Place-Based Information Systems: Textual Location Identification and Visualization

Hanan Samet*

hjs@cs.umd.edu

Department of Computer Science

University of Maryland

College Park, MD 20742, USA

*Based on Joint Work with Marco Adelfio, Brendan Fruin, Mike Lieberman, Gianluca Quercini, Jagan Sankaranarayanan, and Ben Teitler

Extend GIS Notions to Textually-Specified Spatial Data

- Spatial data specification
  1. Usually geometrically
  2. But could also be done textually
     - Advantage: text is a polymorphic type
       - Ex: “Los Angeles” can denote either an area or a point
     - Disadvantage: ambiguity
       - Ex: “Paris, France” or “Paris, Texas”
- Location-based vs: feature-based queries
  1. Location-based: all documents/topics mentioning location/region R
     - Equivalently, top K topics in location/region R
     - Specify R by direct manipulation like a rectangular window
  2. Feature-based: all locations/regions mentioned in topic T articles
     - Equivalently, top K locations mentioned in articles about topic T
     - T is not necessarily known a priori
     - Topics are ranked by importance which could be defined by the number of documents that comprise them
- Extend further to spatial data specified by direct manipulation actions such as pointing, pan, and zoom

Example 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 people to search for information using a map query interface
  - Advantage: a map, coupled with an ability to vary the zoom level at which it is viewed, provides an inherent granularity to the search process that facilitates an approximate search
  - Distinguished from today's prevalent keyword-based conventional search methods that provide a very limited facility for approximate searches
    - Realized by permitting a match via use of a subset of keywords
    - Users have no grasp of which keyword to use, and thus would welcome the capability for the search to also take synonyms into account
    - Map query interface is a step in this direction
    - Act of pointing at a location (e.g., by the appropriate positioning of a pointing device) and making the interpretation of the precision of this positioning specification dependent on the zoom level is equivalent to permitting the use of 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” or “New York City” are good answers when no such events can be found in “Manhattan” as they correspond to approximate synonyms:
    - “Harlem” by virtue of proximity, and
    - “New York City” by virtue of a containment relationship

Conventional Search Engines and Spatial Synonyms

- Conventional search engines use the page rank method and are good at finding documents containing keywords that we are looking for, but they cannot be easily modified to handle spatial proximity query
- Primary utility is based on grounds of popularity in the sense that the page rank algorithm ensures that the web pages provided to the user as part of the response are ordered by a measure that incorporates some aspect related to their frequency, thereby ensuring that the results are the same as those provided to other users
  - "Democratization of search"
  - All users are treated equally
  - They all get the same bad (or good!) answers
- Use of page rank algorithm to order the results (thereby effectively choosing which results to present to the user) 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 the user
- In case of synonyms, if no links to similar pages on account of being equivalent but for the use of the same words, then the similarity will never be found by the search engine as the page ranking algorithm will never be able to find the similar pages as it crawls the web when building the index to the web pages

Understanding News

- Some related key questions include where are the top stories (i.e., topics) and how do we find them
  - Want to know what is happening around the world and to be able to tunnel down (i.e., using zooming) to specific areas such as
    - South Asia
    - India-Pakistan border
    - Specific neighborhood such as the one from which the reader hails
- 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)
  - Usually present articles grouped by keyword or topic, rather than by geography
- Output of NewsStand, instead, can be summarized as using “What” and “When” to identify “Where” and, to a lesser extent in in terms of our emphasis, “Who”

NewsStand:Spatio-Textual Aggregation of News and Display

1. Crawls the web looking for news sources and feeds
   - Indexing 8,000 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 people, disease, etc.
   - User-generated news (e.g., Social networks such as Twitter)

Goal: Change News Reading Paradigm

- Use a map to read news for all media (e.g., text, photos, 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 the map the medium for presenting all information with a spatial component

Mapping the News

1. 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
2. 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
3. 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. Microsoft Bing
   - Rather primitive and top stories presented linearly
   - Little or no classification by topic
2. 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 (9 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

General Geotagging Issues

1. Toponym recognition: identify geographical references in text
   - Does “Jefferson” refer to a person or a geographical location?
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?
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”

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

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.”

