault — the worst fighting since the 1992-94 separatist war in the region.

"Georgian forces are controlling the entire territory of South Ossetia except Java," a city north of Tskhinvali, Georgian President Mikheil Saakashvili said in a televised address.

"We are fully controlling Tskhinvali," he added, although the rebels shortly after said that they were in control, according to the Interfax news agency.

Geo/Geo Ambiguity: Vancouver



Oops, wrong Vancouver

Thu, Feb 4 2010

By Teresa Carson

VANCOUVER, Washington (Reuters) - Sallie Reavey picked up the phone at her charming Briar Rose Inn and the caller asked about rooms in mid-February. "We have a nice selection of rooms for those dates," she replied, to which the caller gasped: "You still have rooms during the Olympics?"

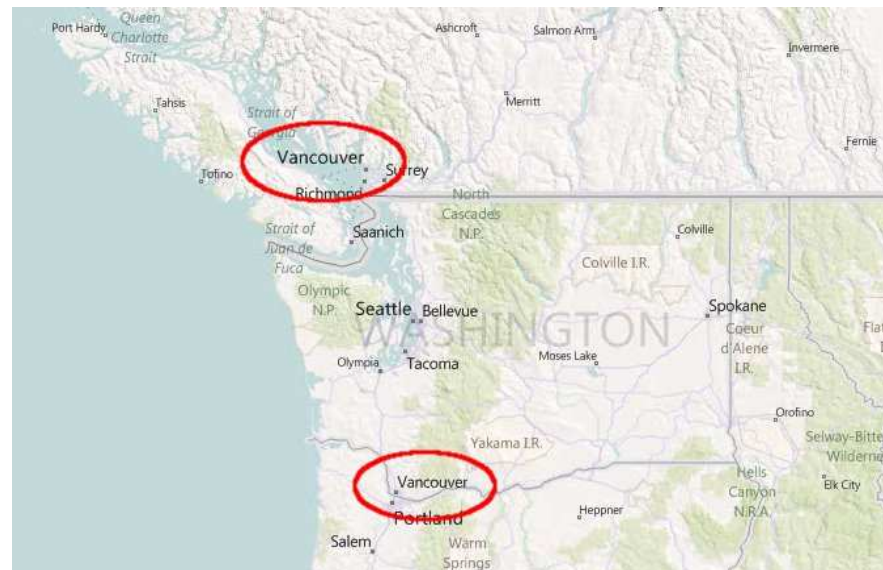
Reavey had to tell him: wrong Vancouver.

The Briar Rose is in Vancouver, Washington, not Vancouver, British Columbia, the Canadian city that will host the 2010 Winter Olympics starting on February 12.

"America's Vancouver," as a former town mayor liked to describe it, sits 250 miles south of the Olympic host Vancouver and has a population of some 165,000 people -- far fewer than the Canadian city.

The Hilton Vancouver Washington has also fielded Olympic enquiries and trained its reservations staff to be sensitive to the possible mistake and, naturally, turn it into a marketing opportunity.

"We absolutely want them to come here," Gerry Link, the hotel's general manager said, adding of the Vancouver mix-ups: "So far it has all been pretty good-natured."

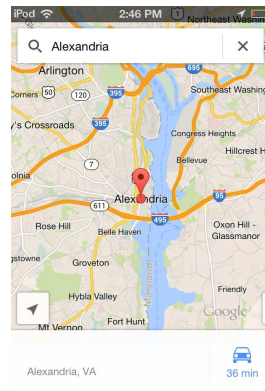


Geotagging

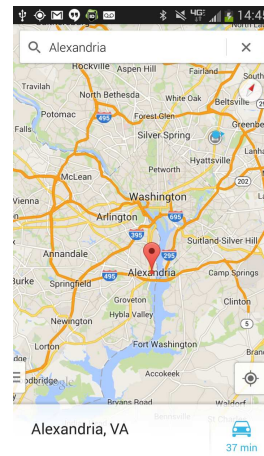
- Geotagging: Understanding textual references to spatial data
 1. Identifying or recognizing
 2. Classifying (is “Michigan” a state or a lake?)
 3. Disambiguating or resolving
 4. Localizing (geocoding to GPS coordinates)
- Context of textual references
 1. Queries - use prior queries and location
 - Ex: Query “Alexandria” when in “College Park, MD”
 2. Underlying data being queried - need context



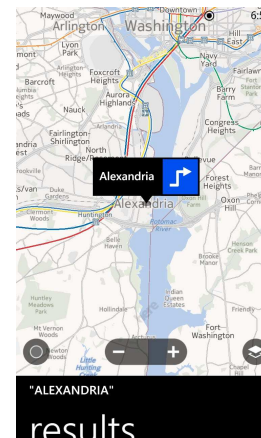
Apple iOS5
Maps by
Google



iOS Maps by
Google



Android
Maps by
Google



Here Maps
on Windows
Phone



Apple iOS6
and iOS7
Maps

Mechanics of Geotagging

1. Goal: high recall in toponym recognition (i.e., not missing toponyms) at expense of precision
 - Rectify by subsequent use of toponym resolution which can (and will) also be used to filter erroneous location interpretations
2. Toponym recognition: 2 stages
 - Finding toponyms
 - Filtering toponyms: postprocessing to remove errors in recognition
3. Toponym resolution
 - Use local lexicons containing locations that can be specified without all of their containers (derived from articles from a particular news source) to determine spatial reader scopes for particular sources
 - E.g., "Dublin" implies "Dublin, Ohio" for readers of a news source in "Columbus, Ohio"
 - Use Wikipedia articles to find concepts related to particular locations so that the presence of these concepts in conjunction with an ambiguous reference to a location can be properly resolved
 - E.g., mention of "White House" in conjunction with "Washington" to provide evidence for resolving as "Washington, D.C."

