

# **Unit 2: Big Data Collection and Process**

Dr. Ming-Hsiang Tsou

San Diego State University

## What is Data Science? (Recap last lecture)

- Data science enables the creation of **data products**.
- **Using data effectively** requires something different from traditional statistics.
- Today's "big" is certainly tomorrow's "medium" and next week's "small." -- The most meaningful definition I've heard: *"big data" is when the size of the data itself becomes part of the problem.*
- We are trying to build "information platforms" (with APIs, tools, and graphics).
- **Making data tell its story.**
- The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades.

# The Fourth Paradigm of Science:

## Data-Driven or **Data-Intensive Science**

tsou

(In Additional Reading Week-2)

Tansley, S., & Tolle, K. M. (Eds.). (2009). The fourth paradigm: data-intensive scientific discovery.

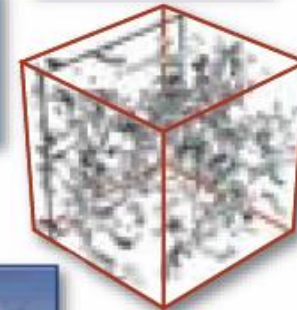
# In the complete book (4<sup>th</sup> paradigm, 2009) –chapter 1.

## Science Paradigms

- Thousand years ago:  
science was **empirical**  
*describing natural phenomena*
- Last few hundred years:  
**theoretical** branch  
*using models, generalizations*
- Last few decades: <sup>tsou</sup>  
a **computational** branch  
*simulating complex phenomena*
- Today: **data exploration** (eScience)  
*unify theory, experiment, and simulation*
  - Data captured by instruments  
or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files  
using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$





## #4: Jim Gray's Fourth Paradigm

- Who is Jim Gray? (work at IBM, DEC,...Microsoft in 1995). SQL relational databases, TerraServer-USA, [http://en.wikipedia.org/wiki/Jim\\_Gray\\_\(computer\\_scientist\)](http://en.wikipedia.org/wiki/Jim_Gray_(computer_scientist))
- **Lost at sea, Jan 28, 2007.**
- Paper written by Clifford Lynch (director of the Coalition for Networked Information (CNI)).
- Gray's paradigm joins the classic pair of opposed but mutually supporting the **second scientific paradigms<sup>tsou</sup>**: **theory** and **experimentation**. The third paradigm—that of large-scale **computational simulation (3)**—emerged through the work of **John von Neumann** and others in the mid-20th century.
- Who is **John von Neumann**? (Father of Computing, a computer architecture – CPU, Storage, Input, Outputs)
- [http://en.wikipedia.org/wiki/John\\_von\\_Neumann](http://en.wikipedia.org/wiki/John_von_Neumann)



# Gray's Fourth Paradigm: Data-intensive Science (Not Data-driven ... Why?)

- The scientific record is intended to do a number of things. First and foremost, it is intended to *communicate* findings, hypotheses, and insights from one person to another, across space and across time.
- *Reproducibility* of scientific results.
- The output of simulations and experiments became large and complex datasets that **could only be summarized**, rather than fully documented, in traditional publications.
- The **data-intensive** computing paradigm: **data and software must be integral parts of the record**—  
tsou
- With computational tools that allow scientists to move beyond the paper to engage the underlying science and data much more effectively and to move from paper to paper, or between paper and reference data collection.
- --Linkage to **eScience** and **Cyberinfrastructure** (to host and archive very large scientific data sets and computational models).

## WHY NOW? (When is the starting of the data-intensive science?)

- The invention of computers -→ 3<sup>rd</sup> paradigm (ENIAC – 1946)
- The invention of Internet, World Wide Web, and Wireless communication → 4<sup>th</sup> paradigm
- Internet → 1987 (TCP/IP protocol)
- WWW → 1992 (HTTP<sup>tsou</sup> protocol)
- Wireless Communication (Wi-Fi) → 1999 (IEEE 802.11a)
- Wireless 3G (GSM, UMTS, and CDMA2000) → 2001 or 2002
- **Smart Phones → 2007 (iPhone and Android phone).**
- **Wireless 4G (LTE) → 2009**
- **The significant progress of computer storage, hardware, and software.**

## Google Flu Trend <https://www.google.org/flutrends/us/#US>

- Video Link Here: <https://www.youtube.com/watch?v=6111nS66Dpk>

google.org Flu Trends
Language: English (United States)

[Google.org home](#)

[Denque Trends](#)

**Flu Trends**

[Home](#)

United States ▼

Cities (Experimental) ▼

Major cities ▼

[Download data](#)

---

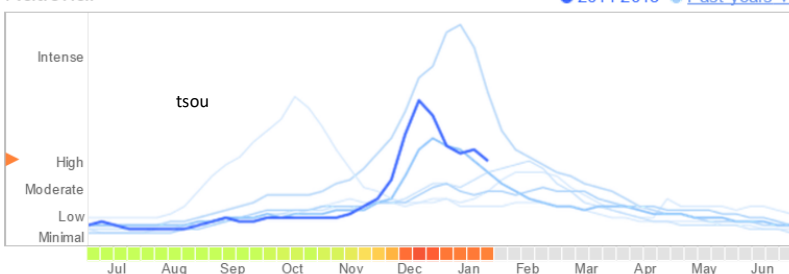
[How does this work?](#)

[FAQ](#)

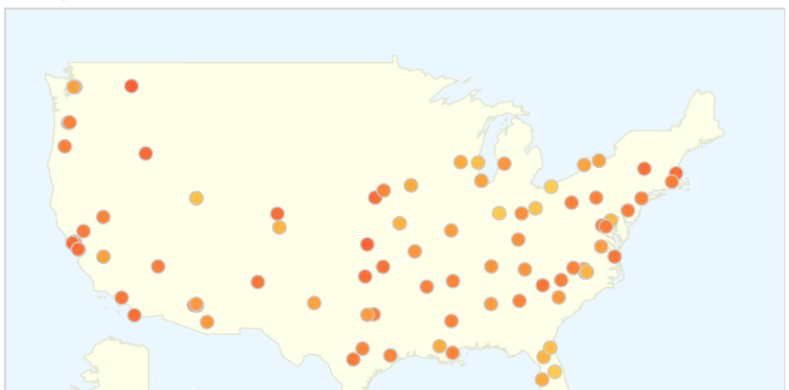
### Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)

**National** ● 2014-2015 ● Past years ▼




States | Cities (Experimental) - Click on a city below to chart the flu trend above.



#### Fight influenza

CDC urges you to take these steps to protect yourself and others from the flu:

1. Get vaccinated against flu – it's your best defense.
2. Cover your cough, wash hands often.
3. Take antiviral drugs if your doctor recommends them.

 [Centers for Disease Control and Prevention](#)

#### Animated Flu Trends in Google Earth

[Download and explore](#) Flu Trends data in Google Earth. Need Google Earth? [Download it here.](#)

#### Embed this chart

Use [this embed code](#) to show this chart on your website.

## Google Trend Exercise (15 mins):

- Use the Web Browser to open: <https://www.google.com/trends/>
- Compare the search result for “Big Data” and “Geography”. What’s their trends? And Seasonal Patterns?
- Choose **two comparable terms** and use Google Trend to compare their results. What are your finding?

tsou

- What are the “strength” of Google Trend?
- What are the potential problems and errors of Google Trend?
- What are the “weakness” of Google Trend?



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



**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. **GOOGLE TREND DATA?**

tsou



**Transportation and human traffic data:** GPS tracks (from taxi, buses, **Uber**, **bike sharing** programs, and mobile phones), traffic censor data (from subways, trolleys, buses, bike lanes, highways), connected vehicles (V2V, GPS tracks), 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 (iNaturalist)**, volunteered geographic information, and census data.

**Different data have different collection methods and APIs.**

- **Public Domain Data (Free cost and Free use)**
  - Census Data (limit to census tracks). <http://www.census.gov/data.html>
  - National Spatial Data Infrastructure). <https://www.geoplatform.gov/>
  - Open Data and Open Government (2013): <https://www.data.gov/>  
<https://www.whitehouse.gov/open>
  - Voting Records (San Diego County Registrar of Voters  
<http://www.sdvote.com/content/rov/en/reportquery.html>
- **Free Cost Data (not necessary public domain – limited use)**
  - **Public Twitter Data APIs** (Stream-API or Search API). Users can download, but **can not share the downloaded data to others (in database format)**. (Data are still owned by Twitter).  
tsou
  - Other Social Media or Web Services Data collected via APIs (similar to Twitter).
  - Google Search Engine Results and Google Trend.
  - (**Data are collectable, but no allowed legally** – such as **YikYak** Data.  
[https://en.wikipedia.org/wiki/Yik\\_Yak](https://en.wikipedia.org/wiki/Yik_Yak) ). (Shutdown in April 28, 2017).
  - Some Data will require specialized programs or “web crawlers” to collect.
  - (A **Web crawler** is an [Internet bot](#) which systematically browses the [World Wide Web](#),  
cited from Wikipedia).

- **Purchasable Data (private or value-added)**
  - Twitter Firehose (GNIP – only for very specific partners ): <http://support.gnip.com/apis/firehose/overview.html>
  - Twitter PowerTrack API (GNIP): search for historical tweets (estimated cost: \$1000 for 100,000 tweets) – expensive?
  - AirSage (CDR data – cell phone data): [www.airsage.com/](http://www.airsage.com/)
  - ESRI Tapestry Data (combine American Community Survey (ACS) data and other business data – value added data). <http://www.esri.com/landing-pages/tapestry>
  - Business Data: MLS (multiple listing service – for real estate), **others?**
- **Governmental-protected Data**
  - Cancer Registry Data (need to apply for and require IRB approval).
  - Census Data: **non-public Census microdata** (at Federal Statistical Research Data Centers<sup>tsou</sup>): California Census Research Data Center: <http://www.ccrdc.ucla.edu/>
- **Private-own Data (not purchasable).**
  - Business Data: Zillow is an online real estate database company (<http://zillow.com> ).
  - Electronic medical records (EMR) in hospitals or health insurance companies.
  - Facebook Data (non public posts).
  - Uber Data
  - Amazon Transaction Data



## Social Media Data via **API (Application Programming Interface)**:

What is an API? **A set of data communication protocols and formats to allow computer programs or applications to request or provide data products.** (modified from wikipedia and others' definition).

-- like a **Power Plug** -- receiving data automatically – required different formats.

- **Twitter REST / Search APIs:** <https://dev.twitter.com/rest/public/search>
  - RESTful API (representational state transfer) using HTTP (get, post, put, delete) and URI. Popular data format is **JSON (JavaScript Object Notation)** or **XML**. (One request each time, not continue, it can collect **historical tweets back to 7 or 9 days**).
- **Twitter Streaming APIs:** <https://dev.twitter.com/streaming/overview> Real-time data update and stream. Can not request historical tweets.
  - Public streams (usually with the limitation of 1% data).
    - **Streaming APIs can use “keywords” or “bounding box” to search – but it can not use both together!**
  - User streams (from a single user's tweets)
  - Site streams (connect to multiple users).

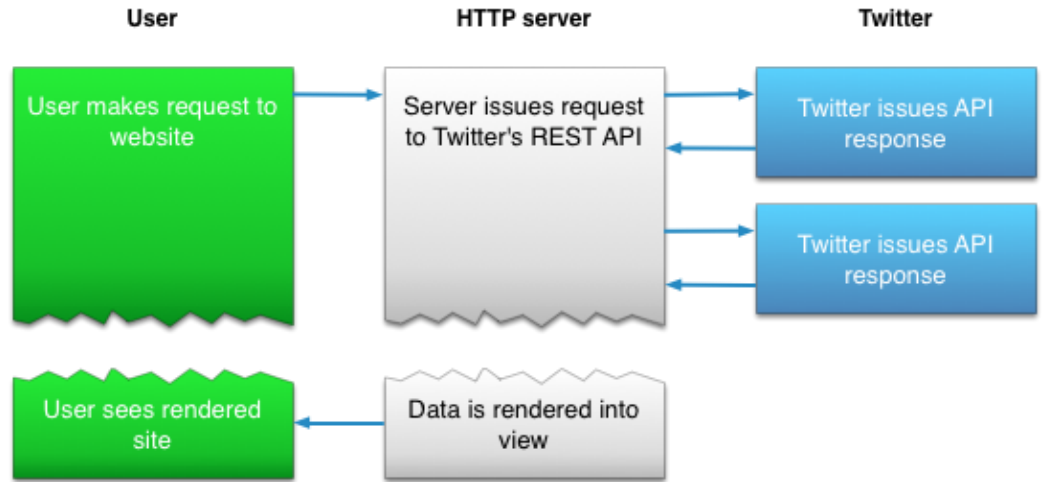
80% academic researchers are using Twitter APIs to get their social media data.

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. tsou
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.

Twitter REST / Search APIs  
 (Example: SMART dashboard)



tsou

Twitter Streaming APIs  
 (using Python's Tweepy library:  
**StreamListener**)  
 (Example: GeoViewer dashboard)

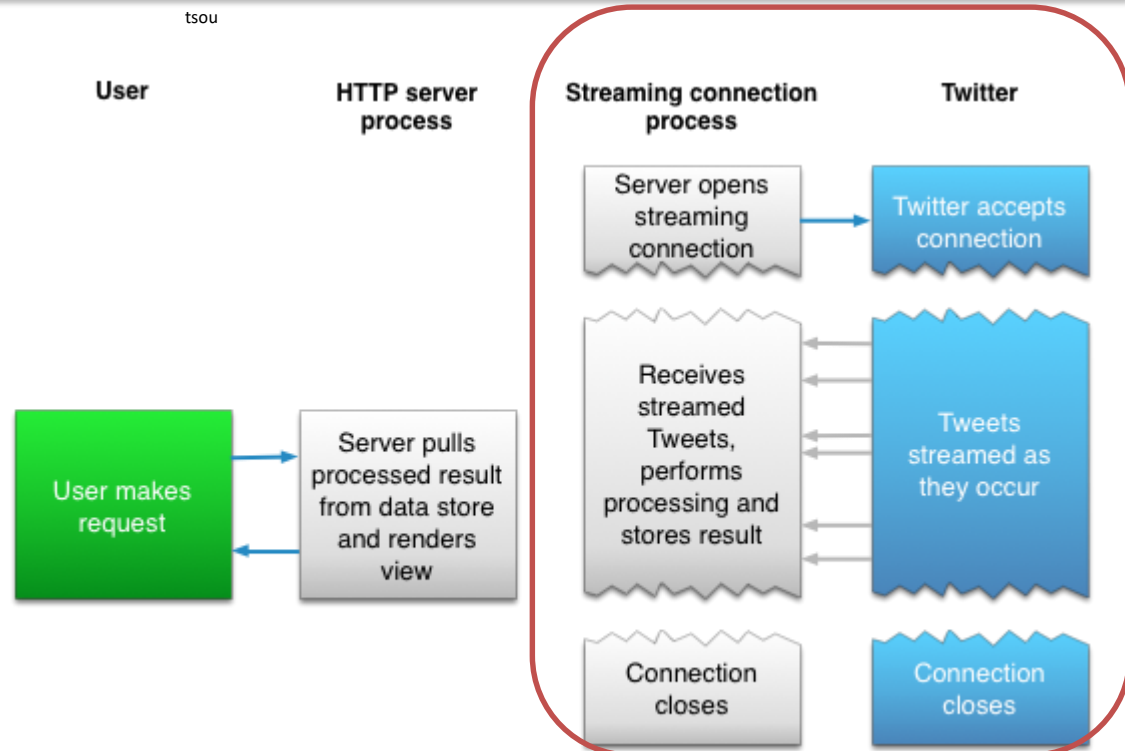
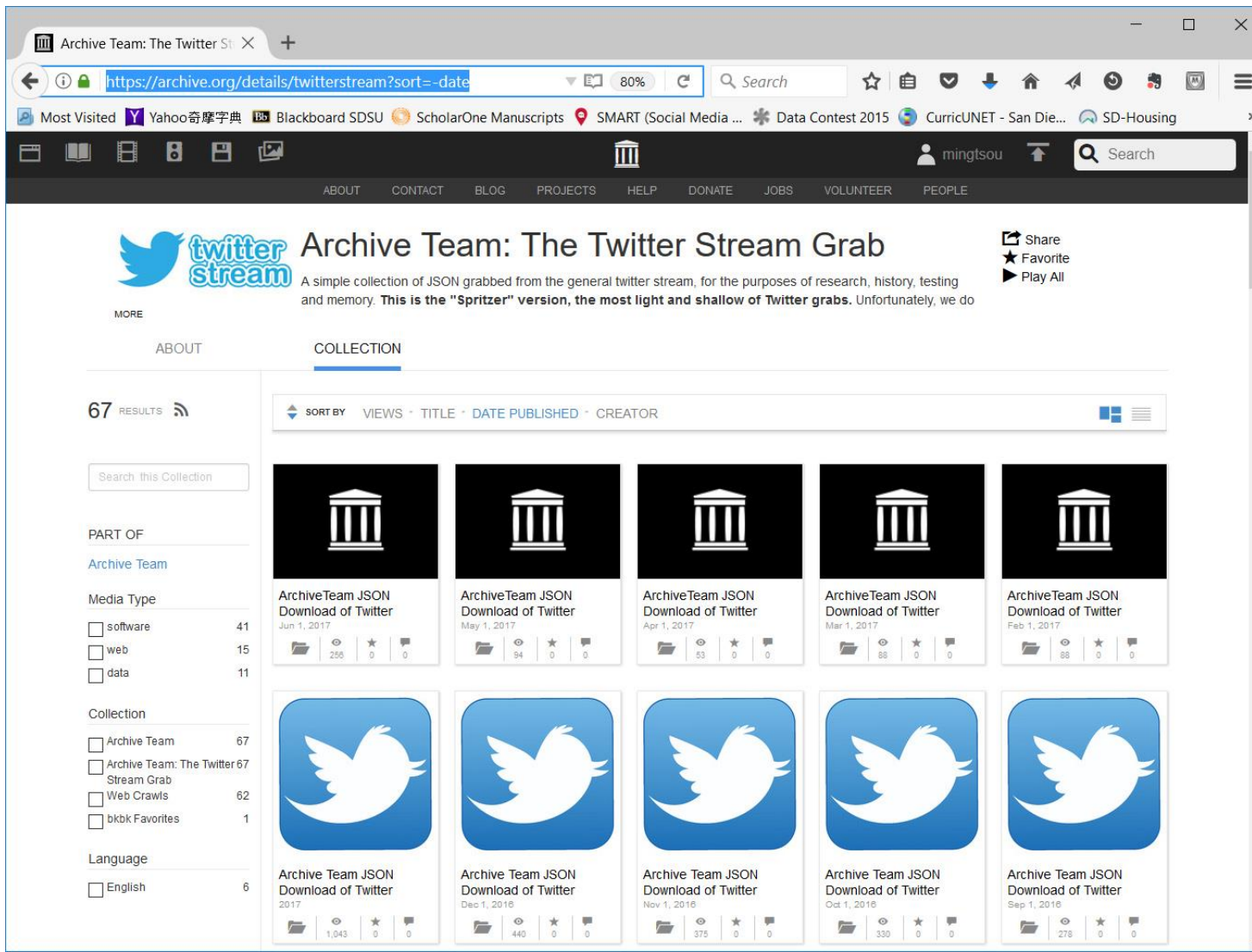


Image source:  
<https://dev.twitter.com/streaming/overview>

# The Internet Archive:

<https://archive.org/details/twitterstream?sort=-date>



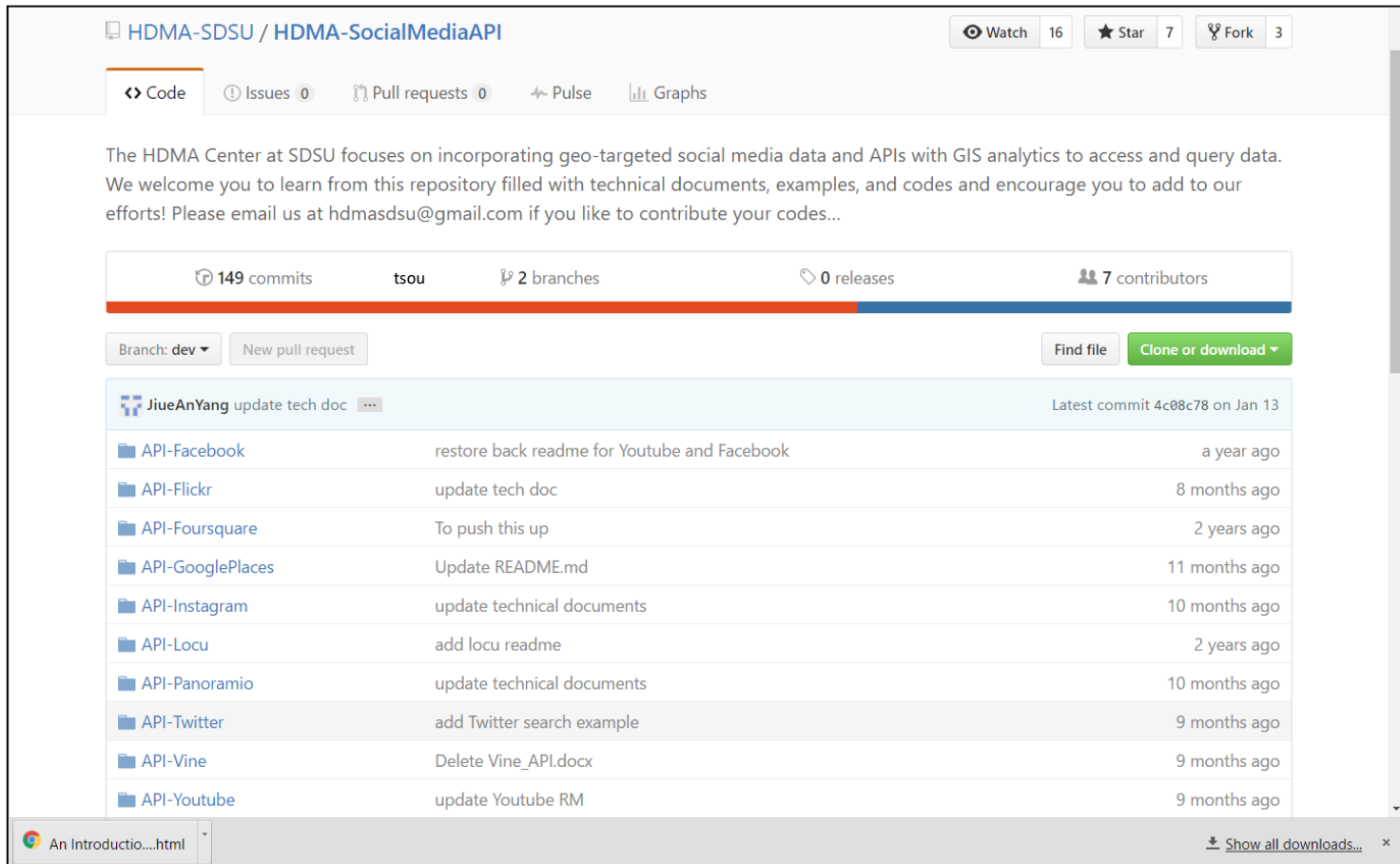
The screenshot shows a web browser window displaying the Internet Archive page for "Archive Team: The Twitter Stream Grab". The browser's address bar shows the URL <https://archive.org/details/twitterstream?sort=-date>. The page features a navigation menu with links for ABOUT, CONTACT, BLOG, PROJECTS, HELP, DONATE, JOBS, VOLUNTEER, and PEOPLE. The main content area includes the Twitter Stream logo and the title "Archive Team: The Twitter Stream Grab". Below the title, there is a description: "A simple collection of JSON grabbed from the general twitter stream, for the purposes of research, history, testing and memory. This is the 'Spritzer' version, the most light and shallow of Twitter grabs. Unfortunately, we do". The page is sorted by "DATE PUBLISHED" and displays 67 results. The results are organized into two rows of five items each. Each item is represented by a thumbnail with a title, a date, and a set of icons for views, stars, and comments. The first row of thumbnails has a black background with a white classical building icon, while the second row has a blue background with a white Twitter bird icon. The titles for all items are "Archive Team JSON Download of Twitter". The dates for the items in the first row are Jun 1, 2017, May 1, 2017, Apr 1, 2017, Mar 1, 2017, and Feb 1, 2017. The dates for the items in the second row are 2017, Dec 1, 2016, Nov 1, 2016, Oct 1, 2016, and Sep 1, 2016. On the left side of the page, there is a sidebar with a search box and filter options for "PART OF", "Media Type", "Collection", and "Language".