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")

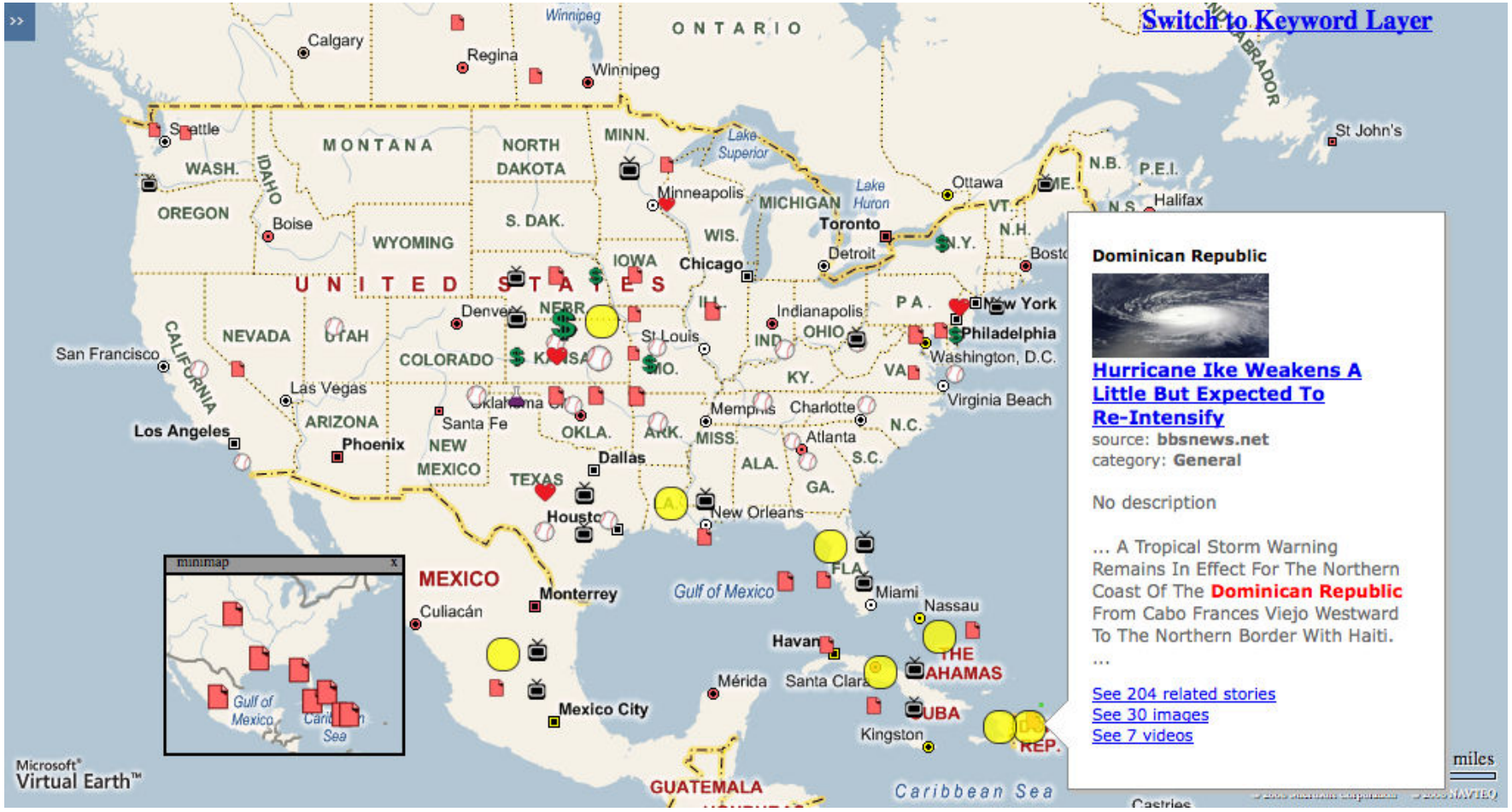
Filtering Toponyms

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

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, Brooklyin, Manhattan
4. Location/Container
   - Ex: “College Park, MD”
5. Local lexicon
   - “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

