Overview

- Social media usage in the United States

- What do people disclose on social media?

- How do we collect social media data?

- Example uses of social media data as a surveillance tool
Social Media Usage

- Nearly two-thirds of American adults (65%) use social networking sites

- Young adults (ages 18 to 29) are more likely to use social media – 90%
  - 35% of those 65 and older report using social media.

- Women and men use social media at similar rates
Social Media Usage

- 56% of those living in the lowest-income households use social media.

- There are not notable differences by racial or ethnic group:
  - 65% of whites
  - 65% of Hispanics
  - 56% of African-Americans
Social Media Usage

- Those who live in rural areas are less likely than those in suburban and urban communities to use social media.
  - 58% of rural residents
  - 68% of suburban residents
  - 64% of urban residents

Social Media Usage

- 64% patients in Community IOP/OP Treatment Centers (Philadelphia) use social media
  - 50% Daily
- Majority have multiple social media accounts:
  - 67% Facebook
  - 39% Instagram
  - 19% Twitter
What Do We Talk About on Social Media

❖ Social media and Google search query data correlates with:
  • Tobacco tax avoidance
  • Electronic cigarettes usage
  • Alcohol consumption
  • Drug usage
  • Emerging drugs of abuse (e.g., synthetic marijuana, bath salts)
Why Use Social Media as a Data Source

- Sheer volume of social media content provides a glimpse into the thoughts of the general population.

- Prior studies have demonstrated that social media data provides an early indication of health trends.

- Unlike surveys, there is no respondent burden...passive collection of data.
Collecting Social Media Data

1. Identify the substance(s) of interest.
2. Mine the stream of search queries and social media postings.
   - Who is using it? (demographics)
   - How much are people talking about this?
     - How has this varied over time?
   - How is the substance portrayed? (pos/neg)
   - Where are people obtaining it?
   - What types of questions do people have about it?
   - Are people asking for help to stop using it?
Collecting Social Media Data

- Social media platforms, such as Twitter and Facebook, are an attractive source of data for public health surveillance.
  - The majority of Twitter data are “public” meaning anyone can access it.
  - Twitter and Facebook data often are geo-coded meaning you can select specific location level data.

- Facebook requires users to “opt-in” to sharing their data
Collecting Social Media Data

♦ To collect social media data, you can use commercial vendors OR you can develop your own “API”

♦ The API, or Application Programming Interface, allows external developers to develop technology which rely on social media data.
  • APIs allow you to “pull” the data
Types of Twitter APIs

- **Firehose**: Returns 100% of the tweets that match your criteria.
  - Two data providers: GNIP and DataSift

- **Search API**: gives you access to a data set that already exists from tweets that have occurred

- **Streaming API**: gives access to tweets in near real-time.
  - With Twitter’s Streaming API, users register a set of criteria (keywords, usernames, locations, named places, etc.) and as tweets match the criteria, they are pushed directly to the user.
Twitter Streaming API

User makes request

Server pulls processed result from data store and renders view

Server opens streaming connection

Twitter accepts connection

Tweets streamed as they occur

Connection closes

Connection closes
Collecting Social Media Data

- With Twitter’s Streaming API you can expect to receive anywhere from 1% of the tweets to over 40% of tweets in near real-time.

- Real-time capabilities to monitor and analyze social media data using keywords, sentiment indicators, text analytics to identify relevant topics, and other data mining techniques.
  - Content can be tracked weekly, daily or even hourly.
Google Search Trend Examples
Example: Synthetic Marijuana Emerging Use

- Determine keywords typically used to find information about synthetic marijuana on the internet (Nov 28, 2011)

- Provided Google Adwords with the seed terms:
  - “synthetic marijuana,” “synthetic weed,” “K2 Spice,” and “herbal incense”

- Conducted searches with seed terms to validate “herbal incense” yielding most search results across three search engines
  - Google, Bing, and Yahoo

Curtis, B., et.al., 2015
Total Number of Results for Each of the Four Search Terms by Search