## HDMA Github - Social Media APIs:

<https://github.com/HDMA-SDSU/HDMA-SocialMediaAPI>

- Flickr and Four Square API demos:

[http://vision.sdsu.edu/ychuang/Flickr\\_InstagramAPI/socialMedia\\_API.html](http://vision.sdsu.edu/ychuang/Flickr_InstagramAPI/socialMedia_API.html)



The screenshot shows the GitHub repository page for HDMA-SDSU / HDMA-SocialMediaAPI. The repository has 16 watchers, 7 stars, and 3 forks. It contains 149 commits, 2 branches, 0 releases, and 7 contributors. The main branch is 'dev'. A table lists various API folders and their commit history:

Folder Name	Commit Message	Time Ago
API-Facebook	restore back readme for Youtube and Facebook	a year ago
API-Flickr	update tech doc	8 months ago
API-Foursquare	To push this up	2 years ago
API-GooglePlaces	Update README.md	11 months ago
API-Instagram	update technical documents	10 months ago
API-Locu	add locu readme	2 years ago
API-Panoramio	update technical documents	10 months ago
API-Twitter	add Twitter search example	9 months ago
API-Vine	Delete Vine_API.docx	9 months ago
API-Youtube	update Youtube RM	9 months ago

- **Online Forum**

- **Public** online forum: <https://www.patientslikeme.com/> other examples?
  - <https://csn.cancer.org/forum> ,
  - <https://www.blogforacure.com/>
- **Private** online forum (need passwords): Facebook Closed Group. Members only forum (political groups, or others).
- Use **Web Scraper** to collect data (web harvesting). Potential legal issues. [https://en.wikipedia.org/wiki/Web\\_scraping](https://en.wikipedia.org/wiki/Web_scraping) (Google Search Engine is a web scraper?).
  - Example: Python with BeautifulSoup4.  
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
  - <https://www.import.io/>
  - <http://scrapy.org/> (opensource)

- **Web search engines (and their APIs)**
  - Google Search Engine: *Google Custom Search* (<https://developers.google.com/custom-search/>) is the current API recommended by Google for web search. This API allows **100 results** for every inquiry. Google custom search lists a number of options which allow developers to customize their search settings.
  - Bing (Microsoft) Search Engine: *Bing Search API* has been moved to Microsoft Azure Market recently as an integral part of Microsoft online service. Bing Search API can return **1,000 results at maximum**. It also requires authentication, similar to Google Search API. The only difference, though, is that given a language Bing Search API requires users to specify the region to retrieve search results. Bing Search API provides 58 language-region pairs.
  - **Yahoo Search Engine: Yahoo BOSS APIs were discontinued on March 31, 2016.**



## Examples of Web Search Engine API results (Search for “Obamacare” in Google)

Rank	Search Engine	Keyword	Search Date	URL	Title
1	Google	Obamacare	40874	<a href="http://en.wikipedia.org/wiki/Patient_Protection_and_Affordable_Care_Act">http://en.wikipedia.org/wiki/Patient_Protection_and_Affordable_Care_Act</a>	Patient Protection and Affordable Care Act - Wikipedia
2	Google	Obamacare	40874	<a href="http://newsbusters.org/other-topics/obama-watch">http://newsbusters.org/other-topics/obama-watch</a>	ObamaCare   NewsBusters.org
3	Google	Obamacare	40874	<a href="http://obamacarewatch.org/">http://obamacarewatch.org/</a>	ObamaCare Watch
4	Google	Obamacare	40874	<a href="http://fixhealthcarepolicy.com/tag/obamacare/">http://fixhealthcarepolicy.com/tag/obamacare/</a>	Fix Health Care Policy   ObamaCare
5	Google	Obamacare	40874	<a href="http://blogs.investors.com/capitalhill/index.php/h">http://blogs.investors.com/capitalhill/index.php/h</a>	20 Ways ObamaCare Will Take Away Our Freedom
6	Google	Obamacare	40874	<a href="http://www.conservapedia.com/ObamaCare">http://www.conservapedia.com/ObamaCare</a>	ObamaCare - Conservapedia
7	Google	Obamacare	40874	<a href="http://blog.heritage.org/2011/09/28/obamacare-">http://blog.heritage.org/2011/09/28/obamacare-</a>	Obamacare Has Arrived in the Supreme Court
8	Google	Obamacare	40874	<a href="http://online.wsj.com/article/SB1000142405297">http://online.wsj.com/article/SB1000142405297</a>	Adler and Cannon: Another ObamaCare Glitch - WSJ
9	Google	Obamacare	40874	<a href="http://biggovernment.com/tag/obamacare/">http://biggovernment.com/tag/obamacare/</a>	ObamaCare - Big Government
10	Google	Obamacare	40874	<a href="http://www.time.com/time/magazine/article/0,91">http://www.time.com/time/magazine/article/0,91</a>	READ The Fatal Flaw of Obamacare - Time Magazine
11	Google	Obamacare	40874	<a href="http://michellemalkin.com/2009/06/19/the-obama">http://michellemalkin.com/2009/06/19/the-obama</a>	Michelle Malkin » The Obamacare horror story you
12	Google	Obamacare	40874	<a href="http://www.weeklystandard.com/keyword/obama">http://www.weeklystandard.com/keyword/obama</a>	Obamacare   The Weekly Standard
13	Google	Obamacare	40874	<a href="http://www.naturalnews.com/ObamaCare.html">http://www.naturalnews.com/ObamaCare.html</a>	Obamacare news and articles
14	Google	Obamacare	40874	<a href="http://obamacare411.wordpress.com/">http://obamacare411.wordpress.com/</a>	ObamaCare 411
15	Google	Obamacare	40874	<a href="http://washingtonexaminer.com/taxonomy/term/2">http://washingtonexaminer.com/taxonomy/term/2</a>	Topic: obamacare News   Washington Examiner
16	Google	Obamacare	40874	<a href="http://dailycaller.com/2011/10/15/americans-just">http://dailycaller.com/2011/10/15/americans-just</a>	CLASS Act   Americans just dodged an ObamaCare
17	Google	Obamacare	40874	<a href="http://www.washingtonpost.com/blogs/ezra-klein">http://www.washingtonpost.com/blogs/ezra-klein</a>	Taking back 'Obamacare'? - The Washin
18	Google	Obamacare	40874	<a href="http://www.forbes.com/sites/davidwhelan/2011/0">http://www.forbes.com/sites/davidwhelan/2011/0</a>	Florida Judge Rules Against ObamaCare, Calls Ind
19	Google	Obamacare	40874	<a href="http://www.nationalreview.com/articles/280402/o">http://www.nationalreview.com/articles/280402/o</a>	Obamacare's Great Unraveling - Rich Lowry -
20	Google	Obamacare	40874	<a href="http://www.businessweek.com/magazine/repal">http://www.businessweek.com/magazine/repal</a>	Repeal Obamacare? Good Luck - Businessweek
21	Google	Obamacare	40874	<a href="http://www.newsmax.com/InsideCover/ObamaCa">http://www.newsmax.com/InsideCover/ObamaCa</a>	Obamacare Foes Rejoice Over CLASS Act Demise
22	Google	Obamacare	40874	<a href="http://www.thenewamerican.com/usnews/health">http://www.thenewamerican.com/usnews/health</a>	Ohio Votes to Nullify ObamaCare
23	Google	Obamacare	40874	<a href="http://www.huffingtonpost.com/2011/11/03/obarr">http://www.huffingtonpost.com/2011/11/03/obarr</a>	Obamacare Repeal Not Nearly As Easy As GOP C
24	Google	Obamacare	40874	<a href="http://www.humanevents.com/article.php?id=465">http://www.humanevents.com/article.php?id=465</a>	ObamaCare Reaches the Supreme Court - HUMAN



- **Electronic medical records (EMR):** <https://www.healthit.gov/providers-professionals/electronic-medical-records-emr> “An electronic medical record (EMR) is a digital version of a **paper chart** that contains all of a **patient’s medical history** from **one practice**. An EMR is mostly used by providers for diagnosis and treatment.”  
(**EHR : Electronic Health Record** – similar to EMR, but more advanced, integrated – link to individuals rather than a provider).  
EMR can provide longitudinal electronic record of patient health information. **But EMR data collected for clinical and billing purposes, NOT for research purpose.** (challenges: in/out migration, errors, ambiguities, omissions, biases).
  - NextGen Health Information System: <https://www.nextgen.com/>
  - [https://en.wikipedia.org/wiki/NextGen Healthcare Information Systems](https://en.wikipedia.org/wiki/NextGen_Healthcare_Information_Systems)
- **Personal health records (PHR):** “A personal health record (PHR) is an electronic application used by patients to maintain and manage their health information in a private, secure, and confidential environment.” <https://www.healthit.gov/providers-professionals/faqs/what-personal-health-record> (Managed by Patients, rather than providers). Early example: Google Health – discontinued on 2012. WHY?).
  - Microsoft HealthVault, Apple’s Health and HealthKit, Dossia (open source).
  - <http://dossia.com/products/health-manager.html#overview-video> (watch video)
  - <https://www.youtube.com/watch?v=nRc87EwsSgI> (HealthVault 5 mins)



## Sample Medical Record: Monica Latte

Previous Page

Table of Contents

Use for April 2011 abstraction

### WeServeEveryone Clinic

1111 First Street California  
111-111-1111 Fax: 111-111-1111

Chart Summary

### Monica Latte

Home: 444-444-4444  
Female DOB: 04/04/1950 0000-44444

Ins: Commercial xxxxx

### Patient Information

Name: Monica Latte

Home Phone: 444-444-4444

Address: 4444 Coffee Ave  
Chocolate, California

Office Phone:

Patient ID: 0000-44444

Fax:

Birth Date: 04/04/1950

Status: Active

Gender: Female

Marital Status: Divorced

Contact By: Phone

Race: Black

Soc Sec No: 444-444-4444

Language: English

Resp Prov: Carl Savem

MRN: MR-111-1111

Referred by:

Emp. Status: Full-time

Email:

Sens Chart: No

Home LOC: WeServeEveryone

External ID: MR-111-1111

### Problems

DIABETES MELLITUS (ICD-250.)  
HYPERTENSION, BENIGN ESSENTIAL (ICD-401.1)

### Medications

PRINIVIL TABS 20 MG (LISINOPRIL) 1 po qd  
Last Refill: #30 x 2 : Carl Savem MD (08/27/2010)  
HUMULIN INJ 70/30 (INSULIN REG & ISOPHANE (HUMAN)) 20 units ac breakfast  
Last Refill: #600 u x 0 : Carl Savem MD (08/27/2010)

### Directives

Allergies and Adverse Reactions (! = critical) tsou

### Services Due

FLU VAX, PNEUMOVAX, MICROALB URN

3/18/2011 - Office Visit: F/u Diabetes  
Provider: Carl Savem MD  
Location of Care: WeServeEveryone Clinic

### OFFICE VISIT


History of Present Illness  
Reason for visit: Routine follow up  
Chief Complaint: No complaints

### History

Diabetes Management  
Hyperglycemic Symptoms  
Polyuria: no  
Polydipsia: no  
Blurred vision: no

Internet Citation: Sample Medical Record: Monica Latte. Content last reviewed May 2013. Agency for Healthcare Research and Quality, Rockville, MD. <http://www.ahrq.gov/professionals/prevention-chronic-care/improve/system/pfhandbook/mod8appbmonicalatte.html>


< My page EDIT PROFILE





Walking on December 30, 2014


**Ming Tsou**  
You have walked 3327 steps per day on average while using S Health for the last month.

Rewards >


  
Aug 18


  
Jul 18


  
Jul 10


  
Jul 10


Personal bests tsou

  
**17627**  
steps  
Most steps  
56 days ago

  
**35**  
min  
Longest duration  
Dec 30, 2014

  
**111**  
Cal  
Most calories burned  
Jan 21, 2015


  
**1.34**  
mi


  
**11.2**  
mph

Dr. Ming-Hsiang Tsou, 鄧明祥, SDSU


S Health [User Icon] [Menu Icon]

ME TOGETHER DISCOVER


 **27**/10000 steps




12 AM 6 AM Now 6 PM 12 AM


 Sleep



Were you asleep between 12:00 AM and 6:00 AM? Tap here to record your sleep.





12 AM 6 AM 12 PM 6 PM 12 AM


  
  
Stress  
**MEASURE**

  
**96%**  
SpO<sub>2</sub>  
**MEASURE**

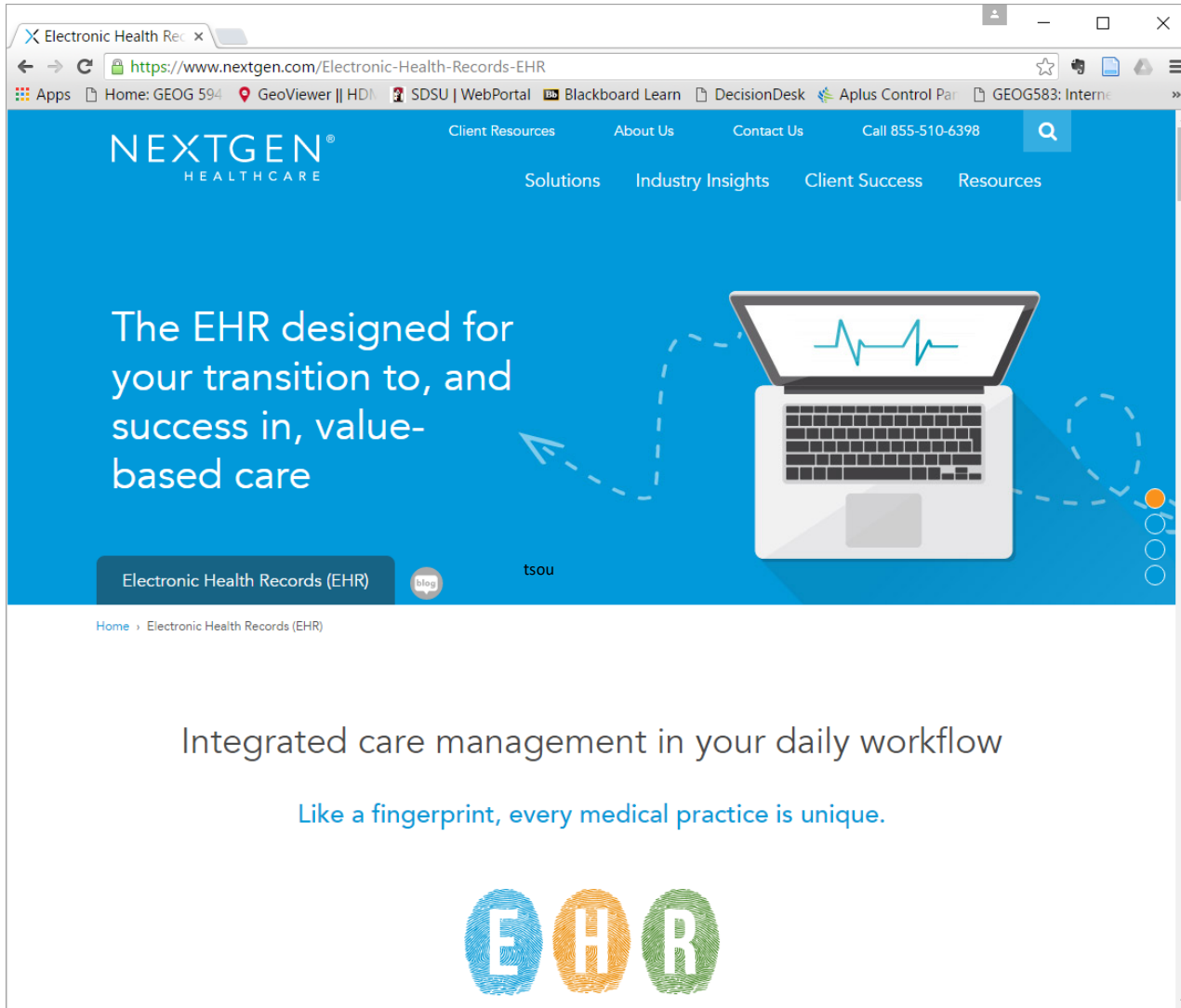
  
Low  
  
UV  
**MEASURE**

  
**84 bpm**  
Heart rate  
**MEASURE**

  
Running  
**START**

  
**14 min**  
Walking  
**START**

Dr. Ming-Hsiang Tsou, 鄧明祥, SDSU



The screenshot shows a web browser window displaying the NextGen Healthcare website. The browser's address bar shows the URL <https://www.nextgen.com/Electronic-Health-Records-EHR>. The website has a blue header with the NextGen Healthcare logo and navigation links: Client Resources, About Us, Contact Us, Call 855-510-6398, Solutions, Industry Insights, Client Success, and Resources. The main content area features a large blue banner with the text "The EHR designed for your transition to, and success in, value-based care" and an illustration of a laptop displaying a heart rate monitor. Below the banner, there is a breadcrumb trail: Home > Electronic Health Records (EHR). The main heading reads "Integrated care management in your daily workflow" with the tagline "Like a fingerprint, every medical practice is unique." At the bottom, the letters "EHR" are displayed, where each letter is contained within a fingerprint graphic.

<https://www.nextgen.com/Electronic-Health-Records-EHR>

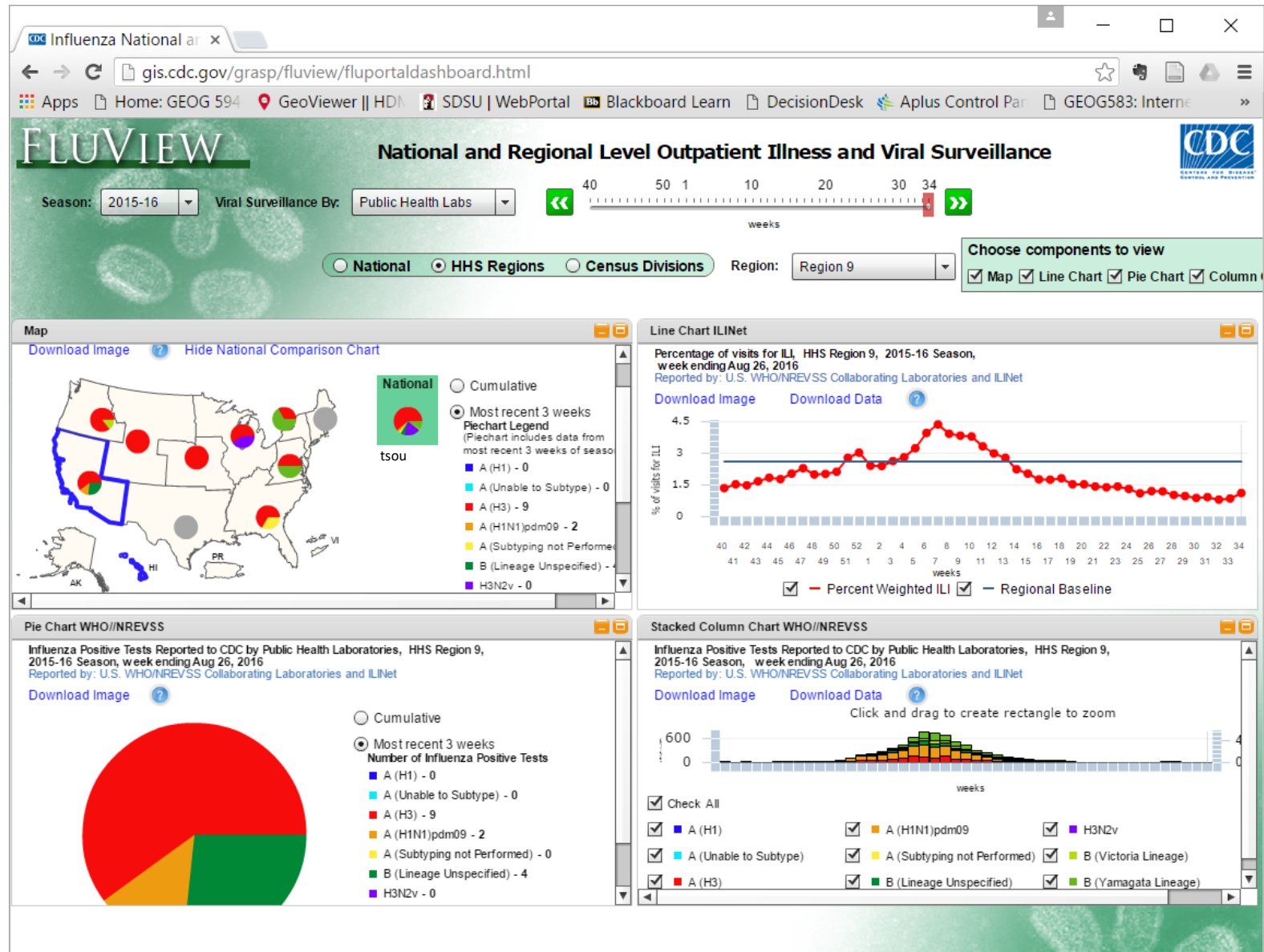
## Cancer Registry Data:

- CDC National Program of Cancer Registries (NPCR):  
<https://www.cdc.gov/cancer/npcr/> in all 50 states.
- SEER (NCI Surveillance, Epidemiology, and End Results Program).  
<http://seer.cancer.gov/>
- California Cancer Registry: <http://www.ccrca.org/>
- San Diego County Live Well Data Portal: <https://data.livewellsd.org/>

