How Can Geospatial “Big Data” Help Disaster Response and Track Disease Outbreaks?
Transform Innovative Geospatial Technology to Solve Real World Problems.

Dr. Ming-Hsiang Tsou
mtsou@mail.sdsu.edu, Twitter @mingtsou
Director of the Center for Human Dynamics in the Mobile Age
Professor, Department of Geography, San Diego State University
Assistant Director, National GeoTech Center
August 19, 2015

What is Big Data?

Image source: http://visual.ly/big-data (definition from IBM, and WIPRO)
Is this a good definition of “Big Data”?

Big Data is Human-Centered Data

Big Data is a large dynamic dataset created by or derived from human activities, communications, movements, and behaviors. (Tsou, 2015).

The term, Big Data, refers to big ideas, big impacts, and big changes for our society in addition to a big volume of datasets.

The Challenge of Big Data Analytics:

Big Data are very Messy, Noisy, and Unstructured!


Require collaboration efforts from linguistics, geographers (GIS experts), computer scientists, data mining experts, statisticians, physicists, modelers, and domain experts.

Big Data Category (Tsou, 2015).

**Social life data:** social media services (Twitter, Flickr, Snapchat, YouTube, Foursquare, etc.), online forums, online video games, and web blogs.

**Health data:** electronic medical records (EMR) from hospitals and health centers, cancer registry data, disease outbreak tracking and epidemiology data.

**Business and commercial data:** credit card transactions, online business reviews (such as Yelp and Amazon reviews), supermarket membership records, shopping mall transaction records, credit card fraud examination data, enterprise management data, and marketing analysis data.

**Transportation and human traffic data:** GPS tracks (from taxi, buses, Uber, bike sharing programs, and mobile phones), traffic sensor data (from subways, trolleys, buses, bike lanes, highways), and mobile phone data (from data transmission records and cellular network data).

**Scientific research data** include earthquakes sensors, weather sensors, satellite images, crowd sourcing data for biodiversity research, volunteered geographic information, and census data.

**Geography (place and time) is the KEY for understanding Big Data!**

Data Integration / Data Fusion

Explore their **spatiotemporal relationships** in both network space and geographical space.

Image provided by Dr. Atshushi Nara (Associate Director of HDMA Center).
Research Showcase #1:  
Geo-Targeted Social Media (Twitter) Analytics for Tracking Flu Outbreaks in U.S.

Question #2:
Do you have a Twitter Account? (YES/NO)
Do you send out tweets regularly? (more than once per week). (YES/NO).

Why Choose Twitter?
80% academic researchers are using Twitter APIs to get their social media data.

1. Free and Open Access Data from APIs (you can write a program in your desktop to download Twitter data (tweets) automatically). But the free APIs has the 1% data limit.
2. Large User Base (+500 million users) and very popular in U.S., Europe, and Japan. But not in China, Taiwan, and Korea (China has a similar platform called "Weibo").
3. Easy to program in Python or PHP (Tweepy, TwitterSearch, etc.). Many available API libraries to use now.
4. Historical data and 100% data can be purchased from Twitter (but very expensive).
5. Rich [Metadata] tags in each tweet (time stamp, user, follower, platform, time zone, text, URL, Retweet, language, devices).

Other possible social media APIs: Flickr, Instagram, Foursquare, Yelp, YouTube. Why not Facebook? (Facebook Graph APIs are VERY LIMITED and PROTECTIVE. No Public data feed). You need to have “internal connections” to Facebook staff to conduct research.

Why not Facebook? (Facebook Graph APIs are VERY LIMITED and PROTECTIVE. No Public data feed). You need to have “internal connections” to Facebook staff to conduct research.

What we can get from Twitter data? Where to find geospatial information?

Example: Use Twitter Search API to search for keyword “HIV test” or “HIV testing”
Only 1% - 7% of Tweets have X, Y GEO-coordinates (from GPS or Geo-tagged).
But 50% - 70% Tweets have city-level locations provided by their user profile.
Time Zone (spatial meaning)
Geo-Targeting Data Collection (Twitter APIs)

Data Filtering, Mining, and Visualization

How to Collect Geo-targeted Social Media Data?

Source: Twitter Search APIs Spatial Search Function to collect public tweets (Geo-targeted) within a city.

Keywords: “flu” and “influenza”.

Region: 17 or 20 miles radius from the center of 31 U.S. Cities.