NewsStand



http://newsstand.umiacs.umd.edu

http://newsstand.umiacs.umd.edu/news/light

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

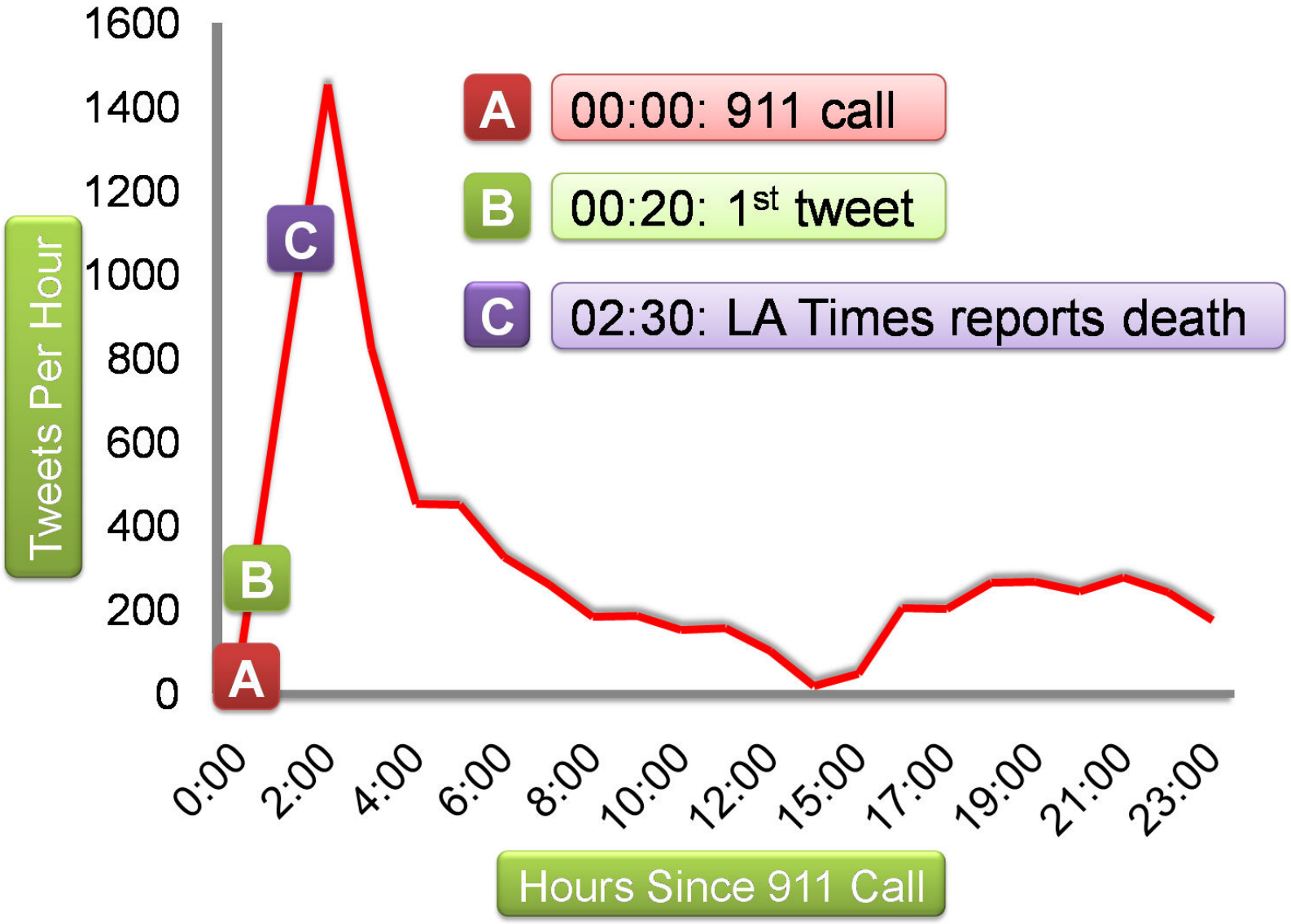
TwitterStand: News from Tweets

- 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

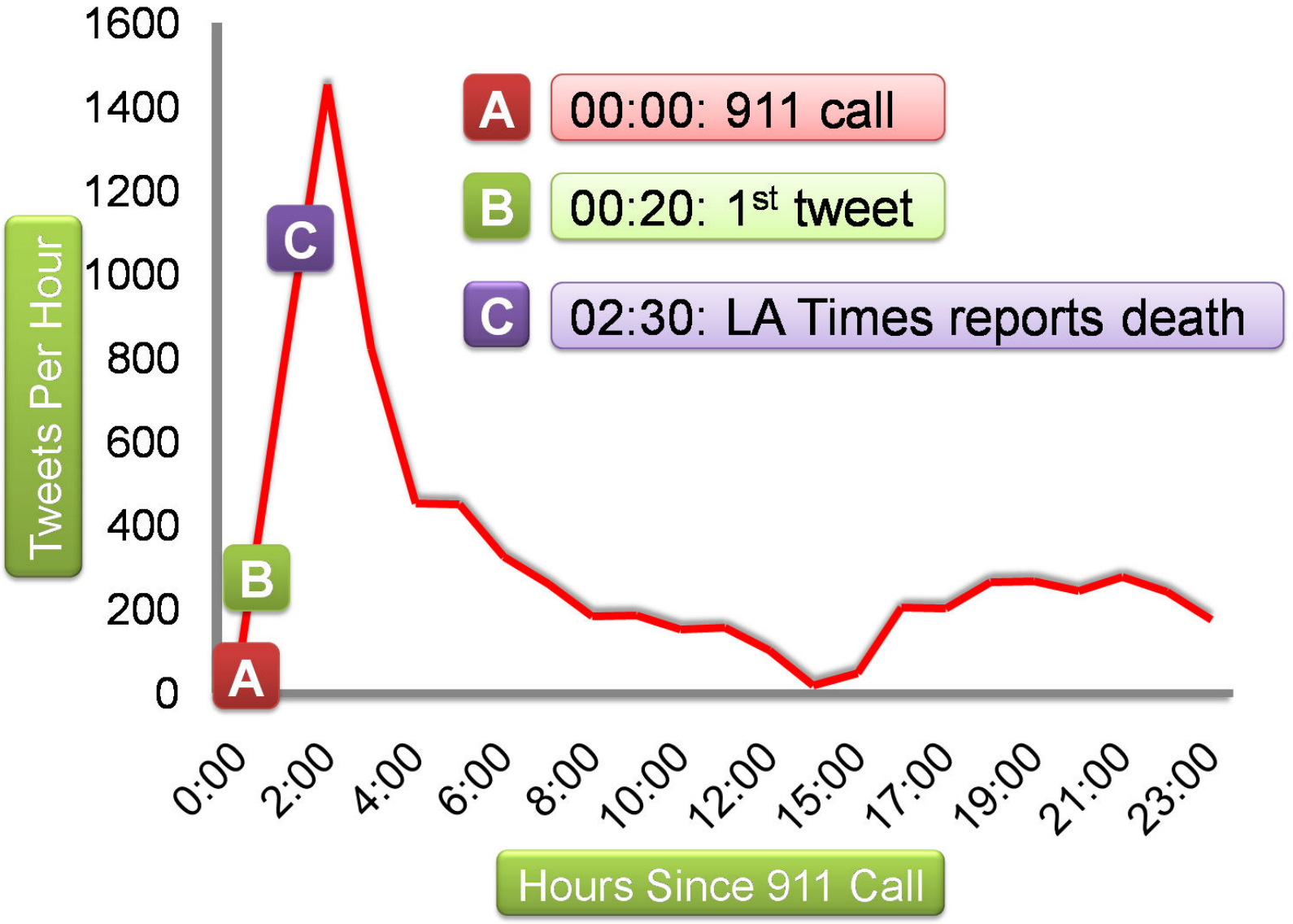