## Disease Outbreak and Epidemiology Data:

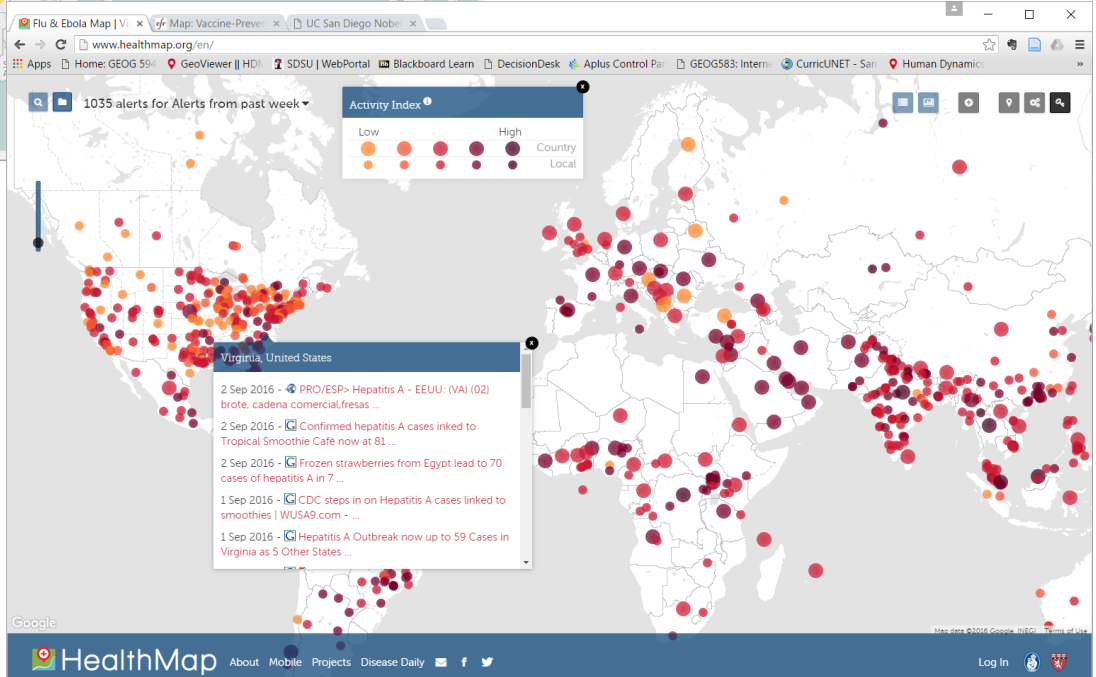
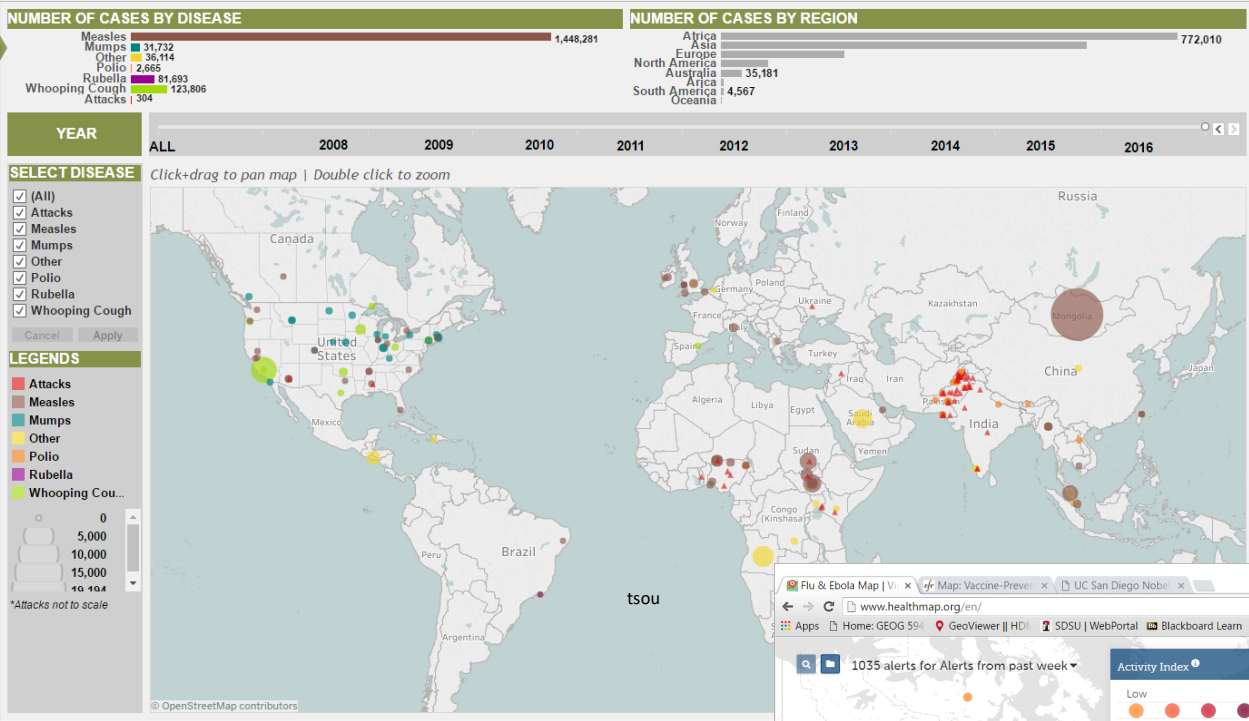
- CDC Flu Outbreak Monitoring:  
<http://www.cdc.gov/flu/weekly/fluactivitysurv.htm>
- WHO Disease Outbreak News (DONs): <http://www.who.int/csr/don/en/>
- HealthMap (Boston, Dr. John Brownstein) <http://www.healthmap.org/en/>
- Vaccine-Preventable Outbreaks (**Laurie Garrett**) :  
[http://www.cfr.org/interactives/GH\\_Vaccine\\_Map/index.html#map](http://www.cfr.org/interactives/GH_Vaccine_Map/index.html#map)
- SMART dashboard Flu Monitoring: <http://vision.sdsu.edu/hdma/smart/flu2>

<http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>



Vaccine-Preventable Outbreaks

EMBED DOWNLOAD DATA [social media icons] 146



What are the differences between the two web maps?



## Business Data:

- Credit card transactions (credit score): three major [credit bureaus](#) : [Experian](#), [TransUnion](#), and [Equifax](#).
  - **Experian's** principal lines of business are **credit services, marketing services, decision analytics and consumer services**. The company collects information on people, businesses, motor vehicles and insurance. It also collects 'lifestyle' data from on- and off-line surveys.)
  - **Equifax** has operated primarily in the business-to-business sector, selling consumer credit and insurance reports and related analytics to businesses in a range of industries (cited from Wikipedia).
  - **Yelp Review and Amazon Review**: Yelp develops and publish [crowd-sourced](#) reviews about local businesses (**Yelp APIs don't provide review contents, just the individual business info** and the summarized ranks.
  - Locu API: <https://dev.locu.com/documentation/>



ESRI Business Analytics Online (BAO): Require ArcGIS online accounts and BAO subscription: <http://www.esri.com/software/businessanalyst>  
<https://bao.arcgis.com/esriBAO/login/>

tsou

Hello, Ming-Hsiang Tsou | Preferences | Help | Support

USA

Home | Maps | Reports

## Access 2016/2021 US Demographics

Up-to-date data on US population, Tapestry Segmentation, Retail MarketPlace, Consumer Spending, and Market Potential are now available in the data browser and reports.

GET STARTED NOW

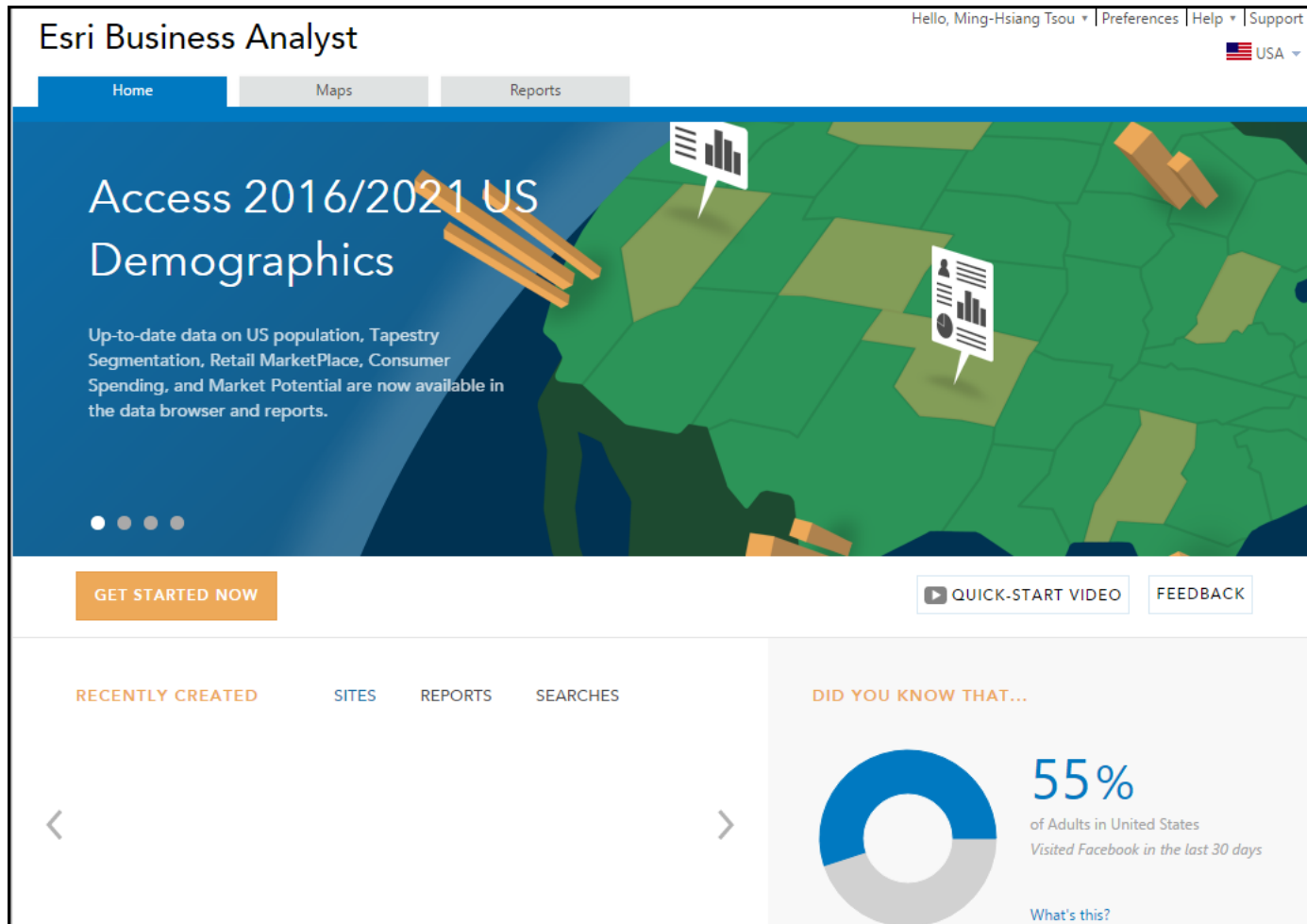
QUICK-START VIDEO | FEEDBACK

RECENTLY CREATED | SITES | REPORTS | SEARCHES

DID YOU KNOW THAT...

55%  
of Adults in United States  
Visited Facebook in the last 30 days

What's this?



## Transportation Data:

- Public NYC Taxicab Database:  
[http://www.nyc.gov/html/tlc/html/about/trip\\_record\\_data.shtml](http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) (Many transportation research papers have used this great datasets).
- NYC Open Data: <https://data.cityofnewyork.us/data?cat=transportation> (including NYC Subway Entrances).
- Bike Share Data: Capital Bikeshare (Washington DC):  
<http://www.capitalbikeshare.com/system-data> (need to install Silverlight).
- San Diego Traffic volumes: tsou  
[http://data.sandiego.gov/search/field\\_topic/transportation-611](http://data.sandiego.gov/search/field_topic/transportation-611)
- CDR data (Call detail record):  
[https://en.wikipedia.org/wiki/Call\\_detail\\_record](https://en.wikipedia.org/wiki/Call_detail_record)
  - [AirSage: http://www.airsage.com/](http://www.airsage.com/)
  - Mobile Phone flow maps: <http://www.worldpop.org.uk/ebola/>
  - Open Big Data: <https://dandelion.eu/datamine/open-big-data/>
- Bike Map: <https://bikemaps.org/>

The screenshot shows the NYC Taxi & Limousine Commission website. The main heading is "TLC Trip Record Data". Below the heading is a map of New York City with yellow and green markers representing taxi trip records. The text on the page states: "The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data." Below this, it says: "The For-Hire Vehicle ('FHV') trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary."

File size is very big  
 (One month: 1.6GB)

The taskbar shows the following download progress:

- green\_tripdata\_...csv: 22.2/220 MB, 1 min...
- yellow\_tripdata\_...csv: 0.0/1.6 GB, 17 mins...
- yellow\_tripdata\_...csv: Canceled
- nycTaxiTrip\_...torrent

Buttons: Show all downloads...

**INSIDE A CONNECTED VEHICLE**

- 1.** An under-the-hood box (a processor with memory) collects and transmits data between the vehicle's onboard equipment (OBE) and between OBE on near-by connected vehicles and safety devices along the roadside.
- 2.** A display panel, sitting in the vehicle's center console opposite the driver's dashboard, displays audio or visual safety warnings to the driver.
- 3.** A radio and antenna, using dedicated short-range communications (DSRC) and a GPS receiver, receive and transmit data about the vehicle's position to other vehicles and to safety devices along the roadway.
- 4.** Sensors collect additional information that improves the accuracy of the data being collected and transmitted by the vehicle.



- **Vehicle-to-vehicle (V2V):** Bi-directional information sharing between vehicles
- **Vehicle-to-infrastructure (V2I):** Bi-directional information sharing between a vehicle and the roadway
- **V2X (vehicle-to-everything):** Bi-directional information sharing between a vehicle and X (pedestrians, cyclists, trains, etc.)
- **Dedicated short-range communications (DSRC)**
  - Low-latency, robust, secure information (<.5 s latencies)
  - **Short range (< 300 meters)**

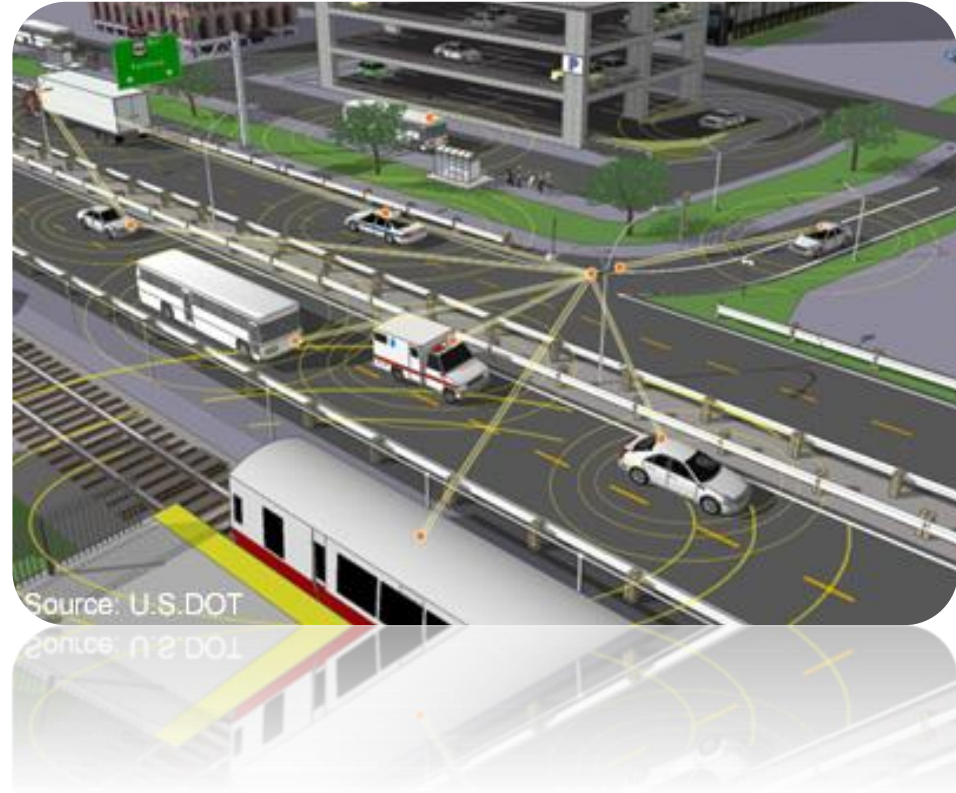
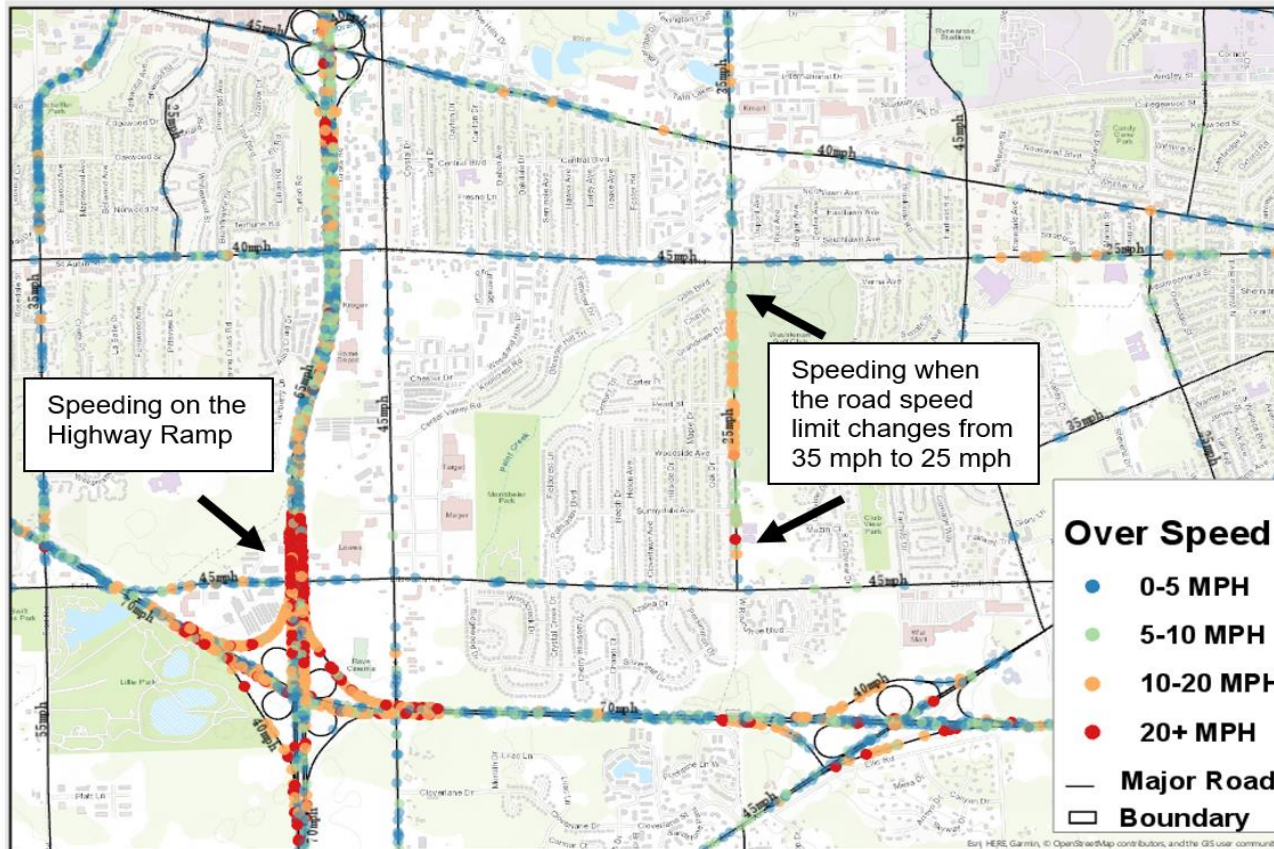


Image provided by Leslie Harwood, Virginia Tech Transportation Institute

**WHO wants to share their vehicle information?**

# Analyzing the Aggressive Driving (Speeding) Behaviors

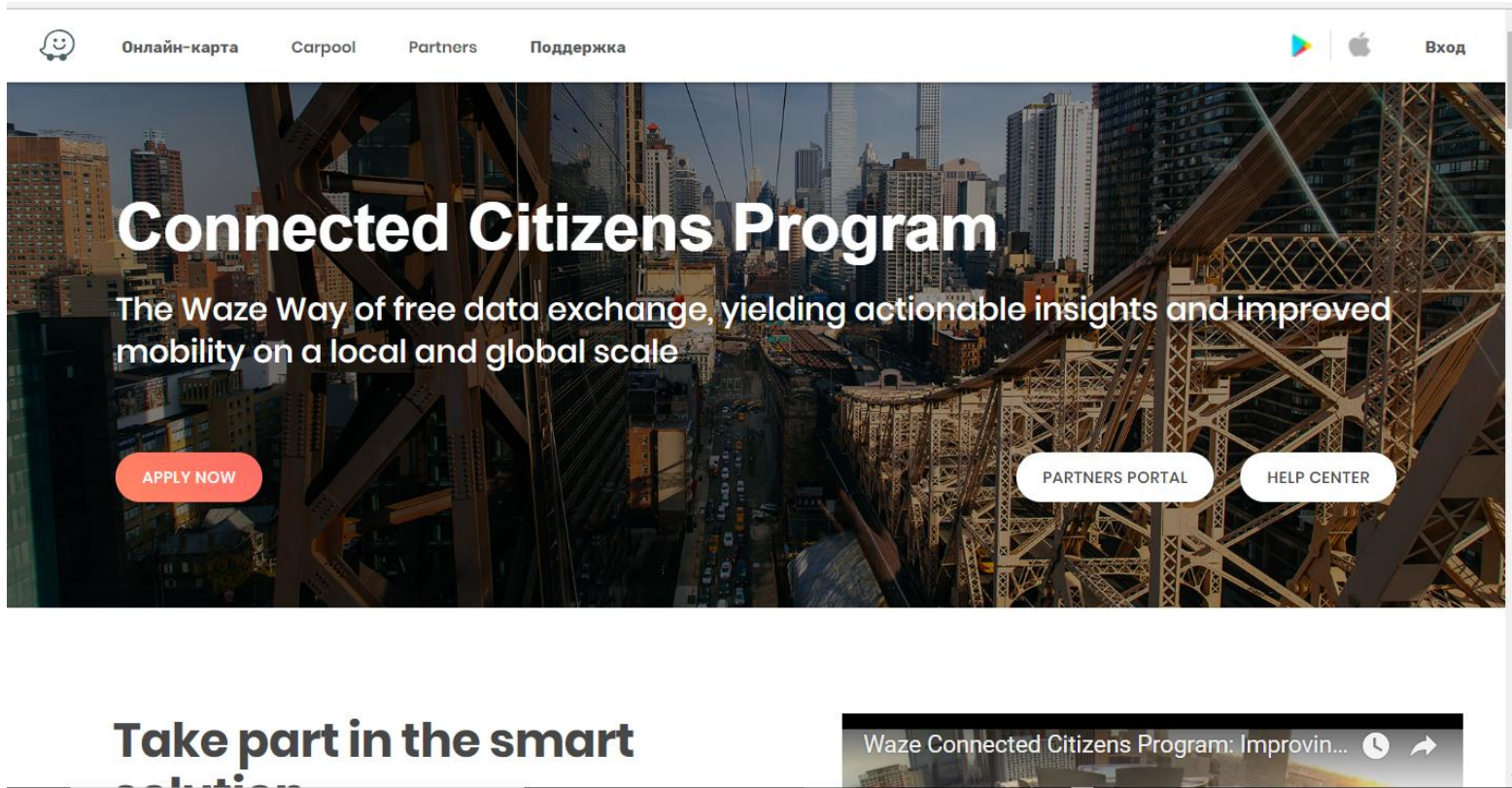


SAFE-D (2018). Big Data Visualization and Spatiotemporal Modeling of Aggressive Driving:  
 URL: <https://www.vtti.vt.edu/utc/safe-d/index.php/projects/big-data-visualization-and-spatiotemporal-modeling-of-aggressive-driving/>