Local Lexicon Example



Finding Toponyms

1. Use entity tables of well-known locations (e.g., names of continents, countries, etc.), abbreviations (e.g., "CA", "FL", etc.), and demonyms (words used to refer to people from particular places such as "German")
2. Use entity dictionaries containing names of entities that appear frequently in news thereby precluding their interpretation as toponyms (e.g., "Apple")
3. Use a Part of Speech (POS) tagger to find proper noun phrases which could denote names even with possessives like "Prince George's County"
4. Use Named Entity Recognition (NER) package which helps avoid geo/non-geo errors by making use of entity types such as name, place, organization, etc.
5. Compensate for NER errors
 - Boundary expansion (e.g., "Guinea" and "Equatorial Guinea")
 - Fragmented references such as names where parts can be interpreted as locations (e.g., "Paul Washington" and "Washington")

Running Example (1)

- Excerpt from an article in the Paris News about a local politician campaigning in Paris, Texas
- Mentions multiple places in Texas

*Democratic candidate for **Texas** Railroad Commissioner Jeff **Weems** stumped in **Paris** late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark **Homer**, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.*

Running Example (1)

- Excerpt from an article in the Paris News about a local politician campaigning in Paris, Texas
- Mentions multiple places in Texas

*Democratic candidate for **Texas** Railroad Commissioner Jeff **Weems** stumped in **Paris** late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark **Homer**, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.*

- True toponyms: **Texas, Paris, Houston, Lamar County, Austin, Dish**

Running Example (1)

- Excerpt from an article in the Paris News about a local politician campaigning in Paris, Texas
- Mentions multiple places in Texas

*Democratic candidate for **Texas** Railroad Commissioner Jeff **Weems** stumped in **Paris** late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark **Homer**, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.*

- True toponyms: Texas, Paris, Houston, Lamar County, Austin, Dish
- Potential mistakes: Weems, Homer, Friday (all in Texas)

Running Example (2)

Democratic candidate for *Texas Railroad Commissioner* Jeff Weems stumped in *Paris* late *Friday* in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the *Railroad Commission* who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for *Democrats* winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the *Midcontinent Express Pipeline* compressor station south of *Paris*.

Running Example (2)

Democratic candidate for *Texas Railroad Commissioner* Jeff Weems stumped in *Paris* late *Friday* in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, *D-Paris*, said it will be refreshing to have someone on the *Railroad Commission* who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for *Democrats* winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the *Midcontinent Express Pipeline* compressor station south of *Paris*.

1. *Initial text*
2. *Entity tables*: [*LOC Texas*], [*PER Jeff Weems*], [*DAY Friday*], [*PER Mark Homer*]

Running Example (2)

Democratic candidate for *Texas Railroad Commissioner* Jeff Weems stumped in *Paris* late *Friday* in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the *Railroad Commission* who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for *Democrats* winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the *Midcontinent Express Pipeline* compressor station south of *Paris*.

1. *Initial text*
2. **Entity tables:** [_{LOC} *Texas*], [_{PER} Jeff Weems], [_{DAY} *Friday*], [_{PER} Mark Homer]
3. **Cue words:** Rep. [_{PER} *Mark Homer*], D-[_{LOC} *Paris*], [_{LOC} Lamar County]

Running Example (2)

Democratic candidate for *Texas Railroad Commissioner* Jeff Weems stumped in *Paris* late *Friday* in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the *Railroad Commission* who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for *Democrats* winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the *Midcontinent Express Pipeline* compressor station south of *Paris*.

1. *Initial text*
2. **Entity tables:** [*LOC Texas*], [*PER Jeff Weems*], [*DAY Friday*], [*PER Mark Homer*]
3. **Cue words:** Rep. [*PER Mark Homer*], D-[*LOC Paris*], [*LOC Lamar County*]
4. **Proper noun phrases:** [*NP Democratic*], [*NP Railroad Commissioner Jeff Weems*], [*NP Paris*], [*NP Rep. Mark Homer*], [*NP Railroad Commission*], [*NP Houston*], [*NP Weems*], [*NP Lamar County*], [*NP Democrats*], [*NP Austin*], [*NP Dish*], [*NP Texas*], [*NP Midcontinent Express Pipeline*]

Running Example (2)

Democratic candidate for *Texas Railroad Commissioner* Jeff Weems stumped in *Paris* late *Friday* in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the *Railroad Commission* who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for *Democrats* winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the *Midcontinent Express Pipeline* compressor station south of *Paris*.

1. Initial text

2. **Entity tables:** [LOC *Texas*], [PER Jeff Weems], [DAY *Friday*], [PER Mark Homer]

3. **Cue words:** Rep. [PER *Mark Homer*], D-[LOC *Paris*], [LOC Lamar County]

4. **Proper noun phrases:** [NP *Democratic*], [NP *Railroad Commissioner Jeff Weems*], [NP *Paris*], [NP *Rep. Mark Homer*], [NP *Railroad Commission*], [NP *Houston*], [NP *Weems*], [NP *Lamar County*], [NP *Democrats*], [NP *Austin*], [NP *Dish*], [NP *Texas*], [NP *Midcontinent Express Pipeline*]

5. **Named-entity recognition:**

[PER <i>Jeff Weems</i>]	0.999	[LOC <i>Houston</i>]	0.917	[LOC <i>Paris</i>]	0.997	[PER <i>Weems</i>]	0.849
[ORG <i>Railroad Commission</i>]	0.995	[LOC <i>Lamar County</i>]	0.737	[LOC <i>Austin</i>]	0.995	[LOC <i>Texas</i>]	0.557
[ORG <i>Midcont. Expr. Ppln.</i>]	0.973	[ORG <i>Democratic</i>]	0.539	[PER <i>Mark Homer</i>]	0.920		

Filtering Toponyms

1. Toponym refactoring:
 - Account for different suffixes and prefixes for same entity
 - Ex: "Fort" and "Ft", "County Kildare" and "Kildare County", "Fairfax Hi" and "Fairfax High School", etc.
2. Active verbs
 - People are active while locations are passive
 - Account for metonymy where an entity like a government is referenced by its location (e.g., "Washington expects ...") and is active but there are usually other references to the location in the text so no harm in ignoring some instances
3. Use Knowledge of noun adjuncts to avoid mistaken container relationships such as "In Russia, U.S. officials ..." due to presence of comma
4. Type propagation to make unknown types consistent within a group as long as there is just one known type in the group
 - E.g., name of streets "Federalist", "Market", "Edgewood" while the type entity of "Paul Revere" and "First" are not identified and thus could interpret them as names of streets

Running Example (3)

Democratic candidate for *Texas* Railroad Commissioner *Jeff Weems* stumped in *Paris* late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-*Paris*, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of *Paris*.