Access to Twitter

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

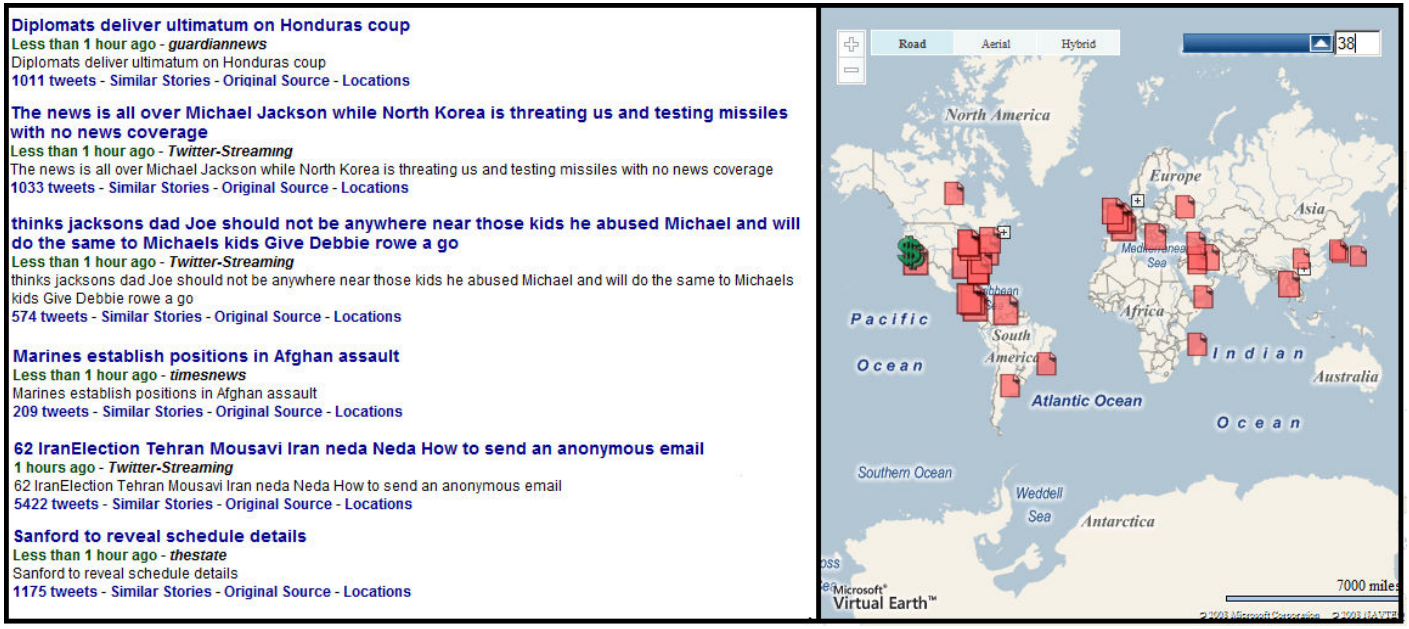


Ex: Tweets about Michael Jackson's Death



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

Live Demo: TwitterStand System



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

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

STEWARD: A Spatio-Textual Search Engine

1. Spatio-Textual Extraction on the Web Aiding Retrieval of Documents
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

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

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:

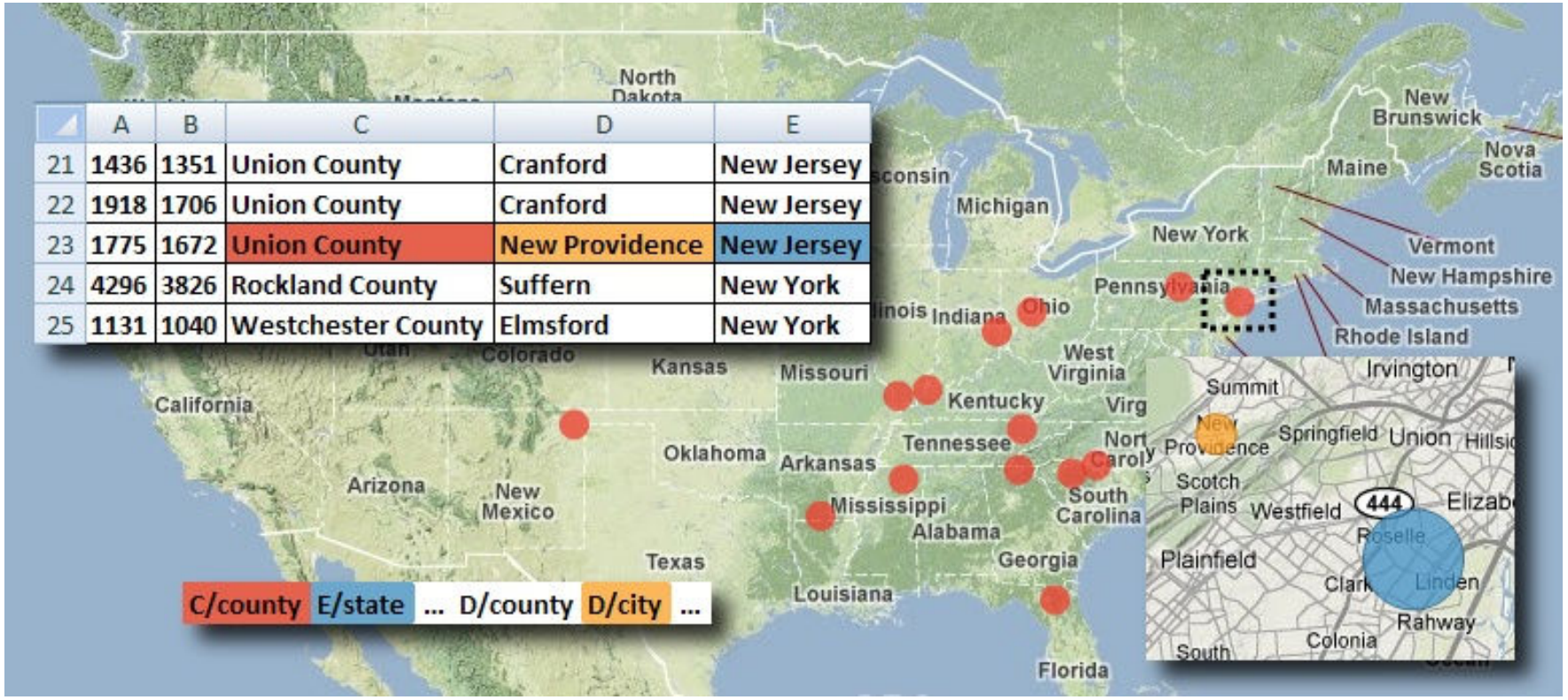
```
SELECT population FROM SCHEMA_WITH(country,
population) WHERE country = 'India';
```

  - 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

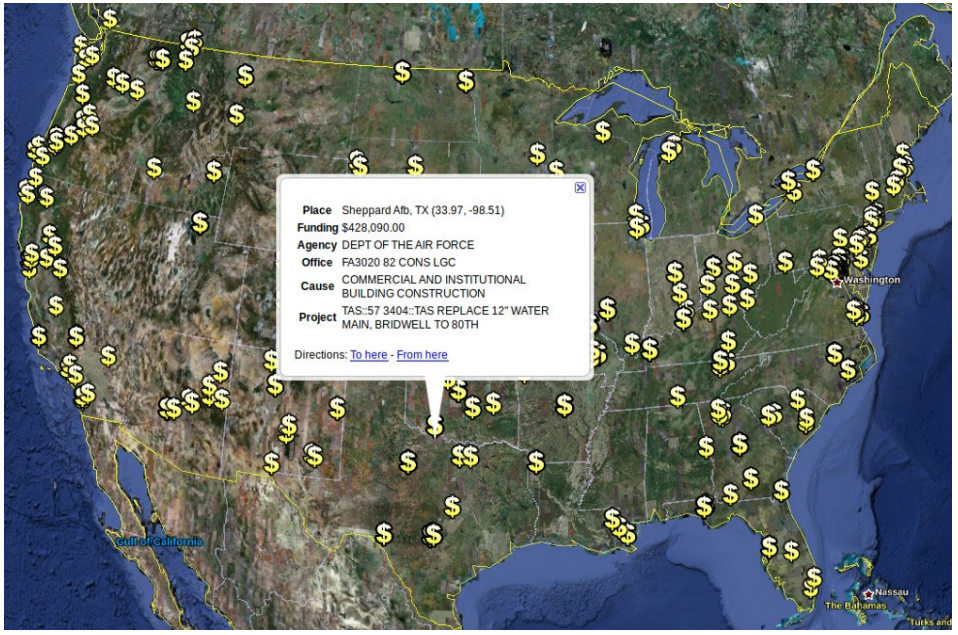


- Generate a consistent location for a row entry
  1. 23 instances of “Union County” (red)
  2. One “New Jersey” (blue) consistent with a “Union County”
  3. One “New Providence” (orange) consistent with both “Union County” and “New Jersey”
- Notice use of colors to differentiate attributes