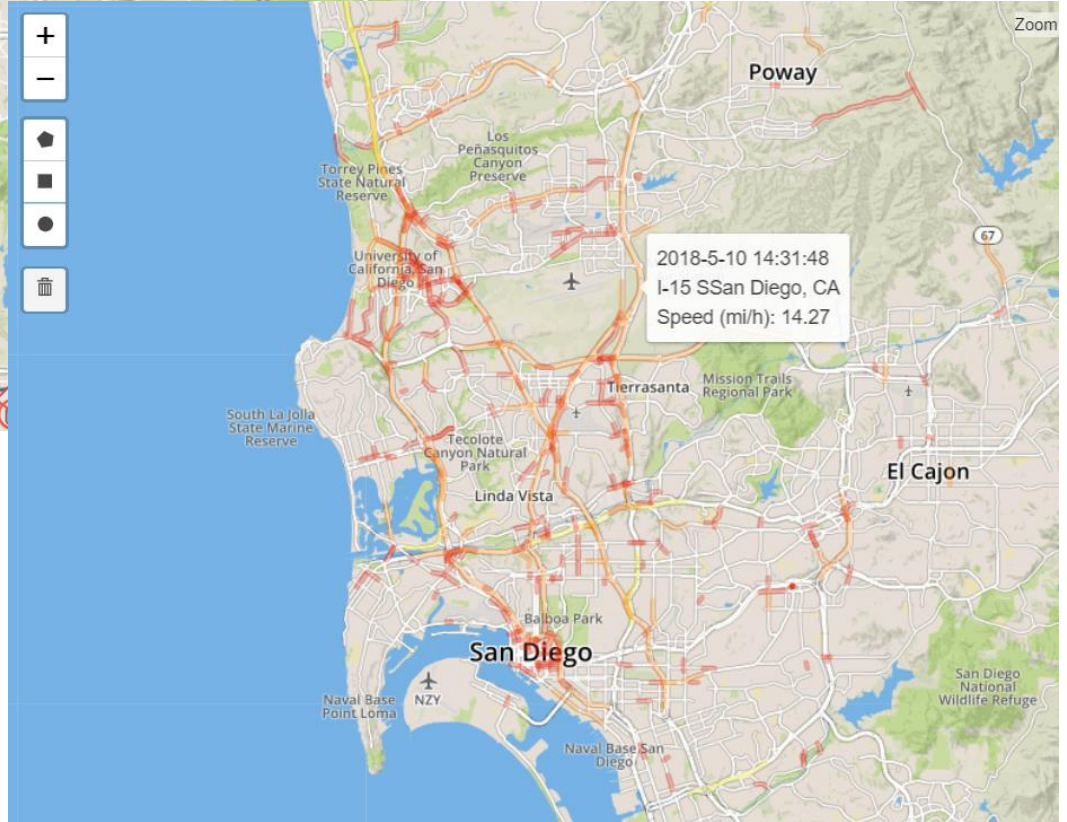
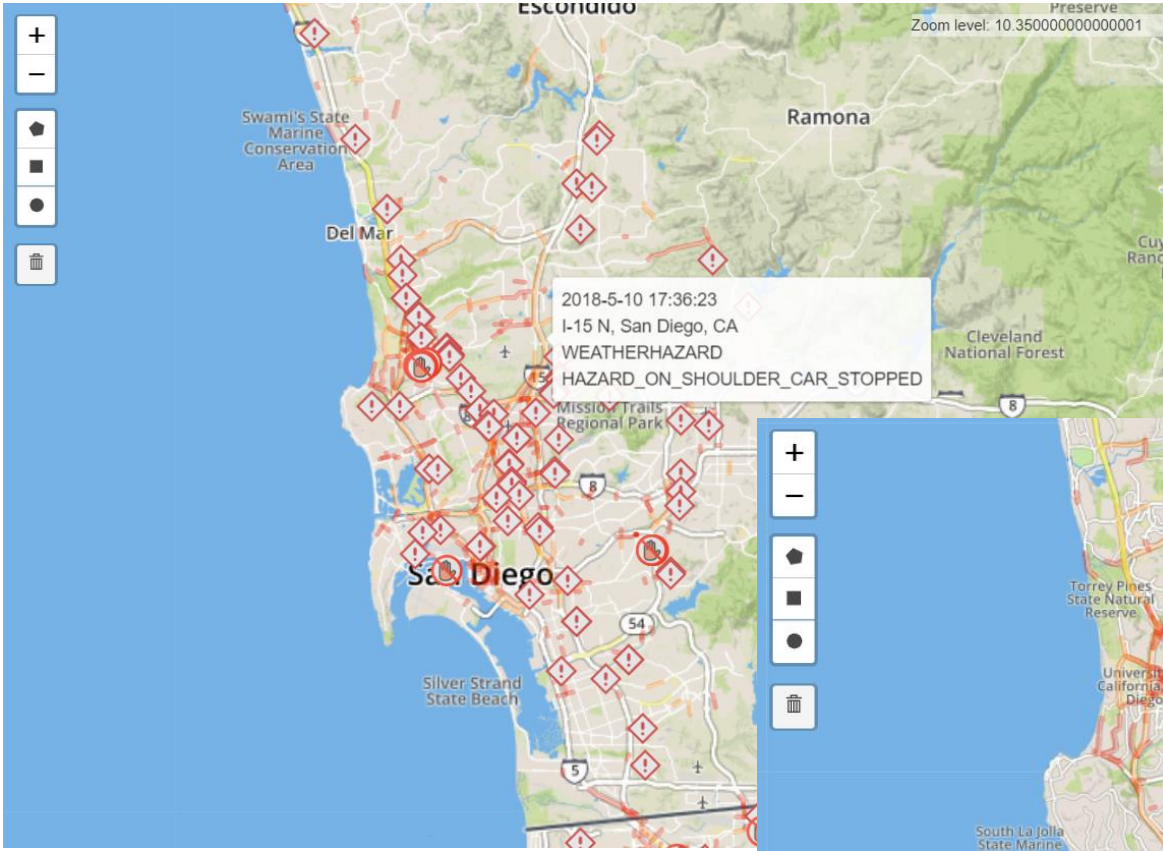


**WAZE is a “crowd-sourcing” GPS navigation software app.**

[https://wiki.waze.com/wiki/Connected\\_Citizens Program](https://wiki.waze.com/wiki/Connected_Citizens_Program)

A screenshot of the Waze Connected Citizens Program website. The page features a large, high-angle photograph of a city street with a prominent steel bridge structure. Overlaid on the image is the text "Connected Citizens Program" in a large, white, bold font. Below this, a subtitle reads "The Waze Way of free data exchange, yielding actionable insights and improved mobility on a local and global scale". There are three buttons: "APPLY NOW" in a red pill-shaped button on the left, and "PARTNERS PORTAL" and "HELP CENTER" in white pill-shaped buttons on the right. The top navigation bar includes a Waze logo, "Онлайн-карта", "Carpool", "Partners", "Поддержка", and "Вход". The bottom of the page shows the start of a video player with the title "Waze Connected Citizens Program: Improvin...".

## Waze Alerts



## Waze Jams



# Waze APIs Data Collection (Within San Diego County)

**Chart 1-** Shows the two different types of *titles* there corresponding *types* and *data formats*.

Title	Type	Data Format
Alert	ROAD_CLOSED	Point
	WEATHERHAZARD	Point
	JAM	Point
	Accident	Point
JAM	NONE	Line

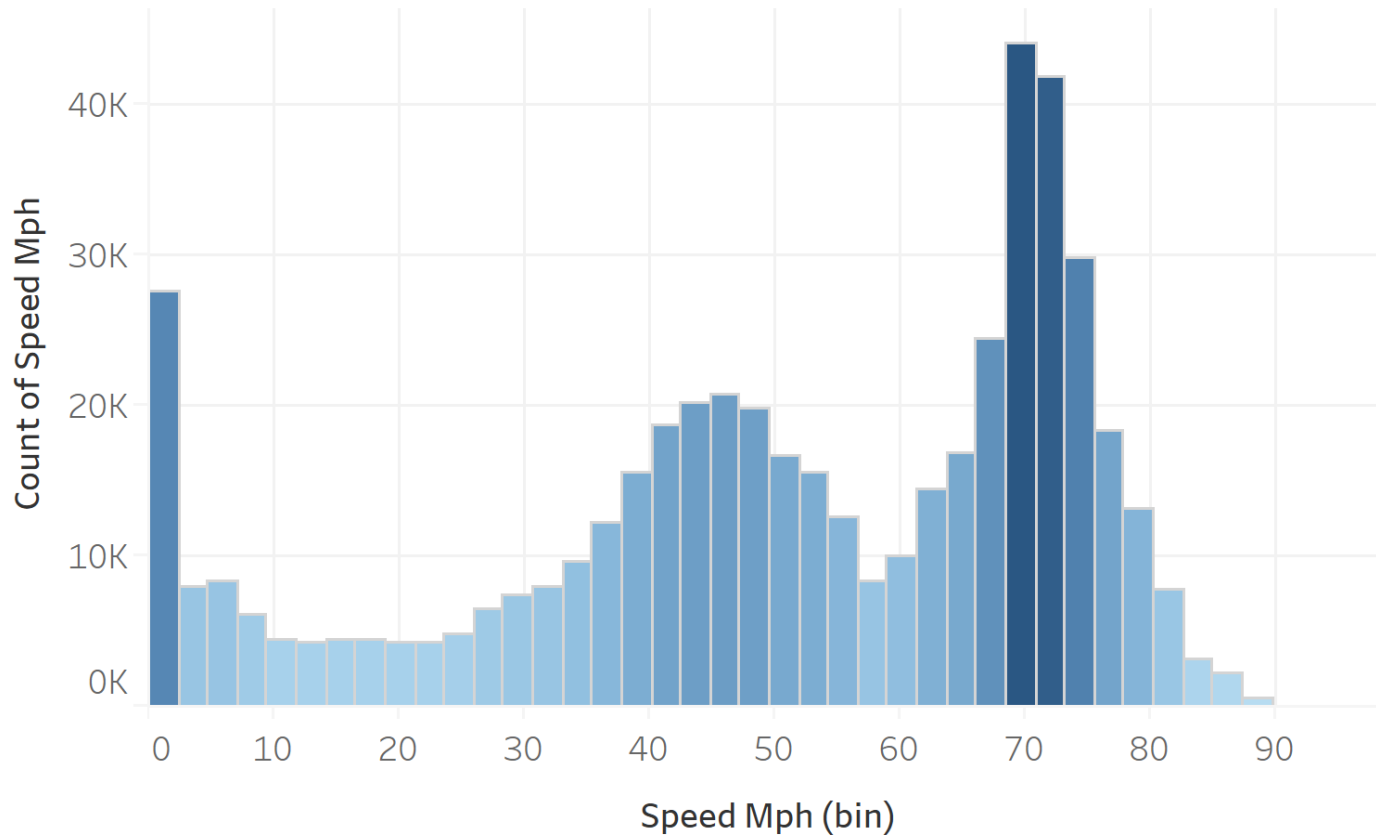
# DATASET - SPMD\_BSM\_P1\_20130415\_01GB

- Data Size: 91.0 MB
- Number of data: 500,000 observations, 24 attributes
- Feature Selection: Focus on latitude, longitude, speed, heading, yawrate, and confidence for visualization.

Field Name	Description
Speed	Vehicle speed.
Heading	Vehicle heading/direction.
Yawrate	Vehicle yaw rate.
Confidence	Signals the accuracy and non-steady state and steady state of curvature estimate. In steady state (straight roadways or curves with constant radius of curvature), a high confidence value is reported.

# FREQUENCY OF SPEED

## Frequency of Speed



CNT(Speed Mph)



3

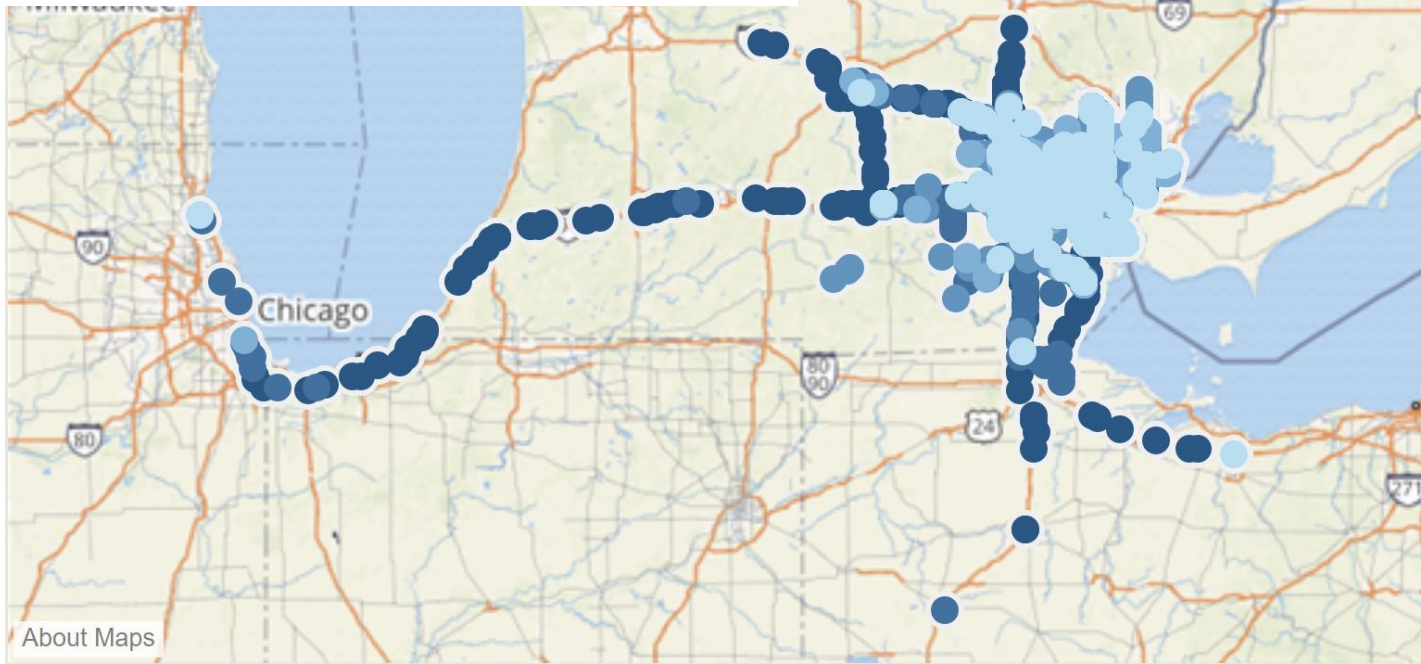
44,117

# SPEED AT DIFFERENT LOCATION

Clusters	Number of Items	Centers Speed Mph
Cluster 1	65901	32.664
Cluster 2	164771	73.62
Cluster 3	79674	62.64
Cluster 4	117097	46.961
Cluster 5	42300	9.6465

Clusters (2)





- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4
- Cluster 5



About Maps

Dandelion API

[Try me!](#)
[Docs](#)
[Pricing](#)
[FAQ](#)
[Support](#)
[Blog](#)
[Ciao mingtsou!](#)

## Open Big Data / Telecommunications - SMS, Call, Internet - MI

Description
Tabular Preview
API
Resources

### Schema

<ol style="list-style-type: none"> <li>1. <b>Square id</b>: the id of the square that is part of the <a href="#">Milano GRID</a>; TYPE: numeric</li> <li>2. <b>Time interval</b>: the beginning of the time interval expressed as the number of millisecond elapsed from the Unix Epoch on January 1st, 1970 at UTC. The end of the time interval can be obtained by adding 600000 milliseconds (10 minutes) to this value. TYPE: numeric</li> <li>3. <b>Country code</b>: the phone country code of a nation. Depending on the measured activity this value assumes different meanings that are explained later. TYPE: numeric</li> <li>4. <b>SMS-in activity</b>: the activity in terms of received SMS inside the Square id, during the Time interval and sent from the nation identified by the Country code. TYPE: numeric</li> <li>5. <b>SMS-out activity</b>: the activity in terms of sent SMS inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric</li> <li>6. <b>Call-in activity</b>: the activity in terms of received calls inside the Square id, during the Time interval and issued from the nation identified by the Country code. TYPE: numeric</li> <li>7. <b>Call-out activity</b>: the activity in terms of issued calls inside the Square id, during the Time interval and received by the nation identified by the Country code. TYPE: numeric</li> <li>8. <b>Internet traffic activity</b>: the activity in terms of performed internet traffic inside the Square id, during the Time interval and by the nation of the users performing the connection identified by the Country code . TYPE: numeric</li> </ol>	<p>tsou</p>
---	-------------

### Important notes

Files are in tsv format. If no activity was recorded for a field specified in the schema above then the corresponding value is missing from the file. For example, if for a given combination of the *Square id* *s*, the *Time interval* *i* and the *Country code* *c* no SMS was sent the corresponding record looks as follows:

```
s | i | t | c | t | SMSout | t | Callin | t | Callout | t | Internettraffic
```

where *t* corresponds to the tab character, *SMSout* is the value corresponding to the *SMS-out activity*, *Callin* is the value corresponding to the *Call-in activity*, *Callout* is the value corresponding to the *Call-out activity* and *Internettraffic* is the


These links have been generated privately for yourself, and will expire in **3 minutes**, and **54 seconds**


Resuming a download with `wget` or `curl` is easy peasy, just use the provided functionalities:

```
wget -c URL
```

```
curl -C - -O DEST_FILE URL
```

Remember that download URLs expire, but no worries: get a fresh link reloading this page, and you will be able to resume the download even with a different URL.

 sms-call-interne...zip  
1.6/79.9 MB, 42 mi...

 full.zip  
0.0/2.5 GB, 21 hour...

[Show all downloads...](#)

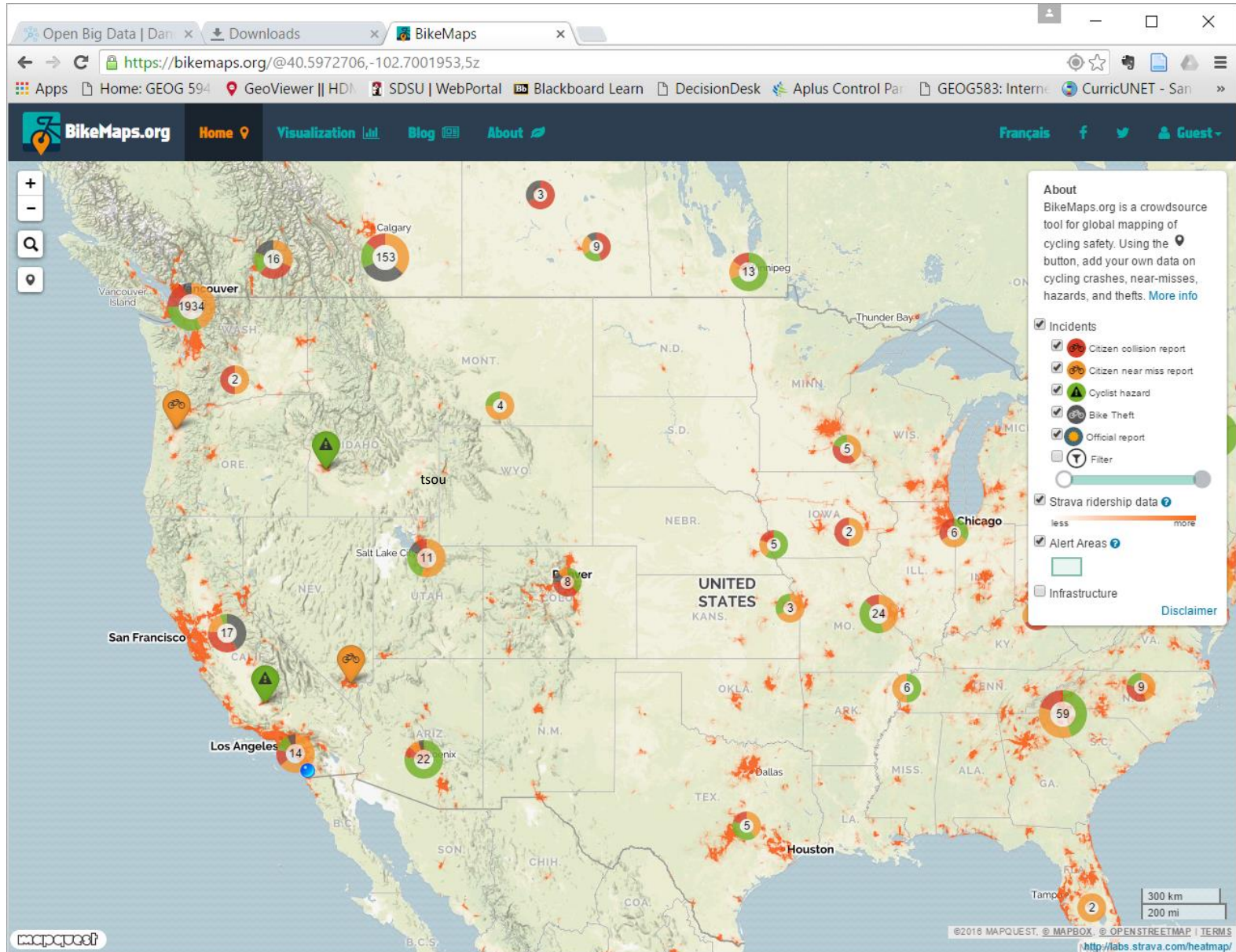
Support

Ziliang Zhao, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen & Ling Yin (2016)  
**Understanding the bias of call detail records in human mobility research**, International  
 Journal of Geographical Information Science, 30:9, 1738-1762, DOI:  
 10.1080/13658816.2015.1137298

**Table 1.** Summary of event codes.

Code	Event	Description	Avg. no. of records per subscriber
RU	Regular update	Regular update triggered by moving from the service area of a cell tower to that of another tower.	12.51
PU	Periodic update	Periodic update triggered by tower pinging if a subscriber has been 'silent' (i.e., no other events listed in this table is detected) for a certain time period. However, the specific condition (e.g., duration of silence) that triggers periodic update is irregular. In addition, mobile phones which are turned off or disconnected from the cellular network do not receive pinging signals from the cellular network.	4.88
OT	Phone communication (outbound)	Subscriber makes a phone call or sends a text message.	4.45
ON	Power on	Mobile phone is turned on and connected to cellular network.	0.62
OF	Power off	Mobile phone is turned off and disconnected from cellular network.	0.39
IN	Phone communication (inbound)	Subscriber receives a phone call or a text message.	14.67
CH	Cellular handover	Transfer of an ongoing phone call from one cell tower to another due to a subscriber's movements.	5.45





## Scientific Research Data

- **Socioeconomic Data:**

- Census Data and American Community Survey (ACS).

<https://www.census.gov/programs-surveys/acs/>

- Survey Data: National Center for Health Statistics

<https://www.cdc.gov/nchs/>

- **Censor Network Data:**

- Weather Data: U.S. National Weather Services (GIS Data portal)

<http://www.weather.gov/> , <http://www.nws.noaa.gov/gis/> (resolution 5km x 5km).

- Earthquake Data (U.S. Geological Survey)

<http://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php>

- Satellite Images (MODIS data for wildfire monitoring).