Running Example (3)

Democratic candidate for **Texas** Railroad Commissioner **Jeff Weems** stumped in **Paris** late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]

Running Example (3)

Democratic candidate for **Texas** Railroad Commissioner **Jeff Weems** stumped in **Paris** late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said

Running Example (3)

Democratic candidate for **Texas** Railroad Commissioner **Jeff Weems** stumped in **Paris** late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said
3. **Noun adjuncts**: [LOC Houston] attorney

Running Example (3)

Democratic candidate for **Texas** Railroad Commissioner **Jeff Weems** stumped in **Paris** late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said
3. **Noun adjuncts**: [LOC Houston] attorney
4. **Final location entities**: Texas, Paris, Houston, Lamar County, Austin, Dish

Toponym Resolution

1. Dateline
2. Relative geography which is usually vague
 - Ex: "Just outside Lewiston"
3. Comma group where use prominence, proximity, or sibling where share a parent in a geographic hierarchy
 - Prominence: Ex: New York, Philadelphia, Chicago
 - Proximity: Ex: Milwaukee, Chicago, Minneapolis, St. Paul
 - Sibling: Queens, Brooklyn, Manhattan
4. Location/Container – Ex: “College Park, MD”
5. Local lexicon – Ex: “Dublin” in the case of “Columbus, Ohio”
6. Global lexicon
 - Gazetteer with names of places that are known regardless of their geographic location
7. One sense
 - Consistency with previously resolved instances of same name in same source article

News-Specific Geotagging Issues

1. Name of news source
 - Identify a geographic focus (also known as a “spatial reader scope”) for a particular news source in terms of the container(s) of the articles in the source and use this to resolve geotagging ambiguities
2. Perform some preliminary clustering by focusing on the headline
3. Multiple vs: a single interpretation as a geographic location
 - Multiple: evidence that it is a geographic location
 - Single: may be an error, verify by checking
 - population
 - presence of containers
 - presence of proximate locations

Near-Duplicate Images

- Images with slight modification are similar to each other
- News Images about the same event are often near-duplicate



- Use hierarchical color histograms
 - Pros
 - Retains certain color layout information
 - Compressed and efficient to compute
 - Cons
 - Not robust to occlusions or significant cropping, which can dramatically affect the intensity and color layout of the image

Sources of Near-Duplicate Images

- Different news sources may use the same “official” photo with some slight modifications (e.g., cropping, rotation, resizing)



- Photographer may produce many photos in short time period



- Different photographers capture the same scene from different views or different lighting conditions



Detection Result

- 1st Row: Similar time instances
- 2nd Row: Different image croppings
- 3rd Row: Changes in image brightness, contrast, and hue
- 4th Row: Similar grayscale and color images



- Example of similar images not detected by hierarchical color histogram



Hard Cases (Misclassified as Near-Duplicate)

- Backgrounds match and face recognition might be useful



2013-08-24 16:39:47

President Barack Obama



2013-08-28 03:45:25

White House Press Secretary Jay Carney speaks about Syria during a press briefing at the White House

- Hard for pure vision algorithms to detect the difference
 - Context textual information should be used



2012-12-27 15:46:10

Syrian President Bashar al-Assad (R) meets with peace envoy Lakhdar Brahimi in the capital Damascus



2013-02-03 09:06:32

A picture released by the official Syrian Arab News Agency shows Syrian president Bashar al-Assad talking with Iran's Saeed Jalili. Picture: AFP

TwitterStand: News from Tweets

- News gathering system using Twitter
- Twitter is a popular social networking website
 - Tweets are 140 character messages akin to SMS
 - Mostly non-news, often frivolous
- TwitterStand is a spontaneous news medium
 - Idea: users of Twitter help to gather news
 - Distributed news gathering
 - Scooping tool bypassing reporters or newspapers
 - E.g., Michael Jackson's death, Iranian election, Haitian earthquake