Time: September 30, 2013 (Week 40) – March 23, 2014 (Week 12)

Twitter Search API: based on the user profile (locations) and gazetteers (San Diego: include La Jolla, La Mesa, Chula Vista) (FREE for Search APIs)

Collect Tweets from Top 31 U.S. Cities (17 miles radius)

31 different cities across the United States (chosen based on their population sizes): Atlanta, Austin, Baltimore, Boston, Chicago, Cleveland, Columbus, Dallas, Denver, Detroit, El Paso, Fort Worth, Houston, Indianapolis, Jacksonville, Las Vegas, Memphis, Milwaukee, Nashville-Davidson, New Orleans, New York, Oklahoma City, Philadelphia, Phoenix, Portland, San Antonio, San Diego, San Francisco, San Jose, Seattle, and Washington, D.C.

Filter and Refine Big Data (Remove Noises)

Machine Learning

Number of tweets:

- Total Flu tweets collected: 307,070.
- Final valid flu tweets: 88,979.

RED Line: National ILI
Purple Line: Weekly Tweeting Rate

(R) value = 0.8494

Trend Analysis at the Municipal Scale (San Diego) with the Lab-tested confirmed flu cases

San Diego: Lab confirmed Flu Cases vs Tweeting Rate: (R) value = 0.9331

Next Step:
Monitor Flu Outbreaks in Real-Time?

The HDMA Center developed the SMART Dashboard:

SMART Dashboard: Social Media Analytic and Research Testbed
http://vision.sdsu.edu/hdma/smart/ (Beta)
CDC Influenza Positive Tests, National Data Summary, through Weeks 40-3, 2014-2015 Season

Problems!!! Twitter broke its Search APIs on 11/20/2014 and only returned Geo-tagged tweets only. (Reduce 90%-95% of tweets collected)

# of Filtered ILI Tweets, Top 30 US Cities, as of February 9, 2015 (from SMART dashboard)

Only 1%-4% tweets has Geo-tagged coordinates.

CDC Influenza Positive Tests, National Data Summary, through Weeks 40-3, 2014-2015 Season

2014-2015 Comparison between ILI and Geotagged-only Tweets among 30 U.S. Cities

Figure 1. The comparison between National ILI Rate and the 31 Cities Tweeting Rate, with prediction up to Week 14. Red: National ILI, Purple: GPS Only Tweets Tweeting Rate, multiplied by 10 for 2014-2015.

Different regions will have different “Voices” and “Opinions”

Question #3:

Does any of your friends get flu last year? (YES/NO).

Does any of your friends get Ebola last year? (YES/NO).
Nepal Earthquake
Monitoring social media voices and discussion about this disaster.
http://humandynamics.sdsu.edu/NepalEarthquake.html

SMART Dashboard Won the BEST METHOD PAPER Award in the 2015 International Conference of Social Media and Society, Toronto, Canada. http://dl.acm.org/citation.cfm?doid=2789187.2789396

SMART Dashboard Examples and Links:
http://humandynamics.sdsu.edu/SMART.html

Research Showcase #2:
Real-time Situation Awareness Viewer for Monitoring Disaster Impacts Using Location-Based Social Media Messages (Twitter).

Geo-targeted Event Observation (GEO) Viewer 1.0
http://vision.sdsu.edu/hdma/wildfire/
Provide dynamic map display about event ground truth observation -- linking GPS locations, texts, photos, and time.

San Diego County: Office of Emergency Services (OES)
How to find out critical information from thousands of GPS-tagged tweets or hundreds of thousands of Non-GPS-tagged tweets?

Nepal Earthquake Example: (keyword search: “trap”)

One Possible Solution: Manual labeling (first 1000 tweets by volunteers) + Machine Learning Classification (built-in).

Digital Volunteers may help us identify and select important Tweets (for machine learning) during and after the disaster events.

Need Some programming and design help from OES, RedCross, and 211:

1. How to combine multiple volunteers’ inputs and integration systems (ranking system).
2. Which category and color schemes/labels should we use for each types of disasters (flooding, wildfires, earthquakes, hurricanes).
3. Which tags might be useful?
4. Who are the target users? What kinds of “Output” system should we create? (for OES staff? For RedCross staff?)
5. Other suggestions?

San Diego Wildfire 2014, May 13, 14, Case Study
(Real time + People’s Need + Public Opinion + Communications)
Social Media (Twitter) Data Collection from May 10 to May 19, 2014

- Total tweets collected

<table>
<thead>
<tr>
<th>Keyword / Topic</th>
<th>Total Tweets</th>
<th>Tweets with GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fire (many unrelated tweets)</td>
<td>51,495</td>
<td>4 % with GPS coordinates</td>
</tr>
<tr>
<td>wildfire</td>
<td>2,885</td>
<td>3 % with GPS</td>
</tr>
<tr>
<td>evacuation</td>
<td>6,477</td>
<td>3 % with GPS</td>
</tr>
<tr>
<td>san marcos</td>
<td>11,781</td>
<td>1 % with GPS</td>
</tr>
<tr>
<td>bernardo</td>
<td>2,871</td>
<td>4 % with GPS</td>
</tr>
<tr>
<td>carlsbad</td>
<td>10,513</td>
<td>2 % with GPS</td>
</tr>
</tbody>
</table>

Tweet Intensity (tweets per day) by Topics

Most relevant tweets started on May 13, 2014. The peak date is May 14, 2014. Then the number of tweets decreased after May 15, 2014.