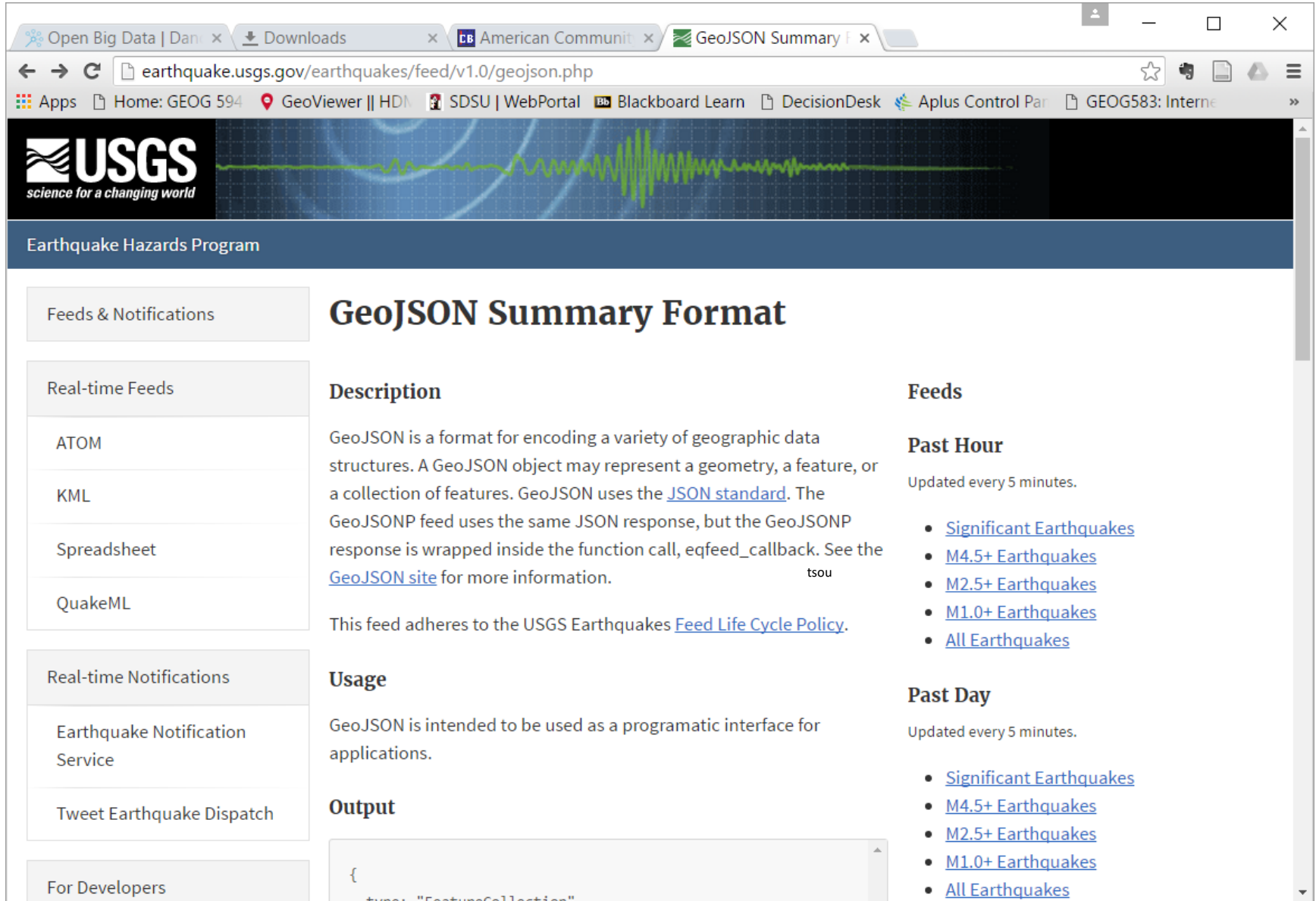
<http://activefiremaps.fs.fed.us/index.php>

- **Citizen Science Data**

- eBird: <http://ebird.org/ebird/explore>

- iNaturalist.org <http://www.inaturalist.org/> (BioBliz event)





The screenshot shows a web browser window displaying the USGS Earthquake Hazards Program GeoJSON Summary Format page. The browser tabs include 'Open Big Data | Dan...', 'Downloads', 'American Communit...', and 'GeoJSON Summary F...'. The address bar shows the URL 'earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php'. The browser's address bar also shows several bookmarks: 'Home: GEOG 594', 'GeoViewer || HD...', 'SDSU | WebPortal', 'Blackboard Learn', 'DecisionDesk', 'Aplus Control Par', and 'GEOG583: Intern...'. The page header features the USGS logo with the tagline 'science for a changing world' and a green seismic waveform graphic. Below the header is a dark blue navigation bar with the text 'Earthquake Hazards Program'. The main content area is divided into a left sidebar and a main content column. The sidebar contains several menu items: 'Feeds & Notifications', 'Real-time Feeds' (with sub-items: ATOM, KML, Spreadsheet, QuakeML), 'Real-time Notifications' (with sub-items: Earthquake Notification Service, Tweet Earthquake Dispatch), and 'For Developers'. The main content column has a large heading 'GeoJSON Summary Format'. Below this heading are three sections: 'Description', 'Usage', and 'Output'. The 'Description' section explains that GeoJSON is a format for encoding geographic data structures and mentions the 'JSON standard' and 'GeoJSON site'. The 'Usage' section states that GeoJSON is intended to be used as a programmatic interface for applications. The 'Output' section shows a code block with a JSON object structure. To the right of the main content are two sections: 'Feeds' and 'Past Hour'. The 'Feeds' section lists several links: 'Significant Earthquakes', 'M4.5+ Earthquakes', 'M2.5+ Earthquakes', 'M1.0+ Earthquakes', and 'All Earthquakes'. The 'Past Hour' section indicates that the data is updated every 5 minutes and lists the same set of links as the 'Feeds' section. The 'Past Day' section also indicates that the data is updated every 5 minutes and lists the same set of links.

Open Big Data | Dan
Downloads
American Communit
Active Fire Mapping

activefiremaps.fs.fed.us/index.php

Apps
Home: GEOG 594
GeoViewer || HDN
SDSU | WebPortal
Blackboard Learn
DecisionDesk
Aplus Control Par
GEOG583: Intern
CurricUNET - San

USDA FOREST SERVICE
REMOTE SENSING APPLICATIONS CENTER

## Active Fire Mapping Program

- Current Large Incidents (Home)
- New Large Incidents
- Fire Detection Maps
- MODIS Satellite Imagery
- VIIRS Satellite Imagery
- Fire Detection GIS Data
- Fire Data in Google Earth
- Fire Data Web Services
- Latest Detected Fire Activity
- Other MODIS Products
- Frequently Asked Questions
- About Active Fire Maps

**Remote Sensing Applications Center**

2222 West 2300 South  
Salt Lake City, UT  
84119 - 2020

voice: (801) 975-3737  
fax: (801) 975-3478

## Active Fire Mapping Program

Fire locations are based on data provided by the National Interagency Coordination Center and are subject to change.

### Current Large Incidents September 04, 2016

① TULLEY	⑦ RAIL	⑬ GRAPE	⑰ BERRY
② GAP	⑧ CAYUSE MTN.	⑭ WEST GOVERNMENT CREEK	⑱ BROADWAY
③ SOBRANES	⑨ MAGGIE	⑮ RATTLESNAKE	⑲ BEAVER CREEK
④ CHIMNEY	⑩ PIONEER	⑯ PETERSON HOLLOW	⑳ GRIFFEN GULCH
⑤ REY	⑪ COPPER KING	⑰ TIE	
⑥ CEDAR	⑫ CLEAR CREEK	⑱ MAPLE	

**IMSR Summary**  
August 29th, 2016

**National Preparedness Level**

Level 4  
National Fire Activity  
Initial attack activity: Light (78 new fires)  
New large incidents: 5  
Large fires contained: 3  
Uncontained large fires: 29  
Area Command Teams Committed: 0  
NIMOs committed: 1  
Type 1 IMTs committed: 8  
Type 2 IMTs committed: 12

Source:  
[Incident Management Situation Report](#)

**Active Fire Mapping News**  
August 24, 2016

**Data Access Alert:** Due to the significant wildfire activity, access to WMS and WFS services will be restricted to ensure their availability for operational fire management needs and maintain availability of other active fire mapping/data products.

[View map with Greater Sage-Grouse habitat layer.](#)



Open Big Data | Downloads | Explore Hotspots

ebird.org/ebird/hotspots#

Home: GEOG 594 | GeoViewer | SDSU | WebPortal | Blackboard Learn | DecisionDesk | Apluser Control Panel | GEOG583: Intern...

Submit Observations | Explore Data | My eBird | Help | Sign In or Register | Language

Hotspot:  Date: Year-round, All Years Location:

**Famosa Slough**  
San Diego, US-CA

Year-round, All Years

**213** SPECIES | **1357** CHECKLISTS

Bar Charts | High Counts | Directions

Submit Data | View Details

Close All Windows | Send Feedback | Zoom Tool

MAP TYPE  
 Street  
 Terrain  
 Satellite  
 Hybrid

FILTER BY RECENT ACTIVITY  
 All Hotspots  
 Past Month  
 Past Week

SPECIES OBSERVED  
 500+  
400  
300  
250  
200  
150  
100  
50  
15  
1  
0

tsou

# eBird Hotspots

<http://ebird.org/ebird/hotspots#>

Famosa Slough | eBird

SDSU | WebPortal | Blackboard Learn | DecisionDesk | Apluser Control Panel | GEOG583: Intern...

Sign In or Register | Language

< Hotspot Map

**Famosa Slough**  
San Diego County, California, US — Get Directions

All Months | All Years | **Set** | Submit Data

Overview | Recent Visits

**213** Species | 1357 Checklists Updated 10 sec ago.

Last Seen | First Seen | High Counts | Bar Charts | Printable Checklist | Show All Details

	SPECIES NAME	COUNT	DATE	BY
1	Gadwall	2	2 Sep 2016	John Bruin
2	American Wigeon	4	2 Sep 2016	John Bruin
3	Mallard	4	2 Sep 2016	John Bruin
4	Blue-winged Teal	9	2 Sep 2016	John Bruin
5	Northern Shoveler	1	2 Sep 2016	John Bruin
6	Double-crested Cormorant	3	2 Sep 2016	John Bruin
7	Great Blue Heron	4	2 Sep 2016	John Bruin
8	Great Egret	1	2 Sep 2016	John Bruin
9	Snowy Egret	8	2 Sep 2016	John Bruin
10	Little Blue Heron	1	2 Sep 2016	John Bruin

Every bird counts. Be a part of it.

**GLOBAL BIG DAY**

**MAY 14 2016**

Learn More

Recent Visits Checklists submitted within the last hour are not shown.

OBSERVER	DATE	SPECIES
John Bruin	2 Sep 2016	24
Johnny Galt	2 Sep 2016	16
Sam Fellows	1 Sep 2016	23
Bill Truitt	30 Aug 2016	22



- **JSON (JavaScript Object Notation)** is a lightweight **data-interchange format**. It is easy for **humans to read** and write. It is **easy for machines to parse** and generate. (Better than XML – more readable) (used for asynchronous browser/server communication (AJAJ) file extension “.json” (<http://www.json.org/> and wikipedia).
- **What is “GeoJSON”? Geo + JSON**
- GeoJSON is a geospatial **data interchange format** based on JavaScript Object Notation (JSON). It defines several types of JSON objects and the manner in which they are combined to **represent data about geographic features, their properties, and their spatial extents**. GeoJSON uses a geographic coordinate reference system, **World Geodetic System 1984**, and **units of decimal degrees**.  
<http://geojson.org/>
- WGS84 = used by all GPS devices (different from traditional GIS: NAD83)

In JSON, they take on these forms:

**An *object*** is an unordered set of name/value pairs. An object begins with **{ (left brace)** and ends with **} (right brace)**. Each name is followed by **: (colon)** and the name/value pairs are separated **by , (comma)**.

tsou

**An *array*** is an ordered collection of values. An array begins with **[ (left bracket)** and ends with **] (right bracket)**. Values are separated by **(comma)**.

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```



## A GeoJSON feature collection:

```
{ "type": "FeatureCollection",
  "features": [
    { "type": "Feature",
      "geometry": { "type": "Point", "coordinates": [102.0, 0.5] },
      "properties": { "prop0": "value0" }
    },
    { "type": "Feature",
      "geometry": {
        "type": "LineString",
        "coordinates": [
          [102.0, 0.0], [103.0, 1.0], [104.0, 0.0], [105.0, 1.0]
        ]
      },
      "properties": {
        "prop0": "value0",
        "prop1": 0.0
      }
    },
    { "type": "Feature",
      "geometry": {
        "type": "Polygon",
        "coordinates": [
          [ [100.0, 0.0], [101.0, 0.0], [101.0, 1.0],
            [100.0, 1.0], [100.0, 0.0] ]
        ]
      },
      "properties": {
        "prop0": "value0",
        "prop1": { "this": "that" }
      }
    }
  ]
}
```

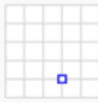
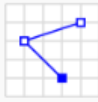
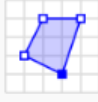
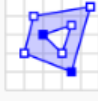


## Geometries [\[ edit \]](#)

New Standard: August 2016 (replacing 2008 specification).

<https://tools.ietf.org/html/rfc7946>

### Geometry primitives

Type	Examples	
Point		<pre>{ "type": "Point",   "coordinates": [30, 10] }</pre>
LineString	 <small>tsou</small>	<pre>{ "type": "LineString",   "coordinates": [     [30, 10], [10, 30], [40, 40]   ] }</pre>
Polygon		<pre>{ "type": "Polygon",   "coordinates": [     [[30, 10], [40, 40], [20, 40], [10, 20], [30, 10]]   ] }</pre>
		<pre>{ "type": "Polygon",   "coordinates": [     [[35, 10], [45, 45], [15, 40], [10, 20], [35, 10]],     [[20, 30], [35, 35], [30, 20], [20, 30]]   ] }</pre>

## Latitude, Longitude

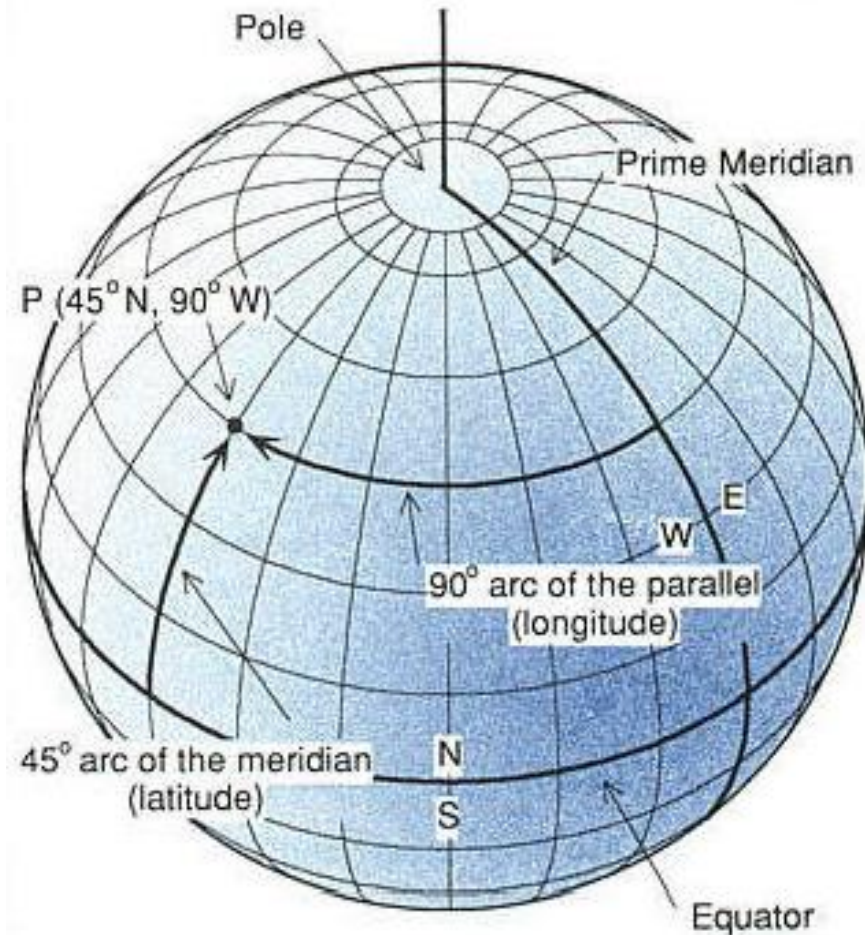
**32° 20' 15" N (North)**

**130° 42' 30" W (West)**

Angular Coordinate System:

**Degree, Minutes, Seconds**

- 360 degree in a circle tsou
- 1 degree = 60 minutes
- 1 minute = 60 seconds
- Longitude: 0 to 180 east and west
- Latitude: 0 to 90 north and south
- Circumference of the earth = 24,900 miles (40,075 km) at the equator
- **130° 42' 30" W = 130.70833 (decimal degree)**



How to convert from a degree/minutes/second format to a decimal degree format? (positive or negative numbers?)

**Latitude: N (+), S (-), Longitude: E (+), W (-)**

**130° 42' 30 " W (West). = - 130.70833.**

1. Convert the [seconds] to minutes:  $30''$  (seconds) =  $30 / 60 = 0.5'$  (minute)

2. Add the value (0.5) back to the minutes (42).  $42 + 0.5 = 42.5$  (minutes)

3. Convert the [minutes] to [degree]:  $42.5'$  (minutes) =  $42.5 / 60 = 0.70833$  (degree).

4. Add the result (0.70833) to the degree number (130):  $130 + 0.70833 = 130.70833$  (degree).

5. Since the longitude is West. The value of the decimal degree will be negative --> **- 130.70833**

**130° 42' 30 " W (West). = - 130.70833 (degree)**

## When you try to process **Decimal Degree Data**:

1. Which format? “**Latitude, Longitude**” format (for Web Map, Google Maps, Twitter GEO), or “**Longitude, Latitude**” format (for GIS software, **GeoJSON**, Twitter Coordinates, and **KML** use Long/Lat).
2. Project Datum should be **WGS84** (default GPS data settings for Datum). If other GIS data uses NAD83 (another popular projection datum), you will see the data location shifted by 1 or 2 meters.
3. Use **Web Mapping Tools** or **GIS** software (Google Maps, GeoJSON + Leaflet, MapBox, CartoDB, ArcGIS Online, StoryMaps, etc.).

## Big Data Sampling Problems, Biases, and Noises

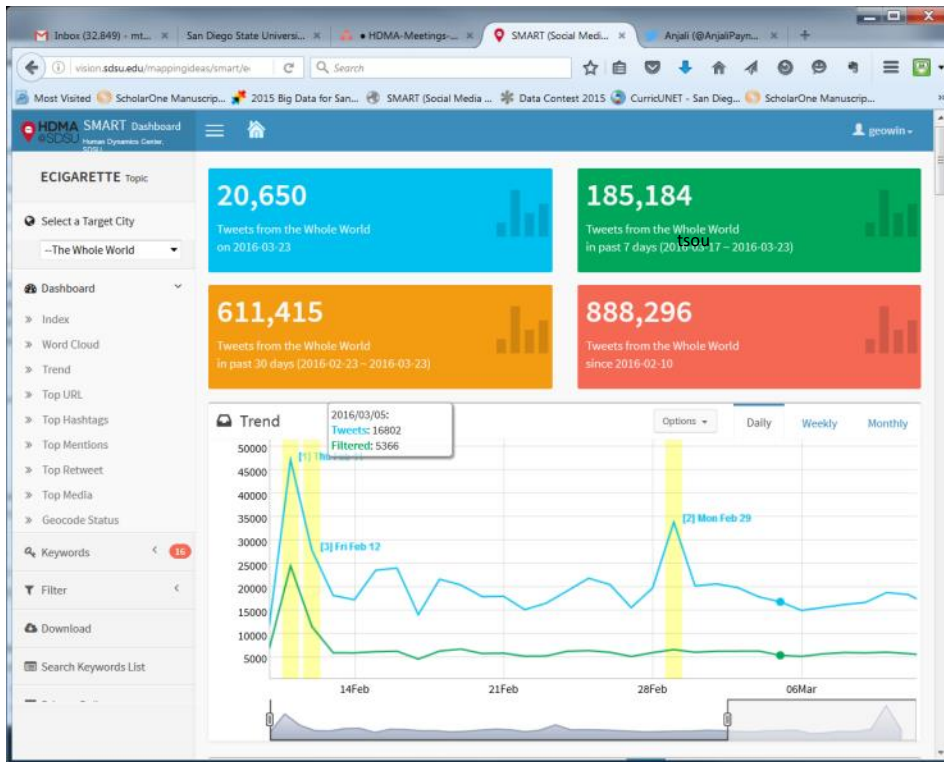
Sometimes, it is difficult to define “Noises” and “Errors” in Big Data Analytics<sup>tsou</sup>. Different Tasks and Goals will define different criteria for “Noises” and “Errors”.

**Someone’s trash might be someone’s treasure.**

# Who are the “Noises” or “Errors”? Humans or robots (bots)?

## Use SMART dashboard to track “E-cigarette” topics

### High Peak on Feb 11, 2016 (Why?)



**Anjali** @AnjaliPayne

i saw 4 ppl from my school at the mall and they were all at the vape pens section i gtg!!!!

RETWEET 1 LIKES 5

**Meagan** @Meagan\_Hardin

i saw 4 ppl from my school at the mall and they were all at the vape pens section i gtg!!!!

RETWEET 1

8:59 PM - 11 Feb 2016

**Kelli** @Kelli\_Mosley

i saw 4 ppl from my school at the mall and they were all at the vape pens section i gtg!!!!

RETWEET 1 LIKES 5



# 1,553 Twitter Accounts

## Said the Exact Sentence! In One Day (2/11/2016),

From 11114 to 9561 = 1553 (Mummy or Ghost Twitter Accounts?) for Advertisement?

### Tweets on 2016-2-11 about eCigarette

Showing 4 to 17 of 1,000 entries (Only show Top 1000 entries)

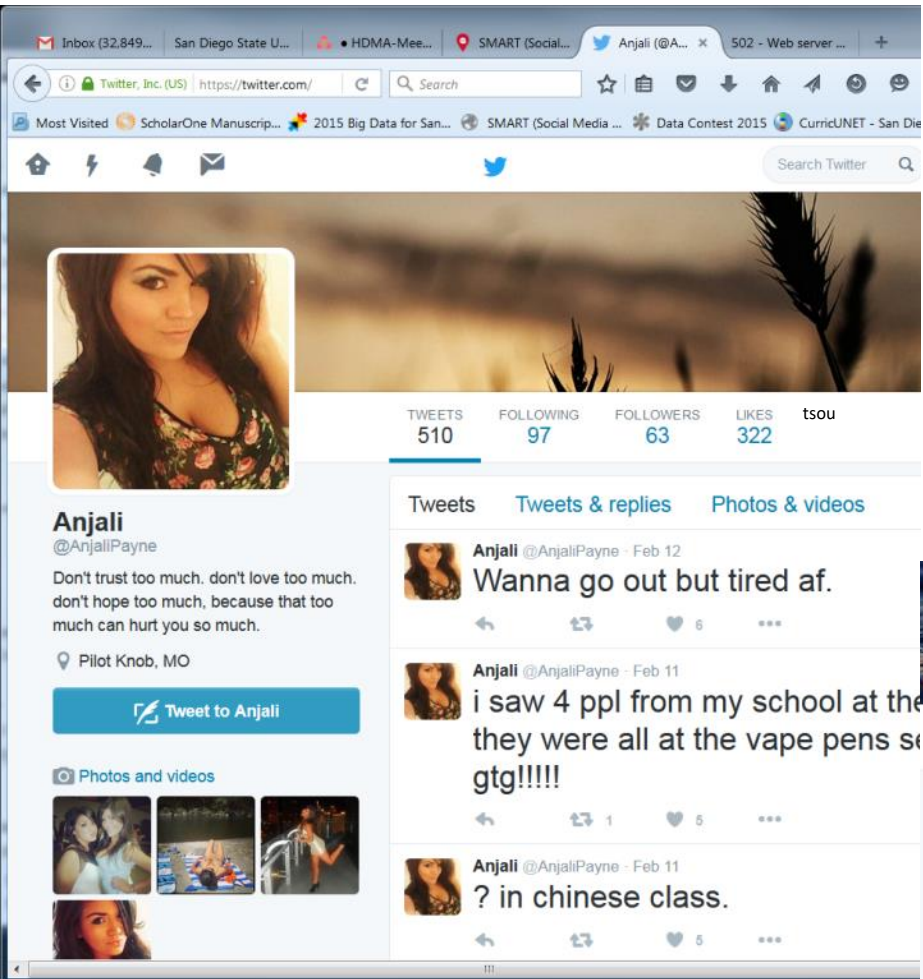
Download

Filter data results by keyword

#	CREATED_AT_LOCAL	USERNAME	TEXT
4	2016-02-11 23:59:35	Anjali	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
5	2016-02-11 23:59:34	janny hass ars	Why am I hanging in a Vape shop? #WeirdQuestionsToAskGod
6	2016-02-11 23:59:32	Ximena Fischer	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
7	2016-02-11 23:59:29	Chrissy Larsen	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
8	2016-02-11 23:59:29	Roberta_mhmmm	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
9	2016-02-11 23:59:26	Kelli	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
10	2016-02-11 23:59:19	Zelida	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
11	2016-02-11 23:59:18	Adelina Rios	when in doubt Vape it out
12	2016-02-11 23:59:17	Tanner Martin	Apparently the key to getting women is Vaping.. It's apparently the cool things the kids are doing
13	2016-02-11 23:59:16	Rania	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!
14	2016-02-11 23:59:08	Nala	i saw 4 ppl from my school at the mall and they were all at the Vape pens section i gtg!!!!