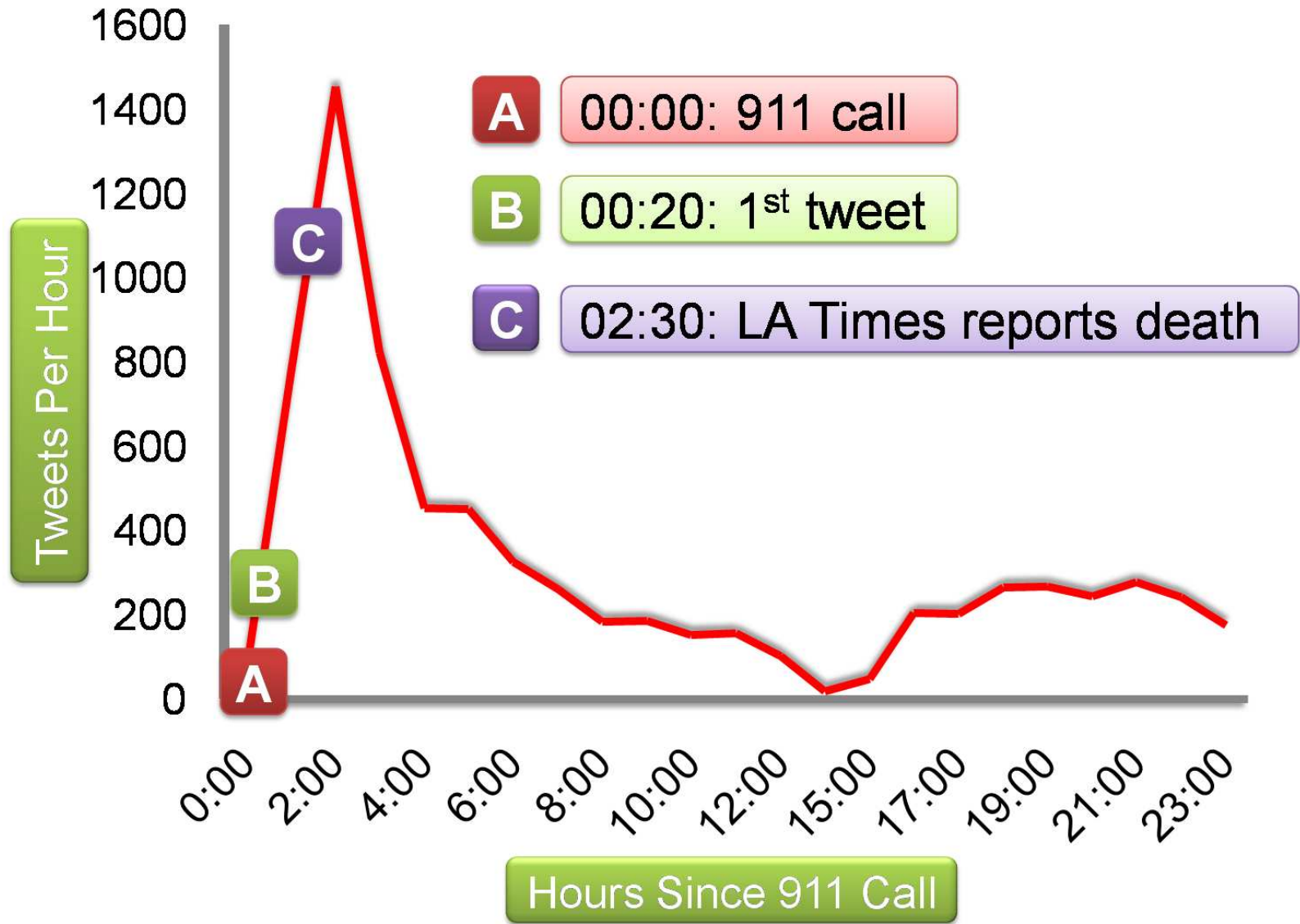
TwitterStand: News from Tweets

- News gathering system using Twitter
- Twitter is a popular social networking website
 - Tweets are 140 character messages akin to SMS
 - Mostly non-news, often frivolous
- TwitterStand is a spontaneous news medium
 - Idea: users of Twitter help to gather news
 - Distributed news gathering
 - Scooping tool bypassing reporters or newspapers
 - E.g., Michael Jackson's death, Iranian election, Haitian earthquake
- Key challenges:
 - Managing the deluge
 - Twitter is a noisy medium as most of the Tweets are not news
 - Challenge: extract news Tweets from mountain of non-news Tweets
 - Tweets are coming at a furious pace
 - Tweets capture the pulse of the moment
 - So, not a good strategy to store and process them in batches
 - TwitterStand uses online algorithms
 - Works without access to entire dataset (i.e., being offline)
 - Determine spatial focus of stories enabling news reading on map