Temporal Change of Geo-tagged Tweets Related to “fire”, “wildfire”, and “evacuation”.

Spatial clusters (hotspots) of tweets are nearby the actual locations of wildfires (events).
Social Network Analysis (SNA)

- Identify the network influence for each individual (who are the opinion leaders?)
- Predicting the Spreads (Speed, Scale, and Range) of Social Media Messages in Different Social Networks. (following, retweets, and mentions relationships)

SMART Dashboard for Tracking CA Wildfires:
http://vision.sdsu.edu/hdma/smart/wildfire_ca

The Limitations and Challenges of Social Media For Public Health Research
Question #4:

Have you ever used Facebook/Foursquare/Instagram/Twitter to “Check-in” places and restaurants? (YES/NO)

Do your friends “tag” you in their photos or messages? (YES/NO).

Social Media User Profiles

Social Media messages can NOT represent all population, but it can provide warning signals and real-time updates.

Twitter Users are:
- Young (75% are between 15 – 25 years old).
- More urban residents than rural
- Higher adoption % in African Americans
- Many journalists and Mass Media staff.
- 20% are not real “human beings” (robots): many advertisement and marketing activities.

Using Different Keywords can get different demographic groups:
- #Healthcare: include more senior people (Very few teenagers will tweet about “healthcare”). (We need more background study).
- “Keywords” as a sampling tool for social media users.

User Privacy Issue

- Concerns about “Big Brother”.
- Although all the tweets collected from APIs are “public tweets” (everyone can search them and retrieve them).
- Some content of tweets may contain personal private information (real names, locations of homes, offices, private conversations, medical situations, etc.)

* HDMA center conceals tweet locations by randomly selecting a coordinate in a 100m radius of the original location to protect Twitter users’ privacy.

Additional Information and Learning Resources

The Center for Human Dynamics in the Mobile Age
http://humandynamics.sdsu.edu/educationSource.html
Github and YouTube Channels

https://github.com/HDMA-SDSU/HDMA-SocialMediaAPI
https://www.youtube.com/channel/UCKiEJPhr3yRagfIfsa049WA

HDMA Center Activities (Fall 2015)

Annual Big Data Science Symposium (Oct 2): project strong academic influences at national and international levels.


Build a collaborative community for the future development of the Open Government and Open Data Initiative in San Diego (City of San Diego, County of San Diego, SANDAG, SanGIS).

Thank You Q & A

Director: Dr. Ming-Hsiang Tsou
mtsou@mail.sdsu.edu

Funded by

• NSF Cyber-Enabled Discovery and Innovation (CDI) program, Award # 1028177, (2010-2015). http://mappingideas.sdsu.edu/

• NSF Interdisciplinary Behavioral and Social Science (IBSS) Program, Award #1416509 (2014-2018), "Spatiotemporal Modeling of Human Dynamics Across Social Media and Social Networks," http://socialmedia.sdsu.edu/
Backup Slides

Next Step: Syndromic Surveillance (Underdevelopment)
(tracking multiple Symptoms: fever, cold, cough, vomiting, etc.)
http://vision.sdsu.edu/hdma/smart/syndromic

Designed for Early Detection of "unknown" disease outbreaks, such as Swine Flu and SARS

The HDMA Center has built our own Internal Geocoder Engine for User Location Profile:
using GeoNames.org gazetteers (Creative Commons Data) + User defined rules.

Enable Flexible or Self-defined Geo-Target Boundaries (California, Santa Barbara, Los Angeles, San Diego – bounding boxes, or State boundaries)

However, Twitter Search API is not working since Nov, 2014.

Develop manual coding methods and automatic machine learning process for both SMART dashboard and GeoViewer. (Underdevelopment).

From the Wildfires CA collection (the Whole World): Select 500 Original Tweets (Remove RT and URL).
Conduct a manual Coding for (Based on Red Cross Suggestions)
DS – seeking information
DS – Need for help/service (Urgent)
DS – Need for help/service (Non-Urgent)
DS – Offer of help/service
DS – Situation reports
DS – In kind donations
DS – Emotional support
Not-DS relevant

Then Run Machine Learning to test the auto-classification results (For Wildfire Disaster).
Advancing Interdisciplinary Research on BIG DATA, Human Dynamics, and the Social Web

http://humandynamics.sdsu.edu/