# Are They “Mummies and Ghosts (Zombie)” ?

## Who are they? How they post the messages?

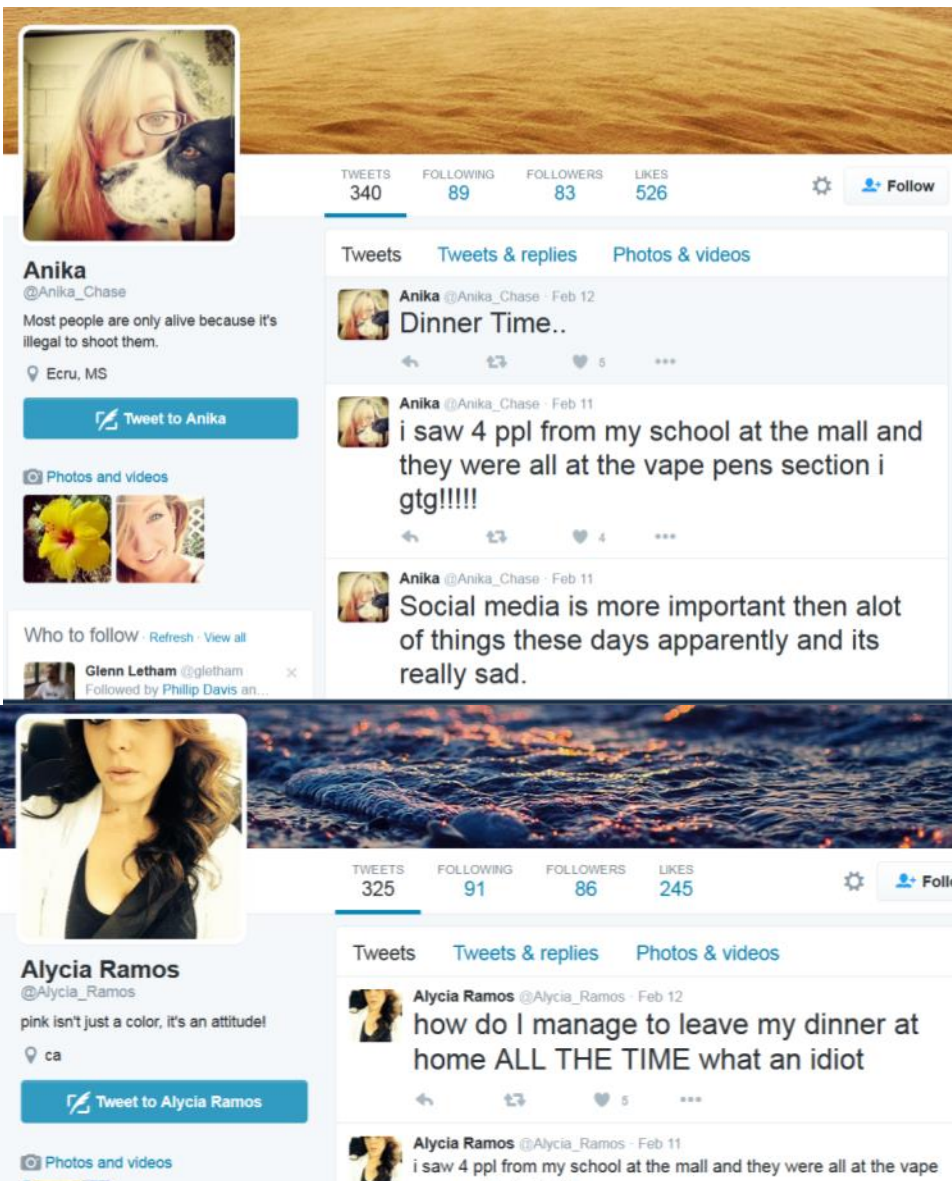


**Anjali**  
@AnjaliPayne  
Don't trust too much. don't love too much. don't hope too much, because that too much can hurt you so much.  
Pilot Knob, MO

TWEETS 510 FOLLOWING 97 FOLLOWERS 63 LIKES 322

**Tweets** Tweets & replies Photos & videos

- Anjali @AnjaliPayne - Feb 12  
Wanna go out but tired af. (6 likes)
- Anjali @AnjaliPayne - Feb 11  
i saw 4 ppl from my school at the they were all at the vape pens se gtg!!!! (5 likes)
- Anjali @AnjaliPayne - Feb 11  
? in chinese class. (5 likes)



**Anika**  
@Anika\_Chase  
Most people are only alive because it's illegal to shoot them.  
Ecu, MS

TWEETS 340 FOLLOWING 89 FOLLOWERS 83 LIKES 526

**Tweets** Tweets & replies Photos & videos

- Anika @Anika\_Chase - Feb 12  
Dinner Time.. (5 likes)
- Anika @Anika\_Chase - Feb 11  
i saw 4 ppl from my school at the mall and they were all at the vape pens section i gtg!!!! (4 likes)
- Anika @Anika\_Chase - Feb 11  
Social media is more important then alot of things these days apparently and its really sad.

**Alycia Ramos**  
@Alycia\_Ramos  
pink isn't just a color, it's an attitude!  
ca

TWEETS 325 FOLLOWING 91 FOLLOWERS 86 LIKES 245

**Tweets** Tweets & replies Photos & videos

- Alycia Ramos @Alycia\_Ramos - Feb 12  
how do I manage to leave my dinner at home ALL THE TIME what an idiot (5 likes)
- Alycia Ramos @Alycia\_Ramos - Feb 11  
i saw 4 ppl from my school at the mall and they were all at the vape pens section i gtg!!!!

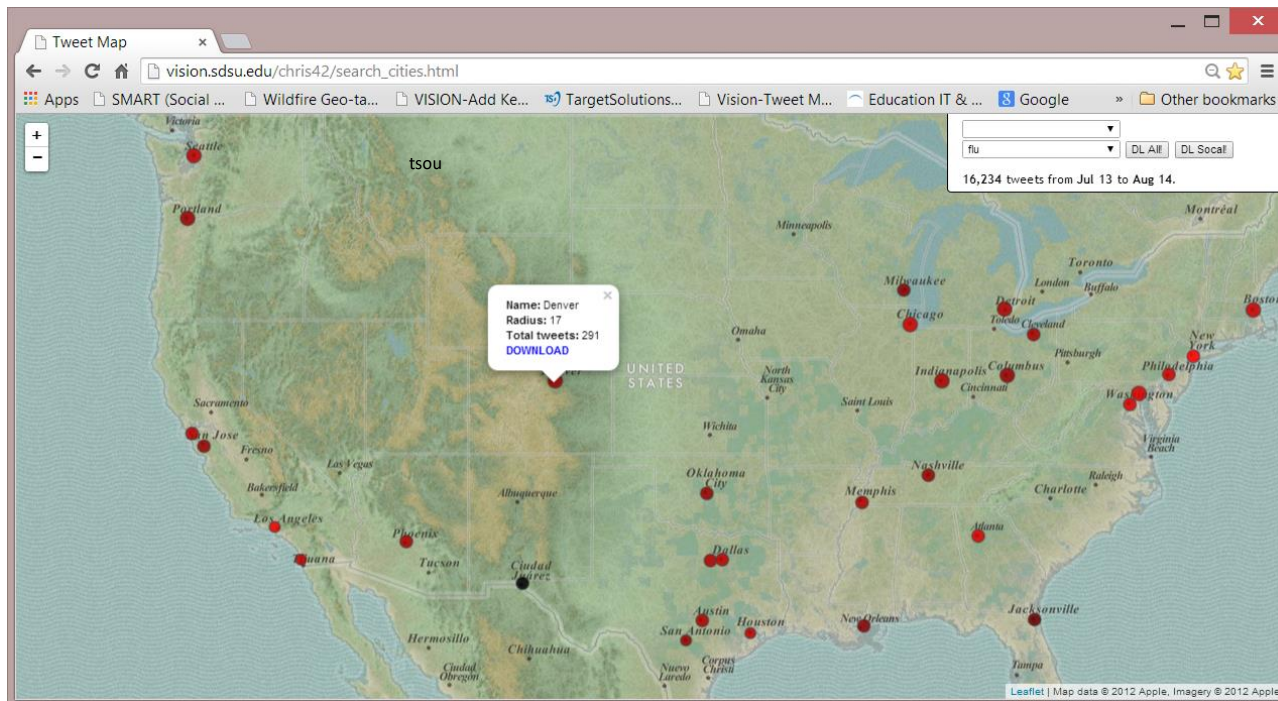
## Data Filtering and Data Process (Removing Noises).

- Should we remove these “bots” accounts and their tweets from our data analysis? Why? Why Not?
- Which regions will you analysis focus on? The whole world? Or U.S. or just California? (Regional selection).
- When ? Temporal selection.<sup>tsou</sup>

# Monitoring Flu Outbreaks in U.S. (using Twitter Messages)

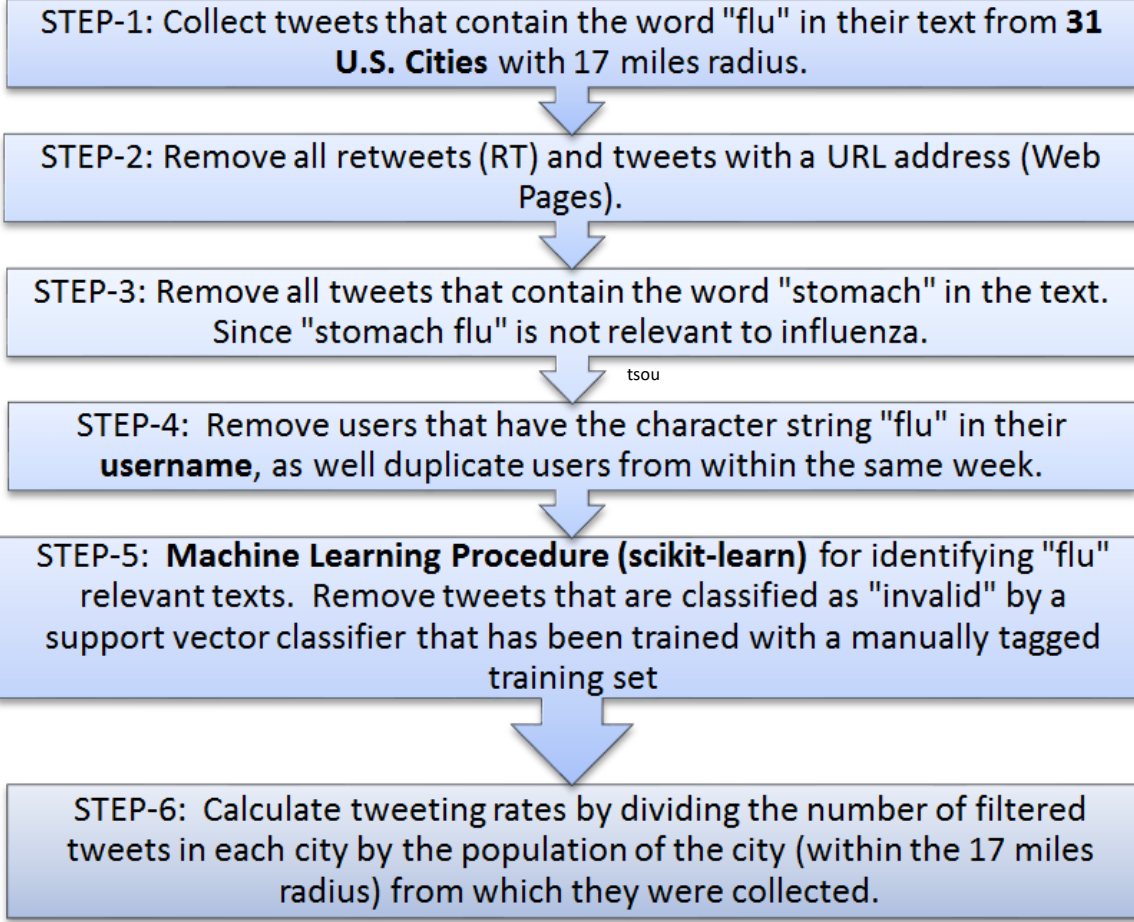
Collect Tweets from Top 31 U.S. Cities (**17 miles radius**) with “flu” and “influenza” keyword search.

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, Los Angeles, 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.*





KEYWORD	CITY	CREATED_AT_LOCAL	TEXT	LOCATION	URLS	HASHTAGS	FOLLOWER	FRIENDS	STATUS	TIME_ZONE
flu	San_Diego	2013-12-02 00:20:28	RT @grobbins: @SDSU monitoring flu outbreaks via Twitter. http://bit.ly/18Tutsandiego_sdsu	Mission Viejo, CA	http://bit.ly/18Tutsandiego_sdsu		46	78	1636	Pacific Time
flu	San_Diego	2013-12-01 23:10:12	This is what I get for not getting my flu shot on Tuesday. -_-	Sunny San Diego			25	57	902	Pacific Time
flu	San_Diego	2013-12-01 22:32:54	I never catch a cold or the flubut catchy tunes, the most contagiou	San Diego			528	961	1932	Pacific Time
flu	San_Diego	2013-12-01 22:08:31	Flu vaccine and the flu! Rishi is a pediatric infectious disease phy	San Diego, CA	http://www.khanacademy.org/vic		825	918	2769	Pacific Time
flu	San_Diego	2013-12-01 21:26:22	SDSU monitoring flu via Twitter: Researchers are looking for quic	San Diego, Calif	http://q.gs/58S_sandiego		588	737	74052	Arizona
flu	San_Diego	2013-12-01 21:08:00	@SDSU monitoring flu outbreaks via Twitter. http://t.co/a8CmkKP	San Diego, Calif	http://bit.ly/18Tutsandiego_sdsu		2632	298	7372	Pacific Time
flu	San_Diego	2013-12-01 18:21:35	RT @swineflu911: Bird Flu Vaccine Approved By FDA: First Adju	VIRAL - Disseminate Globally			199	1397	1481	Pacific Time
flu	San_Diego	2013-12-01 13:33:18	No hangover but I feel like I have the flu. #sweet	San Diego		sweet	82	130	642	
flu	San_Diego	2013-11-30 20:13:24	My flu symptoms are back.... C'mon I just started feeling better! ☺☺☺				306	299	5443	Pacific Time
flu	San_Diego	2013-11-29 16:48:51	I feel the flu coming in.... this can't be!!! There's Finals coming up!	San Diego, CA			323	326	28780	Pacific Time



Number of tweets
10,678
5,398
4,947
4,944
3279

**Machine Learning**

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

## Questions:

- **When should we remove “RT” (Retweets)? When should we keep “RT”?**
- **When should we remove “URL”? When should we keep “URL”?**  
tsou
- **How will you define other data filtering procedures?**
- **Verify the actual messages to create these additional rules.**



## Real-Time Monitoring of Flu Outbreaks in U.S.

(National Scale – combined 31 Cities), 2013 – 2014 flu season

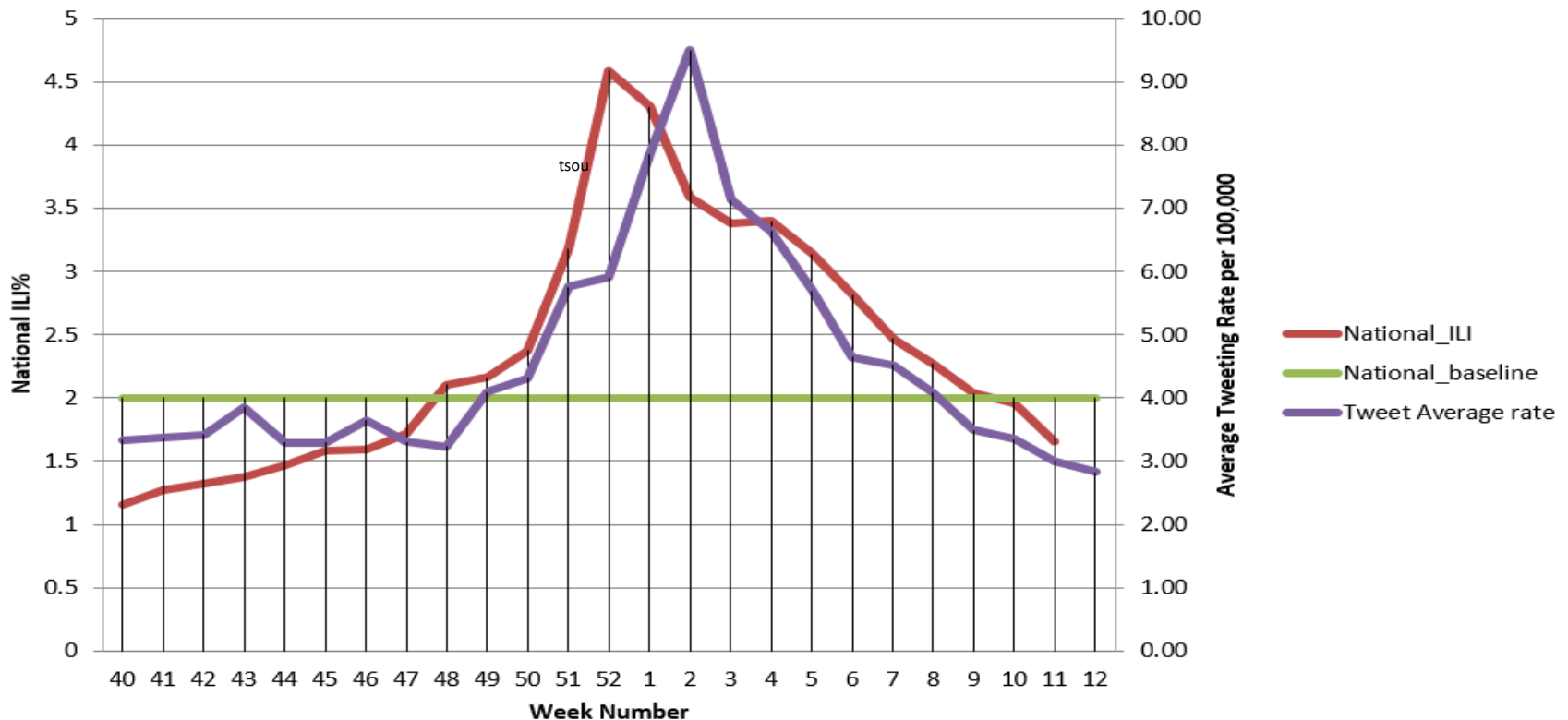
**RED Line: National ILI data (Influenza-like illness) (provided by CDC)**

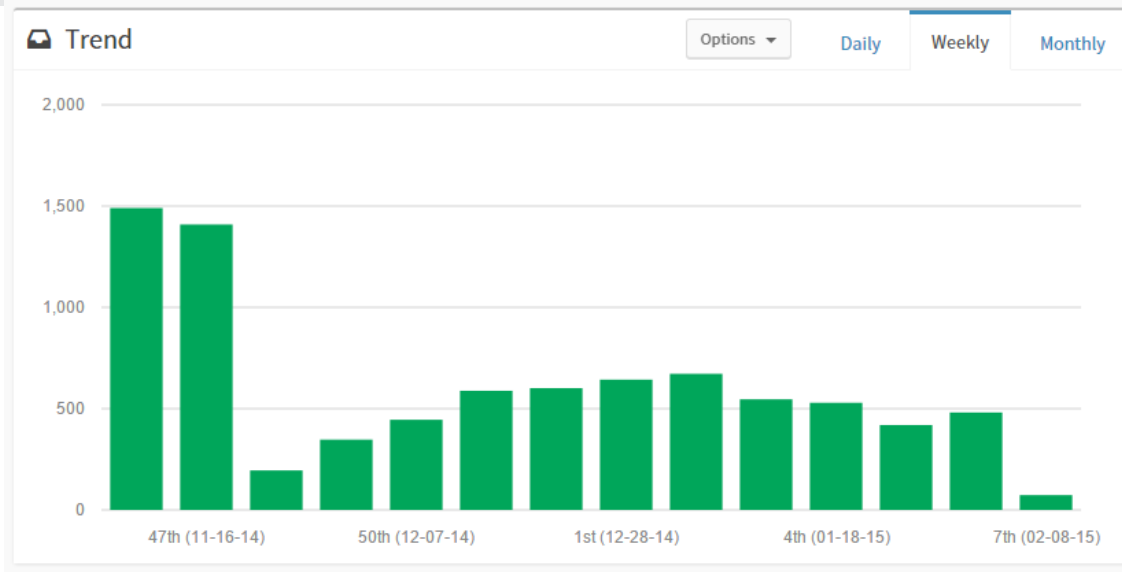
**Purple Line: Weekly Tweeting Rate (two weeks earlier than CDC data)**

ILI: Influenza-like Illness

(R) value = **0.8494**

**Average Tweeting Rate and National ILI**





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

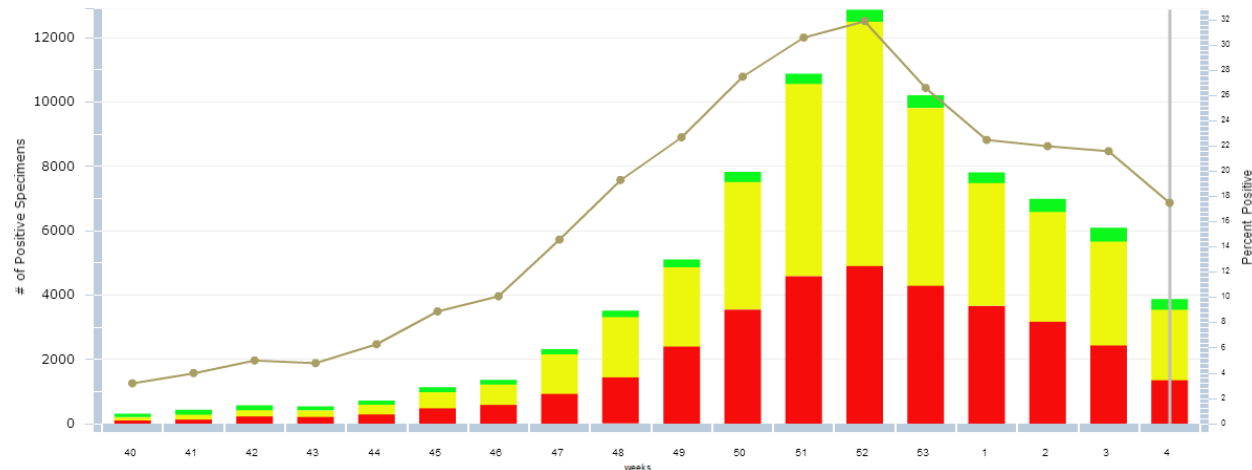
Only 1% -4% tweets has Geo-tagged coordinates.