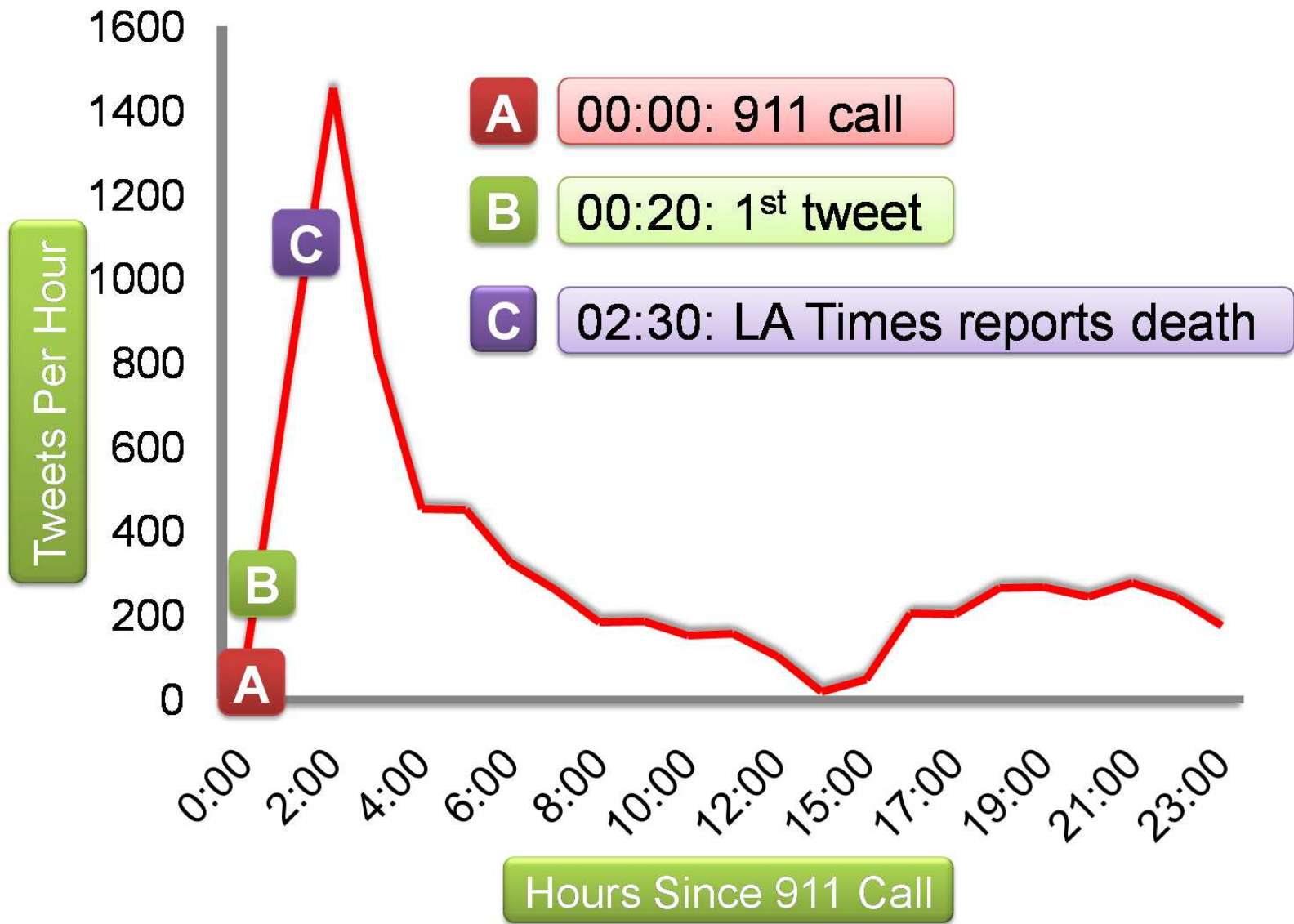
Access to Twitter

1. Whitelisted which means TwitterStand can access Twitter 20K times per hour
2. Access to Gardenhose which yields many Tweets but not clear what percentage
3. Birddog enables TwitterStand to obtain feeds from up to 200K users
4. Seeders are 2000 handpicked users who are known to publish news

Ex: Tweets about Michael Jackson's Death



Ex: Tweets about Michael Jackson's Death



■ Notice that Twitter beat the LA Times by more than two hours

Live Demo: TwitterStand System

The screenshot displays the TwitterStand interface. On the left, there are five news snippets, each with a title, source, time, and tweet count. On the right, a world map shows red location markers corresponding to the news items. The map includes navigation controls at the top and a scale bar at the bottom.

Diplomats deliver ultimatum on Honduras coup
Less than 1 hour ago - *guardiannews*
Diplomats deliver ultimatum on Honduras coup
1011 tweets - Similar Stories - Original Source - Locations

The news is all over Michael Jackson while North Korea is threatening us and testing missiles with no news coverage
Less than 1 hour ago - *Twitter-Streaming*
The news is all over Michael Jackson while North Korea is threatening us and testing missiles with no news coverage
1033 tweets - Similar Stories - Original Source - Locations

thinks jacksons dad Joe should not be anywhere near those kids he abused Michael and will do the same to Michaels kids Give Debbie rowe a go
Less than 1 hour ago - *Twitter-Streaming*
thinks jacksons dad Joe should not be anywhere near those kids he abused Michael and will do the same to Michaels kids Give Debbie rowe a go
574 tweets - Similar Stories - Original Source - Locations

Marines establish positions in Afghan assault
Less than 1 hour ago - *timesnews*
Marines establish positions in Afghan assault
209 tweets - Similar Stories - Original Source - Locations

62 IranElection Tehran Mousavi Iran neda Neda How to send an anonymous email
1 hours ago - *Twitter-Streaming*
62 IranElection Tehran Mousavi Iran neda Neda How to send an anonymous email
5422 tweets - Similar Stories - Original Source - Locations

Sanford to reveal schedule details
Less than 1 hour ago - *thestate*
Sanford to reveal schedule details
1175 tweets - Similar Stories - Original Source - Locations

<http://twitterstand.umiacs.umd.edu/>

- What people are tweeting about rather than where they are tweeting from

STEWARD: A Spatio-Textual Search Engine

1. **S**patio-**T**extual **E**xtraction on the **W**eb **A**iding **R**etrieval of **D**ocuments
2. Sample spatio-textual query:
 - Keyword: “rock concert”
 - Location: near “College Park, MD”
3. Result documents are relevant to both keyword and location
 - Mention of rock concert
 - Spatial focus near “College Park, MD”
4. Issues with results from conventional search engines:
 - Is it the intended “College Park”?
 - What about spatial synonyms such as rock concerts in “Hyattsville” or “Greenbelt”?
 - Don’t usually understand the various forms of specifying geographic content
 - More than just postal addresses!
 - Results often based on other measures, e.g., link structure
5. Applied to HUD USER, PubMed, ProMED-mail, and news

STEWARD Is Not Google Local

1. Google Local geocodes postal addresses into points on the map
 - Address strings are well-formatted
 - Most results drawn from online yellow pages
2. STEWARD works on unstructured text documents
 - Document is a bag of words
3. STEWARD goals:
 - More than searching for addresses in documents, which is easier
 - Identify all geographic locations mentioned in document (i.e., Geotagging)
 - Identify geographic focus of document
 - Retrieve documents by spatio-textual proximity

STEWARD is Different from NewsStand

1. STEWARD focuses on determining the geographic focus or foci of single documents
2. NewsStand focuses on finding clusters of articles on a single topic and associating them with the geographic locations that they are about and to a lesser extent that they mention
3. NewsStand may choose to ignore some locations as being irrelevant to the central topic of the article
4. The common topic of the cluster is used to improve the geographic foci determination process in NewsStand
5. In STEWARD, the user selects the keywords that determine the documents (could be news articles) that are retrieved
6. In NewsStand, the topics are more general than keywords and are determined by the clustering process independent of the user
7. NewsStand uses the functionality of STEWARD to enhance the process of reading particular articles in the cluster
 - Search the cluster for keywords
 - Browse the geographical foci of elements of the clustering

Live Demo: STEWARD System

 Spatio-textual Advanced [Help](#) [Acknowledgements](#)

Keyword(s):

Location (optional):