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Google</th>
<th>Yahoo</th>
<th>Bing</th>
</tr>
</thead>
<tbody>
<tr>
<td>“herbal incense”</td>
<td>2,730,000</td>
<td>1,550,000</td>
<td>1,500,000</td>
</tr>
<tr>
<td>“synthetic marijuana”</td>
<td>1,170,000</td>
<td>763,000</td>
<td>767,000</td>
</tr>
<tr>
<td>“K2 spice”</td>
<td>247,000</td>
<td>119,000</td>
<td>119,000</td>
</tr>
<tr>
<td>“synthetic weed”</td>
<td>122,000</td>
<td>39,100</td>
<td>37,100</td>
</tr>
</tbody>
</table>

Curtis, B., et.al., 2015
US Internet Search Interest in “Herbal Incense”
Jan 2008-Jan 2012 (Google Trends)

Google Trends

Curtis, B., et.al., 2015
Impact of States Legal Sanction Policies

Example of **strict** policy sanctions: Louisiana; illegal to have, use or sell synthetic marijuana

August 2010
Impact of States Legal Sanction Policies

Example of weaker policy sanctions: Missouri; illegal to possess synthetic marijuana
Example: Google Trends & Flakka (Philadelphia)
Twitter Examples
General Population: (over 218 million Twitter users)

Twitter Messages Containing "Drunk"

- 2012: 0.25%
- 2013: 0.20%
- 2014: 0.20%
- 2015: 0.15%
“Hey Everyone, I’m Drunk” An Evaluation of Drinking-Related Twitter Chatter

Universe of Tweets from March 13 to April 11, 2014
\[ N = \text{approximately } 15,000,000,000 \text{ Tweets} \]

Select Tweets with the terms “drunk,” “alcohol,” “beer,” “liquor,” “vodka,” “hangover” (each used in ≥500,000 Tweets in a month)
\[ N = 11,966,381 \text{ Tweets} \]

Restrict to Tweets in the top 25th percentile of both Klout score and number of followers
\[ N = 1,606,701 \text{ Tweets} \]

Random sample of 5,000 Tweets coded for themes

Cavazos-Rehg et. al., 2015
“Hey Everyone, I’m Drunk” An Evaluation of Drinking-Related Twitter Chatter

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Number of Tweets&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>drunk or #drunk&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5,336,372</td>
</tr>
<tr>
<td>beer or #beer&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3,444,778</td>
</tr>
<tr>
<td>alcohol or #alcohol&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1,565,258</td>
</tr>
<tr>
<td>vodka or #vodka&lt;sup&gt;e&lt;/sup&gt;</td>
<td>752,988</td>
</tr>
<tr>
<td>liquor or #liquor</td>
<td>566,266</td>
</tr>
<tr>
<td>hangover or #hangover&lt;sup&gt;f&lt;/sup&gt;</td>
<td>517,959</td>
</tr>
</tbody>
</table>

<sup>a</sup>Sum does not equal the total of 11,966,381 Tweets because some Tweets contained more than one term;<sup>b</sup>excluded Tweets with references to “drunk in/with love” (including the popular song by Beyoncé, “Drunk in Love”), “drunk in/with/on power,” and “punch drunk”;<sup>c</sup>excluded Tweets with references to “beer belly,” “root beer,” “beer batter,” “Madison Beer,” “beer cheese,” “beer bar neon (signs),” and “beer stein”;<sup>d</sup>excluded Tweets with references to hand sanitizer/gel (also Germ-X and Purell) and rubbing alcohol;<sup>e</sup>excluded Tweets with references to “vodka sauce”;<sup>f</sup>excluded Tweets referencing The Hangover series of movies.

Cavazos-Rehg et. al., 2015
“Hey Everyone, I’m Drunk” An Evaluation of Drinking-Related Twitter Chatter

Cavazos-Rehg et. al., 2015
Adderall: Nonmedical Use—College Students

Methods:
- Public-facing Twitter status messages containing the term “Adderall” were monitored from November 2011 to May 2012.
- GPS data were identified with clusters of nearby colleges and universities for regional comparison.

Results:
- 213,633 tweets from 132,099 unique user accounts mentioned “Adderall.”
- Adderall tweets peaked during traditional college and university final exam periods.
- Rates of Adderall tweeters were highest among college and university clusters in the northeast and south regions of the United States.
- 12.9% mentioned an alternative motive (eg, study aid) in the same tweet.
- The most common substances mentioned with Adderall were alcohol (4.8%) and stimulants (4.7%), and the most common side effects were sleep deprivation (5.0%) and loss of appetite (2.6%).