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

tsou

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



# 2014-2015 Comparison between ILI and Geo-tagged-only Tweets (4%) among 30 U.S. Cities

ILI Activity, Week 13 Update, CDC Data and SMART Dashboard

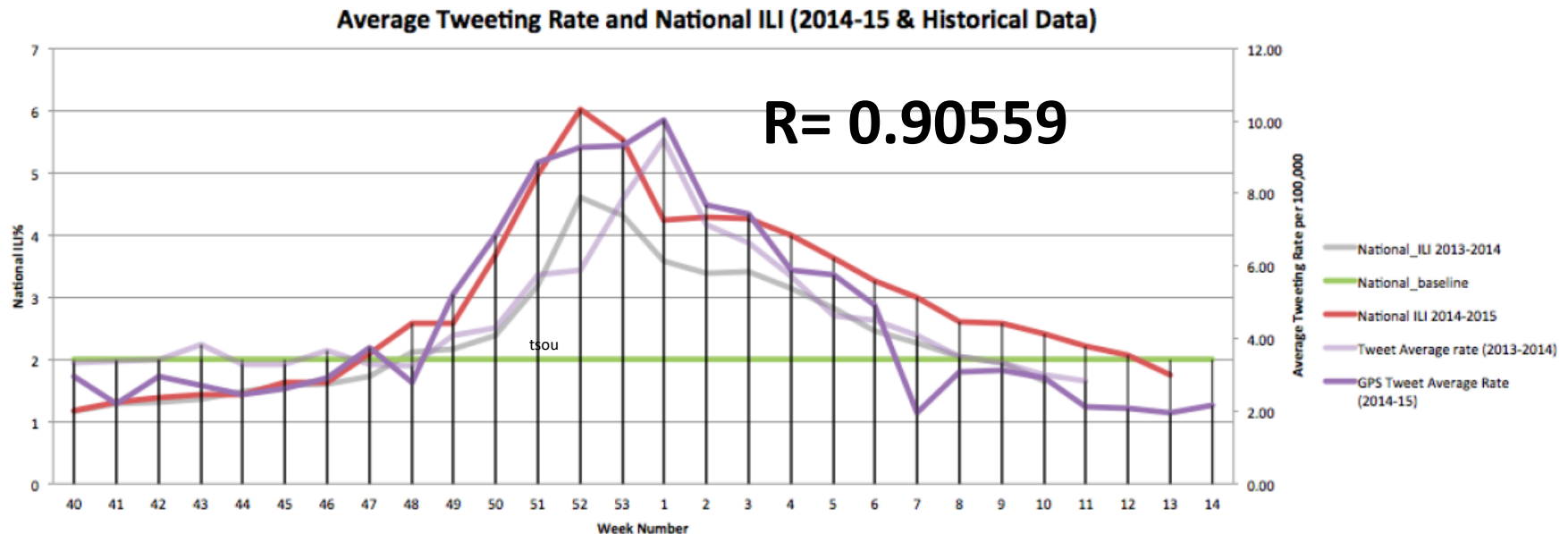


Figure 1. The comparison between National ILI Rate and the 31 Cities Tweeting Rate, with prediction up to Week 14. **Red: National ILI**, **Purple: GPS Only Tweets Tweeting Rate, multiplied by 10 for 2014-2015.**

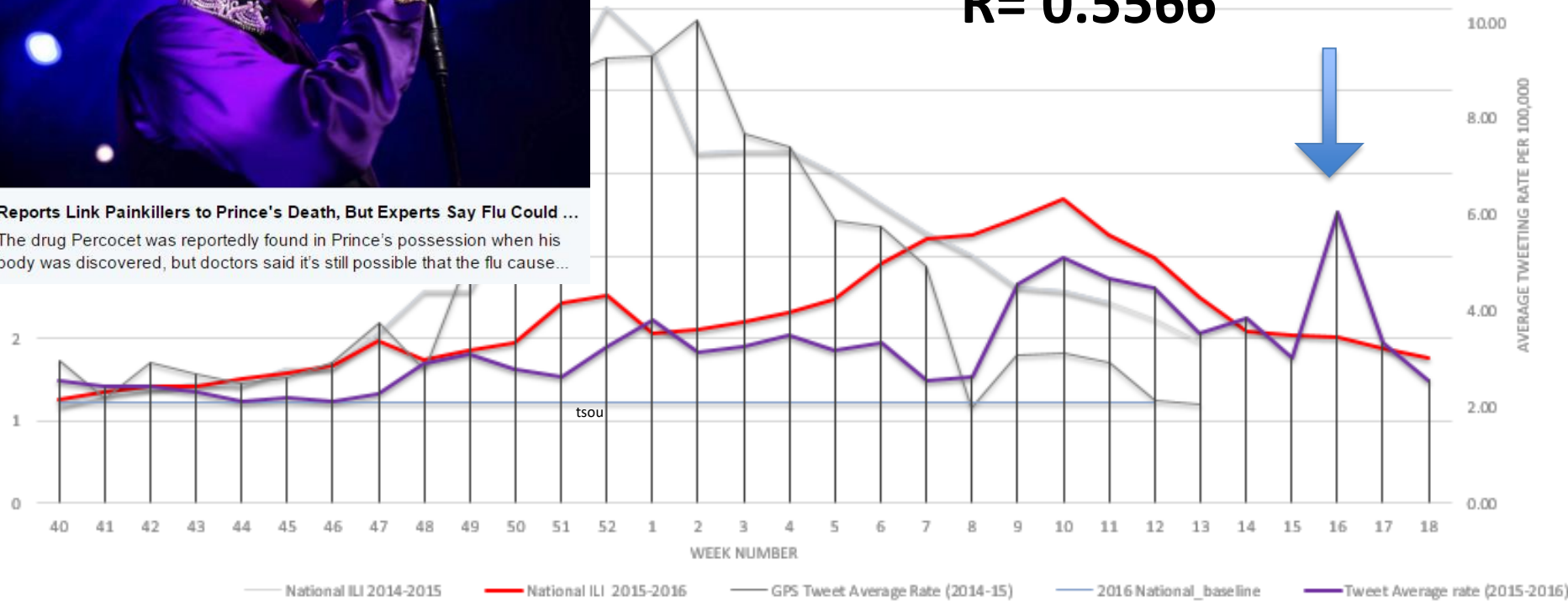
*\*NOTE: Week 7 GPS Tweet Average (2014-15) is missing tweets from 2/15-2/19 due to internal server error, the week 7 Tweeting rage is significantly lower and incomplete. Week 8 is back to the normal collection process.*

# 2016 Flu Tweets vs CDC ILI data



**Reports Link Painkillers to Prince's Death, But Experts Say Flu Could ...**  
 The drug Percocet was reportedly found in Prince's possession when his body was discovered, but doctors said it's still possible that the flu cause...

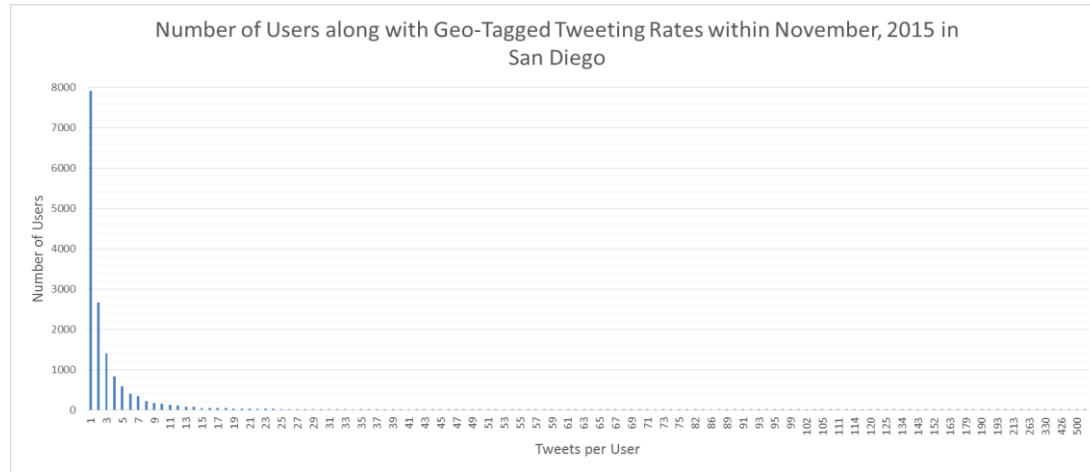
**R= 0.5566**



The comparison between National ILI Rate and the 32 Cities Tweeting Rate, with prediction up to Week 15. **Red National ILI, Purple Tweet Rate for 2015-2016.**

# Few Users with Big Voices

This Figure reveals the number of users along with their geo-tagged rates throughout the month of November, 2015. Over **7,900 users only had one tweet** during the whole month, which consists up to 49% of total users. More than 80% of Twitter users created less than 5 tweets in the whole month. But **1% of Twitter users created 23% of total Tweets**. Meanwhile, the person, who tweeted most in the month of November, sent out 903 tweets.



tsou

			Ratio of Users who tweeted		Percentage of tweets created by Top		
	Human Tweets	Human Users	1 time	1-5 times	1% active users	5% active users	10% active users
San Diego	69317	15916	49.00%	84.20%	22.80%	44.40%	56.80%
Columbus	29902	8758	58.00%	89.50%	24.40%	45.40%	56.60%

Table 1. A side by side comparison between San Diego and Columbus

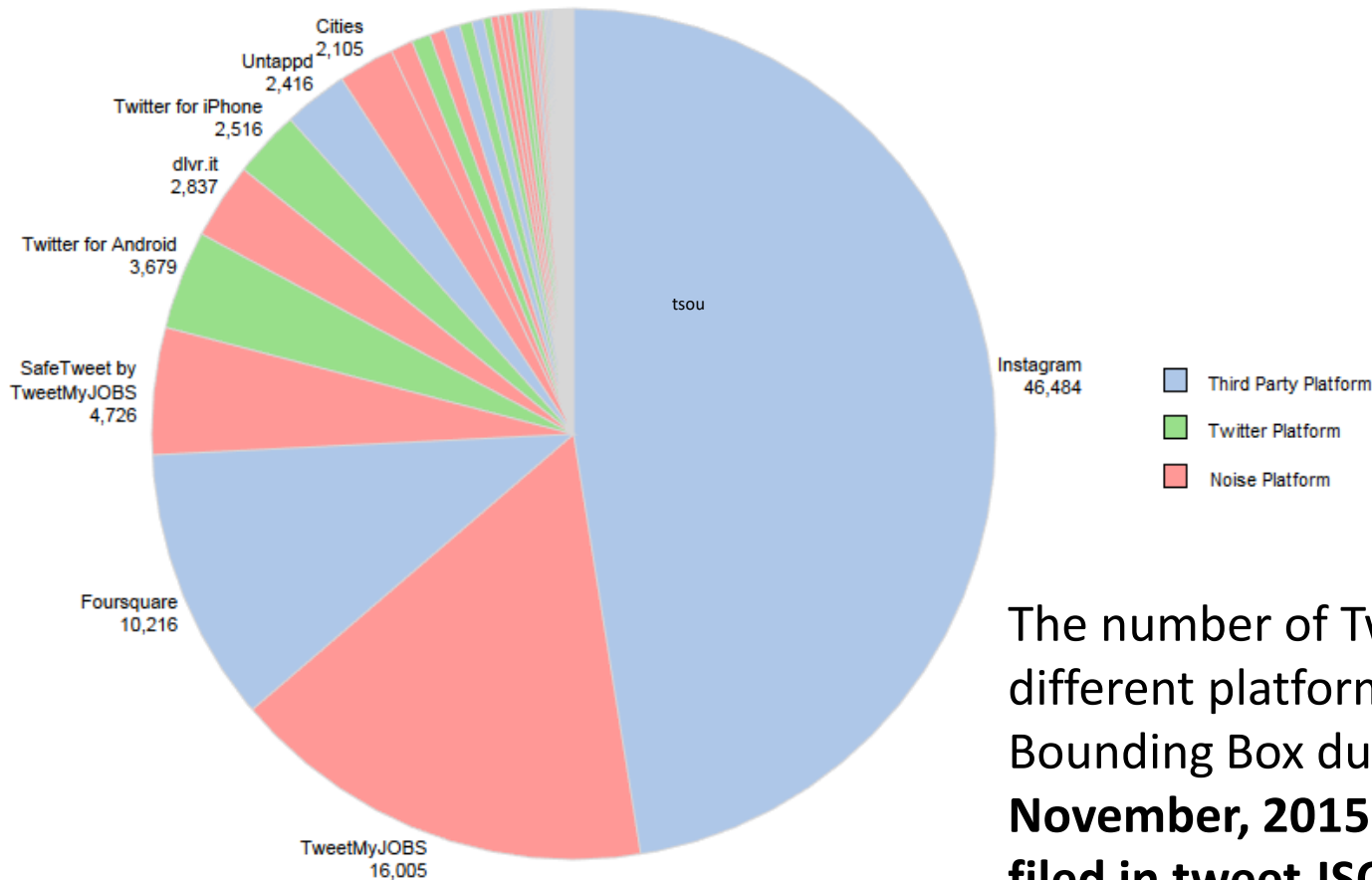
**How to adjust the “voices” to represent all users’ opinions?**

# Potential Errors and Noises in Geotagged Tweets

	Source category	Source name	Hashtag	Tweet number	Percentage
	Job	TweetMyJOBS		16005	
		SafeTweet by TweetMyJOBS		4726	
		CareerCenter		6	
<b>Total</b>				<b>20737</b>	<b>21.17%</b>
	Advertisement	dlvr.it		2837	
		Golfstar		269	
		dine here		182	
		Simply Best Coupons		77	
		Auto City Sales		56	
		sp_california	Coupon	41	
<b>Total</b>				<b>3421</b>	<b>3.49%</b>
	Weather	Cities		2105	
		iembot		24	
		Sandaysoft Cumulus		7	
<b>Total</b>				<b>2136</b>	<b>2.18%</b>
	Earthquake		Earthquake	762	
		everyEarthquake		203	
		EarthquakeTrack.com		69	
		QuakeSOS		9	
<b>Total</b>				<b>1043</b>	<b>1.06%</b>
	News	San Diego Trends		843	
		WordPress.com		111	
<b>Total</b>				<b>954</b>	<b>0.97%</b>
	Traffic	TTN SD traffic		512	
		TTN LA traffic		11	
<b>Total</b>				<b>523</b>	<b>0.53%</b>
				<b>Percentage of Noise:</b>	<b>29.42%</b>

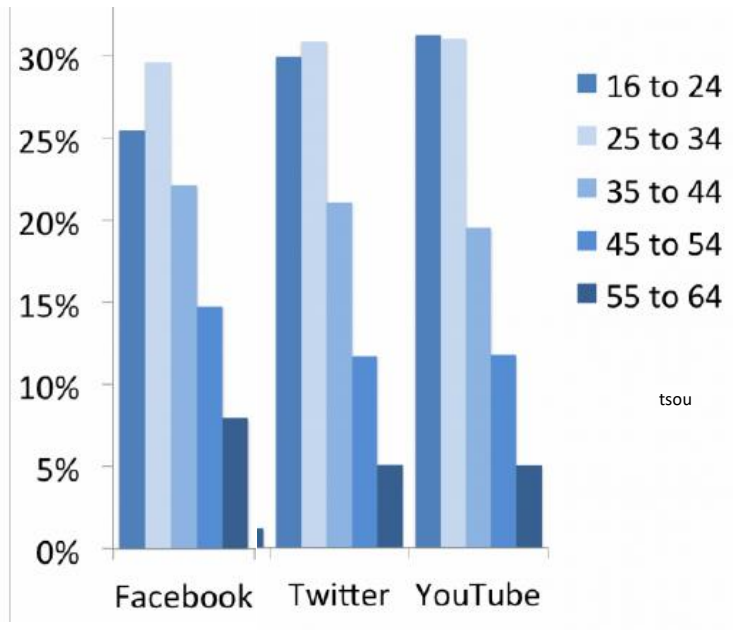


Detect **robot tweets** or **advertisement tweets** (noises) in geo-tagged tweets by examining the “**source**” metadata field. The portion of data noises is significant (**29.42%**) in our case study.



The number of Tweets produced by different platforms inside San Diego Bounding Box during the month of **November, 2015**. In the [Source] filed in tweet JSON documents.

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



2014 Survey (Business Insider)

Twitter Users are

- **Young** (60% are between 16 – 34 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” could be used as a sampling tool for social media users.**

## **Textbook: Chapter 2.**

Statistical Inference, Exploratory Data Analysis (EDA), and the Data Science Process

(O'Neil, C., & Schutt, R. (2013). ***Doing Data Science: Straight Talk from the Frontline.*** O'Reilly Media, Inc.

tsou



- “**Big Data is a point of view**, or philosophy, about **how decisions will be—** and perhaps should be— **made in the future.**” (Steve Lohr, The New York Times).
- Statistical inference is the process of **drawing conclusions** about **populations** or **scientific truths from data**. There are many modes of **performing inference** including statistical modeling, data oriented strategies and explicit use of designs and randomization in analyses. (cited from <https://www.coursera.org/learn/statistical-inference>). Example: **predicting presidential election results** or **weather prediction models**.
- Data **represents** the traces of the real-world **processes**, and exactly which traces we gather are decided by our data collection or **sampling method**. You, the data scientist, the observer, are turning the world into data, and **this is an utterly subjective, not objective, process**.
- Statistical inference is the discipline that concerns itself with the development of **procedures, methods, and theorems** that allow us to extract meaning and information from data that has been generated by **stochastic (random) processes**.

- **$N$**  represents the **total number of observations** in the **population**. (Population is **the entire collection** of similar items or events which can be used to answer research questions or hypothesis) (modified from multiple online definitions).
- When we take a **sample**, we take **a subset of the units of size  $n$**  in order to examine the observations to **draw conclusions and make inferences** about the population.
- The sampling mechanism can introduce **biases** into the data, and distort it, so that the subset is not a “mini-me” shrunk-down version of the population.
- **Biases** (major problems in the Twitter Data Analytics<sup>tsou</sup>) mentioned before.
  - Discussion: Any other Biases in Twitter Data? Or Facebook Data or Instagram Data or Yelp Data?
- The uncertainty created by such a sampling process has a name: **the sampling distribution**.
- **Different types of data will need different sampling methods.**
- Big Data Can Mean Big Assumptions.

- **How much data you need to sample really depends on what your goal is.**
- Examples in analyzing Twitter messages during Hurricane Sandy: The only conclusion you can actually draw is that this is what Hurricane Sandy was like for the subset of Twitter users (who themselves are not representative of the general US population), whose situation was ~~not~~ so bad that they didn't have time to tweet. **(Any other examples? Wildfire Tweets in San Diego?)**
- **Can N = ALL ?**
- **(Not Really) – Election polls example. Does everyone vote?**  
tsou
- **Data is no objective!**
- **Data doesn't speak for itself! (Data needs "data scientists" (human beings) to analyze and explain.)**



- **A model is our attempt to understand and represent the nature of reality** through a particular lens, be it architectural, biological, or mathematical. A model is an artificial construction where all extraneous detail has been removed or abstracted. (Examples: GIS data model: vector data vs. raster data, or statistical models: linear relationship  $\rightarrow Y = aX + b$  )
- **Probability distributions** are the foundation of statistical models.
- The classical example of probability distribution is **the height of humans**,
  - following a **normal distribution**—a bell-shaped curve, also called a **Gaussian distribution**, named after Gauss.
  - (Is the Age of humans a normal distribution? Are the housing prices in San Diego a normal distribution? )
- Not *all* processes generate data that looks like a *named* distribution, but many do. We can use these functions as building blocks of our models.

## Different statistical models

### “probability distributions”

- Normal Distribution
- Chi-Square Distribution
- **Exponential Distribution**
- **Weibull Distribution** (many business models adopt this).
- **Power Law Distribution (Pareto distribution)**



Power-law (long tail – 80-20 rule) <sup>tsou</sup>

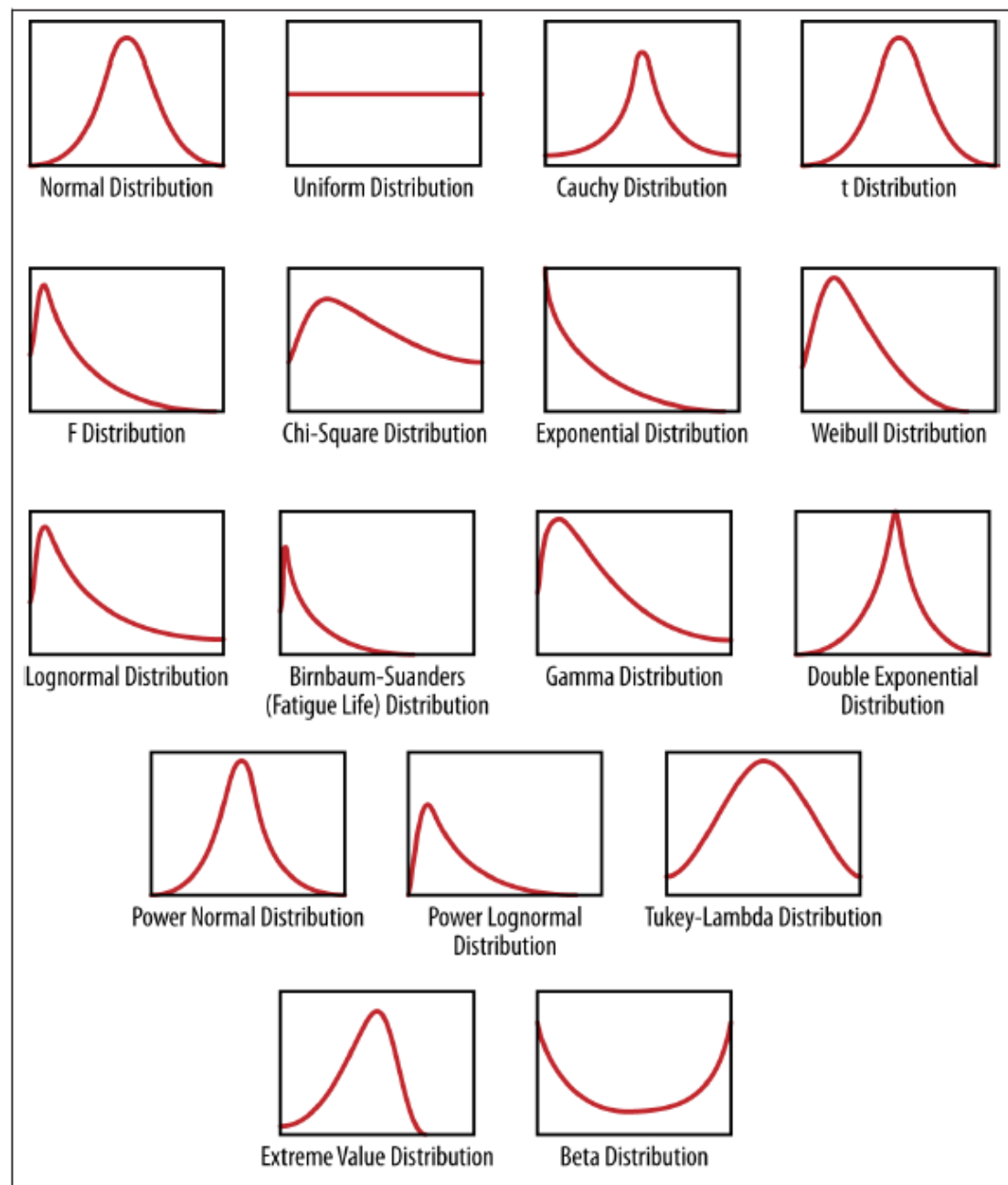
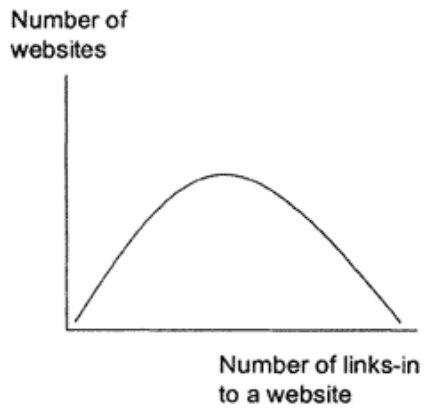


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

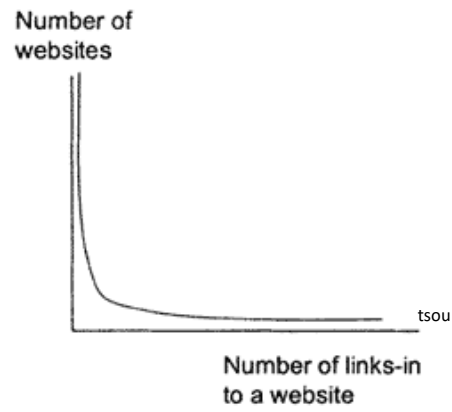
# The differences between Power Law Distribution vs. Exponential Distribution

power law:  $y = x^{(\text{constant})}$

exponential:  $y = (\text{constant})^x$



Normal distribution: A small number of sites have few or no links, a large number of sites have a moderate number of links into them, and a small number have a large number of links pointing to them.