Lat/Long: Dataset: ■■■■

HUD USER

Results 1-10 of 22

[Previous Results](#) [Next Results](#)

1 [Capacity Building and Governance in El Cenizo](#)

Score: 0.10 [Georefs: 21](#) [Exit Focus Mode](#)

Doc: [Original](#)  [Highlighted](#)

1 of 34 extracts

of its local government. Following a description of the unique development challenges of the **colonias**, the article describes the participatory action research model adopted in this project and

2 [SOUTHWEST HOUSING TRADITIONS](#)

Score: 0.10 [Georefs: 55](#) [Focus](#)

Doc: [Original](#)  [Highlighted](#)

1 of 19 extracts

and Urban Development is committed to meeting the unique housing needs of the citizens of the **"colonias,"** those rural communities and neighborhoods located close to the U.S.-Mexico border that lack

3 [PROBLEMS AND SOLUTIONS](#)



<http://steward.umiacs.umd.edu>

Live Demo: Using STEWARD to Monitor Disease Reporting over Time

The screenshot displays the STEWARD web application interface. At the top left is the STEWARD logo, a globe with a red and yellow checkered pattern. To its right are three tabs: "Spatio-textual", "Advanced", and "Temporal". Further right are links for "Help" and "Acknowledgements". Below the tabs is a timeline slider from 2003 to 2012, with a play button on the left and navigation arrows on the right. The text "from 2005.7 and 2006.3" is positioned above the timeline.

The main content area is split into two panels. The left panel, titled "Results 1-100 of 5875", shows a list of search results. The first result is "Avian influenza, human - East Asia (C)" with a score of 0.10 and 4 Georefs. It includes links for "Original" and "Highlighted" documents and indicates "1 of 128 extracts". The second result is "Avian influenza, human - East Asia (C)" with a score of 0.10 and 10 Georefs, also with "Original" and "Highlighted" links and "1 of 128 extracts". The third result is "Avian influenza, WHO fact sheet".

The right panel is a world map in "Terrain" mode. It features navigation controls (up, down, left, right arrows and zoom in/out buttons) on the left side. The map shows continents labeled: NORTH AMERICA, SOUTH AMERICA, AUSTRALIA, ASIA, and ANTARCTICA. Ocean names include Atlantic Ocean, Pacific Ocean, and Indian Ocean. Numbered blue callout boxes are placed on the map, with numbers 22, 93, 52, 21, 2, 5, 4, 43, 18:A, 16, and 98. A scale bar at the bottom left shows 5000 miles and 5000 kilometers. A small window at the bottom right contains a globe icon and the letters L, W, and E.

<http://steward.umiacs.umd.edu>

Spatio-Textual Spreadsheets

■ Motivation

1. Web is full of structured tables with spatial information in the cells
2. Google's ranking algorithm cannot index this spatial information
3. Understanding the structure of these spatio-textual tables enables a more intelligent search engine

■ Objectives:

1. Identify spatial attributes in spreadsheets
2. Enable web crawlers to take advantage of spatial information in spreadsheets
3. Enable web-based queries on the tuples of spreadsheets
4. Visualize spreadsheets based on their spatial attributes
5. Process spreadsheets in contrast to HTML relational tables as in Google's WebTables

Spatial Coherence in Spreadsheets

- Column coherence: cells in a spatial column share the same spatial type
- Row coherence: containment relationships among spatial data in a row
- Spreadsheet coherence: locations in adjacent rows are usually geographically proximate

State	Zip Code	County Name	Project or Program Type Book 1 - 3	Loan	Grant
AR	725429471	Sharp	City of Highland - Sewer	\$128,000	\$297,000
AR	726539699	Baxter	City of Salesville - Sewer	\$832,000	\$1,479,000
AZ	853620727	Yavapai	Yarnell Water Improvement Assoc.	\$767,000	\$533,000
CA	936090218	Fresno	Caruthers CSD	\$1,515,000	\$988,000
CA	959482117	Butte	City of Gridley	\$2,750,000	\$2,300,850
CA	936152125	Tulare	Cutler PUD	\$1,761,000	\$1,169,000
CA	961309786	Lassen	Leavitt Lake CSD	\$182,000	\$0
CA	952520284	Calaveras	Valley Springs Utility District	\$1,300,000	\$130,000
CA	961370319	Lassen	Westwood Community Services District	\$500,000	\$59,000
CT	62601831	Windham	Town of Putnam	\$7,511,000	\$5,989,000
CT	62601831	Windham	Town of Putnam Wellfield Impr.	\$3,680,000	
CT	60760101	Tolland	Town of Stafford	\$6,566,000	\$5,333,700
FL	32463	Washington	Town of Wausau - water	\$664,000	\$1,691,000
ID	835390126	Idaho	City of Kooskia	\$425,000	
IL	623121303	Pike	City of Barry	\$747,000	\$0

Applications

■ Tuple retrieval from spreadsheets

■ Find the population of India:

```
SELECT population FROM SCHEMA_WITH(country,  
population) WHERE country = 'India';
```

■ Find closest restaurants to a lat/long point:

```
SELECT business_name, phone FROM  
SCHEMA_WITH(business_name, type, address, phone)  
WHERE type = 'restaurant' ORDER BY  
distance(address, '(x,y)') LIMIT 10;
```