Hanson et al., 2013
Methods:

- Twitter statuses mentioning prescription drugs were collected from November 2011 to November 2012.
- From this set, 25 Twitter users were selected who discussed topics indicative of prescription drug abuse.
- Social circles of 100 people were discovered around each of these Twitter users; the tweets of the Twitter users in these networks were collected and analyzed according to prescription drug abuse discussion and interaction with other users about the topic.

Hanson et al., 2013
Prescription Drug Abuse & Twitter

Results:

- 3 million mentions of prescription drug terms were observed.
- For the 25 social circles, on average 53.96% of the Twitter users used prescription drug terms at least once in their posts.

<table>
<thead>
<tr>
<th>Table 3. Number of prescription drug tweets by drug category.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
</tr>
<tr>
<td>Drug total</td>
</tr>
<tr>
<td>Larger doses / overdose</td>
</tr>
<tr>
<td>Co-ingestion</td>
</tr>
<tr>
<td>More frequent doses</td>
</tr>
<tr>
<td>Alternative motives /</td>
</tr>
<tr>
<td>dependence</td>
</tr>
<tr>
<td>Alternative routes of</td>
</tr>
<tr>
<td>admission</td>
</tr>
<tr>
<td>Legitimacy of obtaining</td>
</tr>
<tr>
<td>Trading / selling</td>
</tr>
<tr>
<td>Seeking</td>
</tr>
</tbody>
</table>

- Strong correlation was found between the kinds of drugs mentioned by the index user and his or her network.

Hanson et al., 2013
Mining Twitter & Language Analysis
Using Social Media Text Data

- We use an open vocabulary approach and create data driven “topics”
- Topics are created through a process call Latent Dirichlet Allocation (LDA)
- The topics often give us face valid results, tie in with existing research and suggest new hypotheses

Schwartz, et al., 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLOS ONE, 8(9), e73791
Twitter Language Predicts Excessive Drinking

  - 50 states, plus DC, Puerto Rico, Guam and US Virgin Islands
  - Excessive drinking
    - Binge (drinking 5 or more drinks on an occasion for men or 4 or more drinks on an occasion for women)
    - Heavy Drinking (15 or more drinks per week for men or 8 or more drinks per week for women)

Twitter Dataset
- Twitter allows access to a random 1% of their streaming data
- Often tweets come with location information such as latitude / longitude or a user specified location
- We used this to map tweets to US counties, ~8% of total tweets can be mapped
- Twitter data set
  - Tweets from 2011 - 2013
  - 5.4 million tweets per month
  - 1.7 million users per month
Top 5 topics negatively correlated with Excess Drinking (Twitter alone)

\[ R = 0.441 \]

\[ R = 0.408 \]

\[ R = 0.399 \]

\[ R = 0.394 \]

\[ R = 0.387 \]
Top 5 topics **positively** correlated with Excess Drinking (Twitter alone)

\[ R = 0.320 \]
\[ R = 0.316 \]
\[ R = 0.303 \]
\[ R = 0.301 \]
\[ R = 0.287 \]
Can we predict Excess Drinking from language?

- Using a 10 fold, cross validated ridge regression, we predicted excess drinking from
  - Twitter language alone
  - Demographic and SES variables alone
  - Twitter, demographics and SES combined

<table>
<thead>
<tr>
<th></th>
<th>Twitter Alone</th>
<th>DEM + SES</th>
<th>Twitter + DEM + SES</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>0.645</td>
<td>0.635</td>
<td>0.676</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.42</td>
<td>0.40</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Possibilities

What Possibilities

Do you see?
Acknowledgements

Research Team:
- Drs. Kimberly Kirby, Lyle Ungar, Andy Schwartz & Anneke Buffone (investigators)
- Sal Giorgi & Sphoorti Mirji (developers)
- Casey Hamilton & Dan Summers (research assistants)
- Center on the Continuum of Care in the Addictions (Penn)
- Center for Studies of Addiction (Penn)

Research Funding Source:
- NIH/NIDA grant: R01 DA039457
Possibilities

What Possibilities
Do you see?