Power law distribution: A very large number of sites have few or no links, a small number have a moderate number of links into them, a tiny number have a very large number of links pointing to them.

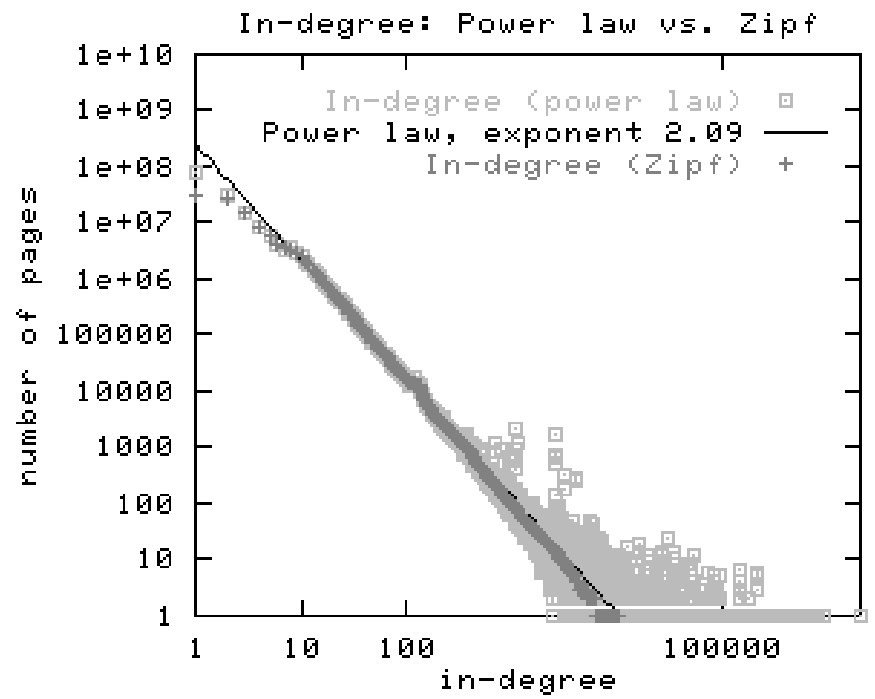


Image source: <http://www.climate-change-two.net/wealth-of-networks/ch-07.htm>

- T-test** for testing and validating the value collected from **small samples** (sub-group) from the total population. (variable should be “numerical”). Degree of freedom =  $n$  (sub-group numbers) - 1 (two tails or one tail). Such as the average testing scores in one class comparing the whole grades in a high school. **Examples: student average GPA in this class – comparing to the whole university (total population).**
- Chi-square test  $\chi^2$**  (for **categorical (nominal) data**) to compare two samples (or one sample with the expected values) and their variations.
  - $\chi^2 = \text{Sum (square[Ob. - Ex.] / Ex. )}$  (image from Wikipedia).
 

tsou

### Calculating the test-statistic [\[edit\]](#)

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^n \frac{(O_i/N - p_i)^2}{p_i}$$

where

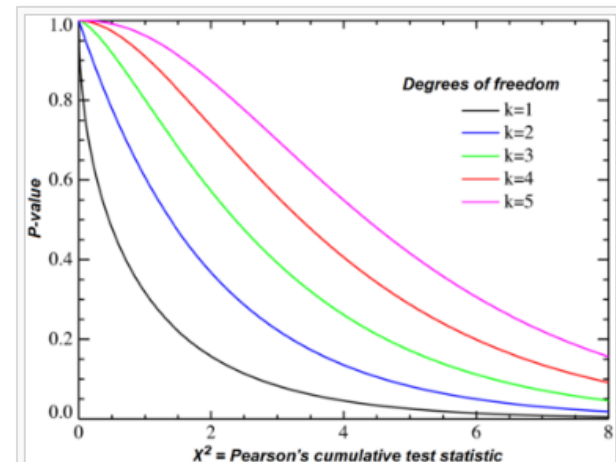
$\chi^2$  = Pearson's cumulative test statistic, which asymptotically approaches a  $\chi^2$  distribution.


$O_i$  = the number of observations of type  $i$ .

$N$  = total number of observations

$E_i = Np_i$  = the expected (theoretical) frequency of type  $i$ , asserted by the null hypothesis that the fraction of type  $i$  in the population is  $p_i$

$n$  = the number of cells in the table.



Chi-squared distribution, showing  $\chi^2$  on the x-axis and P-value on the y-axis. 

*Measurement level (scale):*

**Nominal**  
**(categorical)**  
Male/Female

**Ordinal**  
**(rank – order)**  
Gold/Silver/Brown

**Interval/ratio**  
**(numerical)**  
Height/ Revenue

tsou

*Statistical descriptor:*

***Mode***

***Median***

***Mean***

*Statistical testing*

***Chi-square Test***

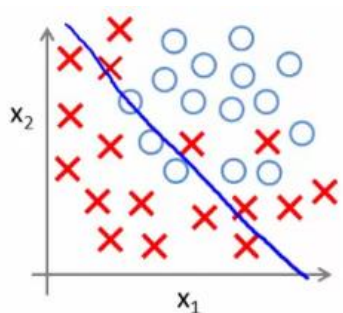
***Chi-square Test?***

***T-test or ANOVA***

***Logistic regression***

***correlation,  
regression***

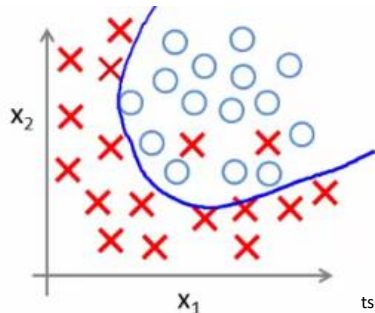
- Fitting a model** means that you estimate the parameters of the model using the observed data. You are using your data as evidence to help approximate the real-world mathematical process that generated the data. Fitting the model often involves **optimization methods** and **algorithms**, such as *maximum likelihood estimation*, to help get the parameters. (example: linear relationship  $Y = 3 + 5X$ ).
- Overfitting:** Overfitting is the term used to mean that you used a dataset to estimate the parameters of your model, but **your model isn't that good at capturing reality beyond your sampled data.**



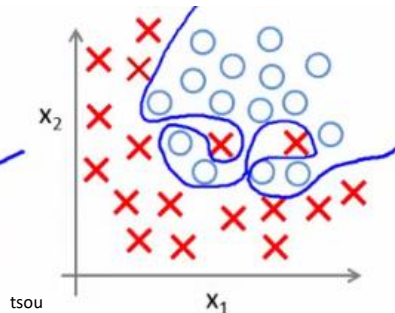
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

( $g$  = sigmoid function)

**UNDERFITTING**  
 (high bias)



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

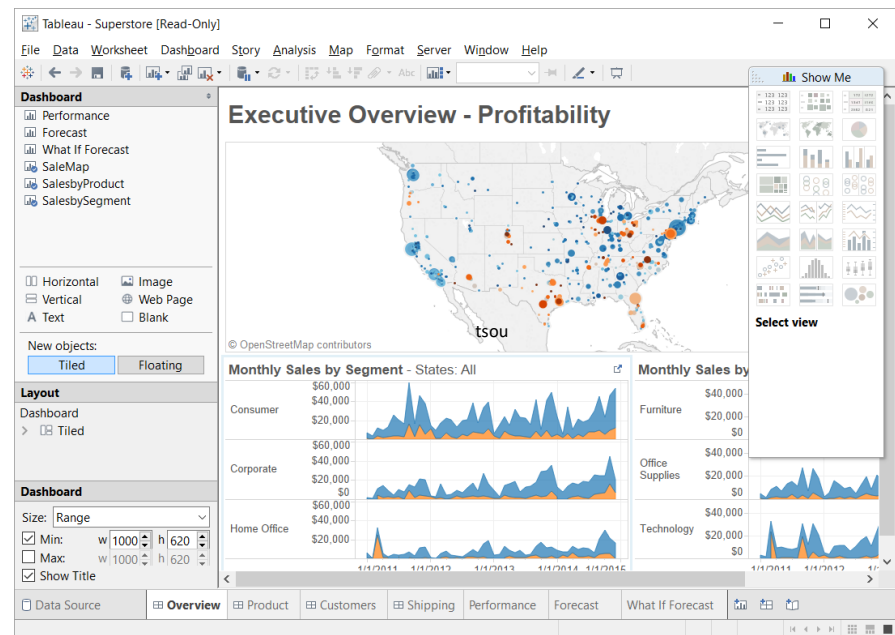
**OVERFITTING**  
 (high variance)

Image source:  
[http://www.holehouse.org/mlclass/07\\_Regularization.html](http://www.holehouse.org/mlclass/07_Regularization.html)



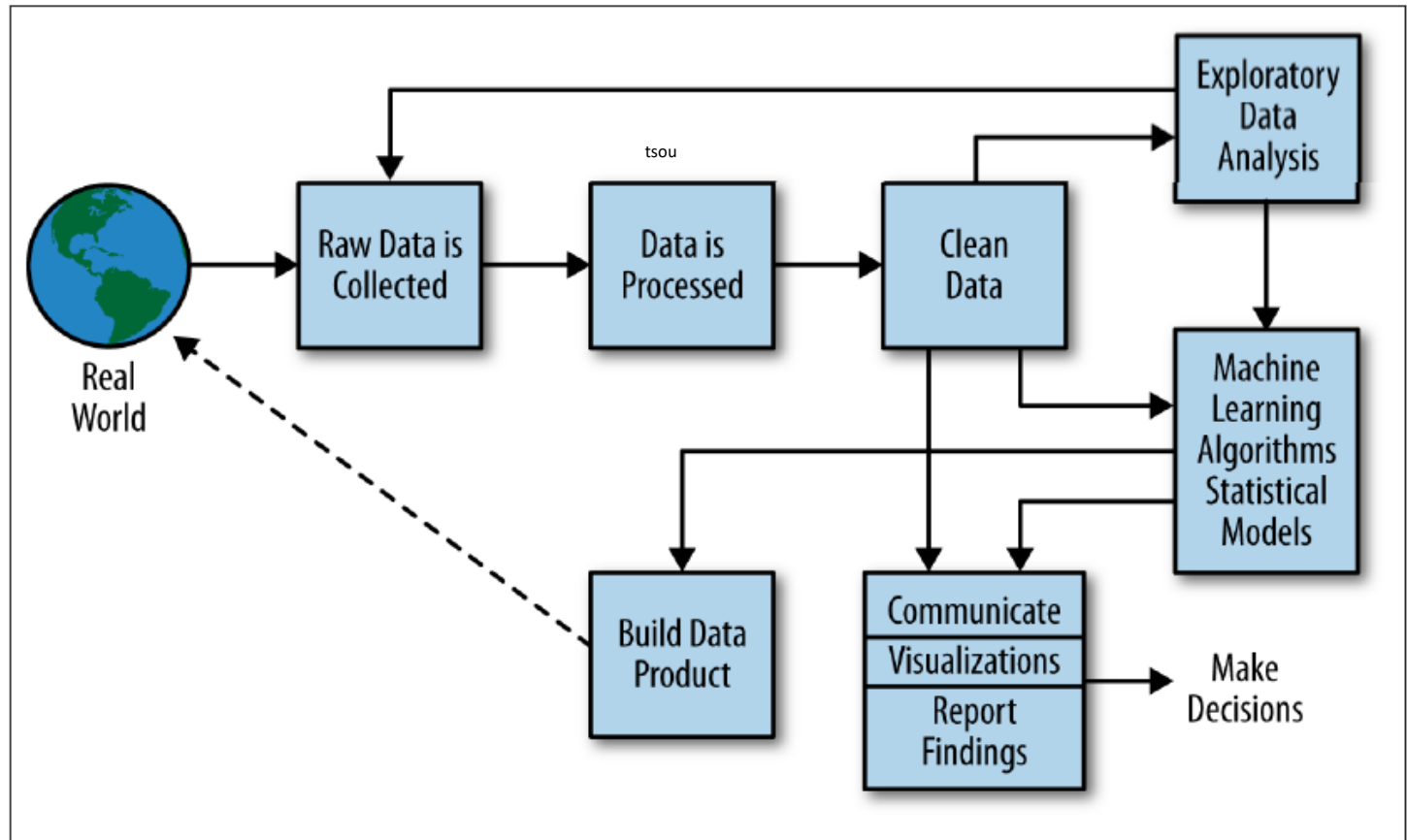
- Exploratory data analysis (EDA) as **the first step** toward building a statistical model.
- In EDA, **there is no hypothesis and there is no model**. The “exploratory” aspect means that your understanding of the problem you are solving, or might solve, is changing as you go.
- **The basic tools of EDA are plots, graphs and summary statistics.**
- You want to understand the data—**gain intuition, understand the shape of it,** and try to **connect your understanding of the process** that generated the data to the data itself.

## Example: Tableau Software



# The Data Science Process

Let's put it all together into what we define as the data science process. The more examples you see of people doing data science, the more you'll find that they fit into the general framework shown in **Figure 2-2**. As we go through the book, we'll revisit stages of this process and examples of it in different ways.



- Our goal may be to build or prototype a “**data product**”; e.g., a spam classifier, or a search ranking algorithm, or a recommendation system. Now the key here that makes data science special and distinct from statistics is that this **data product then *gets incorporated back* into the real world**, and users interact with that product, and that **generates more data**, which creates a feedback loop. (Examples: Stock Market Analysis, Housing Price from Zillow.com). tsou
- Human Dynamics → Enable the “**feedback loop**” from data product to users and from users to data product.
- **A Data Scientist’s Role in This Process:** Data Scientists have to make the decisions about what data to collect, and why. They need to be formulating questions and hypotheses and making a plan for how the problem will be attacked.

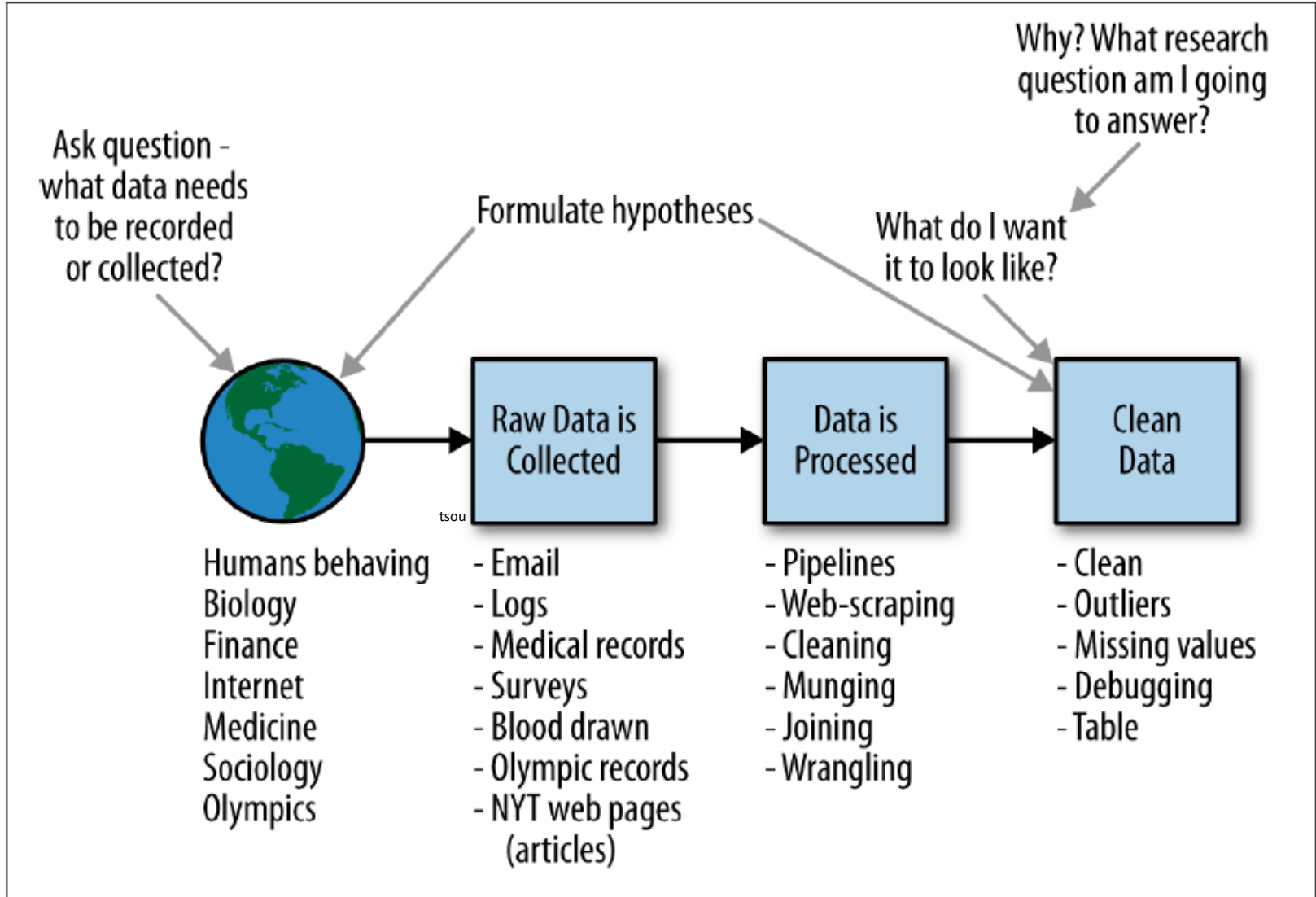


Figure 2-3. The data scientist is involved in every part of this process

- **Ask a question.** (WHY? What? How? When? Where?)
- **Do background research.** (Anyone has analyzed this types of data before?)
- **Construct a hypothesis** (to support your research goals or help you to answer the questions).
- **Test your hypothesis by doing an experiment.**  
(Choose which methods or models to test...)
- **Analyze your data and draw a conclusion.**
- **Communicate your results** (Visualization, Statistic Finding – Who are your audience? ).

## Additional Reading (Unit-2):

Lohr, Steve (2014). In Big Data, Shepherding Comes First. The New York Times, 12/15/2014.

(URL: <http://www.nytimes.com/2014/12/15/technology/in-big-data-shepherding-comes-first.html>) .



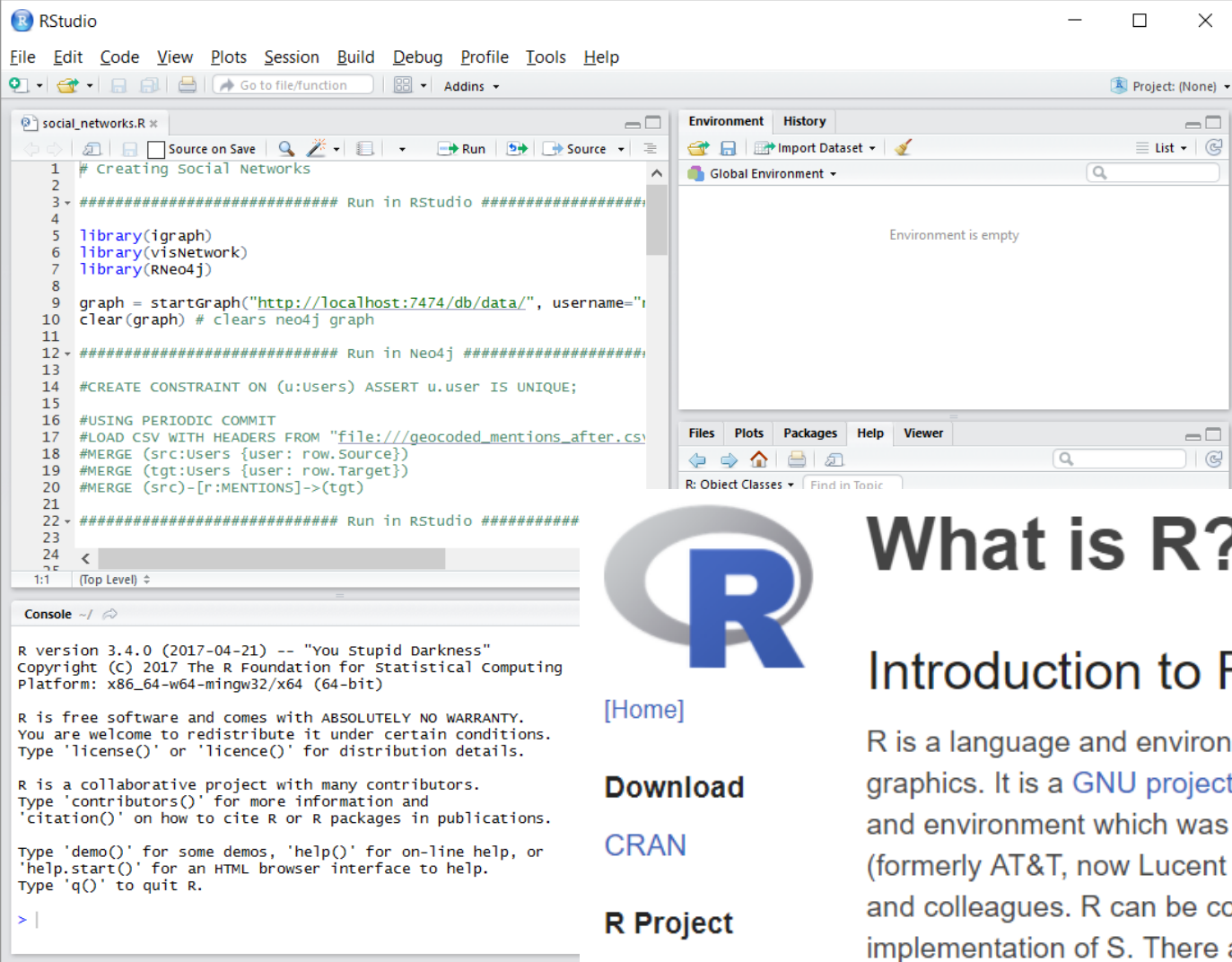
## Key points:

- **Building big data businesses** is proving to be anything but a get rich quick game, and to **require both agility and patience**.
- Companies knew they had a problem, knew they had data, but **not how to devise projects to explore** and experiment with data. “So we had to move up to a higher level with clients to work on data strategy, identifying a road map.
- The programmers that work in banks, retailers, health care providers, media companies and elsewhere will be critical. “The industry experts will be the ones building these new applications. (**Requiring Domain Knowledge**).
- Revenue is coming from helping corporate customers start writing big data applications. Cask, he said, works with corporate developers, often **building the first half of a pilot project and handing off the second half of the project to them**.

# Questions & Answers ?

# Web Exercise-02:

Introduction of R and R Studio



# R and RStudio



## What is R?

### Introduction to R

[\[Home\]](#)

[Download](#)

[CRAN](#)

[R Project](#)

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting](#)

[Bugs](#)

[Development](#)

[Site](#)

R is a language and environment for statistical computing and graphics. It is a [GNU project](#) which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.