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

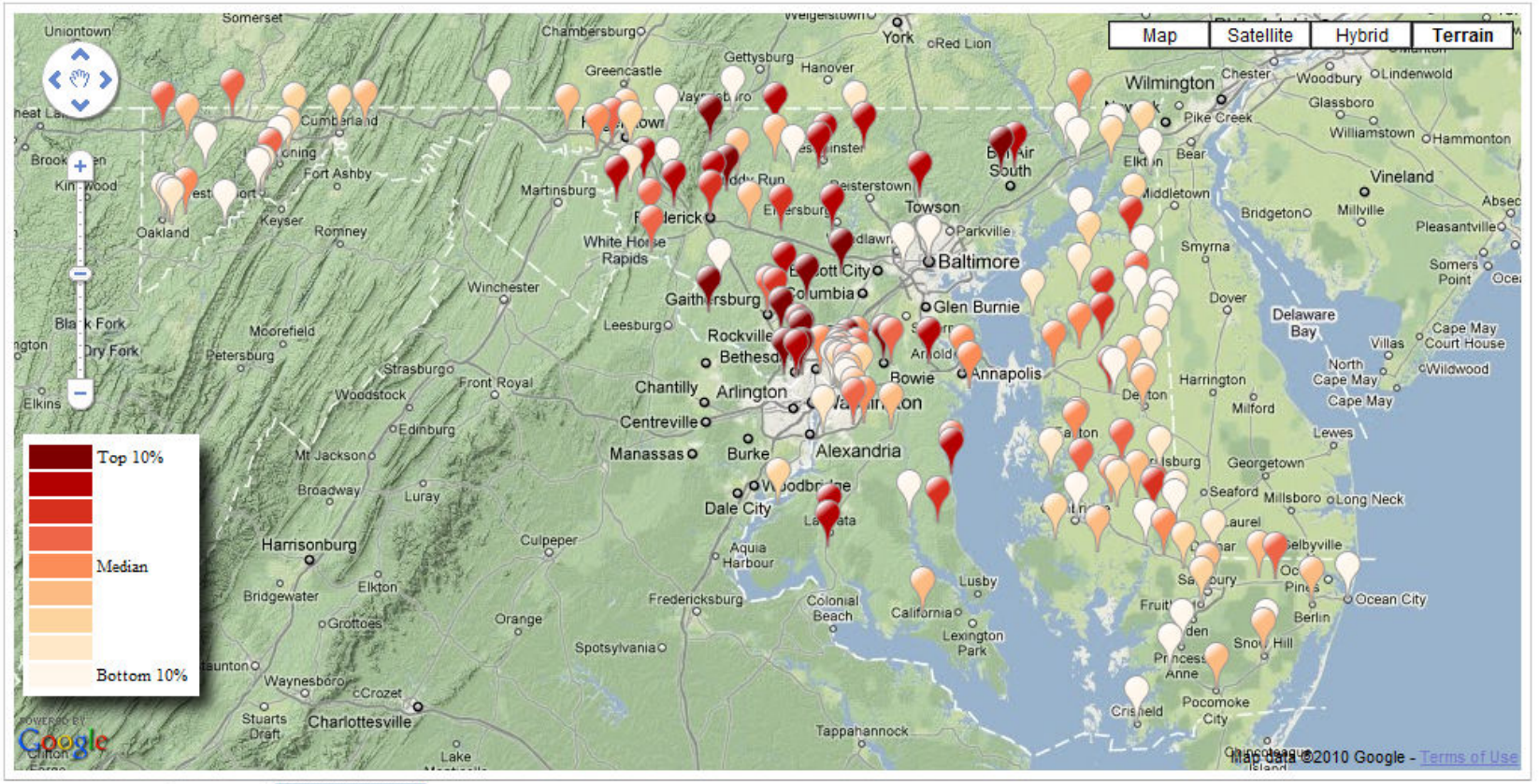
- $ indicate locations where stimulus money has been spent



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 Refresh Map

Concluding Remarks and Future Work

- Power of spatio-textual location specification
- Future work:
  1. Improve precision of toponym resolution
  2. Devise corpuses for evaluating geotagging process
  3. Improve clustering methods and possibly parallelize
  4. Cloud-based implementation
  5. Quality of service for many users