■ Mining spreadsheets

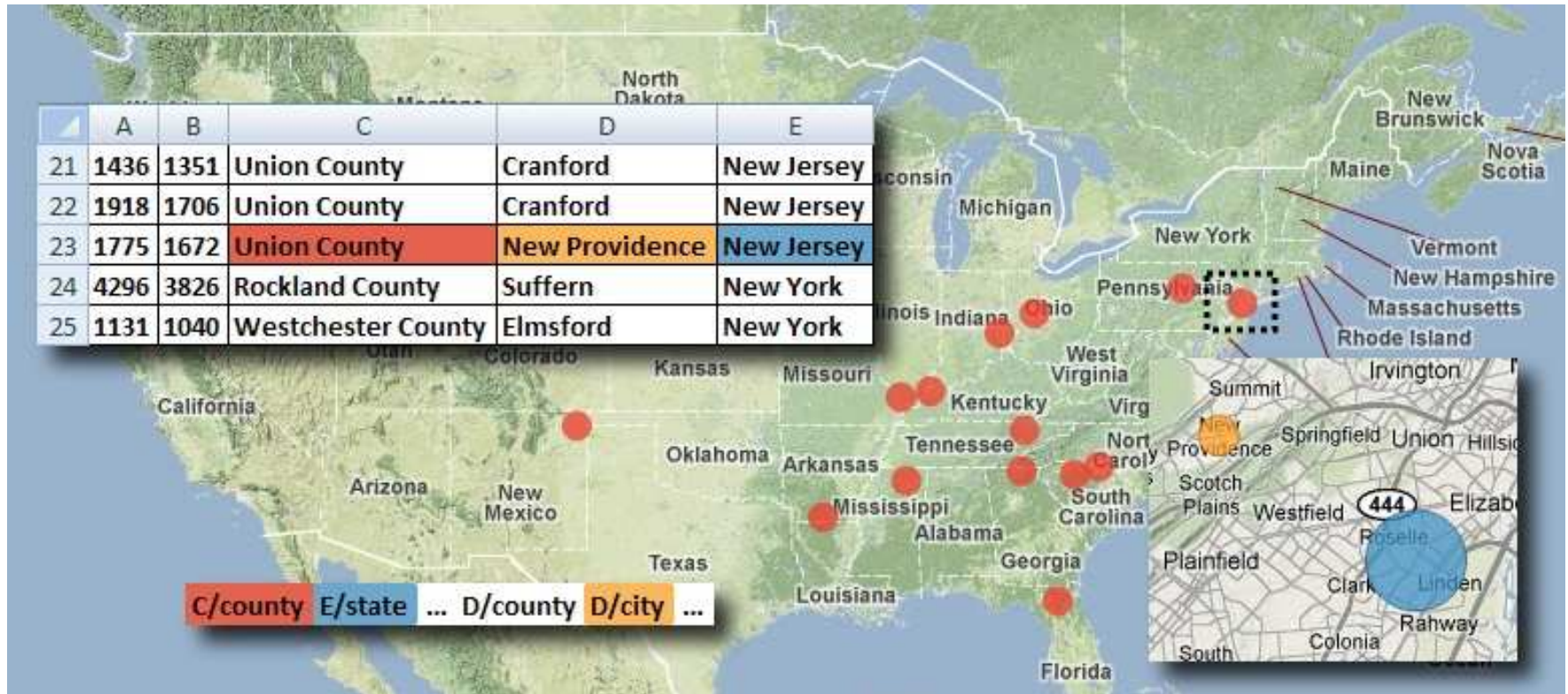
■ Given an attribute (ZIP code, GDP), find its type (number, percentage)

■ Find aliases of spatial column names (“state name”, “State_name”, “StName”, ...)

■ Use spatial attributes as join keys to merge tuples from different spreadsheets

■ Gazetteer generator: gather names and related neighborhood names of large cities in the state of Maryland

Ex: Mapping US County Rent Information

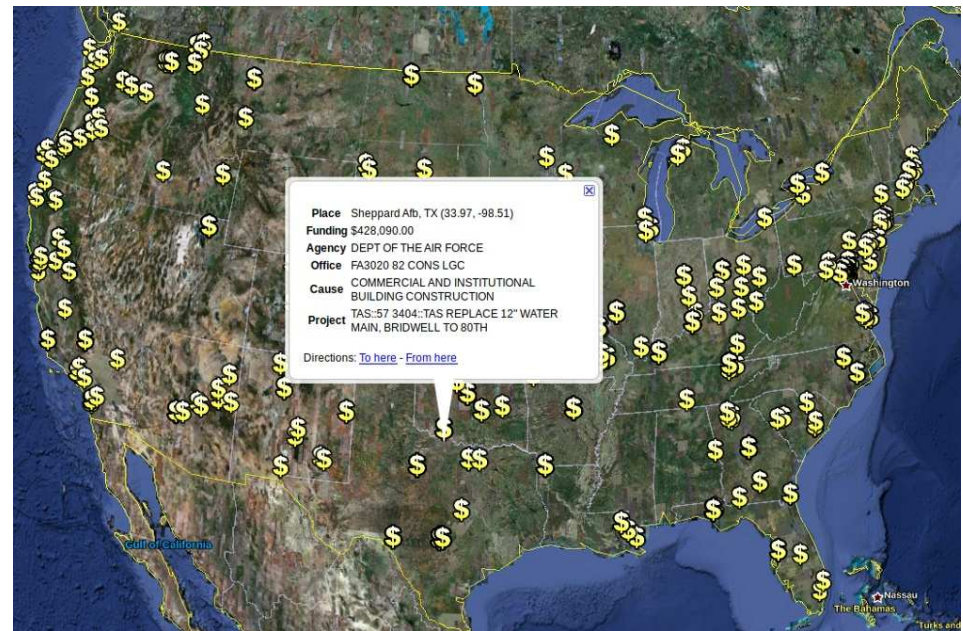


- Generate a consistent location for a row entry
 1. 23 instances of “Union County” (red)
 2. One “New Jersey” (blue) consistent with a “Union County”
 3. One “New Providence” (orange) consistent with both “Union County” and “New Jersey”
- Notice use of colors to differentiate attributes

Ex: Mapping Stimulus Money Spending

DEPT OF THE AIR FORCE	MARION	INDIANAPOLIS	IN	462414812	\$974,988.00
DEPT OF THE AIR FORCE	MARION	INDIANAPOLIS	IN	462414812	\$744,880.00
DEPT OF THE AIR FORCE	DAVIDSON	NASHVILLE	TN	372011815	\$21,982.21
DEPT OF THE ARMY	SANTA BARBARA	LOMPOC	CA	934371499	\$249,951.00
DEPT OF THE AIR FORCE	WICHITA	SHEPPARD AFB	TX	763112716	\$245,783.00
DEPT OF THE AIR FORCE	WICHITA	SHEPPARD AFB	TX	763112716	\$428,090.00
DEPT OF THE AIR FORCE	WICHITA	SHEPPARD AFB	TX	763112746	\$772,141.00
DEPT OF THE AIR FORCE	BEXAR	LACKLAND AFB	TX	782365253	\$1,570,941.63
DEPT OF THE AIR FORCE	WICHITA	SHEPPARD AFB	TX	763112746	\$375,796.37
DEPT OF THE AIR FORCE	LOWNDES	MOODY AFB	GA	316991794	\$68,761.26
DEPT OF THE AIR FORCE	WICHITA	SHEPPARD AFB	TX	763112746	\$519,029.00

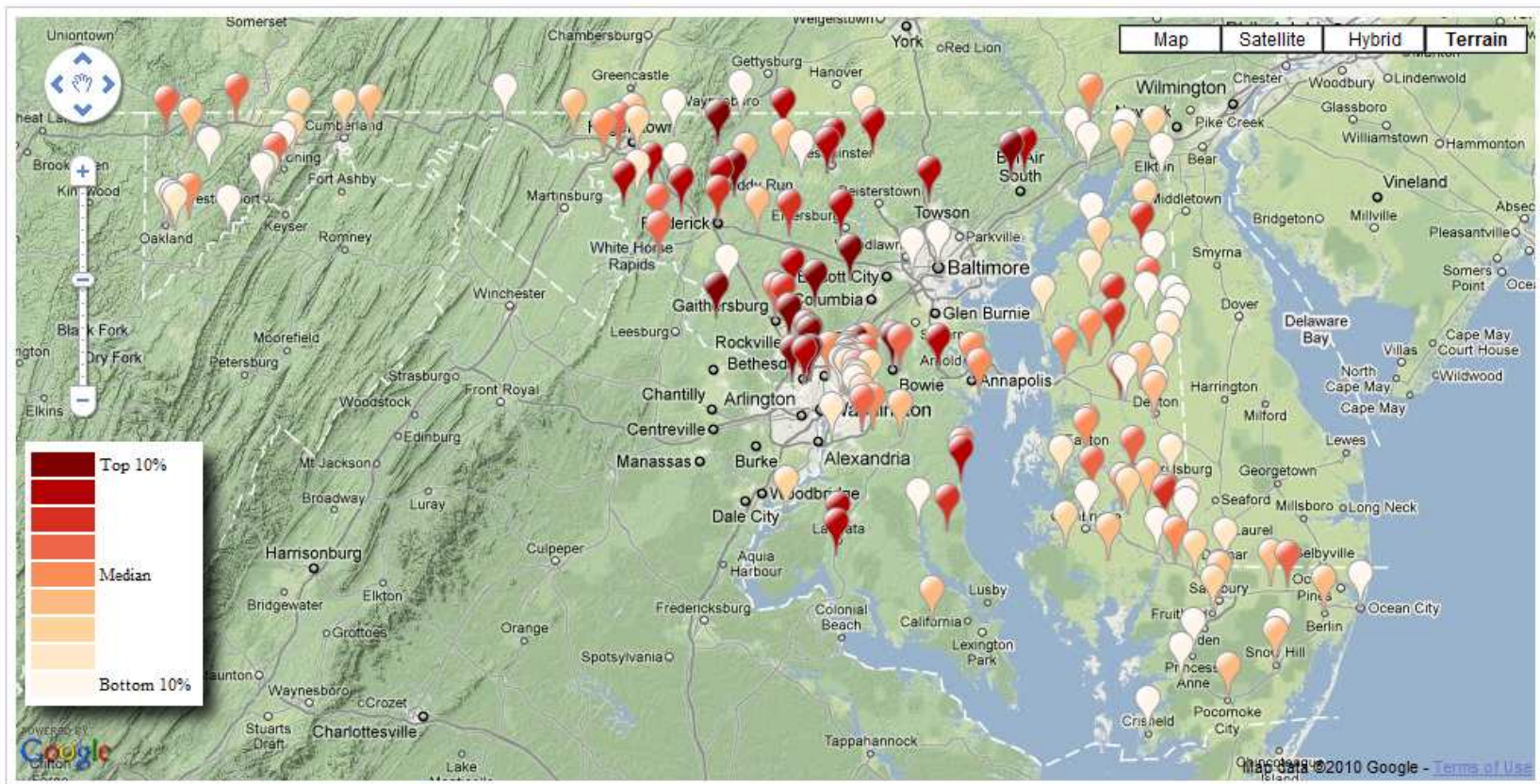
- \$ indicate locations where stimulus money has been spent



Ex: Mapping Census Response Rates

- Spreadsheet only contains location names and values
- Display reveals that locations are in Maryland
- Colors indicate ranges of Census response rates

15	Frederick County	74
16	Garrett County	55
17	Harford County	75
18	Howard County	80
19	Kent County	62
20	Montgomery County	77
21	Prince George's County	68



Color by column: Percent

Future Research Topics/Projects

- Machine learning for individual location classifiers
- Error feedback for learning better location classifiers
- Geocrowdsourcing for geotagging and building corpuses
- Window query efficiency and caching
- Near duplicate image detection
- Incorporation of foreign languages, translation, and clustering using them
- Cluster creation/death, measurements, and GPU implementation
- Keyword searching in NewsStand using pyramid model of news so use first paragraph instead of just headline as is currently the case
- TwitterStand seeder identification
- Improve precision of toponym resolution
- Devise corpuses for evaluating geotagging process
- Cloud-based implementation and quality of service for many users
- Android and Windows 8 ports
- Spatiotemporal visualization such as disease spreading
- Automatic incorporation of ontologies as done for diseases and people